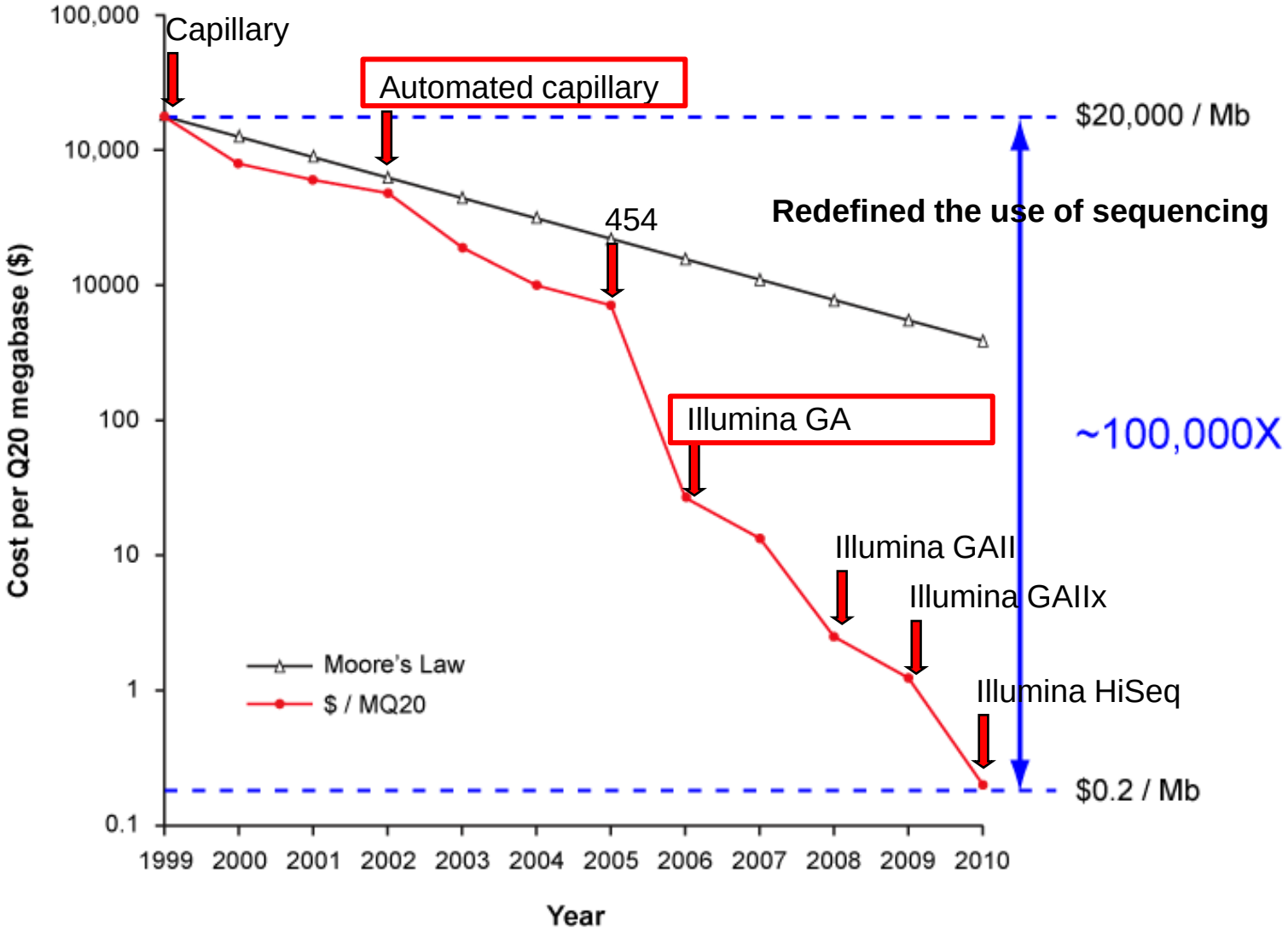
UMASS CCTS Seminar Series
The Bioinformatics Core
and
The Garber Lab

Manuel Garber

Nov 29th 2012

Plummeting cost of DNA sequencing



Estimated cost per Mb in large scale centers

Illumina Length, quality, yield continue to improve



HiSeq

Which has drastically change informatics, from

Achievements

- Reference human and other species genomes
- Reference transcriptome
- Reference variation map (HapMap)

Two step approach – sequence is expensive

- Sequence once, a reference
- Build arrays to explore samples

➢ Toolkits

- Affy/Agilent expression arrays
- Affy genotyping arrays
- Conservation databases

To DNA sequence being general purpose tool

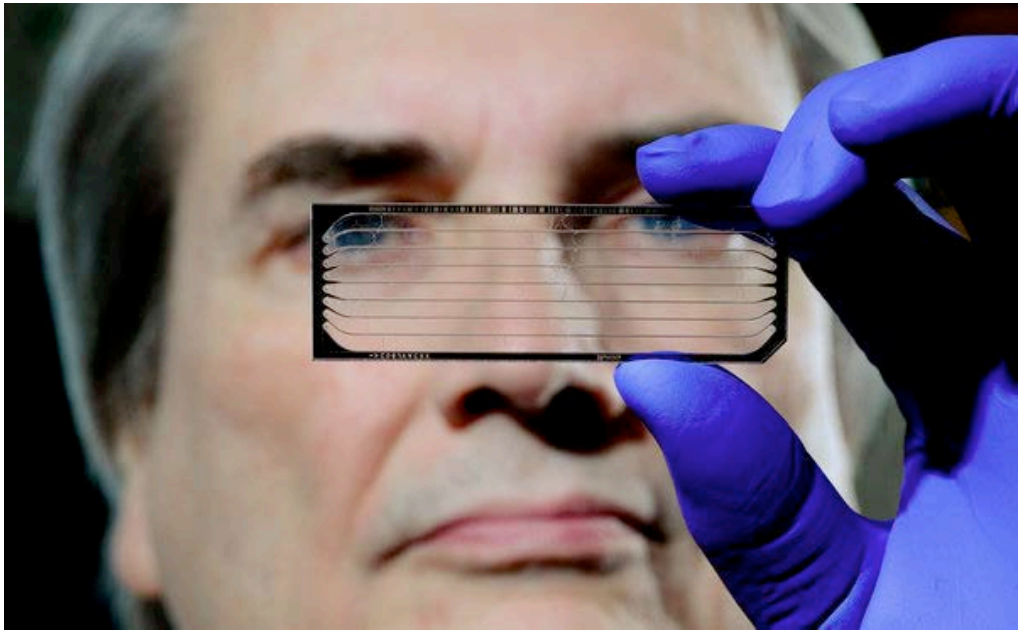
- Normal human variation and association studies
- Human genetics and gene discovery
- Cancer genomics
  - Map translocations, CNVs, structural changes
  - Profile somatic mutations
- Genome assembly
  - Virus
  - Bacteria/fungi
  - Mammals
- Transcriptomics
  - Comprehensive genome annotation
  - Expression dynamics (DGE)
  - micro- and small RNAs
  - Immunogenomics
- Epigenomics
  - Map histone modifications
  - Map DNA methylation
- Polymorphism/mutation discovery (SNPs and structural)
  - Bacteria
  - Genome dynamics/directed evolution
  - Exon (and other target) sequencing
  - Disease gene sequencing
- Ancient DNA (Neanderthal)
- Pathogen discovery
- Metagenomics
  - Human microbiome



The New York Times

DNA Sequencing Caught in Deluge of Data

November 30, 2011



"The result is that the ability to determine DNA sequences is starting to outrun the ability of researchers to store, transmit and **especially to analyze the data**."

"Data handling is now the bottleneck," said David Haussler, director of the center for biomolecular science and engineering at the University of California, Santa Cruz. "**It costs more to analyze a genome than to sequence a genome."**

Challenges of big data

- Large datasets are “easy”, “fast”, and “cheap” to generate. But they are time consuming and expensive to analyze.
- Looking at data is **crucial** to data analysis.
- Thinking about **how to analyze** the data is crucial.
- *Data analysis involves many **similar steps** with only variations on the approach*

My Goal is to have a core focused on enabling sequence data analysis.

Steps for analyzing NGS data

Biological insight

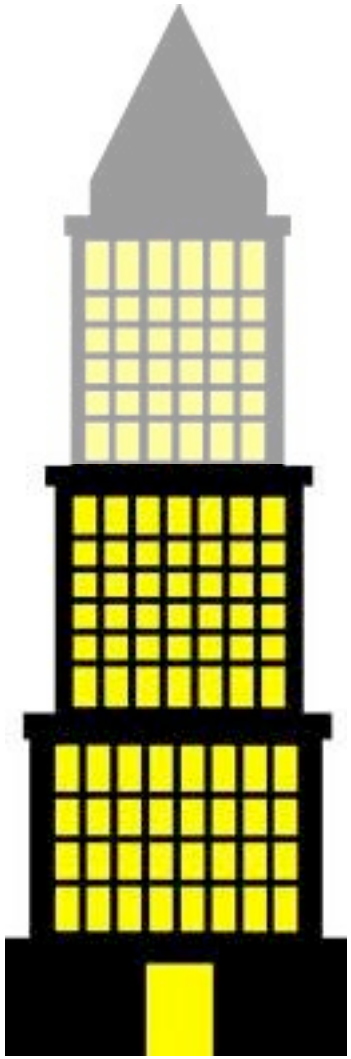
Finding signal, what are my differentially expressed genes, which peaks are in samples vs. controls etc.

Data processing
Sequences → Mapping, assembly, peak calling, transcript quantification → tables, browser tracks



Data setup: format and make data accessible to bioinformatics programs

Most of the technical, computationally intensive is generic and “core”



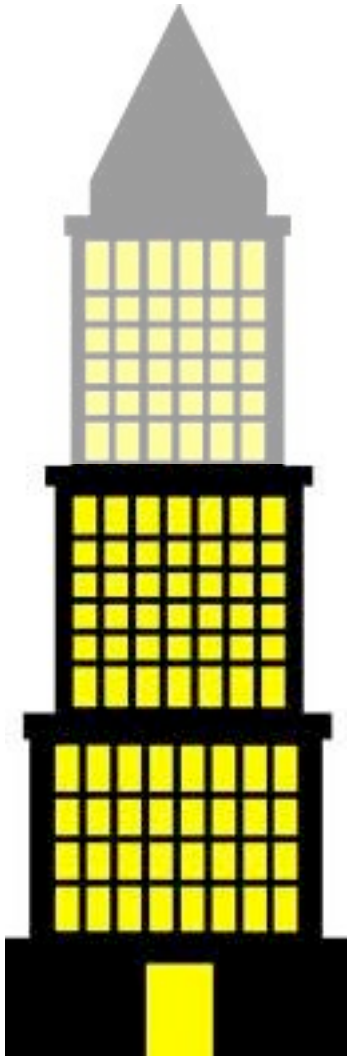
Biological insight

Finding signal, what are my differentially expressed genes, which peaks are in samples vs controls,etc

Data processing
Sequences → Mapping, assembly, peak calling, transcript quantification → tables, browser tracks

Data set up accessible to bioinformatics programs

Maximum impact



Biological Insight → Results, grants, papers

Provides standard analysis options and supports analysis developed at UMASS

Implements best of breed, UMASS specific data processing pipelines

Eases data access and manipulation

An informatics community around the core

Education:
- How to interpret processed data
- Statistics
- Visualization

Training sessions:
- Hands on training
- Consultation
- Q&A





Development
- New tools
- New pipelines

Community
- Regular presentations
- User discussion forum

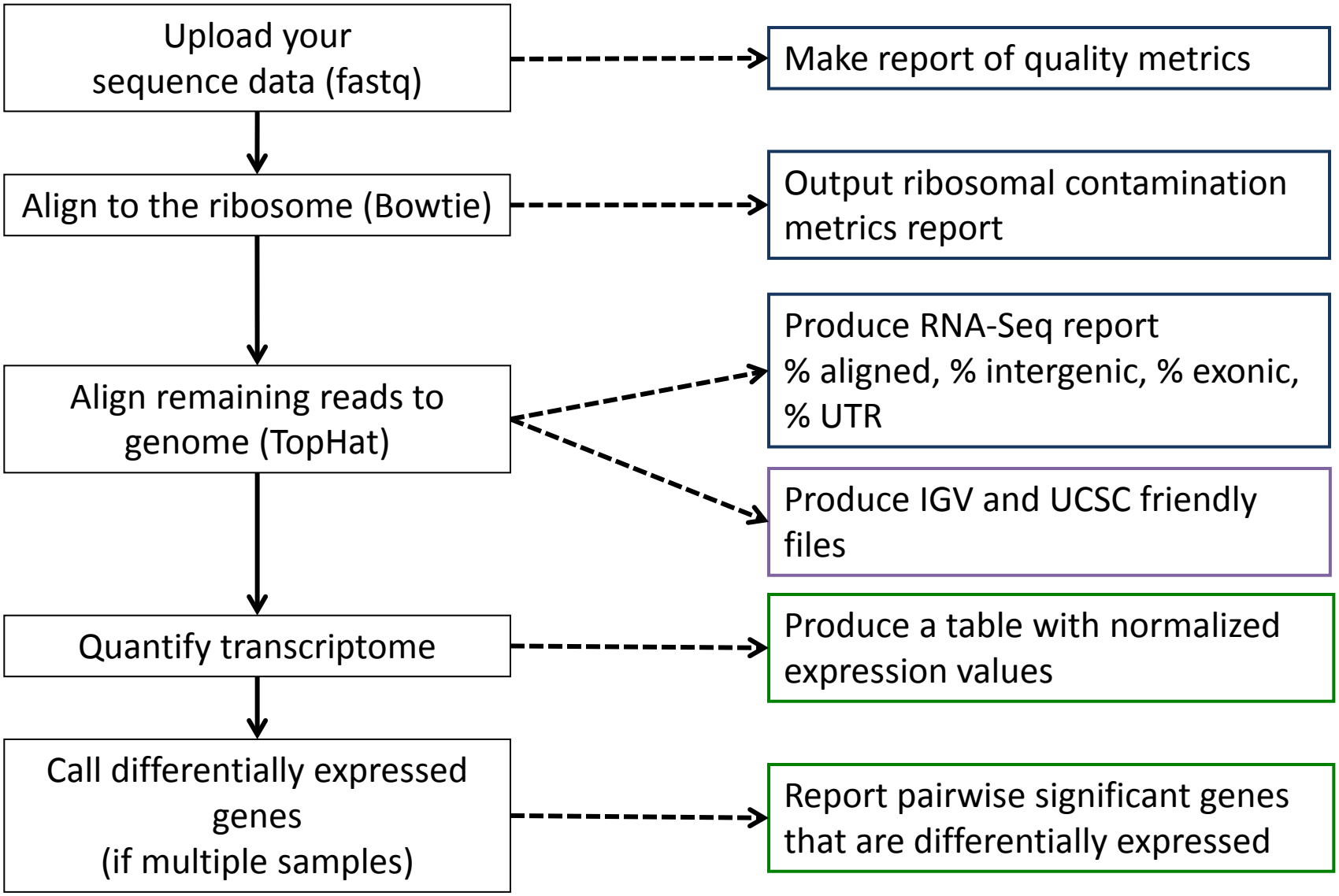
Current Progress

- Bioinformatics seminar (We've had 2 so far, all are welcomed)
  - Occurs the second Friday of every month at 11:00 am
  - Two short talks
    - A computational talk:
      - Algorithm
      - Pipeline
      - Method
    - A data centric analysis:
      - An integrative analysis
      - A preliminary analysis that looks for feedback
      - Data from novel techniques
- Redesigned Website
  - Dynamic website with documentation on pipelines
  - Hot-ticketing system when users experience problems with core-supported pipelines
- One Hire and one more being recruited
- Listserv for online discussions: bioinfo@list.umassmed.edu

Short Term Goals

- Integrate user Galaxy and similar tools with the HPC cluster
- Implement standard pipelines for
  - RNA sequencing analysis
    - including small RNA
  - ChIP sequencing analysis
  - Variant calling from deep sequence or exome data
- Make this pipelines available through Galaxy so that most users can take advantage of them

A typical pipeline (e.g. RNA-Seq)



Pipelines will be available in the HPCC cluster

- For those unafraid of UNIX, pipelines will be available to execute from the command line:
  - Write a script

```
#!/bin/bash
# Now set up some environment stuff

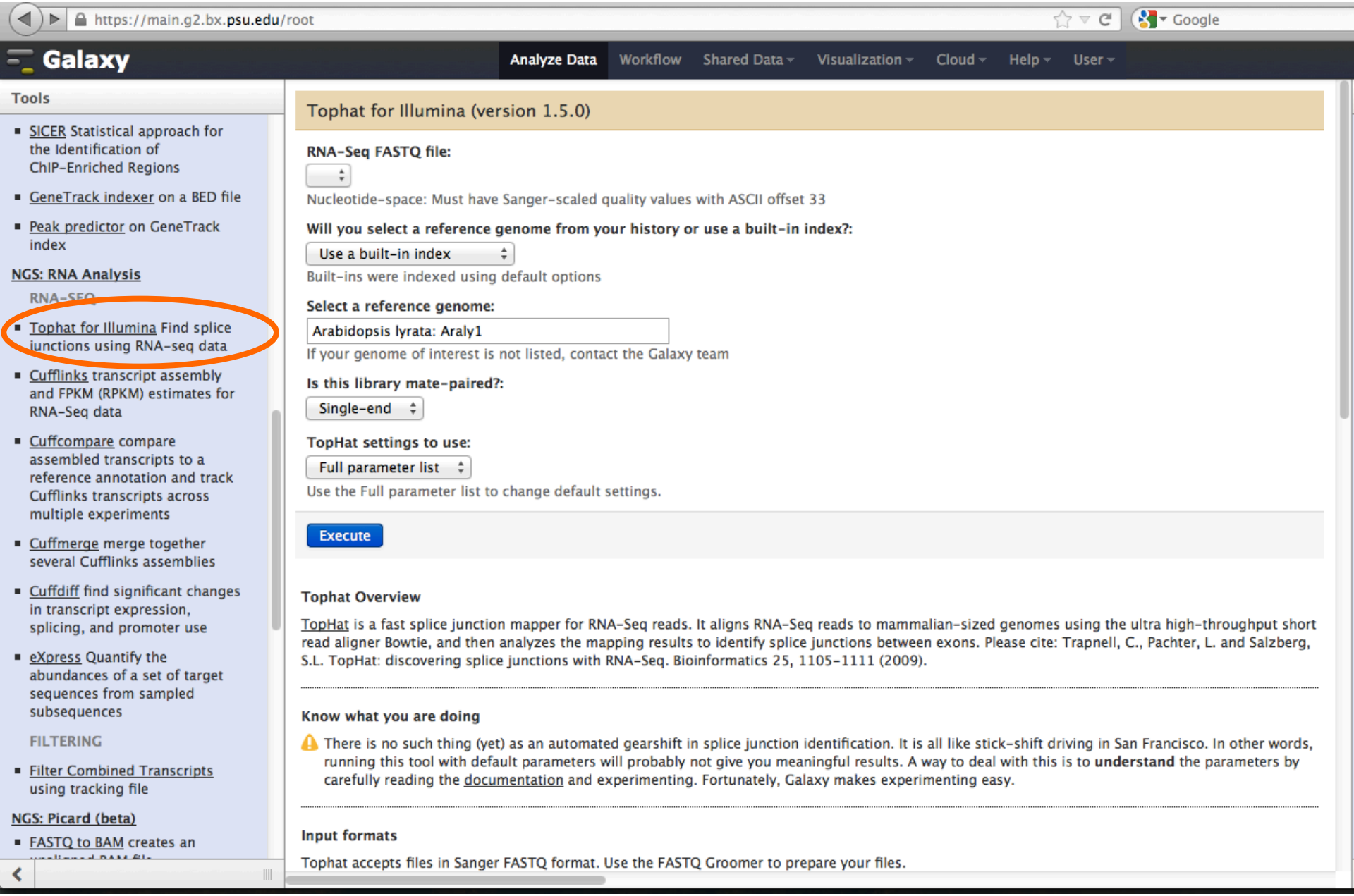
 export PATH=/share/apps/bin:$PATH

 tophat --num-threads 4 --GTF /seq/lincRNA/data/mm9.mrna.
10.31.gtf --prefilter-multihits --output-dir
ACTAAG.th1.4.1.g15 --segment-length 20 --max-multihits 15
/seq/lincRNA/data/mm9.nonrandom.bowtie
/seq/dcchip/mouse/DC/rnaSeq/DGE/hiseq_1-20-
12/split_fqs/ACTAAG.fq
```

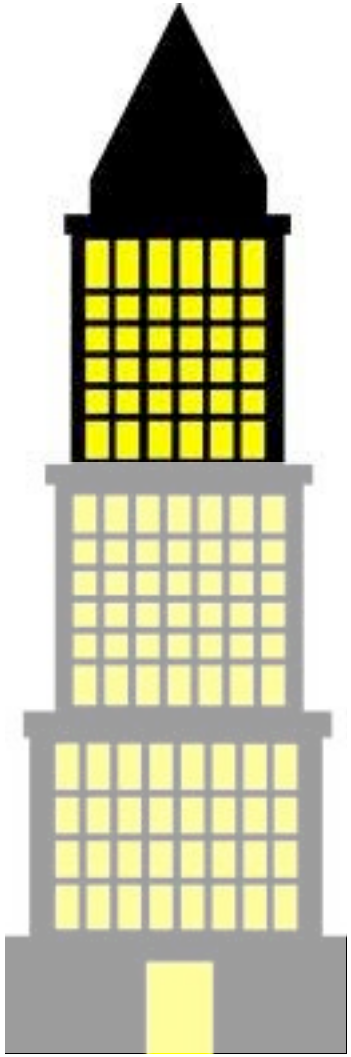
  - And submit to the server farm

Of course not averyone is comfortable with UNIX and scripting!

And there is a solution for this …



Goal: abstract the technical complexity, let labs leverage their intuition



Repeat analysis using different parameter settings
Specific analysis incorporating biological insight
Custom analysis
Custom figures
Writing my paper

Provides standard analysis options and supports analysis developed at UMASS

Implements best of breed, UMASS specific data processing pipelines

Eases data access and manipulation

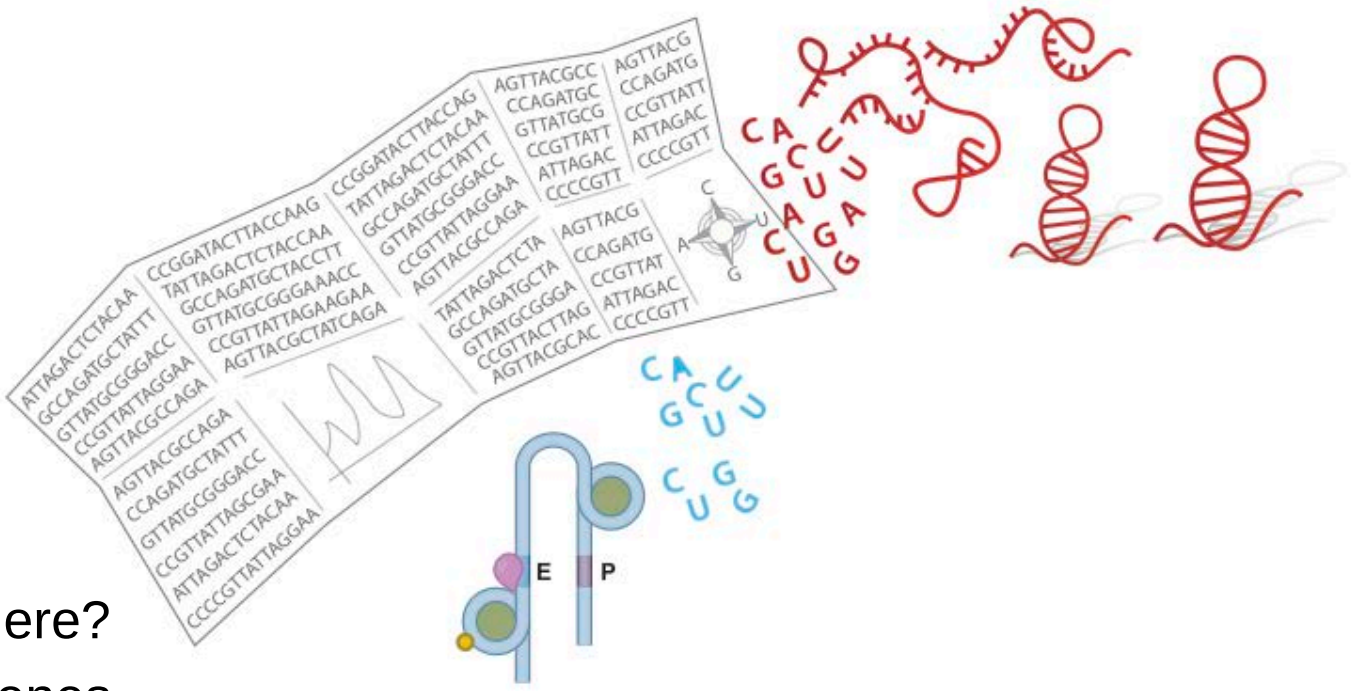
Help for the last step

- Discussion:
  - Monthly meeting, 2nd Friday of every month at 11:00 am all are welcomed.
    - Next meeting: methods to annotate and quantify piRNAs
  - Mailing list: bioinfo@list.umassmed.edu
- Training:
  - Invited speakers:
    - Genome Space team (February)
    - R workshop (possibly in March)
    - Planning an RNA-Seq analysis workshop
  - Training on supported tools and methods

Questions on the bioninformatics core?

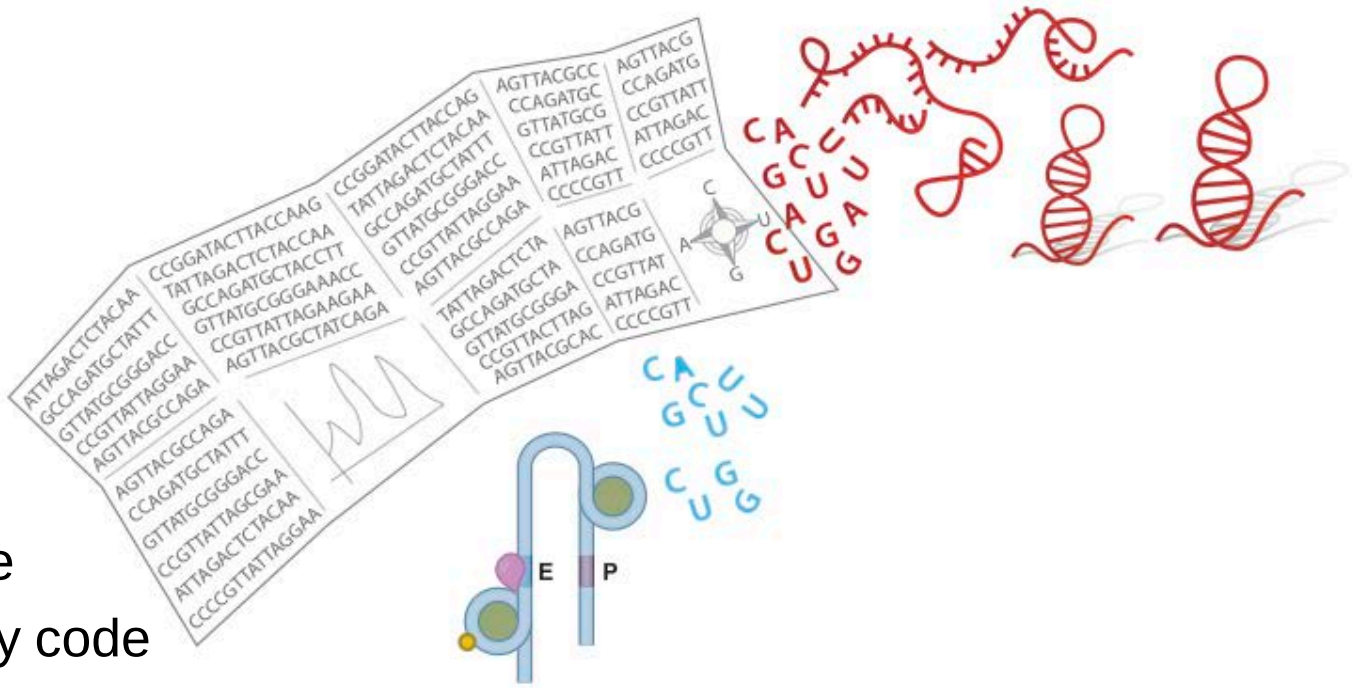
WHAT DOES MY LAB DO?

The genome encodes many different elements



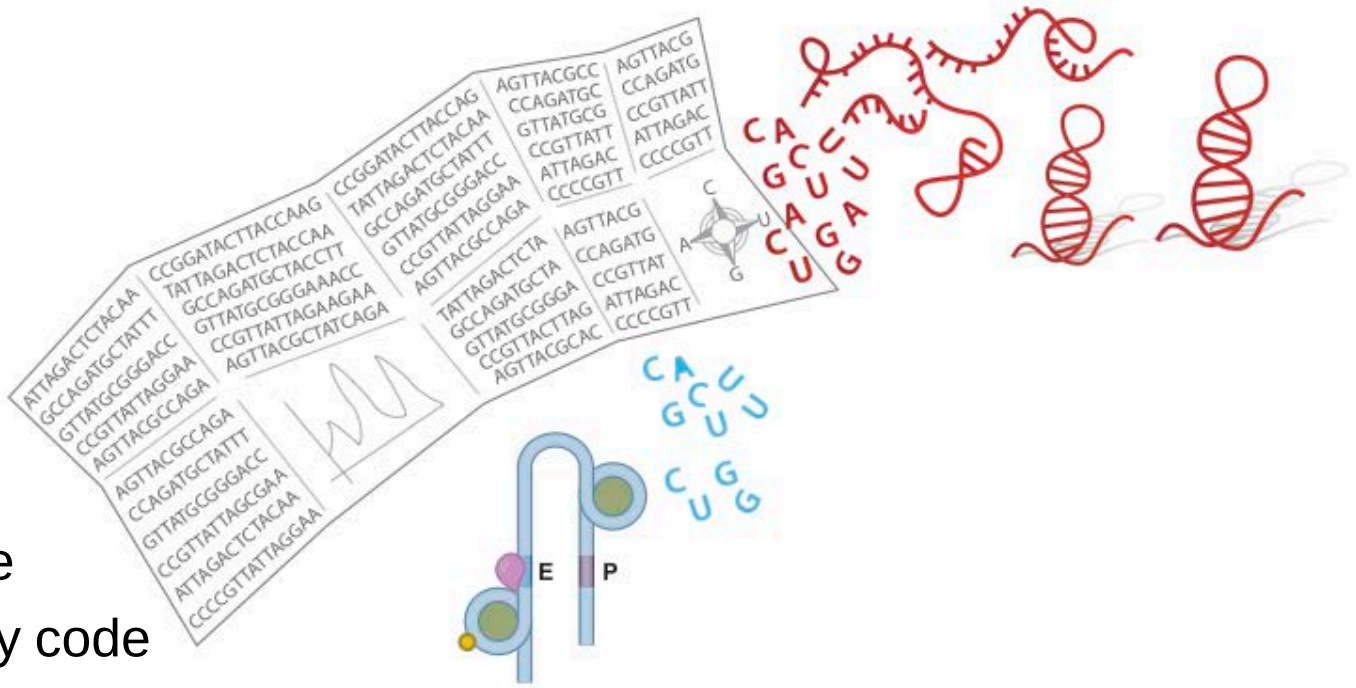
- What is out there?
- Non-coding genes
  - Finding them
  - *Characterizing function*
  - *Mechanism and evolution*
- Regulatory elements

Each element set has a unique code



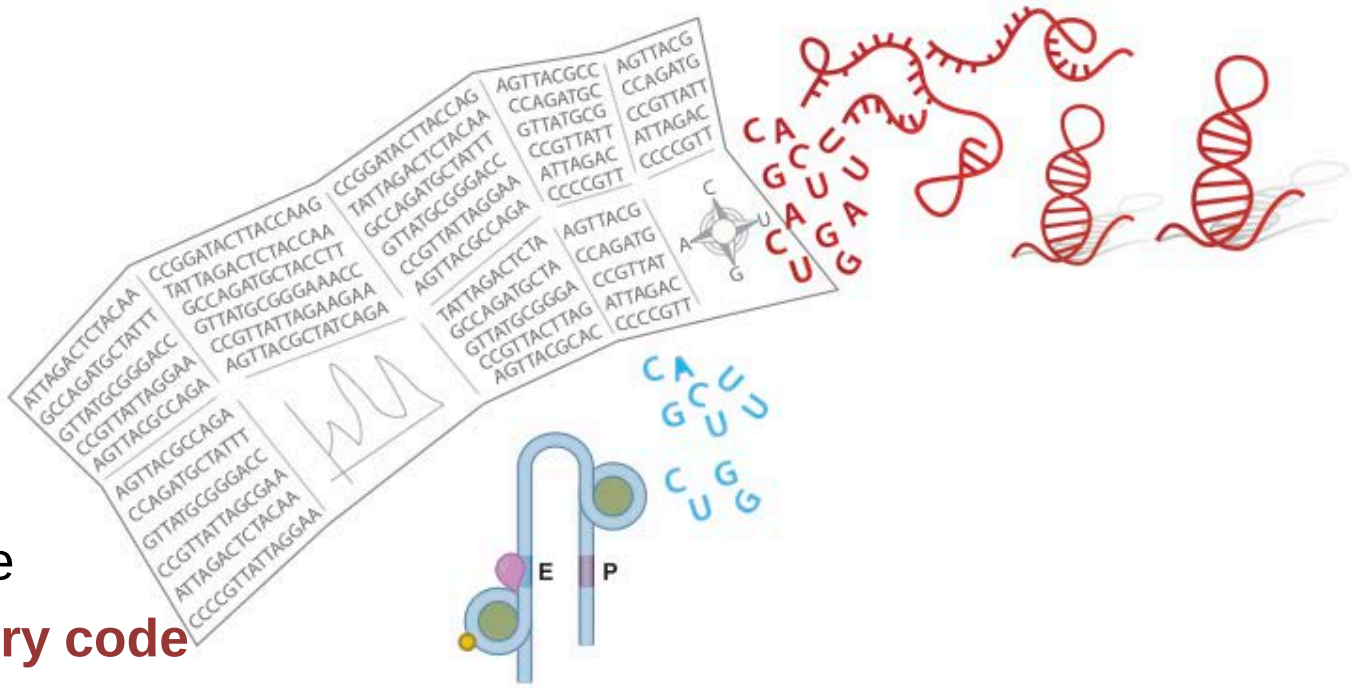
- Genetic Code
- Cis-regulatory code
- RNAi and miRNA codes
- *RNA Code*
- Histone Code

Most of which we do not understand



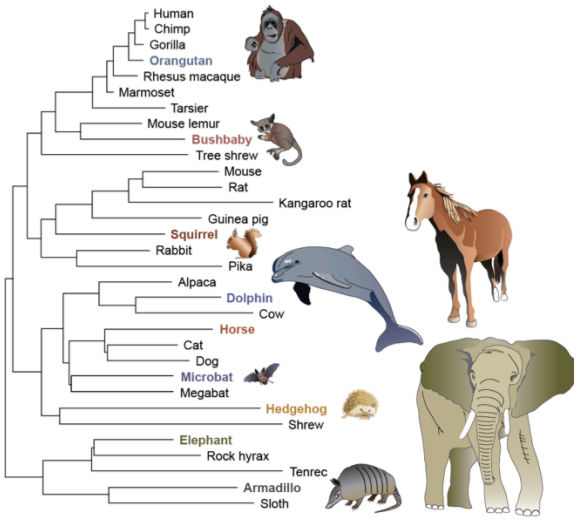
- ✓ Genetic Code
- ? Cis-regulatory code
- ? RNAi and miRNA codes
- ? *RNA Code*
- ? Histone Code

Most of which we do not understand

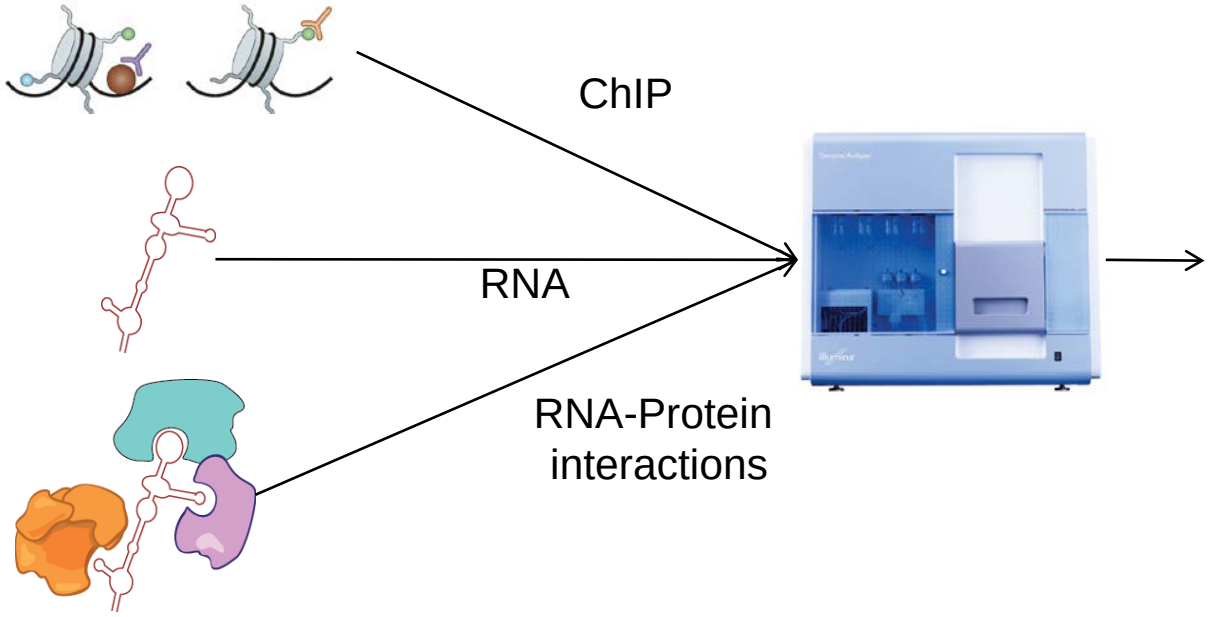


- ✓ Genetic Code
- ? **Cis-regulatory code**
- ? RNAi and miRNA codes
- ? *RNA Code*
- ? Histone Code

Our work



Estimate the “functional genome” by finding what is under selection



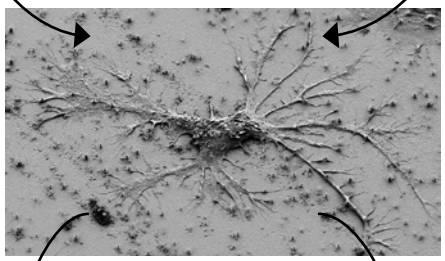
- Develop informatics tools for new methods
- Develop models of transcriptional regulation
- Develop models of epigenetic interactions
- Evolution of large non-coding RNAs

Project: Transcription regulation in DCs

Input

Bacteria

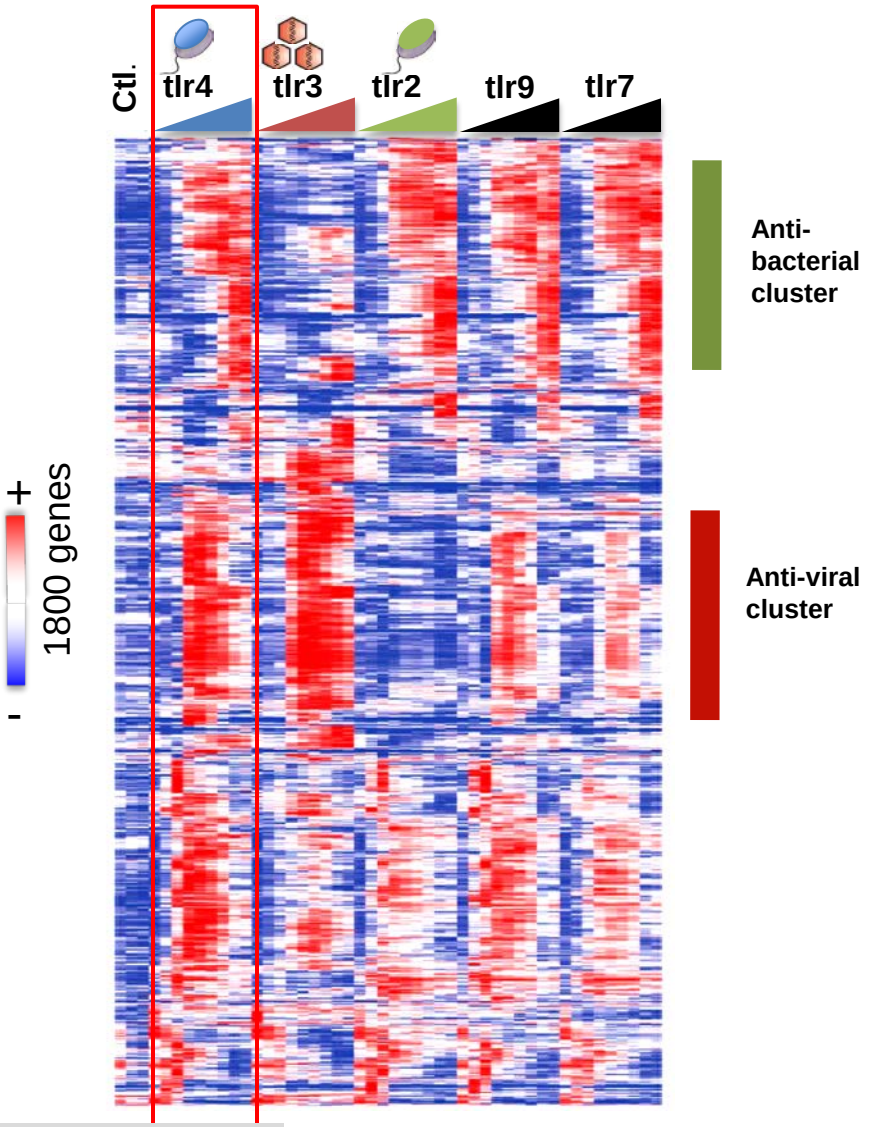
Viruses



Output

Anti-bacterial program (inflammation)

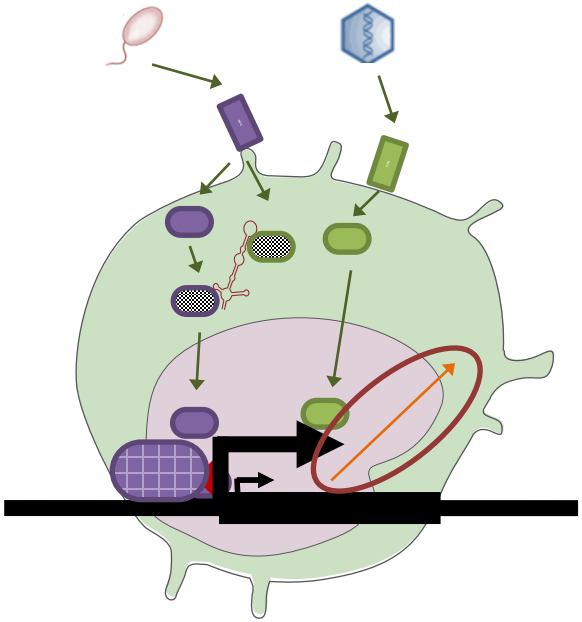
Anti-viral program (interferon)

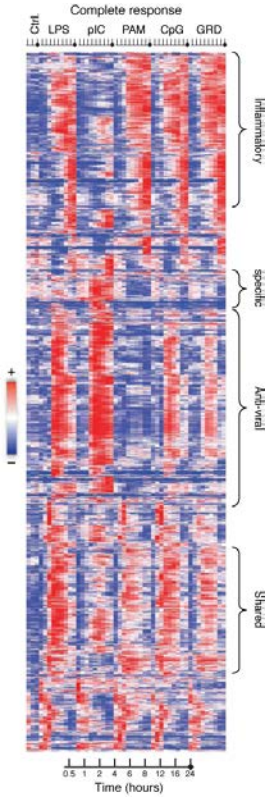


How is this response controlled?

Amit, Garber et al, *Science* 2010

Strategy: Genetic + physical mapping

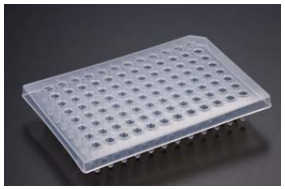




What are the direct targets of transcription factors?

Only possible with High throughput ChIP-Sequencing

Crosslink → Sonicate → Precipitate → Purify DNA → Indexed library → Sequence



Magnetic beads



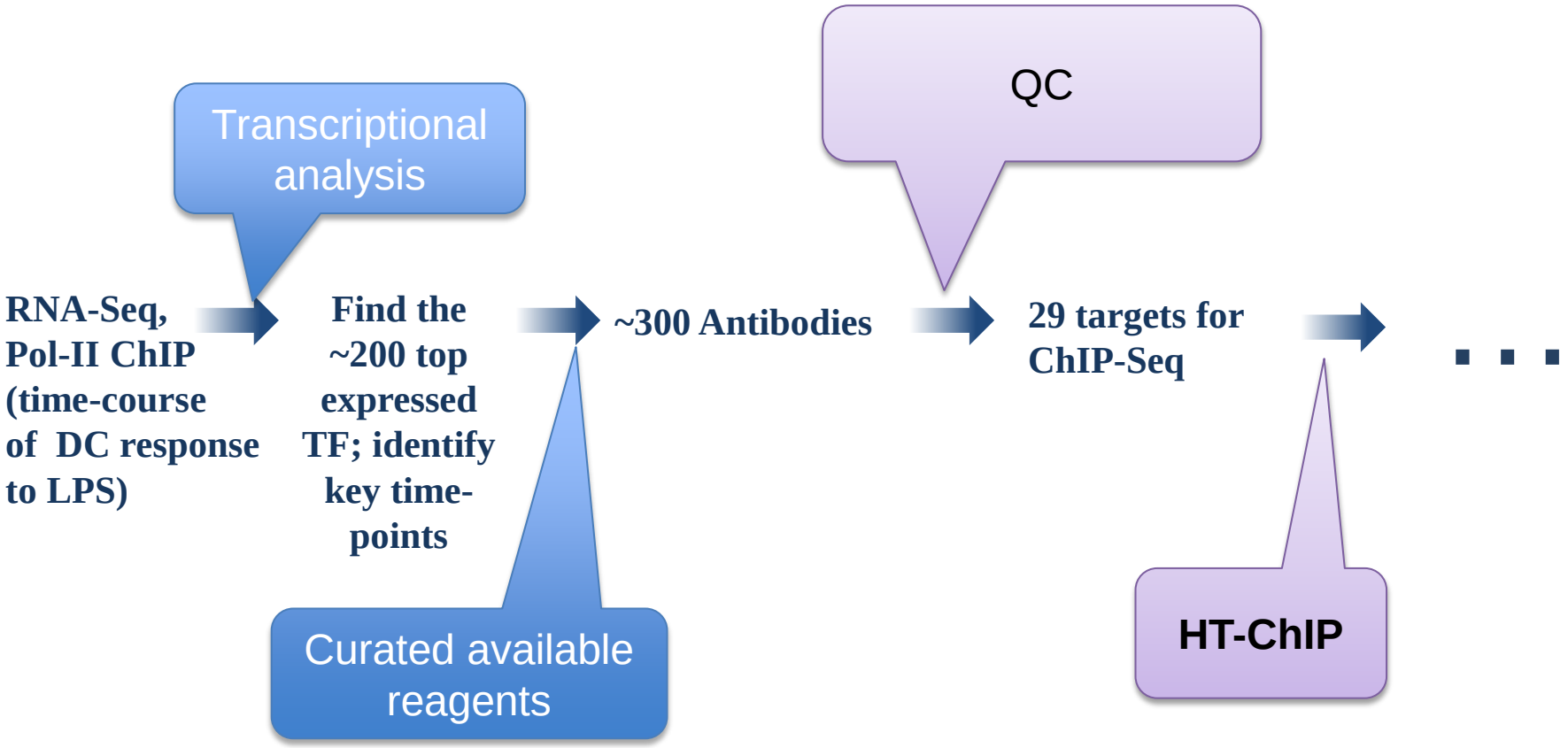
Robot-automated



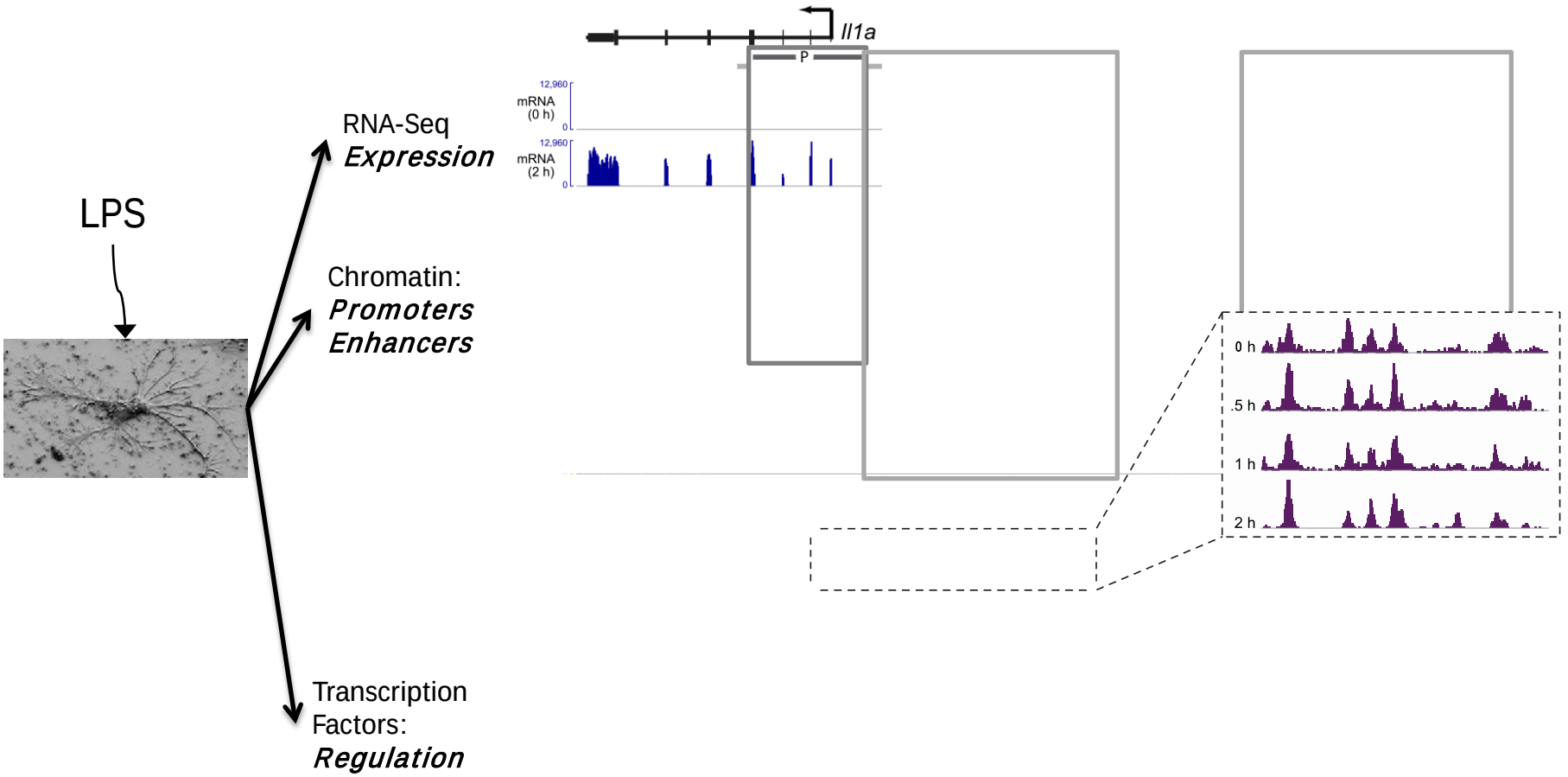
24 libraries/lane



Systematic mapping of the DC LPS-response network

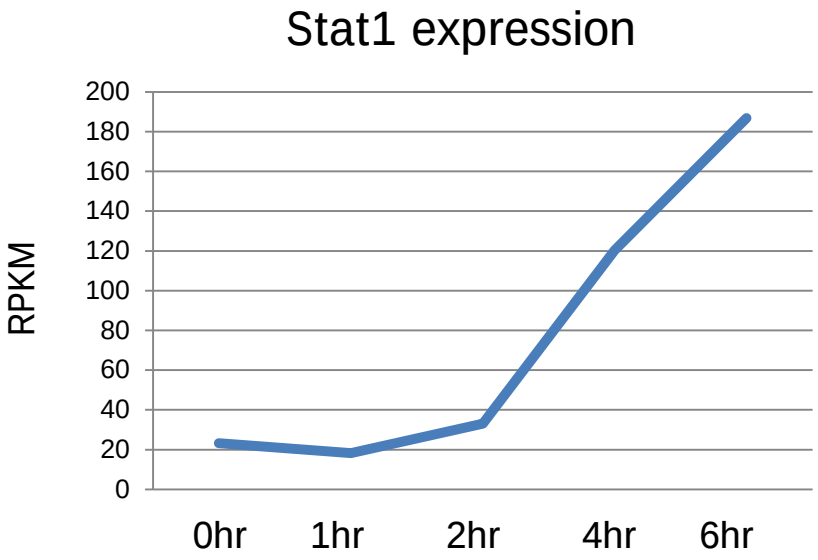


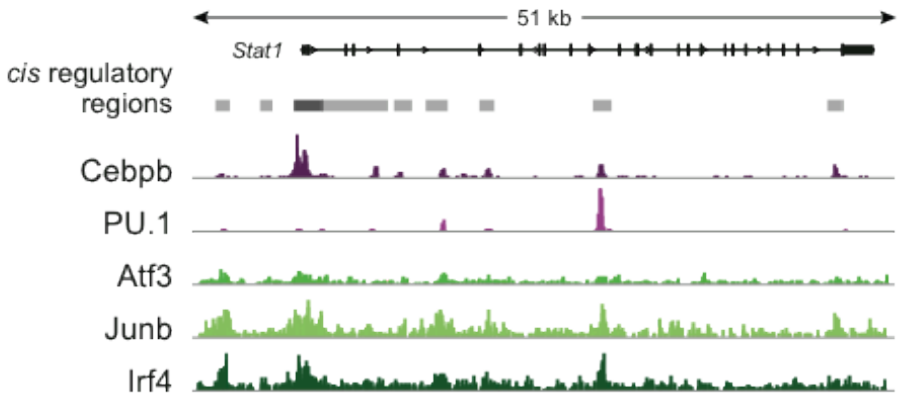
Dataset: temporal view of expression and state



85% of high scoring TF peaks fall within annotated *cis*-regulatory regions

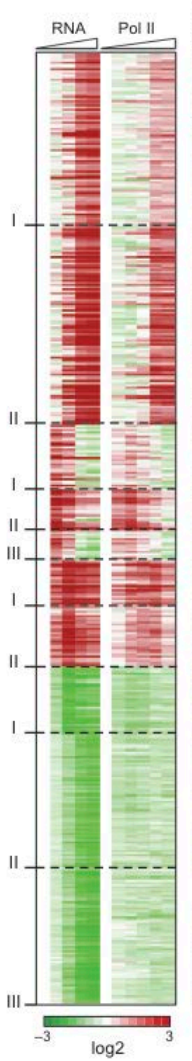
An example: Stat1, a late induced gene





Stat1 expression is a combination of pre-binding and dynamic binding

Transcription factors control specific pathways



Cd274
Cd38
Tnfrsf14

Inflammation:
B, T activation

Stat1
Irf7
Mx2

Anti-viral

TNF
Cxcl2
Nfkbia

Inflammation

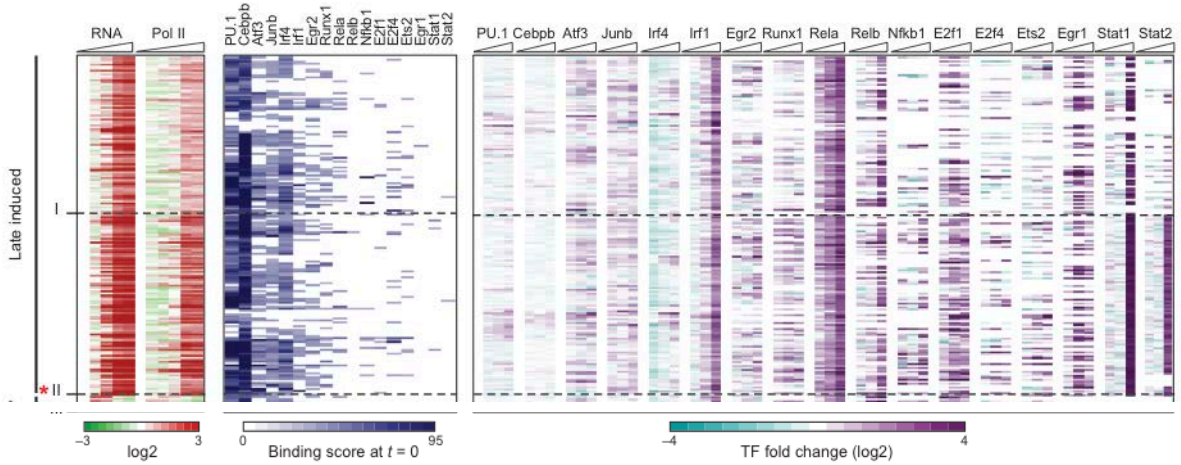
Inflammation

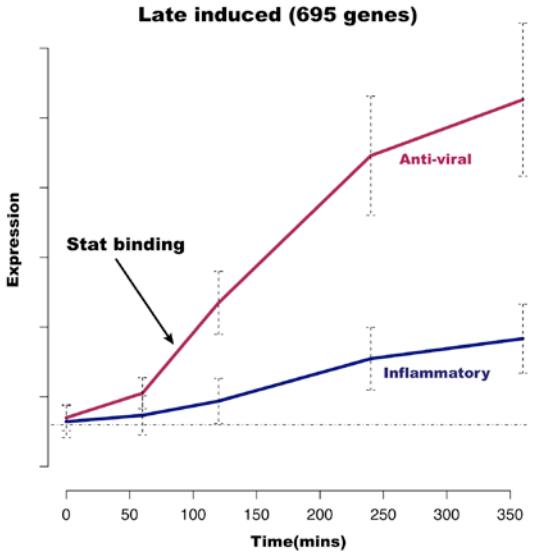
Anti-apoptotic

Cell cycle

What are the differences between sub-clusters?

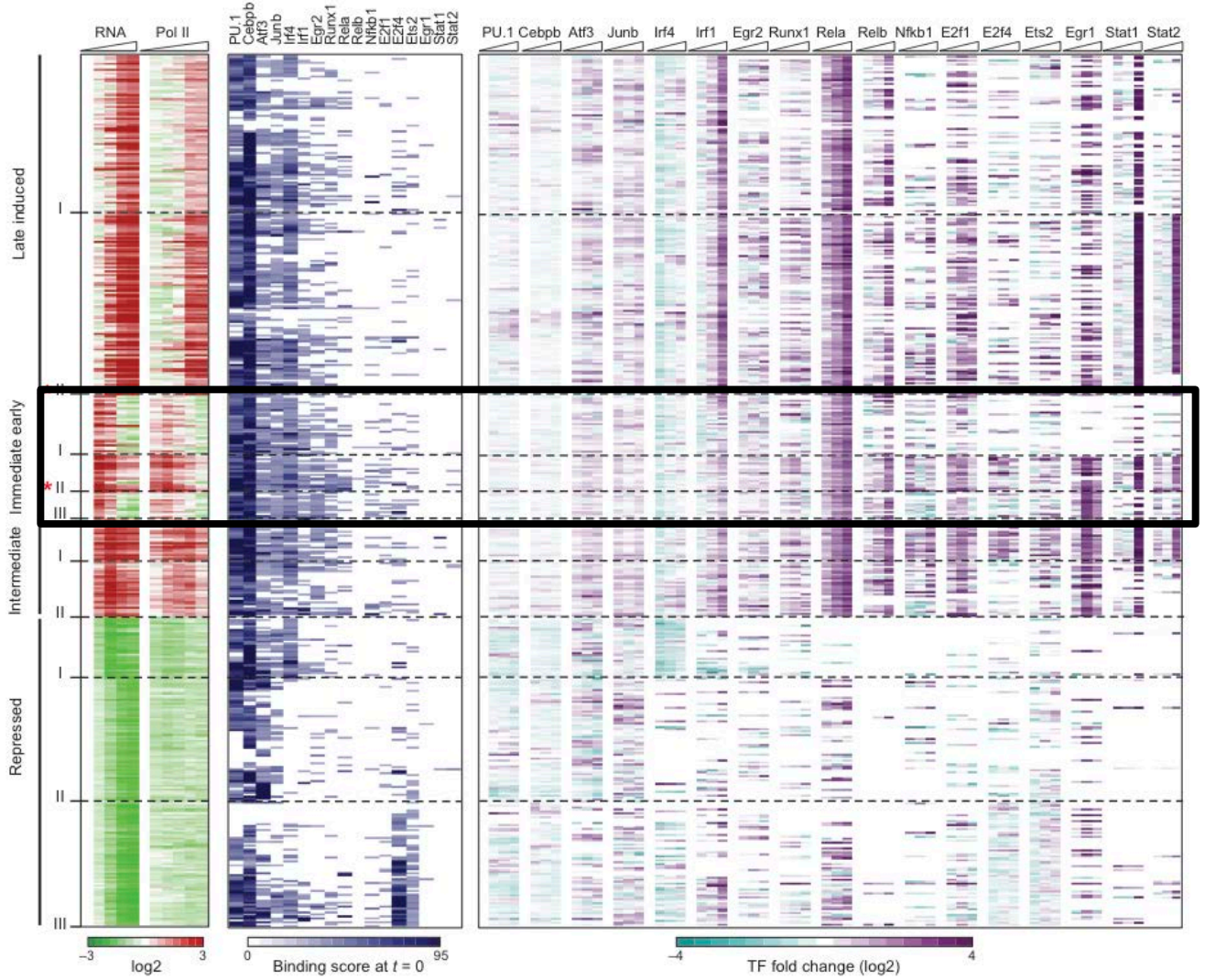
Specific factors control amplitude of expression





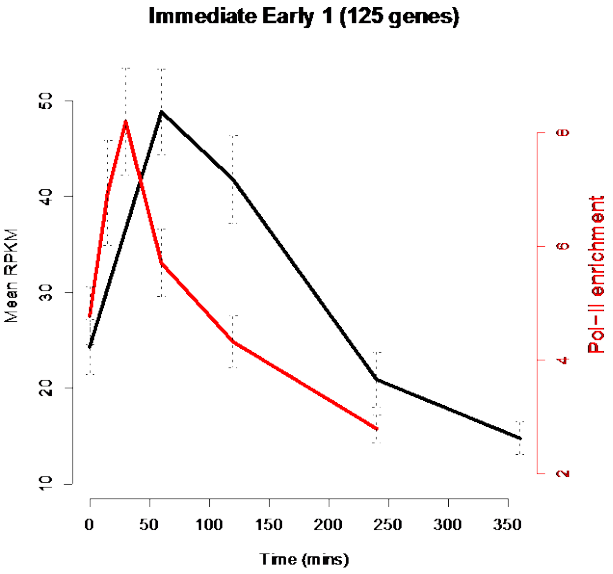
Binding of stat1/2 controls inductions levels

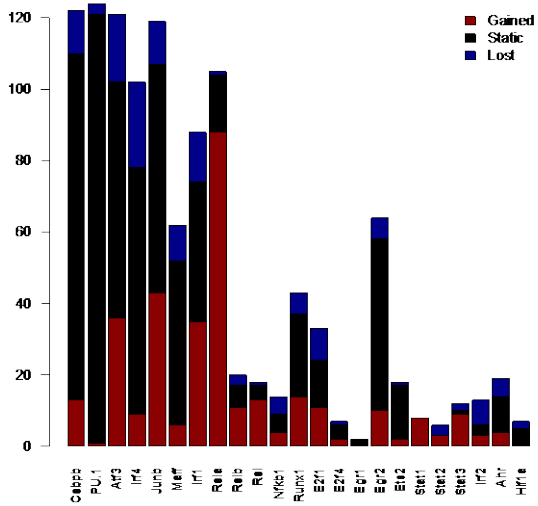
Immediate early genes are highly bound

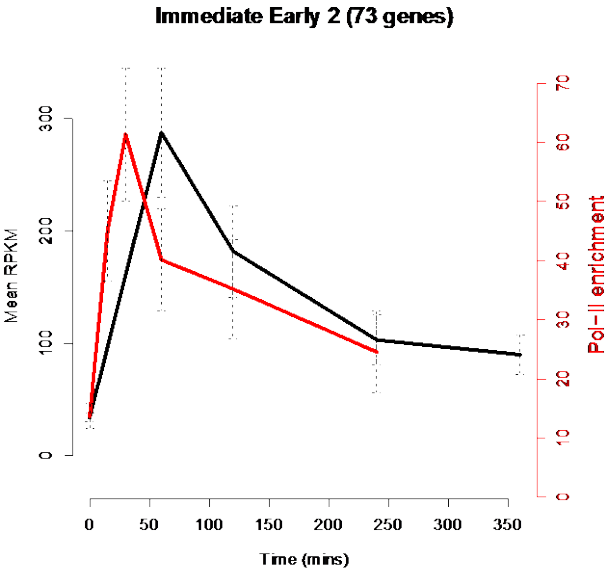


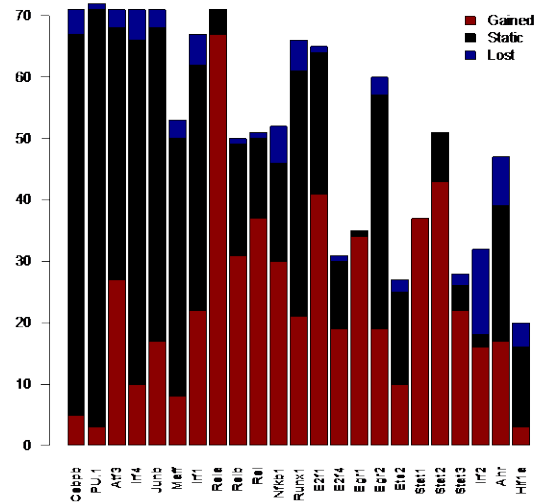
What are the differences between sub-clusters?

Immediate early gene programs



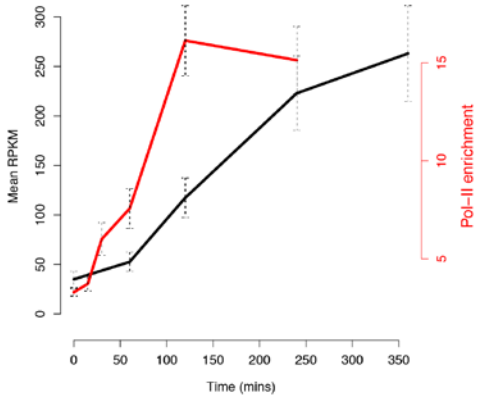




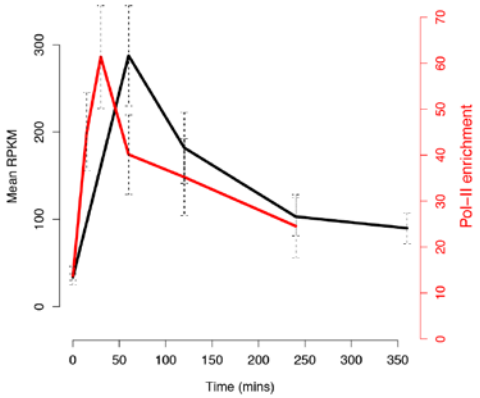


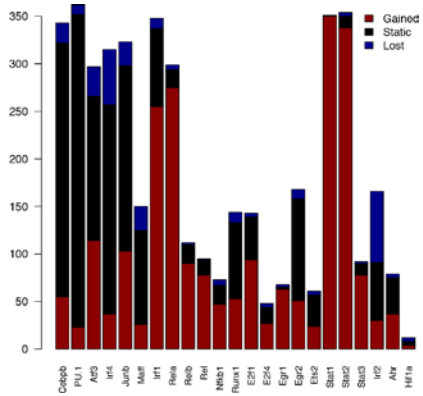
Current models under consideration

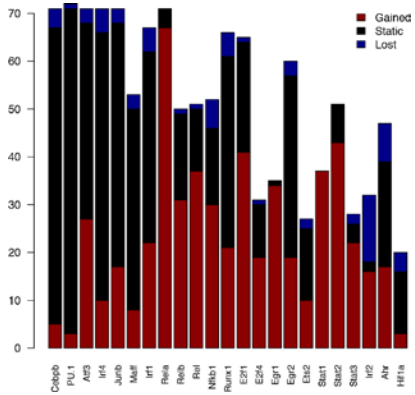
Late induced
Stat regulated



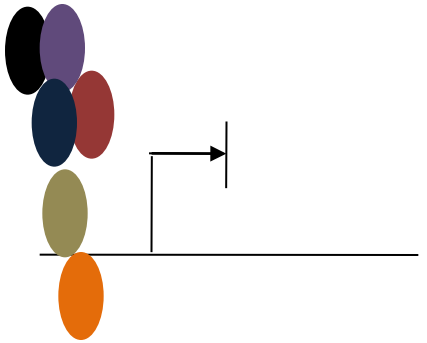
Immediate early
highly induced



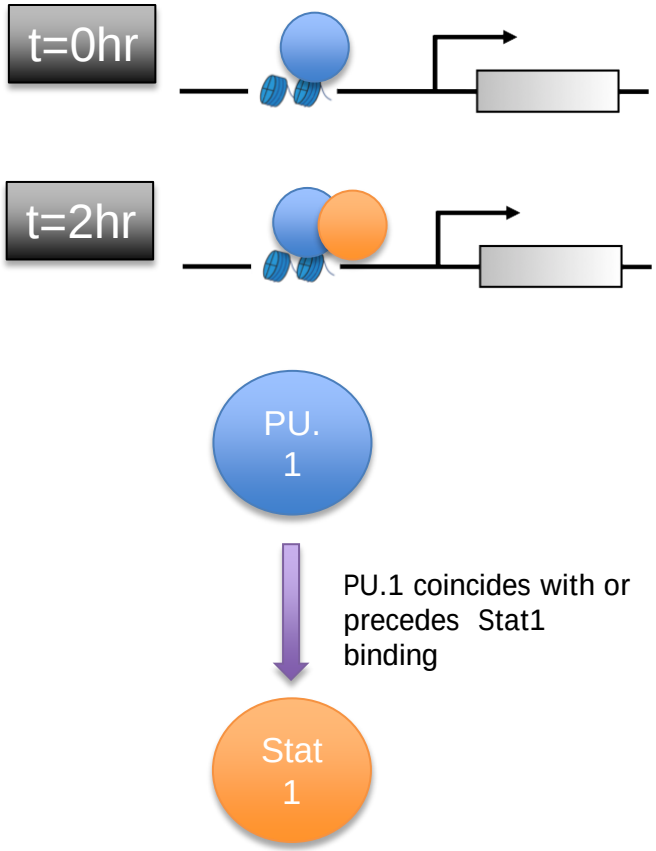


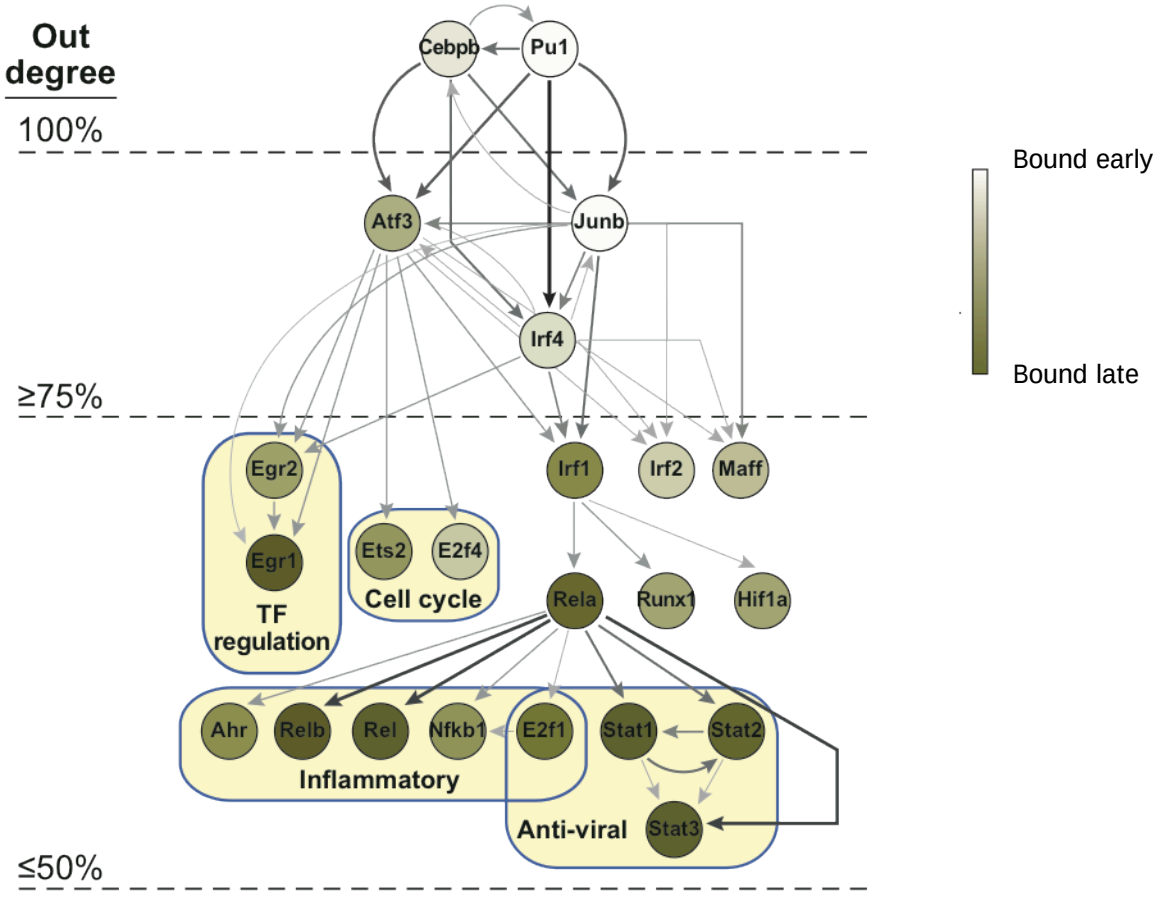


Two forms of regulation?



Regulatory modes are established hierarchically

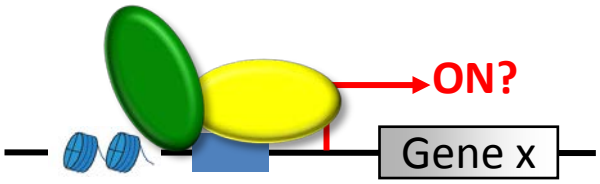




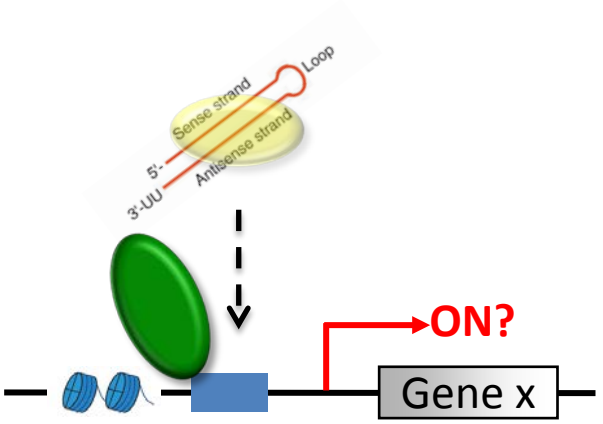
Conclusions and considerations

- A large fraction of binding exist prior to stimulus
- Immediate vs. late regulation is quite distinct:
  - Early induced genes regulators are more redundant
  - Late induced regulators are less redundant
  - *Are the early inflammation pathways evolutionary more malleable?*
- Factors act in layers, consistent with previous reports
- Genomic approaches like this are applicable to many systems
  - Protocols can handle smaller input material (Alon Goren, Oren Ram)
- *Test models using a genome wide genetic screen*
- *Map TFs with no available antibodies*
- *Currently building maps of another 20 factors for which antibodies became available*

Next steps: Perturbing each factor



TF binding map



Loss of function screen

An expensive proposition…

~100 genes KD * replicates = LARGE NUMBER of samples

- ✗ high cost
- ✗ limited starting material

An expensive proposition…

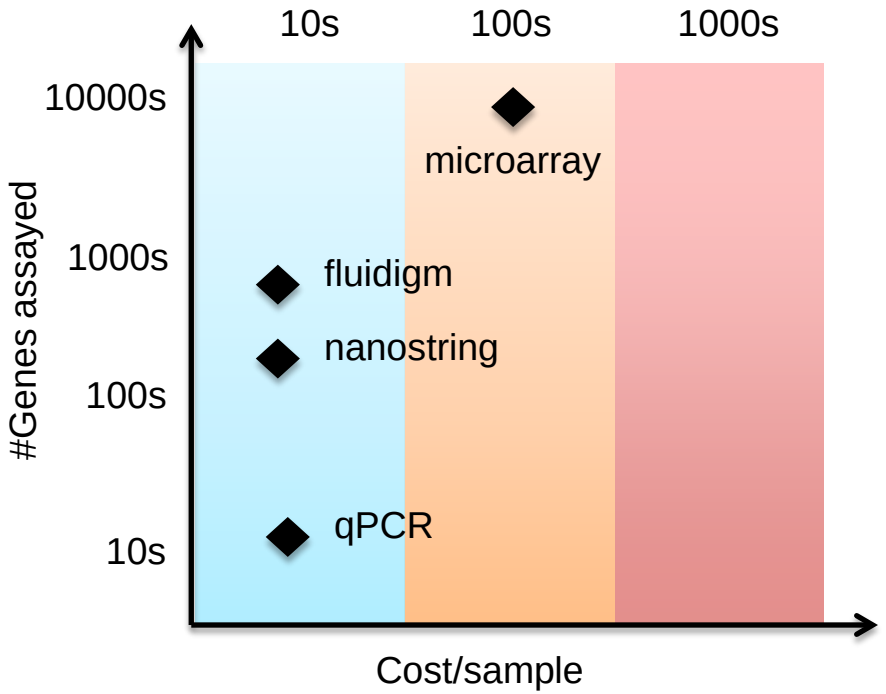
~100 genes KD * replicates = LARGE NUMBER of samples

- ✗ high cost
- ✗ limited starting material

Previous solution

- qPCR
- Fluidigm
- Luminex
- NanoString
- microarray

Problem: need to chose your genes in advance and limited #genes assayed



An expensive proposition…

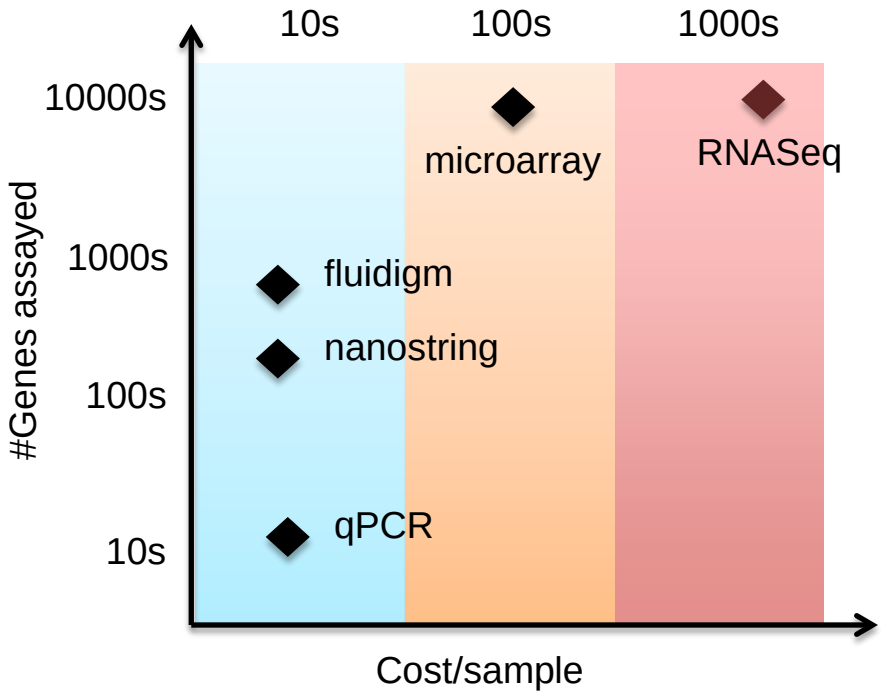
~100 genes KD * replicates = LARGE NUMBER of samples

- ✗ high cost
- ✗ limited starting material

Previous solution

- qPCR
- Fluidigm
- Luminex
- NanoString
- microarray

Problem: need to chose your genes in advance and limited #genes assayed



Goal: Cheap RNA-Seq for quantification

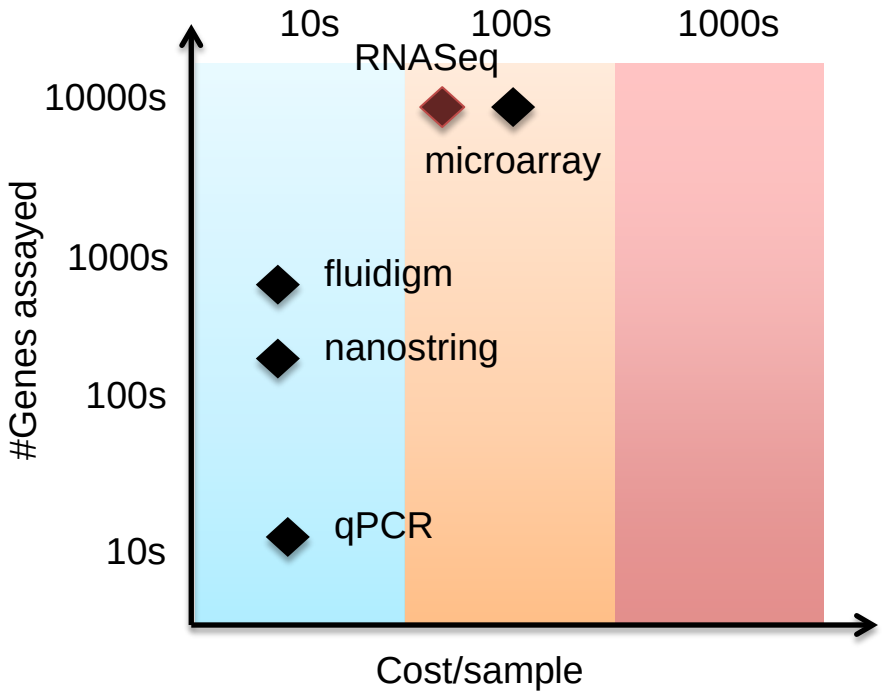
~100 genes KD * replicates = LARGE NUMBER of samples

- ✗ high cost
- ✗ limited starting material

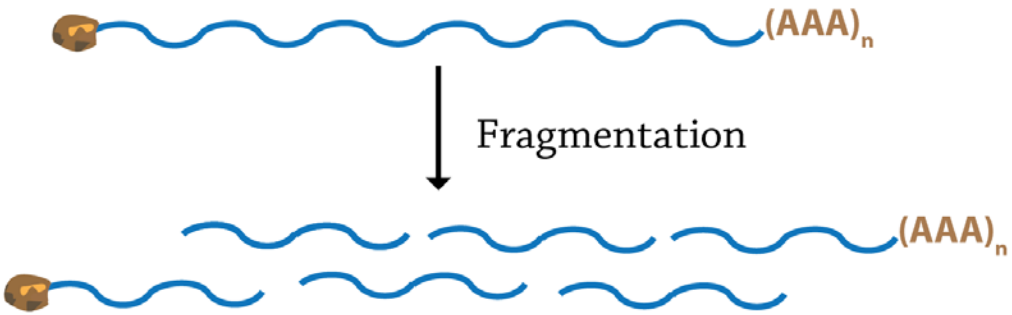
Previous solution

- qPCR
- Fluidigm
- Luminex
- NanoString
- microarray

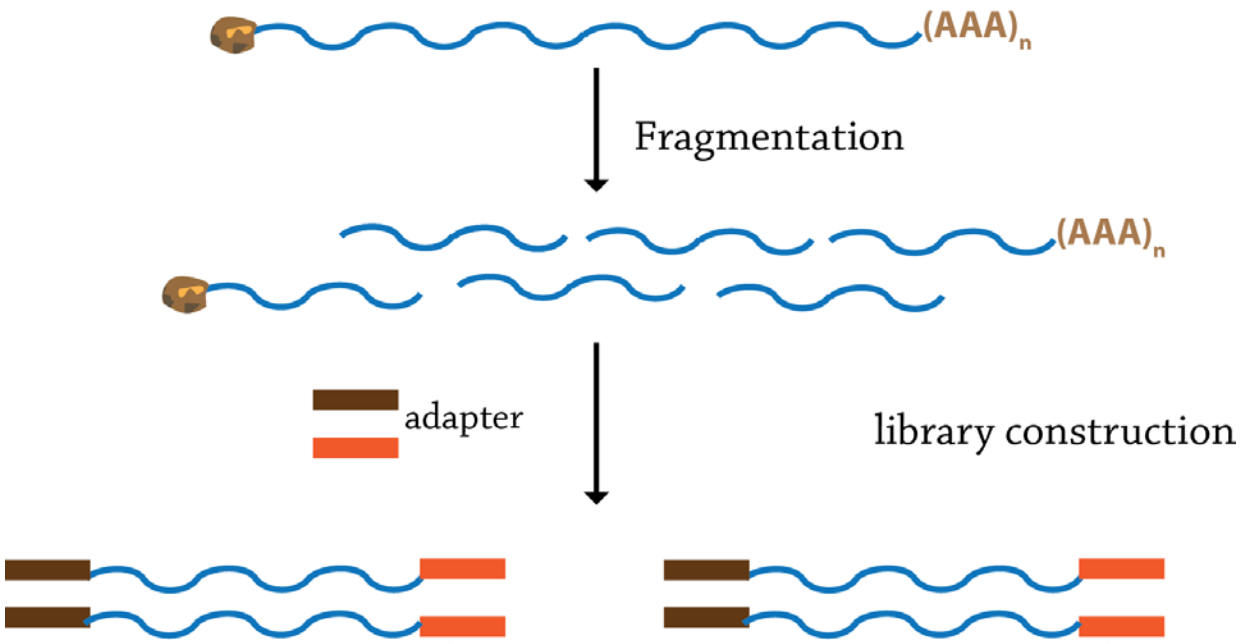
Problem: need to chose your genes in advance and limited #genes assayed



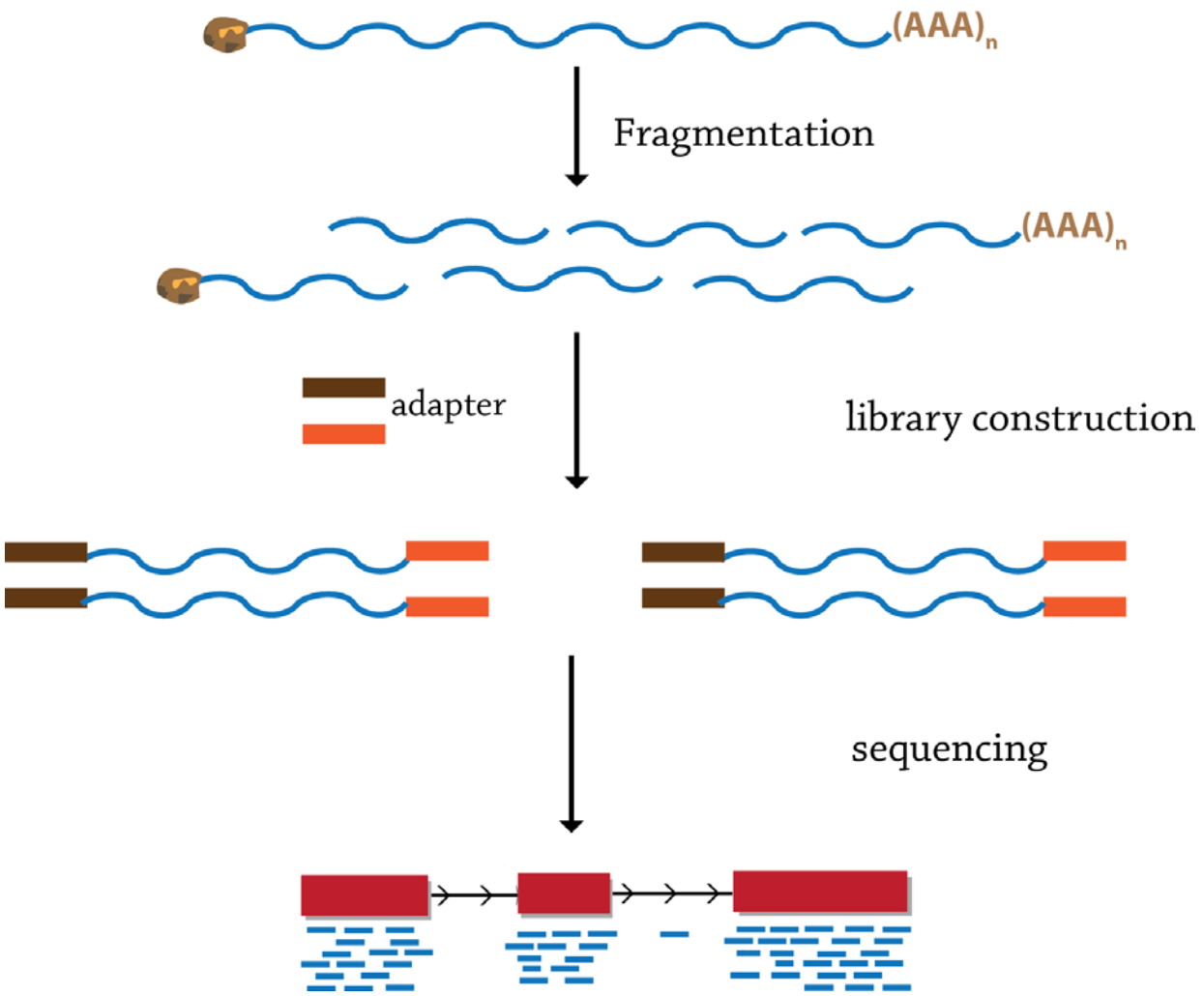
Full length RNA-Sequencing



Full length RNA-Sequencing



Full length RNA-Sequencing



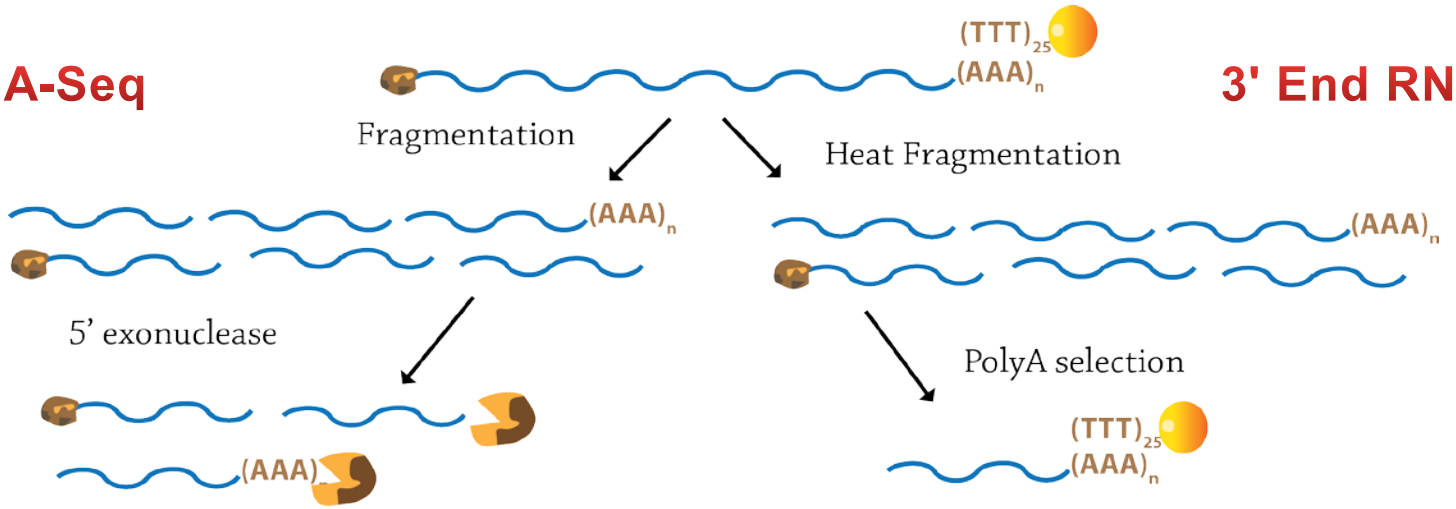
End RNA-Seq



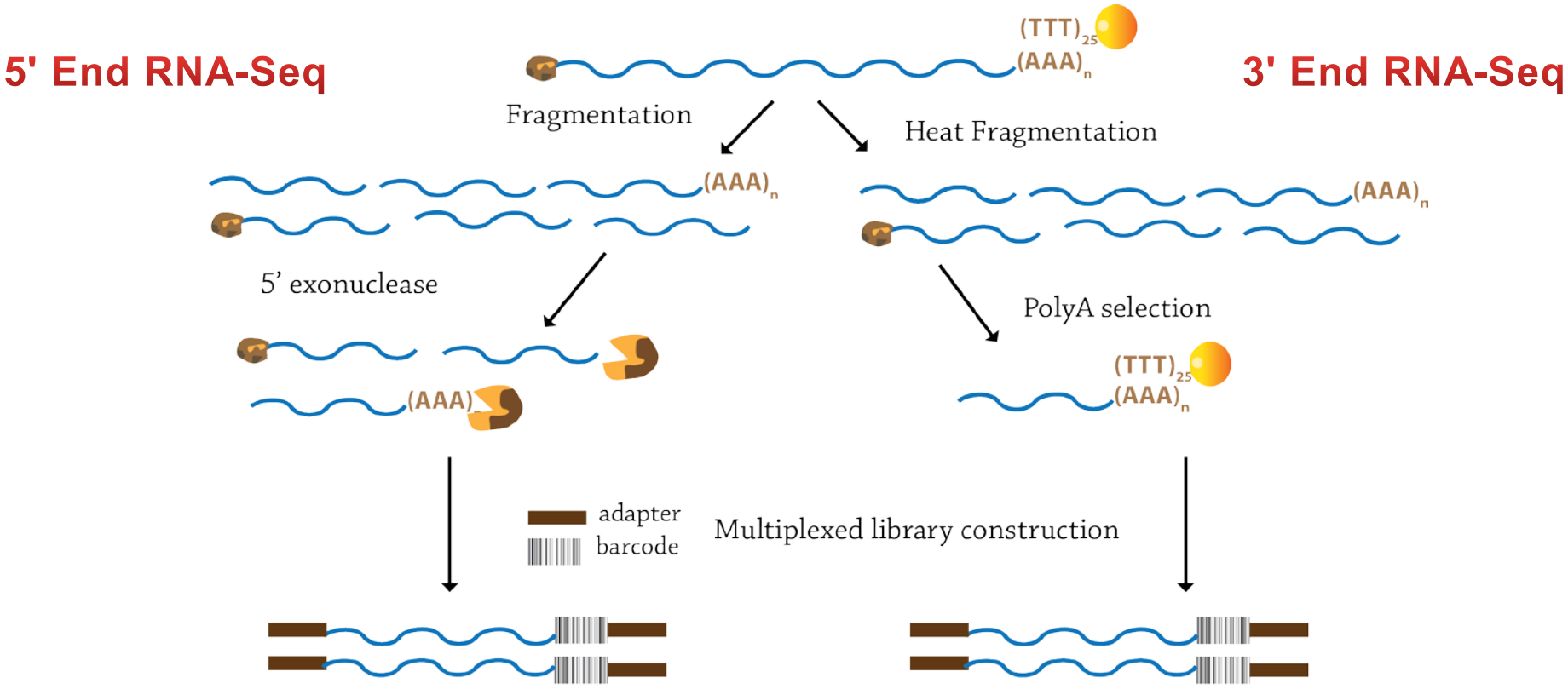
End RNA-Seq

5' End RNA-Seq

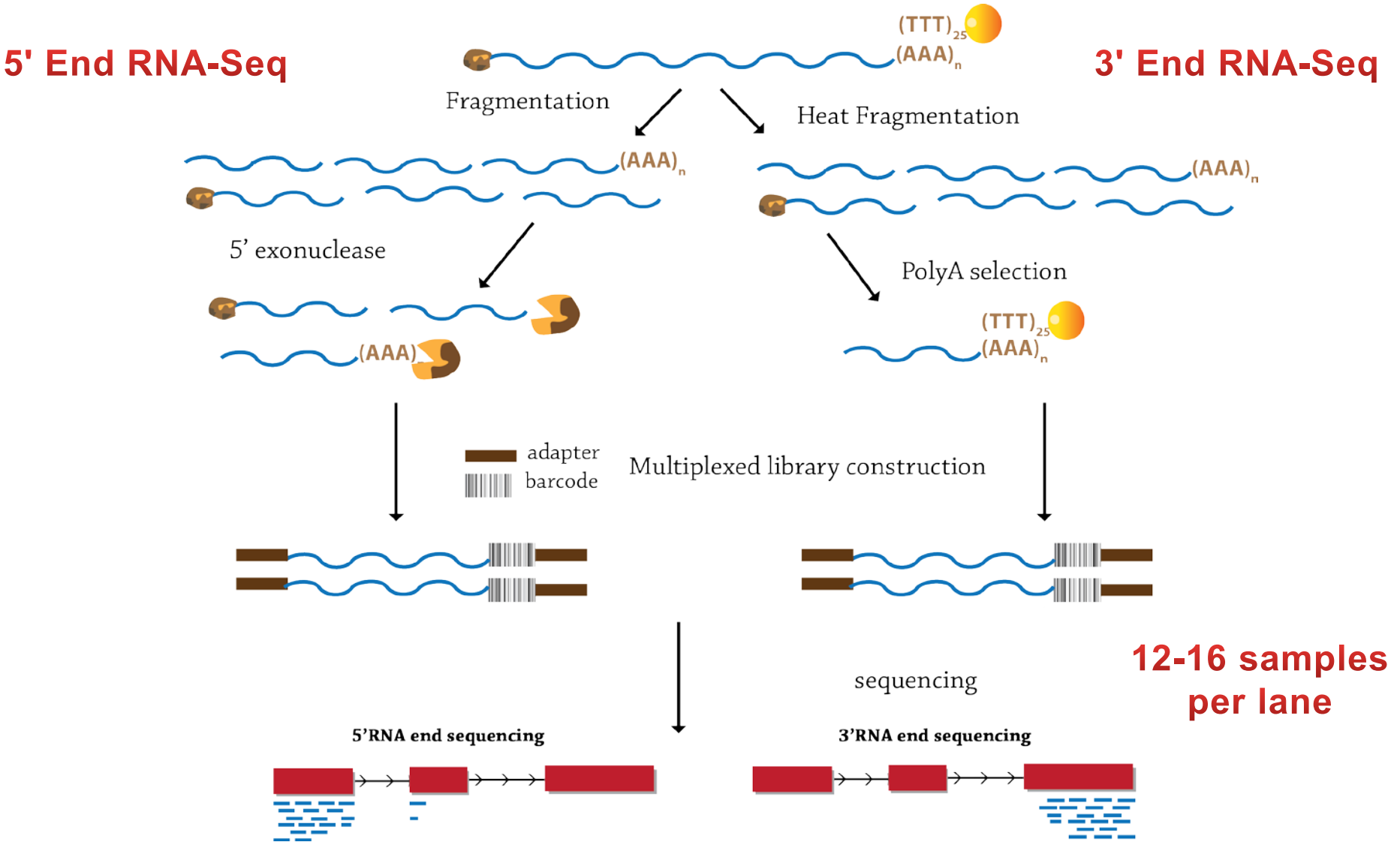
3' End RNA-Seq



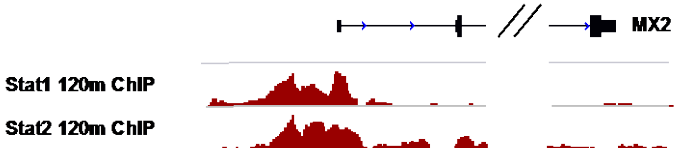
End RNA-Seq

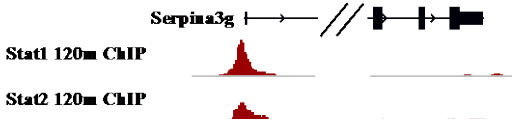

5' End RNA-Seq
3' End RNA-Seq
(TTT)25
(AAA)n
Fragmentation
Heat Fragmentation
(AAA)n
(AAA)n
5' exonuclease
PolyA selection
(AAA)n
(TTT)25
(AAA)n
adapter
barcode
Multiplexed library construction

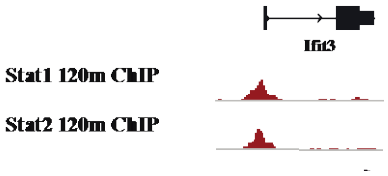
End RNA-Seq

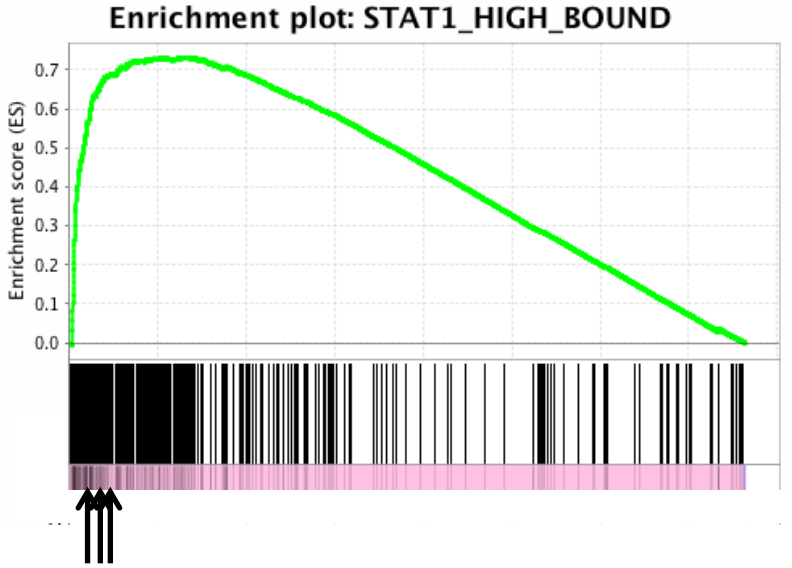

5' End RNA-Seq
3' End RNA-Seq
(TTT)25
(AAA)n
Fragmentation
Heat Fragmentation
(AAA)n
(AAA)n
5' exonuclease
PolyA selection
(AAA)
(TTT)25
(AAA)n
adapter
barcode
Multiplexed library construction
sequencing
12-16 samples per lane
5'RNA end sequencing
3'RNA end sequencing

Current work: Generating timecourses of KDs









Acknowledgements

Ido Amit



Nir Yosef



Jim Robinson, Helga Thorvaldsdottir, Bang Wong (IGV)
New Visualization for time series ChIP data

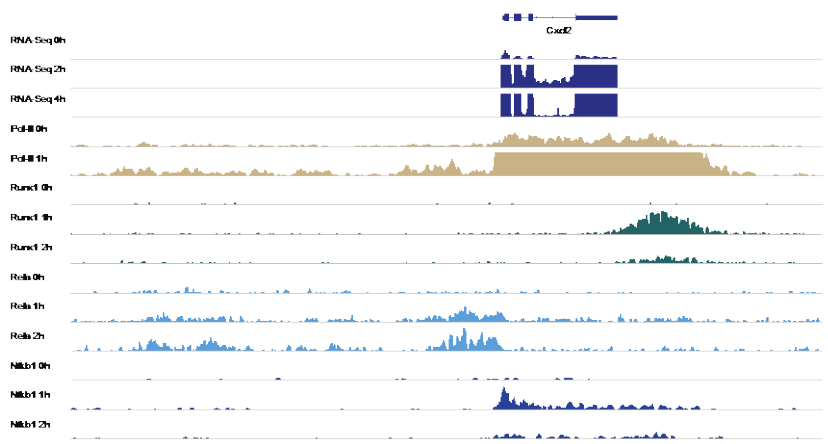
Raktima Raychowdhury and Anne Thielke
Automation, Library preparation, cell culture

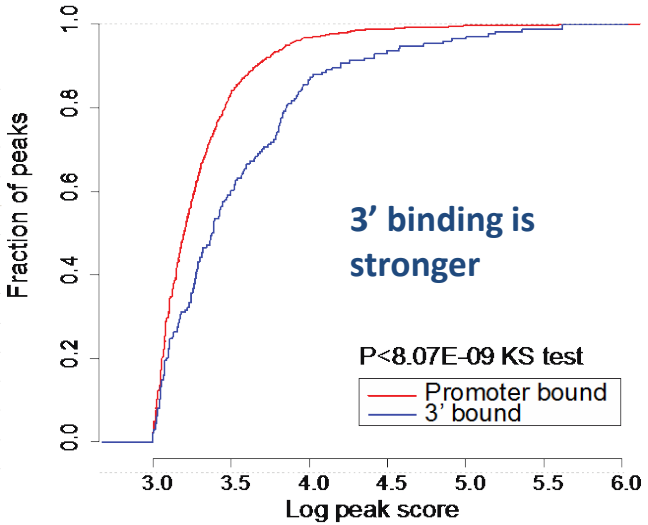
Brian Minie, Dennis Friedrich, Jim Meldrim, Andrew Barry, Chad Nusbaum (GSAP)
Automation, High Throughput protocols

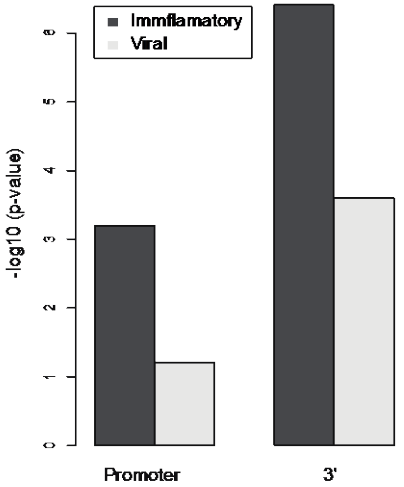
Oren Ram, Alon Goren
ChIP String, Low input, many interactions

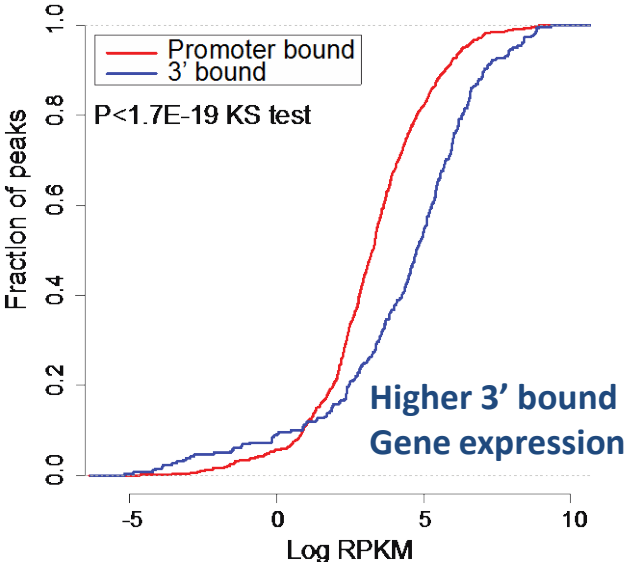
Jim Bochicchio
Christine Cheng
Nir Hacohen
Brad Bernstein
Aviv Regev

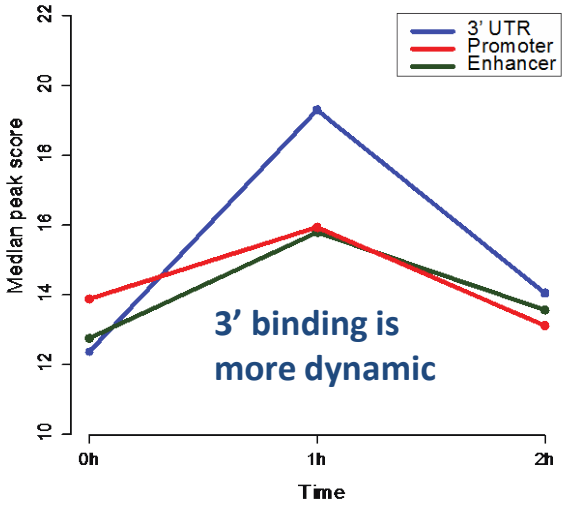
A non-canonical binder: Runx1

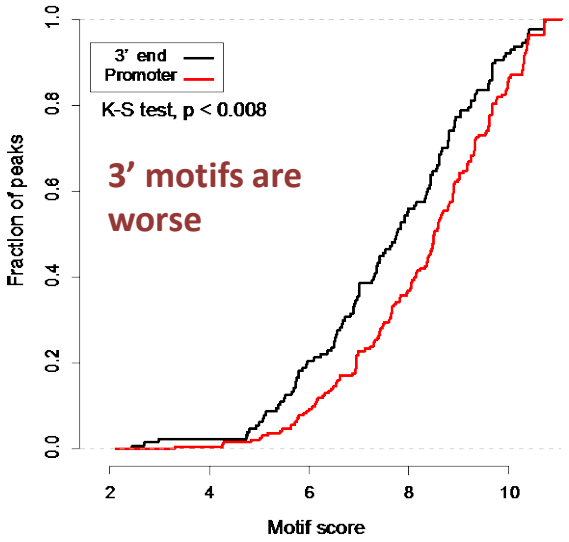






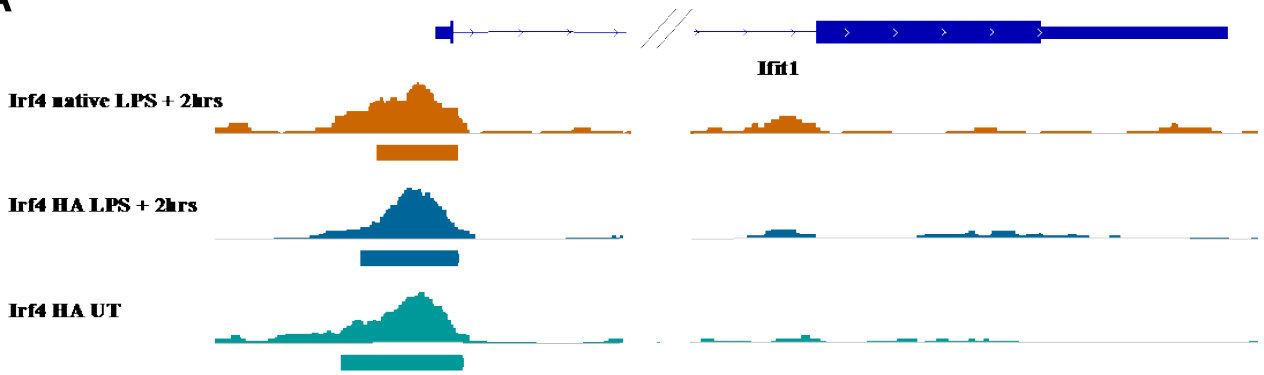




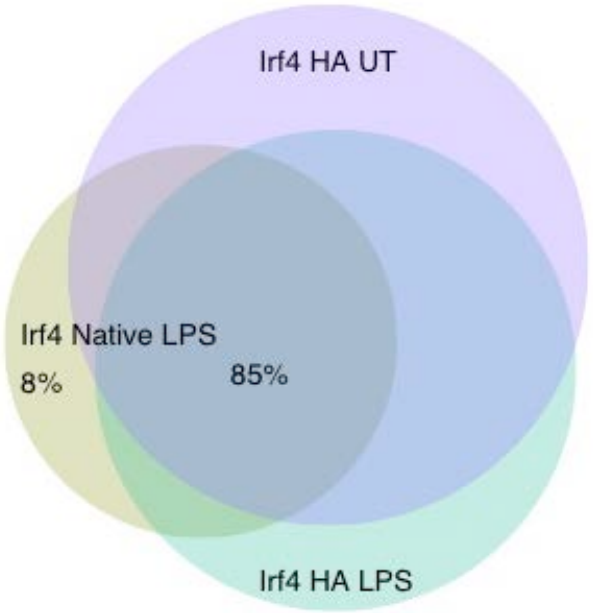


Mechanism unknown, different partner? Different complex?

A


Ifit1
Irf4 native LPS + 2hrs
Irf4 HA LPS + 2hrs
Irf4 HA UT

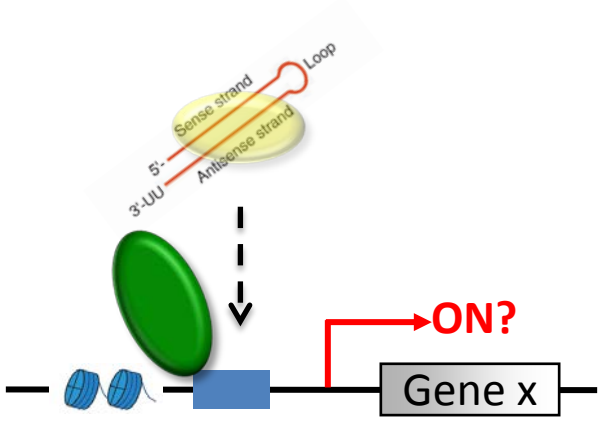
B


Irf4 HA UT
Irf4 Native LPS
8%
85%
Irf4 HA LPS

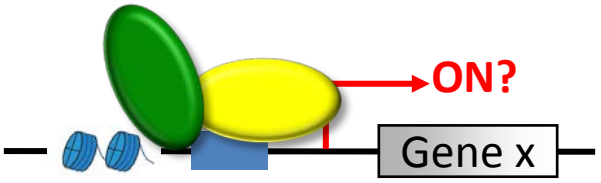
C


Rps19
Cd79a
Arhgef1
Rabac1
Atp1a3
Grik5
Zfp574
Pou2f2
NFKBIZ_Ab_1
NFKBIZ_Ab_2
NFKBIZ_Ab_3
NFKBIZ_Ab_4

Understanding the cis-regulome

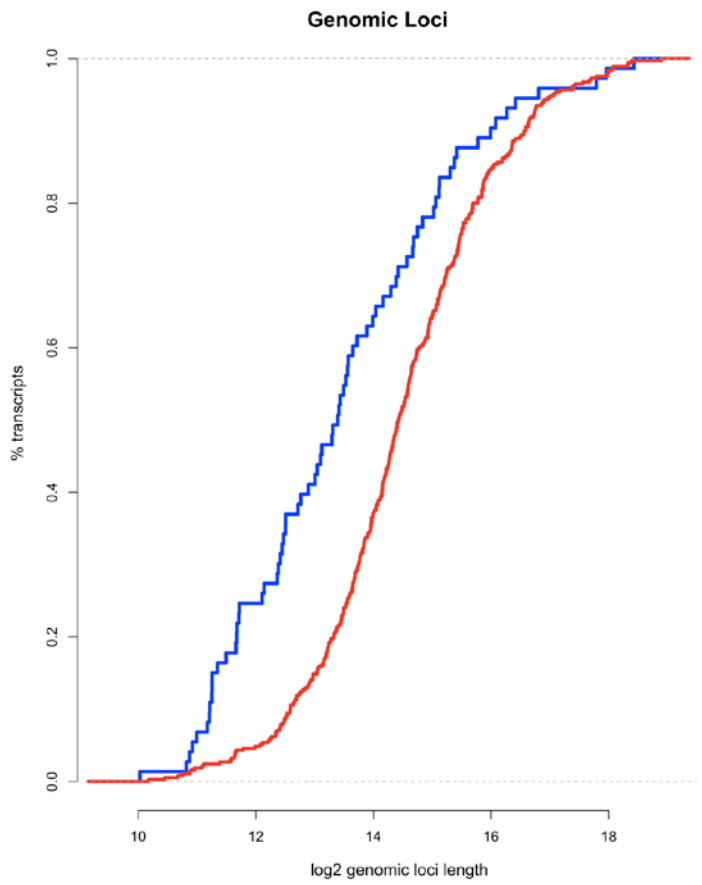


Loss of function screen



TF binding map

Early inflammatory genes are smaller, have larger enhancers and are farther away from other genes



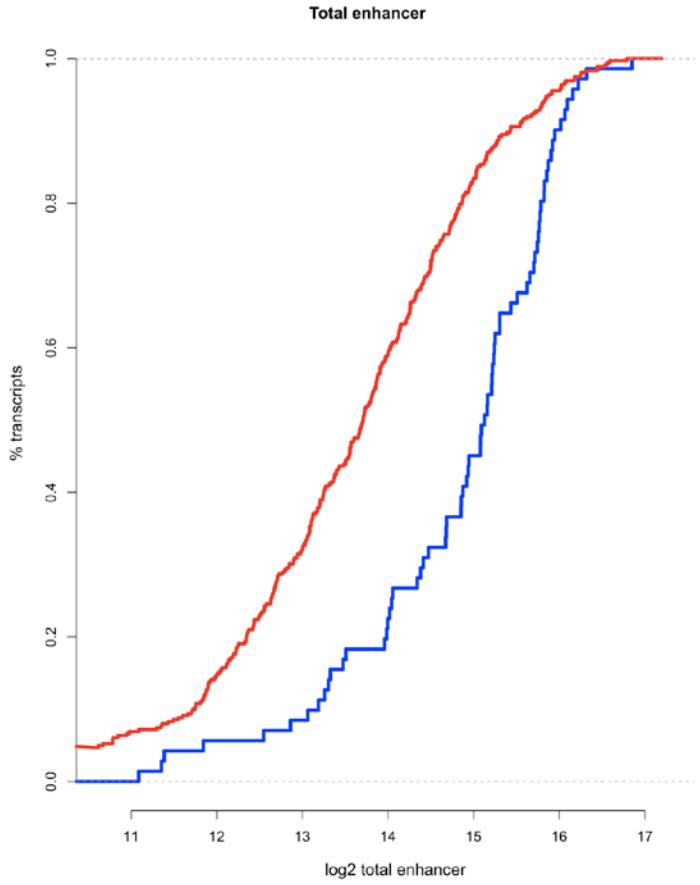


Figure S1 (part-II)

E

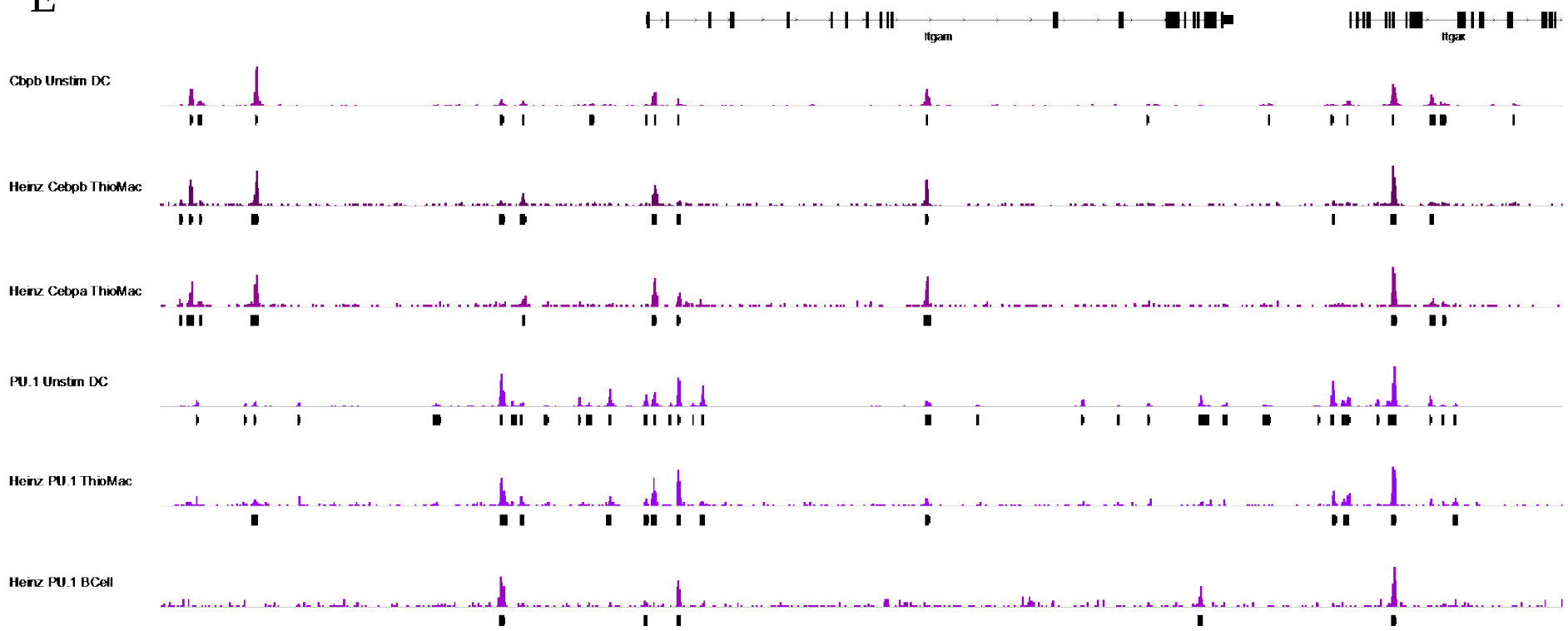


Figure S1 (part-I)

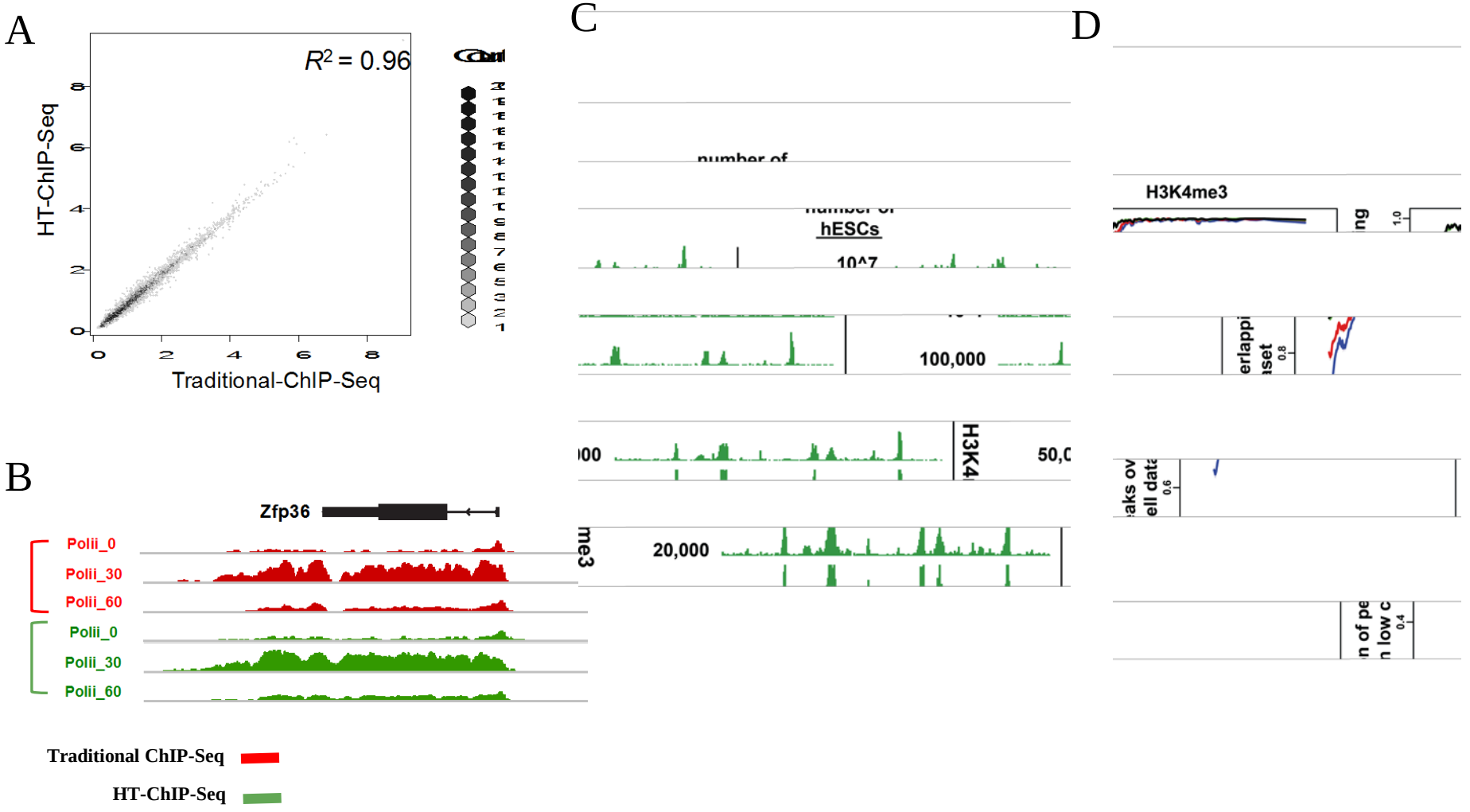
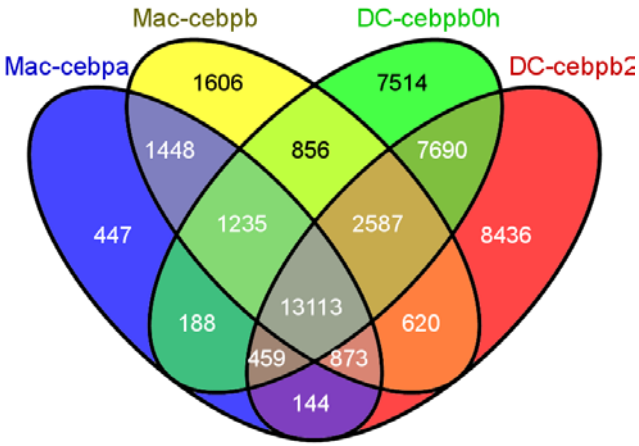


Figure S1 (part-III)

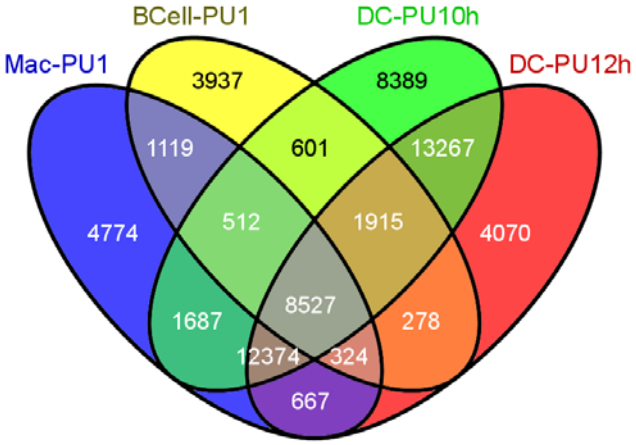
F

	DC Cebpb 0hr	DC PU1 0hr	Mac Cebpb	Mac PU1	Bcell PU1	Mac Cebpa	DC PU1 2hr	DC Cebpb 2hr
DC Cebpb 0hr	34467	14574	18060	11883	4601	15196	13360	24328
DC PU1 0hr	14448	47806	9846	23321	11644	8340	36487	14132
Mac Cebpb	18060	9884	22733	10718	3465	16895	9175	17457
Mac PU1	11859	23516	10734	30612	10604	9143	22283	12037
Bcell PU1	4534	11606	3437	10504	17328	2668	11094	4805
Mac Cebpa	15203	8361	16917	9139	2675	18201	7747	14781
DC PU1 2hr	13242	36532	9149	22116	11143	7731	41982	13874
DC Cebpb 2hr	24059	14198	17335	11993	4892	14680	13955	34643

G



H



I

