Reinterpreting Variation in Light of GRCh38

Personal Genomes 2014

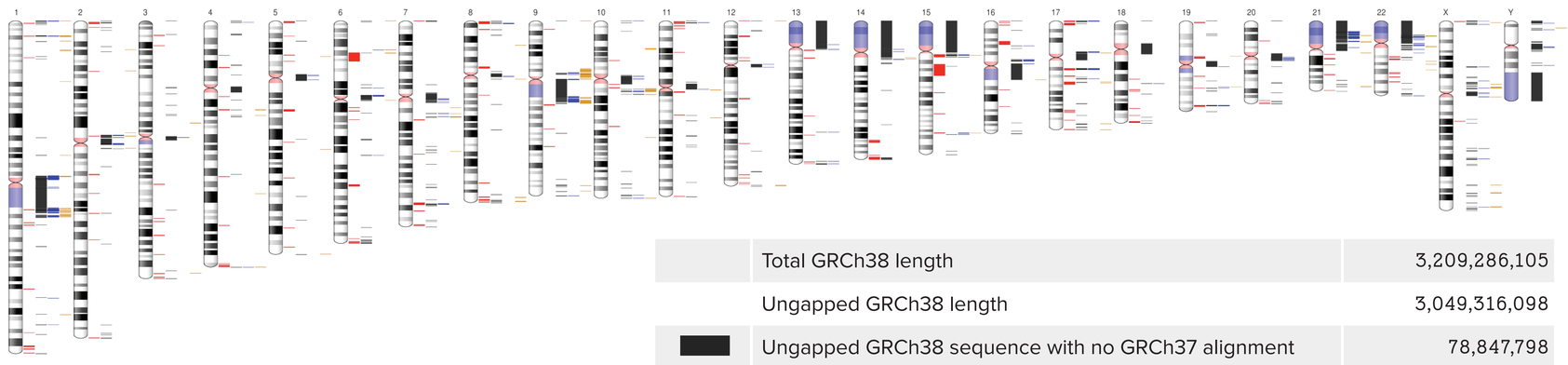
Deanna M. Church, Jason Harris, Stephen Chervitz, Gabor Bartha, Anil Patwardhan, Scott Kirk, Michael J. Clark, Sarah Garcia, John West and Richard Chen

Contact: deanna.church@personalis.com

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

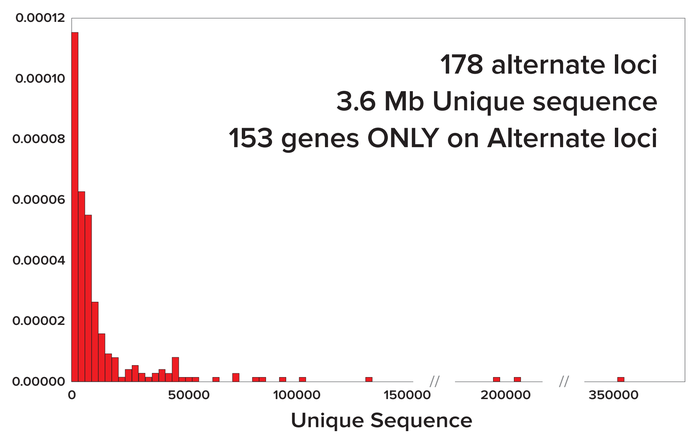
GRCh38 *vs.* GRCh37

There where many changes to GRCh38 relative to GRCh37. The change most likely to affect our understanding of variation is the addition of paralogs missing in GRCh37 and the removal of false duplication present in GRCh37 due to haplotype expansion in GRCh37.



- Regions with alternate loci
- GRCh38 regions with no GRCh37 alignment
- GRCh38 regions expanded compared to GRCh37
- GRCh38 regions collapsed compared to GRCh37

	Total GRCh38 length	3,209,286,105
	Ungapped GRCh38 length	3,049,316,098
	Ungapped GRCh38 sequence with no GRCh37 alignment	78,847,798
	GRCh38 regions expanded compared to GRCh37	16,630,147
	GRCh38 regions collapsed compared to GRCh37	7,759,372



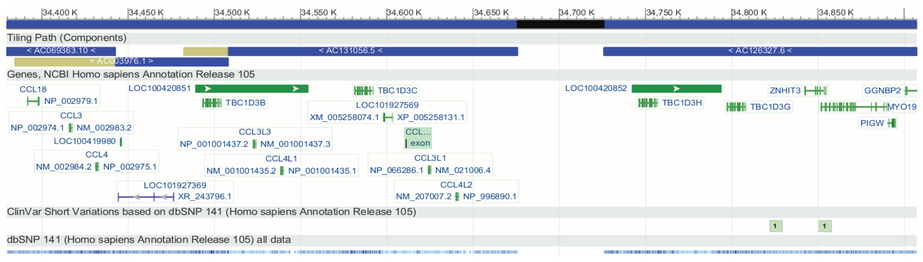
Contribution of unique sequence per alternate locussequence. There is a wide contribution of unique sequence per locus.

Distribution of genomic regions containing novel, expanded or collapsed sequence and alternate loci in GRCh38 when compared to GRCh37. Common used analysis tools and reporting formats do not support the alternate loci robustly. There is an effort to improve tools to use the full GRCh38 sequence; https://github.com/GenomeRef/SoftwareDevTrack.

Accurate Representation of Genomic Regions

Structural variation and complex repeats complicates assemblies. When using an assembly model that requires the production of a single, haploid consensus sequence you can often end up with sequence representations not present in any human. To produce GRCh38, the GRC retiled many such complex regions using a single haplotype resource (CHM1), in many cases producing a single, high quality representation on the chromosome. Often an additional high quality representation is also available as an alternate locus (red ticks in above diagram) at these regions.

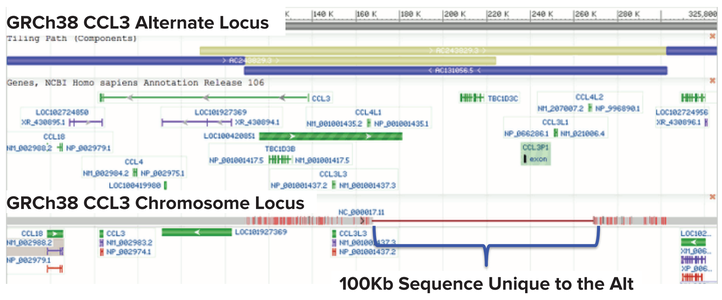
GRCh37



The CCL3 region contains a polymorphic 90Kb repeat. In GRCh37 the region was constructed from clones from 4 different donors leading to a false gap, haplotype expansion and a genomic structure not likely present in anyone.

Re-tile Region

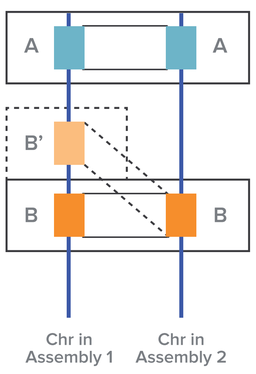
GRCh38



The CCL3 chromosome region was retiled in a single haplotype resource (CHM1) producing a deletion allele. An insertion allele containing 100Kb of sequence not found on the chromosome was added as an alternate locus.

Using GRCh38 to Improve GRCh37 Analysis

While *de novo* gene annotation has been performed on GRCh38, no large-scale projects have used this for variant calling yet. To start to get a picture of variation on GRCh38, we've used the NCBI remap tool to project variant locations from GRCh37->GRCh38.



Unique well aligned region in both assemblies

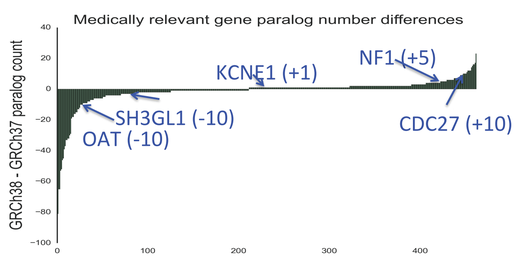
Expansion in Assembly 1

Collapse in Assembly 2

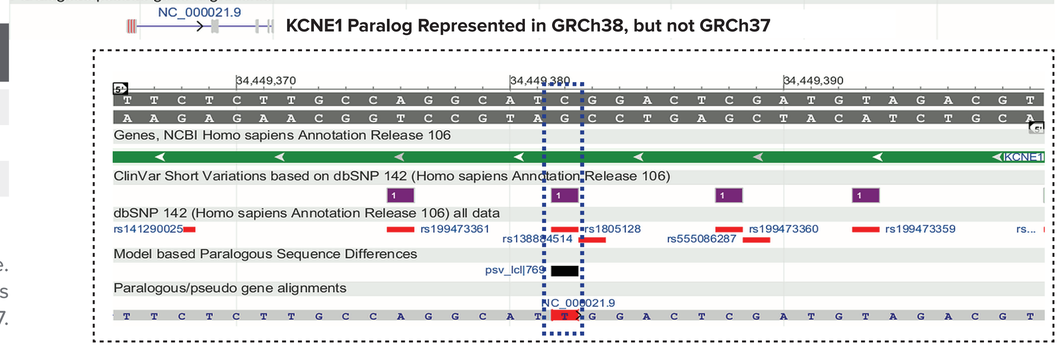
The NCBI assembly alignment process is optimized to identify paralogous sequence expansion and collapse when comparing two assemblies.

Dataset	Starting Loci	Failure	Unique to Primary	Unique to Alts	Collapse in GRCh37	Collapse in GRCh38
GWAS catalog	7,991	0	7,827	0	14	0
ClinVar	88,343	3	86,549	5	278	4
GO-ESP 6500	1,982,177	180	1,920,864	339	5,792	324
GIAB	2,915,713	274	2,874,786	47	1,662	4

Feature remapping of commonly used variant sets. ClinVar is based on the Sept, 2014 release. Remap alignments were from Sep 20, 2014 using software version 1.7. The highlighted columns shows variants that are candidate false positive variant calls due to missing paralogs in GRCh37. 88 of the 278 such ClinVar variants are annotated as pathogenic.



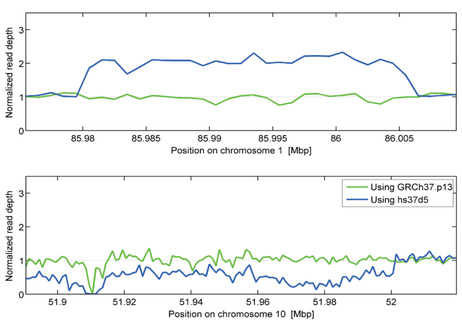
Global assessment of paralogous sequences in GRCh37 and GRCh38. The graph above shows the difference between the two for a set of biomedically relevant genes. Genes on the left have lost paralogs in GRCh38, while genes on the right have gained paralogs in GRCh38. Some of the new paralogs are only present on alternate loci. (NCBI annotation 105 and 106)



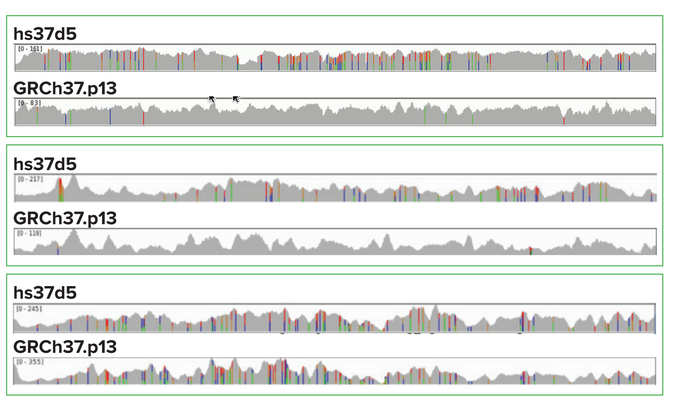
KCNE1 (top) is associated with Long-QT syndrome. All 35 pathogenic variants in ClinVar are in the terminal coding exon. The zoomed in view shows the overlap of a pathogenic variant with a paralogous sequence difference.

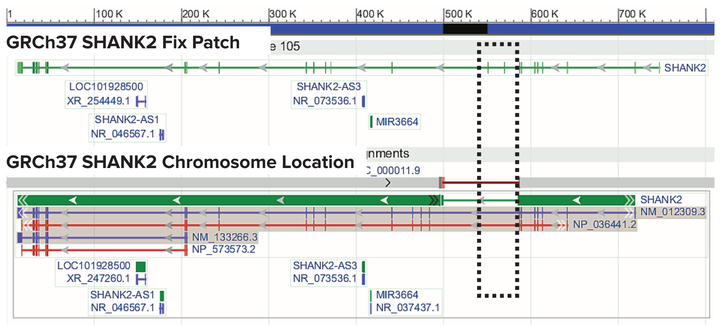
Improving GRCh37 Analysis

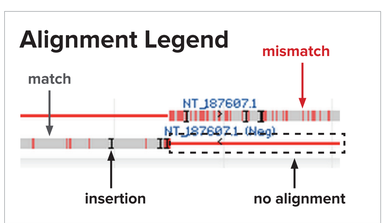
- Identify genes with missing paralogs
- Mark up variants that map to more than one location in GRCh38
- Use fix patches to incorporate some sequence improvements when using GRCh37 as a reference.



In addition to improvements in alignments within FIX patch regions we also see improved alignments outside of FIX patch regions as shown in the normalized read depth plots above. This often leads to improved variant calls (see the IGV plots to the right).







While working towards GRCh38, the GRC released patches representing fixedregions of the genome (FIX-patches). Tothe left is a patch for the SHANK2 genethat is missing 2 coding exons in GRCh37 representation. We are building a pipelineto use the FIX patches. This allows us to get more out of GRCh37 while we work towards tools for GRCh38. Additionally, the GRC has just released the first 13 fixpatches for GRCh38. To implement this pipeline, we redact the incorrect chromosome sequence to force read alignments to the correct fix-patch sequence, thus allowing us to avoid theduplicate sequence problem.

GRC: http://genomereference.org
CHM1: http://dx.doi.org/10.1101/006841
NCBI Remap: http://www.ncbi.nlm.nih.gov/genome/tools/remap