Nevertheless, these new microarray methods are notable for their simplicity and their ability to be scaled up. This is illustrated by the construction of 'whole-genome scan' arrays, in which probes representing the vast majority of all exons predicted by Genscan in the draft human genome (over 442,000) were synthesized on a set of 50 arrays and tested under two pairs of conditions[1]. This feat probably represents the largest experimental test of a bioinformatics prediction ever undertaken. It also illustrates the increasingly close coupling of bioinformatics and functional genomics: a bioinformatics tool (Genscan) is used to design a functional genomics experiment (exon-scanning array) that is interpreted by another bioinformatics tool (the clustering algorithm used to infer EVGs). Perhaps surprisingly, the authors choose not to enter into the controversy about exactly how many genes the human genome contains[12], apparently waiting for the results of additional whole-genome array experiments.

Clearly, hybridization of whole-genome scan arrays to mRNA derived from a larger number of tissues and cell lines is in the works and will increase the sensitivity of this approach. Construction of similar arrays that represent other genomes is sure to follow. Increases in the sensitivity of the experimental and computational aspects of this technology should allow detection of alternative splicing, increasingly recognized as a widespread and important mode of gene regulation[13,14]. □


1. Shoemaker, D.D. *et al. Nature* **409**, 922–927 (2001).
2. Hillier, L.D. *et al. Genome Res.* **6**, 807–828 (1996).
3. Altschul, S.F. *et al. Nucleic Acids Res.* **25**, 3389–3402 (1997).
4. Birney, E. & Durbin, R. *Genome Res.* **10**, 547–548 (2000).
5. Roest Crollius, H. *et al. Nature Genet.* **25**, 235–238 (2000).
6. Batzoglou, S., Pachter, L., Mesirov, J.P., Berger, B. & Lander, E.S. *Genome Res.* **10**, 950–958 (2000).
7 Burge, C. & Karlin, S. *J. Mol. Biol.* **268**, 78–94 (1997).
8. Dunham, I. *et al. Nature* **402**, 489–495 (1999).
9. de Souza, S.J. *et al. Proc. Natl. Acad. Sci. USA* **97**, 12690–12693 (2000).
10. Penn, S.G., Rank, D.R., Hanzel, D.K. & Barker, D.L. *Nature Genet.* **26**, 315–318 (2000).
11. Burge, C.B. & Karlin, S. *Curr. Opin. Struct. Biol.* **8**, 346–354 (1998).
12. Dunham, I. *Yeast* **17**, 218–224 (2000).
13. Mironov, A.A., Fickett, J.W. & Gelfand, M.S. *Genome Res.* **9**, 1288–1293 (1999).
14. The International Human Genome Sequencing Consortium *Nature* **409**, 860–921 (2001).
15. Carninci, P. *et al. Genome Res.* **10**, 1617–1630 (2000).




# Variation is the spice of life

Leonid Kruglyak[1] & Deborah A. Nickerson[2]

[1]*Division of Human Biology, Program in Genetics, Fred Hutchinson Cancer Research Center, Seattle, Washington, USA.* [2]*Department of Molecular Biotechnology, University of Washington, Seattle, Washington, USA.*
*e-mail: leonid@fhcrc.org and debnick@u.washington.edu*

**Assembling a comprehensive catalog of common human sequence polymorphisms is a key goal of the Human Genome Project. The International SNP Map Working Group has compiled a map of 1.4 million single-nucleotide polymorphisms. What fraction of all human sequence variation is captured in this collection, and what further advances are needed before the map becomes a useful tool for genetic studies?**

'The Human Genome Project' has always been something of a misnomer, implying the existence of a single human genome. Of course, every person on the planet—with the exception of identical twins—has a unique genome, and even though any two genomes are roughly 99.9% identical, that still leaves millions of differences among the 3.2 billion base pairs. It is precisely these differences that account for heritable variation among individuals, including susceptibility to disease. Most of the differences take the form of substitutions at a single base pair, dubbed single-nucleotide polymorphisms (SNPs). The importance of sequence variation has not been lost on the genomics community[1,2], and The International SNP Map Working Group has now reported the results of an ambitious survey, providing a map of 1.4 million candidate SNPs (ref. 3). This effort increases the number of known human polymorphisms by more than an order of magnitude, and promises great advances in medical and evolutionary genetics. How comprehensive is the map, and what must happen before its promise can be realized?

## How many SNPs in the human genome?

The number 1.4 million is impressive in itself, but how does it compare with the total amount of variation in our species? The question is not as straightforward as it might seem. If we ask for the total number of nucleotides that vary in today's population, a quick calculation based on 6 billion people and a mutation rate of $2\times10^{-8}$ per base pair per generation shows that every site at which mutations are compatible with life has been mutated an average of 240 times in just the most recent generation (and many more times in human history). Of course, most of these mutations are vanishingly rare. A more reasonable basis for comparison is obtained if we restrict our consideration to SNPs with both alleles occurring in the population at or above a minimal frequency. The traditional definition of 'polymorphism' sets this frequency at 1%. The Working Group observes that two haploid genomes differ at 1 nucleotide per 1,331 bp (ref. 3), a rate that is consistent with previous estimates[4] (but is expected to vary somewhat between ethnic groups). Classic neutral theory of population genetics allows us to infer from this rate the existence of 11 million sites in a genome of 3.2

**Table 1 • Occurrence of SNPs in the human population and their representation in the current collection**

| Minimal allele frequency | Expected SNP number (millions) | Expected SNP frequency (bp) | Expected % in collection |
|---|---|---|---|
| 1% | 11.0 | 290 | 11–12 |
| 5% | 7.1 | 450 | 15–17 |
| 10% | 5.3 | 600 | 18–20 |
| 20% | 3.3 | 960 | 21–25 |
| 30% | 2.0 | 1,570 | 23–27 |
| 40% | 0.97 | 3,280 | 24–28 |

**Table 2 • Detection rate for SNPs with a given minimal allele frequency**

| n | 1% | 5% | 10% | 20% | 30% | 40% |
|---|---|---|---|---|---|---|
| 2 | .21 | .30 | .36 | .43 | .47 | .49 |
| 3 | .32 | .46 | .55 | .65 | .71 | .74 |
| 4 | .39 | .56 | .66 | .77 | .83 | .86 |
| 5 | .44 | .62 | .73 | .84 | .90 | .93 |
| 6 | .48 | .68 | .78 | .89 | .94 | .96 |
| 7 | .52 | .72 | .83 | .92 | .96 | .98 |
| 8 | .55 | .75 | .86 | .94 | .98 | .99 |
| 9 | .57 | .78 | .88 | .96 | .98 | .99 |
| 10 | .59 | .80 | .90 | .97 | .99 | .997 |
| 16 | .69 | .89 | .96 | .99 | .999 | >.999 |
| 24 | .76 | .95 | .99 | .999 | >.999 | >.999 |
| 48 | .87 | .99 | .999 | >.999 | >.999 | >.999 |
| 96 | .95 | .999 | >.999 | >.999 | >.999 | >.999 |
| 192 | .99 | >.999 | >.999 | >.999 | >.999 | >.999 |

billion bp that vary in at least 1% of the world's population (Box 1). Because most mapping strategies require SNPs with alleles of higher frequency, the expected total number of SNPs for different minimal allele frequencies is given in Table 1.

## Putting SNPs on the map

So far, experimental validation of the mapped SNPs has been limited to a very small subset. Of these, 95% were verified to be polymorphic in the original discovery sample of 24 individuals, and 82% were found to be polymorphic in a new sample by a method insensitive to low-frequency alleles[3]. The fraction of candidate SNPs that are true polymorphisms with a minimal allele frequency of 1% probably lies between these two rates, leading to an estimate of 1.16–1.35 million such sites. Thus, the reported collection is expected to comprise roughly 11%–12% of human polymorphic nucleotide variation (Table 1). A discovery effort based on a small number of haploid genomes (here mostly two) should detect a higher fraction of common SNPs. For example, 60% or 700,000–800,000 SNPs in the catalog should have a minimal allele frequency of 20%, representing 21%–25% of all such SNPs in the genome (Table 1).

What accounts for the incompleteness of SNP discovery? There are three primary factors. First, SNPs can be assigned to unique locations only in the 2.7 billion bp (84%) of the genome for which draft sequence has been assembled[5]. Second, the entire 2.7 billion bp has not been screened for polymorphism. The SNP discovery rate was 1 per 1,331 bp in a 1.5 billion bp subset of high-quality sequence covered by two haploid genomes, as compared with 1 per 1,900 bp in the assembly as a whole[3]. So, only 70% of the assembly has been assayed for polymorphisms. A more important factor is that for most of the sequence, only two haploid genomes have been compared. Such a comparison, even if error-free and extended to the entire genome, detects only a fraction of all SNPs—the roughly 20% that happen to differ between the two particular haploid genomes (Box 1 and Table 2). In principle, 2.4 million SNPs could have been discovered using this strategy; the current collection contains 50%–60% of this number.

## What would it take to complete the SNP catalog?

A future goal of the SNP program is "to obtain a nearly complete catalogue of common variants"[5]. The difficulty of this task depends on the meanings of "nearly complete" and "common." Table 2 shows the fraction of all SNPs at or above a given allele frequency that can be detected in a comparison of $n$ haploid genomes. A screen with $n$=6 will detect over 95% of SNPs with minimal allele frequency of 40%, and a screen with $n$=8 will detect 99%. In contrast, detection of 95% of 'polymorphic' sites (minimal allele frequency of 1%) would require $n$=96, and detection of 99% would require $n$=192.

Such deep surveys may only make sense for regions of the genome in which polymorphisms would be expected to alter gene function through amino-acid substitution or other mechanisms, and thereby contribute to phenotypic variation. It is worth noting that polymorphisms with functional consequences are expected to have lower allele frequencies and, in fact, the majority of coding region SNPs (cSNPs) that change an amino acid have allele frequencies below 5% (refs. 6,7). For variants below a frequency of 1%, which may nonetheless have an important role in human health and variation, centralized cataloging efforts may be less effective than direct mutation screening in samples with relevant phenotypes. This is particularly true because rare variants are more likely to be restricted to specific ethnic groups[6,7].

## Putting the map to work

The current catalog makes it clear that our ability to find SNPs already outstrips our capacity to genotype them by at least 1,000-fold. Although many approaches for SNP genotyping have emerged over the past

### Box 1 • Behind the numbers

The standard neutral model of population genetics assumes a random-mating population of constant size, with all mutations uniquely arising and selectively neutral[4]. Using the Ewens sampling formula[22] for this model, it can be shown that the number of sites $S_f$ with both alleles at frequency $\geq f$ in the population is related to the number of sites $S_2$ that vary between two haploid genomes by approximately

$$S_f = S_2 \ln\left(\frac{1-f}{f}\right)$$

This relationship was used to compute the numbers in the second and third columns of Table 1. To compute the last column, we note that an easy consequence of the Ewens sampling formula is that when two haploid genomes are compared to find SNPs, the distribution of the minor allele frequency of discovered SNPs is uniform. That is, the fraction of discovered SNPs that have both alleles at frequency $\geq f$ in the population is $1–2f$. The SNP detection rates in Table 2 are readily computed by combining the SNP allele frequency spectrum (which can be derived from $S_f$) with simple binomial sampling statistics[23].

The standard neutral model, although highly useful for first-order estimates, makes a number of simplifying assumptions. Deviations from the model and complications posed by natural selection and human demographic history (including population subdivision) are discussed in ref. 4 A detailed mathematical treatment, which includes the effects of population expansion and arrives at similar numbers, can be found in ref. 24.

decade[8], even the most parallel systems—ordered oligonucleotide arrays—have been tested with only a few hundred SNPs (ref. 9). Comprehensive applications of the map clearly await an increase in parallelism of SNP genotyping using arrays or other strategies, elimination of the requirement for PCR amplification of each locus (perhaps using strategies involving reduced representation or whole-genome amplification) and a radical decrease in the cost of these analyses.

What is the likely scale of future genotyping efforts? Simply validating and determining the population frequency of SNPs in the current catalog in different ethnic populations, as well as linkage disequilibrium relationships between SNPs, will entail typing the 1.4 million candidate polymorphisms in hundreds of samples. The resulting linkage disequilibrium map of the human genome would lay a foundation for future genotype–phenotype exploration[10]. Studies of complex phenotypic variation will involve thousands or tens of thousands of individuals.

How many SNPs will need to be typed? For direct association studies, the relevant number is that of SNPs that alter gene function. Previous studies have found about four cSNPs per gene[6,7]. This observation, together with the new estimates of roughly 30,000 genes in the human genome[5,11,12], indicates the existence of a total of approximately 120,000 cSNPs. Of these, 40% are expected to change an amino acid[6,7]. These 50,000 non-synonymous cSNPs, together with an unknown number of regulatory and other non-coding but functional polymorphisms, comprise the bulk of common molecular variation with potential phenotypic consequences (the fraction of all cSNPs represented in the current catalog is difficult to estimate but probably lies between 15% and 30%). In the future, genotyping this complete set of functional variants will be a minimal requirement for direct association studies[1,13,14].

The current map is clearly geared toward indirect association studies based on linkage disequilibrium. Although the number of SNPs required for such studies has yet to be settled definitively, even low-end estimates are in the range of 50,000–100,000 SNPs (refs. 15,16; at the high end, all the common SNPs on the current map will be pressed into service[17,18]). Thus, any whole-genome association studies will demand assays of on the order of $10^5$–$10^6$ polymorphisms, with accurate determination of $10^8$–$10^{10}$ individual genotypes. Although SNP genotyping of pooled individual samples is made attractive by the decrease in the number of genotypes, this strategy has important drawbacks. Pooling loses statistical power, particularly for rarer alleles. Haplotypes at multiple loci cannot be resolved, precluding some powerful mapping strategies. Finally, clinical samples are less readily stratified by phenotypic differences and environmental factors, and such analyses may be key to understanding disease susceptibility.

With time and sufficient demand, most of the technical hurdles we now face will fall away, and we can come to grips with the underlying biological complexities[19–21]. Already, the new catalog allows us to contemplate studying the entire range of common molecular variation present in our species, and to test whether these common polymorphisms account for human phenotypic diversity. Genetics is the study of variation, and the prospect of carrying it to this level of resolution is a heady one. □


1. Collins, F.S., Guyer, M.S. & Chakravarti, A. *Science* **278**, 1580–1581 (1997).
2. Collins, F.S. *et al. Science* **282**, 682–689 (1998).
3. The International SNP Map Working Group *Nature* **409**, 928–933 (2001).
4. Przeworski, M., Hudson, R.R. & Di Rienzo, A. *Trends Genet.* **16**, 296–302 (2000).
5. International Human Genome Sequencing Consortium *Nature* **409**, 860–921 (2001).
6. Cargill, M. *et al. Nature Genet.* **22**, 231–238 (1999).
7. Halushka, M.K. *et al. Nature Genet.* **22**, 239–247 (1999).
8. Landegren, U., Nilsson, M. & Kwok, P.-Y. *Genome Res.* **8**, 769–776 (1998).
9. Wang, D.G. *et al. Science* **280**, 1077–1082 (1998).
10. Kruglyak, L. *Proc. Natl. Acad. Sci. USA* **96**, 1170–1172 (1999).
11. Ewing, B. & Green, P. *Nature Genet.* **25**, 232–234 (2000).
12. Roest Crollius, H. *et al. Nature Genet.* **25**, 235–238 (2000).
13. Lander, E.S. *Science* **274**, 536–539 (1996).
14. Risch, N. & Merikangas, K. *Science* **273**, 1516–1517 (1996).
15. Boehnke, M. *Nature Genet.* **25**, 246–247 (2000).
16. Collins, A., Lonjou, C. & Morton, N.E. *Proc. Natl. Acad. Sci. USA* **96**, 15173–15177 (1999).
17. Dunning, A.M. *et al. Am. J. Hum. Genet.* **67**, 1544–1554 (2000).
18. Kruglyak, L. *Nature Genet.* **22**, 139–144 (1999).
19. Weiss, K.M. & Terwilliger, J.D. *Nature Genet.* **26**, 151–157 (2000).
20. Altshuler, D., Daly, M. & Kruglyak, L. *Nature Genet.* **26**, 135–137 (2000).
21. Horikawa, Y. *et al. Nature Genet.* **26**, 163–175 (2000).
22. Ewens, W.J. *Theor. Popul. Biol.* **3**, 87–112 (1972).
23. Eberle, M.A. & Kruglyak, L. *Genet. Epidemiol.* **19**, S29–35 (2000).
24. Durrett, R. & Limic, V. *Stoch. Processes Appl.* (in press).


# γ-H2AX illuminates meiosis


Neil Hunter[1], G. Valentin Börner[1], Michael Lichten[2] & Nancy Kleckner[1]

[1]*Molecular and Cellular Biology, Harvard University, Cambridge, Massachusetts 02138, USA.* [2]*Laboratory of Biochemistry, Division of Basic Science, National Cancer Institute, Bethesda, Maryland 20892-4255, USA. e-mail: nhunter@fas.harvard.edu*



**The temporal and functional relationships between the DNA events of meiotic recombination and the synaptonemal complex (SC), a meiosis-specific structure formed between homolog axes, are subjects of intense discussion and investigation. A new study provides evidence that initiation of recombination (through programmed double-strand breaks (DSBs)) precedes initiation of SC formation, and further suggests that progression of recombination is required for formation of SC on a region-by-region basis. These conclusions derive from immunocytological analysis of a phosphorylated histone variant, γ-H2AX, previously found to be characteristic of DSB repair in mitotic cells, and shown here to be recruited for specialized use during meiosis.**


During meiosis, chromosome replication is followed by an extended prophase devoted to interhomolog interactions that culminate in formation of one or a few selectively placed chiasmata. Chiasmata correspond to crossovers at both DNA and structural levels, and provide the connections that permit accurate disjunction of homologous chromosomes at the first meiotic division[1]. During prophase, homologs become progressively more closely juxtaposed until they comprise a single morphological unit. Chromosomes then become diffuse and, when they remerge as individualized units, homologs are connected only by their chiasmata. Interhomolog interactions include recombination-independent recognition (analogous to somatic pairing), the biochemical events of DNA recombination, and a dramatic series of cytologically visible structural changes