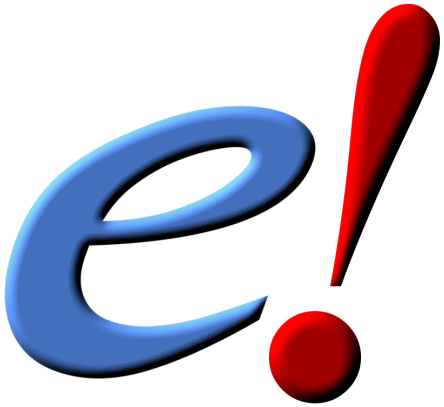
ENCODE data in Ensembl

EMBL-EBI

wellcome trust sanger institute

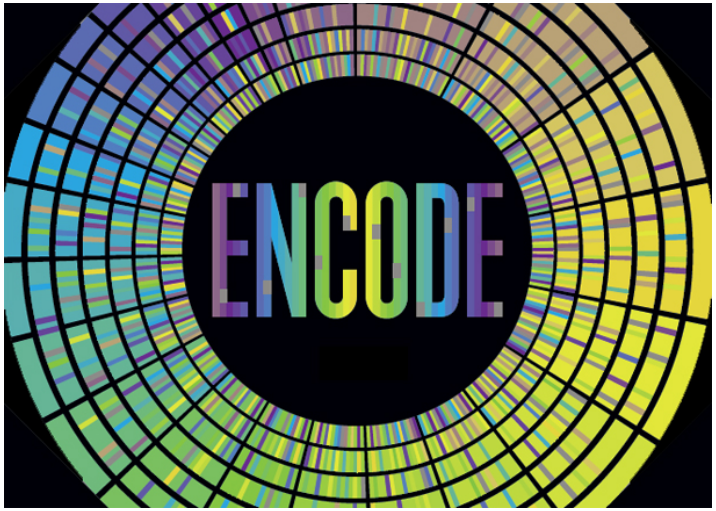


Ensembl Regulation

- The goal of Ensembl Regulation team is to annotate the genome with features that may play a role in the transcriptional regulation of genes.

- Predicted open/closed chromatin
- Transcription factor binding sites
- DNase I sensitivity
- Epigenetic marks
- RNA Pol binding

Data sources



ROADMAP epigenomics PROJECT

BLUEPRINT epigenome

cisRED

APPLIED MICROARRAYS

Agilent Technologies

VISTA

OneArray® by Phalanx Biotech Group

Affymetrix

An ENCODE subset

- Only a subset of ENCODE data is displayed in Ensembl.
- We display cell lines that have, at a minimum:
  - CTCF binding
  - DNase or FAIRE data
  - H3K4me3, H3K27me3, H3K36me3 data
- We display all TFBS and histone modification data known in these cell lines.

- Further ENCODE data can be added using the ENCODE track hub.

ENCODE data in Ensembl

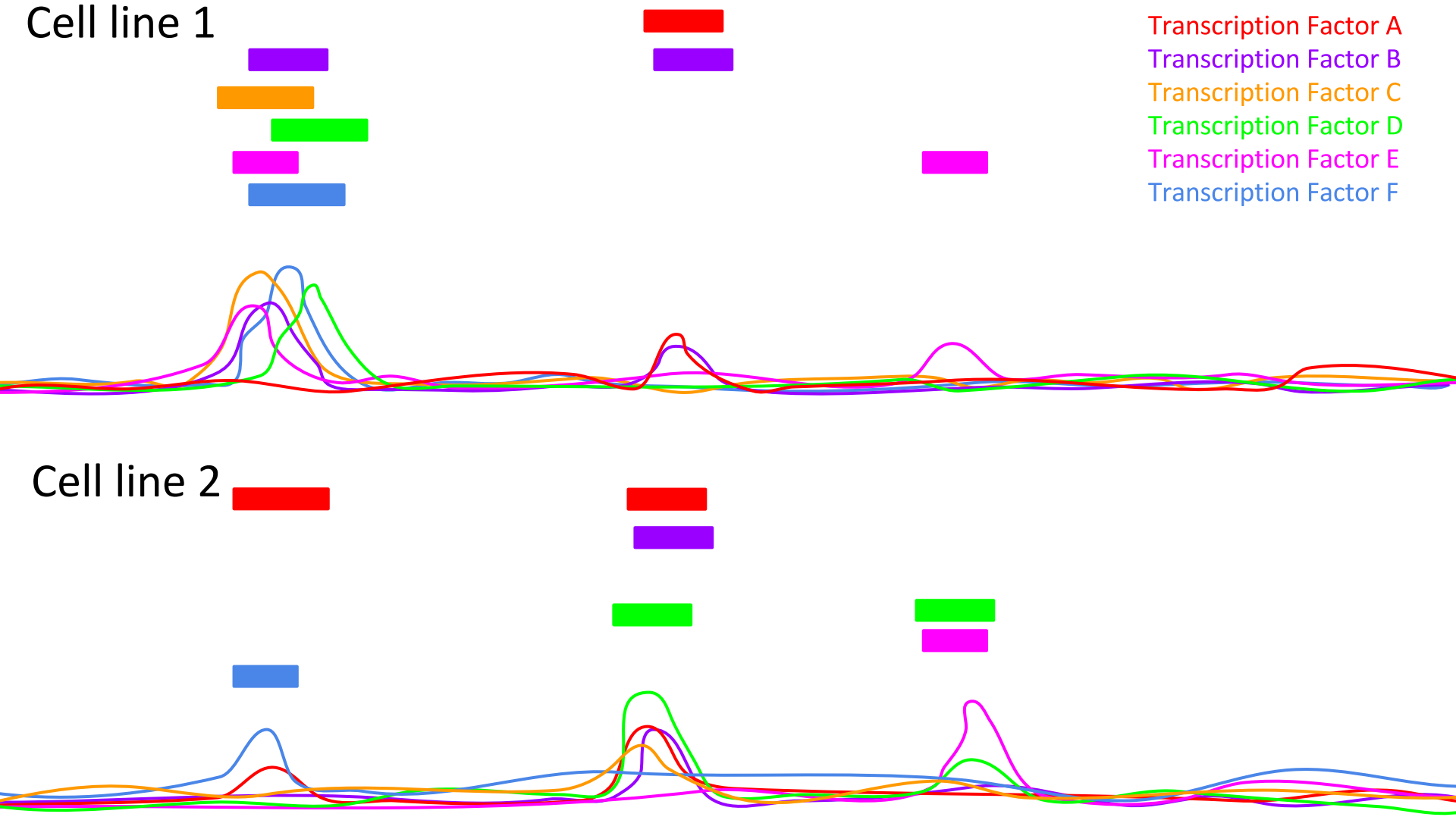
- The raw data is taken from ENCODE.
- This is processed to predict the positions of regulatory features, such as proximal and distal elements and insulators.
- The activity of these features is predicted in the different ENCODE cell lines.

- All of this can be viewed in the genome browser.

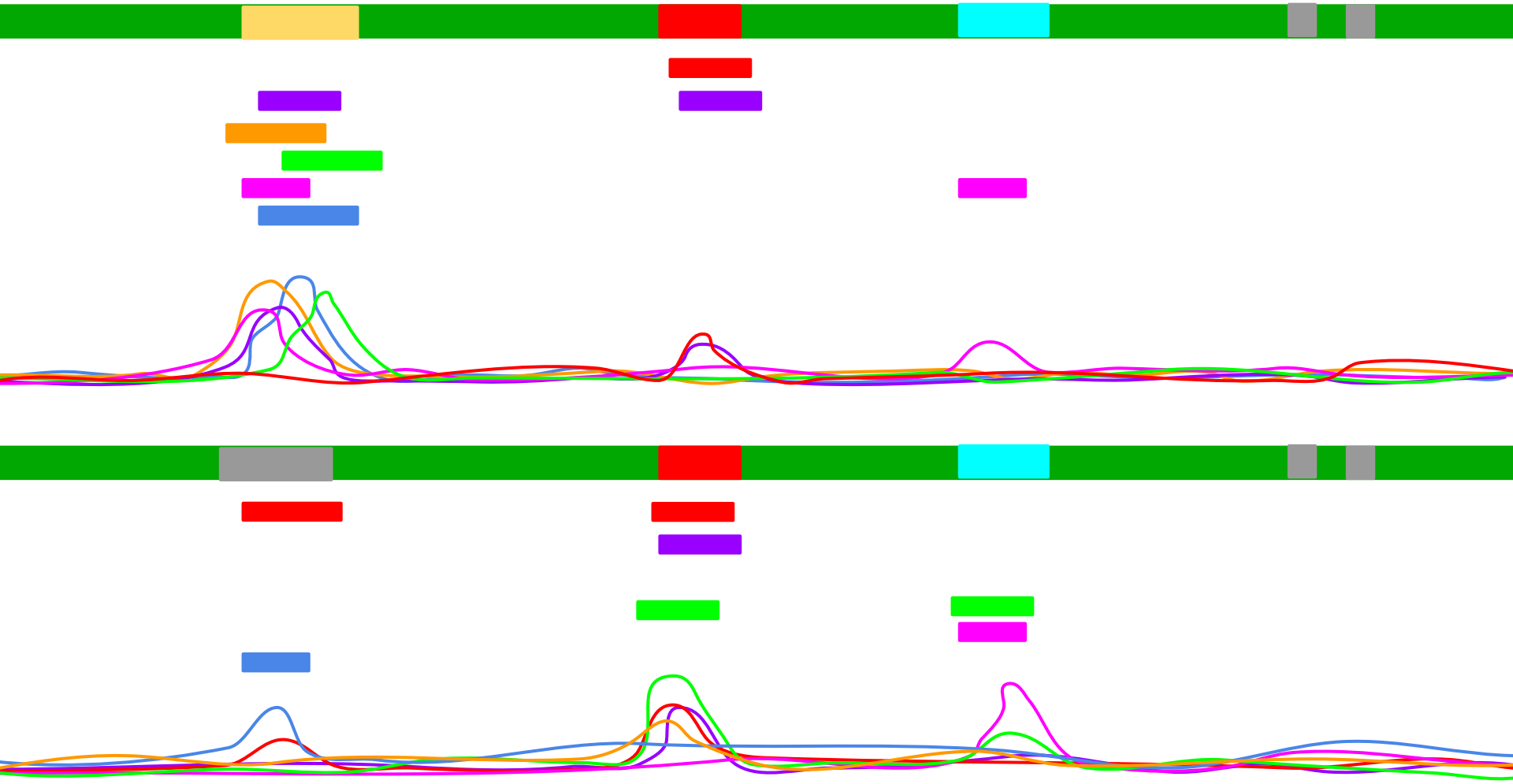
Changes are afoot...

- Ensembl are currently overhauling the way we display and process regulatory data.
- This is be complete and released in full in July.
- I'd like to show you a preview of these displays.

We start with the raw data



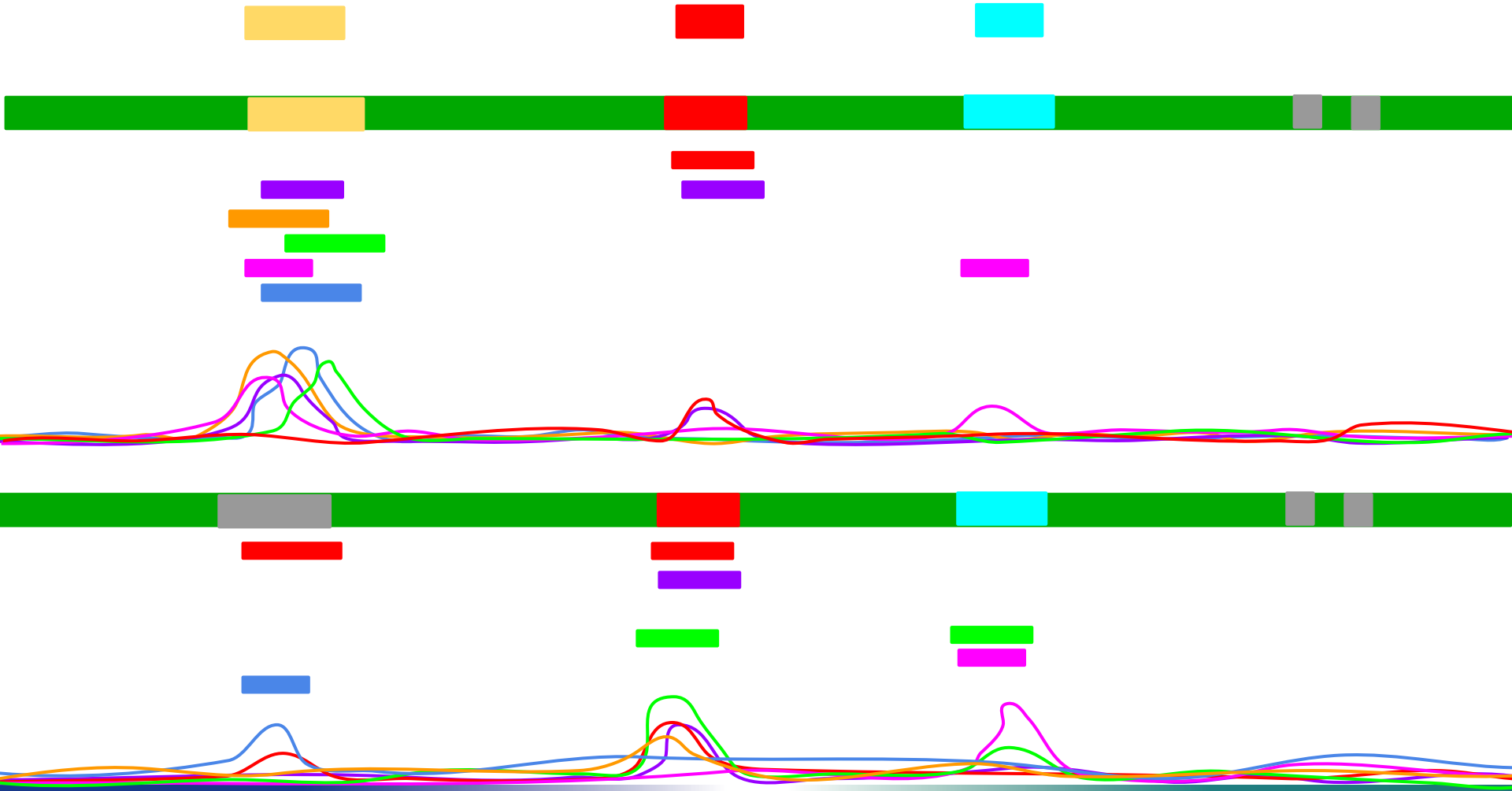
We process the data in each cell line to produce segments



How does segmentation work?

- Raw regulatory data (eg histone mods, TFBSs, DNase) is plotted onto the genome.
- We search for patterns – what tends to colocalise with what?
- These colocalisations are categorised depending on known functions.

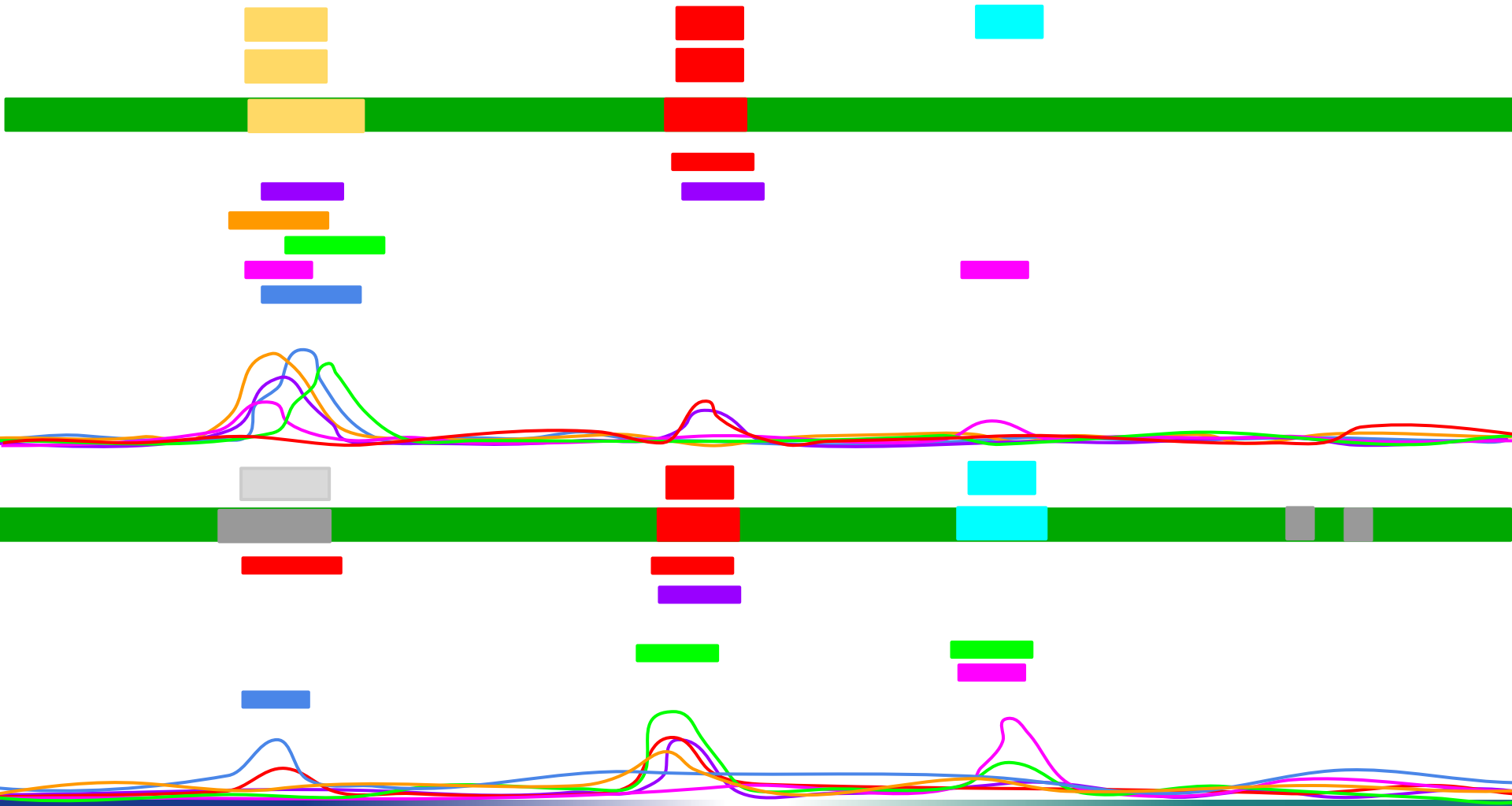
We compare segments in all cell lines to give a genome build



The genome build

- The segments in each cell line are compared.
- Where we see similar features in many cell lines, we say that a feature is present in the genome.

We project back to assign a state to each feature in each cell line



Projection

- For each segment in the genome build, we compare to the cell line segmentation.
- We use this to predict the activity of the segments in each cell line.

Coverage

Label	Count	Mean length (bp)	Max length (bp)	Total length (Mbp)
TSS	40,249	973.2	11,400	39.2
Proximal Reg.	101,206	1005.5	15,000	101.8
Distal Reg.	209,081	526.1	8,400	110.0
CTCF	108,284	550.1	5,200	59.6
Unannotated TFBS	163,528	155.8	1,630	25.5
Union				299.2

Demo

- In order to demonstrate this new build, I'll use:
  - The ENCODE data already in Ensembl
  - A track hub of the updated build

- From July onwards these data will be integrated together in a more easy to use format

Help and documentation

Course online http://www.ebi.ac.uk/training/online/subjects/11

Tutorials www.ensembl.org/info/website/tutorials

Flash animations

www.youtube.com/user/EnsemblHelpdesk

http://u.youku.com/Ensemblhelpdesk

Email us helpdesk@ensembl.org

Ensembl public mailing lists dev@ensembl.org, announce@ensembl.org

Follow us


Facebook www.facebook.com/Ensembl.org


Twitter https://twitter.com/Ensembl

Blog www.ensembl.info

Publications

http://www.ensembl.org/info/about/publications.html

- Flicek, P. *et. al.*
  Ensembl 2014
  Nucleic Acids Research
  http://nar.oxfordjournals.org/content/early/2013/12/06/nar.gkt1196.full

- Xosé M. Fernández-Suárez and Michael K. Schuster
  Using the Ensembl Genome Server to Browse Genomic Sequence Data.
  Current Protocols in Bioinformatics 1.15.1-1.15.48 (2010)
  www.ncbi.nlm.nih.gov/pubmed/20521244

- Giulietta M Spudich and Xosé M Fernández-Suárez
  Touring Ensembl: A practical guide to genome browsing
  BMC Genomics **11**:295 (2010)
  www.biomedcentral.com/1471-2164/11/295

Acknowledgements

Ensembl 2014

Paul Flicek[1,2,*], M. Ridwan Amode[2], Daniel Barrell[2], Kathryn Beal[1], Konstantinos Billis[2], Simon Brent[2], Denise Carvalho-Silva[1], Peter Clapham[2], Guy Coates[2], Stephen Fitzgerald[1], Laurent Gil[1], Carlos García Girón[2], Leo Gordon[1], Thibaut Hourlier[2], Sarah Hunt[1], Nathan Johnson[1], Thomas Juettemann[1], Andreas K. Kähäri[2], Stephen Keenan[1], Eugene Kulesha[1], Fergal J. Martin[2], Thomas Maurel[1], William M. McLaren[1], Daniel N. Murphy[2], Rishi Nag[2], Bert Overduin[1], Miguel Pignatelli[1], Bethan Pritchard[2], Emily Pritchard[1], Harpreet S. Riat[2], Magali Ruffier[1], Daniel Sheppard[2], Kieron Taylor[1], Anja Thormann[1], Stephen J. Trevanion[2], Alessandro Vullo[1], Steven P. Wilder[1], Mark Wilson[2], Amonida Zadissa[1], Bronwen L. Aken[2], Ewan Birney[1], Fiona Cunningham[1], Jennifer Harrow[2], Javier Herrero[1], Tim J.P. Hubbard[2], Rhoda Kinsella[1], Matthieu Muffato[1], Anne Parker[2], Giulietta Spudich[1], Andy Yates[1], Daniel R. Zerbino[1] and Stephen M.J. Searle[2]

Funding

wellcome trust

EMBL

National Human Genome Research Institute

BBSRC bioscience for the future

European Commission
Framework Programme 7

SEVENTH FRAMEWORK PROGRAMME

BLUEPRINT epigenome

EURATRANS

HELIX NEBULA THESCIENCECLOUD