4DN Policy on Genome Assembly

Approved by the 4DN Omics Data Standards Committee: 7/10/2017
Approved by the 4DN Steering Committee: 7/18/2017

The 4D Nucleome Network will provide uniformly processed data for production grade experiment types. It is important that data from different experiments and experiment types are processed with a consistent set of reference annotations to ensure integrability. This document describes the Human Genome Reference assembly selected by the 4DN Omics Standards WG and the DCIC, and outlines the motivation for this selection.

We propose to use the GRCh38 Genome Reference Consortium Human Reference 38 Patch 15 (GCA_000001405.15) including the 25 assembled chromosomes (1-22, X, Y, M), the 127 unplaced contigs, and the 42 unlocalized contigs, but excluding the 261 alternative haplotypes. Since Epstein-Barr virus (EBV) very often turns up in human DNA sequencing and since this decoy assembly is included by the ENCODE consortium, we also propose to include the EBV assembly AC:AJ507799.2. This proposal will align our genome assembly exactly to the assembly used by ENCODE3 and ENCODE4.

Why GRCh38 rather than hg19?

- It is a more accurate assembly. In particular, ENCODE has found GRCh38 handles repeat sequences better, such that blacklists to filter repeats become less important.
- It is more future proof.

Why exclude alternative haplotypes?

- Standard aligners cannot utilize the extra information effectively. Reads mapping in these regions may be reported as non-unique alignment.
- We would like to use the same assembly as ENCODE.

Why not use hg19 also?

- Providing hg19 also will double our data processing costs (roughly $50k in fiscal year3 and going up in later years.)
- Providing hg19 also will require additional infrastructure development (UI tools to clearly denote which files are from which assembly) which will cost the DCIC software development team two months that can be used for other development.
- It will slow down the transition that the field is already going through.

How do we handle legacy data sets?

- Data from big consortia such as ENCODE are already available with respect to GRCh38.
- For datatypes for which we have implemented a uniform processing pipeline, the DCIC can process data from selected landmark publications to generate files in GRCh38. The OMICS WG has asked all members of the consortium to nominate such papers for consideration by OMICS WG. In addition, Over the next year, the DCIC will be able to

process legacy data from other production omics experiment types (Repli-seq, ChIP-seq, ATAC-seq, DNAse-seq, RNA-seq, ChIAPET/PLAC-seq) once standards have been approved.