European Journal of Human Genetics (2005) 13, 265–270

www.nature.com/ejhg

NEWS AND COMMENTARIES

Genomics

The human genome, revisited

Gert-Jan B van Ommen

European Journal of Human Genetics (2005) **13,** 265–267.
doi:10.1038/sj.ejhg.5201348
Published online 8 December 2004

In a recent issue of Nature, the landmark paper has appeared that describes and discusses the finished version ('build 35') of the human genome.[1]

This International Human Genome Sequencing Consortium's new build of the genome sequence is as finished as current technology allows, which is a major improvement on the initial draft sequence. The finished version of the genome sequence covers 2.85 Gbp or ~99% of the euchromatic region, with only 341 gaps, compared to the 150 000 in the working draft. It has a long-range continuity of 39.5 Mb, which is 475 times better than the 81 kb in the working draft, and the sequence has been delivered at an error rate of 1×10^{-5}: almost 10 times better than the quality standard of 10^{-4} originally aimed for.

The authors of this landmark paper emphasize the extremely rigorous quality control applied to achieve and validate this result. And rightly so, this sequence will be the raw material upon which future generations of scientists and healthcare workers will base studies that depend on highly reliable data. To this aim, ~40 Mb was independently resequenced, and an overlap study was carried out on 4235 clones from one large insert library. As hoped for, the latter showed a bimodal distribution of base differences, consistent with half of these coming from one haplotype and half from the other. These results confirm that the sequencing error rate is 20–100 times lower than the human polymorphism rate. This confirms the finished human genome sequence as a robust resource for large-scale evolutionary, functional and comparative analyses. The Consortium also resequenced 750 000 clones (8× coverage) from an independent fosmid library to validate over 97% of the junctions of the large insert clones. This analysis also suggests the presence of 50–100 erroneous deletions (average ~5 kb).

Of the 341 gaps in the sequence, 308 encompass 28 Mb or ~1% of the euchromatic DNA and 33 cover ~200 Mb of heterochromatic DNA. Interestingly, the euchromatic gaps are mostly adjacent to segmental duplications and strongly cluster near centromeres and telomeres: the ~3 Mb of DNA that flank each chromosome arm, and which in total add up to only 4.7% of the euchromatic DNA, contain no less than 41% of the gaps, chromosome 9 leading with 0.3% DNA and 13.3% gaps. In terms of gene loci, a reassuring 99.74% of the published cDNA sequence (925 Mb) was identified in the finished genomic sequence. Partially (0.23%) and completely (0.06%) missing cDNAs are mostly located near segmental duplications: clearly, these regions are the biggest stumbling blocks to an even more 'finished' sequence.[2] Several approaches are discussed to address the gaps, but by and large these have to become the subject of targeted, focused research efforts rather than brute-force high-throughput technology. This issue is more specifically addressed in a parallel paper from Eichler and collaborators.[3]

Owing to the emphasis on thoroughness, validation and QC, the flavor of the paper is rather more technical than the mainly biology-oriented working draft publication, which was therefore, unsurprisingly, more electrifying. However, additional annotation can be found on the accompanying poster, in on-line supplementary information, and on genome browsers like http://genome.ucsc.edu/; http://www.ensembl.org/; and http://www.ncbi.nlm.nih.gov/genome/guide/human/. In addition, rather than reiterating points made in the draft paper, several more advanced biological issues that could be dealt with are highlighted, only now the sequence is much more robust and co-linear.

Gene count

One revisited issue is the human gene count. This is corrected to an estimate of 22 500 (range 20–25 000): downwards from the already surprisingly low estimate of 31 000 in the working draft. One almost wonders what, other than genes that make humans embark on sequencing genomes, *does* set us apart from flies and worms… .

Several differences are responsible for the further decrease. The predominant cause is erroneous gene splitting: mapping errors, missing exons and erroneous stop codons turn out to have falsely split single genes into two or more in the working draft. Sequencing errors in pseudogenes account for another group of these predictions.

The authors note that this figure concerns protein genes, excluding noncoding RNA genes like rRNAs, tRNAs and microRNAs and large nontranslated RNAs, as well as other potentially functional elements embedded in conserved noncoding regions. Indeed, a sequence this robust enables for the first time a large-scale study of these biologically and evolutionarily highly relevant aspects, including a rigorous search for ancient, less conserved pseudogenes.

Gene birth and death

A first pass into this type of large-scale evolutionary biology is presented with the study of gene birth and death in humans. Of course, this newly discovered genomic plasticity, notably the duplicated genes within the segmental duplications, presents the ideal 'evolutionary putty' for positive selection.[4] Very recent duplications appear to account for about 1200 genes, most of these occurring within larger gene clusters (~3300 genes in total). Not unexpectedly, olfactory and

immunology-related genes are over-represented in these clusters, together with reproduction-related genes. Interestingly, an excess of gene birth events is found for the last ~3–4 million years. Besides, the obvious explanations of an enhanced gene radiation in primates, or a more extensive copy-editing through gene conversion within the human species, the authors suggest a third explanation: some may just be 'passenger genes', without functional benefit but still too 'young' to have mutated into pseudogenes.

The authors analyze the occurrence of gene death, too. Out of 34 well-studied very recent pseudogenes, 19 have two or more mutations and were pseudogenes already in the chimpanzee. Of the 14 with one mutation, eight are shared with the chimp, and five are functional in the chimp, while one is a segregating polymorphism in humans. Of the 32 fixed human pseudogenes, 22 cover a wide (ex)functional variety, while 10 are olfactory receptors, further underlining the dynamics in this family, as also in evidence from a recent study among humans.[5]

Genome plasticity

A main issue that can now be addressed more solidly than with the preliminary data, is that of the presence and implications of segmental duplications. Clearly, these duplications are at the heart of the remaining quality and gap issues. Looking at it from a biological perspective, segmental duplication is most likely both a main cause and a main consequence of genome plasticity. These duplications amount to ~5.3% of the euchromatic DNA, most of which is sequenced but with ~10% in the unsequenced gaps. Recently, it became clear that the 'fixed' segmental duplications have a swathe of younger relatives, when three groups[6–8] reported segmental copy number polymorphism (CNP) last summer. While the extent of these CNPs has yet to be established, two hardly overlapping sets of over 200 elements (median size 0.46 Mb) have been described.[5,7] Moreover, since only ~10% of these CNPs achieved an allele frequency of 10%, we may be in for much more extensive individual-specific ('private'), and common-heredity, group-specific variability of gene dosage than considered thus far.

Both the fixed and polymorphic segmental duplications have significant relevance for human genetic pathology. Examples mentioned in the paper are the Charcot–Marie–Tooth/HNPP region on 17p, the Williams syndrome region on 7q and the DiGeorge regions on 22q. However, ample additional examples exist, with the extreme of Down's syndrome on the one end and the polymorphic SMN region on chromosome 5, which influences the severity of spinal muscular atrophy (SMA), on the other. Incidentally, the latter is dealt with in more detail in another recent paper, providing the full annotation of the finished chromosome 5 sequence.[9] Chromosome 5 is a classroom example of what segmental duplication may do to an everyday piece of single-copy DNA. I still remember the lively debate on the Mediterranean Ile des Embiez, in the early nineties, among members of the SMA community on the clinical criteria to distinguish SMA types I–III, or alternatively types I–VIII. If this community had known what they were getting themselves into, one wonders if they would have had the courage.

Specifically, now we know that CNPs could be a source – and mechanism – of quantitative variation present in a disease like SMA, one might ask how much of the variation in complex genetic diseases could these kinds of differences account for.[6] This is not a trivial question to address. Most large SNP association studies to find common disease risk factors typically are based on the hypothesis of qualitative and not quantitative polymorphism. So far, the prevailing idea, among scientists as well as the public at large, is that the source of common diseases is the combined effect of 'poorer' and 'better' genes. It would significantly ease our task as geneticists in communicating to the public, if we could put this in the perspective of more or fewer copies of perfectly normal genes.

Finally, in the light of gene birth and death, one might even speculate that the occasional duplicated gene or genetic segment, in different ethnic backgrounds and environments, might have started to diverge, at least in copy number, before – or even without – fixation in all humans.

Data mining: from biology to innovation

The authors of this landmark paper anticipate the increasingly rewarding postgenomic research that is now becoming possible and highlight the importance of multiple genomes being sequenced over the coming years. However, the real work, of course, has already begun: all human genetics journals have seen a significant increase in the amount of interesting papers. They range from basic to clinical, and from large-scale, discovery-oriented to very focused, hypothesis-driven. Indeed, one can validly conclude that this is the human genome project really at work: allowing more people, in more labs, in more countries, to contribute to the discovery process, and thus eventually to improving healthcare. This trend clearly vindicates the wise decision of the Public Consortium to put their work in the public domain with highest priority.

Finally, the specific impact of the genome-sequence-driven expansion of human genetics in Europe should be considered. This is a highly data-intensive research era, requiring tremendous multidisciplinary crosstalk between genomic, proteomic, biological, clinical and epidemiological databases. However, the infrastructure of databases in Europe is in a far from well-supported state. Despite the good intentions of many bodies – and the efforts of some – it is amazing how difficult it is to establish a visionary, stable and sufficient long-term funding structure in Europe. The European integrated FP6-project 'BioSapiens' (http://www.ebi.ac.uk/biosapiens/) is at least one example of a first step towards an integrated, multidisciplinary bioinformatics infrastructure in Europe. However, this is a project with a finite (5-year) duration, so it is, in nature, a temporary solution. Similarly, we do have highly valuable data resources such as EMBL-EBI (http://www.ebi.ac.uk) and Swissprot (http://www.ebi.ac.uk/swissprot/), but their continuity, size, scale and business model is a continuing topic of debate. With the

scaling up of biology, the lagging of a well-funded European central database infrastructure undermines the core of European medical and biological research: the easy, continued access to a rising tide of high-density data. Without flourishing, well-accessible resources, which are being actively codeveloped in parallel to other regions in the world, we will not gain the required momentum in turning data into insights. And it is the insights that will be fuelling the engine of European biotech innovation■

Gert-Jan B van Ommen is at the Center for Medical Systems Biology and the Center of Human and Clinical Genetics, Leiden University Medical Center, PO Box 9503, 2300 RA Leiden, The Netherlands.
E-mail: gjvo@lumc.nl

References

1 International Human Genome Sequencing Consortium: Finishing the euchromatic sequence of the human genome. *Nature* 2004; **431**: 931–945.
2 Eichler EE, Clark RA, She X: An assessment of the sequence gaps: unfinished business in a finished human genome. *Nat Rev Genet* 2004; **5**: 345–354.
3 She X, Jiang Z, Clark RA *et al*: Shotgun sequence assembly and recent segmental duplications within the human genome. *Nature* 2004; **431**: 927–930.
4 Johnson ME, Viggiano L, Bailey JA *et al*: Positive selection of a gene family during the emergence of humans and African apes. *Nature* 2001; **413**: 514–519.
5 Menashe I, Man O, Lancet D, Gilad Y: Different noses for different people. *Nat Genet* 2003; **34**: 143–144.
6 Sebat J, Lakshmi B, Troge J *et al*: Large-scale copy number polymorphism in the human genome. *Science* 2004; **305**: 525–528.
7 Fredman D, White SJ, Potter S, Eichler EE, Den Dunnen JT, Brookes AJ: Complex SNP-related sequence variation in segmental genome duplications. *Nat Genet* 2004; **36**: 861–866.
8 Iafrate AJ, Feuk L, Rivera MN *et al*: Detection of large-scale variation in the human genome. *Nat Genet* 2004; **36**: 949–951.
9 Schmitz J, Martin J, Terry A *et al*: The DNA sequence and comparative analysis of human chromosome 5. *Nature* 2004; **431**: 268–274.

Evolutionary Genetics

Genetics of lactase persistence – fresh lessons in the history of milk drinking

Edward Hollox

European Journal of Human Genetics (2005) **13,** 267–269.
doi:10.1038/sj.ejhg.5201297
Published online 15 December 2004

Most people cannot drink milk as adults without the symptoms of lactose intolerance, and most lactose intolerance is due to absence of the lactase enzyme in the gut. This presence/absence is a genetic polymorphism commonly called lactase persistence/nonpersistence, depending on whether or not lactase activity persists from childhood into adulthood.[1] In Northern Europe, lactase persistence is common and many people not only drink milk, but culturally it is seen as a healthy and nutritious food. How this happened is now becoming clearer.

Lactase nonpersistence is the ancestral state, and lactase persistence only became advantageous after the invention of agriculture, when milk from domesticated animals became available for adults to drink. As expected, lactase persistence is strongly correlated with the dairying history of the population. This genetic ability to digest milk has been regarded as a classic example of gene-culture coevolution, where the culture of dairying creates a strong selective advantage to those who can drink milk as adults, for only they can nutritionally benefit from the milk. A recent paper confirmed this link by analysing the diversity in bovine milk protein genes and showing that the highest gene diversity (and by implication the largest historical population size) is in cows from areas of the world where dairy farming is practised and the people are lactose tolerant.[2] In humans, epidemiological analysis has shown that the cultural development of dairying preceded selection for lactase persistence.[3] Since dairying is thought to have originated around 10 000 years ago, the selective pressure has been only for the past 400 generations. Despite this short time, there is suggestive evidence of recent positive selection: lactase persistence is associated with one haplotype, which is very common only in northern Europeans, and is distant from the ancestral haplotype.[4,5] Discovery of the possible molecular basis of this polymorphism – a single nucleotide change 14 kb away from the gene, has allowed further analysis of genetic variation associated with lactase persistence/nonpersistence.[6–8]

Proving that the lactase gene has been under recent positive selection in Northern Europe is difficult. As it is a recent regulatory change, codon-based methods that examine the different substitution patterns across a gene are not suitable. Instead, methods relying on allele frequency must be used – which are vulnerable to the fact that frequency patterns produced by selection can also be produced by demographic processes such as changes in population size and genetic drift. A statistic called 'relative extended haplotype homozygosity' (REHH) has been developed, which relies on the fact that a selected haplotype (ie a haplotype