Evolutionary history of regulatory variation in human populations

Tuuli Lappalainen and Emmanouil T. Dermitzakis

Department of Genetic Medicine and Development, University of Geneva Medical School, 1411-4 Geneva, Switzerland

Received August 5, 2010; Revised and Accepted September 15, 2010

Genetic variation in the regulation of gene expression is likely to be a major contributor to phenotypic variation in humans, and it also constitutes an important target of recent natural selection in human populations and plays a major role in morphological evolution. The increasing amount of data of genome and transcriptome variation is now leading to a better annotation of regulatory elements and a growing understanding of how the evolution of gene regulation has shaped human diversity. In this review, we discuss the evolutionary history of the variation in the expression of protein-coding genes in humans. We outline the current methodology for mapping regulatory variants and their distribution in human populations. General mechanisms of regulatory evolution are discussed with a special emphasis on different selective processes targeting gene regulation in humans.

INTRODUCTION

Analysis of regulatory variation has been motivated by a quest for understanding the sources of phenotypic variation in humans, including variation in susceptibility to disease. Genetic differences in the regulation of gene expression may also underlie some of evolutionary adaptive phenotypic differences between human populations, and from a longer evolutionary perspective, the evolution of human-specific traits that distinguish us from other primates has been a major focus of research. The importance of gene regulation in morphological evolution has been acknowledged and debated for decades (1–7). Recently, the field of evolutionary genetics has witnessed an accumulation of evidence of regulatory changes underlying phenotypic differences within and between species—first through case examples, but increasingly through genome-wide analysis of genomes and transcriptomes. Now, we are increasing our understanding of how the information in the genetic code is transferred to the transcriptome, to proteins, and thereon to phenotypes at the cellular, systemic and organismal levels. We are learning how different types of genetic variations alter these pathways, and how variants of different functional categories are being shuffled by the evolutionary process. In this review article, we will discuss population genetics of regulatory variation affecting the expression of protein-coding genes in humans, and the evolutionary history of this variation. The non-coding part of the transcriptome and its evolution has been discussed elsewhere (8–12).

ANALYSING GENOMES AND TRANSCRIPTOMES

The analysis of regulatory variation requires data of both genetic and transcriptome data. Levels of gene expression have been analyzed now for almost 10 years in a genome-wide manner by expression arrays, and genome-wide analysis of human genetic variation was made possible about 5 years ago through the development of array-based genotyping of hundreds of thousands of single nucleotide polymorphisms (SNPs).

Despite the wide variety of approaches relying on these techniques, the recent advance in sequencing technologies has opened a range of new exciting possibilities for the analysis of the genome and its function. Sequencing of mRNA offers a much more accurate analysis of splicing patterns, expression levels and allele-specific expression than array-based technologies (13–18). Transcriptome sequencing has also revolutionized the comparison of gene expression patterns between species by eliminating the need to rely on pre-designed probes that have been available only for a limited set of species with established genome annotation.
The analysis of genetic variation is also shifting from genotyping of pre-selected SNPs to genomic re-sequencing, which offers not only a denser coverage of variants and accurate genotyping of structural variation but also a more even coverage of the frequency spectrum (17) (www.1000genomes.org) without ascertainment bias towards well-studied populations in array SNP selection, which has been a concern in population genetic studies (19). The new technologies are also being used for de novo sequencing as well for population-based re-sequencing of non-human species to discover genetic variation in other organisms (20,21) (www.sanger.ac.uk/modelorgs/mousegenomes).

Furthermore, an increasing understanding of the mechanisms of genome function and its evolution is being gained through sequencing applications for assaying, for example, transcription factor binding, methylation patterns and chromatin structure. Integrating these data into knowledge of genetic and transcriptome analysis will shed light on regulatory networks and the annotation of regulatory elements (12,17,22–26).

**MAPPING REGULATORY VARIATION**

Several approaches have been developed to find genetic variants that affect gene expression. The most common method has been testing for association between the genotype classes of a genetic variant and gene expression levels to map expression quantitative trait loci (eQTLs), mostly in cis close to the target gene, but also in trans (13,14,27–35). eQTL analysis captures only common regulatory variation, because statistical power to detect association to expression levels decreases sharply with minor allele frequency. Another approach has been to study allele-specific expression, where allelic imbalance in the mRNA production between coding heterozygous polymorphisms is used as a signal of cis-regulatory variation (36–40). This method has its own limitations and sources of error, but has better power to find rare regulatory variants in cis, accessible especially in RNA-sequencing data (13,14,18). Altogether, these studies have shed light on the general patterns of regulatory variation in human populations and provided the tools for further studies of the role of regulatory variation, for example, in human disease and evolution.

These approaches do not, however, give direct information of the causal variants that alter gene expression, because typically a large number of variants show significant association to expression values of a gene because of the linkage disequilibrium (LD) in the human genome, i.e. the strong correlation of genetic variants located up to tens or even hundreds of kilobases from each other. This makes defining the causal variant difficult even when full information of all the genetic variation is available through genomic resequencing (Montgomery et al., in preparation). Figure 1 illustrates an example of this: SNPs from array (Fig. 1A) as well as resequencing (Fig. 1B) data sets have several markers that show significant association due to being in LD with each other. However, the peak region is easier to distinguish in the denser resequencing data, especially from the African population where the extent of LD is lower than in the Europeans due to differences in population history. Additional information of the location of the variants relative to functional elements such as the transcription start site, transcription factor binding sites and splice sites, as well as evolutionary conservation and differences between populations can also be used to model the most likely causal variant (41).

Many types of mutation can affect gene regulation and lead to variation in expression patterns within and between species. Even though SNPs have been studied most, structural changes of different sizes are likely to contribute as well: in human populations, 20% of eQTLs appear to be caused by large copy number variations typically of >100 kb in length (29). However, this proportion is likely to rise when data from small insertions and deletions, obtained from resequencing data, are added to the analysis—although at the same time, this will complicate the inference of causality (Montgomery et al., in preparation). The importance of structural changes has been observed in interspecies comparisons, too: there is a positive correlation between the density of structural mutations and the extent of expression level changes between species (42,43). The mechanism of how structural variation affects gene expression appears to be independent of a simple dosage effect both within and between species, suggesting that structural variation in proximal regulatory elements is often causing the change in expression patterns (29,42,44).

**PATTERNS OF REGULATORY VARIATION IN HUMAN POPULATIONS**

Human populations show significant differences in gene expression levels: in analyses of cell lines from different populations, 17–29% of genes have shown expression differences between European, African and Asian populations (27,30,45), and about 15% of the total expression variation between an European American and an African population could be attributed to differences between the populations (45,46). A large part of this variation is genetic: overall heritability of gene expression in humans has been calculated to be 0.43 by using an admixed population (46), and in different studies expression levels for 13–31% of human genes have been estimated to have heritability over 0.2 (30–33,47,48). Also splicing of exons is known to have genetically determined variation between individuals and populations, although this has only recently become apparent from large-scale studies (13,14,49–52). However, environmental and technical contribution to the observed gene expression variation is not to be overlooked (53,54). In particular, if gene expression is measured not from cell lines but from individuals exposed to different environments, even genetically similar populations have been shown to have large expression differences (55,56), and thus the heritability values obtained from cell lines are likely to be overestimates of the true values in human populations.

Over 1000 eQTLs have been mapped across the human genome in different populations, mostly cis-eQTLs, but also some trans-eQTLs and a much smaller number of splicing QTLs (13,14,27–35). Studies of allele-specific expression capture some of the same loci but also others, with up to
30% of genes having signs of common regulatory variation in cis (36–38,40). Analyses of different populations generally show a significant overlap with about one-third of eQTLs being shared between populations from Africa, Europe and Asia (30). In general, population differences in gene expression and eQTL sharing appear to follow genetic differentiation across the genome as well as in a genome-wide scale across a large set of populations from different continents (Stranger et al., in preparation, 27,46). Shared eQTLs appear to have nearly always the same direction of effect in different populations as well as similar fold changes, suggesting that while the extent to which individual eQTLs affect gene expression depend on allele frequencies that vary between populations, the underlying regulatory mechanisms are shared (Stranger et al., in preparation, 27,30).

Mapping of regulatory variation in cis has been much easier than in trans due to higher effect sizes and easier control of multiple testing problems (30,31,33,47,57), and thus patterns of cis-variation are much better understood. However, regulatory variation in cis probably represents a minority of the total heritable proportion of variation in gene expression levels: estimates derived from an analysis of an admixed human population suggests that it accounts for only 12% of the total variation (46), which is consistent with a higher contribution of trans-variation observed in Drosophila (58), although defining cis- and trans-variation is often difficult and varies between studies (57,59,60). Most human studies have relied on transformed lymphoblastoid cell lines, but it is now known that regulatory differences between different cell types are much bigger than between populations, with common tissue-specific effects, where an eQTL affects gene expression only in a part of the tissues where the gene is expressed (32,49,61,62).

MECHANISMS OF REGULATORY EVOLUTION

The relative contribution of regulatory and coding changes in evolutionary change is one of the big open questions in evolutionary genetics. In general, compared with changes in the amino acid sequences, the functional consequences of regulatory mutations, especially in cis, are thought to provide more flexible material for natural selection with fewer functional trade-offs: while a non-synonymous mutation generally leads to a qualitative change in protein structure wherever the gene is expressed, regulatory changes are quantitative, and may often be activated only in particular environmental conditions, developmental stages or tissues. Furthermore, differences in patterns of selection may arise from cis-regulatory variants being usually codominant, whereas coding and trans-regulatory variants are more often recessive (31,60,63). However, despite many comparisons of the evolution of regulatory and coding regions, no consensus of their relative importance has been reached (2–4,6,7,53,64–67). Unbiased analysis is challenging due to the difficulty of predicting functional consequences of non-coding mutations, and an additional complication arises from LD between coding and
flanking regions, which may lead to correlated patterns even in the absence of similar selective effects. Interaction between cis-regulatory and coding non-synonymous variants is also not to be ignored as the expression of the two allelic variants is often imbalanced (36,68), but the phenotypic and evolutionary consequences of such interactions are still poorly known. Furthermore, several studies suggest that natural selection has targeted cis- and trans-regulatory elements differently: cis-elements appear to be more frequent targets of positive selection and contribute more to interspecies differences in gene expression compared with more-constrained trans-variation (58,63,65,69)—possibly because cis-elements have more tissue-specific effects and are thus less likely to have pleiotropic effects across tissues than trans-variants.

REGULATORY VARIATION AS A TARGET OF NATURAL SELECTION

Although gene expression has been suggested to be under less strict purifying selection than coding variation (1), and even selective neutrality has been proposed (66), gene expression levels both between and within populations often show evolutionary constraint (15,53,67,70). Regulatory regions of the genome are often strongly conserved between species, and have a lower degree of variation also within species (12,71–74). In human populations, SNPs in 5′ and 3′ regions show an enrichment of low differentiation between populations suggesting purifying selection (75). Also, variation in transcription factor binding between humans and between human and chimpanzee is much more common in sites far away from transcription start sites of genes, suggesting that natural selection eliminates variation that would change expression patterns of genes (22).

Because of the lack of power to detect rare eQTLs, the genes regulated by known eQTLs generally tolerate common regulatory variation in humans. Thus, their regulation is likely to be less constrained, and it has been shown that the carriers of the ancestral haplotype of common human eQTLs do not show expression levels closer to the chimpanzees, suggesting low constraint with several rounds of regulatory variants fixing in both species (Montgomery and Dermitzakis, unpublished data). This is supported by cis-variation being rarer in gene-dense regions that are likely to be more constrained (69), and by the significant overlap between human and mouse genes that show allele-specific expression, which suggests low constraint in these genes in mammals (76). Thus, eQTL data are inherently biased towards common variants in genes whose expression is not heavily constrained, but the novel possibilities for mapping rare regulatory variation from RNA sequencing data (13) will now yield a more complete catalogue of different regulatory variants and enable more comprehensive analyses of evolutionary processes affecting human gene expression.

Even though a combination of purifying selection and selective neutrality probably predominates the evolution of regulatory and other types of functional variation, gene regulation constitutes a potential target for adaptive evolution, and eQTLs appear to be enriched for signs of recent positive selection in humans (77). Figure 1 provides an example of such a gene: the SNPs with the highest association to gene expression levels also have large allele frequency differences between populations, which is often a sign of recent positive selection that drives a beneficial allele to a higher frequency. Also 5′ regions that often harbor regulatory elements appear enriched for signs of positive selection (64,75). There are several case examples of recent adaptations through regulatory variation, for instance the convergent emergence of lactase tolerance via continued expression of the LCT gene after childhood (78), and resistance to malaria through a tissue-specific inactivation of the Duffy antigen (79,80). Macro-evolutionary differences between species may also derive from regulatory changes, and much work has been dedicated to characterizing events of positive selection that underlie human-specific adaptations. A large number of genes show a change in the expression level in humans compared with other primates, which may sometimes be a sign of directional selection (67).

Whether the regulation of genes that are expressed across multiple tissues are more or less frequent targets of selection than tissue-specific ones is a matter of debate, with results suggesting more regulatory constraint on ubiquitously expressed genes (66,67,83), or less constraint (84), as well as little overall correlation (64). Evolution of the human brain has been a topic of particular interest, and several studies have found an enrichment of regulatory adaptation in the brain (64,66,67). Additionally, the evolution of male reproduction, the brain and the dietary system appears to have been dominated by regulatory change, as opposed to an enrichment of coding changes in the evolution of, for instance, the immunological system (64,85). To date, little is known of the systemic targets of recent positive selection in humans due to the unavailability of expression data from multiple tissues from a variety of populations. In general, the modest degree of overlap between different studies suggests that the analyses of systemic targets of regulatory evolution have not been very robust. While straightforward grouping of genes according to tissue of expression has provided a good starting point for understanding the evolution of gene expression, it lacks the resolution that, for example, network-based approaches may have (86).

Changes in gene expression are likely to be one of the major mechanisms underlying differences in susceptibility to disease between individuals, for both Mendelian and common disease (31–33,36,87–90). Disease-associated genes have been observed to be enriched for negative selection both in the coding and regulatory regions, but there are also several examples of disease genes with signs of positive selection in their regulatory regions (64,77,91–93). However, it remains mostly unknown how often disease-causing coding and regulatory genetic variation has arisen through inefficiency of purifying selection, and to what degree it is a result of evolutionary trade-offs or past positive selection.
The effects of coding variation have been studied for decades on several levels, ranging from the cell to tissues and organisms, and further to the role of coding variation in disease and evolution. Now, this spectrum is being studied also for regulatory variation. Although mapping the causal variants of cis-regulatory variation in the eQTL region still remains a challenge due to the correlation structure of genetic variants, the annotation of proximal regulatory elements has progressed rapidly, and will become even more accurate through the many applications of novel sequencing technologies. Thus, finding loci with regulatory variants of high effect sizes in cis is starting to become straightforward. However, this likely accounts only for a small proportion of all the heritable variation of gene expression, and the biggest challenge now lies in understanding the contributions of common regulatory polymorphisms in trans, and rare regulatory variants. The different layers of transcriptional regulation and the complexity of feedback networks are difficult to untangle, given the statistical challenge of testing interactions across the entire genome. Furthermore, many genetic regulatory effects are likely not stable and ubiquitous, but are mediated through modifications of regulatory networks in a cell-type specific manner during different developmental stages or as a response to particular environmental conditions.

Yet, information of all of the aspects of the regulatory landscape is essential if we are to understand how variation in the genomes has given rise to the biological complexity we observe around us, within and between species (Fig. 2).

Evolutionary analysis of gene regulation has now moved from case examples to genome-wide approaches. Studies of individual genes and regulatory elements have provided intriguing examples of processes of evolutionary adaptation, but it is unfavourable to collect sufficient experimentally validated case examples to get an unbiased view of general evolutionary mechanisms and their relative importance. Genome-wide scanning is another approach to find targets of natural selection. However, it often yields long lists of candidate genes whose validation is difficult and that are often biased towards specific types and ages of selection. The search for general trends behind these gene lists has often been based on categorization according to gene ontology or tissue of expression, but these approaches lack resolution and have often yielded relatively inconsistent results between different studies. A general problem in evolutionary genetics is that it is relatively easy to come up with attractive stories of possible adaptive mechanisms even in the absence of real evidence. Some degree of uncertainty is inevitable because the evolutionary history cannot be rerun to obtain a truly independent replication, but especially now in the era of massive genomic data sets, we must aim to design and conduct studies that test well-defined hypotheses and answer specific questions about the evolution of genomes.

Conflict of Interest statement. None declared.

FUNDING
The funding was provided by the Louis-Jeantet foundation, the Academy of Finland and the Emil Aaltonen foundation.

REFERENCES


