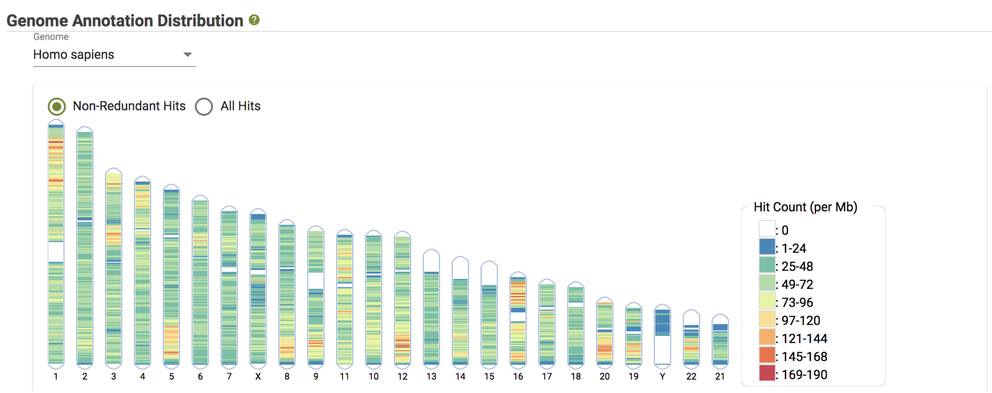
Developing Consensus Sequence Score Thresholds for Transposable Element Families

Eric Yeh, Robert Hubley, Arian Smit
Institute for Systems Biology, Seattle, WA, USA

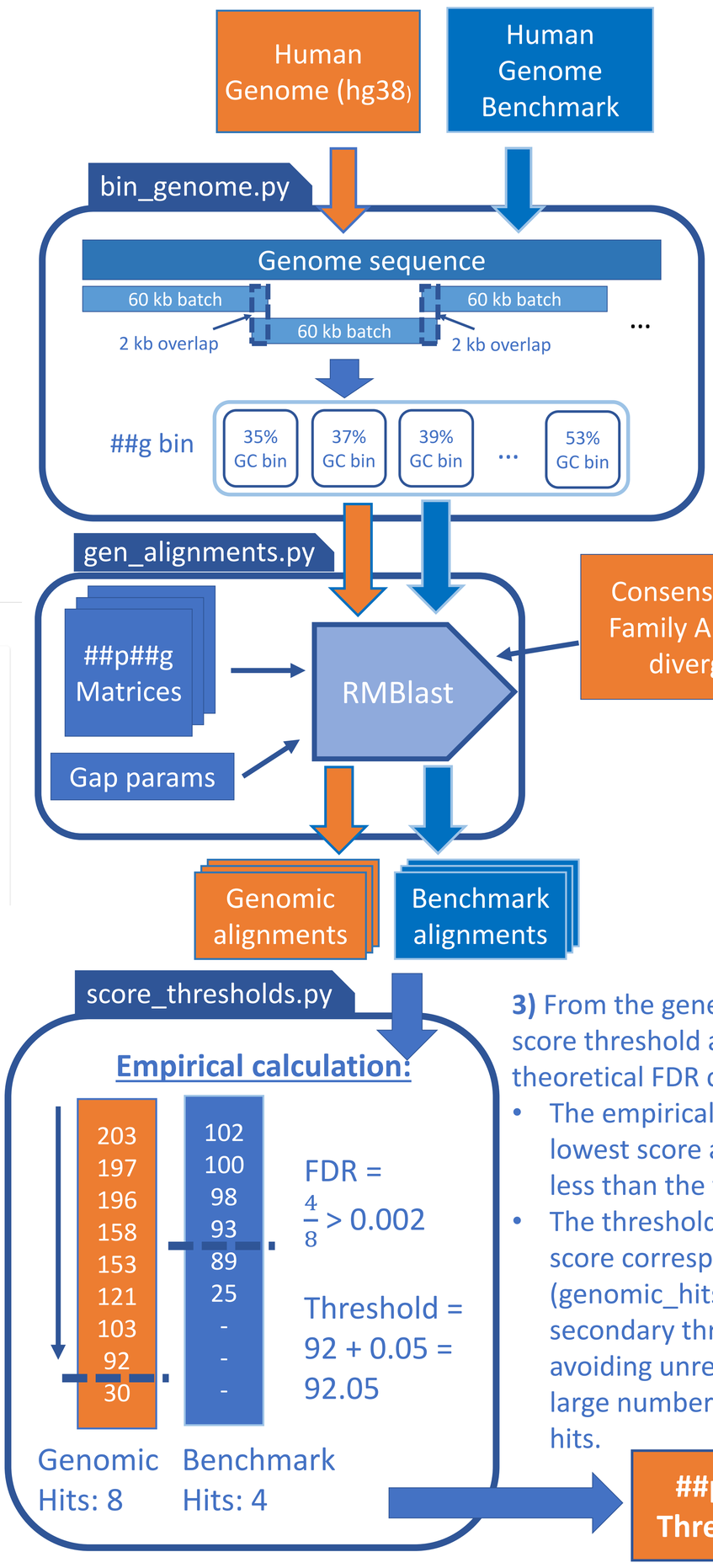
Background

- **Transposable elements (TEs):** DNA sequences that encode the ability to replicate themselves (autonomous) or rely on the products of other TE families to replicate (non-autonomous). Due to their minimal direct contribution to host function, TEs have been formerly been labeled as "selfish" or "junk" DNA, often having deleterious effects like inserting themselves into functional DNA segments. However, the success of TEs in many organisms (over 50% of the human genome is derived from TE copies) is an indication of the major role they have played in genome evolution, and their individual copies provide a rich dataset for many areas of research.
- **Dfam:** a database that stores TE family entries consisting of consensus sequences and profile hidden Markov models (HMMs) to classify and annotate families within genomes. The use of Dfam with profile HMMs identified an additional 2.9% of the human genome as being derived from TEs.



- **False Discovery Rate (FDR):** the proportion of sequence alignments (annotations) that are expected to be false positives (Type I errors). To account for mutations over time, we identify TE copies with a sequence similarity search algorithm, which aligns the consensus sequence of a TE family with a sequence of DNA and computes a similarity score. In the past, we have a used a small set of conservative fixed score thresholds to maintain a low overall FDR, but applying the same set of thresholds to large groups of TE families risks being overly conservative with some families and misses true matches. To maintain high sensitivity in our annotation, we should find a minimal score threshold(s) that can find more matches while keeping the FDR reasonably low, in our case below 0.2%. To estimate the alignment threshold for a fixed FDR, we align a family consensus sequence against a benchmark genome (a realistic sequence devoid of TE content, generate using GARLIC), and a reference genome. The threshold would then be the lowest reference genome alignment score which satisfies the target FDR.

Workflow



1) Split genome and benchmark sequences into 60kb batches, group batches into 10 bins based on the batch's GC background

$$\text{GC background} = \frac{G + C}{G + C + A + T}$$

2) Run the consensus sequence against each bin in RMBlast, using the appropriate matrix and gap parameters

3) From the generated alignments of RMBlast, compute the score threshold as the more conservative of the empirical and theoretical FDR calculations

- The empirical FDR threshold is determined by finding the lowest score at which the benchmark_hits/genomic_hits is less than the target FDR of 0.2%.
- The threshold was also required to be at least as high as the score corresponding to an E-value of min(1000, (genomic_hits – benchmark_hits) * FDR_threshold). This secondary threshold was chosen as a simple method of avoiding unreasonably liberal scores for families with a large number of genomic hits and few, if any benchmark hits.

Objective

The goal of this project is to create a pipeline that can compute sensitive but conservative score thresholds for each TE family consensus sequence per GC-tuned scoring matrix. By selecting score thresholds maintaining a false discovery rate below 0.2%, we can develop a robust dataset that can be used for Dfam and other projects.

From the resulting thresholds we can answer the following questions:

- Is it sufficient to use one set of thresholds for all taxa in which the model is applicable?
- How significant are the differences between isochores?
- What is the expected improvement to annotation if we switch to using per-family thresholds?

Next Steps

- Run several consensus sequences through this workflow to develop database of score thresholds
- Analyze calculated score thresholds to see if we can draw any conclusions
- Find ways to optimize and automate code for running several consensus sequences in a cluster environment
- Use alignments as test data for adjudication project

References

[1] Bonchev, G.N. Useful parasites: the evolutionary biology and biotechnology applications of transposable elements. *J Genet* 95, 1039-1052 (2016). https://doi.org/10.1007/s12041-016-0702-6

[2] Wheeler, Travis J., et al. "Dfam: a database of repetitive DNA based on profile hidden Markov models." *Nucleic acids research* 41.D1 (2012): D70-D82.

[3] Hubley, Robert, et al. "The Dfam database of repetitive DNA families." *Nucleic acids research* 44.D1 (2016): D81-D89.

[4] Smit, AFA, Hubley, R & Green, P. *RepeatMasker Open-4.0.* 2013-2015 <http://www.repeatmasker.org>.