Journal of Genetic Syndromes & Gene Therapy

Cooper, J Genet Syndr Gene Ther 2012, 3:4
DOI: 10.4172/2157-7412.1000e113

Editorial Open Access

Exploring the Role of Genomic Architecture and the Local DNA Sequence Environment in Mediating Gene Mutations Underlying Human Inherited Disease

David N Cooper*

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

Different types of human gene mutation vary dramatically in size, from gross structural variants (SVs) to subtle single base-pair substitutions. What they all have in common, however, is that their nature, location and frequency are often determined either by specific characteristics of the local DNA sequence environment or by higher-order features of the genomic architecture. It is now recognized that the human genome contains 'pervasive architectural flaws' [1] in that certain DNA sequences are inherently mutation-prone by virtue of their base composition, sequence repetitivity and/or epigenetic modification [2,3]. The mutability of a given gene or genomic region may also be influenced indirectly by a variety of non-canonical (non-B) secondary structures whose formation is facilitated by the underlying DNA sequence [4,5]. Since these non-B DNA structures can interfere with subsequent DNA replication and repair, and may serve to increase mutation frequencies in generalized fashion (i.e. both in the context of subtle mutations and SVs), they have the potential to serve as a unifying concept in studies of mutational mechanisms underlying human inherited disease [6]. Our task is to come to understand the ground rules that characterize the different mechanisms of mutagenesis in order to apply this knowledge in the context not only of the analysis and diagnosis of human genetic disease, but also eventually perhaps, in the cause of its therapeutic correction.

To these ends, we have, over the last 20 years, compiled a comprehensive core collection of data on germ-line mutations in nuclear genes underlying or associated with human inherited disease in order to study the nature and underlying mechanisms of human gene mutation. This resource, the Human Gene Mutation Database (HGMD; [7]) has been made publicly available since 1996 (http://www.hgmd.org). Data catalogued include single base-pair substitutions in coding, regulatory, and splicing-relevant regions, micro-deletions and micro-insertions, indels and triplet repeat expansions, as well as gross gene deletions, insertions, duplications and complex rearrangements. Each mutation is entered into HGMD only once, in order to avoid confusion between recurrent and identical-by-descent lesions. By October 2012, the database contained in excess of 130,000 different lesions (HGMD Professional release 2012.3; http://www.biobase-international.com/product/hgmd) detected in ~5,000 different nuclear genes, with new entries currently accumulating at a rate in excess of 10,000 per annum; ~6,000 of these entries constitute disease-associated and functional polymorphisms.

HGMD data have been used by its curators and their collaborators to perform an extensive series of meta-analyses of different types of gene mutation causing human inherited disease [3]. In particular, HGMD data have been utilized to study the role that repetitive sequence elements, sequence homologies, and specific motifs play in promoting mutagenesis, and to explore in detail the underlying mutational mechanisms [8-14]. New insights have been obtained into the phenotypic/clinical consequences of several entirely novel types of human gene mutation *viz.* mutations giving rise to gains of glycosylation [15] mutations which disrupt predicted exon splice enhancers [16,17] and closely spaced multiple mutations that may constitute signatures of transient hypermutability in human genes [18].

The recognition that certain DNA sequences are inherently hypermutable has been accompanied by an emerging understanding of how DNA sequence influences (and indeed often underpins) secondary structure formation, how certain local DNA structures can themselves be mutagenic, and how the type and frequency of the resulting mutations can in turn help to explain the nature and prevalence of specific human genetic diseases [19,20]. Studies of hypermutable sequences have also provided important insights into the endogenous nature of many of the known mechanisms of mutagenesis, for example CpG deamination [21,22] or slipped mispairing at the DNA replication fork [23], that are responsible for quite different types of recurring micro-lesion. Recurrent mutation ensures that some missense mutations and micro-deletions/micro-insertions are shared between the germ-line and the soma: these lesions are characterized by higher mutability rates, greater physicochemical differences between wild-type and mutant residues, and a tendency to occur in evolutionarily conserved residues and within CpG/CpHpG oligonucleotides [24].

HGMD data made possible the identification of potential disease-causing variants in the first two human diploid genomes ever sequenced [25,26]. In collaboration with the 1000 Genomes Project, we have contributed HGMD data to aid (i) the construction of a map of the location, allele frequency and local haplotype structure of ~15 million human single nucleotide polymorphisms, allowing direct estimation of the rate of *de novo* germ-line base substitution [27], (ii) the definition of the functional spectrum of human low-frequency coding variation [28] and (iii) the identification of putative loss of function variants in 185 human genomes thereby demonstrating that human genomes typically contain ~100 genuine loss of function variants with ~20 genes completely inactivated [29]. More recently, we have assessed the numbers of potentially deleterious variants in the genomes of apparently healthy humans by cross-comparing whole-genome sequence data from 179 individuals in the 1000 Genomes Pilot Project with HGMD data [30]. Each individual was found to harbour between 40 and 110 variants classified by the HGMD as disease-causing

***Corresponding author:** David N Cooper, Institute of Medical Genetics, School of Medicine, Cardiff University, Heath Park, Cardiff CF14 4XN, UK, Tel: +44 2920 744062; E-mail: cooperDN@cardiff.ac.uk

Received October 12, 2012; **Accepted** October 12, 2012; **Published** October 13, 2012

Citation: Cooper DN (2012) Exploring the Role of Genomic Architecture and the Local DNA Sequence Environment in Mediating Gene Mutations Underlying Human Inherited Disease. J Genet Syndr Gene Ther 3:e113. doi:10.4172/2157-7412.1000e113

mutations (DMs), 3-24 in the homozygous state, as well as many polymorphisms putatively associated with disease [30]. Whereas many of these DMs could represent disease annotation errors, between 0 and 8 DMs per individual (0-1 homozygous) are predicted to be highly damaging and could provide information of direct medical relevance to the individuals concerned.

The sequencing of the human genome was completed some time ago and its structural and functional annotation is now well underway. Considerable numbers of human genome sequences are now becoming publicly available almost by the week. It is in this context that personalized genomic medicine is beginning to come to the fore, and human gene mutation data are assuming ever greater importance. Indeed, the availability of HGMD data has facilitated the study of gene mutation by making possible:

1. A first description of the spectrum of genetic variation underlying human inherited disease;
2. The identification of the factors that determine the propensity of certain DNA sequences to undergo germ-line mutation;
3. An improved understanding of the mechanisms underlying the different types of human gene mutation and the reasons why different types of mutational lesion occur with different frequencies in different genes;
4. The identification of disease states that exhibit incomplete mutational spectra, prompting the search for novel gene lesions associated with different clinical phenotypes;
5. The optimization of mutational screening strategies;
6. The meaningful comparison between the potentially different mechanisms of mutagenesis underlying inherited and somatic disease;
7. Studies of human genetic disease in the context of the evolutionary conservation of the affected nucleotide sequences or encoded amino acid residues;
8. Extrapolation towards the genetic basis of other, more complex traits and diseases.
9. A source of data for clinical interpretive use in next generation sequencing/exome screening studies.

For all these reasons, we should endeavor to follow the advice of the founder of modern human genetics, William Bateson, who, in the context of collecting plant mutants, exhorted us to "treasure our exceptions."

References

1. Avise JC (2010) Colloquium paper: footprints of nonsentient design inside the human genome. Proc Natl Acad Sci U S A 107 Suppl 2: 8969-8976.
2. Nakken S, Rødland EA, Hovig E (2010) Impact of DNA physical properties on local sequence bias of human mutation. Hum Mutat 31: 1316-1325.
3. Cooper DN, Bacolla A, Férec C, Vasquez KM, Kehrer-Sawatzki H, et al. (2011) On the sequence-directed nature of human gene mutation: the role of genomic architecture and the local DNA sequence environment in mediating gene mutations underlying human inherited disease. Hum Mutat 32: 1075-1099.
4. Bacolla A, Jaworski A, Larson JE, Jakupciak JP, Chuzhanova N, et al. (2004) Breakpoints of gross deletions coincide with non-B DNA conformations. Proc Natl Acad Sci U S A 101: 14162-14167.
5. Chuzhanova N, Chen JM, Bacolla A, Patrinos GP, Férec C, et al. (2009) Gene conversion causing human inherited disease: evidence for involvement of non-B-DNA-forming sequences and recombination-promoting motifs in DNA breakage and repair. Hum Mutat 30: 1189-1198.
6. Bacolla A, Wang G, Jain A, Chuzhanova NA, Cer RZ, et al. (2011) Non-B DNA-forming sequences and WRN deficiency independently increase the frequency of base substitution in human cells. J Biol Chem 286: 10017-10026.
7. Stenson PD, Mort M, Ball EV, Howells K, Phillips AD, et al. (2009) The Human Gene Mutation Database: 2008 update. Genome Med 1: 13.
8. Krawczak M, Ball EV, Cooper DN (1998) Neighboring-nucleotide effects on the rates of germ-line single-base-pair substitution in human genes. Am J Hum Genet 63: 474-488.
9. Krawczak M, Chuzhanova NA, Stenson PD, Johansen BN, Ball EV, et al. (2000) Changes in primary DNA sequence complexity influence the phenotypic consequences of mutations in human gene regulatory regions. Hum Genet 107: 362-365.
10. Chuzhanova NA, Anassis EJ, Ball EV, Krawczak M, Cooper DN (2003) Meta-analysis of indels causing human genetic disease: mechanisms of mutagenesis and the role of local DNA sequence complexity. Hum Mutat 21: 28-44.
11. Chuzhanova N, Abeysinghe SS, Krawczak M, Cooper DN. (2003). Translocation and gross deletion breakpoints in human inherited disease and cancer II: Potential involvement of repetitive sequence elements in secondary structure formation between DNA ends. Hum. Mutat 22: 245-251.
12. Ball EV, Stenson PD, Abeysinghe SS, Krawczak M, Cooper DN, et al. (2005) Microdeletions and microinsertions causing human genetic disease: common mechanisms of mutagenesis and the role of local DNA sequence complexity. Hum Mutat 26: 205-213.
13. Chen JM, Chuzhanova N, Stenson PD, Férec C, Cooper DN (2005) Meta-analysis of gross insertions causing human genetic disease: novel mutational mechanisms and the role of replication slippage. Hum Mutat 25: 207-221.
14. Chen JM, Chuzhanova N, Stenson PD, Férec C, Cooper DN (2005) Complex gene rearrangements caused by serial replication slippage. Hum Mutat 26: 125-134.
15. Vogt G, Chapgier A, Yang K, Chuzhanova N, Feinberg J, et al. (2005) Gains of glycosylation comprise an unexpectedly large group of pathogenic mutations. Nat Genet 37: 692-700.
16. Sanford JR, Wang X, Mort M, Vanduyn N, Cooper DN, et al. (2009) Splicing factor SFRS1 recognizes a functionally diverse landscape of RNA transcripts. Genome Res 19: 381-394.
17. Sterne-Weiler T, Howard J, Mort M, Cooper DN, Sanford JR (2011) Loss of exon identity is a common mechanism of human inherited disease. Genome Res 21: 1563-1571.
18. Chen JM, Férec C, Cooper DN (2009) Closely spaced multiple mutations as potential signatures of transient hypermutability in human genes. Hum Mutat 30: 1435-1448.
19. Bacolla A, Larson JE, Collins JR, Li J, Milosavljevic A, et al. (2008) Abundance and length of simple repeats in vertebrate genomes are determined by their structural properties. Genome Res 18: 1545-1553.
20. Arnheim N, Calabrese P (2009) Understanding what determines the frequency and pattern of human germline mutations. Nat Rev Genet 10: 478-488.
21. Krawczak M, Ball EV, Cooper DN (1998) Neighboring-nucleotide effects on the rates of germ-line single-base-pair substitution in human genes. Am J Hum Genet 63: 474-488.
22. Cooper DN, Mort M, Stenson PD, Ball EV, Chuzhanova NA (2010) Methylation-mediated deamination of 5-methylcytosine appears to give rise to mutations causing human inherited disease in CpNpG trinucleotides, as well as in CpG dinucleotides. Hum Genomics 4: 406-410.
23. Chen JM, Chuzhanova N, Stenson PD, Férec C, Cooper DN (2005) Intrachromosomal serial replication slippage in trans gives rise to diverse genomic rearrangements involving inversions. Hum Mutat 26: 362-373.
24. Ivanov D, Hamby SE, Stenson PD, Phillips AD, Kehrer-Sawatzki H, et al. (2011) Comparative analysis of germline and somatic microlesion mutational spectra in 17 human tumor suppressor genes. Hum Mutat 32: 620-632.
25. Levy S, Sutton G, Ng PC, Feuk L, Halpern AL, et al. (2007) The diploid genome sequence of an individual human. PLoS Biol 5: e254.
26. Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, et al. (2008) The complete genome of an individual by massively parallel DNA sequencing. Nature 452: 872-876.
27. 1000 Genomes Project Consortium (2010) A map of human genome variation from population-scale sequencing. Nature 467: 1061-1073.
28. Marth GT, Yu F, Indap AR, Garimella K, Gravel S, et al. (2011) The functional spectrum of low-frequency coding variation. Genome Biol 12: R84.

29. MacArthur DG, Balasubramanian S, Frankish A, Huang N, Morris J, et al. (2012) A systematic survey of loss-of-function variants in human protein-coding genes. Science 335: 823-828.

30. Xue Y, Chen Y, Ayub Q, Huang N, Ball EV, et al. (2012) Deleterious and disease allele prevalence in healthy individuals: insights from current predictions, mutation databases and population-scale resequencing. Am J Hum Genet 91: 1022-1032.