OPEN ACCESS Freely available online

PLoS BIOLOGY

Community Page

Modernizing Reference Genome Assemblies

Deanna M. Church[1]*, Valerie A. Schneider[1], Tina Graves[2], Katherine Auger[3], Fiona Cunningham[4], Nathan Bouk[1], Hsiu-Chuan Chen[1], Richa Agarwala[1], William M. McLaren[4], Graham R.S. Ritchie[4], Derek Albracht[2], Milinn Kremitzki[2], Susan Rock[2], Holland Kotkiewicz[2], Colin Kremitzki[2], Aye Wollam[2], Lee Trani[2], Lucinda Fulton[2], Robert Fulton[2], Lucy Matthews[3], Siobhan Whitehead[3], Will Chow[3], James Torrance[3], Matthew Dunn[3], Glenn Harden[3], Glen Threadgold[3], Jonathan Wood[3], Joanna Collins[3], Paul Heath[3], Guy Griffiths[3], Sarah Pelan[3], Darren Grafham[3], Evan E. Eichler[5,6], George Weinstock[2], Elaine R. Mardis[2], Richard K. Wilson[2], Kerstin Howe[3], Paul Flicek[4], Tim Hubbard[3]

1 National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland, United States of America, **2** The Genome Institute at Washington University, St. Louis, Missouri, United States of America, **3** The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, United Kingdom, **4** The European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, United Kingdom, **5** Department of Genome Sciences, University of Washington School of Medicine, Seattle, Washington, United States of America, **6** Howard Hughes Medical Institute, Seattle, Washington, United States of America

The Rationale for the GRC

The availability of a high quality human genome assembly has revolutionized biomedical research. Genomics has now entered the realm of clinical genetics, with many groups using either whole genome sequencing [1,2] or whole exome sequencing [3] to identify variants underlying diseases and informing treatment options [4]. Advances in technology have increased the number of sequenced human genomes; however, de novo assembly of next generation sequencing reads is still problematic. The alignment of sequencing reads from these new genomes to a high quality reference genome remains a critical aspect of data interpretation [5].

While the human reference assembly is the highest quality mammalian assembly available, it is not without shortcomings. The "finished" assembly [6] contained over 300 gaps in the euchromatic portion of the genome, tiling path errors and regions represented by uncommon alleles. Furthermore, assessment of genome-wide variation revealed regions of the genome with complex, structurally diverse, allelic representations [7–9] that were insufficiently represented in the reference genome. Other analyses identified sequences that failed to align to the reference assembly either because the reference assembly contained a valid deletion allele or underrepresented multi-copy genes [10–13]. The Genome Reference Consortium (GRC) was formed to address these issues.

The GRC (the GRC consists of The Genome Institute at Washington University, The Wellcome Trust Sanger Institute, The European Bioinformatics Institute, and The National Center for Biotechnology Information) is an international consortium with expertise in genome mapping, sequencing, and informatics. The goal of the GRC is to provide high quality genome assemblies that will allow a user to place any sequence greater than 500 bp into a chromosome context. While this report focuses largely on recent GRC advances concerning the human reference assembly, the GRC is also responsible for the mouse and zebrafish reference assemblies. Continued improvement of the human reference assembly is critical as we move towards an era of clinical and personal genomics. The reference genomes of mouse and zebrafish are similarly critical in light of their importance as model organisms and the significant investments made in creating community resources such as gene knockout collections.

Assembly Management

Two major problems faced the GRC at the outset of this project, the decentralized nature of the Human Genome Project and the lack of a suitable data model for representing complex genomes. Much of the data underlying curation decisions had not been captured nor standardized. The human reference assembly had never been submitted to the International Nucleotide Sequence Database Collaboration (INSDC) [14] and thus lacked stable, trackable sequence identifiers that could be accessed from any INSDC database.

Initial efforts at assembling the human genome were guided by the concept of "a golden path" [15], a single clone tiling path that could be reduced to one non-redundant haploid representation of the human genome. While this model fit well with the prediction that single nucleotide variants (SNVs) would be the predominant source of variation in the population, it is now clear that structural variation is a much larger source of genomic diversity than previously recognized [16,17]. Additionally, this model did not deal robustly with sequences that were not part of chromosome assemblies. These often represent sequences that cannot be easily ordered or oriented on the chromosome

The Community Page is a forum for organizations and societies to highlight their efforts to enhance the dissemination and value of scientific knowledge.

Citation: Church DM, Schneider VA, Graves T, Auger K, Cunningham F, et al. (2011) Modernizing Reference Genome Assemblies. PLoS Biol 9(7): e1001091. doi:10.1371/journal.pbio.1001091

Published July 5, 2011

Funding: Support for this work came from the Intramural Research Program of the NIH, The National Library of Medicine, the European Molecular Biology Laboratory, the Wellcome Trust (grant number 077198), and the Howard Hughes Medical Institute (EEE). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Competing Interests: I have read the journal's policy and have the following conflicts: Paul Flicek is married to the deputy editor of *PLoS Medicine*, Melissa Norton. Evan Eichler is on the board of Pacific Biosciences.

Abbreviations: GRC, Genome Reference Consortium; INSDC, International Nucleotide Sequence Database Collaboration

* E-mail: church@ncbi.nlm.nih.gov

assembly due to structural complexity but frequently contain genes that may be of biological interest [18] or represent alternate haplotypes of regions in the chromosome assembly [9,19]. Earlier versions of the reference genome assembly included some of these allelic variants (such as at the MHC region) but the sequences themselves often were not used because they had no relation to the chromosome sequence and could not be easily distinguished from sequences reflecting biological or artificial duplication.

The GRC has addressed these problems by establishing common tools and standard operating procedures (SOPs) so that the genome assembly is now constructed in a regularized fashion. We have developed a single database to store all data underlying the genome assembly. Finally, we have developed a system to track individual regions that are under review. All of these data are made publicly available through our Web site (http://genomereference.org/).

Additionally, the GRC has formalized an assembly model (Figure 1 and Box 1) that provides for improved accounting for all sequences, including those that are not part of chromosome assemblies, and facilitates

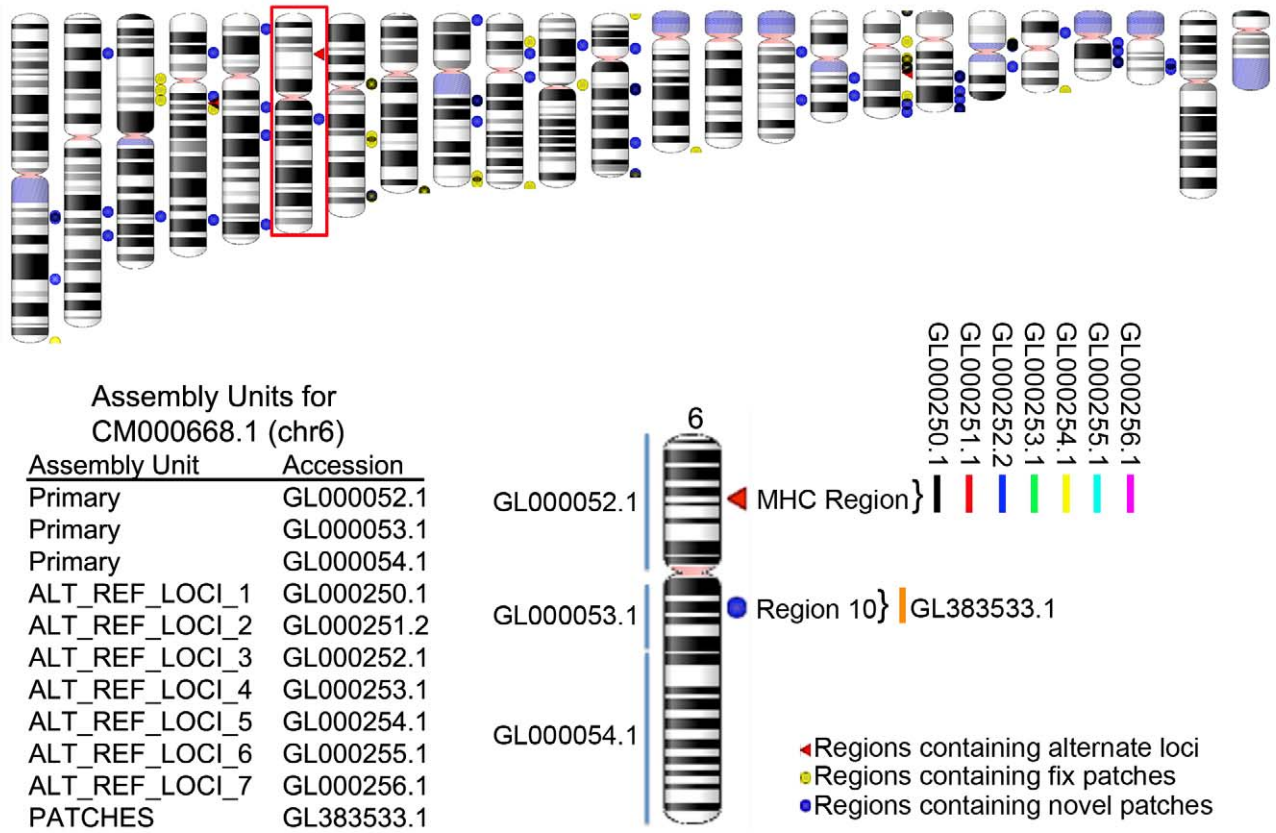

Figure 1. Assembly representation for GRCh37.p3. The top panel shows an ideogram representation of the human genome. The primary assembly unit contains sequences for the non-redundant haploid assembly; this includes the scaffolds that make up the chromosome sequence as well as unplaced and unlocalized scaffolds that are thought to represent novel sequence (not shown in this picture). Alternate loci and patches are placed in separate assembly units to facilitate annotation. Note the seven alternate scaffolds in the MHC region are all placed in different assembly units, as they all represent different representations of the same sequences. Other alternate loci can be added to these assembly units at the next major release if they don't overlap the existing alternates. All patches are placed in the PATCHES assembly unit and minor releases are cumulative such that the latest minor release will contain all patches. The red triangle, yellow circles, and blue circles represent regions that contain additional sequences that are not given actual chromosome coordinates, but rather are given a chromosome context via alignment to the primary assembly. The red triangles represent regions' alternate loci; these are sequences that provide an additional tiling path to the one given in the chromosome representation and are essential for representing structurally complex loci. The circles represent patch sequences; these are minor updates made to the assembly outside of the major build cycle. Yellow circles represent "fix" patches: regions of the chromosome assembly that will change with the next major assembly update. Blue circles represent "novel" patches: these are sequences that represent new alternate loci in the next major assembly update. Unlocalized and unplaced sequences are not represented in this figure. Sequences within the assembly are placed within containers known as assembly units. Note: a region can point to more than one type of extra chromosomal sequence; for example, a region could point to an alternate locus and to a fix or novel patch.
doi:10.1371/journal.pbio.1001091.g001

Box 1. Assembly Definitions

AGP: A file used to describe the instructions for building a contig, scaffold, or chromosome sequence. This file specifies the order, orientation, and switch points for each component.

Alternate Locus: A sequence that provides an alternate representation of a locus found in a largely haploid assembly. These sequences don't represent a complete chromosome sequence, although there is no hard limit on the size of the alternate locus; currently these are less than 5 Mb.

Assembly: A set of sequences (chromosomes, unlocalized, unplaced, and alternate loci) used to represent an organism's genome.

Assembly Unit: Collections of sequences used to define discrete parts of an assembly.

Component: The basic genomic level sequence used to construct the genome; typically these are clone sequences, Whole Genome Shotgun sequences, or PCR fragments. These sequences must be submitted to GenBank/EMBL/DDBJ.

Contig: A contiguous sequence generated from determining the non-redundant path along an ordered set of component sequences. A contig should contain no gaps.

Patch: A genome patch is a scaffold sequence that is part of a minor genome release. These sequences either correct errors in the assembly (a FIX patch) or add additional alternate loci (a NOVEL patch). These sequences allow us to update the assembly information without disrupting the chromosome coordinate system. FIX patches will be removed at the next major assembly release, as the changes will be rolled into the new assembly. NOVEL patches will be moved from the PATCHES assembly unit to a proper assembly unit.

Primary Assembly Unit: Represents the collection of sequences that, when combined, represent a non-redundant haploid genome.

Scaffold: An ordered and oriented set of contigs. A scaffold will contain gaps, but there is typically some evidence to support the contig order, orientation, and gap size estimates.

TPF: Tiling Path File; this provides the order of the component sequences that are used to build a higher order sequence (contig, scaffold, or chromosome).

Switch Point: The base at which the contig sequence stops being generated from one component sequence and switches to using the next component sequence. There must be at least one switch point between adjacent component sequences in a contig.

Unlocalized sequence: A sequence found in an assembly that is associated with a specific chromosome, but that cannot be ordered or oriented on that chromosome.

Unplaced sequence: A sequence found in an assembly that is not associated with any chromosome.

genome annotation by placing additional structure on those sequences. Structurally complex regions can be represented by more than one tiling path; one of which will be integrated into the chromosome assembly while the others will be instantiated as an independent sequence that, by alignment to the chromosome, provides the chromosome context for the alternate allele.

We have also introduced the concept of a "minor" assembly update, in the form of genome patches. This mechanism provides users with timely access to genome improvements without inducing frequent changes to the coordinate system upon which assembly annotations are based. Because genome patches take the same form as alternate loci the two forms of data can be similarly managed.

The release cycle for major assembly updates will not occur on a fixed schedule. In order to minimize the need for frequent re-annotation, major assembly updates will occur infrequently when we have produced at least 100 fix patches or affected >1% of the euchromatic sequence. The GRC will announce planned updates on their Web site at least 6 months in advance of any major assembly release. Additional, detailed information regarding major releases will be publicly announced via the Web site as data freeze dates approach. Minor assembly updates will be made quarterly.

Assembly Quality and Improvement

We have produced a major release of the human reference assembly, GRCh37, which was submitted in June of 2009 to the INSDC (GCA_000001405.1), and four minor assembly updates, with the last patch, GRCh37.p4 (GCA_000002405.5), released in April 2011. Detailed information concerning genome assembly construction is on our Web site (http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/info/index.shtml).

The top part of Figure 2 shows the distribution of issue types that were resolved for these assembly releases. Some assembly updates are relatively minor, involving the correction of a single nucleotide discrepancy in the assembly (e.g., HG-445; http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/issue_detail.cgi?id=HG-445) while others involved multiple components and required generation of new, region-specific tiling paths (e.g., HG-2; http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/issue_detail.cgi?id=HG-2). (Figure 2) [20].

While the model changes described above facilitated our assembly management and reporting, we also wished to investigate whether these updates would allow for improved genome analysis. To investigate this, we first tried to recover sequence identified as novel in a personal genome, theYH1 human assembly [12]. Roughly 25% could be placed in a chromosome context using GRCh37.p2 (see supplemental table 1 and supplemental figure 1 at http://www.ncbi.nlm.nih.gov/genome/assembly/grc/supplement/). The remaining sequences are being investigated to determine if they warrant inclusion in a future assembly release.

We also wished to investigate the impact on alignment of next generation sequencing reads. We selected two samples from the 1,000 Genomes project [21], NA12156 and NA12878, (SRA accessions ERX000125 and ERX000080, respectively) and aligned their reads to GRCh37, with and without

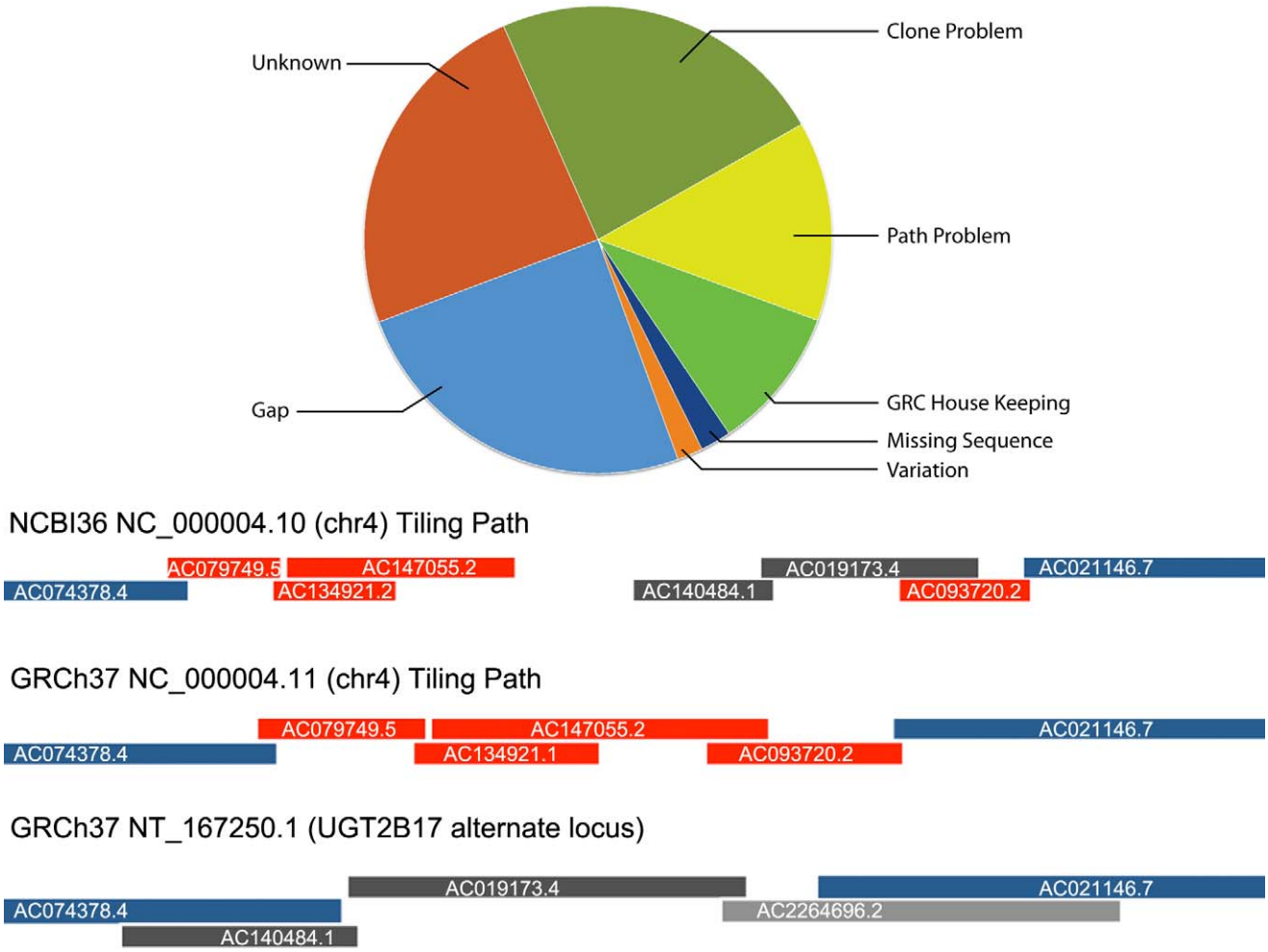

Figure 2. Distribution of issues addressed and an example region. (Top Panel) Issues for GRCh37, GRCh37.p1, and GRCh37.p2, broken down by type. Issue types are: Clone Problem: The issue is contained within a single clone. This may be a single nucleotide difference or a clone mis-assembly. Path Problem: There is evidence that the tiling path within a given region is incorrect and we will need to update the path. GRC Housekeeping: Changes use to help regularize the tiling path. Missing Sequence: Sequence that we can't yet place on the assembly. Mapping studies are ongoing to help place these sequences. Variation: There is evidence to suggest that complex variation is complicating a region and an alternate allele may need to be produced. Gap: The issue concerns filling a gap. Unknown: Issue is still under investigation for classification. (Bottom Panel) Details for issue HG-2, a Path Problem. The representation in NCBI36 was a mixed haplotype. The tiling paths for NCBI36 and GRCh37 are shown. Blue clones are anchor clones that are in NCBI36, the GRCh37 chr4 path, and the GRCh37 alternate locus path. Red clones represent the UGT2B17 insertion path and dark gray clones represent the UGT2B17 deletion path. The light gray clone was not used in NCBI36, but was used in GRCh37 to complete the alternate locus.
doi:10.1371/journal.pbio.1001091.g002

the alternate loci. We demonstrated that removal of the alternate loci leads to misalignment of approximately two-thirds of the alternate-locus specific reads (see supplemental table 2, supplemental figure 2 at http://www.ncbi.nlm.nih.gov/genome/assembly/grc/supplement/). These data clearly demonstrate that that inclusion of alternate representations for genomic loci can improve alignment quality and thus avoid spurious variation calls.

Policy Implications

We envision the high quality reference assemblies generated by the GRC having a long-term role in biomedical research because they most accurately capture all forms of human genetic variation and facilitate investigation of human disease in model organisms. With this in mind, we have built a reference assembly infrastructure to support transparent curation and assembly production. We have also updated the assembly model so that it better represents our current understanding of genome structure and diversity. We will use this model to encompass new discoveries and ultimately capture all significant variations in the human population structure as discovered through projects such as 1,000 genomes. Additionally, we wish to engage the research and clinical communities to identify regions that require targeted effort and to incorporate information from groups performing detailed work on specific loci. The GRC can only be truly successful with community input. Users can report problems directly to the GRC via our Web page (http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/ReportAnIssue.shtml).

It is difficult to overstate the importance of the human reference assembly, even in the age of personal genomics. Given current sequencing and assembly technology, there is a clear need for a high quality reference that can represent structural

diversity across all populations. Providing a representation of this diversity is critical for next generation sequence analysis. Even using an assembly with only three regions with alternative alleles, we show improved alignment quality and by extension variation calling, which is the primary product of personal genomics. More genomic alignment tools that can take the alternate representations into account need to be developed.

Understanding how genotype influences phenotype necessitates an accurate and complete picture of all loci in multiple populations. For many genomic regions, this can be denoted by a sequence with annotated SNPs and small indels, but other loci will require multiple sequence instances for complete representation. Some human loci, such as the 1q21 region, which remains misassembled in GRCh37.p2, are sufficiently complex that significant effort is needed to obtain even one correct sequence for the region. Additional work is required to sort out the haplotypes segregating among various populations, many of which contribute to phenotypes associated with multiple developmental disorders [22].

While assemblies using next generation sequencing are beginning to approach the quality of long-read Whole Genome Shotgun assemblies [23], they continue to fail in complex regions. While it is likely that sequencing and assembly technology will improve such that de novo assembly of individual genomes will approach the quality of the human reference, it is not clear when this will happen. However, even when this is a common occurrence, we see a role for the GRC in integrating the data from thousands of human genomes to produce a "gold-standard" reference assembly. We anticipate a continued need for a high quality reference assembly that will allow any human sequence to be placed into a chromosome context quickly and easily. As we march down the path of personal genomics it is critical that we devote resources to the current reference assembly in order to support clinical applications. As we continue to understand how genotype influences phenotype, the best possible reference assembly available must be made available to the research community.

Acknowledgments

The GRC would like to acknowledge the following contributors to this project: David C. Schwartz, Jane Rogers, Mario Caccamo, Paul Kitts, Michael DiCuccio, Françoise Thibaud-Nissen, Avi Kimchi, Jonathan Mudge, Richard Clark, Andrew Dearlove, Michelle Smitth, Britt Kilian, Karen McLaren, James Gilbert, Laurens Wilming, Darren Ware, Sharmin Begum, Karen Davey, Diana Kidger, Kim Brugger, Tony Gaige, and Jason Walker.

References

1. Ley TJ, Mardis ER, Ding L, Fulton B, McLellan MD, et al. (2008) DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. Nature 456: 66–72.
2. Pelak K, Shianna KV, Ge D, Maia JM, Zhu M, et al. (2010) The characterization of twenty sequenced human genomes. PLoS Genet 6: e1001111. doi: 10.1371/journal.pgen.1001111.
3. Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, et al. (2009) Targeted capture and massively parallel sequencing of 12 human exomes. Nature 461: 272–276.
4. Lupski JR, Reid JG, Gonzaga-Jauregui C, Rio Deiros D, Chen DCY, et al. (2010) Whole-genome sequencing in a patient with Charcot-Marie-Tooth neuropathy. N Engl J Med 362: 1181–1191.
5. Alkan C, Sajjadian S, Eichler EE (2011) Limitations of next-generation genome sequence assembly. Nat Methods 8: 61–65.
6. International Human Genome Sequencing Consortium (2004) Finishing the euchromatic sequence of the human genome. Nature 431: 931–945.
7. Kidd JM, Cooper GM, Donahue WF, Hayden HS, Sampas N, et al. (2008) Mapping and sequencing of structural variation from eight human genomes. Nature 453: 56–64.
8. Schmutz J, Martin J, Terry A, Couronne O, Grimwood J, et al. (2004) The DNA sequence and comparative analysis of human chromosome 5. Nature 431: 268–274.
9. Zody MC, Jiang Z, Fung H, Antonacci F, Hillier LW, et al. (2008) Evolutionary toggling of the MAPT 17q21.31 inversion region. Nat Genet 40: 1076–1083.
10. Kidd JM, Sampas N, Antonacci F, Graves T, Fulton R, et al. (2010) Characterization of missing human genome sequences and copy-number polymorphic insertions. Nat Methods 7: 365–371.
11. Itsara A, Wu H, Smith JD, Nickerson DA, Romieu I, et al. (2010) De novo rates and selection of large copy number variation. Genome Res 20: 1469–1481.
12. Li R, Li Y, Zheng H, Luo R, Zhu H, et al. (2010) Building the sequence map of the human pan-genome. Nat. Biotechnol 28: 57–63.
13. Sudmant PH, Kitzman JO, Antonacci F, Alkan C, Malig M, et al. (2010) Diversity of human copy number variation and multicopy genes. Science 330: 641–646.
14. Brunak S, Danchin A, Hattori M, Nakamura H, Shinozaki K, et al. (2002) Nucleotide sequence database policies. Science 298: 1333.
15. Kent WJ, Haussler D (2001) Assembly of the working draft of the human genome with GigAssembler. Genome Res 11: 1541–1548. doi:10.1101/gr.183201.
16. Mefford HC, Eichler EE (2009) Duplication hotspots, rare genomic disorders, and common disease. Curr Opin Genet Dev 19: 196–204.
17. Feuk L, Carson AR, Scherer SW (2006) Structural variation in the human genome. Nat Rev Genet 7: 85–97.
18. Doggett NA, Xie G, Meincke LJ, Sutherland RD, Mundt MO, et al. (2006) A 360-kb interchromosomal duplication of the human HYDIN locus. Genomics 88: 762–771.
19. Antonacci F, Kidd JM, Marques-Bonet T, Teague B, Ventura M, et al. (2010) A large and complex structural polymorphism at 16p12.1 underlies microdeletion disease risk. Nat Genet 42: 745–750.
20. Xue Y, Sun D, Daly A, Yang F, Zhou X, et al. (2008) Adaptive evolution of UGT2B17 copy-number variation. Am J Hum Genet 83: 337–346.
21. Durbin RM, Abecasis GR, Altshuler DL, Auton A, Brooks LD, et al. (2010) A map of human genome variation from population-scale sequencing. Nature 467: 1061–1073.
22. Mefford HC, Sharp AJ, Baker C, Itsara A, Jiang Z, et al. (2008) Recurrent rearrangements of chromosome 1q21.1 and variable pediatric phenotypes. N Engl J Med 359: 1685–1699.
23. Gnerre S, Maccallum I, Przybylski D, Ribeiro FJ, Burton JN, et al. (2011) High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proc Natl Acad Sci U S A 108: 1513–1518.