REVIEW

Human Mutation

OFFICIAL JOURNAL
HGVS
HUMAN GENOME VARIATION SOCIETY
www.hgvs.org

Genes, Mutations, and Human Inherited Disease at the Dawn of the Age of Personalized Genomics

David N. Cooper,[1]* Jian-Min Chen,[2] Edward V. Ball,[1] Katy Howells,[1] Matthew Mort,[1] Andrew D. Phillips,[1] Nadia Chuzhanova,[3] Michael Krawczak,[4] Hildegard Kehrer-Sawatzki,[5] and Peter D. Stenson[1]

[1]Institute of Medical Genetics, School of Medicine, Cardiff University, Heath Park, Cardiff, United Kingdom; [2]Institut National de la Santé et de la Recherche Médicale (INSERM), U613 and Etablissement Français du Sang (EFS) – Bretagne, Brest, France; [3]School of Science and Technology, Nottingham Trent University, Nottingham, United Kingdom; [4]Institut für Medizinische Informatik und Statistik, Christian-Albrechts-Universität, Kiel, Germany; [5]Institut für Humangenetik, Universität Ulm, Ulm, Germany

Communicated by Richard G.H. Cotton

Received 21 January 2010; accepted revised manuscript 26 March 2010.
Published online 13 April 2010 in Wiley InterScience (www.interscience.wiley.com). DOI 10.1002/humu.21260

ABSTRACT: The number of reported germline mutations in human nuclear genes, either underlying or associated with inherited disease, has now exceeded 100,000 in more than 3,700 different genes. The availability of these data has both revolutionized the study of the morbid anatomy of the human genome and facilitated "personalized genomics." With ~300 new "inherited disease genes" (and ~10,000 new mutations) being identified annually, it is pertinent to ask how many "inherited disease genes" there are in the human genome, how many mutations reside within them, and where such lesions are likely to be located? To address these questions, it is necessary not only to reconsider how we define human genes but also to explore notions of gene "essentiality" and "dispensability." Answers to these questions are now emerging from recent novel insights into genome structure and function and through complete genome sequence information derived from multiple individual human genomes. However, a change in focus toward screening functional genomic elements as opposed to genes *sensu stricto* will be required if we are to capitalize fully on recent technical and conceptual advances and identify new types of disease-associated mutation within noncoding regions remote from the genes whose function they disrupt. Hum Mutat 31:631–655, 2010. © 2010 Wiley-Liss, Inc.

KEY WORDS: Human Gene Mutation Database; HGMD; inherited mutations; human genome; gene number; gene definition; disease genes; gene essentiality; noncoding regions; functionome; mutome; mutation detection

Introduction

What man that sees the ever-whirling wheele
Of Change, the which all mortall things doth sway,
But that thereby doth find, and plainly feele,
How mutability in them doth play
Her cruell sports, to many men's decay?
(Edmund Spenser, *The Faerie Queene*, Book VII,

"Two Cantos of Mutabilitie," Canto VI, stanza 1, published posthumously in 1609).

*Correspondence to: David N. Cooper, Institute of Medical Genetics, School of Medicine, Cardiff University, Heath Park, Cardiff CF14 4XN, UK.
E-mail: CooperDN@cardiff.ac.uk

Just over 30 years ago, the first heritable human gene mutations were characterized at the DNA level: gross deletions of the human α-globin (*HBA*; MIM# 141800) and β-globin (*HBB*; MIM# 141900) gene clusters giving rise to α- and β- thalassaemia [Orkin et al., 1978] and a single base-pair substitution (Lys17Term) in the human β-globin (*HBB*) gene causing β-thalassaemia [Chang and Kan, 1979]. With the number of known germline mutations in human nuclear genes either underlying or associated with inherited disease now exceeding 100,000 in over 3,700 different genes (Human Gene Mutation Database [HGMD]; http://www.hgmd.org; March 2010 update) [Stenson et al., 2009], the characterization of the spectrum of human germline mutations has reached a symbolic landmark.

Newly described human gene mutations are currently accumulating at a rate of ~10,000 per annum, with ~300 new "inherited disease genes" being recognized every year. It is therefore pertinent to pose the double question: how many inherited disease genes are there in the human genome and how many mutations are likely to be found within them? A first bold estimate of the "number of mutations causing inherited disease" (20 million mutations apportioned between 20,000 different human genes) has recently been put forward [Cotton, 2009], but these numbers appear to constitute only rough estimates that have not been justified in any formal way.

In principle, the number of human "disease genes" may well be estimable, albeit approximately. However, although the number of different mutations that could *potentially* cause human inherited disease is clearly almost limitless (if, e.g., one were to include all possible frameshift microdeletions and microinsertions), the number of mutations *actually* in existence and available to be identified and characterized is a complex function of the mutability of each inherited disease gene, the prevalence and ease of ascertainment of the consequent clinical phenotype(s), the demographic history of the human population, as well as the technical means at our disposal to locate and identify the pathological mutation(s) in any one individual.

Reich and Lander [2001] concluded that, with a "typical" (pathological) gene mutation rate of 3.2×10^{-6} per generation, the average number of mutations underlying a rare inherited disease would equal 77,000 at mutation-drift equilibrium. These authors also opined that the kinetics of the mutation process are such that, for diseases characterized by an overall population frequency of pathological mutations <1%, this equilibrium is likely to have been reached in the extant human population. Based upon these considerations, the number of different mutations *actually* underlying inherited human disease is likely to be one to

two orders of magnitude higher than that suggested by Cotton [2009], potentially totalling between 600 million and 2.4 billion (average: 1.2 billion) depending upon the number of genes (estimated to lie somewhere between 7,750 and 30,770, with an average of 15,300; see below) adjudged to qualify as "inherited disease genes." However, most of these mutations will be extremely rare. Indeed, it can be calculated from the approximate distribution function of allele numbers at mutation-drift equilibrium [Gale, 1990] that, given an overall population frequency of pathological mutations of 1% in a given gene, fewer than four mutations will have a relative frequency $> 5 \times 10^{-4}$ in the pool of pathological mutations of that gene. Thus, in terms of those inherited disease mutations that are in practice actually detectable, the above figures are likely to represent gross overestimates, and the number of mutations detected in a given gene will depend mostly upon the number of patients studied rather than on the diversity of the underlying mutational spectrum of that gene.

In attempting to collate all inherited human pathological gene mutations as they emerge in the literature [Stenson et al., 2009], HGMD has to some extent embarked on an open ended project whose eventual scale and scope was quite impossible to assess from the outset. Daunting as this prospect is, it is nevertheless appropriate at this juncture to take stock and try to assess where we are in terms of the indubitably massive task of identifying, annotating, and cataloguing the human germline mutational spectrum ("mutome"). We shall argue that, although the question of the "number of mutations causing inherited disease" may well be akin to asking "How long is a piece of string?," there are several related questions that appear to be worthwhile addressing on account of their practical importance: *How many genes are there in the human genome? How many of these are inherited disease genes (i.e., genes harboring mutations that are capable of causing inherited disease)? What proportion of the universe of possible mutations within these inherited disease genes is likely to be of pathological significance? Where, in genomic terms, are these mutations likely to be found? How many deleterious mutations are there on average per individual?* The answers to these questions should shed some light on the likely size of the task facing us as we attempt to document the spectrum of mutations causing (or associated with) human inherited disease.

How Many Genes are There in the Human Genome?

Defining the Gene in a Complex Genome

The answer to the question of how many genes there are in the human genome is in large part dependent upon how we opt to define the term "gene." Initial annotation data indicated that the human genome encodes at least 20,000–25,000 protein-coding genes with an indeterminate number of additional "computationally derived genes" supported by somewhat weaker in silico evidence [International Human Genome Sequencing Consortium, 2004; Venter et al., 2001]. Many genes are now known to encode RNAs rather than proteins as their final products [Griffiths-Jones, 2007; see below] but many still remain unannotated [Kapranov et al., 2007b]. In the latest assembly of the human genome (Genome Reference Consortium, release GRCh37, Feb. 2009), the Genebuild published by Ensembl (database version 56.37a) includes 23,616 protein-coding genes, 6,407 putative RNA genes and 12,346 pseudogenes (http://www.ensembl.org/Homo_sapiens/Info/StatsTable). The HUGO Human Gene Nomenclature Committee (http://www.genenames.org/index.html) has so far approved more than 28,000 human gene symbols, although some of these may yet turn out to correspond to functionally meaningless open reading frames [Clamp et al., 2010]. It is nevertheless encouraging that at least 17,052 human genes have been shown to have orthologous counterparts in the mouse genome, suggesting that they do indeed correspond to real proteins [Pruitt et al., 2009]. However, the definition of what constitutes a gene is still fairly fluid, and hence, depending upon the precise definition adopted, it may be that many additional human "genes" still remain to be described and annotated.

To appreciate why definition is an issue here, one need only be aware of the many exceptions to genes being contiguous (as well as functionally and spatially distinct) entities, as classically envisaged. Thus, some genes are known to occur within the introns of other genes [Herzog et al., 1996; Vuoristo et al., 2001; Yu et al., 2005]. Some genes can overlap with each other either on the same or on different DNA strands [Denoeud et al., 2007], resulting in the sharing of some of their coding and/or regulatory elements [van Bokhoven et al., 1996; Yang and Elnitski, 2008]. In addition, the vast majority of human genes are now known to undergo alternative splicing [Pan et al., 2008], leading in some cases to quite different proteins being encoded by the same gene. For example, the human *CDKN2A* gene (MIM# 600160) encodes an alternatively spliced variant ($p14^{ARF}$) that, through the inclusion of an alternative first exon, acquires an altered reading frame so as to specify a protein product that is structurally unrelated to the other p16 isoforms encoded by this gene.

Bicistronic genes (e.g., *MOCS1*; MIM# 603707) [Gross-Hardt and Reiss, 2002] are also atypical, with transcription initiating from one gene and continuing in *cis* through a neighboring downstream gene to yield a precursor protein that is then cleaved to generate different proteins. Such "transcription-mediated gene fusion" may well not be an infrequent occurrence in the human genome [Akiva et al., 2006; Parra et al., 2006]. Moreover, there is now persuasive evidence for the occurrence of *trans*-splicing in human cells, involving the cleavage and joining of entirely separate RNA transcripts [Gingeras et al., 2009].

Many protein-coding genes have been found to possess alternative transcriptional initiation sites, some of which may be quite remote from the gene itself, in some instances even residing within the bounds of another gene [Carninci et al., 2006; Denoeud et al., 2007]. Other genes exhibit differential polyadenylation site usage leading to length heterogeneity of the 3′ untranslated region [Kwan et al., 2008].

Should distant *cis*-acting regulatory sequences be included within the boundaries of the gene they serve to regulate? If so, then it would make the concept of the gene that much more flexible. Indeed, if we are prepared to redefine what constitutes a gene, should we perhaps entertain the concept of an extended gene whose component parts are not necessarily contiguous on the same DNA strand or even on the same chromosome? In exploring further the complexity of human genes below, it will be seen how difficult it has become to put forward a general definition of the gene, either structurally or functionally, that will withstand close scrutiny in the context of the diversity displayed by the many thousands of known human genes. In the light of recent conceptual advances, the inherent limitations of the gene-centric strategies routinely employed to detect disease-associated mutations will be all too evident.

Transcripts of Unknown Function and Unannotated Transcripts

The *ENCyclopedia of DNA Elements* (ENCODE) project, designed to analyze 30 megabases (Mb) of DNA from 44 genomic

regions (thereby sampling 1% of the genome) to characterize the functional elements present, has identified complex patterns of regulation and "pervasive transcription" of the human genome [ENCODE Project Consortium, 2007]. Although >90% of the human genome appears to be represented in nuclear primary transcripts, it has become clear that only 35–50% of processed transcripts have so far been annotated as genes, implying that many genes may not yet have been recognized as such [ENCODE Project Consortium, 2007; Gingeras, 2007; Rozowsky et al., 2007; Sultan et al., 2008]. Thus, large numbers of hitherto unannotated transcripts may well yet turn out to be of functional significance [Gingeras, 2007]. Such transcripts have been collectively classified as *transcripts of unknown function* (TUFs) and are thought to include (1) antisense transcripts of protein-coding genes, (2) isoforms of protein-coding genes, and (3) transcripts that either overlap introns of annotated gene transcripts (on the same strand) or which are derived entirely from intergenic regions. Although both the complexity and abundance of TUFs are quite remarkable, it should be realized that there is often no firm evidence for these transcripts being of functional significance. Indeed, unannotated nonpolyadenylated transcripts originating from intergenic regions have been found to represent the bulk of the >90% of the human genome that now appears to be transcribed [Gingeras, 2007; Kapranov et al., 2002, 2007a]. Although the functional significance of "pervasive transcription" remains unclear, it is much more extensive than had previously been realized [Dinger et al., 2009].

In both humans and the mouse, up to 70% of genomic loci exhibit evidence of transcription from the antisense strand as well as the sense strand [Grinchuk et al., 2010; Katayama et al., 2005; Werner et al., 2009]. These naturally occurring antisense transcripts may modulate the level of expression of their associated sense transcripts (or otherwise influence their processing) thereby adding another level of complexity to the regulation of gene expression [Faghihi and Wahlestedt, 2009; He et al., 2008]. Although there is, as yet, no suggestion that the genomic sources of such antisense transcripts should be regarded as genes in their own right, their prevalence clearly renders our task of defining the gene that much more difficult.

RNA Genes

A large proportion of the human transcriptome still remains to be annotated [Peters et al., 2007]. Although some of the overall transcriptional activity may simply be "transcriptional noise" [Louro et al., 2009; Ponjavic et al., 2007], at least a portion of it is likely to be associated with functional nonprotein-coding RNA genes, many of which are located in regions previously regarded as intergenic and/or noncoding [ENCODE Project Consortium, 2007]. Noncoding RNA genes are as widespread as they are diverse [Borel et al., 2008], are transcribed from both strands of the genome, and may well exceed protein-coding genes in terms of their number [Fejes-Toth et al., 2009; Washietl et al., 2005]. Nonprotein-coding RNAs of known function include structural RNAs such as transfer RNAs, ribosomal RNAs, and small nuclear RNAs, but also putative regulatory RNAs (microRNAs, small interfering RNAs [siRNAs], Piwi-interacting RNAs, transcription initiation RNAs [tiRNAs], transcription start site-associated RNAs [TSSa-RNAs], promoter upstream transcripts [PROMPTs], promoter-associated sRNAs [PASRs and PALRs], and longer noncoding RNAs such as *XIST*), which are involved in the sequence-specific transcriptional and posttranscriptional modulation of gene expression [Collins and Penny, 2009; Kawaji and Hayashizaki, 2008; Mattick, 2009; Mercer et al., 2009; Preker et al., 2008; Seila et al., 2008; Taft et al., 2009]. Thus, more than 700 microRNA genes have already been identified in the human genome with many more probably awaiting discovery (miRBase; http://www.mirbase.org/cgi-bin/mirna_summary.pl?org=hsa) [Li et al., 2010b]. In total, at least 1,500 nonprotein-coding RNA genes have already been annotated in the human genome reference sequence with up to 5,000 more predicted by homology-based methods [Griffiths-Jones, 2007] (see Ensembl, database version 56.37a). Indeed, large intergenic noncoding RNAs (lincRNAs) have recently been found to represent a novel category of evolutionarily conserved RNAs, with a diverse array of functions ranging from embryonic stem cell pluripotency to cellular proliferation [Guttman et al., 2009; Khalil et al., 2009]; lincRNAs appear to number at least 3,000 in the human genome.

Pseudogenes

Whether processed or nonprocessed (duplicational), it has become clear that pseudogenes are almost as abundant as genes ("classical" or otherwise) in the human genome, with ~20% of known pseudogenes being transcribed [Harrison et al., 2005; Sakai et al., 2007; Zheng et al., 2007]. It should, however, be appreciated that although some pseudogenes may well be readily identifiable as lacking protein-coding potential by virtue of the interruption of their open-reading frames by premature stop codons or frameshift mutations, others will be less easily recognizable, especially if they are transcribed. The recent identification of short ($\leq$300 bp) human pseudogenes generated via the retrotransposition of mRNAs [Terai et al., 2010], however, suggests that pseudogenes may be even more common in the human genome than previously appreciated. Intriguingly, some of these pseudogenes are polymorphic in that they have functional as well as nonfunctional alleles segregating in the extant human population [Zhang et al., 2010].

With the realization that pseudogene-derived RNA transcripts may harbor functional elements [Khachane and Harrison, 2009; Zheng et al., 2007], the distinction between genes and pseudogenes has become somewhat blurred [Zheng and Gerstein, 2007]. Indeed, some "pseudogenes" appear to have a regulatory role [Hirotsune et al., 2003; Svensson et al., 2006], providing additional examples of the potential functional significance of noncoding RNAs. It is at present unclear what proportion of pseudogenes identified to date have either retained or acquired a function via their noncoding RNAs.

Transposable Elements

Transposable elements, including LINE-1, *Alu,* and SVA elements, make up ~40% of the human genome [Mills et al., 2007] and constitute a major source of interindividual structural variability [Xing et al., 2009]. Some of these transposable elements have contributed gene coding sequence to the human genome via "exonization" [Lin et al., 2009]. Other transposable elements have contributed functional noncoding sequence, for example, as regulatory elements [Jordan et al., 2003; Thornburg et al., 2006], microRNAs [Piriyapongsa et al., 2007] or naturally occurring antisense transcripts [Conley et al., 2008]. Many more are likely to have functional significance as suggested by their evolutionary conservation [Lowe et al., 2007; Nishihara et al., 2006].

Regulatory Noncoding Sequences

Extensive evolutionary conservation of noncoding DNA sequence is very evident in the human genome, because only

~40% of the evolutionarily constrained sequence occurs within protein-coding exons or their associated untranslated regions [ENCODE Project Consortium, 2007]. Studies of evolutionarily conserved noncoding sequences [Asthana et al., 2007; Drake et al., 2006; Parker et al., 2009; Ponting and Lunter, 2006] have suggested that 5–20% of the genome may be of functional importance rather than just the ~2% associated with the protein-coding portion [Eory et al., 2010; Pheasant and Mattick, 2007]. Some noncoding regions contain "ultraconserved elements" which not only exhibit enhancer function, but are also transcribed and often appear to have been subject to selection to the same extent as protein-coding regions [Katzman et al., 2007; Licastro et al., 2010; McLean and Bejerano, 2008]. Some noncoding regions contain CpG islands, far from the transcriptional initiation sites of genes, which may nevertheless have some regulatory significance [Medvedeva et al., 2010]. It should be appreciated, however, that the absence of evolutionary conservation does not necessarily denote lack of function. Indeed, human-specific functional elements have been shown to be present within rapidly evolving noncoding sequences [Bird et al., 2007; Prabhakar et al., 2006].

Toward a New Definition of the Gene

It is clear from the above that precisely what constitutes a gene has become somewhat contentious. The quite unanticipated scale of the extent of transcription in the genome, coupled with the widespread occurrence of overlapping genes and shared functional elements, hampers attempts to demarcate precisely and unambiguously where one gene ends and another one begins. As a consequence, the notion of the gene has become quite diffuse [Gerstein et al., 2007; Gingeras, 2007]. Indeed, as Kapranov et al. [2005] opined, "it is not unusual that a single base-pair can be part of an intricate network of multiple isoforms of overlapping sense and antisense transcripts, the majority of which are unannotated." Gene regulatory elements that are often quite distant from the genes they regulate [Kleinjan and Lettice, 2008], the existence of *trans*- as well as *cis*-regulatory elements [Morley et al., 2004], the formation of noncolinear transcripts through *trans*-splicing [Gingeras, 2009], taken together with the abundance of noncoding RNA genes [Zhang, 2008] and evolutionarily conserved noncoding regions [Drake et al., 2006; Ponting and Lunter, 2006], have combined to challenge the classical notion of the gene.

On the basis of the findings of the ENCODE project, Gerstein et al. [2007] proposed an updated definition of the gene as "a union of genomic sequences encoding a coherent set of potentially overlapping functional products." An alternative less heterodox definition of the gene as "a discrete genomic region whose transcription is regulated by one or more promoters and distal regulatory elements and which contains the information for the synthesis of functional proteins or noncoding RNAs, related by the sharing of a portion of genetic information at the level of the ultimate products (proteins or RNAs)" has been proposed by Pesole [2008]. Irrespective of its precise definition, it is clear that the concept of the gene is inadequate to the task of building a lexicon of those functional genomic sequences that could harbor mutations causing human inherited disease. It is indeed likely that, in the context of mutation detection, we shall eventually have to consider the universe of functional genetic elements in the human genome (the "functionome" see Fig. 1) as our hunting ground rather than simply genes per se.

How Many Inherited Disease Genes are There in the Human Genome?

The Concept of Gene Essentiality Lies at the Heart of any Discussion of Human Disease Genes

The question of how many inherited disease genes there are in the human genome should essentially be couched in terms of the proportion of human genes whose mutation would come to clinical attention in a nonnegligible proportion of cases by conferring a discernible clinical phenotype upon the individual concerned. As López-Bigas et al. [2006] have expressed it, "a gene is involved in a hereditary disease when its sequence has been subjected to a mutation that impairs its function or expression strongly enough to produce a certain pathological phenotype that is classified as a disease." However, although necessarily deleterious, such a mutation must not be lethal to the individual at an early stage of development because this would militate against its detection. Hence, disease genes are not, and cannot be, synonymous with "essential genes." Indeed, they exhibit very different properties [Feldman et al., 2008; Goh et al., 2007; López-Bigas et al., 2006]. The above notwithstanding, disease genes appear to be distinguishable from "nondisease genes" (in reality, the latter can only be defined as genes that are not yet known to cause inherited disease) in terms of a range of features including gene structure, gene expression, physicochemical properties, protein structure, and evolutionary conservation [Aerts et al., 2006; Cai et al., 2009; Domazet-Lošo and Tautz, 2008; Huang et al., 2004; Jimenez-Sanchez et al., 2001; Lage et al., 2008; López-Bigas and Ouzounis, 2004; Ng and Henikoff, 2006; Smith and Eyre-Walker, 2003; Subramanian and Kumar, 2006; Tu et al., 2006]. In this context, it should be appreciated that many disease genes will not have been identified as such simply because the underlying mutations have not yet appeared in the homozygous/compound heterozygous/hemizygous state required to manifest a clinical phenotype [Furney et al., 2006; Osada et al., 2009].

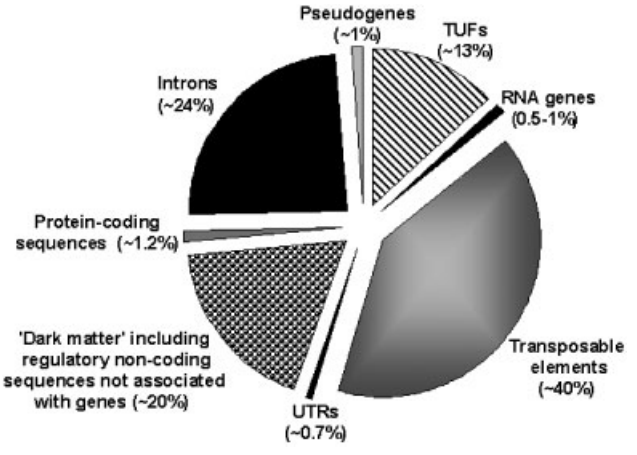

Figure 1. The "functionome": types of functional or potentially functional DNA sequences in the human genome that may harbor disease-causing mutations. Relative proportions of the human genome sequence are according to the International Human Genome Sequencing Consortium [2004] (protein-coding sequences, transposable elements, untranslated regions of genes [UTRs]); Ensembl GRCh37, Feb 2009 database version 57.37b (pseudogenes, RNA genes); Venter et al. [2001] (introns); Kopranov et al. [2002] (transcripts of unknown function (TUFs)); Pheasant and Mattick [2007], Evory et al. [2010] (regulatory noncoding sequences not associated with genes).

Although ~15% of mouse gene knockouts are developmentally lethal [Turgeon and Meloche, 2009] (Mouse Genome Informatics Resource (http://www.informatics.jax.org/), any definition of gene essentiality based exclusively on developmental lethality would be unnecessarily restrictive. Disease genes should therefore be understood in terms of a spectrum of gene "essentiality" that stretches from the truly essential genes on the one hand to almost dispensable genes on the other. Although essential genes have been quite reasonably defined as those genes that are "absolutely required for survival, or [which] strongly contribute to fitness and robust competitive growth" [Park et al., 2008], it should be appreciated that definitions of gene essentiality can differ quite widely between studies [Gerdes et al., 2006]. Using 2,789 disease genes from the HGMD gene set, Park et al. [2008] explored the likelihood of a gene being linked to human inherited disease in relation to its level of essentiality in mouse (4,004 genes) as adjudged by the results of gene disruption and knock-out experiments. Twice as many genes with abnormal effects when disrupted in mouse (1,311/3,635; 36%) were found to have a human disease gene homologue than genes that displayed no overt phenotypic abnormality when disrupted (63/369; 17%). Somewhat surprisingly, when the genes with abnormal effects in mouse were subdivided into genes with lethal effects and nonlethal effects, the frequencies of disease gene homologues among them were comparable (38% [728/1,904] and 34% [583/1,731], respectively). However, when Park et al. [2008] further subdivided the genes with lethal effects in mouse, they found human disease gene homologues to be 1.4 times more frequent among genes categorized as being "postnatal lethal" in the mouse than among "embryonic lethal" genes. Thus, almost counterintuitively at first glance, the more essential murine genes (which are embryonic lethal in mouse) appear to be less likely to be disease genes in human. This finding confirms the above mentioned dictum that disease genes are not, and cannot be, synonymous with "essential genes." Interestingly, Park et al. [2008] also observed that the type of disease mutation in the human homologue varies depending upon the essentiality of the mouse gene involved, with nonsense mutations, splicing mutations, microinsertions/microdeletions, and gross insertions/deletions being disproportionately associated with the mouse genes displaying abnormal effects when disrupted. We may also speculate that although a mild mutation in an "essential" gene may be sufficient to cause disease, a much more severe mutation may be necessary in a "dispensable" gene. Clearly, concepts of gene essentiality and dispensability are likely to be context-dependent.

Although ~91% of the murine genes employed in the study described above were deemed to belong to the "essential" category (i.e., the group of genes that display abnormal phenotypic effects when mutated), we should be wary of making direct inferences in the human context. This is not only because those mouse genes with a known mutational phenotype comprise fewer than 20% of the total number of genes in this organism, but also because it may be somewhat hazardous to extrapolate to the human genome where both gene function [Liao and Zhang, 2008] and copy number [Cutler and Kassner, 2008] may differ quite markedly from the mouse.

Inspection of the entry record history of HGMD [Stenson et al., 2009] reveals a constant increase in the rate at which newly reported disease genes have been entered into HGMD every year, with 293 recorded for 2009 (Fig. 2). Because this increase has to cease at some point in time, simply because the number of human genes is finite, we ventured to fit the various four- or five-parameter saturation models provided by SigmaPlot v.8.02 (SPSS Inc., 2002) to the annual cumulative gene number in HGMD since 1978. The results of these admittedly highly speculative projections (which nevertheless yielded an $R^2 > 0.9999$ for all models) point to a total number of inherited disease genes of between 7,750 (five-parameter Weibull model) and 30,750 (five-parameter Chapman model). The remaining four models (sigmoid, logistic, Gompertz and Hill) yielded estimates in a very narrow range of between 11,850 and 15,100 inherited disease genes, and the average taken over all six models equalled 15,300, that is, 46% of the 33,000 genes currently estimated to be present in the human genome (HuRef NCBI build 37.1).

Concepts of Human Gene Essentiality and Dispensability Are Necessarily Bound Up with Gene Copy Number

The loss of a particular gene/allele is not invariably associated with a readily discernible clinical phenotype [Waalen and Beutler, 2009]. This assertion is supported by the identification of more than 1,000 putative nonsense single nucleotide polymorphisms (SNPs) (i.e., nonsense mutations that have attained polymorphic frequencies) in human populations [Ng et al., 2008; Yngvadottir et al., 2009b]. About half of these nonsense SNPs have been validated by dbSNP (http://www.ncbi.nlm.nih.gov/projects/SNP), a process that involves the exclusion of mutations in pseudogenes and of artefacts caused by sequencing errors. Bona fide nonsense SNPs are expected either to lead to the synthesis of a truncated protein product or alternatively

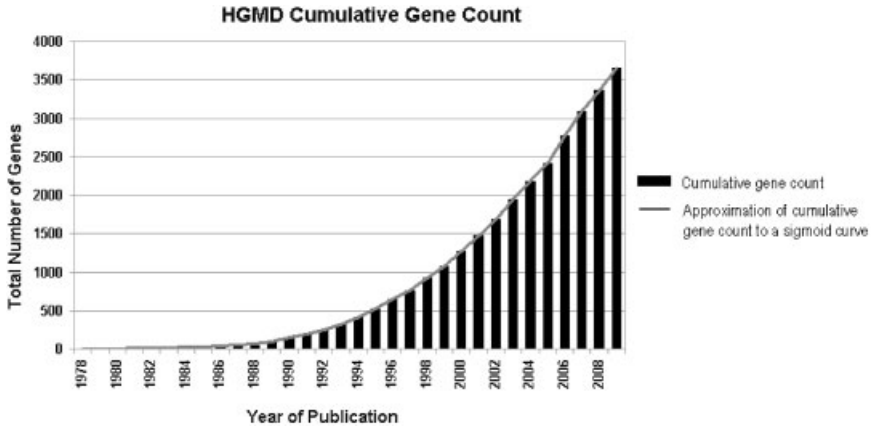

Figure 2. Annual cumulative gene count in the Human Gene Mutation Database (HGMD). Shown is the cumulative number of different human "disease genes" present in the HGMD. The line represents an approximation to a sigmoid curve.

to the greatly reduced synthesis of the truncated protein product (if the mRNA bearing them is subject to nonsense-mediated mRNA decay [NMD]). Based upon the relative locations of the nonsense SNPs and the exon–intron structures of the affected genes, Yamaguchi-Kabata et al. [2008] concluded that 49% of nonsense SNPs would be predicted to elicit NMD, whereas 51% would be predicted to yield truncated proteins. Some of these nonsense SNPs have been found to occur in the homozygous state in normal populations [Yngvadottir et al., 2009b], attesting to the likely functional redundancy of the corresponding genes. At the very least, genes harboring nonsense SNPs may be assumed to be only under weak selection [Ng et al., 2008].

It should be appreciated that nonsense SNPs may even occur in "essential" genes yet still fail to come to clinical attention (or give rise to a detectable phenotype) if these genes are subject to copy number variation (see *CNVs and copy number mutations* below) that masks any deleterious consequences by ensuring an adequate level of gene expression from additional wild-type copies either in *cis* or in *trans*. Thus, copy number variation might serve to "rescue" the full or partial loss of gene function brought about by the nonsense mutations, thereby accounting for the occurrence of the latter at polymorphic frequencies. Consistent with this postulate, Ng et al. [2008] reported that ~30% of nonsense SNPs occur in genes residing within segmental duplications, a proportion some threefold larger than that noted for synonymous SNPs. Genes harboring nonsense SNPs were also found to belong to gene families of higher than average size [Ng et al., 2008], suggesting that some functional redundancy may exist between paralogous human genes. In support of this idea, Hsiao and Vitkup [2008] reported that those human genes that have a homologue with $\geq$90% sequence similarity are approximately three times less likely to harbor disease-causing mutations than genes with less closely related homologues. They interpreted their findings in terms of "genetic robustness" against null mutations, with the duplicated sequences providing "back-up" by potentiating the functional compensation/complementation of homologous genes in the event that they acquire deleterious mutations. Potential examples of such functional redundancy in the human genome involve the genes for *CCL4* and *CCL4L1* chemokines [Howard et al., 2004] and the Rab GTPase genes, *RAB27A* and *RAB27B* [Barral et al., 2002]. In the mouse, the proportion of essential genes among gene duplicates is ~7% lower than among singletons, implying that ~15% of single gene deletions that would otherwise be lethal (or infertile) are actually viable (or fertile) as a consequence of functional compensation by the duplicate gene copy [Liang and Li, 2009]. This level of functional redundancy may be even more pronounced for the most recently duplicated genes [Su and Gu, 2008].

What Proportion of the Possible Mutations Within Inherited Disease Genes is Likely to be of Pathological Significance?

Human gene mutation is a highly sequence-specific process, irrespective of the type of lesion involved. This has had important implications, not only for the nature and prevalence, but also for the diagnosis of human genetic disease [Antonarakis and Cooper, 2007; Mefford and Eichler, 2009; Zhang et al., 2009]. Certain DNA sequences have been found to be hypermutable, thereby providing important clues as to the nature of the endogenous mechanisms underlying different types of human gene lesion, but also emphasizing the nonuniform nature of mutagenesis [Antonarakis and Cooper, 2007]. Of course, human gene mutations also lack a uniform distribution within genes for functional reasons that are bound up with the nature of the gene product in question [Miller et al., 2003; Subramanian and Kumar, 2006].

The vast majority of mutations listed in HGMD reside within the coding region (86%), the remainder being located in either intronic (11%) or regulatory (3%, promoter, untranslated or flanking regions) sequences. The question of the proportion of possible mutations within human disease genes that are likely to be of pathological significance is very difficult to address because it is dependent not only upon the type and location of the mutation but also upon the functionality of the nucleotides involved (itself dependent in part upon the amino acid residues that they encode) which is often hard to assess [Arbiza et al., 2006; Capriotti et al., 2008; Ferrer-Costa et al., 2002; Kumar et al., 2009; Li et al., 2009a; Miller and Kumar, 2001]. In addition, some types of mutation are likely to be much more comprehensively ascertained than others, making observational comparisons between mutation types an inherently hazardous undertaking.

Recently, it has been demonstrated that multiple mutations may not be an infrequent cause of human genetic disease [Chen et al., 2009b]. Such multiple mutations may constitute the signatures of transient hypermutability in human genes. This has raised serious concerns regarding current practices in mutation screening, practices that are likely to have resulted in either the neglect, or even the complete failure to detect, many potentially important secondary mutations linked in *cis* to the putative primary pathological lesion. Because interactions may well occur between genetic variants linked in *cis*, inadequacies in the current practice of mutation screening could easily have contributed to the frequently observed inconsistencies in the genotype–phenotype relationship.

The above notwithstanding, the question of the proportion of all possible mutations within human disease genes that are likely to be of pathological significance is clearly of paramount importance to medical diagnostics. However, the corollary to this question is the issue of whether some mutations may have been overlooked in mutation screening programs because they are located at some very considerable distance from the genes whose function they disrupt. These related questions will be addressed in some detail below.

Coding Sequence Mutations

The first study to attempt to partition human amino acid substitutions with respect to their phenotypic consequences was that of Fay et al. [2001], which was based on common ($f \geq 0.15$) polymorphism and sequence divergence data from human genes. These workers estimated that 60% of missense mutations were deleterious, 20% were slightly deleterious, and 20% were neutral. More recently, from a combined analysis of disease-causing mutations logged in HGMD, mutations driving human–chimpanzee sequence divergence, and systematic data on neutral human genetic variation, Kryukov et al. [2007] concluded that ~20% of new missense mutations in humans result in a loss of function, whereas ~53% have mildly deleterious effects and ~27% are effectively neutral with respect to phenotype. Their estimates have received independent support, at least qualitatively, from a study of human coding SNPs by Boyko et al. [2008], who predicted that 27–29% of missense mutations would be neutral or near neutral, 30–42% would be moderately deleterious, with most of the rest (i.e., 29–43%) being highly deleterious or lethal.

Although it has been estimated that only ~1.6% of disease-causing missense substitutions in human genes also affect mRNA splicing [Krawczak et al., 2007], the actual proportion of exonic missense mutations that disrupt splicing, and which are therefore of pathological significance, may be substantially higher [López-Bigas et al., 2005; Sanford et al., 2009].

In addition, one must be aware in this context that synonymous ("silent") mutations, although not altering the amino acid sequence of the encoded protein directly, can still influence splicing accuracy or efficiency [Cartegni et al., 2002; Gorlov et al., 2006; Hunt et al., 2009; Sanford et al., 2009; Wang and Cooper, 2007]. Recently, it has been realized that apparently silent SNPs may also become distinctly "audible" in the context of mRNA stability or even protein structure and function. Thus, three common haplotypes of the human catechol-O-methyltransferase gene (*COMT*; MIM# 116790), which differ in terms of two synonymous and one nonsynonymous substitution, confer differences in COMT enzymatic activity and pain sensitivity [Nackley et al., 2006, 2009]. The major *COMT* haplotypes differed with respect to the stability of the *COMT* mRNA local stem–loop structures, the most stable being associated with the lowest levels of COMT protein and enzymatic activity [Nackley et al., 2006]. In a similar vein, synonymous SNPs in the ATP-binding cassette, subfamily B, member 1 gene (*ABCB1*; MIM# 171050) have been shown to alter ABCB1 protein structure and activity [Kimchi-Sarfaty et al., 2007], possibly by changing the timing of protein folding following extended ribosomal pause times at rare codons [Tsai et al., 2008].

Finally, it should be understood that whereas the deleteriousness of the average synonymous mutation is always likely to be less than that of that of a nonsynonymous (missense) mutation [Boyko et al., 2008], the higher prevalence of synonymous mutations means that they may actually make a significantly greater contribution to the phenotype than nonsynonymous mutations [Goode et al., 2010].

Mutations and Functional Polymorphisms

With the realization that a sizeable proportion of gene-associated polymorphisms serve to alter the structure, function, or expression of their host genes, drawing a sharp distinction between functional polymorphisms, disease-associated polymorphisms and pathological mutations has become increasingly difficult. In practical terms, such a distinction is generally made in the context of the prevalence of the variant in the population under study as well as its penetrance (i.e., the probability with which a specific genotype manifests itself as a given clinical phenotype). Variants with a minor allele frequency of $\geq 1\%$ in the population of interest are, by convention, termed polymorphisms, and an increasing number have been found to play a role in complex disease [Frazer et al., 2009]. Currently, over 5,000 disease-associated or functional polymorphisms have been reported in a total of over 1,800 different human genes (see HGMD). This number is predicted to increase quite dramatically over the coming years (as promoter regions, untranslated regions, and introns are more and more systematically screened for such variants), although distinguishing them from neutral polymorphisms is unlikely to be a trivial undertaking [Li et al., 2009a; Mort et al., 2010].

Although the vast majority (90%) of disease-associated or functional polymorphic variants listed in HGMD are SNPs, a sizeable number are of the insertion/deletion type. Disease-associated or functional polymorphic variants are generally located in either gene regulatory (~25%) or gene coding regions (~60%), although it should be noted that variants occurring outside of these regions may still have consequences for gene expression, splicing, transcription factor binding, etc. In addition, some functionally important SNPs are associated with nonprotein-coding genes [Borel and Antonarakis, 2008; Yang et al., 2008a].

At present, ~55% of the polymorphic variants recorded in HGMD are "disease associated." However, even in cases where no disease association has yet been demonstrated, functional polymorphisms that alter the expression of a gene or the structure/function of the gene product are potentially very important. Although such a polymorphism may not appear to have any direct and/or immediate clinical relevance, the respective data in HGMD could yet prove very valuable in terms of understanding interindividual differences in disease susceptibility.

Intronic Mutations

Mutations that occur within the extended consensus sequences of exon–intron splice junctions account for ~10% of all reported mutations logged in HGMD and are frequently encountered in mutation screening studies [Krawczak et al., 2007]. However, mutations residing in other intronic locations (including the canonical branchpoint sequence) may often go undetected unless they induce aberrant splicing (e.g., exon skipping or cryptic splice-site utilization) that is readily distinguishable qualitatively or quantitatively from both normal and alternative splicing [Coulombe-Huntington et al., 2009]. Introns probably represent substantially larger targets for functional mutations than has hitherto been recognized [Lynch, 2010] on account of their harboring a multiplicity of functional elements including intron splice enhancers and silencers, *cis*-acting RNA elements that regulate alternative splicing [Tress et al., 2007; Wang et al., 2009a], and potentially also *trans*-splicing elements [Akiva et al., 2006; Gingeras 2009; Shao et al., 2006], as well as other regulatory elements some of which may be deeply embedded within very large introns [Solis et al., 2008]. In terms of identifying intronic functional elements, it may be helpful that they are often characterized by a reduced level of genetic variation [Lomelin et al., 2010].

Deep intronic mutations generally appear to comprise less than 1% of known splicing mutations (Table 1), but this figure is very likely to be an underestimate owing to the inherent difficulty in detecting splicing mutations located outside of (and distant from) exon–intron splice junctions. Thus, for example, when the *NF1* gene (MIM# 162200) was methodically screened for mutations that altered splicing, ~5% of the identified lesions that altered splicing were deep intronic mutations [Pros et al., 2008]. Among disease-causing lesions, inclusion of a pseudoexon as a consequence of cryptic splice-site activation appears to be the most common consequence of deep intronic mutation [Dhir and Buratti, 2010]. If we also consider the deep intronic polymorphic variants that have the potential to confer susceptibility to disease [Choi et al., 2008; Emison et al., 2005; Fraser and Xie, 2009; Grant et al., 1996; Mann et al., 2001; Susa et al., 2008], it is very likely that splicing-relevant intronic variation will have been seriously underascertained thus far. Consistent with this statement, Goode et al. [2010] have recently reported that the vast majority of putatively functional variants in the human genome actually reside in either intronic or intergenic locations.

Table 1. Selected Examples of Deep Intronic Mutations Identified as Causing Human Inherited Disease[a]

Gene (MIM#)	Disease	Chromosomal location	Mutation (HGVS Nomenclature)	Mutation (traditional name)	Consequences for mRNA splicing	Reference
ATM (607585)	Ataxia telangiectasia	11q22–q23	NM_000051.3:c.3994−159A>G	IVS28−159A>G	Weakens 5′ splice site of alternative exon 28a and activates 5′ cryptic splice site 83 nt downstream	Coutinho et al. [2005]
CDKN2A (600160)	Melanoma, predisposition	9p21	NM_000077.3:c.458−105A>G	IVS2−105A>G	Activates cryptic splice site 105 nt 5′ to exon 3 resulting in aberrant splicing	Harland et al. [2001]
DMD (300377)	Dystrophinopathy, asymptomatic	Xp21.2	NM_004006.2:c.93+5591T>A	IVS2+5591T>A	Activates two 5′ cryptic splice sites 132 nt or 46 nt downstream	Yagi et al. [2003]
FGB (134830)	Afibrinogenemia	4q28	NM_005141.3:c.115−600A>G	IVS1−600A>G	Creates consensus sequence for splicing factor SF2/ASF leading to inclusion of cryptic exon	Dear et al. [2006]
FGG (134850)	Afibrinogenemia	4q28	NM_000509.4:c.667−320A>T	IVS6−320A>T	Activates cryptic splice leading to inclusion of cryptic exon carrying a premature Stop codon	Spena et al. [2007]
HADHB (143450)	Mitochondrial trifunctionalprotein deficiency	2p23	NM_000183.2:c.442+614A>G	IVS7+614A>G	Activates cryptic splice leading to inclusion of two cryptic exons	Purevsuren et al. [2008]
MTRR (602568)	Homocystinuria	5p15.31	NM_002454.2:c.903+469T>C	IVS6+469T>C	Creates an SF2/ASF-binding exon splice enhancer which leads to pseudoexon activation	Homolova et al. [2010]
MUT (609058)	Methylmalonic aciduria	6p12.3	NM_000255.2:C1957−892C>A	IVS11−892C>A	Activates cryptic splice site leading to the inclusion of pseudoexon	Rincón et al. [2007]
NF1 (162200)	Neurofibromatosis type 1	17q11.2	NM_000267.3:c.288+2025T>G	IVS3+2025T>G	Activates cryptic splice site leading to inclusion of a cryptic exon	Pros et al. [2008]
OTC (300461)	Ornithine transcarbamylase deficiency	Xp21.1	NM_000531.5:c.540+265G>A	IVS5+265G>A	Activates cryptic splice site leading to inclusion of cryptic exon	Ogino et al. [2007]
PCCA (232000)	Propionic acidaemia	13q32	NM_000282.2:c.1285−1416A>G	IVS14−1416A>G	Activates cryptic splice site leading to the inclusion of pseudoexon	Rincón et al. [2007]
PCCB (232050)	Propionic acidaemia	3q21–q22	NM_000532.3:c.654+462A>G	IVS6+462A>G	Activates cryptic splice site leading to the inclusion of pseudoexon	Rincón et al. [2007]
PMM2 (601785)	Congenital disorder of glycosylation type Ia	16p13	NM_000303.2:c.640−15479C>T	IVS7−15479C>T	Activates cryptic splice site leading to the inclusion of pseudoexons	Schollen et al. [2007]
PRPF31 (606419)	Retinitis pigmentosa, autosomal dominant	19q13.4	NM_015629.3:c.1374+654C>G	IVS13+654C>G	Activates cryptic splice site leading to the inclusion of pseudoexons	Rio Frio et al. [2009]
RB1 (180200)	Retinoblastoma	13q14.2	NM_000321.2:c.2490−1398A>G	IVS23−1398A>G	Activates cryptic splice site leading to inclusion of cryptic exon	Dehainault et al. [2007]
SLC12A3 (600968)	Gitelman syndrome	16q13	NM_000339.2:c.1670−191C>T	IVS13−191C>T	Activates cryptic splice site leading to inclusion of cryptic exon	Nozu et al. [2009]

IVS, intron number.
[a]Located within an intron at least 100 bp from the nearest splice site.

Mutations Residing within Proximal Gene Regulatory Regions

Microlesions within proximal gene regulatory regions currently comprise only ~1.7% of known mutations causing or associated with human inherited disease (see HGMD). Their relative rarity may be in part because not all regulatory elements occur immediately 5′ to the genes that they regulate. Indeed, many such elements are located within the first exon, within introns [Cecchini et al., 2009] or within 5′ or 3′ untranslated regions (UTRs); regulatory elements within intronic sequences and UTRs are less likely to be screened for mutations [Chatterjee and Pal, 2009; Chen et al., 2006a,b], particularly if the UTRs are large (e.g., *MECP2*; MIM# 300005) [Coutinho et al., 2007]. Recently, a whole new category of pathological mutation has been recognized within the 3′ UTRs of human genes: genetic variants located in microRNA target sites either causing, or associated with an increased risk of, inherited disease [Bandiera et al., 2010; Martin et al., 2007; Rademakers et al., 2008; Sethupathy and Collins, 2008; Simon et al., 2010].

In the same vein, upstream open reading frames (uORFs), present in ~50% of human genes, often impact upon the expression of the primary ORFs; indeed, both mutations and polymorphisms have been reported within uORFs that can modulate or even abolish the expression of the downstream gene [Calvo et al., 2009; Wen et al., 2009].

Microsatellites located within introns, promoter, or flanking regions often play a regulatory role in the expression of human genes [Li et al., 2004], and some of them have been highly conserved over evolutionary time [Buschiazzo and Gemmell, 2010]. It therefore comes as no surprise to find that genetic variants in such sequences are increasingly being found to cause, or to be associated with, human inherited disease [Brouwer et al., 2009; Wilkins et al., 2009].

Mutations Residing within Remote Gene Regulatory Regions

One reason for the relative paucity of regulatory mutations is that our knowledge of transcriptional regulatory elements (i.e., core promoters, proximal promoters, distal enhancers, repressors/silencers, insulators/boundary elements, and locus control regions), is still fairly rudimentary, particularly in the case of remote regulatory elements that act at a distance [Attanasio et al., 2008; Heintzman and Ren, 2009; Kleinjan and Coutinho, 2009; Maston et al., 2006; Pennacchio et al., 2006; Visel et al., 2009; Zhang et al., 2007] so that the appropriate elements are often simply not recognized, let alone screened for mutation. It is therefore scarcely surprising that the number of known regulatory mutations decays quite rapidly with distance from the gene, mutations within remote regulatory elements being few and far between. Table 2 lists known microlesions that occur >10 kb 5′ upstream of human genes causing inherited disease. These include a total of nine mutations within a 1-kb region (termed the long-range or limb-specific enhancer, ZRS) ~979 kb 5′ to the transcription initiation site of the sonic hedgehog gene (*SHH*; MIM# 600725) [Gordon et al., 2009].

Far upstream polymorphic variants that influence gene expression and that are relevant to disease are also beginning to be documented. Thus, for example, the C>T functional SNP 14.5 kb upstream of the interferon regulatory factor 6 gene (*IRF6*; MIM# 607199), associated with cleft palate, alters the binding of transcription factor AP-2α [Rahimov et al., 2008]. Similarly, a functional SNP ~6 kb upstream of the α-globin-like *HBM* gene (MIM# 609639) serves to create a binding site for the erythroid-specific transcription factor GATA1 and interferes with the activation of the downstream α-globin genes [De Gobbi et al., 2006]. A functional SNP ~335 kb upstream of the *MYC* gene (MIM# 190080) increases the risk of colorectal and prostate cancer by increasing the expression of the *MYC* gene by altering the binding strength of transcription factors TCF4 and/or TCF7L2 to a transcriptional enhancer [Haiman et al., 2007; Pomerantz et al., 2009; Tuupanen et al., 2009; Wright et al., 2010]. Finally, in the context of pointing out the shortcomings of the gene-centric approach to mutation detection, we should be aware that functional SNP rs4988235, located 13.9 kb upstream of the lactase gene (*LCT*; MIM# 603202) and associated with adult-type hypolactasia, actually resides deep within intron 13 (c.1917+326C>T) of the minichromosome maintenance complex component 6 gene (*MCM6*; MIM# 601806) [Enattah et al., 2002; Lewinsky et al., 2005; Olds and Sibley, 2003]. Given that up to 5%

Table 2. Examples of Regulatory Mutations, Located far Upstream of Gene Sequences, Known to Cause Human Inherited Disease

Gene (MIM#)	Disease	Chromosomal location	Mutation (chromosomal coordinate)[c]	Mutation (relative location)[a]	Reference
SOX9 (608160)	Cleft palate, Pierre Robin sequence	17q24.3–q25.1	Chr17:66187898T>C	−1440858T>C	Benko et al. [2009]
SHH (600725)	Triphalangeal thumb-polysyndactyly syndrome	7q36.3	Chr7:156277624C>T	−979896C>T	Wang et al. [2007]
SHH (600725)	Preaxial polydactyly	7q36.3	Chr7:156277226C>G	−979498C>G	Lettice et al. [2003]
SHH (600725)	Triphalangeal thumb	7q36.3	Chr7:156277036T>C	−979308T>C	Furniss et al. [2008]
SHH (600725)	Preaxial polydactyly	7q36.3	Chr7:156277026A>T	−979298A>T	Lettice et al. [2003]
SHH (600725)	Preaxial polydactyly	7q36.3	Chr7:156277002T>C	−979274T>C	Lettice et al. [2003]
SHH (600725)	Preaxial polydactyly	7q36.3	Chr7:156276927G>A	−979199G>A	Lettice et al. [2003]
SHH (600725)	Werner mesomelic syndrome	7q36.3	Chr7:156276927G>C	−979199G>C	Wieczorek et al. [2010]
SHH (600725)	Triphalangeal thumb-polysyndactyly syndrome	7q36.3	Chr7:156276710C>G	−978982C>G	Gurnett et al. [2007]
SHH (600725)	Triphalangeal thumb-polysyndactylysyndrome	7q36.3	Chr7:156276592A>G	−978864A>G	Gurnett et al. [2007]
POU6F2 (609062)	Wilms' tumour	7p14.1	Chr7:38984140C>G	−28793C>G	Perotti et al. [2004]
EPHX1 (132810)	Hypercholanemia	1q42.12	Chr1:224075359T>A	−4240T>A[b]	Zhu et al. [2003]
PSEN1 (104311)	Alzheimer disease, early onset	14q24.2	Chr14:72670114A>G	−2818A>G	Theuns et al. [2000]

[a]Location given is relative to the transcriptional initiation site of the specified gene. Only mutations >2 kb 5′ to the transcriptional initiation site of the associated gene are listed.
[b]Mutation is located in a recognition site for hepatocyte nuclear factor 3 (HNF-3).
[c]Chromosomal coordinates were obtained from NCBI build 36.3.

of quantitative trait loci for gene expression (eQTLs) lie >20 kb upstream of transcriptional initiation sites [Veyrieras et al., 2008], many more far upstream polymorphic variants that influence gene expression are likely to be identified in the coming years.

Rather fewer pathological mutations are known to be located at a considerable distance downstream of human genes. One example is the C>G transversion 2528 nt 3′ to the Term codon of the *CDK5R1* gene (MIM# 603460), which has been postulated to play a role in nonspecific mental retardation [Venturin et al., 2006]. Perhaps more dramatic is the A>G SNP (rs2943641), 565981 bp 3′ to the Term codon of the *IRS1* gene (MIM# 147545), which is associated with type 2 diabetes, insulin resistance, and hyperinsulinemia; the G allele was found to be associated with a reduced basal level of IRS1 protein [Rung et al., 2009].

Remote regulatory elements have sometimes come to attention as a consequence of their removal by gross deletions located at some considerable distance (from 10 kb to several megabases) from the genes whose expression they disrupt (Table 3). Thus, for example, a 960-kb deletion of noncoding sequence, lying between 1.477 Mb and 517 kb upstream of the *SOX9* gene (MIM# 608160) gives rise to the acampomelic form of campomelic dysplasia [Lecointre et al., 2009]. Such pathological deletions, however, are not necessarily always so large. Indeed, a 7.4 kb deletion, located 283 kb upstream of the *FOXL2* gene (MIM# 605597) has been identified as a cause of blepharophimosis syndrome; it disrupts a long noncoding RNA (*PISRT1*) as well as eight conserved noncoding sequences [D'haene et al., 2009] (Table 3). For some conditions, such lesions may actually occur quite frequently, as in the case of the *SHOX* gene (MIM# 312865) where ~22% of Leri-Weill syndrome patients (MIM# 127300) and ~1% of individuals with idiopathic short stature (MIM# 300582) harbor a microdeletion spanning the upstream enhancer region that leaves the coding region of the *SHOX* gene intact [Chen et al., 2009a].

In this context, it is interesting to note that developmental genes appear to be disproportionately represented among those human genes located within "gene deserts" (i.e., those chromosomal regions that are devoid of annotated genes) [Ovcharenko et al., 2005; Taylor, 2005] and are often separated from their regulatory elements by up to several hundred kilobases.

The remote regulatory elements of several such genes (viz. *BMP2, PAX6, SHH, SHOX,* and *SOX9*) are known to be subject to deletion or gross rearrangement resulting in inherited disease (Table 3).

Given that the number of transcriptional initiation sites in the human genome is much greater than the number of genes [Carninci et al., 2006], it may well be that the number of regulatory sequences associated with human genes has been seriously underestimated. Further, both *cis*- and *trans*-acting variation within regulatory regions may serve to modify gene expression and/or the functional effects of protein coding variants [Dimas et al., 2008, 2009; Kasowski et al., 2010; Pastinen et al., 2006; Stranger et al., 2005, 2007]. The underascertainment of disease-associated mutations within regulatory regions is therefore likely to be quite substantial but can potentially be rectified by emerging high-throughput entire genome sequencing protocols [Chorley et al., 2008].

CNVs and Copy Number Mutations

No mention of the human germline mutational spectrum would be complete without making reference to copy number variants (CNVs). CNVs are a recently discovered form of genomic diversity involving DNA sequences ≥1 kb in length that are present in the human genome in a variable number of copies [Iafrate et al., 2004; Redon et al., 2006; Sebat et al., 2004]. Such gross duplications/deletions are not only rather abundant but also often occur at polymorphic frequencies. The Database of Genomic Variants (http://projects.tcag.ca/variation; August 2009) currently lists 8,410 CNV loci (CNV loci represent genomic regions that harbor CNVs) and their number is increasing steadily, fuelled by refined analytical methods and the ongoing characterization of this type of genomic variation in different human populations [Kidd et al., 2008]. Conrad et al. [2010] have generated a comprehensive map of >8,500 validated CNVs >500 bp (detected in 41 Europeans/West Africans) that together cover a total of 112.7 Mb (3.7% of the genome). These authors estimated that 39% of the validated CNVs overlapped 13% of RefSeq genes (NCBI mRNA reference sequence collection). Further, they concluded that the CNVs detected resulted in the "unambiguous loss of function" of alleles for 267 different genes.

It is important to note that the mutation rate (per locus and per generation) is considerably higher for CNVs (3×10^{-2} to 10^{-7}) than for SNPs (10^{-7} to 10^{-8}) [Conrad et al., 2010; Redon et al., 2006], no doubt due to the very different mutational mechanisms involved. In their very comprehensive treatment of this issue, Conrad et al. [2010] attempted to estimate the average per-generation rate of CNV formation. However, rate estimates were found to vary by several orders of magnitude between sites. Conrad et al. [2010] further noted that these estimates did not allow for purifying selection, and so they probably represent "a lower bound on the true rate." There is also an ascertainment bias to contend with, duplications being significantly harder to identify than deletions [Quemener et al., 2010].

It has been estimated that on average, 73 to 87 genes vary in copy number between any two individuals [Alkan et al., 2009]. This high degree of interindividual variability with regard to gene copy number has challenged traditional definitions of wild-type and "normality," and even the very concept of a "reference genome" itself [Dear, 2009]. High resolution breakpoint mapping is a prerequisite for the accurate assessment of CNV size, the identification of the genes and regulatory elements affected, and hence, for the determination of the consequences of copy number variation for gene expression and the phenotypic *sequelae* [Beckmann et al., 2008; de Smith et al., 2008]. This notwithstanding, it is already becoming clear that these consequences may go far beyond the physical bounds of a given CNV. Thus, a CNV involving the human *HBA* gene (MIM# 141800) has a dramatic influence on the expression of the *NME4* gene (MIM# 601818) some 300 kb distant [Lower et al., 2009]. In addition, a 5.5 kb microduplication of a conserved noncoding sequence with demonstrated enhancer function, about 110 kb downstream of the bone morphogenic protein 2 gene (*BMP2*; MIM# 112261), has been found to cause brachydactyly type 2A in two families [Dathe et al., 2009].

It may well be that the precise extent and/or location of many CNVs will vary between individuals, thereby further increasing both the mutational and phenotypic heterogeneity. The extent to which CNVs are likely to contribute to the diversity of human phenotypes, including "single gene defects," genomic disorders, and complex disease, is increasingly being recognized. Indeed, CNVs are now being widely recruited to genome-wide association studies with the aim of assessing their influence on human disease causation/susceptibility [Beckmann et al., 2008; McCarroll, 2008; Merikangas et al., 2009]. To date, 37 human disease conditions have been identified, which are either caused by CNVs or whose relative risk is increased by CNVs [Beckmann et al., 2008; Lee and Scherer, 2010, and references therein]. Remarkably, an excess of both rare and de novo CNVs has been identified in patients with psychiatric disorders and obesity [Bochukova et al., 2010; Elia

Table 3. Examples of Genomic Deletions and Other Rearrangements Causing Human Inherited Disease but Located at Some Considerable Distance from the Genes Whose Function they Disrupt

Gene(MIM#)	Disease	Chromosomal location	Genomic rearrangement	Location (5′ or 3′) relative to gene	Reference
BMP2 (112261)	Autosomal dominant brachydactyly type A2	20p12.3	Duplication (5.5 kb)	~110 kb 3′ to gene	Dathe et al. [2009]
DLX6 (600030)	Hearing loss and craniofacial defects	7q21.3	Inversion breakpoint	~65 kb 5′ to gene	Brown et al. [2010]
FOXC2 (602402)	Lymphedema-distichiasis syndrome	16q24.1	Translocation breakpoint	120 kb 3′ to gene	Fang et al. [2000]
FOXF1 (601089)	Alveolar capillary dysplasia	16q24.1	Deletions (524 kb, 145 kb)	52 kb and 259 kb 5′ to gene	Stankiewicz et al. [2009]
FOXL2 (605597)	Blepharophimosis syndrome	3q22.3	Deletion (7.4 kb)	283 kb 5′ to gene	D'haene et al. [2009]
GJB2 (121011)	Nonsyndromic sensorineural hearingloss	13q12.11	Deletion (131.4 kb)	>100 kb 5′ to gene	Wilch et al. [2010]
HBA2 (141850)	α-thalassaemia	16p13.3	Deletions, various	>20 kb 5′ to gene	Hatton et al. [1990] Romao et al. [1991] Viprakasit et al. [2003]
HBB (141900)	β-thalassaemia	11p15.5	Deletions, various	>50 kb 5′ to gene	Driscoll et al. [1989] Harteveld et al. [2005] Koenig et al. [2009]
PAX6 (607108)	Aniridia	11p13	Deletions (975 kb, 1105 kb)	11.6 kb and 22.1 kb 3′ to gene	Lauderdale et al. [2000]
PITX2 (601542)	Rieger syndrome	4q25	Translocation breakpoint	~90 kb 5′ to gene	Flomen et al. [1998]
POU3F4 (300039)	X-linked deafness type 3 (DFN3)	Xq21.1	Deletions, various	~900 kb 5′ to gene	de Kok et al. [1996]
SHH (600725)	Preaxial polydactyly	7q36	De novo reciprocal t(5,7) (q11,q36) translocation breakpoint	~1 Mb 5′ to gene	Lettice et al. [2002]
SOX9 (608160)	Acampomelic campomelic dysplasia	17q24.3	Deletion (960 kb) De novo balanced complex chromosomal rearrangement with a 17q breakpoint. Balanced translocation, t(4;17)(q28.3;q24.3) breakpoint	1.477 Mb and 517 kb 5′ to gene ~1.3 Mb 3' to gene ~900 kb 5′ to gene	Lecointre et al. [2009] Velagaleti et al. [2005] Velagaleti et al. [2005]
SHOX (312865)	Leri-Weill dyschondrosteosis	Xp22.33	Deletions, various	30–250 kb 3′ to gene	Benito-Sanz et al. [2005]
TRPS1 (604386)	Ambras syndrome	8q23.3	Inversion breakpoint	7.3 Mb 3′ to gene	Fantauzzo et al. [2008]
TWIST (601622)	Saethre-Chotzen syndrome	7p21.1	Inversion and translocation breakpoints	>260 kb 3′ to gene	Cai et al. [2003]

et al., 2009; Glessner et al., 2009; Sebat et al., 2007; Stefansson et al., 2008; Walsh et al., 2008; Walters et al. 2010]. These recent findings point to genetic heterogeneity in these conditions thereby illustrating the likely complexity inherent in identifying all disease-causing CNVs. Intriguingly, Shlien et al. [2008] have reported a highly significant increase in CNV number among patients with Li-Fraumeni syndrome (MIM# 151623), carriers of inherited *TP53* mutations. Hence, it would appear that heritable genetic variants have the potential to modulate the rate of germline CNV formation.

It is already clear that the disease relevance of CNVs represents a continuum, stretching from "neutral" polymorphisms on the one hand to directly pathogenic copy number changes on the other [Beckmann et al., 2008]. Between these two extremes may lie those CNVs that are capable of acting as predisposing (or protective) factors in relation to complex disease [Fanciulli et al., 2010]. Thus, for example, a 117 kb deletion encompassing the UDP glucuronosyltransferase 2 family, polypeptide B17 (*UGT2B17*) gene (MIM# 601903) has been found to be associated with an increased risk of osteoporosis [Yang et al., 2008b]. Intriguingly, some germline CNVs appear to predispose to disease even although no known genes reside within their boundaries [Liu et al., 2009; Thean et al., 2010]. Importantly, a 520 kb microdeletion has been identified at 16p12.1, which predisposes to various neuropsychiatric phenotypes as a single copy number mutation and aggravates neurodevelopmental disorders if it co-occurs together with other large deletions and duplications [Girirajan et al., 2010]. It remains to be seen whether "CNV equivalents" <1 kb in size (also termed "indels"), that actually occur rather more frequently than true CNVs (>1 kb) [Conrad et al., 2010], will also be relevant to disease. What is already clear is that, over the coming years, a large number of important CNV-disease associations are going to come to light [Stankiewicz and Lupski, 2010].

Mutations in Nonprotein-Coding Genes

In contrast to the plethora of mutations identified in protein-coding genes, the identification of mutations in nonprotein-coding genes is still very much in its infancy. A number of disease-causing or disease-associated mutations have already been reported in various small nucleolar RNA genes and microRNA genes (Table 4). In addition, mutations have also been documented in the longer noncoding RNA genes (*XIST* [MIM# 314670], *TERC* [MIM# 602322], *H19* [MIM# 103280], *RMRP* [MIM# 157660]; see HGMD for details). A putative pathological mutation has been described in a "gene" encoding a paternally expressed antisense transcript of the GNAS complex locus (*GNASAS*; MIM# 610540) [Bastepe et al., 2005], whereas a functional polymorphism has been reported within an enhancer at the 3′ end of the *CDKN2BAS* "gene" (MIM# 600431), which encodes an antisense RNA transcript [Jarinova et al., 2009]. A *CRYGEP1* (MIM# 123660) pseudogene-reactivating mutation associated with hereditary cataract formation [Brakenhoff et al., 1994] probably also falls into this category.

The above examples are likely to comprise only the tip of a fairly large iceberg that still remains essentially unexplored. Thus, for example, both single nucleotide polymorphism and copy number variation are both likely to impact significantly on microRNA gene expression with myriad potential pathological consequences [Bandiera et al., 2010].

Table 4. Disease-Causing Mutations and Disease-Associated Polymorphisms in microRNA and Small Nucleolar RNA Genes

Gene (MIM#)	Disease/disease association	Chromosomal location	Mutation[c]	Nature and relative location of mutation[a]	Reference
MIR16-1 (609704)	Chronic lymphocytic leukemia, association with	13q14.3	NR_029486.1:r.89+7C>T	+96C>T[b]	Calin et al. [2005]
MIR17 (609416)	Breast cancer, association with	13q31.3	NR_029487.1:r.84+9C>T	+93C>T	Shen et al. [2009]
MIR30C1	Breast cancer, association with	1p34.2	NR_029833.1:r.48C>T	+48C>T	Shen et al. [2009]
MIR96 (611606)	Hearing loss, progressive	7q32.2	NR_029512.1:r.13G>A	+13G>A	Mencía et al. [2009]
MIR96 (611606)	Hearing loss, progressive	7q32.2	NR_029512.1:r.14C>A	+14C>A	Mencía et al. [2009]
MIR125A (611191)	Breast cancer, association with	19q13.33	NR_029693.1:r.22G>T	+22G>T[b]	Li et al. [2009c]
MIR146A (610566)	Papillary thyroid carcinoma, association with	5q33.3	NR_029701.1:r.60G>C	+60G>C[b]	Jazdzewski et al. [2008]
MIR191	Ovarian cancer, predisposition to	3p21.31	NR_029690.1:r.15G>C	+15G>C	Shen et al. [2010]
MIR196A2 (609687)	Nonsmall-cell lung cancer survival, associated with	12q13.13	NR_029617.1:r.78C>T	+78C>T[b]	Hu et al. [2008]
MIR206 (611599)	Cancers, reduced expression in association with	6p12.2	NR_029713.1:r.86+35C>T	+121C>T	Wu et al. [2008]
MIR499	Breast cancer, increased risk, association with	20q11.22	NR_030223.1:r.73A>G	+73A>G[b]	Hu et al. [2009]
MIR502	Schizophrenia, association with	Xp11.23	NR_030226.1:r.13C>T	+13C>T[b]	Sun et al. [2009]
MIR510	Schizophrenia, association with	Xp27.3	NR_030237.1:r.48T>C	+48T>C[b]	Sun et al. [2009]
Mir2861	Osteoporosis, primary	2	N/A	+33C>G	Li et al. [2009d]
MIRLET7E (611250)	Cancers, reduced expression in association with	19q13.41	NR_029482.1:r.79+19G>A	+98G>A	Wu et al. [2008]
SNORD50A (613117)	Breast and prostate cancer, association with	6q14.3	NR.002743.2:r.54_55delTT	ΔTT	Dong et al. [2008]
SNORD116@ gene cluster (605436)	Prader-Willi syndrome	15q11.2	–	Gross deletion	Sahoo et al. [2008]

N/A, no reference sequence available.
[a]Location given is relative to the transcriptional initiation site of the specified gene.
[b]Disease-associated polymorphism.
[c]HGVS nomenclature was adapted for use with microRNA genes.

Mutations in Noncoding Regions of Functional Significance

By adopting a gene-centric view, we have until now largely ignored the extensive nonprotein-coding portion of the human genome in our quest for mutations of pathological significance. As a consequence, we have not only seriously underestimated the extent of the functional component of the genome, but may also have overlooked many mutations within this genomic "dark matter" [Collins and Penny, 2009]. As we increasingly adopt "genotype-first" strategies to characterizing genetic defects in patients with diverse clinical phenotypes [Mefford, 2009], many more mutations are likely to be identified in nonprotein-coding genes.

In both the human and the mouse genomes, many noncoding regions exhibit a similar level of evolutionary conservation to that evident in protein-coding regions [Asthana et al., 2007; Kryukov et al., 2005]. As yet, however, little is known of the effect that mutations in these regions might have on either the phenotype or on overall fitness. Studies of the most evolutionarily conserved noncoding regions have yielded results that are consistent with the view that most mutations in noncoding regions are only slightly deleterious [Chen et al., 2007; Kryukov et al., 2005]. The conservation observed may thus be due to variations in the mutation rate rather than selective constraints [Gorlov et al., 2008; Keightley et al., 2005]. Indeed, Keightley et al. [2005] have shown that selection in conserved noncoding sequences is significantly weaker in hominids compared to murids, probably a consequence of the low effective population size of hominids resulting in the reduced effectiveness of selection.

To obtain a first, necessarily rather crude, estimate of the contribution of variation in human noncoding sequences to phenotypic and/or disease traits, Visel et al. [2009] performed a meta-analysis of ~1,200 SNPs that have been identified as the most significantly associated variants in published genome-wide association studies. They found that, in 40% of cases, neither the SNP in question nor its associated haplotype block overlapped with any known exons. These authors therefore concluded that in at least one-third of detected disease associations, variation in noncoding sequence rather than coding sequence could have causally contributed to the trait in question. We suspect that this could be because the common disease-common variant' hypothesis [Schork et al., 2009] may be much more likely to apply to noncoding sequence than to coding sequence, owing to the selectional constraints impacting upon sufficiently frequent functional variation in the latter. In similar vein, others have also estimated that 39–43% of trait/disease-associated SNPs in GWAS are located within intergenic regions [Glinskii et al., 2009; Hindorff et al., 2009]. This notwithstanding, it should be appreciated that any given variant apparently detected within a noncoding region may actually reside within a hitherto undiscovered exon [Denoeud et al., 2007]. We should however also be aware that rare variants, in *cis* to those found to be associated with a given disease or trait in GWAS studies, may simply by chance give rise to "synthetic associations" that are then attributed to much more common variants [Dickson et al., 2010].

Compensated Pathogenic Deviations

The intriguing idea that two individually deleterious mutations might be capable of restoring normal fitness when they occur in combination may be traced back to Kimura [1985], who suggested that "compensatory neutral mutations" might play an important role in evolution. More recently, Kondrashov et al. [2002] compared pathological missense mutations in 32 human proteins to the amino acid substitutions that occurred during the course of evolution of these same proteins, and estimated that ~10% of all amino acid sequence differences between a human protein and its nonhuman (mammalian) orthologue could represent what they termed "compensated pathogenic deviations" (CPDs). Because such amino acid substitutions are pathogenic in humans, Kondrashov et al. [2002] surmised that the normal functioning of a CPD-containing protein in the nonhuman species must be due to other ("compensatory") amino acid sequence deviations from the human sequence.

Numerous examples of CPDs have now been reported from comparative genome sequencing studies. CPDs represent human pathological missense mutations where the substituting amino acids have been found to be identical to the wild-type amino acid residues at the orthologous positions in, for example, the mouse [Gao and Zhang, 2003], macaque [Gibbs et al., 2007], and chimpanzee [Azevedo et al., 2006; Suriano et al., 2007]. In principle, these compensatory changes could be either allelic to the CPD (and, hence, closely linked genetically) or nonallelic (e.g., involving the coevolution of a ligand and its receptor encoded by unlinked genes) [Liu et al., 2001]. The above notwithstanding, in evolution, compensatory mutations are unlikely to occur singly; indeed, Poon et al. [2005] have suggested that, on average, 11.8 compensatory mutations may interact epistatically with a given deleterious mutation so as to restore wild-type levels of fitness.

CPDs tend to be less severe in terms of the difference in physicochemical properties between the substituted and substituting amino acids than is normally the case for pathological mutations [Barešić et al., 2010; Ferrer-Costa et al., 2007]. In the context of human disease, Suriano et al. [2007] have provided a good example of the influence of compensated and compensating mutations in the *OTC* gene. The human and chimpanzee OTC sequences differ at only two positions, amino acid residues 125 and 135. Amino acid replacements Thr135Ala and Thr125Met have respectively occurred in the human and chimpanzee lineages since their divergence from their common ancestor. The Thr135Ala substitution appears to be human-specific, whereas the Thr125Met substitution was chimpanzee-specific (both Thr125 and Thr135 were found to be ancestral residues). When the derived Met125 is associated with the ancestral Thr135 (in chimpanzee), no abnormal phenotype is evident. However, when Met125 occurs on a background containing the human-specific Ala135 residue, this results in a clinical phenotype (neonatal hyperammonemia). Suriano et al. [2007] demonstrated in vitro that human OTC bearing the Thr125Met mutant is inactive, whereas the chimpanzee version of OTC (with Met at residue 125) possesses an enzymatic activity comparable with the wild-type human OTC. The presence of Thr at position 135 in chimpanzees therefore rescues the deleterious effect of Met at position 125 through intralocus compensation.

The high proportion of disease-associated/functional SNPs that constitute CPDs in nonhuman primates may have important implications for the study of complex disease in humans. With mendelian diseases, the norm is for the pathological mutations to be new (i.e., derived), and in many cases, this paradigm can be extended to common disease. However, there are some curious examples in which the alleles that increase the risk of common disease are ancestral, whereas the derived alleles are "protective" [Di Rienzo and Hudson, 2005]. This reversal of the standard model is consistent with the idea that some forms of common disease susceptibility may be a consequence of ancient human adaptations to a long-term stable environment ("thrifty alleles"); with a changed environment consequent to the recent shift to a

modern lifestyle, these ancestral alleles have now come to increase the risk of common disease [Di Rienzo and Hudson, 2005]. Thus, the ancestral alleles represent the recapitulation of ancient states that may once have been protective, but which now result in adverse consequences for human health. On the other hand, some ancestral alleles may be weakly deleterious mutations that have become fixed by genetic drift [Kryukov et al., 2007], a process that may be facilitated by small effective population size). Viewed within this evolutionary framework, new (derived) alleles may be expected to confer "protection" against disease. Although ancestral alleles constitute only a minority of all putative risk variants, their number nevertheless appears to be sufficiently high for us to conclude that they are likely to account for a sizeable proportion of inherited susceptibility to common disease.

A Mutation in Search of a Gene

As is evident from the above, mutation hunting has so far been almost invariably gene-centric. Once a disease gene is discovered, the identification and characterization of pathological mutations within this gene usually follows apace. Generally speaking, the occasional exception serves only to prove the rule. Such an exception is fascioscapulohumeral muscular dystrophy (FSHD; MIM# 158900). The mutation responsible for this disease has long been known to be the deletion of a critical number of units of a repeat sequence (D4Z4) on chromosome 4q35. This deletion appears to correlate with the derepression of transcription of muscle-expressed genes in the vicinity of the D4Z4 repeats. However, although various candidates have been proposed [Dixit et al., 2007; Klooster et al., 2009; Snider et al., 2009], the identity and location of the FSHD gene (or genes) still remain elusive, as does the disease mechanism. It is anticipated that further examples of disease-associated mutations lacking an immediately obvious relationship to a specific gene or genes will come to light as our mutation-searching procedures become less gene-centric and more all-genome encompassing.

Refocusing Our Attention on the "Functionome"

In the context of identifying genetic variants responsible for human inherited disease, we believe that it will be increasingly important to consider functional elements in the genome (the "functionome") rather than simply genes per se. We employ the term "functionome" here to describe the totality of the biologically functional nucleotide sequences in the human genome, irrespective of whether they are associated with genes or not. A number of novel techniques, such as chromatin immunoprecipitation (ChIP) [Wong and Wei, 2009] and ChIP-sequencing (ChIP-Seq) [Park, 2009], which are capable of exploring protein–DNA interactions at a genome-wide (and protein–RNA interactions at a transcriptome-wide) level, are in the vanguard of attempts to characterize the human "functionome." Because conserved noncoding sequences in the human genome appear to be ~10-fold more abundant than known genes [Attanasio et al., 2008], it is likely that (1) currently known mutations within coding regions are unlikely to be fully representative of the universe of pathological mutations (which would imply that any extrapolation from HGMD data would be highly speculative), and (2) a whole new grouping of disease-causing mutations may await identification and characterization. Once again, a paradigm shift in our thinking may well be required if we are to maximize the potential of the emerging high-throughput technology to detect new (hitherto latent) types of human gene mutation.

The above notwithstanding, it is rather unlikely that the functional nonprotein-coding portion of the human genome will prove to be quite as mutation-dense as the protein-coding portion. For most inherited disorders, the mutation detection rate is already fairly high (>90%), although this success rate is often achieved by combining different mutation detection methodologies, for example, to screen for exon deletions and copy number variants as well as more subtle lesions [Quemener et al., 2010]. At least some of the "missing lesions" may nevertheless be found by screening extragenic functional elements.

How Many Deleterious Mutations are There on Average per Individual?

It has long been appreciated that every individual is heterozygous for a certain number of deleterious mutations that, if homozygous, would lead to the premature death of that individual [Bittles and Neel, 1994]. Based upon the average prevalence of recessive diseases in the human population, Morris [2001] estimated that there might be, on average, some 23 deleterious mutations in the protein coding region of a single individual. This estimate would receive additional support by reference to the expected disease allele frequency, $q = \mu/hs$ at mutation selection equilibrium: assuming a heterozygosity effect of $hs = 1.5 \times 10^{-3}$ for null mutations [Gillespie, 1998] and an average gene mutation rate of $\mu = 3 \times 10^{-6}$, the population frequency of the disease allele class of a given gene would amount to 2×10^{-3}, or 0.2%. Depending upon the number of inherited disease genes assumed to exist in the human genome (7,750 to 30,750; see above), the average number of deleterious (i.e., null) mutations in any given individual would therefore be expected to lie between 15 and 60. If we assume that the number of inherited disease genes is 15,300 (i.e., the average outcome of extrapolations based upon annual HGMD inclusion rates; see above), our best guess would be 31 deleterious mutations per individual. Depending upon whether the gene mutation rate and heterozygosity effect mentioned above cover nonsense SNPs and CNVs as well, such variants would either be included in this estimate, or not.

With the advent of whole genome sequencing, predictive mathematical modelling has largely given way to direct molecular analysis. Ng et al. [2008] employed the *SIFT* program to predict that 14% of 10,400 nonsynonymous variants (~1,500) detected in the Venter genome [Levy et al., 2007] would impact adversely upon protein function. Wheeler et al. [2008] employed *PolyPhen* to predict that some 20% of 3,898 nonsynonymous variants (~780) detected in the Watson genome would be "probably or possibly damaging" to protein function. Ng et al. [2009] reported 2,227 "probably damaging" and 3,368 "possibly damaging" variants predicted by *PolyPhen* from 13,295 nonsynonymous variants detected in 12 human "exomes" (comprising 180,000 exons per genome and corresponding to the 30 Mb protein coding region). *PolyPhen* has also been used to identify 765 "possibly damaging" SNPs and 454 "probably damaging" SNPs in the genome of a Yoruba (Nigerian) individual [McKernan et al., 2009]. Using a likelihood ratio test, Chun and Fay [2009] examined the genomes of Venter, Watson, and a Han Chinese male (whose sequence had been reported by [Wang et al., 2008a]) and identified between 796 and 837 deleterious mutations per genome, ~15% of all nonsynonymous variants assessed; most of these deleterious mutations were found to be heterozygous

(76–83%) and individual-specific (~60%). Chun and Fay [2009] estimated that their likelihood ratio test had been successful in identifying 62% of the "rare deleterious mutations" in the Venter genome. They also identified a further 838 deleterious mutations in the reference human genome [International Human Genome Sequencing Consortium, 2004], 474 of which were specific to that (artificial multisource) genome sequence and absent from the other three genomes examined [Chun and Fay, 2009]. Interestingly, some 435 (23%) of the 1,928 putatively deleterious variants found in the Venter, Watson, and Han Chinese genomes were present in more than one of these genomes. Existing compilations of mutation data, OMIM [Kim et al., 2009; Levy et al., 2007; McKernan et al., 2009; Ng et al., 2008; Schuster et al., 2010; Wang et al., 2008a], HGMD [Kim et al., 2009; McKernan et al., 2009; Rasmussen et al., 2010; Venter et al., 2001; Wheeler et al., 2008], SWISS-PROT [Chun and Fay, 2009] or SNPedia [Kim et al., 2009], have also been used directly to identify potential disease-associated mutations in the various sequenced genomes. However, because current genome sequencing protocols typically do not assemble whole human genomes but rather identify variants relative to a reference sequence [Snyder et al., 2010], it should be appreciated that not all variants detected are going to be bona fide because the original reference genome sequence exhibits an error rate of ~0.01% [Lander et al., 2001]. This problem is likely to be compounded by the finding that individual human genomes contain many "novel" sequences, corresponding to polymorphic structural variants, that cannot be readily mapped onto the reference genome sequence. Indeed, Li et al. [2010a] have estimated that a complete "human pan-genome" would contain an additional 19–40 Mb of novel sequence, over and above the reference genome sequence, that is both population- and individual-specific. However, Li et al. [2010a] believe that the vast majority of common structural sequence polymorphisms (those occurring with a frequency of >1% in the human population) should be defined by the complete sequencing of about 100–150 individuals randomly selected from the world population.

For a variety of reasons, it is very hard to compare directly the numbers of putatively disease-relevant variants detected in the different genome-wide sequencing studies and even more so to relate these numbers to the model-based, theoretical predictions mentioned above. First and foremost, the total number of deleterious variants present in a given genome will of course depend very much upon what we mean by "deleterious." A "deleterious" variant may well reduce or "damage" protein function but this is not to say that it will markedly alter the phenotype, let alone cause disease. This notwithstanding, what is striking is the remarkably similar number of "deleterious" variants reported from the different genomes studied to date and the fact that these numbers were between one and two orders of magnitude larger than those arrived at via theoretical considerations. The obvious explanation for this discrepancy is that the latter were focused upon recessive (or null) mutations. In practice, any increase in the fitness of (homozygous or heterozygous) carriers of the mutations in question would serve to increase the expected number of such mutations to be identified in a given individual by a corresponding amount. In fact, if the heterozygosity effect were an order of magnitude smaller than assumed above, then the number of deleterious SNPs would be an order of magnitude larger, and hence, would be of the same order of magnitude as the results obtained by genome/exome sequencing. The same result would pertain if selection against homozygous carriers were to be substantially weaker than in the case of recessive lethals. However, we doubt that there is solid evidence for such small effects being the rule rather than the exception with deleterious recessives (even though they may not be lethal). Moreover, such small effects would be difficult (if not impossible) to estimate directly, but would have to come from model-based studies.

Regarding the different outcomes of the sequencing studies, it must be noted that the levels of sequencing coverage ($7\times$ to $40\times$) [Yngvadottir et al., 2009a] differed quite dramatically between studies as did the portions of the genomes sequenced (i.e., "entire" genome vs. exome). Furthermore, the different sequencing platforms employed exhibit very different error patterns and rates [Smith et al., 2008; Wheeler et al., 2008]. Also, the different deleterious variant prediction tools used for functional profiling can differ quite markedly in terms of their sensitivity and specificity [Ng and Henikoff, 2006]. Finally, it should be appreciated that the question of ethnicity may impact significantly on the question of deleterious gene diversity. Thus, although African-Americans exhibit a higher level of heterozygosity for both "possibly damaging" and "probably damaging" SNPs than European-Americans, European-Americans possess significantly more genotypes that are homozygous for the putatively damaging allele of "probably damaging" SNPs than do African-Americans [Lohmueller et al., 2008].

To optimize their practical utility, the bioinformatics tools available for the prediction of deleterious mutations [Karchin, 2009] will need to be improved by the inclusion of data on specific sites of structural and/or functional interest [Mort et al., 2010] and by consideration of such key issues as mutation penetrance [Waalen and Beutler, 2009] and interactions between allelic and nonallelic mutations/polymorphic variants [Dimas et al., 2008]. However, it is most encouraging that existing bioinformatics tools have already been successfully applied in the context of filtering whole-exome/genome sequencing data to identify the pathological mutations underlying rare Mendelian disorders of previously unknown cause [Lupski et al., 2010; Ng et al., 2010]. Finally, the use of the same source of disease causing/disease-associated mutation and functional polymorphism data (e.g., HGMD) between studies could also introduce some uniformity into the pathological annotation of individual genomes thereby ensuring that valid crosscomparisons can be made.

Can we Estimate the Number of Mutations Causing Human Inherited Disease that Still Remain to be Characterized?

> *"There are known knowns; there are things we know we know. We also know there are known unknowns; that is to say we know there are some things we do not know. But there are also unknown unknowns—the ones we don't know we don't know."* (Donald Rumsfeld, Feb. 12, 2002, Department of Defense news briefing).

Because we still only have an approximate idea of the number of human genes, and a fairly crude estimate of the size and location of the functional portion of the human genome, the known unknowns would seem at present to outnumber the known knowns. Thus, any reliable estimate of the number of different functionally significant mutations yet to be identified in the extant human population is likely to remain a guessing game for the foreseeable future. What is clear, is that with the advent not only of massively parallel sequencing of the human exome [Choi et al., 2009; Hedges et al., 2009; Jones et al., 2009; Ng et al., 2009; Tucker

et al., 2009] and high-throughput targeted resequencing of defined genomic regions [Kryukov et al., 2009; Nikopoulos et al., 2010; Prabhu and Pe'er, 2009], but also of the successful application of direct RNA sequencing of the human transcriptome [Ozsolak et al., 2009; Wang et al., 2009b] and whole-genome sequencing [Lupski et al., 2010; Roach et al., 2010], the identification of inherited pathological mutations is entering a new era. This will be an era in which, for each patient, many genomic variants "will be called but few will be chosen." Hence, the development of bioinformatics techniques, sufficiently powerful to identify, with a high degree of certainty, pathological needles in the human genomic haystack, will be paramount. However, in deploying these emerging techniques, we should be wary of being constrained by outmoded overly gene-centric approaches to mutation screening. Once again, in terms of mutation hunting, we should not focus exclusively on genes *per se* but rather shift our emphasis so as to include the sequence elements that characterize a potentially larger (and yet still functional) portion of the genome. Expanding our horizons through the inclusion of new types of functional element among our screening targets should serve to extend the known germline mutational spectrum very significantly. We predict that entirely new types of pathological gene lesion (the unknown unknowns!) are likely to become apparent whose characterization should provide new insights not only into the morbid anatomy of the human genome but also its normal structure and function.

Concluding Remarks

In summary, the number of germline mutations in human nuclear genes known to either cause or to be associated with inherited disease now exceeds 100,000 in over 3,700 different genes. Newly described human gene mutations are currently being reported at a rate of ~10,000 per annum, with ~300 new "inherited disease genes" being recognized every year. As the human "mutome" passes the historic 100,000 landmark, we have posed the double question: how many inherited disease genes are there in the human genome and how many mutations are likely to be found within them? The total number of genes present in the human genome is dependent in part upon one's operating definition of a gene but appears to be at least 25,000 and may yet be found to exceed 33,000. We estimate that among these, there are likely to be at least 7,750 "disease genes," with our best guess being ~15,300. We further estimate that the total number of different mutations underlying inherited human disease may well exceed one billion although, in practice, most of these are going to occur too infrequently for them to be detectable. The question of the proportion of possible mutations within inherited human disease genes that are likely to be of pathological significance is very difficult to address because it is dependent not only upon the type and location of the mutation but also upon the functionality of the nucleotides involved. As to how many deleterious mutations there are on average per individual, if we assume that the total number of inherited disease genes is 15,300, then our best guess would be 31 such mutations per individual.

We surmise that, given current mutation screening techniques, it is very likely that many pathological mutations will have been overlooked as a consequence of their being located at some considerable distance from the genes whose function they disrupt. To avoid such oversights, we believe that it is important not to screen for mutations in an overly gene-centric way. Indeed, by coining here the term "functionome" to describe the universe of biologically functional nucleotide sequences in the human genome, we hope to encourage researchers to leave, when required, "the narrow roads of gene land" and to consider the totality of functional elements in the genome rather than simply opting for the increasingly well-trammelled path of analysing coding sequence or genes per se. We believe that this change of tack will amply repay us with the identification of novel types of pathological gene lesion whose characterization should yield new insights into human genome structure and function.

As we contemplate the future of mutation identification and characterization in a human context, we should not omit to mention that the term "mutation" in its broadest sense could, in principle, be extended beyond the traditional confines of DNA sequence-based changes so as to include heritable (germline) alterations of DNA methylation ("epimutations") that result in abnormal transcriptional silencing [Cropley et al., 2008]. The best example of this phenomenon is provided by the constitutional epimutations in the human *MLH1* gene (MIM♯ 120436), which cause hereditary nonpolyposis colorectal cancer [Hitchins and Ward, 2009]. With the determination of the human methylome [Lister et al., 2009] and the recent recognition that DNA sequence polymorphisms can exert an effect on gene function via allele-specific methylation in *cis* [Schalkwyk et al., 2010], the number of recognized epimutations should rise quite significantly in the coming years. If eventually shown to be both of pathological significance and heritable, some examples of histone modification [Luco et al., 2010; VerMilyea et al., 2009; Wang et al., 2008b] or RNA editing [Li et al., 2009b; Lualdi et al., 2010] could also turn out to represent "honorary mutations."

Irrespective of how the human germline mutational spectrum is transmogrified over the coming years, we must remain committed to collating human gene mutation data as they emerge, endeavoring as we do so to follow the advice of the founder of modern human genetics, William Bateson, who, in the context of collecting plant mutants over a century ago, exhorted us to "treasure your exceptions."

References

Aerts S, Lambrechts D, Maity S, Van Loo P, Coessens B, De Smet F, Tranchevent LC, De Moor B, Marynen P, Hassan B, Carmeliet P, Moreau Y. 2006. Gene prioritization through genomic data fusion. Nat Biotechnol 24:537–544.

Akiva P, Toporik A, Edelheit S, Peretz Y, Diber A, Shemesh R, Novik A, Sorek R. 2006. Transcription-mediated gene fusion in the human genome. Genome Res 16:30–36.

Alkan C, Kidd JM, Marques-Bonet T, Aksay G, Antonacci F, Hormozdiari F, Kitzman JO, Baker C, Malig M, Mutlu O, Sahinalp SC, Gibbs RA, Eichler EE. 2009. Personalized copy number and segmental duplication maps using next-generation sequencing. Nat Genet 41:1061–1067.

Antonarakis SE, Cooper DN. 2007. Mutations in human genetic disease. Nature and consequences. In: Principles and practice of medical genetics. 5th ed. Rimoin DL, Connor JM, Pyeritz RE, Korf BR. edotprs. Edinburgh: Churchill Livingstone, p 101–128.

Arbiza L, Duchi S, Montaner D, Burguet J, Pantoja-Uceda D, Pineda-Lucena A, Dopazo J, Dopazo H. 2006. Selective pressures at a codon-level predict deleterious mutations in human disease genes. J Mol Biol 358:1390–1404.

Asthana S, Noble WS, Kryukov G, Grant CE, Sunyaev S, Stamatoyannopoulos JA. 2007. Widely distributed noncoding purifying selection in the human genome. Proc Natl Acad Sci USA 104:12410–12415.

Attanasio C, Reymond A, Humbert R, Lyle R, Kuehn MS, Neph S, Sabo PJ, Goldy J, Weaver M, Haydock A, Lee K, Dorschner M, Dermitzakis ET, Antonarakis SE, Stamatoyannopoulos JA. 2008. Assaying the regulatory potential of mammalian conserved non-coding sequences in human cells. Genome Biol 9:R168.

Azevedo L, Suriano G, van Asch B, Harding RM, Amorim A. 2006. Epistatic interactions: how strong in disease and evolution? Trends Genet 22:581–585.

Bandiera S, Hatem E, Lyonnet S, Henrion-Caude A. 2010. microRNAs in diseases: from candidate to modifier genes. Clin Genet 77:306–313.

Barešić A, Hopcroft LE, Rogers HH, Hurst JM, Martin AC. 2010. Compensated pathogenic deviations: analysis of structural effects. J Mol Biol 396:19–30.

Barral DC, Ramalho JS, Anders R, Hume AN, Knapton HJ, Tolmachova T, Collinson LM, Goulding D, Authi KS, Seabra MC. 2002. Functional redundancy of Rab27 proteins and the pathogenesis of Griscelli syndrome. J Clin Invest 110:247–257.

Bastepe M, Fröhlich LF, Linglart A, Abu-Zahra HS, Tojo K, Ward LM, Jüppner H. 2005. Deletion of the NESP55 differentially methylated region causes loss of maternal *GNAS* imprints and pseudohypoparathyroidism type Ib. Nat Genet 37:25–27.

Beckmann JS, Sharp AJ, Antonarakis SE. 2008. CNVs and genetic medicine (excitement and consequences of a rediscovery). Cytogenet Genome Res 123:7–16.

Benito-Sanz S, Thomas NS, Huber C, Gorbenko del Blanco D, Aza-Carmona M, Crolla JA, Maloney V, Rappold G, Argente J, Campos-Barros A, Cormier-Daire V, Heath KE. 2005. A novel class of pseudoautosomal region 1 deletions downstream of *SHOX* is associated with Leri-Weill dyschondrosteosis. Am J Hum Genet 77:533–544.

Benko S, Fantes JA, Amiel J, Kleinjan DJ, Thomas S, Ramsay J, Jamshidi N, Essafi A, Heaney S, Gordon CT, McBride D, Golzio C, Fisher M, Perry P, Abadie V, Ayuso C, Holder-Espinasse M, Kilpatrick N, Lees MM, Picard A, Temple IK, Thomas P, Vazquez MP, Vekemans M, Crollius HR, Hastie ND, Munnich A, Etchevers HC, Pelet A, Farlie PG, Fitzpatrick DR, Lyonnet S. 2009. Highly conserved non-coding elements on either side of *SOX9* associated with Pierre Robin sequence. Nat Genet 41:359–364.

Bird CP, Stranger BE, Liu M, Thomas DJ, Ingle CE, Beazley C, Miller W, Hurles ME, Dermitzakis ET. 2007. Fast-evolving noncoding sequences in the human genome. Genome Biol 8:R118.

Bittles AH, Neel JV. 1994. The costs of human inbreeding and their implications for variations at the DNA level. Nat Genet 8:117–121.

Bochukova EG, Huang N, Keogh J, Henning E, Purmann C, Blaszczyk K, Saeed S, Hamilton-Shield J, Clayton-Smith J, O'Rahilly S, Hurles ME, Farooqi IS. 2010. Large, rare chromosomal deletions associated with severe early-onset obesity. Nature 463:666–670.

Borel C, Antonarakis SE. 2008. Functional genetic variation of human miRNAs and phenotypic consequences. Mamm Genome 19:503–509.

Borel C, Gagnebin M, Gehrig C, Kriventseva EV, Zdobnov EM, Antonarakis SE. 2008. Mapping of small RNAs in the human ENCODE regions. Am J Hum Genet 82:971–981.

Boyko AR, Williamson SH, Indap AR, Degenhardt JD, Hernandez RD, Lohmueller KE, Adams MD, Schmidt S, Sninsky JJ, Sunyaev SR, White TJ, Nielsen R, Clark AG, Bustamante CD. 2008. Assessing the evolutionary impact of amino acid mutations in the human genome. PLoS Genet 4:e1000083.

Brakenhoff RH, Henskens HA, van Rossum MW, Lubsen NH, Schoenmakers JG. 1994. Activation of the gamma E-crystallin pseudogene in the human hereditary Coppock-like cataract. Hum Mol Genet 3:279–283.

Brouwer JR, Willemsen R, Oostra BA. 2009. Microsatellite repeat instability and neurological disease. Bioessays 31:71–83.

Brown KK, Reiss JA, Crow K, Ferguson HL, Kelly C, Fritzsch B, Morton CC. 2010. Deletion of an enhancer near DLX5 and DLX6 in a family with hearing loss, craniofacial defects, and an inv(7)(q21.3q35). Hum Genet 127:19–31.

Buschiazzo E, Gemmell NJ. 2010. Conservation of human microsatellites across 450 million years of evolution. Genome Biol Evol 2010:153–165.

Cai J, Goodman BK, Patel AS, Mulliken JB, Van Maldergem L, Hoganson GE, Paznekas WA, Ben-Neriah Z, Sheffer R, Cunningham ML, Daentl DL, Jabs EW. 2003. Increased risk for developmental delay in Saethre-Chotzen syndrome is associated with *TWIST* deletions: an improved strategy for *TWIST* mutation screening. Hum Genet 114:68–76.

Cai JJ, Borenstein E, Chen R, Petrov DA. 2009. Similarly strong purifying selection acts on human disease genes of all evolutionary ages. Genome Biol Evol 2009:131–144.

Calin GA, Ferracin M, Cimmino A, Di Leva G, Shimizu M, Wojcik SE, Iorio MV, Visone R, Sever NI, Fabbri M, Iuliano R, Palumbo T, Pichiorri F, Roldo C, Garzon R, Sevignani C, Rassenti L, Alder H, Volinia S, Liu CG, Kipps TJ, Negrini M, Croce CM. 2005. A microRNA signature associated with prognosis and progression in chronic lymphocytic leukemia. N Engl J Med 353:1793–1801.

Calvo SE, Pagliarini DJ, Mootha VK. 2009. Upstream open reading frames cause widespread reduction of protein expression and are polymorphic among humans. Proc Natl Acad Sci USA 106:7507–7512.

Capriotti E, Arbiza L, Casadio R, Dopazo J, Dopazo H, Marti-Renom MA. 2008. Use of estimated evolutionary strength at the codon level improves the prediction of disease-related protein mutations in humans. Hum Mutat 29:198–204.

Carninci P, Sandelin A, Lenhard B, Katayama S, Shimokawa K, Ponjavic J, Semple CA, Taylor MS, Engström PG, Frith MC, Forrest AR, Alkema WB, Tan SL, Plessy C, Kodzius R, Ravasi T, Kasukawa T, Fukuda S, Kanamori-Katayama M, Kitazume Y, Kawaji H, Kai C, Nakamura M, Konno H, Nakano K, Mottagui-Tabar S, Arner P, Chesi A, Gustincich S, Persichetti F, Suzuki H, Grimmond SM, Wells CA, Orlando V, Wahlestedt C, Liu ET, Harbers M, Kawai J, Bajic VB, Hume DA, Hayashizaki Y. 2006. Genome-wide analysis of mammalian promoter architecture and evolution. Nat Genet 38:626–635.

Cartegni L, Chew SL, Krainer AR. 2002. Listening to silence and understanding nonsense: exonic mutations that affect splicing. Nat Rev Genet 3:285–298.

Cecchini KR, Raja Banerjee A, Kim TH. 2009. Towards a genome-wide reconstruction of cis-regulatory networks in the human genome. Semin Cell Dev Biol 20:842–848.

Chang JC, Kan YW. 1979. β^0 thalassemia, a nonsense mutation in man. Proc Natl Acad Sci USA 76:2886–2889.

Chatterjee S, Pal JK. 2009. Role of 5′- and 3′-untranslated regions of mRNAs in human diseases. Biol Cell 101:251–262.

Chen CT, Wang JC, Cohen BA. 2007. The strength of selection on ultraconserved elements in the human genome. Am J Hum Genet 80:692–704.

Chen J, Wildhardt G, Zhong Z, Röth R, Weiss B, Steinberger D, Decker J, Blum WF, Rappold G. 2009a. Enhancer deletions of the *SHOX* gene as a frequent cause of short stature: the essential role of a 250 kb downstream regulatory domain. J Med Genet 46:834–839.

Chen JM, Férec C, Cooper DN. 2006a. A systematic analysis of disease-associated variants in the 3′ regulatory regions of human protein-coding genes I: general principles and overview. Hum Genet 120:1–21.

Chen JM, Férec C, Cooper DN. 2006b. A systematic analysis of disease-associated variants in the 3′ regulatory regions of human protein-coding genes II: the importance of mRNA secondary structure in assessing the functionality of 3′ UTR variants. Hum Genet 120:301–333.

Chen JM, Férec C, Cooper DN. 2009b. Closely spaced multiple mutations as potential signatures of transient hypermutability in human genes. Hum Mutat 30:1435–1448.

Choi JW, Park CS, Hwang M, Nam HY, Chang HS, Park SG, Han BG, Kimm K, Kim HL, Oh B, Kim Y. 2008. A common intronic variant of *CXCR3* is functionally associated with gene expression levels and the polymorphic immune cell responses to stimuli. J Allergy Clin Immunol 122: 1119–1126.e7.

Choi M, Scholl UI, Ji W, Liu T, Tikhonova IR, Zumbo P, Nayir A, Bakkaloglu A, Ozen S, Sanjad S, Nelson-Williams C, Farhi A, Mane S, Lifton RP. 2009. Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. Proc Natl Acad Sci USA 106:19096–19101.

Chorley BN, Wang X, Campbell MR, Pittman GS, Noureddine MA, Bell DA. 2008. Discovery and verification of functional single nucleotide polymorphisms in regulatory genomic regions: current and developing technologies. Mutat Res 659:147–157.

Chun S, Fay JC. 2009. Identification of deleterious mutations within three human genomes. Genome Res 19:1553–1561.

Clamp M, Fry B, Kamal M, Xie X, Cuff J, Lin MF, Kellis M, Lindblad-Toh K, Lander ES. 2010. Distinguishing protein-coding and noncoding genes in the human genome. Proc Natl Acad Sci USA 104:19428–19433.

Collins LJ, Penny D. 2009. The RNA infrastructure: dark matter of the eukaryotic cell? Trends Genet 25:120–128.

Conley AB, Miller WJ, Jordan IK. 2008. Human *cis* natural antisense transcripts initiated by transposable elements. Trends Genet 24:53–56.

Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, Zhang Y, Aerts J, Andrews TD, Barnes C, Campbell P, Fitzgerald T, Hu M, Ihm CH, Kristiansson K, Macarthur DG, Macdonald JR, Onyiah I, Pang AW, Robson S, Stirrups K, Valsesia A, Walter K, Wei J, The Wellcome Trust Case Control Consortium, Tyler-Smith C, Carter NP, Lee C, Scherer SW, Hurles ME. 2010. Origins and functional impact of copy number variation in the human genome. Nature 464:704–712.

Cotton RGH. 2009. Collection of variation causing disease—the Human Variome Project. Hum Genomics 3:301–303.

Coulombe-Huntington J, Lam KC, Dias C, Majewski J. 2009. Fine-scale variation and genetic determinants of alternative splicing across individuals. PLoS Genet 5:e1000766.

Coutinho AM, Oliveira G, Katz C, Feng J, Yan J, Yang C, Marques C, Ataíde A, Miguel TS, Borges L, Almeida J, Correia C, Currais A, Bento C, Mota-Vieira L, Temudo T, Santos M, Maciel P, Sommer SS, Vicente AM. 2007. *MECP2* coding sequence and 3′UTR variation in 172 unrelated autistic patients. Am J Med Genet B Neuropsychiatr Genet 144B:475–483.

Coutinho G, Xie J, Du L, Brusco A, Krainer AR, Gatti RA. 2005. Functional significance of a deep intronic mutation in the *ATM* gene and evidence for an alternative exon 28a. Hum Mutat 25:118–124.

Cropley JE, Martin DI, Suter CM. 2008. Germline epimutation in humans. Pharmacogenomics 9:1861–1868.

Cutler G, Kassner PD. 2008. Copy number variation in the mouse genome: implications for the mouse as a model organism for human disease. Cytogenet Genome Res 123:297–306.

Dathe K, Kjaer KW, Brehm A, Meinecke P, Nürnberg P, Neto JC, Brunoni D, Tommerup N, Ott CE, Klopocki E, Seemann P, Mundlos S. 2009. Duplications

involving a conserved regulatory element downstream of *BMP2* are associated with brachydactyly type A2. Am J Hum Genet 84:483–492.

Dear A, Daly J, Brennan SO, Tuckfield A, George PM. 2006. An intronic mutation within *FGB* (IVS1+2076 a→g) is associated with afibrinogenemia and recurrent transient ischemic attacks. J Thromb Haemost 4:471–472.

Dear PH. 2009. Copy-number variation: the end of the human genome? Trends Biotechnol 27:448–454.

De Gobbi M, Viprakasit V, Hughes JR, Fisher C, Buckle VJ, Ayyub H, Gibbons RJ, Vernimmen D, Yoshinaga Y, de Jong P, Cheng JF, Rubin EM, Wood WG, Bowden D, Higgs DR. 2006. A regulatory SNP causes a human genetic disease by creating a new transcriptional promoter. Science 312:1215–1217.

de Kok YJ, Vossenaar ER, Cremers CW, Dahl N, Laporte J, Hu LJ, Lacombe D, Fischel-Ghodsian N, Friedman RA, Parnes LS, Thorpe P, Bitner-Glindzicz M, Pander HJ, Heilbronner H, Graveline J, den Dunnen JT, Brunner HG, Ropers HH, Cremers FP. 1996. Identification of a hot spot for microdeletions in patients with X-linked deafness type 3 (DFN3) 900 kb proximal to the DFN3 gene *POU3F4*. Hum Mol Genet 5:1229–1235.

de Smith AJ, Walters RG, Froguel P, Blakemore AI. 2008. Human genes involved in copy number variation: mechanisms of origin, functional effects and implications for disease. Cytogenet Genome Res 123:17–26.

Dehainault C, Michaux D, Pagès-Berhouet S, Caux-Moncoutier V, Doz F, Desjardins L, Couturier J, Parent P, Stoppa-Lyonnet D, Gauthier-Villars M, Houdayer C. 2007. A deep intronic mutation in the *RB1* gene leads to intronic sequence exonisation. Eur J Hum Genet 15:473–477.

Denoeud F, Kapranov P, Ucla C, Frankish A, Castelo R, Drenkow J, Lagarde J, Alioto T, Manzano C, Chrast J, Dike S, Wyss C, Henrichsen CN, Holroyd N, Dickson MC, Taylor R, Hance Z, Foissac S, Myers RM, Rogers J, Hubbard T, Harrow J, Guigó R, Gingeras TR, Antonarakis SE, Reymond A. 2007. Prominent use of distal 5′ transcription start sites and discovery of a large number of additional exons in ENCODE regions. Genome Res 17:746–759.

D'haene B, Attanasio C, Beysen D, Dostie J, Lemire E, Bouchard P, Field M, Jones K, Lorenz B, Menten B, Buysse K, Pattyn F, Friedli M, Ucla C, Rossier C, Wyss C, Speleman F, De Paepe A, Dekker J, Antonarakis SE, De Baere E. 2009. Disease-causing 7.4 kb *cis*-regulatory deletion disrupting conserved non-coding sequences and their interaction with the *FOXL2* promotor: implications for mutation screening. PLoS Genet 5:e1000522.

Dhir A, Buratti E. 2010. Alternative splicing: role of pseudoexons in human disease and potential therapeutic strategies. FEBS J 277:841–855.

Di Rienzo A, Hudson RR. 2005. An evolutionary framework for common diseases: the ancestral-susceptibility model. Trends Genet 21:596–601.

Dickson SP, Wang K, Krantz I, Hakonarson H, Goldstein DB. 2010. Rare variants create synthetic genome-wide associations. PLoS Biol 8:e1000294.

Dimas AS, Deutsch S, Stranger BE, Montgomery SB, Borel C, Attar-Cohen H, Ingle C, Beazley C, Gutierrez Arcelus M, Sekowska M, Gagnebin M, Nisbett J, Deloukas P, Dermitzakis ET, Antonarakis SE. 2009. Common regulatory variation impacts gene expression in a cell type-dependent manner. Science 325:1246–1250.

Dimas AS, Stranger BE, Beazley C, Finn RD, Ingle CE, Forrest MS, Ritchie ME, Deloukas P, Tavaré S, Dermitzakis ET. 2008. Modifier effects between regulatory and protein-coding variation. PLoS Genet 4:e1000244.

Dinger ME, Amaral PP, Mercer TR, Mattick JS. 2009. Pervasive transcription of the eukaryotic genome: functional indices and conceptual implications. Brief Funct Genomic Proteomic 8:407–423.

Dixit M, Ansseau E, Tassin A, Winokur S, Shi R, Qian H, Sauvage S, Mattéotti C, van Acker AM, Leo O, Figlewicz D, Barro M, Laoudj-Chenivesse D, Belayew A, Coppée F, Chen YW. 2007. *DUX4*, a candidate gene of facioscapulohumeral muscular dystrophy, encodes a transcriptional activator of PITX1. Proc Natl Acad Sci USA 104:18157–18162.

Domazet-Lošo T, Tautz D. 2008. An ancient evolutionary origin of genes associated with human genetic diseases. Mol Biol Evol 25:2699–2707.

Dong XY, Rodriguez C, Guo P, Sun X, Talbot JT, Zhou W, Petros J, Li Q, Vessella RL, Kibel AS, Stevens VL, Calle EE, Dong JT. 2008. SnoRNA U50 is a candidate tumor-suppressor gene at 6q14.3 with a mutation associated with clinically significant prostate cancer. Hum Mol Genet 17:1031–1042.

Drake JA, Bird C, Nemesh J, Thomas DJ, Newton-Cheh C, Reymond A, Excoffier L, Attar H, Antonarakis SE, Dermitzakis ET, Hirschhorn JN. 2006. Conserved noncoding sequences are selectively constrained and not mutation cold spots. Nat Genet 38:223–227.

Driscoll MC, Dobkin CS, Alter BP. 1989. γδβ-thalassemia due to a *de novo* mutation deleting the 5′ β-globin gene activation-region hypersensitive sites. Proc Natl Acad Sci USA 86:7470–7474.

Elia J, Gai X, Xie HM, Perin JC, Geiger E, Glessner JT, D'arcy M, Deberardinis R, Frackelton E, Kim C, Lantieri F, Muganga BM, Wang L, Takeda T, Rappaport EF, Grant SF, Berrettini W, Devoto M, Shaikh TH, Hakonarson H, White PS. 2009. Rare structural variants found in attention-deficit hyperactivity disorder are preferentially associated with neurodevelopmental genes. Mol Psychiatry (in press).

Emison ES, McCallion AS, Kashuk CS, Bush RT, Grice E, Lin S, Portnoy ME, Cutler DJ, Green ED, Chakravarti A. 2005. A common sex-dependent mutation in a *RET* enhancer underlies Hirschsprung disease risk. Nature 434:857–863.

Enattah NS, Sahi T, Savilahti E, Terwilliger JD, Peltonen L, Järvelä I. 2002. Identification of a variant associated with adult-type hypolactasia. Nat Genet 30:233–237.

ENCODE Project Consortium, 2007. 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature 447:799–816.

Eory L, Halligan DL, Keightley PD. 2010. Distributions of selectively constrained sites and deleterious mutation rates in the hominid and murid genomes. Mol Biol Evol 27:177–192.

Faghihi MA, Wahlestedt C. 2009. Regulatory roles of natural antisense transcripts. Nat Rev Mol Cell Biol 10:637–643.

Fanciulli M, Petretto E, Aitman TJ. 2010. Gene copy number variation and common human disease. Clin Genet 77:201–213.

Fantauzzo KA, Tadin-Strapps M, You Y, Mentzer SE, Baumeister FA, Cianfarani S, Van Maldergem L, Warburton D, Sundberg JP, Christiano AM. 2008. A position effect on *TRPS1* is associated with Ambras syndrome in humans and the Koala phenotype in mice. Hum Mol Genet 17:3539–3551.

Fay JC, Wyckoff GJ, Wu CI. 2001. Positive and negative selection on the human genome. Genetics 158:1227–1234.

Fejes-Toth K, Sotirova V, Sachidanandam R, Assaf G, Hannon GJ, Kapranov P, Foissac S, Willingham AT, Duttagupta R, Dumais E, Gingeras TR. 2009. Post-transcriptional processing generates a diversity of 5′-modified long and short RNAs. Nature 457:1028–1032.

Feldman I, Rzhetsky A, Vitkup D. 2008. Network properties of genes harboring inherited disease mutations. Proc Natl Acad Sci USA 105:4323–4328.

Ferrer-Costa C, Orozco M, de la Cruz X. 2002. Characterization of disease-associated single amino acid polymorphisms in terms of sequence and structure properties. J Mol Biol 315:771–786.

Ferrer-Costa C, Orozco M, de la Cruz X. 2007. Characterization of compensated mutations in terms of structural and physico-chemical properties. J Mol Biol 365:249–256.

Flomen RH, Vatcheva R, Gorman PA, Baptista PR, Groet J, Barisić I, Ligutic I, Nizetić D. 1998. Construction and analysis of a sequence-ready map in 4q25: Rieger syndrome can be caused by haploinsufficiency of *RIEG*, but also by chromosome breaks approximately 90 kb upstream of this gene. Genomics 47:409–413.

Fraser HB, Xie X. 2009. Common polymorphic transcript variation in human disease. Genome Res 19:567–575.

Frazer KA, Murray SS, Schork NJ, Topol EJ. 2009. Human genetic variation and its contribution to complex traits. Nat Rev Genet 10:241–251.

Furney SJ, Albà MM, López-Bigas N. 2006. Differences in the evolutionary history of disease genes affected by dominant or recessive mutations. BMC Genomics 7:165.

Furniss D, Lettice LA, Taylor IB, Critchley PS, Giele H, Hill RE, Wilkie AO. 2008. A variant in the sonic hedgehog regulatory sequence (ZRS) is associated with triphalangeal thumb and deregulates expression in the developing limb. Hum Mol Genet 17:2417–2423.

Gale JS. 1990. Theoretical population genetics. London: Unwin Hyman, p 325–343.

Gao L, Zhang J. 2003. Why are some human disease-associated mutations fixed in mice? Trends Genet 19:678–681.

Gerdes S, Edwards R, Kubal M, Fonstein M, Stevens R, Osterman A. 2006. Essential genes on metabolic maps. Curr Opin Biotechnol 17:448–456.

Gerstein MB, Bruce C, Rozowsky JS, Zheng D, Du J, Korbel JO, Emanuelsson O, Zhang ZD, Weissman S, Snyder M. 2007. What is a gene, post-ENCODE? History and updated definition. Genome Res 17:669–681.

Gibbs RA, Rogers J, Katze MG, Bumgarner R, Weinstock GM, Mardis ER, Remington KA, Strausberg RL, Venter JC, Wilson RK, Batzer MA, Bustamante CD, Eichler EE, Hahn MW, Hardison RC, Makova KD, Miller W, Milosavljevic A, Palermo RE, Siepel A, Sikela JM, Attaway T, Bell S, Bernard KE, Buhay CJ, Chandrabose MN, Dao M, Davis C, Delehaunty KD, Ding Y, Dinh HH, Dugan-Rocha S, Fulton LA, Gabisi RA, Garner TT, Godfrey J, Hawes AC, Hernandez J, Hines S, Holder M, Hume J, Jhangiani SN, Joshi V, Khan ZM, Kirkness EF, Cree A, Fowler RG, Lee S, Lewis LR, Li Z, Liu YS, Moore SM, Muzny D, Nazareth LV, Ngo DN, Okwuonu GO, Pai G, Parker D, Paul HA, Pfannkoch C, Pohl CS, Rogers YH, Ruiz SJ, Sabo A, Santibanez J, Schneider BW, Smith SM, Sodergren E, Svatek AF, Utterback TR, Vattathil S, Warren W, White CS, Chinwalla AT, Feng Y, Halpern AL, Hillier LW, Huang X, Minx P, Nelson JO, Pepin KH, Qin X, Sutton GG, Venter E, Walenz BP, Wallis JW, Worley KC, Yang SP, Jones SM, Marra MA, Rocchi M, Schein JE, Baertsch R, Clarke L, Csürös M, Glasscock J, Harris RA, Havlak P, Jackson AR, Jiang H, Liu Y, Messina DN, Shen Y, Song HX, Wylie T, Zhang L, Birney E, Han K, Konkel MK, Lee J, Smit AF, Ullmer B, Wang H, Xing J, Burhans R, Cheng Z,

Karro JE, Ma J, Raney B, She X, Cox MJ, Demuth JP, Dumas LJ, Han SG, Hopkins J, Karimpour-Fard A, Kim YH, Pollack JR, Vinar T, Addo-Quaye C, Degenhardt J, Denby A, Hubisz MJ, Indap A, Kosiol C, Lahn BT, Lawson HA, Marklein A, Nielsen R, Vallender EJ, Clark AG, Ferguson B, Hernandez RD, Hirani K, Kehrer-Sawatzki H, Kolb J, Patil S, Pu LL, Ren Y, Smith DG, Wheeler DA, Schenck I, Ball EV, Chen R, Cooper DN, Giardine B, Hsu F, Kent WJ, Lesk A, Nelson DL, O'Brien WE, Prüfer K, Stenson PD, Wallace JC, Ke H, Liu XM, Wang P, Xiang AP, Yang F, Barber GP, Haussler D, Karolchik D, Kern AD, Kuhn RM, Smith KE, Zwieg AS. 2007. Evolutionary and biomedical insights from the rhesus macaque genome. Science 316:222–234.

Gingeras TR. 2007. Origin of phenotypes: genes and transcripts. Genome Res 17:682–690.

Gingeras TR. 2009. Implications of chimaeric non-co-linear transcripts. Nature 461:206–211.

Girirajan S, Rosenfeld JA, Cooper GM, Antonacci F, Siswara P, Itsara A, Vives L, Walsh T, McCarthy SE, Baker C, Mefford HC, Kidd JM, Browning SR, Browning BL, Dickel DE, Levy DL, Ballif BC, Platky K, Farber DM, Gowans GC, Wetherbee JJ, Asamoah A, Weaver DD, Mark PR, Dickerson J, Garg BP, Ellingwood SA, Smith R, Banks VC, Smith W, McDonald MT, Hoo JJ, French BN, Hudson C, Johnson JP, Ozmore JR, Moeschler JB, Surti U, Escobar LF, El-Khechen D, Gorski JL, Kussmann J, Salbert B, Lacassie Y, Biser A, McDonald-McGinn DM, Zackai EH, Deardorff MA, Shaikh TH, Haan E, Friend KL, Fichera M, Romano C, Gécz J, DeLisi LE, Sebat J, King MC, Shaffer LG, Eichler EE. 2010. A recurrent 16p12.1 microdeletion supports a two-hit model for severe developmental delay. Nat Genet 2010 42:203–209.

Glessner JT, Wang K, Cai G, Korvatska O, Kim CE, Wood S, Zhang H, Estes A, Brune CW, Bradfield JP, Imielinski M, Frackelton EC, Reichert J, Crawford EL, Munson J, Sleiman PM, Chiavacci R, Annaiah K, Thomas K, Hou C, Glaberson W, Flory J, Otieno F, Garris M, Soorya L, Klei L, Piven J, Meyer KJ, Anagnostou E, Sakurai T, Game RM, Rudd DS, Zurawiecki D, McDougle CJ, Davis LK, Miller J, Posey DJ, Michaels S, Kolevzon A, Silverman JM, Bernier R, Levy SE, Schultz RT, Dawson G, Owley T, McMahon WM, Wassink TH, Sweeney JA, Nurnberger JI, Coon H, Sutcliffe JS, Minshew NJ, Grant SF, Bucan M, Cook EH, Buxbaum JD, Devlin B, Schellenberg GD, Hakonarson H. 2009. Autism genome-wide copy number variation reveals ubiquitin and neuronal genes. Nature 459:569–573.

Glinskii AB, Ma J, Ma S, Grant D, Lim CU, Sell S, Glinsky GV. 2009. Identification of intergenic trans-regulatory RNAs containing a disease-linked SNP sequence and targeting cell cycle progression/differentiation pathways in multiple common human disorders. Cell Cycle 8:3925–3942.

Goh KI, Cusick ME, Valle D, Childs B, Vidal M, Barabási AL. 2007. The human disease network. Proc Natl Acad Sci USA 104:8685–8690.

Goode DL, Cooper GM, Schmutz J, Dickson M, Gonzales E, Tsai M, Karra K, Davydov E, Batzoglou S, Myers RM, Sidow A. 2010. Evolutionary constraint facilitates interpretation of genetic variation in resequenced human genomes. Genome Res 20:301–310.

Gordon CT, Tan TY, Benko S, Fitzpatrick D, Lyonnet S, Farlie PG. 2009. Long-range regulation at the *SOX9* locus in development and disease. J Med Genet 46:649–656.

Gorlov IP, Kimmel M, Amos CI. 2006. Strength of the purifying selection against different categories of the point mutations in the coding regions of the human genome. Hum Mol Genet 15:1143–1150.

Gorlov IP, Gorlova OY, Amos CI. 2008. Relative effects of mutability and selection on single nucleotide polymorphisms in transcribed regions of the human genome. BMC Genomics 9:292.

Grant SF, Reid DM, Blake G, Herd R, Fogelman I, Ralston SH. 1996. Reduced bone density and osteoporosis associated with a polymorphic Sp1 binding site in the collagen type I alpha 1 gene. Nat Genet 14:203–205.

Griffiths-Jones S. 2007. Annotating noncoding RNA genes. Annu Rev Genomics Hum Genet 8:279–298.

Grinchuk OV, Jenjaroenpun P, Orlov YL, Zhou J, Kuznetsov VA. 2010. Integrative analysis of the human *cis*-antisense gene pairs, miRNAs and their transcription regulation patterns. Nucleic Acids Res 38:534–547.

Gross-Hardt S, Reiss J. 2002. The bicistronic *MOCS1* gene has alternative start codons on two mutually exclusive exons. Mol Genet Metab 76:340–343.

Gurnett CA, Bowcock AM, Dietz FR, Morcuende JA, Murray JC, Dobbs MB. 2007. Two novel point mutations in the long-range *SHH* enhancer in three families with triphalangeal thumb and preaxial polydactyly. Am J Med Genet A 143:27–32.

Guttman M, Amit I, Garber M, French C, Lin MF, Feldser D, Huarte M, Zuk O, Carey BW, Cassady JP, Cabili MN, Jaenisch R, Mikkelsen TS, Jacks T, Hacohen N, Bernstein BE, Kellis M, Regev A, Rinn JL, Lander ES. 2009. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature 458:223–227.

Haiman CA, Le Marchand L, Yamamato J, Stram DO, Sheng X, Kolonel LN, Wu AH, Reich D, Henderson BE. 2007. A common genetic risk factor for colorectal and prostate cancer. Nat Genet 39:954–956.

Harland M, Mistry S, Bishop DT, Bishop JA. 2001. A deep intronic mutation in *CDKN2A* is associated with disease in a subset of melanoma pedigrees. Hum Mol Genet 10:2679–2686.

Harrison PM, Zheng D, Zhang Z, Carriero N, Gerstein M. 2005. Transcribed processed pseudogenes in the human genome: an intermediate form of expressed retrosequence lacking protein-coding ability. Nucleic Acids Res 33:2374–2383.

Harteveld CL, Voskamp A, Phylipsen M, Akkermans N, den Dunnen JT, White SJ, Giordano PC. 2005. Nine unknown rearrangements in 16p13.3 and 11p15.4 causing α- and β-thalassaemia characterised by high resolution multiplex ligation-dependent probe amplification. J Med Genet 42:922–931.

Hatton CS, Wilkie AO, Drysdale HC, Wood WG, Vickers MA, Sharpe J, Ayyub H, Pretorius IM, Buckle VJ, Higgs DR. 1990. α-Thalassemia caused by a large (62 kb) deletion upstream of the human α-globin gene cluster. Blood 76:221–227.

He Y, Vogelstein B, Velculescu VE, Papadopoulos N, Kinzler KW. 2008. The antisense transcriptomes of human cells. Science 322:1855–1857.

Hedges DJ, Burges D, Powell E, Almonte C, Huang J, Young S, Boese B, Schmidt M, Pericak-Vance MA, Martin E, Zhang X, Harkins TT, Züchner S. 2009. Exome sequencing of a multigenerational human pedigree. PLoS One 4:e8232.

Heintzman ND, Ren B. 2009. Finding distal regulatory elements in the human genome. Curr Opin Genet Dev 19:541–549.

Herzog H, Darby K, Hort YJ, Shine J. 1996. Intron 17 of the human retinoblastoma susceptibility gene encodes an actively transcribed G protein-coupled receptor gene. Genome Res 6:858–861.

Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA. 2009. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci USA 106:9362–9367.

Hirotsune S, Yoshida N, Chen A, Garrett L, Sugiyama F, Takahashi S, Yagami K, Wynshaw-Boris A, Yoshiki A. 2003. An expressed pseudogene regulates the messenger-RNA stability of its homologous coding gene. Nature 423:91–96.

Hitchins MP, Ward RL. 2009. Constitutional (germline) *MLH1* epimutation as an aetiological mechanism for hereditary non-polyposis colorectal cancer. J Med Genet 46:793–802.

Homolova K, Zavadakova P, Doktor TK, Schroeder LD, Kozich V, Andresen BS. 2010. The deep intronic c.903+469T>C mutation in the *MTRR* gene creates an SF2/ASF binding exonic splicing enhancer, which leads to pseudoexon activation and causes the cblE type of homocystinuria. Hum Mutat 31:437–444.

Howard OM, Turpin JA, Goldman R, Modi WS. 2004. Functional redundancy of the human *CCL4* and *CCL4L1* chemokine genes. Biochem Biophys Res Commun 320:927–931.

Hsiao T-L, Vitkup D. 2008. Role of duplicate genes in robustness against deleterious human mutations. PLoS Genet 4:e1000014.

Hu Z, Chen J, Tian T, Zhou X, Gu H, Xu L, Zeng Y, Miao R, Jin G, Ma H, Chen Y, Shen H. 2008. Genetic variants of miRNA sequences and non-small cell lung cancer survival. J Clin Invest 118:2600–2608.

Hu Z, Liang J, Wang Z, Tian T, Zhou X, Chen J, Miao R, Wang Y, Wang X, Shen H. 2009. Common genetic variants in pre-microRNAs were associated with increased risk of breast cancer in Chinese women. Hum Mutat 30:79–84.

Huang H, Winter EE, Wang H, Weinstock KG, Xing H, Goodstadt L, Stenson PD, Cooper DN, Smith D, Albà MM, Ponting CP, Fechtel K. 2004. Evolutionary conservation and selection of human disease gene orthologs in the rat and mouse genomes. Genome Biol 5:R47.

Hunt R, Sauna ZE, Ambudkar SV, Gottesman MM, Kimchi-Sarfaty C. 2009. Silent (synonymous) SNPs: should we care about them? Methods Mol Biol 578:23–39.

Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y, Scherer SW, Lee C. 2004. Detection of large-scale variation in the human genome. Nat Genet 36:949–951.

International Human Genome Sequencing Consortium. 2004. Finishing the euchromatic sequence of the human genome. Nature 431:931–945.

Jarinova O, Stewart AF, Roberts R, Wells G, Lau P, Naing T, Buerki C, McLean BW, Cook RC, Parker JS, McPherson R. 2009. Functional analysis of the chromosome 9p21.3 coronary artery disease risk locus. Arterioscler Thromb Vasc Biol 29:1671–1677.

Jazdzewski K, Murray EL, Franssila K, Jarzab B, Schoenberg DR, de la Chapelle A. 2008. Common SNP in pre-miR-146a decreases mature miR expression and predisposes to papillary thyroid carcinoma. Proc Natl Acad Sci USA 105:7269–7274.

Jimenez-Sanchez G, Childs B, Valle D. 2001. Human disease genes. Nature 409:853–855.

Jones S, Hruban RH, Kamiyama M, Borges M, Zhang X, Parsons DW, Lin JC, Palmisano E, Brune K, Jaffee EM, Iacobuzio-Donahue CA, Maitra A,

Parmigiani G, Kern SE, Velculescu VE, Kinzler KW, Vogelstein B, Eshleman JR, Goggins M, Klein AP. 2009. Exomic sequencing identifies *PALB2* as a pancreatic cancer susceptibility gene. Science 324:217.

Jordan IK, Rogozin IB, Glazko GV, Koonin EV. 2003. Origin of a substantial fraction of human regulatory sequences from transposable elements. Trends Genet 19:68–72.

Kapranov P, Cawley SE, Drenkow J, Bekiranov S, Strausberg RL, Fodor SP, Gingeras TR. 2002. Large-scale transcriptional activity in chromosomes 21 and 22. Science 296:916–919.

Kapranov P, Cheng J, Dike S, Nix DA, Duttagupta R, Willingham AT, Stadler PF, Hertel J, Hackermüller J, Hofacker IL, Bell I, Cheung E, Drenkow J, Dumais E, Patel S, Helt G, Ganesh M, Ghosh S, Piccolboni A, Sementchenko V, Tammana H, Gingeras TR. 2007b. RNA maps reveal new RNA classes and a possible function for pervasive transcription. Science 316:1484–1488.

Kapranov P, Drenkow J, Cheng J, Long J, Helt G, Dike S, Gingeras TR. 2005. Examples of the complex architecture of the human transcriptome revealed by RACE and high-density tiling arrays. Genome Res 15:987–997.

Kapranov P, Willingham AT, Gingeras TR. 2007a. Genome-wide transcription and the implications for genomic organization. Nat Rev Genet 8:413–423.

Karchin R. 2009. Next generation tools for the annotation of human SNPs. Brief Bioinform 10:35–52.

Kasowski M, Grubert F, Heffelfinger C, Hariharan M, Asabere A, Waszak SM, Habegger L, Rozowsky J, Shi M, Urban AE, Hong MY, Karczewski KJ, Huber W, Weissman SM, Gerstein MB, Korbel JO, Snyder M. 2010. Variation in transcription factor binding among humans. Science 328:232–235.

Katayama S, Tomaru Y, Kasukawa T, Waki K, Nakanishi M, Nakamura M, Nishida H, Yap CC, Suzuki M, Kawai J, Suzuki H, Carninci P, Hayashizaki Y, Wells C, Frith M, Ravasi T, Pang KC, Hallinan J, Mattick J, Hume DA, Lipovich L, Batalov S, Engström PG, Mizuno Y, Faghihi MA, Sandelin A, Chalk AM, Mottagui-Tabar S, Liang Z, Lenhard B, Wahlestedt C, RIKEN Genome Exploration Research Group, Genome Science Group (Genome Network Project Core Group), FANTOM Consortium. 2005. Antisense transcription in the mammalian transcriptome. Science 309:1564–1566.

Katzman S, Kern AD, Bejerano G, Fewell G, Fulton L, Wilson RK, Salama SR, Haussler D. 2007. Human genome ultraconserved elements are ultraselected. Science 317:915.

Kawaji H, Hayashizaki Y. 2008. Exploration of small RNAs. PLoS Genet 4:e22.

Keightley PD, Kryukov GV, Sunyaev S, Halligan DL, Gaffney DJ. 2005. Evolutionary constraints in conserved nongenic sequences of mammals. Genome Res 15:1373–1378.

Khachane AN, Harrison PM. 2009. Assessing the genomic evidence for conserved transcribed pseudogenes under selection. BMC Genomics 10:435.

Khalil AM, Guttman M, Huarte M, Garber M, Raj A, Rivea Morales D, Thomas K, Presser A, Bernstein BE, van Oudenaarden A, Regev A, Lander ES, Rinn JL. 2009. Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. Proc Natl Acad Sci USA 106:11667–11672.

Kidd JM, Cooper GM, Donahue WF, Hayden HS, Sampas N, Graves T, Hansen N, Teague B, Alkan C, Antonacci F, Haugen E, Zerr T, Yamada NA, Tsang P, Newman TL, Tüzün E, Cheng Z, Ebling HM, Tusneem N, David R, Gillett W, Phelps KA, Weaver M, Saranga D, Brand A, Tao W, Gustafson E, McKernan K, Chen L, Malig M, Smith JD, Korn JM, McCarroll SA, Altshuler DA, Peiffer DA, Dorschner M, Stamatoyannopoulos J, Schwartz D, Nickerson DA, Mullikin JC, Wilson RK, Bruhn L, Olson MV, Kaul R, Smith DR, Eichler EE. 2008. Mapping and sequencing of structural variation from eight human genomes. Nature 453:56–64.

Kim JI, Ju YS, Park H, Kim S, Lee S, Yi JH, Mudge J, Miller NA, Hong D, Bell CJ, Kim HS, Chung IS, Lee WC, Lee JS, Seo SH, Yun JY, Woo HN, Lee H, Suh D, Lee S, Kim HJ, Yavartanoo M, Kwak M, Zheng Y, Lee MK, Park H, Kim JY, Gokcumen O, Mills RE, Zaranek AW, Thakuria J, Wu X, Kim RW, Huntley JJ, Luo S, Schroth GP, Wu TD, Kim H, Yang KS, Park WY, Kim H, Church GM, Lee C, Kingsmore SF, Seo JS. 2009. A highly annotated whole-genome sequence of a Korean individual. Nature 460:1011–1015.

Kimchi-Sarfaty C, Oh JM, Kim IW, Sauna ZE, Calcagno AM, Ambudkar SV, Gottesman MM. 2007. A "silent" polymorphism in the *MDR1* gene changes substrate specificity. Science 315:525–528.

Kimura M. 1985. The role of compensatory neutral mutations in molecular evolution. J Genet 64:7–19.

Kleinjan DA, Lettice LA. 2008. Long-range gene control and genetic disease. Adv Genet 61:339–388.

Kleinjan DJ, Coutinho P. 2009. Cis-ruption mechanisms: disruption of cis-regulatory control as a cause of human genetic disease. Brief Funct Genomic Proteomic 8:317–332.

Klooster R, Straasheijm K, Shah B, Sowden J, Frants R, Thornton C, Tawil R, van der Maarel S. 2009. Comprehensive expression analysis of FSHD candidate genes at the mRNA and protein level. Eur J Hum Genet 17:1615–1624.

Koenig SC, Becirevic E, Hellberg MS, Li MY, So JC, Hankins JS, Ware RE, McMahon L, Steinberg MH, Luo HY, Chui DH. 2009. Sickle cell disease caused by heterozygosity for Hb S and novel LCR deletion: Report of two patients. Am J Hematol 84:603–606.

Kondrashov AS, Sunyaev S, Kondrashov FA. 2002. Dobzhansky-Muller incompatibilities in protein evolution. Proc Natl Acad Sci USA 99:14878–14883.

Krawczak M, Thomas NS, Hundrieser B, Mort M, Wittig M, Hampe J, Cooper DN. 2007. Single base-pair substitutions in exon-intron junctions of human genes: nature, distribution, and consequences for mRNA splicing. Hum Mutat 28:150–158.

Kryukov GV, Pennacchio LA, Sunyaev SR. 2007. Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. Am J Hum Genet 80:727–739.

Kryukov GV, Schmidt S, Sunyaev S. 2005. Small fitness effect of mutations in highly conserved non-coding regions. Hum Mol Genet 14:2221–2229.

Kryukov GV, Shpunt A, Stamatoyannopoulos JA, Sunyaev SR. 2009. Power of deep, all-exon resequencing for discovery of human trait genes. Proc Natl Acad Sci USA 106:3871–3876.

Kumar S, Suleski MP, Markov GJ, Lawrence S, Marco A, Filipski AJ. 2009. Positional conservation and amino acids shape the correct diagnosis and population frequencies of benign and damaging personal amino acid mutations. Genome Res 19:1562–1569.

Kwan T, Benovoy D, Dias C, Gurd S, Provencher C, Beaulieu P, Hudson TJ, Sladek R, Majewski J. 2008. Genome-wide analysis of transcript isoform variation in humans. Nat Genet 40:225–231.

Lage K, Hansen NT, Karlberg EO, Eklund AC, Roque FS, Donahoe PK, Szallasi Z, Jensen TS, Brunak S. 2008. A large-scale analysis of tissue-specific pathology and gene expression of human disease genes and complexes. Proc Natl Acad Sci USA 105:20870–20875.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann N, Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin JC, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston RH, Wilson RK, Hillier LW, McPherson JD, Marra MA, Mardis ER, Fulton LA, Chinwalla AT, Pepin KH, Gish WR, Chissoe SL, Wendl MC, Delehaunty KD, Miner TL, Delehaunty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng JF, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier M, Gibbs RA, Muzny DM, Scherer SE, Bouck JB, Sodergren EJ, Worley KC, Rives CM, Gorrell JH, Metzker ML, Naylor SL, Kucherlapati RS, Nelson DL, Weinstock GM, Sakaki Y, Fujiyama A, Hattori M, Yada T, Toyoda A, Itoh T, Kawagoe C, Watanabe H, Totoki Y, Taylor T, Weissenbach J, Heilig R, Saurin W, Artiguenave F, Brottier P, Bruls T, Pelletier E, Robert C, Wincker P, Smith DR, Doucette-Stamm L, Rubenfield M, Weinstock K, Lee HM, Dubois J, Rosenthal A, Platzer M, Nyakatura G, Taudien S, Rump A, Yang H, Yu J, Wang J, Huang G, Gu J, Hood L, Rowen L, Madan A, Qin S, Davis RW, Federspiel NA, Abola AP, Proctor MJ, Myers RM, Schmutz J, Dickson M, Grimwood J, Cox DR, Olson MV, Kaul R, Raymond C, Shimizu N, Kawasaki K, Minoshima S, Evans GA, Athanasiou M, Schultz R, Roe BA, Chen F, Pan H, Ramser J, Lehrach H, Reinhardt R, McCombie WR, de la Bastide M, Dedhia N, Blöcker H, Hornischer K, Nordsiek G, Agarwala R, Aravind L, Bailey JA, Bateman A, Batzoglou S, Birney E, Bork P, Brown DG, Burge CB, Cerutti L, Chen HC, Church D, Clamp M, Copley RR, Doerks T, Eddy SR, Eichler EE, Furey TS, Galagan J, Gilbert JG, Harmon C, Hayashizaki Y, Haussler D, Hermjakob H, Hokamp K, Jang W, Johnson LS, Jones TA, Kasif S, Kaspryzk A, Kennedy S, Kent WJ, Kitts P, Koonin EV, Korf I, Kulp D, Lancet D, Lowe TM, McLysaght A, Mikkelsen T, Moran JV, Mulder N, Pollara VJ, Ponting CP, Schuler G, Schultz J, Slater G, Smit AF, Stupka E, Szustakowski J, Thierry-Mieg D, Thierry-Mieg J, Wagner L, Wallis J, Wheeler R, Williams A, Wolf YI, Wolfe KH, Yang SP, Yeh RF, Collins F, Guyer MS, Peterson J, Felsenfeld A, Wetterstrand KA, Patrinos A, Morgan MJ, de Jong P, Catanese JJ, Osoegawa K, Shizuya H, Choi S, Chen YJ, International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. Nature 409:860–921.

Lauderdale JD, Wilensky JS, Oliver ER, Walton DS, Glaser T. 2000. 3′ deletions cause aniridia by preventing *PAX6* gene expression. Proc Natl Acad Sci USA 97:13755–13759.

Lecointre C, Pichon O, Hamel A, Heloury Y, Michel-Calemard L, Morel Y, David A, Le Caignec C. 2009. Familial acampomelic form of campomelic dysplasia caused by a 960 kb deletion upstream of *SOX9*. Am J Med Genet 149A:1183–1189.

Lee C, Scherer SW. 2010. The clinical context of copy number variation in the human genome. Expert Rev Mol Med 12:e8.

Lettice LA, Heaney SJ, Purdie LA, Li L, de Beer P, Oostra BA, Goode D, Elgar G, Hill RE, de Graaff E. 2003. A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. Hum Mol Genet 12:1725–1735.

Lettice LA, Horikoshi T, Heaney SJ, van Baren MJ, van der Linde HC, Breedveld GJ, Joosse M, Akarsu N, Oostra BA, Endo N, Shibata M, Suzuki M, Takahashi E, Shinka T, Nakahori Y, Ayusawa D, Nakabayashi K, Scherer SW, Heutink P, Hill RE, Noji S. 2002. Disruption of a long-range *cis*-acting regulator for Shh causes preaxial polydactyly. Proc Natl Acad Sci USA 99:7548–7553.

Levy S, Sutton G, Ng PC, Feuk L, Halpern AL, Walenz BP, Axelrod N, Huang J, Kirkness EF, Denisov G, Lin Y, MacDonald JR, Pang AW, Shago M, Stockwell TB, Tsiamouri A, Bafna V, Bansal V, Kravitz SA, Busam DA, Beeson KY, McIntosh TC, Remington KA, Abril JF, Gill J, Borman J, Rogers YH, Frazier ME, Scherer SW, Strausberg RL, Venter JC. 2007. The diploid genome sequence of an individual human. PLoS Biol 5:e254.

Lewinsky RH, Jensen TG, Møller J, Stensballe A, Olsen J, Troelsen JT. 2005. T-13910 DNA variant associated with lactase persistence interacts with Oct-1 and stimulates lactase promoter activity in vitro. Hum Mol Genet 14:3945–3953.

Li B, Krishnan VG, Mort ME, Xin F, Kamati KK, Cooper DN, Mooney SD, Radivojac P. 2009a. Automated inference of molecular mechanisms of disease from amino acid substitutions. Bioinformatics 25:2744–2750.

Li H, Xie H, Liu W, Hu R, Huang B, Tan YF, Xu K, Sheng ZF, Zhou HD, Wu XP, Luo XH. 2009d. A novel microRNA targeting *HDAC5* regulates osteoblast differentiation in mice and contributes to primary osteoporosis in humans. J Clin Invest 119:3666–3677.

Li JB, Levanon EY, Yoon JK, Aach J, Xie B, Leproust E, Zhang K, Gao Y, Church GM. 2009b. Genome-wide identification of human RNA editing sites by parallel DNA capturing and sequencing. Science 324:1210–1213.

Li R, Li Y, Zheng H, Luo R, Zhu H, Li Q, Qian W, Ren Y, Tian G, Li J, Zhou G, Zhu X, Wu H, Qin J, Jin X, Li D, Cao H, Hu X, Blanche H, Cann H, Zhang X, Li S, Bolund L, Kristiansen K, Yang H, Wang J, Wang J. 2010a. Building the sequence map of the human pan-genome. Nat Biotechnol 28:57–63.

Li SC, Chan WC, Hu LY, Lai CH, Hsu CN, Lin WC. 2010b. Identification of homologous microRNAs in 56 animal genomes. Genomics (in press).

Li W, Duan R, Kooy F, Sherman SL, Zhou W, Jin P. 2009c. Germline mutation of microRNA-125a is associated with breast cancer. J Med Genet 46:358–360.

Li YC, Korol AB, Fahima T, Nevo E. 2004. Microsatellites within genes: structure, function, and evolution. Mol Biol Evol 21:991–1007.

Liang H, Li W-H. 2009. Functional compensation by duplicated genes in mouse. Trends Genet 25:441–442.

Liao BY, Zhang J. 2008. Null mutations in human and mouse orthologs frequently result in different phenotypes. Proc Natl Acad Sci USA 105:6987–6992.

Licastro D, Gennarino VA, Petrera F, Sanges R, Banfi S, Stupka E. 2010. Promiscuity of enhancer, coding and non-coding transcription functions in ultraconserved elements. BMC Genomics 11:151.

Lin L, Jiang P, Shen S, Sato S, Davidson BL, Xing Y. 2009. Large-scale analysis of exonized mammalian-wide interspersed repeats in primate genomes. Hum Mol Genet 18:2204–2214.

Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, Nery JR, Lee L, Ye Z, Ngo QM, Edsall L, Antosiewicz-Bourget J, Stewart R, Ruotti V, Millar AH, Thomson JA, Ren B, Ecker JR. 2009. Human DNA methylomes at base resolution show widespread epigenomic differences. Nature 462:315–322.

Liu JC, Makova KD, Adkins RM, Gibson S, Li WH. 2001. Episodic evolution of growth hormone in primates and emergence of the species specificity of human growth hormone receptor. Mol Biol Evol 18:945–953.

Liu W, Sun J, Li G, Zhu Y, Zhang S, Kim ST, Sun J, Wiklund F, Wiley K, Isaacs SD, Stattin P, Xu J, Duggan D, Carpten JD, Isaacs WB, Grönberg H, Zheng SL, Chang BL. 2009. Association of a germ-line copy number variation at 2p24.3 and risk for aggressive prostate cancer. Cancer Res 69: 2176–2179.

Lohmueller KE, Indap AR, Schmidt S, Boyko AR, Hernandez RD, Hubisz MJ, Sninsky JJ, White TJ, Sunyaev SR, Nielsen R, Clark AG, Bustamante CD. 2008. Proportionally more deleterious genetic variation in European than in African populations. Nature 451:994–997.

Lomelin D, Jorgenson E, Risch N. 2010. Human genetic variation recognizes functional elements in non-coding sequence. Genome Res 20:311–319.

López-Bigas N, Audit B, Ouzounis C, Parra G, Guigó R. 2005. Are splicing mutations the most frequent cause of hereditary disease? FEBS Lett 579:1900–1903.

López-Bigas N, Blencowe BJ, Ouzounis CA. 2006. Highly consistent patterns for inherited human diseases at the molecular level. Bioinformatics 22:269–277.

López-Bigas N, Ouzounis CA. 2004. Genome-wide identification of genes likely to be involved in human genetic disease. Nucleic Acids Res 32:3108–3114.

Louro R, Smirnova AS, Verjovski-Almeida S. 2009. Long intronic noncoding RNA transcription: expression noise or expression choice? Genomics 93:291–298.

Lowe CB, Bejerano G, Haussler D. 2007. Thousands of human mobile element fragments undergo strong purifying selection near developmental genes. Proc Natl Acad Sci USA 104:8005–8010.

Lower KM, Hughes JR, De Gobbi M, Henderson S, Viprakasit V, Fisher C, Goriely A, Ayyub H, Sloane-Stanley J, Vernimmen D, Langford C, Garrick D, Gibbons RJ, Higgs DR. 2009. Adventitious changes in long-range gene expression caused by polymorphic structural variation and promoter competition. Proc Natl Acad Sci USA 106:21771–21776.

Lualdi S, Tappino B, Di Duca M, Dardis A, Anderson CJ, Biassoni R, Thompson PW, Corsolini F, Di Rocco M, Bembi B, Regis S, Cooper DN, Filocamo M. 2010. Enigmatic in vivo iduronate-2-sulfatase (*IDS*) mutant transcript correction to wild-type in Hunter syndrome. Hum Mutat 31:E1261–E1285.

Luco RF, Pan Q, Tominaga K, Blencowe BJ, Pereira-Smith OM, Misteli T. 2010. Regulation of alternative splicing by histone modifications. Science 327:996–1000.

Lupski JR, Reid JG, Gonzaga-Jauregui C, Rio Deiros D, Chen DC, Nazareth L, Bainbridge M, Dinh H, Jing C, Wheeler DA, McGuire AL, Zhang F, Stankiewicz P, Halperin JJ, Yang C, Gehman C, Guo D, Irikat RK, Tom W, Fantin NJ, Muzny DM, Gibbs RA. 2010. Whole-genome sequencing in a patient with Charcot-Marie-Tooth neuropathy. New Engl J Med 362:1181–1191.

Lynch M. 2010. Rate, molecular spectrum, and consequences of human mutation. Proc Natl Acad Sci USA 107:961–968.

McCarroll SA. 2008. Extending genome-wide association studies to copy-number variation. Hum Mol Genet 17(R2):R135–142.

McKernan KJ, Peckham HE, Costa GL, McLaughlin SF, Fu Y, Tsung EF, Clouser CR, Duncan C, Ichikawa JK, Lee CC, Zhang Z, Ranade SS, Dimalanta ET, Hyland FC, Sokolsky TD, Zhang L, Sheridan A, Fu H, Hendrickson CL, Li B, Kotler L, Stuart JR, Malek JA, Manning JM, Antipova AA, Perez DS, Moore MP, Hayashibara KC, Lyons MR, Beaudoin RE, Coleman BE, Laptewicz MW, Sannicandro AE, Rhodes MD, Gottimukkala RK, Yang S, Bafna V, Bashir A, MacBride A, Alkan C, Kidd JM, Eichler EE, Reese MG, De La Vega FM, Blanchard AP. 2009. Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. Genome Res 19:1527–1541.

McLean C, Bejerano G. 2008. Dispensability of mammalian DNA. Genome Res 18:1743–1751.

Mann V, Hobson EE, Li B, Stewart TL, Grant SF, Robins SP, Aspden RM, Ralston SH. 2001. A *COL1A1* Sp1 binding site polymorphism predisposes to osteoporotic fracture by affecting bone density and quality. J Clin Invest 107:899–907.

Martin MM, Buckenberger JA, Jiang J, Malana GE, Nuovo GJ, Chotani M, Feldman DS, Schmittgen TD, Elton TS. 2007. The human angiotensin II type 1 receptor +1166 A/C polymorphism attenuates microrna-155 binding. J Biol Chem 282:24262–24269.

Maston GA, Evans SK, Green MR. 2006. Transcriptional regulatory elements in the human genome. Annu Rev Genomics Hum Genet 7:29–59.

Mattick JS. 2009. The genetic signatures of noncoding RNAs. PloS Genet 5:e1000459.

Medvedeva YA, Fridman MV, Oparina NJ, Malko DB, Ermakova EO, Kulakovskiy IV, Heinzel A, Makeev VJ. 2010. Intergenic, gene terminal, and intragenic CpG islands in the human genome. BMC Genomics 11:48.

Mefford HC. 2009. Genotype to phenotype-discovery and characterization of novel genomic disorders in a "genotype-first" era. Genet Med 11:836–842.

Mefford HC, Eichler EE. 2009. Duplication hotspots, rare genomic disorders, and common disease. Curr Opin Genet Dev 19:196–204.

Mencía A, Modamio-Hŷbjør S, Redshaw N, Morín M, Mayo-Merino F, Olavarrieta L, Aguirre LA, del Castillo I, Steel KP, Dalmay T, Moreno F, Moreno-Pelayo MA. 2009. Mutations in the seed region of human miR-96 are responsible for nonsyndromic progressive hearing loss. Nat Genet 41: 609–613.

Mercer TR, Dinger ME, Mattick JS. 2009. Long non-coding RNAs: insights into functions. Nat Rev Genet 10:155–159.

Merikangas AK, Corvin AP, Gallagher L. 2009. Copy-number variants in neurodevelopmental disorders: promises and challenges. Trends Genet 25:536–544.

Miller MP, Kumar S. 2001. Understanding human disease mutations through the use of interspecific genetic variation. Hum Mol Genet 10:2319–2328.

Miller MP, Parker JD, Rissing SW, Kumar S. 2003. Quantifying the intragenic distribution of human disease mutations. Ann Hum Genet 67:567–579.

Mills RE, Bennett EA, Iskow RC, Devine SE. 2007. Which transposable elements are active in the human genome? Trends Genet 23:183–191.

Morley M, Molony CM, Weber TM, Devlin JL, Ewens KG, Spielman RS, Cheung VG. 2004. Genetic analysis of genome-wide variation in human gene expression. Nature 430:743–747.

Morris JA. 2001. How many deleterious mutations are there in the human genome? Med Hypotheses 56:646–652.

Mort M, Evani US, Krishnan VG, Kamati KK, Baenziger PH, Bagchi A, Peters B, Sathyesh R, Li B, Sun Y, Xue B, Shah N, Kann M, Cooper DN, Radivojac P, Mooney SD. 2010. *In silico* functional profiling of human disease-associated and polymorphic amino acid substitutions. Hum Mutat 31:335–346.

Nackley AG, Shabalina SA, Lambert JE, Conrad MS, Gibson DG, Spiridonov AN, Satterfield SK, Diatchenko L. 2009. Low enzymatic activity haplotypes of the human catechol-O-methyltransferase gene: enrichment for marker SNPs. PLoS One 4:e5237.

Nackley AG, Shabalina SA, Tchivileva IE, Satterfield K, Korchynskyi O, Makarov SS, Maixner W, Diatchenko L. 2006. Human catechol-O-methyltransferase haplotypes modulate protein expression by altering mRNA secondary structure. Science 314:1930–1933.

Ng PC, Henikoff S. 2006. Predicting the effects of amino acid substitutions on protein function. Annu Rev Genomics Hum Genet 7:61–80.

Ng PC, Levy S, Huang J, Stockwell TB, Walenz BP, Li K, Axelrod N, Busam DA, Strausberg RL, Venter JC. 2008. Genetic variation in an individual human exome. PLoS Genet 4:e1000160.

Ng SB, Buckingham KJ, Lee C, Bigham AW, Tabor HK, Dent KM, Huff CD, Shannon PT, Jabs EW, Nickerson DA, Shendure J, Bamshad MJ. 2010. Exome sequencing identifies the cause of a mendelian disorder. Nat Genet 42:30–35.

Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, Lee C, Shaffer T, Wong M, Bhattacharjee A, Eichler EE, Bamshad M, Nickerson DA, Shendure J. 2009. Targeted capture and massively parallel sequencing of 12 human exomes. Nature 461:272–276.

Nikopoulos K, Gilissen C, Hoischen A, van Nouhuys CE, Boonstra FN, Blokland EA, Arts P, Wieskamp N, Strom TM, Ayuso C, Tilanus MA, Bouwhuis S, Mukhopadhyay A, Scheffer H, Hoefsloot LH, Veltman JA, Cremers FP, Collin RW. 2010. Next-generation sequencing of a 40 Mb linkage interval reveals *TSPAN12* mutations in patients with familial exudative vitreoretinopathy. Am J Hum Genet 86:240–247.

Nishihara H, Smit AF, Okada N. 2006. Functional noncoding sequences derived from SINEs in the mammalian genome. Genome Res 16:864–874.

Nozu K, Iijima K, Nozu Y, Ikegami E, Imai T, Fu XJ, Kaito H, Nakanishi K, Yoshikawa N, Matsuo M. 2009. A deep intronic mutation in the *SLC12A3* gene leads to Gitelman syndrome. Pediatr Res 66:590–593.

Ogino W, Takeshima Y, Nishiyama A, Okizuka Y, Yagi M, Tsuneishi S, Saiki K, Kugo M, Matsuo M. 2007. Mutation analysis of the ornithine transcarbamylase (*OTC*) gene in five Japanese OTC deficiency patients revealed two known and three novel mutations including a deep intronic mutation. Kobe J Med Sci 53:229–240.

Olds LC, Sibley E. 2003. Lactase persistence DNA variant enhances lactase promoter activity *in vitro*: functional role as a *cis* regulatory element. Hum Mol Genet 12:2333–2340.

Orkin SH, Alter BP, Altay C, Mahoney MJ, Lazarus H, Hobbins JC, Nathan DG. 1978. Application of endonuclease mapping to the analysis and prenatal diagnosis of thalassemias caused by globin-gene deletion. N Engl J Med 299:166–172.

Osada N, Mano S, Gojobori J. 2009. Quantifying dominance and deleterious effect on human disease genes. Proc Natl Acad Sci USA 106:841–846.

Ovcharenko I, Loots GG, Nobrega MA, Hardison RC, Miller W, Stubbs L. 2005. Evolution and functional classification of vertebrate gene deserts. Genome Res 15:137–145.

Ozsolak F, Platt AR, Jones DR, Reifenberger JG, Sass LE, McInerney P, Thompson JF, Bowers J, Jarosz M, Milos PM. 2009. Direct RNA sequencing. Nature 461:814–818.

Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ. 2008. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat Genet 40:1413–1415.

Park D, Park J, Park SG, Park T, Choi SS. 2008. Analysis of human disease genes in the context of gene essentiality. Genomics 92:414–418.

Park PJ. 2009. ChIP-seq: advantages and challenges of a maturing technology. Nat Rev Genet 10:669–680.

Parker SC, Hansen L, Abaan HO, Tullius TD, Margulies EH. 2009. Local DNA topography correlates with functional noncoding regions of the human genome. Science 324:389–392.

Parra G, Reymond A, Dabbouseh N, Dermitzakis ET, Castelo R, Thomson TM, Antonarakis SE, Guigó R. 2006. Tandem chimerism as a means to increase protein complexity in the human genome. Genome Res 16:37–44.

Pastinen T, Ge B, Hudson TJ. 2006. Influence of human genome polymorphism on gene expression. Hum Mol Genet 15 Spec No 1:R9–R16.

Pennacchio LA, Ahituv N, Moses AM, Prabhakar S, Nobrega MA, Shoukry M, Minovitsky S, Dubchak I, Holt A, Lewis KD, Plajzer-Frick I, Akiyama J, De Val S, Afzal V, Black BL, Couronne O, Eisen MB, Visel A, Rubin EM. 2006. In vivo enhancer analysis of human conserved non-coding sequences. Nature 444:499–502.

Perotti D, De Vecchi G, Testi MA, Lualdi E, Modena P, Mondini P, Ravagnani F, Collini P, Di Renzo F, Spreafico F, Terenziani M, Sozzi G, Fossati-Bellani F, Radice P. 2004. Germline mutations of the *POU6F2* gene in Wilms tumors with loss of heterozygosity on chromosome 7p14. Hum Mutat 24:400–407.

Pesole G. 2008. What is a gene? An updated operational definition. Gene 417:1–4.

Peters BA, St Croix B, Sjöblom T, Cummins JM, Silliman N, Ptak J, Saha S, Kinzler KW, Hatzis C, Velculescu VE. 2007. Large-scale identification of novel transcripts in the human genome. Genome Res 17:287–292.

Pheasant M, Mattick JS. 2007. Raising the estimate of functional human sequences. Genome Res 17:1245–1253.

Piriyapongsa J, Mariño-Ramírez L, Jordan IK. 2007. Origin and evolution of human microRNAs from transposable elements. Genetics 176:1323–1337.

Pomerantz MM, Ahmadiyeh N, Jia L, Herman P, Verzi MP, Doddapaneni H, Beckwith CA, Chan JA, Hills A, Davis M, Yao K, Kehoe SM, Lenz HJ, Haiman CA, Yan C, Henderson BE, Frenkel B, Barretina J, Bass A, Tabernero J, Baselga J, Regan MM, Manak JR, Shivdasani R, Coetzee GA, Freedman ML. 2009. The 8q24 cancer risk variant rs6983267 shows long-range interaction with *MYC* in colorectal cancer. Nat Genet 41:882–884.

Ponjavic J, Ponting CP, Lunter G. 2007. Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs. Genome Res 17:556–565.

Ponting CP, Lunter G. 2006. Signatures of adaptive evolution within human non-coding sequence. Hum Mol Genet 15:R170–R175.

Poon A, Davis BH, Chao L. 2005. The coupon collector and the suppressor mutation: estimating the number of compensatory mutations by maximum likelihood. Genetics 170:1323–1332.

Prabhakar S, Noonan JP, Pääbo S, Rubin EM. 2006. Accelerated evolution of conserved noncoding sequences in humans. Science 314:786.

Prabhu S, Pe'er I. 2009. Overlapping pools for high-throughput targeted resequencing. Genome Res 19:1254–1261.

Preker P, Nielsen J, Kammler S, Lykke-Andersen S, Christensen MS, Mapendano CK, Schierup MH, Jensen TH. 2008. RNA exosome depletion reveals transcription upstream of active human promoters. Science 322:1851–1854.

Pros E, Gómez C, Martín T, Fábregas P, Serra E, Lázaro C. 2008. Nature and mRNA effect of 282 different *NF1* point mutations: focus on splicing alterations. Hum Mutat 29:E173–E193.

Pruitt KD, Harrow J, Harte RA, Wallin C, Diekhans M, Maglott DR, Searle S, Farrell CM, Loveland JE, Ruef BJ, Hart E, Suner MM, Landrum MJ, Aken B, Ayling S, Baertsch R, Fernandez-Banet J, Cherry JL, Curwen V, Dicuccio M, Kellis M, Lee J, Lin MF, Schuster M, Shkeda A, Amid C, Brown G, Dukhanina O, Frankish A, Hart J, Maidak BL, Mudge J, Murphy MR, Murphy T, Rajan J, Rajput B, Riddick LD, Snow C, Steward C, Webb D, Weber JA, Wilming L, Wu W, Birney E, Haussler D, Hubbard T, Ostell J, Durbin R, Lipman D. 2009. The consensus coding sequence (CCDS) project: identifying a common protein-coding gene set for the human and mouse genomes. Genome Res 19:1316–1323.

Purevsuren J, Fukao T, Hasegawa Y, Fukuda S, Kobayashi H, Yamaguchi S. 2008. Study of deep intronic sequence exonization in a Japanese neonate with a mitochondrial trifunctional protein deficiency. Mol Genet Metab 95:46–51.

Quemener S, Chen JM, Chuzhanova N, Bénech C, Casals T, MacekJr M, Bienvenu T, McDevitt T, Farrell PM, Loumi O, Messaoud T, Cuppens H, Cutting GR, Stenson PD, Giteau K, Audrézet MP, Cooper DN, Férec C. 2010. Complete ascertainment of intragenic copy number mutations (CNMs) in the *CFTR* gene and its implications for CNM formation at other autosomal loci. Hum Mutat 31:421–428.

Rademakers R, Eriksen JL, Baker M, Robinson T, Ahmed Z, Lincoln SJ, Finch N, Rutherford NJ, Crook RJ, Josephs KA, Boeve BF, Knopman DS, Petersen RC, Parisi JE, Caselli RJ, Wszolek ZK, Uitti RJ, Feldman H, Hutton ML, Mackenzie IR, Graff-Radford NR, Dickson DW. 2008. Common variation in the miR-659 binding-site of *GRN* is a major risk factor for TDP43-positive frontotemporal dementia. Hum Mol Genet 17:3631–3642.

Rahimov F, Marazita ML, Visel A, Cooper ME, Hitchler MJ, Rubini M, Domann FE, Govil M, Christensen K, Bille C, Melbye M, Jugessur A, Lie RT, Wilcox AJ, Fitzpatrick DR, Green ED, Mossey PA, Little J, Steegers-Theunissen RP, Pennacchio LA, Schutte BC, Murray JC. 2008. Disruption of an AP-2α binding site in an *IRF6* enhancer is associated with cleft lip. Nat Genet 40:1341–1347.

Rasmussen M, Li Y, Lindgreen S, Pedersen JS, Albrechtsen A, Moltke I, Metspalu M, Metspalu E, Kivisild T, Gupta R, Bertalan M, Nielsen K, Gilbert MT, Wang Y, Raghavan M, Campos PF, Kamp HM, Wilson AS, Gledhill A, Tridico S, Bunce M, Lorenzen ED, Binladen J, Guo X, Zhao J, Zhang X, Zhang H, Li Z, Chen M, Orlando L, Kristiansen K, Bak M, Tommerup N, Bendixen C, Pierre TL, Grønnow B, Meldgaard M, Andreasen C, Fedorova SA, Osipova LP, Higham TF, Ramsey CB, Hansen TV, Nielsen FC, Crawford MH, Brunak S, Sicheritz-Pontén T, Villems R, Nielsen R, Krogh A, Wang J, Willerslev E. 2010. Ancient human genome sequence of an extinct Palaeo-Eskimo. Nature 463:757–762.

Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, Fiegler H, Shapero MH, Carson AR, Chen W, Cho EK, Dallaire S, Freeman JL, González JR, Gratacòs M, Huang J, Kalaitzopoulos D, Komura D, MacDonald JR, Marshall CR, Mei R, Montgomery L, Nishimura K, Okamura K, Shen F, Somerville MJ, Tchinda J, Valsesia A, Woodwark C, Yang F, Zhang J, Zerjal T, Zhang J, Armengol L, Conrad DF, Estivill X, Tyler-Smith C, Carter NP, Aburatani H, Lee C, Jones KW, Scherer SW,

Hurles ME. 2006. Global variation in copy number in the human genome. Nature 444:444–454.

Reich DE, Lander ES. 2001. On the allelic spectrum of human disease. Trends Genet 17:502–510.

Rio Frio T, McGee TL, Wade NM, Iseli C, Beckmann JS, Berson EL, Rivolta C. 2009. A single-base substitution within an intronic repetitive element causes dominant retinitis pigmentosa with reduced penetrance. Hum Mutat 30:1340–1347.

Rincón A, Aguado C, Desviat LR, Sánchez-Alcudia R, Ugarte M, Pérez B. 2007. Propionic and methylmalonic acidemia: antisense therapeutics for intronic variations causing aberrantly spliced messenger RNA. Am J Hum Genet 81:1262–1270.

Roach JC, Glusman G, Smit AF, Huff CD, Hubley R, Shannon PT, Rowen L, Pant KP, Goodman N, Bamshad M, Shendure J, Drmanac R, Jorde LB, Hood L, Galas DJ. 2010. Analysis of genetic inheritance in a family quartet by whole-genome sequencing. Science 328:636–639.

Romao L, Osorio-Almeida L, Higgs DR, Lavinha J, Liebhaber SA. 1991. α-Thalassemia resulting from deletion of regulatory sequences far upstream of the alpha-globin structural genes. Blood 78:1589–1595.

Rozowsky JS, Newburger D, Sayward F, Wu J, Jordan G, Korbel JO, Nagalakshmi U, Yang J, Zheng D, Guigó R, Gingeras TR, Weissman S, Miller P, Snyder M, Gerstein MB. 2007. The DART classification of unannotated transcription within the ENCODE regions: associating transcription with known and novel loci. Genome Res 17:732–745.

Rung J, Cauchi S, Albrechtsen A, Shen L, Rocheleau G, Cavalcanti-Proença C, Bacot F, Balkau B, Belisle A, Borch-Johnsen K, Charpentier G, Dina C, Durand E, Elliott P, Hadjadj S, Järvelin MR, Laitinen J, Lauritzen T, Marre M, Mazur A, Meyre D, Montpetit A, Pisinger C, Posner B, Poulsen P, Pouta A, Prentki M, Ribel-Madsen R, Ruokonen A, Sandbaek A, Serre D, Tichet J, Vaxillaire M, Wojtaszewski JF, Vaag A, Hansen T, Polychronakos C, Pedersen O, Froguel P, Sladek R. 2009. Genetic variant near *IRS1* is associated with type 2 diabetes, insulin resistance and hyperinsulinemia. Nat Genet 41:1110–1115.

Sahoo T, del Gaudio D, German JR, Shinawi M, Peters SU, Person RE, Garnica A, Cheung SW, Beaudet AL. 2008. Prader-Willi phenotype caused by paternal deficiency for the HBII-85 C/D box small nucleolar RNA cluster. Nat Genet 40:719–721.

Sakai H, Koyanagi KO, Imanishi T, Itoh T, Gojobori T. 2007. Frequent emergence and functional resurrection of processed pseudogenes in the human and mouse genomes. Gene 389:196–203.

Sanford JR, Wang X, Mort M, Vanduyn N, Cooper DN, Mooney SD, Edenberg HJ, Liu Y. 2009. Splicing factor SFRS1 recognizes a functionally diverse landscape of RNA transcripts. Genome Res 19:381–394.

Schalkwyk LC, Meaburn EL, Smith R, Dempster EL, Jeffries AR, Davies MN, Plomin R, Mill J. 2010. Allelic skewing of DNA methylation is widespread across the genome. Am J Hum Genet 86:196–212.

Schollen E, Keldermans L, Foulquier F, Briones P, Chabas A, Sánchez-Valverde F, Adamowicz M, Pronicka E, Wevers R, Matthijs G. 2007. Characterization of two unusual truncating *PMM2* mutations in two CDG-Ia patients. Mol Genet Metab 90:408–413.

Schork NJ, Murray SS, Frazer KA, Topol EJ. 2009. Common vs. rare allele hypotheses for complex diseases. Curr Opin Genet Dev 19:212–219.

Schuster SC, Miller W, Ratan A, Tomsho LP, Giardine B, Kasson LR, Harris RS, Petersen DC, Zhao F, Qi J, Alkan C, Kidd JM, Sun Y, Drautz DI, Bouffard P, Muzny DM, Reid JG, Nazareth LV, Wang Q, Burhans R, Riemer C, Wittekindt NE, Moorjani P, Tindall EA, Danko CG, Teo WS, Buboltz AM, Zhang Z, Ma Q, Oosthuysen A, Steenkamp AW, Oostuisen H, Venter P, Gajewski J, Zhang Y, Pugh BF, Makova KD, Nekrutenko A, Mardis ER, Patterson N, Pringle TH, Chiaromonte F, Mullikin JC, Eichler EE, Hardison RC, Gibbs RA, Harkins TT, Hayes VM. 2010. Complete Khoisan and Bantu genomes from southern Africa. Nature 463:943–947.

Sebat J, Lakshmi B, Malhotra D, Troge J, Lese-Martin C, Walsh T, Yamrom B, Yoon S, Krasnitz A, Kendall J, Leotta A, Pai D, Zhang R, Lee YH, Hicks J, Spence SJ, Lee AT, Puura K, Lehtimäki T, Ledbetter D, Gregersen PK, Bregman J, Sutcliffe JS, Jobanputra V, Chung W, Warburton D, King MC, Skuse D, Geschwind DH, Gilliam TC, Ye K, Wigler M. 2007. Strong association of *de novo* copy number mutations with autism. Science 316:445–449.

Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin P, Månér S, Massa H, Walker M, Chi M, Navin N, Lucito R, Healy J, Hicks J, Ye K, Reiner A, Gilliam TC, Trask B, Patterson N, Zetterberg A, Wigler M. 2004. Large-scale copy number polymorphism in the human genome. Science 305:525–528.

Seila AC, Calabrese JM, Levine SS, Yeo GW, Rahl PB, Flynn RA, Young RA, Sharp PA. 2008. Divergent transcription from active promoters. Science 322:1849–1851.

Sethupathy P, Collins FS. 2008. MicroRNA target site polymorphisms and human disease. Trends Genet 24:489–497.

Shao X, Shepelev V, Fedorov A. 2006. Bioinformatic analysis of exon repetition, exon scrambling and trans-splicing in humans. Bioinformatics 22:692–698.

Shen J, Ambrosone CB, Zhao H. 2009. Novel genetic variants in microRNA genes and familial breast cancer. Int J Cancer 124:1178–1182.

Shen J, DiCioccio R, Odunsi K, Lele SB, Zhao H. 2010. Novel genetic variants in miR-191 gene and familial ovarian cancer. BMC Cancer 10:47.

Shlien A, Tabori U, Marshall CR, Pienkowska M, Feuk L, Novokmet A, Nanda S, Druker H, Scherer SW, Malkin D. 2008. Excessive genomic DNA copy number variation in the Li-Fraumeni cancer predisposition syndrome. Proc Natl Acad Sci USA 105:11264–11269.

Simon D, Laloo B, Barillot M, Barnetche T, Blanchard C, Rooryck C, Marche M, Burgelin I, Coupry I, Chassaing N, Gilbert-Dussardier B, Lacombe D, Grosset C, Arveiler B. 2010. A mutation in the 3′-UTR of the *HDAC6* gene abolishing the post-transcriptional regulation mediated by hsa-miR-433 is linked to a new form of dominant X-linked chondrodysplasia. Hum Mol Genet 19:2015–2027.

Smith NG, Eyre-Walker A. 2003. Human disease genes: patterns and predictions. Gene 318:169–175.

Smith AD, Xuan Z, Zhang MQ. 2008. Using quality scores and longer reads improves accuracy of Solexa read mapping. BMC Bioinformatics 9:128.

Snider L, Asawachaicharn A, Tyler AE, Geng LN, Petek LM, Maves L, Miller DG, Lemmers RJ, Winokur ST, Tawil R, van der Maarel SM, Filippova GN, Tapscott SJ. 2009. RNA transcripts, miRNA-sized fragments and proteins produced from D4Z4 units: new candidates for the pathophysiology of facioscapulohumeral dystrophy. Hum Mol Genet 18:2414–2430.

Snyder M, Du J, Gerstein M. 2010. Personal genome sequencing: current approaches and challenges. Genes Dev 24:423–431.

Solis AS, Shariat N, Patton JG. 2008. Splicing fidelity, enhancers, and disease. Front Biosci 13:1926–1942.

Spena S, Asselta R, Platé M, Castaman G, Duga S, Tenchini ML. 2007. Pseudo-exon activation caused by a deep-intronic mutation in the fibrinogen gamma-chain gene as a novel mechanism for congenital afibrinogenaemia. Br J Haematol 139:128–132.

Stankiewicz P, Sen P, Bhatt SS, Storer M, Xia Z, Bejjani BA, Ou Z, Wiszniewska J, Driscoll DJ, Maisenbacher MK, Bolivar J, Bauer M, Zackai EH, McDonald-McGinn D, Nowaczyk MM, Murray M, Hustead V, Mascotti K, Schultz R, Hallam L, McRae D, Nicholson AG, Newbury R, Durham-O'Donnell J, Knight G, Kini U, Shaikh TH, Martin V, Tyreman M, Simonic I, Willatt L, Paterson J, Mehta S, Rajan D, Fitzgerald T, Gribble S, Prigmore E, Patel A, Shaffer LG, Carter NP, Cheung SW, Langston C, Shaw-Smith C. 2009. Genomic and genic deletions of the FOX gene cluster on 16q24.1 and inactivating mutations of *FOXF1* cause alveolar capillary dysplasia and other malformations. Am J Hum Genet 84:780–791.

Stankiewicz P, Lupski JR. 2010. Structural variation in the human genome and its role in disease. Annu Rev Med 61:437–455.

Stefansson H, Rujescu D, Cichon S, Pietiläinen OP, Ingason A, Steinberg S, Fossdal R, Sigurdsson E, Sigmundsson T, Buizer-Voskamp JE, Hansen T, Jakobsen KD, Muglia P, Francks C, Matthews PM, Gylfason A, Halldorsson BV, Gudbjartsson D, Thorgeirsson TE, Sigurdsson A, Jonasdottir A, Jonasdottir A, Bjornsson A, Mattiasdottir S, Blondal T, Haraldsson M, Magnusdottir BB, Giegling I, Möller HJ, Hartmann A, Shianna KV, Ge D, Need AC, Crombie C, Fraser G, Walker N, Lonnqvist J, Suvisaari J, Tuulio-Henriksson A, Paunio T, Toulopoulou T, Bramon E, Di Forti M, Murray R, Ruggeri M, Vassos E, Tosato S, Walshe M, Li T, Vasilescu C, Mühleisen TW, Wang AG, Ullum H, Djurovic S, Melle I, Olesen J, Kiemeney LA, Franke B; GROUP, Sabatti C, Freimer NB, Gulcher JR, Thorsteinsdottir U, Kong A, Andreassen OA, Ophoff RA, Georgi A, Rietschel M, Werge T, Petursson H, Goldstein DB, Nöthen MM, Peltonen L, Collier DA, St Clair D, Stefansson K. 2008. Large recurrent microdeletions associated with schizophrenia. Nature 455:232–236.

Stenson PD, Mort M, Ball EV, Howells K, Phillips AD, Thomas NS, Cooper DN. 2009. The Human Gene Mutation Database: 2008 update. Genome Med 1:13.

Stranger BE, Forrest MS, Clark AG, Minichiello MJ, Deutsch S, Lyle R, Hunt S, Kahl B, Antonarakis SE, Tavaré S, Deloukas P, Dermitzakis ET. 2005. Genome-wide associations of gene expression variation in humans. PLoS Genet 1:e78.

Stranger BE, Nica AC, Forrest MS, Dimas A, Bird CP, Beazley C, Ingle CE, Dunning M, Flicek P, Koller D, Montgomery S, Tavaré S, Deloukas P, Dermitzakis ET. 2007. Population genomics of human gene expression. Nat Genet 39:1217–1224.

Su Z, Gu X. 2008. Predicting the proportion of essential genes in mouse duplicates based on biased mouse knockout genes. J Mol Evol 67:705–709.

Subramanian S, Kumar S. 2006. Evolutionary anatomies of positions and types of disease-associated and neutral amino acid mutations in the human genome. BMC Genomics 7:306.

Sultan M, Schulz MH, Richard H, Magen A, Klingenhoff A, Scherf M, Seifert M, Borodina T, Soldatov A, Parkhomchuk D, Schmidt D, O'Keeffe S, Haas S, Vingron M, Lehrach H, Yaspo ML. 2008. A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. Science 321:956–960.

Sun G, Yan J, Noltner K, Feng J, Li H, Sarkis DA, Sommer SS, Rossi JJ. 2009. SNPs in human miRNA genes affect biogenesis and function. RNA 15:1640–1651.

Suriano G, Azevedo L, Novais M, Boscolo B, Seruca R, Amorim A, Ghibaudi EM. 2007. In vitro demonstration of intra-locus compensation using the ornithine transcarbamylase protein as model. Hum Mol Genet 16:2209–2214.

Susa S, Daimon M, Sakabe J, Sato H, Oizumi T, Karasawa S, Wada K, Jimbu Y, Kameda W, Emi M, Muramatsu M, Kato T. 2008. A functional polymorphism of the TNF-α gene that is associated with type 2 DM. Biochem Biophys Res Commun 369:943–947.

Svensson O, Arvestad L, Lagergren J. 2006. Genome-wide survey for biologically functional pseudogenes. PLoS Comput. Biol 2:e46.

Taft RJ, Glazov EA, Cloonan N, Simons C, Stephen S, Faulkner GJ, Lassmann T, Forrest AR, Grimmond SM, Schroder K, Irvine K, Arakawa T, Nakamura M, Kubosaki A, Hayashida K, Kawazu C, Murata M, Nishiyori H, Fukuda S, Kawai J, Daub CO, Hume DA, Suzuki H, Orlando V, Carninci P, Hayashizaki Y, Mattick JS. 2009. Tiny RNAs associated with transcription start sites in animals. Nat Genet 41:572–578.

Taylor J. 2005. Clues to function in gene deserts. Trends Biotechnol 23:269–271.

Terai G, Yoshizawa A, Okida H, Asai K, Mituyama T. 2010. Discovery of short pseudogenes derived from messenger RNAs. Nucleic Acids Res 38:1163–1171.

Thean LF, Loi C, Ho KS, Koh PK, Eu KW, Cheah PY. 2010. Genome-wide scan identifies a copy number variable region at 3q26 that regulates *PPM1L* in *APC* mutation-negative familial colorectal cancer patients. Genes Chrom Cancer 49:99–106.

Theuns J, Del-Favero J, Dermaut B, van Duijn CM, Backhovens H, Van den Broeck MV, Serneels S, Corsmit E, Van Broeckhoven CV, Cruts M. 2000. Genetic variability in the regulatory region of presenilin 1 associated with risk for Alzheimer's disease and variable expression. Hum Mol Genet 9:325–331.

Thornburg BG, Gotea V, Makałowski W. 2006. Transposable elements as a significant source of transcription regulating signals. Gene 365:104–110.

Tress ML, Martelli PL, Frankish A, Reeves GA, Wesselink JJ, Yeats C, Olason PI, Albrecht M, Hegyi H, Giorgetti A, Raimondo D, Lagarde J, Laskowski RA, López G, Sadowski MI, Watson JD, Fariselli P, Rossi I, Nagy A, Kai W, Størling Z, Orsini M, Assenov Y, Blankenburg H, Huthmacher C, Ramírez F, Schlicker A, Denoeud F, Jones P, Kerrien S, Orchard S, Antonarakis SE, Reymond A, Birney E, Brunak S, Casadio R, Guigo R, Harrow J, Hermjakob H, Jones DT, Lengauer T, Orengo CA, Patthy L, Thornton JM, Tramontano A, Valencia A. 2007. The implications of alternative splicing in the ENCODE protein complement. Proc Natl Acad.Sci USA 104:5495–5500.

Tsai CJ, Sauna ZE, Kimchi-Sarfaty C, Ambudkar SV, Gottesman MM, Nussinov R. 2008. Synonymous mutations and ribosome stalling can lead to altered folding pathways and distinct minima. J Mol Biol 383:281–291.

Tu Z, Wang L, Xu M, Zhou X, Chen T, Sun F. 2006. Further understanding human disease genes by comparing with housekeeping genes and other genes. BMC Genomics 7:31.

Tucker T, Marra M, Friedman JM. 2009. Massively parallel sequencing: the next big thing in genetic medicine. Am J Hum Genet 85:142–154.

Turgeon B, Meloche S. 2009. Interpreting neonatal lethal phenotypes in mouse mutants: insights into gene function and human diseases. Physiol Rev 89:1–26.

Tuupanen S, Turunen M, Lehtonen R, Hallikas O, Vanharanta S, Kivioja T, Björklund M, Wei G, Yan J, Niittymäki I, Mecklin JP, Järvinen H, Ristimäki A, Di-Bernardo M, East P, Carvajal-Carmona L, Houlston RS, Tomlinson I, Palin K, Ukkonen E, Karhu A, Taipale J, Aaltonen LA. 2009. The common colorectal cancer predisposition SNP rs6983267 at chromosome 8q24 confers potential to enhanced Wnt signaling. Nat Genet 41:885–890.

van Bokhoven H, Rawson RB, Merkx GF, Cremers FP, Seabra MC. 1996. cDNA cloning and chromosomal localization of the genes encoding the alpha- and beta-subunits of human Rab geranylgeranyl transferase: the 3′ end of the alpha-subunit gene overlaps with the transglutaminase 1 gene promoter. Genomics 38:133–140.

Velagaleti GV, Bien-Willner GA, Northup JK, Lockhart LH, Hawkins JC, Jalal SM, Withers M, Lupski JR, Stankiewicz P. 2005. Position effects due to chromosome breakpoints that map approximately 900 Kb upstream and approximately 1.3 Mb downstream of *SOX9* in two patients with campomelic dysplasia. Am J Hum Genet 76:652–662.

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, Zhang Q, Kodira CD, Zheng XH, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang J, Gabor Miklos GL, Nelson C, Broder S, Clark AG, Nadeau J, McKusick VA, Zinder N, Levine AJ, Roberts RJ, Simon M, Slayman C, Hunkapiller M, Bolanos R, Delcher A, Dew I, Fasulo D, Flanigan M, Florea L, Halpern A, Hannenhalli S, Kravitz S, Levy S, Mobarry C, Reinert K, Remington K, Abu-Threideh J, Beasley E, Biddick K, Bonazzi V, Brandon R, Cargill M, Chandramouliswaran I, Charlab R, Chaturvedi K, Deng Z, Di Francesco V, Dunn P, Eilbeck K, Evangelista C, Gabrielian AE, Gan W, Ge W, Gong F, Gu Z, Guan P, Heiman TJ, Higgins ME, Ji RR, Ke Z, Ketchum KA, Lai Z, Lei Y, Li Z, Li J, Liang Y, Lin X, Lu F, Merkulov GV, Milshina N, Moore HM, Naik AK, Narayan VA, Neelam B, Nusskern D, Rusch DB, Salzberg S, Shao W, Shue B, Sun J, Wang Z, Wang A, Wang X, Wang J, Wei M, Wides R, Xiao C, Yan C, Yao A, Ye J, Zhan M, Zhang W, Zhang H, Zhao Q, Zheng L, Zhong F, Zhong W, Zhu S, Zhao S, Gilbert D, Baumhueter S, Spier G, Carter C, Cravchik A, Woodage T, Ali F, An H, Awe A, Baldwin D, Baden H, Barnstead M, Barrow I, Beeson K, Busam D, Carver A, Center A, Cheng ML, Curry L, Danaher S, Davenport L, Desilets R, Dietz S, Dodson K, Doup L, Ferriera S, Garg N, Gluecksmann A, Hart B, Haynes J, Haynes C, Heiner C, Hladun S, Hostin D, Houck J, Howland T, Ibegwam C, Johnson J, Kalush F, Kline L, Koduru S, Love A, Mann F, May D, McCawley S, McIntosh T, McMullen I, Moy M, Moy L, Murphy B, Nelson K, Pfannkoch C, Pratts E, Puri V, Qureshi H, Reardon M, Rodriguez R, Rogers YH, Romblad D, Ruhfel B, Scott R, Sitter C, Smallwood M, Stewart E, Strong R, Suh E, Thomas R, Tint NN, Tse S, Vech C, Wang G, Wetter J, Williams S, Williams M, Windsor S, Winn-Deen E, Wolfe K, Zaveri J, Zaveri K, Abril JF, Guigó R, Campbell MJ, Sjolander KV, Karlak B, Kejariwal A, Mi H, Lazareva B, Hatton T, Narechania A, Diemer K, Muruganujan A, Guo N, Sato S, Bafna V, Istrail S, Lippert R, Schwartz R, Walenz B, Yooseph S, Allen D, Basu A, Baxendale J, Blick L, Caminha M, Carnes-Stine J, Caulk P, Chiang YH, Coyne M, Dahlke C, Mays A, Dombroski M, Donnelly M, Ely D, Esparham S, Fosler C, Gire H, Glanowski S, Glasser K, Glodek A, Gorokhov M, Graham K, Gropman B, Harris M, Heil J, Henderson S, Hoover J, Jennings D, Jordan C, Jordan J, Kasha J, Kagan L, Kraft C, Levitsky A, Lewis M, Liu X, Lopez J, Ma D, Majoros W, McDaniel J, Murphy S, Newman M, Nguyen T, Nguyen N, Nodell M, Pan S, Peck J, Peterson M, Rowe W, Sanders R, Scott J, Simpson M, Smith T, Sprague A, Stockwell T, Turner R, Venter E, Wang M, Wen M, Wu D, Wu M, Xia A, Zandieh A, Zhu X. 2001. The sequence of the human genome. Science 291:1304–1351.

Venturin M, Moncini S, Villa V, Russo S, Bonati MT, Larizza L, Riva P. 2006. Mutations and novel polymorphisms in coding regions and UTRs of *CDK5R1* and *OMG* genes in patients with non-syndromic mental retardation. Neurogenetics 7:59–66.

VerMilyea MD, O'Neill LP, Turner BM. 2009. Transcription-independent heritability of induced histone modifications in the mouse preimplantation embryo. PLoS One 4:e6086.

Veyrieras JB, Kudaravalli S, Kim SY, Dermitzakis ET, Gilad Y, Stephens M, Pritchard JK. 2008. High-resolution mapping of expression-QTLs yields insight into human gene regulation. PLoS Genet 4:e1000214.

Viprakasit V, Kidd AM, Ayyub H, Horsley S, Hughes J, Higgs DR. 2003. *De novo* deletion within the telomeric region flanking the human α-globin locus as a cause of α-thalassaemia. Br J Haematol 120:867–875.

Visel A, Rubin EM, Pennacchio LA. 2009. Genomic views of distant-acting enhancers. Nature 461:199–205.

Vuoristo JT, Berrettini WH, Ala-Kokko L. 2001. *C18orf2*, a novel, highly conserved intronless gene within intron 5 of the *GNAL* gene on chromosome 18p11. Cytogenet Cell Genet 93:19–22.

Waalen J, Beutler E. 2009. Genetic screening for low-penetrance variants in protein-coding genes. Annu Rev Genomics Hum Genet 10:431–450.

Walsh T, McClellan JM, McCarthy SE, Addington AM, Pierce SB, Cooper GM, Nord AS, Kusenda M, Malhotra D, Bhandari A, Stray SM, Rippey CF, Roccanova P, Makarov V, Lakshmi B, Findling RL, Sikich L, Stromberg T, Merriman B, Gogtay N, Butler P, Eckstrand K, Noory L, Gochman P, Long R, Chen Z, Davis S, Baker C, Eichler EE, Meltzer PS, Nelson SF, Singleton AB, Lee MK, Rapoport JL, King MC, Sebat J. 2008. Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. Science 320:539–543.

Walters RG, Jacquemont S, Valsesia A, de Smith AJ, Martinet D, Andersson J, Falchi M, Chen F, Andrieux J, Lobbens S, Delobel B, Stutzmann F, El-Sayed Moustafa JS, Chèvre JC, Lecoeur C, Vatin V, Bouquillon S, Buxton JL, Boute O, Holder-Espinasse M, Cuisset JM, Lemaitre MP, Ambresin AE, Brioschi A, Gaillard M, Giusti V, Fellmann F, Ferrarini A, Hadjikhani N, Campion D, Guilmatre A, Goldenberg A, Calmels N, Mandel JL, Le Caignec C, David A, Isidor B, Cordier MP, Dupuis-Girod S, Labalme A, Sanlaville D, Béri-Dexheimer M, Jonveaux P, Leheup B, Ounap K, Bochukova EG, Henning E, Keogh J, Ellis RJ, Macdermot KD, van Haelst MM, Vincent-Delorme C, Plessis G, Touraine R, Philippe A, Malan V, Mathieu-Dramard M, Chiesa J, Blaumeiser B, Kooy RF, Caiazzo R, Pigeyre M, Balkau B, Sladek R, Bergmann S, Mooser V, Waterworth D, Reymond A, Vollenweider P, Waeber G, Kurg A, Palta P, Esko T, Metspalu A, Nelis M, Elliott P, Hartikainen AL, McCarthy MI, Peltonen L, Carlsson L, Jacobson P, Sjöström L, Huang N, Hurles ME, O'Rahilly S, Farooqi IS, Männik K, Jarvelin MR, Pattou F, Meyre D, Walley AJ, Coin LJ, Blakemore AI, Froguel P, Beckmann JS. 2010. A new highly penetrant form of obesity due to deletions on chromosome 16p11.2. Nature 463:671–675.

Wang GS, Cooper TA. 2007. Splicing in disease: disruption of the splicing code and the decoding machinery. Nat Rev Genet 8:749–761.

Wang J, Wang W, Li R, Li Y, Tian G, Goodman L, Fan W, Zhang J, Li J, Zhang J, Guo Y, Feng B, Li H, Lu Y, Fang X, Liang H, Du Z, Li D, Zhao Y, Hu Y, Yang Z, Zheng H, Hellmann I, Inouye M, Pool J, Yi X, Zhao J, Duan J, Zhou Y, Qin J, Ma L, Li G, Yang Z, Zhang G, Yang B, Yu C, Liang F, Li W, Li S, Li D, Ni P, Ruan J, Li Q, Zhu H, Liu D, Lu Z, Li N, Guo G, Zhang J, Ye J, Fang L, Hao Q, Chen Q, Liang Y, Su Y, San A, Ping C, Yang S, Chen F, Li L, Zhou K, Zheng H, Ren Y, Yang L, Gao Y, Yang G, Li Z, Feng X, Kristiansen K, Wong GK, Nielsen R, Durbin R, Bolund L, Zhang X, Li S, Yang H, Wang J. 2008a. The diploid genome sequence of an Asian individual. Nature 456:60–65.

Wang X, Wang K, Radovich M, Wang Y, Wang G, Feng W, Sanford JR, Liu Y. 2009a. Genome-wide prediction of *cis*-acting RNA elements regulating tissue-specific pre-mRNA alternative splicing. BMC Genomics 10(Suppl 1):S4.

Wang Z, Gerstein M, Snyder M. 2009b. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 10:57–63.

Wang Z, Zang C, Rosenfeld JA, Schones DE, Barski A, Cuddapah S, Cui K, Roh TY, Peng W, Zhang MQ, Zhao K. 2008b. Combinatorial patterns of histone acetylations and methylations in the human genome. Nat Genet 40:897–903.

Wang ZQ, Tian SH, Shi YZ, Zhou PT, Wang ZY, Shu RZ, Hu L, Kong X. 2007. A single C to T transition in intron 5 of *LMBR1* gene is associated with triphalangeal thumb-polysyndactyly syndrome in a Chinese family. Biochem Biophys Res Commun 355:312–317.

Washietl S, Hofacker IL, Lukasser M, Hüttenhofer A, Stadler PF. 2005. Mapping of conserved RNA secondary structures predicts thousands of functional noncoding RNAs in the human genome. Nat Biotechnol 23:1383–1390.

Wen Y, Liu Y, Xu Y, Zhao Y, Hua R, Wang K, Sun M, Li Y, Yang S, Zhang XJ, Kruse R, Cichon S, Betz RC, Nöthen MM, van Steensel MA, van Geel M, Steijlen PM, Hohl D, Huber M, Dunnill GS, Kennedy C, Messenger A, Munro CS, Terrinoni A, Hovnanian A, Bodemer C, de Prost Y, Paller AS, Irvine AD, Sinclair R, Green J, Shang D, Liu Q, Luo Y, Jiang L, Chen HD, Lo WH, McLean WH, He CD, Zhang X. 2009. Loss-of-function mutations of an inhibitory upstream ORF in the human hairless transcript cause Marie Unna hereditary hypotrichosis. Nat Genet 41:228–233.

Werner A, Carlile M, Swan D. 2009. What do natural antisense transcripts regulate? RNA Biol 6:43–48.

Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, McGuire A, He W, Chen YJ, Makhijani V, Roth GT, Gomes X, Tartaro K, Niazi F, Turcotte CL, Irzyk GP, Lupski JR, Chinault C, Song XZ, Liu Y, Yuan Y, Nazareth L, Qin X, Muzny DM, Margulies M, Weinstock GM, Gibbs RA, Rothberg JM. 2008. The complete genome of an individual by massively parallel DNA sequencing. Nature 452:872–876.

Wieczorek D, Pawlik B, Li Y, Akarsu NA, Caliebe A, May KJ, Schweiger B, Vargas FR, Balci S, Gillessen-Kaesbach G, Wollnik B. 2010. A specific mutation in the distant sonic hedgehog (*SHH*) cis-regulator (ZRS) causes Werner mesomelic syndrome (WMS) while complete ZRS duplications underlie Haas type polysyndactyly and preaxial polydactyly (PPD) with or without triphalangeal thumb. Hum Mutat 31:81–89.

Wilch E, Azaiez H, Fisher RA, Elfenbein J, Murgia A, Birkenhäger R, Bolz H, da Silva-Costa SM, Del Castillo I, Haaf T, Hoefsloot L, Kremer H, Kubisch C, Le Marechal C, Pandya A, Sartorato EL, Schneider E, Van Camp G, Wuyts W, Smith RJ, Friderici KH. 2010. A novel DFNB1 deletion allele supports the existence of a distant *cis*-regulatory region that controls *GJB2* and *GJB6* expression. Clin Genet (in press).

Wilkins JM, Southam L, Mustafa Z, Chapman K, Loughlin J. 2009. Association of a functional microsatellite within intron 1 of the *BMP5* gene with susceptibility to osteoarthritis. BMC Med Genet 10:141.

Wong E, Wei CL. 2009. ChIP'ing the mammalian genome: technical advances and insights into functional elements. Genome Med 1:89.

Wright JB, Brown SJ, Cole MD. 2010. Upregulation of c-*MYC* in cis through a large chromatin loop linked to a cancer risk-associated single-nucleotide polymorphism in colorectal cancer cells. Mol Cell Biol 30:1411–1420.

Wu M, Jolicoeur N, Li Z, Zhang L, Fortin Y, L'Abbe D, Yu Z, Shen SH. 2008. Genetic variations of microRNAs in human cancer and their effects on the expression of miRNAs. Carcinogenesis 29:1710–1716.

Xing J, Zhang Y, Han K, Salem AH, Sen SK, Huff CD, Zhou Q, Kirkness EF, Levy S, Batzer MA, Jorde LB. 2009. Mobile elements create structural variation: analysis of a complete human genome. Genome Res 19:1516–1526.

Yagi M, Takeshima Y, Wada H, Nakamura H, Matsuo M. 2003. Two alternative exons can result from activation of the cryptic splice acceptor site deep within intron 2 of the dystrophin gene in a patient with as yet asymptomatic dystrophinopathy. Hum Genet 112:164–170.

Yamaguchi-Kabata Y, Shimada MK, Hayakawa Y, Minoshima S, Chakraborty R, Gojobori T, Imanishi T. 2008. Distribution and effects of nonsense polymorphisms in human genes. PLoS One 3:e3393.

Yang MQ, Elnitski LL. 2008. Diversity of core promoter elements comprising human bidirectional promoters. BMC Genomics 9(Suppl 2):S3.

Yang R, Frank B, Hemminki K, Bartram CR, Wappenschmidt B, Sutter C, Kiechle M, Bugert P, Schmutzler RK, Arnold N, Weber BH, Niederacher D, Meindl A, Burwinkel B. 2008a. SNPs in ultraconserved elements and familial breast cancer risk. Carcinogenesis 29:351–355.

Yang TL, Chen XD, Guo Y, Lei SF, Wang JT, Zhou Q, Pan F, Chen Y, Zhang ZX, Dong SS, Xu XH, Yan H, Liu X, Qiu C, Zhu XZ, Chen T, Li M, Zhang H, Zhang L, Drees BM, Hamilton JJ, Papasian CJ, Recker RR, Song XP, Cheng J, Deng HW. 2008b. Genome-wide copy-number-variation study identified a susceptibility gene, *UGT2B17*, for osteoporosis. Am J Hum Genet 83:663–674.

Yngvadottir B, MacArthur DG, Jin H, Tyler-Smith C. 2009a. The promise and reality of personal genomics. Genome Biol 10:237.

Yngvadottir B, Xue Y, Searle S, Hunt S, Delgado M, Morrison J, Whittaker P, Deloukas P, Tyler-Smith C. 2009b. A genome-wide survey of the prevalence and evolutionary forces acting on human nonsense SNPs. Am J Hum Genet 84:224–234.

Yu P, Ma D, Xu M. 2005. Nested genes in the human genome. Genomics 86:414–422.

Zhang C. 2008. MicroRNomics: a newly emerging approach for disease biology. Physiol. Genomics 33:139–147.

Zhang F, Carvalho CM, Lupski JR. 2009. Complex human chromosomal and genomic rearrangements. Trends Genet 25:298–307.

Zhang ZD, Frankish A, Hunt T, Harrow J, Gerstein M. 2010. Identification and analysis of unitary pseudogenes: historic and contemporary gene losses in humans and other primates. Genome Biol 11:R26.

Zhang ZD, Paccanaro A, Fu Y, Weissman S, Weng Z, Chang J, Snyder M, Gerstein MB. 2007. Statistical analysis of the genomic distribution and correlation of regulatory elements in the ENCODE regions. Genome Res 17:787–797.

Zheng D, Frankish A, Baertsch R, Kapranov P, Reymond A, Choo SW, Lu Y, Denoeud F, Antonarakis SE, Snyder M, Ruan Y, Wei CL, Gingeras TR, Guigó R, Harrow J, Gerstein MB. 2007. Pseudogenes in the ENCODE regions: consensus annotation, analysis of transcription, and evolution. Genome Res 17: 839–851.

Zheng D, Gerstein MB. 2007. The ambiguous boundary between genes and pseudogenes: the dead rise up, or do they? Trends Genet 23: 219–224.

Zhu QS, Xing W, Qian B, von Dippe P, Shneider BL, Fox VL, Levy D. 2003. Inhibition of human m-epoxide hydrolase gene expression in a case of hypercholanemia. Biochim Biophys Acta 1638:208–216.