BlobToolKit – Interactive quality assessment of genome assemblies

Richard Challis [1,2*], Edward Richards [3], Jeena Rajan[3], Guy Cochrane [3], Mark Blaxter [1,2]

1 Institute of Evolutionary Biology, University of Edinburgh, Edinburgh EH9 3JT, UK

2 Wellcome Sanger Institute, Cambridge CB10 1SA, UK

3 European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Cambridge CB10 1SD, UK

Supplementary File S1

Figure captions and URLs

Most of the figures presented in this paper were generated using the BlobToolKit Viewer and can be reproduced using the following URLs. Where applicable, these are presented alongside the automatically generated captions.

Figure 1A

Caption: Square-binned blob plot of base coverage in SRR026696 against GC proportion for scaffolds in assembly ACVV01. Scaffolds are coloured by phylum and binned at a resolution of 30 divisions on each axis. Coloured squares within each bin are sized in proportion to the sum of individual scaffold lengths on a square-root scale, ranging from 452 to 36,531,624. Histograms show the distribution of scaffold length sum along each axis.

URL: https://blobtoolkit.genomehubs.org/view/ACVV01/dataset/ACVV01/blob

Figure 1B

Caption: Cumulative scaffold length for assembly ACVV01. The grey line shows cumulative length for all scaffolds. Coloured lines show cumulative lengths of scaffolds assigned to each phylum using the bestsumorder taxrule.

URL: https://blobtoolkit.genomehubs.org/view/ACVV01/dataset/ACVV01/cumulative

Figure 1C

Caption: Snail plot summary of assembly statistics for assembly ACVV01. The main plot is divided into 1,000 size-ordered bins around the circumference with each bin representing 0.1% of the 253,560,284 bp assembly. The distribution of scaffold lengths is shown in dark grey with the plot radius scaled to the longest scaffold present in the assembly (7,262,926 bp, shown in red). Orange and pale-orange arcs show the N50 and N90 scaffold lengths (23,589 and 3,544 bp), respectively. The pale grey spiral shows the cumulative scaffold count on a log scale with white scale lines showing successive orders of magnitude. The blue, pale-blue and white area around the outside of the plot shows the distribution of GC, AT and N percentages in the same bins as the inner plot. A summary of complete, fragmented, duplicated and missing BUSCO genes in the diptera_odb9 set is shown in the top right.

URL: https://blobtoolkit.genomehubs.org/view/ACVV01/dataset/ACVV01/snail

Figure 1D

URL:

https://blobtoolkit.genomehubs.org/view/ACVV01/dataset/ACVV01/busco

Figure 2:

A representation of the INSDC-pipeline directed acyclic graph (DAG) can be generated using Snakemake. From a local copy of the pipeline directory, a PDF format image of the DAG can be created using the command:

```
snakemake -p -n --configfile example.yaml --rulegraph
--forceall | dot -Tpdf > DAG.pdf
```

Figure 3A

Caption: Square-binned blob plot of base coverage in SRR026696 against GC proportion for scaffolds in assembly ACVV01. Scaffolds are coloured by phylum and binned at a resolution of 30 divisions on each axis. Coloured squares within each bin are sized in proportion to the sum of individual scaffold lengths on a logarithmic scale, ranging from 867 to 40,536,114. Histograms show the distribution of scaffold length sum along each axis.

URL: https://blobtoolkit.genomehubs.org/view/ACVV01/dataset/ACVV01/blob?zScale=scaleLog&catField=bestsumorder_phylum&otherLimit=2&palette=user&color1=rgba%28191%2C191%2C191%2C1%29&color0=rgba%28255%2C127%2C0%2C1%29&showTotal=false&bestsumorder_phylum--Active=true&bestsumorder_phylum--Order=%2CProteobacteria#Lists

Selection: To highlight the set of Proteobacterial scaffolds with BUSCOs, upload File S2 to the above URL then deactivate the selection using the Filters menu.

Figure 3B

Caption: Kite-shaped blob plot of base coverage in SRR026696 against GC proportion for scaffolds in assembly ACVV01. Scaffolds are coloured by phylum. Kite shapes summarise the core distribution of scaffolds. Horizontal and vertical lines represent a range spanning 2 standard deviations about the weighted mean value for each axis. The lines intersect at a point representing the weighted median value. Histograms show the distribution of scaffold length sum along each axis.

URL: https://blobtoolkit.genomehubs.org/view/ACVV01/dataset/ACVV01/blob?zScale=scaleLog&catField=bestsumorder_phylum&otherLimit=2&palette=user&color1=rgba%28191%2C191%2C191%2C1%29&color0=rgba%28255%2C127%2C0%2C1%29&showTotal=false&bestsumorder_phylum--Active=true&bestsumorder_phylum--Order=%2CProteobacteria&plotShape=kite

Figure 3C

Caption: Square-binned blob plot of base coverage in SRR026696 against GC proportion for scaffolds in assembly ACVV01. Scaffolds are coloured by genus and binned at a resolution of 30 divisions on each axis. Coloured squares within each bin are sized in proportion to the sum of individual scaffold lengths on a square-root scale, ranging from 1,005 to 771,195. The assembly has been filtered to exclude scaffolds with phylum in no-hit, Arthropoda, Ascomycota, Mollusca, undef, Streptophyta, Nematoda, Chordata, Platyhelminthes, Cnidaria, Echinodermata, Firmicutes, Annelida, Cyanobacteria, Viruses-undef, Eukaryota-undef, Actinobacteria, Tardigrada, Hemichordata, Porifera or Basidiomycota. Histograms show the distribution of scaffold length sum along each axis.

URL: https://blobtoolkit.genomehubs.org/view/ACVV01/dataset/ACVV01/blob?zScale=scaleSqrt&catField=bestsumorder_genus&otherLimit=4&palette=user&color1=rgba%2831%2C120%2C180%2C1%29&color0=rgba%28255%2C127%2C0%2C1%29&showTotal=false&bestsumorder_phylum--Active=true&bestsumorder_phylum--Order=%2CProteobacteria&bestsumorder_phylum--Keys=0%2C2%2C4%2C6%2C3%2C7%2C13%2C5%2C8%2C9%2C11%2C10%2C12%2C16%2C17%2C14%2C15%2C18%2C19%2C20%2C21&bestsumorder_genus--Active=true&bestsumorder_genus--Order=Acetobacter%2CGluconobacter%2CWolbachia&color2=rgba%2851%2C160%2C44%2C1%29&plotShape=square&color3=rgba%28191%2C191%2C191%2C1%29

Figure 3D

Caption: Kite-shaped blob plot of base coverage in SRR026696 against GC proportion for scaffolds in assembly ACVV01. Scaffolds are coloured by genus. Kite shapes summarise the core distribution of scaffolds. Horizontal and vertical lines represent a range spanning 2 standard deviations about the weighted mean value for each axis. The lines intersect at a point representing the weighted median value. Histograms show the distribution of scaffold length sum along each axis.

URL: https://blobtoolkit.genomehubs.org/view/ACVV01/dataset/ACVV01/blob?zScale=scaleSqrt&catField=bestsumorder_genus&otherLimit=4&palette=u

ser&color1=rgba%2831%2C120%2C180%2C1%29&color0=rgba%28255%2C127%2C0%2C1%29&showTotal=false&bestsumorder_phylum--Active=true&bestsumorder_phylum--Order=%2CProteobacteria&bestsumorder_genus--Active=true&bestsumorder_genus--Order=Acetobacter%2CGluconobacter%2CWolbachia&color2=rgba%2851%2C160%2C44%2C1%29&plotShape=kite&color3=rgba%28191%2C191%2C191%2C1%29

Figure 4A

Caption: Square-binned blob plot of base coverage in SRR1714990 against GC proportion for scaffolds in assembly SDAX01. Scaffolds are coloured by phylum and binned at a resolution of 30 divisions on each axis. Coloured squares within each bin are sized in proportion to the sum of individual scaffold lengths on a square-root scale, ranging from 201 to 183,673,465. Histograms show the distribution of scaffold length sum along each axis.

URL:
https://blobtoolkit.genomehubs.org/view/Conus%20consors/dataset/SDAX01/blob?staticThreshold=2800000000&nohitThreshold=2800000000

Figure 4B

Caption: Square-binned blob plot of base coverage in SRR1719763 against base coverage in SRR1712902 for scaffolds in assembly SDAX01. Scaffolds are coloured by phylum and binned at a resolution of 30 divisions on each axis. Coloured squares within each bin are sized in proportion to the sum of individual scaffold lengths on a square-root scale, ranging from 201 to 33,053,995. The assembly has been filtered to exclude scaffolds with base coverage in SRR1714990 > 0.01. Histograms show the distribution of scaffold length sum along each axis.

URL:
https://blobtoolkit.genomehubs.org/view/Conus/dataset/SDAX01/blob?SRR1714990_cov--Max=0.01&SRR1719763_cov--Active=true&yField=SRR1719763_cov&xField=SRR1712902_cov&SRR1712902_cov--Active=true&staticThreshold=2800000000&nohitThreshold=2800000000

Figure 4C

Caption: Square-binned blob plot of base coverage in SRR1714990 against GC proportion for scaffolds in assembly SDAX01. Scaffolds are coloured by phylum and binned at a resolution of 30 divisions on each axis. Coloured squares within each bin are sized in proportion to the sum of individual scaffold lengths on a square-root scale, ranging from 201 to 21,767,682. The assembly has been filtered to exclude scaffolds with phylum matches no-hit. Histograms show the distribution of scaffold length sum along each axis.

URL: https://blobtoolkit.genomehubs.org/view/Conus%20consors/dataset/SDAX01/blob?bestsumorder_phylum--Keys=6&staticThreshold=2800000000

Figure 4D

Caption: Blob plot of base coverage in SRR1714990 against GC proportion for scaffolds in assembly SDAX01. Scaffolds are coloured by phylum. Circles are sized in proportion to scaffold length on a square-root scale, ranging from 201 to 148,256. The assembly has been filtered to exclude scaffolds with phylum matches no-hit. Histograms show the distribution of scaffold length sum along each axis.

URL: https://blobtoolkit.genomehubs.org/view/Conus%20consors/dataset/SDAX01/blob?bestsumorder_phylum--Keys=6&plotShape=circle&plotGraphics=svg&staticThreshold=2800000000

Figure 4E

Caption: Kite-shaped blob plot of base coverage in SRR1714990 against GC proportion for scaffolds in assembly SDAX01. Scaffolds are coloured by phylum. Kite shapes summarise the core distribution of scaffolds. Horizontal and vertical lines represent a range spanning 2 standard deviations about the weighted mean value for each axis. The lines intersect at a point representing the weighted median value. Histograms show the distribution of scaffold length sum along each axis. The assembly has been filtered to exclude scaffolds with phylum matches no-hit.

URL: https://blobtoolkit.genomehubs.org/view/Conus%20consors/dataset/SDAX01/blob?bestsumorder_phylum--Keys=6&plotShape=kite&staticThreshold=2800000000

Figure 5A

Caption: Blob plot of base coverage in SRR6918124 against GC proportion for scaffolds in assembly PTEZ01. Scaffolds are coloured by phylum. Circles are sized in proportion to scaffold length on a square-root scale, ranging from 1,000 to 14,341,824. Histograms show the distribution of scaffold length sum along each axis.

URL: https://blobtoolkit.genomehubs.org/view/Aves/dataset/PTEZ01/blob?plotShape=circle&plotResolution=42&yField=SRR6918124_cov&SRR6918124_co

v--Active=true&palette=user&color1=rgba%28191%2C191%2C191%2C1%29&color2=rgba%28255%2C127%2C0%2C1%29&color7=rgba%28178%2C223%2C138%2C1%29&color0=rgba%2831%2C120%2C180%2C1%29#Settings

Figure 5B

Caption: Blob plot of base coverage in SRR6918124 against GC proportion for scaffolds in assembly PTEZ01. Scaffolds are coloured by family. Circles are sized in proportion to scaffold length on a square-root scale, ranging from 1,000 to 613,323. The assembly has been filtered to exclude: scaffolds with base coverage in SRR6918124 > 2; or family not in no-hit, Physeteridae, Tinamidae, Odontophoridae or Sarcocystidae. Histograms show the distribution of scaffold length sum along each axis.

URL: https://blobtoolkit.genomehubs.org/view/Aves/dataset/PTEZ01/blob?plotShape=circle&plotResolution=42&yField=SRR6918124_cov&SRR6918124_cov--Active=true&palette=user&color1=rgba%28191%2C191%2C191%2C1%29&color2=rgba%28255%2C127%2C0%2C1%29&color7=rgba%28178%2C223%2C138%2C1%29&bestsumorder_family--Active=true&catField=bestsumorder_family&color4=rgba%2851%2C160%2C44%2C1%29&color3=rgba%2851%2C160%2C44%2C1%29&color6=rgba%2831%2C120%2C180%2C1%29&color0=rgba%28166%2C206%2C227%2C1%29&bestsumorder_family--Keys=13%2C5%2C6%2C9%2C10&bestsumorder_family--Inv=true&SRR6918124_cov--Max=2

Figure 5C

Caption: Square-binned blob plot of base coverage in SRR6918124 against GC proportion for scaffolds in assembly PTEZ01. Scaffolds are coloured by phylum and binned at a resolution of 42 divisions on each axis. Coloured squares within each bin are sized in proportion to the sum of individual scaffold lengths on a square-root scale, ranging from 1,000 to 172,797,723. Histograms show the distribution of scaffold length sum along each axis.

URL: https://blobtoolkit.genomehubs.org/view/Aves/dataset/PTEZ01/blob?plotResolution=42&yField=SRR6918124_cov&SRR6918124_cov--Active=true&palette=user&color1=rgba%28191%2C191%2C191%2C1%29&color2=rgba%28255%2C127%2C0%2C1%29&color7=rgba%28178%2C223%2C138%2C1%29&color0=rgba%2831%2C120%2C180%2C1%29#Lists

Selection: To highlight the set of scaffolds with BUSCO annotations in any of the reference gene sets, upload File S3 to the above URL then deactivate the selection using the Filters menu.

Table reproduction

Tables were produced using code available in the **INSDC-pipeline** repository or obtained directly from the Viewer, using the commands and URLs presented below.

Table 1

Current values for the numbers of analysed and available assemblies presented in Table 1 can be obtained using the **INSDC-pipeline** script `available_assemblies.py` and a local copy of the NCBI taxonomy `new_taxdump` directory:

```
scripts/available_assemblies.py --taxdump
/path/to/new_taxdump
```

Table 2

Values in Table 2 were obtained from the BUSCO view of the unfiltered PTEZ01 dataset (https://blobtoolkit.genomehubs.org/view/Aves/dataset/PTEZ01/busco). Numbers in parentheses were obtained by subtracting the BUSCO scores with a minimum coverage filter of 2 applied to the base coverage in read set SRR6918124 (https://blobtoolkit.genomehubs.org/view/Aves/dataset/PTEZ01/busco?SRR6918124_cov--Active=true&SRR6918124_cov--Min=2) from the unfiltered values.

Table 3

Values in Table 3 were obtained from https://blobtoolkit.genomehubs.org/view/Chordata by using the *Customise table* option and adding *Apicomplexa span* and *Chordata span* to the displayed assembly statistics.