Human Molecular Genetics, 2001, Vol. 10, No. 20 **2195–2198**

The birth and death of human single-nucleotide polymorphisms: new experimental evidence and implications for human history and medicine

Raymond D. Miller and Pui-Yan Kwok*

Division of Dermatology, Washington University School of Medicine, St Louis, MO 63110, USA

Received June 21, 2001; Accepted July 16, 2001

Extensive, new databases of single-nucleotide polymorphisms (SNPs) provide a powerful resource for disease gene discovery, and they will be even more useful as more frequency data become available. Interesting observed genomic patterns include SNP deserts (regions of low SNP incidence) and lengthy regions of linkage disequilibrium containing only a few haplotypes. A variety of genetic studies will benefit from SNP resources.

The desire to test the promising approach of studying complex traits by genetic association has spurred the recent explosion in discovery of human single-nucleotide polymorphisms (SNPs). For Mendelian diseases with high penetrance due to mutations in a single gene, pedigree studies have been very successful in mapping the disease loci. However, for complex diseases, those caused by multiple disease genes of modest effect, the number of families required is not practical and the results obtained for the families are not additive. In theory, association studies comparing the prevalence of a set of markers in a moderate number of affected and unaffected individuals, could localize high frequency, complex disease alleles due to underlying non-random association in short stretches of the genome, if enough genetic markers were available (1). A major motivation to find human SNPs and record them in public databases was to provide freely available markers for genome-wide association studies (2).

Major, successful SNP search efforts were mounted, and the past 2 years have truly been an era of SNP discovery and seen the birth of a new SNP collection. As part of the publication of the draft human genome sequence in February, 2001, 1.42 million SNPs were reported (1.69 million as of June, 2001) (3). These are available in several databases including dbSNP at NCBI (http: //www.ncbi.nlm.nih.gov/SNP/) and ENSEMBL (http://www.ensembl.org). The SNPs designated 'rs' (for reference SNP) in dbSNP represent a non-redundant set, since many have been found by more than one strategy. One highly efficient discovery method was to compare DNA sequences from overlapping clones sequenced for the human genome project, using statistical methods to identify candidate SNPs versus sequencing noise. If the clones were derived from the same chromosome, there should be no differences, but if derived from maternal and paternal chromosomes or from different individuals, one finds SNPs at the rate of approximately 1 per 1000 bp (4,5). The other major effort was lead by The SNP Consortium (TSC), a group formed by the Wellcome Trust and a number of companies. The TSC funded academic groups to sequence 'reduced representation' genomic DNA libraries and compare the reads with each other and with the draft genomic sequence to find SNPs (6,7). The TSC SNPs cover the genome with an average spacing of 1 SNP per 3000 bp. The combined TSC and overlap SNPs cover the human genome at an average spacing of 1 SNP per 2000 bp. Based on the coverage of these SNPs on well-characterized exons, it was estimated that 85% of all exons are within 5 kb of at least one of the SNPs in this collection (3).

While the SNP collection is a major resource, it is not the final product that is needed to map complex diseases. Indeed, it is useful to refer to the SNPs as candidate SNPs because for most nothing is known about the frequency in populations. Several groups, including ours, have begun characterization of the candidate SNPs. A pilot study showed that ~17% of the candidate SNPs in the collection had no detectable variation in any of three major populations in the US (African-Americans, Asians and Caucasians) and would not be useful for genetic studies. The majority of the 'monomorphic' candidates represent private SNPs, i.e. real variants in the discovery DNAs, but with no appreciable population frequency. About 6% are uncommon SNPs (rare allele detected but always <20% in any population), ~53% of candidates are common SNPs (rare allele ≥20%) in any one population and ~27% of SNP candidates are common in all three populations (8). Any investigator using the SNP database should bear these distributions in mind.

Concurrent with SNP discovery, information and questions about SNPs have grown, as well as technology to type SNPs (9). In this review, we consider the life cycle of SNPs, the incidence of SNPs within the genome, the connection between patterns observed in SNPs and human history, and SNP haplotypes.

LIFE CYCLE OF A SNP

The 'life cycle' of an SNP can be divided into four phases: (i) appearance of a new variant allele by nucleotide mutation; (ii) survival against odds of the allele through early generations; (iii) increase to substantial frequency including survival

*To whom correspondence should be addressed. Tel: +1 314 362 8236; Fax: +1 314 362 8159; Email: kwok@genetics.wustl.edu

through population fluctuations; and (iv) fixation. We discuss these in order.

Despite amazingly efficient DNA repair mechanisms (10), nucleotide mutation creates both normal variation and disease alleles in humans. Regardless of the method one uses to find variation in humans, one has to compare DNA sequences derived from different chromosomes. Typically, a common SNP (one with the minor allele frequency of >20%) is found about every 1000 bp. In humans, all combinations of substitution polymorphisms are observed, with A/G substitution SNPs (including reverse complement T/C) being the most prevalent (11,12).

The hypothesis that variation between species is an extension of variation within species is a long-standing one in population genetics (13). Our group tested this hypothesis for humans by comparing the types of SNPs found while scanning on the long arm of the human X chromosome, with the divergence in similarly sequenced DNA from a Sumatran orangutan. We found that the hypothesis fits quite well, both for regions of high and low GC content (12). The 2.9% divergence provided great statistical power for estimating the mutational spectrum in humans (and orangutans). The spectrum of SNP types was nearly the same in regions of high and low GC content, although regional subtleties in the mutation spectrum must ultimately give rise to the substantial differences observed in GC content. Due to the generality of the mutation process, the spectrum is also an estimate of kinds of disease-causing mutations.

The most frequent mutation in humans is the mutation from methyl CpG to TpG (a transition accounting for 25% of all mutations) (4,14). This molecular mechanism must be a major cause of the great deficiency of CG dinucleotides observed in genomic sequence since many will eventually become TG, whereas new CpG sites will be created by other, less frequent mutations. Since the genome contains only a few percent coding sequence (15), the vast majority of SNPs are likely to have little functional consequence. However, many missense coding mutations are expected to be deleterious, the majority causing destabilization of the folded protein (16).

The second phase of the SNP life cycle we call 'survival against odds', because most new mutations are expected to be lost in early generations due to sampling. For example, if a person heterozygous for a new, selectively neutral, autosomal mutation has two children, the probability the mutation will be found in at least one of the children is 0.75. As long as the mutant allele copies are carried in the population by heterozygotes each having two children, the probability the new mutation will be lost is $1 - (0.75)^g$, where g = generations, with the mutation occurring in g = 0. For example, there is a 94% probability of loss in 10 generations (200 years if one assumes 20 years per generation) and even if each person has three children, a 74% probability of loss in the same time.

If a mutation survives early generations and increases in frequency until it becomes homozygous in some individuals, the potentially lengthy third phase is entered and the risk of loss is reduced. The frequency of a new allele is expected to show variation due to stochastic processes and hitchhiking, but severe bottlenecks in population size will tend to favor survival of the highest frequency alleles and loss of rare alleles.

In the fourth phase of the lifecycle, a few SNPs will eventually cease to be SNPs because the 'new' allele will reach 100% in the population and become a difference between humans and a closely related species. In regions of both high and low GC content in the X chromosome, we found that 2.9% of nucleotide sites differ between humans and the Sumatran orangutan. We found no evidence of shared SNPs between the species, i.e. the lifetime of SNPs is shorter that the divergence time of humans and orangutans. Based on the divergence and variability within humans, we estimated the lifetime of an SNP destined to become fixed for a new allele to be 284 000 years (12).

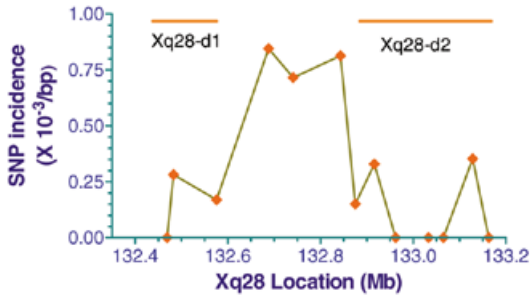

Figure 1. Two SNP deserts in Caucasians. Common SNPs were detected by scanning 65 STS, and the incidence was averaged for STS within ~50 kb intervals (12). The labeled bars at the top show the extent of the deserts. The region is contained within contig NT_011597.3 in Xq28, and the map distances from the telomer are based upon NCBI build 22.

VARIATION IN SNP INCIDENCE

The incidence of SNPs at sites causing coding changes is reduced compared with silent sites, probably due to selection against deleterious alleles (17–19). For the non-coding genome, an initial working hypothesis was that the incidence of SNPs should be approximately constant, but this working hypothesis is clearly not true. The species-wide value of π, the probability that a nucleotide position is heterozygous when two chromosomes are compared, was estimated as 7.65×10^{-4} for autosomes. In contrast, the values were much lower for the X (4.69×10^{-4}) and Y chromosome (1.51×10^{-4}). As part of the explanation for the lower values in X and Y, it is noted that the effective population sizes are lower than for autosomes and the population of X chromosomes spends more time in females, where the mutation rate is lower, than in males. Regions of highest GC content had ~15% higher heterozygosity than regions of lowest GC content, possibly because of the higher incidence of CpG sites in the former. In analyzing 200 kb windows for diversity, the variation was larger than expected by simple models, and 12 regions were identified that contained no SNP candidates (3).

To identify SNPs on the long arm of the X chromosome, our group scanned for SNPs in CEPH families. While we were successful in constructing a map of common SNPs (20), we had considerable difficulty finding SNPs in some regions. Further investigation showed that the difficulty in finding SNPs was due to long regions of very low SNP density, which we called 'SNP deserts'. We estimated the SNP deserts represented 28% of the investigated sequence, and the longest one was 1.66 Mb in length. Details of two deserts in Xq28 are shown (Fig. 1). Comparative sequence of these regions in the orangutan showed that the deserts were not due to regionally

altered mutation rates. The best explanation for an SNP desert is that the region had a recent, single ancestral sequence (recent coalescent event). Based on observed variation in some deserts, they had coalescent times of ~17 000–39 000 years ago. Part of the explanation for the observation that there is variation in variation is that the human genome (at least the X chromosome in European-derived populations) is peppered with recent coalescent events (12).

SNPS AND HUMAN HISTORY

Our SNPs are a product of the past and surely may provide clues about human history. In the last 500 years, there has been a massive increase in human populations and very substantial geographic migrations. Recent population increases should be reflected in the incidence of private SNPs. Recent migrations often lead to admixture; for example, European ancestry in African-Americans living in South Carolina ranging from 3.5% (Gullah Sea Islanders) to 17.7% (Columbia residents) (21). Earlier modern humans can be thought of as three major groups centered in Africa, Asia (including immigrants to the South Sea Islands, Australia and the Americas) and Europe. Two major, unresolved questions are as follows. (i) What was the relationship among the three groups? (ii) Did any group experience large population changes that may have affected their genetic composition?

LINKAGE DISEQUILIBRIUM AND HAPLOTYPES

As discussed, a major motivation for SNP discovery has been the potential search for complex disease loci using association studies. These studies assume there is a correlation between the disease mutation and some marker SNP. Often this is called linkage disequilibrium (LD) and is measured by 'D' or a number of other measures (22).

Based on certain assumptions about human history, a recent paper predicted that LD would be limited to ~3 kb and, consequently, a minimum of 500 000 SNPs would be required for a whole-genome association study (23). In contrast, our group found two large blocks of LD (one extending 1 Mb) present on Xq25 and Xq28 in three European populations of very different histories, together with small regions lacking LD. Each of the blocks of LD consisted of a small number of high-frequency haplotypes (24). Additional studies found similar results (25–30), including a systematic study of LD beginning at a starter site in 19 genes in a European-derived population from Utah, which estimated LD on average extends 60 kb (≥50% of maximal possible value). However, many genes showed an abrupt decrease in LD at some point, and there was much variation between genes. Similar patterns were observed in a population from Sweden. Comparative study of the Yoruban population from Nigeria showed that LD was less than in the other populations, but the same haplotypes were in high frequency (27).

In theory, the breakdown of LD over time is related to the recombination value between the two loci. A correlation between the rank in extent of LD and the local recombination rate has been shown (27). Less clear is what may have created the LD now observed. Since an initial mutation must occur on a specific haplotype, there will be LD between that mutation and nearby SNPs. However, common SNPs should be long past this 'birth' LD. LD can also be caused by admixture between very different populations or by severe restriction in population size that reduce the number of haplotypes. The LD data of the Utah samples fit a model of severe restriction in population size (e.g. 50 people for 40 generations) in Europe ~27 000–53 000 years ago, aided possibly by hotspots of recombination every ~60 kb (27). A severe restriction in population size would explain the recent coalescent events observed on the X chromosome (12). Better understanding major demographic events in our history will aid understanding of patterns of variation in humans.

Three kinds of genetic resolution are now available in humans: (i) meiotic recombination in pedigree studies; (ii) admixture disequilibrium in key populations; and (iii) association mapping. Admixture mapping has great potential for whole-genome scans for disease genes using few SNPs (31). Results from LD studies suggest the block hypothesis for the human genome: the genome consists of a series of blocks, each of which is predominantly made up of a very few high-frequency haplotypes, and short regions with many haplotypes separate the blocks. Identifying the extent and allelic composition of each block would greatly aid the search for complex disease genes.

REFERENCES

1. Risch, N. and Merikangas, K. (1996) The future of genetic studies of complex human diseases. *Science*, **273**, 1516–1517.
2. Collins, F.S., Guyer, M.S. and Chakravarti, A. (1997) Variations on a theme: cataloging human DNA sequence variation. *Science*, **278**, 1580–1581.
3. The International SNP Map Working Group (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature*, **409**, 928–933.
4. Taillon-Miller, P., Gu, Z.J., Li, Q., Hillier, L. and Kwok, P.Y. (1998) Overlapping genomic sequences – a treasure trove of single-nucleotide polymorphisms. *Genome Res.*, **8**, 748–754.
5. Marth, G.T., Korf, I., Yandell, M.D., Yeh, R.T., Gu, Z., Zakeri, H., Stitziel, N.O., Hillier, L., Kwok, P.Y. and Gish, W.R. (1999) A general approach to single-nucleotide polymorphism discovery. *Nat. Genet.*, **23**, 452–456.
6. Mullikin, J.C., Hunt, S.E., Cole, C.G., Mortimore, B.J., Rice, C.M., Burton, J., Matthews, L.H., Pavitt, R., Plumb, R.W., Sims, S.K. *et al.* (2000) An SNP map of human chromosome 22. *Nature*, **407**, 516–520.
7. Altshuler, D., Pollara, V.J., Cowles, C.R., Van Etten, W.J., Baldwin, J., Linton, L. and Lander, E.S. (2000) An SNP map of the human genome generated by reduced representation shotgun sequencing. *Nature*, **407**, 513–516.
8. Marth, G., Yeh, R., Minton, M., Donaldson, R., Li, Q., Duan, S., Davenport, R., Miller, R.D. and Kwok, P.Y. (2001) Single-nucleotide polymorphisms in the public domain: how useful are they? *Nat. Genet.*, **27**, 371–372.
9. Kwok, P.Y. (2000) High-throughput genotyping assay approaches. *Pharmacogenomics*, **1**, 95–100.
10. Wood, R.D., Mitchell, M., Sgouros, J. and Lindahl, T. (2001) Human DNA repair genes. *Science*, **291**, 1284–1289.
11. Taillon-Miller, P., Piernot, E.E. and Kwok, P.Y. (1999) Efficient approach to unique single-nucleotide polymorphism discovery. *Genome Res.*, **9**, 499–505.
12. Miller, R.D., Taillon-Miller, P. and Kwok, P.Y. (2001) Regions of low single-nucleotide polymorphism incidence in human and orangutan Xq: deserts and recent coalescences. *Genomics*, **71**, 78–88.
13. Dobzhansky, T.G. (1970) *Genetics of the Evolutionary Process*. Columbia University Press, New York, pp. 1–505.
14. Wang, D.G., Fan, J.B., Siao, C.J., Berno, A., Young, P., Sapolsky, R., Ghandour, G., Perkins, N., Winchester, E., Spencer, J., *et al.* (1998) Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science*, **280**, 1077–1082.

15. Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W. *et al*. (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
16. Wang, Z. and Moult, J. (2001) SNPs, protein structure, and disease. *Hum. Mutat.*, **17**, 263–270.
17. Cargill, M., Altshuler, D., Ireland, J., Sklar, P., Ardlie, K., Patil, N., Shaw, N., Lane, C.R., Lim, E.P., Kalyanaraman, N. *et al*. (1999) Characterization of single-nucleotide polymorphisms in coding regions of human genes. [Published erratum appears in *Nat. Genet.* (1999) **23**, 373]. *Nat. Genet.*, **22**, 231–238.
18. Chakravarti, A. (1999). Population genetics—making sense out of sequence. *Nat. Genet.*, **21**, 56–60.
19. Halushka, M.K., Fan, J.B., Bentley, K., Hsie, L., Shen, N., Weder, A., Cooper, R., Lipshutz, R. and Chakravarti, A. (1999) Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. *Nat. Genet.*, **22**, 239–247.
20. Taillon-Miller, P. and Kwok, P.Y. (2000) A high-density single-nucleotide polymorphism map of Xq25–q28. *Genomics*, **65**, 195–202.
21. Parra, E.J., Kittles, R.A., Argyropoulos, G., Pfaff, C.L., Hiester, K., Bonilla, C., Sylvester, N., Parrish-Gause, D., Garvey, W.T., Jin, L. *et al*. (2001) Ancestral proportions and admixture dynamics in geographically defined African Americans living in South Carolina. *Am. J. Phys. Anthropol.*, **114**, 18–29.
22. Morton, N.E., Zhang, W., Taillon-Miller, P., Ennis, S., Kwok, P.Y. and Collins, A. (2001) The optimal measure of allelic association. *Proc. Natl Acad. Sci. USA*, **98**, 5217–5221.
23. Kruglyak, L. (1999) Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nat. Genet.*, **22**, 139–144.
24. Taillon-Miller, P., Bauer-Sardina, I., Saccone, N.L., Putzel, J., Laitinen, T., Cao, A., Kere, J., Pilia, G., Rice, J.P. and Kwok, P.Y. (2000) Juxtaposed regions of extensive and minimal linkage disequilibrium in human Xq25 and Xq28. *Nat. Genet.*, **25**, 324–328.
25. Abecasis, G.R., Noguchi, E., Heinzmann, A., Traherne, J.A., Bhattacharyya, S., Leaves, N.I., Anderson, G.G., Zhang, Y., Lench, N.J., Carey, A. *et al*. (2001). Extent and distribution of linkage disequilibrium in three genomic regions. *Am. J. Hum. Genet.*, **68**, 191–197.
26. Mateu, E., Calafell, F., Lao, O., Bonne-Tamir, B., Kidd, J.R., Pakstis, A., Kidd, K.K. and Bertranpetit, J. (2001) Worldwide genetic analysis of the CFTR region. *Am. J. Hum. Genet.*, **68**, 103–117.
27. Reich, D.E., Cargill, M., Bolk, S., Ireland, J., Sabeti, P.C., Richter, D.J., Lavery, T., Kouyoumjian, R., Farhadian, S.F., Ward, R. *et al*. (2001) Linkage disequilibrium in the human genome. *Nature*, **411**, 199–204.
28. Moffatt, M.F., Traherne, J.A., Abecasis, G.R. and Cookson, W.O. (2000) Single nucleotide polymorphism and linkage disequilibrium within the TCR α/δ locus. *Hum. Mol. Genet.*, **9**, 1011–1019.
29. Bonnen, P.E., Story, M.D., Ashorn, C.L., Buchholz, T.A., Weil, M.M. and Nelson, D.L. (2000) Haplotypes at ATM identify coding-sequence variation and indicate a region of extensive linkage disequilibrium. *Am. J. Hum. Genet.*, **67**, 1437–1451.
30. Kidd, J.R., Pakstis, A.J., Zhao, H., Lu, R.B., Okonofua, F.E., Odunsi, A., Grigorenko, E., Tamir, B.B., Friedlaender, J., Schulz, L.O. *et al*. (2000) Haplotypes and linkage disequilibrium at the phenylalanine hydroxylase locus, PAH, in a global representation of populations. *Am. J. Hum. Genet.*, **66**, 1882–1899.
31. Pfaff, C.L., Parra, E.J., Bonilla, C., Hiester, K., McKeigue, P.M., Kamboh, M.I., Hutchinson, R.G., Ferrell, R.E., Boerwinkle, E. and Shriver, M.D. (2001) Population structure in admixed populations: effect of admixture dynamics on the pattern of linkage disequilibrium. *Am. J. Hum. Genet.*, **68**, 198–207.