

Core resources for genome analysis

Matthieu Muffato, Mark Blaxter

Tree of Life, Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, CB10 1SA, UK.



To facilitate genome analysis, we will compute a set of elementary sequence analyses on every genome that comes out of the institute and provide the tracks on a publicly available server through Track Hubs. The goals is to save the community efforts and resources by providing a uniform dataset, complement submissions to the public archives (ENA, EVA), and supply additional process data to further improve genomes. Additional submissions will include aligned reads (as CRAM files) and variant calls (as VCF files).

We will use standard file formats:

- Tab-separated
 - BED
 - bedGraph
 - VCF
 - GFF
 - TSV
- Binary indexed
 - bigBed
 - bigWig
 - BCF
 - CRAM
- JSON

Assembly metrics
Count, N50, sizes
Chromosomes, scaffolds and contigs
BUSCO scores (for all matching datasets)
Base content
N%, GC%, GC skew
k-mer analyses
k-mer spectra
Assembly QV stats
di/tri/tetra-nucleotides, hexamers
Read alignment (all read sets)
Read coverage
Variant calls (germline mutations)
Joint-calling if multiple samples
Density
Variants
Repeats (for each repeat type)
Genes (for each biotype)
Hi-C plot
Gene sets and homologies
Ensembl
BUSCO
Genome structure
Telomeric repeats
Synteny
GDA (Genome Decomposition Analysis)



```
hub.txt
stats/
read_coverage/
  hifi
  hic
  10x
  rnaseq
fastk/
  tab/
  bed/
base_content/
  k1/
  k2/
  k3/
  k4/
  hexamer/
variant/
  hifi/
  illumina/
busco/
  [buscoset1]/
  [buscoset2]/
repeat/
  ensembl/
  gda /
gene/
  busco/
  ensembl/
```

Standard, documented, paths, e.g.

```
/insects/Celastrina_argiolus/analysis/ilCelArgi3.1/...
.../hub.txt
.../base_content/k1/assembly.GC.1k.bw
.../busco/insecta_odb10_mtaeuk.scores.json
```

Data and track-hubs hosted at
<https://darwin.cog.sanger.ac.uk>
(S3-compatible)



This work was supported by Wellcome through core funding to the Wellcome Sanger Institute (206194) and the Darwin Tree of Life Discretionary Award (218328)

