Published in final edited form as:

Nat Rev Genet. 2017 July ; 18(7): 425–440. doi:10.1038/nrg.2017.19.

Comparative transcriptomics in human and mouse

Alessandra Breschi[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

[3]Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11742

Abstract

Cross-species comparisons of genomes, transcriptomes and gene regulation are now feasible at unprecedented resolution and throughput, enabling the comparison of human and mouse biology at the molecular level. Insights have been gained into the degree of conservation between human and mouse at the level of not only gene expression, but also epigenetics and interindividual variation. However, a number of limitations exist, including incomplete transcriptome characterization and difficulties in identifying orthologous phenotypes and cell types, which are beginning to be addressed by emerging technologies. Ultimately, these comparisons will help identify the conditions under which the mouse is a suitable model of human physiology and disease and optimize the use of animal models.

For decades, the laboratory mouse (*Mus Musculus*) has been the preferred model organism for the study of human biology and diseases. Humans and mice share a very similar genetic background, and around 90% of both genomes can be partitioned into regions of conserved synteny[1]. 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 experience[2,3] and technological advances[4] in the generation of mutated mouse strains, hundreds of mouse models are currently available to mimic many human diseases[5], even those that are not naturally found in mice, such as cystic fibrosis and Alzheimer's. Recently, the creation of mouse model has largely improved through the CRISPR-Cas9 technology (clustered regularly interspaced palindromic repeat and CRISPR-associated endonuclease), which allows highly efficient genome editing by site- directed DNA endonucleases and can be performed directly on the zygotes, circumventing the need for a germline-competent embryonic stem cell line[6]. Mouse models are commonly used for research in diverse fields of biology (Box 1), ranging from neuro science and behavioural research to physiology and cancer research. The most recent official statistics from the European Committee[7] report that just under 11.5 million laboratory animals were used in Europe in 2011, 61% of which were mice. A UK governmental report shows that 1.16

Competing interests
The authors declare no competing interests.

million mice were used in the United Kingdom in 2014, which equates to 60% of the 1.93 million experimental procedures completed that year[8], with usage reported to be consistently at a similar level from 2005 onwards.

It is unsurprising that the mouse is the most commonly used species for scientific purposes. Clinical trials, in particular, rely heavily on non-human organisms, before testing a drug on patients, as proven efficacy in *in vivo* preclinical studies is essential for a drug to enter further clinical phases. Nonetheless, drugs often fail along the phases of clinical trials; for instance, 40% of the drugs investigated between 2003 and 2011 did not proceed to the second phase of testing, and only 10.4% of drug candidates are likely to get FDA approval[9]. In cancer research, specifically, the average rate of successful translation from animal models to human clinical trials is less than 8%[10], which mimics the difficulties in using mice as xenograft models of cancer[11].

The above highlights that although many core biological processes and genetic elements are conserved between human and mouse, other biological features have diverged substantially, leading to phenotypic differences and poorly correlated physiological responses between species. Diverging features can be genomic differences (such as retrotransposition events, gene expansions or gene losses, genomic rearrangements, differences in coding and non-coding sequences) or regulatory differences that affect gene expression and, ultimately, protein levels (such as alternative splicing, enhancer activity, structural elements such as chromatin domains, and post-translational modifications).

With the continuously decreasing cost and technical challenges of high-throughput sequencing technologies has come a growing effort to functionally characterize the human and mouse genomes, to identify what is shared and what has diverged between these two species. To this end, a series of large-scale projects has analysed a vast array of human and mouse samples, with the dual aim of understanding the principles of genomic regulation across different conditions and of comparing them between species. These projects include, but are not limited to, the Genotype-Tissue Expression (GTEx) project[12], which aims to establish a resource database and associated tissue bank to study the relationship between genetic variation and gene expression in human tissues, as well as the Roadmap Epigenomics project[13] and the Blueprint project[14], which aim to build a public resource of human epigenomic data. Other projects that are collecting human and mouse data simultaneously include the FANTOM project[15], which focuses mostly on Cap Analysis of Gene Expression (CAGE) profiles of human and mouse tissue and cell lines, and the human and mouse ENCODE projects[16,17], the scope of which is to catalogue all functional elements in the human and mouse genomes, respectively.

Characterizing gene expression profiles across multiple samples and species is instrumental to determine to what extent the biology of a given organism can be extrapolated to another. Thus, this Review centres on presenting an overview of the main findings of comparative molecular studies between human and mouse, with a focus on comparative transcriptomics, and how these studies illuminate the cases and conditions under which mouse is a suitable model of human biology. We also discuss the limitation of current approaches, which include incomplete transcriptome characterization, and difficulties in identifying homologous phenotypes and cell types, and how these can be addressed using emerging technologies.

1 Human and mouse genomes

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[18–20]. The most recent genome assemblies (GRC38) include 3.1 Gb and 2.7 Gb for human and mouse, respectively (Table 1), with the murine genome being 12% smaller than the human one. Around 90% of each genome can be partitioned into conserved syntenic regions, and 40% of the nucleotides in human can be aligned to mouse[20]. The remaining 60% of unalignable nucleotides might be attributed to lineage-specific deletion of repeated elements from the ancestral genome, nucleotide-level insertions and deletions and lineage- specific duplications[20].

1.1 Protein-coding genes

According to the latest release of GENCODE annotation[21] (v25, Ensembl86), which recently started to also curate the mouse genome[22] (vM11, Ensembl86), the human genome encodes 58,037 genes, of which approximately one-third are protein-coding (19,950), and which yields 198,093 transcripts. By comparison, the mouse genome encodes 48,709 genes, of which about one half are protein-coding (22,018 genes), and yields 118,925 transcripts overall (Table 1). For both species, the current number of protein-coding genes is about 10,000 genes lower than was estimated from early genome assembly drafts[18,20].

The discrepancy in the total number of annotated genes between the two species is unlikely to reflect differences in underlying biology, and can be attributed mostly to the less advanced state of the mouse annotation. The number of protein-coding and long non-coding RNAs (lncRNAs) encoded in the human and mouse genomes is expected to be very similar, and differences in the total genome length do not result from differences in the number of genes, but probably from differences in the lengths of introns and intergenic space[20] (Figure 1). Indeed, when including predicted gene models from RNA sequencing (RNA-seq) and CAGE data, the mouse annotation is expanded to a size that is similar to the human annotation[23]. 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[24] (15,893) (Figure 1), a number which is highly similar to the 15,736 orthologous genes derived after extensive curation efforts by the ENCODE consortium[25]. The remaining 20-30% protein-coding genes are either in one-to-many or many-to-many orthologous relationships, are members of gene families that have undergone species-specific expansions or reductions, or contain species-specific open reading frames (ORFs). These genes might contribute to human disease phenotypes and should therefore be taken into account when engineering mouse models[20]. For example, the human-specific gene saitohin (*STH*), which contains a single nucleotide polymorphism (Q7R) that is associated with susceptibility to several neurodegenerative diseases[26], has no orthologous gene in mice.

1.2 Long non-coding RNAs

Evidence for the importance of lncRNAs is continuously growing, and an increasing number of lncRNAs related to human diseases is discovered every year[27–29]. Identifying the possible mouse orthologues of human lncRNAs would greatly assist in the elucidation of their biological role.

Currently, there are 15,767 and 9,989 lncRNAs annotated by GENCODE in human and mouse, respectively[21,22]. The discrepancy, again, is a consequence of the less complete state of the mouse genome annotation. lncRNAs are usually expressed at a lower level than protein-coding genes and often in a very tissue-specific manner, which hinders their identification and leads to a requirement for additional resources to build a comprehensive annotation[30,31]. Finding orthologous relationships and conservation estimates for lncRNAs is also more challenging as their sequence is less conserved than that of protein-coding genes[30] and not constrained by amino-acid translation. In fact, the definition itself of lncRNA orthology is not as clear as for protein-coding genes and has so far been considered a combination of sequence and/or functional conservation and synteny [32]. Whereas RNA secondary structure might be useful to identify short non-coding RNAs and their degree of conservation, only few lncRNAs identified thus far have distinct structural domains as defined in Rfam[33,34]. Thus, current catalogues of orthologous lncRNAs are still highly incomplete and inaccurate[34], and the development of methods to identify lncRNA orthology constitutes an active field of investigation.

A number of studies in the past few years have attempted to identify novel lncRNAs in mice and other species and identify their orthologs in humans[23,35–37]. Although the gene sets may vary amongst the different studies, they produce a consistent estimate of approximately 1,000–2,000 orthologous lncRNAs between human and mouse. Necsulea et al.[36] report the highest number of human-mouse orthologous lncRNAs (2,720), based on sequence similarity of both novel and annotated transcripts, whereas Washietl et al.[37] identify 1,100 orthologous lncRNAs based on genome-wide chain alignments. Pervouchine et al.[23] reported 851 lncRNAs orthologs on the basis of a mixed approach including both genome alignments and sequence homology. A more recent study, which includes the information on syntenic blocks to call the orthology, reports 1,587 human-mouse orthologous lncRNAs[38]. However, the overlap between these studies is quite low: Pervouchine and colleagues[23] reported that only 189 orthologous lncRNAs are in common between their study and that of Necsulea et al.[36]. In all of these studies, orthologous lncRNAs represent only a small fraction of all annotated lncRNAs in both species, especially when compared to protein-coding genes.

About 5,000 lncRNA transcripts are in antisense orientation with respect to protein-coding genes in both mouse and human[39], and antisense transcription is known to have a role in the regulation of expression of the sense gene in a number of cases[40]. For example, an antisense transcript of the tumour suppressor gene *CDKN1A* recruits a regulatory complex that induces trimethylation of Lys27 of histone H3 (H3K27me3) to suppress the sense promoter region[40]. Although antisense transcription is largely present in both species, the proportion of orthologous sense–antisense pairs relative to all sense–antisense pairs is low (less than

20%, around 1,000 pairs[23,39]), suggesting low conservation of antisense transcription, and consequently of the corresponding biology.

1.3 Small non-coding RNAs

Compared to protein-coding and long non-coding RNAs, small non-coding RNAs, which include microRNAs (miRNAs), transfer RNAs (tRNAs), small nuclear RNAs (snRNAs) and small nucleolar RNAs (snoRNAs), have received less attention in comparative transcriptomics studies, partially because they are more difficult to monitor, with analyses limited to only a handful of tissues, such as brain, liver, kidney, heart and testis[41–44]. Small non-coding RNAs are known to be involved in the regulation of RNA processing, expression and translation[45,46], and there is growing evidence of their involvement in human diseases[29]. For example, alterations in miRNA expression can lead to several diseases, ranging from immune-related diseases, such as multiple sclerosis, to neurodegenerative diseases, such as Parkinson's disease[47], and cancer[48]. Thus, the use of specific mouse models to understand the mechanisms of small non-coding RNA involvement in diseases will certainly be beneficial[49]. For example, obese mice deficient in miR-375 developed severe insulin-deficient diabetes, suggesting that miR-375 is essential for mediating metabolic stress[49].

Currently, almost 3,000 and 2,000 miRNAs are annotated in the human and mouse genome, respectively[50] (Table 1). However, only a small fraction (300 miRNAs) of them has a defined ortholog in the other species[51].

tRNAs have a peculiar secondary structure that allows them to recognize mRNA codons by pairing to their anticodon and to carry an amino acid cognate to the tRNA[52]. Because of codon degeneracy for the 21 amino acids (including selenocysteine), multiple anticodons are related to the same amino acid (tRNAs isoacceptors). Human and mouse share 46 isoacceptors[43]. The number of predicted tRNA genes is similar between human and mouse (631 and 471 tRNA genes, respectively, Table 1)[53], as is the number of tRNA genes detected in human and mouse liver (223 and 224 tRNA genes, respectively)43. Although tRNA expression is conserved between the two species at the isotype level (tRNA isoacceptors related to the same amino acid), 34% of mouse tRNA genes cannot be aligned to human homologues, and only 79 tRNA genes are commonly expressed in liver samples[43], which suggests a certain degree of divergence in the evolution of tRNA genes.

snRNAs are essential elements of the spliceosome, and their expression levels are overall conserved between human and mouse[44]. snoRNAs contribute to biochemically modify specific sites of ribosomal RNA, tRNA and snRNA[45]. Of the 944 and 1,508 annotated human and mouse snoRNA genes[21], respectively (Table 1), at least 208 are conserved between the two species[44]. Of these, 63 snoRNA genes (30%) have distinct expression profiles[44], which indicates that the regulation of snoRNA genes has diverged considerably between the two species.

Further studies will certainly improve our understanding of the regulatory role and evolution of these RNA families, and hopefully of their involvement in diseases. Particularly relevant will be the understanding of the conservation between human and mouse of the relationship between the precursor long RNA molecules and the small functional RNA products. This, in particular, would extend the possibility of therapeutic interventions along the entire molecular path involved in the synthesis of small RNA molecules.

2 Conservation of transcriptomes

Similarities in the gene sets between two species do not necessarily reflect transcriptomic similarities, as the expression pattern of a gene across tissues and conditions can be very different across species. With the development of microarray technologies, and subsequently of RNA-seq, which enable 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 and splicing (Box 2) have been globally conserved between human and mouse.

2.1 Microarray studies

Most of the early microarray studies focused primarily on the expression of orthologous protein-coding genes in a variety of homologous tissues, such as brain, heart, muscle and liver. Under the assumption that mouse is a good model of human biology, one would expect higher similarity of gene expression in homologous organs between the species, than in different organs from the same species[54]. In other words, human liver would have an expression profile resembling that of mouse liver more than that of the human heart.

The relationship of transcriptomes from multiple RNA samples is usually visually rep resented using methods related to hierarchical clustering. In this approach, samples are given as the leaves of a dendrogram that is built on the basis of a given similarity measure between transcriptomes. This measure is usually the Euclidean distance between individual gene expression levels or the correlation coefficient across all genes between samples (FIG. 2). Alternative methods to visualize transcriptome relationships include dimensionality reduction techniques, such as principle component analysis (PCA [55]), multidimensional scaling (MDS[56]), or the more recently developed t-distributed Stochastic Neighbor Embedding (t-SNE[57]). These approaches project samples onto a two-dimensional (2D) or 3D space, where their distance is related to overall transcriptome similarity (FIG. 2e-g).

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 that emphasize proper use of normalization methods report a high conservation of expression between human and mouse tissues[54,58], such as brain, muscle, liver, kidney, lung and spleen, after correcting for array-specific differences in expression. By contrast, inaccurate normalization — for instance, failing to account for species specific systematic bias in signal intensity values in microarray probe sets — has been shown to spuriously exacerbate differences between species[59,60].

It is still under debate whether these results, supporting transcriptional conservation between humans and mice, but obtained in a limited number of samples, are generally applicable to any type of samples and to the whole transcriptome. For example, although induction and repression of major transcriptional regulators of erythropoiesis are conserved between mouse and human, significant transcriptional divergence between the two species has been detected at the transcriptome level[61]. Many transcriptional differences were also reported at the level of the immune and inflammatory response. These might be explained by *cis*-regulatory differences. For instance, although the macrophage response to lipopolysaccharide (LPS) is overall conserved between the two species, differential sets of genes are activated and repressed in mouse and human, a transcriptional plasticity that might be conferred by TATA-enriched and CpG island-depleted promoters[62]. By contrast, intraspecies differences in macrophage transcriptional response to glucocorticoid seem to be associated with gain and losses of glucocorticoid response elements[63]. In another study, it was shown that mouse transcriptional responses to different inflammatory stresses, including trauma, burns and endotoxemia, correlate poorly with the human ones, even though human transcriptional responses to different inflammatory stresses correlated well with each other[64]. This finding raised the serious question whether mouse is a good clinical model to study such conditions. This conclusion was challenged by a reanalysis of the same data that was restricted to a smaller set of genes with changes in expression levels that were conserved between human and mouse[65]. However, it has been noted that this approach introduces a bias in the results, and that the low percentage of genes with conserved changes in expression (12%) may itself be indicative of poor reproducibility of the human response in mice[66,67].

2.2 RNA-seq studies

The introduction of RNA-seq technology prompted more comparative transcriptomics studies at a deeper resolution and including larger numbers of species, since RNA-seq does not depend on first fabricating a species-specific spotted microarray (see[68] and[69] for reviews). Advantages of RNA-seq over microarray technology include its greater sensitivity, broader range of detection from lowly to higlhy expressed genes, and that it allows for annotation-independent detection of RNA abundances[70].

The Mouse ENCODE consortium[71] has been collecting around one hundred RNA-seq datasets for a range of mouse tissue and cell types, to create a comprehensive reference for future studies[17]. The profiled samples included almost 30 tissues, from adult mice and brain, nervous system, limbs and liver from embryos, as well as mouse cell lines, such as embryonic stem cells, murine erythroleukemia cells (MEL), and mouse lymphoma cells (CH12). Depending on the sample, they were collected and sequenced at different centres, and at least 2 replicates were sequenced for each sample. As in the case of microarray, clustering of mouse and human gene expression profiles from homologous tissues strongly depended on the normalization method applied[17]. However, as human data from comparable experimental conditions is not available (the bulk of human ENCODE transcriptome data was obtained in cell lines[16] whereas the mouse data were obtained from primary tissues), it is hard to disentangle the gene expression variation attributable to the species from that resulting from other biological factors or technical effects[17]. Simultaneous analysis of the human and mouse RNA data uncovered a large fraction of orthologous protein-coding genes (about 50%) with fairly constrained expression inde pendent from the sampled cell type in both human and mouse[23]. Analysis of human and mouse gene expression from a more homogeneous experimental setting, where samples are collected, processed and sequenced similarly at the same centre, however, argued 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 showed strong conservation between the two species[72–75]. By contrast, a study that also included organs expressing fewer tissue-specific genes[72], such as fat and stomach, showed that transcriptional patterns are overall different between human and mouse, separating the species more than the organs. This conclusion led to another highly charged debate, suggesting that other factors and biases, such as sequencing site, time of sequencing and the sequencing instrument used, need to be taken into account when undertaking comparative transcriptomics[76].

The analysis of additional vertebrate species at different phylogenetic distances to human and mouse, such as macaque, chimpanzee, opossum, platypus and chicken, affirmed the original conclusion that transcriptional patterns are more similar between orthologous organs of different species than between different organs from the same species[77,78,79]. These studies, however, were again based on organs expressing the largest numbers of organ-specific genes.

Taken together, these studies suggest that the question of whether mouse is overall a suitable model of human biology, based on transcriptome comparisons, is ill-posed. These works implicitly assume an average behaviour for genes, ignoring that each gene has a characteristic pattern of expression variation across species and organs (FIG. 1). This pattern has been recently investigated both between human and mouse[17] and across multiple species[80] (FIG. 2). In both studies linear models were used to decom pose the variation of gene expression in a set of homologous adult tissues across human and mouse only or across multiple mammals, including human and mouse, and chicken. Each gene exhibits its own pattern of variation across tissues and species. For example, the expression of the uromodulin gene (*UMOD*) is variable across tissues, but stable between human and mouse as it shows kidney-specific expression in both species[17]. By contrast, the gene encoding for the calcium-binding and coiled-coil domain-containing protein 2 (*CALCOCO2*) has a relatively constant expression across tissues in human, whereas is not detected in adult tissues in mouse, albeit it is expressed during embryonic developement[17,81], thus having more variation across species. Thus, a subset of genes was identified that varies a lot across tissues, but little across species, leading to tissue-dominated clustering, whereas another subset of genes varies a lot across species, but little across tissues, leading to a species-dominated clustering [82](FIG. 2c). Vertebrate (mouse) models of human biology may be particularly appropriate for the genes in the former set[83]. Remarkably, these genes are more likely to be associated with diseases than are genes whose expression varies predominantly across species [82].

2.3 lncRNA expression conservation

Most of the large-scale comparative studies of gene expression have been centred on orthologous protein-coding genes. Only in the last decade have comparative surveys of non-coding transcriptomes emerged, owing to the continuous expansion of lncRNA annotation[21,22]. Overall, orthologous lncRNAs between human and mouse have conserved levels of expression[23,35]. However, clustering analysis and PCA based on lncRNAs show more rapid evolution of expression patterns compared to protein-coding genes[36]. In addition, the breadth of expression is conserved not only between human and mouse but also in other mammals: ubiquitously expressed lncRNAs in human are ubiquitous across all species analysed, and tissue-specific lncRNAs in human are tissue-specific in all species[35,37]. However, these results might be influenced by the relatively low number of orthologous lncRNAs (less than 10% of all annotated lncRNAs) compared with orthologous protein-coding genes (75%). Most lncRNAs appear to be testis-specific in both species[35,37], especially the less conserved ones[36]. This is hypothetically related to a more permissive chromatin conformation during spermatogenesis[84], which could potentially contribute to the rapid evolution of testis transcriptomes. Therefore, organ-specific evolutionary rates of gene expression must be considered when evaluating whether the mouse transcriptome is a good model of the human transcriptome.

2.4 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. Overall, average gene expression levels correlate well between human and mouse, such that highly expressed genes in humans tend to be highly expressed in mice[85,86], even when very heterogeneous samples, such as cell lines and tissues, are considered[23]. Conservation of gene expression is to some extent reflected by sequence conservation in the gene body[85,87]. Promoter sequences, however, have diverged more than gene body sequences between mouse and human[86]. Depending on the method in which promoter sequence conservation is quantified -global or local sequence alignments or preserved presence of TF binding motifs - only a slight correlation between promoter conservation and gene expression conservation is observed[86]. Gene expression is predominantly conserved, even if the sequence of regulatory regions has diverged[88,89]. This might be due to compensatory mechanisms, for instance, two different transcription factors (TFs) in the two species acting on the expression of the same gene might activate it at comparable levels despite binding to two different regions[90].

3 Comparative gene regulation

Over the past 5 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[91]. However, the combinatorial complexity of gene regulatory factors (for example, histone modifications and TFs), the use of different sample types (tissues or cell lines), and the difficulties in associating specific regulatory regions with the regulated genes (which may be distal) make it extremely challenging to reach a comprehensive genome-wide map of regulatory elements. Most comparative experiments between human and mouse have been confined to a handful of TFs in a few cell types[92–96]. Nonetheless, these studies have revealed principles of cis-regulation which have subsequently been confirmed by larger studies. The Mouse ENCODE consortium has been collecting chromatin immunoprecipitation followed by sequencing (ChIP-seq) data for histone modifications and TF binding sites, DNAse-seq data for chromatin accessibility sites and replication timing data for chromatin domains for hundreds of different mouse tissue and cell types[17]. Although chromatin states inferred from histone modifications[97] and chromatin domains were highly similar between the two species, patterns of TF binding, as measured by ChIP-seq and inferred from DNase I hypersensitive sites (‘footprints’), were more diverged with only 22% TF footprints conserved[17].

3.1 TF binding

The primary consensus sequence motif for orthologous TFs is virtually the same in human and mouse[92,98], but secondary motifs often differ[99]. Of the 4 human secondary motifs with the strongest enrichment in the peaks inferred from TF ChIP-seq experiments, only some, if any, have a conserved sequence with mouse secondary motifs[98]. As secondary motifs often represent the consensus motifs for other TFs, the identity of associated factors might be lineage-specific[98]. Thus, the most commonly used motifs in one species may have binding capacity in the other species, with the caveat that the presence of the 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 a homologous sequence in the other species[98,100,101] and widely share the same relative distance to the transcription start site (TSS)[98]. Yet, only 10-20% of the TF-bound sites in one species are also bound in the other species[98,102].

Species-specific binding sites may arise from species-specific innovations or losses (FIG. 1). Novel TF binding sites and enhancers can arise from transposition of repeated elements[17,95,103] or by DNA exaptation[104]. Surprisingly, it has been shown that up to 40% of binding sites for the TF CCAAT/enhancer-binding protein alpha (CEBPA) that have no binding in human but are present in mouse have an unchanged sequence [94] By contrast, the loss of TF binding occupancy in aligned regions is, in about 50% of cases, compensated by another binding motif within 10 kb[94], so the main regulatory circuits of gene regulatory networks are maintained. Indeed, tissue and cell type specific gene regulatory networks of TFs in mouse, inferred from genomic DNase I footprints, are highly similar to the networks in human homologous tissues and cell types, with more than 40% conserved TF-to-TF regulatory connections[105]. This finding suggests that conservation of functional regulatory circuitry is considerably greater than indicated by sequence conservation alone[105,106] (FIG. 1). In addition, TF binding sites in one species are often repurposed in other species; it has been computed that 48% and 57% of homologous sites are bound in the other species for human and mouse, respectively, such that a sequence is bound either by the same TF in a different cell type or by different TFs in the same cell type[100]. Furthermore, 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[98].

Taken together these findings suggest that although the relationships between TFs and their targets are conserved between human and mouse, the activity of specific regulatory DNA elements, such as enhancers and promoters, in one species cannot be inferred from sequence homology and consensus motifs in the other species alone. This is also the case for many lncRNAs, which show species-specific expression even if located in regions of conserved synteny[38]. In fact, only functional validation experiments can confirm the reliability of cross-species-predicted TF binding sites[107]. Screening strategies have been developed for testing the *in vivo* activity of enhancers using transgenic mouse embryos, which also allows the assessment of their tissue specificity[108]. Over the years, a database has been assembled containing the results for almost 3,000 tested enhancers that are orthologous between human and mouse[109], as a freely available resource for the scientific community.

3.2 Inferring human SNP causality from mouse regulatory regions

Ultimately, enhancers and TF binding sites in mouse can be a good proxy to find functional genomic regions implicated in human traits, for instance, genome-wide association studies (GWAS) single nucleotide polymorphisms (SNPs)[110]. Specifically, if a human variant identified in a human GWAS study can be mapped to an orthologous region within the mouse genome, its overlap with functional elements in mouse, such as enhancers, can be investigated. Promisingly, more than 4,000 SNPs from human GWAS studies have been mapped uniquely onto the mouse genome [17]. As an encouraging example, SNPs associated with traits related to liver function (such as HDL cholesterol levels and alcohol dependence) in humans reside in liver-specific enhancers in mouse [17]. Similarly, SNPs associated with traits related to urate levels in humans reside in kidney-specific enhancers in mouse[17]. Thus, mouse can be a useful model to gain better insights into the causality of SNPs identified in human GWAS.

4 Intraspecies expression variation

Between 4 and 5 million SNPs differentiate each person from the human reference genome[111] and a conservative estimate postulates that the genomes of any two individuals differ by at least 0.5%[112]. How this variation affects molecular features, such as gene expression, and ultimately phenotypes, is currently a topic of active research, especially within consortium-led projects like the Geuvadis[113], the GTEx[114], and others. For instance, the GTEx project (http://gtexportal.org) identified 199,362 mutations that affected the expression of 27,159 genes in at least one of 44 human tissues (release V6p). The major stratification of variation within the human species is at the level of populations. The concept of interindividual variation in laboratory mice is less straightforward, since the *Mus musculus* species have multiple layers of stratification due to human intervention. Laboratory strains can be classified into classical inbred strains and wild-derived strains[115], with the former being characterized by at least 98.6% homozygous loci in each individual[116]. Classical inbred strains are mosaics of a handful of haplotypes derived from mice generated from wild subspecies[117], with more than 90% of their genetic background coming from *M. m. domesticus*[115,118].

To quantify the genetic variation between strains, the Mouse Genomes Project sequenced and catalogued a number of classical inbred and wild-derived strains[119]. Variation within the reference genome strain is negligible as it is virtually indistinguishable from the sequencing error rate[120]. Also the variation between mice of the same strain, but created from different centres, is very low (fewer than 10,000 SNPs[119]), although phenotypic differences in behaviour have been reported, due both to subtle genetic differences between substrains and to environmental factors, such as the order of testing and inter-test interval[121,122]. Interstrain variation, however, is more pronounced, with around 4-5 million SNPs between the mouse reference genome and any other classical inbred strain[119,123]; considering that these SNPs are limited to the 85% of uniquely mappable genomic sequences and that the mouse genome size is smaller than that of human, this variation is higher than interindividual variation amongst humans. Finally, the mouse reference genome differs from other wild-derived strains by at least 17 million SNPs, with the exception of strains derived from *M. m. domesticus*[119].

In analogy to genetic variation, there is relatively little variation in terms of gene expression both between classical inbred strains[124–126] and within the same strain[127] in different tissues. These differences in gene expression are not necessarily related to the diverse genetic background, as many environmental factors (for example, progressive removal of littermates from the cage) can temporarily alter the gene expression profiles of individual mice[127]. Thus, it is very important to select a proper mouse population to understand mouse intraspecific variation, possibly from outbred wild-caught mice, and compare it to human. The Mouse Phenome Database, which originally integrated phenotype data from 40 inbred strains, has recently started to introduce data from the Collaborative Cross and Diversity Outbred mice[128]. These mice present extensive genetic variation from eight founder inbred strains, and a variety of molecular data is collected from them to understand the impact of genotypic diversity in mice[129]. This approach recently led to the identification of 4,188 mouse expression QTLs (eQTLs) [130]: the identification of causal variants in mice can help tailoring mouse models with specific mutations for human-relevant phenotypes inserted into a defined genetic background., However, this will require a comprehensive mapping of eQTLs from one species to the other, which is still lacking, to the best of our knowledge, although significant overlap of orthologous genes affected by eQTLs in CD4+ T cells from healthy humans and from a panel of the most common inbred mouse strains has been reported[126].

The use of inbred strains to uncover relationships between genotype and gene expression is more suited to experiments on allele-specific expression than comparative transcriptomes. In hybrid mice generated from two distinct inbred strains, maternal and paternal genotypes can be readily tracked. In fact, with more than 450 inbred strains[116], carefully annotated by the Jackson Laboratory[5] (http://www.informatics.jax.org), RNA production from only one allele can be easily detected and compared across multiple tissues[119,131.]

5 Cellular complexity of mammalian organs

A vast proportion of transcriptomics studies in human and mouse, especially the comparative ones, has been focused on profiling gene expression at the organ or 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 of an organized mixture of different cell types, whose concerted genomic activity establishes the proper functioning of the organ 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[132], 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 of unwanted variation in gene expression between organs and between species[133] (Figure 3). 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. For example, clustering analysis have revealed that populations of human and mouse primary cells of a given type have distinctive expression profiles[134,135]. Therefore, it is extremely important to deconvolute qualitatively and quantitatively which cell populations contribute to the global expression patterns of organs[136].

Most transcriptomics studies on mammalian primary cells are based on meta-analyses, largely of microarray data from disparate sources, which, despite the use of normalization methods, carries technical noise and reduced sensitivity. The FANTOM consortium released the largest organized atlas of promoter (and gene) expression data[15] in hundreds of human and mouse primary cells and tissues. However, a systematic comparative analysis between the two species, including a large panel of cell types and conditions, is still lacking at the resolution of cell populations. Such an analysis 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 that are expressed in the alpha cells and beta cells of pancreatic islets, *GC* (encoding group-specific component (vitamin D binding protein)) and *DLK1* (encoding delta like non-canonical Notch ligand 1), have opposite cell-specific expression in human and mouse[137].

Expression data from purified populations of primary cells provide higher resolution than whole-tissue transcriptomes, being robust to stochastic variability between cells[138]. Recent advancements in single-cell technologies[139,140], such as single-cell RNA-seq, enable researchers to obtain gene expression data for rare cell types — the signals of which are usually masked at the population level — to identify novel cell types with previously unknown markers, and to characterize cell differentiation stages[141]. Due to noticeable experimental challenges in disaggregating solid tissues, especially of human samples, most single-cell RNA-seq research has focused on mouse solid tissues, including brain[142], lung[143] and intestine [144], although a small number of studies have analysed human samples from pancreatic islets[137], brain[145] and blood[146,147]. In this regard, the Human Cell Atlas consortium (https://www.humancellatlas.org/) is being formed to create comprehensive reference maps of all human cells using multiple molecular assays, including RNA-seq. Additionally, single-cell RNA-seq has been applied to investigate RNA dynamics over time, especially in the early stages of life. For example, more than 1,000 single cells from the mouse epiblast were collected in a study from early gastrulation at embryonic day 6.5 to day 7.75 to investigate mesodermal lineage differentiation towards the hematopoietic system[148], just days after fertilization [148,149].

Despite the growing bulk of projects employing single-cell RNA-seq, as with cell population data, very few compare human and mouse single-cell expression. One complication may be the intrinsic difficulty of obtaining comparable samples from homologous organs or identifying homologous dynamic processes. A recent study compared the genetic programmes of human and mouse early embryos, in the developmental stages between oocytes and morula[150]. The authors observed that while global gene expression profiles were conserved, the actual developmental timing of expression differed between the two species[151]. Ultimately, comparing human and mouse transcriptomes at the single-cell level will help to identify previously undescribed conserved cell types, overcome the biases of different cell type composition and help to understand conserved and diverged elements of temporal dynamics. Albeit promising, this will require the development of specific computational methods that deal with the complexity of single-cell data and integrate it with the additional dimension of cross-species comparison.

6 Conclusions and perspectives

The rise of next-generation sequencing technologies in the past years has considerably advanced the field of comparative genomics, transcriptomics and epigenomics.

These approaches are particularly important to study the evolution of gene regulation in model organisms, to gain deeper insights into the degree of their conservation with human at the molecular level, and how this conservation correlates with conservation at the phenotypic level. Ultimately, this knowledge can help to understand to what extent a given animal model is suitable for the study of a specific biological process or condition.

A considerable amount of work, including efforts from international consortium projects such as mouse ENCODE[17] and FANTOM[15], has been centred on the laboratory mouse owing to its indisputable relevance as a model for human biology and diseases. Emerging from this wealth of data is a complex picture that underlies the difficulties associated with mapping the conservation of transcriptional patterns to the conservation of phenotypic traits. At the root of the problem is the difficulty of matching phenotypes across species, and therefore of quantifying phenotypic differences between species, which can then be correlated to transcriptional differences. This is even the case for apparently straightforward phenotypes, such as those affecting individual tissues, organs or anatomical sites. Indeed, tissues are complex structures composed of many primary cell types, and it is unclear whether equivalent cell types remain orthologous, to what extent the relative abundances of the populations of these cells types have been conserved among the species, or whether the tissue sample sectioned in the different species retains the same underlying tissue substructure.

Moreover, gene expression is affected by an almost unlimited number of biological factors, including sex[152], age[153], circadian rhythms (that is, recent research suggests that about half of all mammalian genes are subject to circadian regulation[154]), ischemic time and RNA integrity[155] or environmental factors. Many of these biological factors are very difficult to control, even when analysing apparently orthologous tissues. If, for instance, the biological age or the time of day at which the tissues have been collected differs between species, this may artificially exaggerate transcriptional differences, beyond those that can be uniquely attributed to the species.

This problem may be exacerbated in the case of more complex phenotypes, such as developmental or differentiation processes, response to external stimuli or insults, behaviour or systemic diseases. Hence, because it is technically very difficult to identify orthologous phenotypes, transcriptomes monitored in different species will likely overestimate the true interspecies transcriptional differences.

Single-cell genomics may contribute to addressing some of these issues. The unbiased identification of populations of cells sharing a similar phenotype could help to match these populations across species (that is, by using orthologous specific markers). In addition, new methodologies are emerging that preserve spatial information about the tissue context or subcellular localization of analysed nucleic acids[156]. Although spatial transcriptomics is still in its early days[157,158], it carries the promise of revolutionizing the way multicellular complexes, such as organs, are studied and might reveal new insights into the conservation of how these complexes are organized between human and mouse.

This should lead to biologically more meaningful transcriptome comparisons.

By contrast, most comparative transcriptome studies have focused on the patterns of gene expression of protein-coding genes, that is, on the genomic elements most strongly conserved across species. However, lncRNAs, as well as other non-coding transcriptional elements, such as small RNAs, pseudogenes, repetitive elements and others, are emerging as important players in the biology of organisms. These elements are less conserved across species than protein-coding genes, and orthology is difficult to determine or simply does not exist. Generally, they are poorly characterized from the transcriptional standpoint. As the expression patterns of the non-coding transcriptome are known to be more species-specific than those of protein-coding genes[30], transcriptional comparisons based on the latter (the vast majority, so far) are likely to overestimate interspecies similarities. Not accounting for this non-coding transcription may partially underlie the poor extrapolability of some mouse models. Remarkably, although the prevalent view is still that proteins are the main effectors of biological function, a comprehensive proteomics comparison between human and mouse is still lacking, with available studies being so far limited to a few specific samples[159].

Increasing the number of transcriptomic elements monitored, as well as that of orthologous conditions and phenotypes will in practice generate a large (almost infinite) data matrix[160], in which rows can be seen as conditions and columns as genomic elements (such as genes, transcripts and other transcriptional elements, but also epigenomic elements, such as TF binding sites or histone modifications), with a third dimension representing the species, and a fourth dimension representing dynamic processes (FIG. 4). The matrix is currently quite sparse, even if considering only human and mouse. For instance, there is little comparative data about transcriptional changes associated with processes occurring over time, such as differentiation and development[161,162], or with cellular and organismic responses to external stimuli. Indeed, there is some evidence that inducible genes might be responsible for gene expression divergence between species[63], although such genes are more challenging to be identified because similar perturbations need to be applied on homologous systems. This could be especially important for clinical studies, for example, to study the time of physiological responses to drugs or the progression of a disease. The deconvolution of such a data matrix, which is certainly challenging from the analytical standpoint, will contribute to understanding the transcriptome determinants underlying phenotypic similarities and differences between species. While still far from such a goal, data currently available, which we have reviewed here, strongly suggest that the question of whether mouse is overall a good model of human biology is an ill-posed question that does not have a binary answer. It clearly depends on the phenotype of interest, the genes involved in the phenotype and the tissues and organs in which these genes are expressed.

In the era of precision medicine, each individual may come to have his or her genome sequenced, and possibly be subjected to multiple genomics assays analysing 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 basis, and customized mouse models might be generated that are tailored to an individual. Understanding what part of mouse biology (or of the biology of any model organism) can be extrapolated to humans, and under which circumstances, is of crucial importance not only for improving therapeutic interventions, but also to optimize the use of animal models and decrease the economical and ethical costs associated with animal research. We caution that as many factors as possible should be matched when mouse models are used to study human physiology or disease[67].

Acknowledgements

We acknowledge Giuseppina Schiavo and Davide Cirillo for insightful discussion. We acknowledge support of the Spanish Ministry of Economy and Competitiveness, ”Centro de Excelencia Severo Ochoa 2013-2017”, of the CERCA Programme / Generalitat de Catalunya, and of Programa de Ayudas FPI del Ministerio de Economia y Competitividad-BES-2012-055848.

Biography

Alessandra Breschi

Alessandra Breschi received her Ph.D. in 2016 from the Pompeu Fabra University (UPF) in Barcelona, and is now a postdoctoral researcher at the Center for Genomic Regulation (CRG) in Barcelona in the group of Dr. Roderic Guigó. During her Ph.D. she participated in several consortium projects, such ENCODE and BluePrint. Her main research interest is understanding how regulation of gene expression affects phenotypic differences across cell types, individuals and species.

Thomas R. Gingeras

Thomas R. Gingeras received his Ph.D. in 1976 from the New York University, and his postdoctoral studies at Cold Spring Harbor Laboratory under Dr. Richard J. Roberts. While there, he sequenced the first DNA tumor virus and wrote a collection of the first bioinformatics software tools including one of the first genome sequence assemblers. In 1993, he moved to Affymetrix, Inc. and developed high density tiling arrays to study whole human genome transcriptional profiles. In 2010, his discovery of the pervasive transcription in human cells (the Dark Genome) using such arrays was cited as the Scientific Breakthrough and Insights of Decade by Science Magazine. He currently holds the position of Professor and Head of Functional Genomics at Cold Spring Harbor Laboratory.

Roderic Guigó

Roderic Guigó obtained his PhD in 1988 for work on Evolutionary Ecology from the University of Barcelona. He did postdoctoral research in Computational Genomics with

Temple F. Smith at Harvard and Boston Universities, and with James W. Fickett at Los Alamos National Laboratory in New Mexico (US). In 1994 he joined City Institute for Medical Research (IMIM) in Barcelona. Currently, he coordinates the Bioinformatics and Genomics program at the Center for Genomic Regulation (CRG), and he is a Bioinformatics Professor at Pompeu Fabra University in Barcelona. He participate in the human genome project, as well as in many other genomic and functional genomics projects, such as ENCODE, GTEx and BluePrint. Dr Guigó ‘s main research interests are in the understanding and modeling the regulated production of RNA in eukaryotic cells.

Glossary

Synteny

Preserved genomic order and orientation of genes or other elements between species

Xenograft models of cancer

Are created when cancerous tissue from a patient’s primary tumor is implanted directly into an immunodeficient mouse

Cap Analysis of Gene Expression (CAGE) profiles

In CAGE short (˜20 nucleotide) sequence tags originating from the 5’ end of full-length mRNAs are sequenced to identify transcription events on a genome-wide scale

Orthologous

Pertains to homologous genes in different species that have evolved from a common ancestral gene by speciation

GENCODE annotation

The GENCODE project produces high quality reference gene annotation and experimental validation for human and mouse genomes

Long non-coding RNAs (lncRNAs)

Non-protein coding transcripts longer than 200 nucleotides. This somewhat arbitrary limit distinguishes llncRNAs from small regulatory RNAs

MicroRNAs (miRNAs)

Derived from primary transcripts with features similar to mRNAs that are enzymatically processed to their mature length of 21-24 nucleotides by Drosha and Dicer enzymes

Transfer RNAs (tRNAs)

Adaptor RNA molecules (long 76-90 nucleotides) which serve as the physical link between the mRNA and the amino acid sequence of proteins, by carrying an amino acid to the ribosome as directed by the codon in a messenger RNA

Small nuclear RNAs (snRNAs) and small nucleolar RNAs (snoRNAs)

Classes of short non-coding RNAs (100-200 nt) that have important regulatory roles in nuclear ribonucleoprotein complexes

Homologues

A pair of genes that descended from a common ancestral gene

Hierarchical clustering

A statistical method in which objects (for example, gene expression profiles for different individuals or tissue samples) are grouped into a hierarchy, which is visualized in a dendrogram. Objects close to each other in the hierarchy, measured by tracing the branch heights, are also close by some measure of distance — for example, between gene expression profiles. Individuals or samples with similar expression profiles will be close together in terms of branch lengths

Euclidean distance

The Euclidean distance between points p and q is the length of the line segment connecting them in a multi-dimensional space In gene expression analysis, p and q are usually vectors of expression values in two samples/conditions

Dimensionality reduction techniques

Reduce multidimensional data to a minimal number of dimensions for visualization by identifying those dimensions that capture the most important information underlying the data structure

Principal Component Analysis (PCA)

Orthogonal linear transformation that transforms the original data to a new coordinate system such that the greatest variance of the projected data comes to lie on the first coordinate (called the first principal component), the second greatest variance on the second coordinate, and so on

Multidimensional Scaling (MDS)

Technique used to display the information contained in a distance matrix, which aims to place each object in N-dimensional space such that the between-object distances are preserved as well as possible

t-distributed Stochastic Neighbor Embedding (t-SNE)

Nonlinear dimensionality reduction technique based on the probability distribution over pairs of high-dimensional objects which are embedded into a space of two or three dimensions. Similar objects are modeled by nearby points and dissimilar objects are modeled by distant points

Normalization

Methods used to adjust measurements so that they can be appropriately compared among samples. For example, in microarray analysis, methods such as quantile normalization manipulate common characteristics of the data

Chromatin domains

Functionally distinct chromosomal regions, which confer structural organization to eukaryotic genomes, representing regulatory units for gene expression and chromosome behavior

DNA exaptation

The shift in the function of a DNA sequence during evolution

Allele-specific expression

Expression variation between the two haplotypes of a diploid individual distinguished by heterozygous sites

Ischemic time

In the case of organ donors, the time elapsed between the r donor death and the organ extraction

Pseudogenes

Segments of DNA that originate from functional genes, but have lost at least some of the ability of the parent gene in terms of expression or coding potential

Precision medicine

Emerging approach for disease treatment and prevention that takes into account individual variability in genes, environment, and lifestyle for each person

Expression QTL (eQTL)

Genomic locus that contribute to variation in expression levels of mRNAs

References

1. Chinwalla AT et al. Initial sequencing and comparative analysis of the mouse genome. Nature 420, 520–562 (2002) [PubMed: 12466850] This article comprehensively characterizes the initial sequence of the mouse genome and is still a valuable reference for comparative genomics.
2. Adams DJ & van der Weyden L Contemporary approaches for modifying the mouse genome. Physiological genomics 34, 225–238 (2008). [PubMed: 18559964]
3. Bedell MA, Jenkins NA & Copeland NG Mouse models of human disease. Part I: techniques and resources for genetic analysis in mice. Genes and Development 11, 1–10 (1997). [PubMed: 9000047]
4. Singh P, Schimenti JC & Bolcun-Filas E A mouse geneticist's practical guide to CRISPR applications. Genetics 199, 1–15 (2015). [PubMed: 25271304]
5. Bult CJ et al. Mouse genome database 2016. Nucleic acids research 44, D840–D847 (2016). [PubMed: 26578600]
6. Qin W et al. Generating Mouse Models Using CRISPR-Cas9-Mediated Genome Editing. Current protocols in mouse biology, 39–66 (2016). [PubMed: 26928663]
7. http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52013DC0859&from=EN.
8. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/469508/spanimals14.pdf.
9. Hay M, Thomas DW, Craighead JL, Economides C & Rosenthal J Clinical development success rates for investigational drugs. Nature biotechnology 32, 40–51 (2014)Comprehensive survey of clinical success rates across the drug industry.
10. Mak I, Evaniew N & Ghert M Lost in translation: animal models and clinical trials in cancer treatment. Am J Transl Res 6, 114–8 (2014). [PubMed: 24489990]
11. Morgan RA Human tumor xenografts: the good, the bad, and the ugly. Molecular Therapy 20, 882 (2012). [PubMed: 22549804]
12. Lonsdale J et al. The genotype-tissue expression (GTEx) project. Nature genetics 45, 580–585 (2013). [PubMed: 23715323]
13. Kundaje A et al. Integrative analysis of 111 reference human epigenomes. Nature 518, 317–330 (2015). [PubMed: 25693563]

14. Abbott A Europe to map the human epigenome. Nature 477, 518 (2011). [PubMed: 21956305]
15. The FANTOM Consortium et al. A promoter-level mammalian expression atlas. Nature 507, 462–470 (2014). [PubMed: 24670764]
16. ENCODE Project Consortium et al. An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57–74 (2012). [PubMed: 22955616]
17. Yue F et al. A comparative encyclopedia of DNA elements in the mouse genome. Nature 515, 355–364 (2014). [PubMed: 25409824]
18. Lander ES et al. Initial sequencing and analysis of the human genome. Nature 409, 860–921 (2001). [PubMed: 11237011]
19. Venter JC et al. The sequence of the human genome. science 291, 1304–1351 (2001). [PubMed: 11181995]
20. Chinwalla AT et al. Initial sequencing and comparative analysis of the mouse genome. Nature 420, 520–562 (2002). [PubMed: 12466850]
21. Harrow J et al. GENCODE: the reference human genome annotation for The ENCODE Project. Genome research 22, 1760–1774 (2012). [PubMed: 22955987]
22. Mudge JM & Harrow J Creating reference gene annotation for the mouse C57BL6/J genome assembly. Mammalian Genome 26, 366–378 (2015). [PubMed: 26187010]
23. Pervouchine D et al. Enhanced Transcriptome Maps from Multiple Mouse Tissues Reveal Evolutionary Constraint in Gene Expression for Thousands of Genes. Nat Commun 6 (2015).
24. Herrero J et al. Ensembl comparative genomics resources. Database 2016, bav096 (2016). [PubMed: 26896847]
25. Yue F et al. A comparative encyclopedia of DNA elements in the mouse genome. Nature 515, 355–364 (2014) [PubMed: 25409824] This paper from the Mouse ENCODE consortium presents an extensive catalog of mouse DNA elements identified through hundreds of NGS assays.
26. Wang Y, Gao L, Conrad CG & Andreadis A Saitohin, which is nested within the tau gene, interacts with tau and Abl and its human-specific allele influences Abl phosphorylation. Journal of cellular biochemistry 112, 3482–3488 (2011). [PubMed: 21769920]
27. Shi X, Sun M, Liu H, Yao Y & Song Y Long non-coding RNAs: a new frontier in the study of human diseases. Cancer letters 339, 159–166 (2013). [PubMed: 23791884]
28. Wapinski O & Chang HY Long noncoding RNAs and human disease. Trends in cell biology 21, 354–361 (2011). [PubMed: 21550244]
29. Esteller M Non-coding RNAs in human disease. Nature Reviews Genetics 12, 861–874 (2011).
30. Derrien T et al. The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. Genome research 22, 1775–1789 (2012). [PubMed: 22955988]
31. Cabili MN et al. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes & development 25, 1915–1927 (2011). [PubMed: 21890647]
32. Ulitsky I Evolution to the rescue: using comparative genomics to understand long non-coding RNAs. Nature Reviews Genetics 17, 601–614 (2016)Review on current strategies to identify lncRNAs and their function through comparative analysis across different species.
33. Nawrocki EP et al. Rfam 12.0: updates to the RNA families database. Nucleic acids research, gku1063 (2014).
34. Pignatelli M et al. ncRNA orthologies in the vertebrate lineage. Database 2016, bav127 (2016). [PubMed: 26980512]
35. Hezroni H et al. Principles of long noncoding RNA evolution derived from direct comparison of transcriptomes in 17 species. Cell reports 11, 1110–1122 (2015). [PubMed: 25959816]
36. Necsulea A et al. The evolution of lncRNA repertoires and expression patterns in tetrapods. Nature 505, 635–640 (2014). [PubMed: 24463510]
37. Washietl S, Kellis M & Garber M Evolutionary dynamics and tissue specificity of human long noncoding RNAs in six mammals. Genome research 24, 616–628 (2014). [PubMed: 24429298]
38. Chen J et al. Evolutionary analysis across mammals reveals distinct classes of long non-coding RNAs. Genome biology 17, 1 (2016). [PubMed: 26753840]

39. Engström PG et al. Complex loci in human and mouse genomes. PLoS Genet 2, e47 (2006). [PubMed: 16683030]

40. Faghihi MA & Wahlestedt C Regulatory roles of natural antisense transcripts. Nature reviews Molecular cell biology 10, 637–643 (2009). [PubMed: 19638999]

41. Roux J, Gonza` lez-Porta M & Robinson-Rechavi M Comparative analysis of human and mouse expression data illuminates tissue-specific evolutionary patterns of miRNAs. Nucleic acids research 40, 5890–5900 (2012). [PubMed: 22457063]

42. Meunier J et al. Birth and expression evolution of mammalian microRNA genes. Genome research 23, 34–45 (2013). [PubMed: 23034410]

43. Kutter C et al. Pol III binding in six mammals shows conservation among amino acid isotypes despite divergence among tRNA genes. Nature genetics 43, 948–955 (2011). [PubMed: 21873999]

44. Zhang B et al. Changes in snoRNA and snRNA Abundance in the Human, Chimpanzee, Macaque, and Mouse Brain. Genome biology and evolution 8, 840–850 (2016). [PubMed: 26926764]

45. Matera AG, Terns RM & Terns MP Non-coding RNAs: lessons from the small nuclear and small nucleolar RNAs. Nature reviews Molecular cell biology 8, 209–220 (2007). [PubMed: 17318225]

46. Huang Y et al. Molecular functions of small regulatory noncoding RNA. Biochemistry (Moscow) 78, 221–230 (2013). [PubMed: 23586714]

47. Li Y & Kowdley KV MicroRNAs in common human diseases. Genomics, proteomics & bioinformatics 10, 246–253 (2012).

48. Lin S & Gregory RI MicroRNA biogenesis pathways in cancer. Nature Reviews Cancer 15, 321–333 (2015). [PubMed: 25998712]

49. Park CY, Choi Y & McManus MT Analysis of microRNA knockouts in mice. Human molecular genetics, ddq367 (2010).

50. Kozomara A & Griffiths-Jones S miRBase: integrating microRNA annotation and deep-sequencing data. Nucleic acids research, gkq1027 (2010).

51. Landgraf P et al. A mammalian microRNA expression atlas based on small RNA library sequencing. Cell 129, 1401–1414 (2007). [PubMed: 17604727]

52. Goodenbour JM & Pan T Diversity of tRNA genes in eukaryotes. Nucleic acids research 34, 6137–6146 (2006). [PubMed: 17088292]

53. Chan PP & Lowe TM GtRNAdb: a database of transfer RNA genes detected in genomic sequence. Nucleic acids research 37, D93–D97 (2009). [PubMed: 18984615]

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

55. Jolliffe I Principal component analysis (Wiley Online Library, 2002).

56. Cox MA & Cox TF in Handbook of data visualization 315–347 (Springer, 2008).

57. Maaten L. v. d. & Hinton G Visualizing data using t-SNE. Journal of Machine Learning Research 9, 2579–2605 (2008).

58. McCall MN, Uppal K, Jaffee HA, Zilliox MJ & Irizarry RA The Gene Expression Barcode: leveraging public data repositories to begin cataloging the human and murine transcriptomes. Nucleic acids research 39, D1011–D1015 (2011). [PubMed: 21177656]

59. Liao B-Y & Zhang J Evolutionary conservation of expression profiles between human and mouse orthologous genes. Molecular biology and evolution 23, 530–540 (2006) [PubMed: 16280543] First paper highlighting the importance of normalization in comparative transcriptomics studies.

60. Yanai I, Graur D & Ophir R Incongruent expression profiles between human and mouse orthologous genes suggest widespread neutral evolution of transcription control. Omics: a journal of integrative biology 8, 15–24 (2004). [PubMed: 15107234]

61. Pishesha N et al. Transcriptional divergence and conservation of human and mouse erythropoiesis. Proceedings of the National Academy of Sciences 111, 4103–4108 (2014).

62. Schroder K et al. Conservation and divergence in Toll-like receptor 4-regulated gene expression in primary human versus mouse macrophages. Proceedings of the National Academy of Sciences 109, E944–E953 (2012).

63. Jubb AW, Young RS, Hume DA & Bickmore WA Enhancer turnover is associated with a divergent transcriptional response to glucocorticoid in mouse and human macrophages. The Journal of Immunology 196, 813–822 (2016). [PubMed: 26663721]

64. Seok J et al. Genomic responses in mouse models poorly mimic human inflammatory diseases. Proceedings of the National Academy of Sciences 110, 3507–3512 (2013).

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

66. Warren HS et al. Mice are not men. Proceedings of the National Academy of Sciences 112, E345–E345 (2015).

67. Shay T, Lederer JA & Benoist C 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, E346–E346 (2015).

68. Romero IG, Ruvinsky I & Gilad Y Comparative studies of gene expression and the evolution of gene regulation. Nature Reviews Genetics 13, 505–516 (2012).

69. Necsulea A & Kaessmann H Evolutionary dynamics of coding and non-coding transcriptomes. Nature Reviews Genetics 15, 734–748 (2014)Review on comparative transcriptomics studies in vertebrates.

70. Mortazavi A, Williams BA, McCue K, Schaeffer L & Wold B Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nature methods 5, 621–628 (2008). [PubMed: 18516045]

71. Stamatoyannopoulos JA et al. An encyclopedia of mouse DNA elements (Mouse ENCODE). Genome biology 13, 1 (2012).

72. Lin S et al. Comparison of the transcriptional landscapes between human and mouse tissues. Proceedings of the National Academy of Sciences 111, 17224–17229 (2014).

73. Sudmant PH, Alexis MS & Burge CB Meta-analysis of RNA-seq expression data across species, tissues and studies. Genome biology 16, 1–11 (2015). [PubMed: 25583448]

74. Su AI et al. Large-scale analysis of the human and mouse transcriptomes. Proceedings of the National Academy of Sciences 99, 4465–4470 (2002).

75. Chan ET et al. Conservation of core gene expression in vertebrate tissues. Journal of biology 8, 1 (2009).

76. Gilad Y & Mizrahi-Man O A reanalysis of mouse ENCODE comparative gene expression data. F1000Research 4 (2015).

77. Brawand D et al. The evolution of gene expression levels in mammalian organs. Nature 478, 343–348 (2011). [PubMed: 22012392]

78. Barbosa-Morais NL et al. The evolutionary landscape of alternative splicing in vertebrate species. Science 338, 1587–1593 (2012). [PubMed: 23258890]

79. Merkin J, Russell C, Chen P & Burge CB Evolutionary dynamics of gene and isoform regulation in Mammalian tissues. Science 338, 1593–1599 (2012) [PubMed: 23258891] References 78 and 79 are the first systematic studies on alternative splicing evolution across vertebrates by RNA-seq.

80. Breschi A et al. Gene-specific patterns of expression variation across organs and species. Genome Biology 17, 151 (2016) [PubMed: 27391956] This study investigated the pattern of gene expression variation across tissues and species individually for each gene along the set of vertebrate orthologues.

81. Bortvin A et al. Incomplete reactivation of Oct4-related genes in mouse embryos cloned from somatic nuclei. Development 130, 1673–1680 (2003). [PubMed: 12620990]

82. Breschi A et al. Gene-specific patterns of expression variation across organs and species. Genome Biology 17, 151 (2016). [PubMed: 27391956]

83. Hardison RC A guide to translation of research results from model organisms to human. Genome Biology 17, 161 (2016). [PubMed: 27459999]

84. Soumillon M et al. Cellular source and mechanisms of high transcriptome complexity in the mammalian testis. Cell reports 3, 2179–2190 (2013). [PubMed: 23791531]

85. Liao B-Y & Zhang J Low rates of expression profile divergence in highly expressed genes and tissue-specific genes during mammalian evolution. Molecular biology and evolution 23, 1119–1128 (2006). [PubMed: 16520335]

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

87. Koonin EV & Wolf YI Constraints and plasticity in genome and molecular-phenome evolution. Nature Reviews Genetics 11, 487–498 (2010).

88. Vakhrusheva OA, Bazykin GA & Kondrashov AS Genome-level analysis of selective constraint without apparent sequence conservation. Genome biology and evolution 5, 532–541 (2013). [PubMed: 23418180]

89. Carvunis A-R et al. Evidence for a common evolutionary rate in metazoan tran scriptional networks. eLife 4, e11615 (2015). [PubMed: 26682651]

90. Weirauch MT & Hughes TR Conserved expression without conserved regu latory sequence: the more things change, the more they stay the same. Trends in Genetics 26, 66–74 (2010). [PubMed: 20083321]

91. Pai AA & Gilad Y Comparative studies of gene regulatory mechanisms. Current opinion in genetics & development 29, 68–74 (2014). [PubMed: 25215415]

92. Odom DT et al. Tissue-specific transcriptional regulation has diverged significantly between human and mouse. Nature genetics 39, 730–732 (2007). [PubMed: 17529977]

93. Johnson R et al. Evolution of the vertebrate gene regulatory network controlled by the transcriptional repressor REST. Molecular biology and evolution 26, 1491–1507 (2009). [PubMed: 19318521]

94. Schmidt D et al. Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. Science 328, 1036–1040 (2010). [PubMed: 20378774]

95. Kunarso G et al. Transposable elements have rewired the core regulatory network of human embryonic stem cells. Nature genetics 42, 631–634 (2010). [PubMed: 20526341]

96. Ballester B et al. Multi-species, multi-transcription factor binding highlights conserved control of tissue-specific biological pathways. Elife 3, e02626 (2014). [PubMed: 25279814]

97. Ernst J & Kellis M ChromHMM: automating chromatin-state discovery and characterization. Nature methods 9, 215–216 (2012). [PubMed: 22373907]

98. Cheng Y et al. Principles of regulatory information conservation between mouse and human. Nature 515, 371–375 (2014). [PubMed: 25409826]

99. Cheng Y et al. Principles of regulatory information conservation between mouse and human. Nature 515, 371–375 (2014) [PubMed: 25409826] Comparative analysis of binding sites of 32 transcription factors through ChIP-sequencing in human and mouse cell lines.

100. Denas O et al. Genome-wide comparative analysis reveals human-mouse regulatory landscape and evolution. BMC genomics 16, 1 (2015). [PubMed: 25553907]

101. Vierstra J et al. Mouse regulatory DNA landscapes reveal global principles of cis-regulatory evolution. Science 346, 1007–1012 (2014). [PubMed: 25411453]

102. Vierstra J et al. Mouse regulatory DNA landscapes reveal global principles of cis-regulatory evolution. Science 346, 1007–1012 (2014) [PubMed: 25411453] Analysis of open chromatin regions in 45 mouse cell and tissue types by DNase-seq and comparison to human.

103. Bourque G et al. Evolution of the mammalian transcription factor binding repertoire via transposable elements. Genome research 18, 1752–1762 (2008). [PubMed: 18682548]

104. Villar D et al. Enhancer evolution across 20 mammalian species. Cell 160, 554–566 (2015) [PubMed: 25635462] This study compares the liver enhancer landscape of 20 mammals inferred from ChIP-seq of H3K27ac and H3K4me3.

105. Stergachis AB et al. Conservation of trans-acting circuitry during mammalian regulatory evolution. Nature 515, 365–370 (2014). [PubMed: 25409825]

106. Young RS Lineage-specific genomics: Frequent birth and death in the human genome. BioEssays (2016).

107. Visel A et al. Ultraconservation identifies a small subset of extremely constrained developmental enhancers. Nature genetics 40, 158–160 (2008). [PubMed: 18176564]

108. Pennacchio LA et al. In vivo enhancer analysis of human conserved non-coding sequences. Nature 444, 499–502 (2006). [PubMed: 17086198]

109. Visel A, Minovitsky S, Dubchak I & Pennacchio LA VISTA Enhancer Browser-a database of tissue-specific human enhancers. Nucleic acids research 35, D88–D92 (2007). [PubMed: 17130149]

110. Welter D et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic acids research 42, D1001–D1006 (2014). [PubMed: 24316577]

111. 1000 Genomes Project Consortium et al. A global reference for human genetic variation. Nature 526, 68–74 (2015) [PubMed: 26432245] Milestone paper on human genetic variation from the 1000 Genomes Project Consortium.

112. Levy S et al. The diploid genome sequence of an individual human. PLoS Biol 5, e254 (2007). [PubMed: 17803354]

113. Lappalainen T et al. Transcriptome and genome sequencing uncovers functional variation in humans. Nature 501, 506–511 (2013). [PubMed: 24037378]

114. GTEx Consortium et al. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. Science 348, 648–660 (2015). [PubMed: 25954001]

115. Yang H et al. Subspecific origin and haplotype diversity in the laboratory mouse. Nature genetics 43, 648–655 (2011). [PubMed: 21623374]

116. Beck JA et al. Genealogies of mouse inbred strains. Nature genetics 24, 23–25 (2000). [PubMed: 10615122]

117. Wade CM et al. The mosaic structure of variation in the laboratory mouse genome. Nature 420, 574–578 (2002) [PubMed: 12466852] This paper explains how the most common laboratory mouse strains were created from wild mice.

118. Yang H, Bell TA, Churchill GA & de Villena FP-M On the subspecific origin of the laboratory mouse. Nature genetics 39, 1100–1107 (2007). [PubMed: 17660819]

119. Keane TM et al. Mouse genomic variation and its effect on phenotypes and gene regulation. Nature 477, 289–294 (2011). [PubMed: 21921910]

120. Wade CM et al. The mosaic structure of variation in the laboratory mouse genome. Nature 420, 574–578 (2002). [PubMed: 12466852]

121. Matsuo N et al. Behavioral profiles of three C57BL/6 substrains. Frontiers in behavioral neuroscience 4, 29 (2010). [PubMed: 20676234]

122. Kiselycznyk C & Holmes A All (C57BL/6) mice are not created equal. Frontiers in neuroscience 5, 10 (2011). [PubMed: 21390289]

123. Keane T et al. Deep genome sequencing and variation analysis of 13 inbred mouse strains defines candidate phenotypic alleles, private variation, and homozygous truncating mutations. bioRxiv, 039131 (2016).

124. Turk R et al. Gene expression variation between mouse inbred strains. BMC genomics 5, 1 (2004). [PubMed: 14704093]

125. Holgersen K et al. 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, jjv050 (2015).

126. Mostafavi S et al. Variation and genetic control of gene expression in primary immunocytes across inbred mouse strains. The Journal of Immunology 193, 4485–4496 (2014). [PubMed: 25267973]

127. Pritchard CC, Hsu L, Delrow J & Nelson PS Project normal: defining normal variance in mouse gene expression. Proceedings of the National Academy of Sciences 98, 13266–13271 (2001).

128. Bogue MA, Churchill GA & Chesler EJ Collaborative Cross and Diversity Outbred data resources in the Mouse Phenome Database. Mammalian Genome 26, 511–520 (2015) [PubMed: 26286858] This paper illustrates the current status of the Mouse Phenome Database as an established resource for studying mouse genetic variation.

129. Bogue MA, Churchill GA & Chesler EJ Collaborative Cross and Diversity Outbred data resources in the Mouse Phenome Database. Mammalian Genome 26, 511–520 (2015). [PubMed: 26286858]

130. Chick JM et al. Defining the consequences of genetic variation on a proteome-wide scale. Nature 534, 500–505 (2016) [PubMed: 27309819] Paper describing the relationship between eQTLs and pQTLs in outbred mice.

131. Deng Q, Ramsköld D, Reinius B & Sandberg R Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. Science 343, 193–196 (2014). [PubMed: 24408435]

132. Vickaryous MK & Hall BK Human cell type diversity, evolution, development, and classification with special reference to cells derived from the neural crest. Biological reviews 81, 425–455 (2006). [PubMed: 16790079]

133. Abdulreda M, Caicedo A & Berggren P A natural body window to study human pancreatic islet cell function and survival. CellR4 1, 111–122 (2013).

134. Mabbott NA, Baillie JK, Brown H, Freeman TC & Hume DA An expression atlas of human primary cells: inference of gene function from coexpression networks. BMC genomics 14, 632 (2013). [PubMed: 24053356]

135. Hume DA, Summers KM, Raza S, Baillie JK & Freeman TC Functional clustering and lineage markers: insights into cellular differentiation and gene function from large-scale microarray studies of purified primary cell populations. Genomics 95, 328–338 (2010). [PubMed: 20211243]

136. Lee Y. s., Krishnan A, Zhu Q & Troyanskaya OG Ontology-aware classification of tissue and cell-type signals in gene expression profiles across platforms and technologies. Bioinformatics 29, 3036–3044 (2013). [PubMed: 24037214]

137. Li J et al. Single-cell transcriptomes reveal characteristic features of human pancreatic islet cell types. EMBO reports 17, 178–187 (2016). [PubMed: 26691212]

138. Saliba A-E, Westermann AJ, Gorski SA & Vogel J Single-cell RNA-seq: advances and future challenges. Nucleic acids research 42, 8845–8860 (2014). [PubMed: 25053837]

139. Macosko EZ et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell 161, 1202–1214 (2015). [PubMed: 26000488]

140. Kolodziejczyk AA, Kim JK, Svensson V, Marioni JC & Teichmann SA The technology and biology of single-cell RNA sequencing. Molecular cell 58, 610–620 (2015). [PubMed: 26000846]

141. Trapnell C Defining cell types and states with single-cell genomics. Genome research 25, 1491–1498 (2015). [PubMed: 26430159]

142. Zeisel A et al. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. Science 347, 1138–1142 (2015). [PubMed: 25700174]

143. Treutlein B et al. Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. Nature 509, 371–375 (2014). [PubMed: 24739965]

144. Grün D et al. Single-cell messenger RNA sequencing reveals rare intestinal cell types. Nature 525, 251–255 (2015). [PubMed: 26287467]

145. Darmanis S et al. A survey of human brain transcriptome diversity at the single cell level. Proceedings of the National Academy of Sciences 112, 7285–7290 (2015).

146. Jaitin DA et al. Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. Science 343, 776–779 (2014). [PubMed: 24531970]

147. Paul F et al. Transcriptional heterogeneity and lineage commitment in myeloid progenitors. Cell 163, 1663–1677 (2015). [PubMed: 26627738]

148. Scialdone A et al. Resolving early mesoderm diversification through single-cell expression profiling. Nature 535, 289–293 (2016). [PubMed: 27383781]

149. Ohnishi Y et al. Cell-to-cell expression variability followed by signal reinforcement progressively segregates early mouse lineages. Nature cell biology 16, 27–37 (2014). [PubMed: 24292013]

150. Xue Z et al. Genetic programs in human and mouse early embryos revealed by single-cell RNA [thinsp] sequencing. Nature 500, 593–597 (2013) [PubMed: 23892778] Dynamic transcriptional study of human and mouse early embryo development at single-cell resolution.

151. Xue Z et al. Genetic programs in human and mouse early embryos revealed by single-cell RNA [thinsp] sequencing. Nature 500, 593–597 (2013). [PubMed: 23892778]

152. Blekhman R, Marioni JC, Zumbo P, Stephens M & Gilad Y Sex-specific and lineage-specific alternative splicing in primates. Genome research 20, 180–189 (2010). [PubMed: 20009012]

153. Melé M et al. The human transcriptome across tissues and individuals. Science 348, 660–665 (2015). [PubMed: 25954002]

154. Zhang R, Lahens NF, Ballance HI, Hughes ME & Hogenesch JB A circadian gene expression atlas in mammals: implications for biology and medicine. Proceedings of the National Academy of Sciences 111, 16219–16224 (2014).

155. Romero IG, Pai AA, Tung J & Gilad Y RNA-seq: impact of RNA degradation on transcript quantification. BMC biology 12, 42 (2014). [PubMed: 24885439]

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

157. Chen KH, Boettiger AN, Moffitt JR, Wang S & Zhuang X Spatially resolved, highly multiplexed RNA profiling in single cells. Science 348, aaa6090 (2015). [PubMed: 25858977]

158. Satija R, Farrell JA, Gennert D, Schier AF & Regev A Spatial reconstruction of single-cell gene expression data. Nature biotechnology 33, 495–502 (2015).

159. Gharib SA et al. Of mice and men: comparative proteomics of bronchoalveolar fluid. European Respiratory Journal 35, 1388–1395 (2010). [PubMed: 20032019]

160. Maher B ENCODE: The human encyclopaedia. Nature 489, 46 (2012). [PubMed: 22962707]

161. Von Meyenn F et al. Comparative Principles of DNA Methylation Reprogramming during Human and Mouse In Vitro Primordial Germ Cell Specification. Developmental Cell 39, 104–115 (2016). [PubMed: 27728778]

162. Nakamura T et al. A developmental coordinate of pluripotency among mice, monkeys and humans. Nature 537, 57–62 (2016). [PubMed: 27556940]

163. Rosenthal N & Brown S The mouse ascending: perspectives for human-disease models. Nature cell biology 9, 993–999 (2007). [PubMed: 17762889]

164. Onos KD, Rizzo SJS, Howell GR & Sasner M Toward more predictive genetic mouse models of Alzheimer's disease. Brain research bulletin 122, 1–11 (2016). [PubMed: 26708939]

165. Blesa J, Phani S, Jackson-Lewis V & Przedborski S Classic and new animal models of Parkinson's disease. BioMed Research International 2012 (2012).

166. Ribeiro FM, Camargos E. R. d. S., Souza L. C. d. & Teixeira AL Animal models of neurodegenerative diseases. Revista Brasileira de Psiquiatria 35, S82–S91 (2013). [PubMed: 24271230]

167. Antonarakis SE Down syndrome and the complexity of genome dosage imbalance. Nature Reviews Genetics (2016).

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

169. Steimer T Animal models of anxiety disorders in rats and mice: some conceptual issues. Dialogues Clin Neurosci 13, 495–506 (2011). [PubMed: 22275854]

170. Schweinfurth N & Lang UE Behavioral Testing of Mice Concerning Anxiety and Depression. Zeitschrift für Psychologie (2015).

171. Lynch WJ, Nicholson KL, Dance ME, Morgan RW & Foley PL Animal models of substance abuse and addiction: implications for science, animal welfare, and society. Comparative medicine 60, 177–188 (2010). [PubMed: 20579432]

172. Ellacott KL, Morton GJ, Woods SC, Tso P & Schwartz MW Assessment of feeding behavior in laboratory mice. Cell metabolism 12, 10–17 (2010). [PubMed: 20620991]

173. Mestas J & Hughes CC Of mice and not men: differences between mouse and human immunology. The Journal of Immunology 172, 2731–2738 (2004). [PubMed: 14978070]

174. Karpel ME, Boutwell CL & Allen TM BLT humanized mice as a small animal model of HIV infection. Current opinion in virology 13, 75–80 (2015). [PubMed: 26083316]

175. Chayama K et al. Animal model for study of human hepatitis viruses. Journal of gastroenterology and hepatology 26, 13–18 (2011). [PubMed: 21175788]

176. Silverman JL, Yang M, Lord C & Crawley JN Behavioural phenotyping assays for mouse models of autism. Nature Reviews Neuroscience 11, 490–502 (2010). [PubMed: 20559336]

177. Vanhooren V & Libert C The mouse as a model organism in aging research: usefulness, pitfalls and possibilities. Ageing research reviews 12, 8–21 (2013). [PubMed: 22543101]

178. Bult CJ et al. Mouse Tumor Biology (MTB): a database of mouse models for human cancer. Nucleic acids research 43, D818–D824 (2015). [PubMed: 25332399]

179. Justice MJ & Dhillon P Using the mouse to model human disease: increasing validity and reproducibility 2016.

180. Rangarajan A & Weinberg RA Comparative biology of mouse versus human cells: modelling human cancer in mice. Nature Reviews Cancer 3, 952–959 (2003). [PubMed: 14737125]

181. Egan ME How useful are cystic fibrosis mouse models? Drug Discovery Today: Disease Models 6, 35–41 (2009).

182. Fisher JT, Zhang Y & Engelhardt JF Comparative biology of cystic fibrosis animal models. Cystic Fibrosis: Diagnosis and Protocols, Volume II: Methods and Resources to Understand Cystic Fibrosis, 311–334 (2011).

183. McGreevy JW, Hakim CH, McIntosh MA & Duan D Animal models of Duchenne muscular dystrophy: from basic mechanisms to gene therapy. Disease Models and Mechanisms 8, 195–213 (2015). [PubMed: 25740330]

184. Modrek B & Lee CJ Alternative splicing in the human, mouse and rat genomes is associated with an increased frequency of exon creation and/or loss. Nature genetics 34, 177–180 (2003). [PubMed: 12730695]

185. Abril JF, Castelo R & Guigó R Comparison of splice sites in mammals and chicken. Genome research 15, 111–119 (2005). [PubMed: 15590946]

186. Sorek R & Ast G Intronic sequences flanking alternatively spliced exons are conserved between human and mouse. Genome Research 13, 1631–1637 (2003). [PubMed: 12840041]

187. Merkin J, Russell C, Chen P & Burge CB Evolutionary dynamics of gene and isoform regulation in Mammalian tissues. Science 338, 1593–1599 (2012). [PubMed: 23258891]

188. Zambelli F, Pavesi G, Gissi C, Horner DS & Pesole G Assessment of orthologous splicing isoforms in human and mouse orthologous genes. BMC genomics 11, 1 (2010). [PubMed: 20044946]

189. Tilgner H et al. Comprehensive transcriptome analysis using synthetic long-read sequencing reveals molecular co-association of distant splicing events. Nature biotechnology 33, 736–742 (2015).

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

191. Hay M, Thomas DW, Craighead JL, Economides C & Rosenthal J Clinical development success rates for investigational drugs. Nature biotechnology 32, 40–51 (2014).

192. Ulitsky I Evolution to the rescue: using comparative genomics to understand long non-coding RNAs. Nature Reviews Genetics 17, 601–614 (2016).

193. Liao B-Y & Zhang J Evolutionary conservation of expression profiles between human and mouse orthologous genes. Molecular biology and evolution 23, 530–540 (2006). [PubMed: 16280543]

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

195. Villar D et al. Enhancer evolution across 20 mammalian species. Cell 160, 554–566 (2015). [PubMed: 25635462]

196. 1000 Genomes Project Consortium et al. A global reference for human genetic variation. Nature 526, 68–74 (2015). [PubMed: 26432245]

197. Chick JM et al. Defining the consequences of genetic variation on a proteome-wide scale. Nature 534, 500–505 (2016). [PubMed: 27309819]

Box1 - Mice as models for diseases

Since the early days of mouse research, mice have been engineered to generate models for a variety of human diseases and conditions[163]. The Jackson Laboratory has generated more than 5,000 mouse models with different genotypes for almost 1,500 human diseases[5]. Their range of application is very broad, including neurological and muscular disorders, genetic illnesses, behavioural and cognitive abilities, response to viruses and cancer research.

Genetic mouse models of neurodegenerative disorders, such as Alzheimer's disease (Online Mendelian Inheritance in Man (OMIM https://www.omim.org/)) 104300)[164] and Parkinson's disease (OMIM 168600)[165], which recapitulate the essential features of each disease, have significantly advanced our understanding of the molecular basis of disease progression. However, their translational impact remains limited, as neurodegenerative human diseases are heterogeneous in both pathological and clinical (or behavioral) domains and the non-hereditary causes (affecting the majority of the cases) are unknown[166].

As another example, several mouse models for Down syndrome (also known as trisomy 21, OMIM 190685) have been generated based on the homology of the human chromosome 21 and the mouse chromosomes 10, 16 and 17 [167]. These models exhibit many of the behavioural, learning and physiological defects associated with the syndrome in humans, and as such have proved useful to test therapies that rescue these alterations[168].

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 applications of behavioural tests include studies of anxiety[169,170], substance abuse and addiction[171], and diet[172].

Despite acknowledged discrepancies between the human and murine immune systems[173], mouse models exist to also investigate viral infections and limit the ethical and practical costs of primate research. For instance, humanized mice derived from the combination of transplantation of human fetal pluripotent hematopoietic stem cells with surgical engraftment of human fetal thymic tissue (BLT mice) have been used to study many aspects of HIV infection, including prevention, transmission and therapies[174]. Similarly, human hepatocytes are transplanted into immunodeficient mice to develop humanized chimeric mice, which enable the study of viral replication and cellular changes caused by the human hepatitis viruses[175].

Finally, mice have also been widely used for the research of very complex multifactorial conditions, such as autism[176] and ageing[177], where it is crucial to be able to account for one individual factor at a time. Among complex diseases, cancer research is certainly prompting the development of several mouse models to study the relationship between mutations and tumour biology[178].

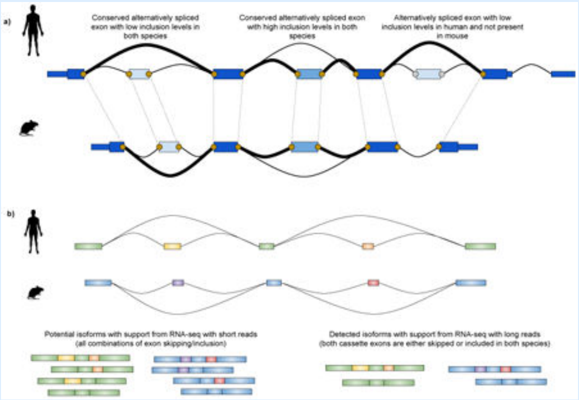
However, current limitations of mouse models are well known [179]. The use of mice to study the intricacies of human cancer pathogenesis, for example, is limited by many species differences, including cell duplication time, lifespan and cancer susceptibility[180], amongst others. Other examples include mouse models of cystic fibrosis (CF; OMIM 219700), a hereditary lung disease caused by a mutation in the gene encoding the membrane protein CFTR. Although these have proven useful to discover ways to correct this defect[181], CF mouse models have a limited ability to recapitulate spontaneous lung disease[182]. Similarly, mouse models for the progressive muscle-wasting disorder Duchenne muscular dystrophy (DMD; OMIM 310200), that is, *mdx* mice, have been engineered to study potential gene therapies, but a caveat is that they show only minimal clinical symptoms[183]. Alternative animal models are being investigated to potentiate translational research, and larger mammalian species, such as pigs, ferrets and dogs, are proving beneficial to scale up initial results obtained in mouse models[182,183].

Box2 - Splicing

Splicing is the mechanism through which exons and introns of genes are processed into mature coding and non-coding transcripts. Different combinations of exonic and intronic sequences can be arranged through alternative splicing to expand the range of processed isoforms from a relatively limited pool of genes.

Exon structure and splicing are very similar between humans and mice, in terms of number and order of exons per gene, exon length, precise boundaries and sequence[184,185]. The exact number of orthologous exons is heavily dependent on the genome assemblies, the annotation status and on the set of analysed genes: the mouse ENCODE consortium has annotated over 150,000 orthologous internal exons[17], a noticeable increase compared to the 2,000 exons[186] identified in the earliest reports right after publication of the first complete mouse genome draft. Although alternatively spliced exons with low proportion of inclusion tend to be more species-specific[184], exon inclusion levels are overall highly correlated between the two species even across very distant sample types[23] (see the figure, part a). Indeed, alternative splicing was shown to be less evolutionarily conserved than gene expression in comparative studies including multiple species and organs[78,187].

However, comparative analyses of exon inclusion are usually limited to a few hundred conserved exons[78,187] and are tied to local splicing events, not considering the whole isoform structure. Determining orthology at the isoform level, for complete gene structures of exons and introns, is particularly challenging, due to the presence of non-coding exons, which have less sequence constraints than coding sequences, and to the redundancy of exonic elements between multiple isoforms of the same gene[188]. Novel transcriptomic sequencing strategies, for example, synthetic long-read sequencing[189] and single-molecule long-read sequencing[190], enable detection of full-length transcripts and preserve the relationship between distant exons (see the figure, part b). 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 and of its regulation across species.


a)
Conserved alternatively spliced exon with low inclusion levels in both species
Conserved alternatively spliced exon with high inclusion levels in both species
Alternatively spliced exon with low inclusion levels in human and not present in mouse
b)
Potential isoforms with support from RNA-seq with short reads (all combinations of exon skipping/inclusion)
Detected isoforms with support from RNA-seq with long reads (both cassette exons are either skipped or included in both species)

Databases

Ensembl Compara http://www.ensembl.org/info/genome/compara/index.html

Rfam http://rfam.xfam.org

VISTA Enhancer Browser https://enhancer.lbl.gov

OMIM https://www.omim.org

GTEx project http://www.gtexportal.org

Gencode http://www.gencodegenes.org

Key points

- Mouse is the most widely used model organism to study human disease, but often mouse biology cannot be extrapolated to human. A deep comparison of mouse and human physiology at the molecular level is essential for understanding under which circumstances mouse can be a good model of human biology and for creating better mouse models.
- Advances in next generation sequencing technologies fostered genome-wide annotation of functional DNA elements enabling extensive comparison of human and mouse genomes.
- At the transcriptional level human and mouse gene expression profiles are overall conserved, although the degree of conservation varies depending on the tissues and the genes that are compared. Therefore the question whether the human and mouse transcriptomes cluster preferentially by tissue/organ or by species does not have an answer overall, and it will depend specifically on the genes being considered.
- Conservation of expression is not a direct consequence of conservation in regulatory sequences, including promoters and enhancers. Although gene regulatory networks are overall preserved between human and mouse, transcription binding sites are often not conserved.
- Interindividual genetic variation can affect human gene expression, but such variation cannot be modelled in inbred strains of laboratory mice because their genetic variation is small compared to the human population. An expansion of the current studies on the relationship between genetic variation and gene expression in outbred mice might provide helpful insights to understand the same relationship in humans.
- New emerging technologies, such single-cell genomics and spatial transcriptomics, and time-series experiments will improve the annotation of human and mouse genomes, refine the current definitions of homologous cell types, as well as of homologous (molecular) phenotypes and ultimately help scientists to identify which mouse models are the most appropriate to address a given biological question.

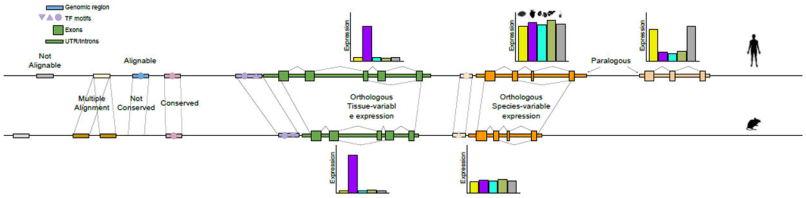


Figure 1. Homology of human and mouse genes and genomic elements. Orthologous genes between human and mouse can be identified based on sequence homology of coding exons.

Orthologous genes tend to have conserved exonic structure and exon lengths, but introns are generally shorter in mouse. There is some degree of conservation of alternative splicing patterns (Box 2), but species specific splicing events exist (green gene). Orthologous genes may have conserved expression profiles between the two species (green) or diverged expression (orange). The bar chart represents expression levels of the genes in different organs. Genes with homologous sequence within the same species are called paralogous. Paralogous genes may originate from gene duplication events and their exonic structure, sequence and expression may diverge with evolutionary time. Promoter sequences (upstream from genes) are less conserved than gene body sequences. Regulatory motifs may differ although regulatory networks may be conserved. Orthologous genomic regions (and elements) can be identified through whole genome alignments (pink). However, some elements cannot be aligned to the other species (different shades of grey), or can map in multiple locations (brown). Finally, some genomic regions can be aligned, but their function may not be conserved (blue).

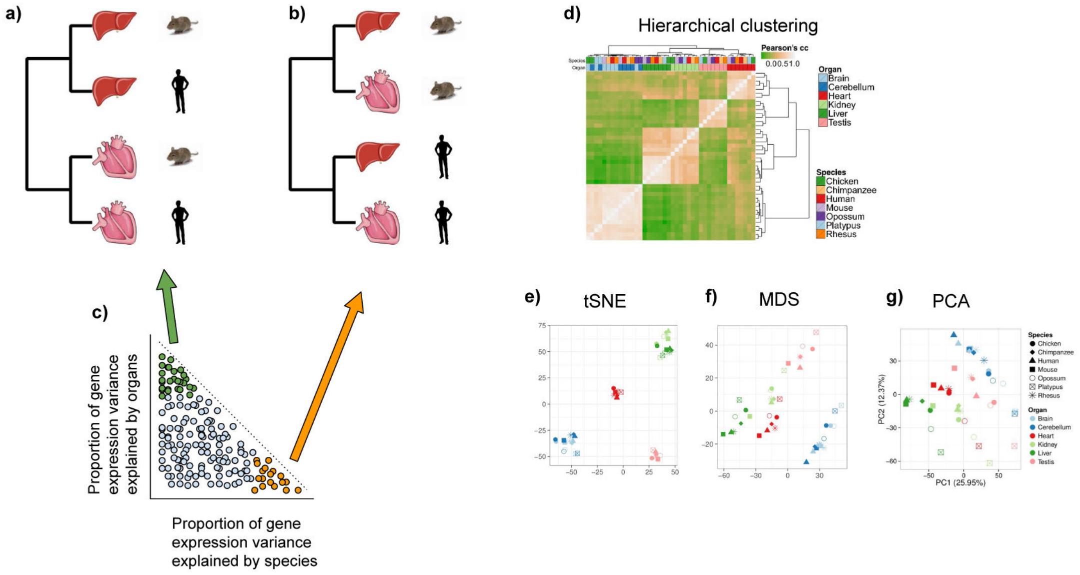


Figure 2. Simplified clustering of human and mouse tissue samples and variance decomposition of gene expression.

Samples can be clustered based on their transcriptional profiles. If a human organ (for example, liver or heart) has a more similar gene expression profile to the homologous mouse organ than to another human organ, the clustering is organ-dominated (**a**). Vice versa, if human organs have more similar gene expression profiles between each other than compared to their homologous mouse organs, the clustering is species-dominated (**b**). The variation of expression for each gene can be decomposed into the most contributing factors, in this case species and organs (**c**). Genes are distributed in a continuous way along these proportions of variation. Nonetheless, genes at the extremes of this distribution can be identified as genes with proportionally higher variation across species and lower across organs (orange) and genes with proportionally higher variation across species and lower across organs (green). If only the expression of one or the other set of genes is used for clustering, genes with proportionally higher variation across species or organs lead to a more species-dominated clustering, or organ-dominated clustering, respectively. **d**| Hierarchical clustering based on real gene expression data from different organs across mammals and chicken, performed with the entire set of orthologous genes across species, reveal organ-dominated clustering[82]. Distances between samples can be visually represented also on a 2-D space through several dimensionality reduction techniques, such tSNE (**e**, same input as **d**, perplexity=4, iterations=1000), MDS (**f**, same input as **d**, euclidean distance) and PCA[82] (**g**).

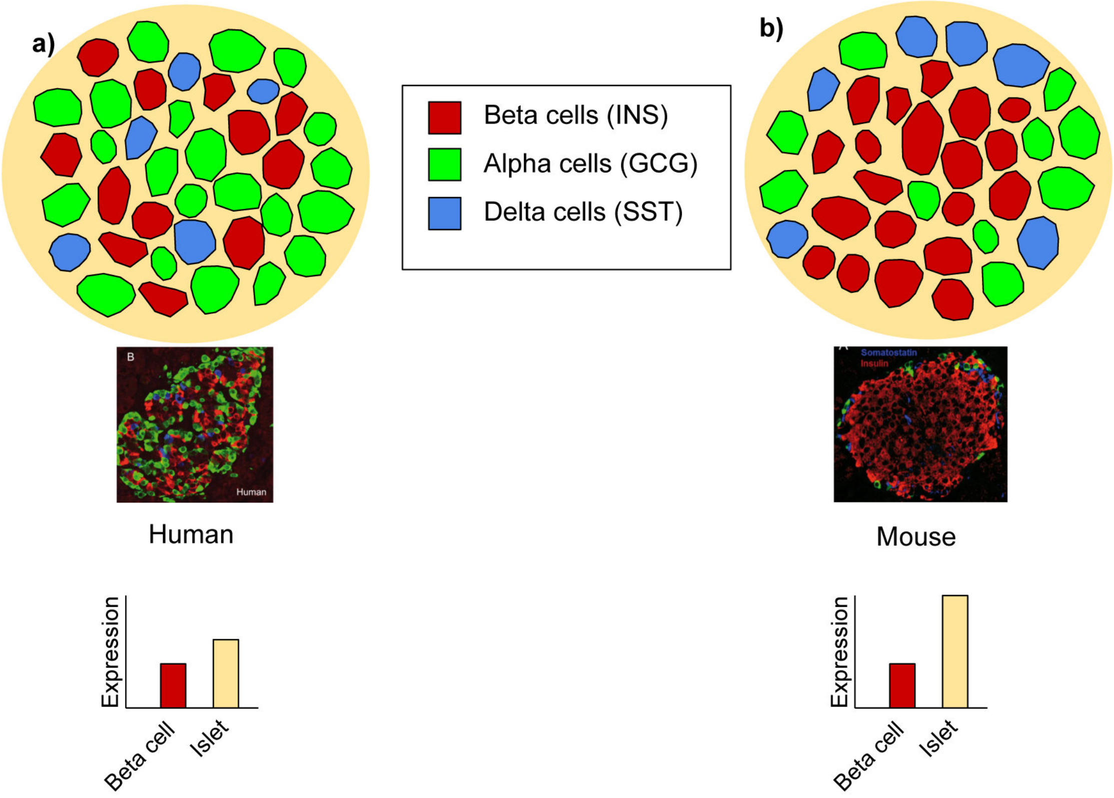


Figure 3. Cellular composition of human (a) and mouse (b) pancreatic islets.

Humans and mice have a different composition of pancreatic islets of Langerhans. Insulin-producing beta cells make up to 80% of mouse islets, whereas they constitute only up to 50% of the human islets. By contrast, glucagon-producing alpha cells compose up to 40% of the human islets. Fluorescent-stained images are taken from[133]. The expression of a given gene may appear different when the whole anatomical structure is profiled, whereas what actually changes is the relative abundance of cells of different types expressing that gene, and not the expression of a gene in a particular cell type.

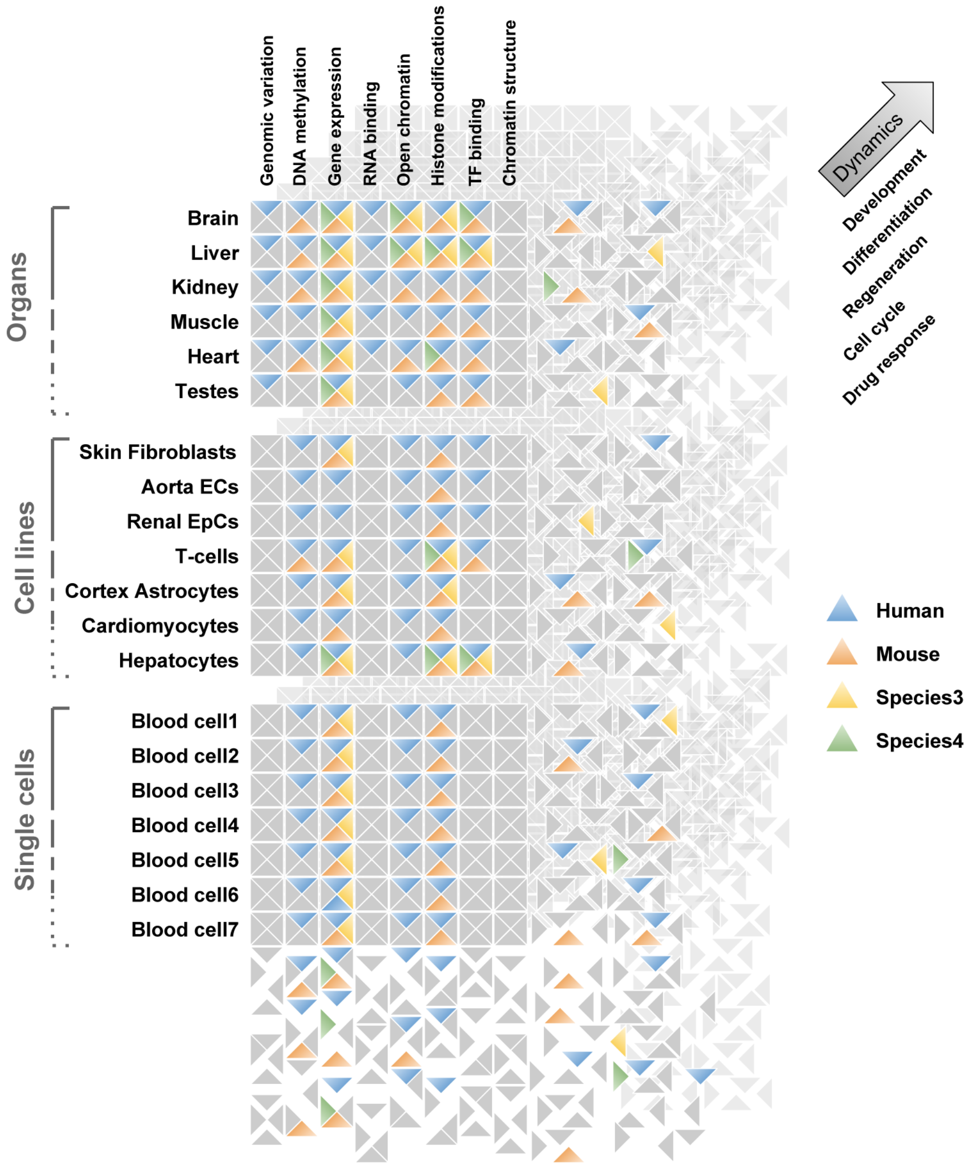


Figure 4. Multidimensional complexity of omics-layers integration across species.

Four-dimensional matrix illustrating possible experimental combinations of genomics features profiled in different sample types, across species and in dynamic conditions. Colored triangles represent combinations of factors for which experiments are already available. This information is just figurative and might not reflect the current status of published experiments across all public or private databases. This figure is adapted from[160].

Table 1.

Summary statistics of human and mouse genomes and gene sets.

Annotation counts are retrieved from the Gencode website (http://www.gencodegenes.org/, v25 for human and vM11 for mouse). The number of microRNAs is obtained from miRBase v21[50]. The number of tRNAs is obtained from GtRNAdb[53]. The number of protein-coding orthologues is taken from Ensembl Compara[24] (v86), while the numbers of orthologous long non-coding RNAs were obtained from different sources[23,36–38].

	Human (GRCh38)	**Mouse (GRCm38)**
Genome size (nt)	3,088,269,832	2,725,521,370
Unplaced scaffolds (nt)	11,464,317	5,334,105
Number of chromosomes	22 + X +Y	19 + X + Y
Chain alignments (nt)	2,735,135,097	2,465,275,732
Number of genes	58,037	48,709
Number of transcripts	198,093	118,925
Protein-coding		
- genes	19,950	22,018
- 1 to 1 orthologs	15,893	
- transcripts	80,087	52,382
Long non-coding RNAs		
- genes	15,767	9,989
- orthologs	2,720 [36], 1,587 [38], 1,100 [37], 851 [23]	
- transcripts	27,692	13,904
Pseudogenes	14,650	10,096
Small RNAs	7,258	6,110
- miRNAs [42]	2,588	1,915
- snRNAs	1,900	1,383
- snoRNAs	944	1,508
- tRNAs [46]	631	471