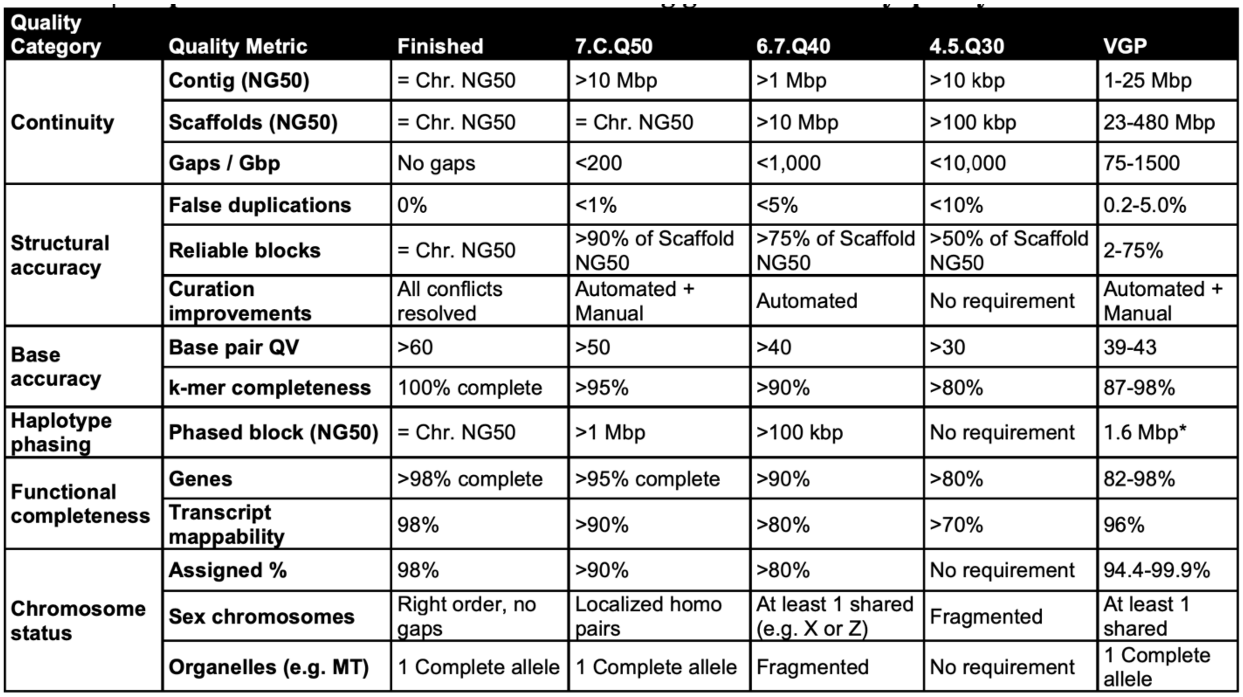
# Standards Recommendations for the Earth BioGenome Project

## Supplementary Tables and Legends

**Table S1 EBP assembly quality standards.** We recommend the “6.7.Q40” standard as a minimum for EBP to ensure a solid basis for future biological research. Table from (10) with permission.



| Quality Category | Quality Metric | Finished | 7.C.Q50 | 6.7.Q40 | 4.5.Q30 | VGP |
|---|---|---|---|---|---|---|
| **Continuity** | **Contig (NG50)** | = Chr. NG50 | >10 Mbp | >1 Mbp | >10 kbp | 1-25 Mbp |
| | **Scaffolds (NG50)** | = Chr. NG50 | = Chr. NG50 | >10 Mbp | >100 kbp | 23-480 Mbp |
| | **Gaps / Gbp** | No gaps | <200 | <1,000 | <10,000 | 75-1500 |
| **Structural accuracy** | **False duplications** | 0% | <1% | <5% | <10% | 0.2-5.0% |
| | **Reliable blocks** | = Chr. NG50 | >90% of Scaffold NG50 | >75% of Scaffold NG50 | >50% of Scaffold NG50 | 2-75% |
| | **Curation improvements** | All conflicts resolved | Automated + Manual | Automated | No requirement | Automated + Manual |
| **Base accuracy** | **Base pair QV** | >60 | >50 | >40 | >30 | 39-43 |
| | **k-mer completeness** | 100% complete | >95% | >90% | >80% | 87-98% |
| **Haplotype phasing** | **Phased block (NG50)** | = Chr. NG50 | >1 Mbp | >100 kbp | No requirement | 1.6 Mbp* |
| **Functional completeness** | **Genes** | >98% complete | >95% complete | >90% | >80% | 82-98% |
| | **Transcript mappability** | 98% | >90% | >80% | >70% | 96% |
| **Chromosome status** | **Assigned %** | 98% | >90% | >80% | No requirement | 94.4-99.9% |
| | **Sex chromosomes** | Right order, no gaps | Localized homo pairs | At least 1 shared (e.g. X or Z) | Fragmented | At least 1 shared |
| | **Organelles (e.g. MT)** | 1 Complete allele | 1 Complete allele | Fragmented | No requirement | 1 Complete allele |

**Table S2. Annotation features.**

| Features to be annotated in all genomes | Features that may be annotated in some genomes |
|---|---|
| 1. Simple repeats and transposable elements<br>2. Functional sequence features such as CpG islands<br>3. Protein-coding genes<br>4. Non-protein coding genes including small RNA (sRNA) and lncRNA genes | 1. Pseudogenes<br>2. Regulatory regions including promoters, enhancers, regions of open chromatin and locations of DNA binding proteins.<br>3. Chromosomal features such as banding patterns<br>4. Homology relationships between genes or other features<br>5. Horizontal and lateral gene transfer |

**Table S3: Gene annotation evidence categories.**

| | |
|---|---|
| *Ab initio* | Predictions based on the sequence features of a single genome (only). |
| *De novo* | Predictions based on sequence models within one of multiple species or comparative simultaneous annotation of multiple genomes. No expressed sequence is used. |
| Projection | Transfer of annotation from one species to another via a scaffold/chromosome level assembly to assembly alignment. |
| Protein sequence alignment | Identification of coding regions by alignment of observed or inferred protein sequence from another species. |
| cDNA sequence alignment | Identification of coding regions and untranslated regions by alignment of predicted or confirmed transcript sequences from another species. |
| Short read transcriptomics | Gene structure annotation based on alignment of short read RNA-seq or assembled transcripts from the same or closely related species. |
| Long read transcriptomics | Extended (possibly full length) transcript annotation from alignment of long read transcriptomic data from the same or closely related species. |
| Other molecular data | Other data providing insight into genome annotation including, but not limited to, EST sequences, proteomics data, RiboSeq and other expressed or functional data. |
| Expert manual curation | Annotation evaluated via a systematic expert process by human annotators. |

**Table S4 Genome reference analyses: Approaches methods and resources.**

| Analysis category | Areas addressed | Example methods |
|---|---|---|
| Assembly and scaffolding | Long read assembly, haplotypic duplicate removal, scaffolding, polishing, quality assessment | HiFiAsm, HiCanu (PacBio HiFi)<br>Flye, Shasta (ONT)<br>Purge_dups<br>SALSA2, JuiceBox, PreText (HiC), Solve (BioNano)<br>Winnowmap/FreeBayes/Merfin<br>Merqury, yak |
| Alignments of genomes and synteny analysis | Alignments form the basis for comparative genomics. Alignments can be generated using TBLASTN or blast reciprocal best matches at both the nucleotide level for evolutionarily close species, and the protein level for wider divergence. CACTUS generates large reference-free multispecies alignments. | LastZ, MultiZ, Mashmap<br>CACTUS<br>Ragout<br>SynMap<br>HalSynteny<br>Circos<br>Genomicus<br>Evolution Highway |
| Repeat content and evolution | Catalogs of simple sequence repeats and transposable elements will be generated as part of the genome annotation. Repeats can change genome size, gene content and gene regulation of a genome. | Repeatmasker and Repeat Modeler<br>REPET<br>MITE-hunter and LTRharvest (de novo discovery) |
| Partial or whole-genome duplication | Genome size is also a function of loss and addition based on gene and genome duplication. Partial or whole-genome duplications allow divergent evolution of duplicated genes.  Synteny analysis of internal similarity of genome sections enables analysis of gene gain and loss. | Read depth coverage<br>Alignments and curation<br>Element lengths, homology and copy number<br>Ks Plots<br>Synteny Analysis (e.g. via CoGe) |
| Species trees | Species trees are required for analyses such as calling evolutionary constraint, detecting positive selection, delineating species boundaries/hybridization events, correlating phenotypic | TreeBASE<br>Open Tree of Life<br>FigShare (data repository)<br>Fast Tree |

| | | |
|---|---|---|
| | change with genetic change and inferring evolutionary relationships. | RAxML<br>IQ-Tree<br>PhyloBayes, HyPhy, |
| Evolutionary constraint and accelerated evolution | Depending on power of the sample set, evolutionary constraint can be detected as low at single-base resolution, or at less power with fewer species. Constraint scores help identify coding regions, ultra-conserved elements, and enhancers, promoters, and insulators. | GERP<br>Phastcons<br>PhyloP<br>SweeD |
| Analysis of gene content, gene family expansion and selection on protein-coding genes | Understanding gene content change and underlying sequence selection in a phylogenetic phenotypic context is critical to understand species evolution. | Orthodb<br>PhylomeDB<br>CodeML<br>Conserved motif identification by the local multiple Em (expectation Maximization) for motif elicitation (MEME) |
| Analysis of non-coding transcripts | Non-coding transcripts have a key role in genome regulation and function. These include both IncRNAs and miRNAs and will be identified as part of the genome annotation process. Non-coding transcripts typically evolve more rapidly than protein-coding genes, usually requiring species specific transcript data for identification. | FEELnc<br>Rfam Covariance Models (v13.0)<br>tRNAscan-SE (v1.3.1) |
| Intraspecific variation, conservation, biodiversity and adaptation | Genome reference and population data are required to guide conservation efforts. Historical population sizes can be estimated, and together with additional sequence data, can facilitate the accurate definition of populations and geographic regions most in need of protection. | Runs of homozygosity<br>Heterozygosity GENHET<br>tests for HWE ARLEQUIN<br>Fst<br>Estimation of population histories<br>Reconstruction of multi-generational pedigrees SEQUOIA<br>Genetic structure and admixture STRUCTURE<br>Marker generation |
| | | eDNA metabarcoding |

| | | |
|---|---|---|
| Supporting environmental DNA and/or ecological samples | eDNA analysis allows the characterization and analysis of threatened and non-threatened species within ecosystems. The EBP accelerates this by generating a high-quality digital reference library enabling identification of eDNA sequences. | Shotgun sequencing<br>UPARSE<br>DADA2<br>Blast |