



HAL
open science

Genes, mutations and human inherited disease at the dawn of the age of personalized genomics

David N. Cooper, Jian-Min Chen, Edward V. Ball, Katy Howells, Matthew Mort, Andrew D Phillips, Nadia Chuzhanova, Michael Krawczak, Hildegard Kehrer-Sawatzki, Peter Stentson

► To cite this version:

David N. Cooper, Jian-Min Chen, Edward V. Ball, Katy Howells, Matthew Mort, et al.. Genes, mutations and human inherited disease at the dawn of the age of personalized genomics. *Human Mutation*, 2010, 31 (6), pp.631. 10.1002/humu.21260 . hal-00552385

HAL Id: hal-00552385

<https://hal.science/hal-00552385>

Submitted on 6 Jan 2011

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Genes, mutations and human inherited disease at the dawn of the age of personalized genomics



Journal:	<i>Human Mutation</i>
Manuscript ID:	humu-2010-0027.R1
Wiley - Manuscript type:	Review
Date Submitted by the Author:	16-Mar-2010
Complete List of Authors:	Cooper, David; Cardiff University, Institute of Medical Genetics, College of Medicine Chen, Jian-Min; INSERM U613, Etablissement Français du Sang ?Bretagne Ball, Edward; Cardiff University, Institute of Medical Genetics, College of Medicine Howells, Katy; Cardiff University, Institute of Medical Genetics, College of Medicine Mort, Matthew; School of Medicine, Cardiff University, Institute of Medical Genetics Phillips, Andrew; Cardiff University, Institute of Medical Genetics, College of Medicine Chuzhanova, Nadia; University of Central Lancashire, School of Computing, Engineering and Physical Sciences Krawczak, Michael; University of Kiel, Institut für Medizinische Informatik und Statistik Kehrer-Sawatzki, Hildegard; University of Ulm Stentson, Peter; Cardiff University, Institute of Medical Genetics, College of Medicine
Key Words:	Human Gene Mutation Database, HGMD, inherited mutations, gene number, gene definition, gene essentiality, non-coding regions, functionome, mutation detection, human genome



Genes, mutations and human inherited disease at the dawn of the age of
personalized genomics,

DAVID N. COOPER^{1,6}, JIAN-MIN CHEN², EDWARD V. BALL¹, KATY HOWELLS¹,
 MATTHEW MORT¹, ANDREW D. PHILLIPS¹, NADIA CHUZHANOVA³, MICHAEL
KRAWCZAK⁴, HILDEGARD KEHRER-SAWATZKI⁵ AND PETER D. STENSON¹

¹Institute of Medical Genetics, School of Medicine, Cardiff University, Heath Park, Cardiff CF14 4XN,
 UK.

²Institut National de la Santé et de la Recherche Médicale (INSERM), U613 and Etablissement
Français du Sang (EFS) – Bretagne, Brest, France.

³School of Computing, Engineering and Physical Sciences, University of Central Lancashire, Preston PR1
 2HE, UK

⁴Institut für Medizinische Informatik und Statistik, Christian-Albrechts-Universität, Brunswiker Straße
 10, 24105 Kiel, Germany.

⁵Institut für Humangenetik, Universität Ulm, Albert-Einstein-Allee 11, 89081 Ulm, Germany.

⁶To whom all correspondence should be addressed. Tel: +44 2920 744062, Fax: +44 2920 746551

Email: CooperDN@cardiff.ac.uk

Formatted: Left: 72 pt, Right: 72 pt, Width: 612 pt, Height: 792 pt, Header distance from edge: 35.3 pt, Footer distance from edge: 35.3 pt

Deleted: 100K

Formatted: Right: 0 pt, Line spacing: Double

Deleted: ¶ mutome era¶

Formatted: Line spacing: Double

Deleted: ,

Formatted: Right: 0 pt, Line spacing: Double

1
2
3
4
5
6
7
8 **ABSTRACT:** The number of reported germline mutations in human nuclear genes, either
9 underlying or associated with inherited disease, has now exceeded 100,000 in more than 3,700
10 different genes. The availability of these data has both revolutionized the study of the morbid
11 anatomy of the human genome and facilitated ‘personalized genomics’. With ~300 new
12 ‘inherited disease genes’ (and ~10,000 new mutations) being identified annually, it is pertinent to
13 ask how many ‘inherited disease genes’ there are in the human genome, how many mutations
14 reside within them and where such lesions are likely to be located? To address these questions, it
15 is necessary not only to reconsider how we define human genes but also to explore notions of
16 gene ‘essentiality’ and ‘dispensability’. Answers to these questions are now emerging from
17 recent novel insights into genome structure and function and through complete genome sequence
18 information derived from multiple individual human genomes. However, a change in focus
19 towards screening functional genomic elements as opposed to genes *sensu stricto* will be
20 required if we are to capitalize fully on recent technical and conceptual advances and identify
21 new types of disease-associated mutation within non-coding regions remote from the genes
22 whose function they disrupt.
23
24
25
26
27
28
29
30
31
32
33
34
35
36

Deleted: is set to

Deleted: 6

Deleted: there

Deleted: are

Deleted: '

Deleted: '

37
38
39 **KEY WORDS:** Human Gene Mutation Database; HGMD; inherited mutations; human genome;
40 gene number; gene definition; disease genes; gene essentiality; non-coding regions; functionome;
41 mutation detection
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

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).

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 recognised every year. It is therefore pertinent to pose the double question: how many inherited disease genes are there in the human

Formatted: Line spacing: Double

Formatted: Right: 0 pt, Line spacing: Double

Deleted: set to

Deleted: 6

Deleted: -

1
2 genome and how many mutations are likely to be found within them? A first bold estimate of the
3
4 “number of mutations causing inherited disease” (20 million mutations apportioned between
5
6 20,000 different human genes) has recently been put forward [Cotton, 2009] but these numbers
7
8 appear to constitute only rough estimates that have not been justified in any formal way.

Deleted: are not well

9
10 In principle, the number of human ‘disease genes’ may well be estimable, albeit
11
12 approximately. However, whilst the number of different mutations that could *potentially* cause
13
14 human inherited disease is clearly almost limitless (if, for example, one were to include all
15
16 possible frameshift micro-deletions and micro-insertions), the number of mutations *actually* in
17
18 existence and available to be identified and characterized is a complex function of the mutability
19
20 of each inherited disease gene, the prevalence and ease of ascertainment of the consequent
21
22 clinical phenotype(s), the demographic history of the human population, as well as the technical
23
24 means at our disposal to locate and identify the pathological mutation(s) in any one individual.

25
26 Reich and Lander [2001] concluded that, with a ‘typical’ (pathological) gene mutation rate of
27
28 3.2×10^{-6} per generation, the average number of mutations underlying a rare inherited disease
29
30 would equal 77,000 at mutation-drift equilibrium. These authors also opined that the kinetics of
31
32 the mutation process are such that, for diseases characterized by an overall population frequency
33
34 of pathological mutations <1%, this equilibrium is likely to have been reached in the extant
35
36 human population. Based upon these considerations, the number of different mutations *actually*
37
38 underlying inherited human disease is likely to be one to two orders of magnitude higher than
39
40 that suggested by Cotton [2009], potentially totalling between 600 million and 2.4 billion
41
42 (average: 1.2 billion) depending upon the number of genes (estimated to lie somewhere between
43
44 7,750 and 30,770, with an average of 15,300; see below) adjudged to qualify as ‘inherited
45
46 disease genes’. However, most of these mutations will be extremely rare. Indeed, it can be
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2 calculated from the approximate distribution function of allele numbers at mutation-drift
3
4 equilibrium [Gale, 1990] that, given an overall population frequency of pathological mutations of
5
6 1% in a given gene, fewer than four mutations will have a relative frequency $>5 \times 10^{-4}$ in the pool
7
8 of pathological mutations of that gene. Thus, in terms of those inherited disease mutations that
9
10 are in practice actually detectable, the above figures are likely to represent gross overestimates,
11
12 and the number of mutations detected in a given gene will depend mostly upon the number of
13
14 patients studied rather than on the diversity of the underlying mutational spectrum of that gene.
15

16 | In attempting to collate all inherited human pathological gene mutations as they emerge in the
17 | literature [Stenson et al., 2009], HGMD has to some extent embarked on an open ended project
18 | whose eventual scale and scope was quite impossible to assess from the outset. Daunting as this
19 | prospect is, it is nevertheless appropriate at this juncture to take stock and try to assess where we
20 | are in terms of the indubitably massive task of identifying, annotating and cataloguing the human
21 | germline mutational spectrum ('mutome'). We shall argue that, although the question of the
22 | 'number of mutations causing inherited disease' may well be akin to asking "How long is a piece
23 | of string?", there are several related questions which appear to be worthwhile addressing on
24 | account of their practical importance: *How many genes are there in the human genome? How*
25 | *many of these are inherited disease genes (i.e. genes harbouring mutations that are capable of*
26 | *causing inherited disease)? What proportion of the universe of possible mutations within these*
27 | *inherited disease genes is likely to be of pathological significance? Where, in genomic terms, are*
28 | *these mutations likely to be found? How many deleterious mutations are there on average per*
29 | *individual?* The answers to these questions should shed some light on the likely size of the task
30 | facing us as we attempt to document the spectrum of mutations causing (or associated with)
31 | human inherited disease.
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

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 [Venter et al., 2001; International Human Genome Sequencing Consortium, 2004]. 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., 2007]. 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.

1
2 To appreciate why definition is an issue here, one need only be aware of the many exceptions
3 to genes being contiguous (as well as functionally and spatially distinct) entities, as classically
4 envisaged. Thus, some genes are known to occur within the introns of other genes [Herzog et al.,
5
6 1996; Vuoristo et al., 2001]. Some genes can overlap with each other either on the same or on
7
8 different DNA strands [Denoeud et al., 2007] resulting in the sharing of some of their coding
9
10 and/or regulatory elements [van Bokhoven et al., 1996; Yang and Elnitski, 2008]. In addition, the
11
12 vast majority of human genes are now known to undergo alternative splicing [Pan et al., 2008],
13
14 leading in some cases to quite different proteins being encoded by the same gene. For example,
15
16 the human *CDKN2A* gene (MIM# 600160) encodes an alternatively spliced variant (p14^{ARF}) that,
17
18 through the inclusion of an alternative first exon, acquires an altered reading frame so as to
19
20 specify a protein product that is structurally unrelated to the other p16 isoforms encoded by this
21
22 gene.
23
24

Deleted: Other

Deleted: T

Deleted: protein

25
26 Bicistronic genes [e.g. *MOCSI*; MIM# 6003707; Gross-Hardt and Reiss, 2002] are also
27
28 atypical, with transcription initiating from one gene and continuing in *cis* through a neighbouring
29
30 downstream gene to yield a precursor protein that is then cleaved to generate different proteins.
31
32 Such 'transcription-mediated gene fusion' may well not be an infrequent occurrence in the
33
34 human genome [Akiva et al., 2006; Parra et al., 2006]. Moreover, there is now persuasive
35
36 evidence for the occurrence of *trans*-splicing in human cells, involving the cleavage and joining
37
38 of entirely separate RNA transcripts [Gingeras et al., 2009].
39

Deleted: In addition

40 Many protein-coding genes have been found to possess alternative transcriptional initiation
41
42 sites, some of which may be quite remote from the gene itself, in some instances even residing
43
44 within the bounds of another gene [Carninci et al., 2006; Denoeud et al., 2007]. Other genes
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2 exhibit differential polyadenylation site usage leading to length heterogeneity of the 3'
3 untranslated region [Kwan et al., 2008].

4
5
6 Should distant *cis*-acting regulatory sequences be included within the boundaries of the gene
7 they serve to regulate? If so, then it would make the concept of the gene that much more flexible.
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

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 come up with a general definition of the gene, either structurally or functionally, that will withstand close scrutiny in the context of many thousands of different 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.

Deleted: define

Transcripts of unknown function and unannotated transcripts

The *ENCyclopedia Of DNA Elements* (ENCODE) project, designed to analyse 30 megabases (Mb) of DNA from 44 genomic regions (thereby sampling 1% of the genome) in order to characterize the functional elements present, has identified complex patterns of regulation and 'pervasive transcription' of the human genome [ENCODE Project Consortium, 2007]. Whilst >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

Deleted: L

[Gingeras, 2007]. Such transcripts have been collectively classified as *transcripts of unknown function* (TUFs) and are thought to include (i) antisense transcripts of protein-coding genes, (ii) isoforms of protein-coding genes, and (iii) transcripts that either overlap introns of annotated gene transcripts (on the same strand) or which are derived entirely from intergenic regions.

~~Whilst both the complexity and abundance of TUFs are quite remarkable, it should be realised that there is often no firm evidence for these transcripts being of functional significance.~~ Indeed,

unannotated non-polyadenylated 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

[Kapranov et al., 2002; 2007a; Gingeras, 2007]. Although the functional significance of 'pervasive transcription' remains unclear, it is much more extensive than had previously been realised [Dinger et al., 2009].

In both human and mouse, up to 70% of genomic loci exhibit evidence of transcription from the antisense strand as well as the sense strand [Katayama et al., 2005; Werner et al., 2009; Grinchuk et al., 2010]. 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 [He et al., 2008; Faghihi and Wahlestedt, 2009]. 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

Deleted: B

Deleted: that

Deleted: are

Deleted: clearly

Deleted: (cisNATs)

1
2 noise' [Ponjavic et al., 2007; Louro et al., 2009], at least a portion of it is likely to be associated
3
4 with functional non-protein-coding RNA genes, many of which are located in regions previously
5
6 regarded as intergenic and/or non-coding [ENCODE Project Consortium, 2007]. Non-coding
7
8 RNA genes are as widespread as they are diverse [Borel et al., 2008], are transcribed from both
9
10 strands of the genome, and may well exceed protein-coding genes in terms of their number
11
12 [[Washietl et al., 2005](#); Fejes-Toth et al., 2009]. Non-protein-coding RNAs of known function
13
14 include structural RNAs such as transfer RNAs, ribosomal RNAs and small nuclear RNAs, but
15
16 also putative regulatory RNAs [microRNAs, small interfering RNAs (siRNAs), Piwi-interacting
17
18 RNAs, transcription initiation RNAs (tiRNAs), transcription start site-associated RNAs (TSSa-
19
20 RNAs), promoter upstream transcripts (PROMPTs), promoter-associated sRNAs (PASRs and
21
22 PALRs), and longer non-coding RNAs such as *XIST*] which are involved in the sequence-
23
24 specific transcriptional and post-transcriptional modulation of gene expression [Kawaji and
25
26 Hayashizaki, 2008; Seila et al., 2008; Collins and Penny, 2009; Mercer et al., 2009; Taft et al.,
27
28 2009; Mattick, 2009]. Thus, more than 700 microRNA genes have already been identified in the
29
30 human genome with many more probably awaiting discovery (miRBase;
31
32 http://www.mirbase.org/cgi-bin/mirna_summary.pl?org=hsa). In total, at least 1,500 non-protein-
33
34 coding RNA genes have already been annotated in the human genome reference sequence with
35
36 up to 5,000 more predicted by homology-based methods [Griffiths-Jones, 2007; see Ensembl,
37
38 database version 56.37a]. Indeed, large intergenic non-coding RNAs (lincRNAs) have recently
39
40 been found to represent a novel category of evolutionarily conserved RNAs, with a diverse array
41
42 of functions ranging from embryonic stem cell pluripotency to cellular proliferation [Guttman et
43
44 al., 2009; Khalil et al., 2009]; lincRNAs appear to number at least 3,000 in the human genome.
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Pseudogenes

Whether processed or non-processed (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; Zheng et al., 2007; Sakai et al., 2007]. It should however be appreciated that while 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 \(<300 bp\) 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 non-functional alleles segregating in the extant human population \(Zhang et al., 2010\).](#)

Deleted: them

Deleted: by

Deleted: they

Deleted: both

Deleted: and

Deleted:

With the realization that pseudogene-derived RNA transcripts may harbour functional elements [Zheng et al., 2007; Khachane and Harrison, 2009], 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 non-coding RNAs. It is at present unclear what proportion of pseudogenes identified to date have either retained or acquired a function via their non-coding RNAs.

Transposable elements

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

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 inter-individual 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 non-coding 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., 2007]. Many more are likely to have functional significance as suggested by their evolutionary conservation [Nishihara et al., 2006; Lowe et al., 2007].

Deleted: to the human genome

Regulatory non-coding sequences

Extensive evolutionary conservation of non-coding DNA sequence is very evident in the human genome, since 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 non-coding sequences [Ponting and Lunter, 2006; Drake et al., 2006; Asthana et al., 2007; Parker et al., 2009] have suggested that 5-20% of the genome may be of functional importance rather than just the ~2% associated with the protein-coding portion [Pheasant and Mattick, 2007; Eory et al., 2010]. Some non-coding regions contain 'ultraconserved elements' [Katzman et al., 2007] which appear to have been subject to similarly strong selection as protein-coding regions. [Some non-coding 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 however be appreciated that the absence of evolutionary conservation does not necessarily denote lack of function. Indeed,

1
2 human-specific functional elements have been shown to be present within rapidly evolving non-
3 coding sequences [Prabhakar et al., 2006; Bird et al., 2007].
4
5

6
7
8 *Towards a new definition of the gene*
9

10 It is clear from the above that precisely what constitutes a gene has become somewhat
11 contentious. The quite unanticipated scale of the extent of transcription in the genome, coupled
12 with the widespread occurrence of overlapping genes and shared functional elements, hampers
13 attempts to demarcate precisely and unambiguously where one gene ends and another one
14 begins. As a consequence, the notion of the gene has become quite diffuse [Gerstein et al., 2007;
15 Gingeras, 2007]. Indeed, as Kapranov et al. [2005] opined, “it is not unusual that a single base-
16 pair can be part of an intricate network of multiple isoforms of overlapping sense and antisense
17 transcripts, the majority of which are unannotated”. Gene regulatory elements which are often
18 quite distant from the genes they regulate [Kleinjan and Lettice, 2008], the existence of *trans*- as
19 well as *cis*-regulatory elements [Morley et al., 2004], the formation of non-colinear transcripts
20 through *trans*-splicing [Gingeras, 2009], taken together with the abundance of non-coding RNA
21 genes [Zhang, 2008] and evolutionarily conserved non-coding regions [Ponting and Lunter,
22 2006; Drake et al., 2006] have combined to challenge the classical notion of the gene.
23
24
25
26
27
28
29
30
31
32
33
34
35

36 On the basis of the findings of the ENCODE project, Gerstein et al. [2007] proposed an
37 updated definition of the gene as “a union of genomic sequences encoding a coherent set of
38 potentially overlapping functional products”. An alternative less heterodox definition of the gene
39 as “a discrete genomic region whose transcription is regulated by one or more promoters and
40 distal regulatory elements and which contains the information for the synthesis of functional
41 proteins or non-coding RNAs, related by the sharing of a portion of genetic information at the
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2 level of the ultimate products (proteins or RNAs)” has been proposed by Pesole [2009].
3
4 Irrespective of its precise definition, it is clear that the concept of the gene is inadequate to the
5
6 task of building a lexicon of those functional genomic sequences that could harbour mutations
7
8 causing human inherited disease. It is *indeed* likely that, in the context of mutation detection, we
9
10 shall eventually have to consider the universe of functional genetic elements in the human
11
12 genome (the ‘functionome’; see Fig. 1) as our hunting ground rather than *simply* genes *per se*.
13
14
15

16 **How many inherited disease genes are there in the human genome?**

17
18 *The concept of gene essentiality lies at the heart of any discussion of human disease genes*
19

20 The question of how many inherited disease genes there are in the human genome should
21
22 essentially be couched in terms of the proportion of human genes whose mutation would come to
23
24 clinical attention in a non-negligible proportion of cases by conferring a discernible clinical
25
26 phenotype upon the individual concerned. As López-Bigas et al. [2006] have expressed it, “a
27
28 gene is involved in a hereditary disease when its sequence has been subjected to a mutation that
29
30 impairs its function or expression strongly enough to produce a certain pathological phenotype
31
32 that is classified as a disease”. However, although necessarily deleterious, such a mutation must
33
34 not be lethal to the individual at an early stage of development since this would militate against
35
36 its detection. Hence, disease genes are not, and cannot be, synonymous with ‘essential genes’.
37
38 Indeed, they exhibit very different properties [López-Bigas et al., 2006; Goh et al., 2007;
39
40 Feldman et al., 2008]. The above notwithstanding, disease genes appear to be distinguishable
41
42 from ‘non-disease genes’ (in reality, the latter can only be defined as genes that are not yet
43
44 known to cause inherited disease) in terms of a range of features including gene structure, gene
45
46 expression, physicochemical properties, protein structure and evolutionary conservation
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2 [Jimenez-Sanchez et al., 2001; Smith and Eyre-Walker, 2003; López-Bigas and Ouzounis, 2004;
3
4 Huang et al., 2004; Tu et al., 2006; Subramanian and Kumar, 2006; Ng and Henikoff, 2006;
5
6 Aerts et al., 2006; Lage et al., 2008; Domazet-Lošo and Tautz, 2008; Cai et al., 2009]. In this
7
8 context, it should be appreciated that many disease genes will not have been identified as such
9
10 simply because the underlying mutations have not yet appeared in the homozygous/compound
11
12 heterozygous/hemizygous state required to manifest a clinical phenotype [Furney et al., 2006;
13
14 Osada et al., 2009].

15
16 | Although ~15% of mouse gene knockouts are developmentally lethal [Turgeon and Meloche,
17
18 2009; Mouse Genome Informatics Resource (<http://www.informatics.jax.org/>)], any definition of
19
20 gene essentiality based exclusively on developmental lethality would be unnecessarily
21
22 restrictive. Disease genes should therefore be understood in terms of a spectrum of gene
23
24 'essentiality' that stretches from the truly essential genes on the one hand to almost dispensable
25
26 genes on the other. Although essential genes have been quite reasonably defined as those genes
27
28 which are “absolutely required for survival, or [which] strongly contribute to fitness and robust
29
30 competitive growth” [Park et al., 2008], it should be appreciated that definitions of gene
31
32 essentiality can differ quite widely between studies [Gerdes et al., 2006]. Using 2,789 disease
33
34 genes from the HGMD gene set, Park et al. [2008] explored the likelihood of a gene being linked
35
36 to human inherited disease in relation to its level of essentiality in mouse (4,004 genes) as
37
38 adjudged by the results of gene disruption and knock-out experiments. Twice as many genes
39
40 with abnormal effects when disrupted in mouse (1,311/3,635; 36%) were found to have a human
41
42 disease gene homologue than genes which displayed no overt phenotypic abnormality when
43
44 disrupted (63/369; 17%). Somewhat surprisingly, when the genes with abnormal effects in
45
46 mouse were sub-divided into genes with lethal effects and non-lethal effects, the frequencies of
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2 disease gene homologues among them were comparable [38% (728/1,904) and 34% (583/1,731)
3
4 respectively]. However, when Park et al. [2008] further sub-divided the genes with lethal effects
5
6 in mouse, they found human disease gene homologues to be 1.4 times more frequent among
7
8 genes categorized as being 'post-natal lethal' in the mouse than among 'embryonic lethal' genes.
9

10 Thus, almost counter-intuitively at first glance, the more essential murine genes (which are
11 embryonic lethal in mouse) appear to be less likely to be disease genes in human. This finding
12 Deleted: the
13
14 confirms the above mentioned dictum that disease genes are not, and cannot be, synonymous
15
16 with 'essential genes'. Interestingly, Park et al. [2008] also observed that the type of disease
17
18 mutation in the human homologue varies depending upon the essentiality of the mouse gene
19
20 involved, with nonsense mutations, splicing mutations, micro-insertions/micro-deletions and
21
22 gross insertions/deletions being disproportionately associated with the mouse genes displaying
23
24 abnormal effects when disrupted. We may also speculate that whilst a mild mutation in an
25
26 'essential' gene may be sufficient to cause disease, a much more severe mutation may be
27
28 necessary in a 'dispensable' gene. Clearly, concepts of gene essentiality and dispensability are
29
30 likely to be context-dependent.
31

32 Although ~91% of the murine genes employed in the study described above were deemed to
33
34 belong to the 'essential' category (i.e. the group of genes that display abnormal phenotypic
35
36 effects when mutated), we should be wary of making direct inferences in the human context.
37
38 This is not only because those mouse genes with a known mutational phenotype comprise fewer
39
40 than 20% of the total number of genes in this organism, but also because it may be somewhat
41
42 hazardous to extrapolate to the human genome where both gene function [Liao and Zhang, 2008]
43
44 and copy number [Cutler and Kassner, 2008] may differ quite markedly from mouse.
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2 Inspection of the entry record history of HGMD [Stenson et al., 2009] reveals a constant
3 increase in the rate at which newly reported disease genes have been entered into HGMD every
4 year, with 293 recorded for 2009 [Fig. 2]. Since this increase has to cease at some point in time,
5 simply because the number of human genes is finite, we ventured to fit the various 4- or 5-
6 parameter saturation models provided by SigmaPlot v.8.02 (SPSS Inc., 2002) to the annual
7 cumulative gene number in HGMD since 1978. The results of these admittedly highly
8 speculative projections (which nevertheless yielded an $R^2 > 0.9999$ for all models) point to a total
9 number of inherited disease genes of between 7,750 (5-parameter Weibull model) and 30,750 (5-
10 parameter Chapman model). The remaining four models (sigmoid, logistic, Gompertz and Hill)
11 yielded estimates in a very narrow range of between 11,850 and 15,100 inherited disease genes,
12 and the average taken over all six models equalled 15,300, i.e. 46% of the 33,000 genes currently
13 estimated to be present in the human genome (HuRef NCBI build 37.1) .
14
15
16
17
18
19
20
21
22
23
24
25
26

Deleted: 1

Deleted: 1

27
28 *Concepts of human gene essentiality and dispensability are necessarily bound up with gene copy*
29 *number*
30

31
32 The loss of a particular gene/allele is not invariably associated with a readily discernible
33 clinical phenotype (Waaen and Beutler, 2009). This assertion is supported by the identification
34 of more than 1,000 putative nonsense SNPs (i.e. nonsense mutations that have attained
35 polymorphic frequencies) in human populations [Ng et al., 2008; Yngvadottir et al., 2009].
36 About half of these nonsense SNPs have been validated by dbSNP
37 [<http://www.ncbi.nlm.nih.gov/projects/SNP>], a process that involves the exclusion of mutations
38 in pseudogenes and of artefacts caused by sequencing errors. *Bona fide* nonsense SNPs are
39 expected either to lead to the synthesis of a truncated protein product or alternatively to the
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2 greatly reduced synthesis of the truncated protein product [if the mRNA bearing them is subject
3 to nonsense-mediated mRNA decay (NMD)]. Based upon the relative locations of the nonsense
4 SNPs and the exon-intron structures of the affected genes, Yamaguchi-Kabata et al. [2008]
5
6 concluded that 49% of nonsense SNPs would be predicted to elicit NMD whereas 51% would be
7
8 predicted to yield truncated proteins. Some of these nonsense SNPs have been found to occur in
9
10 the homozygous state in normal populations [Yngvadottir et al., 2009] attesting to the likely
11
12 functional redundancy of the corresponding genes. At the very least, genes harbouring nonsense
13
14 SNPs may be assumed to be only under weak selection [Ng et al., 2008].
15
16

17
18 It should be appreciated that nonsense SNPs may even occur in 'essential' genes yet still fail to
19
20 come to clinical attention (or give rise to a detectable phenotype) if these genes are subject to
21
22 copy number variation (see *CNVs and copy number mutations* below) that masks any deleterious
23
24 consequences by ensuring an adequate level of gene expression from additional wild-type copies
25
26 either in *cis* or in *trans*. Thus, copy number variation might serve to 'rescue' the full or partial
27
28 loss of gene function brought about by the nonsense mutations, thereby accounting for the
29
30 occurrence of the latter at polymorphic frequencies. Consistent with this postulate, Ng et al.
31
32 [2008] reported that ~30% of nonsense SNPs occur in genes residing within segmental
33
34 duplications, a proportion some three-fold larger than that noted for synonymous SNPs. Genes
35
36 harbouring nonsense SNPs were also found to belong to gene families of higher than average
37
38 size [Ng et al., 2008] suggesting that some functional redundancy may exist between paralogous
39
40 human genes. In support of this idea, Hsiao and Vitkup [2008] reported that those human genes
41
42 which have a homologue with $\geq 90\%$ sequence similarity are ~3 times less likely to harbour
43
44 disease-causing mutations than genes with less closely related homologues. They interpreted
45
46 their findings in terms of 'genetic robustness' against null mutations, with the duplicated
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2 sequences providing 'back-up' by potentiating the functional compensation/complementation of
3
4 homologous genes in the event that they acquire deleterious mutations. Potential examples of
5
6 such functional redundancy in the human genome involve the genes for *CCL4* and *CCL4L1*
7
8 chemokines [Howard et al., 2004] and the Rab GTPase genes, *RAB27A* and *RAB27B* [Barral et
9
10 al., 2002]. In the mouse, the proportion of essential genes among gene duplicates is ~7% lower
11
12 than among singletons, implying that ~15% of single gene deletions that would otherwise be
13
14 lethal (or infertile) are actually viable (or fertile) as a consequence of functional compensation by
15
16 the duplicate gene copy (Liang and Li, 2009). This level of functional redundancy may be even
17
18 more pronounced for the most recently duplicated genes (Su and Gu, 2008).
19

20
21
22 **What proportion of the possible mutations within inherited disease genes is likely to be of**
23
24 **pathological significance?**
25

26 Human gene mutation is a highly sequence-specific process, irrespective of the type of lesion
27
28 involved. This has had important implications, not only for the nature and prevalence, but also
29
30 for the diagnosis of human genetic disease [Antonarakis and Cooper, 2007; Zhang et al., 2009a;
31
32 Mefford and Eichler, 2009]. Certain DNA sequences have been found to be hypermutable,
33
34 thereby providing important clues as to the nature of the endogenous mechanisms underlying
35
36 different types of human gene lesion, but also emphasizing the non-uniform nature of
37
38 mutagenesis [Antonarakis and Cooper, 2007]. Of course, human gene mutations also lack a
39
40 uniform distribution within genes for functional reasons [that are](#) related to the nature of the gene
41
42 product in question [Miller et al., 2003; Subramanian and Kumar, 2006].
43

44 [The vast majority of mutations listed in HGMD reside within the coding region \(86%\), the](#)
45
46 [remainder being located in either intronic \(11%\) or regulatory \(3%, promoter, untranslated or](#)
47

1
2 flanking regions) sequences. The question of the proportion of possible mutations within human
3
4 disease genes that are likely to be of pathological significance is very difficult to address because
5
6 it is dependent not only upon the type and location of the mutation but also upon the
7
8 functionality of the nucleotides involved (dependent in part upon the amino acid residues that
9
10 they encode) which is often hard to assess [Miller and Kumar, 2001; Ferrer-Costa et al., 2002;
11
12 Arbiza et al., 2006; Capriotti et al., 2008; Kumar et al., 2009; Li et al., 2009a]. In addition, some
13
14 types of mutation are likely to be much more comprehensively ascertained than others, making
15
16 observational comparisons between mutation types an inherently hazardous undertaking.

17
18 Recently, it has been demonstrated that multiple mutations may not be an infrequent cause of
19
20 human genetic disease (Chen et al., 2009b). Such multiple mutations may constitute the
21
22 signatures of transient hypermutability in human genes. This has raised serious concerns
23
24 regarding current practices in mutation screening, practices which are likely to have resulted in
25
26 either the neglect, or even the complete failure to detect, many potentially important secondary
27
28 mutations linked in *cis* to the putative primary pathological lesion. Since interactions may well
29
30 occur between genetic variants linked in *cis*, inadequacies in the current practice of mutation
31
32 screening could easily have contributed to the frequently observed inconsistencies in the
33
34 genotype-phenotype relationship.

35
36 The above notwithstanding, the question of the proportion of all possible mutations within
37
38 human disease genes that are likely to be of pathological significance is clearly of paramount
39
40 importance to medical diagnostics. However, the corollary to this question is the issue of whether
41
42 some mutations may have been overlooked in mutation screening programs because they are
43
44 located at some very considerable distance from the genes whose function they disrupt. These
45
46 related questions will be addressed in some detail below.

Deleted: Nevertheless

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 [Lopez-Bigas et al., 2005; Sanford et al., 2009]. In addition, one must be aware in this context that synonymous 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; Wang and Cooper, 2007; Sanford et al., 2009; Hunt et al., 2009]. Finally, it should be understood that whereas the deleteriousness of the average

Deleted: F

Deleted: of

Deleted: of

Deleted: however

Deleted: ilst

synonymous mutation is always likely to be less than that of that of a non-synonymous (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 non-synonymous 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 single nucleotide polymorphisms (SNPs), a sizeable number are of the insertion/deletion type. Disease-associated or functional polymorphic variants are generally

Deleted: 4

Deleted: 800

Deleted: 700

Deleted: T

Deleted: but

1
2 located in either gene regulatory (~25%) or coding regions (~60%) although it should be noted
3
4 that variants occurring outside of these regions may still have consequences for gene expression,
5
6 splicing, transcription factor binding etc. In addition, some functionally important SNPs are
7
8 associated with non-protein-coding genes [Borel and Antonarakis, 2008; Yang et al., 2008a].
9

Deleted: nevertheless

10 At present, ~55% of the polymorphic variants recorded in HGMD are 'disease-associated'.
11
12 However, even in cases where no disease association has yet been demonstrated, functional
13
14 polymorphisms that alter the expression of a gene or the structure/function of the gene product
15
16 are potentially very important. Although such a polymorphism may not appear to have any direct
17
18 and/or immediate clinical relevance, the respective data in HGMD could yet prove very valuable
19
20 in terms of understanding inter-individual differences in disease susceptibility.
21

22 23 24 *Intronic mutations*

25
26 Mutations that occur within the extended consensus sequences of exon-intron splice junctions
27
28 account for ~10% of all reported mutations logged in HGMD and are frequently encountered in
29
30 mutation screening studies [Krawczak et al., 2007]. However, mutations residing in other
31
32 intronic locations (including the canonical branchpoint sequence) may often go undetected
33
34 unless they induce aberrant splicing (e.g. exon skipping or cryptic splice site utilization) that is
35
36 readily distinguishable qualitatively or quantitatively from both normal and alternative splicing
37
38 (Coulombe-Huntington et al., 2009). Introns probably represent substantially larger targets for
39
40 functional mutations than has hitherto been recognized on account of their harbouring a
41
42 multiplicity of functional elements including intron splice enhancers and silencers, *cis*-acting
43
44 RNA elements that regulate alternative splicing [Wang et al., 2009a; Tress et al., 2007], and
45
46 potentially also *trans*-splicing elements [Akiva et al., 2006; Shao et al., 2006; Gingeras 2009], as
47
48
49

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).

Deleted: them

Deleted: intronic functional elements

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 exon-intron splice junctions. Thus, for example, in the case of the *NFI* gene (MIM# 162200), Pros et al. [2008] noted that ~5% of

[identified](#) pathological mutations that altered splicing were deep intronic mutations. [Among disease-causing lesions, pseudoexon activation is one 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 [Grant et al., 1996; Mann et al., 2001; Emison et al., 2005; Susa et al., 2008; Choi et al., 2008; [Fraser and Xie, 2009](#)], it is very likely that splicing-relevant intronic mutations will have been seriously under-ascertained 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.](#)

Deleted: P

Deleted: as a

Deleted: is also likely to be a more common cause of inherited disease than previously appreciated

Mutations residing within remote gene regulatory regions

Micro-lesions 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 or within introns [Cecchini et al., 2009]. [Some mutations are known to occur within 5' or 3' untranslated regions, and these latter elements may be less likely to be screened for mutations \[Chen et al.,](#)

Deleted: the

Deleted: or 3' untranslated regions

Deleted:

2006a; Chen et al., 2006b; Chatterjee and Pal, 2009] particularly if they are large [e.g. *MECP2*, MIM# 300005; Coutinho et al., 2007]. In the same vein, upstream open reading frames (uORF), 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].

~~Another 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 [Pennacchio et al., 2006; Maston et al., 2006; Zhang et al., 2007; Attanasio et al., 2008; Kleinjan and Coutinho, 2009; Visel et al., 2009; Heintzman and Ren, 2009] 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 micro-lesions 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 (*SHH*; MIM# 600725) gene [Gordon et al., 2009].

Far upstream polymorphic variants that influence gene expression and which 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 (*IRF6*; MIM# 607199) gene, associated with cleft palate, alters the binding of transcription factor AP-2 α [Rahimov et al., 2008]. Similarly, a functional SNP ~6 kb upstream of the α -globin-like *HBM* (MIM# 609639) gene serves to create

Deleted: In addition

Deleted: ,

Deleted: eight

1
2 a binding site for the erythroid-specific transcription factor GATA1 and interferes with the
3
4 activation of the downstream α -globin genes [De Gobbi et al., 2006]. A functional SNP ~335 kb
5
6 upstream of the *MYC* (MIM# 190080) gene increases the risk of colorectal and prostate cancer
7
8 by [increasing the expression of the *MYC* gene by](#) altering the binding strength of transcription
9
10 factors TCF4 and/or TCF7L2 to a transcriptional enhancer [Haiman et al., 2007; Tuupanen et al.,
11
12 2009; Pomerantz et al., 2009; [Wright et al., 2010](#)]. Finally, in the context of pointing out the
13
14 shortcomings of the gene-centric approach to mutation detection, we should be aware that
15
16 functional SNP rs4988235, located 13.9 kb upstream of the lactase (*LCT*; MIM# 603202) gene
17
18 and associated with adult-type hypolactasia, actually resides deep within intron 13
19
20 (c.1917+326C>T) of the minichromosome maintenance complex component 6 (*MCM6*; MIM#
21
22 601806) gene [Enattah et al., 2002; Olds and Sibley, 2003; Lewinsky et al., 2005]. [Given that up
23
24 to 5% of quantitative trait loci for gene expression \(eQTLs\) lie >20 kb upstream of
25
26 transcriptional initiation sites \(Veyrieras et al., 2008\), many more far upstream polymorphic
27
28 variants that influence gene expression are likely to be identified in the coming years.](#)
29
30 [Rather](#) fewer pathological mutations are known to be located at a considerable distance
31
32 downstream of human genes. One example is the C>G transversion 2528 nt 3' to the Term codon
33
34 of the *CDK5RI* gene, which has been postulated to play a role in non-specific mental retardation
35
36 [Venturin et al., 2006]. [Perhaps more dramatic is the A>G SNP \(rs2943641\), 565981 bp 3' to the
37
38 Term codon of the *IRS1* gene, which is associated with type 2 diabetes, insulin resistance and
39
40 hyperinsulinemia; the G allele was found to be associated with a reduced basal level of *IRS1*
41
42 protein \[Rung et al., 2009\].](#)

43
44 Remote regulatory elements [have](#) sometimes come to attention as a consequence of their
45
46 removal by gross deletions located at some considerable distance (from 10 kb to several
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Deleted: Even

1
2 megabases) from the genes whose expression they disrupt (Table 3). Thus, for example, a 960 kb
3
4 deletion of non-coding sequence, lying between 1.477 Mb and 517 kb upstream of the *SOX9*
5
6 gene (MIM# 608160) gives rise to the acampomelic form of campomelic dysplasia [Lecointre et
7
8 al., 2009]. Such pathological deletions are however not necessarily always so large. Indeed, a 7.4
9
10 kb deletion, located 283 kb upstream of the *FOXL2* gene (MIM# 605597) has been identified as
11
12 a cause of blepharophimosis syndrome; it disrupts a long non-coding RNA (*PISRT1*) as well as
13
14 eight conserved non-coding sequences [D'haene et al., 2009; Table 3]. For some conditions, such
15
16 lesions may actually occur quite frequently, as in the case of the *SHOX* gene (MIM# 312865)
17
18 where ~22% of Leri-Weill syndrome patients (MIM# 127300) and ~1% of individuals with
19
20 idiopathic short stature (MIM# 300582) harbour a microdeletion spanning the upstream enhancer
21
22 region that leaves the coding region of the *SHOX* gene intact [Chen et al., 2009a].

23
24 In this context, it is interesting to note that developmental genes appear to be
25
26 disproportionately represented among those human genes located within 'gene deserts' [i.e. those
27
28 chromosomal regions which are devoid of annotated genes; Taylor, 2005; Ovcharenko et al.,
29
30 2005] and are often separated from their regulatory elements by up to several hundred kilobases.
31
32 The remote regulatory elements of several such genes (viz. *BMP2*, *PAX6*, *SHH*, *SHOX* and
33
34 *SOX9*) are known to be subject to deletion or gross rearrangement resulting in inherited disease
35
36 (Table 3).

37
38 Given that the number of transcriptional initiation sites in the human genome is much greater
39
40 than the number of genes [Carninci et al., 2006], it may well be that the number of regulatory
41
42 sequences associated with human genes has been seriously under-estimated. Further, both *cis*-
43
44 and *trans*-acting variation within regulatory regions may serve to modify gene expression and/or
45
46 the functional effects of protein coding variants [Stranger et al., 2005; Pastinen et al., 2006;
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2 Stranger et al., 2007; Dimas et al., 2008; Dimas et al., 2009]. The under-ascertainment of
3
4 disease-associated mutations within regulatory regions is therefore likely to be quite substantial
5
6 but can potentially be rectified by emerging high-throughput entire genome sequencing protocols
7
8 [Chorley et al., 2008].
9

10 11 *CNVs and copy number mutations*

12
13
14 No mention of the human germline mutational spectrum would be complete without making
15
16 reference to copy number variants (CNVs). CNVs are a recently discovered form of genomic
17
18 diversity involving DNA sequences ≥ 1 kb in length that are present in the human genome in a
19
20 variable number of copies [Iafate et al., 2004; Sebat et al., 2004, Redon et al., 2006]. Such gross
21
22 duplications/deletions are not only rather abundant but also often occur at polymorphic
23
24 frequencies. The Database of Genomic Variants (<http://projects.tcag.ca/variation>; August 2009)
25
26 currently lists 8,410 CNV loci (CNV loci represent genomic regions that harbour copy number
27
28 variants) and their number is increasing steadily, fuelled by refined analytical methods and the
29
30 ongoing characterization of this type of genomic variation in different human populations [Kidd
31
32 et al., 2008]. Conrad et al. [2010] have generated a comprehensive map of >8,500 validated
33
34 CNVs >500 bp (detected in 41 Europeans/West Africans) that together cover a total of 112.7 Mb
35
36 (3.7% of the genome). These authors estimated that 39% of the validated CNVs overlapped 13 %
37
38 of RefSeq genes (NCBI mRNA reference sequence collection). Further, they concluded that the
39
40 CNVs detected resulted in the “unambiguous loss of function” of alleles for 267 different genes.

41
42 It is important to note that the mutation rate (per locus and per generation) is considerably
43
44 higher for CNVs (3×10^{-2} to 10^{-7}) than for SNPs (10^{-7} to 10^{-8}) [Redon et al., 2006; Conrad et al.,
45
46 2010], no doubt due to the very different mutational mechanisms involved. In their very
47

Deleted: 2009

Deleted: 09

1
2 [comprehensive treatment of this issue, Conrad et al. \[2010\] attempted to estimate the average](#)
3 [per-generation rate of CNV formation. However, rate estimates were found to vary by several](#)
4 [orders of magnitude between sites. Conrad et al. \[2010\] further noted that these estimates did not](#)
5 [allow for purifying selection, and so they probably represent “a lower bound on the true rate”.](#)
6 [There is also an ascertainment bias to contend with, duplications being significantly harder to](#)
7 [identify than deletions \[Quemener et al., 2010\].](#)

Deleted: 09

Deleted: 09

8
9
10
11
12
13
14 It has been estimated that on average, 73 to 87 genes vary in copy number between any two
15 individuals [Alkan et al., 2009]. This high degree of inter-individual variability with regard to
16 gene copy number has challenged traditional definitions of wild-type and ‘normality’, and even
17 the very concept of a ‘reference genome’ itself [Dear, 2009]. High resolution breakpoint
18 mapping is a prerequisite for the accurate assessment of CNV size, the identification of the genes
19 and regulatory elements affected, and hence for the determination of the consequences of copy
20 number variation for gene expression and the phenotypic *sequelae* [Beckmann et al., 2008; de
21 Smith et al., 2008]. This notwithstanding, it is already becoming clear that these consequences
22 may go far beyond the physical bounds of a given CNV. Thus, a CNV involving the human *HBA*
23 (MIM# 141800) gene has a dramatic influence on the expression of the *NME4* (MIM# 601818)
24 gene some 300 kb distant (Lower et al., 2009). [In addition, a 5.5 kb micro-duplication of a](#)
25 [conserved noncoding sequence with demonstrated enhancer function, about 110 kb downstream](#)
26 [of the *BMP2* \(bone morphogenic protein 2\) gene, has been found to cause brachydactyly type 2A](#)
27 [in two families \[Dathe et al., 2009\].](#)

28
29
30
31
32
33
34
35
36
37
38
39
40
41
42 It may well be that the precise extent and/or location of many CNVs will vary between
43 individuals, thereby further increasing both [the](#) mutational and phenotypic heterogeneity. The
44 extent to which CNVs are likely to contribute to the diversity of human phenotypes, including
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

'single gene defects', genomic disorders and complex disease, is increasingly being recognized.

Deleted: and

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; Scherer and Lee, 2010 and references therein]. Remarkably, an excess

of both rare and *de novo* CNVs has been identified in patients with psychiatric disorders and

obesity [Sebat et al., 2007; Stefansson et al., 2008; Walsh et al., 2008; Elia et al., 2009; Glessner

et al., 2009; Bochukova et al., 2010; Walters et al. 2010]. These recent findings point to genetic

heterogeneity in these conditions ~~thereby illustrating the likely complexity inherent in identifying~~

Deleted: and hence serve to

Deleted: e

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

1
2
3 [which predisposes to various neuropsychiatric phenotypes as a single copy number mutation and](#)
4 [aggravates neurodevelopmental disorders if it co-occurs together with other large deletions and](#)
5 [duplications \[Girijaan et al., 2010\].](#) It remains to be seen whether ‘CNV equivalents’ <1 kb in
6 size (also termed ‘indels’), that actually occur rather more frequently than true CNVs (>1 kb)
7
8 [Conrad et al., 2010], will also be relevant to disease. What is already clear is that, over the
9 coming years, a large number of important CNV-disease associations are going to come to light
10 [Stankiewicz et al., 2010].
11
12
13
14
15

Deleted: ;

Deleted: 09

Deleted:)

16 17 18 *Mutations in non-protein-coding genes*

19
20 In contrast to the plethora of mutations identified in protein-coding genes, the identification of
21 mutations in non-protein-coding genes is still very much in its infancy. A [number](#) of disease-
22 causing or disease-associated mutations have [already](#) been reported in various small nucleolar
23 RNA genes and microRNA genes (Table 4). In addition, mutations have also been documented
24 in the longer non-coding RNA genes [*XIST* (MIM# 314670), *TERC* (MIM# 602322), *H19*
25 (MIM# 103280), *RMRP* (MIM# 157660); see HGMD for details]. A putative pathological
26 mutation has been described in a ‘gene’ encoding a paternally expressed antisense transcript of
27 the *GNAS* complex locus [*GNASAS*; MIM# 610540; Bastepe et al., 2005] whilst a functional
28 polymorphism has been reported within an enhancer at the 3’ end of the *CDKN2BAS* ‘gene’
29 (MIM# 600431) which encodes an antisense RNA transcript [Jarinova et al., 2009]. A *CRYGEP1*
30 (MIM# 123660) pseudogene-reactivating mutation associated with hereditary cataract formation
31 [Brakenhoff et al., 1994] probably also falls into this category.
32
33
34
35
36
37
38
39
40
41
42
43

Deleted: handful

Deleted: so far

44 The above examples are likely to comprise only the tip of a fairly large iceberg that still
45 remains essentially unexplored. [Thus, for example, both single nucleotide polymorphism and](#)
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2 [copy number variation are both likely to impact significantly on microRNA gene expression with](#)
3
4 [myriad potential pathological consequences \(Bandiera et al., 2010\).](#)

5
6
7
8 *Mutations in non-coding regions of functional significance*

9
10 By adopting a gene-centric view, we have until now largely ignored the extensive non-protein-
11 coding portion of the human genome in our quest for mutations of pathological significance. As
12 a consequence, we have not only seriously under-estimated the extent of the functional
13 component of the genome, but may also have overlooked many mutations within this genomic
14 'dark matter' [Collins and Penny, 2009]. As we increasingly adopt 'genotype-first' strategies to
15 characterizing genetic defects in patients with diverse clinical phenotypes (Mefford, 2009), many
16 more mutations are likely to be identified in non-protein-coding genes.

Deleted: ¶

17
18 In both the human and mouse genomes, many non-coding regions exhibit a similar level of
19 evolutionary conservation to that evident in protein-coding regions [Kryukov et al., 2005;
20 Asthana et al., 2007]. As yet, however, little is known of the effect that mutations in these
21 regions might have on either the phenotype or on overall fitness. Studies of the most
22 evolutionarily conserved non-coding regions have yielded results which are consistent with the
23 view that most mutations in non-coding regions are only slightly deleterious [Kryukov et al.,
24 2005; Chen et al., 2007]. The conservation observed may thus be due to variations in the
25 mutation rate rather than selective constraints [Keightley et al., 2005; Gorlov et al., 2008].
26
27 Indeed, Keightley et al. [2005] have shown that selection in conserved non-coding sequences is
28 significantly weaker in hominids as compared to murids, probably a consequence of the low
29 effective population size of hominids resulting in the reduced effectiveness of selection.

Deleted: ed

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

In order to obtain a first, necessarily rather crude, estimate of the contribution of variation in human non-coding sequences to phenotypic and/or disease traits, Visel et al. [2009] performed a meta-analysis of ~1200 single nucleotide polymorphisms (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 non-coding 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 non-coding 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 [Hindorff et al. 2009; Glinskii et al., 2009]. This notwithstanding, it should be appreciated that any given variant apparently detected within a non-coding region may actually reside within a hitherto undiscovered exon [Denoeud et al., 2007]. We should however be aware that rare variants, in cis to those found to be associated with a given disease or trait in GWAS studies, may by chance give rise to 'synthetic associations' that are then attributed to much more common variants [Dickson et al., 2010].

Deleted: may

Deleted: (

Deleted:)

Deleted: may

Deleted: Hindorff et al. [2009] reported

Deleted: were

Deleted: however

Deleted: through chance

Deleted: associations

Deleted: credited

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

1
2 recently, Kondrashov et al. [2002] compared pathological missense mutations in 32 human
3 proteins to the amino acid substitutions that occurred during the course of evolution of these
4 same proteins, and estimated that ~10% of all amino acid sequence differences between a human
5 protein and its non-human (mammalian) orthologue could represent what they termed
6 'compensated pathogenic deviations' (CPDs). Since such amino acid substitutions are
7 pathogenic in humans, Kondrashov et al. [2002] surmised that the normal functioning of a CPD-
8 containing protein in the non-human species must be due to other ('compensatory') amino acid
9 sequence deviations from the human sequence.
10

11 Numerous examples of CPDs have now been reported from comparative genome sequencing
12 studies. CPDs represent human pathological missense mutations where the substituting amino
13 acids have been found to be identical to the wild-type amino acid residues at the orthologous
14 positions in e.g. mouse [Gao and Zhang, 2003], macaque [Gibbs et al., 2007] and chimpanzee
15 [Azevedo et al., 2006; Suriano et al., 2007]. In principle, these compensatory changes could be
16 either allelic to the CPD (and hence closely linked genetically) or non-allelic (e.g. involving the
17 coevolution of a ligand and its receptor encoded by unlinked genes; Liu et al., 2001). The above
18 notwithstanding, in evolution, compensatory mutations are unlikely to occur singly; indeed, Poon
19 et al. [2005] have suggested that, on average, 11.8 compensatory mutations may interact
20 epistatically with a given deleterious mutation so as to restore wild-type levels of fitness.
21

22 CPDs tend to be less severe in terms of the difference in physicochemical properties between
23 the substituted and substituting amino acids than is normally the case for pathological mutations
24 [Ferrer-Costa et al., 2007; Barešić et al., 2010]. In the context of human disease, Suriano et al.
25 [2007] have provided a good example of the influence of compensated and compensating
26 mutations in the *OTC* gene. The human and chimpanzee *OTC* sequences differ at only two
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2 positions, amino acid residues 125 and 135. Amino acid replacements Thr135Ala and
3
4 Thr125Met have respectively occurred in the human and chimpanzee lineages since their
5
6 divergence from their common ancestor. The Thr135Ala substitution appears to be human-
7
8 specific whilst the Thr125Met substitution was chimpanzee-specific (both Thr125 and Thr135
9
10 were found to be ancestral residues). When the derived Met125 is associated with the ancestral
11
12 Thr135 (in chimpanzee), no abnormal phenotype is evident. However, when Met125 occurs on a
13
14 background containing the human-specific Ala135 residue, this results in a clinical phenotype
15
16 (neonatal hyperammonemia). Suriano et al. [2007] demonstrated *in vitro* that human OTC
17
18 bearing the Thr125Met mutant is inactive, whereas the chimpanzee version of OTC (with Met at
19
20 residue 125) possesses an enzymatic activity comparable with the wild-type human OTC. The
21
22 presence of Thr at position 135 in chimpanzees therefore rescues the deleterious effect of Met at
23
24 position 125 through intra-locus compensation.

25
26 The high proportion of disease-associated/functional SNPs that constitute CPDs in non-human
27
28 primates may have important implications for the study of complex disease in humans. With
29
30 mendelian diseases, the norm is for the pathological mutations to be new (i.e. derived) and in
31
32 many cases, this paradigm can be extended to common disease. However, there are some curious
33
34 examples in which the alleles that increase the risk of common disease are ancestral whilst the
35
36 derived alleles are 'protective' [Di Rienzo & Hudson, 2005]. This reversal of the standard model
37
38 is consistent with the idea that some forms of common disease susceptibility may be a
39
40 consequence of ancient human adaptations to a long-term stable environment ('thrifty alleles');
41
42 with a changed environment consequent to the recent shift to a modern lifestyle, these ancestral
43
44 alleles have now come to increase the risk of common disease [Di Rienzo & Hudson, 2005].
45
46 Thus, the ancestral alleles represent the recapitulation of ancient states that may once have been
47
48

1
2 protective, but which now result in adverse consequences for human health. On the other hand,
3
4 some ancestral alleles may be weakly deleterious mutations that have become fixed by genetic
5
6 drift [Kryukov et al., 2007], a process that may be facilitated by small effective population size).
7
8 Viewed within this evolutionary framework, new (derived) alleles may be expected to confer
9
10 'protection' against disease. Whilst ancestral alleles constitute only a minority of all putative risk
11
12 variants, their number nevertheless appears to be sufficiently high for us to conclude that they are
13
14 likely to account for a sizeable proportion of inherited susceptibility to common disease.
15

Deleted: ¶

16
17
18 *A mutation in search of a gene*
19

20 As is evident from the above, mutation hunting has so far been almost invariably gene-centric.
21
22 Once a disease gene is discovered, the identification and characterization of pathological
23
24 mutations within this gene usually follows apace. Generally speaking, the occasional exception
25
26 serves only to prove the rule. Such an exception is fascioscapulohumeral muscular dystrophy
27
28 (FSHD; MIM# 158900). The mutation responsible for this disease has long been known to be the
29
30 deletion of a critical number of units of a repeat sequence (D4Z4) on chromosome 4q35. This
31
32 deletion appears to correlate with the derepression of transcription of muscle-expressed genes in
33
34 the vicinity of the D4Z4 repeats. However, although various candidates have been proposed
35
36 [Dixit et al., 2007; Snider et al., 2009; Klooster et al., 2009], the identity and location of the
37
38 FSHD gene (or genes) still remain elusive, as does the disease mechanism. It is anticipated that
39
40 further examples of disease-associated mutations lacking an immediately obvious relationship to
41
42 a specific gene or genes will come to light as our mutation-searching procedures become less
43
44 gene-centric and more all-genome encompassing.
45

1
2 *Refocusing our attention on the 'functionome'*

3
4 In the context of identifying genetic variants responsible for human inherited disease, we
5
6 believe that it will be increasingly important to consider functional elements in the genome (the
7
8 'functionome') rather than simply genes *per se*. We employ the term 'functionome' here to
9
10 describe the totality of the biologically functional nucleotide sequences in the human genome,
11
12 irrespective of whether they are associated with genes or not. A number of novel techniques,
13
14 such as chromatin immunoprecipitation [ChIP; Wong and Wei, 2009] and ChIP-sequencing
15
16 [ChIP-Seq; Park, 2009], which are capable of exploring protein-DNA interactions at a genome-
17
18 wide (and protein-RNA interactions at a transcriptome-wide) level, are in the vanguard of
19
20 attempts to characterize the human 'functionome'. Since conserved non-coding sequences in the
21
22 human genome appear to be ~10-fold more abundant than known genes [Attanasio et al., 2008],
23
24 it is likely that (i) currently known mutations within coding regions are unlikely to be fully
25
26 representative of the universe of pathological mutations (which would imply that any
27
28 extrapolation from HGMD data would be highly speculative) and (ii) a whole new grouping of
29
30 disease-causing mutations may await identification and characterization. Once again, a paradigm
31
32 shift in our thinking may well be required if we are to maximize the potential of the emerging
33
34 high-throughput technology to detect new (hitherto latent) types of human gene mutation.

Deleted: c

35
36 The above notwithstanding, it is rather unlikely that the functional non-protein-coding portion
37
38 of the human genome will prove to be quite as mutation-dense as the protein-coding portion. For
39
40 most inherited disorders, the mutation detection rate is already fairly high (>90%) although this
41
42 success rate is often achieved by combining different mutation detection methodologies, for
43
44 example to screen for exon deletions and copy number variants as well as more subtle lesions
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

[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

Deleted: The credibility of the above calculations notwithstanding, w

1
2 al., 2007] would impact adversely upon protein function. Wheeler et al. [2008] employed
3
4 *PolyPhen* to predict that some 20% of 3,898 nonsynonymous variants (~780) detected in the
5
6 Watson genome would be “probably or possibly damaging” to protein function. Ng et al. [2009]
7
8 reported 2,227 “probably damaging” and 3,368 “possibly damaging” variants predicted by
9
10 *PolyPhen* from 13,295 non-synonymous variants detected in 12 human ‘exomes’ (comprising
11
12 180,000 exons per genome and corresponding to the 30 Mb protein coding region). *PolyPhen* has
13
14 also been used to identify 765 “possibly damaging” SNPs and 454 “probably damaging” SNPs in
15
16 the genome of a Yoruba (Nigerian) individual [McKernan et al., 2009]. Using a likelihood ratio
17
18 test, Chun and Fay [2009] examined the genomes of Venter, Watson and a Han Chinese male
19
20 (whose sequence had been reported by Wang et al., 2008a) and identified between 796 and 837
21
22 deleterious mutations per genome, ~15% of all nonsynonymous variants assessed; most of these
23
24 deleterious mutations were found to be heterozygous (76-83%) and individual-specific (~60%).
25
26 Chun and Fay [2009] estimated that their likelihood ratio test had been successful in identifying
27
28 62% of the “rare deleterious mutations” in the Venter genome. They also identified a further 838
29
30 deleterious mutations in the reference human genome [International Human Genome Sequencing
31
32 Consortium, 2004], 474 of which were specific to that (artificial multi-source) genome sequence
33
34 and absent from the other three genomes examined [Chun and Fay, 2009]. Interestingly, some
35
36 435 (23%) of the 1,928 putatively deleterious variants found in the Venter, Watson and Han
37
38 Chinese genomes were present in more than one of these genomes. Existing compilations of
39
40 mutation data, OMIM [Levy et al., 2007; Ng et al., 2008; Wang et al., 2008a; Kim et al., 2009;
41
42 McKernan et al., 2009; [Schuster et al., 2010](#)], HGMD [Venter et al., 2001; Wheeler et al., 2008;
43
44 Kim et al., 2009; McKernan et al., 2009; [Rasmussen et al., 2010](#)], SWISS-PROT [Chun and Fay,
45
46 2009] or SNPedia [Kim et al., 2009], have also been used directly to identify potential disease-
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

associated mutations in the various sequenced genomes. However, since 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].

Deleted: (

Deleted:)

Deleted: (

Deleted:)

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, 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. If the heterozygosity effect were an order of magnitude smaller, 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. The above 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

Deleted: W

Deleted: , however,

Deleted: One

Deleted: could be

1
2 increase in the fitness of (homozygous or heterozygous) carriers of the mutations in question
3
4 would serve to increase the expected number of such mutations to be identified in a given
5
6 individual by a corresponding amount.
7

8 Regarding the different outcomes of the sequencing studies, it must be noted that the levels of
9
10 sequencing coverage (7x to 40x; Yngvadottir et al., 2009) differed quite dramatically between
11
12 studies as did the portions of the genomes sequenced (i.e. 'entire' genome v. exome).
13

14 Furthermore, the different sequencing platforms employed exhibit very different error patterns

15 and rates [[Smith et al., 2008](#); [Wheeler et al., 2008](#)]. Also, the different deleterious variant

Deleted: (

Deleted:)

16 prediction tools used for functional profiling can differ quite markedly in terms of their

17 sensitivity and specificity [Ng and Henikoff, 2006]. Finally, it should be appreciated that the

18 question of ethnicity may impact significantly on question of deleterious gene diversity. Thus,

19 whilst African-Americans exhibit a higher level of heterozygosity for both 'possibly damaging'

20 and 'probably damaging' SNPs than European-Americans, European-Americans possess

21 significantly more genotypes which are homozygous for the putatively damaging allele of

22 'probably damaging' SNPs than do African-Americans [Lohmueller et al., 2008].
23
24
25
26
27
28
29

30 To [optimize their](#) practical utility, the bioinformatics tools available for the prediction of

Deleted: acquire true

31 deleterious mutations [[Karchin, 2009](#)] will need to be improved by the inclusion of data on

32 specific sites of structural and/or functional interest [[Mort et al., 2010](#)] and by consideration of

Deleted: e.g. protein binding sites, sites of post-translational modification etc;

Deleted: the

33 such key issues as mutation penetrance [Waalén and Beutler, 2009] and interactions between

34 allelic and non-allelic mutations/polymorphic variants [Dimas et al., 2008]. However, it is most

35 encouraging that [existing](#) bioinformatics tools have already been successfully applied in the

36 context of filtering whole-exome/[genome](#) sequencing data to identify the pathological mutations

Deleted: a

37 underlying [rare Mendelian disorders](#) of previously unknown cause [Ng et al., 2010; [Lupski et al.](#),
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

[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 cross-comparisons can be made.

Deleted: , preferably comprehensive,

Can we estimate the number of mutations causing human inherited disease that still remain to be characterized?

Formatted: Line spacing: Double

“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].

Formatted: Right: 0 pt, Line spacing: Double

Since 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 outweigh 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 [Jones et al., 2009; Ng et al., 2009; Choi et al., 2009; Tucker et al., 2009] and high-throughput targeted resequencing of defined genomic regions [Prabhu and Pe'er, 2009; Kryukov et al., 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.

Deleted: (

Deleted:)

1
2 This will be an era in which, for each patient, many genomic variants 'will be called but few will
3 be chosen'. Hence, the development of bioinformatics techniques, sufficiently powerful to
4 identify, with a high degree of certainty, pathological needles in the human genomic haystack,
5 will be paramount. However, in deploying these emerging techniques, we should be wary of
6 being constrained by outmoded overly gene-centric approaches to mutation screening. Once
7 again, in terms of mutation hunting, we should not focus exclusively on genes *per se* but rather
8 shift our emphasis so as to include the sequence elements that characterize a potentially larger
9 (and yet still functional) portion of the genome. Expanding our horizons through the inclusion of
10 new types of functional element among our screening targets should serve to extend the known
11 germline mutational spectrum very significantly. We predict that entirely new types of
12 pathological gene lesion (the unknown unknowns!) are likely to become apparent whose
13 characterization should provide new insights not only into the morbid anatomy of the human
14 genome but also its normal structure and function.

30 Concluding remarks

31 In summary, the number of germline mutations in human nuclear genes known to either cause
32 or to be associated with inherited disease now exceeds 100,000 in over 3,700 different genes.
33 Newly described human gene mutations are currently being reported at a rate of ~10,000 per
34 annum, with ~300 new 'inherited disease genes' being recognised every year. As the human
35 'mutome' passes the historic 100,000 landmark, we have posed the double question: how many
36 inherited disease genes are there in the human genome and how many mutations are likely to be
37 found within them? The total number of genes present in the human genome is dependent in part
38 upon one's operating definition of a gene but appears to be at least 25,000 and may yet be found

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

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

Deleted: Finally

1
2 principle be extended beyond the narrow confines of DNA sequence-based changes so as to
3
4 include heritable (germline) alterations of DNA methylation ('epimutations') that result in
5
6 abnormal transcriptional silencing [Cropley et al., 2008]. The best example of this phenomenon
7
8 is provided by the constitutional epimutations in the human *MLH1* (MIM# 120436) gene which
9
10 cause hereditary non-polyposis colorectal cancer [Hitchins and Ward, 2009]. With the
11
12 determination of the human methylome [Lister et al., 2009], the number of recognized
13
14 epimutations should rise significantly in the coming years. If eventually shown to be both of
15
16 pathological significance and heritable, some examples of histone modification [Wang et al.,
17
18 2008b; VerMilyea et al., 2009; [Luco et al., 2010](#)] or RNA editing [Li et al., 2009b; [Lualdi et al.,](#)
19
20 [2010](#)] could also turn out to represent 'honorary mutations'.

21
22 Irrespective of how the human germline mutational spectrum transmogrifies over the coming
23
24 years, we must remain committed to collating human gene mutation data as they emerge,
25
26 endeavouring as we do so to follow the advice of the founder of modern human genetics,
27
28 William Bateson, who, in the context of collecting plant mutants over a century ago, exhorted us
29
30 to "treasure your exceptions".
31

32 33 34 35 36 **References**

37
38
39
40 Aerts S, Lambrechts D, Maity S, Van Loo P, Coessens B, De Smet F, Tranchevent LC, De Moor
41
42 B, Marynen P, Hassan B, Carmeliet P, Moreau Y. 2006. Gene prioritization through genomic
43
44 data fusion. *Nat Biotechnol* 24:537-544.
45

1
2 Akiva P, Toporik A, Edelheit S, Peretz Y, Diber A, Shemesh R, Novik A, Sorek R. 2006.

3
4 Transcription-mediated gene fusion in the human genome. *Genome Res* 16:30-36.

5
6
7
8 Alkan C, Kidd JM, Marques-Bonet T, Aksay G, Antonacci F, Hormozdiari F, Kitzman JO,

9
10 Baker C, Malig M, Mutlu O, Sahinalp SC, Gibbs RA, Eichler EE. 2009. Personalized copy
11 number and segmental duplication maps using next-generation sequencing. *Nat Genet* 41:1061-
12 1067.

13
14
15
16
17
18 Antonarakis SE, Cooper DN. 2007. Mutations in human genetic disease. Nature and
19 consequences. Chap. 7 in *Principles and Practice of Medical Genetics*. 5th Ed. DL Rimoim, JM
20 Connor, RE Pyeritz, BR Korf (Eds), Churchill Livingstone, Edinburgh, pp101-128.

21
22
23
24
25
26 Arbiza L, Duchi S, Montaner D, Burguet J, Pantoja-Uceda D, Pineda-Lucena A, Dopazo J,
27 Dopazo H. 2006. Selective pressures at a codon-level predict deleterious mutations in human
28 disease genes. *J Mol Biol* 358:1390-1404.

29
30
31
32
33
34 Asthana S, Noble WS, Kryukov G, Grant CE, Sunyaev S, Stamatoyannopoulos JA. 2007.
35 Widely distributed noncoding purifying selection in the human genome. *Proc Natl Acad Sci*
36 USA 104:12410-12415.

37
38
39
40
41
42 Attanasio C, Reymond A, Humbert R, Lyle R, Kuehn MS, Neph S, Sabo PJ, Goldy J, Weaver M,
43 Haydock A, Lee K, Dorschner M, Dermitzakis ET, Antonarakis SE, Stamatoyannopoulos JA.

1
2 2008. Assaying the regulatory potential of mammalian conserved non-coding sequences in
3
4 human cells. *Genome Biol* 9:R168.

5
6
7
8 [Azevedo L, Suriano G, van Asch B, Harding RM, Amorim A. \(2006\) Epistatic interactions: how
9 strong in disease and evolution? *Trends Genet* 22:581-585.](#)

Deleted:

10
11
12
13
14 [Bandiera S, Hatem E, Lyonnet S, Henrion-Caude A. 2010. microRNAs in diseases: from
15 candidate to modifier genes. *Clin Genet* In press. \[doi>10.1111/j.1399-0004.2010.01370.x,
16 PMID: 20132241\]](#)

Deleted:

17
18 [Barešić A, Hopcroft LE, Rogers HH, Hurst JM, Martin AC. 2009. Compensated pathogenic
19 deviations: analysis of structural effects. *J Mol Biol* 396:19-30.](#)

20
21
22
23
24
25
26
27
28 Barral DC, Ramalho JS, Anders R, Hume AN, Knapton HJ, Tolmachova T, Collinson LM,
29
30 Goulding D, Authi KS, Seabra MC. 2002. Functional redundancy of Rab27 proteins and the
31
32 pathogenesis of Griscelli syndrome. *J Clin Invest* 110:247-257.

33
34
35
36 Bastepe M, Fröhlich LF, Linglart A, Abu-Zahra HS, Tojo K, Ward LM, Jüppner H. 2005.
37
38 Deletion of the NESP55 differentially methylated region causes loss of maternal *GNAS* imprints
39
40 and pseudohypoparathyroidism type Ib. *Nat Genet* 37:25-27.

41
42
43
44 Beckmann JS, Sharp AJ, Antonarakis SE. 2008. CNVs and genetic medicine (excitement and
45
46 consequences of a rediscovery). *Cytogenet Genome Res* 123:7-16.

1
2
3
4 Benito-Sanz S, Thomas NS, Huber C, Gorbenko del Blanco D, Aza-Carmona M, Crolla JA,
5
6 Maloney V, Rappold G, Argente J, Campos-Barros A, Cormier-Daire V, Heath KE. 2005. A
7
8 novel class of pseudoautosomal region 1 deletions downstream of *SHOX* is associated with Leri-
9
10 Weill dyschondrosteosis. *Am J Hum Genet* 77:533-544.

11
12
13
14 Benko S, Fantes JA, Amiel J, Kleinjan DJ, Thomas S, Ramsay J, Jamshidi N, Essafi A, Heaney
15
16 S, Gordon CT, McBride D, Golzio C, Fisher M, Perry P, Abadie V, Ayuso C, Holder-Espinasse
17
18 M, Kilpatrick N, Lees MM, Picard A, Temple IK, Thomas P, Vazquez MP, Vekemans M,
19
20 Crollius HR, Hastie ND, Munnich A, Etchevers HC, Pelet A, Farlie PG, Fitzpatrick DR, Lyonnet
21
22 S. 2009. Highly conserved non-coding elements on either side of *SOX9* associated with Pierre
23
24 Robin sequence. *Nat Genet* 41:359-364.

25
26
27
28 Bird CP, Stranger BE, Liu M, Thomas DJ, Ingle CE, Beazley C, Miller W, Hurles ME,
29
30 Dermitzakis ET. 2007. Fast-evolving noncoding sequences in the human genome. *Genome Biol*
31
32 8:R118.

33
34
35
36 Bittles AH, Neel JV. 1994. The costs of human inbreeding and their implications for variations at
37
38 the DNA level. *Nat Genet* 8:117-121.

39
40
41
42 [Bochukova EG, Huang N, Keogh J, Henning E, Purmann C, Blaszczyk K, Saeed S, Hamilton-](#)
43
44 [Shield J, Clayton-Smith J, O'Rahilly S, Hurles ME, Farooqi IS. 2010. Large, rare chromosomal](#)
45
46 [deletions associated with severe early-onset obesity. *Nature* 463:666-670.](#)
47

1
2
3
4 Borel C, Antonarakis SE. 2008. Functional genetic variation of human miRNAs and phenotypic
5 consequences. *Mamm Genome* 19:503-509.
6
7

8
9
10 Borel C, Gagnebin M, Gehrig C, Kriventseva EV, Zdobnov EM, Antonarakis SE. 2008.
11 Mapping of small RNAs in the human ENCODE regions. *Am J Hum Genet* 82:971-981.
12
13

14
15
16 Boyko AR, Williamson SH, Indap AR, Degenhardt JD, Hernandez RD, Lohmueller KE, Adams
17 MD, Schmidt S, Sninsky JJ, Sunyaev SR, White TJ, Nielsen R, Clark AG, Bustamante CD.
18
19
20 2008. Assessing the evolutionary impact of amino acid mutations in the human genome. *PLoS*

21
22 Genet 4:e1000083.

Deleted: ¶

23
24
25
26 Brakenhoff RH, Henskens HA, van Rossum MW, Lubsen NH, Schoenmakers JG. 1994.

27
28 Activation of the gamma E-crystallin pseudogene in the human hereditary Coppock-like cataract.

29
30 *Hum Mol Genet* 3:279-283.

Deleted: ¶

31
32
33
34 Brown KK, Reiss JA, Crow K, Ferguson HL, Kelly C, Fritsch B, Morton CC. 2010. Deletion of
35 an enhancer near *DLX5* and *DLX6* in a family with hearing loss, craniofacial defects, and an

36
37
38 *inv(7)(q21.3q35)*. *Hum Genet* 127:19-31.

Deleted: ¶

39
40
41
42 Cai J, Goodman BK, Patel AS, Mulliken JB, Van Maldergem L, Hoganson GE, Paznekas WA,

43
44 Ben-Neriah Z, Sheffer R, Cunningham ML, Daentl DL, Jabs EW. 2003. Increased risk for
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2 developmental delay in Saethre-Chotzen syndrome is associated with *TWIST* deletions: an
3 improved strategy for *TWIST* mutation screening. Hum Genet 114:68-76.

Deleted: ¶

4
5
6
7
8 Cai JJ, Borenstein E, Chen R, Petrov DA. 2009. Similarly strong purifying selection acts on
9 human disease genes of all evolutionary ages. Genome Biol Evol 2009:131-144;
10 doi:10.1093/gbe/evp013.

Deleted: ¶
¶

11
12
13
14
15
16 Calin GA, Ferracin M, Cimmino A, Di Leva G, Shimizu M, Wojcik SE, Iorio MV, Visone R,
17 Sever NI, Fabbri M, Iuliano R, Palumbo T, Pichiorri F, Roldo C, Garzon R, Sevignani C,
18 Rassenti L, Alder H, Volinia S, Liu CG, Kipps TJ, Negrini M, Croce CM. 2005. A microRNA
19 signature associated with prognosis and progression in chronic lymphocytic leukemia. N Engl J
20 Med 353:1793-1801.

21
22
23
24
25
26
27
28 [Calvo SE, Pagliarini DJ, Mootha VK. 2009. Upstream open reading frames cause widespread](#)
29 [reduction of protein expression and are polymorphic among humans. Proc Natl Acad Sci USA](#)
30 [106:7507-7512.](#)

Deleted: ¶

31
32
33
34
35
36 Capriotti E, Arbiza L, Casadio R, Dopazo J, Dopazo H, Marti-Renom MA. 2008. Use of
37 estimated evolutionary strength at the codon level improves the prediction of disease-related
38 protein mutations in humans. Hum Mutat 29:198-204.

39
40
41
42
43
44 Carninci P, Sandelin A, Lenhard B, Katayama S, Shimokawa K, Ponjavic J, Semple CA, Taylor
45 MS, Engström PG, Frith MC, Forrest AR, Alkema WB, Tan SL, Plessy C, Kodzius R, Ravasi T,
46
47

1
2 Kasukawa T, Fukuda S, Kanamori-Katayama M, Kitazume Y, Kawaji H, Kai C, Nakamura M,
3
4 Konno H, Nakano K, Mottagui-Tabar S, Arner P, Chesi A, Gustincich S, Persichetti F, Suzuki H,
5
6 Grimmond SM, Wells CA, Orlando V, Wahlestedt C, Liu ET, Harbers M, Kawai J, Bajic VB,
7
8 Hume DA, Hayashizaki Y. 2006. Genome-wide analysis of mammalian promoter architecture
9
10 and evolution. *Nat Genet* 38:626-635.

Deleted:

11
12
13
14 Cartegni L, Chew SL, Krainer AR. 2002. Listening to silence and understanding nonsense:
15
16 exonic mutations that affect splicing. *Nat Rev Genet* 3:285-298.

17
18
19
20 Cecchini KR, Raja Banerjee A, Kim TH. 2009. Towards a genome-wide reconstruction of *cis*-
21
22 regulatory networks in the human genome. *Semin Cell Dev Biol* 20:842-848.

Deleted: ¶
¶

23
24
25
26 Chang JC, Kan YW. 1979. β^0 thalassemia, a nonsense mutation in man. *Proc Natl Acad Sci USA*
27
28 76:2886-2889.

29
30
31
32 [Chatterjee S, Pal JK. 2009. Role of 5'- and 3'-untranslated regions of mRNAs in human diseases.](#)
33
34 [Biol Cell 101:251-262.](#)
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2 [Chen JM, Férec C, Cooper DN. 2006a. A systematic analysis of disease-associated variants in](#)
3 [the 3' regulatory regions of human protein-coding genes I: general principles and overview. Hum](#)
4 [Genet 120:1-21.](#)

Deleted: ¶

5
6
7
8
9 [Chen JM, Férec C, Cooper DN. 2006b. A systematic analysis of disease-associated variants in](#)
10 [the 3' regulatory regions of human protein-coding genes II: the importance of mRNA secondary](#)
11 [structure in assessing the functionality of 3' UTR variants. Hum Genet 120:301-333.](#)

12
13
14
15
16
17 [Chen CT, Wang JC, Cohen BA. 2007. The strength of selection on ultraconserved elements in](#)
18 [the human genome. Am J Hum Genet 80:692-704.](#)

19
20
21
22
23 Chen J, Wildhardt G, Zhong Z, Röth R, Weiss B, Steinberger D, Decker J, Blum WF, Rappold
24 G. 2009a. Enhancer deletions of the *SHOX* gene as a frequent cause of short stature: the essential
25 role of a 250 kb downstream regulatory domain. J Med Genet 46:834-839.

26
27
28
29
30
31 [Chen JM, Férec C, Cooper DN. 2009b. Closely spaced multiple mutations as potential signatures](#)
32 [of transient hypermutability in human genes. Hum Mutat 30:1435-1448.](#)

33
34
35
36
37 Choi JW, Park CS, Hwang M, Nam HY, Chang HS, Park SG, Han BG, Kimm K, Kim HL, Oh
38 B, Kim Y. 2008. A common intronic variant of *CXCR3* is functionally associated with gene
39 expression levels and the polymorphic immune cell responses to stimuli. J Allergy Clin Immunol
40
41
42
43 122: 1119-1126.e7.

Deleted: ¶

1
2 Choi M, Scholl UI, Ji W, Liu T, Tikhonova IR, Zumbo P, Nayir A, Bakkaloglu A, Ozen S,
3 Sanjad S, Nelson-Williams C, Farhi A, Mane S, Lifton RP. 2009. Genetic diagnosis by whole
4 exome capture and massively parallel DNA sequencing. Proc Natl Acad Sci USA 106:19096-
5
6 19101.
7
8 Deleted: ¶
9 ¶

10
11 Chorley BN, Wang X, Campbell MR, Pittman GS, Nouredine MA, Bell DA. 2008. Discovery
12 and verification of functional single nucleotide polymorphisms in regulatory genomic regions:
13 current and developing technologies. Mutat Res 659:147-157.
14
15

16
17 Chun S, Fay JC. 2009. Identification of deleterious mutations within three human genomes.
18
19 Genome Res 19:1553-1561.
20
21

22
23 [Clamp M, Fry B, Kamal M, Xie X, Cuff J, Lin MF, Kellis M, Lindblad-Toh K, Lander ES. 2010.](#)
24 [Distinguishing protein-coding and noncoding genes in the human genome. Proc Natl Acad Sci](#)
25 [USA 104:19428-19433.](#)
26
27

28
29 Collins LJ, Penny D. 2009. The RNA infrastructure: dark matter of the eukaryotic cell? Trends
30
31 Genet 25:120-128.
32
33

34
35 Conley AB, Miller WJ, Jordan IK. 2008. Human *cis* natural antisense transcripts initiated by
36 transposable elements. Trends Genet 24:53-56.
37
38 Deleted:
39 ¶
40 ¶
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

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* In press.

Deleted: 09

[[doi 10.1038/nature08516](https://doi.org/10.1038/nature08516), [PMID: 19812545](https://pubmed.ncbi.nlm.nih.gov/19812545/)]

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 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.

[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.](#)

1
2 | Cropley JE, Martin DI, Suter CM. 2008. Germline epimutation in humans. *Pharmacogenomics*
3 |
4 | 9:1861-1868.

5
6
7
8 | Cutler G, Kassner PD. 2008. Copy number variation in the mouse genome: implications for the
9 | mouse as a model organism for human disease. *Cytogenet Genome Res* 123:297-306.

Deleted: .

10
11
12
13
14 | Dathe K, Kjaer KW, Brehm A, Meinecke P, Nürnberg P, Neto JC, Brunoni D, Tommerup N, Ott
15 |
16 | CE, Klopocki E, Seemann P, Mundlos S. 2009. Duplications involving a conserved regulatory
17 | element downstream of BMP2 are associated with brachydactyly type A2. *Am J Hum Genet*
18 |
19 | 84:483-492.

20
21
22
23
24 | De Gobbi M, Viprakasit V, Hughes JR, Fisher C, Buckle VJ, Ayyub H, Gibbons RJ, Vernimmen
25 |
26 | D, Yoshinaga Y, de Jong P, Cheng JF, Rubin EM, Wood WG, Bowden D, Higgs DR. 2006. A
27 | regulatory SNP causes a human genetic disease by creating a new transcriptional promoter.
28 |
29 | *Science* 312:1215-1217.

30
31
32
33
34 | de Kok YJ, Vossenaar ER, Cremers CW, Dahl N, Laporte J, Hu LJ, Lacombe D, Fischel-
35 |
36 | Ghodsian N, Friedman RA, Parnes LS, Thorpe P, Bitner-Glindzicz M, Pander HJ, Heilbronner
37 |
38 | H, Graveline J, den Dunnen JT, Brunner HG, Ropers HH, Cremers FP. 1996. Identification of a
39 | hot spot for microdeletions in patients with X-linked deafness type 3 (DFN3) 900 kb proximal to
40 | the DFN3 gene *POU3F4*. *Hum Mol Genet* 5:1229-1235.
41 |
42 |
43 |
44 |
45 |
46 |
47 |
48 |
49 |
50 |
51 |
52 |
53 |
54 |
55 |
56 |
57 |
58 |
59 |
60 |

1
2 de Smith AJ, Walters RG, Froguel P, Blakemore AI. 2008. Human genes involved in copy
3 number variation: mechanisms of origin, functional effects and implications for disease.
4

5
6 Cytogenet Genome Res 123:17-26.
7

8
9
10 Dear A, Daly J, Brennan SO, Tuckfield A, George PM. 2006. An intronic mutation within *FGB*
11 (IVS1+2076 a-->g) is associated with afibrinogenemia and recurrent transient ischemic attacks. J
12 Thromb Haemost 4:471-472.
13
14

15
16
17
18 Dear PH. 2009. Copy-number variation: the end of the human genome? Trends Biotechnol
19 27:448-454.
20

21
22
23
24 Denoeud F, Kapranov P, Ucla C, Frankish A, Castelo R, Drenkow J, Lagarde J, Alioto T,
25 Manzano C, Chrast J, Dike S, Wyss C, Henrichsen CN, Holroyd N, Dickson MC, Taylor R,
26 Hance Z, Foissac S, Myers RM, Rogers J, Hubbard T, Harrow J, Guigó R, Gingeras TR,
27 Antonarakis SE, Reymond A. 2007. Prominent use of distal 5' transcription start sites and
28 discovery of a large number of additional exons in ENCODE regions. Genome Res 17:746-759.
29
30

31
32
33
34
35
36 D'haene B, Attanasio C, Beysen D, Dostie J, Lemire E, Bouchard P, Field M, Jones K, Lorenz B,
37 Menten B, Buysse K, Pattyn F, Friedli M, Ucla C, Rossier C, Wyss C, Speleman F, De Paepe A,
38 Dekker J, Antonarakis SE, De Baere E. 2009. Disease-causing 7.4 kb *cis*-regulatory deletion
39 disrupting conserved non-coding sequences and their interaction with the *FOXL2* promoter:
40 implications for mutation screening. PLoS Genet 5:e1000522.
41
42
43
44 Deleted:

1
2 [Dhir A, Buratti E. 2010. Alternative splicing: role of pseudoexons in human disease and](#)
3 [potential therapeutic strategies. FEBS J 277:841-855.](#)
4
5

6
7
8 [Di Rienzo A, Hudson RR. 2005. An evolutionary framework for common diseases: the ancestral-](#)
9 [susceptibility model. Trends Genet 21:596-601.](#)
10
11

12
13
14 [Dickson SP, Wang K, Krantz I, Hakonarson H, Goldstein DB. 2010. Rare variants create](#)
15 [synthetic genome-wide associations. PLoS Biol 8:e1000294.](#)
16
17

18
19
20 Dimas AS, Stranger BE, Beazley C, Finn RD, Ingle CE, Forrest MS, Ritchie ME, Deloukas P,
21 Tavaré S, Dermitzakis ET. 2008. Modifier effects between regulatory and protein-coding
22 variation. PLoS Genet 4:e1000244.
23
24

25
26
27
28 Dimas AS, Deutsch S, Stranger BE, Montgomery SB, Borel C, Attar-Cohen H, Ingle C, Beazley
29 C, Gutierrez Arcelus M, Sekowska M, Gagnebin M, Nisbett J, Deloukas P, Dermitzakis ET,
30 Antonarakis SE. 2009. Common regulatory variation impacts gene expression in a cell type-
31 dependent manner. Science 325:1246-1250.
32
33

34
35
36
37 Dinger ME, Amaral PP, Mercer TR, Mattick JS. 2009. Pervasive transcription of the eukaryotic
38 genome: functional indices and conceptual implications. Brief Funct Genomic Proteomic 8:407-
39 423.
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Dixit M, Anseau 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. \$\gamma\delta\beta\$ -thalassemia due to a *de novo* mutation deleting the 5' \$\beta\$ -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.](#)

1
2 [Devoto M, Shaikh TH, Hakonarson H, White PS. 2009. Rare structural variants found in](#)
3 [attention-deficit hyperactivity disorder are preferentially associated with neurodevelopmental](#)
4 [genes. Mol Psychiatry In press. \[doi>10.1038/mp.2009.57, PMID: 19546859\]](#)
5
6
7
8

9
10 Emison ES, McCallion AS, Kashuk CS, Bush RT, Grice E, Lin S, Portnoy ME, Cutler DJ, Green
11 ED, Chakravarti A. 2005. A common sex-dependent mutation in a *RET* enhancer underlies
12 Hirschsprung disease risk. *Nature* 434:857-863.
13
14
15

16
17
18 Enattah NS, Sahi T, Savilahti E, Terwilliger JD, Peltonen L, Järvelä I. 2002. Identification of a
19 variant associated with adult-type hypolactasia. *Nat Genet* 30:233-237.
20
21
22

23
24 ENCODE Project Consortium, 2007. 2007. Identification and analysis of functional elements in
25 1% of the human genome by the ENCODE pilot project. *Nature* 447:799-816.
26
27
28

29
30 Eory L, Halligan DL, Keightley PD. 2010. Distributions of selectively constrained sites and
31 deleterious mutation rates in the hominid and murid genomes. *Mol Biol Evol* 27:177-192.
32
33
34

35
36 Eyre-Walker A, Keightley PD. 1999. High genomic deleterious mutation rates in hominids.
37 *Nature* 397:344-347.
38
39

40
41 Faghihi MA, Wahlestedt C. 2009. Regulatory roles of natural antisense transcripts. *Nat Rev Mol*
42 *Cell Biol* 10:637-643.
43
44
45
46
47
48

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Fanciulli M, Petretto E, Aitman TJ. 2010. Gene copy number variation and common human disease. *Clin Genet* [77:201-213](#).

Deleted: In press. doi10.1111/j.1399-0004.2009.01342.x

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.

1
2 [Ferrer-Costa C, Orozco M, de la Cruz X. 2007. Characterization of compensated mutations in](#)
3 [terms of structural and physico-chemical properties. J Mol Biol 365:249-256.](#)
4
5
6
7

8 Flomen RH, Vatcheva R, Gorman PA, Baptista PR, Groet J, Barisić I, Ligutic I, Nizetić D. 1998.
9
10 Construction and analysis of a sequence-ready map in 4q25: Rieger syndrome can be caused by
11 haploinsufficiency of *RIEG*, but also by chromosome breaks approximately 90 kb upstream of
12 this gene. *Genomics* 47:409-413.
13
14 Deleted:

15
16
17
18 [Fraser HB, Xie X. 2009. Common polymorphic transcript variation in human disease. *Genome*](#)
19 [Res 19:567-575.](#)
20
21
22
23

24 Frazer KA, Murray SS, Schork NJ, Topol EJ. 2009. Human genetic variation and its contribution
25 to complex traits. *Nat Rev Genet* 10:241-251.
26
27

28
29
30 Frio TR, McGee TL, Wade NM, Iseli C, Beckmann JS, Berson EL, Rivolta C. 2009. A single-
31 base substitution within an intronic repetitive element causes dominant retinitis pigmentosa with
32 reduced penetrance. *Hum Mutat* 30:1340-1347.
33
34
35

36
37
38 Furney SJ, Albà MM, López-Bigas N. 2006. Differences in the evolutionary history of disease
39 genes affected by dominant or recessive mutations. *BMC Genomics* 7:165.
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

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*. Unwin Hyman, London, pp. 325-343.

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, Chandrasekhar 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,](#)

1
2 [Nelson JO, Pepin KH, Qin X, Sutton GG, Venter E, Walenz BP, Wallis JW, Worley KC, Yang](#)
3 [SP, Jones SM, Marra MA, Rocchi M, Schein JE, Baertsch R, Clarke L, Csürös M, Glasscock J,](#)
4 [Harris RA, Havlak P, Jackson AR, Jiang H, Liu Y, Messina DN, Shen Y, Song HX, Wylie T,](#)
5 [Zhang L, Birney E, Han K, Konkel MK, Lee J, Smit AF, Ullmer B, Wang H, Xing J, Burhans R,](#)
6 [Cheng Z, Karro JE, Ma J, Raney B, She X, Cox MJ, Demuth JP, Dumas LJ, Han SG, Hopkins J,](#)
7 [Karimpour-Fard A, Kim YH, Pollack JR, Vinar T, Addo-Quaye C, Degenhardt J, Denby A,](#)
8 [Hubisz MJ, Indap A, Kosiol C, Lahn BT, Lawson HA, Marklein A, Nielsen R, Vallender EJ,](#)
9 [Clark AG, Ferguson B, Hernandez RD, Hirani K, Kehrer-Sawatzki H, Kolb J, Patil S, Pu LL,](#)
10 [Ren Y, Smith DG, Wheeler DA, Schenck I, Ball EV, Chen R, Cooper DN, Giardine B, Hsu F,](#)
11 [Kent WJ, Lesk A, Nelson DL, O'brien WE, Prüfer K, Stenson PD, Wallace JC, Ke H, Liu XM,](#)
12 [Wang P, Xiang AP, Yang F, Barber GP, Haussler D, Karolchik D, Kern AD, Kuhn RM, Smith](#)
13 [KE, Zwiig AS. 2007. Evolutionary and biomedical insights from the rhesus macaque genome.](#)
14 [Science 316:222-234.](#)
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29

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

31
32
33
34 Gingeras TR. 2009. Implications of chimaeric non-co-linear transcripts. *Nature* 461:206-211.

35
36
37
38 [Girirajan S, Rosenfeld JA, Cooper GM, Antonacci F, Siswara P, Itsara A, Vives L, Walsh T,](#)
39 [McCarthy SE, Baker C, Mefford HC, Kidd JM, Browning SR, Browning BL, Dickel DE, Levy](#)
40 [DL, Ballif BC, Platky K, Farber DM, Gowans GC, Wetherbee JJ, Asamoah A, Weaver DD,](#)
41 [Mark PR, Dickerson J, Garg BP, Ellingwood SA, Smith R, Banks VC, Smith W, McDonald MT,](#)
42 [Hoo JJ, French BN, Hudson C, Johnson JP, Ozmore JR, Moeschler JB, Surti U, Escobar LF, El-](#)
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 [Khechen D, Gorski JL, Kussmann J, Salbert B, Lacassie Y, Biser A, McDonald-McGinn DM,](#)
4 [Zackai EH, Deardorff MA, Shaikh TH, Haan E, Friend KL, Fichera M, Romano C, Gécz J,](#)
5 [DeLisi LE, Sebat J, King MC, Shaffer LG, Eichler EE. 2010. A recurrent 16p12.1 microdeletion](#)
6 [supports a two-hit model for severe developmental delay. Nat Genet. 2010 42:203-209.](#)
7
8

9
10
11 [Glessner JT, Wang K, Cai G, Korvatska O, Kim CE, Wood S, Zhang H, Estes A, Brune CW,](#)
12 [Bradfield JP, Imielinski M, Frackelton EC, Reichert J, Crawford EL, Munson J, Sleiman PM,](#)
13 [Chiavacci R, Annaiah K, Thomas K, Hou C, Glaberson W, Flory J, Otieno F, Garris M, Soorya](#)
14 [L, Klei L, Piven J, Meyer KJ, Anagnostou E, Sakurai T, Game RM, Rudd DS, Zurawiecki D,](#)
15 [McDougle CJ, Davis LK, Miller J, Posey DJ, Michaels S, Kolevzon A, Silverman JM, Bernier](#)
16 [R, Levy SE, Schultz RT, Dawson G, Owley T, McMahon WM, Wassink TH, Sweeney JA,](#)
17 [Nurnberger JI, Coon H, Sutcliffe JS, Minshew NJ, Grant SF, Bucan M, Cook EH, Buxbaum JD,](#)
18 [Devlin B, Schellenberg GD, Hakonarson H. 2009. Autism genome-wide copy number variation](#)
19 [reveals ubiquitin and neuronal genes. Nature 459:569-573.](#)
20
21

Formatted: Line spacing: Double

22
23
24 [Glinskii AB, Ma J, Ma S, Grant D, Lim CU, Sell S, Glinsky GV. 2009. Identification of](#)
25 [intergenic trans-regulatory RNAs containing a disease-linked SNP sequence and targeting cell](#)
26 [cycle progression/differentiation pathways in multiple common human disorders. Cell Cycle](#)
27 [8:3925-3942.](#)
28
29

Formatted: Right: 0 pt, Line spacing: Double

30
31
32
33
34
35
36
37
38
39
40
41
42 Goh KI, Cusick ME, Valle D, Childs B, Vidal M, Barabási AL. 2007. The human disease
43
44 network. Proc Natl Acad Sci USA 104:8685-8690.
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2 [Goode DL, Cooper GM, Schmutz J, Dickson M, Gonzales E, Tsai M, Karra K, Davydov E,](#)

3
4 [Batzoglou S, Myers RM, Sidow A. 2010. Evolutionary constraint facilitates interpretation of](#)
5
6 [genetic variation in resequenced human genomes. Genome Res 20:301-310.](#)

7
8
9
10 Gordon CT, Tan TY, Benko S, Fitzpatrick D, Lyonnet S, Farlie PG. 2009. Long-range regulation
11
12 at the *SOX9* locus in development and disease. J Med Genet 46:649-656.

13
14
15
16 [Gorlov IP, Kimmel M, Amos CI. 2006. Strength of the purifying selection against different](#)
17
18 [categories of the point mutations in the coding regions of the human genome. Hum Mol Genet](#)
19
20 [15:1143-1150.](#)

21
22
23
24 [Gorlov IP, Gorlova OY, Amos CI. 2008. Relative effects of mutability and selection on single](#)
25
26 [nucleotide polymorphisms in transcribed regions of the human genome. BMC Genomics 9:292.](#)

27
28
29
30 Grant SF, Reid DM, Blake G, Herd R, Fogelman I, Ralston SH. 1996. Reduced bone density and
31
32 osteoporosis associated with a polymorphic Sp1 binding site in the collagen type I alpha 1 gene.
33
34 Nat Genet 14:203-205.

35
36
37
38 Griffiths-Jones S. 2007. Annotating noncoding RNA genes. Annu Rev Genomics Hum Genet
39
40 8:279-298.

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

[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, Yamamoto 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.

1
2 Harrison PM, Zheng D, Zhang Z, Carriero N, Gerstein M. 2005. Transcribed processed
3 pseudogenes in the human genome: an intermediate form of expressed retrosequence lacking
4 protein-coding ability. *Nucleic Acids Res* 33:2374-2383.
5
6
7

8
9
10 [Harteveld CL, Voskamp A, Phylipsen M, Akkermans N, den Dunnen JT, White SJ, Giordano](#)
11 [PC. 2005. Nine unknown rearrangements in 16p13.3 and 11p15.4 causing \$\alpha\$ - and \$\beta\$ -thalassaemia](#)
12 [characterised by high resolution multiplex ligation-dependent probe amplification. *J Med Genet*](#)
13 [42:922-931.](#)
14
15 Deleted:

16
17
18
19
20 [Hatton CS, Wilkie AO, Drysdale HC, Wood WG, Vickers MA, Sharpe J, Ayyub H, Pretorius](#)
21 [IM, Buckle VJ, Higgs DR. 1990. \$\alpha\$ -thalassemia caused by a large \(62 kb\) deletion upstream of](#)
22 [the human \$\alpha\$ -globin gene cluster. *Blood* 76:221-227.](#)
23
24
25

26
27
28 He Y, Vogelstein B, Velculescu VE, Papadopoulos N, Kinzler KW. 2008. The antisense
29 transcriptomes of human cells. *Science* 322:1855-1857.
30 Deleted:

31
32
33
34 [Heintzman ND, Ren B. 2009. Finding distal regulatory elements in the human genome. *Curr*](#)
35 [Opin Genet Dev](#) 19:541-549.
36
37

38
39
40 Herzog H, Darby K, Hort YJ, Shine J. 1996. Intron 17 of the human retinoblastoma susceptibility
41 gene encodes an actively transcribed G protein-coupled receptor gene. *Genome Res* 6:858-861.
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2 | Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA. 2009.
3
4 | Potential etiologic and functional implications of genome-wide association loci for human
5
6 | diseases and traits. Proc Natl Acad Sci USA 106:9362-9367.
7

8
9
10 | Hirotsune S, Yoshida N, Chen A, Garrett L, Sugiyama F, Takahashi S, Yagami K, Wynshaw-
11
12 | Boris A, Yoshiki A. 2003. An expressed pseudogene regulates the messenger-RNA stability of
13
14 | its homologous coding gene. Nature 423:91-96.
15

16
17
18 | Hitchins MP, Ward RL. 2009. Constitutional (germline) *MLH1* epimutation as an aetiological
19
20 | mechanism for hereditary non-polyposis colorectal cancer. J Med Genet 46:793-802.
21

22
23
24 | [Homolova K, Zavadakova P, Doktor TK, Schroeder LD, Kozich V, Andresen BS. 2010. The](#)
25
26 | [deep intronic c.903+469T>C mutation in the *MTRR* gene creates an SF2/ASF binding exonic](#)
27
28 | [splicing enhancer, which leads to pseudoexon activation and causes the cbIE type of](#)
29
30 | [homocystinuria. Hum Mutat In press. \[doi>10.1002/humu.21206, PMID: 20120036\]](#)
31

32
33
34 | Howard OM, Turpin JA, Goldman R, Modi WS. 2004. Functional redundancy of the human
35
36 | *CCLA* and *CCL4L1* chemokine genes. Biochem Biophys Res Commun 320:927-931.
37

38
39
40 | Hsiao T-L, Vitkup D. 2008. Role of duplicate genes in robustness against deleterious human
41
42 | mutations. PLoS Genet 4:e1000014.
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2 Hu Z, Chen J, Tian T, Zhou X, Gu H, Xu L, Zeng Y, Miao R, Jin G, Ma H, Chen Y, Shen H.
3
4 2008. Genetic variants of miRNA sequences and non-small cell lung cancer survival. *J Clin*
5
6 *Invest* 118:2600-2608.
7

8
9
10 Hu Z, Liang J, Wang Z, Tian T, Zhou X, Chen J, Miao R, Wang Y, Wang X, Shen H. 2009.
11
12 Common genetic variants in pre-microRNAs were associated with increased risk of breast cancer
13
14 in Chinese women. *Hum Mutat* 30:79-84.
15

16
17
18 Huang H, Winter EE, Wang H, Weinstock KG, Xing H, Goodstadt L, Stenson PD, Cooper DN,
19
20 Smith D, Albà MM, Ponting CP, Fechtel K. 2004. Evolutionary conservation and selection of
21
22 human disease gene orthologs in the rat and mouse genomes. *Genome Biol* 5:R47.
23

24
25
26 Hunt R, Sauna ZE, Ambudkar SV, Gottesman MM, Kimchi-Sarfaty C. 2009. Silent
27
28 (synonymous) SNPs: should we care about them? *Methods Mol Biol* 578:23-39.
29

30
31
32 [Iafate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y, Scherer SW, Lee C. 2004.](#)
33
34 [Detection of large-scale variation in the human genome. *Nat Genet* 36:949-951.](#)
35

36
37
38 International Human Genome Sequencing Consortium. 2004. Finishing the euchromatic
39
40 sequence of the human genome. *Nature* 431:931-945.
41

42
43
44 [International Schizophrenia Consortium. 2008. Rare chromosomal deletions and duplications](#)
45
46 [increase risk of schizophrenia. *Nature* 455:237-241.](#)
47

1
2
3
4 Jarinova O, Stewart AF, Roberts R, Wells G, Lau P, Naing T, Buerki C, McLean BW, Cook RC,
5
6 Parker JS, McPherson R. 2009. Functional analysis of the chromosome 9p21.3 coronary artery
7
8 disease risk locus. *Arterioscler Thromb Vasc Biol* 29:1671-1677.
9

10
11 Jazdzewski K, Murray EL, Franssila K, Jarzab B, Schoenberg DR, de la Chapelle A. 2008.
12
13 Common SNP in pre-miR-146a decreases mature miR expression and predisposes to papillary
14
15 thyroid carcinoma. *Proc Natl Acad Sci USA* 105:7269-7274.
16

17
18
19
20 Jimenez-Sanchez G, Childs B, Valle D. 2001. Human disease genes. *Nature* 409:853-855.
21

22
23
24 Jones S, Hruban RH, Kamiyama M, Borges M, Zhang X, Parsons DW, Lin JC, Palmisano E,
25
26 Brune K, Jaffee EM, Iacobuzio-Donahue CA, Maitra A, Parmigiani G, Kern SE, Velculescu VE,
27
28 Kinzler KW, Vogelstein B, Eshleman JR, Goggins M, Klein AP. 2009. Exomic sequencing
29
30 identifies *PALB2* as a pancreatic cancer susceptibility gene. *Science* 324:217.
31

32
33
34 Jordan IK, Rogozin IB, Glazko GV, Koonin EV. 2003. Origin of a substantial fraction of human
35
36 regulatory sequences from transposable elements. *Trends Genet* 19:68-72.

Deleted: ¶

37
38
39
40 [Kapranov P, Cawley SE, Drenkow J, Bekiranov S, Strausberg RL, Fodor SP, Gingeras TR. 2002.](#)
41
42 [Large-scale transcriptional activity in chromosomes 21 and 22. *Science* 296:916-919.](#)
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2 | Kapranov P, Drenkow J, Cheng J, Long J, Helt G, Dike S, Gingeras TR. 2005. Examples of the
3 |
4 | complex architecture of the human transcriptome revealed by RACE and high-density tiling
5 |
6 | arrays. *Genome Res* 15:987-997.
7 |

8 |
9 |
10 | Kapranov P, Willingham AT, Gingeras TR. 2007a. Genome-wide transcription and the
11 |
12 | implications for genomic organization. *Nat Rev Genet* 8:413-423.
13 |

14 |
15 |
16 | Kapranov P, Cheng J, Dike S, Nix DA, Duttagupta R, Willingham AT, Stadler PF, Hertel J,
17 |
18 | Hackermüller J, Hofacker IL, Bell I, Cheung E, Drenkow J, Dumais E, Patel S, Helt G, Ganesh
19 |
20 | M, Ghosh S, Piccolboni A, Sementchenko V, Tammana H, Gingeras TR. 2007b. RNA maps
21 |
22 | reveal new RNA classes and a possible function for pervasive transcription. *Science* 316:1484-
23 |
24 | 1488.
25 |

26 |
27 |
28 | [Karchin R. 2009. Next generation tools for the annotation of human SNPs. *Brief Bioinform*](#)
29 |
30 | [10:35-52.](#)
31 |

32 |
33 |
34 | Katayama S, Tomaru Y, Kasukawa T, Waki K, Nakanishi M, Nakamura M, Nishida H, Yap CC,
35 |
36 | Suzuki M, Kawai J, Suzuki H, Carninci P, Hayashizaki Y, Wells C, Frith M, Ravasi T, Pang KC,
37 |
38 | Hallinan J, Mattick J, Hume DA, Lipovich L, Batalov S, Engström PG, Mizuno Y, Faghihi MA,
39 |
40 | Sandelin A, Chalk AM, Mottagui-Tabar S, Liang Z, Lenhard B, Wahlestedt C; RIKEN Genome
41 |
42 | Exploration Research Group; Genome Science Group (Genome Network Project Core Group);
43 |
44 | FANTOM Consortium. 2005. Antisense transcription in the mammalian transcriptome. *Science*
45 |
46 | 309:1564-1566.
47 |

1
2
3
4 Katzman S, Kern AD, Bejerano G, Fewell G, Fulton L, Wilson RK, Salama SR, Haussler D.
5
6 2007. Human genome ultraconserved elements are ultraselected. *Science* 317:915.
7

8
9
10 Kawaji H, Hayashizaki Y. 2008. Exploration of small RNAs. *PLoS Genet* 4:e22.
11

12
13
14 [Keightley PD, Kryukov GV, Sunyaev S, Halligan DL, Gaffney DJ. 2005. Evolutionary](#)
15 [constraints in conserved nongenic sequences of mammals. *Genome Res* 15:1373-1378.](#)
16
17

18
19
20 Khachane AN, Harrison PM. 2009. Assessing the genomic evidence for conserved transcribed
21
22 pseudogenes under selection. *BMC Genomics* 10:435.
23

24
25
26 Khalil AM, Guttman M, Huarte M, Garber M, Raj A, Rivea Morales D, Thomas K, Presser A,
27
28 Bernstein BE, van Oudenaarden A, Regev A, Lander ES, Rinn JL. 2009. Many human large
29
30 intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene
31
32 expression. *Proc Natl Acad Sci USA* 106:11667-11672.
33

34
35
36 Kidd JM, Cooper GM, Donahue WF, Hayden HS, Sampas N, Graves T, Hansen N, Teague B,
37
38 Alkan C, Antonacci F, Haugen E, Zerr T, Yamada NA, Tsang P, Newman TL, Tüzün E, Cheng
39
40 Z, Ebling HM, Tusneem N, David R, Gillett W, Phelps KA, Weaver M, Saranga D, Brand A,
41
42 Tao W, Gustafson E, McKernan K, Chen L, Malig M, Smith JD, Korn JM, McCarroll SA,
43
44 Altshuler DA, Peiffer DA, Dorschner M, Stamatoyannopoulos J, Schwartz D, Nickerson DA,
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2 Mullikin JC, Wilson RK, Bruhn L, Olson MV, Kaul R, Smith DR, Eichler EE. 2008. Mapping
3 and sequencing of structural variation from eight human genomes. *Nature* 453:56-64.
4
5

6
7
8 Kim JI, Ju YS, Park H, Kim S, Lee S, Yi JH, Mudge J, Miller NA, Hong D, Bell CJ, Kim HS,
9
10 Chung IS, Lee WC, Lee JS, Seo SH, Yun JY, Woo HN, Lee H, Suh D, Lee S, Kim HJ,
11
12 Yavartanoo M, Kwak M, Zheng Y, Lee MK, Park H, Kim JY, Gokcumen O, Mills RE, Zaranek
13
14 AW, Thakuria J, Wu X, Kim RW, Huntley JJ, Luo S, Schroth GP, Wu TD, Kim H, Yang KS,
15
16 Park WY, Kim H, Church GM, Lee C, Kingsmore SF, Seo JS. 2009. A highly annotated whole-
17
18 genome sequence of a Korean individual. *Nature* 460:1011-1015.
19
20

21
22 [Kimura M. 1985. The role of compensatory neutral mutations in molecular evolution.](#)

23
24 [J Genet 64:7-19.](#)
25
26
27

28 Kleinjan DA, Lettice LA. 2008. Long-range gene control and genetic disease. *Adv Genet*
29
30 61:339-388.
31

32
33
34 Kleinjan DJ, Coutinho P. 2009. Cis-rupture mechanisms: disruption of *cis*-regulatory control as a
35
36 cause of human genetic disease. *Brief Funct Genomic Proteomic* 8:317-332.
37
38

39
40 Klooster R, Straasheijm K, Shah B, Sowden J, Frants R, Thornton C, Tawil R, van der Maarel S.
41
42 2009. Comprehensive expression analysis of FSHD candidate genes at the mRNA and protein
43
44 level. *Eur J Hum Genet* 17:1615-1624.
45
46
47
48

1
2 [Koenig SC, Becirevic E, Hellberg MS, Li MY, So JC, Hankins JS, Ware RE, McMahon L,](#)
3
4 [Steinberg MH, Luo HY, Chui DH. 2009. Sickle cell disease caused by heterozygosity for Hb S](#)
5
6 [and novel LCR deletion: Report of two patients. Am J Hematol 84:603-606.](#)
7
8

9
10 Kondrashov AS. 1995. Contamination of the genome by very slightly deleterious mutations: why
11
12 have we not died 100 times over? J Theor Biol 175:583-594.
13

14
15
16 [Kondrashov AS, Sunyaev S, Kondrashov FA. 2002. Dobzhansky-Muller incompatibilities in](#)
17
18 [protein evolution. Proc Natl Acad Sci USA 99:14878-14883.](#)
19

20
21 Kondrashov AS. 2003. Direct estimates of human per nucleotide mutation rates at 20 loci
22
23 causing Mendelian diseases. Hum Mutat 21:12-27.
24

25
26
27
28 Krawczak M, Thomas NS, Hundrieser B, Mort M, Wittig M, Hampe J, Cooper DN. 2007. Single
29
30 base-pair substitutions in exon-intron junctions of human genes: nature, distribution, and
31
32 consequences for mRNA splicing. Hum Mutat 28:150-158.
33

34
35
36 [Kryukov GV, Schmidt S, Sunyaev S. 2005. Small fitness effect of mutations in highly conserved](#)
37
38 [non-coding regions. Hum Mol Genet 14:2221-2229.](#)
39

40
41
42 Kryukov GV, Pennacchio LA, Sunyaev SR. 2007. Most rare missense alleles are deleterious in
43
44 humans: implications for complex disease and association studies. Am J Hum Genet 80:727-739.
45

1
2 Kryukov GV, Shpunt A, Stamatoyannopoulos JA, Sunyaev SR. 2009. Power of deep, all-exon
3
4 resequencing for discovery of human trait genes. Proc Natl Acad Sci USA 106:3871-3876.
5

6
7
8 Kumar S, Suleski MP, Markov GJ, Lawrence S, Marco A, Filipinski AJ. 2009. Positional
9
10 conservation and amino acids shape the correct diagnosis and population frequencies of benign
11
12 and damaging personal amino acid mutations. Genome Res 19:1562-1569.
13

14
15
16 [Kwan T, Benovoy D, Dias C, Gurd S, Provencher C, Beaulieu P, Hudson TJ, Sladek R,](#)
17
18 [Majewski J. 2008. Genome-wide analysis of transcript isoform variation in humans. Nat Genet](#)
19
20 [40:225-231.](#)
21

22
23
24 Lage K, Hansen NT, Karlberg EO, Eklund AC, Roque FS, Donahoe PK, Szallasi Z, Jensen TS,
25
26 Brunak S. 2008. A large-scale analysis of tissue-specific pathology and gene expression of
27
28 human disease genes and complexes. Proc Natl Acad Sci USA 105:20870-20875.
29

30
31
32 [Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle](#)
33
34 [M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J,](#)
35
36 [LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J,](#)
37
38 [Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann N, Stojanovic N,](#)
39
40 [Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee](#)
41
42 [C, Carter N, Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L,](#)
43
44 [Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, McMurray A,](#)
45
46 [Matthews L, Mercer S, Milne S, Mullikin JC, Mungall A, Plumb R, Ross M, Shownkeen R,](#)
47

1
2 [Sims S, Waterston RH, Wilson RK, Hillier LW, McPherson JD, Marra MA, Mardis ER, Fulton](#)
3 [LA, Chinwalla AT, Pepin KH, Gish WR, Chissole SL, Wendl MC, Delehaunty KD, Miner TL,](#)
4 [Delehaunty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW, Hawkins T,](#)
5 [Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng JF, Olsen A,](#)
6 [Lucas S, Elkin C, Uberbacher E, Frazier M, Gibbs RA, Muzny DM, Scherer SE, Bouck JB,](#)
7 [Sodergren EJ, Worley KC, Rives CM, Gorrell JH, Metzker ML, Naylor SL, Kucherlapati RS,](#)
8 [Nelson DL, Weinstock GM, Sakaki Y, Fujiyama A, Hattori M, Yada T, Toyoda A, Itoh T,](#)
9 [Kawagoe C, Watanabe H, Totoki Y, Taylor T, Weissenbach J, Heilig R, Saurin W, Artiguenave](#)
10 [F, Brottier P, Bruls T, Pelletier E, Robert C, Wincker P, Smith DR, Doucette-Stamm L,](#)
11 [Rubenfield M, Weinstock K, Lee HM, Dubois J, Rosenthal A, Platzer M, Nyakatura G, Taudien](#)
12 [S, Rump A, Yang H, Yu J, Wang J, Huang G, Gu J, Hood L, Rowen L, Madan A, Qin S, Davis](#)
13 [RW, Federspiel NA, Abola AP, Proctor MJ, Myers RM, Schmutz J, Dickson M, Grimwood J,](#)
14 [Cox DR, Olson MV, Kaul R, Raymond C, Shimizu N, Kawasaki K, Minoshima S, Evans GA,](#)
15 [Athanasίου M, Schultz R, Roe BA, Chen F, Pan H, Ramser J, Lehrach H, Reinhardt R,](#)
16 [McCombie WR, de la Bastide M, Dedhia N, Blöcker H, Hornischer K, Nordsiek G, Agarwala R,](#)
17 [Aravind L, Bailey JA, Bateman A, Batzoglou S, Birney E, Bork P, Brown DG, Burge CB,](#)
18 [Cerutti L, Chen HC, Church D, Clamp M, Copley RR, Doerks T, Eddy SR, Eichler EE, Furey](#)
19 [TS, Galagan J, Gilbert JG, Harmon C, Hayashizaki Y, Haussler D, Hermjakob H, Hokamp K,](#)
20 [Jang W, Johnson LS, Jones TA, Kasif S, Kasprzyk A, Kennedy S, Kent WJ, Kitts P, Koonin EV,](#)
21 [Korf I, Kulp D, Lancet D, Lowe TM, McLysaght A, Mikkelsen T, Moran JV, Mulder N, Pollara](#)
22 [VJ, Ponting CP, Schuler G, Schultz J, Slater G, Smit AF, Stupka E, Szustakowski J, Thierry-](#)
23 [Mieg D, Thierry-Mieg J, Wagner L, Wallis J, Wheeler R, Williams A, Wolf YI, Wolfe KH,](#)
24 [Yang SP, Yeh RF, Collins F, Guyer MS, Peterson J, Felsenfeld A, Wetterstrand KA, Patrinos A,](#)
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2 [Morgan MJ, de Jong P, Catanese JJ, Osoegawa K, Shizuya H, Choi S, Chen YJ; International](#)
3 [Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human](#)
4 [genome. Nature 409:860-921.](#)
5
6
7

8
9
10 Lauderdale JD, Wilensky JS, Oliver ER, Walton DS, Glaser T. 2000. 3' deletions cause aniridia
11 by preventing *PAX6* gene expression. Proc Natl Acad Sci USA 97:13755-13759.
12
13

14
15
16 Le WD, Xu P, Jankovic J, Jiang H, Appel SH, Smith RG, Vassilatis DK. 2003. Mutations in
17 *NR4A2* associated with familial Parkinson disease. Nat Genet 33:85-89.
18
19

20
21
22 Lecoindre C, Pichon O, Hamel A, Heloury Y, Michel-Calemard L, Morel Y, David A, Le
23 Caignec C. 2009. Familial acampomelic form of campomelic dysplasia caused by a 960 kb
24 deletion upstream of *SOX9*. Am J Med Genet 149A:1183-1189.
25
26
27

28
29
30
31 [Lee C, Scherer SW. 2010. The clinical context of copy number variation in the human genome.](#)
32 [Expert Rev Mol Med 12:e8.](#)
33
34
35

36
37 Lettice LA, Horikoshi T, Heaney SJ, van Baren MJ, van der Linde HC, Breedveld GJ, Joesse M,
38 Akarsu N, Oostra BA, Endo N, Shibata M, Suzuki M, Takahashi E, Shinka T, Nakahori Y,
39 Ayusawa D, Nakabayashi K, Scherer SW, Heutink P, Hill RE, Noji S. 2002. Disruption of a
40 long-range *cis*-acting regulator for *Shh* causes preaxial polydactyly. Proc Natl Acad Sci USA
41 99:7548-7553.
42
43
44
45
46
47
48
49

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

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.

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 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 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.

1
2
3
4 [Li H, Xie H, Liu W, Hu R, Huang B, Tan YF, Xu K, Sheng ZF, Zhou HD, Wu XP, Luo XH.](#)
5 [2009d. A novel microRNA targeting *HDAC5* regulates osteoblast differentiation in mice and](#)
6 [contributes to primary osteoporosis in humans. *J Clin Invest* 119:3666-3677.](#)
7
8
9

10
11
12 Liang H, Li W-H. 2009. Functional compensation by duplicated genes in mouse. *Trends Genet*
13 25:441-442.
14
15

16
17
18 Liao BY, Zhang J. 2008. Null mutations in human and mouse orthologs frequently result in
19 different phenotypes. *Proc Natl Acad Sci USA* 105:6987-6992.
20
21
22

23
24 [Lin L, Jiang P, Shen S, Sato S, Davidson BL, Xing Y. 2009. Large-scale analysis of exonized](#)
25 [mammalian-wide interspersed repeats in primate genomes. *Hum Mol Genet* 18:2204-2214.](#)
26
27

Deleted: ¶

28
29
30 Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, Nery JR, Lee L, Ye Z,
31 Ngo QM, Edsall L, Antosiewicz-Bourget J, Stewart R, Ruotti V, Millar AH, Thomson JA, Ren
32 B, Ecker JR. 2009. Human DNA methylomes at base resolution show widespread epigenomic
33 differences. *Nature* 462:315-322.
34
35
36

37
38
39
40 [Liu JC, Makova KD, Adkins RM, Gibson S, Li WH. 2001. Episodic evolution of growth](#)
41 [hormone in primates and emergence of the species specificity of human growth hormone](#)
42 [receptor. *Mol Biol Evol* 18:945-953.](#)
43
44
45
46
47
48

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

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](#).

Deleted: In press.
doi10.1101/gr.094151.109

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.

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.

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

1
2 | Lowe CB, Bejerano G, Haussler D. 2007. Thousands of human mobile element fragments
3 |
4 | undergo strong purifying selection near developmental genes. Proc Natl Acad Sci USA
5 |
6 | 104:8005-8010.
7 |

8 |
9 |
10 | Lower KM, Hughes JR, De Gobbi M, Henderson S, Viprakasit V, Fisher C, Goriely A, Ayyub
11 |
12 | H, Sloane-Stanley J, Vernimmen D, Langford C, Garrick D, Gibbons RJ, Higgs DR. 2009.
13 |
14 | Adventitious changes in long-range gene expression caused by polymorphic structural variation
15 |
16 | and promoter competition. Proc Natl Acad Sci USA 106:21771-21776.
17 |

18 |
19 |
20 | [Lualdi S, Tappino B, Di Duca M, Dardis A, Anderson CJ, Biassoni R, Thompson PW, Corsolini](#)
21 | [F, Di Rocco M, Bembi B, Regis S, Cooper DN, Filocamo M. 2010. Enigmatic *in vivo* iduronate-](#)
22 | [2-sulfatase \(*IDS*\) mutant transcript correction to wild-type in hunter syndrome. Hum Mutat In](#)
23 | [press. \[doi>10.1002/humu.21208, PMID: 20104590\]](#)
24 |

25 |
26 |
27 | [Luco RF, Pan Q, Tominaga K, Blencowe BJ, Pereira-Smith OM, Misteli T. 2010. Regulation of](#)
28 | [alternative splicing by histone modifications. Science 327:996-1000.](#)
29 |

30 | [Lupski JR, Reid JG, Gonzaga-Jauregui C, Rio Deiros D, Chen DC, Nazareth L, Bainbridge M,](#)
31 | [Dinh H, Jing C, Wheeler DA, McGuire AL, Zhang F, Stankiewicz P, Halperin JJ, Yang C,](#)
32 | [Gehman C, Guo D, Irikat RK, Tom W, Fantin NJ, Muzny DM, Gibbs RA. 2010. Whole-genome](#)
33 | [sequencing in a patient with Charcot-Marie-Tooth neuropathy. New Engl J Med In press.](#)
34 | [\[doi>10.1056/NEJMoa0908094, PMID: 20220177\]](#)
35 |

Formatted: Line spacing: Double

Formatted: Right: 0 pt, Line spacing: Double

1
2 | McCarroll SA. 2008. Extending genome-wide association studies to copy-number variation.
3
4 | Hum Mol Genet 17(R2):R135-142.
5

6
7
8 | McKernan KJ, Peckham HE, Costa GL, McLaughlin SF, Fu Y, Tsung EF, Clouser CR, Duncan
9
10 | C, Ichikawa JK, Lee CC, Zhang Z, Ranade SS, Dimalanta ET, Hyland FC, Sokolsky TD, Zhang
11
12 | L, Sheridan A, Fu H, Hendrickson CL, Li B, Kotler L, Stuart JR, Malek JA, Manning JM,
13
14 | Antipova AA, Perez DS, Moore MP, Hayashibara KC, Lyons MR, Beaudoin RE, Coleman BE,
15
16 | Laptewicz MW, Sannicandro AE, Rhodes MD, Gottimukkala RK, Yang S, Bafna V, Bashir A,
17
18 | MacBride A, Alkan C, Kidd JM, Eichler EE, Reese MG, De La Vega FM, Blanchard AP. 2009.
19
20 | Sequence and structural variation in a human genome uncovered by short-read, massively
21
22 | parallel ligation sequencing using two-base encoding. Genome Res 19:1527-1541.
23

24
25
26 | Mann V, Hobson EE, Li B, Stewart TL, Grant SF, Robins SP, Aspden RM, Ralston SH. 2001. A
27
28 | *COL1A1* Sp1 binding site polymorphism predisposes to osteoporotic fracture by affecting bone
29
30 | density and quality. J Clin Invest 107:899-907.
31

32
33
34 | Maston GA, Evans SK, Green MR. 2006. Transcriptional regulatory elements in the human
35
36 | genome. Annu Rev Genomics Hum Genet 7:29-59.
37

38
39
40 | Mattick JS. 2009. The genetic signatures of noncoding RNAs. PloS Genet 5:e1000459
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Deleted:

1
2 [Medvedeva YA, Fridman MV, Oparina NJ, Malko DB, Ermakova EO, Kulakovskiy IV, Heinzl](#)

3
4 [A, Makeev VJ. Intergenic, gene terminal, and intragenic CpG islands in the human genome.](#)

5
6 [BMC Genomics 11:48.](#)

7
8
9
10 Mefford HC. 2009. Genotype to phenotype-discovery and characterization of novel genomic
11 disorders in a "genotype-first" era. *Genet Med* 11:836-842.

12
13
14
15
16 Mefford HC, Eichler EE. 2009. Duplication hotspots, rare genomic disorders, and common
17 disease. *Curr Opin Genet Dev* 19:196-204.

18
19
20
21 Mencía A, Modamio-Høybjør S, Redshaw N, Morín M, Mayo-Merino F, Olavarrieta L, Aguirre
22 LA, del Castillo I, Steel KP, Dalmay T, Moreno F, Moreno-Pelayo MA. 2009. Mutations in the
23 seed region of human miR-96 are responsible for nonsyndromic progressive hearing loss. *Nat*
24 *Genet* 41:609-613.

25
26
27
28
29
30
31
32 Mercer TR, Dinger ME, Mattick JS. 2009. Long non-coding RNAs: insights into functions. *Nat*
33 *Rev Genet* 10:155-159.

34
35
36
37 Merikangas AK, Corvin AP, Gallagher L. 2009. Copy-number variants in neurodevelopmental
38 disorders: promises and challenges. *Trends Genet* 25:536-544.

39
40
41
42
43
44 Miller MP, Kumar S. 2001. Understanding human disease mutations through the use of
45 interspecific genetic variation. *Hum Mol Genet* 10:2319-2328.

1
2
3
4 Miller MP, Parker JD, Rissing SW, Kumar S. 2003. Quantifying the intragenic distribution of
5 human disease mutations. *Ann Hum Genet* 67:567-579.
6
7

8
9
10 Mills RE, Bennett EA, Iskow RC, Devine SE. 2007. Which transposable elements are active in
11 the human genome? *Trends Genet* 23:183-191.
12
13

14
15
16 Morley M, Molony CM, Weber TM, Devlin JL, Ewens KG, Spielman RS, Cheung VG. 2004.
17 Genetic analysis of genome-wide variation in human gene expression. *Nature* 430:743-747.
18
19

20
21
22 Morris JA. 2001. How many deleterious mutations are there in the human genome? *Med*
23 *Hypotheses* 56:646-652.
24
25

26
27
28 Mort M, Evani US, Krishnan VG, Kamati KK, Baenziger PH, Bagchi A, Peters B, Sathyesh R,
29 Li B, Sun Y, Xue B, Shah N, Kann M, Cooper DN, Radivojac P, Mooney SD. 2010. *In silico*
30 functional profiling of human disease-associated and polymorphic amino acid substitutions. *Hum*
31 *Mutat* [31:335-346](#).
32
33
34
35

Deleted: .

Deleted: ion

Deleted: In press

36
37
38 Nachman MW, Crowell SL. 2000. Estimate of the mutation rate per nucleotide in humans.
39 *Genetics* 156:297-304.
40
41

42
43
44 Ng PC, Henikoff S. 2006. Predicting the effects of amino acid substitutions on protein function.
45 *Annu Rev Genomics Hum Genet* 7:61-80.
46
47

1
2
3
4 Ng PC, Levy S, Huang J, Stockwell TB, Walenz BP, Li K, Axelrod N, Busam DA, Strausberg
5
6 RL, Venter JC. 2008. Genetic variation in an individual human exome. PLoS Genet 4:e1000160.
7

8
9
10 Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, Lee C, Shaffer T, Wong M,
11
12 Bhattacharjee A, Eichler EE, Bamshad M, Nickerson DA, Shendure J. 2009. Targeted capture
13
14 and massively parallel sequencing of 12 human exomes. Nature 461:272-276.
15

16
17
18 Ng SB, Buckingham KJ, Lee C, Bigham AW, Tabor HK, Dent KM, Huff CD, Shannon PT, Jabs
19
20 EW, Nickerson DA, Shendure J, Bamshad MJ. 2010. Exome sequencing identifies the cause of a
21
22 mendelian disorder. Nat Genet 42:30-35.
23

24
25
26 Nishihara H, Smit AF, Okada N. 2006. Functional noncoding sequences derived from SINEs in
27
28 the mammalian genome. Genome Res. 16:864-874.
29

30
31
32 Nozu K, Iijima K, Nozu Y, Ikegami E, Imai T, Fu XJ, Kaito H, Nakanishi K, Yoshikawa N,
33
34 Matsuo M. 2009. A deep intronic mutation in the *SLC12A3* gene leads to Gitelman syndrome.
35
36 Pediatr Res 66:590-593.
37

38
39
40 Olds LC, Sibley E. 2003. Lactase persistence DNA variant enhances lactase promoter activity *in*
41
42 *vitro*: functional role as a *cis* regulatory element. Hum Mol Genet 12:2333-2340.
43
44
45
46
47
48

1
2 Orkin SH, Alter BP, Altay C, Mahoney MJ, Lazarus H, Hobbins JC, Nathan DG. 1978.

3
4 Application of endonuclease mapping to the analysis and prenatal diagnosis of thalassemias
5 caused by globin-gene deletion. *N Engl J Med* 299:166-172.
6
7

8
9
10 Osada N, Mano S, Gojobori J. 2009. Quantifying dominance and deleterious effect on human
11 disease genes. *Proc Natl Acad Sci USA* 106:841-846.
12
13

14
15
16 Ovcharenko I, Loots GG, Nobrega MA, Hardison RC, Miller W, Stubbs L. 2005. Evolution and
17 functional classification of vertebrate gene deserts. *Genome Res* 15:137-145.
18
19

20
21
22 Ozsolak F, Platt AR, Jones DR, Reifengerger JG, Sass LE, McInerney P, Thompson JF, Bowers
23 J, Jarosz M, Milos PM. 2009. Direct RNA sequencing. *Nature* 461:814-818.
24
25

26
27
28 Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ. 2008. Deep surveying of alternative splicing
29 complexity in the human transcriptome by high-throughput sequencing. *Nat Genet* 40:1413-
30 1415.
31
32

33
34
35
36 Park D, Park J, Park SG, Park T, Choi SS. 2008. Analysis of human disease genes in the context
37 of gene essentiality. *Genomics* 92:414-418.
38
39

40
41
42 Park PJ. 2009. ChIP-seq: advantages and challenges of a maturing technology. *Nat Rev Genet*
43 10:669-680.
44
45

1
2 Parker SC, Hansen L, Abaan HO, Tullius TD, Margulies EH. 2009. Local DNA topography
3
4 correlates with functional noncoding regions of the human genome. *Science* 324:389-392.
5

6
7
8 Parra G, Reymond A, Dabbouseh N, Dermitzakis ET, Castelo R, Thomson TM, Antonarakis SE,
9
10 Guigó R. 2006. Tandem chimerism as a means to increase protein complexity in the human
11
12 genome. *Genome Res* 16:37-44.
13

14
15
16 Pastinen T, Ge B, Hudson TJ. 2006. Influence of human genome polymorphism on gene
17
18 expression. *Hum Mol Genet* 15 Spec No 1:R9-16.
19

20
21
22 Pennacchio LA, Ahituv N, Moses AM, Prabhakar S, Nobrega MA, Shoukry M, Minovitsky S,
23
24 Dubchak I, Holt A, Lewis KD, Plajzer-Frick I, Akiyama J, De Val S, Afzal V, Black BL,
25
26 Couronne O, Eisen MB, Visel A, Rubin EM. 2006. *In vivo* enhancer analysis of human
27
28 conserved non-coding sequences. *Nature* 444:499-502.
29

30
31
32 Perotti D, De Vecchi G, Testi MA, Lualdi E, Modena P, Mondini P, Ravagnani F, Collini P, Di
33
34 Renzo F, Spreafico F, Terenziani M, Sozzi G, Fossati-Bellani F, Radice P. 2004. Germline
35
36 mutations of the *POU6F2* gene in Wilms tumors with loss of heterozygosity on chromosome
37
38 7p14. *Hum Mutat* 24:400-407.
39

40
41
42 Pesole G. 2008. What is a gene? An updated operational definition. *Gene* 417:1-4.
43
44
45
46
47
48
49

1
2 Peters BA, St Croix B, Sjöblom T, Cummins JM, Silliman N, Ptak J, Saha S, Kinzler KW,
3
4 Hatzis C, Velculescu VE. 2007. Large-scale identification of novel transcripts in the human
5
6 genome. *Genome Res* 17:287-292.
7

8
9
10 Pheasant M, Mattick JS. 2007. Raising the estimate of functional human sequences. *Genome Res*
11
12 17:1245-1253.
13

14
15
16 Piriyaopongsa J, Mariño-Ramírez L, Jordan IK. 2007. Origin and evolution of human microRNAs
17
18 from transposable elements. *Genetics* 176:1323-1337.
19

20
21
22 Pomerantz MM, Ahmadiyeh N, Jia L, Herman P, Verzi MP, Doddapaneni H, Beckwith CA,
23
24 Chan JA, Hills A, Davis M, Yao K, Kehoe SM, Lenz HJ, Haiman CA, Yan C, Henderson BE,
25
26 Frenkel B, Barretina J, Bass A, Taberero J, Baselga J, Regan MM, Manak JR, Shivdasani R,
27
28 Coetzee GA, Freedman ML. 2009. The 8q24 cancer risk variant rs6983267 shows long-range
29
30 interaction with *MYC* in colorectal cancer. *Nat Genet* 41:882-884.
31

32
33
34 Ponjavic J, Ponting CP, Lunter G. 2007. Functionality or transcriptional noise? Evidence for
35
36 selection within long noncoding RNAs. *Genome Res* 17:556-565.
37

38
39
40 Ponting CP, Lunter G. 2006. Signatures of adaptive evolution within human non-coding
41
42 sequence. *Hum Mol Genet* 15:R170-R175.
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2 [Poon A, Davis BH, Chao L. 2005. The coupon collector and the suppressor mutation: estimating](#)
3 [the number of compensatory mutations by maximum likelihood. Genetics 170:1323-1332.](#)
4
5
6
7

8 Prabhakar S, Noonan JP, Pääbo S, Rubin EM. 2006. Accelerated evolution of conserved
9 noncoding sequences in humans. Science 314:786.
10
11

12
13
14 Pros E, Gómez C, Martín T, Fábregas P, Serra E, Lázaro C. 2008. Nature and mRNA effect of
15 282 different *NFI* point mutations: focus on splicing alterations. Hum Mutat 29:E173-E193.
16
17

18
19
20 Prabhu S, Pe'er I. 2009. Overlapping pools for high-throughput targeted resequencing. Genome
21 Res 19:1254-1261.
22
23

24
25
26 Preker P, Nielsen J, Kammler S, Lykke-Andersen S, Christensen MS, Mapendano CK, Schierup
27 MH, Jensen TH. 2008. RNA exosome depletion reveals transcription upstream of active human
28 promoters. Science 322:1851-1854.
29
30
31

32
33
34 Pros E, Gómez C, Martín T, Fábregas P, Serra E, Lázaro C. 2008. Nature and mRNA effect of
35 282 different *NFI* point mutations: focus on splicing alterations. Hum Mutat 29:E173-E193.
36
37

38
39
40 Prabhu S, Pe'er I. 2009. Overlapping pools for high-throughput targeted resequencing. Genome
41 Res 19:1254-1261.
42
43
44
45
46
47
48
49

[Pruitt KD, Harrow J, Harte RA, Wallin C, Diekhans M, Maglott DR, Searle S, Farrell CM, Loveland JE, Ruff 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, Madaik 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.](#)

Formatted: Line spacing: Double

[Quemener S, Chen JM, Chuzhanova N, Bénech C, Casals T, Macek M Jr, 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* In press. \[doi>10.1002/humu.21196, PMID: 20052766\]](#)

Formatted: Right: 0 pt, Line spacing: Double

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-2 α binding site in an *IRF6* enhancer is associated with cleft lip. *Nat Genet* 40:1341-1347.

Deleted:

[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,](#)

1
2 [Kamp HM, Wilson AS, Gledhill A, Tridico S, Bunce M, Lorenzen ED, Binladen J, Guo X, Zhao](#)
3 [J, Zhang X, Zhang H, Li Z, Chen M, Orlando L, Kristiansen K, Bak M, Tommerup N, Bendixen](#)
4 [C, Pierre TL, Grønnow B, Meldgaard M, Andreassen C, Fedorova SA, Osipova LP, Higham TF,](#)
5 [Ramsey CB, Hansen TV, Nielsen FC, Crawford MH, Brunak S, Sicheritz-Pontén T, Villems R,](#)
6 [Nielsen R, Krogh A, Wang J, Willerslev E. 2010. Ancient human genome sequence of an extinct](#)
7 [Palaeo-Eskimo. Nature 463:757-762.](#)

8
9
10
11
12
13
14
15
16 [Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, Fiegler H, Shapero MH,](#)
17 [Carson AR, Chen W, Cho EK, Dallaire S, Freeman JL, González JR, Gratacòs M, Huang J,](#)
18 [Kalaitzopoulos D, Komura D, MacDonald JR, Marshall CR, Mei R, Montgomery L, Nishimura](#)
19 [K, Okamura K, Shen F, Somerville MJ, Tchinda J, Valsesia A, Woodwark C, Yang F, Zhang J,](#)
20 [Zerjal T, Zhang J, Armengol L, Conrad DF, Estivill X, Tyler-Smith C, Carter NP, Aburatani H,](#)
21 [Lee C, Jones KW, Scherer SW, Hurles ME. 2006. Global variation in copy number in the human](#)
22 [genome. Nature 444:444-454.](#)

23
24
25
26
27
28
29
30
31
32 Reich DE, Lander ES. 2001. On the allelic spectrum of human disease. Trends Genet 17:502-
33 510.

34
35
36
37
38 [Roach JC, Glusman G, Smit AF, Huff CD, Hubley R, Shannon PT, Rowen L, Pant KP,](#)
39 [Goodman N, Bamshad M, Shendure J, Drmanac R, Jorde LB, Hood L, Galas DJ. 2010. Analysis](#)
40 [of genetic inheritance in a family quartet by whole-genome sequencing. Science In press.](#)
41
42 [\[doi>10.1126/science.1186802, PMID: 20220176\]](#)
43
44
45
46
47

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

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.

1
2 | Sakai H, Koyanagi KO, Imanishi T, Itoh T, Gojobori T. 2007. Frequent emergence and
3 functional resurrection of processed pseudogenes in the human and mouse genomes. *Gene*
4 389:196-203.
5
6
7

8
9
10 | Sanford JR, Wang X, Mort M, Vanduyn N, Cooper DN, Mooney SD, Edenberg HJ, Liu Y. 2009.
11 Splicing factor SFRS1 recognizes a functionally diverse landscape of RNA transcripts. *Genome*
12 Res 19:381-394.
13
14
15

16
17
18 | Schollen E, Keldermans L, Foulquier F, Briones P, Chabas A, Sánchez-Valverde F, Adamowicz
19 M, Pronicka E, Wevers R, Matthijs G. 2007. Characterization of two unusual truncating *PMM2*
20 mutations in two CDG-Ia patients. *Mol Genet Metab* 90:408-413. 

21
22
23 | Schork NJ, Murray SS, Frazer KA, Topol EJ. 2009. Common vs. rare allele hypotheses for
24 complex diseases. *Curr Opin Genet Dev* 19:212-219. 

25
26
27 | [Schuster SC, Miller W, Ratan A, Tomsho LP, Giardine B, Kasson LR, Harris RS, Petersen DC,](#)
28 [Zhao F, Qi J, Alkan C, Kidd JM, Sun Y, Drautz DI, Bouffard P, Muzny DM, Reid JG, Nazareth](#)
29 [LV, Wang Q, Burhans R, Riemer C, Wittekindt NE, Moorjani P, Tindall EA, Danko CG, Teo](#)
30 [WS, Buboltz AM, Zhang Z, Ma Q, Oosthuysen A, Steenkamp AW, Oostuisen H, Venter P,](#)
31 [Gajewski J, Zhang Y, Pugh BF, Makova KD, Nekrutenko A, Mardis ER, Patterson N, Pringle](#)
32 [TH, Chiaromonte F, Mullikin JC, Eichler EE, Hardison RC, Gibbs RA, Harkins TT, Hayes VM,](#)
33 [2010. Complete Khoisan and Bantu genomes from southern Africa. *Nature* 463:943-947.](#) 

1
2 [Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin P, Månér S, Massa H, Walker M,](#)
3 [Chi M, Navin N, Lucito R, Healy J, Hicks J, Ye K, Reiner A, Gilliam TC, Trask B, Patterson N,](#)
4 [Zetterberg A, Wigler M. 2004. Large-scale copy number polymorphism in the human genome.](#)
5 [Science 305:525-528.](#)

Deleted: ¶

6
7
8
9
10
11 [Sebat J, Lakshmi B, Malhotra D, Troge J, Lese-Martin C, Walsh T, Yamrom B, Yoon S,](#)
12 [Krasnitz A, Kendall J, Leotta A, Pai D, Zhang R, Lee YH, Hicks J, Spence SJ, Lee AT, Puura K,](#)
13 [Lehtimäki T, Ledbetter D, Gregersen PK, Bregman J, Sutcliffe JS, Jobanputra V, Chung W,](#)
14 [Warburton D, King MC, Skuse D, Geschwind DH, Gilliam TC, Ye K, Wigler M. 2007. Strong](#)
15 [association of *de novo* copy number mutations with autism. Science 316:445-449.](#)

Deleted: ¶
¶

16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

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.

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.](#)

1
2 [Shlien A, Tabori U, Marshall CR, Pienkowska M, Feuk L, Novokmet A, Nanda S, Druker H,](#)

3
4 [Scherer SW, Malkin D. 2008. Excessive genomic DNA copy number variation in the Li-](#)

5
6 [Fraumeni cancer predisposition syndrome. Proc Natl Acad Sci USA 105:11264-11269.](#)

7
8
9
10 Sleiman PM, Healy DG, Muqit MM, Yang YX, Van Der Brug M, Holton JL, Revesz T, Quinn

11
12 NP, Bhatia K, Diss JK, Lees AJ, Cookson MR, Latchman DS, Wood NW. 2009.

13
14 Characterisation of a novel *NR4A2* mutation in Parkinson's disease brain. *Neurosci Lett* 457:75-

15
16 79.

17
18
19
20 Smith NG, Eyre-Walker A. 2003. Human disease genes: patterns and predictions. *Gene* 318:169-

21
22 175.

23
24
25
26 [Smith AD, Xuan Z, Zhang MQ. 2008. Using quality scores and longer reads improves accuracy](#)

27
28 [of Solexa read mapping. BMC Bioinformatics 9:128.](#)

29
30
31
32 Snider L, Asawachaicharn A, Tyler AE, Geng LN, Petek LM, Maves L, Miller DG, Lemmers

33
34 RJ, Winokur ST, Tawil R, van der Maarel SM, Filippova GN, Tapscott SJ. 2009. RNA

35
36 transcripts, miRNA-sized fragments and proteins produced from D4Z4 units: new candidates for

37
38 the pathophysiology of facioscapulohumeral dystrophy. *Hum Mol Genet* 18:2414-2430.

39
40
41
42 [Snyder M, Du J, Gerstein M. 2010. Personal genome sequencing: current approaches and](#)

43
44 [challenges. Genes Dev 24:423-431.](#)

1
2 Solis AS, Shariat N, Patton JG. 2008. Splicing fidelity, enhancers, and disease. *Front Biosci*
3
4 13:1926-1942.
5

6
7
8 Spena S, Asselta R, Platé M, Castaman G, Duga S, Tenchini ML. 2007. Pseudo-exon activation
9
10 caused by a deep-intronic mutation in the fibrinogen gamma-chain gene as a novel mechanism
11
12 for congenital afibrinogenaemia. *Br J Haematol* 139:128-132.
13

14
15
16 Stankiewicz P, Sen P, Bhatt SS, Storer M, Xia Z, Bejjani BA, Ou Z, Wiszniewska J, Driscoll DJ,
17
18 Maisenbacher MK, Bolivar J, Bauer M, Zackai EH, McDonald-McGinn D, Nowaczyk MM,
19
20 Murray M, Hustead V, Mascotti K, Schultz R, Hallam L, McRae D, Nicholson AG, Newbury R,
21
22 Durham-O'Donnell J, Knight G, Kini U, Shaikh TH, Martin V, Tyreman M, Simonic I, Willatt L,
23
24 Paterson J, Mehta S, Rajan D, Fitzgerald T, Gribble S, Prigmore E, Patel A, Shaffer LG, Carter
25
26 NP, Cheung SW, Langston C, Shaw-Smith C. 2009. Genomic and genic deletions of the *FOX*
27
28 gene cluster on 16q24.1 and inactivating mutations of *FOXF1* cause alveolar capillary dysplasia
29
30 and other malformations. *Am J Hum Genet* 84:780-791.
31

32
33
34 Stankiewicz P, Lupski JR. 2010. Structural variation in the human genome and its role in
35
36 disease. *Annu Rev Med* 61:437-455.
37

38
39
40 [Stefansson H, Rujescu D, Cichon S, Pietiläinen OP, Ingason A, Steinberg S, Fossdal R,](#)
41
42 [Sigurdsson E, Sigmundsson T, Buizer-Voskamp JE, Hansen T, Jakobsen KD, Muglia P, Francks](#)
43
44 [C, Matthews PM, Gylfason A, Halldorsson BV, Gudbjartsson D, Thorgeirsson TE, Sigurdsson](#)
45
46 [A, Jonasdottir A, Jonasdottir A, Bjornsson A, Mattiasdottir S, Blondal T, Haraldsson M,](#)
47

1
2 [Magnusdottir BB, Giegling I, Möller HJ, Hartmann A, Shianna KV, Ge D, Need AC, Crombie](#)
3
4 [C, Fraser G, Walker N, Lonnqvist J, Suvisaari J, Tuulio-Henriksson A, Paunio T, Toulopoulou](#)
5
6 [T, Bramon E, Di Forti M, Murray R, Ruggeri M, Vassos E, Tosato S, Walshe M, Li T, Vasilescu](#)
7
8 [C, Mühleisen TW, Wang AG, Ullum H, Djurovic S, Melle I, Olesen J, Kiemeny LA, Franke B;](#)
9
10 [GROUP, Sabatti C, Freimer NB, Gulcher JR, Thorsteinsdottir U, Kong A, Andreassen OA,](#)
11
12 [Ophoff RA, Georgi A, Rietschel M, Werge T, Petursson H, Goldstein DB, Nöthen MM,](#)
13
14 [Peltonen L, Collier DA, St Clair D, Stefansson K. 2008. Large recurrent microdeletions](#)
15
16 [associated with schizophrenia. Nature 455:232-236.](#)
17
18
19
20 Stenson PD, Mort M, Ball EV, Howells K, Phillips AD, Thomas NS, Cooper DN. 2009. The
21
22 Human Gene Mutation Database: 2008 update. Genome Med 1:13.
23
24
25
26 Stranger BE, Forrest MS, Clark AG, Minichiello MJ, Deutsch S, Lyle R, Hunt S, Kahl B,
27
28 Antonarakis SE, Tavaré S, Deloukas P, Dermitzakis ET. 2005. Genome-wide associations of
29
30 gene expression variation in humans. PLoS Genet 1:e78.
31
32
33
34 Stranger BE, Nica AC, Forrest MS, Dimas A, Bird CP, Beazley C, Ingle CE, Dunning M, Flicek
35
36 P, Koller D, Montgomery S, Tavaré S, Deloukas P, Dermitzakis ET. 2007. Population genomics
37
38 of human gene expression. Nat Genet 39:1217-1224.
39
40
41
42 Su Z, Gu X. 2008. Predicting the proportion of essential genes in mouse duplicates based on
43
44 biased mouse knockout genes. J Mol Evol 67:705-709.
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 Subramanian S, Kumar S. 2006. Evolutionary anatomies of positions and types of disease-
4 associated and neutral amino acid mutations in the human genome. *BMC Genomics* 7:306.
5

6
7
8 Sultan M, Schulz MH, Richard H, Magen A, Klingenhoff A, Scherf M, Seifert M, Borodina T,
9 Soldatov A, Parkhomchuk D, Schmidt D, O'Keeffe S, Haas S, Vingron M, Lehrach H, Yaspo
10 ML. 2008. A global view of gene activity and alternative splicing by deep sequencing of the
11 human transcriptome. *Science* 321:956-960.
12
13

14
15
16
17
18 [Sun G, Yan J, Noltner K, Feng J, Li H, Sarkis DA, Sommer SS, Rossi JJ. 2009. SNPs in human
19 miRNA genes affect biogenesis and function. *RNA* 15:1640-1651.](#)
20
21

22
23
24 [Suriano G, Azevedo L, Novais M, Boscolo B, Seruca R, Amorim A, Ghibaudi EM. 2007. In
25 vitro demonstration of intra-locus compensation using the ornithine transcarbamylase protein as
26 model. *Hum Mol Genet* 16:2209-2214.](#)
27
28

29
30
31
32 Susa S, Daimon M, Sakabe J, Sato H, Oizumi T, Karasawa S, Wada K, Jimbu Y, Kameda W,
33 Emi M, Muramatsu M, Kato T. 2008. A functional polymorphism of the TNF- α gene that is
34 associated with type 2 DM. *Biochem Biophys Res Commun* 369:943-947.
35
36

37
38
39
40 Svensson O, Arvestad L, Lagergren J. 2006. Genome-wide survey for biologically functional
41 pseudogenes. *PLoS Comput. Biol* 2:e46.
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2 Taft RJ, Glazov EA, Cloonan N, Simons C, Stephen S, Faulkner GJ, Lassmann T, Forrest AR,
3
4 Grimmond SM, Schroder K, Irvine K, Arakawa T, Nakamura M, Kubosaki A, Hayashida K,
5
6 Kawazu C, Murata M, Nishiyori H, Fukuda S, Kawai J, Daub CO, Hume DA, Suzuki H, Orlando
7
8 V, Carninci P, Hayashizaki Y, Mattick JS. 2009. Tiny RNAs associated with transcription start
9
10 sites in animals. *Nat Genet* 41:572-578.

11
12
13
14 Taylor J. 2005. Clues to function in gene deserts. *Trends Biotechnol* 23:269-271.

15
16
17
18 [Terai G, Yoshizawa A, Okida H, Asai K, Mituyama T. 2010. Discovery of short pseudogenes](#)
19
20 [derived from messenger RNAs. *Nucleic Acids Res* 38:1163-1171.](#)

21
22
23
24 Thean LF, Loi C, Ho KS, Koh PK, Eu KW, Cheah PY. 2010. Genome-wide scan identifies a
25
26 copy number variable region at 3q26 that regulates *PPM1L* in *APC* mutation-negative familial
27
28 colorectal cancer patients. *Genes Chrom Cancer* 49:99-106.

29
30
31
32 Theuns J, Del-Favero J, Dermaut B, van Duijn CM, Backhovens H, Van den Broeck MV,
33
34 Serneels S, Corsmit E, Van Broeckhoven CV, Cruts M. 2000. Genetic variability in the
35
36 regulatory region of presenilin 1 associated with risk for Alzheimer's disease and variable
37
38 expression. *Hum Mol Genet* 9:325-331.

39
40
41
42 Thornburg BG, Gotea V, Makalowski W. 2006. Transposable elements as a significant source of
43
44 transcription regulating signals. *Gene* 365:104-110.

1
2 | Tress ML, Martelli PL, Frankish A, Reeves GA, Wesselink JJ, Yeats C, Olason PI, Albrecht M,
3 |
4 | Hegyi H, Giorgetti A, Raimondo D, Lagarde J, Laskowski RA, López G, Sadowski MI, Watson
5 |
6 | JD, Fariselli P, Rossi I, Nagy A, Kai W, Størling Z, Orsini M, Assenov Y, Blankenburg H,
7 |
8 | Huthmacher C, Ramírez F, Schlicker A, Denoeud F, Jones P, Kerrien S, Orchard S, Antonarakis
9 |
10 | SE, Reymond A, Birney E, Brunak S, Casadio R, Guigo R, Harrow J, Hermjakob H, Jones DT,
11 |
12 | Lengauer T, Orengo CA, Patthy L, Thornton JM, Tramontano A, Valencia A. 2007. The
13 |
14 | implications of alternative splicing in the ENCODE protein complement. *Proc Natl Acad.Sci*
15 |
16 | USA 104:5495-5500.

17
18 |
19 |
20 | Tu Z, Wang L, Xu M, Zhou X, Chen T, Sun F. 2006. Further understanding human disease genes
21 |
22 | by comparing with housekeeping genes and other genes. *BMC Genomics* 7:31.

23
24 |
25 |
26 | Tucker T, Marra M, Friedman JM. 2009. Massively parallel sequencing: the next big thing in
27 |
28 | genetic medicine. *Am J Hum Genet* 85:142-154.

29
30 |
31 |
32 | Turgeon B, Meloche S. 2009. Interpreting neonatal lethal phenotypes in mouse mutants: insights
33 |
34 | into gene function and human diseases. *Physiol Rev* 89:1-26.

35
36 |
37 |
38 | Tuupanen S, Turunen M, Lehtonen R, Hallikas O, Vanharanta S, Kivioja T, Björklund M, Wei
39 |
40 | G, Yan J, Niittymäki I, Mecklin JP, Järvinen H, Ristimäki A, Di-Bernardo M, East P, Carvajal-
41 |
42 | Carmona L, Houlston RS, Tomlinson I, Palin K, Ukkonen E, Karhu A, Taipale J, Aaltonen LA.
43 |
44 | 2009. The common colorectal cancer predisposition SNP rs6983267 at chromosome 8q24
45 |
46 | confers potential to enhanced Wnt signaling. *Nat Genet* 41:885-890.

1
2
3
4 van Bokhoven H, Rawson RB, Merkx GF, Cremers FP, Seabra MC. 1996. cDNA cloning and
5
6 chromosomal localization of the genes encoding the alpha- and beta-subunits of human Rab
7
8 geranylgeranyl transferase: the 3' end of the alpha-subunit gene overlaps with the
9
10 transglutaminase 1 gene promoter. *Genomics* 38:133-140.

11
12
13
14 Velagaleti GV, Bien-Willner GA, Northup JK, Lockhart LH, Hawkins JC, Jalal SM, Withers M,
15
16 Lupski JR, Stankiewicz P. 2005. Position effects due to chromosome breakpoints that map
17
18 approximately 900 Kb upstream and approximately 1.3 Mb downstream of *SOX9* in two patients
19
20 with campomelic dysplasia. *Am J Hum Genet* 76:652-662.

21
22
23
24 Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans
25
26 CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, Zhang Q,
27
28 Kodira CD, Zheng XH, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang J, Gabor
29
30 Miklos GL, Nelson C, Broder S, Clark AG, Nadeau J, McKusick VA, Zinder N, Levine AJ,
31
32 Roberts RJ, Simon M, Slayman C, Hunkapiller M, Bolanos R, Delcher A, Dew I, Fasulo D,
33
34 Flanigan M, Florea L, Halpern A, Hannenhalli S, Kravitz S, Levy S, Mobarry C, Reinert K,
35
36 Remington K, Abu-Threideh J, Beasley E, Biddick K, Bonazzi V, Brandon R, Cargill M,
37
38 Chandramouliswaran I, Charlab R, Chaturvedi K, Deng Z, Di Francesco V, Dunn P, Eilbeck K,
39
40 Evangelista C, Gabrielian AE, Gan W, Ge W, Gong F, Gu Z, Guan P, Heiman TJ, Higgins ME,
41
42 Ji RR, Ke Z, Ketchum KA, Lai Z, Lei Y, Li Z, Li J, Liang Y, Lin X, Lu F, Merkulov GV,
43
44 Milshina N, Moore HM, Naik AK, Narayan VA, Neelam B, Nusskern D, Rusch DB, Salzberg S,
45
46 Shao W, Shue B, Sun J, Wang Z, Wang A, Wang X, Wang J, Wei M, Wides R, Xiao C, Yan C,
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2 Yao A, Ye J, Zhan M, Zhang W, Zhang H, Zhao Q, Zheng L, Zhong F, Zhong W, Zhu S, Zhao
3 S, Gilbert D, Baumhueter S, Spier G, Carter C, Cravchik A, Woodage T, Ali F, An H, Awe A,
4 Baldwin D, Baden H, Barnstead M, Barrow I, Beeson K, Busam D, Carver A, Center A, Cheng
5 ML, Curry L, Danaher S, Davenport L, Desilets R, Dietz S, Dodson K, Doup L, Ferriera S, Garg
6 N, Gluecksmann A, Hart B, Haynes J, Haynes C, Heiner C, Hladun S, Hostin D, Houck J,
7 Howland T, Ibegwam C, Johnson J, Kalush F, Kline L, Koduru S, Love A, Mann F, May D,
8 McCawley S, McIntosh T, McMullen I, Moy M, Moy L, Murphy B, Nelson K, Pfannkoch C,
9 Pratts E, Puri V, Qureshi H, Reardon M, Rodriguez R, Rogers YH, Romblad D, Ruhfel B, Scott
10 R, Sitter C, Smallwood M, Stewart E, Strong R, Suh E, Thomas R, Tint NN, Tse S, Vech C,
11 Wang G, Wetter J, Williams S, Williams M, Windsor S, Winn-Deen E, Wolfe K, Zaveri J,
12 Zaveri K, Abril JF, Guigó R, Campbell MJ, Sjolander KV, Karlak B, Kejariwal A, Mi H,
13 Lazareva B, Hatton T, Narechania A, Diemer K, Muruganujan A, Guo N, Sato S, Bafna V, Istrail
14 S, Lippert R, Schwartz R, Walenz B, Yooseph S, Allen D, Basu A, Baxendale J, Blick L,
15 Caminha M, Carnes-Stine J, Caulk P, Chiang YH, Coyne M, Dahlke C, Mays A, Dombroski M,
16 Donnelly M, Ely D, Esparham S, Fosler C, Gire H, Glanowski S, Glasser K, Glodek A,
17 Gorokhov M, Graham K, Gropman B, Harris M, Heil J, Henderson S, Hoover J, Jennings D,
18 Jordan C, Jordan J, Kasha J, Kagan L, Kraft C, Levitsky A, Lewis M, Liu X, Lopez J, Ma D,
19 Majoros W, McDaniel J, Murphy S, Newman M, Nguyen T, Nguyen N, Nodell M, Pan S, Peck
20 J, Peterson M, Rowe W, Sanders R, Scott J, Simpson M, Smith T, Sprague A, Stockwell T,
21 Turner R, Venter E, Wang M, Wen M, Wu D, Wu M, Xia A, Zandieh A, Zhu X. 2001. The
22 sequence of the human genome. *Science* 291:1304-1351.
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

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 *CDK5RI* 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.](#)

[Viprakit V, Kidd AM, Ayyub H, Horsley S, Hughes J, Higgs DR. 2003. De novo deletion within the telomeric region flanking the human \$\alpha\$ -globin locus as a cause of \$\alpha\$ -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.

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

1
2
3
4 [Walsh T, McClellan JM, McCarthy SE, Addington AM, Pierce SB, Cooper GM, Nord AS,](#)
5
6 [Kusenda M, Malhotra D, Bhandari A, Stray SM, Rippey CF, Roccanova P, Makarov V, Lakshmi](#)
7
8 [B, Findling RL, Sikich L, Stromberg T, Merriman B, Gogtay N, Butler P, Eckstrand K, Noory L,](#)
9
10 [Gochman P, Long R, Chen Z, Davis S, Baker C, Eichler EE, Meltzer PS, Nelson SF, Singleton](#)
11
12 [AB, Lee MK, Rapoport JL, King MC, Sebat J. 2008. Rare structural variants disrupt multiple](#)
13
14 [genes in neurodevelopmental pathways in schizophrenia. Science 320:539-543.](#)
15
16
17
18 [Walters RG, Jacquemont S, Valsesia A, de Smith AJ, Martinet D, Andersson J, Falchi M, Chen](#)
19
20 [F, Andrieux J, Lobbens S, Delobel B, Stutzmann F, El-Sayed Moustafa JS, Chèvre JC, Lecoeur](#)
21
22 [C, Vatin V, Bouquillon S, Buxton JL, Boute O, Holder-Espinasse M, Cuisset JM, Lemaitre MP,](#)
23
24 [Ambresin AE, Brioschi A, Gaillard M, Giusti V, Fellmann F, Ferrarini A, Hadjikhani N,](#)
25
26 [Campion D, Guilmatre A, Goldenberg A, Calmels N, Mandel JL, Le Caignec C, David A, Isidor](#)
27
28 [B, Cordier MP, Dupuis-Girod S, Labalme A, Sanlaville D, Béri-Dexheimer M, Jonveaux P,](#)
29
30 [Leheup B, Ounap K, Bochukova EG, Henning E, Keogh J, Ellis RJ, Macdermot KD, van Haelst](#)
31
32 [MM, Vincent-Delorme C, Plessis G, Touraine R, Philippe A, Malan V, Mathieu-Dramard M,](#)
33
34 [Chiesa J, Blaumeiser B, Kooy RF, Caiazzo R, Pigeyre M, Balkau B, Sladek R, Bergmann S,](#)
35
36 [Mooser V, Waterworth D, Reymond A, Vollenweider P, Waeber G, Kurg A, Palta P, Esko T,](#)
37
38 [Metspalu A, Nelis M, Elliott P, Hartikainen AL, McCarthy MI, Peltonen L, Carlsson L, Jacobson](#)
39
40 [P, Sjöström L, Huang N, Hurles ME, O'Rahilly S, Farooqi IS, Männik K, Jarvelin MR, Pattou F,](#)
41
42 [Meyre D, Walley AJ, Coin LJ, Blakemore AI, Froguel P, Beckmann JS. 2010. A new highly](#)
43
44 [penetrant form of obesity due to deletions on chromosome 16p11.2. Nature 463:671-675.](#)
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2 | Wang GS, Cooper TA. 2007. Splicing in disease: disruption of the splicing code and the
3 | decoding machinery. *Nat Rev Genet* 8:749-761.
4 |
5 |

6 |
7 |
8 | Wang ZQ, Tian SH, Shi YZ, Zhou PT, Wang ZY, Shu RZ, Hu L, Kong X. 2007. A single C to T
9 | transition in intron 5 of *LMBR1* gene is associated with triphalangeal thumb-polysyndactyly
10 | syndrome in a Chinese family. *Biochem Biophys Res Commun* 355:312-317.
11 |
12 |

13 |
14 |
15 |
16 | Wang J, Wang W, Li R, Li Y, Tian G, Goodman L, Fan W, Zhang J, Li J, Zhang J, Guo Y, Feng
17 | B, Li H, Lu Y, Fang X, Liang H, Du Z, Li D, Zhao Y, Hu Y, Yang Z, Zheng H, Hellmann I,
18 | Inouye M, Pool J, Yi X, Zhao J, Duan J, Zhou Y, Qin J, Ma L, Li G, Yang Z, Zhang G, Yang B,
19 | 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,
20 | Ye J, Fang L, Hao Q, Chen Q, Liang Y, Su Y, San A, Ping C, Yang S, Chen F, Li L, Zhou K,
21 | Zheng H, Ren Y, Yang L, Gao Y, Yang G, Li Z, Feng X, Kristiansen K, Wong GK, Nielsen R,
22 | Durbin R, Bolund L, Zhang X, Li S, Yang H, Wang J. 2008a. The diploid genome sequence of
23 | an Asian individual. *Nature* 456:60-65.
24 |
25 |

26 |
27 |
28 | Wang Z, Zang C, Rosenfeld JA, Schones DE, Barski A, Cuddapah S, Cui K, Roh TY, Peng W,
29 | Zhang MQ, Zhao K. 2008b. Combinatorial patterns of histone acetylations and methylations in
30 | the human genome. *Nat Genet* 40:897-903.
31 |
32 |

33 |
34 | Wang X, Wang K, Radovich M, Wang Y, Wang G, Feng W, Sanford JR, Liu Y. 2009a.
35 | Genome-wide prediction of *cis*-acting RNA elements regulating tissue-specific pre-mRNA
36 | alternative splicing. *BMC Genomics* 10 Suppl. 1:S4.
37 |
38 |
39 |
40 |
41 |
42 |
43 |
44 |
45 |
46 |
47 |
48 |
49 |
50 |
51 |
52 |
53 |
54 |
55 |
56 |
57 |
58 |
59 |
60 |

1
2
3
4 Wang Z, Gerstein M, Snyder M. 2009b. RNA-Seq: a revolutionary tool for transcriptomics. *Nat*
5
6 *Rev Genet* 10:57-63.
7

8
9
10 [Washietl S, Hofacker IL, Lukasser M, Hüttenhofer A, Stadler PF. 2005. Mapping of conserved](#)
11 [RNA secondary structures predicts thousands of functional noncoding RNAs in the human](#)
12 [genome. *Nat Biotechnol* 23:1383-1390.](#)
13
14

15
16
17
18 [Wen Y, Liu Y, Xu Y, Zhao Y, Hua R, Wang K, Sun M, Li Y, Yang S, Zhang XJ, Kruse R,](#)
19 [Cichon S, Betz RC, Nöthen MM, van Steensel MA, van Geel M, Steijlen PM, Hohl D, Huber M,](#)
20 [Dunnill GS, Kennedy C, Messenger A, Munro CS, Terrinoni A, Hovnanian A, Bodemer C, de](#)
21 [Prost Y, Paller AS, Irvine AD, Sinclair R, Green J, Shang D, Liu Q, Luo Y, Jiang L, Chen HD,](#)
22 [Lo WH, McLean WH, He CD, Zhang X. 2009. Loss-of-function mutations of an inhibitory](#)
23 [upstream ORF in the human hairless transcript cause Marie Unna hereditary hypotrichosis. *Nat*](#)
24 [Genet](#) 41:228-233.
25
26
27
28
29
30
31
32
33

34 Werner A, Carlile M, Swan D. 2009. What do natural antisense transcripts regulate? *RNA Biol*
35
36 6:43-48.
37

38
39
40 Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, McGuire A, He W, Chen YJ,
41
42 Makhijani V, Roth GT, Gomes X, Tartaro K, Niazi F, Turcotte CL, Irzyk GP, Lupski JR,
43
44 Chinault C, Song XZ, Liu Y, Yuan Y, Nazareth L, Qin X, Muzny DM, Margulies M, Weinstock
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2 GM, Gibbs RA, Rothberg JM. 2008. The complete genome of an individual by massively
3 parallel DNA sequencing. Nature 452:872-876.
4

5
6
7
8 [Wieczorek D, Pawlik B, Li Y, Akarsu NA, Caliebe A, May KJ, Schweiger B, Vargas FR, Balci](#)
9 [S, Gillessen-Kaesbach G, Wollnik B. 2010. A specific mutation in the distant sonic hedgehog](#)
10 [\(*SHH*\) *cis*-regulator \(*ZRS*\) causes Werner mesomelic syndrome \(*WMS*\) while complete *ZRS*](#)
11 [duplications underlie Haas type polysyndactyly and preaxial polydactyly \(*PPD*\) with or without](#)
12 [triphalangeal thumb. Hum Mutat 31:81-89.](#)
13
14
15
16
17

18
19
20 Wong E, Wei CL. 2009. ChIP'ing the mammalian genome: technical advances and insights into
21 functional elements. Genome Med 1:89.
22
23

24
25
26 [Wright JB, Brown SJ, Cole MD. 2010. Upregulation of *c-MYC* in *cis* through a large chromatin](#)
27 [loop linked to a cancer risk-associated single-nucleotide polymorphism in colorectal cancer cells.](#)
28 [Mol Cell Biol 30:1411-1420.](#)
29
30
31

32
33
34 [Wu M, Jolicoeur N, Li Z, Zhang L, Fortin Y, L'Abbe D, Yu Z, Shen SH. 2008. Genetic](#)
35 [variations of microRNAs in human cancer and their effects on the expression of miRNAs.](#)
36 [Carcinogenesis 29:1710-1716.](#)
37
38
39

40
41
42 Xing J, Zhang Y, Han K, Salem AH, Sen SK, Huff CD, Zhou Q, Kirkness EF, Levy S, Batzer
43 MA, Jorde LB. 2009. Mobile elements create structural variation: analysis of a complete human
44 genome. Genome Res 19:1516-1526.
45
46
47
48
49

1
2
3
4 Yagi M, Takeshima Y, Wada H, Nakamura H, Matsuo M. 2003. Two alternative exons can
5 result from activation of the cryptic splice acceptor site deep within intron 2 of the dystrophin
6 gene in a patient with as yet asymptomatic dystrophinopathy. *Hum Genet* 112:164-170.
7
8
9

10
11 Yamaguchi-Kabata Y, Shimada MK, Hayakawa Y, Minoshima S, Chakraborty R, Gojobori T,
12 Imanishi T. 2008. Distribution and effects of nonsense polymorphisms in human genes. *PLoS*
13 *One* 3:e3393.
14
15
16
17

18
19
20 Yang MQ, Einitski LL. 2008. Diversity of core promoter elements comprising human
21 bidirectional promoters. *BMC Genomics* 9 Suppl. 2:S3.
22
23
24

25
26 Yang R, Frank B, Hemminki K, Bartram CR, Wappenschmidt B, Sutter C, Kiechle M, Bugert P,
27 Schmutzler RK, Arnold N, Weber BH, Niederacher D, Meindl A, Burwinkel B. 2008a. SNPs in
28 ultraconserved elements and familial breast cancer risk. *Carcinogenesis* 29:351-355.
29
30
31

32
33
34 Yang TL, Chen XD, Guo Y, Lei SF, Wang JT, Zhou Q, Pan F, Chen Y, Zhang ZX, Dong SS, Xu
35 XH, Yan H, Liu X, Qiu C, Zhu XZ, Chen T, Li M, Zhang H, Zhang L, Drees BM, Hamilton JJ,
36 Papasian CJ, Recker RR, Song XP, Cheng J, Deng HW. 2008b. Genome-wide copy-number-
37 variation study identified a susceptibility gene, *UGT2B17*, for osteoporosis. *Am J Hum Genet*
38 83:663-674.
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2 | Yngvadottir B, Xue Y, Searle S, Hunt S, Delgado M, Morrison J, Whittaker P, Deloukas P,
3 | Tyler-Smith C. 2009. A genome-wide survey of the prevalence and evolutionary forces acting on
4 | human nonsense SNPs. *Am J Hum Genet* 84:224-234.
5
6
7

8
9 |
10 | Yngvadottir B, MacArthur DG, Jin H, Tyler-Smith C. 2009. The promise and reality of personal
11 | genomics. *Genome Biol* 10:237.
12
13

14
15 |
16 | Zhang C. 2008. MicroRNomics: a newly emerging approach for disease biology. *Physiol.*
17 | *Genomics* 33:139-147.
18
19

20
21 |
22 | Zhang ZD, Paccanaro A, Fu Y, Weissman S, Weng Z, Chang J, Snyder M, Gerstein MB. 2007.
23 | Statistical analysis of the genomic distribution and correlation of regulatory elements in the
24 | ENCODE regions. *Genome Res* 17:787-797.
25
26
27

28
29 |
30 | Zhang F, Carvalho CM, Lupski JR. 2009a. Complex human chromosomal and genomic
31 | rearrangements. *Trends Genet* 25:298-307.
32
33

34
35 |
36 | [Zhang D, Cheng L, Qian Y, Alliey-Rodriguez N, Kelsoe JR, Greenwood T, Nievergelt C, Barrett](#)
37 | [TB, McKinney R, Schork N, Smith EN, Bloss C, Nurnberger J, Edenberg HJ, Foroud T, Sheftner](#)
38 | [W, Lawson WB, Nwulia EA, Hipolito M, Coryell W, Rice J, Byerley W, McMahon F, Schulze](#)
39 | [TG, Berrettini W, Potash JB, Belmonte PL, Zandi PP, McInnis MG, Zöllner S, Craig D,](#)
40 | [Szelingier S, Koller D, Christian SL, Liu C, Gershon ES. 2009b. Singleton deletions throughout](#)
41 | [the genome increase risk of bipolar disorder. *Mol Psychiatry* 14:376-380.](#)
42
43
44
45
46
47

1
2
3
4 [Zhang ZD, Frankish A, Hunt T, Harrow J, Gerstein M. 2010. Identification and analysis of](#)
5 [unitary pseudogenes: historic and contemporary gene losses in humans and other primates.](#)
6
7 [Genome Biol 11:R26.](#)
8
9

10
11
12 Zheng D, Gerstein MB. 2007. The ambiguous boundary between genes and pseudogenes: the
13
14 dead rise up, or do they? Trends Genet 23:219-224.
15

16
17
18 Zheng D, Frankish A, Baertsch R, Kapranov P, Reymond A, Choo SW, Lu Y, Denoeud F,
19
20 Antonarakis SE, Snyder M, Ruan Y, Wei CL, Gingeras TR, Guigó R, Harrow J, Gerstein MB.
21
22 2007.) Pseudogenes in the ENCODE regions: consensus annotation, analysis of transcription,
23
24 and evolution. Genome Res 17:839-851.
25

26
27
28 Zhu QS, Xing W, Qian B, von Dippe P, Shneider BL, Fox VL, Levy D. 2003. Inhibition of
29
30 human m-epoxide hydrolase gene expression in a case of hypercholanemia. Biochim Biophys
31
32 Acta 1638:208-216.
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Figure legends:

Figure 1: The 'functionome': types of functional or potentially functional DNA sequences in the human genome that may harbour 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 non-coding sequences not associated with genes)

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.

- Deleted: Shown are t
- Deleted: s
- Deleted: listed
- Deleted: in
- Deleted: on an annual basis

Table 1 Selected examples of deep intronic mutations identified as causing human inherited disease*

Gene (MIM#)	Disease	Chromosomal location	Mutation	Consequences for mRNA splicing	Reference
<i>ATM</i> (MIM# 607585)	Ataxia telangiectasia	11q22-q23	IVS28-159A>G	Activates 5' cryptic splice site 83 nt downstream	Coutinho et al. (2005)
<i>CDKN2A</i> (MIM# 600160)	Melanoma, predisposition to	9p21	IVS2-105A>G	Activates cryptic splice site 105 nt 5' to exon 3	Harland et al. (2001)
<i>DMD</i> (MIM# 300377)	Dystrophinopathy, asymptomatic	Xp21.2	IVS2+5591T>A	Activated two 5' cryptic splice sites 132 nt or 46 nt downstream	Yagi et al. (2003)
<i>FGB</i> (MIM# 134830)	Afibrinogenemia	4q28	IVS1-600A>G	Creates consensus sequence for splicing factor SF2/ASF leading to inclusion of cryptic exon	Dear et al. (2006)
<i>FGG</i> (MIM# 134850)	Afibrinogenemia	4q28	IVS6-320A>T	Activates cryptic splice leading to inclusion of cryptic exon	Spena et al. (2007)
<u><i>MTRR</i></u> <u>(MIM# 602568)</u>	<u>Homocystinuria</u>	<u>5p15.31</u>	<u>IVS6+469T>C</u>	<u>Creates an SF2/ASF-binding exon splice enhancer which leads to pseudoexon activation</u>	<u>Homolova et al. (2010)</u>
<i>NF1</i> (MIM# 162200)	Neurofibromatosis type 1	17q11.2	IVS3+2025T>G	Activates cryptic splice leading to inclusion of cryptic exon	Pros et al. (2008)
<i>PMM2</i> (MIM# 601785)	Congenital disorder of glycosylation type Ia	16p13	IVS7-15479C>T	Activates cryptic splice site leading to the inclusion of pseudoexons	Schollen et al. (2007)
<i>PRPF31</i> (MIM# 606419)	Retinitis pigmentosa, autosomal dominant	19q13.4	IVS13+654C>G	Activates cryptic splice site leading to the inclusion of pseudoexons	Frio et al. (2009)
<i>SLC12A3</i> (MIM# 600968)	Gitelman syndrome	16q13	IVS13-191C>T	Activates cryptic splice leading to inclusion of cryptic exon	Nozu et al. (2009)

*located within an intron at least 100bp from the nearest splice site.

Table 2 Examples of regulatory mutations, located far upstream of gene sequences, known to cause human inherited disease

Gene (MIM#)	Disease	Chromosomal location	Mutation and relative location*	Reference
<i>SOX9</i> (MIM# 608160)	Cleft palate, Pierre Robin sequence	17q24.3-q25.1	-1441231T>C	Benko et al. (2009)
<i>SHH</i> (MIM# 600725)	Triphalangeal thumb-polysyndactyly syndrome	7q36.3	-979896C>T	Wang et al. (2007)
<i>SHH</i> (MIM# 600725)	Preaxial polydactyly	7q36.3	-979498C>G	Lettice et al. (2003)
<i>SHH</i> (MIM# 600725)	Triphalangeal thumb	7q36.3	-979308T>C	Furniss et al. (2008)
<i>SHH</i> (MIM# 600725)	Preaxial polydactyly	7q36.3	-979298A>T	Lettice et al. (2003)
<i>SHH</i> (MIM# 600725)	Preaxial polydactyly	7q36.3	-979275T>C	Lettice et al. (2003)
<i>SHH</i> (MIM# 600725)	Preaxial polydactyly	7q36.3	-979199G>A	Lettice et al. (2003)
SHH (MIM# 600725)	Werner mesomelic syndrome	7q36.3	-979199G>C	Wieczorek et al. (2010)
<i>SHH</i> (MIM# 600725)	Triphalangeal thumb-polysyndactyly syndrome	7q36.3	-978982C>G	Gurnett et al. (2007)
<i>SHH</i> (MIM# 600725)	Triphalangeal thumb-polysyndactyly syndrome	7q36.3	-978864A>G	Gurnett et al. (2007)
<i>POU6F2</i> (MIM# 609062)	Wilms' tumour	7p14.1	-28793C>G	Perotti et al. (2004)
<i>EPHX1</i> (MIM# 132810)	Hypercholanemia	1q42.12	-4328T>A ¹	Zhu et al. (2003)
<i>PSEN1</i> (MIM# 104311)	Alzheimer disease, early onset	14q24.2	-2818A>G	Theuns et al. (2000)
<i>NR4A2</i> (MIM# 601828)	Parkinson disease	2q24.1	-2478C>T	Sleiman et al. (2009)
<i>NR4A2</i> (MIM# 601828)	Parkinson disease	2q24.1	-2414T>G	Le et al. (2003)

*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.¹Mutation is located in a recognition site for hepatocyte nuclear factor 3 (HNF-3).

Deleted: ¶

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
<i>BMP2</i> (MIM# 112261)	Autosomal dominant brachydactyly type A2	20p12.3	Duplication (5.5 kb)	~110 kb 3' to gene	Dathe et al. (2009)
<i>DLX6</i> (MIM# 600030)	Hearing loss and craniofacial defects	7q21.3	Inversion breakpoint	~65 kb 5' to gene	Brown et al. (2010)
<i>FOXC2</i> (MIM# 602402)	Lymphedema-distichiasis syndrome	16q24.1	Translocation breakpoint	120 kb 3' to gene	Fang et al. (2000)
<i>FOXF1</i> (MIM# 601089)	Alveolar capillary dysplasia	16q24.1	Deletions (524 kb, 145 kb)	52 kb and 259 kb 5' to gene	Stankiewicz et al. (2009)
<i>FOXL2</i> (MIM# 605597)	Blepharophimosis syndrome	3q22.3	Deletion (7.4 kb)	283 kb 5' to gene	D'haene et al. (2009)
<i>HBA2</i> (MIM# 141850)	α -thalassaemia	16p13.3	Deletions, various	>20 kb 5' to gene	Hatton et al. (1990) Romao et al. (1991) Viprakasit et al. (2003)
<i>HBB</i> (MIM# 141900)	β -thalassaemia	11p15.5	Deletions, various	>50 kb 5' to gene	Driscoll et al. (1989) Harteveld et al. (2005) Koenig et al. (2009)
<i>PAX6</i> (MIM# 607108)	Aniridia	11p13	Deletions (975 kb, 1105 kb)	11.6 kb and 22.1 kb 3' to gene	Lauderdale et al. (2000)
<i>PITX2</i> (MIM# 601542)	Rieger syndrome	4q25	Translocation breakpoint	~90 kb 5' to gene	Flomen et al. (1998)
<i>POU3F4</i> (MIM# 300039)	X-linked deafness type 3 (DFN3)	Xq21.1	Deletions, various	~900 kb 5' to gene	de Kok et al. (1996)
<i>SHH</i> (MIM# 600725)	Preaxial polydactyly	7q36	<i>De novo</i> reciprocal t(5,7) (q11,q36) translocation breakpoint	~1 Mb 5' to gene	Lettice et al. (2002)
<i>SOX9</i> (MIM# 608160)	Acampomelic campomelic dysplasia	17q24.3	Deletion (960 kb) <i>De novo</i> 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)
<i>SHOX</i>	Leri-Weill	Xp22.33	Deletions, various	30-250 kb 3' to gene	Benito-Sanz et al. (2005)

Deleted: (~105 kb)

Deleted: 7

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47

(MIM# 312865)	dyschondrosteosis				
<i>TRPS1</i> (MIM# 604386)	Ambras syndrome	8q23.3	Inversion breakpoint	7.3 Mb 3' to gene	Fantauzzo et al. (2008)
<i>TWIST</i> (MIM# 601622)	Saethre-Chotzen syndrome	7p21.1	Inversion and translocation breakpoints	>260 kb 3' to gene	Cai et al. (2003)

For Peer Review

Table 4 Disease-causing mutations and disease-associated polymorphisms in microRNA and small nucleolar RNA genes

Gene (MIM#)	Disease/disease association	Chromosomal location	Nature and relative location of mutation*	Reference
<i>MIR16-1</i> (MIM# 609704)	Chronic lymphocytic leukemia, association with	13q14.3	+7C>T ¹	Calin et al. (2005)
<i>MIR17</i> (MIM# 609416)	Breast cancer, association with	13q31.3	+137C>T	Shen et al. (2009)
<i>MIR30C1</i>	Breast cancer, association with	1p34.2	+48G>A	Shen et al. (2009)
<i>MIR96</i> (MIM# 611606)	Hearing loss, progressive	7q32.2	+13G>A	Mencía et al. (2009)
<i>MIR96</i> (MIM# 611606)	Hearing loss, progressive	7q32.2	+14C>A	Mencía et al. (2009)
<i>MIR125A</i> (MIM# 611191)	Breast cancer, association with	19q13.33	+8G>T ¹	Li et al. (2009c)
<i>MIR146A</i> (MIM# 610566)	Papillary thyroid carcinoma, association with	5q33.3	+60G>C ¹	Jazdzewski et al. (2008)
<i>MIR191</i>	Ovarian cancer, predisposition to	3p21.31	+15G>C	Shen et al. (2010)
<i>MIR196A2</i> (MIM# 609687)	Non-small cell lung cancer survival, associated with	12q13.13	+78C>T ¹	Hu et al. (2008)
<i>MIR206</i> (MIM# 611599)	Cancers, reduced expression in association with	6p12.2	+121C>T	Wu et al. (2008)
<i>MIR499</i>	Breast cancer, increased risk, association with	20q11.22	+73A>G ¹	Hu et al. (2009)
<i>MIR502</i>	Schizophrenia, association with	Xp11.23	+13C>T ¹	Sun et al. (2009)
<i>MIR510</i>	Schizophrenia, association with	Xp27.3	+48T>C ¹	Sun et al. (2009)
<i>Mir2861</i>	Osteoporosis, primary	2	+33C>G	Li et al. (2009d)
<i>MIRLET7E</i> (MIM# 611250)	Cancers, reduced expression in association with	19q13.41	+98T>A	Wu et al. (2008)
<i>SNORD50A</i> (MIM# 613117)	Breast and prostate cancer, association with	6q14.3	ΔTT	Dong et al. (2008)
<i>SNORD116@</i> gene cluster (MIM# 605436)	Prader-Willi syndrome	15q11.2	Gross deletion	Sahoo et al. (2008)

*Location given is relative to the transcriptional initiation site of the specified gene.

¹Disease-associated polymorphism.

Deleted: phenotype

Deleted: Reduced expression

Deleted: Breast cancer, association with

Deleted: *MIR17* (MIM# 609416)Deleted: *MIR30C1*

Deleted: Breast cancer, association with

Deleted: *MIR16-1* (MIM# 609704)

... [1]

Deleted: *MIR96* (MIM# 611606)

... [2]

Deleted: ¶

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Page 1: [1] Deleted		INSRV	3/15/2010 1:45:00 PM	
<i>MIR16-1</i> (MIM# 609704)	Reduced expression	13q14.3	+7C>T ¹	Calin et al. (2005)
Page 1: [2] Deleted		INSRV	3/15/2010 1:47:00 PM	
<i>MIR96</i> (MIM# 611606)	Hearing loss, progressive	7q32.2	+13G>A	Mencía et al. (2009)
<i>MIR96</i> (MIM# 611606)	Hearing loss, progressive	7q32.2	+14C>A	Mencía et al. (2009)

For Peer Review

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

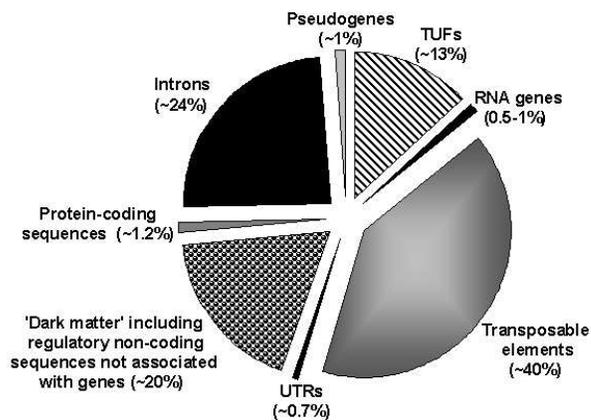


Figure 1

81x60mm (300 x 300 DPI)

Review

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

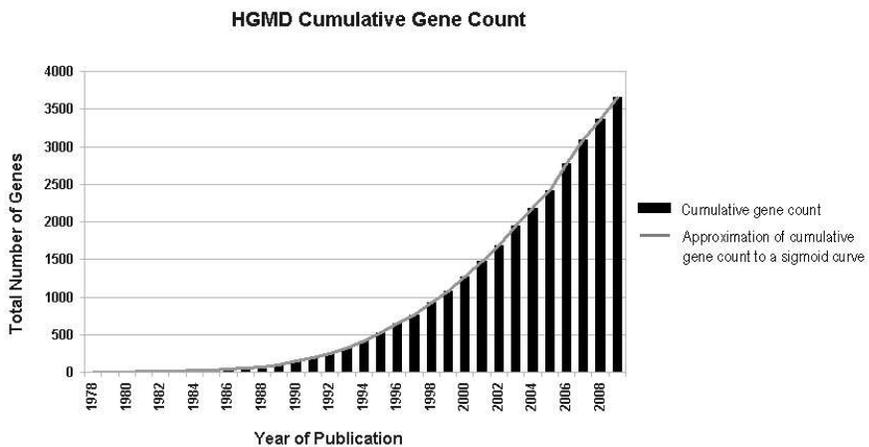


Figure 2

81x60mm (300 x 300 DPI)