Major changes in our DNA lead to major changes in our thinking

Jonathan Sebat

Variability in the human genome has far exceeded expectations. In the course of the past three years, we have learned that much of our naturally occurring genetic variation consists of large-scale differences in genome structure, including copy-number variants (CNVs) and balanced rearrangements such as inversions. Recent studies have begun to reveal that structural variants are an important contributor to disease risk; however, structural variants as a class may not conform well to expectations of current methods for gene mapping. New approaches are needed to understand the contribution of structural variants to disease.

A subject that has gained much attention in the field of human genetics has been the discovery that structural variation of the genome including large insertions and deletions of DNA, collectively termed copy-number variants (CNVs), as well as balanced chromosomal rearrangements, such as inversions, contribute to a major proportion of genetic difference in humans. Following the first studies to report the widespread abundance of CNVs in humans[1,2], knowledge of structural variation has grown rapidly, owing to steady improvements in oligonucleotide microarray technology and the development of new sequencing-based[3] and SNP-based[4,5] structural variant detection methods, and their use in large-scale projects to map structural variation in different populations[6,7].

It is now recognized that the genomes of any two individuals in the human population differ more at the structural level than at the nucleotide sequence level. Conservative estimates suggest that CNVs between individuals amount to 4 Mb (1/800 bp) of genetic difference[3], and less conservative estimates put this figure in the range of 5–24 Mb[7]. By either measure, CNVs account for more nucleotide variation on average than single nucleotide polymorphisms (SNPs), which account for approximately 2.5 Mb (1/1,200 bp)[8–10]. Therefore, the total genomic variability between humans is significantly greater than previously thought, amounting to a difference of at least 0.2%, >0.12% at the structural level and 0.08% at the nucleotide level.

In retrospect, perhaps it should not have been so surprising to find our genome riddled with deletions, duplications and inversions. Remarkable genomic plasticity had been observed in model organisms much earlier, for example when cytogenetic studies by Barbara McClintock found that transposition events explained nonmendelian patterns of segregation for certain maize phenotypes[11]. Later, studies of the human genome revealed the presence of cytogenetically visible polymorphisms in heterochromatin length[12]. Nevertheless, this aspect of human variability was not unmistakable. The proverbial lamp post was firmly fixed in the opposite direction because reliable methods did not exist for ascertaining CNVs genome wide, and because prevailing methods for gene mapping worked best in the context of a static genome.

Technological innovations have opened the door to a fundamental aspect of human genomic variation that was previously unrecognized and have opened a new window into the genetic basis of disease. Methods for detecting CNVs genome-wide have the power to identify risk factors for disease directly, and thereby overcome some key limitations of traditional gene mapping approaches. What these studies have begun to reveal is that structural variants contribute to disease and the risk factors involved often do not conform to the expectations of prevailing association-based methods. This has consequences for what methods should be used to study CNVs, and it also has implications for the respective contribution of common and rare CNVs in disease.

Much of what was previously known about the role of CNVs in disease comes from a rich literature on 'genomic disorders'[13]. Genomic disorders are defined as a diverse group of genetic diseases that are each caused by an alteration in DNA copy number. These mutations can be relatively large, microscopically visible imbalances, such as in Prader-Willi syndrome[14], or they may be much smaller, requiring higher resolution detection methods, such as in Williams Syndrome[15]. Genomic disorders are typically sporadic in nature because the CNV in most cases is a *de novo* mutation with nearly complete penetrance, and because the affected individuals have severe developmental problems and are unlikely to have offspring. However, there are notable examples of mendelian disease traits associated with CNVs. For example, duplications of the gene for peripheral myelin protein 22 (*PMP22*) cause the dominant neuropathy Charcot-Marie Tooth disease type 1A[16], and deletions of the α-globin gene cluster cause the recessive anemia α-thalassemia[17].

Previous knowledge of genomic disorders was limited by the available methods: that is, limited primarily to disorders that form

Jonathan Sebat is at the Cold Spring Harbor Laboratory, One Bungtown Road, Cold Spring Harbor, New York 11724, USA
e-mail: sebat@cshl.edu

Published online 27 June 2007; doi:10.1038/ng2095

a distinct clinical entity and where genomic imbalances are often cytogenetically visible or inherited in a dominant fashion. The application of high resolution genome-wide methods to sporadic disorders promises to greatly improve the power to detect CNVs that cause disease[18]. In addition, these genetic findings are proving helpful in informing physicians about the clinical features of a disorder. For example, by identifying new clinically relevant CNVs and correlating these changes with phenotypic information, new genomic disorders have been defined that had not been previously recognized as distinct clinical entities[19–21].

Because each genomic disorder is a clinically defined syndrome linked with a single locus, and each is nearly 100% penetrant, these diseases are individually quite rare in the human population. However, it is not a great stretch of the imagination to envisage another type of genomic disorder that is similar in many respects to those described above, but is instead a common disease. Consider, for instance, a disorder where the clinically defined phenotype is not associated with a single locus, but is instead associated with the occurrence of a single dominant mutation involving any one of 50 autosomal genes. Assuming a spontaneous CNV mutation rate of 1/10,000 per locus on average, a 'complex genomic disorder' of this kind would be relatively common, with a population prevalence of 1/200.

Spontaneous copy-number mutation has recently emerged as a relevant issue in common disease, for example in autism spectrum disorders (ASD) where the prevalence is estimated at 1/150 (ref. 22). A high frequency of spontaneous copy-number mutation has been reported in ASD[23]. In this study, 10% (12/118) of sporadic cases were associated with a *de novo* CNV, a significantly higher rate than in families with more than one affected child (3%) or in healthy controls (1%). In a separate study focusing on a subset of individuals with syndromic autism (combined with dysmorphic features and mental retardation), Jacquemont *et al.* found the frequency of *de novo* CNVs to be 24% (7/26)[24]. The frequency of *de novo* mutation found in these studies is an underestimate. Considering that microarray analysis at a resolution of ≤85,000 probes detects fewer than 10% of all CNVs, the total frequency of *de novo* copy-number changes in autism could be several-fold higher than what has been reported, raising the possibility that spontaneous structural mutations may contribute to disease in a majority of patients. The mutations identified in these studies occurred at many loci throughout the genome, and no individual CNV was found in more than 1% of cases. This high degree of heterogeneity is consistent with the findings of early genetic studies of autism that found evidence for linkage at many locations in the genome[25]. An important implication of the recent findings in autism is that the genetic component of certain common disorders may consist largely of a constellation of rare, highly penetrant mutations. This line of evidence also favors the notion that much of the sporadic nature of autism can be attributed to spontaneous mutation at individual loci, in contrast to models that explain the lack of mendelian segregation by the additive or multiplicative effects of alleles at multiple loci[26].

A high rate of structural mutation is not a property of autism or other neurodevelopmental disorders; it is a property of the human genome. Therefore, frequent spontaneous copy-number mutation may play a prominent role in adult-onset neuropsychiatric disorders or indeed in any heritable disease whose effect on reproductive fitness and its prevalence in the population seem to defy darwinian logic[27]. There are several examples of familial genomic disorders[28]; but one fact that is not well appreciated is that they are invariably a result of spontaneous mutation (occurring in recent ancestry). For example, autosomal dominant and sporadic forms of Charcot Marie-Tooth disease type 1 are caused by identical duplications of the gene *PMP22*, and are typically inherited in the dominant pedigrees and *de novo* in the sporadic cases[29]. The common α-globin gene deletions found in different isolated populations each occur on a different haplotype background, implying that the deletions arose independently in each group[17], and recently a high rate of spontaneous α-globin mutation in sperm was confirmed by Lam *et al.*[30]. Thus, the persistence of some diseases in the global population may be due to a high rate of random mutation and a large number of potential sites in the genome which, when altered, can produce a similar disease phenotype.

It is certain that common copy-number polymorphisms (CNPs) will underlie heritable human traits. Deletions are known to underlie some relatively common human traits, such as the Rh-negative blood type[31] and color blindness[32,33]. More recently, CNPs have been shown to contribute to disease risk. For example duplications of the gene *CCL3L1* have been found to influence susceptibility to infectious disease[34], and CNPs of *FCGR3B* predispose to systemic autoimmune disease[35,36].

Although the variation in the above cases is common, for a variety of reasons, SNP-based methods may fail to ascertain much of the structural variation at these and other loci. Population-based studies have shown that CNPs as a class have reduced linkage disequilibrium with neighboring SNPs[5,7]. Potential reasons for this effect could include reduced SNP coverage in CNP regions and in regions rich in segmental duplications, or recurrent copy-number mutations at individual loci. Recurrent mutation is certainly evident at some CNP loci, based on the existence of several common alleles. For example, quantitative PCR measurements of *FCGR3B* in a cohort of European ancestry showed four distinct distributions of diploid copy number, indicating that at least three distinct genomic structures, consisting of zero, one or two copies per chromosome, are common in the population[35]. The distribution of *CCL3L1* copy-number alleles was found to be greater still, varying between zero and seven copies per genome[34]. In both of the previous examples, disease risk was associated primarily with the dosage of a gene, rather than with any single allele. Thus, some CNPs constitute common variation that segregates independently of SNPs.

In the past three years, it has become obvious that the structure of the human genome is not static. Furthermore, it is becoming increasingly evident that copy-number variability differs from nucleotide variability in terms of the rate at which copy-number mutations occur spontaneously in the genome[37] and the allelic diversity that may occur as a result. Therefore, CNVs require special consideration in large-scale genetic studies of disease[38]. For loci with the highest mutation rates, linkage disequilibrium–based methods of association are not effective[39]; therefore, direct methods of CNV detection are required. In addition, for some diseases a family-based study may have advantages over a case-control design because it would allow the identification of *de novo* mutations. Lastly, confirming the association of candidate genes originally identified from genome-wide CNV scans will surely require methods that are different from conventional approaches for fine-mapping candidate regions identified in whole-genome association studies, and are likely to involve a more comprehensive analysis of CNVs and SNPs, for example using a combination of tiling-resolution oligonucleotide arrays and high-throughput sequencing technology. When candidate loci originally identified from CNV studies are examined more closely, a new surprise may be in store in terms of the number of genes and diversity of causative alleles that contribute to disease.

ACKNOWLEDGMENTS

Special thanks to M. Wigler, M.-C. King and D. Levy for helpful discussions and to J. Lupski for his critical reading of the manuscript. My laboratory is funded by the Simons Foundation, Lattner Foundation, Stanley Foundation, the US National Institutes of Health (National Institute of Mental Health, National Human Genome Research Institute), Autism Speaks and the Southwest Autism Research and Resource

Center.

COMPETING INTERESTS STATEMENT
The author declares no competing financial interests.

Published online at http://www.nature.com/naturegenetics
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/

1. Iafrate, A.J. *et al.* Detection of large-scale variation in the human genome. *Nat. Genet.* **36**, 949–951 (2004).
2. Sebat, J. *et al.* Large-scale copy number polymorphism in the human genome. *Science* **305**, 525–528 (2004).
3. Tuzun, E. *et al.* Fine-scale structural variation of the human genome. *Nat. Genet.* **37**, 727–732 (2005).
4. McCarroll, S.A. *et al.* Common deletion polymorphisms in the human genome. *Nat. Genet.* **38**, 86–92 (2006).
5. Conrad, D.F., Andrews, T.D., Carter, N.P., Hurles, M.E. & Pritchard, J.K. A high-resolution survey of deletion polymorphism in the human genome. *Nat. Genet.* **38**, 75–81 (2006).
6. Eichler, E.E. *et al.* Completing the map of human genetic variation. *Nature* **447**, 161–165 (2007).
7. Redon, R. *et al.* Global variation in copy number in the human genome. *Nature* **444**, 444–454 (2006).
8. Altshuler, D. *et al.* An SNP map of the human genome generated by reduced representation shotgun sequencing. *Nature* **407**, 513–516 (2000).
9. Wang, D.G. *et al.* Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* **280**, 1077–1082 (1998).
10. Sachidanandam, R. *et al.* A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409**, 928–933 (2001).
11. McClintock, B. The origin and behavior of mutable loci in maize. *Proc. Natl. Acad. Sci. USA* **36**, 344–355 (1950).
12. Jacobs, P.A. Human chromosome heteromorphisms (variants). *Prog. Med. Genet.* **2**, 251–274 (1977).
13. Lupski, J.R. Genomic disorders: structural features of the genome can lead to DNA rearrangements and human disease traits. *Trends Genet.* **14**, 417–422 (1998).
14. Ledbetter, D.H. *et al.* Deletions of chromosome 15 as a cause of the Prader-Willi syndrome. *N. Engl. J. Med.* **304**, 325–329 (1981).
15. Ewart, A.K. *et al.* Hemizygosity at the elastin locus in a developmental disorder, Williams syndrome. *Nat. Genet.* **5**, 11–16 (1993).
16. Lupski, J.R. *et al.* DNA duplication associated with Charcot-Marie-Tooth disease type 1A. *Cell* **66**, 219–232 (1991).
17. Higgs, D.R. *et al.* A review of the molecular genetics of the human α-globin gene cluster. *Blood* **73**, 1081–1104 (1989).
18. Shaw-Smith, C. *et al.* Microarray based comparative genomic hybridisation (array-CGH) detects submicroscopic chromosomal deletions and duplications in patients with learning disability/mental retardation and dysmorphic features. *J. Med. Genet.* **41**, 241–248 (2004).
19. Koolen, D.A. *et al.* A new chromosome 17q21.31 microdeletion syndrome associated with a common inversion polymorphism. *Nat. Genet.* **38**, 999–1001 (2006).
20. Shaw-Smith, C. *et al.* Microdeletion encompassing MAPT at chromosome 17q21.3 is associated with developmental delay and learning disability. *Nat. Genet.* **38**, 1032–1037 (2006).
21. Sharp, A.J. *et al.* Discovery of previously unidentified genomic disorders from the duplication architecture of the human genome. *Nat. Genet.* **38**, 1038–1042 (2006).
22. Centers for Disease Control and Prevention. Prevalence of autism spectrum disorders–autism and developmental disabilities monitoring network, 14 sites, United States, 2002. *MMWR Surveill. Summ.* **56**, 12–28 (2007).
23. Sebat, J. *et al.* Strong association of de novo copy number mutations with autism. *Science* **316**, 445–449 (2007).
24. Jacquemont, M.L. *et al.* Array-based comparative genomic hybridization identifies high frequency of cryptic chromosomal rearrangements in patients with syndromic autism spectrum disorders. *J. Med. Genet.* **43**, 843–849 (2006).
25. Risch, N. *et al.* A genomic screen of autism: evidence for a multilocus etiology. *Am. J. Hum. Genet.* **65**, 493–507 (1999).
26. Pickles, A. *et al.* Latent-class analysis of recurrence risks for complex phenotypes with selection and measurement error: a twin and family history study of autism. *Am. J. Hum. Genet.* **57**, 717–726 (1995).
27. Bassett, A.S., Bury, A., Hodgkinson, K.A. & Honer, W.G. Reproductive fitness in familial schizophrenia. *Schizophr. Res.* **21**, 151–160 (1996).
28. Lee, J.A. & Lupski, J.R. Genomic rearrangements and gene copy-number alterations as a cause of nervous system disorders. *Neuron* **52**, 103–121 (2006).
29. Hoogendijk, J.E. *et al.* De-novo mutation in hereditary motor and sensory neuropathy type I. *Lancet* **339**, 1081–1082 (1992).
30. Lam, K.W. & Jeffreys, A.J. Processes of copy-number change in human DNA: the dynamics of α-globin gene deletion. *Proc. Natl. Acad. Sci. USA* **103**, 8921–8927 (2006).
31. Blunt, T., Steers, F., Daniels, G. & Carritt, B. Lack of RH C/E expression in the Rhesus D–phenotype is the result of a gene deletion. *Ann. Hum. Genet.* **58**, 19–24 (1994).
32. Vollrath, D., Nathans, J. & Davis, R.W. Tandem array of human visual pigment genes at Xq28. *Science* **240**, 1669–1672 (1988).
33. Nathans, J., Piantanida, T.P., Eddy, R.L., Shows, T.B. & Hogness, D.S. Molecular genetics of inherited variation in human color vision. *Science* **232**, 203–210 (1986).
34. Gonzalez, E. *et al.* The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. *Science* **307**, 1434–1440 (2005).
35. Aitman, T.J. *et al.* Copy number polymorphism in *Fcgr3* predisposes to glomerulonephritis in rats and humans. *Nature* **439**, 851–855 (2006).
36. Fanciulli, M. *et al. FCGR3B* copy number variation is associated with susceptibility to systemic, but not organ-specific, autoimmunity. *Nat. Genet.* **39**, 721–723 (2007).
37. Lupski, J.R. Genomic rearrangements and sporadic disease. *Nat. Genet.* **39**, S43–S47 (2007).
38. McCarroll, S.A. & Altshuler, D. M. Copy-number variation and association studies of human disease. *Nat. Genet.* **39**, S37–S42 (2007).
39. Pritchard, J.K. Are rare variants responsible for susceptibility to complex diseases? *Am. J. Hum. Genet.* **69**, 124–137 (2001).