Sea Lice Research Centre

WP5
LiceBase Introductory Course

Part 2: Term and definitions

Michael Dondrup

Bergen 03.10.2014

Oslo 6.10.2014

sfi Senter for forskningsdrevet innovasjon

Terms and Definitions

- Assembly: LSalAtl2s, scaffolds, contigs, chromosomes, landmarks, errors
- Genome annotation, gene prediction, errors: Ensembl
- Gene models: genes, mRNA, exons, UTR's etc.
- Functional annotation
- Evidence
- Genome curation

Assembly

- The result of best effort to reconstruct the genomic DNA or transcript sequence of an organism from the short sequence fragments from whole genome/transcriptome shotgun sequencing, NGS, etc. (LSaIATL2s)
- **reads**: the smallest single, contiguous sequence fragments from the sequencer
- **contigs**: consensus sequences of assembled overlapping DNA sequence reads (ACGT)
- **super contigs, scaffolds**: collections of contigs with relative orientation and gaps (NNNN), main landmarks used in LSaIAtl2s, often also addressed as chromosome
- **chromosome**: eukaryote replicon, scaffolds not yet placed
- **strand**: notation of genomic sequence always 5'->3' (forward, +, +1), notation of opposite strand sequence: 3'->5' (reverse, -, -1) strand for scaffolds is (more or less, unless placed) arbitrary
- **errors**: unavoidable imperfections of the assembled sequence. potential errors in the assembly include, gaps, missing regions, scaffolding errors, mis-assemblies, sequence errors, chimera
- **coverage, depth**: number of times a single base position or interval is covered by sequence reads (e.g. 180X ('fold')), 'depth' ~= 'average coverage over whole genome'

Genome annotation

- The process of detecting important regions and assigning functions to them

- Gene prediction: (mostly in-silico) process of determining functional regions (aka. protein coding genes, etc.) in the sequence

- Function prediction: (partially in -silico) process of assigning potential gene function to regions

Evidence

- Scientific evidence consists of observations and experimental results that serve to support, refute, or modify a scientific hypothesis or theory, when collected and interpreted in accordance with the scientific method. (wikipedia:evidence)

- Bioinformatics predictions (Blast, HMMER, PFAM, InterPro,...)

- Experimental data: rt-PCR, RACE, RNA-seq, ChIP-seq, proteomics, metabolomics, knock-down positive phenotype (RNAi, CRISPR)

- negative phenotype is **not** evidence

- negative results are **not** evidence (e.g. no RNA-seq coverage does not imply gene does not exist)

Genome curation

- The process of improving or maintaining the genome sequence and annotation
- Curation must always be based on evidence
- Examples: updating assembly, changing gene models, adding exons, UTR's, novel genes, changing functional annotation