Universitat Autònoma de Barcelona - Facultat de Medicina

Doctorat en Salut Pública i Mètodes en Recerca Biomèdica

Mètode Bayesià per a l'anàlisi d'Haplotips en estudis d'Associació Genètica

Aplicació a dades d'Esquizofrènia i Càncer

TESI DOCTORAL REALITZADA PER **RAQUEL INIESTA BENEDICTO** SOTA LA DIRECCIÓ DEL

DR. VÍCTOR MORENO AGUADO I LES TUTORIES DEL **DR. FERRAN TORRES BENÍTEZ**

DEPARTAMENT DE PEDIATRIA, OBSTETRÍCIA, GINECOLOGIA I MEDICINA PREVENTIVA

INSTITUT CATALÀ D'ONCOLOGIA - PARC SANITARI SANT JOAN DE DÉU

BARCELONA, 2010

Als meus pares

Agraïments

Per això malgrat la boira cal...

...caminar.

En aquest espai vull tenir un record per a cadascú de vosaltres, els qui d'una manera o altra, en un o altre moment del camí, ja hagi estat amb el coneixement o simplement amb el cor, heu posat el vostre granet de sorra en aquesta tesi.

Tengo una debilidad...

Aquest treball va dedicat als meus pares, perquè és en gran part gràcies al seu esforç que jo he pogut arribar fins aquí. El meu més gran agraïment per haver lluitat per a què tant jo com els meus germans haguem pogut estudiar fins allà on hem triat i per la confiança que sempre he sentit que tenen en mi. Gracias a los dos, mame.

Cau es sol de s'horabaixa dins s'horitzó...

Gràcies al Xavier Bosch, per decidir fer footing pel carrer on jo em trobava fent ràdio i després haver confiat en mi durant tant de temps. També vull agrair al meu director de tesi, el Víctor Moreno, l'haver-me ofert triar entre coses fàcils i difícils el meu primer dia de feina, i haver-me guiat entre les difícils. Gràcies també al Ferran Torres per haver estat el tutor d'aquest treball. Gràcies als meus companys de l'ICO, Esther, Toni, David, Oscar i a la resta del servei, per fer que el record d'aquesta etapa sigui un somriure.

Look at the stars, look how they shine for you...

Gràcies infinites, molt especialment, al Xavi Solé: has viscut aquesta tesi literalment amb mi des del primer dia, amb els meus problemes amb els mallocs que resolíem entre riures i fins al final, amb les simulacions. Has estat allà sempre que t'he necessitat, m'has ajudat moltíssim, m'has animat, m'has fet costat, m'has aconsellat...em quedaria curta posant exemples. Simplement no existeixen paraules per a que jo pugui agrair el teu suport, que ha anat molt més enllà de la informàtica. Gràcies, de tot cor.

...paraules que no s'esborren, imatges que no se'n van.

Gràcies al Josep Maria Haro per recordar-me en tot moment des de la meva arribada a Sant Joan de Déu que tot i haver canviat de feina la meva prioritat havia de ser la tesi. Gràcies també a la Susana Ochoa per interessar-se en els meus avenços i animar-me sempre a seguir endavant amb un somriure. Gràcies als meus companys de la unitat de recerca: Aidi, mil gràcies floreta per les converses i pel teu "fot-li canyaaaaa" diari que m'ha omplert d'energia. Gracias también Pipi por los ánimos, la compañía y las xarlas, eres una mina recopada! Ferranet, acaba el PIR que repetim mojitos ;-) i també gràcies a la resta dels meus companys pel recolzament i els consells que m'heu donat: Christian, Iris, Maria, Ana, Judith, Victoria V., Raquel L, Jaume A, Jordan, Bea, Iria, Elena H, Lluís, Elena R.,... El fet de poder conviure amb alguns de vosaltres el procés d'acabar una tesi ha estat molt reconfortant.

I wish I was a fisherman...

Gràcies a les meves nenes, Olga i Gemma, per fer-me costat. Gràcies també Domi, per transmetre'm la teva confiança i il·lusió. Milions de gràcies al Jaumini, l'Edgarini, el Jordi, i a les meves precioses Lau i Ire per haver convertit el Bon Rotllo en amistat. Durant aquests anys m'heu permès fer un *pit-stop* bàsic per mi cada finde.

Com t'ho podria dir perquè em fos senzill...

Les paraules matemàtiques, LateX i Linux no tindrien sentit sense tu. Fa tretze anys vam començar plegats aquesta aventura, vam acabar la carrera gràcies l'un a l'altre i jo ara acabo el recorregut que tu acabaràs en breu. El teu ajut moral i "talibàn" en tot això ha estat per

mi imprescindible, de fet deus ser la persona del planeta que més presentacions sobre haplotips ha vist ;-) T'ho agraeixo moltíssim David.

It started with a low light...

Tu si que no sabies on et ficaves amb tants "haplotypes" :-) Gràcies Marc per la teva "energyyy" constant i infinita, que has estat capaç de transmetre'm amb tanta força i que per mi ha estat tan important durant l'etapa final d'aquesta feina. Gràcies per compartir-ho amb mi, per les teves ganes, per la il·lusió i la passió que poses en tot i que m'encomanes dia rera dia. I sobretot, gràcies per sumar.

And You know I'm fine but I hear those voices at night...sometimes...they justify my claim!!

Per acabar, s'endú el meu més profund agraïment el Dr Gasulla: la feina que durant aquests anys hem fet plegats, tot el que hem construït i que va molt més enllà d'aquesta tesi, ha estat per mi indispensable per arribar fins aquí. Moltíssimes gràcies pel teu ajut.

El camí més curt no és sempre el més recte.

El camí amb més gent no sempre és el correcte.

Roger Mas

Pròleg

Els avenços que a les darreres dècades han protagonitzat les tècniques de genotipatge i de seqüenciació, unit al desenvolupament de tècniques estadístiques especialitzades i sofisticades, han permès elaborar noves vies de recerca per comprendre la etiologia de malalties complexes l'origen de les quals, en molts casos, és multifactorial. Així com s'han establert factors ambientals que poden modular el risc de patir certes malalties, també s'han detectat variants genètiques que hi poden estar involucrades. Patologies com la diabetis, el càncer, les malalties cardiovasculars, l'esquizofrènia o l'asma es veuen influenciades per factors genètics en interacció amb factors ambientals.

Al capdavant d'aquestes investigacions es troben els mapes de polimorfismes. El polimorfisme més comú al genoma humà és la variació en una sola base de la seqüència genòmica, l'anomenat *Single Nucleotide Polimorphism* i conegut per les seves inicials "SNP". Degut a la seva abundància, els SNPs són molt adients per generar mapes genètics i han esdevingut els marcadors més utilitzats en estudis d'associació genètica.

Si bé des de fa dècades l'estudi del genoma humà s'ha centrat principalment en analitzar les variacions en la seqüència genòmica, des d'inicis de l'any 2000 sabem per diversos estudis que aquestes variacions tendeixen a donar-se en bloc. D'altra banda, també s'ha demostrat que les recombinacions genètiques que es donen al llarg del genoma no es produeixen de manera uniforme. Per aquest motiu, el genoma presenta zones que es transmeten en bloc, de progenitors a descendents, i que poden incloure blocs de variacions. Aquestes zones de

baixa recombinació que es segreguen en bloc són els anomenats haplotips. Els haplotips poden facilitar el descobriment de gens relacionats amb malalties que pateixen els éssers humans.

L'interès en l'assignació d'haplotips i l'anàlisi de l'associació entre haplotips i malaltia en mostres d'individus no relacionats ha crescut incommensurablement als darrers anys degut a l'èmfasi que projectes com HapMap i d'altres iniciatives han situat sobre l'anàlisi d'haplotips. Ara bé, la determinació dels haplotips donada una mostra de genotips per un conjunt d'individus no sempre és immediata, havent de recórrer en alguns casos a tècniques específiques per tal de separar els cromosomes. Les tècniques de tipus molecular són les que aporten menys error però desafortunadament són cares i això dificulta el seu ús, sobretot en estudis poblacionals que tracten amb mostres grans. Per superar aquesta limitació, les investigacions han tendit a utilitzar la inferència estadística com a via més usual a l'hora de determinar els haplotips. La inferència sobre les freqüències haplotípiques és una bona solució per reconstruir la mostra haplotípica, però cal tenir present els efectes que el fet de treballar amb estimacions comportarà sobre tots els càlculs que es realitzin amb la mostra. En aquest sentit, resulta interessant dedicar esforços per tal d'intentar minimitzar la propagació d'aquests errors en les anàlisis d'associació genètica amb haplotips, qüestió que encara és oberta.

Tot i que existeix diversitat de programes per fer anàlisis haplotípiques aplicables a mostres d'individus no relacionats, molts d'ells presenten limitacions que esdevenen una bona motivació per intentar cercar d'altres alternatives teòriques i computacionals per tractar més eficientment la problemàtica dels haplotips.

En aquesta tesi doctoral es presenta el desenvolupament i la implementació informàtica d'un mètode per estimar haplotips i els efectes associats a diversos tipus de fenotips. El marc teòric amb que s'ha treballat és la inferència Bayesiana combinada amb tècniques de Markov Chain Monte Carlo que optimitzin les qüestions computacionals.

La present tesi està estructurada en 7 parts i un apèndix que conté 3 annexos. Cadascuna

de les parts la conformen diferents capítols.

Pel que fa a la part introductòria, està formada per un primer capítol on s'expliquen els conceptes bàsics biològics que són necessaris per comprendre el treball. Es recomana passar directament al capítol 2 a aquells que tinguin assolits aquests coneixements. Al segon capítol es presenta amb detall la rellevància de l'anàlisi haplotípica als estudis d'associació genètica. Tanca la part I un tercer capítol on s'exposa amb detall la problemàtica associada a la pròpia definició de la mostra haplotípica, juntament amb una revisió dels mètodes i softwares existents per fer anàlisi haplotípica. Un cop explicitades a la part II les hipòtesis en què basem aquest treball i els objectius que ens plantegem, arriba la part metodològica (part III de la tesi) on s'introdueix el concepte d'inferència Bayesiana, els mètodes de Monte Carlo i les Cadenes de Markov, fins a descriure amb detall les diferents tècniques de Markov Chain Monte Carlo i com aquestes poden adequar-se i aplicar-se a la qüestió dels haplotips. A la quarta part de la tesi, es presenta el mètode d'anàlisi haplotípica que s'ha dissenyat i implementat informàticament en aquest treball. Es descriu l'algorisme teòric que s'ha programat així com el paquet estadístic en l'entorn R de lliure utilització que l'implementa. La cinquena part es destina a mostrar els resultats obtinguts en aplicar el programa sobre escenaris simulats i sobre dades reals. L'avaluació dels resultats es troba recollida a la sisena part, la discussió, on es fa una valoració del mètode i una comparativa respecte d'altres programes ja en ús, basant-se en els resultats obtinguts i en la literatura existent. Finalment, tanca la tesi un apartat on s'exposen les principals conclusions extretes d'aquest treball.

A l'apartat d'annexos es troben diferents documents d'interès, com són tres articles en que he participat activament, emmarcats en aquest mateix camp, també una exposició ampliada i detallada sobre els aspectes matemàtics relacionats amb les propietats de les cadenes de Markov i un conjunt de taules que resumeixen les característiques de la majoria dels mètodes d'anàlisi haplotípica que existeixen.

Aquesta tesi ha rebut finançament del Ministerio de Salud, formant part del projecte anomenat *Papel de los polimorfismos en genes reparadores del ADN en el cáncer colorrectal esporádico y familiar* (PI030114) desenvolupat al servei d'Epidemiologia de l'Institut Català d'Oncologia amb el Dr.Víctor Moreno com a Investigador Principal. Per dur a terme aquest treball he gaudit d'una beca pre-doctoral concedida per l'Institut d'Investigacions Biomèdiques de Bellvitge (IDIBELL).

Vull agrair l'amabilitat i la disponibilitat del Dr.Julio Sanjuán i de la Dra.Dolores Moltó del Departament de Genètica de la Facultat de Biologia de la Universitat de València en cedirme un conjunt de bases de dades que m'han permès completar aquest treball.

També vull agrair l'ajut rebut per part del Dr.David Tregouet del grup de genòmica cardiovascular del *Institut National française de recherche en santé et medècine* - Universitat Pierre i Marie Curie de París.

Raquel Iniesta Benedicto

Barcelona, Setembre 2010

Summary

Nowadays, haplotypic information has become vitally important to clarify the genetic basis of the etiology of some common diseases. Comparing DNA of healthy and diseased individuals let us to describe changes in the genomic sequence that could modify the risk of suffering from the disease. Association studies are the framework where this class of analysis are carried out.

The DNA variations more often analyzed due to its high frequency along the genome are the Single Nucleotide Polimorphisms. One "SNP" is the change in only one nucleotide between individuals at the same position of their genomes.

Is well known that there are zones in the genomic sequence with a low rate of recombinations, that are inherited as a block by the offspring ([1], [2], [3] and [4]). These zones are called haplotypes, and everyone carries two of them. On the other hand, in the last decade researchers have stated that mutations as SNPs are also transmitted in blocks, situated in haplotypic zones [5]. For all of this, the knowledge of haplotypes corresponding to a sample of genotypes observed for some SNPs of a set of unrelated individuals could be very helpful to better understand the genetic association with a phenotype of interest.

Initiatives as the international HapMap project ([6],[7],[8],[9] and [10]) have strongly motivated the scientific community to use haplotypes in association analysis.

Unfortunately, in the absence of family data, obtaining haplotypic information is not straightforward. Since every cell of the human organism contains 22 pairs of homologous

chromosomes, plus the sexual chromosome, for each chromosomical location at the autosomal chromosomes there are two bases, one for each homologous chromosome at the same position of the DNA sequence. Given that current lab techniques usually only report genotypic data and do not provide the chromosome for each base, individuals with two or more heterozygous sites have uncertain haplotypes because there is more than one possible haplotype pair compatible with their genotype.

Methods of Haplotypic Reconstruction

In the last years several methods of haplotypic reconstruction have been developed in order to overcome this lack of information. Since Clark, in 1990 [11], developed a parsimony algorithm to estimate haplotype frequencies from a sample of genotypes, quite a large number of methods have been developed. Most of them rely on the use of different techniques to calculate the Maximum Likelihood Estimator (MLE).

In 1995, Excoffier and Slatkin [12] adapted the Expectation-Maximization algorithm, an iterative algorithm of maximization developed by Dempster in 1977 [13] to maximize the likelihood function of the haplotypes given the genotypes at specific loci. This method has some limitations and convergence to a local maximum may occur in some situations (Celeux and Diebolt,[14]).

Some authors have attempted to minimize these limitations in their works, like Qin *et al.* [15] using *Divide and conquer* strategies, or David Clayton, implementing an EM-algorithm (*snphap* software) which adds SNPs one by one and estimates haplotype frequencies, discarding haplotypes with low frequency as it progresses [16]. Besides, other techniques have been considered, too. In the context of Bayesian statistics, Stephens *et al.* in 2001 proposed an algorithm based on coalescent theory [17] with a especial prior based on the general mutational model. Niu *et al.* [18] implemented another Bayesian approach using a Markov Chain Monte Carlo method. In general, algorithms dealing with Bayesian models are suit-

able to infer haplotypes from genotypes having a large number of polymorphisms.

Once the frequencies have been estimated by any of the methods mentioned above, the next goal is to test the association between haplotypes and a disease. The most accurate strategy in order to take into account the uncertainty of the sample is to estimate simultaneously haplotype frequencies and haplotype effects. There are some works in this sense (Tanck *et al.*[19], Tregouet *et al.*[20]).

Methods

The algorithm we have developed makes the simultaneous estimation of haplotype frequencies and haplotype effects within the frame of Bayesian models. We aim to compute the Maximum Likelihood Estimator of the parameters using Markov Chain Monte Carlo techniques. To do so, it is first required to define the models for both cases in order to deduce the two associated likelihood functions.

Notation

Consider a sample of individuals of size N, and let be G_i the genotype for the i-th individual, $i = 0, \ldots, N$. Each individual has a finite number of haplotypes compatible with his genotype G_i. If this genotype has at most 1 heterozygous locus, there is only one possible pair of haplotypes compatible with it and there is no uncertainty. Let be m the number of heterozygous loci. If $m \geq 2$, the genotype has 2^m different haplotypes compatibles with it. Let be H_i, $i = 1, \ldots, 2^m$ the set of compatible haplotypes with the genotype of each individual. Assuming that in the whole sample there are M possible haplotypes, h_j denotes the j-th haplotype, with $j = 0, \ldots, M$. The sample frequency for each haplotype is denoted by f_{h_j}.

Likelihood for Genotypes Sample

Now, assuming *Hardy-Weinberg equilibrium*, the sample frequency for each G_i can be expressed by the product of the frequencies of every haplotype in H_i. For example, if an individual is certain, H_i only has two elements h_r and h_s, $r, s \in \mathbf{(1, \ldots, 2^m)}$, then $F_{G_i} = f_{h_r} \times f_{h_s}$. But for individuals with uncertain haplotypes, we have to consider the sum over all the possible pairs:

$$F_{G_i} = \sum_{h_r, h_s \in H_i} c_{rs} f_{h_r} f_{h_s} \tag{0.1}$$

where c_{rs} is a constant value, equal to 1 if $h_r = h_s$ and 2 if $h_r \neq h_s$. Now, taking the product of (0.1) over all the individuals, the likelihood function $\ell(F)$ of the sample of genotypes can be written as Excoffier and Slatkin stated in [12]:

$$\ell(F) = \prod_{i=1}^{N} F_{G_i} = \prod_{i=1}^{N} \sum_{h_r, h_s \in H_i} c_{rs} f_{h_r} f_{h_s} \tag{0.2}$$

where $F = \{F_{G_i}, i = 0, \ldots, N\}$

Estimation of Haplotype Effects. Linear, Logistic and Weibull Regression Models

The estimation of haplotype effects can be done with several designs. A case-control study is a very recommended solution, due to its good cost-effectiveness perform. For this design two samples, one of cases and other of controls are required. The suitable model to assess association between haplotypes and a binary response is the Logistic Regression model, which has related to its coefficients the definition of a useful measure of association, the odds ratio. Otherwise, for a longitudinal design, with a cohort of persons being followed during a period of time, survival analysis is more appropriate and measures like the Risk Ratio could be computed using models as the Weibull Regression. These measures of risk quantify the effect of a given haplotype over the response by comparison with the effect of the reference haplotype (usually the most frequent in the sample). For both designs is possible to analyze the association of a continuous outcome considering a simple Linear

regression model.

For all models there will be a parameter vector β of coefficients to be estimated, that are taking part in the likelihood function associated with each model.

Estimating Parameters

To estimate the parameters of every likelihood function, the haplotypical and the one associated to the chosen regression model, independence among the parameters for the two models is assumed. Then, two Markov Chains are created, one for each likelihood function, with stationary distribution the distribution of the unknown parameters. The method used to create the chains depends on the model:

- For the estimation of the haplotype frequencies in (0.2), a particular case of the Metropolis-Hastings algorithm, the *Random walk*, is a simple and efficient method.
- To estimate the parameters of the Linear, Logistic or Weibull regression model, the sampling will be generated using another particular case of the Metropolis-Hastings algorithm, the *Gibbs Sampler*.

The Algorithm

Rebuilding the Haplotypes Sample

It starts with a sample of genotypes of N individuals, with a known phenotype for each one Y_i. The algorithm begins taking an initial seed for the haplotype frequencies and for the regression coefficients. Then, the i-th step of the algorithm is described as follows:
Let be $f^{(i-1)} = (f_{h_1}^{(i-1)}, f_{h_2}^{(i-1)}, \ldots, f_{h_M}^{(i-1)})$ the previous state of the chain. Then, a new state $f^{(i)}$ is generated using Random Walk sampling, with invariant distribution proportional to (0.2):

1. $f^{(i)} = f^{(i-1)} + u$ where $u = (u_1, \ldots, u_M)$ such as $u_i \sim Unif(0, s)$ or $u_i \sim N(0, s)$ $i = 1, \ldots, M$ where s is chosen empirically.
2. Then, a value v is generated from a $Unif(0, 1)$ distribution.
3. if $v < \ell(f^{(i)})/\ell(f^{(i-1)}$ where ℓ is defined as in (0.2), the new state is accepted. If it is not, $f^{(i)} = f^{(i-1)}$.

After that, haplotypes for the uncertain individuals are rebuilt, drawing a value from a categorical distribution taking the frequencies of the previous state. For example, if an individual has a genotype compatible with the haplotypic pair $H_1 = (h_1, h_2)$ and also with $H_2 = (h_3, h_4)$, then $p_1 = P(H_1)$ and $p_2 = P(H_2)$. Now, a value from a $cat(p_1, p_2)$ is drawn, where $p_1 = f_{h_1} f_{h_2}/(f_{h_1} f_{h_2} + f_{h_3} f_{h_4})$ and $p_2 = f_{h_3} f_{h_4}/(f_{h_1} f_{h_2} + f_{h_3} f_{h_4})$

Estimation of Haplotype Effects

After having the rebuilt haplotypes for the whole sample, they are passed as a covariate inside the regression model and a new state of the chain for its coefficients is generated. This new state $\beta^{(i)}$ is sampled with a Gibbs sampler simulation:

1. The Gibbs sampler is a sampling method which draws values from the full conditional distribution of the model. Let be $\pi(\cdot \mid \beta)$ the full conditional function for the regression model. Then, the Gibbs Sampler makes $2^m + 1$ samples to generate the new state β^i of the chain, i.e.:

$$\beta_{k_j}^{(i)} \sim \pi(\beta_{k_j} | \alpha^{(i)}, \ldots, \beta_{k_{j-1}}^{(i)}, \beta_{k_{j+1}}^{(i-1)}, \ldots, \beta_{k_{2^m}}^{(i-1)})$$

   Notice that drawing the value $\beta_{k_j}^{(i)}$ is not straightforward. Several methods have been tested and finally Slice sampling have proved to be the faster and the most efficient sampling method ([21]) for these models.
2. Hence, $\beta^{(i)}$ is a new state of the chain.

This is a complete stage of the algorithm. Now, return to the first step and generate a new value for the chain of the haplotype frequencies.

Limiting distribution

The constructed Markov Chains are both irreducible and ergodic (i.e. aperiodic and positive recurrent), and so the limiting distribution is unique. This limiting distribution is the stationary distribution of the chain, and so it is the distribution of our parameters. Since the chain values are a sample of the parameters distribution, the posterior mean for f and β can be estimated by the arithmetic average of sample values and it can be taken as the MLE for the parameters. Furthermore, sample values allow us to calculate different estimators such as the median, the symmetry, etc. The variances for these estimators can also be calculated from the chain.

BayHap: The Bayesian package to analyse Haplotypes

This algorithm has been implemented in a C program and can be used through an R package called "BayHap". The package is formed by a set of routines that allows users to perform association analysis between haplotypes and three different type of outcomes: binary, survival and continuous. The package also allows to adjust with other covariates and with interaction terms between covariate and haplotypes. Several inheritance models can be selected too. The package also contain functions to print results, plot graphs and to evaluate the convergence of the generated chains.

Results

Performed simulations with BayHap show that with a burn-in period of about 500 iterations and a sample of 1000, the convergence of the chains is remarkably good. The curvature computed for parameters is good enough, even for haplotypes for low frequency. Results show estimation and curvature differences between results reported by BayHap and EM-algorithm, with a better performing for the Bayesian one.

Discussion

Although there are a lot of programs to estimate haplotypic frequencies, most of them do not perform association analysis or are following poor strategies to do it. The scheme consider in the present work seems to perform quite well in a varity of scenarios. A first good feature to point out is that for haplotypes with low frequency ($< 1/100$), the MCMC algorithm implemented in BayHap seems to be able to make a good estimation of the effect, while other commonly used algorithms of numerical optimization may have more difficulties to solve it. Results have also shown that the simultaneous algorithm diminishes the possibility of converging to a local minimum. Moreover, the considered simultaneous method of sampling gives a good estimation for the variance of β parameter, which is capturing the uncertainty of the haplotype sample. The alternative generation of two chains could make every rebuilding of the haplotype sample different at each step of the algorithm. Thus, individuals with more than two elements in H_i may be rebuilt in a different way depending on the f generated and the covariate value inside the model will then change. Therefore, for samples with a great number of ambiguous individuals, the variance of the β distribution generated with the MCMC algorithm is larger than with non-simultaneous methods. Hence, the latter ones may resolve an odds ratio as significant, while the former may not do it.

BayHap is robust regarding assumptions, and includes survival analysis in the R context.

Conclusions

Markov Chain Monte Carlo techniques and Bayesian inference can be successfully applied in the context of haplotype effects estimation. These techniques allow us to generate the distribution for each parameter and to have all the information about each one improving results given by other commonly used methods like the EM algorithm. Furthermore, for small sample sizes, estimations made with MCMC capture the possible asymmetry of the

sample distribution, while methods based on asymptotic estimators do not. MCMC also seems to perform quite well for haplotypes having low frequency in the sample. Finally, the simultaneous estimation we have considered diminishes the possibility of convergence to a local minimum, so it makes the algorithm suitable to be applied over samples with a large number of polymorphisms.

Although the implemented package BayHap requires users have a minimal previous R knowledge, the volume of information returned by BayHap and the precision of its results, set the program as a good alternative for haplotypic analysis.

Índex

Agraïments III

Pròleg VII

Summary XI

Part I INTRODUCCIÓ

1 Conceptes biològics 3

1.1 Processos biològics 4

1.1.1 Mitosi 5

1.1.2 Meiosi 5

1.1.3 Recombinacions 7

1.2 Polimorfismes 7

1.2.1 SNPs 8

1.3 Equilibri de Hardy-Weinberg 10

1.4 Desequilibri de Lligament 11

1.5 Haplotips 13

2 Estudis d'associació genètica. Paper dels Haplotips. 15

2.1 Estudis d'associació genètica 16

2.2 Tipus d'estudis d'associació genètica 17

2.2.1 Polimorfisme Candidat 18
2.2.2 Gen candidat 18
2.2.3 Regió candidata 19
2.2.4 Rastreig Complet (*Whole Genome Association Studies - WGAS*) 19
2.3 Quines metodologies d'estudi s'utilitzen? 19
2.4 Tècniques estadístiques adients per cada disseny i tipus d'estudi 21
2.5 Paper dels Haplotips als estudis d'associació genètica 23
2.5.1 Avantatges de l'anàlisi d'Haplotips 28

3 Problema Haplotípic i el seu tractament metodològic 31
3.1 Haplotips sense incertesa 31
3.2 Haplotips amb incertesa 32
3.3 Mètodes estadístics per l'anàlisi d'Haplotips amb incertesa 34
3.3.1 Mètode de la Parsimònia 36
3.3.2 Mètodes basats en la Funció de Versemblança 39
3.4 Eines per fer inferència sobre Haplotips incerts 44
3.5 Mètodes estadístics per l'anàlisi d'associació amb Haplotips 46
3.5.1 Mètode de les puntuacions estadístiques (*Scores*) 47
3.5.2 Models de Regressió per Haplotips incerts 49
3.6 Eines per fer l'anàlisi d'associació amb haplotips 55

4 Què podem aportar a la metodologia Haplotípica? 59

Part II HIPÒTESIS DE TREBALL I OBJECTIUS

5 Hipòtesis de treball 65

6 Objectius d'aquesta tesi 67

Part III MÈTODES

7 Mètodes Bayesians 71

7.1 En què es basa l'enfocament Bayesià? 72

7.1.1 Teorema de Bayes 74

8 MCMC: Integració per Monte Carlo i Cadenes de Markov 77

8.1 Integració per Monte Carlo 78

8.2 Cadenes de Markov 79

8.3 Métodes de Markov Chain Monte Carlo 80

8.3.1 Idea intuitiva 81

8.3.2 Algorisme de Metropolis-Hastings 82

8.3.3 Algorisme de Metropolis 85

8.3.4 Gibbs Sampling 86

8.3.5 Mètodes per mostrejar de funcions de densitat no estàndards: DFARS i Slice Sampling. 89

9 Punt de trobada entre MCMC, l'estadística Bayesiana i el problema haplotípic . 103

9.1 Funció de versemblança per les freqüències haplotípiques 104

9.2 Models estadístics segons el tipus de disseny i funcions de versemblança associades 105

9.2.1 Model Lineal generalitzat: Regressió Lineal, Regressió Logística i Regressió de Weibull 106

9.3 Distribucions a priori per a cadascun dels models 112

9.4 Aplicació de tècniques MCMC per l'estimació dels paràmetres 112

9.4.1 Algorisme de Metropolis per estimar les freqüències haplotípiques . . . 113

9.4.2 DFARS i Slice Sampling per estimar l'associació amb fenotip 114

9.5 Els haplotips com a factor de risc: estimació simultània 115

Part IV ALGORISME DISSENYAT EN AQUESTA TESI. IMPLEMENTACIÓ INFORMÀTICA

10 L'algorisme que hem creat 119

10.1 L'algorisme pas a pas 121

10.1.1 Descripció teòrica de l'algorisme 121

10.2 Què hem obtingut? 124

11 BayHap, el paquet Bayesià d'anàlisi d'associació amb haplotips 127

11.1 R i la programació de paquets 127

11.2 BayHap 128

11.2.1 Funcions del paquet 129

11.2.2 Ús del paquet 130

11.2.3 Arguments modificables 132

Part V RESULTATS

12 Aplicació de BayHap sobre escenaris simulats. Comparació amb d'altres programes. 137

12.1 Escenaris en que s'han simulat les bases de dades 138

12.1.1 Descripció numèrica dels escenaris 140

12.2 Resultats de les simulacions 144

13 Algorisme EM vs BayHap en l'anàlisi del gen DRD2 153

13.1 Component genètic en la etiologia de l'Esquizofrènia i del Càncer Colorectal esporàdic 154

13.1.1 Paper del gen DRD2 154

13.2 Anàlisi d'associació en dos estudis 156

13.3 Estudi cas-control en pacients amb esquizofrènia 157

13.3.1 Polimorfismes del gen DRD2 analitzats en aquest estudi 157

13.3.2 Resultats de l'anàlisi d'associació 158

13.4 Estudi cas-control en càncer de còlon 168

13.4.1 Polimorfismes del gen DRD2 analitzats en aquest estudi 169

13.4.2 Resultats de l'anàlisi d'associació 169

13.4.3 Resultats de l'anàlisi de supervivència 182

14 Diferents consideracions de distribucions a priori 189

15 Diferents tractaments de la incertesa haplotípica a l'anàlisi d'associació 191

Part VI DISCUSSIÓ

16 Funcionament de BayHap respecte de la resta de programes 197

16.1 Comparació punt per punt 198

16.1.1 Mètodes i algorismes 198

16.1.2 Precisió 199

16.1.3 Assumpcions 202

16.1.4 Nombre i tipus de marcadors 207

16.1.5 Mida de la mostra 208

16.1.6 Característiques del Software 209

16.1.7 Anàlisi d'associació 212

16.2 Inferència Bayesiana vs Freqüentista 215

17 Consideracions Finals d'aquesta Tesi Doctoral 219

18 Limitacions 223

Part VII CONCLUSIONS

19 Conclusions 229

Part VIII APÈNDIX

A Articles publicats 233

B Taula de programes d'estimació haplotípica 237

C Especificacions matemàtiques 241

Referències 257

Índex alfabètic 277

Índex de taules

12.1 Taula de resultats per freqüències a l'escenari 1 segons BayHap 145
12.2 Taula de resultats per OR a l'escenari 1 segons BayHap 145
12.3 Taula de resultats per freqüències a l'escenari 1 segons Haplo.Stats 145
12.4 Taula de resultats per OR a l'escenari 1 segons Haplo.Stats 146
12.5 Taula de resultats per freqüències a l'escenari 2 segons BayHap 146
12.6 Taula de resultats per OR a l'escenari 2 segons BayHap 147
12.7 Taula de resultats per freqüències a l'escenari 2 segons Haplo.Stats 147
12.8 Taula de resultats per OR a l'escenari 2 segons Haplo.Stats 148
12.9 Taula de resultats per freqüències a l'escenari 3 segons BayHap 148
12.10 Taula de resultats per OR a l'escenari 3 segons BayHap 148
12.11 Taula de resultats per freqüències a l'escenari 4 segons BayHap 149
12.12 Taula de resultats pels coeficients de la regressió lineal a l'escenari 4 segons BayHap 149
12.13 Taula de resultats per freqüències a l'escenari 4 segons Haplo.Stats 150
12.14 Taula de resultats pels coeficients de la regressió lineal a l'escenari 4 segons Haplo.Stats 150
12.15 Taula de resultats per freqüències a l'escenari 5 segons BayHap 151
12.16 Taula de resultats per l'escenari 5 segons BayHap 151

13.1 Freqüències al·lèliques i genotípiques pels polimorfismes del gen DRD2 per l'estudi d'esquizofrènia. 159
13.2 P valors de Hardy-Weinberg 160
13.3 Models d'associació amb Esquizofrènia per cada polimorfisme del gen DRD2 160
13.4 Valors de D' per la mostra general 161
13.5 Valors d'r per la mostra general 161
13.6 P Valors per la mostra general 161
13.7 Freqüència haplotípica i OR amb intervals de confiança segons BayHap i Haplo.Stats (H.S). Haplotips referents als SNPs per ordre: -241, -141, TaqIB, rs1800499, Ser311Cys, His313His, 6277, TaqIA 162
13.8 Freqüències al·lèliques i genotípiques pels polimorfismes del gen DRD2 analitzats a la mostra de CCR. 172
13.9 P valors de Hardy-Weinberg 173
13.10 Models d'associació amb càncer colorectal per cada polimorfisme analitzat del gen DRD2 173
13.11 Freqüència haplotípica i OR segons PHASE i BayHap. Haplotips referents als SNPs per ordre: -141, TaqIB, 3208T, Ser311Cys, rs6277, 1412G, TaqIA 182
13.12 Freqüència haplotípica i HR segons BayHap i THESIAS amb intervals de confiança per l'estudi de CCR. Haplotips referents als SNPs per ordre: -141, TaqIB, 3208T, Ser311Cys, rs6277, 1412G, TaqIA 182

Índex de figures

1.1 Passes que conformen el procés de la meiosi 6
1.2 Canvi en una sola base (SNP) 9

3.1 Genotips obtinguts al laboratori 33

8.1 Funció de densitat multidimensional 82
8.2 Passeig d'una cadena via Gibbs Sampling 83
8.3 Un pas de l'slice sampling utilitzant procediments de stepping-out i shrinkage. 99
8.4 El procediment de doubling. 100

10.1 Esquema simplificat de l'algorisme iteratiu 125
10.2 A cada iteració es genera cadascun dels paràmetres creant una cadena de Markov que es resumeix mitjançant la teoria ergòdica. 126

11.1 Imatge de la consola d'R amb una execució de BayHap, juntament amb alguns dels resultats numèrics i gràfics obtinguts 131
11.2 Imatge d'una pàgina del help del programa BayHap. 132

13.1 Polimorfismes del gen DRD2 155
13.2 Mitjanes ergòdiques per cada coeficient de la regressió logística corresponent a cada haplotip en la mostra d'esquizofrènia. 164

13.3 Densitats del mostreig realitzat per cada coeficient de la regressió en la mostra d'esquizofrènia. 165
13.4 Autocorrelacions parcials de cadascuna de les cadenes en la mostra d'esquizofrènia. 166
13.5 Sèries per a cada coeficient de la regressió en la mostra d'esquizofrènia. 167
13.6 Mitjanes del mostreig realitzat per cada freqüència haplotípica. 174
13.7 Autocorrelacions parcials del mostreig realitzat per cada freqüència haplotípica en la mostra de càncer. 175
13.8 Densitats del mostreig realitzat per cada freqüència haplotípica en la mostra de càncer. 176
13.9 Seqüència mostrejada per cada freqüència haplotípica en la mostra de càncer. 177
13.10 Mitjanes del mostreig realitzat per cada coeficient de la regressió Logística en la mostra de càncer. 178
13.11 Densitats del mostreig realitzat per cada coeficient de la regressió Logística en la mostra de càncer. 179
13.12 Autocorrelacions del mostreig realitzat per cada coeficient de la regressió Logística en la mostra de càncer. 180
13.13 Termes de la serie temporal pel mostreig realitzat per cada coeficient de la regressió Logística en la mostra de càncer. 181
13.14 Mitjanes del mostreig realitzat per cada coeficient de la regressió de Weibull en la mostra de càncer. 184
13.15 Autocorrelacions parcials del mostreig realitzat per cada coeficient de la regressió de Weibull en la mostra de càncer. 185
13.16 Densitats del mostreig realitzat per cada coeficient de la regressió de Weibull en la mostra de càncer. 186
13.17 Densitats del mostreig realitzat per cada coeficient de la regressió de Weibull en la mostra de càncer. 187

15.1 Freqüències pels 6 haplotips més freqüents i pels estranys (”rare”) 192

15.2 Estimacions i variances de les estimacions pels coeficients del model logístic amb covariables els haplotips. 193

B.1 Taula de programes de reconstrucció haplotípica. 239

B.2 Taula de programes que inclouen mètodes d’anàlisi d’associació. 240

Acrònims

ARS *Adaptive Rejection Sampling*

CCR Càncer Colorectal

DFARS *Derivative Free Adaptive Rejection Sampling*

EE *Estimating Equation*

ECM *Expectation Conditional Maximization algorithm*

ELB *Excoffier-Laval-Balding Algorithm, Bayesian*

EM *Expectation Maximization algorithm*

EM Issues Que pot ser sensible a les desviacions de HWE, temps d'execució elevat i convergència a un màxim local i no global, requerint repeticions amb diverses llavors.

HF Freqüència haplotípica estimada

HA Assignació haplotípica individual

HWE Equilibri de Hardy-Weinberg

HR *Hazard Ratio*

IC Interval de Confiança

IP Mètode basat en filogènia imperfecta

JRE *Java Runtime Environment*

LD *Linkage Disequilibrium*

MAC Programa que s'executa en un ordinador Apple

MC Algorisme de Monte Carlo, algorisme Bayesià

MCMC Algorisme de Markov Chain Monte Carlo, algorisme Bayesià

MC-VL *Monte Carlo-Variable Length Chain algorithm*, algorisme Bayesià

MLE *Maximum Likelihood Estimation*

OR Odds Ratio

PC Ordinador personal compatible amb IBM

PL *Partition Ligation*

PP Mètode basat en fliogènia perfecta

P-L Limit pràctic computacional dels programes sobre el nombre de marcadors i/o individus.

RR Risc Relatiu

S-EM Algorisme EM estocàstic

SNP *Single Nucleotide Polimorphism*

TRV Test de Raó de Versemblança

UNIX Sistema operatiu que inclou Linux, FORTRAN, Solaris i d'altres

WGAS *Whole Genome Association Studies*

Part I

INTRODUCCIÓ

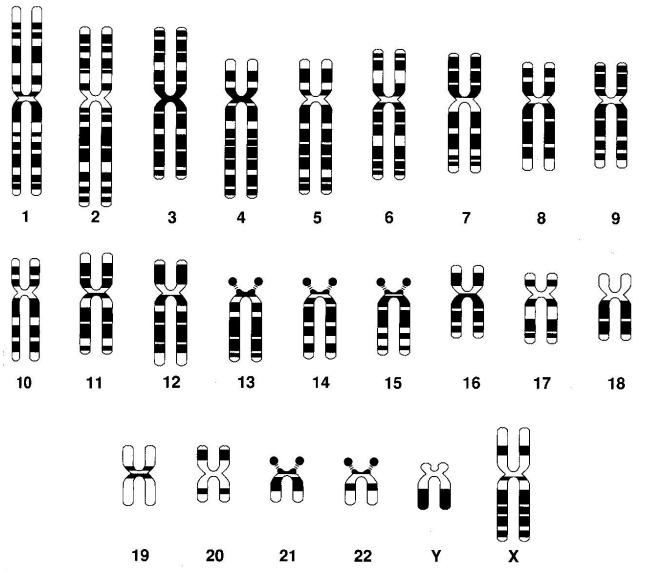
1

Conceptes biològics

La composició química de qualsevol organisme consta principalment d'aigua i de proteïnes. Les proteïnes són tan abundants perquè tenen dos papers fonamentals a la vida de les cèl·lules: la creació d'estructures internes i el control de les reaccions químiques que s'hi produeixen a l'interior.

La informació sobre quines proteïnes pot fabricar cada cèl·lula es troba codificada a l'ADN. L'ADN està format per una seqüència de molècules anomenades *nucleòtids*. D'aquestes molècules n'existeixen 4 tipus: *Adenina* (A), *Citosina* (C), *Timina* (T) i *Guanina* (G) i el seu ordre al llarg de la seqüència determinarà les proteïnes que codificarà la cèl·lula i quina serà la funció que desenvoluparan.

L'ADN de les cèl·lules eucariotes (les humanes ho són) es troba al nucli cel·lular, fragmentat en una sèrie de cadenes allargades que es situen sobre unes proteïnes anomenades *histones* que ajuden a mantenir la forma de l'ADN. El conjunt d'ADN i histones rep el nom de *cromosoma* i només és visible durant l'etapa de divisió cel·lular. És en aquesta etapa quan els cromosomes es dupliquen i es disposen en forma de *X*. Cada cèl·lula humana porta 22 parelles de cromosomes homòlegs (anomenats autosòmics) i una parella més que correspon als cromosomes sexuals. Els *gens* són segments d'ADN que codifiquen almenys una proteïna. En el seu conjunt els gens conformen el *genoma* de l'individu. Qualsevol variació en la seqüència de nucleòtids per un gen en concret pot implicar un canvi en la síntesi de proteïnes per part de la cèl·lula.



23 parelles de cromosomes

La posició que ocupa un determinat gen al llarg d'un cromosoma es denomina *locus*. Gens diferents al mateix locus són denominats *al·lels*. Per a cada locus tenim informació doblada, la corresponent a cada cromosoma. Quan dos loci presenten idèntics al·lels es diu que l'individu és *homozigot* en aquest locus. En cas de presentar dos al·lels diferents, l'individu es diu *heterozigot*. La combinació al·lèlica que porta un individu al llarg del seu genoma s'anomena *genotip*. Aquesta variabilitat al·lèlica que pot donar-se en mateixos loci, en combinació amb factors ambientals en alguns casos, dóna lloc a expressions diferents del mateix caràcter. A aquestes manifestacions externes se les anomena *fenotip*.

1.1 Processos biològics

Per entendre com arribem fins a la situació cromosòmica que analitzarem, cal tenir clars dos processos cabdals a la vida de la cèl·lula: la mitosi i la meiosi.

1.1.1 Mitosi

Cada cromosoma de les cèl·lules humanes, excepte els situats en cèl·lules que desenvoluparan gàmetes sexuals, és creat fent una còpia d'un cromosoma ja existent. Això té lloc durant el procés de divisió cel·lular anomenat *mitosi*. Just abans de la divisió, durant l'etapa de mitosi la cèl·lula crea una còpia idèntica de cada cromosoma i per tant cadascuna de les dues noves cèl·lules rep un conjunt complet de 46 cromosomes. Per tant, cada nova cèl·lula té el mateix conjunt de cromosomes i la mateixa informació genètica que la cèl·lula inicial. Això explica perquè cada cèl·lula del nostre cos té la mateixa informació genètica.

1.1.2 Meiosi

Un procés lleugerament diferent té lloc durant la producció de les gàmetes (masculines o femenines). El nucli d'un espermatozoide formarà part del nucli d'un zigot humà. I el mateix per l'òvul. Però si el procés de divisió previ a la creació d'aquestes gàmetes fos una mitosi, el zigot humà arribaria a tenir $46 + 46$ cromosomes! Per evitar aquesta anomalia, en comptes d'una mitosi, el que es dóna és un procés anomenat *meiosi*.

El procés de la meiosi parteix d'una sola cèl·lula (amb 46 cromosomes). Els cromosomes homòlegs s'uneixen, es dupliquen (n'arribem a tenir 92) i se separen. És en aquesta separació on es dóna la combinació genètica, perquè els cromosomes resultants no són els mateixos que els inicials. La cèl·lula s'acaba dividint dues vegades, donant 4 cèl·lules reproductives que duen cadascuna 23 cromosomes. Un fet important és que la combinació de gens que porten als seus 23 cromosomes és resultat de la barreja dels gens que la cèl·lula inicial portava. Una cèl·lula de la mare i una altra del pare formaran el Zigot que esdevindrà un nou ésser. Per tant, cada parella de cromosomes homòlegs del nou ésser estarà formada per un cromosoma matern i un cromosoma patern per a cada parella, però no seran cromosomes exactes als que duien els pares. Així doncs, la descendència s'assembla als seus pares, perquè la meitat de la informació que porten els seus gens, prové de la seva mare,

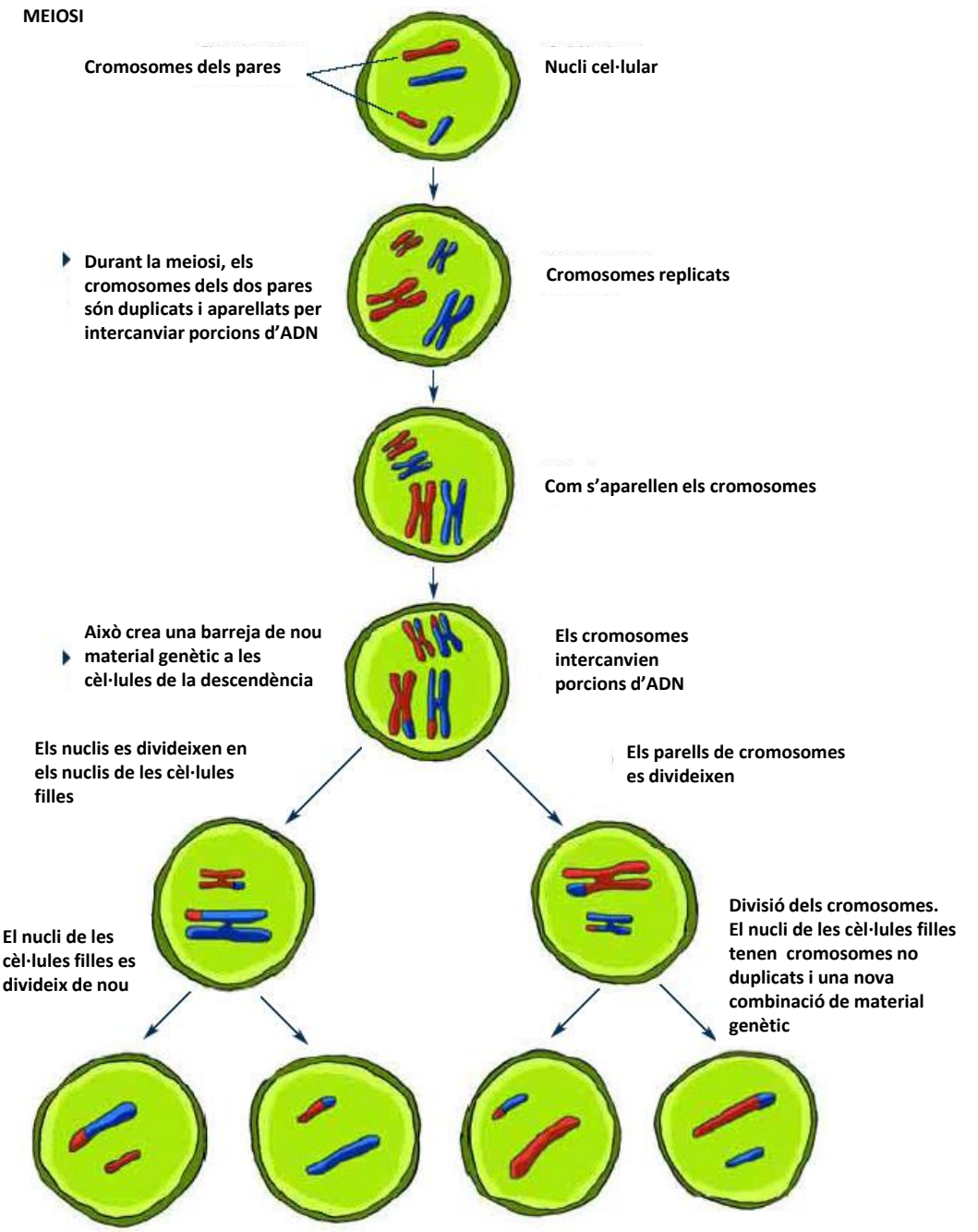


Figura 1.1. Passes que conformen el procés de la meiosi

i l'altre meitat del seu pare, però ni són idèntics a ells, ni ho són amb els seus germans, ja que els processos de recombinació són diferents en cada cas.

1.1.3 Recombinacions

La *Recombinació genètica* és un procés d'intercanvi genètic que es dóna entre les seqüències d'ADN de dos cromosomes homòlegs. Aquest intercanvi es produeix a base de entrecreuaments entre seqüències d'ADN de dos progenitors diferents. Conjuntament amb les mutacions, les recombinacions són les causants que existeixi variabilitat genètica. Per a que apareguin nous genotips com a conseqüència de les recombinacions, és essencial que les dues seqüències homòlogues siguin genèticament diferents. Aquest és el cas que ens ocupa, en tractar-se de cromosomes de pares diferents.

Els entrecreuaments a l'ADN poden causar que al·lels que prèviament es trobaven en el mateix cromosoma siguin separats. Quant més lluny es troben els al·lels entre sí, més probable és que es produeixi una recombinació entre ells i siguin separats. Aquest concepte està molt lligat amb el de Desequilibri de lligament, que definirem a la secció 1.4.

La *Freqüència de Recombinació* és la freqüència amb que tenen lloc entrecreuaments entre dos loci (o gens) durant la meiosis. Es tracta d'una mesura de lligament genètic molt utilitzada a l'hora de fer mapes de lligament. La freqüència d'entrecreuaments per cromosoma és petita, d'1 a 4 i depèn de la mida del cromosoma. La freqüència entre dos loci propers és molt baixa i per això s'observa que la dependència estadística entre loci tendeix a disminuir en successives generacions fins a arribar a la independència.

1.2 Polimorfismes

Els *polimorfismes genètics* són variants de gens que apareixen per mutacions espontànies a la població i que es transmeten a la descendència, prenent certa freqüència dins la població, després de múltiples generacions. S'ha estimat que al genoma cada 1000 parells de bases

dels 3.000 milions de bases que el configuren, apareix una variant. Els polimorfismes són la base de l'evolució i poden o bé no tenir repercussió funcional, poden proporcionar avantatges als individus, o bé poden ser responsables de malalties. Es coneixen moltes malalties determinades genèticament per mutacions o variants, denominades d'alta penetrància, perquè els portadors de la variant solen manifestar la malaltia amb alta probabilitat. Aquestes variants acostumen a ser de baixa freqüència en la població general. Els punts on genomes diferents varien s'anomenen *marcadors genètics*. Per tant els polimorfismes són marcadors genètics.

A l'actualitat molts investigadors centren els seus treballs en identificar gens amb polimorfismes que es donen en la població en major freqüència i que influeixen en el risc de patir una malaltia, però amb baixa probabilitat. Són els anomenats *polimorfismes de baixa penetrància*. Les variants més freqüents són les que es donen en una sola base (SNP). D'altres polimorfismes són repeticions d'una seqüència curta d'ADN. Aquests es denominen VNTR (sl Variable tandem repeat), d'altres es basen en delecions o insercions de seqüències curtes de nucleòtids.

1.2.1 SNPs

Un SNP (*Single Nucleotide Polymorphism*) és un polimorfisme genètic que correspon a la variació en un sol nucleòtid.

En mostres amb mida rellevant per fer recerca biomèdica, la gran majoria dels SNPs tenen dos al·lels. L'SNP representa la substitució d'una base per una altra. Per un sol SNP designarem l'al·lel major al que es presenti amb major freqüència a la població. Així doncs, donat que els humans som diploids amb cromosomes materns i paterns en el seu origen, donat un SNP concret una persona pot tenir diversos genotips: homozigot per l'al·lel major, heterozigot o homozigot per l'al·lel menor.

Els SNPs poden ser identificats a la seqüència d'ADN mitjançant diferents tècniques

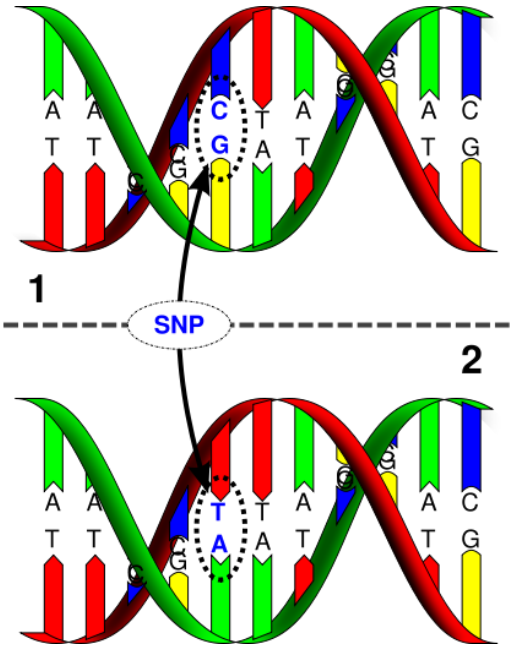


Figura 1.2. Canvi en una sola base (SNP)

([22],[23], [24],[25],[26],[27],[28] i [29]). Encara que molts SNPs són identificats d'aquestes maneres, les característiques com ara la freqüència al·lèlica, la freqüència genotípica i la poblacional de cada SNP no poden ser determinades únicament amb aquestes estratègies. La tècnica més directa i menys esbiaixada per caracteritzar-los és la de seqüenciar la mateixa regió en diferents poblacions ([30],[31]). La mida mostral de la població que és reseqüenciada és important. En general es necessita una mostra gran per identificar SNPs en relació a l'al·lel menor. Per exemple, la teoria de genètica poblacional prediu que per tenir una detecció d'un SNP del 99% es requereixen 48 cromosomes per un SNP que tingui una freqüència per l'al·lel menor del 5% o més. Per una detecció similar, es necessitarien 192 cromosomes, si la freqüència de l'al·lel menor fos del 1% o major [32]. Obtenir una col·lecció d'al·lels comuns és possible, però obtenir el conjunt de tots els SNPs, incloent els estranys, requereix esforços majors ([33],[34],[35]). En termes generals, els SNPs es donen un cop cada 200 parelles de bases ([36],[37],[38],[39]) al genoma humà. Aquells SNPs cat-

alogats com a estranys poden donar-se tan sols un o dos cops en la mostra a estudi. La definició de "comú" depèn de l'aplicació, però es trobaria entre més d'un 5% a més d'un 20% com a freqüència per l'al·lel menor. El caire subjectiu d'aquesta definició es deu a l'ampli rang reportat a la literatura [40].

Si el canvi en un únic nucleòtid es dóna en una zona codificant pot provocar un canvi en la proteïna resultant i això pot implicar una modificació de la seva activitat o funció. Els canvis també es poden donar en zones del promotor d'un gen i modificar la seva expressió. Aquestes zones promotores modulen el procés de transcripció de l'ADN en ARN, el primer pas de la descodificació d'un gen en una proteïna. El mateix pot passar si el canvi es dóna en un intró. Tot i que els introns no es tradueixen en una proteïna, canvis en la seva estructura poden modular l'expressió del gen.

D'altres cops, probablement la majoria, els canvis són silents i no tenen repercussions funcionals. Tot i que només estudis moleculars específics poden posar de manifest si els polimorfismes són funcionals, els estudis epidemiològics són fonamentals per valorar si hi ha efectes en la salut de la població ([41],[42],[43]).

1.3 Equilibri de Hardy-Weinberg

Abans de procedir a l'anàlisi d'associació entre un polimorfisme i una malaltia, és important avaluar si les freqüències genotípiques es corresponen amb el valor esperat, suposant que la transmissió de cada al·lel (del pare i de la mare) és independent. A aquesta propietat se la denomina Equilibri de Hardy-Weinberg.

Considerem un locus bial·lèlic on hi participen els al·lels (A_1, A_2). Les possibles combinacions al·lèliques observables per aquest locus seran (A_1, A_1),(A_1, A_2),(A_2, A_1), o bé (A_2, A_2). Siguin p_1 i p_2 les freqüències pels dos al·lels, respectivament, on $p_2 = 1 - p_1$ donat que només es poden donar dues possibilitats al·lèlilques. En cas de cumplir-se HWE, donada la independència de transmissió, la probabilitat que una de les combinacions d'al·lels

es dongui a un locus concret coincideix amb el producte de les dues freqüències al·lèliques. La següent taula mostra totes les proporcions:

	A_1	A_2
A_1	p_1p_1	p_1p_2
A_2	p_2p_1	p_2p_2

(1.1)

L'equilibri de Hardy-Weinberg ens pot ser de gran utilitat a l'hora de calcular certes probabilitats per parelles d'al·lels. En general, sempre que la mostra sigui d'individus no relacionats parentalment podrem suposar que es compleix. Per tant suposarem que els entrecreuaments entre individus es donen a l'atzar.

Per testar-lo, les freqüències esperades sota compliment de HWE es poden comparar amb les observades utilitzant un test de χ^2.

1.4 Desequilibri de Lligament

Entre diferents polimorfismes localitzats al mateix cromosoma se sol observar un cert grau de correlació o associació estadística denominada desequilibri de lligament, en anglès *Linkage Disequilibrium* i abreviat com LD. Aquest grau de correlació és degut a que, com hem vist a la secció 1.1.2, durant el procés de la meiosi en que es generen les gàmetes, els cromosomes que es transmetran no seran còpies exactes dels cromosomes dels progenitors, degut als entrecreuaments que generaran recombinació. La probabilitat que entre dos loci propers es dongui recombinació és petita, per això s'observa desequilibri de lligament. És a dir, al·lels de loci propers en cromosomes parentals tendeixen a viatjar units cap a la descendència.

El Desequilibri de Lligament tendeix a desaparèixer en successives generacions, fins arribar a l'equilibri, que correspon a la independència estadística.

Suposem que partim de dos loci situats en un segment qualsevol de cromosoma que corresponen a dos marcadors genètics bial·lèlics (per exemple, dos SNP's), que denominarem A i B, amb dos al·lels cadascun: els al·lels A_1 i A_2 pel primer SNP i els al·lels B_1 i B_2 pel segon. La freqüència de l'al·lel A_1 és p_1, de A_2 és p_2, de B_1 és q_1 i de B_2 és q_2. Ara ens preguntem per la probabilitat que en un cromosoma aparegui una parella concreta d'al·lels (un haplotip). Als cromosomes de la població podem esperar trobar les quatre combinacions genètiques possibles d'aquests al·lels, és a dir: $(A_1, B_1), (A_1, B_2), (A_2, B_1)$ o (A_2, B_2). En cas de donar-se equilibri, la freqüència d'aquestes combinacions es calcula mitjançant el producte de les freqüències de cada al·lel, igual que pel cas d'equilibri de Hardy-Weinberg. Si denotem la probabilitat d'A_1 com p_1, la d'A_2 com p_2, la de B_1 com q_1 i la de B_2 com q_2. Ara la taula seria:

	A_1	A_2
B_1	p_1q_1	p_2q_1
B_2	p_1q_2	p_2q_2

(1.2)

on $p_2 = 1 - p_1$ i $q_2 = 1 - q_1$.

Al cas ideal en que cada al·lel tingués una freqüència de $0,5$, trobaríem cada combinació al·lèlica en un 25% dels cromosomes analitzats.

Però suposem ara que aquests gens no es trobessin en equilibri de lligament; és a dir, que trobéssim en excés algunes combinacions i en faltessin d'altres. Per exemple, pot ser que trobem haplotips (A_1,B_1) i (A_2,B_2) amb freqüències més elevades que les que podríem esperar i (A_1,B_2) i (A_2,B_1) amb freqüències menors.

La magnitud d'aquest desequilibri de lligament (denominada D) és variable entre marcadors genètics i entre poblacions, i apareix als càlculs de la següent manera:

	A_1	A_2
B_1	p_1q_1+D	p_2q_1-D
B_2	p_1q_2-D	p_2q_2+D

(1.3)

on $p_2 = 1 - p_1$ i $q_2 = 1 - q_1$ i $D \in (0, 1)$.

Al cas extrem en que dos marcadors estiguessin tan fortament lligats que sempre es transmetessin junts D valdria gairebé 1. En cas contrari, si no hi ha desequilibri, D tendeix a 0. D'altra banda, D disminueix a mida que transcorren las generacions i tendeix lentament a 0. Si no actua cap altre factor, aquesta disminució depèn del temps (a més temps, més recombinacions) i de la freqüència de recombinació existent entre els marcadors considerats.

1.5 Haplotips

Un *haplotip* és la constitució al·lèlica de múltiples loci per un mateix cromosoma. Les investigacions han constatat que els SNPs (definició a 1.2.1) s'hereten en grups que es troben estretament relacionats a l'ADN, en contrast amb la idea sostinguda que plantejava la segregació a l'atzar, degut a les recombinacions genètiques. A aquest conjunt d'SNPs que s'hereten en bloc és al que es denomina haplotip.

S'anomena *fase* a la configuració en que es troben disposats els al·lels en un mateix cromosoma. En concret, es diu que els al·lels que formen un haplotip estan *en fase*.

En una definició més general, un haplotip és el genotip d'un cromosoma simple o d'un grup haploide de cromosomes. Actualment l'haplotip és la nova unitat funcional de la genòmica. Es coneix que més de 10000 nucleòtids s'hereten en bloc, i degut a la quantitat d'SNPs que hi ha al genoma humà, en aquest bloc hi ha un gran nombre d'SNPs. Aquests SNPs que estan presents en un haplotip poden trobar-se en la seqüència d'un gen o en la de múltiples gens, permetent determinar el context en el qual actuen els gens.

A l'hora de determinar els haplotips que duu un individu pot passar que el genotip no defineixi unívocament els seus haplotips. Per exemple, considerem un organisme diploide i dos loci bial·lèlics que siguin SNPs. El primer locus té al·lels A i T amb tres possibles

genotips: AA, AT i TT. El segon locus té al·lels G i C, donant lloc de nou a tres possibles genotips GG, GC i CC. Per un individu donat, imaginem que dugui dos loci heterozigots, AT i GC. Fixem-nos que si el laboratori no ens ha informat sobre el cromosoma que conté cada al·lel, aquest genotip permet fer dues possibles separacions en cromosomes: AG en un cromosoma i TC en l'altre, o bé, AC i GT per cada cromosoma respectivament. Per individus homozigots a ambdós loci no hi ha problema de determinació, però per dos loci heterozigots hi ha incertesa haplotípica.

La resolució de la fase haplotípica pot dur-se a terme mitjançant tècniques de laboratori, però desafortunadament es tracta de mètodes poc cost-efectius i que impliquen força temps. Aquest fet ha motivat la necessitat de desenvolupar diferents tècniques de reconstrucció haplotípica basant-se en enfocaments diversos, com veurem més endavant en aquest treball.

L'estudi d'haplotips s'ha convertit en una eina molt útil a l'hora de determinar la relació genètica entre individus i per tant en l'estudi de l'origen de mutacions causants de diversos fenotips. Amb freqüència són més d'un els polimorfismes que s'analitzen simultàniament en un gen o regió candidata i és especialment interessant que així sigui, ja que el fet de considerar més d'un locus facilita identificar polimorfismes relacionats amb certs fenotips d'interès. És aquí on els haplotips prenen rellevància. El motiu és que el polimorfisme associat al fenotip a estudi pot ser desconegut però trobar-se en LD amb d'altres polimorfismes. Per això identificar haplotips ens pot ser de gran utilitat per localitzar variants funcionals. Si diferents individus amb mateix valor per un fenotip concret són portadors dels mateixos haplotips en una zona polimòrfica, aquest fet pot ser un indici que en la zona considerada pot trobar-se una variant causal.

2

Estudis d'associació genètica. Paper dels Haplotips.

Els estudis d'associació genètica han esdevingut la principal via per localitzar les zones del genoma que confereixen risc moderat de patir malalties que presenten component genètic ([44],[45],[46],[47]). La informació que aporta l'anàlisi d'haplotips als estudis que involucren múltiples marcadors és cabdal per assolir els objectius de l'estudi d'associació donat que permeten entendre les correlacions entre marcadors i determinar variants funcionals que modifiquin el risc associat al fenotip a estudi. Així doncs, a les darreres dues dècades els haplotips han tingut un paper clau en l'estudi de la base genètica que presenten certes malalties comuns i d'altres més complexes com és el cas del càncer, les malalties cardiovasculars, l'asma, la diabetis o l'esquizofrènia.

Des del punt de vista clínic, s'ha demostrat que existeix associació entre el conjunt d'al·lels transferits en bloc per part de cadascun dels progenitors, els haplotips, i diverses malalties ([1],[2],[3],[4]). A més, s'ha constatat que aquesta associació no s'observa si es consideren els SNPs individualment ([48],[49],[50]). Entre aquests articles es troben exemples del pes que pot representar el fet de ser portador d'un haplotip a l'hora de determinar l'associació genètica amb cert fenotip ([51],[52]). És per exemple el cas del gen COMT, variacions del qual s'han associat amb una modificació en la susceptibilitat de patir trastorn psicòtic [53] o del gen ZDHHC8 que també s'ha associat amb aquesta malaltia [54]. Aquesta associació tan pot ser indicadora d'una modificació del risc de malaltia atribuïble al propi fet de ser portador d'un haplotip concret, o bé pot estar suggerint l'associació amb d'altres SNPs que

es trobin en LD amb els estudiats. Per tant, els haplotips s'utilitzen habitualment com a localitzadors de gens o loci associats a una malaltia.

A banda d'aquest interès, una altra àrea on els haplotips també estan mostrant validesa clínica significativa és en el camp de la farmacogenòmica. És ben conegut que la variació individual en la resposta a un fàrmac és atribuïble a algunes variants genètiques específiques ([55],[56]).

En aquest capítol introduirem els estudis d'associació i ens centrarem en entendre la funció que estan tenint els haplotips en aquest tipus d'investigació.

2.1 Estudis d'associació genètica

Els estudis d'associació genètica poblacional tenen com a objectiu principal identificar patrons de polimorfismes que varien sistemàticament entre individus que tenen un estat de malaltia diferent i així poder descriure al·lels o grup d'al·lels que modifiquen el risc de patir la malaltia. Es tracta d'estudis útils per avaluar l'associació entre una malaltia i un o més factors genètics.

En primer lloc, és important disposar de certa evidència que almenys una part de la malaltia ve determinada genèticament. Per aquest motiu, són útils els estudis d'agregació familiar, els de bessons i els d'emigrants. En segon lloc, cal que s'identifiqui on són els gens d'interès per la malaltia. En aquesta fase es realitzen estudis anomenats de lligament (en anglès *linkage*) que utilitzen com a marcadors genètics una sèrie de polimorfismes repartits per tot el genoma. En aquests estudis se solen triar famílies grans amb diversos membres afectats per la malaltia a estudi, permetent identificar zones del genoma d'interès per la comprensió de la malaltia. Tot i així, aquests estudis tenen poca resolució: a les zones identificades poden haver centenars de gens interessants i milers de polimorfismes candidats. Per determinar amb major precisió els gens d'interès i dins d'aquests gens, el o els polimorfismes responsables, s'utilitzen estudis d'associació en els que es compara la freqüència relativa

de les diferents variants d'una sèrie de polimorfismes entre individus afectats i un grup control adequat. Aquests estudis acostumen a triar gens candidats que podrien tenir la seva funció relacionada amb la malaltia a estudi, i dins d'aquests gens es genotipen diferents polimorfismes en individus afectats i no afectats. És d'esperar que les variacions que es donen especialment en aquells individus malalts o sans, o bé contribueixin d'alguna manera a modificar el risc de patir-la o bé es trobin en una zona on algun altre SNP sigui el que modifiqui el risc. Aquests polimorfismes acostumen a ser SNPs tals que alguna de les seves variants codifiquen proteïnes que poden alterar funcions que poden influenciar el fenotip d'interès.

2.2 Tipus d'estudis d'associació genètica

Existeixen diferents estratègies a l'hora d'identificar la relació entre un polimorfisme o variant en un gen i certa malaltia. Cada tipus d'estudi difereix en el nombre d'SNPs a analitzar i també els diferencia la necessitat d'informació prèvia abans d'iniciar l'anàlisi. Els estudis d'un sol polimorfisme, gen o regió candidata són adients per detectar gens que estan relacionats amb malalties comuns i d'altres més complexes, tals que el risc degut al factor genètic és relativament petit. Així doncs, per aquests tipus d'estudi el primer pas crític a l'hora de dur-los a terme serà la tria adequada del gen o de la zona. En canvi els estudis de rastreig complet analitzen tot el genoma per tal de detectar un marcador associat al fenotip a estudi.

També cal tenir en compte que en qualsevol d'aquests estudis es podria donar una associació fals-positiva degut a un efecte d'estratificació de la població, és a dir, situació en que les freqüències al·lèliques difereixen en les subpoblacions de casos i de controls, per un incorrecte aparellament de casos i controls o per efecte de l'atzar. És important tenir present que el genoma és tan llarg que patrons que podrien suggerir associació amb una malaltia,

podrien ser únicament fruit de l'atzar ([57],[58]).

2.2.1 Polimorfisme Candidat

Els estudis que analitzen polimorfismes candidats es basen en l'anàlisi d'un SNP individual que és suspecte d'estar implicat en la malaltia. Es tracta d'un tipus d'estudi que requereix informació prèvia sobre quin SNP triar. L'anàlisi de l'SNP ens aportarà informació sobre l'efecte que té l'SNP individualment sobre el fenotip que s'estigui estudiant. A més, utilitzant les tècniques estadístiques adients podrem quantificar la magnitud de l'associació, com veurem a 2.4. Aquestes tècniques permeten ajustar els resultats per possibles variables de confusió i per termes d'interacció entre el polimorfisme i d'altres factors. A [59] Iniesta et al. presentem una estratègia estàndard d'anàlisi d'SNPs.

2.2.2 Gen candidat

Es tracta d'un tipus d'estudi d'associació genètica en que es considera més d'un SNP. Als estudis d'associació de gens candidats, es tria un gen basat en coneixement previ que habitualment prové de resultats d'un estudi de famílies o bé de models animals. Aquest estudis involucren entre 5 i 50 SNPs aproximadament pertanyents al gen. Podria donar-se el cas que cap dels SNPs analitzats sigui causal però que sigui d'interès per la presència de desequilibri de lligament entre ells i l'SNP causal. En aquest cas es poden dur a terme estudis de cadascun dels SNPs per separat i també un anàlisi de múltiples SNPs (aquest treball 2.4).
En aquest cas, a més de l'anàlisi de cada polimorfisme podrem testar l'associació del conjunt d'SNPs, així com també serà possible fer una anàlisi d'haplotips que ens permetrà localitzar d'altres SNPs causals al mateix gen que potser no han estat genotipats.

2.2.3 Regió candidata

Aquest estudis són duts a terme sobre regions candidates d'entre 1-10Mb. La zona ha d'haver estat identificada per estudis de lligament i pot arribar a contenir entre 5 i 50 gens. El nombre d'SNPs que s'acostuma a genotipar es troba entre 10 i 100. Les tècniques d'anàlisi seran les mateixes que les exposades per l'estudi d'un gen candidat.

2.2.4 Rastreig Complet (*Whole Genome Association Studies - WGAS*)

Un inconvenient d'aquests tipus d'estudi que acabem de descriure és el fet que l'investigador ha d'inicialment fer la tria del gen o la regió que vol investigar. La gran diferència entre els estudis de regió candidata i els de *whole-genome* és que els darrers no requereixen un candidat com a gen o regió causal. Acostant-se més a l'estil del disseny d'estudi de lligament, el genoma sencer és testat per detectar la relació entre un marcador i un fenotip. Aquest seria un exemple d'un enfocament indirecte, donat que l'investigador es recolza en el desequilibri de lligament entre el presumpte marcador no funcional (o funcionalment no relacionat) i l'SNP causal [46]. En aquest cas però, el nombre d'SNPs que cal genotipar és més gran que en un estudi de lligament. Caldria genotipar entre 170000 i més d'un milió d'SNPs en funció del grau de desequilibri de lligament que presenti la població ([60],[61],[4]).

2.3 Quines metodologies d'estudi s'utilitzen?

Pel que fa a la metodologia de l'estudi, s'utilitzen dissenys epidemiològics clàssics basats en individus no relacionats. També es poden considerar dissenys basats en famílies en què els individus control són parents dels casos, com per exemple els dissenys de casos i germans sans o trios (cas i pares) ([62],[63],[64]). Tot i que els dos tipus d'estudi, el de famílies i el d'associació amb individus no relacionats, se centren en identificar la zona que pot contenir

un locus causal, cadascun dels estudis pren una aproximació diferent a l'hora de mesurar les recombinacions sobre els individus a estudi. En un estudi de lligament amb famílies, les recombinacions específiques poden ser directament mesurades, donat que són les pròpies recombinacions les que separen els marcadors genotipats del locus causal, si no es troben prou a prop del locus. En canvi, en estudis d'associació les recombinacions es mesuren indirectament mitjançant l'estudi del desequilibri de lligament, un reflex o producte de les recombinacions històriques en el temps, en individus relacionats llunyanament [62].

El disseny més simple per tractar amb individus no relacionats és el transversal, que recull dades referents a fenotips i SNPs per una mostra aleatòria d'individus. Aquest disseny és adient si la malaltia d'interès és una malaltia comuna o bé si l'investigador està interessat en estudiar algun tret relacionat amb la malaltia (com pot ser per exemple la pressió arterial).

Per l'estudi de malalties rares, és més adient utilitzar l'estudi de cas-control. Es tracta d'un disseny d'estudi molt potent a l'hora d'identificar associacions entre una variant i cert fenotip, per variants que confereixen risc moderat. En aquest estudi es recol·lecten dades retrospectivament en una mostra de casos (individus que pateixen la malaltia) i en una mostra de controls (individus que no presenten la malaltia). Aquest disseny és molt habitual en els estudis d'associació genètica degut al seu cost-efectivitat en la recollida de dades. A més, en un disseny d'aquest tipus, els investigadors no han de fer assumpcions sobre el mode exacte en que la malaltia va ser transmesa. El major problema del disseny de cas-control és que pot dur a associacions falses degut a una mala sel·lecció dels controls en relació a la raça o a d'altres factors que influenciïn la composició genètica dels individus.

Si la característica d'interès per exemple és l'edat de diagnòstic de la malaltia a estudi, aleshores és preferible realitzar el seguiment d'una cohort d'individus a risc de malaltia en el temps, potser exposant a part dels individus a unes condicions concretes que es volen analitzar com a associades al fet de desenvolupar la malaltia. Durant aquest seguiment es registra el temps que triga cada individu fins a desenvolupar la malaltia a estudi, en cas

que arribi a desenvolupar-la.

Els estudis de cohort ofereixen diverses avantatges en relació als estudis de cas-control [65]. Per exemple, algunes característiques com ara l'edat de diagnòstic aporten més informació per entendre la etiologia de malalties complexes que el fet de saber únicament si l'individu pateix o no la malaltia. Podem veure diversos exemples a [66]. Ara bé, les dades genotípiques haurien de ser conegudes en tota la cohort i això de vegades pot resultar molt car en cohorts de gran nombre d'individus. En aquests casos també és possible considerar un altre tipus d'estudi anomenat de cas-cohort [67] en que només cal genotipar un subconjunt dels membres de la cohort.

2.4 Tècniques estadístiques adients per cada disseny i tipus d'estudi

A l'hora de plantejar un estudi d'associació cal tenir present que la qualitat de les dades és una qüestió de gran importància. Les dades s'han de testar pel que fa a problemes d'estratificació, a efectes d'altres variables com pot ser el centre de recollida de dades i també testar la possible presència de patrons inusuals de valors perduts. També es necessari comprovar el supòsit d'equilibri de Hardy-Weinberg definit a 1.1. En condicions habituals, si la transmissió dels al·lels de progenitors a descendents és independent i no es donen fenòmens distorsionadors com l'aparició freqüent de noves mutacions o la sel·lecció d'al·lels, s'ha de complir Hardy-Weinberg. Abans de realitzar una anàlisi d'associació s'ha de comprovar que es compleix aquest principi com a mostra representant de la població general. Les desviacions de HWE poden ser degudes a un excés d'heterozigosi o d'homozigosi en un locus concret. En cas que s'observés una desviació caldria revisar el mètode de genotipació. També podria passar que els individus no siguin independents, que estiguem sel·leccionant un al·lel associat amb alguna característica de la mostra o bé que per atzar estiguem al 5% d'error inherent al test estadístic d'independència que es duu a terme.

En relació a l'estudi transversal i al de cas-control, l'avaluació de l'associació entre un SNP o múltiples SNPs i la malaltia es pot dur a terme mitjançant un model de regressió Logística com es pot veure a [68]. Aquest model no assumeix cap distribució per les covariables, que són tractades no-paramètricament. El model també permet la inclusió de termes d'interacció entre les variables genètiques i les variables ambientals. Tot i que sovint no s'explicita, la principal condició que ens porta a utilitzar la regressió Logística en un disseny de cas-control és que es compleixi l'equilibri de Hardy-Weinberg tant pels casos com pels controls. El model de regressió Logística a més permet estimar de manera no esbiaixada l'Odds Ratio (aquest treball 9.2.1). Es tracta d'una mesura adient per descriure com de gran és l'associació entre els factors genètics i la malaltia, per quantificar-ne l'efecte.

En cas de l'estudi de cohorts, la informació genètica també pot ser incorporada als diferents models. Aquesta anàlisi es pot fer de manera paramètrica, no-paramètrica o semi-paramètrica. Si triem la manera paramètrica hem de tenir present que les distribucions que habitualment s'apliquen a d'altres àrees de l'estadística, i molt en particular la distribució normal, no són vàlides en una anàlisi de supervivència. Per aquestes anàlisis necessitem distribucions definides sobre la recta real positiva i amb un coeficient d'asimetria negatiu. Una distribució adequada que acostuma a descriure bé el temps de supervivència és la distribució de Weibull (capítol 9.2.1), una distribució que inclou la exponencial com a cas particular i que s'adapta molt bé al truncament. Els mètodes no-paramètrics, com poden ser les taules de la vida i l'estimador de Kaplan-Meier, són molt populars en anàlisis de supervivència donat que algunes característiques especials de les dades de supervivència no s'aconsegueixen modelar fàcilment mitjançant distribucions. Com a model semi-paramètric destaquem el model de Cox, model de tipus multivariant que consisteix en establir una relació paramètrica entre la variable dependent i les covariables. El model de Cox és l'equivalent en supervivència al model de regressió lineal.

Com veurem al capítol següent, a l'hora de testar l'associació entre haplotips i fenotip augmenta la complexitat degut a la dificultat de definir els haplotips per alguns genotips en

concret. Si els haplotips són observables directament, qualsevol d'aquestes tècniques que s'acaben de citar seran adients. En cas que per alguns individus hi hagi incertesa haplotípica, haurem de considerar algunes de les solucions proposades a la secció 3.3.

2.5 Paper dels Haplotips als estudis d'associació genètica

Com ja s'ha definit en aquest treball a la secció 1.5, un haplotip és la combinació d'al·lels de diferents loci propers que es troben en un mateix cromosoma i que presenten certa correlació entre ells, de tal manera que tendeixen a viatjar conjuntament cap a la descendència. Donat que els humans som organismes diploids, al conjunt de loci genotipats li correspon dos haplotips, on cada haplotip o bloc d'al·lels correspondrà a un i altre cromosoma, el transmès per part del pare i el transmès per part de la mare. Aquests al·lels hauran estat transmesos en bloc des dels cromosomes originals materns o paterns, si en aquests cromosomes es trobaven en LD (aquest treball 1.4), propietat que permet assumir que els al·lels no han estat separats per recombinació. Al·lels de loci propers, per exemple, segueixen aquesta propietat. Actualment, no existeix millor manera per entendre els patrons de LD que la de conèixer els haplotips. Els haplotips ens informen directament sobre com s'organitzen els al·lels al llarg dels cromosomes, reflectint els patrons d'herència que han dut a l'evolució. Daly et al. [1] ofereixen un clar exemple que demostra com el coneixement dels haplotips pot ser vital en l'anàlisi del LD.

Als estudis d'associació genètica el rol dels haplotips variarà segons la hipòtesi que es pretengui testar. En aquests estudis s'analitzen els haplotips formats per al·lels de loci propers i polimòrfics. Es tracta del genotipatge de zones que donen lloc a diferents possibilitats haplotípiques entre els individus de la mostra. D'un cantó, els haplotips poden representar un efecte sobre el fenotip a estudi, resultat de la combinació de diverses zones al llarg del mateix cromosoma que no podria ser detectat si s'analitzessin els SNPs un per un. D'altra

banda, el fet que un sol SNP aparegui associat amb una malaltia significa que o bé l'al·lel està contribuint al risc de patir la malaltia, o bé es troba en desequilibri de lligament amb un altre SNP que hi està contribuint. Per això, una associació positiva entre un fenotip i un haplotip, pot indicar que una zona no directament genotipada però associada a d'altres al·lels en el mateix cromosoma (haplotip) contribueix al fenotip. En aquest cas, la investigació acostuma a focalitzar-se en descobrir i genotipar d'altres variacions per determinar el grau d'associació que presenta l'haplotip. En cas de disposar de la seqüència completa de variacions, si hi ha molt desequilibri de lligament, els efectes individuals de cada SNP poden quedar sense resoldre, tot i els grans esforços que la investigació apliqui.

Aquest enfocament sobre com testar gens candidats en els estudis d'associació genètica ha millorat durant els darrers anys gràcies a l'existència de bases de dades públiques que contenen milions de marcadors útils per estudis d'associació genètica. Més encara, és possible obtenir descripcions detallades de les recombinacions en relació a les variacions ([5],[69]) i al desequilibri de lligament [70] per molt gens.

Tot i que entre poblacions es comparteix una proporció d'haplotips, hi ha diferències entre freqüències [71] que poden ser rellevants en un estudi d'associació. Alhora, cal tenir present també el possible efecte d'estratificació de població, que pot engrandir les estimacions del desequilibri de lligament [72].

Donada la gran quantitat d'SNPs que com s'ha vist requereixen els estudis d'associació *whole-genome* és d'esperar que l'interès recaigui en desenvolupar mètodes que ajudin a triar el conjunt òptim d'SNPs a genotipar. En aquest sentit i lligat als estudis en què es genotipen o bé un gen candidat [73], un cromosoma sencer [3] o bé una àmplia regió del genoma en població de mida moderada, apareix el concepte de blocs d'haplotips com descriuen Daly et al. a [1]. Aquests blocs d'haplotips s'han demostrat molt útils en els estudis d'associació *whole-genome*. En general, a mesura que el nombre de marcadors augmenta el nombre d'haplotips també s'incrementa, formant eventualment haplotips que són únics

en alguns individus. Daly et al. constaten que la regió 500-kb del gen 5q31 genotipat en una població d'ascendència europea té regions discretes de baixa diversitat haplotípica. Aquestes regions, denominades "blocs", estan formades per fins a 100kb de llargada i generalment consisteixen en conjunts de 2 a 4 haplotips que representen més del 90% dels cromosomes estudiats. Dins dels blocs, s'observa molt poca o cap recombinació (resultat d'estar en gran desequilibri de lligament). Entre els blocs, s'observa agrupament de recombinacions, resultant en un desglossament del desequilibri de lligament. Aquestes troballes duen a la conclusió que aquests punts de recombinació formen els límits dels blocs d'haplotips ([1],[2]). En aquests articles podem veure com Gabriel et al [4] formalitzen una definició de blocs haplotípics utilitzant D' (aquest treball 1.4) mesura de desequilibri de lligament. A més, els autors també demostren que els blocs d'haplotips existeixen al llarg del genoma humà en diverses poblacions.

Impulsats per la perspectiva que el genoma humà pot ser descrit per blocs d'haplotips, el *National Human Genome Research Institute* (NHGRI) del *National Institutes of health* (NIH) iniciaren el Projecte Internacional HapMap. Diversos països (Japó, Regne Unit, Canadà, Xina, Nigèria i els Estats Units) s'uneixen per fer realitat aquest projecte que pretén descriure els patrons comuns de variació en la seqüència d'ADN (freqüències, patrons de desequilibri de lligament, etc), identificar i catalogar similituds i diferències genètiques entre humans, basades en determinar haplotips comuns formats per SNPs i a més, fer aquestes dades de domini públic per tal que els investigadors interessats en dur a terme estudis d'associació *whole-genome* puguin utilitzar-les ([6],[7],[8],[9]i [10]). Per fer això, el *International HapMap Consortium* proposà un enfocament jeràrquic de genotipatge i d'anàlisis. Així doncs, el consorci genotipa més de 3.000.000 d'SNPs amb una freqüència de l'al·lel menys freqüent superior al 5% i espaiats aproximadament en 5kb.

Malgrat tot, les tècniques de genotipatge són cares, i això ha fet que s'hagi destinat especial èmfasi en identificar marcadors que eliminin d'altres marcadors redundants, és a dir, marcadors que estiguin en gran desequilibri de lligament entre ells. Es tracta de triar

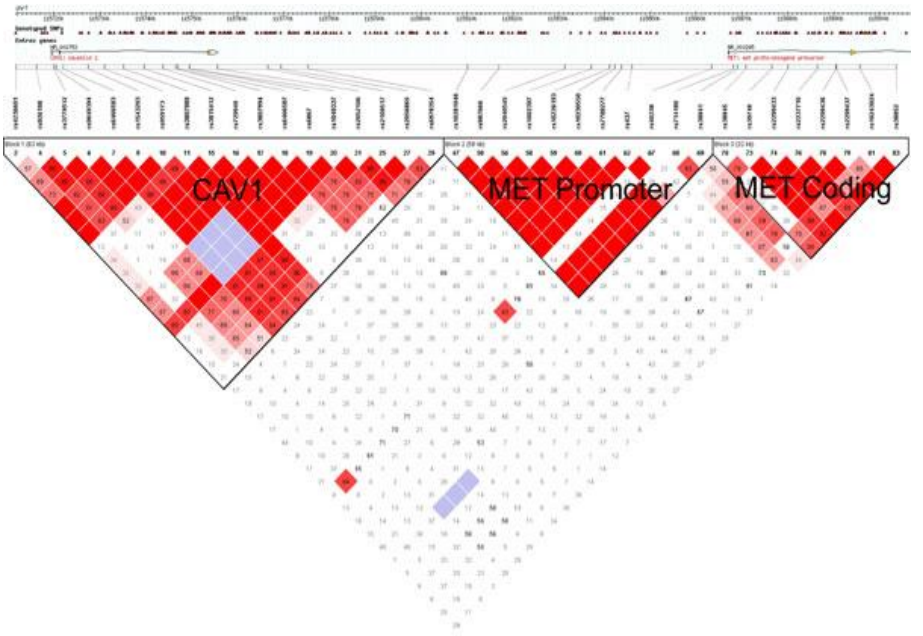


Figura de hapmap que representa l'estructura en blocs de LD d'una regió del cromosoma 7.

marcadors que puguin representar de la millor manera possible la variació genètica del gen candidat a estudi. Aquests SNPs són els anomenats tagSNPs. Als estudis que utilitzen tagSNPs, es determinen i genotipen SNPs que identifiquen haplotips de manera única [74]. Des que es va introduir el concepte de tagSNP, s'han desenvolupat diversos mètodes estadístics per identificar SNPs que capturin tota la diversitat haplotípica observada a la població ([75],[76]). La taula 2.1 mostra diferents aplicacions que resolen aquesta qüestió. Un aspecte esperançador és que només amb una petita mostra de la població a estudi ja és possible identificar els tagSNPs ([77],[78]) fent de l'estratègia una via eficient i alhora econòmica.

Nom prog	Lloc Web	Referència
Haploview	www.broad.mit.edu/personal/jcbarret/haplo	119
LDSelect	droog.gs.washington.edu/ldSelect.html	29
SNPtagger	www.well.ox.ac.uk/ xiayi/haplotype/index.html	117
TagIT	popgen.biol.ucl.ac.uk/software.html	118
TagSNPs	www-rcf.usc.edu/ stram/tagSNPs.html	90

(2.1)

La incertesa que en alguns casos presenta la determinació dels haplotips provoca que algunes d'aquestes aplicacions presentin limitacions. Molts d'aquests algorismes requereixen haplotips [75] però no tenen en compte els haplotips que poden haver estat inferits incorrectament [76]. Alhora, aquests algorismes assumeixen que els haplotips conformen un patró de bloc [75] o bé imposen aquest patró com a part de l'algorisme [76]. Una altra limitació es que molts dels gens candidats i de les regions del genoma presenten diferents haplotips [71]. Aquesta variabilitat en la diversitat d'haplotips que es poden donar en gens candidats limita la eficiència d'aquests algorismes. Degut a totes aquestes limitacions resulta més recomanable triar tagSNPs basats en el desequilibri de lligament de dades seqüenciades que no pas en haplotips inferits [38]. Una altra limitació que cal tenir en compte és que tagSNPs triats en una població, per exemple la Europea, no són apropiats per genotipar en una població diferent, com per exemple l'Africana. Per això han calgut poblacions diferents (Europea, Africana i Asiàtica) per determinar els tag SNPs. Les dades són analitzades segons diversos mètodes, incloent l'enfocament basat en blocs d'haplotips ([6],[7],[8],[9]). En tot cas, el principal avantatge de la baixa diversitat haplotípica o blocs haplotípics pels estudis d'associació *whole-genome* és que per representar els haplotips dins d'un bloc només és necessari genotipar un nombre reduït de tagSNPs.

La creació de HapMap sense dubte enriqueix diverses àrees d'investigació. Es tracta d'un gran avenç per conèixer l'estructura en bloc del genoma humà, que a més pot ser aplicada

al disseny dels estudis d'associació *whole-genome* i a l'anàlisi. Per exemple, encara que diversos estudis han constatat l'estructura en blocs en regions del genoma diferents a 5q31 ([79],[80]) els límits dels blocs podrien haver estat generats per d'altres causes i no per punts de recombinació ([80],[81]). Aquesta idea que d'altres forces poden haver influït als límits dels blocs i en la seva mida té importants repercussions a l'hora de triar el nombre d'SNPs necessari per dur a terme un estudi d'associació *whole-genome* en diverses poblacions. En efecte, és ben sabut que les poblacions amb ascendència Africana tenen un nombre de blocs curts superior a les poblacions amb ascendència Europea [70]. Per tant el mapa per les poblacions cal que sigui més dens. També Wall i Pitchard [82] determinen en diverses poblacions que tot i que el genoma humà exhibeix estructura en blocs, aquesta és desigual. El nombre i la mida dels blocs depèn de la densitat d'SNPs [83], la freqüència triada com a punt de tall per l'al·lel menys freqüent ([84],[85]) i també de l'algorisme triat per definir els blocs. Tot i així, més estudis de patrons de lligament han assegurat que el projecte international HapMap és una eina d'utilitat pública en la cerca dels gens i loci causals de malaltia ([8],[6]).

2.5.1 Avantatges de l'anàlisi d'Haplotips

L'anàlisi d'un sol SNP pot presentar poc poder per detectar associació donat que alguns SNPs poden estar altament correlacionats. En cas que entre els SNPs genotipats es dongui poc desequilibri de lligament degut a una gran distància entre ells o en cas que tots els SNPs siguin genotipats (i per tant en cas d'haver-ne un de causal, també serà genotipat) l'estudi de cada SNP individualment pot resoldre el nostre objectiu. Ara bé, a la pràctica, analitzar SNPs d'un en un pot provocar una pèrdua d'informació sobre la distribució conjunta dels SNPs. La majoria dels estudis es basen en analitzar SNPs genotipats propers en el cromosoma i no amb tota la densitat d'SNPs existents a la regió candidata. Per tant, els estudis de més d'un SNP tenen avantatges substancials envers els d'un únic SNP. Per això, una estratègia molt habitual motivada per l'estructura en bloc del genoma humà és

utilitzar haplotips per intentar capturar l'estructura de correlacions entre SNPs en regions de baixa recombinació. Els haplotips formats per SNPs que poden ser o no funcionals poden aportar més informació que les anàlisis d'un sol marcador a l'hora de determinar associació genètica amb una malaltia ([86],[87],[88]). Aquest fet és degut a que la distribució haplotípica captura l'estructura ancestral, com es pot veure a ([89]). La literatura que tracta sobre la comparació d'efectivitats entre analitzar haplotips respecte d'analitzar marcadors individuals és complicada donat que hi ha diverses característiques implicades en les anàlisis, com el nombre de loci, el nombre de possibles al·lels en cada loci i el grau de desequilibri de lligament entre els al·lels possibles a cada locus. Des del punt de vista estadístic, l'enfocament haplotípic és preferible donat que porta a anàlisis amb menor graus de llibertat.

3

Problema Haplotípic i el seu tractament metodològic

El fet que els haplotips hagin esdevingut tan importants a l'hora d'identificar loci associats a malaltia ha fet créixer considerablement l'interès per desenvolupar mètodes d'assignació d'al·lels a cromosomes. Aquest representa un camp d'investigació molt ampli degut a que el fet de determinar la parella d'haplotips que porta un individu no sempre és immediat. Com s'ha descrit a la secció 1.5 donat el genotip d'un individu, aquest duu dos haplotips, l'un format pels al·lels transmesos en bloc pel pare i l'altre format pels al·lels transmesos per la mare. Així doncs, donat un genotip per determinar els haplotips compatibles amb ell haurem de ser capaços de discernir quins al·lels pertanyen a cadascun dels progenitors. Anem a veure quines tècniques poden resoldre aquesta qüestió.

3.1 Haplotips sense incertesa

Actualment hi ha dues vies que permeten determinar els haplotips sense incertesa: directament genotipant pedigrees i utilitzar mètodes moleculars en combinació amb genotipar mostres d'individus que no tenen informació de pedigree. Els mètodes basats en famílies es fonamenten en el fet que loci diferents al mateix cromosoma (haplotip) seran heretats com una unitat a no ser que siguin separats per un cas de recombinació. La probabilitat d'una recombinació depèn en part de la distància entre els marcadors que s'estiguin tenint en compte. Els marcadors que són propers físicament tenen una probabilitat major d'estar lligats. Els loci es diuen lligats, o *linked* en anglès, si viatgen plegats (si es cosegreguen ple-

gats) quan són transmesos dels pares a la descendència com un haplotip. La recombinació entre dos cromosomes crearà dos nous haplotips que podran ser potencialment transmesos a les següents generacions.

En estudis poblacionals, els mètodes moleculars o experimentals són el mètode "gold standard" per reconstruir haplotips, essent diversos els mètodes moleculars existents per reconstruir haplotips. Dos dels mètodes més utilitzats inclouen *allele-specific polymerase reaction* conegut amb les inicials com AS-PCR i híbrids cel·lulars somàtics ([90],[91]). Aquests mètodes moleculars distingeixen quin al·lel és a cada cromosoma, una passa que generalment no és necessària en estudis familiars, donat que en aquest cas la informació pot ser extreta a partir de determinar els al·lels transmesos pels pares a la seva descendència. Una reacció PCR comú duta a terme en una mostra individual sense informació familiar explicarà quins dos al·lels són presents a la mostra, però un AS-PCR explicarà a més quin al·lel és present en relació a un altre al·lel en el mateix cromosoma. La tècnica dels híbrids cel·lulars somàtics és un mètode que separa físicament els cromosomes patern i matern d'un individu, abans de genotipar-lo. Tant la AS-PCR [92] com els híbrids cel·lulars somàtics [3] són tècniques moleculars que han estat utilitzades per determinar els haplotips en poblacions petites o moderades. Un article publicat aquest mateix any utilitza la microdissecció per realitzar la separació cromosòmica reportant resultats de gran precisió [93].

3.2 Haplotips amb incertesa

Encara que els estudis familiars i els mètodes moleculars eliminen la incertesa en assignar al·lels a cromosomes, les dues tècniques resulten cares i necessiten molt de temps per ser dutes a termes. Generalment, si no s'han fet servir mètodes moleculars de separació de cromosomes, les dades que ens proporciona el laboratori pel que fa al genotip d'un individu

són un seguit de lletres, que representen al·lels, sense especificar en quin dels cromosomes homòlegs es troben cadascuna. És a dir, suposem que estem estudiant dos loci en un cromosoma. El laboratori ens proporcionaria: A/A i B/B. Això vol dir que, en dos loci diferents d'un cromosoma, en un locus hi tenim l'al·lel A per un dels cromosomes i també l'al·lel A pel seu homòleg, i per un altre locus del cromosoma hi tenim l'al·lel B en un dels dos cromosomes, i també B al seu homòleg, al mateix punt. Si només observem un locus, no tenim

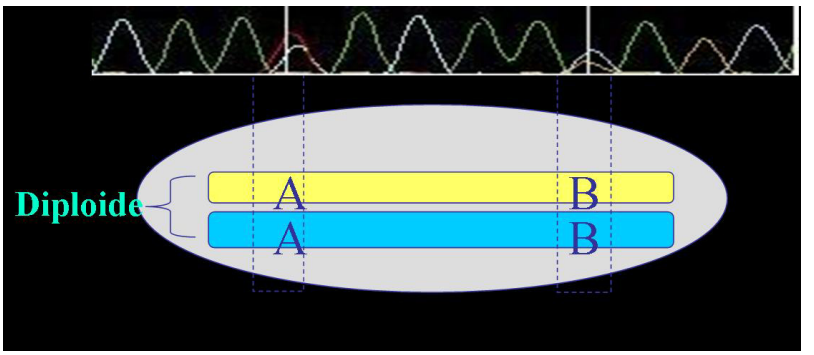


Figura 3.1. Genotips obtinguts al laboratori

cap dubte que un al·lel pertany a un cromosoma i l'altre pertany al cromosoma homòleg. És a dir, l'individu amb genotip A/A té l'al·lel A en un cromosoma i l'A en l'altre. Quan observem més d'un locus, pot passar també que no tinguem problema a l'hora de separar cromosomes: L'individu amb genotip A/A i B/B, només permet una possible separació: els al·lels $A-B$ en un cromosoma i $A-B$ en l'altre. Aquests són els dos únics haplotips possibles per aquest genotip. Ara bé, i si l'individu té al·lels diferents en més d'un locus? És a dir, i si l'individu és heterozigot en més d'un locus? Un individu amb genotip A/a i B/b pot donar lloc a dues separacions diferents en cromosomes: $A-B$ i $a-b$, o bé, $A-b$ i $a-B$. Quina de les dues parelles d'haplotips o reconstruccions haplotípiques escollim? Aquí rau el problema de la incertesa.

Recordem que ens centrarem en estudiar haplotips concrets, pertanyents a locus polimòrfics en un sol nucleòtid (SNPs). Considerem el cas en que tinguem dos loci bial·lèlics, és a dir,

locus en que només podem trobar dos tipus d'al·lels. La següent taula explicita les diferents possibilitats de genotips i d'haplotips:

SNP1	SNP2	haplotip 1	haplotip 2
C/C	A/A	C-A	C-A
C/C	G/A	C-A	C-G
C/C	G/G	C-G	C-G
C/T	A/A	C-A	T-A
C/T	**G/A**	**C-A o C-G?**	**T-G o T-A?**
C/T	G/G	C-G	T-G
T/T	A/A	T-A	T-A
T/T	G/A	T-A	T-G
T/T	G/G	T-G	T-G

(3.1)

Pel cas del genotip heterozigot la separació en cromosomes no és immediata , és a dir, no sabem quins dos haplotips porta l'individu. Pel cas en que estudiem m locus, tot aquell genotip amb més d'un locus heterozigot presentarà incertesa pel que fa als seus haplotips. Un genotip amb n loci heterozigots té 2^{n-1} possibles parelles d'haplotips. En cas de tenir 10 loci heterozigots, l'individu podria dur fins a 512 parelles diferents! Necessitem un criteri per triar.

Aquesta manca de coneixement sobre la fase en que es troben els al·lels, sobre el cromosoma al que pertanyen per alguns individus, és un problema d'imprecisió en les dades. Aquest és un tipus de problemàtica que pot ser tractat mitjançant inferència estadística obtenint bons resultats ([94],[95]).

3.3 Mètodes estadístics per l'anàlisi d'Haplotips amb incertesa

La inferència estadística és l'àrea de la ciència que es basa en el procés deductiu d'assolir unes conclusions generals partint d'unes dades, mitjançant mesures quantitatives. Sovint

existeix incertesa associada a aquestes mesures, ja sigui perquè han estat fetes amb imprecisió o bé perquè el procés a estudi s'ha dut a terme sota unes condicions que o bé són desconegudes o bé no ha estat possible controlar completament. En aquest camp, l'eina utilitzada per quantificar aquestes incerteses és la teoria de la probabilitat, on distribucions de probabilitat s'associen a aquestes mesures incertes. Un model estadístic es definirà com l'especificació de distribucions de probabilitat per aquestes mesures incertes (o variables aleatòries) que poden presentar relacions deterministes entre elles.

Al cas que ens ocupa, el dels haplotips, les quantitats mesurables quantitativament que presenten imprecisió són les freqüències haplotípiques atribuïbles a una mostra de genotips donada. Hem de tenir present que per una mostra d'individus amb haplotips directament identificables, la freqüència de cada haplotip es pot calcular fàcilment, fent un recompte dels cops que apareix cada haplotip a la mostra. Però, en cas que a la mostra hi hagi individus amb genotip com el vist a l'exemple 3.1, com es calculen les freqüències haplotípiques en aquesta mostra, si hi ha individus pels que no sabem del cert quina parella d'haplotips duen?

Per donar resposta a aquesta pregunta ens cal fer una revisió dels mètodes existents per estimar freqüències haplotípiques. Ens remuntem a l'any 1990 per recuperar el treball de Clark [11] basat en el principi de la parsimònia. Aquest fou el primer mètode que trobem de reconstrucció haplotípica i no es basa en la inferència estadística. Des d'aleshores fins al moment, s'han desenvolupat d'altres estratègies més acurades, que en molts casos sí que s'han situat en el context estadístic basant-se en la tècnica de la màxima versemblança. En aquest sentit, molts del mètodes han nascut motivats pel desig d'optimitzar els resultats del procés de maximització de la funció de versemblança, que en el cas de les freqüències haplotípiques, com veurem, no és immediat, donada la complexitat de la funció i la quantitat de variables que poden arribar a participar-hi.

De programes que implementen mètodes d'estimació haplotípica aplicables a mostres d'individus no relacionats, n'hi ha una cinquantena que almenys estimen les freqüències hap-

lotípiques. D'aquests, vora una quinzena també resolen algun tipus d'associació entre els haplotips i el fenotip, la majoria d'ells per estudis de cas-control. Els diferents mètodes que existeixen de resoldre la qüestió de l'anàlisi d'haplotips poden ser classificats en dues famílies:

- Els basats en mètodes combinatoris de parsimònia ([11],[96],[97],[98],[99],[100])
- Els basats en mètodes de Màxima Versemblança: Algorisme EM ([12],[101],[102],[16], [103]) i els mètodes Bayesians ([17],[104],[18],[105])

3.3.1 Mètode de la Parsimònia

El primer algorisme que es va crear per fer reconstrucció haplotípica a partir d'informació genotípica es va basar en el principi de la parsimònia. Aquest és un principi filosòfic segons el qual a l'hora de triar entre dues teories possibles és preferible escollir la teoria més simple en comptes de la més complexa o dit amb d'altres paraules, quan dues teories tenen les mateixes conseqüències és preferible triar la teoria més simple. El principi de parsimònia és un dels principis més bàsics en la natura i ha estat aplicat a nombrosos problemes biològics. Aplicat al cas que ens ocupa, el mètode de reconstrucció de la mostra haplotípica basat en el principi de parsimònia té com a objectiu minimitzar el nombre total d'haplotips observats a la mostra i així reflectir mitjançant models genètics simples l'evolució dels haplotips en la població. L'algorisme, utilitzat primerament per Clark, fou molt utilitzat a la pràctica demostrant la seva utilitat ([92],[106],[51]). L'algorisme arrenca llistant tots els haplotips que apareixen amb certesa en la mostra, és a dir, aquells haplotips pertanyents a individus homozigots en tots els loci, o bé només heterozigots en un locus o bé tals que els seus haplotips han estat inferits prèviament. És a dir, per un conjunt de genotips, es construeix un conjunt més petit d'haplotips H. A l'inici, per a cada genotip G es designen un parell d'haplotips en H que expliquen G. Tot seguit, l'algorisme itera mitjançant un mètode que dóna prioritat als haplotips ja observats i que, segons això, assigna parelles d'haplotips als individus incerts. Un cop resolt cada individu (inicialment amb haplotips incerts) els seus haplotips es

consideren ja observats. Es tracta d'un algorisme senzill i fàcil d'utilitzar. El programa que implementa l'algorisme de Clark s'anomena HAPINFREX. És computacionalment ràpid i eficient, i ha estat fet servir freqüentment en recerca.

L'algorisme de parsimònia de Clark té la limitació que la solució depèn de l'ordre en que es consideren els individus, és a dir, de quins haplotips es consideren observats en el moment en que l'algorisme es disposa a solucionar la fase d'un nou individu. Un altre punt en contra és que la base de dades de la qual parteix necessita tenir almenys un individu amb els haplotips no incerts, i això no sempre té perquè existir en dades de caire complex. L'algorisme tampoc assegura que tots els haplotips es resolguin per cada individu de la mostra i que l'assignació sigui la correcta. A més, diferents execucions del programa poden reportar solucions diferents. Finalment, una altra limitació és que l'aplicació del mètode sobre una mostra amb pocs individus no incerts és *NP-Hard* ([11], [96]). Per superar aquestes limitacions es considerà una extensió de pura parsimònia en l'àmbit de la filogènia perfecta.

Mètode de la Filogènia perfecta

Després de l'algorisme de Clark, Gusfield [99] introdueix un model de perfecta filogènia per resoldre el problema de la inferència d'haplotips. El mètode es basa en dues assumpcions. En primer lloc, el model assumeix que per un conjunt d'SNPs estretament lligats, no han existit recombinacions anteriors. De fet, generalment els resultats experimentals i els models genètics segueixen aquesta assumpció. En segon lloc, el model adopta el supòsit estàndard que diu que a cada posició on es dóna un SNP, una mutació pot donar-se com a molt un cop donat que hi ha infinits llocs de mutació. Sota aquestes dues suposicions, els $2 * n$ haplotips d'una mostra de n individus poden ser organitzats en un arbre amb arrel anomenat perfecta filogènia. Cada fulla d'aquest arbre representa un haplotip. Cadascuna de les arestes interiors està marcada per almenys un SNP i cada SNP conté exactament una

sola aresta. Un camí des de l'arrel fins a una fulla, recorre tots els llocs mutants de l'haplotip corresponent a la fulla. La perfecta filogènia troba, donada una mostra de genotips, un conjunt d'haplotips que admeten una perfecta filogènia. Gusfield dissenyà un algorisme que reduïa la qüestió a un problema de teoria de grafs GPPH, però la implementació és massa complexa per ser pràctica. Des de llavors, trobem diverses propostes: una alternativa simple també basada en anàlisi de grafs s'utilitza a DPPH [100]. Donat que les dades empíriques poden violar les assumpcions que necessita aquest mètode, els supòsits són relaxats en la implementació anomenada HAP [107] i també a BPPH [108].

Parsimònia Pura

L'enfocament de la pura parsimònia ha estat també investigat ([98],[109]) per part de la comunitat dedicada a la biologia computacional. Sota aquest criteri, l'objectiu és el de trobar el conjunt mínim d'haplotips diferents que poden resoldre tots els genotips donats. La raó de ser del principi de parsimònia pel problema dels haplotips es basa també en la observació que, a les poblacions d'humans, el nombre d'haplotips diferents observats és molt inferior al de tots els possibles haplotips. A diferència de la filogènia perfecta que compta amb un algorisme d'òptim temps lineal, el càlcul de minimitzar la diversitat haplotípica esdevé en un alt consum computacional. S'ha demostrat [109] que, en teoria, el problema no només no compta amb algorismes de resolució exactes, si no que ni tan sols compta amb algorismes que ho resolguin de manera aproximada. Gusfield [98] va formular el problema utilitzant l'enfocament de la programació lineal, que pot assolir solucions òptimes en conjunts petits. Wang i Xu [96] proposaren un algorisme *Branch and Bound* que demostrà utilitat en problemes pràctics. Aquests enfocaments basats en la parsimònia han estat revisats en detall per Gusfield [110].

A més dels enfocaments discrets, els models estadístics també han estat àmpliament es-

tudiats a la literatura i molt utilitzats als estudis d'associació genètica. El context al que s'engloben és el de la màxima versemblança.

3.3.2 Mètodes basats en la Funció de Versemblança

La majoria de programes que existeixen per resoldre el problema dels haplotips es basen en la funció de versemblança de la mostra 9.1. Els mètodes que exploten la teoria de la versemblança poden ser classificats en els de Màxima Versemblança i en els mètodes Bayesians.

Mètode de la Màxima Versemblança

Segons l'enfocament de la Màxima Versemblança ([12],[101]) les freqüències haplotípiques poblacionals són considerades com a paràmetres desconeguts que necessiten ser inferits. L'objectiu és estimar valors per aquestes freqüències haplotípiques, tals que maximitzin la probabilitat de veure els genotips donats. Si considerem que tots els individus de la mostra de genotips són independents, aleshores la funció de versemblança de la mostra com veurem a 9.1 es pot expressar com el producte de les probabilitats de cada genotip. Alhora, sota l'assumpció de l'equilibri de Hardy Weinberg (aquest treball 1.1), la probabilitat de cada genotip pot ser expressada en funció dels haplotips compatibles amb cada genotip: exactament serà la suma dels productes de cada parell de freqüències haplotípiques, per totes les parelles haplotípiques compatibles amb el genotip. Quan l'estimador màxim versemblant per aquestes freqüències (denotat en anglès *MLEs*) no pot ser obtingut mitjançant mètodes analítics de derivació de la funció de versemblança, el més habitual és utilitzar mètodes numèrics.

L'algorisme EM

L'algorisme més utilitzat de maximització numèrica per obtenir els *MLEs* és l'Algorisme EM, un mètode preferible a d'altres com pot ser el de Newton Rapshon, gràcies al seu

millor cost computacional. Al 1995 tres grups de recerca programaren i publicaren tres programes que implementaven l'algorisme EM: el 3.LOCUS.PAS [102], HAPLO [101] i el MLHAPFRE [111]. La versió original de l'algorisme data de l'any 1977 (Dempster *et al.* [13] i fou dut al context haplotípic per Excoffier i Slatkin al 1995 [12]. Aquest dos autors discutiren els avenços i les limitacions d'aplicar l'algorisme EM a l'anàlisi d'haplotips. Com es pot veure a 3.3.2 l'algorisme EM és un mètode iteratiu que consisteix en alternar dues passes: la passa "E"i la passa "M". Al context de la inferència haplotípica, l'algorisme considera les freqüències com a paràmetres, i la fase de cada individu com a dades *missing*. En aquestes dues parts, l'algorisme incialment considera la funció de versemblança utilitzant uns primers valors pels paràmetres i calcula conjunts d'haplotips que maximitzin les probabilitats a posteriori dels genotips donats. Les estimacions es van actualitzant a cada iteració per arribar a maximitzar la funció de versemblança. L'algorisme itera fins a convergir o fins que assoleix un nombre màxim d'iteracions permeses pel programador. Llavors, per estimar la parella d'haplotips per cada individu, un pot prendre la parella d'haplotips més probable, basant-se en les freqüències haplotípiques que s'han estimat. L'algorisme EM s'ha demostrat precís mitjançant simulacions [112] i produeix estimacions de les freqüències comparables a les obtingudes utilitzant mètodes moleculars ([113],[114],[115]), millorant a mida que augmenta la mida mostral. També s'ha vist que la majoria de l'error de l'algorisme EM és causat per l'error de mostreig ([114],[97]).

Tot i així, l'algorisme EM presenta algunes limitacions importants: Les freqüències per haplotips poc freqüents poden ser estimades erròniament per aquest mètode. A més, teòricament, l'algorisme EM assegura la convergència a un màxim, que pot ser local, però el nombre de variables (és a dir, les freqüències haplotípiques) pot ser exponencialment gran en comparació amb el nombre de locus que s'estudia. Per això, una de les limitacions de l'algorisme EM és el nombre de loci que pot acceptar i també el nombre d'individus. Una implementació directa de l'algorisme EM habitualment no pot resoldre haplotips per mostres de més de 25 SNPs ([12],[17]). Com ja s'ha esmentat, el fet que l'algorisme EM

pugui convergir a un màxim local en comptes de fer-ho a un de global, comporta que els usuaris hagin de repetir diverses execucions amb diferents llavors ([12],[112] i Celeux and J. Diebolt, [14]). A més, en general l'algorisme EM no retorna les estimacions de les variàncies pels estimadors *MLEs*, a no ser que el nombre de loci sigui petit. D'altra banda, l'algorisme EM necessita suposar equilibri de Hardy-Weinberg a la mostra.

Diverses variants de l'algorisme EM han estat implementades amb l'objectiu d'intentar solucionar les limitacions mencionades. Qiu, Niu i Liu ([15],[116],[117]) intenten posar solució el problema dels màxims locals mitjançant la creació de l'algorisme PL-EM (Partition Ligation EM). Aquesta tècnica diríem que és del tipus *Divide and Conquer*. L'algorisme divideix la regió en blocs d'SNPs i després utilitza l'algorisme EM sobre cada bloc per reconstruir-hi localment els haplotips. En una segona fase, lliga els haplotips resultants de cada part, per obtenir-ne de sencers, tot utilitzant de nou l'algorisme EM. Qiu et al. adverteixen que el fet de mirar el genotip localment pot dur a solucions no òptimes donat que alguns haplotips obtinguts considerant només alguns SNPs poden tenir una probabilitat molt baixa, en canvi mirats en conjunt amb la resta de SNPs poden tenir més pes. Una altra adaptació de l'algorisme EM és la que fa en David Clayton. Sota el nom de SNPHAP [16], trobem implementat un algorisme EM que s'aplica sobre la mostra d'SNPs, però d'una manera peculiar: els SNPs es van considerant d'un en un, s'afegeixen a cada pas. Durant el procés, els haplotips amb baixa probabilitat són descartats, la qual cosa també pot dur a solucions errònies. També el programa THESIAS de D.Tregouet ([20],[118]) proposa una variant estocàstica de l'algorisme EM que resol alguna d'aquestes limitacions. Tot i que també té un màxim de SNPs analitzable, és aplicable a grans bases de dades pel que fa a nombre d'individus.

Tècniques Bayesianes

A diferència del mètodes basats en estadística freqüentista que tracten els paràmetres com punts desconeguts en una espai de paràmetres, els Bayesianistes consideren aquests

paràmetres com a variables aleatòries. L'objectiu de la inferència Bayesiana és, donada la observació d'unes dades, estimar la distribució de probabilitat a posteriori pels paràmetres d'interès havent assumit un coneixement previ sobre aquests paràmetres abans d'observar les dades. Aquesta incorporació que fan els mètodes Bayesians a diferència de la resta de mètodes, serveix de guia per la inferència d'haplotips no observats [104]. Les estimacions puntuals poden alhora ser extretes prenent el valor de la mitjana de la distribució posterior, així com també la variància i qualsevol estadístic que sigui calculable partint de la distribució de probabilitat (mediana, quartils, etc). El càlcul d'aquesta probabilitat a posteriori es duu a terme mitjançant el teorema de Bayes exposat a 7.1. Aquesta fórmula involucra el valor de la probabilitat total del genotip, al qual intervenen integrals multidimensionals o la suma d'un nombre exponencial de termes que en molts casos fa el problema intractable. Per resoldre aquesta qüestió és molt avantatjós la utilització de tècniques de Markov Chain Monte Carlo.

S'han proposat diferents aplicacions Bayesianes per resoldre el problema d'estimació de la mostra haplotípica partint de dades de genotips poblacionals. La tècnica numèrica més utilitzada és la Gibbs Sampling ([17],[18],[119],[104],[111],[120]) explicada en aquest treball a 8.3.4. Els mètodes Bayesians poden ser subdividits en dues subclasses: els simples i els que es basen en teoria coalescent. Els mètodes simples no fan cap assumpció sobre la història de les recombinacions per les poblacions de les quals s'han extret les mostres d'individus a estudi. Alguns dels programes Bayesians simples són l'HAPLOTYPER i l'HAPLOREC. A l'HAPLOTYPER Niu *et al.* [18] utilitzen inferència Bayesiana per fer reconstrucció haplotípica. El seu treball es basa en aplicar la Gibbs Sampling, considerant com a distribució a priori per les freqüències genotípiques una distribució de Dirichlet. L'algorisme de Niu et al. parteix d'una assignació inicial de freqüències haplotípiques. A cada iteració, primer es mostreja una parella d'haplotips compatibles amb el genotip de cada individu de la mostra, i després s'actualitza les freqüències haplotípiques en funció de l'assignació feta a cada in-

dividu. Per la seva part, els autors de l'HAPLOREC implementen un mètode Bayesià que utilitza el mètode de Markov Chain de llargada variable [121].

Els mètodes basats en teoria coalescent essencialment prenen les similituds entre haplotips, assegurant que els haplotips que es generen són similars als que ja han estat generats. Aquest tipus d'algorismes inclou un programa àmpliament utilitzat, creat per Stephens et al. [17] sota el nom de PHASE. Aquests autors proposen un mètode de Markov Chain Monte Carlo per reconstruir els haplotips d'una mostra de genotips. Els autors implementen una Gibbs Sampling i construeixen una cadena de Markov per les freqüències haplotípiques. Així, a cada pas de l'algorisme cal mostrejar de la distribució condicional que té com a variable la freqüència haplotípica d'un individu concret, considerant sabuts els haplotips per la resta d'individus. Aquesta distribució, per la majoria de models mutacionals és desconeguda. Stephens et al. proposen una distribució que aproximi el model mutacional general. A la pràctica, l'algorisme comença amb una solució arbitrària d'haplotips donada una mostra de genotips i iterativament actualitza una mostra aleatòria d'individus assumint que tota la resta d'individus tenen assignada la parella d'haplotips correctament. El programa presenta una segona versió anomenada FastPhase [122] on es milloren les característiques computacionals del programa. Aquesta versió incorpora un algorisme millorat pel que fa a precisió i una estratègia de P-L per millorar el temps d'execució.

Un altre programa que també es basa en el model coalescent és l'ARLEQUIN que utilitza una definició més simplificada de similitud entre haplotips en un enfocament també iteratiu ([111],[123]).

Els mètodes de Niu i el de Stephens difereixen bàsicament en la distribució prior que consideren. Stephens tria una prior que aproxima el model coalescent mentre que Niu tria una distribució de Dirichlet. Sota el model coalescent, els haplotips mostrejats tendeixen a ser similars als haplotips ja mostrejats, una propietat que ja havia estat utilitzada en l'algorisme de Clark. Alguns experiments [104] han demostrat que les estimacions basades en el model coalescent són més acurades que les basades en la priori de Dirichlet, per dades que

responguin a aquest model.

En aquesta secció hem descrit dues classes principals de mètodes per dur a terme inferència haplotípica per poblacions d'individus no relacionats. Els mètodes descrits han estat triats en representació de cada categoria, però per exemple, existeixen diverses variants del mètode de Clark que no han estat exposades. Salem et al. publicà una revisió de mètodes al 2005 [124] de la qual hem extret i exposat la relació de programes existents fins a aquell any. Fins el 2008 hem utilitzat la revisió actualitzada que es troba al capítol 6 de Feng et al. [125].

Des de llavors, diversos estudis han demostrat que els algorismes que existeixen per fer estimació haplotípica són acurats ([124],[94],[95]). Malgrat tot, la inclusió d'informació familiar pot reduir l'ambigüitat haplotípica i millorar la precisió de la inferència haplotípica [126].

3.4 Eines per fer inferència sobre Haplotips incerts

Com es pot observar a la taula que es troba a l'apèndix B, existeix un conjunt molt ampli de programes que resolen l'estimació de les freqüències haplotípiques. L'avaluació d'un conjunt tan ampli de programes és molt complicada, degut a la varietat dels mètodes utilitzats, les mesures d'exactitud dels algorismes que es consideren i les característiques concretes de cada programa. A més, les característiques específiques de cada conjunt de dades, ja siguin determinades molecularment o simulades, determinaran en gran mesura l'èxit d'execució del programa.

El principal desavantatge de tots els programes que infereixen la mostra d'haplotips és que una proporció d'haplotips inferits pot ser incorrecta ([119],[104],[111],[120]). Per exemple, haplotips que només apareixen un cop a la mostra poden no ser mai resolts correctament mitjançant aquests mètodes. Aquesta incertesa en la reconstrucció haplotípica pot dur a

una pèrdua de poder a l'hora de testar l'associació entre els haplotips i una malaltia. Amb l'objectiu de quantificar la imprecisió d'aquests algorismes, diversos estudis han comparat haplotips inferits respecte haplotips determinats molecularment als mateixos gens. Aquest estudis demostren que la majoria d'aquests algorismes poden estimar les freqüències per la majoria dels haplotips eficaçment ([121],[127],[128]) per bases de dades amb poc o cap error de genotipatge [1]. Ara bé, la precisió dels haplotips assignats a cada individu varia. És particularment complicat assignar al·lels estranys a un cromosoma [119] i alguns estudis demostren que la precisió de la inferència sobre haplotips és major per al·lels més freqüents que pels estranys [129].

Com ja s'ha vist a la secció (3.2) cadascun dels mètodes teòrics exposats presenta alguna limitació. De mètodes i algorismes n'hi ha diversos, essent la família més utilitzada la dels mètodes basats en inferència estadística i en particular, en la tècnica de la màxima versemblança. L'algorisme EM i els mètodes Bayesians serien les dues subfamílies més utilitzades, havent donat aquests darrers mètodes els millors resultats, pel que fa a convergència, nombre d'SNPs acceptat i valors que retornen. A més, en conjunt disminueixen les limitacions de l'algorisme EM que com hem vist a la secció 3.3.2 són diverses. Les tècniques d'integració numèriques desenvolupades als darrers anys han fet que a més les tècniques Bayesianes siguin factibles computacionalment. Pel que fa a la precisió dels resultats, la literatura [130] constata que la precisió del programa pel que fa a les reconstruccions i estimacions de freqüències haplotípiques que retorna depèn molt del conjunt de dades on s'aplica. En general, els programes basats en tècniques Bayesianes, EM o filogènia, tenen un rendiment similar, ja sigui amb dades simulades o reals. Pel que fa a les assumpcions, la majoria de programes requereixen HWE i no fan assumpcions sobre LD. Pel que fa al tractament dels missings, els programes que accepten dades amb missings sovint assumeixen que els missings es distribueixen de manera aleatòria. D'aquesta manera es poden introduir haplotips falsos a la mostra [111].

Donat que tant la precisió com el poder de les anàlisis d'associació es veuen afectats

pels valors missings, alguns programes incorporen la incertesa del genotipatge en la inferència haplotípica [131]. La majoria de programes, però, no accepten dades amb missings. Cal tenir present que el fet d'incorporar-los té conseqüències computacionals indesitjables degut a que s'augmenta considerablement la complexitat dels problemes haplotípics. Pel que fa a les qüestions computacionals, no tots els programes estan disponibles en totes les plataformes ni tots els programes són d'accés lliure. D'altra banda, alguns programes per bases de dades grans necessiten processadors d'alt rendiment per a que les execucions siguin computacionalment possibles. La interfície és una component molt important pel que fa a l'ús d'aquests programes. La majoria de programes s'executen via comandes de prompt, una interfície clarament poc còmode i amigable. Pel que fa al temps d'execució dels programes, els programes que treballen amb locus multial·lèlics sovint tenen associats uns temps d'execució excessivament llargs. Tant el nombre d'individus com el nombre de loci són components molt importants a l'hora d'avaluar un programa d'anàlisi d'haplotips. A la taula de l'apèndix es poden observar amb detall els límits sobre la mida mostral i el nombre de loci. A mida que el nombre d'individus creix, millora la precisió de la majoria de programes. Els programes EM accepten un nombre màxim de loci inferior als Bayesians ([12],[112],[17]). Alguns programes consideren tècniques de *Divide and Conquer* que permeten assumir un nombre de SNPs superior ([18],[104],[111],[121]). El nombre d'individus no acostuma a generar problemes sinó millores en precisió, tot i que provoca un augment en el temps d'execució. D'altra banda, l'augment en el nombre de loci pot dur a problemes haplotípics computacionalment irresolubles.

3.5 Mètodes estadístics per l'anàlisi d'associació amb Haplotips

Com ja hem comentat, diversos estudis han provat que els mètodes basats en haplotips poden ser més potents i precisos alhora d'analitzar l'associació entre la malaltia i la genètica de l'individu ([89],[132]). Una diversitat de mètodes han estat proposats per resoldre la qüestió

de les anàlisis d'associació. En funció de les dades, aquests mètodes poden classificar-se segons si són aplicables sobre mostres d'individus no relacionats o relacionats. En aquest treball només considerem estudis amb individus no relacionats, així que la revisió de mètodes que presentarem serà per aquest tipus de mostra.

En aquesta secció ens centrarem en explicar els dos tipus principals de mètodes que permeten dur a terme anàlisi d'associació amb haplotips: els mètodes basats en *scores* estadístics i els mètodes englobats en el marc dels models de regressió.

3.5.1 Mètode de les puntuacions estadístiques (*Scores*)

Si la informació haplotípica és sabuda, existeixen molts mètodes que poden utilitzar-se , ja sigui per comparar les freqüències dels haplotips entre casos i controls, utilitzant molts dels mètodes ja desenvolupats per la comparació de la freqüència d'al·lels [133], o per realitzar l'anàlisi en el context de la regressió, on els haplotips poden ser tractats com a variables categòriques. No obstant això, com s'indica en les seccions anteriors, la informació sobre la fase haplotípica sol ser desconeguda i ha de ser estimada. Els mètodes tradicionals d'associació d'haplotips per als estudis de casos i controls acostumen a utilitzar proves de bondat d'ajust per determinar si la distribució dels haplotips entre els casos i els controls són les mateixes. Normalment, és possible la construcció d'un TRV. Aquest enfocament té algunes limitacions [133].

1. Quan hi ha molts haplotips, hi ha molts graus de llibertat i el poder per detectar associació pot ser feble. A més, amb poques dades, les estimacions per als haplotips rars poden ser problemàtiques i la distribució nul·la pot no seguir una distribució χ^2 com es requereix.
2. No es pot ajustar per altres variables.
3. Només funciona per a variables resposta qualitatives.
4. Assumeix HWE per als parells d'haplotips.

S'han proposat diverses vies per abordar aquestes limitacions.

Haplotips compartits i *clusters*

Intuïtivament, el nombre d'haplotips es pot reduir si s'agrupen alguns haplotips similars entre ells. Molts dels mètodes estadístics que s'han proposat es basen en la recerca de les similituds entre haplotips dins dels casos en comparació amb la observada dins dels controls ([134],[135]). La idea inicial dels haplotips compartits fou de Te Meerman i Van Der Meulen [136] que varen proposar un estadístic sobre haplotips compartits anomenat HSS (*Haplotype Sharing Statistic*) basat en la variància de les longituds dels haplotips compartits que es trobaven localitzats al voltant dels haplotips de la mostra de casos.

Les similituds entre haplotips proporcionen una via natural pe definir grups (o *clusters*) d'haplotips, que ofereixen una solució prometedora a les dificultats que provoca la presència d'alguns haplotips. L'agrupament d'haplotips pot augmentar l'eficiència de l'anàlisi d'haplotips utilitzant un petit nombre de grups d'haplotips que poden reduir els graus de la llibertat i alhora reduir els efectes que poden provocar els haplotips rars. Com que els mètodes de clustering tenen en compte el LD entre múltiples marcadors, poden tenir una bona potència per detectar gens predisposants ([134], [137]). Tzeng i col·laboradors a [138] demostraren que per malalties comunes, els tests d'haplotips compartits poden ser més potents que els de bondat d'ajust, però pel cas d'haplotips rars, passa exactament el contrari. A més, també veieren que el poder dels dos enfocaments millora agrupant de manera apropiada els haplotips rars.

Cal que tinguem present que les tècniques de compartir haplotips i d'agrupament no pertanyen a cap test o mètodes. Per això les veurem de nou més endavant.

Tests Estadístics No-lineals

Zhao et al. a ([139],[140]) proposen millorar el poder de l'estadístic 3.5.1 utilitzant transformacions no lineals que amplifiquin les diferències de les freqüències haplotípiques entre casos i controls, donat que creuen que aquesta és la clau. I ho demostraren a [140], veient

que no només s'incrementa el poder per captar associacions, sinó que a més, el test no incrementa els falsos positius.

Estadístics de puntuació provinent de models de regressió

Schaid et al. [133] conclouen que els mètodes d'anàlisi d'associació amb haplotips basats en els models lineals generalitzats (GLM) aporten una via per construir estadístics *Score* per a la hipòtesi nul·la de no efecte haplotípic. Els estadístics construïts segons aquest criteri poden ser ajustats per d'altres covariables i acceptar fenotips continus a més dels binaris. Aquest mètodes els explicarem a la següent secció dedicada als models de regressió.

3.5.2 Models de Regressió per Haplotips incerts

Com hem vist, l'estimació de les freqüències haplotípiques usualment no és el resultat de principal interès. L'objectiu de la recerca serà qui marcarà quines són les següents anàlisis a realitzar. L'anàlisi de regressió és un marc àmpliament utilitzat en els estudis d'associació amb haplotips per les avantatges que ofereix. Els haplotips jugaran el paper de factor de risc del model, que podrà ser ajustat per covariables i per termes d'interacció. Ara bé, a la pràctica habitual, en la majoria de casos els haplotips no poden ser inferits sense ambigüitat, sigui quin sigui el mètode utilitzat. Així doncs, tots els mètodes de reconstrucció de la mostra haplotípica presenten un cert grau d'error en les assignacions d'haplotips ([141],[103],[142]). Donada la incertesa que com hem vist comporta la informació haplotípica, abans de dur a terme una anàlisi haplotípica amb models de regressió caldrà decidir com es tractarà la incertesa donat que si aquesta incertesa s'ignora en les anàlisis posteriors, les estimacions dels coeficients dels models poden resultar esbiaixades [143],[144]. Només en situacions en que els haplotips inferits tenen gran fiabilitat, els biaixos en les estimacions desapareixen i poden fer-se servir directament anàlisis convencionals [142].

Diverses estratègies han estat proposades per incorporar els haplotips inferits quan l'anàlisi

d'associació es fa sobre dades genotípiques de fase incerta. En aquesta secció revisarem els mètodes d'anàlisi sota l'enfocament de la regressió.

Tractament de la incertesa haplotípica als models de regressió

Un enfocament habitual per tractar la incertesa és el d'utilitzar la parella d'haplotips més probable per cada individu en les anàlisis subseqüents. Aquesta manera de procedir implica considerar els haplotips com si haguessin estat observats.

Diversos estudis ([145],[146],[147],[130],[68],[142], [148],[19]) han demostrat que aquest tractament en dues passes independents no només comporta la pèrdua d'informació rellevant, si no que també introdueix errors de mesura i indueix al biaix en les estimacions dels efectes atribuïbles als haplotips. Aquest biaix es fa encara més palès quan la mida de l'efecte és gran o bé quan la incertesa haplotípica a la mostra és alta ([143],[144]).

Una manera intuïtiva d'intentar resoldre aquest problema és utilitzar totes les possibles parelles d'haplotips consistents amb el genotip observat ([149],[150],[133],[151],[68],[142]).

Una via força més potent és estimar les freqüències haplotípiques i els efectes associats als haplotips de manera simultània amb l'objectiu d'obtenir una millor eficiència en l'estimació dels paràmetres, com es pot veure a qualsevol d'aquestes publicacions ([146],[20],[19],[152]).

La majoria d'aquests mètodes es basen en l'ús d'una versemblança prospectiva ([149],[150],[153],[133],[151],[154],[68], [142]).

Models de Regressió

Lake et al. a [150] explicita la funció de versemblança conjunta que permet l'estimació conjunta de les freqüències d'haplotips i els paràmetres del model de regressió. Zhao et al. [142] utilitzen equacions d'estimació basades en equacions *score* derivades de versemblances prospectives per estimar els paràmetres d'una regressió Logística, considerant com a hipòtesi malalties rares i independència entre els haplotips i variables ambientals. Per

estimar les freqüències haplotípiques que són necessàries per avaluar les *prospective score equations* utilitzen un algorisme EM similar al proposat per Excoffier and Slatkin [12]. Sota l'assumpció d'independència entre els gens i els factors ambientals, la incorporació de factors ambientals en aquest mètode és directa ([150],[142]). Quan la exposició a un factor ambiental extern no és directament controlada pel comportament propi del mateix individu, l'assumpció d'independència és probable que se satisfaci; ara bé, Lin i col·laboradors a [146] fan constar que aquesta assumpció no es dóna a la pràctica i que a més, no és estadísticament eficient. Malgrat tot, encara avui en dia no s'ha aclarit completament aquesta qüestió.

Donat que en general als estudis de cas-control els casos esta sobrerepresentats, les estimacions de les freqüències haplotípiques poden resultar esbiaixades en favor de la hipòtesi alternativa si no es té especial cura amb aquesta qüestió [133]. Aquest fet pot provocar que l'estimació dels efectes també esdevingui esbiaixada. Aquest biaix, induït per la pròpia determinació dels haplotips, no es dóna quan la fase de les dades genotípiques és coneguda [133]. El motiu és que la distribució de les covariables és no paramètrica en aquest cas [146]. La magnitud del biaix dependrà de la precisió amb que s'hagin estimat els haplotips. Aquesta és una limitació dels mètodes que utilitzen mètodes de versemblança prospectiva. Per resoldre aquest fet, Zhao et al. [142] proposa utilitzar només controls per estimar les freqüències d'haplotips. Això només pot funcionar per a les malalties rares i el mètode podria produir biaixos substancials per als paràmetres del model quan el supòsit subjacent de malaltia rara és violat [155]. Stram et al. [154] proposa utilitzar mostreig de ponderacions basat en la prevalença de la malaltia en la població per corregir les estimacions esbiaixades. Epstein i Satten a [156] proposen una versemblança retrospectiva que també permet l'estimació conjunta de les freqüencies haplotípiques i dels paràmetres del model. La funció de versemblança és el producte de les distribucions multinomials de les dades genotípiques observades per casos i per controls independentment, condicionals a ser cas o control. Aquest enfocament requereix suposar HWE només pels controls, tot i que utilitza freqüències

per casos i controls. Aquesta via s'ha demostrat igual o millor que la prospectiva proposada anteriorment [157].

El fet que la versemblança retrospectiva impliqui paràmetres problemàtics en relació a l'especificació de les distribucions dels factors ambientals, la incorporació dels mateixos (i dels factors d'interacció) és complicada en aquest enfocament [155]. En aquest article Spinka et al. estenen la proposta de Chatterjee i Carroll [158] incorporant factors genetics i ambientals, i acceptant la presència de dades *missing* als genotips. Utilitzant un algorisme EM aconsegueixen un procediment d'estimació de paràmetres a relativament senzill que reporta resultats robustos menys sensible a la pèrdua de HWE i a la independència entre els factors ambientals i els genètics.

Sinha et al. [159] utilitzen una versemblança condicional per resoldre la qüestió, considerant també únicament HWE sobre la mostra de controls. Per dur a terme l'estimació conjunta dels paràmetres del model logístic, ells proposen l'algorisme ECM (*Expectation and Conditional Maximization*) i l'apliquen a estudis de cas-control aparellats.

Una característica molt atractiva de l'ús dels GLM és que aquests models accepten diferents tipus de respostes. Lin i Zeng [146] , i Iniesta i Moreno [160] proposem un marc teòric més ampli i general per dur a terme l'anàlisi d'associació basat en models GLM i mètodes de versemblança, tals que poden ser utilitzats en tots els dissenys d'estudi més habituals (cros-seccional, cas-control i cohorts) i on els diversos fenotips (incloent els binaris, els quantitatius i de supervivència) són tractats de manera similar. Els models de regressió que presenten permeten avaluar els efectes associats als haplotips, així com les interaccions entre gen i factors ambientals. Alhora, els models inclouen diferents mecanismes genètics d'herència (models recessiu, dominant, additiu i codominant). Lin and Zeng [146] povaren la identificació dels paràmetres del model, i la consistència, la normalitat asimptòtica, i l'eficiència dels estimadors màxim-versemblants sota certes condicions. Tot i així, aquesta tècnica deixa diverses qüestions per resoldre com és la incorporació de valors missing a

les anàlisis o l'estimació dels efectes atribuïbles a haplotips de baixa freqüència. Durant els darrers anys, en aquesta tesi hem desenvolupat precisament un mètode d'estimació conjunta de freqüències haplotípiques i els efectes associats en el marc dels GLM que intenta posar solució a algunes d'aquestes qüestions. Fins a dia d'avui, l'anàlisi d'associació amb haplotips és un camp d'investigació obert que es troba situat en la utilització de models GLM i la millora en les estimacions dels paràmetres que hi prenen part ([161],[162]).

Haplotips compartits i *clusters*

Com hem comentat amb anterioritat, la tècnica dels haplotips compartits/agrupats no pertany a cap mètode específic. Diversos mètodes estadístics han estat proposats per dur a terme l'anàlisi d'haplotips incorporant als models de regressió informació referent a *clusters* d'haplotips per tal de reduir la dimensionalitat del problema ([135],[163],[164],[165], [166]). L'extensió de l'enfocament cladístic al camp dels GLM va permetre la incorporació de la incertesa haplotípica. La solució contempla utilitzar el mètode de *clusters* jeràrquic habitual per crear un arbre jeràrquic d'haplotips. Com a resultat s'acaba generant un arbre que sorgeix d'anar retallant les branques tals que no ajusten bé un model logístic. [163] també incorporen mètodes probabilístics de clustering als mètodes GLM que havien presentat Schaid i col·laboradors amb anterioritat [151], així com també se sumen d'altres autors recentment ([164],[165],[166]). Aquest mateix any ha aparegut un paquet d'R anomenat SHARE que presenta un mètode d'estimació d'efectes en estudis cas-control mitjançant *clustering* DAI et al. Aquest paquet també pot ser utilitzat per identificar els SNPs que conformen els haplotips que millor discriminen la mostra. Aquest mètode és adient sempre i quan l'objectiu de l'estudi no recaigui en estimar l'associació amb haplotips poc freqüents. La qualitat dels resultats que ofereixen aquests mètodes és encara discutida a la pràctica. Tot i així són força utilitzats per reduir els graus de llibertat i eliminar els haplotips rars i els haplotips que no difereixen entre casos i controls com es pot veure a ([167],[168]). Els principals desavantatges són:

- habitualment aquests mètodes són incapaços de detectar variants rares amb grans efectes, degut a que els haplotips estranys no són mantinguts en l'espai de *clusters* d'haplotips [163].
- La majoria d'ells no treballen bé en estudis de cas-control de malalties complexes [133]
- Aquests mètodes depenen fortament de l'esquema d'agrupament utilitzat, és a dir, de les mesures de similitud utilitzades. Es necessiten més treballs en aquest sentit per definir el millor tipus de mesura de similitud entre haplotips. Volem remarcar que l'ús de *clusters* no implica forçosament millores en les anàlisis. El fet d'agrupar prèviament els haplotips segons l'algorisme d'arbres jeràrquic no millora el poder de detectar associació en comparació amb utilitzar regressió Logística sense agrupació d'haplotips, excepte en cas que les dades presentin patrons de LD molt particulars.

Construcció d'estadístics *Score*

Una avantatge afegida de la utilització dels GLM és que proporcionen una via per construir estadístics d'*score* per testar la hipòtesi nul·la de no associació [133]. Aquest estadístic mesura la covariància dels residus del model GLM que ajusta només les covariables ambientals amb els haplotips esperats. Els pesos que s'utilitzen pels haplotips esperats són les probabilitats posteriors del parell d'haplotips donats els genotips observats [133]. L'estadístic és eficient en tant que es pot obtenir per simulació, que és un mètode habitualment més robust que utilitzar teoria asimptòtica, sobretot en mostres petites.

Una tasca pendent que al llarg d'aquests anys ha presentat dificultats en la majoria dels programes ha estat l'estimació de l'associació per haplotips de baixa freqüència. En l'actualitat aquest és un tema d'estudi com es mostra a ([161],[162]) on els autors utilitzen versemblances retrospectives en l'estimació d'efectes haplotípics per estudis de cas-control.

En aquesta secció hem introduït alguns dels mètodes més representatius basats en els models de regressió. Aquest enfocament basat en la regressió ofereix un conjunt d'avantatges [133] que els constitueix una part primordial en les anàlisis haplotípiques. Ara bé, com hem pogut observar, la majoria de metodologia ha estat destinada a millorar els càlculs per estudis de cas-control.

A la següent secció presentem les implementacions informàtiques que permetran dur a terme a la pràctica alguns dels algorismes teòrics que hem exposat.

3.6 Eines per fer l'anàlisi d'associació amb haplotips

A la taula 3.2 trobem algunes de les aplicacions més utilitzades per fer anàlisis haplotípiques. A la taula s'explicita quin és l'algorisme dut a terme per estimar els haplotips, el tipus d'anàlisi que accepten i amb quines variables fenotípiques tracten. Com es pot observar, existeixen força més programes que estimin haplotips que no pas programes que també estimin efectes.

Nom prog	Algorisme	Caract.	Resposta
FASTEHPLUS	EM	Test LD Test dif	Cas-control
GENECOUNTING	EM	Test dif	Cas-control
HAP	Filogènia Imperfecta	Test dif	Cas-control
HAPLO.STATS	EM	GLM + covar	Cas-control Ordinal Poisson
HAPASSOC	EM	GLM + covar	Cas-control Ordinal Poisson Gamma
HPLUS	EM-PL	Test dif + covar	Cas-control
PHASE	MCMC	Test permutació	Cas-control
THESIAS	S-EM	Test dif + covar	Cas-control Supervivència
WHAP	EM	GLM + Test permutació	Cas-control
BEAGLE	Clustering	Test permutació	Cas-control

(3.2)

A la taula que es troba a l'apèndix B s'amplia la taula anterior i també s'hi afegeix altres programes implementen testos d'hipòtesi. De tots aquests programes, dos d'ells destaquen per resoldre associació amb diversos fenotips i per incorporar la incertesa haplotípica a l'hora d'estimar l'efecte dels haplotips. Tots dos però, tenen com a problemàtica les limitacions de l'algorisme EM. Un és l'Haplo.Stats i l'altre és el THESIAS.

Haplo.Stats

L'haplo.stats és un conjunt de funcions implementades en l'entorn del programari estadístic R que té com a principal utilitat l'anàlisi d'haplotips indirectament mesurats. Les

anàlisis estadístiques que es duen a terme assumeixen que tots els individus són no relacionats i que a la mostra hi ha la possibilitat que hi hagi individus ambigus pel que es desconeix la fase de lligament dels seus marcadors genètics. Els marcadors genètics s'assumeixen com a codominants.

El paquet haplo.stats utilitza l'algorisme EM per estimar les freqüències haplotípiques i les associacions, que són estimades de manera simultània en considerar una funció de versemblança conjunta. A [151] Schaid et al expliquen el mètode basat en l'algorisme EM que permet estimar efectes per haplotips en relació a fenotips binaris, ordinals i quantitatius, i que alhora també ofereix la possibilitat d'incorporar d'altres variables no genètiques d'ajust. Aquest mètode aplicable a estudis transversals i de cas-control, no reconstrueix la mostra haplotípica i a posteriori realitza l'anàlisi d'associació sinó que en el propi algorisme incorpora la incertesa haplotípica com a dada faltant a tractar alhora que estima l'efecte dels haplotips sobre la resposta. El programa accepta valors *missing* i resol l'associació per fenotips discrets i continus, tot i que no contempla el cas de fenotip de supervivència. En estar basat en models lineals generalitzats, l'haplo.stats accepta l'ajust per covariables a l'hora de testar l'associació, permetent controlar els efectes confusors d'altres variables clíniques o ambientals, així com també és possible considerar termes d'interacció entre els haplotips i aquestes variables. L'algorisme que utilitza està basat en el del programa SNPHAP de David Clayton [16]. El temps d'execució és força òptim i la preparació de les dades no durà excessiu problema per aquells acostumats a utilitzar l'entorn R. Pels que no ho estiguin, sempre es pot optar per la opció SNPstats, una aplicació via web que utilitza les funcions d'aquest paquet i que és de fàcil us [169].

Les limitacions d'aquest programa són, per un cantó, la manca d'alguns valors en els resultats que retorna. El mètode no retorna per exemple variàncies per les estimacions de les freqüències haplotípiques, ni intervals de confiança. Tampoc retorna un mostreig per les freqüències ni pels paràmetres, donat que no és Bayesià. El programa presenta els prob-

lemes de convergència propis de l'algorisme EM. I pel que fa als fenotips amb que treballa, presenta la limitació de no permetre dur a terme l'anàlisis de supervivència.

Thesias

El programa THESIAS (*Testing Haplotype EffectS In Association Studies*) també duu a terme anàlisis d'associació amb haplotips com el seu nom indica. El tipus de mostra al que s'adreça també és d'individus no relacionats i els mètodes que implementa es basen en la màxima versemblança. En aquest cas, David Tregouet i col·laboradors, autors de THESIAS, proposen un algorisme EM modificat que anomenen SEM *(Stochastic EM)* com es descriu a [20]. Aquest programa també tria l'opció de simultanietat a l'hora de considerar la incertesa haplotípica de la mostra en l'anàlisi d'associació amb el fenotip d'interès. La implementació actual resol anàlisis amb fenotips discrets i continus, i accepta anàlisis de supervivència. Alhora també permet l'ajust per covariables i per termes d'interacció. El temps d'execució és superior al de l'Haplo.Stats, variant en funció del tipus de fenotip que es consideri i de la mida de la base de dades.

Les limitacions d'aquest programa són les pròpies de l'algorisme EM. A més, la interfície en java de THESIAS no és gaire amigable i no existeix execució via web que la millori. L'única alternativa és l'execució en mode *bathc* que pot resultar encara més farragosa. A més, una altra incomoditat de THESIAS és que l'usuari ha de forçosament executar el programa dos cops si desitja per estimar efectes, havent d'actualitzar paràmetres a mitja execució.

4

Què podem aportar a la metodologia Haplotípica?

L'àrea de la inferència i l'anàlisi dels haplotips ha avançat molt en la darrera dècada com a resultat dels grans esforços dedicats. Tot i així, segueixen encara sense resoldre's algunes qüestions complexes. La gran majoria dels programes avaluats al capítol anterior no ofereixen la possibilitat de dur a terme una anàlisi d'associació amb haplotips per diferents fenotips. Així com la qüestió de l'estimació de les freqüències haplotípiques està força ben resolta, pel que fa a l'anàlisi d'associació encara es poden millorar molts aspectes. El tipus de fenotips que accepten els programes sovint es limita al binari, i en molts casos no ofereixen quantificació de la magnitud de l'associació entre els haplotips i el fenotip. També cal destacar que alguns d'aquests mètodes fan una estimació no simultània dels efectes haplotípics, partint d'una prèvia imputació haplotípica per estimar els efectes o bé considerant tots els possibles haplotips com a variable de risc en un model de regressió amb pesos per cadascun dels haplotips. Com ja hem comentat, és un fet acceptat que l'estimació simultània d'haplotips i efectes és la millor via d'anàlisi.

L'eficiència de tots els mètodes exposats en aquesta introducció depèn en gran mesura de triar la llargada "correcta"pels haplotips. Si els haplotips són massa llargs incloent massa marcadors, els haplotips estaran composats per massa al·lels, donant lloc a un nombre excessiu de configuracions haplotípiques que poden diluir els senyals d'associació amb la malaltia a estudi [133]. Tot i que s'han proposat diversos mètodes per tractar aquesta qüestió, com els blocs d'haplotips, encara avui en dia no existeixen solucions òptimes. La

majoria dels mètodes d'anàlisi haplotípic (incloent la inferència estadística) compten amb l'assumpció de HWE, de genotips sense missings o de missings aleatoris, i també assumeixen la no existència d'errors de genotipatge, malgrat que aquestes assumpcions poden no donar-se a la pràctica. Donat que les poblacions humanes no solen ser resultats d'aparellaments aleatoris, l'assumpció de HWE ha de ser avaluada amb cura en l'anàlisi haplotípica. Inclús amb l'avenç tecnològic, és comú que els estudis genètics hagin de tractar amb genotips amb valors *missings* i amb errors de genotipatge ([126],[77]). Tot i que diversos estudis han detectat que el fet d'ignorar els genotips amb *missings* provoca un decrement en la precisió de les estimacions haplotípiques ([170],[126]), la majoria dels mètodes actuals no els tenen en compte. Aquesta és una altra qüestió que necessita més investigació. Existeixen d'altres temes rellevants, com els haplotips rars, que encara necessiten de nous mètodes per ser tractats correctament.

Per tal de millorar l'eficiència dels mètodes haplotípics en els estudis genètics, caldria seguir treballant en un seguit de qüestions metodològiques que resten per resoldre. És per això que en aquesta tesi ens plantegem la creació d'una eina d'anàlisi d'associació emmarcada en el context dels models GLM basada en estadística Bayesiana. Així com les tècniques Bayesianes han funcionat molt bé per l'estimació de les freqüències haplotípiques [122], creiem que tècniques similars poden ser utilitzades per estimar associacions. D'aquesta manera s'ampliaria el panorama de mètodes dominat pels mètodes freqüentistes. Conscients que existeix gran controvèrsia entre els dos punts de vista estadístics, seria oportú crear una que permetés realitzar anàlisis sota els dos enfocaments. Pel que fa a la qualitat dels resultats, seria interessant dissenyar un mètode que millori les estimacions de les freqüències haplotípiques baixes i dels efectes associats a elles, oferint la possibilitat de col·lapsar aquests valors rars en una sola categoria pels usuaris no interessats en aquesta qüestió. Un mètode que accepti valors missings i que permeti tractar diversos fenotips inclòs el de supervivència, l'ajust per covariables i interaccions amb factors ambientals, i que

alhora ofereixi la opció de considerar diferents models d'herència. Tot això implementat en una aplicació informàtica situada en un entorn de fàcil ús i que sigui factible d'utilitzar en diverses plataformes. Alhora, seria desitjable rebre com a resultat de l'execució del programa no només les estimacions de les quantitats d'interès exclusivament, si no també oferir a l'usuari la opció d'obtenir un mostreig per a aquestes variables, per així reunir més informació sobre el comportament d'aquestes estimacions i fins i tot poder-les graficar i resumir-ne les distribucions. A més, aquesta seria una bona opció per avaluar l'efectivitat del programa i la convergència a punts indesitjables com poden ser els màxims locals. En aquest sentit la majoria de programes són adreçats a usuaris de baixa experiència i no ofereixen els mostrejos dels paràmetres, ni l'opció de modificar els paràmetres bàsics amb que s'executarà el programa i que poden ser de vital importància per a que el mètode assoleixi una bona convergència. Seria una bona opció que usuaris avançats tinguessin la possibilitat de poder modificar aquests valors.

En els següents capítols anem a desenvolupar la metodologia necessària per fonamentar el disseny d'un algorisme Bayesià d'anàlisi haplotípic que compti amb aquestes característiques.

Part II

HIPÒTESIS DE TREBALL I OBJECTIUS

5

Hipòtesis de treball

Les hipòtesis d'aquesta Tesi Doctoral són les següents:

- El conjunt de mètodes i de programes d'anàlisi haplotípica que existeixen a l'actualitat presenten aspectes millorables. Així com existeix més varietat de mètodes i implementacions que estimin i reconstrueixin de manera satisfactòria la mostra d'haplotips, les eines existents per estimar els efectes associats als haplotips són insuficients en alguns escenaris particulars.
- L'estimació simultània de la mostra haplotípica i de l'associació entre els haplotips i el fenotip a estudi sembla ser millor alternativa per incorporar la incertesa a l'anàlisi que la imputació fixa.
- Els models de Regressió GLM són una eina adequada per estimar els efectes associats a una mostra d'haplotips en relació a diversos fenotips.
- Els mètodes Bayesians poden ser de gran utilitat en l'anàlisi haplotípica, permetent una millor avaluació dels resultats i una interpretació més intuïtiva.
- Els mètodes d'estimació basats en la simulació de Monte Carlo ofereixen estimacions més robustes que els basats en teoria asimptòtica especialment en alguns escenaris concrets.
- Les aplicacions que no són de fàcil accés, de fàcil ús, o que requereixen instal·lació resulten incòmodes i són poc utilitzades per part dels investigadors.

6

Objectius d'aquesta tesi

Els objectius que ens plantegem en aquesta Tesi Doctoral són els següents:

- Dissenyar un algorisme matemàtic d'estimació conjunta de freqüències haplotípiques i associació amb fenotips de tipus binari i quantitatiu basat en els models lineals generalitzats (GLM) que millori algunes limitacions dels algorismes existents.
- Desenvolupar i validar una aplicació informàtica basada en l'algorisme dissenyat, que sigui versàtil, de lliure accés i de fàcil maneig tant per usuaris comuns com per usuaris amb coneixements avançats.
- Comprovar que els mètodes Bayesians són una tècnica adient per dur a terme l'anàlisi haplotípica, tant per l'estimació de la mostra haplotípica com per a la realització de les anàlisis d'associació. Comprovar que les estimacions són més acurades que les dutes a terme mitjançant estimadors asimptòtics.
- Comprovar a nivell pràctic que és possible la implementació informàtica de l'algorisme basat en els aspectes teòrics estudiats i que l'aplicació és factible a nivell computacional.
- Posar de manifest les mancances dels programes actuals i els biaixos en els resultats que retornen tant a través de recerca bibliogràfica com duent a terme simulacions amb els propis programes. Comparar les eines pel que fa a les característiques més rellevants, tant teòriques com pràctiques.

Part III

MÈTODES

7

Mètodes Bayesians

La informació haplotípica ocupa un lloc prioritari en els estudis genètics i és per això que als darrers anys s'han fet molts esforços per desenvolupar mètodes estadístics d'anàlisi d'haplotips [133]. El mapa d'haplotips del genoma humà ha esdevingut un recurs molt valuós, no només per a la investigació genètica a nivell pràctic, sinó també pel desenvolupament de la metodologia haplotípica ([8],[7]). El fet de separar els cromosomes per tal d'obtenir haplotips és una tasca complexa que precisa de tècniques de laborartori cares. És per això que la majoria d'esforços s'han dedicat a resoldre la qüestió des de fora del laboratori, havent-se desenvolupat un conjunt de tècniques basades majoritàriament en la inferència estadística ([94],[95]) per resoldre la qüestió. Com hem fet constar a la introducció, existeix un conjunt ampli d'aplicacions que resolen l'estimació de freqüències haplotípiques, i algunes d'elles també tracten de resoldre l'anàlisi d'associació. Donat que les tècniques Bayesianes han donat molt bon rendiment en els estudis d'SNPs individuals tal i com conclouen Lunn i col·laboradors a [171], i també han estat molt útils en l'estimació de les freqüències haplotípiques [104], en aquesta tesi ampliarem el seu ús a l'estimació de l'associació entre diversos tipus de fenotips i haplotips.

Tot i que el punt de vista Bayesià ha comptat sempre amb el suport de molts estadístics, el seu desenvolupament s'ha mantingut sempre lligat a la possibilitat pràctica d'aplicar aquestes teories a problemes reals. L'àmplia i creixent aparició d'equipament computacional cada cop més eficient que ha tingut lloc durant les darreres dècades ha comportat

un increment sense precedents en la investigació sobre el tractament estadístic dels models complexos, fet que ha beneficiat fortament l'àrea de la inferència Bayesiana. Actualment podem dir que la història ha canviat definitivament. La redescoberta i aplicació de tècniques de simulació relativament senzilles, però alhora molt potents, ha permès considerar el paradigma Bayesià pel tractament de diversos problemes pràctics complexos, com el que ens ocupa en aquest treball. A més, l'ús d'aquestes tècniques no requereix la necessitat de comptar amb requisits de coneixement estadístic específics previs.

El mètode d'anàlisi d'associació haplotípica que desenvoluparem en aquesta tesi es fonamenta en la inferència Bayesiana. És per això que passem a introduir els conceptes principals que conformen aquest camp de l'estadística.

7.1 En què es basa l'enfocament Bayesià?

Quan afirmem que en llençar una moneda a l'aire la probabilitat que surti cara és de 0.5 hi ha dues possibles interpretacions. D'un cantó, pot voler dir que si llencem la moneda molts cops esperem obtenir el mateix nombre de cares que de creus. Aquesta és la interpretació freqüentista de la probabilitat. D'altra banda, la interpretació Bayesiana diu que la probabilitat de 0.5 és quelcom subjectiu, és a dir, és allò que un individu concret espera en llençar una moneda a l'aire, però pot no ser el mateix nombre per un altre individu diferent. Per tant, la principal diferència conceptual entre l'estadística freqüentista i l'estadística Bayesiana és la interpretació del que significa una probabilitat.

Històricament, la visió Bayesiana fou predominant al llarg del s.XIX amb els treballs de l'estadístic i astrònom francès Pierre-Simon Laplace. Tot i així, l'enfocament freqüentista ha dominat la ciència estadística del s.XX essent-ne pioner l'estadístic i genetista anglès Ronald A. Fisher. Donat l'auge que estan tenint els mètodes Bayesians als darrers anys, és complicat predir quina serà la perspectiva dominant al s.XXI.

Punts en comú i diferències

Abans d'exposar les diferències entre freqüentistes i Bayesians, és important deixar clars els aspectes comuns. Per un costat, en ambdós casos s'utilitzen models amb paràmetres desconeguts per caracteritzar el món real. D'altra banda, els dos enfocaments requereixen la recol·lecció de dades com a base de l'estimació d'aquests paràmetres desconeguts.

A la pràctica, la principal diferència entre l'estadística freqüentista i Bayesiana és el tractament dels paràmetres desconeguts que volem estimar per caracteritzar el món real a través de models. Els freqüentistes consideren els paràmetres com uns valors fixos però desconeguts. L'estimació es basa en l'elecció d'aquells valors dels paràmetres que maximitzen la probabilitat d'observar les dades. De la seva banda, els Bayesians interpreten els paràmetres com a variables aleatòries tals que la seva distribució de probabilitat ve donada pel Teorema de Bayes. La idea és simple: un Bayesià ha de tenir una distribució dels paràmetres abans de veure les dades (a priori) que modificarà segons les dades que hagi observat per obtenir una distribució a posteriori que resumirà tot el coneixement de l'investigador sobre els paràmetres d'interès, donades les dades i les seves creences a priori.

De distribucions a priori n'existeixen de dues classes: les informatives i les anomenades objectives o no informatives. Una prior informativa és aquella que expressa informació específica i definida sobre la variable. Aquest tipus de priors són també anomenades subjectives, donat que la seva tria sovint té a veure amb la percepció subjectiva que l'investigador té en relació al paràmetre a estudi. En canvi, una prior no informativa expressa informació vaga o general. En aquest cas es tracta d'informació objectiva, no opinable, com per exemple el fet que la variable sigui positiva o inferior a algun valor límit. El mètode més simple per determinar una prior no informativa és el principi de la indiferència, que assigna la mateixa probabilitat a tots els possibles valors. En l'estimació de paràmetres, l'ús d'una prior no informativa provoca que sigui la funció de versemblança la que aporti tota la informació.

El següent teorema ens dóna la clau sobre com combinar les creences a priori amb les dades observades.

7.1.1 Teorema de Bayes

Teorema 7.1.1 *(Teorema de Bayes)* Siguin D el conjunt de valors observats per una variable aleatòria X i θ el model de paràmetres, tals que $P(D) > 0$ i $P(\theta) > 0$. Aleshores es compleix que

$$P(\theta|D) = \frac{P(\theta) \cdot P(D|\theta)}{\int P(\theta) \cdot P(D|\theta) d\theta} \tag{7.1}$$

Aquesta expressió és equivalent a una de més generalitzada, que usarem sovint:

$$\pi(\theta) = P(\theta|D) = \frac{P(\theta) \cdot P(D|\theta)}{P(D)} \tag{7.2}$$

on $\pi(\theta) = P(\theta|D)$ és la probabilitat a posteriori, $P(\theta)$ és la priori i $P(D|\theta)$ coincideix amb la versemblança de la mostra. Assumirem que $\int P(\theta)P(D|\theta)$, la constant de normalització, pot ser desconeguda, i que per tant $P(\theta|D) \propto P(\theta)P(D|\theta)$.

La idea fonamental del Teorema de Bayes es la modificació de les creences un cop s'han observat les dades. Es tracta de l'ordre de les causes i els efectes. Donat un problema relacionat amb una situació d'incertesa, la informació a priori de la que disposem s'incorpora al càlcul de la probabilitat actual. I així, coneixent la probabilitat dels efectes, capgirem l'ordre natural causa-efecte per poder calcular la probabilitat de les causes.

El fet d'obtenir la distribució *a posteriori* és un pas important, però no el definitiu. Donada aquesta distribució, és possible extreure'n informació molt valuosa i traduir-la en termes del seu impacte en l'estudi. Això es troba directament relacionat amb l'avaluació de mesures de resum com són la mitjana, la mediana o la moda, la desviació estàndard i els intervals de probabilitat o credibilitat. Aquests intervals Bayesians i els de confiança freqüentistes han de coincidir en cas que no s'estigui utilitzant informació prior. Tot i així, és important tenir present que la interpretació d'ambdós intervals és diferent; un interval

de credibilitat del $r\%$ per un paràmetres ens indica que hi ha una probabilitat igual al $r\%$ que el valor poblacional pel paràmetre es trobi en aquest interval. En canvi, l'interval de confiança del $r\%$ no ens dóna la probabilitat que el valor poblacional del paràmetre estigui a dins de l'interval. El que ens diu és la proporció d'intervals que, amb la mateixa mida de mostra, contenen el valor real de la població. Es a dir, si prenem 100 mostres de la mateixa mida i calculem per cadascuna d'elles l'interval del $r\%$ de confiança, hauria de passar que a r intervals estigui inclòs el valor real (poblacional) del paràmetre. Però, en concret, no sabem si la proporció està o no inclosa al nostre interval. Hi ha un $(100 - r)\%$ de probabilitat que no hi sigui.

Càlcul d'esperances en espais multidimensionals

Essent (7.1) l'expressió de la distribució posterior pel paràmetre a estudi, el càlcul de mesures de resum com són els moments o els quantils comporten el tractament de l'esperança de la distribució per certa funció $t(\theta)$. Per tant sigui quina sigui la mesura de resum d'interès, l'objectiu serà calcular la següent expressió:

$$E[t(\theta)|D] = \frac{\int t(\theta)P(\theta)P(D|\theta)d\theta}{\int P(\theta)P(D|\theta)d\theta} \tag{7.3}$$

Les integrals que apareixen a (7.3) han estat durant anys la causa de la majoria de les dificultats pràctiques d'aplicar inferència Bayesiana. Especialment en dimensions grans, l'avaluació analítica de $E[t(\theta)|D]$ és literalment impossible. Les alternatives per calcular-ho inclouen avaluació numèrica, que resulta dificultosa i imprecisa per dimensions superiors a 20. Com veurem al capítol 8.1, la integració numèrica per Monte Carlo, incloent els mètodes MCMC, resulta més precisa per alguns escenaris.

8

MCMC: Integració per Monte Carlo i Cadenes de Markov

Avui en dia existeix una gran quantitat de problemes classificats en la categoria de models d'alta dimensionalitat. Els mètodes de Markov Chain Monte Carlo es refereixen a una àrea de l'estadística, habitualment anomenada MCMC, nom que sorgeix de considerar la inicial de cadascuna de les paraules. Es tracta d'una família de tècniques que donen resposta al tan dificultós problema de simular sobre valors desconeguts de distribucions multivariades que apareixen en considerar models complexos en espais de dimensió elevada.

La introducció de les cadenes de Markov en els esquemes de simulació és vital per poder tractar amb distribucions complicades d'aquest estil. En termes molt generals, les cadenes de Markov són processos que descriuen trajectòries tals que quantitats successives es defineixen probabilísticament d'acord amb el valor dels seus predecessors immediats. En alguns casos, aquests processos tendeixen a un equilibri i les quantitats límit segueixen una distribució invariant. Les tècniques MCMC permeten simular d'una distribució, considerant-la com a distribució límit d'una cadena de Markov, i simulant valors de la cadena fins que assoleixin l'equilibri.

D'aquesta idea se'n deriva ràpidament una qüestió: com es pot construir una cadena de Markov tal que la seva distribució límit sigui exactament la distribució d'interès? És fascinant descobrir com aquest fet no només és possible, si no que a més existeixen una diversitat d'esquemes que permeten generar cadenes amb aquesta propietat.

Abans d'entrar de ple als mètodes MCMC és important que tant la integració per Monte

Carlo com les propietats de les cadenes de Markov s'entenguin bé. Per això, en les següents seccions se n'exposaran els resultats més rellevants. Tots els resultats es mostraran per variables contínues, essent igualment vàlids per variables discretes. Els termes "funció de densitat"o "funció de distribució"seran tractats indistintament. Per adaptar els resultats pel cas de variables discretes, només caldrà canviar integrals per sumatoris. Les qüestions de caire més matemàtic, definicions, resultats i demostracions, es poden trobar als annexos d'aquest treball.

8.1 Integració per Monte Carlo

Sigui θ el paràmetre d'una distribució $\pi(\theta)$, entès com una variable aleatòria en el context de la inferència Bayesiana. En voler resumir la informació d'una mostra de dades per aquesta variable, serà necessari resoldre una integral de la forma:

$$I = \int t(\theta)\pi(\theta)d\theta \tag{8.1}$$

En cas que $t(\theta) = \theta$, llavors (8.1) correspon a l'esperança per θ. Si calculem $I(\theta < c) = \frac{1}{2}$ aleshores c és la mediana. Per $t(\theta) = \theta^2$, (8.1) correspon a la variància.

Depenent de la complexitat que presenti la funció de distribució, la resolució analítica d'aquesta expressió pot no ser viable. En aquest cas, podem utilitzar integració de Monte Carlo.

Proposició 8.1.1 Sigui $q(\theta)$ una densitat alternativa per θ amb el mateix suport que $\pi(\theta)$. Aleshores,

$$I = \int \frac{t(\theta)\pi(\theta)}{q(\theta)} q(\theta)d\theta = E_q[\frac{t(\theta)\pi(\theta)}{q(\theta)}]$$

Teorema 8.1.2 Sigui $\theta_1, \ldots, \theta_n$ una mostra per θ que segueix la distribució $q(\theta)$. Aleshores,

$$\hat{I} = \frac{1}{n}\sum_{i=1}^{n} \frac{t(\theta_i)\pi(\theta_i)}{q(\theta_i)} \tag{8.2}$$

Es tracta de l'aplicació del mètode dels moments per estimar I que estima l'esperança poblacional segons la mitjana mostral. Aquest és un estimador amb bones propietats:

- No té biaix
- La seva variància és de la forma $V_q(\hat{I}) = \frac{\sigma^2}{n}$ on σ^2 depèn de π, t i q
- Pel teorema central del límit, $\sqrt{n}\frac{\hat{I}-I}{\sigma} \to N(0,1)$ quan n$\to \infty$
- $\hat{I} \to I$ quan n$\to \infty$ amb probabilitat 1 (l'estimador és consistent).

Quan els valors de la mostra $\{\theta_i\}$ són independents, les lleis dels grans nombres asseguren que l'aproximació de l'esperança descrita a (8.2) es pot fer tan precisa com es desitgi, només incrementant la mida de la mostra n. Notar que en tant que la mostra $\{\theta_i\}$ és generada per l'analista, n està sota el seu control, no és una mida mostral fixada.

En general, mostrejar valors $\{\theta_i\}$ independentment de $q(\theta)$ no és factible donat que no acostuma a tractar-se de distribucions estàndard. Malgrat tot, no és estrictament necessari que els $\{\theta_i\}$ siguin independents. De fet, n'hi ha prou amb que els $\{\theta_i\}$ siguin generats mitjançant qualsevol procés que, amb paraules planeres, mostregi valors al llarg del suport de $q(\theta)$ amb les proporcions correctes. Una manera de fer això és a través d'una cadena de Markov que tingui $q(\theta)$ com a distribució estacionària. Això és exactament al que ens referim quan parlem de "Markov Chain Monte Carlo".

8.2 Cadenes de Markov

Considerem una seqüència de variables alaeatòries $\{\theta_0, \theta_1, \dots\}$ tal que a cada temps $t \geq 0$ el següent estat θ_{t+1} s'obté mostrejant d'una distribució $P(\theta_{t+1}|\theta_t)$ que depèn només de l'estat actual de la cadena, θ_t. Això és, donat θ_t, el següent estat de la cadena θ_{t+1} no depèn dels estats més antics $\theta_0, \theta_1, ...\theta_{t-1}$. En altres paraules, passat i futur són independents. Aquesta seqüència s'anomena Cadena de Markov, i $P(.|.)$ és l'anomenat *transition kernel* o nucli de transició de la cadena. Assumirem que la cadena és homogènia en relació al temps, és a dir

que $P(.|.)$ no depèn de t.

Com afecta l'estat inicial θ_0 a θ_t? Aquesta qüestió implica la distribució de θ_t donat θ_0, que podem denotar per $P^t(\theta_t|\theta_0)$. Aquí no estem considerant les variables intermitges $\theta_1, ...\theta_{t-1}$ pel que θ_t depèn directament de θ_0. Sent fidel a les condicions de regularitat, la cadena gradualment "oblidarà"el seu estat inicial i $P^t(.|\theta_0)$ eventualment convergirà a una distribució única anomenada invariant o estacionària, que no dependrà de t o de θ_0.Denotarem la distribució estacionària com $\pi(.)$. Per tant, a mida que t creixi, els valors mostrejats $\{\theta_t\}$ cada cop s'aproparan més a ser mostrejos dependents de la distribució $\pi(.)$.

Així doncs, amb un *burnin* (o període "d'escalfament"per la cadena) suficientment llarg de m iteracions, les següents $\theta_{m+1}, ..., \theta_n$ seran aproximadament valors dependents mostrejats de $\pi(.)$. Existeixen diferents mètodes per determinar el valor m ([172],[173]).

Ara podem utilitzar els valors sortida de la cadena de Markov per estimar l'esperança $E[t(\theta)]$ on θ es distribueix segons $\pi(.)$. Les mostres de l'espai burnin solen ser descartades per aquest càlcul, donant lloc al següent estimador:

$$\bar{\theta} = \frac{1}{n-m} \sum_{t=m+1}^{n} \theta(\theta_t) \tag{8.3}$$

Aquesta és l'anomenada mitjana ergòdica. La convergència a l'esperança en qüestió és assegurada pel teorema ergòdic. Aquest teorema i una ampliació de les qüestions més tècniques sobre la teoria de cadenes de Markov es troben a l'Apèndix C.

8.3 Métodes de Markov Chain Monte Carlo

Com acabem de veure, la cadena de Markov generada amb distribució límit coincidint amb la d'interès (cadena que per tant representa un mostreig per la distribució) se sumaritza mitjançant el càlcul de mitjanes ergòdiques. Una mitjana ergòdica sobre una mostra és, com hem vist, una aplicació de la integració de Monte Carlo.

Encara segueix en peu la pregunta sobre com generar una cadena de Markov tal que la seva distribució límit sigui exactament una distribució concreta. Com ja s'ha comentat amb ante-

rioritat, hi ha diversos mètodes que permeten generar cadenes així. Un d'aquests mètodes és la Gibbs Sampling, popularitzada per Gelfand i Smith al 1990 [21]. Es basa en una cadena de Markov tal que la dependència del predecessor ve definida per la distribució condicional que prové del mateix model amb que s'està treballant. Pot passar que el model tingui una distribució conjunta complexa però que per construcció la distribució condicional sigui més senzilla. Gibbs sampling explora aquest punt i és capaç de proporcionar solucions simples a problemes complexos. Una altra possibilitat com es veurà la proporcionen els algorismes de Metropolis Hastings, basats en una cadena de Markov tal que la dependència dels estats predecessors es divideix en dues parts:una *proposal* i una acceptació de la *proposal*. Les proposals suggereixen un següent pas arbitrari en la trajectòria de la cadena i l'acceptació assegura si la direcció cap a la distribució límit és apropiada. Alguns dels algorismes de Metropolis-Hastings poden ser vistos com generalitzacions de la Gibbs Sampling. En la present tesi aplicarem tècniques de Metropolis-Hastings i Gibbs Sampling.

8.3.1 Idea intuitiva

Donat un conjunt de paràmetres $(\theta_1, \ldots, \theta_n)$, mitjançant una cadena de Markov es generarà una mostra de valors per a cada component. D'aquesta manera obtindrem la distribució conjunta del vector de paràmetres, donat que cada distribució marginal per cadascun dels paràmetres θ_i està generant globalment una superfície en un espai de dimensió n. Per exemple, en un espai de 2 paràmetres (P, Q) on tinguéssim una funció de densitat com la de la figura 8.1, intuïtivament cada cop que generem un nou valor per la cadena podem pensar que estem "pintant" un punt del suport d'aquesta superfície. la Figura 8.2 ens mostra com la cadena aniria mostrejant de la distribució conjunta a posteriori del vector (P, Q). Quant millor es recobreix el suport, amb les passes de la cadena, millor serà el mostreig, i millor es dibuixarà la distribució.

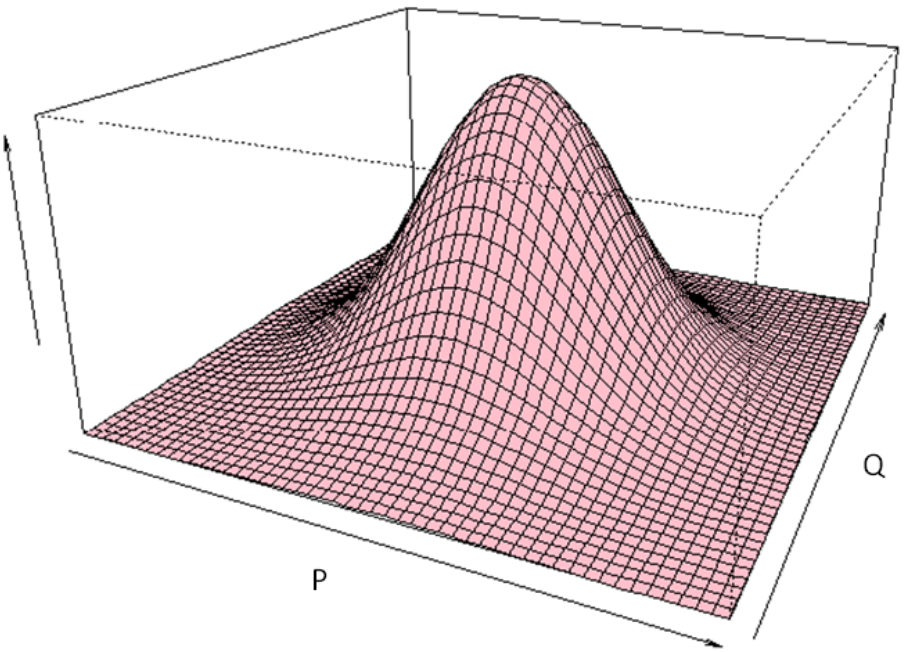


Figura 8.1. Funció de densitat multidimensional

8.3.2 Algorisme de Metropolis-Hastings

L'objectiu que ens ocupa és saber com es poden generar cadenes de Markov de manera que tinguin com a distribució estacionària la que desitgem. Doncs bé, construir una cadena de Markov així és sorprenentment fàcil. Segons l'algorisme de Metropolis-Hastings, fixat un pas n, triem el següent estat de la cadena $\theta^{(n+1)}$ mostrejant un punt candidat Y segons una distribució proposada $q(\cdot \mid \theta^{(n)})$ que depèn del punt actual. Sigui el pas actual $\theta^{(n)} = X$. Llavors, el punt candidat serà acceptat amb probabilitat $\alpha(X, Y)$ on

$$\alpha(X, Y) = \min\left(1, \tfrac{\pi(Y)q(X|Y)}{\pi(X)q(Y|X)}\right). \tag{8.4}$$

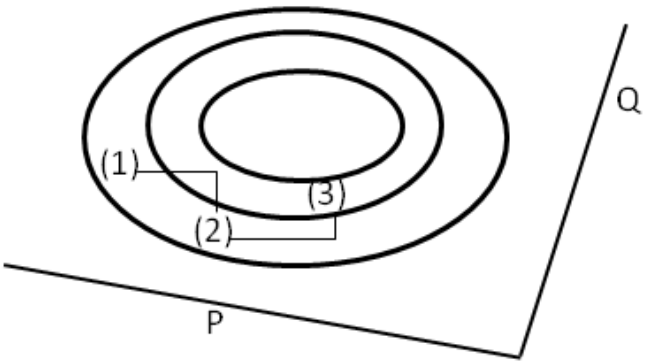


Figura 8.2. Passeig d'una cadena via Gibbs Sampling

Si el candidat és acceptat, l'estat següent serà $\theta^{(n+1)} = Y$. Si el candidat no s'accepta, la cadena no es mou i llavors $\theta^{(n+1)} = \theta^{(n)} = X$. La distribució estacionària per la cadena serà π.

L'Algorisme pas a pas

Esquematitzem el que acabem d'explicar en unes quantes passes:

Inicialitzem la cadena: $\theta^{(0)} = (\theta_1^{(0)}, \theta_2^{(0)}, \ldots, \theta_n^{(0)})$

I iterem:

1. Mostregem un punt Y de $q(\cdot \mid \theta^{(n)})$
2. Mostregem un valor U de $unif(0,1)$
3. Si $U \leq \alpha(\theta^{(n)}, Y)$ llavors $\theta^{(n+1)} = Y$. Si no, $\theta^{(n+1)} = \theta^{(n)}$.
4. $n = n + 1$

Fixem-nos que:

- La distribució $q(\cdot \mid \cdot)$ pot tenir qualsevol forma. La seva tria no afectarà la convergència de la cadena pròpiament, però sí a la velocitat amb que ho faci .
- Aquesta cadena és de Markov. A cada pas la proposada només depèn del pas actual.

Per què funciona?

Tal i com es pot veure a la secció de l'apèndix C dedicada a cadenes de Makov, tot i que la reversibilitat no és una condició necessària per a que la distribució de la cadena convergeixi a una distribució estacionària, sí que és suficient. Per tant, si considerem cadenes reversibles amb un nucli de transició p que satisfaci

$$\pi(\theta)p(\theta,\phi) = \pi(\phi)p(\phi,\theta), \forall(\theta,\phi) \in S \tag{8.5}$$

π serà la distribució estacionària de la cadena.

La cadena generada mitjançant Metropolis-Hastings sorgeix de considerar com a nucli de transició $p(\theta,\phi)$ una expressió depenent de 2 factors: un nucli de transició arbitrari $q(\theta,\phi)$ i una probabilitat $\alpha(\theta,\phi)$ tal que

$$p(\theta,\phi) = q(\theta,\phi)\alpha(\theta,\phi)$$

si $\theta \neq \phi$.

Per tant, el nucli de transició defineix una densitat $p(\theta,\cdot)$ per cada possible valor del paràmetre, diferent de θ. Llavors, la cadena té una probabilitat $1 - \int q(\theta,\phi)\alpha(\theta,\phi)d\phi$ de quedar-se a l'estat θ.

Resultat: Triada $q(\cdot \mid \cdot)$ i prenent el valor d'α descrit a (8.4), tenim que p defineix una cadena reversible amb distribució estacionària π.

Tot i que qualsevol distribució proposada q ens arribarà a donar un mostreig de π, la velocitat de convergència sí que depèn de la tria de q. És recomanable fer anàlisis exploratòries per cada cas, tot i que sovint la tria més senzilla de $q(\cdot \mid \cdot)$ acostuma a donar bons resultats.

8.3.3 Algorisme de Metropolis

L'Algorisme de Metropolis és un cas particular del de Metropolis-Hastings, en el que la distribució proposada q és simètrica, *i.e.*, $q(\theta^{(n+1)} \mid \theta^{(n)}) = q(\theta^{(n)} \mid \theta^{(n+1)})$. Per exemple si θ és contínua, $q(\cdot \mid \theta)$ podria ser una normal amb una mitjana i variància concretes. En aquest cas, la probabilitat d'acceptació no depèn de q. Si recordem (8.4) ara, la q simètrica es cancel·la, i llavors

$$\alpha(\theta^{(n)}, \theta^{(n+1)}) = \min\left(1, \frac{\pi(\theta^{(n+1)})}{\pi(\theta^{(n)})}\right). \tag{8.6}$$

Un cas especial d'algorisme de Metropolis és el *Random Walk Metropolis*. Es tracta d'un esquema molt simple basat en una distribució proposada simètrica (com per exemple la normal) centrada en l'estat actual. Totes les variables poden adaptar-se simultàniament o adaptar alternativament una variable a cada moment de temps. Per aquest mètode $q(\theta^{(n+1)} \mid \theta^{(n)}) = q(|\theta^{(n)} - \theta^{(n+1)}|)$. El nou punt generat per la cadena és el resultat de sumar al punt anterior un nou valor generat per q, *i.e*, $\theta^{(n+1)} = \theta^{(n)} + q(|\theta^{(n)} - \theta^{(n+1)}|)$. Per tant q està generant les distàncies entre els punts de la cadena.

Notem que una distribució proposada q que generi passes molt petites, tindrà una acceptació molt alta (ja que $\frac{\pi(\theta^{(n+1)})}{\pi(\theta^{(n)})}$ és propera a 1). Una distribució més arriscada, que generi grans salts entre el centre i les cues de la distribució, farà que la fracció sigui petita i per tant tinguem baixa acceptació. Això ens obligaria a haver de generar moltes més passes de la cadena per aconseguir convergència. Per tant, una q òptima serà aquella que eviti aquests dos extrems.

Metropolis-Hastings d'una component

Sigui $\theta^{(n)} = (\theta_1^{(n)}, ...\theta_m^{(n)})$ el pas actual de la cadena. En aquest cas particular dels algorismes de Metropolis -hastings, l'actualització a cada pas es fa component a component. Les passes de l'algorisme són les següents:

Sigui $\theta_{-i}^{(\cdot)} = (\theta_1^{(\cdot)}, \ldots, \theta_{i-1}^{(\cdot)}, \theta_{i+1}^{(\cdot)}, \ldots, \theta_m^{(\cdot)})$. Per tenir una actualització del vector sencer, caldrà

fer m actualitzacions, una per a cada component. Sigui $\theta_i^{(n)}$ l'estat de la coordenada i-èssima al pas n-èssim de la cadena. Aquesta coordenada, per l'estat $(n+1)$-èssim de la cadena, es genera usant l'algorisme de Metropolis-Hastings, mitjançant la distribució proposada $q_i(\theta_i^{(\cdot)} \mid \theta_i^{(n)}, \theta_{-i}^{(n)})$ on

$$\theta_{-i}^{(n)} = (\theta_1^{(n+1)}, \theta_2^{(n+1)}, \ldots, \theta_{i-1}^{(n+1)}, \theta_{i+1}^{(n)}, \ldots, \theta_m^{(n)})$$

i les components $1, 2, \ldots, i-1$ ja han estat actualitzades. Així doncs, la i-èssima proposada q_i genera un candidat només per la coordenada i-èssima de θ i pot dependre de qualsevol dels valors que prenguin la resta de components. El candidat Y_i és acceptat amb probabilitat $\alpha(\theta_{-i}^{(n)}, \theta_i^{(n)}, Y_i^{(\cdot)})$ on

$$\alpha(\theta_{-i}^{(n)}, \theta_i^{(n)}, Y_i^{(\cdot)}) = \min\left(1, \frac{\pi(Y_i^{(\cdot)}|\theta_{-i}^{(\cdot)} q_i(\theta_i^{(\cdot)}|Y_i^{(\cdot)}, \theta_{-i}^{(\cdot)}}{\pi(\theta_i^{(\cdot)}|\theta_{-i}^{(\cdot)} q_i(Y_i^{(\cdot)}|\theta_i^{(\cdot)}, \theta_{-i}^{(\cdot)}}\right). \tag{8.7}$$

Aquí, $\pi(\theta_i^{(\cdot)} \mid \theta_{-i}^{(\cdot)})$ és la distribució *full conditional* per $\theta_i^{(\cdot)}$ sota $\pi(\cdot)$. Si Y_i s'accepta, llavors $\theta_i^{(n+1)} = Y_i$. Si no s'accepta, $\theta_i^{(n+1)} = \theta_i^{(n)}$. La resta de components no es toquen. Cada adaptació provoca un pas en la direcció d'un dels eixos de coordenades, com es pot veure a la Figura 8.2.

La distribució full conditional $\pi(\theta_i^{(\cdot)}) \mid \theta_{-i}^{(\cdot)}$ és la distribució de la i-èssima component de θ condicional a la resta de components, on θ té distribució $\pi(\cdot)$:

$$\pi(\theta_i^{(\cdot)} \mid \theta_{-i}^{(\cdot)}) = \frac{\pi(\theta)}{\int \pi(\theta) d\theta_i^{\cdot}} \tag{8.8}$$

Aquest algorisme amb probabilitat d'acceptació (8.7) genera correctament mostres de la distribució objectiu $\pi(\theta)$ perquè aquesta distribució està unívocament determinada pel seu conjunt de full conditionals.

8.3.4 Gibbs Sampling

La Gibbs Sampling és un cas particular dels algorismes de Metropolis-Hastings d'una component. Aquest mètode considera com a distribució proposada per adaptar la component i-èssima de θ a la següent q_i:

$$q_i(Y_i^{(\cdot)} \mid \theta_i^{(\cdot)}, \theta_{-i}^{(\cdot)}) = \pi(Y_i^{(\cdot)} \mid \theta_{-i}^{(\cdot)}) \tag{8.9}$$

Els candidats a ser nou punt de la cadena, generats via la Gibbs Sampler, sempre són acceptats. Només cal substituir (8.9) a (8.7) i ja veiem que s'obté $\alpha = 1$. Així doncs, la Gibbs Sampling consisteix únicament en mostrejar de les full conditionals i anar actualitzant cada component. Com s'escriuen les distribucions condicionals? Anem a descriure-ho, fent un canvi de notació. Reanomenem β als paràmetres ja que en facilitarà la comprensió de l'aplicació que tindrem en compte més endavant pels diferents models de regressió.

Sigui $\beta = (\beta_0, \ldots, \beta_p)$ un vector de $p+1$ paràmetres. Com hem vist la teoria general de la Gibbs Sampling descriu com obtenir-ne un mostreig. Sigui

$$\pi(\beta_i|\beta_0^{(n)}, \ldots, \beta_{i-1}^{(n)}, \beta_{i+1}^{(n-1)}, \ldots, \beta_p^{(n-1)}) = \frac{\pi(\beta_i, \beta_{-i})}{\int \pi(\beta_i, \beta_{-i}) d\beta_i} \tag{8.10}$$

la funció *full conditional* per β_i. La Gibbs Sampler diu que :

$$\beta_i^{(n)} \sim \pi(\beta_i|\beta_0^{(n)}, \ldots, \beta_{i-1}^{(n)}, \beta_{i+1}^{(n-1)}, \ldots, \beta_p^{(n-1)})$$

Per tant, a cada volta de l'algorisme caldrà fer $p+1$ mostrejos dels que obtindrem un nou valor pel vector de β's. En fer consecutives voltes anirem obtenint una cadena de vectors, que a partir d'un lloc dibuixaran un mostreig pel vector.

Com s'escriu la *full conditional*?

Moltes vegades, en desconèixer la distribució del vector β, no som capaços d'escriure directament la distribució *full conditional*. Per aquest motiu, si es coneix la versemblança pels paràmetres i les distribucions a priori, es treballa amb un model Bayesià. Sigui x la variable observada. La distribució conjunta per x i β és

$$P(x, \beta) = \prod_1^N P(x_i|\beta) \mathit{prior}(\beta) \tag{8.11}$$

Quan x és observada la distribució conjunta posterior per β és

$$\pi(\beta_0, \ldots, \beta_p) = P(\beta_0, \ldots, \beta_p|x) = \frac{P(x, \beta)}{\int P(x, \beta) d\beta} \tag{8.12}$$

Unint (8.10),(8.11) i (8.12) podrem demostrar el següent resultat:

$$\pi(\beta_i|\beta_{-i}) \propto P(x,\beta) \tag{8.13}$$

Veiem-ho: Per la definició (8.10) de *full conditional*

$$\pi(\beta_i|\beta_{-i}) = \frac{\frac{P(x,\beta)}{\int P(x,\beta)d\beta}}{\frac{\int P(x,\beta)d\beta_i}{\int P(x,\beta)d\beta}}$$

Simplificant,

$$\pi(\beta_i|\beta_{-i}) = \frac{P(x,\beta)}{\int P(x,\beta)d\beta_i} = \frac{P(x,\beta)}{P(x,\beta_{-i}|\beta i)}$$

Donat que $\pi(\beta_i|\beta_{-i})$ és una funció en β_i, el denominador és una constant (no depèn de β_i). Així, se segueix (10.1). Substituint el valor de $P(x,\beta)$ pel donat pel model Bayesià (8.11) ja tenim l'expressió de qui mostrejar a cada pas de la Gibbs Sampling:

$$\pi(\beta_i|\beta_{-i}) \propto \prod_1^N P(x_i|\beta)prior(\beta) \tag{8.14}$$

Cal tenir en compte que això serà així en cas de no tenir hiperparàmetres. Si les priors pel vector β depenguessin de paràmetres, caldria afegir les probabilitats condicionades corresponents (Richardson,Spiegelhalter, pàg.77)

Fixem-nos també que per construir la *full conditional* per β_i només cal prendre els termes de 8.14 que depenen de β_i.

Exemple: Aplicació en Regressió Logística. Qui és la *full conditional*?

Direm qui és $\pi(\beta_i|\beta_{-i})$ llevat d'una constant, tal i com s'ha descrit a (8.14). Ens cal conèixer la funció de versemblança de la mostra i la distribució a priori pels paràmetres. Aquests paràmetres ara són els coeficients d'una regressió Logística, en la que participa una variable resposta dicotòmica y que pren valors 1 i 0, i una covariable x. El model logístic es troba explicat a la secció 9.2.1. En ser y una variable discreta, el que modelarem serà la proporció d'y, *i.e.*, $E(y=1|x)$. Aquesta esperança correspon a una funció de probabilitat $f(x) = \frac{\exp(x\beta)}{1+\exp(x\beta)}$. D'aquesta manera, si anomenem $p = E(y=1|x) = f(x)$, en fer $logit(p)$ obtenim un model Lineal:

$$logit(p) = \log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

Com ja hem dit, $y = 1$ amb probabilitat p. Per tant, $y = 0$ amb prob $1 - p$. Així doncs, y es distribueix com una *Bernoulli*(p). La funció de versemblança per una mostra de N individus és:

$$P(x_i|\beta) = \prod_{i=1}^{N} p^{y_i}(1-p)^{1-y_i} \tag{8.15}$$

Pel que fa a la prior pels paràmetres és habitual considerar distribucions normals poc informatives, planes (variància gran, precisió petita). Explicitem les condicionals pel cas d'una covariable z:

$$\begin{aligned} y_i &\sim Bernoulli\left(\frac{1}{1+\exp -(\mu+\alpha z_i)}\right) \\ \alpha &\sim N(0,1) \\ \mu &\sim N(0,1) \end{aligned}$$

La *full conditional* per α és :

$$\pi(\alpha|\mu) \propto \exp(-\frac{1}{2}\alpha^2)\prod_{i=1}^{N}\left\{\frac{1}{1+\exp -(\mu+\alpha z_i)}\right\}^{y_i}\left\{\frac{1}{1+\exp(\mu+\alpha z_i)}\right\}^{1-y_i} \tag{8.16}$$

Així mateix, per μ escriuríem:

$$\pi(\mu|\alpha) \propto \exp(-\frac{1}{2}\mu^2)\prod_{i=1}^{N}\left\{\frac{1}{1+\exp -(\mu+\alpha z_i)}\right\}^{y_i}\left\{\frac{1}{1+\exp(\mu+\alpha z_i)}\right\}^{1-y_i} \tag{8.17}$$

8.3.5 Mètodes per mostrejar de funcions de densitat no estàndards: DFARS i Slice Sampling.

La Gibbs sampling és vàlida només si se sap com mostrejar de les diferents funcions de distribució condicionals que s'hi veuen implicades. Aquest fet sovint pot comportar la necessitat de mètodes específics per mostrejar valors d'aquestes funcions. És amb aquesta finalitat que es desenvolupa la *Adaptive Rejection Sampling* (ARS) ([174],[175]), algorisme que permet mostrejar eficientment de qualsevol funció de distribució condicional tal que la seva funció de densitat sigui log-còncava. La diferenciabilitat de la funció pot ajudar però no és imprescindible. El primer pas en aplicar ARS és com veurem el de trobar punts als dos

costats de la moda de la distribució. Això en general implicarà una cerca i la tria d'una interval inicial. Aquest valor pot ser triat retrospectivament després de testar algunes iteracions de la cadena, sense afectar el resultat final, donat que aquesta qüestió només pot modificar la rapidesa de la convergència de la cadena, però no la convergència en sí. A partir d'aquest mètode, es proposen d'altres variacions com l'ARMS (Adaptive Rejection Metropolis Hastings) que allibera la necessitat de densitats log-còncaves i el DFARS (Derivative Free Adaptive Rejection Sampling) que com el seu nom indica, és una adaptació de l'ARS tal que no necessita el supòsit de diferenciabilitat sobre la funció d'on es mostreja.

Tot i que aquests mètodes resulten útils en un ampli rang de situacions, hi ha certs casos com el que ens ocupa en aquest treball, en que s'han mostrat massa costosos a nivell computacional. En aquest sentit, l'*Slice Sampling* [21] és una tècnica alternativa de mostreig que també permet mostrejar de distribucions complexes, resultant molt més eficient. Es tracta d'un mètode que adapta apropiadament l'interval de mostreig de manera recurrent durant les iteracions, en funció de la zona que s'estigui mostrejant. Els algorismes d'Slice Sampling que adapten de manera elaborada aquestes passes, o bé que suprimeixen els random-walks, poden potencialment ser molt més ràpids que mètodes més simples.

Tots aquests mètodes poden servir per mostrejar distribucions multivariants i no requereixen l'avaluació de la constant normalitzadora. Aquest és un punt important, perquè al cas dels haplotips, no tindrem aquesta constant.

DFARS: Derivative Free Adaptive Rejection Sampling

Es tracta un mètode englobat en els anomenats de *Rejection Sampling*. En general, si $g(Y)$ és una funció proporcional a la distribució d'interès $\pi(\beta_i|\beta_{-i})$ la rejection sampling necessita una funció *envelope* $G(Y) > g(Y)\ \forall Y$ de la que mostreja el candidat Y. Aquest valor és acceptat com a punt pertanyent al mostreig de $g(Y)$ amb probabilitat $g(Y)/G(Y)$. Notem que al nostre cas, la funció proporcional a $\pi(\beta_i|\beta_{-i})$ és la donada a (8.14). Per tant, el nostre principal problema és crear-li una *envelope*, un recobriment, $G(Y)$.

L'algorisme utilitzat en l'ARS es basa en construir la funció envelope prenent les tangents per un conjunt d'abscises (tres és suficient). Aquest mètode ens obliga a derivar la funció $g(Y)$, la qual cosa al cas dels haplotips no és desitjable. Per això considerarem una variació de l'ARS que no necessita derivar: DFARS.

Creació de la funció recobriment

Recordem que el nostre objectiu és donar un recobriment per a la funció log-còncava $g(Y)$. Fixem-nos que si som capaços de recobrir el $\log(g(Y))$ funció definida a trossos, mitjançant rectes, podrem dir que la funció recobriment $G(Y)$ és *Piece-wise exponential*, és a dir, exponencial a trossos. Veiem el següent exemple:

Sigui $g(y)$ la funció log-còncava. Sigui $log(g(y))$ la funció definida en $[a, b]$. Siguin $c1$, $c2$ i $c3$ tres abscises pertanyents a $[a, b]$. Considerem dues secants a $log(g(y))$: sigui $r1$ la secant que uneix $(c1, log(g(c1)))$ amb $(c2, log(g(c2)))$, i sigui $r2$ la secant que va d'aquest darrer a $(c3, log(g(c3)))$. Per ser el $log(g(y))$ còncava, sabem que les seves secants queden per sota del gràfic, a l'interval en que es defineixen, i per sobre del gràfic si allarguem els segments tal i com es pot comprovar a la figura 8.3.5.

És a dir, suposem que:

$$\begin{cases} \log(g(Y)) < r1 \ si \ \ a < Y < c_1 \ \ c_2 \leq Y < c_3 \\ \log(g(Y)) < r2 \ si \ c_1 \leq Y < c_2 \ \ c_3 \leq Y < b \end{cases}$$

Llavors, prenent exponencials a ambdós costats de la desigualtat, obtenim l'*envelope*:

$$\begin{cases} g(Y) < \exp(r1) \ si \ \ a < Y < c_1 \ \ c_2 \leq Y < c_3 \\ g(Y) < exp(r2) \ si \ c_1 \leq Y < c_2 \ \ c_3 \leq Y < b \end{cases}$$

Per tant definim,

$$G(Y) = \begin{cases} \exp(r1) \ si \ \ a < Y < c_1 \ \ c_2 \leq Y < c_3 \\ \exp(r2) \ si \ c_1 \leq Y < c_2 \ \ c_3 \leq Y < b \end{cases}$$

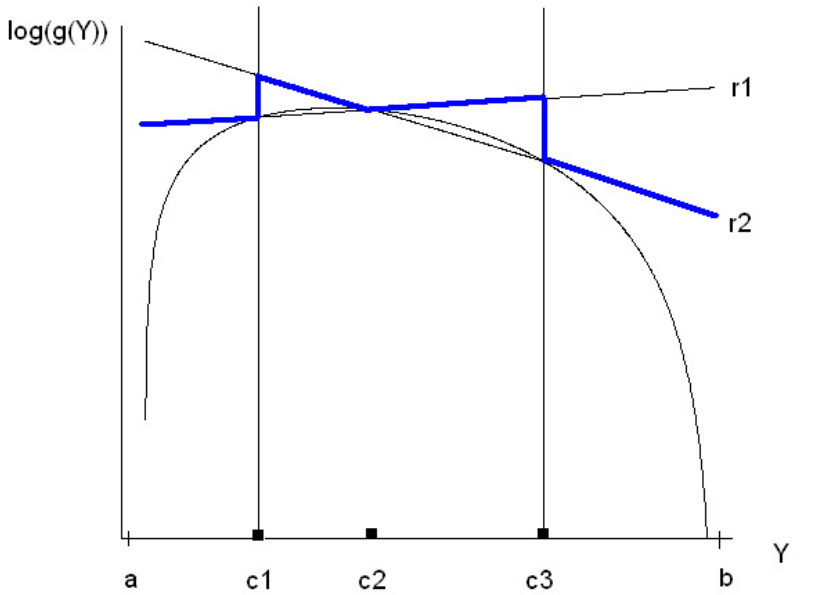


Fig.8.3.5 Els segments blaus ens serviran per definir l'*envelope* a $G(Y)$.

Així, $g(Y) < G(Y)$ com es volia. Per tant $G(Y)$ (figura 8.3.5 és una funció recobridora (l'*envelope* per la funció $g(Y)$ (funció proporcional a la funció d'interès).

Quantes abscises considerem i com les triem?

En general, tres o quatre abscises seran suficients, excepte en casos en que la massa de la funció estigui especialment concentrada. En cas de prendre més de tres abscises, podem observar a la figura 8.3.5 com allargant les secants es donen tres interseccions: dues amb les verticals i una entre dues secants.

Sobre quins punt escollir, caldrà triar dos punts tals que el màxim de la funció $log(g(Y))$ es trobi entre ells, i un tercer (i quart si s'escau) dins de l'interval definit per aquests dos. Si la funció està definida en un interval acotat, podem triar els dos extrems de l'interval, i un punt interior. Si no està acotada, una manera de garantir que la moda de la funció es

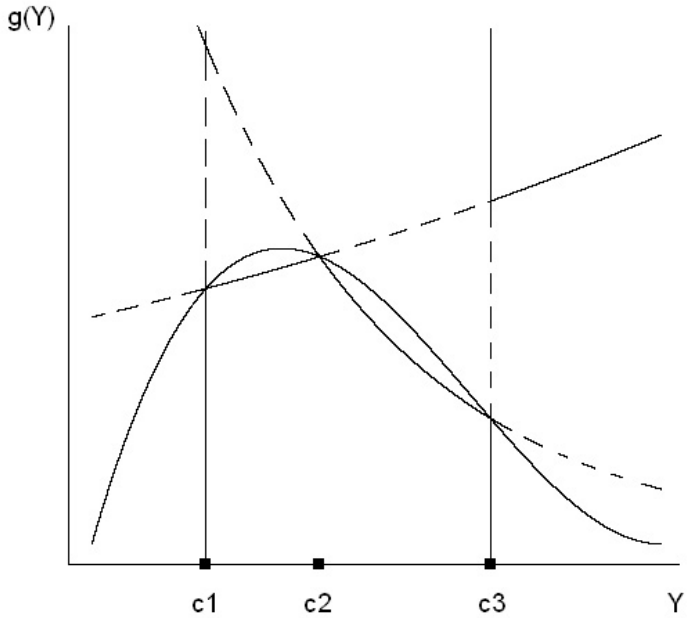


Fig.8.3.5 La linea continua és $g(Y)$. La discontínua correspon a l'*envelope* $G(Y)$.

troba entre els punts triats és prendre un punt per l'extrem esquerra amb derivada positiva, i anàlogament per la dreta, amb derivada negativa. Si és dificultós considerar la derivada (estem precisament en aquest cas) una bona solució és localitzar el màxim de la funció numèricament.

Concretem l'algorisme

Partim d'un conjunt d'abcises S. Sigui G_S la funció recobridora de $g(Y)$ per S. Fem el següent bucle:

repetir{

Mostregem Y de $G_S(Y)$

Mostregem U de $U(0,1)$

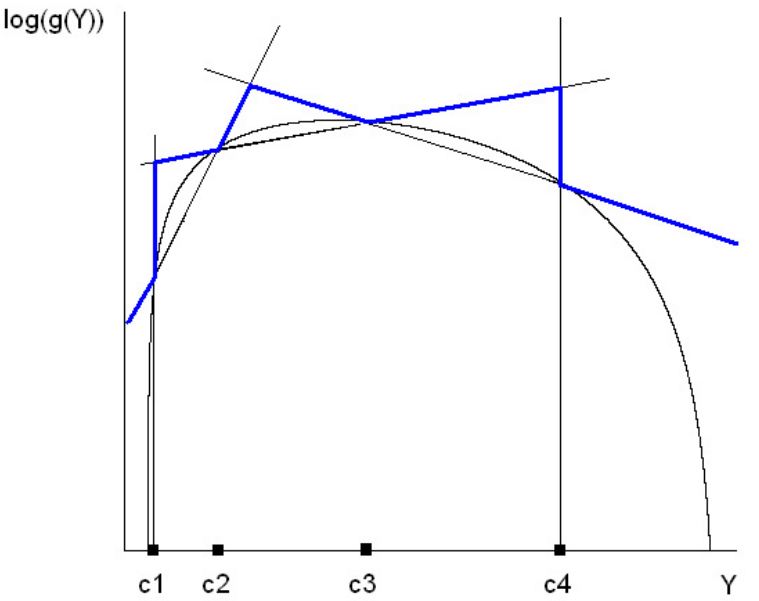


Fig.8.3.5

Si U$\leq g(Y)/G_S(Y)$ *Llavors*

acceptem Y

sino

adjuntem Y a S

}***Fins que acceptem un*** Y

observacions sobre l'algorisme:

1. Fixem-nos que en cas de no acceptar el candidat $\{Y\}$ en una iteració, l'algorisme afegeix aquest $\{Y\}$ al conjunt d'abscises S que ja teníem. Això provoca que en començar la nova iteració pel nou S, cal endreçar S, i refer el recobriment per aquest nou S (s'han de refer les secants, etc). Això ens duu al següent matís de notació.

2. **Notació**: Seria més correcte, pel que fa a notació, marcar Y i S amb el nombre d'iteració, per exemple, $Y^{(n)}$ i $S^{(n)}$. Igualment llavors, notaríem $G_{S^{(n)}}$ al recobriment obtingut pel nou conjunt d'abscises
$S^{(n)} = S^{(n-1)} \bigcup \{Y^{(n-1)}\}$ al que se li acaba d'adjuntar l'Y anterior, no acceptat. Ara que tenim clar qui es qui, seguirem utilitzant la notació simplificada.
3. El fet d'anar afegint abscises a S fa que la nova *envelope* cada cop sigui més propera a la funció g. Així, la probabilitat de rebuig cada cop va sent més petita ($g(Y)/G_S(Y)$ cada cop s'apropa més a 1).
4. Per poder dur a terme cada iteració necessitem saber com mostrejar de l'*envelope* $G(Y)$, funció exponencial a trossos. El mostreig de $G(Y)$ es fa en dues passes: Primer, mostregem un interval pels que tenim definits els consecutius trossos d'exponencial. Un cop triat un interval, hi mostregem un punt que hi pertanyi. Per mostrejar un interval, ho farem via una distribució categòrica. Necessitem saber la probabilitat que té cada interval de ser triat. Com que l'àrea sota $G(Y)$ no és 1, caldrà normalitzar l'àrea de cada sector. Sigui A l'àrea total i sigui A_{sector_i} l'àrea de $G(Y)$ sobre l'interval i-èssim. Llavors, cada interval té com a probabilitat de ser triat $p = \mathrm{A}_{sector_i}/\mathrm{A}$.

Mostreig d'una piece-wise exponential

Només falta precisar com s'escriu l'àrea de la funció recobridora sobre cada interval, i l'àrea total sota $G(Y)$. Calcularem les integrals que toqui.

Comencem explicitant l'àrea total. Sigui S_c el conjunt d'abscises en que canvia la definició de $G(Y)$. Notem que S i S_c coincidiran en cas que $\#S = 3$. Si $\#S > 3$, llavors $S_c = S \bigcup \{\text{interseccions entre les secants}\}$. Suposarem que $\#S_c = n$ i notarem els seus elements com y_i, $i = 1, \ldots, n$. Els n punts sobre la recta real defineixen $n + 1$ intervals que numerarem de 0 a n.

Llavors,

$\int_{-\infty}^{+\infty} G(Y)dY = \int_{-\infty}^{y_1} \exp(\alpha_0 Y + \beta_0)dY + \sum_{i=1}^{n} \int_{y_i}^{y_{i+1}} \exp(\alpha_i Y + \beta_i)dY +$

$\int_{y_n}^{+\infty} \exp(\alpha_n Y + \beta_n) dY = \mathrm{A}$

d'on,

$$\int_{-\infty}^{y_1} \exp(\alpha_1 Y + \beta_1) dY = \exp(\beta_1) \frac{\exp(\alpha_1 y_1)}{\alpha_1} \tag{8.18}$$

$$\int_{y_i}^{y_{i+1}} \exp(\alpha_i Y + \beta_i) dY = \exp(\beta_i) \frac{\exp(\alpha_i y_{i+1}) - \exp(\alpha_i y_i)}{\alpha_i} \; i = 1, \ldots, n \tag{8.19}$$

i,

$$\int_{y_n}^{+\infty} \exp(\alpha_n Y + \beta_n) dY = -\exp(\beta_n) \frac{\exp(\alpha_n y_n)}{\alpha_n} \tag{8.20}$$

Observació: La darrera integral convergeix perquè $\alpha_n < 0$.

Així doncs, la probabilitat de l'interval i-éssim és:

$$p_i = \frac{\mathrm{A}_{sector_i}}{\mathrm{A}} = \frac{(8.19)}{(8.18)+(8.19)+(8.20)} \; i = 0, \ldots, n$$

Un cop mostrejat un dels intervals, per mostrejar un punt de l'interior, generem U$\sim U(0,1)$ i prenem

$$Y = \frac{1}{\alpha_i} \log[\exp(\alpha_i y_i + U(\exp(\alpha_i y_{i+1}) - \exp(\alpha_i y_i))]$$

Així ja tenim el candidat Y.

Aplicació de l'Slice Sampling a la Gibbs sampling

Un altre mètode per mostrejar de funcions de densitat no estàndards, utilitzable a la Gibbs Sampling és la Slice Sampling.

Suposem que volem mostrejar valors d'una variable β que pren valors en cert subconjunt C_n i tal que la seva densitat és proporcional a certa funció $f(\beta)$. Podríem fer-ho mostrejant uniformement de la regió $(n+1)$-dimensional que cau just a sota del gràfic de $f(\beta)$. Aquesta idea es pot formalitzar introduint una variable auxiliar real y i definint la distribució conjunta sobre β i y, que és uniforme sobre la regió $U = \{(\beta, y) : 0 < y < f(\beta)\}$ sota la corba o superfície definida per $f(\beta)$. És a dir, la densitat conjunta per (β, y) serà

$$p(\beta, y) = \begin{cases} 1/Z, & \text{si} \quad 0 < y < f(\beta) \\ 0 & \text{en cas contrari} \end{cases} \tag{8.21}$$

on $Z = \int f(\beta) d\beta$. La densitat marginal per x és aleshores:

$$p(\beta) = \int_0^{f(\beta)} (1/Z) dy = f(\beta)/Z \tag{8.22}$$

Per mostrejar de β podem mostrejar conjuntament de (β, y) i després simplement ignorar y.

Generar punts independents mostrejats uniformement de U pot ser força complicat. Per això, una via per resoldre-ho pot ser generar una cadena de Markov que convergeixi a aquesta distribució uniforme. Aquesta és la idea general de l'Slice Sampling.

L'Slice Sampling és un mètode molt simple sempre i quan sigui aplicat a casos en que només una variable real estigui sent actualitzada. Aquest, per descomptat serà el cas de les distribucions univariades, però més habitualment també serà el cas de mostrejar de distribucions multivariants per $\beta = (\beta_1, \beta_2 ... \beta_n)$ que es mostregen univariadament, circulant per cadascuna de les variables una rera l'altra. Per actualitzar β_i és necessari calcular el valor que pren una funció $f_i(\beta_i)$ proporcional a $p(\beta_i | \{\beta j\}_{ji})$ on $\{\beta j\}_{ji}$ són els valors que prenen la resta de variables. Sovint, la distribució conjunta per $(\beta_1, ..., \beta n)$ serà definida per una funció $f(\beta_1, ..., \beta n)$ que és proporcional a la funció de distribució conjunta. En tal cas només cal prendre $f_i(\beta_i) = f(\dots, \beta i, \dots)$ on les variables diferents de β_i tenen els valors fixats.

Per simplificar notació, escriurem la variable a actualitzar com β sense subíndex, i els subíndex denotaran punts diferents i no components del mateix vector. La funció proporcional a la densitat de probabilitat de β la denotarem per $f(\beta)$. El mètode d'Slice Sampling que descriurem aquí reemplaça el valor actual β_0 per un valor β_1 que es dedueix segons el següent procediment basat en tres passes:

1. Mostrejar un valor real y de manera uniforme en l'interval $(0, f(\beta_0))$, definint un tall horitzontal (una *slice*) $S = \{\beta : y < f(\beta)\}$. Notem que β_0 es troba sempre dins de S.

2. Buscar un interval I=(L,R) al voltant de β_0 que contingui tota o almenys gran part de l'slice.
3. Mostrejar un nou punt β_1 que pertanyi a la part de l'slice dins d'aquest interval.

El primer agafa un valor de la variable auxiliar que és característica de l'Slice Sampling. Fixem-nos que no hi ha cap necessitat de retenir aquest valor entre diferents passes de la cadena de Markov, donat que aquest valor per y és oblidat per la següent iteració. A la pràctica, és molt habitual treballar amb $g(\beta) = log(f(\beta))$ en comptes de fer-ho amb $f(\beta)$ per evitar possibles problemes de valors massa petits. Un pot utilitzar la variable auxiliar $z = log(y) = g(\beta_0) - e$, on e és exponencialment distribuïda, amb mitjana igual a 1, i definint l'slice segons: $S = \{\beta : z < g(\beta)\}$.

La segona i tercera passa es poden implementar de diferent manera. Independentment da la via triada, el resultat serà una cadena de Markov amb distribució invariant $f(\beta)$. La figura (8.3) mostra un mètode aplicable en termes generals, tal que l'interval és trobat mitjançat una tècnica de *stepping out*, i el nou punt es mostreja seguint un procediment de *shrinkage* en anglès o "encongiment"en català. Amb les tres passes que es mostren, s'acaba generant un nou punt β_1 que serà el següent a β_0 en el mostreig. Al pas 1, es mostreja verticalment el punt y de l'interval $(0, f(\beta_0))$. Al pas 2, un interval de llargada w es posiciona aleatòriament al voltant de β_0 i després s'expandeix en passes d'amplada w fins que els dos extrems es troben fora de l'slice. I a la tercera passa, es mostreja un nou punt β_1 uniformement de dins l'interval, fins que es troba un que estigui situat dins de l'slice. Els punts que es troben i estan fora de l'interval s'utilitzen per redimensionar el propi interval, encongint-lo.La figura (8.4) mostra una manera alternativa per trobar l'interval. En (a) l'interval inicial es duplica dos cops, fins que els dos extrems es troben fora de l'slice. A (b), en que l'inici és un altre, no es fa cap duplicació.

Com trobar un interval apropiat?

En aquesta tesi s'han considerat aquestes dues maneres de generar l'interval. El procediment d'Stepping out és apropiat per qualsevol distribució, sempre i quan sigui possible

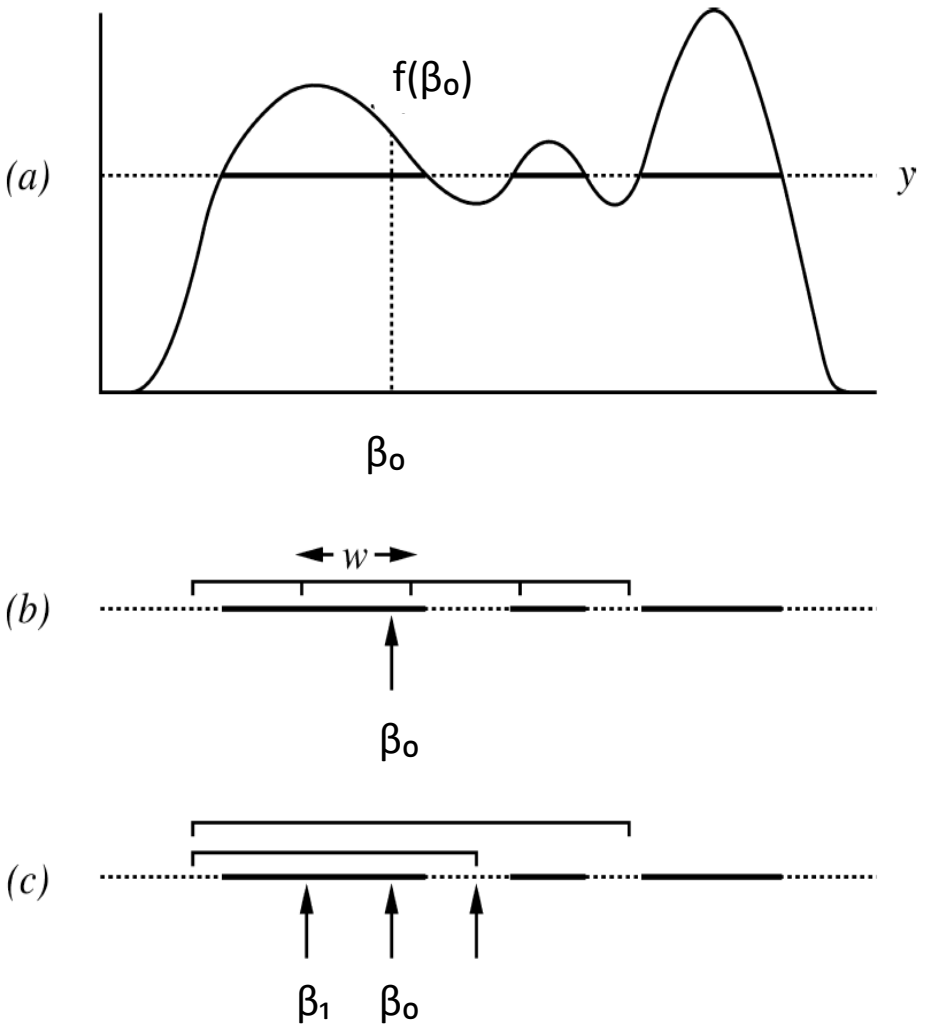


Figura 8.3. Un pas de l'slice sampling utilitzant procediments de stepping-out i shrinkage.

proporcionar un valor w que s'ajusti a l'amplada general de l'slice. La descripció gràfica de com es troba un interval segons aquest procediment ja s'ha vist a la figura (8.3). La descripció detallada de l'algorisme és la següent:

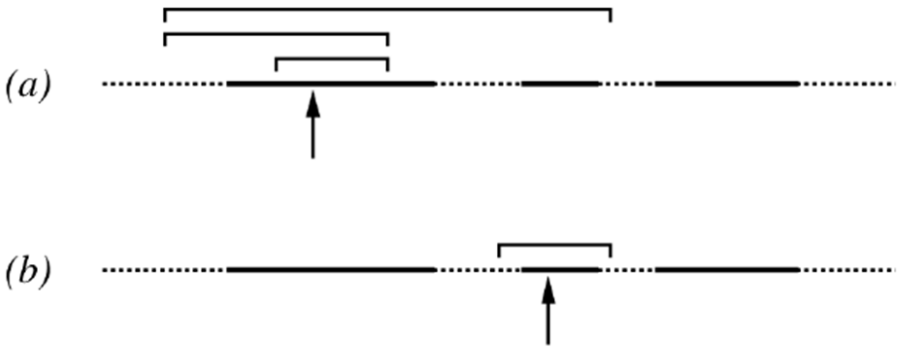


Figura 8.4. El procediment de doubling.

INPUT:

- f = funció proporcional a la densitat
- β_0 = el punt actual
- y = el valor fixat a l'eix vertical que defineix l'slice a la funció
- w = Estimació de la mida habitual de l'slice
- m = enter que limita l'amplada de l'slice a mw.

OUTPUT: (L,R) l'interval buscat.

ALGORISME:

$$\begin{cases} U \sim Unif(0,1) \\ L \leftarrow \beta_0 - w * U \\ R \leftarrow L + w \\ V \sim Unif(0,1) \\ J \leftarrow part - entera(m * V) \\ K \leftarrow (m-1) - J \end{cases}$$

Repetir mentres $J > 0$ i $y < f(L)$:

$$\begin{cases} L \leftarrow l - w \\ J \leftarrow J - 1 \end{cases}$$

Repetir mentres $K > 0$ i $y < f(R)$:

$$\begin{cases} R \leftarrow R + w \\ K \leftarrow K - 1 \end{cases}$$

A diferència del stepping out, el mètode de doubling pot eixamplar l'interval més ràpidament i per tant ser més eficient quan l'estimació de l'amplada w tendeix a ser massa petita. Aquest procediment il·lustrat a la figura (8.4) és descrit tot seguit.

INPUT:

- f = funció proporcional a la densitat
- β_0 = el punt actual
- y = el valor fixat a l'eix vertical que defineix l'slice a la funció
- w = Estimació de la mida habitual de l'slice
- p = enter que limita l'amplada de l'slice a $2^p w$.

OUTPUT: (L,R) l'interval buscat.

ALGORISME:

$$\begin{cases} U \sim Unif(0,1) \\ L \leftarrow \beta_0 - w * U \\ R \leftarrow L + w \\ K \leftarrow P \end{cases}$$

Repetir mentres $K > 0$ i $\{y < f(L)$ o bé $y < f(R)\}$:

$$\begin{cases} V \sim Unif(0,1) \\ \text{if} V < 0.5 \text{ then } L \leftarrow L - (R - L) \\ \text{else } R < -R + (R - L) \\ K < -K - 1 \end{cases}$$

L'algorisme de l'Slice Sampling genera una cadena de Markov que té com a distribució invariant la desitjada, utilitzant qualsevol d'aquests dos mètodes. Aquesta convergència ve garantida pel fet que la cadena resultant és ergòdica. Per veure les demostracions d'aquest fet ens podem adreçar a [21].

9

Punt de trobada entre MCMC, l'estadística Bayesiana i el problema haplotípic

Els mètodes Bayesians tal i com hem vist, permeten treballar amb distribucions per paràmetres que inicialment són desconegudes. Aquest fet ofereix un ampli ventall de possibilitats perquè en cas de conèixer la funció de versemblança per una mostra tal que involucri un vector de paràmetres, acte seguit i segons (7.2) tenim una manera d'escriure l'expressió per la distribució posteriori multivariada del conjunt de paràmetres. Així és, la distribució posterior és proporcional al producte entre la funció de versemblança i una distribució prior. Tenint en compte que sempre podem triar com a distribució prior la menys informativa, podem fins i tot considerar que la distribució posterior és directament proporcional a la funció de versemblança de la mostra. Un cop aquesta funció està definida, les tècniques MCMC detallades en aquest treball permetran aconseguir un mostreig per cadascun dels paràmetres de la distribució.

Per tal de poder aplicar aquesta teoria a la resolució del problema haplotípic ens cal aclarir diverses qüestions:

- Quina serà la distribució a posteriori amb què treballarem i per tant explicitar la funció de versemblança de la mostra haplotípica i les possibles distribucions a priori.
- Quins seran els models que utilitzarem per estimar associació entre haplotips i fenotips, i explicitar les funcions de versemblança implicades.
- Quina és la tècnica MCMC més adient per dur a terme el mostreig de les freqüències haplotípiques i dels coeficients dels models.

9.1 Funció de versemblança per les freqüències haplotípiques

Tots els mètodes d'estimació haplotípica basats en el mètode de la màxima versemblança, incloent el mètode que es presenta en aquest treball, necessiten l'especificació de la funció de versemblança de la mostra haplotípica. Es tracta d'una funció de versemblança complexa, que té com a paràmetres les freqüències haplotípiques de la mostra genotípica. La complexitat és deguda als individus amb haplotips incerts pels que, com es veurà tot seguit, cal considerar totes les possibles parelles d'haplotips compatibles amb el seu genotip i incorporar-les a la funció.

Descripció de la funció

Sigui G el conjunt de genotips d'una mostra de N individus on cada individu té un genotip g_i, $i = 0, \ldots, N$. En funció d'aquest genotip, cada individu pot tenir un nombre finit d'haplotips compatibles amb g_i. Si aquest genotip té com a molt un locus heterozigot, l'individu només pot portar una parella d'haplotips. En cas que tingui més d'un locus heterozigot, l'individu pot dur 2^m haplotips diferents, on m és el nombre de locus heterozigots.

Siguin $f_{h_1}, \ldots, f_{h_{2^m}}$ les freqüències de cada haplotip possible a la mostra. Considerant que es dóna equilibri de lligament, la freqüència de cada genotip F_{g_i} és el producte de les freqüències dels haplotips. En cas que el genotip d'un individu sigui compatible amb una sola parella d'haplotips (h_r, h_s), aleshores la freqüència del genotip és $f_{h_r} f_{h_s}$. Ara bé, si el genotip g_i de l'individu i-èssim és compatible amb més d'una parella d'haplotips, aleshores $F_{g_i} = \sum_{h_r, h_s \in H_i} c_{rs} f_{h_r} f_{h_s}$ on H_i és el conjunt d'haplotips compatibles amb el genotip g_i, i c_{rs} és una constant que val 1 si $h_r = h_s$ i 2 si $h_r \neq h_s$.

Un cop clarificat com escriure la freqüència per cada cas de la mostra, la funció de versemblança serà el productori d'aquestes freqüències sobre el total de la mostra de genotips:

$$\ell(F) = \prod_{i=1}^{N} F_{g_i} = \prod_{i=1}^{N} \sum_{h_r, h_s \in H_i} c_{rs} f_{h_r} f_{h_s} \tag{9.1}$$

on $F = \{F_{g_i}, i = 0, \ldots, N\}$.

Donada la complexitat de la maximització analítica d'aquesta funció, computacionalment s'ha optat per mètodes d'estimació numèrica, com l'algorisme EM o les tècniques de Markov Chain Monte Carlo que han estat les triades en aquest treball.

9.2 Models estadístics segons el tipus de disseny i funcions de versemblança associades

El context dels estudis d'associació permet considerar diverses classes de dissenys que proporcionaran diferents tipus i quantitat de dades resultants. Com s'ha comentat a l'apartat introductori, el disseny d'estudi més utilitzat degut a la seva potència a l'hora d'identificar associacions entre una variant i cert fenotip i degut també al seu cost-efectivitat en la recollida de dades, és l'estudi de cas-control. En aquest estudi es recullen dades retrospectivament en una mostra de casos (individus que pateixen la malaltia) i en una mostra de controls (individus que no presenten la malaltia). Un altre disseny que a diferència del de cas control permet establir ordre temporal entre esdeveniments, és l'estudi longitudinal de cohorts, adient per estudiar l'aparició d'esdeveniments en funció del temps.

Cadascun d'aquests estudis té associat un model estadístic concret, que ve definit pel tipus de variable resposta. Donat que per a aquests estudis, la variable resposta no és quantitativa ni es distribueix de manera normal, s'utilitzen els anomenats Models Lineals Generalitzats (GLM).

9.2.1 Model Lineal generalitzat: Regressió Lineal, Regressió Logística i Regressió de Weibull

Model de Regressió Lineal

El model Lineal habitual s'escriu com

$$y_i = \beta x_i + \epsilon_i$$

amb Y una variable contínua, X un conjunt de covariables i complint-se un conjunt d'hipòtesis de centralitat, normalitat, independència i homocedasticitat pels errors ϵ_i. Com a conseqüència que els errors tinguin esperança zero, passa que $E(Y|X) = X\beta$. Per aquest model la funció de versemblança és la següent:

$$\prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{\left(-\frac{(y_i - x_i\beta)^2}{2\sigma^2}\right)}$$

Segons la distribució que segueix la resposta Y és possible aplicar-li transformacions mitjançant una funció "link" g que sigui contínua i invertible, de tal manera que sigui possible escriure $g(E(Y|X)) = X\beta$.

Model de Regressió Logística

El model logístic s'utilitza pel cas de variables resposta binàries, com és el cas dels estudis de cas-control.

Sigui $Y = \{y_i\}_{1 \leq i \leq N}$ la variable resposta que pren valors 1 o 0. Sigui X la matriu de covariables de dimensió $N \times M$ i $\beta = (\beta_0, \ldots, \beta_{M-1})$ el vector de coeficients. Sigui p la proporció $p = P(Y = 1|X)$. Notem que $p \in (0, 1)$ i que la combinació de covariables i coeficients $X\beta$ no té perquè pertànyer a aquest rang. Per això, triem una funció link tal que g^{-1} porti $X\beta$ a $(0, 1)$. La funció link per aquest model és $g(p) = \log\left(\frac{p}{1-p}\right)$ i per tant:

Definició 9.2.1 El model logístic s'escriu com:

$$\log\left(\frac{p}{1-p}\right) = X\beta \tag{9.2}$$

i per un individu concret escrivim:

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 X_{i0} + \cdots + \beta_{M-1} X_{iM-1} \tag{9.3}$$

on X_{ij} representa l'entrada ij-èssima de la matriu de regressores, és a dir, es tracta del valor de la variable j-èssima observada per l'individu i-èssim.

De fet, podem comprovar que la probabilitat que $y_i = 1$, és

$$p_i = \frac{e^{X_i\beta}}{1 + e^{X_i\beta}} \tag{9.4}$$

Per veure-ho només cal aïllar p_i de (9.3) i veure que efectivament, $p_i \in (0, 1)$.

Funció de versemblança

Donada una mostra de N individus la variable Y pot prendre els valors 0 o 1. Per tant, $Y \sim Bernoulli(p)$

$$\begin{cases} y_i = 1 \ amb \ p_i \\ y_i = 0 \ amb \ 1 - p_i \end{cases}$$

i la funció de versemblança és:

$$\prod_{i=1}^{N} p_i^{y_i}(1-p_i)^{1-y_i} \tag{9.5}$$

on p_i és la de (9.4).

Estimació dels coeficients

En aquest treball, els coeficients del model logístic els estimem via mètodes MCMC.

Interpretació dels coeficients: L'Odds Ratio

Els coeficients d'una regressió Logística quantifiquen el risc que la variable resposta prengui valor 1 en funció del valor que prengui el factor de risc considerat. Això és així donat que aquests coeficients permeten definir l'*Odds Ratio*.

La odds d'un esdeveniment és el quocient entre la probabilitat de que passi l'esdeveniment, i la probabilitat de que no passi. És a dir, si p és aquesta probabilitat,

$$odds = \frac{p}{1-p}$$

La odds és una mesura de risc. Notem que $(1-p) \times odds = p$. Per tant, la odds ens diu quantes vegades més probable és que passi l'esdeveniment respecte de que no passi. Si considerem ara un factor de risc amb diferents nivells, podem calcular la Odds sobre els diferents valors d'aquest factor. El quocient entre Odds calculades per dos d'aquests nivells es coneix com l'*Odds Ratio* conegut amb les inicials "OR´´.

$$OR = \frac{odds(Y = 1| + X)}{odds(Y = 1|\neg X)}$$

Donat un model logístic amb coeficients (α, β) tenim que

$$OR = e^{\beta}$$

Per comprovar-ho només cal substituir les definicions d'odd a l'OR i recordar (9.4). Per tant, el coeficient β quantifica la magnitud de l'associació entre la resposta i el factor de risc d'interès.

Al cas dels haplotips, el coeficient quantificarà l'aportació sobre el risc de patir una malaltia que fa el fet de dur un haplotip respecte el fet de dur-ne un altre de referència (habitualment, el més freqüent a la mostra).

Una condició bàsica que cal que es compleixi per tal de poder utilitzar regressió Logística en un disseny de cas-control és que es compleixi l'equilibri de Hardy Weinberg tant pels casos com pels controls. Això es tradueix a tenir penetrància multiplicativa, és a dir, cada còpia de l'haplotip i contribueix al risc de malaltia tal que $OR_{ij} = OR_{1i}OR_{1j}$, d'on OR_{ij} és l'odds ratio que compara l'haplotip (i, j) respecte el de referència.

Model de Regressió de Weibull

Les dades recollides segons un estudi longitudinal de seguiment d'una cohort de persones es poden analitzar segons diferents vies. En aquest treball considerarem la opció paramètrica i prendrem un model Lineal generalitzat (GLM) amb funció link la distribució de Weibull.

Definicions bàsiques

Siguin:

a)

$$S(t) = P(T > t)$$

amb $t \geq 0$ la *Funció de Supervivència*. És la funció que mesura la probabilitat de sobreviure a un esdeveniment més temps que t.

b)

$$f(t) = \lim_{\triangle t \to 0} P(t \leq T < t + \triangle t)$$

És la *Funció de densitat* i s'interpreta com la probabilitat que l'esdeveniment es dongui a temps t.

c)

$$\lambda(t) = \lim_{\triangle t \to 0} \frac{1}{\triangle t} P(t \leq T < t + \triangle t \mid T \geq t)$$

correspon a la *Funció de risc* que calcula la probabilitat de que un individu d'edat t (és a dir, un individu viu fins aquell moment) pateixi l'esdeveniment.

Relació bàsica

De les 3 definicions se'n deriva la següent relació:

$$f(t) = \lambda(t)S(t)$$

Censures

Sigui C_R la data fixada per finalitzar un estudi. Suposem que no tots els individus han entrat al mateix temps, i que per tant cadascun té un temps màxim d'estada en l'estudi

diferent (el que va del moment que entra fins a C_R). Sigui C_i aquest temps d'observació per a cada individu. Direm que l'individu no està censurat si pateix l'esdeveniment abans de C_i. Si no, direm que està censurat. Sigui T_i el temps en què pateix l'esdeveniment. Definim l'indicador de no censura:

$$\delta_i = \begin{cases} 1 \; si \; T_i \leq C_i \; no - censura \\ 0 \; si \; T_i > C_i \; censura \end{cases}$$

Funció de versemblança

Mitjançant la relació bàsica, la funció de versemblança en cas que no hi hagin dades censurades s'escriuria com:

$$L(t) = \prod^{n} f(t) = \prod^{n} \lambda(t) S(t)$$

La informació per cada individu es representarà amb un parell (Y, δ) on Y representa el valor per la variable temps i prendrà el valor $\min\{T_i, C_i\}$. En cas que l'individu no estigui censurat $(Y = y, \delta = 1)$, la contribució de l'individu a la versemblança serà:

$$P(y, \delta = 1) = P(Y = y, T_i \leq C_i) = P(T_i = y, T_i \leq C_i) = P(T_i = y) P(C_i \geq y) \tag{9.6}$$

Estem suposant independència entre el temps en que passa l'esdeveniment i el temps de censura. Si l'individu presenta censura per la dreta $(Y = y, \delta = 0)$, la seva contribució ve donada per:

$$P(y, \delta = 0) = P(Y = y, T_i > C_i = P(C_i = y, T_i > y) = P(C_i = y) P(T_i > y) \tag{9.7}$$

1. Les probabilitats resultants en ambdues expressions (9.6) i (9.7) corresponen a funcions de densitat i de supervivència.

2. A les darreres igualtats s'utilitza que el temps assimilat com a temps d'esdeveniment per l'individu censurat és el temps final d'observació.

3. Unint les dues expressions obtenim una expressió general per la contribució de cada individu:

$$P(y,\delta) = (P(T_i = y)P(T_i \leq C_i))^{\delta_i}(P(C_i = y)P(T_i > y))^{1-\delta_i}$$

que podem expressar en funció de les funcions de densitat i de supervivència de T i C.

Siguin f i g les funcions de densitat i S i G les de supervivència per T i C respectivament. Escriuríem:

$$P(y,\delta_i) = (f(y)G(y))^{\delta_i}(g(y)S(y))^{1-\delta_i}$$

Ara ja podem escriure la funció de versemblança per una mostra d'n individus:

$$\prod_{i=1}^{n} P(y_i,\delta_i) = (f(y_iG(y_i))^{\delta_i}(g(y_i)S(y_i))^{1-\delta_i}$$

Si ara suposem que:

1. C no censura informativament a T
2. El suports per C i T són diferents

Llavors com que el que volem estimar és la distribució dels temps T aquesta no dependrà de la distribució de C i per tant podem escriure la versemblança com:

$$L = \prod_{i=1}^{n} f(y_i)^{\delta_i} S(y_i)^{1-\delta_i} \tag{9.8}$$

o bé, aplicant la relació bàsica:

$$L = \prod_{i=1}^{n} \lambda(y_i)^{\delta_i} S(y_i) \tag{9.9}$$

Per tant un individu censurat per la dreta, tindrà $\delta = 0$ i contribuirà amb $S(y_i)$ on $y_i = C_i$.

Ara només ens falta aplicar tot això al nostre cas:

Distribució de Weibull

Per aquesta distribució de temps tenim:

- Funció de supervivència: $S(t) = e^{-(\rho t)^k}$
- Funció de risc: $\lambda(t) = k\rho(\rho t)^{k-1}$ on $t > 0$ i $k > 0$
- Funció de densitat: $f(t) = k\rho(\rho t)^{k-1}e^{-(\rho t)^k}$

Per tant la funció de versemblança s'obté multiplicant $\lambda(t)$ i $f(t)$:

$$L = \prod_{i=1}^{n} k\rho_i(\rho_i t_i)^{k-1} e^{-(\rho_i t_i)^k} \tag{9.10}$$

Sigui z_i el valor d'una covariable per l'individu i-èssim i sigui β els coeficients de la regressió. Fem les parametritzacions $\mu_i = {\rho_i}^k$ i seguidament $\mu_i = e^{\beta z_i}$. Substituint ens queda:

$$L = \prod_{i=1}^{n} k e^{\beta z_i} {t_i}^{k-1} e^{-e^{\beta z_i} t^k} \tag{9.11}$$

I aquesta serà la funció de versemblança que utilitzarem en aquest treball per crear una cadena de Markov que circularà pels paràmetres (β, k).

9.3 Distribucions a priori per a cadascun dels models

El problema de l'elecció de la distribució a priori és una de els principals qüestions que hom ha d'afrontar quan decideix fer una anàlisi Bayesiana. En cas que l'investigador compti amb una creença prèvia sobre els paràmetres a estimar i vulgui incloure aquesta informació a l'anàlisi podrà fer-ho mitjançant la distribució prior (7.1). En principi, aquesta distribució pot ser qualsevol que l'investigador cregui oportuna.

En aquest treball en cas que es desitgi introduir informació a priori s'ha considerat la familia de distribucions normals per a cada coeficient del model de regressió:

$$\beta_i \sim N(\mu_i, \sigma_i^2)$$

Es tracta d'una classe de distribucions a priori molt flexible pel cas de models de regressió com assenyala Geisser al seu llibre [176].

En cas que no es disposi de cap coneixement previ sobre els paràmetres, en aquest treball es considera per defecte una distribució uniforme no informativa que dóna mateixa probabilitat a tots els possibles valors.

9.4 Aplicació de tècniques MCMC per l'estimació dels paràmetres

Tot i que els resultats sobre MCMC a nivell teòric indiquen que les diferents tècniques podrien ser aplicades de forma gairebé indistinta a l'estimació de paràmetres, a la pràctica

ens trobem que la convergència teòrica pot no assolir-se en un interval de temps computacionalment òptim per l'usuari. És per això que no totes les tècniques són adients per a cadascuna de les versemblances de les quals haurem de mostrejar. El fonament d'aquesta tria s'ha basat en implementar diferents mètodes i comprovar si la seva aplicació pràctica era possible a nivell de temps d'execució. Les tècniques testades han estat:

- Algorisme de Metropolis
- Algorisme de la Gibbs Sampling
- Mètodes DFARS
- Mètode Slice Sampling

Tots ells són mètodes a priori adients per les versemblances amb que s'ha treballat: la referent als haplotips (9.1), al model Lineal (9.2.1), al Logístic (9.2.1) i al de Weibull (9.8). L'únic supòsit que necessitàvem pel cas de DFARS era la log-concavitat de les funcions a mostrejar, i efectivament, les condicionals del model Lineal, Logístic i de Weibull la compleixen [175]. Aquestes propietats també es compleixen en cas que s'incorpori informació a priori a l'anàlisi i per tant aquestes versemblances vinguin multiplicades per les distribucions prior pels paràmetres que han estat especificades a la secció anterior.

Alguns mètodes convergeixen amb poques iteracions, però la quantitat de càlculs que cal dur a terme per generar cada component de la cadena és tan costós computacionalment, que el mètode no resulta útil a la pràctica. Amb d'altres mètodes passa el contrari, es necessita un nombre superior d'iteracions per a que la cadena de Markov convergeixi, però la creació de cada estat de la cadena té pocs requeriments a nivell informàtic i permet implementar-la en un temps òptim.

9.4.1 Algorisme de Metropolis per estimar les freqüències haplotípiques

Per estimar els valors de les freqüències haplotípiques s'ha utilitzat l'algorisme de Metropolis (8.3.3) en la seva versió de Random Walk. Es tracta d'una aplicació senzilla que reporta

molt bons resultats per aquesta funció de versemblança. Necessita pocs termes per la cadena, amb 1000 termes n'acostuma a haver prou. Així, mitjançant el mètode de Random Walk obtenim una cadena de 1000 termes per cadascuna de els freqüències. Cadascuna d'aquestes cadenes té per distribucìó invariant la distribució posterior de cadascuna de les freqüències enteses com a variables aleatòries. Per tant, s'obté un mostreig.

Per facilitar l'entesa, considerarem una variació de la notació de la secció 9.1 i notarem $f_r = f_{h_r}$ com la freqüència de l'haplotip r-èssim a la població. Sigui M el nombre d'haplotips possibles a la població. La variable a qui volem donar una densitat serà $f = (f_1, f_2, \cdots, f_M)$. Per tant crearem una cadena de Markov multivariada per aquesta variable, és a dir, es construiran M cadenes de Markov.

Per aquest mètode cal definir quina serà la distribució proposada. S'han testat dues distribucions, una uniforme i una normal, i en tots dos casos la distribució límit acaba coincidint.

Com hem vist a la observació 8.3.3, tant al cas de la normal com al de la uniforme, la rapidesa de la convergència de la cadena depèn en part de la desviació que triem (és a dir, de si fem el salt més o menys gran).

9.4.2 DFARS i Slice Sampling per estimar l'associació amb fenotip

L'associació amb el fenotip es duu a terme amb els models ja exposats, tenint com a variable de risc la reconstrucció haplotípica. En aquest análisi d'associació com hem vist podran prendre part tres models: el Lineal, el Logístic i el de Weibull. Després d'haver considerat per a aquests models el mateix mètode que per les freqüències haplotípiques, el Random-Walk, es constata que la convergència no és gens òptima i el temps fins la convergència és excessivament alt. El mètode que acaba donant millors resultats és la Gibbs sampling. L'algorisme funcionarà exactament igual pel model Lineal, pel logístic i pel model Weibull, només que per la continua i pel Weibull a més dels coeficients haurem d'estimar el paràmetre σ^2 referent a la variància i el paràmetre k referent a l'escala, respectivament.

Així doncs, en genèric crearem $M + 1$ cadenes ($M + 2$ pel cas Weibull i Lineal) cadascuna d'elles referents a cada component del vector de paràmetres $\beta = (\beta_1, \ldots, \beta_M)$. Recordem que la Gibbs Sampling mostreja de les distribucions condicionals de les versemblances de cadascun dels models. En aquest cas, el mostreig de la distribució condicional no ha estat immediat, havent d'implementar diversos mètodes de mostreig per densitats multivariades complexes com els que s'han vist a la secció 8.3.5. La log-concavitat de les funcions en qüestió ens ha permès la utilització d'aquests mètodes.

En primer lloc es considerà el mètode ARS i en particular la seva versió lliure de derivades, la DFARS. La construcció de la funció recobriment a cada pas de l'algorisme requereix diverses avaluacions de les funcions que intervenen, fent del mètode un via molt poc òptima a nivell computacional. Per això, va ser substituït en favor de l'Slice Sampling. Aquest mètode millora notablement l'anterior, en la rapidesa de convergència a nivell de nombre d'iteracions i en el temps que triga per cada iteració. A nivell de programació, la complexitat d'un envers l'altre és incomparable. L'Slice Sampling és un algorisme més senzill i amb menys requeriment computacional que el DFARS. En particular es tria el mètode d'Stepping out exposat gràficament a la figura 8.3.

9.5 Els haplotips com a factor de risc: estimació simultània

En aquest treball, la parella d'haplotips que dugui cada individu juga el paper de factor de risc del model que hagi estat considerat. Però com tractem la incertesa haplotípica? Mateixos individus poden tenir més d'una parella haplotípica. Com podem introduir aquesta informació en el model? Fixem-nos que si utilitzem el mètode d'imputació haplotípica, àmpliament utilitzat encara en l'actualitat, en un primer pas reconstruiríem la mostra d'haplotips i després, estudiaríem l'associació entre aquests haplotips i la malaltia. Aquí cal aturar-nos i posar especial èmfasi en el següent fet: la mostra d'haplotips aconseguida per imputació no és única. Cal recordar que s'ha trobat mitjançant inferència estadística, i

que per tant, arrossega un error. Això vol dir que potser en altre cas, els individus amb haplotips incerts se'ls hagués resolt amb una altra fase, se'ls hagués assignat una altra parella d'haplotips. Aquesta qüestió ha estat tractada àmpliament a **??**.

El mètode de tractament de la incertesa que utilitzarem és un mètode que com ja hem vist s'ha demostrat eficient en relació a d'altres d'existents pel tractament de la incertesa haplotípica, que es basa en fer una estimació simultània de les freqüències haplotípiques i dels efectes associats a cada haplotip.

Part IV

ALGORISME DISSENYAT EN AQUESTA TESI.

IMPLEMENTACIÓ INFORMÀTICA

10

L'algorisme que hem creat

El principal objectiu d'aquesta tesi és el de dissenyar un mètode Bayesià per analitzar l'associació entre una mostra haplotípica i diverses classes de fenotip d'interès. Si la informació haplotípica fos coneguda, la qüestió no tindria més interès que el de realitzar una anàlisi d'associació similar a la que es duu a terme pel cas dels SNPs, codificant la informació haplotípica en categories i analitzant l'associació mitjançant el model més adient. Ara bé, com ja s'ha exposat en aquest treball, les tècniques de laboratori per separar cromosomes resulten poc cost-efectives i el més habitual és que la mostra genotípica no diferenciï en quin cromosoma es troba cadascun dels al·lels genotipats per SNP. Així doncs, la incertesa inherent a la mostra haplotípica fa que l'anàlisi de l'associació entre fenotips i haplotips no sigui immediata.

Expressant-nos en termes pràctics, considerem que partim d'una mostra d'individus pels que tenim genotipats un conjunt d'SNPs. A més de la informació genètica de cada individu, suposem que també tenim recollida informació sobre si han desenvolupat certa malaltia o no, potser també sabem si durant un interval de temps han estat lliures de malaltia o bé coneixem alguna mesura quantitativa que ens interessa estudiar en relació a la genètica de l'individu. Aquestes dades ens permeten realitzar una anàlisi d'associació entre les diferents característiques i els SNPs, estudiar mesures de recombinació i LD. Però donat que els individus amb dos o més locus heterozigots no tenen la seva parella d'haplotips definida prèviament, per fer una anàlisi d'associació en relació als haplotips cal que primer els re-

construïm.

En aquest context, utilitzarem els mètodes MCMC per:

1. Estimar les freqüències haplotípiques per salvar la incertesa de la mostra i així poder-ne reconstruir els haplotips.
2. En funció d'aquesta reconstrucció, estimar el risc de malaltia o la supervivència associada als haplotips.

El mètode d'estimació i anàlisi haplotípica que presentem en aquesta tesi és Bayesià. Ho és en tant que utilitza conceptes Bayesians en el tractament de la informació i en tant que els resultats que retorna són propis de la inferència Bayesiana. Pel que fa a la utilització de funcions prior, l'algorisme permet la introducció d'aquestes distribucions. Un cop definides la versemblança per les freqüències haplotípiques i pels tres models estadístics considerats (9.2.1,**??**,9.10), des del punt de vista Bayesià la distribució de la que mostrejarem serà proporcional a cadascuna d'aquestes versemblances en cas que considerem una distribució prior igual a 1. En cas que considerem una distribució prior diferent, com ara la distribució normal amb paràmetres mitjana i variància fixats coneguts, caldrà considerar el producte d'aquesta distribució prior per la funció de versemblança.

El tractament de la incertesa haplotípica és una qüestió clau en l'algorisme. Com hem vist a la part de mètodes secció 9.5 dedicada a aquesta qüestió i com ja s'ha fonamentat a la introducció, l'algorisme realitzarà l'estimació simultània de freqüències haplotípiques i dels paràmetres d'associació. Aquest fet es contrastarà mitjançant diferents aplicacions exposades a l'apartat de resultats.

En essència, l'algorisme que hem creat és iteratiu, i a cada pas reconstrueix la mostra haplotípica i calcula l'associació entre la reconstrucció actual i el fenotip fixat, construint pas a pas, amb cadascuna d'aquestes estimacions, una cadena de Markov per cadascun dels paràmetres implicats. És així com l'algorisme acaba generant un mostreig per cadascun dels paràmetres. En l'estimació d'aquests paràmetres és on intervenen els mètodes MCMC.

10.1 L'algorisme pas a pas

Fins aquí ja ho hem explicat gairebé tot. Hem vist quins són els paràmetres que ens permeten resoldre l'anàlisi d'associació entre fenotips i haplotips. Hem vist com s'expressen les funcions on intervenen aquests paràmetres i com els podem estimar. També hem reflexionat sobre com podem tractar la incertesa haplotípica. Així doncs, arribats a aquest punt, el que queda per fer és unir-ho tot plegat. Així s'ha dissenyat un algorisme iteratiu que es basa en repetir les següents passes tants cops com termes necessitem per assolir les convergències de les cadenes de Markov implicades.

10.1.1 Descripció teòrica de l'algorisme

L'algorisme necessita partir d'una llavor inicial pels valors de les freqüències i del vector de paràmetres del model. A partir d'aquí les tres passes que es van iterant són les següents:

1. Mitjançant una cadena multivariant de Markov basada en la funció de versemblança de les freqüències haplotípiques, generem les freqüències pel nou pas.
2. Segons aquestes noves freqüències reconstruïm els haplotips de cada individu. Això ho fem simulant valors segons una distribució categòrica amb tantes categories com haplotips possibles tingui cada individu. D'aquesta manera si un individu té més d'una parella d'haplotips compatible amb el seu genotip, segons la distribució categòrica, amb força seguretat se li assignarà la parella d'haplotips més probable. Però per casos menys extrems, o fins i tot propers a la equiprobabilitat, pot ser que en diferents moments de l'algorisme se li assignin parelles diferents.
3. Un cop reconstruïda la mostra d'haplotips, passem aquesta variable al model que haguem considerat. Ara, generem un nou pas de la segona cadena multivariada creada pels coeficients del model.

L'algorisme en notació matemàtica

Comencem donant uns valors inicials

Es tracta de valors qualssevol que fan de llavor pel primer pas de la cadena per f i per β:

$$f^{(0)} = (f_1^{(0)}, f_2^{(0)}, \ldots, f_M^{(0)})$$

$$\beta^{(0)} = (\beta_1^{(0)}, \beta_2^{(0)}, \ldots, \beta_M^{(0)})$$

Generem un següent candidat per la cadena de les freqüències segons Random Walk Sigui $u = (u_1, \ldots, u_M)$ tal que $u_i \sim Unif(0, s)$ o bé $u_i \sim N(0, s)$ $i = 1, \ldots, M$. Llavors,

$$f^{(1)} = f^{(0)} + u$$

La desviació s es tria experimentalment. Testem si ens quedem aquest candidat.

Sigui ℓ_1 qualsevol de les tres versemblances descrites a (9.2.1)(**??**) o (9.10). Ara, per (7.2) tenim que si P és una prior concreta, es compleix:

$$\frac{\pi(f^{(1)})}{\pi(f^{(0)})} = \frac{\ell_1(f^{(1)})P}{\ell_1(f^{(0)})P} = \frac{\ell_1(f^{(1)})}{\ell_1(f^{(0)})}$$

Seguint el procediment descrit a la secció 8.3.2, generem un valor $v \sim Unif(0, 1)$ i comprovem si

$$v < \frac{\ell_1(f^{(1)})}{\ell_1(f^{(0)})}$$

Si passa, llavors acceptem el candidat. Si no, $f^{(1)} = f^{(0)}$.

Un cop actualitzat el valor de les freqüències, reconstruïm els haplotips per cada individu.

Reconstrucció dels haplotips

Pels genotips que no presenten incertesa, sabem amb seguretat la parella que porten. En canvi, per aquells que poden dur més d'una parella, els hi assignem una parella resultant de mostrejar d'una distribució categòrica amb probabilitats equivalents a les freqüències f. És a dir, considerem a tall d'exemple un individu que pot dur dues parelles d'haplotips:

$H_1 = (h_1, h_2)$ o bé $H_2 = (h_3, h_4)$. Coneixent $f = (f_1, \ldots, f_M)$, passa que $P(H_1) = 2f_1 * f_2$ i $P(H_2) = 2f_3 * f_4$. Aleshores per decidir quina parella imputar-li a l'individu, mostrejaríem d'una $cat(p_1, p_2)$ on $p_1 = \frac{f_1 f_2}{f_1 f_2 + f_3 f_4}$ i $p_2 = \frac{f_3 f_4}{f_1 f_2 + f_3 f_4}$.

Un cop feta la reconstrucció, ja tenim la variable $H = (H^1, \ldots, H^N)$ on H^i representa la parella haplotípica de l'individu i-èssim. Ara, traduïm H a variables indicadores, i les introduïm com a regressora al model. Aquests valors apareixeran al càlcul de la funció de versemblança del model amb que s'estigui treballant, que a partir d'ara notarem ℓ_2.

Generem un nou candidat pels coeficients del model de regressió

Ja estem en condicions de generar un nou candidat per la cadena de les β's segons la Gibbs Sampler i el mètode de l'Slice Sampler.

Per al vector:

$$\beta^{(0)} = (\beta_1^{(0)}, \beta_2^{(0)}, \ldots, \beta_{M-1}^{(0)})$$

considerem la seva distribució posterior:

$$\pi(\beta^{(0)}) \propto p \times L(\beta^{(0)}) \tag{10.1}$$

on $L(\beta^{(0)})$ és la funció de versemblança que depèn del model i p és la distribució prior que triem. Per a cada β_i, prenem la distribució condicional $\pi(\beta_i|\beta_{-i})$ tal i com diu la teoria de la Gibbs Sampler. El mostreig univariat per a aquesta distribució el fem aplicant el mètode de tipus slice exposat a 8.3.5 de manera univariada per a cada component. Així ja obtenim un nou pas de la cadena de les betes:

$$\beta^{(1)} = (\beta_1^{(1)}, \beta_2^{(1)}, \ldots, \beta_M^{(1)})$$

Tornem a començar

Tot seguit, partint del punt anterior, tornaríem a generar un nou candidat per les freqüències, les actualitzaríem, generaríem una nova mostra d'haplotips, li passaríem al model, tornaríem a generar un candidat per les β's, i així iterativament fins que la mitjana ergòdica comenci a

ser estable, i per tant, es pugui considerar que les cadenes ja tenen distribució estacionària, i que per tant, podem aplicar els estimadors ergòdics corresponents donat que ja haurem generat una mostra.

10.2 Què hem obtingut?

Amb aquest algorisme com podem veure a la figura 10.2 hem obtingut M cadenes, una per a cada freqüència haplotípica i M cadenes més, una per cada paràmetre del model.
Aquestes cadenes de Markov, per la teoria que ja hem exposat tenen com a distribució invariant la de cadascun dels paràmetres. Per tant, mitjançant la teoria ergòdica ara podem resumir les distribucions segons la mitjana ergòdica marginal calculant:

$$\bar{f}_i = \frac{1}{n}\sum_{j=1}^{n} f_i^{(j)} \tag{10.2}$$

o bé:

$$\bar{\beta}_i = \frac{1}{n}\sum_{j=1}^{n} \beta_i^{(j)} \tag{10.3}$$

així com també podem calcular d'igual manera la variància marginal de cada component:

$$\sigma^2_{\bar{f}_i} = \frac{1}{n}\sum_{j=1}^{n} (f_i^{(j)} - \bar{f}_i)^2 \tag{10.4}$$

o bé:

$$\sigma^2_{\bar{\beta}_i} = \frac{1}{n}\sum_{j=1}^{n} (\beta_i^{(j)} - \bar{\beta}_i)^2 \tag{10.5}$$

Com veurem a l'apartat de resultats, podrem graficar aquestes distribucions i extreure la informació que ens sembli pertinent. El comportament de l'algorisme ha estat validat mitjançant simulacions informàtiques que es mostren al capítol de resultats.

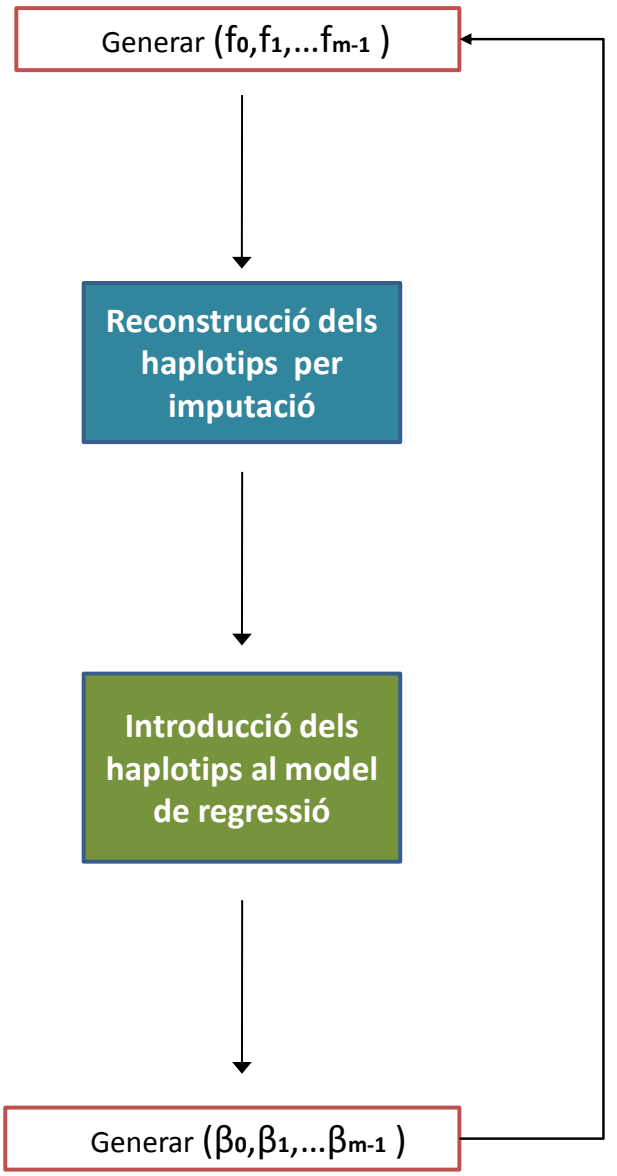


Figura 10.1. Esquema simplificat de l'algorisme iteratiu

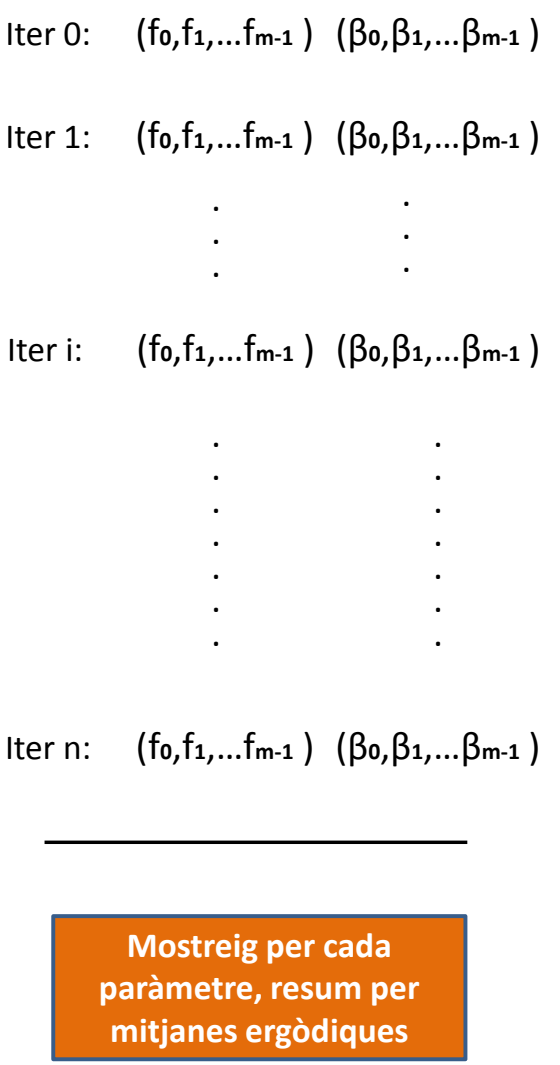


Figura 10.2. A cada iteració es genera cadascun dels paràmetres creant una cadena de Markov que es resumeix mitjançant la teoria ergòdica.

11

BayHap, el paquet Bayesià d'anàlisi d'associació amb haplotips

Per tal de fer factible la utilització del mètode que hem dissenyat, l'algorisme s'ha implementat informàticament. La implementació ve acompanyada d'una interfície per tal de facilitar als usuaris l'execució del programa. El programa s'ha desenvolupat amb llenguatge de programació C, havent de ser especialment curosos ja que la programació dels mètodes que hi intervenen són susceptibles de generar nombrosos problemes numèrics. La interfície s'ha situat en R, entorn de programació per anàlisis estadístiques i gràfiques. R es distribueix sota la llicència GNU i està disponible pels sistemes operatius Windows, Macintosh, Unix i GNU/Linux.

11.1 R i la programació de paquets

R és un dels entorns més flexibles, potents i professionals que existeixen a l'actualitat per realitzar tasques estadístiques de tot tipus, des de les més elementals fins les més avançades. Probablement, R és el llenguatge més utilitzat en investigació por la comunitat estadística, sent a més molt popular en el camp de la investigació biomèdica, la bioinformàtica i les matemàtiques financeres. En particular, està desenvolupat i mantingut per alguns dels estadístics més prestigiosos del moment. Compta, a més, amb l'avantatge de ser un projecte de software lliure gratuït i senzill pel que fa a descarrega i instal·lació. R proporciona un ampli ventall d'eines estadístiques (models lineals i no lineals, tests estadístics, anàlisi de sèries temporals, algorismes de classificació i agrupament, etc.) i la capacitat de generar

gràfics molt complerts. A tot això se suma la possibilitat de carregar diferents llibreries o paquets amb finalitats específiques de càlcul o gràfic. Existeix un repositori oficial que actualment ja supera la xifra dels 2000 paquets. Donada la gran quantitat de nous paquets, s'han organitzat per temes que permeten agrupar-los segons la seva naturalesa i funcionalitat. Per exemple, hi ha grups de paquets relacionats amb estadística Bayesiana, econometria, series temporals, etc.

Gran part de les funcions que s'executen en l'entorn R estan escrites amb el mateix R, però per algorismes computacionalment més exigents, és possible desenvolupar llibreries en C, C++ o Fortran que es carreguen dinàmicament. Els usuaris més avançats també poden manipular els objectes d'R directament des de codi desenvolupat en C. Aquest fet és el que s'ha explotat en aquesta tesi.

11.2 BayHap

BayHap és la llibreria d'R formada per una família de funcions escrites en R i per una llibreria dinàmica escrita en C que en el seu conjunt permeten a l'usuari preparar dades genètiques, executar l'algorisme que hem presentat, i resumir i graficar els resultats obtinguts. El paquet BayHap implementa l'estimació simultània de les freqüències haplotípiques per haplotips coneguts i incerts, i també computa l'associació entre aquests haplotips i fenotips basant-se en els models lineals generalitzats. Els fenotips poden ser de classe contínua, binària o de supervivència. La inferència Bayesiana i les tècniques de Markov Chain Monte Carlo són el marc teòric on s'engloben els mètodes d'estimació que s'inclouen en aquest paquet. El paquet permet incloure distribucions prior pels paràmetres dels models, a més d'oferir diferents tests de convergència i anàlisis estadístic i gràfic del mostreig resultant. Aprofitant el fet de programar en un entorn lliure, BayHap inclou algunes funcions ja existents en els paquets 'genetics' i 'Boa'.

11.2.1 Funcions del paquet

Les funcions que conformen el paquet són les següents:

- autocorr: Funció que calcula les autocorrelacions d'una seqüència de MCMC per cada paràmetre tenint en compte el conjunt d'iteracions que queden excloses de l'anàlisi (les referents al 'lag' explicat als arguments modificables.)
- bayhapFreq: Aquesta funció implementa l'estimació de les freqüències d'haplotips incerts. L'estadística Bayesiana i les tècniques de MCMC són el marc teòric on s'inclou el mètode implementat en aquesta funció. El mostreig per les freqüències d'haplotips es duu a terme mitjançant un *Random Walk* per les freqüències d'haplotips. La funció retorna l'estimació dels paràmetres amb la seva desviació estàndard i interval de confiança.
- bayhapReg: La principal funció d'aquest paquet és la funció *bayhapReg*. Donada una mostra de genotips, aquesta funció duu a terme estimacions simultànies de les freqüències d'haplotips i les estimacions dels paràmetres del model lineal generalitzat triat, duent la variable d'haplotips com a factor de risc. Trets quantitatius, binaris i de supervivència són acceptats per aquesta funció i modelats a través de regressió lineal, regressió Logística i regressió de Weibull. Els models accepten termes d'interacció entre les variables haplotípiques i covariables d'interès. Així com també es possible triar entre tres models d'herència diferents: additiu, dominant o recessiu.
- BIC: Aquesta funció calcula el *Bayesian Information Criterion* pels models estimats amb la funció bayhapReg.
- conv.test: Calcula els diagnòstics de convergència de Heidleberger i Welch convergence pels paràmetres d'una seqüència MCMC.
- correl: Calcula la matriu de correlacions pels paràmetres d'una seqüència MCMC.
- plotACF: Crea un gràfic per les autocorrelacions als lags per un paràmetre específic.
- plotDensity: Estima i fa el gràfic de la funció de densitat pels paràmetres d'interès.

- plotFreq: Aquesta funció retorna conjuntament els gràfics per les autocorrelacions, la mitjana ergòdica, les funcions de densitat i les seqüències creades per cadascuna de les freqüències haplotípiques.
- plotReg: Aquesta funció retorna conjuntament els gràfics per les autocorrelacions, la mitjana ergòdica, les funcions de densitat i les seqüències creades per cadascun dels coeficients del model de regressió considerat.
- plotRmean: Calcula i grafica la mitjana ergòdica dels paràmetres pels que es construeix la cadena de Markov.
- plotTrace: Rutina que retorna el gràfic de la seqüència que el programa genera per cada paràmetre.
- setupData: Aquesta funció comprova que el tipus i el format de les dades originals sigui apropiat per l'anàlisi.

11.2.2 Ús del paquet

La principal funció d'aquest paquet és la funció *bayhapReg*. Abans d'executar aquesta funció, en primer lloc els usuaris han d'executar la funció *setupData* i així obtenir un objecte de tipus data.frame per ser inserit en *bayhapReg*. A l'apèndix es poden observar diversos exemples. En cas que l'usuari desitgi incloure informació prèvia, abans de l'execució de *bayhapReg* cal executar la funció *bayhapFreq*, i obtenir així les etiquetes per a cada haplotip existents a la mostra de genotips.

Un cop s'hagi executat *bayhapReg* el següent pas és utilitzar el seguit de funcions que el paquet inclou per mostrar els resultats resumits numèricament i gràficament. Per avaluar la convergència del mètode i per tant, la validesa dels resultats, cal fer el diagnòstic de la cadena. Per aquest fet són útils les funcions *autocorr, conv.test, correl, plotACF, plotDensity, plotRmean* i *plotTrace*. Executar plotRmean serà útil per observar l'estabilitat de la mitjana durant l'execució, com d'encertat ha estat el burnin i el nombre total d'iteracions triat. Les autocorrelacions graficades mitjançant plotACF són útils per comprovar la seva disminu-

ció a mida que la cadena es va generant.

En cas que s'hagin provat diferents models, el paquet ofereix la mesura BIC per triar el que millor ajusti. Un punt a favor d'R i de l'ús dels paquets és la facilitat de comprensió del fun-

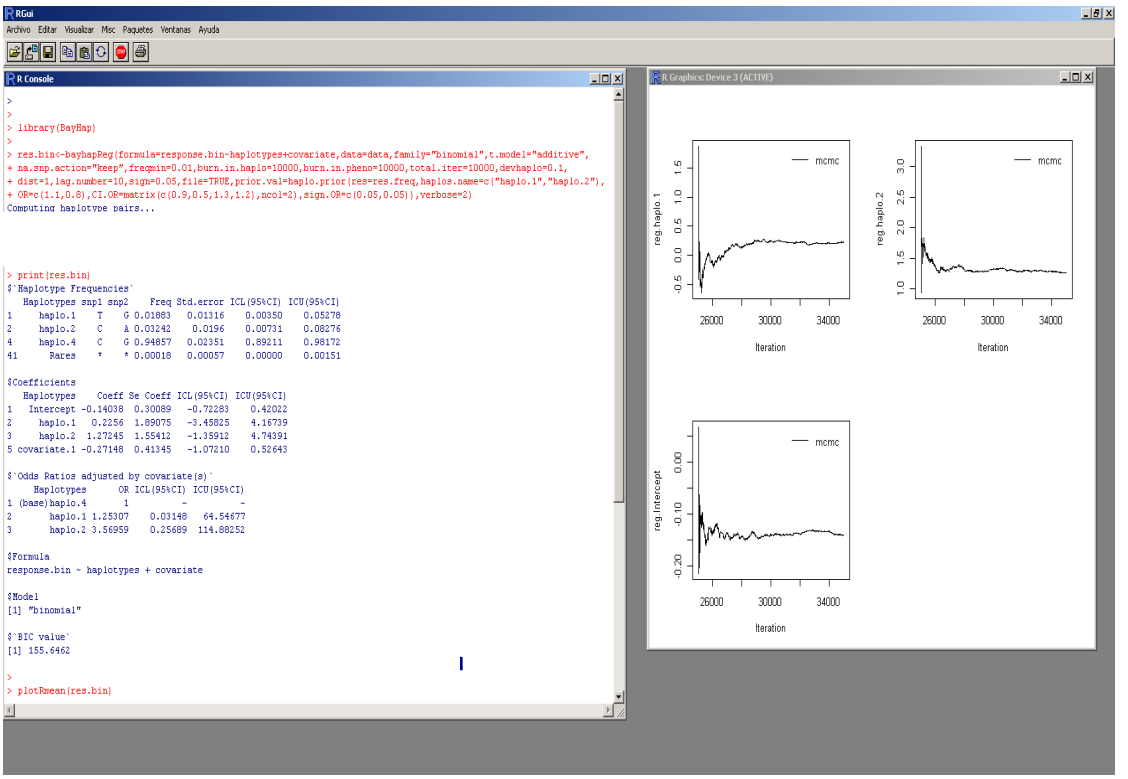


Figura 11.1. Imatge de la consola d'R amb una execució de BayHap, juntament amb alguns dels resultats numèrics i gràfics obtinguts

cionament de les diferents rutines, gràcies al sistema d'ajudes que ofereix aquest entorn. En aquest cas, BayHap també ve acompanyat de pàgines d'ajuda per a cada funció, incloent exemples de dades i d'execució per facilitar a l'usuari la utilització del paquet.

Figura 11.2. Imatge d'una pàgina del help del programa BayHap.

11.2.3 Arguments modificables

Un dels punts forts d'aquest paquet és que en funció de les dades, del coneixement previ, o bé després de la no convergència d'una execució anterior, l'usuari pot adaptar el valor d'un quants paràmetres amb l'objectiu d'optimitzar l'execució i la precisió dels resultats.

- burnin: Aquest és un argument clau per a que el resum dels paràmetres sigui precís. El burnin és la quantitat de primeres iteracions de qualsevol de les cadenes que encara fan oscil·lar la mitjana ergòdica, prèvies a la estabilització d'aquesta. Aquestes són les

iteracions que seran descartades a l'hora de fer els càlculs sobre la cadena. El valor per defecte que porta el programa s'hauria d'adaptar si s'observa que la cadena no ha convergit correctament. BayHap ofereix dos possibles burnins, un per les cadenes de les freqüències i un altre per les cadenes dels paràmetres del model.

- devhaplo: Aquest argument té a veure amb el procés de mostreig de Random Walk utilitzat per la generació de les cadenes haplotípiques. Aquest valor defineix com de gran és el salt que es realitza en aquest mostreig entre diferents passes. Influeix directament en la convergència de la cadena i hauria de ser incrementat en cas d'observar-se convergència a màxims locals.
- Model d'herència: Hi ha tres possibles models d'herència, l'additiu, el dominant o el recessiu. L'additiu dóna el mateix pes als dos haplotips que porta cadascun dels individus. Pel que fa al model dominant, l'aparició un sol cop d'un haplotip té el mateix pes que si apareix dos cops. I segons el model recessiu, si els dos haplotips que duu l'individu són diferents, no tindran cap aportació al model. Només tindran el mateix efecte que a un model dominant en cas que la parella estigui formada per dos haplotips iguals. La tria del model d'herència es pot dur a terme segons BIC.
- freqmin: Punt de tall tal que el conjunt d'haplotips amb freqüència per sota d'aquest nivell entraran al model agrupats en una categoria anomenada 'rares'. Per sobre d'aquest valor tots els haplotips s'introduiran independentment al model. El valor per defecte és de 0.01.
- Lag: Per evitar que zones de la distribució no es visitin i d'altres es visitin massa, es pot definir un nombre d'iteracions anomenat *Lag* o *Thinning interval*. Aquest nombre determinarà cada quant guardem com a membre de la cadena el valor generat. És a dir, si posem un thinning interval de 10, només guardarem un de cada 10 termes que generi la cadena. Aquest valor també es recomana trobar-lo empíricament.

- Distribució de Random Walk: L'usuari pot triar la distribució que determinarà el següent pas del mostreig segons Random Walk. Les dues opcions són la distribució Uniforme i la distribució Normal.

Part V

RESULTATS

12

Aplicació de BayHap sobre escenaris simulats. Comparació amb d'altres programes.

Als darrers anys l'ús de simulacions computacionals en l'àmbit de la recerca ha experimentat un creixement notable. Dominis com l'astrofísica, l'enginyeria, la química, la biologia i els estudis ambientals s'estan beneficiant d'aquesta important capacitat de resoldre una gran varietat de problemes científics. Malgrat tot, simular proporciona una enorme quantitat de dades que s'han de saber tractar, analitzar i interpretar.

En aquest treball s'han utilitzat tècniques de simulació per tal de determinar el comportament del programa BayHap i de validar els resultats que retorna. Exactament, el que s'ha fet ha estat generar un gran nombre de bases de dades, cadascuna d'elles amb mateixes característiques fixades i conegudes pels que les generàvem, i a posteriori s'ha comprovat la probabilitat amb que BayHap i d'altres programes estimen amb correcció aquests valors coneguts amb antelació. Cadascuna de les bases de dades que es genera representa una mostra d'una ''població'' fictícia amb unes característiques teòriques fixades. D'aquesta manera ha estat possible avaluar els resultats retornats per BayHap segons diverses característiques de la mostra, i comparar-ho amb els resultats obtinguts sobre les mateixes dades amb d'altres programes. El programa BayHap s'ha executat amb una distribució prior no informativa.

Pel que fa a la tria de les característiques amb que s'han generat les dades, cal tenir present que executar cada programa sobre un conjunt tan nombrós de bases de dades té un cost de temps molt elevat. Així doncs, s'han simulat conjunts de dades variant algunes de les

característiques que la literatura destaca com a més rellevants. Es tracta de propietats que s'associen a l'aplicabilitat del programa i a la precisió dels resultats.

Més concretament, per validar BayHap s'han generat 25.000 conjunts de dades per les quals varia el nombre d'SNPs, el nombre d'individus, el tipus de disseny considerat i per tant varia el tipus de fenotip analitzat i de model estadístic utilitzat, la incertesa de les dades, la freqüència haplotípica i la mida dels efectes associats. Això ha suposat un total de 5 escenaris que combinen aquesta varietat de característiques. Per cadascun dels escenaris s'han generat 5000 bases de dades a les quals s'ha aplicat el programa BayHap. En alguns casos també s'ha executat el paquet d'R Haplo.Stats, un programa estàndard, per tal de poder comparar resultats i efectivitat del programa BayHap. El programa es troba explicat a (3.6). La idea original de simular també amb el programa THESIAS pel cas de dades referents a un estudi de supervivència no s'ha pogut dur a terme degut als entrebancs informàtics que suposa simular amb el programa, tant en la versió de línia de comandes com amb la versió en java.

12.1 Escenaris en que s'han simulat les bases de dades

Les 25.000 bases de dades que s'han generat per testar el programa BayHap contenen informació genotípica referent a un seguit d'SNPs bial·lèlics per un conjunt d'individus. Les dades no han comptat amb valors missing. Els conjunts de genotips s'han generat sota equilibri de Hardy-Weinberg. Les variables fenotípiques que s'han generat estan associades amb una certa magnitud a alguns haplotips. Així doncs, per exemple pel que fa a les dades referents a un estudi cas-control, s'han generat fenotips binaris de tal manera que certs haplotips de la mostra tenen associats uns valors d'OR concrets coneguts.

Les característiques en què s'ha basat la simulació de les dades són les següents:

1. Mida mostral: S'han considerat diverses mides mostrals. Dos de reduïts de 200 i 300 individus, i un altra mida de 1000 individus.

2. Incertesa: S'han considerat incerteses altes, ja que és en aquests casos en que l'estimació d'haplotips resulta més interessant. En cas de no incertesa les solucions no tenen especial interès i els resultats entre programes són similars. S'ha considerat bases de dades amb una incertesa aproximada del 40%. Aquest tant per cent es refereix al percentatge d'individus a la mostra que presenten un genotip amb dos o més SNPs heterozigots.
3. Nombre d'SNPs: S'han fet simulacions amb un nombre reduït d'SNPs i també amb una quantitat moderadament més elevada, però tenint en compte que fos factible el temps d'execució per poder realitzar el nombre de simulacions estipulat. Per això s'han generat bases de dades amb 3 SNPs, 4 SNPs i 8 SNPs.
4. Freqüència haplotípica: Als diversos escenaris s'ha generat dades genotípiques que continguessin un haplotip majoritari i un altre amb freqüència <0.1 per avaluar l'estimació d'aquestes freqüències petites i també dels efectes atribuïbles a aquestes freqüències. També s'ha considerat l'aparició a la mostra haplotípica d'haplotips amb freqüències similars per estimar la precisió amb que els diversos programes són capaços d'estimar aquestes freqüències.

Cada base de dades ha estat generada mitjançant funcions programades amb llenguatge R.

12.1.1 Descripció numèrica dels escenaris

A continuació es resumeixen les característiques numèriques detallades dels cinc escenaris que s'han generat, incloent les freqüències de cada haplotip i els valors de les mesures d'associació que s'han simulat:

Escenari	N	Incertesa (%)	Nombre d'SNPs	Fenotip
1	200	38.5	3	Binari
2	1000	40	8	Binari
3	1000	22	8	Binari
4	300	35	4	Continu
5	600	35	3	Supervivència

Escenari número 1

- Mida de la mostra: 200 individus
- Nombre d'SNPs: 3 SNPs
- Incertesa: 38,5%
- Nombre de base de dades generat: 5000
- Total d'haplotips possibles a la mostra: 8 haplotips, n'apareixen 4
- Disseny: Cas-control
- Valors de les freqüències haplotípiques i ORs referents a les 5000 bases de dades:

Haplotip	Mostra General	Mostra de Casos	Mostra de Controls	OR
AAA	0.6	0.6	0.6	1.0
AAB	0.25	0.25	0.25	1.0
ABA	0.1	0.11	0.087	1.3
ABB	0.05	0.06	0.03	2.0

Escenari número 2

- Mida de la mostra: 1000 individus
- Nombre d'SNPs: 8 SNPs
- Incertesa: 40%
- Nombre de base de dades generat: 5000
- Total d'haplotips possibles a la mostra: 256 haplotips, n'apareixen 6
- Disseny: Cas-control
- Valors de les freqüències haplotípiques i OR's referents a les 5000 bases de dades:

Haplotip	Mostra General	Mostra de Casos	Mostra de Controls	OR
AAAAAAAA	0.45	0.45	0.45	1.0
AAAAAAAB	0.2	0.2	0.2	1.0
AAAAAABB	0.11	0.11	0.11	1.0
BBBBBBBB	0.12	0.144	0.096	1.5
AAAAABAA	0.07	0.094	0.047	2.0
AAAAABAB	0.05	0.075	0.025	3.0

Escenari número 3

- Mida de la mostra: 1000 individus
- Nombre d'SNPs: 8 SNPs, n'apareixen 6

- Incertesa: 22%
- Nombre de base de dades generat: 5000
- Total d'haplotips possibles a la mostra: 256
- Disseny: Cas-control
- Valors de les freqüències haplotípiques i OR's referents a les 5000 bases de dades:

Haplotip	**Mostra General**	**Mostra de Casos**	**Mostra de Controls**	**OR**
AAAAAAAA	0.6	0.6	0.6	1.0
AAAAAABA	0.12	0.12	0.12	1.0
AAAAABAA	0.1	0.12	0.08	1.5
AAAAABBB	0.07	0.046	0.094	2.0
AAAAAAAB	0.06	0.06	0.06	1.0
AAAAABAB	0.05	0.075	0.025	3.0

Escenari número 4

- Mida de la mostra: 300 individus
- Nombre d'SNPs: 4 SNPs
- Incertesa: 35%
- Nombre de base de dades generat: 5000
- Total d'haplotips possibles a la mostra: 16, n'apareixen 3
- Disseny: Resposta quantitativa
- Valors de les freqüències haplotípiques i ORs referents a les 5000 bases de dades:

Haplotip	**Mostra General**	**Dif de mitjanes**
BABA	0.57	0
AAAA	0.33	1.0
ABAB	0.10	0

Escenari número 5

- Mida de la mostra: 600 individus
- Nombre d'SNPs: 3 SNPs
- Incertesa: 35%
- Nombre de base de dades generat: 5000
- Total d'haplotips possibles a la mostra: 8, n'apareixen 3
- Disseny: Anàlisi de Supervivència
- Valors de les freqüències haplotípiques i HR's referents a les 5000 bases de dades:

Haplotip	Mostra General	Hazard Ratio
AAA	0.75	1.0
ABB	0.166	2.3
ABA	0.083	3.3

12.2 Resultats de les simulacions

A continuació s'exposen els resultats de les 55.000 execucions que s'han realitzat en els diferents escenaris i amb els diversos programes ja comentats. El model d'herència triat en tots els casos ha estat l'additiu. El model estadístic ha anat variant segons el tipus de fenotip.

Per a cada escenari i conjunt de 5000 execucions d'un mateix programa s'han calculat:

- Estimadors: Mitjana i desviació típica obtingudes en el conjunt de simulacions pels diferents paràmetres.
- Biaix: es considera la mitjana de les diferències obtingudes entre l'estimador mitjana retornat per cadascuna de les 5000 simulacions i el valor real del paràmetre. També es calcula la desviació d'aquest biaix al llarg de les diferents execucions.
- Cobertura: S'ha computat el tant per cent de cops que l'interval retornat pel programa (per cada base de dades) inclou el valor real del paràmetre. Es vol testar si l'interval de confiança inclou el 95% de vegades el valor real.

Resultats BayHap per simulacions en l'escenari número 1 (200 individus i 3 SNPs)

A la taula 12.1 podem observar com el biaix de les freqüències és nul. Com veurem, aquesta serà la tònica general de tots els resultats de les execucions pel que fa al biaix de les freqüències que retorna BayHap. Les desviacions típiques són petites la qual cosa implica que les estimacions es desvien poc del valor real de les freqüències. Pel que fa a les estimacions dels OR, la taula 12.2 mostra com el biaix més gran el trobem per l'OR associat a

Haplotip	Mitjana Freq	Sd Freq	Biaix Freq	Sd Biaix	Cober Freq
AAA	0.6	0.02	0.0	0.02	94.52
AAB	0.25	0.02	0.0	0.02	95.10
ABA	0.1	0.02	0.0	0.02	94.78
ABB	0.05	0.01	0.0	0.01	94.44

Taula 12.1. Taula de resultats per freqüències a l'escenari 1 segons BayHap

Haplotip	Mitjana OR	Sd OR	Biaix OR	Sd Biaix	Cober OR
AAA	-	-	-	-	-
AAB	1.00	0.24	0.027	0.25	95.50
ABA	1.31	0.36	0.103	0.54	94.35
ABB	2.16	0.56	0.56	1.79	95.25

Taula 12.2. Taula de resultats per OR a l'escenari 1 segons BayHap

l'haplotip menys freqüent. Tot i així, la cobertura és bona tant per les freqüències com pels OR, mantenint-se al voltant del 95%.

Resultats Haplo.Stats per simulacions en l'escenari número 1 (200 individus i 3 SNPs)

Haplotip	Mitjana Freq	Sd Freq	Biaix Freq	Sd Biaix	Cober Freq
AAA	0.6	0.02	0.0	0.02	-
AAB	0.25	0.02	0.0	0.02	82.48
ABA	0.1	0.02	0.0	0.02	99.40
ABB	0.05	0.01	0.0	0.01	92.86

Taula 12.3. Taula de resultats per freqüències a l'escenari 1 segons Haplo.Stats

Les simulacions amb les mateixes dades de l'escenari 1 resultat d'aplicar el programa Haplo.Stats. Com es pot observar a la taula 12.3 els estimadors per les freqüències són no esbiaixats. Pel que fa als ORs, el referent a l'haplotip menys freqüent és el més esbiaixat i amb biaix més dispers. En aquest cas, a diferència dels resultats de BayHap, hi ha algunes

Haplotip	Mitjana OR	Sd OR	Biaix OR	Sd Biaix	Cober OR
AAA	-	-	-	-	-
AAB	0.91	0.2	0.027	0.26	95.06
ABA	1.35	0.4	0.09	0.53	95.28
ABB	2.22	0.9	0.46	1.69	96.48

Taula 12.4. Taula de resultats per OR a l'escenari 1 segons Haplo.Stats

cobertures molt baixes.

En aquest escenari el programa Haplo.Stats no ha convergit en 5 execucions, el que representa un 0.1% dels casos.

Resultats BayHap per simulacions en l'escenari número 2 (1000 individus i 8 SNPs)

Haplotip	Mitjana Freq	Sd Freq	Biaix Freq	Sd Biaix	Cober Freq
AAAAAAAA	0.45	0.011	0.0	0.011	95.00
AAAAAAAB	0.2	0.009	0.0	0.009	94.76
BBBBBBBB	0.12	0.007	0.0	0.007	94.18
AAAAAABB	0.11	0.007	0.0	0.007	94.92
AAAAABAA	0.07	0.006	0.0	0.006	95.14
AAAAABAB	0.05	0.005	0.0	0.005	95.02

Taula 12.5. Taula de resultats per freqüències a l'escenari 2 segons BayHap

En aquest escenari, amb un major nombre d'SNPs, podem observar segons les taules 12.5 i 12.6 que els resultats de BayHap tenen característiques similars als del primer escenari.

Per aquest cas, donada la mida mostral, el biaix s'ha reduït. Les cobertures són correctes tant per les freqüències com pels OR.

Haplotip	Mitjana OR	Sd OR	Biaix OR	Sd Biaix	Cober OR
AAAAAAAA	-	-	-	-	-
AAAAAAAB	1.02	0.11	0.03	0.12	97.58
BBBBBBBB	1.51	0.15	0.13	0.04	95.07
AAAAAABB	1.0	0.15	0.01	0.15	95.17
AAAAABAA	2.12	0.17	0.16	0.39	96.84
AAAAABAB	2.86	0.21	0.07	0.65	98.23

Taula 12.6. Taula de resultats per OR a l'escenari 2 segons BayHap

Resultats Haplo.Stats per simulacions en l'escenari número 2 (1000 individus i 8 SNPs)

En relació a la taula 12.7 destacar que per freqüències baixes, els biaixos i les desviacions típiques són superiors que les retornades pel programa BayHap. Les cobertures per les freqüències d'un dels haplotips es troba molt per sota del 95%.

Haplotip	Mitjana Freq	Sd Freq	Biaix Freq	Sd Biaix	Cober Freq
AAAAAAAA	0.45	0.011	0.002	0.011	94.32
AAAAAAAB	0.2	0.009	0.0	0.009	95.08
BBBBBBBB	0.12	0.006	0.0	0.006	96.66
AAAAAABB	0.11	0.006	0.0	0.006	95.06
AAAAABAA	0.07	0.005	0.001	0.005	91.46
AAAAABAB	0.05	0.005	0.001	0.005	99.26

Taula 12.7. Taula de resultats per freqüències a l'escenari 2 segons Haplo.Stats

Haplotip	Mitjana OR	Sd OR	Biaix OR	Sd Biaix	Cober OR
AAAAAAAA	-	-	-	-	-
AAAAAAAB	1.0	0.13	0.01	0.12	95.42
BBBBBBBB	1.49	0.15	0.013	0.16	94.78
AAAAAABB	1.0	0.22	0.06	0.46	95.33
AAAAABAA	2.01	0.28	0.19	0.96	94.86
AAAAABAB	3.00	0.15	0.02	0.23	95.62

Taula 12.8. Taula de resultats per OR a l'escenari 2 segons Haplo.Stats

Resultats BayHap per simulacions en l'escenari número 3 (1000 individus i 8 SNPs, menor incertes

Haplotip	Mitjana Freq	Sd Freq	Biaix Freq	Sd Biaix	Cober Freq
AAAAAAAA	0.6	0.01	0.0	0.01	95.14
AAAAAABB	0.12	0.008	0.0	0.008	95.01
AAAAABAA	0.1	0.008	0.0	0.008	94.76
ABABAAAA	0.07	0.006	0.0	0.006	96.4
AABAAAAB	0.06	0.006	0.0	0.006	94.90
BAAAABAB	0.05	0.005	0.0	0.005	94.64

Taula 12.9. Taula de resultats per freqüències a l'escenari 3 segons BayHap

Haplotip	Mitjana OR	Sd OR	Biaix OR	Sd Biaix	Cober OR
AAAAAAAA	-	-	-	-	-
AAAAAABB	1.0	0.1	0.0	0.14	97.06
AAAAABAA	1.6	0.1	0.1	0.22	96.72
ABABAAAA	2.0	0.2	0.0	0.36	98.20
AABAAAAB	1.1	0.2	0.1	0.20	96.48
BAAAABAB	2.8	0.2	0.1	0.65	96.66

Taula 12.10. Taula de resultats per OR a l'escenari 3 segons BayHap

Pel que fa a l'escenari 3 a les taules 12.9 i 12.10 observem biaixos petits tant per OR's com per freqüències. La desviació típica més alta correspon als haplotips menys freqüents, però la cobertura segueix sent bona per aquests haplotips.

Resultats BayHap per simulacions en l'escenari número 4 (300 individus i 4 SNPs)

A les taules 12.11 i 12.12 s'inclouen els resultats per les dades de l'escenari 4 resolt segons el programa BayHap. En aquest cas per l'haplotip més freqüent la cobertura apareix lleugerament per sota del desitjat.

Haplotip	Mitjana Freq	Sd Freq	Biaix Freq	Sd Biaix	Cober Freq
BABA	0.57	0.01	0.0	0.01	93.43
AAAA	0.33	0.0	0.0	0.0	99.97
ABAB	0.10	0.01	0.0	0.01	99.89

Taula 12.11. Taula de resultats per freqüències a l'escenari 4 segons BayHap

Haplotip	Mitjana Coef	Sd Coef	Biaix Coef	Sd Biaix	Cober Coef
BABA	-	-	-	-	-
AAAA	0.5	0.007	0.0	0.007	95.7
ABAB	0.9	0.006	0.13	0.005	96.88

Taula 12.12. Taula de resultats pels coeficients de la regressió lineal a l'escenari 4 segons BayHap

Resultats Haplo.Stats per simulacions en l'escenari número 4 (300 individus i 4 SNPs)

En aplicar el programa Haplo.Stats (taules 12.13 i 12.14) també s'observa una cobertura per sota del desitjable referent al segon haplotip més freqüent de la mostra.

Haplotip	Mitjana Freq	Sd Freq	Biaix Freq	Sd Biaix	Cober Freq
BABA	0.57	0.01	0.0	0.011	91.58
AAAA	0.33	0.0	0.0	0.0	96.2
ABAB	0.10	0.01	0.0	0.011	99.98

Taula 12.13. Taula de resultats per freqüències a l'escenari 4 segons Haplo.Stats

Haplotip	Mitjana Coef	Sd Coef	Biaix Coef	Sd Biaix	Cober Coef
BABA	-	-	-	-	-
AAAA	0.5	0.008	0.0	0.008	95.08
ABAB	0.9	0.006	0.0	0.006	94.50

Taula 12.14. Taula de resultats pels coeficients de la regressió lineal a l'escenari 4 segons Haplo.Stats

Resultats BayHap per simulacions en l'escenari número 5 (600 individus i 3 SNPs)

A les taules 12.15 i 12.16 es mostren els resultats referents a les simulacions dutes a terme sobre dades de supervivència. Com es pot observar el biaix tant pel que fa a freqüències com a coeficients són força petits. Les cobertures són en general bones. Només l'HR referent a l'haplotip menys freqüent queda lleugerament per sota de l'esperat.

Haplotip	Mitjana Freq	Sd Freq	Biaix Freq	Sd Biaix	Cober Freq
AAA	0.75	0.0	0.0	0.0	99
ABB	0.17	0.0	0.0	0.0	98.4
ABA	0.08	0.0	0.0	0.0	98.2

Taula 12.15. Taula de resultats per freqüències a l'escenari 5 segons BayHap

Haplotip	Mitjana HR	Sd HR	Biaix HR	Sd Biaix	Cober HR
AAA	-	-	-	-	-
ABB	2.28	0.26	0.02	0.26	95.1
ABA	3.28	0.53	0.02	0.53	93.5

Taula 12.16. Taula de resultats per l'escenari 5 segons BayHap

13

Algorisme EM vs BayHap en l'anàlisi del gen DRD2

En aquest capítol aplicarem el programa BayHap a dues bases de dades reals. Els polimorfismes que analitzarem pertanyen al gen DRD2, el gen dels receptors de la dopamina. Diversos polimorfismes d'aquest gen s'han associat a l'alcoholisme, a l'abús de substàncies i a d'altres malalties de tipus psiquiàtric. Alguns estudis també han suggerit que aquest gen podria estar modulant el risc de patir càncer de còlon.

Les dades amb que treballarem provenen de dos estudis cas-control duts a terme en dues mostres independents d'individus. Ambdós estudis han analitzat diversos polimorfismes del gen DRD2: un d'ells en relació al risc de patir esquizofrènia i l'altre respecte el de patir càncer colorectal. Pel que fa al primer dels estudis, les dades han estat analitzades dins d'una tesi doctoral [177]. L'anàlisi d'haplotips que s'ha realitzat es basa en la imputació d'haplotips i no ha generat cap resultat significatiu. Veurem com l'estimació simultània de BayHap ofereix d'altres resultats i els compararem amb els que reporta l'estimació també simultània del paquet Haplo.Stats (3.6), utilitzant l'algorisme EM. Pel que fa a les dades referents a l'estudi de CCR, s'han publicat resultats [178] en relació a l'estudi de cas-control, però no pel que fa a l'anàlisi de pronòstic, tot i que per la mostra de pacients de CCR es tenen recollides dades sobre l'evolució en el temps dels pacients. Afegirem els resultats de l'anàlisi de supervivència que proporciona BayHap i els compararem amb els reportats pel software THESIAS (3.6). També compararem els resultats de BayHap amb l'altre programa utilitzat més habitualment al context Bayesià, el PHASE (3.3.2). Les execucions de BayHap

es realitzaran sense informació a priori, donat que no es té cap creença prèvia sobre la distribució dels paràmetres a estudi.

13.1 Component genètic en la etiologia de l'Esquizofrènia i del Càncer Colorectal esporàdic

Diversos estudis realitzats en famílies amb diferent grau de parentiu, incloent bessons, semblen indicar que tant l'esquizofrènia com el trastorn bipolar estarien fortament influenciats per factors genètics. Alguns autors inclús han suggerit que aquestes malalties podrien compartir gens que conferirien susceptibilitat a patir-les. Estudis farmacogenòmics de lligament així com diversos estudis d'associació de gens candidats han identificat diverses regions cromosòmiques que podrien tenir certa implicació tant en l'esquizofrènia ([179],[180],[181],[182],[183]) com en el trastorn bipolar [184].

Pel que fa al càncer colorectal esporàdic, es tracta d'una malaltia associada a múltiples factors. Es considera que múltiples exposicions interaccionen de manera complexa amb la genètica particular de cada individu, modulant el risc de patir la malaltia. S'han dut a terme diversos estudis de cas-control, focalitzant-se en gens que intervenen en el metabolisme dels agents carcinògens dietètics ([185],[186]). Malgrat tot, se sap poc sobre els factors endògens que poden modificar la fisiologia del còlon, duent a un augment de risc de càncer.

13.1.1 Paper del gen DRD2

El gen que codifica per al receptor D2 humà (DRD2) va ser clonat per primera vegada per Grandy el 1989. Es localitza al braç llarg del cromosoma 11 (11q22-23) i consisteix en vuit exons separats per set introns. S'han descrit dues isoformes del gen, D2 long i D2 short, segons la presència o no de 29 aminoàcids en el tercer bucle citoplasmàtic del receptor [187]. Des del clonatge del gen DRD2 s'han descrit diversos polimorfismes.

La Figura 13.1 mostra alguns d'aquests polimorfismes. Entre els SNPs descrits a la bibli-

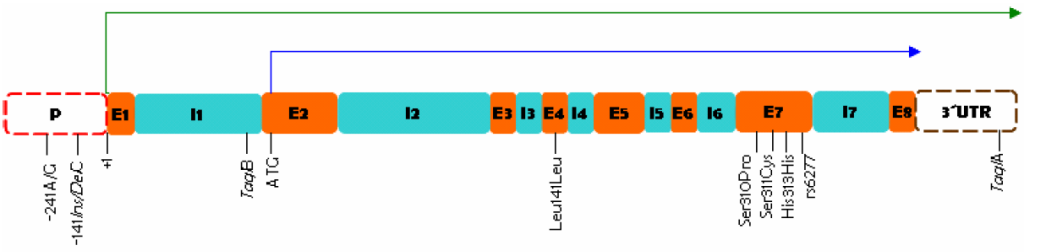


Figura 13.1. Polimorfismes del gen DRD2

ografia es troben -241 A / G i -141 Ins / Del C (a la regió promotora), TaqIB (a l'intró 1), Ser311Cys (a l'exó 7), TaqI (a la regió 3'UTR), entre d'altres. D'aquesta manera s'han realitzat estudis d'associació dels polimorfismes TaqI i -141 Ins / Del C [188], i TaqIB [189] amb l'alcoholisme; Ser311Cys [190], -141 Ins / Del C [191] entre d'altres, [192], [193],[194], amb l'esquizofrènia. Ara bé, la major part dels estudis que han tractat de trobar relació entre aquests polimorfismes i l'esquizofrènia o el trastorn bipolar han estat negatius. Per als polimorfismes DRD2 Taqi i TaqIB només hi ha un estudi realitzat en població esquizofrènica francesa que troba associació positiva amb els al·lels A2 i B2, relacionant-los amb l'excés de transmissió dopaminèrgica . Per al polimorfisme DRD2-141C, en canvi, hi ha més treballs on es suggereix que l'al·lel Del conferiria protecció davant l'esquizofrènia, tant en població japonesa com en població caucàsica, tot i un treball de meta-anàlisi en població britànica no aconsegueix replicar aquests resultats. Els estudis que han intentat relacionar polimorfismes del gen DRD2 amb el trastorn bipolar han resultat negatius . Altres estudis han suggerit una possible relació entre polimorfismes del gen DRD2 trastorns com l'obesitat, migranya o trastorns de la personalitat [195].

Pel que fa a l'associació entre el gen DRD2 i el risc de CCR, fins fa pocs anys s'havia prestat poca atenció a la dopamina i als receptors de dopamina , tot i saber-se que la dopamina pot regular el creixement de cèl lules del tracte gastrointestinal [196] i exercir un efecte

protector per l'estómac i l'intestí contra agents carcinògens com demostraven els models animals [197]. També s'ha demostrat que el teixit maligne de còlon humà té una disminució del contingut de dopamina en comparació amb el teixit normal. S'ha suggerit que aquesta disminució podria estar vinculada a una disminució en l'expressió de receptors de dopamina, com les del tipus D2 [198]. Per tant, en cercar nous mecanismes en l'etiologia del Càncer Colorectal, s'ha investigat si el risc de desenvolupar aquesta malaltia és modulada per variacions genètiques en el gen receptor de dopamina DRD2. En particular, hi ha diversos estudis que assenyalen que el gen D2 del receptor de la dopamina té polimorfismes que afecten la funció de la proteïna o la seva expressió ([199],[200],[201]), i apart de ser associats com ja hem dit amb una àmplia gamma de trastorns neurològics, psiquiàtrics o condicions de comportament (incloent la malaltia de Parkinson, l'esquizofrènia, conducta esquizoide i l'addicció al tabaquisme i al alcohol [195] algunes d'aquestes variacions també apareixen consistentment associades al risc de patir CCR [178].

13.2 Anàlisi d'associació en dos estudis

En aquest treball comptem amb les dades de dos estudis de cas-control que analitzen respectivament l'associació entre diversos polimorfismes del gen DRD2 i el risc de patir esquizofrènia i càncer colorectal. L'anàlisi es basa en 8 i 7 SNPs respectivament d'aquest gen, cinc d'ells comuns en tots dos estudis. En tots dos casos es realitzarà una anàlisi d'associació complerta que incloure l'anàlisi individual de cadascun dels SNPs i l'anàlisi d'haplotips. Aquest darrer anàlisi es farà amb el programa BayHap, amb l'algorisme PHASE, amb el programa THESIAS i amb l'algorisme EM implementat al paquet Haplo.Stats.

13.3 Estudi cas-control en pacients amb esquizofrènia

La mostra total per aquest estudi compta amb 422 individus, 164 dels quals són controls i la resta són malalts d'esquizofrènia. Tots els pacients compleixen el criteri DSM IV-R per l'esquizofrènia i tota la informació recollida prové d'ells mateixos, de la seva família, dels que en tenen cura d'ells i del metge encarregat de cada cas a la Unitat de Psiquiatria de l'Hospital Clínic de València. Per a cada individu participant a l'estudi s'ha recollit dades sociodemogràfiques com l'edat, el gènere, l'estat civil i el nivell d'estudis. També es tenen dades clíniques com els antecedents psiquiàtrics, el tractament que prenen, l'edat d'inici de la malaltia, l'estat clínic general i la valoració de la presència d'al·lucinacions. Aquest estudi ha estat aprovat pel Comitè d'Ètica local i tots els pacients han donat el consentiment informat per escrit.

Com a controls es trien individus tals que les característiques ètniques i demogràfiques s'assemblin el més possible a la dels pacients per evitar estratificació en la mostra. Prèviament a l'extracció de sang, se'ls va demanar emplenar un mini-qüestionari per tal de descartar presència d'antecedents psiquiàtrics i alteracions perceptives. Es valoraren a més d'altres factors de risc, com el consum d'estupefaents. Les dades recollides han estat valorades exclusivament pels psiquiatres a càrrec de la investigació.

13.3.1 Polimorfismes del gen DRD2 analitzats en aquest estudi

Els polimorfismes analitzats en la mostra són: -241 A / G, -141 Ins / Del C, TaqIB, rs1800499, Ser311Cys, His313His, rs6277, Pro310Ser i TaqIA. Tots ells provinents de la bibliografia excepte l'SNP rs1800499. El polimorfisme Pro310Ser va ser monomòrfic a la mostra analitzada, per la qual cosa no es van realitzar les anàlisis estadístiques d'associació. La correspondència entre aquesta nomenclatura i la de la dbSNP és:-241A / G (rs1799978), -141 Ins / Delco (rs1799732), TaqIB (rs1079597), Leu141Leu (rs1800499), Pro310Ser (rs1800496), Ser311Cys (rs1801028), His313His (rs6275) i TaqIA (rs1800497).

13.3.2 Resultats de l'anàlisi d'associació

A la taula 13.1 es mostren les freqüències al·lèliques i genotípiques per cadascun dels SNPs, a la mostra general, a la de controls i a la dels casos. I a la taula 13.2 es mostren els p valors del test d'independència que prova l'equilibri de Hardy-Weinberg. Observem que tots els polimorfismes estudiats es troben en equilibri de Hardy-Weinberg, tant en controls com en casos, excepte pel -141 Ins/Del i TaqIA pels que s'ha trobat desviacions significatives als casos, i també a la mostra total pel -141 Ins/Del. En controls, tots els SNPs compleixen HWE.

L'associació de cadascun dels polimorfismes es mostra a la taula 13.3. Els models s'han ajustat per sexe i per edat. Els SNPs que s'associen a una variació de risc de patir esquizofrènia són TaqIB, His313His i rs6277. Per TaqIB, l'heterozigot és protector. Per His313His, segons el model additiu, portar per cada còpia de l'al·lel variant augmenta el risc. Pel polimorfisme rs6277, els homozigots variants dupliquen el risc respecte de la resta. Si s'aplica la correcció de Bonferroni, cap d'ells es troba per sota de 0,00625, pel que es perden aquestes significacions.

Passem a l'anàlisi de múltiples SNPs. En primer lloc, descriurem la presència de blocs de LD. La regió analitzada pel gen DRD2 s'estén al llarg de 75.523pb, i comprèn pràcticament la totalitat del gen. Als controls no hi ha blocs de LD i pel que fa a la mostra dels pacients, es defineix un bloc discontinu que inclou els polimorfismes TaqIB, His313His y rs6277, interromput per una zona amb LD baix que inclou els loci rs1800499 i Ser311Cys. Però a les dues mostres els valors de r^2 són baixos (menors de 0.7). Destaquen valors de r^2 propers a 0.5 entre els polimorfismes TaqIB-TaqIA i His313His-rs6277B tant pel que fa a la mostra de pacients com a la dels controls.

Taula 13.1. Freqüències al·lèliques i genotípiques pels polimorfismes del gen DRD2 per l'estudi d'esquizofrènia.

SNP	AL·LEL	TOTAL	CTROLS	CASOS	GENO	TOTAL	CTROLS	CASOS
-241	A	0,95	0,94	0,95	A/A	0,89	0,88	0,9
	G	0,05	0,06	0,05	A/G	0,11	0,12	0,1
					G/G	0	0	0
-141 Ins/Del	I	0,93	0,95	0,91	I/I	0,87	0,89	0,85
	D	0,07	0,5	0,09	I/D	0,12	0,11	0,12
					D/D	0,02	0	0,03
TaqIB	G	0,87	0,86	0,88	A/A	0,03	0,01	0,04
	A	0,13	0,14	0,12	A/G	0,21	0,26	0,17
					G/G	0,77	0,73	0,79
rs1800499	G	0,97	0,96	0,98	A/A	0	0	0
	A	0,03	0,04	0,02	A/G	0,06	0,09	0,05
					G/G	0,94	0,91	0,95
Ser311Cys	C	0,97	0,98	0,97	C/C	0,95	0,95	0,95
	G	0,03	0,02	0,03	C/G	0,05	0,05	0,05
					G/G	0	0	0
His313His	C	0,75	0,79	0,72	C/C	0,57	0,61	0,54
	T	0,25	0,21	0,28	C/T	0,36	0,36	0,36
					T/T	0,07	0,03	0,1
rs6277	T	0,6	0,64	0,58	T/T	0,37	0,39	0,36
	C	0,4	0,36	0,42	T/C	0,46	0,49	0,43
					C/C	0,17	0,11	0,21
TaqIA	G	0,79	0,78	0,8	G/G	0,63	0,6	0,65
	A	0,21	0,22	0,2	G/A	0,33	0,37	0,3
					A/A	0,04	0,03	0,05

Taula 13.2. P valors de Hardy-Weinberg

SNP	TOTAL	CTROLS	CASOS
-241	0,62	1	1
-141 Ins/Del	**0,012**	1	**0,0039**
TaqIB	0,11	0,31	**0,0025**
rs1800499	1	1	1
Ser311Cys	1	1	1
His313His	0,35	0,47	0,1
rs6277	0,4	0,4	0,082
TaqIA	1	0,35	0,41

Taula 13.3. Models d'associació amb Esquizofrènia per cada polimorfisme del gen DRD2

SNP	MODEL	GENOTIP	CONTROLS	CASOS	OR(95%IC)	P VALOR	AIC	BIC
-241	–	A/A	85(90,4%)	202(90,6%)	1	0,61	367,5	382,5
		A/G	9(9,6%)	21(9,4%)	1,25(0,53-2,98)			
-141	Dominant	I/I	83(87.4%)	191(85.7%)	1.00	0,57	368	383
		I/D-D/D	12(12,6%)	32(14,3%)	1,24(0,59-2,62)			
TaqIB	Sobredominant	G/G-A/A	71 (75,5%)	186 (83,4%)	1	0,034	362,5	377,5
		A/G	23(24.5%)	37(16.6%)	**0.50(0.27-0.94)**			
rs1800499	–	G/G	85(89.5%)	209(95%)	1	0,14	363,4	378,4
		A/G	10(10.5%)	11 (5%)	0,49(0,19-1,24)			
Ser311Cys	–	C/C	89(93,7%)	213(96%)	1,00	0,44	366,3	381,3
		C/G	6(6,3%)	9 (4%)	0.64 (0.21-1.94)			
His313His	Aditiu	–	–	–	**1,57(1,01-2,42)**	0,038	361,1	376,2
rs6277	Recessiu	T/T-C/T	86(90.5%)	176 (80%)	1	0,028	361,7	376,7
		C/C	9(9.5%)	44(20%)	**2,32(1,05-5,10)**			
TaqIA	Sobredominant	G/G-A/A	58 (61%)	155 (69.8%)	1	0,28	367,4	382,5
		A/G	37(39%)	67(30.2%)	0,75(0,44-1,27)			

Taula 13.4. Valors de D′ per la mostra general

	-241	-141	TaqIB	rs1800499	Ser311Cys	His313His	rs6277	TaqIA
-241	.	0.7592	0.0939	0.9657	0.0922	0.3774	0.5257	0.0307
-141	.	.	0.3894	0.4036	0.0023	0.6998	0.8107	0.4476
TaqIB	.	.	.	0.6121	0.9758	0.8954	0.9993	0.9863
rs1800499	.	.	.	.	0.0382	0.9927	0.8361	0.6837
Ser311Cys	.	.	.	.	.	0.6586	0.7174	0.2602
His313His	.	.	.	.	.	.	0.9608	0.8433
rs6277	.	.	.	.	.	.	.	0.4744

Taula 13.5. Valors d'r per la mostra general

	-241	-141	TaqIB	rs1800499	Ser311Cys	His313His	rs6277	TaqIA
-241	.	−0.0507	0.058	−0.0415	−0.0035	0.1546	0.1533	0.0143
-141	.	.	−0.0421	−0.0206	0.0013	0.3411	0.2812	−0.0641
TaqIB	.	.	.	−0.0425	−0.0605	−0.1985	0.4715	0.7446
rs1800499	.	.	.	.	0.034	−0.1041	−0.1231	−0.0629
Ser311Cys	.	.	.	.	.	0.184	0.1426	−0.0214
His313His	.	.	.	.	.	.	0.6838	−0.2477
rs6277	.	.	.	.	.	.	.	0.2965

Taula 13.6. P Valors per la mostra general

	-241	-141	TaqIB	rs1800499	Ser311Cys	His313His	rs6277	TaqIA
-241	.	0.1557	0.1037	0.2455	0.9213	0	0	0.6876
-141	.	.	0.2393	0.5647	0.9705	0	0	0.0728
TaqIB	.	.	.	0.2342	0.0913	0	0	0
rs1800499	.	.	.	.	0.3423	0.0037	$6e-04$	0.078
Ser311Cys	.	.	.	.	.	0	$1e-04$	0.5504
His313His	.	.	.	.	.	.	0	0
rs6277	.	.	.	.	.	.	.	0

Pel que fa a l'anàlisi amb múltiples SNPs, el primer que es mostra són les freqüències haplotípiques calculades mitjançant el programa Haplo.Stats i mitjançant el programa BayHap. Com es pot observar a la taula 13.7, un haplotip és el més freqüent de la mostra amb

Taula 13.7. Freqüència haplotípica i OR amb intervals de confiança segons BayHap i Haplo.Stats (H.S). Haplotips referents als SNPs per ordre: -241, -141, TaqIB, rs1800499, Ser311Cys, His313His, 6277, TaqIA

Haplotip	Freq	IC-Freq 95%	OR BayHap	IC-OR 95% BayHap	OR H.S	IC-OR 95% H.S
haplo.208=AIGGCCTG	0.478	(0.441, 0.515)	1	−−	1	−−
haplo.176=AIGGCTCG	0.143	(0.117, 0.170)	1.4669	(0.92884, 2.34314)	**1.88**	**(1.02,3.45)**
haplo.12=AIAGCCCA	0.110	(0.087, 0.136)	0.83199	(0.51452, 1.35013)	0.88	(0.50, 1.56)
haplo.80=AIGGCCTA	0.064	(0.047, 0.081)	1.00699	(0.54553, 1.82225)	1.05	(0.48, 2.32)
haplo.174=ADGGCTCG	0.056	(0.040, 0.075)	**2.16072**	**(1.00755,4.8235)**	2.11	(0.77, 5.77)
haplo.81=GIGGCTCG	0.026	(0.015, 0.039)	0.75903	(0.28866, 2.18021)	0.83	(0.27, 2.59)
haplo.200=AIGACCTG	0.029	(0.018, 0.042)	0.61185	(0.23707, 1.50649)	0.48	(0.15, 1.47)
haplo.192=AIGGGTCG	0.016	(0.008, 0.025)	2.589	(0.7265, 12.21922)	NA	(NA, NA)
haplo.144=AIGGCCCG	0.015	(0.007, 0.024)	0.5223	(0.13014, 1.82668)	1.25	(0.23, 6.74)
rares (freq<0.01)	0.064	(0.047, 0.083)	−−	−−	−−	−−

una freqüència del 48%. Un 5% de la mostra haplotípica està formada per haplotips que es presenten amb una freqüència inferior al 1%.

Pel que fa a l'anàlisi d'associació amb els haplotips, a la tesi on aquestes dades ja havien estat analitzades, s'havia fet imputació haplotípica i posteriorment un test d'independència de χ^2. Afegim les estimacions dels valors d'OR que retornen BayHap i Haplo.Stats, ajustant els models per sexe i edat. L'algorisme EM de Haplo.Stats no convergeix per a alguns haplotips. Els resultats obtinguts són clarament diferents i de fet, els de l'Haplo.Stats són poc fiables donada la no convergència. BayHap ha convergit com es pot observar al test de convergència que retorna el valor "passed"pel test d'estacionarietat, i també observant

els gràfics de mitjana ergòdica (13.2). Per comprovar que el mètode ha funcionat correctament, també és necessari observar els gràfics de les densitats (13.3), de les autocorrelacions (13.4) i la variabilitat de la seqüència (13.5). Com mostren els gràfics, la mitjana ergòdica està estabilitzada, els gràfics de densitat per cada coeficient del model de regressió es distribueixen aproximadament de manera normal, les correlacions de les cadenes són nul·les i la variabilitat de la sèrie és constant.

Segons els resultats de BayHap, les conclusions de l'estudi varien.

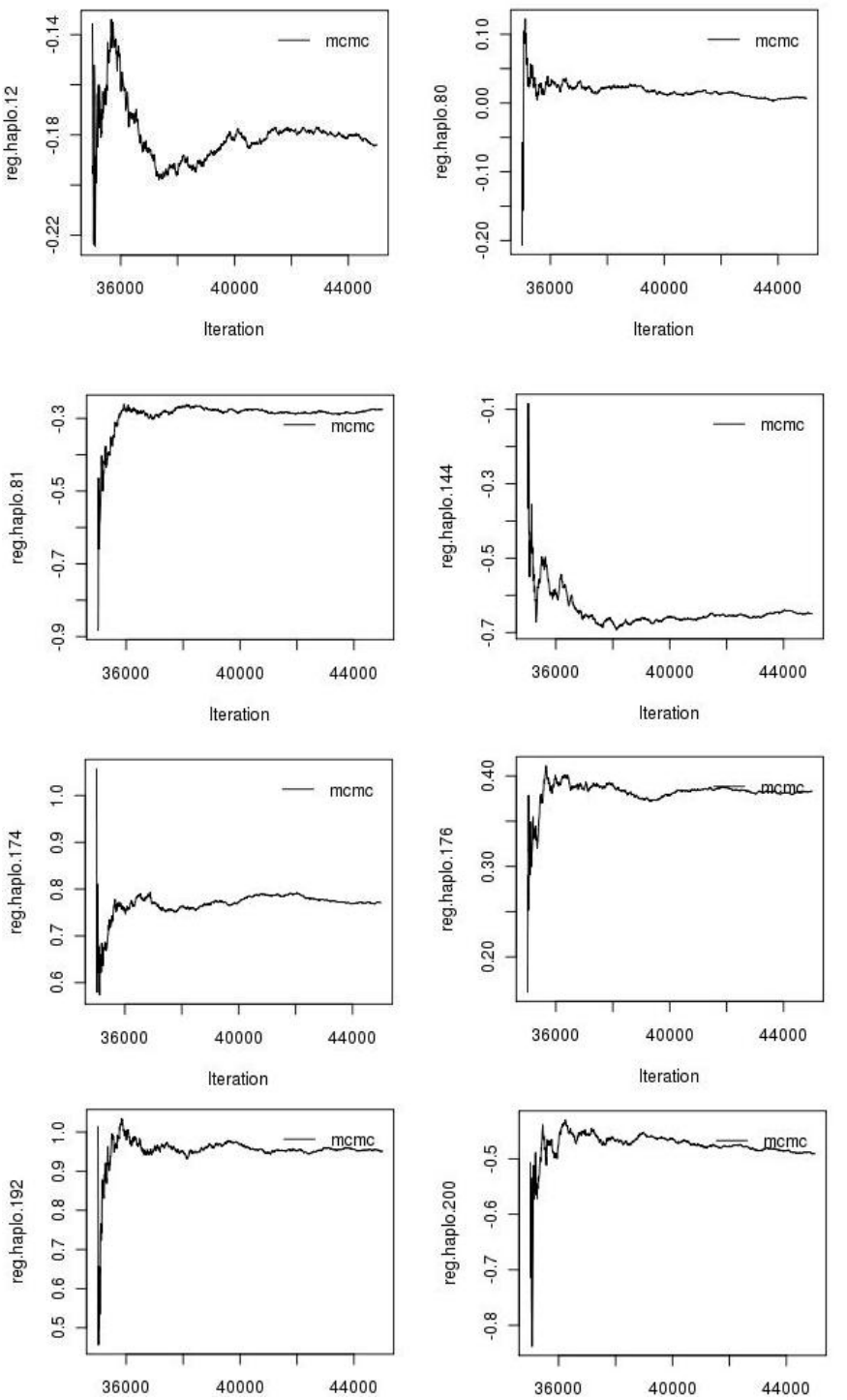


Figura 13.2. Mitjanes ergòdiques per cada coeficient de la regressió logística corresponent a cada haplotip en la mostra d'esquizofrènia.

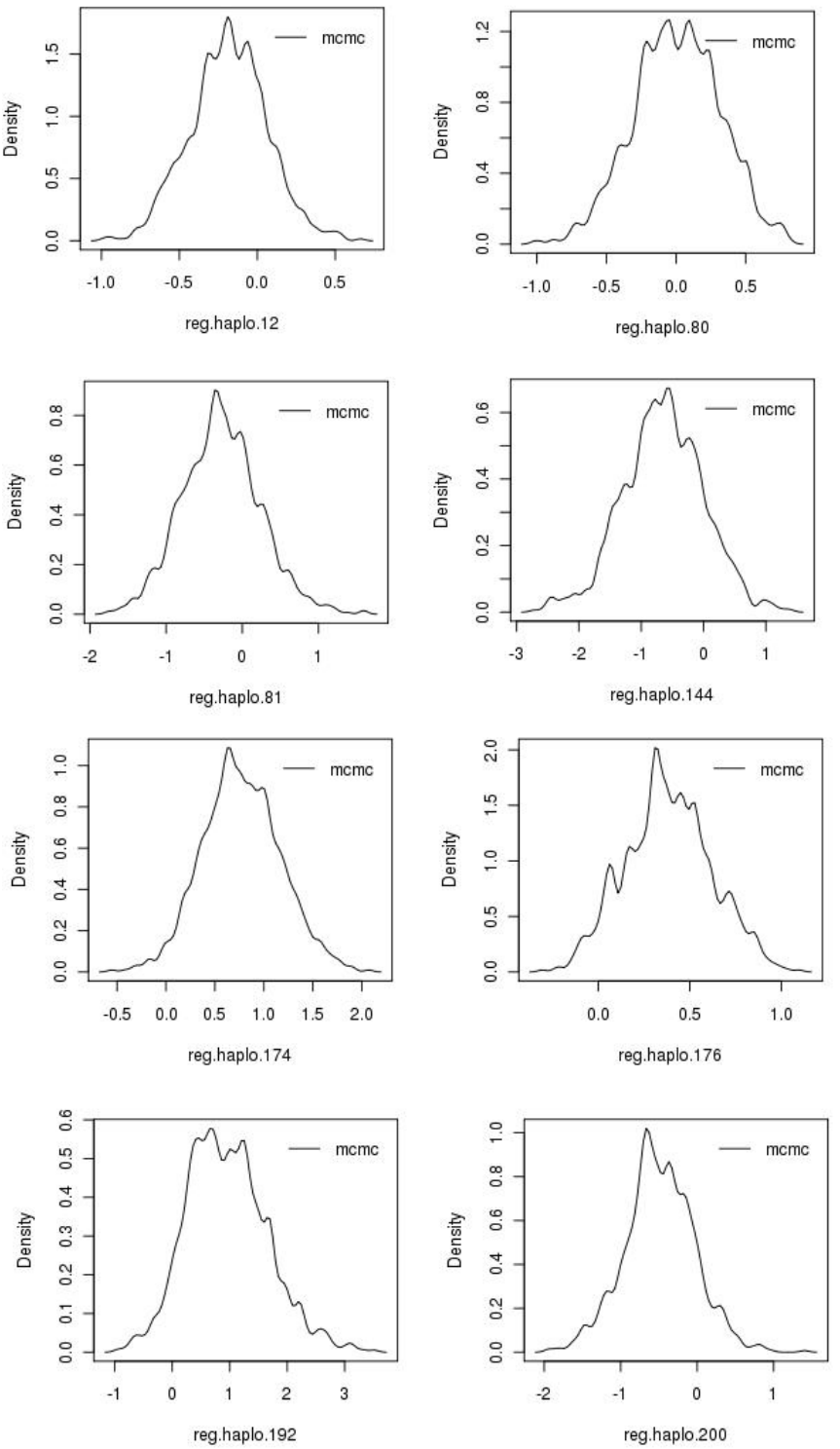


Figura 13.3. Densitats del mostreig realitzat per cada coeficient de la regressió en la mostra d'esquizofrènia.

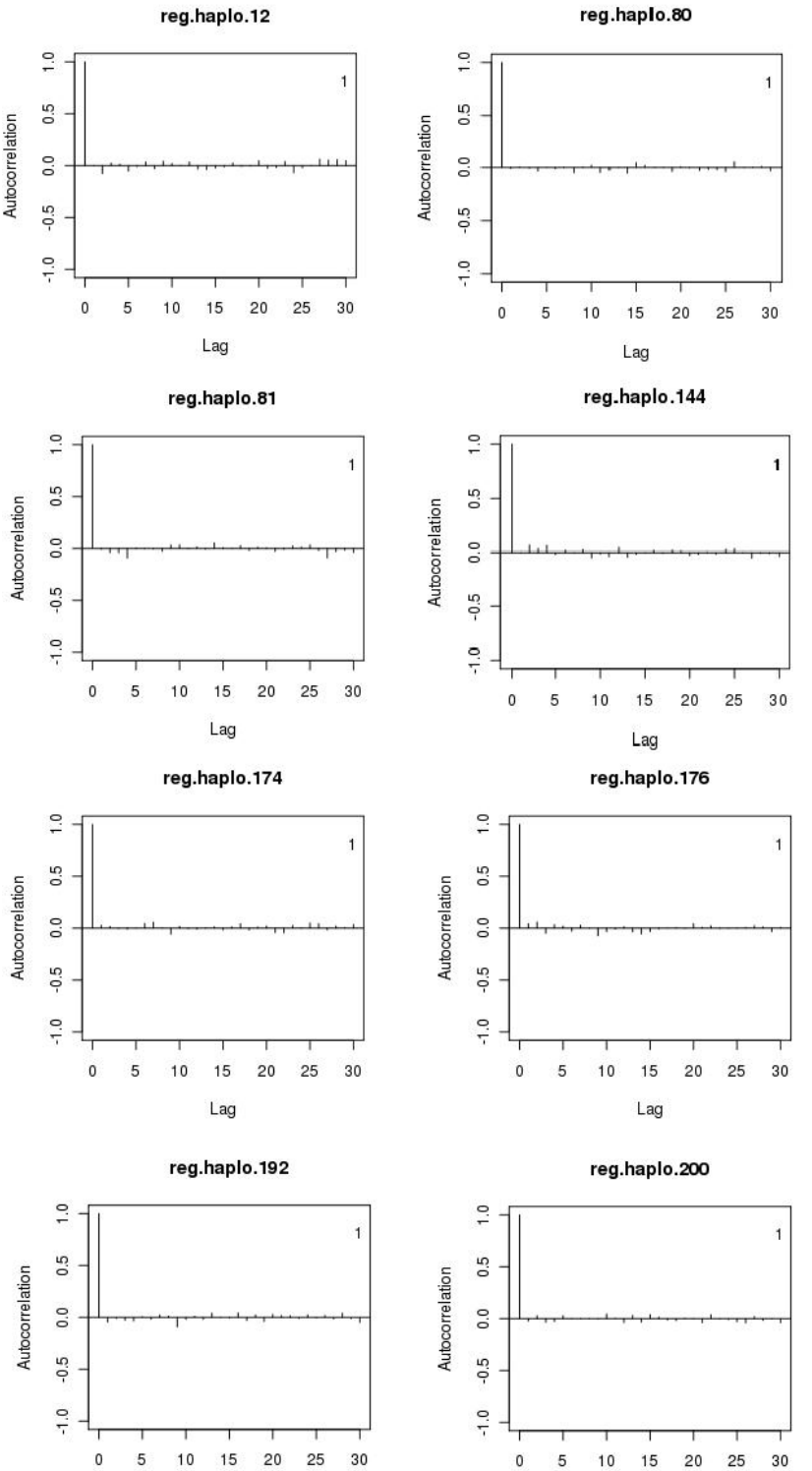


Figura 13.4. Autocorrelacions parcials de cadascuna de les cadenes en la mostra d'esquizofrènia.

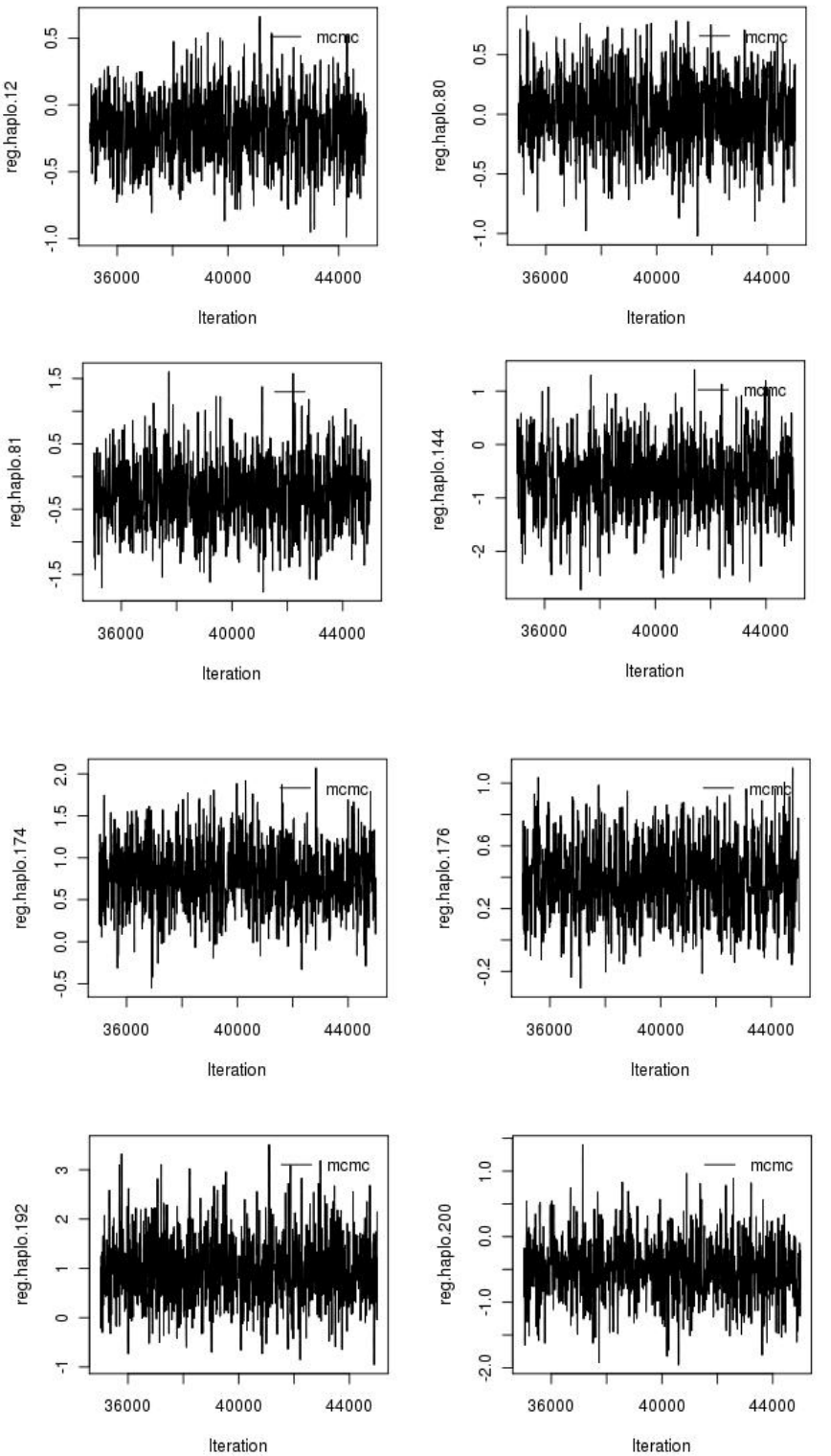


Figura 13.5. Sèries per a cada coeficient de la regressió en la mostra d'esquizofrènia.

13.4 Estudi cas-control en càncer de còlon

Mitjançant un estudi de cas-control s'han avaluat interaccions ambientals en relació al risc de patir càncer colorectal. Els casos van ser pacients amb un diagnòstic nou de adenocarcinoma colorectal que assistiren a un Hospital Universitari de Barcelona entre gener de 1996 i desembre de 1998. Aquest estudi inclou els 370 individus (72 % dels elegibles) que van poder ser entrevistats i que van proporcionar mostres biològiques de les anàlisis genètiques de suficient qualitat. Les negatives van ser un 2% dels elegibles, mentre que el 14% no va poder ser entrevistat perquè o bé havien mort, hi havia algun tipus d'impediment mental o bé se'ls va donar d'alta i no van poder ser localitzats. Finalment, un 12 % van ser entrevistats, però no van oferir mostres biològiques. Aquests casos perduts van ser similars als inclosos pel que fa a l'edat, sexe, localització del tumor i l'extensió. Per evitar biaixos de selecció, el criteri per a la inclusió dels casos va ser que el motiu de l'ingrés a l'hospital fou un nou cas de malaltia (no diagnosticat prèviament). Aquest criteri es va utilitzar per evitar la inclusió de pacients amb malalties cròniques, que podrien ser ingressats a l'hospital en diverses ocasions i modificar els seus hàbits a causa de la seva malaltia.

Els casos i els controls van ser entrevistats per personal entrenat, mitjançant un qüestionari estructurat. Es tracta d'una qüestionari sobre història dietètica, prèviament elaborat i validat dins el marc de Recerca Prospectiva Europea, en un estudi sobre Càncer i Nutrició que es basa en avaluar els aliments consumits durant l'any abans del diagnòstic. Els Grups d'Aliments es van basar en les propietats bromatològiques i varen ser calculats a partir dels productes consumits. Altres factors de risc mesurats van ser l'índex de massa corporal al moment del diagnòstic i 10 anys abans, la història del pacient pel que fa als fàrmacs presos, amb especial èmfasi en fàrmacs antiinflamatoris no esteroïdals , el consum de tabac i el d'alcohol. També es va recollir informació sobre antecedents familiars de neoplàsies de primer i segon grau. Els casos pertanyents a la poliposi adenomatosa familiar es van excloure però tres casos que complien amb els criteris d'Amsterdam per el Càncer Colorectal

hereditari sense poliposi no es van excloure.

Els controls (n = 327, 69,4% dels elegibles) van ser persones que vivien a la mateixa zona i que eren representatives de la població general, triats a l'atzar entre els pacients ingressats al mateix hospital durant el mateix període de temps. Les negatives van ser de 7% dels elegibles, mentre que el 5% no es va poder entrevistar a causa de deficiència mental o altres impediments. Finalment, 87 (18,6%) van ser entrevistats, però no van oferir una mostra de sang.

13.4.1 Polimorfismes del gen DRD2 analitzats en aquest estudi

Per investigar si els polimorfismes funcionals dins de DRD2 poden tenir un paper en la modulació del risc del càncer colorectal esporàdic, s'analitzen els genotips obtinguts en 370 casos i 327 controls per a set SNPs de DRD2 (141Cdel, TaqIB, TaqIA, S311Cys, rs6277,1412G i 3208T).

13.4.2 Resultats de l'anàlisi d'associació

Per cada polimorfisme es testa l'equilibri de Hardy-Weinberg als controls. Per provar la hipòtesi d'associació entre polimorfismes genètics i càncer colorectal, s'utilitzen mètodes multivariats basats en la regressió logística obtenint-se l'Odds ratio (OR) i els intervals de confiança al 95%. La categoria de referència són els de menor nivell d'exposició. Pels polimorfismes, els homozigots per l'al·lel més freqüent entre els controls s'estableix com la categoria de referència. Les proves de tendència lineal i dels OR es van calcular mitjançant després d'assignar un score lineal a cada categoria endreçada. Per polimorfismes, a l'homozigot per l'al·lel més freqüent (el de referència) se li dona una puntuació d'1, 2 als heterozigots, i 3 als homozigots per l'al·lel menys freqüent. Els p valors es calculen mitjançant el test de raó de versemblança. L'anàlisi es fa sota un model codominant (tres genotips separats). També es van considerar el model dominant (heterozigots agrupats amb els homozigots per l'al·lel menys freqüent) o el model recessiu (heterozigots agrupat amb els

homozigots per l'al·lel comú) en cas que les similituds d'OR suggerissin un millor ajust per aquests models que pel codominant. Totes les anàlisis estan ajustades per edat i sexe. El nivell de significació va ser del 5% (a dues cues). Els haplotips es reconstrueixen i s'analitzen primerament utilitzant el programa PHASE Versió 2 [122] i després segons BayHap. Els resultats que mostrarem fan referència a les anàlisis publicades a [178]. Utilitzant BayHap, afegirem les freqüències haplotípiques estimades per comparar respecte PHASE, recalcularem les associacions mitjançant BayHap, i a més, s'afegirà un anàlisi de pronòstic utilitzant THESIAS i BayHap.

A la taula 13.8 es mostren les freqüències al·lèliques i genotípiques pels SNPs analitzats. Els resultats de HWE es poden consultar a la taula 13.4.2. Tots els SNPs es troben en HWE tan per la mostra de casos com per la de controls. Els resultats de les anàlisis d'assoiació amb SNPs es presenten a la taula 13.10. Observem associació entre 141Cdel, TaqIB i 957C de DRD2 i el càncer colorectal. El polimorfisme 141Cdel és el que presenta menor pvalor. Aquesta variant és rara i només quatre homozigots (dos casos i dos controls) van ser detectats a la mostra. El model dominant va confirmar l'associació (OR=2.8; 95% IC, 1.38-3.76). Per confirmar aquests resultats per 141Cdel, es va genotipar de nou tots els casos i controls amb l'assaig de nucleasa 5V (TaqMan), i es van obtenir els mateixos resultats. L'augment del risc de càncer es va seguir observant en estratificar les mostres segons còlon i recte. (OR=3.35, IC 95%(1.67,6.7) i OR=2.22 IC 95%(0.97,5.09) respectivament).

En el conjunt de mostres, el polimorfisme TaqIB també ha aparegut associat a un increment de risc de càncer colorectal, mostrant un major OR per als homozigots variants (OR, 1,41; 95% IC, 1,01-1,96).

Per investigar més a fons aquestes associacions, s'analitzen els haplotips de DRD2 composats pels 7 SNPs, presos en el seu ordre físic. A la taula 13.11 es pot veure com només l'haplotip DGGCCGC es troba significativament associats amb el càncer colorectal segons el model de regressió logística OR=2.86 IC95%(1.58,5.18). BayHap troba resultats similars,

però retorna un interval de confiança més ampli OR=2.72 IC95%(1.41,5.74). Aquest haplotip inclou al·lels 141Cdel, 957C, i 1412G. Com que l'haplotip CGGCCGC també porta el 957C, les variants i 1412G, però no 141Cdel i aquest no apareix relacionat amb càncer colorectal, sembla que el risc podria estar associat amb 141Cdel o un efecte cooperatiu d'aquestes variants.L'associació entre 957C i el càncer colorectal observat en els models dominants es podria deure a un desequilibri de lligament amb el polimorfisme 141Cdel. El polimorfisme de TaqIB es va trobar en un sol haplotip, que apareix dèbilment associat segons resultats de PHASE i sense associació segons BayHap (OR: 1.33 IC 95% (0.93,1.91) i OR=1.23 IC95%(0.78,1.93). En afegir l'anàlisi de supervivència observem que l'haplotip CGGCTAT que en l'estudi de cas-control no havia donat significatiu, mostra certa significació en l'anàlisi de supervivència. Ara bé, l'interval de confiança reportat per BayHap té un límit molt proper a 1 i això podria suggerir que aquest és un resultat degut a l'atzar. El programa THESIAS no el retorna com a significatiu.

L'associació entre el polimorfisme 141Cdel i el càncer colorectal va ser explorat en relació amb altres per excloure efectes confusors i detectar interaccions. El risc per al càncer colorectal va ser igualment alt, amb independència de sexe, grup d'edat, localització tumoral (Còlon o el recte), i l'estadi tumoral dels individus.

Pel que fa als resultats de BayHap, s'ha comprovat que la convergència és correcta, analitzant els gràfics corresponents i els testos de convergència que implementa BayHap. Observant la taula 13.11 podrem extreure diverses conclusions sobre el fet d'imputar haplotips o bé fer una estimació simultània.

Taula 13.8. Freqüències al·lèliques i genotípiques pels polimorfismes del gen DRD2 analitzats a la mostra de CCR.

SNP	AL·LEL	TOTAL	CTROLS	CASOS	GENO	TOTAL	CTROLS	CASOS
-141 Ins/Del	C	0,93	0,95	0,91	C/C	0,87	0,91	0,83
	T	0,07	0,5	0,09	C/T	0,12	0,08	0,16
					T/T	0,01	0,01	0,01
TaqIB	G	0,87	0,89	0,85	A/A	0,02	0,01	0,03
	A	0,13	0,11	0,15	A/G	0,22	0,21	0,23
					G/G	0,76	0,79	0,74
1412A>G	A	0,74	0,75	0,73	A/A	0,56	0,59	0,53
	G	0,26	0,25	0,27	A/G	0,36	0,32	0,39
					G/G	0,08	0,08	0,08
Ser311Cys	C	0,98	0,97	0,98	C/C	0,96	0,95	0,97
	G	0,02	0,02	0,03	C/G	0,04	0,04	0,03
					G/G	0,01	0,01	0
3208G>T	G	0,9	0,91	0,89	G/G	0,81	0,82	0,8
	T	0,1	0,09	0,11	G/T	0,17	0,17	0,18
					T/T	0,01	0,01	0,02
rs6277	T	0,6	0,63	0,57	T/T	0,37	0,41	0,33
	C	0,4	0,37	0,42	T/C	0,46	0,44	0,49
					C/C	0,17	0,16	0,18
TaqIA	C	0,82	0,83	0,81	C/C	0,68	0,7	0,67
	T	0,18	0,17	0,19	C/T	0,28	0,27	0,29
					T/T	0,04	0,03	0,04

SNP	TOTAL	CTROLS	CASOS
-141 Ins/Del	**0.52**	0.12	1
TaqIB	0.73	0.4	0.28
1412A>G	0.14	0.05	0.89
Ser311Cys	0.0024	0.058	0.083
3208G>T	0.36	1	0.23
rs6277	0.4	0.31	0.91
TaqIA	0.36	0.69	0.39

Taula 13.9. P valors de Hardy-Weinberg

Taula 13.10. Models d'associació amb càncer colorectal per cada polimorfisme analitzat del gen DRD2

SNP	MODEL	GENOTIP	OR(95%IC)	P VALOR
-141	Dominant	C/C	1	–
		C/T-T/T	2.28(1.38-3.76)	<0.001
TaqIB	–	G/G	1.00	–
		A/A	**4.90(1.07-22.54)**	0.046
3208G>T	Dominant	G/G	1	–
		G/T-T/T	1.13(0.75-1.72)	0.558
Ser311Cys	Dominant	C/C	1,00	–
		C/G-G/G	0.60 (0.26-1.38)	0.229
rs6277	Dominant	T/T	1	–
		T/C-C/C	**1.41(1.01-1.96)**	0.042
1412A>G	Dominant	A/A	1	–
		A/G-G/G	1.29(0.93-1.79)	0.126
TaqIA	Dominant	C/C	1	–
		C/T-T/T	1.16(0.84-1.61)	0.372

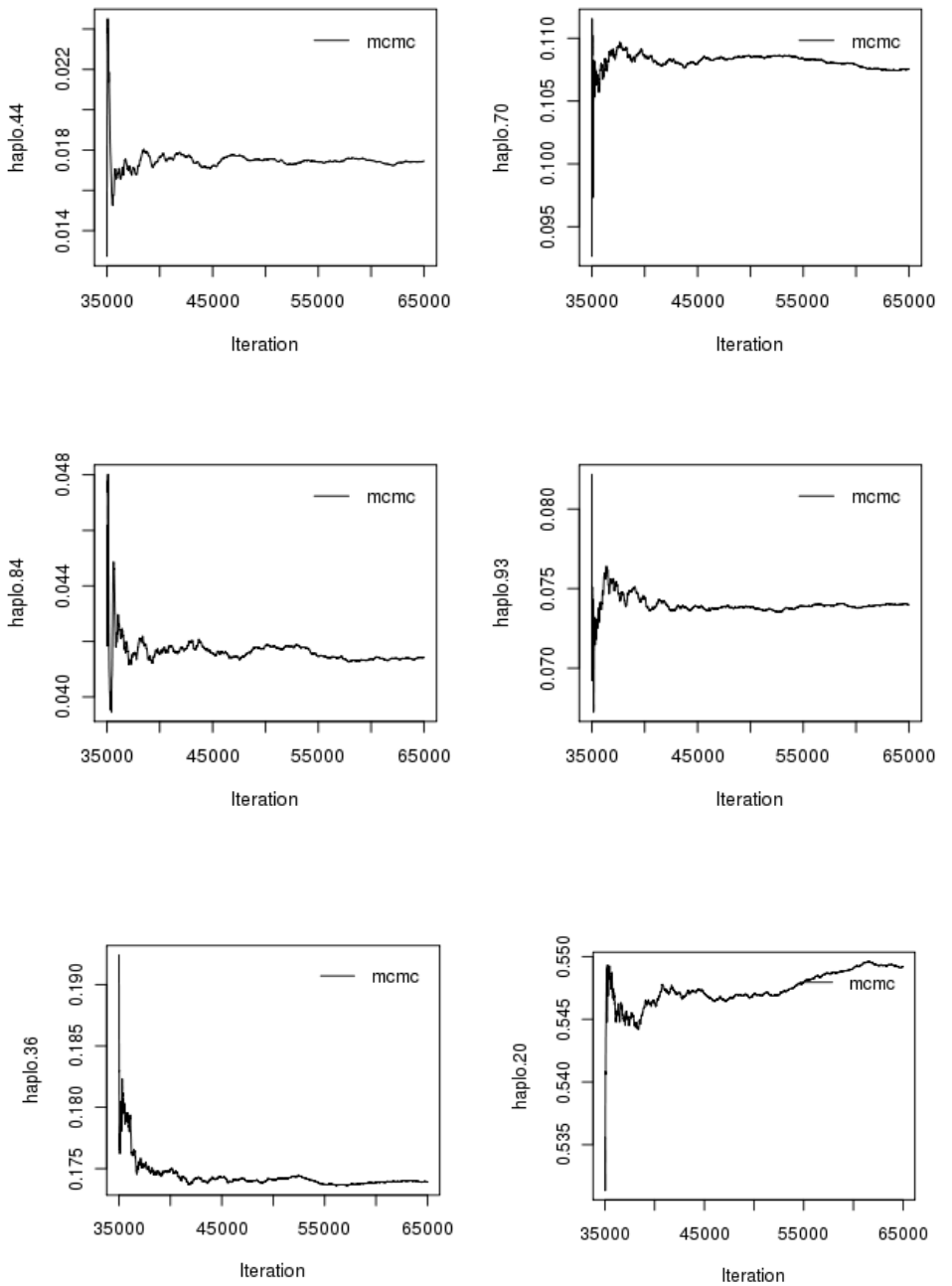


Figura 13.6. Mitjanes del mostreig realitzat per cada freqüència haplotípica.

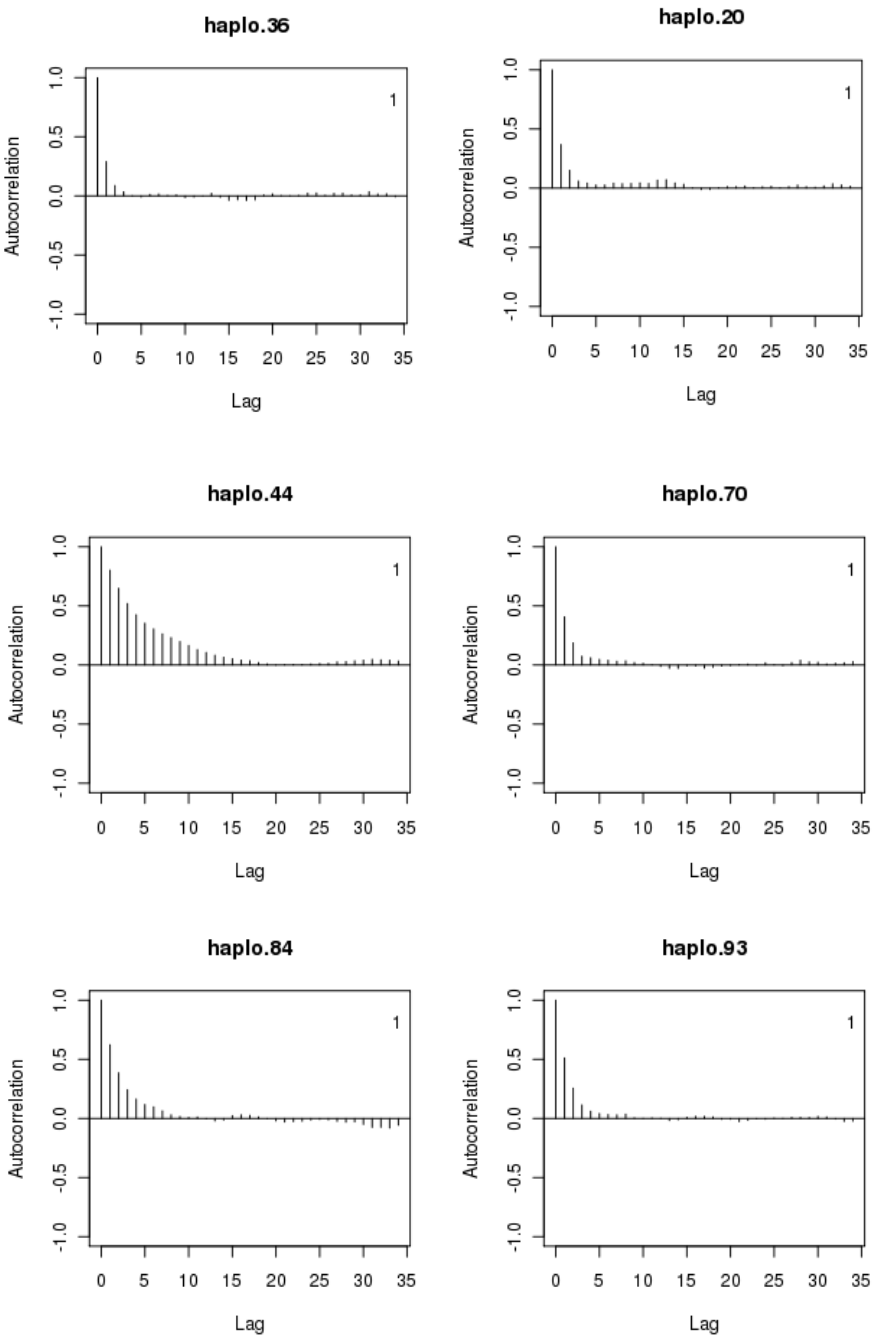


Figura 13.7. Autocorrelacions parcials del mostreig realitzat per cada freqüència haplotípica en la mostra de càncer.

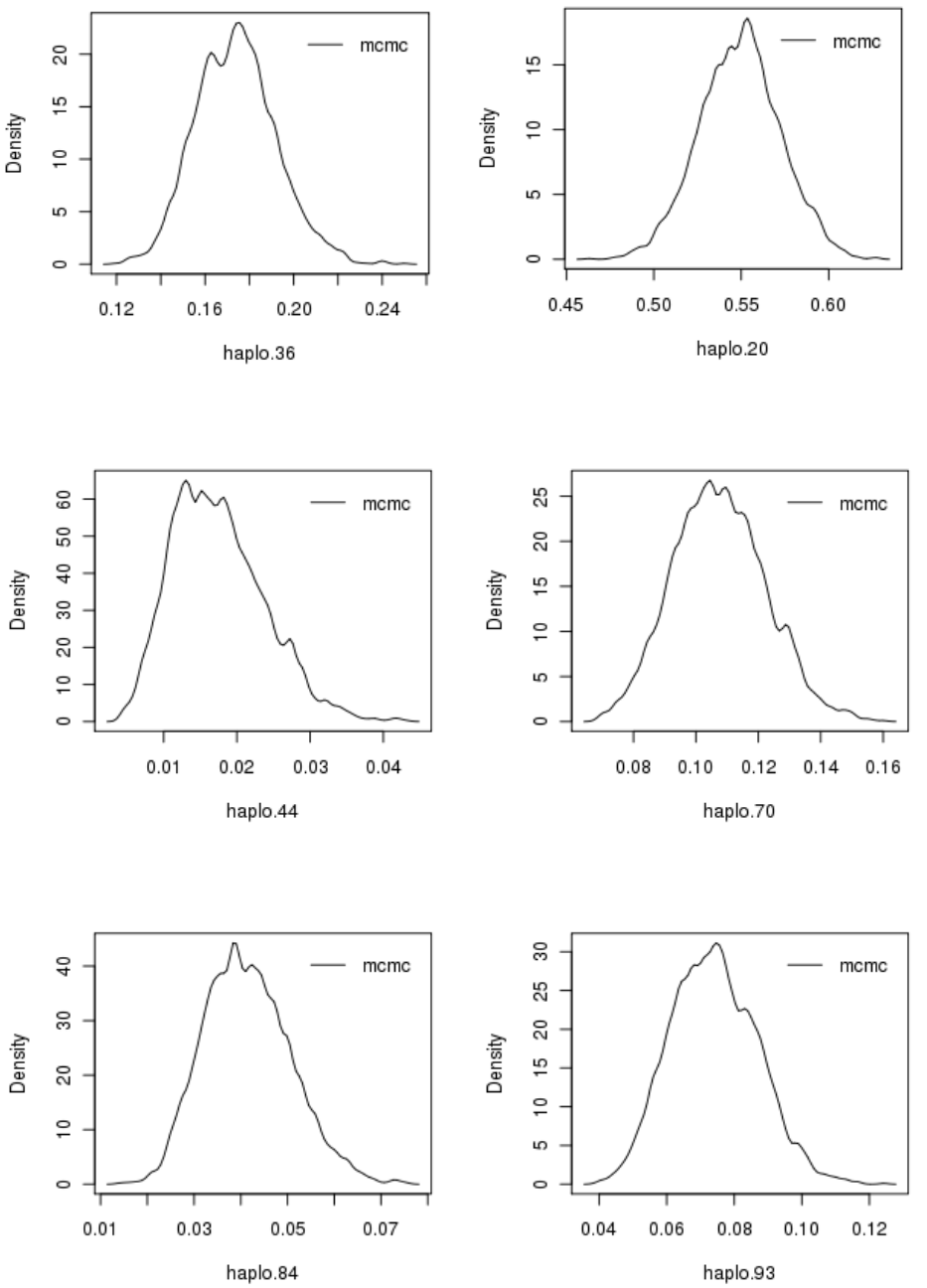


Figura 13.8. Densitats del mostreig realitzat per cada freqüència haplotípica en la mostra de càncer.

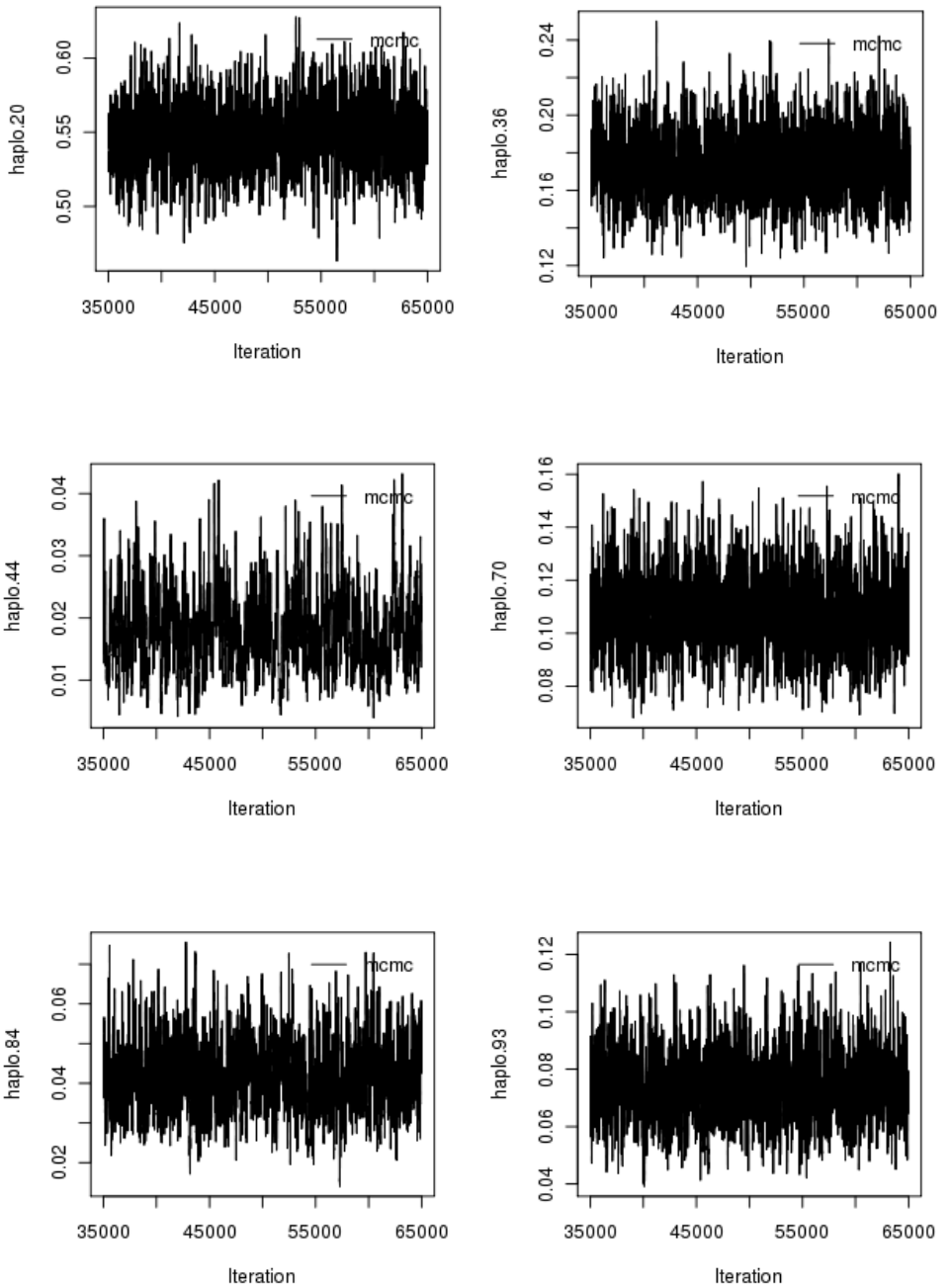


Figura 13.9. Seqüència mostrejada per cada freqüència haplotípica en la mostra de càncer.

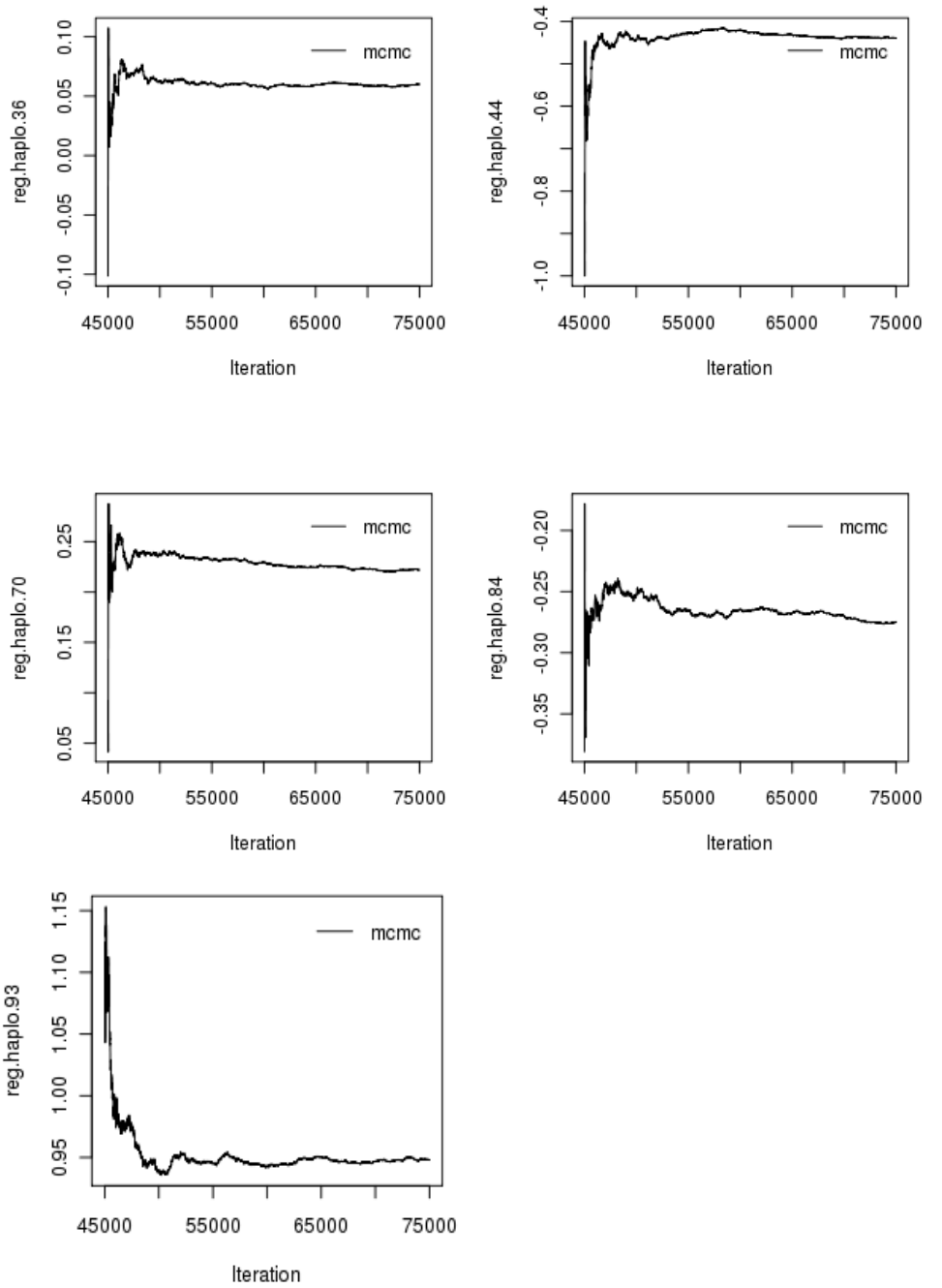


Figura 13.10. Mitjanes del mostreig realitzat per cada coeficient de la regressió Logística en la mostra de càncer.

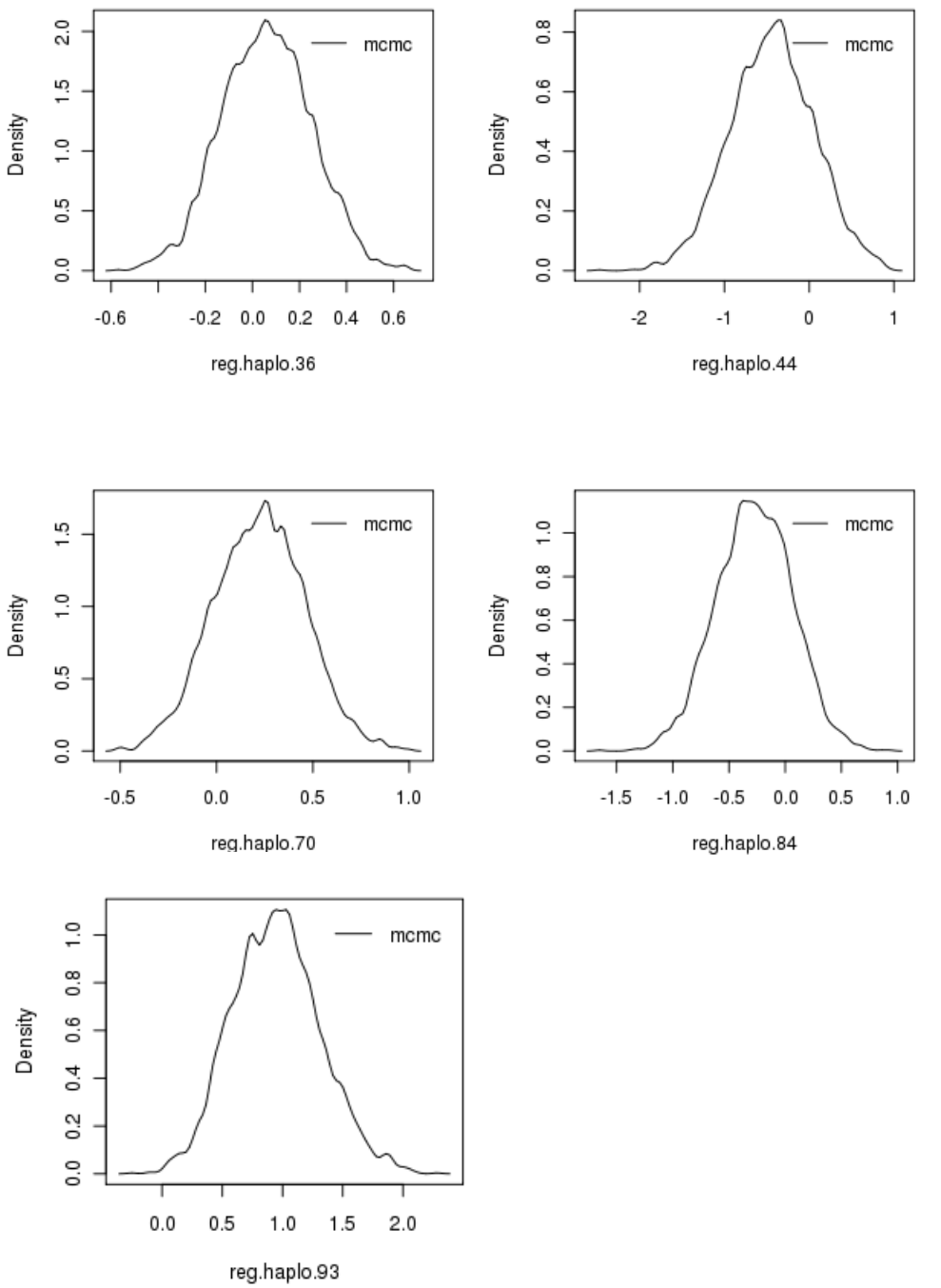


Figura 13.11. Densitats del mostreig realitzat per cada coeficient de la regressió Logística en la mostra de càncer.

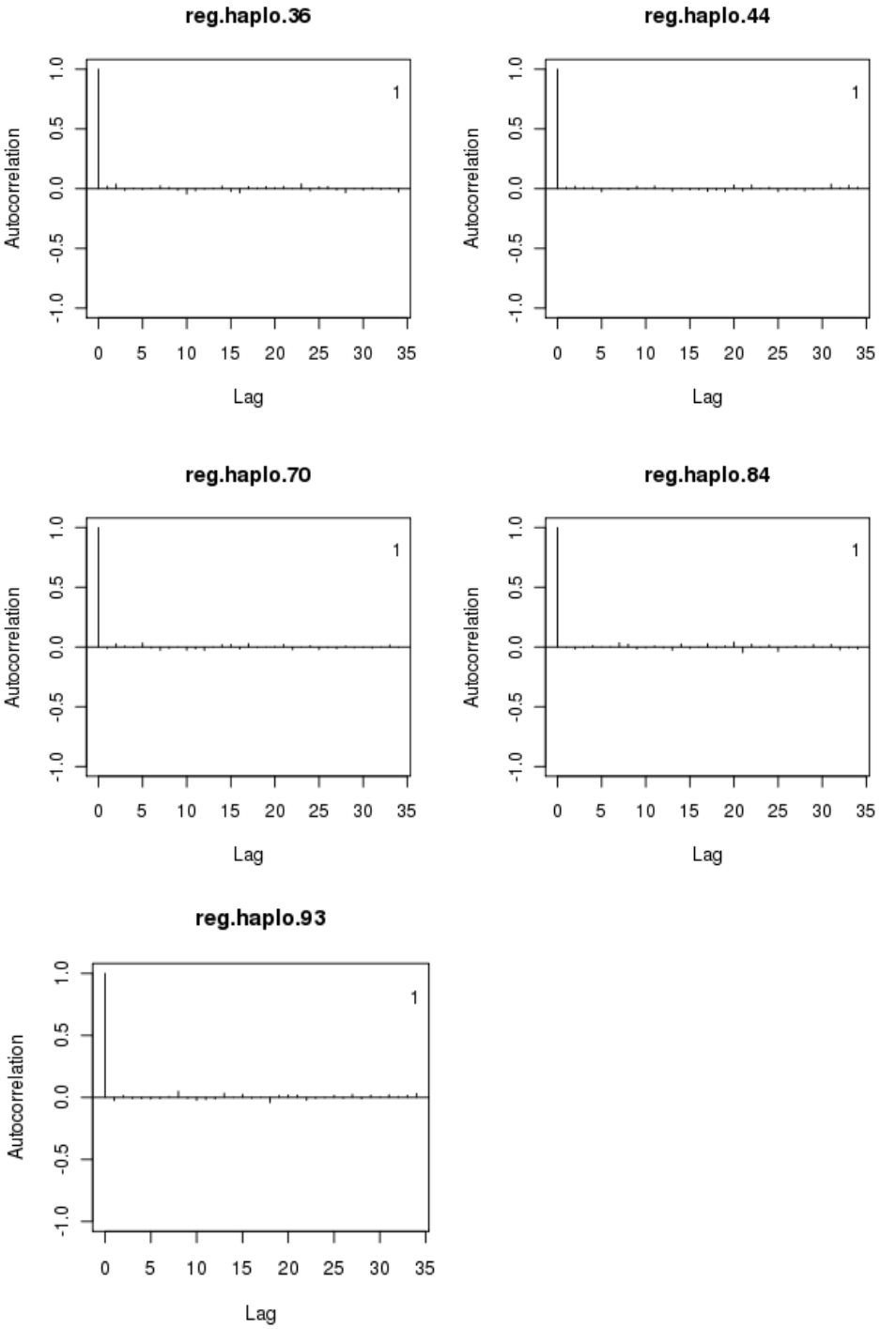


Figura 13.12. Autocorrelacions del mostreig realitzat per cada coeficient de la regressió Logística en la mostra de càncer.

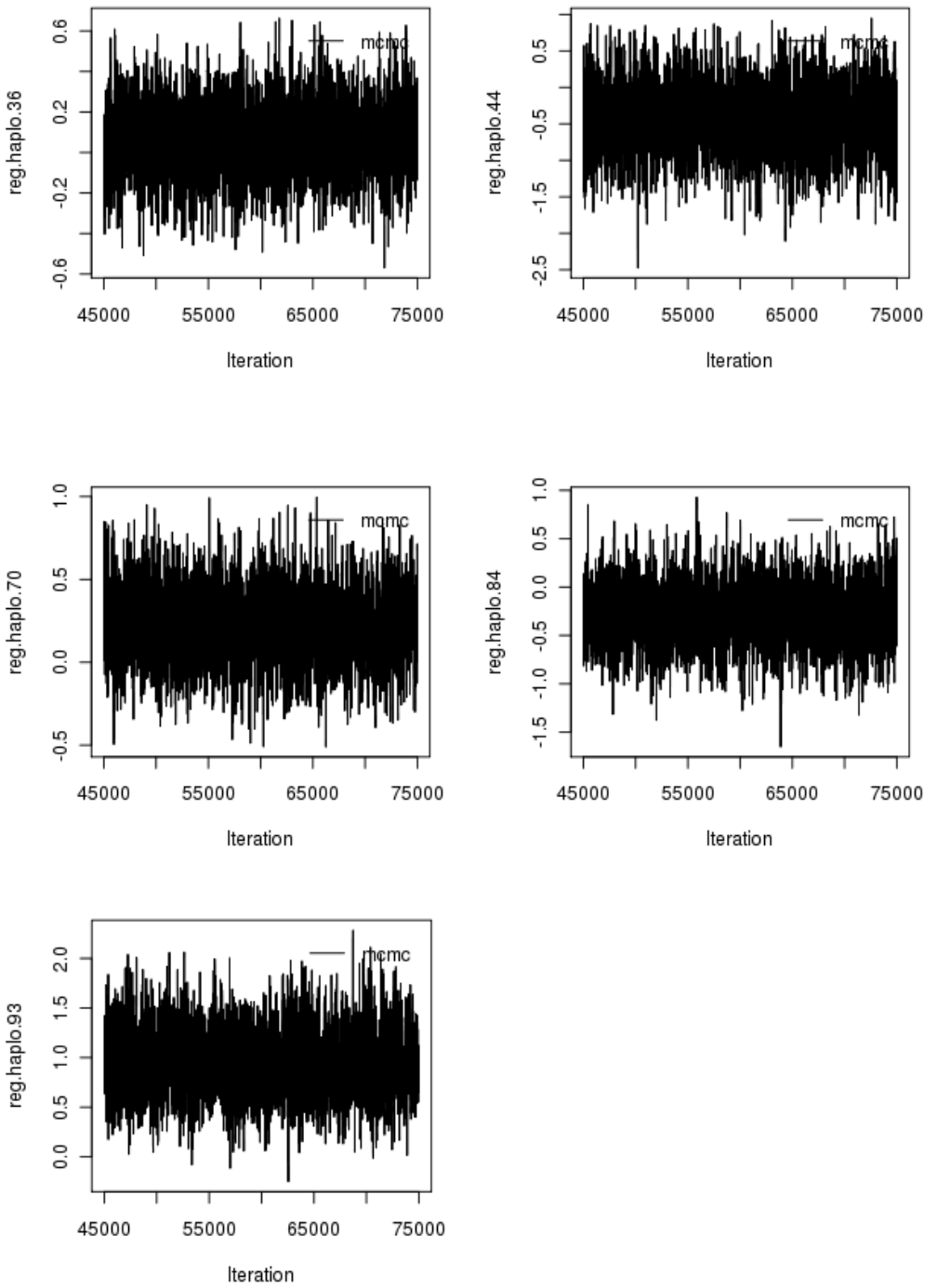


Figura 13.13. Termes de la serie temporal pel mostreig realitzat per cada coeficient de la regressió Logística en la mostra de càncer.

Taula 13.11. Freqüència haplotípica i OR segons PHASE i BayHap. Haplotips referents als SNPs per ordre: -141, TaqIB, 3208T, Ser311Cys, rs6277, 1412G, TaqIA

Haplotip	Freq BayHap (IC95%)	Freq PHASE	OR IC(95%) rec.PHASE	OR IC(95%)BayHap
haplo.20=CGGCTAC	0.56(0.53, 0.60)	0.56	1.00	1.00
haplo.36=CGGCCGC	0.18(0.15, 0.20)	0.19	1.04(0.78, 1.40)	1.03(0.72, 1.48)
haplo.44=CGGGCGC	0.02(0.01, 0.03)	0.02	0.69(0.27, 1.73)	0.63(0.24, 1.64)
haplo.70=CATCCAT	0.10(0.08, 0.12)	0.12	1.33(0.93,1.91)	1.23(0.78, 1.93)
haplo.84=CGGCTAT	0.05(0.04, 0.06)	0.04	0.84(0.49, 1.46)	0.73(0.37, 1.37)
haplo.93=DGGCCGC	0.05(0.04, 0.07)	0.05	**2.86(1.58,5.18)**	**2.72(1.41,5.74)**
rares	0.03(0.02, 0.05)	0.02	–	–

13.4.3 Resultats de l'anàlisi de supervivència

Taula 13.12. Freqüència haplotípica i HR segons BayHap i THESIAS amb intervals de confiança per l'estudi de CCR. Haplotips referents als SNPs per ordre: -141, TaqIB, 3208T, Ser311Cys, rs6277, 1412G, TaqIA

Haplotip	Freq BayHap (IC95%)	Freq THESIAS	HR IC(95%) BayHap	HR IC(95%)THESIAS
haplo.20=CGGCTAC	0.56(0.53, 0.60)	0.55	1.00	1.00
haplo.36=CGGCCGC	0.18(0.15, 0.20)	0.17	0.80(0.51, 1.23)	0.88(0.56, 1.37)
haplo.44=CGGGCGC	0.02(0.01, 0.03)	0.02	0.60(0.10, 2.13)	0.80(0.19, 3.41)
haplo.70=CATCCAT	0.10(0.08, 0.12)	0.11	0.82(0.47, 1.37)	0.86(0.51, 1.45)
haplo.84=CGGCTAT	0.05(0.04, 0.06)	0.04	**0.33(0.08,0.99)**	0.43(0.13, 1.38)
haplo.93=DGGCCGC	0.05(0.04, 0.07)	0.07	0.65(0.29, 1.30)	0.68(0.32, 1.45)
rares	0.03(0.02, 0.05)	0.02	–	–

En general els resultats de BayHap i de THESIAS són similars, excepte per l'haplotip haplo.44, de baixa freqüència (0.02). Per aquest haplotip, BayHap retorna un HR de 0.60 i THESIAS de 0.80. Aquest és el cas en que les estimacions disten més. Pel que fa a la resta

d'haplotips, BayHap retorna com a significant l'haplotip haplo.84 a diferència de THESIAS. Tot i així, l'interval de confiança té límit superior molt proper a 1. Les convergències per aquests coeficients segons BayHap es poden consultar als següents gràfics:

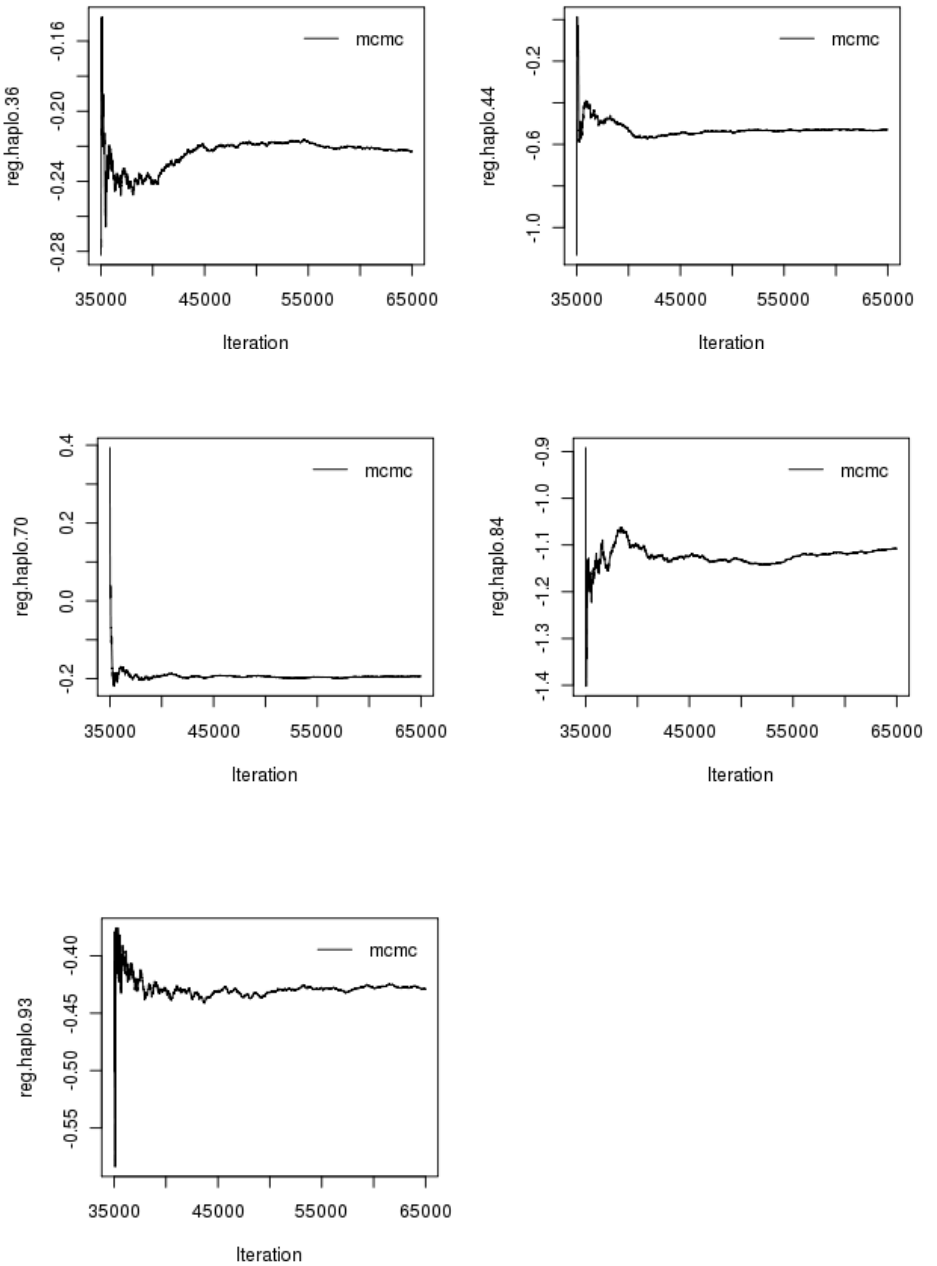


Figura 13.14. Mitjanes del mostreig realitzat per cada coeficient de la regressió de Weibull en la mostra de càncer.

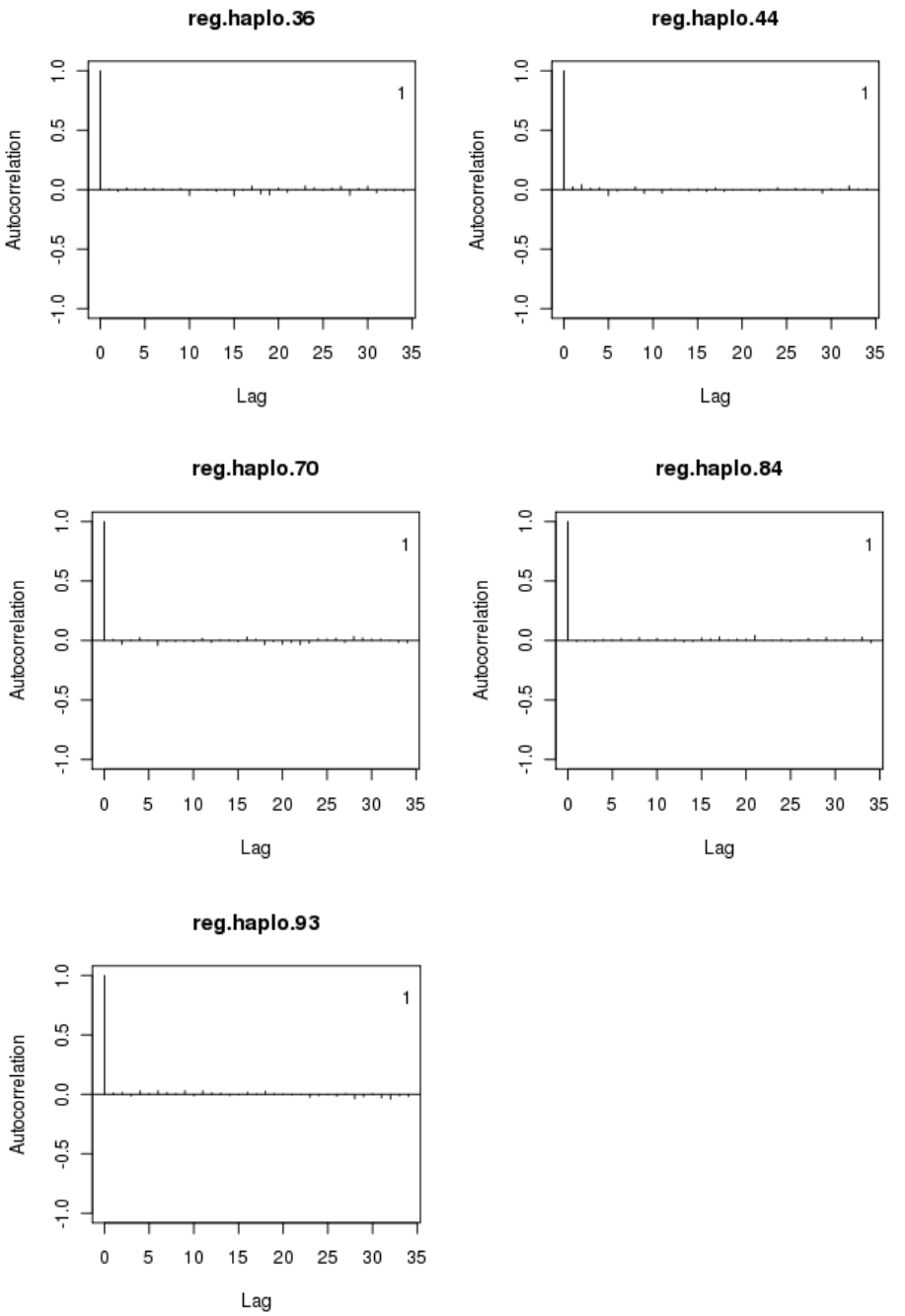


Figura 13.15. Autocorrelacions parcials del mostreig realitzat per cada coeficient de la regressió de Weibull en la mostra de càncer.

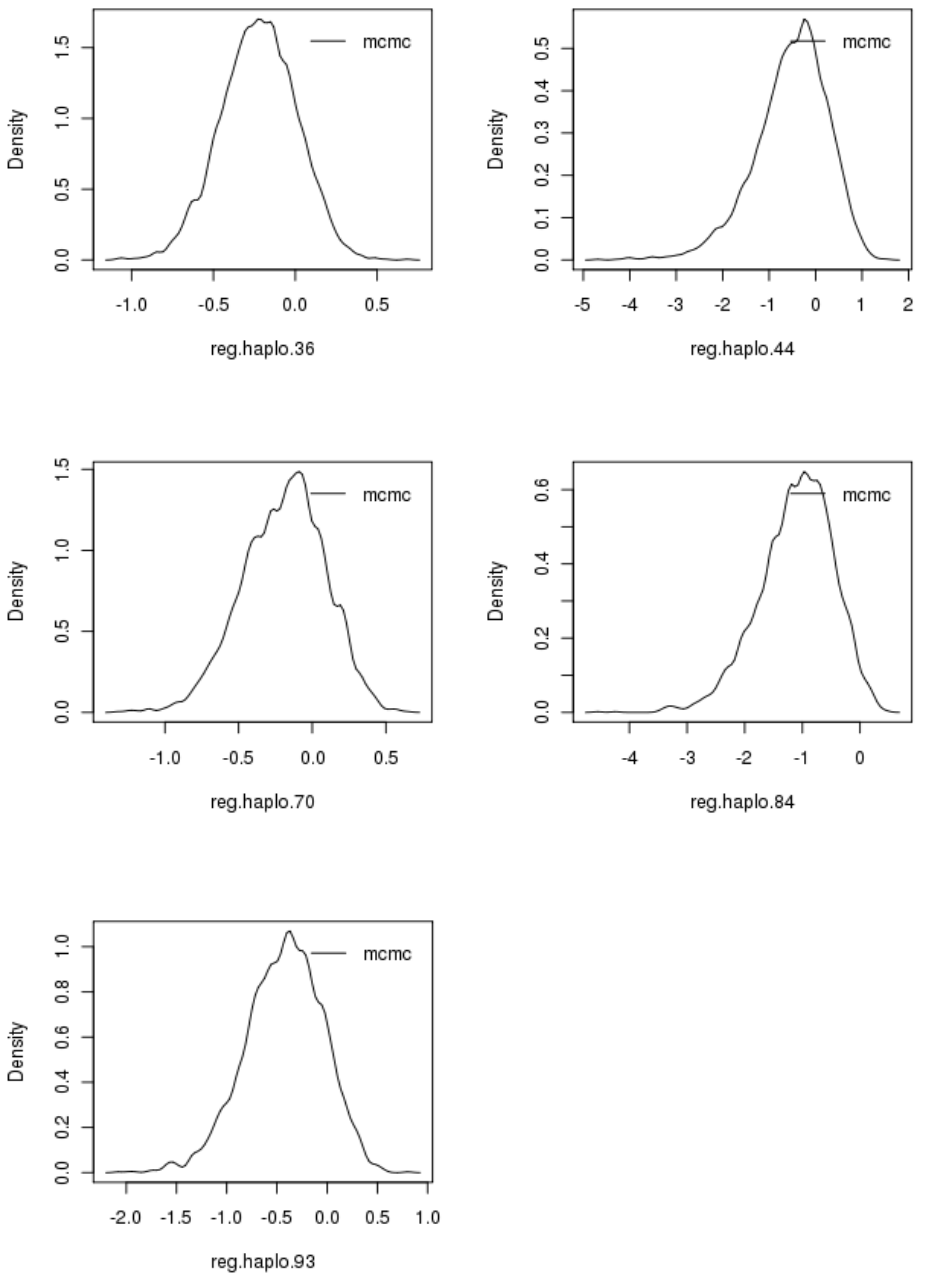


Figura 13.16. Densitats del mostreig realitzat per cada coeficient de la regressió de Weibull en la mostra de càncer.

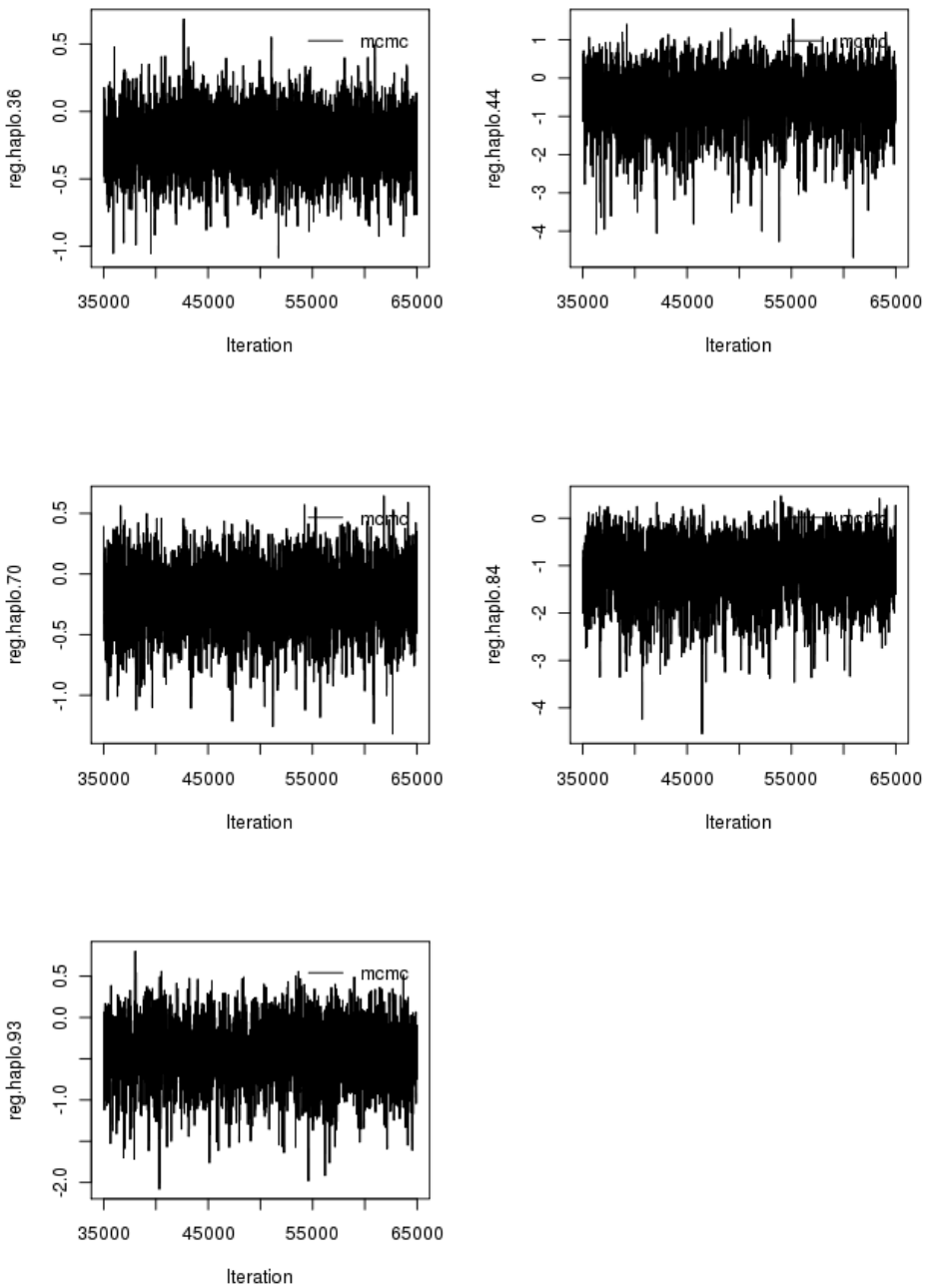


Figura 13.17. Densitats del mostreig realitzat per cada coeficient de la regressió de Weibull en la mostra de càncer.

14

Diferents consideracions de distribucions a priori

El programa BayHap a diferència de la resta de programes ofereix la possibilitat de considerar distribucions a priori pels paràmetres d'interès. La consideració d'una distribució a priori es basa en el coneixement per part de l'investigador del comportament d'aquests paràmetres. Donat que no sempre es disposa d'aquesta informació, el programa també s'executa per defecte amb una distribució prior no informativa. En aquest apartat, es mostren únicament els resultats de diferents anàlisis on es consideren la prior no informativa i de la normal multivariada.

La base de dades que s'ha simulat conté dos SNPs per a 50 individus, i dos fenotips de tipus continu i binari respectivament. Com es pot observar a les següents taules, les estimacions dels coeficients varien sensiblement en un cas respecte de l'altre.

Haplotypes	snp1	snp2	Freq	Std.error	ICL(95%CI)	ICU(95%CI)
haplo.1	T	G	0,02	0,02	0,00	0,06
haplo.2	C	A	0,03	0,02	0,01	0,08
haplo.4	C	G	0,95	0,03	0,89	0,98
Rares	*	*	0,00	0,00	0,00	0,00

Freqüències haplotípiques

Haplotypes	Coeff	Se.Coeff	ICL(95%CI)	ICU(95%CI)
Intercept	**-0,14***	**0,30**	**-0,75**	**0,44**
haplo.1	**0,30***	**1,82**	**-3,36**	**4,03**
haplo.2	**1,21***	**1,55**	**-1,53**	**4,75**
Intercept	-0,13	0,31	-0,74	0,49
haplo.1	0,26	1,81	-3,36	4,03
haplo.2	1,17	1,49	-1,50	4,43

***Coef Reg. Logística amb prior N((0.1,-0.22),(0.05^2,0.01^2))**

Haplotypes	Coeff	Se.Coeff	ICL(95%CI)	ICU(95%CI)
Intercept	**0,19***	**0,14**	**-0,08**	**0,46**
haplo.1	**1,19***	**0,95**	**-0,68**	**3,12**
haplo.2	**0,94***	**0,67**	**-0,35**	**2,23**
Intercept	0,19	0,14	-0,10	0,46
haplo.1	1,26	0,95	-0,53	3,14
haplo.2	0,95	0,67	-0,32	2,28

***Coef Reg. Lineal amb prior N((1.1,0.8),(0.1^2,0.1^2))**

15

Diferents tractaments de la incertesa haplotípica a l'anàlisi d'associació

En aquesta secció es vol posar de manifest les diferències que es poden obtenir en tractar un mateix problema haplotípic des de diferents punts de vista teòrics pel que fa a la incorporació de la incertesa a l'anàlisi d'associació.

El fet que els genotips d'aquells individus amb dos o més loci heterozigots no tinguin una definició directa dels seus haplotips pot ser tractat de diferents maneres a l'hora de quantificar l'associació entre els haplotips de la mostra i cert fenotip a estudi.

Aquí reproduïm els resultats per una base de dades real, provinent del mateix estudi presentat en l'apartat anterior, ara amb l'objectiu d'analitzar diversos SNPs del gen COX2 en relació al risc de patir Càncer de Còlon. La base de dades està formada per 417 individus, 193 casos i 224 controls, pels quals s'han genotipat vuit SNPs. Per aquest exemple no es mostrarà tot l'estudi d'associació sinó que només ens centrarem en els resultats referents a l'anàlisi d'haplotips, dut a terme mitjançant diferents maneres de tractar la incertesa.

En primer lloc es mostra la taula de freqüències haplotípiques obtinguda en aplicar el programa BayHap a les dades. En la mostra es donen 6 haplotips amb una freqüència superior a 0.01. La resta d'haplotips, que apareixen a la mostra però amb una freqüència menor a 0.01, els englobem en una sola categoria d'haplotips estranys, anomenada "rare". A la Figura 15.1 veiem les freqüències haplotípiques, calculades mitjançant el nostre mètode.

A la segona taula recollim les estimacions dels coeficients del model logístic segons el

```
Haplotypes:
           d401 d926 d1629 d3050 d5209 d8473 d9850 d10335 hap.freq
Haplo.1       0    1     0     1     1     1     0      1   0.4889
Haplo.2       0    1     0     0     1     1     0      1   0.1822
Haplo.3       1    0     0     1     0     0     0      1   0.1470
Haplo.4       0    1     0     1     1     0     0      1   0.1167
Haplo.5       0    0     0     1     0     0     1      0   0.0252
Haplo.6       0    1     1     0     1     1     0      1   0.0155

Pool rare     *    *     *     *     *     *     *      *   0.0246
```

Figura 15.1. Freqüències pels 6 haplotips més freqüents i pels estranys ("rare")

mètode MCMC implementat a BayHap i segons d'altres mètodes, per poder comparar els resultats. Els mètodes considerats han estat:

1. El mètode *naïf* que consisteix en imputar a cada individu la parella d'haplotips més freqüent a la mostra, d'entre les que pot dur. Amb aquest mètode fixem la parella d'haplotips abans de procedir a l'anàlisi d'associació. Si existien d'altres parelles d'haplotips compatibles amb el genotip d'un individu incert, aquestes no seran considerades en l'anàlisi posterior.
2. Regressió Logística amb pesos. Primer s'estimen les freqüències haplotípiques per cada individu, i després es consideren aquestes freqüències com pesos per cada individu dins d'un model de regressió Logística.
3. El mètode Bayesià implementat a BayHap, duent a terme estimació simultània.

A la Figura 15.2 hi trobem les estimacions dels coeficients de la regressió pels tres mètodes considerats.

Com es pot observar les estimacions puntuals difereixen en funció del mètode utilitzat.

	Naïve		EM weighted		MCMC	
	Estimate	Std. Error	Estimate	Std. Error	Estimate	Std. Error
Intercept	-0,207	0,100	-0,270	0,174	-0,209	0,100
01001101	0,128	0,190	0,122	0,191	0,125	0,191
10010001	-0,138	0,208	-0,109	0,205	-0,139	0,209
01011001	0,145	0,226	0,153	0,220	0,166	0,226
00010010	0,900	0,473	0,857	0,468	0,931	0,485
01101101	0,677	0,579	0,634	0,556	0,741	0,621
rare	0,207	0,458	0,225	0,434	0,188	0,461

Figura 15.2. Estimacions i variances de les estimacions pels coeficients del model logístic amb covariables els haplotips.

També difereixen els errors estàndards de les estimacions, essent més grans els retornats pel mètode MCMC.

Part VI

DISCUSSIÓ

16

Funcionament de BayHap respecte de la resta de programes

De programes que combinin l'estimació d'haplotips i l'anàlisi d'associació propi dels estudis d'associació genètica, com fa BayHap, n'hi ha relativament pocs. En aquest apartat ens centrarem en comparar BayHap respecte la resta d'aplicacions existents. Així doncs, contrastarem els resultats que hem obtingut en aplicar el programa sobre escenaris simulats exposats a l'apartat de resultats, així com els que hem obtingut en aplicar-lo sobre dos conjunts de dades reals. També discutirem l'efecte que té sobre els paràmetres el fet de considerar o no l'estimació simultània en les anàlisis. D'aquesta manera destacarem els punts febles i els forts que hem pogut copsar en relació als mètodes de Markov Chain Monte Carlo i a l'estadística Bayesiana en front dels mètodes freqüentistes i dels basats en l'algorisme EM.

Abans d'abordar aquesta comparativa, cal tenir molt present que la comparació entre mètodes resulta francament complexa. De fet, fent una recerca bibliogràfica exhaustiva basada en les fetes al 2005 per Salem et al. [124] i al 2008 per Liu et al. [202], trobem articles en que només es comparen un conjunt petit de programes així com també s'utilitzen una gran varietat d'indicadors per avaluar la precisió general. Aquests fets dificulten la comparació.

Per elaborar aquesta discussió ens hem basat en els propis resultats generats mitjançant la simulació del programa BayHap i dels programes Haplo.Stats i THESIAS (ambdós programes utilitzant l'algorisme EM) i PHASE (Bayesià) sobre mateixos conjunts de dades, així

com també inclourem resultats extrets de la literatura on es realitzen comparacions (amb les limitacions esmentades) amb d'altres programes rellevants pel seu ús dins la comunitat científica com són HAPINFREX, Haplotyper, PL-EM, EM-Decoder i SNPHAP. Tots ells, programes basats en Filogènia, Algorisme EM o Inferència Bayesiana.

16.1 Comparació punt per punt

Tot seguit s'exposa la comparativa entre BayHap i alguns programes representatius de les principals famílies de mètodes que existeixen. Aquesta comparativa la farem avaluant els trets que hem considerat que millor descriuen un programa: analitzarem els mètodes i els algorismes utilitzats, la precisió dels resultats de BayHap respecte la resta pel que fa a l'estimació de les freqüències i els efectes, comprovarem la necessitat d'assumpcions que té cada programa i l'efecte que pot provocar sobre els resultats el fet de violar cadascun dels supòsits. També veurem com pot afectar les anàlisis la incorporació de la història de l'evolució poblacional, l'error de genotipatge i les dades missing. Pel que fa a l'anàlisi d'associació, es compararà les opcions que ofereix BayHap amb les que ofereixen d'altres programes. Finalment, des d'un punt de vista més pràctic, es compararan les característiques del software que repercuteixen pròpiament en la facilitat d'accés i d'ús dels programes.

16.1.1 Mètodes i algorismes

Com ja s'apuntava a ([17],[18]) i els resultats de BayHap que hem pogut observar, podem corroborar que els mètodes Bayesians són adients per l'estimació de freqüències haplotípiques millorat diversos aspectes de l'algorisme EM com ja s'havia indicat prèviament. Alhora, hem pogut observar com aquests mètodes també es mostren eficients per dur a terme l'estimació de l'associació entre els haplotips i un fenotip continu o binari. Com anirem explicant de manera més detallada, el mètode Bayesià ha reportat millors resultats que l'algorisme EM en alguns escenaris concrets. L'algorisme Bayesià també millo-

ra els resultats de l'algorisme de parsimònia, donat que aquests reporten diferents resultats en funció de l'execució, cosa que no passa amb els mètodes Bayesians. Els mètodes Bayesians accepten un major nombre d'SNPs i són més robustos pel que fa a convergència i a cobertura, essent menys dependents de la llavor inicial que l'algorisme EM [17]. A més, com s'ha pogut observar amb els gràfics mostrats a l'apartat de resultats, els mètodes Bayesians retornen més informació que els freqüentistes en donar a més de l'estimació del paràmetre d'interès (OR, freqüències...) les distribucions d'aquests paràmetres, en tant que els considera variables aleatòries. Per això els mètodes Bayesians permeten una execució per part d'usuaris comuns no experts en mètodes de Monte Carlo ni Bayesians, i alhora també són molt oportuns per usuaris avançats, ja que permeten estudiar la convergència de les cadenes generades, així com en cas de ser necessari, permeten modificar i ajustar els paràmetres d'execució per tal de garantir i millorar aquesta convergència. Mentre l'usuari comú pot aturar-se davant d'una cadena no convergida, resultat que també retornarien d'altres programes, amb BayHap l'usuari avançat pot modificar els valors entrants en funció del que observa als gràfics (modificar burnins, lag...) amb l'objectiu d'obtenir el resultat desitjat.

També és molt interessant destacar que mentre d'altres algorismes fallen en l'estimació d'algunes freqüències baixes com es pot observar a la taula 13.7 on Haplo.Stats utilitzant EM no es capaç de convergir, mentre la solució Bayesiana de BayHap resol satisfactòriament el problema. En aquesta mateixa aplicació, ens trobem el cas d'un haplotip no significatiu segons Haplo.Stats i significatiu segons BayHap. Cal tenir present que la convergència d'Haplo.Stats amb aquestes dades no està sent bona en canvi la de BayHap sí que ho és. Amb BayHap tenim la possibilitat d'observar els gràfics i comprovar aquesta convergència.

16.1.2 Precisió

Observant els resultats obtinguts a l'escenari 1 per BayHap i per Haplo.Stats veiem que en relació a les estimacions de les freqüències haplotípiques, el mètode Bayesià i l'Algo-

risme EM en general retornen el mateix estimador puntual i la mateixa desviació estàndard. La mitjana dels biaixos també és igualment nul·la per les estimacions de les freqüències pels dos mètodes. En relació a les desviacions estàndards dels biaixos, BayHap és més estable en la precisió que Haplo.Stats, particularment pel cas dels haplotips menys freqüents. En aquest cas la sd del biaix que retorna Haplo.Stats dobla la de BayHap. Pel que fa als paràmetres del models, en aquest cas els OR's, l'estimació puntual de BayHap és més precisa que la d'Haplo.stats. En particular, la desviació estàndard dels biaixos per l'OR referent a l'haplotip de freqüència més baixa es el doble per Haplo.Stats que per BayHap. Això suggereix que el conjunt de biaixos provocats per l'algorisme EM és més dispers i que per tant l'algorisme EM és menys estable en la precisió dels seus resultats que BayHap. Pel que fa a la cobertura, BayHap presenta en general millor cobertura per les freqüències, especialment pels haplotips de baixa freqüència. Les cobertures pels paràmetres dels models són similars. Aquests resultats ens fan pensar que efectivament, amb mida mostral reduïda, els estimadors asimptòtics que utilitza Haplo.Stats, unit a les limitacions de convergència de l'algorisme EM, fan que aquest darrer mètode funcioni de manera menys precisa que el mètode Bayesià.

Ara bé, si ara comparem els dos mètodes per tamany de mostra superior (escenari 2) observem que no es donen diferències tan clares a favor d'un o altre mètode. Per uns haplotips les estimacions són millors segons l'EM, per uns altres son millors pel Bayesià. Igualment, els valors de les cobertures són en general bons pels dos mètodes.

Observant els resultats sobre dades reals, per la mostra d'esquizofrènia que representaria un cas amb pocs individus passa exactament el mateix. Els haplotips amb baixa freqüència no convergeixen mitjançant Haplo.Stats. En canvi BayHap permet obtenir l'estimació per les freqüències i els efectes. I a més en aquest cas, gràcies a això s'obté un haplotip significatiu. Veiem doncs que la precisió varia en un o altra programa, però que és un fet també lligat a la mida mostral. BayHap funciona de manera més precisa en mostres de mida reduïda.

Pel que fa al funcionament de BayHap respecte la resta de programes, podem recuperar la comparació que es duu a terme en quatre articles entre l'algorisme EM i el Bayesià PHASE, programa que també implementa una Gibbs Sampling. PHASE millora els resultats respecte HAPINFREX i un EM estàndard ([203],[17]). També millora Haplotyper i PL-EM [127]. Per tant, el Bayesià PHASE que segueix el model coalescent, milloraria l'algorisme EM pel que fa a l'estimació de freqüències haplotípiques, coincidint aquest fet amb el que s'ha constatat en aquest treball en comparar la cobertura de BayHap respecte la de l'algorisme EM. Ara bé, alguns articles destaquen que aquest fet es compleix sobre dades simulades i no sobre dades determinades molecularment [18]. En les nostres comparacions sobre dades reals, fixem-nos que tot i haver diferències entre les estimacions puntuals de BayHap i PHASE, els valors de PHASE cauen dins de l'interval de confiança de BayHap per cada haplotip. Això concorda amb les conclusions a que arriben Stephens et al. a [104]. Els autors comparen el funcionament de PHASE respecte d'altres programes sobre els mateixos conjunts de dades i arriben a la conclusió que els programes basats en tècniques Bayesianes, EM o Filogènia tenen un rendiment similar ja sigui en dades simulades o determinades molecularment.

Incertesa

La incertesa és un factor clau en la precisió de les estimacions reportades pels programes. Si aquesta és baixa, la qüestió dels haplotips perd interès donat que la determinació de la parella d'haplotips pels genotips de la mostra es converteix en directa per tots els individus no ambigus. Per tant, tota la teoria existent per estimar haplotips només pren rellevància en bases de dades amb un nombre alt d'individus amb fase haplotípica incerta. Tant en programes basats en EM com en Bayesians, a mida que augmenta el nombre d'individus amb haplotips ambigus disminueix la precisió dels resultats donat que la mostra haplotípica guanya en incertesa. Pel que fa a les simulacions realitzades amb BayHap, hem vist com amb una alta incertesa BayHap estima amb correcció les freqüències haplotípiques inclús si aquestes són petites, i amb bons valors de cobertura. Si comparem els resultats de Bay-

Hap pels escenaris 2 i 3, amb nombre d'individus alt i 8 SNPs, només diferenciant-los la incertesa d'un 22% a un 40%, el programa segueix reportant resultats amb precisió similar. Aquests resultats suggereixen que el programa és robust pel que fa a la incertesa de les dades.

16.1.3 Assumpcions

Pel que fa a les assumpcions anem a descriure quins són els efectes que pot provocar la violació dels diferents supòsits. Les assumpcions sovint estan relacionades les unes amb les altres i pot passar que el fet de violar una dugui a violar-ne una segona. Per a clarificar l'exposició, farem un repàs de cada assumpció una per una.

Equilibri de Hardy-Weinberg

Un gran nombre de programes, juntament amb BayHap, necessiten que les dades segueixin l'equilibri de Hardy-Weinberg com es pot veure a la taula adjuntada a l'apèndix. En particular, tots els programes basats en la funció de versemblança, siguin resolts mitjançant l'algorisme EM o via mètodes Bayesians, assumeixen HWE. S'ha demostrat que la desviació que poden tenir els resultats en cas de no complir-se aquesta assumpció afecta l'estimació de les freqüències, però de manera específica segons com sigui aquesta desviació. En cas que les dades presentin una desviació de HWE deguda a un excés d'homozigosi decreixerà el nombre d'individus ambigus, la qual cosa s'ha demostrat que té petit impacte en la precisió de la majoria de mètodes, incloent EM i Bayesians ([112],[128]). Per contra, com és d'esperar la precisió decreix tant per Bayesians com per programes basats en EM si la desviació de HWE és deguda a un excés d'heterozigosi. HAPINFREX és el que es mostra més vulnerable segons [18].

Desequilibri de lligament

La investigació duta a terme fins el moment suggereix que els segments de cromosoma amb alts nivells de recombinació tendeixen a ser separats en blocs d'haplotips amb molt poca recombinació dins d'ells i un alt desequilibri de lligament. Aquesta estructura de desequilibri

de lligament és habitual al genoma humà ([4],[204],[1]). Un nivell molt alt de recombinacions en un fragment petit del genoma podria violar les assumpcions dels programes basats en el model coalescent ([17],[119]). Malgrat tot, tots els mètodes, Bayesians inclosos i per tant també el programa BayHap, poden presentar problemes a l'hora de construir haplotips en zones amb grans nivells de recombinacions ([18],[111]) i baix desequilibri de lligament [205]. Tot i que ni BayHap ni la majoria de programes no fan assumpcions explícites sobre LD, els resultats dels mètodes basats en EM ([114],[205],[12],[128]) i els Bayesians [17] milloren en augmentar el nivell de LD. En presència de recombinacions, Arlequin s'ha mostrat el més precís [111]. Alguns programes incorporen un test de LD per tal d'identificar els blocs d'haplotips [206].

Així doncs, l'avaluació del LD i de les recombinacions és un pas rellevant a l'hora de dur a terme una anàlisi d'haplotips. Dades que continguin recombinació seran un repte pels programes que no considerin recombinació. El decrement en LD s'associa amb un increment de l'error en les estimacions [205] i magnifica els efectes de l'error de genotipatge [207]. Encara que deduir els haplotips en zones amb baix LD és important, les estimacions haplotípiques per aquest tipus de dades poden ser poc fiables. Com és d'esperar, les recombinacions porten a un increment en el nombre d'haplotips, incloent haplotips de baixa freqüència que són difícils d'estimar amb precisió. En aquest sentit, BayHap persenta un punt fort amb aquest tipus de dades donada la seva propietat de poder estimar haplotips amb baixa freqüència i per tant, pot reportar millors resultats en aquest escenari que d'altres programes, com ja s'ha vist a la secció de resultats en relació a haplotips poc freqüents. A més, el programa BayHap s'ha concebut per ser executat en l'entorn R on ja existeixen diversos paquets que estimen LD i recombinacions i també per a ser en un futur executat mitjançant l'aplicació via web SNPstats, que ja incorpora el test de LD i de recombinacions. En cas que el nivell de LD sigui molt baix, pot ser recomanable augmentar la mida de la mostra d'individus per millorar la precisió en presència d'alta recombinació. Analitzar el segment del cromosoma a cada banda dels punts de recombinació sembla ser la opció més

viable [208].

Història de l'evolució poblacional

Diversos programes necessiten partir de certes assumpcions sobre la història evolutiva de la població de la qual s'extreu la mostra. Aquest supòsit té per objectiu millorar l'eficiència del programa i simplificar l'anàlisi d'haplotips. El programa PHASE per exemple, incorpora un model de coalescència. D'altres programes es basen en variants d'aquest model o bé es basen en el concepte de perfecta o imperfecta filogènia. El benefici d'incorporar un model evolutiu com aquests és que l'algorisme treu avantatge del fet que existeixin similituds entre haplotips. Es considera que s'obté millors estimacions que amb d'altres mètodes ([17],[104]). La desavantatge és que el comportament dels al·lels en un plaç curt d'evolució cromosòmica pot violar el model induint a errors. En contrast, d'altres programes com Haplotyper, HAPINFREX, Hapar, no imposen història evolutiva. La precisió d'aquests programes es pot veure afectada en conjunts de dades que ajustin algun dels models i no s'estigui considerant per part del programa. Així com si les dades no ajusten a un model concret, aquests programes ajusten millor que els que suposen el model [17]. Per exemple, quan les dades violen el model coalescent, la resta de programes que no el suposen funcionen millor que PHASE que sí el suposa [111]. Tot i així, la tria del model dependrà del tipus de dades. El model coalescent sembla adient per poblacions estables que hagin evolucionat durant llargs períodes de temps, però és menys adequat per poblacions amb fluix de gens, estratificació i/o emigració. Tot i així existeix discussió sobre aquesta qüestió ([203],[17],[18]).

BayHap no basa les seves estimacions en cap model concret. Aquest tipus de programa s'han de fer servir amb cura, ja que desviacions del model poden tenir un impacte molt important en la precisió de les estimacions haplotípiques, i donada la manca de coneixement que en moltes ocasions es té sobre el model evolutiu de la població amb la que es treballa, sembla preferible triar programes que no basin les seves estimacions en cap model concret, a no ser que es compti amb aquesta informació.

Error de genotipatge i dades amb valors *missing*

L'error de genotipatge és una forma d'error de classificació que pot portar a efectes perjudicials en les anàlisis d'associació, en les mesures de LD i de recombinacions ([209],[210]) i que per tant pot dur a anàlisis haplotípiques errònies ([111],[207],[131],[211],[212]). El poder dels estudis d'associació amb SNPs decreix inclús amb errors de genotipatge de magnitud relativament petita. Una tendència similar la trobem en els estudis d'associació entre fenotips i haplotips. Els requeriments pel que fa al nombre d'individus a analitzar en funció dels errors de genotipatge als SNPs es poden trobar explicitats al lloc web PAWE (*Power Association With Error*) ([213],[170]).

La majoria d'errors de genotipatge són deguts a la pèrdua d'SNPs, donant lloc a un problema de tractament de dades missing. Aquests errors també acostumen a ser deguts a la dificultat de genotipatge que presenten els genotips heterozigots. Aquesta dificultat duu a una infrarepresentació a la mostra de genotips heterozigots i per tant a un biaix a favor de l'increment de la proporció de genotips homozigots ([170],[214]).

Els programes que accepten dades amb missings sovint assumeixen que els missings es troben repartits aleatòriament. BayHap, haplo.stats i THESIAS fan aquesta assumpció. S'ha de tenir present que alguns haplotips falsos poden ser introduïts a la mostra per aquest sistema de considerar tots el al·lels possibles pels loci faltants [111],. Aquest error de genotipatge i aquesta falsa assumpció d'igualtat d'oportunitats pels diferents al·lels que poden ser atribuïts a un locus no informat, pot dur a una pèrdua de precisió, particularment quan el LD és baix i existeixen alguns haplotips rars ([211],[215]). Una estratègia comú és genotipar dos cops un subconjunt de la població a estudi per determinar el grau d'error. A l'hora d'estimar l'associació, la precisió i el poder d'aquests anàlisis poden ser millorats incorporant la incertesa del genotipatge en la inferència haplotípica per evitar els efectes d'aquests errors de genotipatge, com es descriu a [170]. En aquest sentit, BayHap es troba en aquest conjunt de programes havent considerat l'estimació simultània dels efectes referents a tots els haplotips compatibles amb cada genotip, augmentant aquest conjunt d'haplotips

en totes les possibilitats que es poden donar en cas que existeixin valors missings. Aquesta estratègia duta a terme per BayHap té per objectiu reduir l'efecte d'aquests errors respecte altres programes. Cal tenir present que la majoria de programes no accepten dades amb missings com es pot veure a la taula sobre mètodes haplotípics de l'apèndix. Es tracta de programes que en la seva majoria exclouen de les anàlisis els individus pels que falta alguna dada. Aquests programes poden donar lloc a un efecte de desviació cap a la homozigosi a la mostra genotípica.

Cal tenir present alhora, que acceptar dades amb missings comporta una pèrdua d'efectivitat computacional rellevant. En bases de dades on faltin dades, augmenta el temps d'execució, augmenten els requeriments de memòria i s'incrementa la incertesa. S'han proposat diverses estratègies per intentar posar solució a aquesta qüestió. L'algorisme EM es pot adaptar per tal que accepti dades amb missings [216]. En l'entorn dels mètodes Bayesians, PHASE accepta dades incomplertes fent una imputació aleatòria [104]). Haplotyper també s'ha demostrat estable en presència de dades missing, tot i que cal anar amb cura [18]. BayHap ha estat programat sota els mateixos criteris que Haplo.stats i com s'ha pogut observar a les aplicacions als conjunts de dades reals que contenien dades mancants, funciona de manera similar a PHASE i a Haplo.Stats.

Es poden trobar discussions molt complertes i interessants sobre el tractament de dades missing i l'anàlisi d'haplotips com per exemple ([18],[103]). La inclusió d'individus amb gran quantitat de dades missing ($> 10\%$) pot tenir un efecte negatiu en la reconstrucció de la fase dels individus que no presenten missings. Finalment, marcadors que no compleixin els patrons aleatoris d'error de genotipatge haurien de ser exclosos del conjunt de genotips a estudi.

16.1.4 Nombre i tipus de marcadors

La majoria dels programes d'anàlisi haplotípica que existeixen estan limitats a l'ús de locus bial·lèlics. BayHap no n'és una excepció i aquesta primera versió del programa també presenta aquesta limitació. Això és degut a que els programes que accepten locus multial·lèlics sovint presenten temps d'execució molt elevats i aquest fet els converteix en programes poc òptims a nivell pràctic. Alguns programes també presenten limitacions pel que fa al nombre de loci com es pot observar a la taula de l'apèndix. BayHap no estipula un nombre de loci màxim, donat que aquest nombre va lligat a d'altres factors també influents com la mida mostral o el nombre de covariables d'ajust, i si es consideren termes d'interacció o no. Si analitzem les execucions exposades a l'apartat de resultats, observant l'escenari 2 en que es treballa amb 8 SNPs podem veure que els biaixos per les freqüències calculades amb el programa Bayesià són més petits que els obtinguts amb l'algorisme EM. Tot i així en magnitud aquests biaixos són molt petits, i creiem que la diferència en nombre d'SNPs considerada no permet establir grans diferències entre el mètode Bayesià i l'Algorisme EM. Segons la literatura, els programes basats en l'algorisme EM a la pràctica tenen un límit de 25 loci, degut a requeriments de memòria de processador i a mala convergència ([12],[112],[17]). HAPINFREX no té cap límit pràctic, tot i que en l'inici el programa podria fallar si es parteix d'un nombre de marcadors molt gran [11]. L'altre programa basat en parsimònia HAPAR supera HAPINFREX i la seva precisió millora també en augmentar la mida mostral. Per la seva banda, l'estratègia divide and conquer programada al software PL-EM també és efectiva a l'hora de tractar amb grans nombres de marcadors [18]. Esquemes similars s'han implementat també en programes bayesians ([18],[104],[111],[121]). Recentment, dos mètodes han millorat la pèrdua de poder que l'augmentar el nombre de marcadors, provoca en les estimacions. Aquests mètodes que inclouen a la regressió la distància entre locus podrien disminuir la pèrdua ([217],[218])

Pel que fa a la quantitat d'SNPs heterozigots, quant més baixa sigui la quantitat de mar-

cadors d'aquest tipus, més acurada serà la precisió donat el decrement d'incertesa en les dades. Les simulacions de BayHap i Haplo.Stats s'han dut a terme en escenaris amb alta incertesa (de 20 al 40%) perquè s'ha considerat que aquests eren escenaris interessants per comparar la precisió de les execucions donat que escenaris amb baixa incertesa no proposen cap repte afegit al d'un anàlisi de variables categòriques habitual. Els resultats mostren que tot i en escenaris incerts, BayHap ha recuperat correctament les freqüències haplotípiques que havien estat simulades amb gairebé biaix nul i una cobertura molt correcte. Haplo.Stats també estima les freqüències sense biaix, però l'interval de confiança que reporta té una cobertura pitjor que BayHap.

16.1.5 Mida de la mostra

Tant el nombre de loci com el nombre d'individus que conformarà la mostra de genotips són components influents en l'execució dels programes d'anàlisi d'haplotips. A la taula de mètodes haplotípics de l'apèndix es poden consultar els detalls sobre el límit de mida mostral que accepten els diferents softwares. Així com la mida de la mostra creix, el temps d'execució dels programes augmenta. La precisió dels programes basats en l'algorisme EM també augmenta a mida que s'incrementa la quantitat d'individus ([219],[20]). De la mateixa manera, la precisió de HAPAR, Haplotyper i PHASE, programes Bayesians, també millora en aquest cas [96] així com també millora l'estimació de les freqüències baixes [115].

Les simulacions dutes a terme amb el programa BayHap suggereixen que aquest és un programa que reporta estimacions de freqüències vàlides tant en mostres de mida petita (200 individus) com gran (1000 individus), tant per haplotips més freqüents com poc freqüents. Aquest resultat és similar a l'aconseguit amb l'algorisme EM, tot i que com ja s'ha esmentat, tot i tenir mida mostral més gran Haplo.Stats reporta cobertures inferiors a l'esperat per les estimacions de les freqüències d'alguns haplotips poc freqüents i en alguns casos per aquests haplotips pot arribar a fallar la convergència. Pel que fa a l'estimació dels coeficients, els

biaixos són superiors pels efectes associats a haplotips poc freqüents i a mida que el nombre d'individus a la mostra augmenta, la precisió de l'estimació d'aquests efectes també millora, tant en BayHap com per Haplo.Stats.

16.1.6 Característiques del Software

En aquest punt es discuteixen diverses qüestions relacionades amb l'ús dels diferents programes. L'accessibilitat i el fàcil maneig són qüestions molt rellevants a l'hora de triar un programa. Així com els requeriments computacionals que tingui el software, que també determinaran la necessitat de màquina per poder-lo executar, el format en que s'hagin d'introduir les dades, la interfície d'accés al programa, el format dels resultats que retorna cada programa i el temps d'execució del programa.

Requeriments computacionals

La columna anomenada *platform* de la taula de mètodes haplotípics de l'apèndix mostra els requeriments de sistema operatiu de cada programa. Com es pot observar, no tots els programes estan disponibles per diferents sistemes operatius. Aquest és un tema molt rellevant donat que la selecció d'un programa en concret pot requerir una inversió en un nou equip informàtic i incomoditats diverses per l'usuari. Per un usuari de windows pot resultar poc pràctica la tria d'un software que s'executi en Linux. Pel que fa a aquesta qüestió, BayHap es pot executar tant en windows com en linux donat que l'entorn R existeix pels dos sistemes operatius i el programa s'ha compilat per funcionar en ambdós sistemes.

Format de les dades

Desafortunadament no hi ha un format estàndard per les dades genotípiques i les variables fenotip. Manipular les dades d'un format a un altre pot resultar incòmode, dificultós i farragós. HIT i HAPLOSCOPE són plataformes de programes que incorporen diversos programes d'anàlisi d'haplotips en una mateixa interfície. BayHap també pretén facilitar el seu ús en aquest sentit i per això el format de dades és molt similar a l'utilitzat per d'altres programes com Haplo.Stats i THESIAS.

Interfície

La interfície és de nou una component bàsica en relació a l'ús dels programes. La tria d'un programa dependrà en forta mesura de com de fàcil i ràpid li sigui a l'usuari accedir a l'aplicació i entendre com funciona el programa en sí. La majoria de programes s'executen a través de comandes de prompt, una interfície poc amicable i que tendeix a intimidar els usuaris novells o poc experts en qüestions informàtiques. Afortunadament, existeixen programes que tenen interfície gràfica com Arlequin, Haploview, Haploscope, Hplus o THESIAS. BayHap, igual que d'altres llibreries per anàlisi genètic pertany a l'entorn estadístic R, d'accés lliure i molt present entre aquells que practiquen recerca biomèdica. L'ús de BayHap serà especialment fàcil per usuaris d'R i d'S-PLUS.

La majoria de programes són força hermètics pel que fa als valors dels arguments que utilitzen. Com per exemple THESIAS que no permet que l'usuari accedeixi ni modifiqui cap dels valors d'execució. A BayHap s'ofereixen un seguit de valors per defecte que han de funcionar per la majoria d'ocasions. En cas que no sigui així, l'usuari els pot modificar segons convingui.

Valors de sortida

A més de les estimacions de les freqüències haplotípiques, molts programes també retornen mesures que avaluen la bondat d'ajust dels haplotips construïts. Alguns programes basats en l'algorisme EM com ara el Genecounting, HPLUS, Haplo.Stats, LD-SUPPORT, MLOCUS, el PL-EM o el SNPHAP, ofereixen les probabilitats posteriors de les assignacions haplotípiques. Les probabilitats posteriors són útils per l'avaluació de les assignacions haplotípiques, ja que en la reconstrucció de la mostra les estimacions de les freqüències es poden fer servir com a pesos per cada haplotip ([147],[68]). Alguns programes retornen clarament les variàncies per les freqüències haplotípiques estimades (HAPLO, HPLUS i PL-EM). Haplo.Stats les retorna però no és immediat accedir-hi, no es mostren de manera senzilla amb el gruix de resultats. BayHap retorna les probabilitats a posteriori per la freqüència de cada haplotip, estimacions puntuals i variància per les freqüències, així com

l'interval de confiança. Alhora, BayHap també afegeix com a resultat aquesta distribució de probabilitat i estimadors puntuals amb interval de confiança per les estimacions dels efectes associats a cada haplotip. També permet generar gràfics de sortida per avaluar la convergència, les característiques de les cadenes i les distribucions a posteriori per cada paràmetre en el model.

Pel que fa a la sortida, també és molt rellevant el format en que els diferents programes entreguen els resultats. És de valorar que siguin fàcilment exportables i manipulables, en format de taula. Haplo.stats retorna els resultats com a un objecte dins d'aquest entorn. La versió de THESIAS amb interfície en Java retorna els resultats en una pàgina html amb els resultats incrustats i de difícil exportació. Haplotyper, entre d'altres, retorna els resultats en un arxiu de text, així com EM-DeCODER té una sortida en java. Com es pot observar hi ha varietat de formats en la sortida. Els resultats de BayHap són com els d'Haplo.Stats i es guarden en un objecte dins l'entorn R. El paquet compta amb funcions que retornen taules amb els valors principals i els gràfics són fàcilment exportables.

Temps d'execució

El temps d'execució va estretament lligat a la complexitat del problema haplotípic, que empitjora amb el nombre d'SNPs considerat ([12],[17]). Tot i que l'algorisme EM teòricament pot funcionar amb un nombre infinit de loci polimòrfics, a la pràctica es veu limitat per l'increment exponencial que l'augment d'SNPs suposa a nivell de requeriment de memòria ([12],[112]). Més encara, l'algorisme EM necessita diverses execucions amb diverses llavors per evitar la convergència local i això incrementa el temps que es requereix per inferir haplotips [12].Tot i que com ja hem dit utilitzar Gibbs Sampling, com fa BayHap i PHASE, comporta una determinació de la fase haplotípica més eficient que la reportada per l'algorisme EM i reconstrueix un nombre de marcadors superiors, les execucions són més lentes donat que es tracta d'algorismes no paral·lelitzables ([17],[119]). PHASE reconegut com un dels més utilitzats compta amb unes execucions molt lentes ([17],[116],[119],[111]). Per exemple, si comparem programes Bayesians amb 50 individus i de 14 a 119 SNPs, Haplo-

typer estima els haplotips en segons, Arlequin en minuts i PHASE en hores [111]. Tot i que PHASE també presenta la versió fast PHASE conscients que aquesta és una feblesa important d'aquest programa, els programes que modifiquen l'algorisme EM com el SNPHAP, el PL-EM o l'implementat a Haplo.Stats tenen menor temps d'execució que PHASE per conjunts de dades grans [104]. Els programes basats en metodologia filogènica es mostren més ràpids que la resta en diversos escenaris [107]. El temps d'execució augmentarà en presència de dades missing i de marcadors multial·lèlics ([15],[111],[121]).

Pel que fa al temps d'execució de BayHap, es presenta sensible a la mida mostral, i molt especialment als factors que determinen el nombre d'elements pels que haurà de circular la cadena de Markov: el nombre d'haplotips possibles a la mostra que ve determinat en gran mesura pel nombre d'SNPs heterozigots i el nombre de covariables d'ajust i termes d'interacció.

Accés

La majoria dels programes que hem anomenat són programari lliure, d'ús gratuït per interessos no comercials, així com també ho és BayHap. Alguns d'ells són d'ús públic però necessiten previ registre d'usuari.

16.1.7 Anàlisi d'associació

Com ja hem pogut veure en d'altres punts d'aquest treball, l'estimació de les freqüències haplotípiques no acostuma a ser l'objectiu final d'un estudi. Habitualment, l'estimació de freqüències haplotípiques s'emmarca dins d'estudis d'associació genètica que es duen a terme just després d'aquesta determinació haplotípica. De programes que combinen estimació d'haplotips i l'anàlisi d'associació propi dels estudis d'associació genètica n'hi ha relativament pocs. Es pot observar la llista dels programes existents a 3.2 o bé a la llista ampliada a l'apèndix.

Cal partir de la idea que tots els mètodes de reconstrucció de la mostra haplotípica assignen els haplotips amb cert error ([141],[103],[142]) degut a la incertesa que presenten

alguns haplotips. Aquesta incertesa no pot ser ignorada en les anàlisis posteriors, donat que això podria dur a a estimacions esbiaixades dels paràmetres i a sobreestimar el nombre de resultats fals-positius ([147],[130],[68],[142]). Per tal de no ignorar aquesta incertesa, BayHap implementa l' estimació simultània de freqüències haplotípiques i efectes associats a un fenotip binari o continu, segons models de regressió Logística, de regressió Lineal i de regressió de Weibull. A l'apartat de resultats a la taula 15.2 hem comparat l'estimació simultània respecte el mètode d'imputació i el de regressió Logística amb pesos. Hem observat com efectivament l'interval de confiança que retorna Bayhap és més ampli, degut a la incorporació de la incertesa. Aquesta diferència pot ser molt rellevant, donat que diferents programes poden donar associacions significatives o no significatives pel mateix haplotip. Aquest fet s'ha donat al comparar els resultats retornats per PHASE amb imputació fixa d'haplotips en l'estudi de CCR (13.11). En aquest cas, tot i eixamplar-se l'interval de confiança, no s'ha perdut la significació per aquest fet. Pel cas de l'haplotip CATCCAT que per PHASE es queda molt a prop de la significació estadística, pel cas de BayHap es queda més lluny donat que l'IC s'amplia.

Com es pot veure a l'apartat de resultats, les simulacions realitzades amb BayHap demostren que el programa recupera efectivament els valors simulats pels tres models estadístics: el Logístic, el Lineal i el de Weibull. Per tant, les cadenes estan convergint als valors teòrics poblacionals que toca. En general les cobertures són bones, tot i ser un pèl inferiors per haplotips poc freqüents, però encara acceptables.

Pel que fa a les aplicacions sobre bases de dades reals, les conclusions a les que s'arriben en un i altre anàlisi varien lleugerament. En l'anàlisi d'esquizofrènia com s'ha vist, el fet de poder inferir efectes per haplotips de freqüència menor ha descobert un haplotips significatiu. Pel cas de CCR, fixem-nos també que en la taula 13.12 on es mostra l'anàlisi de supervivència, els resultats de BayHap i THESIAS varien, especialment pel cas d'haplotips amb freqüències petites. Aquest cas de supervivència presenta especial interès donat que existeixen pocs programes que realitzin aquesta anàlisi. En aquest cas es genera certa incertesa

sobre quin resultat és ”correcte´´ o millor dit, és ”més correcte´´. La diferència en els límits dels Intervals de confiança segurament siguin degudes a les diferències dels mètodes d'estimació. Consultant la convergència de BayHap podem dir que és bona segons els gràfics, la de THESIAS no la podem comprovar perquè no ofereix aquesta opció.

Pel que fa als models d'herència, BayHap permet la tria del model d'herència més adient (additiu, dominant o recessiu) essent l'únic software a l'actualitat que permet executar per exemple un anàlisi de supervivència, amb un model d'herència recessiu, amb ajust per covariables i interaccions. També en aquest sentit és l'únic software que permet tenir resultats gràfics per aquestes estimacions.

El programa HAP basat en Filogènia imperfecta, s'ha demostrat precís a l'hora d'assignar haplotips a la mostra de genotips [121]. Aquest programa duu a terme l'anàlisi d'associació amb fenotips discrets i continus, tot i que el perill de biaix existeix degut a la incertesa de l'assignació haplotípica. BayHap per la seva banda no retorna una mostra d'haplotips reconstruïts donat que això topa amb la filosofia del mètode: no fixar la parella d'haplotips pels individus incerts, i permetre que la mostra variï en funció de les freqüències haplotípiques estimades a cada pas de la cadena. Tot i així, si l'usuari ho desitja, sempre pot assumir com a pesos les freqüències que BayHap estima i imputar els haplotips segons aquestes freqüències. Seguint un criteri similar, diversos programes eviten la imputació d'haplotips comparant directament les freqüències entre dos grups ([148],[220]) en el disseny cas-control. Entre aquests es troben EH, EHPLUS, Genecounting, PHASE, el mòdul de SAS genetics i el SNPEM. Fallin et al. [132] demostren les avantatges d'aquest enfocament utilitzant aquest darrer software. Aquesta metodologia, però, no accepta ajust per covariables. Hi ha programes com el de Zaykin [68] que utilitza el *Likelihood ratio test* per testar l'associació entre haplotips i fenotips. Haplo.stats ([151],[150]) i THESIAS [118] són programes basats en l'algorisme EM que també inclouen tests sobre les interaccions amb covariables utilitzant models de regressió però amb els inconvenients de l'algorisme EM.

Discussions addicionals sobre tests d'associació amb haplotips es poden trobar a ([68],[151],[149],[156],[157],[153],[154]).

Els resultats obtinguts en aplicar BayHap sobre bases de dades reals i simulades ens han demostrat que el programa és vàlid a l'hora d'estimar freqüències haplotípiques i l'associació entre els haplotips i un fenotip continu o binari. Tal i com es mostra a l'apartat de Resultats, les execucions realitzades amb BayHap, Haplo.Stats, THESIAS i PHASE, ens suggereixen que el programa BayHap és una eina útil en aquest camp, aportant millores en l'anàlisi d'haplotips. Particularment BayHap funciona millor en les anàlisis de mostres de mida reduïda i en l'estimació de freqüències haplotípiques petites, tant en l'estimació d'aquestes freqüències com en l'estimació dels efectes associats a aquests haplotips poc freqüents. En aquest sentit, un punt a destacar és que BayHap ofereix una via més àmplia d'avaluació dels resultats retornant un gruix d'informació superior al retornat per d'altres programes. A més BayHap permet analitzar associació i supervivència amb ajust de covariables, interaccions i diferents models d'herència utilitzant estadística Bayesiana en l'entorn estadístic R.

16.2 Inferència Bayesiana vs Freqüentista

Com hem introduït a la secció 7.1, la idea fonamental del Teorema de Bayes (7.1) és la de modificar la creença a priori que podríem tenir sobre certs paràmetres abans de veure cap dada mitjançant les dades que s'han observat. D'aquí sorgeix la principal crítica dels freqüentistes: basar l'anàlisi en unes creences subjectives de l'investigador i fer que el resultat depengui de manera crucial d'aquestes creences sembla poc rigorós. Malgrat tot, quan

veritablement hi ha creences fortes i consensuades sobre determinats paràmetres, com per exemple el coneixement que cert coeficient prengui valors negatius, perquè no fer-les explícites i transparents a través de l'anàlisi Bayesià? D'altra banda, sabem que si tenim una mostra suficientment gran (aquesta mida mostral dependrà de la complexitat del problema analitzat) la creença a priori de l'investigador es veu dominada per les dades, i la seva influència al resultat final disminueix fins a fer-se inexistent per una mostra amb infinites observacions. També sabem que una gran quantitat de resultats freqüentistes poden obtenir-se des d'una perspectiva Bayesiana tot i que la filosofia subjacent sigui diferent. Per exemple, el mínims quadrats ordinaris és un estimador freqüentista que coincideix exactament amb la mitjana de la distribució Bayesiana sota unes creences concretes al marc del model lineal.

Tenint en compte la connexió entre ambdós enfocaments a la pràctica i també la possibilitat d'obtenir els mateixos resultats sota tots dos tractaments, els Bayesians argumenten que la interpretació del problema sota el seu enfocament sempre és més intuïtiva i natural. Recordem que un Bayesià proporcionarà conclusions del tipus: hi ha un 95% de probabilitat que el paràmetre estigui entre 0.3 i 0.8. En canvi un freqüentista afirmaria: si generem 100 mostres aleatòries de la mateixa mida i repetim l'estimació 100 vegades, en 95 d'elles el paràmetre estimat es trobarà entre 0.3 i 0.8.

Així doncs, els contrastos d'hipòtesis semblen més naturals al marc Bayesià. Un Bayesià convençut no calcularà mai un p-valor, donat que l'únic que necessitarà per contrastar hipòtesis és tenir la distribució a posteriori dels paràmetres. En la gran majoria d'ocasions els freqüentistes basen el seu contrast en l'anàlisi asimptòtic, és a dir, en calcular p-valors de la distribució asimptòtica de l'estimador (no del paràmetre veritable, que és un valor fix). Aquesta distribució, majoritàriament normal gràcies a nombrosos teoremes centrals del límit, és la que l'estimador tindria si l'investigador tingués moltes més dades de les que, en la majoria de casos, veritablement té. Ens podem plantejar si aquesta és una bona manera de fer inferència en general. Contràriament, el Bayesià es basa en la distribució dels

paràmetres donades les seves dades, que poden en principi tenir qualsevol forma no gaussiana.

Com s'ha exemplificat al capítol 14 BayHap permet dur a terme anàlisis Bayesianes o freqüentistes, mitjançant la distribució prior que es triï, obtenint resultats diferenciats tant pel que fa als propis valors numèrics com a la interpretació d'aquests. Cal destacar que les distribucions priori proposades pel programa són de caire conservador.

17

Consideracions Finals d'aquesta Tesi Doctoral

L'anàlisi haplotípica és una part bàsica i molt prometedora en l'estudi de la base genètica que presenten algunes malalties complexes. Es tracta d'un camp en constant evolució i estudi com demostra el gran nombre de publicacions que se li han dedicat al llarg de 20 anys i que se li segueixen dedicant a l'actualitat ([86],[145],[144],[63],[148],[161],[212]). Malgrat tot aquest esforç, l'eficiència d'utilitzar haplotips en relació a utilitzar marcadors individuals no és sempre clara. De la revisió duta a terme, concloem que cap mètode és superior als altres pel que fa a precisió dels resultats. La majoria de programes comparteixen diverses similituds, però també presenten diferències substancials que en general van lligades a característiques concretes de la població. Podríem dir que cada programa presenta la seva combinació "única"de punts forts i de limitacions. Seria desitjable que els investigadors interessats en l'anàlisi haplotípica consultessin les diferents i complertes revisions que existeixen ([133][221],[124],[202]) i triessin el mètode haplotípic que millor s'adeqüi a les característiques de les seves dades i als interessos del seu anàlisi. Aquest criteri de selecció s'hauria de fer en funció de quins són els objectius de la recerca, de les hipòtesis que es pretenen testar, de les assumpcions que les dades compleixen, dels errors de genotipatge, de la presència de *missings* a les dades i de l'experiència informàtica a l'hora d'executar programes. Perquè, al cap i a la fi, un bon programa d'anàlisi haplotípica és aquell que reporta els resultats desitjats pel que fa a les freqüències haplotípiques i a l'anàlisi d'associació. Els programes són més o menys eficients bàsicament en funció del compliment de les assump-

cions. S'ha vist que desviacions de les assumpcions porten en general a una pitjor qualitat dels resultats. Per tant, un pas primordial per dur a terme un bon anàlisi haplotípic és la comprovació de les assumpcions. Tot i així, el tractament Bayesià dels paràmetres com hem vist en aquesta tesi doctoral aporta avantatges respecte el punt de vista freqüentista, permetent la inclusió de coneixement a priori, basant la inferència en la distribució dels propis paràmetres i no en distribucions asimptòtiques i oferint a més una interpretació dels intervals de probabilitat més intuïtiva que la dels intervals de confiança. Aquests són alguns punts fort de l'estadística Bayesiana envers la freqüentista.

La selecció d'un programa també es basarà en la facilitat del seu ús. L'avaluació d'aquest criteri és complexa, i es basa en subcriteris més específics que han estat discutits en el capítol anterior. Els programes basats en entorns gràfics coneguts o bé que ofereixin execució via web, semblen ser els més senzills i còmodes d'utilitzar. Desafortunadament, en aquest sentit només un reduït grup de programes satisfan les necessitats dels investigadors. La utilització d'un o altre programa també dependrà fortament de l'experiència informàtica de l'investigador. En resum, la tria del programa s'hauria de basar en identificar les necessitats particulars de la recerca i triar aquell que millor les resolgui, sense oblidar-se del compliment de les assumpcions i de les limitacions de cada mètode.

La majoria de programes són revisats, mantinguts i actualitzats regularment. L'anàlisi d'haplotips és un camp de ràpida evolució, amb força activitat i en què apareixen nous programes i mètodes amb prou rapidesa. De fet, el nombre de mètodes i programes d'anàlisi haplotípica han augmentat en nombre i han millorat amb molta rapidesa durant la darrera dècada. Tot i així, el conjunt de programes presenta algunes qüestions que encara queden per millorar o resoldre, com l'estimació de les freqüències baixes, o el tractament de dades amb missings, qüestions a les que BayHap aconsegueix donar una millor solució. BayHap també s'ha centrat en millorar la qüestió relacionada amb les dades missing, proporcionant una eina que accepta aquest tipus de dades. La filosofia global de BayHap ha estat la d'oferir una eina més complerta que d'altres, acceptant diversos fenotips i poden realitzar

una bona avaluació dels resultats, tot amb la mateixa aplicació. Tot i així, queden d'altres qüestions obertes que han quedat fora dels objectius plantejats en aquesta tesi per BayHap. Futures versions d'aquest i d'altres programes hauran de resoldre encara diversos temes, com ara avaluar l'efecte sobre les estimacions dels diferents nivells de LD o avaluar més minuciosament els efectes del no compliment de les assumpcions. Idealment, estudis futurs haurien de comparar un conjunt encara més gran de programes entre els més utilitzats, aplicats sobre mateixos escenaris per avaluar els més eficients. Més enllà d'això, seria molt còmode per l'ús dels programes l'instaurar un format estàndard de dades que fos vàlid per totes les aplicacions existents. Aquestes serien algunes de les qüestions cap a on encaminar futures investigacions en el camp dels haplotips.

18

Limitacions

A continuació es llisten un seguit de qüestions que es consideren limitacions del mètode d'anàlisi d'haplotips que s'ha desenvolupat en aquesta tesi:

Accés a l'entorn R

Tot i les clares avantatges que ofereix l'entorn R, per usuaris no habituats a treballar-hi a l'inici pot resultar poc amigable, i per tant l'ús d'aquest paquet pot quedar reduït a aquells usuaris que habitualment hi treballin. Per tal de fer l'ús del paquet extensible a d'altres usuaris, la següent fase del projecte on s'engloba el desenvolupament d'aquest programa preveu l'execució de l'aplicació via web, inclosa a la plataforma SNPstats [169] desenvolupada pel mateix grup de recerca i que de moment utilitza les funcions del paquet Haplo.Stats. L'aplicació via web serà una bona opció que permetrà executar-lo sense necessitat de recórrer a l'execució directa del paquet, ni exigirà tenir coneixements d'R. SNPstats és una interfície molt senzilla d'utilitzar, en què l'usuari només ha d'introduir les dades i marcar les anàlisis que desitja realitzar.

Informació retornada per BayHap

Durant aquesta tesi s'ha destacat com quelcom positiu el fet que BayHap ofereixi un gruix d'informació de resultat superior al retornat per d'altres programes. Però aquest fet pot ser poc útil per persones no coneixedores de els tècniques que aquí s'apliquen. Versions futures del programa intentaran que l'aplicació sigui més autònoma respecte alguns paràmetres,

però conservant la filosofia primària de permetre a l'usuari modificar els valors si així ho creu necessari.

Haplotips de baixa freqüència

Com s'ha destacat, BayHap és eficient a l'hora de resoldre l'anàlisi d'haplotips de baixa freqüència. El paper d'aquests haplotips als estudis d'associació és discutit. En mostres petites podem tenir molts pocs individus que els duguin i pot ser complicat extreure conclusions pel que fa a associacions donada una baixa potència. En mostres més grans l'estimació d'aquests haplotips està més justificada.

Execucions fallides

El programa té un ratio de fallida de sobre un 0.7% que té a veure amb dades tals que els valors fixats d'amplada d'interval de l'Slice Sampling no permet avançar i convergir en un temps realista per l'usuari. El programa és forçat a acabar i ofereix un missatge d'error. Aquest és un valor intern que l'usuari no pot modificar.

Weibull i no Cox

El programa realitza una anàlisi de supervivència mitjançant un mètode paramètric, mentre que un model de Cox semiparamètric podria ser adient per un conjunt de casos més ampli. Tot i així, encara que el model de Cox i el de Weibull són força diferents pel que fa a formulació matemàtica i assumpcions, ambdós s'han mostrat similars a l'hora de produir resultats en un ampli ventall de situacions [118].

Burnin

Els burnins oferts com a argument al paquet, en són dos, un vàlid per a totes les freqüències i un altre pels paràmetres del model estadístic. Podria ser que cada paràmetre necessites un burnin diferent i que el programa obligui a circular per tots els paràmetres el valor màxim per tal que totes les cadenes convergeixin bé. Igual com el nombre d'iteracions també es tria per freqüències i coeficients, però en grup. Aquest fet pot fer augmentar el temps d'execució del programa.

Clustering

La versió actual de BayHap no inclou la possibilitat de fer una anàlisi amb *clusters* d'individus. Això ha estat triat així conseqüentment amb un dels objectius d'aquest treball, l'estimació d'haplotips de baixa freqüència i els seus efectes. En cas que l'usuari tingui altres interessos sempre pot reduir la dimensionalitat del problema col·lapsant els haplotips estranys en una sola categoria. En cas de voler col·lapsar haplotips de freqüències superiors, sempre es pot executar prèviament a l'anàlisi amb BayHap, una eina alternativa que retorni el millor nombre d'SNPs a seleccionar (és a dir, el conjunt mínim d'SNPs que conformen els haplotips que millor discriminen la mostra) com per exemple el paquet d'R presentat molt recentment per Dai i col·laboradors anomenat SHARE [222].

Missings

Tot i que BayHap accepta dades amb valors faltants i aquest és un gran avenç, el programa assumeix que aquests valors són repartits de manera aleatòria al llarg de les dades genotípiques i que qualsevol possibilitat al·lèlica pot donar-s'hi. Com hem vist, aquesta aproximació, tot i ser clarament millor que el fet d'ignorar els valors missings, no és la via de tractament més òptima. Estudis molt recents apunten d'altres vies interessants per adreçar aquesta qüestió, com la de Liu et al. [215].

Part VII

CONCLUSIONS

19

Conclusions

Les conclusions que es deriven d'aquesta Tesi Doctoral són les següents:

- L'algorisme dissenyat en aquesta tesi per a l'estimació simultània de freqüències haplotípiques i associació entre haplotips i malaltia millora les solucions reportades per d'altres mètodes, especialment pel que fa a l'estimació d'haplotips poc freqüents a la mostra.
- BayHap, l'aplicació informàtica que implementa l'algorisme dissenyat en aquesta tesi, és un programa vàlid per estimar freqüències haplotípiques i avaluar associació amb haplotips. L'entorn estadístic R ha resultat un recurs apropiat per situar-hi un programa d'aquestes característiques, donat que hi tenen cabuda aplicacions d'aquest tipus i ofereix el lliure accés als usuaris.
- Tot i que cap dels programes d'anàlisi d'haplotips estudiats es mostri globalment superior a la resta, l'enfocament Bayesià en que s'ha basat BayHap ofereix avantatges respecte del conjunt de programes freqüentistes pel que fa a la interpretació i el diagnòstic dels resultats.
- Els mètodes d'integració de Markov Chain Monte Carlo permeten treballar de manera computacionalment òptima amb mètodes d'estimació Bayesians. En particular, pel problema haplotípic la combinació de Random Walk i Slice Sampling és una bona solució a nivell numèric.

- La majoria de programes presenten la seva combinació de punts forts i febles. La tria del programa s'ha de fer en funció dels requeriments de l'anàlisi i les característiques particulars de la mostra.

Part VIII

APÈNDIX

A

Articles publicats

En aquest apèndix es mostra la primera pàgina de tres articles en els que he participat des de l'any 2005 en l'àmbit de l'epidemiologia genètica.

El primer article s'anomena "Análisis estadístico de polimorfismos genéticos en estudios epidemiológicos" i està publicat a la revista "Gaceta Sanitaria". D'aquest article en som autors Raquel Iniesta, Elisabet Guinó i Victor Moreno. En ell exposem la nomenclatura bàsica que s'utilitza en estudis d'epidemiologia genètica així com expliquem diferents estratègies bàsiques d'anàlisi de polimorfismes genètics mitjançant models de regressió Logística i diferents models d'herència.

Pel que fa al segon article, s'anomena "Assessment of Genetic Association using Haplotypes inferred with Uncertainty via Markov Chain Monte Carlo" i es troba publicat a mode de capítol en el llibre MCQMC Proceedings, editat per l'editorial Springer. Els autors som Raquel Iniesta i Victor Moreno. En ell centrem tota la qüestió haplotípica, descrivint el tractament que ha rebut el tema amb anterioritat i presentem el mètode que s'ha dissenyat en aquesta tesi.

En relació al tercer article, du per títol "SNPstats: a web tool for the analysis of association studies" i està publicat a la revista "Bioinformatics". Els autors som Xavier Solé, Elisabet Guinó, Joan Valls, Raquel Iniesta i Victor Moreno. En aquesta publicació presentem una aplicació via web que permet dur a terme anàlisis d'associació genètica, tant amb SNPs com amb haplotips.

NOTA METODOLÓGICA

Análisis estadístico de polimorfismos genéticos en estudios epidemiológicos

Raquel Iniesta[a] / Elisabet Guinó[a] / Víctor Moreno[a,b]
[a]Servicio de Epidemiología y Registro del Cáncer, IDIBELL, Instituto Catalán de Oncología, L'Hospitalet de Llobregat, Barcelona, España; [b]Unidad de Bioestadística, Facultad de Medicina, Universidad Autónoma de Barcelona, Barcelona, España.

(Statistical analysis of genetic polymorphisms in epidemiological studies)

Resumen

El análisis de los polimorfismos genéticos permite identificar genes que confieren susceptibilidad a presentar enfermedades. En este trabajo se presenta la nomenclatura utilizada en la bibliografía de epidemiología genética y una estrategia básica de análisis estadístico de estudios epidemiológicos que incorporan estos marcadores. En primer lugar, se presenta el análisis descriptivo de un único polimorfismo y la evaluación del equilibrio Hardy-Weinberg. A continuación se presentan los métodos para evaluar la asociación con la enfermedad. Para ello se emplean modelos de regresión logística y se estudian los posibles modelos de herencia. Por último, se presentan métodos para el análisis simultáneo de múltiples polimorfismos: estimación de las frecuencias de haplotipos y análisis de asociación con la enfermedad.

Palabras clave: Epidemiología genética. Polimorfismo. Genotipo. Haplotipo. Análisis estadístico.

Abstract

Analysis of genetic polymorphisms allows the genes that confer susceptibility to diseases to be analyzed. This paper presents the nomenclature used in genetic epidemiology literature and a basic strategy for statistical analysis of epidemiological studies that use genetic markers. First, a descriptive analysis of a single nucleotide polymorphism is presented, with assessment of Hardy-Weinberg equilibrium. Next, methods to assess the association with disease are presented. To do this, logistic regression models are used and alternative models of inheritance are explored. Finally, methods for the simultaneous analysis of multiple polymorphisms are presented: haplotype frequency estimation and analysis of disease association.

Key words: Genetic epidemiology. Polymorphism. Genotype. Haplotype. Statistical analysis.

Introducción

Los polimorfismos genéticos son variantes del genoma que aparecen por mutaciones en algunos individuos, se transmiten a la descendencia y adquieren cierta frecuencia en la población tras múltiples generaciones. Se ha estimado que hay una variante en cada 1.000 pares de bases de los 3.000 millones que configuran el genoma humano. Los polimorfismos son la base de la evolución y los que se consolidan, bien pueden ser silentes o proporcionar ventajas a los individuos, aunque también pueden contribuir a causar enfermedades[1]. Se conocen muchas enfermedades determinadas genéticamente por mutaciones o variantes denominadas de «alta penetrancia», ya que los portadores de la variante suelen manifestar la enfermedad con una alta probabilidad. Estas variantes suelen ser de baja frecuencia en la población general, por ejemplo, las mutaciones heredadas en el gen supresor de tumores *APC* determinan la aparición de la poliposis familiar adenomatosa que a menudo degenera en carcinomas en el colon, pero esta entidad no explica más de un 1% del total de tumores de colon.

En la actualidad muchos investigadores centran sus trabajos en identificar genes con polimorfismos que se dan en la población con mayor frecuencia y que influyen en el riesgo de padecer una enfermedad, pero con baja probabilidad (son los llamados polimorfismos de «baja penetrancia»). También se les denomina variantes que confieren susceptibilidad genética a la enfermedad, y para que dicha variante genética se expre-

Correspondencia: Dr. Víctor Moreno.
Servicio de Epidemiología y Registro del Cáncer.
Instituto Catalán de Oncología.
Gran Vía, km 2,7. 08970 L'Hospitalet de Llobregat. Barcelona. España.
Correo electrónico: v.moreno@iconcologia.net

Recibido: 5 de octubre de 2004. *Aceptado:* 12 de enero de 2005.

Assessment of Genetic Association using Haplotypes Inferred with Uncertainty via Markov Chain Monte Carlo

Raquel Iniesta and Victor Moreno

Cancer Epidemiology Service, Catalan Institute of Oncology, Barcelona, Spain
riniesta@iconcologia.net

Summary. In the last years, haplotypic information has become an important subject in the context of molecular genetic studies. Assuming that some genetic mutations take part in the etiology of some diseases, it could be of great interest to compare sets of genetic variations among di erent unrelated individuals, inherited in block from their parents, in order to conclude if there is some association between variations and a disease. The main problem is that, in the absence of family data, obtaining haplotypic information is not straightforward: individuals having more than one polymorphic heterozygous locus have uncertain haplotypes.
We have developed a Markov Chain Monte Carlo method to estimate simultaneously the sample frequency of each possible haplotype and the association between haplotypes and a disease.

1 Introduction

Nowadays, haplotypic information has become vitally important in the context of association studies. Association studies deal with the relationship between genetic information and the etiology of some particular disease. Comparing DNA of healthy and diseased individuals, we can find changes in the sequence that could modify the risk of su ering from the disease [Bal06].

DNA variations we are going to deal with are the changes in only one nucleotide, called SNP (Single Nucleotide Polymorphism).

The knowledge of haplotypes corresponding to a sample of genotypes observed for some SNPs of a set of unrelated individuals is very helpful to better describe this association. Unfortunately, in the absence of family data, obtaining haplotypic information is not straightforward. Since every cell of the human organism contains 22 pairs of homologous chromosomes, plus the sexual chromosomes, for each chromosomical location at the autosomal chromosomes there are two bases, one for each homologous chromosome at the same position of the DNA sequence. Given that current lab techniques usually only report genotypic data and do not provide the chromosome for each base, individuals

BIOINFORMATICS APPLICATIONS NOTE Vol. 22 no. 15 2006, pages 1928–1929
doi:10.1093/bioinformatics/btl268

Genetics and population analysis

SNPStats: a web tool for the analysis of association studies

Xavier Solé[1], Elisabet Guinó[1], Joan Valls[1,2], Raquel Iniesta[1] and Víctor Moreno[1,2,*]

[1]Catalan Institute of Oncology, IDIBELL, Epidemiology and Cancer Registry, L'Hospitalet, Barcelona, Spain and [2]Autonomous University of Barcelona, Laboratory of Biostatistics and Epidemiology, Bellaterra, Barcelona, Spain

Received on March 6, 2006; revised on May 16, 2006; accepted on May 18, 2006
Advance Access publication May 23, 2006
Associate Editor: Charlie Hodgman

ABSTRACT

Summary: A web-based application has been designed from a genetic epidemiology point of view to analyze association studies. Main capabilities include descriptive analysis, test for Hardy–Weinberg equilibrium and linkage disequilibrium. Analysis of association is based on linear or logistic regression according to the response variable (quantitative or binary disease status, respectively). Analysis of single SNPs: multiple inheritance models (co-dominant, dominant, recessive, over-dominant and log-additive), and analysis of interactions (gene–gene or gene–environment). Analysis of multiple SNPs: haplotype frequency estimation, analysis of association of haplotypes with the response, including analysis of interactions.

Availability: http://bioinfo.iconcologia.net/SNPstats. Source code for local installation is available under GNU license.

Contact: v.moreno@iconcologia.net

Supplementary Information: Figures with a sample run are available on *Bioinformatics* online. A detailed online tutorial is available within the application.

The analysis of association between genetic polymorphisms and diseases allows identifying susceptibility genes (Cordell and Clayton, 2005). The proper analysis of these studies can be performed with general purpose statistical packages, but the researcher usually needs the assistance of additional software to perform specific analysis, like haplotype estimation, and results from different packages are difficult to integrate.

We present a free web-based tool to help researchers in the analysis of association studies based on SNPs or biallelic markers. Both the selection of analysis and the output have been designed from a genetic epidemiology perspective. This application can also be used for learning purposes. We have written (in Spanish) an analysis guide with detailed explanations (Iniesta *et al.*, 2005). A similar extensive help in English can also be found on the website.

The software is used following three steps, with the possibility of performing multiple analyses in one session. The steps are as follows.

(1) *Data entry*. Raw data in tabular form can be pasted in a window or uploaded from a text file. Variables can be named and the user can choose the field delimiter and the missing value code (Supplementary Figure 1). SNPs should be coded as genotypes with each allele separated by a slash (e.g. 'T/T', 'T/C', 'C/C').

(2) *Data processing*. A list with the variables read by the application is presented with an initial suggestion about the type: quantitative, categorical or SNP, which can be modified (Supplementary Figure 2). The user is prompted to select those needed for the analysis and to specify which one is the response, which may be binary (disease status) or quantitative. For categorical variables, including SNPs, the user can reorder the categories. The first one will be treated as reference category in the analysis. The application assumes that the main interest is the analysis of the SNPs in relation to the response. Other variables selected with type quantitative or categorical will be added to the regression models for analysis as covariates and treated as potential confounders.

(3) *Analyses customization*. The third step requests the selection of the desired statistical analyses that will be described later in this article (Supplementary Figure 3).

Regarding the statistical analysis, the association with disease is modeled depending on the response variable. If binary, the application assumes an unmatched case–control design and unconditional logistic regression models are used. If the response is quantitative, then a unique population is assumed and linear regression models are used to assess the proportion of variation in the response explained by the SNPs.

The association for each SNP is analyzed in turn and adjusted for the selected covariates. If more than one SNP are selected, then the application assumes that haplotype analysis is appropriate. Haplotype frequencies are estimated using the implementation of the EM algorithm coded into the *haplo.stats* package (Sinnwell and Schaid, 2005, http://mayoresearch.mayo.edu/mayo/research/biostat/schaid.cfm). Association between haplotypes and disease appropriately accounts for the uncertainty in the estimation of haplotypes for individuals with multiple heterozygous when phase is unknown or when missing values are present (Schaid *et al.*, 2002). Individuals with missing values in the response, in all SNPs or in any covariate are excluded from analysis.

The software main page can be found online at http://bioinfo.iconcologia.net/SNPstats. The application uses PHP server programming language to build the input forms, upload data, call the statistical analysis procedures and process the output. The statistical analyses are performed in a batch call to the R package (R Development Core Team, 2005, http://www.R-project.org). The contributed packages *genetics* (Warnes and Leisch, 2005) and *haplo.stats* (Sinnwell and Schaid, 2005, http://mayoresearch.mayo.edu/mayo/research/biostat/schaid.cfm) are called to perform some of the analysis. Anonymous use is guaranteed and data are

*To whom correspondence should be addressed.

© 2006 The Author(s)
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.0/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

B

Taula de programes d'estimació haplotípica

Program Name	Algorithm	Outputa	Missing Datab	Assumptions	Key Features	Limitations	MAX # Subjects, Loci, & Type	Platform
Simple Parsimony								
HAPAR	Parsimony	HA	No	None	·Overcomes limitations of HAPINFREX ·Increasing sample size improve accuracy	·May be susceptible to HWE departures	No Max, Biallelic	PC / UNIX
HAPINFERX	Clark's	HA	No	None	·Intuitive method, fast ·Reduced number haplotypes ·No limit on number of loci	·May fail to start ·Sensitive to data order ·Unstable and erroneous estimates	No Max, Biallelic / Multiallelic	UNIX
Phylogeny								
BPPH	IP	HA	No	Imperfect Phylogeny	·Similar to HAP-H ·Speed	·User Interface	No Max, Biallelic	MAC
DPPH	PP	HA	No	Perfect Phylogeny	·Handles large datasets ·Speed	·Theoretical ·Strict population assumptions	No Max, Biallelic	MAC
GPPH	PP	HA	No	Perfect Phylogeny	·Handles large datasets ·Speed	·Theoretical ·Strict population assumptions	No Max, Biallelic	MAC / PC / UNIX
HAP - H	IP	HA / HF	Yes	HWE, Imperfect Phylogeny	·Predicts haplotype blocks ·Constructs haplotypes within blocks ·Identifies Block Structure ·Web-Based	·No probability of haplotype assignments	Max 500 loci, biallelic	Web Based

Maximum Likelihood								
Arlequin v2.0	EM	HA / HF	No	HWE	·Includes numerous population genetic analysis tools	·EM Issues	EM Practical Limits, Biallelic / Multallelic	JRE on MAC / PC / UNIX
CHAPLIN	ECM	HF	Yes	HWE	·Graphical interface ·Association tests ·HWE assumption relaxed in Case sample	·ECM algorithm needs to be compared to standard EM methods	ECM Practical Limits, Biallelic / Multallelic	PC
EH	EM	HF	No	HWE	·Estimates haplotype frequency ·Compare Case-Control HF under different assumptions	·EM Issues ·Must specify mode of inheritance and pentrance of disease	No Max, 3-4 practical max, Biallelic/ Multiallelic	PC
EHPLUS	EM	HF	No	HWE	·Improves EH, more Loci and polymorphic markers ·Incorporates model free analysis	·Long runtimes for Permutation calculations	Max 5 loci, 15 alleles in analysis	PC / UNIX
EM-DeCODER	EM	HA / HF	No	HWE	·Program with standard EM algorithm	·EM Issues	Max 15 loci, biallelic	UNIX
FASTEHPLUS	EM	HF	No	HWE	·Similar to EHPLUS, with speed improvements	·EM Issues	Max 5 loci, 15 alleles in analysis	PC / UNIX
GENECOUNTING	EM	HA / HF	Yes	HWE	·Provides posterior probabilities for assigned haplotypes ·Compares Global and specific haplotypes between groups	·Missing data limited to biallelic loci ·EM Issues	10-15 loci Practical limit, Biallelic / Multiallelic	PC / UNIX
GCHAP	EM	HA / HF	YES	HWE	·Haplotypes with Zero Likelihood dropped to improve speed and accuracy ·Similar to SNPHAP	·EM Issues	20 loci Practical limit, Biallelic	JRE on PC / UNIX
GS-EM	EM	HA / HF	Yes	HWE	·Includes algorithm for assigning probability to genotype calls from several genotyping methods ·Haplotypes are constructed using genotypes with assigned probability ·Web-Based	·EM Issues ·Limited to Biallelic SNPs	Practical limit, Biallelic	Web Based
HAP - Z	EH	HA / HF	Yes	HWE	·Modified version of SNPHA that accommodates multiallelic loci	·EM Issues	No Max, Biallelic / Multiallelic	PC / UNIX
HAPMAX	MLE	HF	No	HWE	·Ease of use	·Accommodates a limited number of SNPs	8 Loci, Biallelic	PC
HAPLO-H	EM	HF	Yes	HWE	·Handles some missing data ·Utilizes pedigree data, if available ·Calculates standard error	·EM Issues	10 Loci, 40 alleles max, Biallelic/ Multiallelic	UNIX
HAPLOSCOPE	EM / MCMC	▲	▲	▲	·Platform program, incorporates SNPHAP and PHASE v1.0 ·Facilitates comparison/testing ·Graphical Interface, identifies tagging SNPs and LD blocks	·See individual programs for limitations/features	▲	UNIX / Windows
HAPLOVIEW	EM + PL	HA / HF	Yes	HWE	·Calculates pairwise LD ·Checks for recombination ·Identifies tagging SNPs ·Accepts pedigree and unrelated genotype data	·EM Issues	100's, practical limit, bialleic	JRE on MAC / PC / UNIX
HAPLO.STATS	EM	HA / HF	Yes	HWE	·Incorporates method similar to SNPHAP, with user inputs ·Separate programs that: (1) assign haplotypes with posterior probability of assignments (2) allow linear regression for trait to haplotype analysis (3) calculates score statistic for haplotype phenotype association	·Requires Knowledge of S-Plus 6.0 or R ·EM Issues	Practical limit, Biallelic / Multiallelic	S-PLUS 6.0 on UNIX / R on UNIX & PC
HIT	EM / MCMC / MC + PL	▲	▲	▲	·Platform program, incorporates SNPHAP and PHASE v1.0 ·Facilitates comparison ·Graphical Interface, identifies tagging SNPs and LD blocks	·See individual programs for limitations/features	▲	*
HPLUS	EM + EE + PL	HA / HF	Yes	HWE	·Provides posterior probabilities for assigned haplotypes ·Compares Haplotype frequencies between groups, adjusts for covariates ·Utilizes pedigree data, if available	·Requires Matlab ·EM Issues	100 Loci, Biallelic	MATLAB on PC / UNIX
LDSUPPORT	EM	HA / HF	Yes	HWE	·Provides posterior probabilities for assigned haplotypes ·Identifies LD-Blocks for haplotype reconstruction ·Examines association with disease, automation speeds process	·EM Issues	*	UNIX
LOGINSERM ESTIHAPLO	EM	HA / HF	Yes	HWE	·Program uses ML method to infer haplotypes for individuals with missing data ·Offers option to exclude individuals with missing data	·EM Issues	Practical limit, Biallelic / Multiallelic	PC / UNIX
MLHAPFRE	EM	HF	Yes	HWE	·Performance improves with presence of LD ·Performs well with large sample size	·Incorporated into Arlequin ·EM Issues	16 loci, Biallelic	JRE on Mac / PC / UNIX
MLOCUS	EM	HA / HF	Yes	HWE	·Provides posterior probabilities for assigned haplotypes ·Notes observed vs. Inferred haplotypes ·Calculates pairwise LD	·EM Issues	11 Loci, Biallelic / Multiallelic	PC
OSLEM	EM	Yes	No	HWE	·Modified EM alg that runs faster	·EM Issues	Practical limit, Biallelic	Web Based

PL-EM	EM + PL	HA / HF	Yes	HWE	·Combines PL with EM ·EM based version of HAPLOTYPER ·Calculates variance of haplotype frequency estimates	·EM Issues	100's, practical limit, bialleic	PC / UNIX
SAS Genetics	EM	HA / HF	Yes	HWE	·Provides posterior probabilities for assigned haplotypes ·Incorporates statistical tests and procedures	·Requires SAS ·EM Issues	Practical limit, Biallelic / Multiallelic	SAS on PC / UNIX
SNPEM	EM	HF	No	HWE	·Estimates haplotype frequency by population ·Compares global and specific haplotype between 2 groups	·EM Issues	10 Loci, Biallelic	UNIX
SNPHAP	EM	HA / HF	Yes	HWE	·Uses posterior and prior trimming to handle large number loci ·Provides posterior probabilities for assigned haplotypes	·EM Issues	Practical limit, Biallelic	UNIX
THESIAS	S-EM	HF	Yes	HWE	·Stochastic EM avoids issues of standard EM programs ·Includes tests for haplotype-phenotype association ·Handles large sample sizes	·S-EM algorithm needs to be compared to standard EM methods	Practical limit, 20 loci, Biallelic	PC / UNIX
WHAP	EM	▵	▵	▵	·Uses haplotype output from SNPHAP for association testing ·Allows weighted association analysis	·EM issues ·Requires separate haplotyping program	▵	PC / UNIX
Zaykin et al.	EM	HF	No	HWE	·Program on analysis of haplotype-phenotype association	·EM Issues ·Subjects with missing data ignored	Practical limit, Biallelic / Multiallelic	PC / UNIX
Zou and Zhao	MLE / EM	HF	Yes	HWE	·Adjust haplotype frequency estimates for Genotyping Error ·Program also works for nuclear families	·Assumes genotyping errors are random ·Assumes error rates are known	EM Practical Limits, Biallelic / Multallelic	*
3locus.PAS	EM	HF	Yes	HWE	·Handles some missing data ·Various tests available ·Improves with increasing sample size	·EM Issues	3 loci, Biallelic/ Multiallelic	PC / UNIX
Simple Bayesian								
HAPLOTYPER	MC + PL	HA / HF	Yes	HWE	·Use PL algorithm to construct haplotypes with many loci ·Provides posterior probabilities for assigned haplotypes	·Long run times ·Posterior probabilities may be difficult to interpret	256 max, Biallelic	UNIX
HAPLOREC	MC-VL	HA / HF	Yes	HWE	·Uses variable length chain based on maximizing LD ·Handle large number loci	·Restarts avoid non-global optimum	No Max, Biallelic	Java virtual machine, v1.4 or newer
Coalescent-Based Bayesian								
Arlequin v3.0	ELB	HA / HF	No	Ad Hoc Coalescent	·Includes numerous population genetic analyses ·Handles recombination	·Long run times	1000's, Biallelic / Multiallelic	JRE on LINUX / PC/ Mac
PHASE v2.0	MCMC + PL	HA / HF	Yes	Coalescent / HWE	·Improve run time ·Comparison haplotype frequency between groups ·Handles Recombination ·Provides posterior probabilities for assigned haplotypes	·Departures for coalescent model may impact performance ·Posterior probabilities may be difficult to interpret	No Max, Biallelic / Multiallelic	PC / MAC / UNIX
PHASE v1.0	MCMC	HA / HF	No	Coalescent / HWE	·Incorporates pop-genetics and coalescence ideas ·Incorporates known phase and trios pedigrees into analysis ·Provides posterior probabilities for assigned haplotypes	·Departures for coalescent model may impact performance ·Slow run times ·Posterior probabilities may be difficult to interpret	No Max, Biallelic / Multiallelic	UNIX
SLHAP v1.0	MCMC	HA / HF	Yes	Neutral Coalescent / HWE	·Similar to PHASE v1.0 ·Missing data ·Improved run time	·Departures for coalescent model may impact performance	No Max, Biallelic / Multiallelic	UNIX

Figura B.1. Taula de programes de reconstrucció haplotípica.

Program Name	Haplotyping Algorithm	Key Analysis Feature(s)	Discrete Outcome	Continuous Outcome
CHAPLIN	ECM	·Includes Likelihood Ratio statistic and Score statistic for haplotype - phenotype analysis, uses permutation test to determine significance ·Includes AIC for model selection, does not accommodate covariates	Yes, Case-Control	No
EH	EM	·Test for LD for unrelated and in case-control ·Test for frequency difference between case-control under: H1 association, H2 association for all loci	Yes, Case-Control	No
EHPLUS	EM	·Improves on EH ·Model free analysis and permutation test	Yes, Case-Control	No
FASTEHPLUS	EM	·Implements EH and EHPLUS test ·Significant speed improvements	Yes, Case-Control	No
GENECOUNTING	EM	·Compares overall and specific haplotype frequency between cases and controls	Yes, Case-Control	No
HAP^{H}	IP	*Phylogeny based haplotyping method *Uses information from phylogeny for analysis, includes parametric and non-parametric tests for qualitative and quantitative phenotypes	Yes, Case-Control	Yes
HAPLO.STATS	EM	·Score statistic for haplotype - phenotype analysis ·GLM for regression of trait on haplotype, adjustment for covariates and interaction	Yes, Binary, Ordinal, & Poisson	Yes
HPLUS	EE + PL + EM	·Compares haplotypes frequency between cases and controls, option to adjust for covariates, and interaction assessment ·Reports OR, Confidence Interval, and identifies haplotype blocks	Yes, Case-Control	No
HAPASSOC	EM	·Uses likelihood method to calculate risk of developing disease phenotype from diplotype configuration	Yes, Case-Control, gaussian, Poisson and Gamma	Yes
PHASE v2.0	MCMC	·Allows comparison of haplotype frequency between populations	Yes, Case-Control	No
THESIAS	SEM	·Compares haplotypes frequency between cases and controls, survival analysis, option to adjust for covariates, and interaction assessment ·Uses chi-square statistics/t-test for analysis	Yes, Case-Control, Survival Analysis	Yes
SAS Genetics	EM	·Allows comparison of haplotype frequency between populations ·Haplotype Trend Regression (HTR) and several population Genetic tests ·TDT test for family data	Yes, Case-Control	Yes
SNPEM	EM	·Compares overall and specific haplotype frequency between cases and controls ·Includes batch feature for sliding windows analysis	Yes, Case-Control	No
WHAP	EM	·Uses SNPHAP for Regression based haplotype association test on SNPs, provides beta estimates of effects ·Includes haplotype weighted likelihood analysis, permutation tests and sliding windows analysis	Yes, Case-Control	Yes
Zaykin et al.	EM	·Likelihood Ratio statistic for haplotype - phenotype analysis ·Allows sliding windows analysis	Yes, Case-Control	Yes
3locus.PAS	EM	·Test for global disequilibrium, including pairwise and three way disequilibrium for an unrelated sample	No	No
Other Analysis Programs				
Arlequin v2.0/3.0	EM / ELB	·Several population genetic tests		
Zou and Zhao	EM	·Adjust haplotype frequency estimates for genotyping error		

Figura B.2. Taula de programes que inclouen mètodes d'anàlisi d'associació.

C

Especificacions matemàtiques

Algorisme EM

En aquesta secció passem a descriure els aspectes teòrics de l'algorisme EM, una de les eines que com ja hem vist a la introducció, ha estat àmpliament utilitzada per tractar la qüestió haplotípica.

L'algorisme EM (*Expectation Maximization*) és un mètode general que té per objectiu calcular el MLE (*Maximum Likelihood Estimator*) pels paràmetres d'una funció de versemblança. L'algorisme s'aplica principalment en les dues situacions següents:

1. Quan no és possible maximitzar la versemblança analíticament
2. Quan es tenen dades incomplertes, ja sigui a causa d'incertesa inherent a la naturalesa de les dades, o bé per l'existència de missings.

Al nostre cas, el paràmetre a estimar és la freqüència relativa d'haplotips en una població. Aquest paràmetre s'estima mitjançant la funció de versemblança descrita al capítol 9.1, una funció de difícil maximització analítica.

Aplicació de l'algorisme al cas dels haplotips

Donat un genotip, considerarem que ve definit unívocament segons els haplotips compatibles amb ell. És a dir, entendrem un genotip com la possibilitat de transportar una parella

concreta d'haplotips. Notem que tot i que donat un genotip, aquest pot ser compatible amb diverses parelles d'haplotips, a l'inrevés no és cert, es compleix unicitat:

Observació: Donada una parella d'haplotips, hi ha un i només un genotip possible compatible amb la parella haplotípica.

Així doncs, la probabilitat de dur un genotip podrà ser expressada com la probabilitat de dur parelles concretes d'haplotips .

Pas E: Aquest pas de l'algorisme consisteix en calcular l'esperança de cada genotip en funció dels haplotips que porta, utilitzant les freqüències d'haplotips actuals.

Sigui g_i un genotip tal que no presenta incertesa pel que fa als seus haplotips. Sigui (h_r,h_s) la única parella d'haplotips compatible amb g_i. L'esperança del genotip serà:

$$F_{g_i} = p(h_r h_s) = \frac{n_i}{n} \tag{C.1}$$

on n_i és el nombre de cops que apareix el genotip i-éssim a la mostra. En cas que el genotip g_i pugui dur més d'una parella d'haplotips, la freqüència del genotip es pot descomposar pels diferents casos de parelles possibles. És a dir, la probabilitat que un genotip porti la parella d'haplotips (h_r,h_s) és:

$$p(h_r h_s) = \frac{n_i}{n} \frac{c_{rs} f_{h_r} f_{h_s}}{\sum_{h_r,h_s \in H_i} c_{rs} f_{h_r} f_{h_s}}$$

Per tant la freqüència total del genotip s'expressa com a suma de les diferents freqüències, obtingudes a partir de les diferents parelles d'haplotips compatibles amb g_i:

$$F_{g_i} = \sum_{h_r,h_s \in H_i} p(h_r h_s) = \frac{n_i}{n} \sum_{h_r,h_s \in H_i} \frac{c_{rs} f_{h_r} f_{h_s}}{\sum_{h_r,h_s \in H_i} c_{rs} f_{h_r} f_{h_s}} = \frac{n_i}{n} \tag{C.2}$$

Per exemple, suposem el cas que el genotip pugui dur les parelles (h_1,h_2),(h_3,h_4). Aquest genotip tant es pot entendre com el que porta la primera parella o com el que porta la segona. Fixem-nos que segons (C.2) la freqüència s'expressa en dues parts, en funció se si suposem que el genotip porta una parella d'haplotips o l'altra:

$$p(h_1 h_2) = \frac{n_i}{n} \frac{f_{h_1} f_{h_2}}{f_{h_1} f_{h_2} + f_{h_3} f_{h_4}}$$

$$p(h_3h_4) = \frac{n_i}{n} \frac{f_{h_3} f_{h_4}}{f_{h_1} f_{h_2} + f_{h_3} f_{h_4}}$$

I per tant,

$$F_{g_i} = \frac{n_i}{n} \frac{f_{h_1} f_{h_2}}{f_{h_1} f_{h_2} + f_{h_3} f_{h_4}} + \frac{n_i}{n} \frac{f_{h_3} f_{h_4}}{f_{h_1} f_{h_2} + f_{h_3} f_{h_4}} = \frac{n_i}{n}$$

Aquest és el pas E de l'algorisme. Fins aquí sabem calcular les freqüències dels diferents genotips, incerts o no, en funció dels haplotips.

Pas M: El pas M es basa en calcular unes noves freqüències haplotípiques. Bàsicament, es realitza un recompte dels cops que apareix cada haplotip a la mostra, usant les freqüències del pas anterior:

$$p_t^{(g+1)} = \frac{1}{2} \sum_{j=1}^{n} \sum_{i=1}^{2^m} \delta_{it} f_{h_r} f_{h_s}{}^{(g)} \tag{C.3}$$

on $p_t^{(g+1)}$ és la freqüència de l'haplotip t dins la mostra, al pas $(g+1)$ de l'algorisme. L'$\frac{1}{2}$ és necessari donat que cada individu porta dos haplotips i per tant la mostra haplotípica té el tamany doblat respecte la d'individus. n és el nombre total de genotips diferents a la mostra, m és el nombre de loci heterozigots per un genotip concret i per tant, 2^m és el nombre total d'haplotips diferents que pot tenir un genotip amb m locus heterozigots. Per acabar, δ_{it} és una variable indicadora que pren valors 0, 1 o 2 segons si el genotip j-éssim porta l'haplotip t, 0, 1 o 2 vegades.

L'expressió, doncs, està calculant la freqüència de cada haplotip. Per cada genotip de la mostra, suma la freqüència de l'haplotip segons els cops que hi pot aparèixer al genotip. Si és incompatible amb el genotip, directament δ_{it} val 0.

L'algorisme EM es basa en anar iterant i alternant les passes E i M fins que els valors convergeixin. Pel primer pas, cal donar una llavor per les freqüències haplotípiques.

Teoria referent a les cadenes de Markov

Per començar, una cadena de Markov és un tipus especial de procés estocàstic:

Definició C.0.1 Un procés estocàstic és una família de variables aleatòries $\{\theta^{(t)} \in S : t \in T\}$ on S i T són dos conjunts.

Considerarem que el conjunt T és numerable. Per tant, treballarem amb processos estocàstics discrets. L'espai S s'anomena espai d'estats i acostuma a ser un subconjunt de ^{d}, però també pot ser discret. El procés estocàstic es pot entendre com un conjunt de variables aleatòries que depenen del temps.

Definició C.0.2 Siguin $A_1, \ldots, A_{n-1}, A \subset S$. Una cadena de Markov és un procés estocàstic que cumpleix:

$$P(\theta^{(n+1)} \in A | \theta^{(n)} = x, \theta^{(n-1)} \in A_{n-1}, \ldots, \theta^{(0)} \in A_0 = P(\theta^{(n+1)} \in A | \theta^{(n)} = x) \qquad \text{(C.4)}$$

Per tant, una cadena de Markov es caracteritza perquè donat l'estat actual, passat i futur són independents. En general, la probabilitat (C.4 depèn d'x, A i n. Però si la probabilitat de l'estat futur (que només depèn de l'actual) és sempre la mateixa, *i.e.* no depèn de n, diem que la cadena és *homogènia*.

Probabilitat de transició

Definició C.0.3 Sigui C una cadena homogènia. Definim el *transition kernel* $P(x, A)$ com:

1. $\forall x \in S, P(x, \cdot)$ és una distribució de probabilitat sobre S.
2. $\forall A \subset S$, $x \mapsto P(x, A)$ està ben definida.

Per espais d'estats S discrets, s'acostuma a identificar

$$P(x, A) = P(x, \{y\} = P(x, y)$$

Tot i que l'aplicació que farem nosaltres serà contínua, és recomenable entendre el funcionament pel cas discret i després fer-lo extensiu al cas continu. Per tant entendrem el *transition Kernel* com la probabilitat de salt d'un estat de la cadena a un altre. És a dir, la probabilitat

que té un estat (futur) de ser visitat, partint d'un altre (l'actual), però independentment de l'anterior (passat).

Exemple C.0.4 Passeig aleatori:

Considerem una partícula movent-se independentment a dreta i esquerra sobre una recta. Sigui f la funció de probabilitat sobre els enters que regeix aquests moviments. Sigui $C = \{\theta^{(n)} : n \in \mathbb{N}\}$ la cadena que representa la posició de la partícula a l'instant $t = n$, amb $n \in \mathbb{N}$. Inicialment $\theta^{(0)}$ es distribueix segons una $\pi^{(0)}$. Per tant, S és l'espai de posicions possibles, que es poden escriure com

$$\theta^{(n)} = \theta^{(n-1)} + \omega_n = \omega_1 + \omega_2 + \cdots + \omega_{n-1} + \omega_n$$

on les ω_i són variables aleatòries independents amb funció de probabilitat f que poden prendre valor 1 (dreta), -1 (esquerra) o bé 0 (no es mou). Per tot això, C és una cadena de Markov sobre $\mathbb{Z}$. Si $f(1) = p$, $f(-1) = q$, i $f(0) = r$ amb $p+q+r = 1$, llavors les probabilitats de transició s'escriuen com:

$$P(x,y) = \begin{cases} p \; si \; y = x+1 \\ q \; si \; y = x-1 \\ r \; si \; y = x \\ 0 \; si \; y \neq x-1, x, x+1 \end{cases}$$

Definició C.0.5 Sigui $S = x_1, \ldots, x_r$ l'espai discret d'estats amb r elements. Definim la matriu de transició P com la matriu que té per entrada (i,j)-éssima la probabilitat de transició $P(x_i, x_j$:

$$P = \begin{pmatrix} P(x_1,x_1 & \cdots & P(x_1,x_r \\ \vdots & \ddots & \vdots \\ P(x_r,x_1 & \cdots & P(x_r,x_r \end{pmatrix}$$

Denotarem per $P(x,y)^m$ la probabilitat de transició després de m passes en la cadena. És a dir, la probabilitat de que, partint de l'estat x, la cadena arribi a l'estat y en m passes.

Proposició C.0.6 Siguin $x_1, \ldots, x_{m-1}$ els $m-1$ estats pel que passa la cadena de Markov abans d'arribar a l'estat y. Aleshores,

$$P^m(x,y) = \sum_{x_1} \cdots \sum_{x_{m-1}} P(x, x_1 P(x_1, x_2 \cdots P(x_{m-1}, y)$$

Demostració.

$$P^m(x,y) = Pr(\theta^{(m)} = y | \theta^{(0)} = x) =$$

$$= \sum_{x_1} \cdots \sum_{x_{m-1}} Pr(\theta^{(m)} = y, \theta^{(m-1)} = x_{m-1}, \ldots, \theta^{(1)} = x_1 | \theta^{(0)} = x) =^{(1)}$$

$$= \sum_{x_1} \cdots \sum_{x_{m-1}} Pr(\theta^{(m)} = y | \theta^{(m-1)} = x_{m-1}, \ldots, Pr(\theta^{(1)} = x_1 | \theta^{(0)} = x) =$$

$$= \sum_{x_1} \cdots \sum_{x_{m-1}} P(x, x_1 P(x_1, x_2 \cdots P(x_{m-1}, y)$$

Observacions C.0.7

- (1) és certa per ser cadena de Markov.
- La darrera igualtat ens diu que P^m s'aconsegueix multiplicant P per si mateixa m cops.

Proposició C.0.8 En aquest context,

$$P^{n+m} = \sum_z P^n(x,z) P^m(z,y) \tag{C.5}$$

Demostració.

$$P^{n+m} = \sum_z Pr(\theta^{(n+m)} = y | \theta^{(n)} = z, \theta^{`(0)} = x) Pr(\theta^n = z | \theta^0 = x) = \sum_z P^n(x,z) P^m(z,y)$$

Corol.lari C.0.9

Com que hem aconseguit identificar la matriu de transició al pas m amb el producte matricial, es cumpleix que $P^{n+1} = P^n P$

Notarem a la distribució marginal de l'n-éssim estat de la cadena com:

$$\pi^{(n)} = (\pi^{(n)}(x_1, \cdots, \pi^{(n)}(x_r)$$

On cadascun del, $\pi^{(n)}(x_i$ s'entén com la probabilitat que té la cadena de prendre l'estat x_i, des de qualsevol estat anterior. Per $n = 0$, coincideix amb la distribució inicial de la cadena.

Proposició C.0.10 En notació matricial, es compleix que $\pi^{(n)} = \pi^{(0)}P^n$. A més, $\pi^{(n)} = \pi^{(n-1)}P$.

Demostració. Sigui $y \in S$ l'estat al que salta la cadena.

$$\pi^{(n)}(y) = Pr(\theta^{(n)} = y) =$$
$$= \sum_{x \in S} Pr(\theta^{(n)} = y | \theta^{(0)} = x) Pr(\theta^{(0)} = x)$$
$$= \sum_{x \in S} P^n(x, y)\pi^{(0)}(x)$$

Per tant,

$$\pi^{(n)} = (\pi^{(n)}(x_1, \ldots, \pi^{(n)}(x_r) = \left(\sum_{x_i, x_j \in S} P^{(n)}(x_i, x_j \pi^{(0)}(x_i \right)$$

I per tant en notació matricial es compleix $\pi^{(n)} = \pi^{(0)}P^n$ que també és vàlid per $n-1$. Així doncs

$$\pi^{(n)} = \pi^{(0)}P^{n-1}P = \pi^{n-1}P$$

Notació 1 *La probabilitat per un esdeveniment $A \subset S$ per una cadena de Markov que comença en x, es denota $Pr_x(A)$.*

Definició C.0.11 Sigui $A \subset S$. Si $\theta^{(n)} \in A$ per algun n, definim el temps d'arribada a A com $T_A = min\{n \geq 1 \mid \theta^{(n)} \in A\}$. Si $\nexists\, n$ llavors $T_A = \infty$

Notació 2 *Si $A = \{a\}$, notarem $T_{\{a\}} = T_a$*

Descomposició de S

Passem a classificar els diferents estats en que es pot trobar una cadena de Markov amb espai d'estats S i matriu de transició P. Per estudiar la cadena ens interessa saber quins estats visita i quants cops ho fa.

Definició C.0.12 La probabilitat de que la cadena que ha començat en un estat x arribi a l'estat y en alguna passa posterior és:

$$\rho_{xy} = \{Pr_x(y) \mid T_y < \infty\} =^{(1)} Pr_x(T_y < \infty)$$

(1) és notació.

Definició C.0.13 El nombre de visites que fa una cadena a l'estat y és

$$N(y) = \#\{n > 0 \mid \theta^{(n)} = y\} = \sum_{n=1}^{\infty} \mathbb{I}(\theta^{(n)} = y)$$

Definició C.0.14 Un estat $y \in S$ s'anomena recurrent si la cadena de Markov començada a y, retorna a y amb probabilitat 1, *i.e*, si $\rho_{yy} = 1$.

Per tant si una cadena comença en un estat recurrent sabem amb seguretat que, per cert n retornarà al punt d'inici.

Definició C.0.15 Un estat $y \in S$ és de transició si $\rho_{yy} < 1$.

Per tant, si la cadena cau en un estat de transició, tenim probabilitat positiva de que la cadena no hi torni a passar.

Observació C.0.16 Un estat absorvent, *i.e*, un estat $t.q$ la cadena no es mou d'ell, és un estat recurrent, ja que

$$Pr_y(T_y = 1) = Pr_y(\theta^{(1)} = y) = P(y, y) = 1$$

Observació C.0.17 Si una cadena de Markov comença en un estat y recurrent, el temps de retorn T_y és una quantitat finita aleatòria a qui li podem calcular l'esperança μ_y.

Definició C.0.18 Sigui y un estat recurrent. Direm que l'estat és recurrent positiu si μ_y és finita. En cas contrari li direm *null* recurrent.

La recurrència positiva és una propietat molt important de les cadenes de Markov com veurem a la propera secció.

Proposició C.0.19 Sigui $y \in S$ un estat de transició, $\forall x \in S$,

$$Pr_x(N(y) < \infty) = 1$$

i,

$$E[N(y) \mid \theta^{(0)} = x] = \frac{\rho_{xy}}{1 - \rho_{xy}} < \infty$$

Demostració. Per definició d'estat de transició, la probabilitat de que una cadena que comença a x arribi a y un nombre finit de vegades és 1, ja que es poden donar dues situacions:

- o bé la cadena no arriba mai a y, i llavors $N(y) = 0$ que és finit.
- o bé la cadena arriba un primer cop a y però com és de transició, té probabilitat positiva de no tornar-hi. Per tant, $N(y) < \infty$.

Per demostrar la segona igualtat, observem que

$$E[N(y) \mid \theta^{(0)} = x] = \sum_{n=1}^{\infty} P^n(x,y) =^{(1)} \frac{\rho_{xy}}{1-\rho_{xy}} \tag{C.6}$$

(1) és cert ja que per cada n fixat, sabem que $P^n(x,y) = P(x,y)^n$. Per tant, com que $P(x,y) = \rho_{xy}$, estem sumant una sèrie geomètrica amb raó < 1, que per tant és convergent i suma això.

Proposició C.0.20 Sigui $y \in S$ un estat recurrent. Llavors,

$$Pr(N(y) = \infty) = 1$$

i,

$$E[N(y)|\theta^{(0)} = y] = \infty$$

Demostració. Com que y és recurrent, sabem que la cadena que passa per y sempre hi retorna, per tant $N(y) = \infty$ amb seguretat. Per provar la segona igualtat només cal considerar (C.6, amb $\rho_{xy} = 1$.

Per tant, els estats recurrents són infinitament visitats amb seguretat. En canvi els estats de transició es visiten un nombre finit de cops. Resulta interessant descomposar l'espai S en subgrups d'estats de transició i recurrents. A partir d'aquesta descomposició, podem estudiar la probabilitat de que la cadena arribi a un d'aquests subgrups.

Definició C.0.21 Siguin x i y dos estats de S, $x \neq y$. Es diu que x arriba a y, denotat $x \to y$ si $\rho_{xy} > 0$.

Definició C.0.22 Un subconjunt $C \subseteq S$ es diu que és tancat si $\rho_{xy} = 0$ per $x \in C$ i $y \notin C$.

Definició C.0.23 Direm que C és irreductible si $x \to y \; \forall x, y \in C$. Una cadena es diu irreductible si S ho és.

Proposició C.0.24 La recurrència defineix una classe d'equivalència respecte la operació $\leftrightarrow$. És a dir,

- Si x és recurrent, $x \to x$ i x és recurrent.
- Si x és recurrent i $x \to y$, aleshores y és recurrent i en aquest cas $y \to x$.
- Si x, y i z són estats recurrents i $x \to y$, $y \to z$ aleshores $x \to z$.

I encara és possible enunciar un resultat més fort:

Teorema C.0.25 La recurrència negativa i positiva també defineixen una classe d'equivalència.

Corol.lari C.0.26

Si $C \subseteq S$ és tancat, finit i irreductible, aleshores tots els estats de C són recurrents.

Observem que la irreductibilitat a C fa que tots els estats es visitin entre ells. Per tant, si $x \to y$, també $y \to x$, i per tant, x és recurrent. Necessitem que sigui tancat, perquè, si per exemple $x \to z$ on $z \notin C$ no sabem que es compleixi irreductibilitat per z i la cadena podria no tornar a entrar a C. D'aquesta manera, tots els estats de C no serien recurrents. Però si afegim que C sigui tancat, sí.

Distribucions estacionàries

Al context de la simulació, un problema fonamental relacionat amb les cadenes de Markov és l'estudi del comportament asimptòtic de la cadena, quan $n \to \infty$. Un concepte clau és el de distribució estacionària.

Definició C.0.27 Sigui π la distribució d'una cadena amb probabilitat de transició $P(x, y)$. Es diu que π és estacionària si

$$\sum_{x \in S} P(x,y)\pi(x) = \pi(y), \forall y \in S \tag{C.7}$$

En notació matricial, $\pi = \pi P$

Si la distribució en un pas qualsevol de la cadena és π llavors la distribució pel pas següent és $\pi P = \pi$. Un cop la cadena assoleix el nombre de passes necessari per a que π sigui la distribució de la cadena, la cadena reté aquesta distribució per la resta de passes de la cadena. Passem a discutir l'existència i unicitat de distribucions estacionàries. Sigui $N_n(y)$ el nombre de visites que rep l'estat y en n passes. Definim $G_n(x,y) = E_x[N_n(y)]$ la mitjana pel nombre de visites que fa la cadena a l'estat y i $m_y = E_y(T_y$ la mitjana pel temps de retorn a l'estat y. Llavors, $G_n(x,y) = \sum_{k=1}^{n} P^k(x,y)$ i $\lim_{n\to\infty} \frac{G_n(x,y)}{n}$ ens donen una idea del nivell d'ocupació de l'estat y quan la cadena porta un nombre molt gran de passes.

Teorema C.0.28 Es cumpleix que:

- Si $y \in S$ és de transició llavors el $\lim_{n\to\infty} \frac{N_n(y)}{n} = 0$ amb probabilitat 1 i $\lim_{n\to\infty} \frac{G_n(x,y)}{n} = 0$ per tot $x \in S$.
- Si $y \in S$ és recurrent llavors $\lim_{n\to\infty} \frac{N_n(y)}{n} = \frac{\mathbb{I}(T_y<\infty)}{m_y}$ amb probabilitat 1, i $\lim_{n\to\infty} \frac{G_n(x,y)}{n} = \frac{\rho_{xy}}{m_y}$ $\forall x \in S$.

El següent resultat ens dóna la clau per caracteritzar les cadenes que tenen distribució estacionària.

Teorema C.0.29 Una cadena de Markov irreductible és positiva recurrent si i només si té una distribució estacionària tal que

$$\lim_{n\to\infty} \frac{\sum_{k=1}^{n} P^k(x,y)}{n} = \lim_{n\to\infty} \frac{G_n(x,y)}{n} = \pi(y) \tag{C.8}$$

Intuitivament, la probabilitat estacionària d'un estat ve donada per la freqüència de visites a l'estat.

Corol.lari C.0.30

Si π és distribució estacionària, llavors $\pi(x) = 0$, si x és de transició o null recurrent ($m_x = \infty$). Si x és recurrent positiu, $\pi(x) = \frac{1}{m_x}$.

Com que el conjunt d'estats positius recurrents S_{R^p}, i nulls recurrents S_{R^n} són tancats si S és finit, llavors $S_{R^n} = \emptyset$. En aquest cas particular, pel Teorema C.0.29 la cadena té distribució estacionària.

Teoremes sobre límits

No sempre les distribucions estacionàries s'aconsegueixen com a distribucions límit. Per tal de poder establir quan aquestes distribucions estacionàries apareixen com a límit, cal introduir el concepte de periodicitat.

Definició C.0.31 El període d'un estat $x \in S$ és $d_x = mcd\{n \geq 1 \mid P^n(x,x) > 0\}$

Propietats C.0.32

$i)$ Si $P(x,x) > 0$, llavors $d_x = 1$. En aquest cas diem que l'estat és aperiòdic.

$iI)$ Si $x \leftrightarrow y$ llavors $d_x = d_y$.

$iii)$ Els estats d'una cadena irreductible tenen tots igual període.

Un estat x aperiòdic i positiu recurrent s'anomena ergòdic. Una cadena es diu periòdica amb període d si tots els seus estats ho són amb període $d > 1$ i aperiòdica, si tots els seus estats són aperiòdics. Igualment, direm que una cadena és ergòdica si tots els seus estats són ergòdics.

Tot i que l'aperiodicitat no determina l'existència de la distribució estacionària, és necessària a l'hora d'establir convergència per les probabilitats de transició. Veurem quin és el seu paper a l'hora de definir unicitat per la distribució.

Sigui $(\theta^{(n)}_{n\geq 0}$ una cadena irreductible, positiva recurrent amb distribució estacionària π.

Teorema C.0.33

$i)$ Si la cadena és aperiòdica, llavors $\lim_{n\to\infty} P^n(x,y) = \pi(y)\ \forall x, y \in S$

$ii)$ Si la cadena és irreductible i ergòdica (aperiòdica i positiva recurrent) llavors $\lim_{n\to\infty} \|P^n(x,\cdot) - \pi(\cdot)\| = 0\ \forall x \in S$

Per tant, hi ha tres propietats que ens asseguren la convergència de la cadena cap a una distribució estacionària. La irreductibilitat, per a que des de qualsevol punt on comenci la cadena, aquesta pugui assolir qualsevol subconjunt no buit d'estats, amb probabilitat positiva. La cadena ha de ser aperiòdica per evitar que la cadena oscili entre alguns subgrups d'estats periòdicament i no convergeixi. I per últim, la cadena ha de ser positiva recurrent, perquè així ens assegurem l'existència de la distribució estacionària (Teorema C.0.29.

Un cop establerta la ergodicitat de la cadena, podem formular alguns teoremes de convergència importants. Primer, però, cal tenir clar el següent concepte:

Definició C.0.34 Sigui $t(\theta)$ una funció sobre $\mathbb{R}$. La mitjana ergòdica per al valor de la funció és

$$\bar{t}_n = \frac{1}{n}\sum_{i=1}^{n} t(\theta^{(i)}$$

Teorema C.0.35 *Teorema ergòdic*

Sigui $(\theta^{(n)}_{n\geq 0}$ una cadena ergòdica i tal que $E_\pi[t(\theta)] < \infty$ per la única distribució límit π. Llavors,

$$\bar{t}_n \to E_\pi[t(\theta)]$$

quan $n \to \infty$ amb probabilitat 1.

Aquesta és la versió de la llei dels grans nombres adaptat al cas de les cadenes de Markov. Ens assegura, doncs, que les mitjanes dels valors de la cadena ens proporcionen estimadors consistents pels paràmetres de la distribució π.

Al cas particular en que $t(\theta) = \mathbb{I}(\theta = x)$, *i.e*, si només comptem les vegades que l'estat x ha estat visitat, el Teorema Ergòdic estableix que aquesta freqüència relativa convergeix a $\pi(x) = \frac{1}{m_x}$.

Veurem que també és possible formular una versió del Teorema central del límit per cadenes de Markov.

Raó de convergència

Definició C.0.36 Una cadena es diu geomètricament ergòdica si és ergòdica (positiva recurrent i aperiòdica) i existeix una constant $0 \leq \lambda < 1$ i una funció sobre els reals integrable $M(x)$ *t.q*

$$\|P^n(x,\cdot) - \pi(\cdot)\| \leq M(x)\lambda^n \tag{C.9}$$

$\forall x \in S$. Si M no depèn de x, la ergodicitat es diu uniforme.

El valor més petit de λ pel que existeix $M(x)$ s'anomena la *Raó de convergència*. La denotem λ^*. Per entendre millor les implicacions de la convergència geomètrica, hauríem de considerar l'anàlisi espectral de les cadenes de Markov. Si la cadena és reversible, aquesta teoria ens dóna poderoses eines d'anàlisi.
Les probabilitats de transició s'escriuen de forma matricial, i per tant, tenen associades una família de valors propis $\{\lambda_0, \lambda_1, ...\}$ amb els seus vectors propis corresponents $\{v_0, v_1, ...\}$. Doncs, λ^* coincideix amb $\sup_{k>0} \mid \lambda_k \mid$.

Abans de passar al teorema central del límit, definim una sèrie de conceptes:

Definició C.0.37 Sigui $t^n = t(\theta^{(n)}$. A aquesta cadena li definim:

- *Autocovariança de lag* $k > 0$: $\gamma_k = Cov_\pi(t^{(n)}, t^{(n+k)}$
- *Variança de* $t^{(n)}$ és $\sigma^2 = \gamma_0$
- *L'autocorrelació de lag K*: és $\rho_k = \frac{\gamma_k}{\sigma^2}$

És important no barrejar conceptes. σ^2 és la variança de $t(\theta)$ sota la distribució límit π. La variança de la mostra aconseguida, notem-la τ^2 no té perquè coincidir, ja que depèn de si el mostreig ha estat independent. Aquest segon valor, recull la incertesa del mètode.

Teorema C.0.38 Si una cadena és geomètricament uniforme ergòdica, llavors

$$\sqrt{n}\frac{\bar{t}_n - E_\pi[t(\theta)]}{\tau} \to N(0,1) \tag{C.10}$$

en distribució.

Gràcies a (C.10 podrem calcular intervals de credibilitat.

Cadenes Reversibles

Sigui $(\theta_{n\geq 0}^{(n)}$ una cadena de Markov homogènia amb probabilitats de transició $P(x,y)$ i distribució estacionària π. Ens interessa estudiar aquelles cadenes tals que en considerar el conjunt d'estats en ordre invers, $\theta^{(n)}, \theta^{(n-1)}, ...$, les propietats originals es segueixen mantenint.

Propietat C.0.39 *Reversibilitat*

Una cadena de Markov es diu reversible si cumpleix

$$\pi(x)P(x,y) = \pi(y)P(y,x) \tag{C.11}$$

$\forall x, y \in S$.

La reversibilitat és útil, pel següent motiu:

Proposició C.0.40 Sigui π una distribució que satisfà (C.11 per una cadena irreductible. Aleshores la cadena, a més de ser reversible, és positiva recurrent amb distribució estacionària π.

Per tant la construcció de cadenes de Markov amb una distribució estacionària donada, es redueix a trobar probabilitats de transició $P(x,y)$ tals que satisfacin (C.11. En aquest fet es basaran les tècniques que estudiarem.

Cadenes de Markov quan S és continu

Anàlogament al cas dels espais discrets, donada una cadena $\{X^n : n \geq 0\}$ amb distribució estacionària π

1. *Transition kernel*:

$$K(X^n, A) = P(X^{n+1} \in A \mid X^n$$

2. *Distribució estacionària*

$$\pi(A) = \int K(x,A)\pi(x)dx$$

per tot A amb $\pi(A) > 0$.

3. *Distribucions límit*

$$\lim_{n\to\infty} K^n(x, A) = \pi(A)$$

$\forall A$ amb $\pi(A) = 0$.

4. *Irreductibilitat* Si per tots els conjunts A amb $\pi(A) > 0$ i per tot $x \in A$, existeix un enter $n \geq 1$ tal que $K^n(x, A) > 0$.

5. *Aperiodicitat i recurrència* se segueixen anàlogament de les definicions per S discret, però amb el concepte de recurrència de Harris substituint la recurrència positiva.

6. *Teorema Ergòdic* La distribució invariant π és única i és la distribució límit d'una cadena de Markov ergòdica.

7. $\forall x, y \in S$, $\pi(x)K(x, y) = \pi(y)K(y, x)$. Les cadenes de Markov reversibles tenen π com a distribució invariant.

Referències

1. M. J. Daly, J. D. Rioux, S. F. Schaffner, T. J. Hudson, and E. S. Lander. High-resolution haplotype structure in the human genome. *Nat Genet*, 29(2):229–32, 2001.
2. A. Jeffreys, L. Kauppi, and R. Neumann. Intensely punctate meiotic recombination in the class ii region of the major histocompatibility complex. *Nature Genetics*, 29(2):217–222, 2001.
3. N. Patil, A.J. Berno, D.A. Hinds, W.A. Barret, et al. Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science*, 294:1719–1723, 2001.
4. S. B. Gabriel, S. F. Schaffner, H. Nguyen, J. Moore, et al. The structure of haplotype blocks in the human genome. *Science*, 296(5576):2225–9, 2002.
5. G.A.T. McVean, Myers S.R., Hunt S., Deloukas P., Bentley D.R., and Donnelly P. The fine-scale structure of recombination rate variation in the human genome. *Am J Hum Genet*, 304:581–584, 2004.
6. The International HapMap Consortium. A second generation human haplotype map of over 3.1 million snps. *Nature*, 449:851–861, 2007.
7. The International HapMap Consortium. A haplotype map of the human genome. *Nature*, 437: 1299–1320, 2005.
8. The International HapMap Consortium. The international hapmap project. *Nature*, 426:789–796, 2003.
9. The International HapMap Consortium. Integrating ethics and science in the international hapmap project. *Nature Reviews Genetics*, 5:467–475, 2004.
10. G.A. Thorisson, A.V. Smith, L. Krishnan, and L.D. Stein. The international hapmap project web site. *Genome Research*, 15:1591–1593, 2005.

11. A. G. Clark. Inference of haplotypes from pcr-amplified samples of diploid populations. *Mol Biol Evol*, 7(2):111–22, 1990.

12. L. Excoffier and M. Slatkin. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol Biol Evol*, 12(5):921–7, 1995.

13. Dempster, Laird, and Rubin. Maximum likelihood from incomplete data via the em-algorithm. *Journal of the Royal Satistical Society*, 39:1–38, 1977.

14. Celeux and J. Diebolt. The sem algorithm: a probabilistic teacher derived from the em algorithm for the mixture problem. *Computer Statistics Quart*, pages 73–82, 1985.

15. Z. S. Qin, T. Niu, and J. S. Liu. Partition-ligation-expectation-maximization algorithm for haplotype inference with single-nucleotide polymorphisms. *Am J Hum Genet*, 71(5):1242–7, 2002.

16. D. Clayton. Snphap a program for estimating frequencies of haplotypes of large numbers of diallelic markers from unphased genotype data from unrelated subjects. *version 1.3*, 2001. URL `http://www-gene.cimr.cam.ac.uk/clayton/software`.

17. M. Stephens, N. J. Smith, and P. Donnelly. A new statistical method for haplotype reconstruction from population data. *Am J Hum Genet*, 68(4):978–89, 2001.

18. T. Niu, Z. S. Qin, X. Xu, and J. S. Liu. Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms. *Am J Hum Genet*, 70(1):157–69, 2002.

19. M.W.T. Tanck, J.W. Jukema, A.H.E.M. Klerkx, Kuivenhoven, J.A., et al. A novel method to estimate haplotype effects in patient populations. *Circulation*, 104:179–90, 2001.

20. D. A. Tregouet, S. Escolano, L. Tiret, A. Mallet, and J. L. Golmard. A new algorithm for haplotype-based association analysis: the stochastic-em algorithm. *Ann Hum Genet*, 68(Pt 2): 165–77, 2004.

21. Radford M. Neal. Slice sampling. *The Annals of Statistics*, 31(3):705–767, 2003.

22. R. Sachidanandam, D. Weissman, S. C. Schmidt, J. M. Kakol, et al. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature*, 409(6822): 928–33, 2001.

23. J. C. Venter, M. D. Adams, E. W. Myers, P. W. Li, et al. The sequence of the human genome. *Science*, 291(5507):1304–51, 2001.

24. P. Taillon-Miller, Z. Gu, Q. Li, L. Hillier, and P. Y. Kwok. Overlapping genomic sequences: a treasure trove of single-nucleotide polymorphisms. *Genome Res*, 8(7):748–54, 1998.

25. K. H. Buetow, M. N. Edmonson, and A. B. Cassidy. Reliable identification of large numbers of candidate snps from public est data. *Nat Genet*, 21(3):323–5, 1999.

26. G. T. Marth, I. Korf, M. D. Yandell, R. T. Yeh, Z. Gu, H. Zakeri, N. O. Stitziel, L. Hillier, P. Y. Kwok, and W. R. Gish. A general approach to single-nucleotide polymorphism discovery. *Nat Genet*, 23(4):452–6, 1999.

27. K. Garg, P. Green, and D. A. Nickerson. Identification of candidate coding region single nucleotide polymorphisms in 165 human genes using assembled expressed sequence tags. *Genome Res*, 9(11):1087–92, 1999.

28. K. Irizarry, V. Kustanovich, C. Li, N. Brown, S. Nelson, W. Wong, and C. J. Lee. Genome-wide analysis of single-nucleotide polymorphisms in human expressed sequences. *Nat Genet*, 26(2): 233–6, 2000.

29. D. Altshuler, V. J. Pollara, C. R. Cowles, W. J. Van Etten, J. Baldwin, L. Linton, and E. S. Lander. An snp map of the human genome generated by reduced representation shotgun sequencing. *Nature*, 407(6803):513–6, 2000.

30. C. Schlotterer. The evolution of molecular markers–just a matter of fashion? *Nat Rev Genet*, 5(1): 63–9, 2004.

31. Z. Yang, G. K. Wong, M. A. Eberle, M. Kibukawa, D. A. Passey, W. R. Hughes, L. Kruglyak, and J. Yu. Sampling snps. *Nat Genet*, 26(1):13–4, 2000.

32. L. Kruglyak and D. A. Nickerson. Variation is the spice of life. *Nat Genet*, 27(3):234–6, 2001.

33. C. E. Glatt, J. A. DeYoung, S. Delgado, S. K. Service, K. M. Giacomini, R. H. Edwards, N. Risch, and N. B. Freimer. Screening a large reference sample to identify very low frequency sequence variants: comparisons between two genes. *Nat Genet*, 27(4):435–8, 2001.

34. C. S. Carlson, M. A. Eberle, L. Kruglyak, and D. A. Nickerson. Mapping complex disease loci in whole-genome association studies. *Nature*, 429(6990):446–52, 2004.

35. F. S. Collins, E. D. Green, A. E. Guttmacher, and M. S. Guyer. A vision for the future of genomics research. *Nature*, 422(6934):835–47, 2003.

36. J. C. Stephens, J. A. Schneider, D. A. Tanguay, J. Choi, et al. Haplotype variation and linkage disequilibrium in 313 human genes. *Science*, 293(5529):489–93, 2001.

37. M. K. Halushka, J. B. Fan, K. Bentley, L. Hsie, N. Shen, A. Weder, R. Cooper, R. Lipshutz, and A. Chakravarti. Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. *Nat Genet*, 22(3):239–47, 1999.

38. C. S. Carlson, M. A. Eberle, M. J. Rieder, Q. Yi, L. Kruglyak, and D. A. Nickerson. Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *Am J Hum Genet*, 74(1):106–20, 2004.

39. R. J. Livingston, A. von Niederhausern, A. G. Jegga, D. C. Crawford, et al. Pattern of sequence variation across 213 environmental response genes. *Genome Res*, 14(10A):1821–31, 2004.

40. K. T. Zondervan and L. R. Cardon. The complex interplay among factors that influence allelic association. *Nat Rev Genet*, 5(2):89–100, 2004.

41. A. E. Guttmacher and F. S. Collins. Genomic medicine–a primer. *N Engl J Med*, 347(19):1512–20, 2002.

42. N. E. Caporaso. Why have we failed to find the low penetrance genetic constituents of common cancers? *Cancer Epidemiol Biomarkers Prev*, 11(12):1544–9, 2002.

43. H. K. Tabor, N. J. Risch, and R. M. Myers. Candidate-gene approaches for studying complex genetic traits: practical considerations. *Nat Rev Genet*, 3(5):391–7, 2002.

44. E. S. Lander. The new genomics: global views of biology. *Science*, 274(5287):536–9, 1996.

45. N. Risch and K. Merikangas. The future of genetic studies of complex human diseases. *Science*, 273(5281):1516–7, 1996.

46. F. S. Collins, M. S. Guyer, and A. Charkravarti. Variations on a theme: cataloging human dna sequence variation. *Science*, 278(5343):1580–1, 1997.

47. J. K. Pritchard and N. J. Cox. The allelic architecture of human disease genes: common disease-common variant...or not? *Hum Mol Genet*, 11(20):2417–23, 2002.

48. R. L. Nagel, M. E. Fabry, J. Pagnier, I. Zohoun, H. Wajcman, V. Baudin, and D. Labie. Hematologically and genetically distinct forms of sickle cell anemia in africa. the senegal type and the benin type. *N Engl J Med*, 312(14):880–4, 1985.

49. R. L. Nagel, S. Erlingsson, M. E. Fabry, H. Croizat, S. M. Susuka, H. Lachman, M. Sutton, C. Driscoll, E. Bouhassira, and H. H. Billett. The senegal dna haplotype is associated with the amelioration of anemia in african-american sickle cell anemia patients. *Blood*, 77(6):1371–5, 1991.

50. J. H. Stengard, A. G. Clark, K. M. Weiss, S. Kardia, D. A. Nickerson, V. Salomaa, C. Ehnholm, E. Boerwinkle, and C. F. Sing. Contributions of 18 additional dna sequence variations in the gene encoding apolipoprotein e to explaining variation in quantitative measures of lipid metabolism. *Am J Hum Genet*, 71(3):501–17, 2002.

51. C. M. Drysdale, D. W. McGraw, C. B. Stack, J. C. Stephens, R. S. Judson, K. Nandabalan, K. Arnold, G. Ruano, and S. B. Liggett. Complex promoter and coding region beta 2-adrenergic receptor haplotypes alter receptor expression and predict in vivo responsiveness. *Proc Natl Acad Sci U S A*, 97(19):10483–8, 2000.

52. J. H. Lee, J. H. Choi, W. Namkung, J. W. Hanrahan, et al. A haplotype-based molecular analysis of cftr mutations associated with respiratory and pancreatic diseases. *Hum Mol Genet*, 12(18): 2321–32, 2003.

53. D.R. Pamela, B. Funke, K.E. Burdicka, T. Lencza, et al. Comt genotype and manic symptoms in schizophrenia. *Schizophrenia Research*, 87(1-3):28–31, 2006.

54. M. Xu, D. S. Clair, and L. He. Testing for genetic association between the zdhhc8 gene locus and susceptibility to schizophrenia: An integrated analysis of multiple datasets. *Am J Med Genet B Neuropsychiatr Genet*, 2010.

55. W. E. Evans and H. L. McLeod. Pharmacogenomics–drug disposition, drug targets, and side effects. *N Engl J Med*, 348(6):538–49, 2003.

56. R. Weinshilboum. Inheritance and drug response. *N Engl J Med*, 348(6):529–37, 2003.

57. K. T. Zondervan, L. R. Cardon, and S. H. Kennedy. What makes a good case-control study? design issues for complex traits such as endometriosis. *Hum Reprod*, 17(6):1415–23, 2002.

58. L. R. Cardon and L. J. Palmer. Population stratification and spurious allelic association. *Lancet*, 361(9357):598–604, 2003.

59. R. Iniesta, E. Guinó, and V. Moreno. Análisis estadístico de polimorfismos genéticos en estudios epidemiológicos. *Gac Sanit*, 19(4):333–41, 2005.

60. L. Kruglyak. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nat Genet*, 22(2):139–44, 1999.

61. R. Judson, B. Salisbury, J. Schneider, A. Windemuth, and J. C. Stephens. How many snps does a genome-wide haplotype map require? *Pharmacogenomics*, 3(3):379–91, 2002.

62. L. R. Cardon and J. I. Bell. Association study designs for complex diseases. *Nat Rev Genet*, 2(2): 91–9, 2001.

63. H. Zhao. Family-based association studies. *Stat Methods Med Res*, 9(6):563–87, 2000.

64. W. J. Gauderman, J. S. Witte, and D. C. Thomas. Family-based association studies. *J Natl Cancer Inst Monogr*, (26):31–7, 1999.

65. N. E. Breslow and N. E. Day. *Statistical methods in cancer research. Volume II–The design and analysis of cohort studies*. IARC Sci Publ, 1987.

66. L.P. Fried, N.O. Borhani, P. Enright, C.D. Furberg, et al. The cardiovascular health study: Design and rationale. *Annals of Epidemiology*, 1(3):263–276, 1991.

67. J.D. Kalbfleisch and R.L. Prentice. *The statistical Analysis of Failure Time Data*. Second Edition. Wiley, 2002.

68. D. V. Zaykin, P. H. Westfall, S. S. Young, M. A. Karnoub, M. J. Wagner, and M. G. Ehm. Testing association of statistically inferred haplotypes with discrete and continuous traits in samples of unrelated individuals. *Hum Hered*, 53(2):79–91, 2002.

69. D. C. Crawford, T. Bhangale, N. Li, G. Hellenthal, M. J. Rieder, D. A. Nickerson, and M. Stephens. Evidence for substantial fine-scale variation in recombination rates across the human genome. *Nat Genet*, 36(7):700–6, 2004.

70. J. D. Wall and J. K. Pritchard. Haplotype blocks and linkage disequilibrium in the human genome. *Nat Rev Genet*, 4(8):587–97, 2003.

71. D.C. Crawford, C.S. Carlson, M.J. Rieder, D.P. Carrington, et al. Haplotype diversity across 100 candidate genes for inflammation, lipid metabolism, and blood pressure regulation in two populations. *Am J Hum Genet*, 74(4):610–622, 2004.

72. J. K. Pritchard and M. Przeworski. Linkage disequilibrium in humans: models and data. *Am J Hum Genet*, 69(1):1–14, 2001.

73. L. Subrahmanyan, M. A. Eberle, A. G. Clark, L. Kruglyak, and D. A. Nickerson. Sequence variation and linkage disequilibrium in the human t-cell receptor beta (tcrb) locus. *Am J Hum Genet*, 69(2):381–95, 2001.

74. G.C. Johnson, L. Esposito, B.J. Barratt, A.N. Smith, et al. Haplotype diversity across 100 candidate genes for inflammation, lipid metabolism, and blood pressure regulation in two populations. *Nature Genetics*, 29(2):233–7, 2001.

75. D. O. Stram, C. A. Haiman, J. N. Hirschhorn, D. Altshuler, L. N. Kolonel, B. E. Henderson, and M. C. Pike. Choosing haplotype-tagging snps based on unphased genotype data using a preliminary sample of unrelated subjects with an example from the multiethnic cohort study. *Hum Hered*, 55(1):27–36, 2003.

76. K. Zhang and L. Jin. Haploblockfinder: haplotype block analyses. *Bioinformatics*, 19(10):1300–1, 2003.

77. D. Thompson, D. Stram, D. Goldgar, and J. S. Witte. Haplotype tagging single nucleotide polymorphisms and association studies. *Hum Hered*, 56(1-3):48–55, 2003.

78. K. Zhang, Z. S. Qin, J. S. Liu, T. Chen, M. S. Waterman, and F. Sun. Haplotype block partitioning and tag snp selection using genotype data and their applications to association studies. *Genome Res*, 14(5):908–16, 2004.

79. E. Dawson, G. R. Abecasis, S. Bumpstead, Y. Chen, et al. A first-generation linkage disequilibrium map of human chromosome 22. *Nature*, 418(6897):544–8, 2002.

80. M. S. Phillips, R. Lawrence, R. Sachidanandam, A. P. Morris, et al. Chromosome-wide distribution of haplotype blocks and the role of recombination hot spots. *Nat Genet*, 33(3):382–7, 2003.

81. N. Wang, J.M. Akey, K. Zhang, R. Chakraborty, and L. Jin. Distribution of recombination crossovers and the origin of haplotype blocks: the interplay of population history, recombination, and mutation. *Am J Hum Genet*, 73(5):1227–34, 2002.

82. J.D. Wall and J.K. Pritchard. Assessing the performance of the haplotype block model of linkage disequilibrium. *Am J Hum Genet*, 73(3):502–15, 2003.

83. X. Ke, S. Hunt, W. Tapper, R. Lawrence, G. Stavrides, J. Ghori, P. Whittaker, A. Collins, A.P. Morris, D. Bentley, L.R. Cardon, and P. Deloukas. The impact of snp density on fine-scale patterns of linkage disequilibrium. *Hum Mol Genet*, 13(6):577–88, 2004.

84. T.G. Schulze, K. Zhang, Y.S. Chen, N. Akula, F. Sun, and F.J. McMahon. Defining haplotype blocks and tag single-nucleotide polymorphisms in the human genome. *Hum Mol Genet*, 13(3): 335–42, 2004.

85. M.P. Stumpf. Haplotype diversity and snp frequency dependence in the description of genetic variation. *Eur J Hum Genet*, 12(6):469–77, 2004.

86. A. S. Allen and G. A. Satten. Association mapping via a class of haplotype-sharing statistics. *BMC Proc*, 1 Suppl 1:S123, 2007.

87. A. Dempfle, R. Hein, L. Beckmann, A. Scherag, T. T. Nguyen, H. Schafer, and J. Chang-Claude. Comparison of the power of haplotype-based versus single- and multilocus association methods for gene x environment (gene x sex) interactions and application to gene x smoking and gene x sex interactions in rheumatoid arthritis. *BMC Proc*, 1 Suppl 1:S73, 2007.

88. V. C. Sandrim and J. E. Tanus-Santos. Haplotype analysis can provide improved clinical information than single genotype analysis. *Thromb Res*, 120(5):779, 2007.

89. J. Akey, L. Jin, and M. Xiong. Haplotypes vs single marker linkage disequilibrium tests: what do we gain? *Eur J Hum Genet*, 9(4):291–300, 2001.

90. H. Yan, N. Papadopoulos, G. Marra, and C. Perrera. Conversion of diploidy to haploidy. *Nature*, 403(6771):723–4, 2000.

91. J.A. Douglas, M. Boehnke, E. Gillanders, J.M. Trent, and S.B. Gruber. Experimentally-derived haplotypes substantially increase the efficiency of linkage disequilibrium studies. *Nat Genet*, 28 (4):361–4, 2001.

92. A. G. Clark, K. M. Weiss, D. A. Nickerson, S. Taylor, et al. Haplotype structure and population genetic inferences from nucleotide-sequence variation in human lipoprotein lipase. *Am J Hum Genet*, 63(2):595–612, 1998.

93. L. Ma, Y. Xiao, H. Huang, Q. Wang, W. Rao, Y. Feng, K. Zhang, and Q. Song. Direct determination of molecular haplotypes by chromosome microdissection. *Nat Methods*, 7(4):299–301, 2010.

94. D. H. Bos, S. M. Turner, and J. A. Dewoody. Haplotype inference from diploid sequence data: evaluating performance using non-neutral mhc sequences. *Hereditas*, 144(6):228–34, 2007.

95. M. Pirinen, S. Kulathinal, D. Gasbarra, and M. J. Sillanpaa. Estimating population haplotype frequencies from pooled dna samples using phase algorithm. *Genet Res*, 90(6):509–24, 2008.

96. L. Wang and Y. Xu. Haplotype inference by maximum parsimony. *Bioinformatics*, 19(14):1773–80, 2003.

97. D. Gusfield. Inference of haplotypes from samples of diploid populations: complexity and algorithms. *J Comput Biol*, 8(3):305–23, 2001.

98. G. Zou and H. Zhao. Haplotype inference by pure parsimony. *UC Davis Computer Science Engineering Technical Report*, 2002. URL `http://www.cs.ucdavis.edu/research/techreports/2003/CSE-2003-2.pdf`i.

99. G. Zou and H. Zhao. Haplotyping as perfect philogeny: Conceptual framework and efficient solutions. *Annual conference on Research in Computational Molecular Biology*, 2002. URL `http://www.csif.cs.ucdavis.edu/rgusfield/paperlist.html`.

100. V. Bafna, D. Gusfield, G. Lancia, and S. Yooseph. Haplotyping as perfect phylogeny: a direct approach. *J Comput Biol*, 10(3-4):323–40, 2003.

101. M. E. Hawley and K. K. Kidd. Haplo: a program using the em algorithm to estimate the frequencies of multi-site haplotypes. *J Hered*, 86(5):409–11, 1995.

102. J. C. Long, R. C. Williams, and M. Urbanek. An e-m algorithm and testing strategy for multiple-locus haplotypes. *Am J Hum Genet*, 56(3):799–810, 1995.

103. M. N. Chiano and D. G. Clayton. Fine genetic mapping using haplotype analysis and the missing data problem. *Ann Hum Genet*, 62(Pt 1):55–60, 1998.

104. M. Stephens and P. Donnelly. A comparison of bayesian methods for haplotype reconstruction from population genotype data. *Am J Hum Genet*, 73(5):1162–9, 2003.

105. J. Zhang, M. Vingron, and M. Hoehe. On haplotype reconstruction for diploid populations. *EURANDOM Report*, pages 2001–026, 2001.

106. M. J. Rieder, S. L. Taylor, A. G. Clark, and D. A. Nickerson. Sequence variation in the human angiotensin converting enzyme. *Nat Genet*, 22(1):59–62, 1999.

107. E. Halperin and E. Eskin. Haplotype reconstruction from genotype data using imperfect phylogeny. *Bioinformatics*, 20(12):1842–9, 2004.

108. R.H. Chung and D. Gusfield. *Empirical explanation of perfect phylogeny halotyping and haplotypes*. Lecture Notes in Computer Science. Springer, 2003.

109. G. Lancia, M.C. Pinotti, and R. Rizzi. Haplotyping populations by pure parsimony: Complexity of exact and approximation algorithms. *INFORMS Journal on Computing archive*, 16(4):348–359, 2004.

110. D. Gusfield. An overview of combinatorial methods for haplotype inference. In S. Istrail, M. Waterman, and A. Clark, editors, *Computational Methods for SNP and Haplotype Inference*, pages 9–25. Springer-Verlag, 2004.

111. L. Excoffier, G. Laval, and D. Balding. Gametic phase estimation over large genomic regions using an adaptive window approach. *Hum Genomics*, 1(1):7–19, 2003.

112. D. Fallin and N. J. Schork. Accuracy of haplotype frequency estimation for biallelic loci, via the expectation-maximization algorithm for unphased diploid genotype data. *Am J Hum Genet*, 67 (4):947–59, 2000.

113. J. Tost, O. Brandt, F. Boussicault, D. Derbala, C. Caloustian, D. Lechner, and I. G. Gut. Molecular haplotyping at high throughput. *Nucleic Acids Res*, 30(19):e96, 2002.

114. Y. Kitamura, M. Moriguchi, H. Kaneko, H. Morisaki, T. Morisaki, K. Toyama, and N. Kamatani. Determination of probability distribution of diplotype configuration (diplotype distribution) for each subject from genotypic data using the em algorithm. *Ann Hum Genet*, 66(Pt 3):183–93, 2002.

115. S. A. Tishkoff, A. J. Pakstis, G. Ruano, and K. K. Kidd. The accuracy of statistical methods for estimation of haplotype frequencies: an example from the cd4 locus. *Am J Hum Genet*, 67(2): 518–22, 2000.

116. S. S. Li, N. Khalid, C. Carlson, and L. P. Zhao. Estimating haplotype frequencies and standard errors for multiple single nucleotide polymorphisms. *Biostatistics*, 4(4):513–22, 2003.

117. J. Barret, B. Fry, and M.J. Daly. Haploview: analysis and visualization of ld and haplotype maps. *Bioinformatics*, 2005. URL `http://www.broadinstitute.org/haploview`.

118. D. A. Tregouet and L. Tiret. Cox proportional hazards survival regression in haplotype-based association analysis using the stochastic-em algorithm. *Eur J Hum Genet*, 12(11):971–4, 2004.

119. S. Lin, D. J. Cutler, M. E. Zwick, and A. Chakravarti. Haplotype inference in random population samples. *Am J Hum Genet*, 71(5):1129–37, 2002.

120. Lin S., Chakravarti A., and Cutler D.J. Haplotype and missing data inference in nuclear families. *Genome Res*, 14(8):1624–32, 2004.

121. L. Eronen, F. Geerts, and H. Toivonen. A markov chain approach to reconstruction of long haplotypes. *Pacific Symposium on Biocomputing*, 2004. URL `http://helix-web.stanford.edu/psb04/eronen.pdf`.

122. P. Scheet and M. Stephens. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am J Hum Genet*, 78(4):629–44, 2006.

123. S. Schneidler, D. Roessli, and L. Excoffier. Arlequin: A software for population genetics data analysis. *Genetics and Biometry Laboratory*, University of Geneva:Switzerland, 2002.

124. R.M. Salem, J. Wessel, and N.J. Schorck. A comprehensive literature review of haplotyping software and methods for use with unrelated individuals. *Hum Genomics*, 2(1):39 – 66, 2005.

125. Z. Feng, N. Liu, and H. Zhao. Haplotype inference and association analysis in unrelated samples. In H.W. Deng, H. Shen, Y.J. Liu, and H. Hu, editors, *Current topics in Human Genetics: Studies in Complex Diseases*, pages 135–176. World Scientific Publishing Company, Singapore, 2008.

126. P.Y. Liu, Y. Lu, and H.W. Deng. Accurate haplotype inference for multiple linked single nucleotide polymorphisms using sibship data. *Genetics*, 174(1):499–509, 2006.

127. M. Stephens, N.J. Smith, and P. Donelly. Reply to zhang et al. *Am J Hum Genet.*, 69(4):912–914, 2001.

128. R. M. Single, D. Meyer, J. A. Hollenbach, M. P. Nelson, J. A. Noble, H. A. Erlich, and G. Thomson. Haplotype frequency estimation in patient populations: the effect of departures from hardy-weinberg proportions and collapsing over a locus in the hla region. *Genet Epidemiol*, 22(2):186–95, 2002.

129. Goldstein D.B., Ahmadi K.R., Weale M.E., and Wood N.W. Genome scans and candidate gene approaches in the study of common diseases and variable drug responses. *Trends Genet.*, 19(11): 615–622, 2003.

130. M. E. Weale. A survey of current software for haplotype phase inference. *Hum Genomics*, 1(2): 141–4, 2004.

131. H. Kang, Z. S. Qin, T. Niu, and J. S. Liu. Incorporating genotyping uncertainty in haplotype inference for single-nucleotide polymorphisms. *Am J Hum Genet*, 74(3):495–510, 2004.

132. D. Fallin, A. Cohen, L. Essioux, I. Chumakov, M. Blumenfeld, D. Cohen, and N. J. Schork. Genetic analysis of case/control data using estimated haplotype frequencies: application to apoe locus variation and alzheimer's disease. *Genome Res*, 11(1):143–51, 2001.

133. D. J. Schaid. Evaluating associations of haplotypes with traits. *Genet Epidemiol*, 27(4):348–64, 2004.

134. L. Beckmann, D. C. Thomas, C. Fischer, and J. Chang-Claude. Haplotype sharing analysis using mantel statistics. *Hum Hered*, 59(2):67–78, 2005.

135. J. Y. Tzeng. Evolutionary-based grouping of haplotypes in association analysis. *Genet Epidemiol*, 28(3):220–31, 2005.

136. M.A. Van der Meulen and G.J. te Meerman. Haplotype sharing analysis in affected individuals from nuclear families with at least one affected offspring. *Genet Epidemiol*, 14:915–920, 1997.

137. K. Yu, J. Xu, D. C. Rao, and M. Province. Using tree-based recursive partitioning methods to group haplotypes for increased power in association studies. *Ann Hum Genet*, 69(Pt 5):577–89, 2005.

138. J. Y. Tzeng, B. Devlin, L. Wasserman, and K. Roeder. On the identification of disease mutations by the analysis of haplotype similarity and goodness of fit. *Am J Hum Genet*, 72(4):891–902, 2003.

139. Y. Y. Zhao, L. Y. Wu, J. H. Zhang, R. S. Wang, and X. S. Zhang. Haplotype assembly from aligned weighted snp fragments. *Comput Biol Chem*, 29(4):281–7, 2005.

140. Z. Zhao, N. Yu, Y. X. Fu, and W. H. Li. Nucleotide variation and haplotype diversity in a 10-kb noncoding region in three continental human populations. *Genetics*, 174(1):399–409, 2006.

141. R. Judson and J. C. Stephens. Notes from the snp vs. haplotype front. *Pharmacogenomics*, 2(1): 7–10, 2001.

142. L. P. Zhao, S. S. Li, and N. Khalid. A method for the assessment of disease associations with single-nucleotide polymorphism haplotypes and environmental variables in case-control studies. *Am J Hum Genet*, 72(5):1231–50, 2003.

143. P. Kraft, D.G. Cox, R.A. Paynter, D. Hunter, and I. De Vivo. Accounting for haplotype uncertainty in matched association studies: A comparison of simple and flexible techniques. *Am J Hum Genet*, 28(3):261–272, 2005.

144. H. Zhang, Z. Li, and G. Zheng. Statistical methods for haplotype-based matched case-control association studies. *Genet Epidemiol*, 31(4):316–326, 2007.

145. E. Lin, Y. Hwang, K. H. Liang, and E. Y. Chen. Pattern-recognition techniques with haplotype analysis in pharmacogenomics. *Pharmacogenomics*, 8(1):75–83, 2007.

146. D.Y. Lin and D. Zeng. Likelihood-based inference on haplotype effects in genetic association studies. *J Am Stat Assoc*, 101:89–104, 2006.

147. D. J. Schaid. Relative efficiency of ambiguous vs. directly measured haplotype frequencies. *Genet Epidemiol*, 23(4):426–43, 2002.

148. J. H. Zhao, D. Curtis, and P. C. Sham. Model-free analysis and permutation tests for allelic associations. *Hum Hered*, 50(2):133–9, 2000.

149. M. N. Chiano and D. G. Clayton. Genotypic relative risks under ordered restriction. *Genet Epidemiol*, 15(2):135–46, 1998.

150. S. L. Lake, H. Lyon, K. Tantisira, E. K. Silverman, S. T. Weiss, N. M. Laird, and D. J. Schaid. Estimation and tests of haplotype-environment interaction when linkage phase is ambiguous. *Hum Hered*, 55(1):56–65, 2003.

151. D. J. Schaid, C. M. Rowland, D. E. Tines, R. M. Jacobson, and G. A. Poland. Score tests for association between traits and haplotypes when linkage phase is ambiguous. *Am J Hum Genet*, 70(2):425–34, 2002.

152. A.H. Klerkx, M.W. Tanck, J.J. Kastelein, H.O. Molhuizen, J.W. Jukema, A.H. Zwinderman, and J.A. Kuivenhoven. Haplotype analysis of the cetp gene: not taqib, but the closely linked -629c–¿a polymorphism and a novel promoter variant are independently associated with cetp concentration. *Hum Mol Genet*, 12(2):111–23, 2003.

153. D. Y. Lin. Haplotype-based association analysis in cohort studies of unrelated individuals. *Genet Epidemiol*, 26(4):255–64, 2004.

154. D. O. Stram, C. Leigh Pearce, P. Bretsky, M. Freedman, J. N. Hirschhorn, D. Altshuler, L. N. Kolonel, B. E. Henderson, and D. C. Thomas. Modeling and e-m estimation of haplotype-specific relative risks from genotype data for a case-control study of unrelated individuals. *Hum Hered*, 55(4):179–90, 2003.

155. C. Spinka, R. J. Carroll, and N. Chatterjee. Analysis of case-control studies of genetic and environmental factors with missing genetic information and haplotype-phase ambiguity. *Genet Epidemiol*, 29(2):108–27, 2005.

156. M. P. Epstein and G. A. Satten. Inference on haplotype effects in case-control studies using unphased genotype data. *Am J Hum Genet*, 73(6):1316–29, 2003.

157. G. A. Satten and M. P. Epstein. Comparison of prospective and retrospective methods for haplotype inference in case-control studies. *Genet Epidemiol*, 27(3):192–201, 2004.

158. N. Chatterjee and R.J. Carroll. Semiparametric maximum likelihood estimation exploiting gene-environment independence in case-control studies. *Biometrika*, 92:399–418, 2005.

159. S. Sinha, S. B. Gruber, B. Mukherjee, and G. Rennert. Inference of the haplotype effect in a matched case-control study using unphased genotype data. *Int J Biostat*, 4(1):Article6, 2008.

160. R. Iniesta and V. Moreno. Assessment of genetic association using haplotypes inferred with uncertainty via markov chain monte carlo. In A. Keller, S. Heinrich, and H. Niederreiter, editors, *Monte Carlo and Quasi Monte Carlo Methods*, pages 529–535. Springer-Verlag, Berlin, 2006.

161. N. Chatterjee, Y. H. Chen, S. Luo, and R. J. Carroll. Analysis of case-control association studies: Snps, imputation and haplotypes. *Stat Sci*, 24(4):489–502, 2009.

162. W. Guo, C. Y. Liang, and S. Lin. Haplotype association analysis of north american rheumatoid arthritis consortium data using a generalized linear model with regularization. *BMC Proc*, 3 Suppl 7:S32, 2009.

163. J. Y. Tzeng, C. H. Wang, J. T. Kao, and C. K. Hsiao. Regression-based association analysis with clustered haplotypes through use of genotypes. *Am J Hum Genet*, 78(2):231–42, 2006.

164. C. Pattaro, I. Ruczinski, D. M. Fallin, and G. Parmigiani. Haplotype block partitioning as a tool for dimensionality reduction in snp association studies. *BMC Genomics*, 9:405, 2008.

165. Z. Yu and D. J. Schaid. Application of sequential haplotype scan methods to case-control data. *BMC Proc*, 1 Suppl 1:S21, 2007.

166. R.P. Jr Igo, D. Londono, K. Miller, A.R. Parrado, et al. Density-based clustering in haplotype analysis for association mapping. *BMC Proc*, Suppl:1–27, 2008.

167. W. Guo and S. Lin. Generalized linear modeling with regularization for detecting common disease rare haplotype association. *Genet Epidemiol*, 33(4):308–16, 2009.

168. Z. Wang and M. S. McPeek. An incomplete-data quasi-likelihood approach to haplotype-based genetic association studies on related individuals. *J Am Stat Assoc*, 104(487):1251–1260, 2009.

169. X. Sole, E. Guino, J. Vall, R. Iniesta, and V. Moreno. Snpstats: a web tool for the analysis of association studies. *Bioinformatics*, 22(15):1928–1929, 2006.

170. S. J. Kang, D. Gordon, and S. J. Finch. What snp genotyping errors are most costly for genetic association studies? *Genet Epidemiol*, 26(2):132–41, 2004.

171. D. J. Lunn, J. C. Whittaker, and N. Best. A bayesian toolkit for genetic association studies. *Genet Epidemiol*, 30(3):231–47, 2006.

172. G.O. Roberts. Markov chain monte carlo concepts related to sampling algorithms. In W.R. Gilks, S. Richardson, and D.J. Spiegelhalter, editors, *Markov Chain Monte Carlo in practice*, pages 45–57. London: ChapmanHall, 1995.

173. M.K. Cowles and B.P. Carlin. Markov chain monte carlo convergence diagnostics: a comparative review. In *Technical Report*, pages 94–008. Division of Biostatistics, School of Public Health, University of Minessota, 1994.

174. W.R. Gilks. Derivative-free adaptive rejection sampling for gibbs sampling. In J. Bernardo, J. Berger, A. P. Dawid, and A. F. M. Smith, editors, *Bayesian Statistics 4*, pages 641–649. Oxford University Press, 1992.

175. W.R. Gilks and P. Wild. Adaptive rejection sampling for gibbs sampling. *Applied Statistics*, 41(2): 337–348, 1992.

176. S. Geisser. *Predictive Inference: An introduction*. Chapman and Hall, 1993.

177. M.I. Toirac López. Análisis genético de los sistemas colecistoquinérgico y dopaminérgico en pacientes esquizofrénicos con alucinaciones auditivas. *Tesis Doctoral dirigida per Rosa De Frutos Illán i codirigida per Julio Sanjuan Arias*, Universitat de València, 2008.

178. F. Gemignani, S. Landi, V. Moreno, L. Gioia-Patricola, A. Chabrier, E. Guino, M. Navarro, M. Cambray, G. Capella, and F. Canzian. Polymorphisms of the dopamine receptor gene drd2 and colorectal cancer risk. *Cancer Epidemiol Biomarkers Prev*, 14(7):1633–8, 2005.

179. G. Kirov, M. C. O'Donovan, and M. J. Owen. Finding schizophrenia genes. *J Clin Invest*, 115(6): 1440–8, 2005.

180. M. J. Owen. Genomic approaches to schizophrenia. *Clin Ther*, 27 Suppl A:S2–7, 2005.

181. M. J. Owen, N. Craddock, and M. C. O'Donovan. Schizophrenia: genes at last? *Trends Genet*, 21 (9):518–25, 2005.

182. M. J. Owen, N. Craddock, and M. C. O'Donovan. Schizophrenia: genes at last? *Trends Genet*, 21 (9):518–25, 2005.

183. M. J. Owen, M. C. O'Donovan, and P. J. Harrison. Schizophrenia: a genetic disorder of the synapse? *BMJ*, 330(7484):158–9, 2005.

184. D. H. Blackwood, P. M. Visscher, and W. J. Muir. Genetic studies of bipolar affective disorder in large families. *Br J Psychiatry Suppl*, 41:s134–6, 2001.

185. R. S. Houlston and I. P. Tomlinson. Polymorphisms and colorectal tumor risk. *Gastroenterology*, 121(2):282–301, 2001.

186. M. M. de Jong, I. M. Nolte, G. J. te Meerman, W. T. van der Graaf, E. G. de Vries, R. H. Sijmons, R. M. Hofstra, and J. H. Kleibeuker. Low-penetrance genes and their involvement in colorectal cancer susceptibility. *Cancer Epidemiol Biomarkers Prev*, 11(11):1332–52, 2002.

187. D. K. Grandy, M. A. Marchionni, H. Makam, R. E. Stofko, M. Alfano, L. Frothingham, J. B. Fischer, K. J. Burke-Howie, J. R. Bunzow, A. C. Server, and et al. Cloning of the cdna and gene for a human d2 dopamine receptor. *Proc Natl Acad Sci U S A*, 86(24):9762–6, 1989.

188. H. Ishiguro, T. Arinami, T. Saito, S. Akazawa, et al. Systematic search for variations in the tyrosine hydroxylase gene and their associations with schizophrenia, affective disorders, and alcoholism. *Am J Med Genet*, 81(5):388–96, 1998.

189. K. Blum, E. P. Noble, P. J. Sheridan, A. Montgomery, T. Ritchie, T. Ozkaragoz, R. J. Fitch, R. Wood, O. Finley, and F. Sadlack. Genetic predisposition in alcoholism: association of the d2 dopamine receptor taqi b1 rflp with severe alcoholics. *Alcohol*, 10(1):59–67, 1993.

190. T. Arinami, M. Itokawa, H. Enguchi, H. Tagaya, S. Yano, H. Shimizu, H. Hamaguchi, and M. Toru. Association of dopamine d2 receptor molecular variant with schizophrenia. *Lancet*, 343(8899):703–4, 1994.

191. K. Ohara, M. Nagai, K. Tani, Y. Nakamura, and A. Ino. Functional polymorphism of -141c ins/del in the dopamine d2 receptor gene promoter and schizophrenia. *Psychiatry Res*, 81(2): 117–23, 1998.

192. T. Lencz, D. G. Robinson, K. Xu, J. Ekholm, S. Sevy, H. Gunduz-Bruce, M. G. Woerner, J. M. Kane, D. Goldman, and A. K. Malhotra. Drd2 promoter region variation as a predictor of sustained response to antipsychotic medication in first-episode schizophrenia patients. *Am J Psychiatry*, 163(3):529–31, 2006.

193. M. J. Parsons, I. Mata, M. Beperet, F. Iribarren-Iriso, B. Arroyo, R. Sainz, M. J. Arranz, and R. Kerwin. A dopamine d2 receptor gene-related polymorphism is associated with schizophrenia in a spanish population isolate. *Psychiatr Genet*, 17(3):159–63, 2007.

194. C. C. Zai, R. W. Hwang, V. De Luca, D. J. Muller, N. King, G. C. Zai, G. Remington, H. Y. Meltzer, J. A. Lieberman, S. G. Potkin, and J. L. Kennedy. Association study of tardive dyskinesia and twelve drd2 polymorphisms in schizophrenia patients. *Int J Neuropsychopharmacol*, 10(5):639–51, 2007.

195. E. P. Noble. The drd2 gene in psychiatric and neurological disorders and its phenotypes. *Pharmacogenomics*, 1(3):309–33, 2000.

196. G. B. Glavin and S. Szabo. Dopamine in gastrointestinal disease. *Dig Dis Sci*, 35(9):1153–61, 1990.

197. M. A. Shibata, M. Hirose, M. Yamada, M. Tatematsu, S. Uwagawa, and N. Ito. Epithelial cell proliferation in rat forestomach and glandular stomach mucosa induced by catechol and analogous dihydroxybenzenes. *Carcinogenesis*, 11(6):997–1000, 1990.

198. S. Basu and P. S. Dasgupta. Decreased dopamine receptor expression and its second-messenger camp in malignant human colon tissue. *Dig Dis Sci*, 44(5):916–21, 1999.

199. J. Duan, M. S. Wainwright, J. M. Comeron, N. Saitou, A. R. Sanders, J. Gelernter, and P. V. Gejman. Synonymous mutations in the human dopamine receptor d2 (drd2) affect mrna stability and synthesis of the receptor. *Hum Mol Genet*, 12(3):205–16, 2003.

200. T. Li, M. Arranz, K. J. Aitchison, C. Bryant, X. Liu, R. W. Kerwin, R. Murray, P. Sham, and D. A. Collier. Case-control, haplotype relative risk and transmission disequilibrium analysis of a dopamine d2 receptor functional promoter polymorphism in schizophrenia. *Schizophr Res*, 32(2):87–92, 1998.

201. T. Ritchie and E. P. Noble. Association of seven polymorphisms of the d2 dopamine receptor gene with brain receptor-binding characteristics. *Neurochem Res*, 28(1):73–82, 2003.

202. N. Liu, K. Zhang, and H. Zhao. Haplotype-association analysis. *Adv Genet*, 60:335–405, 2008.

203. S. Zhang, A. J. Pakstis, K. K. Kidd, and H. Zhao. Comparisons of two methods for haplotype reconstruction and haplotype frequency estimation from population data. *Am J Hum Genet*, 69 (4):906–14, 2001.

204. D. E. Reich, M. Cargill, S. Bolk, J. Ireland, P. C. Sabeti, et al. Linkage disequilibrium in the human genome. *Nature*, 411(6834):199–204, 2001.

205. C. F. Xu, K. Lewis, K. L. Cantone, P. Khan, C. Donnelly, N. White, N. Crocker, P. R. Boyd, D. V. Zaykin, and I. J. Purvis. Effectiveness of computational methods in haplotype prediction. *Hum Genet*, 110(2):148–56, 2002.

206. P. Fearnhead and P. Donnelly. Estimating recombination rates from population genetic data. *Genetics*, 159(3):1299–318, 2001.

207. G. Zou and H. Zhao. Haplotype frequency estimation in the presence of genotyping errors. *Hum Hered*, 56(1-3):131–8, 2003.

208. R. Judson, J. C. Stephens, and A. Windemuth. The predictive power of haplotypes in clinical response. *Pharmacogenomics*, 1(1):15–26, 2000.

209. J. M. Akey, K. Zhang, M. Xiong, and L. Jin. The effect of single nucleotide polymorphism identification strategies on estimates of linkage disequilibrium. *Mol Biol Evol*, 20(2):232–42, 2003.

210. G. Zou and H. Zhao. The impacts of errors in individual genotyping and dna pooling on association studies. *Genet Epidemiol*, 26(1):1–10, 2004.

211. K. M. Kirk and L. R. Cardon. The impact of genotyping error on haplotype reconstruction and frequency estimation. *Eur J Hum Genet*, 10(10):616–22, 2002.

212. C. Lamina, H. Kuchenhoff, J. Chang-Claude, B. Paulweber, H. E. Wichmann, T. Illig, M. R. Hoehe, F. Kronenberg, and I. M. Heid. Haplotype misclassification resulting from statistical reconstruction and genotype error, and its impact on association estimates. *Ann Hum Genet*, 2010.

213. D. Gordon, S. J. Finch, M. Nothnagel, and J. Ott. Power and sample size calculations for case-control genetic association tests when errors are present: application to single nucleotide polymorphisms. *Hum Hered*, 54(1):22–33, 2002.

214. K. R. Ewen, M. Bahlo, S. A. Treloar, D. F. Levinson, B. Mowry, J. W. Barlow, and S. J. Foote. Identification and analysis of error types in high-throughput genotyping. *Am J Hum Genet*, 67 (3):727–36, 2000.

215. N. Liu, R. Bucala, and H. Zhao. Modeling informatively missing genotypes in haplotype analysis. *Commun Stat Theory Methods*, 38(18):3445–3460, 2009.

216. P. A. Gourraud, E. Genin, and A. Cambon-Thomsen. Handling missing values in population data: consequences for maximum likelihood estimation of haplotype frequencies. *Eur J Hum Genet*, 12(10):805–12, 2004.

217. J. Wessel and N.J. Schork. Generalized genomic distance-based regression methodology for multilocus association analysis. *Am J Hum Genet*, 79(5):792–806, 2006.

218. W. Y. Lin and D. J. Schaid. Power comparisons between similarity-based multilocus association methods, logistic regression, and score tests for haplotypes. *Genet Epidemiol*, 33(3):183–97, 2009.

219. R.A. Gibbs, J.W. Belmont, and P. Hardenbol. The international hapmap project. *Nature*, 426 (6968):789–96, 2003.

220. Xie X. and Ott J. Testing linkage disequilibrium between a disease gene and marker loci. *Am J Hum Genet*, 53:1107, 1993.

221. D. C. Crawford and D. A. Nickerson. Definition and clinical importance of haplotypes. *Annu Rev Med*, 56:303–20, 2005.

222. J. Y. Dai, M. Leblanc, N. L. Smith, B. Psaty, and C. Kooperberg. Share: an adaptive algorithm to select the most informative set of snps for candidate genetic association. *Biostatistics*, 10(4): 680–93, 2009.

Índex alfabètic

ADN, 3

al·lel, 4

Bayes, Teorema de, 74

Bayesià, 72

BayHap, 128

Cadena de Markov, 79

cluster, 53

cromosoma, 3

DFARS, 91

EM, 39

Equilibri de Hardy Weinberg, 10

estratificació, 17

estudi, 15

- d'associació genètica, 15
- de cas-control, 20
- de cohort, 21
- de lligament, 16, 20
- transversal, 20
- *Whole-Genome*, 19

fase, 13

fenotip, 4

filogènia, 37

- perfecta, 37

gen, 3

- candidat, 18
- COX2, 191
- DRD2, 153

genoma, 3

genotip, 4

Gibbs Sampling, 42, 86

haplo.stats, 57

haplotip, 13

- cluster d', 48
- incert, 27

HapMap, 25

heterozigot, 4

homozigot, 4

inferència, 34

- Bayesiana, 42

Freqüentista, 41

Linkage Disequilibrium, 11

locus, 4

marcador genètic, 8

MCMC, 42, 77, 80

meiosi, 5

Metropolis, 85

Metropolis-Hastings, 82

mitosi, 5

model de regressió, 50

Lineal, 106

Logístic, 106

Weibull, 109

Monte Carlo, 78

nucleòtid, 3

Odds Ratio, 107

parsimònia, 36

de Clark, 37

pura, 38

polimorfisme, 7

candidat, 18

priori, 73, 74, 112

Random Walk, 85

Recombinació, 7

score, 47, 54

Slice Sampling, 96

SNP, 8

tagSNP, 26

THESIAS, 58

versemblança, 39

haplotípica, 104

Mètode de la màxima, 39

prospectiva, 50

retrospectiva, 51