Impacts of Updating the Reference Assembly on Genome Interpretation

Contact: deanna.church@personalis.com

Deanna M. Church, Jason Harris, Stephen Chervitz, Gabor Bartha, Shujun Luo, Mirian Karbelashvili, Ming Li, Amy Huang, Parin Sripakdeevong, Scott Kirk, Michael Clark, Sarah Garcia, Mark Pratt, John West, and Richard Chen

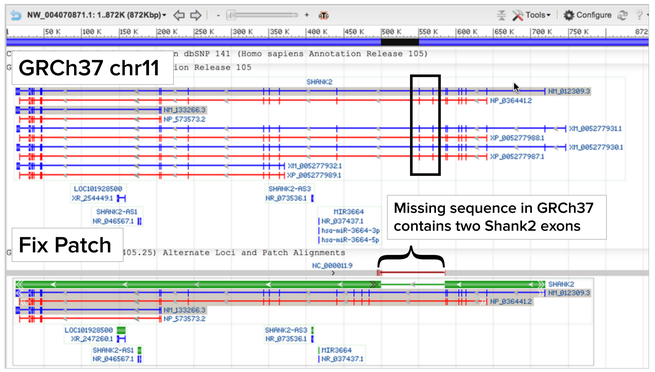
Personalis, Inc. | 1350 Willow Road, Suite 202 | Menlo Park, CA 94025

The Path to GRCh38

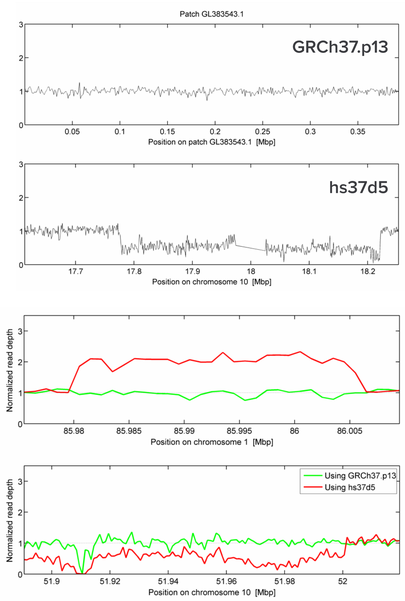
GRCh38 Lacks Deep Annotation

GRCh37 is the coordinate system that has been used by most major projects in the last 4 years (1000 Genomes, GO-ESP, etc) and is the standard in most clinical and research labs. Thus, while analytical validity may be improved by GRCh38, it is difficult to interpret the data without other biological context. We are taking an approach that lets us take advantage of some GRCh38 improvements in the context of GRCh37 knowledge. We are investigating the usage of the 131 Fix patches released by the GRC.

Fix Patch Content



The 131 Fix patches in GRCh37.p13 have gene annotation from NCBI and Ensembl and contain 652 genes of biomedical interest (NCBI annotation 105). While not all of these will be substantially different between the patch and the GRCh37 chromosome, many of these fix assembly errors. The above figure shows the correction of the SHANK2 gene (associated with 20 genetic tests in GTR) which is missing 2 exons on the GRCh37 chr11 representation but is corrected in the Fix patch.



Preliminary Data

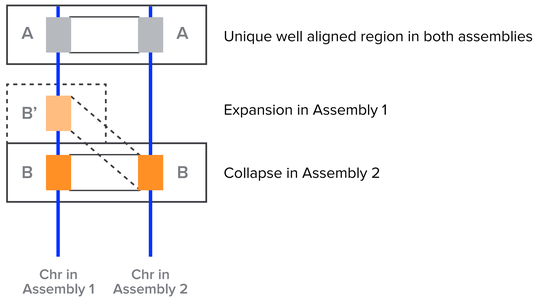
Our approach involves redacting the chromosome sequence to force read alignments to the patch. The top panel shows normalized read depth for a region on GRCh37 chr10 that is very rearranged in GRCh38. The top panel shows the alignment to GRCh37.p13 while the bottom panel shows the alignment to the equivalent region in hs37d5 (the 1000 genomes reference). The bottom panels show two different regions that don't overlap patch regions, demonstrating the patch inclusion can improve alignments across the assembly, not just in patch regions.

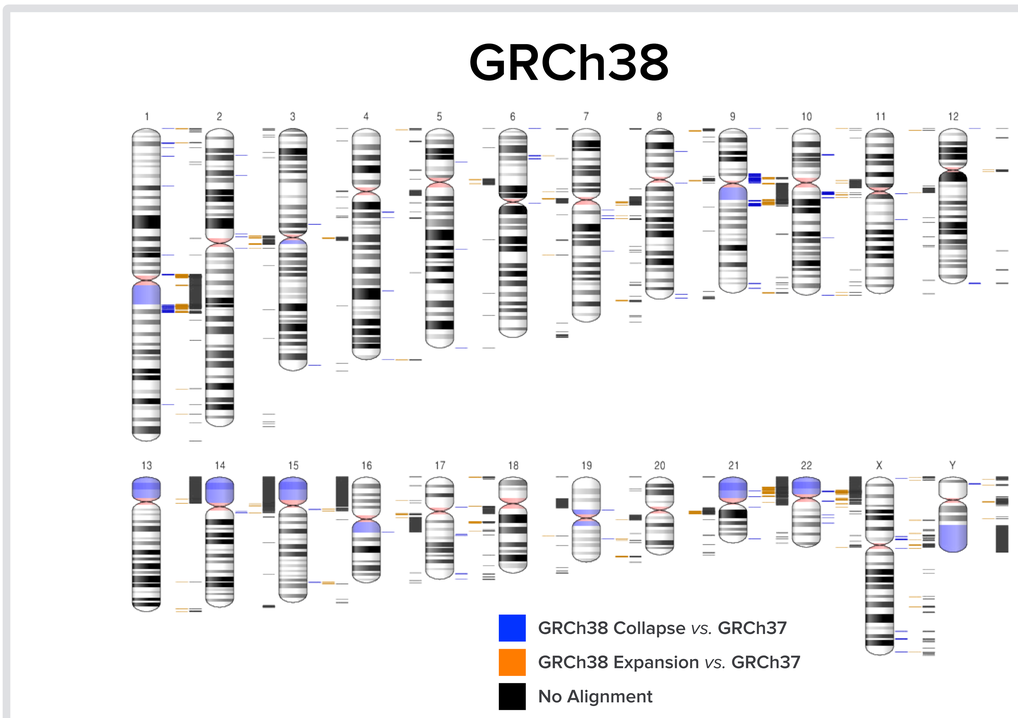
Assembly Update

GRCh38 was Released in December 2013

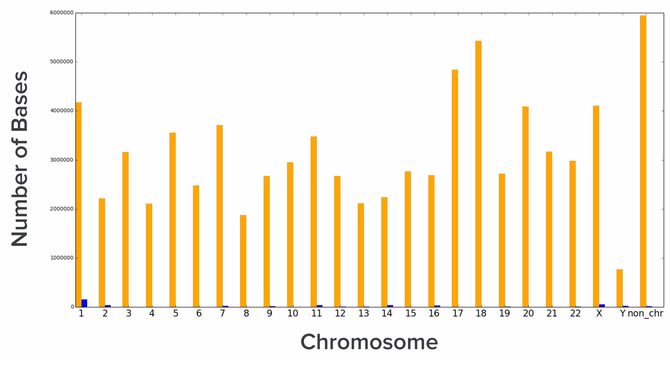
There as been some *de novo* analysis performed, such as gene annotation, but much of what we know comes from the NCBI Remap process. The process aligns the two assemblies using a procedure that allows us to more robustly capture regions of expansion and contraction that often occur at loci with complex repeats or structural allelic diversity. The cartoon to the right illustrates expansion and collapse in the assembly alignments.

http://www.ncbi.nlm.nih.gov/genome/tools/remap





Bases in GRCh38 with No Alignment to GRCh37



Novel Sequence in GRCh38

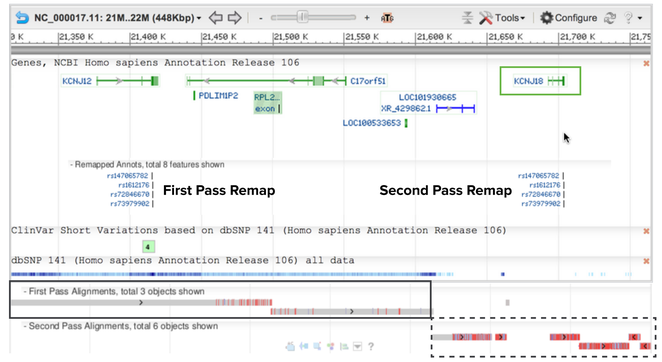
Several megabases of new sequence has been added to GRCh38. The graph to the left shows the amount of sequence (bp) in each assembly with no alignment to the other assembly. While there is a small amount of sequence in GRCh37 with no equivalent in GRCh38, this is dwarfed by the sequence added by the new assembly. Note the 'non_chr' column. This is largely sequence by alternate loci. This sequence is distinct from the 'expansions' shown in the ideogram.

For more assembly information
http://genomereference.org

Old Data in a New Light

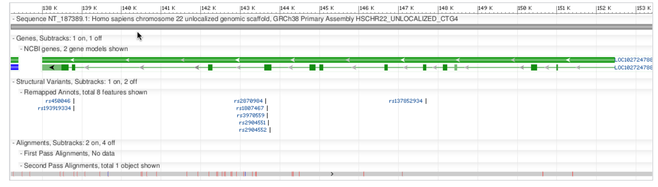
Using NCBI Remap to Move Annotation

In an effort to understand GRCh37 content in the context of GRCh38 we have used the NCBI remap program to map content from ClinVar. At the time of remapping, there were 55,461 unique positions in the ClinVar VCF file. Eleven of these failed to remap, but manual review of these indicated this was an error in the alignment process. Of interest are the 100 ClinVar variants that are in regions of collapse in GRCh37 that have been expanded in GRCh38. These variants need to be re-evaluated as they might identify paralogous sequence variants rather than allelic sequence variants.



GRCh38 chr17: KCNJ12-KCNJ18 Region

KCNJ18 was missing in GRCh37. ClinVar variants annotated in KCNJ12in GRCh37 map both to KCNJ12 and KCNJ18 (boxed in green) in GRCh38.Of note is the fact that KCNJ18 has a phenotype association in OMIM whileKCNJ12 does not. Both the validity of these variants as well as the validity of this association may need to be re-evaluated in light of the new assembly.



PRODH Paralog Added

The above figure shows the secondary mapping of a cluster of ClinVar variants in the PRODH gene. Of note the sequence is unlocalized, though associated with chromosome 22, underscoring the importance of analyzing the entire assembly rather than just the chromosome sequences.

Further Information from Personalis

Wednesday, June 11

3:30 – 5:00 PM

Panels vs. Exomes: An Interactive Panel Discussion

Table 5

Garcia *et al.*, The Clinical Exome: Personalis' Experience Using an Enhanced Exome and Genome-wide Structural Variant Detection for the Diagnosis of Diseases of Unknown Genetic Etiology

Chen *et al.*, Top 10 Misconceptions of Clinical NGS Sequencing

Personalis®
Pioneering Genome-Guided Medicine

POS-GIDC-1.0 June 2014

www.personalis.com | info@personalis.com | +1 855-436-6634 | +1 650-752-1300