2× genomes––Does depth matter?

Phil Green

Genome Res. 2007 17: 1547-1549
Access the most recent version at doi:10.1101/gr.7050807

Commentary

2× genomes—Does depth matter?

Phil Green

Department of Genome Sciences and Howard Hughes Medical Institute, University of Washington, Seattle, Washington 98195, USA

This issue of *Genome Research* marks the publication by collaborators at the NIH, Agencourt Bioscience, and the Broad Institute of a genome sequence for the domestic cat *Felis catus* (Pontius et al. 2007). It was obtained by a whole-genome shotgun approach in which, on average, each genomic base is represented in roughly two sequence reads ("2×" redundancy), a level at which there remain many gaps in the sequence due to statistical fluctuations in read placement, biases in subclone libraries, and assembly difficulties. In all, the NHGRI is sponsoring 24 such "low redundancy" mammalian genome sequences, 17 of which have already been assembled and released (Table 1). Complete coverage of every genome would obviously be preferable; the decision to acquire multiple incomplete sequences represents a compromise balancing phylogenetic breadth (inclusion of many species) against redundancy depth for particular species. In this commentary, I discuss some of the issues involved in this compromise, and touch on the uses, characteristics, and limitations of 2× assemblies.

The tradeoff between breadth and depth has actually been a recurring theme in genome sequencing, because the choice between spreading available data-generating capacity broadly over a larger extent of targeted DNA and obtaining deeper coverage of a more limited target arises in several contexts. The appropriate balance in each case is far from obvious and depends on a number of factors, including the relative difficulty of obtaining DNA sources of various types, available data analysis tools, and the implications of sequence errors and gaps for downstream utility. Early in the Human Genome Project, the high level of redundancy required by the shotgun method was often viewed as unduly wasteful, and considerable effort was expended toward developing more efficient directed strategies. These all foundered, due to a combination of logistical complexity, the problems posed by interspersed repeats, and the recognition that variation in data quality makes it impossible to get highly accurate sequence without considerable redundancy. Once the superiority of the shotgun approach was accepted, a key issue became the breadth of the targeted region: is it better to shotgun a whole genome, or a series of individual BAC clones (Green 1997; Weber and Myers 1997)? Although not immediately obvious, this choice also involves a breadth versus depth tradeoff, in two ways: with whole genome targets, typically 25% of reads cannot be assembled (Adams et al. 2000; Venter et al. 2001), thereby reducing the overall effective depth, and, in addition, the inability to control redundancy depth in local regions (as can readily be done in a clone-by-clone approach) implies that a larger fraction of the genome will be poorly covered. The whole-genome approach is also unable to resolve many segmental duplications (She et al. 2004) and interspersed repeats (Adams et al. 2000). Nonetheless, the (considerable) advantage of simpler logistics, combined with a willingness to tolerate less accurate and less complete sequence, have made the whole-genome strategy the current method of choice for sequencing most genomes.

With completion of the human genome sequence, the breadth versus depth issue has now shifted to how best to apply available sequencing capacity across organisms. Broad phylogenetic representation accelerates research on a wide variety of species and provides general insights into the core evolutionary processes of mutation and selection. However, from the perspective of the human genome, its main benefit is to help identify functional features as regions with a reduced frequency of substitution differences between organisms due to purifying selection. The statistical power to detect such regions depends on overall sequence divergence: there is an analogy to shotgun-sequence assembly, in that one seeks "coverage" of human bases by multiply aligned orthologous sequences instead of reads, and the relevant "depth" is total branch length (expected number of mutations per neutral site) of the phylogenetic tree relating the species. Theoretical analysis (Eddy 2005) indicates that a surprisingly large branch length is required to resolve small features with reasonable power—for example, the mouse genome, which among placental mammals is one of the most highly diverged from human, yields only enough information to identify conserved features longer than 50 bp. While there is substantial branch length available from sequenced non-mammalian vertebrates, they provide no information regarding mammal-specific features. The 24 NHGRI 2× targets, combined with the seven or so other placental mammals (human, mouse, rat, dog, chimpanzee, macaque, cow) having more complete sequences, are predicted to provide adequate branch length to resolve features of the size of a typical protein-binding site (6–8 bp) at a false positive rate of one per 10 kb (http://www.genome.gov/25521745). This resolution could not be achieved by allocating the same number of sequence reads to six genomes at a more conventional 8× redundancy, or even 12 genomes at 4×.

It is worth noting that effective delineation of conserved elements via this strategy is not yet a completely solved problem. Multiple-genome alignments are error prone even with relatively complete sequences (Prakash and Tompa 2007). Perhaps more seriously, comprehensive analysis of local genomic regions (ENCODE Project Consortium 2007) suggests that, at least with current methods, sequence conservation fails to detect some important features, even when alignments appear reliable. This likely reflects, at least in part, the intrinsic plasticity of the sequence signals involved, and it remains unclear to what extent improvements in computational strategies or additional sequence data will be able to address this.

What are the consequences of reduced depth for individual genomes? The primary determinants of sequence utility are assembly accuracy (correctness of read overlaps and of contig order and orientation) and percent coverage of the genome. For low-redundancy (~0.5× to ~3×) shotguns, both accuracy and coverage can be improved by 'assisting' the assembly using available near-complete "reference" genomes from other species. Cat is a

E-mail phg@u.washington.edu; fax (206) 685-9720.
Article is online at http://www.genome.org/cgi/doi/10.1101/gr.7050807. Freely available online through the *Genome Research* Open Access option.

Table 1. NHGRI-sponsored 2× genome sequences

Common name	Scientific name	2× Status	Deeper coverage in progress?
African savannah elephant	*Loxodonta africana*	Assembled and released	Yes
Nine-banded armadillo	*Dasypus novemcinctus*	Assembled and released	Yes
European rabbit	*Oryctolagus cuniculus*	Assembled and released	Yes
Lesser hedgehog (Tenrec)	*Echinops telfairi*	Assembled and released	No
European common shrew	*Sorex araneus*	Assembled and released	No
Guinea pig	*Cavia porcellus*	Assembled and released	Yes
European hedgehog	*Erinaceus europeaus*	Assembled and released	No
Cat	*Felis catus*	Assembled and released	Yes
Little brown bat	*Myotis lucifugus*	Assembled and released	Yes
Ground squirrel	*Spermophilis tridecemlineatus*	Assembled and released	No
Bushbaby	*Otolemur garnetti*	Assembled and released	Yes
Tree shrew	*Tupaia belangeri*	Assembled and released	Yes
Horse	*Equus caballus*	Assembled and released	Yes
Pika	*Ochotona princeps*	Assembled and released	No
Mouse lemur	*Microcebus murinus*	Assembled and released	Yes
Hyrax	*Procavia capensis*	In process	No
Megabat	*Pteropus vampyrus*	Assembled and released	No
Dolphin	*Tursiops truncatus*	Assembled and released	No
Tarsier	*Tarsier syrichta*	In process	Yes
Kangaroo rat	*Dipodomys spp.*	In process	No
Chinese pangolin	*Manis pentadactyla*	In process	No
Two-toed sloth	*Choloepus hoffmanni*	In process	No
Alpaca	*Vicugna pacos*	In process	No
Flying lemur	*Dermoptera spp.*	In process	No

Sources: http://www.genome.gov/25521745; http://www.broad.mit.edu/mammals/; A. Felsenfeld (pers. comm.).

favorable case (relative to other low-redundancy sequences that are in progress) in having a reasonably close reference sequence, that of the dog, to which it is ~80% identical at the nucleotide level, and Pontius et al. (2007, this issue) make effective use of it. In its most extreme version, assisted assembly consists simply of aligning reads to the reference genome, using read mate-pair information to reduce the risk of misassembly due to structural differences between the two genomes. In this approach, some reads are effectively lost because they cannot be uniquely placed, or because of insertions or deletions in one lineage, but experiments by Margulies et al. (2005) suggest that the vast majority should find their orthologous regions. As redundancy increases, many reads not directly alignable to the reference sequence can be incorporated into the assembly by virtue of overlaps with other reads. For redundancies of 2× or more, reasonable de novo (unassisted) assembly becomes possible (Margulies et al. 2005), but there does not seem to be a strong reason to prefer this to assisted assembly.

For both a simulated unassisted 2× mouse genome assembly (Margulies et al. 2005) and the assisted 1.9× cat genome assembly of Pontius et al. (2007) euchromatic genome coverage by assembled contigs was only ~65%, significantly less than the theoretical Poisson expectation (Lander and Waterman 1988) of ~85%. The shortfall presumably reflects some combination of subcloning biases and assembly difficulties caused by repetitive sequence. It will be important to sort this out and, if possible, alleviate it, because the reduced coverage compromises power to identify functional elements in the human genome (which is as noted above the main rationale for the 2× sequences). On the other hand, accuracy of both assemblies appears reasonably good. However, there do seem to be many ambiguities in the order of contigs along the chromosome (although this has apparently not been quantified), and only 54% of the cat genome is covered by chromosomally mapped contigs.

The analyses by Pontius et al. (2007) and an earlier "survey sequence" analysis of the dog (Kirkness et al. 2003) reveal the types of biological information that can be gleaned from low-redundancy genomes: a reasonably comprehensive assessment of interspersed repeat content (although not their specific locations) including the identification of lineage-specific families, partial sequences of most genes and other evolutionarily conserved segments together with a rough orthology/synteny map of their chromosomal arrangement relative to related species, and average estimates of mutation rates of various types (including chromosome rearrangements). In addition, allelic differences between overlapping reads yield a large number of polymorphisms, providing information about population genetic history as well as a resource for phenotype mapping. This is probably the main benefit of 2× versus even lower redundancies.

On the other hand, the low-percent coverage significantly limits many applications of the sequence. Inferences about lineage-specific losses of genes or other functional features are not possible, and it is hard to distinguish genes from pseudogenes. More seriously, since very few features of any appreciable size (e.g., genes) will be completely covered, analyses requiring complete features cannot be carried out. In addition, as was noted above, whole-genome assemblies (of any depth) often fail to incorporate a significant fraction of the repetitive sequence in the genome. This is often considered to be a relatively minor deficiency, which may be true so long as the primary research focus is on broadly shared biological features. However, it is now apparent that repetitive sequence is a key agent of evolutionary change: Segmental duplications are likely the primary source of new genes (Ohno 1970; Sharp et al. 2006), and recent evidence strongly suggests that transposable elements are important mechanisms of regulatory innovation (Kamal et al. 2006; Nishihara et al. 2006; Lowe et al. 2007; Mikkelsen et al. 2007). As researchers' attention turns toward understanding differences among organisms rather than similarities, inadequate coverage of these types of sequences will become increasingly problematic.

What are the prospects for correcting these deficiencies? Fortunately, 11 of the 24 2× genomes (including cat) are already slated for deeper sequencing (Table 1). For the others, technical improvements in assembly, including better discrimination between different repeat copies and more aggressive assisted assembly strategies, should help somewhat. One hope is that most gaps could be closed with large numbers of cheap short reads generated using newer technologies (Bentley 2006). This approach has been successfully applied to microbial genomes (Goldberg et al. 2006), but it is likely to be less effective with mammalian assemblies, since a large fraction of the gap edges tend to lie in repeats that are unbridgeable by short reads (and even when one is not interested in the repeats themselves, it is usually impossible to be

sure that an assembly gap consists entirely of repetitive sequence except by filling it). Short reads also will not resolve segmental duplications, although the additional redundancy they provide should increase power to detect specific duplications that have been artificially collapsed in the assembly. It seems likely that regions of particular interest will instead have to be completed primarily by targeted cloning and sequencing (Thomas et al. 2003; http://bacpac.chori.org/) using the existing assembly as a scaffold. A major function of the 2× genomes will no doubt prove to be whetting users' appetites for more complete sequences.

Acknowledgments

I thank E. Eichler, E. Green, R. Waterston, A. Felsenfeld, and two anonymous reviewers for suggestions.

References

Adams, M.D., Celniker, S.E., Holt, R.A., Evans, C.H., Gocayne, J.D., Amanatides, P.G., Sherer, S.E., Li, P.W., Hoskins, R.A., Galle, R.F., et al. 2000. The genome sequence of *Drosophila melanogaster*. *Science* **287:** 2185–2195.

Bentley, D.R. 2006. Whole-genome re-sequencing. *Curr. Opin. Genet. Dev.* **16:** 545–552.

Eddy, S.R. 2005. A model of the statistical power of comparative genome sequence analysis. *PLoS Biol.* **3:** e10. doi: 10.1371/journal.pbio.0030010.

The ENCODE Project Consortium. 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447:** 799–816.

Goldberg, S.M., Johnson, J., Busam, D., Feldblyum, T., Ferriera, S., Friedman, R., Halpern, A., Khouri, H., Kravitz, S.A., Lauro, F.M., et al. 2006. A Sanger/pyrosequencing hybrid approach for the generation of high-quality draft assemblies of marine microbial genomes. *Proc. Natl. Acad. Sci.* **103:** 11240–11245.

Green, P. 1997. Against a whole-genome shotgun. *Genome Res.* **7:** 410–417.

Kamal, M., Xie, X., and Lander, E.S. 2006. A large family of ancient repeat elements in the human genome is under strong selection. *Proc. Natl. Acad. Sci.* **103:** 2740–2745.

Kirkness, E.F., Bafna, V., Halpern, A.L., Levy, S., Remington, K., Rusch, D.B., Delcher, A.L., Pop, M., Wang, W., Fraser, C.M., et al. 2003. The dog genome: Survey sequencing and comparative analysis. *Science* **301:** 1898–1903.

Lander, E.S. and Waterman, M.S. 1988. Genomic mapping by fingerprinting random clones: A mathematical analysis. *Genomics* **2:** 231–239.

Lowe, C.B., Bejerano, G., and Haussler, D. 2007. Thousands of human mobile element fragments undergo strong purifying selection near developmental genes. *Proc. Natl. Acad. Sci.* **104:** 8005–8010.

Margulies, E.H., Vinson, J.P., NISC Comparative Sequencing Program, Miller, W., Jaffe, D.B., Lindblad-Toh, K., Chang, J.L., Green, E.D., Lander, E.S., Mullikin, J.C., et al. 2005. An initial strategy for the systematic identification of functional elements in the human genome by low-redundancy comparative sequencing. *Proc. Natl. Acad. Sci.* **102:** 4795–4800.

Mikkelsen, T.S., Wakefield, M.J., Akin, B., Amemeya, C.T., Chang, J.L., Duke, S., Garber, M., Gentles, A.J., Goodstadt, L., Heger, A., et al. 2007. Genome of the marsupial *Monodelphis domestica* reveals innovation in non-coding sequences. *Nature* **447:** 167–177.

Nishihara, H., Smit, A.F., and Okada, N. 2006. Functional noncoding sequences derived from SINEs in the mammalian genome. *Genome Res.* **16:** 864–874.

Ohno, S. 1970. *Evolution by gene duplication*. Springer-Verlag, New York.

Pontius, J.U., Mullikin, J.C., Smith, D., Agencourt Sequencing Team, Lindblad-Toh, K., Gnerre, S., Clamp, M., Chang, J., Stephens, R., Neelam, B., et al. 2007. Initial sequence and comparative analysis of the cat genome. *Genome Res.* (this issue). doi: 10.1101/gr.6380007.

Prakash, A. and Tompa, M. 2007. Measuring the accuracy of genome-size multiple alignments. *Genome Biol.* **8:** R124. doi: 10.1186/gb-2007-8-6-r124.

Sharp, A.J., Cheng, Z., and Eichler, E.E. 2006. Structural variation of the human genome. *Annu. Rev. Genomics Hum. Genet.* **7:** 407–442.

She, X., Jiang, Z., Clark, R.A., Liu, G., Cheng, Z., Tuzun, E., Church, D.M., Sutton, G., Halpern, A.L., and Eichler, E.E. 2004. Shotgun sequence assembly and recent segmental duplications within the human genome. *Nature* **431:** 927–930.

Thomas, J.W., Touchman, J.W., Blakesley, R.W., Boufford, G.G., Beckstrom-Sternberg, S.M., Margulies, E.H., Blanchette, M., Siepel, A.C., Thomas, P.J., McDowell, J.C., et al. 2003. Comparative analyses of multi-species sequences from targeted genomic regions. *Nature* **424:** 788–793.

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandells, M., Evans, C.A., Holt, R.A., et al. 2001. The sequence of the human genome. *Science* **291:** 1304–1391.

Weber, J.L. and Myers, E.W. 1997. Human whole-genome shotgun sequencing. *Genome Res.* **7:** 401–409.