Universitat Autònoma de Barcelona

Departament de Bioquímica i Biologia Molecular

Towards Objective Human Brain Tumours Classification using DNA microarrays

Dissertation for the degree of
Doctor of Biochemistry and Molecular Biology
presented by

Xavier Castells Domingo

This work was performed at the Department of Biochemistry and Molecular Biology of the Universitat Autònoma de Barcelona under the supervision of Dr. Carles Arús Caraltó, Dr. Joaquín Ariño Carmona and Dr. Anna Barceló Vernet

Undersigned by

Dr. Carles Arús Caraltó

Dr. Joaquín Ariño Carmona

Dr. Anna Barceló Vernet

Xavier Castells Domingo

Cerdanyola del Vallès, 7 May 2009

En agraïment a totes les persones que van decidir
donar un tros de biòpsia al noste grup.
Especialment vull dedicar aquesta tesi a la memòria de
les persones que han mort durant el transcurs
de la meva tesi, i de les quals he tingut el gran honor de
poder extreure RNA de les seves biòpsies.

Contents

List of Figures

List of Tables

Abbreviations

Ag Anaplastic glioma

a.u. Arbitrary units

AUC Area under the curve

cDNA Complementary DNA

CGH Comparative genomic hybridisation

CIBER-BNN Centro de Investigación Biomédica en Red-Bioingeniería, Biomateriales y Nanotecnología

CNIO Centro Nacional de Investigaciones Oncológicas

CNS Central nervous system

COT1 Cotyledon trichome 1

Cr Creatine

cRNA Complementary RNA

CV Coefficient of variation

Cy3 Cyanine 3

Cy5 Cyanine 5

DAVID Database for Annotation, Visualization and Integrated Discovery

DMEM Dubelcco's modified Eagle's medium

DNA Deoxyribonucleic acid

DSP Desmoplakin

dUTP Deoxy-uridine tryphosphate

ECD eTUMOUR consensus diagnosis

EGF Epidermal growth factor

EGFR Epidermal growth factor receptor

EGFRvIII EGFR variant 3

EtBr Ethidium bromide

EU European union

FDR False discovery rate

FN False negative

Fn14 Type 1 transmembrane protein Fn14 (TWEAK receptor)

FNR False negative rate

FP False positive

FPR False positive rate

FWER Family-wise error rate

GABRMN Grup d'Aplicacions Biomèdiques de la Ressonància Magnètica Nuclear

GAGs Glioblastoma-associated genes

GAPDH Glyceraldehyde-3-phosphate deshydrogenase

Gb Glioblastoma

Gbm Glioblastoma multiforme

GCOS GeneChip Operating System

GEM Genetically engineered mice

GEO Gene Expression Omnibus

GFAP Glial fibrillary acidic protein

Gly Glycine

GPM6B Glycoprotein M6B

HBT Human brain tumour

HRMAS High resolution magic angle spinning spectroscopy

ICD-O International Classification of Diseases for Oncology

iNOS Inducible nitric oxid synthase

JAK Janus kinase

KUL Katholieke Universiteit Leuven

LDA Linear discriminant analysis

Lgg Low-grade glioma

limma Linear models for microarray data

LOH Loss of heterezigosity

LOOCV Leave-one-out cross validation

MAPK Mitogen-activated protein kinase

MAQC MicroArray Quality Control

MAS Microarray Suite

MCV Monte Carlo cross validation

MEDIVO2 Mejora del diagnóstico y de la valoración pronóstica de tumores cerebrales humanos in vivo. Modelos animales y celulares para la metabolómica de la progresión tumoral. Fase 2

Mg Meningioma

mIno myo-Inositol

Mm Meningothelial meningioma

MM Miss match

MRI Magnetic resonance imaging

MRS Magnetic resonance spectroscopy

mTOR Mammalian target of rapamycin

MW Mann-Whitney

NMR Nuclear magnetic resonance

NF-κB Necrosis factor-κB

NF Normalisation factor

NF2 Neurofibromatosis 2

NSCP Neural stem cell progenitor

NUSE Normalised unscaled standard errors

OPD Originating pathologist diagnosis

PBS Phosphate buffer saline

PCA Principal component analysis

PCR Polymerase chain reaction

PDGF Platelet derived growth factor

PDGFR Platelet derived growth factor receptor

pFDR Positive false discovery rate

Phospholipase C-γ PLC-γ

PI3K Phosphatidylinositol 3-kinase

PIP3 Phosphatidylinositol 3-phosphate

PM Perfect match

PRELP Proline arginine-rich end leucine-rich repeat protein

PRTPRZ1 Proteine tyrosine phosphatase receptor-type Z1

PTEN Phosphatase and tensin homolog

randF Random forest

RIN RNA integrity number

RLE Relative log expression

RMA Robust microarray analysis

RNA Ribonucleic acid

rRNA Ribosomal RNA

RT-PCR Real time-PCR

RuBisCO Ribulose-1,5-bisphosphate carboxylase/oxygenase

SGHMS Saint George's Hospital Medical School

SLRPs Small leucine rich proteoglycans

SS Split sample

SSC Sodium saline citrate

STAT Signal transducer and activator of transcription

SVM Support vector machine

SW Stepwise

TGT Target value

TN True positive

TNF Tumour necrosis factor

TP True negative

TP53 Tumour protein 53

tRNA Transference RNA

TWEAK TNF-related weak inducer of apoptosis

UAB Universitat Autònoma de Barcelona

UPC Universitat Politècnica de Catalunya

UPVLC Universitat Politècnica de València

UV Ultra-violet

VFCV V-fold cross validation

WHO World Health Organization

WP Work package

5FCV 5-fold cross validation

Chapter 1

INTRODUCTION

1.1 Overview on human brain tumours (HBT)

1.1.1 Incidence and mortality of HBT

Cancer genetics and cancer progression are processes not completely understood yet. Genetic and epigenetic alterations result in a molecular cascade of events leading to several subtypes of tumours[2]. HBT are not an exception. Its study at the molecular level still remains poorly developed, even though great improvement has been achieved lately[2, 3].

By definition, HBT are tumours arising from central nervous system cells (CNS), or their meningeal covering[3]. Diagnoses of these tumours are devastating for the patient, and prognoses are yet difficult to determine[2, 4, 5, 6]. In 2002, the world-wide incidence of HBT was 189485 cases, and 141650 the estimated mortality[7], which represents the 1.7% of new cancers and 2.1% of cancer deaths, respectively. The highest rates are observed in developed areas (Australia/New Zealand, Europe, and North America) and the lowest in Africa and the Pacific islands[7]. In Europe, the incidence in 2004 was 37200 cases and 28600 the estimated mortality[8]. Both incidence and mortality are higher in males than in females[7, 8].

Although HBT are less common than other cancers, interest in these tumours has increased in the preceding years[3] (see figure 1.1). This is explained by four main reasons:

1. Nowadays, HBT are the leading cause of death from cancer in children under the age of 15, and the second one in young people from 15 to 34[3, 9].
2. HBT are among the most aggressive and intractable tumours[6].
3. Neuroncology is increasingly attracting the interest of neuroscientists, typically focused in the past on neurodegenerative diseases and developmental disorders[3].
4. The aging population in developed countries increases incidence of adult brain cancers.

1.1.2 Description of HBT

1.1.2.1 Diagnosis of HBT in current clinical practice

Commonly HBT are detected by neurological examination, unfortunately once the tumour mass is widely developed[10]. The difficulty of HBT diagnosis relies on the

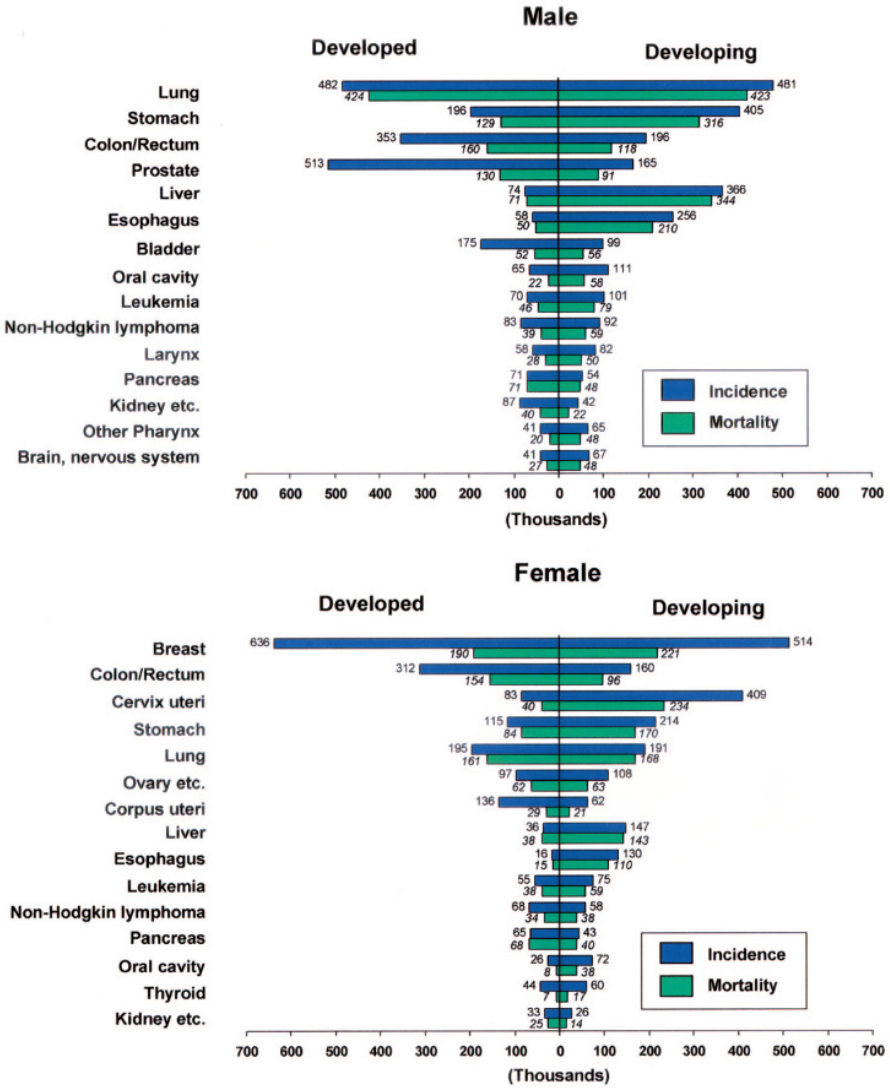


Figure 1.1: *Worldwide cancer statistics from GLOBALCAN 2002.* Estimated incidence and mortality of cancer types for developed and developing countries. Separated statistics by sex are shown. Figure extracted from reference [7].

lack of early symptoms detectable by clinicians. Typical symptoms in advanced stages are headache, seizures, fatigue and focal deficits[4, 10]. Currently, diagnosis of HBT is assessed in a first step by magnetic resonance imaging (MRI)[11]. In the clinical practice, this examination enables to delimit the extent of the tumour mass and to determine relevant morphological parameters to propose a preliminary diagnosis. Nevertheless, diagnosis is usually confirmed by histopathological examination, which is considered the gold-standard to classify HBT[12].

1.1.2.2 Overview on HBT classification

HBT are classified by the WHO in twelve histological types[13]. As can be seen in figure 1.2, there are three tumours types accounting in average for 75% of the prevalence of HBTs: gliomas (45%), meningiomas (15%) and metastasis (15%). The complete list of HBTs is depicted in table 1.1.

Although HBT account for a relative small percentage of worldwide cancer cases, their usual malignant transformation and their dramatic clinical course for both the patient and its family have driven to thorough investigation to improve patient healthcare. As in other cancers, the formation of HBT is a complex process involving an accumulation of genetic alterations[14]. Since specific alterations have been described in almost each HBT, the concept of different molecular pathways leading to different types of tumors has gained general acceptance. This means that molecular information from surgically resected tumours may have diagnostic value. Even more, it could be a substitute for the traditional histopathological diagnosis, if some technical and ethical problems could be overcome. A more extensive explanation on this topic is developed in section 1.5.

1.1.3 World Health Organization (WHO) classification criteria

1.1.3.1 Historical overview

The international classification of human tumours published by the WHO was promoted during the decade of 1950s with a clear objective valid until today: to establish a classification and grading of human tumours that is accepted worldwide. Definition of histological and clinical diagnostic criteria were indispensable for epidemiological studies and clinical trials to be conducted beyond institutional and national boundaries. The WHO publishes the classification of HBT and other tumour types in the WHO Blue Book Series[13].

Since the first edition on the histological typing of tumours of the nervous system appeared in 1979, posterior editions have progressively incorporated immunochemistry and genetics profiles into diagnostic[16, 17]. Precisely, the third edition published in 2000 included concise sections on epidemiology, clinical signs and symptoms, imaging, prognosis and predictive factors[18]. The classification was based on the consensus of an international Working Group in all editions[19].

The WHO classification of HBT covers tumours of the central system, including

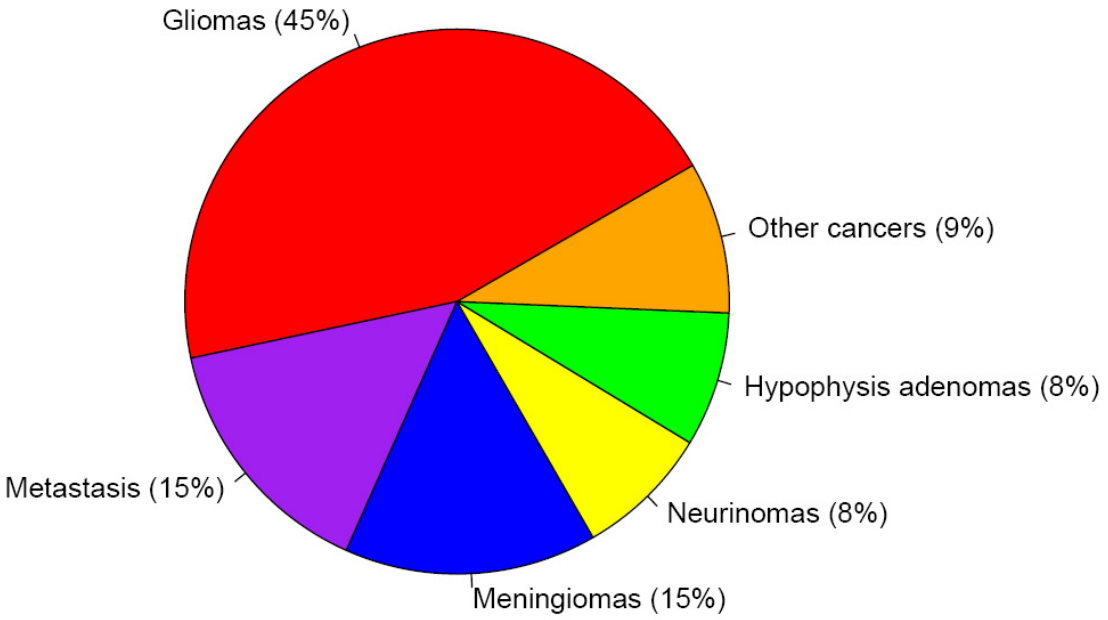


Figure 1.2: *Prevalence of HBT types.* Estimated prevalence of the most frequents HBT types. Gliomas comprise almost half of the diagnosed cases. Meningiomas and metastasis are the second tumours most frequently diagnosed with similar prevalence. Also, neurinomas and hypophysis adenomas have a similar prevalence. Data obtained from reference [15].

tumours of cranial and paraspinal nerves, whereas those of the peripheral nervous system are covered in other volumes of the WHO classification books. In the context of the WHO, the International Classification of Diseases for Oncology (ICD-O) assures that incidence and mortality data of population stratified histopathologically become available for epidemiological and oncological studies. The codes of ICD-O are an interface between pathologists and cancer registries. The fourth edition of the Blue Book Series, which appeared in 2007, introduced preliminary codes for several new entities, variants and patterns[19].

1.1.3.2 Entities, variants and patterns

A new **entity** must be characterised by distinctive morphology, location, age distribution and biologic behaviour, and not simply by an unusual histopathological pattern. **Variants** were defined as being reliably identified histologically and having some relevance for clinical outcome, but as still being part of a previously defined, overarching entity. Finally, **patterns** of differentiation were considered identifiable histological appearances, but did not have a clinical or pathological significance. Two or more reports from different institutions were considered mandatory for a new entity, variant or pattern to be included in the WHO classification. The fourth edition includes 8 new entities, 3 new variant and 2 patterns of differentiation[19].

Histological grading is a means of predicting the biological behaviour of a neoplasm with the purpose of facilitating the choice of a therapy and predicting the outcome. Each human tumour has an associated malignancy grade (see table 1.1). The grading scheme included in the WHO classification of tumours of the nervous system accounts for the malignancy of the neoplasms rather than being a strict histological grading system[16, 20]. The WHO classification is widely used, but not a mandatory application.

1.1.3.3 Malignancy grade schemes

The assignment of malignancy grades for the new entities proposed by the WHO Working Group in the 2007 edition remains preliminary, since publication of additional data and long-term follow-up is pending[13]. Grade I applies to potentially proliferative lesions and the possibility of cure following surgical resection alone. Grade II are generally infiltrative tumours in nature and tend to progress to higher grades of malignancy. Grade III are commonly lesions with histological evidence of malignancy, including atypia and brisk mitotic activity. Diagnosed patients with a tumour of this grade receive adjuvant radiation and/or chemotherapy. WHO grade IV is assigned to cytologically malignant, mitotically active, necrosis-prone neoplasms typically associated with rapid pre- and post-surgery disease evolution and a fatal outcome[13].

1.1.3.4 Differences between classification schemes

Being the WHO classification the most widely used in the clinical practise, there are other valid classification schemes that slightly differs from the WHO one. In particular, the WHO 2007 scheme grades diffusely infiltrative astrocytic tumours in a three-tiered system similar to that of the Ringertz[21], St Anne/Mayo[22] and the previously published WHO schemes[16]. The major difference between WHO and St Anne/Mayo classification lies in grade I[22]. The WHO classification assigns grade I to the more circumscribed pilocytic astrocytoma, whereas the St Anne/Mayo classification assigns grade I to an exceedingly rare diffuse astrocytoma without atypia[13]. Separation of grade II from grade III tumours following the WHO scheme, is supposed to be reliably achieved by determination of Mindbomb 1 (MIB-1) labelling indices rather than determination of mitosis[23, 24]. Similarly, some authors only accept the criterion of endothelial proliferation to assign grade IV, while WHO classification also accepts glomeruloid microvascular proliferation and any type of necrosis[13].

Astrocytic tumours	I	II	III	IV	ICD-O
Subependymal giant cell astrocytoma	•				9384/1
Pilocytic astrocytoma	•				9421/1[1]
Pilomyxoid astrocytoma		•			9425/3*
Diffuse astrocytoma		•			9400/3
Pleomorphic xanthroastrocytoma		•			9424/3
Anaplastic astrocytoma			•		9401/3
Glioblastoma				•	9440/3
Giant cell glioblastoma				•	9441/3
Gliosarcoma				•	9442/3
Oligodendroglial tumours					
Oligodendroglioma		•			9450/3
Anaplastic oligodendroglioma			•		9451/3
Oligoastrocytic tumours					
Oligoastrocytoma		•			9382/3
Anaplastic oligoastrocytoma			•		9382/3
Ependymal tumours					
Subependymoma	•				9383/1
Myxopapyllary ependymoma	•				9394/1
Ependymoma			•		9391/3
Anaplastic ependymoma			•		9392/3
Choroid plexus tumours					
Choroid plexus papilloma	•				9390/0
Atypical choroid plexus papilloma		•			9390/1*
Choroid plexus carcinoma			•		9390/3
Other neuroepithelial tumours					
Angiocentric glioma	•				9431/1*
Choroid glioma of the third ventricle		•			9444/1
Neuronal and mixed neuronal-glial tumours					
Gangliocytoma	•				9492/0
Ganglioglioma	•				9505/1
Anaplastic ganglioglioma			•		9505/3
Desmoplastic infantile astrocytoma/gangliglioma	•				9412/1
Dysembryoplastic neuroepithelial tumour	•				9413/0
Central neurocytoma		•			9506/1*
Extraventricular neurocytoma		•			9506/1*
Cerebellar liponeurocytoma		•			9509/1[1]
Paraganglioma of the spinal cord	•				8680/1
Papillary glioneuronal tumour		•			9509/1*
Rosette-forming glioneuronal tumour of the fourth ventricle		•			9509/1*
Pineal tumours					
Pineocytoma	•				9361/1
Pineal parenchymal tumour of intermediate differentiation	•	•			9362/3
Pineoblastoma			•		9362/3
Papillary tumour of the pineal region		•	•		9395/3*
Embryonal tumours					
Medulloblastoma				•	9470/3
CNS primitive neuroectodermal tumour (PNET)				•	9473/3
Atypical teratoid/rhabdoid tumour				•	9508/3
Tumours of the cranial and paraspinal nerves					
Schwanoma	•				9560/0
Neurofibroma	•				9540/0
Perineurinoma	•	•	•		9540/0
Malignant peripheral nerve sheath tumour (MPNST)		•	•	•	9540/3
Meningeal tumours					
Meningioma					9530/0
Atypical meningioma		•			9539/1
Anaplastic/malignant meningioma			•		9530/3
Haemangiopericytoma		•			9150/1
Anaplastic haemangiopericytoma			•		9150/3
Haemangioblastoma	•				9161/1
Tumours of the sellar region					
Chraniopharyngioma		•			9350/1
Granullar cell tumour of the neurohypophysis	•				9582/0
Pituicytoma	•				9432/1*
Spindle cell oncocytoma of the adenohypophysis	•				8291/0*

Table 1.1: *WHO grading of tumours of the Central Nervous System reprinted from* [19]. Summary of histological tumour types of the CNS and malignancy grade for each specimen (I–IV). The International Code of Diseases for Oncology (ICD-O) is also depicted.

[1] Morphology code of the International Classification of Diseases for Oncology (ICD-O) 614A and the Systematized Nomenclature of Medicine (http://snomed.org). Behaviour is coded /0 for benign tumours, /3 for malignant tumours and /1 for borderline or uncertain behaviour.

* Provisional codes proposed for the 4^{th} edition of ICD-O. While they are expected to be incorporated into the next ICD-O edition, they currently remain subject to change.

1.1.3.5 Survival of patients suffering from HBT

The WHO grade contributes to an overall estimation of prognosis combined with a set of clinical findings, such as age of patient, neurologic performance status and tumour location; radiological features such as contrast enhancement, extent of surgical resection; proliferation indices; and genetic alterations. Despite these variables, patients with WHO grade II tumours typically survive more than 5 years and those with grade III tumours survive 2-3 years[13]. The prognosis of patients with WHO grade IV tumours depends largely upon availability of effective treatment regimens. Most of glioblastoma patients succumb to the disease within a year, whereas for the other grade IV neoplasms the outlook may be considerably better.

1.2 HBT targets for genetic diagnosis implementation

Among HBTs, the most representative candidates for genetic diagnosis in the clinical practise appears to be oligodendroglial tumors, astrocytic tumours leading to glioblastoma, and meningiomas[14]. They have been extensively investigated in the past decade because they are the most frequent tumours. The huge amount of data obtained was compared with histopathological and clinical features, to produce models of multi-step carcinogenesis.

1.2.1 Glial tumours

1.2.1.1 Overview on gliomas

Human gliomas are the most frequent primary tumour of the CNS[12], but their incidence is low compared to other human cancers. However, an increase in glioma cases have been reported in the last years, in part due to the improvement of diagnostic techniques and the increase of life expectancy[25]. Human gliomas are classified according to their hypothesized line of differentiation. That is, whether they display features of astrocytic, oligodendroglial, or ependymal cells[13, 26].

The paradigm of glioma development was based on the progressively dedifferentiation of any of the formerly cited cellular type from its mature form until resembling their precursor cells[27] (see figure 1.3). Nonetheless, such a paradigm is under revision since new molecular events have been described and neural stem cells discovered. Moreover, histological stand-alone classification can not predict neither the clinical course of the pathology, nor the response to therapy of diagnosed gliomas[12, 27].

1.2.1.2 Malignancy grades

Gliomas affecting the cerebral hemispheres of adults are termed diffuse gliomas because of their propensity to infiltrate throughout the brain[26]. The diffuse gliomas are classified histologically as astrocytomas, oligodendrogliomas, or tumours with morphological features of both astrocytes and oligodendrocytes, termed oligoastrocytomas[26]. Astrocytic tumors are subsequently graded as pilocytic astrocytoma, grade I; astrocytoma grade II; anaplastic astrocytoma, grade III; and

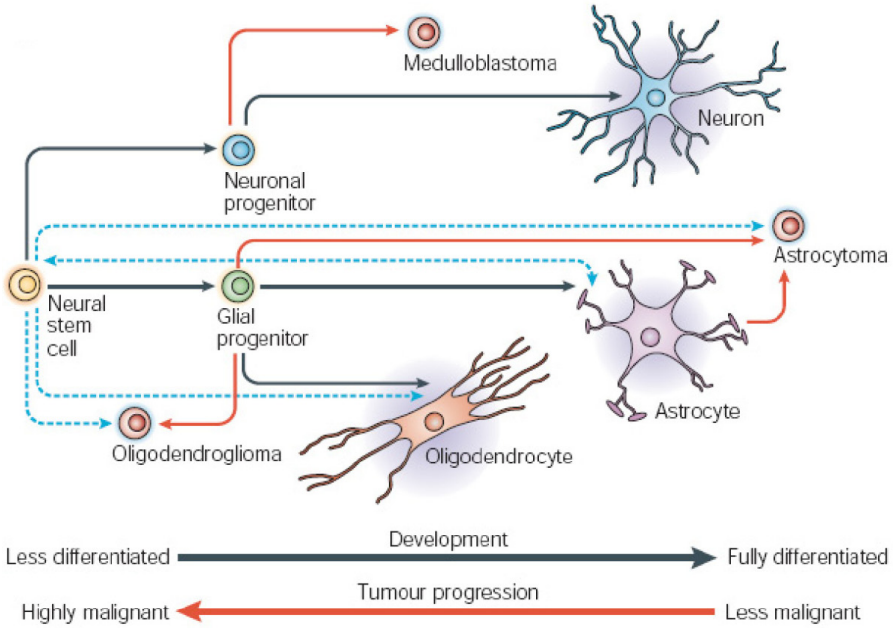


Figure 1.3: *Normal development of glial cells and pathological transformation.* The panoply of glial tumours is suposed to arise from a common neural stem cell progenitor that produces a glial progenitor to form the glial lineage, which can turn into malignant progression yielding gliomas. Black arrows indicate the hypothesized normal development and red arrows the supposed progression of CNS tumours. Dotted blue arrows emphasizes the role of neural stem cells in normal development and potentially in the formation of brain tumours. Figure extracted from reference [3].

glioblastoma (Gbm), grade IV. Oligodendrogliomas and oligoastrocytomas are subsequently graded as grade II and anaplastic, grade III.

Seventy percent of grade II gliomas transform into grade III and IV tumors within 5-10 years of diagnosis and then behave clinically like the higher-grade tumors[26]. This particular feature of gliomas implies that grade I and II tumours should be well characterized at both histological and molecular level to predict their malignant transformation when detected in the clinical practice.

Currently, there are two clinically-distinguishable types of grade IV glioma: the primary and the secondary Gbm. The primary appears to arise *de novo* by accumulation of genetic alterations of a progenitor glial cell with stem-cell properties[26, 27]. The secondary Gbm would correspond to a malignant transformation of a diffuse glioma within 5-10 years after diagnosis of the low grade tumour[26] (see figure 1.4).

Despite genetic alterations producing both Gbm types differ, molecular pathways triggered by these alterations seem to be the same[26]. Corroborating this similitude, both primary and secondary Gbm are clinically indistinguishable since the survival associated with their diagnoses is almost identical[26]. The fact that

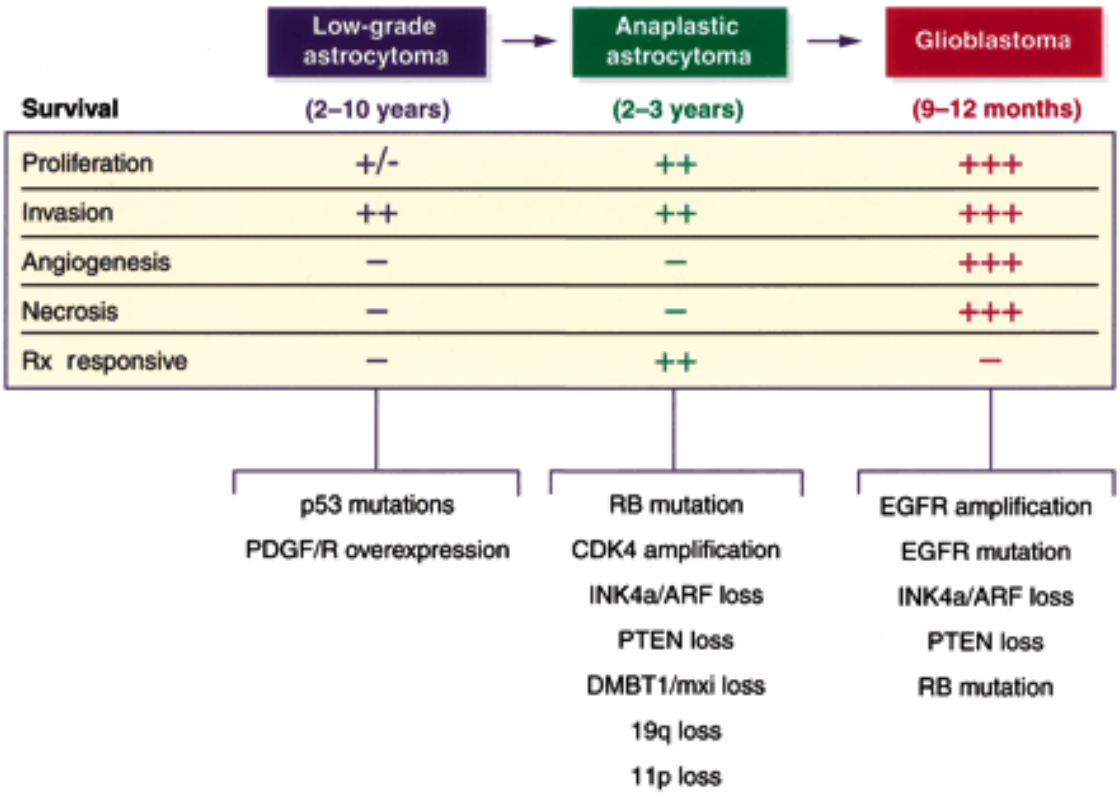


Figure 1.4: *Genetic and histological alterations from low to high grade gliomas.* The development of glial tumours from differentiated glial cells to high grades of malignancy is a sequential process. Genetic alterations (mutations, loss of heterozigosity, amplifications,...) are accumulated, distinct therapy responses are displayed and a decreased survival time is expected. Figure extracted from reference [26].

two Gbm types can share the same molecular pathways driving to malignant transformation may be explained as a result of the capacity of mature astrocytes to dedifferentiate into their stem-like progenitor cells, the radial glia[28, 29]. The scheme of genetic alterations leading to the transformation from low grade glioma to Gbm is depicted at figure 1.4.

1.2.1.3 Molecular genetics

Platelet-derived growth factor (PDGF) and epidermal growth factor (EGF) are elements thought to be important in malignant transformation of gliomas: PDGF in glioma development, EGF in neural stem cell proliferation and survival[26, 2]. The prominent overexpression of PDGF in low-grade gliomas and EGF receptor (EGFR) in Gbms suggests that these receptor tyrosine kinase (RTK) signalling pathways are critical targets in gliomagenesis[26].

Primary Gbm often affects the elderly population, it generally shows overexpression of EGFR and lack of mutation of tumour protein 53 (TP53), whilst the

secondary Gbm behaves inversely[26, 27, 30]. EGFR is the starting point of the EGFR/PTEN/Akt/mTOR pathway, which is a key signalling pathway in the development of primary glioblastoma[30]. EGFR becomes activated through the binding of growth factors (epidermal growth factor, transforming growth factor-α) to its extracellular domain, resulting in recruitment of PI3K to the cell membrane (see figure 1.5).

PI3K phosphorylates phosphatidynositol-4,5-bisphosphate to the respective 3,4,5-trisphosphate form (PIP3), which activates downstream effector molecules such as AKT (protein kinase B) and mTOR, the mammalian target of rapamycin. This results in cell proliferation and increased cell survival by blocking apoptosis. PTEN inhibits the PIP3 signal[31], thereby inhibiting cell proliferation (see figure 1.5).

In addition, the EGFR variant 3 (EGFRvIII) is exclusively expressed in Gbm and constitutively activated[30]. The EGFRvIII is the most abundant truncated form of EGFR, and lacks exons 2 and 7.

Concerning PDGF, during embryogenesis, neurons and astrocytes express PDGF [32], whereas glial progenitors and neurons express the PDGF-receptor (PDGFR) [33]. During the postnatal period, as glial progenitors differentiate into oligodendrocytes, PDGFR expression is down-regulated. As an example of the relevance of this receptor, PDGFR-α-expressing cells seems to be present in oligodendrogliomas[26], raising the possibility that these cells may be precursors of some oligodendrogliomas.

Furthermore, low-grade gliomas usually overexpress PDGF ligand and receptor, as well as harbor the TP53 mutation. This results in important genetic interactions between PDGF and p53[26]. The proliferative stimulus provided by PDGF signaling through the Ras/MAPK pathway is capable of promoting reentry into the cell cycle. This allows to speculate that simultaneous mutational inactivation of p53 and PDGF overexpression serves to promote the survival of aberrantly cycling premalignant cells[26].

Nevertheless, overexpression of this potent growth factor is associated with a very low proliferative rate in low-grade gliomas *in vivo*[26]. This contradictory finding may be explained by different reasons such as the presence of inhibitors operating at the level of the core cell-cycle machinery; an insufficient stimulation of PDGF to promote a robust cell-cycle entry in glioma cells; and involvement of several pathways (PI3K, JAK-STAT or PLC-γ) in addition to Ras[26].

Perhaps the most important role of PDGF in low-grade tumors may be to induce tumor cell migration through activation of PI3-K and PLC-γ. Both proteins have

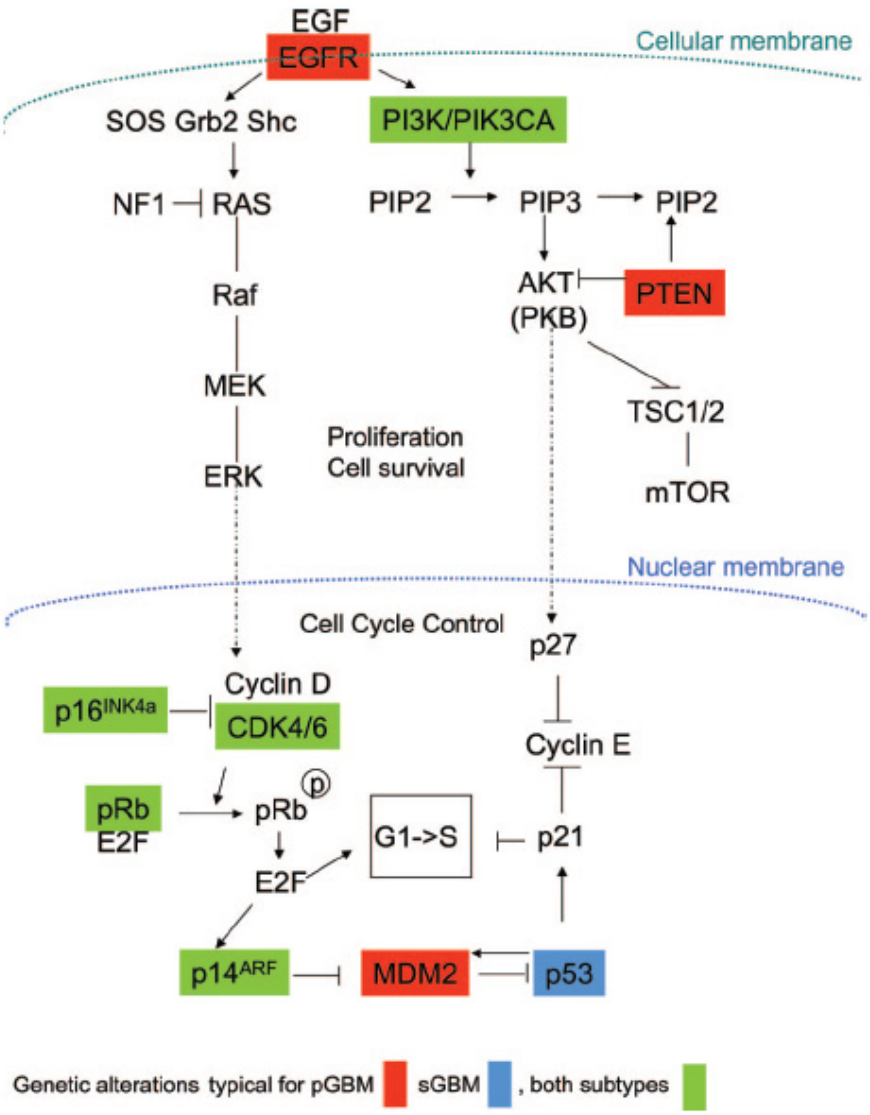


Figure 1.5: *Main signalling pathways involved in glioblastoma formation.* The EGFR signalling pathway starts at the plasmic membrane and transduces the signal up to the nucleus, where transcription factors are activated. The particular combination of genetic alterations at different levels of the signaling cascade characterizes each glioblastoma subtype. pGBM indicates primary glioblastoma and sGBM corresponds to secondary glioblastoma. Figure extracted from reference [30].

been implicated in migration and scattering[34], notably in oligodendrogliomas rather than astrocytomas[35], for which further investigation is necessary[26].

1.2.1.4 Treatment of gliomas

Surgical resection can be curative in some brain tumours, but not for malignant gliomas[36, 37]. Malignant gliomas usually infiltrate into the brain tissue surrounding the tumour mass. Thus, surgical removal appears not to be a curative

treament[37]. A combination of chemotherapy and radiotherapy is undergone prior and/or after tumour removal[37, 38]. The chemotherapeutic agent and the radiotherapy dose are selected depending on the tumour type.

Whilst temozoladamide is the most used drug to treat Gbm, grade III gliomas are more likely to be treated with carmustine, carbazine, lomustine and/or vincristine [38, 39, 40]. At any rate, delivery of drugs into the brain is difficult, since it produces systemic side effects[41]. Despite the continuous investigation about malignant glioma therapy, there was not an evident increase of patient survival in the past four decades[37]. Inhibitors targeting molecular key points of the malignant transformation process of gliomas, mainly angiogenesis and EGFR signalling pathway, have been tested in phase II of clinical trials without a noticeable amelioration[39].

1.2.2 Meningeal tumours

1.2.2.1 Overview on meningiomas

Meningiomas are the second HBT with the highest incidence after glial tumours, ranging between the 20 and 30% cases of intracranial tumours[42, 43]. The neoplasia in meningiomas arises from arachnoidal (meningothelial) cap cells, which are cells composing the meninges[42, 43, 44].

They remains in benign stages in approximately 90% of diagnosed cases, and predominantly appear in the elderly population and affect more females than males (2:1)[42, 43]. Unlike gliomas, meningiomas display well defined edges that permit complete resection, curative depending on location[44, 45].

1.2.2.2 Malignancy grades

Classification is most often based on WHO criteria, which establishes three grades of malignity: benign (grade I), atypical (grade II) and anaplastic (grade III)[13, 44]. The first editions of the WHO classification graded meningiomas based on qualitative criteria. In contrast, the more recent editions (2000 and 2007), incorporate quantitative parameters to estimate the tumour grade, which were proposed at the end of the past decade by investigators at the Mayo clinic[44].

1.2.2.3 Genetic alterations

Meningiomas have long been a subject of intense genetic and biological interest[46]. As a result, meningiomas are among the cytogenetically best characterised cancers [46]. The most widely accepted cytogenetic abnormality that meningiomas exhibit is the loss of heterozigosity (LOH) of the long arm of the chromosome 22[43, 44, 46].

Extensive molecular studies in meningiomas enabled the identification of the tumour suppressor gene *NF2*, responsible for neurofibromatosis 2 disease. *NF2* encodes a cytoeskeleton-associated protein, called schwannomin/merlin protein (also known as moesin-, ezrin-, radixinlike protein). The reduced expression of the *NF2* gene in meningiomas implies a decrease in cell adhesion and increased cell tumourogenesis[43]. Furthermore, mutations in *NF2* gene within grade I meningiomas can be detected from 25% to 80% of cases depending on the histological type[43]. As atypical and anaplastic meningiomas show a mutation rate in *NF2* similar to benign meningiomas, the *NF2* mutation appears to be an important event for tumour formation, rather than for tumour progression (see figure 1.6).

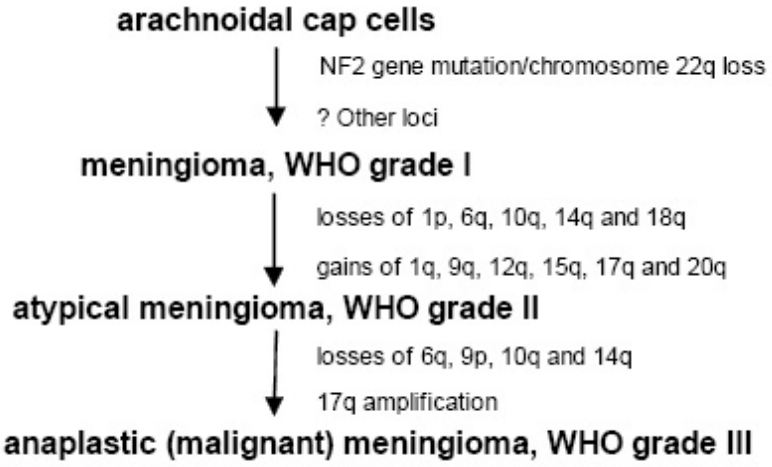


Figure 1.6: *Molecular and genetic alterations leading the meningioma progression.* Accumulation of sequential molecular alterations (mutations, losses and gains of heterozigosity) appear to be the key mechanism for meningioma formation and progression. Figure modified from reference [46].

On the contrary, in sporadic meningiomas, mutations of *NF2* or deletion of its locus, occurs notably less often than LOH of chromosome 22. This demonstrates lack of complete correlation of the two events and points to another altered locus with a role in meningioma biology[43]. Cytogenetic studies revealed that in proximity to the *NF2* locus, a series of genes presented abnormalities: Beta-adaptin-meningioma gene on chromosome 22 (*BAM22*), Acetylglucosaminyl transferase-like

protein (*LARGE*), Meningioma 1 gene (*MN1*) and Integrase interactor 1 (*INI1*).

Aside from the LOH of chromosome 22, alterations on other chromosomes have also been described in meningiomas. This results in an onset of changes, which all conjointly characterize formation and progression of meningioma (see figure 1.6). The chromosome 1 is the second chromosome accumulating more alterations in such tumour type[43, 46]. It appears to be involved, together with deletions of chromosomes 10 and 14, in malignant progression of meningioma[47, 48]. Frequency of deletions in chromosome 1 increase whith the malignancy grade, occurring in 13% to 26% of grade I, in 40% to 76% of grade II, and 70% to 100% of grade III meningiomas[46].

1.2.2.4 Treatment of meningiomas

The mainstay of treatment for meningiomas is complete resection, being curative in most cases, notably in low grades. Also, there is a dependency on both patient-related factors (age, performance status, medical comorbidities) and treatment-related factors (reasons for symptoms, patient respectability, and goals of surgery)[42, 49]. Radiation after the surgery was shown to improve progression-free survival of patients when partial removal of tumour mass is performed[42].

Nonetheless, the decision to undertake adjuvant radiotherapy should be weighed against the potential for symptomatic recurrence (considering the slow growth rate of most meningiomas) in the patient lifetime, versus potential side effects of radiation (for example, leukoencephalopathy and cognitive symptoms, necrosis, and focal neurological injury)[42, 49].

Chemotherapeutical agents are principally administrated to patiens suffering from meningioma recurrence, being the compounds used in clinical trials hydroxyurea, temozolomide, somastatin analogs and multidrug treatment, amongst others[42].

1.2.3 Mouse models to study HBT

1.2.3.1 Xenograft tumour models

Development of effective drugs to treat HBT requires a profound knowledge of the underlying molecular events driving to tumour formation, progression and recurrence[50]. A combination of cell cultures and mouse models are used to test the experimental hypothesis, prior a candidate drug can be subjected to clinical trials[50].

Xenograft tumour models consists in implantation or injection of primary tumour cells or cell lines into the animal model[51, 52]. It can be performed subcutaneously or orthotopically (into the native tumour site), to immunosuppressed, immunodeficient, or newborn immunonaive target mice[51, 52].

In vitro cell cultures together with *in vivo* xenograft brain tumour models provides a quick and efficient way of testing novel therapeutic agents and targets. The knowledge generated can be translated and tested in more sophisticated models such as genetically engineered mice (GEM). This particular mice model is expected to result in high quality clinical trials, which provide better treatment outcomes and reduced drug toxicities for patients[50].

1.2.3.2 Genetically engineered mice (GEM)

GEM are strains of mice forming spontaneous tumors due to mutations in the characteristic genes of malignancy[50]. They are an opportunity to discern the involvement of a certain gene, or a combination of genes, in the molecular and physiological events occurring in HBT tumourogenesis (angiogenesis, tumour-host interactions and metastasis to distant sites), because of predictability of the tumour-initiating lesion(s), immunocompetence, and tumour development at the appropriate site[53]. Furthermore, GEM enable to enlarge the dataset of those studies, in which there is a limited patient population[50].

1.2.3.3 Application of mice xenografts to model ischaemia in HBT

Schwaninger and collaborators generated a series of mice models to investigate the role of NF-κB signalling pathway in the onset of cerebral ischaemia[54]. Among other findings, they identified the pro-apoptotic cytokine TWEAK that binds Fn14, a member of the TNF receptor familly[55], which activates in turn the IKK complex and thus the translocation of NF-κB from the cytosol to the nucleus[56, 57].

Alternatively, Iadecola and collaborators characterised the function of inducible nitric oxid synthase (iNOS) producing cerebral ischaemia by oclusion of the middle cerebral artery in mice lacking expression of the iNOS gene[58]. Reduction of cerebral ischaemia injuries were demonstrated by a major recovery of motor functions of iNOS knockout mice compared to wild-type, although the glial response and the upturn of the cerebral blood flow were comparable in both animal conditions[58].

1.3 Microarray technology

1.3.1 An insight on microarray technology

1.3.1.1 Historical remarks

The begining of microarray experiments dates back to the second half of the 1980s when the first assays using fluorescent labelled antibodies were generated to detect protein levels in a "multi-analyte" approach[59, 60]. Translation of such initial approaches to nucleic acid hybridisation was first performed by Fodor and collaborators in 1993 by creating a biological chip with *in situ* synthesized probes[61]. Later, Brown and collaborators developed in 1995 a quantitative method to determine the levels of mRNA spotting cDNA probes onto glass-slides[62]. Nevertheless, the sequence of the human genome presented in 2001 by the Human Genome Project Consortium together with the Celera Genetics company, largely enhanced the use of the microarray technology broadly opening the "post-genome era"[63, 64].

1.3.1.2 Principle of the technology

The principle of gene-expression microrray or microchip technology is based on the binding capacity by sequence complementarity of single chains of nucleic acids molecules[65]. A microarray is a solid surface with immobilized gene-probes covering completely or partially a certain genome[65]. In a microarray experiment the messenger RNA (mRNA) is copied into complementary DNA (cDNA) or RNA (cRNA), and labelled with a fluorochrome or other stainning method depending on the technology used[65]. Labelled material is hybridized onto the gene-probes immobilyzed at the microarray and the signal quantified by a laser scanner[65] (see figure 1.7).

1.3.1.3 Impact of gene-expression microarrays in research

The development of high throughput technologies, such as mass spectometry, CGH arrays and protein arrays among others, conjointly with gene-expression microarrays have changed the paradigm of experimental molecular biology. Investigation about biological molecular mechanisms in a high throughput manner needs the collaboration of scientists covering a wide range of disciplines, which is the case of gene-expression microarrays[66].

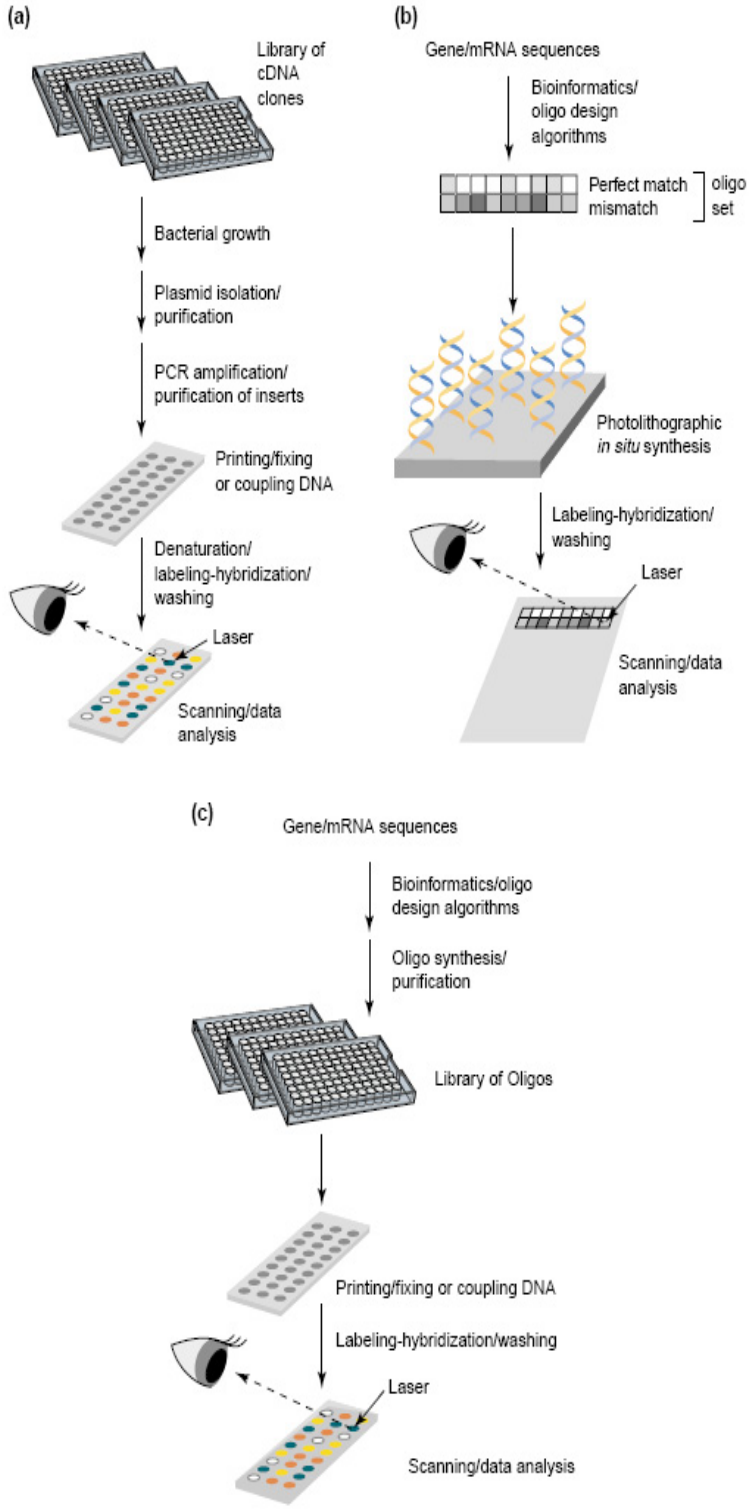


Figure 1.7: *Microarray protocols.* Microarray technology can be divided into two main groups: *in situ* synthesized and spotted-based microarrays. The former consists in growing up computationally designed sequences as can be seen in picture **(b)**. For spotted-based microarrays, the gene-probes are cDNA molecules from a library **(a)**, or synthesized oligonucleotides **(c)**, which are spotted onto the glass surface. Figure extracted from reference [65].

The ability of gene-expression microarrays to screen simultaneously the transcript level of the whole genome in a single experiment has attracted the interest of clinicians to search for gene signatures able to improve the understanding of molecular mechanisms of complex diseases, like neurological disorders and cancer[67]. Currently, several studies have revealed cancer subtypes[68], correlated gene signatures with patient survival[69] and allowed to determine the response of the organism to the treatment received[70].

1.3.1.4 Validation of gene-expression microarrays

Standardisation among microarray technologies remains a key point under discussion, noticeably when whishing to extrapollate its results for clinical application[71, 72]. Normally, experimental validation of microarray data is performed through real time-PCR (RT-PCR)[73]. However, compliance of clinical trials conditions requires of further validation. Otherwise, the results generated can not be accepted as a testable target with therapeutic purposes[74, 75, 76]. This point is more extensively considered in section 1.5.3.

1.3.1.5 Type of microarray experiments

There are two main possible types of microarray experiments:

1. Single-labelling.

   A single-labelling experiment consists in hybridizing onto a microarray cDNA or cRNA from a single condition labelled with a single dye (i.e. Cy3). From this experiment a single fluorescence intensity from each probe is obtained as gene-expression measurement.

2. Double-labelling.

   A double-labelling experiment is a competitive experiment between cDNA or cRNA from two different conditions, which are labelled with two different fluorochromes (i.e., Cy3 and Cy5). Labelled cDNA or cRNA solutions are mixed and hybridised onto a microarray.

   Double labelling experiments permit pairwise comparisons of each RNA sample (or condition) versus the rest of RNA samples (or conditions). Also, each RNA sample can be compared to a reference sample, which is RNA isolated from a sample considered a neutral condition with respect to the biological conditions evaluated.

See next sections for a more detailed explanation on the application of each approach.

1.3.2 Spotted-based microarrays

1.3.2.1 Introductory insights

As explained in the previous section, since the beginning of the microarray technonology two methods has prevailed in their manufacturing: *in situ* synthesized and spotted-based microarrays[62, 61]. The former was developed by the Affymetrix company (*www.affymetrix.com*); and its manufacturing method is explained at section 1.3.3. Spotted-based microarrays were proposed by Brown and collaborators at the Stanford University Medical Center with the aim of creating a technology to be self-manufactured by researchers that could be accessible worldwide and economically affordable[62].

1.3.2.2 Manufacturing

The manufacture of the spotted-based microarrays begins by generating a library of cDNA clones. After growing up the clones, DNA fragments are purified and amplified by PCR. Finally, they are robotically spotted (also called printed) onto nylon fiber or a modified-glass slide[65]. Alternatively, computer assisted oligonucleotide-design and synthesis of sequences is a possible method of microarray manufacture[65]. Printed sequences are electrostatically fixed onto the modified-glass slide, and crosslinked by heat or UV[65]. Covalent binding of the 5'-end sequences with the amine or other active groups on the modified-glass slides is feasible[65]. Such a procedure results in a microarray containing up to 50,000 features with a diameter ranging from 20 to 200 μm and spaced each other 50 μm, which enables printing of sequences that cover all human genes[65, 77]. Despite the first spotted microarrays were made in research laboratories, nowadays there is great availability of commercial products[65].

1.3.2.3 Sample labelling, hybridisation and image scanning

Labelling is initiated by retrotranscribing the mRNA into cDNA with an oligod(T)-primer, which enables incorporation of a dye into the growing sequence by using a labelled nucleotide. Usually nucleotides are labelled with a fluorochrome, but could be some other dye. This method is called direct or first strand labelling[78].

Conversely, rather than being labelled, the nucleotide can be chemically modified with an amino-allyl group, which in a further step binds the dye. This method is called indirect or second strand labelling[78].

At any rate, two different experiments can be performed when dealing with spotted-microarrays: single- or double-labelling[78, 79]. That is, labelling one RNA sample to hybridize onto one microarray, or separatelly labelling two different RNA samples with two different fluorochromes (usually Cy3 and Cy5), mix the solutions containing the labelled cDNA and do competitive hybridisation onto one microarray. Spotted-microarrays are typically performed with the double-labelling method[78].

Finally, after overnight hybridisation, the intensity signals are quantified by a laser scanner, and a specific software transforms into images the fluorescence intensity signals of fluorochromes such as cyanine 3 (Cy3) and cyanine 5 (Cy5)[77]. The gain of both the laser and the photomultiplier can be selected by the user. This allows a manual optimization of fluorescence signals within the detection range of the experiment. In contrast, optimization of signals is an automated process in Affymetrix microchips.

1.3.3 *In situ* synthesized-based microarrays

1.3.3.1 Introductory insights

Two main microarray types based on *in situ* synthesis are currently available: Affymetrix and Agilent technologies microarrays. In the case of the pioneers Affymetrix, back in 1991[77], the *in situ* synthesis was an adaptation of the photolithography production of computer chips to the gene expression studies, which resulted in the GeneChip name.

The second approach to *in situ* synthesize oligonucleotide was inspired in the ink-jet technology used on electronic printing devices that was adapted to gene expression studies by Rosetta Inpharmatics and licensed to Agilent technologies[77].

1.3.3.2 Manufacturing of Affymetrix microchips

Synthetic linkers with photolabile protecting groups are attached to a glass substrate, and a mask is used to direct light to predetermined areas on the substrate to remove the exposed groups. These de-protected groups are then available for reaction with bi-functional deoxynucleosides, resulting in chemical coupling. A new

mask is used to direct coupling at other sites, and the step is repeated until the desired sequence and length of oligonucleotide is synthesized[65] (see figure 1.8).

As a result, sequences of 25 bases in length are produced and grouped together on an area of 18 x 18 μm, which defines a probe cell. A probeset in the microchip is composed of 11-20 perfect match (PM) probe cells and 11-20 miss match (MM) counterparts. MM sequences are the same than their corresponding PM but with the central nucleotide changed with the purpose of detecting false positive hybridisations[80]. High-density oligonucleotide microchips can contain between 10^6 and 10^7 probe cells[65].

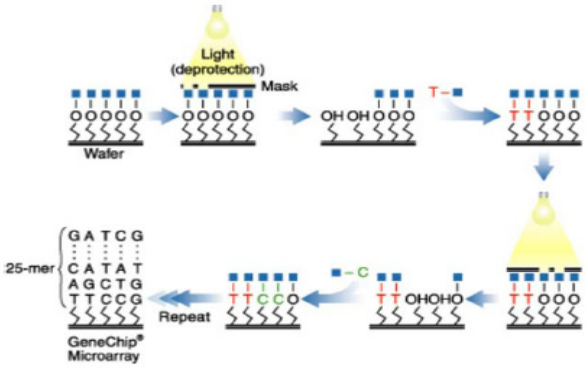


Figure 1.8: *Photolitography-based oligonucleotide microarrays from Affymetrix technologies.* The photolithographic process was inspired in the manufacturing of electronic microchips. The oligonucleotide sequence is synthesized by solid phase chemistry, and protected with lithographic masks. By applying light on the protected sequences, the mask is removed and the synthesis can continue. Repeated cycles are performed until obtaining the complete sequence. Figure extracted from the Affymetrix webpage (http://www.affymetrix.com/technology/manufacturing/index.affx).

1.3.3.3 Manufacturing of Agilent technologies microchips

In this method, modified ink-jet pumps, similar to those used in printers, are used to dispense 100-picoliter reagent droplets onto a hydrophobic surface containing chemically active hydroxyl groups. The droplets contain phosphoramidite DNA monomers that react and are covalently bounded. After washing and de-protection, the process is repeated until the desired oligonucleotide length is reached (see figure 1.9). The synthesis of oligonucleotides based on phosphoramidite chemistry is the standard technique since the 1980s, and it has other applications than microarrays.

The advantages of the in situ inkjet method are that no masks are required, synthesis is faster because each cycle attaches one base (four cycles per base are required with photolithography), and new arrays can be created by simply programming the computer with directions on how to synthesize the new set of oligonu-

cleotide sequences[65, 81, 82]. The disadvantage is the reduced number of elements, 176,000 in the latest chips, that can be synthesized compared to the Affymetrix technology, which can achieve 10^7 elements.

1.3.3.4 Sample labelling, hybridisation and image scanning

For Affymetrix experiments, labelling begins as in the case of spotted-based microarrays by generating a double-stranded cDNA, but carrying a transcriptional start site for the T7 RNA polymerase. cDNA molecules are transcribed *in vitro*, and biotin-labelled nucleotides are incorporated into the synthesized cRNA molecules. Each target sample is hybridized to a separate microarray and target binding is detected by staining with a fluorescent dye coupled to streptavidin[83, 84]. Signal intensities of probesets on different microchips are used to calculate relative mRNA abundance for each evaluated condition[65]. Affymetrix provides the one- and two- cycle amplification assays, which are selected depending on the amount of starting material available.

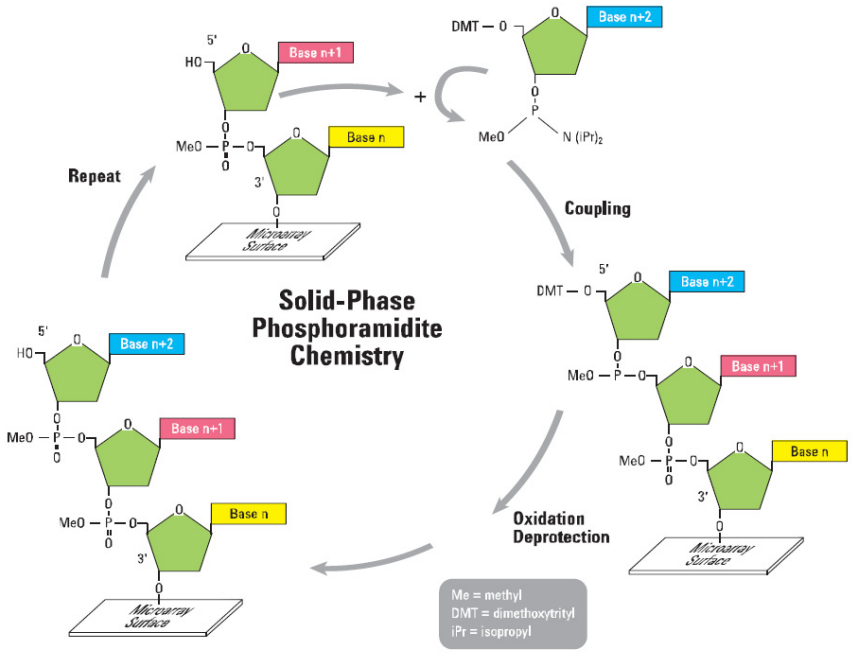


Figure 1.9: *Ink-jet-based oligonucleotide microarrays from Agilent technologies.* The ink-jet technology to manufacture microarrays is based in the same principles that governs the paper printing. In this case, repetitive cycles of synthesis based on solid-phase phosphoramidite chemistry are performed onto the glass surface, yielding the gene-probes composing the microarray. Figure extracted from reference[81].

Similarly, labelling for Agilent microchips consists in converting the mRNA into double strand cDNA using a oligod(T)-T7 primer, which serves as a promoter

for the amplification step using a T7 polymerase. cDNA is converted into cRNA incorporating cytosines labelled with either Cy3 or Cy5[85]. Unlike Affymetrix labelling, Agilent microchips allows both single- and double-labelling experiments. In both cases, an overnight incubation is required. However, intensity signals of fluorescence from Affymetrix microchips can only be quantified by the scanner provided by the company, while intensity signals from Agilent microchips can be quantified by most commercial scanners for microarrays.

1.3.3.5 Other *in situ* synthesis-based microchips

Aside from the Affymetrix and Agilent microchips technologies, there are three additional methods to produce microchips. First, GE Healthcare developed a microarray based in applying oligonucleotides of 30 bases in length to a three-dimensional polyacrylamide gel matrix by way of a non-contact, propietary piezoelectric dispensing method. Through covalent attachment, the oligonucleotides are immobilized to the active functional groups of the slide surface[86].

Also, an evolution of pholitographic masks used by Affymetrix for *in situ* oligonucleotides synthesis is the digital mirror device (DMD, or digital light processor, DLP), which synthesized oligonucleotides without requirement of a mask. Such a technology is commercialized by NimbleGen Systems, Febit and Xeotron companies [78].

1.4 Microarray data analysis

1.4.1 R language

1.4.1.1 Language definition

R is a programming language derived from the S language, which was designed in the 1980s and both R and S are widely used by the statistical community. Since 1998, the popularity of R was increased when its principal designer, John M. Chambers, was awarded with the ACM Software Systems Award for S[87].

R provides an integrated suite of software facilities for data manipulation, calculation and graphical display[88]. R must be understood as an "environment" affording a fully planned and coherent system, rather than an incremental accretion of very specific and inflexible tools, as is frequently the case with other data analysis software.

1.4.1.2 Advantages

In comparison with other statistical software packages such as SAS or SPSS, R has the advantage of performing the statistical analysis in steps, resulting in intermediate results stored in objects. R being an object-oriented programming language, gives minimal output and stores the results in a fit object for subsequent interrogation by further R functions. In contrast, SAS and SPSS will give copious outputs, which complicate their interrogation in further analysis[88].

1.4.1.3 Use of R

The most convenient way to use R is with a graphic workstation running on a windowing system[88]. R can run under UNIX, Windows and MacOS computers. In any of these operating systems, computation is performed through commandlines, which permits a larger interaction with the system as compared to other softwares, and gives the possibility of creating functions to systemize repetitive work tasks[87].

The R environment is supplied with a series of packages, which are a collection of functions to work on specific topics. The great flexibility of R has allowed the development of packages covering a wide range of scientific areas requiring statistical assessment. The aim underlying R is to generate a vehicle for newly

developing methods of interactive data analysis, which can be freely accessible worldwide.

R developers can contribute to add-on packages to the repository at the Comprehensive R Archive Network (CRAN) (http://cran.r-project.org/) webpage, where new packages are stored. In parallel, some projects have created repositories storing packages focusing on specific topics.

1.4.1.4 The Bioconductor repository

Bioconductor (http://www.bioconductor.org/) is an R project that provides a package repository for the analysis and comprehension of genomic data, which is continuously under development. There are two releases each year that corresponds to the released version of R. Packages are mainly developed to cover gene-expression microarray-data analysis.

Nonetheless, an increased number of packages have been uploaded to deal with other high throughput technologies such as SNP microarrays and mass spectometry. Regarding the topic concerning this thesis, Bioconductor contains a large number of packages allowing statistical analysis of microarrays that ranges from the processing of fluorescense intensity signals from scan images to the development of classifiers for different biological groups.

In addition, experimental data from different laboratories are available in Bioconductor, which usually exemplifies methods proposed in packages. Therefore, the available experimental data increases the facility of users to rapidly apply packages to their own data.

1.4.2 Data pre-processing

1.4.2.1 Spotted-based microarrays image processing

The processing of images generated using spotted-based microarrays can be divided in three steps: *adressing* or *gridding*, *segmentation* and *intensity extraction*[89]. There are several softwares enabling these steps, in a manual and/or automatic manner (e.g., Genepix, ImaGene, ScanAlyze and QuantArray)[89].

Gridding

This step consists in assigning the coordinates to each spot onto the image. For that, a grid composed of empty circles is fitted to the spots of the microarray

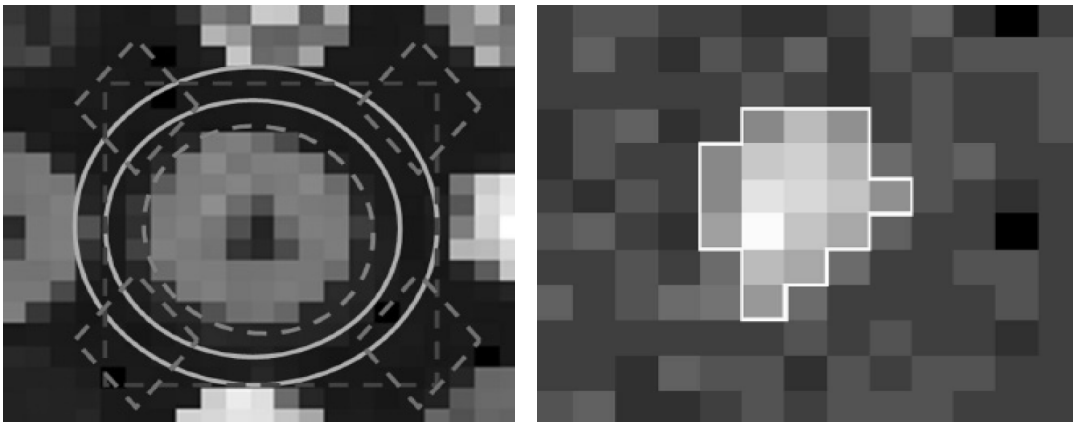


Figure 1.10: *Segmentation and Background estimation.* On the left, segmentation based on adaptative shape is shown by the region inside the white line. On the right, the region inside the circle delimited by the dashed line represents the foreground signal obtained by fix circle segmentation, and the other regions bounded by lines represent local background estimation by different methods. Images extracted from reference [89].

image.

Segmentation

Segmentation of a microarray image is the classification of pixels as a foreground or background signals, so that fluorescence intensities can be calculated as measures of transcript abundance[89]. The foreground signal pixels can be delimited by four principal methods: fixed circle, adaptative circle, adaptative shape and histogram *segmentation* (see figure 1.10).

The two first consist in delimiting each spot with a circle that can be fix for all spots or automatically adapted. Adaptative shape is an evolution of the adaptative circle *segmentation*, since the edges of the mask are adapted to the spot shape, thus eliminating unspecific signal that could arise from using a circle mask for a non-circular spot.

The histogram *segmentation* method is based on using for all spots a *circle mask* with the size of the spot largest on the microarray. A histogram of pixel intensities is computed for each spot with the aim of arbitrary selecting a low intensity rank to assign the background signal and a high intensity rank corresponding to the foreground signal[89].

Background estimation

On the other hand, delimitation of an area to estimate background intensity signal is slightly more sophisticated. A general procedure consists in considering as background the signal that can be detected on the area surrounding or nearby to a determined spot (local background), with the aim of removing it from the foreground signal in either natural or logarithmic scale. Less frequently, the background is not estimated and uncorrected foreground signals are used[89]. Nevertheless, these assumptions are not always satisfactory[89]. There are three different methods of background estimation that can be considered the most widely accepted:

- local background: the background area can be delimited depending on the software by two concentric circles around the spot, the outer part of a square centered at the spot center or four rombs as depicted at figure 1.10.

- morphological opening: a non-linear filter is obtained by computing a form of local minimum filter (an erosion) followed by a form of local maximum filter (a dilation), producing a background image from the raw image.

- constant background: it assumes that a better measure of background signal is to use the average intensity signals of negative controls spotted onto the microarray, for example specific gene sequences of plants onto a human microarray.

Additionally, Edwards and collaborators developed a background correction method for single-labelling microarray experiments sensitive to the difference between foreground and background quantified signals, thus, avoiding negative signals that can generate problems in data normalization[90]. That is, when foreground is larger than background signal, the difference is computed, otherwise the intensity signal is computed by interpolation of a smooth monotonic function that is linear to background intensities in logarithmic scale.

Intensity extraction

Finally, *intensity extraction* or *quantification* of both the foreground and background signal enables the estimation of the gene expression measure for each spot. The intensity of each spot is determined as the sum or mean (both are possible) of pixel intensities contained within the area delimited during the *segmentation*

step. When working with double-labelling microarray experiments, calculations of expression measures are based in ratios, which can be computed as the quotient between the sum or mean of pixels from each channel. Alternatively, the ratio of medians is not associated with any biological meaning but can be seen as a robust variant of ratios of means[89].

Quality of hybridisation

The quality of cDNA microarrays or batch reproducibility can be assessed by comparison of the intensity signal of all microarrays hybridised for a certain study.

High signal-to-noise ratios, foreground to background signal, is a measure of a correct hybridisation[91]. However, artefactually increased signals can arise from cross-hybridisation or the unspecific binding of fluorochromes to cDNA probes[92].

Finally, *spiking external controls* at different stages of a microarray experiments allows monitoring of possible pitfalls during the whole process[91]. The basis of such an approach consists in spotting a series of dilutions of a gene not expressed in the studied biological condition, for instance the large subunit of *RuBisCO* onto a human microarray. The cDNA sequence of the spiked control is included in the labelling step, which serves to monitor any downstream alteration, and specifically enables detection of cross-hybridisations[93].

1.4.2.2 Affymetrix GeneChip image processing

Gridding

The focus of the data pre-processing will be on the Affymetrix GeneChips because they are the *in situ*-synthesized microarrays used in this thesis. Affymetrix provides the GeneChip Operating System (GCOS) software, formerly named Microarray Suite (MAS), which controls the cRNA labelling, hybridisation and scanning processes. The raw image obtained by the scanner is also transformed by GCOS to a *.dat* file by means of an automatic process that consists of placing a grid on the image. Gridding is guided by the signal of the hybridisation control $\beta 2$ sequences on the corners of the microchip[94, 95].

Intensity extraction

The grid is divided into squares corresponding to each probe cell of the microchip (see figure 1.11). The pixels placed at the perimeter of a probe cell are discarded, since the optimal hybridisation occurs in its central zone. Furthermore,

misalignment of a probe cell carrying signal from other probe cells is more likely to occur on the edges[95]. As a result, the 75^{th} percentile of total pixels is reported as the estimate intensity of each probe cell.

Each probe cell has thousands of a certain 25-mer oligonucleotide sequence, which is called perfect match (PM). Each PM has an associated miss match (MM) probel cell, which has the same sequence than the PM, but with the central nucleotide changed. A probe set is composed of 11-20 probe pairs PM-MM. Quantification of all probe pair PM-MM from a probeset results in the generation of the *.cel* file[95]. Such a computation is performed by the GCOS Affymetrix software through a two-step process: the detection and the signal algorithms[94].

Detection algorithm

The detection algorithm consists of assigning the vote of present, absent or marginal call for each probe pair based on the computation of the discriminant score (R). Evaluation of the statistical significance of the R score is performed by comparison with an arbitrary parameter *Tau*, which by default is set to 0.015 and must be a small positive value. Considering the MM a measure of unspecific hybridisations, R accounts for the relevance of PM signal in each probe pair:

$$R = \frac{(PM - MM)}{(PM + MM)}$$

By comparing the R values from a probeset with the *Tau* parameter defined value by the user, a *p-value* is computed using a One-Side Wilcoxon's Signed Rank test. The value of the *Tau* parameter is adjusted depending on wanting to increase or decrease specificity and/or sensitivity(see figure 1.11 and 1.12). The higher the Tau, the higher the specificity but the lower the sensitivity.

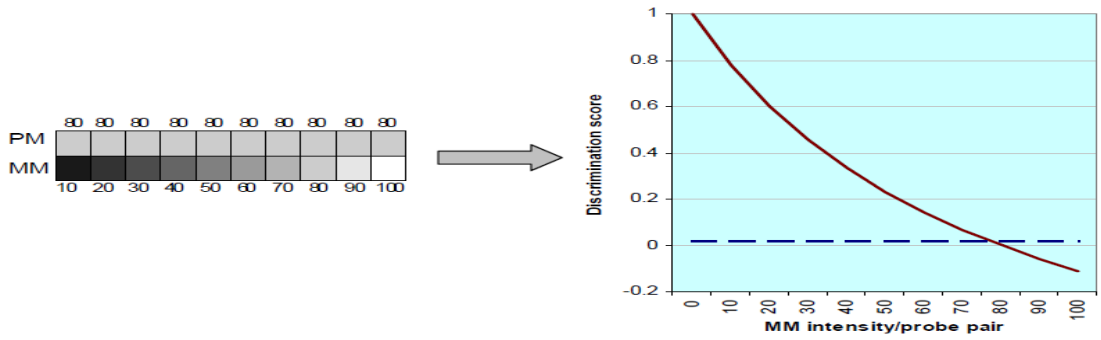


Figure 1.11: *Intensity extraction in Affymetrix microchips.* At the right side, a hypothetical probeset composed of 10 probe pairs PM-MM. Each probe pair is composed of one probe cell PM and MM. The curve represented at the right side is derived from the intensities obtained from each probe pair. Image extracted from reference [94].

Additionally, two more arbitrary parameters, $\alpha 1$ and $\alpha 2$, are selected by GCOS among the computed *p values* to define the marginal vote or range. Probe pairs that are above or below the marginal vote for a determined probe set will be considered present or absent calls, respectively (see figure 1.12).

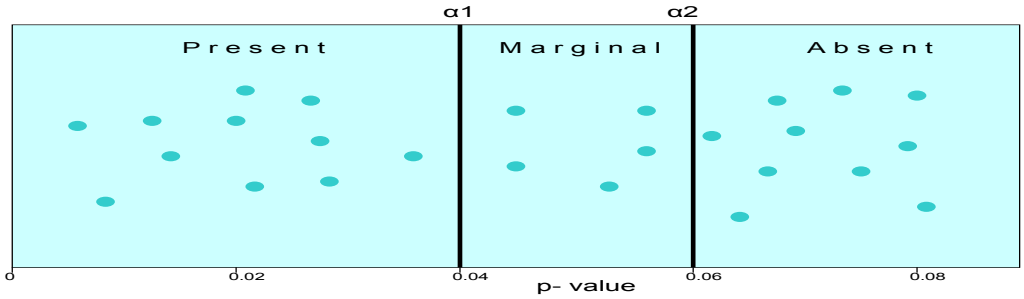


Figure 1.12: *Graphical representation of the detection algorithm.* Diagram of *p-values* computed from the *Tau* parameter defined by the user. The additional $\alpha 1$ and $\alpha 2$ paramers serves to set the marginal region. Image extracted from reference [94].

Signal algorithm

On the other hand, the signal algorithm operates in a similar way to the detection algorithm. However, the purpose in this case is to provide an estimation of the intensity signal for each probe set that represents the relative level of expression of a transcript[94]. Signal is calculated using the One-Step Tukey's Biweight Estimate which yields a robust weighted mean that is relatively insensitive to outliers, even when extreme. The signal for a probe set is computed depending on the level of PM and MM intensities of each probe pair following three rules:

1. **Rule 1:** If the MM value is less than the PM value, then the MM value is considered informative and the intensity value is used directly as an estimate of background signal.

2. **Rule 2:** If the MM probe cells are generally informative across the probe set, with the exception of only a few MMs, an adjusted MM value is used for uninformative MMs based on the biweight mean of the PM and MM ratio.

3. **Rule 3:** If the MM probe cells are generally uninformative, the uninformative MMs are replaced with a value that is slightly smaller than the PM. These probe sets are generally called Absent by the Detection algorithm.

Quality of hybridisation

When dealing with multiple chips during the *.dat* file generation a target value (TGT) must be fixed by the user, by default 500 a.u., in order for the intensities

of the microchips to be comparable[94]. As a result, a .cel file is generated for each microchip, which is scaled to the intensity set with the TGT value.

Moreover, there are a series of parameters and internals controls within the microchip to ensure the quality of the experiment and the comparability of all microchips of a certain study. Such an information is stored at the *.chp* file and it is summarized at the *.rpt* file. The most important parameters or steps are:

- **Visual inspection** of the image derived from the *.dat* file must be performed to detect possible artifacts and to verify that the $\beta 2$ control sequences display an intensity signal close to saturation.

- The **average background** and the **noise** value provide respectively a measure of the unsignificative signal and the electrical noise of the scanner. The average background values must range between 20 and 100 a.u. and the noise must not vary significantly between experiments, since it is an inherent parameter of the scanner.

- The number of probe sets called **"Present"** relative to the total number of probe sets on the array should be similar between experiments and never too low, which may indicate poor sample quality.

- The **scale factor (SF)** accounts for the variability among a set of experiments. The SF is the coefficient to be applied to the trimmed mean signal (a method to compute the average signal) of an experiment to fit the TGT, set by the user, for all the considered experiments of a study[96].

- The **normalisation factor (NF)** performs similarly than the SF. In this case, the NF is a coefficient used to normalize the trimmed mean of a single experiment to an arbitrary normalization value[96]. Large discrepancies among scaling/normalization factors (e.g., three-fold or greater) may indicate significant assay variability or sample degradation leading to noisier data[94].

- The **poly-A RNA controls** can be used to monitor the entire process of labelling and hybridisation. *Lys*, *phe*, *ther* and *dap* genes from *B. subtilis* were modified to include a poly-A tail, which serves as starting transcription-site for T3 RNA polymerase. Labelled poly-A controls are included into the hybridisation cocktail and must be called "Present" with increasing signal values from *Lys* to *dap* genes.

- The **hybridisation controls** *BioB*, *BioC*, *BioD* and *Cre* genes are provided by the manufacturer and spiked in the hybridisation cocktail. They must display increasing signal values from *BioB* to *Cre*, since increased concentrations are furnished with the kit.

- The **internal controls** of *GAPDH* and *β-actin* measure the efficiency of labelling by computation of the 3'-end to 5'-end ratio. They indicate degraded RNA or inefficient transcription of double-strain cDNA or biotinylated cRNA, when the ratio is higher than 3 for the one-cycle amplification protocol.

1.4.2.3 Alternative methods for image processing for Affymetrix GeneChips

Once the *.cel* file is generated and the set of experiments accomplishes the quality controls explained above, alternative background correction methods have been proposed to optimize the detection of the foreground signal. The R language-based software offers a collection of methods to correct background within the *affy* and *affyPLM* packages. They allow to apply the same methods used by the Affymetrix software, which is named MAS5 within the mentioned packages. Additionally, alternative correction methods, such as the robust microarray analysis (RMA) and gcRMA, were developped by scientists making them available worldwide through the R repository Bioconductor.

While MAS5 uses both the PM and MM to estimate the individual signal of probe sets as explained above, RMA and gcRMA neglects the MM signals. The expression values are computed by inferring a linear model that accounts for the binding affinity of the sequence synthesized in each probe cell and the combination of probe cells signals calculated using a median polish[97, 98]. Likewise, gcRMA also considers the guanine and cytosine content of the sequence synthesized in each probe cell, in an attempt to correct the signal provided by unspecific hybridisations[99].

1.4.3 Data normalisation

1.4.3.1 Scope of normalisation

The purpose of data normalisation consists in correcting intensity bias within and between microarray experiments[90, 100]. Such a bias may arise from the intrinsic variability of the microarray technology, RNA isolation, labelling and hybridisation,

rather than from biological differences, which are of interest to preserve. The general hypothesis for microarray data normalisation is that the level of expression of measured genes in a certain experiment does not vary neither within nor between experiments[101].

1.4.3.2 Global or scale normalisation

Each experiment can be scaled to an arbitrary intensity value and consequently all experiments should be comparable with each other. Typically, this strategy is named *globlal median normalisation* and used to normalise single-labelling microarray experiments, both spotted- and *in situ* synthesized-based microarrays. To note, Affymetrix, through the GCOS software, adjusts intensity of microchip experiments to an arbitrary scale factor (TGT) as explained in section 1.4.2.2. However, such an approach can not correct the bias that usually appears at low intensity signal range, so more sophisticated normalization methods have been developed to account for this[102, 103].

1.4.3.3 Local weighted and scatterplot smoothing (lowess) normalisation

Originally designed to normalise double-labelling cDNA microarray experiments, this normalisation process can be visualized through an MA plot (see figure 1.13). Considering G as the intensity signals from the sample labelled with Cy3, normally depicted green, and R as those from the labelled with Cy5, usually depicted red, the MA plot displays along the y *axis* the

$$M = \log R - \log G$$

, and the

$$A = \frac{(\log R + \log G)}{2}$$

along the x *axis*[100].

A linear regression is locally fitted within selected intensity signal ranges as to cover the whole intensity range of the experiment. For each range, the intensity signal of I is the result of a linear function applied to G:

$$I = g(G) + \varepsilon_i$$

where g is the local regression function and ε_i is a random error value[104]. Since

the lowess method enables to select discrete ranges of fluorescence signals, print-tip lowess normalization can be performed over groups of spots printed with the same tip. Thus, spots with a great probability of showing a similar variation due to technological effects can be normalised separately[100].

An extension of this method to single-labelling cDNA microarrays was proposed by Edwards and collaborators, who replaced the second labelled sample of a double-labelling experiment by an average chip computed from all the microarrays considered in their experiments[90]. Another lowess method to deal with single-labelling experiments is the cyclic lowess[105, 106]. This is an iterative method fitting the lowess regression in pairs of experiments as to perform all pairwise comparisons.

1.4.3.4 Quantile normalisation

Similarly to lowess, quantile normalisation corrects intensity bias by means of a linear model. In this case, intensity signals of the considered microarrays are increasingly or decreasingly ordered. Consequently, a matrix composed of as many columns as considered microarrays is obtained. Rows are ranked by value of intensity signal, thus it may not correspond to the same gene or feature in each column. For each row, the average is computed across all columns and the value of the whole row is replaced by such an average. Finally, undoing the order applied to each column the normalized matrix is obtained[105, 106].

1.4.3.5 Non-linear normalisation

The idea underlying non-linear methods is approximately the same than the linear methods, but differing in their computation. To scale the intensity signal of the microarrays, a baseline is set and rather than linear regression or scaling, non-linear methods as smoothing splines[108] or a piecewise running median line[109], are applied.

1.4.4 Feature selection

1.4.4.1 Introductory remarks

Transcriptomic analysis using microarray experiments has the advantage of generating a huge amount of gene-expression data expected to improve accuracy of tumour classification. Nevertheless, proper gene or feature selection among the thousands of genes available in a microarray experiment must be performed. The

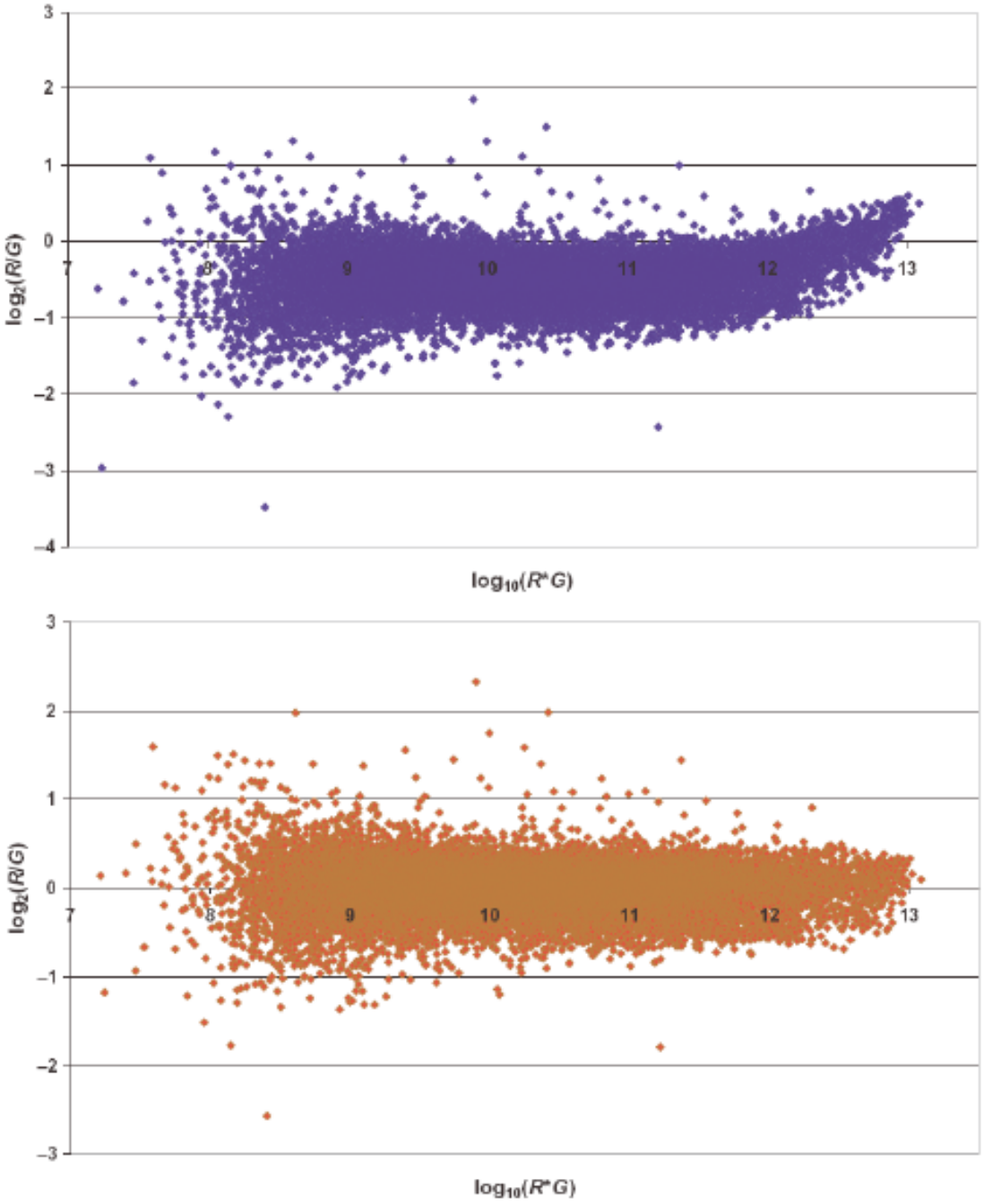


Figure 1.13: *Display of lowess normalisation.* The MA plot enables visualization of the normalisation produced by the lowess method. The upper plot represents non-normalised data, whereas the lower one depicts data after lowess normalisation. R indicates the intensities values obtained from the sample labelled with Cy5, whereas G those values obtained from the sample labelled with Cy3. At both plots, the x axis harbors the sum of logarithmic intensities, or $log_{10}R * G$, for each gene and labelled sample (A values). In contrast, along the y axis, the difference of logarithmic intensities, or $log_2 R/G$, is plotted (M values). Images extracted from reference [107].

major drawback of the feature selection relies on detecting those genes that are differentially expressed between the classes compared.

1.4.4.2 Fold-change ratio

The fold-change ratio between the averaged expression of genes between two classes can be used to select features. Usually, a cutoff value is set and genes with a fold-change higher than 2 or lower than 0.5, respectively, are selected for further analysis. Feature selection by means of fold-change must be complemented with the assessment of statistical significance[107]. That is, computation of a statistic with an associated probability accounting for the significance of the gene-expression between two classes.

1.4.4.3 Statistical significance analysis

Common statistical tests assume normal distribution and equal variance-covariance between the two groups, as in the case of the t-test. However, such conditions are rarely accomplished when dealing with microarray experiments[110].

Microarray experiments generates data with high dimensionality. The number of genes in each experiment is much higher than the cases analysed. This produces that the above mentioned conditions to perform common significance tests can not be accomplished[111]. Furthermore, this may produce false detection of differentially expressed genes, since a high statistical significance can result from random effects when simultaneously performing multiple tests[111]. As a consequence, several statistical tests addressing the particularities of high throughput experiments have been developed.

Non-parametric tests

Non-parametric statistical tests, such as the Wilcoxon/Mann-Whitney, allow to compute a statistic irrespective of the distribution of data. To evaluate the difference between two groups, this test ranks the values of each group[112]. If all the ranked values of one group are smaller than the other, this indicates that the groups analysed are different and a p-value is calculated based on these grounds[112].

1.4.4.4 Multiple-test adjustment

P-values must be adjusted to reduce the error derived from a multiple-test analysis. For this purpose, developed methods can be divided in three groups: those controlling the family-wise-error-rate (FWER), the false discovery rate (FDR), and the *positive* false discovery rate (pFDR)[113, 114, 115].

Family-wise error rate (FWER)

Similarly to uncorrected p-values, FWER accounts for the false positive rate (FPR). That is, the rate for truly null genes (truly non-differentially expressed genes) to be called significant[114, 115], but considering the multiple test conditions to compute the statistic[113]. The most popular methods for this are the Bonferroni correction and resampling-based tests[113]. The former consists in dividing the cutoff p-value to call a gene differentially expressed, usually $p<0.05$, by the number of genes being analysed[113].

Resampling tests are based on permuting the columns of the gene-expression matrix, where columns are cases and rows are genes, without regard to the their class[114]. Such a procedure results in a determination of the p-values irrespective with the distribution of data and minimizing the FPR.

False discovery rate (FDR) methods

The FDR method is thought to estimate the expected proportion of false positive genes among the significantly expressed genes[114]. This method proposed by Benjamini and Hochberg[116] coincides with the FWER when the amount of significantly expressed genes is equal to the amount of non-significantly expressed genes[114].

By contrast, the pFDR method proposed by Storey[115] is defined as the conditional expectation of finding false positive genes among the genes called significant through a Bayesian approach[114]. It undertakes the modelling of an *a priori* probability for a gene to be called significantly expressed[114]. The derived adjusted p-values are called q-values when using FDR and pFDR.

1.4.4.5 Principal component analysis (PCA)

Finally, aside from the feature selection by means of siginificance tests, principal component analysis (PCA) provides a robust manner to determine those features accounting for the greatest variability. PCA opperates reducing the high dimensionality of microarray data, thus simplifying the management of the analysis, through computation of *eigenvectors*, or *eigengenes* in the biological context[117]. Such a methodology has been reported to yield good classification results dealing with microarray experiments[118, 119].

1.4.5 Classification methods

1.4.5.1 Introductory remarks

In the oncological framework, researchers are motivated to employ genetic profiles derived from microarray experiments with the aim of improving classification of cancer types. As previously mentioned, a proper identification of candidate genes to produce adequate classification results must be performed. In addition, development of classifiers with enough confidence to reach a high clinical relevance is mandatory[120]. In this sense, a strict procedure must be followed not to overestimate the ability of the classifier generated[120].

1.4.5.2 Unsupervised classification

Cancer classification using microarrays is often addressed in a first step by performing unsupervised classification, for which the most popular method is hierarchical clustering[76, 120]. Such an unsupervised approach is based on computation of differences between genes and cases through selected metrics.

The computation method for this will vary with the algorithm used but, broadly speaking, clusters of genes that have a similar level of expression are achieved by means of an iterative process. Such a procedure can be simultaneously performed for several cases. The result is a graphical display optimally expected to allow visualization of cases clustered into the correct group, if labels are known, with a series of genes clusters defining each group of cases[121].

Gene signatures derived from a hierarchical clustering can lead to false conclusions when the signatures are directly correlated with the outcome[120]. Although methods have been proposed to calculate the accuracy of the clustering[122], the most appropriate method to determine genes related with the classification outcome is through supervised methods[120].

1.4.5.3 Supervised classification

In the context of supervised classification, a dataset composed of samples belonging to two classes, or more, is split into a training and a test set. The former is used to generate a mathematical function that serves to predict the class of the test samples left apart. The *classification accuracy* can be defined as the percentage of correctly assigned cases that are predicted by a mathematical algorithm.

Several mathematical algorithms for carrying out supervised classification of

cancer classes have been reported[120, 123, 124]. Briefly, all algorithms search for a function that enables separation of cases into their corresponding class group (see figure 1.14). Three representative strategies are described as follows:

- **Linear discriminant analysis (LDA)** builds a linear function minimizing the distances within group and maximizing the distance between groups[125].
- **Support vector machines (SVM)** seeks for a hyperplane than can separate the groups, assuming that the data is linearly separable[126]. Otherwise, transformation of data to a higher dimensional space is performed, which results in a non-linear separation when transforming data to the input space.
- **Random Forest (randF)** is a method that classifies test samples by construction of classification trees using the input variables (genes in our case) from the training set. Each classification tree provides a vote for the sample to be predicted. The algorithm chooses for prediction the class having more votes[127].

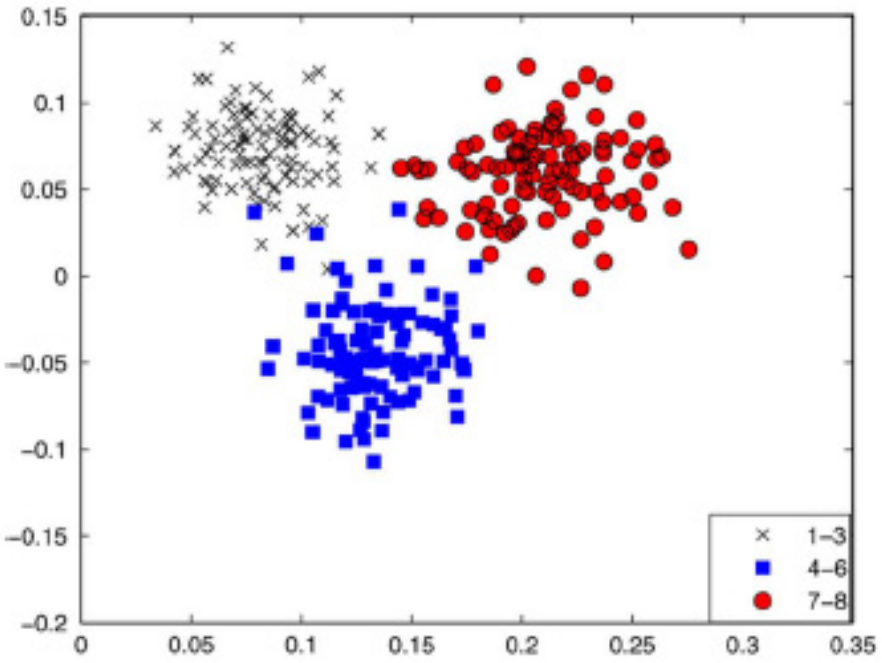


Figure 1.14: *Discriminant plot for a three-classes LDA classifier.* The display of LDA values for each sample is a means to visualize the "physical" separation between classes considered. Along x axis, the values of LDA1 component are displayed, whereas those values for LDA2 component are displayed along the y axis. Figure extracted from reference [128].

1.4.5.4 Procedure to estimate the classification accuracy

When wanting to define a set of candidate genes to distinguish among cancer classes, there are a series of steps that can not be neglected. Otherwise, an overestimation of the outcome may ensue. First, the indications explained in section 1.4.4, regarding gene selection must be implemented through a resampling classification approach[120]. Namely, considering a classification problem, it is mandatory to divide the available dataset in one part for training and another one for test purposes.

Such a random partition must be repeated and the feature selection performed on the training set at each iteration(see figure 1.15). To remark, the feature selection must exclusively be performed on the training set. That is, samples included for training purposes can not be used as a test set to calculate a correct classification accuracy. In an iterative process, the classification accuracy of test sample(s) is computed at each iteration, which allows computation of an averaged accuracy over all iterations and a confidence interval[120].

Partition methods

There is no unique rule for the partition of the data set into training and test set[123]. Some representative strategies can be:

- *Split sample (SS)* usually selects 2/3 of the dataset for training and 1/3 for test purposes.
- *v-fold cross validation (VFCV)* leaves apart from the training set a proportion of samples defined by $1/v$, where v is positive integer number. In the case of *leave-one-out cross validation (LOO)*, only one sample is not used for training.
- *Monte Carlo cross validation (MCCV)* and *.632+ Bootstrap*, works similarly than the previous methods, but requiring a higher computational capacity. As the case of *SS* and *VFCV*, several training sizes can be evaluated.

Sample size estimation

At this stage, it is also relevant to consider how to determine the correct sample size. Certainly, a homogeneous size of class groups must be considered[120] and the proportion of samples reserved for the training set must be selected through

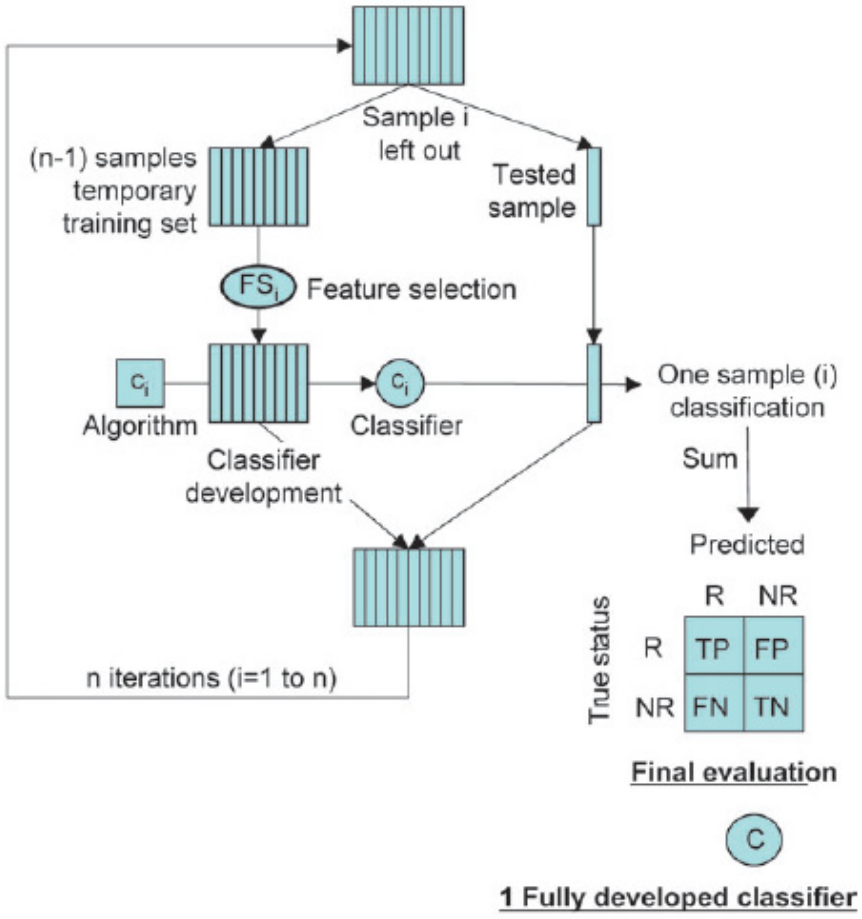


Figure 1.15: *Classification procedure.* This scheme is intended to visualize the required procedure to correctly estimate the classification accuracy and related parameters (sensitivity, specificity,...). The data set must be split into training and test set. Such a partition can be performed in different ways (see section 1.4.5.4). In this figure, only leave-one-out (LOO) is represented. Feature selection is only performed on the training set, and the classification function extracted is used to predict the class of the sample left apart. This procedure is repeated for all cases, and the accuracy mean of classification is computed. Figure extracted from reference [120]. R is a hypothetical condition and NR the non-R condition. TP is true positive, TN indicates true negative, FP denotes false positive and FN is false negative.

objective criteria when using *VFCV*, *MCCV* or *.632+ Bootstrap*. For this aim, a series of classifications using several training set sizes can be tested and the accuracy calculated for each one. The training set size yielding the greatest accuracy must be selected to optimize the best classifier[124].

Statistical significance of models

At any rate, the statistical significance of the classification accuracy must be assessed[120, 124]. To do this, an iterative procedure is performed over the dataset as described above, but cases of the training set for each class are randomly as-

signed. Therefore, such a procedure permits computation of a random accuracy that is contrasted with the true accuracy to generate a *p-value* accounting for its statistical significance[124]. The final validation of a classifier must be performed testing an independent test set when enough samples are available[120].

Classification parameters of clinical interest

A rigorous classification scheme with clinical interest must contemplate the computation of sensitivity and specificity, since the prediction accuracy alone has no clinical use[120]. Let consider a hypothetical classification example for two classes, healthy (n=15) and disease (n=15). If we have a "gold-standard" to refer the classification provided by our model, we could build a table to report the cases correctly or uncorrectly classified:

	Healthy **"gold-standard"**	**Disease** **"gold-standard"**
Healthy **predicted**	10 TP	5 FN
Disease **predicted**	7 FP	8 TN

where TP is the amount of true positives (correctly assigned healthy cases) and TN is the amount of true negatives (correctly assigned disease cases). If we consider, the amount of false negatives (FN) (uncorrectly assigned healthy cases), and the amount of false positives (FN) (uncorrectly assigned disease cases), we can derive the parameters of clinical interest:

- $Sensitivity = \frac{TP}{TP+FN} = \frac{10}{10+5} = 0.67$
- $Specificity = \frac{TN}{TN+FP} = \frac{8}{8+7} = 0.53$
- $False\ negative\ rate\ (FNR) = 1 - Sensitivity = 1 - 0.67 = \frac{FN}{FN+TP} = \frac{5}{5+10} = 0.33$
- $False\ positive\ rate\ (FPR) = 1 - Specificity = 1 - 0.53 = \frac{FP}{FP+TN} = \frac{7}{7+8} = 0.47$

This hypothetical classification scheme produces both low sensitivity and specificity. That is, the classification scheme used would not detect neither the TP cases, nor the TN ones. Similarly would happen with both the FP and FN values.

1.5 Implementation of microarray-data results in clinical practise

1.5.1 RNA stability in human biopsies

1.5.1.1 RNA quality

The biological source for the detection of gene-expression based in the microarray technology is the messenger RNA (mRNA). The quality of the RNA is a critical point for a successful hybridisation [129]. The RNA quality can be measured by the absence of DNA and protein contamination[130], jointly with the evaluation of its integrity, which can be defined as the degree of fragmentation that an RNA specimen shows [129, 130]. An RNA sample with a high degree of fragmentation would not properly hybridize with the probes immobilized onto a microarray.

1.5.1.2 RNA integrity

Usually, the RNA integrity is measured by the ratio of the EtBr-stained 28S and 18S ribosomal peaks. Those specimens having a 28S/18S ratio approximately equal to 2 and with absence of bands running prior, between and/or after the ribosomal peaks in an agarose gel electrophoresis have acceptable RNA integrity[130]. However, the quantity of RNA (3-10 μg) required for agarose gel analysis can not be obtained from small biological samples.

As a result, new microfluidics stations were designed to reduce the amount of sample needed for analysis and software was developed to refine the detection of degraded RNA[130, 131]. One of these devices is the 2100 Bioanalyzer from Agilent Technologies.

2100 Bioanalyzer (Agilent Technologies, USA)

Based on a microcapillary electrophoretic system, twelve samples containing between 25 and 500 ng of RNA can be simultaneously analyzed. This device generates an electropherogram and a virtual gel derived from the fluorescence signal detected by a laser. The profile of the electropherogram serves to compute the 28S/18S ratio and to generate the RNA integrity number (RIN). The RIN ranges from 0 to 10, completely degraded and undegraded, respectively[129] (see figure 1.16).

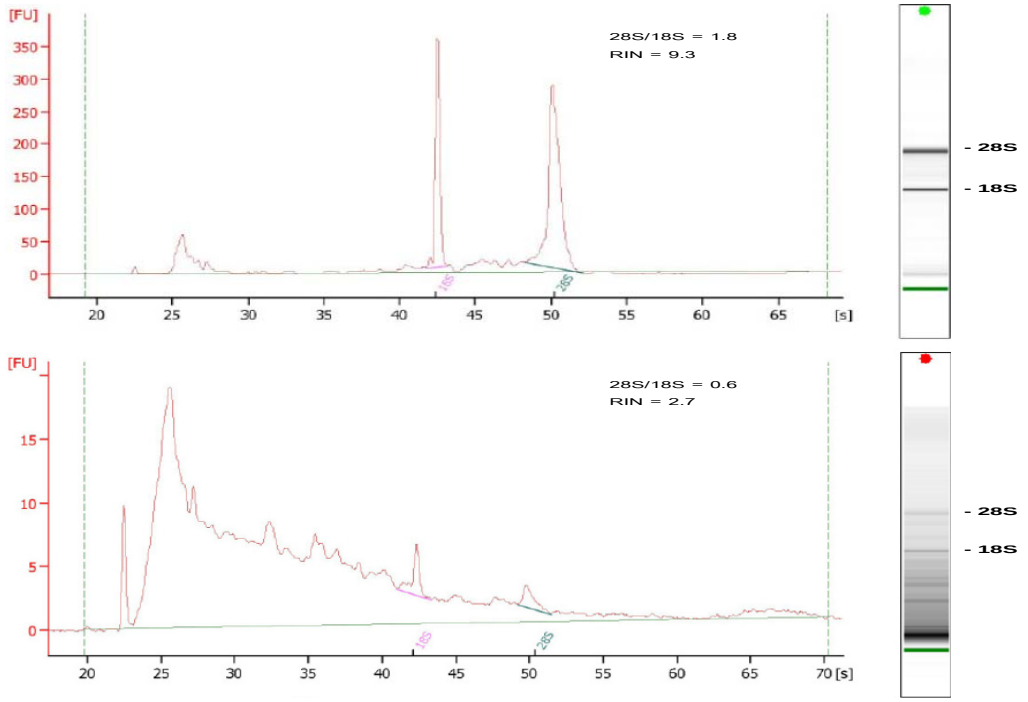


Figure 1.16: *2100 Bioanalyzer profiles.* A profile of an undegraded RNA (top) and a degraded RNA (bottom) are shown. The images at the right side corresponds to the virtual gel generated from the electropherogram.

The developers of the RIN algorithm argued that the 28S/18S ratio is not sensitive enough to detect fragmentation of RNA, since the computation of the ratio is restricted to the integration of the peaks, and does not consider the entire electrophoretic profile, as RIN does. Briefly, the RIN algorithm was obtained by computing the area under the 28S and 18S ribosomal peaks, and also under the regions placed before, between and after these peaks (see figure 1.17). The most informative features were selected and were used to train a neural network model [129].

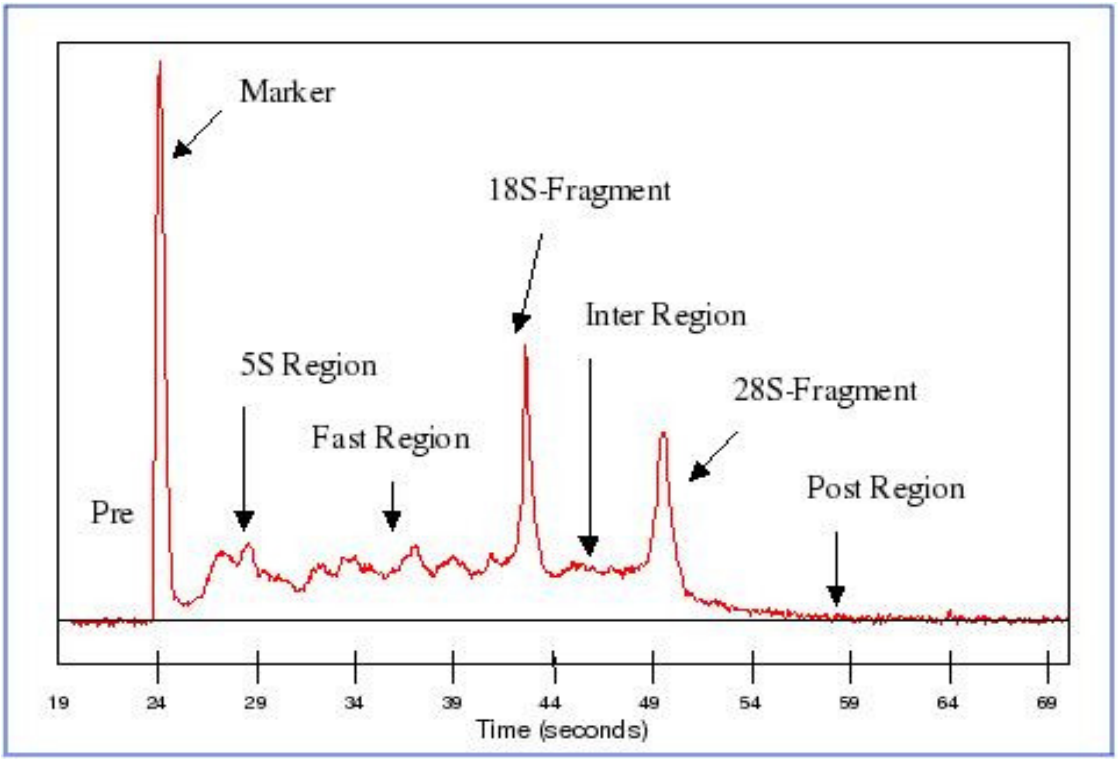


Figure 1.17: *RNA areas used for RIN computation.* The values of intensity of fluorescence obtained from the areas indicated in the figure (between 23 and 58 seconds) are used to computed the RIN number. The peak of the marker serves as a reference to identify the rest of peaks in the electropherogram, but it is not used for RIN computation.

The development of the 0 to 10 range of RIN was generated by comparing the area of the mentioned regions between RNA samples covering the whole range of degradation. The series of comparisons was the input for a Bayesian learning procedure resulting in a classifier that can identify the degree of degradation of a RNA profile [129].

1.5.1.3 Studies on RNA integrity

Effects of collection media

Micke and collaborators studied the effect of various collection media in a time course experiment to evaluate their capacity to preserve RNA[132]. In this experiment human tonsil and normal colon tissue were extracted from patients representing the malignant and benign pathological conditions, respectively.

Immediately after surgical removal samples were cut in cubes and four pieces snap frozen with isopenthane/dry (-120℃) ice and transferred to a -80℃ freezer as the reference 0h time-point. The rest of pieces were respectively placed in ice, RNA*later* (Ambion, Applied Biosystems, USA), 0.9% NaCl (only for tonsil

samples) or left at room temperature. After a period of 0.5, 1, 3, 6 and 16 hours, two samples were removed from each collection medium at each time point and frozen as above until RNA isolation.

Evaluation of the 28S/18S ribosomal peaks ratio did not show evident signs of degradation at any storage condition. Additionally, they evaluated the 28S/18S ratio of 47 biopsies from their local liquid nitrogen biobank accounting for a representative spectrum of tissues and conditions, which resulted in only 2 samples not showing clearly defined ribosomal peaks. Similarly, a previous study comparing gene expression differences between samples preserved after surgical operation at either room temperature, snap frozen or left in RNA*later* for 24 or 72h, resulted in no significative statistical changes[133].

Similar results were obtained by Scicchitano an collaborators[134], and Blackwall and collaborators[135], in their respective studies.

Assessment of different evaluation methods on RNA integrity

Strand and collaborators evaluated three methods of classifying RNA integrity using the 2100 Bioanalyzer for 24 snap frozen breast cancer biopsies: a) visual inspection (visible 28S and 18S peaks and flat baseline), b) 28S/18S ratio ($\geq$0.65) and c) RIN ($\geq$6)[136].

The visual inspection and the RIN method yielded the highest amount of samples classified with good integrity. To assess the dependency of the gene-expression on the RNA integrity, the RNA isolated from each breast cancer biopsy was hybridised onto a cDNA microarray. A hierarchical cluster was generated from the 24 breast cancer samples and all features of the microarray (16,641 genes). Poor or good RNA integrity was assigned to samples based on the three described methods.

Two groups were detected: one including most of degraded samples and the other one most of undegraded samples. However, the number of samples misclassified (expected to be degraded, but grouped by gene profile with undegraded, or *vice versa*) varied depending on the method. The visual inspection and RIN method produced better results, since they had only 1 and 2 misclassifications, respectively. In contrast, the cluster generated by labelling the samples with the ratio method produced a total of 5 misclassified samples[136].

1.5.2 Improvement of diagnosis and prognosis of HBT by microarray-based data

In recent years, the use of gene-expression microarrays has strongly increased the knowledge about molecular signatures underlying HBTs. Naturally, the tumour types receiving greatest attention have been the malignant forms of gliomas.

Anaplastic gliomas (Ags) and glioblastomas (Gbs) progression leads to a fatal outcome for the patient, although with heterogeneous time courses. Therefore, several studies have tried to correlate gene signatures with survival. For instance, Nutt and collaborators generated a classification model from 50 Gbs and 22 Ag[137]. Such model was used to correctly predict 18 out of 21 test samples (14 Gbs and 7 Ags). Surprisingly, significant correlation with survival was only obtained when using gene signatures from the developed model, rather than from the histological classification.

Similar results were reported by Freije and collaborators[69]. In this case, a dataset including 24 grade III and 50 grade IV gliomas was used to delineate two survival groups. The one with poorer life expectancy was enriched with 4/5 parts of Gbm cases. Among the 595 genes with fold-change >2 (two-sided t test with $P \leq 0.01$), four gene clusters correlating with survival were found: genes related to neurogenesis, genes involved in synaptic transmission, genes involved in mitosis and extracellular matrix components and regulatory genes.

Finally, an attempt to decipher the molecular mechanisms underlying HBT was performed by Tso and collaborators describing gene signatures characterizing primary and secondary glioblastomas[68]. In this study, authors selected 46 primary and 14 secondary glioblastomas. As a result, 73 glioblastoma-associated genes (GAGs) characterizing primary glioblastomas and 36 GAGs for the secondary glioblastomas were detected. However, 15 out of the total 113 GAGs belonged to a common functional category between primary and secondary glioblastomas. These 15 genes shared some functional categorization and are involved in mitosis and extracellular response-associated genes. In contrast, the secondary glioblastomas showed higher expression in several mitotic cell cycle-associated genes, whereas primary glioblastomas exhibited higher expression of several extracellular response-associated genes.

1.5.3 Developed clinical trials based on microarray-data

1.5.3.1 Relevance of clinical trials

The milestone to corroborate the applicability in the clinical practise of data generated from any new technology is the clinical trial. The molecular profile that can be derived from microarray experiments makes this high throughput technology an interesting option to detect a large set of differentially expressed genes to which target therapeutical action. The wide genetic variability of cancers represents a great difficulty to propose effective treatments that minimize side effects.

Determination of gene signatures for each patient is an attractive procedure that has evolved during the last decade. The possibility of conceiving the treatment to be individually managed has gained great acceptance in the scientific community. Such a conception completely transforms the current paradigm of medicine and, although this decision scheme seems far away to be implemented in present clinical practice, the preliminary steps translated to clinical trials have begun.

1.5.3.2 Steps in clinical trials

The process that leads gene signatures detected by microarray data to reach the clinical trials can be divided into three phases[76]:

1. Description of gene signatures related to cancer specimens.

2. Validation of the gene signatures by an independent test set.

3. Expansion of the cancer gene signatures to predict patient outcome and to guide the use of cancer therapeutics.

As an example of the first phase, several studies have described gene signatures in HBT that can account for the molecular characteristics of different high-grade gliomas[30, 68]. The second phase requires a higher amount of samples to discern the validity of the proposed gene signatures. In this sense, some valuable examples of high prediction accuracy have been reported in HBT research[138, 139, 140]. As an example of third phase, prediction of patient outcome has been addressed for HBT[69, 1]. Therefore, there are published examples that demonstrate the potential use of gene signatures from microarray experiments for clinical trials.

1.5.3.3 Standardization of microarray data prior to clinical trials

Despite that guidelines for gene signatures to reach the clinical trials stage seem well defined, there are several points to address regarding the validation of microarray results prior to those trials. At the beginning of the current decade, a key question, still in discussion, arose about how to translate the detected gene signatures into an user-friendly application in the clinical practise[141, 71]. The main proposals consisted in selecting a set of markers to be screened with a more accessible technology (i.e., immunochemistry or RT-PCR), or in contrast, to design a disease-specific microarray to use as a clinical test.

Another issue the microarray technology has to confront is the standardisation of the results[71, 141]. Due to the large availability of different microarray technologies, there is often a low reproducibility of gene expression levels or gene signatures[141]. Such a discrepancy can simply arise from unmatching gene sequences between microarray types[71, 141]. However, standardization is a topic being rigorously addressed. Several studies demonstrates that results obtained from distinct technologies can be compared[142, 143, 144].

The MicroArray Quality Control (MAQC) project assessed the intra- and inter-laboratory reproducibility of microarray results by using four different technologies (Affymetrix, Agilent, Applied Biosystems and GE HealthCare). This project was developed within a consortium in the United States promoted by the Food and Drug Administration, and it was composed of fifty-one centers representing the academia, the industry and the US government[142].

Comparability assessment

With regard to the comparability of microarray results, there is a request of high-impact factor journals to make raw data publicly available. This allows researchers to mutually compare gene signatures and/or improve the output by combination of data from different centres[71, 120]. Currently, the Minimum Information About a Microarray Experiment (MIAME)-compliant form is the usual document enabling worldwide distribution of data[145]. The MIAME-forms are asked to be submitted prior to publication to at least one publicly available repository such as the Gene Expression Omnibus (GEO) at the National Center of Biotechnology and Informatics (NCBI)(http://www.ncbi.nlm.nih.gov/geo/) or the Array Express at the European Bioinformatics Institute (EBI)(http://www.ebi.ac.uk/arrayexpress).

1.5.3.4 Reported clinical trials based on microarray data

There are some reported clinical trials in HBTs that were designed based on microarray data[146, 147], although not reaching the degree of clinical implementation as those developed for breast cancer[72, 148, 149]. Nonetheless, gene signatures from microarray data are used in studies with clinical implication to better describe the pathogenesis of HBTs. For instance, to elucidate response to therapy[150, 151, 152] or to identify prognostic markers[69, 1, 153].

At any rate, gene signatures must overcome the clinical trials before being implemented in the clinical practice. First, gene signatures must be tested on a representative population in the context of a phase II clinical trials. Second, gene signatures would be conducted in definitive phase III trials with relatively modest sample size, but large screening population[141]. That is, the selected population for phase III clinical trials must include people that accounts for a wide range of clinical factors (sex, age, ethnical origin,...). These steps appears not to have been addressed for microarray-based gene signatures yet.

1.5.4 Contribution of eTUMOUR, HealthAgents and MEDIVO2 projects to improve diagnosis and prognosis of HBT

The intensive study of HBTs during the recent years is expected to achieve an effective therapy, rather than the current palliative treatment improving patient healthcare. In Europe, the scientific policy of the European Union (EU) has fostered the development of projects with multiple partners. The multicentric studies undertaken in the EU have the advantage of collecting a great number of patient samples, which is not conceivable for single centers. Thus, this increases the reliability of the results obtained.

1.5.4.1 The eTUMOUR project

One of these projects is eTUMOUR (http://www.etumour.net), in which context this thesis has been developed. Its aim is to create a comprehensive Web-accessible Decision Support System (DSS) for analysis and interpretation of Magnetic Resonance Spectroscopy and Imaging (MRS & MRI) data of brain tumours, together with transcriptomic and metabolomic data. The DSS is expected to provide clinicians with a user-friendly diagnostic and prognostic tool, which will be implemented

in the clinical routine. The DSS can facilitate the decision about the treatment, based on the data generated from previous patients.

As all european projects, eTUMOUR is divided into work packages (WPs) that serve to distribute partners in groups to perform a determined task. There are nineteen partners involved in eTUMOUR from eight EU states and one from Argentina. Concerning this thesis, transcriptomic data was acquired and stored in the eTUMOUR database (https://dbtest.etumour.net:9091/eTumour/). The UAB also provided MRS, MRI and High Resolution Magic Angle Spining (HRMAS) data.

1.5.4.2 The HealthAgents project

Another project to which the UAB is providing resonance and microarrays data is the EU-funded HealthAgents (http://www.healthagents.net) project, which is also intended to improve HBT classification through a DSS. However, the manner to approach the subject is distinct from eTUMOUR. HealthAgents aims to create a data warehouse furnished with data from involved partners to build classifiers improving diagnoses and prognoses[154]. The data warehouse does not store all the information, but the local databases. As a consequence, a certain partner can not use all available data in the warehouse to refine its classifications. This is the innovating feature of the project since the exchange of data between partners is performed through the agents technology, which decides the amount of data that a partner can receive from another one, depending on its contributed data. Additionally, HealthAgents attempts to improve HBT classification by incorporating text mining tools, aside from the common machine learning tools.

1.5.4.3 The MEDIVO2 project

Finally, the *Mejora del diagnóstico y de la valoración prognóstica de tumores cerebrales humanos in vivo. Modelos animales para la metabolómica de la progresión tumoral. Fase 2* (MEDIVO2) was designed by the GABRMN group to improve the sensitivity of the non-invasive diagnosis and prognosis of HBT *in vivo*. Extraction of a metabolomic phenotype from single and multivoxel proton magnetic resonance (^{1}HMRS) is expected to allow the characterisation of diagnostic markers for HBT that could be incorporated into a DSS, which can assist clinicians to refine diagnosis and improve patient healthcare. In this sense, transcriptomic and HRMAS data is included to further improve diagnoses.

Another focus of MEDIVO2 using high field MRS/MRI consists in modelling tumour progression making use of genetically modified mice. Concerning this, characterisation of mobile lipids in cells and animal models permits detection of cell proliferation markers. Unlike the eTUMOUR and HealthAgents projects, MEDIVO2 is a Spanish government-funded grant, which is developed in collaboration with clinicians and surgeons from the *Hospital Universitari de Bellvitge-IDIBELL*, as well as with researchers from the *Universitat Politècnica de València* (UPVLC), the *Universitat Politècnica de Catalunya* (UPC), the *Katholieke Universiteit Leuven* (KUL) and the *Saint George's Hospital Medical School* (SGHMS).

Chapter 2

OBJECTIVES

1. **Collection of transcriptomic data from DNA microarrays**
   (a) Collect biopsies from intracranial human brain tumours at the hospitals from the Barcelona metropolitan area and establish a local biobank.
   (b) Isolate RNA from collected biopsies and analyze its integrity.
   (c) Hibrydize and analyze cDNA microarrays (CNIO).
   (d) Populate with gene expression microchips the eTUMOUR database.
2. **Characterization of RNA integrity from HBT biopsies**
   (a) Evaluate parameters that can influence the RNA quality of HBT biopsies.
3. **Development of a proof of principle by prediction of glioblastoma multiforme (Gbm) and meningothelial meningioma (Mm) biopsies using microarrays-based gene profiling**
   (a) Generate a prediction formula to distinguish Gbm and Mm biopsies using microarrays-based gene profiling.
   (b) Assess the ability of gene signatures to predict Gbm and Mm biopsies based on the histopathological diagnosis.
4. **Development of prediction models for various HBT types using Affymetrix microchips-based gene profiling**
   (a) Generate prediction models for those histopathological classes with highest prevalence.

(b) Develop an automatic strategy to determine gene signatures from Affymetrix data.

Chapter 3

MATERIALS AND METHODS

3.1 Collection, storage and histopathology analysis of samples

3.1.1 Collection and storage of samples

Biopsy samples collected for this work were obtained at different hospitals of the Barcelona area: *Hospital Universitari de Bellvitge*, *Hospital Universitari Germans Trias i Pujol*, *Hospital Clínic* and *Hospital Sant Joan de Déu*. Collection of biopsies was performed through 3 research projects: eTUMOUR, HealthAgents and MEDIVO2. The study was approved by the local Ethics Committee and informed consent was obtained from all patients (see annex A-1).

All collected biopsy samples were stored in liquid nitrogen in the surgery room after surgical removal from the patient brain. They were monthly taken to our laboratory and stored in liquid nitrogen at our local biopsy bank. For 33 of collected biopsies, an aliquot was also collected in RNA*later* in the surgery room immediately after surgical operation and stored at 4℃ until RNA isolation was performed.

3.1.2 Histopathologycal analysis of samples

Biopsy tumour samples were fixed in 4% buffered formalin and embedded in paraffin in the originating center. For routine histological examination 4-μm thick sections were stained with hematoxylin and eosin. The WHO 2000 Nervous System Classification criteria were used for diagnosis[18]. Moreover, at least one additional tissue section was prepared from the biopsy samples collected through the eTUMOUR project. Such a section was analysed by the Clinical Subcommittee of the Committee for Quality Control of Data as described at the deliverable 3 of the eTUMOUR project (see annex A-1).

3.1.3 Storage of data at the eTUMOUR and HealthAgents databases

The eTUMOUR database (eTDB, https://dbtest.etumour.net:9091/eTumour/) is the data warehouse of the project. All information available from patients must be entered into the eTDB to make it available to partners of the project. The eTDB can store clinical information, images of tissue slices and MRI, spectra obtained from single voxel MRS, multi voxel MRS and HRMAS, and transcriptomic data.

Each type of data is stored into a different section, in which there are several fields to describe the experiment performed (see annex A-2 for the case of quality control fields for transcriptomic data).

In the case of transcriptomic analysis, the fields included describe the quality of RNAs isolated and the quality of the hybridisation. Of course, information from hybridisation was only available for those RNA samples that displayed a sufficient quality, as agreed in the eTUMOUR project (see annex A-3).

The HealthAgents database (HADB, http://158.109.50.115:8091/haGUI/) approximately contains the same fields than the eTDB. However, the structure of this database differs from the eTDB one. In the eTDB each type of data is stored in a section independent from the rest. In contrast, in the HADB experimental data (MRI, MRS, HRMAS and transcriptomic data) is stored as a subsection of clinical data.

3.2 RNA isolation

3.2.1 Isothiocyanate-based RNA isolation (Qiagen)

RNA isolation based on isothiocyanate denaturation was performed by using the RNeasy Midi kit for those biopsies collected to perform cDNA microarray experiments, at the inital stages of the MEDIVO2 project. A specialized high-salt buffer system allows up to 1 mg of RNA longer than 200 bases to be adsorbed by the RNeasy silica-gel membrane. RNA molecules shorter than 200 bases (such as micro RNA, small-interfering RNA, 5.8S rRNA, 5S rRNA and tRNAs, which together comprise 15-20% of total RNA) are discarded.

Biological samples are first lysed and homogenized in the presence of a highly denaturing guanidine isothiocyanate containing buffer, which immediately inactivates RNases to ensure isolation of intact RNA. Ethanol is added to provide appropriate binding conditions, and the sample is then applied to the RNeasy column where the total RNA binds and contaminants are washed away. High-quality RNA is then eluted in RNase-free water[155].

3.2.2 Acid-Phenol:Chloroform-based RNA isolation (Ambion)

RNA isolation based on Acid-Phenol:Chloroform was performed by using the *mir*Vana miRNA Isolation kit[156] for those biopsies collected to perform Affymetrix microchip experiments. This part was performed during the eTUMOUR, HealthAgents and MEDIVO2 projects. This technology combines an organic extraction with a solid-phase extraction that allows isolation of total RNA including small RNA molecules (micro RNA, small-interfering RNA, 5.8S rRNA, 5S rRNA and tRNAs). The first step was to disrupt samples in a denaturing lysis buffer. Next, samples were subjected to Acid-Phenol:Chloroform extraction that removed most DNA[157]. At this point there were separate protocols for purification of either total RNA, including very small RNA species, or for purifying RNA highly enriched by small RNA species, which contained very little RNA larger than about 200 bases. In this thesis, only isolation of total RNA was performed.

Ethanol was added to samples, and they were passed through a Filter Cartridge containing a glass-fiber filter, which immobilized the RNA. The filter was then washed a few times, and finally the RNA was eluted with a low ionic-strength solution. To isolate RNA that was highly enriched for small RNA species, absolute ethanol was added to bring the samples to 25% ethanol. When this lysate/ethanol mixture was passed through a glass-fiber filter, large RNAs were immobilized, and the small RNA species were collected in the filtrate. The ethanol concentration of the filtrate was then increased to 55%, and it was passed through a second glass-fiber filter where the small RNAs become immobilized. This RNA was washed a few times, and eluted in a low ionic strength solution.

3.2.3 Evaluation of RNA quality

RNA was characterised using a NanoDrop spectrophotometer (NanoDrop Technologies). For RNA samples isolated using both RNeasy Midi kit and *mir*Vana miRNA Isolation kit, absence of protein contamination was monitored by the 260 nm/280 nm ratio of absorbance. In the former case, samples with a ratio ranging between 1.6 and 2.0 were accepted for further processing. In contrast, for RNA samples isolated using *mir*Vana miRNA Isolation kit, the accepted range was between 1.6 and 2.3 as agreed in the eTUMOUR project quality control document (see annex A-3).

Integrity of the RNA was assessed by using the capillary electrophoretic system

2100 Bioanalyzer (Agilent). For RNA samples isolated using RNeasy Midi kit, only those producing a 28S/18S ratio higher than 1.1 or an RNA integrity number (RIN) number higher than 5 were used for further analysis. For the other RNA samples isolated with the *mir*Vana kit, only those producing a 28S/18S ratio equal or higher than 1.2 or an RIN number equal or higher than 6 were selected as agreed in the eTUMOUR project protocol (see annex A-3).

3.3 Labelling and scanning

3.3.1 Single-Cy3 cDNA microarray labelling

Thirty-five RNA samples (17 glioblastoma multiforme, Gbm, and 18 meningothelial meningioma, Mg) isolated from those biopsies collected during the MEDIVO2 project were labelled and hybridised through the protocol described as follows. cDNA labelling was performed using the Cy3-fluorescent dye and the CyScribe First Strand labelling kit (GE Healthcare, UK). The starting material was approximately 14 μg of total RNA. Starting RNA was copied into cDNA using a reverse transcriptase and an oligo(dT) primer incorporating Cy3-dUTP into the growing cDNA sequence. Alkaline treatment was performed to eliminate the RNA template. Then, the cDNA labelled product was purified from the reaction mixture using the CyScribe^{TM} GFX^{TM} purification kit. Labelled cDNA was resuspended in 100 μl of the hybridisation solution, composed of 50% deionised formamide, 5x sodium saline citrate (SSC) and 0.1% SDS. Two μl human COT1-DNA (1 μg/μl), 2 μl polyadenilic acid (6 μg/μl) and 0.4 μl salmon sperm DNA (10 μg/μl) were added to avoid unspecific hybridisations.

The final solution was denatured for 2 minutes at 95℃ and immediately placed on ice. The solution containing the labelled cDNA was hybridised onto a prehybridised human CNIO oncochip for an overnight period in an incubator ArrayBooster (Advalytix, Munich, Germany). The human CNIO oncochip is a 12K cDNA microarray produced at the Spanish National Cancer Research Centre (CNIO Genomics Unit, ArrayExpress acc. no. A-MEXP-261) that contains 11,500 cDNA clones representing 9,300 loci. After incubation, slides were washed and Cy3-dye fluorescence was measured using a ScanArray 4000 (Perkin Elmer, Waltham, USA) detection system. Signal was quantified by the Genepix 6.0 software (Molecular Devices, Sunnyvale, USA).

3.3.2 Affymetrix labelling

RNA samples isolated from those biopsies collected during the eTUMOUR project were labelled and hybridised through the protocol described as follows. The whole procedure described in this section was performed at the Affymetrix core facility of the *Institut de Recerca de la Vall d'Hebron* (Barcelona, Catalunya).

Labelling was performed using the One-Cycle Target Labeling and Control Reagents kit (Affymetrix, USA). The starting material for the labelling protocol ranged from 0.3 to 5 μg of total RNA. First, the total RNA was reverse transcribed using a T7-Oligo(dT) promoter primer in the first-strand cDNA synthesis reaction, and four poli-A spike-in controls (poli-*lys*, -*phe*, -*thr* and -*dap*) included in each reaction sample to assess the batch-to-batch reproducibility of hybridisation.

Second, T4 DNA polymerase produced the double-stranded cDNA, which served as a template for *in vitro* transcription (IVT). The IVT reaction was carried out in the presence of T7 RNA polymerase and a biotinylated nucleotide analog/ribonucleotide mix for complementary RNA (cRNA) amplification and biotin labelling. The biotinylated cRNA targets were then cleaned up, fragmented, and hybridized onto the HG-U133 plus 2.0 GeneChip. Prior to hybridization, profiles of both amplified and fragmented material were monitored using the Bionalyzer (Agilent, USA).

For hybridization, 15 μg of fragmented cRNA were added to the hybridisation mix composed of 5 μl control oligonucleotide B2, 15 μl 20X Eukaryotic Hybridization Controls (*bioB*, *bioC*, *bioD*, *cre*), 3 μl herring sperm DNA (10 mg/ml), 3 μl BSA (50 mg/ml), 150 μl 2X hybridization buffer and water up to a final reaction volume of 300 μl. The hybridization mix was denatured at 99℃ for 5 minutes and transferred to 45℃ for 5 additional minutes. The hybridisation mix was spun down for 5 minutes to remove any insoluble material from the solution and 200 μl loaded into the HG-U133 plus 2.0 GeneChip. After an incubation period of 16 hours in the hybridization oven at 45℃ and 60 rpm, microchips were washed and stained by adding a solution that contained Streptavidin-phycoerythrin Biotinalyted anti-streptavidin antibody. Images were obtained by the software provided with the GeneChip Scanner 3000. This software automatically adjusts the intensity of the laser and the photomultiplier.

For a more detailed explanation about labelling and hybridisation Affymetrix protocols see reference [84].

3.4 Prediction of Gbm and Mm using single-labelling cDNA microarrays data

3.4.1 Foreword

The study to predict Gbm and Mm using single-labelling cDNA microarrays data was performed through a collaboration between UAB and its associated clinical centres and UPVLC. The article derived from this study and accepted for publication in *Diagnostic Molecular Pathology*, displays Xavier Castells and Juan-Miguel García-Gómez, both as first co-authors. To the effect of their respective PhD thesis, both co-authors agree in that Xavier Castells performed the RNA isolation of biopsies, hybridization of isolated RNA, scanning and analysis of microarrays, the RT-PCR experiments and the functional analysis of differentially detected genes. They also agree in that Juan-Miguel García-Gómez performed the statistical analysis of data including background correction and normalisation, detection of differentially expressed genes and development of a prediction formula.

3.4.2 Data pre-processing

Prior to the computations to obtain the predictor on our dataset, a pre-processing step to make the expression values comparable among microarrays was performed. Due to the specific protocol used in this study (single-labelling cDNA-based microarrays), non-standard pre-processing methods derived from adaptations of the Affymetrix pre-processing methods were set up and applied to our data.

First, a visual inspection of the scan images, discarded experiments having an artefactual signal in at least one microarray experiment. Artefactual signals were considered to be those signals on a spot that came spreading from other close by spots or were generated by dust or other contaminants sticking to the microarray. Foreground values were corrected using the background smoothing procedure defined by Edwards in reference [90]. Genes with negative intensity signal (foreground minus background) in more than 20% of cases in each of Gbm and Mm groups were also removed. Data was normalized using the *average reference loess* [90]. Afterwards, genes that were not validated by the microchip manufacturer (CNIO) by PCR evidence (single band) and sequence verification were removed. Finally, signals corresponding to genes spotted more than once in the microarray were averaged.

3.4.3 Feature selection and sample classification

Statistical significance was assessed by the non-parametric Mann-Whitney (MW) test on data from the training dataset (10 Gbm and 11 Mm). Afterwards, p-values of the MW test were corrected for the false discovery rate (FDR) control obtaining the so called q-values[115].

Genes with q-value lower than 0.02 were considered to be differentially expressed. Starting from the set of differentially expressed genes, selection of 3 or 4 genes through a stepwise (SW) procedure was performed.

Linear models based on Rank Reduced Linear Discriminant Analysis (LDA) were fitted to our data. Hence, given a set of samples a projection that maximized the separation between projected values of both classes was searched for. Prediction accuracy was evaluated by randomly sampling the training dataset 200 times. That is, 15 samples, following the distribution frequency of the classes in the dataset, were selected to train the predictor and 6 samples to validate its result. Such a resampling procedure provided an estimation of the prediction accuracy. The final evaluation of the predictors performance was carried out in a totally independent test dataset (7 Gbm and 7 Mm) with the labels blind to the testers. The ability to produce a single predictor for direct use in clinical routine was demonstrated by generating an LDA-based predictor with the four most selected genes across the 200 iterations. Such an LDA-based predictor was developed over the training dataset (10 Gbm and 11 Mm) and its performance tested over the independent dataset (7 Gbm and 7 Mm).

3.4.4 Functional analysis of gene signatures

Aiming to determine a gene signature that may characterize each tumour type based on the expression levels, a hierarchical cluster was performed with the 629 genes with the q-value lower than 0.02. Furthermore, the selected gene subset was submitted to the web-based Database for Annotation, Visualization and Integrated Discovery tool (DAVID)[158] with the purpose of detecting statistically significant functional gene groups with differential expression between classes. In our study, we chose the highest stringency level among the five stringency levels provided by DAVID for a set of genes to be called a functional group.

3.4.5 RT-PCR validation

Total RNA (100 ng) was used as starting template RNA for reverse transcription in a total volume of 25 μl, which included 2.5 μl of primers. We used the validated primers Quantitect Primer Assays (Qiagen) and the one-step Quantitect SYBR Green RT-PCR kit (Qiagen), on a Smart Cycler (Cepheid) system. Sixty cycles composed of 3 steps were performed: denaturation for 15 seconds at 95℃, annealing for 30 seconds at 50℃and elongation at 72℃. The Gbm/Mm ratio was calculated using the $2^{-\Delta Ct}$ method[159].

3.5 Exploratory analysis of meningioma and glial tumours using Affymetrix data

3.5.1 Data pre-processing

3.5.1.1 Background correction and data normalisation

We uploaded the *.cel* files into a SGI Altix 350 remote cluster composed of 30 processors Intel Itanium of 64 bits at 1.5 GHz with 64 Gb of shared memory RAM (Suse SLES9 / SGI ProPack 4). For more details see webpage of the cluster (*http://cibercluster.upf.edu/EN/Pages/que_es.aspx*), which is maintained by the group lead by Dr. Alejandro Frangi. The subsequent steps and the development of the prediction models were run into the described cluster, to which we had access as a partner of the scientific network *Centro Investigación Biomédica en Red-Bioingeniería, Biomateriales y Nanomedicina* (CIBER-BNN).

We processed the *.cel* files processed using the *affy* and *affyPLM* R packages as described in the annex A-5. Briefly, an *AffyBatch* object was created by using the *ReadAffy* function. The *AffyBatch* object contained all probesets prior combination of replicates (summarisation). As probesets are composed of 11-20 probes (25-mer oligonucleotides), a comparison of the intensities at 5'-end versus those at the 3'-end, can provide an estimation of the integrity of transcripts. The *affy* package enables such a verification through the degradation plots.

After such a verification, the *AffyBatch* was used to test three different combinations of background correction and normalisation methods:

1. **Robust Microarrays Analysis (RMA)** background correction and **quantile** normalisation.

2. **Microchip Analysis Suite 5 (MAS5)** background correction and **scaling** normalisation.

3. **No** background correction and **scaling** normalisation.

These three approaches were generated by using the *fitPLM* function, included in the *affyPLM* package (see annex A-5). The approach yielding the lowest variability between cases, was selected to develop prediction models. Data variability was assessed by plotting a boxplot, an MA, a Relative Log Expression (RLE) values and a Normalised Unscaled Standard Errors (NUSE) and a density plot for each approach.

See A-5 for a more detailed explanation on how this analysis was run in the R software.

3.5.2 Generation of prediction models

3.5.2.1 Grouping of samples

From the cases considered to develop prediction models, 4 main groups were created comprising the HBT types and subtypes of highest incidence:

- **Glioblastoma (Gb).**
- **Anaplastic glioma (Ag).**
- **Low grade gliomas (Lgg).**
- **Meningiomas (Mg).**

Pairwise predictors for all possible combinations among these 4 groups were performed. Furthermore, three-class predictors were generated for two additional discrimination problems: **Mg-Lgg-Gb** and **Lgg-Ag-Gb**.

3.5.2.2 Statistical analysis

Splitting of samples

The optimization of the analysis to reduce prediction overfitting was the main objective in this part of the work. For such a reason, resampling procedures, based on leave-one-out (LOOCV) and 5-fold cross validation (5FCV), were implemented to split data into training and test set. In the case of LOOCV, data splitting was repeated as many times as samples included in the complete dataset (training and

test). In the case of 5FCV, the splitting was repeated five times the total number of samples included in the dataset. Performing all possible combinations by leaving 1/5 of samples apart from training would have been highly time consuming and it was avoided. Moreover, the frequency of samples per tumour type was maintained equal to the one in the complete dataset at each iteration of the 5FCV approach.

As a result, for each iteration of cross validation, the accuracy of the prediction model was computed. Therefore, a vector with length equal to the number of iterations performed depending on the cross validation method was generated.

Feature selection

Regardless the cross validation approach, feature selection was performed only on data selected for training, as described by [120]. Two methods were used:

- **P-values:** computation of p-values was performed following the multiple-test correction method described by Benjamini and Yekutieli[160].
- **PCA:** reduction of variables was performed by computing the principal components of the considered cases and the 54,675 probesets.

The corrected p-values (or q-values) were computed using the *linear models for microarray data (limma)* package. Only those genes with fold-change equal or higher than 2 and q-value <0.05 were considered as input for the prediction algorithms tested. In case there were no probesets below the cutoff, the 100 genes of highest fold-change and lowest p-values were selected. For the three class comparison problems, the p-values were first computed for each pairwise comparison. Second, those genes with p-value <0.05 across all three pairwise comparisons were selected. If there were no common genes with p-value <0.05, the union of genes with p-value <0.05 at each pairwise comparison were selected and their q-values computed.

The principal components were computed for the whole dataset (training and test) by the *prcomp* function, which is included in the *stats* package.

Prediction algorithms

To assess the relevance of algorithms in supervised class prediction[120], three different methods were tested:

- **Linear discriminant analysis (LDA)** from the *MASS* package.
- **Support vector machines (SVM)** from the *e1071* package.

- **Random Forests (randF)** from the *randomForest* package.

Statistical significance of prediction models

As a result of the combination of all resampling, feature selection and algorithm methods, several prediction models were generated. Apart from the prediction accuracy to asses the performance of each model, the statistical significance of prediction was computed in each case. For that, the class of cases was randomly assigned and all prediction models generated again. Using a *Wilcoxon-test*, the prediction p-value was computed by comparing the prediction accuracies obtained from the correct labelling with those from the random labelling.

Integration into an R function

The three R functions from prediction algorithms (LDA, SVM and randF) are not implemented into an overarching one, which could be used to automatically run the explained strategy. For such a reason, an R function was developed to integrate a proper resampling and feature selection procedure, as well as to test three different prediction algorithms.

Accordingly, the *MultiClassPred* function was developed. Such a function enables a proper estimation of the prediction accuracy for the combination of LOOCV or 5FCV with p-values or PCA-based feature selection. For any of those combinations, *LDA*, *svm* and *randF* predictions algorithms were computed. Moreover, six different sets of input variables (probesets or PCA variable) were used.

From the object generated by the *MultiClassPred* function, different prediction parameters of clinical interest were computed:

- The prediction accuracy mean based on the area under the curve (AUC), as described in reference[161].
- The p-value derived from the comparison of prediction values obtained from correctly and randomly labelling of cases.
- The maximum and minimum prediction accuracy obtained across the performed iterations.
- The sensitivity and specificity for each tumour type when performing a 3-class predictor. In the case of a pairwise predictor, only one sensitivity and specificity are computed.

- The false negative rate (FNR) and false positive rate (FPR). The class dependency of FNR and FPR computation is identical to that for the sensitivity and specificity.

See annex A-5 and A-6 for a more detailed explanation on how this analysis was performed in the R software.

3.5.3 Glioblastoma subtypes

3.5.3.1 Assessment of statistically significance of clusters

A hierarchical cluster based on the euclidean distance was calculated by using the *heatmap_2* function from the *Heatplus* R package. As a first step to verify the reliability of clusters visually detected in the hierarchical cluster, a *k-means* cluster composed of 2, 3, 4 and 5 clusters was computed.

The *silhouette* statistics from the *cluster* R package was computed for each of the generated *k-means* clusters. This *silhouette* statistics is a measure of dissimilarity of a determined cluster with respect to its neighbour clusters. Its value ranges from 0 to 1, being 1 the highest dissimilarity.

See annex A-5 for a more detailed explanation on how this analysis was performed with the R software.

3.5.3.2 Data pre-processing of NMR data

Single voxel data was acquired at the *Centre Diagnòstic Pedralbes-Institut d'Altes Tecnologies* using a General Electric (GE) spectometer. Raw data was pre-processed as described in reference [162].

HRMAS data was acquired and pre-processed by Dr. Daniel Valverde Saubí, as described in his Phd thesis[163].

3.6 Generation of a murine glial tumour model to simulate *ex vivo* ischaemia at normal body temperature in brain tumour biopsies

3.6.1 Animals and cells

A total of 29 C57BL/6 female mice, 20-23 g weight, were used in this study. These were obtained from Charles River Laboratories (France) and housed at the animal facility of the *Universitat Autònoma de Barcelona*. All animal studies were approved by the local ethics committee, according to the regional and state legislation (protocol DARP-3255/CEEAH-530). GL261 mouse glioma cells were obtained and cultured exactly as described by Quintero and collaborators[164].

3.6.2 Inoculation of the mice brain with GL261 tumour glial cells

Tumors were induced in 29 mice by intracranial sterotactic injection of 10^5 GL261 cells in the caudate nucleus. About 15 min after being given a dosis of analgesia (Meloxicam subcutaneous, s.c., 1.0 mg/Kg), animals were anesthetized (Ketamine-Xylazine, 80-10 mg/kg intraperitoneal, i.p.) and then immobilized in a stereotactic holder (Kopf Instruments, Tujunga, USA). The skull was exposed and a high speed micro-driller (Fine Science Tools, Heidelberg, Germany) used to make a small hole in its surface (1mm): 2.3 mm to the right of the midline, as measured from the Bregma. A 26 G Hamilton syringe (Hamilton, Reno, USA), positioned on a digital push-pull microinjector (KD Scientific, Hollisto, USA), was advanced through this hole, 2.3mm from the cortical surface into the striatum, to deliver 10^5 GL261 cells (in 4 μl RPMI medium) at a rate of 2 μl/min. The syringe was slowly removed 3-5 min after the injection had finished and the scission site closed with suture silk (5.0). Animals were left to recover from anaesthesia in a warm environment ($\approx$25 ℃) and, as they began to wake up, a stronger analgesic (opioid) was given: Buprenorphine s.c., 0.1 mg/kg. Meloxicam analgesia was repeatedly administrated at 24 and 48 hours post-surgery.

Formation of the tumour mass was detected 3-5 days after inoculation by MRI, and necrosis 2 weeks after inoculation, approximately.

For the purpose of the experiment, 9 mice were sacrificed when necrosis was

monitored by MRI (more than 2 weeks post-inoculation) using a Bruker Biospec 7T spectrometer (Wissembourg, France), which was fitted with a specific probe for mouse brain, essentially as described in reference[165]. The remaining 20 mice were sacrificed when a non-necrotic intracranial tumour mass was detected by MRI (less than 2 weeks post-inoculation), essentially as described in reference[165].

3.6.3 Experimental procedure to simulate *ex vivo* ischaemia at normal body temperature of brain tumour samples

3.6.3.1 Animal sacrifice and encephalon removal

Simulation of ex vivo ischemia at normal body temperature was performed in 4 out of the 9 necrotic tumour mice and in 14 out of the 20 non-necrotic tumour mice. Animals were sacrificed by an intraperitoneal injection of sodium pentobarbital (60 mg/ml) at a dose of 200 mg/kg. When the animal did not respond to mechanical stimulus in the legs, the head was sectioned from the rest of the body by cutting with sterile scissors. With the same scissors, the upper part of the skull was removed by an incision at each occipital condyle and cutting in anterior direction up to the nasal cavity. The encephalon was removed by lifting it up with sterile dissection tweezers at the resulting cavity from the process of sectioning the occipital condyles.

3.6.3.2 Dissection of the tumour mass

The tumour cells-inoculated hemisphere and a thin layer of the other hemisphere were separated by using a sterile scalp and dissection tweezers. The layer of the non-inoculated hemisphere was included to obtain the maximal tumour mass, in the case contra-lateral hemisphere invasion occurred. Taking as a reference the point left on the encephalon by the inoculation puncture, cerebral parenchyma was progressively removed down. The tumour mass was characterised by its mucous appearance.

In the case of necrotic tumours, the tumour mass was clearly identifiable due to its darker colour. In contrast, the identification of the tumour mass in the case of non-necrotic tumours was more difficult, since their colour was closely similar to the non-tumour parenchyma.

The whole procedure described in this section was performed at room temperature. The time elapsed from animal death until the tumour mass was extracted,

ranged between 5 and 7 minutes. All steps requiring manipulation of mice and GL271 cell cultures were performed by Rui Simoes, Teresa Delgado and Milena Acosta from GABRMN.

3.6.3.3 Simulation of *ex vivo* ischaemia at normal body temperature

Immediately after tumour mass resection, an aliquot (<1 mm^3) was submerged into formol for posterior histological verification of necrosis in the investigated tumours. Such verification was performed by Professor Martí Pumarola (Àrea de Medicina i Cirurgia Animal, Facultat de Veterinària, UAB) using standard protocols (paraffin-embedded and hematoxylin/eosin-stained tissue slides).

To simulate body temperature, tumour masses were introduced into separated 1.8 ml criotubes pre-filled with PBS at 37ºC. Samples were incubated for 30 minutes and snap-frozen in liquid nitrogen after this period. Furthermore, 7 out of the 14 non-necrotic mice tumours were incubated 15 minutes instead of 30 minutes. Those specimens not subjected to normal body temperature incubation were snap-frozen in liquid nitrogen immediately after dissection.

3.7 Simulation of *ex vivo* ischaemia at normal body temperature in C6 cells

3.7.1 Culture and harvesting of C6 cells

Cells were cultured essentially as described by Valverde and collaborators[166]. Culture medium was removed from the plate by aspiration with a Pasteur pipette connected to a vacuum-water pump. To remove any trace of medium, 10 ml of PBS pre-heated at 37℃ was added. Cells were detached from the plate by enzymatic digestion with 2 ml trypsin-EDTA (0.5 g porcine trypsin and 0.2 g EDTA per 100 ml) (Sigma, USA), pre-heated at 37℃. When cells were clearly detached from the plate, 8 ml of culture medium pre-heated at 37℃ was added to stop trypsin digestion.

The cell suspension was centrifuged at 4000g for 2 minutes. The supernatant was removed and 5 ml of PBS pre-heated at 37℃ added to remove any trace of trypsin. A second centrifugation was performed, and the supernatant was discarded.

3.7.2 Simulation of *ex vivo* normal body temperature ischaemia

The cell pellet was transfered to a 1.8 ml cryogenic tube containing PBS pre-heated at 37℃. Incubation for 30 minutes at 37℃ was performed for cell pellets at logarithmic phase (n=3) and cell pellets at post-confluence (n=3). Furthermore, 3 additional cell pellets at logarithmic phase were incubated for 15 minutes instead of 30 minutes. After incubation at 37℃, the supernatant was discarded and cell pellets snap frozen in liquid nitrogen after removal of the supernatant. Other cell pellets were transfered to an empty cryotube and snap frozen in liquid nitrogen after removal of the supernatant from the previous centrifugation.

Chapter 4

RESULTS AND DISCUSSION

4.1 Discrimination of Gbm and Mm using cDNA-microarrays data

4.1.1 Results

4.1.1.1 Collection of biopsies

A total of 78 biopsy samples were collected at the *Hospital Universitari de Bellvitge* in the context of the MEDIVO2 research project. Among them, 38 samples were diagnosed by the anatomopathology service of the hospital as glioblastoma multiforme, 31 as meningothelial meningioma, 6 as carcinoma metastasis, 4 as adenocarcinoma metastasis, 1 as anaplastic astrocytoma, 1 as astrocytoma WHO grade II and 1 as schwanoma. For the object of this section, the RNA isolated from 35 biopsies (17 glioblastoma multiforme (Gbm) and 18 meningothelial meningioma (Mm)) accomplished the criteria of RNA integrity described in section 3.2.3. The percentage of samples with RNA degraded was 55.3% in Gbm and 41.9% in Mm.

4.1.1.2 Pre-processing and prediction results

The methodology described in the data pre-processing section 3.4.2 was applied to both training and test datasets to attenuate the effect of possible variability due to non-biological causes in CNIO microarrays. In our experiments, the total number of analysed probes per microarray was 27,648. After pre-filtering, 23,652 features remained in the expression matrix. The next step consisted in discarding a gene from further processing when more than 20% of samples produced negative signals (foreground minus background) in each group, Gbm and Mm. After background correction and the normalization steps, removal of genes that were not validated by the microchip manufacturer (CNIO) by PCR evidence (single band) and sequence verification yielded 15,584 features. Averaging of feature replicates gave rise to a final expression matrix of 7,218 features for the 35 samples investigated (training and test datasets).

Starting from this number of genes, those differentially expressed in each tumour type were investigated. Considering that Gbm and Mm are both histological and pathologically highly distinct brain tumour types, graphical discrimination of these two tumour types by simply plotting values of differentially expressed genes was expected. For this purpose, the Gbm/Mm ratio for each gene was computed and genes displaying the highest and lowest Gbm/Mm ratios (see table 4.1) were used

to create a graphical classifier (see figure 4.1**A** and **B**). Interestingly, the gene with the highest Gbm/Mm ratio was the glial fibrilar acidic protein (GFAP) with a value higher than 400. The protein encoded by this gene is a well known biological marker of glial cells. Furthermore, desmoplakin (DSP) showed the lowest Gbm/Mm ratio, being 250-fold more expressed in Mm compared to Gbm (Gbm/Mm=0.004), see table 4.1.

A predictor based on LDA was developed (see figure 4.1**C**). Our aim was two-fold, in the first place, to profit from the panoply of genes available to build a better and potentially more robust predictor than the expression of a single gene product (GFAP or DSP). Secondly, we were interested in detection of gene signatures providing biological information about the underlying molecular mechanisms, which may characterise Gbm and Mm in such a pairwise comparison. The predictor was built by splitting the full dataset in twenty-one samples (10 Gbm and 11 Mm) for training and cross-validation purposes. Additionally, fourteen totally independent and blinded samples (7 Gbm and 7 Mm) for the testers were used for the final evaluation purposes. Statistical significance was computed by using the Mann-Whitney nonparametric test on the training set and genes with a corrected q-value less than 0.02 were selected. As a result 629 genes were found differentially expressed from the initial 7,218 gene set.

This set of 629 genes was used to generate a prediction model with three or four randomly selected genes from a SW selection procedure. Two hundred repetitions were performed, from which two hundred LDA different predictors were obtained. As a result, a 95% prediction accuracy mean was obtained. Concerning our blind test dataset of fourteen samples, an observed accuracy of 100% with a [70%, 100%] confidence interval was obtained across the 200 iterations.

To demonstrate the ability to produce a predictor that could be used in an automated way once developed, all training samples (21) were used to fit the final models. Therefore, the final predictor shown in figure 4.1C was calculated selecting the four more selected genes across the 200 prediction iterations: *GFAP*, *PTPRZ1*, *GPM6B* and *PRELP* (see table 4.1). Such a predictor produces an objective and automated prediction result by simply introducing the pre-processed and normalised gene expression values into the LDA formula:

$$\begin{aligned} DSC = &-0.394 * GFAP - 0.397 * PTPRZ1 \\ &- 0.397 * GPM6B + 0.365 * PRELP \end{aligned} \tag{4.1}$$

where DSC is the discriminant score.

Gene symbol	Accession number	Gene description	Gbm/Mm ratio	Selection frequency
GFAP	AA069414	*Glial fibrillary acidic protein*	413	16
PTPRZ1	AA476460	*Protein tyrosine phosphatase, receptor-type, Z polypeptide 1*	357	9
GPM6B	AA284329	*Glycoprotein M6B*	133	9
PRELP	AA131664	*Proline/arginine-rich end leucine-rich repeat protein*	0.042	9
FABP7	W72051	*Fatty acid binding protein 7, brain*	221	7
EGFL3	AA975413	*EGF-like-domain, multiple 3*	0.299	7
PDE4B	AA453293	*Phosphodiesterase 4B, cAMP-specific (phosphodiesterase E4 dunce homolog, Drosophila)*	26	5
OMD	N32201	*Osteomodulin*	0.011	5
LAPTM4A	AA398233	*Lysosomal-associated protein transmembrane 4 α*	0.346	5
USP25	AA479313	*Ubiquitin specific peptidase 25*	0.110	5
NFATC3	AA293819	*Nuclear factor of activated T-cells, cytoplasmic, calcineurin-dependent 3*	0.399	4
CTGF	AA598794	*Connective tissue growth factor*	0.106	4
PIGT	H83225	*Phosphatidylinositol glycan, class T*	0.508	4
FLJ39155	R08141	*Hypothetical protein FLJ39155*	0.019	4
DSP	H90899	*Desmoplakin*	0.004	4
GAS1	AA025819	*Growth arrest-specific 1*	0.084	4
PLK1	AA629262	*POLO-like kinase 1 (Drosophila)*	4	4
NEK6	AA463188	*NIMA (never in mitosis gene a)-related kinase 6*	4	4
TNXB	T58430	*Similar to tenascin XB isoform 1; tenascin XB1; tenascin XB2; hexabrachion-like [Pan troglodytes]*	0.118	4
LHX2	AA018276	*LIM homeobox 2*	36	4
MGC21621	W52061	*MAS-related GPR, member F*	0.041	4
PDGFD	AI005125	*Platelet derived growth factor D*	0.034	4
IL27RA	AI088984	*Interleukin 27 receptor α*	2	4
-	AI249137	*Transcribed locus*	0.205	4
CCND1	R81200	*Cyclin D1*	0.109	4
NCAM2	AI306467	*Neural cell adhesion molecule 2*	6	4
SH3GL3	AI359676	*SH3-domain GRB2-like 3*	0.079	4
MT2A	BF131311	*Metallothionein 2A*	11	4
NUDT1	AA443998	*Nudix (nucleoside diphosphate linked moiety X)-type motif 1*	3	4
RARRES2	AA481944	*Retinoic acid receptor responder (tazarotene induced) 2*	0.196	3
SMARCD3	AA035796	*SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily d, member 3*	2	3
HYAL1	AA464791	*Hyaluronoglucosaminidase 1*	0.159	3
CDK2AP1	R78607	*CDK2-associated protein 1*	2	3
CTNND2	H04985	*Catenin (cadherin-associated protein), δ 2 (neural plakophilin-related arm-repeat protein)*	44	3
CYB5	R92281	*Cytochrome b5 type A (microsomal)*	0.333	3
CA2	H23187	*Carbonic anhydrase II*	17	3
OAT	AA446819	*Ornithine aminotransferase (gyrate atrophy)*	0.383	3
GPM6A	AA448033	*Glycoprotein M6A*	90	3
HSPC195	R63735	*CXXC finger 5*	4	3
ZMYM6	W81504	*Zinc finger, MYM-type 6*	0.567	3
PPARGC1A	N89673	*Peroxisome proliferative activated receptor, gamma, coactivator 1, α*	0.257	3
TEK	H02848	*TEK tyrosine kinase, endothelial (venous malformations, multiple cutaneous and mucosal)*	0.140	3
APM2	AA478298	*Chromosome 10 open reading frame 116*	0.034	3
FGL2	H56349	*Fibrinogen-like 2*	0.022	3
CDH2	W49619	*Cadherin 2, type 1, N-cadherin (neuronal)*	20	3
CANPL1	H15456	*Calpain 1, (μ/I) large subunit*	0.532	3
GPR17	R44664	*G protein-coupled receptor 17*	45	3
LOC119504	AA004832	*Chromosome 10 open reading frame 104*	0.504	3
-	W52340	-	0.355	3
DHRS3	AA171606	*Dehydrogenase/reductase (SDR family) member 3*	0.217	3

Table 4.1: *Genes with highest discriminant capacity.* We show the fifty more selected genes across the 200 iterations of the SW resampling approach over the training dataset to estimate the prediction accuracy. The two genes harbouring the highest Gbm/Mm gene-expression ratios are the most selected genes in the training. The complete list of selected genes is avalailable at annex A-7.

The cut-off point at 0 enables objective prediction between the two tumour types. Negative values are Gbms, while positive values denote Mm (see annex A-8). Using this predictor, a 100% prediction of the independent test set was obtained.

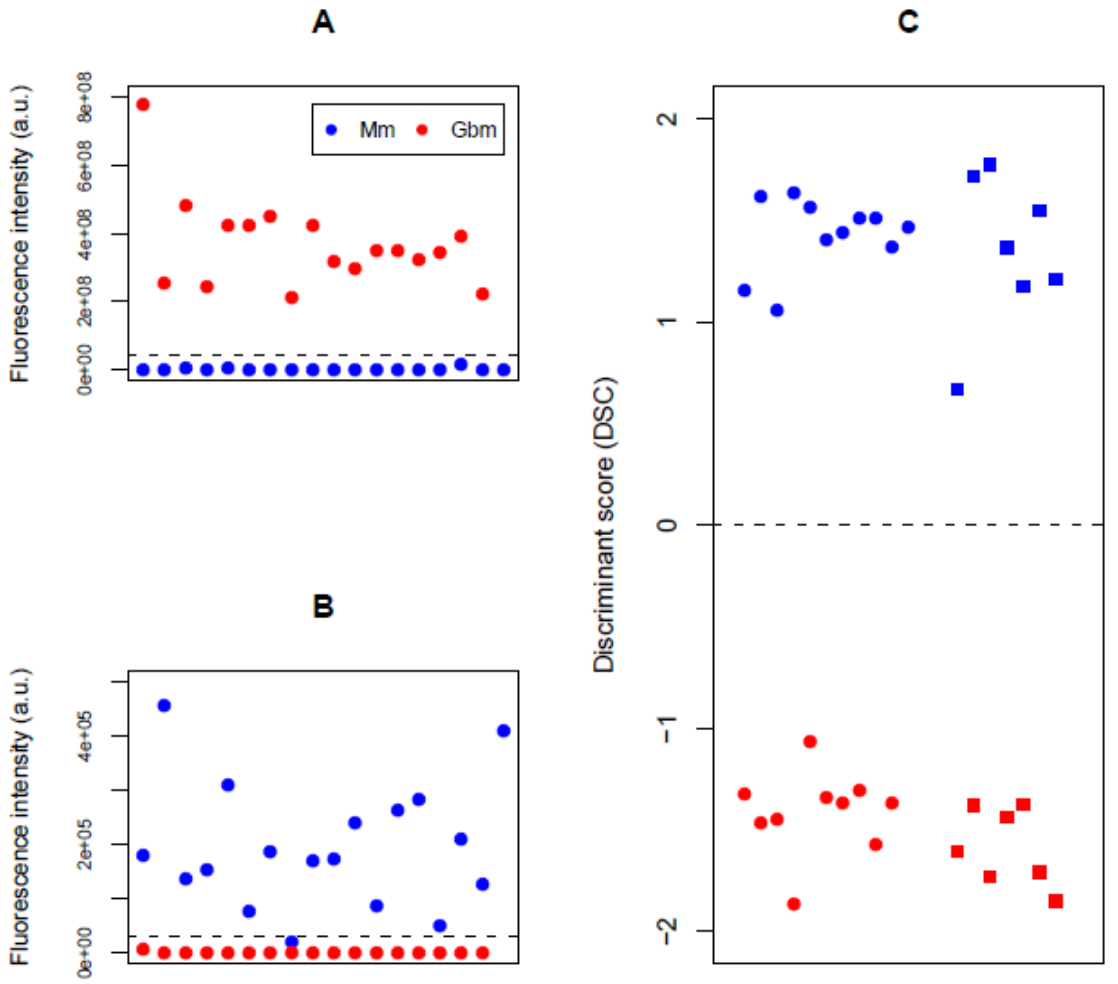


Figure 4.1: *Graphical representation of normalised data and LDA-based predictor.* Red symbols correspond to Gbm samples, while blue symbols denote Mm samples. **A, B)** Scatter plot of normalised expression values of genes showing the highest and lowest Gbm/Mm ratio, GFAP and DSP respectively for all Gbm and Mm cases (see table 4.1 for abbreviations meaning). Samples were arbitrarily distributed along the x axis, while along the y axis fluorescence intensity values (a.u.) were plotted. **C)** Discriminant scores obtained from the LDA-based predictor generated using normalised expression values from *GFAP*, *PTPRZ1*, *GPM6B* and *PRELP* genes, those most selected across the 200 iterations, which perfectly separated these two tumours class members. Circles are training samples, while squares are test samples. Along the y axis discriminant scores at the latent space are shown.

4.1.1.3 Molecular characterization of Gbm and Mm biopsy cases

Aiming to detect a broader gene signature that could also characterise and differentiate Gbms and Mms, the subset of genes with q-value lower than 0.02, was used to perform the hierarchical cluster shown in figure 4.2. Genes were initially grouped in sixteen clusters clearly defining a specific profile for each tumour type. Furthermore, determination of functionally-related groups of genes was assessed by subjecting the mentioned gene subset to the DAVID tool. A total of eleven functional groups with p-value lower than 0.05 accounting for eighty genes were obtained. Interestingly, three out the eleven clusters were exclusively composed by genes either overexpressed in Gbm or in Mm (see table 4.2). Functional group 2 contained genes overexpressed in Mm that belong to the family of the small leucine rich proteoglycans (SLRPs): *FMOD* (**J**), *PRELP* (**G**), *OMD* (**G**), *BGN* (**J**) and *OGN* (**G**).

Similarly, functional group 11 was composed of five members belonging to the cytochrome family: *CYP1B1* (**J**), *CYP4Z1* (**B**), *CYB5* (**J**), *CYP4B1* (**B**) and *CYP3A5* (**B**). On the other hand, functional group 6 was composed of several isoforms of genes encoding tubulins overexpressed in Gbm: *TUBA1* (**I**), *TUBA2* (**I**), *TUBA3* (**F**), *TUBB* (**F**), *TUBB2* (**F**), *TUBB4* (**I**), *TUBA4A* (**I**) and *TUBA8* (**I**). Also detected by the DAVID tool when setting the medium stringency level, a large number of metallothionein isoforms were highly expressed in Gbm (see figure 4.2): *MT1H* (**K**), *MT1F* (**I**), *MT1X* (**K**), *MT2A* (**F** and **I**) and *MT3* (**I**). Concerning the remaining nine functional groups composed of genes overexpressed in both Gbm and Mm, the cluster with highest and lowest statistical significance harboured a collection of cadherin and cytochrome isoforms, respectively (see table 4.2). Interestingly, functional groups 7, 8, 9 and 10 were mainly composed of genes encoding proteins somewhat related to cell signalling: signal receptors (group 7), G-protein receptors (group 8), Ras proteins (group 9) and tyrosine kinases (group 10). Groups 3, 4 and 5 were apparently enriched with proteins related to the extra-cellular matrix and the cell-cell adhesion complexes.

Gene symbol	Accession number	Gene description	Gbm/Mm ratio	Selection frequency
Functional group 1		**P-value: 0.000047**		
CDH3	AA425556	*Cadherin 3, type 1, P-Cadherin (Placental)*	0.014	1
DSG2	W37448	*Desmoglein 2*	0.041	-
PCDH17	AA669075	*Protocadherin 17*	8	-
CDH2	W49619	*Cadherin 2, type 1, N-Cadherin (Neuronal)*	19.7	3
PCDH1	R77512	*Protocadherin 1 (Cadherin-like 1)*	5	1
CDH10	R14164	*Cadherin 10, type 2 (T2-Cadherin)*	4	-
PCDH9	R38168	*Protocadherin 9*	17	-
CDH5	H02884	*Cadherin 5, type 2, VE-Cadherin (Vascular epithelium)*	0.20	-
CDH1	AI671174	*Cadherin 1, type 1, E-Cadherin (Epithelial)*	0.052	2
CDH11	AA136983	*Cadherin 11, type 2, OB-Cadherin (Osteoblast)*	0.22	-
Functional group 2		**P-value: 0.000080**		
FMOD	AA486471	*Fibromodulin*	0.14	-
PRELP	AA131664	*Proline/Arginine-rich end Leucine-rich repeat protein*	0.042	9
OMD	N32201	*Osteomodulin*	0.011	5
BGN	BE262957	*Biglycan*	0.41	-
OGN	AA045327	*Osteoglycin (Osteoinductive factor, Mimecan)*	0.0050	1
Functional group 3		**P-value: 0.000090**		
DSCAM	N64532	*Down syndrome cell adhesion molecule*	10	1
GHR	AA775738	*Growth hormone receptor*	0.34	1
IL6ST	AA775738	*Interleukin 6 signal transducer (GP130, Oncostatin M receptor)*	0.52	2
IFNGR1	BE973918	*Interferon γ receptor 1*	0.21	-
LEPR	H51066	*Leptin receptor*	0.35	1
Functional group 4		**P-value: 0.00014**		
AGTRL1	R58969	*Angiotensin II receptor-like 1*	21	1
GPM6A	AA448033	*Glycoprotein M6A*	90	3
GPR4	AI492409	*G protein-coupled receptor 4*	0.15	-
TMA4SF7	AA100696	*Tetraspanin 4*	0.13	2
SDC2	H64346	*Syndecan 2 (heparan sulfate proteoglycan 1, cell surface-associated, fibroglycan)*	0.042	1
TMA4SF13	W86202	*Tetraspanin 13*	14	2
Functional group 5		**P-value: 0.00017**		
NCAM2	AI306467	*Neural cell adhesion molecule 2*	6	4
DSCAM	N64532	*Down syndrome cell adhesion molecule*	10	1
ALCAM	R13558	*Activated leucocyte cell adhesion molecule*	0.092	2
JAM2	AA410345	*Junctional adhesion molecule 2*	0.094	3
JAM3	H73479	*Junctional adhesion molecule 3*	0.065	-
Functional group 6		**P-value: 0.00027**		
TUBB4	BX100915	*Tubulin β 4*	8	1
TUBA8	BF195571	*Tubulin α 8*	6	-
TUBA1	AA180912	*Tubulin α 1 (Testis specific)*	4	1
TUBA2	AA426374	*Tubulin α 2*	5	1
TUBA4A	AA626698	*α-Tubulin isotype H2-α*	55	-
TUBB	AI672565	*Tubulin β 2A*	14	-
TUBB2	AI000256	*Tubulin β 2C*	2	2
TUBA3	AI865469	*Tubulin α 3*	6	2
Functional group 7		**P-value: 0.00083**		
INSR	T47312	*Insulin receptor*	0.17	-
PDGFRA	H23235	*Platelet-derived growth factor receptor, α polypeptide*	19	-
BMPR1A	AA991180	*Bone morphogenetic protein receptor type Iα*	0.42	2
KIT	H23235	*Hardy-Zuckerman 4 feline sarcoma viral oncogene homolog*	0.23	-
TEK	H02848	*Tyrosine kinase endothelial (venous malformations, multiple cutaneous and mucosal)*	0.14	3
EPHA7	N91461	*EPH receptor A7*	0.031	-
ERBB3	AA664212	*Erythroblastic leukemia viral oncogene homolog 3 (avian)*	14	2
RAGE	N77779	*Renal tumour antigen*	0.48	1
RYK	T77810	*receptor-like tyrosine kinase*	0.27	-
TYRO3	BM665421	*Protein tyrosine kinase*	4	2
FGFR4	AA446994	*Fibroblast growth factor receptor 4*	0.18	1

Continued from last page

Gene symbol	Accession number	Gene description	Gbm/Mm ratio	Selection frequency
Functional group 7		**P-value: 0.00083**		
INSR	T47312	*Insulin receptor*	0.17	-
PDGFRA	H23235	*Platelet-derived growth factor receptor, α polypeptide*	19	-
BMPR1A	AA991180	*Bone morphogenetic protein receptor type Iα*	0.42	2
KIT	H23235	*Hardy-Zuckerman 4 feline sarcoma viral oncogene homolog*	0.23	-
TEK	H02848	*Tyrosine kinase endothelial (venous malformations, multiple cutaneous and mucosal)*	0.14	3
EPHA7	N91461	*EPH receptor A7*	0.031	-
ERBB3	AA664212	*Erythroblastic leukemia viral oncogene homolog 3 (avian)*	14	2
RAGE	N77779	*Renal tumour antigen*	0.48	1
RYK	T77810	*receptor-like tyrosine kinase*	0.27	-
TYRO3	BM665421	*Protein tyrosine kinase*	4	2
FGFR4	AA446994	*Fibroblast growth factor receptor 4*	0.18	1
Functional group 8		**P-value: 0.0026**		
GPR4	AI492409	*G protein-coupled receptor 4*	0.15	-
GPR17	R44664	*G protein-coupled receptor 17*	45	-
P2RY5	R91539	*Purinergic receptor P2Y, G-protein coupled 5*	0.36	1
FZD4	AA677200	*Frizzled homolog 4 (Drosophila)*	0.13	2
CCRL2	AI288845	*Chemokine (C-C motif) receptor-like 1*	8	1
AGTRL1	R58969	*Angiotensin II receptor-like 1*	21	1
FZD7	H71474	*Frizzled homolog 7 (Drosophila)*	0.12	2
GPR153	AA777493	*G protein-coupled receptor 153*	3	2
MGC21621	W52061	*MAS-related GPR, member F*	0.041	4
RAMP1	BE262882	*Receptor (Calcitonin) activity modifying protein 1*	19	1
Functional group 9		**P-value: 0.0044**		
ARL7	N35301	*ADP-ribosylation factor-like 4C*	12	1
ARF4L	AA878652	*ADP-ribosylation factor-like 4D*	0.082	1
RAB9A	H98534	*Member Ras oncogene family*	2	-
RAB31	AA432084	*Member Ras oncogene family*	2	1
RAB33A	AI360342	*Member Ras oncogene family*	6	1
RRAS2	R21415	*Related Ras viral oncogene homolog 2*	0.21	2
RASD1	BM674708	*Ras dexamethasone-induced 1*	19	-
ARHN	AI027909	*Rho family GTPase 2*	13	1
RALB	W15297	*v-Ral simian leukemia viral oncogene homolog B (Ras related; GTP binding protein)*	2	2
Functional group 10		**P-value: 0.011**		
PLK1	AA629262	*POLO-like kinase 1 (Drosophila)*	4	4
RIPK1	AA426324	*Receptor (TNFRSF)-interacting Serine-Threonine kinase 1*	0.63	2
NEK6	AA463188	*Never in mitosis gene A-related kinase 6*	4	4
PRKCN	AA417816	*Protein kinase D3*	0.37	-
PRKACB	AA459980	*Protein kinase, cAMP-dependent, catalytic β*	5	2
RAGE	N77779	*Renal tumour antigen*	0.48	1
PRKCM	N53380	*Protein kinase D1*	5	-
PRKCD	AA496360	*Protein kinase C δ*	0.39	1
PRKCH	AA128274	*Protein kinase C η*	0.19	1
Functional group 11		**P-value: 0.036**		
CYP1B1	AA448157	*Cytochrome P450, family 1, submafily B, polypeptide 1*	0.093	-
CYP4Z1	H21977	*Cytochrome P450, family 4, submafily Z, polypeptide 1*	0.22	-
CYB5	R92281	*Cytochrome B5, type A (microsomal)*	0.33	3
CYP4B1	AA291484	*Cytochrome P450, family 4, submafily B, polypeptide 1*	0.028	-
CYP3A5	BF062953	*Cytochrome P450, family 3, submafily A, polypeptide 5*	0.13	-

Table 4.2: *Functional analysis of gene with q-value lower than 0.02.* The eleven functional clusters arising from the DAVID tool are depicted. The mentioned tool enables a stringency range for a set of genes to be considered a differentially expressed functional group. The depicted table was computed using the highest stringency and selecting those groups with p-value lower than 0.02.

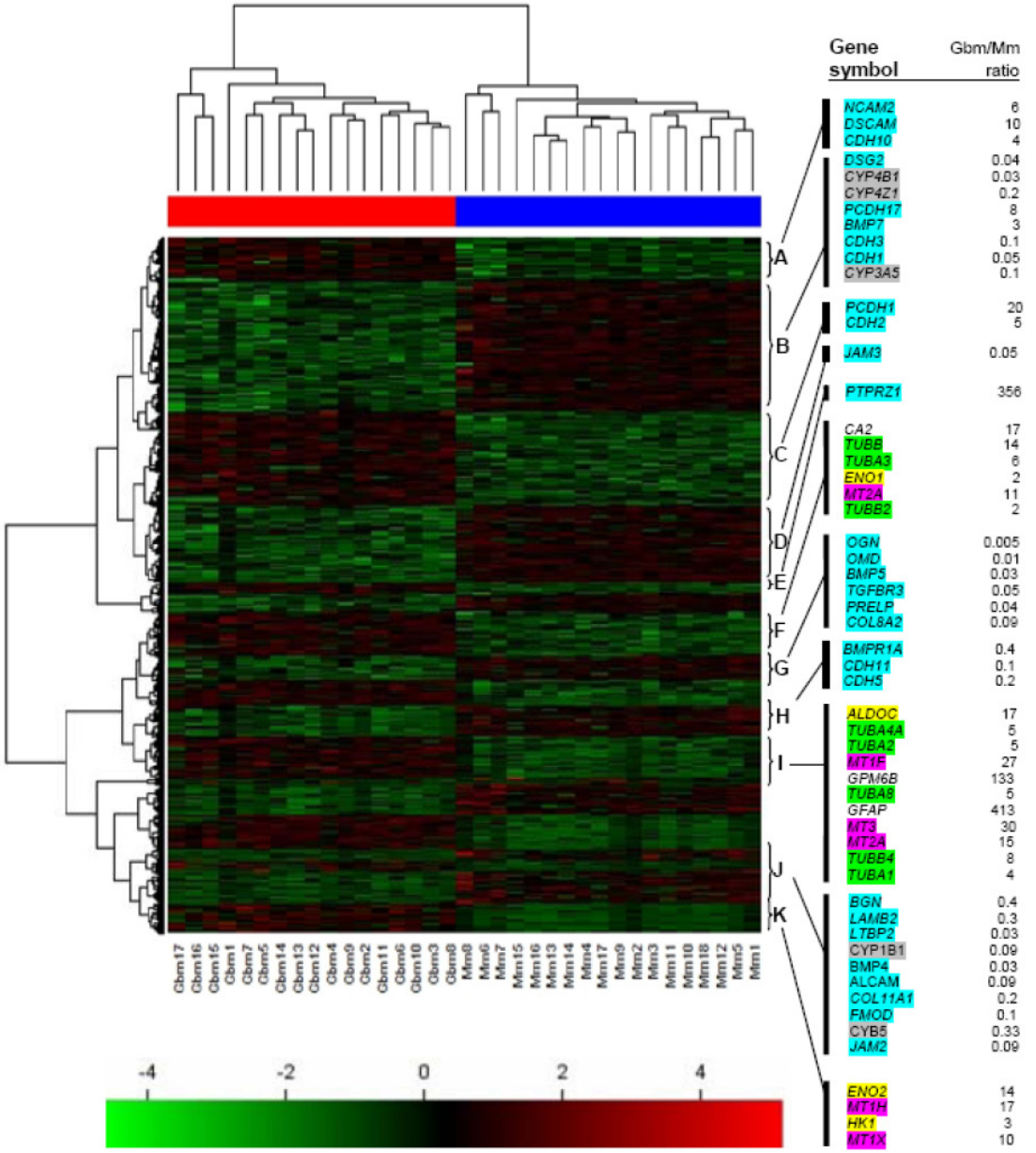


Figure 4.2: *Hierarchical cluster of differentially expressed genes from the training set.* Graphical illustration of the hierarchical cluster performed across samples and genes computed using Euclidean distance. Columns are samples and rows are genes. The 629 genes with a p-value lower than 0.02 were used to generate this cluster with the 35 samples of the full dataset. At the top of the figure, Gbm and Mm samples are denoted by red and blue bars respectively. At the right margin, letters (A-K) indicate clusters of genes belonging to functional families or related to different signalling pathways. Groups or gene-families are specifically coloured: metallothioneins (purple), tubulins (green), glycolysis-related genes (yellow), cell membrane and/or the extracellular matrix (blue) and cytochrome-related genes (grey). *GFAP*, *PTPRZ1*, *GPM6B* and *PRELP* are also included within the hierarchical cluster. See annex A-7 for description of genes.

4.1.1.4 RT-PCR expression results

A subset composed of 6 samples, 3 Gbm and 3 Mm, were subjected to RT-PCR for validation purposes of transcriptomic levels detected from microarray experiments. Selection of these samples was based on RNA integrity and concentration, as well as absence of protein contamination. Among differentially expressed genes in our study, a subset to be subjected to RT-PCR was selected based on four criteria: maximal or minimal Gbm/Mm ratio and minimal FDR corrected p-value, maximal selection at the re-sampling of the training set and biological relevance for tumour progression. Using such criteria selected genes were: *GFAP*, *PTPRZ1*, *GPM6B*, *MT3*, *CA2*, *TUBB*, *APM2*, *PRELP*, *OGN* and *DSP*. The six first genes were far more expressed in Gbm than in Mm, and the four last genes showed opposite behaviour. As observed in table 4.3, RT-PCR confirmed the expression profile of the mentioned genes according to the type of tumour evaluated.

		cDNA microarrays	RT-PCR
Gene symbol	**Accession number**	**Gbm/Mm ratio**	**Gbm/Mm fold-change**
GFAP	AA069414	813	50419
PTPRZ1	AA476460	378	517
GPM6B	AA284329	149	157
MT3	AI362950	27	61
CA2	H23187	17	28
TUBB	AI672565	9	42
PRELP	AA131664	0.022	0.04
APM2	AA478298	0.019	0.03
DSP	H90899	0.0016	0.0007
OGN	AA045327	0.00076	0.0005

Table 4.3: *Comparison of gene-expression values between microarrays and RT-PCR.* This table shows the Gbm/Mm ratios derived from gene-expression values from cDNA microarrays and RT-PCR. Ratios were obtained from expression values originating from the three Gbm and the three Mm selected samples that were subjected to RT-PCR. For cDNA microarrays, the ratio was obtained by dividing the average of expression values from the 3 Gbm by that from the 3 Mm. For RT-PCR this ratio was computed as the power of 2 to the negative difference between the Ct average from the 3 Gbm and 3 Mm samples subjected to RT-PCR ($2^{-(Ct(Gbm)-Ct(Mm))}$).

4.1.1.5 Expression level of *GFAP*, *PTPRZ1*, *GPM6B* and *PRELP* in Affymetrix-based hybridisation cases

Subsequent to demonstrate the ability of a prediction formula based on the expression level of 4 genes obtained from cDNA microarray experiments, we explored the expression level of the 4 mentioned genes in glioblastomas (Gbs) and meningiomas (Mgs), whose gene-profile was obtained from Affymetrix microchips.

We considered 3 Affymetrix microchip datasets:

1. **UAB1**: 32 Gbs and 12 Mgs collected at the UAB for the eTUMOUR project and hybridised onto the HG-U133 plus 2.0 microchip.

2. **UAB2**: 17 Gbs and 19 Mgs collected at the UAB for the eTUMOUR project and hybridised onto the HG-U133 plus 2.0 microchip.

3. **Pubmed**: 67 Gbs (GDS1976) and 31 Mgs (GSE9438) made publicly available at the *Gene Expression Omnibus* DataSets from the *National Center for Biotechnology Information* (NCBI) database (http://www.ncbi.nlm.nih.gov/sites/entrez?db=gds). Gbs had been hybridised onto HG-U133 A and B Affymetrix microchips, whereas Mgs had been hybridised onto HG-U133 plus 2.0 microchips.

As can be seen in table 4.4, multiple Affymetrix probesets are represented for each gene, except for the *PTPRZ1*. Therefore, we also explored the probeset providing the highest or lowest Gb/Mg expression ratio.

Among the three probesets representing the *GFAP* gene, the 203540_at provided the highest Gb/Mg ratio across the 3 Affymetrix datasets. Curiously, the mentioned probeset codes for the isoform 1 of the *GFAP* gene, whereas the additional two probesets are hypothetical alternative splicings.

Concerning the *GPM6B* gene, the four first probesets depicted in table 4.4 showed similar Gb/Mg expression ratios across the three Affymetrix datasets. Strikingly, only the gene product of the 209170_s_at probeset has been characterised.

With respect to the *PRELP* gene, the 204223_at probeset provided the lowest Gb/Mg expression ratio and its gene product is known.

In summary, the probesets 203540_at (*GFAP*), 204469_at (*GPM6B*), 209170_s_at (*GPM6B*) and 204223_at (*PRELP*) were selected to verify the prediction formula robustness for Affymetrix-based hybridisation cases.

Accession Number	ProbeSet Affymetrix	Gene symbol	Locus Link	UniGene	Gb/Mg ratio			Sequence mRNA length (bp)	Sequence protein length (aa)
					Affymetrix eTUMOUR UAB 1	Affymetrix eTUMOUR UAB 2	Affymetrix Pubmed		
J04569	203539_s_at	*GFAP*	2670	Hs.514227	0.88	0.95	1.22	3017	432
NM_002055	203540_at	*GFAP*	2670	Hs.514227	40.12	327.78	37.65	3081	432
AL133013	229259_at	*GFAP*	2670	Hs.514227	51.66	46.31	27.00	3279	438
NM_002851	204469_at	*PTPRZ1*	5803	Hs.489824	156.52	544.20	125.81	8169	2314
AI419030	209167_at	*GPM6B*	2824	Hs.495710	58.90	57.53	28.19	458	-
AW148844	209168_at	*GPM6B*	2824	Hs.495710	74.86	46.35	36.59	415	-
N63576	209169_at	*GPM6B*	2824	Hs.495710	72.07	60.04	25.41	396	-
AF016004	209170_s_at	*GPM6B*	2824	Hs.495710	68.16	53.63	23.93	1642	265
AA194253	236116_at	*GPM6B*	2824	Hs.495710	1.02	0.83	0.98	232	-
AL041745	240286_at	*GPM6B*	2824	Hs.495710	1.22	0.87	1.39	533	-
NM_002725	204223_at	*PRELP*	5549	Hs.632481	0.084	0.069	0.096	5833	382
AA573140	228224_at	*PRELP*	5549	Hs.632481	0.17	0.14	0.24	470	-
AI190575	231366_at	*PRELP*	5549	Hs.632481	1.09	1.04	0.98	451	-
U41344	37022_at	*PRELP*	5549	Hs.632481	0.52	0.45	0.52	924	382

Table 4.4: *Affymetrix probesets representing the 4 genes of the prediction Gbm/Mm formula (see equation 4.1).* Biological information about the probesets representing the 4 genes (*GFAP*, *PTPRZ1*, *GPM6B* and *PRELP*) selected to compute the Gbm/Mm formula are depicted. From left to right, the accession number, the Affymetrix probeset, the gene symbol, the locus link and the unigene identifiers are given. The next three columns are the Gb/Mg expression ratio for each Affymetrix dataset. Finally, the length of the mRNA sequence that each probeset represents and the length of the corresponding protein are shown. A dash indicates that the protein sequence is unknown.

4.1.1.6 Verification of the formula robustness by prediction of Affymetrix-based hybridisation cases

To further test the robustness of the developed discriminant formula using cDNA microarrays gene-profile, we predicted the class of a set of 12 meningiomas (Mg) and 32 glioblastomas (Gb), whose gene profile was obtained from HG-U133 plus 2.0 Affymetrix microchips (see also section 4.2). As can be seen at figure 4.3, all samples were correctly classified in their class group.

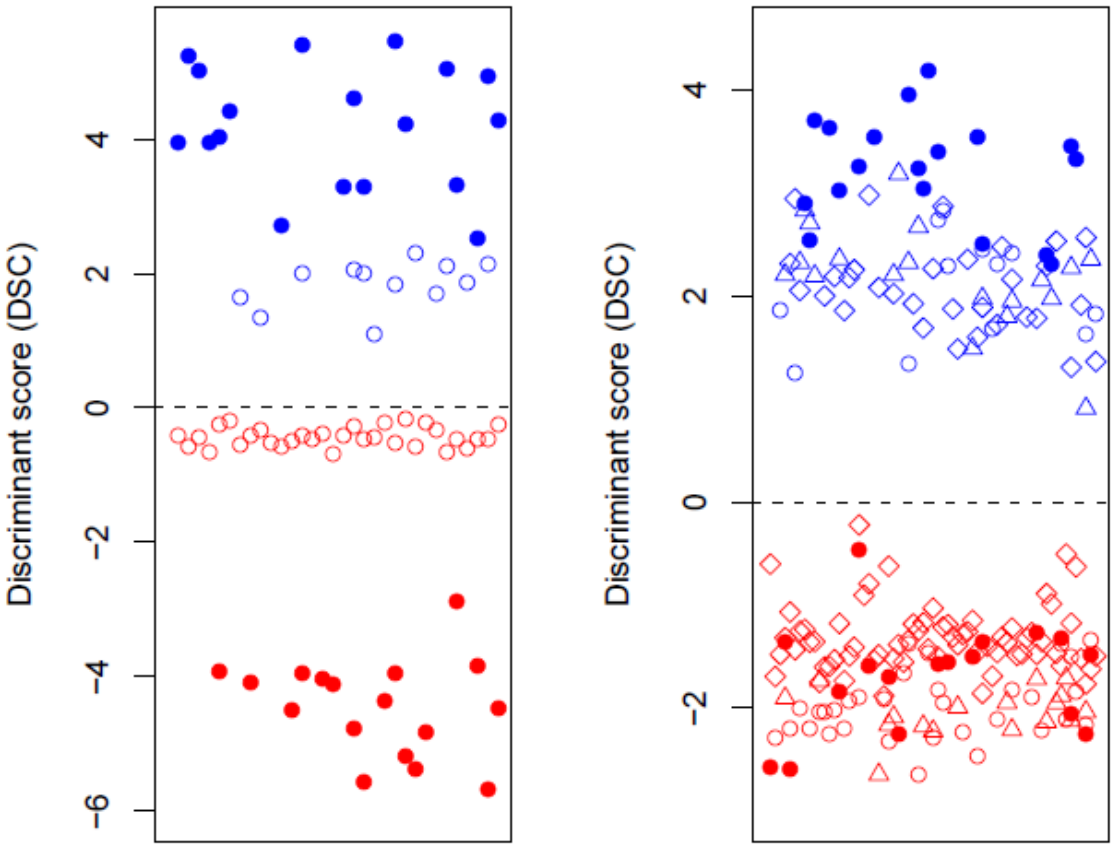


Figure 4.3: *Prediction of Affymetrix-based gene-profile Gb and Mg cases.* Robustness of the developed formula using cDNA microarrays and based on 4 genes (*GFAP*, *PTPRZ1*, *GPM6B* and *PRELP*), was assessed by prediction of 3 datasets. First, 32 Gb and 12 Mg cases, for which the gene-profile was obtained from Affymetrix microchips (plot on the left). Red symbols are glioblastomas and blue symbols indicate meningiomas. Solid symbols denote Gbm and Mm cases, from which the gene expression profile was obtained by using cDNA microarrays. Correspondingly, open circles denote those 32 Gb and 12 Mg Affymetrix microchips-hybridised samples. On the right, the DSCs were computed using the optimized formula. The second independent test set composed of Affymetrix cases (17 Gbs and 19 Mgs) is denoted with open triangles. Third, the 67 Gbs and 31 Mgs obtained from publicly available data are denoted by empty blue rhombus for Mg cases, while empty red rhombus indicate Gb cases. Along x axis cases are arbitrarily distributed and the discriminant score (DSC) plotted along the y axis.

It can be seen that the glioblastoma cases from Affymetrix data were placed close to the discriminant threshold, and above those Gbm cases from cDNA microarrays data. Taking this result into account, we applied a correction factor to the discriminant coefficients of the formula, so that a better grouping could be achieved. As a consequence, the discriminant coefficients of the developed formula (see equation 4.1) were adjusted as described in annex A-8 and summarized herein:

$$\begin{aligned} DSC = {} & 0.078 * GFAP - 0.6207 * PTPRZ1 \\ & - 0.670 * GPM6B + 0.660 * PRELP \end{aligned} \tag{4.2}$$

In doing so, an increased grouping was obtained for glioblastomas and enlarged the distance with respect to meningiomas (see figure 4.3).

The increase of prediction ability by doing such a correction was verified with a second test set composed of 17 Gbs and 19 Mgs (see figure 4.3).

Furthermore, we predicted 31 Mgs (GSE9438) and 67 Gbs (GDS1976) publicly available cases from the *Gene Expression Omnibus* DataSets from the *National Center for Biotechnology Information* (NCBI) database. The gene-expression profile of Mgs had been obtained from the HG-U133 plus 2.0 Affymetrix microchip, whereas the gene expression profile of Gbs were from the HG-U133 A and B Affymetrix microchip. Downloaded *.cel* files were normalised as described in section 3.5.1.1. As depicted in figure 4.3, all publicly available cases were correctly predicted.

These results demonstrate that the developed formula (see equation 4.2) can completely predict glioblastoma and meningioma cases, regardles the microarray technology used to obtain the gene expression profile.

4.1.2 Discussion

4.1.2.1 Development of an automated predictor based on gene signatures of brain tumours

Several studies have demonstrated usefulness of data generated from gene-expression based microarrays to classify brain tumours when they cannot be properly discriminated by using histological and image-based morphologic examinations alone[167, 168, 1, 169]. These studies have focused on recognizing tumour molecular subtypes of Gbm and classification of histologically distinct Mm, but no automated predic-

tor for classical histological WHO types was made available for public use. This issue was considered highly relevant within the context of the European project eTUMOUR, as a required proof of principle prior to attempting automated and objective recognition of tumour types or grades difficult to ascertain by classical histology (i.e. glioblastoma multiforme molecular subtypes)[2, 3, 5]. In our study we have demonstrated that 100% successful automated prediction between glioblastoma multiforme and meningothelial meningioma tumours is achievable without subjective data judgement. Furthermore, such a prediction was performed using single-labelling cDNA microarrays. Being a protocol of lower cost and greater design flexibility than other genomic technologies, it points to an attractive experimental option for routine use.

The particular type of microarray protocol used (single-labelling cDNA microarrays) forced us to develop a novel algorithm for data pre-processing to correct background and normalise experiments. Discrimination of Gbm and Mm was first performed by plotting only normalised intensity values of genes with highest or lowest Gbm/Mm ratios (figure 4.1A and B). Separation of Gbm and Mm samples is better defined when plotting GFAP expression values rather than when plotting those from DSP. This fact is not surprising since GFAP is specifically expressed in astrocytes and astrocytomas, and it is considered the molecular marker reference of this cellular type in the central nervous system[170, 171], whereas such well defined cell type specific marker does not exist for Mm. Regarding GFAP, variable expression among samples seen in figure 4.1A could be justified by a heterogeneous population of glioblastoma in our study concerning both tumour progression stage and cell stemness characteristics, since expression of GFAP seems to be modulated depending on these factors in Gbm tumours[172, 173].

Nevertheless, a search for a potentially more robust prediction formula by using linear discriminant analysis (LDA) of the most differentially expressed genes was performed. Fully successful prediction of the blindly analysed independent test set was achieved. The result obtained provides evidence of a 100% of sensitivity and specificity by means of a completely objective method, in which intervention of operator bias is strongly reduced. To our knowledge, complete discrimination of two HBT by developing an LDA predictor based on gene signatures arisen from microarray data that could be used by other laboratories had not been previously reported.

4.1.2.2 Molecular signature characteristics of Gbm and Mm

Development of an automated-predictor based on microarray experiments was also useful to study the gene signatures underlying the biology of both Gbm and Mm. For this purpose, an unsupervised hierarchical cluster for each pre-processing procedure using as input those genes within a threshold q-value lower than 0.02 was generated (figure 4.2). This computation corroborated the previous LDA-based prediction because all samples were correctly clustered within its tumour group. In addition, a clear gene signature was achieved for both Gbm and Mm tumour types, which was functionally characterized by analyzing those genes with the DAVID tool. From our results, a high expression of tubulins seems to characterise Gbm, while SLRPs and cytochrome-related genes seems to characterise Mm.

Functional group 2 (see table 4.2) is a paradigm of genes belonging to the SLRPs family and overexpressed in Mms. *FMOD* and *PRELP* genes are located at the q fragment of the 1 chromosome, *OGN* and *OMD* at the q fragment of the 9 chromosome and *BGN* at the q fragment of the X chromosome. However, their promoter region may be similarly regulated[174]. Noticeably, the described involvement of the SLRPs family in collagen fibrillogenesis, cellular growth, differentiation and migration revealed the relevance of this family in extracellular matrix modelling[174]. Specifically, cleavage of OGN precursors by the bone morphogenetic-1 protein (BMP1) producing the mature OGN forms was proposed as a mechanism by which formation of collagen fibrils is controlled[175]. Curiously, we found a set of *BMP* genes among those of q-value lower than 0.02, although it was not detected by the DAVID tool as a functional significant group. Interestingly, the genes coding for receptor *BMP1R*, *BMP4* and *BMP5* were found overexpressed in Mm (see annex A-7). Furthermore, OGN has been recently described as one of the main components of the human amniotic membrane that promotes the development of limbal stem cell niches[176]. Also, interaction of SLRPs members with TGF-β facilitates signal transduction inside the cell, resulting in an increase of SLRPs gene-expression [174, 177, 178]. In our study, we found overexpressed in Mms two genes encoding SLRP proteins involved in the recruitment of TGF-β from the extracellular space to the membrane, FMOD and BGN[177], a receptor of TGF-β, TGFBR2, and a protein modulating the secretion and activation of TGF-β, LTBP2 [178]. From these results, an apparent modulation of the extracellular matrix through SLRPs may characterize tumourigenesis of Mms. Incidentally, *LTBP2* is downregulated when benign meningiomas progress into atypical or anaplastic stages[169].

Functional group 11 contains a set of genes encoding cytochrome proteins overexpressed in meningiomas. Interestingly, there is a consolidated bibliography describing the involvement of cytochrome P450 in cancer drug metabolism [179], and their crucial role in sterol and androgen synthesis, as well as in retinoic acid metabolism [179]. Nevertheless, the cytochrome P450 isoforms found overexpressed in our Mm samples, had not apparently been associated in the biology of this tumour, although other isoforms implication had been linked to both meningioma and glioma progression [180, 181, 182].

The functional group 6 contains a group of tubulins overexpressed in Gbms. Tubulins are structural components of microtubules, which take part in cell motility and intracellular transport, and whose overexpression seems needed in malignant progression of gliomas. Nitration of tubulins is more acute in grade IV than in grade I gliomas[183]. Specifically, gene expression of TUBA3 is induced by PI3K in human glioblastoma cells under stimulation with KCl, a well known differentiation inducer[184].

The functional group with highest statistical significance (group 1) harbours a set of cadherins, among which one half are overexpressed in Mms and the other half overexpressed in Gbms. Such a result may suggest a cadherin sub-type link with the tumour grade and/or histological type. As expected, E-cadherin was found overexpressed in Mms, in agreement with previous findings of E-cadherin detection by histochemistry in meningiomas[185, 186]. Furthermore, E-cadherin is normally not expressed in gliomas, which, instead, express the neural isoform N-cadherin[187, 188], in agreement with our results (see table 4.2). Therefore, we show here a specific expression of cadherins by histological tumour type, rather than an agressivity-linked expression. This may be first exemplified by the differential expression of E- and N-cadherin, and secondly, by those isoforms specifically overexpressed in each tumour type. Likewise, functional group 5 with a set of genes related to cell adhesion would also sustain the important and tumour specific role of the extracellular matrix in cancer.

A group of metallothioneins was significantly detected by the DAVID tool when setting the medium stringency level (data not shown). Metallothioneins are involved in cell detoxification, growth and redox balance, among other cellular roles[189], and were previously found overexpressed in Gbm compared to Mm by immunohistochemistry [190]. Likewise, genes related to glycolytic metabolism were also overexpressed in Gbm (see figure 4.2 and annex A-7), in agreement with references[191, 192], and also in agreement with the well known correlation be-

tween glycolytic phenotype and malignity[193, 194].

In summary, we propose herewith a signature for Mms composed of SLRPs and cytochrome-related genes, which had not been previously described. With regard to Gbms, we confirm the important role of tubulins in malignant progression of this tumour type. Finally, we corroborate the specific expression of E- and N-cadherin in Mm and Gbm, respectively. This is due to the different embryonic origin that characterize meningeal and glial cells. Therefore, a specifc cadherin signature for Gbm and Mm may be characterised by those isoforms overexpressed in each tumour type. In this sense, we may propose a signature for human brain tumour benignity and malignity based on the expression level of SLRPS, cytochrome-related, tubulins and cadherin genes.

4.2 Exploratory analysis of meningioma and glial tumours using Affymetrix data

4.2.1 Results and Discussion

4.2.1.1 Collection of biopsies

A total of 255 biopsy samples were accrued during the eTUMOUR, HealthAgents and MEDIVO2 projects, accounting for several types and subtypes of human brain tumours. They were collected mainly at *Hospital Universitari de Bellvitge IDIBELL* (234), but also at *Hospital Universitari Germans Trias i Pujol* (n=9), *Hospital Clínic* (n=8) and *Hospital Sant Joan de Déu* (n=4).

4.2.1.2 Selection of cases

Among the RNA samples isolated from the 255 biopsies collected during the eTUMOUR, HealthAgents and MEDIVO2 projects, 185 samples fulfilled the quality criteria standard agreed in the eTUMOUR project (see annex A-4), for a RNA sample to be accepted for hybridisation. Among those 185 cases, 86 cases were considered to develop prediction models, since when the analysis was performed, they were those cases with either diagnosis available or accomplishing the quality criteria of hybridisation agreed in eTUMOUR (see annex A-3).

The 86 considered cases comprised 32 glioblastomas (including 1 gliosarcoma), 10 anaplastic astrocytomas, 4 anaplastic oligoastrocytomas, 2 anaplastic oligodendrogliomas, 1 anaplastic ependymoma, 7 diffuse astrocytomas, 3 pilocytic astrocytomas, 2 gemistocytic astrocytomas, 6 oligoastrocytomas, 3 oligodendrogliomas, 12 meningiomas (including 8 meningothelial and 2 fibrous variants), 1 transitional meningioma, 2 atypical meningiomas and 1 hypophysis adenoma.

4.2.1.3 Grouping of samples

From the 86 cases considered to develop prediction models, 4 main groups were created comprising the HBT types and subtypes of highest incidence:

- **Glioblastoma (Gb):** 30 glioblastomas and 1 gliosarcoma.
- **Anaplastic glioma (Ag):** 10 anaplastic astrocytomas, 4 anaplastic oligoastrocytomas and 2 anaplastic oligodendrogliomas.

- **Low grade gliomas (Lgg):** 7 diffuse astrocytomas, 2 gemistocytic astrocytomas, 6 oligoastrocytomas and 3 oligodendrogliomas.
- **Meningiomas (Mg):** 11 meningiomas, comprising 8 of meningothelial type and 3 of fibrous type.

The glioblastoma case **et3223** and the meningioma cases **et3011** and **et3196** showed an unexpected expression level of *GFAP*, *PTPRZ1*, *GPM6B* and *PRELP* genes. The **et3223** case showed values of meningioma, while those two meningioma cases showed values of glioblastoma. This unexpected value was corroborated by RT-PCR. For that reason, we decided discard them from supervised analysis, since some mislabeling of RNAs could occurs at some step. Posteriorly, the RNA from these samples was hybridised again and the expected level of these genes was detected. Unfortunately, the microchip data was obtained after the supervised analysis was performed and we could not include them in the supervised analysis.

4.2.1.4 Evaluation of RNA integrity of collected biopsies

From the 255 biopsies collected during the eTUMOUR project, microchip analysis was performed only for 185 biopsies. The RNA from the remaining 70 biopsies did not fulfill the minimum requirement of integrity for a sample to be accepted for hybridisation (see section 3.2.3 and annex A-4).

Considering that the 27.5% of RNA samples obtained could not be subjected to hybridysation, a search for experimental and clinical parameters that could explain such a consistent loss of samples was performed. First, the frequency of tumour types in each case was evaluated (see figure 4.4).

No visually different distribution of HBT types was found between hybridised and non-hybridised samples (see figure 4.4). As the tumour type did not apparently affect RNA integrity, the blood content of the biopsy was approximated from the visual appearance for each sample at the initial homogenization step for RNA extraction (section 3.2.2). More blood content was assumed to produce a brown homogenate, whereas low or no blood content should produce an uncoloured homogenate. Boxplots for hybridised and non-hybridised did not visually reveal any difference between cases (see figure 4.5).

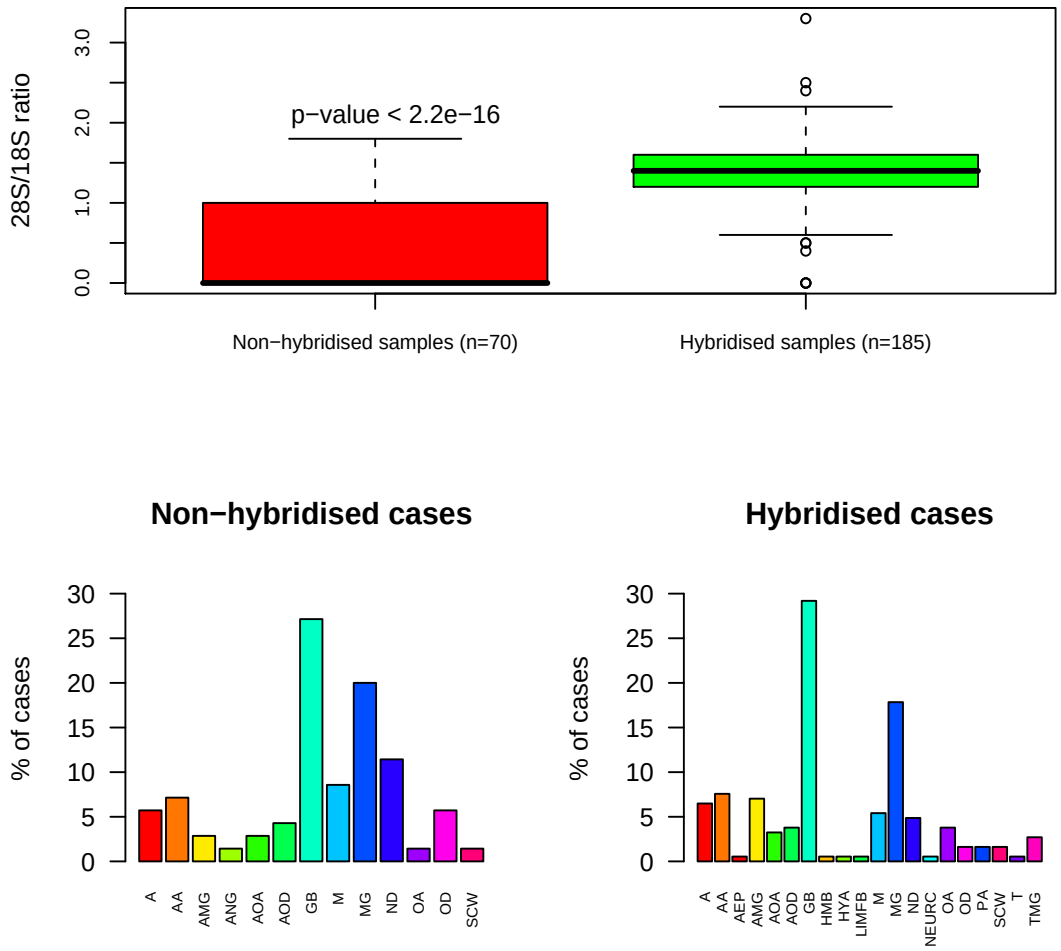


Figure 4.4: *Characterization of RNA integrity on HBT samples.* This figure is an attempt to evaluate a hypothetical dependency on HBT types of the RNA integrity of extracted samples. The ribosomal peaks 28S/18S ratio was first plotted for both hybridised and non-hybridised samples (top figure). The number of samples per tumour type was plotted for each condition (hybridised and non-hybridised samples). The p-value plotted was computed using the rank-based Wilcoxon test. No statistically significant difference between hybridised and non-hybridised samples was found by the paired-rank test of Wilcoxon, when comparing the percentage of diagnoses in each condition. Abbreviations: A (low grade astrocytoma), AA (anaplastic astrocytoma), AEP (anaplastic ependymoma), AMG (atypical meningioma), ANG (angioma), AOA (anaplastic oligoastrocytoma), AOD (anaplastic oligodendroglioma), GB (glioblastoma), HMB (haemangioblastoma), HYA (hypophysis adenoma), LIMFB (B lymphoma), M (metastasis), MG (meningioma), ND (no diagnosis), NEURC (neurocytoma), OA (low grade oligoastrocytoma), OD (low grade oligodendroglioma), PA (pilocytic astrocytoma), SCW (schwanoma), T (teratoma) and TMG (transitional meningioma).

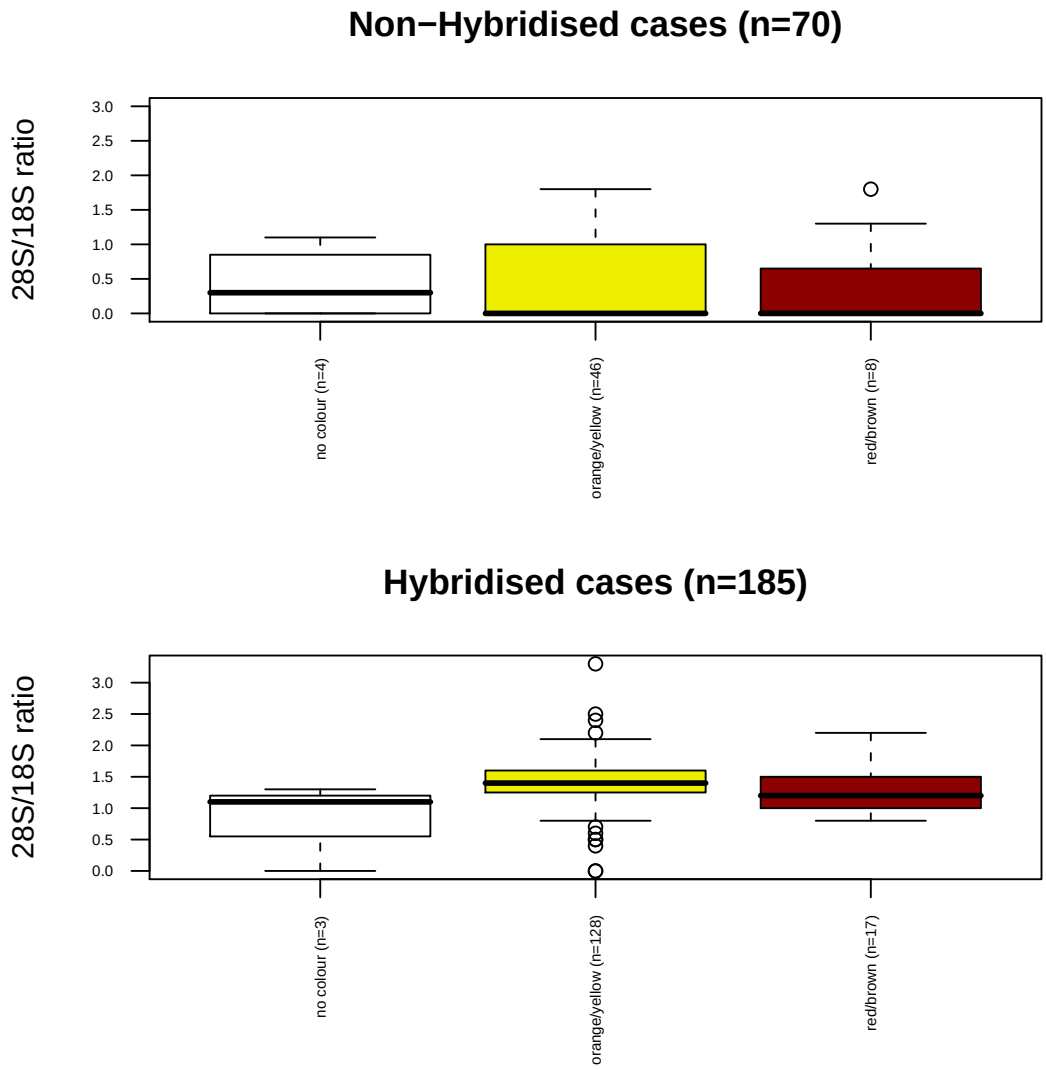


Figure 4.5: *Evaluation of apparent blood content correlation with RNA integrity of biopsies.* To assess the correlation of the blood content on RNA integrity, a boxplot of ribosomal peaks 28S/18S ratio values was represented.

As a further step to investigate the role of blood content of biopsies, we performed the same boxplot than figure 4.6, but for the 3 tumour types that accumulated more than the 50% of total cases: glioblastoma, meningioma and metastasis.

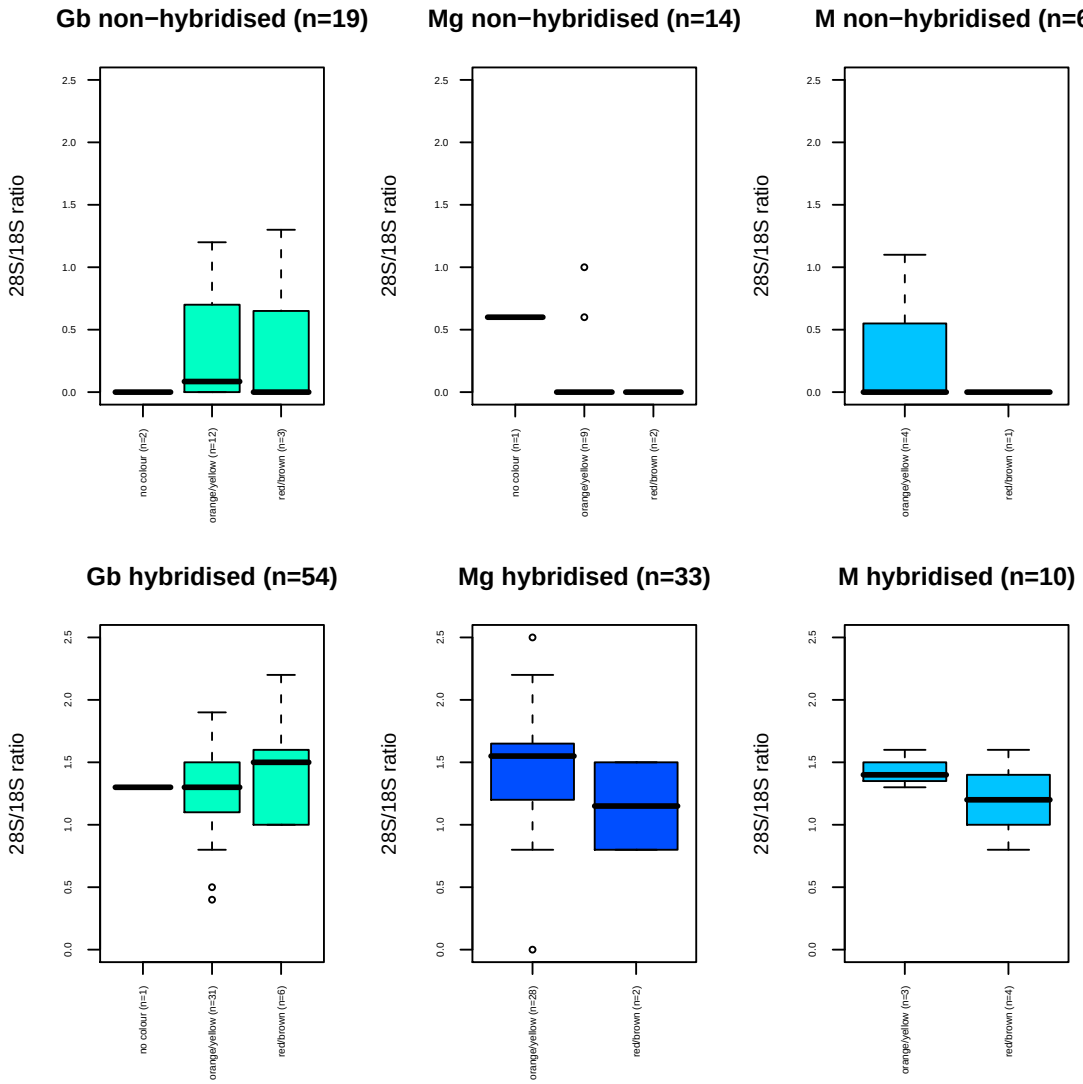


Figure 4.6: *Evaluation of apparent blood content correlation with RNA integrity of biopsies for the 3 most frequent tumour types.* The possible correlation of the apparent blood content of biopsies on RNA integrity was assessed for the 3 most frequent cases of the studied dataset. Along the *y axis* the 28S/18S values are plotted. Single cases are represented by a black horizontal line.

In doing so, we visually verified that the three different diagnosis do not determine a difference on RNA integrity between hybridised and non-hybridised samples, since similar 28S/18S ratio values were found across the 3 tumour types, regardless the degree of biopsy irrigation approximated from apparent blood content(see figure 4.6). Only a metastasis with high blood content from the hybridised cases had an RNA of lower integrity than the rest (see figure 4.6).

Comparison of RNA quality of biopsies collected in RNA*later* and liquid nitrogen

In front of the above described results, the reliability of liquid nitrogen as collection media for biopsies was evaluated. For that, we compared the RNA integrity of 33 biopsies simultaneously collected in both RNA*later* (Ambion, Applied Biosystems, USA) and liquid nitrogen at the surgery room by the Dr. Juan José Acebes and his surgery team, at the *Hospital Universitari de Bellvitge-IDIBELL* (see figure 4.7).

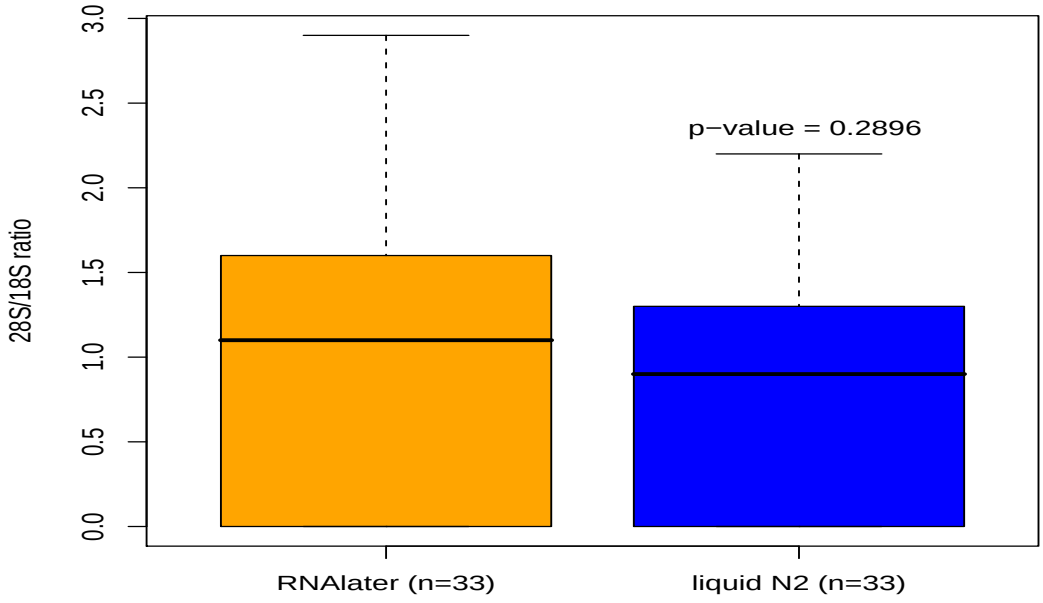


Figure 4.7: *Effect of biopsy collection media on RNA integrity.* Comparison of 28S/18S ratio for 33 biopsies simultaneously collected at the surgery room is depicted in this boxplot. As indicated, no significant difference was detected. A two-sided t-test was used to compute the p-value.

The results obtained indicated that there is no difference in RNA integrity (28S/18S ratio) between biopsy collection in either RNA*later* or liquid nitrogen (p-value=0.2896, 95% confidence interval, two-sided t-test). Therefore, we discarded the collection medium protocol as a possible factor compromising the RNA integrity of accrued samples, as some authors had already reported [133, 132].

Being unable to decipher the reason by which an approximately 30% of RNAs were not valid for hybridisation, a murine model of brain tumour was used for further hypothesis testing, as explained in section 3.6 and discussed in section 4.4.

4.2.1.5 Data pre-processing

Quality of hybridisation

As explained in section 3.5.1.1, a degradation plot was generated to visualize, across microchips, possible differences in labelling between the 5'- and 3'-ends (see 4.8). Development of prediction models based on Affymetrix data was performed on the 86 cases available at the time of the analysis, as described in section 4.2.1.2.

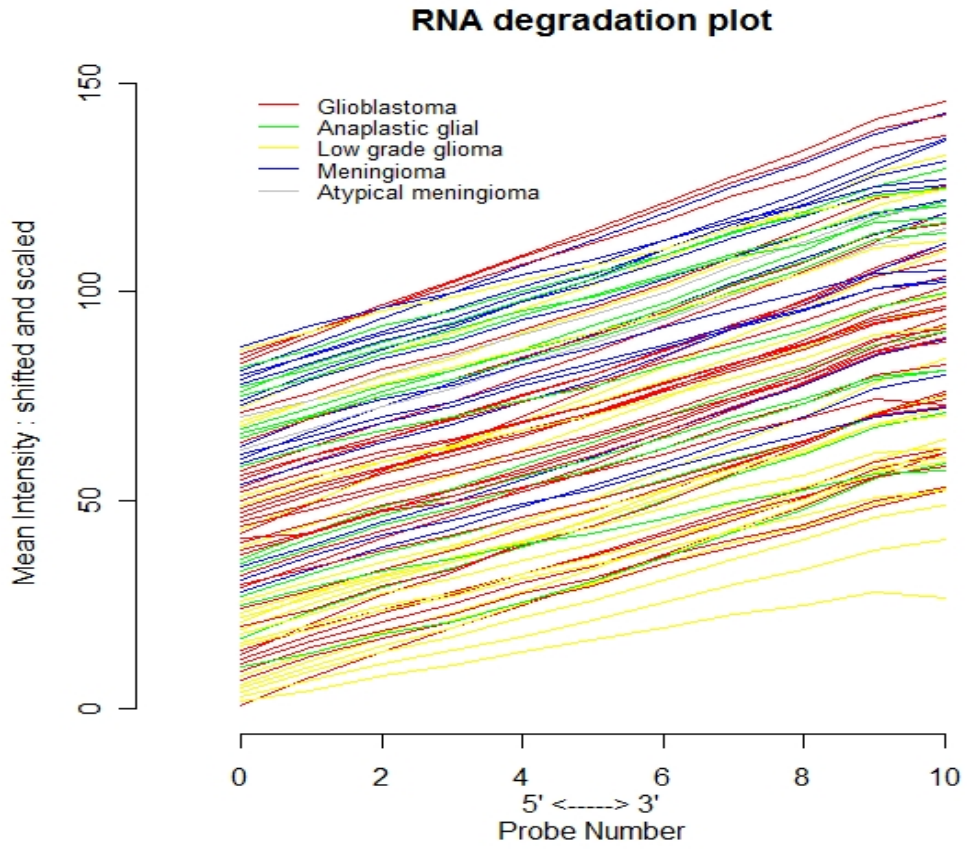


Figure 4.8: *Degradation plot for the 86 Affymetrix microchips performed.* The degradation plot is a measure of fluorescence bias between the 5'- and 3'-ends. Ideally, similar biases must be found for all microchips considered in a project. That is, the intensity slope should not differ between microchips. The x axis indicates the position of the first 10 probes in 5'- to 3'-end direction in each probeset. Along the y axis, averaged fluorescence intensity for each probe across microchips. The legend indicates the main tumour types considered in this work.

Assessment of data pre-processing approaches

The degradation plot provided evidence that similar 3'/5'-ends biases between microchips were obtained. Therefore, no microchip was considered for possible removal from the analysis, and all were included for further testing of three different combinations of background correction and normalisation methods: **RMA**, **MAS5** and **No background correction** (see figure 4.9).

From the generated plots, a slight variability across samples can be visualized from the boxplots and NUSE plots, regardless of the approach used for normalisation and background correction. In contrast, the RLE plot showed an increased variability across samples for the **MAS5** and **No background correction** approaches compared to the **RMA** approach. As RLE plots are a measurement of probesets variability, our results would suggest that a better normalisation was obtained when applying the **RMA** approach for data pre-processing.

Correspondingly, the density plots revealed a higher variability across samples when data was pre-processed using the **MAS5** and **No background correction** approaches compared to the **RMA** approach. In contrast, no striking differences were visualised in the MA plots, although the best adjustment was provided by the **RMA** approach (red line better shaped and median closer to 0).

In front of these results, the derived data from the **RMA** pre-preprocessing approach was selected to undertake further analysis.

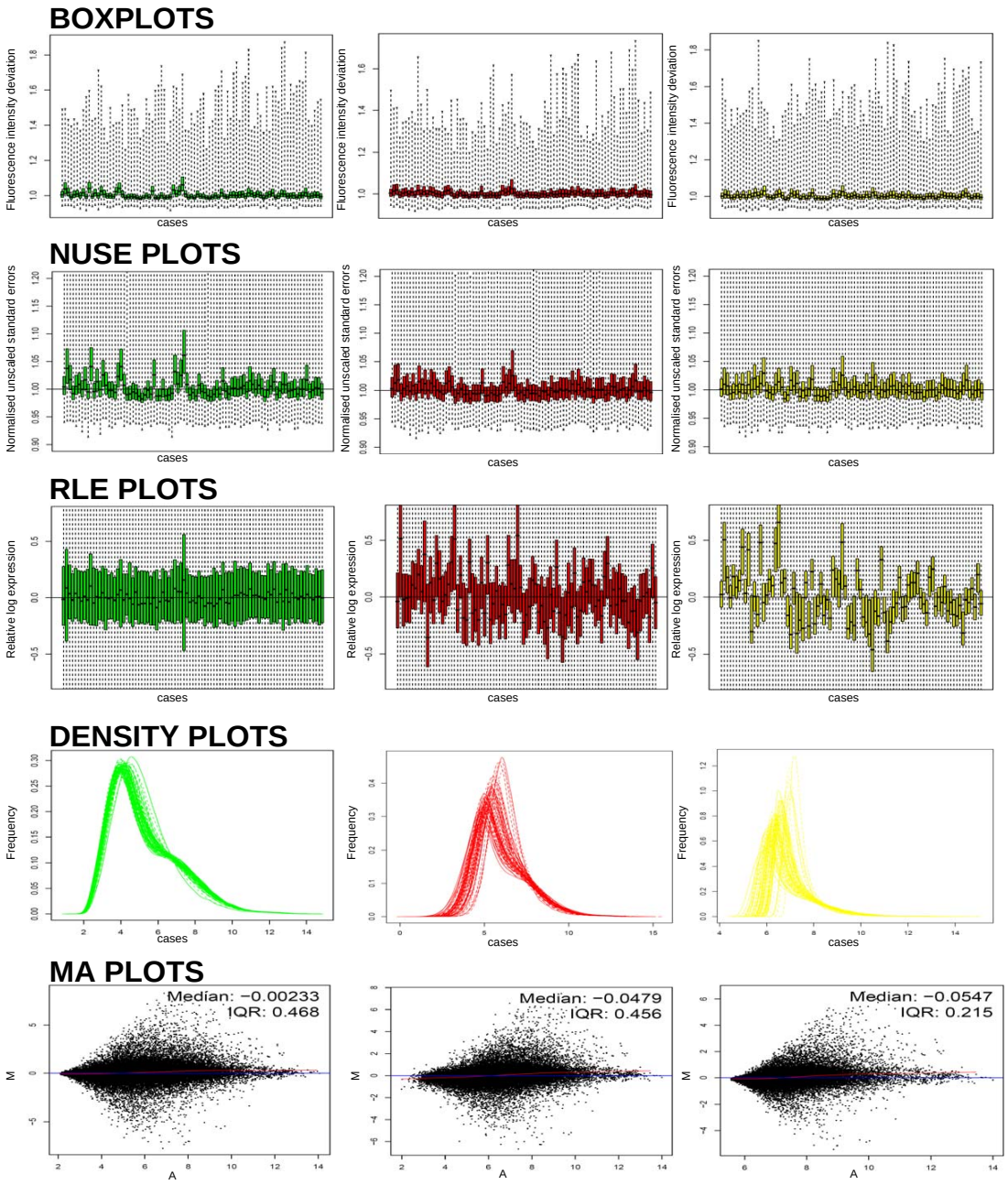


Figure 4.9: *Assessment of data pre-processing approaches.* As a measurement of data-bias produced by the **RMA** (green), **MAS5** (red) and **No background correction** (yellow) pre-processing approaches, five plots were generated. For each of the three approaches a boxplot (top figure), NUSE, RLE, density and MA plots (bottom figure) are depicted. Boxplot is a representation of fluorescence intensity deviation (y axis) for each sample, where the mean of all samples is scaled to 1. The idea underlying the NUSE plot is approximately the same than the boxplot. Nonetheless, in this case the normalised unscaled standard errors (NUSE) for each microchip across genes are plotted, rather than fluorescence intensities. Similarly, the RLE plot represents along the y axis, the relative log expression (RLE) values scaled to 0. Such values are computed for each probeset by comparing the expression value on each array against the median expression value for that probeset across all arrays. The density plot is a histogram of logarithmic-transformed fluorescence intensities for each microchip. Finally, the MA plot along the y axis shows for each probeset, the logarithmic difference between the microchip et3513 and an average of all microchips (M). Along the x axis, the sum is represented (A). At the top right corner, the median and the interquartile range (IQR) are shown.

4.2.1.6 Data prediction

As described in section 4.2.1.3, prediction models were developed for all pairwise comparisons among Mg, Lgg, Ag and Gb tumour groups. Furthermore, two additional three-class prediction models were generated: Mg-Lgg-Gb and Lgg-Ag-Mg. Splitting of samples for training was based on LOOCV and 5FCV, feature selection based on computation of p-values and PCA, and three prediction algorithms tested: linear discriminant analysis (LDA), support vector machines (SVM) and random Forests (randF).

In doing so, several prediction models were generated, which are described in annex A-9. Moreover, tumour class was assigned to each case using two different available diagnosis:

1. Originating pathologist diagnosis (OPD).
2. eTUMOUR consensus diagnosis (ECD).

Diagnosis discrepancies between OPD and ECD provided evidence of the histopathological examination difficulty to classify samples from complex tumour problems. Evaluation of prediction models based on two different diagnoses was intended to characterise the diagnosis that can better predict a classifier developed from DNA microarrays data. Supervised prediction based on molecular profile could clarify the apparent confusion detected between diagnoses. Prediction models using both available diagnoses were generated for three out of six discrepant samples:

1. **et2953** diffuse astrocytoma by the OPD and Glioblastoma by the ECD.
2. **et2034** oligoastroastrocytoma by the OPD and anaplastic oligodendroglioma by the ECD.
3. **et2870** oligoastroastrocytoma by the OPD and anaplastic oligodendroglioma by the ECD.

Three additional samples were uniquely labelled with the OPD diagnosis, since the malignancy grade assigned by the OPD was higher than the ECD:

1. **et2354** glioblastoma by the OPD and diffuse astrocytoma by the ECD.
2. **et3509** glioblastoma by the OPD and diffuse astrocytoma by the ECD.

This decision was taken following the recommendation of the committee of histopathologists that coordinates clinical validation in eTUMOUR. A low-grade glioma assigned by the ECD, but high-grade glioma assigned by the OPD, could indicate that some relevant tissue feature was lacking in the slide received by the eTUMOUR clinical committee, while it was present in the locally reviewed slides.

Only LOOCV models were plotted, since all possible combinations of training sets and test sample were evaluated. That is, each sample of the whole dataset was left apart only once as test sample. Thus, only a discriminant score (DSC) was obtained for each sample. In contrast, the 5FCV models produced more than one discriminant score per sample, which made more complicated to plot final results.

The prediction models (LOOCV and 5FCV) that provided the highest prediction accuracy for each comparison are shown in table 4.5. To visualize the separation between tumour groups, a scatter plot of discriminant values from LOOCV models was performed for the best predictor of each comparison (see figures 4.10 and 4.11).

From the generated models, Mgs can be perfectly predicted for all pairwise comparisons, regardless the diagnosis (OPD or ECD) used to label samples (see table 4.5 and figure 4.10). In contrast, an increased difficulty to predict samples and a highest dependency on diagnosis origin was detected when analysing glial tumour grades (see table 4.5 and figure 4.10).

With the exception of Lgg-Gb model that reached 92% of prediction accuracy when using eTUMOUR consensus diagnosis, models for the rest of pairwise comparisons yielded between 70% and less than 90% of prediction accuracy. Such low prediction accuracy rates may be indicative of varying molecular characteristics across samples that histopathological diagnosis can not detect.

Best prediction models for each comparison

		Mg vs Lgg	Mg vs Ag	Mg vs Gb	Lgg vs Gb	Lgg vs Ag	Ag vs Gb	Mg vs Lgg vs Gb	Lgg vs Ag vs Gb
Originating pathologist diagnosis	*LOOCV*	% accuracy (AUC)	100	100	100	89.6	77.1	79.6	91.7	61.2
		p-value	6.1E-07	1.3E-08	2.3E-06	5.7E-04	4.7E-02	6.7E-03	2.0E-08	8.1E-05
		number of variables	PCA for 60% variability	15 genes	10 genes	2 genes	2 genes	10 genes	PCA for 80% variability	PCA for 90% variability
		Prediction algorithm	LDA	svm	randF	randF	randF	svm	LDA	LDA
	5FCV	% accuracy (AUC)	100	100	100	84	81	77.5	89.6	56.1
		p-value	1.0E-07	1.0E-06	1.0E-08	1.0E-05	4.4E-16	4.3E-13	1.0E-05	1.0E-06
		number of variables	PCA for 50% variability	PCA for 50% variability	PCA for 50% variability	PCA for 50% variability	40 genes	20 genes	PCA for 80% variability	PCA for 70% variability
		Prediction algorithm	LDA	LDA	LDA	LDA	svm	svm	LDA	LDA
eTUMOUR consensus diagnosis	*LOOCV*	% accuracy (AUC)	100	100	100	92	70	81.8	93.5	63.1
		p-value	1.1E-07	2.2E-07	2.4E-06	1.5E-03	2.4E-02	5.1E-06	1.3E-08	1.9E-02
		number of variables	PCA for 50% variability	25 genes	10 genes	5 genes	PCA for 90% variability	5 genes	PCA for 80% variability	PCA for 90% variability
		Prediction algorithm	LDA	LDA	LDA	svm	svm	svm	LDA	LDA
	5FCV	% accuracy (AUC)	100	100	100	86.6	75	80.9	91.5	57.5
		p-value	1.0E-07	1.0E-06	1.0E-07	1.0E-05	4.6E-15	1.0E-06	1.0E-05	1.0E-05
		number of variables	PCA for 50% variability	PCA for 50% variability	PCA for 50% variability	5 genes	40 genes	10 genes	PCA for 80% variability	PCA for 80% variability
		Prediction algorithm	LDA	LDA	LDA	svm	svm	svm	LDA	LDA

Table 4.5: *Best prediction models for each tumour comparison(s).* The prediction accuracy for each model (LOOCV and 5FCV) is shown for each tumour comparison(s). At the upper part of the table, results from prediction models labelled with originating pathologist diagnosis are described. At the bottom part, the same but labelled with the eTUMOUR consensus diagnosis.

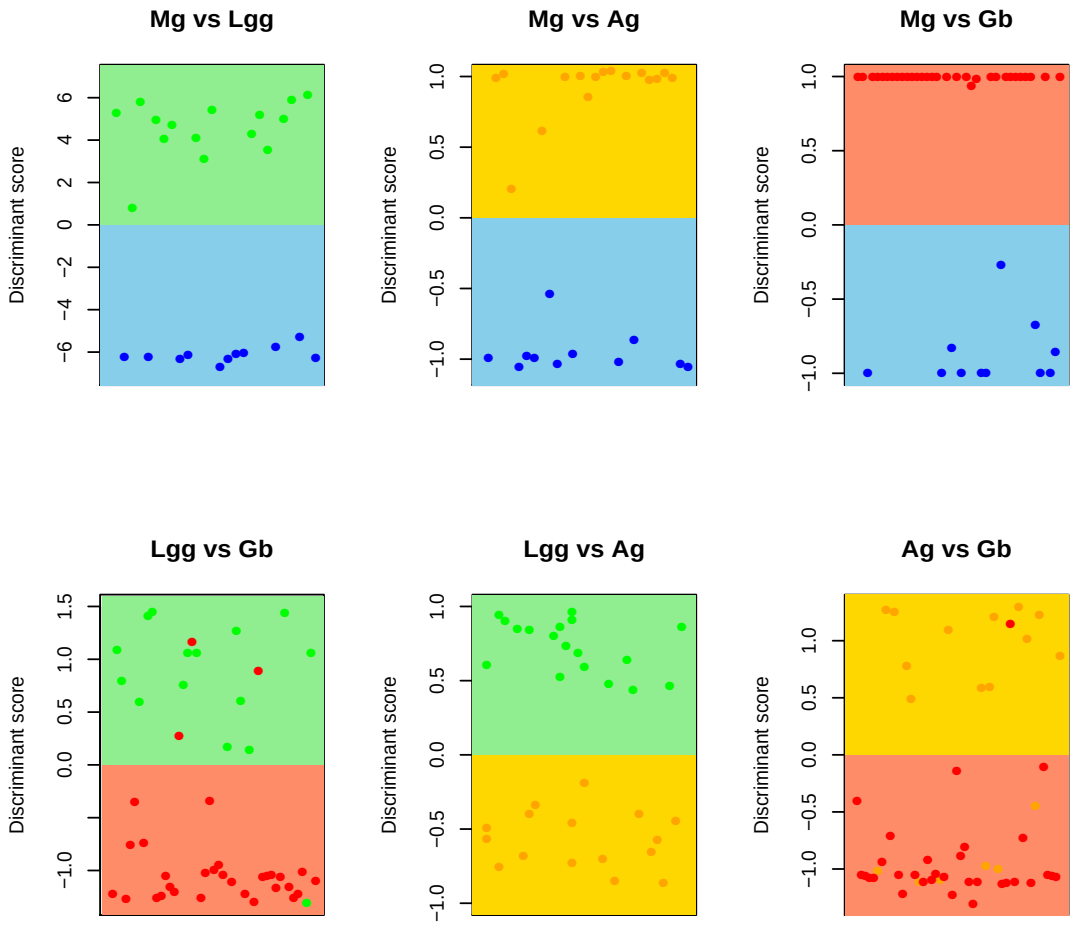


Figure 4.10: *Discriminant plots for pairwise comparisons.* For each performed tumour comparison, the discriminant score of test samples is plotted. That is, at each LOO iteration only one sample is left apart to test the prediction ability of the prediction formula, which is obtained from training samples. When all iterations are performed, a discriminant score for each sample is obtained. An exception was performed for the Lgg-Ag comparison, since the discriminant scores of training and test samples were plotted for one of the 14 models yielding 100% prediction accuracy. Only the best prediction model is plotted, regardless the diagnosis used to label samples. Blue symbols are meningioma, green low-grade glioma, orange anaplastic glioma and red glioblastoma samples. Colour zones depict borders between classes defined by the classifier.

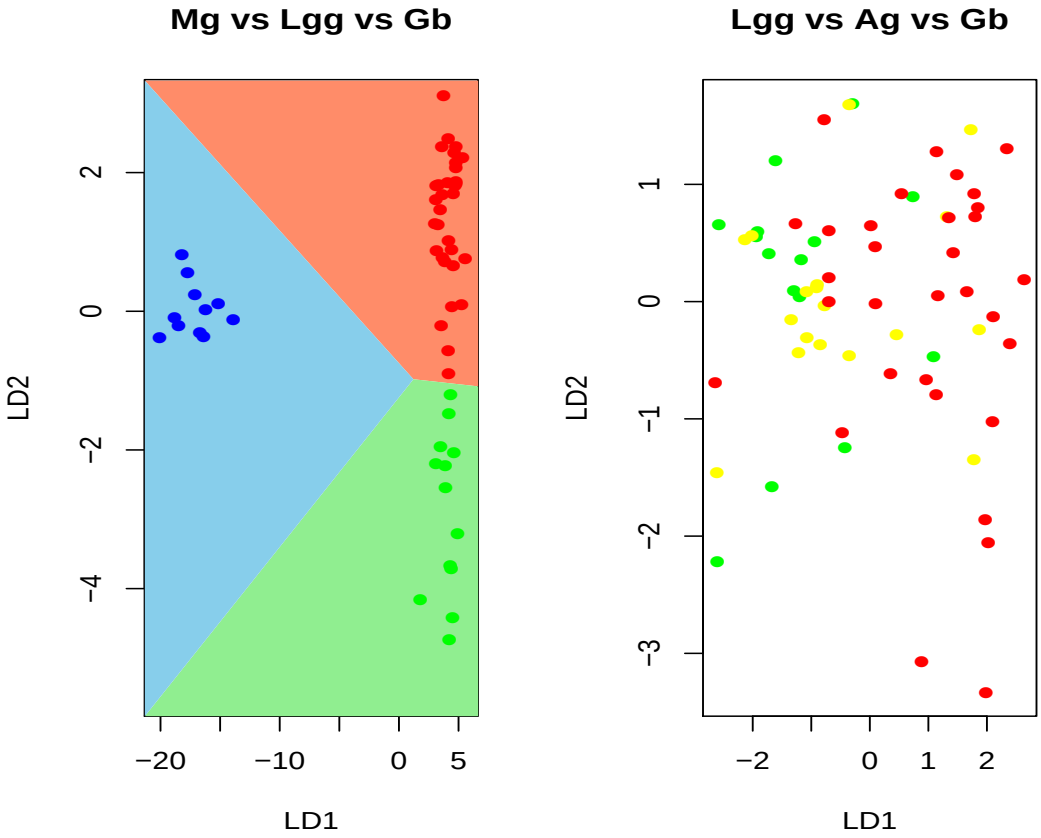


Figure 4.11: *Discriminant plots for three class comparisons.* For each performed tumour comparison, the discriminant score of test samples is plotted in Lgg-Ag-Gb comparison. Similarly to Lgg-Ag in figure 4.10, an exception was performed for the Mg-Lgg-Gb comparison, since the discriminant scores of training and test samples were plotted for one of the 4 models yielding 100% prediction accuracy. Blue symbols are meningioma, green low-grade glioma, orange anaplastic glioma and red glioblastoma samples. Colour zones depict borders between classes defined by the classifier. No colour background was displayed for Lgg-Ag-Gb comparison, since no clear boundary limits could be defined among classes.

4.2.1.7 Molecular signature characteristics of meningeal and glial tumours

Those unique genes that fitted the best prediction models described in section 4.2.1.6, were used to investigate molecular signatures characterising meningeal and glial tumours. Those probesets with a fold-change equal or higher than 2 in any of the six possible comparisons (Mg-Lgg, Mg-Ag, Mg-Gb, Lgg-Gb, Lgg-Ag and Ag-Gb) were selected.

In doing so, 2097 probesets with p-value equal or lower than 0.05 were selected. To remark, p-value was used for selection, although it does not fulfill multiple test conditions. The use of the p-value was due to the few probesets with a q-value equal or lower than 0.05 for comparisons between glial tumours (Lgg-Gb, Lgg-Ag and Ag-Gb). Evidently, a p-value equal or lower than 0.05 always corresponded to a q-value equal or lower than 0.05 for a large number of probesets when comparing Mg to either Lgg, Ag or Gb (see annex A-10).

These probesets were submitted to the DAVID tool, from which resulted 25 statistically significant functional groups, which accounts for 282 genes and 379 probesets (see table 4.6 and annex A-10).

Comparison of Mg and Gb molecular signatures obtained from Affymetrix and cDNA microarrays data

Among the 11 functional groups identified in the comparison of Gbm and Mm using cDNA microarrays data (see section 4.1.2.2), the group containing several cadherin isoforms and the group with various members of the SLRPs family were also found among the 25 functional groups derived from Affymetrix data (see functional groups 4 and 6 in table 4.6 and annex A-10).

Interestingly, a higher expression of SLRPS genes in Mgs compared to glial tumours, and expression of specific cadherins for Mgs and glial tumours was detected again. In contrast, none of these functional groups displayed a expression difference with statistically significance, when comparing glial tumours each other. This result would indicate that differences in histopathological types of glial tumours do reflect in differential expression of genes from these functional groups.

Similarly, those genes encoding membrane proteins or related with the extracellular space, genes encoding proteins involved in transduction of external signals or cell signalling were also found (see functional groups 1-3, 7-11, 13-15, 17-21 and 23-25 in table 4.6 and annex A-10).

Functional group	Representative genes of the group
1	Leucine-Rich Repeat Containing 8 Family genes (*LRCC8* isoforms), Major Facilitator Superfamily Domain 4 gene (*MFSD4*) and Transmembrane Protein 30B gene (*TMEM30B*)
2	Glypican 6 (*GPC6*) and neurotrimin genes (*HNT*)
3	Leucine-Rich Repeat genes (*LGR5*, *LINGO2*, *LRFN5*, *LRIG1*, *LRIG3*, *LRRC3B*, *LRRN1*, *LRRN3*, *LRRTM2*)
4	Fibromodulin (*FMOD*), lumican (*LUM*), osteoglycin (*OGN*), osteomodulin (*OMD*) and Proline/Arginine-Rih End Leucine-Rich Repeat protein (*PRELP*)
5	Familly with sequence similarity 77 gene (FAM77D) and synaptogyrin 3 (*SYNGR3*)
6	Cadherin (*CDH1*, *CDH10*, *CDH18* and *CDH19*) and protocadherin genes (*PCDH7*, *PCDH8*, *PCDH9* and *PCDH17*, among others)
7	Axon guidance receptor-related genes (*ROBO1* and *ROBO2*), cell adhesion-related genes (*NCAM*, *NCAM1* and *VCAM*), contactin 1 (*CNTN1*) and neuronal growth regulator 1 (*NEGR1*)
8	Sparc-related genes (*SMOC1*, *SMOC2*, *SPOCK1* and *SPOCK3*)
9	Genes encoding various types of receptors (*CNR1*,*GABBR2*, *GPR17*, *GPR22*, *GPR27*, *LGR5*, *PTGDR* and *P2RY12*)
10	Genes encoding various isoforms of collagen (*COL1A1*, *COL1A2*, *COL3A1* and *COL6A2*, among others)
11	Solute carrier gene (*SLC44A5*), Phosphatidic acid phophatse gene (*PPAPDC1A*), T-cell lymphoma-related gene (TCBA1)
12	Genes encoding various isoforms of immunoglobulins (*IGHA1*, *IGHG2*, *IGHG3*, *IGLA2*, *IGLJ3* and *IGHKC*)
13	Genes encoding various isoforms of γ-aminobutyric acid, glutamate and glycine (*GABRA1*, *GABRA2*, *GABRA4*, *GABRA5*, *GRIK2*, *GRIA2*, *GRIN2A* and *GLRB*)
14	Membrane-related genes (*MS4A7*, *MS4A4*, *MS4A46A*, *SLC44A5*, *TMEM47* and *TM6SF1*, among others)
15	Bestrophin 3 (*BEST3*) and membrane related-genes (*FXYD7* and *PLLP*)
16	Genes related to synapsis (*SV2A*, *SV2A*, *SV2C*, *SVOP* and *SYNPR*)
17	Genes encoding ligands of chemokines (*CXCL2*, *CXCL3*, *CXCL5* and *CXCL6*, among others)
18	Genes encoding various transporters of ions (*KCNK2*, *KCNN2*, *SCN1A* and *SCN2A*, among others)
19	Genes encoding metallopeptidases (*ADAMTS6* and *ADAMTS9*, among others) and transporters of ions (*KCNE4* and *SCN3B*, among others)
20	Genes encoding various transporters of ions (*KCNE4*, *SLC4A4*, *SLC12A5*, *SLC24A3* and *SLC10A4*)
21	Genes encoding various transporters of clhoride (*CLCNKB*, *CLIC3*, *CLIC5* and *CLIC6*)
22	Genes involved in the metabolism of galactose (*B3GALT6*, *GALNTL2*, *GAL3ST2* and *ST6GAL2*, among others)
23	Genes related to the metabolism of ATP (*ABCA8*, *ABCA13*, *ATP8A2* and *ATP13A4*, among others)
24	Genes encoding various regulators of G-protein signalling (*RGS1*, *RGS4*, *RGS5* and *RGS7*, among others)
25	Genes with capacity of binding calcium (*CABP1*, *CALN1*, *TNNC1* and *VSNL1*)

Table 4.6: *Description of functional genes detected from Affymetrix data.* This table describes the 25 functional groups detected from Affymetrix data, which accounts for 379 probesets. These functional groups were obtained by submitting to the DAVID tool those 2097 probesets with p-value equal or lower than 0.05 across all pairwise comparisons between Mg, Lgg, Ag and Gb (see text for more details). The complete description of functional groups can be seen in annex A-10.

Curiously, the 1-3, 7-11, 14, 18, 19 and 24 functional groups were composed of several probesets detected differentially expressed (q-value<0.05) when comparing Lgg and Ag cases. Neither the Ag-Gb, nor the Lgg-Gb comparisons displayed differentially expressed probesets in these functional groups.

In contrast, the group containing genes encoding several isoforms of cytochromes and the group harboring several isoforms of tubulins were not identified in Affyme-

trix data. This may indicate a decreased role of these genes in Mg and Gb biology, or maybe the strategy used to select genes in the prediction models discarded these genes. Interestingly, group 10 encoding collagen proteins and group 17 encoded various ligands of chemokines, which were found overexpressed in meningiomas and glial tumours, respectively.

All this taken into account, SLRPs, collagen, cytochrome-related and specific cadherin genes overexpressed in meningiomas may provide a specific signature of benignity for HBTs, which to our knowledge had not been previously described. Similarly, a signature of malignity for HBTs may be provided by ligands of chemokine, tubulin and specific cadherin genes overexpressed in glial tumours.

On the other hand, few functional groups with genes differentially expressed between glial tumours (Lgg, Ag and Gb) were detected (see annex A-10). Differential expression with statistical significance was found between Lgg and Ag tumours. This may suggest that progression from malignacy grade II to grade IV is not a progressive process. That is, if the development of glioma were a progressive process, differential expression between Lgg and Gb groups should have been detected. As it is not, there may be some Lgg tumours that would progress directly to Gb tumours, while other Lgg tumours would progress to grade IV through grade III of malignancy. This could explain the high misclassification of Lggs as Ags, but not the opposite (see table 4.7). This hypothesis somewhat agrees with the accepted existence of primary and secondary glioblastomas[26, 27]. Also, it agrees with the three proposed molecular types of malignant gliomas (proneural, proliferative and mesenchymal), each one of them including tumours of grade II, III and IV of malignancy[1].

Comparison	Misprediction	Cases mispredicted when labelled with OPD	Cases mispredicted when labelled with ECD
Mg vs Lgg		-	-
Mg vs Ag		-	-
Mg vs Gb		-	-
Lgg vs Gb	Lgg as Gb	et2953 and et3217	et3217
	Gb as Lgg	et2354, et3202 et3207 and et3509	et2354, et3207 and et3509
Lgg vs Ag	Lgg as Ag	et2041, et2915, et3208, et3217, et3247 and et3251	et2030, et2035, et2435, et2915, et3208, et3215, et3217, et3247 and et3251
	Ag as Lgg	et2952 and et3245	-
Ag vs Gb	Ag as Gb	et2262, et2425, et3201, et3225, et3246 and et3254	et2262, et2425, et3201, et3225, et3246 and et3254
	Gb as Ag	et2354	et2354
Mg vs Lgg vs Gb	Mg as other	-	-
	Lgg as Gb	et2953	et2435 and et3217
	Gb as Lgg	et2354, et2951, et3197, et3203, et3218 and et3509	et2354 and et2951
Lgg vs Ag vs Gb	Lgg as within brackets	et2870(Gb), et2915(Ag), et2953(Gb), et3217(Gb) and et3247(Ag)	et2035(Ag), et2435(Ag) and et2915(Ag)
	Ag as within brackets	et2262(Gb), et2425(Gb), et2952(Lgg), et3008(Lgg), et3225(Lgg), et3245(Lgg), et3248(Lgg) and et3256(Lgg)	et2034(Lgg), et2870(Gb), et2262(Gb), et2425(Gb), et2952(Lgg), et3201(Lgg), et3245(Gb), et3248(Lgg) and et3254(Gb)
	Gb as within brackets	et2353(Ag), et2354(Lgg), et2357(Lgg), et2951(Ag), et3202(Ag), et3203(Ag), et3205(Lgg), et3212(Lgg), et3218(Ag), et3243(Ag), et3250(Ag) and et3507(Ag)	et2353(Ag), et2354(Lgg), et2357(Lgg), et2951(Ag), et3202(Ag), et3203(Ag), et3205(Lgg), et3207(Lgg), et3212(Lgg), et3218(Ag), et3243(Ag), et3250(Ag) and et3507(Ag)

Table 4.7: *Mispredicted samples across all comparisons.* Samples that were mispredicted in each comparison for both originating pathologist and eTUMOUR consensus diagnosis are shown.

4.3 Exploratory analysis of molecular subtypes of glioblastoma

4.3.1 Determination of clusters of glioblastoma

After assigning the most probable tumour class for the discrepant diagnosis cases (see section 4.2.1.6), a total of 47 Gb cases were finally used to explore the existence of molecular subtypes. This amount included 17 new cases acquired at the UAB after the analysis at the section 4.2 had been performed.

A key step for discovering molecular subtypes of Gbs relies on proper gene selection. For that reason, 4 different approaches were performed to select genes:

1. The 555 probesets included within the 629 genes differentially expressed between Gbm and Mm using cDNA microarrays.

2. The 20 probesets with highest and lowest ratio between Gb and the rest of tumour groups (Mg, Lgg and Ag). This resulted in a total of 120 probesets.

3. The 27 matching genes proposed by **Phillips** and collaborators[1].

4. The 10% of probesets with highest coefficient of variation among the 47 considered Gb cases.

A hierarchical cluster based on the euclidean distance was performed for each approach (see figures 4.12, 4.14, 4.16 and 4.18).

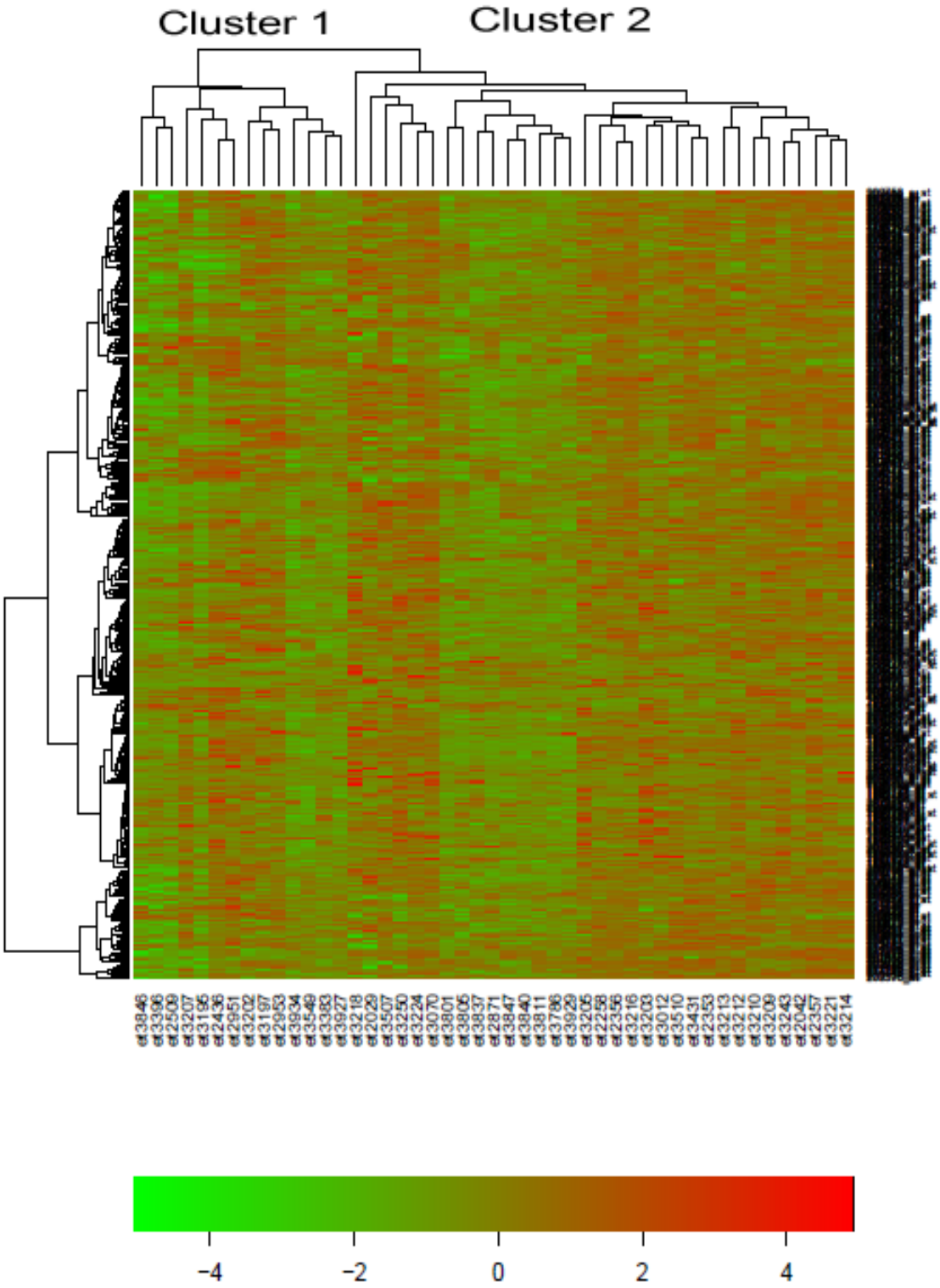


Figure 4.12: *Hierarchical cluster of 47 Gbs based on the 555 probesets differentially expressed between Gbm and Mm using cDNA microarrays.* This figure displays the molecular profile of the 555 probesets matching the 629 genes differentially expressed between Gbm and Mm using cDNA microarrays. Columns are Gb cases and rows probesets. The bottom bar indicates the intensity (arbitrary scale) of probesets per each sample.

Concerning the first approach, two main clusters grouping Gb cases can be seen

from the column dendogram (see figure 4.12). However, such a defined column dendogram does not seems to clearly define molecular-profiles. To assess the relevance of clusters, the *silhouette* statistics (see figure 4.13) was computed as described in section 3.5.3.

The closer to 1 the statistics is, the higher the cluster reliability. Considering that the maximum statistics mean is 0.15, the generated cluster seems to be unreliable. This is not surprising, since the genes used are optimal only to distinguish between Gbm and Mm.

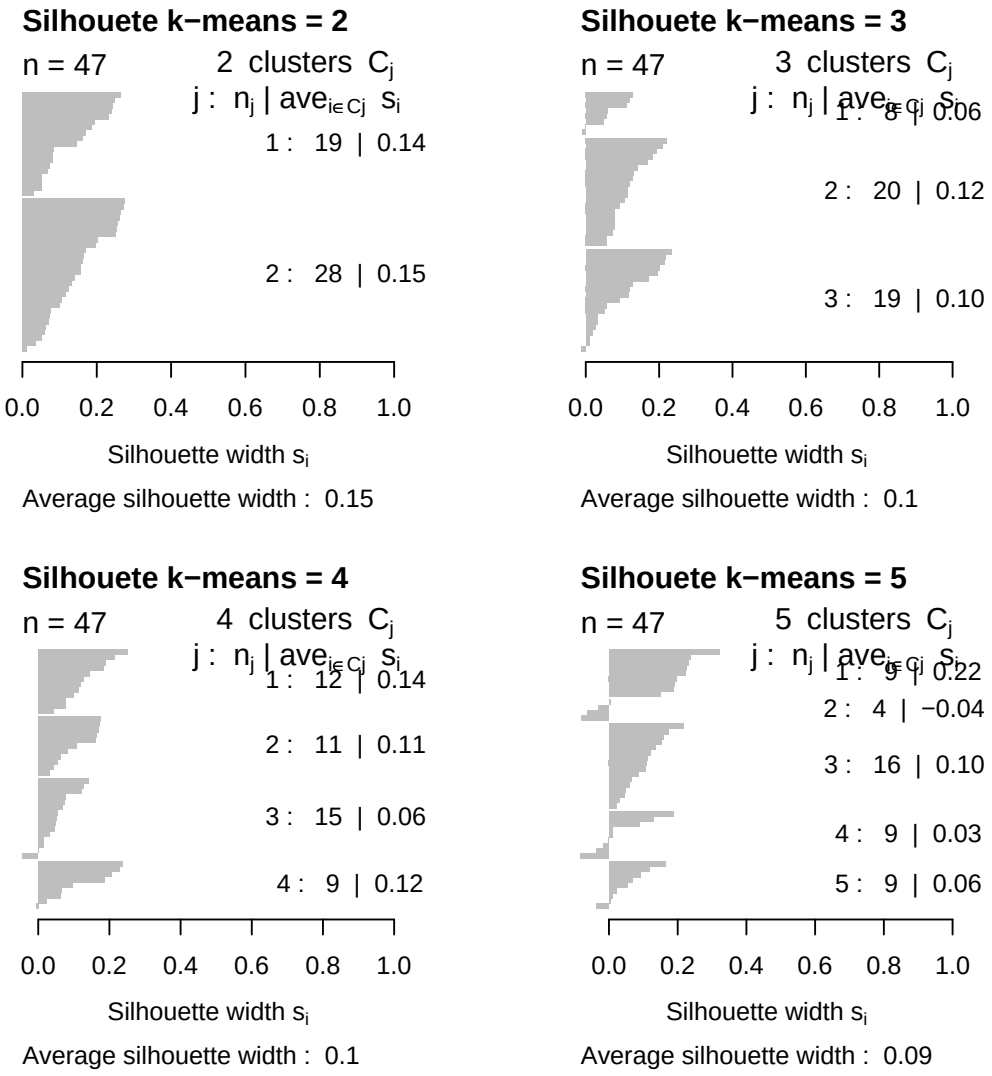


Figure 4.13: *Silhouette plot to determine the number of clusters from figure 4.12.* Four number of clusters (2, 3, 4 and 5) were assessed by the *silhouette* statistics.

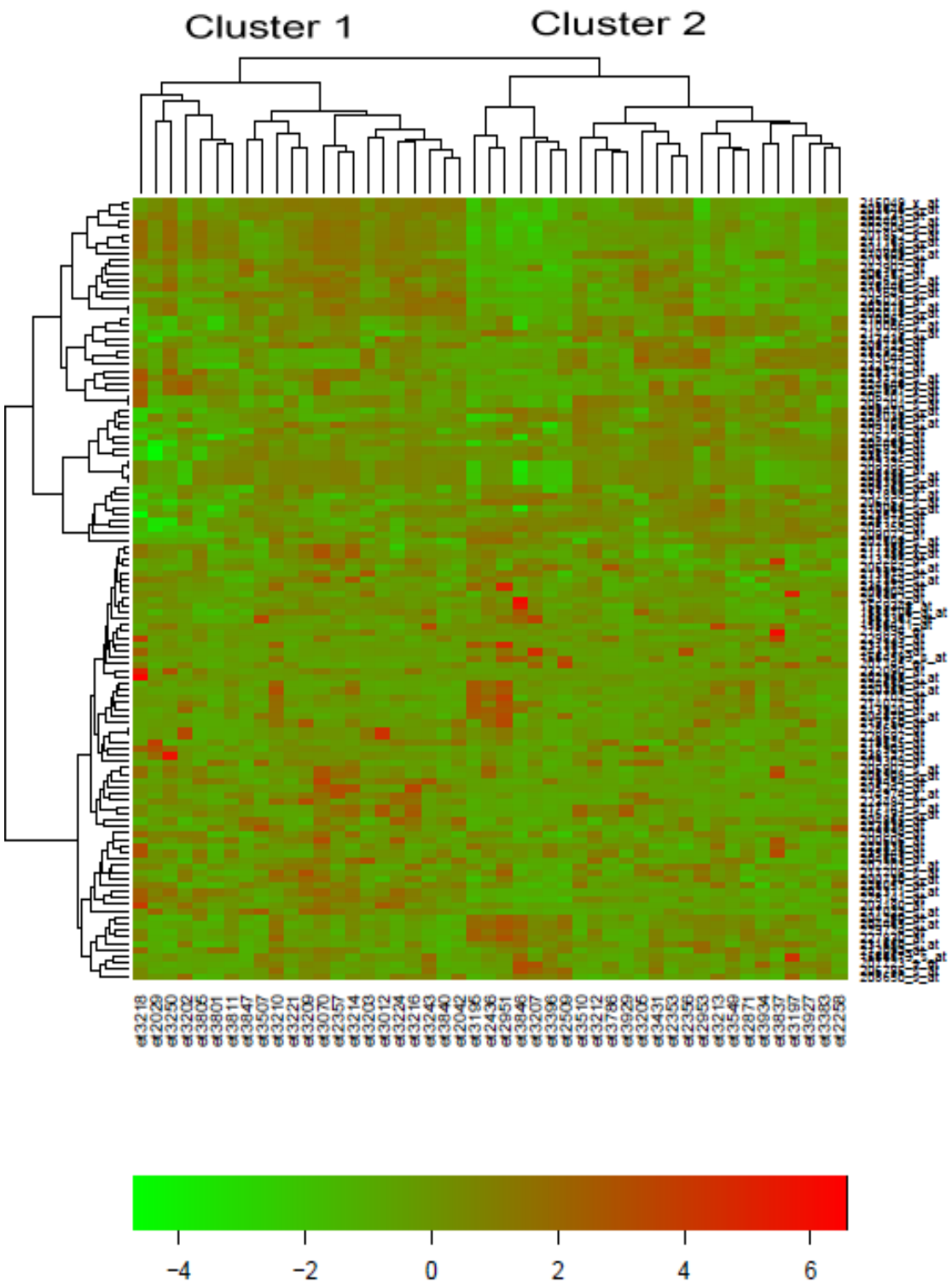


Figure 4.14: *Hierarchical cluster of 47 Gbs based on the 120 probesets differentially expressed between Gb and the rest of tumour groups.* This figure display the molecular profile of the 120 probesets differentially expressed between Gb and the rest of tumour groups(Mg, Lgg and Ag). Columns are Gb cases and rows probesets. The bottom bar indicates the intensity (arbitrary scale) of probesets per each sample.

Concerning the second approach to select genes, also two main clusters can be seen from the column dendogram (see figure 4.14). In this case, a more defined molecular-profile compared to figure 4.12 can be seen for each cluster. Furthermore, the maximal *silhouette* statistics is 0.2 (see figure 4.15), which even not being very reliable, it is higher than the one obtained in the previous approach.

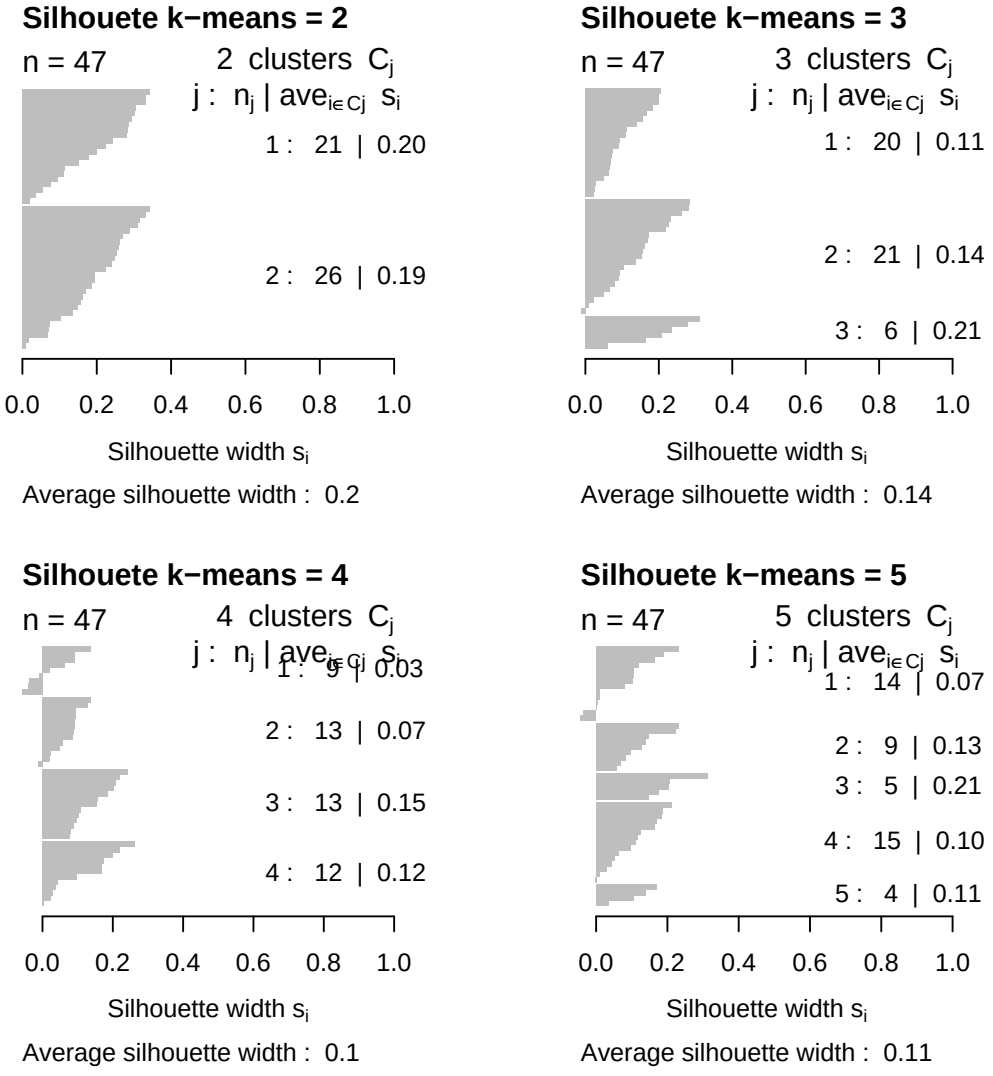


Figure 4.15: *Silhouette plot to determine the number of clusters from figure 4.14.* Four cluster combinations (2, 3, 4 and 5) were predicted by the *silhouette* statistics.

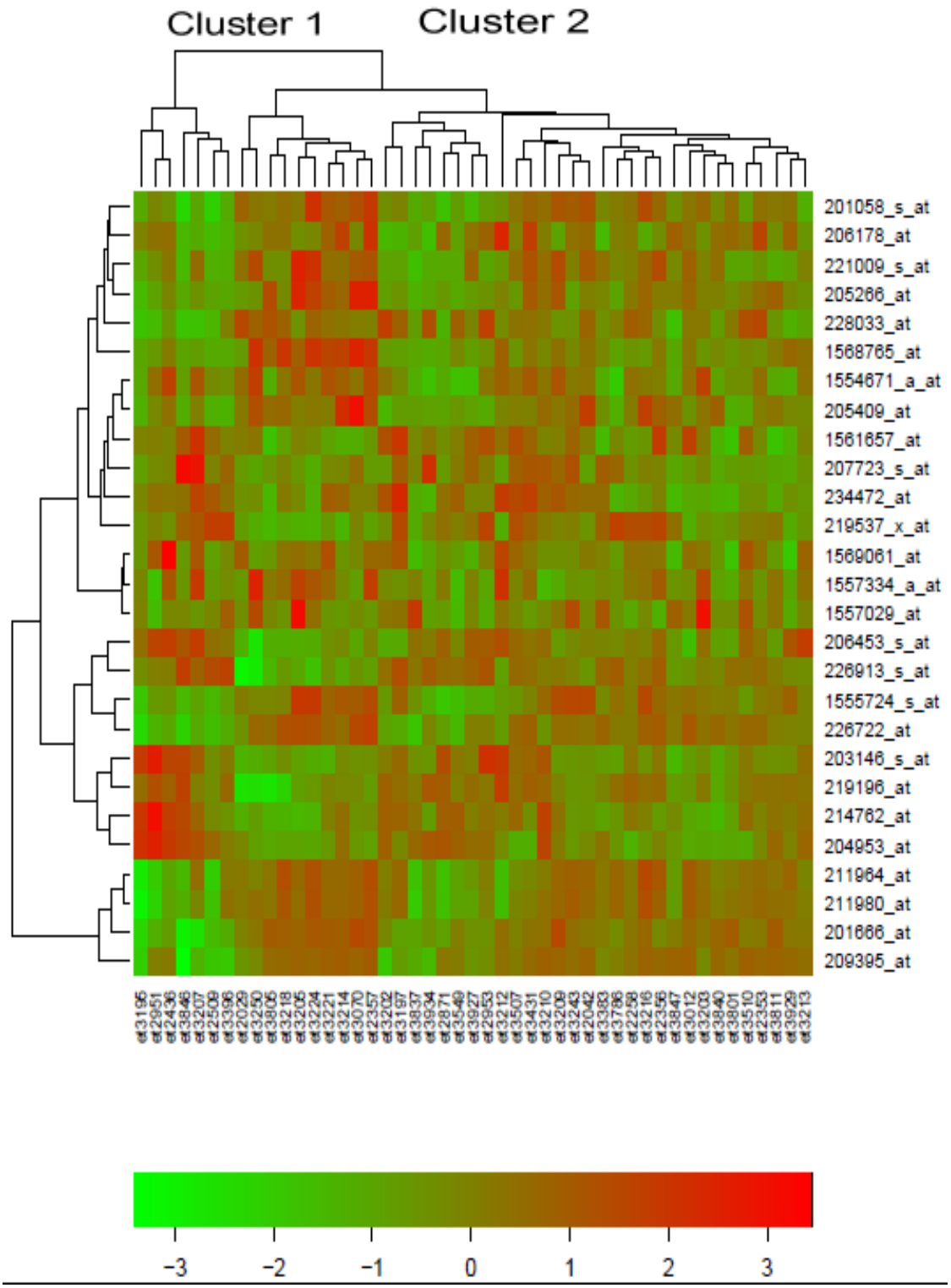


Figure 4.16: *Hierarchical cluster of 47 Gbs based on the 27 genes from* ***Phillips et al***[1]. This figure displays the molecular profile of 47 Gbs based on the 27 genes from reference **Phillips et al**[1]. Columns are Gb cases and rows probesets. The bottom bar indicates the intensity (arbitrary scale) of probesets per each sample.

With regard to the third approach, two clusters seems to be delimited by the columns dendogram (see figure 4.16). A less balanced size of clusters was found. In addition, no molecular-profile can be clearly detected as in figure 4.12. The *silhouette* statistics also demonstrate that no reliable clusters can be detected (see figure 4.17).

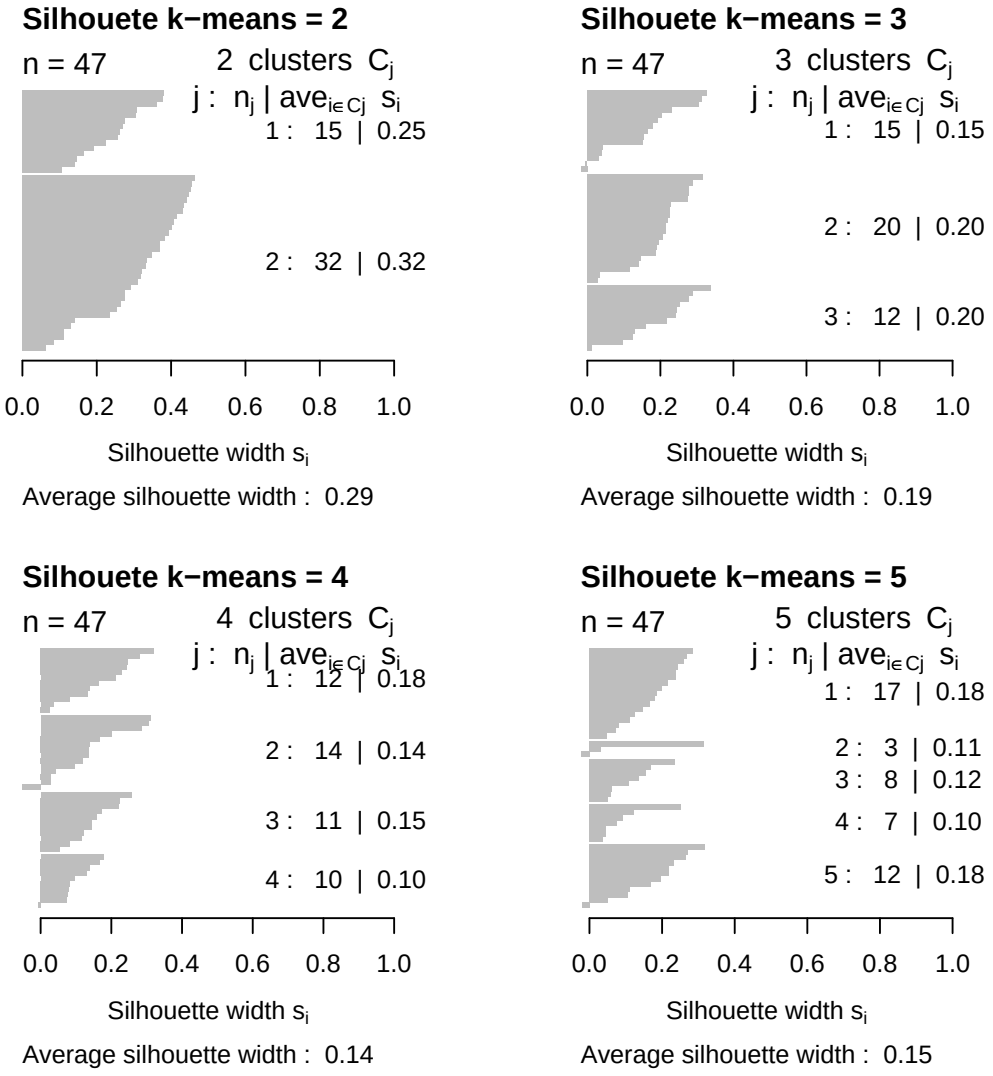


Figure 4.17: *Silhouette plot to determine the number of clusters from figure 4.16.* Four cluster combinations (2, 3, 4 and 5) were predicted by the *silhouette* statistics.

Phillips and collaborators proposed the used genes in this approach as a molecular-signature to identify the resemblance of glial tumours to their precursor cells. Therefore, the obtained result may be interpreted such as that the considered Gb cases would have dedifferentiated into a similar development stage.

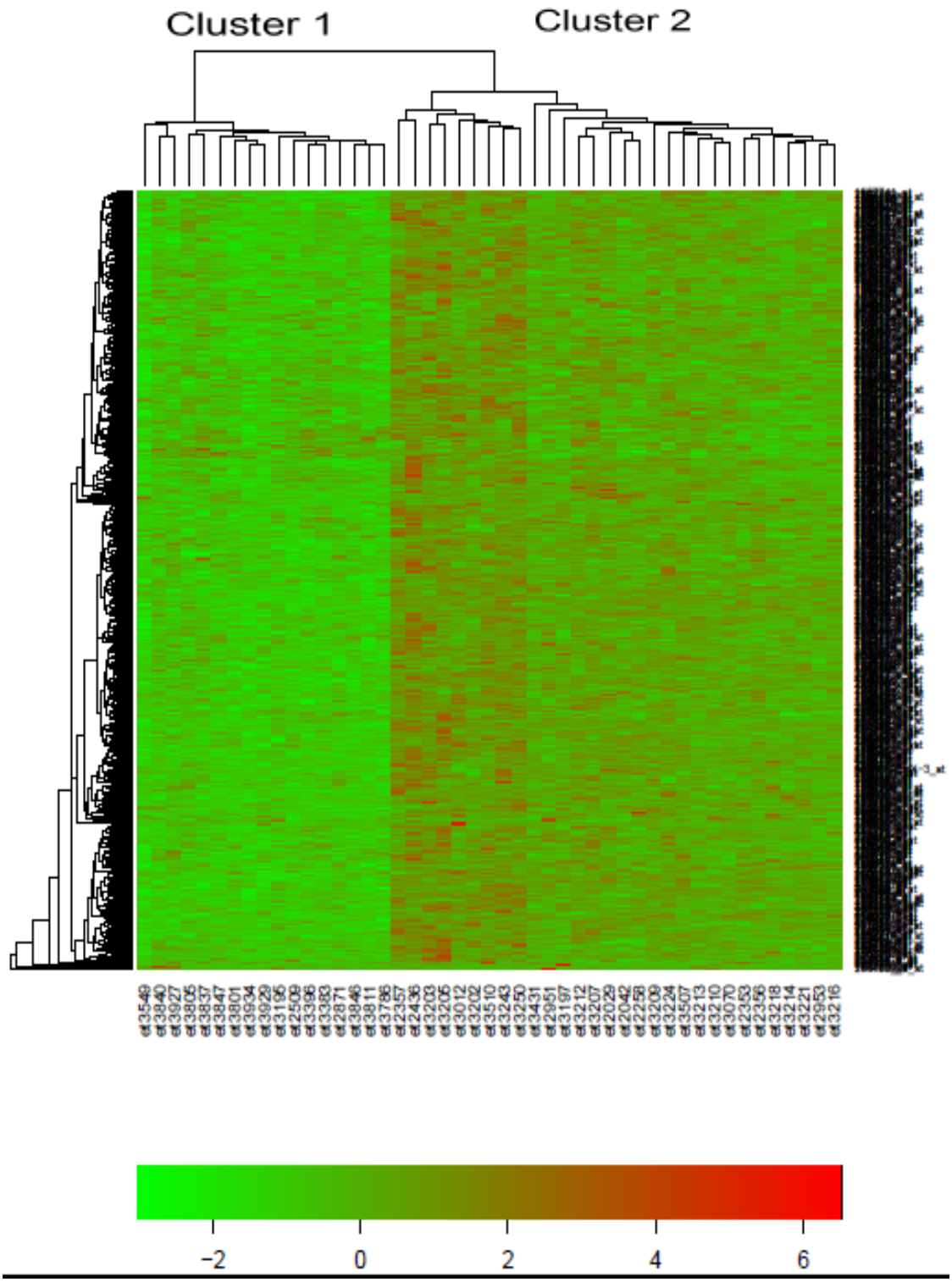


Figure 4.18: *Hierarchical cluster of 47 Gbs based on the 10% of highest CV probesets.* This figure display the molecular profile of the 10% of highest CV probesets. Columns are Gb cases and rows probesets. The bottom bar indicates the intensity (arbitrary scale) of probesets per each sample.

Finally, an interesting result was obtained by performing a hierarchical cluster

with the 10% of highest CV probesets (see figure 4.18). Apart from displaying two clear clusters from the columns dendogram, the associated molecular-profile is well defined for each cluster. In fact, the largest cluster seems to harbor two molecular-profiles, but the *silhouette* statistics rejected such hypothesis (see figure 4.19).

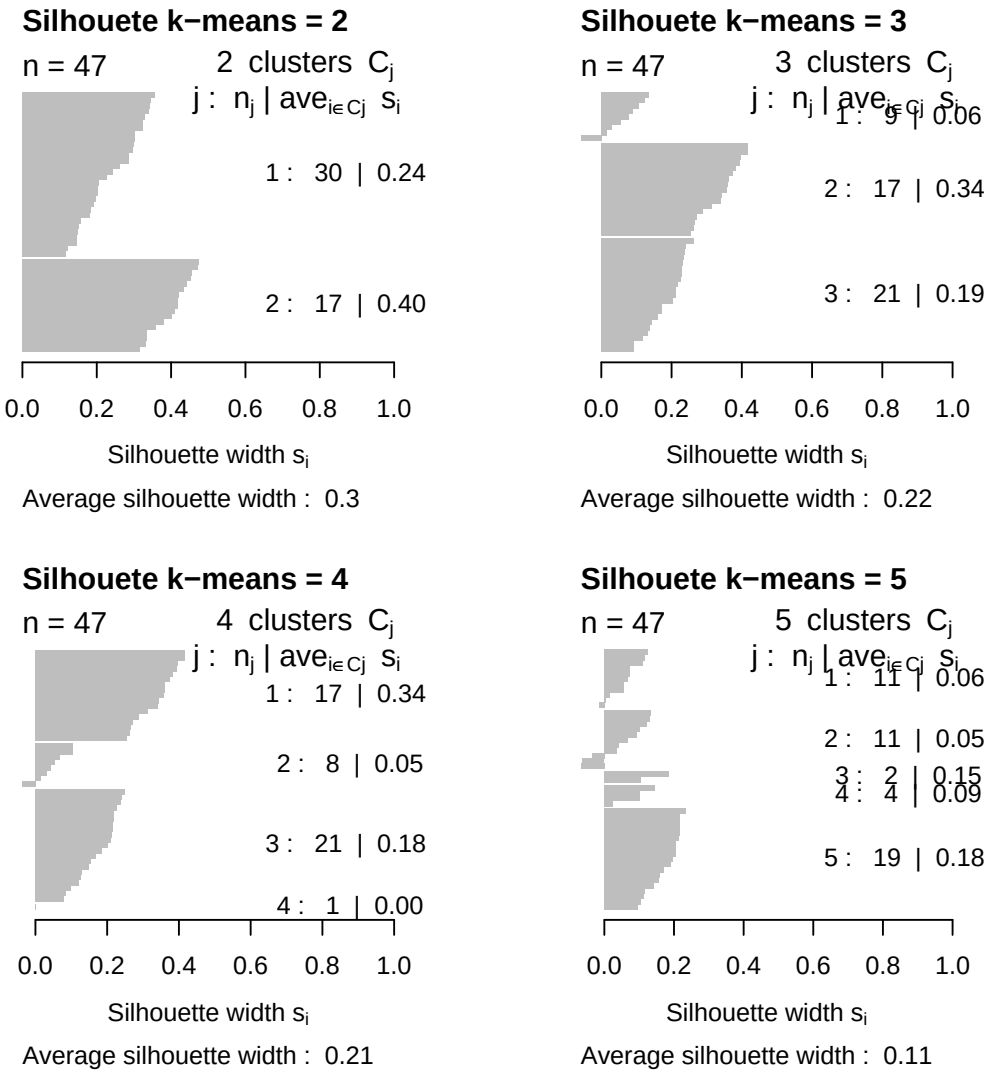


Figure 4.19: *Silhouette plot to determine the number of clusters from figure 4.18.* Four cluster combinations (2, 3, 4 and 5) were predicted by the *silhouette* statistics.

At any rate, this last approach provided the highest average mean for the *silhouette* statistics. However, the probesets selected displayed low signals of fluorescence. For that reason, we repeated the hierarchical cluster, but we selected among the genes with highest CV, those probesets with fluorescence signals higher than 1000 a.u. in at least 15% of cases. This selection approach produced a hierarchical cluster with more defined groups (see figure 4.20).

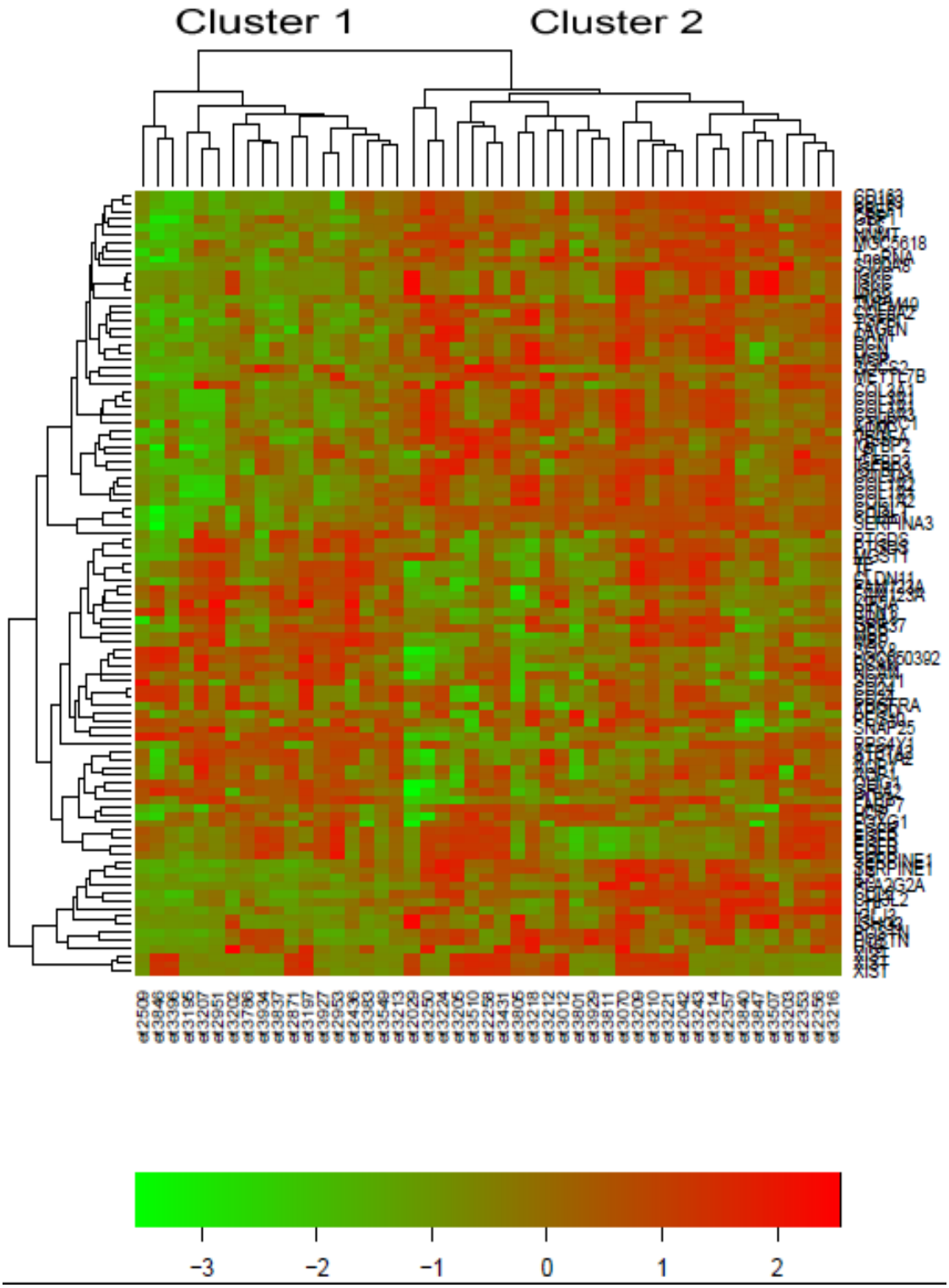


Figure 4.20: *Hierarchical cluster of 47 Gbs based on the 100 probesets of highest CV and with high fluorescence signals.* This figure display the molecular profile of the 100 probesets of highest CV and with fluorescence signals higher than 1000 a.u. in at least 15% of cases. Columns are Gb cases and rows probesets. The bottom bar indicates the intensity (arbitrary scale) of probesets per each sample.

Two clusters of Gbs were predicted by the *silhouette* statistics (see figure 4.21).

One composed of 18 Gbs (cluster 1) and the other one composed of 29 Gbs (cluster 2). The averaged signal of fluorescence of the 100 probesets used for this hierarchical cluster was lower in the cluster 1 than in the cluster 2 (720.6 and 1200.5, respectively, p-value=$1.67x10^{-06}$, ranks-based Wilcoxon test). Among the 100 probesets, genes that encodes proteins involved in various biological functions relevant in tumour progression were detected: proliferative factors (*EGFR*, *IGFBP2*,*IGFBP3*, *TGFBI*, *PDGFRA* and *VEGFA*), collagen isoforms (*COL1A1*, *COL1A2*,*COL3A1* and *COL6A3*) and transmembrane proteins (*CD24* and *CD163*).

These genes or their isoforms have been described in previous work as belonging to a molecular signature for glial tumours[69, 1, 195]. Similarly, gene signatures for glioblastoma subtypes also included isoforms of these genes[68, 167, 196, 197]. Therefore, the proposed gene signature to identify clusters of glioblastoma agrees with previous work.

Interestingly, Tso and collaborators defined in primary glioblastomas a set of significantly overexpressed genes, which included probesets that represent *CD163*, *CHI3L2*, *CHI3L1*, *COL6A2*, *COL5A1*, *EGFR*, *FABP5*, *IGFBP2*, *SERPINA3* and *VEGF*[68]. These genes or similar isoforms were found overexpressed in cluster 2, but not those probesets that represent the *EGFR* gene (see table 4.8 and annex A-11).

This result suggests that cluster 2 could be a group enriched in primary glioblastomas. Interestingly, a small difference in the averaged age of patients in each cluster was found (cluster 1= 54.6 $\pm$ 20 years and cluster 2= 61.5 $\pm$ 9.9 years, p-value=0.1995, ranks-based Wilcoxon test). Although the difference was not statistically significant, those patients younger than 30 years were only included in cluster 1, while all patients except one in cluster 2 were older than 40 years. This also agrees with the averaged age of patients that display a molecular profile of primary glioblastoma, as described by Tso and collaborators[68].

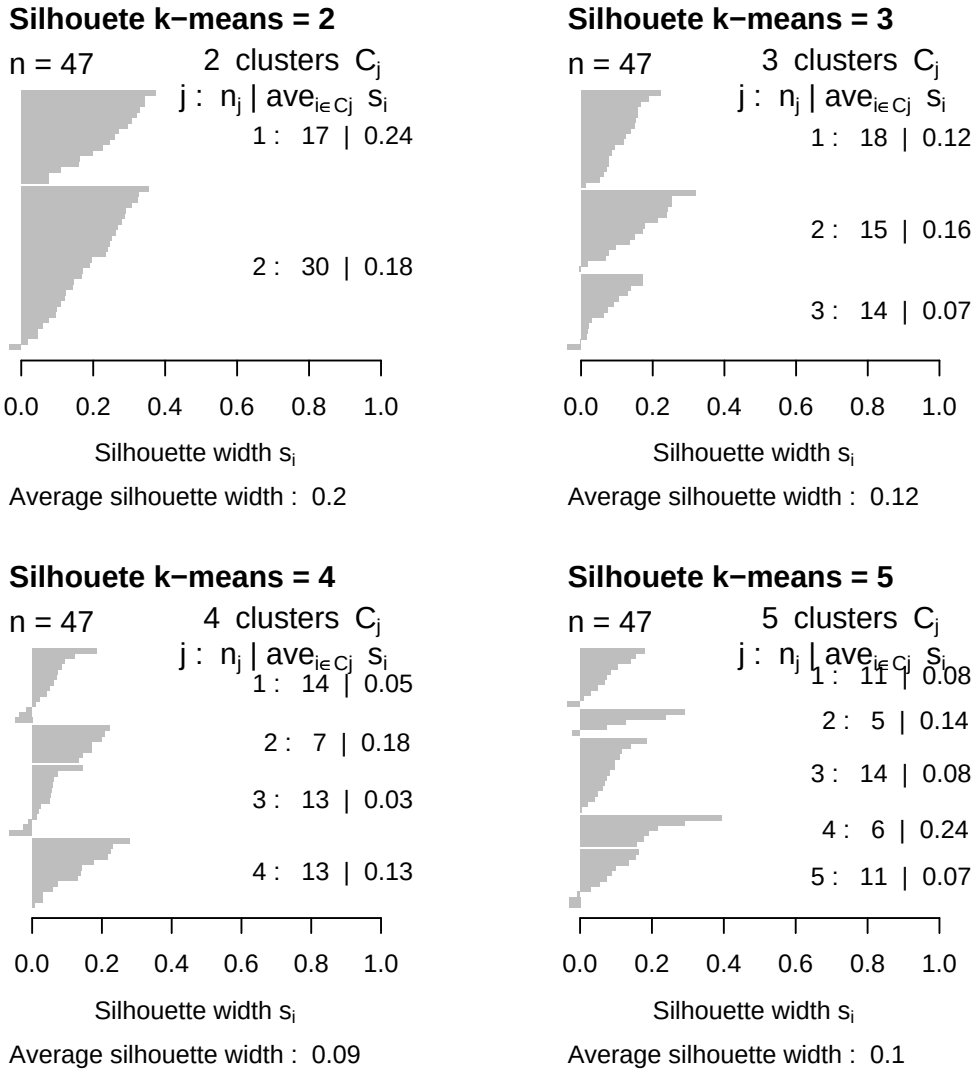


Figure 4.21: *Silhouette plot to determine the number of clusters from figure 4.18.* Four cluster combinations (2, 3, 4 and 5) were predicted by the *silhouette* statistics.

Gene symbol	Probeset	Cluster 1/Cluster 2	q-value
XIST	224588_at	0.69	$7.7x10^{-02}$
XIST	227671_at	0.56	$7.0x10^{-02}$
XIST	221728_x_at	0.75	$2.9x10^{-01}$
CHI3L2	213060_s_at	0.12	$2.4x10^{-03}$
SERPINE1	202628_s_at	0.07	$8.1x10^{-05}$
SERPINE1	202627_s_at	0.14	$1.0x10^{-04}$
EGFR	201984_s_at	0.98	$2.8x10^{-01}$
EGFR	232541_at	1.12	$1.4x10^{-01}$
EGFR	224999_at	0.93	$2.0x10^{-01}$
EGFR	201983_s_at	0.93	$2.4x10^{-01}$
FABP7	205029_s_at	0.51	$2.2x10^{-02}$
PDGFRA	203131_at	1.40	$1.9x10^{-01}$
CD24	216379_x_at	2.22	$1.1x10^{-01}$
CD24	209771_x_at	2.07	$1.1x10^{-01}$
SERPINA3	202376_at	0.44	$2.3x10^{-03}$
CHI3L1	209396_s_at	0.21	$6.2x10^{-05}$
CHI3L1	209395_at	0.20	$5.0x10^{-05}$
COL1A2	202403_s_at	0.17	$6.1x10^{-04}$
COL1A1	1556499_s_at	0.19	$2.4x10^{-04}$
COL1A2	202404_s_at	0.16	$3.6x10^{-04}$
COL3A1	215076_s_at	0.13	$1.6x10^{-04}$
IGFBP3	212143_s_at	0.22	$1.0x10^{-03}$
IGFBP3	210095_s_at	0.24	$1.2x10^{-03}$
IGFBP2	202718_at	0.49	$1.2x10^{-02}$
VEGFA	212171_x_at	0.35	$2.3x10^{-03}$
COL6A3	201438_at	0.06	$7.0x10^{-05}$
COL1A1	202310_s_at	0.10	$1.8x10^{-04}$
COL3A1	211161_s_at	0.11	$1.7x10^{-04}$
COL3A1	201852_x_at	0.09	$1.5x10^{-04}$
COL5A2	221729_at	0.25	$1.3x10^{-04}$
CD163	203645_s_at	0.15	$4.9x10^{-05}$
CD163	215049_x_at	0.18	$6.6x10^{-05}$

Table 4.8: *Summary table of genes in common with Tso and collaborators.* This table shows some genes more expressed in primary glioblastoma as described by Tso and collaborators[68]. Most of these genes were also found more expressed in cluster 2 from figure 4.20. The three first genes in the table were not described by Tso and collaborators, but may justify the low proportion of females in cluster 1.

Another interesting factor that was not identified by Tso and collaborators was the gender of patients. The 28.6% of patients in cluster 1 were females, while a similar percentage of females (48.3%) and males (51.7%) was detected in cluster 2. The difference in proportion of females between clusters was confirmed to

be significant by the Pearson's chi-square test (p-value=0.03704). Interestingly, three probesets that represent the X chromosome inactivation factor (*XIST*) gene showed lower signals of fluorescence in cluster 1 than in cluster 2. The difference between clusters showed tendency for signification (p=0.07 and 0.08) in two out of the three probesets (see table 4.8), but it was due to the unbalanced number of females and males in each cluster (see annex A-11). As *XIST*, there are 31 probesets that were selected by the variance associated to the gender in our unbalanced population composed of 18 females and 29 males (see A-11). This indicates that those genes represented by 59 probesets without bias of gender may provide a signature to identify two groups of glioblastoma. These results would suggest that the two clusters of glioblastoma detected may be correlated to clinical parameters. Therefore, we would like to propose two profiles that could differentiate each cluster of glioblastoma and potentially primary glioblastoma tumours:

Cluster 1

1. Adult and young population.
2. It would affect more males than females.
3. Low expression values of the 59 probesets without bias of gender.

Cluster 2

1. Adult and elderly population.
2. It would similarly affect males and females.
3. High expression values of the 59 probesets without bias of gender.

4.3.2 Correlation with *in vivo* and *ex vivo* ^{1}H-magnetic resonance data

To attempt to approach the identification of these possible subtypes of Gb in the clinical practice, we searched for possible differential patterns in nuclear magnetic resonance (NMR) data: single voxel at both short and long echo time (SV long and short TE), and HRMAS. The averaged spectra available for cases with HRMAS data in each cluster (see figures 4.22 and 4.23) and for paired cases in each cluster were computed for *in vivo* NMR data (see figures 4.24, 4.25, 4.26 and 4.27).

4.3.3 HRMAS data

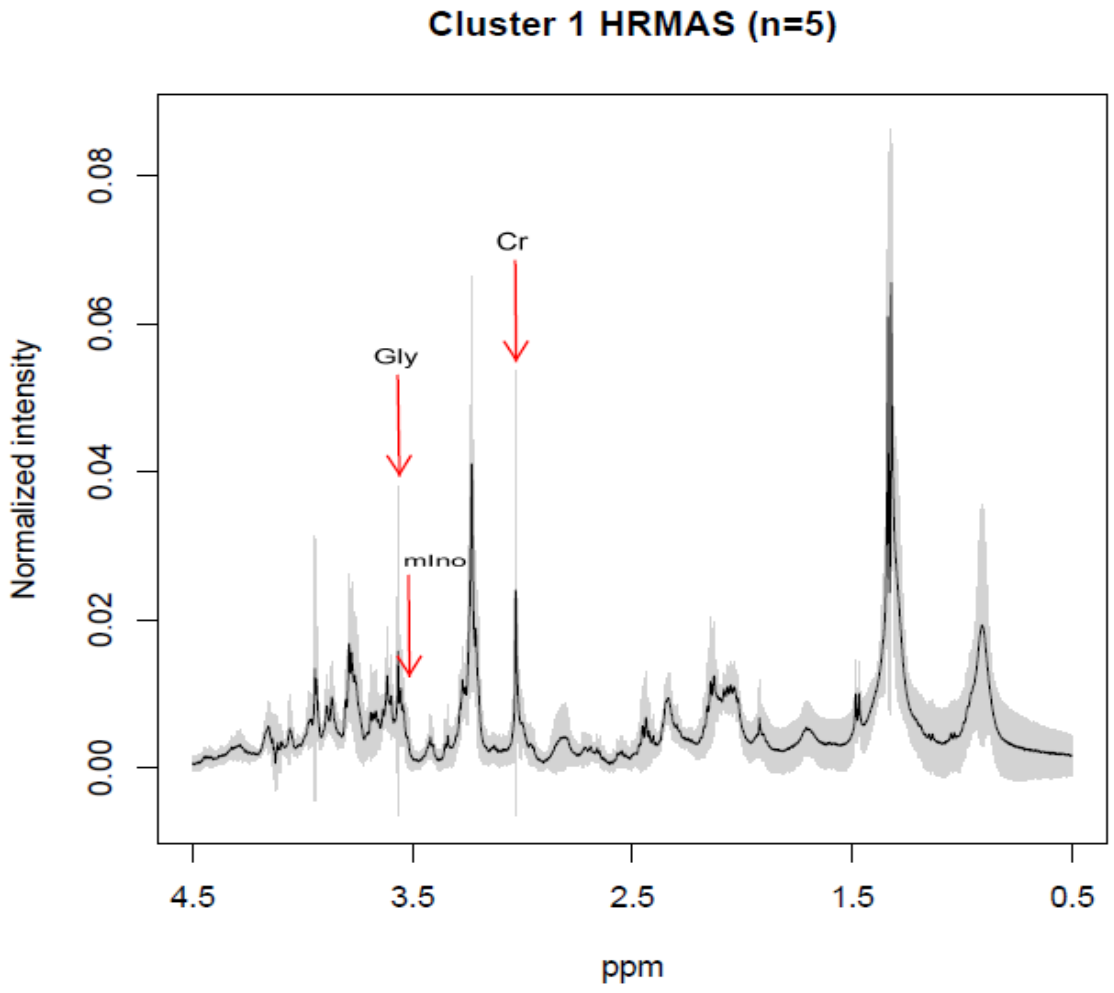


Figure 4.22: *Averaged spectrum for the cluster 1 of Gbs (fig. 4.20) using HRMAS data.* Each point in this spectrum is the mean value of the 5 Gbs for which HRMAS data was available. The averaged spectrum is represented by the black line. The gray area corresponds to the standard deviation of each point of the spectrum. Along the x axis parts per million (ppm) are depicted, while along the y axis the normalised intensity is shown.

A visual inspection of the averaged spectra revealed that in cluster 1 the region of myo-Inositol (mIno) at 3.55-3.56 ppm was higher than in cluster 2, whereas the region of glycine (Gly) at 3.53-3.55 ppm was higher in cluster 2 than in cluster 1 (see figures 4.22 and 4.23). The area under these peaks was computed using the *bining* function developed by Dr. Daniel Valverde[163]. Also, we computed the area under the peak of creatine (Cr) at 3.03 ppm as a reference to which compare both the mIno and the Gly regions. Statistical significance was found between the mIno/Cr and Gly/Cr ratio for the cluster 2 (mean(mIno/Cr)= 4.6 ± 5.1, mean(Gly/Cr)= 8.0 ± 9.4, p-value=0.03667, rank-based Wilxoxon test for paired samples), but not for cluster 1 (mean(mIno/Cr)= 3.6 ± 2.2, mean(Gly/Cr)= 4.3 ± 4.0,p-value=0.5896, rank-based Wilxoxon test for paired samples).

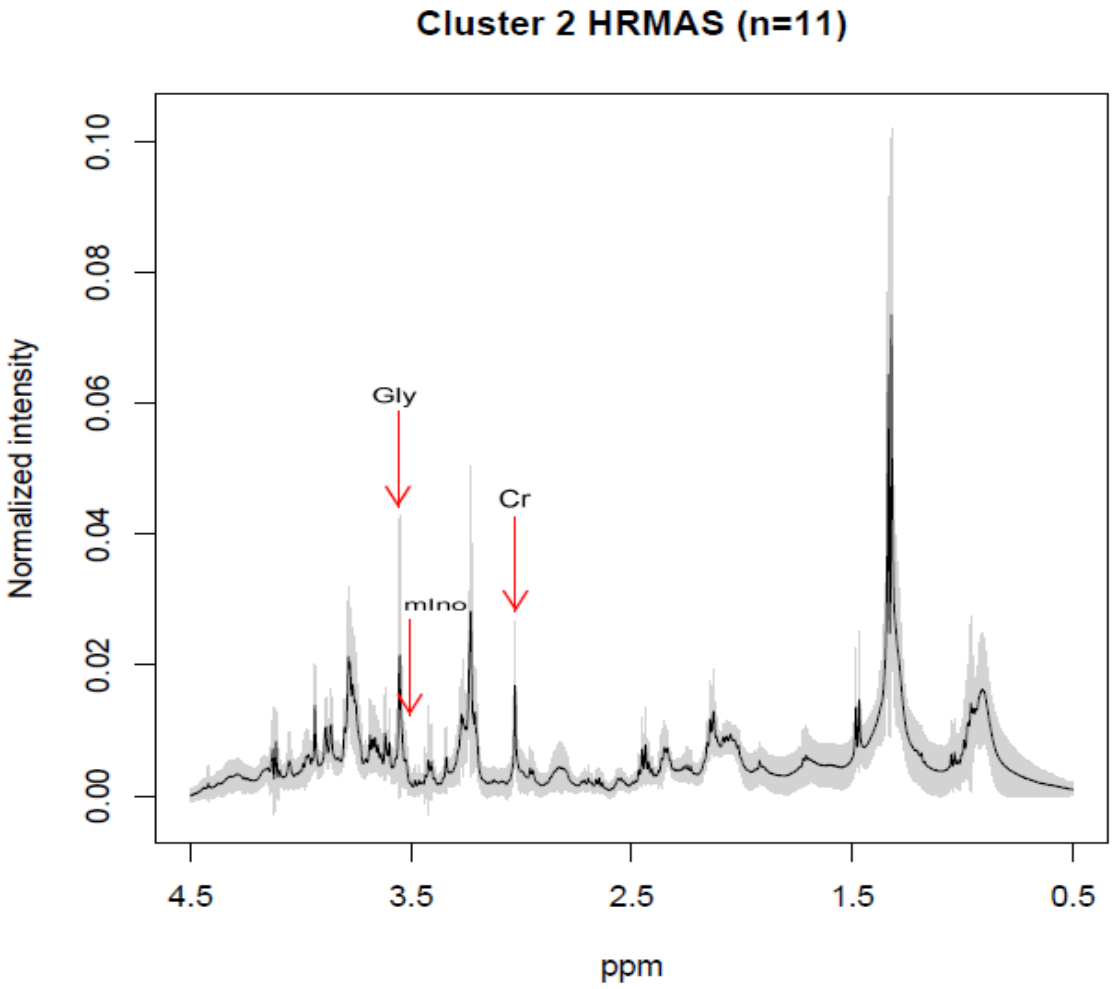


Figure 4.23: *Averaged spectrum for the cluster 2 of Gbs (fig. 4.20) using HRMAS data.* Each point in this spectrum is the mean value of the 11 Gbs for which HRMAS data was available. The averaged spectrum is represented by the black line. The gray area corresponds to the standard deviation of each point of the spectrum. Along the x axis parts per million (ppm) are depicted, while along the y axis the normalised intensity is shown.

4.3.4 *In vivo* NMR data

A visual inspection on the averaged *in vivo* NMR spectra did not apparently reveal differences between clusters of glioblastoma. However, as a higher Gly/Cr ratio than mIno/Cr was detected in cluster 2 using the HRMAS data, we computed the contribution of these signals for *in vivo* NMR data. Unlike HRMAS, the regions of Gly and mIno appear as a unique signal in the *in vivo* NMR spectra at 3.55 ppm. Differential contribution of Gly or mIno can be assessed by computation of the ratio of mIno-Gly/Cr index as described by Candiota and collaborators[198]. This index results from the division of the ratio mIno-Gly/Cr at short and long echo time. A high mIno-Gly/Cr index means a higher contribution of mIno than Gly. As the averaged mIno-Gly/Cr index was relatively high in both clusters (mean(Cluster1)= 3.03 ± 1.61 and mean(Cluster2)= 3.91 ± 3.55), the high contribution of Gly in cluster 2 detected from HRMAS data does not seem to be reproduced here. However, direct comparison of the two types of NMR data can be misleading. It has been described that HRMAS acquisition may increase the visibility of certain metabolites (i.e., creatine)[199]. In the case of Gly, it has been described that it may bind to macromolecular structures[200], which could reduce the glycine visibility in *in vivo* spectra.

Future analysis by pattern recognition methodologies of the *in vivo* patterns[162] using a larger number of cases may provide better *in vivo* biomarkers for the transcriptomically proposed glioblastoma subtypes.

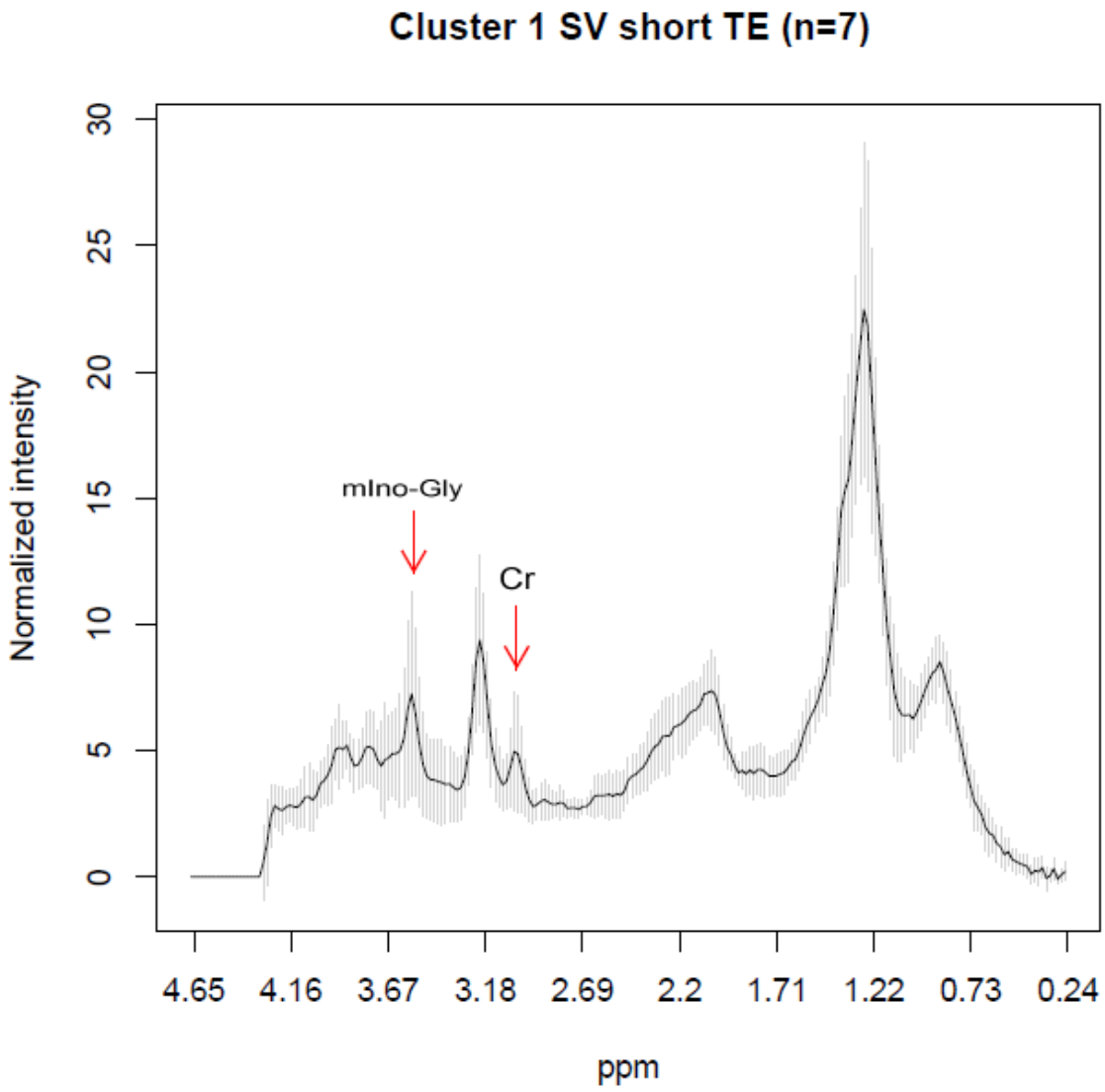


Figure 4.24: *Averaged spectrum for the cluster 1 of Gbs (fig. 4.20) using SV at short TE data.* Each point in this spectrum is the mean value of the 9 Gbs for which SV at short TE data was available. The averaged spectrum is represented by the black line. The gray area corresponds to the standard deviation of each point of the spectrum. Along the x axis parts per million (ppm) are depicted, while along the y axis the normalised intensity is shown.

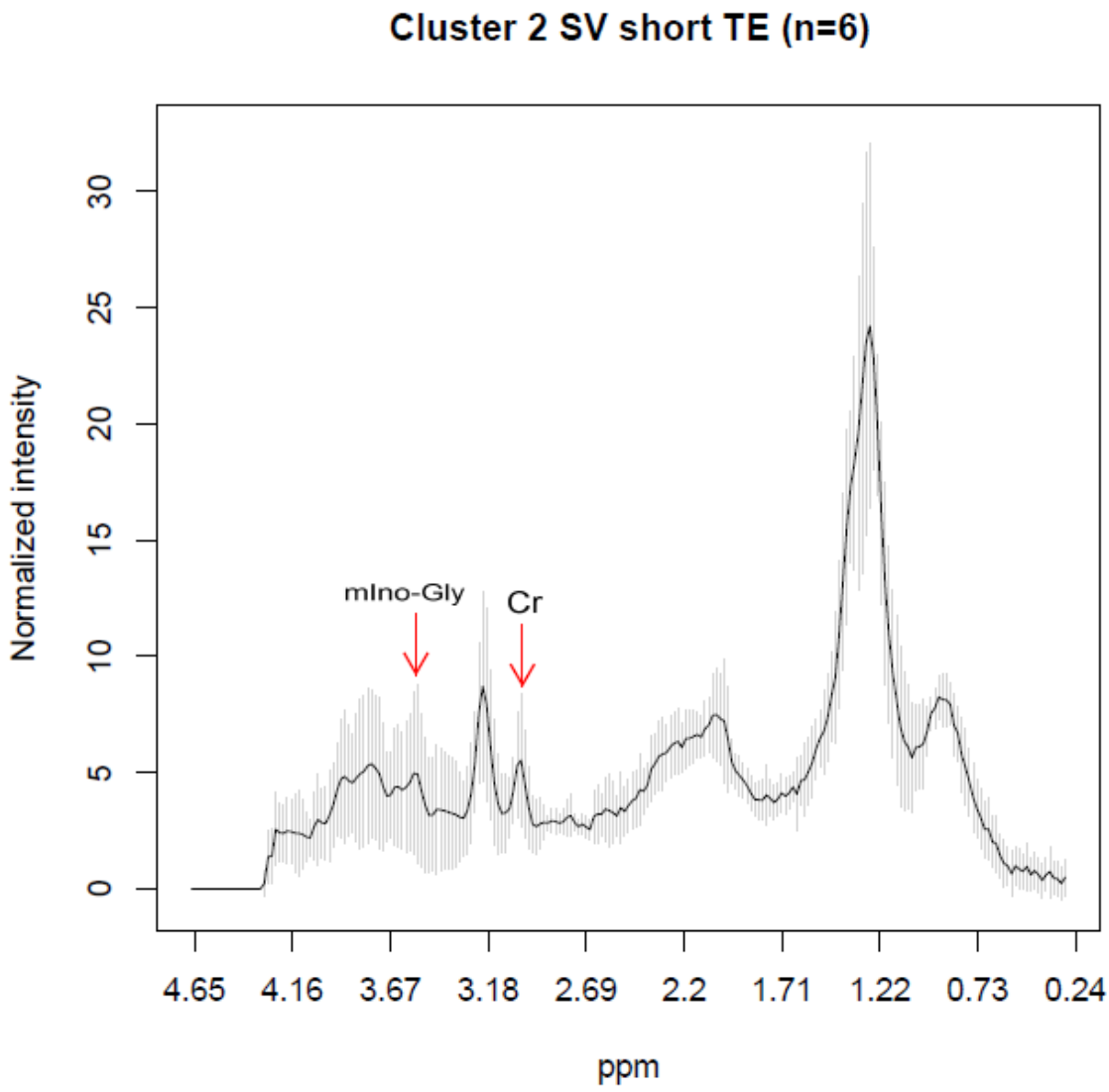


Figure 4.25: *Averaged spectrum for the cluster 2 of Gbs (fig. 4.20) using SV at short TE data.* Each point in this spectrum is the mean value of the 8 Gbs for which SV at short TE data was available. The averaged spectrum is represented by the black line. The gray area corresponds to the standard deviation of each point of the spectrum. Along the x axis parts per million (ppm) are depicted, while along the y axis the normalised intensity is shown.

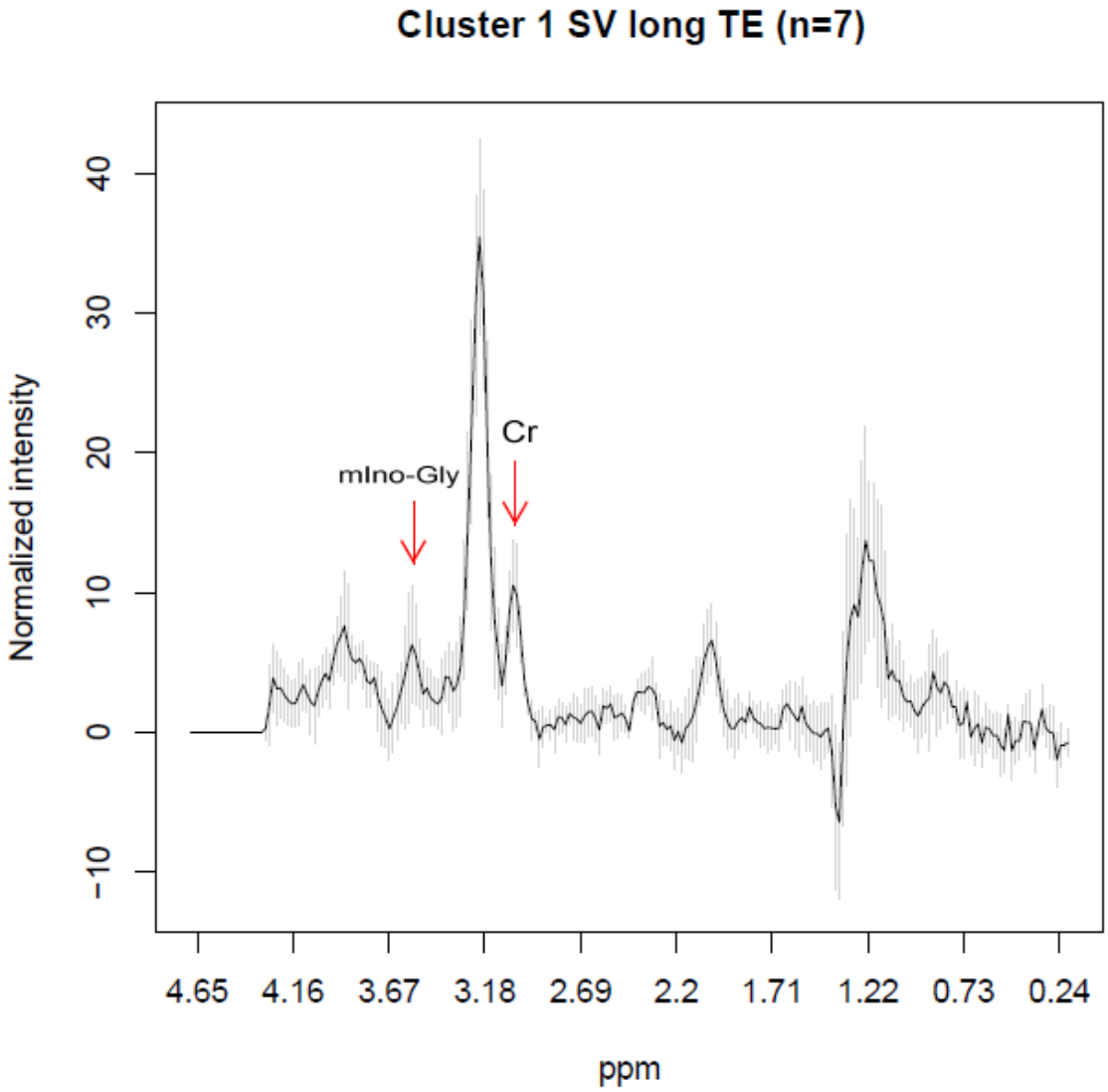


Figure 4.26: *Averaged spectrum for the cluster 1 of Gbs (fig. 4.20) using SV at long TE data.* Each point in this spectrum is the mean value of the 7 Gbs for which SV at short TE data was available. The averaged spectrum is represented by the black line. The gray area corresponds to the standard deviation of each point of the spectrum. Along the x axis parts per million (ppm) are depicted, while along the y axis the normalised intensity is shown.

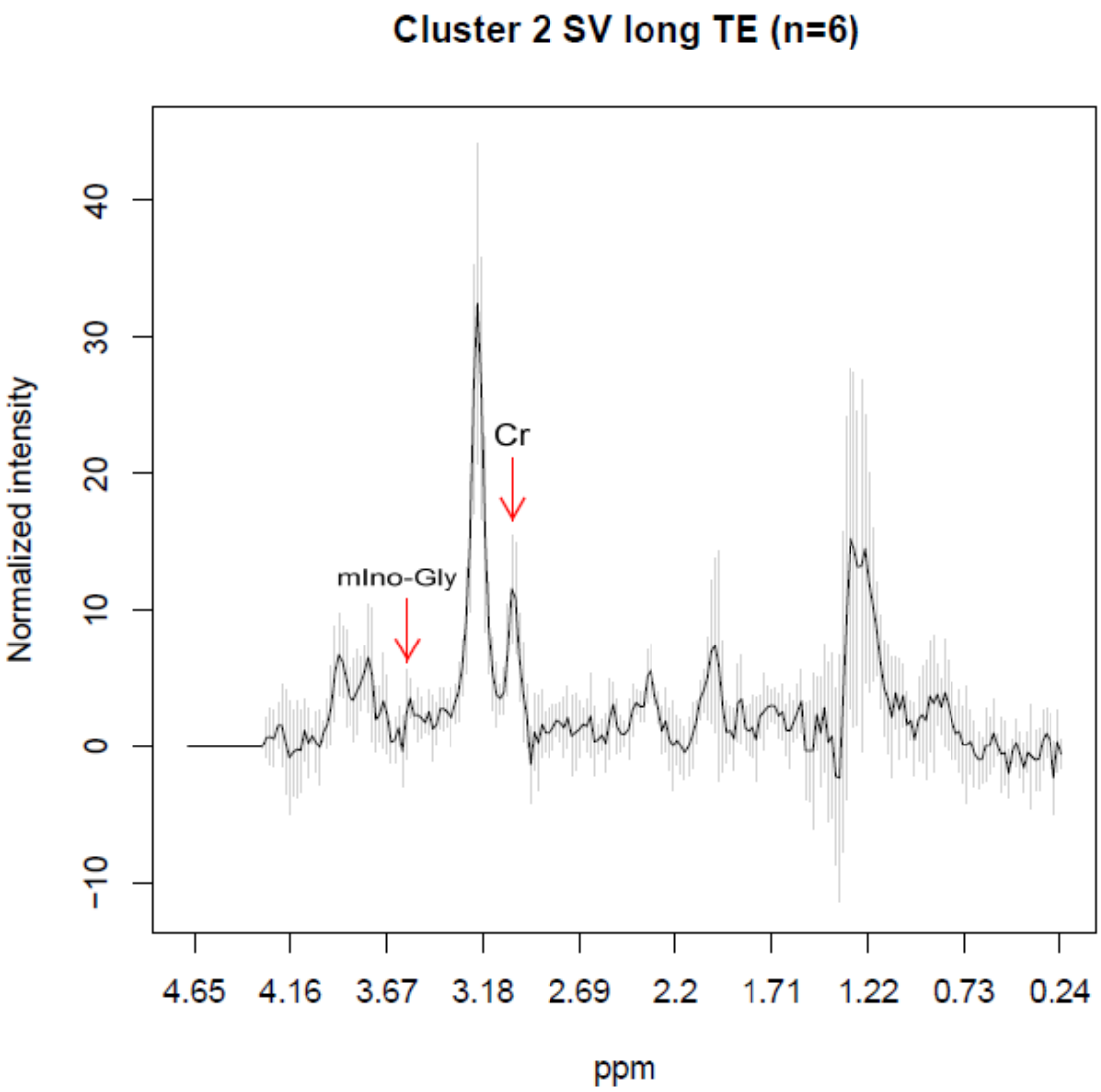


Figure 4.27: *Averaged spectrum for the cluster 2 of Gbs (fig. 4.20) using SV at long TE data.* Each point in this spectrum is the mean value of the 7 Gbs for which SV at short TE data was available. The averaged spectrum is represented by the black line. The gray area corresponds to the standard deviation of each point of the spectrum. Along the x axis parts per million (ppm) are depicted, while along the y axis the normalised intensity is shown.

4.4 Simulation of *ex vivo* ischaemia at normal body temperature in brain tumour samples of mice and C6 cells

Incubation for 30 minutes in PBS at 37ºC, did not produce differences in RNA degradation for necrotic tumours (see figure 4.28 for the evolution of tumours), compared to snap frozen in liquid nitrogen (see figure 4.29). In contrast, non-necrotic tumours showed a statistically significant decrease of the 28S/18S ratio for cases incubated for both 15 and 30 minutes in PBS at 37ºC(see figure 4.29). Thus, RNA degradation as measured from the 28S/18S ratio occurs for the non-necrotic tumour specimens, while it does not in the necrotic ones, when simulating an *ex vivo* ischaemia period at normal body temperature.

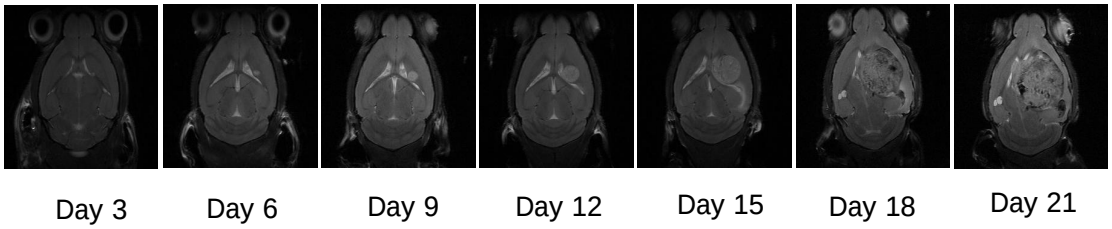


Figure 4.28: *Evolution of tumour mass induced in mice with GL261cells.* The growth of the brain tumours induced in mice by inoculation of GL261 rat glioma cells is shown. The tumour mass can be seen at day 3 after inoculation and necrosis detected after day 18 after inoculation. Images acquired by Teresa Delgado and Milena Acosta at the NMR facility of the Universitat Autònoma de Barcelona (Cerdanyola del Vallès, Catalunya), using a 7 T horizontal magnet (BioSpec 70/30, Bruker BioSpin, Ettlingen, Germany) equipped with actively shielded gradients (B-GA12 gradient coil inserted into a B-GA20S gradient system) and a quadrature receive surface coil, actively decoupled from a volume resonator with 72 mm inner diameter.

Translating these findings into the usual clinical practice, an ischaemia time at normal body temperature may cause RNA degradation and may explain the 25-30% of human biopsies showing RNA degradation, and unusable for microarray hybridization. Proper surgical practice during HBT removal requires surgeons stopping blood flow to the tumour by cauterization of visible blood vessels. This practice avoids excessive excessive bleeding, but it also results in a variable ischaemia time in patients at body temperature prior to biopsy removal (between 5 and 30 minutes). Therefore, our working hypothesis is that the time elapsed between the blood flow halt and biopsy removal, may cause RNA degradation (about 30% of tumours and perhaps the less necrotic ones).

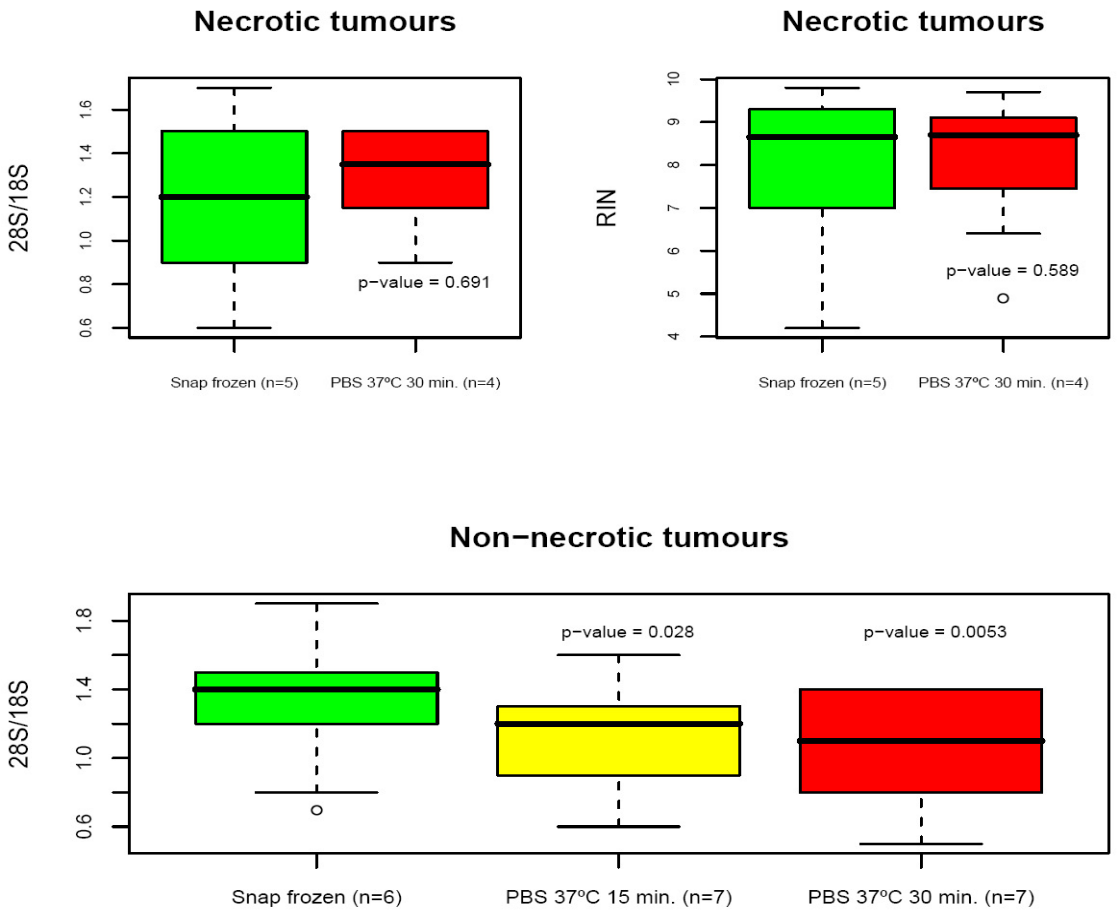


Figure 4.29: *Simulation of ischaemia at normal body temperature in mice.* Effect of ischaemia at normal body temperature on RNA degradation for both necrotic and non-necrotic mice tumours is shown. At the upper part, the box plots of both 28S/18S ratio and RIN number for necrotic-tumours are shown. At the bottom figure, 28/18S ratio for non-necrotic tumours at three time ponts of ischaemia at normal body temperature (snap frozen, 15 min. and 30 min.) is shown. In this case, only 28S/18S ratio values was evaluated, since RIN number was not computed by the 2100 Bioanalyzer for all cases. Statistical difference between conditions was assessed using the non-parametric test of Wilcoxon. Displayed p-values were computed between snap frozen samples and samples from each time point (15 and 30 min.).

Molecular and cellular explanation for our results are yet difficult to provide. Even though plenty of studies have adressed the mechanisms of mRNA turnover[201, 202, 203, 204], there is an apparent lack of published work that evaluates the involvement of ischaemia in RNA degradation.

To try to evaluate whether ischaemia is a general mechanism that induces RNA degradation in cells, a similar experiment than the one performed with mice was performed with C6 rat glioma cells. In this case, necrotic and non-necrotic stages of intracranial tumours were partially mimicked in C6 cells as post-confluence (partial proliferation arrest, 7 days culture) and logarithmic phase (5 days culture) cells, respectively. If RNA degradation could be detected in post-confluence cells after 30 minutes at 37ºC, but not in logarithmic phase cells, the RNA instability could

be tentatively correlated to an ischaemia effect onto fast proliferating cells.

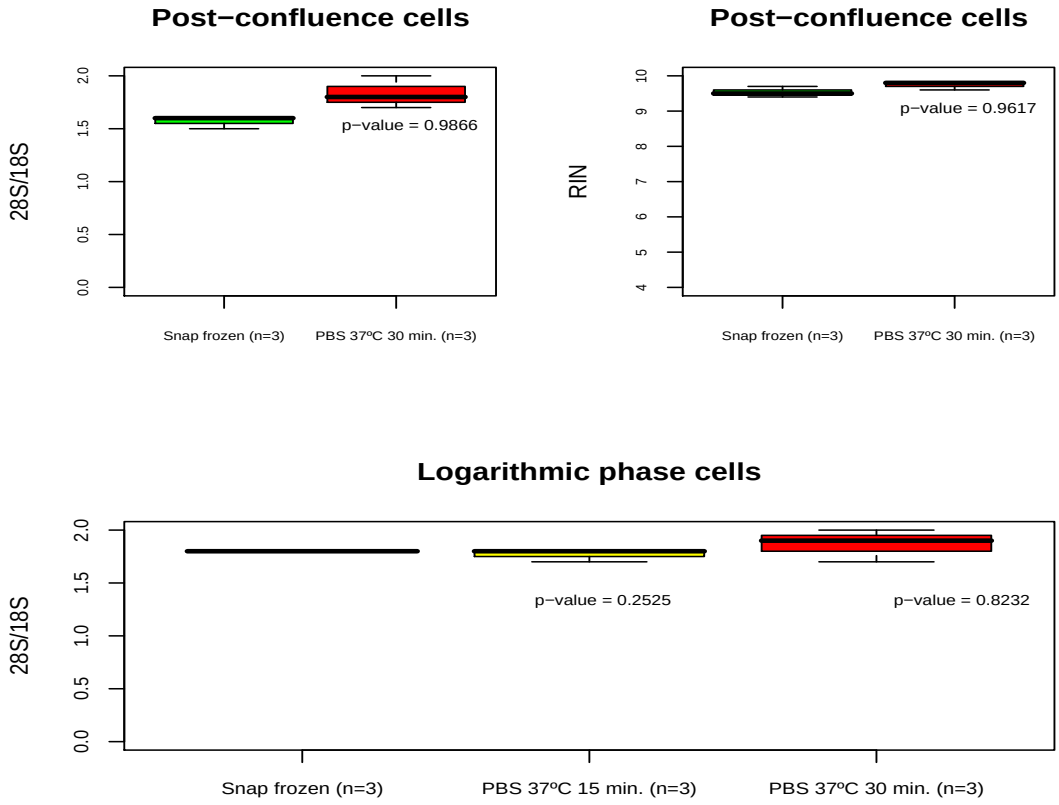


Figure 4.30: *Simulation of ischaemia at normal body temperature in C6 cells.* Effect of ischaemia at normal body temperature on RNA degradation for both post-confluence and logarithmic phase C6 cells is shown. At the upper part, the box plots of both 28S/18S ratio and RIN number for post-confluence cells are shown. At the bottom figure, 28/18S ratio for logarithmic phase cells at three time ponts of ischaemia at normal body temperature (snap frozen, 15 min. and 30 min.) is shown. In this case, only 28S/18S ratio values was evaluated, since RIN number was not computed by the 2100 Bioanalyzer for all cases. Statistical difference between conditions was assessed using using the non-parametric test of Wilcoxon. Displayed p-values were computed between snap frozen samples and samples from each time point (15 and 30 min.).

There was no significant difference for post-confluence and logarithmic phase cells at any of the evaluated conditions (see figure 4.30). This result clearly demonstrates that the ischaemia time may explain RNA degradation detected in intact GL261 tumours, but it can not be considered a general mechanism in cultured cells. In that case, we conclude that RNA degradation may possibly be induced by some other factor, induced by ischaemia, present in the tumour mass, but absent in cultured cells.

Chapter 5

CONCLUSIONS

1. 255 biopsy cases of primary and metastasic human brain tumours have been accrued from the hospitals of the Barcelona metropolitan area for the eTUMOUR, HealthAgents and MEDIVO2 projects and processed to allow transcriptomic analysis.

2. The original diagnosis, the amount of blood in the biopsy and the collection medium (liquid nitrogen or RNA*later*) could not explain the percentage of biopsies (27.5%) with unusable RNA for transcriptomic analysis. Simulation of an *ex vivo* period of ischaemia in intracranial tumours induced in C57 mice demonstrated that a 15-30 minutes period can cause significant RNA degradation in non-necrotic tumours, but does not affect RNA integrity of specimens containing necrotic regions. This would suggest that the blood flow arrest in tumours induced by surgeons before biopsy removal, may induce RNA degradation in non-necrotic regions of tumours, and could explain why RNA of both benign and malignant tumours was found degraded in similar percentages.

3. A formula based on the expression profile of only 4 genes (*GFAP*, *PTPRZ1*, *GPM6B* and *PRELP*) obtained from cDNA microarrays can perfectly and objectively predict Gbm and Mm biopsy cases. This has provided a proof-of-principle to use microarray-based gene profiling as an objective predictor technology in the clinical practice. Perfect prediction of both Affymetrix data accrued at the UAB and publicly available datasets suggests that microarray data could be made compatible across different platforms for predictor development.

4. We propose a gene signature of benignity and malignity for HBTs based on

the 629 genes with q-value lower than 0.02 from cDNA microarray data. The signature of benignity would be composed by SLRP, cytochrome-related and specific cadherin genes overexpresed in Mms, whereas the signature of malignity would be composed by tubulins and specific cadherins overexpressed in Gbms.

5. *RMA* was the best method for background correction and data normalisation of Affymetrix experiments. Selection of probesets based on their differential expression produced correct classification results for pairwise comparisons, but PCA selection was better for three-class classifiers. The LDA algorithm produced the best results in most cases and all predictors were statistically different from random assignation of classes.

6. Among the models developed using data from Affymetrix microchips, the highest percentage of correct classification was 92% for Lgg-Gb, 81.8% for Ag-Gb and 77.1% for Lgg-Ag comparison. As samples misclassified in Lgg-Ag comparison were only Lggs, it may indicate that the molecular profile of Lggs is similar to Ags, although they show different histological features.

7. Development of prediction models based on OPD and ECD has indicated that a diagnosis based on the examination of various specialists may be more accurate, when discrepancies exists between both diagnoses. However, a bias in diagnosis could be produced in the case that each pathologist examined a tissue slice obtained from different parts of the tumour.

8. Comparison of the gene expression profile of Mgs and Gbs obtained from Affymetrix microchips data with the gene profile obtained from cDNA microarrays data confirmed the signatures of benignity and malignity for HBTs. Moreover, these signatures were increased with a group of genes encoding chemokines and another group encoding collagen proteins, which were found overexpressed in glial tumours and meningiomas, respectively.

9. The highest number of genes differentially expressed across comparisons of glial tumours (Lgg-Ag, Lgg-Gb and Ag-Gb) were detected in the Lgg-Ag comparison. This may indicate that progression from Lgg to Gb tumours do not necessarily requires an Ag intermediate stage. If so, a large number of genes would be differentially expressed between Lgg and Gb tumours. These result agrees with previous work[68, 1, 197].

10. The hierarchical cluster of glioblastomas based on the expression level of 100 probesets with both highest CV and signals of intensity higher than 1000 a.u. in at least 15% of cases defined 2 groups of glioblastomas. Most probesets selected were overexpressed in cluster 2 and they coded for some genes previously found highly expressed in primary glioblastomas[68]. On the basis of our data and the existing literature, it can be proposed that cluster 2 could represent primary glioblastomas and affected patients may have lower survival.

Bibliography

[1] Phillips HS, Kharbanda S, Chen R, Forrest WF, Soriano RH, Wu TD, et al. Molecular subclasses of high-grade glioma predict prognosis, delineate a pattern of disease progression, and resemble stages in neurogenesis. Cancer Cell. 2006 Mar;9(3):157–173. Available from: `http://dx.doi.org/10.1016/j.ccr.2006.02.019`. xiv, 52, 54, 91, 112, 114, 119, 124, 140

[2] Behin A, Hoang-Xuan K, Carpentier AF, Delattre JY. Primary brain tumours in adults. Lancet. 2003;361:323–331. 3, 12, 92

[3] Mischel PS, Cloughesy TF, Nelson SF. DNA-microarray analysis of brain cancer: molecular classification for therapy. Nat Rev Neurosci. 2004;5:782–792. 3, 11, 92

[4] Armstrong TS, Cohen MZ, Eriksen LR, Hickey JV. Symptom clusters in oncology patients and implications for symptom research in people with primary brain tumors. J Nurs Scholarsh. 2004;36:197–206. 3, 4

[5] Noble M, Dietrich J. The complex identity of brain tumors: emerging concerns regarding origin, diversity and plasticity. Trends Neurosci. 2004 Mar;27:148–154. 3, 92

[6] Oliver TG, Wechsler-Reya RJ. Getting at the root and stem of brain tumors. Neuron. 2004 Jun;42:885–888. 3

[7] Parkin DM, Bray F, Ferlay J, Pisani P. Global cancer statistics, 2002. CA Cancer J Clin. 2005;55:74–108. 3, 4

[8] Boyle P, Ferlay J. Cancer incidence and mortality in Europe, 2004. Ann Oncol. 2005 March;16:481–488. 3

[9] Baldwin RT, Preston-Martin S. Epidemiology of brain tumors in childhood–a review. Toxicol Appl Pharmacol. 2004 Sep;199:118–131. 3

[10] Lovely MP. Symptom management of brain tumor patients. Semin Oncol Nurs. 2004 Nov;20:273–283. 3, 4

[11] Patel MR, Tse V. Diagnosis and staging of brain tumors. Semin Roentgenol. 2004 Jul;39:347–360. 4

[12] Figarella-Branger D, Bouvier C. Histological classification of human gliomas: state of art and controversies. Bull Cancer. 2005 Apr;92:301–309. 4, 10

[13] Louis DN, Ohgaki H, Wiestler OD, Cavenee WK, Burger PC, Jouvet A, et al. The 2007 WHO classification of tumours of the central nervous system. Acta Neuropathol. 2007 Aug;114:97–109. 5, 7, 9, 10, 15

[14] Hashimoto N, Murakami M, Sasajima H, Takahashi Y, Mineura K. Diagnostic molecular genetics and its application to clinical decision in brain tumors. Internat Congress Series. 2002;1247:221–229. 5, 10

[15] Vaquero J, Coca S. Patologia tumoral del Sistema nervioso. Edimsa, S.A.; 2004. 6

[16] Zülch KJ. Histological typing of tumours of the central nervous system. World Health Organization; 1979. 5, 7

[17] Kleihues P, Burger PC, Scheithauer BW. Histological typing of tumours of the nervous system. World Health Organization international histological classification of tumours. Springer; 1993. 5

[18] Kleihues P, Cavenee WK. Pathology and genetics of tumours of the nervous system. IARC Press; 2000. 5, 61

[19] Louis DN, Ohgaki H, D WO, Cavenee WK. WHO Classification of tumours of the central nervous system. IARC; 2007. 5, 6, 8

[20] Kleihues P, Burger PC, Scheithauer BW. The new WHO classification of brain tumours. Brain Pathol. 1993 Jul;3:255–268. 7

[21] Ringertz J. Grading of gliomas. Acta Pathol Microbiol Scand. 1950;. 7

[22] Daumas-Duport C, Varlet P. Dysembryoplastic neuroepithelial tumors. Rev Neurol (Paris). 2003 Jul;159:622–636. 7

[23] Giannini C, Scheithauer BW, Steinberg J, Cosgrove TJ. Intraventricular perineurioma: case report. Neurosurgery. 1998 Dec;43:1478–1481. 7

[24] Neder L, Colli BO, Machado HR, Carlotti CG, Santos AC, Chimelli L. MIB-1 labeling index in astrocytic tumors–a clinicopathologic study. Clin Neuropathol. 2004;23:262–270. 7

[25] Bauchet L, Rigau V, Mathieu-Daude H, Fabbro M, Campello C, Figarella-Branger D, et al. Groupe de neuro-oncologie du Languedoc-Roussillon, avec la participation de l'Association des neuro-oncologues d'expression française et la Société française de neuro-chirurgie. Recensement national des tumeurs primitives du système nerveux central. In: Livre du Congrès des 2es Journées Françaises de Neurochirurgie. vol. 72; 2003. . 10

[26] Maher EA, Furnari FB, Bachoo RM, Rowitch DH, Louis DN, Cavenee WK, et al. Malignant glioma: genetics and biology of a grave matter. Genes Dev. 2001 Jun;15:1311–1333. 10, 11, 12, 13, 14, 112

[27] Louis DN, Holland EC, Cairncross JG. Glioma classification: a molecular reappraisal. Am J Pathol. 2001 Sep;159:779–786. 10, 11, 13, 112

[28] Hunter KE, Hatten ME. Radial glial cell transformation to astrocytes is bidirectional: Regulation by a diffusible factor in embryonic forebrain. Proc Natl Acad Sci USA. 1995 March;92(6):2061–2065. 12

[29] Soriano E, Alvarado-Mallart RM, Dumesnil N, Río JAD, Sotelo C. Cajal-Retzius cells regulate the radial glia phenotype in the adult and developing cerebellum and alter granule cell migration. Neuron. 1997 Apr;18(4):563–577. 12

[30] Ohgaki H, Kleihues P. Genetic pathways to primary and secondary glioblastoma. Am J Pathol. 2007 May;170:1445–1453. 13, 14, 52

[31] Mellinghoff IK, Wang MY, Vivanco I, Haas-Kogan DA, Zhu S, Dia EQ, et al. Molecular determinants of the response of glioblastomas to EGFR kinase inhibitors. N Engl J Med. 2005 Nov;353:2012–2024. 13

[32] Yeh HJ, Ruit KG, Wang YX, Parks WC, Snider WD, Deuel TF. PDGF A-chain gene is expressed by mammalian neurons during development and in maturity. Cell. 1991;64:209–216. 13

[33] Yeh HJ, Silos-Santiago I, Wang YX, George RJ, Snider WD, Deuel TF. Developmental expression of the platelet-derived growth factor -receptor gene in mammalian central nervous system. Proc Natl Acad Sci. 1993;90:1952–1956. 13

[34] Heldin CH. Protein tyrosine kinase receptors. Cancer Surv. 1996;27:7–24. 14

[35] Wolswijk G, Riddle PN, Noble M. Platelet-derived growth factor is mitogenic for O-2A adult progenitor cells. Glia. 1991;4:495–503. 14

[36] Howe FA, Opstad KS. 1HMR spectroscopy of brain tumours and masses. NMR Biomed. 2003 May;16:123–131. 14

[37] Pang BC, Wan WH, Lee CK, Khu KJ, Ng WH. The role of surgery in high-grade glioma–is surgical resection justified? A review of the current knowledge. Ann Acad Med Singap. 2007 May;36:358–363. 14, 15

[38] Stupp R, Hegi ME, van den Bent MJ, Mason WP, Weller M, Mirimanoff RO, et al. Changing paradigms–an update on the multidisciplinary management of malignant glioma. Oncologist. 2006 Feb;11:165–180. 15

[39] Brada M. Radiotherapy in malignant glioma. Ann Oncol. 2006 Sep;17 Suppl 10:x183–185. 15

[40] Friedman HS, Kerby T, Calvert H. Temozolomide and treatment of malignant glioma. Clin Cancer Res. 2000 Jul;6:2585–2597. 15

[41] Lawson HC, Sampath P, Bohan E, Park MC, Hussain N, Olivi A, et al. Interstitial chemotherapy for malignant gliomas: the Johns Hopkins experience. J Neurooncol. 2007 May;83:61–70. 15

[42] Rockhill J, Mrugala M, Chamberlain MC. Intracranial meningiomas: an overview of diagnosis and treatment. Neurosurg Focus. 2007;23(4):E1. Available from: `http://dx.doi.org/10.3171/foc.2007.23.4.2`. 15, 17

[43] Ragel BT, Jensen RL. Molecular genetics of meningiomas. Neurosurg Focus. 2005 Nov;19(5):E9. 15, 16, 17

[44] Commins DL, Atkinson RD, Burnett ME. Review of meningioma histopathology. Neurosurg Focus. 2007;23(4):E3. Available from: `http://dx.doi.org/10.3171/foc.2007.23.4.4`. 15, 16

[45] Goel A, Muzumdar D, Desai KI. Tuberculum sellae meningioma: a report on management on the basis of a surgical experience with 70 patients. Neurosurgery. 2002 Dec;51(6):1358–63; discussion 1363–4. 15

[46] Lamszus K. Meningioma pathology, genetics, and biology. J Neuropathol Exp Neurol. 2004 Apr;63(4):275–286. 16, 17

[47] Antinheimo J, Sankila R, Carpén O, Pukkala E, Sainio M, Jääskeläinen J. Population-based analysis of sporadic and type 2 neurofibromatosis-associated meningiomas and schwannomas. Neurology. 2000 Jan;54(1):71–76. 17

[48] Gutmann DH, Donahoe J, Perry A, Lemke N, Gorse K, Kittiniyom K, et al. Loss of DAL-1, a protein 4.1-related tumor suppressor, is an important early event in the pathogenesis of meningiomas. Hum Mol Genet. 2000 Jun;9(10):1495–1500. 17

[49] Soyuer S, Chang EL, Selek U, Shi W, Maor MH, DeMonte F. Radiotherapy after surgery for benign cerebral meningioma. Radiother Oncol. 2004 Apr;71(1):85–90. Available from: http://dx.doi.org/10.1016/j.radonc.2004.01.006. 17

[50] Fomchenko EI, Holland EC. Mouse models of brain tumors and their applications in preclinical trials. Clin Cancer Res. 2006 Sep;12(18):5288–5297. Available from: http://dx.doi.org/10.1158/1078-0432.CCR-06-0438. 17, 18

[51] Shapiro WR, Basler GA, Chernik NL, Posner JB. Human brain tumor transplantation into nude mice. J Natl Cancer Inst. 1979 Mar;62(3):447–453. 18

[52] Lassman AB, Holland EC. Mouse models of human cancers. Chapter 12: Central nervous system tumors. Wiley J, Sons, editors. Holland EC; 2004. 18

[53] Weiss B, Shannon K. Mouse models of human cancers. Chapter 26: Preclinical trials in mouse cancer models. EC H, editor. John Wiley and Sons; 2004. 18

[54] Schwaninger M, Inta I, Herrmann O. NF-kappaB signalling in cerebral ischaemia. In: International Symposium on Neurodegeneration and Neuroprotection. vol. 34; 2006. p. 1291–1294. 18

[55] Potrovita I, Zhang W, Burkly L, Hahm K, Lincecum J, Wang MZ, et al. Tumor necrosis factor-like weak inducer of apoptosis-induced neurodegeneration. J Neurosci. 2004 Sep;24(38):8237–8244. Available from: http://dx.doi.org/10.1523/JNEUROSCI.1089-04.2004. 18

[56] Saitoh T, Nakayama M, Nakano H, Yagita H, Yamamoto N, Yamaoka S. TWEAK induces NF-kappaB2 p100 processing and long lasting NF-kappaB activation. J Biol Chem. 2003 Sep;278(38):36005–36012. Available from: http://dx.doi.org/10.1074/jbc.M304266200. 18

[57] Perkins ND, Gilmore TD. Good cop, bad cop: the different faces of NF-kappaB. Cell Death Differ. 2006 May;13(5):759–772. Available from: http://dx.doi.org/10.1038/sj.cdd.4401838. 18

[58] Iadecola C, Zhang F, Casey R, Nagayama M, Ross ME. Delayed reduction of ischemic brain injury and neurological deficits in mice lacking the inducible nitric oxide synthase gene. J Neurosci. 1997 Dec;17(23):9157–9164. 18

[59] Ekins R, Chu F, Micallef J. High specific activity chemiluminescent and fluorescent markers: their potential application to high sensitivity and 'multi-analyte' immunoassays. J Biolumin Chemilumin. 1989 Jul;4(1):59–78. Available from: http://dx.doi.org/10.1002/bio.1170040113. 19

[60] Ekins R, Chu FW. Microarrays: their origins and applications. Trends Biotechnol. 1999 Jun;17(6):217–218. 19

[61] Fodor SP, Rava RP, Huang XC, Pease AC, Holmes CP, Adams CL. Multiplexed biochemical assays with biological chips. Nature. 1993 Aug;364(6437):555–556. Available from: http://dx.doi.org/10.1038/364555a0. 19, 22

[62] Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science. 1995 Oct;270(5235):467–470. 19, 22

[63] Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature. 2001 Feb;409(6822):860–921. 19

[64] Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, et al. The sequence of the human genome. Science. 2001 Feb;291(5507):1304–1351. Available from: http://dx.doi.org/10.1126/science.1058040. 19

[65] Barrett JC, Kawasaki ES. Microarrays: the use of oligonucleotides and cDNA for the analysis of gene expression. Drug Discov Today. 2003 Feb;8(3):134–141. 19, 20, 22, 24, 25

[66] Morris CM, Wilson KE. High throughput approaches in neuroscience. Int J Dev Neurosci. 2004 Nov;22(7):515–522. Available from: http://dx.doi.org/10.1016/j.ijdevneu.2004.07.010. 19

[67] Geschwind DH. DNA microarrays: translation of the genome from laboratory to clinic. Lancet Neurol. 2003 May;2(5):275–282. 21

[68] Tso CL, Freije WA, Day A, Chen Z, Merriman B, Perlina A, et al. Distinct transcription profiles of primary and secondary glioblastoma subgroups. Cancer Res. 2006 Jan;66(1):159–167. Available from: http://dx.doi.org/10.1158/0008-5472.CAN-05-0077. 21, 51, 52, 124, 126, 140, 141

[69] Freije WA, Castro-Vargas FE, Fang Z, Horvath S, Cloughesy T, Liau LM, et al. Gene expression profiling of gliomas strongly predicts survival. Cancer Res. 2004 Sep;64(18):6503–6510. Available from: http://dx.doi.org/10.1158/0008-5472.CAN-04-0452. 21, 51, 52, 54, 124

[70] Bild AH, Yao G, Chang JT, Wang Q, Potti A, Chasse D, et al. Oncogenic pathway signatures in human cancers as a guide to targeted therapies. Nature. 2006 Jan;439(7074):353–357. Available from: http://dx.doi.org/10.1038/nature04296. 21

[71] Abdullah-Sayani A, de Mesquita JMB, van de Vijver MJ. Technology Insight: tuning into the genetic orchestra using microarrays–limitations of DNA microarrays in clinical practice. Nat Clin Pract Oncol. 2006 Sep;3(9):501–516. Available from: http://dx.doi.org/10.1038/ncponc0587. 21, 53

[72] Cardoso F, Veer LV, Rutgers E, Loi S, Mook S, Piccart-Gebhart MJ. Clinical application of the 70-gene profile: the MINDACT trial. J Clin Oncol. 2008 Feb;26(5):729–735. Available from: http://dx.doi.org/10.1200/JCO.2007.14.3222. 21, 54

[73] Pusztai L, Hess KR. Clinical trial design for microarray predictive marker discovery and assessment. Ann Oncol. 2004 Dec;15(12):1731–1737. Available from: http://dx.doi.org/10.1093/annonc/mdh466. 21

[74] Fu J, Jeffrey SS. Transcriptomic signatures in breast cancer. Mol Biosyst. 2007 Jul;3(7):466–472. Available from: http://dx.doi.org/10.1039/b618163e. 21

[75] Sotiriou C, Piccart MJ. Taking gene-expression profiling to the clinic: when will molecular signatures become relevant to patient care? Nat Rev Cancer. 2007 Jul;7(7):545–553. Available from: http://dx.doi.org/10.1038/nrc2173. 21

[76] Wang TH, Chao A. Microarray analysis of gene expression of cancer to guide the use of chemotherapeutics. Taiwan J Obstet Gynecol. 2007 Sep;46(3):222–229. 21, 41, 52

[77] Schulze A, Downward J. Navigating gene expression using microarrays, a technology review. Nature Cell Biology. 2001 August;3:E190–195. 22, 23

[78] Hardiman G. Microarray platforms–comparisons and contrasts. Pharmacogenomics. 2004 Jul;5(5):487–502. Available from: http://dx.doi.org/10.1517/14622416.5.5.487. 22, 23, 26

[79] Yauk CL, Berndt ML, Williams A, Douglas GR. Comprehensive comparison of six microarray technologies. Nucleic Acids Res. 2004;32(15):e124. Available from: http://dx.doi.org/10.1093/nar/gnh123. 23

[80] Affymetrix I. GeneChip Arrays Provide Optimal Sensitivity and Specificity for Microarray Expression Analysis; 2001-07. 24

[81] Team RIC. Agilent SurePrint Technology (Manual); 2003. 25

[82] Hughes TR, Mao M, Jones AR, Burchard J, Marton MJ, Shannon KW, et al. Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. Nature Biotechnology. 2001 April;19(4):342–247. 25

[83] Affymetrix I. Standardized Assays and Reagents for GeneChip Expression Analysis; 2003-04. 25

[84] Affymetrix I. GeneChip Expression Analysis Technical Manual With Specific Protocols for Using the GeneChip Hybridization, Wash, and Stain Kit; 2005-06. 25, 65

[85] Agilent. One-Color Microarray-Based Gene Expression Analysis; 2008. 26

[86] Ramakrishnan R, Dorris D, Lublinsky A, Nguyen A, Domanus M, Prokhorova A, et al. An assessment of Motorola CodeLink microarray performance for gene expression profiling applications. Nucleic Acids Res. 2002 Apr;30(7):e30. 26

[87] Team RDC. R Language Definition; 2005. 27

[88] Venables WN, Smith DM, the R Development Core Team. An Introduction to R, Notes on R: A Programming Environment for Data Analysis and Graphics (electronic edition); 2005. 27

[89] Yang YH, Buckley MJ, Speed TP. Analysis of cDNA microarray images. Brief Bioinform. 2001 Dec;2(4):341–349. 28, 29, 30, 31

[90] Edwards D. Non-linear normalization and background correction in one-channel cDNA microarray studies. Bioinformatics. 2003 May;19(7):825–833. 30, 35, 37, 66

[91] van Bakel H, Holstege FCP. In control: systematic assessment of microarray performance. EMBO Rep. 2004 Oct;5(10):964–969. Available from: http://dx.doi.org/10.1038/sj.embor.7400253. 31

[92] Chuaqui RF, Bonner RF, Best CJM, Gillespie JW, Flaig MJ, Hewitt SM, et al. Post-analysis follow-up and validation of microarray experiments. Nat Genet. 2002 Dec;32 Suppl:509–514. Available from: http://dx.doi.org/10.1038/ng1034. 31

[93] Tong W, Lucas AB, Shippy R, Fan X, Fang H, Hong H, et al. Evaluation of external RNA controls for the assessment of microarray performance. Nat Biotechnol. 2006 Sep;24(9):1132–1139. Available from: http://dx.doi.org/10.1038/nbt1237. 31

[94] Affymetrix I. GeneChip Expression Analysis. Data Analysis Fundamentals; 2002-04. 31, 32, 33, 34

[95] Arteaga-Salas JM, Zuzan H, Langdon WB, Upton GJG, Harrison AP. An overview of image-processing methods for Affymetrix GeneChips. Brief Bioinform. 2008 Jan;9(1):25–33. Available from: http://dx.doi.org/10.1093/bib/bbm055. 31, 32

[96] Affymetrix I. Affymetrix GeneChip Operating Software With AutoLoader, Version 1.4; 2003-05. 34

[97] Tukey J. Exploratory Data Analysis. Addison-Wesley, Reading; 1977. 35

[98] Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, Speed TP. Summaries of Affymetrix GeneChip probe level data. Nucleic Acids Res. 2003 Feb;31(4):e15. 35

[99] Wu Z, Irizarry RA, Gentleman R, Martinez-Murillo F, Spencer F. A Model-Based Background Adjustment for Oligonucleotide Expression Arrays. Journal of the American Statistical Association. 2004 December;99(468):909–917. 35

[100] Smyth GK, Speed T. Normalization of cDNA microarray data. Methods. 2003 Dec;31(4):265–273. 35, 36, 37

[101] Kroll TC, Wölfl S. Ranking: a closer look on globalisation methods for normalisation of gene expression arrays. Nucleic Acids Res. 2002 Jun;30(11):e50. 36

[102] Fan J, Tam P, Woude GV, Ren Y. Normalization and analysis of cDNA microarrays using within-array replications applied to neuroblastoma cell response to a cytokine. Proc Natl Acad Sci U S A. 2004 Feb;101(5):1135–1140. Available from: http://dx.doi.org/10.1073/pnas.0307557100. 36

[103] Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J, et al. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. Nucleic Acids Res. 2002 Feb;30(4):e15. 36

[104] Fujita A, Sato JR, de Oliveira Rodrigues L, Ferreira CE, Sogayar MC. Evaluating different methods of microarray data normalization. BMC Bioinformatics. 2006;7:469. Available from: http://dx.doi.org/10.1186/1471-2105-7-469. 36

[105] Wu W, Dave N, Tseng GC, Richards T, Xing EP, Kaminski N. Comparison of normalization methods for CodeLink Bioarray data. BMC Bioinformatics. 2005;6:309. Available from: http://dx.doi.org/10.1186/1471-2105-6-309. 37

[106] Bolstad BM, Irizarry RA, Astrand M, Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics. 2003 Jan;19(2):185–193. 37

[107] Quackenbush J. Microarray data normalization and transformation. Nat Genet. 2002 Dec;32 Suppl:496–501. Available from: http://dx.doi.org/10.1038/ng1032. 38, 39

[108] Schadt EE, Li C, Ellis B, Wong WH. Feature extraction and normalization algorithms for high-density oligonucleotide gene expression array data. J Cell Biochem Suppl. 2001;Suppl 37:120–125. 37

[109] Li C, Wong WH. Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. Proc Natl Acad Sci U S A. 2001 Jan;98(1):31–36. Available from: http://dx.doi.org/10.1073/pnas.011404098. 37

[110] Xu X, Tian L, Wei LJ. Combining dependent tests for linkage or association across multiple phenotypic traits. Biostatistics. 2003 Apr;4(2):223–229. Available from: http://dx.doi.org/10.1093/biostatistics/4.2.223. 39

[111] Jung SH, Jang W. How accurately can we control the FDR in analyzing microarray data? Bioinformatics. 2006 Jul;22(14):1730–1736. Available from: http://dx.doi.org/10.1093/bioinformatics/btl161. 39

[112] Applegate KE, Tello R, Ying J. Hypothesis testing III: counts and medians. Radiology. 2003 Sep;228(3):603–608. Available from: http://dx.doi.org/10.1148/radiol.2283021330. 39

[113] Wu B, Guan Z, Zhao H. Parametric and nonparametric FDR estimation revisited. Biometrics. 2006 Sep;62(3):735–744. Available from: http://dx.doi.org/10.1111/j.1541-0420.2006.00531.x. 39, 40

[114] Ge Y, Dudoit S, Speed TP. Resampling-based multiple testing for microarray data analysis (electronic edition). Department of Statistics, Division of Biostatistics, University of California, Berkeley Division of Genetics and Bioinformatics, The Walter and Eliza Hall Institute of Medical Research, Australia; 2003. 39, 40

[115] Storey JD, Tibshirani R. Statistical significance for genomewide studies. Proc Natl Acad Sci U S A. 2003 Aug;100(16):9440–9445. Available from: http://dx.doi.org/10.1073/pnas.1530509100. 39, 40, 67

[116] Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Statist Soc. 1995;B 57:289–300. 40

[117] Jackson J. A Users Guide to Principal Components. Wiley & Sons; 1991. 40

[118] Alter O, Brown PO, Botstein D. Singular value decomposition for genome-wide expression data processing and modeling. Proc Natl Acad Sci U S A. 2000 Aug;97(18):10101–10106. 40

[119] Raychaudhuri S, Stuart JM, Altman RB. Principal components analysis to summarize microarray experiments: application to sporulation time series. Pac Symp Biocomput. 2000;p. 455–466. 40

[120] Dupuy A, Simon RM. Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting. J Natl Cancer Inst. 2007 Jan;99(2):147–157. Available from: http://dx.doi.org/10.1093/jnci/djk018. 41, 42, 43, 44, 45, 53, 70

[121] D'haeseleer P. How does gene expression clustering work? Nat Biotechnol. 2005 Dec;23(12):1499–1501. Available from: http://dx.doi.org/10.1038/nbt1205-1499. 41

[122] Dougherty ER, Barrera J, Brun M, Kim S, Cesar RM, Chen Y, et al. Inference from clustering with application to gene-expression microarrays. J Comput Biol. 2002;9(1):105–126. Available from: http://dx.doi.org/10.1089/10665270252833217. 41

[123] Molinaro AM, Simon R, Pfeiffer RM. Prediction error estimation: a comparison of resampling methods. Bioinformatics. 2005 Aug;21(15):3301–3307. Available from: http://dx.doi.org/10.1093/bioinformatics/bti499. 42, 43

[124] Ancona N, Maglietta R, Piepoli A, D'Addabbo A, Cotugno R, Savino M, et al. On the statistical assessment of classifiers using DNA microarray data. BMC Bioinformatics. 2006;7:387. Available from: http://dx.doi.org/10.1186/1471-2105-7-387. 42, 44, 45

[125] Solberg HE. Discriminant analysis. CRC Crit Rev Clin Lab Sci. 1978;9(3):209–242. 42

[126] Jaeger J, Sengupta R, Ruzzo WL. Improved gene selection for classification of microarrays. Pac Symp Biocomput. 2003;p. 53–64. 42

[127] Leea JW, Leea JB, Parkb M, Songa SH. An extensive comparison of recent classification tools applied to microarray data. Computational Statistics and Data Analysis. 2005;48:869–885. 42

[128] Ji S, Ye J. Generalized linear discriminant analysis: a unified framework and efficient model selection. IEEE Trans Neural Netw. 2008 Oct;19(10):1768–1782. Available from: http://dx.doi.org/10.1109/TNN.2008.2002078. 42

[129] Schroeder A, Mueller O, Stocker S, Salowsky R, Leiber M, Gassmann M, et al. The RIN: an RNA integrity number for assigning integrity values to RNA measurements. BMC Mol Biol. 2006;7:3. Available from: http://dx.doi.org/10.1186/1471-2199-7-3. 47, 48, 49

[130] Sambrook J, Fritsch E, Maniatis T. Molecular Cloning, a laboratory manual. 2nd ed. Cold Spring Harbor Laboratory Press; 1989. 47

[131] Auer H, Lyianarachchi S, Newsom D, Klisovic MI, Marcucci G, Marcucci U, et al. Chipping away at the chip bias: RNA degradation in microarray analysis. Nat Genet. 2003 Dec;35(4):292–293. Available from: http://dx.doi.org/10.1038/ng1203-292. 47

[132] Micke P, Ohshima M, Tahmasebpoor S, Ren ZP, Ostman A, Pontén F, et al. Biobanking of fresh frozen tissue: RNA is stable in nonfixed surgical specimens. Lab Invest. 2006 Feb;86(2):202–211. Available from: http://dx.doi.org/10.1038/labinvest.3700372. 49, 101

[133] Mutter GL, Zahrieh D, Liu C, Neuberg D, Finkelstein D, Baker HE, et al. Comparison of frozen and RNALater solid tissue storage methods for use in RNA expression microarrays. BMC Genomics. 2004 Nov;5(1):88. Available from: http://dx.doi.org/10.1186/1471-2164-5-88. 50, 101

[134] Scicchitano MS, Dalmas DA, Bertiaux MA, Anderson SM, Turner LR, Thomas RA, et al. Preliminary comparison of quantity, quality, and microarray performance of RNA extracted from formalin-fixed, paraffin-embedded, and unfixed frozen tissue samples. J Histochem Cytochem. 2006 Nov;54(11):1229–1237. Available from: http://dx.doi.org/10.1369/jhc.6A6999.2006. 50

[135] Blackhall FH, Pintilie M, Wigle DA, Jurisica I, Liu N, Radulovich N, et al. Stability and heterogeneity of expression profiles in lung cancer specimens harvested following surgical resection. Neoplasia. 2004;6(6):761–767. Available from: http://dx.doi.org/10.1593/neo.04301. 50

[136] Strand C, Enell J, Hedenfalk I, Fernö M. RNA quality in frozen breast cancer samples and the influence on gene expression analysis–a comparison of three evaluation methods using microcapillary electrophoresis traces. BMC Mol Biol. 2007;8:38. Available from: http://dx.doi.org/10.1186/1471-2199-8-38. 50

[137] Nutt CL, Mani DR, Betensky RA, Tamayo P, Cairncross JG, Ladd C, et al. Gene expression-based classification of malignant gliomas correlates better with survival than histological classification. Cancer Res. 2003 Apr;63(7):1602–1607. 51

[138] Kim S, Dougherty ER, Shmulevich I, Hess KR, Hamilton SR, Trent JM, et al. Identification of combination gene sets for glioma classification. Mol Cancer Ther. 2002 Nov;1(13):1229–1236. 52

[139] Godard S, Getz G, Delorenzi M, Farmer P, Kobayashi H, Desbaillets I, et al. Classification of human astrocytic gliomas on the basis of gene expression: a correlated group of genes with angiogenic activity emerges as a strong predictor of subtypes. Cancer Res. 2003 Oct;63(20):6613–6625. 52

[140] Shirahata M, Iwao-Koizumi K, Saito S, Ueno N, Oda M, Hashimoto N, et al. Gene expression-based molecular diagnostic system for malignant gliomas is superior to histological diagnosis. Clin Cancer Res. 2007 Dec;13(24):7341–7356. Available from: http://dx.doi.org/10.1158/1078-0432.CCR-06-2789. 52

[141] Pusztai L, Ayers M, Stec J, Hortobágyi GN. Clinical application of cDNA microarrays in oncology. Oncologist. 2003;8(3):252–258. 53, 54

[142] Consortium MAQC, Shi L, Reid LH, Jones WD, Shippy R, Warrington JA, et al. The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. Nat Biotechnol. 2006 Sep;24(9):1151–1161. Available from: http://dx.doi.org/10.1038/nbt1239. 53

[143] Chen JJ, Hsueh HM, Delongchamp RR, Lin CJ, Tsai CA. Reproducibility of microarray data: a further analysis of microarray quality control (MAQC) data. BMC Bioinformatics. 2007;8:412. Available from: http://dx.doi.org/10.1186/1471-2105-8-412. 53

[144] Shi L, Perkins RG, Fang H, Tong W. Reproducible and reliable microarray results through quality control: good laboratory proficiency and appropriate data analysis practices are essential. Curr Opin Biotechnol. 2008 Feb;19(1):10–18. Available from: http://dx.doi.org/10.1016/j.copbio.2007.11.003. 53

[145] Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, et al. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. Nat Genet. 2001 Dec;29(4):365–371. Available from: http://dx.doi.org/10.1038/ng1201-365. 53

[146] Wakabayashi T, Natsume A, Hashizume Y, Fujii M, Mizuno M, Yoshida J. A phase I clinical trial of interferon-beta gene therapy for high-grade glioma: novel findings from gene expression profiling and autopsy. J Gene Med. 2008 Apr;10(4):329–339. Available from: http://dx.doi.org/10.1002/jgm.1160. 54

[147] Jarboe JS, Johnson KR, Choi Y, Lonser RR, Park JK. Expression of interleukin-13 receptor alpha2 in glioblastoma multiforme: implications for targeted therapies. Cancer Res. 2007 Sep;67(17):7983–7986. Available from: http://dx.doi.org/10.1158/0008-5472.CAN-07-1493. 54

[148] de Azambuja E, Cardoso F, Meirsman L, Straehle C, Dolci S, Vantongelen K, et al. [The new generation of breast cancer clinical trials: the right drug for the right target]. Bull Cancer. 2008 Mar;95(3):352–357. Available from: http://dx.doi.org/10.1684/bdc.2008.0587. 54

[149] Natowicz R, Incitti R, Horta EG, Charles B, Guinot P, Yan K, et al. Prediction of the outcome of preoperative chemotherapy in breast cancer by DNA probes that convey information on both complete and non complete responses. BMC Bioinformatics. 2008 Mar;9(1):149. Available from: http://dx.doi.org/10.1186/1471-2105-9-149. 54

[150] Hegde PS, Rusnak D, Bertiaux M, Alligood K, Strum J, Gagnon R, et al. Delineation of molecular mechanisms of sensitivity to lapatinib in breast cancer cell lines using global gene expression profiles. Mol Cancer Ther. 2007 May;6(5):1629–1640. Available from: http://dx.doi.org/10.1158/1535-7163.MCT-05-0399. 54

[151] Brewer M, Kirkpatrick ND, Wharton JT, Wang J, Hatch K, Auersperg N, et al. 4-HPR modulates gene expression in ovarian cells. Int J Cancer. 2006 Sep;119(5):1005–1013. Available from: http://dx.doi.org/10.1002/ijc.21797. 54

[152] van de Vijver M. Gene-expression profiling and the future of adjuvant therapy. Oncologist. 2005 Oct;10 Suppl 2:30–34. Available from: http://dx.doi.org/10.1634/theoncologist.10-90002-30. 54

[153] Schuetz CS, Bonin M, Clare SE, Nieselt K, Sotlar K, Walter M, et al. Progression-specific genes identified by expression profiling of matched ductal carcinomas in situ

and invasive breast tumors, combining laser capture microdissection and oligonucleotide microarray analysis. Cancer Res. 2006 May;66(10):5278–5286. Available from: http://dx.doi.org/10.1158/0008-5472.CAN-05-4610. 54

[154] González-Vélez H, Mier M, Julià-Sapé M, Arvanitis TN, García-Gómez JM, Robles M, et al. HealthAgents: distributed multi-agent brain tumor diagnosis and prognosis. Appl Intell. 2007;p. 10.1007/s10489–007–0085–8. 55

[155] Qiagen. RNeasy Midi/Maxi Handbook (electronic version); 2006. 62

[156] Ambion. mirVanaTM miRNA Isolation Kit, Instruction Manual;. 63

[157] Chomczynski P, Sacchi N. Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction. Anal Biochem. 1987 Apr;162:156–159. 63

[158] Dennis G, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, et al. DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome Biol. 2003;4(5):P3. 67

[159] Livak KJ, Schmittgen TD. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. Methods. 2001 Dec;25(4):402–408. Available from: http://dx.doi.org/10.1006/meth.2001.1262. 68

[160] Benjamini Y, Yekutieli D. The control of the false discovery rate in multiple testing under dependency. Ann Stat. 2001;29(4):1165–1188. 70

[161] Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology. 1982 Apr;143(1):29–36. 71

[162] Tate AR, Underwood J, Acosta DM, Julià-Sapé M, Majós C, Moreno-Torres A, et al. Development of a decision support system for diagnosis and grading of brain tumours using in vivo magnetic resonance single voxel spectra. NMR Biomed. 2006 Jun;19(4):411–434. Available from: http://dx.doi.org/10.1002/nbm.1016. 72, 130

[163] Valverde Saubí D. Estudi de possibles marcadors espectroscòpics per RMN: de proliferació en extractes cel.lulars i de tipus i grau tumoral en biòpsies de tumors cerebrals humans. Universitat Autònoma de Barcelona; 2008. 72, 129

[164] Quintero M, Cabañas ME, Arús C. A possible cellular explanation for the NMR-visible mobile lipid (ML) changes in cultured C6 glioma cells with growth. Biochim Biophys Acta. 2007 Jan;1771(1):31–44. Available from: http://dx.doi.org/10.1016/j.bbalip.2006.10.003. 73

[165] Simões RV, García-Martín ML, Cerdán S, Arús C. Perturbation of mouse glioma MRS pattern by induced acute hyperglycemia. NMR Biomed. 2008 Mar;21(3):251–264. Available from: http://dx.doi.org/10.1002/nbm.1188. 74

[166] Valverde D, Quintero MR, Candiota AP, Badiella L, Cabañas ME, Arús C. Analysis of the changes in the 1H NMR spectral pattern of perchloric acid extracts of C6 cells with growth. NMR Biomed. 2006 Apr;19(2):223–230. Available from: http://dx.doi.org/10.1002/nbm.1024. 76

[167] Mischel PS, Shai R, Shi T, Horvath S, Lu KV, Choe G, et al. Identification of molecular subtypes of glioblastoma by gene expression profiling. Oncogene. 2003 Apr;22(15):2361–2373. Available from: http://dx.doi.org/10.1038/sj.onc.1206344. 91, 124

[168] Watson MA, Gutmann DH, Peterson K, Chicoine MR, Kleinschmidt-DeMasters BK, Brown HG, et al. Molecular characterization of human meningiomas by gene expression profiling using high-density oligonucleotide microarrays. Am J Pathol. 2002 Aug;161(2):665–672. 91

[169] Wrobel G, Roerig P, Kokocinski F, Neben K, Hahn M, Reifenberger G, et al. Microarray-based gene expression profiling of benign, atypical and anaplastic meningiomas identifies novel genes associated with meningioma progression. Int J Cancer. 2005 Mar;114(2):249–256. Available from: http://dx.doi.org/10.1002/ijc.20733. 91, 93

[170] Eng LF, Ghirnikar RS, Lee YL. Glial fibrillary acidic protein: GFAP-thirty-one years (1969-2000). Neurochem Res. 2000 Oct;25(9-10):1439–1451. 92

[171] Baba H, Nakahira K, Morita N, Tanaka F, Akita H, Ikenaka K. GFAP gene expression during development of astrocyte. Dev Neurosci. 1997;19(1):49–57. 92

[172] Lee J, Kotliarova S, Kotliarov Y, Li A, Su Q, Donin NM, et al. Tumor stem cells derived from glioblastomas cultured in bFGF and EGF more closely mirror the phenotype and genotype of primary tumors than do serum-cultured cell lines. Cancer Cell. 2006 May;9(5):391–403. Available from: http://dx.doi.org/10.1016/j.ccr.2006.03.030. 92

[173] Zhou R, Skalli O. TGF-alpha differentially regulates GFAP, vimentin, and nestin gene expression in U-373 MG glioblastoma cells: correlation with cell shape and motility. Exp Cell Res. 2000 Feb;254(2):269–278. Available from: http://dx.doi.org/10.1006/excr.1999.4762. 92

[174] Tasheva ES, Klocke B, Conrad GW. Analysis of transcriptional regulation of the small leucine rich proteoglycans. Mol Vis. 2004 Oct;10:758–772. 93

[175] Ge G, Seo NS, Liang X, Hopkins DR, Höök M, Greenspan DS. Bone morphogenetic protein-1/tolloid-related metalloproteinases process osteoglycin and enhance its ability to regulate collagen fibrillogenesis. J Biol Chem. 2004 Oct;279(40):41626–41633. Available from: http://dx.doi.org/10.1074/jbc.M406630200. 93

[176] Baharvand H, Heidari M, Ebrahimi M, Valadbeigi T, Salekdeh GH. Proteomic analysis of epithelium-denuded human amniotic membrane as a limbal stem cell niche. Mol Vis. 2007;13:1711–1721. 93

[177] Droguett R, Cabello-Verrugio C, Riquelme C, Brandan E. Extracellular proteoglycans modify TGF-beta bio-availability attenuating its signaling during skeletal muscle differentiation. Matrix Biol. 2006 Aug;25(6):332–341. Available from: http://dx.doi.org/10.1016/j.matbio.2006.04.004. 93

[178] Oklü R, Hesketh R. The latent transforming growth factor beta binding protein (LTBP) family. Biochem J. 2000 Dec;352 Pt 3:601–610. 93

[179] Purnapatre K, Khattar SK, Saini KS. Cytochrome P450s in the development of target-based anticancer drugs. Cancer Lett. 2008 Jan;259(1):1–15. Available from: http://dx.doi.org/10.1016/j.canlet.2007.10.024. 94

[180] Elexpuru-Camiruaga J, Buxton N, Kandula V, Dias PS, Campbell D, McIntosh J, et al. Susceptibility to astrocytoma and meningioma: influence of allelism at glutathione S-transferase (GSTT1 and GSTM1) and cytochrome P-450 (CYP2D6) loci. Cancer Res. 1995 Oct;55(19):4237–4239. 94

[181] Khalid MH, Tokunaga Y, Caputy AJ, Walters E. Inhibition of tumor growth and prolonged survival of rats with intracranial gliomas following administration of clotrimazole. J Neurosurg. 2005 Jul;103(1):79–86. 94

[182] Wundrack I, Meese E, Müllenbach R, Blin N. Debrisoquine hydroxylase gene polymorphism in meningioma. Acta Neuropathol. 1994;88(5):472–474. 94

[183] Fiore G, Cristo CD, Monti G, Amoresano A, Columbano L, Pucci P, et al. Tubulin nitration in human gliomas. Neurosci Lett. 2006 Feb;394(1):57–62. Available from: http://dx.doi.org/10.1016/j.neulet.2005.10.011. 94

[184] Kim SJ, Chung TW, Jin UH, Suh SJ, Lee YC, Kim CH. Molecular mechanisms involved in transcriptional activation of the human Sia-alpha2,3-Gal-beta1,4-GlcNAc-R:alpha2,8-sialyltransferase (hST8Sia III) gene induced by KCl in human glioblastoma cells. Biochem Biophys Res Commun. 2006 Jun;344(4):1057–1064. Available from: http://dx.doi.org/10.1016/j.bbrc.2006.04.004. 94

[185] Panagopoulos AT, Lancellotti CLP, Veiga JCE, de Aguiar PHP, Colquhoun A. Expression of cell adhesion proteins and proteins related to angiogenesis and fatty acid metabolism in benign, atypical, and anaplastic meningiomas. J Neurooncol. 2008 Aug;89(1):73–87. Available from: http://dx.doi.org/10.1007/s11060-008-9588-3. 94

[186] Shimada S, Ishizawa K, Hirose T. Expression of E-cadherin and catenins in meningioma: ubiquitous expression and its irrelevance to malignancy. Pathol Int. 2005 Jan;55(1):1–7. Available from: http://dx.doi.org/10.1111/j.1440-1827.2005.01786.x. 94

[187] Barami K, Lewis-Tuffin L, Anastasiadis PZ. The role of cadherins and catenins in gliomagenesis. Neurosurg Focus. 2006;21(4):E13. 94

[188] Utsuki S, Sato Y, Oka H, Tsuchiya B, Suzuki S, Fujii K. Relationship between the expression of E-, N-cadherins and beta-catenin and tumor grade in astrocytomas. J Neurooncol. 2002 May;57(3):187–192. 94

[189] Miles AT, Hawksworth GM, Beattie JH, Rodilla V. Induction, regulation, degradation, and biological significance of mammalian metallothioneins. Crit Rev Biochem Mol Biol. 2000;35(1):35–70. 94

[190] Maier H, Jones C, Jasani B, Ofner D, Zelger B, Schmid KW, et al. Metallothionein overexpression in human brain tumours. Acta Neuropathol. 1997 Dec;94(6):599–604. 94

[191] Altenberg B, Greulich KO. Genes of glycolysis are ubiquitously overexpressed in 24 cancer classes. Genomics. 2004 Dec;84(6):1014–1020. Available from: http://dx.doi.org/10.1016/j.ygeno.2004.08.010. 94

[192] Kumanishi T, Watabe K, Washiyama K. An immunohistochemical study of aldolase C in normal and neoplastic nervous tissues. Acta Neuropathol. 1985;67(3-4):309–314. 94

[193] Gatenby RA, Gillies RJ. Why do cancers have high aerobic glycolysis? Nat Rev Cancer. 2004 Nov;4(11):891–899. Available from: http://dx.doi.org/10.1038/nrc1478. 95

[194] Gatenby RA, Gillies RJ. Glycolysis in cancer: a potential target for therapy. Int J Biochem Cell Biol. 2007;39(7-8):1358–1366. Available from: http://dx.doi.org/10.1016/j.biocel.2007.03.021. 95

[195] Li A, Walling J, Ahn S, Kotliarov Y, Su Q, Quezado M, et al. Unsupervised analysis of transcriptomic profiles reveals six glioma subtypes. Cancer Res. 2009 Mar;69(5):2091–2099. Available from: http://dx.doi.org/10.1158/0008-5472.CAN-08-2100. 124

[196] Liang Y, Diehn M, Watson N, Bollen AW, Aldape KD, Nicholas MK, et al. Gene expression profiling reveals molecularly and clinically distinct subtypes of glioblastoma multiforme. Proc Natl Acad Sci U S A. 2005 Apr;102(16):5814–5819. Available from: http://dx.doi.org/10.1073/pnas.0402870102. 124

[197] Lee Y, Scheck AC, Cloughesy TF, Lai A, Dong J, Farooqi HK, et al. Gene expression analysis of glioblastomas identifies the major molecular basis for the prognostic benefit of younger age. BMC Med Genomics. 2008;1:52. Available from: http://dx.doi.org/10.1186/1755-8794-1-52. 124, 140

[198] Candiota AP, Majós C, Julià-Sapé M, Cabañas M, Mercadal G, Acebes JJ, et al. Book of Abstracts ESMRMB 2005 22nd Annual Scientific Meeting Basle/CH, Sept. 15-18, 2005 EPOSTM Posters. MAGMA. 2005 Sep;18 Suppl 7:176–307. Available from: http://dx.doi.org/10.1007/s10334-005-0002-2. 130

[199] Chen JH, Wu YV, DeCarolis P, O'Connor R, Somberg CJ, Singer S. Resolution of creatine and phosphocreatine 1H signals in isolated human skeletal muscle using HR-MAS 1H NMR. Magn Reson Med. 2008 Jun;59(6):1221–1224. Available from: http://dx.doi.org/10.1002/mrm.21604. 130

[200] Zhang K, Weinberg JM, Venkatachalam MA, Dong Z. Glycine protection of PC-12 cells against injury by ATP-depletion. Neurochem Res. 2003 Jun;28(6):893–901. 130

[201] Beelman CA, Parker R. Degradation of mRNA in eukaryotes. Cell. 1995 Apr;81(2):179–183. 136

[202] Moore MJ. Nuclear RNA turnover. Cell. 2002 Feb;108(4):431–434. 136

[203] Jing Q, Huang S, Guth S, Zarubin T, Motoyama A, Chen J, et al. Involvement of microRNA in AU-rich element-mediated mRNA instability. Cell. 2005 Mar;120(5):623–634. Available from: http://dx.doi.org/10.1016/j.cell.2004.12.038. 136

[204] Bolognani F, Perrone-Bizzozero NI. RNA-protein interactions and control of mRNA stability in neurons. J Neurosci Res. 2008 Feb;86(3):481–489. Available from: http://dx.doi.org/10.1002/jnr.21473. 136

Note: The number(s) that appear(s) at the end of the reference correspond to the page(s) of citation.

Chapter 6

ANNEXES

A-1 Collection of clinical and histopathological data in eTUMOUR

Definitive version , 20041008

Project Nº – *FP6-2002-LIFESCIHEALTH 503094*

Project Full Title – *WEB ACCESSIBLE MR DECISION SUPPORT SYSTEM FOR BRAIN TUMOUR DIAGNOSIS AND PROGNOSIS, INCORPORATING IN VIVO AND EX VIVO GENOMIC AND METABOLOMIC DATA*

ACRONYM – *eTUMOUR*

Dissemination Level – *PU*

Deliverable Number – *D3 (D2.2)*

Contractual date of delivery – *MONTH 6*

Actual date of delivery – *MONTH 9 (OCTOBER 2004)*

Title of the deliverable – *SPECIFICATION OF WHAT CLINICAL AND HISTOPATHOLOGICAL DATA MUST BE OBTAINED*

Work package contributing to the deliverable – *WP2*

Nature of the deliverable – *R*

Authors – *M Margarita Julià-Sapé and Carles Arús. Partner 2, UAB*

ABSTRACT

The specification of what clinical and histopathological data must be obtained during the eTUMOUR project serves three purposes: First, to ensure that data from all patients recruited for the project are collected in a standard and compatible way; Second, to ensure that the minimum amount of information needed to validate and use data collected for each patient case will be available; And third, to aid in the logistics of the clinical data quality checking effort for each patient.

For this, documents on advised informed consent documentation, clinical data to be recorded, histopathological and clinical data validation protocols have been discussed and agreed.

INDEX

ACRONYM LIST

CDVP	Clinical Data Validation Protocol
CQCD	Committee for Quality Control of Data
CRF	Clinical Record Form
CSC	Clinical Subcommittee of CQCD
DQCE	Data Quality Checking Effort
DSS	Decision Support System
eTDB	e-Tumour Database
GUI	Graphical User Interface
HPD	Histopathology Diagnosis
HR MAS	High Resolution Magic Angle Spinning
HVP	Histopathological Validation Protocol
IF	Informed Consent Form
KPS	Karnofsky Performance Score
MRI	Magnetic Resonance Imaging
MRS	Magnetic Resonance Spectroscopy
P11, KUL	Partner 11, Katholieke Universiteit Leuven Research & Development
P15, UPVLC	Partner 15, Universidad Politécnica de Valencia
P18, BU	Partner 18, Institute of Child Health, University of Birmingham
P2, UAB	Partner 2, Universitat Autònoma de Barcelona
P3, SGHMS	Partner 3, St George's Hospital Medical School
P4, UMCN	Partner 4, University Medical Center Nijmegen

EXECUTIVE SUMMARY

The specification of what clinical and histopathological data must be obtained will be used to ensure uniform pathology type and grade assignment for cases to be used in classifier development to be incorporated in the Graphical User Interface (GUI) of a Decision Support System (DSS). In order to achieve this, two other requirements have to be met first:

- To have available a standard way of collecting and storing clinical and histopathological data.
- Once data is stored, a Data Quality Checking Effort (DQCE) has to be performed in order to ensure that each case will be usable in the eTUMOUR project from the point of view of medical inclusion criteria.

Essentially, when each patient undergoes a brain scan at one of the participating clinical institutions because the presence of a brain tumour is suspected, a full Ethics Committee Approval and Informed Consent (IF) form will be available. **Annex 1a** shows a guidance and IF consent forms and **Annex 1b** shows a list of minimal concepts to be included. The patient or a responsible person will be asked whether he/she is willing to participate in the eTUMOUR project. If an IF form is signed, when the patient undergoes the surgical procedure to diagnose and/or to remove the brain tumour, a sample of the abnormal tissue will be obtained for histopathological analysis in order to analyse which tumour type the patient has. When possible, part of the excised tumour will also be frozen immediately upon excision, labelled and stored in liquid N_2 for subsequent *ex vivo* analysis by the HR MAS [1] and microarray [2] techniques. The eTUMOUR clinicians who are treating the

patient will collect the clinical information from each patient in order to perform DQCE and ensure that the case is usable for inclusion into the final eTUMOUR DSS.

FULL DESCRIPTION OF DELIVERABLE CONTENT

INTRODUCTION

The lifespan of the European population is increasing and accordingly, diseases that become prevalent in old age, such as brain tumours, will afflict a larger percentage of this population. Brain tumours do not have a lifestyle-associated aetiology hence prevention is not yet possible. Gold standard diagnosis is based on histological analysis of tumour biopsies, which is an invasive procedure that carries risks. Additionally, in slowly evolving tumours (e.g. pilocytic astrocytoma in children) repeated biopsy may not be advisable at all. Diagnosis by magnetic resonance imaging (MRI) is non-invasive, but only achieves 60-90% accuracy depending on the tumour type and grade [3] , therefore can only replace biopsy for particular cases. Magnetic resonance spectroscopy (MRS) provides a non-invasive method to obtain a profile of the biochemical constituents of the tumour and has been shown to improve the accuracy of diagnosis in specific instances [4] . However, MR spectra are complex and require skilled interpretation, for these reasons routine clinical use of MRS is still low. Thus a decision support system that should facilitate the uptake of MRS by clinicians, by providing an automated classification of tumour MR spectra, has already been developed (INTERPRET [5]).

Tumour tissue is generally heterogeneous and there are a large number of different tumour types and grades. Thus, in order to develop automated classification methods that are comprehensive, data from several hospitals must be combined to fully characterise the variability of tumour spectra. Furthermore, the robustness of the classification method must then be validated in a real clinical setting. Furthermore,

the possibility of phenotyping tumours with DNA microarrays may create new subtypes of tumours on molecular grounds. Moreover, the extensive and more precise metabolic analysis of tumours by MRS at high fields (≥ 9.4 T) from tissues (ex vivo) can allow a better understanding of the tumour biochemistry and may also refine the classification of brain tumours. It will then become necessary to correlate MRS data with the tumour gene and metabolic expression profile thus requiring a large database of MR spectra -either in vivo and ex vivo- and microarray analyses. Finally, it is important to look for correlations of patient survival or more precisely, patient performance status, with MRS characteristics, to assess whether there are better prognostic indicators than the current grading system.

METHODS

Clinical personnel are responsible for all matters relating to clinical data collection, informed consent form preparation and collection (**Annexes 1A** and **1B**), eTUMOUR database (eTDB) data upload and validation of clinical and histopathological data. This information is collected locally at each participating centre and stored in the eTDB from the same location. The eTDB is the source from which the members of the Committee for Quality Control of Data (CQCD), through its Clinical Subcommittee (CSC), will perform the DQCE. Figure 1 summarises the general path envisioned for clinical data storage and CQCD.

For this reason the following documents and protocols have to be developed:

1. Histopathological data validation protocol (HVP). The purpose of the histopathological validation protocol is to ensure a standard processing protocol and diagnosis for all biopsy samples of tumours.

2. Clinical data validation protocol (CDVP). The purpose of the clinical data validation protocol is to ensure that inclusion criteria for patient clinical and histopathological data to be used in DSS development are met. This protocol is to be applied by CSC members to all incoming cases.

3. Clinical record form (CRF). The purpose of the CRF is to ensure that raw clinical data and results of HVP and CDVP are recorded in a standard, secure and accessible way to CQDC and CSC members and DSS developers. An associated purpose of the CRF is to be an instrument allowing CDVP, HVP and in general, the full DQCE process.

Once all incoming data has been entered into the eTDB, quality control checks on data consistency will be performed, essentially as in the INTERPRET project (See **Annex 2** and [6]). However, WP6 will deal with this matter in detail.

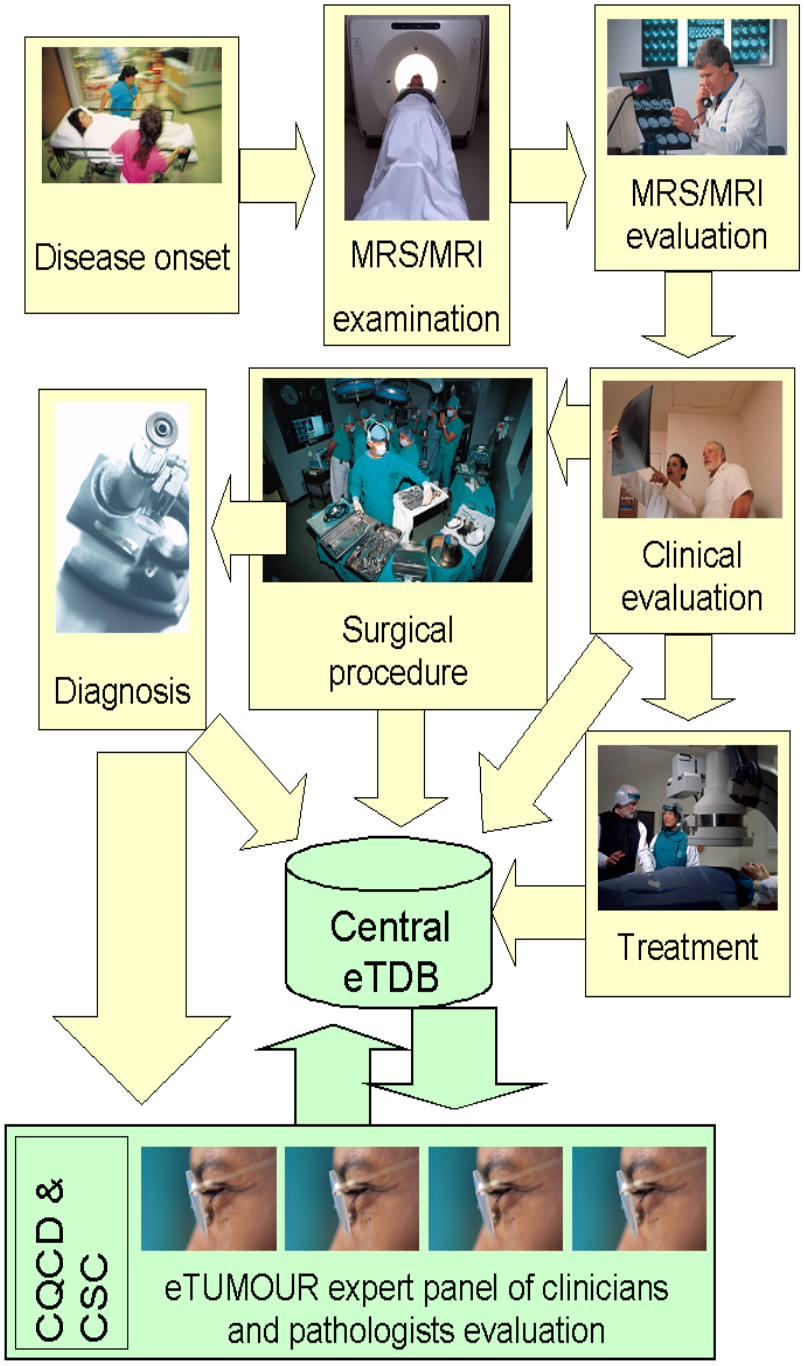


***Figure 1**: Envisioned path for clinical data storage into the eTDB. **Yellow**: originating hospital. **Green**: eTUMOUR*

RESULTS

1. HVP

Of each case, at least one but preferably 5 unstained slides of each paraffin block of the brain tumour have to be sent to Dr. Pieter Wesseling. Based on Haematoxylin-Eosin staining, and when necessary, additional stainings. Histological diagnosis will be made according to the WHO-2000 [7] classification by Dr. Pieter Wesseling (head of CSC) and a second neuropathologist, in case of discrepancy also by a third neuropathologist. Second and third neuropathologists will be Dr. Martin Lammens and Dr. Max Kros, from The Netherlands. If additional immunohistochemistry has to be performed in order to reach a diagnosis, it will be done on the unstained slides that have been received and had not been used for previous Haematoxilin-Eosin staining. The final consensus diagnosis (or, in case of lack of consensus, alternative diagnosis/diagnoses) will be included in the eTDB.

2. CDVP

Electronic data contained in the eTDB relevant for clinical data review (clinical, radiological, histopathological (and when available, molecular) data will be sent to members of CQCD (CSC subsection), including a neurologist, a neurosurgeon, a neuroradiologist, a paediatric oncologist and a neuropathologist. All data will be evaluated by each of the participating members of this group for errors, missing significant information, etc. When all members agree on the absence of errors and missing significant information the case will be flagged in the eTDB as “validated”. If data on the case do not meet the minimum quality standards, the case will be flagged as “not validated”, indicating that it is not usable for further study. Minimum quality standards are to be defined and stored in written format in CQCD meeting minutes. Teleconference by members of the CQCD (CSC subsection) about problematic

cases will also be performed in order to reach a decision on if and how difficult cases can still be used and flagged as “validated”.

3. CRF

Table 1 depicts the list of fields and the possible values defined for each of them.

FIELD NUMBER	FIELD NAME	WAS ALREADY AN INTERPRET [5] FIELD	IS A NEW eTUMOUR FIELD	POSSIBLE VALUES	COMMENTS
1	AGE	X		INTEGER	
2	DATE OF BIRTH		X	DD / MM / YYYY	
3	SEX	X		M / F / NA	
4	DATE OF SPECTROSCOPY	X		DD / MM / YYYY	
5	WEEKS SINCE FIRST SYMPTOM	X		INTEGER	
6	PRESENTING SYMPTOM: EPILEPTIC FIT	X		YES / NO / NA	
7	PRESENTING SYMPTOM: NEUROLOGICAL DEFICIT	X		YES / NO / NA	
8	PRESENTING SYMPTOM: COMA	X		YES / NO / NA	
9	MEDICATION AT TIME OF SPECTROSCOPY: STEROIDS	X		YES / NO / NA	
10	MEDICATION AT TIME OF SPECTROSCOPY: ANTICONVULSANTS	X		YES / NO / NA	
11	MEDICATION AT TIME OF SPECTROSCOPY: GADOLINIUM	X		YES / NO / NA	
12	MEDICATION AT TIME OF SPECTROSCOPY: ANAESTETHIC AGENTS	X		YES / NO / NA	
13	MEDICATION AT TIME OF SPECTROSCOPY: MANNITOL	X		YES / NO / NA	
14	BLEED INTO TUMOUR	X		YES / NO / NA	
15	TUMOUR LOCATION	X		TECTAL PLATE / TEMPORAL /MIDBRAIN / MEDULLA / PONS / PINEAL REGION / OPTIC CHIASM / HYPOTHALAMUS / SUPRASELLAR REGION / VENTRICULAR / FRONTAL / PARIETAL / OCCIPITAL / BASAL GANGLIA / CEREBELLUM ## LEFT / RIGHT	POSSIBLE VALUES MODIFIED FROM FREE TEXT TO MENU
16	CALCIFICATIONS		X	YES / NO / NA	
17	NUMBER OF DETECTED LESIONS		X	INTEGER	
18	TUMOUR SIZE	X		nn x nn x nn (mm) ## nn OF BIGGER AXIS FOR GLIAL TUMOURS (mm)	
19	RADIOLOGICAL DIAGNOSIS 1	X		FREE TEXT	

20	*CONFIDENCE RATING FOR RADIOLOGICAL DIAGNOSIS 1*		X	*0 / 1 / 2 / 3 / 4*	*0: NOT CONFIDENT AT ALL, 4: TOTALLY CONFIDENT*
21	*RADIOLOGICAL DIAGNOSIS 2*		X	*FREE TEXT*	*P2 UAB SUGGESTION: TO ACCEPT UP TO 3 RADIOLOGICAL DIAGNOSTICS*
22	*CONFIDENCE RATING FOR RADIOLOGICAL DIAGNOSIS 2*		X	*0 / 1 / 2 / 3 / 4*	*0: NOT CONFIDENT AT ALL, 4: TOTALLY CONFIDENT*
23	*RADIOLOGICAL DIAGNOSIS 3*		X	*FREE TEXT*	*SAME AS BEFORE*
24	*CONFIDENCE RATING FOR RADIOLOGICAL DIAGNOSIS 3*		X	*0 / 1 / 2 / 3 / 4*	*0: NOT CONFIDENT AT ALL, 4: TOTALLY CONFIDENT*
25	*SITE OF OPERATION*	X		*FREE TEXT*	
26	*TUMOUR REMOVAL*	X	*OPTIONS MODIFIED*	*COMPLETE MACROSCOPIC RESECTION / PARTIAL MACROSCOPIC RESECTION / OPEN BIOPSY / STEREOTACTIC BIOPSY*	*IN ORDER TO SPECIFY THAT DEGREE OF RESECTION IS FROM THE MACROSCOPIC POINT OF VIEW. AND TO SPECIFY TYPE OF BIOPSY IN CASES WHERE THERE IS NO RESECTION*
	SUBTOTAL TUMOUR REMOVAL	*REMOVE*		*YES / NO / NA*	
	STEREOTACTIC BIOPSY	REMOVE	*MERGED WITH FIELD 26*	*YES/NO/NA*	*FOLLOWING P3 SGHMS COMMENTS: IF THE BIOPSY WAS OBTAINED FROM A STEREOTACTIC BIOPSY OR FROM TUMOUR EXCISION*
27	*DATE OF BIOPSY*	X		*DD/MM/YYYY*	
28	*PARAFFIN SECTION WHO CLASSIFICATION (ORIGINATING PATHOLOGIST)*		X	*WHO DIAGNOSES MENU + FREE TEXT*	*RELATED TO PROTOCOL FOR HISTOLOGY VALIDATION. FREE TEXT SUGGESTED BY P2 UAB,, AS MIGHT NOT NECESSARILY BE eTUMOUR PATHOLOGISTS*
29	*PARAFFIN SECTION WHO CLASSIFICATION (PATHOLOGIST A)*		X	*WHO DIAGNOSES MENU*	*RELATED TO PROTOCOL FOR HISTOLOGY VALIDATION.*
30	*PARAFFIN SECTION WHO CLASSIFICATION (PATHOLOGIST B)*		X	*WHO DIAGNOSES MENU*	*RELATED TO PROTOCOL FOR HISTOLOGY VALIDATION.*
31	*PARAFFIN SECTION WHO CLASSIFICATION (PATHOLOGIST C)*		X	*WHO DIAGNOSES MENU*	*RELATED TO PROTOCOL FOR HISTOLOGY VALIDATION.*
32	*PARAFFIN SECTION WHO CLASSIFICATION (CONSENSUS)*		X	*WHO DIAGNOSES MENU*	*RELATED TO PROTOCOL FOR HISTOLOGY VALIDATION.*

	DAUMAS-DUPORT ASTROCYTOMA GRADE	*REMOVE*		*1/2/3/4*	*CONSIDERED REDUNDANT*
33	*HISTOPATHOLOGY VALIDATED*	X		*YES/NO/NA*	
34	*CONSENSUS CLINICAL DIAGNOSIS*		X	*WHO DIAGNOSES MENU + OTHER DISEASES*	*TO ACCOUNT FOR DISEASES THAT ARE NOT DIAGNOSED HISTOLOGICALLY (DIFFUSE PONTINE GLIOMAS, TECTAL PLATE GLIOMAS, SECRETING GERM CELL TUMOURS), OR TO ACCOUNT FOR CASES IN WHICH HISTOLOGY IS NOT DIAGNOSTIC OR WHEN THERE IS NO HISTOLOGY (METASTASES)*
35	*CHEMOTHERAPY DRUGS USED*	X		*YES/NO/NA*	
36	*RADIOTHERAPHY DOSE GIVEN*	X		*FREE TEXT*	
	OUTCOME SCORE AT THREE MONTHS	REMOVE		*1/2/3/4/5/6/7*	*SUBSTITUTED BY KARNOFSKY / LANSKY INDEX*
	OUTCOME SCORE AT TWO YEARS	REMOVE		*1/2/3/4/5/6/7*	*SUBSTITUTED BY KARNOFSKY / LANSKY INDEX*
37	*KARNOFSKY PERFORMANCE SCORE AT DIAGNOSTIC (for ages above 16)*		X	*0/10/20/30/40/50/60/70/80/90/100*	*SUGGESTED BY P2 UAB*
38	*KARNOFSKY PERFORMANCE SCORE AT THREE MONTHS (for ages above 16)*		X	*0/10/20/30/40/50/60/70/80/90/100*	*SUGGESTED BY P2 UAB*
39	*KARNOFSKY PERFORMANCE SCORE AT TWO YEARS (for ages above 16)*		X	*0/10/20/30/40/50/60/70/80/90/100*	*SUGGESTED BY P2 UAB*
40	*KARNOFSKY PERFORMANCE SCORE AT FIVE YEARS (for ages above 16)*		X	*0/10/20/30/40/50/60/70/80/90/100*	*SUGGESTED BY P2 UAB*
41	*LANSKY PERFORMANCE SCORE AT DIAGNOSTIC (for ages below or equal to 16)*		X	*0/10/20/30/40/50/60/70/80/90/100*	*SUGGESTED BY P18 BU*
42	*LANSKY PERFORMANCE SCORE AT THREE MONTHS (for ages below or equal to 16)*		X	*0/10/20/30/40/50/60/70/80/90/100*	*SUGGESTED BY P18 BU*
43	*LANSKY PERFORMANCE SCORE AT TWO YEARS (for ages below or equal to 16)*		X	*0/10/20/30/40/50/60/70/80/90/100*	*SUGGESTED BY P18 BU*
44	*LANSKY PERFORMANCE SCORE AT FIVE YEARS (for ages below or equal to 16)*		X	*0/10/20/30/40/50/60/70/80/90/100*	*SUGGESTED BY P18 BU*
45	*DATE OF RADIOLOGICAL PROGRESSION/RELAPSE*		X	*DD/MM/YYYY*	*SUGGESTED BY P18 BU*
46	*DATE OF DEATH*		X	*DD/MM/YYYY*	*SUGGESTED BY P18 BU*

47	*CAUSE OF DEATH*		X	*FREE TEXT*	*SUGGESTED BY P18 BU AND P3 SGHMS*
48	*CONCOMITANT DISEASE*	X		*FREE TEXT*	
49	*DATE POSTMORTEM EXAM*	X		*DD/MM/YYYY*	
50	*HISTOPATHOLOGY WHO CLASSIFICATION OF TUMOUR ON AUTOPSY*	X		*WHO DIAGNOSES MENU + FREE TEXT*	*FREE TEXT SUGGESTED BY P2 UAB,, AS MIGHT NOT NECESSARILY BE eTUMOUR PATHOLOGISTS*
	DAUMAS-DUPORT ASTROCYTOMA GRADE OF TUMOUR ON AUTOPSY	REMOVE		*1/2/3/4*	
51	*PRIMARY TUMOUR DETECTED*	X		*YES/NO/NA*	
52	*LOCATION OF PRIMARY TUMOUR*	X		*FREE TEXT*	
53	*SPECTRAL LOCALISATION VALIDATED?*	X		*YES/NO/NA*	
	ASSIGNED CDVC (CLINICAL DATA VALIDATION COMMITEE) CLASS	*TRANSFORMED INTO FIELD 65*		*A/B/C/D/E/F*	
	VALIDATED (MEANS THAT HISTOLOGY IS VALIDATED AND ASSIGNED CDVC CLASS IS NOT "F")	REMOVE		*A/B/C/D/E/F*	
54	*OTHER*	X		*FREE TEXT*	
55	*THE CASE HAS BEEN REVIEWED AND ACCEPTED/DISCARDED BY A NEUROPATHOLOGIST*		X	*ACCEPT/DISCARD/NA*	*NOT DIRECTLY RELATED TO CLINICAL INFORMATION, BUT TO CLINICAL VALIDATION*
56	*NEUROPATHOLOGIST'S COMMENTS ON CASE*		X	*FREE TEXT*	*SAME AS FIELD 55*
57	*THE CASE HAS BEEN REVIEWED AND ACCEPTED/DISCARDED BY A NEUROSURGEON*		X	*ACCEPT/DISCARD/NA*	*SAME AS FIELD 55*
58	*NEUROSURGEON COMMENTS ON CASE*		X	*FREE TEXT*	*SAME AS FIELD 55*
59	*THE CASE HAS BEEN REVIEWED AND ACCEPTED/DISCARDED BY A PEDIATRIC NEURORADIOLOGIST*		X	*ACCEPT/DISCARD/NA*	*SAME AS FIELD 55*
60	*PEDIATRIC NEURORADIOLOGIST'S COMMENTS ON CASE*		X	*FREE TEXT*	*SAME AS FIELD 55*
61	*THE CASE HAS BEEN REVIEWED AND ACCEPTED/DISCARDED BY A NEURORADIOLOGIST*		X	*ACCEPT/DISCARD/NA*	*SAME AS FIELD 55*
62	*NEURORADIOLOGIST'S COMMENTS ON CASE*		X	*FREE TEXT*	*SAME AS FIELD 55*
63	*THE CASE HAS BEEN REVIEWED AND ACCEPTED/DISCARDED BY A NEUROLOGIST*		X	*ACCEPT/DISCARD/NA*	*SAME AS FIELD 55*

64	*NEUROLOGIST'S COMMENTS ON CASE*		X	*FREE TEXT*	*SAME AS FIELD 55*
65	*GENERAL CLINICAL CQCD VALIDATION FIELD*		X	*VALIDATED/NOT VALIDATED/NA*	*SAME AS FIELD 55*

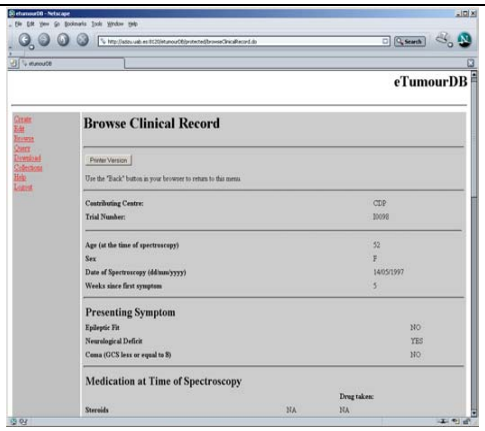
Table 1: *From left to right: field number, field identification, presence/absence of this field in the INTERPRET previous clinical record, whether it is new e-TUMOUR field, defined values for each field and commetns field for justification or explanation of meaning. WHO diagnoses menu [7] depicted in Annex 4. [NA: not available]*

DISCUSSION

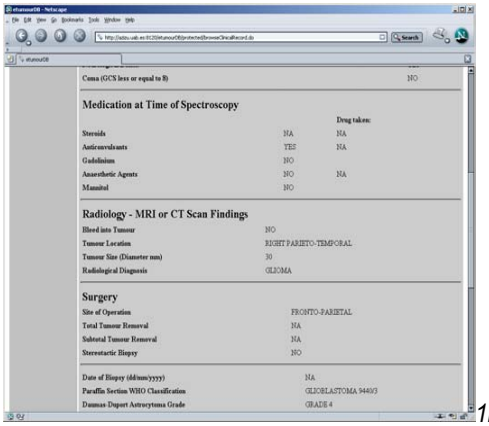
The expertise gained all through the INTERPRET [5] project (January 2000-December 2002) provided eTUMOUR with a web-accessible, secured and quality-control checked database [8] which has been made available to eTUMOUR partners through the following link:

http://azizu.uab.es:8120/etumourDB/

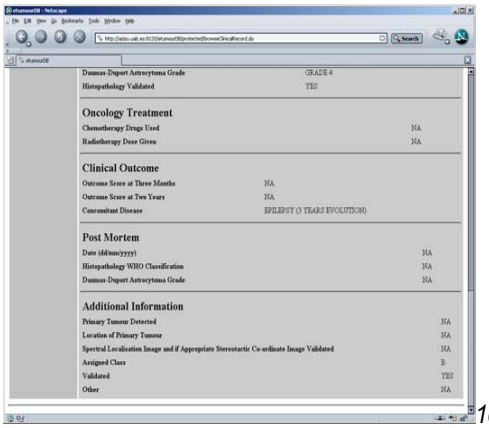
During the INTERPRET project, a CRF was also developed in order to ensure uniform classification of cases to be used in automated classifier and DSS development (Figure 2 and [9]). A public deliverable describing the clinical protocol used was submitted to and approved by the EU-IST [10] and served as starting point to develop the eTUMOUR CRF (Table 1), through the INTERPRET CRF (**Figure 2** and [9]).



1a



1b



1c

Figure 2: *1a, 1b and 1c correspond to the clinical record form as viewed when scrolling down in the CRF part of the preliminary eTumour database adapted from the INTERPRET project.*

In a first stage, a general e-mail survey to all eTUMOUR partners was performed during April 2004 (month 3 of the project). A summary of responses obtained was sent in June 2004 (month 5) to all eTUMOUR partners. From the responses received, there was unanimity among eTUMOUR partners in the INTERPRET model or some other comprehensive way to store the natural history of the brain tumoural disease in each case. An alternative CRF model based on an P2 (UAB) National project (MEDIVO [11]) was also submitted. Essentially, it consisted in storing basic clinical data such as age, sex, diagnosis and Karnofsky performance score (**Annex 3**). This model was not estimated sufficient by eTUMOUR partners as it contained a low number of clinical information data fields.

Following this, the prevailing opinion among eTUMOUR partners was that any field in the CRF should have to be susceptible of statistical correlation with spectral data. P15 UPVLC suggested that missing values could constitute a problem for automated classification systems of brain tumour development and proposed that fields be ordered with respect to their relevance for the analysis, so that the maximum number of cases with complete data will be finally available to WP3 partners.

For this reason, a second survey was sent to (month 5 of project) to be answered by partners involved in data acquisition, quality control and pattern recognition. The general purpose of that was:

- To confirm which fields used in the INTERPRET project CRF were going to be used.
- To confirm which new fields had to be added.

- To agree the format for each field (i.e. free text, menu, and in case of menu, which should be available options).
- To rate their importance for WP3 partners.
- To rate how easily the information was susceptible of quality assurance monitoring in each clinical setting .

The results of the survey, by originating centre, are shown in **Annex 4**.

The main conclusions arising from the survey answers were that most fields included in the CRF were again considered necessary to some extent by the majority of participating groups.

Together with the survey, the following comments were received:

P2 (UAB) proposed to store the following additional information:

- Rate RADIOLOGICAL DIAGNOSIS within a confidence level scale. This would allow that Receiver Operating Characteristic (ROC) curves [12] for evaluating diagnostic performance of the newly developed classifier systems using MRS can be compared to classification performance of MRI alone.
- The originating hospital histopathological diagnosis for each case.

P18 (BU) proposed to store the following information:

- To use the Karnofsky performance score (KPS) [13] instead of the INTERPRET performance score, with a note that the most accepted outcomes are overall survival and event-free survival, implying that date of death and date of progression/relapse should be collected.

P4 (UMCN) by way of Dr. Pieter Wesseling (CSC chairman) proposed to store:

- The diagnostic of the three pathologists as well as the final consensus histopathological diagnostic.

- To perform the clinical validation of cases on-line with the aid of the CRF stored in the eTDB.

Taking all suggestions into account, UAB proposal for the eTUMOUR CRF was that most fields that were available from the INTERPRET project are going to be kept, and that the various possible histopathological diagnoses given to a case will be added to the eTDB. Diagnoses will be stored using the WHO classification of brain tumours (**Annex 5**) and there will also be the possiility for storing diagnoses in free text format in case of diseases that are not diagnosed histologically. Karnofsky performance score (KPS) or Lansky score (for patients below or equal to 16 years of age) (**Annex 6**) will be stored instead of the INTERPRET performance score and also date of death and date of progression/relapse from the radiological point of view, will be recorded. A confidence scale number for the radiological diagnosis field will also be included, several fields allowing the storage of the DQCE results of the CSC of the CDVP will be included. Those will include fields for the different specialists to tick and accept or discard after review. A field for general validation of a case will be included as well. This model was circulated during month 7 of the project and after that period, one final modification was included after integrating further feedback from clinical partners. The final modification was in field “Tumour location”, allowing a menu instead of free text. With this final modification incorporated, Partner UAB as responsible for coordinating the definitive CRF form, considered that the document was agreed by all partners having received no further requests. Minor suggestions received on degree of tumour resection have caused that the INTERPRET fields “total tumour removal”, “subtotal tumour removal” and “stereotactic biopsy” have been merged into one single field “tumour removal” with the following options: complete macroscopic resection, partial macroscopic resection, open biopsy,

stereotactic biopsy. It was considered whether data is to be used only for prognosis where it would not be necessary to ditinguish between open and stereotactic biopsy. However, if data from post operative MR scans is to be added at any point then this information may be important. In an open biopsy a sizable bit of tumour is removed, the blood supply is disrupted and metabolic changes may well occur. In contrast, when a stereotactic biopsy is undertaken only small fragments of tumour are removed with very little disruption to the overall tumour and surrounding tumour. The metabolic status of the tumour may thus be very much more preserved.

The final template CRF is shown as **Annex 7** at the end of this deliverable.

BIBLIOGRAPHY AND REFERENCES

[1] e-TUMOUR deliverable D9 (D2.3)

[2] e-TUMOUR deliverable D13 (D2.4)

[3] Julià-Sapé M, Acosta D, Majós C, Wesseling P, Pujol J, Gajewicz W, Bell BA, Griffiths JR, Arús C (2002) Comparison between radiological and histopathological diagnoses in a multicentre international brain tumour magnetic resonance database. *MAGMA* 15(Suppl 1): 92

[4] Murphy M, Loosemore A, Clifton AG, Howe FA, Tate AR, Cudlip SA, Wilkins PR, Griffiths JR, Bell BA (2004) The contribution of proton magnetic resonance spectroscopy (^{1}HMRS) to clinical brain tumour diagnosis. *Br J Neurosurg* 16(4):329-34

[5] INTERPRET project, IST-1999-10310: http://carbon.uab.es/INTERPRET

[6] INTERPRET Technical report number 161: http://azizu.uab.es/INTERPRET/private/technical_reports.shtml

[7] Pathology and Genetics of Tumours of the Nervous System. P Kleihues & WK Cavenee, Eds. IARC Press, Lyon, France, 2000. ISBN 92 832 2409 4

[8] INTERPRET validated database: http://azizu.uab.es/INTERPRET/int_Disc_FrozenDB.shtml

[9] INTERPRET clinical record form: http://azizu.uab.es/INTERPRET/clinical_data/clinical_data04.html

[10] INTERPRET clinical data protocol deliverable: http://azizu.uab.es/INTERPRET/cdap.html

[11] MEDIVO: http://azizu.uab.es/MEDIVO/

[12] Obuchowski NA (2003) Receiver Operating Characteristic Curves and Their Use in Radiology. *Radiology* 229:3–8

[13] Karnofsky performance score: http://www.clevelandclinicmeded.com/diseasemanagement/neurology/braintumor/table3.htm

ANNEXES INDEX

ANNEX 1a

GUIDANCE IF FORM

Information sheet

Name of Hospital

Title of project: eTUMOUR (WEB ACCESSIBLE MR DECISION SUPPORT SYSTEM FOR BRAIN TUMOUR DIAGNOSIS AND PROGNOSIS, INCORPORATING IN VIVO AND EX VIVO GENOMIC AND METABOLOMIC DATA).

We are trying to find better ways of diagnosing patients using brain scanning. New Magnetic Resonance (MR) techniques provide information on brain structure and chemistry that may help us and we would like to invite you to take part in our research. The results we obtain will not affect your own treatment but could be of benefit to future patients

If you agree to have an MR exam you would be asked to lie in a scanner used for producing routine MR images. We will obtain information on the chemicals in your brain as well as detailed anatomical pictures. The examination will take about an hour and you will be asked to lie as still as possible. At any time while in the magnet you will be able to signal to the MR staff that you wish to come out by pressing a call button.

We may ask to give you an injection of a "contrast agent" (which you may already have had as it is a routine part of normal scanning). This would enable us to more accurately diagnose any abnormality. There is a 1 in 10,000 risk of a slight reaction that may include skin rashes, nausea and vomiting to this agent.

If your treatment were to involve surgery then we would like to. keep a sample for use in a variety of research projects, which will help to improve our understanding of brain cancer. For example analyse tissue taken at this surgery can be used to aid our understanding of its biological and genetic characteristics If we are to improve the treatment for Brain tumours we need to better understand these diseases. This tissue will only be taken for further study once a diagnosis has been made and will in no way compromise your treatment. The Genetic information obtained from our study will not be used for any other purpose than to improve our understanding of brain tumours; will not be disclosed to any other person or organisation except those involved in this study; will be coded so that cannot be related to you

except by ourselves. We will inform your doctor in the event of discoveries that may affect your health. A portion of this abnormal tissue may be sent to other Pathologists and Biologists in Europe who are collaborating on this study, the samples will be coded to maintain your anonymity. There are no extra risks involved in taking abnormal tissue for this study over and above those of the surgery. No material/tissue will be taken or stored for research without your agreement.

The data from our study will be entered into a computer database to create a diagnostic tool that will help future doctors with their diagnoses. Data will be coded and no details will be entered into the database that would enable you to be identified by others except ourselves. We will also review your case notes now and at some time in the future, and will contact doctors that may treat you in the future for the purposes of completing the information our database.

If you have any further enquiries regarding this please contact the Consultant in charge of your treatment, one of their team, or Dr. "X" on Telf "nnnn".

Entry into this project is entirely voluntary and of your own free will, you are free to withdraw any time without giving any reasons. Entry into the study will not affect your ongoing treatment. Similarly, refusal to entry into the study will not imply any loss of right to the best medical assistance we can provide. The same degree of confidentiality applies to the results of these scans as to your usual medical records.

The Local Research Ethics Committee has approved the above statement.

Signed by the person in charge of the project:

_____ Date: _____

Dr. "X"

Signed by the Chair of that Committee: [where required]

_____ Date: _____

Consent form

Name of Hospital

Title of project: eTUMOUR (WEB ACCESSIBLE MR DECISION SUPPORT SYSTEM FOR BRAIN TUMOUR DIAGNOSIS AND PROGNOSIS, INCORPORATING IN VIVO AND EX VIVO GENOMIC AND METABOLOMIC DATA).

This form should be signed by patients/volunteers, undergoing any test, treatment or other procedure connected with clinical research.

	YES	NO	N/A
1. I confirm that I have read and understood the "Patient Information Sheet" which describes this research project and have been given a copy to keep. I have had enough time to decide whether I wish to take part in the above study.			
2. The nature, purpose and possible consequence of taking part in this project have been explained to me by _____ and are acceptable to me.			
3. I am entering this project of my own free will, and understand I am free to withdraw from this study at any time without giving reasons, and that participation or non-participation will not prejudice my treatment in any way.			
4. I agree to have an MR examination and any data obtained may be anonymised, stored, processed and used in a diagnostic tool.			
5. I agree to have an injection of a "contrast agent" to improve diagnosis of the abnormal area and understand that the risks of reaction to this contrast is approximately 1 in 10,000 and includes skin rashes, nausea and vomiting.			
6. I agree for samples of any abnormal tissue obtained during routine treatment to be stored and used in research studies and any data obtained may be anonymised, stored and processed and used in a diagnostic tool.			
7. I do specifically agree to donate tissue samples removed at operation, blood and/or bone marrow samples extra to what is required for medical purposes, for future use in approved research projects. I understand that this includes the storage of small pieces of frozen tissue, wax embedded pathology blocks of tissue and tissue sections on microscope slides.			
8. I understand that I will not be told the results of any tests which may be carried out on the samples I donated and that, if in the future the research shows that there is a test which might be useful to me, the information will be given to my doctor, who will discuss it with me.			

Name of Patient Signature Date

Address

_____

Name of Witness Signature Date

Address

_____

ANNEX 1b

LIST OF MINIMAL CONCEPTS TO BE USED, COMMON TO ALL IF FORMS IN THE ETUMOUR PROJECT

The basic requirements that each local form will have to contain, as

agreed during the first eTUMOUR plenary meeting will be the following:

1-Patient data will be stored in the PROJECT database in an anonymised form.

2-Patient data will be used to develop a decission support system to improve diagnostic of human brain tumours.

3-The patient gives a sample biopsy as a "gift" to the PROJECT.

ANNEX 2

Quality control for incoming clinical data

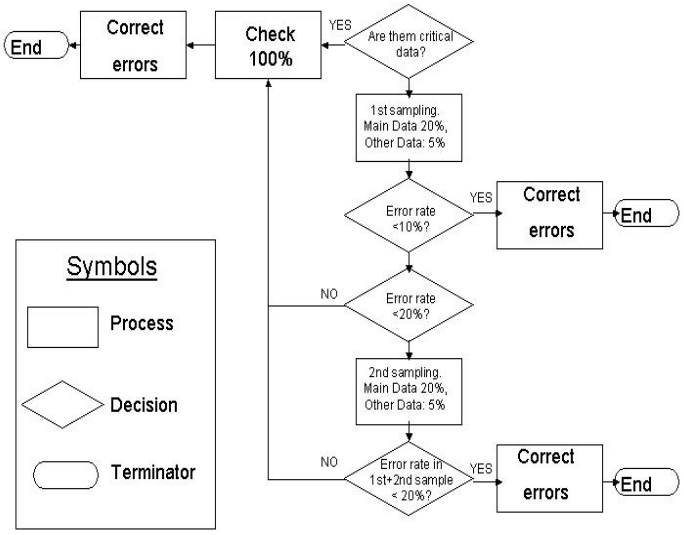
Data included in the database will be submitted to quality control procedures. For this purpose, data will be classified into three groups: critical, main and other data. These groups are defined as follows:

- Critical data: data that allow development of classifiers and the traceability of the data.
- Main data: data that have an influence on critical data.
- Other data: remaining data.

A sampling approach is defined for quality control for each group:

- Critical data: 100%
- Main data: 20%
- Other data: 5%

Testing procedure and acceptance criteria are depicted in the next figure:



ANNEX 3

MEDIVO PROJECT CLINICAL RECORD FORM

CODE: M-02 CLINICAL RECORD FOR DATABASE MEDIVO

	DATA ACCRUAL WORKSHEET FOR THE MEDIVO PROJECT	
		OPTIONS
1	CONTRIBUTING CENTRE	IDI
2	TRIAL NUMBER	EXACT NUMBER
3	ARE THERE PREVIOUS ERM STUDIES?	YES/NO/DO NOT KNOW
4	CODE OF PREVIOUS STUDY	FREE TEXT FORMAT
5	AGE (AT DATE OF SPECTROSCOPY)	exact number
6	SEX	M/F/DO NOT KNOW
7	DATE OF SPECTROSCOPY	DATE (in compatible format)
8	SURGERY	YES/NO/DO NOT KNOW
9	SURGICAL PROCEDURE	TOTAL RESECTION / PARTIAL RESECTION / BIOPSY/DO NOT KNOW
	DIAGNOSTIC	
10	DATE OF ONSET OF CLINICAL SYMPTOMS	DATE (in compatible format)
11	CLINICAL DIAGNOSIS	FREE TEXT FORMAT
12	RADIOLOGICAL DIAGNOSIS	FREE TEXT FORMAT
13	HISTOPATHOLOGICAL DIAGNOSIS	MENU 1
14	COMBINED DIAGNOSIS	FREE TEXT FORMAT
15	DATE WHEN THE HISTOPATHOLOGICAL DIAGNOSIS WAS REACHED	DATE (in compatible format)
16	DATE WHEN THE COMBINED DIAGNOSIS WAS REACHED	DATE (in compatible format)
17	PRIMARY TUMOUR (only if the definitive diagnosis is metastasis)	FREE TEXT FORMAT
	FURTHER ONCOLOGY TREATMENT	
18	CHEMOTHERAPY TREATMENT?	YES/NO/DO NOT KNOW
19	RADIOTHERAPY TREATMENT?	YES/NO/DO NOT KNOW
20	COMMENTS	FREE TEXT FORMAT
	FOLLOW-UP	
21	KPS AT DIAGNOSTIC	10 / 20 / 30 / 40 / 50 / 60 / 70 / 80 / 90 / 100
22	KPS AT THREE MONTHS	10 / 20 / 30 / 40 / 50 / 60 / 70 / 80 / 90 / 100
23	KPS AT TWO YEARS	10 / 20 / 30 / 40 / 50 / 60 / 70 / 80 / 90 / 100
24	KPS AT FIVE YEARS	10 / 20 / 30 / 40 / 50 / 60 / 70 / 80 / 90 / 100
	DEATH	
25	DATE	DATE (in compatible format)
26	CAUSE	FREE TEXT FORMAT
27	**OTHER**	FREE TEXT FORMAT

ANNEX 4

RESULTS OF SURVEY CIRCULATED TO ETUMOUR PARTNERS

	DO YOU CONSIDER THIS FIELD NECESSARY FOR PATTERN RECOGNITION ANALYSIS?			***RATE HOW ESSENTIAL YOU CONSIDER THIS FIELD FOR CLINICAL DATA VALIDATION?***				***RATE HOW EASILY DO YOU THINK THIS DATA CAN BE OBTAINED***	
	P15 UPVLC	*P11 KUL*	*P3 SGHMS*	*P18 BU*	*P4 UMCN*	*P2 UAB*	*P3 SGHMS*	*P18 BU*	*P3 SGHMS*
AGE	3	3	2	4	4	0	4	0	0
SEX	3	3	2	3	4	0	4	0	0
DATE OF SPECTROSCOPY	3	3	3	4	3	4	4	0	0
WEEKS SINCE FIRST SYMPTOM	2	2	2	3	2	0	2	2	3
PRESENTING SYMPTOM: EPILEPTIC FIT	2	2	3	2	1	0	2	2	2
PRESENTING SYMPTOM: NEUROLOGICAL DEFICIT	2	2	3	2	1	0	2	2	2
PRESENTING SYMPTOM: COMA	2	2	3	3	1	0	3	2	0
MEDICATION AT TIME OF SPECTROSCOPY: STEROIDS	2	2	3	3	3	0	3	2	1
MEDICATION AT TIME OF SPECTROSCOPY: ANTICONVULSANTS	2	2	3	3	3	0	2	2	1
MEDICATION AT TIME OF SPECTROSCOPY: GADOLINIUM	2	2	3	3	4	0	3	0	0
MEDICATION AT TIME OF SPECTROSCOPY: ANAESTHETIC AGENTS	2	2	3	3	3	0	2	0	1
MEDICATION AT TIME OF SPECTROSCOPY: MANNITOL	2		3	3	3	0	2	2	2
BLEED INTO TUMOUR	3	3	3	3	3	0	3	2	1
TUMOUR LOCATION	3	3	3	3	4	0	3	1	1
TUMOUR SIZE	3	3	3	3	3	0	2	1	2
RADIOLOGICAL DIAGNOSIS	3	2	3	3	3	3	3	2	2
SITE OF OPERATION	2	2	2	3	3	0	2	1	1
TOTAL TUMOUR REMOVAL	2	2	3	3	3	0	2	2	2
SUBTOTAL TUMOUR REMOVAL	2	2	3	3	3	0	2	1	2
STEREOTACTIC BIOPSY	2	2		3	3	1	2	0	2
DATE OF BIOPSY	1	1	3	4	3	4	4	0	0
PARAFFIN SECTION WHO CLASSIFICATION	4	4	4	4	4	4	4	1	0
DAUMAS-DUPORT ASTROCYTOMA GRADE	2	3	4		0	4	4		0
HISTOPATHOLOGY VALIDATED	4	4	4	3	4	4	3	1	?
CHEMOTHERAPY DRUGS USED	2	2	3	3	3	1	2	1	2
RADIOTHERAPY DOSE GIVEN	2	2	3	3	3	1	2	0	2
OUTCOME SCORE AT THREE MONTHS	3	3	3	4	3	3	3	1	2

OUTCOME SCORE AT TWO YEARS	3	3	3	4	3	3	3	1	2
CONCOMITANT DISEASE	3	2	1	2	3	0	2	2	2
DATE POST-MORTEM EXAM	1	2	0	3	2	0	2	0	2
HISTOPATHOLOGY WHO CLASSIFICATION OF TUMOUR ON AUTOPSY	3	3	1	3	3	0	3	1	2
DAUMAS-DUPORT ASTROCYTOMA GRADE OF TUMOUR ON AUTOPSY	3	3	1		0	0	3		2
PRIMARY TUMOUR DETECTED	3	3	2	3	3	2	2	1	2
LOCATION OF PRIMARY TUMOUR	3	3	2	2	3	3	2	1	2
SPECTRAL LOCALISATION VALIDATED?	4	4	4	3	4	4	3	1	?
ASSIGNED CDVC (CLINICAL DATA VALIDATION COMMITTEE) CLASS	4	4	4	3	4	4	3	0	?
VALIDATED (MEANS THAT HISTOLOGY IS VALIDATED AND ASSIGNED CDVC CLASS IS NOT "F")	4	4	4	3	4	4	3	0	?
OTHER									
NEW FIELDS SUGGESTED AFTER DISCUSSION									
DATE OF DEATH		3	1	4	3		3		1
CAUSE OF DEATH		2	1	3			3		1
DATE OF BIRTH		2	3	3			month and year only if data protection important		
DATE OF RADIOTHERAPY		2		3			2		1
DATE OF CHEMOTHERAPY		2		3			2		1
DATE OF FURTHER SURGERY				4					1
STEREOTACTIC BIOPSY OR BIOPSY DURING SURGERY			4						
MR SYSTEM			1						
MRS PARAMETERS			1						
MRS VOXEL FROM: ENHANCING REGION, NON-ENHANCING, DON'T KNOW			3						
MR IMAGES AVAILABLE?		3							
HR MAS INFO AVAILABLE?		2							
MICROARRAY DATA AVAILABLE?		2							
NECROSIS		2							
FAMILY HISTORY (OTHER FAMILY MEMBERS WITH BRAIN TUMOUR?)		2							
OTHER PRESENTING SYMPTOMS (E.G. HEADACHE, NAUSEA)					1				
MRS IDENTIFIER				4					

ANNEX 5

WHO DIAGNOSES MENU

DIFFUSE ASTROCYTOMA 9400/3
FIBRILLARY ASTROCYTOMA 9420/3
PROTOPLASMIC ASTROCYTOMA 9410/3
GEMISTOCYTIC ASTROCYTOMA 9411/3
ANAPLASTIC ASTROCYTOMA 9401/3
GLIOBLASTOMA 9440/3
GIANT CELL GLIOBLASTOMA 9441/3
GLIOSARCOMA 9442/3
PILOCYTIC ASTROCYTOMA 9421/1
PLEOMORPHIC XANTHOASTROCYTOMA 9424/3
SUBEPENDYMAL GIANT CELL ASTROCYTOMA 9384/1
OLIGODENDROGLIOMA 9450/3
ANAPLASTIC OLIGODENDROGLIOMA 9451/3
OLIGOASTROCYTOMA 9382/3
ANAPLASTIC OLIGOASTROCYTOMA 9382/3
EPENDYMOMA 9391/3
CELLULAR EPENDYMOMA 9391/3
PAPILLARY EPENDYMOMA 9393/3
CLEAR CELL EPENDYMOMA 9391/3
TANYCYTIC EPENDYMOMA 9391/3
ANAPLASTIC EPENDYMOMA 9392/3
MYXOPAPILLARY EPENDYMOMA 9394/1
SUBEPENDYMOMA 9383/1
CHOROID PLEXUS PAPILLOMA 9390/0
CHOROID PLEXUS CARCINOMA 9390/3
ASTROBLASTOMA 9430/3
GLIOMATOSIS CEREBRI 9381/3
CHORDOID GLIOMA OF THE 3RD VENTRICLE 9444/1
GANGLIOCYTOMA 9492/0
DYSPLASTIC GANGLIOCYTOMA OF CEREBELLUM 9493/0
DESMOPLASTIC INFANTILE ASTROCYTOMA/GANGLIOGLIOMA 9412/1
DYSEMBRYOPLASTIC NEUROEPITHELIAL TUMOUR 9413/0
GANGLIOGLIOMA 9505/1
ANAPLASTIC GANGLIOGLIOMA 9505/3
CEREBELLAR LIPONEUROCYTOMA 9506/1
PARAGANGLIOMA OF THE FILUM TERMINALE 8680/1
CENTRAL NEUROCYTOMA 9506/1
OLFACTORY NEUROBLASTOMA 9522/3
OLFACTORY NEUROEPITHELIOMA 9523/3
NEUROBLASTOMAS OF THE ADRENAL GLAND AND SYMPATHETIC NERVOUS SYSTEM 9500/3
PINEOCYTOMA 9361/1
PINEOBLASTOMA 9362/3
PINEAL PARENCHYMAL TUMOUR OF INTERMEDIATE DIFFERENTIATION 9362/3
MEDULLOEPITHELIOMA 9501/3
EPENDYMOBLASTOMA 9392/3
MEDULLOBLASTOMA 9470/3
DESMOPLASTIC MEDULLOBLASTOMA 9471/3
LARGECELL MEDULLOBLASTOMA 9474/3
MEDULLOMYOBLASTOMA 9472/3
MELANOTIC MEDULLOBLASTOMA 9470/3
SUPRATENTORIAL PRIMITIVE NEUROECTODERMAL TUMOUR 9473/3

PNET NEUROBLASTOMA 9500/3
PNET GANGLIONEUROBLASTOMA 9490/3
ATYPICAL TERATOID/RHABDOID TUMOUR 9508/3
SCHWANNOMA 9560/0
CELLULAR SCHWANNOMA 9560/0
PLEXIFORM SCHWANNOMA 9560/0
MELANOTIC SCHWANNOMA 9560/0
NEUROFIBROMA 9540/0
PLEXIFORM NEUROFIBROMA 9550/0
PERINEURINOMA 9571/0
INTRANEURAL PERINEURINOMA 9571/0
SOFT TISSUE PERINEURINOMA 9571/0
MALIGNANT PERIPHERAL NERVE SHEATH TUMOUR 9540/3
EPITHELIOID MPNST 9540/3
MPNST WITH DIVERGENT MESENCHYMAL AND/OR EPITHELIAL DIFFERENTIATION 9540/3
MELANOTIC MPNST 9540/3
MELANOTIC PSAMMOMATOUS MPNST 9540/3
MENINGIOMA 9530/0
MENINGOTHELIAL MENINGIOMA 9531/0
FIBROUS MENINGIOMA 9532/0
TRANSITIONAL MENINGIOMA 9537/0
PSAMMOMATOUS MENINGIOMA 9533/0
ANGIOMATOUS MENINGIOMA 9534/0
MICROCYSTIC MENINGIOMA 9530/0
SECRETORY MENINGIOMA 9530/0
LYMPHOPLASMACYTE-RICH MENINGIOMA 9530/0
METAPLASTIC MENINGIOMA 9530/0
CLEAR CELL MENINGIOMA 9538/1
CHORDOID MENINGIOMA 9538/1
ATYPICAL MENINGIOMA 9539/1
PAPILLARY MENINGIOMA 9538/3
RHABDOID MENINGIOMA 9538/3
ANAPLASTIC MENINGIOMA 9530/3
MESENCHYMAL NON-MENINGOTHELIAL TUMOURS
LIPOMA 8850/0
ANGIOLIPOMA 8861/0
HIBERNOMA 8880/0
LIPOSARCOMA 8850/3
SOLITARY FIBROUS TUMOUR 8815/0
FIBROSARCOMA 8810/3
MALIGNANT FIBROUS HISTIOCYTOMA 8830/3
LEIOMYOMA 8890/0
LEIOMYOSARCOMA 8890/3
RHABDOMYOMA 8900/0
RHABDOMYOSARCOMA 8900/3
CHONDROMA 9220/0
CHONDROSARCOMA 9220/3
OSTEOMA 9180/0
OSTEOSARCOMA 9180/3
OSTEOCHONDROMA 9210/0
HAEMANGIOMA 9120/0
EPITHELIOID HAEMANGIOENDOTHELIOMA 9133/1
ANGIOSARCOMA 9120/3
KAPOSI SARCOMA 9140/3
HAEMANGIOPERICYTOMA 9150/1
DIFFUSE MELANOCYTOSIS 8728/0
MELANOCYTOMA 8728/1

MALIGNANT MELANOMA 8720/3
MENINGEAL MELANOMATOSIS 8728/3
HAEMANGIOBLASTOMA 9161/1
MALIGNANT LYMPHOMAS 9590/3
PLASMACYTOMA 9731/3
GRANULOCYTIC SARCOMA 9930/3
GERMINOMA 9064/3
EMBRYONAL CARCINOMA 9070/3
YOLK SAC TUMOUR 9071/3
CHORIOCARCINOMA 9100/3
TERATOMA 9080/1
MATURE TERATOMA 9080/0
IMMATURE TERATOMA 9080/3
TERATOMA WITH MALIGNANT TRANSFORMATION 9084/3
MIXED GERM CELL TUMOURS 9085/3
CRANIOPHARYNGIOMA 9350/1
ADAMANTINOMATOUS CRANIOPHARYNGIOMA 9351/1
PAPILLARY CRANIOPHARYNGIOMA 9352/1
GRANULAR CEIL TUMOUR 9582/0
PITUITARY ADENOMA 8140/0
PITUITARY CARCINOMA 8140/3
PARAGANGLIOMA 8680/1
CHORDOMA 9370/3
CHONDROMA 9220/0
CHONDROSARCOMA 9220/3
CARCINOMA 8200/3
METASTASIS 8000/6

ANNEX 6

Lansky / Karnofsky scales

LANSKY (1-16 Years)

Fully active, normal	**100**
Minor restrictions in physically strenuous activity	**90**
Active, but tires more quickly	**80**
Both greater restriction of, and less time spent in, active play	**70**
Up and around, but minimal active play, keeps busy with quieter activities	**60**
Gets dressed but lies around much of the day; no active play, able to participate in all quiet play and activities	**50**
Mostly in bed, participates in quiet activities	**40**
In bed, needs assistance even for quiet play	**30**
Often sleeping, play entirely limited to very passive activities	**20**
No play, does not get out of bed	**10**
Unresponsive	**0**

KARNOFSKY (> 16Years)

Normal, no complaints	**100**
Able to carry on normal activities; minor signs or symptoms of disease	**90**
Normal activity with effort	**80**
Cares for self. Unable to carry on normal activity, or to do active work	**70**
Ambulatory. Requires some assistance, but able to care for most of own needs	**60**
Requires considerable assistance and frequent medical care	**50**
Disabled; requires special care and assistance	**40**
Severely disabled, hospitalisation indicated though death not imminent	**30**
Very sick. Hospitalisation necessary. Active supportive treatment necessary	**20**
Moribund, fatal processes in progression	**10**
Dead	**0**

ANNEX 7

DEFINITIVE CRF

FIELD NUMBER	*FIELD NAME*	*POSSIBLE VALUES*
1	*AGE*	*INTEGER*
2	*DATE OF BIRTH*	*DD / MM / YYYY*
3	*SEX*	*M / F / NA*
4	*DATE OF SPECTROSCOPY*	*DD / MM / YYYY*
5	*WEEKS SINCE FIRST SYMPTOM*	*INTEGER*
6	*PRESENTING SYMPTOM: EPILEPTIC FIT*	*YES / NO / NA*
7	*PRESENTING SYMPTOM: NEUROLOGICAL DEFICIT*	*YES / NO / NA*
8	*PRESENTING SYMPTOM: COMA*	*YES / NO / NA*
9	*MEDICATION AT TIME OF SPECTROSCOPY: STEROIDS*	*YES / NO / NA*
10	*MEDICATION AT TIME OF SPECTROSCOPY: ANTICONVULSANTS*	*YES / NO / NA*
11	*MEDICATION AT TIME OF SPECTROSCOPY: GADOLINIUM*	*YES / NO / NA*
12	*MEDICATION AT TIME OF SPECTROSCOPY: ANAESTETHIC AGENTS*	*YES / NO / NA*
13	*MEDICATION AT TIME OF SPECTROSCOPY: MANNITOL*	*YES / NO / NA*
14	*BLEED INTO TUMOUR*	*YES / NO / NA*
15	*TUMOUR LOCATION*	*TECTAL PLATE / TEMPORAL /MIDBRAIN / MEDULLA / PONS / PINEAL REGION / OPTIC CHIASM / HYPOTHALAMUS / SUPRASELLAR REGION / VENTRICULAR / FRONTAL / PARIETAL / OCCIPITAL / BASAL GANGLIA / CEREBELLUM ## LEFT / RIGHT*
16	*CALCIFICATIONS*	*YES / NO / NA*
17	*NUMBER OF DETECTED LESIONS*	*INTEGER*
18	*TUMOUR SIZE*	*nn x nn x nn (mm) ## nn OF BIGGER AXIS FOR GLIAL TUMOURS (mm)*
19	*RADIOLOGICAL DIAGNOSIS 1*	*FREE TEXT*
20	*CONFIDENCE RATING FOR RADIOLOGICAL DIAGNOSIS 1*	*0 / 1 / 2 / 3 / 4*

21	*RADIOLOGICAL DIAGNOSIS 2*	*FREE TEXT*
22	*CONFIDENCE RATING FOR RADIOLOGICAL DIAGNOSIS 2*	*0 / 1 / 2 / 3 / 4*
23	*RADIOLOGICAL DIAGNOSIS 3*	*FREE TEXT*
24	*CONFIDENCE RATING FOR RADIOLOGICAL DIAGNOSIS 3*	*0 / 1 / 2 / 3 / 4*
25	*SITE OF OPERATION*	*FREE TEXT*
26	*TUMOUR REMOVAL*	*COMPLETE MACROSCOPIC RESECTION / PARTIAL MACROSCOPIC RESECTION / OPEN BIOPSY / STEREOTACTIC BIOPSY*
27	*DATE OF BIOPSY*	*DD/MM/YYYY*
28	*PARAFFIN SECTION WHO CLASSIFICATION (ORIGINATING PATHOLOGIST)*	*WHO DIAGNOSES MENU + FREE TEXT*
29	*PARAFFIN SECTION WHO CLASSIFICATION (PATHOLOGIST A)*	*WHO DIAGNOSES MENU*
30	*PARAFFIN SECTION WHO CLASSIFICATION (PATHOLOGIST B)*	*WHO DIAGNOSES MENU*
31	*PARAFFIN SECTION WHO CLASSIFICATION (PATHOLOGIST C)*	*WHO DIAGNOSES MENU*
32	*PARAFFIN SECTION WHO CLASSIFICATION (CONSENSUS)*	*WHO DIAGNOSES MENU*
33	*HISTOPATHOLOGY VALIDATED*	*YES/NO/NA*
34	*CONSENSUS CLINICAL DIAGNOSIS*	*WHO DIAGNOSES MENU + OTHER DISEASES*
35	*CHEMOTHERAPY DRUGS USED*	*YES/NO/NA*
36	*RADIOTHERAPHY DOSE GIVEN*	*FREE TEXT*
37	*KARNOFSKY PERFORMANCE SCORE AT DIAGNOSTIC (for ages above 16)*	*0/10/20/30/40/50/60/70/80/90/100*
38	*KARNOFSKY PERFORMANCE SCORE AT THREE MONTHS (for ages above 16)*	*0/10/20/30/40/50/60/70/80/90/100*
39	*KARNOFSKY PERFORMANCE SCORE AT TWO YEARS (for ages above 16)*	*0/10/20/30/40/50/60/70/80/90/100*

40	*KARNOFSKY PERFORMANCE SCORE AT FIVE YEARS (for ages above 16)*	*0/10/20/30/40/50/60/70/80/90/100*
41	*LANSKY PERFORMANCE SCORE AT DIAGNOSTIC (for ages below or equal to 16)*	*0/10/20/30/40/50/60/70/80/90/100*
42	*LANSKY PERFORMANCE SCORE AT THREE MONTHS (for ages below or equal to 16)*	*0/10/20/30/40/50/60/70/80/90/100*
43	*LANSKY PERFORMANCE SCORE AT TWO YEARS (for ages below or equal to 16)*	*0/10/20/30/40/50/60/70/80/90/100*
44	*LANSKY PERFORMANCE SCORE AT FIVE YEARS (for ages below or equal to 16)*	*0/10/20/30/40/50/60/70/80/90/100*
45	*DATE OF RADIOLOGICAL PROGRESSION/RELAPSE*	*DD/MM/YYYY*
46	*DATE OF DEATH*	*DD/MM/YYYY*
47	*CAUSE OF DEATH*	*FREE TEXT*
48	*CONCOMITANT DISEASE*	*FREE TEXT*
49	*DATE POSTMORTEM EXAM*	*DD/MM/YYYY*
50	*HISTOPATHOLOGY WHO CLASSIFICATION OF TUMOUR ON AUTOPSY*	*WHO DIAGNOSES MENU + FREE TEXT*
51	*PRIMARY TUMOUR DETECTED*	*YES/NO/NA*
52	*LOCATION OF PRIMARY TUMOUR*	*FREE TEXT*
53	*SPECTRAL LOCALISATION VALIDATED?*	*YES/NO/NA*
54	*OTHER*	*FREE TEXT*
55	*THE CASE HAS BEEN REVIEWED AND ACCEPTED/DISCARDED BY A NEUROPATHOLOGIST*	*ACCEPT/DISCARD/NA*
56	*NEUROPATHOLOGIST'S COMMENTS ON CASE*	*FREE TEXT*
57	*THE CASE HAS BEEN REVIEWED AND ACCEPTED/DISCARDED BY A NEUROSURGEON*	*ACCEPT/DISCARD/NA*
58	*NEUROSURGEON COMMENTS ON CASE*	*FREE TEXT*
59	*THE CASE HAS BEEN REVIEWED AND ACCEPTED/DISCARDED BY A PEDIATRIC NEURORADIOLOGIST*	*ACCEPT/DISCARD/NA*
60	*PEDIATRIC NEURORADIOLOGIST'S COMMENTS ON CASE*	*FREE TEXT*

61	*THE CASE HAS BEEN REVIEWED AND ACCEPTED/DISCARDED BY A NEURORADIOLOGIST*	*ACCEPT/DISCARD/NA*
62	*NEURORADIOLOGIST'S COMMENTS ON CASE*	*FREE TEXT*
63	*THE CASE HAS BEEN REVIEWED AND ACCEPTED/DISCARDED BY A NEUROLOGIST*	*ACCEPT/DISCARD/NA*
64	*NEUROLOGIST'S COMMENTS ON CASE*	*FREE TEXT*
65	*GENERAL CLINICAL CQCD VALIDATION FIELD*	*VALIDATED/NOT VALIDATED/NA*

A-2 Description of the eTUMOUR database

The eTUMOUR database (eTDB) is a data warehouse that is accessible through internet. The access is restricted to eTUMOUR partners that contribute or must analyze data. Each data type is stored into an independent section (see figure 6.1).

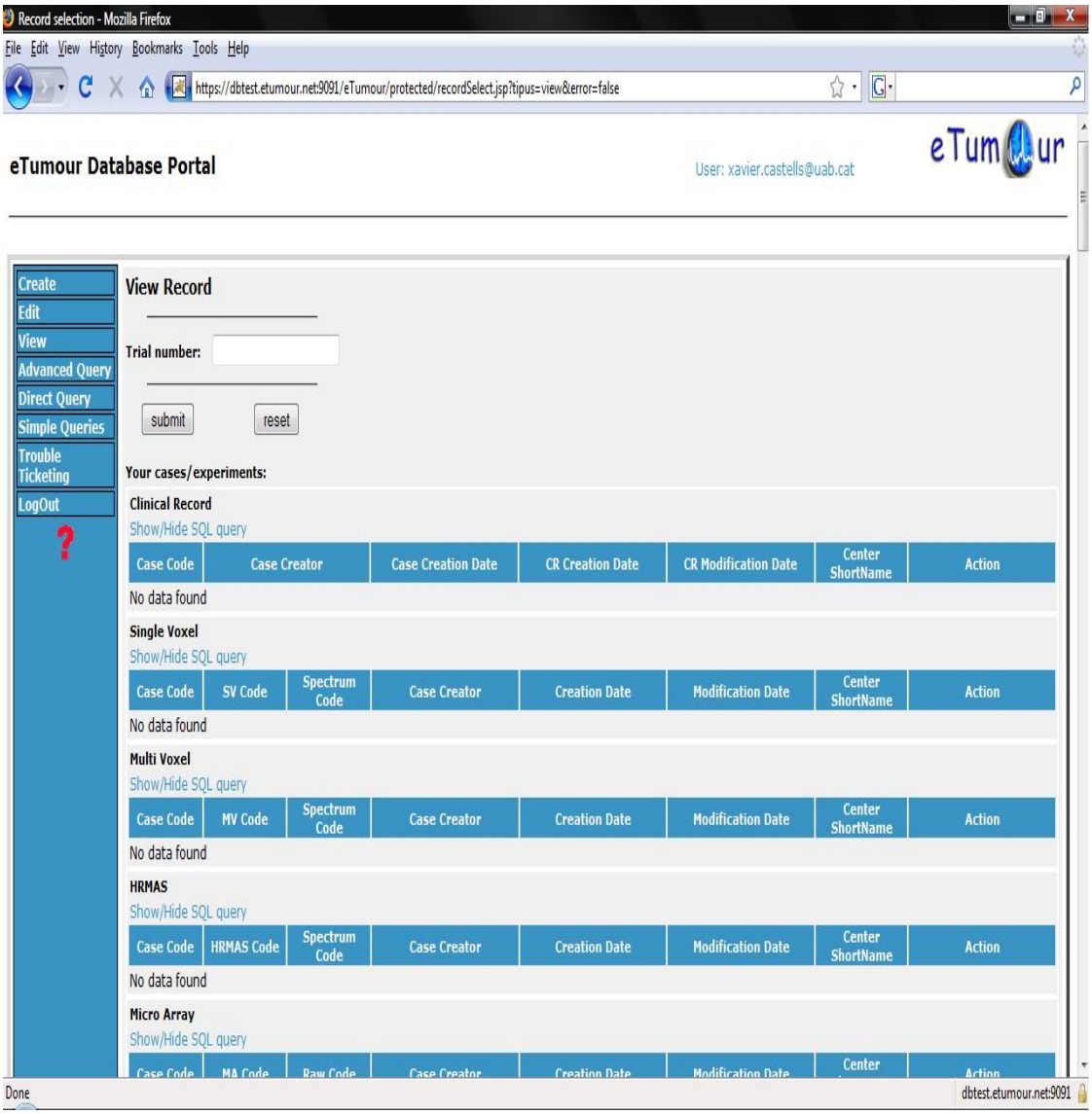


Figure 6.1: *Screenshot for the initial page of the eTDB after successful login.* All sections that compose the eTDB can be seen at this initial page. Each section can be independently accessed.

The section of transcriptomic data has a first part with the information obtained from the RNA isolation (see figure 6.2). The second part of this section displays the information related to the hybridisation of the microchip (see figure 6.3).

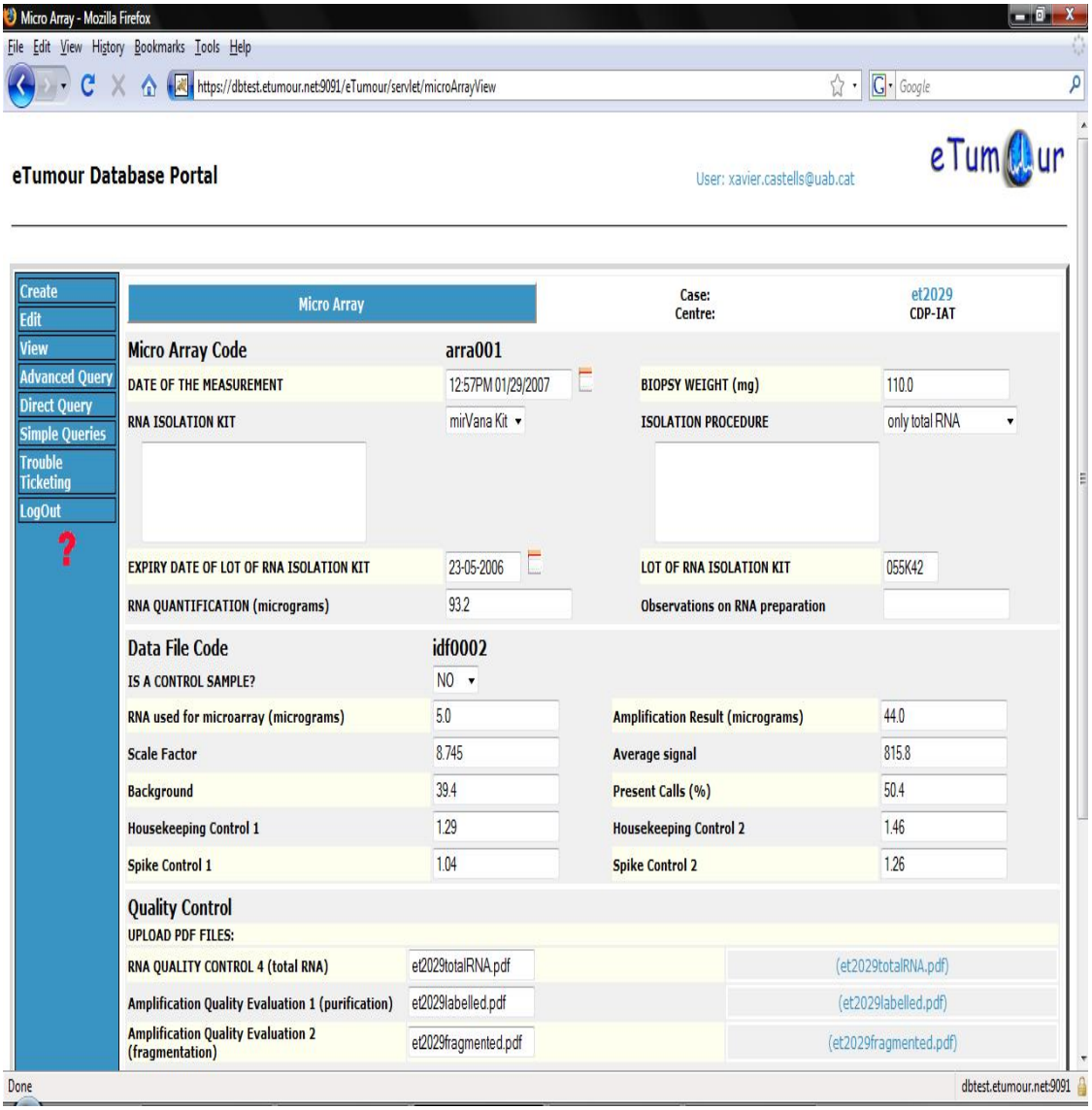


Figure 6.2: *First part of the transcriptomic section of the eTDB.* The information that accounts for the RNA isolation is manually entered into the eTDB.

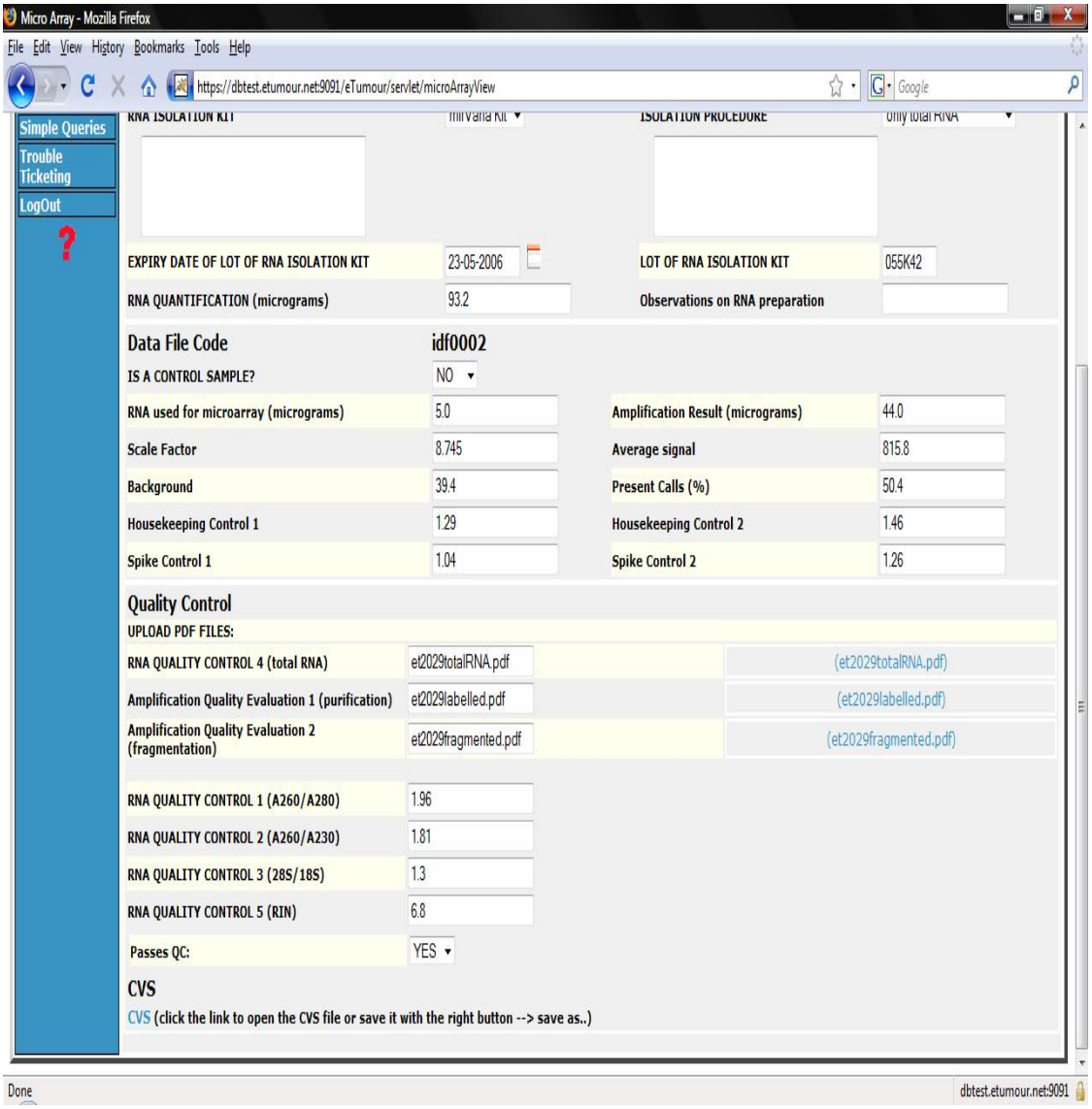


Figure 6.3: *Second part of the transcriptomic section of the eTDB.* The information that accounts for the hybridisation of the microchip is both manually and automatically uploaded into the eTDB. The data that can be extracted from the *.rpt* file is automatically imported (intensity average, SF, percentage of present calls,...).

A-3 Quality control of transcriptomic data in eTUMOUR

A set of parameters to perform automatic validation of microarray data is proposed in this document. Evaluation of the quality of a microarray is carried out at 3 steps of the process:

1) **Quality of RNA isolation (RNA purity and integrity):** no contamination with other macromolecules or organic compounds is checked out through A260/A280 and A230/280 nm ratio of absorbance, measured using a spectrophotomer (see DNA microarrays consensus protocol ("D13-1.doc") included in deliverable 13). Integrity is evaluated through RIN number and 28S/18S ribosomal peak ratio, measured using the Bioanalyzer 2100 (Agilent) (see DNA microarrays consensus protocol ("D13-1.doc") included in deliverable 13).

2) **Quality of amplification:**
   Amplification yield is the quantity of cRNA obtained after amplification of total RNA.

3) **Quality of hybridisation:**

   3.1. Hybridisation descriptive parameters-based validation: parameters considered at this section are average and background signal, scale factor, and number of present calls, which provides general information about the hybridisation.

   3.2. Internal controls-based validation: parameters that provide such information are housekeeping control 1 and 2 (AFFX-HUMRGE/M10098 and AFFX-HUMGDAPGH/M33197 respectively), and spike-in controls 1 and 2 (AFFX-BioB and AFFX-BioC). This set of controls evaluate whether the RNA hybridised is degraded.

The current document is the extended version of the preliminary QC proposal ("D36__27__version_18_12_2005.doc") included into deliverable 36. This preliminary proposal only described the relevant fields for QC purposes, but it did specify neither the ranges, nor the essential fields to validate a case. Next, it is proposed a set of fields for each of the above sections. Each of these fields has a range of values or a threshold associated, which a case must accomplish to be validated.

1) Quality of RNA isolation (RNA purity and integrity):

RNA QUALITY CONTROL 1

A260/A280: 1.6 ≥ ratio ≤ 2 Optimal 1.8

range

RNA QUALITY CONTROL 3

Ribosomal 28S/18S peaks ratio: ratio ≥ 1.2

threshold

RNA QUALITY CONTROL 5

RIN number: value ≥ 6

threshold

These fields are the critical ones to validate a RNA sample before to start labelling. Values of a case for control 1 must be in the indicated range above. At least one of the values for RNA quality controls 3 and 5 must be above the indicated threshold for them.

2) **Quality of amplification**

Amplification yield:

Optimal ≥ 15 micrograms

threshold

3) **Quality of hybridisation (RNA purity and integrity):**

3.1. Hybridisation descriptive parameters-based validation:

Average signal: 700 < value < 1000 fluorescence units

range

Background signal: value < 100 fluorescence units

threshold

Scale factor: Optimal scale factor variation between chips should be as a maximum 3-fold.

threshold

Number of Present calls: 35 < value < 55

range

Unlike background and scale factor parameters, Affymetrix does not recommend a range of values within which the average signal and number of present calls of different microchips can move. Therefore, a range must be established considering data entered into the eTumour DB, which allows removal of only outlier values. The actual range could be changed in the future when more data will be entered into eTDB.

Cases with a scale factor not included in the rank median of scale factors of cases into the eTDB +/- 3 times median of these scale factors are filtered out. Each time a new case is entered into the DB this rank is updated.

3.2. Internal controls-based validation:

Housekeeping control 1

3′/5′ signal ratio for AFFX-HUMRGE/M1009: 0.5 ≥ ratio ≤ 4

range

Housekeeping control 2

3′/5′ signal ratio for AFFX-HUMGDAPGH/M33197: 0.5 ≥ ratio ≤ 4

range

Spike control 1

3′/5′ signal ratio for AFFX-BioB: 0.5 ≥ ratio ≤ 2

range

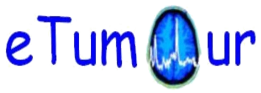
Spike control 2

3′/5′ signal ratio for AFFX-BioC: 0.5 ≥ ratio ≤ 2

range

Note that for housekeeping controls 1 and 2, although only the ratio in the 0.5 to 4 range is allowed, it is proposed that if one of these controls has a ratio above 4, the other must not be above 3 for the sample to be acceptable for classifier development use.

A-4 RNA isolation consensus protocol of the eTUMOUR project



Project no. **LSHC-CT-2004-503094**
Project acronym **E-TUMOUR**
Project title: **WEB ACCESSIBLE MR DECISION SUPPORT SYSTEM FOR BRAIN TUMOUR DIAGNOSIS AND PROGNOSIS, INCORPORATING IN VIVO AND EX VIVO GENOMIC AND METABOLOMIC DATA**

Instrument: **IP**

Thematic Priority
LSH-2002-2.2.0-5 Molecular imaging for early detection of tumours and monitoring of treatment

Task 2.2 Specification of biopsy acquisition, storage and processing protocols for DNA microarray analysis.

D2.3. Month 12. Specification of DNA microarray consensus protocols. Incoming data quality control protocols from CQCD for microarray data.

Due date of deliverable: Month 12
Actual submission date: Month 11

Start date of project: **1 February 2004** Duration: **60 months (until 31 January 2009)**

François Berger – U 318 partner

Modified by Xavier Castells, Anna Barceló and Joaquín Ariño – UAB partner

Date of modification: January 2007

CONTENT

PROTOCOL FOR MICROARRAY PROCESSING

1- Overview of the protocol

This protocol arises from the discussion at the DNA etumour group. Initially different kinds of technologies were evaluated. The choice of Affymetrix for the eTUMOUR protocol [3] was motivated by its high standardization, which provides more accuracy than other transcriptomic technologies for diagnosis purposes. The last human pangenomic DNA chip (Human Genome U133 Plus 2.0 Array) will be used in this project because it provides the largest gene expression set. Using this chip the high resolution scan array 3000 is required.

A single amplification reaction has been agreed because we do not have limitation about the quantity of starting RNA and a linear amplification is ensured. Moreover, we decided to investigate the recently proposed microRNA transcriptional profile. This type of RNA is isolated from total RNA and hybridized onto a specific chip. Characterization of microRNA profile added international competitiveness to the program.

2- RNA Extraction

RNAse free tubes have to be used for all these reactions.

Total RNA will be extracted using mirVana RNA isolation kit (Ambion), which also enables microRNA (miRNA) isolation. mirVana kit is based on an organic extraction followed by immobilization of RNA on glass-fibber filters, allowing purification of both total RNA and RNA enriched in small species.

2-a Lysis and disruption (5 minutes to 15 minutes)

Two different strategies can be used:

(i) Small pieces of frozen tissue are (i.e. <0.5 cm3) homogenized in a potter, being dropped directly into the lysis/binding solution. If this tissue has not been extracted from the same sample used for histological examination, it must be as closer as possible to it. Tissues will be disrupted in 10 volume of lysis buffer into a homogenization vessel on ice (1ml per 0.1 mg of tissue). Dissociation will be done with a motorized rotor-stator homogenizer, or a ground-glass homogenizer for soft tissues.

(ii) Frozen samples are cut on the cryostat. One or two cryostat sections are first kept for further neuropathological analysis (control absence of necrosis, and actual localisation in the tumour area). 0.5 to 250 mg tissue is the optimal range for the kit. Depending of the size of the tumour, 20 to 40-30 µm sections will provide the amount of RNA indispensable for the microarray analysis. Sections are homogenized in the lysis buffer provided by the kit.

For both procedures it is very important to manage tissues quickly and to limit the thawing time. Sample stored in RNA later can be processed. However their fibrous/tough characteristic may impose tissue pulverisation to achieve good cell disruption.

2-b organic extraction (15 minutes)

1/10 volume of miRNA homogenate additive (0.1 ml per ml tissue lysate) is added to the tissue lysate. It is mixed by vortexing several times and is left on ice for 10 minutes. Phenol:chloroform extraction will be done by adding an equivalent volume of phenol:chloroform (300 µl phenol:chloroform per 300 µl homogenate without miRNA homogenate additive), vortex 30-60 seconds, centrifuge 5 minutes at 10000 g at room temperature to separate the aqueous and organic phases. A compact interphase must be observed. At the end the aqueous upper phase is removed without disrupting the lower phase and transfer to a fresh tube.

2-c final RNA isolation (15-20 minutes)

(i) Total RNA isolation procedure
1.25 volumes 100 % ethanol will be added to the aqueous phase (0.125 ml 100 % ethanol per 0.1 ml aqueous phase) and passed through the filter cartridge. Centrifugation at 10000 g allows the mixture to pass through the filter. After 3 washes and elution total RNA is isolated. The volume range for elution must be between 50 and 100 µl of either elution solution or DEPC-treated water.

When RNA concentration is too diluted to carry out labelling, it is allowed performing a precipitation step by adding 2 volumes of absolute ethanol plus 1/10th volume of sodium acetate (from a 3 M stock, pH 5.3) and keeping at -80ºC overnight. Recovery of the RNA must be performed in these steps:

1. Centrifuge the frozen Eppendorf 15 min. at 16100 xg. Discard supernatant and centrifuge 1-2 more minutes. Discard again the supernatant. Keep samples on ice during the rest of the experiment.
2. Add 200 µl of 70% ethanol, vortex and centrifuge 5 min. at 16100 xg. Remove supernatant thoroughly.
3. Keep samples at 37ºC or under vacuum until ethanol is evaporated.
4. Add either DEPC-treated water or elution solution up to obtain the desired concentration and keep it on ice until total resuspension (aprox. 30 minutes).

(ii) Enrichment procedure for small RNAs
1/3 volume of 100 % ethanol is added to aqueous phase (0.33 ml 100 % ethanol per ml aqueous phase). The lower concentration of ethanol, compared to the above protocol enables binding of large RNAs into the filter and absence of retention of small RNA, which are obtained in the flow-through.

treating the filter as described above. microRNA can be purified from the flow-through passing it through an additional filter. Washing and elution is performed as above indicated for total RNA. Thus, two batches of RNA are obtained consisting in total RNA and small RNAs (< 200 nucleotides (nt)).

3- Quality control

Quality control will be performed using the Agilent 2100 analyser system. Absence of degradation is evaluated by the ribosomal RNA 28S/18S ratio, which must be superior to 1.2. Validation of the RNA integrity is mandatory, to subject the RNA to further analysis. To visualize small RNAs, use a 15% denaturing acrylamide gel.
Spectrometric analysis will provide quantification by reading absorbance at 260 and 280 mm.
Small RNA: Concentration=33*A260 µg/ml
Total RNA: Concentration=40*A260 µg/ml
(A260/A280 from 1.8 to 2.1).

4- miRNA analysis by microarray

Microarray probes are 20 nt oligonucleotides with sequences complementary to miRNA. 300 miRNA are present onto the microarray coming from the Sanger miRNA repository and from miRNA recently cloned and published as well as from bioinformatics prediction of new miRNA. Each probe is spotted on a nylon filter using a MWG arrayer. Isolated miRNA are labelled by a kinase incorporating phosphorus 33, hybridized at 55 °C, washed and analyzed using a high resolution phosphor-imager (bass 5000).

5-Expression analysis using Affymetrix system (annexe II and III).

5-1 One cycle amplification

The Genechip used is the Human Genome U133 Plus 2.0 Array. We use the **one-cycle target labelling** kit, which starts from 1 to 15 µg of total RNA (commonly 5 µg works properly). First RNA is reverse transcribed using T7-oligodT promoter primer in the first-strand cDNA synthesis reaction. An RNAse H mediated reaction is performed to degrade RNA, and to enable the second-strand cDNA synthesis. The double strand cDNA is purified and serve as a template for in vitro transcription using T7RNA polymerase. This provides a linear amplification of RNA. Complementary RNA (cRNA) is obtained at the end.
All the reaction are performed in a thermal cycler

We include in the one-cycle cDNA synthesis kit, the preparation of **Poly-A RNA controls** (spike-in-controls) to provide exogenous positive controls to monitor the entire labelling process.

One-cycle amplification includes the following steps:

5.1.1 Starting from 1 to 15 µg (commonly 5 µg properly works) of total RNA first-strand cDNA synthesis is carried out, using a combination of T7-oligodT primer and superscript II enzyme.

5.1.2 Second-strand cDNA synthesis using T4 DNA polymerase

5-2 Target preparation

5-2-1 Sample cleanup of double-strand cDNA on a spin column provided by the kit. RNAse treatment is not recommended at this step.

5-2-2 Synthesis of biotin-labelled cRNA. Genechip in vitro transcription (IVT) labelling kit is used for this step. Overnight IVT reaction is carried out, which has been shown to maximize the labelled cRNA yield with high-quality array results.

5-2-3 Cleanup and quantification of Biotin-labelled cRNA (spectrophotometric analysis at 260 nm, A260/A280 ratio must range from 1.6 to 2.1 to assess adequate purity).

5-2-4 Agilent 2100 bioanalyzer analysis will provide an estimate of the yield, size and distribution of labelled transcripts, providing the last validation step.

5-2-5 Fragmentation of cRNA**.** The cRNA used in the fragmentation procedure must be sufficiently concentrated to maintain a small volume during the procedure.

5-3 Target hybridization

The adequate amount of fragmented cRNA is added to the hybridization buffer. The hybridization cocktail is heated at 99°C, followed by a 45°C step. Centrifugation eliminates the insoluble materials. The probe array is placed into the hybridization oven, set at 45°C after being filled with appropriate volume of the clarified hybridization mixture. Hybridization is performed 16 hours at 45 °C.

5-4 Fluidic station setup, probe array wash and staining

Fluidic station provides an automatic washing.
Setting up, priming the fluidic station, probe array wash, and staining are performed. This step is performed using the appropriate program.
The hybridization cocktail is removed from the probe array. Streptavidin-Phycoerythrin solution, stored in the dark at 4°C, is added at the end. After washing again, the chip is ready for scanning.

5-4 Probe array scan

High resolution scanner is indispensable for the chip. The cartridge is inserted into the scanner and the autofocus is tested. One scan is required with the scanner 3000. The target value must be set to 500.

6-Data analysis (annexe IV)

6-1 quality validation

RNA quality has to be validated as well as cDNA using agilent bioanalyser. Degradation or problem during amplification will induce variability and difficulties in analysis. Hybridization quality, and image quality controlling the entire experiment has to be performed before comparison analysis. Control spike are indispensable for that.

One array per sample will be done, considering that variability is low if quality control is adequate.

6-2 Data analysis

After first order validation, supervised and non supervised analysis will be done, using Eisen clustering analysis. All the data are available as an excel file accessible for our own analysis.

A-5 Statistical analysis of microarray-data in R

R objects

The generic objects in R can be summarised as follows:

- **Vectors:** Data of one dimension, either numeric or character, can be introduced in R as a vector.
- **Matrix:** Numeric data of *nxn* dimensions can be introduced in R as a matrix.
- **Data frame:** When a matrix contains both numeric and character data, R only can deal with it as a data frame.
- **List:** The structure of a list could be pictured as a wardrobe with as many drawers as desired. Moreover each drawer can be filled with any object, of any size.
- **ExpressionSet:** This is a complex object designed to deal with high throughput data. Notably developed to deal with gene-expression microarray datasets, this object enables storage of gene-expression values, gene annotation information and experimental details, among others. Access to each type of data is performed through specific functions.

Moreover, packages have associated objects to perform the analysis. As can be seen below, the *ReadAffy* function from the *affy* package, generates an object of *AffyBatch* class that can be transformed to an *ExpressionSet* object.

Data-preprocessing

The analysis begun by loading the 87 *.cel* files into an R session opened in the remote cluster, previously mentioned in the **"Materials and Methods"** chapter. Once uploaded, an *AffyBatch* object was created:

```
> library(affy)
> UABcel <- ReadAffy()
> UABcel
```

```
AffyBatch object
size of arrays=1164x1164 features (7 kb)
cdf=HG-U133_Plus_2 (54675 affyids)
number of samples=87
number of genes=54675
annotation=hgu133plus2
notes=
```

The *UABcel* object contained the entire probesets without being summarised. As probesets are composed of 11-16 probes (25-mer oligonucleotides), a comparison of the intensities at 5'-end versus those at the 3'-end, can provide an estimation of the integrity of transcripts. The *affy* package enables such a verification through the degradation plots.

For each microarray, the 10 first probes in 5'-3' direction of each probeset are selected. Then the average for each probe across all probesets is computed. Such a computation is performed by the *deg* function and produces a vector of 10 averaged intensities for each microchip. This result can be visualized using the *plotAffyRNAdeg* function.

```
> deg<-AffyRNAdeg(UABcel)
> plotAffyRNAdeg(deg)
```

The *UABcel* object was then subjected to three approaches of data pre-processing:

1. **Robust Microarrays Analysis (RMA)** background correction and **quantile** normalisation.
2. **Microchip Analysis Suite 5 (MAS5)** background correction and **scaling** normalisation.
3. **No** background correction and **scaling** normalisation.

which are implemented in the *affyLM* package:

```
> library(affyPLM)
> eTcelRMAplm <- fitPLM(UABcel, normalize = TRUE,
+ background = TRUE, background.method = "RMA.2",
+ normalize.method = "quantile")
> eTcelMAS5plm <- fitPLM(UABcel, normalize = TRUE,
+ background = TRUE, background.method = "MAS",
+ normalize.method = "scaling")
> eTcelNoBkgplm <- fitPLM(UABcel, normalize = TRUE,
+ background = FALSE, normalize.method = "scaling")
```

The resulting object was of *fitPLM* class, where probes were averaged in each probeset and also averaged those replicated probesets in the microchip. To determine the data-preprocessing method that provided the most unbiased dataset, 5 plots were performed:

```
> RLE(eTcelRMAplm,col="green",xaxt="n",xlab="")
> RLE(eTcelMAS5plm,col="red",xaxt="n",xlab="")
> RLE(eTcelNoBkgplm,col="yellow",xaxt="n",xlab="")
>
> NUSE(eTcelRMAplm,col="green",xaxt="n",xlab="")
> NUSE(eTcelMAS5plm,col="red",xaxt="n",xlab="")
> NUSE(eTcelNoBkgplm,col="yellow",xaxt="n",xlab="")
>
> MAplot(eTcelRMAplm,which=87)
> MAplot(eTcelMAS5plm,which=87)
> MAplot(eTcelNoBkgplm,which=87)
>
> boxplot(eTcelRMAplm,col="green",xaxt="n",xlab="")
> boxplot(eTcelMAS5plm,col="red",xaxt="n",xlab="")
> boxplot(eTcelNoBkgplm,col="yellow",xaxt="n",xlab="")
>
> plotDensity(exprs(PLMset2exprSet(eTcelRMAplm)),
+ col="green",xlab="Log Int (a.u.)")
> plotDensity(exprs(PLMset2exprSet(eTcelMAS5plm)),
+ col="red",xlab="Log Int (a.u.)")
> plotDensity(exprs(PLMset2exprSet(eTcelNoBkgplm)),
+ col="yellow",xlab="Log Int (a.u.)")
```

The **RMA** pre-processing method appeared to be less unbiased, as discussed at the **"Results and Discussion"** chapter. Required by the further processing, the **RMA** data contained in an object of *PLMset* class, was transformed to an object of *ExpressionSet* class.

```
> eTcelRMA<-PLMset2exprSet(eTcelRMAplm)
```

Prior to development of prediction models, a *data.frame* with 5 different annotations for the 54675 probests of the HG-U133 plus2 Affymetrix microchip was created. The incorporated annotations were the *accession number*, the *Affymetrix probeset*, the *gene symbol*, the *locus link* and the *unigene* identifiers.

Using as a reference the *Affymetrix probeset* identifiers, the rest of identifiers were retrieved by applying specific functions from the *annaffy* and *annotate* packages.

```
> library(annotate)
> library(annoaffy)
> eTaffyNames<-featureNames(eTcelRMA)
> eTsymbolNames<-getSYMBOL(eTaffyNames,"hgu133plus2")
> eTLLNames<-getLL(eTaffyNames,"hgu133plus2")
> eTAccNames<-aafGenBank(eTaffyNames,"hgu133plus2")
> eTUnigeneNames<-aafUniGene(eTaffyNames,"hgu133plus2")
>
> HGU133geneID<-cbind(eTAccNames,eTaffyNames,eTsymbolNames
+ ,eTLLNames,eTUnigeneNames)
```

Development of prediction models

Grouping of cases in the 4 main tumour types described in the **"Materials and Methods"** chapter, was performed through a class label vector:

```
> PosUAB<-c('4', '1', '1', '1', '1', '1', '4', '1b', '4',
+ '3', '4', '4gfap', '4', '4', '1', '1', '3', '1', '1',
+ '4', '1b', '1', '1', '4', '3', '1', '3', '0gfap', '4',
+ '4', '0gfap', '4', '3', '0', '3', '3', '4', '4', '1b',
+ '4', '0', '4', '1', '4', '4', '4', '4', '4', '1', '4',
+ '1', '4b', '3b', '0', '4', '4out', '4', '3', '0', '3', '0',
+ '0b', '0', '4', '3', '3', '3', '1', '3', '0b', '4gfap',
+ '1', '0', '0', '3', '3', '3', '0', '4', '0', '0', '5',
+ '4', '4', '4', '1', '0')
```

where **4** represents glioblastomas (Gb), **3** anaplastic gliomas (Ag), **1** low grade gliomas (Lgg) and **0** meningiomas (Mg). The particular cases of the 4 main groups are denoted by "gfap", "out" and "b", which corresponds to unexpected expression values of *GFAP*, sample removed from training and pilocytic astrocytoma, respectively.

The *PosUAB* vector was used to determine the position of each class group within the dataset:

```
> Mg<-which(PosUAB=="0")
> Lgg<-which(PosUAB=="1")
> Aa<-which(PosUAB=="3")
> Gb<-which(PosUAB=="4")
> GbB<-which(PosUAB=="4b")
> GbGfap<-which(PosUAB=="4gfap")
> GbFinal<-c(Gb,GbB,GbGfap)
```

By using the *MultiClassPred* function (described in **annex** A-6), LDA, svm and randF prediction formulas were generated based in all possibles combinations of cross-validation methods, either LOOCV or 5FCV, and feature selection methods, either q-values or PCA. The procedure run in R can be exemplified with the multi-class prediction of Mg, Lgg and Gb cases:

```
> MgLggGb<-c(Mg,Lgg,GbFinal)
> LabelMgLggGb<-c(rep(0,12),rep(1,18),rep(2,31))
>
> MgLggGbLOOCVpval<-MultiClassPred(eTcelRMA[,MgLggGb],
+ LabelMgLggGb, CV="LOOCV",FeatSel="Genes",
+ c(5,10,20,30,40,50),length(LabelMgLggGb))
>
> MgLggGbLOOCVPCA<-MultiClassPred(eTcelRMA[,MgLggGb],
+ LabelMgLggGb, CV="LOOCV",FeatSel="PCA",
+ c(0.4,0.5,0.6,0.7,0.8,0.9),length(LabelMgLggGb))
>
> MgLggGb5FCVpval<-MultiClassPred(eTcelRMA[,MgLggGb],
+ LabelMgLggGb, CV="5FCV",FeatSel="Genes"
+ ,c(5,10,20,30,40,50),305)
>
> MgLggGb5FCVPCA<-MultiClassPred(eTcelRMA[,MgLggGb],
+ LabelMgLggGb, CV="5FCV",FeatSel="PCA"
+ ,c(0.4,0.5,0.6,0.7,0.8,0.9),305)
```

To compute a prediction p-value, the same models were generated but labelling cases randomly. For that, a random vector of classes was generated and prediction models again computed:

```
> LabelMgLggGbRand<-sample(LabelMgLggGb,length(LabelMgLggGb))
>
> MgLggGbLOOCVpvalRand<-MultiClassPred(eTcelRMA[,MgLggGb],
+ LabelMgLggGbRand, CV="LOOCV",FeatSel="Genes",
+ c(5,10,20,30,40,50),length(LabelMgLggGbRand))
>
> MgLggGbLOOCVPCARand<-MultiClassPred(eTcelRMA[,MgLggGb],
+ LabelMgLggGbRand, CV="LOOCV",FeatSel="PCA"
+ ,c(0.4,0.5,0.6,0.7,0.8,0.9),length(LabelMgLggGbRand))
>
> MgLggGb5FCVpvalRand<-MultiClassPred(eTcelRMA[,MgLggGb],
+ LabelMgLggGbRand, CV="5FCV",FeatSel="Genes"
+ ,c(5,10,20,30,40,50),305)
>
```

```
> MgLggGb5FCVPCARand<-MultiClassPred(eTcelRMA[,MgLggGb],
+ LabelMgLggGbRand, CV="5FCV",FeatSel="PCA",
+ c(0.4,0.5,0.6,0.7,0.8,0.9),305)
```

Once all prediction models were computed, the prediction parameters of clinical interest were extracted using the *TableRes* function (see **annex** A-6):

```
> SummaryMgLggGbLOOCVpval<-TableRes(MgLggGbLOOCVpval,
+ MgLggGbLOOCVpvalRand, c("Mg","Lgg","Gb"),
+ CV="LOOCV",c(12,18,31))
>
> SummaryMgLggGbLOOCVPCA<-TableRes(MgLggGbLOOCVPCA,
+ MgLggGbLOOCVPCARand, c("Mg","Lgg","Gb"),
+ CV="LOOCV",c(12,18,31))
>
> SummaryMgLggGb5FCVpval<-TableRes(MgLggGb5FCVpval,
+ MgLggGb5FCVpvalRand, c("Mg","Lgg","Gb"),
+ CV="5FCV",c(12,18,31))
>
> SummaryMgLggGb5FCVPCA<-TableRes(MgLggGb5FCVPCA,
+ MgLggGb5FCVPCARand, c("Mg","Lgg","Gb"),
+ CV="5FCV",c(12,18,31))
```

The output of *TableRes* function is a list that contains three matrices, which correspond to the three prediction method used (LDA, svm or randF). Each matrix provides six parameters of clinical interest, to evaluate the six variables tested as input for the prediction formula:

- The prediction accuracy mean based on the balance error rate (BER).
- The p-value derived from the comparison of prediction values obtained from correctly and random labelling of cases.
- The maximum and minimum prediction accuracy obtained across the performed iterations.
- The sensitivity and specificity for each tumour type when performing a 3-class predictor. In the case of a pairwise predictor, only one sensitivity and specificity are computed.
- The false negative rate (FNR) and false positive rate (FPR). The tumour comparison dependency of FNR and FPR computation is identical to that for the sensitivity and specificity.

From the obtained summary tables, the prediction model of both highest performance and statistical significance was obtained. In the exemplified case of Mg, Lgg and Gb prediction problem, the prediction model based on LOOCV, the number of variables that accounted for 80% of variability, and a dLDA formula yielded the best results. To determine genes that accounted for 80% or more of variability across all the iterations, the *GeneSelFreq* function was applied:

```
> MgLggGbLOOCVPCAgeneList<-GeneSelFreq(MgLggGbLOOCVPCA$FeatSelList,
+ HGU133geneID)
```

As a result, a table with the selection frequency of those genes selected across the LOOCV iterations was obtained. Genes were ordered by decreasing frequency of selection, accompanied with the gene annotations from the *HGU133geneID* object, the ratio of gene-expression for each pairwise comparison and their corresponding p- and q-values. The *GeneSelFreq* worked identically for those models for which feature selection was based on statistical significance of gene expression values.

To note, *GeneSelFreq* was applied to those LOOCV models yielding the highest prediction accuracy mean. The unique selected genes across all tumour comparisons were further analysed (see section 4.2.1.7 and annex A-10).

Computation of glioblastoma subtypes

A hierarchical cluster based on the euclidean distance was performed by using the *heatmap_2* function from the *Heatplus* R package:

```
>ramp <- colorRamp(c("green", "red"))
>rrr<-rgb(ramp(seq(0, 1, length = 256)), max = 255)
>ClusterGb555ps<-heatmap_2(GbeTdataHierClus,legend=1,col=rrr,font.main=8
+,keep.dendro=TRUE)
```

As a first step to verify the reliability of clusters visually detected in the hierarchical cluster, a *k-means* cluster based on the euclidean distance accounting for 2, 3, 4 and 5 clusters was computed:

```
>Distclustps555<-dist(t(GbeTdata[MatchCNIOaffyGeneSymbPos,]),method=
+"euclidean")
>Kmeans2ps555<-kmeans(t(GbeTdata[MatchCNIOaffyGeneSymbPos,]),center=2)
```

```
>Kmeans3ps555<-kmeans(t(GbeTdata[MatchCNIOaffyGeneSymbPos,]),center=3)
>Kmeans4ps555<-kmeans(t(GbeTdata[MatchCNIOaffyGeneSymbPos,]),center=4)
>Kmeans5ps555<-kmeans(t(GbeTdata[MatchCNIOaffyGeneSymbPos,]),center=5)
```

The *silhouette* statistics from the *cluster* R package was computed for each of the generated *k-means* clusters:

```
>SilGb2ps555<-silhouette(Kmeans2ps555$cluster,Distclustps555)
>SilGb3ps555<-silhouette(Kmeans3ps555$cluster,Distclustps555)
>SilGb4ps555<-silhouette(Kmeans4ps555$cluster,Distclustps555)
>SilGb5ps555<-silhouette(Kmeans5ps555$cluster,Distclustps555)
```

The *silhouette* statistics was plotted with the generic *plot* function:

```
>par(mfrow=c(2,2))
>plot(SilGb2ps555,main="Silhouete k-means = 2",cex=0.5)
>plot(SilGb3ps555,main="Silhouete k-means = 3",cex=0.5)
>plot(SilGb4ps555,main="Silhouete k-means = 4",cex=0.5)
>plot(SilGb5ps555,main="Silhouete k-means = 5",cex=0.5)
```

A-6 Description of developed functions

`MultiClassPred`	*Development of prediction models for 2 and 3 class comparisons*

Description

This function enables development of prediction models with reduced overtraining. To this end, models for both 2 and 3 class comparisons are generated by splitting data into training and test using LOOCV or 5FCV, feature selection using PCA components or genes of lowest q-value, and 3 prediction algorithm tested: `lda`, `svm` and `randF`.

To note, feature selection is only performed on training data. For LOOCV, as many iterations as cases contains the whole dataset were performed in this thesis. In contrast, as many iterations as 5 times the number of cases of the whole dataset were performed when using 5FCV. However, the number of iterations can be selected.

Usage

```
MultiClasPred(data,labTr,CV=c(LOOCV,5FCV),FeatSel=c(Genes,PCA),Gnumb,I,N)
```

Description

`data` Object of class `ExpressionSet` containing data to generate prediction models.

`labTr` Numeric vector of class labels for columns of the `data`.

`CV` Cross-validation method. Either `LOOCV` or `5FCV`.

`FeatSel` Feature selection method. Either `Genes` or `PCA`.

`Gnumb` Numeric vector indicating the 6 quantities of variables that will fit the prediction algorithms. If `Genes` is selected as feature selection method, this vector must range from 2 to any desired quantity of genes. In the case of PCA, it must range from a value higher than 0 up to 1.

`I` Number of iterations to perform.

`N` If `Genes` is selected as feature selection method, `N` must correspond to the number of genes with fold-change higher or equal to 2 that will be selected. This parameter can be set as `NULL`.

Value

A list with the below described slots is returned:

`MatPos` Matrix containing as many rows as iterations performed and as many columns as cases in `data`. Each row contains a numeric vector with the column position of samples used for training, followed by the column position of sample(s) left for test.

`GenePos` List with `I` vectors containing the index of selected genes or PCA components in `data`.

`FeatSelList` List with `I` matrices containing the `topTable` (`limma` package) information of selected genes saved in `GenePos`.

`Accuray1-6` Matrix containing as many columns as `I`. Rows are the prediction accuracy mean for each class group of the training set and the test set. A matrix for each quantity of variables and for each prediction algorithm is generated.

`DSC` List of vectors or matrices containing discriminant score(s) of test sample(s) for each prediction algorithm.

TableRes	*Generation of a table with statistics of clinical interest from* MultiClassPred *objects*

Description

This function enables the generation of a summary table from the prediction results obtained by applying MultiClassPred function to a dataset.

Usage

```
TableRes(x,y,z,CV=c(LOOCV,5FCV),label)
```

Description

x MultiClassPred object generated with correct class labels.

y MultiClassPred object generated with random class labels.

z Character vector with the names of considered tumour groups.

CV Cross-validation method used to develop prediction models. Either LOOCV or 5FCV.

label Numeric vector of class labels for columns of the data used as input to MultiClassPred function.

Value

A list of 3 matrices for each prediction algorithm (lda, svm and randF) is returned. Each matrix contains 6 columns that corresponds to the 6 values of tested variables. Rows are statistics of clinical interest:

MeanAc Estimated accuracy mean.

pvalAc P-value based on a Wilcoxon test with the hypothesis that the accuracy mean of randomly labelled models is higher than the correctly labelled models.

Sensitivity Sensitivity estimated from the test samples.

Specificity Specificity estimated from the test samples.

`GeneSelFreq`	*Annotated table of genes selected at least once across the training for a determined tumour comparison using the* `MultiClass` *function*

Description

This function generates a table with annotation of genes selected at least once across the training for a determined tumour comparison using the `MultiClass` function. The selection frequency and the expression ratio between considered tumours is provided with their corresponding p- and q-values.

Usage

```
GeneSelFreq(genes,Annotation,Data,FeatSel=c(Genes,PCA),N,label,Names)
```

Description

`genes` List containing the `GenePos` slot of the prediction model that yields the highest accuracy mean.

`Annotation` Object of class `data.frame` containing annotation data for the type microarray used to obtain `data`. The Affymetrix probeset, the accession number, the locus link and the Unigene identifiers are provided.

`data` The same as in `MultiClassPred` function.

`FeatSel` Feature selection method. Either `Genes` or `PCA`.

`N` Feature selection based on PCA summarises the gene expression levels to a number of components equal to the number of cases (PCA loadings). Thus, `N` indicates the number of genes of highest PCA loadings to be selected.

`label` Numeric vector of class labels for columns of the `data` used as input to `MultiClassPred` function.

`Names` Character vector of names for each class in `label`.

Value

Matrix for which rows are genes selected at least once across the training. Columns are:

`AccN` Accession number of probesets.

`AffyID` Affymetrix probeset identifier.

`GeneSymbol` Gene symbol of probesets.

`LocusLink` Locus link identifier of probesets.

`Unigene` Unigene identifier of probesets.

`FreqSel` Selection frequency of the probeset across the training.

`Ratio` Expression ratio between tumour group(s).

`P-val` P-value of each probe set and tumour comparison(s) based on a Wilcoxon test.

`Q-val` Q-value of each probe set and tumour comparison(s).

A-7 List of the 424 genes selected for the Gbm and Mm tumours predictor using cDNA microarray-data

Gene symbol	Accession number	Gene description	Gbm/Mm ratio	Selection frequency
GFAP	AA069414	Glial fibrillary acidic protein	413	16
PTPRZ1	AA476460	Protein tyrosine phosphatase, receptor-type, Z polypeptide 1	356	9
GPM6B	AA284329	Glycoprotein M6B	133	9
PRELP	AA131664	Proline/arginine-rich end leucine-rich repeat protein	0.042	9
FABP7	W72051	Fatty acid binding protein 7, brain	220	7
EGFL3	AA975413	EGF-like-domain, multiple 3	0.299	7
PDE4B	AA453293	Phosphodiesterase 4B, cAMP-specific (phosphodiesterase E4 dunce homolog, Drosophila)	26	5
OMD	N32201	Osteomodulin	0.011	5
LAPTM4A	AA398233	Lysosomal-associated protein transmembrane 4 alpha	0.346	5
USP25	AA479313	Ubiquitin specific peptidase 25	0.11	5
NFATC3	AA293819	Nuclear factor of activated T-cells, cytoplasmic, calcineurin-dependent 3	0.399	4
CTGF	AA598794	Connective tissue growth factor	0.106	4
PIGT	H83225	Phosphatidylinositol glycan, class T	0.508	4
FLJ39155	R08141	Hypothetical protein FLJ39155	0.019	4
DSP	H90899	Desmoplakin	0.004	4
GAS1	AA025819	Growth arrest-specific 1	0.084	4
PLK1	AA629262	Polo-like kinase 1 (Drosophila)	4.3	4
NEK6	AA463188	NIMA (never in mitosis gene a)-related kinase 6	3.9	4
TNXB	T58430	Similar to tenascin XB isoform 1; tenascin XB1; tenascin XB2; hexabrachion-like [Pan troglodytes]	0.118	4
LHX2	AA018276	LIM homeobox 2	36	4
MGC21621	W52061	MAS-related GPR, member F	0.041	4
PDGFD	AI005125	Platelet derived growth factor D	0.034	4
IL27RA	AI088984	Interleukin 27 receptor alpha	2.1	4
-	AI249137	Transcribed locus	0.205	4
CCND1	R81200	Cyclin D1	0.109	4
NCAM2	AI306467	Neural cell adhesion molecule 2	6.2	4
SH3GL3	AI359676	SH3-domain GRB2-like 3	0.079	4
MT2A	BF131311	Metallothionein 2A	11	4
NUDT1	AA443998	Nudix (nucleoside diphosphate linked moiety X)-type motif 1	2.7	4
RARRES2	AA481944	Retinoic acid receptor responder (tazarotene induced) 2	0.196	3
SMARCD3	AA035796	SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily d, member 3	2.1	3
HYAL1	AA464791	Hyaluronoglucosaminidase 1	0.159	3
CDK2AP1	R78607	CDK2-associated protein 1	2.0	3
CTNND2	H04985	Catenin (cadherin-associated protein), delta 2 (neural plakophilin-related arm-repeat protein)	44	3
CYB5	R92281	Cytochrome b5 type A (microsomal)	0.333	3
CA2	H23187	Carbonic anhydrase II	16	3
OAT	AA446819	Ornithine aminotransferase (gyrate atrophy)	0.383	3
GPM6A	AA448033	Glycoprotein M6A	90	3
HSPC195	R63735	CXXC finger 5	4.2	3
ZMYM6	W81504	Zinc finger, MYM-type 6	0.567	3
PPARGC1A	N89673	Peroxisome proliferative activated receptor, gamma, coactivator 1, alpha	0.257	3
TEK	H02848	TEK tyrosine kinase, endothelial (venous malformations, multiple cutaneous and mucosal)	0.14	3
APM2	AA478298	Chromosome 10 open reading frame 116	0.034	3
FGL2	H56349	Fibrinogen-like 2	0.022	3

CDH2	W49619	Cadherin 2, type 1, N-cadherin (neuronal)	20	3
CANPL1	H15456	Calpain 1, (mu/I) large subunit	0.532	3
GPR17	R44664	G protein-coupled receptor 17	45	3
LOC119504	AA004832	Chromosome 10 open reading frame 104	0.504	3
-	W52340	-	0.355	3
DHRS3	AA171606	Dehydrogenase/reductase (SDR family) member 3	0.217	3
SIPA1L1	AA417567	Signal-induced proliferation-associated 1 like 1	0.351	3
JAM2	AA410345	Junctional adhesion molecule 2	0.094	3
IFITM2	AA862371	Interferon induced transmembrane protein 2 (1-8D)	0.313	3
POLR1C	AA733038	Polymerase (RNA) I polypeptide C, 30kDa	1.9	3
GPC3	AA775872	Glypican 3	0.171	3
TNFSF11	AA504211	Tumor necrosis factor (ligand) superfamily, member 11	0.201	3
NUDC	AA702639	Nuclear distribution gene C homolog (A. nidulans)	2.0	3
TSPAN13	W86202	Tetraspanin 13	13.86	3
EBF	AA488889	Early B-cell factor	0.33	3
TSPYL5	AA626024	TSPY-like 5	0.409	3
-	AA699870	-	0.144	3
LR8	AA987621	LR8 protein	6.1	3
IFITM2	AA985421	Interferon induced transmembrane protein 2 (1-8D)	0.311	3
MT3	AI362950	Metallothionein 3 (growth inhibitory factor (neurotrophic))	30	3
PIK3C2B	AA923518	Phosphoinositide-3-kinase, class 2, beta polypeptide	0.189	3
MAPK8IP1	AI206407	Mitogen-activated protein kinase 8 interacting protein 1	7.6	3
RALGDS	AI131235	Ral guanine nucleotide dissociation stimulator	2.8	3
ELAVL4	AI458073	ELAV (embryonic lethal, abnormal vision, Drosophila)-like 4 (Hu antigen D)	38	3
SOX2OT	AI056507	SOX2 overlapping transcript (non-coding RNA)	30	3
MT2A	AI866473	Metallothionein 2A	15	3
MARCKSL1	AA961735	MARCKS-like 1	6.4	3
PCOLCE	BE259979	Procollagen C-endopeptidase enhancer	0.098	3
C20orf22	R83863	Chromosome 20 open reading frame 22	2.509	3
FOLR1	N91535	Folate receptor 1 (adult)	0.084	3
RTN4RL1	AA453794	Reticulon 4 receptor-like 1	0.104	3
BCAS1	BM666673	Breast carcinoma amplified sequence 1	22	3
TP53I11	BU741540	Tumor protein p53 inducible protein 11	0.07	3
TNFAIP6	W93163	Tumor necrosis factor, alpha-induced protein 6	4.3	2
ZNF286	AA464729	Peroxisome proliferative activated receptor, alpha-like	5.5	2
NID2	AA479199	Nidogen 2 (osteonidogen)	0.098	2
NRCAM	R25521	Neuronal cell adhesion molecule	10	2
UBE2C	AA430504	Ubiquitin-conjugating enzyme E2C	16	2
KCNQ2	H51461	Potassium voltage-gated channel, KQT-like subfamily, member 2	25	2
H2AFX	H95424	H2A histone family, member X	3.2	2
PKP2	H66158	Plakophilin 2	0.022	2
TRIM22	AA083407	Tripartite motif-containing 22	0.383	2
ALDH7A1	AA101299	Aldehyde dehydrogenase 7 family, member A1	0.314	2
ALCAM	R13558	Activated leukocyte cell adhesion molecule	0.092	2
THBS4	AA437064	Thrombospondin 4	11	2
LTBP2	AA424629	Latent transforming growth factor beta binding protein 2	0.035	2
IGFBP6	AA478724	Insulin-like growth factor binding protein 6	0.046	2
CITED1	AA432143	Cbp/p300-interacting transactivator, with Glu/Asp-rich carboxy-terminal domain, 1	19	2
MBTPS1	AA447393	Membrane-bound transcription factor peptidase, site 1	0.527	2
PROCR	T47442	Protein C receptor, endothelial (EPCR)	0.252	2
GCH1	AA443688	GTP cyclohydrolase 1 (dopa-responsive dystonia)	0.321	2
ENO2	AA450189	Enolase 2 (gamma, neuronal)	14	2

RALB	W15297	V-ral simian leukemia viral oncogene homolog B (ras related; GTP binding protein)	2.3	2
FNDC3A	R36431	Fibronectin type III domain containing 3A	0.321	2
SERPINA3	T80924	Serpin peptidase inhibitor, clade A (alpha-1 antiproteinase, antitrypsin), member 3	14	2
ITGA10	H44722	Integrin, alpha 10	0.091	2
FYN	N66144	FYN oncogene related to SRC, FGR, YES	3.7	2
ITM2B	AA453275	Integral membrane protein 2B	0.395	2
SOCS2	AA137031	Suppressor of cytokine signaling 2	3.9	2
PCF11	W73811	PCF11, cleavage and polyadenylation factor subunit, homolog (S. cerevisiae)	0.317	2
PCGF4	T87515	Polycomb group ring finger 4	0.337	2
GOLPH2	AA454597	Golgi phosphoprotein 2	2.6	2
TSPAN4	AA100696	Tetraspanin 4	0.127	2
LOC115098	W69741	Hypothetical protein BC013949	2.2	2
ITM2A	N53447	Integral membrane protein 2A	0.476	2
PRKCD	AA496360	Protein kinase C, delta	0.386	2
ELMO1	H17121	Engulfment and cell motility 1	4.8	2
CNTNAP1	AA028905	Contactin associated protein 1	3.7	2
FCGRT	AA430668	Fc fragment of IgG, receptor, transporter, alpha	0.446	2
CX3CL1	R66139	Chemokine (C-X3-C motif) ligand 1	2.4	2
CRABP1	AA421218	Cellular retinoic acid binding protein 1	0.036	2
BMP4	AA463225	Bone morphogenetic protein 4	0.031	2
ADORA2B	AA055350	Adenosine A2b receptor	0.092	2
TRIM26	AA490855	Tripartite motif-containing 26	0.403	2
MAFG	N21609	V-maf musculoaponeurotic fibrosarcoma oncogene homolog G (avian)	1.5	2
C8orf4	H16793	Chromosome 8 open reading frame 4	26	2
-	N64139	-	0.139	2
C9orf140	AA088458	Chromosome 9 open reading frame 140	15	2
CLIC2	T52201	Chloride intracellular channel 2	0.241	2
EIF4EBP2	H15159	Eukaryotic translation initiation factor 4E binding protein 2	0.283	2
C16orf61	AA181314	Chromosome 16 open reading frame 61	2.2	2
CNTNAP1	H18963	Contactin associated protein 1	3.331	2
TNFRSF11B	AA194983	Tumor necrosis factor receptor superfamily, member 11b (osteoprotegerin)	0.03	2
SPRY2	AA453759	Sprouty homolog 2 (Drosophila)	4.6	2
SFRP2	AA449300	Secreted frizzled-related protein 2	0.212	2
-	R60328	-	1.5	2
RIPK1	AA426324	Receptor (TNFRSF)-interacting serine-threonine kinase 1	0.631	2
-	AA455087	CDNA clone IMAGE:5302158	0.057	2
PRKACB	AA459980	Protein kinase, cAMP-dependent, catalytic, beta	5.0	2
UPP1	AA099568	Uridine phosphorylase 1	11	2
GAS7	R54060	Growth arrest-specific 7	8.7	2
TUBA3	AA865469	Tubulin, alpha 3	6.4	2
MMP2	AA936799	Matrix metallopeptidase 2 (gelatinase A, 72kDa gelatinase, 72kDa type IV collagenase)	0.165	2
ATF1	H54451	Activating transcription factor 1	0.371	2
EIF5	H40023	Eukaryotic translation initiation factor 5	2.3	2
LDB1	AA421335	LIM domain binding 1	0.232	2
PSRC1	N48162	Proline/serine-rich coiled-coil 1	16	2
EFEMP2	AA682527	EGF-containing fibulin-like extracellular matrix protein 2	0.252	2
GPR153	AA777493	G protein-coupled receptor 153	3.1	2
SLC26A2	AA704222	Solute carrier family 26 (sulfate transporter), member 2	0.013	2
FZD4	AA677200	Frizzled homolog 4 (Drosophila)	0.131	2

FBXL7	AA676738	F-box and leucine-rich repeat protein 7	0.216	2
MGMT	AA978354	O-6-methylguanine-DNA methyltransferase	0.307	2
-	H15440	-	44	2
ERBB3	AA664212	V-erb-b2 erythroblastic leukemia viral oncogene homolog 3 (avian)	14	2
-	AA629908	Full-length cDNA clone CS0DN002YM12 of Adult brain of Homo sapiens (human)	4.0	2
DKFZP761M1511	AA776327	Hypothetical protein DKFZP761M1511	3.5	2
DNAJC13	AA778850	DnaJ (Hsp40) homolog, subfamily C, member 13	0.262	2
DUSP22	H42417	Dual specificity phosphatase 22	0.38	2
C10orf116	AA857127	Chromosome 10 open reading frame 116	0.394	2
AP3M1	AA872107	Adaptor-related protein complex 3, mu 1 subunit	0.466	2
FRAP1	AA608530	FK506 binding protein 12-rapamycin associated protein 1	2.5	2
KIAA0391	AA135673	KIAA0391	0.419	2
EDG1	N93476	Endothelial differentiation, sphingolipid G-protein-coupled receptor, 1	8.1	2
RGS16	AA128457	Regulator of G-protein signalling 16	9.1	2
HSPA8	AA620511	Heat shock 70kDa protein 8	1.5	2
TUBB2C	AI000256	Tubulin, beta 2C	2.3	2
ETV1	AA486753	Ets variant gene 1	27	2
TWIST1	AI220198	Twist homolog 1 (acrocephalosyndactyly 3; Saethre-Chotzen syndrome) (Drosophila)	0.173	2
CDH1	AI671174	Cadherin 1, type 1, E-cadherin (epithelial)	0.052	2
BMPR1A	AA991180	Bone morphogenetic protein receptor, type IA	0.415	2
RRAGD	AI095082	Ras-related GTP binding D	7.6	2
CDH11	AI040305	Cadherin 11, type 2, OB-cadherin (osteoblast)	0.23	2
-	AA244506	-	0.238	2
ETV1	AI500327	Ets variant gene 1	20	2
ACTN2	N66231	Actinin, alpha 2	15	2
PRDM2	R73190	PR domain containing 2, with ZNF domain	0.409	2
-	AA744550	-	0.382	2
NCAM1	AA984078	Neural cell adhesion molecule 1	14	2
JUP	AW248439	Junction plakoglobin	0.113	2
-	AW246219	-	3.3	2
CGI-38	BE257080	Brain specific protein	8.2	2
THY1	BE313771	Thy-1 cell surface antigen	4.6	2
MRCL3	BE302683	Myosin regulatory light chain MRCL3	0.414	2
MYL9	BE515089	Myosin, light polypeptide 9, regulatory	0.178	2
CASP9	BE269006	Caspase 9, apoptosis-related cysteine peptidase	1.9	2
TNFAIP8	BE957997	Tumor necrosis factor, alpha-induced protein 8	0.273	2
RRAS2	R64125	Related RAS viral (r-ras) oncogene homolog 2	0.223	2
LAMB2	R73433	Laminin, beta 2 (laminin S)	0.273	2
TMEM64	H22525	Transmembrane protein 64	0.133	2
TUBB2A	R25805	Tubulin, beta 2A	7.2	2
HBEGF	R14663	Heparin-binding EGF-like growth factor	4.9	2
ASTN	R59057	Astrotactin	6.1	2
TGFBR3	H07895	Transforming growth factor, beta receptor III (betaglycan, 300kDa)	0.047	2
TYRO3	BM665421	TYRO3 protein tyrosine kinase	4.2	2
MAD2L2	BM668552	MAD2 mitotic arrest deficient-like 2 (yeast)	3.8	2
ANXA11	BM709344	Annexin A11	0.228	2
CEECAM1	BM712206	Cerebral endothelial cell adhesion molecule 1	3.6	2
IL6ST	BM674517	Interleukin 6 signal transducer (gp130, oncostatin M receptor)	0.517	2
SULT1A1	BE539102	Sulfotransferase family, cytosolic, 1A, phenol-preferring, member 1	0.206	2
MAGEA12	BE542433	Melanoma antigen family A, 12	6.6	2

PTPRG	AW674549	Protein tyrosine phosphatase, receptor type, G	0.467	2
HK1	AA485272	Hexokinase 1	3.2	1
TUBA2	AA426374	Tubulin, alpha 2	5.2	1
COL8A2	AA780815	Collagen, type VIII, alpha 2	0.089	1
RPL10	T67270	Ribosomal protein L10	0.511	1
NFE2L1	AA496576	Nuclear factor (erythroid-derived 2)-like 1	0.31	1
SRI	H60859	Sorcin	4.7	1
CCL2	AA425102	Chemokine (C-C motif) ligand 2	11	1
MAL	AA227594	Mal, T-cell differentiation protein	30	1
LAMP2	N77754	Lysosomal-associated membrane protein 2	3.5	1
KLF4	H45711	Kruppel-like factor 4 (gut)	0.065	1
SPRR2C	AA399674	Small proline-rich protein 2C	0.396	1
RPS6	N91584	Ribosomal protein S6	0.468	1
RAGE	N77779	Renal tumor antigen	0.475	1
SMOX	H93328	Spermine oxidase	2.5	1
ATP6V0B	AA480826	ATPase, H+ transporting, lysosomal 21kDa, V0 subunit b	2.3	1
TOB1	AA490213	Transducer of ERBB2, 1	0.156	1
PIGB	N51166	Phosphatidylinositol glycan, class B	0.234	1
PDGFRL	AA455210	Platelet-derived growth factor receptor-like	0.06	1
SPOCK2	AA398230	Sparc/osteonectin, cwcv and kazal-like domains proteoglycan (testican) 2	6.6	1
ABLIM1	AA406601	Actin binding LIM protein 1	0.155	1
ATP6V0A1	AA427472	ATPase, H+ transporting, lysosomal V0 subunit a1	2.4	1
ALDOC	T77281	Aldolase C, fructose-bisphosphate	17	1
PTPRN	R45941	Protein tyrosine phosphatase, receptor type, N	33	1
CHL1	R40400	Cell adhesion molecule with homology to L1CAM (close homolog of L1)	8.4	1
BNIP3	AA063521	BCL2/adenovirus E1B 19kDa interacting protein 3	5.2	1
TNC	T77595	Tenascin C (hexabrachion)	8.5	1
HLF	W00959	Hepatic leukemia factor	0.097	1
CCL14	R96668	Chemokine (C-C motif) ligand 15	0.192	1
LEPR	H51066	Leptin receptor	0.345	1
MAP7	R77251	Microtubule-associated protein 7	0.079	1
GJB2	AA490688	Gap junction protein, beta 2, 26kDa (connexin 26)	0.026	1
MGST1	AA495936	Microsomal glutathione S-transferase 1	20	1
CHI3L1	AA434115	Chitinase 3-like 1 (cartilage glycoprotein-39)	59	1
BRCA1	H90415	Breast cancer 1, early onset	2.8	1
ANXA4	AA419108	Annexin A4	0.068	1
TUBA1	AA180912	Tubulin, alpha 1 (testis specific)	4.3	1
SDC2	H64346	Syndecan 2 (heparan sulfate proteoglycan 1, cell surface-associated, fibroglycan)	0.042	1
MRLC2	AA487370	Myosin regulatory light chain MRLC2	0.373	1
KIAA0101	W68220	KIAA0101	5.0	1
CUL7	AA479771	Cullin 7	0.339	1
NET1	R24543	Neuroepithelial cell transforming gene 1	0.086	1
PLAGL1	AA463297	Pleiomorphic adenoma gene-like 1	0.332	1
FOXM1	AA129552	Forkhead box M1	10	1
-	AA136125	-	4.9	1
FGFR4	AA446994	Fibroblast growth factor receptor 4	0.178	1
CDH3	AA425556	Cadherin 3, type 1, P-cadherin (placental)	0.114	1
THBS2	H38240	Thrombospondin 2	3.7	1
SLC2A1	H58873	Solute carrier family 2 (facilitated glucose transporter), member 1	3.3	1
-	R70601	Transcribed locus, moderately similar to NP_689672.2 hypothetical protein MGC45438 [Homo sapiens]	0.187	1
EML2	R27680	Echinoderm microtubule associated protein like 2	0.474	1

MLF1	W56360	Myeloid leukemia factor 1	3.0	1
RRAS2	R21415	Related RAS viral (r-ras) oncogene homolog 2	0.206	1
FLJ36748	T66828	Hypothetical protein FLJ36748	0.066	1
-	R63342	-	0.094	1
OLFML1	N55492	Olfactomedin-like 1	0.145	1
GHR	W05000	Growth hormone receptor	0.443	1
-	R25234	-	3.9	1
JAM3	H73479	Junctional adhesion molecule 3	0.065	1
ATAD1	W04668	ATPase family, AAA domain containing 1	0.395	1
OGN	AA045327	Osteoglycin (osteoinductive factor, mimecan)	0.005	1
-	AA026682	-	8.1	1
MED4	AA454015	Mediator of RNA polymerase II transcription, subunit 4 homolog (yeast)	0.513	1
FOXO1A	AA448277	Forkhead box O1A (rhabdomyosarcoma)	0.099	1
CDC42EP2	W81196	CDC42 effector protein (Rho GTPase binding) 2	0.23	1
RIS1	AA127069	Ras-induced senescence 1	6.8	1
-	H72368	-	9.6	1
-	H98688	-	0.553	1
DSCAM	N64532	Down syndrome cell adhesion molecule	9.7	1
LOC132430	N70553	Similar to Polyadenylate-binding protein 4 (Poly(A)-binding protein 4) (PABP 4) (Inducible poly(A)-binding protein) (iPABP) (Activated-platelet protein-1) (APP-1)	0.217	1
ARHGEF5	AA045822	Rho guanine nucleotide exchange factor (GEF) 5	0.175	1
-	AA447514	Transcribed locus, strongly similar to XP_519853.1 PREDICTED: similar to ENSANGP00000014530 [Pan troglodytes]	0.142	1
FZD1	N70776	Frizzled homolog 1 (Drosophila)	0.104	1
CXCL14	W72294	Chemokine (C-X-C motif) ligand 14	27	1
CAMK2N1	AA131299	Calcium/calmodulin-dependent protein kinase II inhibitor 1	3.8	1
NP	AA430382	Nucleoside phosphorylase	1.9	1
MAOB	AA682423	Monoamine oxidase B	45	1
MT1F	T56281	Metallothionein 1F (functional)	27	1
GIYD2	AA398458	Coronin, actin binding protein, 1A pseudogene	0.337	1
CTNNA2	H45976	Catenin (cadherin-associated protein), alpha 2	44	1
DNASE1L3	T73558	Deoxyribonuclease I-like 3	0.005	1
C1orf21	AA676234	Chromosome 1 open reading frame 21	8.4	1
GAD1	AA018457	Glutamate decarboxylase 1 (brain, 67kDa)	10	1
DNM1	AA496334	Dynamin 1	14	1
EXOSC10	AA487064	Exosome component 10	1.8	1
CDK5	AA401479	Cyclin-dependent kinase 5	5.1	1
CLIPR-59	AA488178	CLIP-170-related protein	4.8	1
SPIN	AA428181	Spindlin	0.311	1
P2RY5	R91539	Purinergic receptor P2Y, G-protein coupled, 5	0.355	1
SLC1A2	R15441	Solute carrier family 1 (glial high affinity glutamate transporter), member 2	225	1
-	H20859	-	0.196	1
BAALC	H29251	Brain and acute leukemia, cytoplasmic	62	1
PPP2R2B	H15677	Protein phosphatase 2 (formerly 2A), regulatory subunit B (PR 52), beta isoform	17	1
POLE4	AA400317	Polymerase (DNA-directed), epsilon 4 (p12 subunit)	2.1	1
PLA2R1	W44657	Phospholipase A2 receptor 1, 180kDa	0.099	1
IGSF11	AA490144	Immunoglobulin superfamily, member 11	8.1	1
C20orf23	H23454	Chromosome 20 open reading frame 23	0.231	1
FAM89B	W93891	Family with sequence similarity 89, member B	1.6	1
PSD3	AA460826	Pleckstrin and Sec7 domain containing 3	7.6	1
PSRC2	AA432112	Proline/serine-rich coiled-coil 2	0.459	1
FLJ36090	AA453446	Hypothetical protein FLJ36090	0.081	1

SH3BP5	AA188661	SH3-domain binding protein 5 (BTK-associated)	0.083	1
AGT	H64380	Angiotensinogen (serpin peptidase inhibitor, clade A, member 8)	22	1
DCTD	H68309	DCMP deaminase	0.655	1
PLEKHB1	AA412417	Pleckstrin homology domain containing, family B (evectins) member 1	37	1
PHLDB2	AA479351	Pleckstrin homology-like domain, family B, member 2	0.015	1
C9orf47	AA233892	Endothelial differentiation, sphingolipid G-protein-coupled receptor, 3	6.7	1
TENC1	AA447688	Tensin like C1 domain containing phosphatase (tensin 2)	0.32	1
LOC493869	AA452145	Similar to RIKEN cDNA 2310016C16	0.18	1
FAM11A	R43114	Family with sequence similarity 11, member A	0.547	1
LOC286334	AA425105	Hypothetical protein LOC286334	0.211	1
NBN	AA463450	Nibrin	0.277	1
DNAJC6	AA455940	DnaJ (Hsp40) homolog, subfamily C, member 6	17	1
C2orf17	AA399248	Chromosome 2 open reading frame 17	1.9	1
SLITRK2	R61556	SLIT and NTRK-like family, member 2	14	1
DLK1	AA701996	Delta-like 1 homolog (Drosophila)	1.5	1
GPC6	AA456147	Glypican 6	0.019	1
STMN1	AA873060	Stathmin 1/oncoprotein 18	4.5	1
TLE2	AA873564	Transducin-like enhancer of split 2 (E(sp1) homolog, Drosophila)	0.279	1
ASNS	AA894927	Asparagine synthetase	4.8	1
SORBS3	AA700222	Sorbin and SH3 domain containing 3	0.369	1
PRDX1	AA775803	Peroxiredoxin 1	1.7	1
NET1	H00292	Neuroepithelial cell transforming gene 1	0.311	1
MFSD2	AA774524	Major facilitator superfamily domain containing 2	17	1
RP11-35N6.1	AA700680	Plasticity related gene 3	17	1
ECD	AA701351	Ecdysoneless homolog (Drosophila)	0.418	1
PGM5	AA706788	Phosphoglucomutase 5	0.237	1
TMCC2	AA677167	Transmembrane and coiled-coil domain family 2	10	1
TMEM109	AA504202	Transmembrane protein 109	0.38	1
EHD2	AA708621	EH-domain containing 2	0.314	1
TIMM10	AA670296	Translocase of inner mitochondrial membrane 10 homolog (yeast)	1.7	1
AGTRL1	R58969	Angiotensin II receptor-like 1	21	1
LYPLAL1	AA481256	Lysophospholipase-like 1	0.209	1
ANGPTL2	AA704833	Angiopoietin-like 2	7.0	1
NOV	AA910443	Nephroblastoma overexpressed gene	0.005	1
APLN	AA101878	Apelin, AGTRL1 ligand	17	1
PRKCH	AA128274	Protein kinase C, eta	0.187	1
SCHIP1	AA708955	Schwannomin interacting protein 1	8.4	1
ABHD14B	AA777893	Abhydrolase domain containing 14B	0.25	1
CALCOCO1	AA705325	Calcium binding and coiled-coil domain 1	0.279	1
PCDH1	R77512	Protocadherin 1 (cadherin-like 1)	5.3	1
CTNNA2	R37305	Catenin (cadherin-associated protein), alpha 2	28	1
PSMD14	N67573	Proteasome (prosome, macropain) 26S subunit, non-ATPase, 14	2.3	1
STX12	H91046	Syntaxin 12	2.3	1
FREM1	T96030	FRAS1 related extracellular matrix 1	0.129	1
SUPT5H	AA706107	Suppressor of Ty 5 homolog (S. cerevisiae)	1.7	1
C10orf42	AA884837	Chromosome 10 open reading frame 42	0.338	1
VENTX	AA872096	VENT homeobox homolog (Xenopus laevis)	0.269	1
MARCH5	AA904806	Membrane-associated ring finger (C3HC4) 5	0.162	1
APC2	AA976241	Adenomatosis polyposis coli 2	135	1
PSAT1	AI015679	Phosphoserine aminotransferase 1	8.6	1
KPNA4	AA995784	Karyopherin alpha 4 (importin alpha 3)	1.5	1

MFHAS1	AI017797	Malignant fibrous histiocytoma amplified sequence 1	5.6	1
FRZB	AA454111	Frizzled-related protein	0.063	1
RANGAP1	H98072	Ran GTPase activating protein 1	2.3	1
FZD7	H71474	Frizzled homolog 7 (Drosophila)	0.119	1
CTNNA2	H16079	Catenin (cadherin-associated protein), alpha 2	36	1
DAXX	AA988524	Death-associated protein 6	1.9	1
NCOA6	AI000142	Nuclear receptor coactivator 6	1.5	1
CDA	AA922903	Cytidine deaminase	0.44	1
CCRL2	AI288845	Chemokine (C-C motif) receptor-like 2	7.7	1
AOX1	AI343711	Aldehyde oxidase 1	0.023	1
-	AI361166	Transcribed locus, strongly similar to XP_852136.1 PREDICTED: similar to Spindlin-like protein 2 (SPIN-2) isoform 1 [Canis familiaris]	0.582	1
NOVA1	AI362062	Neuro-oncological ventral antigen 1	7.5	1
PTTG1	AI362866	Pituitary tumor-transforming 1	17	1
RAB33A	AI360342	RAB33A, member RAS oncogene family	5.8	1
CHL1	H15267	Cell adhesion molecule with homology to L1CAM (close homolog of L1)	29	1
SATB1	W72669	Special AT-rich sequence binding protein 1 (binds to nuclear matrix/scaffold-associating DNA's)	3.7	1
IGF1	N67876	Insulin-like growth factor 1 (somatomedin C)	0.146	1
MAML2	AA682512	Mastermind-like 2 (Drosophila)	3.9	1
RAB31	AA432084	RAB31, member RAS oncogene family	1.8	1
ZDHHC5	AI344565	Zinc finger, DHHC-type containing 5	0.467	1
FAM8A1	AI669875	Family with sequence similarity 8, member A1	0.414	1
MCM7	AI688220	MCM7 minichromosome maintenance deficient 7 (S. cerevisiae)	2.9	1
BBC3	AI688112	BCL2 binding component 3	0.328	1
KIAA0408	AI674081	Chromosome 6 open reading frame 174	7.1	1
JAK2	AI376272	Janus kinase 2 (a protein tyrosine kinase)	0.615	1
C10orf11	AA935570	Chromosome 10 open reading frame 11	0.179	1
SCN3A	AA973965	Sodium channel, voltage-gated, type III, alpha	3.8	1
LOC439993	AI000633	LOC439993	0.405	1
RAVER2	AI039422	Hypothetical protein FLJ10770	0.518	1
C9orf32	AI217779	Chromosome 9 open reading frame 32	2.8	1
CTSD	AI285076	Similar to RIKEN cDNA 6330512M04 gene (mouse)	4.1	1
-	AI343669	-	3.8	1
TUBB3	BX100915	Tubulin, beta 3	7.8	1
-	AA989356	CDNA clone IMAGE:4796912	0.385	1
ITGA11	BX119665	Integrin, alpha 11	0.131	1
CYP11A1	T98976	Cytochrome P450, family 11, subfamily A, polypeptide 1	0.112	1
-	W52354	-	0.512	1
RND2	AI027909	Rho family GTPase 2	12	1
TPI1	AA663983	Triosephosphate isomerase 1	1.7	1
-	AA187470	-	0.139	1
ALDH2	AI890849	Aldehyde dehydrogenase 2 family (mitochondrial)	0.178	1
SMAD4	AW410035	SMAD, mothers against DPP homolog 4 (Drosophila)	0.529	1
LAPTM4A	AW411242	Lysosomal-associated protein transmembrane 4 alpha	0.376	1
RAMP1	BE262882	Receptor (calcitonin) activity modifying protein 1	19	1
MGLL	BE261483	Monoglyceride lipase	4.6	1
CCNG1	BE257497	Cyclin G1	0.131	1
MRPL34	BE279280	Mitochondrial ribosomal protein L34	0.696	1
HERPUD1	BE281126	Homocysteine-inducible, endoplasmic reticulum stress-inducible, ubiquitin-like domain member 1	0.26	1
DDB2	BE261143	DNA damage-binding protein 2	0.43	1
ARL4D	BE262902	ADP-ribosylation factor-like 4D	0.082	1

UBE2MP1	BE257314	Hypothetical gene supported by AB012191; BT006754; NM_003969	2.3	1
FCGRT	BE261200	Fc fragment of IgG, receptor, transporter, alpha	0.352	1
-	BE566343	-	3.2	1
CRYAB	BE968687	Crystallin, alpha B	11	1
MT1H	BF674156	Metallothionein 1H	17	1
EFEMP1	T84689	EGF-containing fibulin-like extracellular matrix protein 1	0.355	1
ST6GAL1	H26119	ST6 beta-galactosamide alpha-2,6-sialyltranferase 1	0.287	1
RANGAP1	H52021	Ran GTPase activating protein 1	2.9	1
ENTPD7	H62905	Ectonucleoside triphosphate diphosphohydrolase 7	0.347	1
GLI1	AI473373	Glioma-associated oncogene homolog 1 (zinc finger protein)	0.066	1
GSN	R51491	Gelsolin (amyloidosis, Finnish type)	0.136	1
SCG3	R61070	Secretogranin III	33	1
NBN	H21037	Nibrin	0.187	1
RREB1	AI473516	Ras responsive element binding protein 1	0.215	1
FGD6	AI923117	FYVE, RhoGEF and PH domain containing 6	0.082	1
RGS13	AA767465	Regulator of G-protein signalling 13	0.338	1
HOXD13	AI858239	Homeobox D13	3.2	1
VCL	BM671421	Vinculin	0.215	1
PRKCH	BM668363	Protein kinase C, eta	0.181	1
CDC91L1	BU731832	CDC91 cell division cycle 91-like 1 (S. cerevisiae)	1.5	1
PTPRU	BM667857	Protein tyrosine phosphatase, receptor type, U	0.042	1
IGFBP7	BM676247	Insulin-like growth factor binding protein 7	2.7	1
RASA3	BM715990	RAS p21 protein activator 3	2.0	1
ZNF238	BM677356	Zinc finger protein 238	3.6	1
SP100	BM723015	SP100 nuclear antigen	0.346	1
NUSAP1	BE542067	Nucleolar and spindle associated protein 1	5.2	1
IFITM3	AW675347	Interferon induced transmembrane protein 3 (1-8U)	0.355	1

A-8 Computation of discriminant scores based for discrimination between Gbm and Mm cases

The aim of this appendix is to shortly describe the required computation to obtain the discriminant score (DSC) for a new Gbm or Mm case hybridised onto the CNIO microarray type used in this work. Then, fluorescence signals must be pre-processed and normalised as described in the 3.4.2 section.

A further normalisation is then necessary to transform the fluorescence signals into the variables used for linear discrimination analysis. We can consider a new sample $n(GFAP, PTPRZ1, GPM6B, PRELP)$, where n is the vector containing the normalised fluorescence signals (as described in the article) for *GFAP*, *PTPRZ1*, *GPM6B* and *PRELP* genes, respectively. Also, we have a centering (C) and scaling (S) vectors derived from the developed predictor:

$$C(GFAP, PTPRZ1, GPM6B, PRELP) = (22.59169, 7.58656, 14.77506, 12.11779)$$
$$S(GFAP, PTPRZ1, GPM6B, PRELP) = (5.95949, 4.76408, 4.82909, 2.51886)$$

The final normalised values for each gene are computed as shown:

$$N(GFAP) = \frac{(n(GFAP) - C(GFAP))}{S(GFAP)}$$
$$N(PTPRZ1) = \frac{(n(PTPRZ1) - C(PTPRZ1))}{S(PTPRZ1)}$$
$$N(GPM6B) = \frac{(n(GPM6B) - C(GPM6B))}{S(GPM6B)}$$
$$N(PRELP) = \frac{(n(PRELP) - C(PRELP))}{S(PRELP)}$$

The DSC that would predict a Gbm for negative value of DSC and a Mm for a positive DSC, are computed as follows:

$$DSC = -0.394 * N(GFAP) - 0.397 * N(PTPRZ1)$$
$$-0.397 * N(GPM6B) + 0.365 * N(PRELP)$$

To improve prediction of Affymetrix hybridisation-based cases, we introduced an adjustment to the formula. The centering and scaling vectors are:

$$C(GFAP, PTPRZ1, GPM6B, PRELP) = (8.904749, 7.786328, 9.182813, 7.596874)$$
$$S(GFAP, PTPRZ1, GPM6B, PRELP) = (3.702787, 3.728053, 3.094604, 1.735198)$$

Fluorescence values of each gene are normalised as above and introduced to the discriminant formula:

$$DSC = 0.078 * GFAP - 0.6207 * PTPRZ1$$
$$-0.670 * GPM6B + 0.660 * PRELP$$

A-9 Summary of prediction accuracies for meningioma and glial tumours comparisons

Mg vs Lgg using originating pathologist diagnosis

		Leave-one-out cross validation (LOOCV)						5-fold cross validation (5FCV)					
	Number of variables	% accuracy (AUC)	p-val	Sensitivity (%)	Specificity (%)	FNR (%)	FPR (%)	% accuracy (AUC)	p-val	Sensitivity (%)	Specificity (%)	FNR (%)	FPR (%)
Linear discriminant analysis (LDA)	*2 genes*	100.0	1.1E-05	100.0	100.0	0.0	0.0	99.8	1.0E-08	99.6	100.0	0.4	0.0
	5 genes	100.0	2.6E-05	100.0	100.0	0.0	0.0	99.8	1.0E-06	100.0	100.0	0.0	0.0
	10 genes	100.0	2.6E-05	100.0	100.0	0.0	0.0	100.0	1.0E-06	100.0	100.0	0.0	0.0
	15 genes	100.0	2.6E-05	100.0	100.0	0.0	0.0	100.0	1.0E-06	100.0	100.0	0.0	0.0
	20 genes	100.0	3.0E-01	100.0	100.0	0.0	0.0	98.3	1.0E-08	97.5	99.1	2.5	0.9
	25 genes	100.0	1.1E-05	100.0	100.0	0.0	0.0	100.0	1.0E-06	100.0	100.0	0.0	0.0
	PCA 50% variability	100.0	1.7E-06	100.0	100.0	0.0	0.0	100.0	1.0E-07	100.0	100.0	0.0	0.0
	PCA 60% variability	100.0	6.1E-07	100.0	100.0	0.0	0.0	100.0	1.0E-07	100.0	100.0	0.0	0.0
	PCA 70% variability	100.0	2.6E-05	100.0	100.0	0.0	0.0	100.0	1.0E-07	100.0	100.0	0.0	0.0
	PCA 80% variability	100.0	1.1E-05	100.0	100.0	0.0	0.0	99.8	1.0E-08	100.0	99.6	0.0	0.4
	PCA 85% variability	100.0	8.7E-01	100.0	100.0	0.0	0.0	100.0	1.0E-07	100.0	100.0	0.0	0.0
	PCA 90% variability	95.5	3.0E-03	90.9	100.0	9.1	0.0	98.0	1.0E-07	96.1	100.0	3.9	0.0
Support vector machine (svm)	*2 genes*	100.0	4.3E-06	100.0	100.0	0.0	0.0	99.8	1.0E-08	99.6	100.0	0.4	0.0
	5 genes	100.0	1.4E-04	100.0	100.0	0.0	0.0	99.8	1.0E-06	100.0	100.0	0.0	0.0
	10 genes	100.0	6.3E-04	100.0	100.0	0.0	0.0	100.0	1.0E-06	100.0	100.0	0.0	0.0
	15 genes	100.0	1.3E-03	100.0	100.0	0.0	0.0	100.0	1.0E-06	100.0	100.0	0.0	0.0
	20 genes	100.0	4.0E-01	100.0	100.0	0.0	0.0	100.0	1.0E-06	100.0	100.0	0.0	0.0
	25 genes	100.0	6.3E-04	100.0	100.0	0.0	0.0	100.0	1.0E-06	100.0	100.0	0.0	0.0
	PCA 50% variability	95.5	5.8E-05	90.9	100.0	9.1	0.0	97.5	1.0E-02	95.0	100.0	5.0	0.0
	PCA 60% variability	92.7	1.0E-04	90.9	94.4	9.1	5.6	97.5	1.0E-07	95.0	95.4	5.0	4.6
	PCA 70% variability	88.1	1.9E-03	81.8	94.4	18.2	5.6	90.0	1.0E-07	84.6	95.4	15.4	4.6
	PCA 80% variability	90.9	5.2E-03	81.8	100.0	18.2	0.0	92.0	1.0E-06	83.9	100.0	16.1	0.0
	PCA 85% variability	86.4	9.8E-01	72.7	100.0	27.3	0.0	86.1	1.0E-06	72.1	100.0	27.9	0.0
	PCA 90% variability	77.3	2.3E-02	54.5	100.0	45.5	0.0	76.3	1.0E-10	52.5	100.0	47.5	0.0
Random Forest (randF)	*2 genes*	100.0	4.3E-06	100.0	100.0	0.0	0.0	50.0	5.5E-02	100.0	66.7	0.0	33.3
	5 genes	100.0	6.0E-05	100.0	100.0	0.0	0.0	50.0	5.5E-02	100.0	66.7	0.0	33.3
	10 genes	100.0	6.0E-05	100.0	100.0	0.0	0.0	50.0	3.5E-02	100.0	66.7	0.0	33.3
	15 genes	100.0	1.4E-04	100.0	100.0	0.0	0.0	50.0	3.2E-02	100.0	66.7	0.0	33.3
	20 genes	100.0	8.0E-01	100.0	100.0	0.0	0.0	50.0	3.2E-02	100.0	66.7	0.0	33.3
	25 genes	100.0	6.0E-05	100.0	100.0	0.0	0.0	50.0	3.2E-02	100.0	66.7	0.0	33.3
	PCA 50% variability	100.0	6.0E-05	100.0	100.0	0.0	0.0	50.0	1.1E-03	100.0	66.7	0.0	33.3
	PCA 60% variability	100.0	1.1E-05	100.0	100.0	0.0	0.0	50.0	1.1E-03	98.9	66.7	1.1	33.3
	PCA 70% variability	100.0	1.1E-05	100.0	100.0	0.0	0.0	50.0	9.5E-04	100.0	66.7	0.0	33.3
	PCA 80% variability	100.0	3.0E-04	100.0	100.0	0.0	0.0	50.0	1.2E-03	100.0	66.7	0.0	33.3
	PCA 85% variability	100.0	9.3E-01	100.0	100.0	0.0	0.0	50.4	5.8E-04	100.0	66.8	0.0	33.2
	PCA 90% variability	100.0	1.4E-04	100.0	100.0	0.0	0.0	49.1	5.3E-03	97.5	66.8	2.5	33.2

Mg vs Lgg using eTumour consensus diagnosis

	Leave-one-out cross validation (LOOCV)					5-fold cross validation (5FCV)				
Number of variables	**% accuracy (AUC)**	**p-val**	**Sensitivity (%)**	**Specificity (%)**	**FNR (%)**	**FPR (%)**	**% accuracy (AUC)**	**p-val**	**Sensitivity (%)**	**Specificity (%)**	**FNR (%)**	**FPR (%)**
Linear discriminant analysis (LDA)												
2 genes	100.0	2.9E-06	100.0	100.0	0.0	0.0	99.8	1.0E-08	100.0	99.5	0.0	0.5
5 genes	100.0	1.1E-04	100.0	100.0	0.0	0.0	99.8	1.0E-05	100.0	100.0	0.0	0.0
10 genes	100.0	2.5E-03	100.0	100.0	0.0	0.0	100.0	1.0E-05	100.0	100.0	0.0	0.0
15 genes	100.0	2.5E-03	100.0	100.0	0.0	0.0	99.5	1.0E-07	98.9	100.0	1.1	0.0
20 genes	100.0	4.0E-01	100.0	100.0	0.0	0.0	98.2	1.0E-07	98.6	97.9	1.4	2.1
25 genes	100.0	1.1E-04	100.0	100.0	0.0	0.0	100.0	1.0E-05	100.0	100.0	0.0	0.0
PCA 50% variability	100.0	1.1E-07	100.0	100.0	0.0	0.0	100.0	1.0E-07	100.0	100.0	0.0	0.0
PCA 60% variability	100.0	1.1E-04	100.0	100.0	0.0	0.0	100.0	1.0E-07	100.0	100.0	0.0	0.0
PCA 70% variability	100.0	4.8E-05	100.0	100.0	0.0	0.0	100.0	1.0E-07	100.0	100.0	0.0	0.0
PCA 80% variability	100.0	2.9E-06	100.0	100.0	0.0	0.0	100.0	1.0E-07	100.0	100.0	0.0	0.0
PCA 85% variability	100.0	5.0E-01	100.0	100.0	0.0	0.0	99.1	1.0E-07	98.2	100.0	1.8	0.0
PCA 90% variability	95.5	2.8E-03	90.9	100.0	9.1	0.0	94.9	1.0E-07	91.4	98.3	8.6	1.7
Support vector machine (svm)												
2 genes	100.0	2.0E-05	100.0	100.0	0.0	0.0	99.8	1.0E-08	100.0	99.5	0.0	0.5
5 genes	100.0	4.8E-05	100.0	100.0	0.0	0.0	99.8	1.0E-05	100.0	100.0	0.0	0.0
10 genes	100.0	5.7E-04	100.0	100.0	0.0	0.0	100.0	1.0E-05	100.0	100.0	0.0	0.0
15 genes	100.0	1.2E-03	100.0	100.0	0.0	0.0	100.0	1.0E-05	100.0	100.0	0.0	0.0
20 genes	100.0	9.6E-01	100.0	100.0	0.0	0.0	100.0	1.0E-05	100.0	100.0	0.0	0.0
25 genes	100.0	2.6E-04	100.0	100.0	0.0	0.0	100.0	1.0E-05	100.0	100.0	0.0	0.0
PCA 50% variability	95.5	5.8E-04	90.9	100.0	9.1	0.0	93.4	1.0E-06	86.8	100.0	13.2	0.0
PCA 60% variability	92.1	4.5E-04	90.9	93.3	9.1	6.7	93.4	1.0E-06	86.8	95.0	13.2	5.0
PCA 70% variability	87.6	3.3E-03	81.8	93.3	18.2	6.7	85.9	1.0E-06	76.8	95.0	23.2	5.0
PCA 80% variability	90.9	1.9E-04	81.8	100.0	18.2	0.0	88.1	1.0E-07	77.1	99.0	22.9	1.0
PCA 85% variability	86.4	8.3E-01	72.7	100.0	27.3	0.0	84.1	1.0E-06	68.2	100.0	31.8	0.0
PCA 90% variability	81.8	1.7E-02	63.6	100.0	36.4	0.0	78.3	1.7E-11	57.1	99.5	42.9	0.5
Random Forest (randF)												
2 genes	100.0	7.7E-06	100.0	100.0	0.0	0.0	50.0	2.2E-01	100.0	62.3	0.0	37.7
5 genes	100.0	1.1E-04	100.0	100.0	0.0	0.0	50.0	2.2E-01	100.0	62.5	0.0	37.5
10 genes	100.0	2.0E-05	100.0	100.0	0.0	0.0	50.0	2.7E-01	100.0	62.5	0.0	37.5
15 genes	100.0	2.0E-05	100.0	100.0	0.0	0.0	50.0	3.1E-01	100.0	62.5	0.0	37.5
20 genes	100.0	7.2E-01	100.0	100.0	0.0	0.0	50.0	3.4E-01	100.0	62.5	0.0	37.5
25 genes	100.0	2.9E-06	100.0	100.0	0.0	0.0	50.0	4.2E-01	100.0	62.5	0.0	37.5
PCA 50% variability	100.0	4.8E-05	100.0	100.0	0.0	0.0	50.0	9.2E-01	100.0	62.5	0.0	37.5
PCA 60% variability	100.0	4.8E-05	100.0	100.0	0.0	0.0	50.0	9.2E-01	99.3	62.5	0.7	37.5
PCA 70% variability	100.0	2.0E-05	100.0	100.0	0.0	0.0	50.0	7.8E-01	100.0	62.5	0.0	37.5
PCA 80% variability	100.0	3.4E-07	100.0	100.0	0.0	0.0	48.8	9.6E-01	97.5	62.5	2.5	37.5
PCA 85% variability	100.0	7.2E-01	100.0	100.0	0.0	0.0	50.0	9.5E-01	100.0	62.5	0.0	37.5
PCA 90% variability	100.0	2.0E-05	100.0	100.0	0.0	0.0	50.2	8.3E-01	98.6	62.7	1.4	37.3

Mg vs Ag using originating pathologist diagnosis

		Leave-one-out cross validation (LOOCV)					5-fold cross validation (5FCV)				
	Number of variables	% accuracy (AUC)	p-val	Sensitivity (%)	Specificity (%)	FNR (%)	FPR (%)	% accuracy (AUC)	p-val	Sensitivity (%)	Specificity (%)	FNR (%)	FPR (%)
Linear discriminant analysis (LDA)	*2 genes*	96.9	7.0E-06	100.0	93.8	0.0	6.3	96.7	1.0E-06	99.0	94.5	1.0	5.5
	5 genes	96.9	8.4E-07	100.0	93.8	0.0	6.3	96.7	1.0E-07	99.7	96.6	0.3	3.4
	10 genes	100.0	8.7E-06	100.0	100.0	0.0	0.0	99.0	1.0E-07	99.3	98.6	0.7	1.4
	15 genes	100.0	1.2E-06	100.0	100.0	0.0	0.0	97.4	1.0E-06	97.2	97.5	2.8	2.5
	20 genes	100.0	6.1E-01	100.0	100.0	0.0	0.0	94.3	1.0E-06	93.1	95.4	6.9	4.6
	25 genes	95.5	4.8E-05	90.9	100.0	9.1	0.0	98.9	1.0E-07	99.3	98.4	0.7	1.6
	PCA 50% variability	100.0	5.2E-05	100.0	100.0	0.0	0.0	100.0	1.0E-06	100.0	100.0	0.0	0.0
	PCA 60% variability	100.0	2.2E-05	100.0	100.0	0.0	0.0	100.0	1.0E-06	100.0	100.0	0.0	0.0
	PCA 70% variability	100.0	2.2E-05	100.0	100.0	0.0	0.0	100.0	1.0E-06	100.0	100.0	0.0	0.0
	PCA 80% variability	100.0	8.7E-06	100.0	100.0	0.0	0.0	99.8	1.0E-07	100.0	99.5	0.0	0.5
	PCA 85% variability	96.9	8.0E-01	100.0	93.8	0.0	6.3	97.5	1.0E-06	100.0	94.9	0.0	5.1
	PCA 90% variability	100.0	1.2E-06	100.0	100.0	0.0	0.0	98.1	1.0E-07	96.9	99.3	3.1	0.7
Support vector machine (svm)	*2 genes*	96.9	2.2E-08	100.0	93.8	0.0	6.3	98.0	1.0E-07	99.7	96.3	0.3	3.7
	5 genes	96.9	7.8E-08	100.0	93.8	0.0	6.3	98.0	1.0E-06	100.0	96.6	0.0	3.4
	10 genes	100.0	4.3E-08	100.0	100.0	0.0	0.0	99.3	1.0E-06	100.0	98.6	0.0	1.4
	15 genes	100.0	1.3E-08	100.0	100.0	0.0	0.0	99.7	1.0E-07	100.0	99.3	0.0	0.7
	20 genes	100.0	8.8E-01	100.0	100.0	0.0	0.0	99.7	1.0E-07	100.0	99.3	0.0	0.7
	25 genes	100.0	1.4E-07	100.0	100.0	0.0	0.0	99.9	1.0E-07	100.0	99.8	0.0	0.2
	PCA 50% variability	100.0	2.2E-05	100.0	100.0	0.0	0.0	100.0	1.0E-05	100.0	100.0	0.0	0.0
	PCA 60% variability	95.5	1.2E-04	90.9	100.0	9.1	0.0	100.0	1.0E-05	88.6	100.0	11.4	0.0
	PCA 70% variability	95.5	2.7E-04	90.9	100.0	9.1	0.0	94.1	1.0E-05	88.3	100.0	11.7	0.0
	PCA 80% variability	90.9	2.1E-04	81.8	100.0	18.2	0.0	89.0	1.0E-05	77.9	100.0	22.1	0.0
	PCA 85% variability	90.9	8.9E-01	81.8	100.0	18.2	0.0	87.0	1.0E-07	74.5	99.5	25.5	0.5
	PCA 90% variability	81.8	2.2E-03	63.6	100.0	36.4	0.0	78.0	1.1E-14	57.2	98.9	42.8	1.1
Random Forest (randF)	*2 genes*	92.3	4.1E-07	90.9	93.8	9.1	6.3	50.4	5.5E-01	99.7	63.7	0.3	36.3
	5 genes	96.9	2.5E-06	100.0	93.8	0.0	6.3	50.4	5.5E-01	100.0	63.4	0.0	36.6
	10 genes	100.0	1.2E-06	100.0	100.0	0.0	0.0	50.0	5.8E-01	100.0	63.6	0.0	36.4
	15 genes	100.0	8.7E-06	100.0	100.0	0.0	0.0	50.0	5.9E-01	100.0	63.8	0.0	36.2
	20 genes	100.0	9.3E-01	100.0	100.0	0.0	0.0	50.0	5.3E-01	100.0	63.8	0.0	36.2
	25 genes	100.0	4.2E-07	100.0	100.0	0.0	0.0	50.0	5.5E-01	100.0	63.9	0.0	36.1
	PCA 50% variability	100.0	2.7E-04	100.0	100.0	0.0	0.0	50.0	4.7E-01	100.0	64.0	0.0	36.0
	PCA 60% variability	100.0	3.3E-06	100.0	100.0	0.0	0.0	50.0	4.7E-01	100.0	64.0	0.0	36.0
	PCA 70% variability	100.0	5.2E-05	100.0	100.0	0.0	0.0	50.0	7.2E-01	100.0	64.0	0.0	36.0
	PCA 80% variability	100.0	8.7E-06	100.0	100.0	0.0	0.0	49.5	6.1E-01	99.0	64.0	1.0	36.0
	PCA 85% variability	100.0	8.0E-01	100.0	100.0	0.0	0.0	49.3	7.0E-01	98.6	64.0	1.4	36.0
	PCA 90% variability	95.5	2.5E-06	90.9	100.0	9.1	0.0	45.0	9.9E-01	90.0	64.0	10.0	36.0

Mg vs Ag using eTumour consensus diagnosis

		Leave-one-out cross validation (LOOCV)						5-fold cross validation (5FCV)					
	Number of variables	% accuracy (AUC)	p-val	Sensitivity (%)	Specificity (%)	FNR (%)	FPR (%)	% accuracy (AUC)	p-val	Sensitivity (%)	Specificity (%)	FNR (%)	FPR (%)
Linear discriminant analysis (LDA)	*2 genes*	92.7	5.6E-04	90.9	94.4	9.1	5.6	96.0	1.0E-06	97.7	94.4	2.3	5.6
	5 genes	97.2	1.3E-06	100.0	94.4	0.0	5.6	96.0	1.0E-06	98.1	96.5	1.9	3.5
	10 genes	100.0	1.7E-06	100.0	100.0	0.0	0.0	98.7	1.0E-07	99.4	98.1	0.6	1.9
	15 genes	100.0	4.3E-06	100.0	100.0	0.0	0.0	99.1	1.0E-06	100.0	98.2	0.0	1.8
	20 genes	100.0	3.0E-01	100.0	100.0	0.0	0.0	97.7	1.0E-06	97.1	98.2	2.9	1.8
	25 genes	100.0	2.2E-07	100.0	100.0	0.0	0.0	99.0	1.0E-07	98.7	99.4	1.3	0.6
	PCA 50% variability	100.0	4.3E-06	100.0	100.0	0.0	0.0	100.0	1.0E-06	100.0	100.0	0.0	0.0
	PCA 60% variability	100.0	3.0E-04	100.0	100.0	0.0	0.0	100.0	1.0E-06	100.0	100.0	0.0	0.0
	PCA 70% variability	100.0	6.3E-04	100.0	100.0	0.0	0.0	99.6	1.0E-07	100.0	99.2	0.0	0.8
	PCA 80% variability	100.0	1.1E-05	100.0	100.0	0.0	0.0	99.8	1.0E-06	100.0	99.5	0.0	0.5
	PCA 85% variability	97.2	7.1E-01	100.0	94.4	0.0	5.6	96.1	1.0E-06	96.5	95.8	3.5	4.2
	PCA 90% variability	100.0	4.3E-06	100.0	100.0	0.0	0.0	98.1	1.0E-07	99.4	96.8	0.6	3.2
Support vector machine (svm)	*2 genes*	97.2	3.5E-06	100.0	94.4	0.0	5.6	96.9	1.0E-06	98.7	95.2	1.3	4.8
	5 genes	97.2	3.5E-06	100.0	94.4	0.0	5.6	96.9	1.0E-07	99.4	96.9	0.6	3.1
	10 genes	100.0	4.3E-06	100.0	100.0	0.0	0.0	99.2	1.0E-06	100.0	98.4	0.0	1.6
	15 genes	100.0	4.3E-06	100.0	100.0	0.0	0.0	99.6	1.0E-07	100.0	99.2	0.0	0.8
	20 genes	100.0	1.0E+00	100.0	100.0	0.0	0.0	99.6	1.0E-07	100.0	99.2	0.0	0.8
	25 genes	100.0	6.0E-05	100.0	100.0	0.0	0.0	99.8	1.0E-06	100.0	99.5	0.0	0.5
	PCA 50% variability	100.0	3.0E-04	100.0	100.0	0.0	0.0	100.0	1.0E-05	100.0	100.0	0.0	0.0
	PCA 60% variability	95.5	1.2E-02	90.9	100.0	9.1	0.0	100.0	1.0E-06	91.6	100.0	8.4	0.0
	PCA 70% variability	90.9	2.0E-02	81.8	100.0	18.2	0.0	89.8	1.0E-05	79.7	100.0	20.3	0.0
	PCA 80% variability	90.9	5.2E-03	81.8	100.0	18.2	0.0	88.2	1.0E-05	76.5	100.0	23.5	0.0
	PCA 85% variability	81.8	9.9E-01	63.6	100.0	36.4	0.0	82.3	1.0E-07	64.8	99.8	35.2	0.2
	PCA 90% variability	72.7	4.7E-02	45.5	100.0	54.5	0.0	76.8	8.3E-11	53.9	99.7	46.1	0.3
Random Forest (randF)	*2 genes*	92.7	1.0E-04	90.9	94.4	9.1	5.6	50.1	1.4E-02	99.7	66.2	0.3	33.8
	5 genes	97.2	3.5E-06	100.0	94.4	0.0	5.6	50.1	1.4E-02	100.0	66.1	0.0	33.9
	10 genes	100.0	1.1E-05	100.0	100.0	0.0	0.0	50.0	1.2E-02	100.0	66.3	0.0	33.7
	15 genes	100.0	1.7E-06	100.0	100.0	0.0	0.0	50.0	1.4E-02	100.0	66.3	0.0	33.7
	20 genes	100.0	9.6E-01	100.0	100.0	0.0	0.0	50.0	1.6E-02	100.0	66.4	0.0	33.6
	25 genes	100.0	2.6E-05	100.0	100.0	0.0	0.0	50.0	1.6E-02	100.0	66.4	0.0	33.6
	PCA 50% variability	100.0	6.3E-04	100.0	100.0	0.0	0.0	50.0	1.1E-01	100.0	66.7	0.0	33.3
	PCA 60% variability	100.0	3.0E-04	100.0	100.0	0.0	0.0	50.0	1.1E-01	100.0	66.7	0.0	33.3
	PCA 70% variability	100.0	3.0E-04	100.0	100.0	0.0	0.0	49.7	3.2E-01	99.4	66.7	0.6	33.3
	PCA 80% variability	100.0	1.1E-05	100.0	100.0	0.0	0.0	50.0	2.5E-01	100.0	66.7	0.0	33.3
	PCA 85% variability	100.0	9.6E-01	100.0	100.0	0.0	0.0	49.4	3.4E-01	98.7	66.7	1.3	33.3
	PCA 90% variability	100.0	4.3E-06	100.0	100.0	0.0	0.0	48.4	4.9E-01	96.8	66.7	3.2	33.3

Mg vs Gb using originating pathologist diagnosis

	Number of variables	Leave-one-out cross validation (LOOCV)					5-fold cross validation (5FCV)				
		% accuracy (AUC)	p-val	Sensitivity (%)	Specificity (%)	FNR (%)	FPR (%)	% accuracy (AUC)	p-val	Sensitivity (%)	Specificity (%)	FNR (%)	FPR (%)
Linear discriminant analysis (LDA)	*2 genes*	95.5	9.9E-04	90.9	100.0	9.1	0.0	94.3	1.0E-05	88.6	100.0	11.4	0.0
	5 genes	95.5	2.6E-05	90.9	100.0	9.1	0.0	94.3	1.0E-06	95.3	100.0	4.7	0.0
	10 genes	100.0	5.1E-06	100.0	100.0	0.0	0.0	99.7	1.0E-07	99.3	100.0	0.7	0.0
	20 genes	100.0	1.8E-07	100.0	100.0	0.0	0.0	99.9	1.0E-07	99.8	100.0	0.2	0.0
	30 genes	95.5	1.2E-01	90.9	100.0	9.1	0.0	95.5	1.0E-07	91.2	99.8	8.8	0.2
	40 genes	93.8	8.1E-07	90.9	96.8	9.1	3.2	98.5	1.0E-07	97.2	99.8	2.8	0.2
	PCA 50% variability	95.8	1.0E-04	91.7	100.0	8.3	0.0	100.0	1.0E-08	100.0	100.0	0.0	0.0
	PCA 60% variability	95.8	5.1E-06	91.7	100.0	8.3	0.0	100.0	1.0E-08	100.0	100.0	0.0	0.0
	PCA 70% variability	95.8	2.3E-06	91.7	100.0	8.3	0.0	100.0	1.0E-08	100.0	100.0	0.0	0.0
	PCA 80% variability	95.8	1.8E-07	91.7	100.0	8.3	0.0	100.0	1.0E-08	100.0	100.0	0.0	0.0
	PCA 85% variability	95.8	5.0E-01	91.7	100.0	8.3	0.0	100.0	1.0E-08	100.0	100.0	0.0	0.0
	PCA 90% variability	95.8	7.4E-08	91.7	100.0	8.3	0.0	99.3	1.0E-08	98.6	99.9	1.4	0.1
Support vector machine (svm)	*2 genes*	95.5	5.0E-04	90.9	100.0	9.1	0.0	97.9	1.0E-06	95.8	100.0	4.2	0.0
	5 genes	95.5	2.5E-04	90.9	100.0	9.1	0.0	97.9	1.0E-06	95.8	100.0	4.2	0.0
	10 genes	100.0	2.1E-04	100.0	100.0	0.0	0.0	100.0	1.0E-06	100.0	100.0	0.0	0.0
	20 genes	100.0	1.0E-04	100.0	100.0	0.0	0.0	100.0	1.0E-06	100.0	100.0	0.0	0.0
	30 genes	100.0	1.0E+00	100.0	100.0	0.0	0.0	100.0	1.0E-06	100.0	100.0	0.0	0.0
	40 genes	100.0	2.1E-04	100.0	100.0	0.0	0.0	100.0	1.0E-06	100.0	100.0	0.0	0.0
	PCA 50% variability	91.7	9.9E-04	83.3	100.0	16.7	0.0	95.7	1.0E-06	91.4	100.0	8.6	0.0
	PCA 60% variability	87.5	1.8E-03	75.0	100.0	25.0	0.0	95.7	1.0E-05	82.8	100.0	17.2	0.0
	PCA 70% variability	87.5	1.8E-03	75.0	100.0	25.0	0.0	91.4	1.0E-05	82.8	100.0	17.2	0.0
	PCA 80% variability	87.5	3.6E-03	75.0	100.0	25.0	0.0	88.5	1.0E-05	77.0	100.0	23.0	0.0
	PCA 85% variability	79.2	1.0E+00	58.3	100.0	41.7	0.0	80.8	1.0E-05	61.6	100.0	38.4	0.0
	PCA 90% variability	66.7	1.5E-01	33.3	100.0	66.7	0.0	64.0	4.6E-05	27.9	100.0	72.1	0.0
Random Forest (randF)	*2 genes*	90.9	2.4E-05	81.8	100.0	18.2	0.0	52.4	2.0E-02	99.8	78.3	0.2	21.7
	5 genes	95.5	5.4E-06	90.9	100.0	9.1	0.0	52.4	2.0E-02	100.0	77.8	0.0	22.2
	10 genes	100.0	2.3E-06	100.0	100.0	0.0	0.0	50.0	4.9E-01	100.0	77.5	0.0	22.5
	20 genes	100.0	1.1E-05	100.0	100.0	0.0	0.0	50.0	5.0E-01	100.0	77.5	0.0	22.5
	30 genes	100.0	9.5E-01	100.0	100.0	0.0	0.0	50.0	5.0E-01	100.0	77.5	0.0	22.5
	40 genes	100.0	1.0E-04	100.0	100.0	0.0	0.0	50.0	5.0E-01	100.0	77.5	0.0	22.5
	PCA 50% variability	95.8	2.4E-05	91.7	100.0	8.3	0.0	50.0	2.7E-01	100.0	77.5	0.0	22.5
	PCA 60% variability	95.8	1.0E-04	91.7	100.0	8.3	0.0	50.0	2.7E-01	100.0	77.5	0.0	22.5
	PCA 70% variability	95.8	5.0E-05	91.7	100.0	8.3	0.0	50.0	4.2E-01	100.0	77.5	0.0	22.5
	PCA 80% variability	95.8	5.0E-05	91.7	100.0	8.3	0.0	50.1	4.5E-01	100.0	77.5	0.0	22.5
	PCA 85% variability	95.8	1.0E+00	91.7	100.0	8.3	0.0	50.4	3.5E-01	100.0	77.6	0.0	22.4
	PCA 90% variability	95.8	1.0E-04	91.7	100.0	8.3	0.0	50.5	3.2E-01	100.0	77.7	0.0	22.3

Mg vs Gb using eTumour consensus diagnosis

		Leave-one-out cross validation (LOOCV)					5-fold cross validation (5FCV)				
	Number of variables	% accuracy (AUC)	p-val	Sensitivity (%)	Specificity (%)	FNR (%)	FPR (%)	% accuracy (AUC)	p-val	Sensitivity (%)	Specificity (%)	FNR (%)	FPR (%)
Linear discriminant analysis (LDA)	*2 genes*	95.5	5.9E-05	90.9	100.0	9.1	0.0	94.7	1.0E-05	89.3	100.0	10.7	0.0
	5 genes	95.5	1.2E-04	90.9	100.0	9.1	0.0	94.7	1.0E-06	96.1	100.0	3.9	0.0
	10 genes	100.0	2.4E-06	100.0	100.0	0.0	0.0	99.8	1.0E-07	99.5	100.0	0.5	0.0
	20 genes	100.0	5.4E-06	100.0	100.0	0.0	0.0	100.0	1.0E-05	100.0	100.0	0.0	0.0
	30 genes	95.5	3.1E-01	90.9	100.0	9.1	0.0	96.6	1.0E-06	93.6	99.5	6.4	0.5
	40 genes	87.8	1.4E-04	81.8	93.8	18.2	6.3	98.1	1.0E-07	96.6	99.7	3.4	0.3
	PCA 50% variability	100.0	2.5E-05	100.0	100.0	0.0	0.0	100.0	1.0E-07	100.0	100.0	0.0	0.0
	PCA 60% variability	100.0	5.4E-06	100.0	100.0	0.0	0.0	100.0	1.0E-07	100.0	100.0	0.0	0.0
	PCA 70% variability	100.0	5.4E-06	100.0	100.0	0.0	0.0	100.0	1.0E-07	100.0	100.0	0.0	0.0
	PCA 80% variability	100.0	1.1E-06	100.0	100.0	0.0	0.0	100.0	1.0E-07	100.0	100.0	0.0	0.0
	PCA 85% variability	100.0	4.1E-01	100.0	100.0	0.0	0.0	100.0	1.0E-07	100.0	100.0	0.0	0.0
	PCA 90% variability	100.0	8.3E-08	100.0	100.0	0.0	0.0	99.7	1.0E-07	99.8	99.7	0.2	0.3
Support vector machine (svm)	*2 genes*	95.5	5.9E-05	90.9	100.0	9.1	0.0	97.7	1.0E-06	95.5	100.0	4.5	0.0
	5 genes	95.5	2.5E-04	90.9	100.0	9.1	0.0	97.7	1.0E-06	96.8	100.0	3.2	0.0
	10 genes	100.0	5.1E-05	100.0	100.0	0.0	0.0	100.0	1.0E-06	100.0	100.0	0.0	0.0
	20 genes	100.0	5.1E-05	100.0	100.0	0.0	0.0	100.0	1.0E-06	100.0	100.0	0.0	0.0
	30 genes	100.0	1.0E+00	100.0	100.0	0.0	0.0	100.0	1.0E-06	100.0	100.0	0.0	0.0
	40 genes	100.0	1.1E-04	100.0	100.0	0.0	0.0	100.0	1.0E-06	100.0	100.0	0.0	0.0
	PCA 50% variability	95.5	1.0E-03	90.9	100.0	9.1	0.0	96.7	1.0E-06	93.4	100.0	6.6	0.0
	PCA 60% variability	90.9	3.6E-03	81.8	100.0	18.2	0.0	96.7	1.0E-05	82.7	100.0	17.3	0.0
	PCA 70% variability	90.9	1.9E-03	81.8	100.0	18.2	0.0	91.3	1.0E-02	82.5	100.0	17.5	0.0
	PCA 80% variability	90.9	3.6E-03	81.8	100.0	18.2	0.0	87.2	1.0E-05	74.3	100.0	25.7	0.0
	PCA 85% variability	81.8	1.0E+00	63.6	100.0	36.4	0.0	78.1	2.2E-16	56.1	100.0	43.9	0.0
	PCA 90% variability	68.2	1.5E-01	36.4	100.0	63.6	0.0	59.2	4.1E-03	18.4	100.0	81.6	0.0
Random Forest (randF)	*2 genes*	95.5	5.9E-05	90.9	100.0	9.1	0.0	52.7	3.0E-04	100.0	78.8	0.0	21.2
	5 genes	95.5	2.5E-04	90.9	100.0	9.1	0.0	52.7	3.0E-04	100.0	78.3	0.0	21.7
	10 genes	100.0	2.5E-05	100.0	100.0	0.0	0.0	50.0	4.3E-03	100.0	78.0	0.0	22.0
	20 genes	100.0	1.1E-04	100.0	100.0	0.0	0.0	50.0	3.5E-03	100.0	78.0	0.0	22.0
	30 genes	100.0	9.5E-01	100.0	100.0	0.0	0.0	50.0	3.5E-03	100.0	78.0	0.0	22.0
	40 genes	100.0	1.1E-04	100.0	100.0	0.0	0.0	50.0	3.5E-03	100.0	78.0	0.0	22.0
	PCA 50% variability	100.0	4.2E-04	100.0	100.0	0.0	0.0	50.0	1.2E-02	100.0	78.0	0.0	22.0
	PCA 60% variability	100.0	5.1E-05	100.0	100.0	0.0	0.0	50.0	1.2E-02	100.0	78.1	0.0	21.9
	PCA 70% variability	100.0	1.1E-04	100.0	100.0	0.0	0.0	50.0	1.3E-02	100.0	78.0	0.0	22.0
	PCA 80% variability	100.0	4.2E-04	100.0	100.0	0.0	0.0	50.1	1.3E-02	100.0	78.1	0.0	21.9
	PCA 85% variability	100.0	1.0E+00	100.0	100.0	0.0	0.0	51.2	2.4E-03	100.0	78.4	0.0	21.6
	PCA 90% variability	95.5	1.0E-03	90.9	100.0	9.1	0.0	50.5	7.7E-03	98.6	78.4	1.4	21.6

Lgg vs Ag using originating pathologist diagnosis

		Leave-one-out cross validation (LOOCV)					5-fold cross validation (5FCV)				
	Number of variables	% accuracy (AUC)	p-val	Sensitivity (%)	Specificity (%)	FNR (%)	FPR (%)	% accuracy (AUC)	p-val	Sensitivity (%)	Specificity (%)	FNR (%)	FPR (%)
Linear discriminant analysis (LDA)	*2 genes*	67.7	7.5E-01	66.7	68.8	33.3	31.3	71.9	1.9E-01	74.4	69.3	25.6	30.7
	5 genes	55.2	9.1E-01	66.7	43.8	33.3	56.3	68.6	1.9E-01	72.6	64.6	27.4	35.4
	10 genes	54.9	8.3E-01	72.2	37.5	27.8	62.5	67.7	1.3E-03	69.9	65.5	30.1	34.5
	20 genes	64.6	3.9E-01	66.7	62.5	33.3	37.5	62.7	3.0E-05	63.0	62.5	37.0	37.5
	30 genes	55.9	3.3E-01	55.6	56.3	44.4	43.8	62.1	6.6E-05	61.7	62.5	38.3	37.5
	40 genes	61.5	4.1E-01	66.7	56.3	33.3	43.8	68.1	5.9E-09	67.0	69.1	33.0	30.9
	PCA 50% variability	45.8	9.6E-01	66.7	25.0	33.3	75.0	47.3	1.0E+00	65.6	29.0	34.4	71.0
	PCA 60% variability	42.7	9.6E-01	66.7	18.8	33.3	81.3	44.6	1.0E+00	59.7	29.5	40.3	70.5
	PCA 70% variability	67.4	1.2E-01	72.2	62.5	27.8	37.5	70.0	6.4E-12	73.4	66.7	26.6	33.3
	PCA 80% variability	61.1	1.9E-01	72.2	50.0	27.8	50.0	61.8	1.9E-05	68.6	55.0	31.4	45.0
	PCA 85% variability	70.5	7.5E-01	72.2	68.8	27.8	31.3	68.8	3.1E-12	71.0	66.7	29.0	33.3
	PCA 90% variability	61.5	1.7E-01	66.7	56.3	33.3	43.8	65.2	8.9E-16	66.7	63.6	33.3	36.4
Support vector machine (svm)	*2 genes*	65.6	7.3E-01	50.0	81.3	50.0	18.8	73.9	3.0E-03	66.6	81.1	33.4	18.9
	5 genes	68.4	6.1E-01	55.6	81.3	44.4	18.8	72.3	3.0E-03	68.6	76.0	31.4	24.0
	10 genes	70.1	3.0E-01	77.8	62.5	22.2	37.5	73.1	2.1E-06	73.3	73.0	26.7	27.0
	20 genes	73.3	2.0E-01	77.8	68.8	22.2	31.3	77.8	2.2E-10	81.1	74.5	18.9	25.5
	30 genes	76.0	8.2E-01	83.3	68.8	16.7	31.3	80.1	1.9E-14	82.4	77.7	17.6	22.3
	40 genes	76.0	1.1E-01	83.3	68.8	16.7	31.3	81.0	4.4E-16	83.6	78.5	16.4	21.5
	PCA 50% variability	42.7	8.8E-01	66.7	18.8	33.3	81.3	47.9	1.0E+00	66.1	29.7	33.9	70.3
	PCA 60% variability	42.7	7.6E-01	66.7	18.8	33.3	81.3	43.5	1.0E+00	61.7	25.3	38.3	74.7
	PCA 70% variability	72.6	6.7E-03	88.9	56.3	11.1	43.8	74.2	1.0E-05	83.0	65.3	17.0	34.7
	PCA 80% variability	54.5	5.1E-01	77.8	31.3	22.2	68.8	60.7	4.5E-02	74.9	46.5	25.1	53.5
	PCA 85% variability	49.3	4.2E-01	61.1	37.5	38.9	62.5	69.3	9.1E-04	71.9	66.7	28.1	33.3
	PCA 90% variability	51.7	4.1E-01	72.2	31.3	27.8	68.8	59.5	2.3E-03	75.0	44.0	25.0	56.0
Random Forest (randF)	*2 genes*	77.1	4.7E-02	66.7	87.5	33.3	12.5	76.7	5.6E-01	98.7	54.6	1.3	45.4
	5 genes	67.4	2.5E-01	72.2	62.5	27.8	37.5	76.0	5.6E-01	98.1	53.9	1.9	46.1
	10 genes	69.8	3.3E-01	83.3	56.3	16.7	43.8	75.6	5.7E-01	97.9	53.4	2.1	46.6
	20 genes	76.0	5.8E-02	83.3	68.8	16.7	31.3	75.9	6.3E-01	99.3	52.5	0.7	47.5
	30 genes	70.1	6.8E-01	77.8	62.5	22.2	37.5	76.0	6.4E-01	99.4	52.6	0.6	47.4
	40 genes	70.1	1.3E-01	77.8	62.5	22.2	37.5	75.8	7.5E-01	98.9	52.8	1.1	47.2
	PCA 50% variability	43.4	9.0E-01	55.6	31.3	44.4	68.8	70.8	8.7E-01	89.6	52.0	10.4	48.0
	PCA 60% variability	49.0	8.3E-01	66.7	31.3	33.3	68.8	72.3	8.7E-01	92.7	51.9	7.3	48.1
	PCA 70% variability	69.8	1.5E-01	83.3	56.3	16.7	43.8	73.9	4.8E-01	96.3	51.6	3.7	48.4
	PCA 80% variability	72.9	5.8E-02	83.3	62.5	16.7	37.5	73.2	6.3E-01	95.4	50.9	4.6	49.1
	PCA 85% variability	72.6	6.1E-01	88.9	56.3	11.1	43.8	73.0	6.4E-01	95.3	50.8	4.7	49.2
	PCA 90% variability	60.8	1.3E-01	77.8	43.8	22.2	56.3	71.7	7.7E-01	92.7	50.8	7.3	49.2

Lgg vs Ag using eTumour consensus diagnosis

	Leave-one-out cross validation (LOOCV)						5-fold cross validation (5FCV)					
Number of variables	**% accuracy (AUC)**	**p-val**	**Sensitivity (%)**	**Specificity (%)**	**FNR (%)**	**FPR (%)**	**% accuracy (AUC)**	**p-val**	**Sensitivity (%)**	**Specificity (%)**	**FNR (%)**	**FPR (%)**
Linear discriminant analysis (LDA)												
2 genes	67.2	8.2E-05	73.3	61.1	26.7	38.9	58.0	1.0E+00	57.5	58.5	42.5	41.5
5 genes	76.1	7.7E-04	80.0	72.2	20.0	27.8	58.0	1.0E+00	59.0	57.2	41.0	42.8
10 genes	54.4	1.1E-01	53.3	55.6	46.7	44.4	60.1	8.8E-01	61.6	58.7	38.4	41.3
20 genes	69.4	4.4E-02	66.7	72.2	33.3	27.8	54.7	8.2E-01	57.3	52.2	42.7	47.8
30 genes	60.6	6.9E-01	60.0	61.1	40.0	38.9	59.6	1.3E-01	60.6	58.7	39.4	41.3
40 genes	72.2	4.4E-04	66.7	77.8	33.3	22.2	62.6	3.2E-01	63.1	62.1	36.9	37.9
PCA 50% variability	45.6	9.6E-01	46.7	44.4	53.3	55.6	47.1	1.0E+00	47.1	47.2	52.9	52.8
PCA 60% variability	35.6	9.8E-01	26.7	44.4	73.3	55.6	47.1	1.0E+00	37.6	43.1	62.4	56.9
PCA 70% variability	72.8	4.4E-02	73.3	72.2	26.7	27.8	58.5	1.4E-08	60.0	57.1	40.0	42.9
PCA 80% variability	60.0	1.1E-01	53.3	66.7	46.7	33.3	60.3	1.3E-15	56.9	63.8	43.1	36.2
PCA 85% variability	63.3	5.0E-01	60.0	66.7	40.0	33.3	70.7	1.0E-05	69.4	71.9	30.6	28.1
PCA 90% variability	60.6	1.4E-02	60.0	61.1	40.0	38.9	66.8	1.0E-05	64.7	69.0	35.3	31.0
Support vector machine (svm)												
2 genes	72.2	1.2E-05	66.7	77.8	33.3	22.2	58.2	9.2E-01	49.0	67.4	51.0	32.6
5 genes	71.7	1.4E-02	60.0	83.3	40.0	16.7	58.2	9.2E-01	52.7	64.6	47.3	35.4
10 genes	68.9	4.4E-02	60.0	77.8	40.0	22.2	62.7	2.3E-02	54.9	70.4	45.1	29.6
20 genes	71.7	2.6E-02	60.0	83.3	40.0	16.7	66.4	1.9E-04	57.5	75.4	42.5	24.6
30 genes	68.3	3.1E-01	53.3	83.3	46.7	16.7	70.7	1.7E-06	63.3	78.1	36.7	21.9
40 genes	65.6	4.4E-02	53.3	77.8	46.7	22.2	75.0	1.9E-10	68.4	81.6	31.6	18.4
PCA 50% variability	38.9	9.6E-01	33.3	44.4	66.7	55.6	47.3	1.0E+00	36.9	57.8	63.1	42.2
PCA 60% variability	31.7	9.6E-01	13.3	50.0	86.7	50.0	47.3	1.0E+00	23.9	69.3	76.1	30.7
PCA 70% variability	53.3	5.0E-01	40.0	66.7	60.0	33.3	55.3	4.1E-03	42.2	68.4	57.8	31.6
PCA 80% variability	43.3	3.1E-01	20.0	66.7	80.0	33.3	54.1	9.7E-05	32.0	76.3	68.0	23.7
PCA 85% variability	55.0	5.0E-01	26.7	83.3	73.3	16.7	66.9	1.6E-14	44.5	89.3	55.5	10.7
PCA 90% variability	63.9	2.3E-02	33.3	94.4	66.7	5.6	69.2	5.3E-10	44.1	94.3	55.9	5.7
Random Forest (randF)												
2 genes	66.7	1.2E-03	66.7	66.7	33.3	33.3	37.5	1.0E+00	66.1	56.4	33.9	43.6
5 genes	66.1	7.2E-03	60.0	72.2	40.0	27.8	37.5	1.0E+00	65.3	56.2	34.7	43.8
10 genes	71.7	1.1E-01	60.0	83.3	40.0	16.7	36.2	1.0E+00	65.9	56.6	34.1	43.4
20 genes	66.1	1.7E-01	60.0	72.2	40.0	27.8	35.6	1.0E+00	66.7	56.8	33.3	43.2
30 genes	68.9	6.0E-01	60.0	77.8	40.0	22.2	35.1	1.0E+00	67.1	56.6	32.9	43.4
40 genes	66.1	1.1E-01	60.0	72.2	40.0	27.8	35.6	1.0E+00	68.2	56.5	31.8	43.5
PCA 50% variability	38.3	8.9E-01	26.7	50.0	73.3	50.0	40.5	1.0E+00	64.3	56.2	35.7	43.8
PCA 60% variability	29.4	9.8E-01	20.0	38.9	80.0	61.1	40.5	1.0E+00	62.2	55.5	37.8	44.5
PCA 70% variability	57.2	4.1E-01	53.3	61.1	46.7	38.9	38.8	1.0E+00	62.0	56.0	38.0	44.0
PCA 80% variability	66.7	2.3E-01	66.7	66.7	33.3	33.3	36.4	1.0E+00	63.1	56.1	36.9	43.9
PCA 85% variability	61.7	4.0E-02	40.0	83.3	60.0	16.7	34.6	1.0E+00	61.0	58.7	39.0	41.3
PCA 90% variability	61.1	1.4E-02	33.3	88.9	66.7	11.1	34.8	1.0E+00	61.4	58.9	38.6	41.1

Lgg vs Gb using originating pathologist diagnosis

		Leave-one-out cross validation (LOOCV)					5-fold cross validation (5FCV)				
	Number of variables	% accuracy (AUC)	p-val	Sensitivity (%)	Specificity (%)	FNR (%)	FPR (%)	% accuracy (AUC)	p-val	Sensitivity (%)	Specificity (%)	FNR (%)	FPR (%)
Linear discriminant analysis (LDA)	*2 genes*	88.0	1.5E-03	88.9	87.1	11.1	12.9	81.7	1.0E-05	77.3	86.1	22.7	13.9
	5 genes	80.8	2.6E-02	77.8	83.9	22.2	16.1	79.7	1.0E-05	76.5	82.9	23.5	17.1
	10 genes	76.4	3.1E-02	72.2	80.6	27.8	19.4	79.4	1.0E-05	77.1	81.6	22.9	18.4
	20 genes	74.8	2.1E-02	72.2	77.4	27.8	22.6	74.6	4.4E-16	71.7	77.6	28.3	22.4
	30 genes	76.0	5.8E-01	77.8	74.2	22.2	25.8	69.1	2.4E-12	67.2	71.0	32.8	29.0
	40 genes	70.0	1.3E-01	72.2	67.7	27.8	32.3	62.6	1.4E-04	60.7	64.5	39.3	35.5
	PCA 50% variability	82.4	1.4E-03	77.8	87.1	22.2	12.9	84.0	1.0E-05	78.9	89.1	21.1	10.9
	PCA 60% variability	84.1	1.4E-05	77.8	90.3	22.2	9.7	83.8	1.0E-05	78.7	88.8	21.3	11.2
	PCA 70% variability	85.2	2.4E-05	83.3	87.1	16.7	12.9	81.5	1.0E-05	79.1	84.0	20.9	16.0
	PCA 80% variability	83.2	4.0E-05	88.9	77.4	11.1	22.6	80.3	1.0E-05	82.2	78.4	17.8	21.6
	PCA 85% variability	82.0	5.9E-01	83.3	80.6	16.7	19.4	77.1	1.0E-05	76.3	77.8	23.7	22.2
	PCA 90% variability	73.2	5.8E-04	72.2	74.2	27.8	25.8	72.4	1.0E-05	74.6	70.2	25.4	29.8
Support vector machine (svm)	*2 genes*	84.1	1.5E-03	77.8	90.3	22.2	9.7	80.7	1.0E-05	75.3	86.2	24.7	13.8
	5 genes	80.8	3.5E-03	77.8	83.9	22.2	16.1	81.2	1.0E-05	75.4	87.1	24.6	12.9
	10 genes	76.4	1.7E-02	72.2	80.6	27.8	19.4	79.4	1.0E-05	73.6	85.2	26.4	14.8
	20 genes	79.2	9.3E-04	77.8	80.6	22.2	19.4	79.8	1.0E-05	75.3	84.2	24.7	15.8
	30 genes	76.4	1.0E+00	72.2	80.6	27.8	19.4	80.4	1.0E-05	76.7	84.1	23.3	15.9
	40 genes	82.0	1.4E-03	83.3	80.6	16.7	19.4	80.6	1.0E-05	77.2	83.9	22.8	16.1
	PCA 50% variability	78.0	9.7E-03	72.2	83.9	27.8	16.1	76.8	1.0E-05	67.9	85.7	32.1	14.3
	PCA 60% variability	79.7	3.5E-03	72.2	87.1	27.8	12.9	76.1	1.0E-05	63.3	88.9	36.7	11.1
	PCA 70% variability	72.9	1.3E-02	55.6	90.3	44.4	9.7	74.0	1.0E-06	54.0	94.1	46.0	5.9
	PCA 80% variability	70.2	3.1E-02	50.0	90.3	50.0	9.7	70.8	1.0E-06	47.7	93.9	52.3	6.1
	PCA 85% variability	69.0	9.9E-01	44.4	93.5	55.6	6.5	67.5	1.0E-06	40.1	94.8	59.9	5.2
	PCA 90% variability	63.4	8.2E-02	33.3	93.5	66.7	6.5	62.9	1.0E-06	30.1	95.8	69.9	4.2
Random Forest (randF)	*2 genes*	89.6	5.7E-04	88.9	90.3	11.1	9.7	82.8	9.1E-01	96.8	68.9	3.2	31.1
	5 genes	86.8	5.7E-04	83.3	90.3	16.7	9.7	83.4	9.1E-01	97.4	69.3	2.6	30.7
	10 genes	84.1	2.2E-03	77.8	90.3	22.2	9.7	83.5	9.2E-01	97.4	69.5	2.6	30.5
	20 genes	85.2	2.2E-03	83.3	87.1	16.7	12.9	84.2	8.9E-01	98.9	69.5	1.1	30.5
	30 genes	86.8	9.7E-01	83.3	90.3	16.7	9.7	84.2	8.8E-01	99.1	69.4	0.9	30.6
	40 genes	83.6	5.8E-04	83.3	83.9	16.7	16.1	84.2	9.0E-01	99.2	69.2	0.8	30.8
	PCA 50% variability	70.9	1.4E-02	61.1	80.6	38.9	19.4	82.4	6.4E-02	95.5	69.3	4.5	30.7
	PCA 60% variability	70.9	2.2E-03	61.1	80.6	38.9	19.4	82.8	6.4E-02	96.5	69.1	3.5	30.9
	PCA 70% variability	76.9	9.3E-04	66.7	87.1	33.3	12.9	82.4	1.9E-01	94.7	70.1	5.3	29.9
	PCA 80% variability	75.7	6.4E-03	61.1	90.3	38.9	9.7	82.7	1.4E-01	94.7	70.7	5.3	29.3
	PCA 85% variability	65.8	9.5E-01	44.4	87.1	55.6	12.9	83.1	3.9E-03	94.5	71.8	5.5	28.2
	PCA 90% variability	69.0	3.9E-02	44.4	93.5	55.6	6.5	83.0	1.0E-02	94.0	72.1	6.0	27.9

Lgg vs Gb using eTumour consensus diagnosis

		Leave-one-out cross validation (LOOCV)					5-fold cross validation (5FCV)				
	Number of variables	% accuracy (AUC)	p-val	Sensitivity (%)	Specificity (%)	FNR (%)	FPR (%)	% accuracy (AUC)	p-val	Sensitivity (%)	Specificity (%)	FNR (%)	FPR (%)
Linear discriminant analysis (LDA)	*2 genes*	84.0	2.1E-02	86.7	81.3	13.3	18.8	84.7	1.0E-05	85.4	84.0	14.6	16.0
	5 genes	90.4	3.9E-03	93.3	87.5	6.7	12.5	84.9	1.0E-05	84.8	85.0	15.2	15.0
	10 genes	78.9	4.6E-02	73.3	84.4	26.7	15.6	80.0	1.0E-05	77.4	82.6	22.6	17.4
	20 genes	75.5	6.4E-03	66.7	84.4	33.3	15.6	73.7	1.0E-05	68.4	79.1	31.6	20.9
	30 genes	65.9	6.8E-01	60.0	71.9	40.0	28.1	66.0	1.1E-12	60.6	71.5	39.4	28.5
	40 genes	67.9	6.7E-02	73.3	62.5	26.7	37.5	65.5	7.2E-12	60.7	70.2	39.3	29.8
	PCA 50% variability	78.9	3.0E-03	73.3	84.4	26.7	15.6	85.0	1.0E-05	83.4	86.5	16.6	13.5
	PCA 60% variability	78.9	5.4E-03	73.3	84.4	26.7	15.6	85.8	1.0E-05	83.1	88.4	16.9	11.6
	PCA 70% variability	79.1	1.1E-02	80.0	78.1	20.0	21.9	84.3	1.0E-05	83.1	85.4	16.9	14.6
	PCA 80% variability	80.6	9.4E-03	80.0	81.3	20.0	18.8	85.1	1.0E-05	84.7	85.6	15.3	14.4
	PCA 85% variability	82.2	8.7E-01	80.0	84.4	20.0	15.6	81.2	1.0E-05	81.4	81.0	18.6	19.0
	PCA 90% variability	75.9	3.7E-04	80.0	71.9	20.0	28.1	80.9	1.0E-05	89.2	72.5	10.8	27.5
Support vector machine (svm)	*2 genes*	80.6	7.5E-02	80.0	81.3	20.0	18.8	83.2	1.0E-05	81.7	84.7	18.3	15.3
	5 genes	92.0	1.5E-03	93.3	90.6	6.7	9.4	86.6	1.0E-05	85.4	87.9	14.6	12.1
	10 genes	83.8	1.7E-02	80.0	87.5	20.0	12.5	84.8	1.0E-05	82.0	87.7	18.0	12.3
	20 genes	83.8	3.5E-03	80.0	87.5	20.0	12.5	83.8	1.0E-05	80.4	87.2	19.6	12.8
	30 genes	78.9	9.2E-01	73.3	84.4	26.7	15.6	83.7	1.0E-05	80.0	87.3	20.0	12.7
	40 genes	78.9	2.0E-03	73.3	84.4	26.7	15.6	83.7	1.0E-05	80.4	86.9	19.6	13.1
	PCA 50% variability	75.5	4.8E-02	66.7	84.4	33.3	15.6	77.4	1.0E-05	68.8	86.0	31.2	14.0
	PCA 60% variability	72.2	3.4E-02	60.0	84.4	40.0	15.6	76.8	1.0E-05	64.0	89.6	36.0	10.4
	PCA 70% variability	67.1	3.7E-02	46.7	87.5	53.3	12.5	72.4	1.0E-06	52.3	92.6	47.7	7.4
	PCA 80% variability	66.9	1.3E-01	40.0	93.8	60.0	6.3	67.4	1.0E-06	40.3	94.6	59.7	5.4
	PCA 85% variability	60.2	9.8E-01	26.7	93.8	73.3	6.3	63.7	1.0E-06	32.3	95.0	67.7	5.0
	PCA 90% variability	55.3	2.5E-01	20.0	90.6	80.0	9.4	60.4	1.0E-06	24.8	95.9	75.2	4.1
Random Forest (randF)	*2 genes*	82.2	5.7E-02	80.0	84.4	20.0	15.6	84.8	5.2E-02	97.6	72.0	2.4	28.0
	5 genes	92.0	2.5E-03	93.3	90.6	6.7	9.4	85.3	5.2E-02	98.3	72.2	1.7	27.8
	10 genes	85.5	3.1E-02	86.7	84.4	13.3	15.6	84.8	3.8E-02	97.4	72.2	2.6	27.8
	20 genes	86.9	3.5E-04	80.0	93.8	20.0	6.3	85.3	4.0E-02	98.2	72.4	1.8	27.6
	30 genes	83.8	9.8E-01	80.0	87.5	20.0	12.5	85.0	2.9E-02	97.4	72.5	2.6	27.5
	40 genes	87.1	1.5E-03	86.7	87.5	13.3	12.5	85.2	2.5E-02	98.0	72.4	2.0	27.6
	PCA 50% variability	88.4	4.0E-05	80.0	96.9	20.0	3.1	86.3	4.2E-03	99.0	73.6	1.0	26.4
	PCA 60% variability	88.4	1.8E-04	80.0	96.9	20.0	3.1	86.6	4.2E-03	99.6	73.7	0.4	26.3
	PCA 70% variability	86.7	7.0E-04	73.3	100.0	26.7	0.0	86.1	2.0E-03	97.6	74.6	2.4	25.4
	PCA 80% variability	80.0	7.8E-03	60.0	100.0	40.0	0.0	86.3	1.8E-03	97.6	75.0	2.4	25.0
	PCA 85% variability	76.7	9.8E-01	53.3	100.0	46.7	0.0	85.6	2.8E-03	95.9	75.3	4.1	24.7
	PCA 90% variability	73.3	3.0E-02	46.7	100.0	53.3	0.0	83.6	2.4E-02	91.8	75.5	8.2	24.5

Ag vs Gb using originating pathologist diagnosis

		Leave-one-out cross validation (LOOCV)						5-fold cross validation (5FCV)					
	Number of variables	% accuracy (AUC)	p-val	Sensitivity (%)	Specificity (%)	FNR (%)	FPR (%)	% accuracy (AUC)	p-val	Sensitivity (%)	Specificity (%)	FNR (%)	FPR (%)
Linear discriminant analysis (LDA)	*2 genes*	76.3	2.4E-01	68.8	83.9	31.3	16.1	75.0	1.0E-05	65.4	84.7	34.6	15.3
	5 genes	71.6	2.6E-01	62.5	80.6	37.5	19.4	73.1	1.0E-05	64.3	81.9	35.7	18.1
	10 genes	70.0	4.3E-02	62.5	77.4	37.5	22.6	69.2	6.7E-01	59.7	78.8	40.3	21.3
	20 genes	52.4	4.3E-01	50.0	54.8	50.0	45.2	61.7	9.9E-01	54.2	69.3	45.8	30.7
	30 genes	58.7	9.2E-01	62.5	54.8	37.5	45.2	57.4	6.3E-01	53.5	61.4	46.5	38.6
	40 genes	52.4	5.7E-02	50.0	54.8	50.0	45.2	56.4	2.5E-01	51.3	61.6	48.8	38.4
	PCA 50% variability	73.2	2.4E-02	62.5	83.9	37.5	16.1	73.4	1.0E-05	61.9	84.8	38.1	15.2
	PCA 60% variability	71.6	2.1E-02	62.5	80.6	37.5	19.4	70.8	1.0E-05	60.4	81.3	39.6	18.8
	PCA 70% variability	70.1	1.8E-01	56.3	83.9	43.8	16.1	66.1	8.7E-08	54.0	78.2	46.0	21.8
	PCA 80% variability	63.6	2.3E-01	56.3	71.0	43.8	29.0	64.7	6.0E-08	58.8	70.6	41.3	29.4
	PCA 85% variability	70.0	5.1E-01	62.5	77.4	37.5	22.6	68.8	1.0E-05	63.9	73.8	36.1	26.3
	PCA 90% variability	71.6	1.0E-01	62.5	80.6	37.5	19.4	63.1	1.2E-11	59.4	66.8	40.6	33.2
Support vector machine (svm)	*2 genes*	76.4	2.0E-01	62.5	90.3	37.5	9.7	75.7	1.0E-05	63.2	88.2	36.8	11.8
	5 genes	79.5	8.9E-02	68.8	90.3	31.3	9.7	77.5	1.0E-05	65.4	89.6	34.6	10.4
	10 genes	79.6	6.7E-03	62.5	96.8	37.5	3.2	77.5	1.6E-10	63.5	91.6	36.5	8.4
	20 genes	79.6	7.0E-03	62.5	96.8	37.5	3.2	77.5	4.3E-13	63.1	91.9	36.9	8.1
	30 genes	76.4	1.0E+00	62.5	90.3	37.5	9.7	77.1	1.0E-06	63.2	91.0	36.8	9.0
	40 genes	76.4	2.9E-02	62.5	90.3	37.5	9.7	76.7	1.0E-06	63.2	90.2	36.8	9.8
	PCA 50% variability	66.9	1.8E-01	50.0	83.9	50.0	16.1	69.0	1.0E-05	52.9	85.1	47.1	14.9
	PCA 60% variability	76.5	1.1E-02	56.3	96.8	43.8	3.2	75.2	1.0E-06	56.8	93.6	43.2	6.4
	PCA 70% variability	70.2	4.7E-02	50.0	90.3	50.0	9.7	68.7	1.0E-06	43.8	93.6	56.3	6.4
	PCA 80% variability	65.5	7.1E-02	37.5	93.5	62.5	6.5	63.6	1.0E-06	31.9	95.2	68.1	4.8
	PCA 85% variability	57.7	9.9E-01	25.0	90.3	75.0	9.7	62.0	1.0E-05	28.3	95.6	71.7	4.4
	PCA 90% variability	53.0	4.1E-01	12.5	93.5	87.5	6.5	54.9	3.9E-14	14.2	95.7	85.8	4.3
Random Forest (randF)	*2 genes*	71.6	1.9E-01	62.5	80.6	37.5	19.4	79.2	6.9E-01	88.5	70.0	11.5	30.0
	5 genes	71.7	1.3E-02	56.3	87.1	43.8	12.9	79.6	6.9E-01	89.2	70.1	10.8	29.9
	10 genes	73.3	6.7E-03	56.3	90.3	43.8	9.7	79.4	8.0E-01	88.5	70.4	11.5	29.6
	20 genes	71.7	9.8E-03	56.3	87.1	43.8	12.9	79.4	8.7E-01	88.5	70.3	11.5	29.7
	30 genes	70.1	9.6E-01	56.3	83.9	43.8	16.1	79.4	8.1E-01	88.6	70.2	11.4	29.8
	40 genes	76.4	6.2E-03	62.5	90.3	37.5	9.7	79.4	8.3E-01	88.6	70.3	11.4	29.7
	PCA 50% variability	68.5	3.9E-03	50.0	87.1	50.0	12.9	79.4	6.0E-01	87.8	71.0	12.2	29.0
	PCA 60% variability	59.2	5.7E-02	31.3	87.1	68.8	12.9	78.0	6.0E-01	84.7	71.3	15.3	28.7
	PCA 70% variability	60.8	3.4E-01	31.3	90.3	68.8	9.7	77.2	9.9E-01	82.2	72.1	17.8	27.9
	PCA 80% variability	62.4	9.3E-02	31.3	93.5	68.8	6.5	77.2	8.9E-01	81.3	73.1	18.8	26.9
	PCA 85% variability	53.1	3.6E-02	6.3	100.0	93.8	0.0	77.6	9.0E-01	81.1	74.1	18.9	25.9
	PCA 90% variability	53.1	2.7E-01	6.3	100.0	93.8	0.0	78.0	8.6E-01	81.4	74.6	18.6	25.4

Ag vs Gb using eTumour consensus diagnosis

	Leave-one-out cross validation (LOOCV)						5-fold cross validation (5FCV)					
Number of variables	**% accuracy (AUC)**	**p-val**	**Sensitivity (%)**	**Specificity (%)**	**FNR (%)**	**FPR (%)**	**% accuracy (AUC)**	**p-val**	**Sensitivity (%)**	**Specificity (%)**	**FNR (%)**	**FPR (%)**
Linear discriminant analysis (LDA)												
2 genes	78.6	8.5E-06	66.7	90.6	33.3	9.4	78.9	1.0E-05	68.7	89.1	31.3	10.9
5 genes	78.6	4.7E-05	66.7	90.6	33.3	9.4	76.1	1.0E-05	65.7	86.6	34.3	13.4
10 genes	70.8	2.0E-03	66.7	75.0	33.3	25.0	70.4	5.5E-12	60.3	80.5	39.7	19.5
20 genes	63.7	1.5E-01	55.6	71.9	44.4	28.1	63.8	1.8E-05	56.6	71.0	43.4	29.0
30 genes	57.1	9.8E-01	61.1	53.1	38.9	46.9	58.6	1.3E-05	54.8	62.4	45.2	37.6
40 genes	48.8	1.7E-01	44.4	53.1	55.6	46.9	55.0	4.1E-01	52.1	57.8	47.9	42.2
PCA 50% variability	78.6	8.6E-03	66.7	90.6	33.3	9.4	77.9	1.0E-05	67.2	88.6	32.8	11.4
PCA 60% variability	74.0	1.8E-02	66.7	81.3	33.3	18.8	73.1	1.0E-05	62.0	84.3	38.0	15.7
PCA 70% variability	72.7	9.7E-03	61.1	84.4	38.9	15.6	69.2	1.0E-05	57.4	81.0	42.6	19.0
PCA 80% variability	68.1	2.6E-02	61.1	75.0	38.9	25.0	65.8	1.0E-05	58.5	73.1	41.5	26.9
PCA 85% variability	79.5	5.8E-01	77.8	81.3	22.2	18.8	72.6	1.0E-05	68.5	76.7	31.5	23.3
PCA 90% variability	65.3	2.6E-02	55.6	75.0	44.4	25.0	63.1	1.0E-05	56.5	69.7	43.5	30.3
Support vector machine (svm)												
2 genes	80.2	5.1E-06	66.7	93.8	33.3	6.3	79.4	1.0E-05	66.4	92.5	33.6	7.5
5 genes	81.8	5.1E-06	66.7	96.9	33.3	3.1	80.3	1.0E-05	66.8	93.8	33.2	6.2
10 genes	80.2	4.7E-05	66.7	93.8	33.3	6.3	80.9	1.0E-06	67.0	94.9	33.0	5.1
20 genes	80.2	4.7E-05	66.7	93.8	33.3	6.3	80.5	1.0E-06	66.9	94.1	33.1	5.9
30 genes	80.2	5.1E-01	66.7	93.8	33.3	6.3	80.0	1.0E-06	66.9	93.1	33.1	6.9
40 genes	78.6	8.9E-04	66.7	90.6	33.3	9.4	79.8	1.0E-06	66.9	92.8	33.1	7.2
PCA 50% variability	74.3	5.7E-02	61.1	87.5	38.9	12.5	72.4	1.0E-05	56.0	88.8	44.0	11.2
PCA 60% variability	79.0	7.5E-03	61.1	96.9	38.9	3.1	77.5	1.0E-06	59.6	95.4	40.4	4.6
PCA 70% variability	71.9	5.3E-02	50.0	93.8	50.0	6.3	72.3	1.0E-06	49.1	95.4	50.9	4.6
PCA 80% variability	66.3	1.2E-01	38.9	93.8	61.1	6.3	66.7	1.0E-06	37.4	96.0	62.6	4.0
PCA 85% variability	71.9	1.0E+00	50.0	93.8	50.0	6.3	66.0	1.0E-06	36.0	95.9	64.0	4.1
PCA 90% variability	60.8	2.3E-01	27.8	93.8	72.2	6.3	60.4	1.0E-06	24.7	96.1	75.3	3.9
Random Forest (randF)												
2 genes	74.3	5.1E-05	61.1	87.5	38.9	12.5	78.9	1.0E+00	87.8	69.9	12.2	30.1
5 genes	75.9	1.4E-03	61.1	90.6	38.9	9.4	79.0	1.0E+00	87.6	70.3	12.4	29.7
10 genes	74.3	5.5E-03	61.1	87.5	38.9	12.5	79.1	1.0E+00	87.7	70.5	12.3	29.5
20 genes	75.5	3.5E-03	66.7	84.4	33.3	15.6	79.0	1.0E+00	87.7	70.3	12.3	29.7
30 genes	78.6	6.5E-01	66.7	90.6	33.3	9.4	79.0	1.0E+00	87.7	70.2	12.3	29.8
40 genes	77.1	2.2E-03	66.7	87.5	33.3	12.5	78.9	1.0E+00	87.7	70.2	12.3	29.8
PCA 50% variability	70.0	1.7E-02	55.6	84.4	44.4	15.6	80.2	9.7E-01	89.9	70.6	10.1	29.4
PCA 60% variability	67.2	2.0E-01	50.0	84.4	50.0	15.6	80.4	9.7E-01	89.4	71.4	10.6	28.6
PCA 70% variability	64.8	6.0E-02	38.9	90.6	61.1	9.4	80.8	8.9E-01	89.4	72.2	10.6	27.8
PCA 80% variability	58.9	1.5E-01	33.3	84.4	66.7	15.6	80.6	8.5E-01	88.4	72.8	11.6	27.2
PCA 85% variability	65.1	7.5E-01	33.3	96.9	66.7	3.1	79.2	1.0E+00	84.8	73.6	15.2	26.4
PCA 90% variability	58.0	2.2E-01	22.2	93.8	77.8	6.3	79.3	1.0E+00	84.2	74.4	15.8	25.6

Mg vs Lgg vs Gb using originating pathologist diagnosis

		5-fold cross validation (5FCV)													
		% accuracy		Sensitivity	Specificity	FNR	FPR	Sensitivity	Specificity	FNR	FPR	Sensitivity	Specificity	FNR	FPR
	Number of variables	(AUC)	p-val	Mg (%)	Mg (%)	Mg (%)	Mg (%)	Lgg (%)	Lgg (%)	Lgg (%)	Lgg (%)	Gb (%)	Gb (%)	Gb (%)	Gb (%)
Linear discriminant analysis (LDA)	*5 genes*	83.5	1.0E-05	99.2	100.0	0.8	0.0	61.6	92.2	38.4	7.8	89.9	76.9	10.1	23.1
	10 genes	82.8	1.0E-05	100.0	100.0	0.0	0.0	68.6	85.0	31.4	15.0	79.9	81.2	20.1	18.8
	20 genes	82.5	1.0E-05	100.0	100.0	0.0	0.0	69.4	83.5	30.6	16.5	78.0	81.6	22.0	18.4
	30 genes	82.3	1.0E-05	99.3	100.0	0.7	0.0	69.8	83.1	30.2	16.9	77.6	81.8	22.4	18.2
	40 genes	75.0	1.0E-05	96.1	99.8	3.9	0.2	62.3	74.6	37.7	25.4	66.6	76.7	33.4	23.3
	50 genes	77.2	1.0E-05	97.3	99.9	2.7	0.1	65.8	76.0	34.2	24.0	68.4	79.0	31.6	21.0
	PCA 40% variability	85.2	1.0E-05	100.0	100.0	0.0	0.0	72.4	87.4	27.6	12.6	83.2	81.6	16.8	18.4
	PCA 50% variability	85.8	1.0E-05	100.0	100.0	0.0	0.0	74.8	87.0	25.2	13.0	82.6	83.2	17.4	16.8
	PCA 60% variability	88.9	1.0E-05	100.0	100.0	0.0	0.0	79.0	90.7	21.0	9.3	87.7	86.0	12.3	14.0
	PCA 70% variability	87.2	1.0E-05	100.0	100.0	0.0	0.0	75.2	89.7	24.8	10.3	86.2	83.5	13.8	16.5
	PCA 80% variability	89.6	1.0E-05	100.0	100.0	0.0	0.0	86.0	87.2	14.0	12.8	83.0	90.7	17.0	9.3
	PCA 90% variability	84.5	1.0E-05	99.5	100.0	0.5	0.0	83.2	77.9	16.8	22.1	70.7	88.8	29.3	11.2
Support vector machine (svm)	*5 genes*	66.4	1.0E-05	99.3	100.0	0.7	0.0	0.0	100.0	100.0	0.0	100.0	39.7	0.0	60.3
	10 genes	66.6	1.0E-07	100.0	100.0	0.0	0.0	0.0	100.0	100.0	0.0	99.9	40.0	0.1	60.0
	20 genes	66.7	1.0E-05	100.0	100.0	0.0	0.0	0.0	100.0	100.0	0.0	100.0	40.0	0.0	60.0
	30 genes	66.7	1.0E-08	100.0	100.0	0.0	0.0	0.0	100.0	100.0	0.0	100.0	40.0	0.0	60.0
	40 genes	66.7	1.0E-05	100.0	100.0	0.0	0.0	0.0	100.0	100.0	0.0	100.0	40.0	0.0	60.0
	50 genes	66.7	1.0E-05	100.0	100.0	0.0	0.0	0.0	100.0	100.0	0.0	100.0	40.0	0.0	60.0
	PCA 40% variability	87.5	1.0E-05	100.0	100.0	0.0	0.0	79.1	87.5	20.9	12.5	83.3	86.1	16.7	13.9
	PCA 50% variability	81.3	1.0E-05	91.0	100.0	9.0	0.0	71.4	84.9	28.6	15.1	81.4	79.5	18.6	20.5
	PCA 60% variability	81.3	1.0E-05	89.5	100.0	10.5	0.0	69.3	87.4	30.7	12.6	85.1	78.0	14.9	22.0
	PCA 70% variability	74.3	1.0E-05	82.1	100.0	17.9	0.0	51.1	92.3	48.9	7.7	89.7	61.5	10.3	38.5
	PCA 80% variability	74.9	1.0E-05	78.7	100.0	21.3	0.0	51.1	96.2	48.9	3.8	95.0	60.3	5.0	39.7
	PCA 90% variability	47.6	1.0E-05	11.5	100.0	88.5	0.0	35.8	96.5	64.2	3.5	95.5	27.9	4.5	72.1
Random Forest (randF)	*5 genes*	33.3	7.4E-01	100.0	84.0	0.0	16.0	0.0	74.1	100.0	25.9	0.0	36.9	100.0	63.1
	10 genes	33.3	7.4E-01	100.0	83.9	0.0	16.1	0.0	73.7	100.0	26.3	0.0	37.5	100.0	62.5
	20 genes	33.3	6.8E-01	100.0	83.9	0.0	16.1	0.0	74.0	100.0	26.0	0.0	37.1	100.0	62.9
	30 genes	33.3	6.8E-01	100.0	83.9	0.0	16.1	0.0	74.2	100.0	25.8	0.0	36.9	100.0	63.1
	40 genes	33.3	7.0E-01	100.0	83.9	0.0	16.1	0.0	74.1	100.0	25.9	0.0	37.0	100.0	63.0
	50 genes	33.3	6.9E-01	100.0	83.9	0.0	16.1	0.0	74.0	100.0	26.0	0.0	37.1	100.0	62.9
	PCA 40% variability	34.4	1.8E-01	100.0	84.5	0.0	15.5	0.0	68.2	100.0	31.8	3.2	42.5	96.8	57.5
	PCA 50% variability	34.4	1.8E-01	100.0	84.5	0.0	15.5	0.0	68.1	100.0	31.9	3.2	42.7	96.8	57.3
	PCA 60% variability	34.3	1.3E-01	100.0	84.5	0.0	15.5	0.0	68.1	100.0	31.9	3.0	42.7	97.0	57.3
	PCA 70% variability	34.4	1.4E-01	100.0	84.5	0.0	15.5	0.0	67.9	100.0	32.1	3.2	42.9	96.8	57.1
	PCA 80% variability	34.4	1.0E-01	100.0	84.5	0.0	15.5	0.0	68.7	100.0	31.3	3.3	42.1	96.7	57.9
	PCA 90% variability	34.3	8.2E-02	99.3	84.6	0.7	15.4	0.1	70.3	99.9	29.7	3.4	40.3	96.6	59.7

Mg vs Lgg vs Gb using originating pathologist diagnosis

		Leave-one-out cross validation (LOOCV)													
		% accuracy		Sensitivity	Specificity	FNR	FPR	Sensitivity	Specificity	FNR	FPR	Sensitivity	Specificity	FNR	FPR
	Number of variables	(AUC)	p-val	Mg (%)	Mg (%)	Mg (%)	Mg (%)	Lgg (%)	Lgg (%)	Lgg (%)	Lgg (%)	Gb (%)	Gb (%)	Gb (%)	Gb (%)
Linear discriminant analysis (LDA)	*5 genes*	82.4	1.1E-03	100.0	100.0	0.0	0.0	66.7	85.7	33.3	14.3	80.6	79.3	19.4	20.7
	10 genes	80.3	2.4E-03	100.0	100.0	0.0	0.0	66.7	81.0	33.3	19.0	74.2	79.3	25.8	20.7
	20 genes	82.1	4.8E-05	100.0	100.0	0.0	0.0	72.2	81.0	27.8	19.0	74.2	82.8	25.8	17.2
	30 genes	88.3	4.3E-05	100.0	100.0	0.0	0.0	77.8	90.5	22.2	9.5	87.1	86.2	12.9	13.8
	40 genes	80.3	5.1E-06	100.0	100.0	0.0	0.0	66.7	81.0	33.3	19.0	74.2	79.3	25.8	20.7
	50 genes	74.1	2.8E-05	100.0	100.0	0.0	0.0	61.1	71.4	38.9	28.6	61.3	75.9	38.7	24.1
	PCA 40% variability	85.4	5.9E-07	100.0	100.0	0.0	0.0	72.2	88.1	27.8	11.9	83.9	82.8	16.1	17.2
	PCA 50% variability	87.2	3.4E-08	100.0	100.0	0.0	0.0	77.8	88.1	22.2	11.9	83.9	86.2	16.1	13.8
	PCA 60% variability	87.2	4.7E-05	100.0	100.0	0.0	0.0	77.8	88.1	22.2	11.9	83.9	86.2	16.1	13.8
	PCA 70% variability	88.3	4.1E-06	100.0	100.0	0.0	0.0	77.8	90.5	22.2	9.5	87.1	86.2	12.9	13.8
	PCA 80% variability	91.7	2.0E-08	100.0	100.0	0.0	0.0	94.4	85.7	5.6	14.3	80.6	96.6	19.4	3.4
	PCA 90% variability	86.6	7.1E-09	100.0	100.0	0.0	0.0	88.9	78.6	11.1	21.4	71.0	93.1	29.0	6.9
Support vector machine (svm)	*5 genes*	66.7	1.4E-01	100.0	100.0	0.0	0.0	0.0	100.0	100.0	0.0	100.0	37.9	0.0	62.1
	10 genes	66.7	1.4E-01	100.0	100.0	0.0	0.0	0.0	100.0	100.0	0.0	100.0	37.9	0.0	62.1
	20 genes	66.7	1.8E-01	100.0	100.0	0.0	0.0	0.0	100.0	100.0	0.0	100.0	37.9	0.0	62.1
	30 genes	66.7	1.8E-01	100.0	100.0	0.0	0.0	0.0	100.0	100.0	0.0	100.0	37.9	0.0	62.1
	40 genes	66.7	1.8E-01	100.0	100.0	0.0	0.0	0.0	100.0	100.0	0.0	100.0	37.9	0.0	62.1
	50 genes	66.7	1.4E-01	100.0	100.0	0.0	0.0	0.0	100.0	100.0	0.0	100.0	37.9	0.0	62.1
	PCA 40% variability	88.0	4.7E-05	100.0	100.0	0.0	0.0	83.3	85.7	16.7	14.3	80.6	89.7	19.4	10.3
	PCA 50% variability	81.3	8.2E-05	90.9	100.0	9.1	0.0	72.2	83.3	27.8	16.7	80.6	82.8	19.4	17.2
	PCA 60% variability	82.3	1.7E-05	90.9	100.0	9.1	0.0	72.2	85.7	27.8	14.3	83.9	82.8	16.1	17.2
	PCA 70% variability	74.8	1.3E-03	81.8	100.0	18.2	0.0	55.6	90.5	44.4	9.5	87.1	65.5	12.9	34.5
	PCA 80% variability	75.1	1.2E-03	81.8	100.0	18.2	0.0	50.0	95.2	50.0	4.8	93.5	62.1	6.5	37.9
	PCA 90% variability	50.9	7.1E-02	9.1	100.0	90.9	0.0	50.0	95.2	50.0	4.8	93.5	34.5	6.5	65.5
Random Forest (randF)	*5 genes*	78.0	1.9E-02	100.0	100.0	0.0	0.0	50.0	88.1	50.0	11.9	83.9	69.0	16.1	31.0
	10 genes	76.9	3.0E-02	100.0	100.0	0.0	0.0	50.0	85.7	50.0	14.3	80.6	69.0	19.4	31.0
	20 genes	79.8	1.7E-02	100.0	100.0	0.0	0.0	55.6	88.1	44.4	11.9	83.9	72.4	16.1	27.6
	30 genes	80.6	2.7E-02	100.0	100.0	0.0	0.0	61.1	85.7	38.9	14.3	80.6	75.9	19.4	24.1
	40 genes	83.5	3.7E-02	100.0	100.0	0.0	0.0	66.7	88.1	33.3	11.9	83.9	79.3	16.1	20.7
	50 genes	85.4	4.9E-03	100.0	100.0	0.0	0.0	72.2	88.1	27.8	11.9	83.9	82.8	16.1	17.2
	PCA 40% variability	79.5	7.1E-04	100.0	100.0	0.0	0.0	61.1	83.3	38.9	16.7	77.4	75.9	22.6	24.1
	PCA 50% variability	86.1	5.8E-05	100.0	100.0	0.0	0.0	77.8	85.7	22.2	14.3	80.6	86.2	19.4	13.8
	PCA 60% variability	84.3	1.7E-05	100.0	100.0	0.0	0.0	72.2	85.7	27.8	14.3	80.6	82.8	19.4	17.2
	PCA 70% variability	84.3	1.7E-05	100.0	100.0	0.0	0.0	72.2	85.7	27.8	14.3	80.6	82.8	19.4	17.2
	PCA 80% variability	85.4	5.8E-05	100.0	100.0	0.0	0.0	72.2	88.1	27.8	11.9	83.9	82.8	16.1	17.2
	PCA 90% variability	78.9	1.6E-04	90.9	100.0	9.1	0.0	55.6	92.9	44.4	7.1	90.3	69.0	9.7	31.0

Mg vs Lgg vs Gb using eTumour consensus diagnosis

		5-fold cross validation (5FCV)															
	Number of variables	% accuracy (AUC)	p-val	Sensitivity Mg (%)	Specificity Mg (%)	FNR Mg (%)	FPR Mg (%)	Sensitivity Lgg (%)	Specificity Lgg (%)	FNR Lgg (%)	FPR Lgg (%)	Sensitivity Gb (%)	Specificity Gb (%)	FNR Gb (%)	FPR Gb (%)
Linear discriminant analysis (LDA)	*5 genes*	83.5	1.0E-05	99.2	100.0	0.8	0.0	61.6	92.2	38.4	7.8	89.9	76.9	10.1	23.1
	10 genes	82.8	1.0E-05	100.0	100.0	0.0	0.0	68.6	85.0	31.4	15.0	79.9	81.2	20.1	18.8
	20 genes	82.5	1.0E-05	100.0	100.0	0.0	0.0	69.4	83.5	30.6	16.5	78.0	81.6	22.0	18.4
	30 genes	82.3	1.0E-05	99.3	100.0	0.7	0.0	69.8	83.1	30.2	16.9	77.6	81.8	22.4	18.2
	40 genes	75.0	1.0E-05	96.1	99.8	3.9	0.2	62.3	74.6	37.7	25.4	66.6	76.7	33.4	23.3
	50 genes	77.2	1.0E-05	97.3	99.9	2.7	0.1	65.8	76.0	34.2	24.0	68.4	79.0	31.6	21.0
	PCA 40% variability	84.1	1.0E-05	100.0	100.0	0.0	0.0	65.8	90.0	34.2	10.0	86.6	79.5	13.4	20.5
	PCA 50% variability	87.2	1.0E-05	100.0	100.0	0.0	0.0	77.5	88.1	22.5	11.9	84.1	86.5	15.9	13.5
	PCA 60% variability	89.1	1.0E-05	100.0	100.0	0.0	0.0	81.7	89.3	18.3	10.7	85.7	89.0	14.3	11.0
	PCA 70% variability	91.3	1.0E-05	100.0	100.0	0.0	0.0	84.2	92.3	15.8	7.7	89.8	90.5	10.2	9.5
	PCA 80% variability	91.5	1.0E-05	100.0	100.0	0.0	0.0	84.0	92.9	16.0	7.1	90.6	90.4	9.4	9.6
	PCA 90% variability	87.2	1.0E-05	99.3	100.0	0.7	0.0	84.6	83.0	15.4	17.0	77.5	90.8	22.5	9.2
Support vector machine (svm)	*5 genes*	66.4	1.0E-05	99.3	100.0	0.7	0.0	0.0	100.0	100.0	0.0	100.0	39.7	0.0	60.3
	10 genes	66.6	1.0E-05	100.0	100.0	0.0	0.0	0.0	100.0	100.0	0.0	99.9	40.0	0.1	60.0
	20 genes	66.7	1.0E-05	100.0	100.0	0.0	0.0	0.0	100.0	100.0	0.0	100.0	40.0	0.0	60.0
	30 genes	66.7	1.0E-05	100.0	100.0	0.0	0.0	0.0	100.0	100.0	0.0	100.0	40.0	0.0	60.0
	40 genes	66.7	1.0E-05	100.0	100.0	0.0	0.0	0.0	100.0	100.0	0.0	100.0	40.0	0.0	60.0
	50 genes	66.7	1.0E-05	100.0	100.0	0.0	0.0	0.0	100.0	100.0	0.0	100.0	40.0	0.0	60.0
	PCA 40% variability	87.9	1.0E-05	100.0	100.0	0.0	0.0	79.1	88.5	20.9	11.5	84.7	87.5	15.3	12.5
	PCA 50% variability	81.0	1.0E-05	89.5	100.0	10.5	0.0	70.6	86.2	29.4	13.8	82.8	79.6	17.2	20.4
	PCA 60% variability	79.0	1.0E-05	80.8	99.9	19.2	0.1	68.8	88.6	31.2	11.4	87.4	76.8	12.6	23.2
	PCA 70% variability	75.0	1.0E-05	80.8	100.0	19.2	0.0	52.2	93.9	47.8	6.1	91.8	63.7	8.2	36.3
	PCA 80% variability	69.7	1.0E-05	75.1	100.0	24.9	0.0	36.9	97.8	63.1	2.2	97.1	52.2	2.9	47.8
	PCA 90% variability	43.9	1.0E-05	9.0	100.0	91.0	0.0	25.6	97.8	74.4	2.2	97.1	19.0	2.9	81.0
Random Forest (randF)	*5 genes*	33.3	7.4E-01	100.0	84.0	0.0	16.0	0.0	74.1	100.0	25.9	0.0	36.9	100.0	63.1
	10 genes	33.3	7.4E-01	100.0	83.9	0.0	16.1	0.0	73.7	100.0	26.3	0.0	37.5	100.0	62.5
	20 genes	33.3	6.8E-01	100.0	83.9	0.0	16.1	0.0	74.0	100.0	26.0	0.0	37.1	100.0	62.9
	30 genes	33.3	6.8E-01	100.0	83.9	0.0	16.1	0.0	74.2	100.0	25.8	0.0	36.9	100.0	63.1
	40 genes	33.3	7.0E-01	100.0	83.9	0.0	16.1	0.0	74.1	100.0	25.9	0.0	37.0	100.0	63.0
	50 genes	33.3	6.9E-01	100.0	83.9	0.0	16.1	0.0	74.0	100.0	26.0	0.0	37.1	100.0	62.9
	PCA 40% variability	33.3	2.6E-02	100.0	83.9	0.0	16.1	0.0	73.1	100.0	26.9	0.0	38.1	100.0	61.9
	PCA 50% variability	33.3	2.6E-02	100.0	83.9	0.0	16.1	0.0	72.5	100.0	27.5	0.0	38.7	100.0	61.3
	PCA 60% variability	33.3	4.0E-02	100.0	83.9	0.0	16.1	0.0	73.1	100.0	26.9	0.0	38.1	100.0	61.9
	PCA 70% variability	33.3	4.9E-02	100.0	83.9	0.0	16.1	0.0	73.4	100.0	26.6	0.0	37.8	100.0	62.2
	PCA 80% variability	33.3	1.7E-02	99.8	83.9	0.2	16.1	0.0	74.1	100.0	25.9	0.0	37.0	100.0	63.0
	PCA 90% variability	32.4	3.5E-02	97.1	84.0	2.9	16.0	0.0	75.4	100.0	24.6	0.1	35.5	99.9	64.5

Mg vs Lgg vs Gb using eTumour consensus diagnosis

						Leave-one-out cross validation (LOOCV)							
		% accuracy	p-val	Sensitivity	Specificity	FNR	FPR	Sensitivity	Specificity	FNR Lgg	FPR Lgg	Sensitivity	Specificity	FNR	FPR
	Number of variables	(AUC)		Mg (%)	Mg (%)	Mg (%)	Mg (%)	Lgg (%)	Lgg (%)	(%)	(%)	Gb (%)	Gb (%)	Gb (%)	Gb (%)
Linear discriminant analysis (LDA)	*5 genes*	85.8	1.4E-06	100.0	100.0	0.0	0.0	66.7	93.0	33.3	7.0	90.6	80.8	9.4	19.2
	10 genes	76.3	3.4E-04	100.0	100.0	0.0	0.0	60.0	76.7	40.0	23.3	68.8	76.9	31.3	23.1
	20 genes	80.6	2.9E-04	100.0	100.0	0.0	0.0	66.7	81.4	33.3	18.6	75.0	80.8	25.0	19.2
	30 genes	92.4	8.4E-08	100.0	100.0	0.0	0.0	86.7	93.0	13.3	7.0	90.6	92.3	9.4	7.7
	40 genes	88.1	2.0E-05	100.0	100.0	0.0	0.0	80.0	88.4	20.0	11.6	84.4	88.5	15.6	11.5
	50 genes	74.3	8.1E-05	100.0	100.0	0.0	0.0	66.7	67.4	33.3	32.6	56.3	80.8	43.8	19.2
	PCA 40% variability	83.7	8.1E-05	100.0	100.0	0.0	0.0	66.7	88.4	33.3	11.6	84.4	80.8	15.6	19.2
	PCA 50% variability	87.1	8.4E-05	100.0	100.0	0.0	0.0	80.0	86.0	20.0	14.0	81.3	88.5	18.8	11.5
	PCA 60% variability	88.1	4.3E-06	100.0	100.0	0.0	0.0	80.0	88.4	20.0	11.6	84.4	88.5	15.6	11.5
	PCA 70% variability	89.2	6.6E-07	100.0	100.0	0.0	0.0	80.0	90.7	20.0	9.3	87.5	88.5	12.5	11.5
	PCA 80% variability	93.5	1.3E-08	100.0	100.0	0.0	0.0	86.7	95.3	13.3	4.7	93.8	92.3	6.3	7.7
	PCA 90% variability	89.3	8.5E-05	100.0	100.0	0.0	0.0	86.7	86.0	13.3	14.0	81.3	92.3	18.8	7.7
Support vector machine (svm)	*5 genes*	66.7	1.3E-03	100.0	100.0	0.0	0.0	0.0	100.0	100.0	0.0	100.0	42.3	0.0	57.7
	10 genes	66.7	8.3E-04	100.0	100.0	0.0	0.0	0.0	100.0	100.0	0.0	100.0	42.3	0.0	57.7
	20 genes	66.7	1.3E-03	100.0	100.0	0.0	0.0	0.0	100.0	100.0	0.0	100.0	42.3	0.0	57.7
	30 genes	66.7	1.3E-03	100.0	100.0	0.0	0.0	0.0	100.0	100.0	0.0	100.0	42.3	0.0	57.7
	40 genes	66.7	1.3E-03	100.0	100.0	0.0	0.0	0.0	100.0	100.0	0.0	100.0	42.3	0.0	57.7
	50 genes	66.7	2.0E-03	100.0	100.0	0.0	0.0	0.0	100.0	100.0	0.0	100.0	42.3	0.0	57.7
	PCA 40% variability	88.1	8.5E-05	100.0	100.0	0.0	0.0	80.0	88.4	20.0	11.6	84.4	88.5	15.6	11.5
	PCA 50% variability	81.8	5.3E-04	90.9	100.0	9.1	0.0	73.3	86.0	26.7	14.0	81.3	80.8	18.8	19.2
	PCA 60% variability	78.7	7.6E-04	81.8	100.0	18.2	0.0	66.7	88.4	33.3	11.6	87.5	76.9	12.5	23.1
	PCA 70% variability	74.2	4.6E-03	81.8	100.0	18.2	0.0	53.3	90.7	46.7	9.3	87.5	65.4	12.5	34.6
	PCA 80% variability	69.6	1.2E-02	81.8	100.0	18.2	0.0	33.3	95.3	66.7	4.7	93.8	53.8	6.3	46.2
	PCA 90% variability	42.0	3.0E-01	9.1	100.0	90.9	0.0	20.0	97.7	80.0	2.3	96.9	15.4	3.1	84.6
Random Forest (randF)	*5 genes*	78.1	5.1E-05	100.0	100.0	0.0	0.0	46.7	90.7	53.3	9.3	87.5	69.2	12.5	30.8
	10 genes	79.2	1.2E-05	100.0	100.0	0.0	0.0	53.3	88.4	46.7	11.6	84.4	73.1	15.6	26.9
	20 genes	76.0	1.4E-04	100.0	100.0	0.0	0.0	46.7	86.0	53.3	14.0	81.3	69.2	18.8	30.8
	30 genes	82.5	5.3E-04	100.0	100.0	0.0	0.0	60.0	90.7	40.0	9.3	87.5	76.9	12.5	23.1
	40 genes	84.7	2.2E-04	100.0	100.0	0.0	0.0	66.7	90.7	33.3	9.3	87.5	80.8	12.5	19.2
	50 genes	84.7	3.5E-04	100.0	100.0	0.0	0.0	66.7	90.7	33.3	9.3	87.5	80.8	12.5	19.2
	PCA 40% variability	80.4	1.8E-07	100.0	100.0	0.0	0.0	60.0	86.0	40.0	14.0	81.3	76.9	18.8	23.1
	PCA 50% variability	80.4	6.2E-06	100.0	100.0	0.0	0.0	60.0	86.0	40.0	14.0	81.3	76.9	18.8	23.1
	PCA 60% variability	84.7	1.0E-05	100.0	100.0	0.0	0.0	66.7	90.7	33.3	9.3	87.5	80.8	12.5	19.2
	PCA 70% variability	89.0	1.2E-05	100.0	100.0	0.0	0.0	73.3	95.3	26.7	4.7	93.8	84.6	6.3	15.4
	PCA 80% variability	83.5	8.5E-05	100.0	100.0	0.0	0.0	60.0	93.0	40.0	7.0	90.6	76.9	9.4	23.1
	PCA 90% variability	74.7	1.3E-03	100.0	100.0	0.0	0.0	33.3	93.0	66.7	7.0	90.6	61.5	9.4	38.5

Lgg vs Ag vs Gb using originating pathologist diagnosis

					5-fold cross validation (LOOCV)									
			Sensitivity	Specificity	FNR	FPR	Sensitivity	Specificity	FNR	FPR	Sensitivity	Specificity	FNR	FPR
Number of variables	% accuracy (AUC)	p-val	Lgg (%)	Lgg (%)	Lgg (%)	Lgg (%)	Ag (%)	Ag (%)	Ag (%)	Ag (%)	Gb (%)	Gb (%)	Gb (%)	Gb (%)
Linear discriminant analysis (LDA)														
5 genes	50.7	1.8E-01	49.3	76.4	50.7	23.6	22.4	80.7	77.6	19.3	80.4	78.8	19.6	21.2
10 genes	49.9	1.8E-01	47.4	77.4	52.6	22.6	26.1	77.4	73.9	22.6	76.3	79.3	23.7	20.7
20 genes	48.3	8.0E-01	47.8	77.3	52.2	22.7	29.3	73.2	70.7	26.8	67.9	79.8	32.1	20.2
30 genes	44.9	1.0E+00	46.1	77.0	53.9	23.0	30.1	69.2	69.9	30.8	58.6	77.4	41.4	22.6
40 genes	40.8	1.0E+00	42.4	73.9	57.6	26.1	30.9	68.3	69.1	31.7	49.0	72.6	51.0	27.4
50 genes	37.4	1.0E+00	38.9	69.7	61.1	30.3	30.6	69.0	69.4	31.0	42.7	69.5	57.3	30.5
PCA 40% variability	48.2	1.0E-05	58.4	74.0	41.6	26.0	4.7	87.0	95.3	13.0	81.5	71.5	18.5	28.5
PCA 50% variability	49.3	1.0E-06	49.1	81.1	50.9	18.9	15.4	83.3	84.6	16.7	83.4	68.6	16.6	31.4
PCA 60% variability	53.0	1.0E-06	55.3	80.9	44.7	19.1	26.6	79.8	73.4	20.2	77.2	76.9	22.8	23.1
PCA 70% variability	56.1	1.0E-06	65.2	83.0	34.8	17.0	36.0	80.1	64.0	19.9	67.1	74.7	32.9	25.3
PCA 80% variability	52.2	1.0E-06	63.4	82.1	36.6	17.9	37.8	73.3	62.2	26.7	55.4	75.4	44.6	24.6
PCA 90% variability	52.4	1.0E-06	62.6	76.8	37.4	23.2	44.4	76.0	55.6	24.0	50.2	76.2	49.8	23.8
Support vector machine (svm)														
5 genes	52.6	1.6E-02	65.5	70.5	34.5	29.5	5.1	93.8	94.9	6.2	87.2	75.4	12.8	24.6
10 genes	53.0	1.6E-02	65.2	71.1	34.8	28.9	7.4	93.6	92.6	6.4	86.4	75.0	13.6	25.0
20 genes	53.1	4.7E-03	67.0	70.6	33.0	29.4	6.9	93.9	93.1	6.1	85.6	75.4	14.4	24.6
30 genes	53.0	7.2E-03	68.4	70.0	31.6	30.0	5.1	94.3	94.9	5.7	85.6	75.5	14.4	24.5
40 genes	53.2	5.5E-03	69.4	69.9	30.6	30.1	4.9	94.7	95.1	5.3	85.3	75.5	14.7	24.5
50 genes	53.4	3.8E-03	70.2	69.9	29.8	30.1	4.7	95.0	95.3	5.0	85.1	75.3	14.9	24.7
PCA 40% variability	48.7	1.0E-06	55.2	76.0	44.8	24.0	13.2	84.2	86.8	15.8	77.8	71.6	22.2	28.4
PCA 50% variability	49.9	1.0E-06	50.3	79.3	49.7	20.7	17.1	86.4	82.9	13.6	82.2	67.0	17.8	33.0
PCA 60% variability	53.9	1.0E-06	44.2	83.2	55.8	16.8	30.3	91.4	69.7	8.6	87.2	61.2	12.8	38.8
PCA 70% variability	54.4	1.0E-06	48.9	85.5	51.1	14.5	27.9	93.0	72.1	7.0	86.3	56.8	13.7	43.2
PCA 80% variability	49.8	1.0E-06	45.2	89.6	54.8	10.4	13.4	94.2	86.6	5.8	90.7	45.2	9.3	54.8
PCA 90% variability	46.4	1.0E-06	30.4	93.4	69.6	6.6	16.6	97.0	83.4	3.0	92.4	30.6	7.6	69.4
random Forest (randF)														
5 genes	35.8	1.0E+00	92.2	76.9	7.8	23.1	11.1	75.6	88.9	24.4	4.0	46.6	96.0	53.4
10 genes	35.6	1.0E+00	92.7	76.9	7.3	23.1	10.6	76.0	89.4	24.0	3.6	46.1	96.4	53.9
20 genes	35.6	1.0E+00	93.6	76.7	6.4	23.3	9.6	76.2	90.4	23.8	3.5	46.1	96.5	53.9
30 genes	35.4	1.0E+00	94.3	76.6	5.7	23.4	8.4	76.2	91.6	23.8	3.5	46.2	96.5	53.8
40 genes	35.3	1.0E+00	94.3	76.6	5.7	23.4	7.9	76.2	92.1	23.8	3.6	46.2	96.4	53.8
50 genes	35.4	1.0E+00	95.0	76.5	5.0	23.5	7.7	76.3	92.3	23.7	3.5	46.2	96.5	53.8
PCA 40% variability	33.4	4.5E-05	93.4	76.9	6.6	23.1	1.8	75.6	98.2	24.4	4.9	46.3	95.1	53.7
PCA 50% variability	33.3	4.5E-05	92.0	76.9	8.0	23.1	0.2	75.9	99.8	24.1	7.8	46.0	92.2	54.0
PCA 60% variability	34.6	7.1E-06	94.9	77.4	5.1	22.6	0.2	76.1	99.8	23.9	8.8	45.7	91.2	54.3
PCA 70% variability	35.5	6.4E-07	94.7	77.7	5.3	22.3	1.1	76.6	98.9	23.4	10.7	45.1	89.3	54.9
PCA 80% variability	36.2	1.7E-07	95.7	78.2	4.3	21.8	0.3	77.2	99.7	22.8	12.5	44.1	87.5	55.9
PCA 90% variability	35.8	7.2E-07	92.6	79.2	7.4	20.8	0.0	77.4	100.0	22.6	14.9	42.9	85.1	57.1

Lgg vs Ag vs Gb using originating pathologist diagnosis

				Leave-one-out cross validation (LOOCV)									
	% accuracy		Sensitivity	Specificity	FNR	FPR	Sensitivity	Specificity	FNR	FPR	Sensitivity	Specificity	FNR	FPR
Number of variables	(AUC)	p-val	Lgg (%)	Lgg (%)	Lgg (%)	Lgg (%)	Ag (%)	Ag (%)	Ag (%)	Ag (%)	Gb (%)	Gb (%)	Gb (%)	Gb (%)
Linear discriminant analysis (LDA)														
5 genes	45.6	5.1E-01	50.0	72.3	50.0	27.7	12.5	79.6	87.5	20.4	74.2	76.5	25.8	23.5
10 genes	44.8	5.0E-01	44.4	72.3	55.6	27.7	12.5	79.6	87.5	20.4	77.4	76.5	22.6	23.5
20 genes	49.5	3.7E-01	55.6	72.3	44.4	27.7	18.8	77.6	81.3	22.4	74.2	85.3	25.8	14.7
30 genes	53.6	2.9E-01	55.6	83.0	44.4	17.0	37.5	69.4	62.5	30.6	67.7	85.3	32.3	14.7
40 genes	51.1	3.7E-01	66.7	80.9	33.3	19.1	18.8	73.5	81.3	26.5	67.7	79.4	32.3	20.6
50 genes	43.8	8.1E-01	61.1	70.2	38.9	29.8	25.0	71.4	75.0	28.6	45.2	76.5	54.8	23.5
PCA 40% variability	47.3	1.2E-02	61.1	70.2	38.9	29.8	0.0	89.8	100.0	10.2	80.6	70.6	19.4	29.4
PCA 50% variability	46.7	4.2E-03	50.0	78.7	50.0	21.3	6.3	83.7	93.8	16.3	83.9	67.6	16.1	32.4
PCA 60% variability	53.7	6.8E-04	55.6	78.7	44.4	21.3	25.0	81.6	75.0	18.4	80.6	79.4	19.4	20.6
PCA 70% variability	58.3	2.1E-05	66.7	85.1	33.3	14.9	43.8	77.6	56.3	22.4	64.5	76.5	35.5	23.5
PCA 80% variability	53.3	1.2E-04	61.1	85.1	38.9	14.9	37.5	75.5	62.5	24.5	61.3	70.6	38.7	29.4
PCA 90% variability	61.2	8.1E-05	72.2	80.9	27.8	19.1	50.0	79.6	50.0	20.4	61.3	82.4	38.7	17.6
Support vector machine (svm)														
5 genes	53.1	1.4E-01	72.2	68.1	27.8	31.9	0.0	98.0	100.0	2.0	87.1	73.5	12.9	26.5
10 genes	57.0	8.7E-02	77.8	70.2	22.2	29.8	6.3	98.0	93.8	2.0	87.1	76.5	12.9	23.5
20 genes	50.2	2.4E-01	66.7	63.8	33.3	36.2	0.0	93.9	100.0	6.1	83.9	79.4	16.1	20.6
30 genes	55.7	1.2E-01	83.3	63.8	16.7	36.2	0.0	100.0	100.0	0.0	83.9	79.4	16.1	20.6
40 genes	52.0	1.9E-01	72.2	66.0	27.8	34.0	0.0	95.9	100.0	4.1	83.9	76.5	16.1	23.5
50 genes	52.0	1.9E-01	72.2	66.0	27.8	34.0	0.0	95.9	100.0	4.1	83.9	76.5	16.1	23.5
PCA 40% variability	46.6	3.0E-02	50.0	76.6	50.0	23.4	12.5	79.6	87.5	20.4	77.4	73.5	22.6	26.5
PCA 50% variability	51.7	2.8E-02	55.6	76.6	44.4	23.4	18.8	87.8	81.3	12.2	80.6	70.6	19.4	29.4
PCA 60% variability	56.3	1.2E-03	44.4	85.1	55.6	14.9	37.5	89.8	62.5	10.2	87.1	64.7	12.9	35.3
PCA 70% variability	56.1	1.2E-02	50.0	85.1	50.0	14.9	31.3	93.9	68.8	6.1	87.1	58.8	12.9	41.2
PCA 80% variability	52.8	1.7E-02	55.6	89.4	44.4	10.6	12.5	93.9	87.5	6.1	90.3	50.0	9.7	50.0
PCA 90% variability	45.6	6.0E-02	27.8	91.5	72.2	8.5	18.8	95.9	81.3	4.1	90.3	32.4	9.7	67.6
Random Forest (randF)														
5 genes	43.0	5.1E-01	38.9	70.2	61.1	29.8	6.3	81.6	93.8	18.4	83.9	76.5	16.1	23.5
10 genes	49.9	2.2E-01	50.0	74.5	50.0	25.5	12.5	85.7	87.5	14.3	87.1	76.5	12.9	23.5
20 genes	44.1	4.3E-01	38.9	72.3	61.1	27.7	6.3	83.7	93.8	16.3	87.1	73.5	12.9	26.5
30 genes	50.4	2.4E-01	61.1	72.3	38.9	27.7	6.3	89.8	93.8	10.2	83.9	73.5	16.1	26.5
40 genes	50.4	2.4E-01	61.1	72.3	38.9	27.7	6.3	89.8	93.8	10.2	83.9	73.5	16.1	26.5
50 genes	50.4	2.4E-01	61.1	72.3	38.9	27.7	6.3	89.8	93.8	10.2	83.9	73.5	16.1	26.5
PCA 40% variability	48.5	3.9E-02	55.6	78.7	44.4	21.3	12.5	85.7	87.5	14.3	77.4	64.7	22.6	35.3
PCA 50% variability	49.8	8.4E-03	50.0	80.9	50.0	19.1	18.8	85.7	81.3	14.3	80.6	64.7	19.4	35.3
PCA 60% variability	52.1	4.4E-04	44.4	80.9	55.6	19.1	31.3	91.8	68.8	8.2	80.6	58.8	19.4	41.2
PCA 70% variability	54.0	4.0E-04	50.0	83.0	50.0	17.0	31.3	93.9	68.8	6.1	80.6	55.9	19.4	44.1
PCA 80% variability	53.3	1.0E-03	44.4	85.1	55.6	14.9	25.0	95.9	75.0	4.1	90.3	52.9	9.7	47.1
PCA 90% variability	44.8	8.0E-02	22.2	91.5	77.8	8.5	18.8	93.9	81.3	6.1	93.5	35.3	6.5	64.7

Lgg vs Ag vs Gb using eTumour consensus diagnosis

					5-fold cross validation (LOOCV)									
					FNR	FPR	Sensitivity	Specificity	FNR	FPR	Sensitivity	Specificity	FNR	FPR
	Number of variables	% accuracy (AUC)	p-val	Sensitivity Lgg (%)	Specificity Lgg (%)	Lgg (%)	Lgg (%)	Ag (%)	Ag (%)	Ag (%)	Gb (%)	Gb (%)	Gb (%)	Gb (%)
Linear discriminant analysis (LDA)	*5 genes*	51.1	3.7E-01	39.1	78.6	60.9	21.4	32.9	77.6	67.1	22.4	81.5	79.0	18.5	21.0
	10 genes	48.2	3.7E-01	37.3	76.6	62.7	23.4	31.2	76.3	68.8	23.7	76.2	77.3	23.8	22.7
	20 genes	48.6	9.7E-01	41.0	77.0	59.0	23.0	36.9	73.4	63.1	26.6	67.8	78.1	32.2	21.9
	30 genes	45.7	1.0E+00	39.9	75.4	60.1	24.6	36.7	71.4	63.3	28.6	60.4	76.0	39.6	24.0
	40 genes	41.0	1.0E+00	38.3	73.2	61.7	26.8	34.1	69.8	65.9	30.2	50.7	70.7	49.3	29.3
	50 genes	37.2	1.0E+00	39.5	67.8	60.5	32.2	32.1	69.0	67.9	31.0	39.8	69.5	60.2	30.5
	PCA 40% variability	47.6	1.1E-16	38.8	76.5	61.2	23.5	21.2	81.3	78.8	18.7	82.7	71.6	17.3	28.4
	PCA 50% variability	48.0	1.1E-16	36.2	79.0	63.8	21.0	22.8	80.6	77.2	19.4	85.2	70.7	14.8	29.3
	PCA 60% variability	52.2	1.0E-05	45.3	78.2	54.7	21.8	31.6	79.3	68.4	20.7	79.8	77.9	20.2	22.1
	PCA 70% variability	56.5	1.0E-05	60.6	81.1	39.4	18.9	37.0	80.9	63.0	19.1	71.7	74.5	28.3	25.5
	PCA 80% variability	57.5	1.0E-05	66.6	83.7	33.4	16.3	44.8	76.2	55.2	23.8	61.1	74.7	38.9	25.3
	PCA 90% variability	53.7	1.0E-05	64.0	78.2	36.0	21.8	43.6	75.6	56.4	24.4	53.3	74.9	46.7	25.1
Support vector machine (svm)	*5 genes*	52.0	1.8E-01	28.6	81.2	71.4	18.8	38.8	79.0	61.2	21.0	88.6	78.5	11.4	21.5
	10 genes	51.0	1.8E-01	32.6	78.5	67.4	21.5	32.5	81.5	67.5	18.5	87.8	76.5	12.2	23.5
	20 genes	51.0	4.3E-01	37.2	77.2	62.8	22.8	29.2	83.1	70.8	16.9	86.7	75.5	13.3	24.5
	30 genes	50.7	5.4E-01	39.4	76.1	60.6	23.9	26.1	83.6	73.9	16.4	86.5	75.5	13.5	24.5
	40 genes	50.5	6.0E-01	39.4	75.9	60.6	24.1	25.8	83.5	74.2	16.5	86.3	75.5	13.7	24.5
	50 genes	50.4	6.0E-01	39.9	75.8	60.1	24.2	24.9	83.6	75.1	16.4	86.4	75.4	13.6	24.6
	PCA 40% variability	43.0	6.2E-07	24.0	78.5	76.0	21.5	26.1	75.8	73.9	24.2	78.9	69.1	21.1	30.9
	PCA 50% variability	49.6	6.2E-07	25.4	85.6	74.6	14.4	38.8	79.2	61.2	20.8	84.6	67.1	15.4	32.9
	PCA 60% variability	53.5	1.0E-05	25.9	87.4	74.1	12.6	46.3	86.1	53.7	13.9	88.2	63.4	11.8	36.6
	PCA 70% variability	56.2	1.0E-05	37.6	90.0	62.4	10.0	43.8	89.6	56.2	10.4	87.3	57.9	12.7	42.1
	PCA 80% variability	52.9	1.0E-05	30.8	95.9	69.2	4.1	35.7	91.2	64.3	8.8	92.2	43.9	7.8	56.1
	PCA 90% variability	44.2	1.0E-05	15.2	96.5	84.8	3.5	23.6	95.0	76.4	5.0	93.7	26.1	6.3	73.9
Random Forest (randF)	*5 genes*	31.3	1.0E+00	66.8	78.1	33.2	21.9	10.6	73.2	89.4	26.8	16.6	46.4	83.4	53.6
	10 genes	31.1	1.0E+00	66.0	78.7	34.0	21.3	11.0	73.0	89.0	27.0	16.3	45.9	83.7	54.1
	20 genes	31.5	1.0E+00	66.5	78.8	33.5	21.2	11.5	73.1	88.5	26.9	16.6	45.8	83.4	54.2
	30 genes	31.5	1.0E+00	66.2	78.9	33.8	21.1	11.8	73.1	88.2	26.9	16.7	45.7	83.3	54.3
	40 genes	31.4	1.0E+00	65.8	78.7	34.2	21.3	11.9	73.2	88.1	26.8	16.7	45.8	83.3	54.2
	50 genes	31.8	1.0E+00	66.5	78.8	33.5	21.2	12.1	73.2	87.9	26.8	16.7	45.8	83.3	54.2
	PCA 40% variability	29.9	9.9E-01	67.0	78.7	33.0	21.3	4.3	73.3	95.7	26.7	18.3	45.4	81.7	54.6
	PCA 50% variability	30.0	9.9E-01	62.0	78.9	38.0	21.1	8.3	73.3	91.7	26.7	19.8	45.3	80.2	54.7
	PCA 60% variability	31.6	9.6E-01	65.9	79.3	34.1	20.7	8.1	73.2	91.9	26.8	20.8	45.2	79.2	54.8
	PCA 70% variability	30.4	1.0E+00	64.4	79.6	35.6	20.4	4.8	73.6	95.2	26.4	22.0	44.4	78.0	55.6
	PCA 80% variability	29.7	9.9E-01	60.8	80.9	39.2	19.1	4.3	73.6	95.7	26.4	24.1	43.0	75.9	57.0
	PCA 90% variability	29.6	1.0E+00	58.7	81.7	41.3	18.3	3.0	74.3	97.0	25.7	27.2	41.6	72.8	58.4

Lgg vs Ag vs Gb using eTumour consensus diagnosis

			Leave-one-out cross validation (LOOCV)											
		% accuracy	Sensitivity	Specificity	FNR	FPR	Sensitivity	Specificity	FNR	FPR	Sensitivity	Specificity	FNR	FPR
	Number of variables	(AUC)	p-val	Lgg (%)	Lgg (%)	Lgg (%)	Lgg (%)	Ag (%)	Ag (%)	Ag (%)	Gb (%)	Gb (%)	Gb (%)	Gb (%)
Linear discriminant analysis (LDA)	*5 genes*	55.1	1.7E-02	33.3	84.0	66.7	16.0	44.4	80.9	55.6	19.1	87.5	78.8	12.5	21.2
	10 genes	47.5	9.9E-02	33.3	74.0	66.7	26.0	27.8	80.9	72.2	19.1	81.3	78.8	18.8	21.2
	20 genes	54.1	9.3E-02	46.7	78.0	53.3	22.0	50.0	74.5	50.0	25.5	65.6	84.8	34.4	15.2
	30 genes	45.8	3.7E-01	53.3	70.0	46.7	30.0	27.8	74.5	72.2	25.5	56.3	78.8	43.8	21.2
	40 genes	49.7	3.0E-01	46.7	78.0	53.3	22.0	55.6	70.2	44.4	29.8	46.9	75.8	53.1	24.2
	50 genes	50.2	2.5E-01	66.7	78.0	33.3	22.0	27.8	72.3	72.2	27.7	56.3	75.8	43.8	24.2
	PCA 40% variability	44.4	1.4E-01	26.7	78.0	73.3	22.0	22.2	78.7	77.8	21.3	84.4	72.7	15.6	27.3
	PCA 50% variability	46.6	6.7E-02	33.3	80.0	66.7	20.0	22.2	80.9	77.8	19.1	84.4	69.7	15.6	30.3
	PCA 60% variability	53.4	2.8E-02	40.0	82.0	60.0	18.0	38.9	76.6	61.1	23.4	81.3	81.8	18.8	18.2
	PCA 70% variability	58.0	1.0E-02	60.0	84.0	40.0	16.0	38.9	76.6	61.1	23.4	75.0	81.8	25.0	18.2
	PCA 80% variability	59.1	2.4E-03	73.3	84.0	26.7	16.0	44.4	76.6	55.6	23.4	59.4	75.8	40.6	24.2
	PCA 90% variability	63.1	1.9E-02	80.0	82.0	20.0	18.0	50.0	76.6	50.0	23.4	59.4	84.8	40.6	15.2
Support vector machine (svm)	*5 genes*	38.0	2.5E-01	6.7	76.0	93.3	24.0	16.7	72.3	83.3	27.7	90.6	78.8	9.4	21.2
	10 genes	36.9	3.0E-01	6.7	76.0	93.3	24.0	16.7	72.3	83.3	27.7	87.5	78.8	12.5	24.2
	20 genes	36.9	2.9E-01	6.7	74.0	93.3	26.0	16.7	78.7	83.3	21.3	87.5	69.7	12.5	30.3
	30 genes	37.3	2.9E-01	13.3	72.0	86.7	28.0	11.1	76.6	88.9	23.4	87.5	75.8	12.5	24.2
	40 genes	35.1	3.6E-01	6.7	72.0	93.3	28.0	11.1	74.5	88.9	25.5	87.5	75.8	12.5	24.2
	50 genes	33.2	4.3E-01	6.7	70.0	93.3	30.0	5.6	76.6	94.4	23.4	87.5	72.7	12.5	27.3
	PCA 40% variability	35.7	1.0E-01	6.7	80.0	93.3	20.0	22.2	66.0	77.8	34.0	78.1	72.7	21.9	27.3
	PCA 50% variability	45.9	3.0E-02	20.0	82.0	80.0	18.0	33.3	78.7	66.7	21.3	84.4	69.7	15.6	30.3
	PCA 60% variability	50.3	8.4E-03	13.3	88.0	86.7	12.0	50.0	83.0	50.0	17.0	87.5	63.6	12.5	36.4
	PCA 70% variability	56.9	8.2E-03	33.3	90.0	66.7	10.0	50.0	89.4	50.0	10.6	87.5	60.6	12.5	39.4
	PCA 80% variability	56.5	1.2E-02	40.0	98.0	60.0	2.0	38.9	91.5	61.1	8.5	90.6	45.5	9.4	54.5
	PCA 90% variability	45.0	1.2E-01	13.3	98.0	86.7	2.0	27.8	93.6	72.2	6.4	93.8	27.3	6.3	72.7
Random Forest (randF)	*5 genes*	52.9	3.9E-02	46.7	82.0	53.3	18.0	27.8	80.9	72.2	19.1	84.4	75.8	15.6	24.2
	10 genes	54.0	2.1E-02	46.7	80.0	53.3	20.0	27.8	87.2	72.2	12.8	87.5	72.7	12.5	27.3
	20 genes	46.9	6.7E-02	33.3	76.0	66.7	24.0	16.7	85.1	83.3	14.9	90.6	72.7	9.4	27.3
	30 genes	49.9	4.9E-02	40.0	80.0	60.0	20.0	22.2	80.9	77.8	19.1	87.5	75.8	12.5	24.2
	40 genes	47.7	6.0E-02	33.3	80.0	66.7	20.0	22.2	78.7	77.8	21.3	87.5	75.8	12.5	24.2
	50 genes	54.7	1.2E-02	40.0	86.0	60.0	14.0	33.3	80.9	66.7	19.1	90.6	75.8	9.4	24.2
	PCA 40% variability	42.7	2.1E-02	33.3	80.0	66.7	20.0	16.7	78.7	83.3	21.3	78.1	63.6	21.9	36.4
	PCA 50% variability	45.2	1.2E-02	26.7	84.0	73.3	16.0	27.8	80.9	72.2	19.1	81.3	60.6	18.8	39.4
	PCA 60% variability	42.6	1.5E-02	13.3	80.0	86.7	20.0	33.3	85.1	66.7	14.9	81.3	57.6	18.8	42.4
	PCA 70% variability	55.9	5.5E-03	33.3	88.0	66.7	12.0	50.0	89.4	50.0	10.6	84.4	60.6	15.6	39.4
	PCA 80% variability	50.6	1.0E-02	20.0	92.0	80.0	8.0	44.4	89.4	55.6	10.6	87.5	48.5	12.5	51.5
	PCA 90% variability	52.3	3.2E-02	26.7	94.0	73.3	6.0	33.3	93.6	66.7	6.4	96.9	45.5	3.1	54.5

A-10 Functionally relevant genes for meningeal and glial tumours based on Affymetrix data

Functional group 1							P-value = 7.12E-18			
			Accesion	Ag/Gb			Lgg/Ag		Lgg/Gb	
Affymetrix probeset	Gene symbol	Gene Description	number	ratio	q-value	ratio	q-value	ratio	q-value
213285_at	*TMEM30B*	Transmembrane Protein 30B	AV691491	1.75	4.5E-01	0.87	1.9E-01	0.5	8.2E-01
1559072_a_at	*ELFN2*	Extracellular Leucine-Rich Repeat And Fibronectin Type III Domain Containing 2	BC032083	2.42	1.2E-01	2.96	5.1E-04	1.22	8.2E-01
1558517_s_at	*LRRC8C*	Leucine Rich Repeat Containing 8 Family, Member C	CA773938	0.98	6.0E-01	1.06	4.3E-01	1.08	8.6E-01
1552848_a_at	*PTCHD1*	Patched Domain Containing 1	NM_173495	0.72	2.8E-01	0.7	1.2E-01	0.97	8.9E-01
212976_at	*LRRC8B*	Leucine Rich Repeat Containing 8 Family, Member B	R41498	1.91	2.1E-01	2.06	7.0E-02	1.08	8.9E-01
1557176_a_at	*C14orf37*	Chromosome 14 Open Reading Frame 37	BU074567	0.87	4.4E-01	1.21	2.5E-01	1.38	8.2E-01
1558693_s_at	*C1orf85*	Chromosome 1 Open Reading Frame 85	AW090182	0.78	4.1E-01	0.45	1.6E-04	0.58	8.2E-01
227001_at	*NPAL2*	NIPA-Like Domain Containing 2	AI096706	1.21	4.7E-01	0.65	1.2E-01	0.54	8.2E-01
243756_at	*THSD7A*	Thrombospondin, Type I, Domain Containing 7A	AI057226	1.2	4.7E-01	1.06	4.8E-01	0.89	8.6E-01
229254_at	*MFSD4*	Major Facilitator Superfamily Domain Containing 4	BE550027	2.51	2.4E-01	2.21	1.4E-01	0.88	8.9E-01
242372_s_at	*MFSD4*	Major Facilitator Superfamily Domain Containing 4	AL542291	2.04	2.5E-01	2.26	1.3E-01	1.1	8.9E-01
202016_at	*MEST*	Mesoderm Specific Transcript Homolog (Mouse)	NM_002402	0.55	2.1E-01	0.39	2.7E-02	0.71	8.3E-01
Functional group 2							**P-value = 2.53E-16**		
206773_at	*LY6H*	Lymphocyte Antigen 6 Complex, Locus H	NM_002347	1.6	3.6E-01	2.39	6.5E-02	1.49	8.2E-01
227566_at	*HNT*	Neurotrimin	AW085558	1.68	1.6E-01	1.82	3.1E-03	1.08	8.7E-01
214111_at	*OPCML*	Opioid Binding Protein/Cell Adhesion Molecule-Like	AF070577	1.93	2.3E-01	1.76	8.5E-02	0.91	8.8E-01
223730_at	*GPC6*	Glypican 6	AF111178	0.66	1.4E-01	0.62	1.0E-02	0.94	8.9E-01
227059_at	*GPC6*	Glypican 6	AI651255	0.49	5.5E-02	0.55	8.9E-03	1.1	8.7E-01
207174_at	*GPC5*	Glypican 5	NM_004466	1.7	3.5E-01	1.43	2.6E-01	0.84	8.7E-01
Functional group 3							**P-value = 8.57E-12**		
213880_at	*LGR5*	Leucine-Rich Repeat-Containing G Protein-Coupled Receptor 5	AL524520	1.7	4.4E-01	1.13	4.8E-01	0.66	8.4E-01
236734_at	*SLITRK1*	SLIT and NTRK-Like Family, Member 1	AI565671	3.4	8.2E-02	3.05	2.0E-02	0.9	8.8E-01
222108_at	*AMIGO2*	Adhesion Molecule with Ig-Like Domain 2	AC004010	0.51	1.4E-01	0.33	5.2E-03	0.64	8.2E-01
214930_at	*SLITRK5*	Slit And NTRK-Like Family, Member 5	AW449813	2.36	1.4E-01	2.78	2.5E-03	1.17	8.5E-01
226908_at	*LRIG3*	Leucine-Rich Repeats And Immunoglobulin-Like Domains 3	AI627704	0.22	3.3E-01	0.18	2.0E-01	0.81	8.2E-01
211596_s_at	*LRIG1*	Leucine-Rich Repeats and Immunoglobulin-Like Domains 1	AB050468	1.14	4.9E-01	1.7	7.9E-03	1.49	8.2E-01
230644_at	*LRFN5*	Leucine Rich Repeat And Fibronectin Type III Domain Containing 5	AI375083	2.97	1.8E-01	2.33	7.3E-02	0.78	8.5E-01
209840_s_at	*LRRN3*	Leucine Rich Repeat Neuronal 3	AI221950	0.75	2.5E-01	0.93	4.4E-01	1.25	8.2E-01
209841_s_at	*LRRN3*	Leucine Rich Repeat Neuronal 3	AL442092	0.72	2.0E-01	0.95	4.5E-01	1.31	8.2E-01

Affymetrix probeset	Gene symbol	Gene Description	Accesion number	Ag/Gb ratio	q-value	Lgg/Ag ratio	q-value	Lgg/Gb ratio	q-value
226884_at	*LRRN1*	Leucine Rich Repeat Neuronal 1	N71874	1.96	2.3E-01	2.57	3.1E-02	1.31	8.4E-01
207191_s_at	*ISLR*	Immunoglobulin Superfamily Containing Leucine-Rich Repeat	NM_005545	0.43	1.5E-01	0.32	2.3E-02	0.74	8.2E-01
232720_at	*LINGO2*	Leucine Rich Repeat Neuronal 6C	AL353746	2.29	1.6E-01	1.65	1.8E-01	0.72	8.2E-01
207093_s_at	*OMG*	Oligodendrocyte Myelin Glycoprotein	NM_002544	2.15	1.0E-01	2.39	4.5E-04	1.11	8.5E-01
222853_at	*FLRT3*	Fibronectin Leucine Rich Transmembrane Protein 3	N71923	1.41	3.3E-01	1.47	1.0E-01	1.04	8.9E-01
203476_at	*TPBG*	Trophoblast Glycoprotein	NM_006670	0.56	2.7E-01	0.36	5.7E-02	0.66	8.2E-01
206408_at	*LRRTM2*	Leucine Rich Repeat Transmembrane Neuronal 2	NM_015564	2.23	1.1E-01	2.1	2.4E-03	0.94	8.9E-01
204359_at	*FLRT2*	Fibronectin Leucine Rich Transmembrane Protein 2	NM_013231	1.32	4.2E-01	1.14	4.2E-01	0.86	8.7E-01
1557123_a_at	*RP4-756G23.1*	Hypothetical Protein Bc012882	BC040188	1.14	4.1E-01	1.26	1.5E-01	1.1	8.5E-01
229085_at	*LRRC3B*	Leucine Rich Repeat Containing 3b	AW027879	1.05	6.2E-01	1	5.4E-01	0.95	8.9E-01
205150_s_at	*KIAA0644*	Kiaa0644 Gene Product	AV724192	0.76	4.8E-01	0.9	4.6E-01	1.19	8.6E-01
Functional group 4						**P-value =**	**1.31E-11**		
202709_at	*FMOD*	Fibromodulin	NM_002023	0.63	2.5E-01	0.35	7.4E-03	0.55	8.2E-01
201744_s_at	*LUM*	Lumican	NM_002345	0.31	1.6E-01	0.08	8.3E-03	0.26	8.2E-01
219087_at	*ASPN*	Asporin (LRR Class 1)	NM_017680	0.23	1.2E-01	0.15	2.0E-02	0.65	8.2E-01
204223_at	*PRELP*	Proline/Arginine-Rich End Leucine-Rich Repeat Protein	NM_002725	1.17	4.0E-01	1.17	4.0E-01	1	9.1E-01
205907_s_at	*OMD*	Osteomodulin	AI765819	0.88	5.3E-01	1.11	4.8E-01	1.26	8.6E-01
205908_s_at	*OMD*	Osteomodulin	NM_005014	0.85	4.6E-01	0.99	5.3E-01	1.16	8.7E-01
211896_s_at	*DCN*	Decorin	AF138302	0.47	1.9E-01	0.3	2.2E-02	0.64	8.2E-01
209335_at	*DCN*	Decorin	AI281593	0.46	2.0E-01	0.25	1.7E-02	0.54	8.2E-01
211813_x_at	*DCN*	Decorin	AF138303	0.48	1.9E-01	0.34	2.8E-02	0.72	8.4E-01
201893_x_at	*DCN*	Decorin	AF138300	0.54	2.1E-01	0.36	2.5E-02	0.67	8.2E-01
218730_s_at	*OGN*	Osteoglycin (Osteoinductive Factor, Mimecan)	NM_014057	0.38	4.4E-01	0.48	3.6E-01	1.25	8.9E-01
222722_at	*OGN*	Osteoglycin (Osteoinductive Factor, Mimecan)	AV700059	0.39	4.2E-01	0.43	3.2E-01	1.12	8.9E-01
Functional group 5						**P-value =**	**2.04E-11**		
205691_at	*SYNGR3*	Synaptogyrin 3	NM_004209	2.25	2.7E-01	2.12	1.5E-01	0.94	8.9E-01
242002_at	*TCBA1*	T-Cell Lymphoma Breakpoint Associated Target 1	N62814	2.04	2.6E-01	2.33	6.1E-02	1.14	8.8E-01
228017_s_at	*C20orf58*	Chromosome 20 Open Reading Frame 58	BF593263	1.13	5.5E-01	1.81	4.8E-02	1.61	8.2E-01
228018_at	*C20orf58*	Chromosome 20 Open Reading Frame 58	BF593263	1.05	6.1E-01	1.76	4.1E-02	1.68	8.2E-01
1558387_at	*FAM77D*	Family With Sequence Similarity 77, Member D	R41806	0.79	4.6E-01	2.23	1.7E-02	2.82	6.7E-01
1558388_a_at	*FAM77D*	Family With Sequence Similarity 77, Member D	R41806	1.09	5.9E-01	2.37	5.2E-03	2.18	8.2E-01
209655_s_at	*TMEM47*	Transmembrane Protein 47	AI803181	0.98	6.2E-01	1.36	2.1E-01	1.39	8.2E-01

Functional group 6						P-value = 4.84E-11			
			Accesion	Ag/Gb		Lgg/Ag		Lgg/Gb	
Affymetrix probeset	Gene symbol	Gene Description	number	ratio	q-value	ratio	q-value	ratio	q-value
223629_at	*PCDHB5*	Protocadherin Beta 5	BC001186	1.34	5.1E-01	0.66	2.1E-01	0.49	8.2E-01
1553105_s_at	*DSG2*	Desmoglein 2	NM_001943	1.05	6.3E-01	0.48	1.3E-01	0.46	8.2E-01
217901_at	*DSG2*	Desmoglein 2	BF031829	1.16	6.1E-01	0.18	6.9E-02	0.16	8.2E-01
203440_at	*CDH2*	Cadherin 2, Type 1, N-Cadherin (Neuronal)	M34064	0.66	4.4E-02	0.73	3.4E-02	1.1	8.4E-01
205656_at	*PCDH17*	Protocadherin 17	NM_014459	1.13	4.6E-01	1.33	4.8E-02	1.18	8.2E-01
228863_at	*PCDH17*	Protocadherin 17	N69091	1.1	5.1E-01	1.34	3.4E-02	1.22	8.2E-01
207149_at	*CDH12*	Cadherin 12, Type 2 (N-Cadherin 2)	L33477	1.95	2.9E-01	1.48	3.1E-01	0.76	8.5E-01
231738_at	*PCDHB7*	Protocadherin Beta 7	NM_018940	0.53	2.8E-01	0.31	2.7E-02	0.59	8.4E-01
228635_at	*PCDH10*	Protocadherin 10	AI640307	1.07	5.7E-01	1.36	1.4E-01	1.27	8.2E-01
206935_at	*PCDH8*	Protocadherin 8	NM_002590	1.33	4.9E-01	1.32	3.3E-01	0.99	9.0E-01
210292_s_at	*PCDH11X*	Protocadherin 11 X-Linked	AF332218	2.16	1.7E-01	2.09	1.9E-02	0.97	8.9E-01
219737_s_at	*PCDH9*	Protocadherin 9	AI524125	1.12	5.4E-01	1.07	4.4E-01	0.96	8.9E-01
219738_s_at	*PCDH9*	Protocadherin 9	NM_020403	1.12	5.5E-01	1.18	3.3E-01	1.05	8.9E-01
238919_at	*PCDH9*	Protocadherin 9	R49295	0.97	6.2E-01	1.08	4.2E-01	1.11	8.7E-01
220115_s_at	*CDH10*	Cadherin 10, Type 2 (T2-Cadherin)	NM_006727	1.77	1.4E-01	2.37	1.1E-04	1.34	8.2E-01
228640_at	*PCDH7*	Bh-Protocadherin (Brain-Heart)	BE644809	2.22	1.7E-01	2.53	5.3E-03	1.14	8.7E-01
201131_s_at	*CDH1*	Cadherin 1, Type 1, E-Cadherin (Epithelial)	NM_004360	4	4.4E-01	1.29	3.9E-01	0.32	8.4E-01
206898_at	*CDH19*	Cadherin 19, Type 2	NM_021153	0.95	6.2E-01	0.32	2.7E-02	0.34	8.2E-01
206280_at	*CDH18*	Cadherin 18, Type 2	NM_004934	3.39	1.5E-01	2.66	2.5E-02	0.78	8.5E-01
Functional group 7						**P-value = 1.05E-10**			
213194_at	*ROBO1*	Roundabout, Axon Guidance Receptor, Homolog 1 (Drosophila)	BF059159	1.06	6.0E-01	1.2	3.1E-01	1.13	8.7E-01
226766_at	*ROBO2*	Roundabout, Axon Guidance Receptor, Homolog 2 (Drosophila)	AB046788	0.62	2.3E-01	0.95	5.0E-01	1.55	8.2E-01
216617_s_at	*MAG*	Myelin Associated Glycoprotein	X98405	1.07	6.1E-01	1.33	3.3E-01	1.24	8.5E-01
214111_at	*OPCML*	Opioid Binding Protein/Cell Adhesion Molecule-Like	AF070577	1.93	2.3E-01	1.76	8.5E-02	0.91	8.8E-01
203868_s_at	*VCAM1*	Vascular Cell Adhesion Molecule 1	NM_001078	0.5	1.7E-01	0.65	2.2E-01	1.31	8.7E-01
204105_s_at	*NRCAM*	Neuronal Cell Adhesion Molecule	NM_005010	0.8	2.6E-01	0.79	7.6E-02	0.99	9.0E-01
233401_at	*NRCAM*	Neuronal Cell Adhesion Molecule	BF723605	0.91	5.3E-01	0.98	5.1E-01	1.08	8.9E-01
230045_at	*CNTN2*	Contactin 2 (Axonal)	BF740264	1.32	5.1E-01	1.39	3.1E-01	1.06	8.9E-01
215028_at	*SEMA6A*	Sema Domain, Transmembrane Domain (Tm), And Cytoplasmic Domain, (Semaphorin) 6a	AB002438	0.76	3.3E-01	1.02	5.1E-01	1.35	8.2E-01
225660_at	*SEMA6A*	Sema Domain, Transmembrane Domain (Tm), And Cytoplasmic Domain, (Semaphorin) 6A	W92748	0.73	2.2E-01	1.02	5.1E-01	1.4	8.2E-01
212843_at	*NCAM1*	Neural Cell Adhesion Molecule 1	AA126505	1.25	3.2E-01	1.47	1.7E-02	1.18	8.2E-01
213438_at	*NFASC*	Neurofascin Homolog (Chicken)	AA995925	0.96	6.1E-01	0.99	5.2E-01	1.03	8.9E-01
204584_at	*L1CAM*	L1 Cell Adhesion Molecule	AI653981	3.04	1.7E-01	2.4	5.6E-02	0.79	8.5E-01
227566_at	*HNT*	Neurotrimin	AW085558	1.68	1.6E-01	1.82	3.1E-03	1.08	8.7E-01

Affymetrix probeset	Gene symbol	Gene Description	Accesion number	Ag/Gb ratio	q-value	Lgg/Ag ratio	q-value	Lgg/Gb ratio	q-value
1554784_at	*CNTN1*	Contactin 1	BC036569	2.22	1.8E-01	2.57	1.2E-03	1.16	8.6E-01
227202_at	*CNTN1*	Contactin 1	AW072790	2.07	2.1E-01	2.54	2.1E-03	1.23	8.4E-01
227209_at	*CNTN1*	Contactin 1	AI091445	2.42	1.5E-01	2.86	1.2E-04	1.18	8.5E-01
1553194_at	*NEGR1*	Neuronal Growth Regulator 1	NM_173808	1.52	3.1E-01	2.28	5.3E-02	1.5	8.2E-01
243357_at	*NEGR1*	Neuronal Growth Regulator 1	AA115106	2.7	1.8E-01	2.87	2.6E-02	1.06	8.9E-01
229461_x_at	*NEGR1*	Neuronal Growth Regulator 1	AI123532	2.26	2.1E-01	2.72	3.5E-02	1.2	8.6E-01
229831_at	*CNTN3*	Contactin 3 (Plasmacytoma Associated)	BE221817	2.71	2.1E-01	4.4	6.4E-03	1.62	8.2E-01
204006_s_at	*FCGR3B*	Fc Fragment of Igg, Low Affinity IIIb, Receptor (CD16B)	NM_000570	0.58	3.4E-01	0.49	1.3E-01	0.85	8.9E-01
Functional group 8						**P-value =**	**2.48E-10**		
232010_at	*FSTL5*	Follistatin-Like 5	AA129444	1.74	4.2E-01	2.06	1.6E-01	1.18	8.7E-01
206434_at	*SPOCK3*	Sparc/Osteonectin, CWCV And Kazal-Like Domains Proteoglycan (Testican) 3	NM_016950	1.28	5.3E-01	1.48	2.8E-01	1.16	8.7E-01
235342_at	*SPOCK3*	Sparc/Osteonectin, CWCV And Kazal-Like Domains Proteoglycan (Testican) 3	AI808090	1.53	4.0E-01	1.72	1.4E-01	1.13	8.8E-01
202363_at	*SPOCK1*	Sparc/Osteonectin, CWCV And Kazal-Like Domains Proteoglycan (Testican) 1	AF231124	1.16	5.1E-01	1.4	1.5E-01	1.2	8.4E-01
222784_at	*SMOC1*	Sparc Related Modular Calcium Binding 1	AJ249900	2.7	2.0E-01	4.21	2.6E-03	1.56	8.2E-01
223235_s_at	*SMOC2*	Sparc Related Modular Calcium Binding 2	AB014737	0.31	9.2E-02	0.31	1.9E-02	0.99	9.0E-01
223885_at	*CALN1*	Calneuron 1	AF282250	4.12	1.3E-01	3.71	6.6E-02	0.9	8.9E-01
Functional group 9						**P-value =**	**4.19E-10**		
210220_at	*FZD2*	Frizzled Homolog 2 (Drosophila)	AW072102	3.69	1.3E-01	3.62	3.7E-02	0.98	9.0E-01
205638_at	*BAI3*	Brain-Specific Angiogenesis Inhibitor 3	L37882	0.64	3.0E-01	0.46	2.9E-02	0.71	8.2E-01
203705_s_at	*FZD7*	Frizzled Homolog 7 (Drosophila)	NM_001704	1.95	1.5E-01	2.32	2.1E-04	1.19	8.3E-01
203706_s_at	*FZD7*	Frizzled Homolog 7 (Drosophila)	AI333651	0.38	6.3E-02	0.29	9.6E-04	0.76	8.5E-01
226690_at	*ADCYAP1R1*	Adenylate Cyclase Activating Polypeptide 1 (Pituitary) Receptor Type I	NM_003507	0.41	7.3E-02	0.28	6.8E-04	0.69	8.3E-01
227769_at	*GPR27*	G Protein-Coupled Receptor 27	AW451961	1.25	4.4E-01	2.25	4.2E-03	1.8	8.2E-01
205814_at	*GRM3*	Glutamate Receptor, Metabotropic 3	AI703476	2.39	1.6E-01	3.06	8.6E-04	1.28	8.2E-01
206825_at	*OXTR*	Oxytocin Receptor	NM_000840	1.96	2.7E-01	2.08	1.1E-01	1.06	8.9E-01
214217_at	*GRM5*	Glutamate Receptor, Metabotropic 5	NM_000916	0.42	3.0E-01	0.19	7.8E-02	0.45	8.2E-01
209990_s_at	*GABBR2*	Gamma-Aminobutyric Acid (GABA) B Receptor, 2	D60132	3.09	1.6E-01	3.59	1.3E-02	1.16	8.7E-01
209991_x_at	*GABBR2*	Gamma-Aminobutyric Acid (GABA) B Receptor, 2	AF056085	1.9	2.6E-01	3.37	6.5E-03	1.77	8.2E-01
211679_x_at	*GABBR2*	Gamma-Aminobutyric Acid (GABA) B Receptor, 2	AF069755	2.01	2.1E-01	3.66	1.3E-02	1.83	8.2E-01
217077_s_at	*GABBR2*	Gamma-Aminobutyric Acid (GABA) B Receptor, 2	AF095784	1.86	2.3E-01	3.09	7.7E-03	1.66	8.2E-01
1554008_at	*OSMR*	Oncostatin M Receptor	AF095723	1.93	2.4E-01	3.36	1.1E-02	1.74	8.2E-01
206899_at	*NTSR2*	Neurotensin Receptor 2	BC010943	0.45	6.0E-02	0.33	6.5E-04	0.74	8.2E-01

Affymetrix probeset	Gene symbol	Gene Description	Accesion number	Ag/Gb ratio	q-value	Lgg/Ag ratio	q-value	Lgg/Gb ratio	q-value
1560225_at	*CNR1*	Cannabinoid Receptor 1 (Brain)	NM_012344	2.67	1.9E-01	4.15	5.6E-02	1.56	8.2E-01
213436_at	*CNR1*	Cannabinoid Receptor 1 (Brain)	AI434253	0.44	1.6E-01	0.42	5.1E-02	0.97	8.9E-01
1555533_at	*GPR103*	G Protein-Coupled Receptor 103	U73304	0.5	1.5E-01	0.45	2.5E-02	0.9	8.9E-01
213592_at	*AGTRL1*	Angiotensin II Receptor-Like 1	AF411117	1.03	4.5E-01	1.03	3.2E-01	1	9.0E-01
206785_s_at	*KLRC2*	Killer Cell Lectin-Like Receptor Subfamily C, Member 2	X89271	0.63	2.5E-01	0.75	2.7E-01	1.19	8.6E-01
220005_at	*P2RY13*	Purinergic Receptor P2Y, G-Protein Coupled, 13	NM_002260	7.71	1.4E-01	7.02	8.1E-03	0.91	8.9E-01
215306_at	*STON1*	Stonin 1	NM_023914	1.08	6.0E-01	2.01	5.0E-02	1.87	8.2E-01
232195_at	*GPR158*	G Protein-Coupled Receptor 158	AL049443	0.65	2.5E-01	1.53	2.5E-01	2.35	8.2E-01
209631_s_at	*GPR37*	G Protein-Coupled Receptor 37 (Endothelin Receptor Type B-Like)	R41459	2.12	1.7E-01	2.59	1.5E-03	1.22	8.4E-01
214586_at	*GPR37*	G Protein-Coupled Receptor 37 (Endothelin Receptor Type B-Like)	U87460	0.71	3.1E-01	0.68	1.4E-01	0.97	9.0E-01
213506_at	*F2RL1*	Coagulation Factor II (Thrombin) Receptor-Like 1	T16257	0.64	2.3E-01	0.71	1.9E-01	1.11	8.8E-01
1553316_at	*GPR82*	G Protein-Coupled Receptor 82	BE965369	0.65	3.0E-01	0.47	4.7E-02	0.74	8.5E-01
1553317_s_at	*GPR82*	G Protein-Coupled Receptor 82	NM_080817	0.85	4.4E-01	0.77	2.1E-01	0.91	8.5E-01
206785_s_at	*KLRC2*	Killer Cell Lectin-Like Receptor Subfamily C, Member 1	AL832460	0.77	4.2E-01	0.59	6.7E-02	0.77	8.2E-01
1554706_at	*OR2L13*	Olfactory Receptor, Family 2, Subfamily L, Member 13	NM_002260	7.71	1.4E-01	7.02	8.1E-03	0.91	8.9E-01
219896_at	*DRD1IP*	Dopamine Receptor D1 Interacting Protein	BC028158	1.53	3.2E-01	1.28	2.6E-01	0.83	8.5E-01
206190_at	*GPR17*	G Protein-Coupled Receptor 17	NM_015722	2.44	2.2E-01	2.17	1.4E-01	0.89	8.9E-01
215225_s_at	*GPR17*	G Protein-Coupled Receptor 17	NM_005291	2.3	4.0E-01	2.6	2.2E-01	1.13	8.9E-01
AFFX-HUMRGE/M10098_M_at	*GPR34*	G Protein-Coupled Receptor 34	Z94154	2.4	3.9E-01	2.62	1.9E-01	1.09	8.9E-01
221288_at	*GPR22*	G Protein-Coupled Receptor 22	-	0.31	5.5E-02	0.45	2.4E-02	1.45	8.2E-01
224102_at	*P2RY12*	Purinergic Receptor P2Y, G-Protein Coupled, 12	NM_005295	2.45	2.6E-01	1.79	2.1E-01	0.73	8.5E-01
235885_at	*P2RY12*	Purinergic Receptor P2Y, G-Protein Coupled, 13	AF321815	1.55	3.3E-01	3.02	1.5E-02	1.94	8.2E-01
213880_at	*LGR5*	Leucine-Rich Repeat-Containing G Protein-Coupled Receptor 5	AA810452	1.22	5.2E-01	2.8	2.4E-02	2.29	8.2E-01
215894_at	*PTGDR*	Prostaglandin D2 Receptor (Dp)	AL524520	1.7	4.4E-01	1.13	4.8E-01	0.66	8.4E-01
234165_at	*PTGDR*	Prostaglandin D2 Receptor (Dp)	AI460323	0.99	5.9E-01	1	5.3E-01	1.02	8.9E-01
204563_at	*SELL*	Selectin L (Lymphocyte Adhesion Molecule 1)	AK026202	1.02	5.3E-01	1.05	3.4E-01	1.03	8.7E-01
209469_at	*GPM6A*	Glycoprotein M6A	NM_000655	1.11	6.0E-01	4.06	8.1E-02	3.66	8.2E-01
209470_s_at	*GPM6A*	Glycoprotein M6A	BF939489	1.43	1.4E-01	1.7	1.7E-05	1.18	8.2E-01
236024_at	*GPM6A*	Glycoprotein M6A	D49958	1.35	1.7E-01	1.76	4.1E-05	1.3	8.2E-01
219761_at	*CLEC1A*	C-Type Lectin Domain Family 1, Member A	AW136286	2.21	1.4E-01	2.72	3.9E-03	1.23	8.3E-01

Functional group 10						P-value = 6.51E-10			
Affymetrix probeset	Gene symbol	Gene Description	Accesion number	Ag/Gb ratio	q-value	Lgg/Ag ratio	q-value	Lgg/Gb ratio	q-value
219890_at	*CLEC5A*	C-Type Lectin Domain Family 5, Member A	NM_016511	0.82	1.3E-01	0.88	1.8E-01	1.07	8.5E-01
1556209_at	*CLEC2B*	C-Type Lectin Domain Family 2, Member B	NM_013252	0.3	6.3E-02	0.09	4.5E-04	0.29	8.2E-01
201438_at	*COL6A3*	Collagen, Type VI, Alpha 3	CA447397	1.11	6.1E-01	0.48	4.5E-02	0.43	8.2E-01
201852_x_at	*COL3A1*	Collagen, Type III, Alpha 1 (Ehlers-Danlos Syndrome Type IV, Autosomal Dominant)	NM_004369	0.21	1.4E-01	0.06	1.7E-02	0.27	8.2E-01
215076_s_at	*COL3A1*	Collagen, Type III, Alpha 1 (Ehlers-Danlos Syndrome Type IV, Autosomal Dominant)	AI813758	0.22	1.1E-01	0.06	9.2E-03	0.27	8.2E-01
211161_s_at	*COL3A1*	Collagen, Type III, Alpha 1 (Ehlers-Danlos Syndrome Type IV, Autosomal Dominant)	AU144167	0.27	9.2E-02	0.07	2.8E-03	0.27	8.2E-01
1556499_s_at	*COL1A1*	Collagen, Type I, Alpha 1	AF130082	0.19	8.3E-02	0.06	5.3E-03	0.31	8.2E-01
202310_s_at	*COL1A1*	Collagen, Type I, Alpha 1	BE221212	0.32	7.0E-02	0.1	5.1E-04	0.32	8.2E-01
202311_s_at	*COL1A1*	Collagen, Type I, Alpha 1	K01228	0.23	1.2E-01	0.06	1.0E-02	0.26	8.2E-01
211980_at	*COL4A1*	Collagen, Type IV, Alpha 1	AI743621	0.18	1.3E-01	0.09	2.6E-02	0.5	8.2E-01
204320_at	*COL11A1*	Collagen, Type XI, Alpha 1	AI922605	0.35	4.4E-02	0.19	2.7E-05	0.54	8.2E-01
229271_x_at	*COL11A1*	Collagen, Type XI, Alpha 1	NM_001854	0.67	3.6E-01	0.58	1.0E-01	0.86	8.7E-01
37892_at	*COL11A1*	Collagen, Type XI, Alpha 1	BG028597	0.5	2.8E-01	0.64	2.8E-01	1.29	8.4E-01
202403_s_at	*COL1A2*	Collagen, Type I, Alpha 2	J04177	0.57	2.9E-01	0.49	8.7E-02	0.86	8.8E-01
202404_s_at	*COL1A2*	Collagen, Type I, Alpha 2	AA788711	0.22	9.2E-02	0.1	7.1E-03	0.46	8.2E-01
229218_at	*COL1A2*	Collagen, Type I, Alpha 2	NM_000089	0.21	9.1E-02	0.07	5.9E-03	0.33	8.2E-01
209156_s_at	*COL6A2*	Collagen, Type VI, Alpha 2	AA628535	0.15	1.6E-01	0.09	5.2E-02	0.63	8.2E-01
225664_at	*COL12A1*	Collagen, Type XII, Alpha 1	AY029208	0.38	1.6E-01	0.12	6.5E-03	0.31	8.2E-01
212865_s_at	*COL14A1*	Collagen, Type XIV, Alpha 1 (Undulin)	AA788946	0.26	2.3E-01	0.17	9.7E-02	0.64	8.2E-01
Functional group 11						**P-value = 1.67E-09**			
1552848_a_at	*PTCHD1*	Patched Domain Containing 1	BF449063	0.25	1.6E-01	0.27	7.4E-02	1.11	8.5E-01
242002_at	*TCBA1*	T-Cell Lymphoma Breakpoint Associated Target 1	NM_173495	0.72	2.8E-01	0.7	1.2E-01	0.97	8.9E-01
1552782_at	*SLC44A5*	Solute Carrier Family 44, Member 5	N62814	2.04	2.6E-01	2.33	6.1E-02	1.14	8.8E-01
235763_at	*SLC44A5*	Solute Carrier Family 44, Member 5	AK091400	0.97	6.2E-01	1.18	4.1E-01	1.21	8.5E-01
1552327_at	*ARMCX4*	Armadillo Repeat Containing, X-Linked 4	AA001450	0.9	5.7E-01	1.26	3.6E-01	1.4	8.2E-01
236044_at	*PPAPDC1A*	Phosphatidic Acid Phosphatase Type 2 Domain Containing 1a	NM_152583	1.79	1.9E-01	2.21	3.1E-03	1.23	8.2E-01
1557176_a_at	*C14orf37*	Chromosome 14 Open Reading Frame 37	BF130943	1.15	4.9E-01	1.97	3.2E-02	1.71	8.2E-01
219895_at	*FAM70A*	Family With Sequence Similarity 70, Member A	BU074567	0.87	4.4E-01	1.21	2.5E-01	1.38	8.2E-01

Functional group 12						P-value = 2.14E-09		
			Accesion	Ag/Gb		Lgg/Ag		Lgg/Gb	
Affymetrix probeset	Gene symbol	Gene Description	number	ratio	q-value	ratio	q-value	ratio	q-value
217022_s_at	*IGHA1*	Immunoglobulin Heavy Constant Alpha 1	NM_017938	0.71	2.9E-01	0.67	1.1E-01	0.94	8.9E-01
214677_x_at	*IGLJ3*	Immunoglobulin Lambda Variable 3-25	S55735	0.18	2.0E-01	0.26	1.4E-01	1.46	8.7E-01
215379_x_at	*IGHG2*	Immunoglobulin Heavy Constant Gamma 2 (G2M Marker)	X57812	0.24	2.0E-01	0.56	3.1E-01	2.39	8.2E-01
215121_x_at	*IGLA2*	Immunoglobulin Heavy Constant Alpha 2 (A2M Marker)	AV698647	0.27	1.9E-01	0.66	3.6E-01	2.43	8.2E-01
211430_s_at	*IGHG3*	Immunoglobulin Kappa Variable 1-5	AA680302	0.32	2.1E-01	0.64	3.4E-01	1.97	8.2E-01
217022_s_at	*IGHA1*	Immunoglobulin Heavy Constant Gamma 1 (G1m Marker)	M87789	0.27	2.7E-01	0.54	3.2E-01	1.95	8.5E-01
221651_x_at	*IGKC*	Immunoglobulin Kappa Constant	S55735	0.18	2.0E-01	0.26	1.4E-01	1.46	8.7E-01
221671_x_at	*IGKC*	Immunoglobulin Kappa Constant	BC005332	0.41	2.5E-01	0.58	2.8E-01	1.39	8.7E-01
224795_x_at	*IGKC*	Immunoglobulin Kappa Constant	M63438	0.4	2.3E-01	0.6	2.9E-01	1.48	8.7E-01
211430_s_at	*IGHG3*	Immunoglobulin Heavy Constant Gamma 3 (G3M Marker)	AW575927	0.41	2.4E-01	0.59	2.9E-01	1.46	8.7E-01
214669_x_at	*IGKC*	Immunoglobulin Kappa Constant	M87789	0.27	2.7E-01	0.54	3.2E-01	1.95	8.5E-01
224795_x_at	*IGKC*	Immunoglobulin Kappa Constant	BG485135	0.42	2.4E-01	0.6	2.8E-01	1.43	8.5E-01
221651_x_at	*IGKC*	Immunoglobulin Kappa Constant	AW575927	0.41	2.4E-01	0.59	2.9E-01	1.46	8.7E-01
221671_x_at	*IGKC*	Immunoglobulin Kappa Constant	BC005332	0.41	2.5E-01	0.58	2.8E-01	1.39	8.7E-01
211430_s_at	*IGHG3*	Immunoglobulin Heavy Constant Gamma 3 (G3M Marker)	M63438	0.4	2.3E-01	0.6	2.9E-01	1.48	8.7E-01
Functional group 13						**P-value = 2.27E-09**			
205384_at	*FXYD1*	FXYD Domain Containing Ion Transport Regulator 1 (Phospholemman)	M87789	0.27	2.7E-01	0.54	3.2E-01	1.95	8.5E-01
1560265_at	*GRIK2*	Glutamate Receptor, Ionotropic, Kainate 2	NM_005031	1.39	4.3E-01	1.25	3.4E-01	0.9	8.8E-01
213845_at	*GRIK2*	Glutamate Receptor, Ionotropic, Kainate 2	BQ434382	2.58	1.3E-01	3.14	2.2E-02	1.22	8.5E-01
231384_at	*GRIN2A*	Glutamate Receptor, Ionotropic, N-Methyl D-Aspartate 2a	AL355532	1.82	1.9E-01	2.5	7.6E-03	1.37	8.2E-01
242286_at	*GRIN2A*	Glutamate Receptor, Ionotropic, N-Methyl D-Aspartate 2a	T65537	2.22	3.0E-01	2.16	1.8E-01	0.97	9.0E-01
205279_s_at	*GLRB*	Glycine Receptor, Beta	N48896	2.44	2.3E-01	2.71	1.3E-01	1.11	8.9E-01
205280_at	*GLRB*	Glycine Receptor, Beta	AF094754	1.8	1.7E-01	2.08	2.5E-02	1.16	8.5E-01
205358_at	*GRIA2*	Glutamate Receptor, Ionotropic, AMPA 2	NM_000824	1.77	2.2E-01	1.97	4.6E-02	1.11	8.7E-01
236538_at	*GRIA2*	Glutamate Receptor, Ionotropic, AMPA 2	NM_000826	2.15	9.4E-02	2.75	6.5E-05	1.28	8.2E-01
215634_at	*GRIA1*	Glutamate Receptor, Ionotropic, AMPA 1	BE219628	2.04	1.5E-01	3.03	1.5E-04	1.49	8.2E-01
209793_at	*GRIA1*	Glutamate Receptor, Ionotropic, AMPA 1	AF007137	1.41	5.0E-01	1.37	3.7E-01	0.97	9.0E-01
238663_x_at	*GRIA4*	Glutamate Receptor, Ionotrophic, AMPA 4	AL567302	0.9	5.6E-01	1.3	2.3E-01	1.45	8.2E-01
1555268_a_at	*GRID1*	Glutamate Receptor, Ionotropic, Delta 1	H20055	2.56	1.5E-01	2.94	1.9E-04	1.15	8.6E-01
1554748_at	*CLCNKB*	Chloride Channel KB	BC039263	1.4	1.5E-01	1.57	1.9E-03	1.12	8.3E-01

Affymetrix probeset	Gene symbol	Gene Description	Accesion number	Ag/Gb ratio	q-value	Lgg/Ag ratio	q-value	Lgg/Gb ratio	q-value
206456_at	*GABRA5*	Gamma-Aminobutyric Acid (Gaba) A Receptor, Alpha 5	BC020873	0.87	3.6E-01	0.8	3.6E-02	0.92	8.5E-01
1552943_at	*GABRG1*	Gamma-Aminobutyric Acid (Gaba) A Receptor, Gamma 1	NM_000810	2.32	3.1E-01	2.77	1.5E-01	1.2	8.8E-01
241805_at	*GABRG1*	Gamma-Aminobutyric Acid (Gaba) A Receptor, Gamma 1	NM_173536	1.62	1.9E-01	1.72	7.8E-02	1.07	8.9E-01
233437_at	*GABRA4*	Gamma-Aminobutyric Acid (Gaba) A Receptor, Alpha 4	N92667	2.7	2.0E-01	2.89	3.7E-02	1.07	8.9E-01
206849_at	*GABRG2*	Gamma-Aminobutyric Acid (Gaba) A Receptor, Gamma 2	AF238869	3.33	1.9E-01	2.83	8.2E-02	0.85	8.8E-01
1568612_at	*GABRG2*	Gamma-Aminobutyric Acid (Gaba) A Receptor, Gamma 2	NM_000816	2.19	2.0E-01	2.94	5.2E-02	1.35	8.3E-01
1554308_s_at	*GABRA2*	Gamma-Aminobutyric Acid (Gaba) A Receptor, Alpha 2	BC036030	3.99	1.7E-01	2.98	4.8E-02	0.75	8.5E-01
207014_at	*GABRA2*	Gamma-Aminobutyric Acid (Gaba) A Receptor, Alpha 2	BC022488	1.56	3.2E-01	1.36	2.4E-01	0.87	8.7E-01
206678_at	*GABRA1*	Gamma-Aminobutyric Acid (Gaba) A Receptor, Beta 1	NM_000807	2.5	3.3E-01	2.21	1.4E-01	0.88	8.9E-01
244118_at	*GABRA1*	Gamma-Aminobutyric Acid (Gaba) A Receptor, Beta 1	NM_000806	2.41	2.4E-01	2.7	1.1E-01	1.12	8.9E-01
207010_at	*GABRB1*	Gamma-Aminobutyric Acid (Gaba) A Receptor, Beta 1	AV722228	2.15	3.0E-01	2.2	1.9E-01	1.02	9.0E-01
227690_at	*GABRB3*	Gamma-Aminobutyric Acid (Gaba) A Receptor, Beta 3	NM_000812	2.36	1.7E-01	1.86	6.5E-02	0.79	8.4E-01
229724_at	*GABRB3*	Gamma-Aminobutyric Acid (Gaba) A Receptor, Beta 3	BE502537	4.17	8.2E-02	3.82	2.0E-03	0.92	8.8E-01
227830_at	*GABRB3*	Gamma-Aminobutyric Acid (Gaba) A Receptor, Beta 3	AI693153	4.77	8.2E-02	5.32	9.5E-04	1.12	8.7E-01
221107_at	*CHRNA9*	Cholinergic Receptor, Nicotinic, Alpha 9	AI478781	4.24	6.7E-02	4.11	1.5E-03	0.97	8.9E-01
Functional group 14						**P-value =**	**5.57E-08**		
1552327_at	*ARMCX4*	Armadillo Repeat Containing, X-Linked 4	NM_017581	0.34	1.9E-01	0.19	2.4E-02	0.55	8.4E-01
237802_at	*XKR4*	Xk, Kell Blood Group Complex Subunit-Related Family, Member 4	NM_152583	1.79	1.9E-01	2.21	3.1E-03	1.23	8.2E-01
223343_at	*MS4A7*	Membrane-Spanning 4-Domains, Subfamily A, Member 7	R54212	2.66	1.3E-01	2.54	1.6E-03	0.95	8.9E-01
1558102_at	*TM6SF1*	Transmembrane 6 Superfamily Member 1	AI301935	0.58	2.1E-01	0.4	1.2E-02	0.68	8.2E-01
207723_s_at	*KLRC3*	Killer Cell Lectin-Like Receptor Subfamily C, Member 3	AK055438	1.56	1.5E-01	1.58	1.0E-02	1.01	9.0E-01
213609_s_at	*SEZ6L*	Seizure Related Gene 6 (Mouse)-Like	NM_002261	7.99	2.0E-01	7.59	1.0E-02	0.95	9.0E-01
240709_at	*SEZ6L*	Seizure Related Gene 6 (Mouse)-Like	AB023144	3.27	1.6E-01	5.1	7.2E-04	1.56	8.2E-01
231650_s_at	*SEZ6L*	Seizure Related Gene 6 (Mouse)-Like	AW204757	3.19	2.1E-01	6.1	1.8E-03	1.91	8.2E-01
242301_at	*CBLN2*	Cerebellin 2 Precursor	BE672217	4.4	1.6E-01	5.39	4.5E-04	1.23	8.5E-01
1554689_a_at	*NLGN4X*	Neuroligin 4, X-Linked	R60224	3.04	2.5E-01	2.37	1.4E-01	0.78	8.7E-01
221933_at	*NLGN4X*	Neuroligin 4, X-Linked	BC034018	0.85	4.5E-01	1.01	5.3E-01	1.18	8.3E-01

Affymetrix probeset	Gene symbol	Gene Description	Accesion number	Ag/Gb ratio	Ag/Gb q-value	Lgg/Ag ratio	Lgg/Ag q-value	Lgg/Gb ratio	Lgg/Gb q-value
1555728_a_at	*MS4A4A*	Membrane-Spanning 4-Domains, Subfamily A, Member 4	AI338338	0.74	2.6E-01	0.93	4.4E-01	1.25	8.2E-01
219607_s_at	*MS4A4A*	Membrane-Spanning 4-Domains, Subfamily A, Member 4	AF354928	0.57	3.0E-01	0.28	7.2E-03	0.49	8.2E-01
1554452_a_at	*HIG2*	Hypoxia-Inducible Protein 2	NM_024021	0.56	2.6E-01	0.28	4.3E-03	0.5	8.2E-01
218507_at	*HIG2*	Hypoxia-Inducible Protein 2	BC001863	0.61	3.1E-01	0.17	2.9E-03	0.28	8.2E-01
219410_at	*TMEM45A*	Transmembrane Protein 45a	NM_013332	0.69	4.3E-01	0.18	1.3E-03	0.26	8.2E-01
219890_at	*CLEC5A*	C-Type Lectin Domain Family 5, Member A	NM_018004	0.49	9.1E-02	0.38	7.3E-03	0.76	8.5E-01
227307_at	*TSPAN18*	Tetraspanin 18	NM_013252	0.3	6.3E-02	0.09	4.5E-04	0.29	8.2E-01
214841_at	*CNIH3*	Cornichon Homolog 3 (Drosophila)	AL565381	0.67	2.5E-01	0.74	2.0E-01	1.09	8.8E-01
1559063_at	*C21orf63*	Chromosome 21 Open Reading Frame 63	AF070524	0.73	4.4E-01	0.47	1.1E-02	0.64	8.2E-01
219274_at	*TSPAN12*	Tetraspanin 12	AL355689	0.84	3.3E-01	0.69	2.9E-03	0.83	8.2E-01
215241_at	*TMEM16C*	Transmembrane Protein 16C	NM_012338	0.72	3.9E-01	1.15	4.3E-01	1.6	8.2E-01
218888_s_at	*NETO2*	Neuropilin (Nrp) And Tolloid (Tll)-Like 2	AJ300461	2.86	2.5E-01	2.32	1.4E-01	0.81	8.7E-01
222774_s_at	*NETO2*	Neuropilin (Nrp) And Tolloid (Tll)-Like 2	NM_018092	1.25	4.5E-01	1.14	4.5E-01	0.91	8.8E-01
239575_at	*TMEM10*	Transmembrane Protein 10	AI335263	1.1	5.8E-01	0.98	5.2E-01	0.89	8.8E-01
203963_at	*CA12*	Carbonic Anhydrase XII	N63401	1.74	4.3E-01	2.25	2.0E-01	1.29	8.7E-01
209655_s_at	*TMEM47*	Transmembrane Protein 47	NM_001218	0.45	1.6E-01	0.2	1.2E-03	0.45	8.2E-01
1552736_a_at	*NETO1*	Neuropilin (Nrp) And Tolloid (Tll)-Like 1	AI803181	0.98	6.2E-01	1.36	2.1E-01	1.39	8.2E-01
206785_s_at	*KLRC2*	Killer Cell Lectin-Like Receptor Subfamily C, Member 1	NM_138966	2.17	2.0E-01	1.79	9.5E-02	0.83	8.6E-01
219666_at	*MS4A6A*	Membrane-Spanning 4-Domains, Subfamily A, Member 6a (Cd20-Like Precursor)	NM_002260	7.71	1.4E-01	7.02	8.1E-03	0.91	8.9E-01
223280_x_at	*MS4A6A*	Membrane-Spanning 4-Domains, Subfamily A, Member 6a (Cd20-Like Precursor)	NM_022349	0.69	3.5E-01	0.39	9.0E-03	0.57	8.2E-01
230550_at	*MS4A6A*	Membrane-Spanning 4-Domains, Subfamily A, Member 6a (Cd20-Like Precursor)	AF253977	0.55	1.9E-01	0.34	4.6E-03	0.61	8.2E-01
224356_x_at	*MS4A6A*	Membrane-Spanning 4-Domains, Subfamily A, Member 6a (Cd20-Like Precursor)	AA045175	0.51	1.6E-01	0.35	6.4E-03	0.69	8.4E-01
205691_at	*SYNGR3*	Synaptogyrin 3	AF237908	0.58	2.2E-01	0.33	5.2E-03	0.57	8.2E-01
223557_s_at	*TMEFF2*	Transmembrane Protein With EGF-Like And Two Follistatin-Like Domains 2	NM_004209	2.25	2.7E-01	2.12	1.5E-01	0.94	8.9E-01
224321_at	*TMEFF2*	Transmembrane Protein With EGF-Like And Two Follistatin-Like Domains 2	AB017269	2.44	6.4E-02	2.45	3.4E-03	1	9.1E-01
219602_s_at	*FAM38B*	Family With Sequence Similarity 38, Member B	AB004064	0.87	5.8E-01	0.41	3.7E-02	0.47	8.2E-01
1552782_at	*SLC44A5*	Solute Carrier Family 44, Member 5	NM_022068	1.01	6.4E-01	0.82	2.8E-01	0.81	8.5E-01
235763_at	*SLC44A5*	Solute Carrier Family 44, Member 5	AK091400	0.97	6.2E-01	1.18	4.1E-01	1.21	8.5E-01

Functional group 15						P-value =	8.99E-08		
Affymetrix probeset	Gene symbol	Gene Description	Accesion number	Ag/Gb ratio	q-value	Lgg/Ag ratio	q-value	Lgg/Gb ratio	q-value
1554485_s_at	*TMEM37*	Transmembrane Protein 37	AA001450	0.9	5.7E-01	1.26	3.6E-01	1.4	8.2E-01
213395_at	*MLC1*	Megalencephalic Leukoencephalopathy With Subcortical Cysts 1	BI825302	1.32	5.4E-01	0.52	4.4E-03	0.4	8.2E-01
219415_at	*TTYH1*	Tweety Homolog 1 (Drosophila)	AL022327	0.51	5.5E-02	0.89	3.7E-01	1.75	8.2E-01
204519_s_at	*PLLP*	Plasma Membrane Proteolipid (Plasmolipin)	NM_020659	0.96	6.1E-01	1.49	3.7E-02	1.55	8.2E-01
220131_at	*FXYD7*	Fxyd Domain Containing Ion Transport Regulator 7	NM_015993	1.1	5.8E-01	1.36	2.4E-01	1.24	8.5E-01
205384_at	*FXYD1*	Fxyd Domain Containing Ion Transport Regulator 1 (Phospholemman)	NM_022006	2.33	2.2E-01	2.61	1.1E-01	1.12	8.9E-01
228608_at	*NALCN*	Sodium Leak Channel, Non-Selective	NM_005031	1.39	4.3E-01	1.25	3.4E-01	0.9	8.8E-01
1555492_a_at	*BEST3*	Bestrophin 3	N49852	1.91	1.4E-01	2.69	2.8E-04	1.41	8.2E-01
237003_at	*BEST3*	Bestrophin 3	AF440758	2.38	2.1E-01	2.07	1.1E-01	0.87	8.8E-01
224520_s_at	*BEST3*	Bestrophin 3	AA878383	1.67	3.1E-01	1.77	1.4E-01	1.06	8.9E-01
Functional group 16						**P-value =**	**2.45E-07**		
230303_at	*SYNPR*	Synaptoporin	BC006440	3.52	2.1E-01	3.73	2.7E-02	1.06	8.9E-01
229818_at	*SVOP*	Sv2 Related Protein Homolog (Rat)	H11380	2.46	3.0E-01	2.21	2.0E-01	0.9	8.9E-01
203069_at	*SV2A*	Synaptic Vesicle Glycoprotein 2a	AL359592	3.41	1.2E-01	3.71	4.0E-02	1.09	8.9E-01
216086_at	*SV2C*	Synaptic Vesicle Glycoprotein 2c	NM_014849	1.28	3.3E-01	1.39	1.3E-01	1.09	8.7E-01
205551_at	*SV2B*	Synaptic Vesicle Glycoprotein 2b	AB028977	1.99	3.9E-01	1.13	4.8E-01	0.57	8.2E-01
232426_at	*SV2B*	Synaptic Vesicle Glycoprotein 2b	NM_014848	2.46	2.6E-01	2.69	1.5E-01	1.09	8.9E-01
Functional group 17						**P-value =**	**5.17E-07**		
207850_at	*CXCL3*	Chemokine (C-X-C Motif) Ligand 3	AL109781	2.18	2.7E-01	2.17	1.7E-01	1	9.1E-01
209774_x_at	*CXCL2*	Chemokine (C-X-C Motif) Ligand 2	NM_002090	0.34	2.1E-01	0.34	1.1E-01	0.99	9.1E-01
206336_at	*CXCL6*	Chemokine (C-X-C Motif) Ligand 6 (Granulocyte Chemotactic Protein 2)	M57731	0.76	5.3E-01	1.05	5.2E-01	1.39	8.7E-01
218002_s_at	*CXCL14*	Chemokine (C-X-C Motif) Ligand 14	NM_002993	0.5	3.9E-01	0.17	6.7E-02	0.34	8.2E-01
222484_s_at	*CXCL14*	Chemokine (C-X-C Motif) Ligand 14	NM_004887	1.1	6.1E-01	0.53	9.0E-02	0.48	8.2E-01
214974_x_at	*CXCL5*	Chemokine (C-X-C Motif) Ligand 5	AF144103	1.02	6.3E-01	0.65	2.0E-01	0.64	8.3E-01
205242_at	*CXCL13*	Chemokine (C-X-C Motif) Ligand 13 (B-Cell Chemoattractant)	AK026546	2.02	4.8E-01	0.24	1.0E-01	0.12	8.2E-01
205476_at	*CCL20*	Chemokine (C-C Motif) Ligand 20	NM_006419	0.22	2.1E-01	0.26	1.3E-01	1.2	8.7E-01
Functional group 18						**P-value =**	**2.40E-06**		
207261_at	*CNGA3*	Cyclic Nucleotide Gated Channel Alpha 3	NM_004591	1.01	6.4E-01	0.23	9.5E-02	0.23	8.2E-01
235781_at	*CACNA1B*	Calcium Channel, Voltage-Dependent, L Type, Alpha 1b Subunit	NM_001298	0.47	1.5E-01	0.45	3.5E-02	0.97	9.0E-01
234103_at	*KCNT2*	Potassium Channel, Subfamily T, Member 2	AA448208	2.98	2.1E-01	2.39	1.1E-01	0.8	8.7E-01
228581_at	*KCNJ10*	Potassium Inwardly-Rectifying Channel, Subfamily J, Member 10	AU145191	1.48	3.3E-01	2.29	4.0E-02	1.55	8.2E-01
207103_at	*KCND2*	Potassium Voltage-Gated Channel, Shal-Related Subfamily, Member 2	AW071744	1.54	1.9E-01	1.7	3.4E-03	1.11	8.5E-01

Affymetrix probeset	Gene symbol	Gene Description	Accesion number	Ag/Gb ratio	q-value	Lgg/Ag ratio	q-value	Lgg/Gb ratio	q-value
1552742_at	*KCNH8*	Potassium Voltage-Gated Channel, Subfamily H (Eag-Related), Member 8	NM_012281	1.14	5.5E-01	2.72	2.9E-03	2.37	6.4E-01
227623_at	*CACNA2D1*	Calcium Channel, Voltage-Dependent, Alpha 2/Delta Subunit 1	NM_144633	2.03	1.1E-01	2.41	7.3E-03	1.19	8.4E-01
221107_at	*CHRNA9*	Cholinergic Receptor, Nicotinic, Alpha 9	H16409	1.83	1.6E-01	1.92	3.4E-02	1.05	8.9E-01
221584_s_at	*KCNMA1*	Potassium Large Conductance Calcium-Activated Channel, Subfamily M, Alpha Member 1	NM_017581	0.34	1.9E-01	0.19	2.4E-02	0.55	8.4E-01
1559419_at	*CACNB2*	Calcium Channel, Voltage-Dependent, Beta 2 Subunit	U11058	1.12	4.7E-01	1.05	4.3E-01	0.94	8.7E-01
205902_at	*KCNN3*	Potassium Intermediate/Small Conductance Calcium-Activated Channel, Subfamily N, Member 3	AL162054	1.01	6.4E-01	0.99	5.3E-01	0.98	9.0E-01
1555230_a_at	*KCNIP2*	Kv Channel Interacting Protein 2	AJ251016	0.81	4.7E-01	1.56	1.1E-01	1.92	8.2E-01
220116_at	*KCNN2*	Potassium Intermediate/Small Conductance Calcium-Activated Channel, Subfamily N, Member 2	AF367019	1.93	1.2E-01	4.25	3.1E-02	2.21	8.2E-01
220294_at	*KCNV1*	Potassium Channel, Subfamily V, Member 1	NM_021614	2.51	1.6E-01	2.8	4.0E-03	1.12	8.7E-01
210432_s_at	*SCN3A*	Sodium Channel, Voltage-Gated, Type Iii, Alpha	NM_014379	2.29	2.5E-01	2.56	1.5E-01	1.12	8.9E-01
210454_s_at	*KCNJ6*	Potassium Inwardly-Rectifying Channel, Subfamily J, Member 6	AF225986	3.23	1.3E-01	3.59	2.5E-03	1.11	8.8E-01
205737_at	*KCNQ2*	Potassium Voltage-Gated Channel, Kqt-Like Subfamily, Member 2	U24660	1.97	2.3E-01	1.55	2.1E-01	0.79	8.5E-01
210508_s_at	*KCNK2*	Potassium Channel, Subfamily K, Member 1	NM_004518	2.78	1.5E-01	2.77	3.4E-03	0.99	9.1E-01
204679_at	*KCNK1*	Potassium Voltage-Gated Channel, Shaker-Related Subfamily, Member 1 (Episodic Ataxia With Myokymia)	D82346	0.8	4.1E-01	1.34	2.4E-01	1.67	8.2E-01
230849_at	*KCNJ3*	Potassium Inwardly-Rectifying Channel, Subfamily J, Member 3	NM_002245	2.05	2.2E-01	1.81	1.8E-01	0.88	8.8E-01
233059_at	*SCN6A*	Sodium Channel, Voltage-Gated, Type Vii, Alpha	N64750	1.9	3.2E-01	1.72	1.9E-01	0.91	8.9E-01
228504_at	*SCN6A*	Sodium Channel, Voltage-Gated, Type Vii, Alpha	AK026384	2.28	2.4E-01	1.87	2.0E-01	0.82	8.7E-01
1555246_a_at	*SCN1A*	Sodium Channel, Voltage-Gated, Type I, Alpha	AI828648	0.19	2.6E-01	0.32	2.3E-01	1.67	8.5E-01
210383_at	*SCN1A*	Sodium Channel, Voltage-Gated, Type I, Alpha	AB093548	1.15	3.8E-01	1.43	3.3E-02	1.25	8.2E-01
206381_at	*SCN2A*	Sodium Channel, Voltage-Gated, Type Ii, Alpha 2	AF225985	1.6	2.3E-01	1.64	4.7E-02	1.02	9.0E-01
229057_at	*SCN2A*	Sodium Channel, Voltage-Gated, Type Ii, Alpha 2	NM_021007	2.03	2.1E-01	3.06	2.6E-02	1.51	8.2E-01

Functional group 19						P-value = 2.86E-06			
Affymetrix probeset	Gene symbol	Gene Description	Accesion number	Ag/Gb ratio	q-value	Lgg/Ag ratio	q-value	Lgg/Gb ratio	q-value
204722_at	*SCN3B*	Sodium Channel, Voltage-Gated, Type Iii, Beta	BF432956	2.83	1.8E-01	2.74	5.8E-02	0.97	9.0E-01
204723_at	*SCN3B*	Sodium Channel, Voltage-Gated, Type Iii, Beta	AW007335	2.91	1.4E-01	3.63	1.7E-02	1.25	8.5E-01
219564_at	*KCNJ16*	Potassium Inwardly-Rectifying Channel, Subfamily J, Member 16	AB032984	2.94	1.6E-01	4	2.0E-02	1.36	8.3E-01
1552507_at	*KCNE4*	Potassium Voltage-Gated Channel, Isk-Related Family, Member 4	NM_018658	0.89	5.8E-01	1.28	3.7E-01	1.44	8.2E-01
1552508_at	*KCNE4*	Potassium Voltage-Gated Channel, Isk-Related Family, Member 4	NM_080671	0.49	7.4E-02	0.45	4.7E-03	0.92	8.7E-01
219714_s_at	*CACNA2D3*	Calcium Channel, Voltage-Dependent, Alpha 2/Delta 3 Subunit	NM_080671	0.52	1.1E-01	0.46	6.3E-03	0.88	8.7E-01
1554485_s_at	*TMEM37*	Transmembrane Protein 37	NM_018398	2.09	2.4E-01	1.88	9.9E-02	0.9	8.8E-01
206384_at	*CACNG3*	Calcium Channel, Voltage-Dependent, Gamma Subunit 3	BI825302	1.32	5.4E-01	0.52	4.4E-03	0.4	8.2E-01
237411_at	*ADAMTS6*	Adam Metallopeptidase With Thrombospondin Type 1 Motif, 6	NM_006539	2.85	2.4E-01	2.51	1.5E-01	0.88	8.9E-01
1556989_at	*ADAMTS9*	Adam Metallopeptidase With Thrombospondin Type 1 Motif, 9	N71063	1.42	4.4E-01	1.51	8.9E-02	1.06	8.9E-01
1554697_at	*ADAMTS9*	Adam Metallopeptidase With Thrombospondin Type 1 Motif, 9	AF086069	0.44	6.2E-02	0.58	3.4E-02	1.32	8.2E-01
1555326_a_at	*ADAM9*	Adam Metallopeptidase Domain 9 (Meltrin Gamma)	AF488803	0.4	1.5E-01	0.51	1.1E-01	1.28	8.2E-01
213790_at	*ADAM12*	Adam Metallopeptidase Domain 12 (Meltrin Alpha)	AF495383	0.46	7.0E-02	0.55	7.6E-02	1.19	8.7E-01
226777_at	*ADAM12*	Adam Metallopeptidase Domain 12 (Meltrin Alpha)	W46291	0.38	6.3E-02	0.25	2.7E-04	0.64	8.2E-01
206134_at	*ADAMDEC1*	Adam-Like, Decysin 1	AA147933	0.25	4.4E-02	0.15	6.0E-04	0.58	8.2E-01
1552727_s_at	*ADAMTS17*	Adam Metallopeptidase With Thrombospondin Type 1 Motif, 17	NM_014479	0.59	3.7E-01	0.2	5.7E-02	0.33	8.2E-01
Functional group 20						**P-value = 1.15E-05**			
1552507_at	*KCNE4*	Potassium Voltage-Gated Channel, Isk-Related Family, Member 4	AA022668	1.01	6.4E-01	0.85	3.2E-01	0.84	8.3E-01
1552508_at	*KCNE4*	Potassium Voltage-Gated Channel, Isk-Related Family, Member 4	NM_080671	0.49	7.4E-02	0.45	4.7E-03	0.92	8.7E-01
1554027_a_at	*SLC4A4*	Solute Carrier Family 4, Sodium Bicarbonate Cotransporter, Member 4	NM_080671	0.52	1.1E-01	0.46	6.3E-03	0.88	8.7E-01
203908_at	*SLC4A4*	Solute Carrier Family 4, Sodium Bicarbonate Cotransporter, Member 4	BC030977	0.53	2.7E-01	0.88	4.5E-01	1.66	8.2E-01
210040_at	*SLC12A5*	Solute Carrier Family 12, (Potassium-Chloride Transporter) Member 5	NM_003759	0.87	5.1E-01	1.39	1.2E-01	1.59	8.2E-01

Affymetrix probeset	Gene symbol	Gene Description	Accesion number	Ag/Gb ratio	Ag/Gb q-value	Lgg/Ag ratio	Lgg/Ag q-value	Lgg/Gb ratio	Lgg/Gb q-value
219090_at	*SLC24A3*	Solute Carrier Family 24 (Sodium/Potassium/Calcium Exchanger), Member 3	AF208159	2.66	2.6E-01	2.55	1.5E-01	0.96	9.0E-01
57588_at	*SLC24A3*	Solute Carrier Family 24 (Sodium/Potassium/Calcium Exchanger), Member 3	NM_020689	1.96	1.8E-01	3.04	1.0E-02	1.55	8.2E-01
239913_at	*SLC10A4*	Solute Carrier Family 10 (Sodium/Bile Acid Cotransporter Family), Member 4	R62432	1.7	2.3E-01	2.56	5.5E-03	1.51	8.2E-01
Functional group 21						**P-value = 1.17E-05**			
1554748_at	*CLCNKB*	Chloride Channel Kb	AI421796	0.37	3.1E-01	0.29	1.6E-01	0.78	8.4E-01
219529_at	*CLIC3*	Chloride Intracellular Channel 3	BC020873	0.87	3.6E-01	0.8	3.6E-02	0.92	8.5E-01
227742_at	*CLIC6*	Chloride Intracellular Channel 6	NM_004669	0.98	5.4E-01	1.06	3.6E-01	1.09	8.2E-01
213317_at	*CLIC5*	Chloride Intracellular Channel 5	AI638295	1.1	5.9E-01	0.83	4.2E-01	0.75	8.5E-01
205384_at	*FXYD1*	Fxyd Domain Containing Ion Transport Regulator 1 (Phospholemman)	AL049313	1.1	5.8E-01	1.05	5.0E-01	0.95	8.9E-01
Functional group 22						**P-value = 3.24E-04**			
239144_at	*B3GAT2*	Beta-1,3-Glucuronyltransferase 2 (Glucuronosyltransferase S)	NM_005031	1.39	4.3E-01	1.25	3.4E-01	0.9	8.8E-01
220979_s_at	*ST6GALNAC5*	St6 (Alpha-N-Acetyl-Neuraminyl-2,3-Beta-Galactosyl-1,3)-N-Acetylgalactosaminide Alpha-2,6-Sialyltransferase 5	AA835648	1.76	2.9E-01	2.54	1.2E-02	1.45	8.2E-01
230262_at	*ST8SIA3*	St8 Alpha-N-Acetyl-Neuraminide Alpha-2,8-Sialyltransferase 3	NM_030965	1.52	4.4E-01	1.57	3.1E-01	1.03	9.0E-01
1555123_at	*ST6GAL2*	St6 Beta-Galactosamide Alpha-2,6-Sialyltranferase 2	BF510762	2.68	1.7E-01	2.63	2.6E-02	0.98	9.0E-01
228821_at	*ST6GAL2*	St6 Beta-Galactosamide Alpha-2,6-Sialyltranferase 2	BC008680	1.02	6.3E-01	1.73	1.4E-01	1.7	8.2E-01
1559617_at	*GAL3ST2*	Galactose-3-O-Sulfotransferase 2	AW004016	1.15	5.7E-01	2.33	4.9E-02	2.03	8.2E-01
234472_at	*GALNT13*	UDP-N-Acetyl-Alpha-D-Galactosamine:Polypeptide N-Acetylgalactosaminyltransferase 13 (Galnac-T13)	D55640	1.3	5.3E-01	2.91	1.8E-01	2.24	8.2E-01
228501_at	*GALNTL2*	UDP-N-Acetyl-Alpha-D-Galactosamine:Polypeptide N-Acetylgalactosaminyltransferase-Like 2	AC009227	2.64	1.5E-01	3.22	7.3E-03	1.22	8.5E-01
236361_at	*GALNTL2*	UDP-N-Acetyl-Alpha-D-Galactosamine:Polypeptide N-Acetylgalactosaminyltransferase-Like 2	BF055343	0.61	2.4E-01	0.72	2.0E-01	1.19	8.5E-01
1553959_a_at	*B3GALT6*	UDP-Gal:Betagal Beta 1,3-Galactosyltransferase Polypeptide 6	BF432376	0.55	1.8E-01	0.62	8.8E-02	1.13	8.7E-01
207447_s_at	*MGAT4C*	Mannosyl (Alpha-1,3-)-Glycoprotein Beta-1,4-N-Acetylglucosaminyltransferase, Isozyme C (Putative)	N95564	0.76	1.5E-01	0.7	2.8E-03	0.93	8.6E-01
1559814_at	*CHSY3*	Chondroitin Sulfate Synthase 3	NM_013244	3.75	1.7E-01	6.01	1.0E-02	1.6	8.2E-01

Functional group 23						P-value = 1.35E-02			
Affymetrix probeset	Gene symbol	Gene Description	Accesion number	Ag/Gb ratio	q-value	Lgg/Ag ratio	q-value	Lgg/Gb ratio	q-value
208161_s_at	*ABCC3*	ATP-Binding Cassette, Sub-Family C (Cftr/Mrp), Member 3	AK024712	1.15	5.1E-01	0.89	2.6E-01	0.77	8.2E-01
1553605_a_at	*ABCA13*	ATP-Binding Cassette, Sub-Family A (Abc1), Member 13	NM_020037	0.56	2.9E-01	0.12	4.3E-04	0.21	8.2E-01
204719_at	*ABCA8*	ATP-Binding Cassette, Sub-Family A (Abc1), Member 8	NM_152701	1.75	5.0E-01	0.49	1.3E-01	0.28	8.2E-01
213106_at	*ATP8A1*	ATPase, Aminophospholipid Transporter (Aplt), Class I, Type 8a, Member 1	NM_007168	0.66	2.9E-01	0.73	2.6E-01	1.11	8.8E-01
219659_at	*ATP8A2*	ATPase, Aminophospholipid Transporter-Like, Class I, Type 8a, Member 2	AI769688	2.34	1.1E-01	2.53	1.5E-03	1.08	8.7E-01
1557136_at	*ATP13A4*	ATPase Type 13a4	AU146927	2.34	2.3E-01	1.88	1.4E-01	0.8	8.6E-01
1559571_a_at	*ATP13A4*	ATPase Type 13a4	BG059633	0.68	2.1E-01	1.06	4.7E-01	1.55	8.2E-01
Functional group 24						**P-value = 1.71E-02**			
204337_at	*RGS4*	Regulator of G-Protein Signalling 4	AK095277	0.79	2.1E-01	1.19	2.6E-01	1.51	8.2E-01
204338_s_at	*RGS4*	Regulator of G-Protein Signalling 4	AL514445	1.81	3.7E-01	1.86	2.4E-01	1.03	9.0E-01
204339_s_at	*RGS4*	Regulator of G-Protein Signalling 4	NM_005613	1.33	5.3E-01	2.1	2.3E-01	1.58	8.4E-01
1554500_a_at	*RGS7*	Regulator of G-Protein Signalling 7	BC000737	1.35	4.9E-01	1.61	3.0E-01	1.19	8.8E-01
206290_s_at	*RGS7*	Regulator of G-Protein Signalling 7	AF493931	2.03	2.2E-01	2.49	7.2E-02	1.23	8.6E-01
202988_s_at	*RGS1*	Regulator of G-Protein Signalling 1	NM_002924	2.62	2.1E-01	2.74	4.7E-02	1.05	8.9E-01
216834_at	*RGS1*	Regulator of G-Protein Signalling 1	NM_002922	0.98	6.3E-01	0.27	4.1E-03	0.27	8.2E-01
210138_at	*RGS20*	Regulator of G-Protein Signalling 20	S59049	0.75	4.4E-01	0.34	4.1E-03	0.45	8.2E-01
1554643_at	*RGS11*	Regulator of G-Protein Signalling 11	AF074979	1.72	3.3E-01	1.58	1.7E-01	0.91	8.9E-01
223809_at	*RGS18*	Regulator of G-Protein Signalling 18	BC040504	1.93	2.4E-01	1.5	1.4E-02	0.77	8.4E-01
1555725_a_at	*RGS5*	Regulator of G-Protein Signalling 5	AF076642	0.53	2.6E-01	1.04	5.2E-01	1.94	8.2E-01
Functional group 25						**P-value = 2.28E-02**			
219671_at	*HPCAL4*	Hippocalcin Like 4	AF493929	0.81	3.5E-01	1.07	4.5E-01	1.33	8.2E-01
209904_at	*TNNC1*	Troponin C Type 1 (Slow)	AL136591	2.64	2.6E-01	3.06	6.9E-02	1.16	8.8E-01
203797_at	*VSNL1*	Visinin-Like 1	AF020769	0.92	3.8E-01	0.96	4.1E-01	1.04	8.5E-01
203798_s_at	*VSNL1*	Visinin-Like 1	AF039555	2.29	2.9E-01	2.41	1.6E-01	1.05	8.9E-01
208320_at	*CABP1*	Calcium Binding Protein 1 (Calbrain)	NM_003385	2.57	2.8E-01	2.96	1.7E-01	1.15	8.9E-01
208321_s_at	*CABP1*	Calcium Binding Protein 1 (Calbrain)	NM_004276	2.1	2.9E-01	1.8	2.0E-01	0.86	8.8E-01
210181_s_at	*CABP1*	Calcium Binding Protein 1 (Calbrain)	NM_004276	2.17	2.3E-01	2.22	1.0E-01	1.02	9.0E-01
223885_at	*CALN1*	Calneuron 1	AF169148	1.74	3.5E-01	2.15	1.8E-01	1.24	8.7E-01
230698_at	*CALN1*	Calneuron 1	AF282250	4.12	1.3E-01	3.71	6.6E-02	0.9	8.9E-01

Functional group 1							P-value = 7.12E-18			
				Accesion	Mg/Ag		Mg/Gb		Mg/Lgg	
Affymetrix probeset	Gene symbol	Gene Description	number	ratio	q-value	ratio	q-value	ratio	q-value
213285_at	*TMEM30B*	Transmembrane Protein 30B	AV691491	64.74	1.7E-03	113.56	9.1E-04	129.86	7.7E-04
1559072_a_at	*ELFN2*	Extracellular Leucine-Rich Repeat And Fibronectin Type III Domain Containing 2	BC032083	0.09	1.3E-03	0.21	4.3E-03	0.07	1.9E-06
1558517_s_at	*LRRC8C*	Leucine Rich Repeat Containing 8 Family, Member C	CA773938	3.66	1.6E-03	3.58	9.2E-04	3.39	8.4E-04
1552848_a_at	*PTCHD1*	Patched Domain Containing 1	NM_173495	0.14	2.8E-03	0.1	6.0E-08	0.15	3.5E-04
212976_at	*LRRC8B*	Leucine Rich Repeat Containing 8 Family, Member B	R41498	0.09	3.6E-03	0.17	5.3E-04	0.08	8.3E-04
1557176_a_at	*C14orf37*	Chromosome 14 Open Reading Frame 37	BU074567	0.58	4.3E-03	0.5	6.9E-04	0.42	3.3E-04
1558693_s_at	*C1orf85*	Chromosome 1 Open Reading Frame 85	AW090182	1.61	6.6E-02	1.25	9.7E-02	2.77	4.7E-04
227001_at	*NPAL2*	NIPA-Like Domain Containing 2	AI096706	2.48	1.7E-04	3	2.6E-06	4.64	3.7E-07
243756_at	*THSD7A*	Thrombospondin, Type I, Domain Containing 7A	AI057226	0.19	9.3E-04	0.23	3.8E-04	0.22	1.3E-04
229254_at	*MFSD4*	Major Facilitator Superfamily Domain Containing 4	BE550027	0.06	1.8E-02	0.14	1.5E-02	0.06	6.1E-03
242372_s_at	*MFSD4*	Major Facilitator Superfamily Domain Containing 4	AL542291	0.27	2.7E-02	0.55	5.9E-02	0.25	1.9E-02
202016_at	*MEST*	Mesoderm Specific Transcript Homolog (Mouse)	NM_002402	2.48	2.4E-02	1.37	1.5E-01	3.5	6.1E-03
Functional group 2							**P-value =**	**2.53E-16**		
206773_at	*LY6H*	Lymphocyte Antigen 6 Complex, Locus H	NM_002347	0.18	1.5E-02	0.29	2.3E-02	0.12	1.9E-03
227566_at	*HNT*	Neurotrimin	AW085558	0.32	1.4E-02	0.55	1.0E-01	0.3	2.4E-03
214111_at	*OPCML*	Opioid Binding Protein/Cell Adhesion Molecule-Like	AF070577	0.04	3.5E-03	0.07	1.6E-03	0.04	5.4E-05
223730_at	*GPC6*	Glypican 6	AF111178	14.43	7.9E-04	9.49	4.6E-04	15.33	3.1E-04
227059_at	*GPC6*	Glypican 6	AI651255	19.81	8.0E-04	9.8	5.1E-04	17.98	3.3E-04
207174_at	*GPC5*	Glypican 5	NM_004466	0.1	1.7E-02	0.16	1.6E-03	0.12	1.7E-03
Functional group 3							**P-value =**	**8.57E-12**		
213880_at	*LGR5*	Leucine-Rich Repeat-Containing G Protein-Coupled Receptor 5	AL524520	0.11	4.7E-02	0.19	6.1E-02	0.17	3.6E-04
236734_at	*SLITRK1*	SLIT and NTRK-Like Family, Member 1	AI565671	0.08	1.6E-03	0.27	1.5E-02	0.09	6.2E-04
222108_at	*AMIGO2*	Adhesion Molecule with Ig-Like Domain 2	AC004010	0.82	2.3E-01	0.42	5.0E-03	1.28	1.3E-01
214930_at	*SLITRK5*	Slit And NTRK-Like Family, Member 5	AW449813	0.06	2.0E-03	0.14	3.7E-04	0.05	1.2E-05
226908_at	*LRIG3*	Leucine-Rich Repeats And Immunoglobulin-Like Domains 3	AI627704	4.44	3.6E-03	0.99	3.6E-01	5.48	1.5E-03
211596_s_at	*LRIG1*	Leucine-Rich Repeats and Immunoglobulin-Like Domains 1	AB050468	0.15	8.1E-04	0.17	2.7E-09	0.1	1.6E-06
230644_at	*LRFN5*	Leucine Rich Repeat And Fibronectin Type III Domain Containing 5	AI375083	1.55	1.2E-01	4.61	1.2E-03	1.98	1.9E-02
209840_s_at	*LRRN3*	Leucine Rich Repeat Neuronal 3	AI221950	0.06	1.9E-05	0.04	4.3E-07	0.05	9.3E-08
209841_s_at	*LRRN3*	Leucine Rich Repeat Neuronal 3	AL442092	0.06	1.9E-05	0.04	1.4E-07	0.04	8.4E-09

Affymetrix probeset	Gene symbol	Gene Description	Accesion number	Mg/Ag ratio	q-value	Mg/Gb ratio	q-value	Mg/Lgg ratio	q-value
226884_at	*LRRN1*	Leucine Rich Repeat Neuronal 1	N71874	0.19	1.4E-02	0.37	2.8E-03	0.15	1.3E-03
207191_s_at	*ISLR*	Immunoglobulin Superfamily Containing Leucine-Rich Repeat	NM_005545	27.28	2.7E-03	11.86	2.2E-03	36.77	1.3E-03
232720_at	*LINGO2*	Leucine Rich Repeat Neuronal 6C	AL353746	2.84	1.9E-02	6.5	4.0E-03	3.95	6.1E-03
207093_s_at	*OMG*	Oligodendrocyte Myelin Glycoprotein	NM_002544	0.02	2.6E-04	0.04	6.8E-05	0.02	1.2E-07
222853_at	*FLRT3*	Fibronectin Leucine Rich Transmembrane Protein 3	N71923	4.5	1.3E-03	6.33	4.5E-04	4.32	6.9E-04
203476_at	*TPBG*	Trophoblast Glycoprotein	NM_006670	3.52	1.6E-05	1.95	4.0E-03	5.36	3.1E-07
206408_at	*LRRTM2*	Leucine Rich Repeat Transmembrane Neuronal 2	NM_015564	0.04	6.9E-04	0.08	4.3E-07	0.04	1.1E-06
204359_at	*FLRT2*	Fibronectin Leucine Rich Transmembrane Protein 2	NM_013231	3.97	3.6E-02	5.24	2.3E-02	4.61	2.1E-02
1557123_a_at	*RP4-756G23.1*	Hypothetical Protein Bc012882	BC040188	0.51	7.5E-04	0.58	1.7E-03	0.46	1.1E-04
229085_at	*LRRC3B*	Leucine Rich Repeat Containing 3b	AW027879	0.06	3.1E-03	0.06	2.9E-03	0.06	1.7E-03
205150_s_at	*KIAA0644*	Kiaa0644 Gene Product	AV724192	0.1	8.5E-03	0.07	2.9E-03	0.08	7.6E-05
Functional group 4						**P-value =**	**1.31E-11**		
202709_at	*FMOD*	Fibromodulin	NM_002023	11.58	6.9E-04	7.26	4.3E-04	20.88	2.1E-04
201744_s_at	*LUM*	Lumican	NM_002345	2.14	1.7E-01	0.66	1.8E-01	8.3	4.1E-02
219087_at	*ASPN*	Asporin (LRR Class 1)	NM_017680	14.96	8.3E-02	3.51	1.1E-01	23.04	5.7E-02
204223_at	*PRELP*	Proline/Arginine-Rich End Leucine-Rich Repeat Protein	NM_002725	10.24	1.3E-03	11.95	6.3E-04	10.24	5.8E-04
205907_s_at	*OMD*	Osteomodulin	AI765819	84.31	5.8E-04	74.1	2.5E-04	66.86	2.2E-04
205908_s_at	*OMD*	Osteomodulin	NM_005014	110.57	5.4E-03	94.53	3.7E-03	95.4	3.1E-03
211896_s_at	*DCN*	Decorin	AF138302	5.3	8.9E-03	2.49	2.6E-02	8.27	3.5E-03
209335_at	*DCN*	Decorin	AI281593	6.78	1.0E-02	3.12	2.2E-02	12.61	4.4E-03
211813_x_at	*DCN*	Decorin	AF138303	4.98	1.2E-02	2.4	3.3E-02	6.96	5.1E-03
201893_x_at	*DCN*	Decorin	AF138300	4.4	8.2E-03	2.36	2.3E-02	6.54	3.0E-03
218730_s_at	*OGN*	Osteoglycin (Osteoinductive Factor, Mimecan)	NM_014057	97.93	1.2E-04	37.49	2.5E-05	78.37	2.9E-05
222722_at	*OGN*	Osteoglycin (Osteoinductive Factor, Mimecan)	AV700059	82.84	1.6E-05	32.13	1.1E-06	74.26	2.2E-06
Functional group 5						**P-value =**	**2.04E-11**		
205691_at	*SYNGR3*	Synaptogyrin 3	NM_004209	0.05	1.7E-02	0.11	1.6E-02	0.05	4.5E-03
242002_at	*TCBA1*	T-Cell Lymphoma Breakpoint Associated Target 1	N62814	0.09	1.3E-02	0.17	7.4E-03	0.07	1.0E-03
228017_s_at	*C20orf58*	Chromosome 20 Open Reading Frame 58	BF593263	0.06	3.0E-04	0.07	8.5E-04	0.04	1.1E-05
228018_at	*C20orf58*	Chromosome 20 Open Reading Frame 58	BF593263	0.1	7.1E-04	0.1	4.5E-04	0.06	5.8E-06
1558387_at	*FAM77D*	Family With Sequence Similarity 77, Member D	R41806	0.18	3.2E-03	0.14	8.3E-04	0.06	4.3E-05
1558388_a_at	*FAM77D*	Family With Sequence Similarity 77, Member D	R41806	0.05	3.0E-03	0.05	1.0E-04	0.02	8.6E-06
209655_s_at	*TMEM47*	Transmembrane Protein 47	AI803181	0.41	2.4E-02	0.4	7.9E-04	0.29	1.9E-03

Functional group 6						P-value = 4.84E-11			
				Mg/Ag		Mg/Gb		Mg/Lgg	
Affymetrix probeset	Gene symbol	Gene Description	Accesion number	ratio	q-value	ratio	q-value	ratio	q-value
223629_at	*PCDHB5*	Protocadherin Beta 5	BC001186	1.06	3.7E-01	1.42	1.1E-01	2.14	1.3E-02
1553105_s_at	*DSG2*	Desmoglein 2	NM_001943	4.19	3.6E-02	4.38	2.5E-02	9.13	1.1E-02
217901_at	*DSG2*	Desmoglein 2	BF031829	3.96	4.7E-02	4.61	2.3E-02	25.24	7.9E-03
203440_at	*CDH2*	Cadherin 2, Type 1, N-Cadherin (Neuronal)	M34064	0.11	1.9E-07	0.07	9.0E-12	0.1	8.6E-07
205656_at	*PCDH17*	Protocadherin 17	NM_014459	0.14	6.9E-05	0.16	3.8E-08	0.12	7.3E-09
228863_at	*PCDH17*	Protocadherin 17	N69091	0.1	5.0E-05	0.11	3.8E-08	0.08	6.4E-10
207149_at	*CDH12*	Cadherin 12, Type 2 (N-Cadherin 2)	L33477	0.19	1.7E-02	0.36	7.0E-02	0.24	8.8E-03
231738_at	*PCDHB7*	Protocadherin Beta 7	NM_018940	1.1	3.6E-01	0.58	8.7E-02	1.85	1.0E-01
228635_at	*PCDH10*	Protocadherin 10	AI640307	0.05	3.7E-04	0.06	3.9E-09	0.04	1.9E-05
206935_at	*PCDH8*	Protocadherin 8	NM_002590	0.04	1.8E-02	0.05	4.1E-05	0.04	2.3E-03
210292_s_at	*PCDH11X*	Protocadherin 11 X-Linked	AF332218	0.13	3.0E-03	0.28	1.0E-02	0.13	2.1E-05
219737_s_at	*PCDH9*	Protocadherin 9	AI524125	0.05	1.4E-03	0.05	2.1E-09	0.05	2.0E-05
219738_s_at	*PCDH9*	Protocadherin 9	NM_020403	0.08	1.8E-03	0.09	3.4E-07	0.08	5.7E-05
238919_at	*PCDH9*	Protocadherin 9	R49295	0.04	5.6E-04	0.04	1.9E-09	0.04	5.9E-06
220115_s_at	*CDH10*	Cadherin 10, Type 2 (T2-Cadherin)	NM_006727	0.06	2.6E-04	0.1	3.4E-06	0.04	2.9E-08
228640_at	*PCDH7*	Bh-Protocadherin (Brain-Heart)	BE644809	0.01	2.0E-03	0.02	2.1E-04	0.01	1.3E-05
201131_s_at	*CDH1*	Cadherin 1, Type 1, E-Cadherin (Epithelial)	NM_004360	7.93	3.9E-05	31.68	2.9E-06	24.6	3.2E-06
206898_at	*CDH19*	Cadherin 19, Type 2	NM_021153	0.16	5.4E-02	0.16	2.3E-03	0.48	9.5E-03
206280_at	*CDH18*	Cadherin 18, Type 2	NM_004934	0.1	1.0E-02	0.35	1.8E-02	0.13	7.3E-04
Functional group 7						**P-value = 1.05E-10**			
213194_at	*ROBO1*	Roundabout, Axon Guidance Receptor, Homolog 1 (Drosophila)	BF059159	0.22	1.4E-02	0.24	3.3E-05	0.2	1.6E-04
226766_at	*ROBO2*	Roundabout, Axon Guidance Receptor, Homolog 2 (Drosophila)	AB046788	0.13	7.1E-03	0.08	3.3E-05	0.08	5.8E-04
216617_s_at	*MAG*	Myelin Associated Glycoprotein	X98405	0.19	1.5E-02	0.2	6.5E-03	0.15	3.3E-03
214111_at	*OPCML*	Opioid Binding Protein/Cell Adhesion Molecule-Like	AF070577	0.04	3.5E-03	0.07	1.6E-03	0.04	5.4E-05
203868_s_at	*VCAM1*	Vascular Cell Adhesion Molecule 1	NM_001078	3.48	1.3E-02	1.73	6.0E-02	2.65	1.8E-02
204105_s_at	*NRCAM*	Neuronal Cell Adhesion Molecule	NM_005010	0.01	8.1E-05	0.01	9.7E-13	0.01	9.1E-08
233401_at	*NRCAM*	Neuronal Cell Adhesion Molecule	BF723605	0.06	3.8E-04	0.05	7.8E-08	0.05	7.0E-04
230045_at	*CNTN2*	Contactin 2 (Axonal)	BF740264	0.09	2.4E-02	0.12	4.3E-03	0.09	2.1E-03
215028_at	*SEMA6A*	Sema Domain, Transmembrane Domain (Tm), And Cytoplasmic Domain, (Semaphorin) 6a	AB002438	0.1	1.2E-03	0.07	6.5E-07	0.07	1.7E-05
225660_at	*SEMA6A*	Sema Domain, Transmembrane Domain (Tm), And Cytoplasmic Domain, (Semaphorin) 6A	W92748	0.08	4.3E-04	0.06	2.0E-09	0.06	1.5E-06
212843_at	*NCAM1*	Neural Cell Adhesion Molecule 1	AA126505	0.11	5.3E-05	0.13	2.5E-07	0.09	9.7E-09
213438_at	*NFASC*	Neurofascin Homolog (Chicken)	AA995925	0.05	2.7E-04	0.05	1.1E-07	0.05	9.5E-08
204584_at	*L1CAM*	L1 Cell Adhesion Molecule	AI653981	0.06	9.0E-03	0.18	5.2E-03	0.07	1.1E-03
227566_at	*HNT*	Neurotrimin	AW085558	0.32	1.4E-02	0.55	1.0E-01	0.3	2.4E-03

Affymetrix probeset	Gene symbol	Gene Description	Accesion number	Mg/Ag ratio	q-value	Mg/Gb ratio	q-value	Mg/Lgg ratio	q-value
1554784_at	*CNTN1*	Contactin 1	BC036569	0.19	9.0E-03	0.41	1.1E-02	0.16	5.7E-06
227202_at	*CNTN1*	Contactin 1	AW072790	0.03	4.4E-03	0.06	1.5E-03	0.02	1.4E-06
227209_at	*CNTN1*	Contactin 1	AI091445	0.05	2.7E-03	0.12	9.7E-05	0.04	2.3E-07
1553194_at	*NEGR1*	Neuronal Growth Regulator 1	NM_173808	0.43	2.8E-02	0.66	5.8E-02	0.29	5.3E-03
243357_at	*NEGR1*	Neuronal Growth Regulator 1	AA115106	0.09	9.0E-03	0.24	6.8E-03	0.08	7.9E-04
229461_x_at	*NEGR1*	Neuronal Growth Regulator 1	AI123532	0.04	6.5E-03	0.09	5.2E-03	0.03	5.8E-04
229831_at	*CNTN3*	Contactin 3 (Plasmacytoma Associated)	BE221817	0.45	9.1E-02	1.22	2.7E-01	0.28	3.0E-03
204006_s_at	*FCGR3B*	Fc Fragment of Igg, Low Affinity IIIb, Receptor (CD16B)	NM_000570	0.66	2.5E-01	0.38	1.6E-02	0.78	2.4E-01
Functional group 8						**P-value =**	**2.48E-10**		
232010_at	*FSTL5*	Follistatin-Like 5	AA129444	0.09	2.0E-02	0.15	8.1E-02	0.08	8.3E-04
206434_at	*SPOCK3*	Sparc/Osteonectin, CWCV And Kazal-Like Domains Proteoglycan (Testican) 3	NM_016950	0.16	2.0E-02	0.2	2.9E-02	0.13	1.2E-03
235342_at	*SPOCK3*	Sparc/Osteonectin, CWCV And Kazal-Like Domains Proteoglycan (Testican) 3	AI808090	0.03	7.3E-03	0.05	4.9E-03	0.03	2.2E-04
202363_at	*SPOCK1*	Sparc/Osteonectin, CWCV And Kazal-Like Domains Proteoglycan (Testican) 1	AF231124	0.02	8.3E-04	0.03	2.0E-06	0.02	1.4E-05
222784_at	*SMOC1*	Sparc Related Modular Calcium Binding 1	AJ249900	0.65	1.7E-01	1.75	6.1E-02	0.42	3.8E-03
223235_s_at	*SMOC2*	Sparc Related Modular Calcium Binding 2	AB014737	28.1	2.7E-03	8.75	2.6E-03	28.4	1.4E-03
223885_at	*CALN1*	Calneuron 1	AF282250	0.12	1.0E-02	0.5	7.5E-02	0.13	1.2E-02
Functional group 9						**P-value =**	**4.19E-10**		
210220_at	*FZD2*	Frizzled Homolog 2 (Drosophila)	AW072102	0.06	6.4E-03	0.23	2.1E-02	0.06	2.7E-03
205638_at	*BAI3*	Brain-Specific Angiogenesis Inhibitor 3	L37882	9.69	1.6E-03	6.24	1.3E-03	13.66	6.3E-04
203705_s_at	*FZD7*	Frizzled Homolog 7 (Drosophila)	NM_001704	0.03	6.5E-04	0.05	1.6E-06	0.02	8.4E-08
203706_s_at	*FZD7*	Frizzled Homolog 7 (Drosophila)	AI333651	11.86	4.9E-04	4.46	5.3E-04	15.63	1.6E-04
226690_at	*ADCYAP1R1*	Adenylate Cyclase Activating Polypeptide 1 (Pituitary) Receptor Type I	NM_003507	11.58	5.1E-04	4.75	5.1E-04	16.79	1.6E-04
227769_at	*GPR27*	G Protein-Coupled Receptor 27	AW451961	0.06	6.9E-04	0.07	4.9E-05	0.03	3.8E-06
205814_at	*GRM3*	Glutamate Receptor, Metabotropic 3	AI703476	1.13	3.0E-01	2.7	2.7E-03	0.88	2.1E-01
206825_at	*OXTR*	Oxytocin Receptor	NM_000840	0.08	1.0E-02	0.15	5.8E-03	0.07	2.0E-03
214217_at	*GRM5*	Glutamate Receptor, Metabotropic 5	NM_000916	0.26	6.4E-02	0.11	2.6E-02	0.58	1.4E-02
209990_s_at	*GABBR2*	Gamma-Aminobutyric Acid (GABA) B Receptor, 2	D60132	0.08	7.6E-03	0.25	1.7E-02	0.07	5.8E-04
209991_x_at	*GABBR2*	Gamma-Aminobutyric Acid (GABA) B Receptor, 2	AF056085	0.02	3.9E-03	0.03	4.4E-03	0.01	8.3E-05
211679_x_at	*GABBR2*	Gamma-Aminobutyric Acid (GABA) B Receptor, 2	AF069755	0.11	4.0E-03	0.23	1.2E-02	0.06	5.9E-04
217077_s_at	*GABBR2*	Gamma-Aminobutyric Acid (GABA) B Receptor, 2	AF095784	0.1	3.8E-03	0.18	3.2E-03	0.06	1.4E-04
1554008_at	*OSMR*	Oncostatin M Receptor	AF095723	0.06	4.5E-03	0.11	7.4E-03	0.03	2.3E-04
206899_at	*NTSR2*	Neurotensin Receptor 2	BC010943	0.88	2.7E-01	0.4	3.8E-04	1.18	1.1E-01

Affymetrix probeset	Gene symbol	Gene Description	Accesion number	Mg/Ag ratio	Mg/Ag q-value	Mg/Gb ratio	Mg/Gb q-value	Mg/Lgg ratio	Mg/Lgg q-value
1560225_at	*CNR1*	Cannabinoid Receptor 1 (Brain)	NM_012344	0.11	1.4E-02	0.3	1.2E-02	0.07	7.8E-03
213436_at	*CNR1*	Cannabinoid Receptor 1 (Brain)	AI434253	0.67	1.4E-02	0.29	8.1E-03	0.7	1.3E-03
1555533_at	*GPR103*	G Protein-Coupled Receptor 103	U73304	0.03	4.6E-03	0.01	1.3E-05	0.03	7.2E-04
213592_at	*AGTRL1*	Angiotensin II Receptor-Like 1	AF411117	2.12	5.5E-02	2.18	4.4E-02	2.12	4.0E-02
206785_s_at	*KLRC2*	Killer Cell Lectin-Like Receptor Subfamily C, Member 2	X89271	0.22	1.3E-02	0.14	9.8E-05	0.18	2.1E-03
220005_at	*P2RY13*	Purinergic Receptor P2Y, G-Protein Coupled, 13	NM_002260	0.03	1.6E-02	0.23	2.1E-02	0.03	1.0E-03
215306_at	*STON1*	Stonin 1	NM_023914	3.02	3.2E-02	3.25	2.2E-02	1.62	1.0E-01
232195_at	*GPR158*	G Protein-Coupled Receptor 158	AL049443	0.6	8.7E-02	0.39	2.7E-03	0.25	1.4E-02
209631_s_at	*GPR37*	G Protein-Coupled Receptor 37 (Endothelin Receptor Type B-Like)	R41459	0.1	3.5E-03	0.2	9.0E-06	0.08	5.9E-06
214586_at	*GPR37*	G Protein-Coupled Receptor 37 (Endothelin Receptor Type B-Like)	U87460	0.03	5.3E-03	0.02	5.5E-08	0.04	8.5E-04
213506_at	*F2RL1*	Coagulation Factor II (Thrombin) Receptor-Like 1	T16257	0.17	8.7E-03	0.11	3.4E-06	0.16	2.1E-03
1553316_at	*GPR82*	G Protein-Coupled Receptor 82	BE965369	0.79	3.0E-01	0.51	6.7E-02	1.07	3.2E-01
1553317_s_at	*GPR82*	G Protein-Coupled Receptor 82	NM_080817	2.72	2.5E-02	2.3	3.2E-02	2.98	1.4E-02
206785_s_at	*KLRC2*	Killer Cell Lectin-Like Receptor Subfamily C, Member 1	AL832460	2.8	4.5E-02	2.14	6.3E-02	3.66	2.0E-02
1554706_at	*OR2L13*	Olfactory Receptor, Family 2, Subfamily L, Member 13	NM_002260	0.03	1.6E-02	0.23	2.1E-02	0.03	1.0E-03
219896_at	*DRD1IP*	Dopamine Receptor D1 Interacting Protein	BC028158	0.5	5.5E-02	0.77	1.1E-01	0.6	1.2E-02
206190_at	*GPR17*	G Protein-Coupled Receptor 17	NM_015722	0.16	2.0E-02	0.39	5.2E-02	0.18	1.1E-02
215225_s_at	*GPR17*	G Protein-Coupled Receptor 17	NM_005291	0.04	7.1E-02	0.09	4.5E-02	0.03	3.1E-02
AFFX-HUMRGE/M10098_M_at	*GPR34*	G Protein-Coupled Receptor 34	Z94154	0.11	9.2E-02	0.25	4.5E-02	0.1	3.2E-02
221288_at	*GPR22*	G Protein-Coupled Receptor 22	-	1.57	2.4E-01	0.49	6.3E-02	1.08	3.2E-01
224102_at	*P2RY12*	Purinergic Receptor P2Y, G-Protein Coupled, 12	NM_005295	0.18	3.8E-02	0.44	6.0E-02	0.24	1.4E-02
235885_at	*P2RY12*	Purinergic Receptor P2Y, G-Protein Coupled, 13	AF321815	0.85	3.0E-01	1.32	2.1E-01	0.44	1.7E-02
213880_at	*LGR5*	Leucine-Rich Repeat-Containing G Protein-Coupled Receptor 5	AA810452	0.81	2.9E-01	0.99	3.6E-01	0.35	9.1E-03
215894_at	*PTGDR*	Prostaglandin D2 Receptor (Dp)	AL524520	0.11	4.7E-02	0.19	6.1E-02	0.17	3.6E-04
234165_at	*PTGDR*	Prostaglandin D2 Receptor (Dp)	AI460323	141.57	2.3E-03	139.82	1.4E-03	139.32	1.2E-03
204563_at	*SELL*	Selectin L (Lymphocyte Adhesion Molecule 1)	AK026202	34.38	2.5E-03	35.12	1.5E-03	33.43	1.3E-03
209469_at	*GPM6A*	Glycoprotein M6A	NM_000655	0.48	1.3E-01	0.54	7.0E-02	0.13	1.9E-02
209470_s_at	*GPM6A*	Glycoprotein M6A	BF939489	0.01	1.1E-05	0.02	3.2E-11	0.01	8.8E-12
236024_at	*GPM6A*	Glycoprotein M6A	D49958	0.01	8.4E-06	0.01	9.1E-11	0.01	9.2E-11
219761_at	*CLEC1A*	C-Type Lectin Domain Family 1, Member A	AW136286	0.04	1.0E-03	0.09	5.1E-05	0.03	2.1E-05

Functional group 10						P-value = 6.51E-10			
Affymetrix probeset	Gene symbol	Gene Description	Accesion number	Mg/Ag ratio	q-value	Mg/Gb ratio	q-value	Mg/Lgg ratio	q-value
219890_at	*CLEC5A*	C-Type Lectin Domain Family 5, Member A	NM_016511	21.72	2.3E-04	17.83	6.9E-05	20.3	6.2E-05
1556209_at	*CLEC2B*	C-Type Lectin Domain Family 2, Member B	NM_013252	2.42	5.4E-02	0.71	1.4E-01	8.38	6.1E-03
201438_at	*COL6A3*	Collagen, Type VI, Alpha 3	CA447397	1.77	1.6E-01	1.97	7.1E-02	4.1	1.4E-02
201852_x_at	*COL3A1*	Collagen, Type III, Alpha 1 (Ehlers-Danlos Syndrome Type IV, Autosomal Dominant)	NM_004369	4.92	5.4E-03	1.01	3.6E-01	18.09	1.2E-03
215076_s_at	*COL3A1*	Collagen, Type III, Alpha 1 (Ehlers-Danlos Syndrome Type IV, Autosomal Dominant)	AI813758	9.04	5.2E-03	1.98	5.4E-02	32.93	1.9E-03
211161_s_at	*COL3A1*	Collagen, Type III, Alpha 1 (Ehlers-Danlos Syndrome Type IV, Autosomal Dominant)	AU144167	7.06	1.0E-03	1.92	2.3E-02	26.38	2.8E-04
1556499_s_at	*COL1A1*	Collagen, Type I, Alpha 1	AF130082	10.16	4.2E-03	1.94	5.1E-02	33.11	1.6E-03
202310_s_at	*COL1A1*	Collagen, Type I, Alpha 1	BE221212	6.85	3.2E-05	2.18	5.1E-04	21.68	8.3E-06
202311_s_at	*COL1A1*	Collagen, Type I, Alpha 1	K01228	12.47	5.7E-03	2.83	2.2E-02	47.96	2.4E-03
211980_at	*COL4A1*	Collagen, Type IV, Alpha 1	AI743621	8.19	2.4E-02	1.45	2.0E-01	16.32	1.2E-02
204320_at	*COL11A1*	Collagen, Type XI, Alpha 1	AI922605	1.27	3.0E-01	0.45	1.3E-02	2.34	8.5E-02
229271_x_at	*COL11A1*	Collagen, Type XI, Alpha 1	NM_001854	11.43	3.6E-03	7.69	3.0E-03	13.25	1.8E-03
37892_at	*COL11A1*	Collagen, Type XI, Alpha 1	BG028597	15.95	1.5E-02	8	1.6E-02	12.41	1.0E-02
202403_s_at	*COL1A2*	Collagen, Type I, Alpha 2	J04177	14.22	4.7E-03	8.05	4.3E-03	16.53	2.5E-03
202404_s_at	*COL1A2*	Collagen, Type I, Alpha 2	AA788711	12.69	3.8E-05	2.81	8.9E-05	27.4	7.5E-06
229218_at	*COL1A2*	Collagen, Type I, Alpha 2	NM_000089	9.24	1.9E-03	1.93	3.6E-02	27.74	6.2E-04
209156_s_at	*COL6A2*	Collagen, Type VI, Alpha 2	AA628535	27.43	2.1E-04	4.1	1.9E-04	43.26	5.4E-05
225664_at	*COL12A1*	Collagen, Type XII, Alpha 1	AY029208	10.66	2.4E-04	4.05	1.9E-04	34.78	5.6E-05
212865_s_at	*COL14A1*	Collagen, Type XIV, Alpha 1 (Undulin)	AA788946	5.01	7.2E-02	1.28	2.8E-01	7.77	4.2E-02
Functional group 11						**P-value = 1.67E-09**			
1552848_a_at	*PTCHD1*	Patched Domain Containing 1	BF449063	3.32	5.0E-02	0.81	2.7E-01	2.99	4.2E-02
242002_at	*TCBA1*	T-Cell Lymphoma Breakpoint Associated Target 1	NM_173495	0.14	2.8E-03	0.1	6.0E-08	0.15	3.5E-04
1552782_at	*SLC44A5*	Solute Carrier Family 44, Member 5	N62814	0.09	1.3E-02	0.17	7.4E-03	0.07	1.0E-03
235763_at	*SLC44A5*	Solute Carrier Family 44, Member 5	AK091400	0.46	2.0E-02	0.45	9.5E-03	0.38	1.7E-02
1552327_at	*ARMCX4*	Armadillo Repeat Containing, X-Linked 4	AA001450	0.15	2.1E-03	0.14	1.1E-03	0.11	2.4E-03
236044_at	*PPAPDC1A*	Phosphatidic Acid Phosphatase Type 2 Domain Containing 1a	NM_152583	0.54	4.2E-02	0.96	3.0E-01	0.43	5.4E-04
1557176_a_at	*C14orf37*	Chromosome 14 Open Reading Frame 37	BF130943	0.49	2.3E-02	0.56	3.8E-02	0.29	1.1E-03
219895_at	*FAM70A*	Family With Sequence Similarity 70, Member A	BU074567	0.58	4.3E-03	0.5	6.9E-04	0.42	3.3E-04

Functional group 12						P-value =	2.14E-09		
						Mg/Gb		Mg/Lgg	
Affymetrix probeset	Gene symbol	Gene Description	Accesion number	Mg/Ag ratio	q-value	ratio	q-value	ratio	q-value
217022_s_at	*IGHA1*	Immunoglobulin Heavy Constant Alpha 1	NM_017938	0.05	2.5E-03	0.04	1.9E-07	0.06	3.9E-04
214677_x_at	*IGLJ3*	Immunoglobulin Lambda Variable 3-25	S55735	0.61	2.6E-01	0.11	3.6E-02	0.41	1.5E-01
215379_x_at	*IGHG2*	Immunoglobulin Heavy Constant Gamma 2 (G2M Marker)	X57812	0.76	3.1E-01	0.18	3.5E-02	0.32	1.2E-01
215121_x_at	*IGLA2*	Immunoglobulin Heavy Constant Alpha 2 (A2M Marker)	AV698647	0.84	3.2E-01	0.23	3.5E-02	0.35	1.1E-01
211430_s_at	*IGHG3*	Immunoglobulin Kappa Variable 1-5	AA680302	0.85	3.2E-01	0.28	4.1E-02	0.43	1.2E-01
217022_s_at	*IGHA1*	Immunoglobulin Heavy Constant Gamma 1 (G1m Marker)	M87789	0.51	2.8E-01	0.14	4.4E-02	0.26	1.3E-01
221651_x_at	*IGKC*	Immunoglobulin Kappa Constant	S55735	0.61	2.6E-01	0.11	3.6E-02	0.41	1.5E-01
221671_x_at	*IGKC*	Immunoglobulin Kappa Constant	BC005332	0.61	2.5E-01	0.25	1.7E-02	0.44	1.3E-01
224795_x_at	*IGKC*	Immunoglobulin Kappa Constant	M63438	0.62	2.6E-01	0.25	1.4E-02	0.42	1.3E-01
211430_s_at	*IGHG3*	Immunoglobulin Heavy Constant Gamma 3 (G3M Marker)	AW575927	0.62	2.6E-01	0.25	1.5E-02	0.42	1.3E-01
214669_x_at	*IGKC*	Immunoglobulin Kappa Constant	M87789	0.51	2.8E-01	0.14	4.4E-02	0.26	1.3E-01
224795_x_at	*IGKC*	Immunoglobulin Kappa Constant	BG485135	0.68	2.4E-01	0.28	2.7E-02	0.47	1.2E-01
221651_x_at	*IGKC*	Immunoglobulin Kappa Constant	AW575927	0.62	2.6E-01	0.25	1.5E-02	0.42	1.3E-01
221671_x_at	*IGKC*	Immunoglobulin Kappa Constant	BC005332	0.61	2.5E-01	0.25	1.7E-02	0.44	1.3E-01
211430_s_at	*IGHG3*	Immunoglobulin Heavy Constant Gamma 3 (G3M Marker)	M63438	0.62	2.6E-01	0.25	1.4E-02	0.42	1.3E-01
Functional group 13							**P-value =**	**2.27E-09**	
205384_at	*FXYD1*	FXYD Domain Containing Ion Transport Regulator 1 (Phospholemman)	M87789	0.51	2.8E-01	0.14	4.4E-02	0.26	1.3E-01
1560265_at	*GRIK2*	Glutamate Receptor, Ionotropic, Kainate 2	NM_005031	3.22	1.1E-03	4.47	2.4E-04	3.57	3.4E-04
213845_at	*GRIK2*	Glutamate Receptor, Ionotropic, Kainate 2	BQ434382	0.24	8.1E-03	0.61	6.7E-03	0.19	2.6E-03
231384_at	*GRIN2A*	Glutamate Receptor, Ionotropic, N-Methyl D-Aspartate 2a	AL355532	0.09	1.2E-03	0.16	9.4E-04	0.07	3.8E-05
242286_at	*GRIN2A*	Glutamate Receptor, Ionotropic, N-Methyl D-Aspartate 2a	T65537	0.06	2.4E-02	0.13	2.3E-02	0.06	1.2E-02
205279_s_at	*GLRB*	Glycine Receptor, Beta	N48896	0.12	2.1E-02	0.29	2.2E-02	0.11	1.6E-02
205280_at	*GLRB*	Glycine Receptor, Beta	AF094754	0.08	8.7E-04	0.14	1.7E-05	0.07	9.5E-05
205358_at	*GRIA2*	Glutamate Receptor, Ionotropic, AMPA 2	NM_000824	0.04	2.0E-03	0.07	5.1E-05	0.04	1.2E-04
236538_at	*GRIA2*	Glutamate Receptor, Ionotropic, AMPA 2	NM_000826	0.01	2.4E-04	0.01	2.4E-06	0	7.3E-08
215634_at	*GRIA1*	Glutamate Receptor, Ionotropic, AMPA 1	BE219628	0.01	8.6E-04	0.02	1.1E-05	0.01	3.1E-07
209793_at	*GRIA1*	Glutamate Receptor, Ionotropic, AMPA 1	AF007137	0.32	7.1E-02	0.46	1.2E-01	0.33	1.6E-02
238663_x_at	*GRIA4*	Glutamate Receptor, Ionotrophic, AMPA 4	AL567302	0.12	9.4E-04	0.11	3.1E-04	0.09	1.2E-06
1555268_a_at	*GRID1*	Glutamate Receptor, Ionotropic, Delta 1	H20055	0.04	3.0E-03	0.11	2.3E-04	0.04	4.3E-07
1554748_at	*CLCNKB*	Chloride Channel KB	BC039263	0.45	8.7E-04	0.63	1.2E-04	0.4	1.8E-06

Affymetrix probeset	Gene symbol	Gene Description	Accesion number	Mg/Ag ratio	q-value	Mg/Gb ratio	q-value	Mg/Lgg ratio	q-value
206456_at	*GABRA5*	Gamma-Aminobutyric Acid (Gaba) A Receptor, Alpha 5	BC020873	3.64	5.4E-03	3.17	5.2E-03	3.97	2.7E-03
1552943_at	*GABRG1*	Gamma-Aminobutyric Acid (Gaba) A Receptor, Gamma 1	NM_000810	0.06	2.7E-02	0.15	5.4E-02	0.05	1.6E-02
241805_at	*GABRG1*	Gamma-Aminobutyric Acid (Gaba) A Receptor, Gamma 1	NM_173536	0.51	2.1E-02	0.82	1.1E-01	0.48	1.2E-02
233437_at	*GABRA4*	Gamma-Aminobutyric Acid (Gaba) A Receptor, Alpha 4	N92667	0.09	1.4E-02	0.25	1.4E-02	0.09	1.4E-03
206849_at	*GABRG2*	Gamma-Aminobutyric Acid (Gaba) A Receptor, Gamma 2	AF238869	0.07	1.7E-02	0.24	2.5E-02	0.08	5.8E-03
1568612_at	*GABRG2*	Gamma-Aminobutyric Acid (Gaba) A Receptor, Gamma 2	NM_000816	0.14	5.9E-03	0.31	4.5E-02	0.1	2.8E-03
1554308_s_at	*GABRA2*	Gamma-Aminobutyric Acid (Gaba) A Receptor, Alpha 2	BC036030	0.02	1.4E-02	0.06	2.3E-02	0.02	1.2E-03
207014_at	*GABRA2*	Gamma-Aminobutyric Acid (Gaba) A Receptor, Alpha 2	BC022488	0.37	2.8E-02	0.57	2.0E-02	0.42	7.7E-03
206678_at	*GABRA1*	Gamma-Aminobutyric Acid (Gaba) A Receptor, Beta 1	NM_000807	0.04	4.9E-02	0.1	1.3E-02	0.05	5.7E-03
244118_at	*GABRA1*	Gamma-Aminobutyric Acid (Gaba) A Receptor, Beta 1	NM_000806	0.05	1.3E-02	0.13	3.1E-02	0.05	6.7E-03
207010_at	*GABRB1*	Gamma-Aminobutyric Acid (Gaba) A Receptor, Beta 1	AV722228	0.05	1.8E-02	0.11	3.4E-02	0.05	1.2E-02
227690_at	*GABRB3*	Gamma-Aminobutyric Acid (Gaba) A Receptor, Beta 3	NM_000812	0.03	3.4E-03	0.06	2.3E-04	0.03	1.1E-04
229724_at	*GABRB3*	Gamma-Aminobutyric Acid (Gaba) A Receptor, Beta 3	BE502537	0.06	2.1E-03	0.25	3.2E-02	0.07	4.0E-05
227830_at	*GABRB3*	Gamma-Aminobutyric Acid (Gaba) A Receptor, Beta 3	AI693153	0.1	3.7E-03	0.47	7.4E-02	0.09	5.7E-05
221107_at	*CHRNA9*	Cholinergic Receptor, Nicotinic, Alpha 9	AI478781	0.13	2.3E-03	0.53	7.2E-02	0.13	8.2E-05
Functional group 14						**P-value =**	**5.57E-08**		
1552327_at	*ARMCX4*	Armadillo Repeat Containing, X-Linked 4	NM_017581	0.54	2.2E-01	0.18	9.2E-03	0.98	3.1E-01
237802_at	*XKR4*	Xk, Kell Blood Group Complex Subunit-Related Family, Member 4	NM_152583	0.54	4.2E-02	0.96	3.0E-01	0.43	5.4E-04
223343_at	*MS4A7*	Membrane-Spanning 4-Domains, Subfamily A, Member 7	R54212	0.04	2.3E-03	0.12	2.4E-04	0.05	2.8E-06
1558102_at	*TM6SF1*	Transmembrane 6 Superfamily Member 1	AI301935	2.11	6.3E-02	1.23	2.2E-01	3.11	1.6E-02
207723_s_at	*KLRC3*	Killer Cell Lectin-Like Receptor Subfamily C, Member 3	AK055438	0.48	5.3E-03	0.75	3.6E-02	0.47	1.5E-04
213609_s_at	*SEZ6L*	Seizure Related Gene 6 (Mouse)-Like	NM_002261	0.09	5.5E-02	0.7	1.5E-01	0.09	2.2E-03
240709_at	*SEZ6L*	Seizure Related Gene 6 (Mouse)-Like	AB023144	0.03	8.0E-03	0.11	1.8E-04	0.02	2.2E-05
231650_s_at	*SEZ6L*	Seizure Related Gene 6 (Mouse)-Like	AW204757	0.14	3.3E-02	0.43	5.5E-03	0.07	1.5E-04
242301_at	*CBLN2*	Cerebellin 2 Precursor	BE672217	0.03	1.4E-02	0.15	5.0E-04	0.03	1.5E-05
1554689_a_at	*NLGN4X*	Neuroligin 4, X-Linked	R60224	0.12	4.4E-02	0.36	4.3E-02	0.15	1.3E-02
221933_at	*NLGN4X*	Neuroligin 4, X-Linked	BC034018	0.22	5.3E-04	0.19	8.8E-06	0.19	4.4E-05

Affymetrix probeset	Gene symbol	Gene Description	Accesion number	Mg/Ag ratio	Mg/Ag q-value	Mg/Gb ratio	Mg/Gb q-value	Mg/Lgg ratio	Mg/Lgg q-value
1555728_a_at	*MS4A4A*	Membrane-Spanning 4-Domains, Subfamily A, Member 4	AI338338	0.05	9.7E-05	0.04	2.3E-07	0.04	8.0E-06
219607_s_at	*MS4A4A*	Membrane-Spanning 4-Domains, Subfamily A, Member 4	AF354928	1.71	1.8E-01	0.97	3.5E-01	3.51	3.2E-02
1554452_a_at	*HIG2*	Hypoxia-Inducible Protein 2	NM_024021	1.77	1.7E-01	1	3.6E-01	3.54	3.9E-02
218507_at	*HIG2*	Hypoxia-Inducible Protein 2	BC001863	0.28	5.5E-02	0.17	9.5E-04	0.99	3.4E-01
219410_at	*TMEM45A*	Transmembrane Protein 45a	NM_013332	0.2	6.5E-02	0.14	1.8E-04	0.77	7.2E-02
219890_at	*CLEC5A*	C-Type Lectin Domain Family 5, Member A	NM_018004	1.89	1.1E-02	0.93	2.8E-01	2.48	2.4E-03
227307_at	*TSPAN18*	Tetraspanin 18	NM_013252	2.42	5.4E-02	0.71	1.4E-01	8.38	6.1E-03
214841_at	*CNIH3*	Cornichon Homolog 3 (Drosophila)	AL565381	11.8	8.1E-04	7.97	4.9E-04	10.82	3.5E-04
1559063_at	*C21orf63*	Chromosome 21 Open Reading Frame 63	AF070524	0.2	4.9E-02	0.15	3.1E-05	0.31	1.3E-05
219274_at	*TSPAN12*	Tetraspanin 12	AL355689	3.29	9.5E-03	2.76	1.1E-02	3.98	4.1E-03
215241_at	*TMEM16C*	Transmembrane Protein 16C	NM_012338	0.55	1.7E-01	0.4	1.8E-02	0.35	2.4E-02
218888_s_at	*NETO2*	Neuropilin (Nrp) And Tolloid (Tll)-Like 2	AJ300461	0.14	3.9E-02	0.39	6.9E-02	0.17	1.2E-02
222774_s_at	*NETO2*	Neuropilin (Nrp) And Tolloid (Tll)-Like 2	NM_018092	0.14	7.9E-04	0.17	1.4E-03	0.15	3.7E-03
239575_at	*TMEM10*	Transmembrane Protein 10	AI335263	0.17	3.9E-03	0.19	4.2E-03	0.19	1.1E-02
203963_at	*CA12*	Carbonic Anhydrase XII	N63401	0.11	3.4E-02	0.19	7.7E-02	0.09	1.6E-02
209655_s_at	*TMEM47*	Transmembrane Protein 47	NM_001218	0.55	1.8E-01	0.25	6.4E-04	1.23	2.3E-01
1552736_a_at	*NETO1*	Neuropilin (Nrp) And Tolloid (Tll)-Like 1	AI803181	0.41	2.4E-02	0.4	7.9E-04	0.29	1.9E-03
206785_s_at	*KLRC2*	Killer Cell Lectin-Like Receptor Subfamily C, Member 1	NM_138966	0.2	1.4E-02	0.43	9.5E-03	0.24	1.9E-03
219666_at	*MS4A6A*	Membrane-Spanning 4-Domains, Subfamily A, Member 6a (Cd20-Like Precursor)	NM_002260	0.03	1.6E-02	0.23	2.1E-02	0.03	1.0E-03
223280_x_at	*MS4A6A*	Membrane-Spanning 4-Domains, Subfamily A, Member 6a (Cd20-Like Precursor)	NM_022349	2.48	5.2E-02	1.7	9.2E-02	4.33	1.2E-02
230550_at	*MS4A6A*	Membrane-Spanning 4-Domains, Subfamily A, Member 6a (Cd20-Like Precursor)	AF253977	2.23	4.8E-02	1.22	2.2E-01	3.64	8.8E-03
224356_x_at	*MS4A6A*	Membrane-Spanning 4-Domains, Subfamily A, Member 6a (Cd20-Like Precursor)	AA045175	4.85	3.8E-02	2.48	6.7E-02	7.08	2.0E-02
205691_at	*SYNGR3*	Synaptogyrin 3	AF237908	2.27	4.2E-02	1.31	1.7E-01	3.98	6.3E-03
223557_s_at	*TMEFF2*	Transmembrane Protein With EGF-Like And Two Follistatin-Like Domains 2	NM_004209	0.05	1.7E-02	0.11	1.6E-02	0.05	4.5E-03
224321_at	*TMEFF2*	Transmembrane Protein With EGF-Like And Two Follistatin-Like Domains 2	AB017269	0.02	1.3E-04	0.04	8.4E-05	0.02	5.2E-06
219602_s_at	*FAM38B*	Family With Sequence Similarity 38, Member B	AB004064	1.4	2.8E-01	1.22	2.8E-01	3	7.1E-02
1552782_at	*SLC44A5*	Solute Carrier Family 44, Member 5	NM_022068	3.79	1.3E-02	3.82	1.0E-02	4.68	6.1E-03
235763_at	*SLC44A5*	Solute Carrier Family 44, Member 5	AK091400	0.46	2.0E-02	0.45	9.5E-03	0.38	1.7E-02

Functional group 15						P-value = 8.99E-08			
Affymetrix probeset	Gene symbol	Gene Description	Accesion number	Mg/Ag ratio	q-value	Mg/Gb ratio	q-value	Mg/Lgg ratio	q-value
1554485_s_at	*TMEM37*	Transmembrane Protein 37	AA001450	0.15	2.1E-03	0.14	1.1E-03	0.11	2.4E-03
213395_at	*MLC1*	Megalencephalic Leukoencephalopathy With Subcortical Cysts 1	BI825302	2.02	1.2E-01	2.65	3.0E-02	5.06	9.3E-03
219415_at	*TTYH1*	Tweety Homolog 1 (Drosophila)	AL022327	0.07	5.4E-04	0.04	5.9E-09	0.04	1.7E-05
204519_s_at	*PLLP*	Plasma Membrane Proteolipid (Plasmolipin)	NM_020659	0.04	3.0E-04	0.04	5.6E-08	0.03	4.0E-07
220131_at	*FXYD7*	Fxyd Domain Containing Ion Transport Regulator 7	NM_015993	0.13	7.3E-03	0.14	2.4E-04	0.1	3.3E-04
205384_at	*FXYD1*	Fxyd Domain Containing Ion Transport Regulator 1 (Phospholemman)	NM_022006	0.2	1.8E-02	0.45	7.3E-02	0.17	1.4E-02
228608_at	*NALCN*	Sodium Leak Channel, Non-Selective	NM_005031	3.22	1.1E-03	4.47	2.4E-04	3.57	3.4E-04
1555492_a_at	*BEST3*	Bestrophin 3	N49852	2.8	5.7E-02	5.34	2.3E-02	1.99	7.6E-02
237003_at	*BEST3*	Bestrophin 3	AF440758	0.21	1.8E-02	0.5	7.4E-02	0.24	6.4E-03
224520_s_at	*BEST3*	Bestrophin 3	AA878383	0.16	9.0E-03	0.26	6.8E-03	0.15	1.9E-03
Functional group 16						**P-value =**	**2.45E-07**		
230303_at	*SYNPR*	Synaptoporin	BC006440	0.07	2.7E-02	0.25	1.2E-02	0.07	2.0E-03
229818_at	*SVOP*	Sv2 Related Protein Homolog (Rat)	H11380	0.03	2.9E-02	0.07	3.2E-02	0.03	1.4E-02
203069_at	*SV2A*	Synaptic Vesicle Glycoprotein 2a	AL359592	0.12	5.8E-03	0.42	7.0E-02	0.11	4.4E-03
216086_at	*SV2C*	Synaptic Vesicle Glycoprotein 2c	NM_014849	0.1	2.2E-04	0.13	1.2E-06	0.09	1.6E-05
205551_at	*SV2B*	Synaptic Vesicle Glycoprotein 2b	AB028977	0.1	5.1E-02	0.2	4.9E-02	0.18	3.3E-03
232426_at	*SV2B*	Synaptic Vesicle Glycoprotein 2b	NM_014848	0.04	1.7E-02	0.1	4.2E-02	0.04	1.3E-02
Functional group 17						**P-value =**	**5.17E-07**		
207850_at	*CXCL3*	Chemokine (C-X-C Motif) Ligand 3	AL109781	0.2	2.8E-02	0.43	6.6E-02	0.2	1.9E-02
209774_x_at	*CXCL2*	Chemokine (C-X-C Motif) Ligand 2	NM_002090	0.49	1.3E-01	0.17	1.9E-02	0.49	1.2E-01
206336_at	*CXCL6*	Chemokine (C-X-C Motif) Ligand 6 (Granulocyte Chemotactic Protein 2)	M57731	0.31	1.3E-01	0.23	1.2E-02	0.22	9.8E-02
218002_s_at	*CXCL14*	Chemokine (C-X-C Motif) Ligand 14	NM_002993	0.38	1.9E-01	0.19	3.4E-02	1.11	9.0E-02
222484_s_at	*CXCL14*	Chemokine (C-X-C Motif) Ligand 14	NM_004887	0.04	4.4E-02	0.05	8.1E-05	0.09	5.1E-03
214974_x_at	*CXCL5*	Chemokine (C-X-C Motif) Ligand 5	AF144103	0.05	2.6E-02	0.05	2.2E-05	0.07	6.7E-03
205242_at	*CXCL13*	Chemokine (C-X-C Motif) Ligand 13 (B-Cell Chemoattractant)	AK026546	0.06	1.3E-01	0.12	3.3E-02	0.53	5.2E-02
205476_at	*CCL20*	Chemokine (C-C Motif) Ligand 20	NM_006419	0.86	2.2E-01	0.19	4.8E-02	0.72	1.6E-01
Functional group 18						**P-value =**	**2.40E-06**		
207261_at	*CNGA3*	Cyclic Nucleotide Gated Channel Alpha 3	NM_004591	0.17	1.8E-01	0.17	3.4E-02	0.73	1.8E-01
235781_at	*CACNA1B*	Calcium Channel, Voltage-Dependent, L Type, Alpha 1b Subunit	NM_001298	0.27	4.2E-02	0.13	8.4E-05	0.28	1.4E-02
234103_at	*KCNT2*	Potassium Channel, Subfamily T, Member 2	AA448208	0.11	2.5E-02	0.32	3.5E-02	0.13	7.7E-03
228581_at	*KCNJ10*	Potassium Inwardly-Rectifying Channel, Subfamily J, Member 10	AU145191	1.37	2.8E-01	2.03	1.6E-01	0.89	2.9E-01
207103_at	*KCND2*	Potassium Voltage-Gated Channel, Shal-Related Subfamily, Member 2	AW071744	0.02	1.3E-04	0.03	3.0E-06	0.02	2.0E-09

Affymetrix probeset	Gene symbol	Gene Description	Accesion number	Mg/Ag ratio	Mg/Ag q-value	Mg/Gb ratio	Mg/Gb q-value	Mg/Lgg ratio	Mg/Lgg q-value
1552742_at	*KCNH8*	Potassium Voltage-Gated Channel, Subfamily H (Eag-Related), Member 8	NM_012281	0.06	4.0E-04	0.07	2.3E-03	0.03	5.3E-06
227623_at	*CACNA2D1*	Calcium Channel, Voltage-Dependent, Alpha 2/Delta Subunit 1	NM_144633	0.14	5.6E-04	0.29	1.3E-03	0.12	5.4E-05
221107_at	*CHRNA9*	Cholinergic Receptor, Nicotinic, Alpha 9	H16409	0.29	3.5E-03	0.52	1.5E-02	0.27	5.2E-04
221584_s_at	*KCNMA1*	Potassium Large Conductance Calcium-Activated Channel, Subfamily M, Alpha Member 1	NM_017581	0.54	2.2E-01	0.18	9.2E-03	0.98	3.1E-01
1559419_at	*CACNB2*	Calcium Channel, Voltage-Dependent, Beta 2 Subunit	U11058	15.3	4.4E-04	17.09	1.6E-04	16.28	1.5E-04
205902_at	*KCNN3*	Potassium Intermediate/Small Conductance Calcium-Activated Channel, Subfamily N, Member 3	AL162054	2.1	6.2E-02	2.12	5.1E-02	2.14	4.2E-02
1555230_a_at	*KCNIP2*	Kv Channel Interacting Protein 2	AJ251016	0.15	3.9E-03	0.12	1.3E-04	0.08	5.9E-05
220116_at	*KCNN2*	Potassium Intermediate/Small Conductance Calcium-Activated Channel, Subfamily N, Member 2	AF367019	0.46	1.1E-02	0.88	1.2E-01	0.21	8.7E-03
220294_at	*KCNV1*	Potassium Channel, Subfamily V, Member 1	NM_021614	0.04	3.6E-03	0.1	1.9E-06	0.04	3.1E-05
210432_s_at	*SCN3A*	Sodium Channel, Voltage-Gated, Type Iii, Alpha	NM_014379	0.28	4.1E-02	0.63	1.1E-01	0.25	3.8E-02
210454_s_at	*KCNJ6*	Potassium Inwardly-Rectifying Channel, Subfamily J, Member 6	AF225986	0.01	3.9E-03	0.04	1.3E-04	0.01	3.5E-05
205737_at	*KCNQ2*	Potassium Voltage-Gated Channel, Kqt-Like Subfamily, Member 2	U24660	0.24	1.7E-02	0.47	4.1E-03	0.3	1.1E-02
210508_s_at	*KCNK2*	Potassium Channel, Subfamily K, Member 1	NM_004518	0.04	4.5E-03	0.11	1.1E-03	0.04	1.5E-05
204679_at	*KCNK1*	Potassium Voltage-Gated Channel, Shaker-Related Subfamily, Member 1 (Episodic Ataxia With Myokymia)	D82346	0.23	3.2E-03	0.19	6.5E-06	0.14	6.0E-04
230849_at	*KCNJ3*	Potassium Inwardly-Rectifying Channel, Subfamily J, Member 3	NM_002245	0.44	4.2E-02	0.91	3.1E-01	0.5	5.4E-02
233059_at	*SCN6A*	Sodium Channel, Voltage-Gated, Type Vii, Alpha	N64750	0.18	3.0E-02	0.35	2.7E-02	0.2	6.7E-03
228504_at	*SCN6A*	Sodium Channel, Voltage-Gated, Type Vii, Alpha	AK026384	0.12	1.7E-02	0.28	2.5E-02	0.15	1.1E-02
1555246_a_at	*SCN1A*	Sodium Channel, Voltage-Gated, Type I, Alpha	Al828648	0.52	2.3E-02	0.1	6.4E-02	0.31	1.3E-01
210383_at	*SCN1A*	Sodium Channel, Voltage-Gated, Type I, Alpha	AB093548	0.51	8.1E-04	0.58	1.6E-04	0.41	4.4E-05
206381_at	*SCN2A*	Sodium Channel, Voltage-Gated, Type Ii, Alpha 2	AF225985	0.04	1.0E-03	0.06	4.1E-07	0.04	1.2E-05
229057_at	*SCN2A*	Sodium Channel, Voltage-Gated, Type Ii, Alpha 2	NM_021007	0.16	7.1E-03	0.33	8.3E-03	0.11	1.4E-03

Functional group 19						P-value = 2.86E-06			
Affymetrix probeset	Gene symbol	Gene Description	Accesion number	Mg/Ag ratio	q-value	Mg/Gb ratio	q-value	Mg/Lgg ratio	q-value
204722_at	*SCN3B*	Sodium Channel, Voltage-Gated, Type Iii, Beta	BF432956	0.02	7.7E-03	0.06	7.0E-03	0.02	1.6E-03
204723_at	*SCN3B*	Sodium Channel, Voltage-Gated, Type Iii, Beta	AW007335	0.06	3.8E-03	0.19	1.5E-02	0.05	7.2E-04
219564_at	*KCNJ16*	Potassium Inwardly-Rectifying Channel, Subfamily J, Member 16	AB032984	0.03	4.2E-03	0.1	2.5E-02	0.02	9.3E-04
1552507_at	*KCNE4*	Potassium Voltage-Gated Channel, Isk-Related Family, Member 4	NM_018658	0.02	9.0E-04	0.02	8.6E-03	0.02	3.6E-04
1552508_at	*KCNE4*	Potassium Voltage-Gated Channel, Isk-Related Family, Member 4	NM_080671	3.01	1.4E-02	1.47	1.0E-01	3.28	6.9E-03
219714_s_at	*CACNA2D3*	Calcium Channel, Voltage-Dependent, Alpha 2/Delta 3 Subunit	NM_080671	3.7	6.0E-03	1.93	2.8E-02	4.19	2.8E-03
1554485_s_at	*TMEM37*	Transmembrane Protein 37	NM_018398	3.11	2.5E-03	6.51	3.1E-04	3.47	7.7E-04
206384_at	*CACNG3*	Calcium Channel, Voltage-Dependent, Gamma Subunit 3	BI825302	2.02	1.2E-01	2.65	3.0E-02	5.06	9.3E-03
237411_at	*ADAMTS6*	Adam Metallopeptidase With Thrombospondin Type 1 Motif, 6	NM_006539	0.13	3.3E-02	0.37	6.2E-02	0.15	2.0E-02
1556989_at	*ADAMTS9*	Adam Metallopeptidase With Thrombospondin Type 1 Motif, 9	N71063	0.15	2.9E-02	0.21	1.4E-06	0.14	7.9E-05
1554697_at	*ADAMTS9*	Adam Metallopeptidase With Thrombospondin Type 1 Motif, 9	AF086069	0.93	3.4E-01	0.41	8.3E-04	0.7	6.2E-02
1555326_a_at	*ADAM9*	Adam Metallopeptidase Domain 9 (Meltrin Gamma)	AF488803	0.8	9.6E-02	0.32	1.2E-02	0.63	4.1E-02
213790_at	*ADAM12*	Adam Metallopeptidase Domain 12 (Meltrin Alpha)	AF495383	1.11	3.1E-01	0.51	6.6E-03	0.93	3.0E-01
226777_at	*ADAM12*	Adam Metallopeptidase Domain 12 (Meltrin Alpha)	W46291	8.13	8.3E-03	3.1	1.9E-02	12.63	3.9E-03
206134_at	*ADAMDEC1*	Adam-Like, Decysin 1	AA147933	9.39	4.1E-03	2.39	2.1E-02	16.26	1.8E-03
1552727_s_at	*ADAMTS17*	Adam Metallopeptidase With Thrombospondin Type 1 Motif, 17	NM_014479	0.47	4.8E-02	0.28	3.8E-02	1.4	5.2E-02
Functional group 20						**P-value = 1.15E-05**			
1552507_at	*KCNE4*	Potassium Voltage-Gated Channel, Isk-Related Family, Member 4	AA022668	2.65	4.4E-02	2.67	3.6E-02	3.15	2.2E-02
1552508_at	*KCNE4*	Potassium Voltage-Gated Channel, Isk-Related Family, Member 4	NM_080671	3.01	1.4E-02	1.47	1.0E-01	3.28	6.9E-03
1554027_a_at	*SLC4A4*	Solute Carrier Family 4, Sodium Bicarbonate Cotransporter, Member 4	NM_080671	3.7	6.0E-03	1.93	2.8E-02	4.19	2.8E-03
203908_at	*SLC4A4*	Solute Carrier Family 4, Sodium Bicarbonate Cotransporter, Member 4	BC030977	7.05	1.9E-02	3.74	2.8E-02	4.25	1.9E-02
210040_at	*SLC12A5*	Solute Carrier Family 12, (Potassium-Chloride Transporter) Member 5	NM_003759	2.76	5.4E-03	2.41	6.3E-03	1.73	2.1E-02

Affymetrix probeset	Gene symbol	Gene Description	Accesion number	Mg/Ag ratio	Mg/Ag q-value	Mg/Gb ratio	Mg/Gb q-value	Mg/Lgg ratio	Mg/Lgg q-value
219090_at	*SLC24A3*	Solute Carrier Family 24 (Sodium/Potassium/Calcium Exchanger), Member 3	AF208159	0.04	2.3E-02	0.11	3.1E-02	0.04	1.1E-02
57588_at	*SLC24A3*	Solute Carrier Family 24 (Sodium/Potassium/Calcium Exchanger), Member 3	NM_020689	1.68	1.1E-01	3.3	1.4E-02	1.09	2.9E-01
239913_at	*SLC10A4*	Solute Carrier Family 10 (Sodium/Bile Acid Cotransporter Family), Member 4	R62432	1.73	8.8E-02	2.94	1.4E-02	1.15	2.4E-01
Functional group 21						**P-value =**	**1.17E-05**		
1554748_at	*CLCNKB*	Chloride Channel Kb	AI421796	0.22	1.6E-02	0.08	4.3E-02	0.28	1.6E-04
219529_at	*CLIC3*	Chloride Intracellular Channel 3	BC020873	3.64	5.4E-03	3.17	5.2E-03	3.97	2.7E-03
227742_at	*CLIC6*	Chloride Intracellular Channel 6	NM_004669	41.76	2.9E-02	40.82	2.4E-02	38.42	2.0E-02
213317_at	*CLIC5*	Chloride Intracellular Channel 5	AI638295	26.21	2.3E-03	28.92	1.3E-03	34.85	1.1E-03
205384_at	*FXYD1*	Fxyd Domain Containing Ion Transport Regulator 1 (Phospholemman)	AL049313	26.56	4.5E-02	29.25	3.7E-02	27.82	3.2E-02
Functional group 22						**P-value =**	**3.24E-04**		
239144_at	*B3GAT2*	Beta-1,3-Glucuronyltransferase 2 (Glucuronosyltransferase S)	NM_005031	3.22	1.1E-03	4.47	2.4E-04	3.57	3.4E-04
220979_s_at	*ST6GALNAC5*	St6 (Alpha-N-Acetyl-Neuraminyl-2,3-Beta-Galactosyl-1,3)-N-Acetylgalactosaminide Alpha-2,6-Sialyltransferase 5	AA835648	0.06	8.7E-03	0.1	1.9E-05	0.04	9.1E-05
230262_at	*ST8SIA3*	St8 Alpha-N-Acetyl-Neuraminide Alpha-2,8-Sialyltransferase 3	NM_030965	0.14	3.0E-02	0.21	1.1E-02	0.13	2.2E-02
1555123_at	*ST6GAL2*	St6 Beta-Galactosamide Alpha-2,6-Sialyltranferase 2	BF510762	0.09	7.5E-03	0.23	1.5E-02	0.09	4.0E-04
228821_at	*ST6GAL2*	St6 Beta-Galactosamide Alpha-2,6-Sialyltranferase 2	BC008680	0.43	3.7E-03	0.44	5.5E-03	0.26	5.3E-03
1559617_at	*GAL3ST2*	Galactose-3-O-Sulfotransferase 2	AW004016	0.08	1.2E-02	0.09	1.1E-03	0.04	5.4E-04
234472_at	*GALNT13*	UDP-N-Acetyl-Alpha-D-Galactosamine:Polypeptide N-Acetylgalactosaminyltransferase 13 (Galnac-T13)	D55640	0.29	7.3E-02	0.38	3.2E-02	0.13	4.2E-02
228501_at	*GALNTL2*	UDP-N-Acetyl-Alpha-D-Galactosamine:Polypeptide N-Acetylgalactosaminyltransferase-Like 2	AC009227	0.07	4.3E-03	0.18	1.4E-03	0.06	1.6E-04
236361_at	*GALNTL2*	UDP-N-Acetyl-Alpha-D-Galactosamine:Polypeptide N-Acetylgalactosaminyltransferase-Like 2	BF055343	0.12	9.1E-03	0.07	2.2E-05	0.1	1.1E-04
1553959_a_at	*B3GALT6*	UDP-Gal:Betagal Beta 1,3-Galactosyltransferase Polypeptide 6	BF432376	0.24	1.7E-02	0.13	3.8E-05	0.21	2.7E-04
207447_s_at	*MGAT4C*	Mannosyl (Alpha-1,3-)-Glycoprotein Beta-1,4-N-Acetylglucosaminyltransferase, Isozyme C (Putative)	N95564	0.61	9.6E-03	0.46	6.7E-08	0.65	1.9E-03
1559814_at	*CHSY3*	Chondroitin Sulfate Synthase 3	NM_013244	0.14	2.4E-02	0.51	1.3E-02	0.09	1.6E-03

Functional group 23						P-value = 1.35E-02			
Affymetrix probeset	Gene symbol	Gene Description	Accesion number	Mg/Ag ratio	q-value	Mg/Gb ratio	q-value	Mg/Lgg ratio	q-value
208161_s_at	*ABCC3*	ATP-Binding Cassette, Sub-Family C (Cftr/Mrp), Member 3	AK024712	5.64	5.6E-02	6.46	4.4E-02	7.28	3.5E-02
1553605_a_at	*ABCA13*	ATP-Binding Cassette, Sub-Family A (Abc1), Member 13	NM_020037	0.24	8.6E-02	0.14	8.1E-05	1.14	2.0E-01
204719_at	*ABCA8*	ATP-Binding Cassette, Sub-Family A (Abc1), Member 8	NM_152701	0.24	1.6E-01	0.41	4.2E-02	0.84	3.8E-02
213106_at	*ATP8A1*	ATPase, Aminophospholipid Transporter (Aplt), Class I, Type 8a, Member 1	NM_007168	2.07	3.6E-02	1.36	1.4E-01	1.86	3.9E-02
219659_at	*ATP8A2*	ATPase, Aminophospholipid Transporter-Like, Class I, Type 8a, Member 2	AI769688	0.21	2.2E-03	0.49	4.5E-02	0.19	6.8E-06
1557136_at	*ATP13A4*	ATPase Type 13a4	AU146927	0.21	2.4E-02	0.5	8.2E-02	0.26	7.0E-03
1559571_a_at	*ATP13A4*	ATPase Type 13a4	BG059633	0.37	2.9E-04	0.25	1.8E-04	0.24	9.5E-05
Functional group 24						**P-value = 1.71E-02**			
204337_at	*RGS4*	Regulator of G-Protein Signalling 4	AK095277	0.71	4.5E-03	0.56	4.1E-04	0.47	5.7E-04
204338_s_at	*RGS4*	Regulator of G-Protein Signalling 4	AL514445	0.05	2.4E-02	0.08	7.4E-03	0.04	1.5E-02
204339_s_at	*RGS4*	Regulator of G-Protein Signalling 4	NM_005613	0.09	4.3E-02	0.12	1.6E-02	0.06	2.4E-02
1554500_a_at	*RGS7*	Regulator of G-Protein Signalling 7	BC000737	0.11	1.8E-02	0.15	1.1E-02	0.09	1.8E-02
206290_s_at	*RGS7*	Regulator of G-Protein Signalling 7	AF493931	0.22	1.2E-02	0.44	3.2E-02	0.18	5.4E-03
202988_s_at	*RGS1*	Regulator of G-Protein Signalling 1	NM_002924	0.06	1.3E-02	0.15	1.4E-02	0.05	1.2E-03
216834_at	*RGS1*	Regulator of G-Protein Signalling 1	NM_002922	1.22	3.3E-01	1.19	2.8E-01	4.47	4.4E-02
210138_at	*RGS20*	Regulator of G-Protein Signalling 20	S59049	1.41	2.3E-01	1.05	3.3E-01	3.12	2.6E-02
1554643_at	*RGS11*	Regulator of G-Protein Signalling 11	AF074979	0.08	1.3E-02	0.14	1.4E-03	0.09	5.3E-04
223809_at	*RGS18*	Regulator of G-Protein Signalling 18	BC040504	1.33	2.6E-01	2.56	5.8E-02	1.71	1.1E-01
1555725_a_at	*RGS5*	Regulator of G-Protein Signalling 5	AF076642	2.84	2.7E-02	1.52	1.4E-01	1.47	1.7E-01
Functional group 25						**P-value = 2.28E-02**			
219671_at	*HPCAL4*	Hippocalcin Like 4	AF493929	2.38	2.8E-03	1.92	5.6E-03	1.8	7.3E-03
209904_at	*TNNC1*	Troponin C Type 1 (Slow)	AL136591	0.07	2.8E-02	0.18	2.9E-02	0.06	4.4E-03
203797_at	*VSNL1*	Visinin-Like 1	AF020769	128.45	6.3E-02	117.75	5.3E-02	123.2	4.6E-02
203798_s_at	*VSNL1*	Visinin-Like 1	AF039555	0.02	1.9E-02	0.04	2.1E-02	0.02	8.9E-03
208320_at	*CABP1*	Calcium Binding Protein 1 (Calbrain)	NM_003385	0.02	2.4E-02	0.04	3.7E-02	0.01	2.3E-02
208321_s_at	*CABP1*	Calcium Binding Protein 1 (Calbrain)	NM_004276	0.16	2.9E-02	0.34	3.2E-02	0.19	1.0E-02
210181_s_at	*CABP1*	Calcium Binding Protein 1 (Calbrain)	NM_004276	0.23	2.2E-02	0.49	4.7E-02	0.22	8.7E-03
223885_at	*CALN1*	Calneuron 1	AF169148	0.26	3.1E-02	0.44	7.9E-02	0.21	2.2E-02
230698_at	*CALN1*	Calneuron 1	AF282250	0.12	1.0E-02	0.5	7.5E-02	0.13	1.2E-02

A-11 Expression values of the 100 probesets that determines two Gb clusters

			Intensity signals of fluorescence				
Gene symbol	Probeset	Cluster 1 / Cluster 2	Females Cluster 1	Males Cluster 1	Females Cluster 2	Males Cluster 2	q-value
XIST	224588_at	0.69	3459.2	12.6	2388.7	13.1	7.7E-02
XIST	227671_at	0.56	1429.1	9.7	1250.8	9.8	7.0E-02
XIST	221728_x_at	0.75	1576.9	14.0	1020.5	11.9	2.9E-01
H19	224646_x_at	1.13	452.6	625.6	775.7	332.6	1.3E-01
POSTN	1555778_a_at	0.16	18.6	160.9	956.9	689.8	1.6E-03
POSTN	210809_s_at	0.25	71.6	617.9	2298.0	1794.9	1.4E-03
IGHG3	211430_s_at	0.09	13.8	75.9	324.6	689.3	8.6E-05
IGLJ3	214677_x_at	0.08	26.6	53.6	453.7	472.3	5.4E-04
LTF	202018_s_at	0.03	19.2	76.1	685.0	3122.8	3.6E-04
CHI3L2	213060_s_at	0.12	41.5	166.1	887.6	1420.4	2.4E-03
CCL2	216598_s_at	0.11	42.8	147.8	1296.2	1065.8	5.1E-05
PLA2G2A	203649_s_at	0.04	32.8	33.6	457.2	1081.9	3.4E-04
IL8	202859_x_at	0.06	26.5	54.5	1159.8	528.7	6.0E-05
SERPINE1	202628_s_at	0.07	18.2	61.9	899.1	599.8	8.1E-05
SERPINE1	202627_s_at	0.14	28.3	115.7	805.6	634.9	1.0E-04
EGFR	201984_s_at	0.98	112.4	1022.0	786.7	928.2	2.8E-01
EGFR	232541_at	1.12	642.5	1403.9	1002.2	1259.7	1.4E-01
EGFR	224999_at	0.93	614.6	2141.7	1841.5	2140.3	2.0E-01
EGFR	201983_s_at	0.93	1213.5	3861.8	3312.3	3912.9	2.4E-01
FOXG1	206018_at	0.80	951.3	860.3	1149.0	1132.6	7.2E-02
HOP	211597_s_at	0.75	503.6	1313.1	1782.7	1379.6	1.4E-01
FABP7	205029_s_at	0.51	748.6	974.2	1887.2	1867.4	2.2E-02
PLP1	210198_s_at	1.67	1709.9	2810.0	1853.4	1366.9	1.4E-02
GRIA2	205358_at	2.06	2828.0	1192.0	824.2	744.8	1.2E-02
OLIG1	228170_at	1.93	2915.8	1707.0	1193.9	945.3	2.7E-03
AQP1	207542_s_at	0.94	449.3	586.8	614.9	601.6	2.8E-01
AQP1	209047_at	1.04	1278.7	1837.5	1531.4	1801.8	2.6E-01
ATP1A2	203295_s_at	1.52	444.3	1054.5	638.6	611.6	3.2E-02
ATP1A2	203296_s_at	1.76	1152.9	2153.8	1193.2	1080.8	1.1E-02
RPS4Y1	201909_at	1.57	19.0	1635.1	31.9	1543.5	1.7E-01
NA	213841_at	4.72	2209.6	927.8	181.5	339.3	1.9E-04
SNAP25	202508_s_at	4.30	1001.0	1658.7	423.2	309.5	9.2E-04
PEG10	212094_at	1.01	723.1	466.4	622.8	421.8	1.9E-01
APOD	201525_at	0.47	203.4	504.4	833.4	1081.6	2.3E-01
PDGFRA	203131_at	1.40	1099.9	710.4	593.4	581.4	1.9E-01
CD24	216379_x_at	2.22	1578.6	662.9	342.7	433.5	1.1E-01
CD24	209771_x_at	2.07	1656.9	688.1	385.2	487.4	1.1E-01
SOX11	204914_s_at	1.70	1150.0	743.5	470.1	530.6	1.5E-01
NCAN	205143_at	1.38	957.9	789.0	580.1	653.5	7.2E-02
BCAN	219107_at	1.87	982.2	792.0	385.5	526.5	1.1E-02
LOC650392	1569872_a_at	2.81	2431.1	857.4	426.4	458.2	3.3E-03
SOX8	226913_s_at	3.52	2859.2	1097.1	358.9	505.0	4.4E-04
MBP	207323_s_at	1.83	407.8	1279.7	821.3	435.3	2.6E-02
MBP	209072_at	2.38	1459.8	3750.0	1694.7	1159.9	1.2E-02
GPR37	209631_s_at	0.78	431.0	398.2	606.9	466.4	1.3E-01
EDIL3	225275_at	1.01	262.8	866.5	972.7	549.3	3.1E-01
RTN1	203485_at	3.03	1413.8	1445.4	508.0	461.1	3.0E-03
KIF5A	229921_at	3.19	1926.2	1137.5	595.3	270.5	1.1E-02
FAM123A	235465_at	2.50	742.7	807.1	371.9	283.3	1.2E-03
FAM123A	230496_at	2.79	1006.0	1126.6	456.4	360.3	6.9E-04
CLDN11	228335_at	1.01	644.1	656.2	872.4	495.8	2.4E-01
TF	214063_s_at	1.37	533.4	1012.7	937.1	468.1	2.5E-02
TF	203400_s_at	1.35	682.9	1281.6	1216.5	588.7	3.3E-02
MGST1	231736_x_at	0.51	243.9	565.4	796.1	1049.1	6.0E-02

Gene symbol	Probeset	Cluster 1 / Cluster 2	Intensity signals of fluorescence: Females Cluster 1	Males Cluster 1	Females Cluster 2	Males Cluster 2	q-value
PTGDS	212187_x_at	1.74	386.6	1682.0	927.5	719.4	1.2E-01
PTGDS	211748_x_at	1.80	556.2	2138.9	1112.7	921.6	8.4E-02
NA	AFFX-M27830_5_at	1.81	1703.3	3692.3	1912.8	1643.6	1.4E-01
SERPINA3	202376_at	0.44	340.0	1669.9	2893.5	3522.2	2.3E-03
CHI3L1	209396_s_at	0.21	1298.2	1388.0	6584.6	7120.0	6.2E-05
CHI3L1	209395_at	0.20	1102.3	1503.7	7697.3	6678.4	5.0E-05
COL1A2	202403_s_at	0.17	454.8	538.2	3368.7	2821.1	6.1E-04
COL1A1	1556499_s_at	0.19	530.0	700.8	3320.1	3501.3	2.4E-04
COL1A2	202404_s_at	0.16	372.6	430.8	2895.7	2447.0	3.6E-04
COL3A1	215076_s_at	0.13	310.3	402.1	3102.0	3039.7	1.6E-04
IGFBP3	212143_s_at	0.22	70.9	250.5	1082.6	878.2	1.0E-03
IGFBP3	210095_s_at	0.24	172.8	627.9	2214.2	2238.6	1.2E-03
LPL	203548_s_at	0.39	272.3	409.3	1256.2	763.6	5.1E-03
IGFBP2	202718_at	0.49	399.0	629.0	1470.4	928.7	1.2E-02
VEGFA	212171_x_at	0.35	214.1	393.8	1272.5	784.4	2.3E-03
ADM	202912_at	0.23	152.0	374.0	1883.9	1078.2	2.3E-04
CTHRC1	225681_at	0.27	290.9	144.5	714.8	624.9	1.6E-03
COL6A3	201438_at	0.06	40.7	60.5	1108.8	920.7	7.0E-05
COL1A1	202310_s_at	0.10	106.4	144.9	1610.3	1293.2	1.8E-04
COL3A1	211161_s_at	0.11	105.7	186.2	1615.1	1549.8	1.7E-04
COL3A1	201852_x_at	0.09	143.5	191.2	2203.5	1923.2	1.5E-04
NA	AFFX-HUMRGE/M10098_5_at	0.63	74.7	559.3	598.6	747.9	4.1E-02
METTL7B	227055_at	0.34	58.6	261.8	867.9	454.7	4.8E-04
SOCS2	203373_at	0.61	159.6	490.2	828.5	563.2	1.6E-02
MGP	202291_s_at	0.14	178.5	260.3	2256.7	1296.2	9.1E-05
DCN	211896_s_at	0.19	96.8	195.0	956.4	920.7	5.8E-04
DCN	211813_x_at	0.23	153.0	297.8	1221.5	1135.5	1.4E-03
CAV1	212097_at	0.21	131.0	201.6	881.6	951.0	8.1E-05
TAGLN	205547_s_at	0.23	309.8	225.1	965.2	1132.6	1.1E-04
TGFBI	201506_at	0.18	248.5	226.2	1360.6	1196.9	4.9E-05
COL5A2	221729_at	0.25	196.2	225.4	978.9	832.6	1.3E-04
TMEM49	224917_at	0.16	152.1	186.9	1224.7	1154.3	4.9E-05
HIG2	1554452_a_at	0.22	119.6	160.8	768.4	596.3	1.5E-03
IGKC	221671_x_at	0.20	113.8	201.9	606.2	940.1	1.3E-03
IGKC	221651_x_at	0.19	112.7	207.8	650.2	1025.9	8.7E-04
IGKC	224795_x_at	0.18	114.9	216.0	704.1	1094.5	4.6E-04
S100A8	202917_s_at	0.12	84.0	163.1	893.3	1475.0	7.4E-05
TncRNA	227062_at	0.27	111.1	252.5	960.5	689.9	3.4E-04
NA	225328_at	0.37	161.0	230.4	459.9	725.7	2.1E-03
MGC5618	221477_s_at	0.27	172.9	318.9	1199.4	955.1	1.2E-03
NNMT	202237_at	0.11	162.0	133.8	978.6	1530.8	4.9E-05
C1S	208747_s_at	0.24	158.3	256.5	806.3	1168.5	1.7E-04
GBP1	231577_s_at	0.31	145.2	414.5	1044.8	1223.2	2.6E-04
PBEF1	243296_at	0.13	139.2	223.6	1329.9	1813.1	4.9E-05
CD163	203645_s_at	0.15	90.0	184.3	1060.0	1109.3	4.9E-05
CD163	215049_x_at	0.18	124.6	265.4	1277.1	1378.8	6.6E-05

Scientific communications

- Accepted article in press: "Automated Brain Tumour Prediction". Castells X, García-Gómez JM, Navarro A, Acebes JJ, Godino O, Boluda S, Robles M, Barceló A, Ariño J and Arús C. *Diagnostic Molecular Pathology*, 2008.

- Poster communication: "Objective prediction of human brain tumour types using gene-expression microarrays-based profiling". Castells X, García-Gómez JM, Navarro AT, Acebes JJ, Boluda S, Robles M, Barceló A, Ariño J and Arús C. *Molecular Diagnostics Europe* (Stockholm, Sweden, 19-20 May 2009). Awarded as the best poster presented in the congress.

- Poster and oral communication: "An exploratory analysis of human brain tumour prediction based on DNA microarray profiling". Castells X, Acebes JJ, Godino O, Boluda S, García-Gómez JM, Navarro A, Robles M, Barceló A, Ariño J and Arús C. *Towards Brain Tumour Classification by Molecular Profiling: Imaging, Metabolomic and Genomic tools* (València, Spain, October 1st 2008).

- Poster communication: "Inter-laboratory control study of micro-array data within the EC supported eTUMOUR project". Postma GJ, Boots-Sprenger S, Barcelo A, Castells X, Ferrer R, Rogers H, Grundy R, Celda B, Wesseling P, Ariño J and Buydens LMC. *Towards Brain Tumour Classification by Molecular Profiling: Imaging, Metabolomic and Genomic tools* (València, Spain, October 1st 2008).

- Poster communication: "Automated Classification of Human Brain Tumours and Correlation with Survival using DNA Microarray-based Gene Expression Profiling". Castells X, García-Gómez JM, Gevaert O, Navarro A, Acebes JJ, Godino O, De Moor B, Barceló A, Ariño J, Robles M and Arús C. *XXX Congreso de la Sociedad Española de Bioquímica y Biología Molecular* (Málaga, Spain, 12-15 September 2007).

- Poster communication: "Genomics and metabolomics research for Brain Tumour diagnosis based on Machine Learning". García-Gómez JM, Tortajada S, Vicente J, Sáez C , Castells X, Luts J, Julià-Sapé M, Juan-Císcar A, Van Huffel S, Barceló A, Ariño J, Arús C and Robles M. *9th International Work-Conference on Artificial Neural Networks. Lecture notes in computer sciences.* (Donostia, Spain, 20-22 June 2007).

- Poster communication: "Microarray Gene Expression of a Murine Transgenic Model of the Bovine Spongiform Encephalopathy". Tortosa R, Costa C, Vidal E, Castells X, Barceló A, Torres JM, Ariño J and Pumarola M. *Prion 2007* (Edinburgh, Scotland, UK , 26-27 September 2007).

- Poster communication: "Las metalotioneínas y la encepalopatía espongiforme bovina. Estudio de su expression génica en un modelo murino". Tortosa R, Costa C, Domènech A, Vidal E, Castells X, Barceló A, Ariño J and Pumarola M. *XVIII Reunión de la Sociedad Española de Anatomía Patológica Veterinaria* (Rabat, Morocco, 28-30 June 2006).

- Poster communication: "Estudio de expression génica en un modelo experimental murino de la encefalopatía espongiforme bovina". Tortosa R, Castells X, Márquez M, Ariño J, Pumarola M and Barceló A. *XVII Reunión de la Sociedad Española de Anatomía Patológica Veterinaria* (Jarandilla de la Vera, Spain, 14-16 June 2005).

Acknowledgements

En primer lloc, agraeixo als meus directors de tesi haver-me iniciat en el món de la recerca. De cadascú d'ells he pogut aprendre diferents maneres de veure i abordar els problemes científics.

Agraeixo al Dr. Carles Arús haver-me finançat la tesi a través del projecte europeu eTUMOUR. Voldria agrair-li especialment que malgrat no haver gaudit d'una beca de la Generalitat o de l'Estat, m'he sentit finançat com si ho estigués. En aquest sentit, voldria agrair-li haver-me donat l'oportunitat de realitzar una estada de recerca a la "Katholieke Universitet Leuven".

Agraeixo al Dr. Joaquín Ariño haver-me ensenyat a ser molt rigorós en el món de la recerca científica. Ara que he acabat la tesi i puc mirar els 5 anys de doctorat amb perspectiva, voldria destacar que aquest rigor après segurament és fruit de la perseverança i intensitat amb les que analitza tot allò que té entre mans.

Agraeixo a la Dra. Anna Barceló el seguiment quasi diari que ha fet de la meva feina experimental. Havent estat la meva tesi duta terme a cavall de dos grups de recerca, la Dra. Barceló ha estat la persona clau per a debatre tots mena de dubtes al llarg de la tesi i coordinar el treball experimental amb la resta de directors.

Agraeixo als projectes eTUMOUR, HealthAgents i MEDIVO2 per haver finançat els experiments i la meva tesi.

Agraeixo al Dr. Juan José Acebes haver recollit les biòpsies amb les quals he pogut realitzar la part més important de la meva tesi doctoral. En aquest sentit, també agraeixo a tot el seu equip de neurocirurgians i infermeres.

Agraeixo a la Dra. Susana Boluda haver facilitat els diagnòstics histopatològics de les biòpsies recollides, així com haver compartit la seva experiència diagnosticant tumors de cervell.

Agraeixo al Dr. Àlex Sánchex pels seus consells sobre el processament de les dades d'Affymetrix i el desenvolupament de models de predicció.

I thank Dr. Pieter Wesseling for his valuous suggestions about how to deal with

the available diagnoses within eTUMOUR.

Agraeixo al Dr. Martí Pumarola per haver dedicat temps a explicar-me les bases de l'histologia i per haver-les il.lustrat en una sessió pràctica amb talls histològics.

Al Dr. Antoni Ramis per haver-me acceptat com oient de l'assignatura d'Anatomia Patològica de Veterinària.

Agraeixo a tots els membres del GABRMN la seva constant col.laboració. A la Dra. Ana Paula Candiota per ajudar-me al llarg de tota la tesi i per interessar-se pel treball que anava realitzant. Voldria destacar que la Dra. Candiota és un exemple de força de voluntat i d'autosuperació sigui quin sigui l'obstacle que tingui al davant. A la Dra. Margarida Julià-Sapé per haver estat una coordinadora eficaç entre tots els involucrats a eTUMOUR a la UAB. Al Dr. Daniel Valverde per haver-me iniciat en l'aprenentatge de LaTeX i per haver-me facilitat les funcions i les dades d'HRMAS. A la Dra. Maria Rosa Quintero per estar sempre disposada a donar un cop de mà en allò que calgués. A la Indira Coronel per haver-me guiat en les primeres recollides de biòpsies a Bellvitge. A en Rui Simoes, la Teresa Delgado i la Milena Acosta per haver induït els tumors de cervell als ratolins utilitzats per a la meva tesi. A la Juana Martín per la seva ajuda amb les cèl.lules C6. A la María Ángeles Verdugo per gestionar amb el Dr. Pumarola la detecció de necrosi en els tumors induïts a ratolins. A l'Álex Pérez per comptar amb el seu suport en tasques computacionals. Al Miguel Lurgui per la seva ajuda amb LaTeX i per resoldre amb celeritat qualsevol problema amb l'eTDB. A l'Alina García per la seva ajuda amb la darrera compilació de LaTeX de la tesi. A l'Iván Olier per les seves recomanacions sobre problemes matemàtics. A l'Óscar Tibaduisa i l'Edgar Villarraga per estar disposats a donar un cop de mà en qualsevol moment. A la Silvia Ferrer, la Myriam Dávila i la Miriam Camison per la gestió dels casos de les dades clíniques de les biòpsies recollides. Tot i no ser del GABRMN, a la Montse Gòdia per la seva gestió eficaç durant la meva tesi.

Agraeixo a tota la gent del grup de llevats de Veterinària la seva ajuda. Com ells han estat els que han conviscut el dia a dia al laboratori, espero no haver-los donat molt la tabarra. A la Dra. Laia Viladevall que ha estat el meu exemple a seguir a l'hora de dur a terme experiments de microarrays de cDNA. A l'Ivan Muñoz, l'Amparo Ruiz, la Raquel Serrano per ser l'exemple de lluita constant en el món de la recerca. Al Dr. Casamayor per compartir els seus extensos coneixements cada cop que li he demanat consell. Trobar gent tan pencaire quan comences la tesi fa agafar uns hàbits exemplars per tot allò que vulguis fer. A la Maria Platara per compartir similars males experiències al llarg de la tesi, però que crec que a

tots dos ens ha enfortit. A en Martí Boleda perquè és un gran amic i per les converses sobre el sexe dels àngels que anem tenint de tant en tant. A la Maribel Marquina per haver-me deixat el seu vórtex i la seva P20 quan m'han calgut. A en Jofre Dalmau-Ferrer perquè sempre està pendent que no manqui l'avituallament sòlid durant els experiments de fons que fem a Veterinària. A la Lina Barreto per deixar de fer el que que estigués fent quan li demanes sobre qualsevol cosa. A l'Anna Vilalta, la Montse Robledo i la Maria Jsús Álvarez per ser unes tècniques de laboratori increïblement generoses i trempades. A la Rosa Ramos per preocupar-se amb afany de tot els afers administratius que m'afectaven. En extensió a la resta de secretàries que han passat per Veterinària: la Sílvia, la Glòria, la Mercè i l'Elena.

I thank Dr. Bart De Moor, Dr. Yves Moreau and Dr. Olivier Gevaert for accepting me in the BIOI group from the KUL during a stage of 3 months. Also, the rest of people of the BIOI from which I learned how to deal with computational problems: Anneleen Daemen, Tim Van den Bulcke, Thomas D'Hollander, Wout Van Delm, Raf Van de Plas, Fabián Ojeda, Dr. Kristof Engelen, Dr. Karen Lemmens, Steven Van Vooren, Riet De Smet and Niels Haverbeke.

De la unitat Bioquímica de Veterinària, voldria agrair al Dani, la Clelia, la Laia i la Maria José del grup de la Dra. Anna Bassols perquè amb ells hi he pogut aprendre coneixements sobre el càncer i intercanviar-ne. Al Dr. Néstor Gómez i a en Josep per les xerrrades transcendentals pel futur del país que hem anat tenint i que tindrem.

Seguint a Veterinària, voldria agrair a en Raül Tortosa d'Anatomia Patològica per l'inici plegats en el moment del RNA i dels microarrays i a l'Anna Domènech per poder posar a punt l'extracció de RNA de cervell en ratolins. Així mateix, agraeixo a la Dra. Carme Costa, el Dr. Enric Vidal i la Dra. Anna Serafín per la seva col.laboració.

Baixant cap a la facultat de Biociències, agraeixo a tota la gent de la unitat de Bioquímica l'ambient engrescador i divertit que s'hi respira a tota hora. En particular voldria agrair al Dr. Suau, la Dra. Inma Ponte, la Dra. Alicia Roque i a en Markus Richter per haver-me iniciat en la recerca durant la llicenciatura de Bioquímica. També agrair a l'Helena que prepara un DMEM que fa créixer les cèl.lules només ensenyant-te'ls-hi i a en Salva per la seva sabiesa i paciència en les diverses impressions de la tesi.

Je veux aussi remercier à mon ami chinois Fei Chei, avec qui ont a commencé le doctorat à Strasbourg. Il est un mec génial qui s'intéresse toujours à moi. Je profite de ces lignes pour m'excuser de mes frequents retards à lui répondre. Dans

tout cas, il est pour moi une personne qui sera toujours spécial.

Modu bereziaz, Asierri eskertzen diot bere emandako laguntza. Egia esan, ezta bakarrik bere laguntzarengatik behar nuen une guztietan, baina dagoen lekuan zabaltzen duen bere indarrarengatik. Balio haundia da nirea horrelako lagun bat edukitzeagatik. Gauza pilo bat ikasi dut berarengandik eta garrantzitsuena harnas galdu arte nire herriaren alde borrokatzea.

Poc voldria oblidar-me d'agrair a mons pares haver-me educat. La dificultat que ha suposat aquesta tesi la conec sobretot jo, però segur que no hagués resistit sense que ells m'haguéssin fet dur com un roc del Querol tant per dins, com per fora. També a mon germà que durant la recta final de la tesi va fer de bon germà i em tenia el sopar preparat quan arribava a hores intempestives.

Agraeixo moltíssim a la Mar que quan em mira encara li brillin els ulls com els diamants més purs del món. Sense la seva comprensió i ajuda moral diària aquest darrer any, vés a saber on pararia jo ara.

Un agraïment especial i possiblement estrambòtic és per les meves vaques (baja, de mons pares). Sense elles i sense la cals de la llet amb la que he crescut, no seria la meitat de valent que sóc ara. Fora bromes, les meves vaques són una metàfora (per mi una realitat) del vincle entre el treball i la terra amb la qual t'identifiques. Si cada persona del món pugués viure content amb la feina que fa i aquesta es traduís a la vegada en la conservació d'una terra, possiblement seríem una societat equilibrada. Sóna utópic, però crec sincerament que només cal un petit pas agosarat per aconseguir-ho.

En aquest sentit però força diferent, vull agrair a tota la gent del món que d'una manera totalment altruista desenvolupa software lliure per a que pugui ser utilitzat arreu. Aquest és un dels exemples de justícia més gran que segons el meu parer han sorgit de la societat del coneixement. Sense ells, la meva tesi hagués estat molt més cara i possiblement no hagués arribat a plantejar fer el salt de disciplina i intentar caminar en el món de l'estadística i la programació informàtica.

Software lliure, societat lliure, visca la Terra!