ENCODE: Understanding the Genome

Michael Snyder

May 8, 2013

Conflicts: Personalis, Genapsys, Illumina
Slides From Ewan Birney, Marc Schaub, Alan Boyle

Encyclopedia of DNA Elements (ENCODE)

- NHGRI-funded consortium
- Goal: delineate all "functional" elements in the human genome
- Wide array of experimental assays
- Three Phases: 1) Pilot 2) Scale Up 1.0 3) Scale up 2.0

The ENCODE Project Consortium. An Integrated Encyclopedia of DNA Elements in the Human Genome. *Nature 2012*
Project website: http://encodeproject.org

The ENCODE Consortium Phase 2

Brad Bernstein (Eric Lander, Manolis Kellis, Tony Kouzarides)

Ewan Birney (Jim Kent, Mark Gerstein, Bill Noble, Peter Bickel, Ross Hardison, Zhiping Weng)

Greg Crawford (Ewan Birney, Jason Lieb, Terry Furey, Vishy Iyer)

Jim Kent (David Haussler, Kate Rosenbloom)

John Stamatoyannopoulos (Evan Eichler, George Stamatoyannopoulos, Job Dekker, Maynard Olson, Michael Dorschner, Patrick Navas, Phil Green)

Mike Snyder (Kevin Struhl, Mark Gerstein, Peggy Farnham, Sherman Weissman)

Rick Myers (Barbara Wold)

Scott Tenenbaum (Luiz Penalva)

Tim Hubbard (Alexandre Reymond, Alfonso Valencia, David Haussler, Ewan Birney, Jim Kent, Manolis Kellis, Mark Gerstein, Michael Brent, Roderic Guigo)

Tom Gingeras (Alexandre Reymond, David Spector, Greg Hannon, Michael Brent, Roderic Guigo, Stylianos Antonarakis, Yijun Ruan, Yoshihide Hayashizaki)

Zhiping Weng (Nathan Trinklein, Rick Myers)

Additional ENCODE Participants: Elliott Marguiles, Eric Green, Job Dekker, Laura Elnitski, Len Pennachio, Jochen Wittbrodt

.. and many senior scientists, postdocs, students, technicians, computer scientists, statisticians and administrators in these groups

NHGRI: Elise Feingold, Mike Pazin, Peter Good

The ENCODE Consortium Phase 3

Brad Bernstein (Eric Lander, Manolis Kellis, Tony Kouzarides)

Ewan Birney (Jim Kent, Mark Gerstein, Bill Noble, Peter Bickel, Ross Hardison, Zhiping Weng)

Greg Crawford (Ewan Birney, Jason Lieb, Terry Furey, Vishy Iyer)

Jim Kent (David Haussler, Kate Rosenbloom) **Mike Cherry**

John Stamatoyannopoulos (Evan Eichler, George Stamatoyannopoulos, Job Dekker, Maynard Olson, Michael Dorschner, Patrick Navas, Phil Green)

Mike Snyder (Kevin Struhl, Mark Gerstein, Peggy Farnham, Sherman Weissman)

Rick Myers (Barbara Wold)

Scott Tenenbaum (Luiz Penalva)

Tim Hubbard (Alexandre Reymond, Alfonso Valencia, David Haussler, Ewan Birney, Jim Kent, Manolis Kellis, Mark Gerstein, Michael Brent, Roderic Guigo)

Tom Gingeras (Alexandre Reymond, David Spector, Greg Hannon, Michael Brent, Roderic Guigo, Stylianos Antonarakis, Yijun Ruan, Yoshihide Hayashizaki)

Zhiping Weng (Nathan Trinklein, Rick Myers)

Brenton Graveley (John Rinn, Others)

.. and many senior scientists, postdocs, students, technicians, computer scientists, statisticians and administrators in these groups

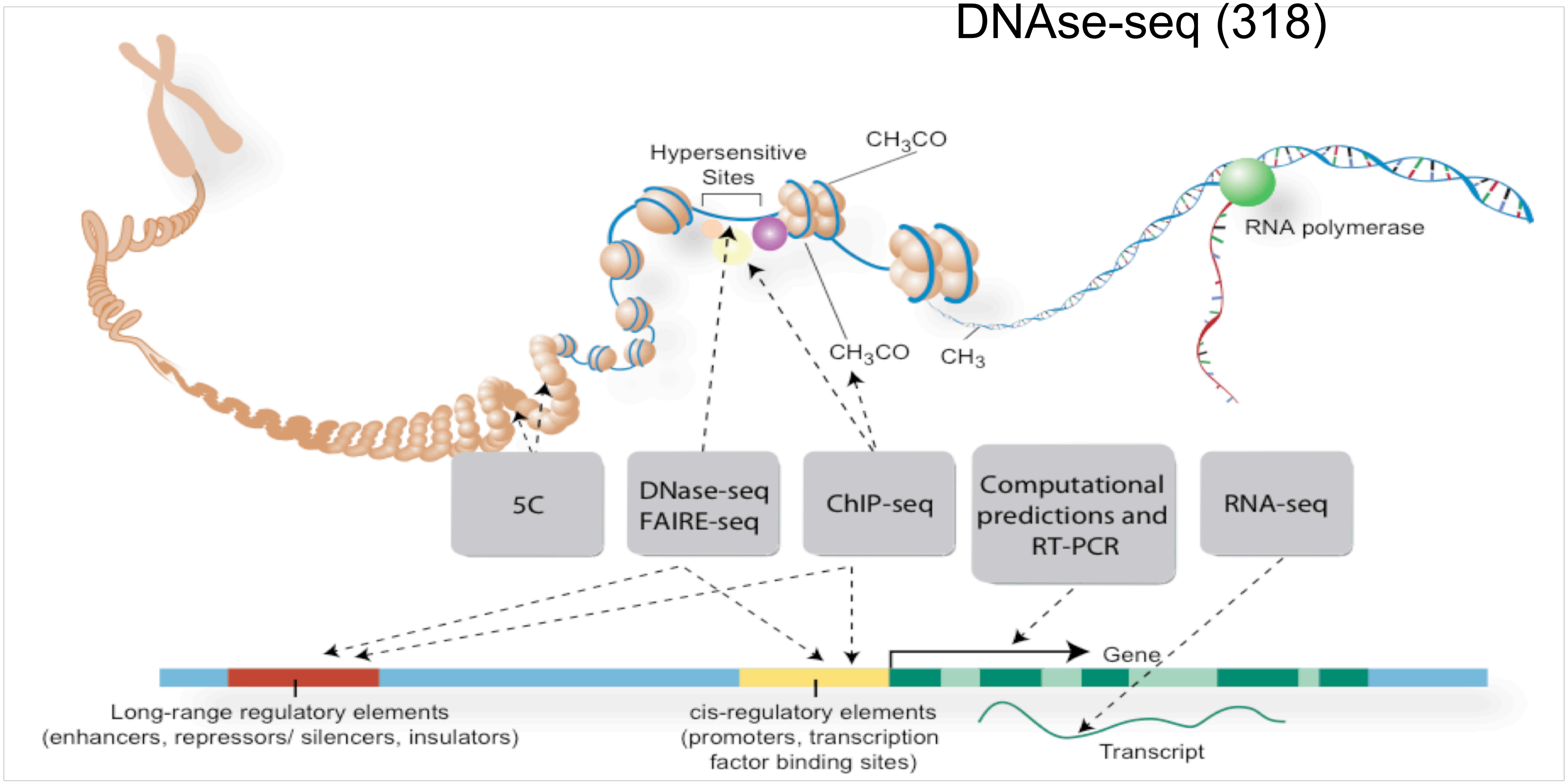
NHGRI: Elise Feingold, Mike Pazin, Peter Good

Experimental Assays

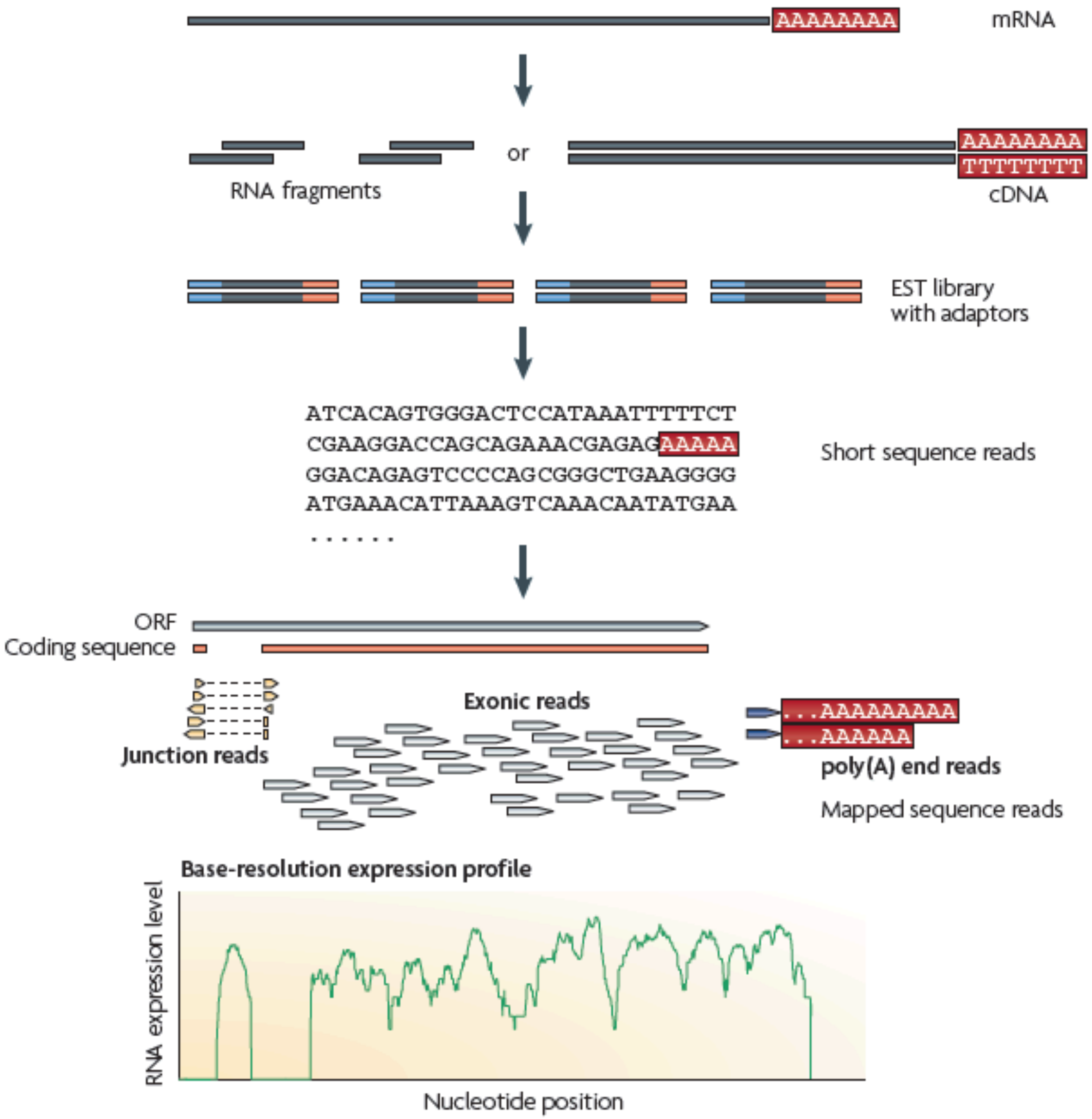
Chip-seq (180 TFs
+ Histone marks;
1770 data sets)

RNA-seq (418)

DNAse-seq (318)

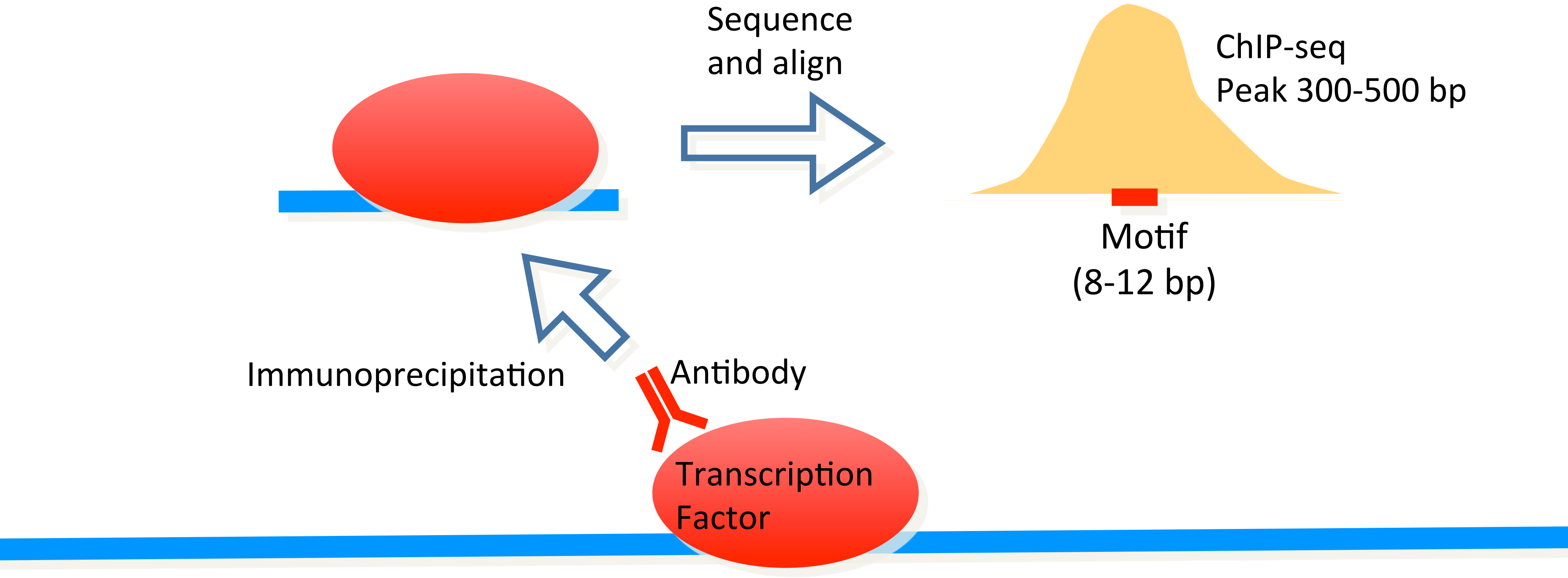


RNA-Sequencing



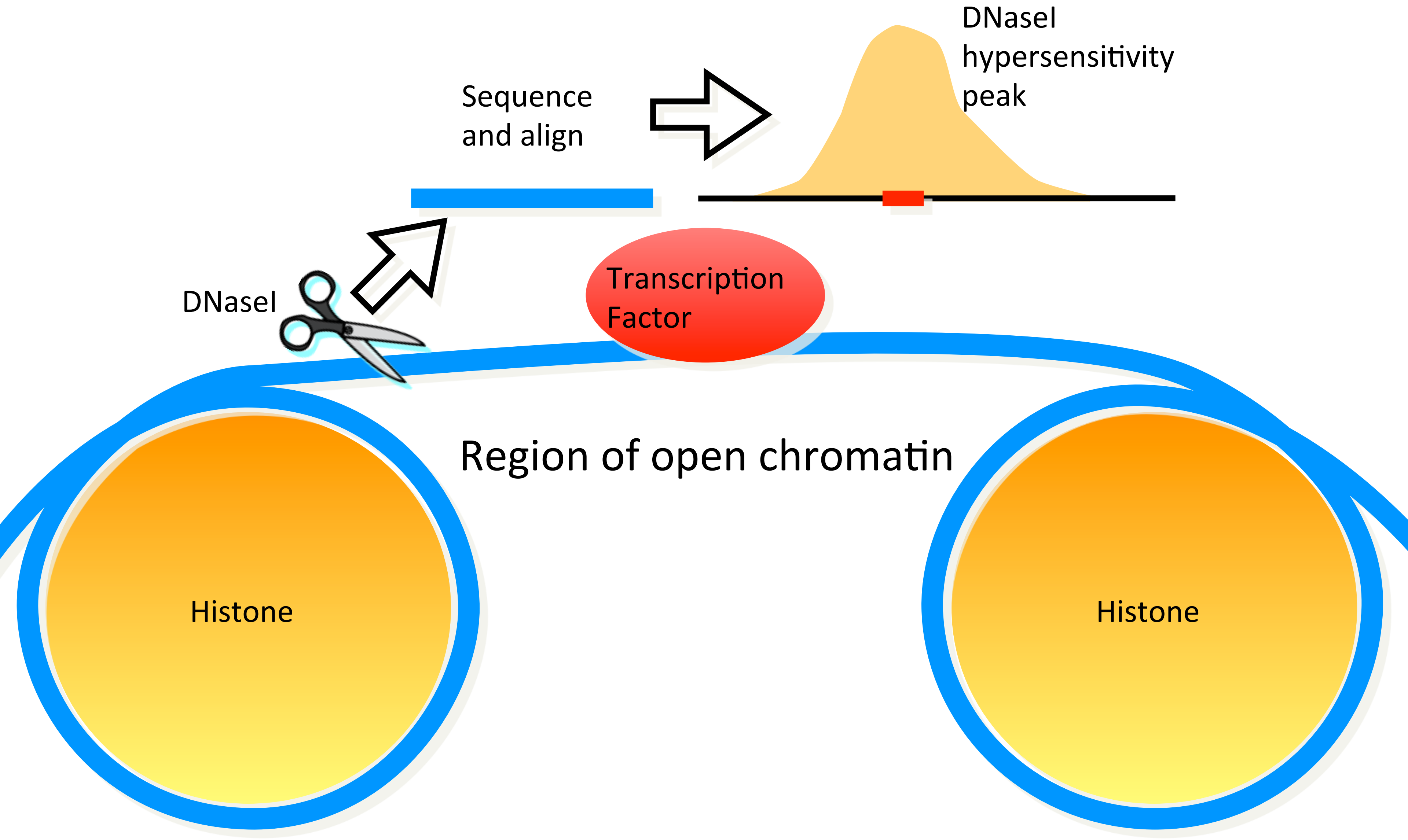
Wang et al. 2009 Nat Gen. Rev.

Functional data: ChIP-seq

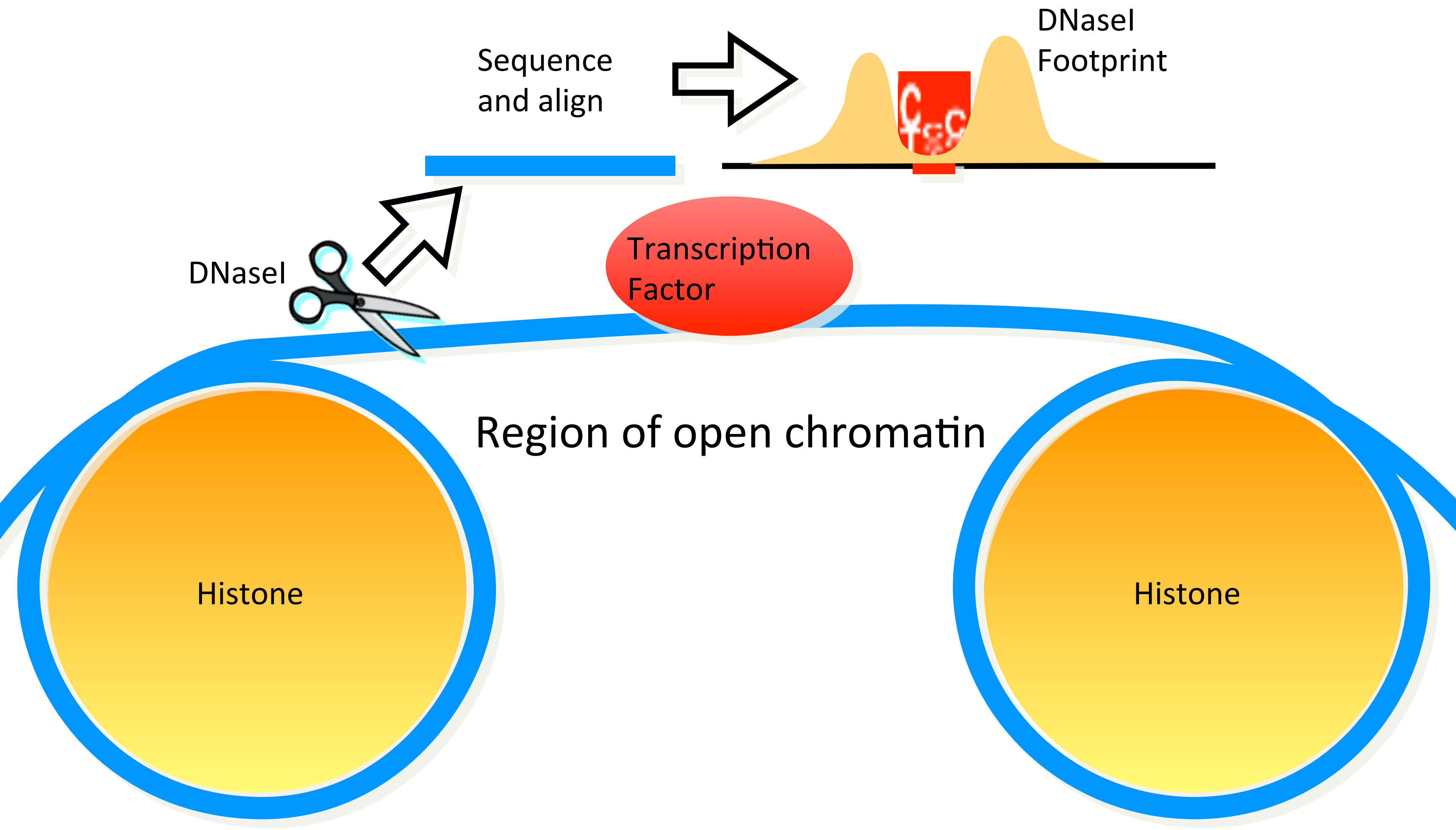


ChIP-exo
Histone Marks

Functional data: DNase-seq



Functional data: DNase footprints



Sequence
and align

DNaseI
Footprint

DNaseI

Transcription
Factor

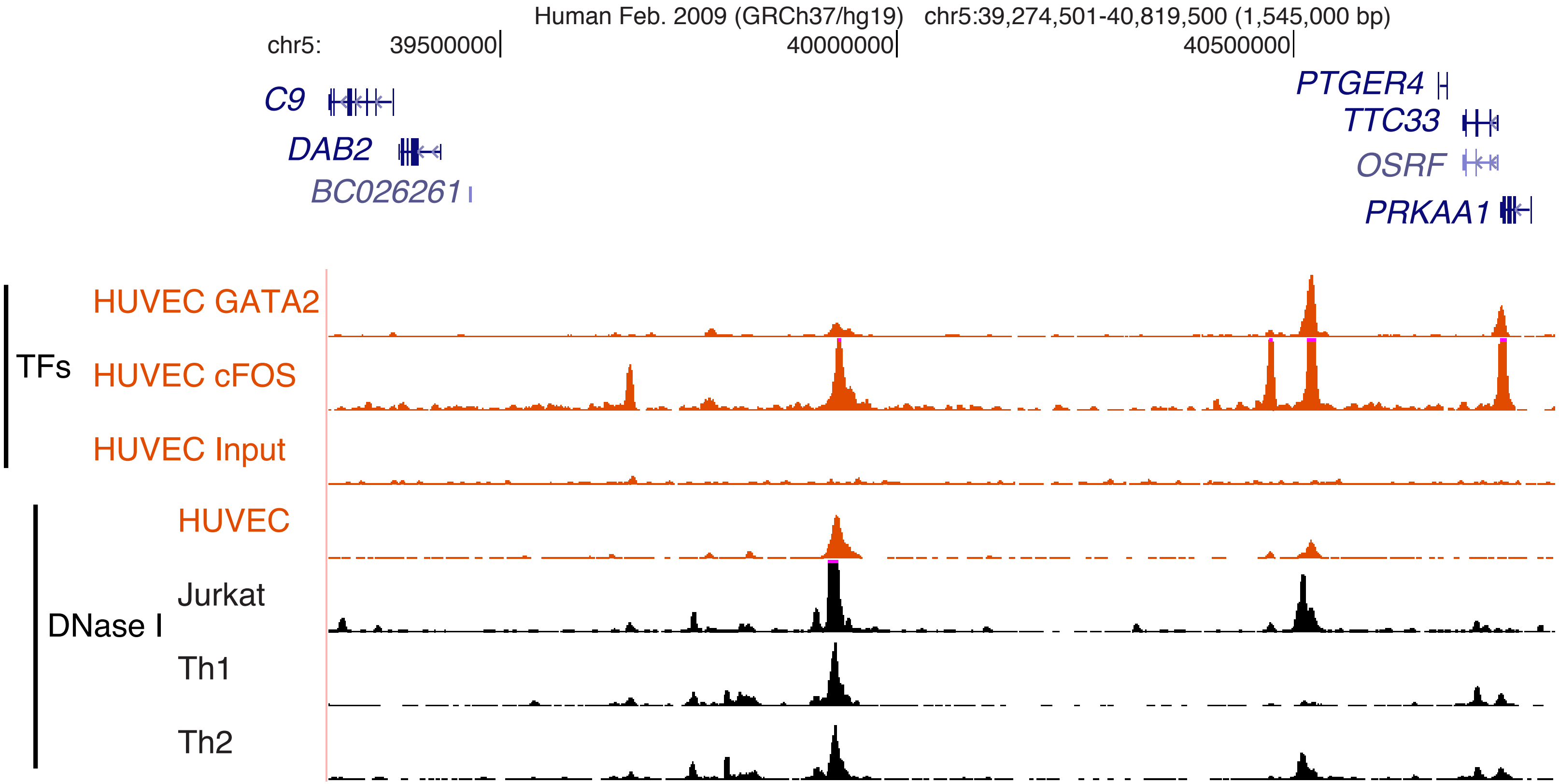
Region of open chromatin

Histone

Histone

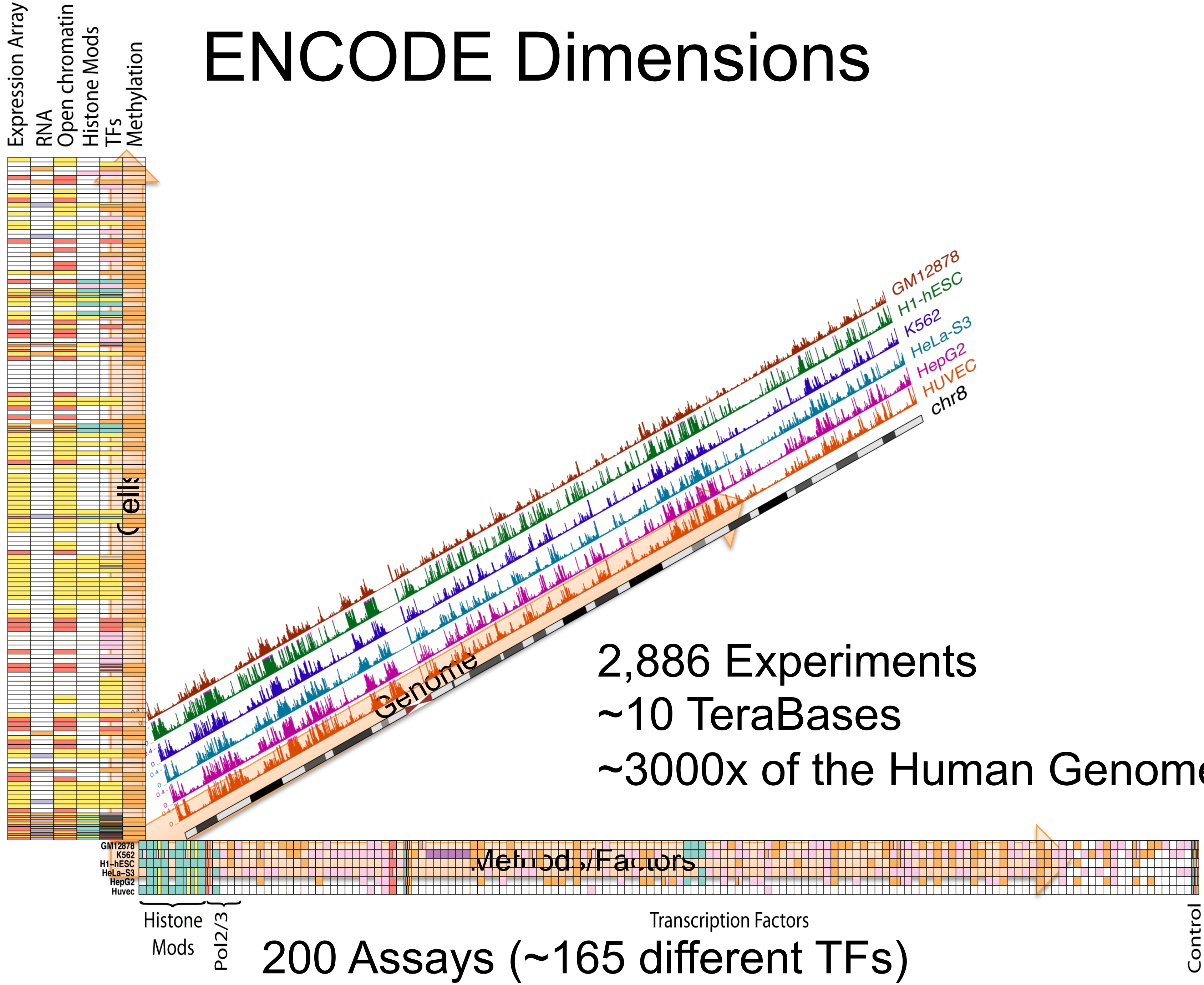
Ross Hardison, Belinda Giardine

Examples of Signal Tracks



ENCODE Dimensions

200 Cell Lines/ Tissues



2,886 Experiments

~10 TeraBases

~3000x of the Human Genome

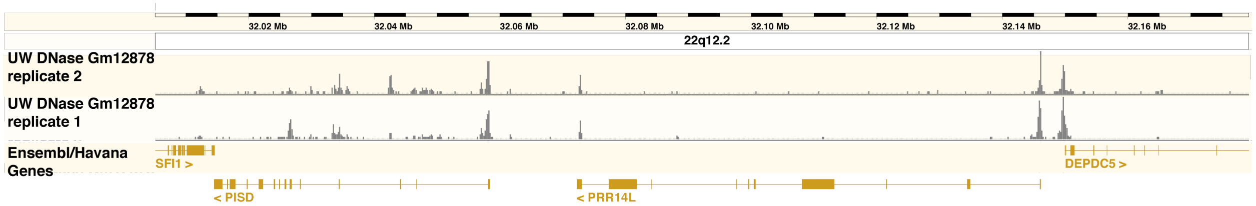
200 Assays (~165 different TFs)

Mouse Datasets Produced and Released

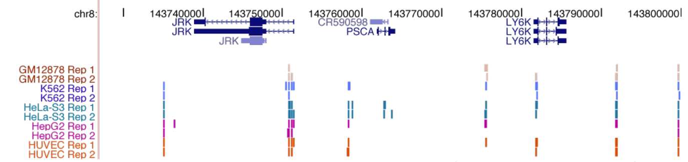
	# tissue or cell types	# experiments	# of data sets
Histone Modifications	33	157	310
Transcription Factor	29	109	299
RNA-Seq	69	104	193
DNase-Seq	55	55	127
Replication Timing	18	18	33
ChIP Controls	34	36	108
Total	**123**	**479**	**1070**

ENCODE Uniform Analysis Pipeline

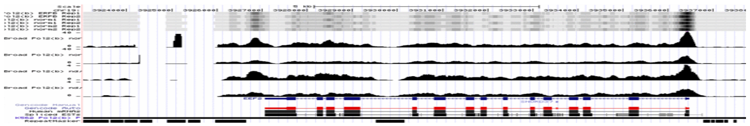
Anshul Kundaje, Qunhua Li, Michael Hoffman, Jason Ernst, Joel Rozowsky, Pouya Kheradpour



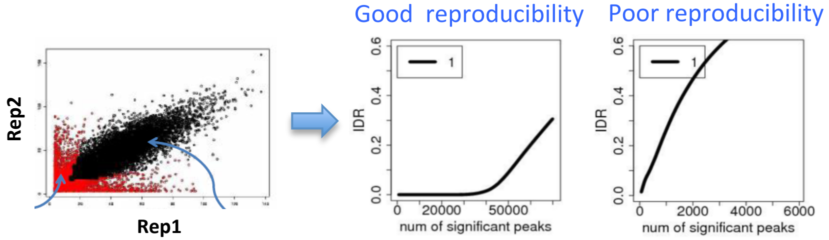
Mapped reads from production (Bam)



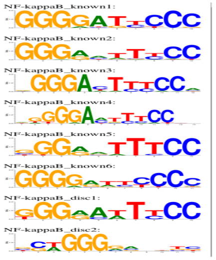
Uniform Peak Calling Pipeline (SPP, PeakSeq)



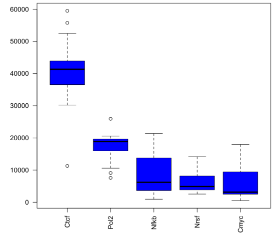
Signal Generation
(read extension and mappability correction)



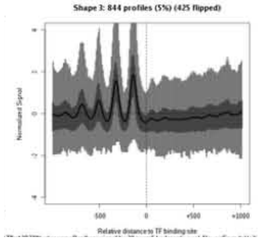
IDR Processing, QC and Blacklist Filtering



Motif Discovery

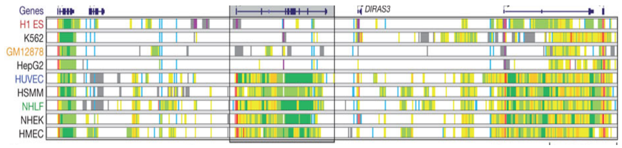


Stats, GSC enrichments, etc.

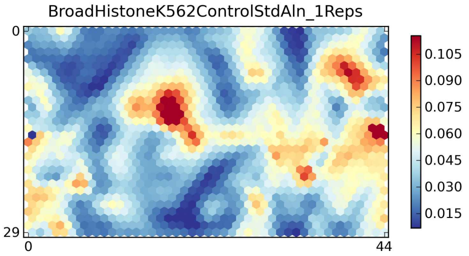


Signal Aggregation over peaks

Segmentation



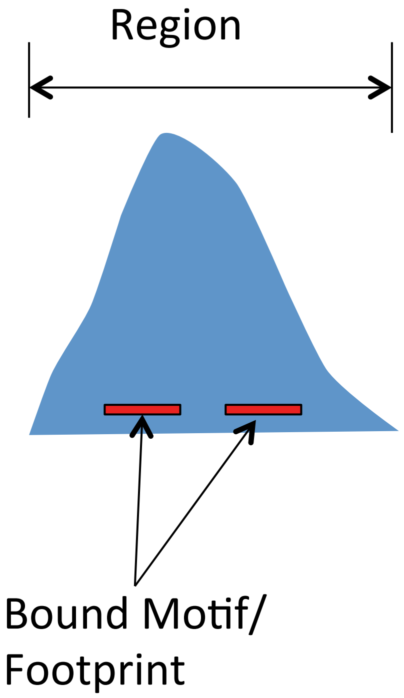
ChromHMM/Segway



Self Organising Maps

Raw genome coverage of elements

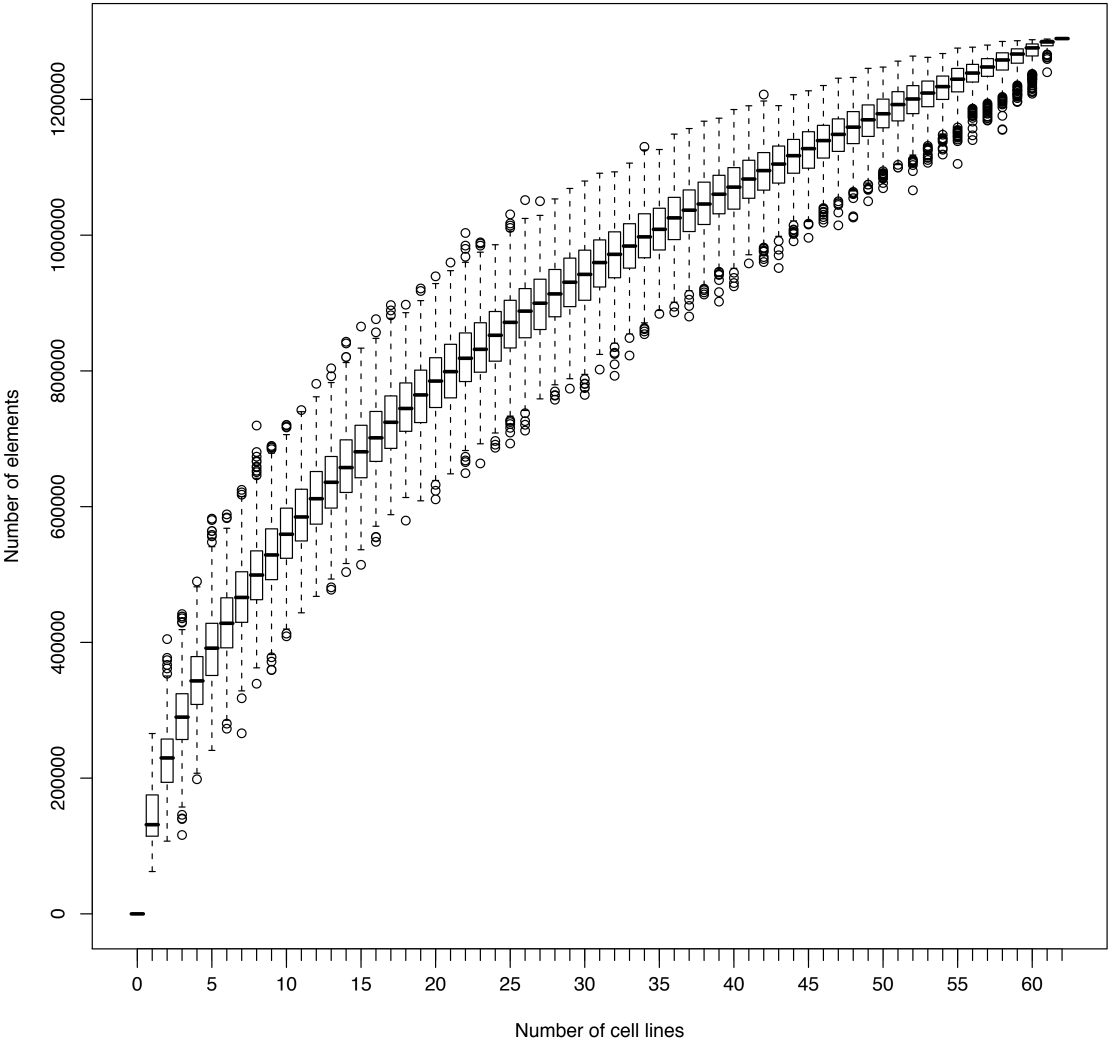
Element Type	Coverage	Cumulative Coverage
Exons	3%	3%
Chip-seq bound motifs	4.5%	5%
DNaseI Footprints	5.7%	9%
Chip-seq bound regions	8.1%	12%
DNaseI HS regions	15.2%	19.4%
Histone Modifications (*)	44%	49%
RNA	62%	80%
(* excluding broad marks)		



(Union over all experiments and cell types)

Saturation

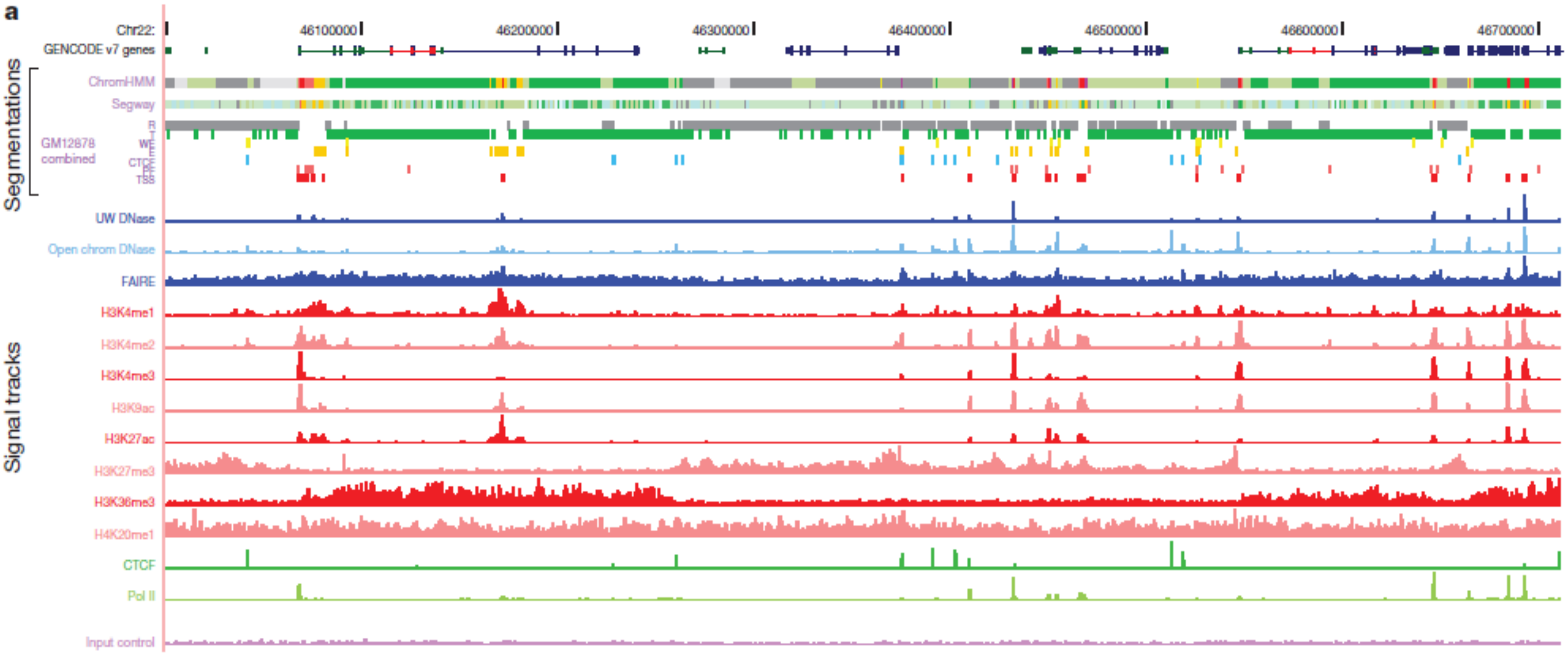
Steve Wilder



Most aggressive fit for saturation suggests a maximum of 50% of elements discovered

Likely to be lower due to inaccessible cell types etc

ENCODE Integrative Segmentations



~7 Major genome segments
25 “elaborations”
1,000s of details

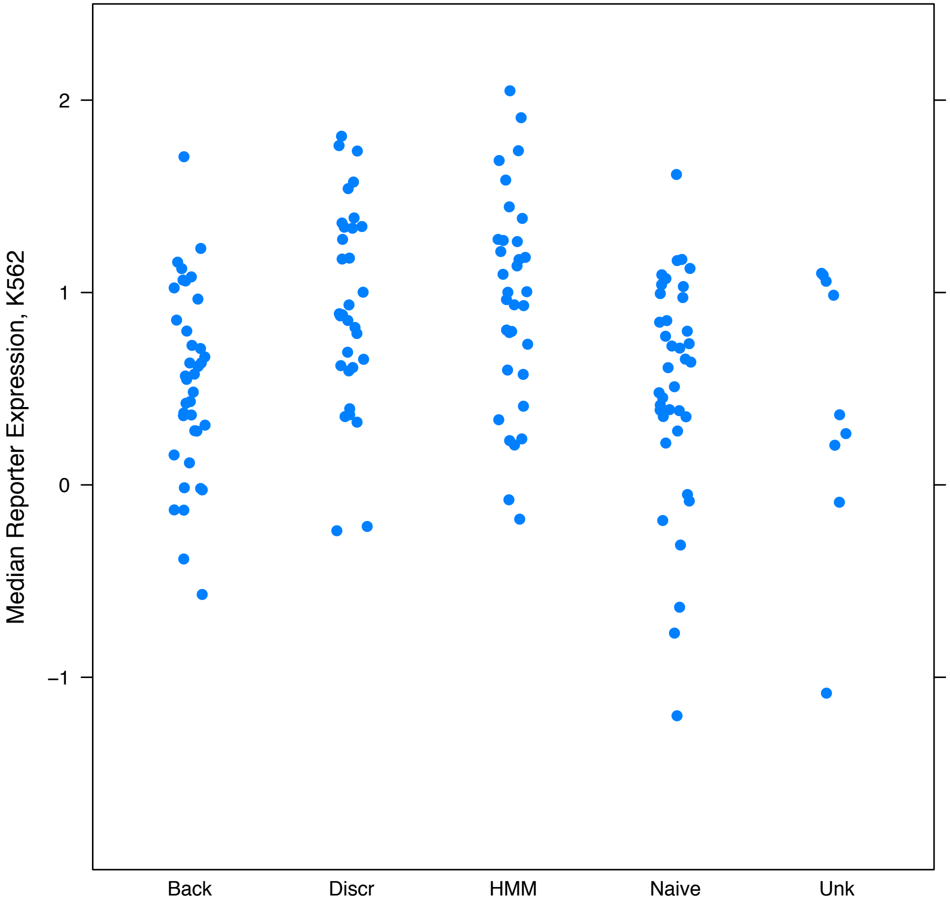
Well Known: TSS, Gene Start, Gene Bodies

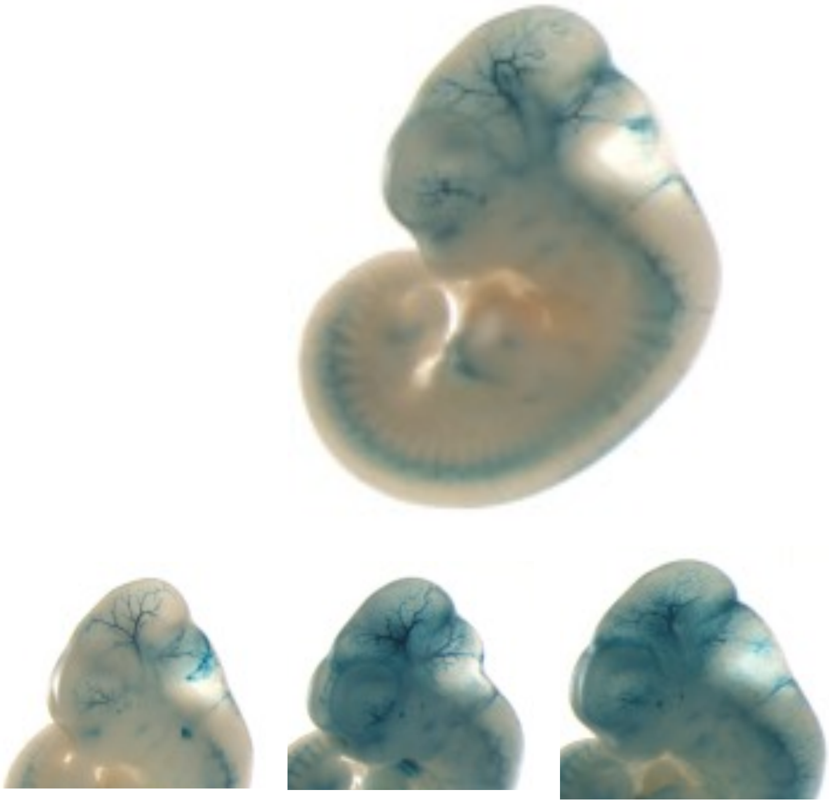
New Info: “Enhancers” (2 states), Insulators

Unexpected: Specific Gene End

Experimental Confirmation of New Enhancers

Jason Gertz, Barbara Wold, Rick Myers, Len Pennacchio

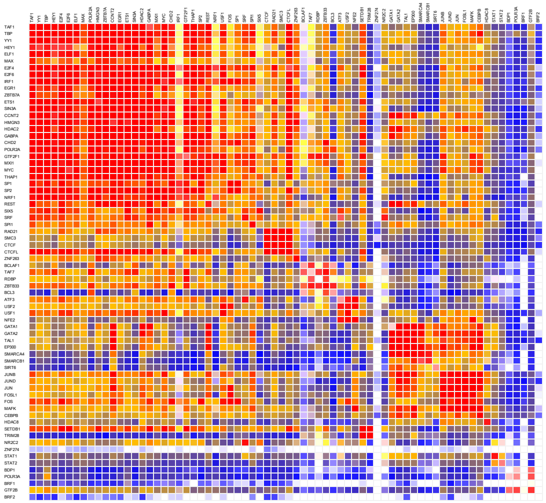




Mann Whitney 0.003 HMM vs Background
1e-7, HMM vs Naïve or Biologist picks
Myers Lab

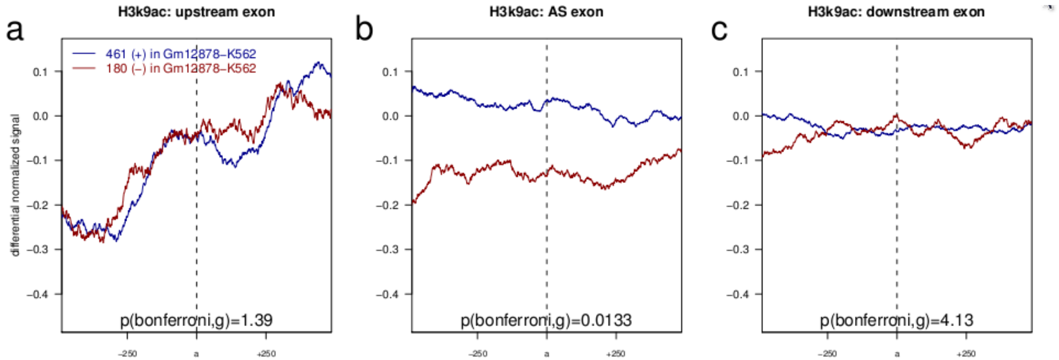
53% hit rate in Mouse Assay
Pennacchio Lab

Many other stories…

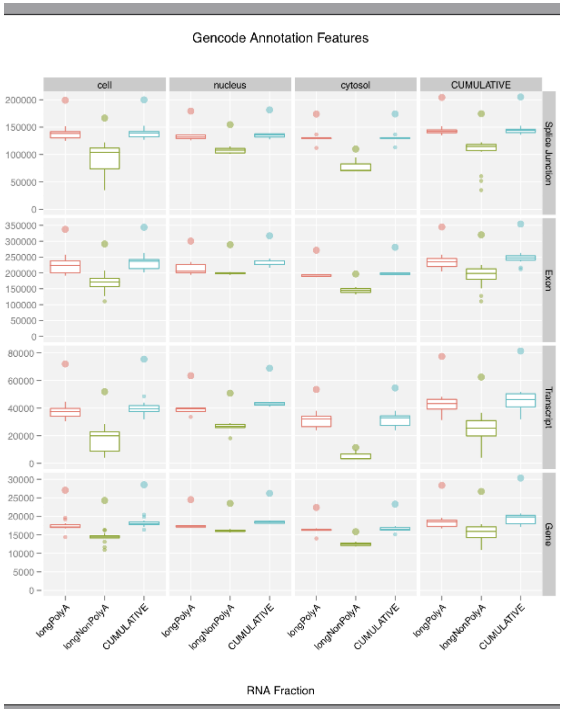

K562 Whole-genome

TF Co association and Regulatory Code
Mike Snyder+Mark Gerstein

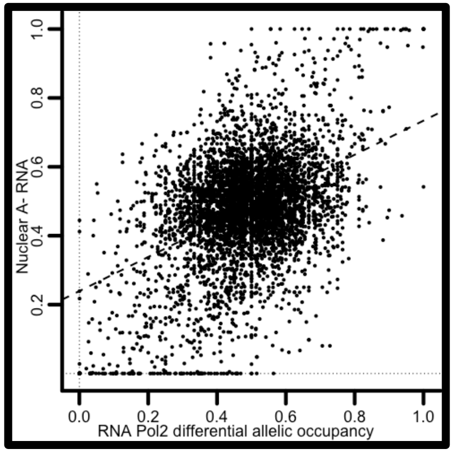
DNaseI footprints – John Stam.
DNA Methylation – Rick Myers



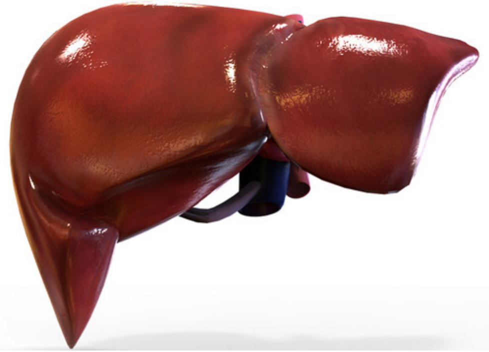
Splicing/Histone interaction (Roderic Guigo)



RNA landscape
Tom Gingeras



What's Next- Enhancing ENCODE



1) Deep analysis of six cell lines/tissues
    - Cancer: K562, MCF7
    - Diploid: GM12878, H1 ES Cells
    - Tissues: Liver, Brain
2) More limited coverage of hundreds of other cell types and tissues (RNA-Seq, DNAaseHS, etc)
3) Some mouse data

Many investigators same as ENCODE2 + Brenton Graveley-a few others

What’s Next- Species Comparisons

1) modENCODE/ENCODE Comparison

   Worms, Flies, Human

   Hundreds of worm and fly datasets (e.g. >250 C. elegans TF ChIP-Seq datasets)

2) MouseENCODE-humanENCODE Comparison

The ENCODE Consortium

Brad Bernstein (Eric Lander, Manolis Kellis, Tony Kouzarides)

Ewan Birney (Jim Kent, Mark Gerstein, Bill Noble, Peter Bickel, Ross Hardison, Zhiping Weng)

Greg Crawford (Ewan Birney, Jason Lieb, Terry Furey, Vishy Iyer)

Jim Kent (David Haussler, Kate Rosenbloom)

John Stamatoyannopoulos (Evan Eichler, George Stamatoyannopoulos, Job Dekker, Maynard Olson, Michael Dorschner, Patrick Navas, Phil Green)

Mike Snyder (Kevin Struhl, Mark Gerstein, Peggy Farnham, Sherman Weissman)

Rick Myers (Barbara Wold)

Scott Tenenbaum (Luiz Penalva)

Tim Hubbard (Alexandre Reymond, Alfonso Valencia, David Haussler, Ewan Birney, Jim Kent, Manolis Kellis, Mark Gerstein, Michael Brent, Roderic Guigo)

Tom Gingeras (Alexandre Reymond, David Spector, Greg Hannon, Michael Brent, Roderic Guigo, Stylianos Antonarakis, Yijun Ruan, Yoshihide Hayashizaki)

Zhiping Weng (Nathan Trinklein, Rick Myers)

Additional ENCODE Participants: Elliott Marguiles, Eric Green, Job Dekker, Laura Elnitski, Len Pennachio, Jochen Wittbrodt

.. and many senior scientists, postdocs, students, technicians, computer scientists, statisticians and administrators in these groups

NHGRI: Elise Feingold, Mike Pazin, Peter Good