Personalis®

Personalis, Inc., Menlo Park, CA

Impacts of updating the reference assembly on genome interpretation

Deanna M. Church, Valerie A. Schneider[1], Sarah Garcia, Mark Pratt, John West and Richard Chen.

[1]. National Center for Biotechnology Information, Bethesda, MD

Contact
deanna.church@personalis.com

Gene and Clinical Variation

References: NCBI Eukaryotic Annotation Pipeline

Robust gene representation in the reference assembly is a critical aspect for analyzing whole genome and whole exome data. Using gene annotation data from the NCBI eukaryotic annotation pipeline, we looked at Gene IDs that are not represented at all in GRCh37 or were partial gene annotations in GRCh37, but complete in GRCh38.

4,989 Gene IDs in GRCh38 that are not in GRCh37 **2,074** Gene IDs in GRCh38 that are partially in GRCh37

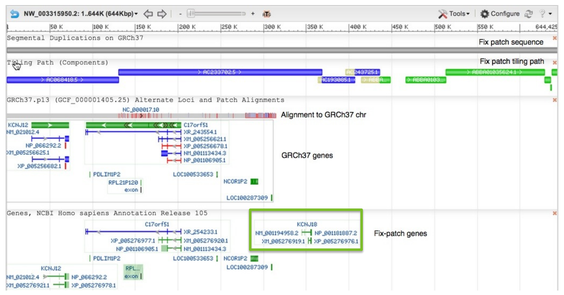


Figure 1. KCNJ18 (mutations in this gene have been associated with thyrotoxic hypokalemic periodic paralysis) is completely missing in GRCh37, but present in GRCh38. The above figure shows an alignment of a fix patch containing KCNJ18 that is now part of GRCh38.

Thousands of genes are added, completed or better represented In GRCh38

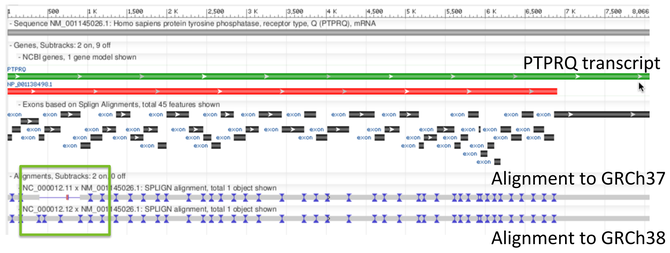


Figure 2. Some gene fixes are more subtle. PTPRQ, associated with autosomal recessive deafness, contained an internal deletion in GRCh37 due to an error in assembling the underlying clone sequence. This has been fixed in GRCh38 by inserting a small piece of the HuRef assembly. The green box highlights the discrepant region.

Centromere Models

References: Miga et al, Genome Research, 2014

The repetitive nature and structural complexity of centromeres meant that These important biological structures were represented by 3 million 'N's in Previous assembly versions. Miga et al. were able to construct graph-based models of the alpha satellite sequences in the HuRef assembly (Figure 6).

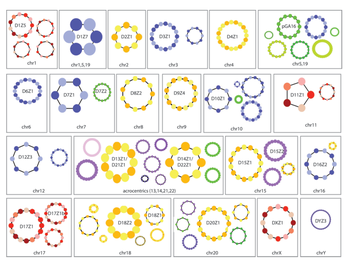


Figure 6. Graphical representations of the centromere graphs constructed by Miga et al.

The graphs represent a model of the relationship between alpha-satellite monomers. But this graph like structure cannot be incorporated into the reference assembly. To do this the graphs had to be linearized into FASTA sequences that represents the average of both autosomes in the HuRef genome. The FASTA could then be used to construct the chromosome sequences that help make up GRCh38. Figure c2 shows an example of this.

Of note, the sequences of some chromosomes could not be separated (chr5 and chr19 collapse into one model; the acrocentric chromosomes collapse into a separate model). This means that some of the model sequence is represented more than one time in the Primary reference assembly. This is a situation that is similar to the Pseudoautosomal region (PAR).

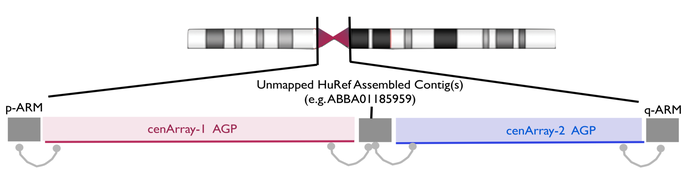
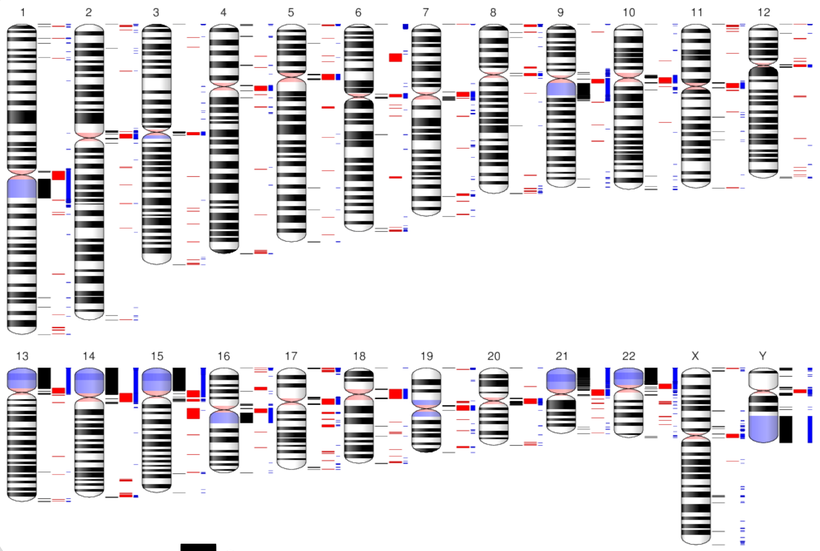


Figure 7. An example of the linearization and integration of a model centromere sequence for GRCh38. Of interest in the above figure is the inclusion of a contig that was previously unplaced in the HuRef assembly. The inclusion of these sequences are often supported by other lines of mapping evidence and show additional utility for including the centromere models.

The inclusion of centromere models allows for improved study of centromeric regions and should improve alignments of NGS sequences containing alpha satellite.

GRCh38



Human Specific Sequences

References: Doggett et al, Genomics 2006
Dennis et al, Cell 2012

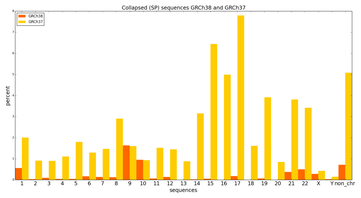


Figure 3. Global assessment of collapsed sequence in GRCh37 and GRCh38. A collapse occurs when different paralogs assembly together due to their sequence identity. Regions of collapse occur in GRCh38 when artificial duplication was removed from GRCh37.

Human specific sequences occur due to genomic duplication that has occurred only in the human lineage. These sequences show a high degree of sequence identity between paralogs and are difficult to sequence and assemble. These sequences are often involved in the expansion and contraction of sequences between assembly builds However, many of these sequences have important biological functions and some have been associated with disease. Missing paralogs may lead to false positive variant identification.

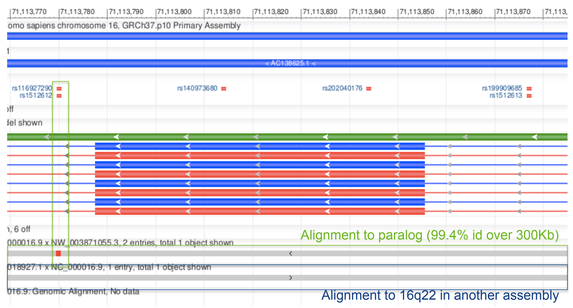


Figure 4. The Hydin locus at 16q22 has been associated with Primary Ciliary Dyskinesia (OMIM: 610812). There has been a human specific partial duplication of this gene, with the paralog residing in 1q21. The chromosome 1 and chromosome 16 loci share 99.4% identity over 300 Kb. In earlier versions of the assembly, some of the chromosome 1 clones were used to construct the chromosome 16 locus. These clones were removed and the Hydin2 paralog was not in the assembly until GRCh37, but then only as an unlocalized scaffold. In GRCh38, Hydin 2 has been added to the chromosome 1 assembly. The figure to the top left shows the alignment of the Hydin locus in GRCh37 to the paralogous sequence on chr1 as well as to a chr 16. locus from another assembly. The figure to the right shows a zoomed in view. The green box highlights a paralogous sequence variant (PSV) that overlaps a submitted SNP. All variation in this region, including clinically associated variants, needs to be reviewed in light of this highly related paralog. Allelic variation in either region will be difficult to identify with current sequencing technologies.

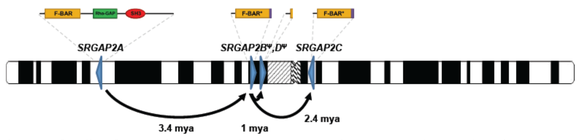


Figure 5. This figure shows a family of genes known as the SRGAPS. Only SRGAP2A is seen in other primates, the partial copies are human specific. No member of this gene family is well represented in GRCh37, but are all fully represented in GRCh38. This gene family is involved in neuronal outgrowth.

GRCh38 adds previously missing human sequences

Alternate Loci

References: Church et al, PLoS Biol, 2011
Kidd et al, PLoS Genetics 2007

Highly diverse regions of the human genome cannot be assembled into a single consensus. In these cases, the GRC constructs individual sequences and incorporates one into the chromosome while any other sequences are scaffolds aligned to the chromosome. In GRCh37, there were 3 genomic regions containing 9 alternate loci. In GRCh38, there are 178 regions of the genome containing 261 alternate loci. These sequences contain genes not represented on the chromosome assembly. Many of these regions, such as the MHC, LRC and the 17q21 region are biomedically important. Exclusion of these regions means omitting **197** genes from your analysis.

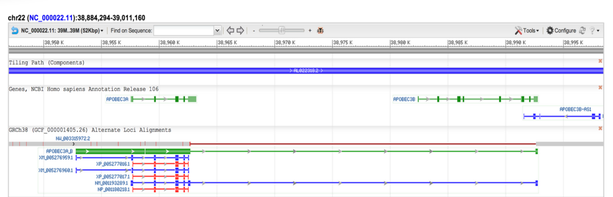


Figure 9. Deletion alternate loci are rare, but this example on chr22 shows an alternate locus in the APOBEC region. The deletion allele creates a new gene that is a combination of the APOBEC3A and APOBEC3B genes. The deletion allele shows population stratification as it is the minor allele (or never seen) in some populations but the major allele in others.

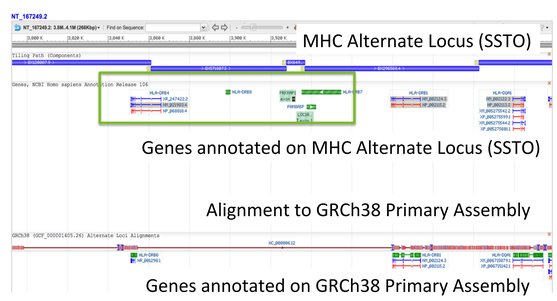


Figure 8. Alignment of the GRCh38 chromosome 6 sequence to one of the alternate loci at the MCH region. The top 2 tracks are the alternate sequence and genes annotated on this in NCBI *Homo sapiens* annotation 106. The bottom two tracks show the alignment to chromosome 6 and genes annotated there. The green box highlights gene models present on the alternate locus that are not on the chromosome assembly.

Alternate loci add gene models in biomedically important regions.

Further Information from Personalis:

- **Presentation: Friday, March 28, 2014, 12:30 pm – 1:00 pm, Theater 1**
- Garcia et al., **The Clinical Exome: Personalis' Experience Using an Enhanced Exomeand Genome-wide Structural Variant Detection for the Diagnosis of Diseases of Unknown Genetic Etiology**, poster #192
- Tirch et al., **User-Friendly Genomic Results: Leveraging a Novel Approach that has the Potential to Decrease Turn-Around Time and Preserve Opportunities for Novel Discoveries,** poster #561
- Chervitz et al., **Accurate Structural Variant Calling for Comprehensive Clinical Interpretation,** poster #291
- Chen et al., **Approaches to Increase Diagnostic Yield for Clinical Genomic Sequencing,** poster #286
- Chandratillake et al., **Exome Sequencing in 20 Probands with Developmental Eye Defects Identifies Causative Mutations in Five Cases and Demonstrates Genetic Heterogeneity,** poster #237
- Clark et al., **Successes Using ACE Exome Sequencing to Identify the Genetic Cause of Retinal Disorders in a Case Series,** poster #280
- **Personalis Booth #312**