Impacts of Variation in the Human Genome on Gene Regulation

Rajini R. Haraksingh and Michael P. Snyder

Department of Genetics, *Stanford University School of Medicine, MC: 5120, 300 Pasteur Drive, M-344, Stanford, CA 94305, USA*

Correspondence to Michael P. Snyder: *mpsnyder@stanford.edu*

http://dx.doi.org/10.1016/j.jmb.2013.07.015

Edited by M. Sternberg

Abstract

Recent advances in fast and inexpensive DNA sequencing have enabled the extensive study of genomic and transcriptomic variation in humans. Human genomic variation is composed of sequence and structural changes including single-nucleotide and multinucleotide variants, short insertions or deletions (indels), larger copy number variants, and similarly sized copy neutral inversions and translocations. It is now well established that any two genomes differ extensively and that structural changes constitute a prominent source of this variation. There have also been major technological advances in RNA sequencing to globally quantify and describe diversity in transcripts. Large consortia such as the 1000 Genomes Project and the ENCODE (ENCyclopedia Of DNA Elements) Project are producing increasingly comprehensive maps outlining the regions of the human genome containing variants and functional elements, respectively. Integration of genetic variation data and extensive annotation of functional genomic elements, along with the ability to measure global transcription, allow the impacts of genetic variants on gene expression to be resolved. There are several well-established models by which genetic variants affect gene regulation depending on the type, nature, and position of the variant with respect to the affected genes. These effects can be manifested in two ways: changes to transcript sequences and isoforms by coding variants, and changes to transcript abundance by dosage or regulatory variants. Here, we review the current state of how genetic variations impact gene regulation locally and globally in the human genome.

Introduction

Human genomes vary as a result of sequence and structural changes. Sequence variation comprises single-nucleotide and multinucleotide variants (SNVs and MNVs, respectively). Another class of variation consists of small insertions or deletions (indels) of a few nucleotides. Structural variation (SV) consists of larger copy number variations (CNVs) including deletions, duplications, and mobile-element insertions; copy neutral inversions and translocations; and chromosomal aneuploidies. Deciphering the functional impacts of variants in the human genome involves measuring their effects on gene expression. Transcription directs the manufacturing of proteins and functional RNAs, which in turn carry out the physical work of a cell. Genetic variants are capable of impacting transcriptional regulation in diverse ways according to the variant's size, nature, and location relative to the coding or regulatory regions of the gene in question. In this review, we will discuss the molecular effects of different types of genetic variants on gene regulation, including the effects on gene expression levels and transcript variability.

Variation in the Human Genome

Genomic variation consists of relative differences in the sequence, or in the arrangement of blocks of sequence between different genomes. These differences constitute a natural phenomenon of human genomes and are major contributors to human phenotypic variation. Genetic variation can result in benign or pathogenic phenotypes. Many genetic variants underlie adaptive traits and have become common in populations where they confer, or once conferred, a selective advantage.

Genetic variants are classified by the size of the varying segments of DNA, measured in base pairs, as well as by the nature and location of the events relative to a reference genome. The type of variant reflects whether DNA material was substituted, gained (duplicated or inserted), lost, or rearranged (inverted, translocated). Each variant is uniquely described by the sequence and the position of the variant relative to the reference sequence.

SNVs or point mutations are single-base-pair changes in one sequence compared to another. MNVs are changes in one sequence with respect to another, which are a few base pairs in length. SNVs and MNVs are substitution mutations where some nucleotides in one genome are replaced with other nucleotides in another, without any net gain or loss of genetic material. SNVs and MNVs that occur in >1% of individuals in a sampled population are usually referred to as single-nucleotide polymorphisms (SNPs) and multinucleotide polymorphisms (MNPs), respectively. The 1000 Genomes Project's latest effort mapped more than 38 million SNPs, 58% of which were previously unknown, in 14 different worldwide populations [1].

Indels are small insertions or deletions in one genome with respect to another and are generally between 1 and 50 bp in size. Indels that are not multiples of 3 bp in length and that lie within coding regions of genomic DNA result in frameshift mutations; not surprisingly, within coding regions of the genome, there is an enrichment of indels that are multiples of 3 bp [2]. A recent study based on data from the 1000 Genomes Project data has mapped 1.6 million indels in three populations using an approach to optimize the sensitivity and specificity of short indel discovery [3,4]. This study found that 21.8% of indel calls were novel to a database that houses all known indels, dbSNP 135. Interestingly, 43–48% of called indels occupy only 4.03% of the genome, clustering in homopolymer runs, tandem repeats, or predicted hotspots. In the rest of the genome, indels are rare, occurring 16 times less often than SNPs. More than 75% of indels arise due to polymerase slippage, while the remaining instances result from simple deletions or forkhead stalling and template switching [4].

Structural variants range in size from hundreds to hundreds of thousands and even millions of base pairs. These comprise both copy number variable changes and copy neutral changes. CNVs consist of deletions or insertions of stretches of DNA in one genome compared to another. Deletions can be either heterozygous or homozygous. However, whole gene insertions have been found to be present from 1 to upwards of 15 times in one genome compared to another [5]. Insertions can be tandem or dispersed in the genome. A predominant class of insertions called mobile-element insertions are those sequences that are derived from ancient transposable elements and persist in the genome, such as short and long interspersed nuclear elements (SINEs such as Alu elements and LINEs) [6]. To date, upwards of 28,000 unique CNVs have been catalogued, several of which are thought to be common in human populations [7]. There are on the order of thousands of CNVs between genomes [3,8]. CNVs that are polymorphic in a population are referred to as copy number polymorphisms.

Structural variation also includes similarly sized copy number neutral events, in which there is no net gain or loss of genomic content. These can be sequence inversions, in which a stretch of DNA sequence has been flipped between two endpoints, or balanced chromosomal translocations, in which two chromosomes have exchanged stretches of DNA sequence in one genome *versus* another.

Successive, increasingly comprehensive efforts to map all the variants in human genomes have demonstrated that CNVs or SVs are responsible for a much greater percentage of the total number of base pairs differing between normal genomes than SNPs [1,3,8–13]. Two normal genomes are estimated to differ by several percent of their length due to CNVs or SVs, but only by 0.1% due to SNPs [9,14]. In fact, a significant proportion of variants lie in regions where structural variation is inherently common [3,7,8].

A clear picture of the genomic and population distributions of genetic variants is beginning to emerge due to massive variant mapping efforts in worldwide populations and especially the 1000 Genomes Project. SVs and in particular CNVs are distributed in a non-random fashion in the genome and are biased towards 'hotspots' such as repetitive sequences. CNVs are enriched in segmental duplications and are biased away from genes [7,12]. The set of genes that is more likely impacted by CNVs is enriched for genes involved in sensory perception and interaction with the environment. Conversely, CNVs are depleted in genes that occupy central nodes in biological networks [8,15].

Variation in the human genome is extensive and is now well understood. Many variants have been shown to influence gene expression by several models, outlined below. Several methods for measuring global gene expression are employed in order to determine the impact of genetic variants on gene regulation.

Methods for Measuring Variation in Genome-Wide Transcription

There are two major methods to quantify genome-wide transcriptional activity in humans: gene expression microarrays and RNA sequencing (RNA-seq). Early attempts to measure locus-specific transcription involved quantitative PCR. However, the invention of DNA expression microarrays allowed for the

first time an efficient protocol for measuring the expression of multiple transcripts at once. These arrays contain single-stranded DNA molecules that tile known genes and splice variants. Total mRNA samples from a test and control are converted to cDNA, differentially labeled, and hybridized to the array. Here, the relative differences in fluorescence are measures of the relative amounts of certain transcripts between the test and control sample. RNA expression arrays were the standard for measuring transcription for many years despite the limitations that only known transcripts could be quantified, quantification was always relative, and cross-hybridization may sometimes render closely related transcripts indistinguishable.

The advent of DNA sequencing technology has allowed for even more precise methods for quantifying transcription. The method of RNA-seq was developed in order to directly sequence the total set of RNA molecules from a sample in order to measure variations in their abundance and sequence [16]. RNA-seq involves isolating all the RNA from a cell, converting it to cDNA, sequencing the cDNA fragments on high-throughput sequencing machines, and mapping the reads back to the genome. RNA-seq allows quantification of transcripts by measuring read depth differences between a test and a control. The number of RNA-seq reads that map to a particular locus can be used as a measure of the relative abundance of that transcript between samples. In single-end RNA-seq transcript abundance is often measured by the number of reads per kilobase per million mapped reads (RPKM). However, paired-end RNA-seq generates two reads per fragment. In this protocol it is only appropriate to count those fragments in which both reads are mappable. A common software suite for analyzing RNA-seq data, Cufflinks, measures transcript abundance in fragments per kilobase of exon per million mapped fragments (FPKM). In addition, the base pair resolution of sequencing technologies also allows for novel transcripts and transcript sequence variants to be quantified. The use of paired-end reads as opposed to single-end reads allows more precise mapping and detection of splice variants. Newly available strand-specific protocols can now detect the strand of a DNA molecule from which a particular transcript is derived [17]. In one method, this is achieved by differentially labeling the second strand in cDNA synthesis by using dUTP instead of dTTP. RNA-seq protocols differ in the methods used for capturing various types of transcripts. For example, some protocols focus on isolating only mRNA using poly-A selection. Other protocols use RiboZero to deplete rRNA in order to study all other RNA species, including lowly expressed transcripts. RNA-seq in its various forms has become the state of the art for quantifying genome-wide transcriptional activity.

Impact of Genetic Variation on Transcript Sequence

The impact of single-nucleotide variations on transcript coding sequence is well established. Such variation may result in synonymous or non-synonymous mutations. Synonymous mutations involve a change in genomic sequence that does not alter the encoded amino acid sequence. Non-synonymous mutations include missense and nonsense mutations, corresponding to a change in the encoded amino acid sequence or the introduction of a premature stop codon, respectively. Introduction of premature stop codons in transcripts will likely either produce shortened peptides or initiate nonsense-mediated decay, resulting in no functional protein being produced. Additionally, SNVs that alter the initial methionine codon of a transcript may render the sequence untranslatable. SNVs that convert stop codons into amino acid codons may alter the length of the encoded peptide. MNVs may produce synonymous, or single or multiple non-synonymous mutations [18].

Indels can affect coding sequence by producing either non-frameshift or frameshift mutations. Non-frameshift mutations occur when the length of the indel is a multiple of 3, and only the part of the amino acid sequence directly covered by the indel is altered. Frameshift mutations occur when the length of the indel is not a multiple of 3, and the entire amino acid sequence to the 5′ end of the genomic variant is affected.

Ten protein-coding genes have been identified with a predicted indel rate greater than 2×10^{-5} per generation across the coding sequence. Three of these genes are known to be associated with disease. *HTT* is associated with Huntington's disease. *AR* is associated with prostate cancer, spinal and bulbar muscular dystrophy, and infertility. *ARID1B* is associated with various forms of neurodevelopmental abormalities [4]. In general, it has been found that indels are under stronger purifying selection than SNPs in functional genomic regions [4]. This is likely due to the strong deleterious effects of frameshift mutations.

The impacts of indels on transcription have not been ascertained to the same extent as that for SNPs and SVs. However, it appears that despite their low rates, the functional impacts of indels may be considerable. In particular, the ~20 repeat expansion diseases identified to date provide the most well understood examples of indel effects on transcription. All repeat expansion diseases that are currently known to contain expansions within exons contain the repeat sequence CAG·CTG. There are at least 10 such diseases known including Huntington's disease. The expansion results in an increase in length of a polyglutamine (poly Q) tract in the protein encoded by the gene [19]. The poly Q tract is thought

to render the protein toxic and lead to various malfunctions [20].

CNVs can affect the coding complement of a transcript by deleting or inserting exonic sequences that may result in frameshift or non-frameshift mutations, or splice variants depending on the length or position of the CNV relative to the exons. CNVs that affect intronic sequences may lead to alternative splicing from the original sequence. Rearrangements involving the red and green opsin gene cluster on chromosome X can lead to the generation of mosaic genes that cause various forms of red–green color distortion. Total deletion of one or the other type of opsin results in red–green color blindness [21].

Other structural variants such as inversions or translocations may affect the coding complement of a gene if the event lies in the appropriate position. For example, inversions that flip exonic sequences will change the sequence of amino acids encoded by the transcript and may induce missense or nonsense mutations, or alter patterns of gene splicing by altering the splice donor and acceptor sites.

Sequence variations may also affect untranslated transcripts such as miRNAs and other small RNAs, thereby affecting their function. For example, SNVs, MNVs, or indels may cause perturbations in RNA secondary structure, which may, in turn, result in altered binding of the RNA in question to other transcripts or genomic regions [22].

Impact of Genetic Variants on Transcript Abundance

There are two main ways in which genetic variants may affect the level of expression of a gene. CNVs may alter gene dosage by altering the number of copies of a gene that is present in the genome. SNVs or SVs may alter gene regulatory elements, thus perturbing the assembly of the transcription machinery. In both cases, variation at the genomic level may result in differences in the amount of expression of a transcript.

There have been several reports in the literature of gene dosage effects due to variations in gene copy number [7,8,11]. This phenomenon was known even before the discovery that CNVs are a major source of human genetic variation. As early as 1993, the number of active copies of the cytochrome P450 *CYP2D6* gene and the amount of metabolism of the CYP2D6 substrate were shown to be correlated [23]. However, there have been few definitive demonstrations in the literature of the direct effects of copy number variants on gene expression. One of the most conclusive studies of a gene dosage effect showed that the salivary amylase (*AMY1*) gene copy number, salivary amylase protein level, and the starch content of the diets of several different populations around the world are positively correlated [5]. The *AMY1* gene copy number can range from 2 to 15 copies in populations worldwide. *AMY1* encodes the enzyme salivary amylase, which metabolizes starch. This work implies that high *AMY1* copy number was selected for in populations with high starch content and that this increased gene dosage leads to more efficient starch metabolism.

CNVs can also affect the amount of gene expression when they occur in the regulatory sequence elements of a gene. The specific effect is determined by the type of regulatory element affected (activator or repressor) and the way in which it is affected, that is, altering the complement of regulatory elements (e.g., deletion or duplication of an activator) or the structure (e.g., novel insertion that nullifies a repressor). A pathological example of a CNV functioning in this way was demonstrated for susceptibility to Crohn's disease (CD). Here, a 20-kb deletion polymorphism immediately upstream of the immunity-related GTPase family M (*IRGM*) gene, and in perfect linkage equilibrium with the most strongly associated CD SNP, causes the gene to segregate in the population with two distinct upstream sequences. The deletion allele and reference allele showed distinct *IRGM* expression patterns [24]. These expression patterns modulate a biological process implicated in CD. The CD association at this locus results from a CNV in the upstream regulatory region of the *IRGM* gene that likely affects the expression of the gene in such a way as to cause the phenotype.

Aside from examples of single loci, there has been little work demonstrating the genome-wide effects of CNV on transcription and in particular gene copy number with gene expression activity. Recently, an accumulation of CNV association studies has produced several strong associations with various phenotypes, indicating that CNVs have profound effects on human health and disease, especially neurological and developmental diseases [24–35]. CNVs have also been shown to be associated with non-pathogenic traits such as height and body mass index [36,37]. However, direct measurement of these CNVs on transcription has been lacking. In 2007, the relationship between gene copy number and gene expression was demonstrated [38]. Array Comparative Genome Hybridization was used for measuring DNA copy number, and expression microarrays were used for measuring transcript expression. A general positive correlation between gene copy number and gene expression was found. However, this study only interrogated 15,000 transcripts and found that only 17% of the variation in expression was explained by CNVs, while 80% was explained by SNPs. Subsequent to this, there have been a handful of studies investigating the genome-wide effects of CNVs on expression. In 2010, a similar trend was found using sequencing-based

technologies for mapping structural variations and measuring transcription genome-wide [39]. The next year, Schlattl *et al.* examined the relationships between different categories of copy number variants based on size, type, and overlap with genes, and gene expression [40]. Using fine-scale CNV mapping and expression data from 129 individuals from the 1000 Genomes Project, they observed the expression of 110 genes to be associated with CNVs. This set of CNVs is enriched for large (>4 kb) events. Again, a general positive correlation between copy number and expression was demonstrated, and this correlation was stronger with CNVs than with nearby SNPs. The authors argue that these data indicate a more causative role for CNVs than SNPs in expression quantitative trait loci. Additionally, pathogenic CNVs are more likely to contain aberrant transcription of genes within or nearby the CNV [41]. Interestingly, although a general trend of positive correlation between gene copy number and gene expression has been observed, there are a minority of cases reported for which the trend is reversed or for which no change in expression is associated with changes in copy number [40,42].

CNVs are also known to have a global effect on the transcriptome [43]. In addition to these gross genome-wide correlations between gene copy number and gene transcription level, several unexpected local effects of CNVs on expression have been observed. It has been well established that most normal genomes contain several very large CNVs [44]. Further, it is now apparent that some of these large CNVs are indeed associated with pathogenic phenotypes [41,45]. In particular, large CNVs have been shown to exert a field effect on the expression of genes within and around the CNV in mice [46] and in human cell lines [41,47]. Such effects are likely due to alterations of regulatory elements for genes that lie outside of the immediate copy variable region.

Copy neutral structural changes can affect transcription either by directly creating breakpoints within transcripts or by rearranging regulatory elements and creating position effects. Such *cis* regulatory position effects have been reported up to 1.5 Mb away from the gene in question [42,48]. Additionally, large structural variations may alter the spatial distribution of chromosomes in the nucleus. Such nuclear reorganization may disrupt *cis* and *trans* regulatory interactions, producing adverse effects on transcription [42].

Noncoding indels are capable of affecting gene regulation by modulating transcription, silencing genes, sequestering proteins involved in splicing and cell architecture, and generating chromosomal fragility. Tandem repeats may act as origins of replication, intrinsic promoter components, transcription enhancers, blocks to transcription elongation, or gene silencers. These effects are reviewed elsewhere [19].

Gene regulatory elements are also affected by point mutations, which in turn disrupt the assembly of transcription machinery and the propagation of transcriptional activating signals. Using a combination of chromatin immunoprecipitation followed by sequencing (ChIP-seq) and deep RNA-seq, Kasowski *et al.* showed that transcription factor binding differences are associated with SNPs in gene regulatory regions, and these differences were correlated with mRNA abundance in 10 individual cell lines [39]. This phenomenon was demonstrated for two different transcription factors, PolII and NFκB. Recently, Reddy *et al.* measured genome-wide allelic differences in gene expression and transcription factor binding in the individual NA12878 of 24 sequence-specific transcription factors. This study leveraged an updated human reference genome that included homozygous and heterozygous sites based on 1000 Genomes Project data. A strong association was found between allelic occupancy and expression within 100 bp of the transcription start site. Additionally, sites showing differential allelic occupancy were significantly enriched for disease-associated and particularly autoimmune-associated variants [49]. These results suggest that disease-associated allelic variants in gene regulatory regions have functional implications due to differences in allelic transcription factor occupancy. The ENCODE (ENCyclopedia Of DNA Elements) Project aims to identify all functional elements in the human genome, including coding and noncoding transcripts, marks of accessible chromatin, and protein binding sites [50–53]. To date, ChIP-seq has been performed for 119 different transcription factors in 147 different cell lines. Expression has also been quantified in many of these cell lines. The ENCODE project represents the largest body of data for studying the interaction between genomic variation in gene regulatory regions with expression. Work to understand the relationship between genomic variation in transcription factor binding sites and gene expression is ongoing.

Conclusions

Massive advances in our ability to map the entire spectrum of genetic variants genome-wide at high resolution have led to a better understanding of the scope and distribution of human genomic variation. In particular, the distribution of genetic variants with respect to the functional regions of the genome including genes and gene regulatory elements is now well characterized. In addition, there have been concurrent advances in our ability to quantify the amount of gene expression as well as transcript variability using allele-specific RNA-seq. Combining genome-wide variant mapping and gene expression analysis has led to the characterization of the

functional implications of the different types of genetic variants on gene expression. SNVs, indels, and SVs are all capable of affecting transcript sequence. CNVs are able to directly affect gene dosage, and SNVs, indels, and SVs that lie in gene regulatory regions have been shown to perturb gene expression in the corresponding genes. However, genetic differences do not successfully explain all the variability of gene expression. Indeed, epigenetic changes such as DNA methylation and chromatin modification also affect gene expression, and the integration of other types of omics data (the epigenome, chromatin folding, regulatory DNA elements, proteome, and metabolome among others) will be useful in obtaining a more comprehensive view by which genetic variation and other modifications affect gene expression. Such efforts are already underway for a handful of genomes [54–56].

It is important to note that the work discussed here pertains mostly to steady-state effects of variation on expression. This is due to limitations in methodology. Most methods can only detect steady-state gene expression or transcription factor binding. However, advances in single-cell technologies and analysis of nascent transcripts are beginning to enable the study of dynamic spatial and temporal expression [57–59]. Furthermore, we note that most of the global studies to date that have examined the relationships between the genome and the transcriptome have done so only in lymphoblastoid cell lines. It remains to be seen if these relationships persist in other cell types.

Many open questions remain regarding the relationship between genetic variant and gene regulation. The full range of transcription factor binding properties in relation to variations in regulatory elements has not been described. In addition, the regions of the genome that do not tolerate CNVs are not fully catalogued, and the relationships among the regions where CNVs exist and their genomic sequence and chromatin structure are yet to be resolved. Thus, the full extent of CNV on transcription is not known. However, it is clear that we now possess the technologies to comprehensively probe the impact of genetic variations on gene regulation.

Received 8 May 2013;
Received in revised form 10 July 2013;
Accepted 10 July 2013
Available online 16 July 2013

Keywords:
genetic variation;
gene expression;
transcription;
structural variation;
gene regulation

Abbreviations used:
SNV, single-nucleotide variant; MNV, multinucleotide variant; SV, structural variation; CNV, copy number variation; SNP, single-nucleotide polymorphism; RNA-seq, RNA sequencing; CD, Crohn's disease.

References

[1] Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, et al. An integrated map of genetic variation from 1,092 human genomes. Nature 2012;491:56–65.
[2] Clark MJ, Chen R, Lam HY, Karczewski KJ, Euskirchen G, Butte AJ, et al. Performance comparison of exome DNA sequencing technologies. Nat Biotechnol 2011;29:908–14.
[3] Consortium GP. A map of human genome variation from population-scale sequencing. Nature 2010;467:1061–73.
[4] Montgomery SB, Goode DL, Kvikstad E, Albers CA, Zhang ZD, Mu XJ, et al. The origin, evolution, and functional impact of short insertion-deletion variants identified in 179 human genomes. Genome Res 2013;23:749–61.
[5] Perry GH, Dominy NJ, Claw KG, Lee AS, Fiegler H, Redon R, et al. Diet and the evolution of human amylase gene copy number variation. Nat Genet 2007;39:1256–60.
[6] Stewart C, Kural D, Strömberg MP, Walker JA, Konkel MK, Stütz AM, et al. A comprehensive map of mobile element insertion polymorphisms in humans. PLoS Genet 2011;7:e1002236.
[7] Mills RE, Walter K, Stewart C, Handsaker RE, Chen K, Alkan C, et al. Mapping copy number variation by population-scale genome sequencing. Nature 2011;470:59–65.
[8] Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, Zhang Y, et al. Origins and functional impact of copy number variation in the human genome. Nature 2010;464:704–12.
[9] Consortium IH. A haplotype map of the human genome. Nature 2005;437:1299–320.
[10] Tuzun E, Sharp AJ, Bailey JA, Kaul R, Morrison VA, Pertz LM, et al. Fine-scale structural variation of the human genome. Nat Genet 2005;37:727–32.
[11] Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, et al. Global variation in copy number in the human genome. Nature 2006;444:444–54.
[12] Korbel JO, Urban AE, Affourtit JP, Godwin B, Grubert F, Simons JF, et al. Paired-end mapping reveals extensive structural variation in the human genome. Science 2007;318:420–6.
[13] Mills Ryan E, Walter Klaudia, Stewart Donald A, Handsaker Robert, Ken Chen, Alkan Can, et al. Mapping structural variation at fine-scale by population-scale genome sequencing; 2010.
[14] Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature 2001;409:860–921.
[15] Schuster-Böckler B, Conrad D, Bateman A. Dosage sensitivity shapes the evolution of copy-number varied regions. PLoS One 2010;5:e9474.
[16] Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 2009;10:57–63.

[17] Levin JZ, Yassour M, Adiconis X, Nusbaum C, Thompson DA, Friedman N, et al. Comprehensive comparative analysis of strand-specific RNA sequencing methods. Nat Methods 2010;7:709–15.
[18] Shastry BS. SNPs: impact on gene function and phenotype. Methods Mol Biol 2009;578:3–22.
[19] Usdin K. The biological effects of simple tandem repeats: lessons from the repeat expansion diseases. Genome Res 2008;18:1011–9.
[20] Orr HT, Zoghbi HY. Trinucleotide repeat disorders. Annu Rev Neurosci 2007;30:575–621.
[21] Deeb SS. Genetics of variation in human color vision and the retinal cone mosaic. Curr Opin Genet Dev 2006;16:301–7.
[22] Cai Y, Yu X, Hu S, Yu J. A brief review on the mechanisms of miRNA regulation. Genomics Proteomics Bioinformatics 2009;7:147–54.
[23] Johansson I, Lundqvist E, Bertilsson L, Dahl ML, Sjöqvist F, Ingelman-Sundberg M. Inherited amplification of an active gene in the cytochrome P450 CYP2D locus as a cause of ultrarapid metabolism of debrisoquine. Proc Natl Acad Sci U S A 1993;90:11825–9.
[24] McCarroll SA, Huett A, Kuballa P, Chilewski SD, Landry A, Goyette P, et al. Deletion polymorphism upstream of IRGM associated with altered IRGM expression and Crohn's disease. Nat Genet; 2008 .
[25] Marshall CR, Scherer SW. Detection and characterization of copy number variation in autism spectrum disorder. Methods Mol Biol 2012;838:115–35.
[26] Levinson DF, Duan J, Oh S, Wang K, Sanders AR, Shi J, et al. Copy number variants in schizophrenia: confirmation of five previous findings and new evidence for 3q29 microdeletions and VIPR2 duplications. Am J Psychiatry 2011;168:302–16.
[27] Malhotra D, McCarthy S, Michaelson JJ, Vacic V, Burdick KE, Yoon S, et al. High frequencies of de novo CNVs in bipolar disorder and schizophrenia. Neuron 2011;72:951–63.
[28] Pankratz N, Dumitriu A, Hetrick KN, Sun M, Latourelle JC, Wilk JB, et al. Copy number variation in familial Parkinson disease. PLoS One 2011;6:e20988.
[29] Kawamura Y, Otowa T, Koike A, Sugaya N, Yoshida E, Yasuda S, et al. A genome-wide CNV association study on panic disorder in a Japanese population. J Hum Genet 2011;56:852–6.
[30] Fernandez TV, Sanders SJ, Yurkiewicz IR, Ercan-Sencicek AG, Kim YS, Fishman DO, et al. Rare copy number variants in tourette syndrome disrupt genes in histaminergic pathways and overlap with autism. Biol Psychiatry 2012;71:392–402.
[31] Pfundt R, Veltman JA. Structural genomic variation in intellectual disability. Methods Mol Biol 2012;838:77–95.
[32] Yeo RA, Gangestad SW, Liu J, Calhoun VD, Hutchison KE. Rare copy number deletions predict individual variation in intelligence. PLoS One 2011;6:e16339.
[33] Gonzalez E, Kulkarni H, Bolivar H, Mangano A, Sanchez R, Catano G, et al. The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. Science 2005;307:1434–40.
[34] Hollox EJ, Huffmeier U, Zeeuwen PL, Palla R, Lascorz J, Rodijk-Olthuis D, et al. Psoriasis is associated with increased beta-defensin genomic copy number. Nat Genet 2008;40:23–5.
[35] Zhang Y, Haraksingh R, Grubert F, Abyzov A, Gerstein M, Weissman S, et al. Child development and structural variation in the human genome. Child Dev 2013;84:34–48.
[36] Dauber A, Yu Y, Turchin MC, Chiang CW, Meng YA, Demerath EW, et al. Genome-wide association of copynumber variation reveals an association between short stature and the presence of low-frequency genomic deletions. Am J Hum Genet 2011;89:751–9.
[37] Hai R, Pei YF, Shen H, Zhang L, Liu XG, Lin Y, et al. Genome-wide association study of copy number variation identified gremlin1 as a candidate gene for lean body mass. J Hum Genet 2012;57:33–7.
[38] Stranger BE, Forrest MS, Dunning M, Ingle CE, Beazley C, Thorne N, et al. Relative impact of nucleotide and copy number variation on gene expression phenotypes. Science 2007;315:848–53.
[39] Kasowski M, Grubert F, Heffelfinger C, Hariharan M, Asabere A, Waszak SM, et al. Variation in transcription factor binding among humans. Science 2010;328:232–5.
[40] Schlattl A, Anders S, Waszak SM, Huber W, Korbel JO. Relating CNVs to transcriptome data at fine resolution: assessment of the effect of variant size, type, and overlap with functional regions. Genome Res 2011;21:2004–13.
[41] Luo R, Sanders SJ, Tian Y, Voineagu I, Huang N, Chu SH, et al. Genome-wide transcriptome profiling reveals the functional impact of rare de novo and recurrent CNVs in autism spectrum disorders. Am J Hum Genet 2012;91:38–55.
[42] Harewood L, Chaignat E, Reymond A. Structural variation and its effect on expression. Methods Mol Biol 2012;838:173–86.
[43] Henrichsen CN, Chaignat E, Reymond A. Copy number variants, diseases and gene expression. Hum Mol Genet 2009;18:R1–8.
[44] Itsara A, Cooper GM, Baker C, Girirajan S, Li J, Absher D, et al. Population analysis of large copy number variants and hotspots of human genetic disease. Am J Hum Genet 2009;84:148–61.
[45] Girirajan S, Brkanac Z, Coe BP, Baker C, Vives L, Vu TH, et al. Relative burden of large CNVs on a range of neurodevelopmental phenotypes. PLoS Genet 2011;7:e1002334.
[46] Henrichsen CN, Vinckenbosch N, Zöllner S, Chaignat E, Pradervand S, Schütz F, et al. Segmental copy number variation shapes tissue transcriptomes. Nat Genet 2009;41:424–9.
[47] Reymond A, Henrichsen CN, Harewood L, Merla G. Side effects of genome structural changes. Curr Opin Genet Dev 2007;17:381–6.
[48] Harewood L, Schütz F, Boyle S, Perry P, Delorenzi M, Bickmore WA, et al. The effect of translocation-induced nuclear reorganization on gene expression. Genome Res 2010;20:554–64.
[49] Reddy TE, Gertz J, Pauli F, Kucera KS, Varley KE, Newberry KM, et al. Effects of sequence variation on differential allelic transcription factor occupancy and gene expression. Genome Res 2012;22:860–9.
[50] Consortium EP. The ENCODE (ENCyclopedia Of DNA Elements) Project. Science 2004;306:636–40.
[51] Birney E, Stamatoyannopoulos JA, Dutta A, Guigó R, Gingeras TR, Margulies EH, et al. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature 2007;447:799–816.
[52] Consortium EP. A user's guide to the encyclopedia of DNA elements (ENCODE). PLoS Biol 2011;9:e1001046.
[53] Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis CA, Doyle F, et al. An integrated encyclopedia of DNA elements in the human genome. Nature 2012;489:57–74.
[54] Chen R, Mias GI, Li-Pook-Than J, Jiang L, Lam HY, Miriami E, et al. Personal omics profiling reveals dynamic molecular and medical phenotypes. Cell 2012;148:1293–307.
[55] Cheng C, Alexander R, Min R, Leng J, Yip KY, Rozowsky J, et al. Understanding transcriptional regulation by integrative analysis of transcription factor binding data. Genome Res 2012;22:1658–67.

[56] Wang J, Zhuang J, Iyer S, Lin X, Whitfield TW, Greven MC, et al. Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. Genome Res 2012;22:1798–812.
[57] Gilad Y. Using genomic tools to study regulatory evolution. Methods Mol Biol 2012;856:335–61.
[58] Core LJ, Waterfall JJ, Lis JT. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. Science 2008;322:1845–8.
[59] Churchman LS, Weissman JS. Nascent transcript sequencing visualizes transcription at nucleotide resolution. Nature 2011;469: 368–73.