Developments in human genome build GRCh38

Highlights

Alternative sequences

As the quantity of sequencing grows, the Genome Resource Consortium have concluded that several human chromosomal regions exhibit sufficient variability to prevent adequate representation by a single sequence. To address this, the GRCh38 assembly contains alternate sequences for selected variant which a presented as separate accessioned sequences. The GRCh38 build contains 261 alternative sequence loci, many of which are located within the LRC/KIR area on chr19 and the MHC region on chr6 (see Figure).

Sequence updates

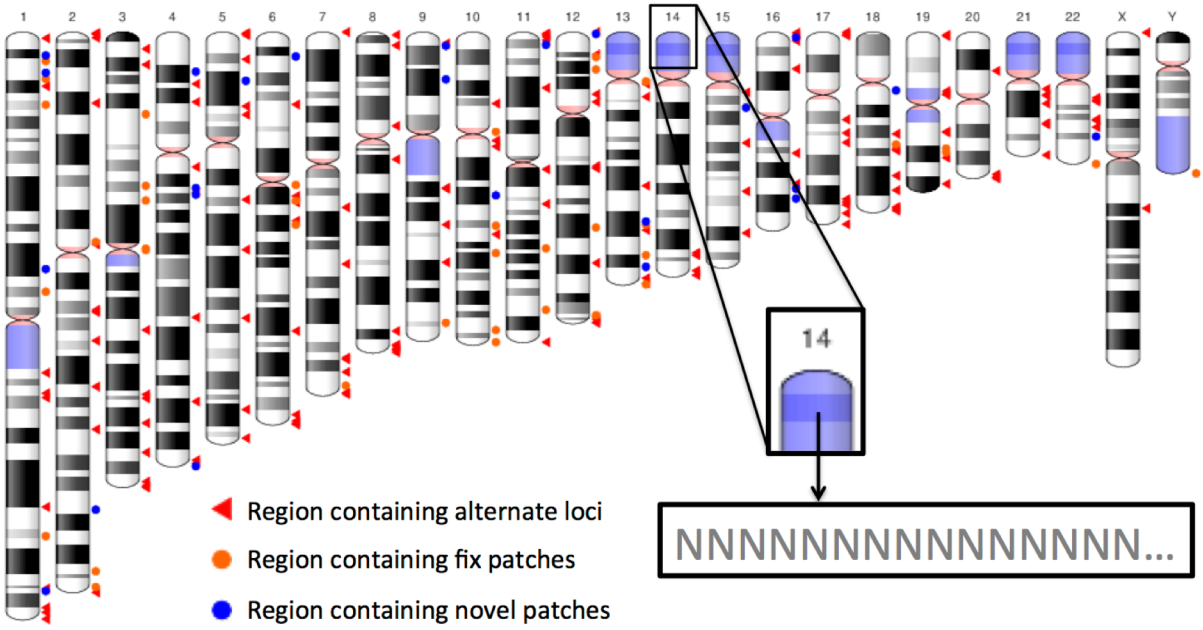
Using data from other genome sequencing projects and the 1000 genomes project, a number of erroneous bases and misassembled regions in GRCh37 have been corrected in the GRCh38 assembly and more than 100 gaps have been filled or reduced. The result of this polishing is an increase in sequenced bases (up from 2.99 to 3.05 Billion bases without N's).

Centromere representation

Coverage of the centromeric regions has been improved using the centromere models and analysis software developed by Karen Miga et al. The models, which provide the approximate repeat number and order for each centromere, will be useful for read mapping and variation studies as they replace the large megabase-sized gaps in previous assemblies.

Mitochondrial genome

The chrM sequence in hg19 (NC_001907) is replaced with the Revised Cambridge Reference Sequence (rCRS) from MITOMAP with GenBank accession number J01415.2.



Analysis Sets

In addition to the definitive genome builds, the GRCh38 assembly offers a number of analysis sets which have been created to accommodate the needs of next generation sequencing analysis.

At the current time, four analysis sets are available:

- no_alt_analysis_set
- full_analysis_set
- full_plus_hs38d1_analysis_set
- no_alt_plus_hs38d1_analysis_set

No-alt analysis sets

The no-alt analysis sets remove the alternate locus and patch scaffolds that cause complications for sequence read alignment programs that are not alt-aware.

Full analysis sets

The full analysis set versions contain hard masking (replacement with N's) of duplicate copies the pseudo-autosomal regions (PAR) and centromeric arrays

hs38d1 analysis sets

The hs38d1 analysis sets contain decoy and EBV sequences that are not part of the human genome assembly, but they are included in the analysis set to serve as read mapping "sinks" for highly repetitive sequences that are difficult to align and foreign reads that are often present in sequencing samples.

The EBV contig can help correct for artifacts stemming from immortalization of human blood lymphocytes with EBV transformation, as well as capture endogenous EBV sequence as EBV naturally infects B cells in ~90% of the world population.

Analysis implications

Tools are slowly evolving to make use of the information contained in the structural changes that GRCh38 introduces. We should expect to see further expansion of these tools as we move from the approach of using the genome without the alternative loci to one where all tools are alt-aware.

There is also the issue that current short-read technologies are not best suited for the sequencing of some of these alternative loci. For example, many of the HLA sequences are very similar, have a number (~10s) of common and many more possible (100s) haplotypes. All of which have extreme polymorphisms (1 SNP every 5bp) and also have pseudogenes which can be expressed. Decoding all this using short reads is a tough, possibly impossible mathematical problem.

At the time of writing BWA-MEM is the only "leading" mapper which is ALT-aware. It essentially computes mapping quality across the non-redundant content of the primary assembly plus the ALT contigs and is free of the problem above. BWA-MEM is carefully engineered such that ALT contigs do not interfere with the alignments to the primary assembly. Non-ALT alignments are only affected by ALT contigs if there are significantly better ALT alignments. For read counting in RNAseq analyses, hisat2 is ALT-aware.

	Number of contigs	Total length (bp)	Percent of primary assembly
Primary assembly	194	3,099,750,718	---
Alternative contigs	786	111,631,854	3.6%
- HLA alternative contigs	525	2,096,467	0.068%
- Other alternative contigs	261	109,535,387	3.5%
EBV and decoy contigs	2386	5,964,345	0.19%