Towards Pangenomics

Fergal Martin

Eukaryotic Annotation Team Leader

EMBL-EBI

What is a pangenome?

What is a pangenome?

- A group of organisms generally contains much more genomic sequence than an individual genome
- A pangenome can be thought of as the collection of genomic sequences that describe the genomes in a particular group
- In eukaryotes this is often though of at the species or population level
- The concept originated in prokaryotes and encompassed both vertical and horizontal transmission

First sight of a pangenome

- In the early 2000s Tettelin et al. realised only 80% the genes of an individual strain of *S. agalactiae* were present other *S. agalactiae* strains
- No single genome could represent *S. agalactiae*
- In fact, this is a general property of prokaryotes

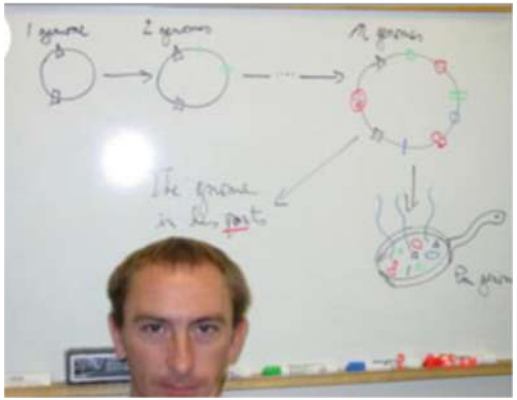
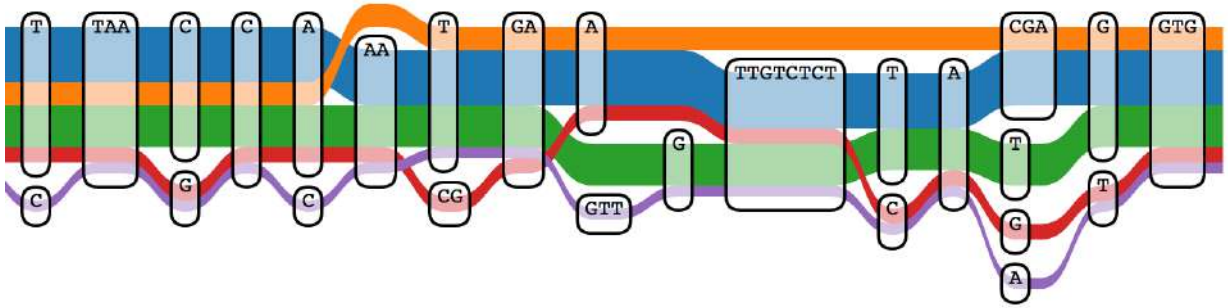


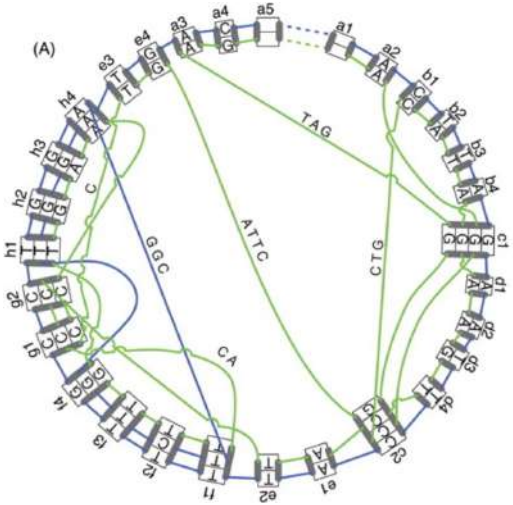
*Photo from Tettelin & Medini "The Pangenome"

Giving structure to a pangenome

- Without modelling the relationships between the sequences and structures, a pangenome is difficult to interpret

- These relationships are often presented as an alignment graph





Why use a pangenome?

- A single reference genome is not adequate to give context to the variation seen across a population/species/clade
- Creates reference bias in downstream analyses
- Reference genomes are generally never perfect
- Even if perfect, some genes or structural features of genes may be entirely absent from the reference

Landscape of Pangenomes

EMBL-EBI

Landscape of pangenomes - Eukaryotic pangenomes

- Prokaryotic pangenomes are a natural by-product of how genetic information flows through prokaryotes

- Eukaryotic pangenomes are less mature, but significant effort on both method development and data generation are well underway

- The major drivers of eukaryotic pangenomes:
  - Human
  - Crop plants
  - Livestock
  - Models

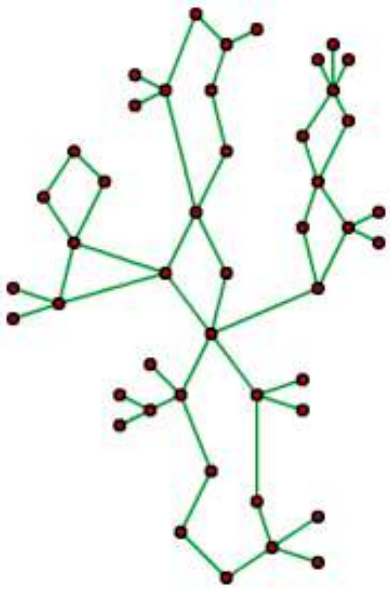
- Current pangenome projects can vary from within a species, within a genus to across families

Landscape of pangenomes - Explicit versus implicit

- Projects such as Human Pangenome Reference Consortium (HPRC) and Pan-Oryza are examples of explicit pangenome efforts
- Other species/groups are implicit pangenome efforts, e.g. dog, pig, chicken
- Different implications, opportunities, many examples of both scenarios already

Landscape of pangenomes - Building graphs

- The concept of a pangenome and pangenome alignment graphs are not the same thing, but heavily linked
- Many of the eukaryotic pangenome efforts have been centred around generating a Cactus alignment
- Not ideal for a species level pangenome
- Not scalable to a dense pangenome
- Efforts in human have looked to find alternative solutions



Landscape of pangenomes - Building graphs

- Minigraph Cactus (Li, Paten):
  - Main analysis for the human pangenome was anchored on Minigraph Cactus
  - Minigraph will quickly model the graph between haplotypes
  - Requires a reference, which has implications
  - Avoids some of the issues of Cactus, which expects a species tree and makes decisions on indels based on it
  - Minigraph cannot model variation < 50bp, running Cactus after Minigraph solves this

- PGGB (Garrison):
  - Also developed as part of HPRC
  - All vs all, reference free approach
  - Output VCF for any genome included in the graph
  - Fast, lots of active development

Landscape of pangenomes - Graph format

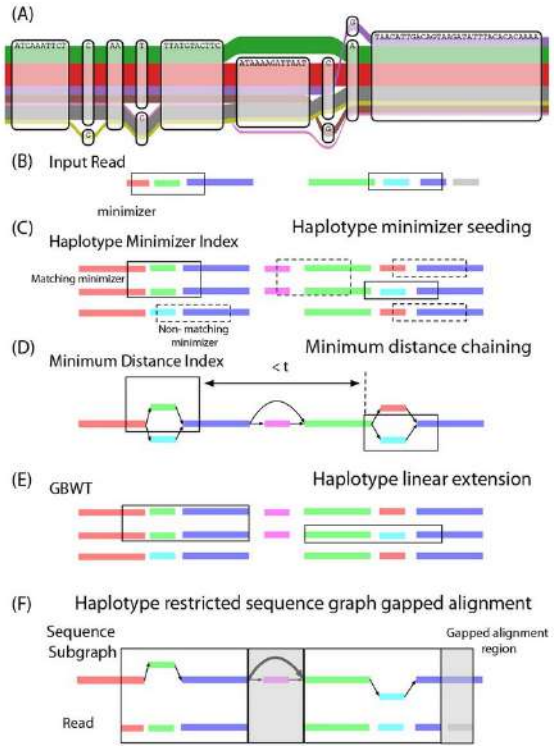
- Graph Fragment Assembly (.gfa) is becoming the canonical graph format
- Adoption in Darwin Tree of Life (complex organelle genomes) and HPRC
- ENA has been working closely with both projects in terms of accepting GFA files

```
H       VN:Z:1.0
S       11      ACCTT
S       12      TCAAGG
S       13      CTTGATT
L       11      +       12      -       4M
L       12      -       13      +       5M
L       11      +       13      +       3M
P       14      11+,12-,13+     4M,5M
```

```
11 ACCTT
12  CCTTGA
13   CTTGATT
14 ACCTTGATT
```

Landscape of pangenomes - Tools

- Tool development still early days in the eukaryotic space
- Lack of intuitive tooling is a significant barrier to transition to pangenomics
- Graph aware tools would provide scalability of analysis
- Not needing to linearise would help storage footprint
- Giraffe, which is a graph-aware short read mapper is an early success



Sirén et al., 2021. PMID: 34914532

EMBL-EBI

Landscape of pangenomes - Users and drivers

- Initial users limited:
  - People who are interested in pangenomes
  - Researchers where the pangenome can answer complex biological questions

- Longer term, clinicians and breeders

- Two clear high impact use cases
  - The human pangenome allows more accurate interpretation of clinical data leading to better outcomes
  - The pangenome of crops/livestock leading to better breeding strategies and greater food security

Landscape of pangenomes - Users and drivers

- Example clinical use case:
  - A clinician has some data relating to a patient's genome
  - These data are searched against the reference pangenome
  - The most appropriate reference path is identified
  - The reference pangenome is layered with functional information to a level we see on GRCh38 now
  - Clinician can confidently and seamlessly use these data to draw conclusion in a manner similar or easier than if they were to perform this on GRCh37/38
  - The use of less biased, more targeted data leads to improved outcomes for human health

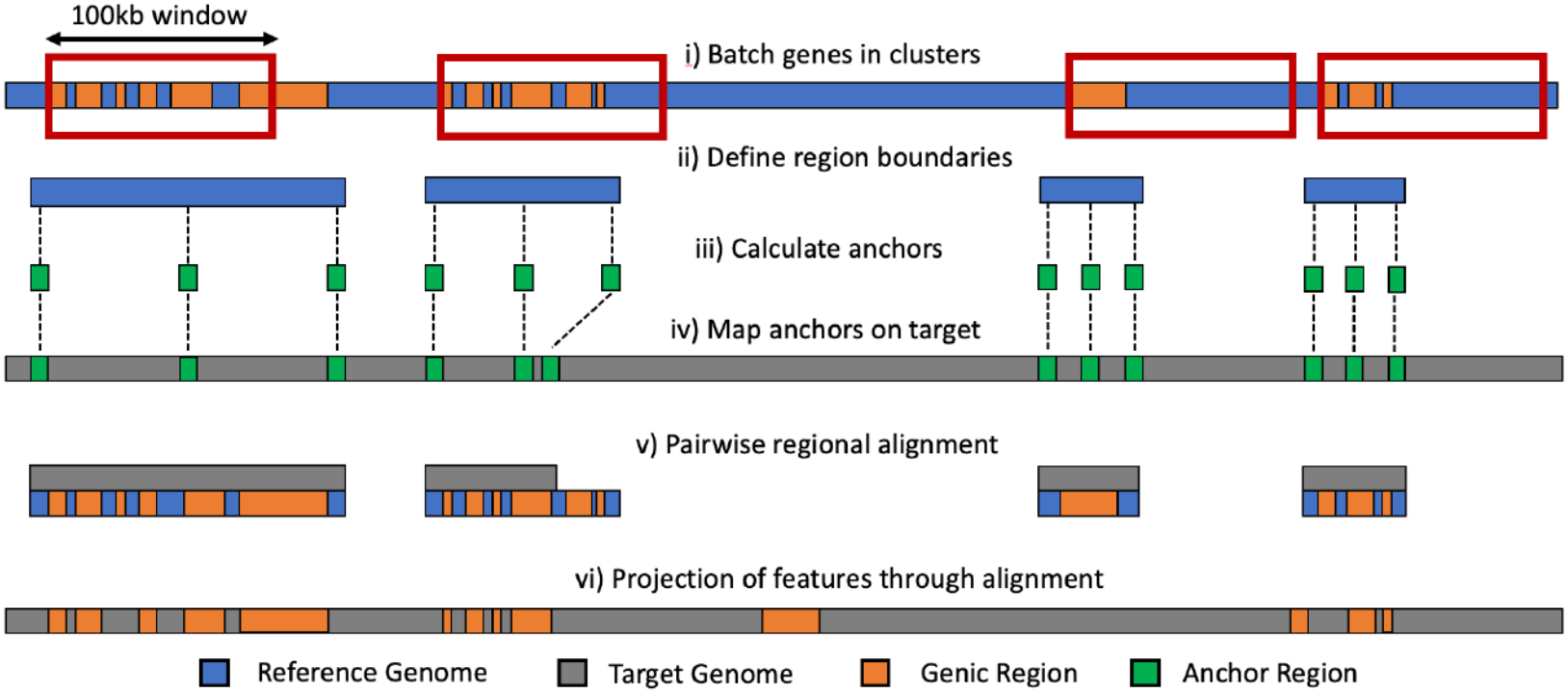
Annotating a pangenome

EMBL-EBI

Annotating a pangenome - approaches

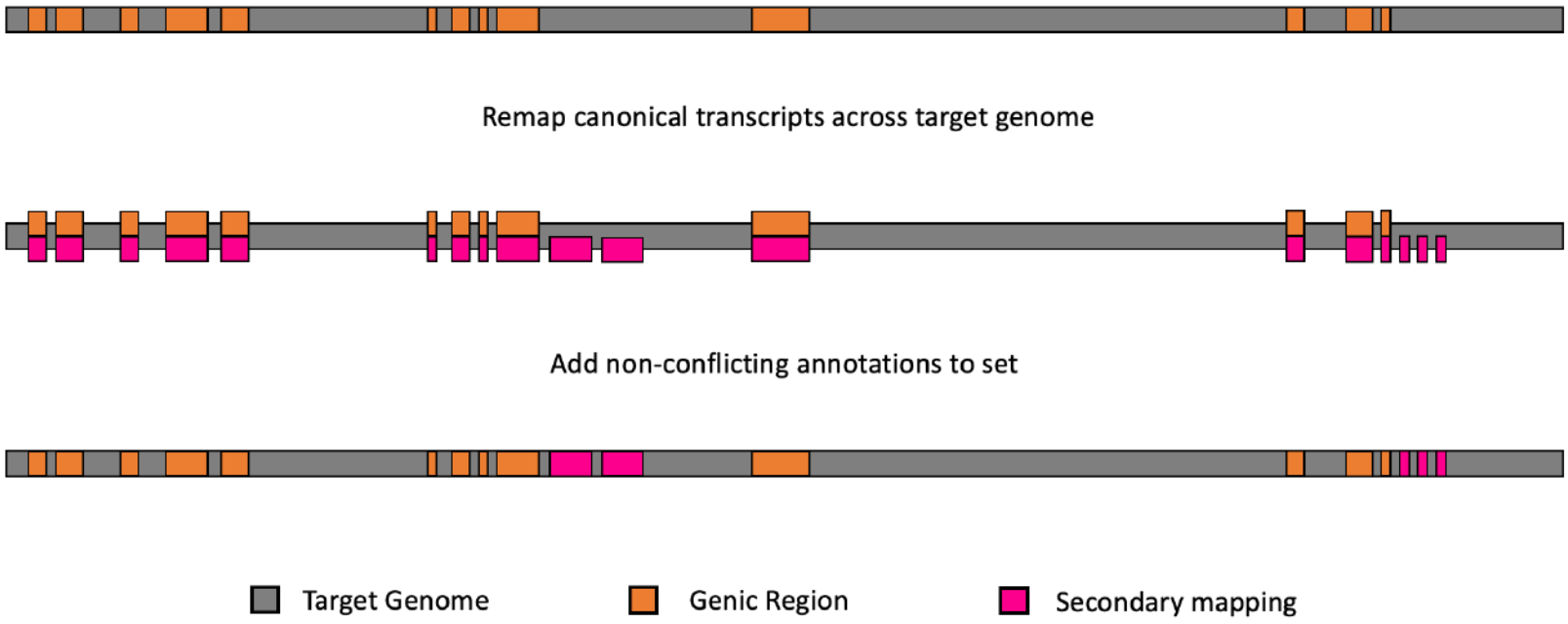
- Most straightforward approach is reference based/biased
- Ground up annotations on individual breeds/haplotypes useful but expensive and more likely to be incomplete
- Best of both worlds approach involves mapping from a well annotated reference while supplementing with targeted transcriptomic and comparative annotation

Annotating a pangenome - primary mapping



EMBL-EBI

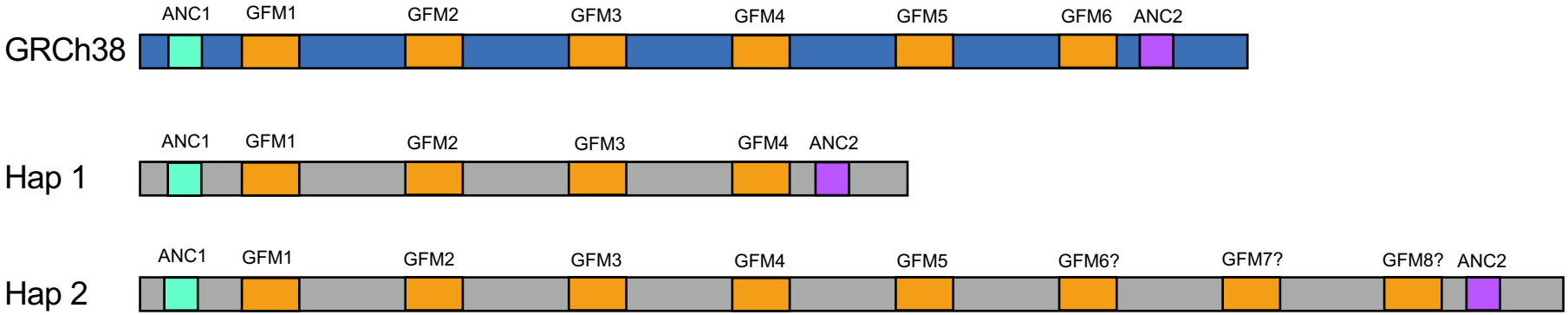
Annotating a pangenome - secondary mapping



EMBL-EBI

Annotating a pangenome - difficulties

- Capturing true novelty
- Assessing what change means
- Gene clusters and CNVs



EMBL-EBI

The Human Pangenome Refence Consortium

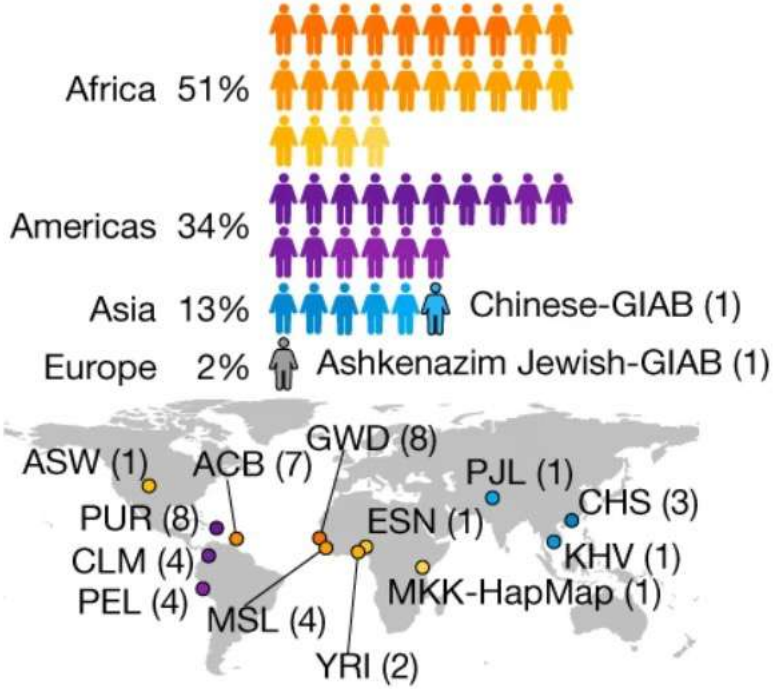
EMBL-EBI

The Human Pangenome Reference Consortium

- An effort to build a draft human pangenome reference
- Currently includes the high-quality genomes 47 individuals
- All genomes are fully phased, diploid assemblies
- The start of addressing the need for better representation of genetic diversity

A Draft Pangenome

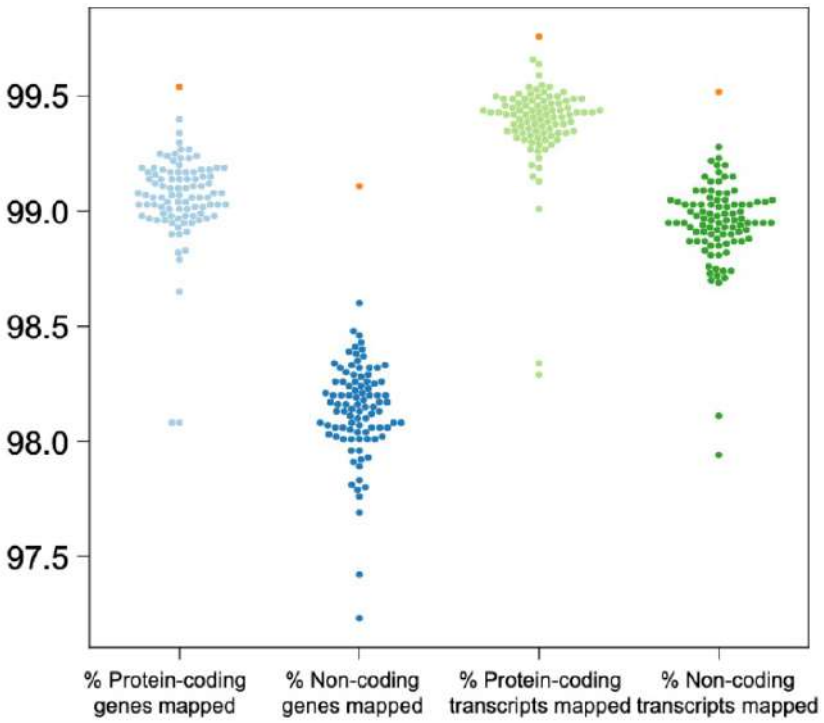
- Several draft pangenome graphs were constructed
- Minigraph-Cactus approach current best in class
- Choice of GRCh38 versus CHM13v2 as reference has some effect mapability to the graph
- ~22M bubbles represent ~20M SNPs, 6.8M indels, ~400K larger SVs
- Significant CNVs in genes related to human health



EMBL-EBI

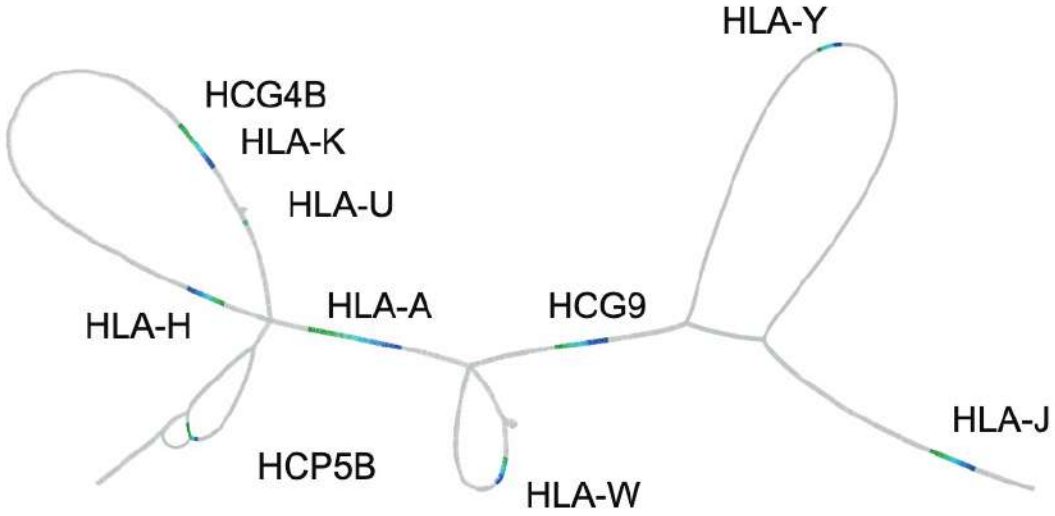
HPRC Annotation Results

- CHM13v2.0 (T2T + Y) shows highest mapping score
- Higher mapping scores for protein-coding genes/transcripts
- Pseudogenes least mappable for both biological and technical reasons
- Gene clusters of paralogous genes cause most issues

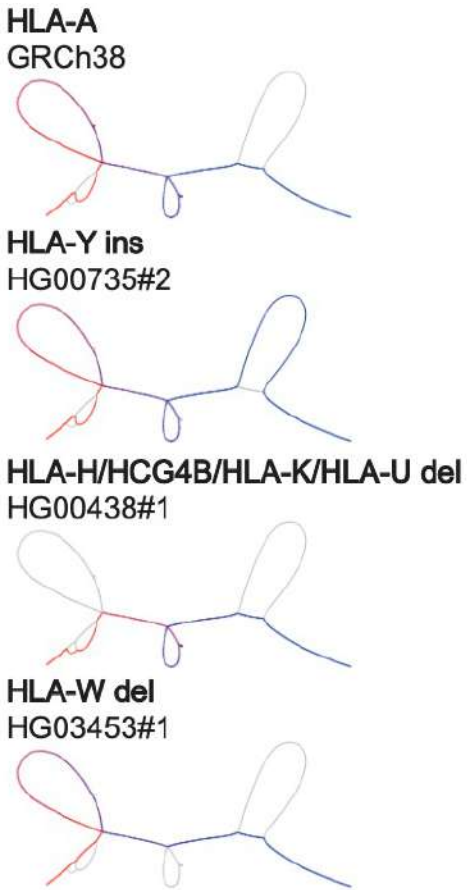


EMBL-EBI

HPRC Annotation Results

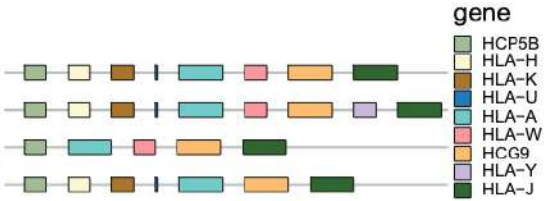


E



F

Count	Frequency	Haplotype name
57	0.63	HLA-A
25	0.28	HLA-Y ins
7	0.08	HLA-H/HCG4B/HLA-K/HLA-U del
1	0.01	HLA-W del



EMBL-EBI

HPRC future directions

- Remainder of phase 1 will go from 47 to 350 individuals (700 haplotypes)
- Phase 2 will add another 200 individuals focusing on genetic diversity in US populations
- An effort to try and push for T2T quality for each haplotype
- Stablilise the pangenome to help with data migration
- High level of interaction with key projects such as GENCODE

Pangenome Annotation Resources

EMBL-EBI

The Eukaryotic Annotation Team

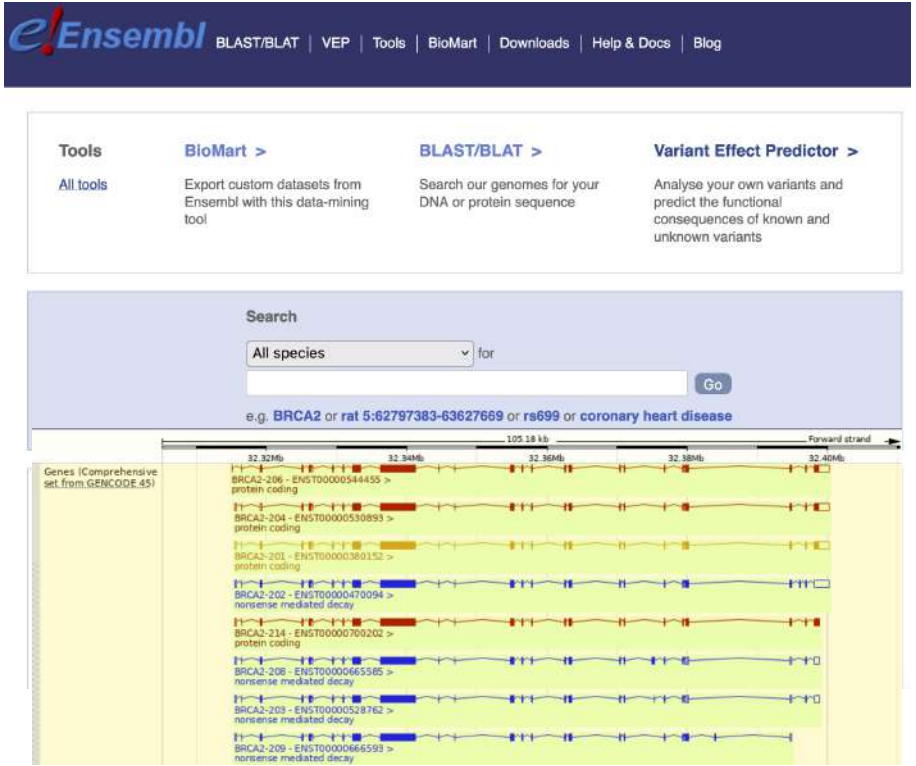
- **Focus:** providing genome annotation and comparative genomics resources for eukaryotes

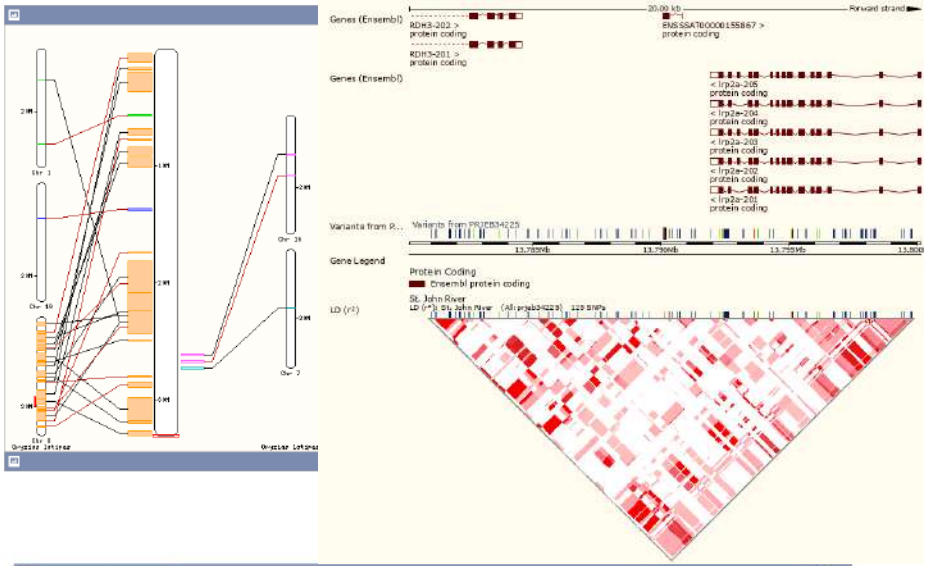
- **Major resources:**
  - GENCODE gene set for human and mouse
  - Automated gene sets for other eukaryotes
  - Repeat libraries and annotations
  - Homologies and gene trees
  - Whole genome alignments

- **Areas of focus:**
  - High quality, expansive resources for popular reference species and pangenomes
  - Scalable support for global biodiversity initiatives

EMBL-EBI

Ensembl





Show/hide columns

Filter

Population	Allele: frequency (count)	Genotype: frequency (count)	Genotypes
ALL:PRJEB34225	G: 0.106 (17) C: 0.894 (143)	CIC: 0.787 (63) CIG: 0.212 (17)	Show
Gaspe of New Brunswick	G: 0.219 (7) C: 0.781 (25)	CIC: 0.562 (9) CIG: 0.438 (7)	Show
Penobscot River	C: 1.000 (22)	CIC: 1.000 (11)	Show
St. John River	G: 0.094 (10) C: 0.906 (96)	CIC: 0.811 (43) CIG: 0.189 (10)	Show

EMBL-EBI

Ensembl and Pangenomes

- Leading annotation efforts for the Human Pangenome Reference Consortium
- Looking at breeds, strains, cultivars, cell lines and haplotypes
- Major area of interest for livestock, agriculture and aquaculture
- Examples include
  - ~20 pigs
  - ~20 sheep
  - 5 chickens
  - ~ 15 medaka
  - 99 human

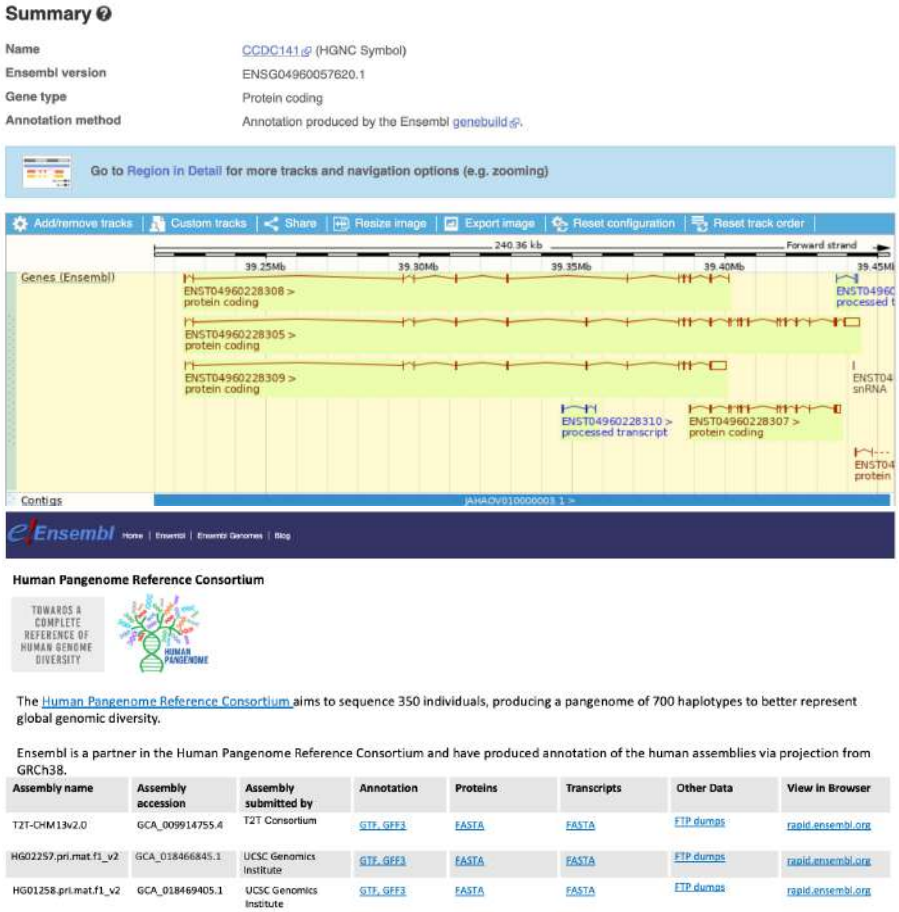
TOWARDS A COMPLETE REFERENCE OF HUMAN GENOME DIVERSITY

HUMAN PANGENOME

EMBL-EBI

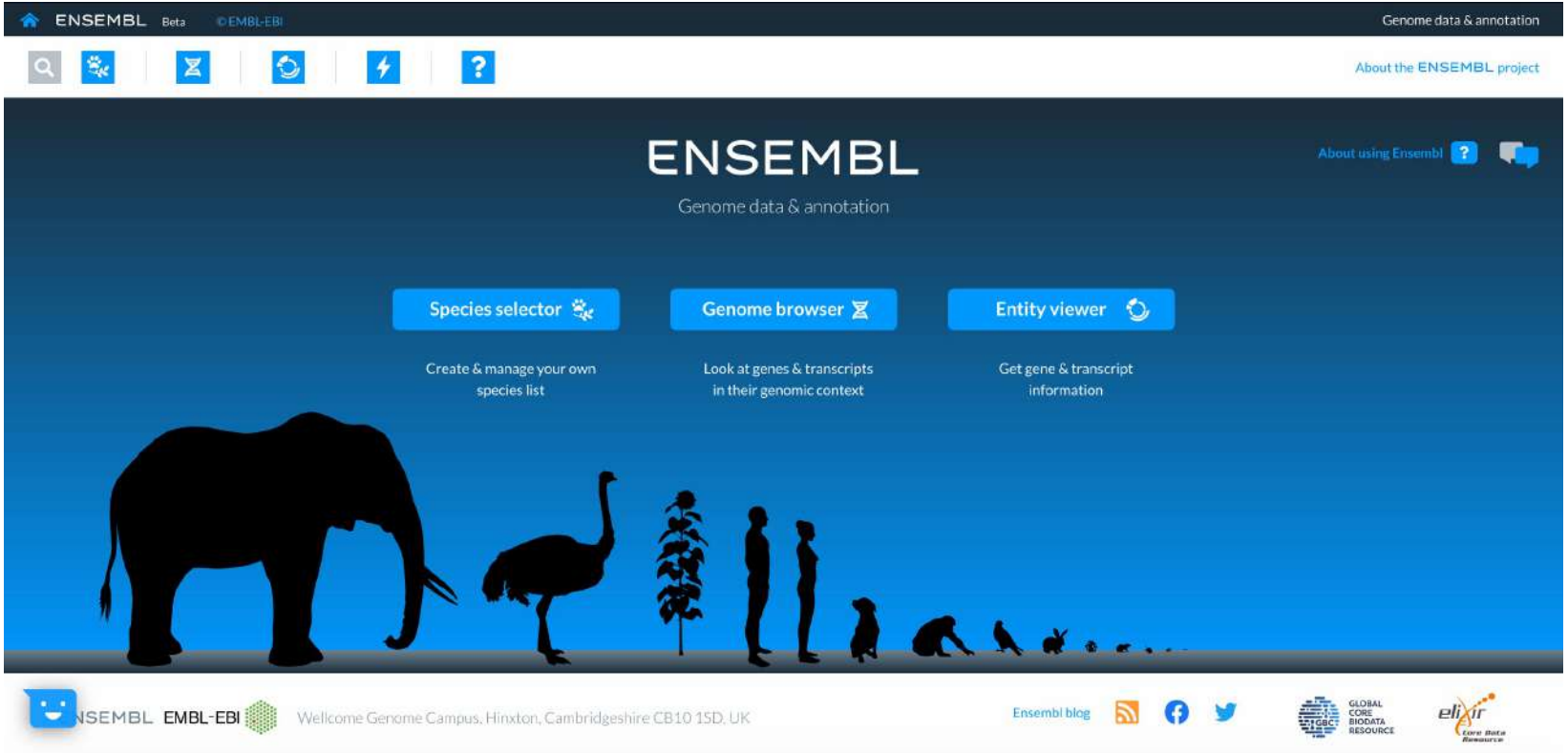
HPRC Data Availability

- Annotated genomes for the 94 haplotypes and CHM13v2.0 assembly can be found on Ensembl Rapid Release
- https://rapid.ensembl.org
- A dedicated HPRC Ensembl project page
- https://projects.ensembl.org/hprc



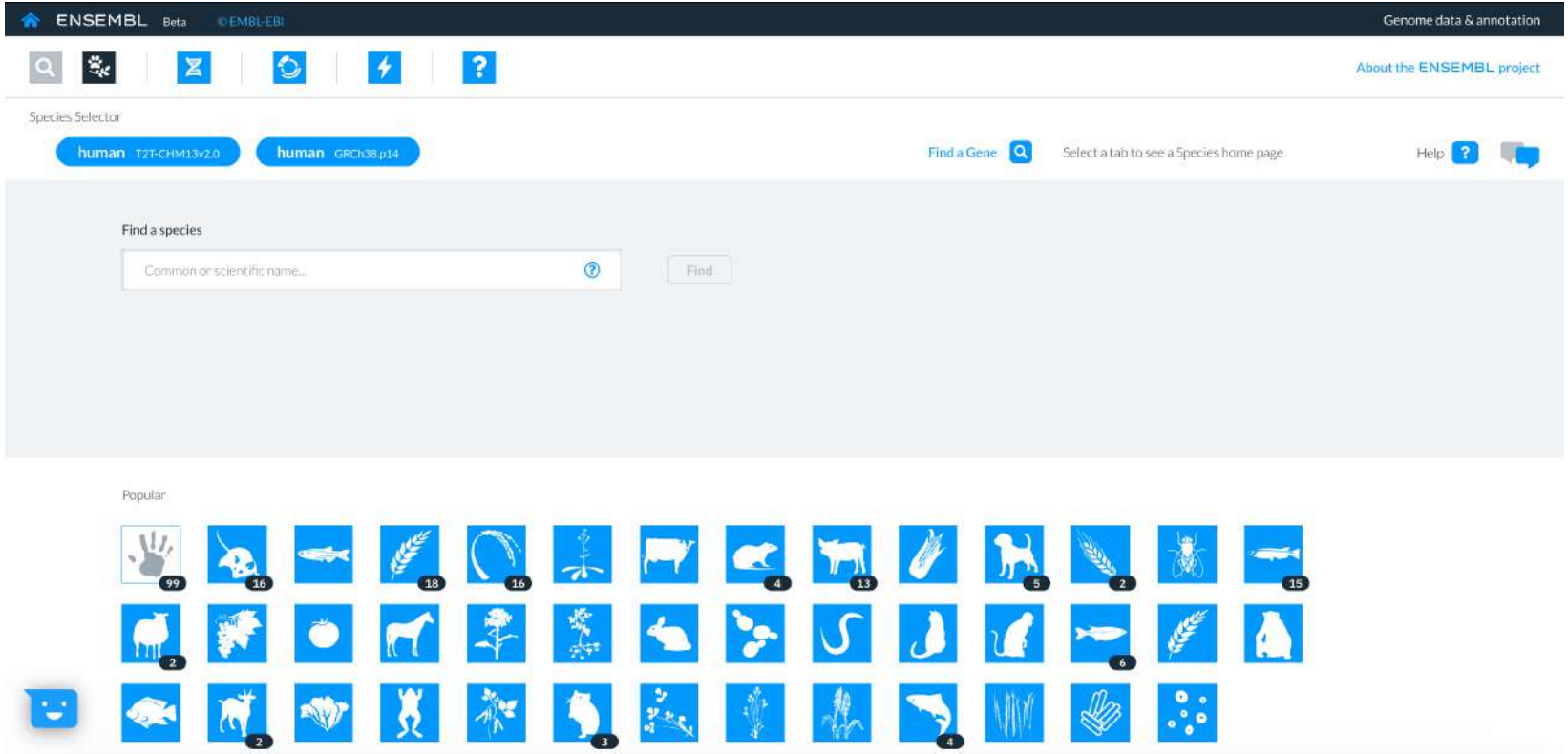
EMBL-EBI

HPRC Data Availability - beta.ensembl.org



EMBL-EBI

HPRC Data Availability - beta.ensembl.org



EMBL-EBI

HPRC Data Availability - beta.ensembl.org

ENSEMBL Beta ©EMBL-EBI

Genome data & annotation

About the ENSEMBL project

Entity viewer

human T2T-CHM13v2.0 | human GRCh38.p14

Change | Help

Gene ZNF277ENSG05220053036.1 | Biotype protein_coding | forward strand | 7:113,525,002-113,662,218

Overview | External references

ENSG05220053036.1

5' 3' forward strand

7 transcripts

View

bp 1 100,000 137,217

Filter & sort | Transcripts | Gene function | Gene relationships

Transcript ID

Ensembl canonical

ENST05220206057.1

Biotype protein_coding | 450 aa | Combined exon length 2,580 bp

7:113,525,002-113,662,218 | ENSP05220092141.1 | Coding exons 12 of 12

View in

Download

ENST05220206062.1

ENST05220206061.1

ENST05220206063.1

ZNF277 ENSG05220053036.1

Gene name

zinc finger protein 277
:parent_gene_display_xref=ZNF277

HGNC:13070

Synonyms

NRIF4, ZNF277P

Attributes

Biotype protein_coding

Find a gene

Function

Other data sets

Publications

Europe PMC

EMBL-EBI

Summary

- Huge growth in terms of pangenome efforts
- Human, primarily via the HPRC, has led to a largescale effort to build a reference pangenome and associated tools
- Already many efforts underway in the agricultural space on pangenomes
- Still very early days for clear use cases, workflows, tools and visualisations
- Adoption of pangenomics will take many more years, needs stable pangenomes, tools and clear use cases

Acknowledgements

- Everyone in the Eukaryotic Annotation Team and Ensembl

- Project partners on HPRC, especially Benedict and the team at UCSC

Co-funded by the European Union

Questions?

Landscape of pangenomes - Users and drivers

- Example agricultural use case:
  - A breeder has a set of variants associated with desirable traits such as drought/disease resistance or yield
  - Breeder assesses paths for breeds/haplotypes within the pangenome that best fit the traits under consideration
  - Creates a strategy for breeding/cross breeding that captures more of these desired traits in the offspring