PanSel: a tool finding conserved/divergent regions

Matthias Zytnicki

PanSel: a tool finding conserved/divergent regions

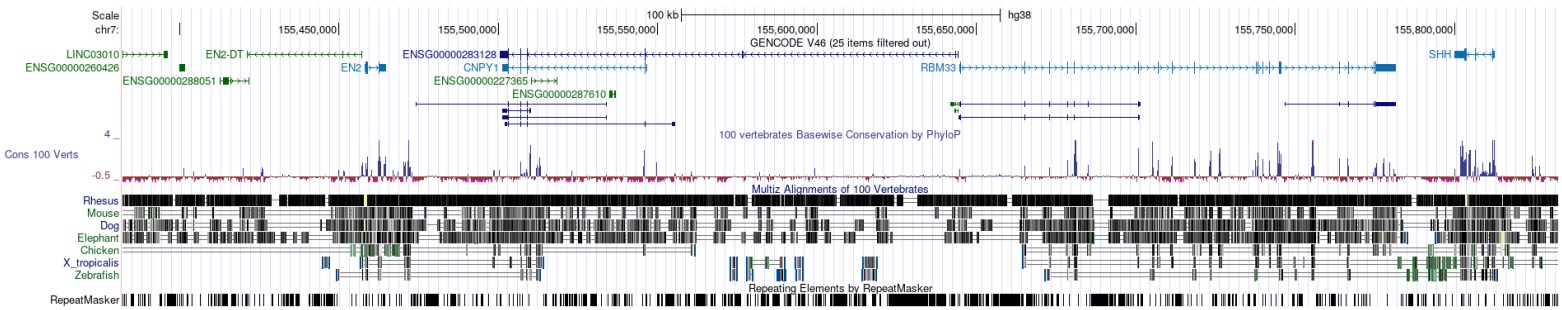
Matthias Zytnicki

MIAT, INRAE

MIGGS 2024

Introduction

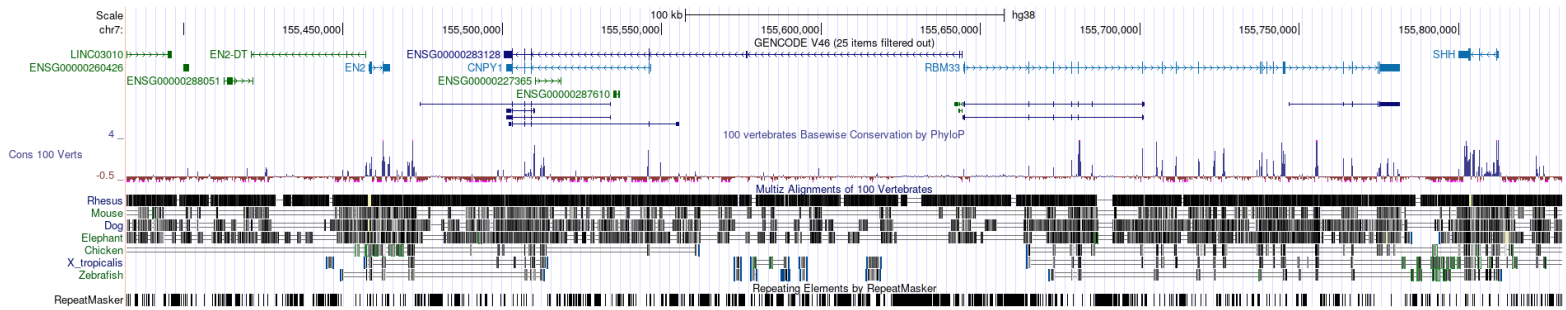
- Transition from a unique (reference) genome to a pangenome requires adaptation of many downstream tools (mapping, visualization, etc.).
- Among them: how to find conserved/divergent regions?
- General aim: is my gene of conserved/divergent in the species?



UCSC genome browser

Current tools GERP/GERP++, PhastCons, PhyloP, etc)

- Use a global inter-species aligment, and a phylogeny.
- Infer the branch length at each position.
- Quantify the conservation.



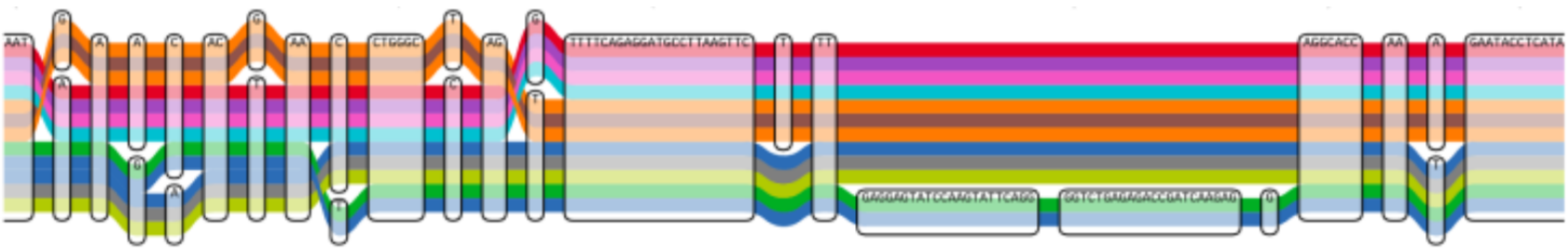
UCSC genome browser

Back to pangenome graph

- Most of the information is in the graph.
- Phylogenetic tree is probably not defined.
- I did not find a way to get a base-level score (please feel free to suggest!).
- Rough idea:
  - use a sliding window
  - compute a simple "edit distance" between each pair of haplotypes
  - compute the mean edit distance: this is your score!
- Used PanSel on the Draft Human Pangenome, with the MiniGraph-Cactus GFA files. Bin sizes of 1k, 10k, 100k.

Step 1: find anchor nodes

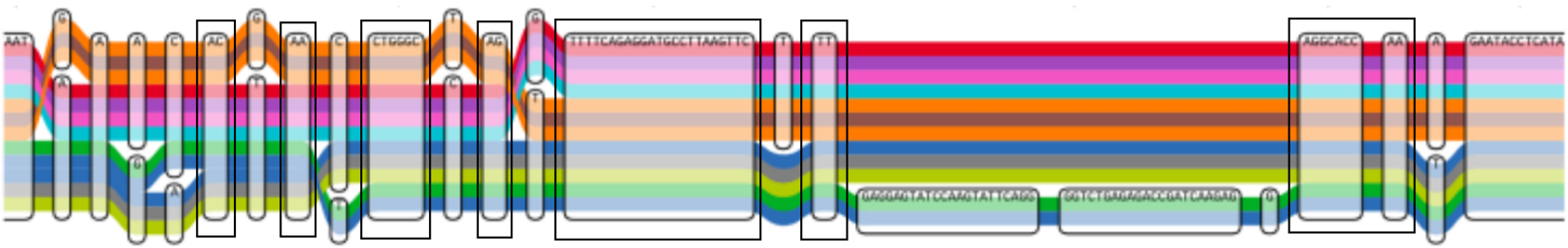
- Parse all the paths.
- Get all the nodes that are shared by all paths.
- Find pairs distant by w (the bin size)



Sequence Tube Map

Step 1: find anchor nodes

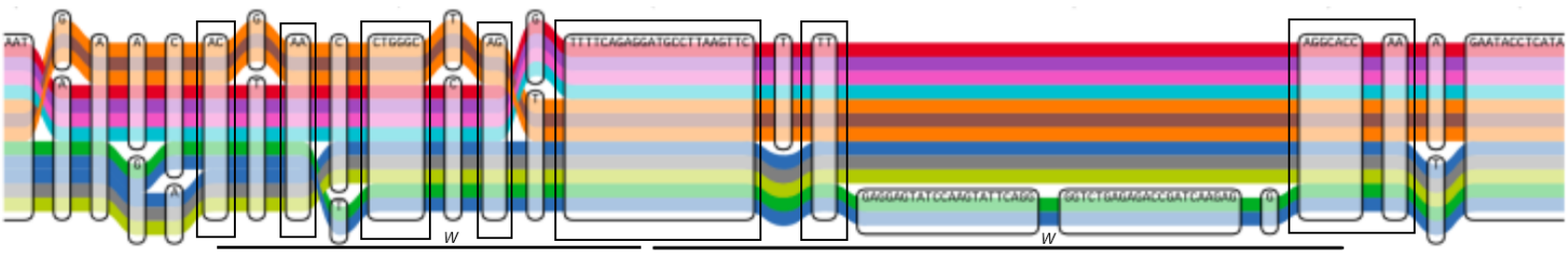
- Parse all the paths.
- Get all the nodes that are shared by all paths.
- Find pairs distant by w (the bin size)



Sequence Tube Map

Step 1: find anchor nodes

- Parse all the paths.
- Get all the nodes that are shared by all paths.
- Find pairs distant by w (the bin size)



Sequence Tube Map

Step 2: compute an weighted Jaccard index

Considering a bin:

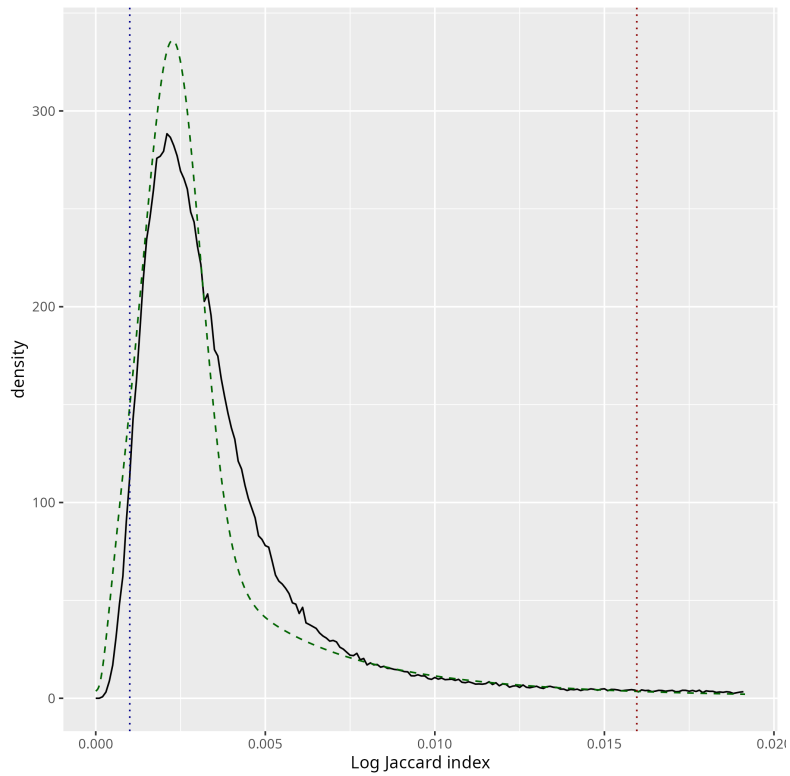
- Extract all the sub-paths.
- For each pair of paths P_1, P_2, compute

$$\frac{\sum_{n \in P_1 \cap P_2} |n|}{\sum_{n \in P_1 \cup P_2} |n|}$$

- Compute the mean of it: this is the score for the bin.

Step 3: fit the distribution

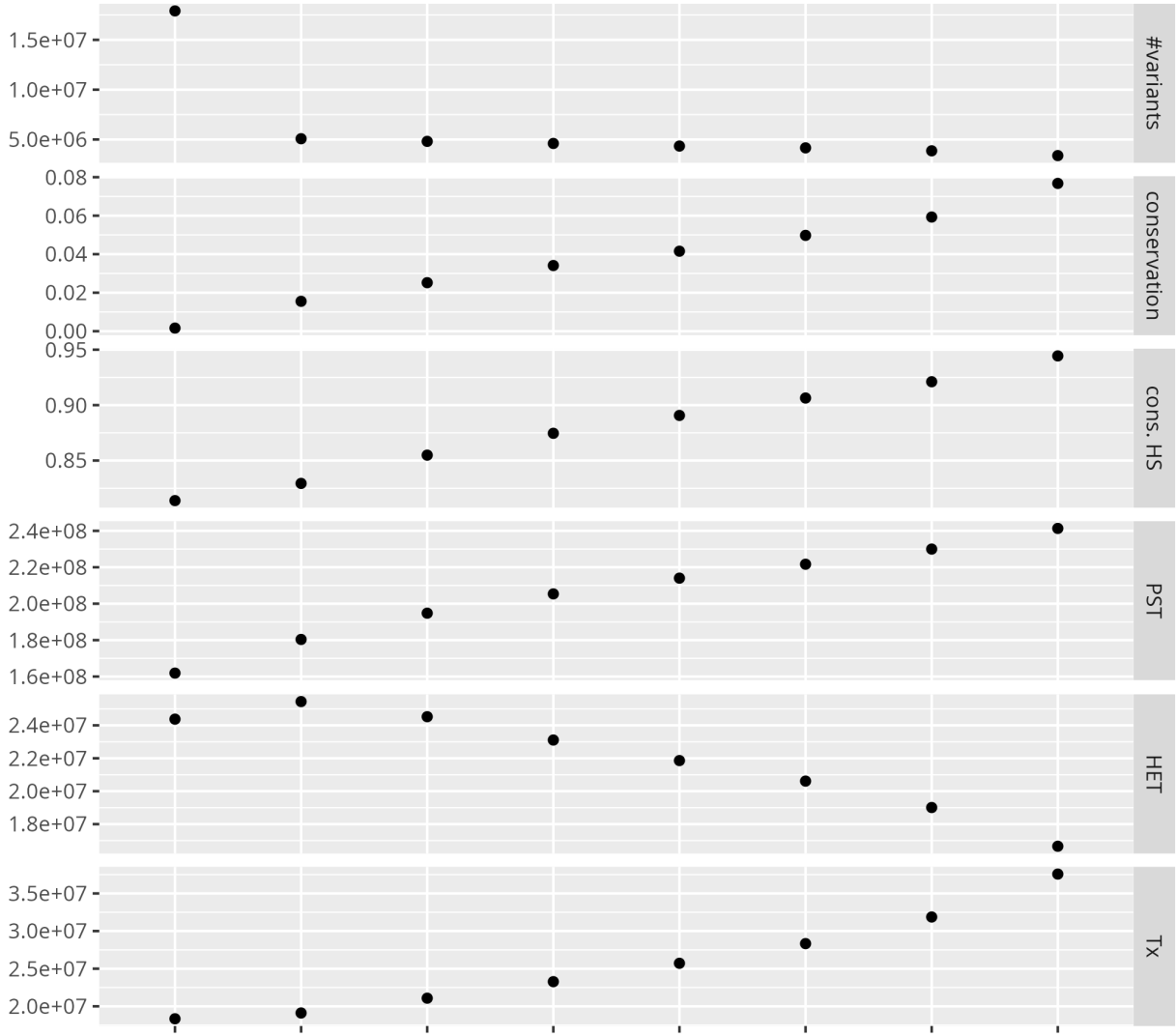
- Transform $[0, 1]$ values to $\mathbb{R}_+$ using $-\log$ transform.
- Fit curve using a mixture of normal and log-normal (divergent distribution).
- Get a p-value for divergent and conserved distributions.



$w = 10\text{k}$

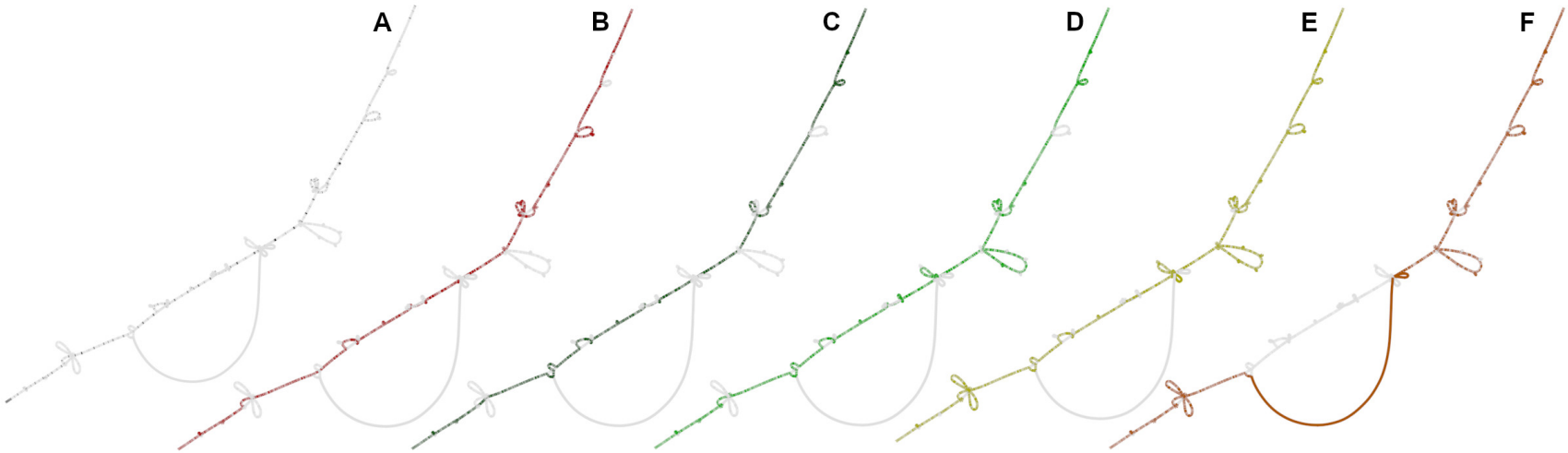
Comparison with other data

Left column: most divergent bins; right column: most conserved bins.



$w = 1\text{k}$

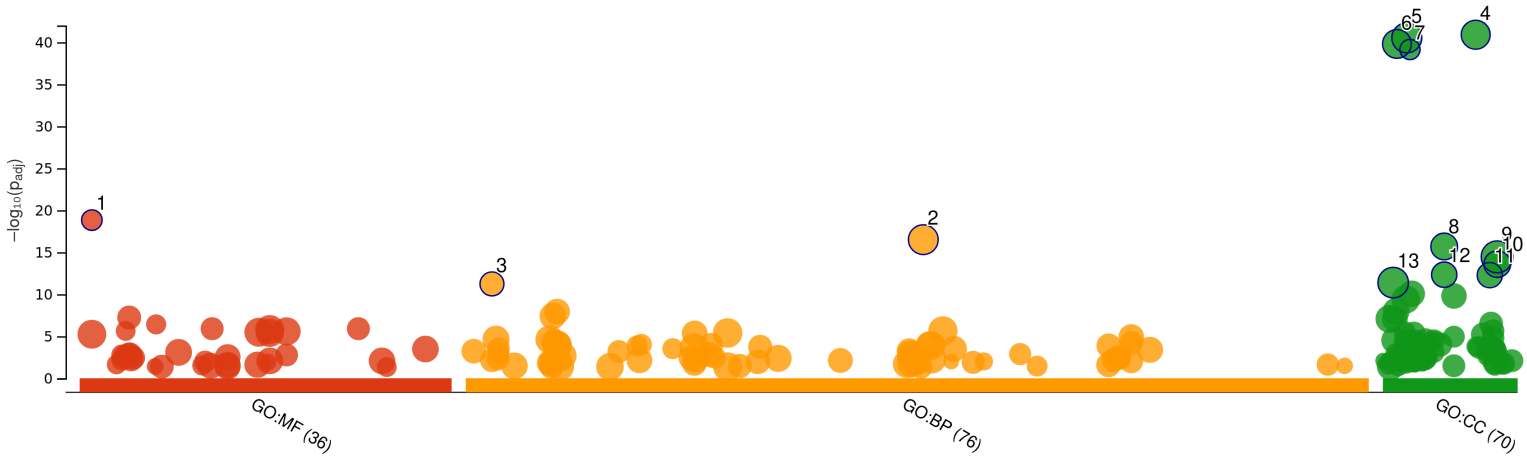
Example of divergent gene: *NBPF20*



A: exons; B: GRCh38; C: CHM13; D-F: haplotypes.

Bandage-NG

GO ontology of divergent genes



ID	Source	Term ID	Term Name	p_{adj} (query_1)
1	GO:MF	GO:0003823	antigen binding	1.491×10^{-19}
2	GO:BP	GO:0050896	response to stimulus	3.170×10^{-17}
3	GO:BP	GO:0002250	adaptive immune response	5.840×10^{-12}
4	GO:CC	GO:0071944	cell periphery	1.275×10^{-41}
5	GO:CC	GO:0016020	membrane	2.962×10^{-41}
6	GO:CC	GO:0005886	plasma membrane	1.564×10^{-40}
7	GO:CC	GO:0019814	immunoglobulin complex	7.383×10^{-40}
8	GO:CC	GO:0042995	cell projection	2.051×10^{-16}
9	GO:CC	GO:0110165	cellular anatomical entity	3.488×10^{-15}
10	GO:CC	GO:0120025	plasma membrane bounded cell projection	2.566×10^{-14}
11	GO:CC	GO:0098590	plasma membrane region	5.360×10^{-13}
12	GO:CC	GO:0043005	neuron projection	4.709×10^{-13}
13	GO:CC	GO:0005737	cytoplasm	4.147×10^{-12}

version	e111_eg58_p18_f463989d
date	8/21/2024, 2:55:15 PM
organism	hsapiens

g:Profiler

gProfiler

Conclusions

- PanSel, a tool for the quantification of conservation using sliding windows.
- What is next?
  - Do not use sliding windows anymore.
  - What to do when there is no anchor node?
  - Score is not very discriminative when bin size is low (many 1's).

Thank you for your attention!