The Golden Jackal (Canis aureus) reference genome assembly and annotation

Endre Barta

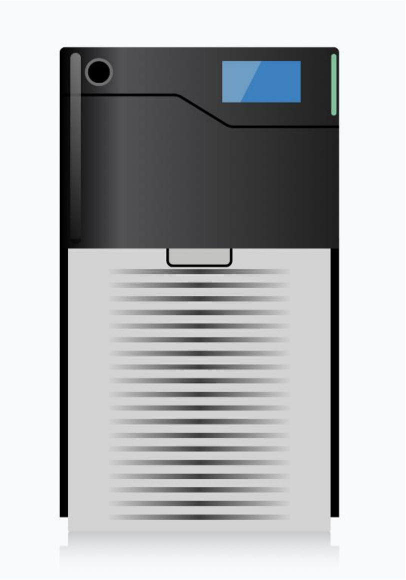
University of Agriculture and Life Sciences
University of Debrecen

Why do we need reference genome sequences?

- Understanding the biology of a species
  - Species specific genes
  - Species specific gene variations
  - Species specific gene regulation
- Understanding the evolution of a species
  - Gene evolution
  - Chromosomal evolution
  - Behavioral evolution etc.
- Population genomics studies
  - Finding markers
  - Using WGS for high resolution population mapping

How to assemble a reference genome

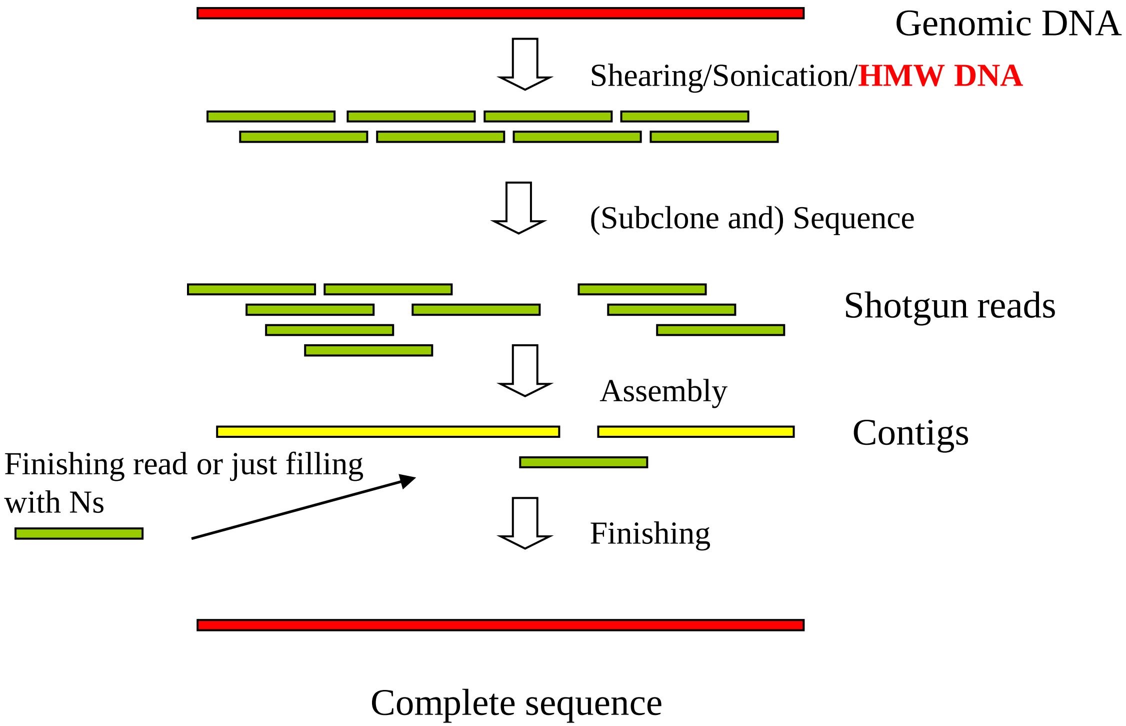
- Human genome (and other subsequent model animal and plant genomes):
  - Intensive genetic and optical mapping followed by Sanger sequencing of BAC clones (Human Genome Project)
  - Shotgun sequencing by Sanger and assembling using supercomputer (Celera genomics, Craig Venter)
- In the post genome (next generation sequencing) era:
  - Illumina short read paired-end and mate pair sequencing and assembling using sophisticated algorithms
  - Nanopore (long read) and illumina (short read) sequencing
  - **Pacific Biosciences (PacBio) High fidelity (hifi) long read sequencing**



Challenges in reference genome sequence assembly

- Which sequencing method to use (how many money we have)?
- Which assembly method to use?
- How to annotate?
- Do we have enough computing capacity?

The process of genome assembly



Genomic DNA

Shearing/Sonication/**HMW DNA**

(Subclone and) Sequence

Shotgun reads

Assembly

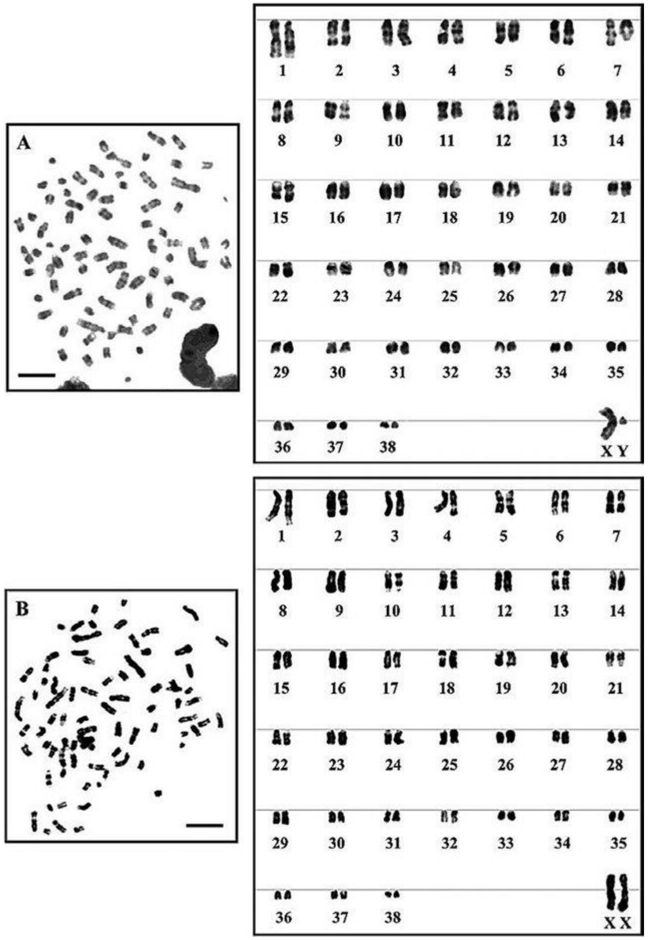
Contigs

Finishing read or just filling with Ns

Finishing

Complete sequence

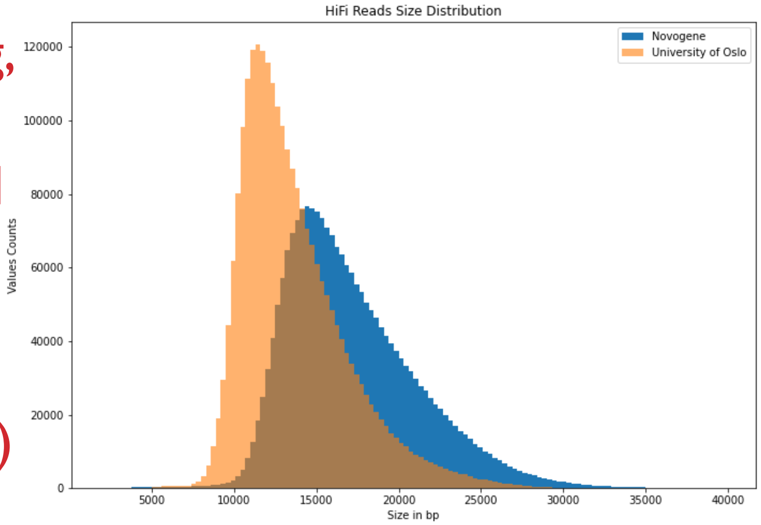
- The final aim is having one sequence for each chromosomes (including the mitochondrion)
- This equals altogether 79 sequences (76 autosomal + 2 sex (X and Y at males) + 1 mitochondrial)
- But in genomics we work with a consensus reference genome sequence (38 autosomal + two sex + the mitochondrial sequence)



Tanomtons et al, Cytology, 2015

Getting the tissue samples for sequencing

- Male animal shot during a regular hunting
- Fresh tissue sample were taken
  - Blood (EDTA tubes) for genome sequencing, Hi-C and RNA-seq
  - Spleen, heart, muscle, lung, testis, liver (liquid nitrogene), RNA-seq
  - HMW DNA were isolated and sequenced using PacBio CCS (hifi) technology (Novogene, and University of Oslo, Norway)
  - DNA was isolated and sequenced using PCR-free illumina 2x150 bp technology (Novogene)
  - Nuclei were isolated from the blood and Hi-C paired-end library were prepared in the University of Debrecen and sequenced in the University of Pécs (iBioSciences)



Samples were provided by Erika Csányi
DNA, and RNA isolation: Viktor Stéger, MATE
Hi-C analysis: Éva Nagy and Lóránt Székvölgyi, University of Debrecen

Reference genome sequence assembly

- The project became the part of the ERGA (European Reference Genome Atlas) pilot project

Ann Mc Cartney and Giulio Formenti
https://www.erga-biodiversity.eu/pilot-project

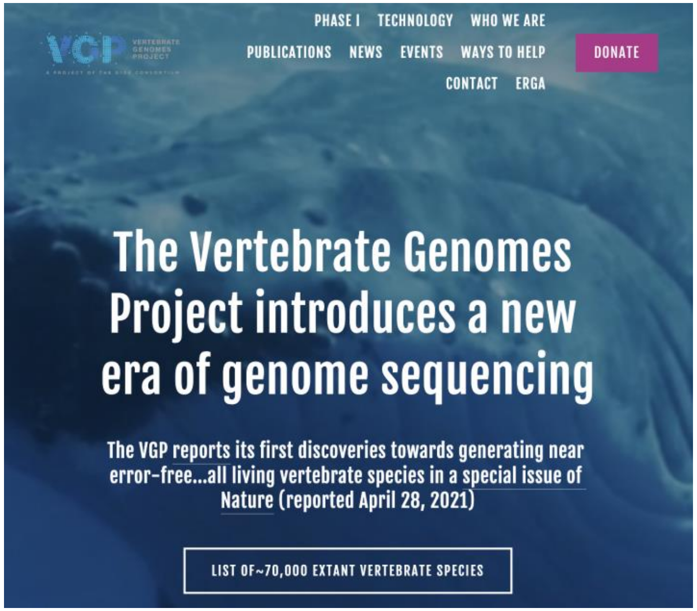
A GENOME ATLAS OF EUROPEAN BIODIVERSITY

The European Reference Genome Atlas (ERGA) initiative is a pan-European scientific response to current threats to biodiversity. Reference genomes provide the most complete insight into the genetic basis that forms each species and represent a powerful resource in understanding how biodiversity functions. With approximately one fifth of the ~200,000 European species at risk of extinction, we need to act fast and together to generate high-quality complete genome resources in large scale.

Tweets from @erga_biodiv

European Reference Ge...
@erga_biodiv · Oct 26

Our #ERGAMemberOfTheWeek is @Leclere_L, researcher @CNRS @cnidevolab. Lucas is the species ambassador for a highly abundant Mediterranean hydroid species. This reference genome will be used to address



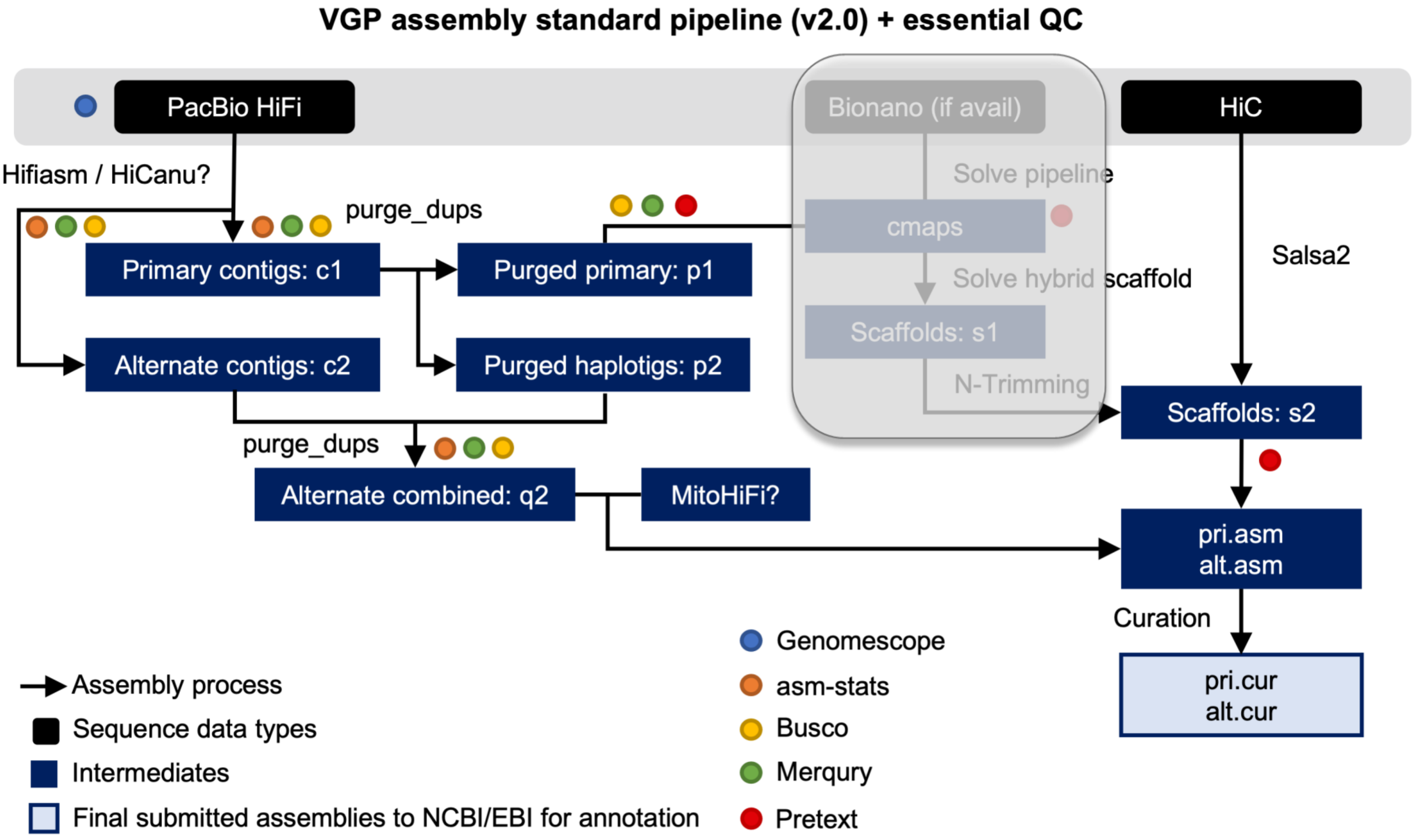
The Vertebrate Genomes Project introduces a new era of genome sequencing

The VGP reports its first discoveries towards generating near error-free...all living vertebrate species in a special issue of Nature (reported April 28, 2021)

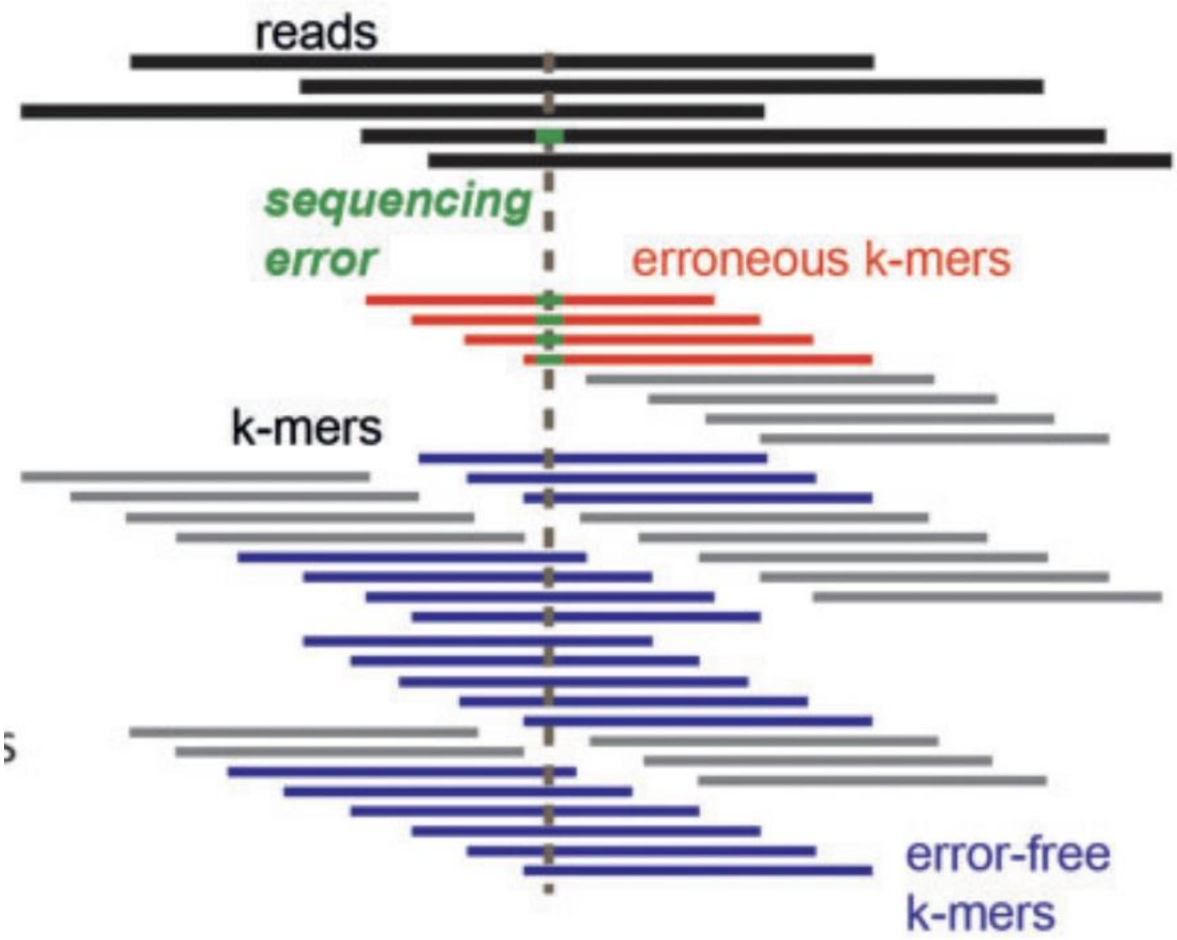
- The genome assembly pipeline was acquired from the Vertebrate Genomes project

Giulio Formenty
https://vertebrategenomesproject.org/

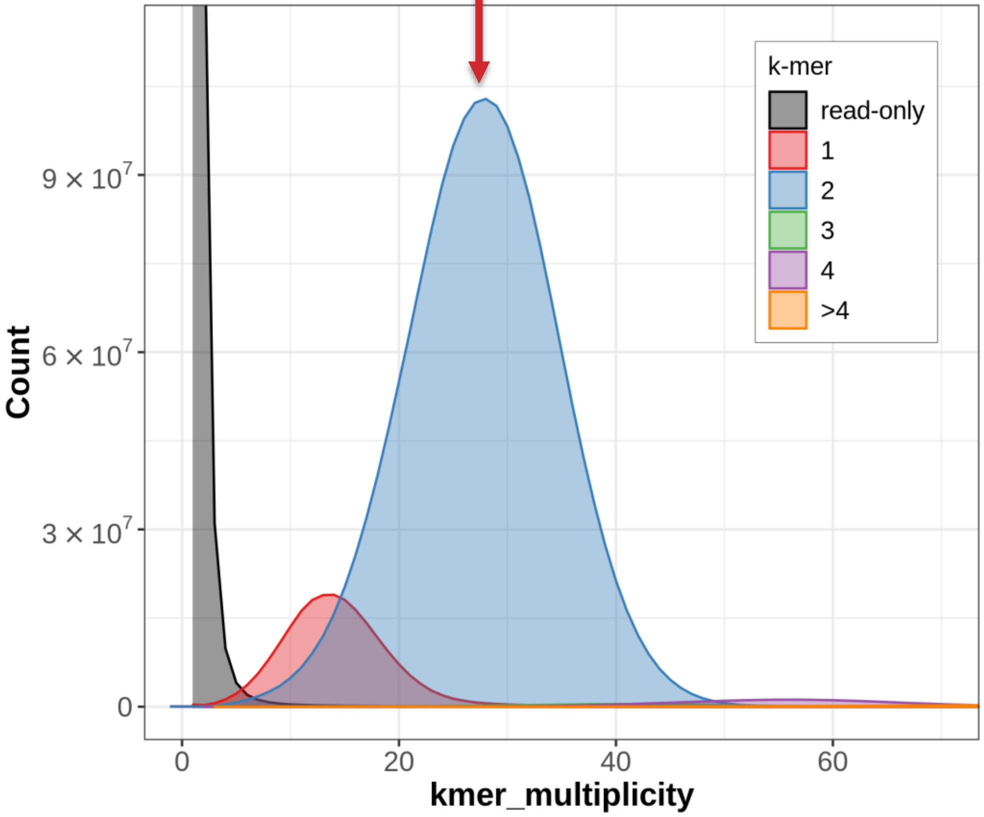
The bioinformatic pipeline used for the assembly



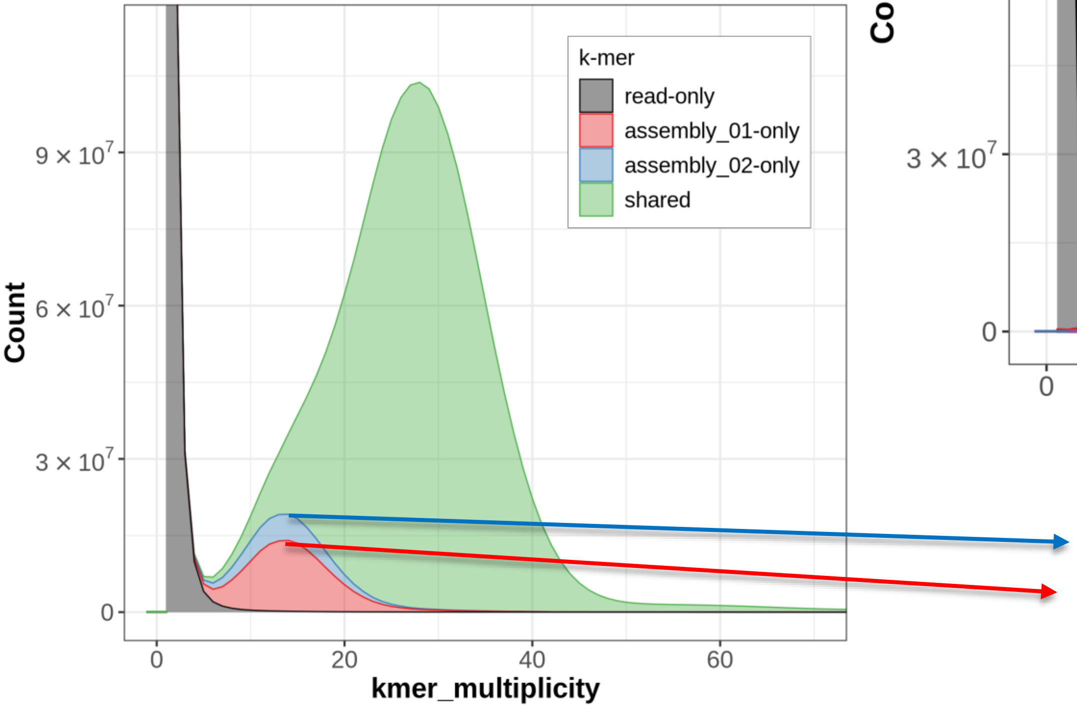
Quality control during the assembly (kmer analysis)



Quality control during the assembly (kmer analysis)

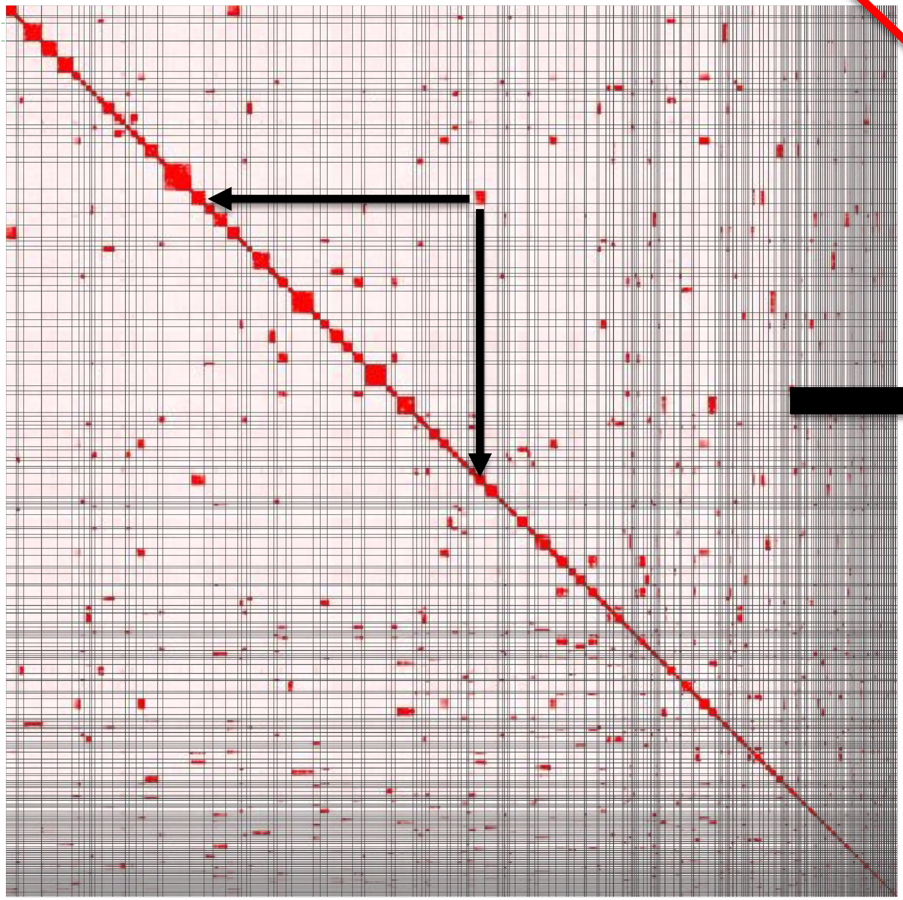


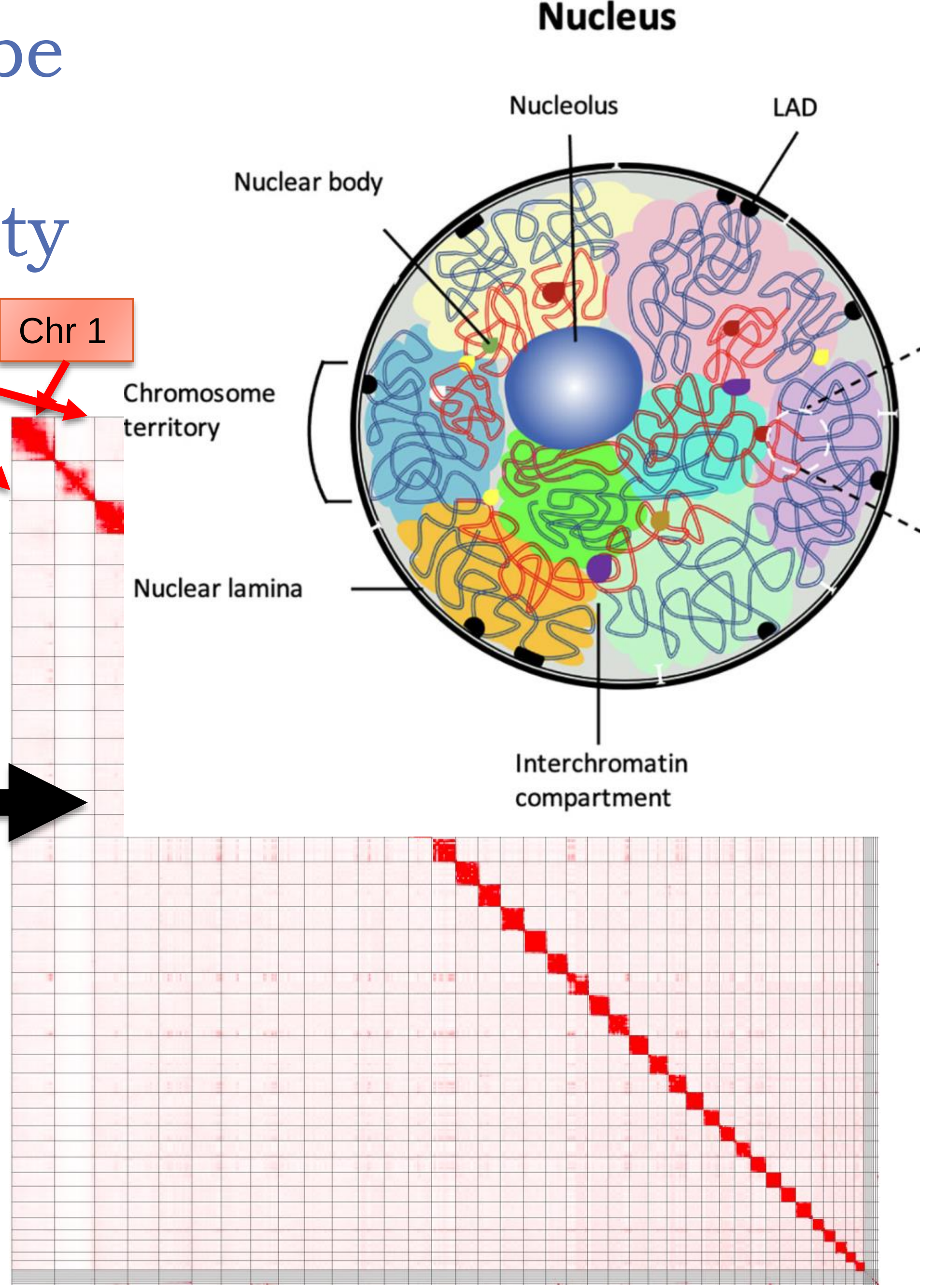
PacBio hifi read coverage (28x)



Unique parental kmers
Haploid from the haploid assemblies

Hi-C contact maps can be used for haplotyping, scaffolding and for quality control



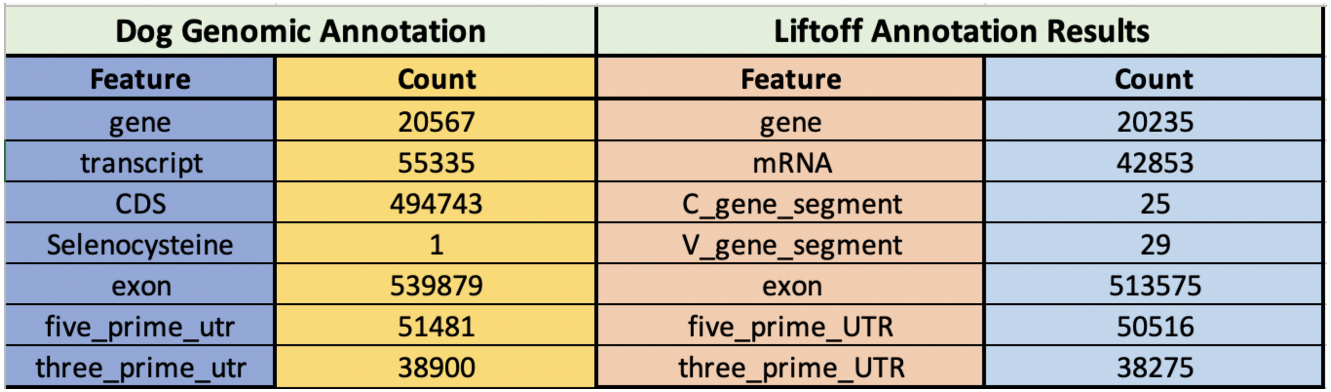


(Almost) final assembly

- 92 scaffolds
  - 39 of them roughly corresponds of the given dog chromosomes
  - Y chromosome is the most problematic
  - The remaining ones are small (<20kb) sequences
- Thanks to the ERGA, the scaffolds will be curated at the EBI by experts for getting the final chromosomal level assembly
- The final assembly will be immediately put into the ENA database (and will be freely accessible)

(Preliminary) annotation of the scaffolds

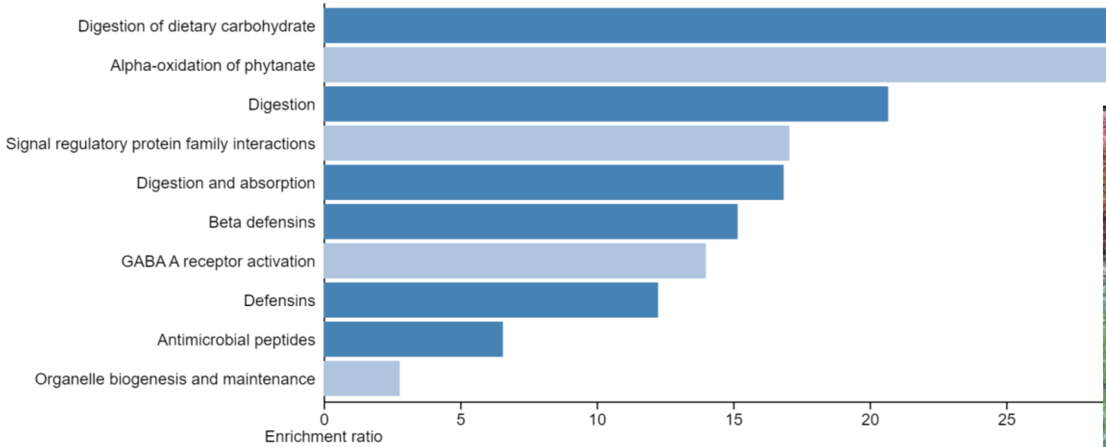
- Annotation means assigning coordinates (chr:start:end) to the different features (Genes, exons, introns, UTRs, repetitive sequences, regulatory regions etc.)
- Three ways of annotation:
  - Based on RNA-seq (problem: not all genes are expressed)
  - Ab initio predicting genes (very difficult, high false positive and false negative ratio)
  - Based on the annotation of a close relative (Dog annotation is obviously available)



Dog Genomic Annotation		Liftoff Annotation Results	
Feature	**Count**	**Feature**	**Count**
gene	20567	gene	20235
transcript	55335	mRNA	42853
CDS	494743	C_gene_segment	25
Selenocysteine	1	V_gene_segment	29
exon	539879	exon	513575
five_prime_utr	51481	five_prime_UTR	50516
three_prime_utr	38900	three_prime_UTR	38275

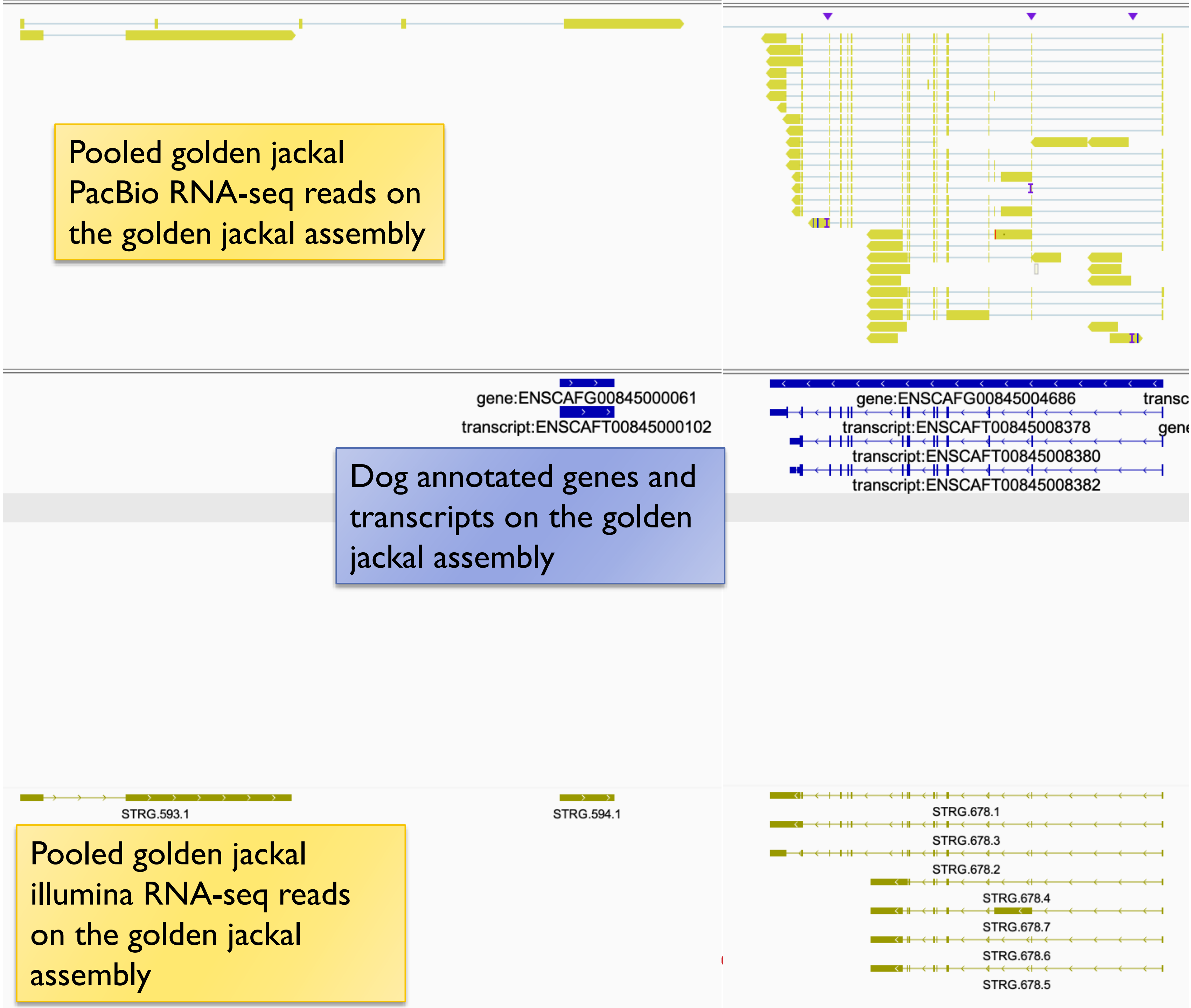
Only 332 dog protein coding genes were not found in the golden jackal assembly

Which genes are present in dog but not in golden jackal?



- 323 dog genes could not be found in the Golden jackal genome
- Many of them related to the digestion (e.g. amylases) and immune defense
- Of course, this is not a surprise as we know that during domestication dogs acquired the ability to digest starch

RNA-seq based annotation



Pooled golden jackal PacBio RNA-seq reads on the golden jackal assembly

Dog annotated genes and transcripts on the golden jackal assembly

Pooled golden jackal illumina RNA-seq reads on the golden jackal assembly

Summary

- We have successfully sequenced the golden jackal genome
- We made the primary, near chromosomal level reference genome assembly
- The assembly can be annotated using the dog annotation and our RNA-seq results

This primary assembly is available upon request.
The final chromosomal assembly and annotation will be available soon

Having the reference genome sequence will allow high resolution population genomics studies at the golden jackal.

We have already collected and whole genome sequenced 11 golden jackal samples from Hungary, Romania and Serbia and got samples from Greece
Thanks to the sample providers (Erika Csányi, Miklós Heltai, Péter Fehér, Viktor, Dusko Cirovoc, Theodoros Kominos and others)

Acknowledments

- MATE Agricultural Genomics and Bioinformatics Group
  - **Tibor Nagy**
  - **Maher Alnajjar**
  - Zsófia Fekete
  - Levente Kontra

- **Erika Csányi** for providing the sample and for the motivation
- **Viktor Stéger** and his group for the wet lab work
- **Lóránt Székvölgyi**'s lab for doing the Hi-C library

Ann Mc Cartney and **Giulio Formenti** from the European Reference Genome Atlas (ERGA)

My group is open for any collaboration in Golden Jackal whole genome sequencing and in the consequent bioinformatic analysis

Barta.Endre@uni-mate.hu