

University of Massachusetts Medical School

eScholarship@UMMS

---

UMass Center for Clinical and Translational  
Science Seminar Series

2012 UMCCTS Seminar Series

---

Nov 29th, 12:00 PM

## The Bioinformatics Core and The Garber Lab

Manuel Garber

*University of Massachusetts Medical School*

Let us know how access to this document benefits you.

Follow this and additional works at: [https://escholarship.umassmed.edu/umccts\\_seminars](https://escholarship.umassmed.edu/umccts_seminars)



Part of the [Bioinformatics Commons](#), [Computational Biology Commons](#), [Genomics Commons](#), and the [Translational Medical Research Commons](#)

---

Garber M. (2012). The Bioinformatics Core and The Garber Lab. UMass Center for Clinical and Translational Science Seminar Series. Retrieved from [https://escholarship.umassmed.edu/umccts\\_seminars/2012/seminars/6](https://escholarship.umassmed.edu/umccts_seminars/2012/seminars/6)

Creative Commons License



This work is licensed under a [Creative Commons Attribution 4.0 License](#).

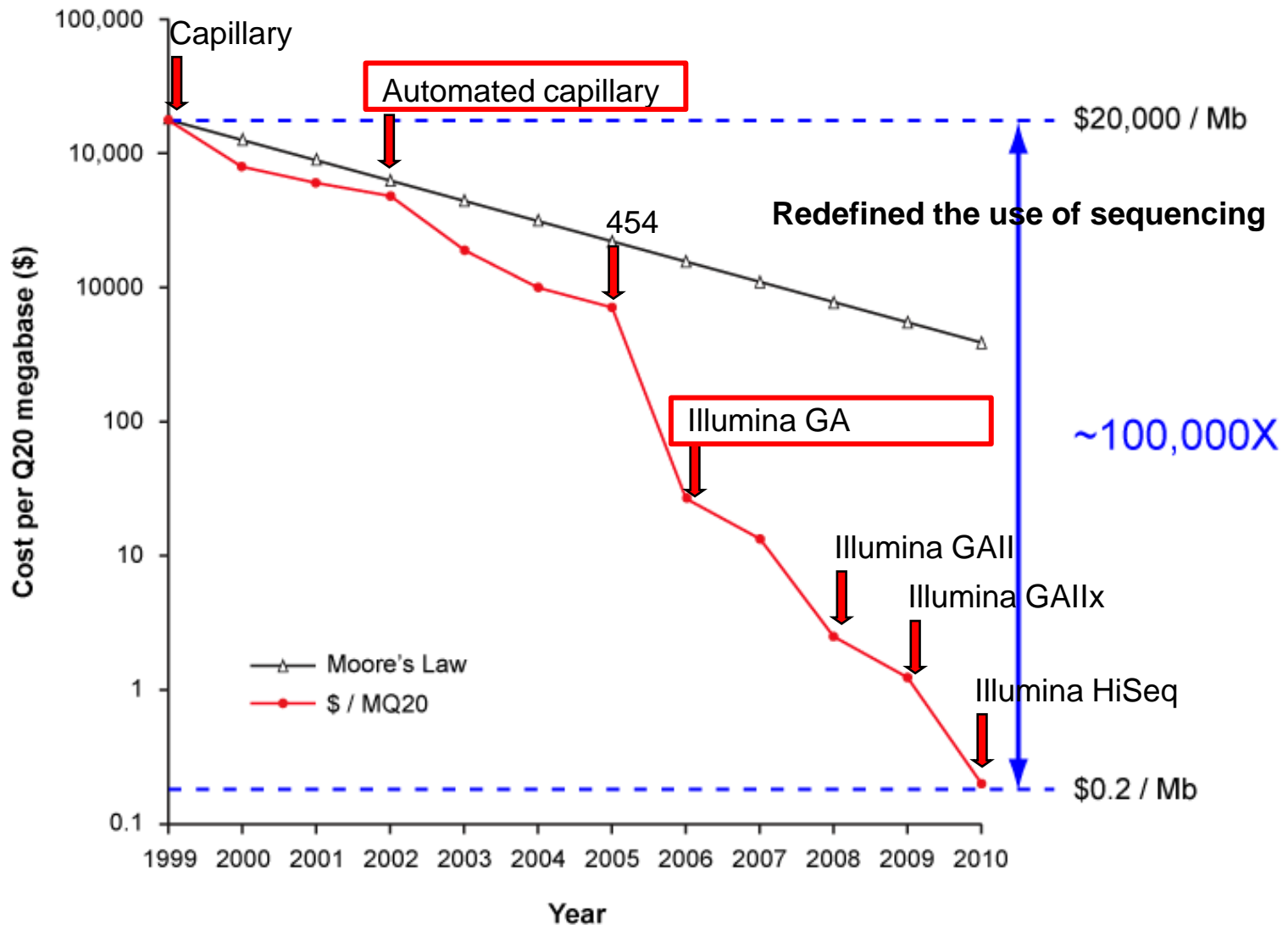
This material is brought to you by eScholarship@UMMS. It has been accepted for inclusion in UMass Center for Clinical and Translational Science Seminar Series by an authorized administrator of eScholarship@UMMS. For more information, please contact [Lisa.Palmer@umassmed.edu](mailto:Lisa.Palmer@umassmed.edu).

UMASS CCTS Seminar Series  
The Bioinformatics Core  
and  
The Garber Lab

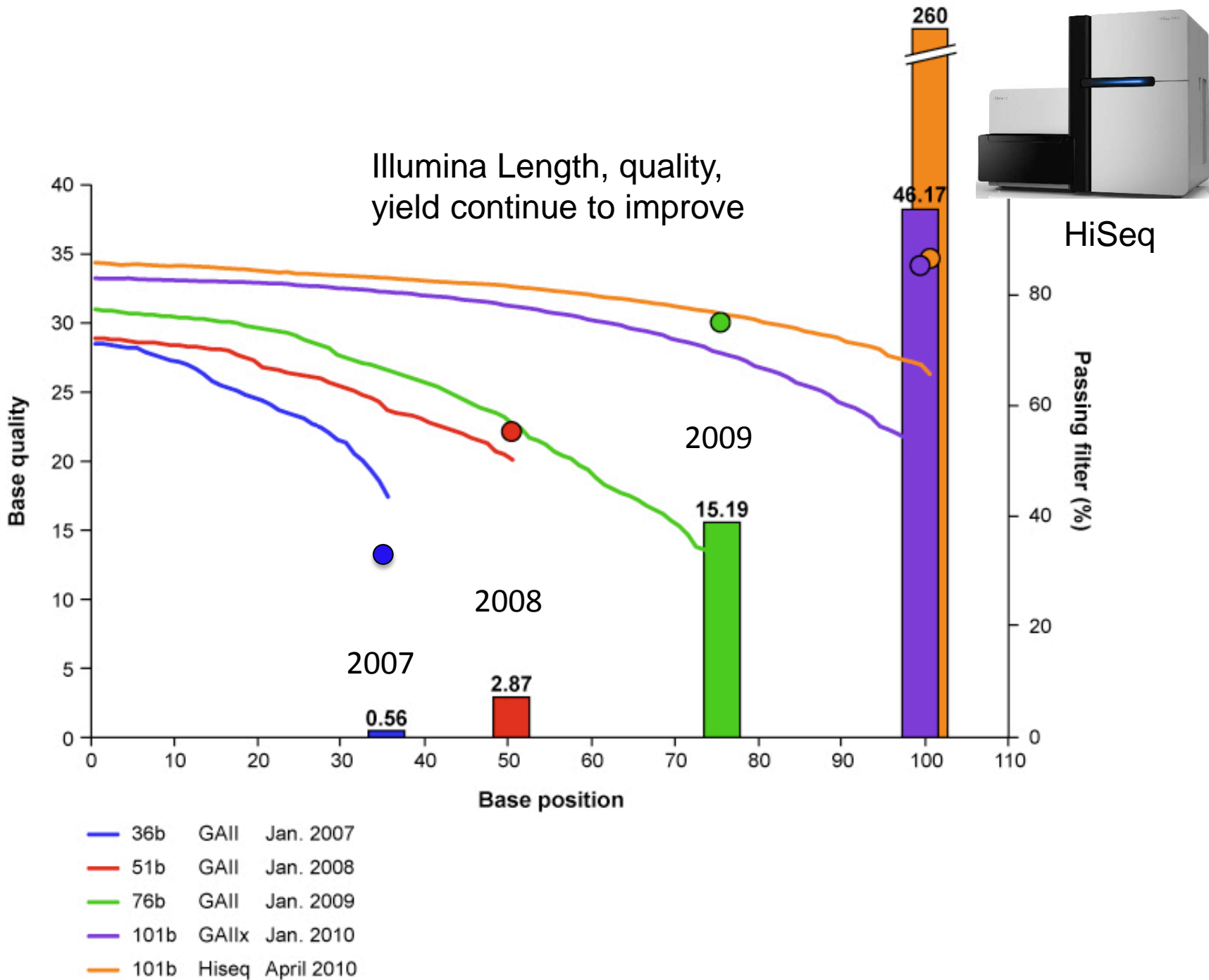
Manuel Garber

Nov 29<sup>th</sup> 2012

# Plummeting cost of DNA sequencing



Estimated cost per Mb in large scale centers



# 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.

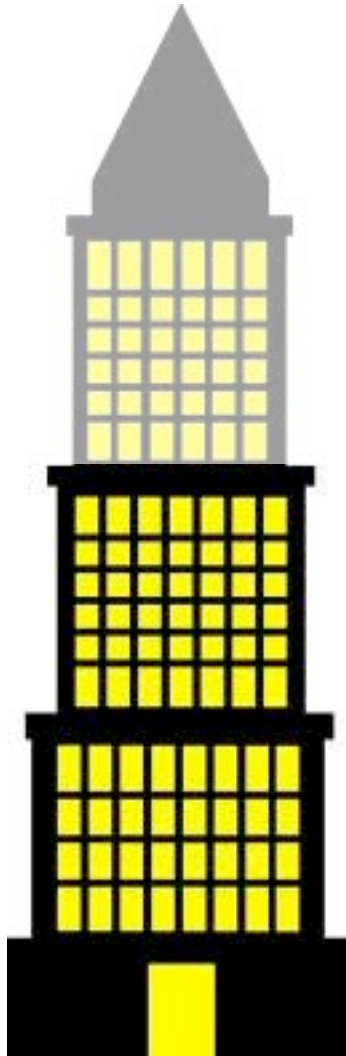
## **Data processing**

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*

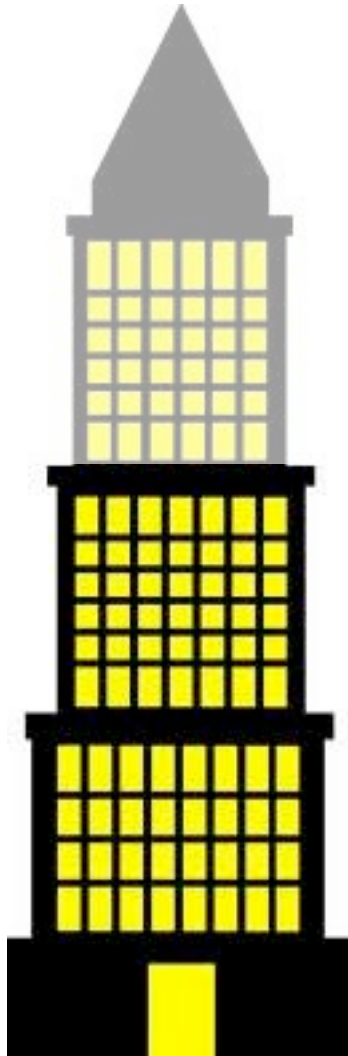
**Finding signal**, what are my differentially expressed genes, which peaks are in samples vs controls, etc

**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

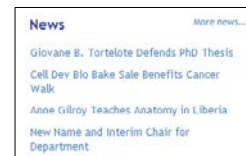
## Education:

- 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

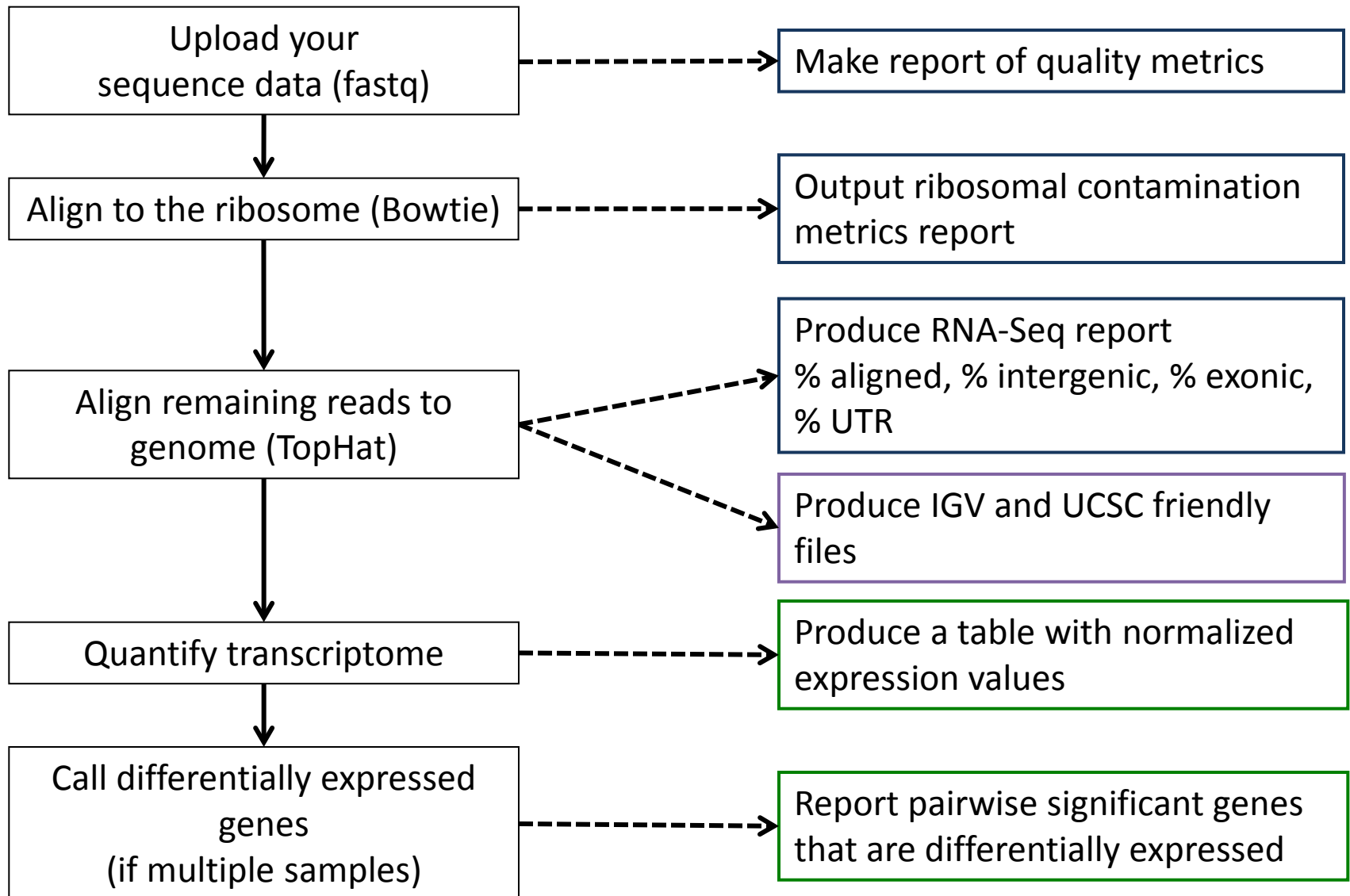
# Current Progress

- 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](mailto: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 HPC 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 everyone is comfortable with UNIX and scripting!**



# And there is a solution for this ...

https://main.g2.bx.psu.edu/root

Galaxy

Analyze Data Workflow Shared Data Visualization Cloud Help User

Tools

- SICER Statistical approach for the Identification of CHIP-Enriched Regions
- GeneTrack indexer on a BED file
- Peak predictor on GeneTrack index

NGS: RNA Analysis

RNA-SEQ

- Tophat for Illumina Find splice junctions using RNA-seq data**
- Cufflinks transcript assembly and FPKM (RPKM) estimates for RNA-Seq data
- Cuffcompare compare assembled transcripts to a reference annotation and track Cufflinks transcripts across multiple experiments
- Cuffmerge merge together several Cufflinks assemblies
- Cuffdiff find significant changes in transcript expression, splicing, and promoter use
- eXpress Quantify the abundances of a set of target sequences from sampled subsequences

FILTERING

- Filter Combined Transcripts using tracking file

NGS: Picard (beta)

- FASTQ to BAM creates an aligned BAM file

## Tophat for Illumina (version 1.5.0)

RNA-Seq FASTQ file:

Nucleotide-space: Must have Sanger-scaled quality values with ASCII offset 33

Will you select a reference genome from your history or use a built-in index?:

Use a built-in index

Built-ins were indexed using default options

Select a reference genome:

Arabidopsis lyrata: Araly1

If your genome of interest is not listed, contact the Galaxy team

Is this library mate-paired?:

Single-end

TopHat settings to use:

Full parameter list

Use the Full parameter list to change default settings.

Execute

### Tophat Overview

TopHat is a fast splice junction mapper for RNA-Seq reads. It aligns RNA-Seq reads to mammalian-sized genomes using the ultra high-throughput short read aligner Bowtie, and then analyzes the mapping results to identify splice junctions between exons. Please cite: Trapnell, C., Pachter, L. and Salzberg, S.L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* 25, 1105-1111 (2009).

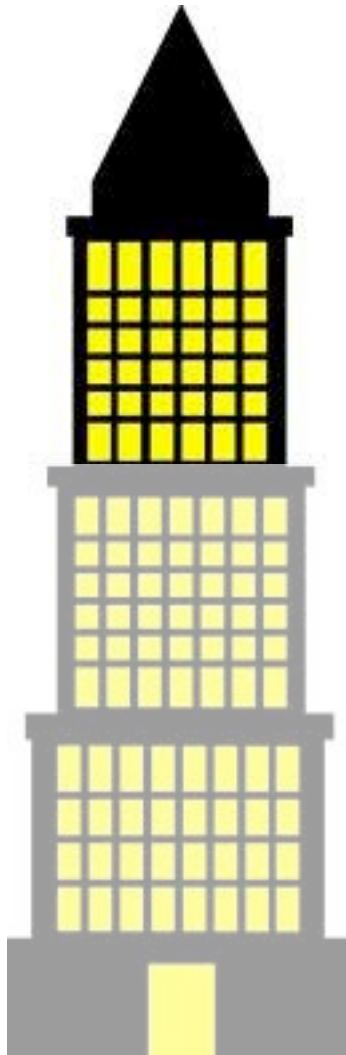
### Know what you are doing

⚠ There is no such thing (yet) as an automated gearshift in splice junction identification. It is all like stick-shift driving in San Francisco. In other words, running this tool with default parameters will probably not give you meaningful results. A way to deal with this is to **understand** the parameters by carefully reading the [documentation](#) and experimenting. Fortunately, Galaxy makes experimenting easy.

### Input formats

Tophat accepts files in Sanger FASTQ format. Use the FASTQ Groomer to prepare your files.

# 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

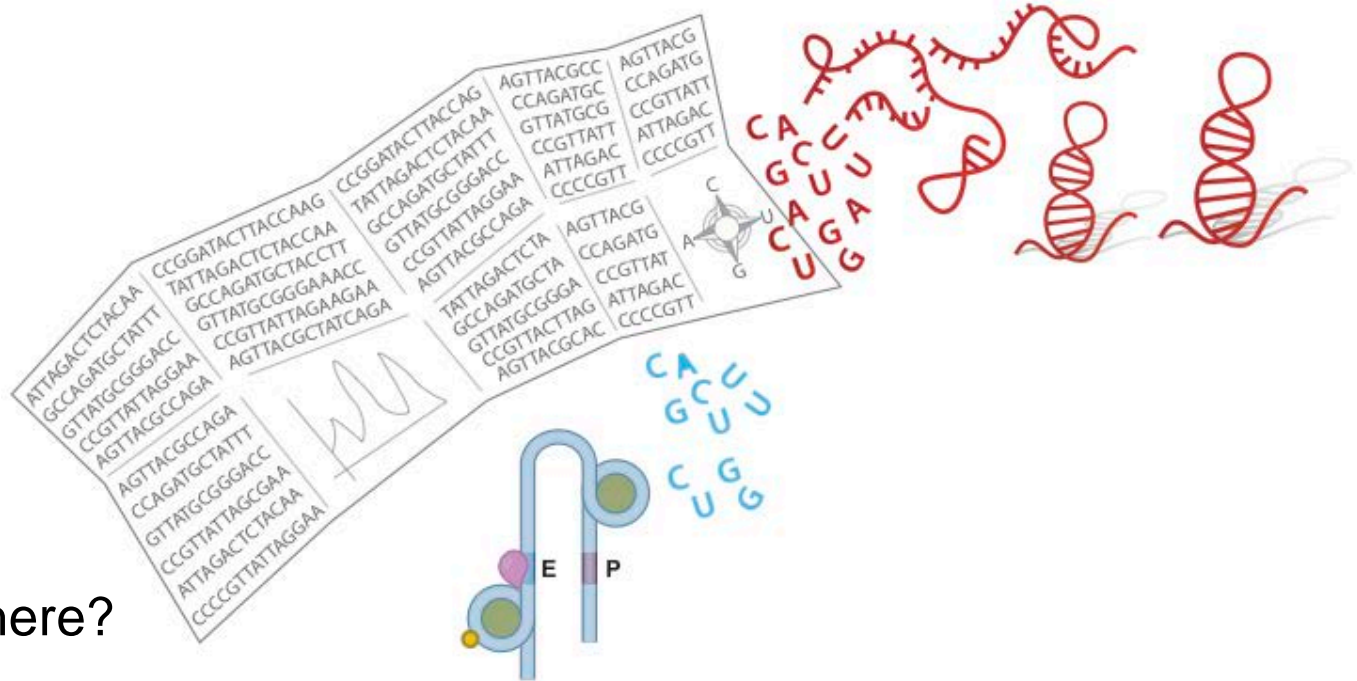
# Help for the last step

- Discussion:
  - Monthly meeting, 2<sup>nd</sup> 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](mailto: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?**

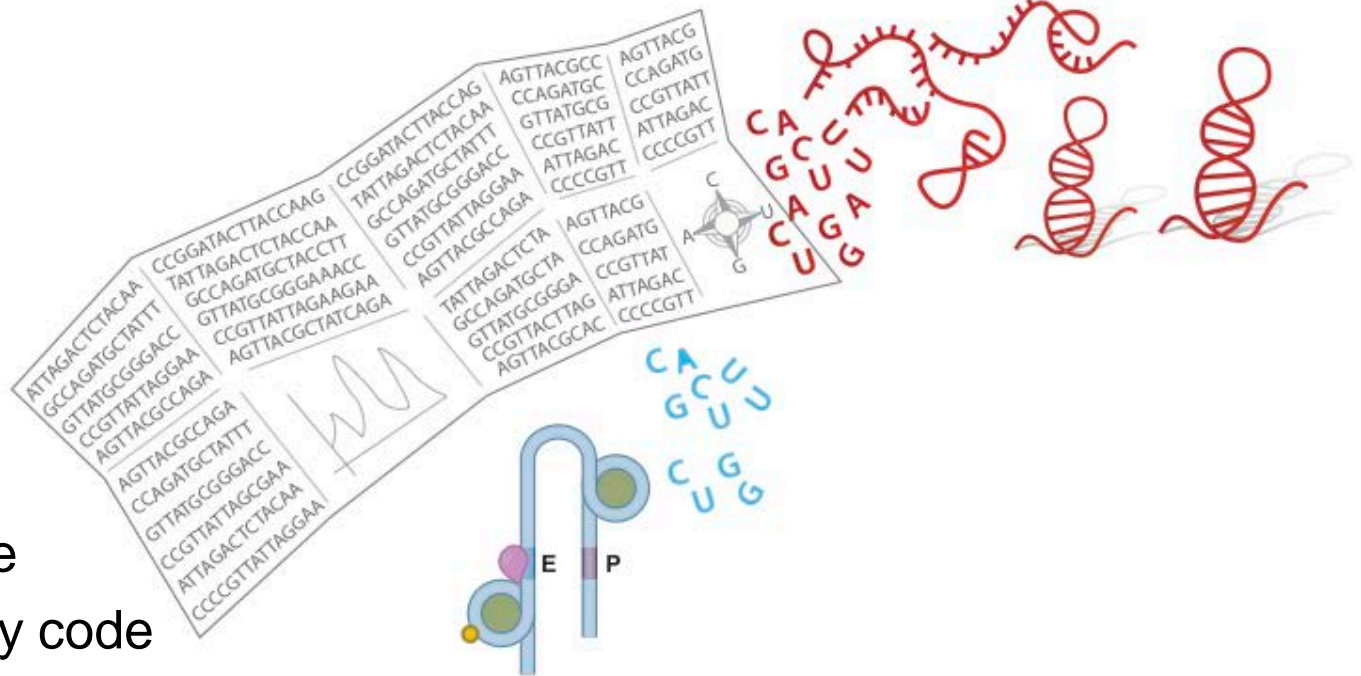
# The genome encodes many different elements



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



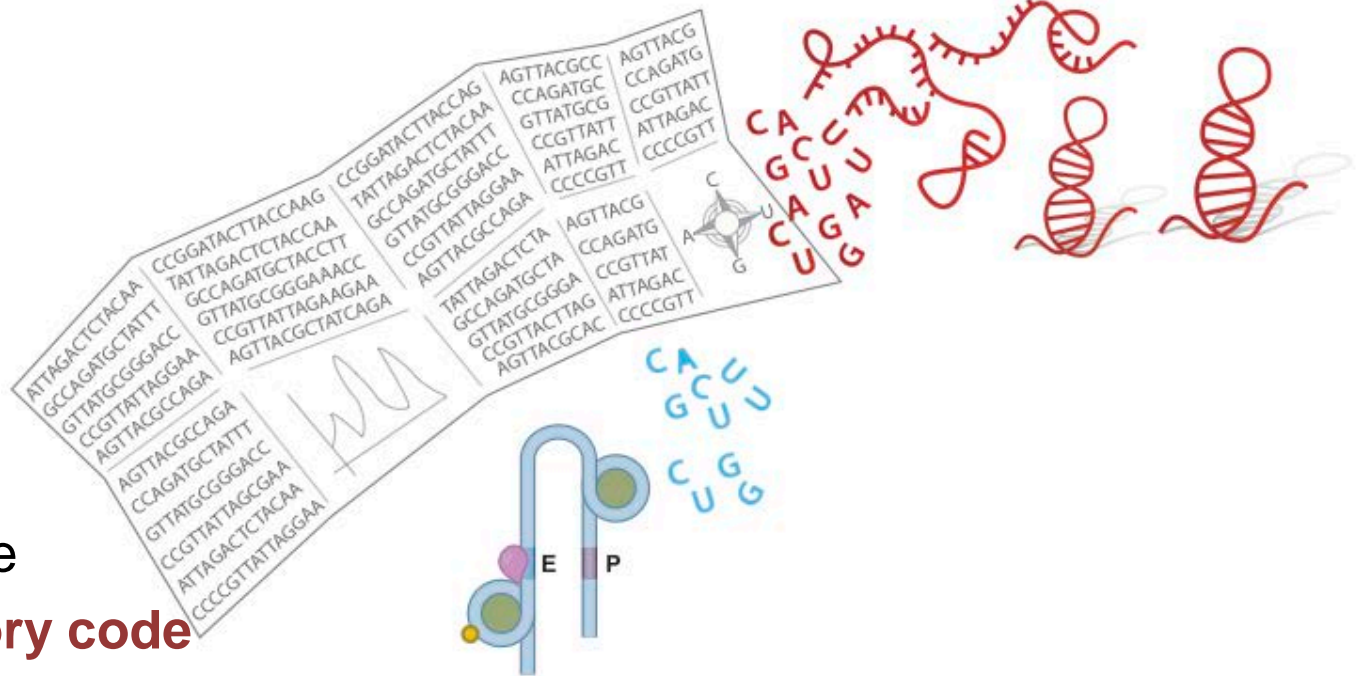
# Most of which we do not understand



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



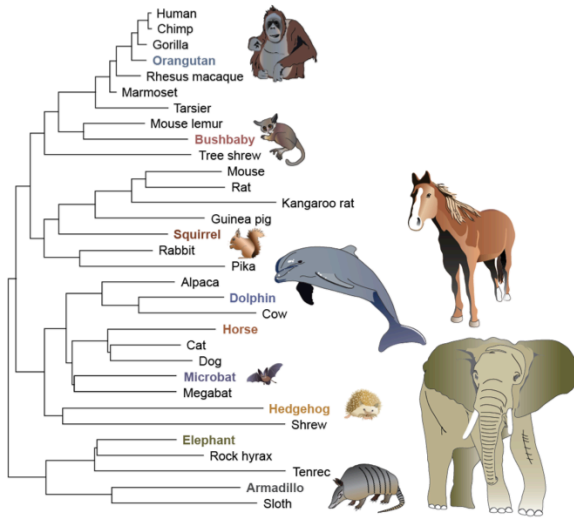
# Most of which we do not understand



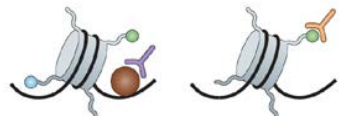
- ✓ Genetic Code
- ? **Cis-regulatory code**
- ? RNAi and miRNA codes
- ? *RNA Code*
- ? Histone Code



# Our work



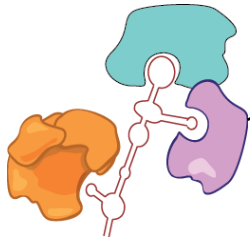
Estimate the “functional genome” by finding what is under selection



ChIP



RNA

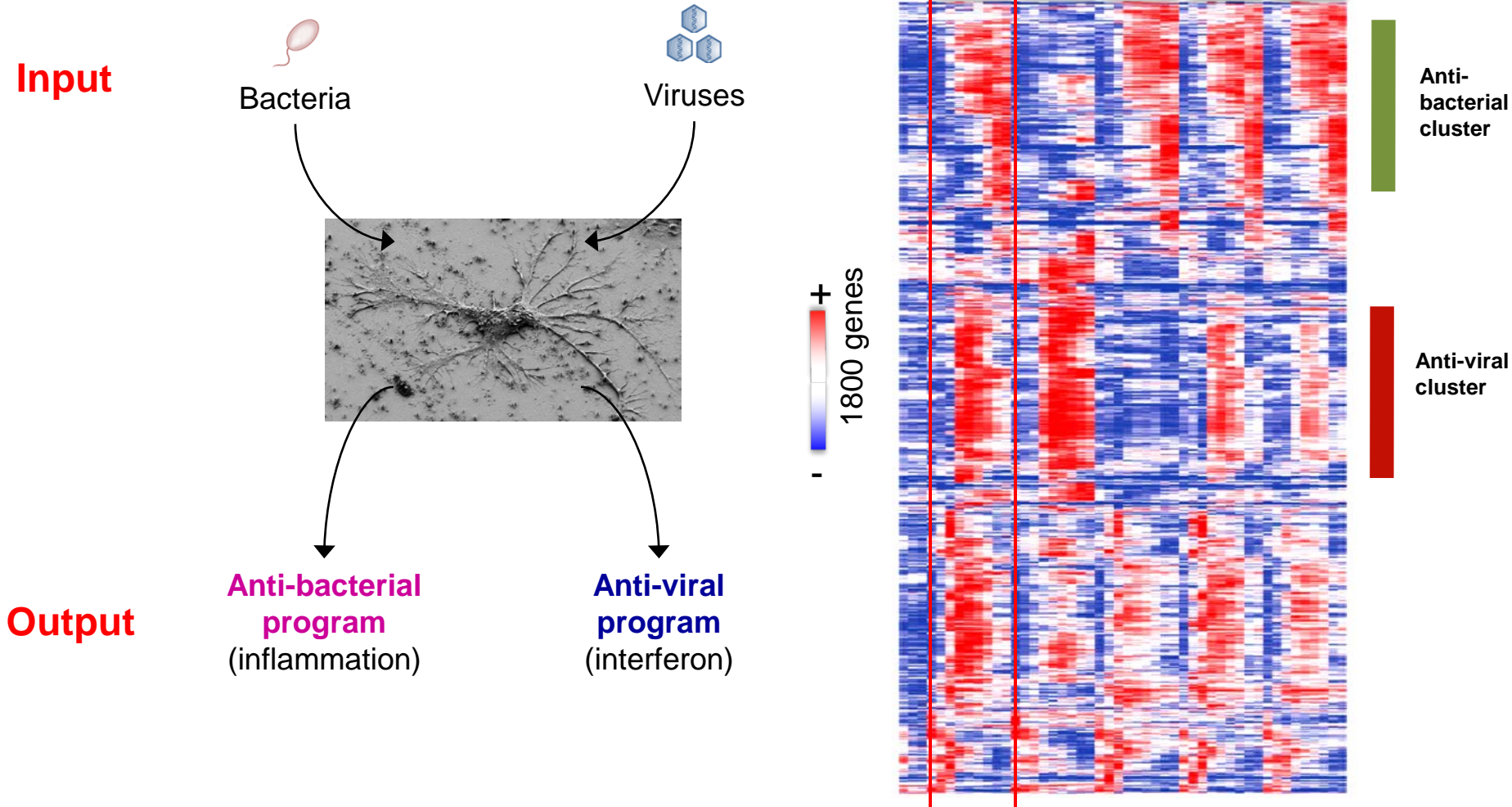


RNA-Protein interactions



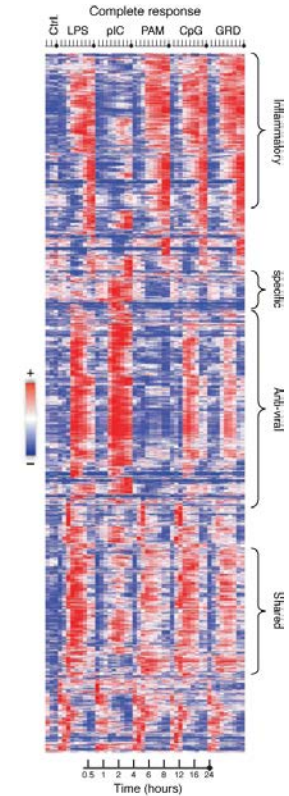
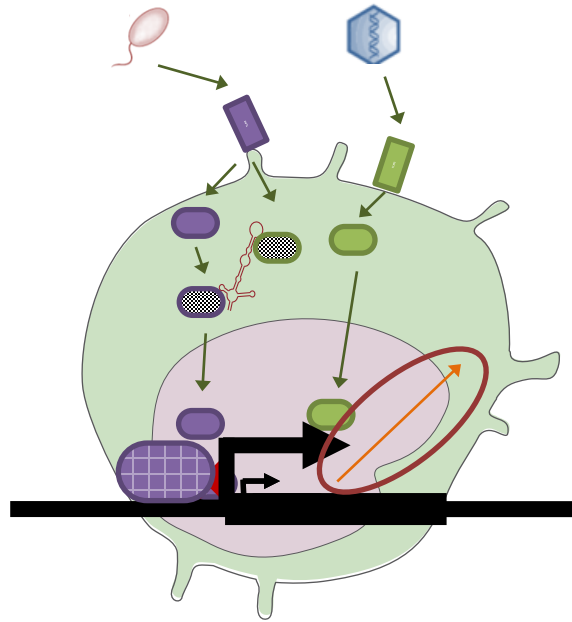
- Develop informatics tools for new methods
- Develop models of transcriptional regulation
- Develop models of epigenetic interactions
- Evolution of large non-coding RNAs

# Project: Transcription regulation in DCs



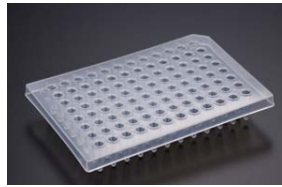
**How is this response controlled?**

# Strategy: Genetic + physical mapping



**What are the direct targets of transcription factors?**

# Only possible with High throughput ChIP-Sequencing



**Magnetic beads**



**Robot-automated**

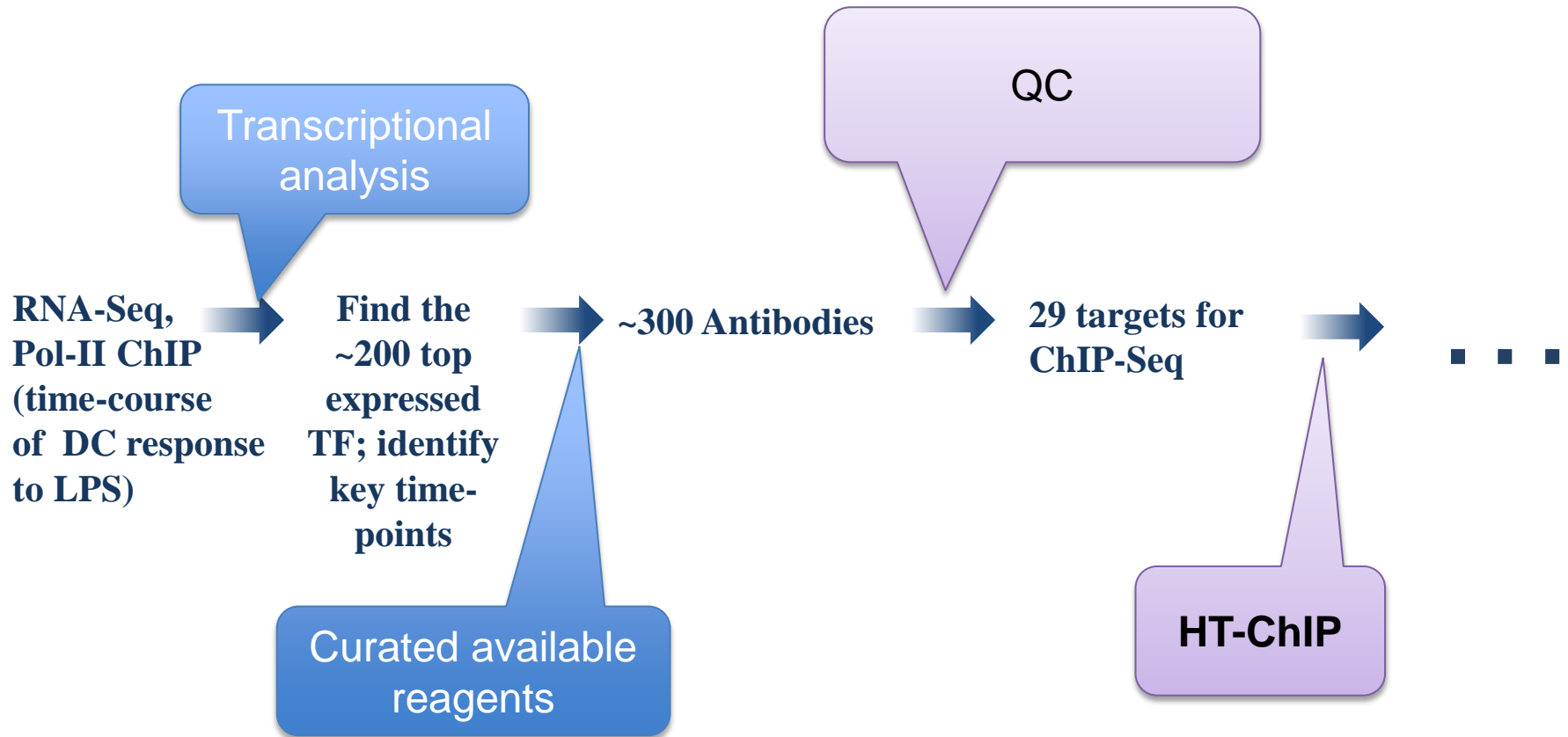


**24 libraries/lane**

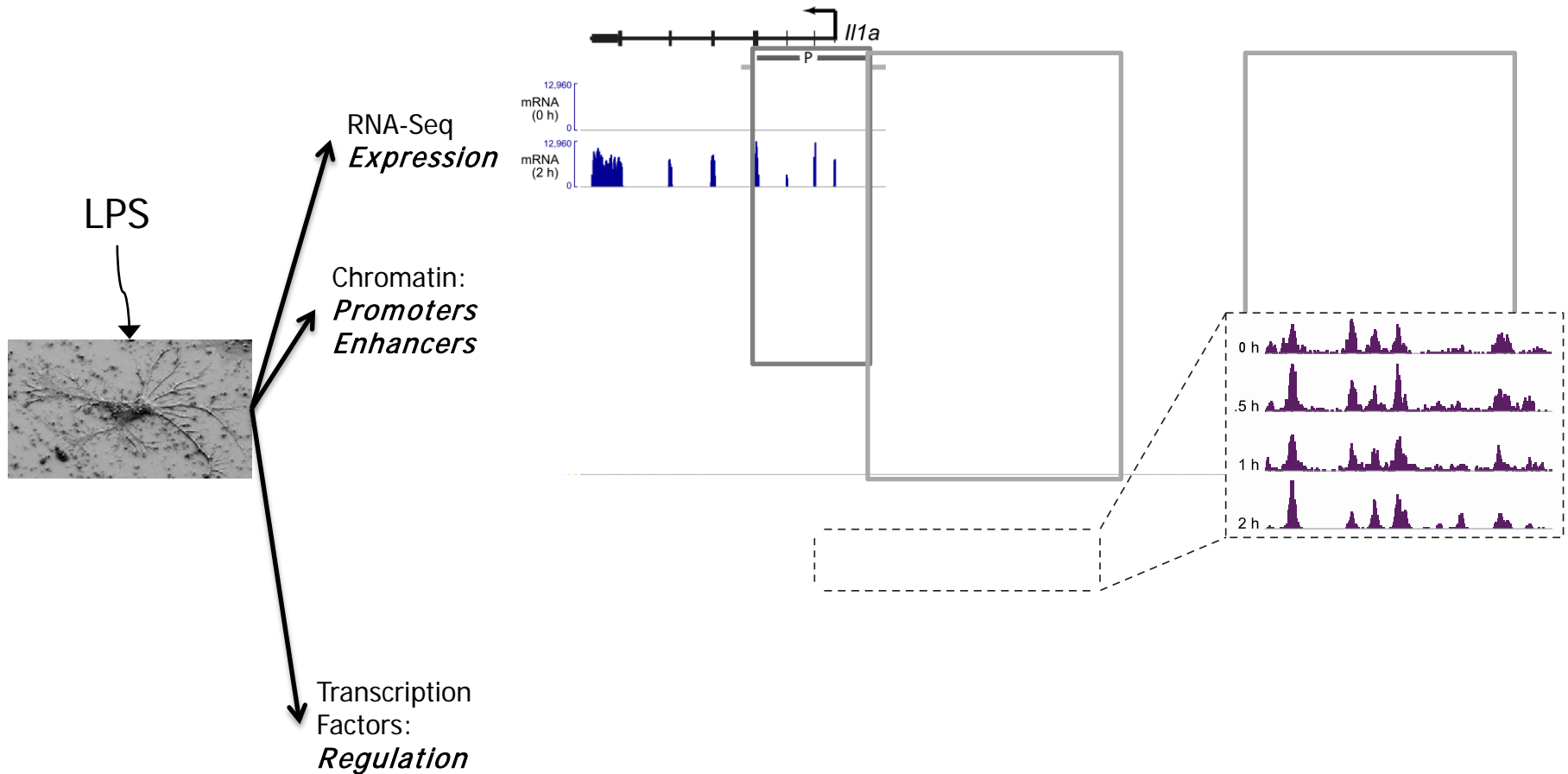


Ido Amit

# Systematic mapping of the DC LPS-response network

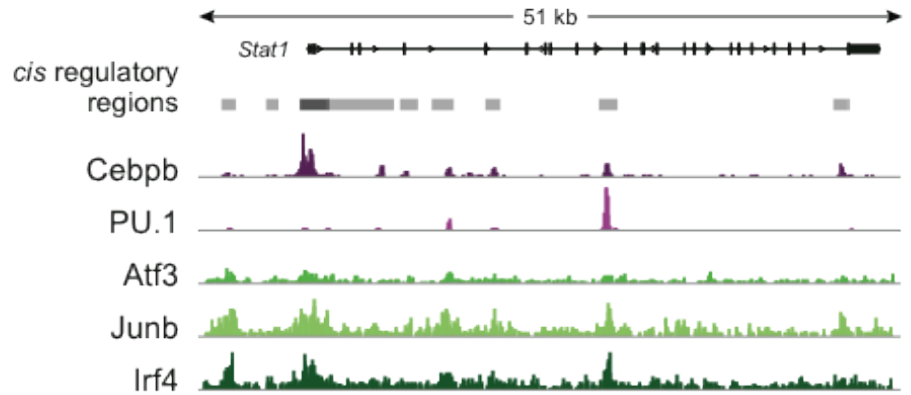
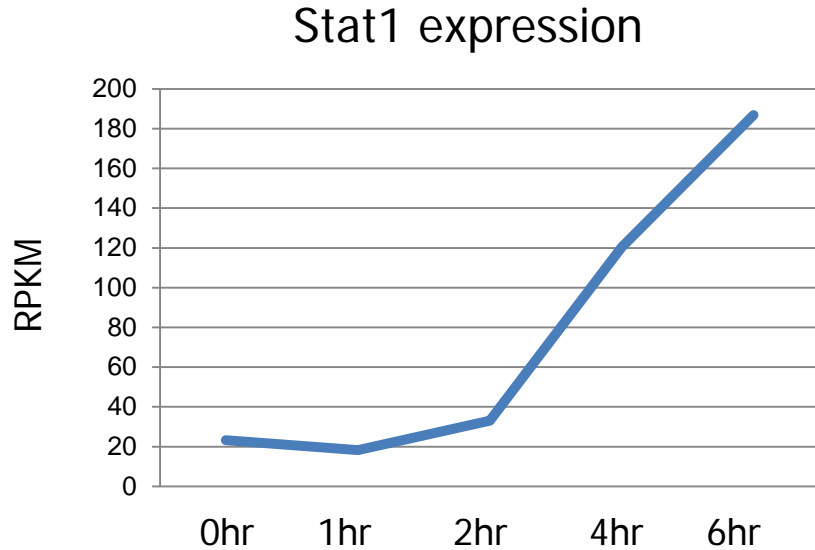


# Dataset: temporal view of expression and state



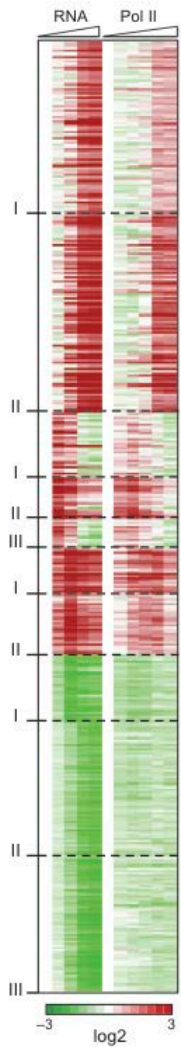
**85% of high scoring TF peaks fall within annotated *cis*-regulatory regions**

# 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

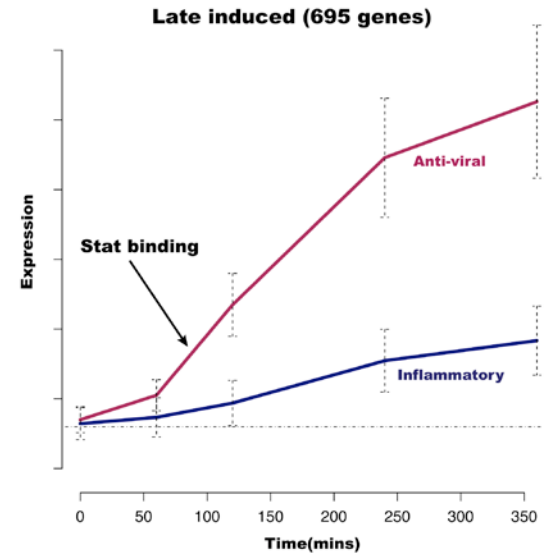
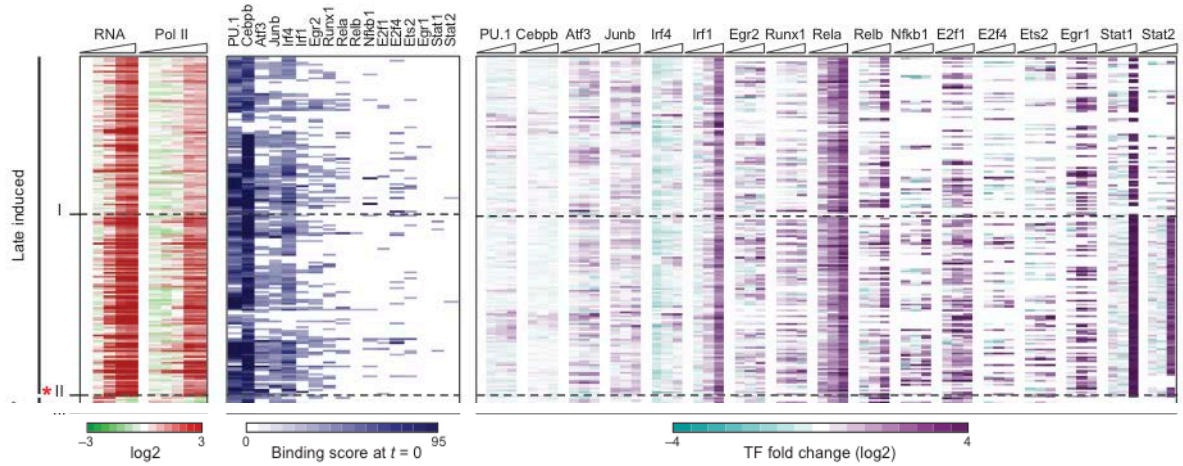
**Inflammation  
Inflammation  
Anti-apoptotic**

**Cell cycle**

**What are the differences between sub-clusters?**

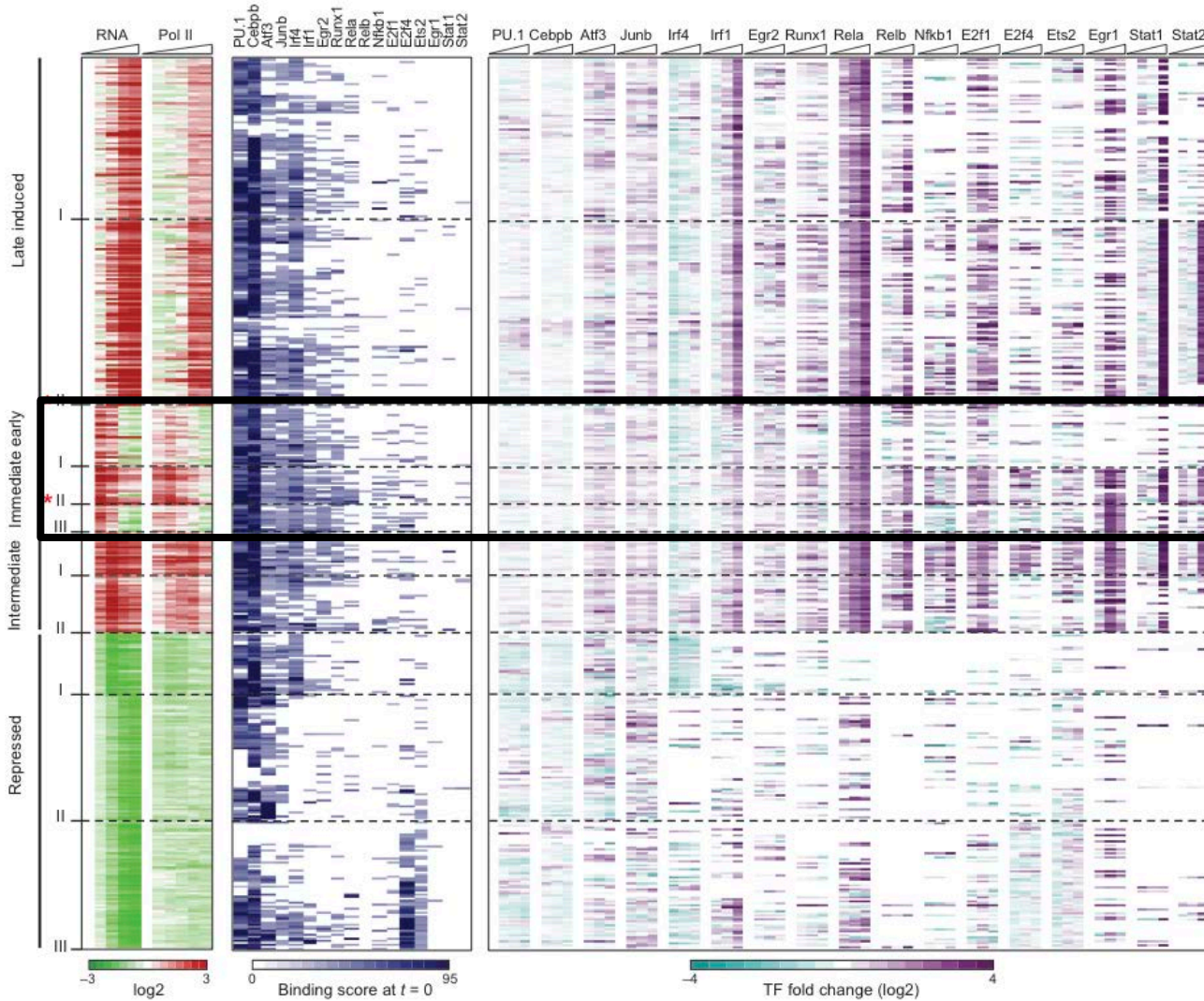


# Specific factors control amplitude of expression



**Binding of stat1/2 controls inductions levels**

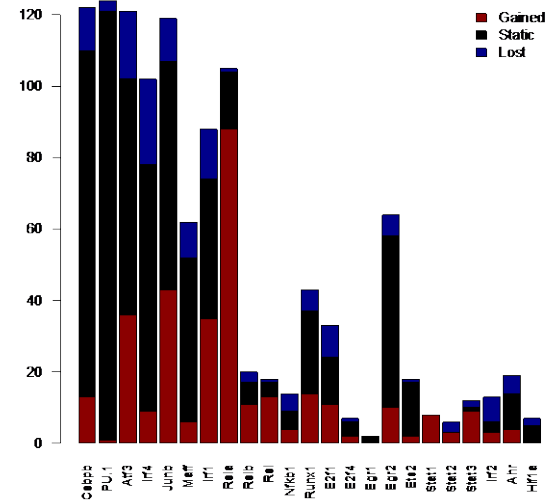
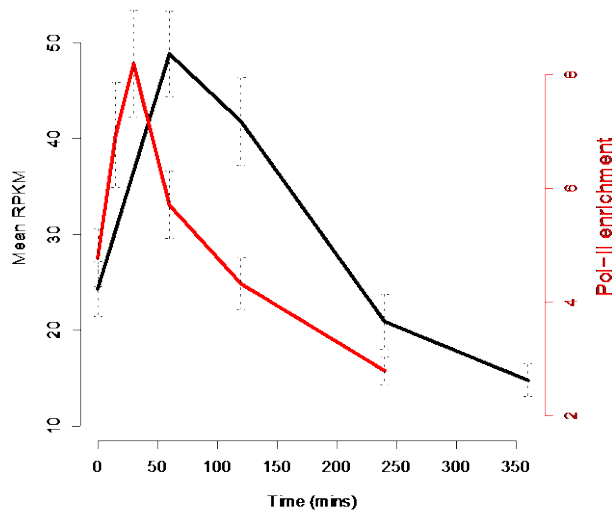
# Immediate early genes are highly bound



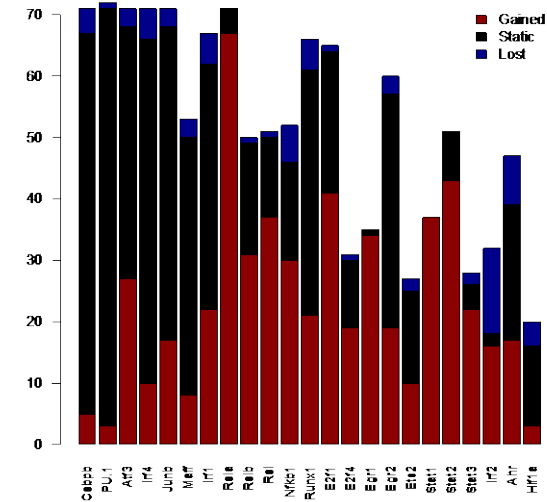
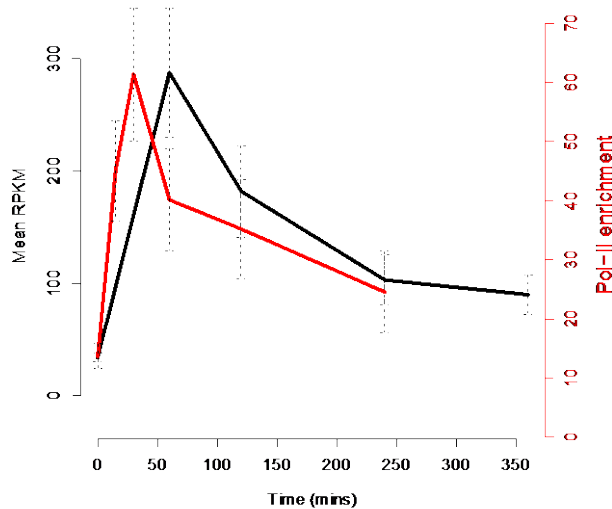
What are the differences between sub-clusters?

# Immediate early gene programs

Immediate Early 1 (125 genes)



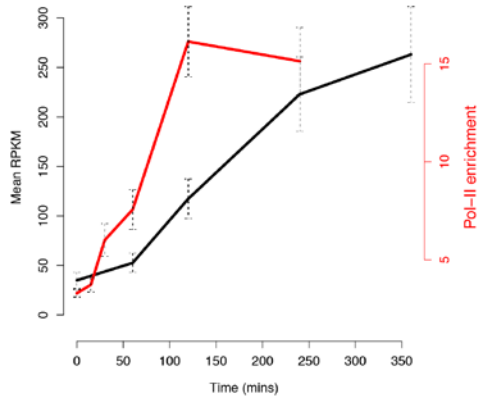
Immediate Early 2 (73 genes)



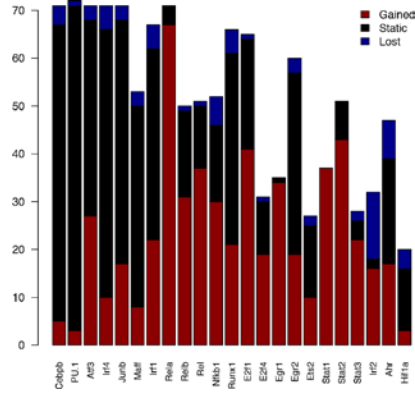
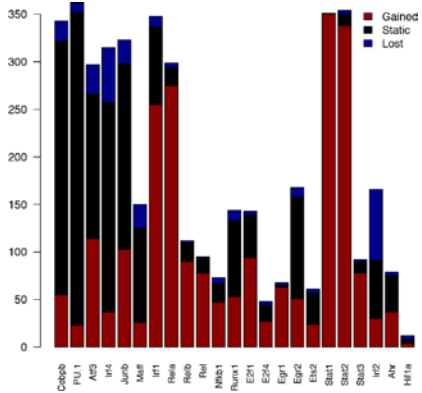
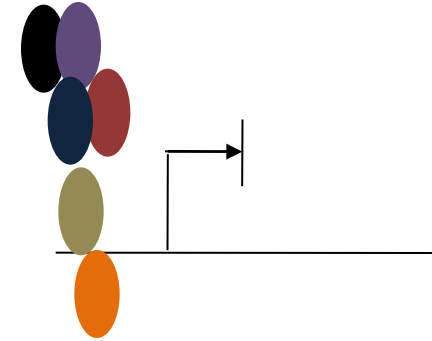
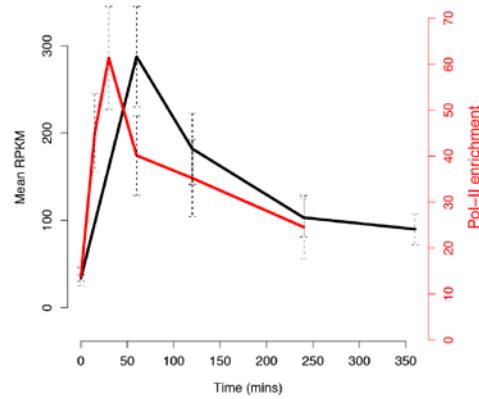
# Current models under consideration

Two forms of regulation?

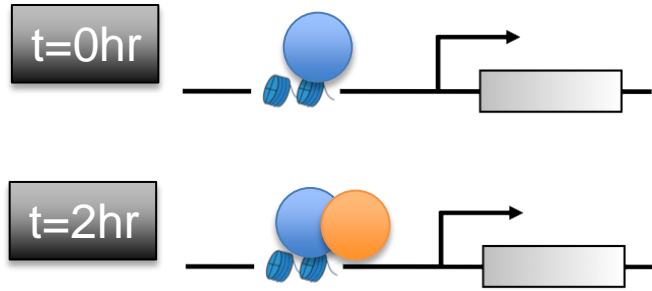
Late induced  
Stat regulated



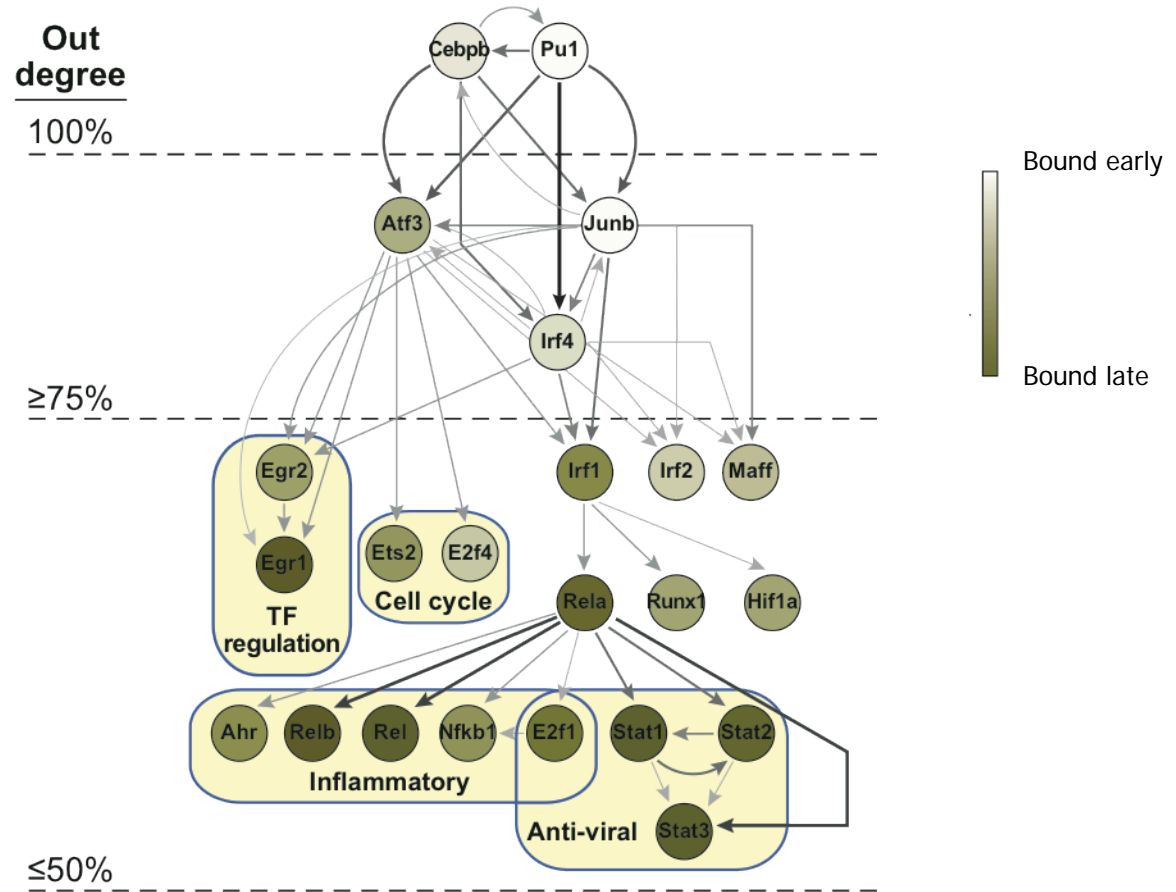
Immediate early  
highly induced



# Regulatory modes are established hierarchically



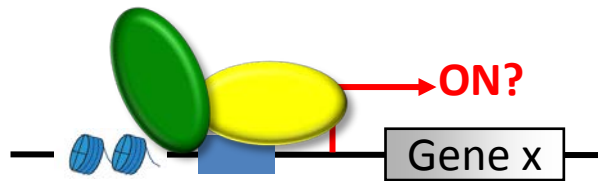
PU.1 coincides with or precedes Stat1 binding



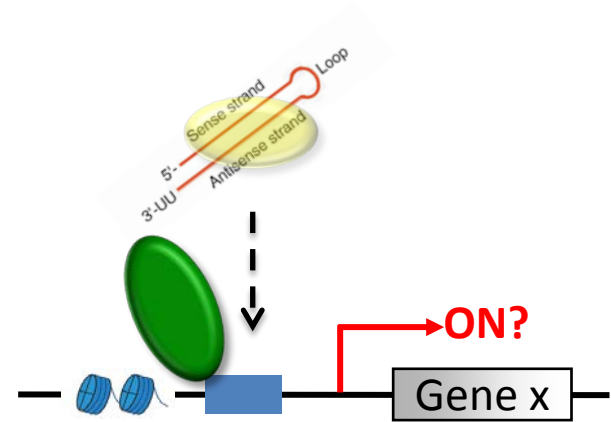
# Conclusions and considerations

- 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

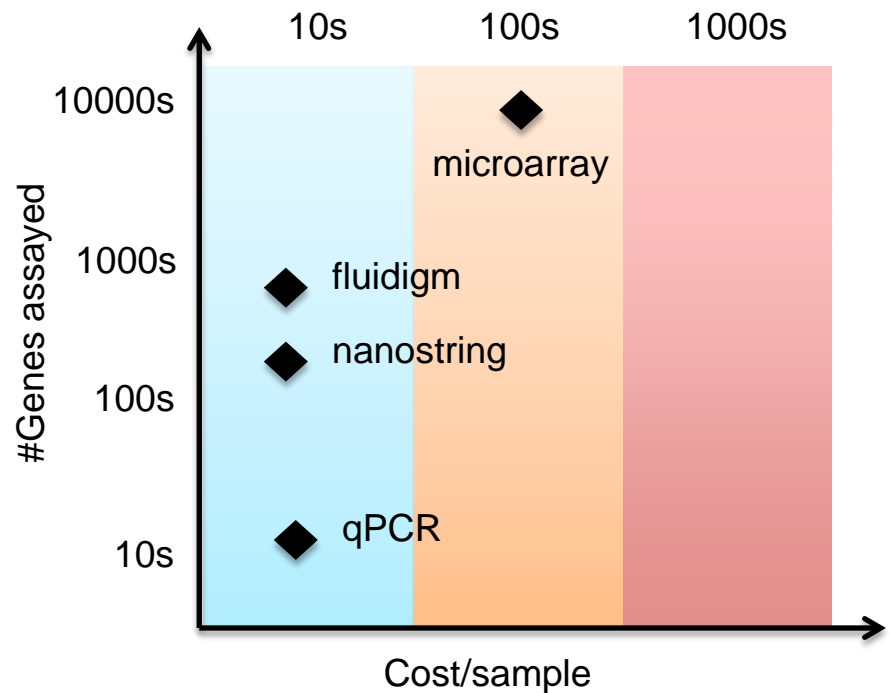
X high cost

X limited starting material

## Previous solution

- qPCR
- Fluidigm
- Luminex
- NanoString
- microarray

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



# An expensive proposition...

~100 genes KD \* replicates = LARGE NUMBER of samples

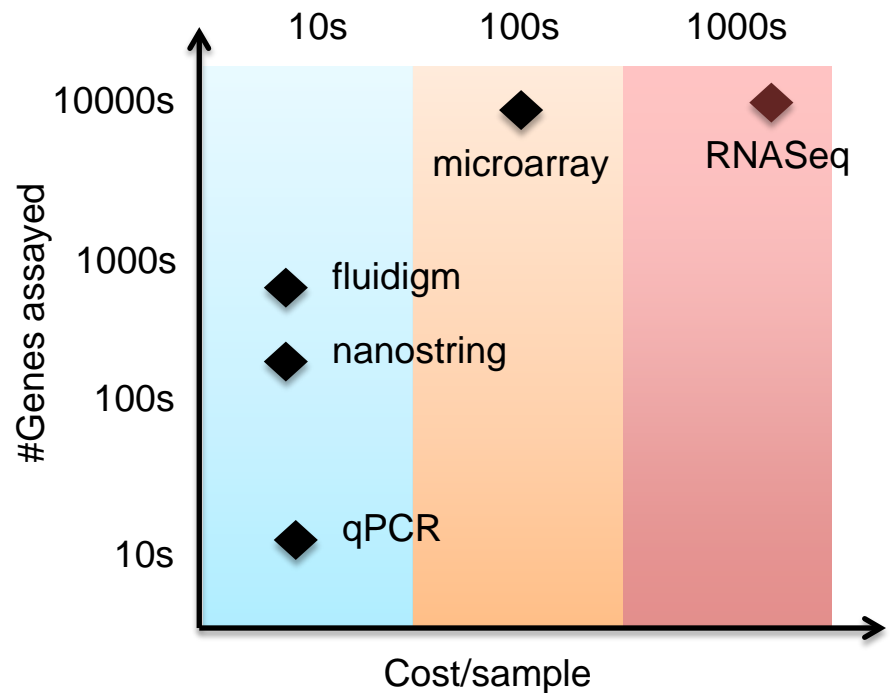
X high cost

X limited starting material

## Previous solution

- qPCR
- Fluidigm
- Luminex
- NanoString
- microarray

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



# Goal: Cheap RNA-Seq for quantification

~100 genes KD \* replicates = LARGE NUMBER of samples

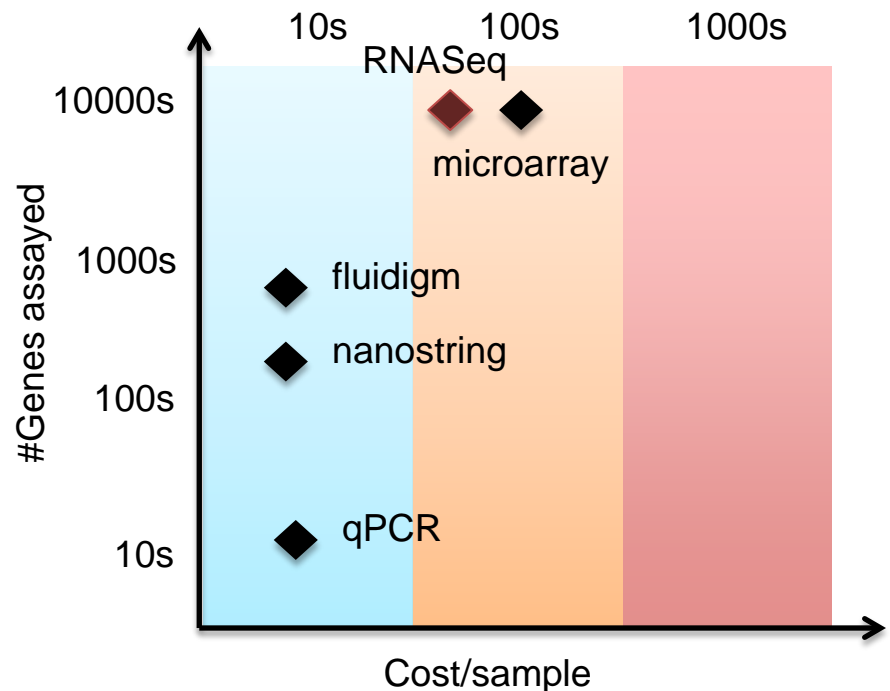
X high cost

X limited starting material

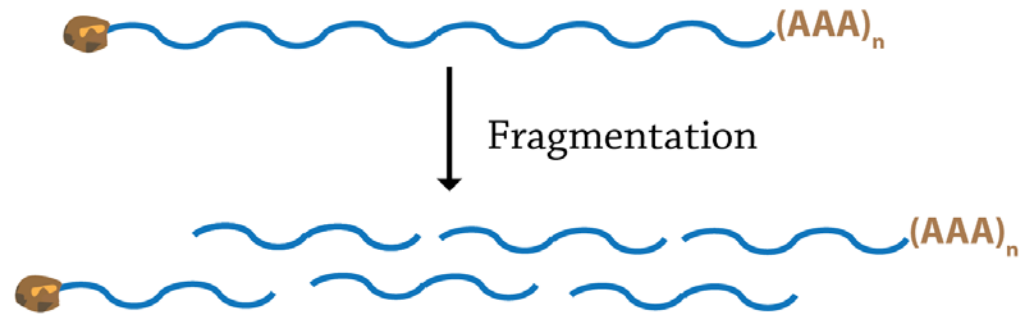
## Previous solution

- qPCR
- Fluidigm
- Luminex
- NanoString
- microarray

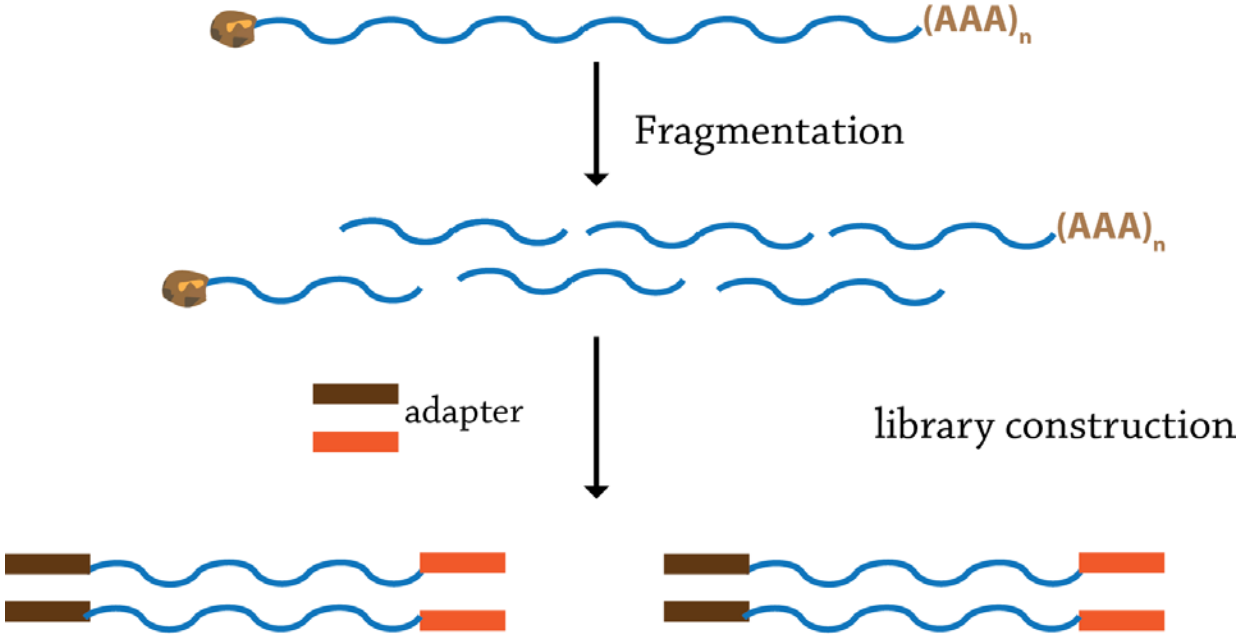
Problem: need to choose your genes in advance and limited #genes assayed



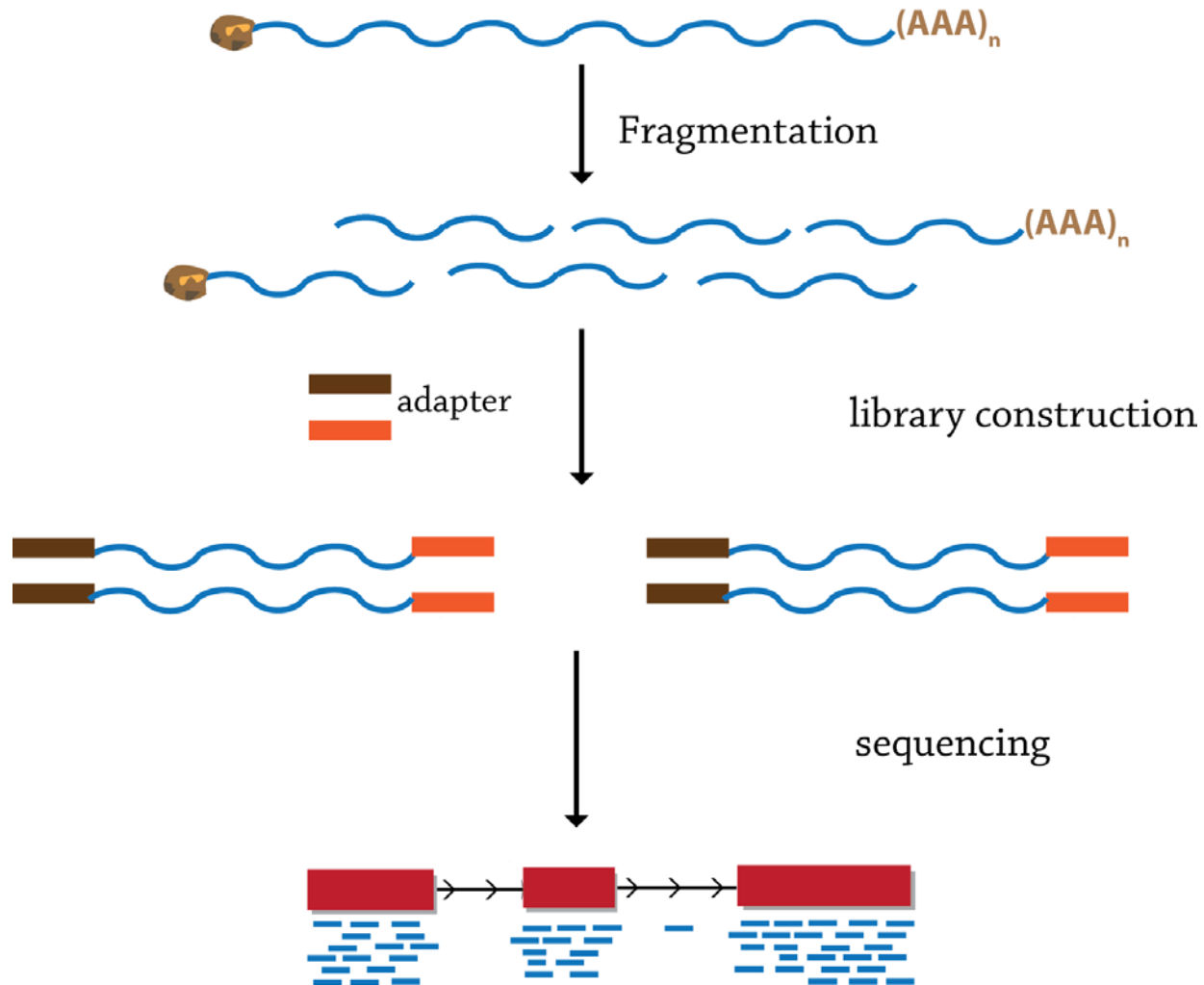
# Full length RNA-Sequencing



# Full length RNA-Sequencing



# Full length RNA-Sequencing



# End RNA-Seq



Motivation

Protocol

Pipeline

Quantification

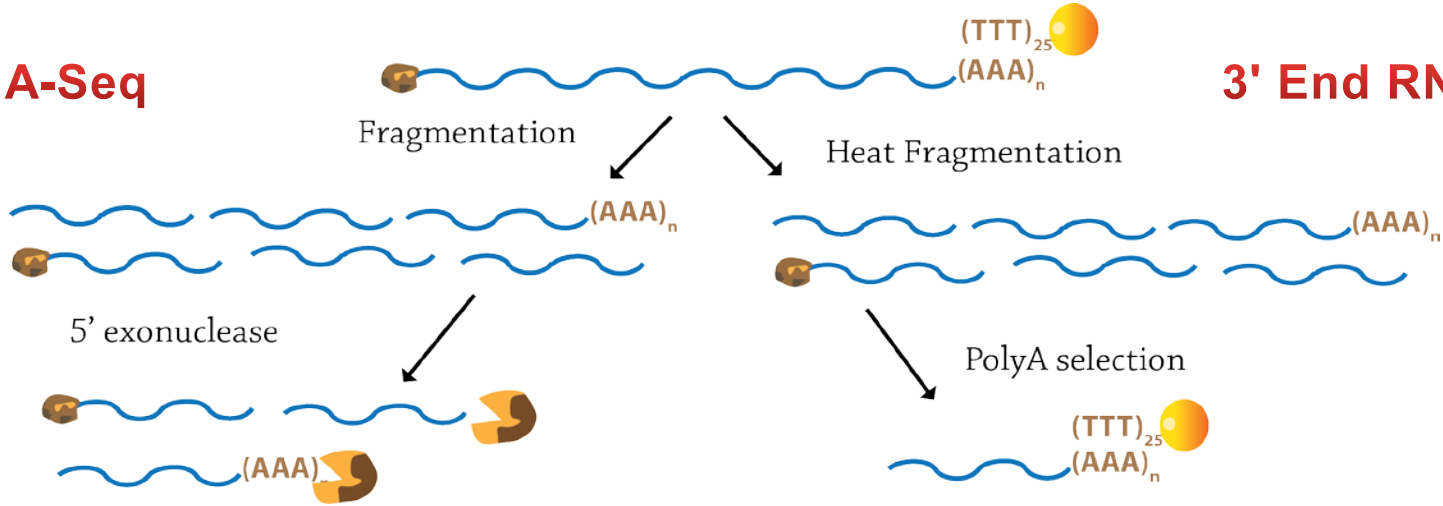
Diff Expr

Annotation

# End RNA-Seq

## 5' End RNA-Seq

## 3' End RNA-Seq



Motivation

Protocol

Pipeline

Quantification

Diff Expr

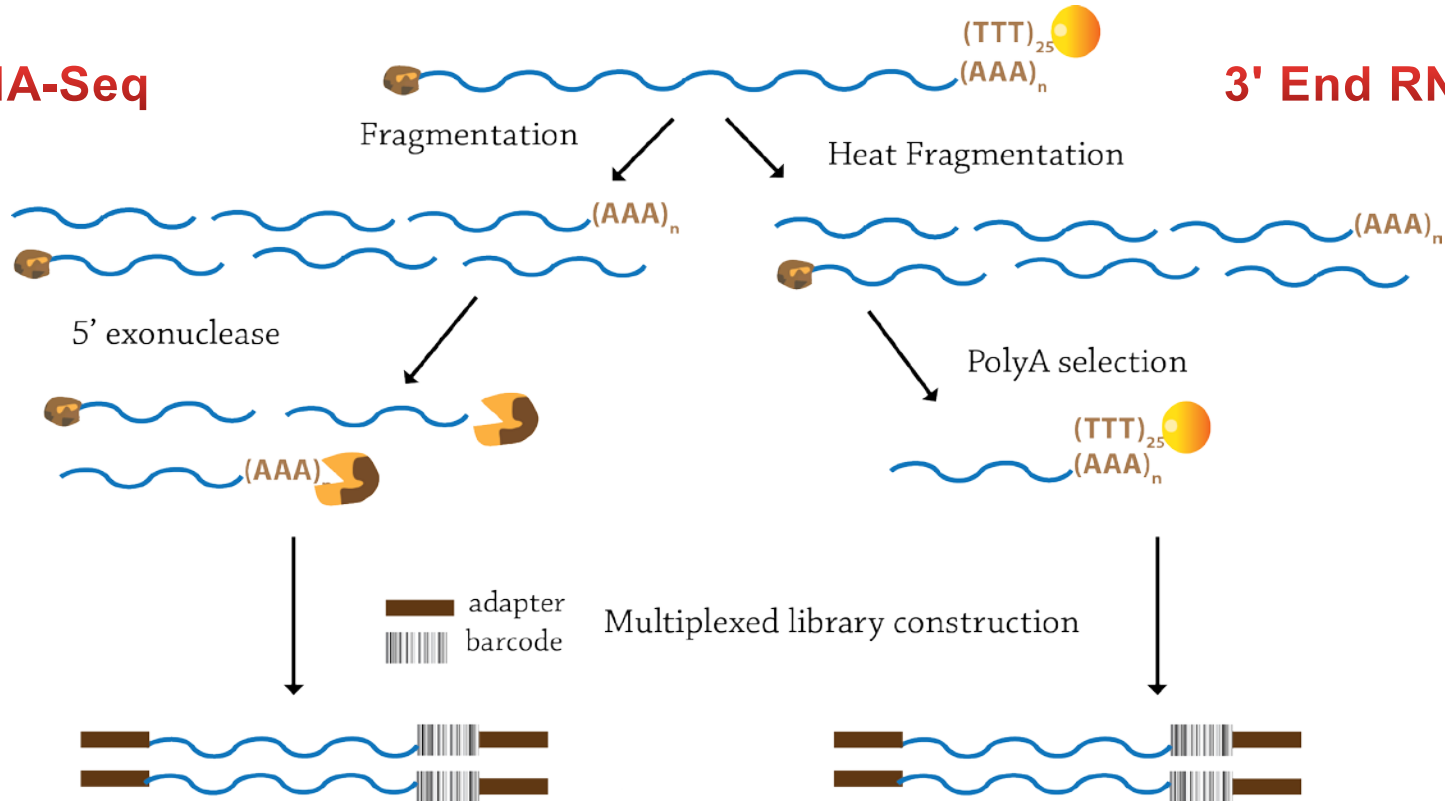
Annotation



# End RNA-Seq

## 5' End RNA-Seq

## 3' End RNA-Seq



Motivation

Protocol

Pipeline

Quantification

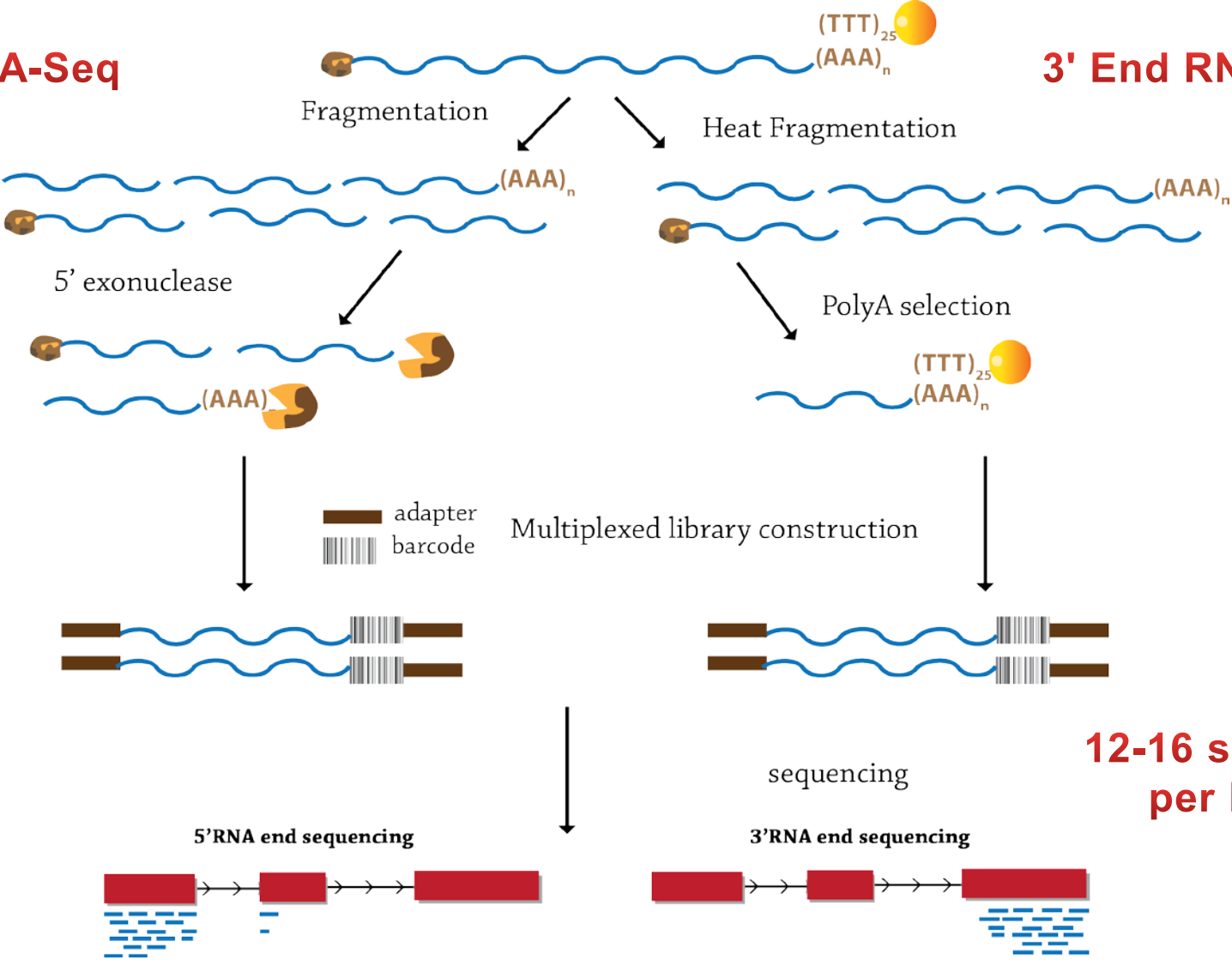
Diff Expr

Annotation

# End RNA-Seq

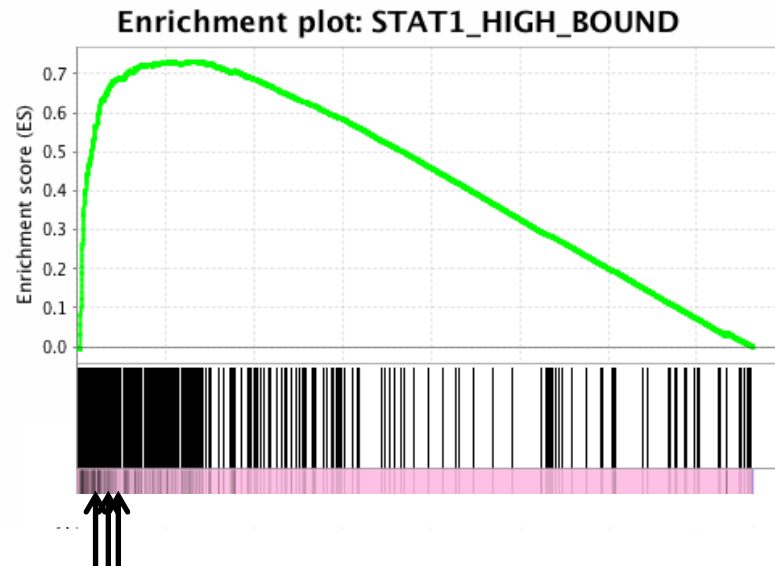
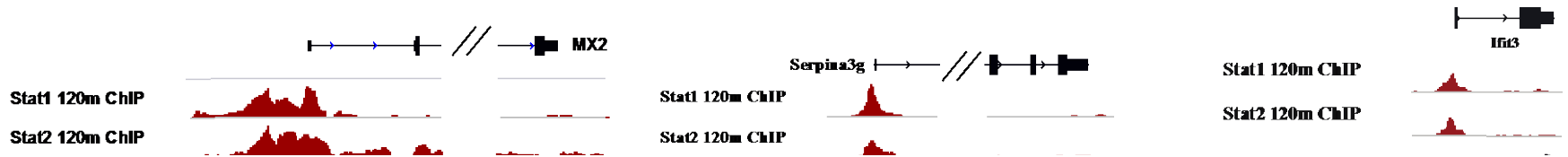
## 5' End RNA-Seq

## 3' End RNA-Seq



**12-16 samples per lane**

# 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)**

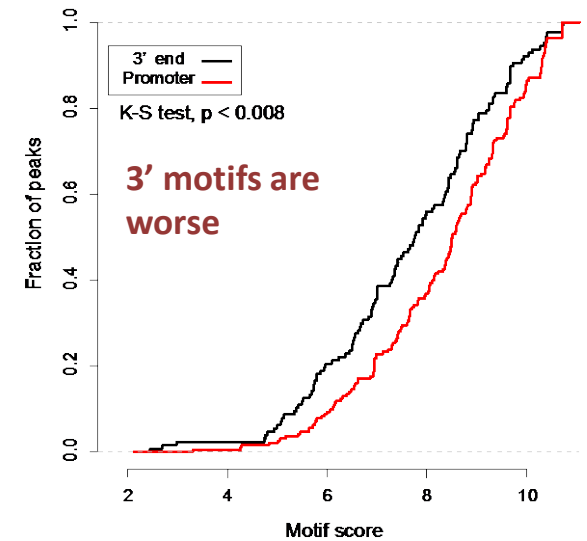
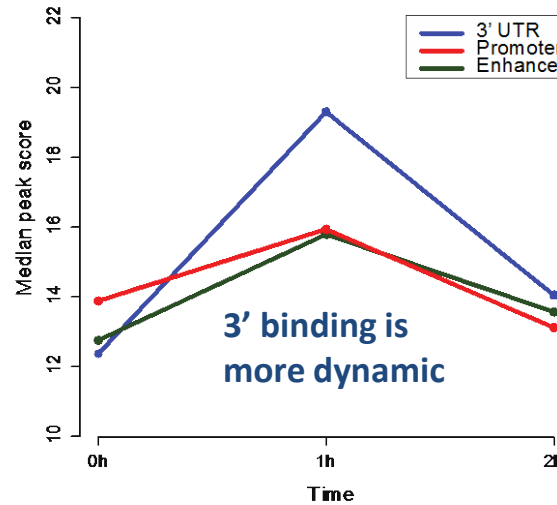
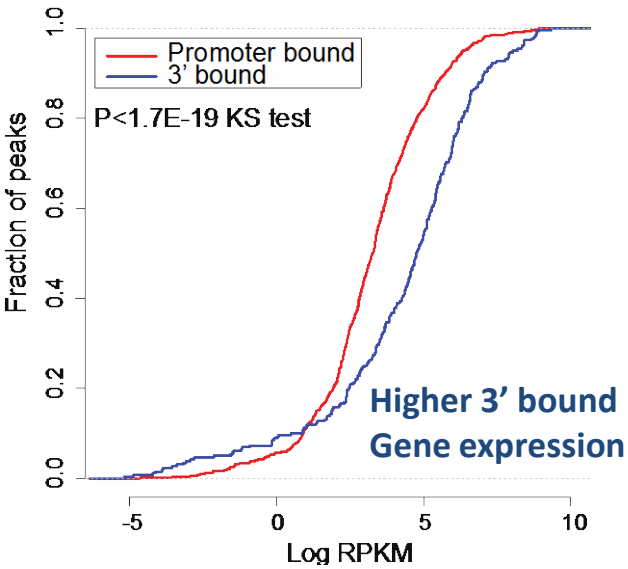
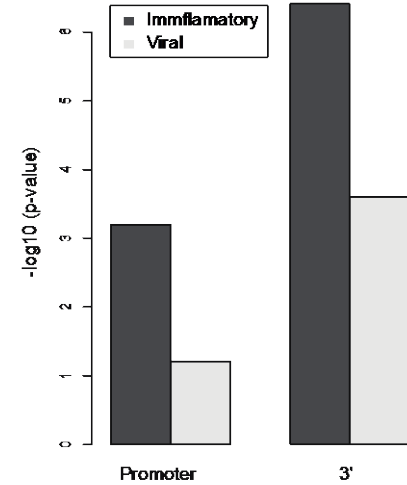
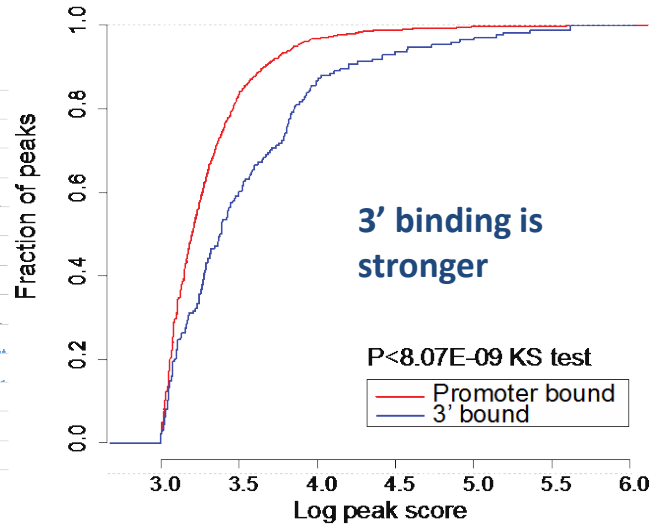
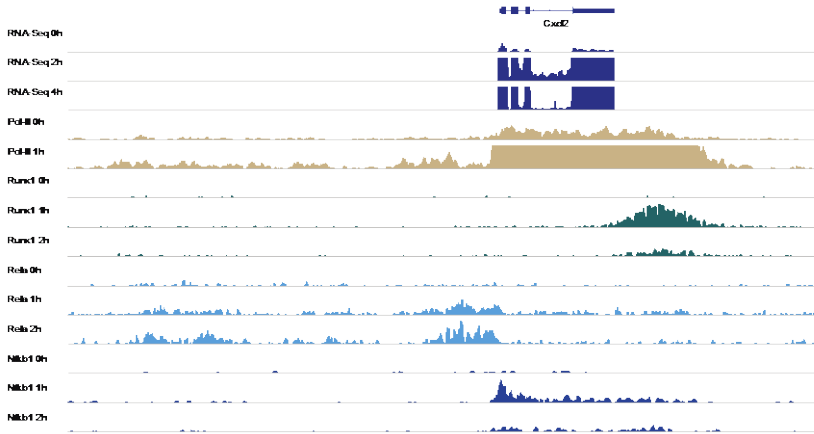
*Automation, High Throughput protocols*

**Oren Ram, Alon Goren**

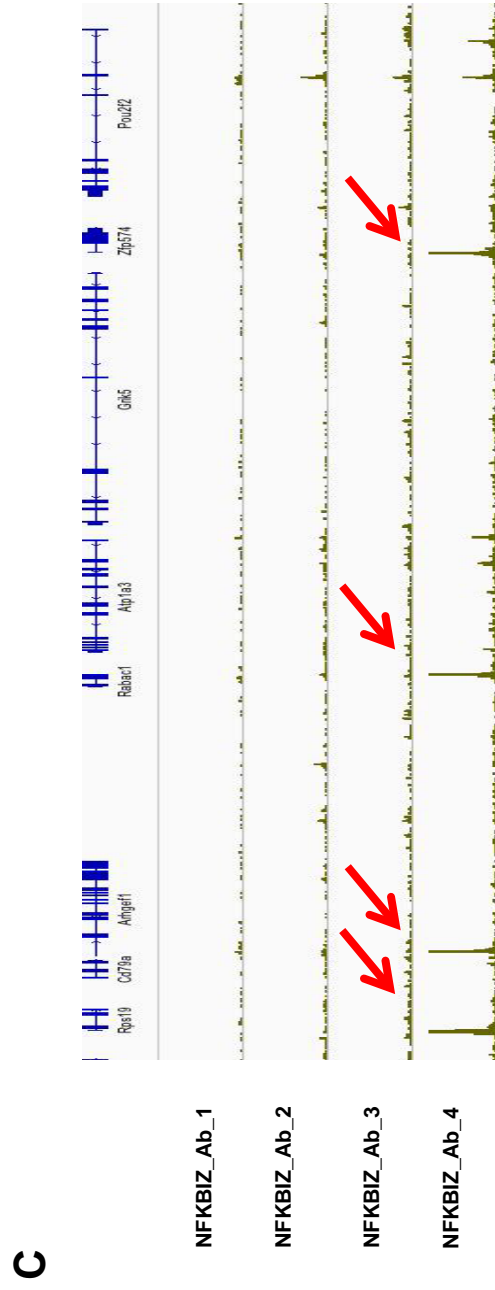
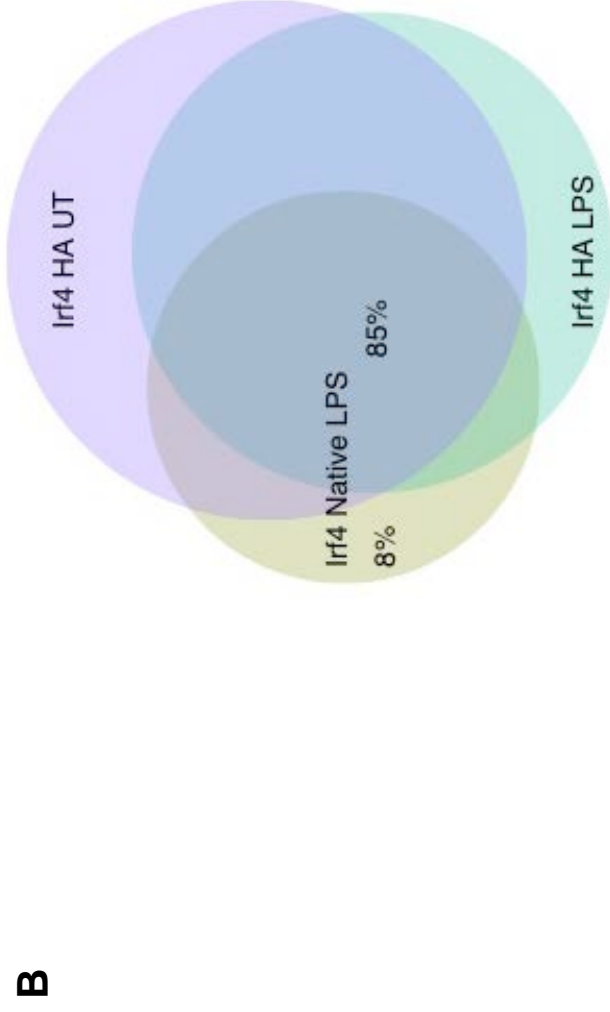
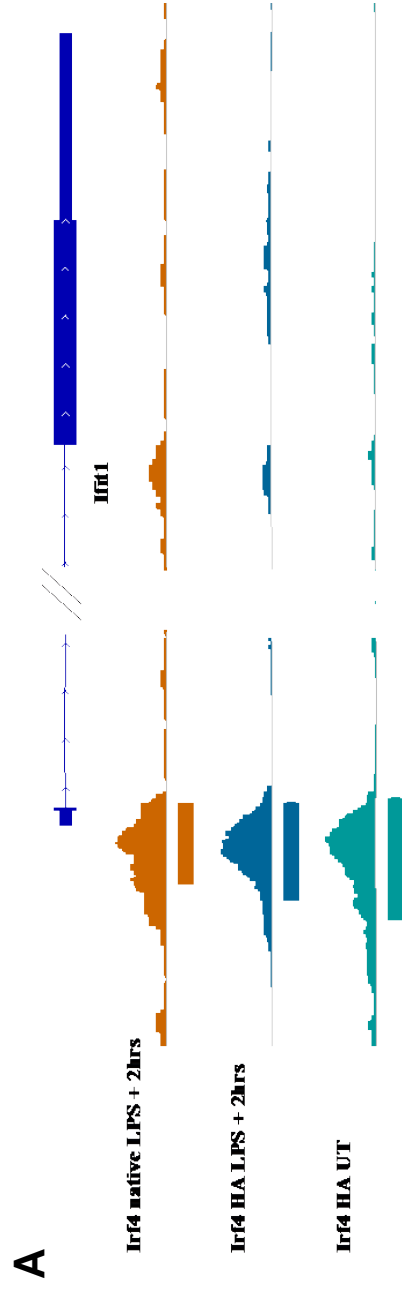
*ChIP String, Low input, many interactions*

**Jim Bochicchio  
Christine Cheng  
Nir Hacohen  
Brad Bernstein  
Aviv Regev**

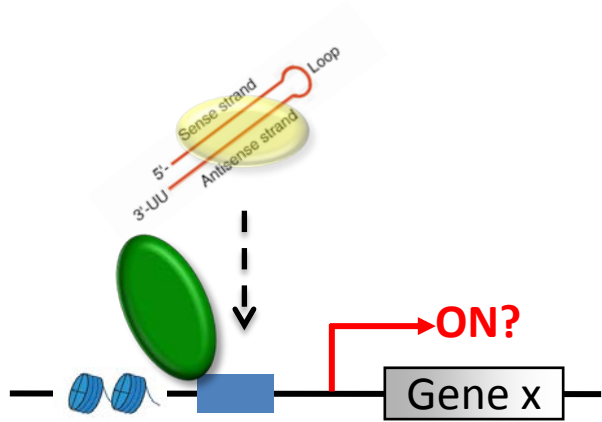
# A non-canonical binder: Runx1



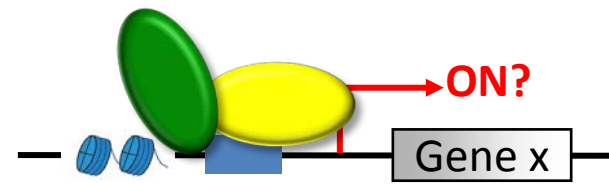
**Mechanism unknown, different partner? Different complex?**



# 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

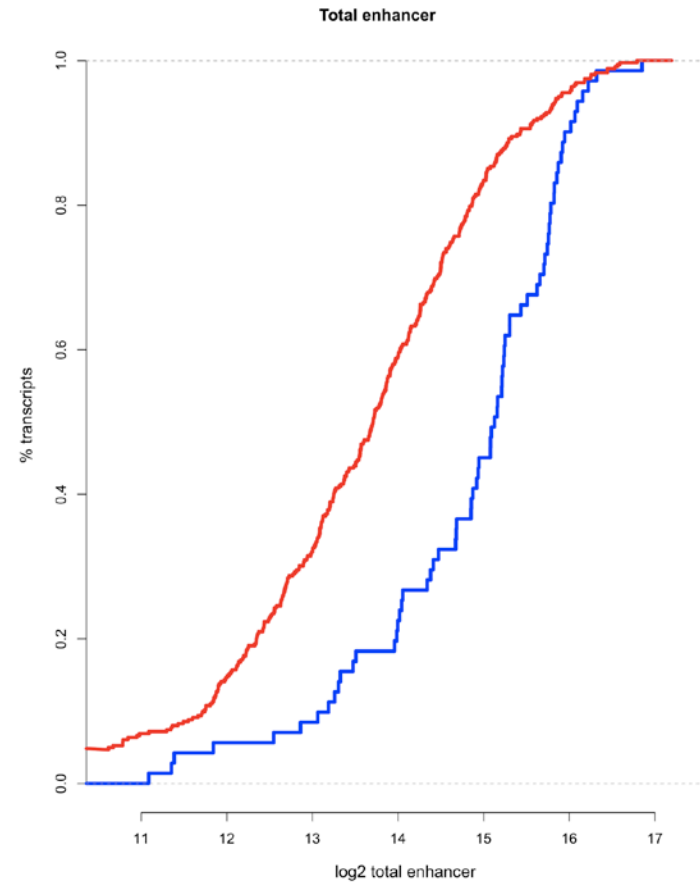
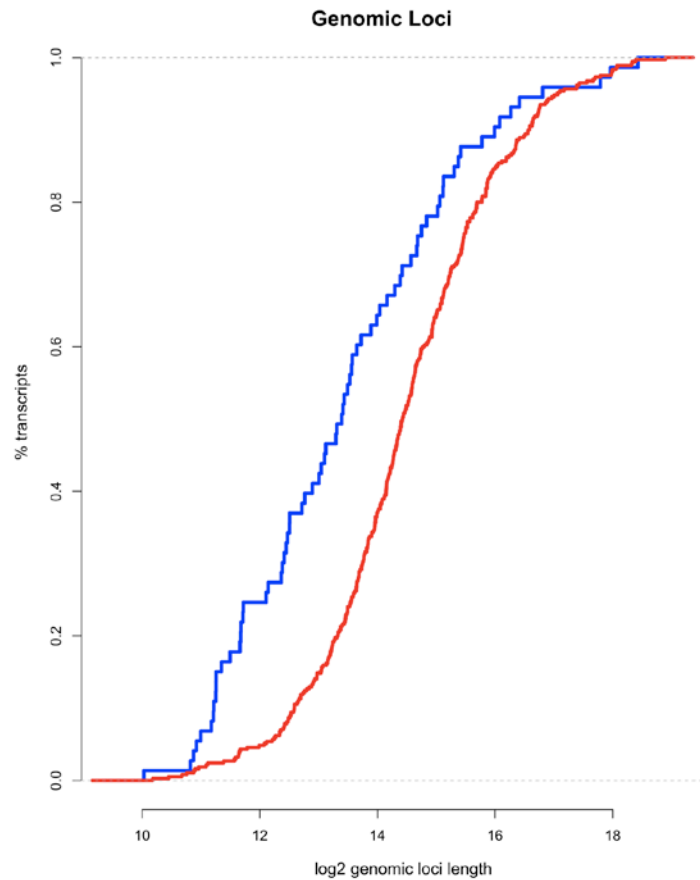




Figure S1 (part-II)

E

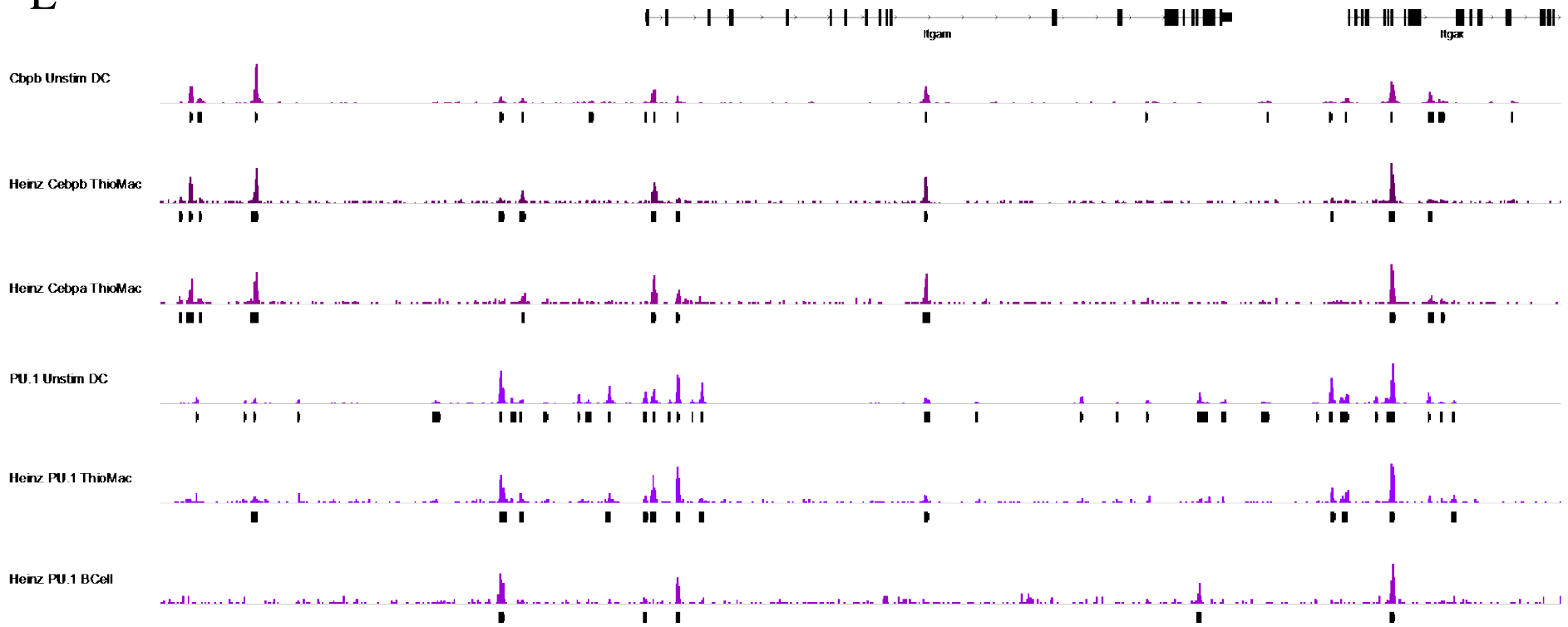


Figure S1 (part-I)

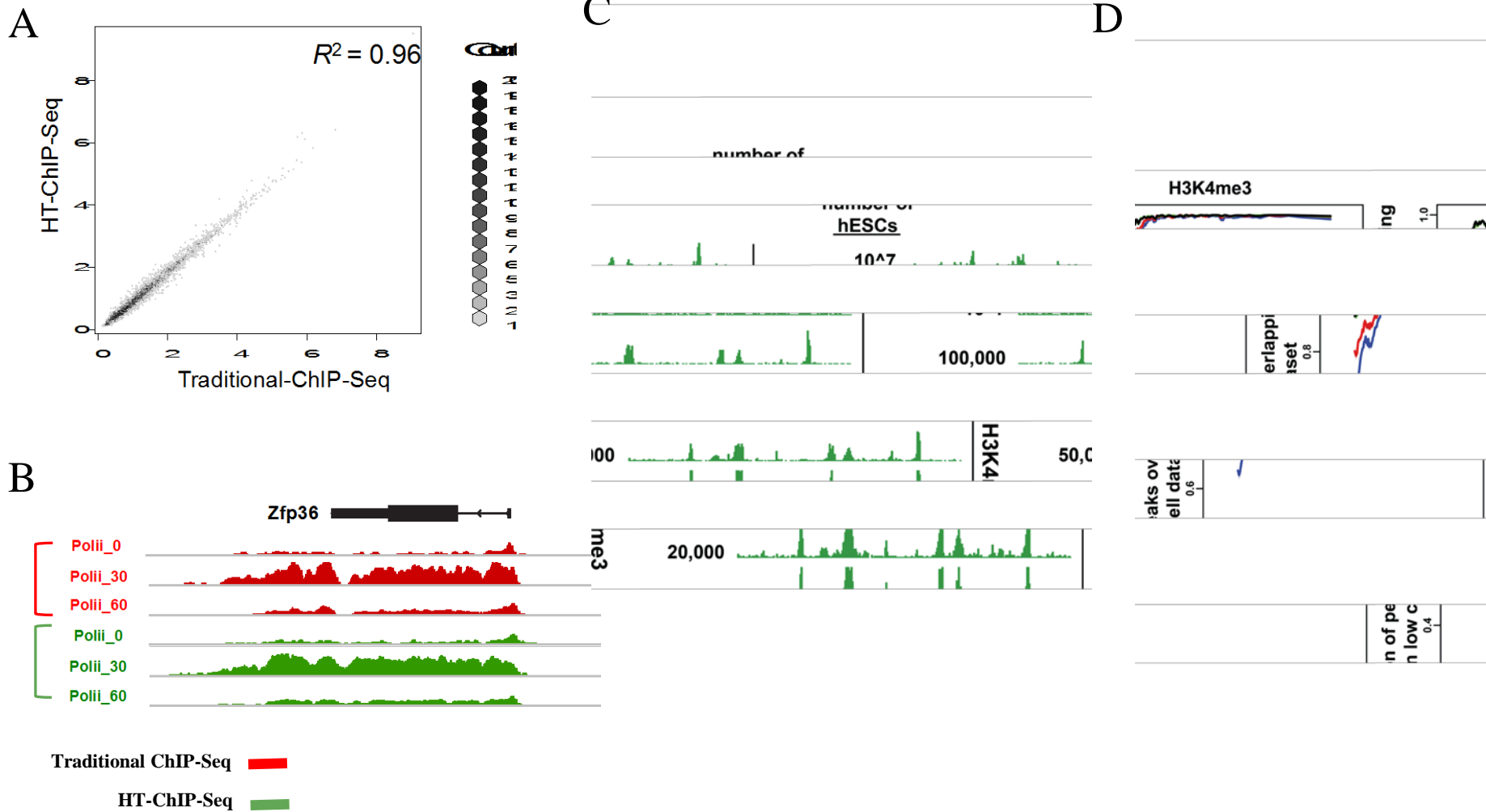
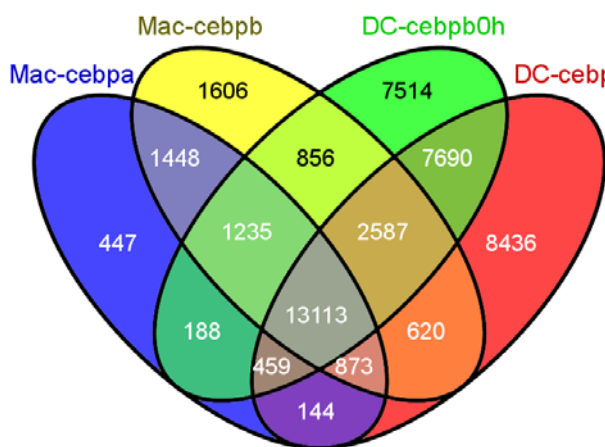


Figure S1 (part-III)

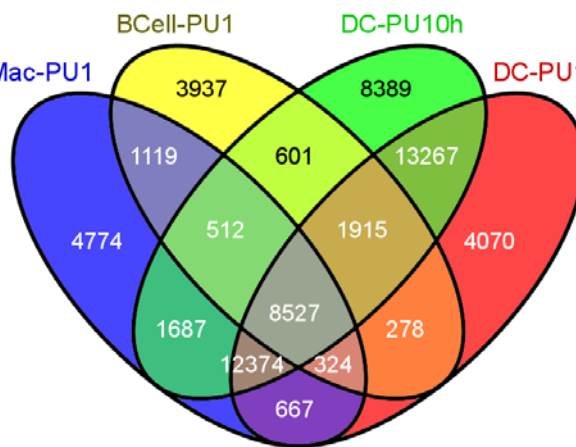
F

|                | DC Cebp<br>0hr | DC PU1<br>0hr | Mac<br>Cebp | Mac<br>PU1 | Bcell<br>PU1 | Mac Cebp<br>a | DC PU1<br>2hr | DC Cebp<br>2hr |
|----------------|----------------|---------------|-------------|------------|--------------|---------------|---------------|----------------|
| DC Cebp<br>0hr | 34467          | 14574         | 18060       | 11883      | 4601         | 15196         | 13360         | 24328          |
| DC PU1<br>0hr  | 14448          | 47806         | 9846        | 23321      | 11644        | 8340          | 36487         | 14132          |
| Mac Cebp       | 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 Cebp<br>a  | 15203          | 8361          | 16917       | 9139       | 2675         | 18201         | 7747          | 14781          |
| DC PU1<br>2hr  | 13242          | 36532         | 9149        | 22116      | 11143        | 7731          | 41982         | 13874          |
| DC Cebp<br>2hr | 24059          | 14198         | 17335       | 11993      | 4892         | 14680         | 13955         | 34643          |

G



H



I

