Genome Biology

Minireview

The promise and reality of personal genomics

Bryndis Yngvadottir, Daniel G MacArthur, Hanjun Jin and Chris Tyler-Smith

Address: The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton CB10 1SA, UK.

Correspondence: Chris Tyler-Smith. Email: cts@sanger.ac.uk

Abstract

The publication of the highest-quality and best-annotated personal genome yet tells us much about sequencing technology, something about genetic ancestry, but still little of medical relevance.

Which country has published the largest per-capita number of personal genomes? The United States, the United Kingdom? Actually, it is Korea. A recent article in *Nature* by Kim *et al.* [1] presents the genome sequence of a Korean male, AK1 - the seventh published sequence of an individual human genome and the second from Korea. The rapid progress in personal genome sequencing is possible because so-called 'next-generation' sequencing technology has decreased costs by orders of magnitude and increased throughput. But those advantages come at a price: short, error-prone reads derived from single molecules that have to be stitched back together to make a best-guess at the starting sequence. We are still at the stage of working out how to apply the available technologies to coax out biological information: the goal of a US$1,000 genome providing life-changing personal medical insights is still some way off.

Genome sequencing is still an imprecise science

The first aim of a genome-sequencing project is to assemble around 6 billion As, Cs, Gs and Ts, comprising the diploid genome of the individual, in the right order. This is a challenge both of scale and because of sequence complexities such as repeated elements. By a series of frankly heroic measures, Kim *et al.* [1] have succeeded in generating a sequence that is likely to be substantially more complete and accurate than any other individual human genome derived so far using the new sequencing technology. Nonetheless, the effort invested in producing such a high-quality sequence, including the cloning and high-coverage sequencing of large segments of the genome from bacterial artificial chromosomes (BACs), is not routinely feasible and the final product is still far from complete. The clear message is that sequencing technology still has a long way to go before we enter the era of cheap, complete and reliable individual genome sequencing.

The high depth of coverage for the AK1 sequence (most parts of the genome were sequenced around 28 times (28x)) meant that most variant sites (arising from heterozygosity within AK1's diploid genome or homozygous differences between this and the reference genome) that could be detected were called accurately. However, the sensitivity of variant detection was low: the authors estimate that they missed 6% of single nucleotide polymorphisms (SNPs) and more than 20% of insertion/deletion variants (indels). Over a whole genome that would add up to 150,000 missed SNPs and perhaps 60,000 unseen indels. The true number missed is likely to be substantially higher, however, as the set of 'true' calls used to derive the SNP sensitivity estimates (from the Illumina 610K genotyping array) is biased towards regions that are likely to be easy to both genotype and sequence. The sensitivity of SNP and indel detection in repetitive or copy-number-variable regions will be lower. Finally, a non-trivial proportion of any genome (150 Mb of the Venter genome [2], and almost 6% of the reads from the first Korean genome [3], for example) is not present in the 'reference human genome' sequence and so reads cannot be mapped to these sections or variants called at all.

The authors invested particular effort in the identification of larger indels, known as copy number variants (CNVs), using both targeted sequencing of BACs and high-resolution chip-based approaches. Large numbers of high-quality CNVs were detected, but it is worth noting that such variants will also have been missed in highly repetitive regions of the genome and that structural rearrangements that do not change DNA copy number - such as inversions - are likely to have been substantially under-called.

Comparison between the individual genomes sequenced so far (Table 1) is complicated by differences in chemistry, coverage, alignment and variant-calling algorithms used, but perhaps most of all by the absence of 'ground truth' large-scale sequence data from which unbiased estimates of error rates could be deduced. So far, only one individual genome has been sequenced using two technologies - the anonymous Nigerian Yoruba male (NA18507) sequenced by both Illumina GA (Solexa) [4] and Applied Biosystem's SOLiD [5] systems - but no explicit comparison of the two versions has yet been published.

Table 1

Summary of seven personal genomes in order of date of publication

Year	Individual	Population	Platform	Coverage	Reference
2007	Craig Venter	European	Capillary	7.5x	[2]
2008	James Watson	European	454	7.4x	[10]
2008	NA18507	Nigerian (Yoruba)	Illumina GA	40.6x	[4]
2008	YH	Han Chinese	Illumina GA	36x	[11]
2008	AML patient	European	Illumina GA	14x,33x*	[12]
2009	Seong-Jin Kim (SJK)	Korean	Illumina GA	29x	[3]
2009	NA18507	Nigerian (Yoruba)	SOLiD	17.9x	[5]
2009	AK1	Korean	Illumina GA	27.8x	[1]

*Normal genome sequenced to 14x, tumor genome to 33x. Note that NA18507 has been sequenced twice using different technologies. AML, acute myelogenous leukemia.

Mitochondrial DNA (mtDNA) provides a useful test case: its high copy number and lack of repeats should lead to a high-quality sequence, and because many thousands of additional mtDNA sequences and their phylogeny are available [6], we can assess a new sequence even without ground truth. In the case of AK1, the sequence passes this test, but assessment of heteroplasmy - variation of this multicopy molecule within an individual - remains problematic, because of alignment difficulties at some positions, such as 3521.

Individual genome sequences as indicators of ancestry

With the first individual human genome sequences now available, what biological information can we hope to gain from them? One area is ancestry: we are curious about our ancestors and as humans are genetically slightly more similar to their geographical neighbors than to people from further away, with enough genetic information we can infer the geographical ancestry of an individual on a remarkably fine scale [7]. It is reassuring to see that the published personal genomes do fit expectations. As Figure 1 reveals, Venter falls within the European region, whereas the Watson genome sequence displays, in addition to the expected major European component, a strong minor ancestry component corresponding to the dominant component in African populations. This could be regarded as support for the notion that Watson has considerable African admixture, a claim made previously in the mainstream media but never (to our knowledge) formally supported in the literature. However, a plausible alternative explanation is that this component is an artifact of the low coverage and poorer sequence quality in the Watson genome. Unsurprisingly, the Yoruba NA18507 genome falls alongside the other HapMap Yoruba (YRI(HapMap) in Figure 1), and the Han YH genome with the other Han groups from HapMap (CHB(HapMap) in Figure 1). The two Korean genomes, SJK and AK1, show close affiliation with populations from East Asia. These conclusions are reinforced by Y-chromosomal and mtDNA analyses. The mtDNA haplogroups (groups of similar haplotypes, characterized by a single SNP, that share a common ancestor) of SJK and AK1 are both D4, respectively, whereas the Y haplogroups are O2b and O3a, all haplogroups prevalent in the Korean population [8].

All of these conclusions could have been obtained by standard genotyping at a price three orders of magnitude lower than the cost of a complete genome sequence, so does the full sequence provide extra insights? Ahn *et al.* [3] emphasize the differences between the SJK ('Korean') and YH ('Chinese') genomes, and we expect that rare variants usually missed by genotyping will provide much more information about fine-scale ancestry. But many more personal genomes will be needed before we can benefit fully from such comparisons.

Kim *et al.* [1] report a strong correlation between regional SNP and indel densities as an unexpected finding, and propose that "unifying molecular or temporal considerations underpin the generation and/or removal of both types of variants". In fact, this correlation is a straightforward prediction from population genetics: SNP and indel densities both depend on the coalescence time (that is, the time since the most recent common ancestor) between AK1 and the reference genome. Because coalescence time varies across the genome, both densities would be expected to vary in a correlated fashion.

How far are we from genome-based personalized medicine?

The primary goal of human genome sequencing is not, of course, to obtain ever-finer insights into genetic ancestry or the consequences of coalescence times, but rather to generate information to inform the practice of individualized medicine. Here, there are two concerns: firstly, the

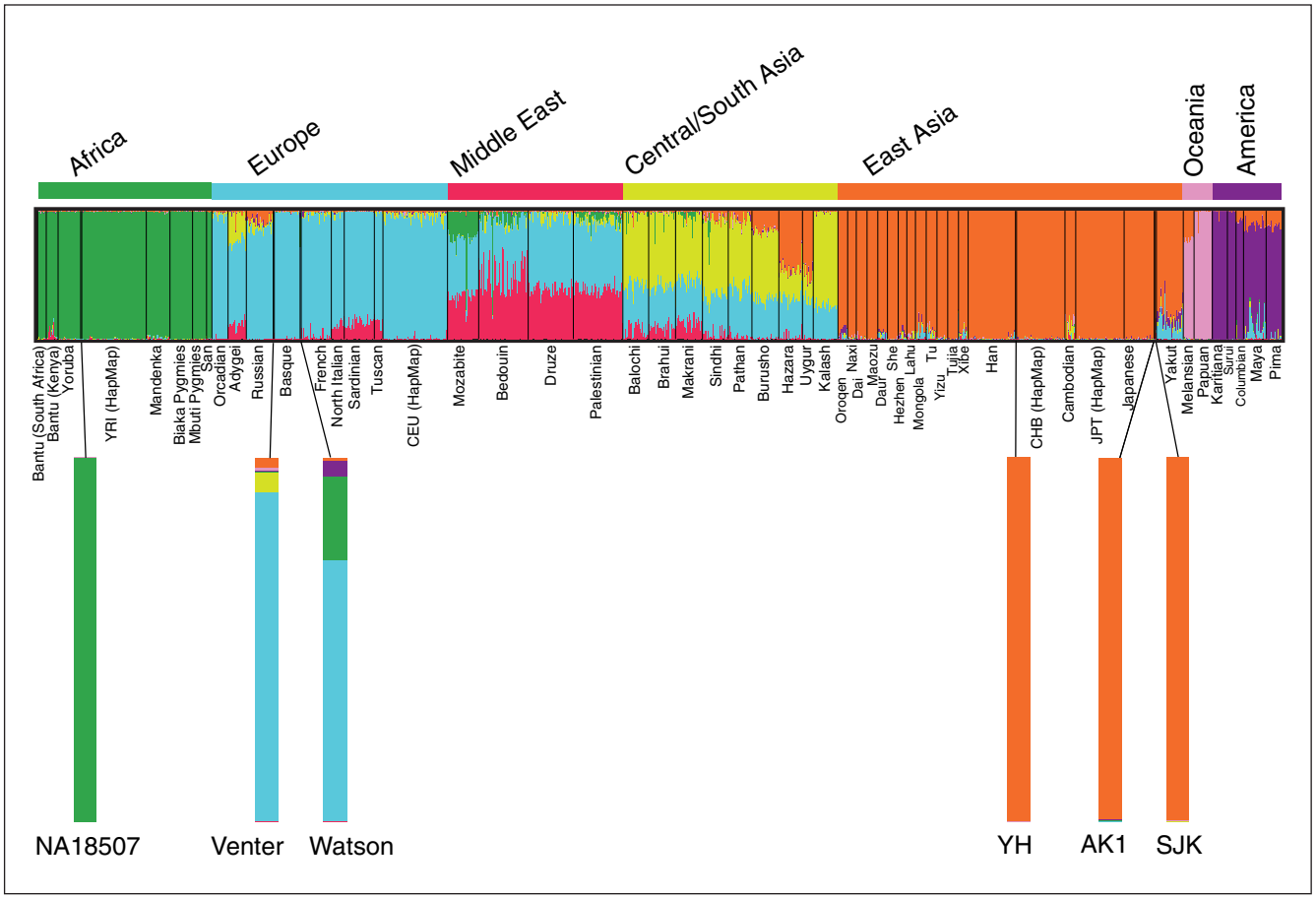

Figure 1

Ancestry inferences from personal genomes in a worldwide context. The program STRUCTURE was used to assign six personal genomes (Table 1) and individuals from the HGDP-CEPH and HapMap panels to seven clusters on the basis of their genotypes as described previously [9]. Upper section: each thin vertical bar represents an individual and is divided into colors corresponding to their inferred ancestry from the seven genetic clusters. Individuals are ordered according to their name or code (personal genomes) or population of origin (HGDP-CEPH and HapMap). Personal genomes fall predominantly into the green (NA18507; African), cyan (Venter, Watson; European) or orange (YH, SJK, AK1; East Asian) clusters. Lower section: expanded view of the personal genome inferred ancestries.

incomplete detection of genetic variants already noted means that a non-trivial fraction of the variants affecting health are missed by current genome sequencing approaches; secondly, our current ability to interpret the medical significance of identified variants is rudimentary.

Kim *et al.* [1] applied an unpublished algorithm (Trait-o-matic) to identify those variants within the AK1 genome that have been associated with phenotypic traits, including increased risk for a wide variety of common diseases, as well as protein-altering variants in positions that are strongly evolutionarily conserved or in genes associated with severe disease. This analysis identified 773 potentially medically relevant variants. As in the ancestry analysis, the common variants highlighted could just as easily have been identified using a SNP genotyping chip. Still, many of those are robustly associated with traits (that is, they have achieved genome-wide significance and independent replication) but also generally have very low predictive value for disease risk. The potentially more interesting variants in the AK1 genome are those that could not have been identified by SNP chips: low-frequency variants that might be expected to disrupt the function of important genes. The authors identified a total of 504 variants in AK1 that alter the protein sequences of genes associated with diseases or traits, but this list illustrates the serious challenges associated with the functional interpretation of such variants.

There are some straightforward results: for instance, the AK1 genome reportedly carries single copies of premature stop-codon mutations in genes associated with severe recessive diseases such as cerebral palsy, retinitis pigmentosa and malonyl-CoA decarboxylase deficiency; these are

unlikely to be associated with a disease phenotype in AK1 himself, but might (if genuine) be important for genetic counseling. In contrast, the clinical importance of the vast majority of the 773 variants highlighted by Trait-o-matic is far from clear. For instance, what is a clinician to make of novel protein-altering variants in genes known to be associated with cancer risk or progression? And how should a sequenced individual respond to a bewildering array of variants associated with increased risk of depression, bipolar disorder and schizophrenia?

A further layer of uncertainty is added by the fact that most of the common variants currently associated with complex traits and diseases have been ascertained and studied only in populations of European origin, and the possibility of altered risk profiles due to different gene-gene and gene-environment interactions in non-European populations is largely unexplored. Clearly, much further work remains to be done before individual genome sequences can serve as a routine source of information for clinical decision-making.

Personal genomics is in its infancy. Like all infants, it makes a lot of noise and attracts a lot of attention, but is poor at communication: readers will not find comparisons of the two Korean genomes, or the two versions of the same Yoruba genome, for example. But the infant will grow, and the publication of AK1 and other individual genomes do represent important milestones on the path towards affordable, medically relevant personal genomics. However, they are also useful reminders of just how far we still have to go before this destination is reached.

Some key steps that we look forward to, in addition to decreasing cost and increasing accuracy, are technical advances such as *de novo* assembly - the stitching together of reads without the use of a reference sequence, a process that will benefit from longer read lengths - and improvements in phenotype interpretation, as noted earlier. Some personal genomics subjects have bravely presented their genomes to the world 'warts and all' along with their names, whereas others have masked certain regions or chosen to remain anonymous. Any position can be criticized and the ethical implications of revealing all this information are just being worked out. We all owe a debt to these pioneers.

Acknowledgments

Our work is supported by the Wellcome Trust.

References

1. Kim JI, Ju YS, Park H, Kim S, Lee S, Yi JH, Mudge J, Miller NA, Hong D, Bell CJ, Kim HS, Chung IS, Lee WC, Lee JS, Seo SH, Yun JY, Woo HN, Lee H, Suh D, Lee S, Kim HJ, Yavartanoo M, Kwak M, Zheng Y, Lee MK, Park H, Kim JY, Gokcumen O, Mills RE, Zaranek AW, *et al.*: **A highly annotated whole-genome sequence of a Korean individual.** *Nature* 2009, **460:**1011-1015.
2. Levy S, Sutton G, Ng PC, Feuk L, Halpern AL, Walenz BP, Axelrod N, Huang J, Kirkness EF, Denisov G, Lin Y, MacDonald JR, Pang AW, Shago M, Stockwell TB, Tsiamouri A, Bafna V, Bansal V, Kravitz SA, Busam DA, Beeson KY, McIntosh TC, Remington KA, Abril JF, Gill J, Borman J, Rogers YH, Frazier ME, Scherer SW, Strausberg RL, Venter JC: **The diploid genome sequence of an individual human.** *PLoS Biol* 2007, **5:**e254.
3. Ahn SM, Kim TH, Lee S, Kim D, Ghang H, Kim DS, Kim BC, Kim SY, Kim WY, Kim C, Park D, Lee YS, Kim S, Reja R, Jho S, Kim CG, Cha JY, Kim KH, Lee B, Bhak J, Kim SJ: **The first Korean genome sequence and analysis: full genome sequencing for a socio-ethnic group.** *Genome Res* 2009, DOI:10.1101/gr.092197.109.
4. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR, Boutell JM, Bryant J, Carter RJ, Keira Cheetham R, Cox AJ, Ellis DJ, Flatbush MR, Gormley NA, Humphray SJ, Irving LJ, Karbelashvili MS, Kirk SM, Li H, Liu X, Maisinger KS, Murray LJ, Obradovic B, Ost T, Parkinson ML, Pratt MR, *et al*: **Accurate whole human genome sequencing using reversible terminator chemistry.** *Nature* 2008, **456:**53-59.
5. McKernan KJ, Peckham HE, Costa GL, McLaughlin SF, Fu Y, Tsung EF, Clouser CR, Duncan C, Ichikawa JK, Lee CC, Zhang Z, Ranade SS, Dimalanta ET, Hyland FC, Sokolsky TD, Zhang L, Sheridan A, Fu H, Hendrickson CL, Li B, Kotler L, Stuart JR, Malek JA, Manning JM, Antipova AA, Perez DS, Moore MP, Hayashibara KC, Lyons MR, Beaudoin RE, *et al*: **Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding.** *Genome Res* 2009, DOI:10.1101/gr.091868.109.
6. Pereira L, Freitas F, Fernandes V, Pereira JB, Costa MD, Costa S, Maximo V, Macaulay V, Rocha R, Samuels DC: **The diversity present in 5140 human mitochondrial genomes.** *Am J Hum Genet* 2009, **84:**628-640.
7. Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ramachandran S, Cann HM, Barsh GS, Feldman M, Cavalli-Sforza LL, Myers RM: **Worldwide human relationships inferred from genome-wide patterns of variation.** *Science* 2008, **319:**1100-1104.
8. Jin HJ, Tyler-Smith C, Kim W: **The peopling of Korea revealed by analyses of mitochondrial DNA and Y-chromosomal markers.** *PLoS One* 2009, **4:**e4210.
9. He M, Gitschier J, Zerjal T, de Knijff P, Tyler-Smith C, Xue Y: **Geographical affinities of the HapMap samples.** *PLoS One* 2009, **4:**e4684.
10. Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, McGuire A, He W, Chen YJ, Makhijani V, Roth GT, Gomes X, Tartaro K, Niazi F, Turcotte CL, Irzyk GP, Lupski JR, Chinault C, Song XZ, Liu Y, Yuan Y, Nazareth L, Qin X, Muzny DM, Margulies M, Weinstock GM, Gibbs RA, Rothberg JM: **The complete genome of an individual by massively parallel DNA sequencing.** *Nature* 2008, **452:**872-876.
11. Wang J, Wang W, Li R, Li Y, Tian G, Goodman L, Fan W, Zhang J, Li J, Zhang J, Guo Y, Feng B, Li H, Lu Y, Fang X, Liang H, Du Z, Li D, Zhao Y, Hu Y, Yang Z, Zheng H, Hellmann I, Inouye M, Pool J, Yi X, Zhao J, Duan J, Zhou Y, Qin J, *et al.*: **The diploid genome sequence of an Asian individual.** *Nature* 2008, **456:**60-65.
12. Ley TJ, Mardis ER, Ding L, Fulton B, McLellan MD, Chen K, Dooling D, Dunford-Shore BH, McGrath S, Hickenbotham M, Cook L, Abbott R, Larson DE, Koboldt DC, Pohl C, Smith S, Hawkins A, Abbott S, Locke D, Hillier LW, Miner T, Fulton L, Magrini V, Wylie T, Glasscock J, Conyers J, Sander N, Shi X, Osborne JR, Minx P, *et al.*: **DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome.** *Nature* 2008, **456:**66-72.

Published: 02 September 2009
doi:10.1186/gb-2009-10-9-237