Using Ensembl tools for browsing ENCODE data

Bert Overduin, Ph.D.
Vertebrate Genomics Team
EMBL- European Bioinformatics Institute
Wellcome Trust Genome Campus
Hinxton, Cambridge CB10 1SD
United Kingdom

Outline

- Presentation
  - Introduction to Ensembl
  - ENCODE data hub
  - Ensembl Regulatory Build
  - Regulatory segmentation
  - Adding custom tracks
  - BioMart
- Worked examples
  - Browser
  - BioMart
- Hands-on exercises
  - Browser / Regulatory Build & segmentation
  - Browser / Adding custom tracks
  - BioMart

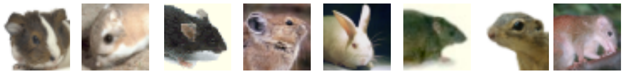
Ensembl - Goal

- To provide automatic annotation of completely sequenced vertebrate genomes
- To integrate this annotation with other available biological data
- To make all this information available to the scientific community
- http://www.ensembl.org

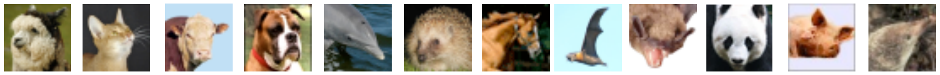
Species

Primates

Rodents etc.



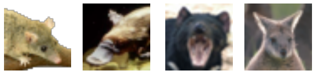
Laurasiatheria



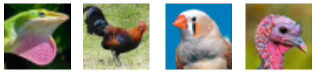
Afrotheria

Xenartha

Other mammals



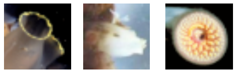
Birds & reptiles



Amphibians

Fish

Other chordates



Other eukaryotes

On *Pre!* Ensembl

68 species total (v66)

Data

- Genomic sequence
- Gene / transcript / protein models
- External references
- Mapped cDNAs, proteins, microarray probes, BAC clones, cytogenetic bands, repeats, markers etc. etc.
- Variation data
- Comparative data
- Regulatory data

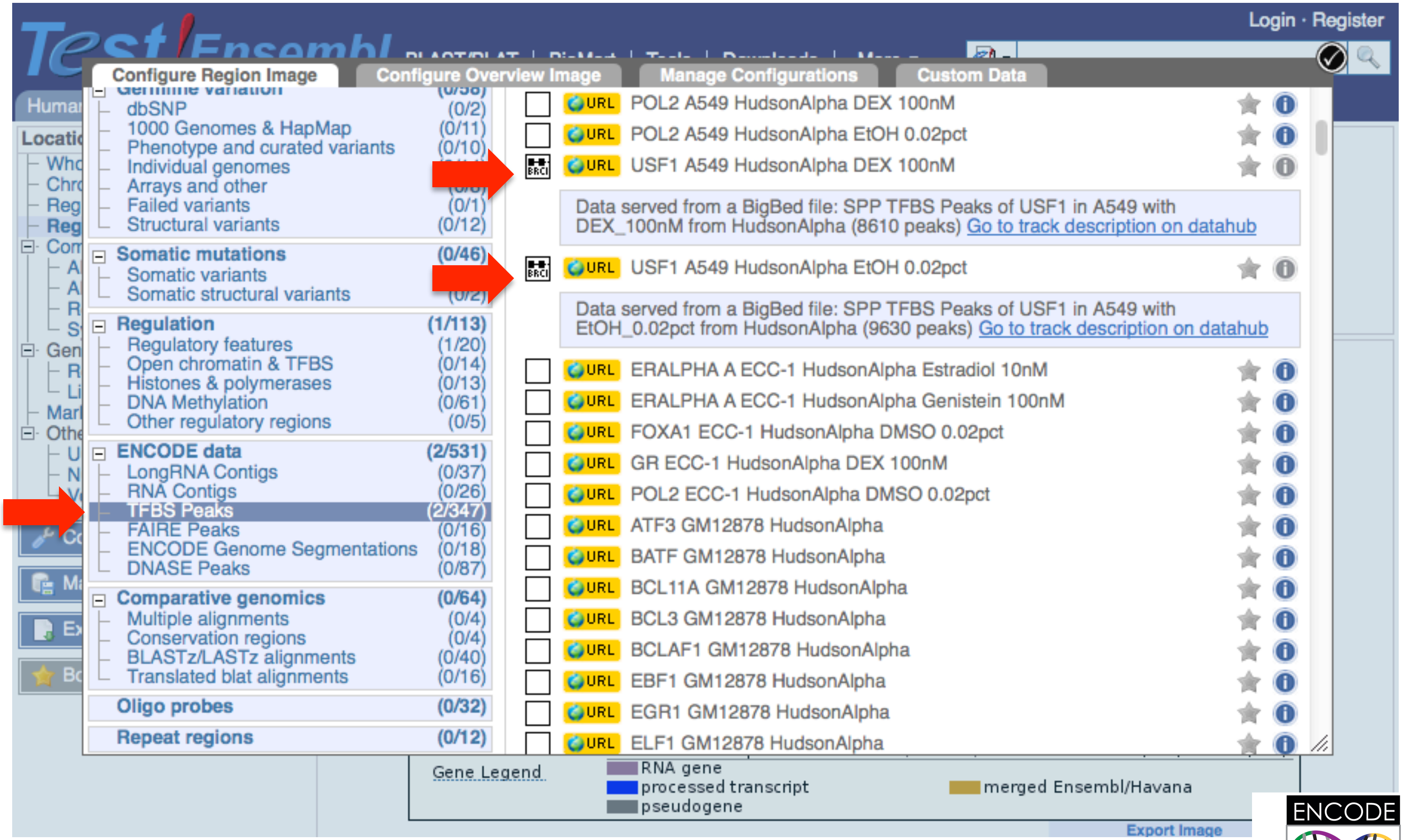
Access to data

• Ensembl web site	http://www.ensembl.org
• *Pre!* web site	http://pre.ensembl.org
• *Archive!* web site	http://archive.ensembl.org
• BioMart	http://www.ensembl.org/biomart/martview
• FTP site	ftp://ftp.ensembl.org
• Amazon Web Services	http://aws.amazon.com/publicdatasets
• MySQL	http://www.ensembl.org/info/data/mysql.html
• Perl API	http://www.ensembl.org/info/data/api.html

Official (cloud-based) mirrors

- United States West Coast
  http://uswest.ensembl.org
- United States East Coast
  http://useast.ensembl.org
- Asia
  http://asia.ensembl.org
- Geo-IP-based redirection

ENCODE data hub



Germline variation	(0/58)
dbSNP	(0/2)
1000 Genomes & HapMap	(0/11)
Phenotype and curated variants	(0/10)
Individual genomes	
Arrays and other	(0/8)
Failed variants	(0/1)
Structural variants	(0/12)
Somatic mutations	**(0/46)**
Somatic variants	
Somatic structural variants	(0/2)
Regulation	**(1/113)**
Regulatory features	(1/20)
Open chromatin & TFBS	(0/14)
Histones & polymerases	(0/13)
DNA Methylation	(0/61)
Other regulatory regions	(0/5)
ENCODE data	**(2/531)**
LongRNA Contigs	(0/37)
RNA Contigs	(0/26)
TFBS Peaks	(2/347)
FAIRE Peaks	(0/16)
ENCODE Genome Segmentations	(0/18)
DNASE Peaks	(0/87)
Comparative genomics	**(0/64)**
Multiple alignments	(0/4)
Conservation regions	(0/4)
BLASTz/LASTz alignments	(0/40)
Translated blat alignments	(0/16)
Oligo probes	**(0/32)**
Repeat regions	**(0/12)**

□ URL POL2 A549 HudsonAlpha DEX 100nM

□ URL POL2 A549 HudsonAlpha EtOH 0.02pct

☑ URL USF1 A549 HudsonAlpha DEX 100nM

Data served from a BigBed file: SPP TFBS Peaks of USF1 in A549 with DEX_100nM from HudsonAlpha (8610 peaks) Go to track description on datahub

☑ URL USF1 A549 HudsonAlpha EtOH 0.02pct

Data served from a BigBed file: SPP TFBS Peaks of USF1 in A549 with EtOH_0.02pct from HudsonAlpha (9630 peaks) Go to track description on datahub

□ URL ERALPHA A ECC-1 HudsonAlpha Estradiol 10nM

□ URL ERALPHA A ECC-1 HudsonAlpha Genistein 100nM

□ URL FOXA1 ECC-1 HudsonAlpha DMSO 0.02pct

□ URL GR ECC-1 HudsonAlpha DEX 100nM

□ URL POL2 ECC-1 HudsonAlpha DMSO 0.02pct

□ URL ATF3 GM12878 HudsonAlpha

□ URL BATF GM12878 HudsonAlpha

□ URL BCL11A GM12878 HudsonAlpha

□ URL BCL3 GM12878 HudsonAlpha

□ URL BCLAF1 GM12878 HudsonAlpha

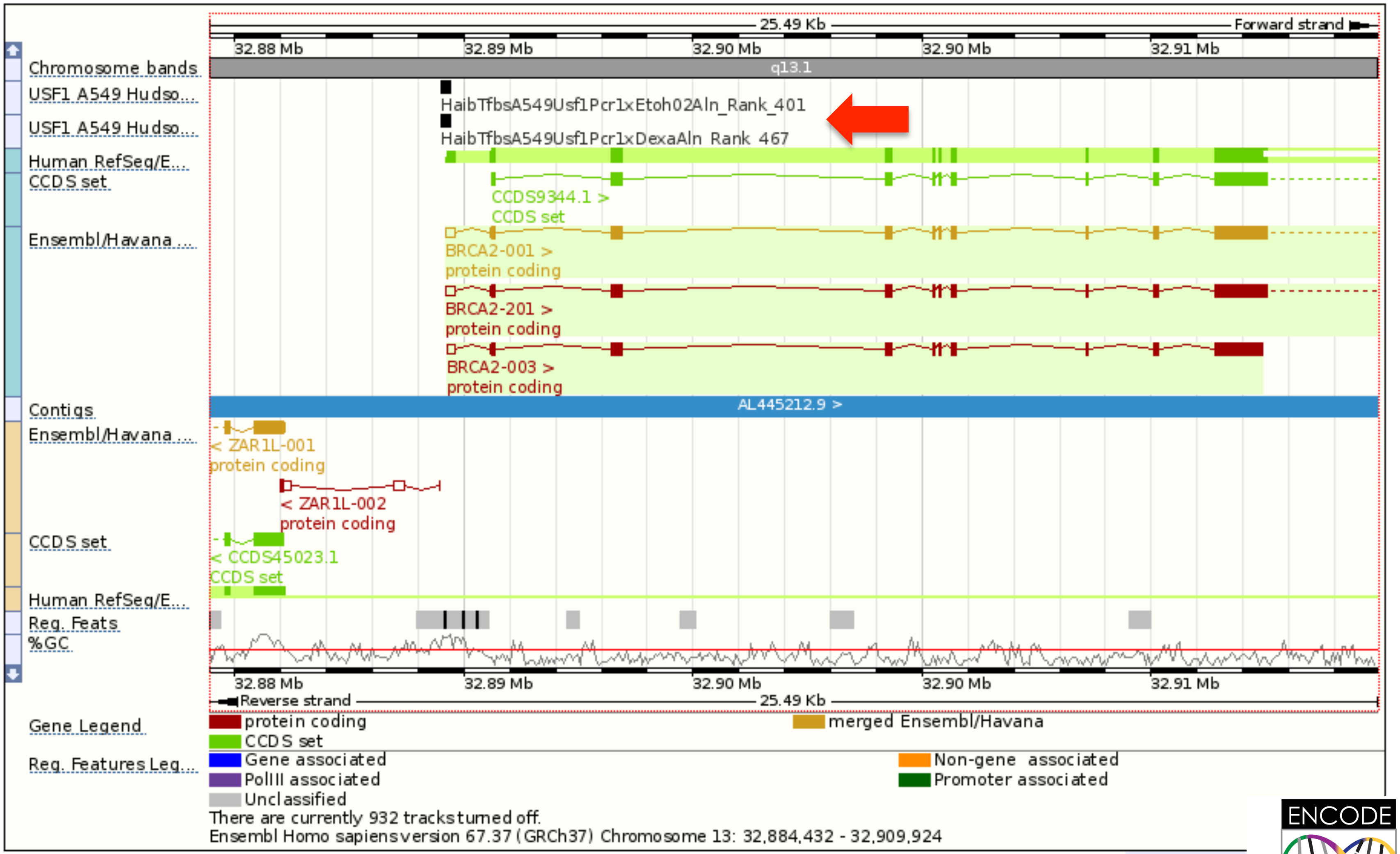
□ URL EBF1 GM12878 HudsonAlpha

□ URL EGR1 GM12878 HudsonAlpha

□ URL ELF1 GM12878 HudsonAlpha

Gene Legend		
	RNA gene	merged Ensembl/Havana
	processed transcript	
	pseudogene	

ENCODE data hub



There are currently 932 tracks turned off.
Ensembl Homo sapiens version 67.37 (GRCh37) Chromosome 13: 32,884,432 - 32,909,924

Ensembl Regulatory Build

- Provides a single “best guess” set of regulatory features
- For human and mouse
- Created by overlap analysis of annotations from genome-wide data sets in a two stage cell type aware manner
- http://www.ensembl.org/info/docs/funcgen/index.html

Regulatory Build data

Focus features (define potential binding sites)

- Open chromatin (DNase1, FAIRE)
- CTCF (insulator/enhancer) binding sites
- Binding sites for 90 transcription factors

Attribute features

- 42 Histone modifications (methylation, acetylation)
- RNA Pol II and III binding sites

13 cell types

ENCODE

Roadmap Epigenomics

Focus features (define potential binding sites)

- Open chromatin (DNase1)
- CTCF (insulator/enhancer) binding sites
- Binding sites for 21 transcription factors

Attribute features

- 8 Histone modifications (methylation)
- RNA Pol II binding sites

5 cell types

ENCODE

- Meta data: http://www.ensembl.org/Homo_sapiens/Experiment/Sources

Regulatory Build procedure

Regulatory feature construction:

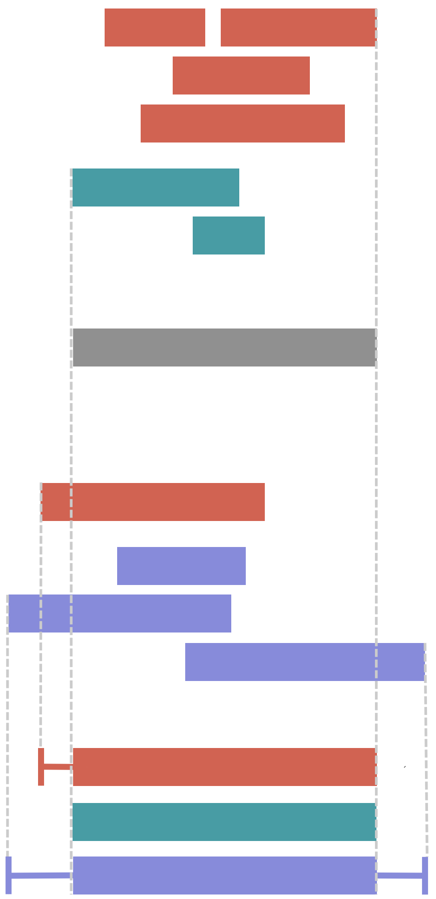
- Identify core regions across all available cell types using focus features
- Extend core regions in a cell type specific manner using attribute features

Regulatory feature annotation:

- Classify regulatory features
- Annotate the position of putative TFBSs using position weight matrices (PWMs) taken from the JASPAR database

ENCODE

Regulatory feature construction



Focus features

DNase1 Cell type 1

CTCF Cell type 1

Taf1 Cell type 1

DNase1 Cell type 2

CTCF Cell type 2

MultiCell reg feature

Attribute features

H3K4me2 Cell type 1

H3K4me2 Cell type 3

H3K4me3 Cell type 3

H3K9ac Cell type 3

Cell type 1 reg feature

Cell type 2 reg feature

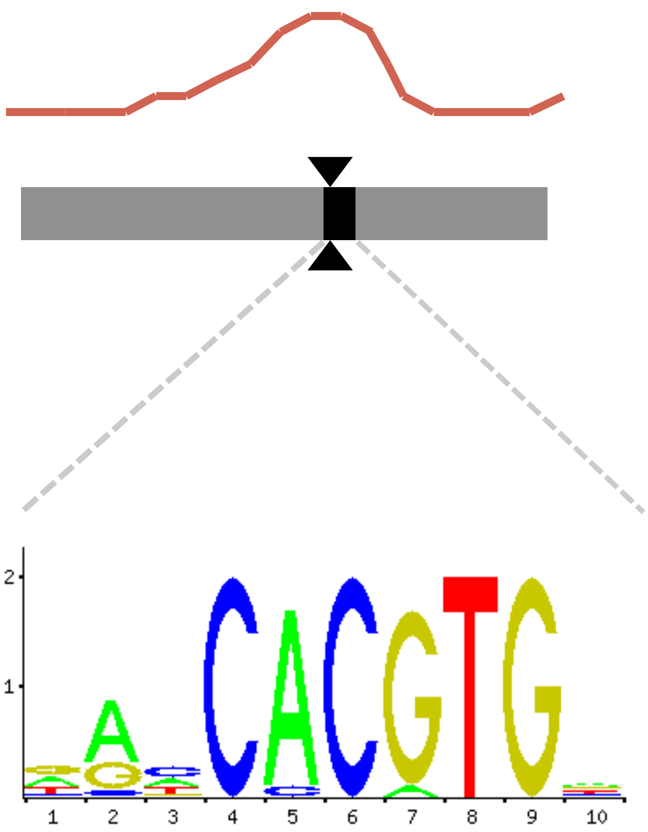
Cell type 3 reg feature

Regulatory feature annotation

🟩	Promoter Associated	Patterns over-represented in the region of the transcription start site plus or minus 2500 bp upstream of protein coding genes, but not in the downstream gene body. Likely to be a 5' proximal promoter.
🟦	Gene Associated	Patterns over-represented in gene bodies. Often represent gene's transcriptional activity (expressed/ repressed).
🟧	Non-gene Associated	Patterns over-represented in non-gene regions. Likely to correspond to a distal regulatory element such as an insulator or enhancer.
🟪	Polymerase III Associated	Patterns over-represented in regions 2500 bp upstream of PolIII transcribed regions e.g. tRNAs. Likely to correspond to a proximal regulatory element specifically associated to Polymerase III transcription.
⬜	Unclassified	Patterns which are currently unclassifiable.

ENCODE

Regulatory feature annotation



ChiP-Seq signal for transcription factor MAX

regulatory feature

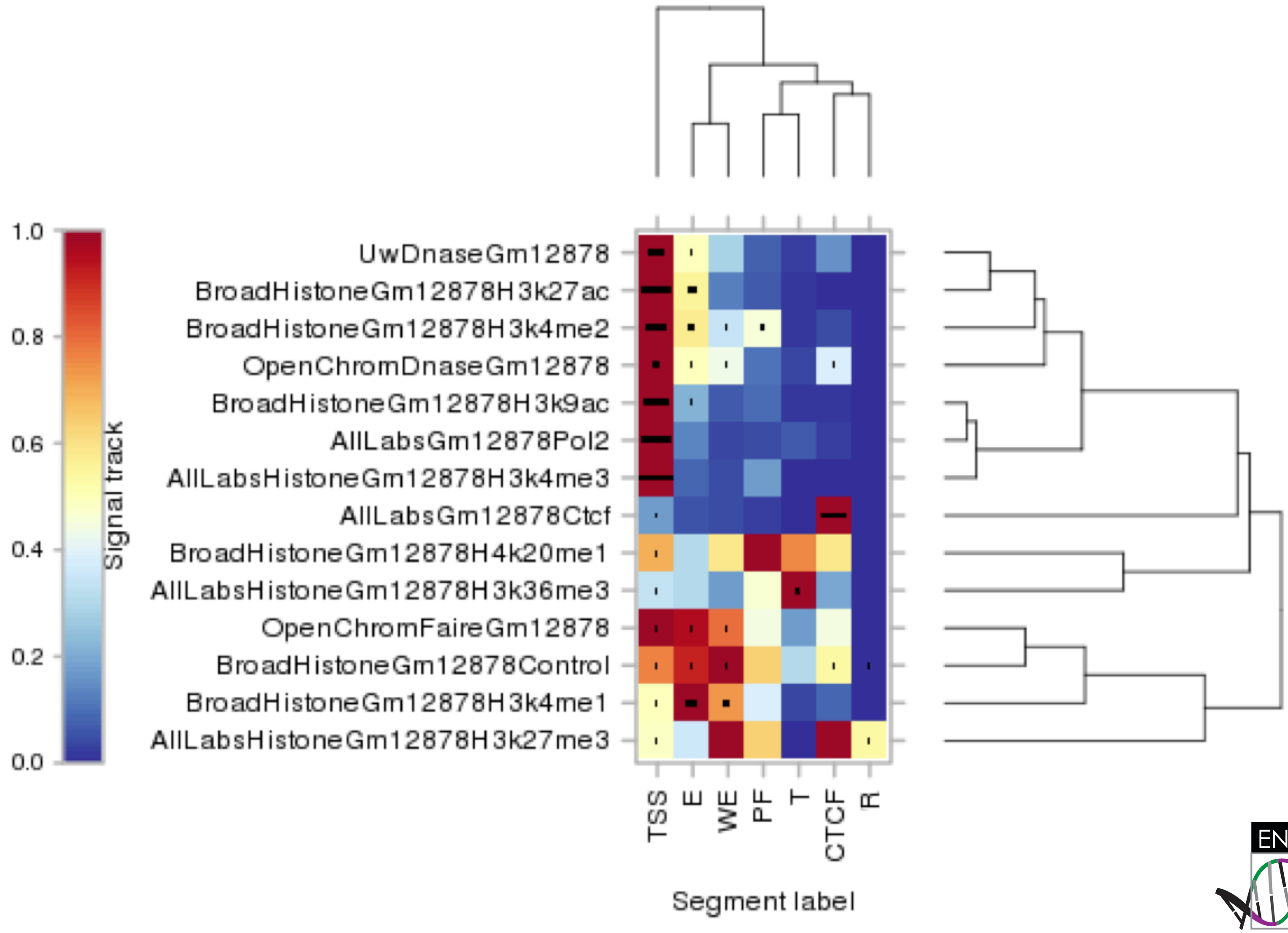
Position Weight Matrix for MAX from JASPAR database

Regulatory segmentation

- Provides a summary of the functional architecture (or “state”) of the human genome
- 6 cell types
- 14 assays, constituting 3 classes of data:
open chromatin, transcription factors, histone modifications
- Produce segmentations using 2 programs:
ChromHMM and Segway
- Classify segments into 7 classes

ENCODE

Regulatory segmentation



Regulatory segmentation

🟦	CTCF	CTCF enriched
🟨	WE	Predicted Weak Enhancer/Cis-reg element
🟩	T	Predicted Transcribed Region
🟧	E	Predicted Enhancer
🟥	PF	Predicted Promoter Flank
⬛	R	Predicted Repressed/Low Activity
🟥	TSS	Predicted Promoter with TSS

Adding custom tracks

Upload data

- 5 MB limit
- Data saved by Ensembl

Attach remote file

- No size limit
- URL-based (http or ftp)
- Data can be updated by the data provider without having to re-upload them
- Data are pulled from remote location every time a view is loaded, so it can take a bit longer time to load

Adding custom tracks

Possible formats:

• BAM	sequence alignments (no upload)
• BED	genes / features
• BedGraph	continuous-valued data
• BigBed	genes / features (no upload)
• BigWig	continuous-valued data (no upload)
• GBrowse	genes / features
• GFF	genes / features
• GTF	genes / features
• PSL	sequence alignments
• VCF	variants (no upload)
• WIG	continuous-valued data

BioMart

- Data retrieval tool
- Originally developed for Ensembl (EnsMart)
- Now used by many large data resources
- Integrated with several widely used software packages
- Joint project between the European Bioinformatics Institute (EBI) and the Ontario Institute for Cancer Research (OICR)
- Central portal: http://www.biomart.org

BioMart

- Step 1 – Dataset

  Choose your dataset and species

- Step 2 – Filters

  Limit your dataset

- Step 3 – Attributes

  Specify what information you want to output

- Step 4 – Results

  Preview and output your results

Help

- Helpdesk
  helpdesk@ensembl.org
- Mailing lists
  http://www.ensembl.org/info/about/contact/mailing.html
- YouTube and YouKu (优酷网) channels:
  http://www.youtube.com/user/EnsemblHelpdesk
  http://u.youku.com/user_show/uid_Ensemblhelpdesk

Keeping in touch

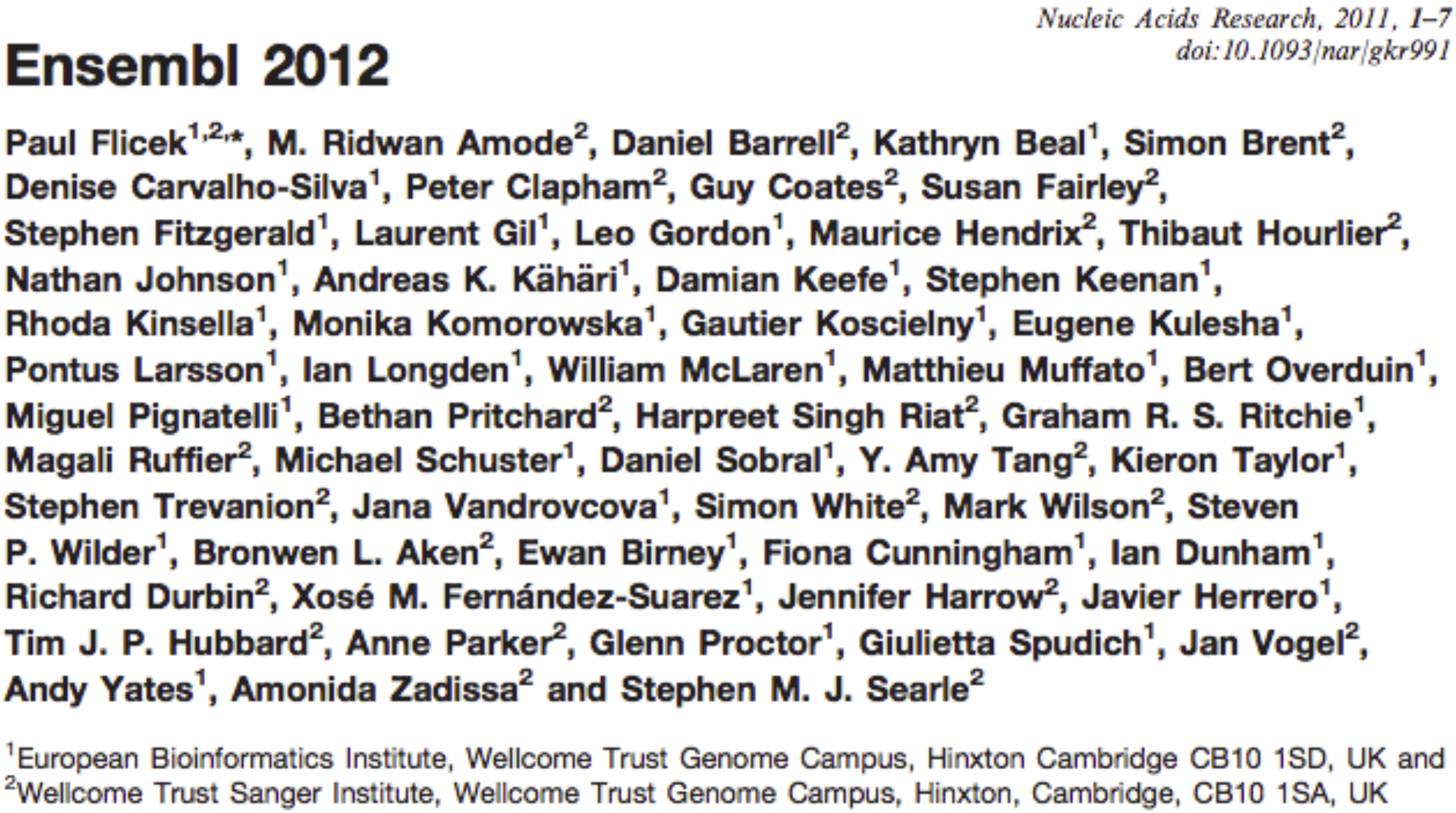
- Blog
  http://www.ensembl.info
- Facebook
  http://www.facebook.com/Ensembl.org
- Twitter
  http://twitter.com/Ensembl

Workshops

- Browser (0.5-2 days) and API (1-3 days) workshops
- Combination of lectures and hands-on exercises
- Advertised on http://www.ensembl.info/workshops/calendar/

- You can host your own workshop!
- For academic institutions there is no fee, apart from the instructor's expenses
- You only need a computer room and participants
- You can get more info from helpdesk@ensembl.org or me (bert@ebi.ac.uk)

Acknowledgements



Nucleic Acids Research, 2011, 1–7
doi:10.1093/nar/gkr991

Ensembl 2012

Paul Flicek[1,2,*], M. Ridwan Amode[2], Daniel Barrell[2], Kathryn Beal[1], Simon Brent[2], Denise Carvalho-Silva[1], Peter Clapham[2], Guy Coates[2], Susan Fairley[2], Stephen Fitzgerald[1], Laurent Gil[1], Leo Gordon[1], Maurice Hendrix[2], Thibaut Hourlier[2], Nathan Johnson[1], Andreas K. Kähäri[1], Damian Keefe[1], Stephen Keenan[1], Rhoda Kinsella[1], Monika Komorowska[1], Gautier Koscielny[1], Eugene Kulesha[1], Pontus Larsson[1], Ian Longden[1], William McLaren[1], Matthieu Muffato[1], Bert Overduin[1], Miguel Pignatelli[1], Bethan Pritchard[2], Harpreet Singh Riat[2], Graham R. S. Ritchie[1], Magali Ruffier[2], Michael Schuster[1], Daniel Sobral[1], Y. Amy Tang[2], Kieron Taylor[1], Stephen Trevanion[2], Jana Vandrovcova[1], Simon White[2], Mark Wilson[2], Steven P. Wilder[1], Bronwen L. Aken[2], Ewan Birney[1], Fiona Cunningham[1], Ian Dunham[1], Richard Durbin[2], Xosé M. Fernández-Suarez[1], Jennifer Harrow[2], Javier Herrero[1], Tim J. P. Hubbard[2], Anne Parker[2], Glenn Proctor[1], Giulietta Spudich[1], Jan Vogel[2], Andy Yates[1], Amonida Zadissa[2] and Stephen M. J. Searle[2]

[1]European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton Cambridge CB10 1SD, UK and [2]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, UK

Funded primarily by the Wellcome Trust,
with additional funding by NIH-NHGRI, EU and EMBL

Ensembl Team Retreat 2012
Norwich, United Kingdom