ENCODE: Understanding the Genome

Michael Snyder

November 6, 2012

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*](http://encodeproject.org)

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)

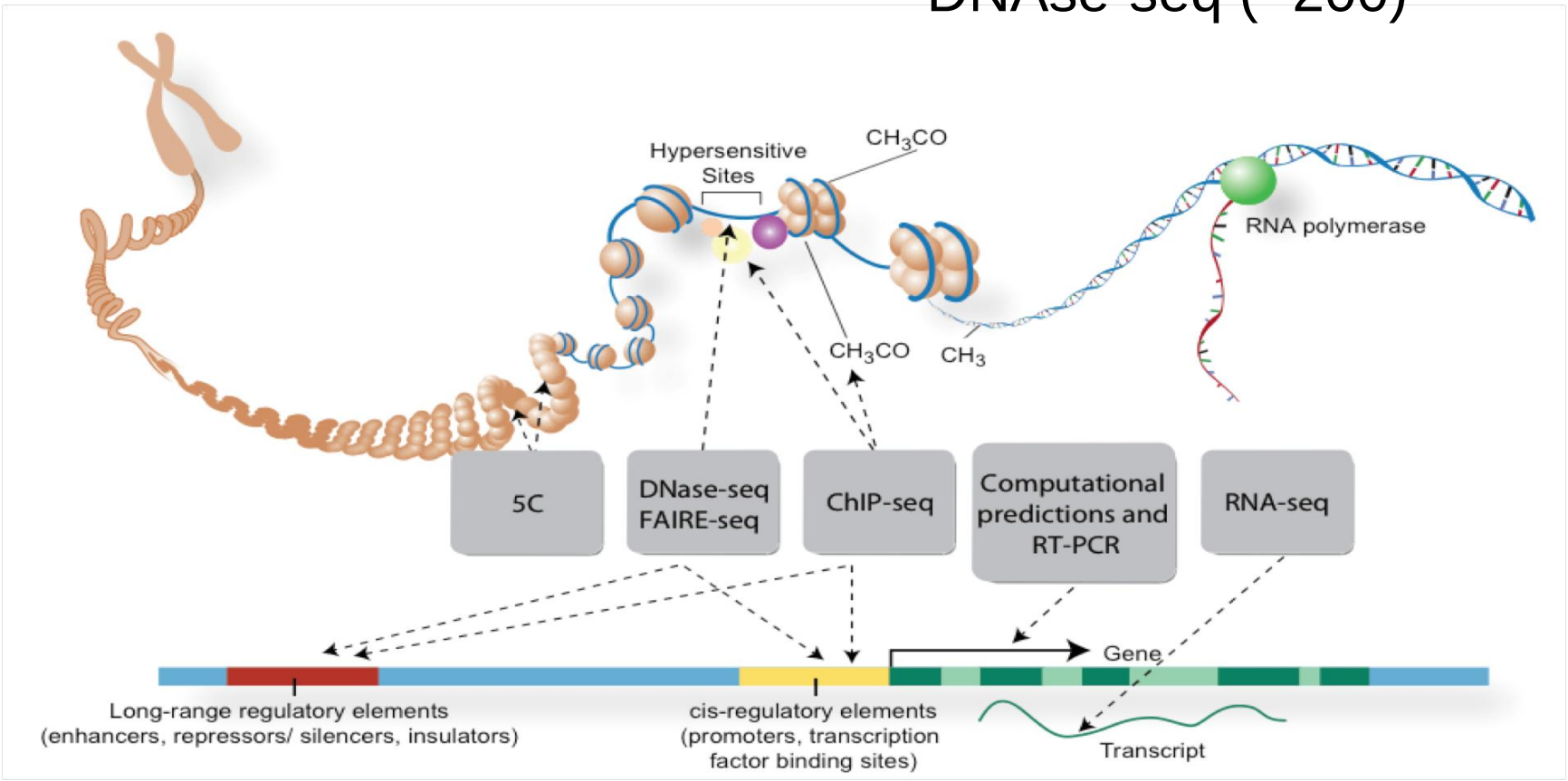
Additional ENCODE Participants: Elliott Marguiles, Eric Green, Job Dekker, Laura Elnitski, Len Pennachio, Jochen Wittbrodt

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

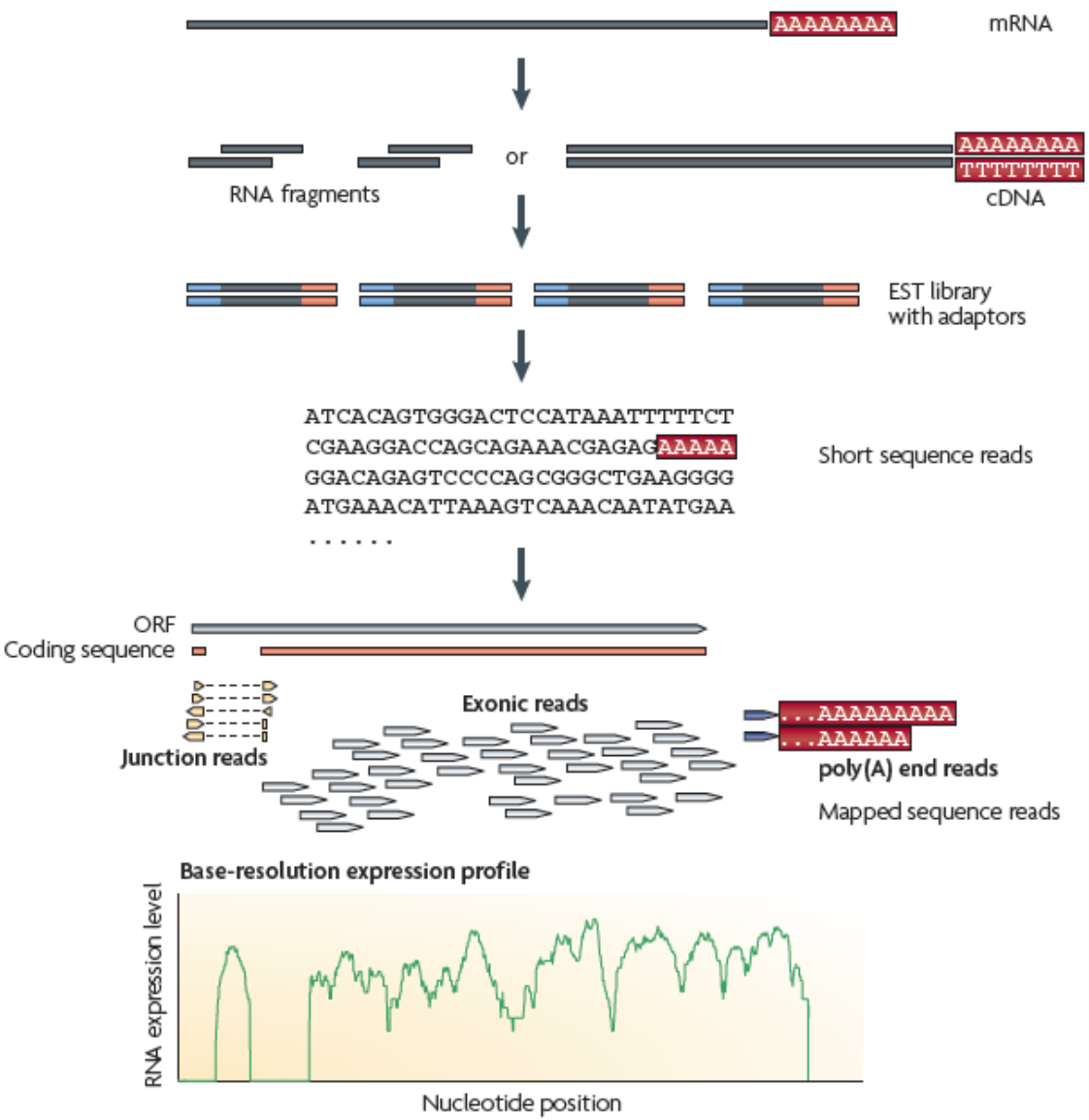
NHGRI: Elise Feingold, Mike Pazin, Peter Good

Experimental Assays

Chip-seq (165 TFs
+ Histone marks)
RNA-seq (292)
DNAse-seq (~200)

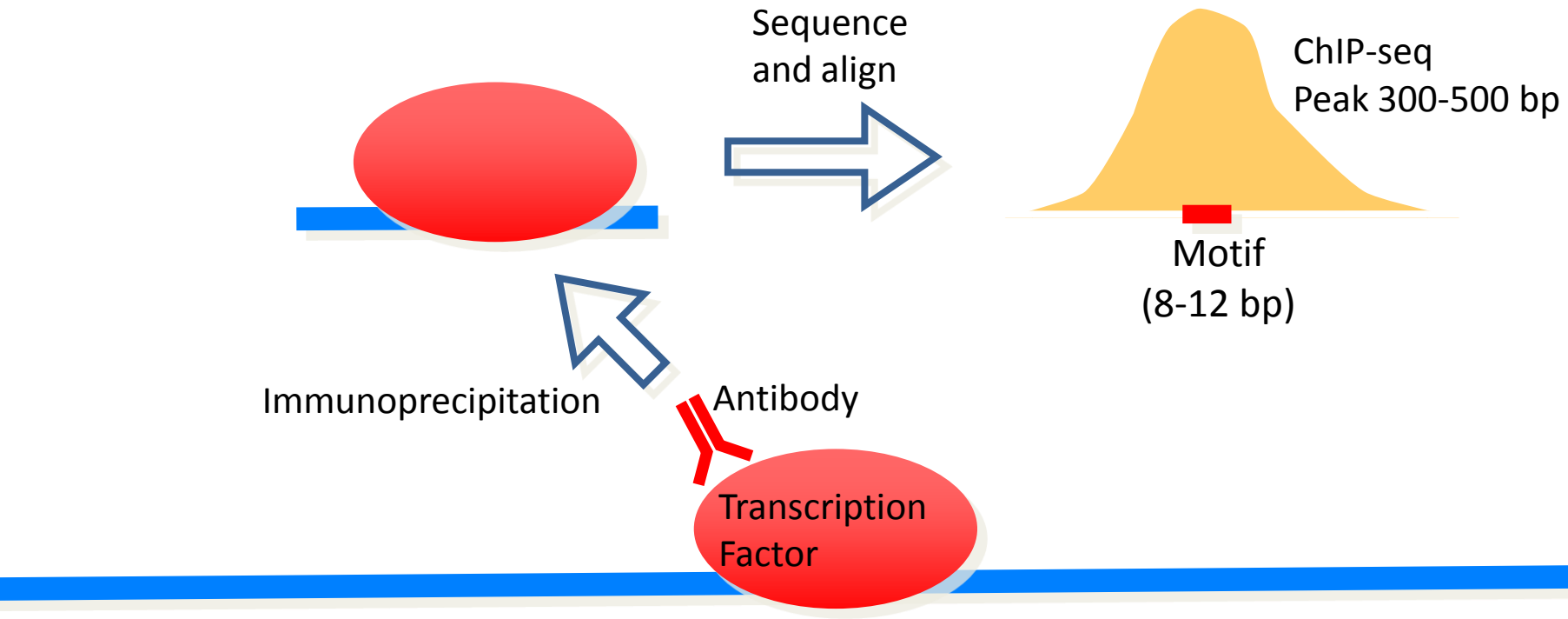


RNA-Sequencing



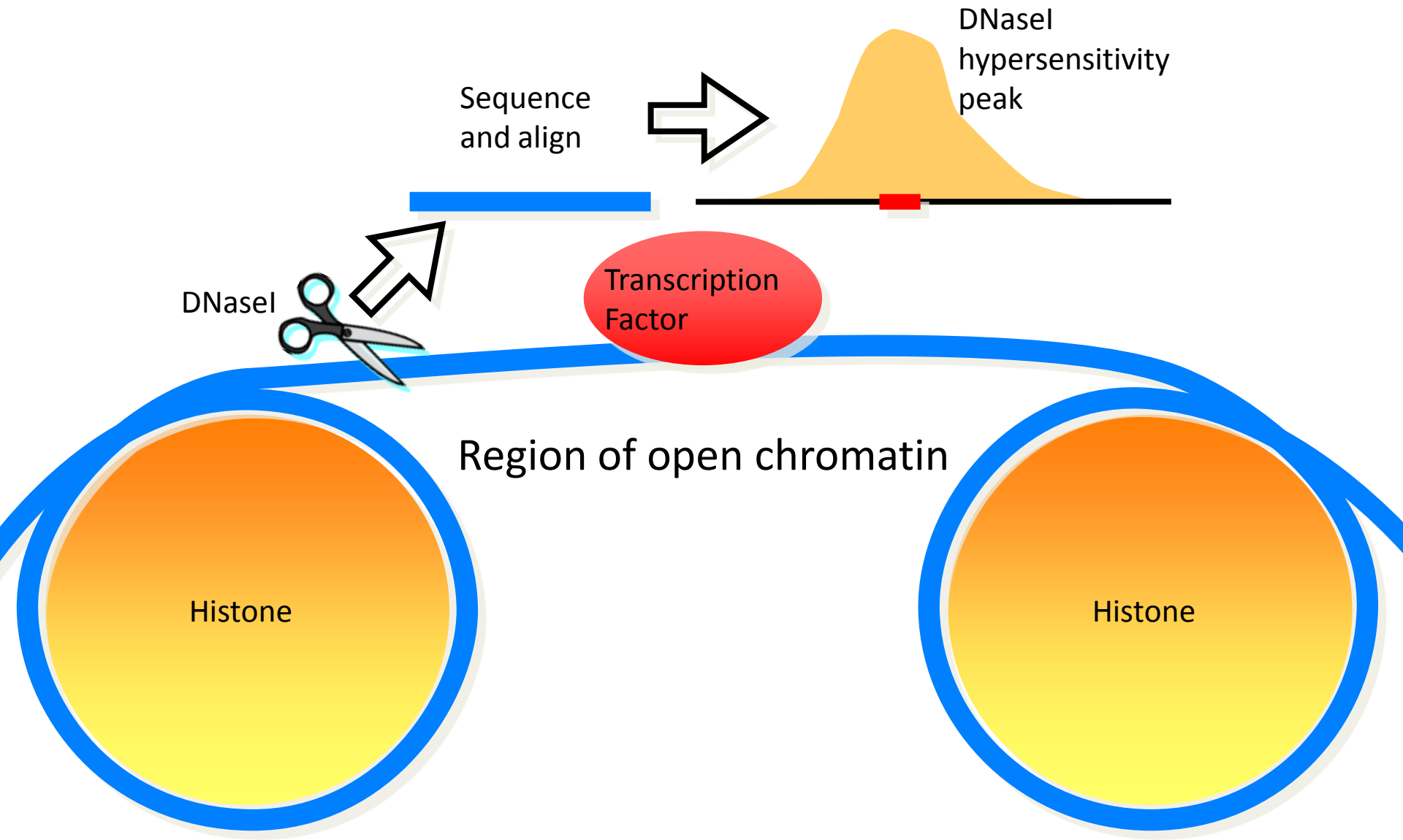
Wang et al. 2009 Nat Gen. Rev.

Functional data: ChIP-seq

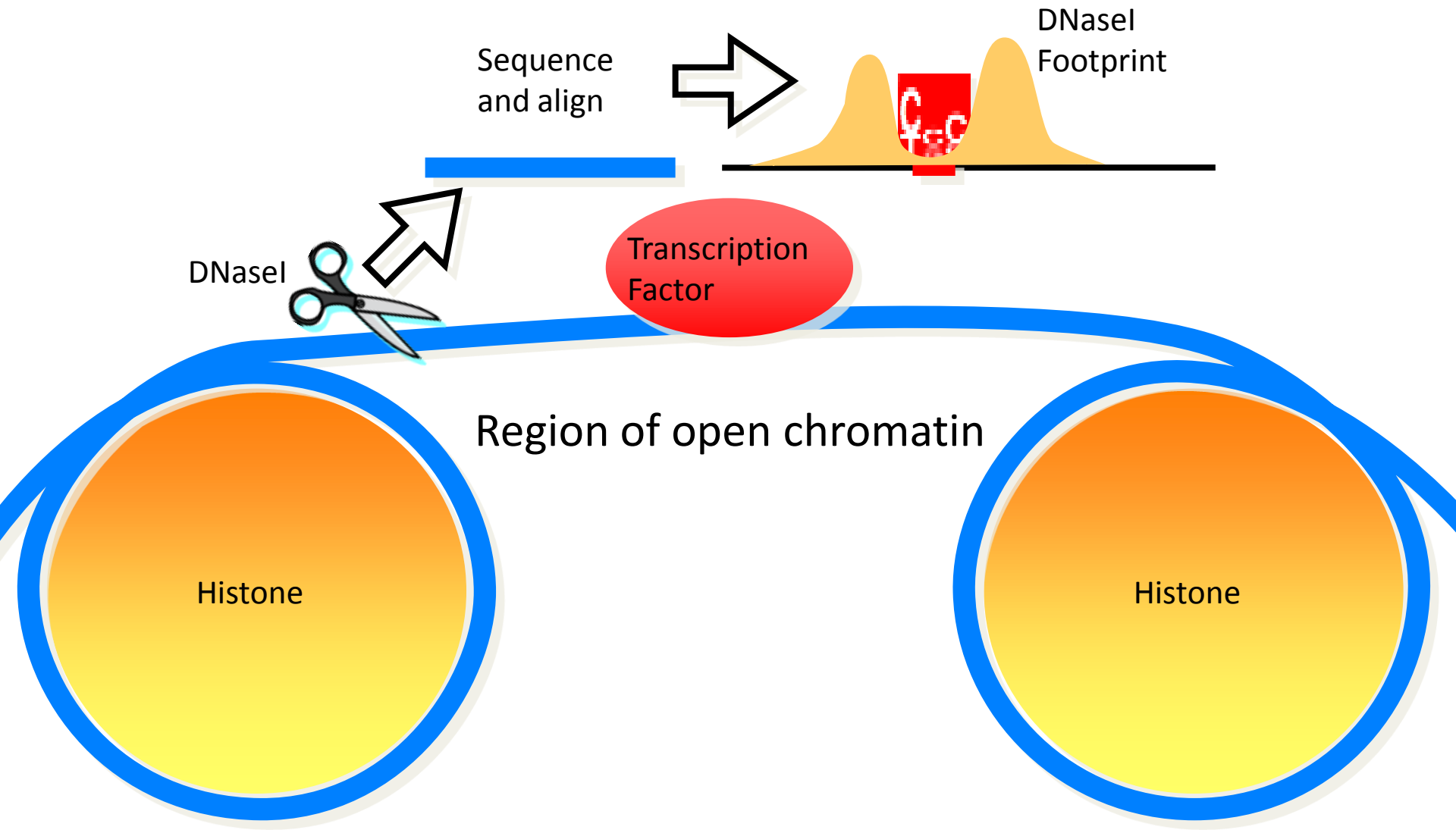


ChIP-exo
Histone Marks

Functional data: DNase-seq

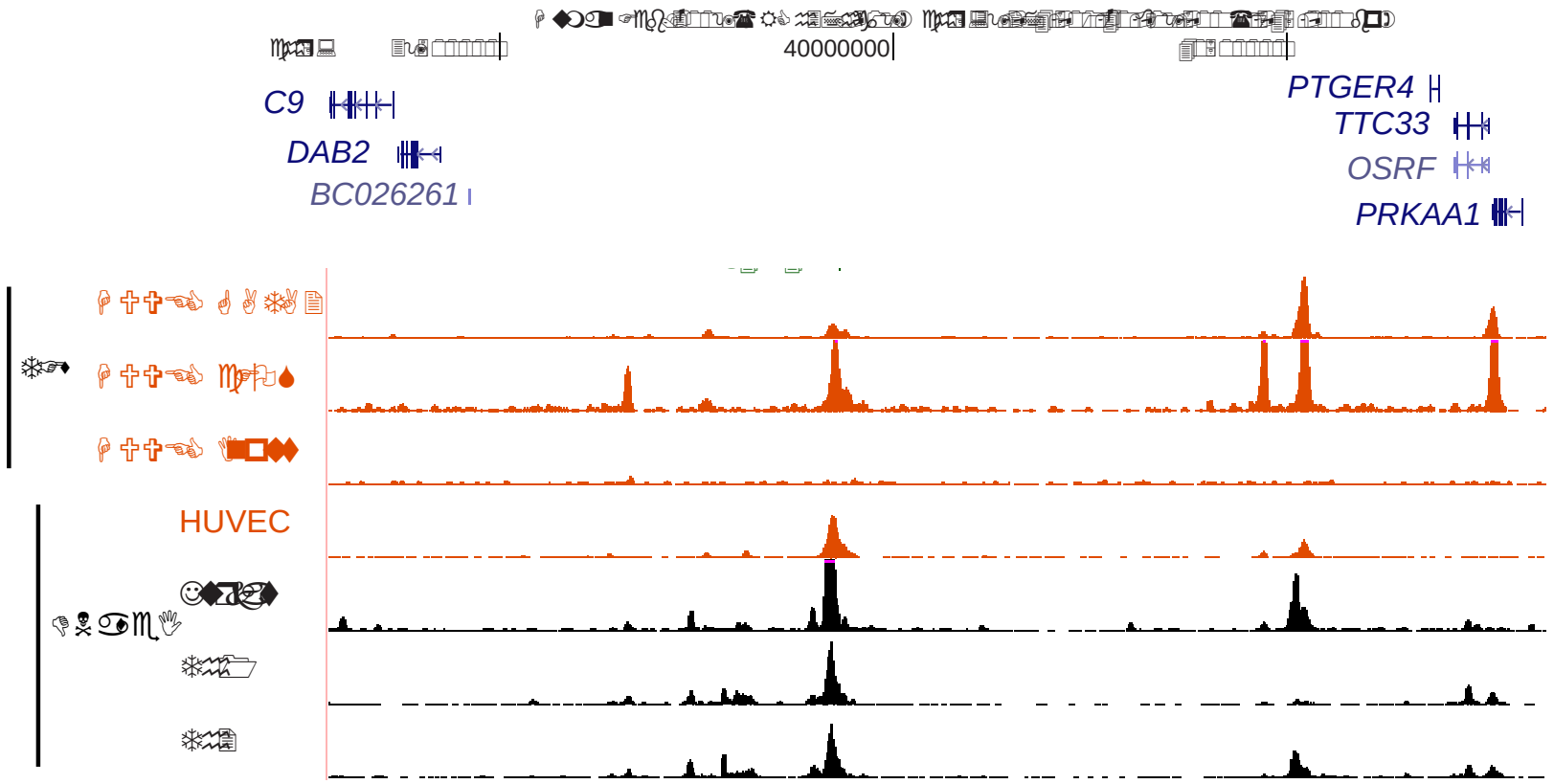


Functional data: DNase footprints



Ross Hardison, Belinda Giardine

Examples of Signal Tracks

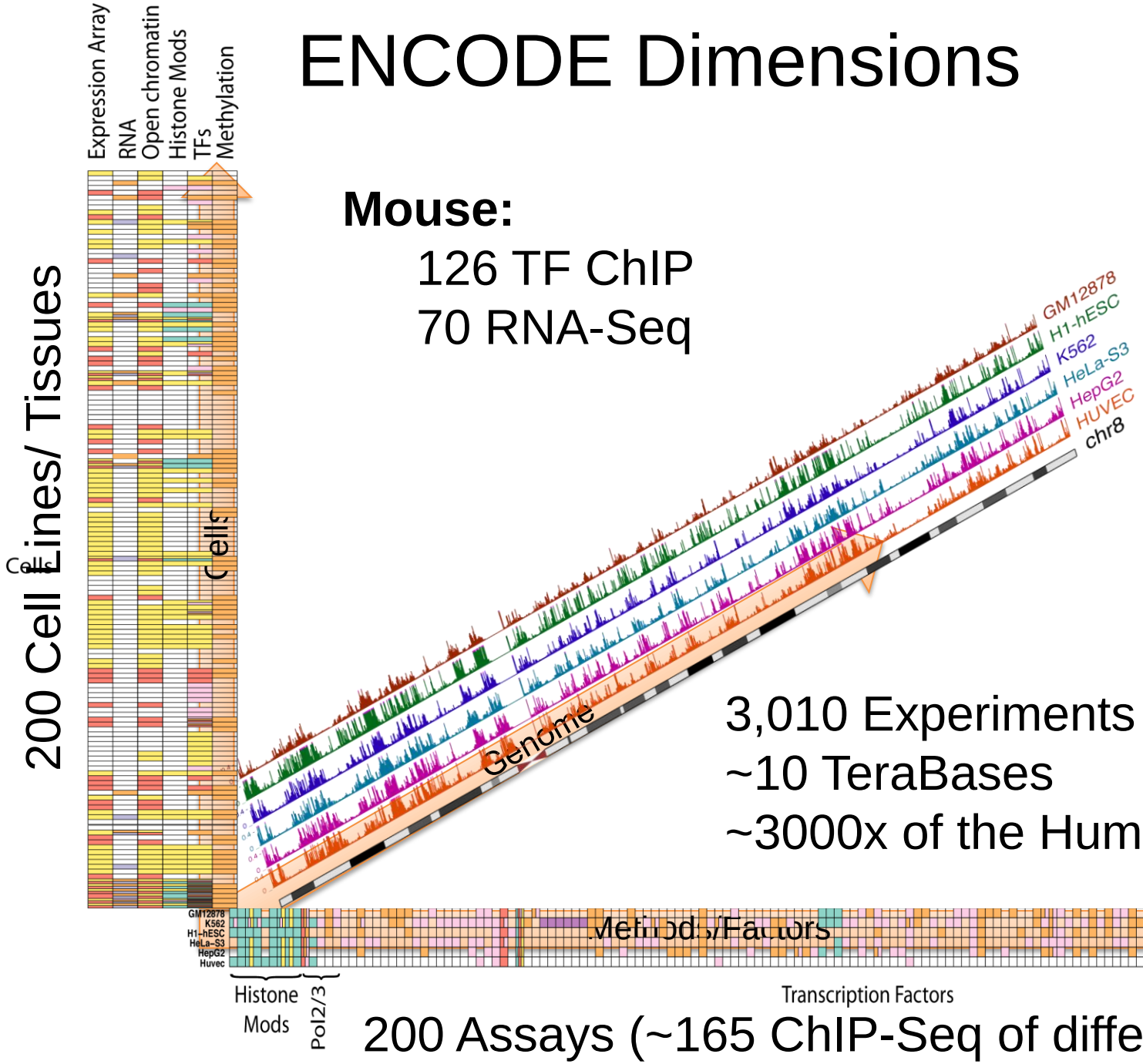


ENCODE Dimensions

Mouse:

126 TF ChIP

70 RNA-Seq



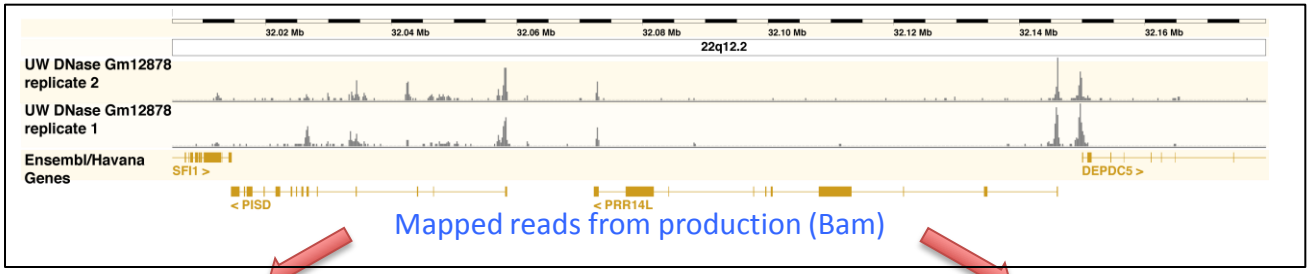
3,010 Experiments

~10 TeraBases

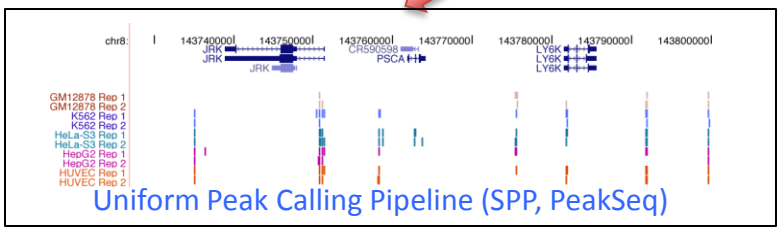
~3000x of the Human Genome

ENCODE Uniform Analysis Pipeline

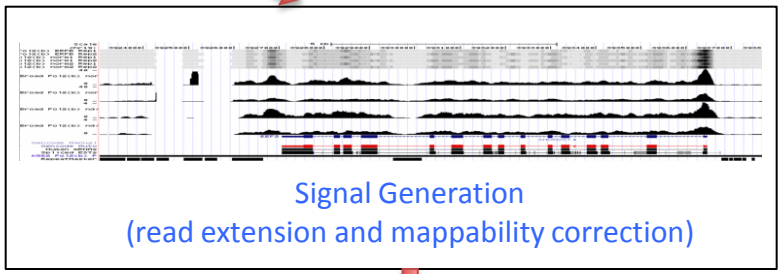
Anshul Kundaje, Qunhua Li, Michael Hoffman, Jason Ernst, Joel Rozowsky, Pouya Kheradpour



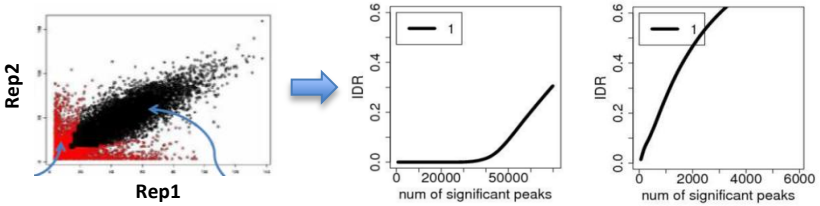
Mapped reads from production (Bam)



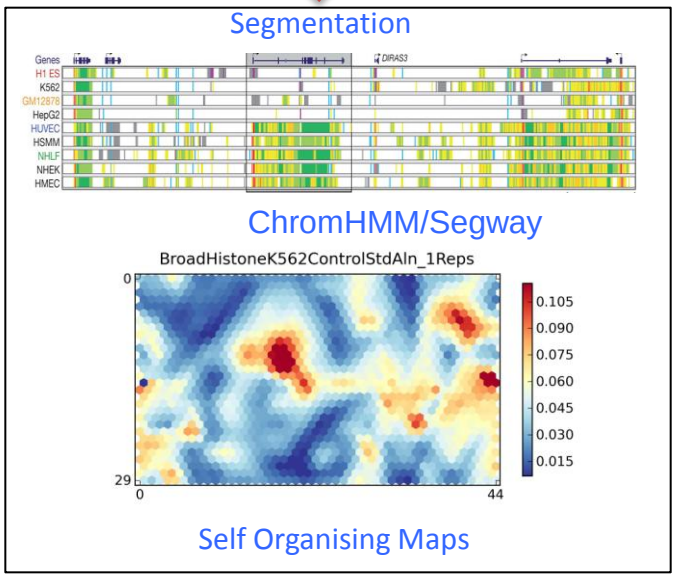
Uniform Peak Calling Pipeline (SPP, PeakSeq)

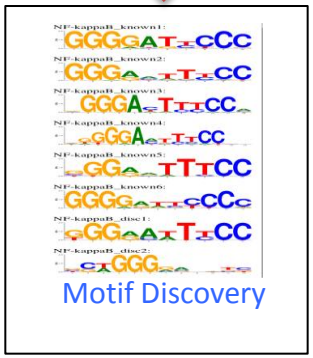


Signal Generation
(read extension and mappability correction)

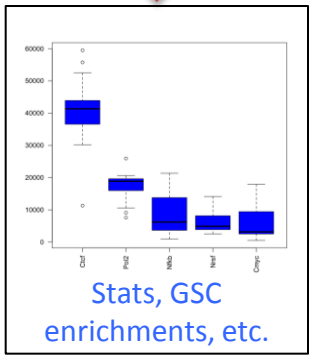


IDR Processing, QC and Blacklist Filtering

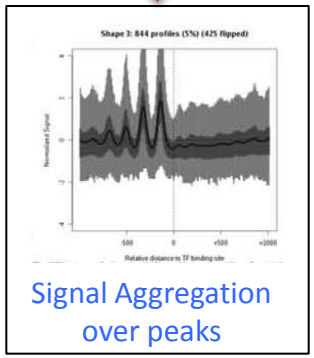




Motif Discovery



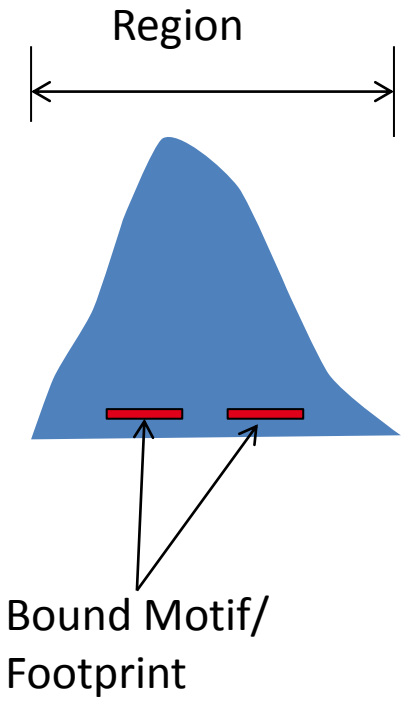
Stats, GSC enrichments, etc.



Signal Aggregation over peaks

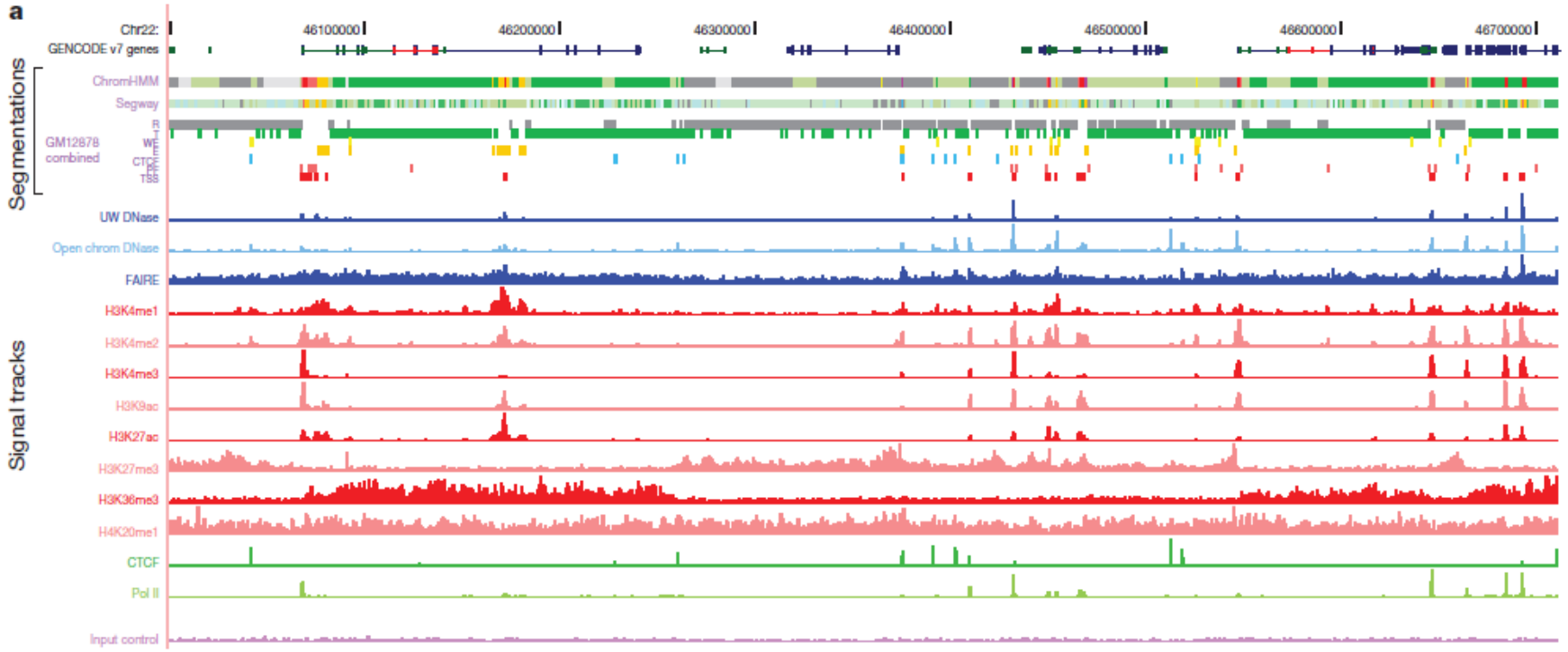
Raw genome coverage of elements

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)

ENCODE Integrative Segmentations



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

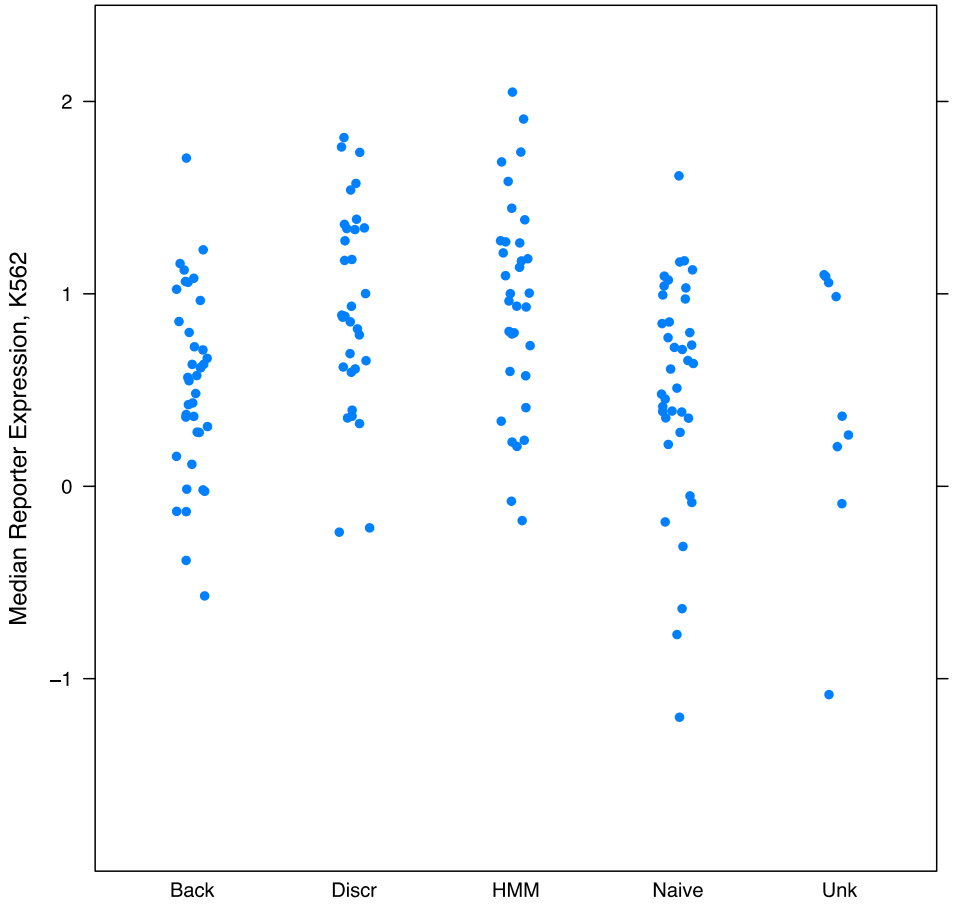
Well Known: TSS, Gene Start, Gene Bodies

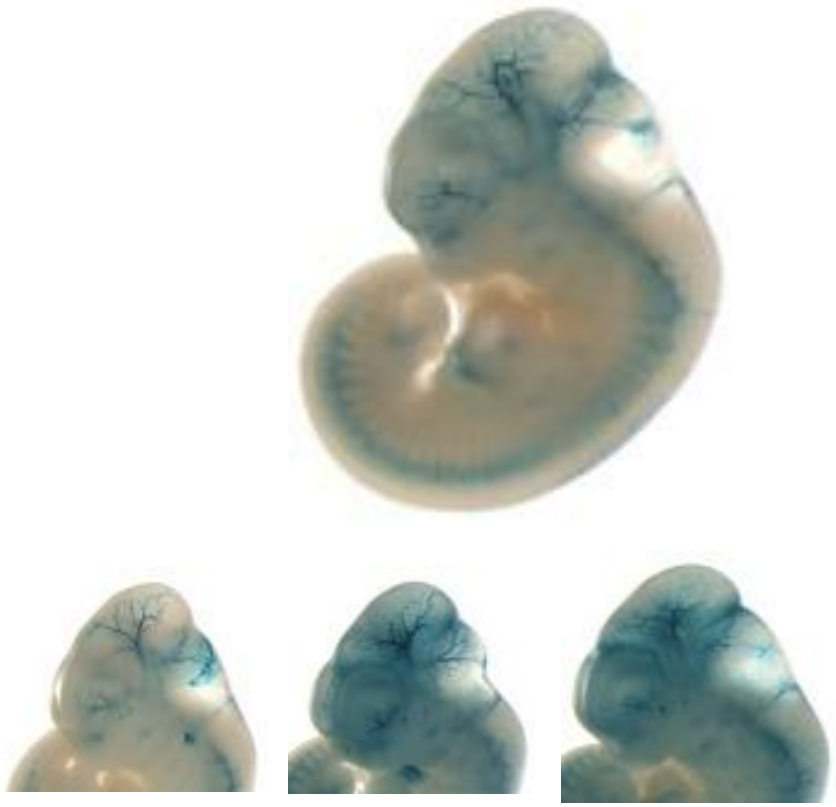
New Info: “Enhancers” (2 states), Insulators

Unexpected: Specific Gene End

Experimental Confirmation of New Enhancers

Jason Gertz, Barbara Wold, Rick Myers, Len Pennacchio

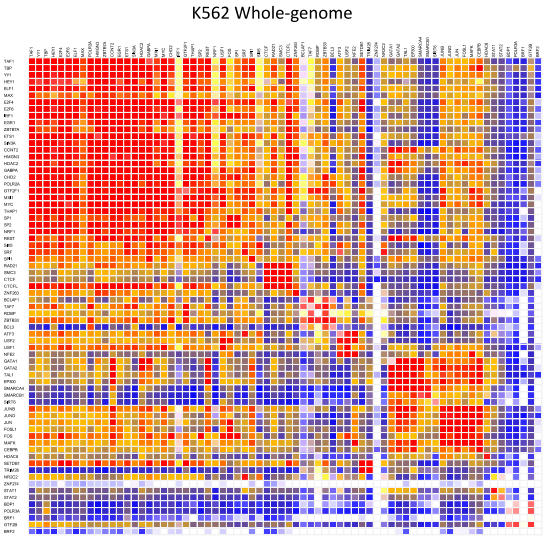




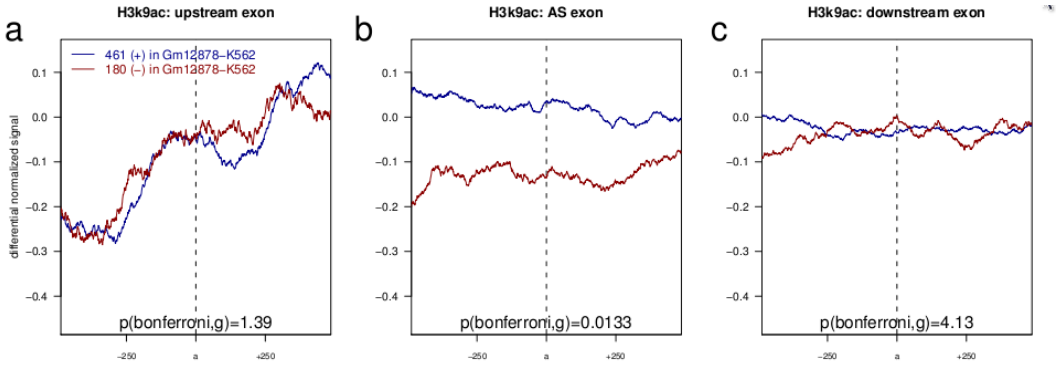
Mann Whitney 0.003 HMM vs Background
1e-7, HMM vs Naïve or Biologist picks
Myers Lab

53% hit rate in Mouse Assay
Pennacchio Lab

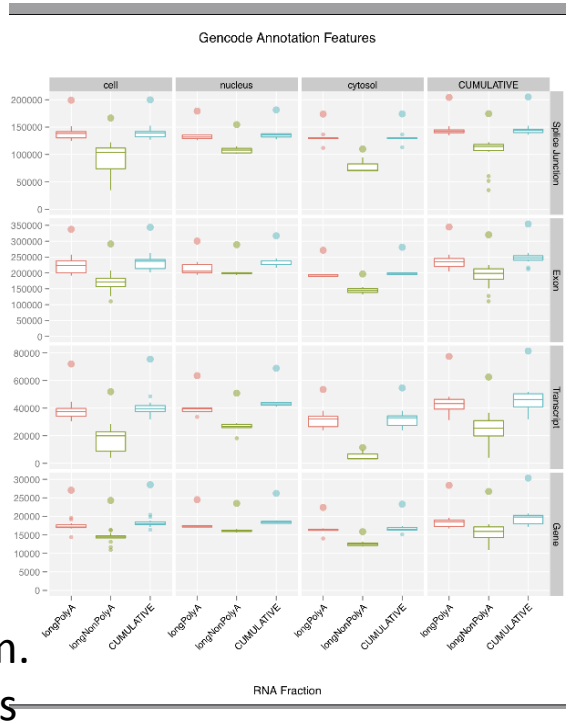
Many



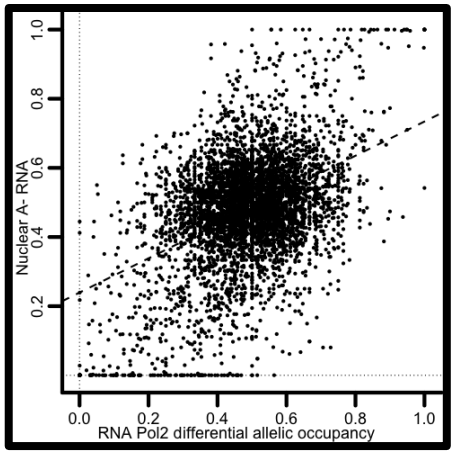
TF Co association and Regulatory Code
Mike Snyder+Mark Gerstein



Splicing/Histone interaction (Roderic Guigo)



RNA landscape
Tom Gingeras



DNAseI footprints – John Stam.
DNA Methylation – Rick Myers

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)

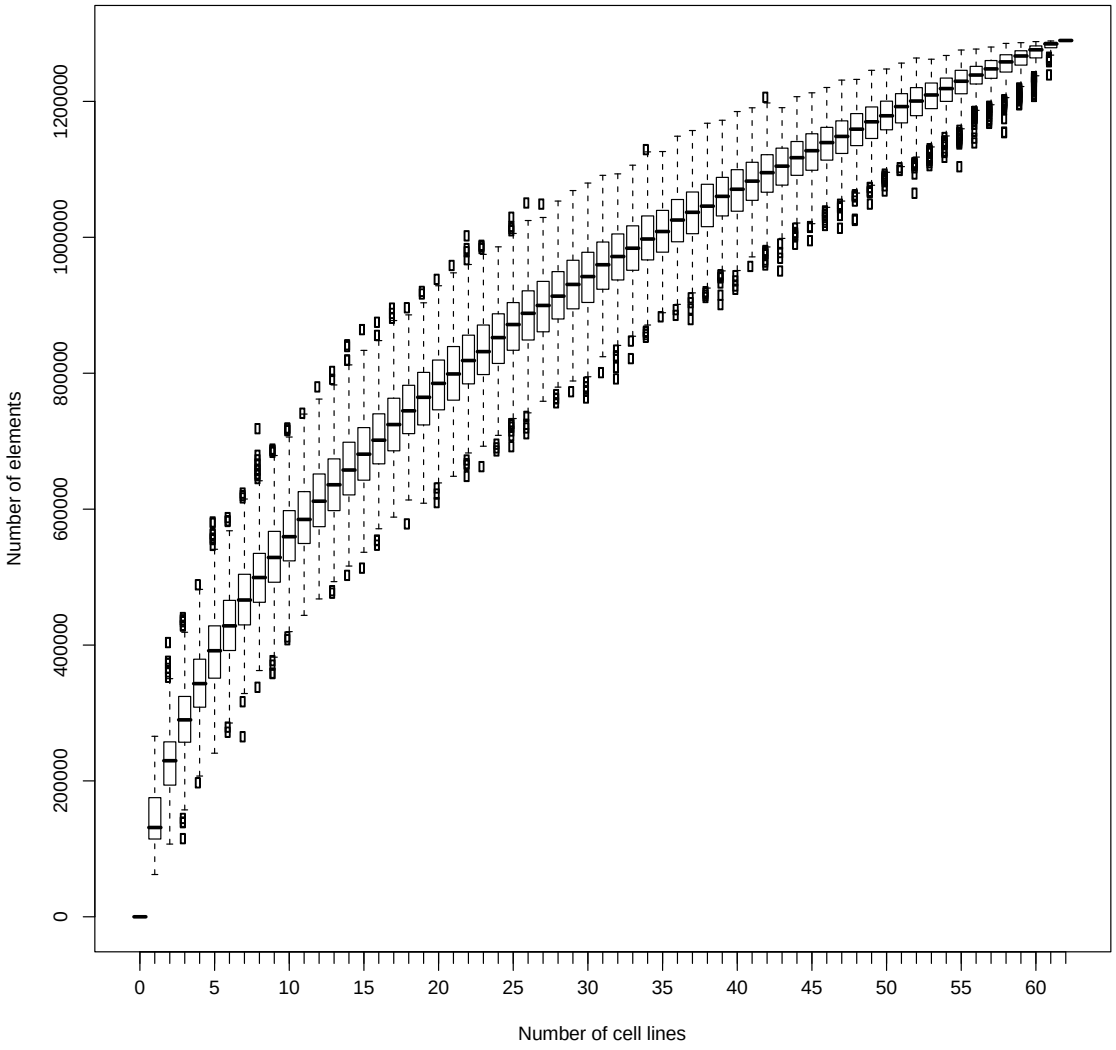
Additional ENCODE Participants: Elliott Marguiles, Eric Green, Job Dekker, Laura Elnitski, Len Pennachio, Jochen Wittbrodt

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

NHGRI: Elise Feingold, Mike Pazin, Peter Good

Saturation

Steve Wilder

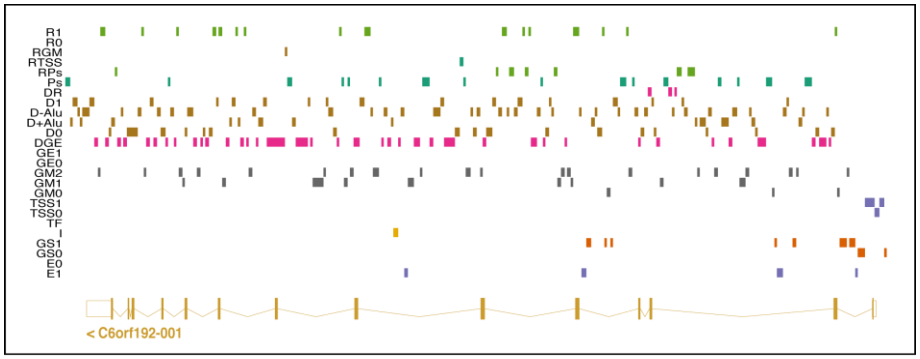


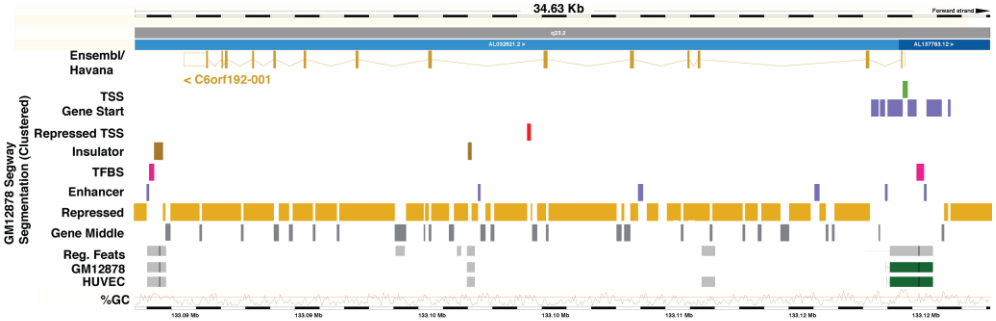
Most aggressive fit for saturation suggests a maximum of 50% of elements discovered

Likely to be lower due to inaccessible cell types etc

Discovering functional genome segments

Michael Hoffman, Jason Ernst, Bill Noble, Manolis Kellis





Well understood:
TSS, Gene Start,
Gene Bodies

Reassuringly Interesting
“Enhancers” (2 states)
Insulators

Definitely There, Unexpected
Specific Gene End

Sub-classification of Repeats

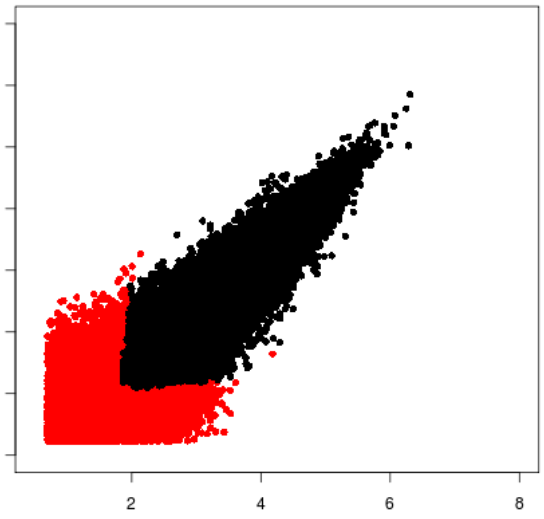
~7 Major segments of the genome
25 “elaborations”
1,000s of details

Irreproducible Discovery Rate (IDR)

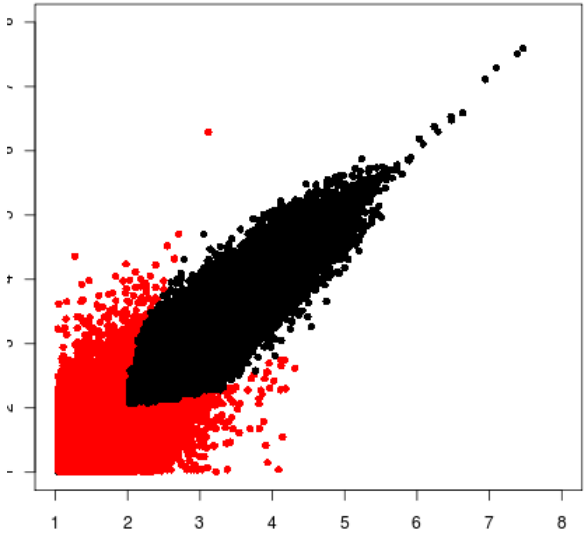
Ben Brown, Qunhau Li, Peter Bickel

If one re-ran the experiment, what is the probability one would observe the same element at this rank or better

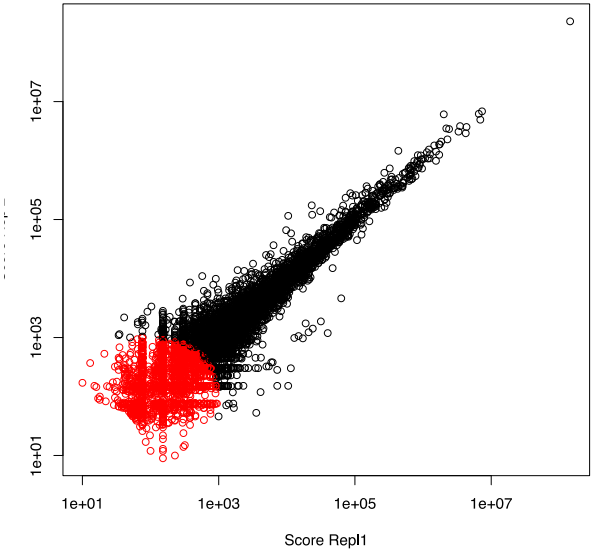
Uses ranked element lists from two replicates, and makes the assumption that there is noise at the bottom of the rank



Chip-seq



Dnase-seq



RNA-seq