Challenges of Moving a Clinical Lab to GRCh38

AGBT 2015

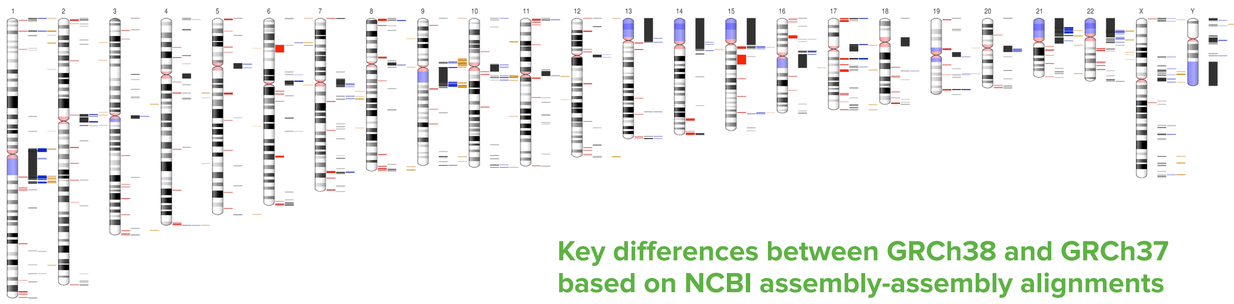
Deanna M. Church, Jason Harris, Stephen Chervitz, Gabor Bartha, Anil Patwardhan, Scott Kirk, Michael J. Clark, Sarah Garcia, John West, and Richard Chen

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

Contact: deanna.church@personalis.com

Motivation

The human reference assembly is one of the most important tools used for genome interpretation. GRCh38 represents the culmination of four years of curation and improvement by the Genome Reference Consortium (GRC). There are several areas of improvement, including the addition of novel sequence, correction of several megabases of mis-assembled sequence and addition of hundreds of alternate loci. Improvements suggest this will be a better substrate for genome analysis.

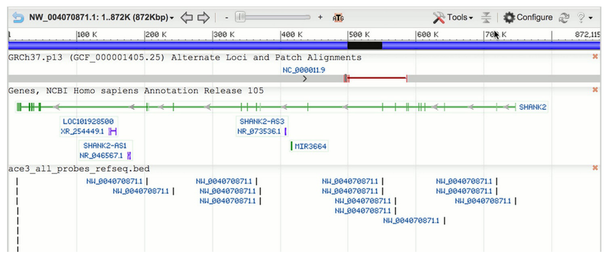


Regions with alternate loci
GRCh38 regions with no GRCh37 alignment
GRCh38 regions expanded compared to GRCh37
GRCh38 regions collapsed compared to GRCh37

Key differences between GRCh38 and GRCh37 based on NCBI assembly-assembly alignments

- 178 Regions containing alternate loci
- >75 Mb of novel sequence
- Several megabases of expanded and contracted paralogous

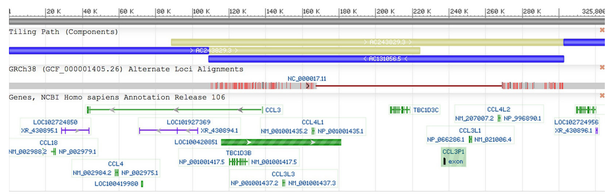
Improved Gene Representation



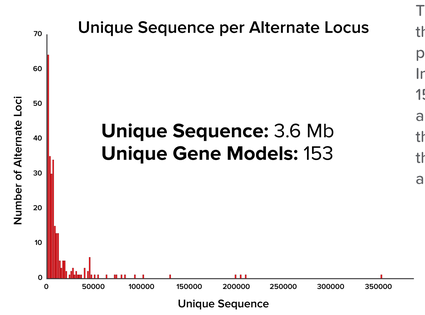
The figure above shows an improvement to the SHANK2 gene. The blue line represents the sequence in a FIX patch. The grey line is an alignment to GRCh37 chr. 11. The thin red line indicates sequence in the patch not found in the chromosome. The new sequence contains two coding exons. A significant focus was put on improving clinically relevant genes.

>2,000 New Coding Transcripts on GRCh38

More Alternate Loci



The figure above shows an alternate representation at the CCL3 locus (sequence in blue). The alignment to the chromosome shows 100 Kb of sequence novel to the alternate locus. This sequence contains genes not present on the chromosome (shown in green).



The figure to the left plots the unique sequence per alternate locus. Importantly, there are 153 genes found only on alternate loci, underscoring the importance of including these sequences in analysis.

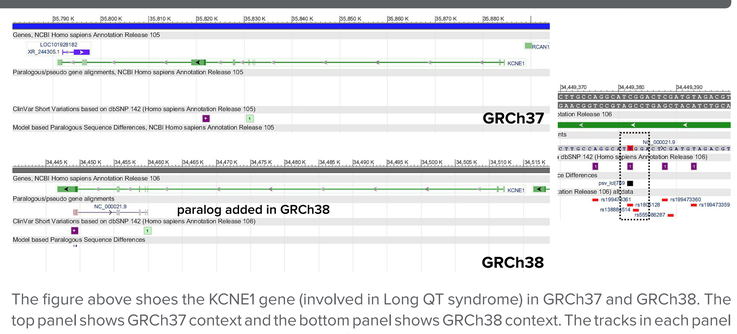
Analysis

Assembly Impact on Variant Interpretation

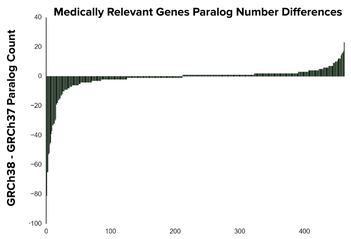
Dataset	Starting loci	No remap*	Unique remap	Map to Primary and Alt	Collapse in GRCh37	Collapse in GRCh38
GWAS catalog	7,994	3	7,861	121	8	1
ClinVar (Aug 2014)	88,265	14	86,896	1,204	147*	4
GO-ESP 6500	1,982,177	623	1,937,085	41,620	2,174	675
GIAB	2,865,730	390	2,837,685 (542 alt only)	27,655	1,191	130

*27 Pathogenic

Large-scale reference projects have not undertaken variant calling on GRCh38. In order to understand the genomic context of reference variant sets, we used the NCBI Remap service to project small variants onto GRCh38. The table above summarizes these results. Of note, a small percentage of variants fail to remap and some variants now only map to alternate loci. Of greater concern are variants that map to regions of collapse in one assembly. These are regions that have changed due to a change paralogous gene content and these variants are candidates for being false positives.

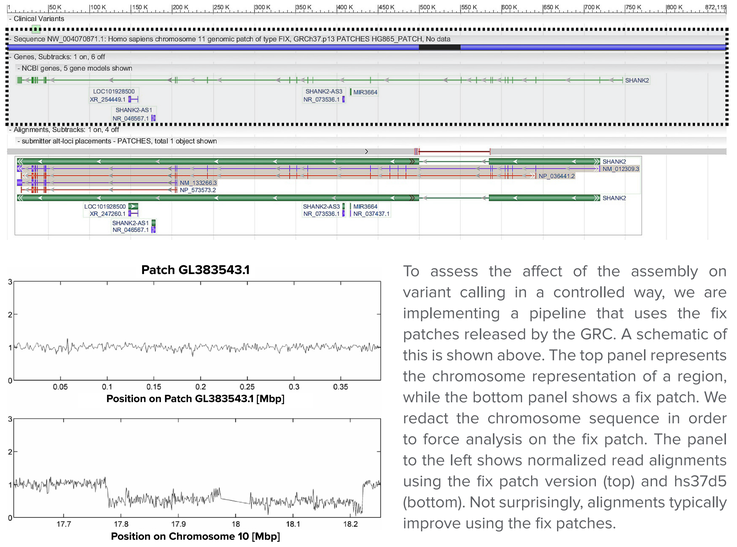


The figure above shoes the KCNE1 gene (involved in Long QT syndrome) in GRCh37 and GRCh38. The top panel shows GRCh37 context and the bottom panel shows GRCh38 context. The tracks in each panel are 'gene', 'paralogous sequence alignment', 'ClinVar' and 'predicted paralogous sequence variants'. The addition of the new paralog complicates variant interpretation. The upper panel to the right shows a zoomed in view of the terminal coding exon, where all of the clinical variants are annotated. Here it is clear that some called variants overlap with locations of paralogous sequence variants.

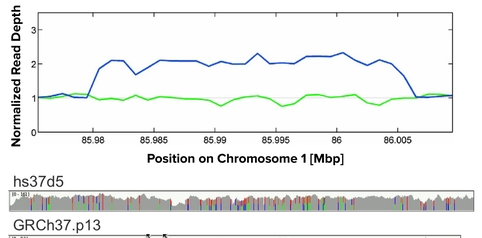


The panel to the left shows the number of paralogs a gene has in GRCh38 subtracted from the number of paralogs a gene has in GRCh37. Data shown are restricted approximately 8000 medically relevant genes. Note some genes lose paralogs, which were likely due to haplotype expansion, making these genes easier to analysis. To the right are genes that had missing paralogs in GRCh37, making them more susceptible to false positive calls.

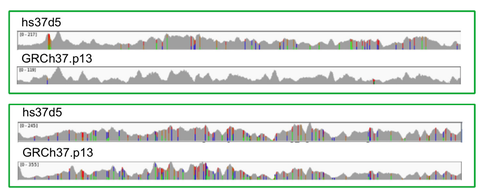
Assembly Impact on Variant Detection



To assess the affect of the assembly on variant calling in a controlled way, we are implementing a pipeline that uses the fix patches released by the GRC. A schematic of this is shown above. The top panel represents the chromosome representation of a region, while the bottom panel shows a fix patch. We redact the chromosome sequence in order to force analysis on the fix patch. The panel to the left shows normalized read alignments using the fix patch version (top) and hs37d5 (bottom). Not surprisingly, alignments typically improve using the fix patches.

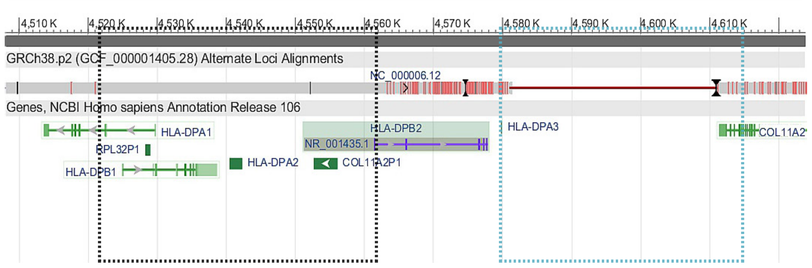


Promisingly, regions outside of fix patches can also show alignment/variant identification improvement. The panel above shows normalized read alignment, with hs37d5 in blue and the fix patch version in green. The IGV plots for each assembly are also shown. Below are IGV plots for additional regions, showing examples of clear alignment improvement (top) and lack of improvement (bottom).



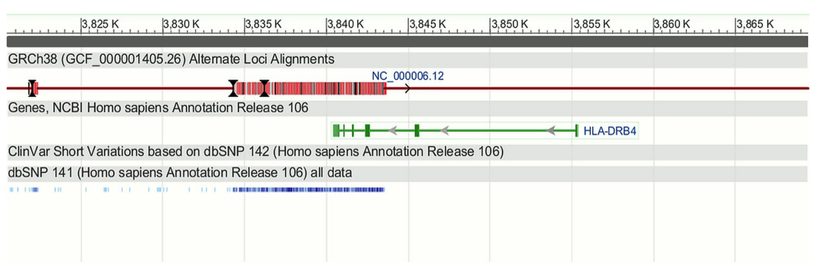
Challenges

Alignment



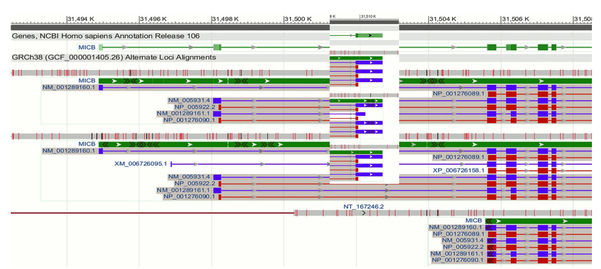
Most modern aligners can't distinguish paralogous duplication from allelic duplication. The box above shows a region of an alt-locus aligned to the chromosome at the HLA region. The black dotted box identifies a region of low diversity between the two alleles where reads are likely to align to both the alt and to the chromosome. Many aligners will lower the mapping score for these reads. The blue box highlights a region unique to the alt, so most reads will aligner here uniquely, but additional work needs to happen to ensure genotypes are represented correctly. BWA-Mem and SRPRism represent two alt-aware aligners, but work needs to be done to understand the implication for variant calling and genotype reconstruction.

Annotation



Alternate locus and patch sequences often contain gene annotation (thanks to NCBI and Ensembl!) but these sequences often have sparse or no variant annotation as they have not typically been included in common variant identification analyses.

Data Representation



Common annotation formats cannot handle allelic duplication. The panel to the left shows the MICB gene, which has multiple allelic copies in the assembly. To represent this in a GFF unique IDs must be fabricated as there is an expectation that a feature will have only 1 location in an assembly.

```
ID=gene13336;Name=MICB;Dbxref=GeneID:4277
ID=gene42005;Name=MICB;Dbxref=GeneID:4277
ID=gene43669;Name=MICB;Dbxref=GeneID:4277
ID=gene44377;Name=MICB;Dbxref=GeneID:4277
ID=gene44827;Name=MICB;Dbxref=GeneID:4277
ID=gene45127;Name=MICB;Dbxref=GeneID:4277
```


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

POS-AGDC-1.0 FEB 2015

Personalis®
Pioneering Genome-Guided Medicine