GRCh38: a new version of the human reference genome sequence

James Torrance and Kerstin Howe on behalf of the Genome Reference Consortium
Wellcome Trust Sanger Institute, Cambridge, UK.

The Genome Reference Consortium

The Genome Reference Consortium (GRC) is the international collaboration responsible for maintaining the assembly of the human reference genome that is deposited with the INSDC. The GRC is also responsible for improving this genome assembly by closing remaining gaps and correcting sequencing errors.

Furthermore, the GRC is working to better represent complex variation in the human genome. A single clone tiling path is insufficient to represent certain variable regions, so the GRC has converted the human genome assembly to a modernised assembly model, which represents each variable region with one or more **alternate loci**: separate tiling paths which are anchored in the reference path by components that are the same in the reference and the variant.

The above improvements allow more complete and accurate read alignments [1] and provide a better basis for gene and feature annotation, which is especially valuable for the medical community.

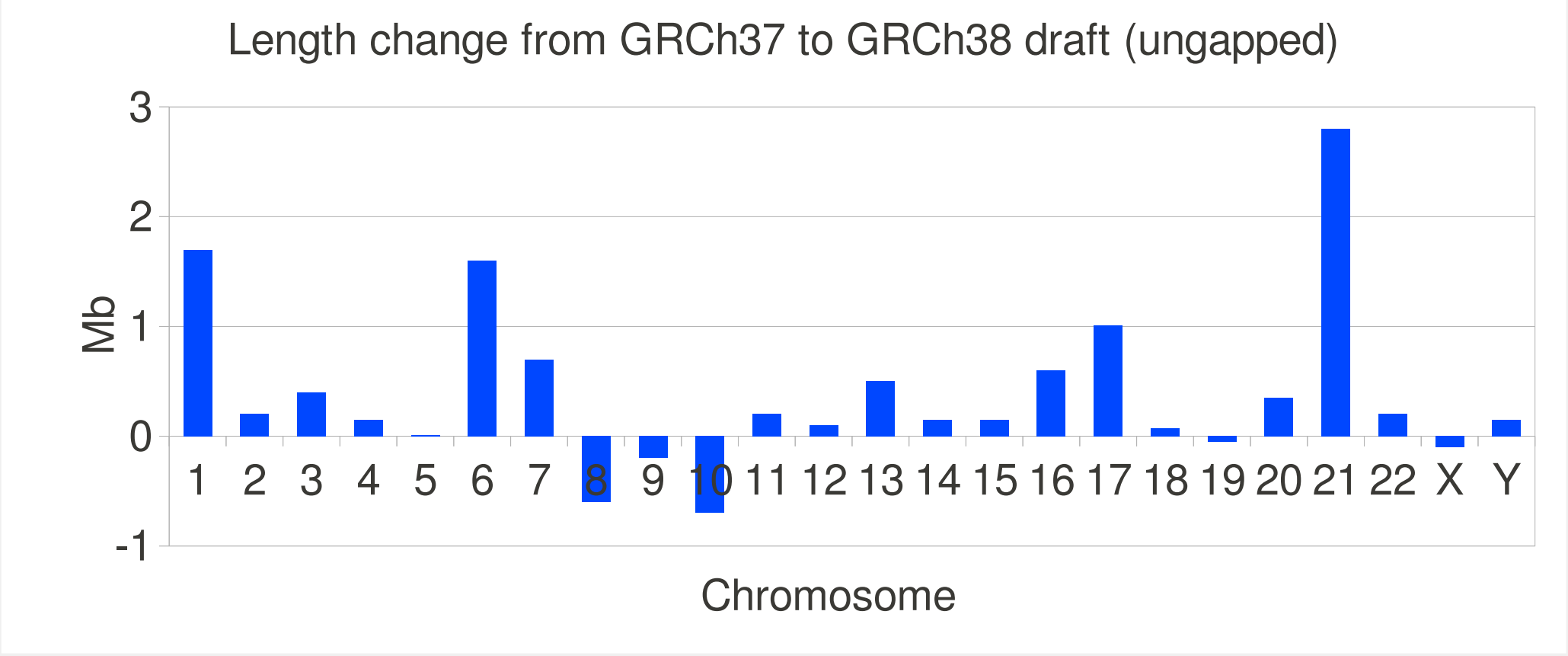
genomereference.org

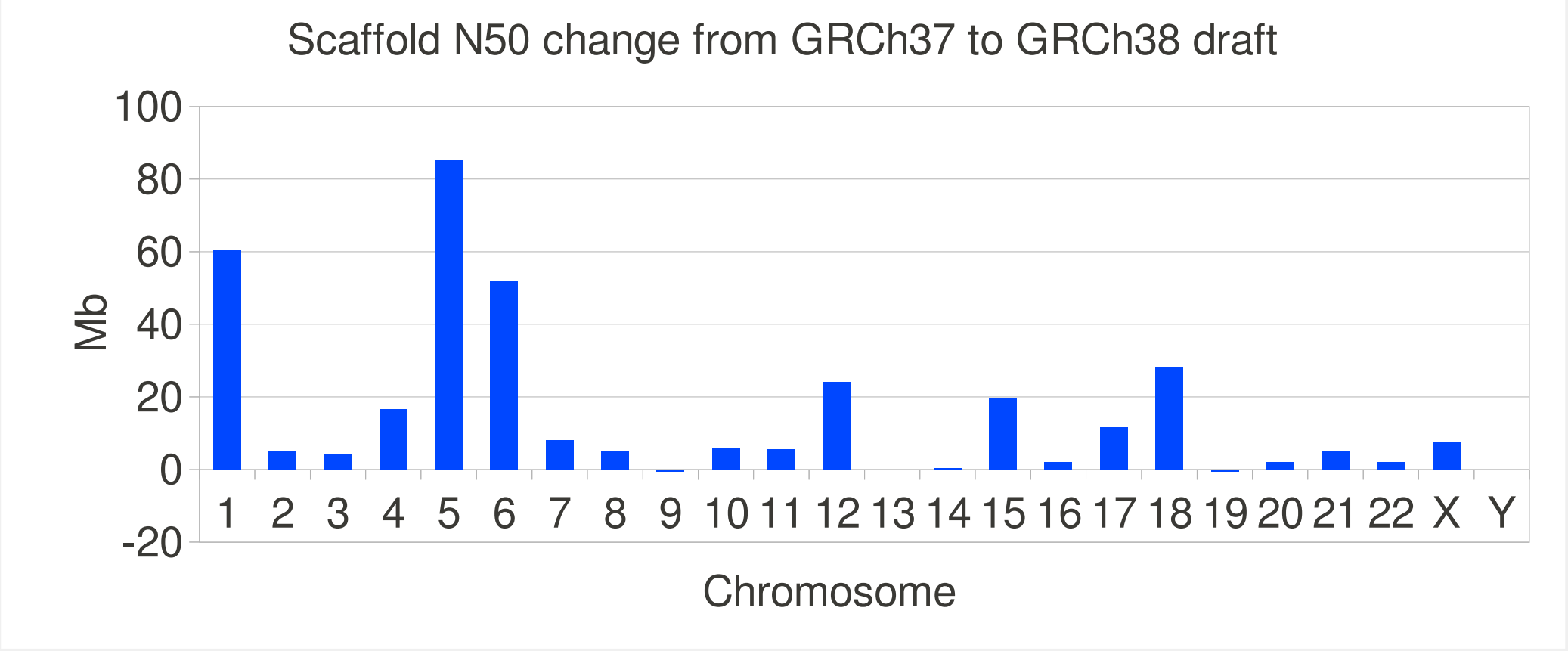
GRCh38

In 2009, we produced a **major release** for the the human genome (GRCh37) that converted it to the new assembly model. Since then, we have we added minor **patch releases** on a quarterly basis, which provided additional alternate loci or corrected errors. These patches have been adopted by the major genome browsers such as UCSC, Ensembl, and NCBI Mapviewer.

Later this year, the GRC will provide a new major release of the human reference genome: **GRCh38**, which will alter the coordinate system. This will incorporate all the improvements already released as patches, as well as further changes, some of which are described in the other sections of this poster. Gaps have been bridged, alternate representations have been added for variant regions, and issues have been fixed both within individual clones and at the level of the ordering of clones and contigs.

We have created a draft version of GRCh38 for internal testing purposes; this shows that GRCh38 adds more sequence to the reference assembly, and improves the N50 of its scaffolds due to gap closures.



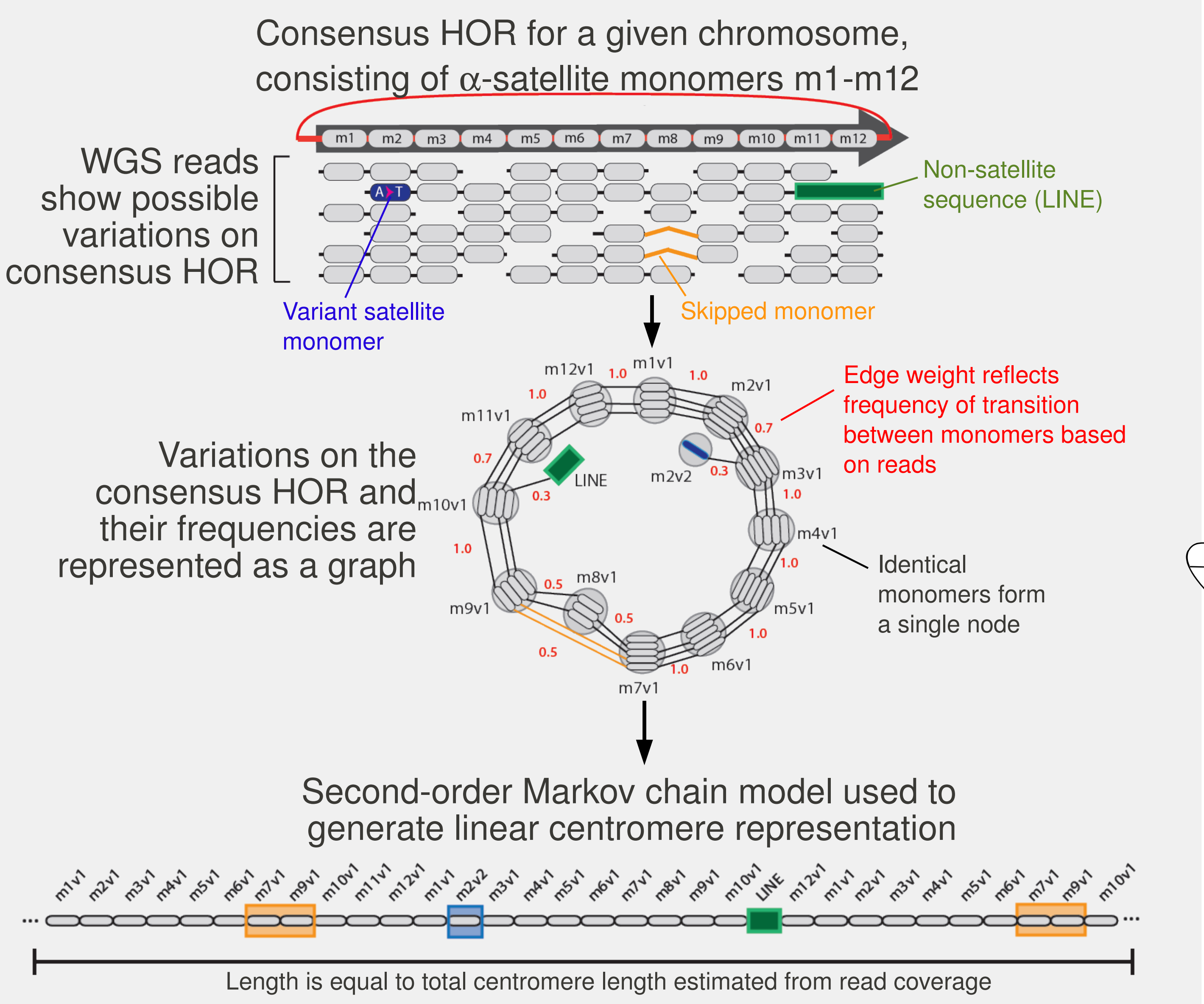


Megabase-scale improvements: Centromeres

Human centromere sequence is not represented in GRCh37 because these regions of highly repetitive α-satellite sequence stretching over several megabases are highly challenging to assemble. The α-satellite monomers may be organised into tandemly repeated multi-monomer arrays called **higher-order repeats (HOR)**.

The Kent group at UCSC has constructed **chromosome-specific representations** of centromere sequences, which will be added to GRCh38 [2]. These will allow reads of centromeric sequence to be mapped to the assembly.

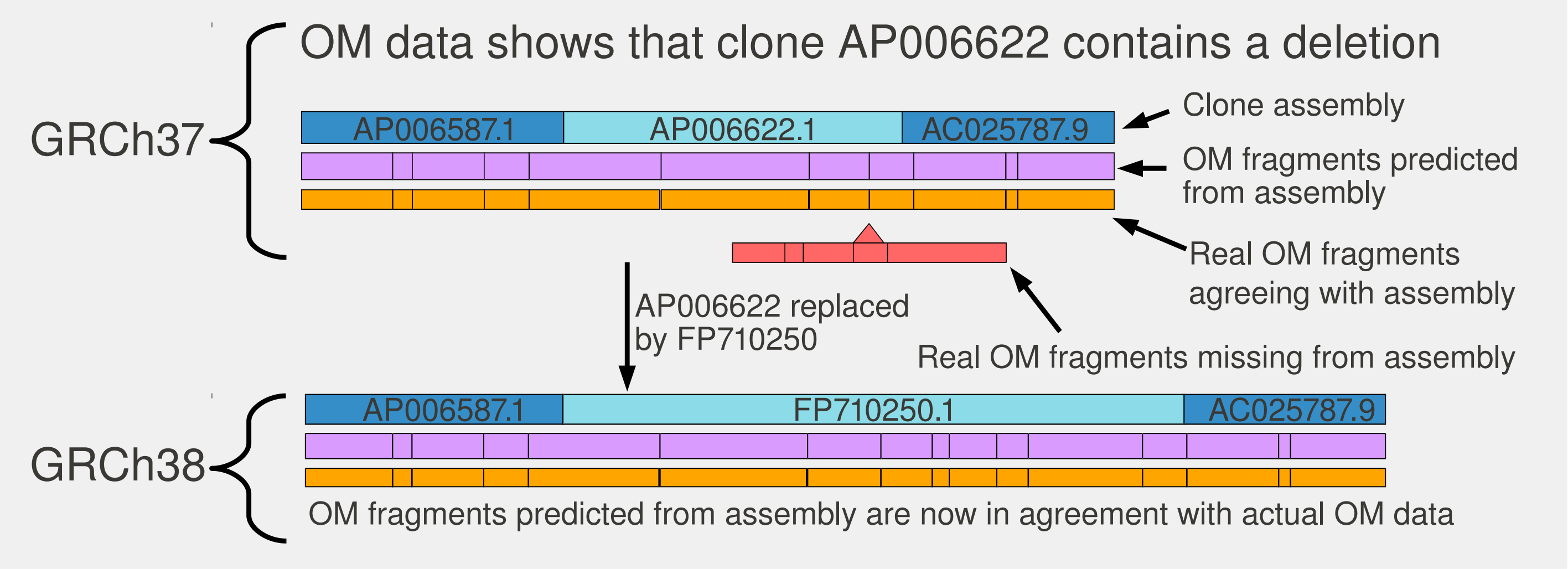
Figure adapted from "Complete sequence representation across human X and Y centromeric regions" by Hayden *et al.* (in press)

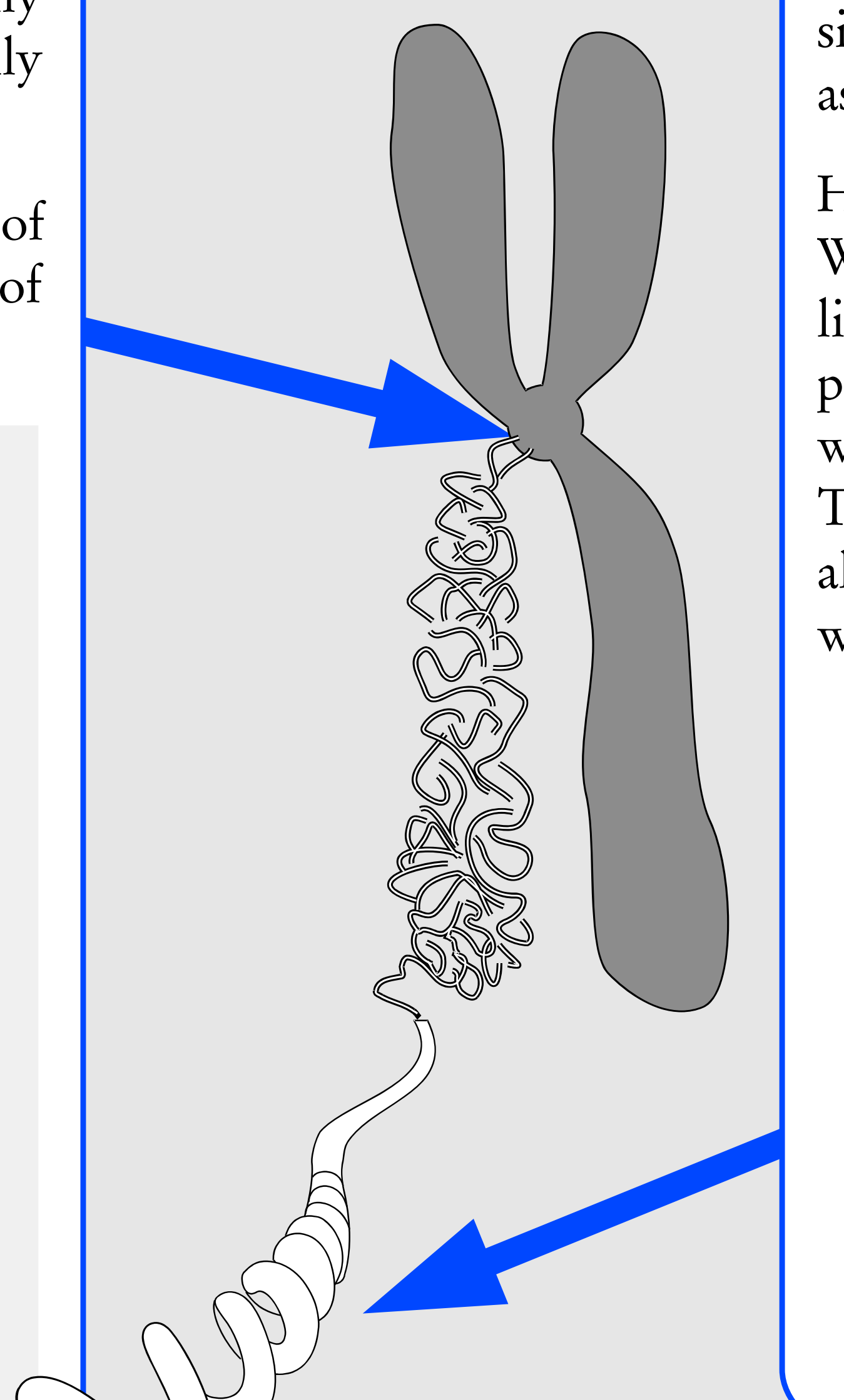

Consensus HOR for a given chromosome, consisting of α-satellite monomers m1-m12
m1 m2 m3 m4 m5 m6 m7 m8 m9 m10 m11 m12
WGS reads show possible variations on consensus HOR
A T
Non-satellite sequence (LINE)
Variant satellite monomer
Skipped monomer
Variations on the consensus HOR and their frequencies are represented as a graph
m1v1
m2v1
m3v1
m4v1
m5v1
m6v1
m7v1
m8v1
m9v1
m10v1
m11v1
m12v1
m2v2
LINE
1.0
0.7
0.3
0.5
Edge weight reflects frequency of transition between monomers based on reads
Identical monomers form a single node
Second-order Markov chain model used to generate linear centromere representation
m1v1 m2v1 m3v1 m4v1 m5v1 m6v1 m7v1 m9v1 m10v1 m11v1 m12v1 m1v1 m2v2 m3v1 m4v1 m5v1 m6v1 m7v1 m8v1 m9v1 m10v1 LINE m12v1 m1v1 m2v1 m3v1 m4v1 m5v1 m6v1 m7v1 m9v1 m10v1
Length is equal to total centromere length estimated from read coverage

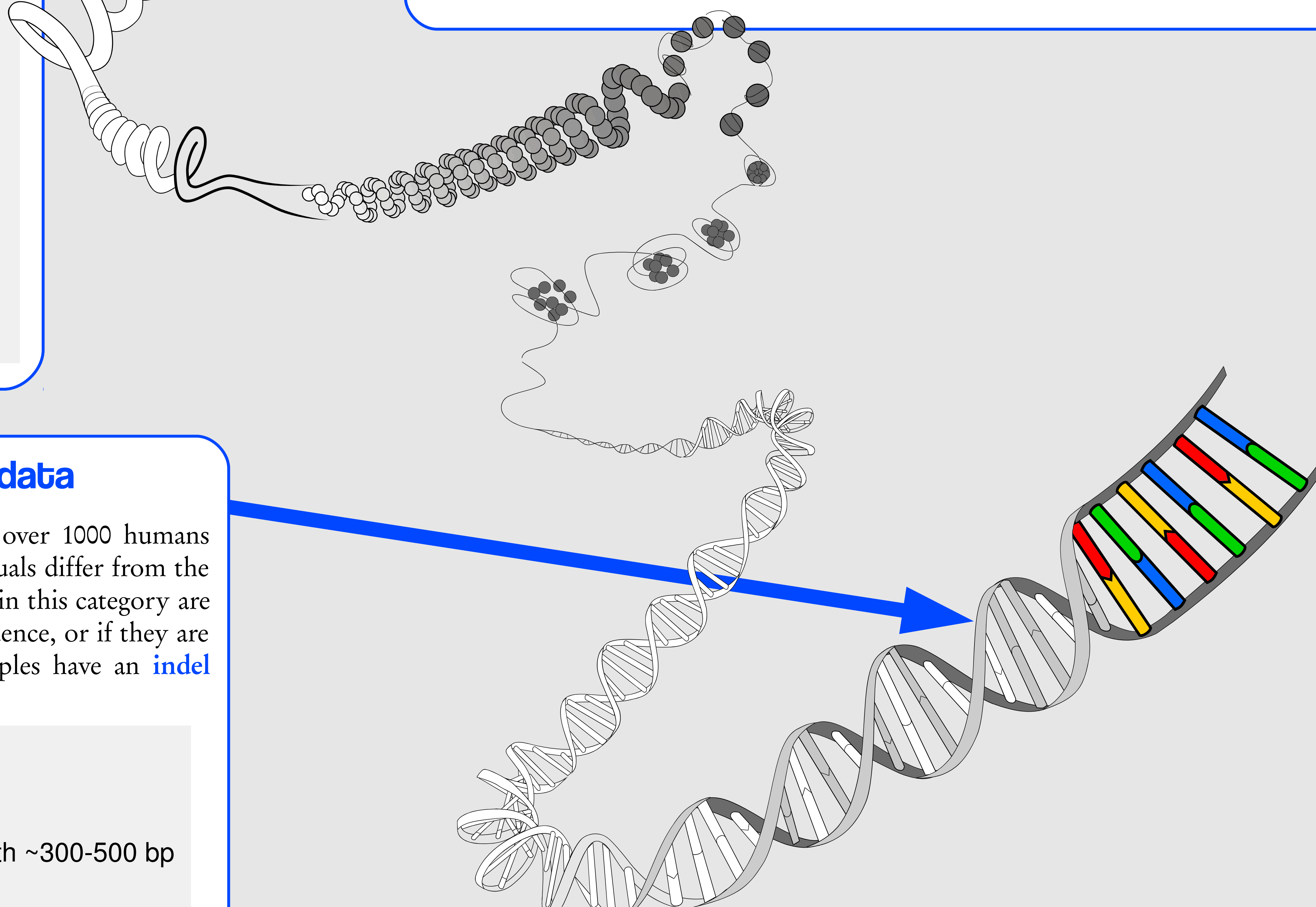
Kilobase-scale improvements: Optical mapping

Optical mapping (OM) is a method for constructing ordered restriction maps from single molecules of DNA. It allows us to identify or refute the presence of large-scale assembly issues such as genuine or artificial duplications, or deletions within clones.

Human optical maps have been created by David Schwartz' group at the University of Wisconsin for a complete hydatidiform mole and for three lymphoblast-derived cell lines [3]. A further optical map of the individual known as NA12878 has been produced by Opgen (www.opgen.com). There are 82 regions in the assembly where we have made changes that bring the assembly into better accord with this OM data. These include **64 gap closures** (in many cases, refuting the possibility that sequence alignments either side of a gap correspond to a duplication), and **12** problems which were resolved by **replacing a clone**.

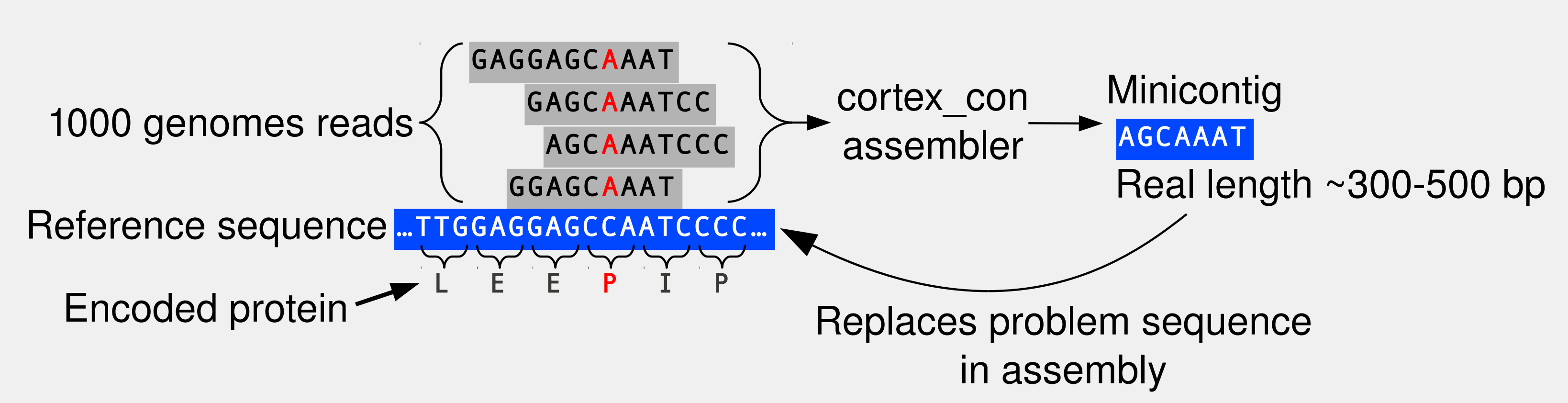

OM data shows that clone AP006622 contains a deletion
GRCh37
AP006587.1
AP006622.1
AC025787.9
Clone assembly
OM fragments predicted from assembly
Real OM fragments agreeing with assembly
Real OM fragments missing from assembly
AP006622 replaced by FP710250
GRCh38
AP006587.1
FP710250.1
AC025787.9
OM fragments predicted from assembly are now in agreement with actual OM data





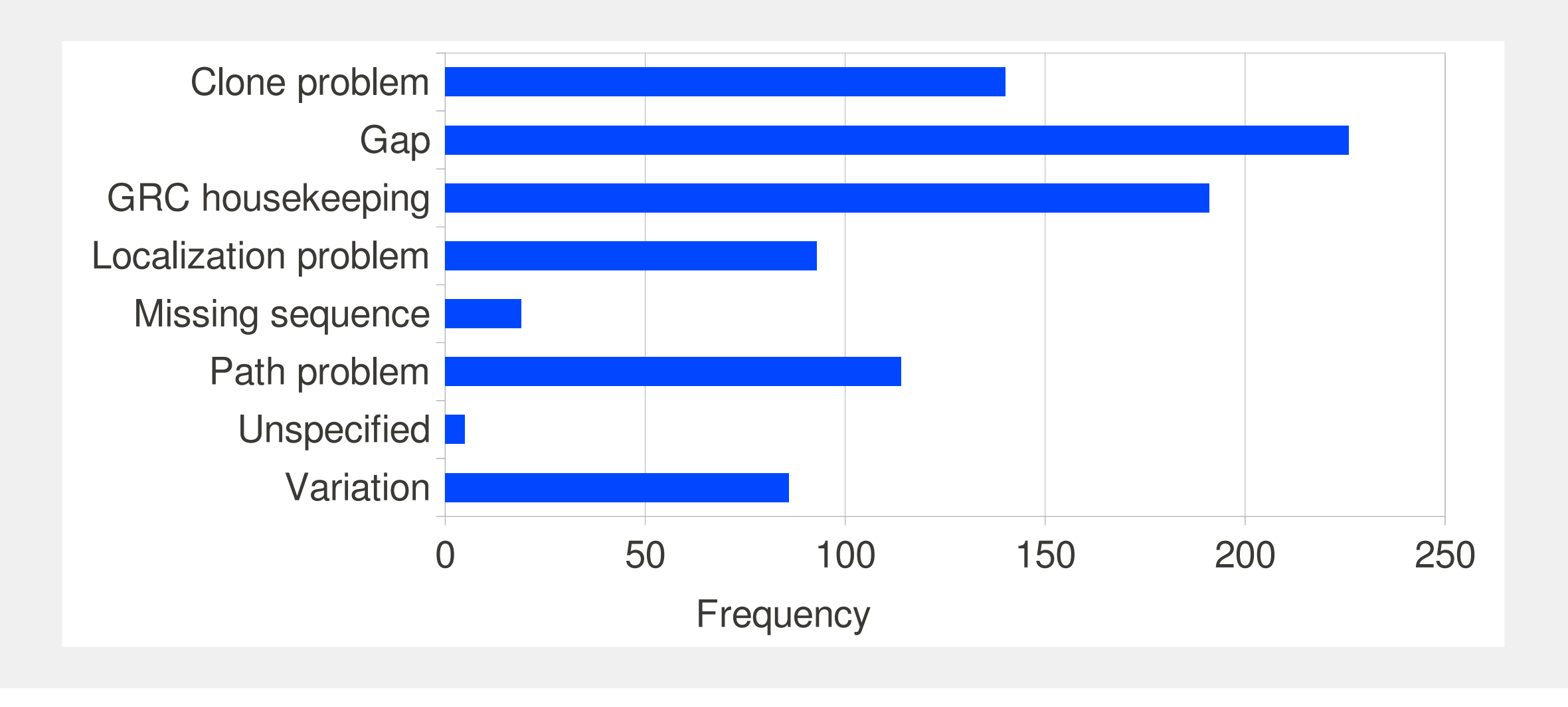
Single-base-scale improvements: Using 1000 genomes data

The 1000 genomes project has assessed human genetic variation by sequencing over 1000 humans from various populations [4]. There are genomic locations where all these individuals differ from the reference genome, which may represent an error in the reference assembly. **SNPs** in this category are being replaced for GRCh38 if they appear to represent errors in the reference sequence, or if they are rare and affect gene function (**over 6,000 cases**). Where all 1000 genomes samples have an **indel** relative to the reference sequence, this will also be replaced (**over 2,000 cases**).



Other types of human genome issue resolved by the GRC

The GRC tracks known issues with the assemblies it maintains, treating these as resolved once they result in a patch or are found not to require one. These are categorised into the types shown below. Progress on these issues can be tracked at genomereference.org.



References

1. Modernizing reference genome assemblies. Church *et al.* (2011) PLoS Biol 9(7): e1001091.
2. Complete sequence representation across human X and Y centromeric regions. Hayden *et al.* In press.
3. High-resolution human genome structure by single-molecule analysis. Teague *et al.* (2010) PNAS 107:10848-10853.
4. An integrated map of genetic variation from 1,092 human genomes. The 1000 Genomes Project Consortium (2012) Nature 491:56-65.

Chromosome image adapted from https://commons.wikimedia.org/wiki/File:Chromosome-es.svg