nature genetics

Genome variation for non-geneticists

Single-nucleotide variation (SNPs or SNVs) in the human genome is now being used by the public and by researchers interested in the functional mechanisms of genetic perturbation for the 3D structure and function of the nucleus in various cells and tissues, and for understanding human–microbiota interactions. We have some requests for authors that may help prevent misunderstanding as familiar genetic markers acquire new users.

Adults differ in their ability to digest milk in part because of genetic differences, but habitual consumption of yoghurt and other fermented milk products also varies, raising interest in the ways that our commensal bacteria help or hinder diet and nutrition. A regulatory variant, in an intron of the *MCM6* gene, was found to be associated with lactase non-persistence—popularly known as lactose intolerance—via a long-range interaction with the neighboring lactase-encoding gene, *LCT* (*Nat. Genet.* **30**, 233–237, 2002). As this variant has recently become of interest as a possible constitutive host factor influencing the composition of the human gut microbiota (pp 1301, 1396 and 1407), we have used it to illustrate some of the potential pitfalls and rather indigestible terminology for useful genetic markers in interdisciplinary research.

We support NCBI's dbSNP (https://www.ncbi.nlm.nih.gov/projects/SNP/) as the recognized community database for single-nucleotide variation. Its rsID identifiers label little exemplar sequences and can be used to index millions of mostly biallelic single-nucleotide substitutions at their complementary positions in the human genome, much like a genomic slide rule. So please call single-nucleotide variant loci by their dbSNP rsIDs. For example, the lactase variant locus introduced above was originally termed *LCT* C/T-13910 and is now known as rs4988235(G>A). In describing the experimental detection or effects of a variant, use the locus identifier together with the allele or genotype to which you are referring: rs4988235[G] or rs4988235[G/A], respectively. The dbSNP database could be improved for a wider range of users in two respects. First, non-human sequences could be put in a separate database, and, second, in listing alternate alleles, the "/" character could be replaced with the Human Genome Variation Society (HGVS) convention using ">", as there is a previous convention in genetics to use a solidus to separate genotypes on the same chromosome.

Please introduce variants that are the subject of major conclusions in your research at first mention or in a supplementary table by their most recent HGVS names: for example, rs4988235(G>A) is NC_000002.12:g.135851076G>A). This HGVS name consists of an exemplar sequence, NC_000002.12, of chromosome 2 in human genome build GRCh38.p7 followed by the genomic signifier ":g.", the position of the variant nucleotide (135851076), and the conventional and alternate alleles (G>A). Equally useful, NC_000002.11:g.136608646G>A corresponds to a chromosome 2 assembly from the genome sequence GRCh37.p13. We also support the use of this name because GRCh37.p13 (build 37) is the source sequence for many genotyping platforms still in use. We think that the wider adoption of chromosomal coordinates for genetic variants will enable research into the cell-specific effects of variants on chromatin interactions and gene expression, without introducing bias as to whether the variants have effects on one or more annotated genes.

rs4988235 has several other synonymous HGVS names in dbSNP, using a range of exemplar sequences. Some of these names are based on subchromosomal assemblies, and some are based on sequences assembled from transcripts. NG_008104.2:g.9094C>T contains a 'locus reference gene', LRG_338, constructed specifically for those most interested in the *LCT* gene. NG_008958.1:g.30366C>T contains a RefSeqGene for the neighboring *MCM6* gene in which the rs4988235 regulatory variant is embedded, and NM_005915.5:c.1917+326C>T has a cDNA sequence for *MCM6* that is a poorer exemplar, as the regulatory variant is in an intron of the *MCM6* gene rather than the actual cDNA sequence, requiring some odd arithmetic.

In GWAS of categorical traits, it is essential to identify not only the variant locus and odds ratio associated with the trait but also the directions of effect for each pair of alleles under comparison (relative risk or relative protection). Likewise, with quantitative traits, it is key to identify the effect and comparator alleles. Not only the rsID sequence but also the origin, age and evolution of variants can require carefully specified context to avoid introducing misinformation. The population under study and its allele frequencies are relevant to claim that an allele is a major or minor allele. For instance, rs4988235[A] is the major allele (70–80% frequency) in northern Europe, but has a frequency of 5–10% in southern Europe (*Nat. Genet.* **37**, 868–872, 2005). From comparison with primate genomes, rs4988235[G] is an ancestral allele tagging an ancestral haplotype (*Am. J. Hum. Genet.* **81**, 615–625, 2007), with the derived lactase persistence (A) allele having arisen recurrently wherever people kept milk cows.

In this editorial we have tried to avoid overusing the word 'reference' because it introduces many ambiguities into the discussion of SNPs and their uses. ■