Advancements in the human genome reference assembly (GRCh38)

Tayebeh Rezaie, Ph.D.
NCBI
16 October 2019
Funding:

• This work was supported in part by the Intramural Research Program of the National Library of Medicine, National Institutes of Health.
• The European Molecular Biology Laboratory.
• The Wellcome Trust, UK.
• The MGI was supported by National Institutes of Health grants 5U54HG003079, 5U41HG007635 and 5U24HG009081.

Thanks to many GRC Collaborators
History of reference assembly

GRCh38/Reference genome:
- A critical resource to the basic & clinical research community, coordinate system, annotation source & discovery of disease-associated variants
- Sanger seq. clone-based from Human Genome Project; multiple individuals

Mosaic haploid: Individual 1, Individual 2, Individual 1

HGP → GRC: reference maintaining, improving and updates

HGP model (2003): each genomic region was represented with one sequence

Current model: ALT LOCI added to represent population genomic diversity

Alt loci: divergent large variation in genomic regions (not SNP/small indels)

Number of ALT LOCI:
- GRCh37: 9
- GRCh38: 261

Population distribution:
- RP11: 70% African-European
- 10 other libraries: 23%
- >50 other libraries: 7%
• Major/coordinate-changing: GRCh38 (Dec 2013)
• Patches/no coordinate-change: GRCh38.p13 (Mar 2019)

- 113 Fix patches: Add >3.88 Mb novel seq
- 72 Novel patches: Add >1.1 Mb novel seq
- 261 ALT Loci: Add 3.6 Mb novel seq

The new version of the reference should capture ALL the updates to GRCh38

The notion for variant representation has started long time ago.
Curation of reference assembly

- Issue sources: GRC assembly evaluation, reports from collaborators, community, literature
- Technology: sequencing, FISH, Optical Mapping
- Data resources: sequences generated by GRC or available in public database (clones, WGS, PCR products)

Evaluation of gaps in GRCh38

- Gap count = 196
- Excluded biological gaps & gaps within WGS scaffolds
- Reports of new assm that can close ref. gaps
- To identify gaps that can be spanned

Alignments of 8 diploid PacBio assemblies to the reference:

- Spanned with the same amount of seq: 26 (missing seq.)
- Spanned with varying amount of seq: 3 (variation)
- Spanned by some not all assemblies: 24 (complex, missing + variation)
- The remaining gaps are under review
Curation of reference assembly: Missing sequences

Evaluation to distinguish error vs. variation

- Find chr. context for missing seq.
- Add variants (>5 kb) as novel patches

Data sources:
- Eichler’s lab (Kidd et al. (2010) PMID: 20440878), structurally variant fosmid seq.
- Heng Li (GCA_000786075.2), a set of non-redundant seq. absent in GRCh38 and ALTs

Reported genome issues = 195
- Resolved no change: 94 (variation < 5 Kb)
- Patches (started adding from p1 in 2014)
  - FIX = 22
  - NOVEL = 43
- Pending action: 36 (Variation 8, sequence error 17, Unknown 11)
The most recent curation to GRCh38:
- FIX patches (43) + NOVEL (2)
- Added >0.5 Mb novel sequence
  - Gap closure: 28
  - Seq. error correction: 8
  - Path: 2
- For p-arm of acrocentric chrs: 5

Highlights of p13:
- Improved clinically important genomic regions
  - Prader-Willi (5.5 Mb, 1.63 Mb unique)
  - CT47A gene cluster
- Improved gene representations:
  SLC5A11, GCNT2, SAMD1, GRCK1, C1R, ECSCR, 5S rRNA

Sequence data sources for updates:
- CHM1 assm: 21
- CHM13 assm: 12
- Other WGS assm: 3
- Clones: 9
Correction of an assembly false gap caused by haplotype incompatibility

Mix haplotype representation of CT47A in GRCh38
Long haplotype: 12 copies

Single haplotype representation of CT47A in GRCh38.p13
Short haplotype: 7 copies

CHM1 Optical Map supporting the updated CT47A haplotype
CYP2D6 haplotypes: genomic diversity of a clinically important region
Involved in metabolizing many prescribed drugs

## Alignment of alt loci and patch scaffolds to the CYP2D6 region of chr. 22

![Alignment of alt loci and patch scaffolds to the CYP2D6 region of chr. 22](image)

## Scaffolds providing alternate sequence representations of CYP2D6 region

<table>
<thead>
<tr>
<th>GenBank Acc.</th>
<th>RefSeq Acc.</th>
<th>Population</th>
<th>CYP2D6</th>
<th>CYP2D7</th>
<th>CYP2D8</th>
</tr>
</thead>
<tbody>
<tr>
<td>KN196485.1</td>
<td>NW_003646207.1</td>
<td>African</td>
<td>Deletion</td>
<td>Single Copy</td>
<td>Single Copy</td>
</tr>
<tr>
<td>KB663609.1</td>
<td>NW_004504305.1</td>
<td>African</td>
<td>Duplication</td>
<td>Single Copy</td>
<td>Single Copy</td>
</tr>
<tr>
<td>KN196486.1</td>
<td>NW_009646208.1</td>
<td>East Asian</td>
<td>Duplication</td>
<td>Single Copy</td>
<td>Single Copy</td>
</tr>
<tr>
<td>KQ458387.1</td>
<td>NW_014040930.1</td>
<td>East Asian</td>
<td>Deletion</td>
<td>Single Copy</td>
<td>Duplication</td>
</tr>
<tr>
<td>KQ458388.1</td>
<td>NW_014040931.1</td>
<td>East Asian</td>
<td>3 Copies</td>
<td>Single Copy</td>
<td>Single Copy</td>
</tr>
<tr>
<td>KQ759761.1</td>
<td>NW_015148968.1</td>
<td>European</td>
<td>Single Copy</td>
<td>Duplication</td>
<td>Single Copy</td>
</tr>
<tr>
<td>GL383582.2</td>
<td>NW_003315971.2</td>
<td>Unknown</td>
<td>Deletion</td>
<td>Single Copy</td>
<td>Single Copy</td>
</tr>
</tbody>
</table>

![Scaffolds providing alternate sequence representations of CYP2D6 region](image)
Genome Reference Consortium

Unresolved genome issues

- Gap: 25%
- Variation: 20%
- Seq. error: 18%
- Localization: 16%
- Missing seq: 7%
- Path: 6%
- Unknown: 5%
- GRC: 3%

Unresolved Issues: n=339

Current curation status

- GRC
- Unknown
- Path
- Missing seq
- Seq. error
- Variation
- Localization
- Gap

Resolution likelihoods as determined by the GRC review

GRCh38.p14 is planned for release in 2020
Conclusion and Future

GRCh38.p14: coming in 2020

The reference has informed its own evolution.

MGI, a GRC member, has been awarded by NHGRI to:
• Produce 350 whole genome phased diploid assm.;
• Identify SVs between samples and current GRCh38;
• Incorporate those SVs into the Reference, likely as a graph representation.

GRCh39 is pending. The GRC is engaged in validation, providing curation tools and support to the pan-genome assemblies.