# Deciphering human non-coding DNA using machine learning approaches

Guillaume Bourque

Department of Human Genetics, McGill University and McGill University and Genome Quebec Innovation Center

MLPM Summer School

Paris, September 17th 2014

McGill

GenomeQuébec

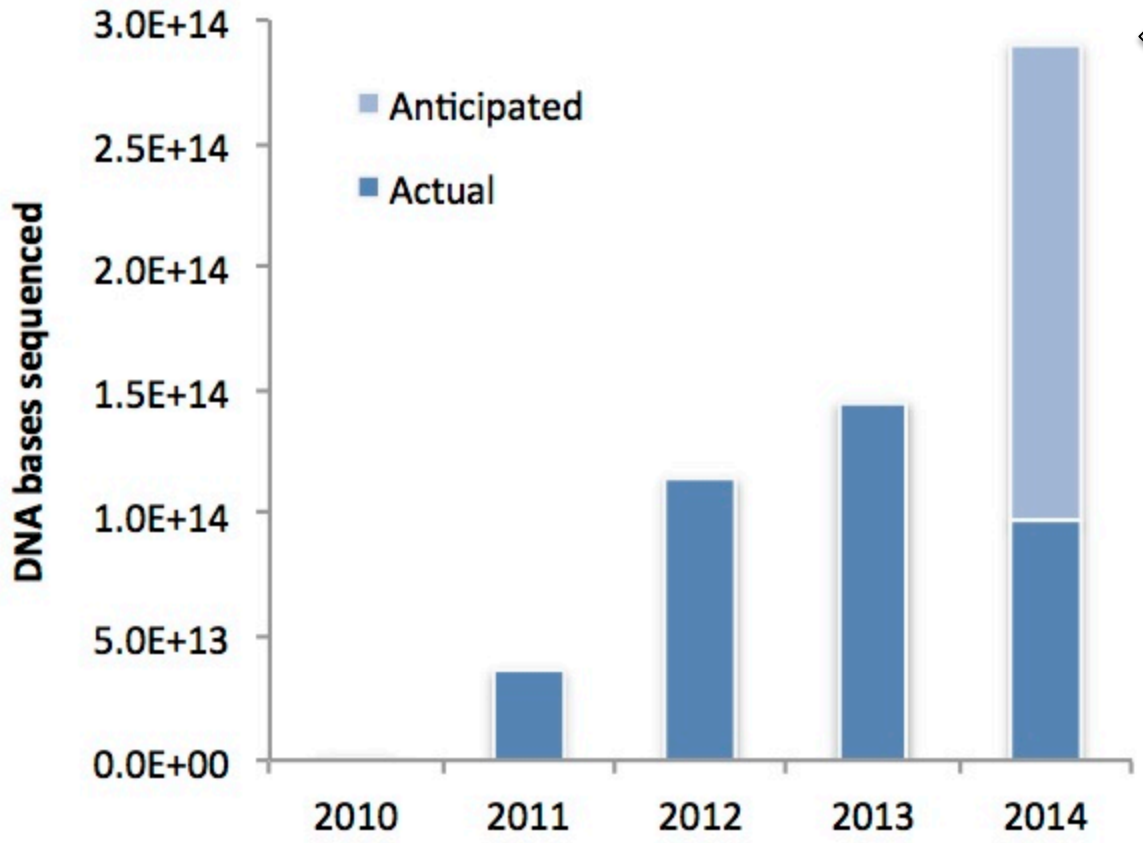
# Outline

- Applications of next-generation sequencing
- Functional genomics
- Example of machine learning approaches in functional genomics
- Role of transposable elements in gene regulation

# Outline

- Applications of next-generation sequencing
- Functional genomics
- Example of machine learning approaches in functional genomics
- Role of transposable elements in gene regulation

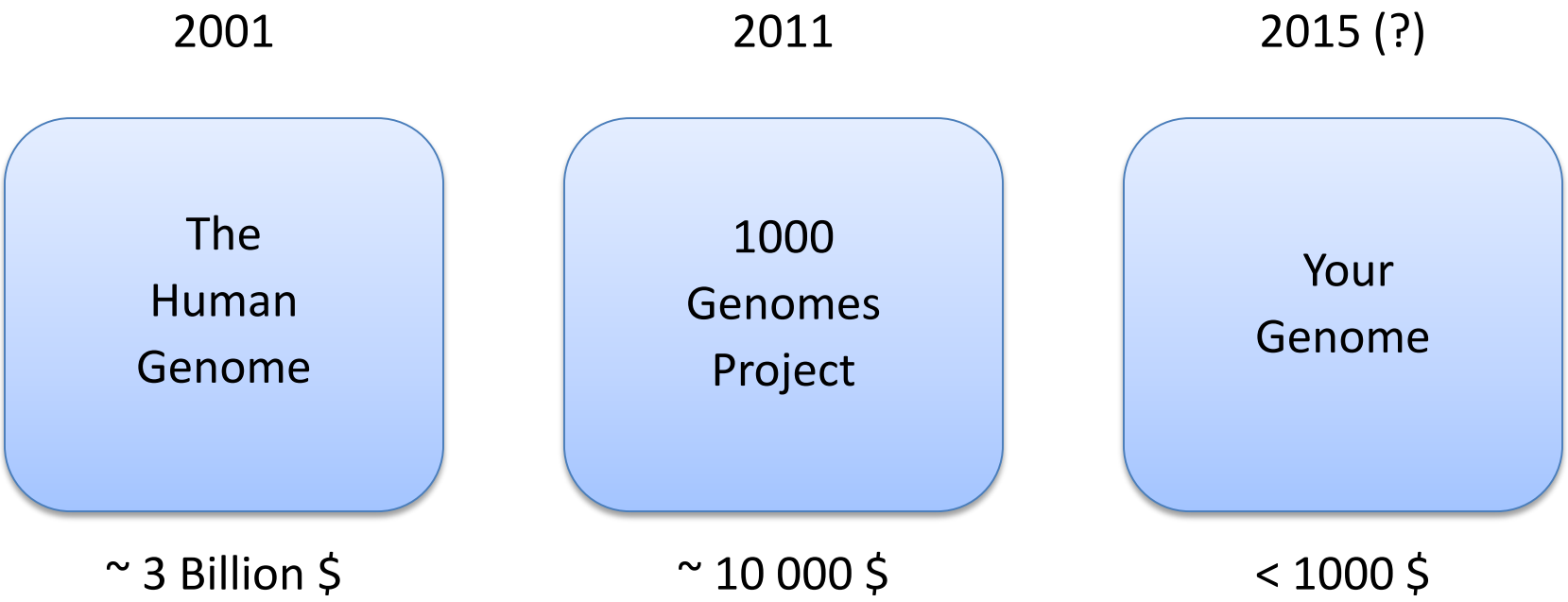
# DNA bases sequenced at the Innovation Center




300 Trillions!

~3000 human genomes at 30X

# Sequencing human genomes



# Big Data

2014

2 X 10 TBytes

1 TBytes

*Intensity files*

*Reads + qualities*

300 TBytes

15 TBytes

30 TB of raw data / month
360 TB of raw data / year

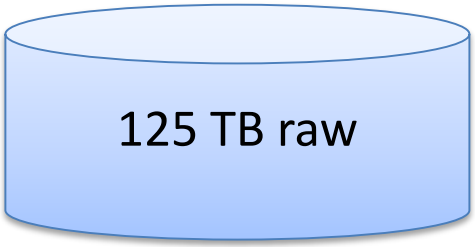
McGill

GenomeQuébec

# Large NGS project

Cancer project with whole genome data:

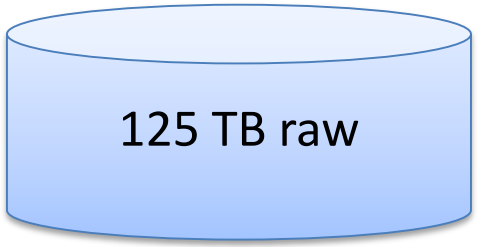
500 tumors




500 X 3 lanes = 500 X 250GB

vs

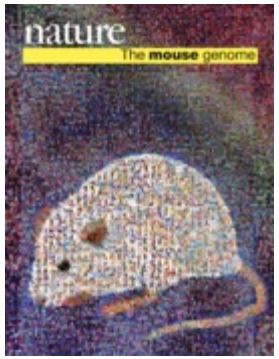
500 matched-normal




500 X 3 lanes = 500 X 250GB

# Applications (I)

- *De novo* sequencing
  - From the human genome… To all model organisms… To all relevant organisms (e.g. extreme genomes)… To "all" organisms?

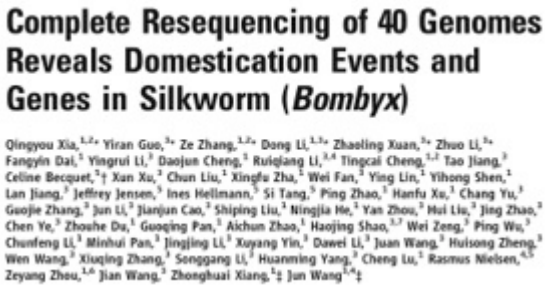




# Applications (II)

- Genome re-sequencing
  - Map genomic structural variations across individuals (to understand genetic disorders and also susceptibility factors)
  - Cancer genome sequencing
  - Agricultural crops

1000 Genomes Project


The Cancer Genome Atlas

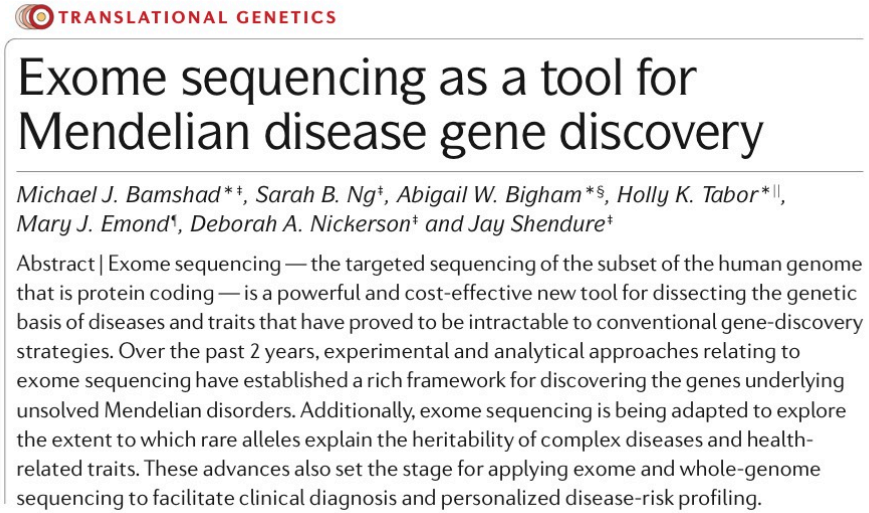


**Complete Resequencing of 40 Genomes Reveals Domestication Events and Genes in Silkworm (*Bombyx*)**

Qingyou Xia,[1,2*] Yiran Guo,[3*] Ze Zhang,[1,2*] Dong Li,[1,3*] Zhaoling Xuan,[3*] Zhuo Li,[3*] Fangyin Dai,[1] Yingrui Li,[3] Daojun Cheng,[1] Ruiqiang Li,[3,4] Tingcai Cheng,[1,2] Tao Jiang,[3] Celine Becquet,[5†] Xun Xu,[3] Chun Liu,[1] Xingfu Zha,[1] Wei Fan,[3] Ying Lin,[1] Yihong Shen,[1] Lan Jiang,[3] Jeffrey Jensen,[3] Ines Hellmann,[3] Si Tang,[3] Ping Zhao,[1] Hanfu Xu,[1] Chang Yu,[3] Guojie Zhang,[3] Jun Li,[3] Jianjun Cao,[3] Shiping Liu,[3] Ningjia He,[1] Yan Zhou,[3] Hui Liu,[3] Jing Zhao,[3] Chen Ye,[3] Zhouhe Du,[1] Guoqing Pan,[1] Aichun Zhao,[1] Haojing Shao,[1,3] Wei Zeng,[3] Ping Wu,[3] Chunfeng Li,[3] Minhui Pan,[3] Jingjing Li,[3] Xuyang Yin,[3] Dawei Li,[3] Juan Wang,[3] Huisong Zheng,[3] Wen Wang,[3] Xiuqing Zhang,[3] Songgang Li,[3] Huanming Yang,[3] Cheng Lu,[1] Rasmus Nielsen,[4,5] Zeyang Zhou,[1,6] Jian Wang,[3] Zhonghuai Xiang,[1‡] Jun Wang[3,6‡]

# Exome sequencing for Mendelian disease

REVIEWS



TRANSLATIONAL GENETICS

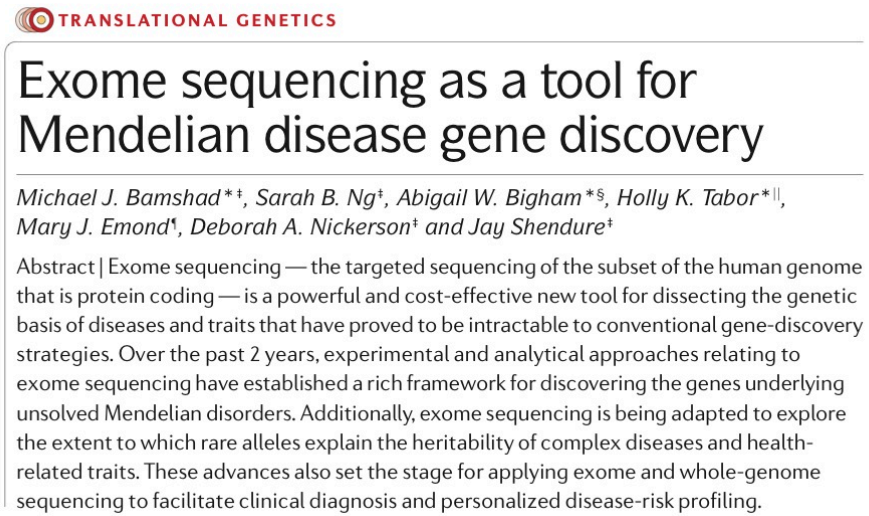
## Exome sequencing as a tool for Mendelian disease gene discovery

*Michael J. Bamshad*[*‡], *Sarah B. Ng*[‡], *Abigail W. Bigham*[*§], *Holly K. Tabor*[*‖], *Mary J. Emond*[¶], *Deborah A. Nickerson*[‡] *and Jay Shendure*[‡]

Abstract | Exome sequencing — the targeted sequencing of the subset of the human genome that is protein coding — is a powerful and cost-effective new tool for dissecting the genetic basis of diseases and traits that have proved to be intractable to conventional gene-discovery strategies. Over the past 2 years, experimental and analytical approaches relating to exome sequencing have established a rich framework for discovering the genes underlying unsolved Mendelian disorders. Additionally, exome sequencing is being adapted to explore the extent to which rare alleles explain the heritability of complex diseases and health-related traits. These advances also set the stage for applying exome and whole-genome sequencing to facilitate clinical diagnosis and personalized disease-risk profiling.

# Exome sequencing for Mendelian disease

REVIEWS



TRANSLATIONAL GENETICS

## Exome sequencing as a tool for Mendelian disease gene discovery

*Michael J. Bamshad*[*‡], *Sarah B. Ng*[‡], *Abigail W. Bigham*[*§], *Holly K. Tabor*[*‖], *Mary J. Emond*[¶], *Deborah A. Nickerson*[‡] *and Jay Shendure*[‡]

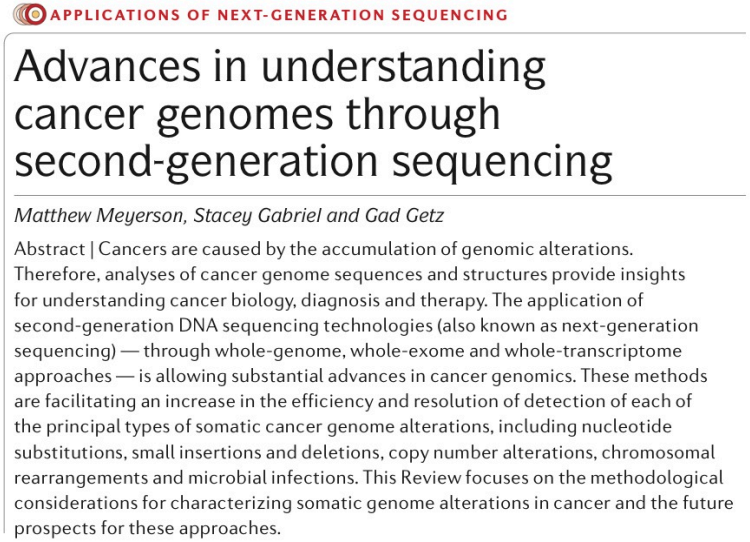
Abstract | Exome sequencing — the targeted sequencing of the subset of the human genome that is protein coding — is a powerful and cost-effective new tool for dissecting the genetic basis of diseases and traits that have proved to be intractable to conventional gene-discovery strategies. Over the past 2 years, experimental and analytical approaches relating to exome sequencing have established a rich framework for discovering the genes underlying unsolved Mendelian disorders. Additionally, exome sequencing is being adapted to explore the extent to which rare alleles explain the heritability of complex diseases and health-related traits. These advances also set the stage for applying exome and whole-genome sequencing to facilitate clinical diagnosis and personalized disease-risk profiling.

"… about one-half to one-third (~3,000) of all known or suspected Mendelian disorders (for example, cystic fibrosis and sickle cell anaemia) have been discovered. However, there is a substantial gap in our knowledge about the genes that cause many rare Mendelian phenotypes."

"Accordingly, we can realistically look towards a future in which the genetic basis of all Mendelian traits is known, …"

# Cancer genome sequencing

REVIEWS



APPLICATIONS OF NEXT-GENERATION SEQUENCING

## Advances in understanding cancer genomes through second-generation sequencing

*Matthew Meyerson, Stacey Gabriel and Gad Getz*

Abstract | Cancers are caused by the accumulation of genomic alterations. Therefore, analyses of cancer genome sequences and structures provide insights for understanding cancer biology, diagnosis and therapy. The application of second-generation DNA sequencing technologies (also known as next-generation sequencing) — through whole-genome, whole-exome and whole-transcriptome approaches — is allowing substantial advances in cancer genomics. These methods are facilitating an increase in the efficiency and resolution of detection of each of the principal types of somatic cancer genome alterations, including nucleotide substitutions, small insertions and deletions, copy number alterations, chromosomal rearrangements and microbial infections. This Review focuses on the methodological considerations for characterizing somatic genome alterations in cancer and the future prospects for these approaches.

# Cancer genome sequencing

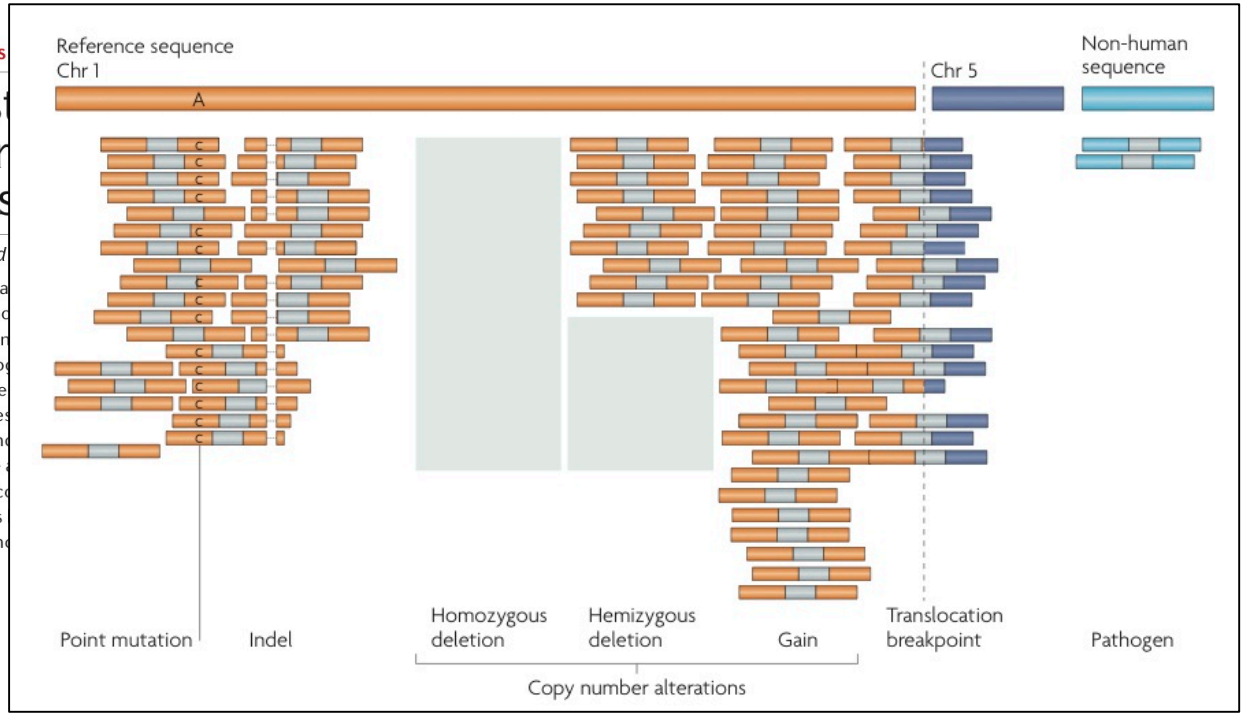
REVIEWS

APPLICATIONS OF NEXT-GENERATION S

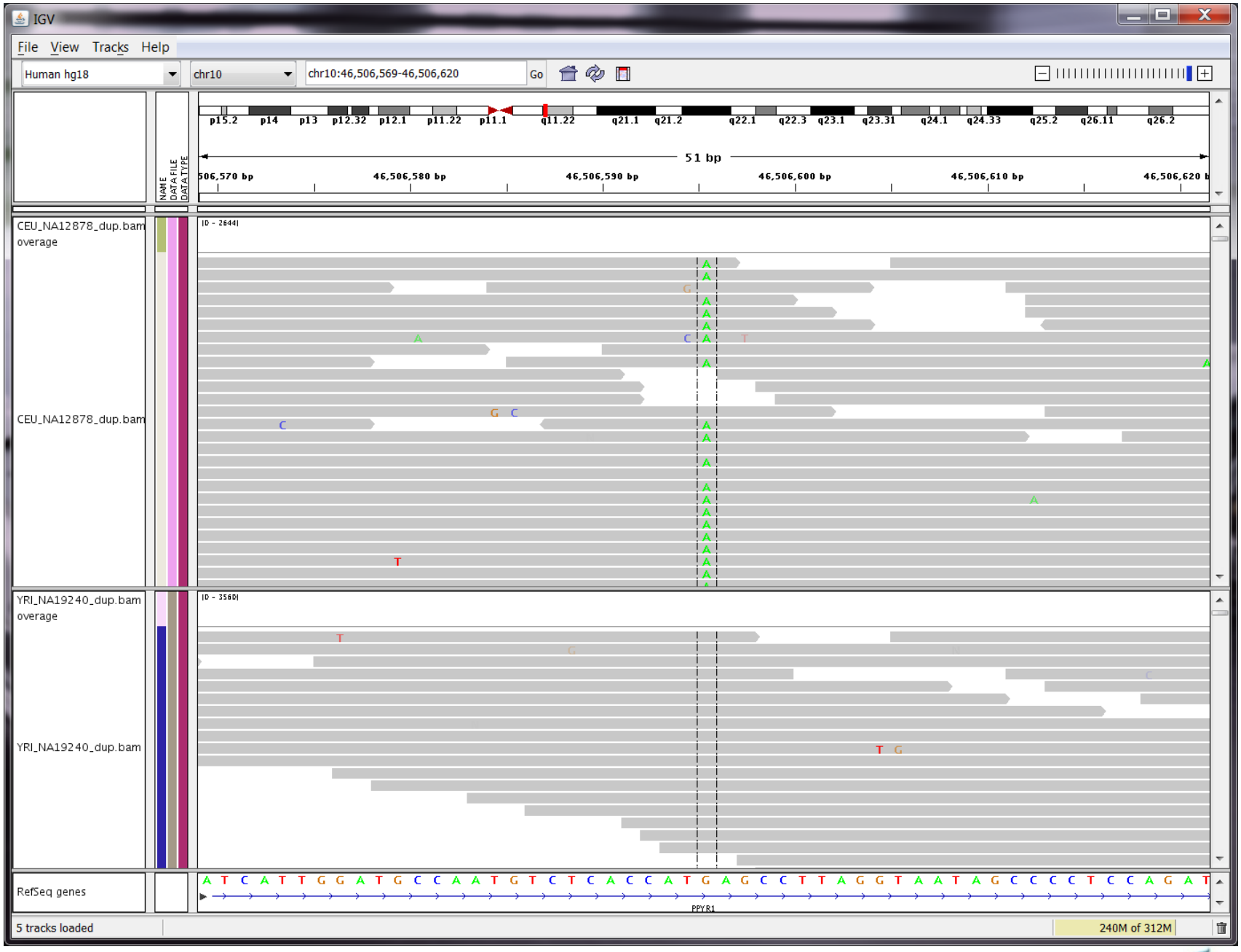
Advances in underst
cancer genomes thr
second-generation s

*Matthew Meyerson, Stacey Gabriel and Gad*

Abstract | Cancers are caused by the accumula
Therefore, analyses of cancer genome sequenc
for understanding cancer biology, diagnosis ar
second-generation DNA sequencing technolo
sequencing) — through whole-genome, whole
approaches — is allowing substantial advance
are facilitating an increase in the efficiency an
the principal types of somatic cancer genome
substitutions, small insertions and deletions, c
rearrangements and microbial infections. This
considerations for characterizing somatic gen
prospects for these approaches.

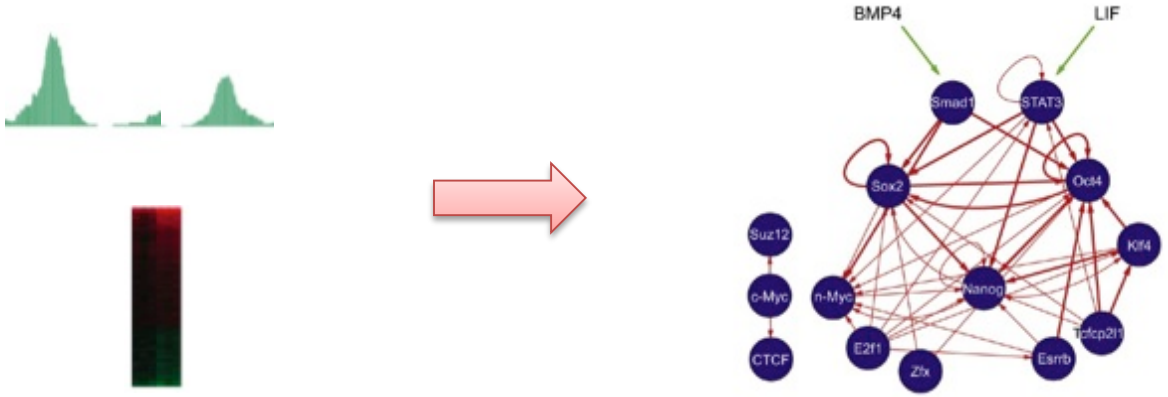



Can obtain a full catalogue of mutations

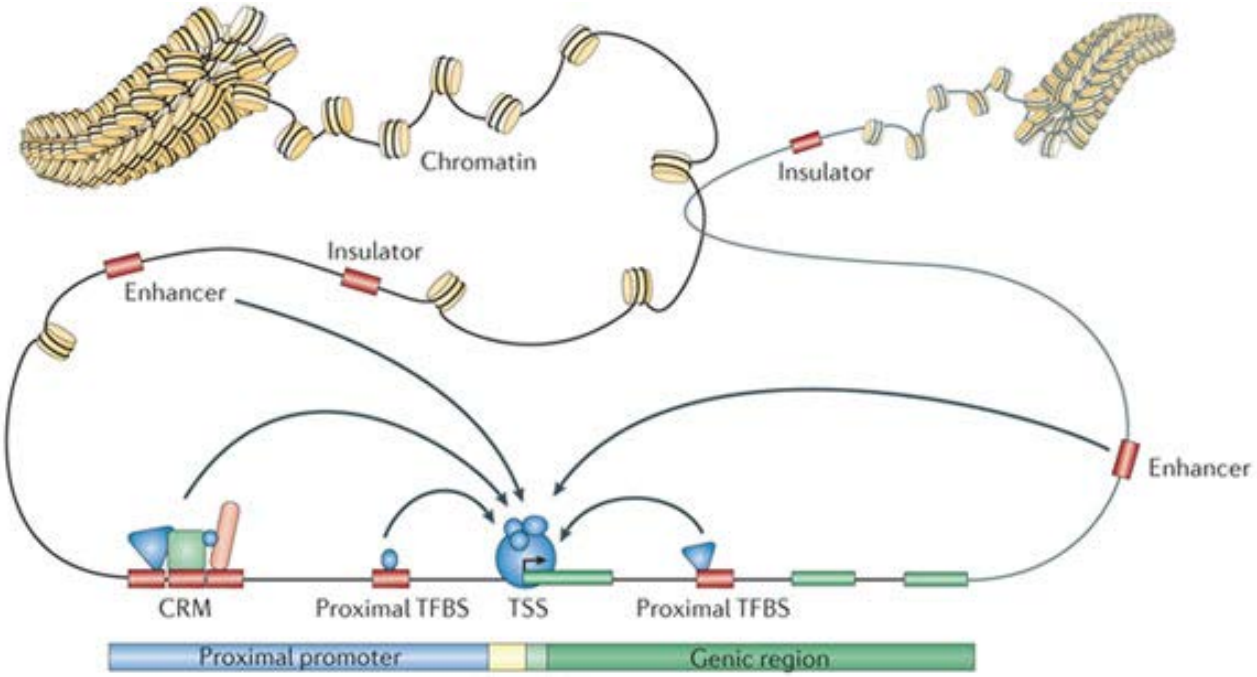

IGV
File View Tracks Help
Human hg18
chr10
chr10:46,506,569-46,506,620
Go
51 bp
CEU_NA12878_dup.bam
overage
CEU_NA12878_dup.bam
YRI_NA19240_dup.bam
overage
YRI_NA19240_dup.bam
RefSeq genes
5 tracks loaded
240M of 312M

# Applications (III)

- Quantitative biology of complex systems
  - New high-throughput technologies in functional genomics: ChIP-Seq, RNA-Seq, ChIA-PET, RIP-Seq, …
  - From single-gene measurements, to thousands of probes on arrays, to profiles covering all 3B bases of the genome



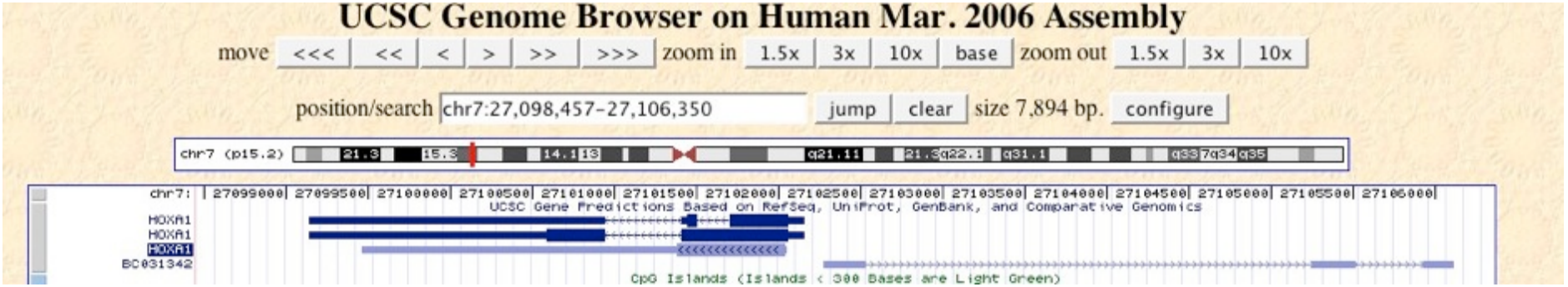
# Transcription Regulation



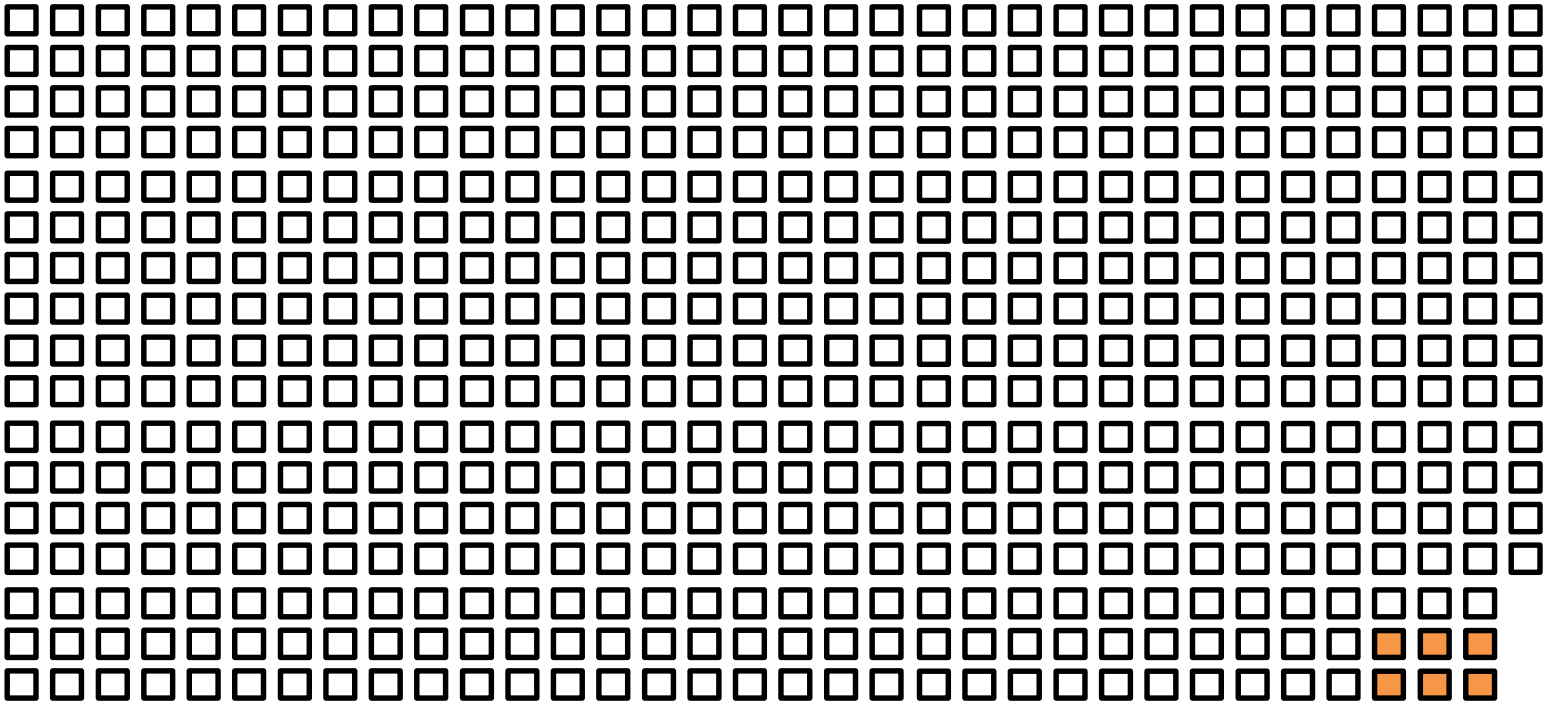

Lenhard, Nat Rev Genet, 2012

# Functional genomics

Locate regulatory regions in the human genome in different cell types, describe their functions, and identify how they differ between different groups (i.e. “disease” vs “healthy”)…

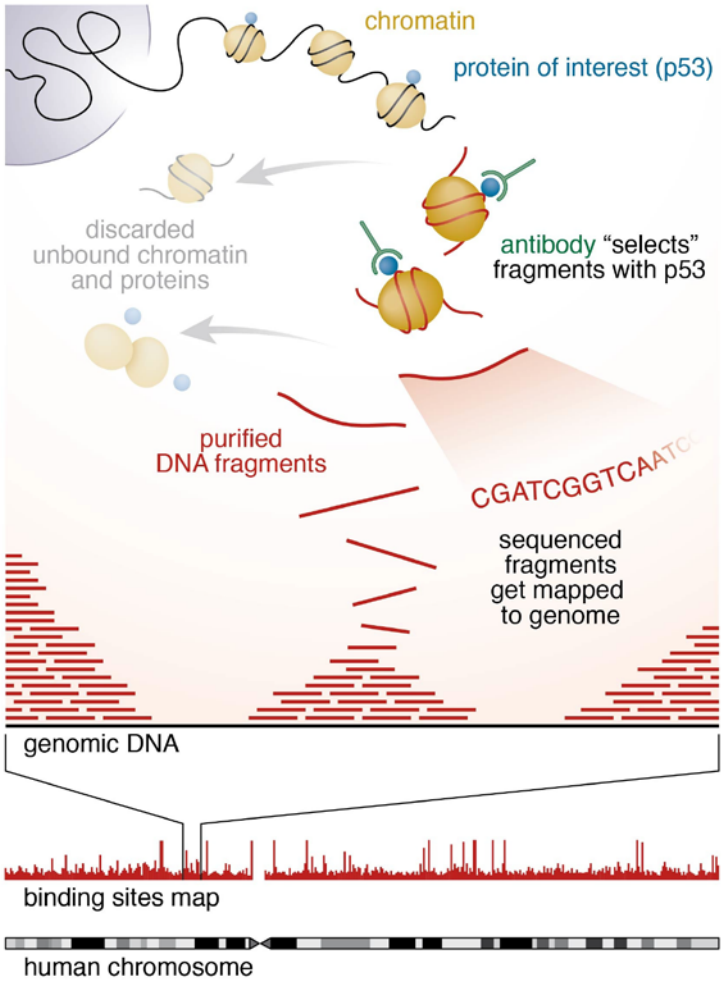


# Somatic mutations in 100 kidney tumors



1000 mutations (Total 575693)
1000 coding mutations (Total 6172)

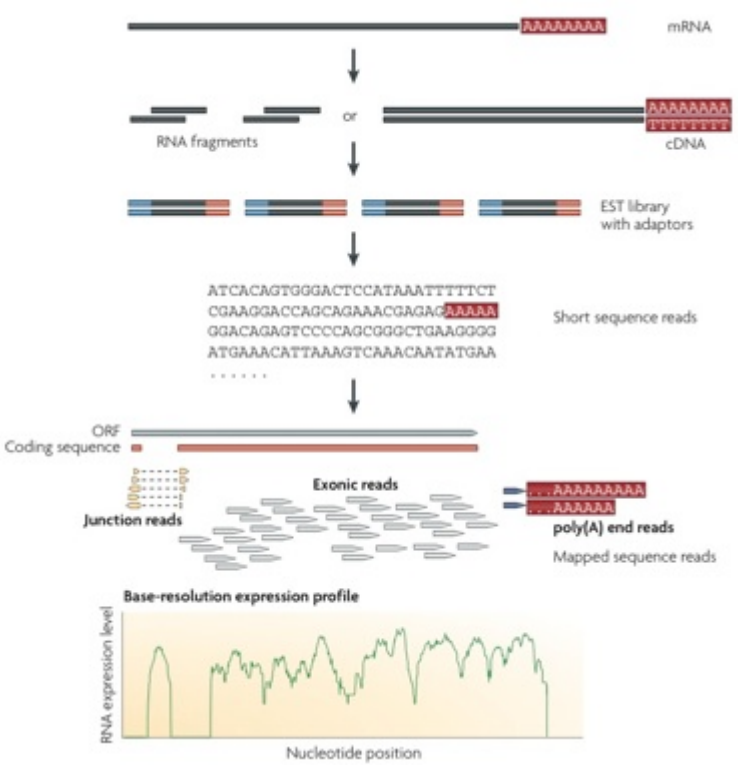
# Chromatin immunoprecipitation – Sequencing (ChIP-Seq)






Source: www.bnl.gov

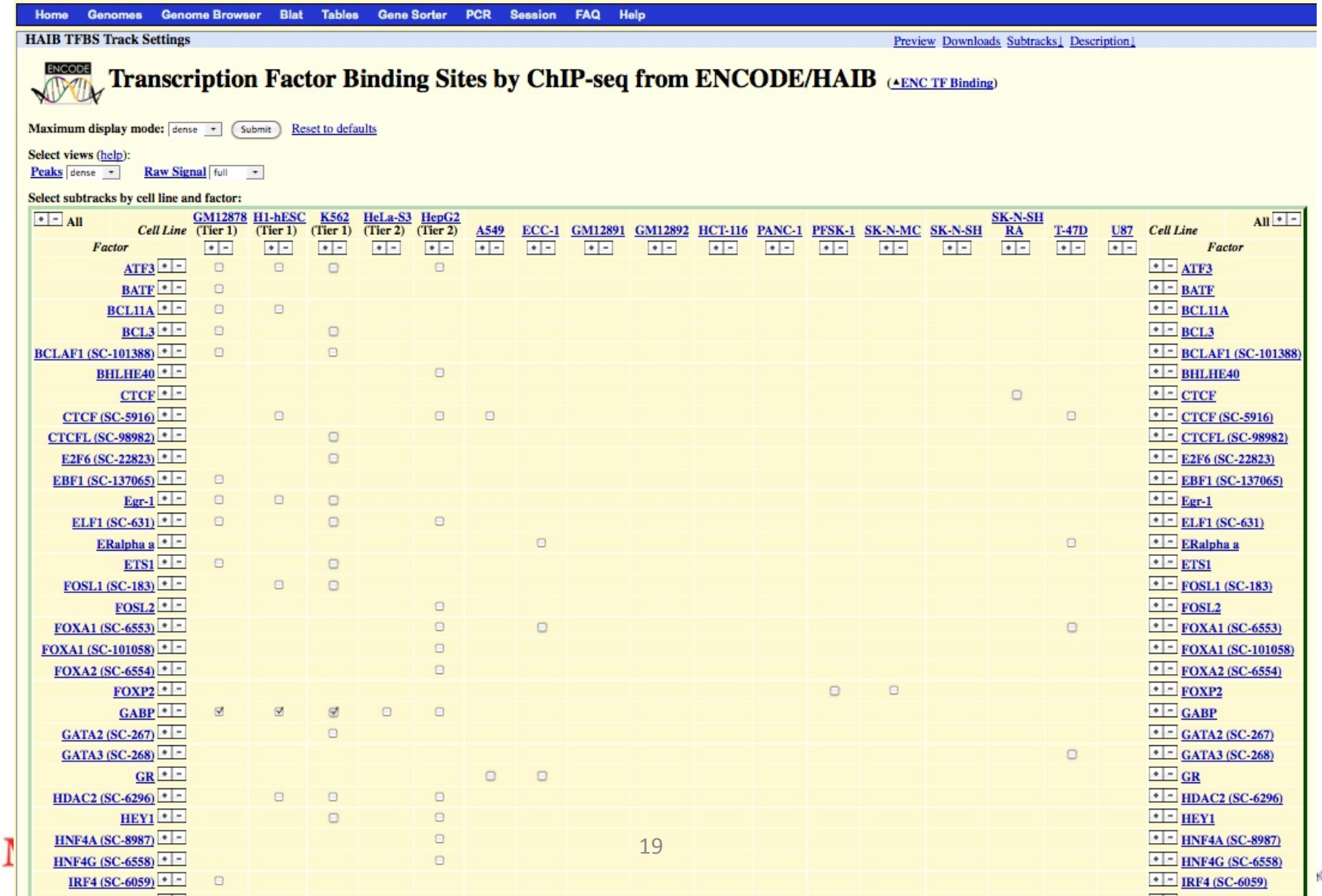
# RNA-Seq: digital expression and much more





Wang et al. *Nat. Rev. Genet.*, 2009

# Some of the ENCODE ChIP-Seq



Home Genomes Genome Browser Blat Tables Gene Sorter PCR Session FAQ Help

HAIB TFBS Track Settings

Preview Downloads Subtracks↓ Description↓

## Transcription Factor Binding Sites by ChIP-seq from ENCODE/HAIB (▲ENC TF Binding)

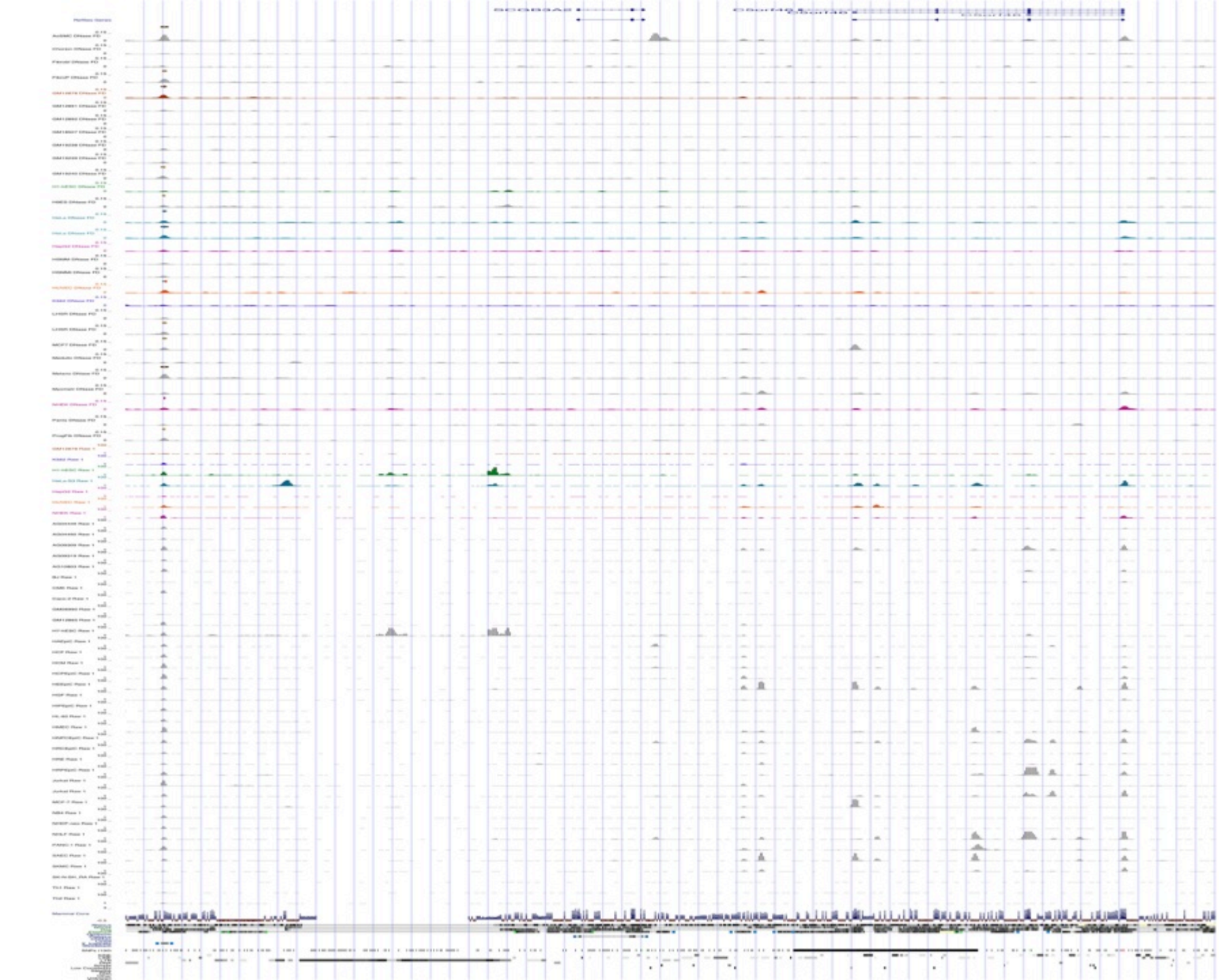
Maximum display mode: dense Submit Reset to defaults

Select views (help):

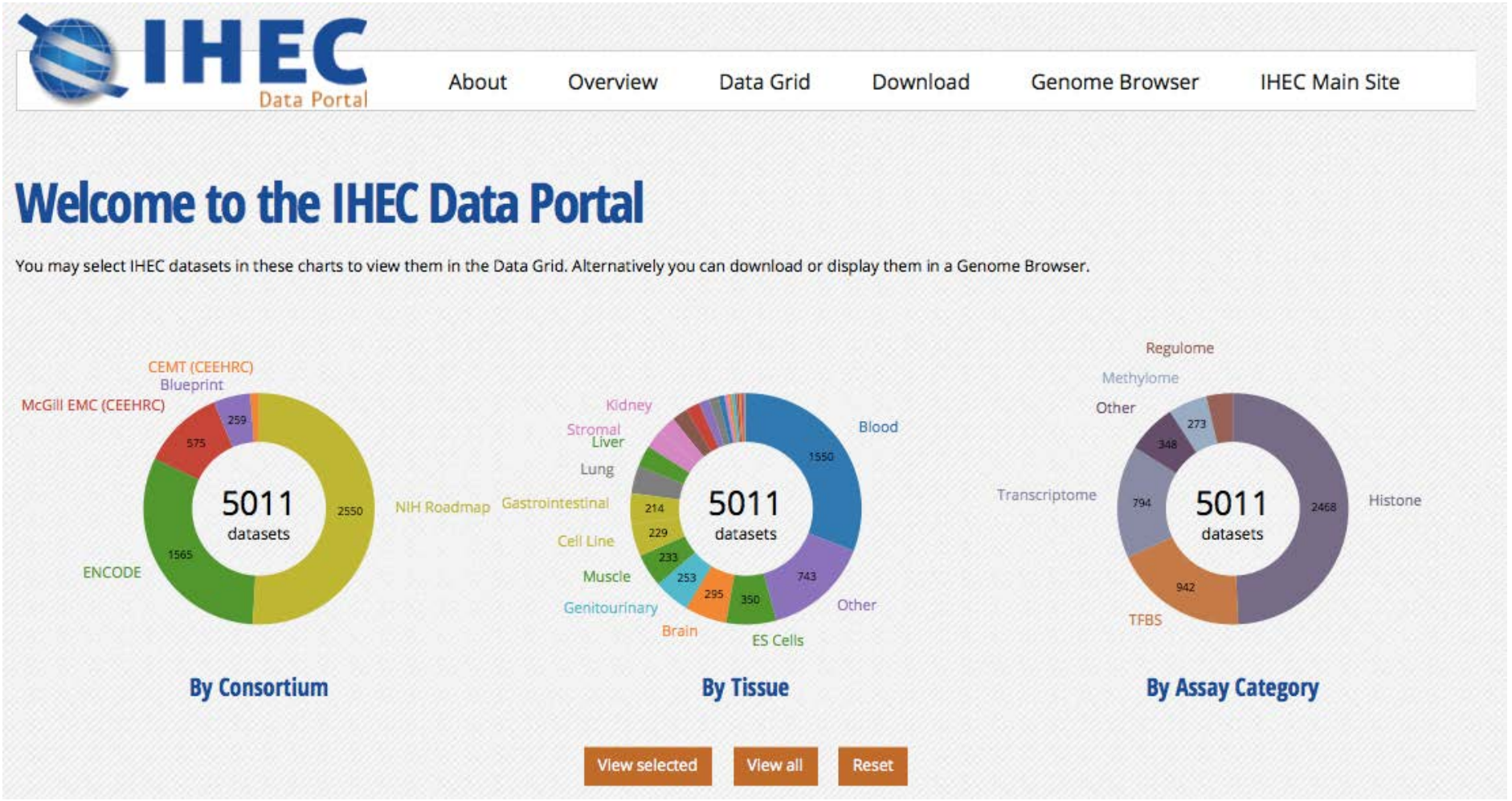
Peaks dense Raw Signal full

Select subtracks by cell line and factor:

| All | Cell Line | GM12878 (Tier 1) | H1-hESC (Tier 1) | K562 (Tier 1) | HeLa-S3 (Tier 2) | HepG2 (Tier 2) | A549 | ECC-1 | GM12891 | GM12892 | HCT-116 | PANC-1 | PFSK-1 | SK-N-MC | SK-N-SH | SK-N-SH RA | T-47D | U87 | Cell Line | All |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Factor | | | | | | | | | | | | | | | | | | | |
| ATF3 | | ☐ | ☐ | ☐ | | ☐ | | | | | | | | | | | | | | ATF3 |
| BATF | | ☐ | | | | | | | | | | | | | | | | | | BATF |
| BCL11A | | ☐ | ☐ | | | | | | | | | | | | | | | | | BCL11A |
| BCL3 | | ☐ | | ☐ | | | | | | | | | | | | | | | | BCL3 |
| BCLAF1 (SC-101388) | | ☐ | | ☐ | | | | | | | | | | | | | | | | BCLAF1 (SC-101388) |
| BHLHE40 | | | | | | ☐ | | | | | | | | | | | | | | BHLHE40 |
| CTCF | | | | | | | | | | | | | | | | ☐ | | | | CTCF |
| CTCF (SC-5916) | | | ☐ | | | ☐ | ☐ | | | | | | | | | | ☐ | | | CTCF (SC-5916) |
| CTCFL (SC-98982) | | | | ☐ | | | | | | | | | | | | | | | | CTCFL (SC-98982) |
| E2F6 (SC-22823) | | | | ☐ | | | | | | | | | | | | | | | | E2F6 (SC-22823) |
| EBF1 (SC-137065) | | ☐ | | | | | | | | | | | | | | | | | | EBF1 (SC-137065) |
| Egr-1 | | ☐ | ☐ | ☐ | | | | | | | | | | | | | | | | Egr-1 |
| ELF1 (SC-631) | | ☐ | | ☐ | | ☐ | | | | | | | | | | | | | | ELF1 (SC-631) |
| ERalpha a | | | | | | | | ☐ | | | | | | | | | ☐ | | | ERalpha a |
| ETS1 | | ☐ | | ☐ | | | | | | | | | | | | | | | | ETS1 |
| FOSL1 (SC-183) | | | ☐ | ☐ | | | | | | | | | | | | | | | | FOSL1 (SC-183) |
| FOSL2 | | | | | | ☐ | | | | | | | | | | | | | | FOSL2 |
| FOXA1 (SC-6553) | | | | | | ☐ | | ☐ | | | | | | | | | ☐ | | | FOXA1 (SC-6553) |
| FOXA1 (SC-101058) | | | | | | ☐ | | | | | | | | | | | | | | FOXA1 (SC-101058) |
| FOXA2 (SC-6554) | | | | | | ☐ | | | | | | | | | | | | | | FOXA2 (SC-6554) |
| FOXP2 | | | | | | | | | | | | | ☐ | ☐ | | | | | | FOXP2 |
| GABP | | ☑ | ☑ | ☑ | ☐ | ☐ | | | | | | | | | | | | | | GABP |
| GATA2 (SC-267) | | | | ☐ | | | | | | | | | | | | | | | | GATA2 (SC-267) |
| GATA3 (SC-268) | | | | | | | | | | | | | | | | | ☐ | | | GATA3 (SC-268) |
| GR | | | | | | | ☐ | ☐ | | | | | | | | | | | | GR |
| HDAC2 (SC-6296) | | | ☐ | ☐ | | ☐ | | | | | | | | | | | | | | HDAC2 (SC-6296) |
| HEY1 | | | | ☐ | | ☐ | | | | | | | | | | | | | | HEY1 |
| HNF4A (SC-8987) | | | | | | ☐ | | | | | | | | | | | | | | HNF4A (SC-8987) |
| HNF4G (SC-6558) | | | | | | ☐ | | | | | | | | | | | | | | HNF4G (SC-6558) |
| IRF4 (SC-6059) | | ☐ | | | | | | | | | | | | | | | | | | IRF4 (SC-6059) |



# International Human Epigenome Consortium



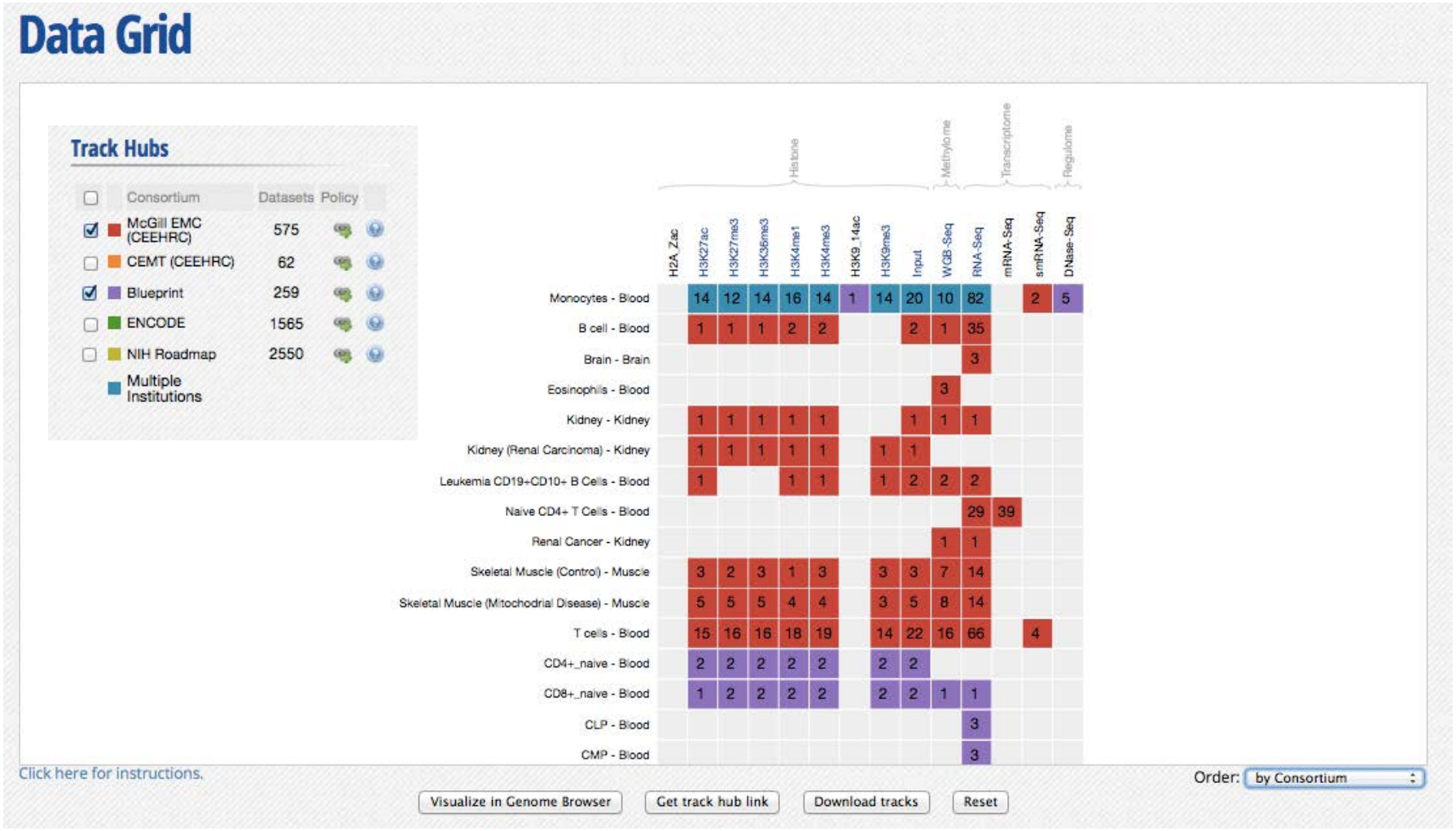

http://epigenomesportal.ca/ihec



McGill

GenomeQuébec

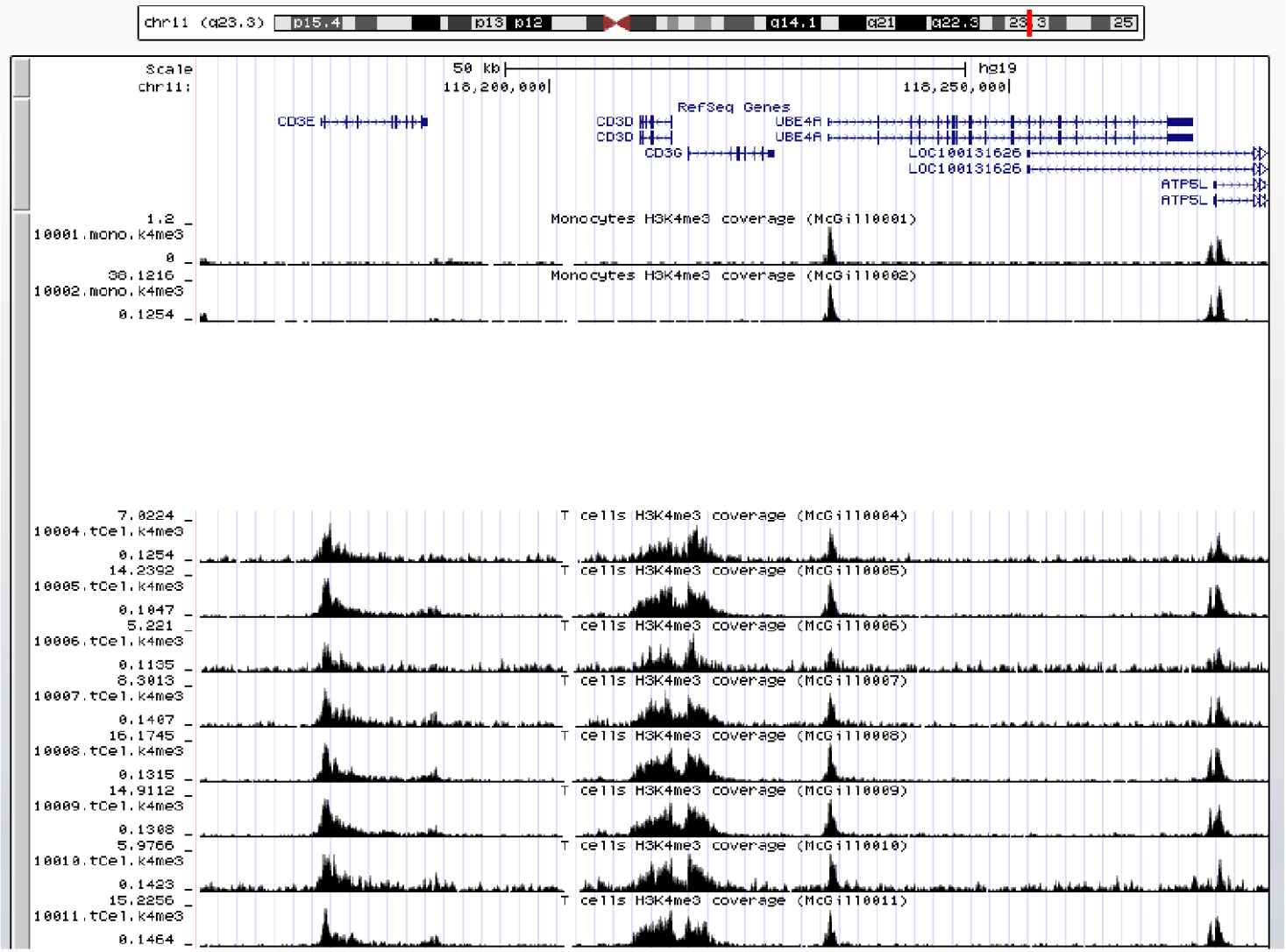
# IHEC Data Grid



# Outline

- Applications of next-generation sequencing
- Functional genomics
- Example of machine learning approaches in functional genomics (work by Toby Hocking)
- Role of transposable elements in gene regulation

# Motivation: visual differences between ChIP-seq profiles

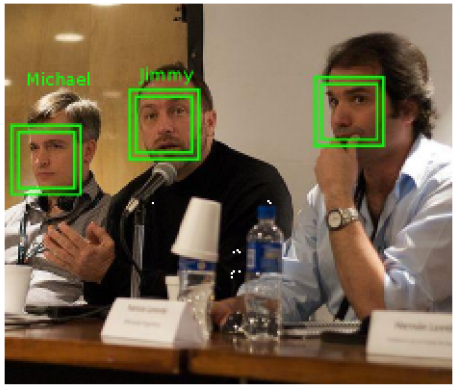
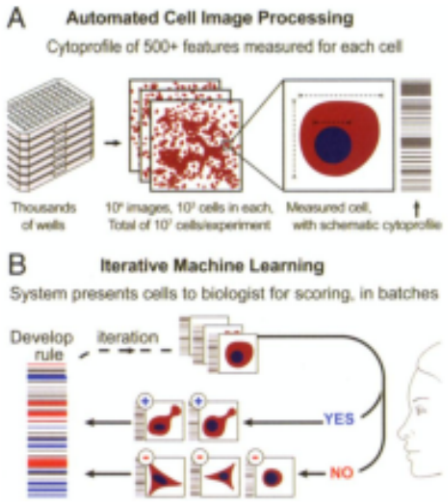
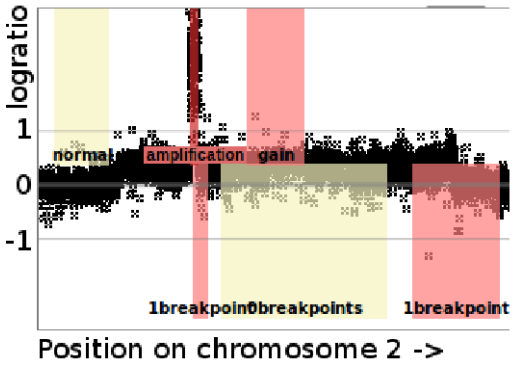


# Challenges with peak calling

- Different types of profiles (narrow peaks or broad signal)
- Many peak caller algorithms…
- Lots of parameters…

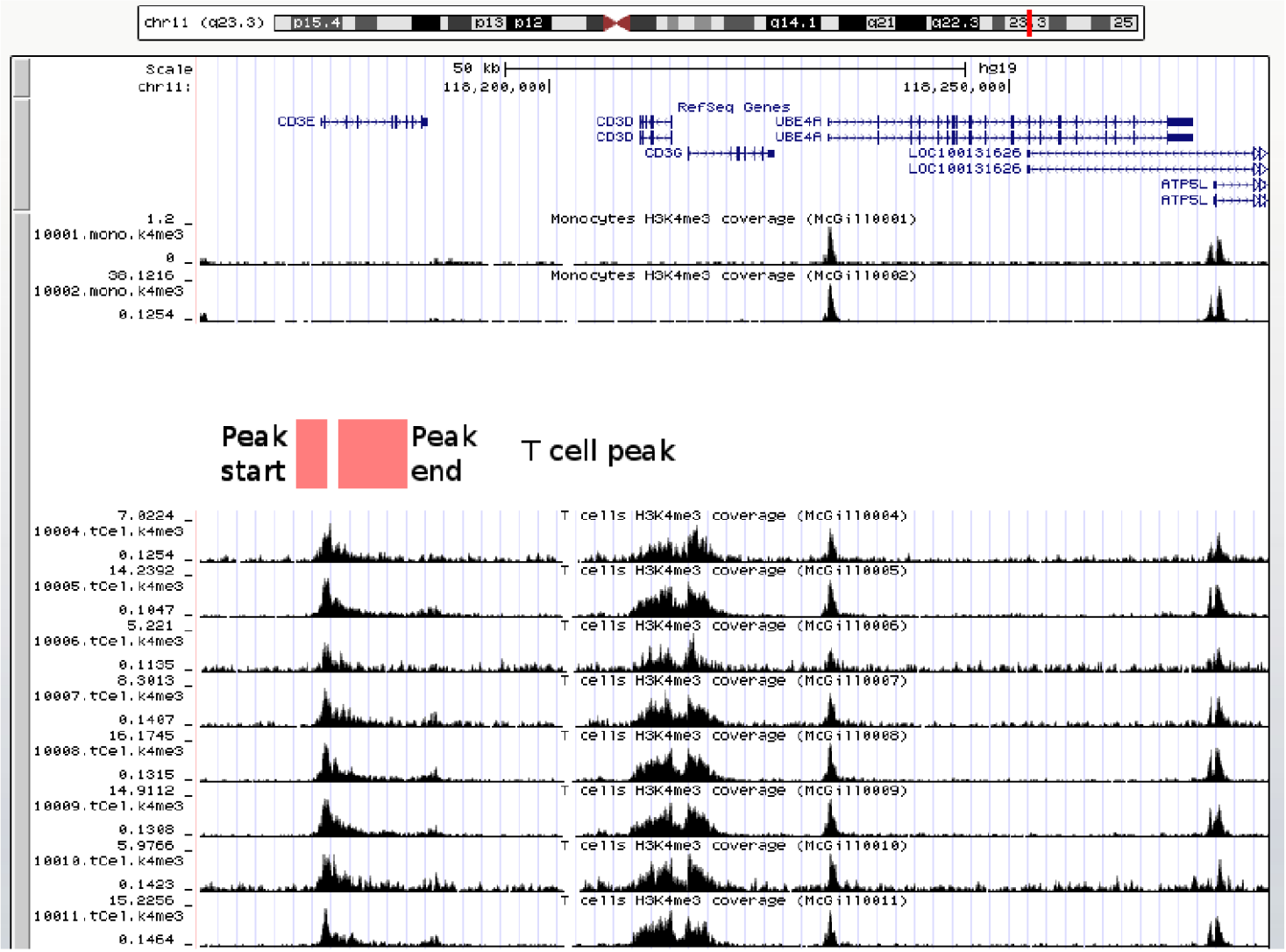
How to choose the right peak caller and optimal parameters?

# Previous work in computer vision: look and add labels to...

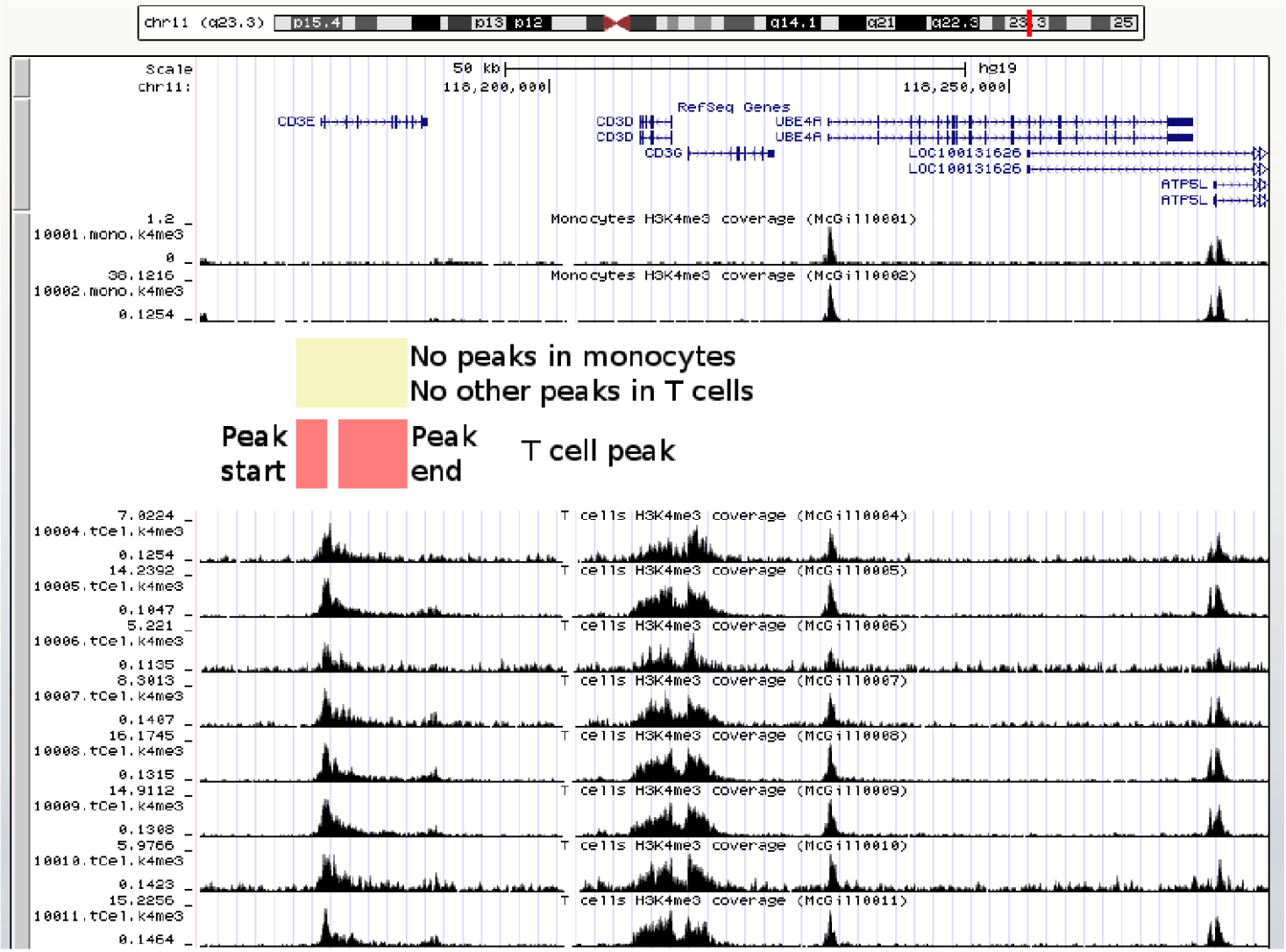
| Photos | Cell images | Copy number profiles |
|---|---|---|
|   |   |   |
| Labels: names | phenotypes | alterations |
| CVPR 2013<br>246 papers | CellProfiler<br>873 citations | SegAnnDB<br>H, *et. al.* 2014. |

Sources: http://en.wikipedia.org/wiki/Face_detection
Jones et al PNAS 2009. Scoring diverse cellular morphologies in image-based screens with iterative feedback and machine learning.

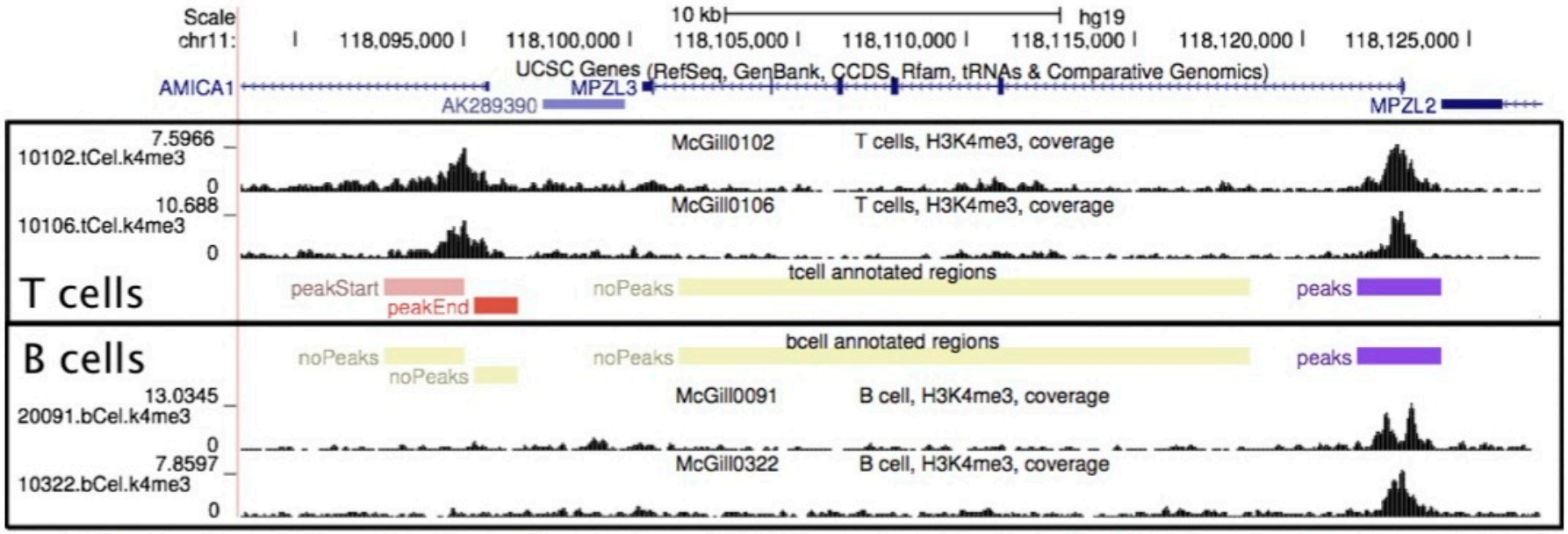
# Annotated regions for possible peak starts and ends



# Implicit specification of negative regions



# Visual peak annotation



# Supervised ML problem

A peak detection function or peak caller $c: \mathbb{R}^d \to \{0,1\}^d$ takes a coverage profile $\mathbf{x} \in \mathbb{R}^d$ as input, and returns a binary peak call prediction $\mathbf{y} = c(\mathbf{x}) \in \{0,1\}^d$ (0 is background noise, 1 is a peak).

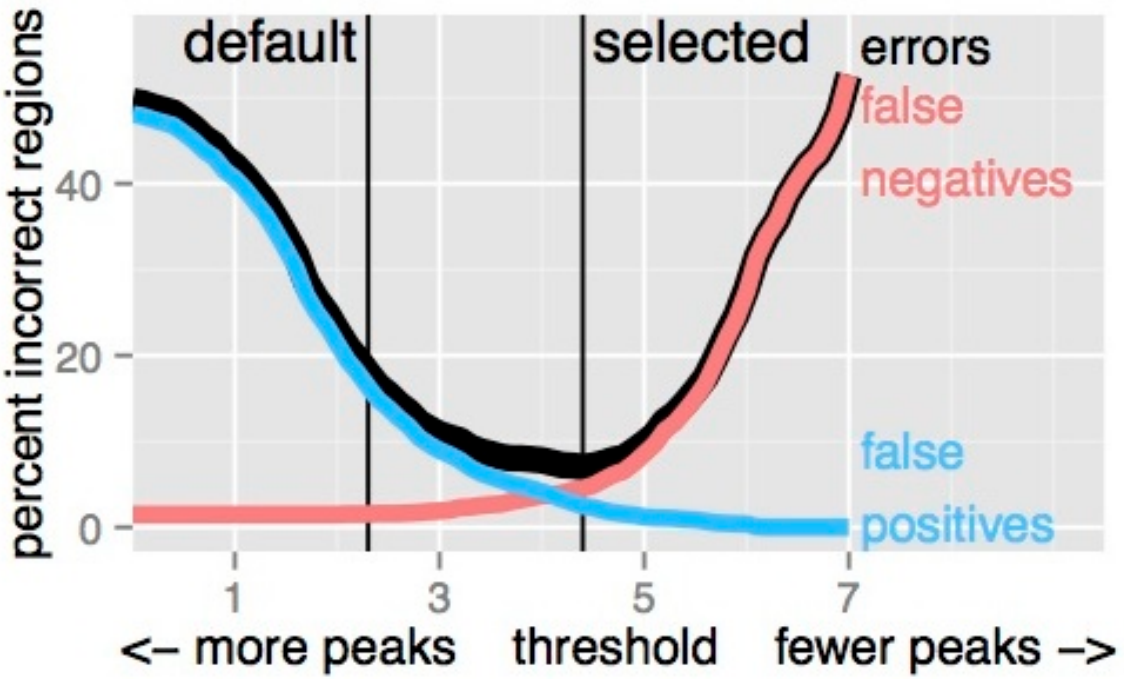
The goal is to learn how to call peaks $c(\mathbf{x}_i)$ which agree with the annotated regions $R_i^+, R_i^-, \underline{R}_i, \overline{R}_i$ for some test samples $i$. To quantify the error of the peak calls with respect to the annotation data, we define the annotation error as the sum of false positive (FP) and false negative (FN) regions:

$$E(\mathbf{y}, \underline{R}_i, \overline{R}_i, R_i^+, R_i^-) = \text{FP}(\mathbf{y}, \underline{R}_i, \overline{R}_i, R_i^-) + \text{FN}(\mathbf{y}, \underline{R}_i, \overline{R}_i, R_i^+). \quad (1)$$

The supervised machine learning problem can be formalized as the following optimization problem. Find the peak caller $c$ with minimal annotation error on a set of test samples:

$$\underset{c}{\text{minimize}} \sum_{i \in \text{test}} E\left[c(\mathbf{x}_i), \underline{R}_i, \overline{R}_i, R_i^+, R_i^-\right]. \quad (2)$$

# Choosing the threshold




$$\hat{\lambda} = \arg\min_{\lambda} \sum_{i \in \{1,\ldots,n\}} E\left[c_\lambda(\mathbf{x}_i), \underline{R}_i, \overline{R}_i, R_i^+, R_i^-\right].$$

# Creating a benchmark dataset

| mark | H3K36me3 | H3K4me3 | H3K36me3 | H3K36me3 | H3K4me3 | H3K4me3 | H3K4me3 |
|---|---|---|---|---|---|---|---|
| annotator | TDH | TDH | AM | TDH | PGP | TDH | XJ |
| sample.set | other | other | immune | immune | immune | immune | immune |
| windows | 4 | 29 | 23 | 4 | 30 | 27 | 12 |
| noPeaks | 72 | 536 | 752 | 230 | 1653 | 1656 | 702 |
| peakStart | 68 | 305 | 294 | 200 | 813 | 796 | 216 |
| peakEnd | 60 | 311 | 294 | 200 | 730 | 933 | 216 |
| peaks | | 218 | 403 | | 638 | 287 | 243 |
| tcell | | | 15 | 15 | 19 | 19 | 19 |
| monocyte | | | 5 | 5 | 6 | 6 | 6 |
| bcell | | | 1 | 1 | 2 | 2 | 2 |
| kidney | 1 | 1 | | | | | |
| kidneyCancer | 1 | 1 | | | | | |
| skeletalMuscleCtrl | 3 | 3 | | | | | |
| skeletalMuscleMD | 3 | 4 | | | | | |
| leukemiaCD19CD10BCells | | 1 | | | | | |

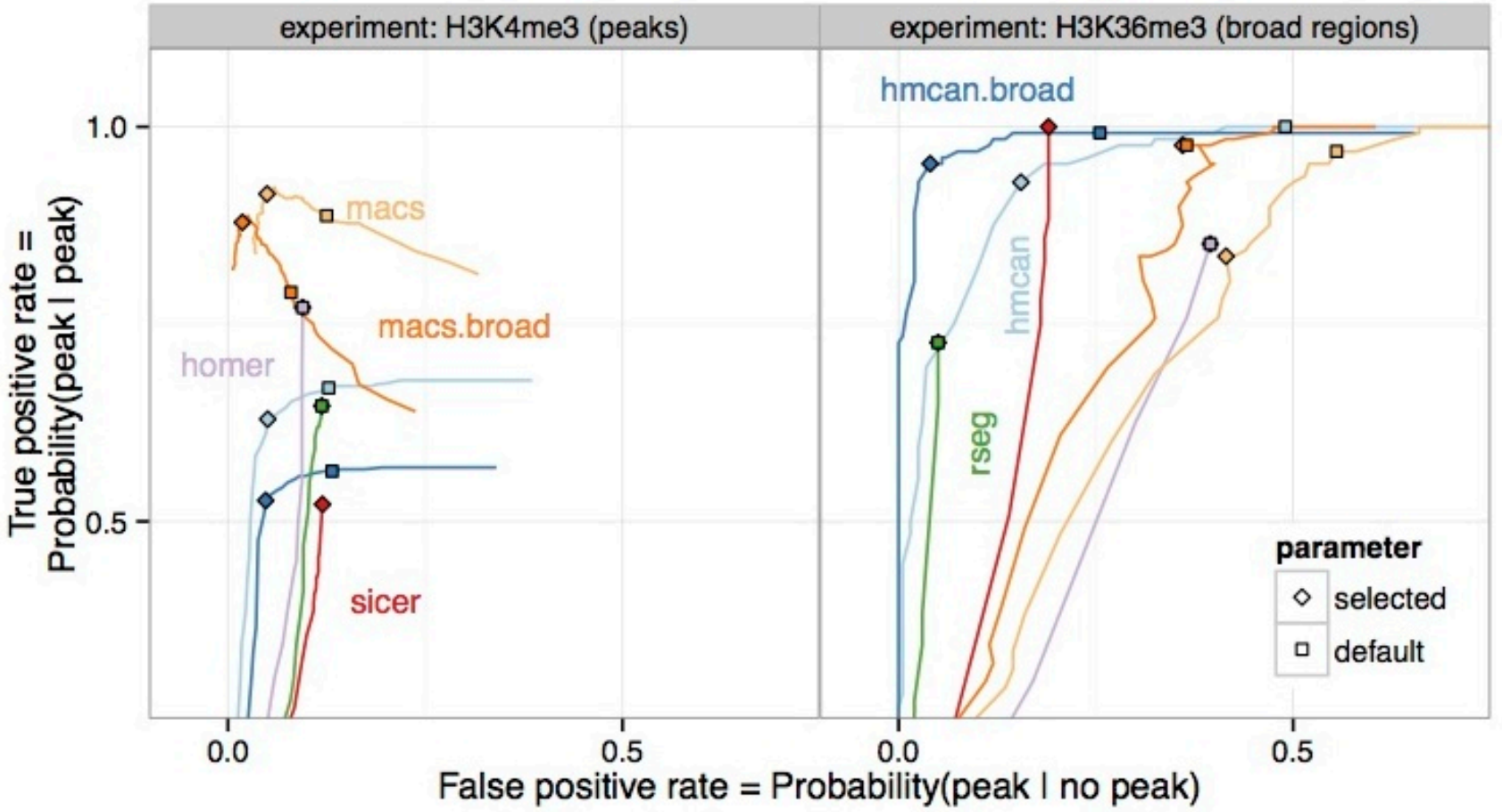
7 genomic regions, 2 histone marks, 4 expert annotators, 8 different cell types

Total of 12,286 annotated regions

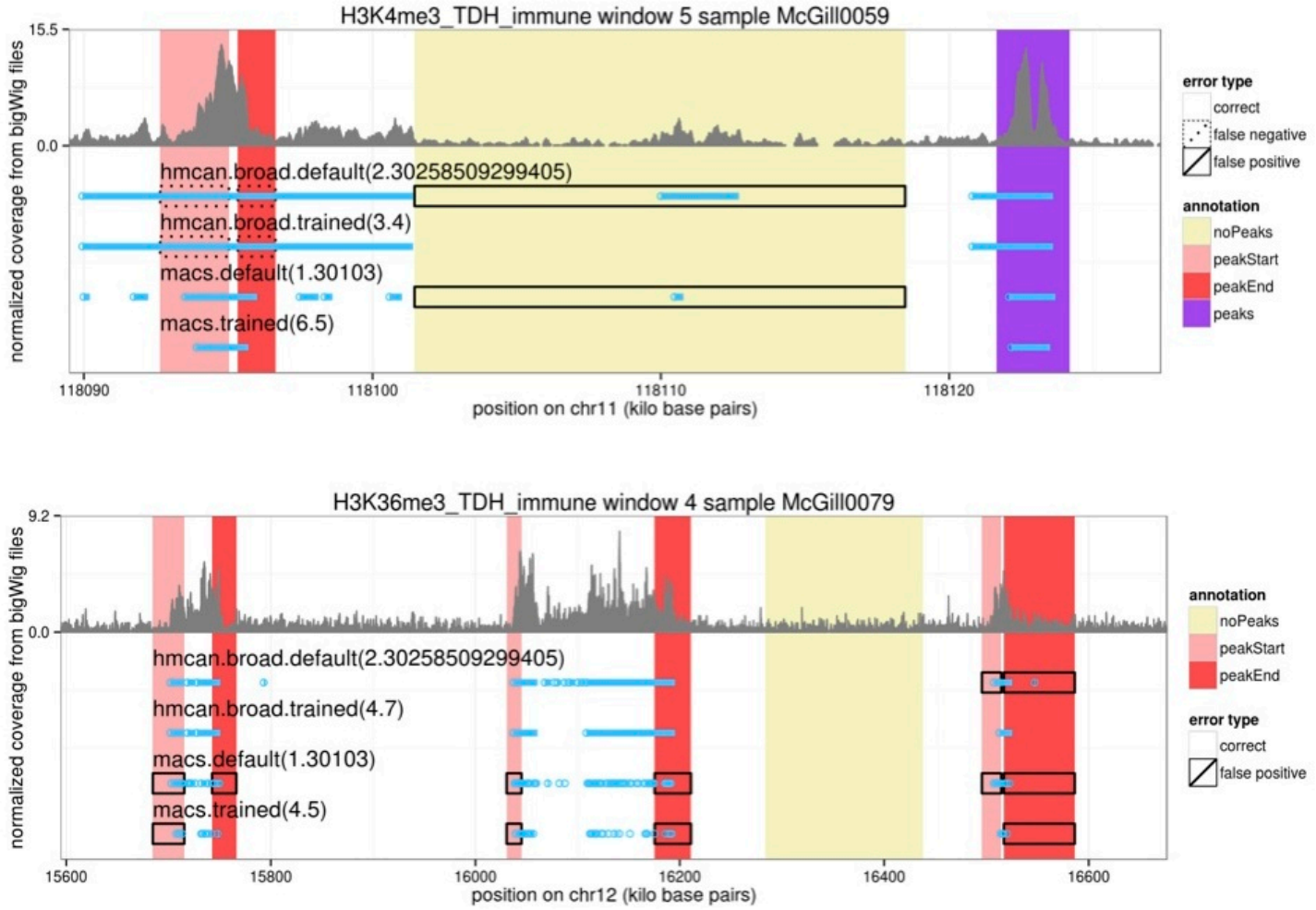
# Training and testing

| Same | Train | Test |
|---|---|---|
| Annotator, cell types | Chr1 | Chr2 |
| Annotator, regions | T cells | Kidney cells |
| Cell types, regions | Annotator 1 | Annotator 2 |

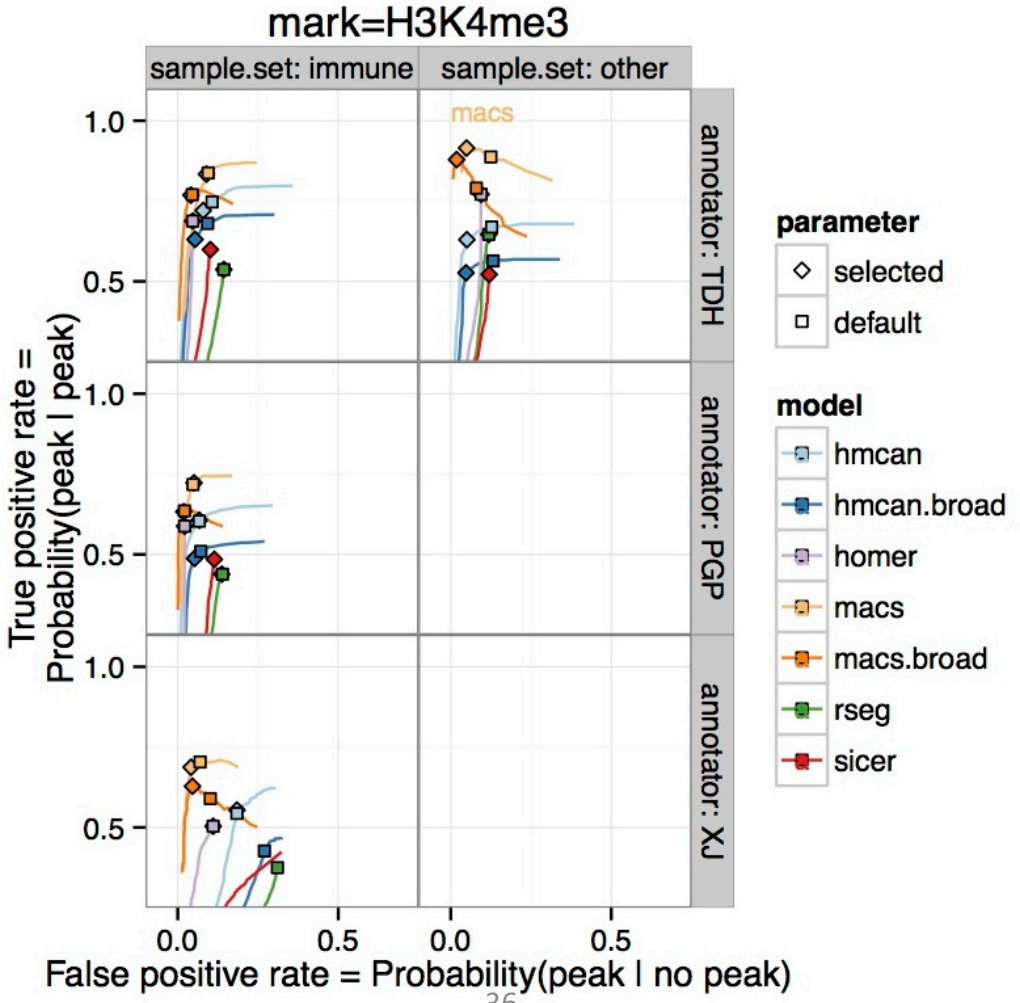
# Performance of different peak callers



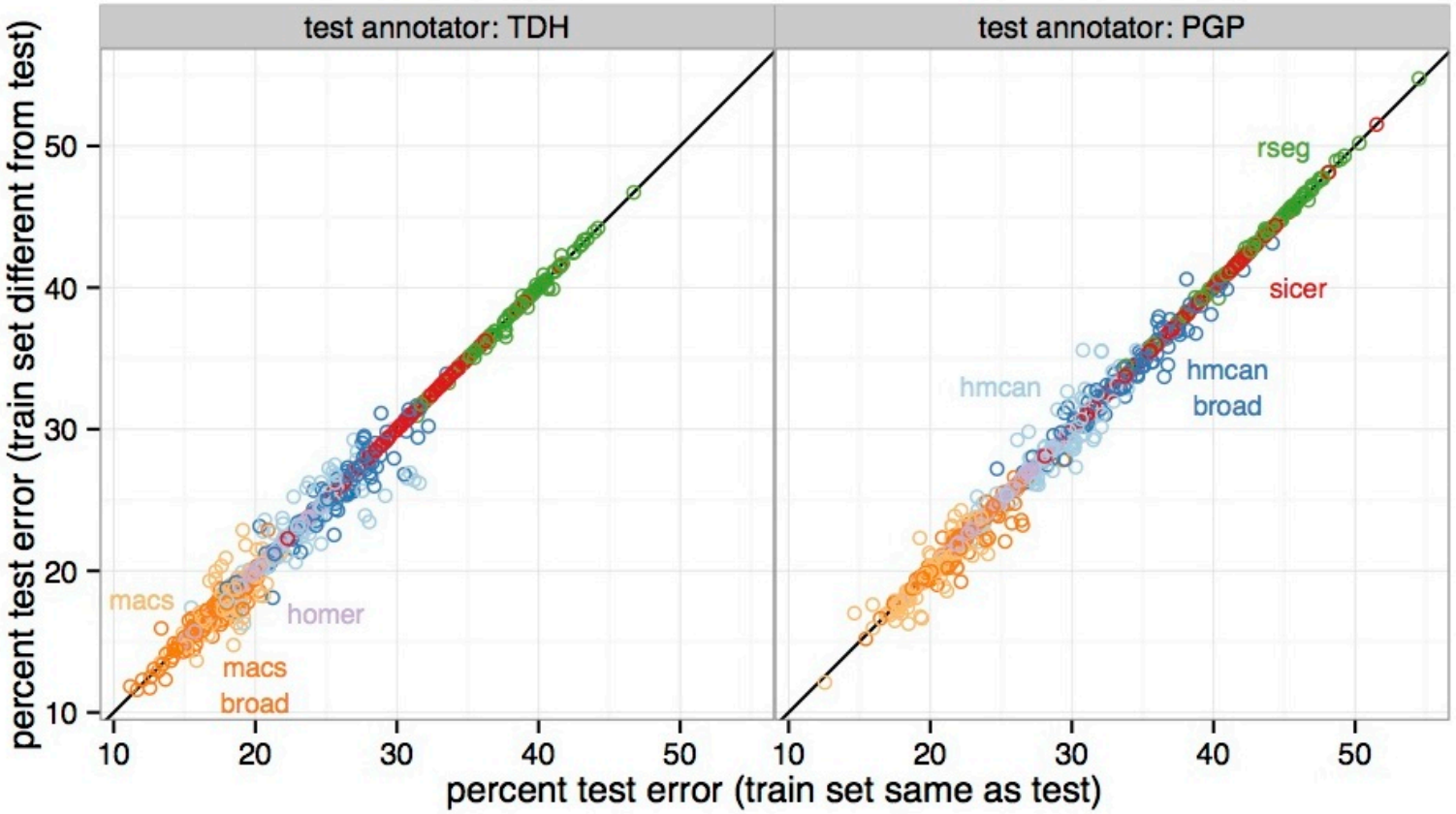
# Example region



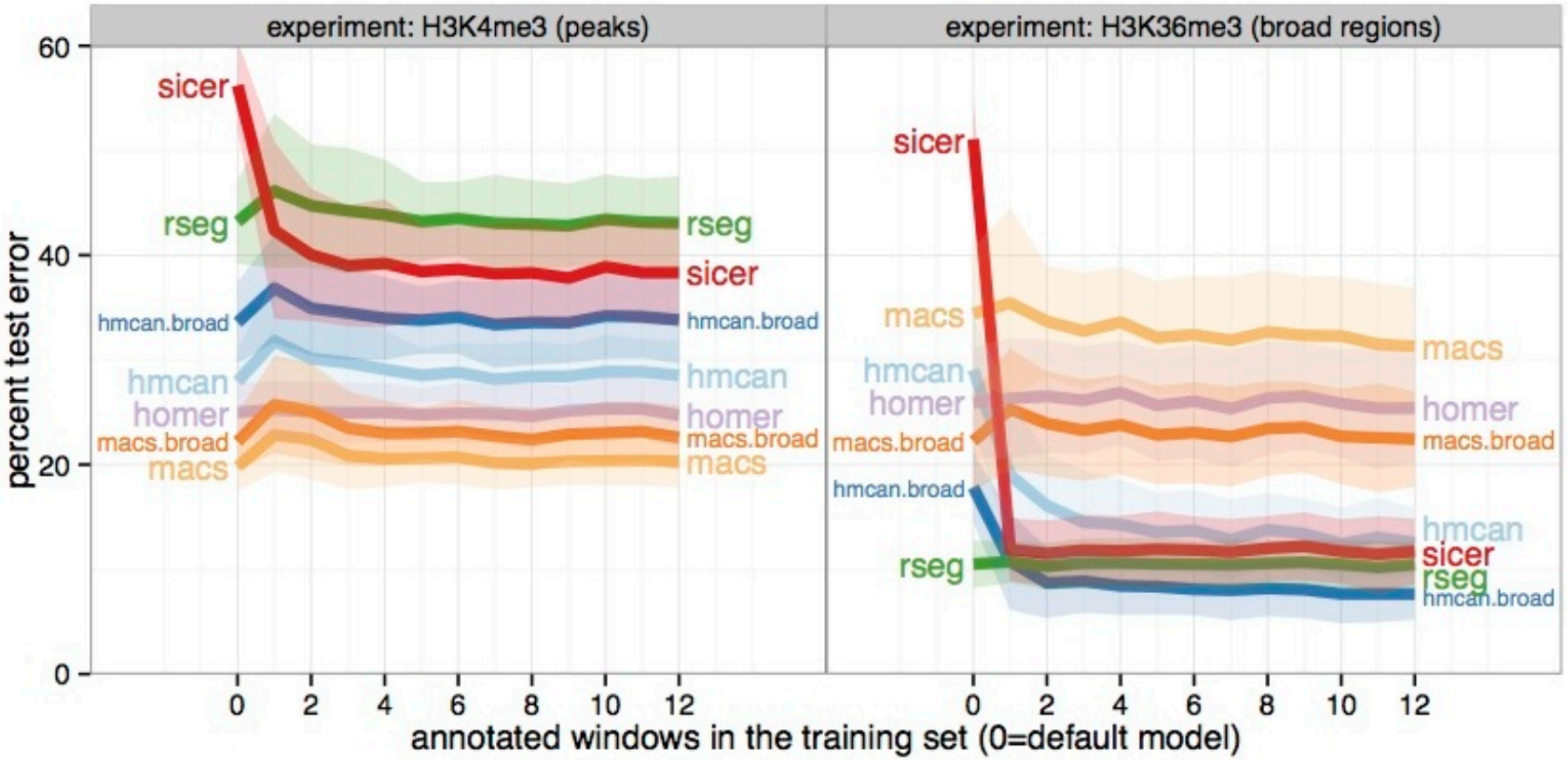
# Performance is stable across annotation datasets



# Test error comparable for different annotators



# Impact of the size of the training set



# Conclusions

- The macs algorithm showed the minimum train error of 9–20% across the 4 annotated “peak” data sets.
- The hmcan.broad algorithm showed the minimum train error of 7–20% across the 3 annotated “broad” data sets.
- Error of the different algorithms and model ordering was independent of the annotator
- Each algorithm quickly achieves its model-specific minimum error, after only about 4 annotated windows in the train set
- For some models the trained model parameters were clearly better than the default model parameters

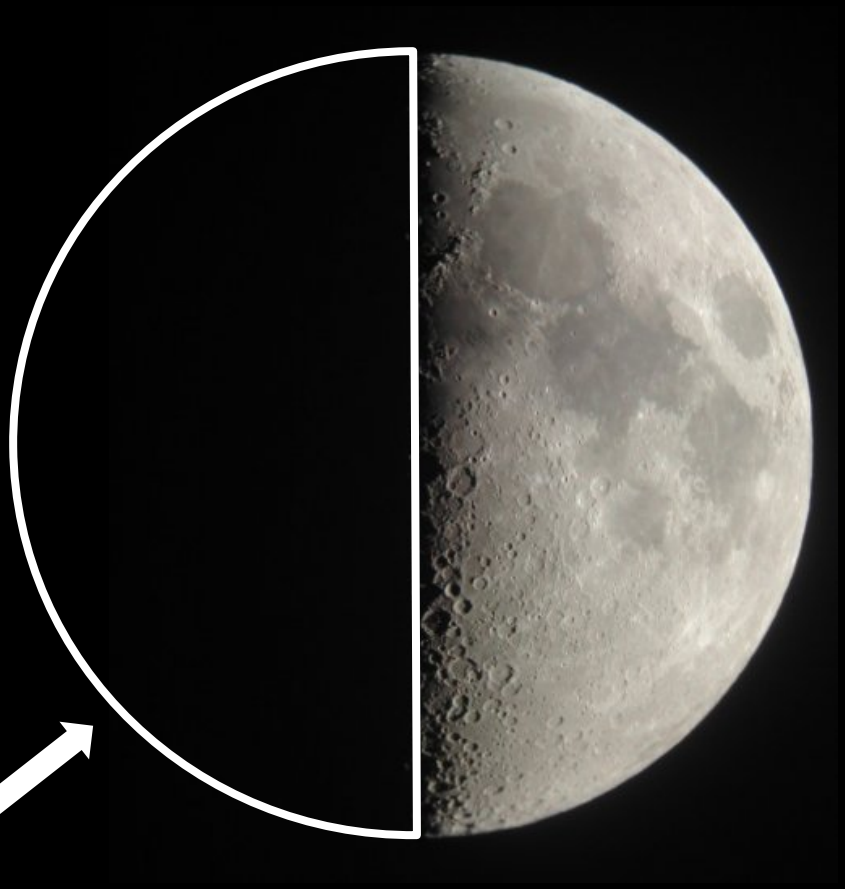
# Future directions

- Calling peaks is an “easy” problem
- The real question that we would like to address is:
  Can we identify genetic variants that have an impact on the “phenotype”?
- This requires combining genetic and epigenomic data
- Previously, there wasn’t enough data the use of ML approaches to answer this question
- But now, we are generating matched genetic and epigenomic data on 100s of individuals…
- Can we learn from this data and build a predictive model?

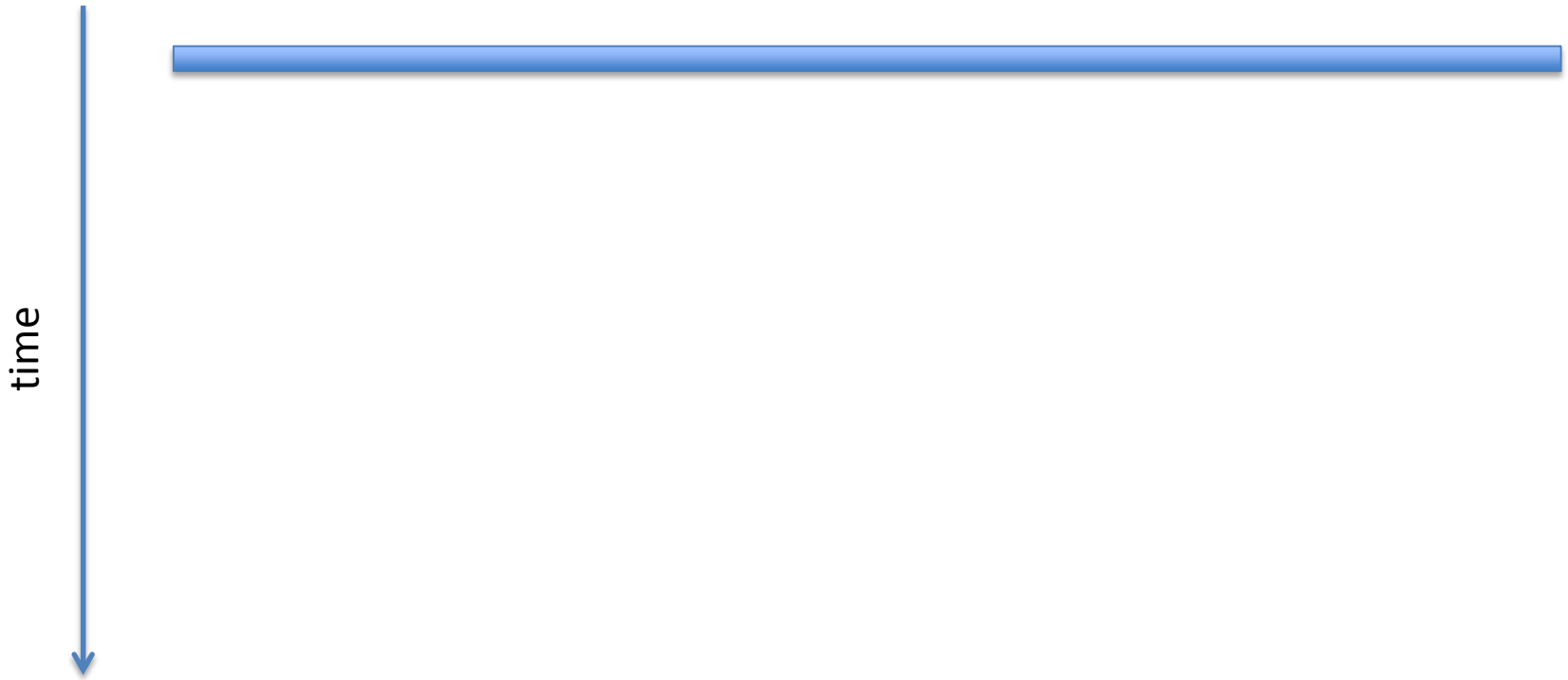
# Outline

- Applications of next-generation sequencing
- Functional genomics
- Example of machine learning approaches in functional genomics
- Role of transposable elements in gene regulation

# 30-50% of every mammalian genome is ‘Dark matter’



# Transposable Elements

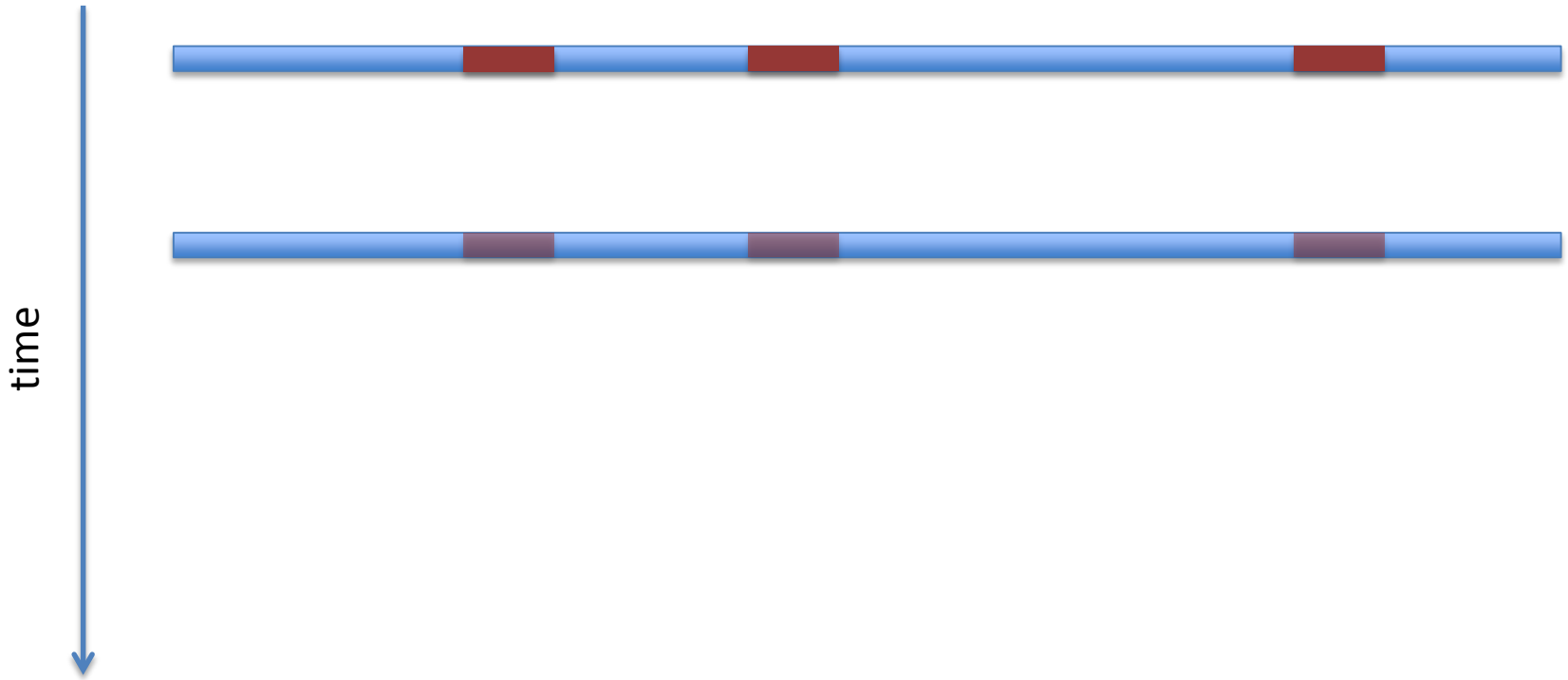


# Transposable Elements

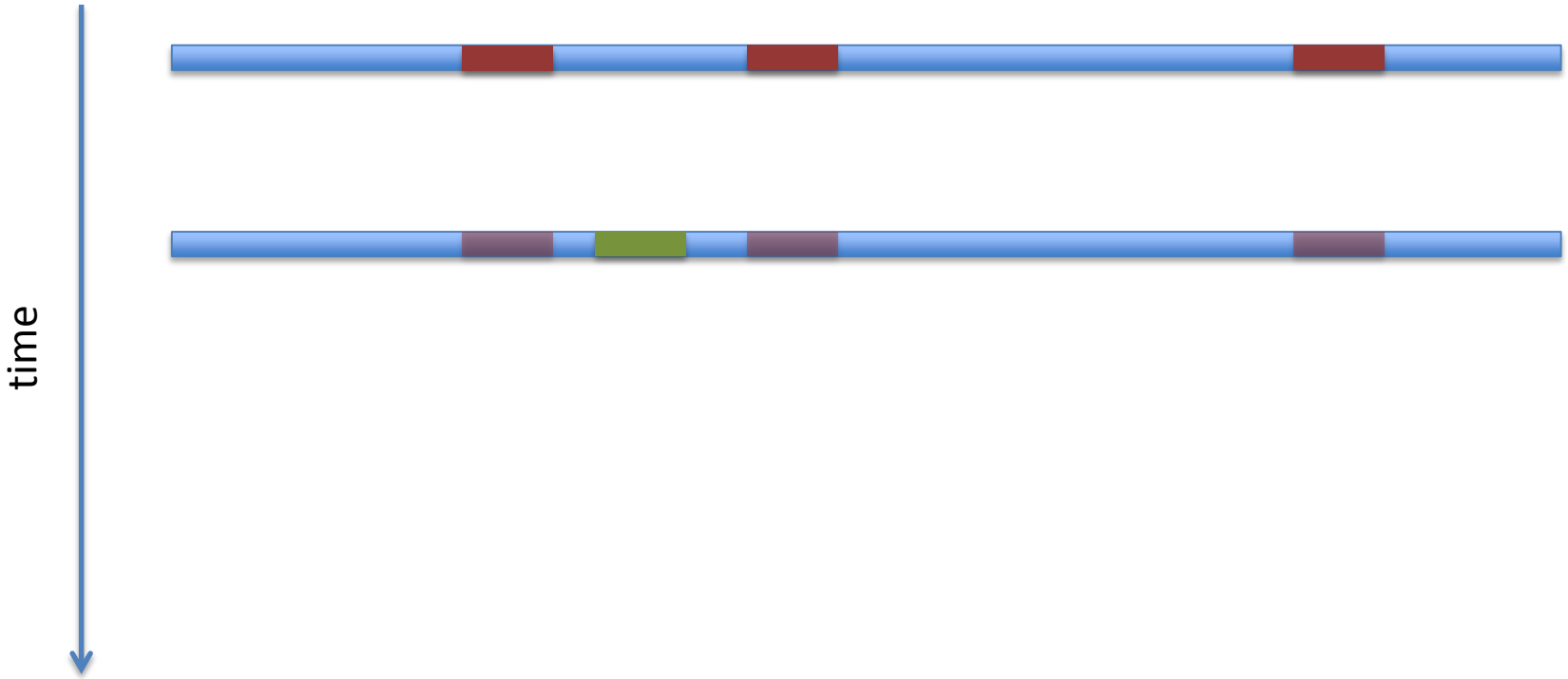
# Transposable Elements

# Transposable Elements

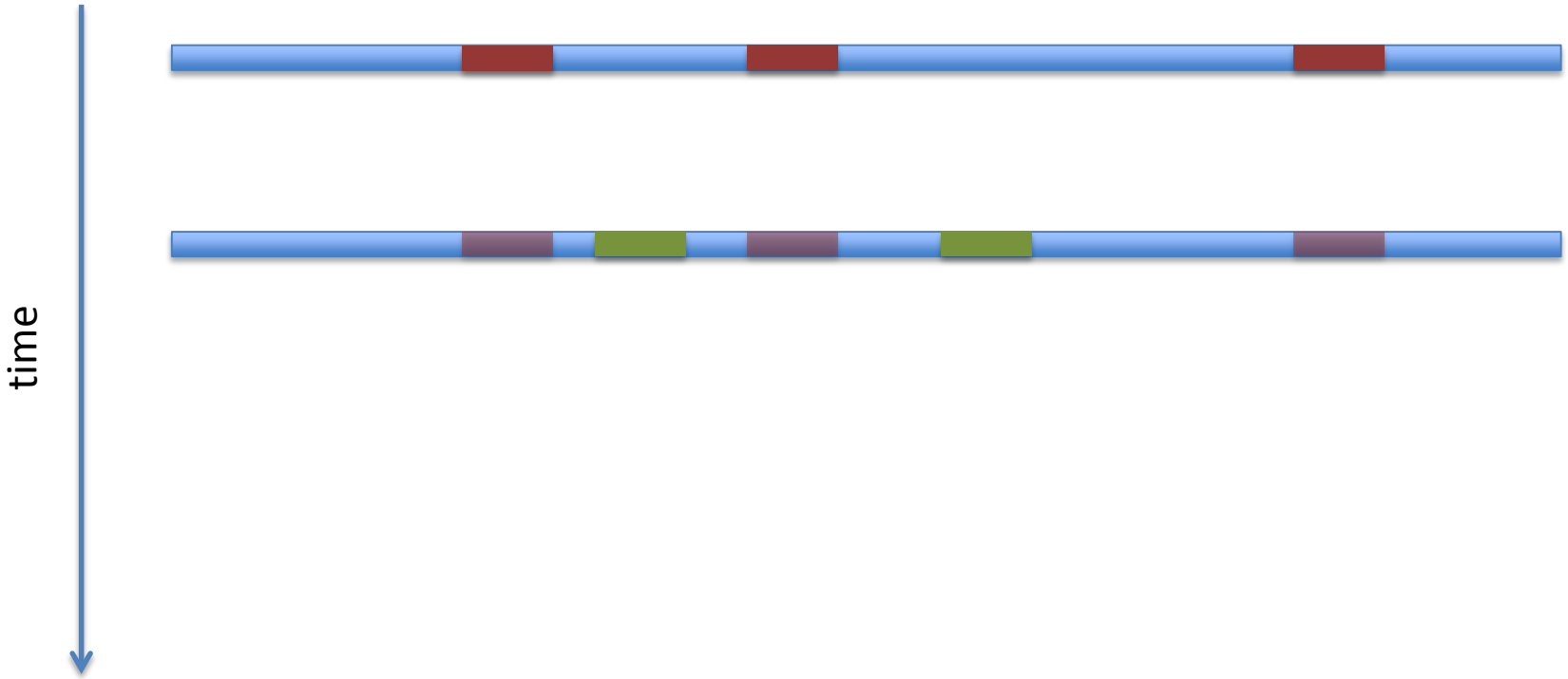
# Transposable Elements



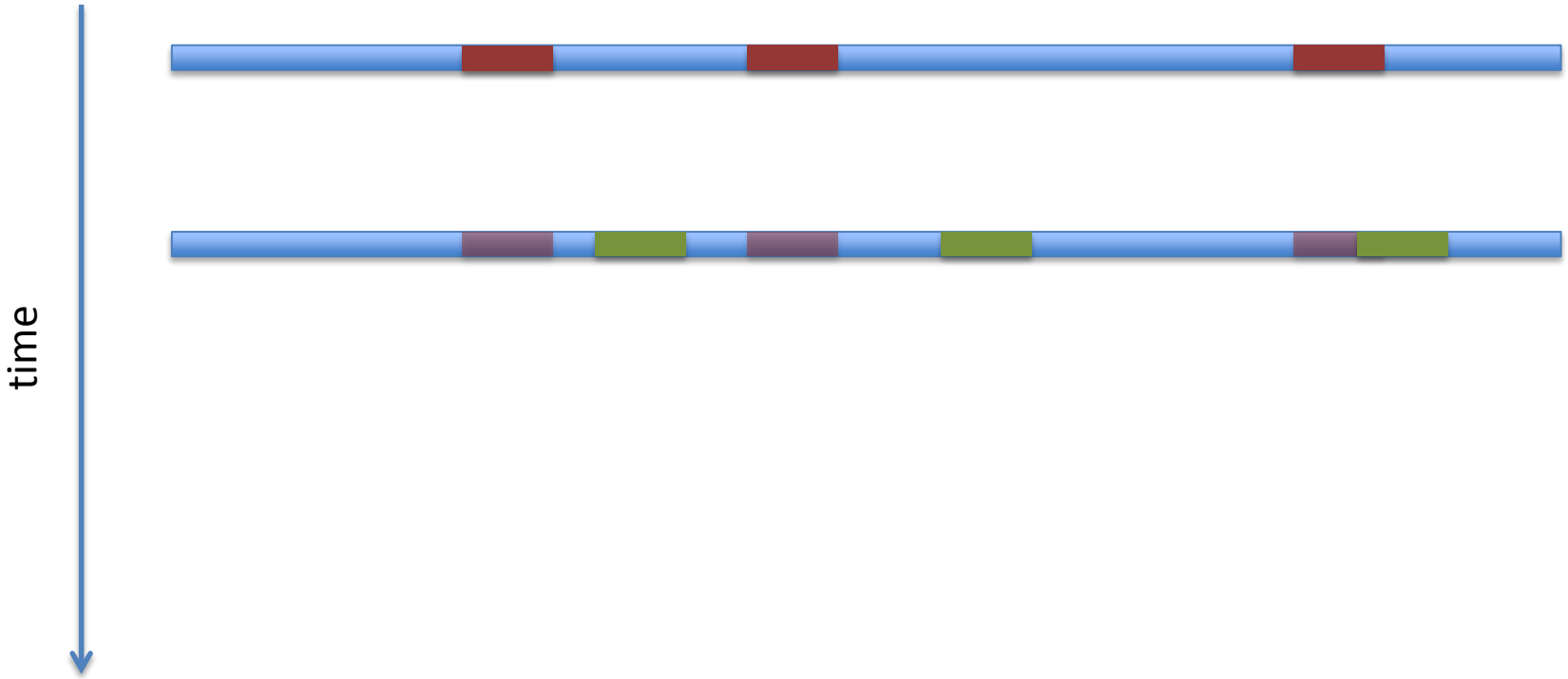
# Transposable Elements



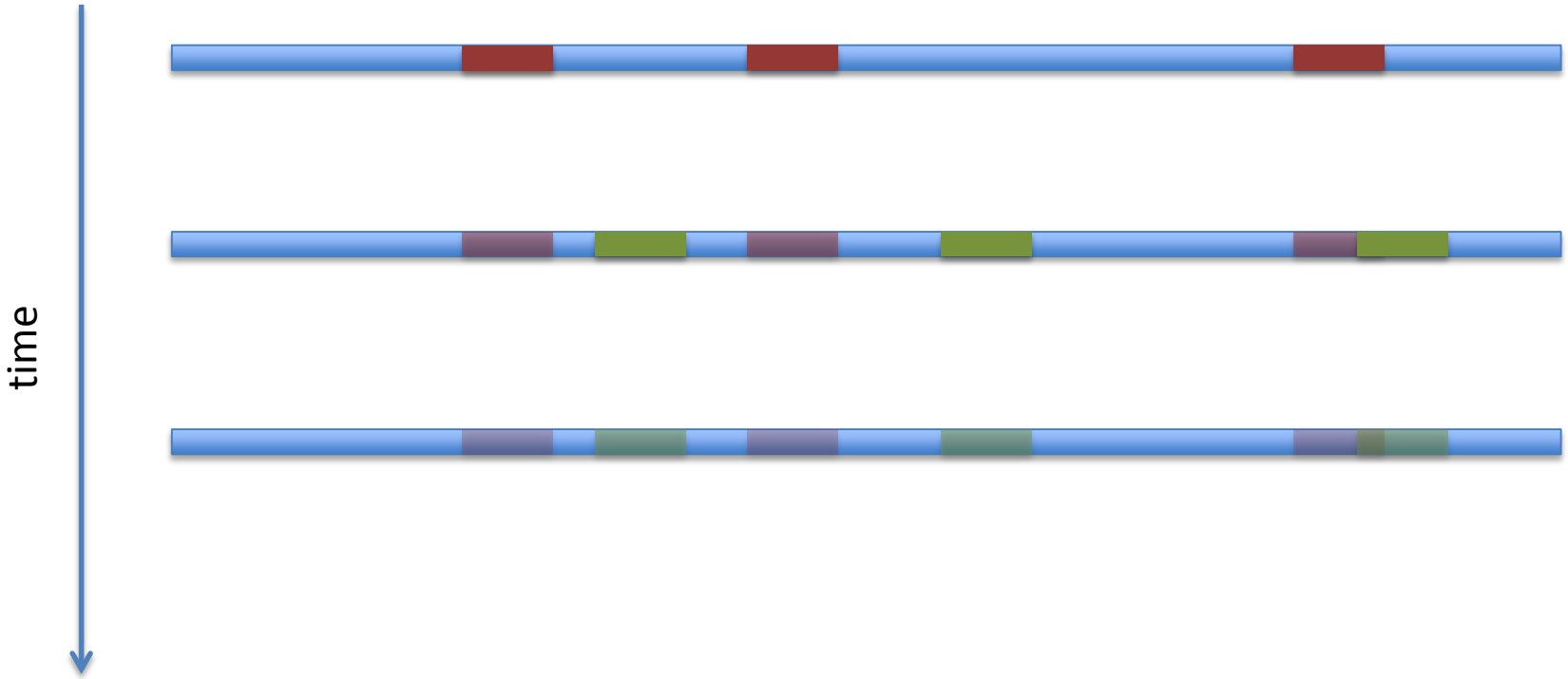
# Transposable Elements



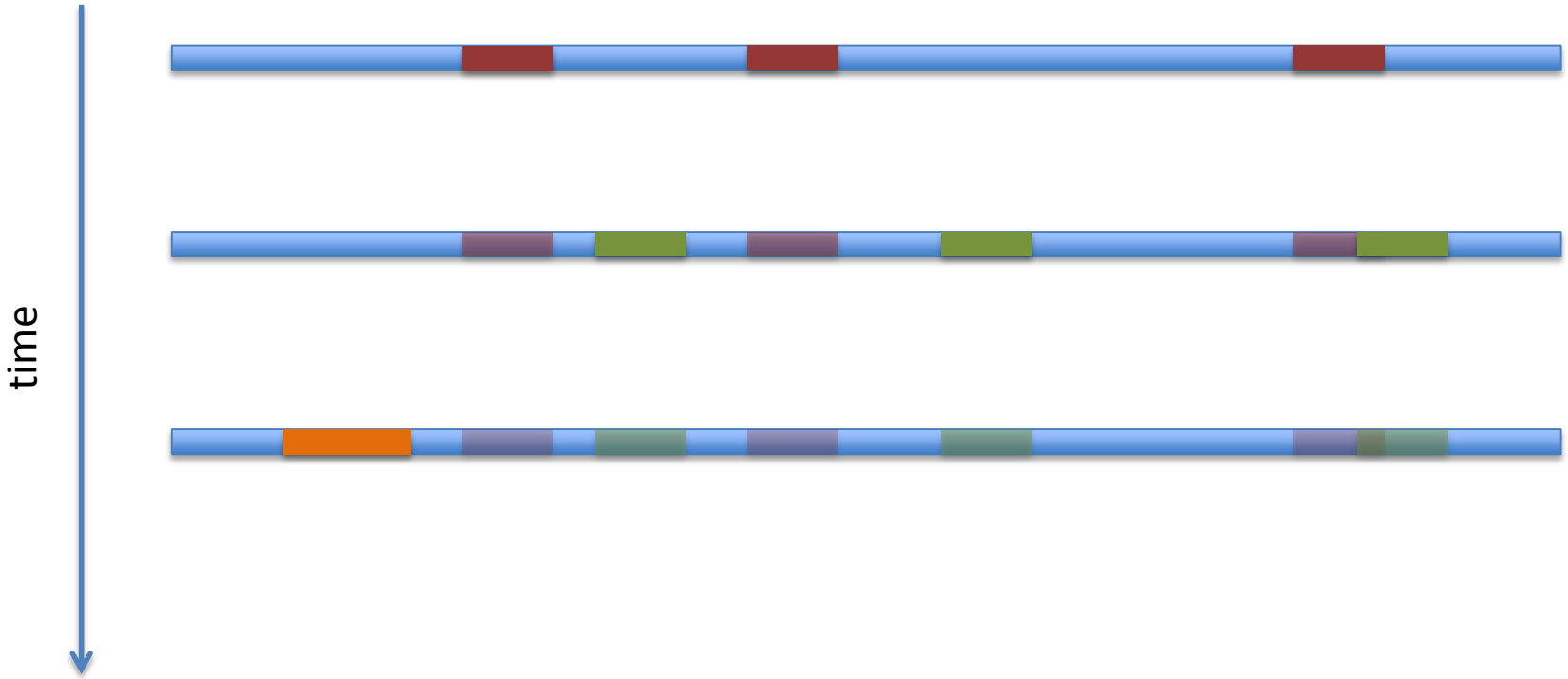
# Transposable Elements



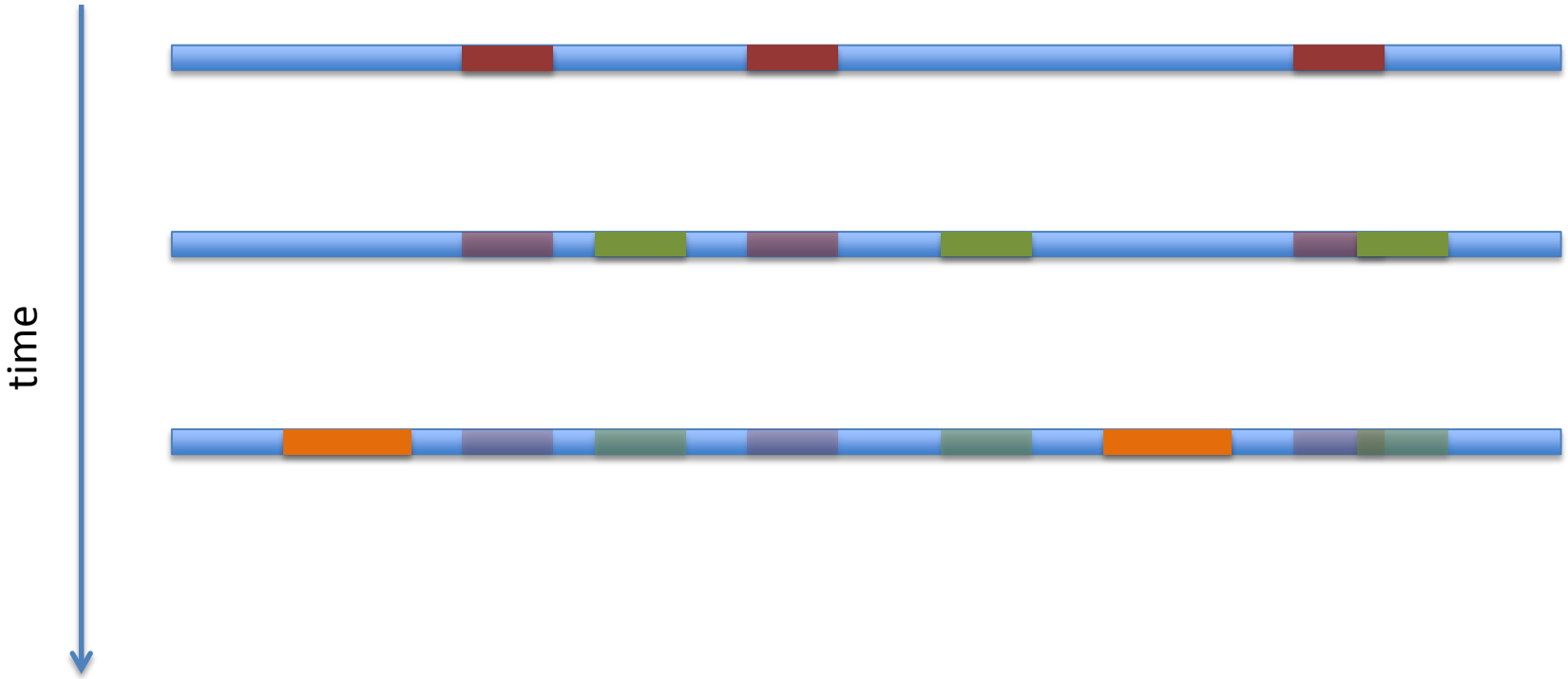
# Transposable Elements



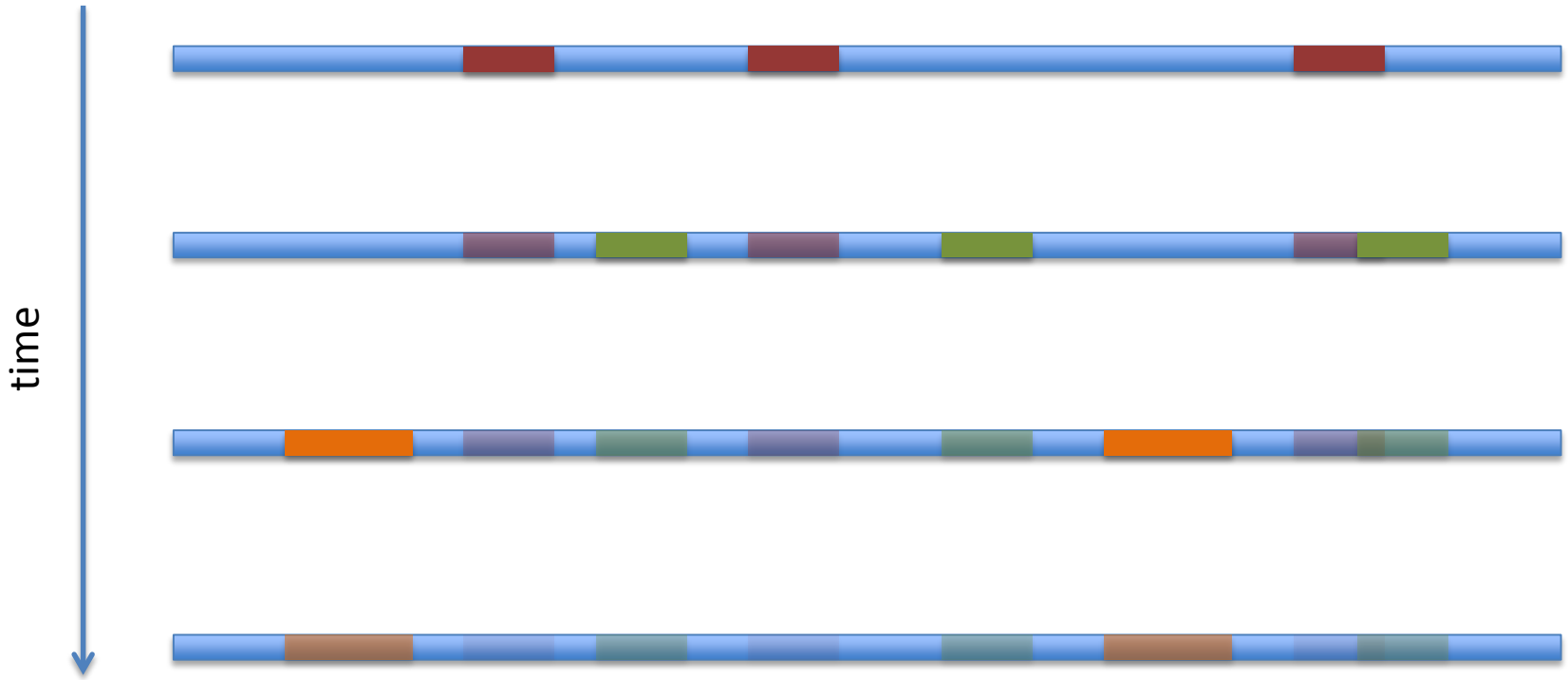
# Transposable Elements



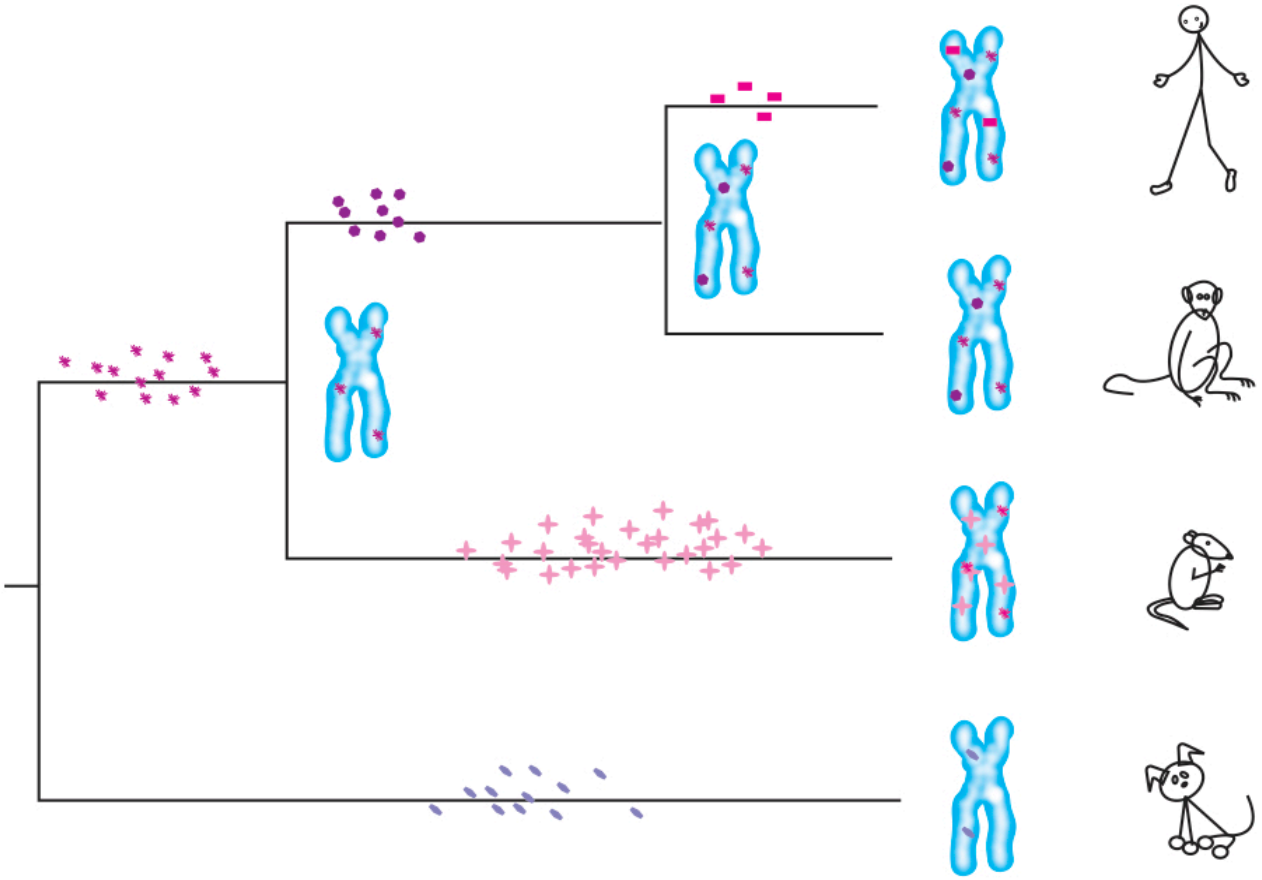
# Transposable Elements



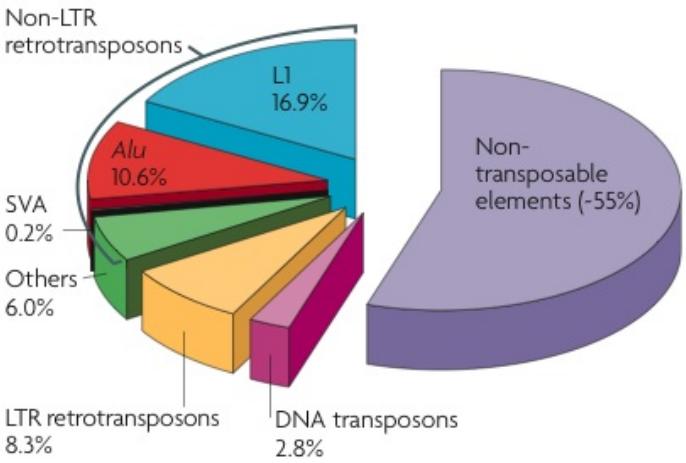
# Transposable Elements



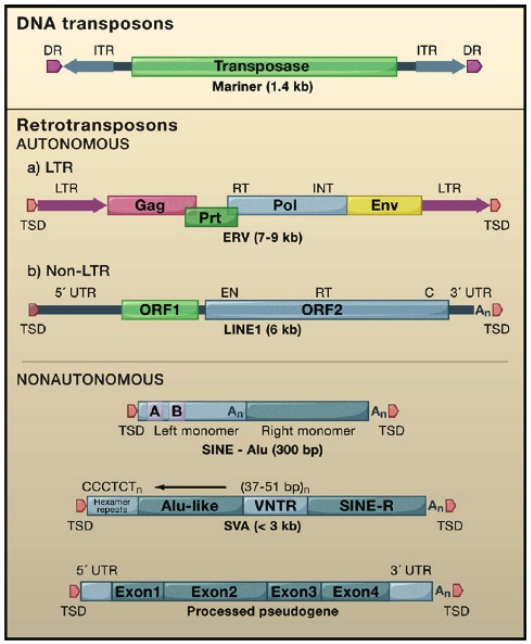
# TEs have changed mammalian genomes



# Transposable Elements (TEs) in the human genome



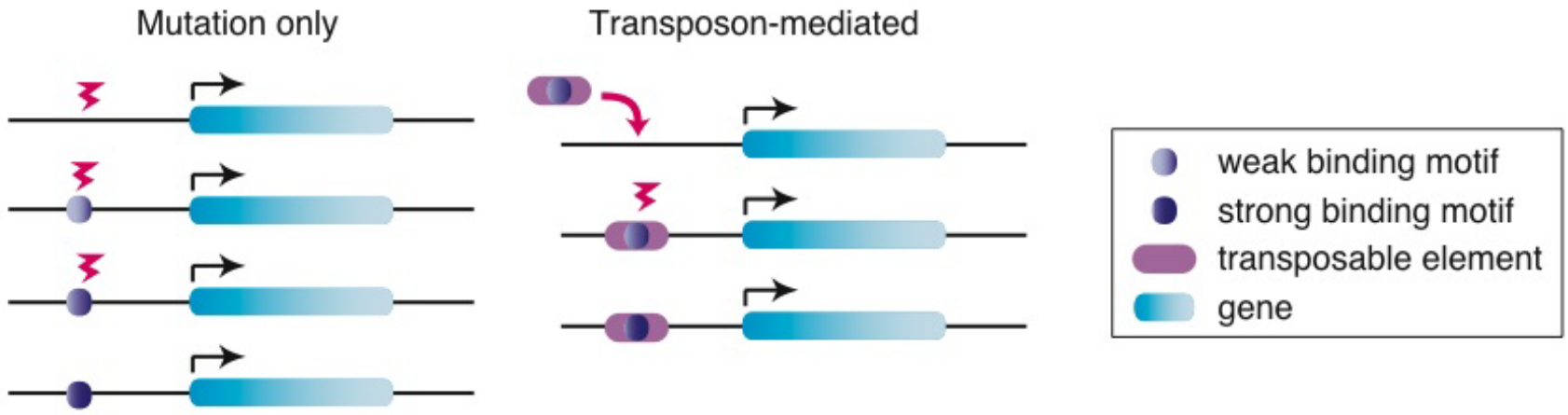

Cordeaux and Batzer,
*Nat Rev Genet*, 2009




Goodier and Kazazian,
*Cell* 2008

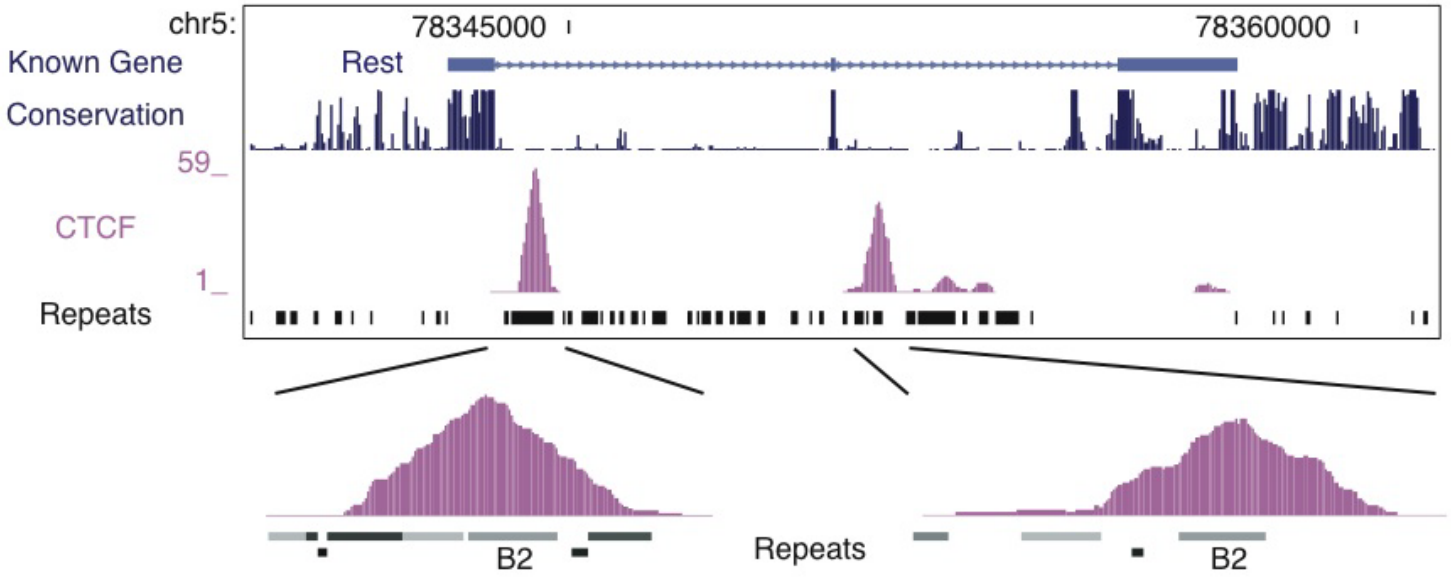
About 25% of the human genome consists of lineage-specific repeats.
For the mouse genome it's 30%.

# Two models for regulatory site acquisition



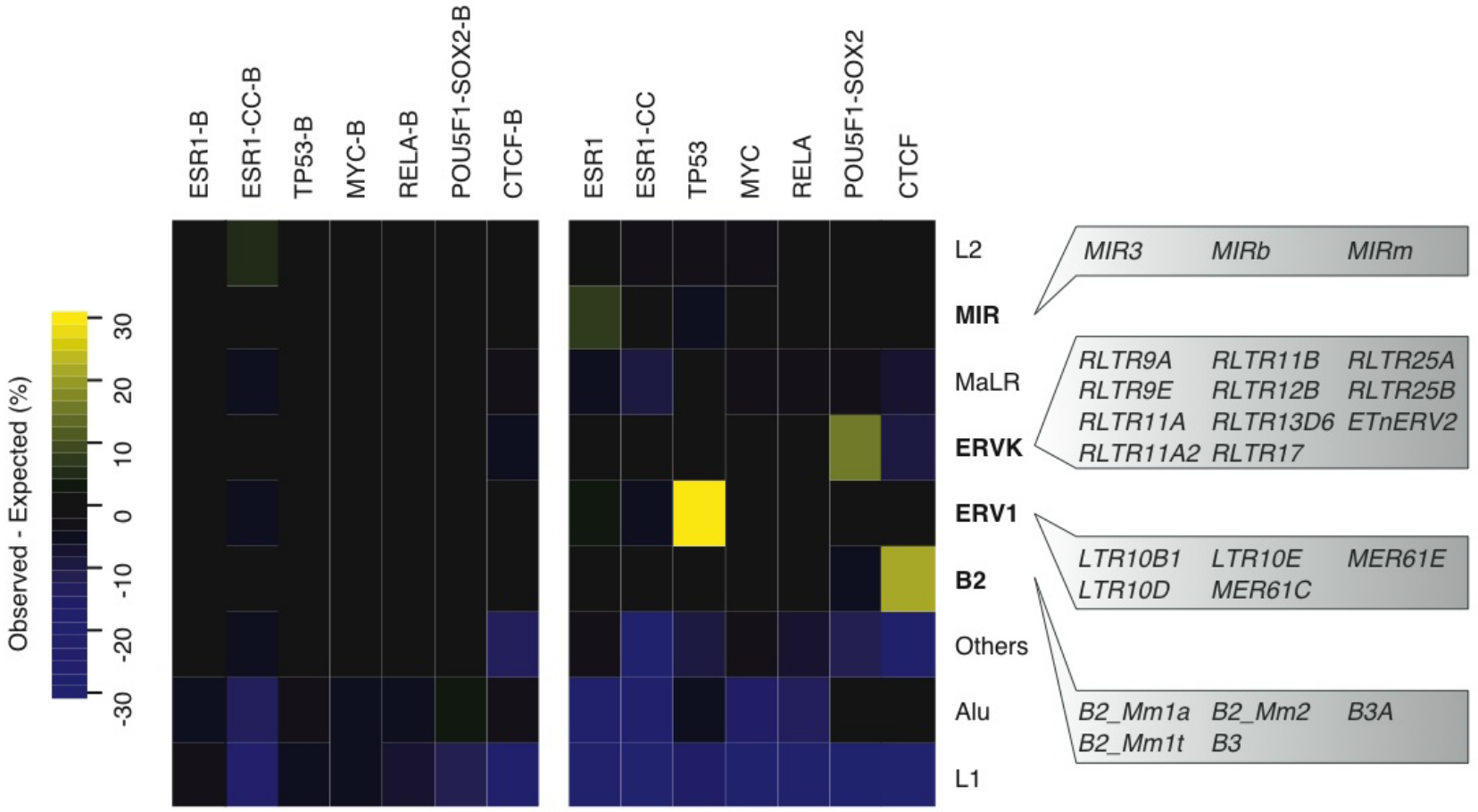

McClintock 1950, Britten & Davidson 1971, Brosius 1991, …

# Genome-wide occupancy maps reveal a strong association to repeats for many TFs



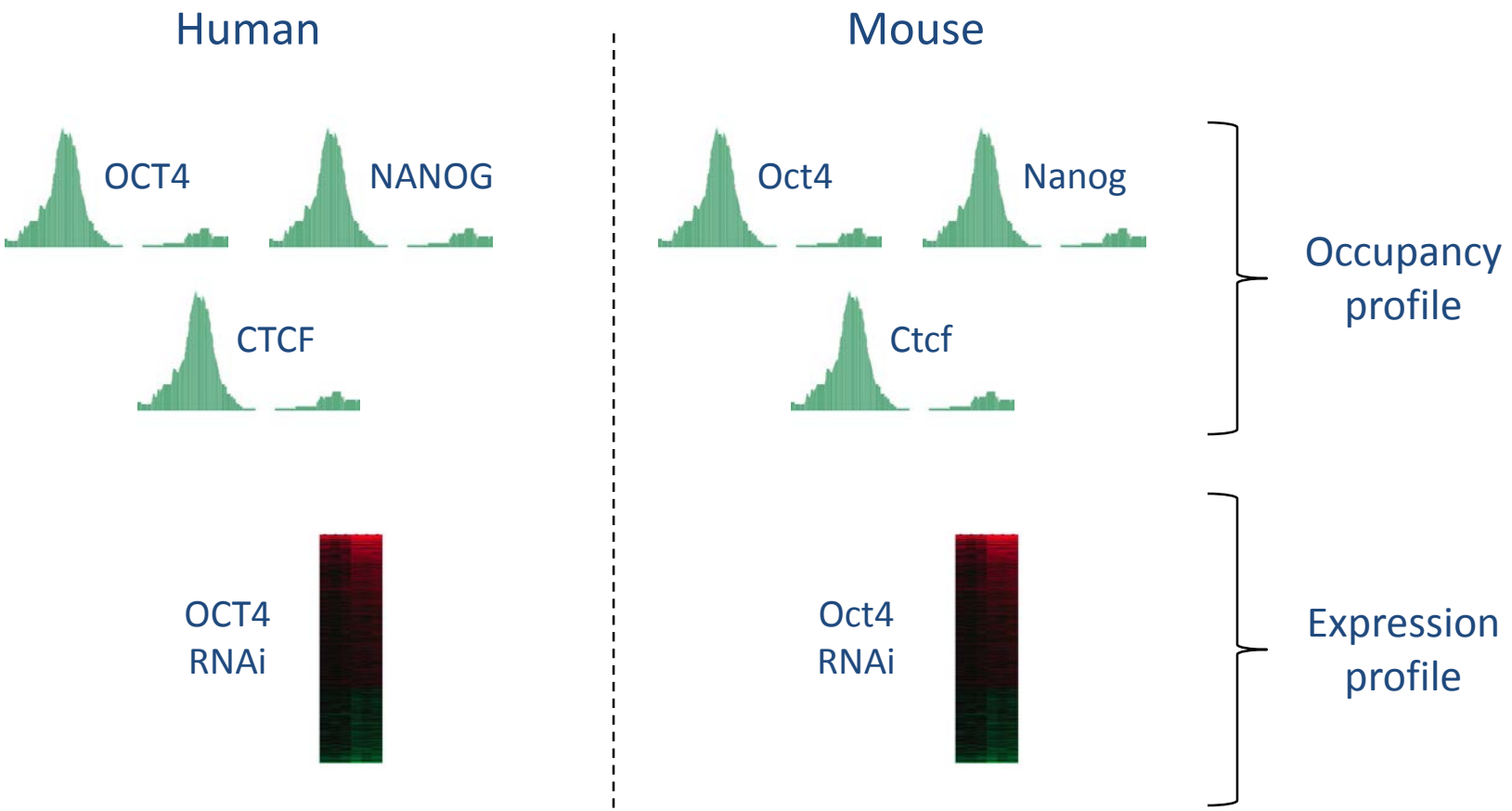

28% of CTCF sites
in B2 repeats

# Repeat Associated Binding Sites (RABS)




Bourque et al. *Genome Res*, 2008

# Human versus mouse ES cells

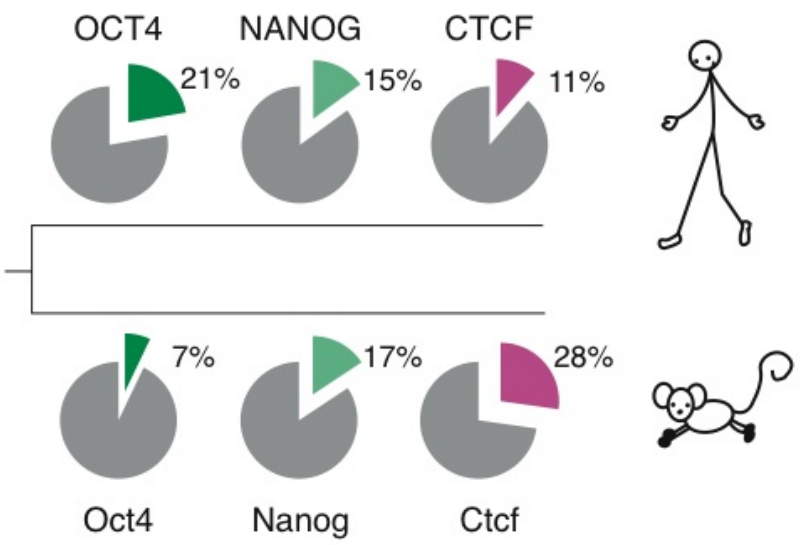


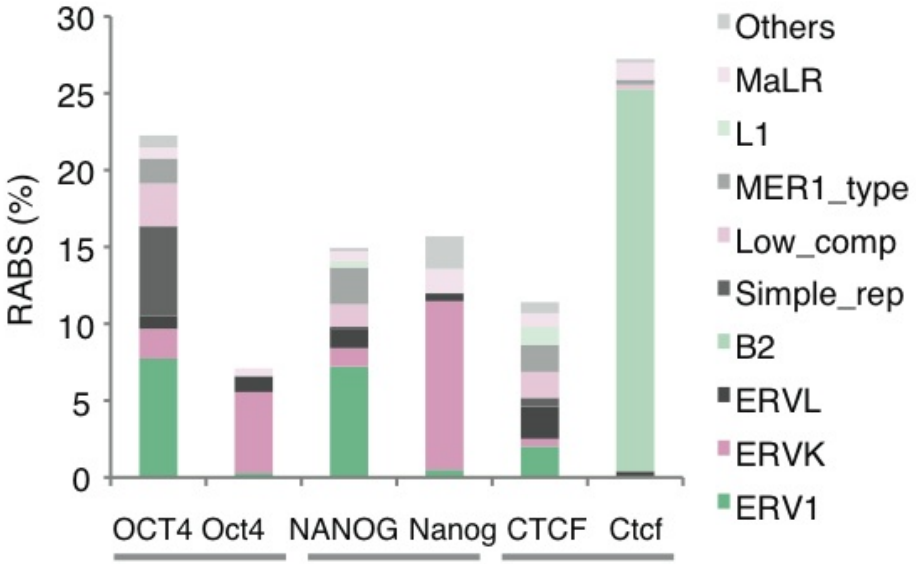

Collaboration with Huck-Hui Ng's lab at GIS

# OCT4 and NANOG RABS in human ES cells

| | Repeat Class | Repeat Family | Repeats | Bound Repeats | Expected | Ratio |
|---|---|---|---|---|---|---|
| NANOG | Low_complex | GC_rich | 13426 | 1215 | 624.6 | 1.9 |
| | **ERV1** | **LTR7** | **2357** | **813** | **32.7** | **24.9** |
| | MER1_type | MER5A | 35661 | 806 | 333.1 | 2.4 |
| | MER1_type | MER5A1 | 14897 | 529 | 136.9 | 3.9 |
| | L1 | L1PA2 | 4789 | 361 | 241.6 | 1.5 |
| | ERVK | LTR5_Hs | 610 | 340 | 27.8 | 12.2 |
| | MER1_type | MER1B | 5131 | 323 | 80.0 | 4.0 |
| | MER1_type | MER5B | 24784 | 321 | 213.8 | 1.5 |
| | ERV1 | MER4A1 | 2190 | 319 | 34.0 | 9.4 |
| | ERV1 | MER21C | 4915 | 235 | 84.8 | 2.8 |
| OCT4 | Simple_repeat | (TG)n | 54553 | 600 | 141.6 | 4.2 |
| | Simple_repeat | (CA)n | 54275 | 548 | 141.0 | 3.9 |
| | Low_complex | GC_rich | 13426 | 315 | 159.4 | 2.0 |
| | ERV1 | LTR9B | 757 | 255 | 3.1 | 81.7 |
| | MER1_type | MER5A | 35661 | 249 | 105.6 | 2.4 |
| | **ERV1** | **LTR7** | **2357** | **208** | **10.6** | **19.7** |
| | MER1_type | MER5A1 | 14897 | 204 | 47.4 | 4.3 |
| | Low_complex | GA-rich | 23936 | 206 | 80.9 | 2.5 |
| | ERV1 | LTR9 | 1962 | 176 | 14.9 | 11.8 |
| | Low_complex | CT-rich | 23713 | 181 | 77.2 | 2.3 |

# RABS in human and mouse ES cells

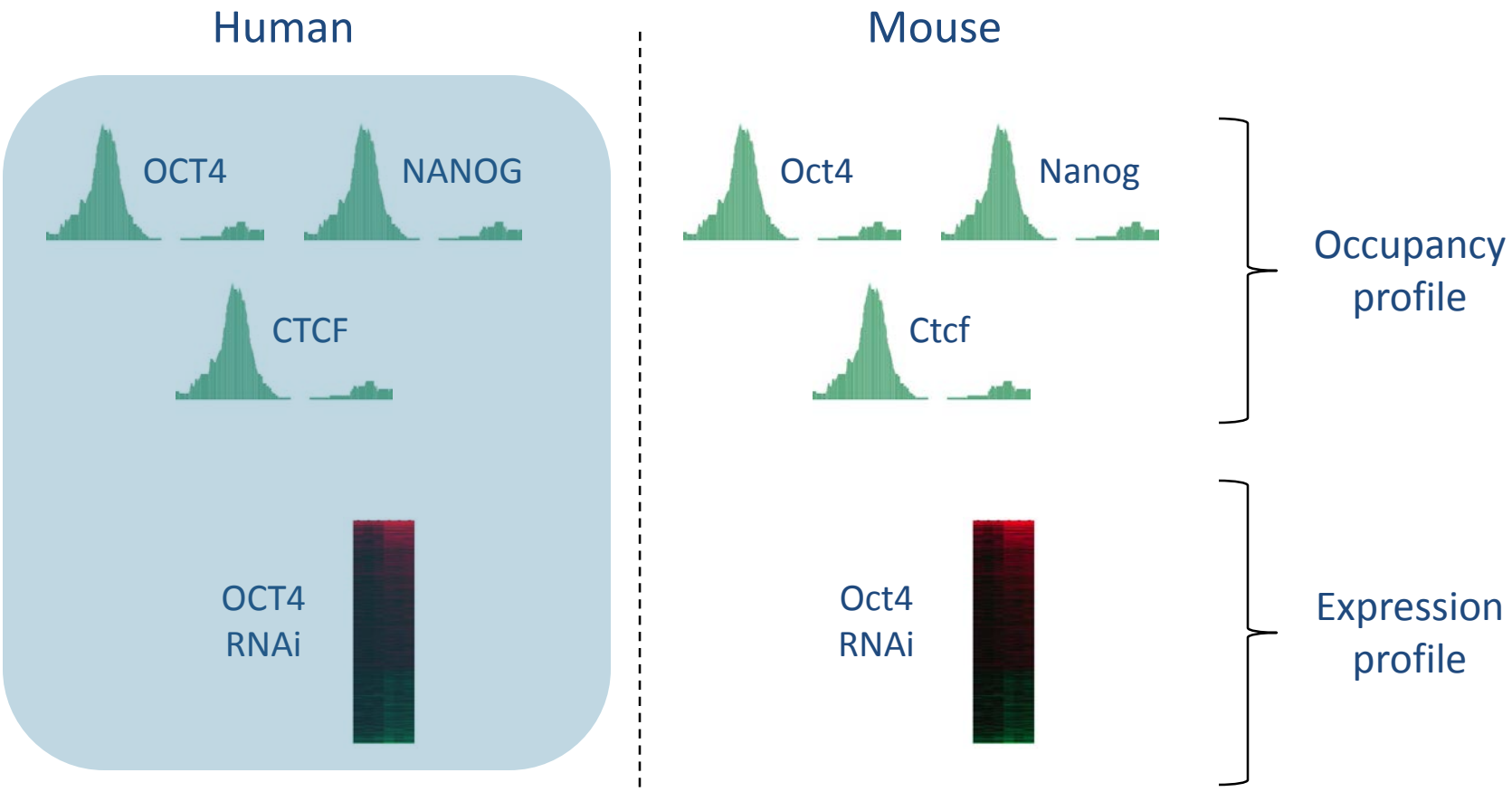


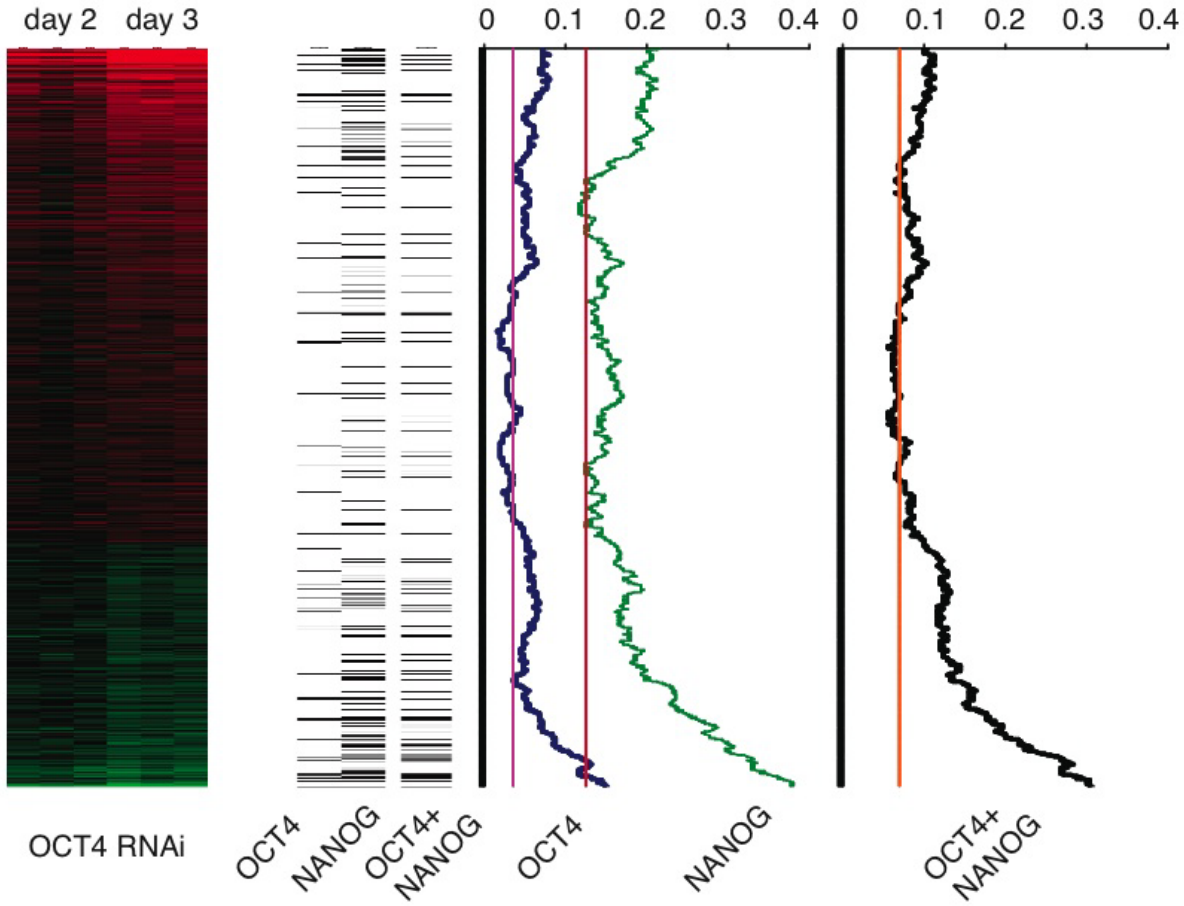

Different families of repeats have contributed a significant fraction of binding sites in both species.

Kunarso et al. *Nat Genet*, 2010

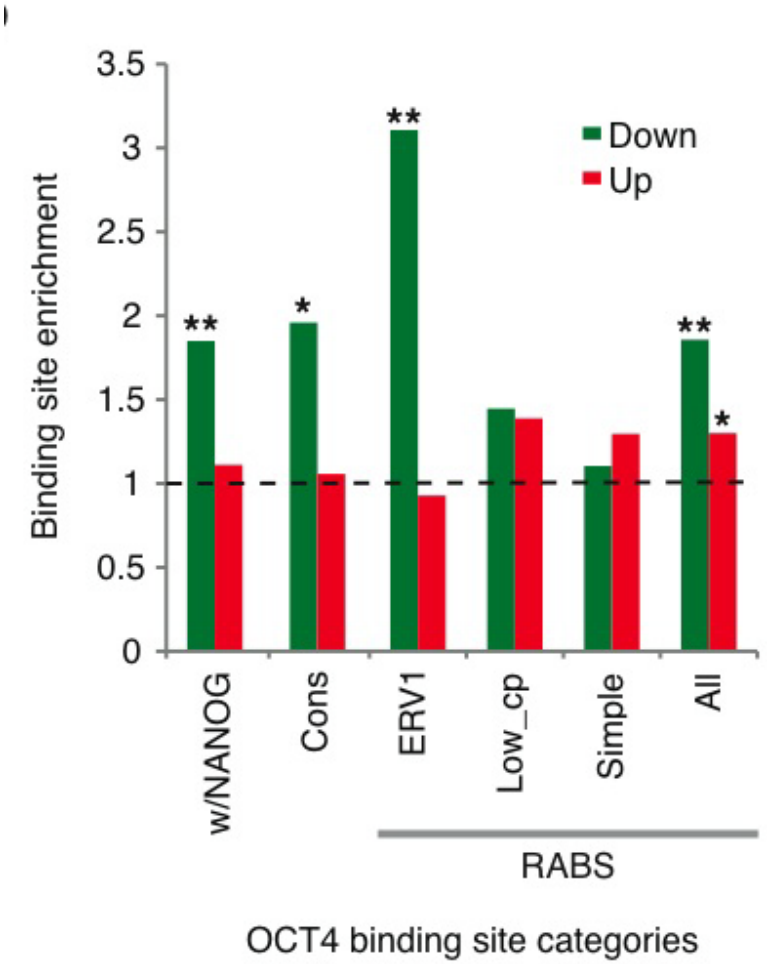
# Human versus mouse ES cells



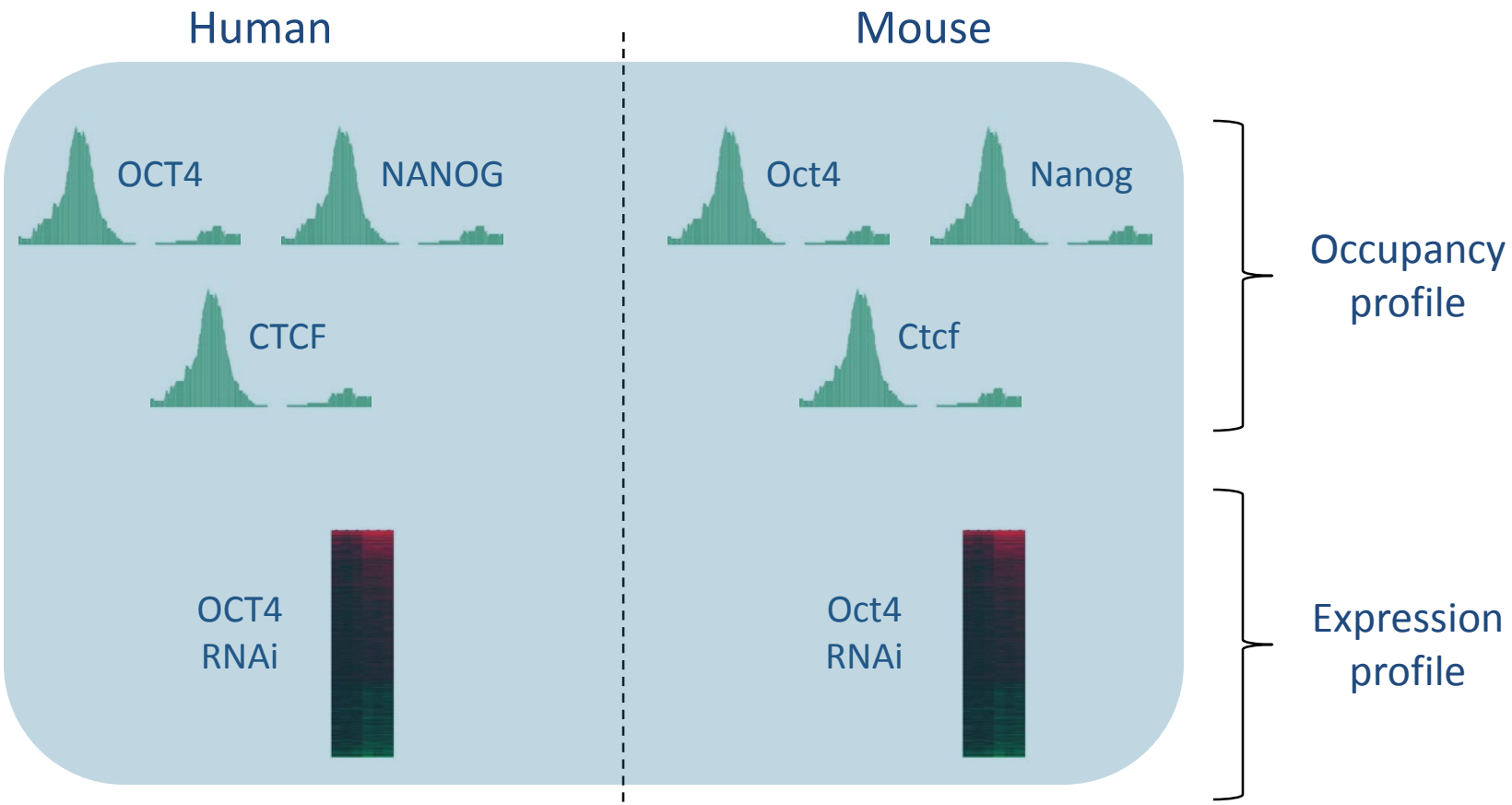
# OCT4 sites are associated with genes that are down regulated following RNAi treatment



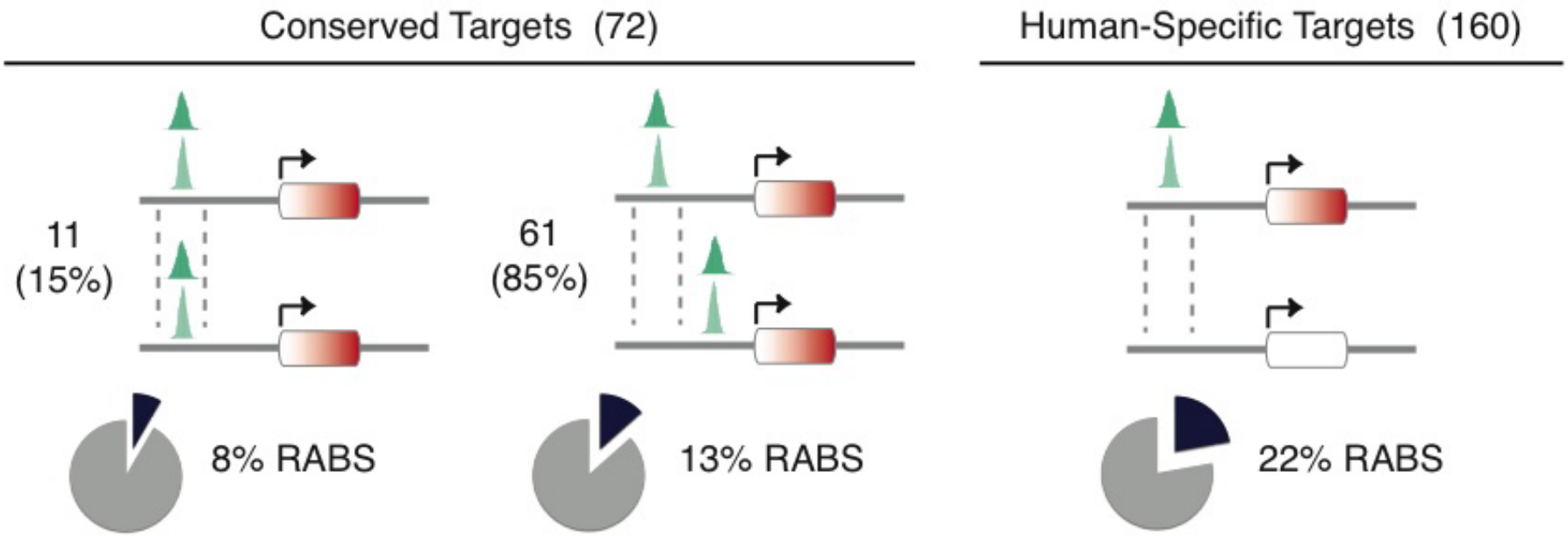
# OCT4 RABS are enriched in proximity of down regulated genes



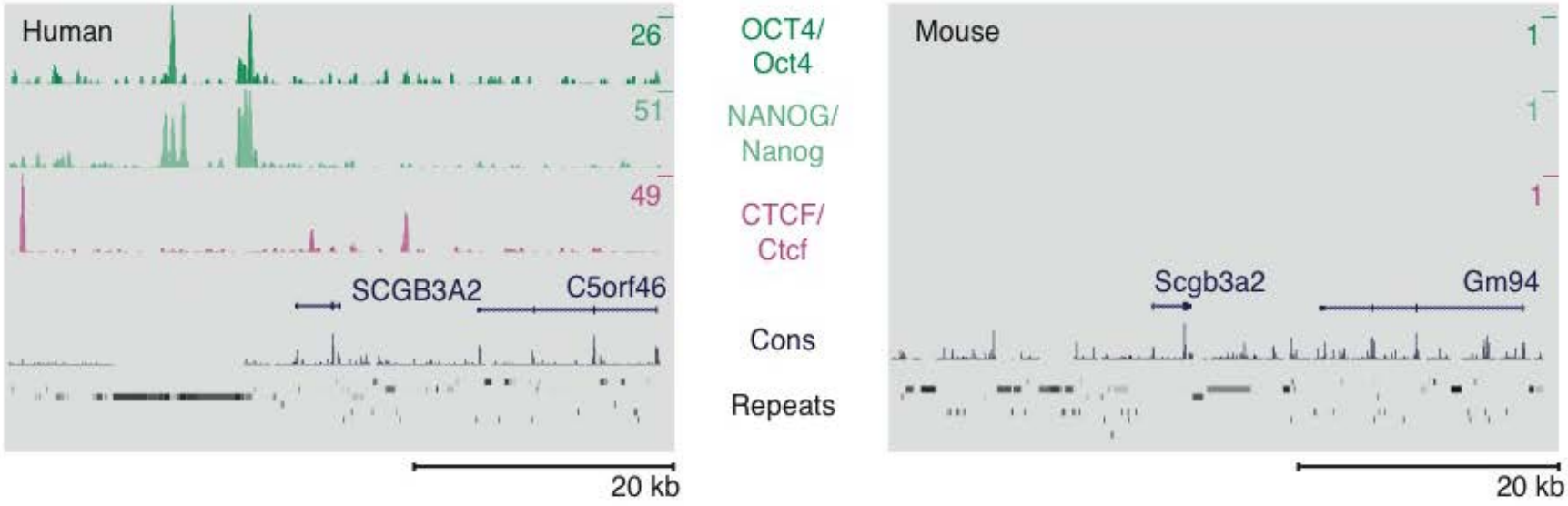
# Human versus mouse ES cells



# Binding sites around genes with conserved expression profiles



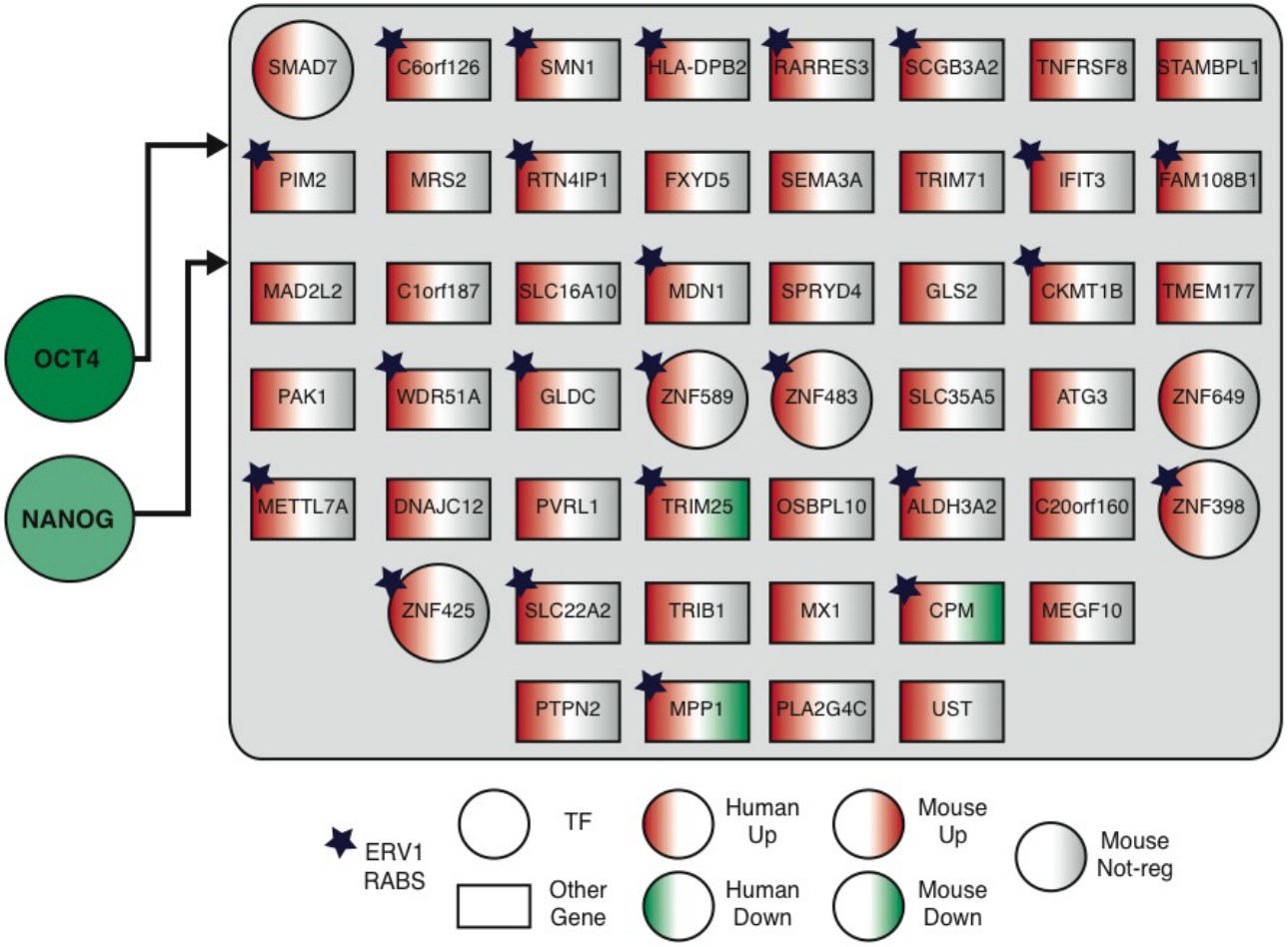
# Human-specific target driven by RABS



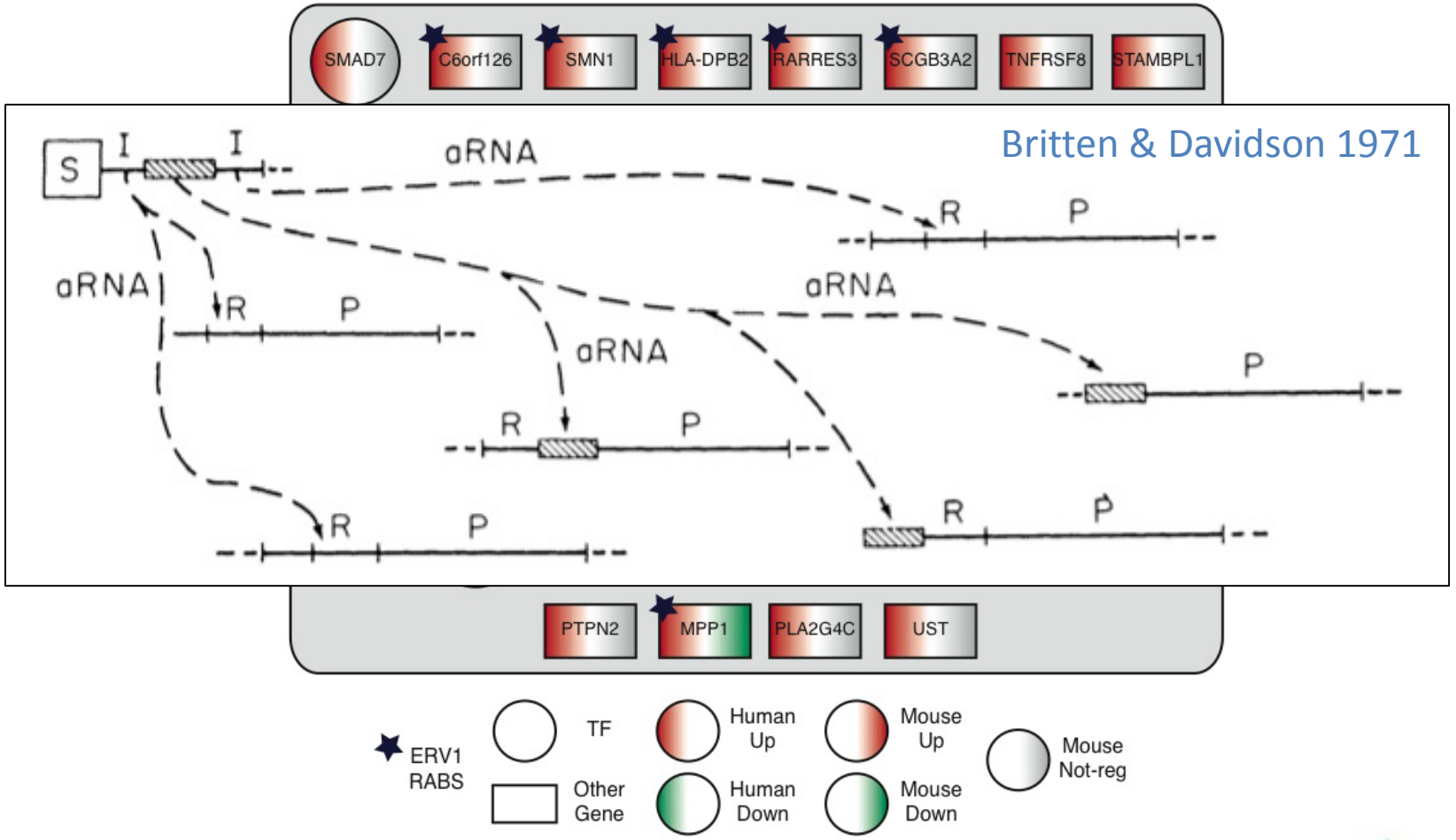

Human locus

Mouse locus

# Transposable Elements have Rewired the Core Regulatory Network of Human ES Cells



# Transposable Elements have Rewired the Core Regulatory Network of Human ES Cells



# Conclusions (Are not)

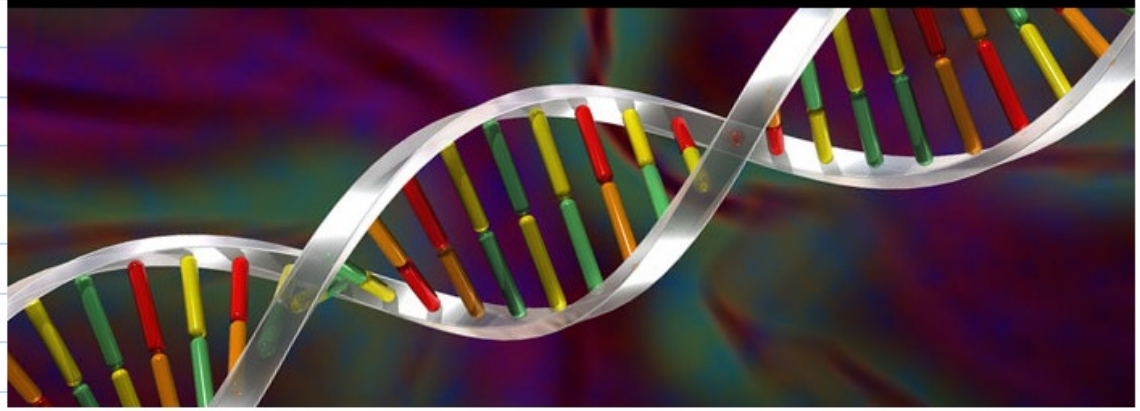
INSTITUTE for CREATION RESEARCH

Search icr.org GO

FREE SUBSCRIPTION · SIGN UP

- About Us
- Evidence for Creation
- Resources
- Search
- Daily Science Updates
- Days of Praise
- Store
- Give
- ICR Home

Articles



Follow Daily Science Updates

**Recent Articles**

- NASA's Ocean Currents Study Confirms Providential Care
- The Plague: Birth of a Killer
- Design in DNA: Flexibility Is Just Right
- Transposon Behavior Negates 'Selfish Gene' Theory
- Is Fossil Really a 'Game Changer'

## Gene 'Jumps' Serve a Purpose, Study Shows

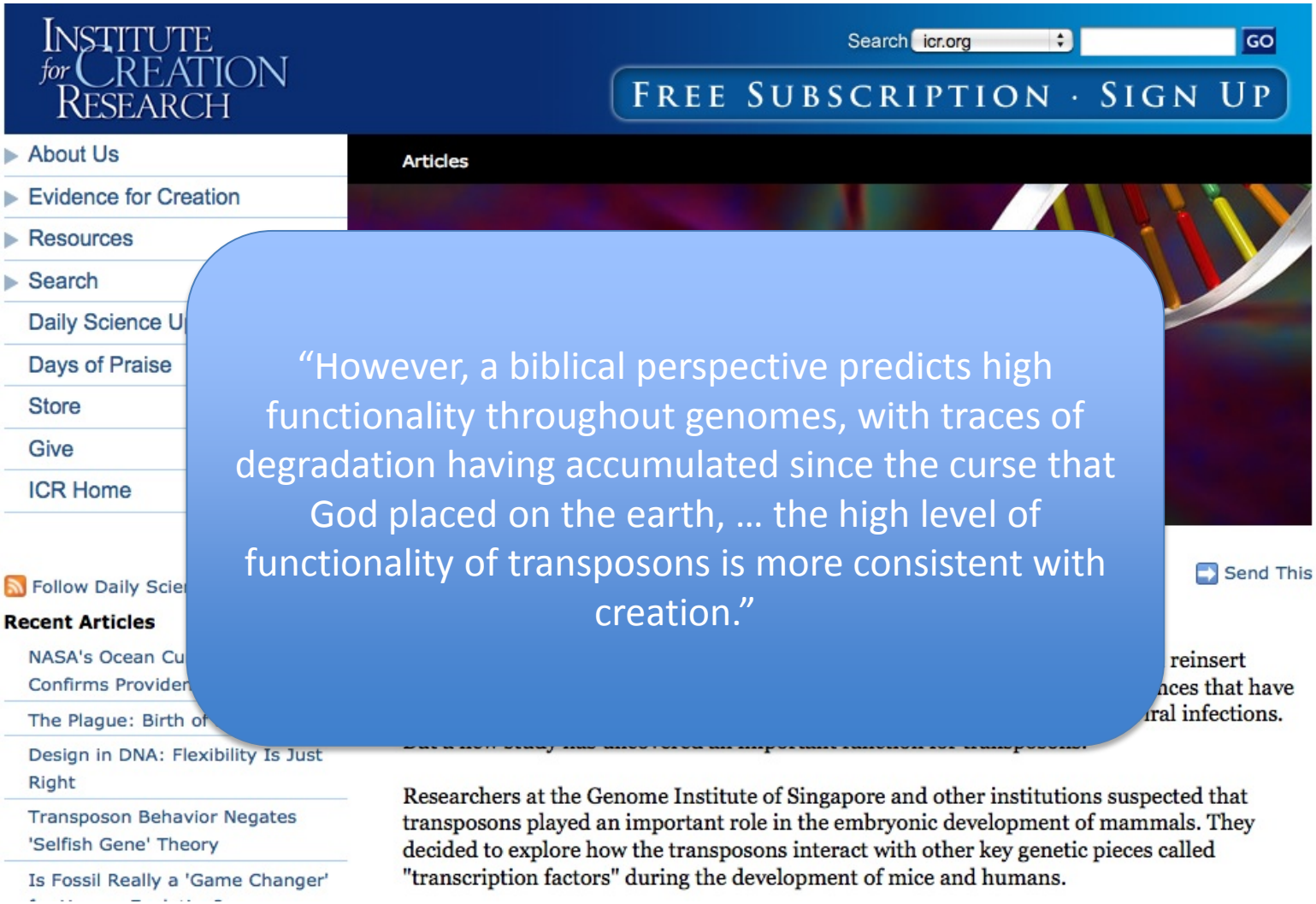
Send This

by Brian Thomas, M.S. *

In the tiny world of the cell, segments of DNA called transposons copy and reinsert themselves into the DNA. They eventually produce large repetitive sequences that have for many years been considered useless "junk" or remnants of ancient viral infections. But a new study has uncovered an important function for transposons.

Researchers at the Genome Institute of Singapore and other institutions suspected that transposons played an important role in the embryonic development of mammals. They decided to explore how the transposons interact with other key genetic pieces called "transcription factors" during the development of mice and humans.

# Conclusions (Are not)




"However, a biblical perspective predicts high functionality throughout genomes, with traces of degradation having accumulated since the curse that God placed on the earth, … the high level of functionality of transposons is more consistent with creation."

# Conclusions

- Up to 25% of new regulatory sites have been contributed by transposable elements in both human and mouse.
- Only 15% of the genes with conserved expression profiles have a conserved binding site.
- A number of human-specific targets have been wired into the core regulatory network of ES cells by repeats.

# Repeats and gene regulation

*Nat Genet 2011*

## Transposon-mediated rewiring of gene regulatory networks contributed to the evolution of pregnancy in mammals

Vincent J Lynch, Robert D Leclerc, Gemma May & Günter P Wagner

*Cell 2012*

## Waves of Retrotransposon Expansion Remodel Genome Organization and CTCF Binding in Multiple Mammalian Lineages

Dominic Schmidt,[1,2,6] Petra C. Schwalie,[3,6] Michael D. Wilson,[1,2] Benoit Ballester,[3] Ângela Gonçalves,[3] Claudia Kutter,[1,2] Gordon D. Brown,[1,2] Aileen Marshall,[1,5] Paul Flicek,[3,4,*] and Duncan T. Odom[1,2,4,*]

*PLoS Genet 2013*

OPEN ACCESS Freely available online

## The Majority of Primate-Specific Regulatory Sequences Are Derived from Transposable Elements

**Pierre-Étienne Jacques[1,2], Justin Jeyakani[1], Guillaume Bourque[3,4]***

**1** Computational and Systems Biology, Genome Institute of Singapore, Singapore, Singapore, **2** Département de Biologie, Université de Sherbrooke, Sherbrooke, Québec, Canada, **3** Department of Human Genetics, McGill University, Montréal, Québec, Canada, **4** McGill University and Génome Québec Innovation Center, Montréal, Québec, Canada

McGill

GenomeQuébec

# Acknowledgements


**Lab**

Galih Kunarso
Pierre-Etienne Jacques
LeeAnn Ramsay
Justin Jeyakani
Jean Monlong
Simon Girard
Toby Hocking
Eric Audemard

**EDCC team**

David Bujold
Bryan Caron
David Morais (UdeS)
Carol Gauthier (UdeS)
Alain Veilleux (UdeS)
ME Rousseau

**Analysis team**

Louis Letourneau
Mathieu Bourgey
Maxime Caron
Gary Lévesque
Robert Eveleigh
Francois Lefebvre
Johanna Sandoval
Pascale Marquis
Joel Fillon
Julien Tremblay

**IT team**

Terrance Mcquilkin
Marc-André Labonté
Genevieve Dancausse
Andras Frankel
Alexandru Guja

**Development team**

Nathalie Émond
Francois Cantin
Catherine Côté
Daniel Guertin
Louis Dumond Joseph
Francois Korbuly
Marc Michaud
Thuong Ngo
Francois Massé


Open positions for students and postdoc!


guil.bourque@mcgill.ca


McGill canarie CIHR IRSC Canadian Institutes of Health Research / Instituts de recherche en santé du Canada GenomeCanada compute • calcul CANADA GenomeQuébec