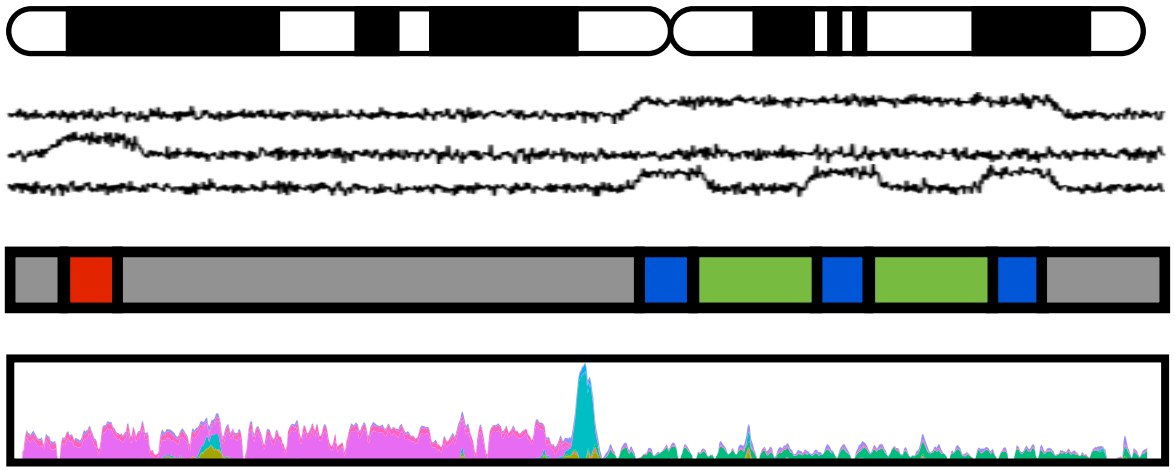
A unified encyclopedia of human functional elements derived from fully automated annotation of 164 human cell types




Maxwell W. Libbrecht

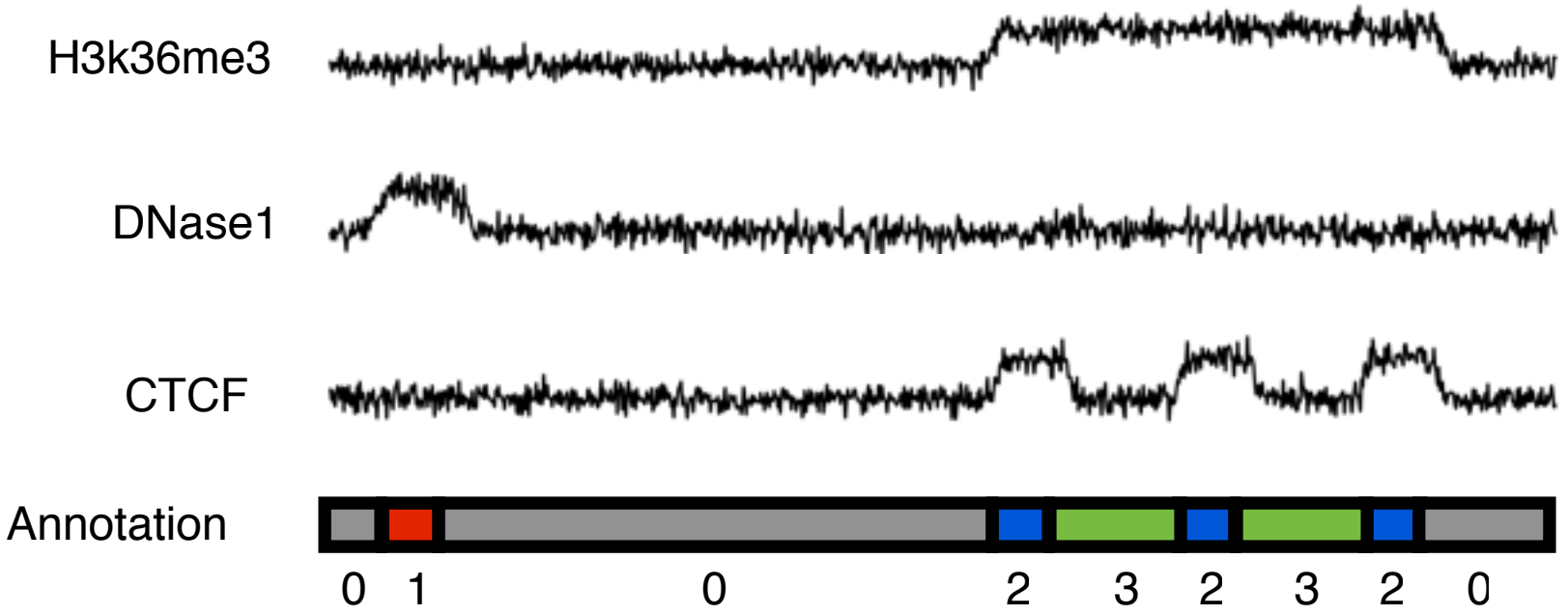

Oscar Rodriguez


Michael M. Hoffman


Jeffrey A. Bilmes


William S. Noble

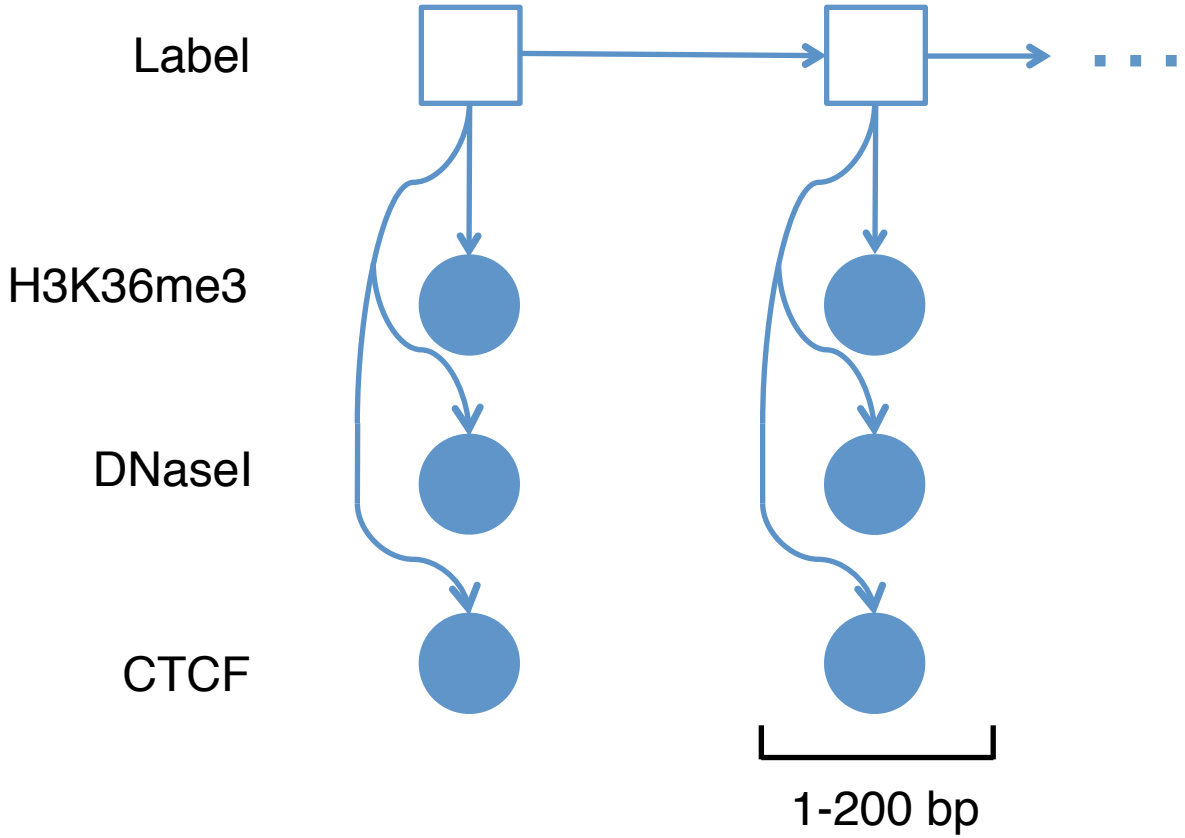
Semi-automated genome annotation algorithms partition and label the genome on the basis of functional genomics tracks



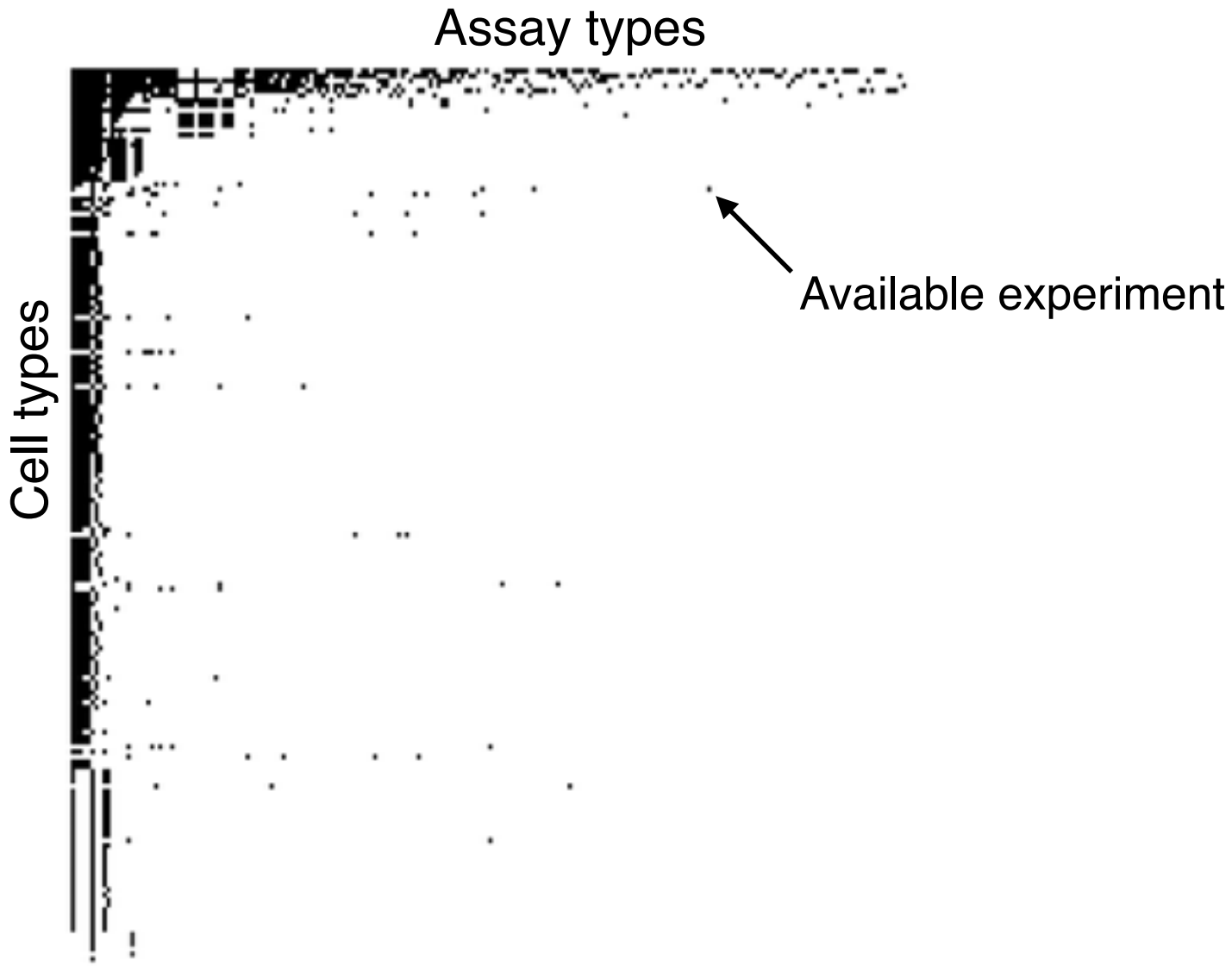
Human interpretation: 1 = “Enhancer”, 2 = “Exon”, …

HMMSeg: Day et al. *Bioinformatics*, 2007
ChromHMM: Ernst, J. and Kellis, M. *Nature Biotechnology*, 2010
Segway: Hoffman, M et al. *Nature Methods*, 2012

Semi-automated genome annotation algorithms use dynamic Bayesian network models

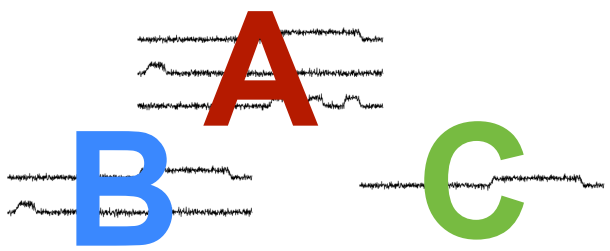


Goal: Annotate all cell types with sufficient data

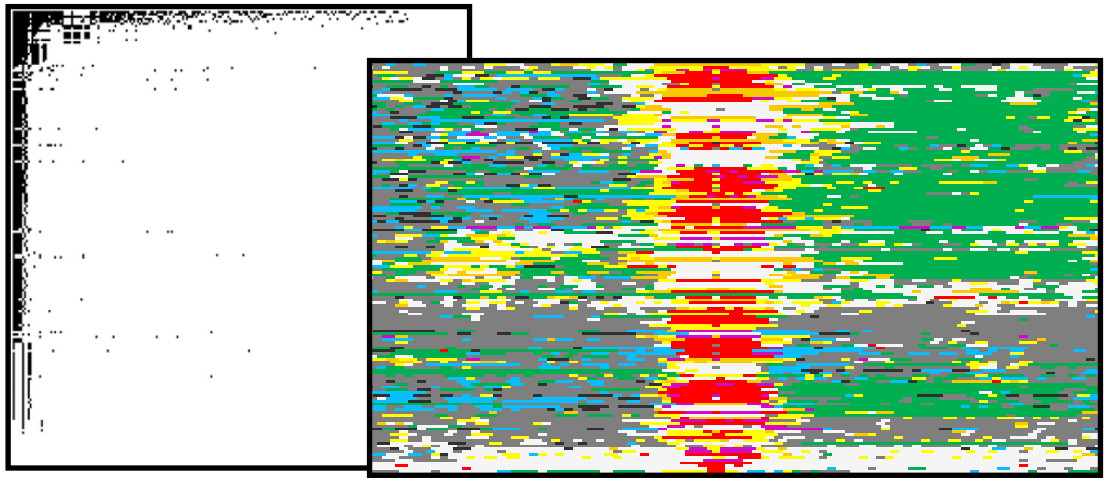


Strategies for multi-cell-type annotation

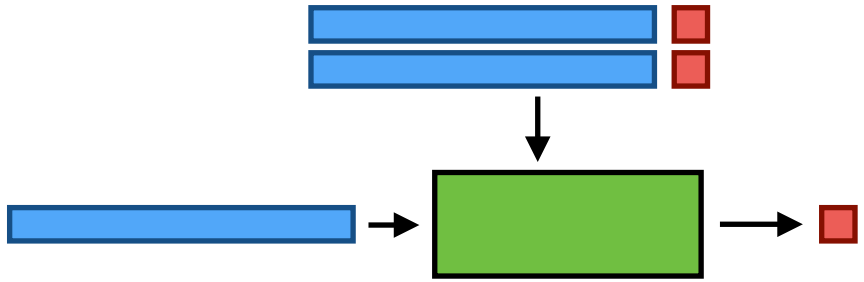
1. Concatenated
   - All cell types must have exactly the same data types.
   - Very sensitive to experimental artifacts between cell types.
2. Impute+concatenated
   - Use of data from other cell types makes interpretation difficult.
3. Separate
   - Requires interpreting labels separately for each cell type.
   - Idea: Use classifier to automate interpretation.



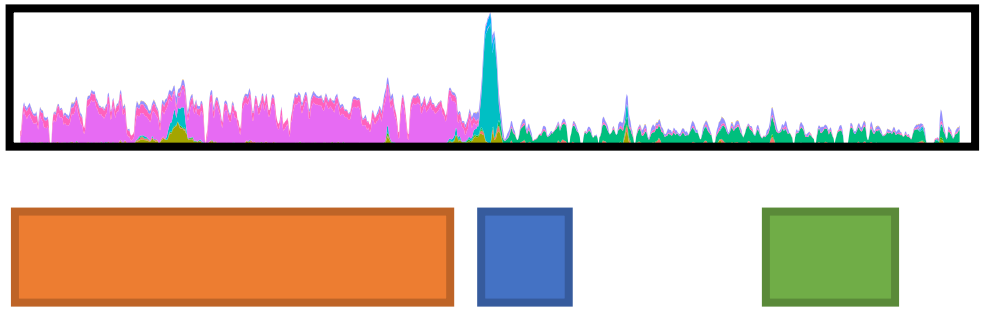
Annotation pipeline



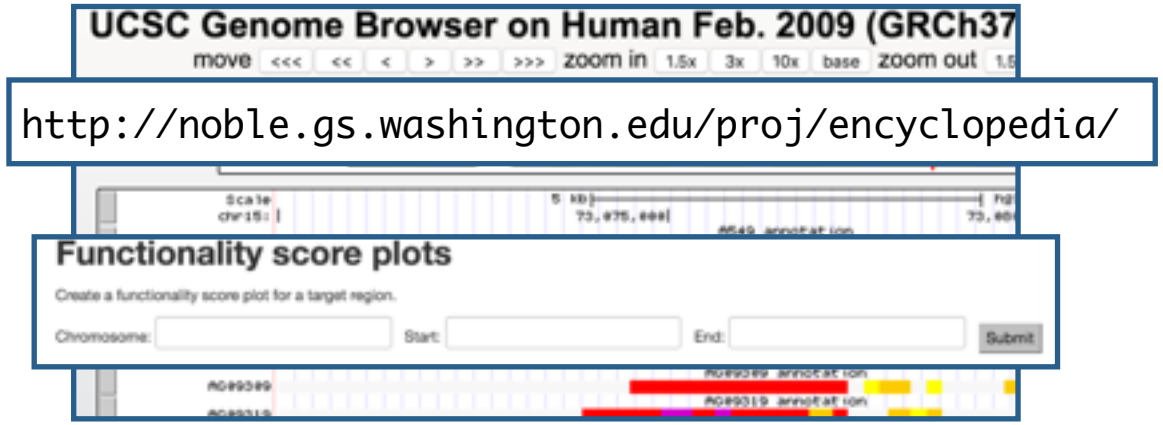
A machine learning classifier recapitulates human interpretation



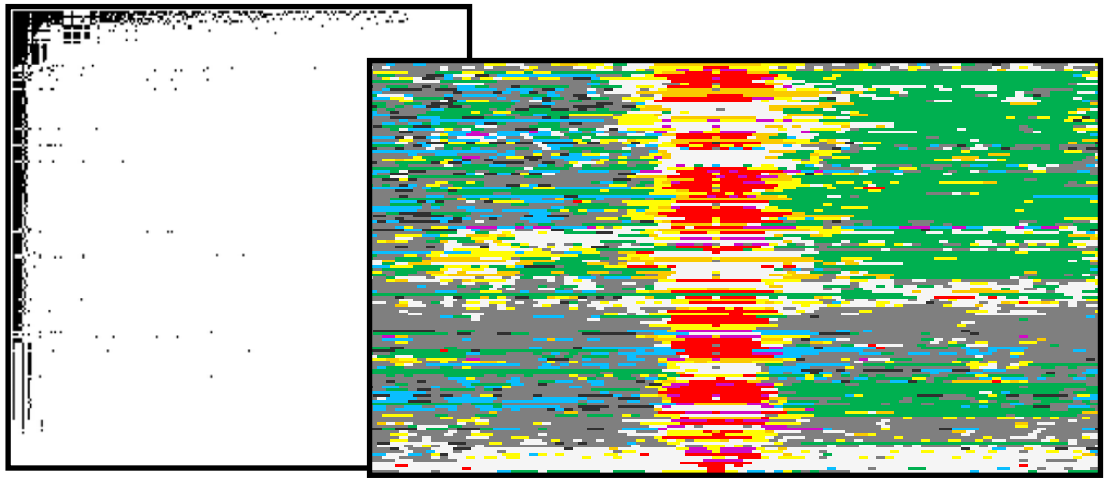
A unified Segway encyclopedia



Accessing the annotations



Annotation pipeline

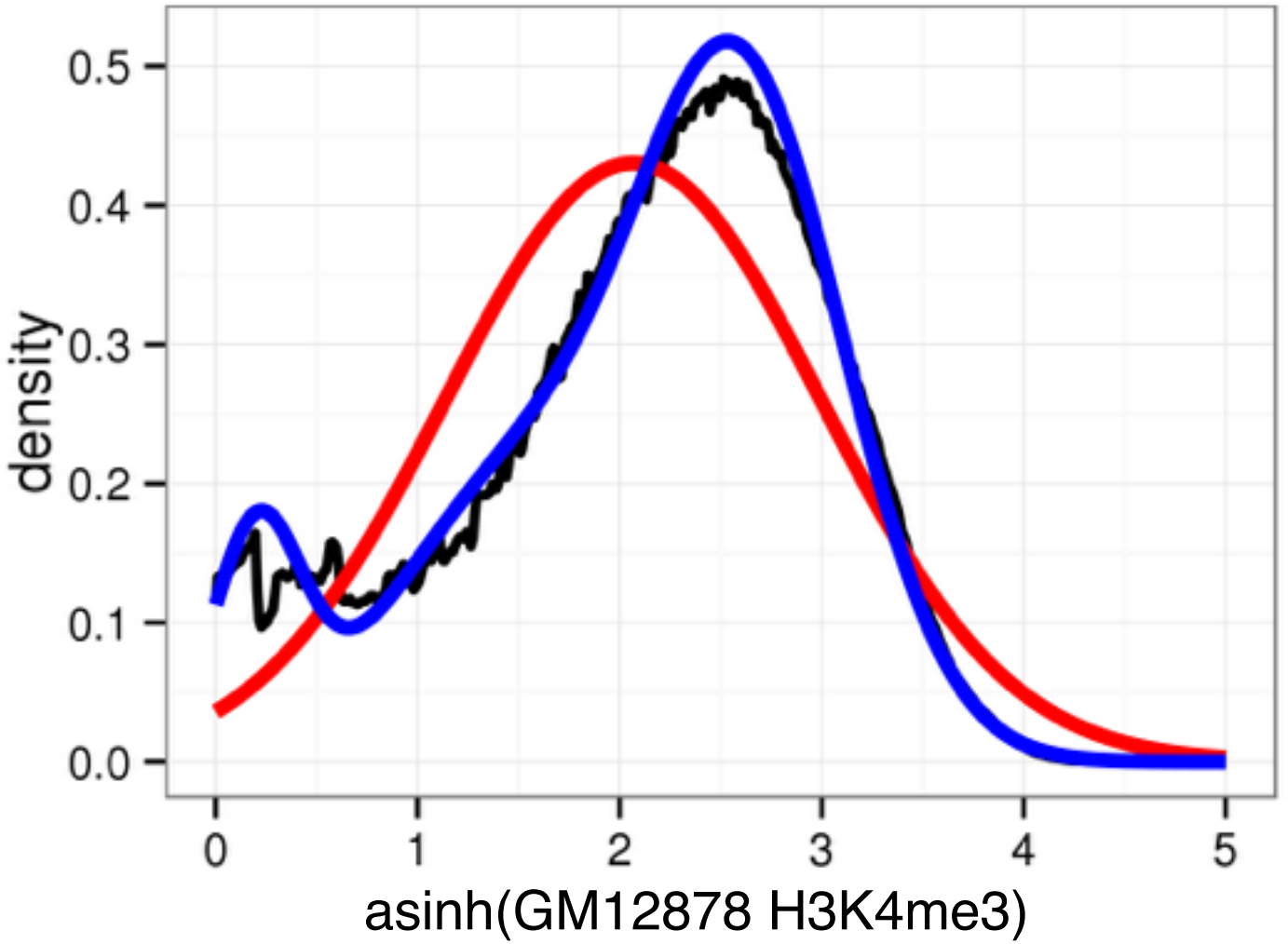


A machine learning classifier recapitulates human interpretation

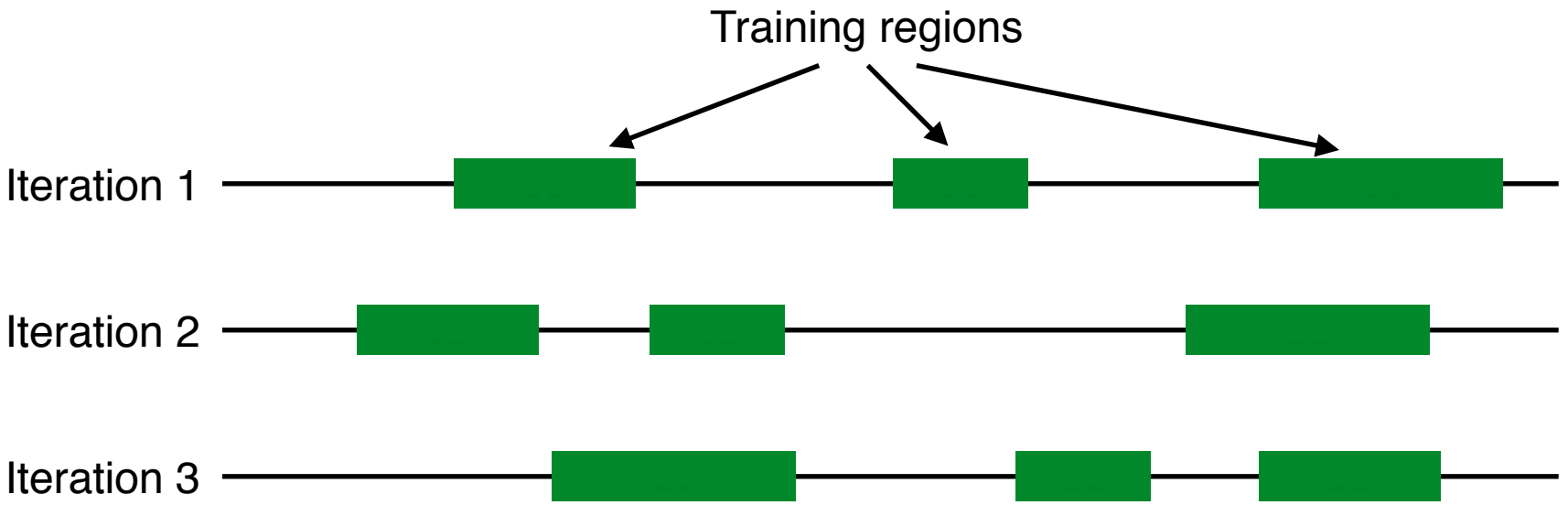
A unified Segway encyclopedia

Accessing the annotations

Three-component mixtures of Gaussians capture biological phenomena better than single-component mixtures

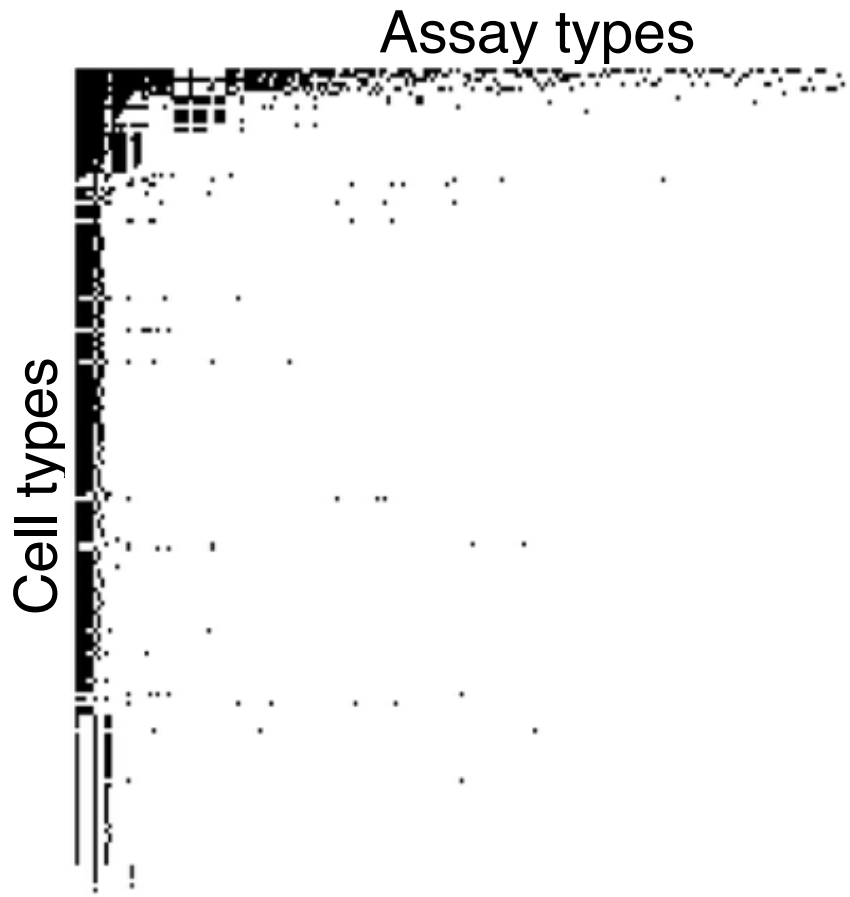


Minibatch training allows efficient optimization while using all available data



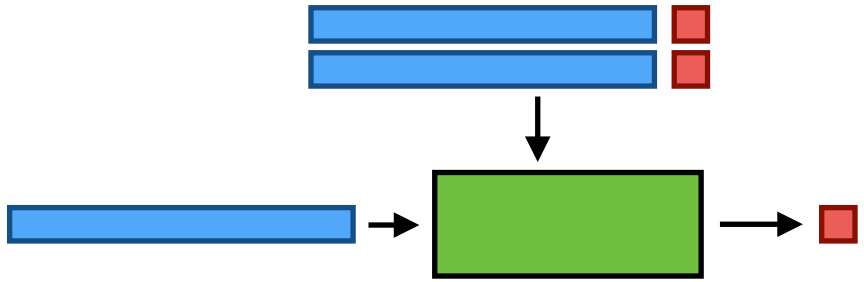
Annotation parameters

- Criterion for annotating a cell type:
  - EITHER: At least one experiment each in two of the following categories: (1) Histone ChIP, (2) TF ChIP, (3) DNase
  - OR: At least six histone ChIP experiments
  - Results in 164 cell types
- asinh-transformed signal
- 100 bp resolution
- Number of labels = 10 + 2*sqrt(number of tracks)
- Emission distributions: mixtures of three Gaussians
- Minibatch training with 1% of the genome in each batch



Annotation pipeline

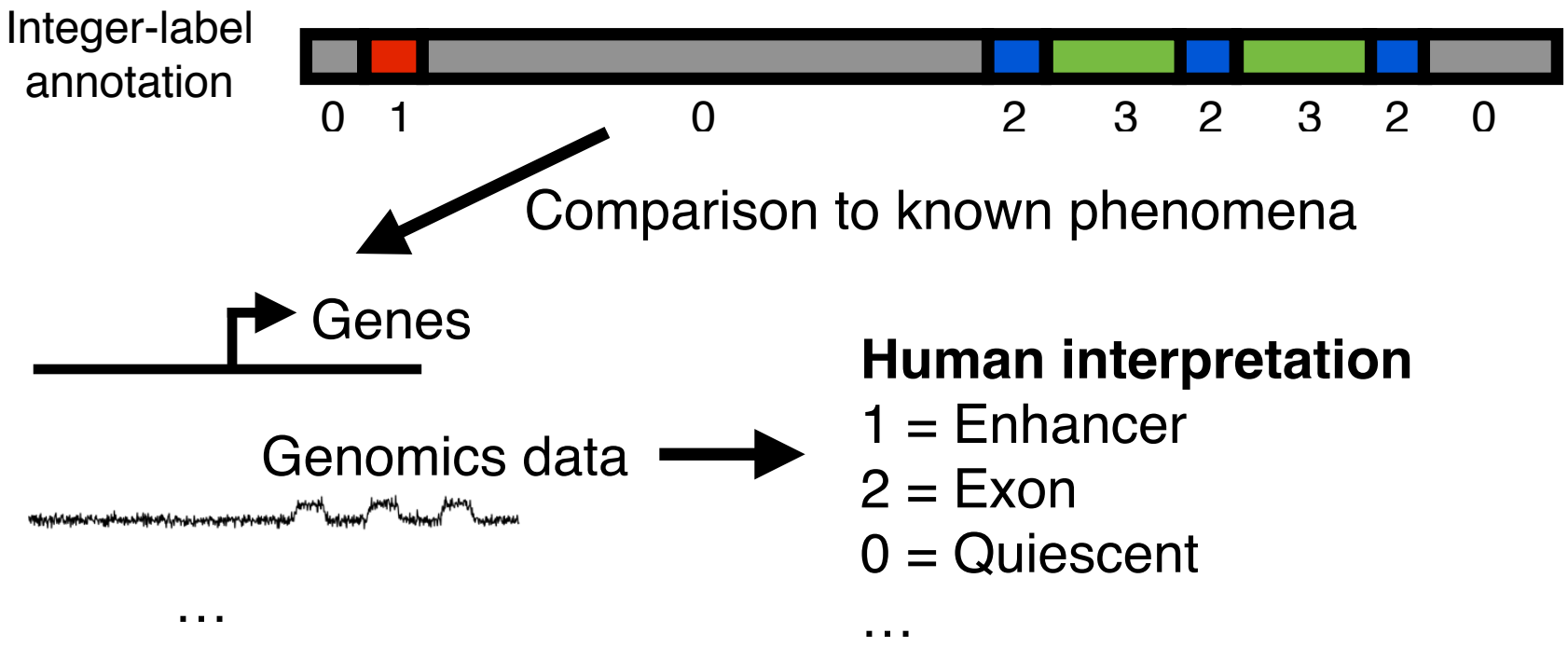
A machine learning classifier recapitulates human interpretation



A unified Segway encyclopedia

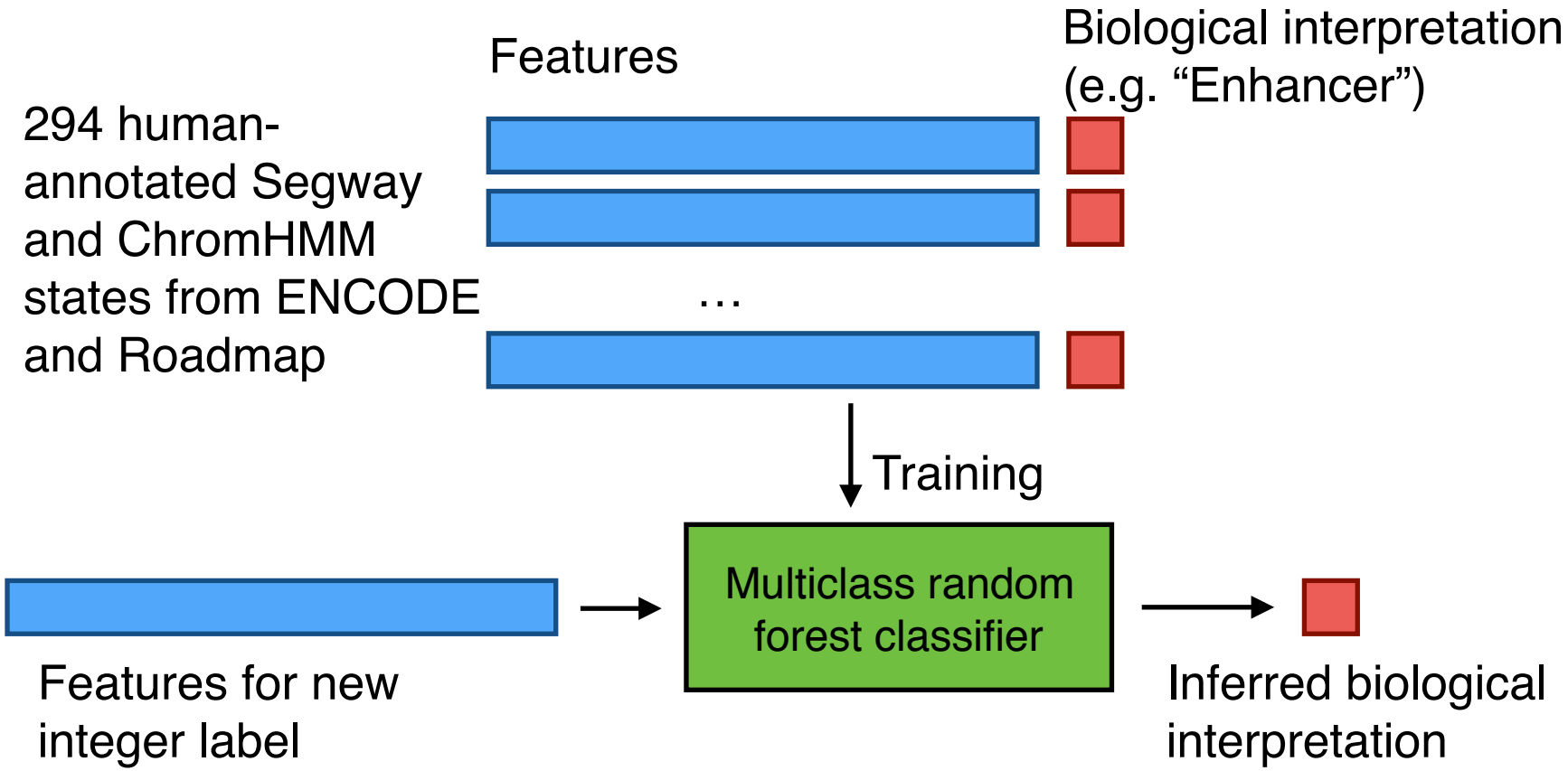
Accessing the annotations

How are semi-automated genome annotations usually interpreted?

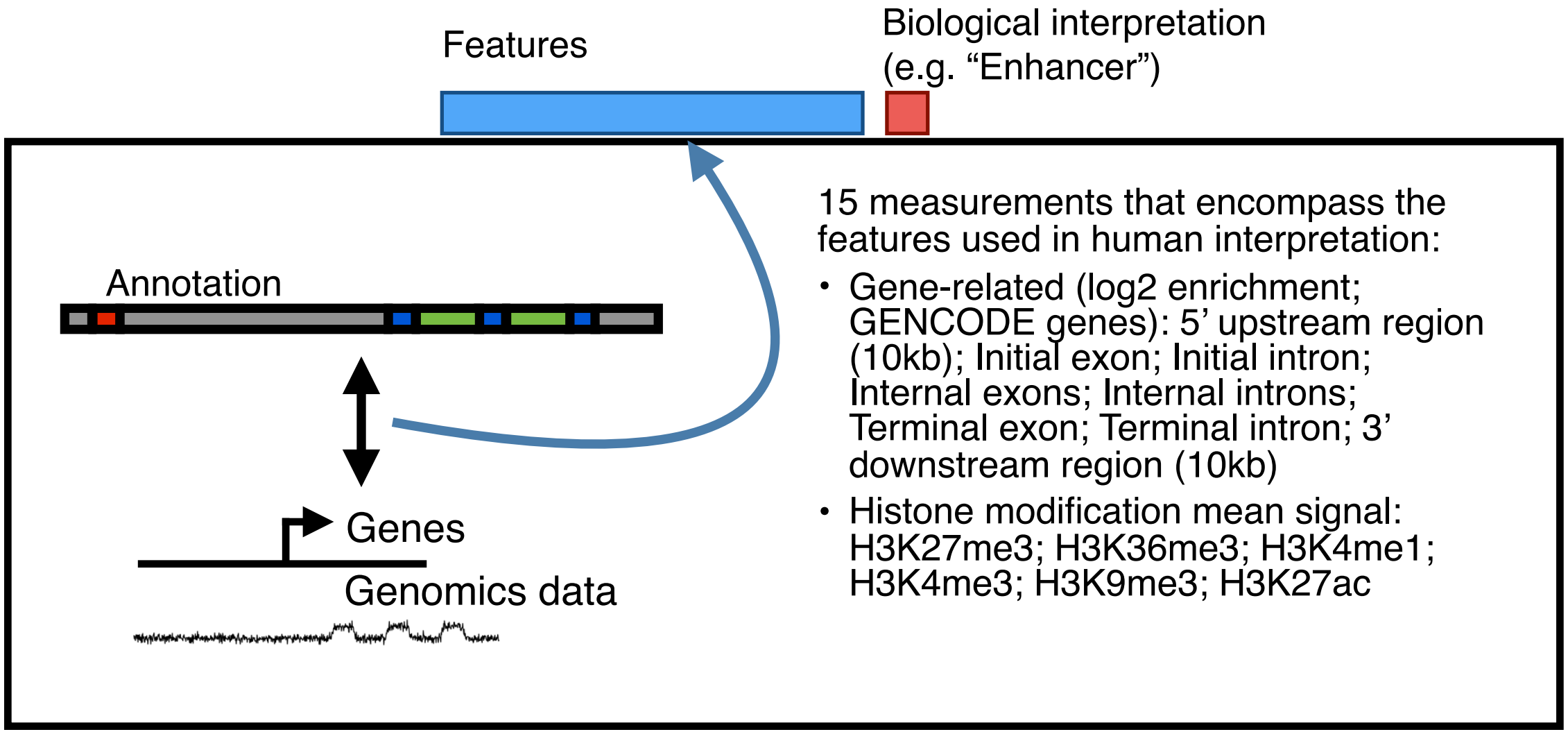


Idea: Use machine learning classifier to reproduce human interpretation.

A classifier-based approach recapitulates human interpretation



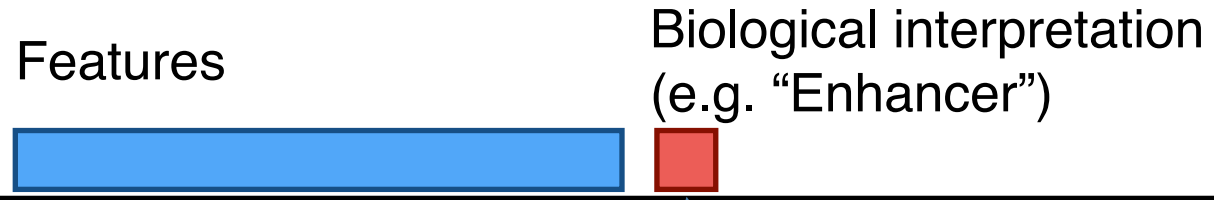
A classifier-based approach recapitulates human interpretation



15 measurements that encompass the features used in human interpretation:

- Gene-related (log2 enrichment; GENCODE genes): 5’ upstream region (10kb); Initial exon; Initial intron; Internal exons; Internal introns; Terminal exon; Terminal intron; 3’ downstream region (10kb)
- Histone modification mean signal: H3K27me3; H3K36me3; H3K4me1; H3K4me3; H3K9me3; H3K27ac

A classifier-based approach recapitulates human interpretation

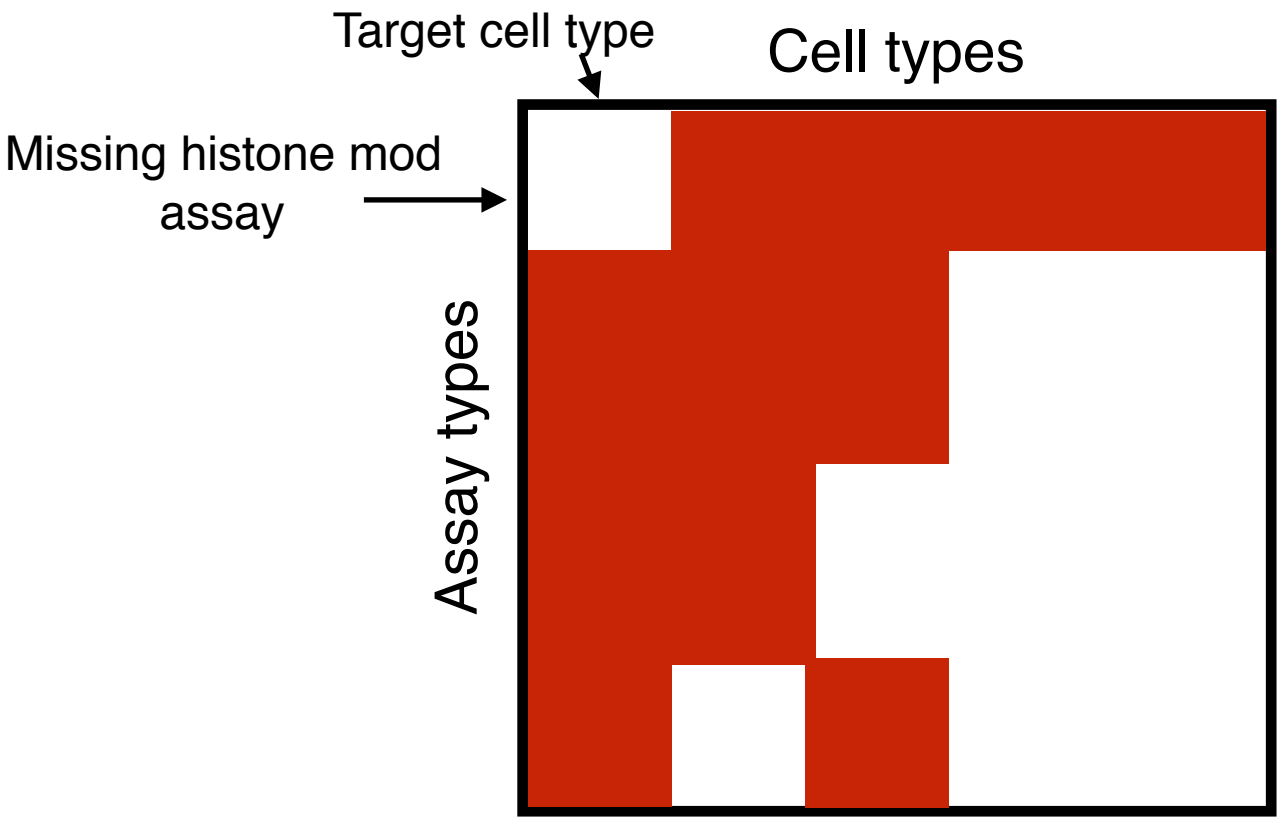


One of 8 biological categories:

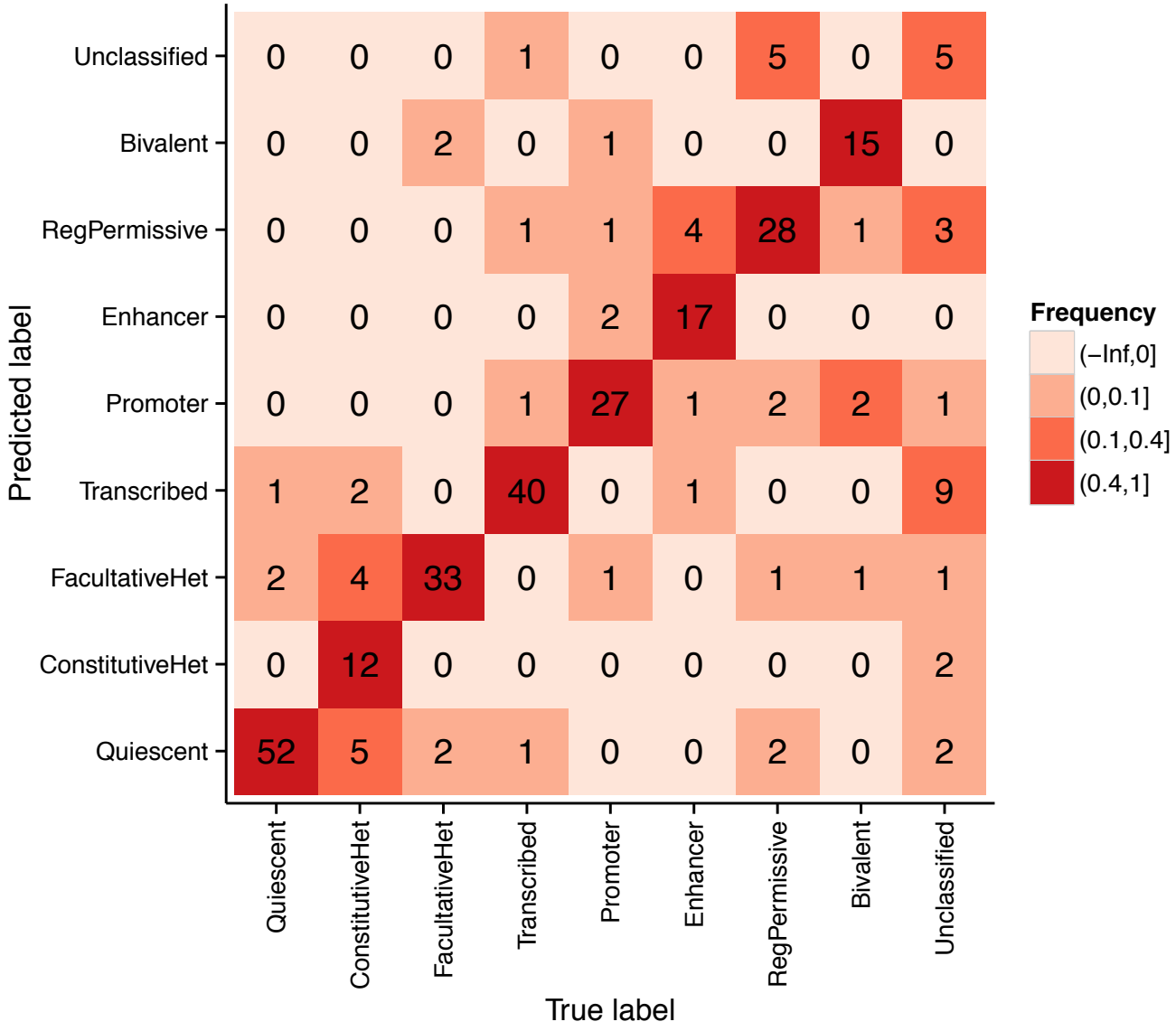
- **Promoter**
- **Enhancer**
- **RegPermissive**
- **Bivalent**
- **Transcribed**
- **ConstitutiveHet** (Heterochromatin that marks permanently silent regions)
- **FacultativeHet** (Cell type-specific heterochromatin)
- **Quiescent**
- **Unclassified** (Does not fit into above categories)

We substitute tracks from related cell types to fill in missing histone mod features

- We substitute the a track from the most similar cell type with that histone mod present.
- Similarity: Defined as average correlation over all shared tracks.

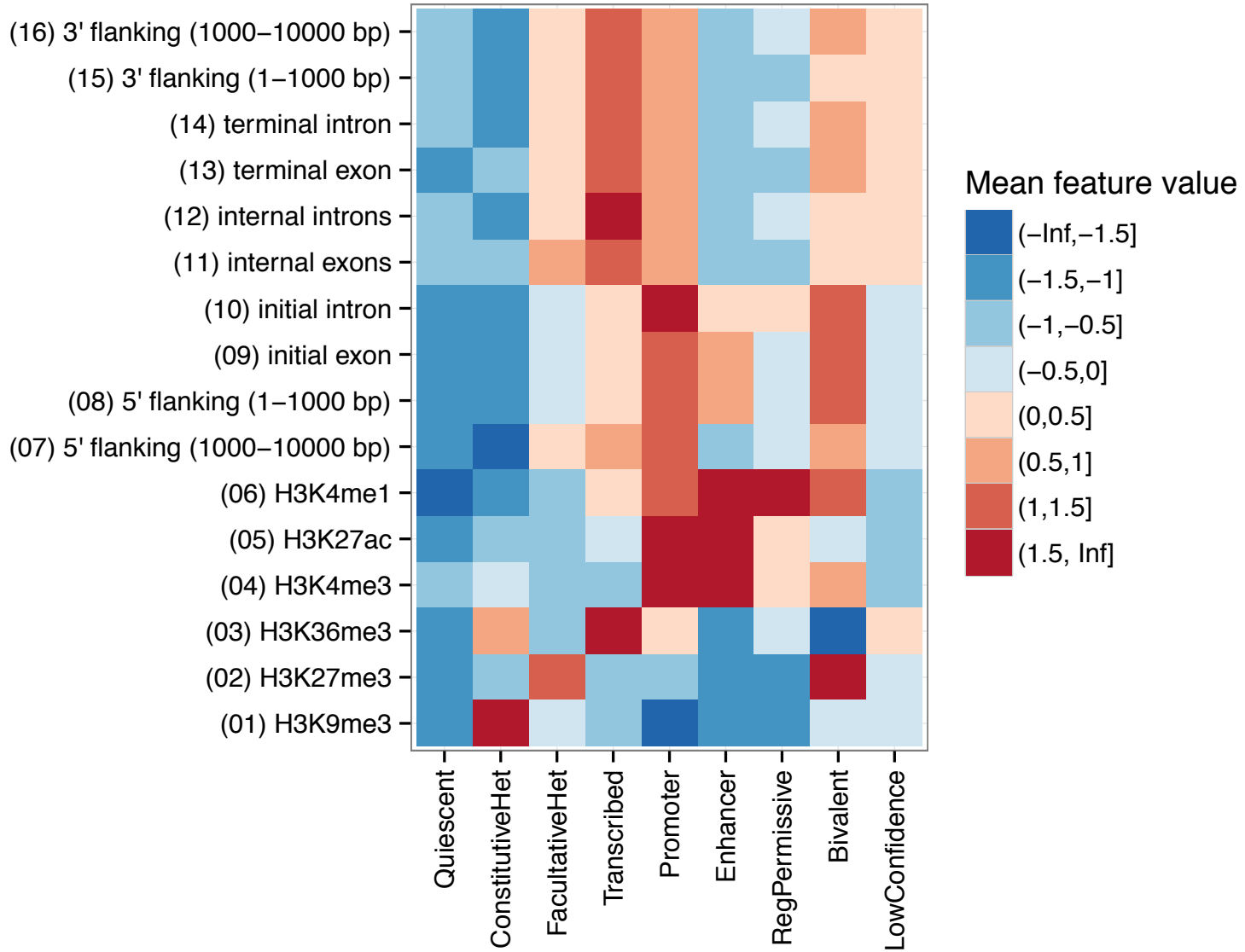


Classifier accurately recapitulates human interpretations

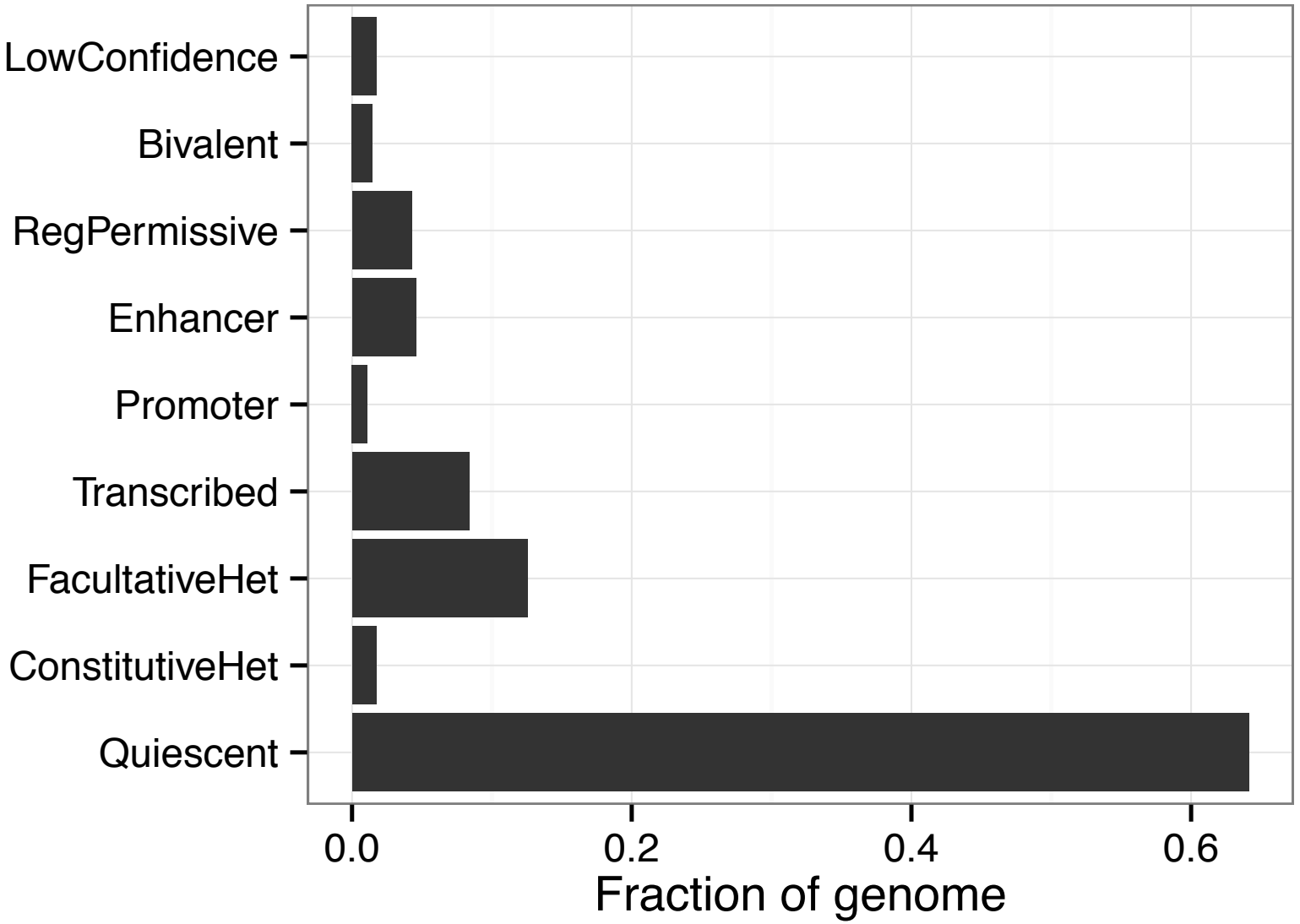


Accuracy = 229/294 = 78%

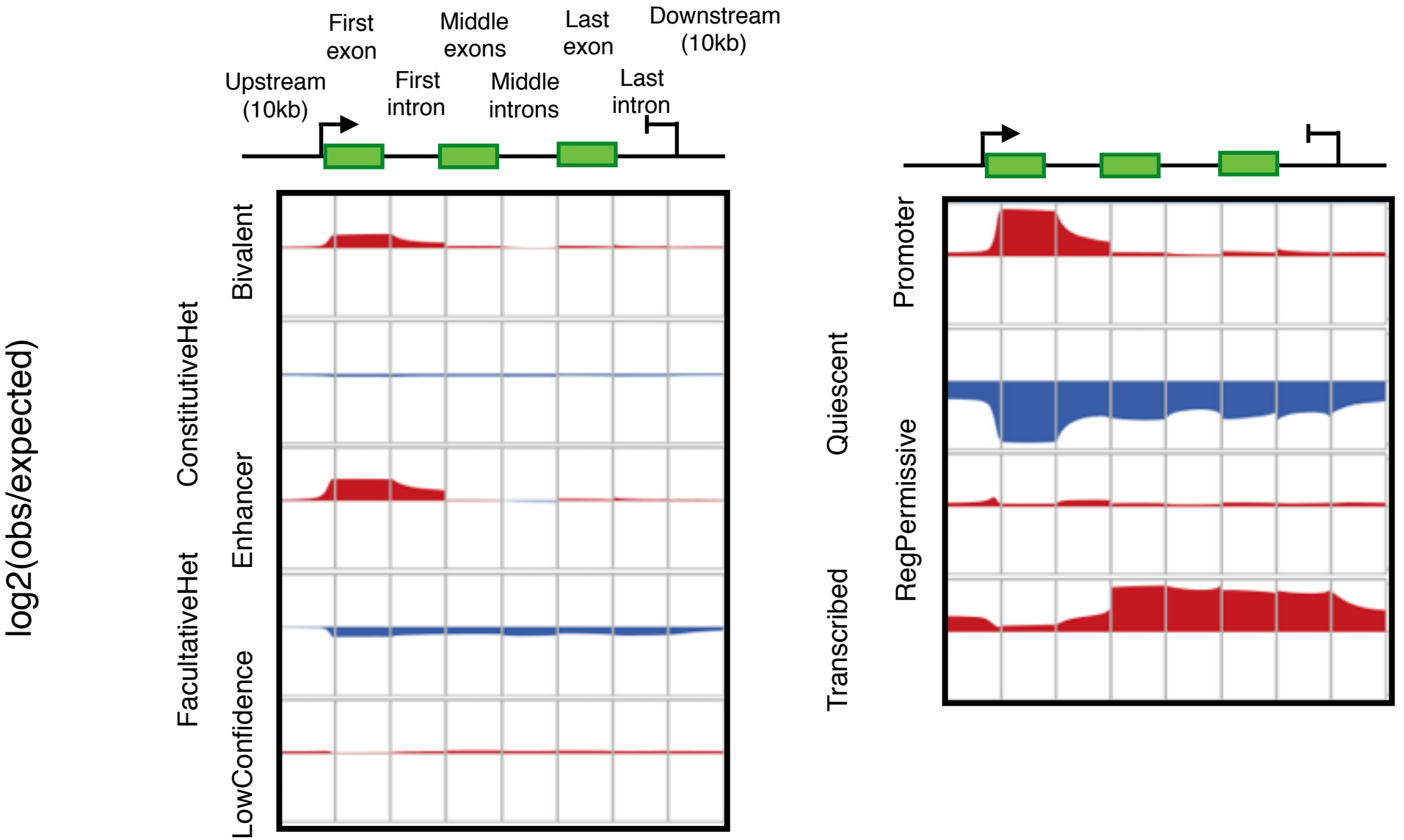
The classifier's inferences are based on the expected features



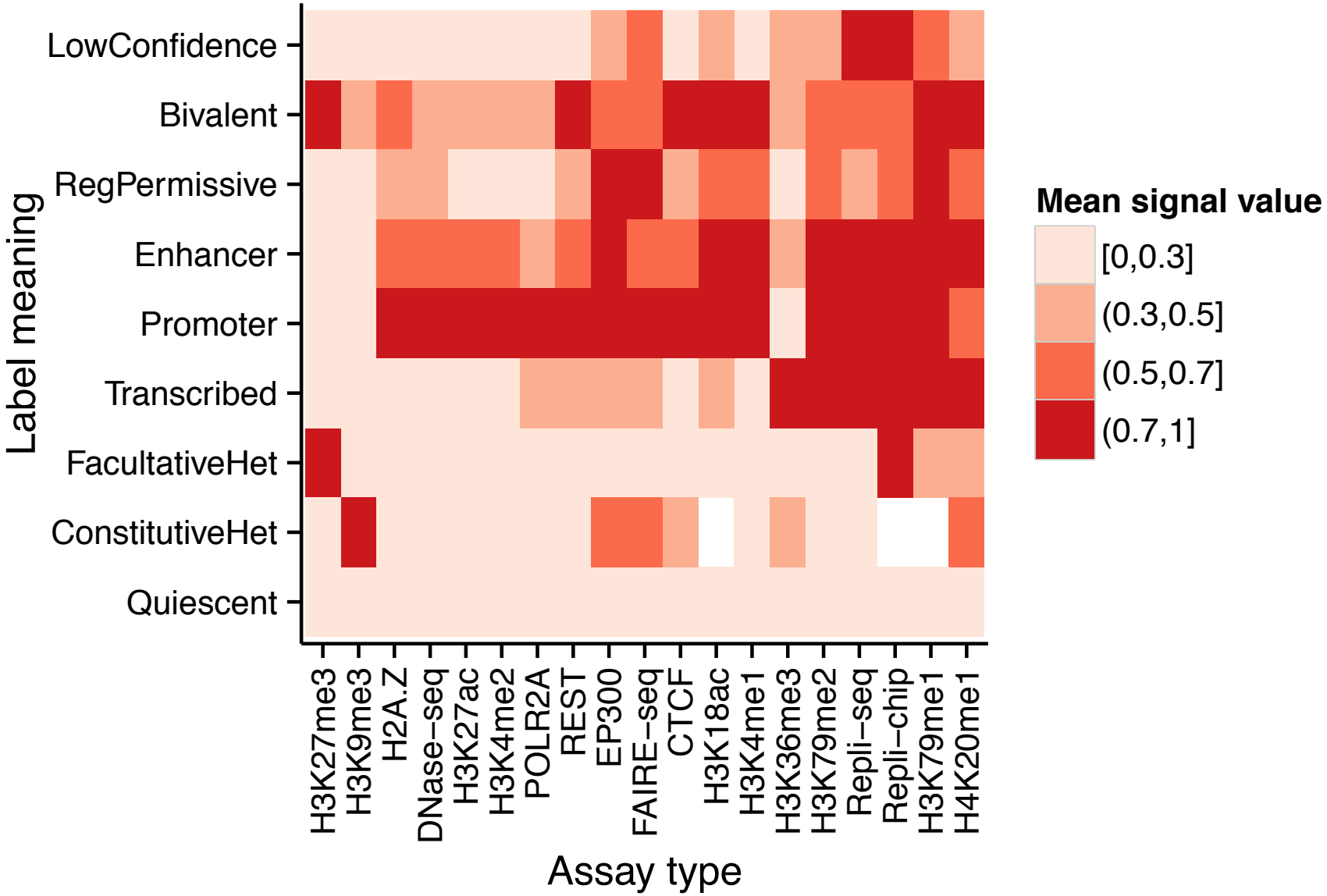
Labels have the expected genomic distribution



Annotations have the expected gene relationships



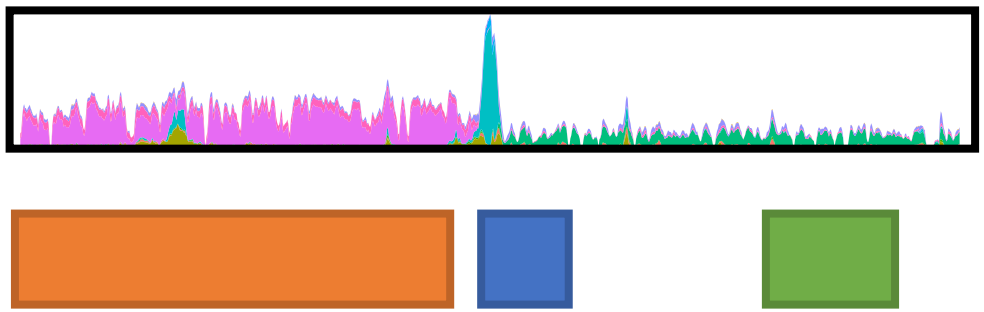
Annotations have the expected relationship to input tracks



Annotation pipeline

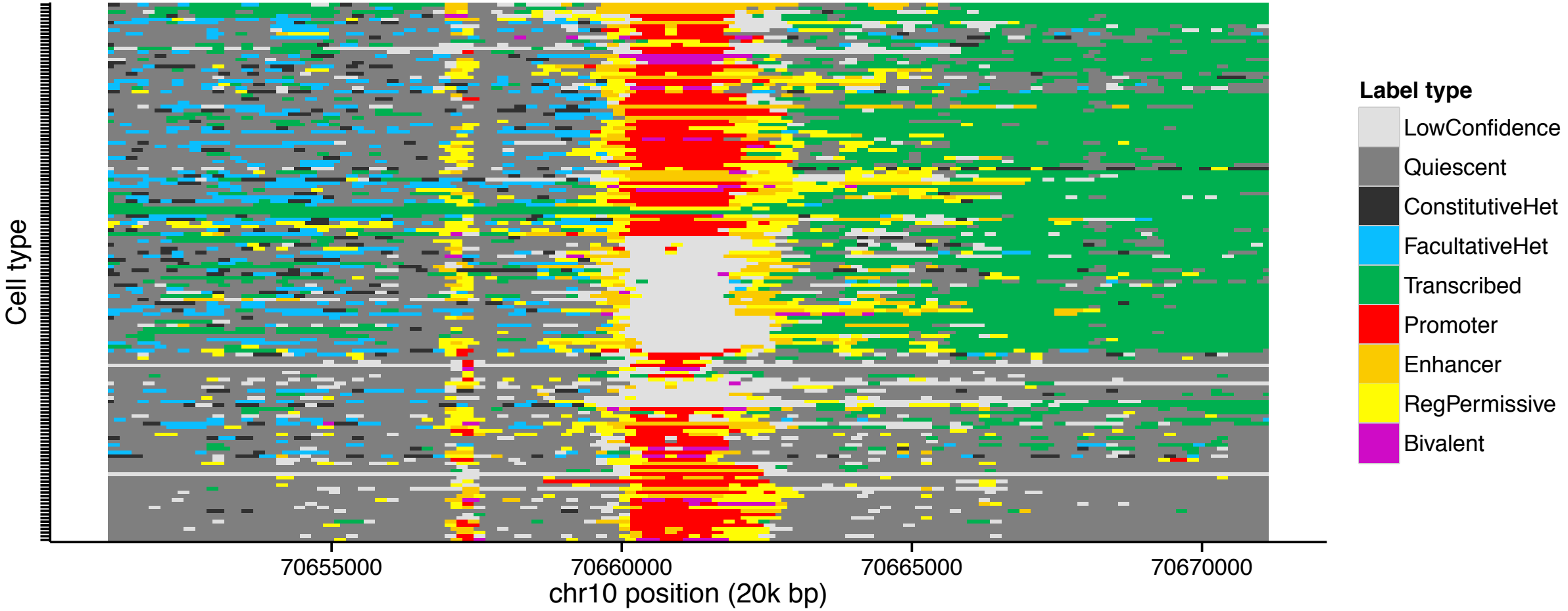
A machine learning classifier recapitulates human interpretation

A unified Segway encyclopedia



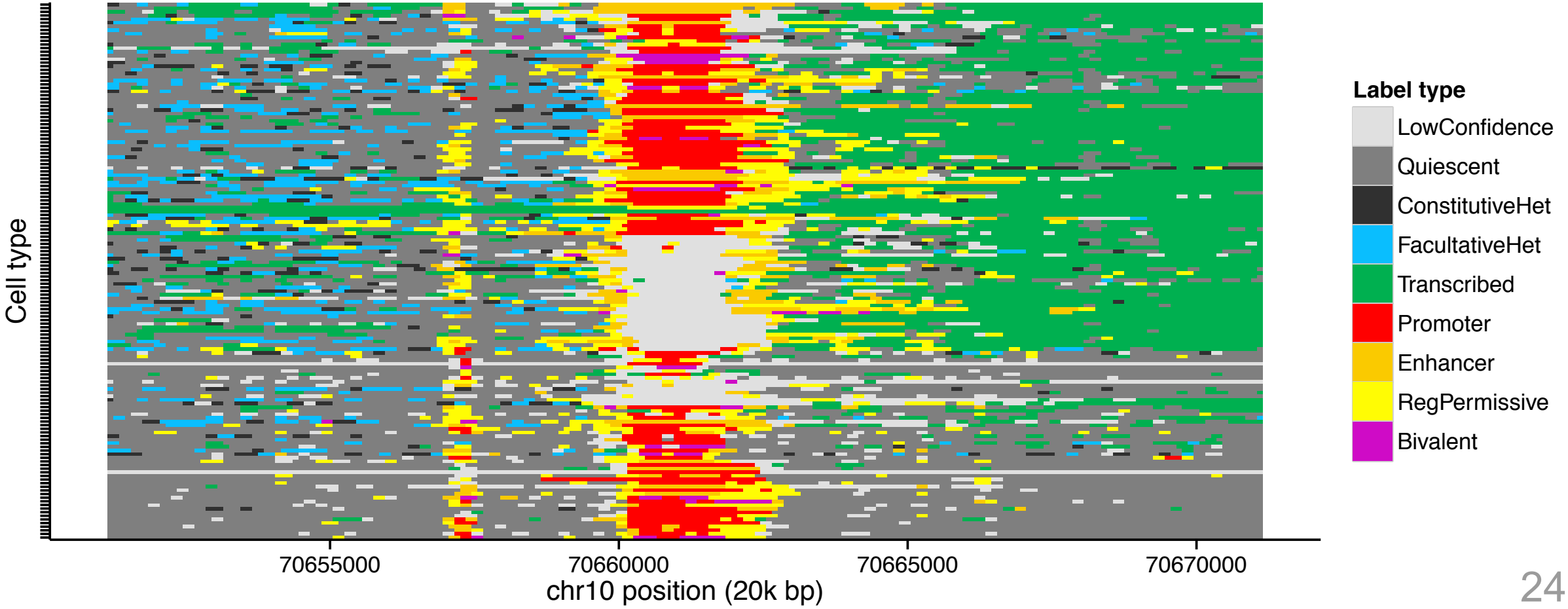
Accessing the annotations

How can we make annotations of hundreds of cell types more understandable?

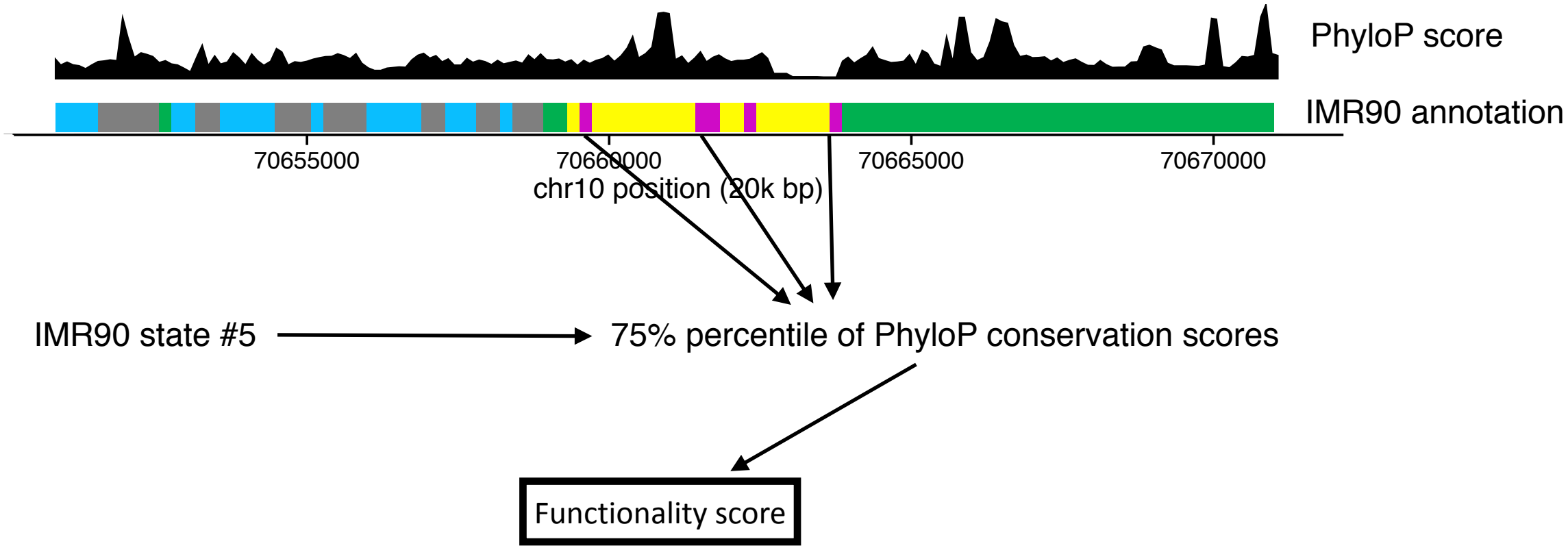


How can we make annotations of hundreds of cell types more understandable?

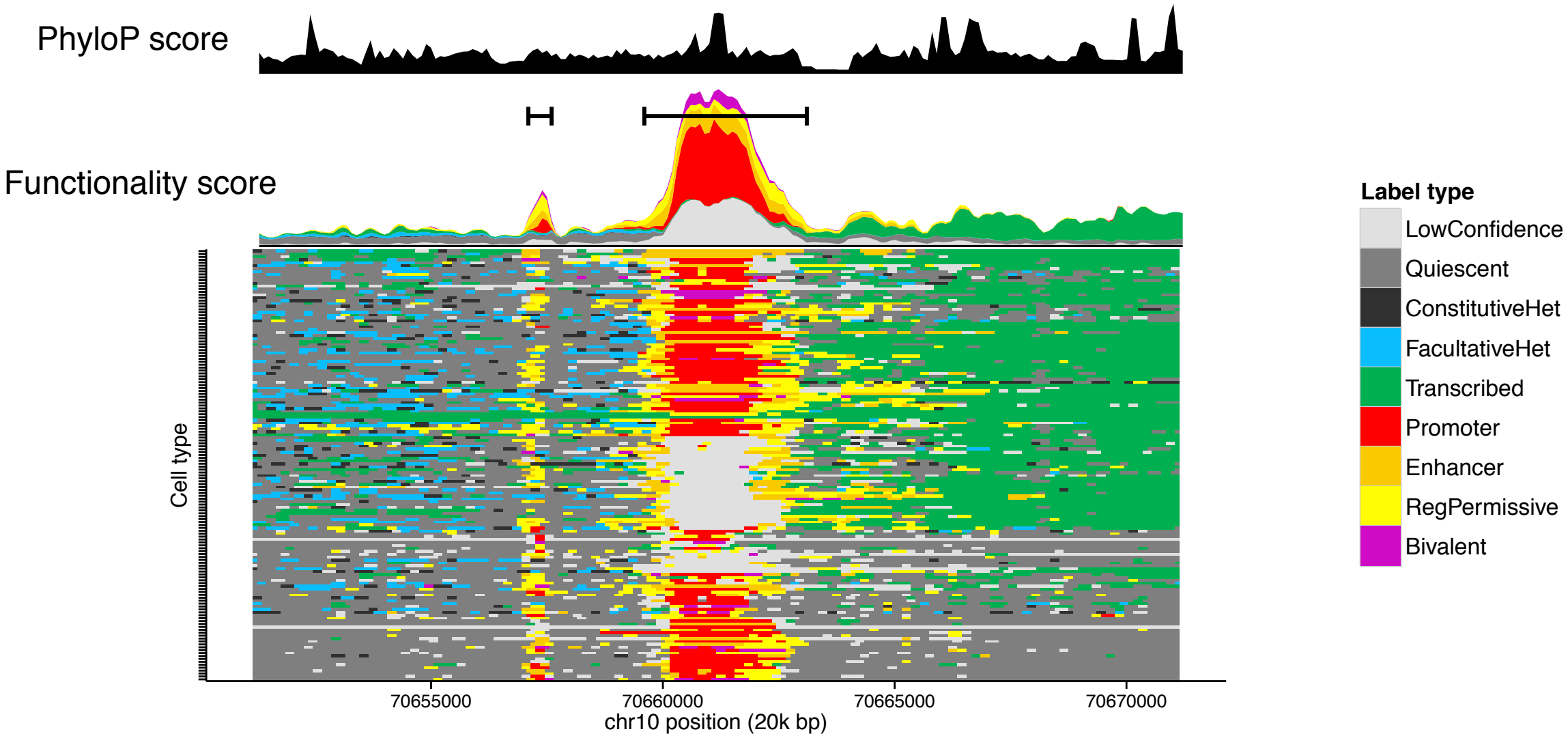
Locus of interest:
e.g. disease variant, human-accelerated region



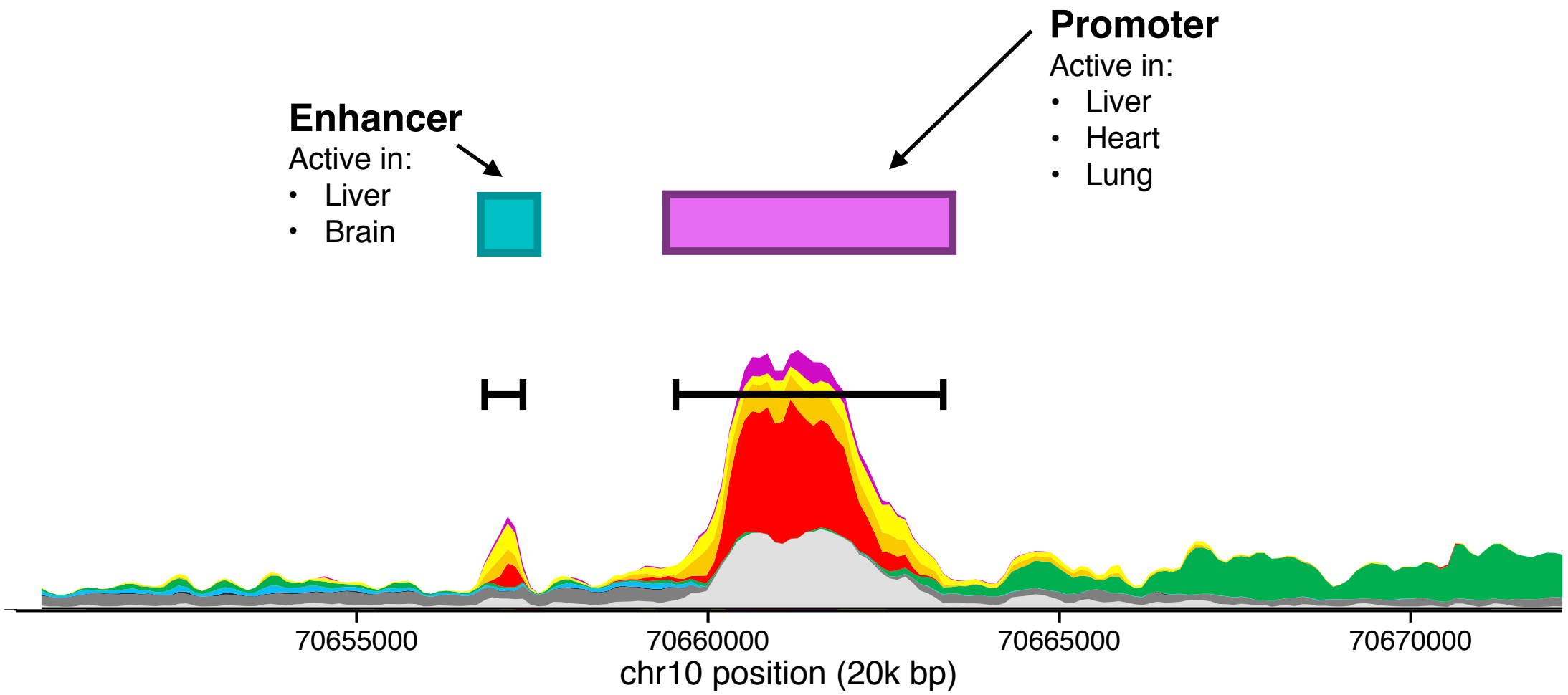
Annotation labels that are enriched for evolutionary conservation mark functional activity



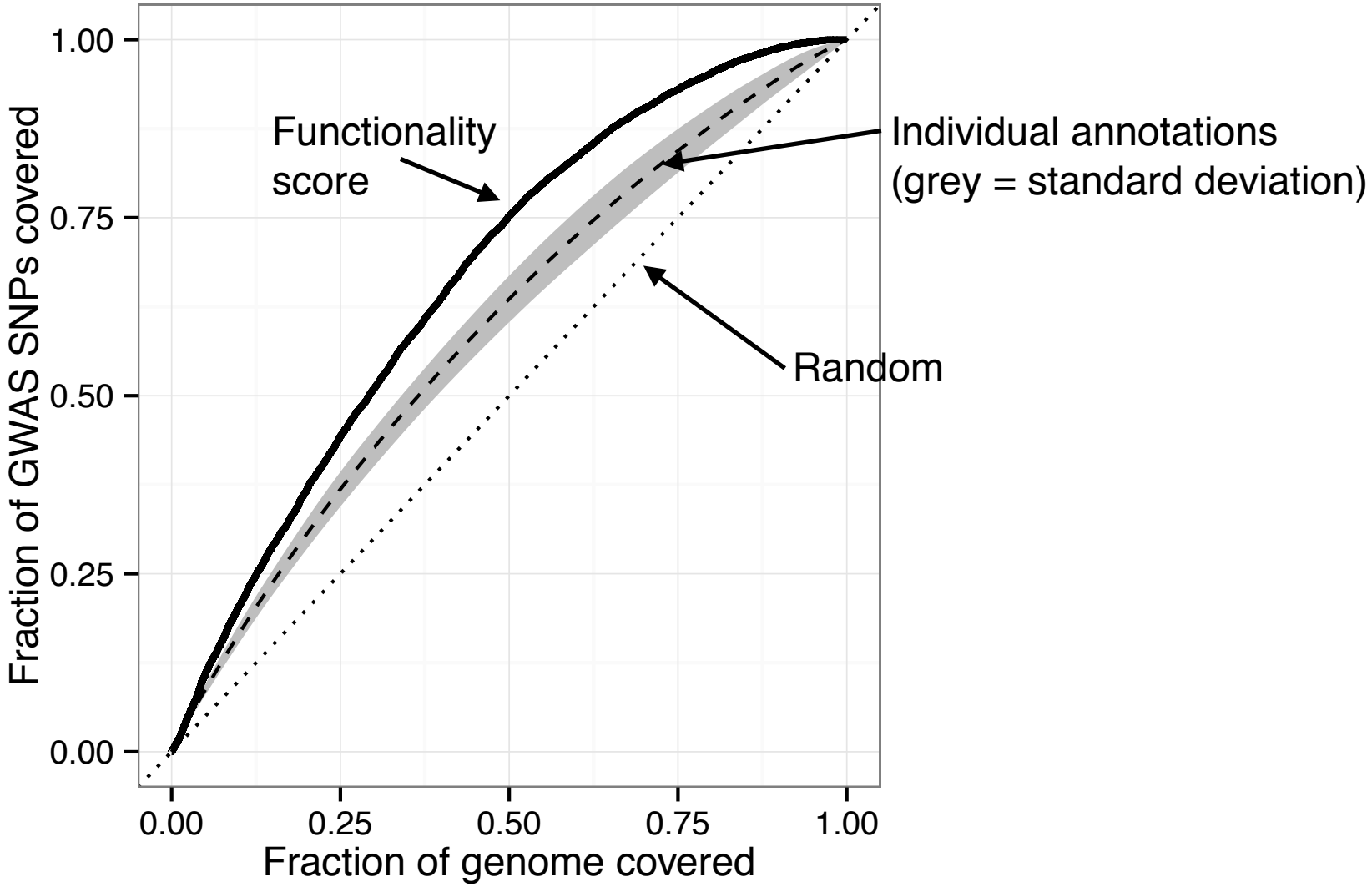
Functionality score view enables easy understanding of activity



A unified Segway encyclopedia of functional activity



Functional regulatory activity explains GWAS associations

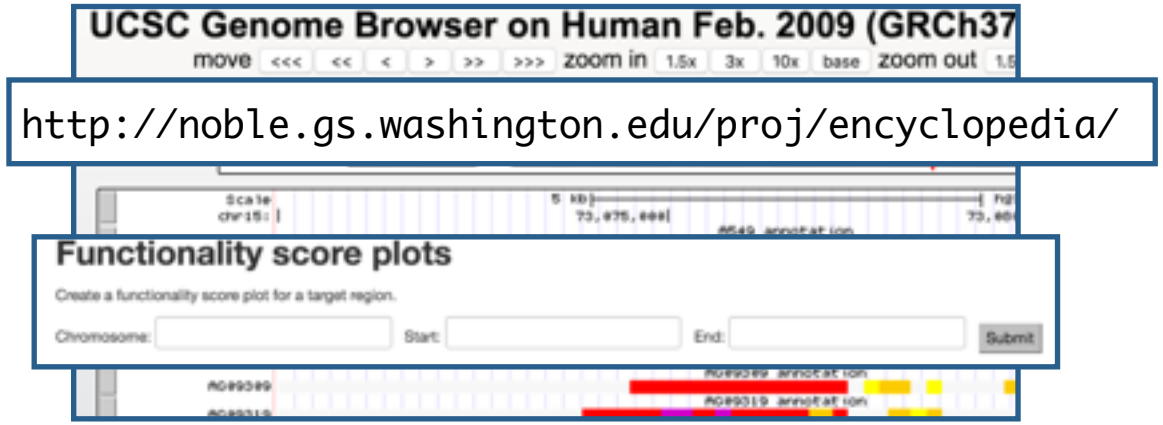


Annotation pipeline

A machine learning classifier recapitulates human interpretation

A unified Segway encyclopedia

Accessing the annotations



http://noble.gs.washington.edu/proj/encyclopedia/

Segway encyclopedia of human regulatory elements and annotation of 164 human cell types

> Libbrecht MW, Rodriguez O, Hoffman MM, Bilmes JA, Noble WS. 2016. A unified encyclopedia of human functional elements through fully automated annotation of 164 human cell types. *In prep*.

Download the annotations

- **Cell type-specific annotations:** Directory. Each annotation is in gzipped BED format, where the fourth column is the annotation label.
- **Encyclopedia:** BED format. Encyclopedia segments are a contiguous regions of high functionality score, generally between 300-20,000 bp. Columns correspond to: (1) chromosome; (2) start; (3) end; (4) sum of functionality score; (5) average base-wise functionality score; (6+) majority label of each cell type in the segment.

View the annotations

- **Individual annotations**: UCSC Genome Browser. Summary reports.
- **Encyclopedia**: UCSC Genome Browser.

Functionality score plots

Create a functionality score plot for a target region.

Chromosome: Start: End: Submit

Label meanings

- **Quiescent:** Inactive region.
- **ConstitutiveHet:** Heterochromatin marking permanently silent regions, characterized by the histone modification H3K9me3.
- **FacultativeHet:** Heterochromatin marking regions of cell type-specific repression, characterized by the histone modification H3K27me3. Also known as Polycomb-repressed heterochromatin.
- **Transcribed:** Transcribed genic region.
- **Promoter:** Regulatory region that occurs directly upstream of transcription start sites.
- **Enhancer:** Gene-distal regulatory element.
- **RegPermissive:** Region with weak marks of regulatory activity such as H3K4me1 or DNase hypersensitivity. May or may not directly control gene expression.
- **Bivalent:** Regulatory element with marks of both activation (such as H3K27ac) and repression (H3K27me3).
- **LowConfidance:** An annotation label that the interpretation classifier could not confidently assign to one of the above categories.

Support

- Information on Segway, including source code and documentation.
- If you have questions please write to the `segway-users` mailing list.

Thank you