Gene expression variation and constraint across organs and species

Alessandra Breschi

TESI DOCTORAL UPF / ANY 2016

DIRECTOR DE LA TESI

Prof. Roderic Guigó

DEPARTMENT OF BIOINFORMATICS AND GENOMICS AT
CENTER FOR GENOMIC REGULATION (CRG)

To Annarita

Abstract

Mice are the premier model organisms to study human biology and disease, but there is still debate about the extent to which molecular mechanisms are conserved between human and mouse. With the advance of next-generation sequencing technologies, comparative transcriptomics can be carried out at unprecedented resolution. In this thesis we confirm findings that human and mouse transcriptomes are globally conserved and we identify and characterize the properties of a core set of genes with constrained expression between the two species. Additionally, we show that clustering of gene expression across humans, mice and other vertebrates across multiple tissues depends on which genes and samples are included. Finally, we analyze gene expression of primary cells in human to understand how functional units of organs contribute to the expression of an organ as a whole. Indeed, profiling entire organs constitutes one of the main limitations of current comparative studies.

Resumen

Los ratones son los principales organismos modelos para estudiar la biología y las enfermedades humanas, pero aún está en debate el nivel de conservación molecular entre humanos y ratones. Con el progreso de las tecnologías de secuenciación masiva, la transcriptómica comparativa ha llegado a una resolución sin precedentes. En esta tesis confirmamos que los transcriptomas de humano y ratón están globalmente conservados y identificamos y caracterizamos las propiedades de un conjunto de genes con expresión parecida entre las dos especies. Además, demostramos que diferentes tejidos de humanos, ratones y otros vertebrados se agrupan en base a su expresión génica segón los genes y las muestras incluidas en el análisis. Finalmente, analizamos la expresión génica de líneas celulares primarias humanas para investigar cómo las unidades funcionales de los órganos afectan la expresión de todo un órgano entero. De hecho, los estudios comparativos actuales tienen como limitación que se basan en datos de órganos enteros.

Preface

The laboratory mouse is currently the most clinically relevant model organism to study human biology and diseases. Although decades of mouse research have consolidated the knowledge that many biological processes and molecular mechanisms are conserved with human, a full understanding of the extent of conservation between the two species is still lacking.

Thanks to the recent technological advancements in next-generation sequencing, the mouse ENCODE consortium generated hundreds of functional genomics data in several mouse tissues and cell lines, including RNA-seq, histone marks and transcription factors ChIP-seq and DNase-seq (Chapter 1). At all the molecular levels, there were both similarities and differences between humans and mice. At the transcriptional level, in particular, we showed that many factors should be taken into account when comparing gene expression profiles across several organs in two different species. In fact, we report that the conservation of expression is stronger for some organs which have more tissue-specific genes, such as brain and testis, while it is more attenuated for others, which has fewer distinctive genes (Chapter 1). Consistently, genes with high variation across organs, but relatively low variation between species, have a stronger pattern of expression conservation, while genes with high variation across species, but relatively low across organs, show a lower degree of expression conservation (Chapter 1). This is also the case when additional vertebrate species are included in the analysis (Chapter 3). Moreover, we showed that also the normalization strategy can influence conclusions on the extent of conservation in gene expression programs (Chapter 1 and 3).

On a global scale, however, we observed very correlated levels of expression between the human and mouse, even when very different sample types, are analyzed, such as human cell lines and mouse tissues (Chapter 2). By filtering for expression variation across very heterogeneous samples, we could define a set of about 6,000 genes with constrained expression between humans and mice (Chapter 2).

However, the vast majority of comparative transcriptomics studies has been conducted on whole organs, which are composed of a mixture of very diverse cell types, possibly each one with very distinct expression profiles. In fact, by analyzing RNA-seq data for over 50 human primary cell lines, we showed that cell type specific transcriptional programs can be identified and play a major role in shaping the transcriptional profile of entire organs (Chapter 4). Thus, gene expression differences between the same organs in two or more species, or even within the same species, can be the result of different cellular composition (Chapter 4).

The work presented in this thesis represents a useful contribution to understanding the level of transcriptional similarities between mice and men, and constitutes a valuable resource to study transcriptional variation across species and across organs.

List of publications during the thesis:

1. **Breschi A**, Djebali S, Gillis J, Pervouchine DD, Dobin A, Davis CA, Gingeras TR, Guigó R. Gene-specific patterns of expression variation across organs and species. Genome Biology. 2016 Jul 8;17(1):1.

2. Pervouchine DD*, Djebali S*, **Breschi A***, Davis CA*, Barja PP, Dobin A, Tanzer A, Lagarde J, Zaleski C, See LH, Fastuca M. Enhanced transcriptome maps from multiple mouse tissues reveal evolutionary constraint in gene expression. Nature communications. 2015 Jan 13;6.

3. Lin S, Lin Y, Nery JR, Urich MA, **Breschi A**, Davis CA, Dobin A, Zaleski C, Beer MA, Chapman WC, Gingeras TR. Comparison of the transcriptional landscapes between human and mouse tissues. Proceedings of the National Academy of Sciences. 2014 Dec 2;111(48):17224-9.

4. Cheng Y, Ma Z, Kim BH, Wu W, Cayting P, Boyle AP, Sundaram V, Xing X, Dogan N, Li J, Euskirchen G. Principles of regulatory information conservation between mouse and human. Nature. 2014 Nov 20;515(7527):371-5.

5. Yue F*, Cheng Y*, **Breschi A***, Vierstra J*, Wu W*, Ryba T*, Sandstrom R,* Ma Z*, Davis C*, Pope BD*, Shen Y*. A comparative encyclopedia of DNA elements in the mouse genome. Nature. 2014 Nov 20;515(7527):355-64.

6. Chen L, Kostadima M, Martens JH, Canu G, Garcia SP, Turro E, Downes K, Macaulay IC, Bielczyk-Maczynska E, Coe

S, Farrow S. Transcriptional diversity during lineage commitment of human blood progenitors. Science. 2014 Sep 26;345(6204):1251033.

Contents

INTRODUCTION

Mouse as a model for human biology

The laboratory mouse (*Mus Musculus*) has been for decades the preferred model organism to study human biology and diseases. Humans and mice share a very similar genetic background, with over 90% of both genomes that can be partitioned into regions of conserved synteny (Chinwalla et al., 2002). Although other organisms, such as yeasts, worms and flies are excellent models for studying basic biological processes, mice are far better tools for probing the complex physiological systems that are shared among mammals. Through years of growing experience (Adams and van der Weyden, 2008, Bedell et al., 1997) and technological advances (Singh et al., 2015) to create mutated mouse strains, several mouse models are currently available to mimic many human diseases, even the ones that are not naturally developed in mice.

Thus, mice are exploited in several fields of biology, from neuroscience, to physiology, from behavioural to cancer research. As mice can be housed in small and controlled spaces, very manageable behavioural tests have been creatively devised to reproduce major human behavioural patterns. Examples of application of behavioural tests include studies of anxiety (Schweinfurth and Lang, 2015, Steimer, 2011), substance abuse and addiction (Lynch et al., 2010) and diet (Ellacott et al., 2010). In the context of neuroscience, specific mutant mice are particularly attractive as models of common neurological disorders, such as Alzheimer's disease (Onos et al., 2016), Down syndrom (Rueda et al., 2012) and autism (Silverman et al., 2010).

Mice have also been widely used in experiments related to aging (Vanhooren and Libert, 2013), which is a very complex multifactorial process, where it is crucial to be able to account for one individual factor at a time. Cancer research has benefitted largely from the generation of genetically engineered mice ((Böck et al., 2014)), which shed light on several aspects of tumor biology and profiling, drug response and biomarker discovery ((Cheon and Orsulic, 2011)). However useful mice have been proven to advancing our knowledge and treatment strategies of cancer, their application as xenograft models is still controversial (Aparicio et al., 2015, Morgan, 2012, Richmond and Su, 2008), one of the main concerns being that the mouse response is not always well predictive of the human one. As with mouse models of human tumours, many differences exist, such as the cell duplication time, lifespan and cancer susceptibility, which can affect the experimental use of mouse to study the intricacy of human cancer pathogenesis (Rangarajan and Weinberg, 2003).

It is not surprising that mouse is the most commonly used species for scientific purposes. The most recent official statistics from the European Committee (http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52013DC0859&from=EN) report that just under 11.5 million animals were used Europe in 2011, 60.96% of which were mice. While a UK governmental report shows 1.16 million mice were used in the United Kingdom in 2014, 60% of the 1.93 million experimental procedures completed that year (https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/469508/spanimals14.pdf), with consistent figures from 2005.

Clinical trials rely heavily on non-human organisms, especially mice, before testing a drug on patients, as proven efficiency in *in vivo* preclinical studies is essential for a drug to enter further clinical phases. Nonetheless, as with xenografts, drugs often fail along the phases of clinical trials. A recent work by BioMedTracker, an institutional research service that identifies investment opportunities in the biotech and pharmaceutical industry by assessing the relative strength of companies' clinical drug pipelines, reports that 60% of the drugs between 2003 and 2011 did not proceed to the second phase of testing, while only 10.4% are likely to get FDA approval (Hay et al., 2014). Although few examples of failed clinical trials initiated after successful preclinical studies are mentioned in the literature (Mak et al., 2014), a comprehensive view of such cases is lacking due to incomplete and unorganized reporting of the outcomes.

All this highlights that despite many core biological processes and genetic elements being conserved between human and mouse, other biological features leading to phenotypic differences and poorly correlated physiological responses diverged substantially. Such features can be genomic differences, like retrotransposition events, gene expansions and deaths, genomic rearrangements, or differences in genomic regulation, like gene expression and alternative splicing divergence, active and silenced enhancers, chromatin domains, structural elements, and ultimately differences in protein expression and posttranslational modifications. As the cost and the technical challenges of high-throughput sequencing technologies are continuously decreasing, there has been a growing effort to functionally characterize the human and mouse genomes, and identify what is shared and what diverged. To this end, research consortia started a series of large-scale projects, in a vast array of human and mouse samples, both to understand the principles of genomic regulation across different conditions and to compare it between the two species. These include, but are not limited to, the Genotype-Tissue Expression (GTEx) project (Lonsdale et al., 2013), which aims at establishing a resource database and associated tissue bank to study the relationship between genetic variation and gene expression in human tissues, the Roadmap Epigenomics project (Kundaje et al., 2015), which aims at building a public resource of human epigenomic data, the FANTOM project (Consortium et al., 2014), which focused mostly on CAGE profile of human and mouse tissues and cell lines, and the human and mouse ENCODE projects (Consortium et al., 2012, Yue et al., 2014), whose scope is to catalogue all functional elements in the human and mouse genomes.

In particular, characterizing gene expression profiles across multiple samples and species is really instrumental to reveal to what extent the biology of a given organism can be extrapolated to another, since much of the organism and organ/tissue biology is determined by regulated production of RNA. Thus, I will center this introduction on presenting an overview of the main findings of comparative studies between human and mouse, with a special focus on comparative transcriptomics. I will also review how this comparison has been extended to multiple species. Finally, I will put these in the context of new emerging technologies, in order to increase our knowledge as to when mouse is a good model for human biology.

The human and mouse genomes and gene sets

As a reflection of its importance as a model organism, the mouse was, in the early 2000s, the second mammalian species to have its genome sequenced after human (Chinwalla et al., 2002, Lander et al., 2001, Venter et al., 2001). The most recent genome assemblies (GRC38) include 3.1 Gb and 2.7 Gb for human and mouse, respectively, with the murine genome being 12% smaller than the human one. Over 90% of each genome can be partitioned into conserved syntenic regions and 40% of the nucleotides in human can be aligned to mouse (Chinwalla et al., 2002).

Protein coding genes

According to the latest release of GENCODE annotation (Harrow et al., 2012) (v25, Ensembl85), which recently started to curate also the mouse genome (Mudge and Harrow, 2015) (vM10, Ensembl85), the human genome encodes 58,037 genes, of which about one third are protein coding (19,950) and 198,093 transcripts, compared to 48,440 genes, of which about one half are protein coding (22,021), and 117,667 transcripts encoded in the mouse genome. For both species, the current number of protein coding genes is about ten thousand less than it was estimated from the early drafts of the genome assemblies (Chinwalla et al., 2002, Lander et al., 2001).

The discrepancy in the total number of annotated genes between the two species it is unlikely to reflect underlying biology, and it can be mostly attributed to the less advanced state of the mouse annotation. The number of protein coding and long non coding RNAs encoded in the human and mouse genomes is expected to be very similar, and differences in the total genome length do not obey to differences in the number of genes, but most likely to differences in the lengths of introns and intergenic space (Chinwalla et al., 2002). Indeed, when including predicted gene models from RNA-seq and Cap Analysis of Gene Expression (CAGE) data, the mouse annotation is expanded to a similar size as the human one (Pervouchine et al., 2015).

There is a high degree of gene orthology between human and mouse: 80% of human and 72% of mouse protein coding genes have a one-to-one orthologous relationship in the automatically derived Ensembl Compara (Herrero et al., 2016) (15,874). A number which is highly similar to the 15,736 orthologous genes derived after extensive curation efforts by the ENCODE consortium (Yue et al., 2014).

Long non-coding RNAs

Evidence for the importance of long non-coding RNAs (lncRNAs) is constantly growing and an increasing number of lncRNAs related to diseases is discovered every year (Esteller, 2011, Shi et al., 2013, Wapinski and Chang, 2011). Identifying the possible mouse orthologous of human lncRNAs would largely contribute to uncover their biological role. Currently, there are 15,767 and 9,856 lncRNAs annotated by GENCODE in human and mouse, respectively (Harrow et al., 2012, Mudge and Harrow, 2015) the discrepancy, again, a consequence of the less complete state of the annotation of the mouse genome. In addition, lncRNAs are usually expressed at a lower level than protein coding genes and often in a very tissue-specific fashion, which penalizes a comprehensive annotation (Cabili et al., 2011, Derrien et al., 2012). However, finding orthologous relationships and conservation estimates for lncRNAs is more challenging than for protein coding genes, since their sequence is less conserved (Derrien et al., 2012) and not constrained by amino-acid translation. While RNA secondary structure might be useful to identify short non-coding RNAs and their degree of conservation, only few lncRNAs have distinct structural domains as defined in Rfam (Nawrocki et al., 2014, Pignatelli et al., 2016). Thus, orthology annotation of lncRNAs is still limited as widespread resource (Pignatelli et al., 2016) and the development of methods to identify lncRNA orthology constitutes an active field of investigation.

A number of papers in the past few years attempted to identify novel lncRNAs in mice and other species and call their orthologs in humans (Hezroni et al., 2015, Necsulea et al., 2014, Pervouchine et al., 2015, Washietl et al., 2014). Although the gene sets may vary amongst the different studies, there is a consistent estimate of approximately one to two thousands orthologous lncRNAs between human and mouse. Necsulea et al. (Necsulea et al., 2014) report the highest number of human-mouse orthologous lncRNAs, 2,720, based on sequence similarity of both novel and annotated transcripts, while Washietl et al. (Washietl et al., 2014) identify 1,100 orthologous lncRNAs based on UCSC chain alignments. Pervouchine et al. (Pervouchine et al., 2015) reported 851 lncRNAs orthologs based on a mixed approach including both genome alignments and sequence homology. A more recent study based on *de novo* transcript reconstruction and sequence similarity identifies 813 orthologs (Hezroni et al., 2015). However, the overlap between these sets is quite low: Pervouchine and colleagues (Pervouchine et al., 2015) computed that only 189 orthologous lncRNAs are in common between their study and Necsulea et al. (Necsulea et al., 2014). In all the studies, orthologous lncRNAs still represent a small fraction of all the annotated lncRNAs in both species, especially when compared to protein coding genes. Since most comparative studies have been so far based on orthologous protein coding genes, and lncRNAs are likely to be heavily under annotated, this may have led to an overestimation of the transcriptomic similarities between human and mouse, and overall of the biological similarities between these organisms. Because of poor orthology, the biology encoded in the lncRNA complement is likely to be very different in human and in mouse.

Conservation of human and mouse transcriptomes

Similarities in the gene sets between two species do not necessarily reflect transcriptomic similarities, since the expression pattern of a gene across tissues and conditions may be very different in the two species. Since the early development of microarray technologies, and of RNA-seq later, made possible for the first time the genome wide survey of the transcriptional activity of genes, there has been much interest in understanding to what extent the patterns of gene expression have been globally conserved between human and mouse.

Microarray studies

Most of the early microarray studies focused primarily on the expression of orthologous protein coding genes in a variety of homologous tissues. Under the assumption that mouse is a good model of human biology, we would expect more similarity of expression in the same organs between the species, than in the different organs from the same species (Zheng-Bradley et al., 2010) (e.g. human liver would have an expression profile closer to mouse liver than to human heart). Global transcriptome relationship between multiple RNA samples are usually visually presented using methods related to hierarchical clustering, where samples are the leaves of a tree, which is built based on a given similarity measure between transcriptomes (usually euclidean distance between individual gene expression levels or correlation coefficient across all genes between samples), or to dimensionality reduction, such as principle component analysis (PCA) or multidimensional scaling (MDS), which project samples on a 2-3 dimensional space where their distance is related to their global transcriptome similarity. However, these statistical methods are heavily dependent on the quality of the input data, how much variation there is between and within samples and how the values are distributed. Indeed, the importance of proper filtering and normalization prior to secondary analysis has been very much stressed for microarray data, which are known to be subject to several technical biases; studies who emphasize a proper use of normalization methods report a high conservation of expression between human and mouse (McCall et al., 2011, Zheng-Bradley et al., 2010). Inaccurate normalization, on the other hand, for instance failing to account for species specific systematic bias in signal intensity values in microarray probe sets, has been shown (Liao and Zhang, 2006a) to spuriously exacerbate differences between species (Yanai et al., 2004).

Nonetheless, it is still under debate whether these results, obtained in a limited number of samples, are generally applicable to any type of samples and on the whole transcriptome. As an example, while induction and repression of major transcriptional regulators of erythropoiesis are conserved, at a global level significant extent of transcriptional divergence has been detected between the two species (Pishesha et al., 2014). In another study, it has been shown that murine transcriptional responses to different inflammatory stresses, including trauma, burns and endotoxemia, have a poor correlation with the human ones, although human responses to them are quite similar (Seok et al., 2013), posing serious questions whether mouse is a good clinical model to study such conditions. These conclusion was challenged by a reanalysis of the same data but restricting only on a smaller set of genes with conserved changes between the human and mouse responses (Takao and Miyakawa, 2015). It has been noted that this approach introduces a bias in the results, and that the low percentage of genes with conserved changes (12%) may itself be indicative of poor reproducibility of the human response in mice (Shay et al., 2015, Warren et al., 2015).

RNA-seq and multiple species

The introduction of RNA-seq technology, which allows for more sensitivity, larger dynamic range and annotation-independent detection of RNA abundances (Mortazavi et al., 2008), prompted more comparative transcriptomics studies at a deeper resolution and including larger numbers of species (since RNA-seq does not depend on a species-specific previously spotted microarray surface, see (Romero et al., 2012) and (Necsulea and Kaessmann, 2014) for reviews). A series of early works concluded that transcriptional patterns are more similar between orthologous organs of different species than between different organs from the same species (Barbosa-Morais et al., 2012, Brawand et al., 2011, Merkin et al., 2012). Regarding specifically mouse, the ENCODE consortium (Stamatoyannopoulos et al., 2012) has been collecting around one hundred RNA-seq datasets for a range of mouse tissues and cell types, in order to create a comprehensive reference for future studies (Yue et al., 2014).

As in the case of microarrays, clustering of mouse and human gene expression profiles from homologous tissues strongly depended on the normalization applied (Yue et al., 2014). However, as human data from comparable experimental conditions was not available, since the bulk of human ENCODE transcriptome data was obtained in cell lines (Consortium et al., 2012) and the mouse in primary tissues, and they were sequenced in different labs, it is hard to disentangle the gene expression variation attributable to the species, to other biological factors, or to technical effects (Yue et al., 2014). However, simultaneous analysis of the human and mouse RNA data uncovered a large fraction of orthologous protein coding genes (about 50%) with relatively constrained expression independent from the cell type in both human and mouse (Pervouchine et al., 2015).

Analysis of human and mouse gene expression from a more homogeneous experimental setting, on the other hand, argumented that different conclusions can be drawn depending on which organs are profiled: organs with more distinct signatures of tissue-specific genes, such as brain, testis, heart, liver and kidney show strong conservation between the two species (Chan et al., 2009, Lin et al., 2014, Su et al., 2002, Sudmant et al., 2015). On the other hand, by using a larger panel of organs in the analysis, that include organs expressing less tissue specific genes, Lin et al. (Lin et al., 2014) show that transcriptional patterns have overall diverged substantially between human and mouse, separating the species more than the organs. This conclusion led to another highly charged debate, where other possible factors and biases were taken into account (Gilad and Mizrahi-Man, 2015).

Individual genes, however, exhibit strong differences in their patterns of expression variation. Therefore, rather than trying to quantify the overall transcriptome similarity between two or more species, Breschi et al. attempted to characterize the pattern of expression variation across tissues and species for each gene individually, between human and mouse only (Yue et al., 2014) and across multiple species (Breschi et al., 2016). Thus, a subset of genes was identified that vary a lot across tissues, but little across species, as well as a set of genes that vary a lot across species, but little across tissues (Breschi et al., 2016). Vertebrate (mouse) models of human biology may be particularly appropriate for the genes in the former set (Hardison, 2016). Remarkably, these genes are more likely to be associated with diseases than are genes whose expression varies predominantly across species.

All this raises the more general issue that although some commonalities exist, global responses as a whole might differ, as it is known to be the case for the immune system (Mestas and Hughes, 2004), and that caution should be applied in carefully matching as many factors as possible when mouse models are applied (Shay et al., 2015).

lncRNA expression conservation

Most of the large-scale comparative studies of gene expression are centered on orthologous protein coding genes. Only in the last decade, comparative surveys of non-coding transcriptomes are emerging, as the annotation of lncRNAs is constantly expanding (Harrow et al., 2012, Mudge and Harrow, 2015). Globally, orthologous lncRNAs between human and mouse have conserved levels of expression (Hezroni et al., 2015, Pervouchine et al., 2015). However, clustering analysis and PCA based on lncRNAs show more rapid evolution of expression patterns compared to protein coding genes (Necsulea et al., 2014). In addition, the breadth of expression is also conserved not only between human and mouse but also in other mammals: ubiquitously expressed lncRNAs in human are ubiquitous across all species and tissue-specific lncRNAs in human are tissue specific in all species (Hezroni et al., 2015, Washietl et al., 2014). However, these results may be influenced by the relatively low number of orthologous lncRNAs (less than 10% of annotated lncRNAs) with respect to orthologous protein coding genes (75%). Most lncRNAs appear to be testis-specific in both species (Hezroni et al., 2015, Washietl et al., 2014), especially the less conserved ones (Necsulea et al., 2014). This is hypothetically related to a more permissive chromatin conformation during spermatogenesis (Soumillon et al., 2013), which could potentially contribute to the rapid evolution of testis transcriptomes. Therefore, organ-specific evolutionary rates of gene expression must be considered to evaluate if the mouse transcriptome is a good model of the human transcriptome.

Expression and sequence conservation

A key question in understanding the evolution of gene expression is how it is related to the evolution of sequences and whether conservation of gene expression is reflected in sequence constraints; which regulatory sequences evolved for those genes with a highly conserved transcript sequence but diverging expression patterns. Overall, average gene expression levels are well correlated between human and mouse: highly expressed genes in humans are also highly expressed in mice (Liao and Zhang, 2006b, Wang and Rekaya, 2009), even when very heterogeneous samples are considered (Pervouchine et al., 2015). This is to some extent reflected at the sequence level in the gene body (Koonin and Wolf, 2010, Liao and Zhang, 2006b). On the other hand, promoter sequences have diverged more, although there is some variation in the reported degree of the divergence (Wang and Rekaya, 2009, Weirauch and Hughes, 2010). One possibility is that compensatory mechanisms act on regulatory regions to maintain conserved gene expression (Vakhrusheva et al., 2013, Weirauch and Hughes, 2010).

Comparative gene regulation

Over the past five years, comparative studies have tried to move beyond characterizations of differences in gene expression levels within and between species to studying variation in regulatory mechanisms (Pai and Gilad, 2014). However, the combinatorial complexity of gene regulatory factors, e.g. histone modifications and transcription factors (TFs), and sample types (tissues/cell lines), and the difficulties in associating specific regulatory regions to the regulated genes, which may be distal, makes it really challenging to reach a comprehensive genome-wide map of regulatory elements. Comparative experiments between human and mouse were usually confined to a handful of TFs in a few cell types (Johnson et al., 2009, Kunarso et al., 2010, Odom et al., 2007, Schmidt et al., 2010). Nonetheless, they revealed principles of cis-regulation which were subsequently confirmed by larger studies. The mouse ENCODE consortium has been collecting hundreds of ChIP-seq data of histone modifications and transcription factor (TF) binding site, DNA-seq data for chromatin accessibility sites and replication timing data for chromatin domains for different mouse tissue/cell types (Yue et al., 2014). Although chromatin states inferred from histone modifications (Ernst and Kellis, 2012) and chromatin domains were highly similar between the two species, patterns of transcription factor binding, as measured by ChIP-seq and inferred from DNase hypersensitive sites, are more diverged (Yue et al., 2014).

The primary consensus sequence motif for orthologous TFs is virtually the same in human and mouse (Cheng et al., 2014, Odom et al., 2007), but the secondary motif is often different (Cheng et al., 2014). Thus, the most represented motifs discovered in one species may be used in the other species, with the caution that motif alone is not indicative of actual binding.

Depending on the sample and the TF, between half and two-thirds of the binding sites in one species can be aligned to an homologous sequence on the other species (Cheng et al., 2014, Denas et al., 2015, Vierstra et al., 2014) and widely share the same relative distance to TSS (Cheng et al., 2014). Yet, only 10-20% of the TF-bound sites in one species are also bound in the other species (Cheng et al., 2014, Vierstra et al., 2014). Species-specific binding sites may arise from species-specific innovations or losses. It has been proposed that novel TF binding sites and enhancers may arise from transposition of repeated elements (Bourque et al., 2008, Kunarso et al., 2010, Yue et al., 2014) or by DNA exaptation (Villar et al., 2015). Surprisingly, it has been shown that up to 40% of TF binding sites lost in human, but present in mouse, have an unchanged sequence (CEBPA, (Schmidt et al., 2010)). On the other hand, the loss of TF binding occupancy in aligned regions is half of the times compensated by another active site within 10kb (Schmidt et al., 2010), so the main regulatory circuits of gene regulatory networks are maintained. In addition, TF binding sites in one species are repurposed in the other species; it has been computed that 47-57% of sites that are bound in one species in one sample are bound in the other species in another sample (Denas et al., 2015). In addition binding sites with non conserved occupancy tend to be more tissue-specific and are usually in a non-permissive chromatin state in the species where they are inactive (Cheng et al., 2014).

Taken together these findings suggest that although the relationships TF-target are conserved between human and mouse, the activity of specific regulatory DNA elements, enhancers and TF binding sites, in one species cannot be inferred from sequence homology and consensus motif alone in the other species. In fact, only functional validation experiments can confirm the reliability of the cross-species predictions (Visel et al., 2008). Pennacchio and colleagues (Pennacchio et al., 2006) developed a method to screen for testing in vivo activity of enhancers using transgenic mouse embryos, which also allows to observe their tissue-specificity. During the years, they collected a database with the result for almost 3,000 tested enhancers, orthologous between human and mouse (Visel et al., 2007), as a freely available resource for the scientific community.

Ultimately, enhancers and TF binding sites in mouse can be a good proxy to find functional genomic regions implicated in human traits, for instance related to genome-wide association studies (GWAS) (Welter et al., 2014). Indeed, TF occupied sites conserved between human and mouse harbor significatively more GWAS SNPs compared to background (Cheng et al., 2014). Promisingly, more than four thousands single nucleotide Polymorphisms (SNPs) from human GWAS studies can be mapped uniquely onto the mouse genome. As an encouraging example, SNPs associated to traits related to the liver function, such as HDL cholesterol and alcohol dependence reside in liver-specific mouse enhancers, and SNPs associated traits related to urate levels reside in kidney-specific mouse enhancers (Yue et al., 2014). Thus, mouse could be a useful model to gain better insights into the causality of human GWAS SNPs.

Intraspecies variation in expression in humans and mice

Between 4 and 5 million SNPs differentiate each person from the human reference genome (Consortium et al., 2015a) and a conservative estimate postulates that the genomes of two individuals differ by at least 0.5% (Levy et al., 2007). How this variation impacts molecular features, such as gene expression, and ultimately phenotypes, is currently a topic of active research, especially within consortium-led projects like the Geuvadis (Lappalainen et al., 2013) and the GTEx projects (Consortium et al., 2015b). The major stratification of variation within the human species is at the level of populations, which is strongly related to people's geographic distribution (Consortium et al., 2015a). This leads to relatively small changes in gene expression affecting only 1% of coding and noncoding genes, especially when compared with approximately 10% genes which change their expression as a function of age (Melé et al., 2015).

The concept of interindividual variation in the laboratory mice is less straightforward, since the *Mus musculus* species have multiple layers of stratification due to human intervention. Three major wild subspecies with distinct geographical ranges, *M. m. domesticus*, *M. m. musculus* and *M. m. castaneus*, are the ancestors of most mouse laboratory strains (Wade et al., 2002), while *M. m. molossinus* subspecies resulted from hybridization between *M. m. musculus* and *M. m. castaneus* in Japan (Wade et al., 2002). Laboratory strains can be classified into classical inbred strains and wild-derived strains depending on their origin (Yang et al., 2011). Inbred strains are, by definition, derived after twenty or more consecutive generations of brother-sister matings, which brings to at least 98.6% homozygous loci in each mouse (Beck et al., 2000). Classical inbred strains are mosaics of a handful of haplotypes derived from fancy mice generated from wild subspecies (Wade et al., 2002), with more than 90% of their genetic background coming from *M. m. domesticus* (Yang et al., 2007, 2011). The laboratory mouse which is most commonly experimentally employed and whose DNA was the first mouse DNA sequenced belongs to the strain C57BL/6J (or black-6 where J stands for the center of origin, The Jackson Laboratory) (Chinwalla et al., 2002). It was bred in the early 20s by Clearance C. Little for studies of the genetics of substance preference given its increased preference for alcohol and narcotics (Beck et al., 2000).

To quantify the genetic variation between strains the Mouse Genomes Project sequenced and catalogued a number of classical inbred and wild-derived strains (Keane et al., 2011). Variation within the reference genome strain is negligible as it is virtually indistinguishable from the sequencing error rate (Wade et al., 2002). Also the variation between mice of the same strain but created from different centers, is very low (less than 10,000 SNPs (Keane et al., 2011)), although phenotypic differences in behaviour have been reported (Kiselycznyk and Holmes, 2011, Matsuo et al., 2010). Interstrain variation, instead, is more pronounced, with around 4-5 million SNPs between the mouse reference genome and any other classical inbred strain (Keane et al., 2016, 2011); considering that these SNPs are limited only to the 85% of accessible genomic sequence and that the mouse genome size is smaller than human, this variation is higher than interindividual variation amongst humans. Finally, the mouse reference genome differs from other wild-derived strain by at least 17 million SNPs, with the exception of strains derived from *M. m. domesticus* (Keane et al., 2011).

Comparatively, there is relatively little variation in terms of gene expression both between classical inbred strains (Holgersen et al., 2015, Turk et al., 2004) and within the same strain (Pritchard et al., 2001), in different tissues. These differences are not necessarily related to the diverse genetic background, as many environmental factors (e.g. progressive removal of littermates from the cage) could temporarily alter gene expression profiles of individual mice (Pritchard et al., 2001). Thus, it is really important to select a good mouse population to understand murine intraspecific variation, possibly from outbred wild-caught mice, and compare it to human. The use of inbred strains to uncover relationships between genotype and gene expression is more suited for experiments on allele-specific expression. In hybrid mice between two distinct inbred strains, maternal and paternal genotypes can be readily tracked. In fact, with more than 450 inbred strains (Beck et al., 2000), carefully annotated by the Jackson Laboratory (Bult et al., 2016) (`http://www.informatics.jax.org`), RNA production from only one allele can be easily detected and compared across multiple tissues (Deng et al., 2014, Keane et al., 2011).

Cellular complexity of mammalian organs

A vast proportion of transcriptomics studies in human and mouse, especially the comparative ones, has been mostly focused on profiling gene expression at the organ/tissue level. Thus, organs have been regarded as the functional units of organisms, each one with its own distinct transcriptional pattern. However, organs are composed by an organized mixture of different cell types, whose concerted genomic activity establishes the proper functioning of organs as a whole. Currently, it is unknown how many different cell types compose mammalian organisms. So far, more than 400 human cell types have been classified (Vickaryous and Hall, 2006), based on multiple criteria including morphology and biochemistry. The diverse composition and relative proportion of cell types within an organ can be a potential source for unwanted variation in gene expression between organs and between species. In fact, theoretically, even two distinct samples from the same biopsy, but from different histological sections, can exhibit distinct gene expression profiles, due to the diversity in cell type composition. Clustering analysis revealed that populations of human and mouse primary cells of a given type have distinctive expression profiles (Hume et al., 2010, Mabbott et al., 2013). Therefore, it is extremely important to deconvolute qualitatively and quantitatively which cell populations contribute to the global expression patterns of organs (Lee et al., 2013).

Most transcriptomics studies on mammalian primary cells are based on meta-analyses of mostly microarray data from disparate sources, which, albeit with normalization methods, carries technical noise and reduced sensitivity. The FANTOM consortium released the largest organized atlas of promoter (and gene) expression data (Consortium et al., 2014) in hundreds human and mouse primary cells and tissues. However, to the best of our knowledge, a systematic comparative analysis between the two species is still lacking at the resolution of cell populations. This could shed light on cell-type-specific differences between human and mouse that are masked by the average behaviour of whole organs. For instance, two genes expressed in pancreatic islets in both human and mouse, group-specific component (vitamin D binding protein) GC and DLK1, specific of alpha cells and beta cells, respectively, in human have opposite cell-specific expression in mouse (Li et al., 2016).

Expression data of purified populations of primary cells provide higher resolution than whole tissue transcriptomes and is more robust to stochastic variability between single cells (Saliba et al., 2014). On the other hand, with the recent advancements in single-cell transcriptomics (Kolodziejczyk et al., 2015, Macosko et al., 2015), single-cell RNA-seq allows to obtain gene expression data for rare cell types, whose signal is usually masked at the population level, to identify novel cell types with previously unknown markers, and to characterize cell differentiation stages (Trapnell, 2015). Due to noticeable experimental challenges in disaggregating solid tissues, especially in human, most single-cell RNA-seq research focused on mouse solid tissues, including brain (Zeisel et al., 2015), lung (Treutlein et al., 2014), intestine (Grün et al., 2015), while fewer studies analyzed human samples from pancreatic islets (Li et al., 2016), brain (Darmanis et al., 2015) and blood (Jaitin et al., 2014, Paul et al., 2015). Additionally, single-cell RNA-seq has been applied to investigate RNA dynamics over time, especially in the early stages of life, just days after fertilization (Ohnishi et al., 2014, Scialdone et al., 2016).

Notwithstanding the growing bulk of projects employing single-cell RNA-seq, as with cell population data, very few compare human and murine single-cell expression. Possibly, one complication being the intrinsic difficulty of obtaining comparable samples from homologous organs or identifying homologous dynamic processes. Xue and colleagues compared the genetic programs of human and mouse early embryos, in the developmental stages between oocytes and morula, and observed that while global gene expression profiles are conserved, the actual timing differs between the two species (Xue et al., 2013). Eventually, comparing human and mouse transcriptomes at the single-cell level can help identifying previously undescribed conserved cell types, overcome the biases of different cell type composition and understand conserved and diverged elements of temporal dynamics. Albeit promising, this will possibly require the development of specific computational methods which would deal with the complexity of single-cell data and integrate it with the additional dimension of cross-species comparison.

CHAPTER 1

Encyclopedia of mouse DNA elements

Similar to the ENCODE project, the Mouse ENCODE project aims at functionally characterize and annotate the mouse genome. A comparative approach has been pursued to identify common and diverged functional elements between the two species. To this end, the Mouse ENCODE consortium is collecting a huge amount of Next-Generation Sequencing data (over 1,000 data sets), in several mouse tissues and cell types, including RNA-seq, ChIP-seq of chromatin marks and transcription factors, replication and DNAse-seq. The integrative analysis of gene expression profiles, chromatin status and trans-acting regulators revealed a variable landscape of conservation. Although gene expression is largely conserved, some genes have dissimilar profiles, and while chromatin states and transcription factor networks are relatively stable between the two species, cis-regulatory sequences seem to be less evolutionarily constrained.

My main contribution has been the analysis of transcriptomics data, which showed that clustering of human and mouse homologous organs is heavily dependent on the normalization method and on the set of genes. In a separate paper, to which again I contributed with clustering analysis of human and mouse homologous organs from a different datasets, it is shown that clustering is also dependent on the organs; organs with the highest number of tissue-specific genes drive an organ-dominated clustering and their expression profiles are the most conserved between the two species

Yue F, Cheng Y, Breschi A, Vierstra J, Wu W, Ryba T, et al. A comparative encyclopedia of DNA elements in the mouse genome. Nature. 2014 Nov 20;515(7527):355–64. DOI: 10.1038/nature13992

Lin S, Lin Y, Nery JR, Urich MA, Breschi A, Davis CA, et al. Comparison of the transcriptional landscapes between human and mouse tissues. Proc Natl Acad Sci U S A. 2014 Dec 2;111(48):17224–9. DOI: 10.1073/pnas.1413624111

CHAPTER 2

Transcriptional comparison of human and mouse genomes

Although some genes show different expression profiles, a s discussed in Chapter 1, human and murine transcriptomes are largely conserved. By comparing RNA-seq data from mouse tissues and human cell lines, we identified a core set of constrained genes which exhibit conserved expression patterns, even between such heterogeneous sample types as tissues and cell lines. Naturally, this set of genes constitutes a substantial fraction of the total RNA production and are involved in housekeeping processes common to all cell types. We also showed that these genes are associated to strong phenotypes and, consistently, to constrained epigenetic marking. In addition, the outstanding sequencing depth of the mouse dataset allowed us to detect novel transcripts and genes, thus enriching the current annotation status.

My main contribution to this work has been the analysis of novel transcripts, conservation of gene expression and antisense transcription and functional characterization of genes with constrained expression and their epigenetics regulation.

Pervouchine DD, Djebali S, Breschi A, Davis CA, Barja PP, Dobin A, et al. Enhanced transcriptome maps from multiple mouse tissues reveal evolutionary constraint in gene expression. Nat Commun. 2015 Jan 13;6(1):5903. DOI: 10.1038/ncomms6903

CHAPTER 3

Gene-specific patterns of expression variation across species and organs

A highly charged debate exists as to what extent transcriptomes are conserved between humans and mice across multiple organs. In Chapter 1 it is shown that different sets of genes and different sets of samples have distinct weights to answer this question. Here, we applied variance decomposition of gene expression, similarly to Chapter 1, to a previously published dataset including RNA-seq experiments from multiple organs across six mammals and chicken. We show that genes with high variability in expression across tissues, but low variability across species, drive to a more tissue-dominated clustering, while genes with high expression variability across species, but low across tissues, drive to a more species-dominated clustering. Genes from the former set, which are more conserved amongst species, might be more suitable to extrapolate experimental data from mouse to human (Hardison, 2016).

.

Breschi A, Djebali S, Gillis J, Pervouchine DD, Dobin A, Davis CA, et al. Gene-specific patterns of expression variation across organs and species. Genome Biol. 2016 Dec 8;17(1):151. DOI: 10.1186/s13059-016-1008-y

CHAPTER 4

Conserved transcriptional programs in human primary cells

Mammalian organs are composed of a very heterogeneous mixture of several cell types. In this chapter, we present the analysis of RNA-seq data from 53 human primary cell lines of different types and isolated from various anatomical locations. We identify four major transcriptional programs which define four major types of cells: endothelial, epithelial, melanocytes, and mixed cells, which include fibroblasts, smooth muscle cells and mesenchymal stem cells. We found 2,873 genes, including 48 very correlated transcription factors, that distinguish these major programs. With the hypothesis that different proportions of these cell types, and consequently of their expression signatures, contribute to shape the expression profile of complex organs, we built a linear model which deconvolutes cellular composition from whole organ expression data. The model shows that each organ has its own distinctive cellular composition and that the same organ can have different cellular compositions, depending on the histological section from which RNA was extracted, which is reflected on its gene expression profile.

Breschi A, Davis CA, Djebali S, Pervouchine DD, Gillis J, Dobin A, Lagarde J, Vlasova A, Gingeras TR, Guigó R. The molecular anatomy of the human body (in preparation).

The molecular anatomy of the human body

Alessandra Breschi[1,2], Carrie A. Davis[3], Sarah Djebali[1,2,4], Dmitri D. Pervouchine[1,2], Jesse Gillis[3], Alex Dobin[3], Julien Lagarde[1,2], Alexandre Esteban[1,2], Anna Vlasova[1,2], Thomas R. Gingeras[*3] and Roderic Guigó[*1,2]

[1]Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Dr. Aiguader 88, 08003 Barcelona, Spain
[2]Universitat Pompeu Fabra (UPF), Barcelona, Spain
[4]GenPhySE, Université de Toulouse, INRA, INPT, INP-ENVT, Castanet Tolosan, France
[3]Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11742

[*] Correspondence should be addressed to E-mail: roderic.guigo@crg.cat (Roderic Guigó) and gingeras@cshl.edu (Thomas R. Gingeras)

Abstract

The adult human body is composed of billions of cells belonging to a yet undefined number of cellular types and subtypes. How the transcriptional profiles of individual cell types are regulated to achieve organ complexity is still under investigation. Here, we present the analysis of RNA-seq data from 53 human primary cell lines extracted from different anatomical regions. We identify four major transcriptional signatures that clearly distinguish endothelial cells, epithelial cells, melanocytes and a mixed set of cells, including fibroblasts, smooth muscle cells and mesenchymal stem cells. About 3,000 genes show distinct patterns of expression between these major cell types. We found that the cell topological origin has little impact on the transcriptome compared to the cell type and there is little overlap between cell type-specific genes and tissue-specific genes. By analysing the expression of cell type specific genes in whole tissue sections from the GTEx consortium, we were able to detect sample-specific differences in cellular composition due to different histological cuts. This study represents the larger collection of RNA-seq of human primary cells to date and it constitutes a great resource to understand organ complexity at the cellular level.

Regulated production and processing of RNA underlies cellular differentiation, and defines cell phenotype and function. Massively parallel sequencing technologies, as well as earlier DNA microarrays, have been extensively used to monitor the transcriptomes of tissues and cell lines, across multiple individuals (healthy and diseased) and species [1–3]. Transcriptomes are usually monitored from RNA extracted from samples including millions of cells ($10^5 - 10^7$). Because tissues and organs are highly heterogeneous structures made from cells of dozens of primary cell types (i.e. populations of cells with common phenotype and biological function), the transcriptomes obtained from tissue samples represent only the average behavior of genes across these heterogeneous cellular collections. A human heart, for instance, is composed of at least seven major cell types [4], while a human lung is composed of at least five major cell types [5], and more cell sub-types exist depending on their differentiation status and anatomical sublocation [6]. Changes in gene expression between tissue samples (for instance, healthy and diseased), therefore, may not necessarily reflect changes in the transcriptional activity of genes in specific cell types, but changes in the cellular composition of the tissues [7].

Recent technological advances in microfluidics and others have made possible the isolation of individual cells from which RNA can be extracted and sequenced. Single-cell RNA sequencing, mostly so far on hematopoietic derived cell populations [8], has revealed larger cellular heterogeneity than expected, leading to the identification of previously unknown cellular types [9–11]. Because of the large number of cells (e.g. 10^11 cells in liver [12]), and of the substantial stress to which they are subjected during organ disaggregation, single-cell RNA-seq studies of solid human tissues are scarce. Moreover, it is unclear to what extent the biological function is exerted by individual cells or by groups of functionally equivalent cells working in concert. If the latter, the consensus transcriptome of primary cells may be the most biologically relevant.

Here, we monitored by RNA-seq the transcriptome of 53 human primary cell lines from multiple body locations and embryological origins (Fig. 1A, Table S1). Primary cells were obtained from PromoCell (`http://www.promocell.com/`) and ScienCell (`http://www.sciencellonline.com/`) providers, and isolated using FACS analysis of surface markers or morphological and functional assays. Total long RNA ($>$200bp) was ribo-depleted, and deeply sequenced in two bio-replicates using Illumina to an average of 156 million paired-end 101 bp reads per replicate. Reads were mapped to the human genome hg19 using STAR [13] and mapped reads were used by Flux Capacitor [14] (Supplementary Information, Table S1) to infer quantification of the genes and transcripts annotated in GENCODE v19 [15]. We believe that the transcriptome of the primary cells monitored here is a key resource to understand tissue biology, serving as interface between tissue and single cell transcriptomes.

We clustered the primary cells based on the filtered expression of protein coding genes (14,475 genes), long non-coding RNAs (lncRNAs, 1,618 genes) and pseudogenes (1,347 genes, Fig. 1B, Supplementary Information). Primary cells do not cluster by body location (or embryological origin), but form quite well defined clusters depending on the broader cell type to which they belong: endothelial, epithelial, melanocytes, and a mixed set of cell types including fibroblasts and mesenchymal and muscle cells. Modularity analysis quantitatively supports the clustering by cell types over body location (or embryological layer, Fig. 1C, Supplementary Information). We recapitulate essentially the same clustering using gene expression quantifications obtained in 107, partially unrelated, primary cells using Cap Analysis of Gene Expression (CAGE)-tags by the FANTOM project [16] (Fig. 1D, Supplementary Information).

These results suggest that cellular transcriptomes adhere to a few basic transcriptional programs, or, in other words, that not any transcriptome is compatible with cellular life. These programs are refined and modulated in the primary cells that constitute organs and tissues, conferring extraordinary cellular heterogeneity. Body location may partially contribute to this modulation, but it is not the primary determinant of the transcriptome of the primary cells. Using linear models we found that the body location explained on average only 4% of the total variance in gene expression across primary cells (Fig. S1).

We identified 2,873 genes that were specific to each of the four transcriptional programs or cell types above (Fig. 2A, Table S2, Fig. S2). The functions of these genes closely match the expected biology of the broad cell types (Fig. S3). As expected, since most tissues include cells from all cell types, there is little overlap between cell type specific genes and tissue specific genes, as inferred from data from the GTEx project (Fig. 2B, Tables S3-4, Supplementary Information). Actually, many cell type specific genes are found expressed in many tissues (Fig. 2C). That tissues are made of cells belonging to a reduced number of common transcriptional programs may explain the finding that there are very few genes that are exclusive to a given tissue [2].

Among cell type specific genes, we identified 156 Transcription Factors (TFs, out of a total of 1,544 TFs annotated in the human genome [17]), 48 of which show strong co-expression patterns (Fig. 2D). These TFs include known cell type-specific transcriptional regulators, such as ERG, which was shown to regulate endothelial cell differentiation [18], and TP63, which is an established regulator of epithelial cell fate and is often altered in tumor cells [19]

The cellular RNA content is the product of both primary RNA transcription and subsequent post-processing to mature functional RNAs—the species mostly monitored through RNA-seq. Among the post-processing steps, splicing is often assumed to play a fundamental role in defining cellular type. However, among cell type specific genes, there is strong depletion of RNA binding proteins (only 45 out of 1,541 [20]). To further disentangle the relative contribution of transcriptional versus post-transcriptional regulation in defining cell type, we estimated for each gene the proportion of the variation in isoform abundance between cell types than can be simply explained by variation in gene expression [21], and found this to be on average 67% (Fig. 2E). In contrast, considering all samples together, we found that variation in gene expression explains only 55% of the variation of isoform abundance between primary cells. All these results strongly suggest that regulation of primary transcription plays the fundamental role in defining the broad transcriptional programs within the human body, while post-transcriptional regulation plays a comparatively more important role in refining the specific cellular transcriptomes of primary cells within each program. In further support of this, we found that the proportion of differentially expressed genes compared to differentially included exons is significantly larger when comparing primary cells within cell types, than when comparing primary cells between cell types (8.2 vs 7.0, p-value=1.159e-39, Mann-Whitney test, Fig. 2F, Supplementary Information). Cell type specific genes belong mostly, as expected, to the class of genes with unconstrained expression as defined in Pervouchine et al. [22] (Fig. 3A). There are notable differences in the patterns of evolution, both of sequence and expression, between cell type specific genes and protein coding genes overall, and between cell type specific genes themselves depending on the type. First, cell type specific genes are relatively ancient within vertebrate evolution compared to protein coding genes overall (Fig. 3B). Indeed, a larger proportion of them were present in amphibians and birds than protein coding genes overall (70-75% compared to 60%). Epithelial specific genes are an exception, since comparatively many of them appear to arise early in mammalian evolution. In contrast, the expression of cell type specific genes is overall less conserved than that of protein coding genes (Fig. 3C). Epithelial specific genes show the lowest conservation of expression.

Since our results suggest that most cells within a particular tissue belong to one of the basic transcriptional programs uncovered here, we used the patterns of expression of cell type specific genes to infer the broad cellular composition of human tissues using the GTEx tissue transcriptome data. We employed constrained linear models to infer the proportion of melanocytes, epithelial, endothelial and mixed cell types in each tissue sample based on the expression of cell type specific genes in the sample (Fig. 4A, Supplementary Information). The estimated proportions reflect the known cellular composition of the tissues (Fig. 4B). We clustered the GTEx samples according to the estimated proportion of cell types. The clustering (which is based in only five values per sample), recapitulates tissue type as strongly as clustering based on gene expression, according to modularity analysis (Fig. S4).

Among all tissues, stomach shows a clear bimodality in the proportion of epithelial and mixed cell types, suggesting sample heterogeneity (Fig. 4C). The GTEx stomach samples are all from the gastric body, the walls of which consist of two broad layers, the mucosa (and submucosa), which is mostly epithelial, and the muscularis, which is mostly smooth muscle tissue (Fig. 4D). GTEx stomach samples do not include these two layers in a consistent proportion, being dominated either by the mucosa or by the muscularis. We used the histological images of the stomach samples, to score them as mostly muscularis or mostly mucosa (Fig. 4D, Supplementary Information). This restored the transcriptome unimodality within each set of samples (Fig. 4C). Then we investigated the patterns of gene expression of the cell type specific genes. We first used the projection score to identify the genes that most contribute to the separation of the samples (this is an unsupervised method, which does not take into account the inferred proportion of cell types in each sample). We observed that, among these genes, epithelial specific genes were exclusively expressed in the mucosa and mixed type specific genes were exclusively expressed in muscularis (Fig. 4E). Even though the cell type specific genes were obtained from the transcriptomes of primary cells, none of which originated from the stomach, their expression pattern when independently measured in heterogeneous tissue samples, such as stomach in GTEx, is indicative of the underlying cellular composition of the tissues. This constitutes a strong validation of our hypothesis that tissues are generally conglomerates of primary cells belonging to a few basic cell types.

The data collected here represents the largest collection to date of RNA-seq experiments in human primary cells. We believe this constitutes a great resource for the scientific community as it is at an intermediate resolution of complexity between single cells and whole organs. We identify a set of about three thousand genes whose expression define cell type specific signatures. Finally, we have shown that transcriptional profiles of complex organs can be reconstructed by the expression levels of cell type specific genes. Extending the variety of profiled cell types and integrating expression data with epigenetics data from matching samples will certainly enrich our understanding of how different cells are regulated to shape the fascinating complexity of entire organs.

Acknowledgements

This project was supported by awards U54HG007004 and U41HG007234 from the National Human Genome Research Institute of the National Institutes of Health, as well as from the Spanish Ministry of Economy and Competitiveness, Centro de Excelencia Severo Ochoa 2013–2017, SEV-2012-0208, and Programa de Ayudas FPI del Ministerio de Economia y Competitividad, BES-2012-055848. We would also like to acknowledge support from the European Research Council (ERC) under the European Union’s Seventh Framework Programme (FP7/2007-2013)/ERC under the European Union’s Seventh Framework Programme (FP7/2007-2013)/ERC grant agreement 294653.

References

[1] M. Lukk, *et al.*, *Nature biotechnology* **28**, 322 (2010).

[2] M. Melé, *et al.*, *Science* **348**, 660 (2015).

[3] S. Djebali, *et al.*, *Nature* **489**, 101 (2012).

[4] M. Xin, E. N. Olson, R. Bassel-Duby, *Nature reviews Molecular cell biology* **14**, 529 (2013).

[5] J. D. Crapo, B. E. Barry, P. Gehr, M. Bachofen, E. R. Weibel, *American Review of Respiratory Disease* **126**, 332 (1982).

[6] B. Treutlein, *et al.*, *Nature* **509**, 371 (2014).

[7] F. Weerkamp, *et al.*, *Journal of Allergy and Clinical Immunology* **115**, 834 (2005).

[8] F. Paul, *et al.*, *Cell* **163**, 1663 (2015).

[9] J. Li, *et al.*, *EMBO reports* **17**, 178 (2016).

[10] C. Trapnell, *Genome research* **25**, 1491 (2015).

[11] B. B. Lake, *et al.*, *Science* **352**, 1586 (2016).

[12] E. Bianconi, *et al.*, *Annals of human biology* **40**, 463 (2013).

[13] A. Dobin, *et al.*, *Bioinformatics* **29**, 15 (2013).

[14] S. B. Montgomery, *et al.*, *Nature* **464**, 773 (2010).

[15] J. Harrow, *et al.*, *Genome research* **22**, 1760 (2012).

[16] T. F. Consortium, *et al.*, *Nature* **507**, 462 (2014).

[17] H.-M. Zhang, *et al.*, *Nucleic acids research* **40**, D144 (2012).

[18] F. McLaughlin, *et al.*, *Blood* **98**, 3332 (2001).

[19] K. Yoh, R. Prywes, *Frontiers in endocrinology* **6**, 51 (2015).

[20] S. Gerstberger, M. Hafner, T. Tuschl, *Nature Reviews Genetics* **15**, 829 (2014).

[21] M. Gonzàlez-Porta, M. Calvo, M. Sammeth, R. Guigó, *Genome research* **22**, 528 (2012).

[22] D. Pervouchine, *et al.*, *Nat Commun* (2015).

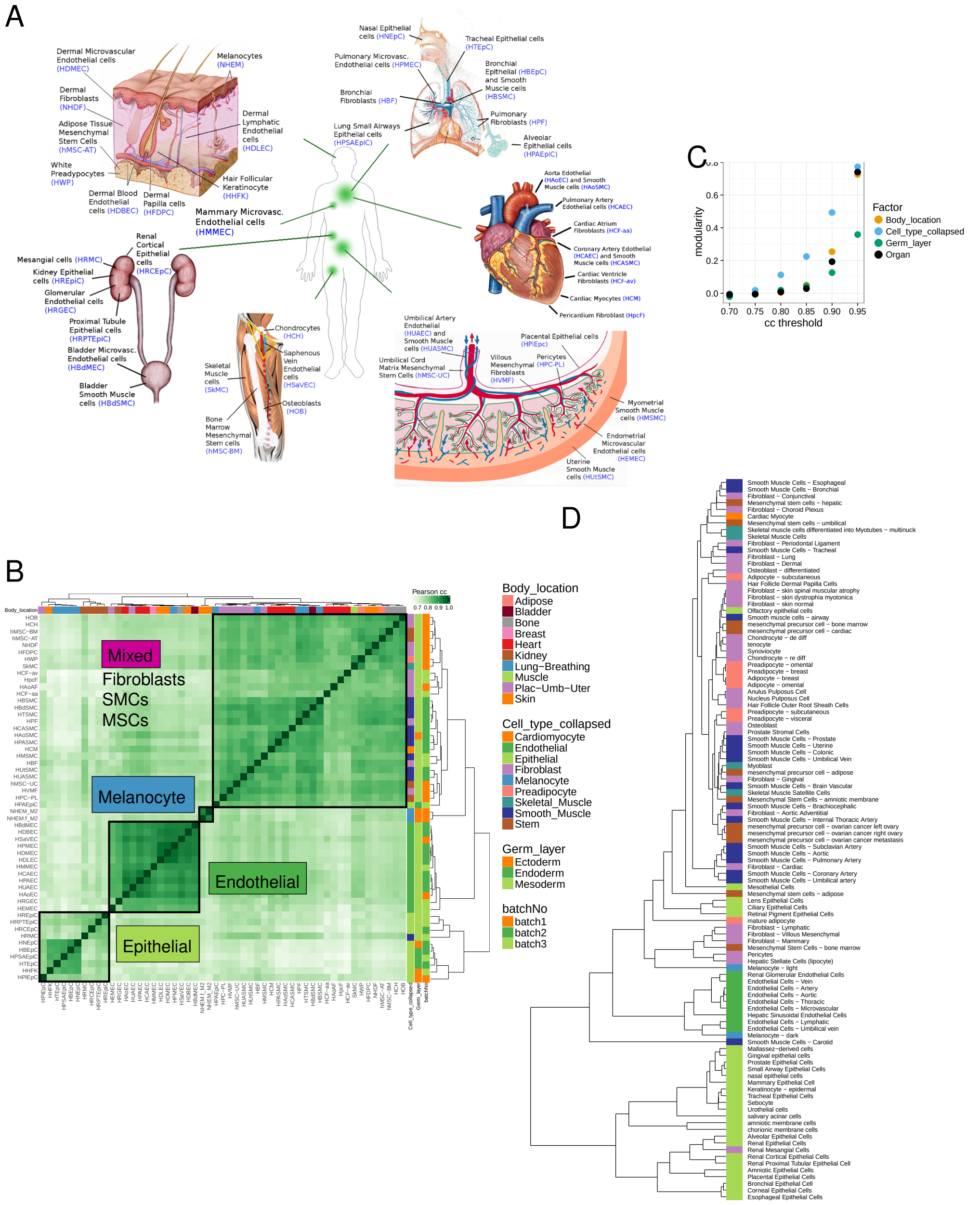


Fig. 1 Transcriptional programs of human primary cells. (**A**) Overview of sequenced primary cells. (**B**) Hierarchical clustering of human primary cells based on gene expression (**C**) Modularity analysis for the network of gene expression correlation. The modularity is given as a function of increasing thresholds of Pearsońcorrelation coefficients between pairs of samples and it is computed for the cell types, the body location, the organ, and the germ layer (**D**) Hierarchical clustering based on CAGE data from FANTOM consortium

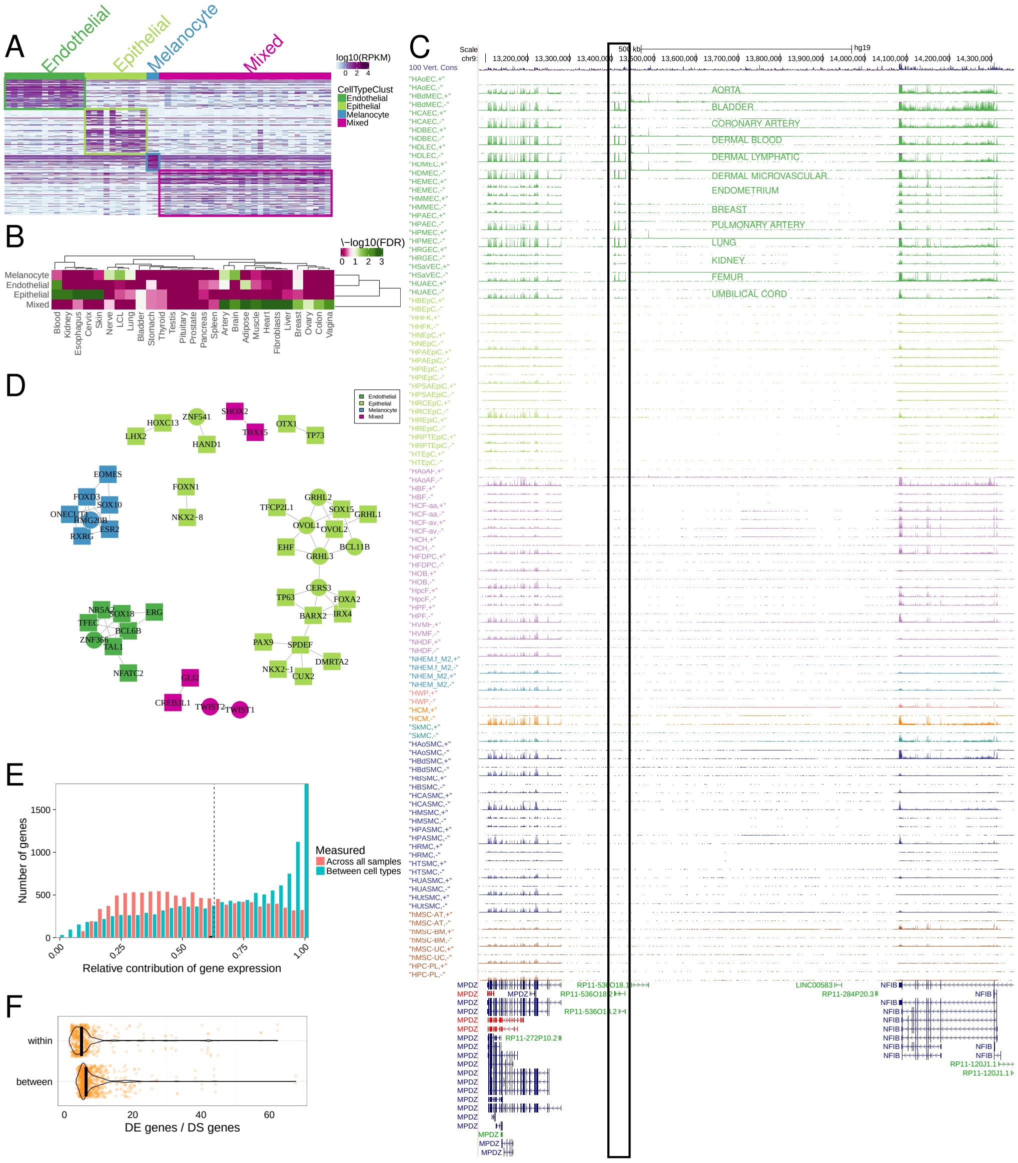


Fig. 2 Cell-cluster-specific genes. (**A**) Expression of 2,873 cell-cluster-specific genes. (**B**) Intersection of cell-cluster-specific genes and tissue-specific genes from GTEx. The significance of the overlap is calculated with hypergeometric test. Negative log10-transformed p-values corrected for multiple testing are shown. (**C**) Expression signal for endothelial-specific lncRNA RP11-536O18.1. (**D**) Network of co-expressed transcription factors. Nodes are colored according to the cell-type-specificity of the TF, and shaped based on the availability of sequence motif (square: available, circle: not available). (**E**) Relative contribution of gene expression to changes in isoform abundance between cell type clusters or all samples. (**F**) Ratio of number of differentially expressed genes (DE) over number of differentially spliced genes (DS), between pairs of samples of the same cell type (within) or different (between).

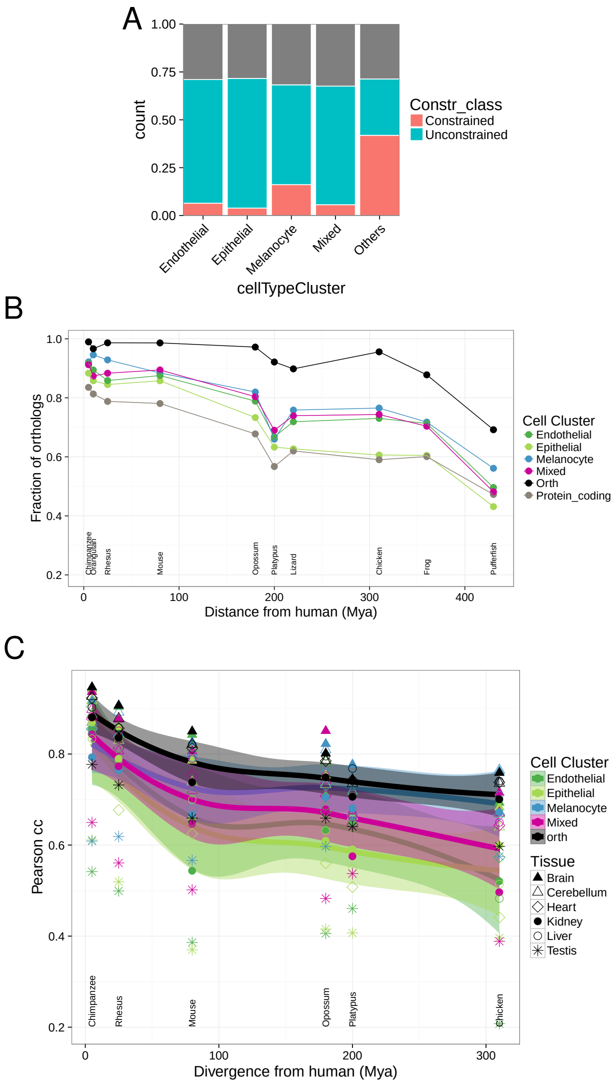


Fig. 3 Sequence and expression conservation of cell type cluster-specific genes. (**A**) Overlap of cell-cluster-specific genes with constrained genes as defined in Pervouchine et al. (**B**) Fraction of one-to-one orthologs between each species and human for the different sets of genes. (**C**) Pearsonś correlation coefficient between gene expression in each human organ and the corresponding one in every other species. The correlation is computed across all the genes in each class separately.

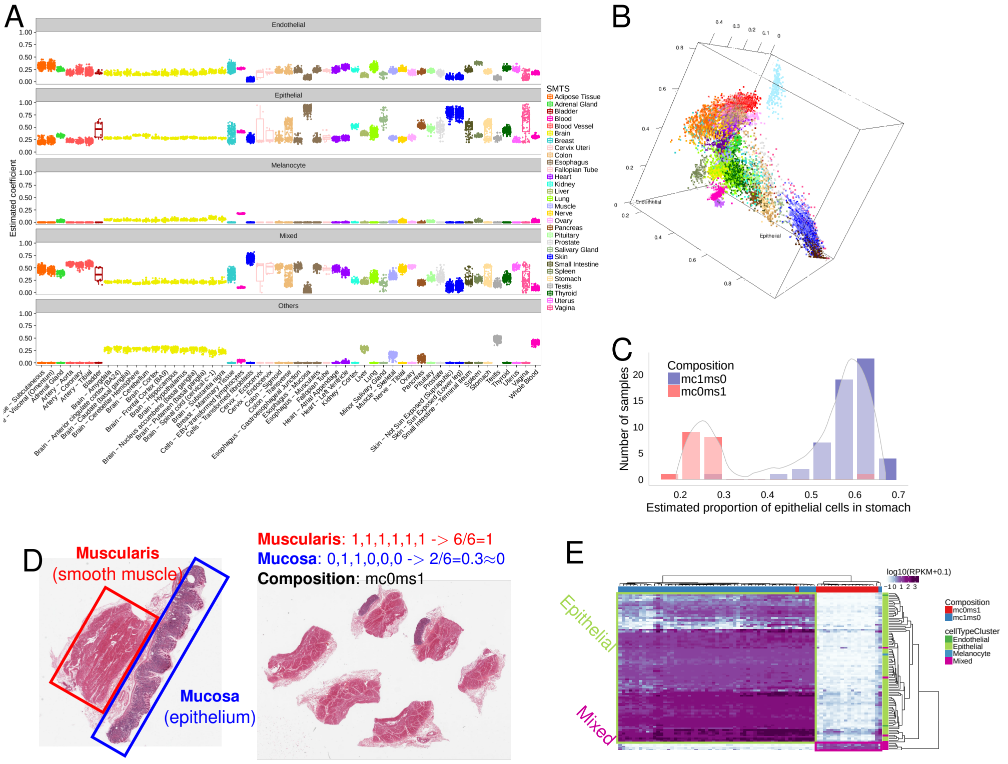


Fig. 4 Expression of cell type cluster-specific genes in GTEx organs. (**A**) Proportions of each cell type, estimated with a constrained linear model of genes expression for each organ in GTEx. An artificial cell type ”Others” was introduced to account for tissue-specific cell types which are not present in this dataset. (**B**) GTEx samples represented in a 3D space where the axes are the estimated proportions of endothelial, epithelial and mixed cells. (**C**) Estimated proportion of epithelial cells in all stomach samples (curve) and specifically in the ones classified as only mucosa (blue bars) or only muscolaris (red bars). (**D**) Example of stomach histological slides which represent the two main tissue layers and the procedure for the manual annotation of the images based on the presence of those layers. (**E**) Expression of cell type-specific genes in stomach samples clearly distinguishes the presence of one or the other tissue layer.

Supplementary Tables and Figures

Supplementary tables S1-3 can be found at:
`http://public-docs.crg.es/rguigo/Data/abreschi/ENCODE/prCells/subm/SupplTables.xlsx`

Table S1. List of ENCODE samples used in this study, with ENCODE ids and GEO ids. Although for some the GEO id is not available, all the data can be accessed through the ENCODE portal `http://www.encodeproject.org`.

Table S2. List of cell type cluster-specific genes.

Table S3. List of randomly selected pairs of GTEx samples. Ids are SRA run ids.

Table S4. List of tissue-specific genes derived from GTEx samples.

Fig. S1. Estimation of proportion of variance explained by the factors body location, cell type, cell cluster, germ layer and organ.

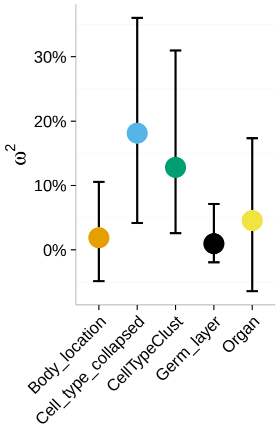


Fig. S2 Distribution of 2,873 cell type cluster specific genes by cell type and gene biotype.

	Endothelial	Epithelial	Melanocyte	Mixed	Total
lncRNA	67	59	46	153	325
protein_coding	531	857	294	729	2411
pseudogene	37	34	13	53	137
Total	635	950	353	935	2873

Fig. S3 GO term enrichments for cell type specific genes. Only the 10 most significant terms for each cell type are shown.

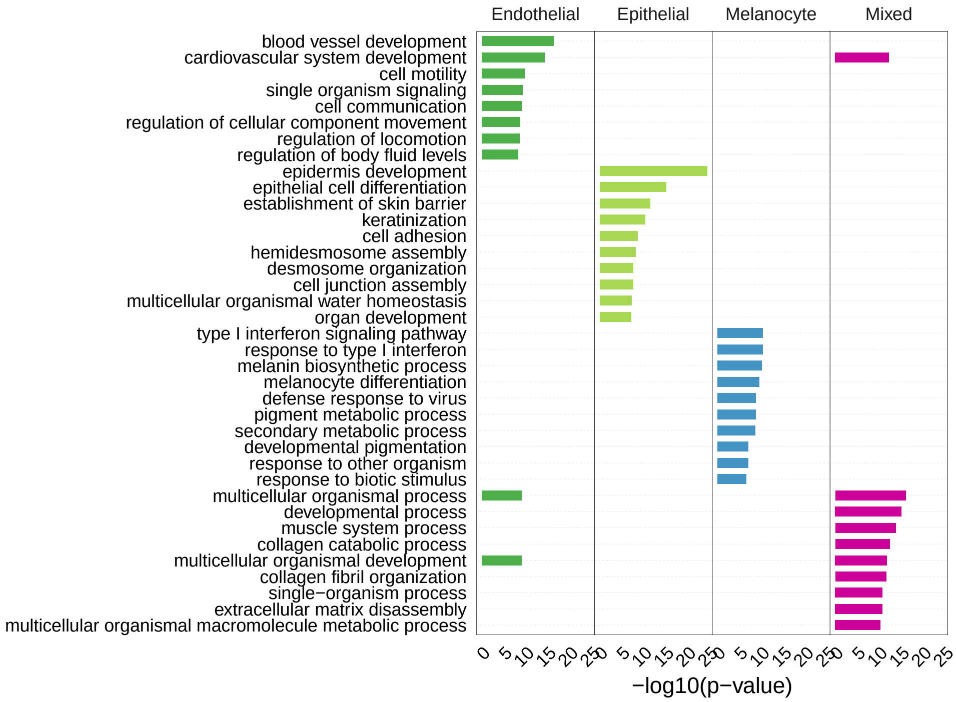
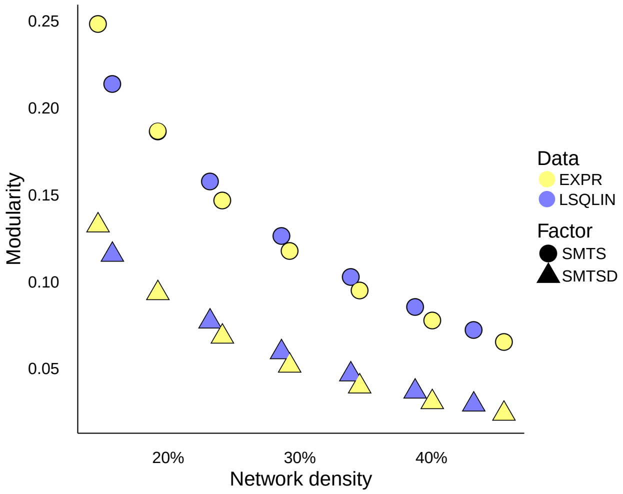


Fig. S4 Network modularity of GTEx samples. The network of GTEx samples is created for increasing network densities, which depend on different thresholds of pairwise Pearson’s correlation coefficients. Network densities are measured as the percentage of edges over the total number of possible edges. Correlation coefficients are computed over gene expression (EXPR) or over the estimated coefficients from the constrained linear model (LSQLIN). The network modularity is computed both with respect to organs (SMTS) or to organ subregions (SMTSD).



Supplementary Information

1 RNA isolation, Library Construction and Sequencing

For each cell type to be made into a library we obtained cell pellets that were stored in RNAlater (Thermofisher) as catalogue items from PromoCell. We ordered 3 vials per cell type per donor for a total of 3 million cells. The 3 vials were combined together and we isolated Total RNA from them using the Ambion mirVana miRNA Isolation kit (cat #AM1561). The rRNA was removed using the RiboZero Gold Protocol (cat #RZG1224). The libraries are made using a homebrew “dUTP” protocol per PMC2764448, which generates stranded libraries. They were sequenced on the Illumina platform in mate-pair fashion and processed though the data processing pipeline at the ENCODE DCC. Additional, information about each of these steps, metadata and files can be found at: `https://www.encodeproject.org/`.

2 RNA-seq processing pipeline

Raw reads from the 106 RNA-seq libraries (see Table S1 for a list of ENCODE library ids and https://www.encodeproject.org for submitted fastq files) were aligned with STAR v2.3.1z [1] to the human genome assembly hg19. Reads mapping to more than 20 multiple positions were discarded. Read counts for all long genes annotated in GENCODE v19 [2] were computed with Flux Capacitor v1.6.1 [3]. Since for most of the analyses we average expression values for a given pair of replicates and sometimes the two biological replicates are from donors of opposite sex, we remove genes on chromosome Y. The lack of an enrichment step for polyadenylated transcripts preserves the presence of some short biotype genes, which are still longer than 200bp. Thus, we remove genes with at least one transcript annotated as short RNA in GENCODE (16 genes, ENSG00000243819.3, ENSG00000270141.2, ENSG00000249352.3, ENSG00000270123.2, ENSG00000228439.3, ENSG00000253143.2, ENSG00000251867.2, ENSG00000254144.2, ENSG00000269900.2, ENSG00000259001.2, ENSG00000258486.2, ENSG00000261519.2, ENSG00000260682.2, ENSG00000264932.2, ENSG00000225978.2, ENSG00000232512.2). These genes are often of repetitive nature which makes the quantification of their expression problematic, this is why we decided to remove them. Read counts which are not reproducible between two replicates (npIDR$>$0.1 [4]) are set to 0. After filtering for reproducibility, read counts are normalized to a slightly modified version of RPKM (reads per kilobase of exon model per million mapped reads [5]). Specifically, read counts were first normalized to cpm (counts per million), where the library sizes are the TMM (trimmed mean of M values [6]) scaled sums of exonic reads, and then normalized by gene length. Finally, RPKM values from the two replicates were averaged, and genes with RPKM$<$1 in all samples were discarded, resulting in 17,440 genes, including 14,475 protein coding, 1,618 long non-coding RNAs and 1,347 pseudogenes. As the samples were prepared and sequenced in three known distinct batches, we used the removeBatchEffect() function from R limma package [7] to build a linear model with the batch information and the cell types on log10-transformed RPKM (with a pseudocount of 0.01), and we regressed out the batch variable.

3 Gene expression analyses

3.1 Hierarchical clustering

Hierarchical clustering based on gene expression (Fig. 1B) was performed on log10-transformed RPKM (with a pseudocount of 0.01) after filtering and batch correction. Complete linkage clustering algorithm is applied to the vectors of Pearson’s correlation coefficients between each pair of samples. The distance between two vectors is computed as abs(1-cc), where cc is again the Pearson’s correlation coefficient between the vectors.

3.2 Network modularity

Modularity was computed similarly to what was described in Breschi et al. [8]. Briefly, we built a graph where vertices (or nodes) are samples and where two vertices, samples, are connected if the Pearson’s correlation coefficient between the corresponding samples, computed on the gene expression values, is higher than a certain threshold (excluding connections of a sample with itself). Like in hierarchical clustering, gene expression values are log10-transformed RPKM after adding a pseudocount of 0.01. The vertex types on which the modularity is computed are either the organ, the body location, the germ layer or the cell type. To compute the modularity we used the modularity() function from the igraph v0.7.1 R package, which implements the following definition [9]:

$$Q = \frac{1}{2m} * \sum_i \sum_j [(A_{ij} - \frac{k_i * k_j}{2m})\delta(c_i, c_j)] \tag{1}$$

where m is the number of edges, A_{ij} is the element of the adjacency matrix A in row i and column j (corresponding to vertices i and j respectively), k_i is the degree of i, k_j is the degree of j, c_i is the type (or component) of i, c_j that of j, the sum goes over all i and j pairs of vertices, and $\delta(x, y) = 1$ if $x = y$ and $\delta(x, y) = 0$ otherwise.

3.3 Estimation of proportions of explained variance

To estimate the proportion of variance explained by the factors cell type, body location, cell cluster, germ layer and organ, we built a separate linear model for each gene and factor. The proportion of explained variance is usually defined as the ratio of the variance across levels of a factor over the total variance. However, to account for different number of levels in each factor, for a gene g and factor f, we compute ω^2_{gf}, which takes into account also the degrees of freedom of that factor:

$$\omega^2_{gf} = \frac{SS_f - DF_f * MSE}{SST + MSE} \tag{2}$$

where SS_f is the sum of squares for factor f, DF_f is the degrees of freedom of factor f, MSE is the mean squared residual variance and SST is the total variance.

3.4 Identification of cell type cluster-specific genes

Cell cluster specificity was surveyed with the edgeR package [10]. Since edgeR relies on a negative binomial model which requires discrete read counts, for this analysis we used the read counts of the filtered 17,440 genes, and kept the counts for individual samples without averaging the replicates. To find genes specific of each major cell type cluster, endothelial, epithelial, melanocytes, and mixed, we performed pairwise differential expression between samples of a given cluster and all the others. Genes with FDR$<$0.01 and at least 4-fold change were considered cell-cluster-specific. The number of cell cluster-specific genes changes depending on the cluster: 635 in the endothelial cluster, 951 epithelial, 353 melanocytes and 935 mixed (2,873 in total).

3.5 Relative contribution of gene expression to variability in isoform abundances

Gene expression contribution in the transcript abundance variation across all samples was computed following the methodology presented in Gonzalez-Porta et al. [11] and further improved in Melé et al. [12]. In a nutshell, for each gene, samples are represented in a multidimensional space using their transcript abundances as coordinates. The contribution of gene expression in the transcript abundance variation is computed by the variation of transcript abundance after projecting the samples into a model of constant splicing (a line in the multidimensional space) divided by the total variation of transcript abundance without projection. If this ratio is close to 1, projecting into the ”no splicing” model didn’t reduce the transcript variation, pointing at mainly gene expression contribution. Inversely, if close to 0, alternative splicing is mostly responsible for the major part of the transcript variation.

A generalization of this approach allows to estimate the effect of a given factor, in this case the cell type, to the contribution of gene expression in transcript abundance variation. Precisely, we asked how much of the transcript variation attributed to tissue is due to changes in gene expression. In practice we compare the proportion of variation explained by the tissue classification after and before projecting the samples into the “no splicing” model. The proportion of variance explained by tissue classification is derived from the classical ANOVA decomposition. The “no splicing” model is represented by a line in the multidimensional space formed by the different transcripts abundances. Like in the analysis across all samples, a value around 1 means that the projection didn’t affect the estimate of variance explained, supporting a full contribution of gene expression. A ratio around 0 means that the variance explained was greatly reduced after projection, supporting a major contribution of alternative splicing.

4 Conservation of cell type cluster specific genes

4.1 Sequence conservation

As a measure of sequence conservation of cell type cluster specific genes, we computed the fraction of genes in each set which have a one-to-one orthology relationship with human and other ten vertebrates. The list of orthologous genes was retrieved from Ensembl Compara [13] v75, which is compatible to Gencode v19 (`http://feb2014.archive.ensembl.org/biomart/martservice`). As a control, the fraction of orthologs was computed also for all protein coding genes (20,731 genes) and for the set of already defined orthologous genes in Barbosa et al. [14] and filtered in Breschi et al. [8]. As this latter set of orthologous genes was from a previous version of Ensembl, we used a subset of them which are retained through the version 75 (6,268 out of 6,283).

4.2 Expression conservation

Using a similar approach to what is described in Breschi et al. [8], expression conservation was measured as the Pearson’s correlation coefficient between expression in human and any other vertebrate species in each organ for the different gene sets. Expression data were obtained from Breschi et al. [8, 14]. Only cell type cluster specific genes with orthologous genes in that dataset were used: 207 of 635 endothelial genes, 277 of 950 epithelial, 130 of 353 melanocyte, 295 of 935 mixed.

5 Analysis of Cap Analysis of Gene Expression data (CAGE)

Gene expression data from CAGE in human primary cells was obtained using FANTOM5 CAGE data [15] and through a private collaboration. Read counts for each gene were obtained by summing up the read counts for all the promoters of that gene and then normalized to cpm. To make it comparable to our RNA-seq dataset, we used only cell lines belonging to a cell type present in our dataset (107 cell lines with multiple biological replicates). Two biological replicates were selected at random, when more than two were available for a given cell line. We applied npIDR filtering [4] on read counts between two replicates and set to 0 the cpm values when the read counts were not reproducible (npIDR$>$0.1), and averaged the cpm between replicates. Genes with cpm$<$1 in all samples were discarded, resulting in 21,282 genes, which were intersected with the filtered 17,440 genes of our dataset, leading to a final set of 15,489 genes. Samples were clustered based on the expression of these genes (Fig. 1D) using log10-transformed cpm after adding a pseudocount of 1. Average linkage clustering algorithm is applied to the vectors of Pearson’s correlation coefficients between each pair of samples. The distance between two vectors is computed as abs(1-cc), where cc is again the Pearson’s correlation coefficient between the vectors.

6 Splicing

6.1 Computation of exon inclusion levels

Exon inclusion levels, measured as percent spliced-in (PSI) values [16], were computed for 439,779 internal exons with IPSA Splicing Analysis Pipeline (`https://github.com/pervouchine/ipsa`). PSI values were computed only for splicing events supported by at least 10 reads. PSI values for exons with an absolute difference in PSI values larger than 0.1 between two replicates were set to NA in both samples. PSI values were averaged between replicates, when available.

6.2 Gene Expression vs Exon inclusion

We compared the differences of gene expression and of exon inclusion within and between cell types. The difference of gene expression within and between cell types was computed as the number of genes differentially expressed (edgeR [10], FDR$<$0.01, log2-fold-change$>$2) between each pair of samples belonging to the same or different cell types, respectively. The difference of exon inclusion within and between cell types, instead, was computed as the number of genes containing at least one exon differentially included between pairs of the same or different cell types, respectively. We defined an exon as differentially included between two samples if the absolute difference between its PSIs in the two samples was larger than 0.2. Finally, we computed the ratio between the number of differentially expressed genes (DE genes) and the number of genes with differentially included exons (DS genes) for each pair of samples. As there was evident bias between PSI values from the first two batches compared to the third, we restricted the analysis to the first and the second batches, which include enough representative samples for the different cell types (Fig. 2F).

7 Analysis of GTEx data

7.1 Identification of tissue-specific genes

Tissue-specific genes were defined from the GTEx consortium RNA-seq data [17]. To have a comparable approach to the one used here, we randomly selected a pair of donors for each tissue subregion, with similar sequencing depth, and reprocessed the raw reads through our pipeline (see Table S3 for a list of SRA run ids). Again, we used edgeR [10] to perform differential expression between samples of a given tissue and the others. This time and because of the high discrepancy in the number of samples between each pairwise differential expression test, we used a quite stringent criteria for differential expression and defined a gene as tissue-specific if its FDR was lower than 0.01 and its log2-fold change was larger than 4. A list of the 2,697 tissue-specific genes and FDR is provided in Table S4.

7.2 Constrained linear model

In this analysis we wanted to estimate the proportion of the four major cell type clusters described so far in whole organs by gene expression data from the GTEx consortium [17]. Thus, we built a linear model for each individual GTEx sample, which uses the average expression of cell type cluster specific genes to estimate the coefficient of each cell type cluster. By constraining the sum of the coefficients to 1, and not allowing negative coefficients, we use these coefficients as a proxy for the relative abundance of a cell type in a given sample. Constrained linear models are implemented with the lsqlin() function from octave optim package. In many cases, we do not expect cell type specific genes to be informative of a certain tissue, e.g. we do not expect high coefficients for blood samples, since we have no blood cells in our dataset. To account for this, we introduce, in each model of a given tissue, the expression of the set of genes specific to that tissue. As an example, when building a model for a blood sample, we add blood-specific genes (see previous section on how we define tissue-specific genes), which are not in the initial set of cell type cluster specific genes. In addition, we add an artificial cell type (”Others”), which in this example would mimic some blood-specific cell type, where the expression of cell type cluster specific genes is 0 (with a pseudocount of 0.1) and the expression of blood-specific genes is the mean of log10-transformed RPKMs (plus a pseudocount of 0.1) across all blood samples. Specifically, the model we build for each sample is:

$$Y_s = \alpha_{endo} * X_{endo} + \alpha_{epi} * X_{epi} + \alpha_{mel} * X_{mel} + \alpha_{mix} * X_{mix} + \alpha_{oth} * X_{oth} \tag{3}$$

where Y_s is the vector of RPKMs for sample s, α_i is the coefficient for cell type i and X_i is the average vector of RPKMs for cell type i. The coefficients for each cell type in all GTEx tissues from all donors are shown in Fig. 4A. Tissues can be represented on a three-dimensional space where the axes are the coefficients for the major cell type clusters, endothelial, epithelial and mixed cells, showing that each tissue has its own characteristic cellular composition (Fig. 4B). The matrix of RPKM values for tissues was downloaded from the GTEx portal (`http://www.gtexportal.org/home/datasets/`, file: `GTEx_Analysis_v6_RNA-seq_RNA-SeQCv1.1.8_gene_rpkm.gct.gz`) for 8,555 samples and 56,319 genes of the v6 release.

7.3 Classification of histological samples

Histological images for the corresponding samples with RNA-seq data can be publicly found on the Biospecimen Reasearch Database website (`https://brd.nci.nih.gov/brd/image-search/searchhome`). Images of histological slides of stomach sections were classified based on the presence of the mucosa (mc) and muscularis (ms) layers (Fig. 4D). To each slide we manually assign two binary vectors, one for the mucosa and one for the muscularis layer, where 1 and 0 indicate presence or absence of the layer, respectively. The length of the vectors depends on the number of tissue sections on the slide, and the order reflect the order of the tissue sections from left to right and from top to bottom. Then, we computed the proportion of sections with a given layer and rounded the proportion to be binary. Thus, each slide will have one of the following possible compositions: mc1ms1, mc1ms0, mc0ms1, mc0ms0. Finally, we focussed on the samples with either one or the other layer, i.e. mc1ms0 and mc0ms1.

7.4 Identification of variable genes among stomach samples

To find the genes at the base of the transcriptional differences we observed amongst stomach samples, we selected in a unsupervised fashion the most variable genes, by using the projection score [18]. We identified 500 most variable genes which maximize the projection score for the first three principal components. Of these, 96 are cell type cluster specific, according to our definition. The expression of these 96 genes clearly discriminates stomach samples with only mucosa or muscularis layer (Fig. 4E).

References

[1] A. Dobin, *et al.*, *Bioinformatics* **29**, 15 (2013).

[2] J. Harrow, *et al.*, *Genome research* **22**, 1760 (2012).

[3] S. B. Montgomery, *et al.*, *Nature* **464**, 773 (2010).

[4] S. Djebali, *et al.*, *Nature* **489**, 101 (2012).

[5] A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, B. Wold, *Nature methods* **5**, 621 (2008).

[6] M. D. Robinson, A. Oshlack, *Genome biology* **11**, 1 (2010).

[7] M. E. Ritchie, *et al.*, *Nucleic Acids Research* **43**, e47 (2015).

[8] A. Breschi, *et al.*, *Genome Biology* **17**, 1 (2016).

[9] A. Clauset, M. E. Newman, C. Moore, *Physical review E* **70**, 066111 (2004).

[10] M. D. Robinson, D. J. McCarthy, G. K. Smyth, *Bioinformatics* **26**, 139 (2010).

[11] M. Gonzàlez-Porta, M. Calvo, M. Sammeth, R. Guigó, *Genome research* **22**, 528 (2012).

[12] M. Melé, *et al.*, *Science* **348**, 660 (2015).

[13] J. Herrero, *et al.*, *Database* **2016**, bav096 (2016).

[14] N. L. Barbosa-Morais, *et al.*, *Science* **338**, 1587 (2012).

[15] T. F. Consortium, *et al.*, *Nature* **507**, 462 (2014).

[16] D. D. Pervouchine, D. G. Knowles, R. Guigó, *Bioinformatics* **29**, 273 (2013).

[17] J. Lonsdale, *et al.*, *Nature genetics* **45**, 580 (2013).

[18] M. Fontes, C. Soneson, *BMC bioinformatics* **12**, 307 (2011).

DISCUSSION

The rise of next-generation sequencing technologies in the past years advanced considerably the field of comparative genomics, transcriptomics and epigenomics. This is particularly important to study the evolution of gene regulation in model organisms, to gain deeper insights on the degree of conservation with human. A considerable amount of work, including efforts from international consortium projects such as Mouse ENCODE (Yue et al., 2014) and FANTOM (Consortium et al., 2014), have been especially centered on the laboratory mouse given its indisputable relevance as model for human biology and diseases.

As presented in Chapter 1, within the Mouse ENCODE consortium we analyzed a vast panel of tissue RNA-seq data in mouse and compared them to human. The conservation of gene expression between the two species is not equivalent for all tissues/organs and genes, and it is heavily dependent on the normalization approach (Gilad and Mizrahi-Man, 2015, Lin et al., 2014, Yue et al., 2014). Organs with more tissue-specific genes, like testes, brain and heart, have more conserved gene expression signatures and clustering the samples based on expression quantification is tissue-dominated (Lin et al., 2014, Sudmant et al., 2015). Extending the analysis to more samples with fewer distinctive tissue-specific genes leads to a more species-dominated clustering (Lin et al., 2014, Sudmant et al., 2015). As presented in Chapter 3, by including more species (human, mouse, macaque, chimpanzee, opossum, platypus and chicken) we decomposed the variance of gene expression across multiple organs into the relative contributions of organ or species and residual variance to the total variance (Breschi et al., 2016). Genes with high proportion of variance across tissues and low proportion of variance across species, which are more tissue-specific, lead to a tissue-dominated clustering. Conversely, genes with high proportion of variance across species and low proportion of variance across tissues have more housekeeping features and lead to a species-dominated clustering.

In Chapter 2 we compared transcriptomes from human cell lines and mouse tissues, to characterize global properties which are conserved despite the heterogeneity of the samples (Pervouchine et al., 2015).

We identified a set of genes with constrained expression, defined as genes whose expression vary less than two orders of magnitude, and we characterized their properties. We noticed that constrained genes are similar to previously described human and murine housekeeping genes, have constrained epigenetic marking and play important roles in determining organismic phenotype.

One of the main limitations of current works on comparative transcriptomics is that entire organs are profiled and analyzed. However, organs are made of mixtures of several cell types and it is possible that a variable cellular composition underlies transcriptional discrepancies between human and mouse homologous organs. Indeed, in Chapter 4 we report a substantial level of gene expression inconsistency amongst samples of the same organ from different patients due to different histological cuts, which favoured one or another cell type. Moreover, in Chapter 4 we describe how certain human primary cells, i.e. endothelial, epithelial, melanocytes and "stromal" cells, including fibroblasts, mesenchymal stem cells (MSCs) and muscle cells, have their own characteristic expression signature, shared amongst cells of the same type, but residing in various anatomical locations. Comparing these transcriptional programs to the mouse ones, albeit certainly relevant and accessible with microarray data, is more challenging with RNA-seq because of the lack of centrally organized datasets and metadata. The FANTOM consortium generated over a thousand CAGE datasets for human and mouse tissues and primary cells (Consortium et al., 2014), but, to our knowledge, no extensive comparative analysis has been described.

Improving annotation of transcripts and genes

In Chapter 2 we have shown how RNA-seq can help improving the current annotation status of coding and noncoding genes, novel isoforms and splicing events. Indeed, having a complete genome annotation is crucial to establish comprehensive orthologous relationships between two genomes. While orthologous protein coding genes are relatively easy to detect thanks to the coding sequence constraints, orthologous noncoding RNAs are more challenging to find. In the past five years several studies, including the one in Chapter 2, reported many novel noncoding transcripts in both the human and the mouse genomes with approximately one thousand orthologous long noncoding RNAs, based on expression levels, sequence homology and synteny (Hezroni et al., 2015, Necsulea et al., 2014, Pervouchine et al., 2015, Washietl et al., 2014). However, discovery of novel lncRNAs is often hindered by their low abundance in the cells. This is recently being addressed by RNA capture techniques, which enrich for targeted sequences and, thus, enhance their coverage (Clark et al., 2015). Targeted RNA-seq of annotated lncRNA loci in human (Clark et al., 2015) and mouse (Bussotti et al., 2016) revealed more complex isoform structures, extended transcript boundaries, and connected previously sparse genes.

Exon structure and splicing are very similar between humans and mice (Abril et al., 2005, Modrek and Lee, 2003), although alternatively spliced exons tend to be less conserved (Modrek and Lee, 2003). As we also describe in Chapter 2, exon inclusion levels are highly correlated between the two species even across very distant sample types. However, comparative analyses of exon inclusion are usually limited to a few hundred conserved exons (Barbosa-Morais et al., 2012, Merkin et al., 2012) and are tight to local splicing events, not considering the whole isoform structure. Novel transcriptomics sequencing strategies, e.g. synthetic long-read sequencing (SLRs) (Tilgner et al., 2015) and single-molecule long-read sequencing (Sharon et al., 2013), enable detection of full-length transcripts and preserve the relationship between distant exons. These techniques, possibly coupled with targeted approaches for lowly abundant loci, will improve the accuracy of isoform detection and might provide new insights on the conservation of isoform usage regulation.

MicroRNAs (miRNAs) are short (approximately 22 nucleotides) noncoding RNA molecules that promote messenger RNA (mRNA) degradation or translational repression, through binding to complementary sequences in target mRNAs (Carthew and Sontheimer, 2009). There is constantly growing evidence that alterations in miRNA expression may lead to several diseases (Li and Kowdley, 2012), including cancer (Lin and Gregory, 2015), thus the use of specific mouse models to understand the mechanisms of miRNA involvement in diseases will certainly be beneficial (Park et al., 2010). Currently, almost three and two thousand miRNAs are annotated in the human and mouse genome, respectively (Kozomara and Griffiths-Jones, 2010). However, only a small fraction (300 miRNAs) of them has a defined ortholog in the other species (Landgraf et al., 2007) and comparison of their expression profiles is still limited to a few studies (Meunier et al., 2013, Roux et al., 2012). We believe that expanding the annotation status and quality of miRNAs in both species can help improving comparative studies and guide the use of mouse as models for miRNA biology.

Biology and big data: filling in the matrix

The development and continuous improvements of high-throughput sequencing technologies pushed forward the field of genomics, and led biology into the expanding world of big data. After the sequencing of the human and mouse genomes, the ENCODE project aimed at determining the function of each sequence in the genome. To do so, the ENCODE and Mouse ENCODE consortia generated a wealth of genome-wide screenings for surveying different functional elements in a plethora of human and mouse samples and conditions (Ecker et al., 2012). These can be seen as a matrix, where each row is a sample in a given condition and each column is an assay (Maher, 2012), with a third dimension representing the species. Although thousands of experiments have been carried out by the ENCODE consortium, by other consortia and by other smaller teams worldwide, one could argue that completing such a matrix would require infinite amount of time, since an infinite number of conditions could be portrayed (Maher, 2012).

Even if the amount of data is steadily growing and more and more cells in this matrix are filled in, integrating all this data is still a current unsolved computational challenge (Libbrecht and Noble, 2015, Ritchie et al., 2015). Indeed, the development of more sophisticated algorithms with optimized performances and the creation of new ways of visualizing highly dimensional data are fields of active research. In addition, not all experiments might need to be physically performed, as new computational methods are emerging to impute whole genome data for new assays or samples from existing ones (Ernst and Kellis, 2015).

Ultimately, the final output of genomic regulation is protein production, which, of course, cannot be assayed with DNA sequencing technologies: therefore, albeit mass-spectrometry allows for large-scale profiling of protein expression, the throughput and the sensitivity of the assay are still inferior to next-generation sequencing (Kim et al., 2014). It has been reported that 10,000-12,000 and 6,000-8,000 proteins are detected in a given human or mouse tissue, respectively (Huttlin et al., 2010, Uhlén et al., 2015, Wilhelm et al., 2014). Like at the other molecular levels, differences and similarities have been found in human and mouse proteomes of very specific samples (Gharib et al., 2010), however, a comprehensive proteomics comparison between the two species is still lacking, to the best of our knowledge.

Most comparative studies of functional genomics are limited to entire organs or populations of cells, whereas it would be relevant to investigate the relationships between single cells within a population or a complex organ. While single-cell genomics has advanced really fast in terms of experimental procedures and bioinformatics analysis and allows to study multicellular structures at an unprecedented resolution, new methodologies are emerging, which preserve spatial information about the tissue context or subcellular localization of analysed nucleic acids (Crosetto et al., 2015). Although spatial transcriptomics is still in its very early days (Chen et al., 2015, Satija et al., 2015), it carries the promise of revolutionizing the way multicellular complexes, e.g. organs, are studied and might reveal new insights into the conservation of how these complexes are organized between human and mouse.

Finally, we could think of adding yet another dimension to this ideal matrix of samples, experiments and species: time. Adding a temporal dimension, such as response to a treatment or differentiation time-courses, to already known steady states could unravel unknown patterns of conservation between humans and mice, and could be especially important for clinical studies, e.g. to study the time of physiological responses to drugs or the progression of a disease.

In the uprising era of precision medicine, each individual will likely have his/her genome sequenced and possibly multiple genomics assays in different anatomical sites and at different life stages. Thus, we can envision that human-mouse comparisons will eventually be done on a person-by-person base and customized mouse models can be tailored in a personalized fashion.

Since both similarities and differences exist between humans and mice, it is not trivial to conclude in absolute terms whether mice are good models for humans. This changes depending on the specific factors (e.g. biological conditions, physiological processes and responses, sample types) that are considered and not all organs and genes showed the same degree of conservation between the two species. We hope the work presented in this thesis contributed to frame this question in a way that moves beyond a global perspective but takes into account some of these specific factors. Other studies besides gene expression comparison, like co-expression network comparison, and comparative proteomics as well as comparative processing dynamics are potential avenues of further research.

CONCLUSIONS

The work presented in this thesis discusses the relevance of the laboratory mouse as a model organism for human, especially at the transcriptional level, and tries to investigate how cellular composition affects gene expression programs of complex organs.

Here is a summary of the main contributions of this thesis:

- Human and mouse transcriptomes have both conserved and diverged patterns.
- Conservation of human and mouse transcriptomes depends on many factors:
  - Number of tissue-specific genes in each organ
  - Gene-specific expression variation across organs and between species
  - Normalization procedures
- These factors affect conservation patterns of expression also in comparisons with more mammals and birds, i.e. chimpanzee, macaque, opossum, platypus, chicken.
- Human and mouse gene expression and splicing are globally conserved even across very heterogeneous sample types, e.g. human cell lines and mouse tissues.
- We defined a set of 6,000 genes with constrained expression both across samples and between human and mouse.
- Constrained genes have stronger epigenetics markings in both species and their perturbations are associated with significant phenotypes including embryonic lethality and cancer.
- RNA-seq analysis of 53 human primary cells shows that similar cell types from different anatomical locations share the same transcriptional programs. In particular, we identified four major programs in this dataset: endothelial, epithelial, melanocytes, and a mixed set of cell types, including fibroblasts, smooth muscle cells and mesenchymal stem cells.

- We defined a set of almost 3,000 genes which clearly distinguish these major groups.
- Gene expression profiles of many organs can be reconstructed based on the expression of these 3,000 genes, and relative cellular composition within the organs can be estimated.

BIBLIOGRAPHY

Abril, J. F., Castelo, R., and Guigó, R. (2005). Comparison of splice sites in mammals and chicken. *Genome research*, 15(1):111–119.

Adams, D. J. and van der Weyden, L. (2008). Contemporary approaches for modifying the mouse genome. *Physiological genomics*, 34(3):225–238.

Aparicio, S., Hidalgo, M., and Kung, A. L. (2015). Examining the utility of patient-derived xenograft mouse models. *Nature reviews Cancer*, 15(5):311–316.

Barbosa-Morais, N. L., Irimia, M., Pan, Q., Xiong, H. Y., Gueroussov, S., Lee, L. J., Slobodeniuc, V., Kutter, C., Watt, S., Çolak, R., et al. (2012). The evolutionary landscape of alternative splicing in vertebrate species. *Science*, 338(6114):1587–1593.

Beck, J. A., Lloyd, S., Hafezparast, M., Lennon-Pierce, M., Eppig, J. T., Festing, M. F., and Fisher, E. M. (2000). Genealogies of mouse inbred strains. *Nature genetics*, 24(1):23–25.

Bedell, M. A., Jenkins, N. A., and Copeland, N. G. (1997). Mouse models of human disease. part i: techniques and resources for genetic analysis in mice. *Genes and Development*, 11(1):1–10.

Böck, B. C., Stein, U., Schmitt, C. A., and Augustin, H. G. (2014). Mouse models of human cancer. *Cancer research*, 74(17):4671–4675.

Bourque, G., Leong, B., Vega, V. B., Chen, X., Lee, Y. L., Srinivasan, K. G., Chew, J.-L., Ruan, Y., Wei, C.-L., Ng, H. H., et al. (2008). Evolution of the mammalian transcription factor binding repertoire via transposable elements. *Genome research*, 18(11):1752–1762.

Brawand, D., Soumillon, M., Necsulea, A., Julien, P., Csárdi, G., Harrigan, P., Weier, M., Liechti, A., Aximu-Petri, A., Kircher, M., et al. (2011). The evolution of gene expression levels in mammalian organs. *Nature*, 478(7369):343–348.

Breschi, A., Djebali, S., Gillis, J., Pervouchine, D. D., Dobin, A., Davis, C. A., Gingeras, T. R., and Guigó, R. (2016). Gene-specific patterns of expression variation across organs and species. *Genome Biology*, 17(1):1.

Bult, C. J., Eppig, J. T., Blake, J. A., Kadin, J. A., Richardson, J. E., Group, M. G. D., et al. (2016). Mouse genome database 2016. *Nucleic acids research*, 44(D1):D840–D847.

Bussotti, G., Leonardi, T., Clark, M. B., Mercer, T. R., Crawford, J., Malquori, L., Notredame, C., Dinger, M. E., Mattick, J. S., and Enright, A. J. (2016). Improved definition of the mouse transcriptome via targeted rna sequencing. *Genome research*, 26(5):705–716.

Cabili, M. N., Trapnell, C., Goff, L., Koziol, M., Tazon-Vega, B., Regev, A., and Rinn, J. L. (2011). Integrative annotation of human large intergenic noncoding rnas reveals global properties and specific subclasses. *Genes & development*, 25(18):1915–1927.

Carthew, R. W. and Sontheimer, E. J. (2009). Origins and mechanisms of mirnas and sirnas. *Cell*, 136(4):642–655.

Chan, E. T., Quon, G. T., Chua, G., Babak, T., Trochesset, M., Zirngibl, R. A., Aubin, J., Ratcliffe, M. J., Wilde, A., Brudno, M., et al. (2009). Conservation of core gene expression in vertebrate tissues. *Journal of biology*, 8(3):1.

Chen, K. H., Boettiger, A. N., Moffitt, J. R., Wang, S., and Zhuang, X. (2015). Spatially resolved, highly multiplexed rna profiling in single cells. *Science*, 348(6233):aaa6090.

Cheng, Y., Ma, Z., Kim, B.-H., Wu, W., Cayting, P., Boyle, A. P., Sundaram, V., Xing, X., Dogan, N., Li, J., et al. (2014). Principles of regulatory information conservation between mouse and human. *Nature*, 515(7527):371–375.

Cheon, D.-J. and Orsulic, S. (2011). Mouse models of cancer. *ANNU REV PATHOL MECH*.

Chinwalla, A. T., Cook, L. L., Delehaunty, K. D., Fewell, G. A., Fulton, L. A., Fulton, R. S., Graves, T. A., Hillier, L. W., Mardis, E. R., McPherson, J. D., et al. (2002). Initial sequencing and comparative analysis of the mouse genome. *Nature*, 420(6915):520–562.

Clark, M. B., Mercer, T. R., Bussotti, G., Leonardi, T., Haynes, K. R., Crawford, J., Brunck, M. E., Lê Cao, K.-A., Thomas, G. P., Chen, W. Y., et al. (2015). Quantitative gene profiling of long noncoding rnas with targeted rna sequencing. *Nature methods*, 12(4):339–342.

Consortium, . G. P. et al. (2015a). A global reference for human genetic variation. *Nature*, 526(7571):68–74.

Consortium, E. P. et al. (2012). An integrated encyclopedia of dna elements in the human genome. *Nature*, 489(7414):57–74.

Consortium, G. et al. (2015b). The genotype-tissue expression (gtex) pilot analysis: Multitissue gene regulation in humans. *Science*, 348(6235):648–660.

Consortium, T. F. et al. (2014). A promoter-level mammalian expression atlas. *Nature*, 507(7493):462–470.

Crosetto, N., Bienko, M., and van Oudenaarden, A. (2015). Spatially resolved transcriptomics and beyond. *Nature Reviews Genetics*, 16(1):57–66.

Darmanis, S., Sloan, S. A., Zhang, Y., Enge, M., Caneda, C., Shuer, L. M., Gephart, M. G. H., Barres, B. A., and Quake, S. R. (2015). A survey of human brain transcriptome diversity at the single cell level. *Proceedings of the National Academy of Sciences*, 112(23):7285–7290.

Denas, O., Sandstrom, R., Cheng, Y., Beal, K., Herrero, J., Hardison, R. C., and Taylor, J. (2015). Genome-wide comparative analysis reveals human-mouse regulatory landscape and evolution. *BMC genomics*, 16(1):1.

Deng, Q., Ramsköld, D., Reinius, B., and Sandberg, R. (2014). Single-cell rna-seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science*, 343(6167):193–196.

Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., Guernec, G., Martin, D., Merkel, A., Knowles, D. G., et al. (2012). The gencode v7 catalog of human long noncoding rnas: analysis of their gene structure, evolution, and expression. *Genome research*, 22(9):1775–1789.

Ecker, J. R., Bickmore, W. A., Barroso, I., Pritchard, J. K., Gilad, Y., and Segal, E. (2012). Genomics: Encode explained. *Nature*, 489(7414):52–55.

Ellacott, K. L., Morton, G. J., Woods, S. C., Tso, P., and Schwartz, M. W. (2010). Assessment of feeding behavior in laboratory mice. *Cell metabolism*, 12(1):10–17.

Ernst, J. and Kellis, M. (2012). Chromhmm: automating chromatin-state discovery and characterization. *Nature methods*, 9(3):215–216.

Ernst, J. and Kellis, M. (2015). Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues. *Nature biotechnology*, 33(4):364–376.

Esteller, M. (2011). Non-coding rnas in human disease. *Nature Reviews Genetics*, 12(12):861–874.

Gharib, S. A., Nguyen, E., Altemeier, W. A., Shaffer, S. A., Doneanu, C. E., Goodlett, D. R., and Schnapp, L. M. (2010). Of mice and men: comparative proteomics of bronchoalveolar fluid. *European Respiratory Journal*, 35(6):1388–1395.

Gilad, Y. and Mizrahi-Man, O. (2015). A reanalysis of mouse encode comparative gene expression data. *F1000Research*, 4.

Grün, D., Lyubimova, A., Kester, L., Wiebrands, K., Basak, O., Sasaki, N., Clevers, H., and van Oudenaarden, A. (2015). Single-cell messenger rna sequencing reveals rare intestinal cell types. *Nature*, 525(7568):251–255.

Hardison, R. C. (2016). A guide to translation of research results from model organisms to human. *Genome Biology*, 17(1):1.

Harrow, J., Frankish, A., Gonzalez, J. M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B. L., Barrell, D., Zadissa, A., Searle, S., et al. (2012). Gencode: the reference human genome annotation for the encode project. *Genome research*, 22(9):1760–1774.

Hay, M., Thomas, D. W., Craighead, J. L., Economides, C., and Rosenthal, J. (2014). Clinical development success rates for investigational drugs. *Nature biotechnology*, 32(1):40–51.

Herrero, J., Muffato, M., Beal, K., Fitzgerald, S., Gordon, L., Pignatelli, M., Vilella, A. J., Searle, S. M., Amode, R., Brent, S., et al. (2016). Ensembl comparative genomics resources. *Database*, 2016:bav096.

Hezroni, H., Koppstein, D., Schwartz, M. G., Avrutin, A., Bartel, D. P., and Ulitsky, I. (2015). Principles of long noncoding rna evolution derived from direct comparison of transcriptomes in 17 species. *Cell reports*, 11(7):1110–1122.

Holgersen, K., Kutlu, B., Fox, B., Serikawa, K., Lord, J., Hansen, A. K., and Holm, T. L. (2015). High-resolution gene expression profiling using rna sequencing in patients with inflammatory bowel disease and in mouse models of colitis. *Journal of Crohn’s and Colitis*, page jjv050.

Hume, D. A., Summers, K. M., Raza, S., Baillie, J. K., and Freeman, T. C. (2010). Functional clustering and lineage markers: insights into cellular differentiation and gene function from large-scale microarray studies of purified primary cell populations. *Genomics*, 95(6):328–338.

Huttlin, E. L., Jedrychowski, M. P., Elias, J. E., Goswami, T., Rad, R., Beausoleil, S. A., Villén, J., Haas, W., Sowa, M. E., and Gygi, S. P. (2010). A tissue-specific atlas of mouse protein phosphorylation and expression. *Cell*, 143(7):1174–1189.

Jaitin, D. A., Kenigsberg, E., Keren-Shaul, H., Elefant, N., Paul, F., Zaretsky, I., Mildner, A., Cohen, N., Jung, S., Tanay, A., et al. (2014). Massively parallel single-cell rna-seq for marker-free decomposition of tissues into cell types. *Science*, 343(6172):776–779.

Johnson, R., Samuel, J., Ng, C. K. L., Jauch, R., Stanton, L. W., and Wood, I. C. (2009). Evolution of the vertebrate gene regulatory network controlled by the transcriptional repressor rest. *Molecular biology and evolution*, 26(7):1491–1507.

Keane, T., Doran, A., Adams, D., Hunter, K., Flint, J., and Wong, K. (2016). Deep genome sequencing and variation analysis of 13 inbred mouse strains defines candidate phenotypic alleles, private variation, and homozygous truncating mutations. *bioRxiv*, page 039131.

Keane, T. M., Goodstadt, L., Danecek, P., White, M. A., Wong, K., Yalcin, B., Heger, A., Agam, A., Slater, G., Goodson, M., et al. (2011). Mouse genomic variation and its effect on phenotypes and gene regulation. *Nature*, 477(7364):289–294.

Kim, M.-S., Pinto, S. M., Getnet, D., Nirujogi, R. S., Manda, S. S., Chaerkady, R., Madugundu, A. K., Kelkar, D. S., Isserlin, R., Jain, S., et al. (2014). A draft map of the human proteome. *Nature*, 509(7502):575–581.

Kiselycznyk, C. and Holmes, A. (2011). All (c57bl/6) mice are not created equal. *Frontiers in neuroscience*, 5:10.

Kolodziejczyk, A. A., Kim, J. K., Svensson, V., Marioni, J. C., and Teichmann, S. A. (2015). The technology and biology of single-cell rna sequencing. *Molecular cell*, 58(4):610–620.

Koonin, E. V. and Wolf, Y. I. (2010). Constraints and plasticity in genome and molecular-phenome evolution. *Nature Reviews Genetics*, 11(7):487–498.

Kozomara, A. and Griffiths-Jones, S. (2010). mirbase: integrating microrna annotation and deep-sequencing data. *Nucleic acids research*, page gkq1027.

Kunarso, G., Chia, N.-Y., Jeyakani, J., Hwang, C., Lu, X., Chan, Y.-S., Ng, H.-H., and Bourque, G. (2010). Transposable elements have rewired the core regulatory network of human embryonic stem cells. *Nature genetics*, 42(7):631–634.

Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M. J., et al. (2015). Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539):317–330.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921.

Landgraf, P., Rusu, M., Sheridan, R., Sewer, A., Iovino, N., Aravin, A., Pfeffer, S., Rice, A., Kamphorst, A. O., Landthaler, M., et al. (2007). A mammalian microrna expression atlas based on small rna library sequencing. *Cell*, 129(7):1401–1414.

Lappalainen, T., Sammeth, M., Friedländer, M. R., AC't Hoen, P., Monlong, J., Rivas, M. A., Gonzàlez-Porta, M., Kurbatova, N., Griebel, T., Ferreira, P. G., et al. (2013). Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, 501(7468):506–511.

Lee, Y.-s., Krishnan, A., Zhu, Q., and Troyanskaya, O. G. (2013). Ontology-aware classification of tissue and cell-type signals in gene expression profiles across platforms and technologies. *Bioinformatics*, 29(23):3036–3044.

Levy, S., Sutton, G., Ng, P. C., Feuk, L., Halpern, A. L., Walenz, B. P., Axelrod, N., Huang, J., Kirkness, E. F., Denisov, G., et al. (2007). The diploid genome sequence of an individual human. *PLoS Biol*, 5(10):e254.

Li, J., Klughammer, J., Farlik, M., Penz, T., Spittler, A., Barbieux, C., Berishvili, E., Bock, C., and Kubicek, S. (2016). Single-cell transcriptomes reveal characteristic features of human pancreatic islet cell types. *EMBO reports*, 17(2):178–187.

Li, Y. and Kowdley, K. V. (2012). Micrornas in common human diseases. *Genomics, proteomics & bioinformatics*, 10(5):246–253.

Liao, B.-Y. and Zhang, J. (2006a). Evolutionary conservation of expression profiles between human and mouse orthologous genes. *Molecular biology and evolution*, 23(3):530–540.

Liao, B.-Y. and Zhang, J. (2006b). Low rates of expression profile divergence in highly expressed genes and tissue-specific genes during mammalian evolution. *Molecular biology and evolution*, 23(6):1119–1128.

Libbrecht, M. W. and Noble, W. S. (2015). Machine learning applications in genetics and genomics. *Nature Reviews Genetics*, 16(6):321–332.

Lin, S. and Gregory, R. I. (2015). Microrna biogenesis pathways in cancer. *Nature Reviews Cancer*, 15(6):321–333.

Lin, S., Lin, Y., Nery, J. R., Urich, M. A., Breschi, A., Davis, C. A., Dobin, A., Zaleski, C., Beer, M. A., Chapman, W. C., et al. (2014). Comparison of the transcriptional landscapes between human and mouse tissues. *Proceedings of the National Academy of Sciences*, 111(48):17224–17229.

Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., Young, N., et al. (2013). The genotype-tissue expression (gtex) project. *Nature genetics*, 45(6):580–585.

Lynch, W. J., Nicholson, K. L., Dance, M. E., Morgan, R. W., and Foley, P. L. (2010). Animal models of substance abuse and addiction: implications for science, animal welfare, and society. *Comparative medicine*, 60(3):177–188.

Mabbott, N. A., Baillie, J. K., Brown, H., Freeman, T. C., and Hume, D. A. (2013). An expression atlas of human primary cells: inference of gene function from coexpression networks. *BMC genomics*, 14(1):632.

Macosko, E. Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A. R., Kamitaki, N., Martersteck, E. M., et al. (2015). Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, 161(5):1202–1214.

Maher, B. (2012). Encode: The human encyclopaedia. *Nature*, 489(7414):46.

Mak, I., Evaniew, N., and Ghert, M. (2014). Lost in translation: animal models and clinical trials in cancer treatment. *Am J Transl Res*, 6(2):114–8.

Matsuo, N., Takao, K., Nakanishi, K., Yamasaki, N., Tanda, K., and Miyakawa, T. (2010). Behavioral profiles of three c57bl/6 substrains. *Frontiers in behavioral neuroscience*, 4:29.

McCall, M. N., Uppal, K., Jaffee, H. A., Zilliox, M. J., and Irizarry, R. A. (2011). The gene expression barcode: leveraging public data repositories to begin cataloging the human and murine transcriptomes. *Nucleic acids research*, 39(suppl 1):D1011–D1015.

Melé, M., Ferreira, P. G., Reverter, F., DeLuca, D. S., Monlong, J., Sammeth, M., Young, T. R., Goldmann, J. M., Pervouchine, D. D., Sullivan, T. J., et al. (2015). The human transcriptome across tissues and individuals. *Science*, 348(6235):660–665.

Merkin, J., Russell, C., Chen, P., and Burge, C. B. (2012). Evolutionary dynamics of gene and isoform regulation in mammalian tissues. *Science*, 338(6114):1593–1599.

Mestas, J. and Hughes, C. C. (2004). Of mice and not men: differences between mouse and human immunology. *The Journal of Immunology*, 172(5):2731–2738.

Meunier, J., Lemoine, F., Soumillon, M., Liechti, A., Weier, M., Guschanski, K., Hu, H., Khaitovich, P., and Kaessmann, H. (2013). Birth and expression evolution of mammalian microrna genes. *Genome research*, 23(1):34–45.

Modrek, B. and Lee, C. J. (2003). Alternative splicing in the human, mouse and rat genomes is associated with an increased frequency of exon creation and/or loss. *Nature genetics*, 34(2):177–180.

Morgan, R. A. (2012). Human tumor xenografts: the good, the bad, and the ugly. *Molecular Therapy*, 20(5):882.

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by rna-seq. *Nature methods*, 5(7):621–628.

Mudge, J. M. and Harrow, J. (2015). Creating reference gene annotation for the mouse c57bl6/j genome assembly. *Mammalian Genome*, 26(9-10):366–378.

Nawrocki, E. P., Burge, S. W., Bateman, A., Daub, J., Eberhardt, R. Y., Eddy, S. R., Floden, E. W., Gardner, P. P., Jones, T. A., Tate, J., et al. (2014). Rfam 12.0: updates to the rna families database. *Nucleic acids research*, page gku1063.

Necsulea, A. and Kaessmann, H. (2014). Evolutionary dynamics of coding and non-coding transcriptomes. *Nature Reviews Genetics*, 15(11):734–748.

Necsulea, A., Soumillon, M., Warnefors, M., Liechti, A., Daish, T., Zeller, U., Baker, J. C., Grützner, F., and Kaessmann, H. (2014). The evolution of lncrna repertoires and expression patterns in tetrapods. *Nature*, 505(7485):635–640.

Odom, D. T., Dowell, R. D., Jacobsen, E. S., Gordon, W., Danford, T. W., MacIsaac, K. D., Rolfe, P. A., Conboy, C. M., Gifford, D. K., and Fraenkel, E. (2007). Tissue-specific transcriptional regulation has diverged significantly between human and mouse. *Nature genetics*, 39(6):730–732.

Ohnishi, Y., Huber, W., Tsumura, A., Kang, M., Xenopoulos, P., Kurimoto, K., Oleś, A. K., Araúzo-Bravo, M. J., Saitou, M., Hadjantonakis, A.-K., et al. (2014). Cell-to-cell expression variability followed by signal reinforcement progressively segregates early mouse lineages. *Nature cell biology*, 16(1):27–37.

Onos, K. D., Rizzo, S. J. S., Howell, G. R., and Sasner, M. (2016). Toward more predictive genetic mouse models of alzheimer's disease. *Brain research bulletin*, 122:1–11.

Pai, A. A. and Gilad, Y. (2014). Comparative studies of gene regulatory mechanisms. *Current opinion in genetics & development*, 29:68–74.

Park, C. Y., Choi, Y., and McManus, M. T. (2010). Analysis of microrna knockouts in mice. *Human molecular genetics*, page ddq367.

Paul, F., Arkin, Y., Giladi, A., Jaitin, D. A., Kenigsberg, E., Keren-Shaul, H., Winter, D., Lara-Astiaso, D., Gury, M., Weiner, A., et al. (2015). Transcriptional heterogeneity and lineage commitment in myeloid progenitors. *Cell*, 163(7):1663–1677.

Pennacchio, L. A., Ahituv, N., Moses, A. M., Prabhakar, S., Nobrega, M. A., Shoukry, M., Minovitsky, S., Dubchak, I., Holt, A., Lewis, K. D., et al. (2006). In vivo enhancer analysis of human conserved non-coding sequences. *Nature*, 444(7118):499–502.

Pervouchine, D., Djebali, S., Breschi, A., Davis, C. A., Barja, P. P., Dobin, A., Tanzer, A., Lagarde, J., Zaleski, C., See, L.-H., et al. (2015). Enhanced transcriptome maps from multiple mouse tissues reveal evolutionary constraint in gene expression for thousands of genes. *Nat Commun*.

Pignatelli, M., Vilella, A. J., Muffato, M., Gordon, L., White, S., Flicek, P., and Herrero, J. (2016). ncrna orthologies in the vertebrate lineage. *Database*, 2016:bav127.

Pishesha, N., Thiru, P., Shi, J., Eng, J. C., Sankaran, V. G., and Lodish, H. F. (2014). Transcriptional divergence and conservation of human and mouse erythropoiesis. *Proceedings of the National Academy of Sciences*, 111(11):4103–4108.

Pritchard, C. C., Hsu, L., Delrow, J., and Nelson, P. S. (2001). Project normal: defining normal variance in mouse gene expression. *Proceedings of the National Academy of Sciences*, 98(23):13266–13271.

Rangarajan, A. and Weinberg, R. A. (2003). Comparative biology of mouse versus human cells: modelling human cancer in mice. *Nature Reviews Cancer*, 3(12):952–959.

Richmond, A. and Su, Y. (2008). Mouse xenograft models vs gem models for human cancer therapeutics. *Disease Models and Mechanisms*, 1(2-3):78–82.

Ritchie, M. D., Holzinger, E. R., Li, R., Pendergrass, S. A., and Kim, D. (2015). Methods of integrating data to uncover genotype-phenotype interactions. *Nature Reviews Genetics*, 16(2):85–97.

Romero, I. G., Ruvinsky, I., and Gilad, Y. (2012). Comparative studies of gene expression and the evolution of gene regulation. *Nature Reviews Genetics*, 13(7):505–516.

Roux, J., Gonzàlez-Porta, M., and Robinson-Rechavi, M. (2012). Comparative analysis of human and mouse expression data illuminates tissue-specific evolutionary patterns of mirnas. *Nucleic acids research*, 40(13):5890–5900.

Rueda, N., Flórez, J., and Martínez-Cué, C. (2012). Mouse models of down syndrome as a tool to unravel the causes of mental disabilities. *Neural plasticity*, 2012.

Saliba, A.-E., Westermann, A. J., Gorski, S. A., and Vogel, J. (2014). Single-cell rna-seq: advances and future challenges. *Nucleic acids research*, 42(14):8845–8860.

Satija, R., Farrell, J. A., Gennert, D., Schier, A. F., and Regev, A. (2015). Spatial reconstruction of single-cell gene expression data. *Nature biotechnology*, 33(5):495–502.

Schmidt, D., Wilson, M. D., Ballester, B., Schwalie, P. C., Brown, G. D., Marshall, A., Kutter, C., Watt, S., Martinez-Jimenez, C. P., Mackay, S., et al. (2010). Five-vertebrate chip-seq reveals the evolutionary dynamics of transcription factor binding. *Science*, 328(5981):1036–1040.

Schweinfurth, N. and Lang, U. E. (2015). Behavioral testing of mice concerning anxiety and depression. *Zeitschrift für Psychologie*.

Scialdone, A., Tanaka, Y., Jawaid, W., Moignard, V., Wilson, N. K., Macaulay, I. C., Marioni, J. C., and Göttgens, B. (2016). Resolving early mesoderm diversification through single-cell expression profiling. *Nature*.

Seok, J., Warren, H. S., Cuenca, A. G., Mindrinos, M. N., Baker, H. V., Xu, W., Richards, D. R., McDonald-Smith, G. P., Gao, H., Hennessy, L., et al. (2013). Genomic responses in mouse models poorly mimic human inflammatory diseases. *Proceedings of the National Academy of Sciences*, 110(9):3507–3512.

Sharon, D., Tilgner, H., Grubert, F., and Snyder, M. (2013). A single-molecule long-read survey of the human transcriptome. *Nature biotechnology*, 31(11):1009–1014.

Shay, T., Lederer, J. A., and Benoist, C. (2015). Genomic responses to inflammation in mouse models mimic humans: we concur, apples to oranges comparisons won’t do. *Proceedings of the National Academy of Sciences*, 112(4):E346–E346.

Shi, X., Sun, M., Liu, H., Yao, Y., and Song, Y. (2013). Long noncoding rnas: a new frontier in the study of human diseases. *Cancer letters*, 339(2):159–166.

Silverman, J. L., Yang, M., Lord, C., and Crawley, J. N. (2010). Behavioural phenotyping assays for mouse models of autism. *Nature Reviews Neuroscience*, 11(7):490–502.

Singh, P., Schimenti, J. C., and Bolcun-Filas, E. (2015). A mouse geneticist’s practical guide to crispr applications. *Genetics*, 199(1):1–15.

Soumillon, M., Necsulea, A., Weier, M., Brawand, D., Zhang, X., Gu, H., Barthès, P., Kokkinaki, M., Nef, S., Gnirke, A., et al. (2013). Cellular source and mechanisms of high transcriptome complexity in the mammalian testis. *Cell reports*, 3(6):2179–2190.

Stamatoyannopoulos, J. A., Snyder, M., Hardison, R., Ren, B., Gingeras, T., Gilbert, D. M., Groudine, M., Bender, M., Kaul, R., Canfield, T., et al. (2012). An encyclopedia of mouse dna elements (mouse encode). *Genome biology*, 13(8):1.

Steimer, T. (2011). Animal models of anxiety disorders in rats and mice: some conceptual issues. *Dialogues Clin Neurosci*, 13(4):495–506.

Su, A. I., Cooke, M. P., Ching, K. A., Hakak, Y., Walker, J. R., Wiltshire, T., Orth, A. P., Vega, R. G., Sapinoso, L. M., Moqrich, A., et al. (2002). Large-scale analysis of the human and mouse transcriptomes. *Proceedings of the National Academy of Sciences*, 99(7):4465–4470.

Sudmant, P. H., Alexis, M. S., and Burge, C. B. (2015). Meta-analysis of rna-seq expression data across species, tissues and studies. *Genome biology*, 16(1):1–11.

Takao, K. and Miyakawa, T. (2015). Genomic responses in mouse models greatly mimic human inflammatory diseases. *Proceedings of the National Academy of Sciences*, 112(4):1167–1172.

Tilgner, H., Jahanbani, F., Blauwkamp, T., Moshrefi, A., Jaeger, E., Chen, F., Harel, I., Bustamante, C. D., Rasmussen, M., and Snyder, M. P. (2015). Comprehensive transcriptome analysis using synthetic long-read sequencing reveals molecular co-association of distant splicing events. *Nature biotechnology*, 33(7):736–742.

Trapnell, C. (2015). Defining cell types and states with single-cell genomics. *Genome research*, 25(10):1491–1498.

Treutlein, B., Brownfield, D. G., Wu, A. R., Neff, N. F., Mantalas, G. L., Espinoza, F. H., Desai, T. J., Krasnow, M. A., and Quake, S. R. (2014). Reconstructing lineage hierarchies of the distal lung epithelium using single-cell rna-seq. *Nature*, 509(7500):371–375.

Turk, R., AC't Hoen, P., Sterrenburg, E., De Menezes, R. X., De Meijer, E. J., Boer, J. M., Van Ommen, G.-J. B., and Den Dunnen, J. T. (2004). Gene expression variation between mouse inbred strains. *BMC genomics*, 5(1):1.

Uhlén, M., Fagerberg, L., Hallström, B. M., Lindskog, C., Oksvold, P., Mardinoglu, A., Sivertsson, Å., Kampf, C., Sjöstedt, E., Asplund, A., et al. (2015). Tissue-based map of the human proteome. *Science*, 347(6220):1260419.

Vakhrusheva, O. A., Bazykin, G. A., and Kondrashov, A. S. (2013). Genome-level analysis of selective constraint without apparent sequence conservation. *Genome biology and evolution*, 5(3):532–541.

Vanhooren, V. and Libert, C. (2013). The mouse as a model organism in aging research: usefulness, pitfalls and possibilities. *Ageing research reviews*, 12(1):8–21.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., et al. (2001). The sequence of the human genome. *science*, 291(5507):1304–1351.

Vickaryous, M. K. and Hall, B. K. (2006). Human cell type diversity, evolution, development, and classification with special reference to cells derived from the neural crest. *Biological reviews*, 81(03):425–455.

Vierstra, J., Rynes, E., Sandstrom, R., Zhang, M., Canfield, T., Hansen, R. S., Stehling-Sun, S., Sabo, P. J., Byron, R., Humbert, R., et al. (2014). Mouse regulatory dna landscapes reveal global principles of cis-regulatory evolution. *Science*, 346(6212):1007–1012.

Villar, D., Berthelot, C., Aldridge, S., Rayner, T. F., Lukk, M., Pignatelli, M., Park, T. J., Deaville, R., Erichsen, J. T., Jasinska, A. J., et al. (2015). Enhancer evolution across 20 mammalian species. *Cell*, 160(3):554–566.

Visel, A., Minovitsky, S., Dubchak, I., and Pennacchio, L. A. (2007). Vista enhancer browser-a database of tissue-specific human enhancers. *Nucleic acids research*, 35(suppl 1):D88–D92.

Visel, A., Prabhakar, S., Akiyama, J. A., Shoukry, M., Lewis, K. D., Holt, A., Plajzer-Frick, I., Afzal, V., Rubin, E. M., and Pennacchio, L. A. (2008). Ultraconservation identifies a small subset of extremely constrained developmental enhancers. *Nature genetics*, 40(2):158–160.

Wade, C. M., Kulbokas, E. J., Kirby, A. W., Zody, M. C., Mullikin, J. C., Lander, E. S., Lindblad-Toh, K., and Daly, M. J. (2002). The mosaic structure of variation in the laboratory mouse genome. *Nature*, 420(6915):574–578.

Wang, Y. and Rekaya, R. (2009). A comprehensive analysis of gene expression evolution between humans and mice. *Evolutionary Bioinformatics*, 5:81.

Wapinski, O. and Chang, H. Y. (2011). Long noncoding rnas and human disease. *Trends in cell biology*, 21(6):354–361.

Warren, H. S., Tompkins, R. G., Moldawer, L. L., Seok, J., Xu, W., Mindrinos, M. N., Maier, R. V., Xiao, W., and Davis, R. W. (2015). Mice are not men. *Proceedings of the National Academy of Sciences*, 112(4):E345–E345.

Washietl, S., Kellis, M., and Garber, M. (2014). Evolutionary dynamics and tissue specificity of human long noncoding rnas in six mammals. *Genome research*, 24(4):616–628.

Weirauch, M. T. and Hughes, T. R. (2010). Conserved expression without conserved regulatory sequence: the more things change, the more they stay the same. *Trends in Genetics*, 26(2):66–74.

Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., Klemm, A., Flicek, P., Manolio, T., Hindorff, L., et al. (2014). The nhgri gwas catalog, a curated resource of snp-trait associations. *Nucleic acids research*, 42(D1):D1001–D1006.

Wilhelm, M., Schlegl, J., Hahne, H., Gholami, A. M., Lieberenz, M., Savitski, M. M., Ziegler, E., Butzmann, L., Gessulat, S., Marx, H., et al. (2014). Mass-spectrometry-based draft of the human proteome. *Nature*, 509(7502):582–587.

Xue, Z., Huang, K., Cai, C., Cai, L., Jiang, C.-y., Feng, Y., Liu, Z., Zeng, Q., Cheng, L., Sun, Y. E., et al. (2013). Genetic programs in human and mouse early embryos revealed by single-cell rna [thinsp] sequencing. *Nature*, 500(7464):593–597.

Yanai, I., Graur, D., and Ophir, R. (2004). Incongruent expression profiles between human and mouse orthologous genes suggest widespread neutral evolution of transcription control. *Omics: a journal of integrative biology*, 8(1):15–24.

Yang, H., Bell, T. A., Churchill, G. A., and de Villena, F. P.-M. (2007). On the subspecific origin of the laboratory mouse. *Nature genetics*, 39(9):1100–1107.

Yang, H., Wang, J. R., Didion, J. P., Buus, R. J., Bell, T. A., Welsh, C. E., Bonhomme, F., Yu, A. H.-T., Nachman, M. W., Pialek, J., et al. (2011). Subspecific origin and haplotype diversity in the laboratory mouse. *Nature genetics*, 43(7):648–655.

Yue, F., Cheng, Y., Breschi, A., Vierstra, J., Wu, W., Ryba, T., Sandstrom, R., Ma, Z., Davis, C., Pope, B. D., et al. (2014). A comparative encyclopedia of dna elements in the mouse genome. *Nature*, 515(7527):355–364.

Zeisel, A., Muñoz-Manchado, A. B., Codeluppi, S., Lönnerberg, P., La Manno, G., Juréus, A., Marques, S., Munguba, H., He, L., Betsholtz, C., et al. (2015). Cell types in the mouse cortex and hippocampus revealed by single-cell rna-seq. *Science*, 347(6226):1138–1142.

Zheng-Bradley, X., Rung, J., Parkinson, H., and Brazma, A. (2010). Large scale comparison of global gene expression patterns in human and mouse. *Genome biology*, 11(12):1.