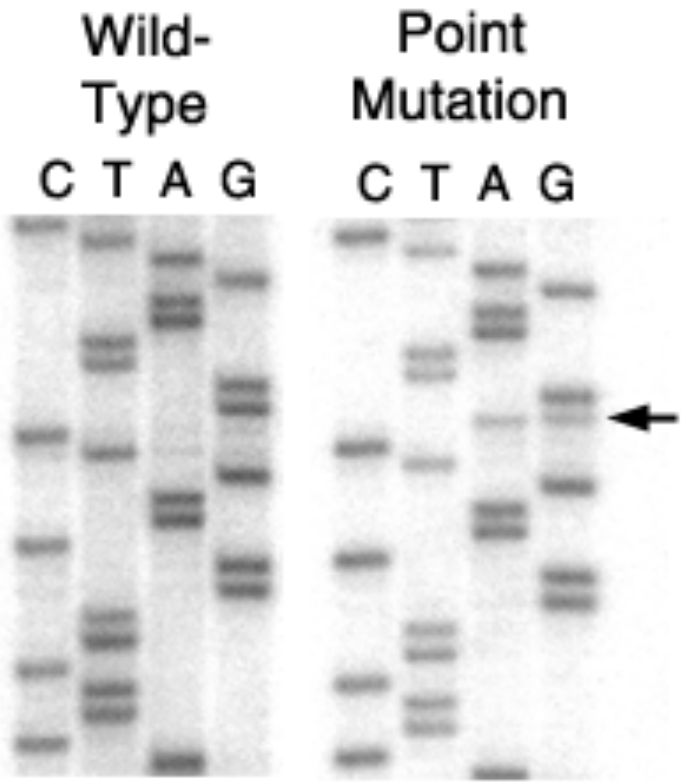
High throughput sequencing methods in systems biology

Bioinformatics 524/525

Module 3, Lecture 2

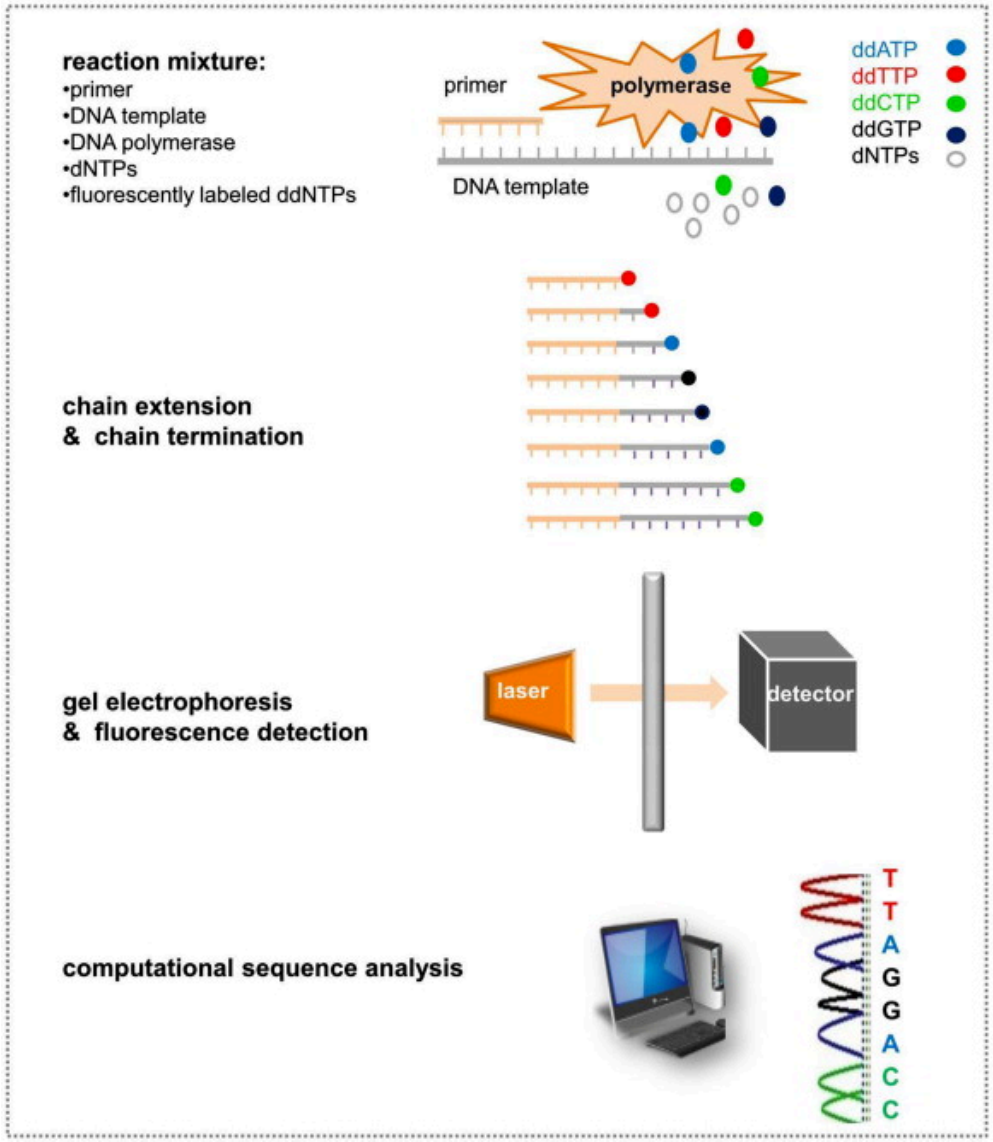
3/28/2017

In the beginning, there were sequencing gels…



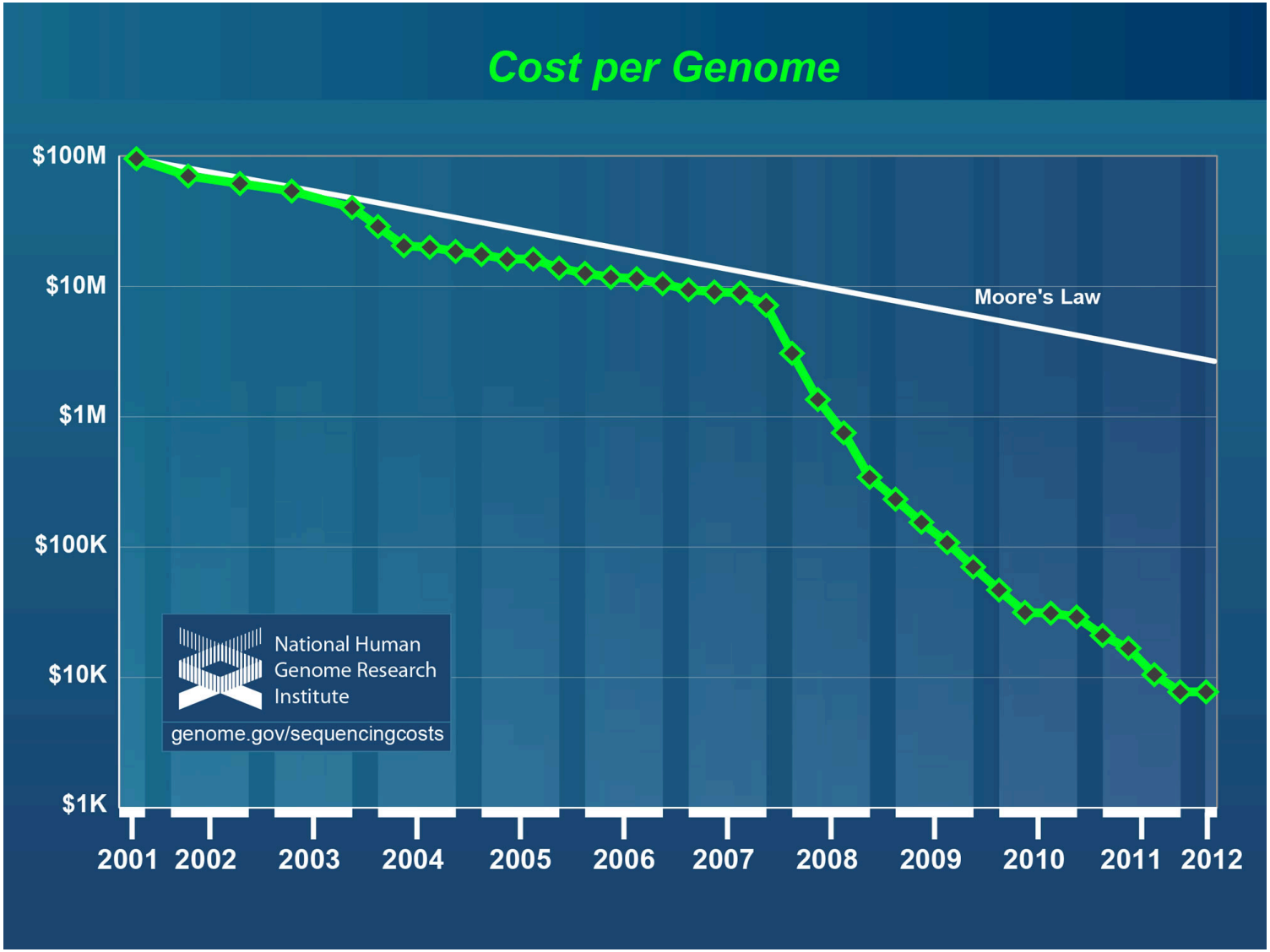
(via Victoria Schulman)

Then there was Sanger sequencing…



From P. Zhang, A. Seth, and H. Fernandes,
Pathobiology of Human Disease

… and then there was ***Next Gen***



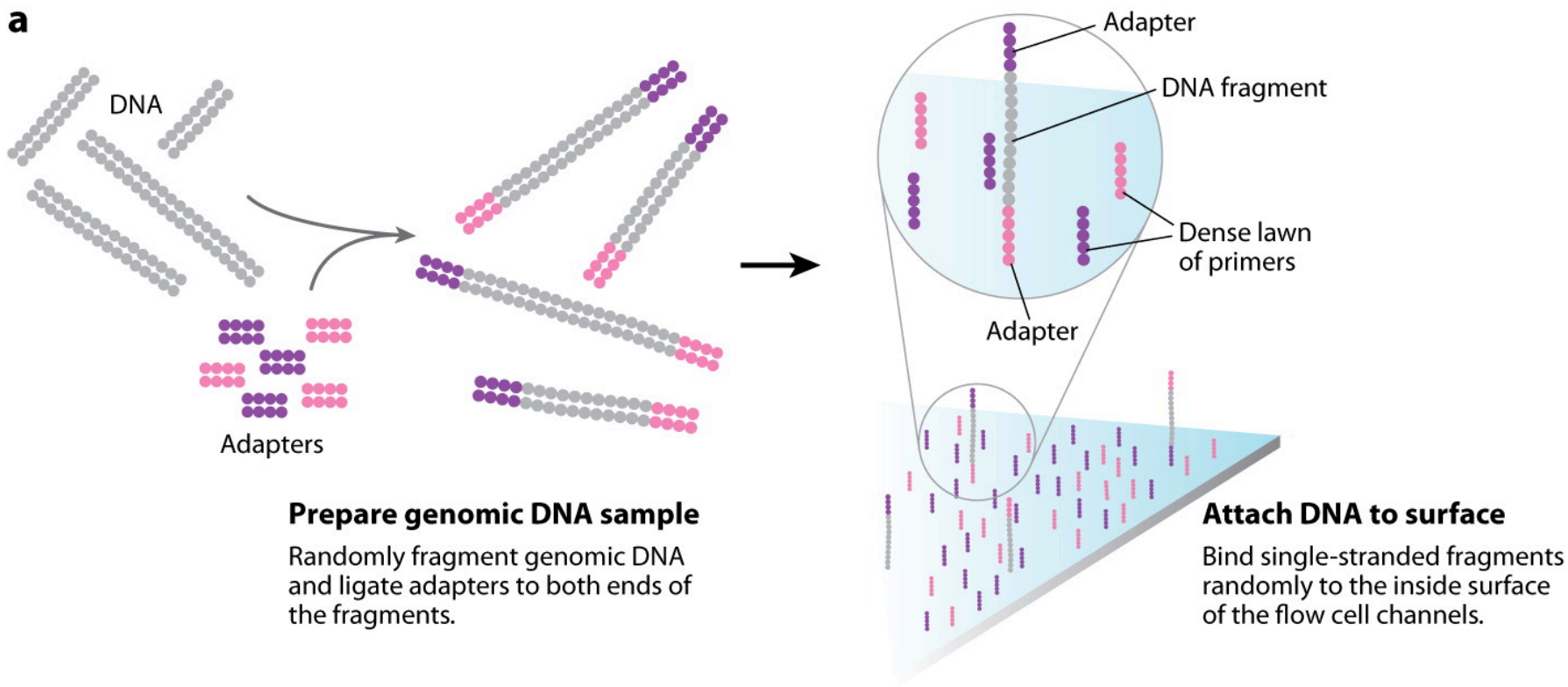
Outline

- Summary of NGS technologies (sequencing and applications)
- Introduction to NGS data analysis
- Commonly available databases
- Workflow integration and making use of existing NGS data

Outline

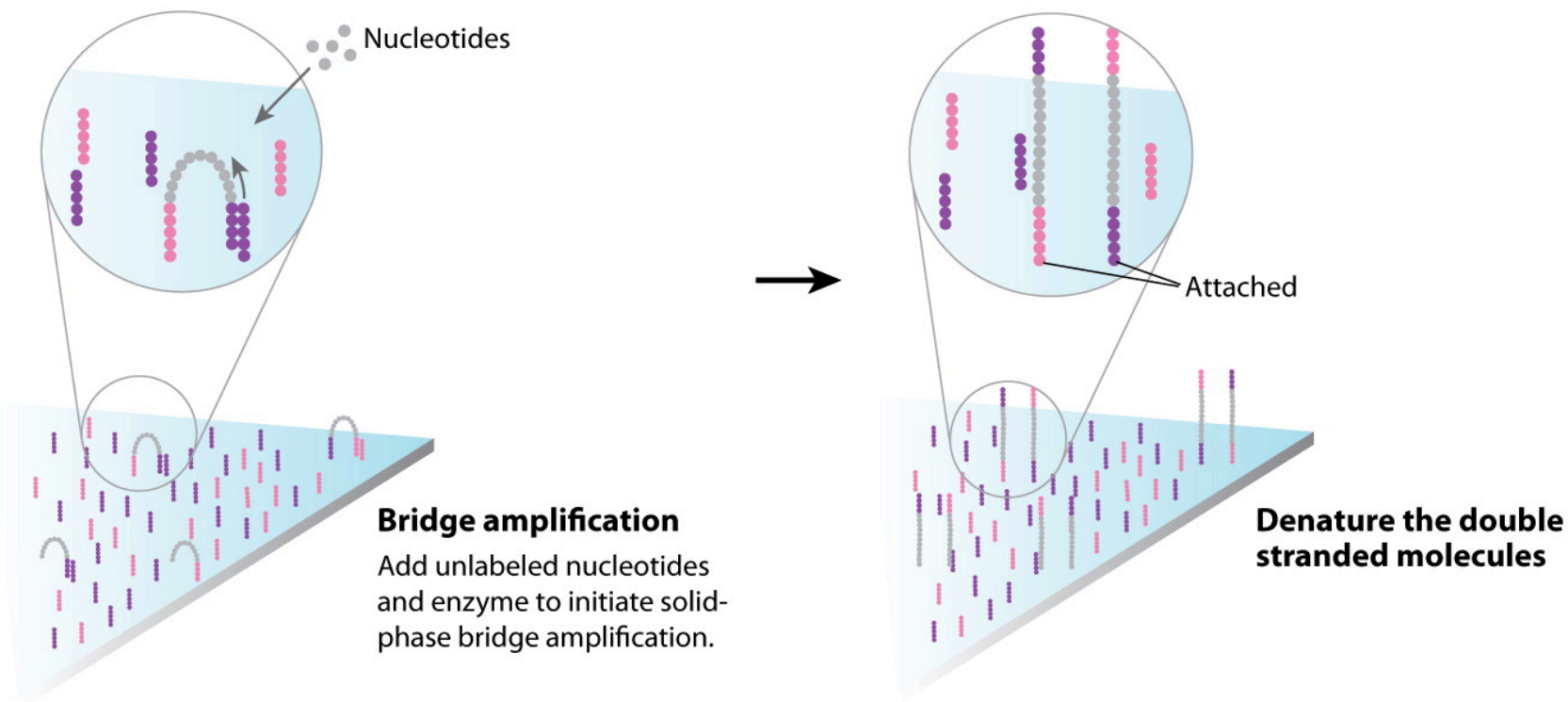
- **Summary of NGS technologies (sequencing and applications)**
- Introduction to NGS data analysis
- Commonly available databases
- Workflow integration and making use of existing NGS data

Illumina sequencing-by-synthesis



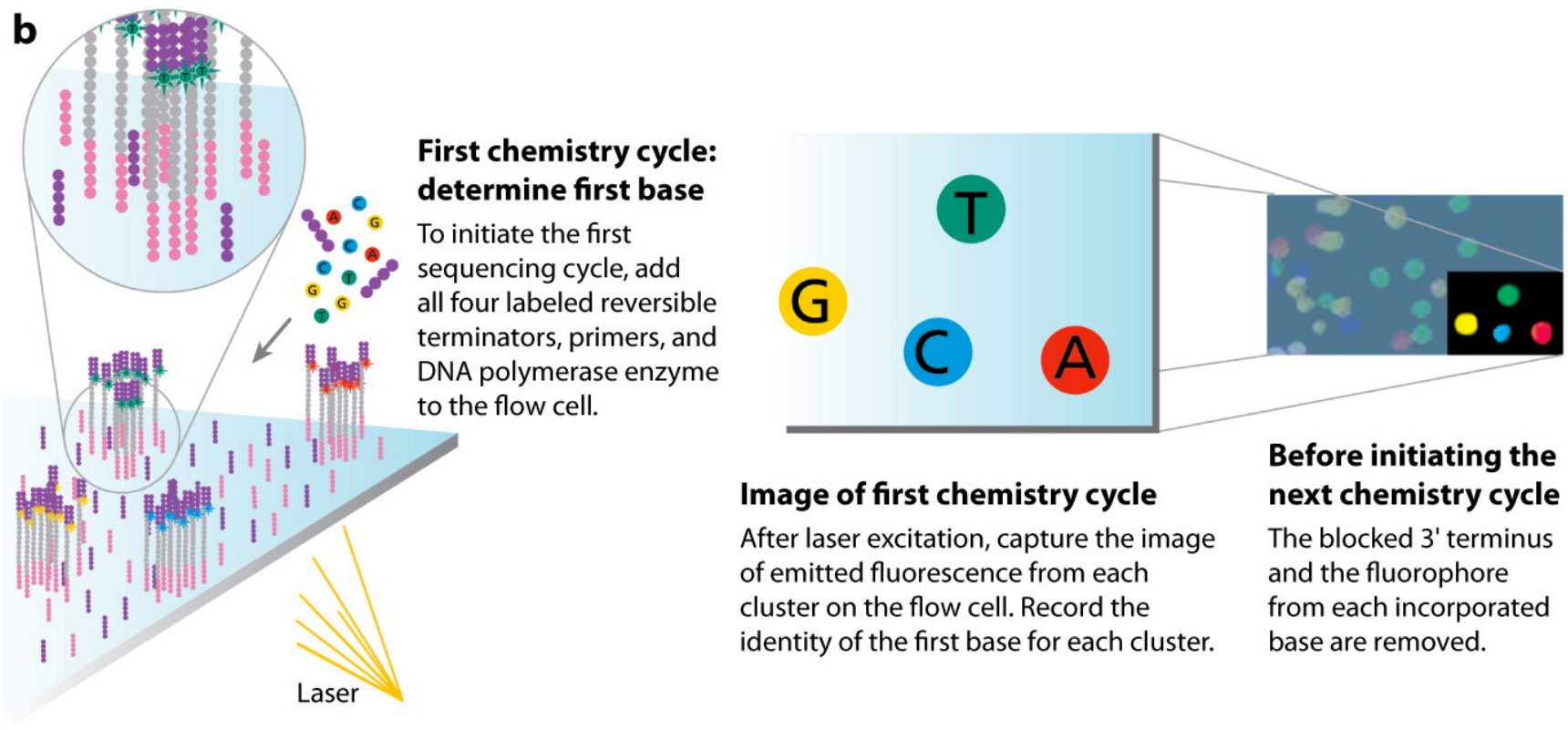
(Mardis, Ann. Rev. Genomics Hum. Genet., 2008)

Illumina sequencing-by-synthesis



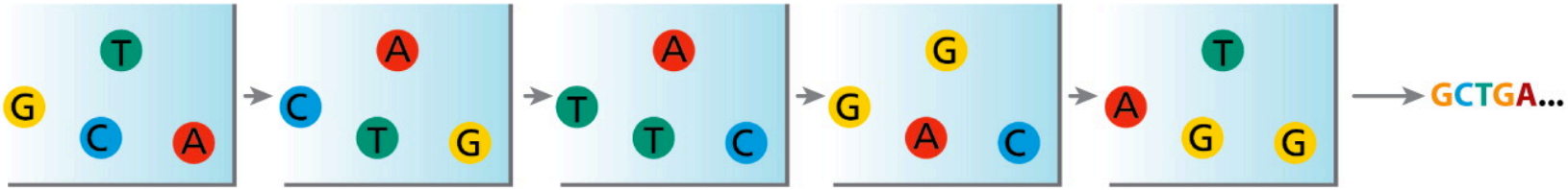
(Mardis, Ann. Rev. Genomics Hum. Genet., 2008)

Illumina sequencing-by-synthesis



(Mardis, Ann. Rev. Genomics Hum. Genet., 2008)

Illumina sequencing-by-synthesis

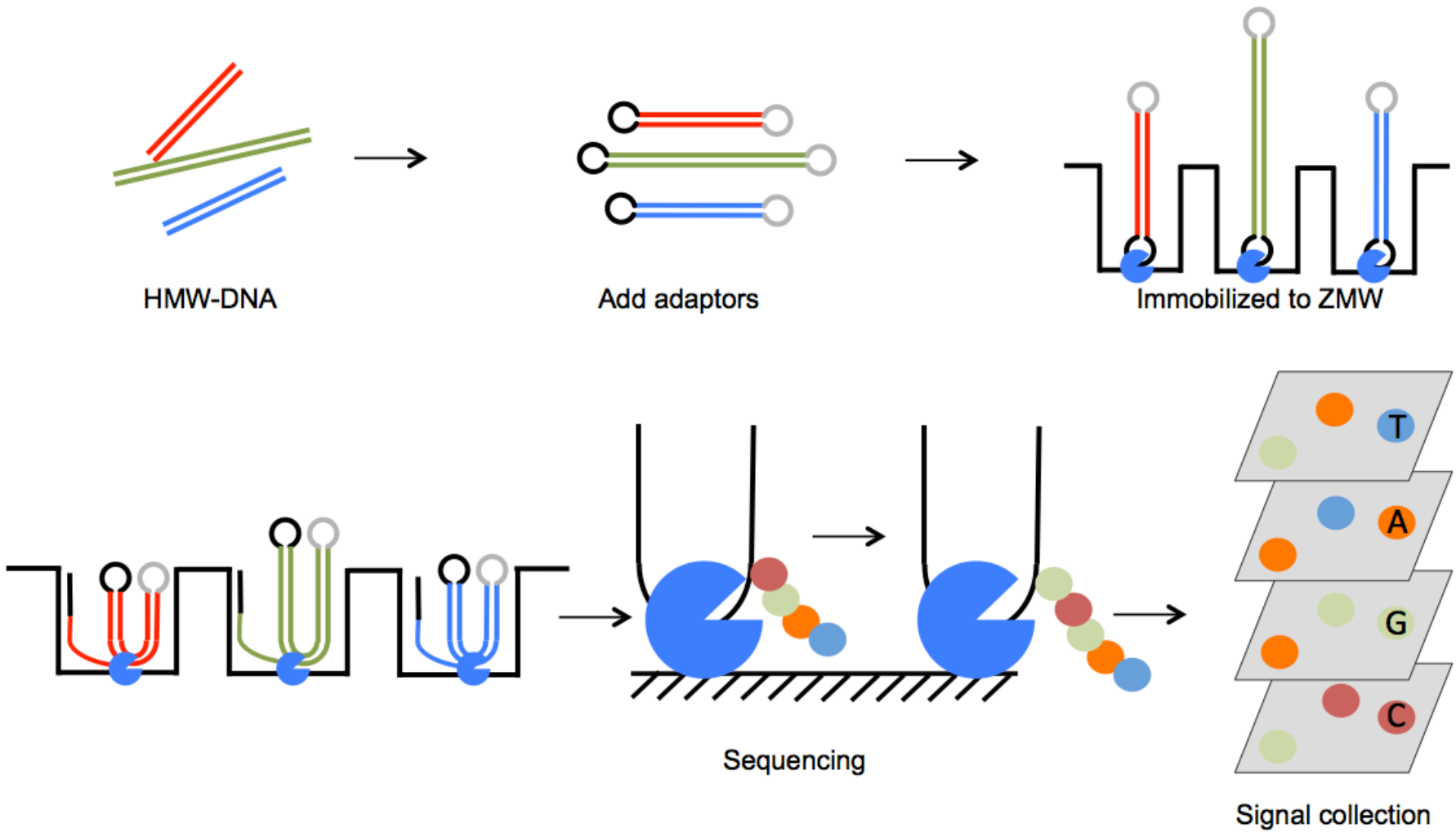


Sequence read over multiple chemistry cycles

Repeat cycles of sequencing to determine the sequence of bases in a given fragment a single base at a time.

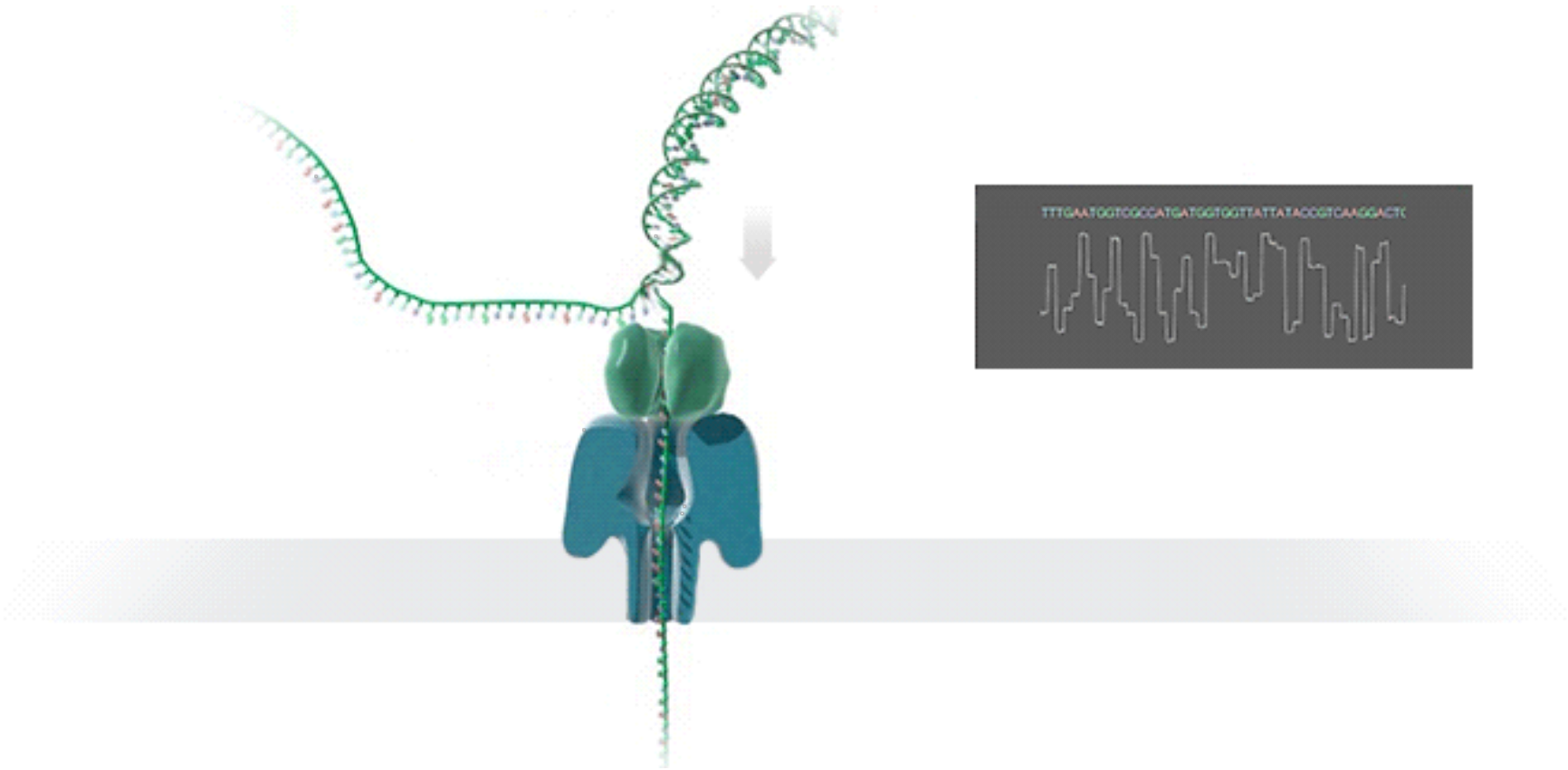
(Mardis, Ann. Rev. Genomics Hum. Genet., 2008)

PacBio SMRT Sequencing



(Image from 3402 Bioinformatics)

Nanopore sequencing



(Image via Oxford Nanopore)

Key considerations for NGS technologies

- Number of reads
- Quality of reads
- Length of reads
- Library preparation
- Cost

Key considerations for NGS technologies

	Illumina	PacBio	Nanopore
Number of reads	***	**	*
Quality of reads	**	*/***	***
Length of reads	*	**	***
Library preparation	Versatile, complex	Moderate	Simple

Library preparation for the Illumina platform

(From “Illumina TruSeq DNA Adapters De-Mystified” by James Schiemer)

Library preparation for the Illumina platform

TruSeq Universal Adapter:
5 AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT 3

TruSeq Indexed Adapter
5 GATCGGAAGAGCACACGTCTGAACTCCAGTCAC-NNNNNN-ATCTCGTATGCCGTCTTCTGCTTG 3

(From "Illumina TruSeq DNA Adapters De-Mystified" by James Schiemer)

Library preparation for the Illumina platform

TruSeq Universal Adapter:

5 AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT 3

TruSeq Indexed Adapter

5 GATCGGAAGAGCACACGTCTGAACTCCAGTCAC-NNNNNN-ATCTCGTATGCCGTCTTCTGCTTG 3

Anneal:

```
5         AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGAC--GCTCTTCCGATC*T 3
3 GTTCGTCTTCTGCCGTATGCTCTA(INDEX)CACTGACCTCAAGTCTGCACA--CGAGAAGGCTAG*P 5
```

(From “Illumina TruSeq DNA Adapters De-Mystified” by James Schiemer)

Library preparation for the Illumina platform

After ligation:

```
LEFT OF INSERT
5        AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGAC--GCTCTTCCGATC*T 3
3 GTTCGTCTTCTGCCGTATGCTCTA(INDEX)CACTGACCTCAAGTCTGCACA--CGAGAAGGCTAG*P 5

RIGHT OF INSERT
5 P*GATCGGAAGAGC--ACACGTCTGAACTCCAGTCAC(INDEX)ATCTCGTATGCCGTCTTCTGCTTG 3
3 T*CTAGCCTTCTCG--CAGCACATCCCTTTCTCACATCTAGAGCCACCAGCGGCATAGTAA        5
```

(From “Illumina TruSeq DNA Adapters De-Mystified” by James Schiemer)

Library preparation for the Illumina platform

PCR Primer 1.0

5 AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGA 3

PCR Primer 2.0

5 CAAGCAGAAGACGGCATACGAGAT 3

Universal Adapter:

5 AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT 3

Indexing Adapter:

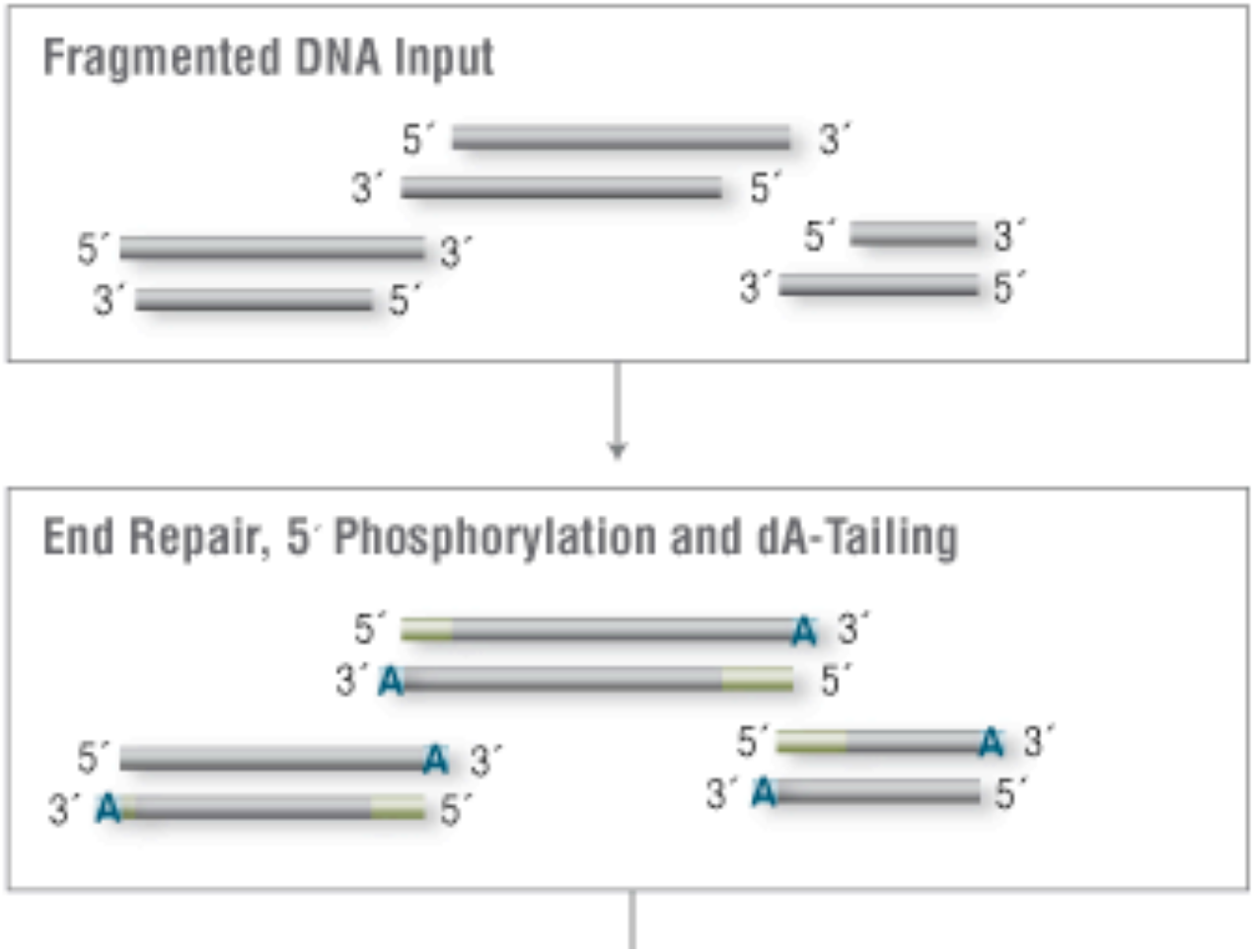
5 GATCGGAAGAGCACACGTCTGAACTCCAGTCAC-NNNNNN-ATCTCGTATGCCGTCTTCTGCTTG 3

(From "Illumina TruSeq DNA Adapters De-Mystified" by James Schiemer)

Library preparation for the Illumina platform

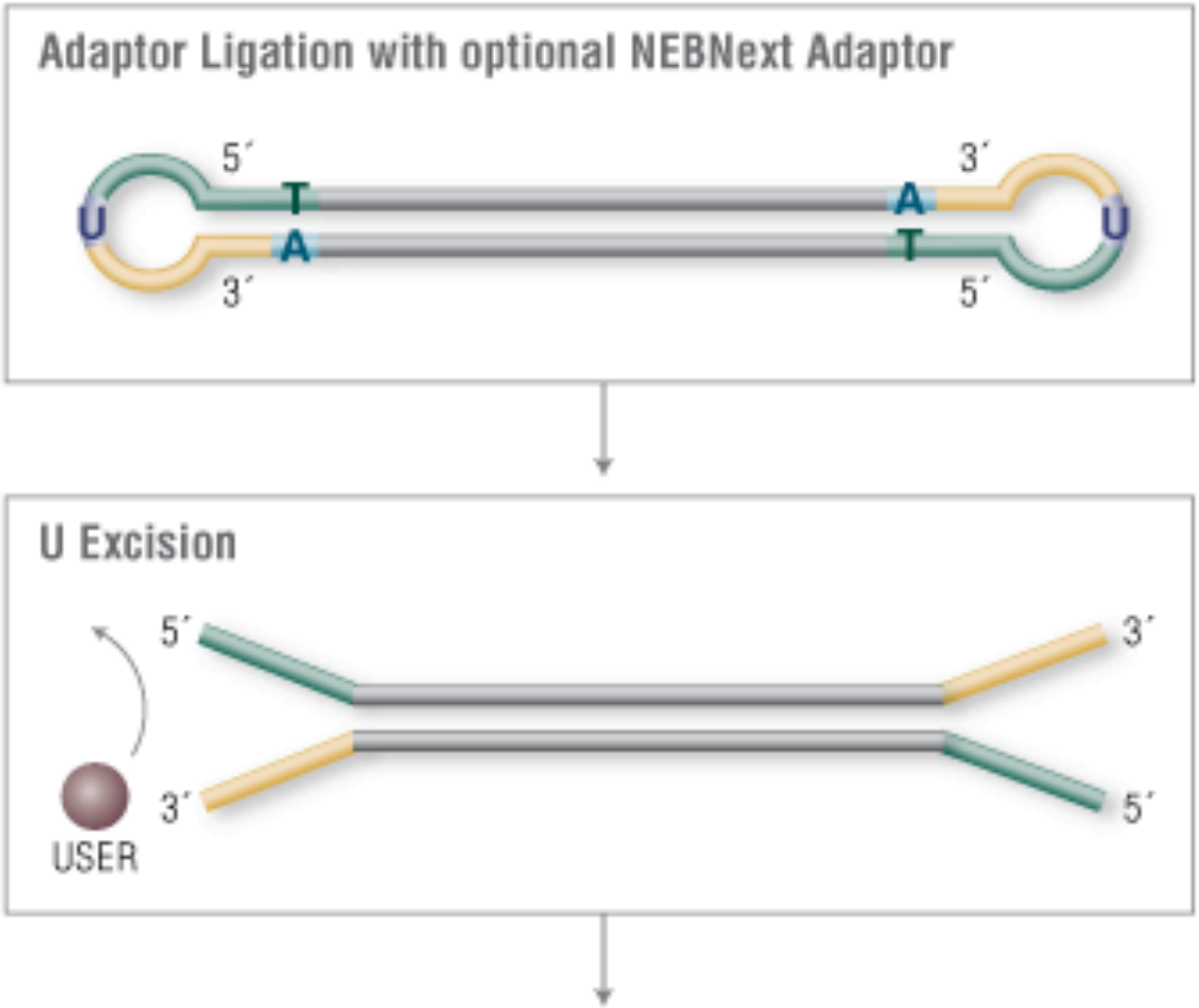
(From “Illumina TruSeq DNA Adapters De-Mystified” by James Schiemer)

Example of a full sequencing prep workflow



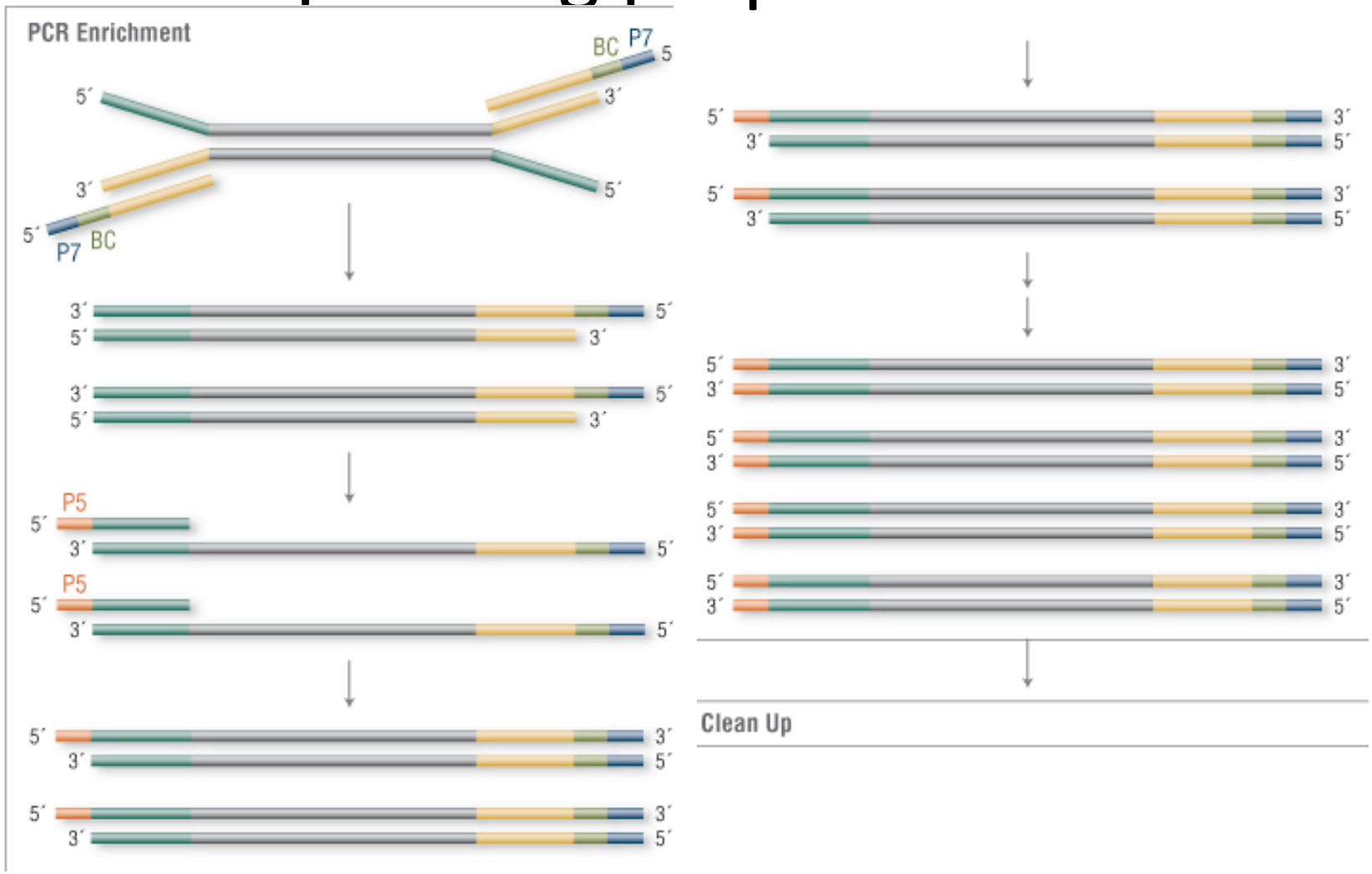
(Image from NEB)

Example of a full sequencing prep workflow



(Image from NEB)

Example of a full sequencing prep workflow

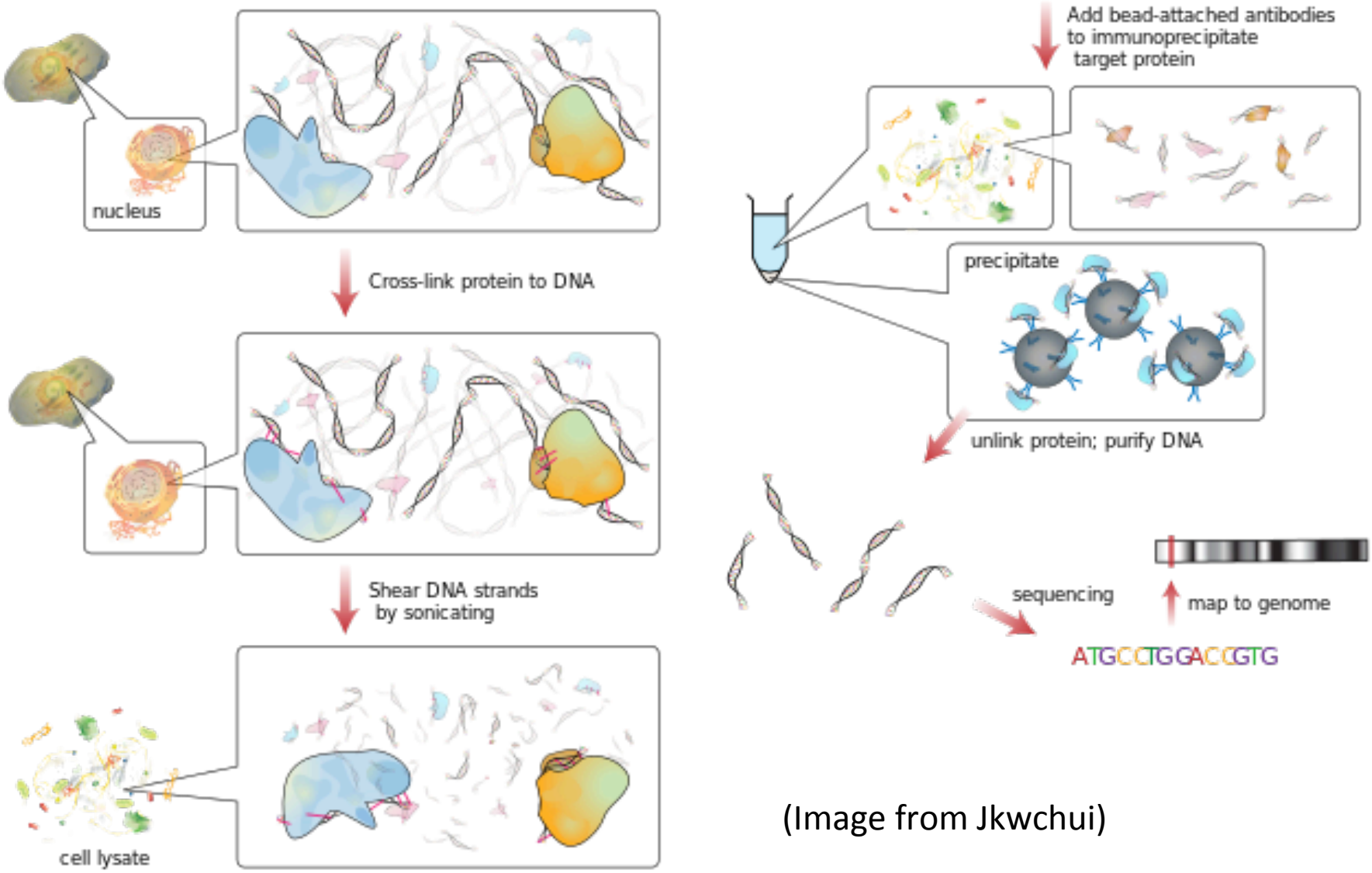


(Image from NEB)

Applications

- Genome sequencing (whole genome, mutations)
- RNA sequencing (transcript quantitation, transcriptome mapping)
- Finding protein-nucleic acid interaction (ChIP-seq, PAR-CLIP)
- Identifying methylation sites (bisulfite sequencing)

Applications



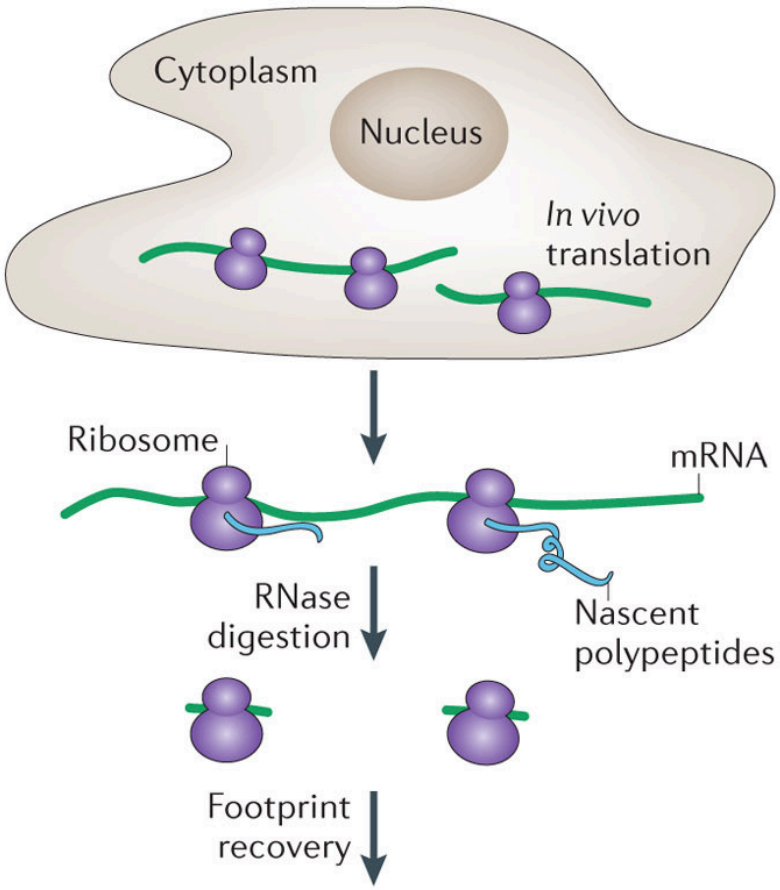
(Image from Jkwchui)

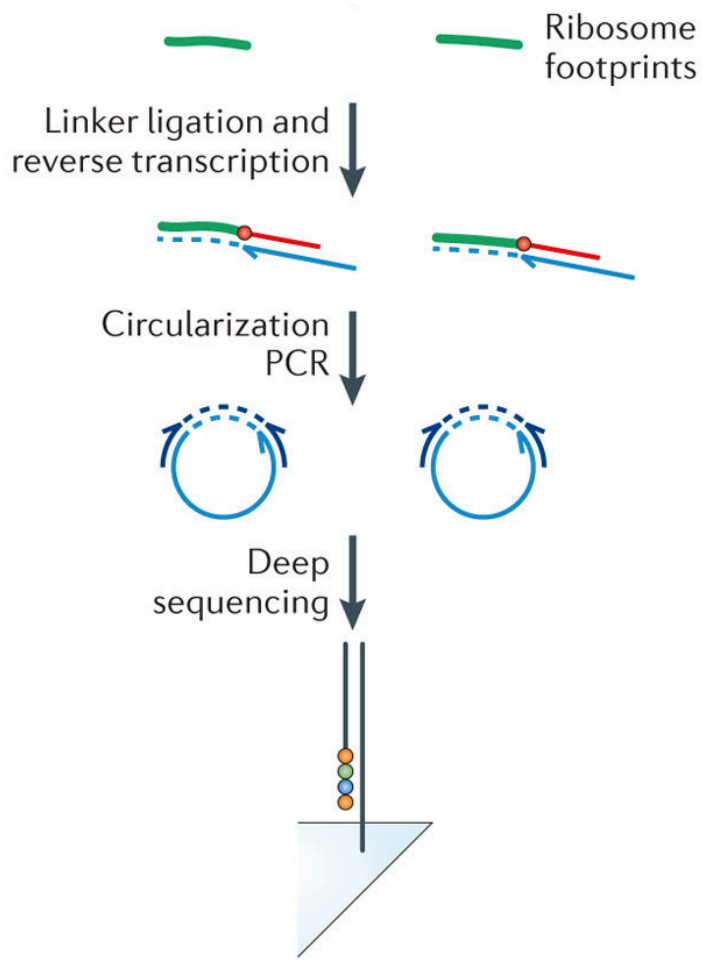
Applications

- Protein translation rates (ribosome profiling)
- Chromosomal conformations (Hi-C)
- Transcript stability (Bru-chase seq)
- Finding DNA-RNA hybrids (Drip-seq)
- Profiling chromatin accessibility (ATAC-seq, Mnase-seq)

And on and on…

Ribosome profiling



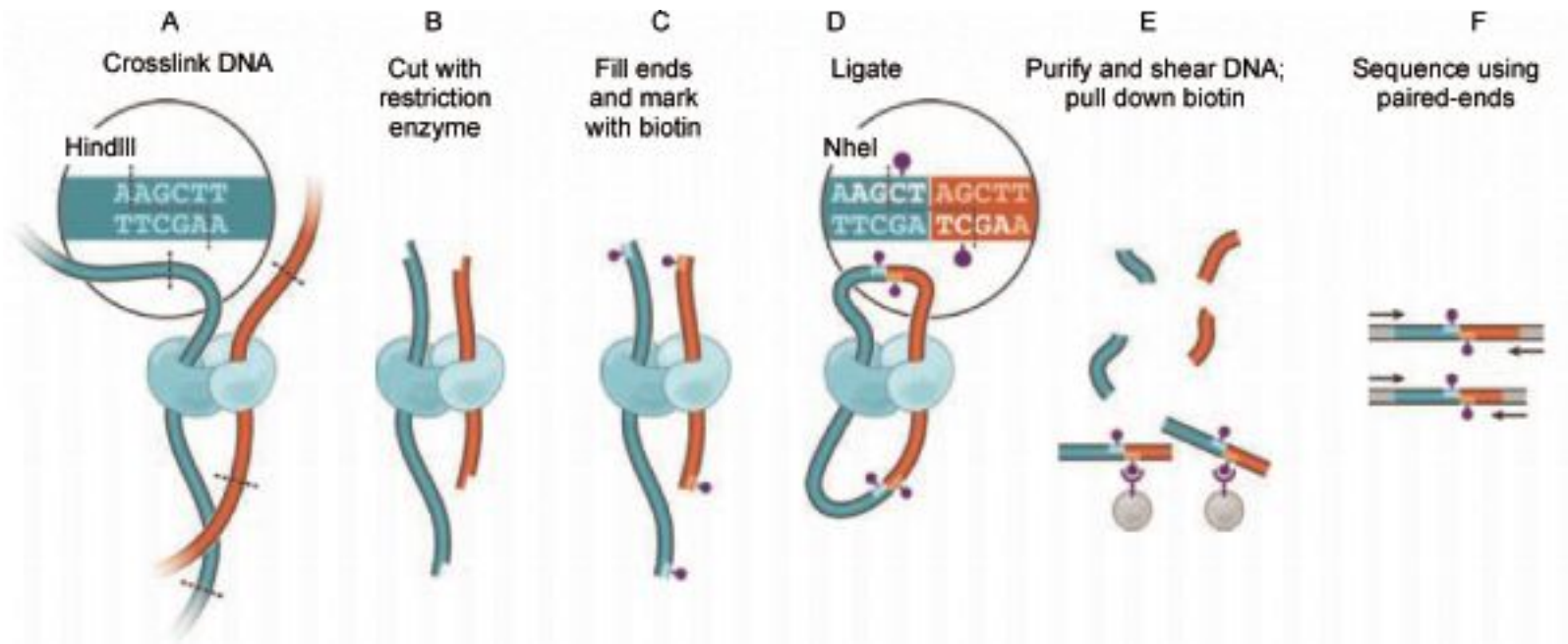


Applications

- Protein translation rates (ribosome profiling)
- Chromosomal conformations (Hi-C)
- Transcript stability (Bru-chase seq)
- Finding DNA-RNA hybrids (Drip-seq)
- Profiling chromatin accessibility (ATAC-seq, Mnase-seq)

And on and on…

Applications



(Lieberman-Aiden et al., Science, 2009)

Applications

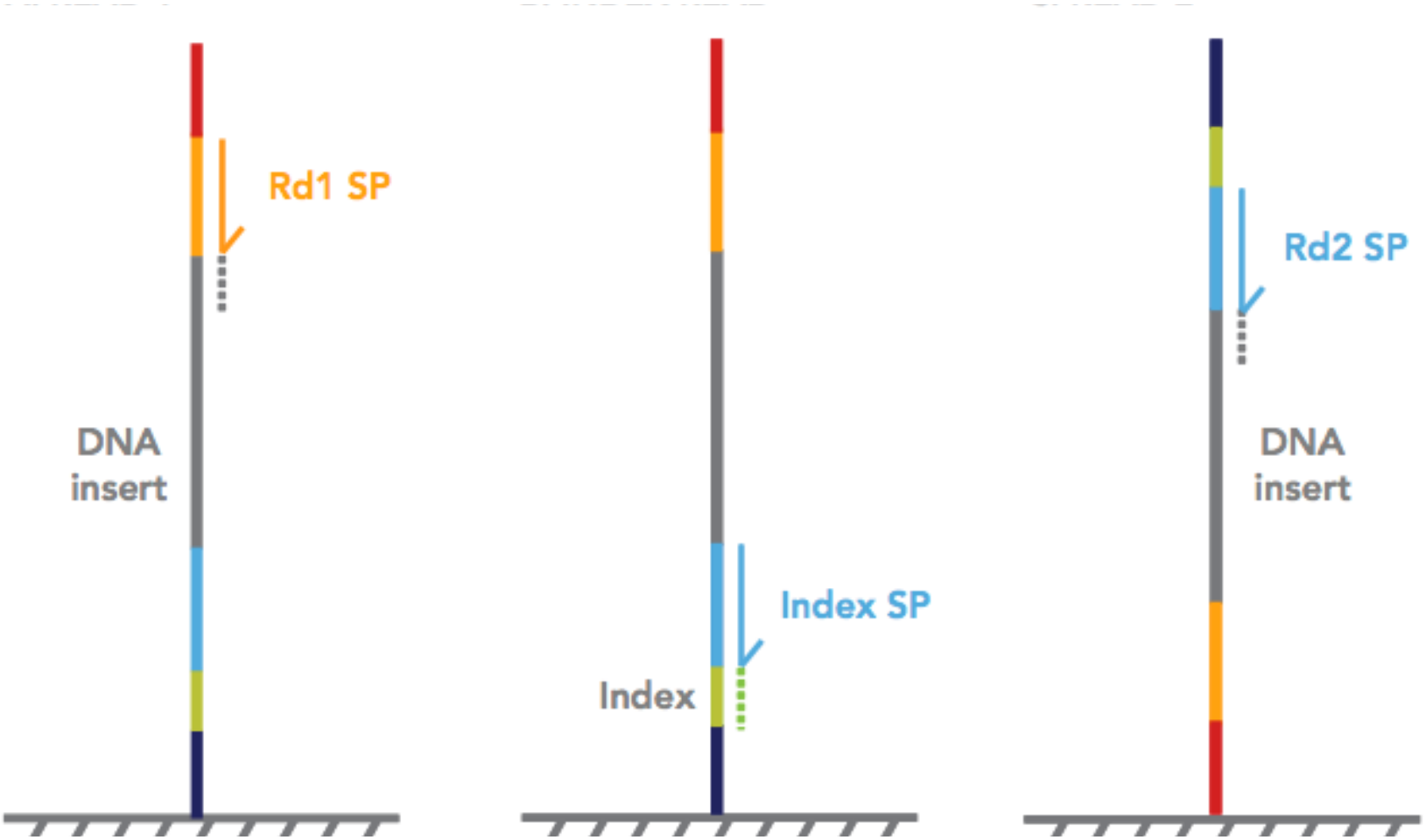
- Protein translation rates (ribosome profiling)
- Chromosomal conformations (Hi-C)
- Transcript stability (Bru-chase seq)
- Finding DNA-RNA hybrids (Drip-seq)
- Profiling chromatin accessibility (ATAC-seq, Mnase-seq)

And on and on…

A few crucial concepts

- Single end vs. paired end reads
- “Coverage”
- Indexing
- FPKM, RPKM, TPM

A few crucial concepts



(Image from Illumina)

A few crucial concepts

- Single end vs. paired end reads
- “Coverage”
- Indexing
- FPKM, RPKM, TPM

Outline

- Summary of NGS technologies (sequencing and applications)
- **Introduction to NGS data analysis**
- Commonly available databases
- Workflow integration and making use of existing NGS data

Raw data: Fastq files

```
@K00135:141:HHJ3TBBXX:1:2228:1661:47383
CTCCTGTTCTTGTGGTTGCTGGGGCTCCAATAG
+
AAA-AAJFF-AFAAF7FFA-AA-77AJ<7A<-7
@K00135:141:HHJ3TBBXX:1:2228:5467:17685
GTATTTTTAGTTCCATACACGCAAGAAGGAG
+
-A-<AF<<FFF<FJFA--<--77<<JF7-7<
```

Raw data: Fastq files

Read name
Sequence
Optional information
Quality scores

Read name
Sequence
Optional information
Quality scores

```
@K00135:141:HHJ3TBBXX:1:2228:1661:47383
CTCCTGTTCTTGTGGTTGCTGGGGCTCCAATAG
+
AAA-AAJFF-AFAAF7FFA-AA-77AJ<7A<-7
@K00135:141:HHJ3TBBXX:1:2228:5467:17685
GTATTTTTAGTTCCATACACGCAAGAAGGAG
+
-A-<AF<<FFF<FJFA--<--77<<JF7-7<
```

Typical steps in analysis workflow

- Quality control
- Adapter clipping
- Quality trimming
- Alignment
- Analysis

Quality control

(Example: FastQC)

FastQC Report

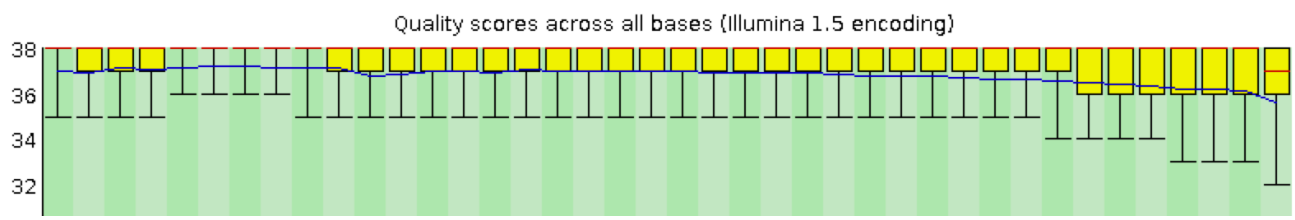
Summary

- Basic Statistics
- Per base sequence quality
- Per tile sequence quality
- Per sequence quality scores
- Per base sequence content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Adapter Content
- Kmer Content

Basic Statistics

Measure	Value
Filename	good_sequence_short.txt
File type	Conventional base calls
Encoding	Illumina 1.5
Total Sequences	250000
Sequences flagged as poor quality	0
Sequence length	40
%GC	45

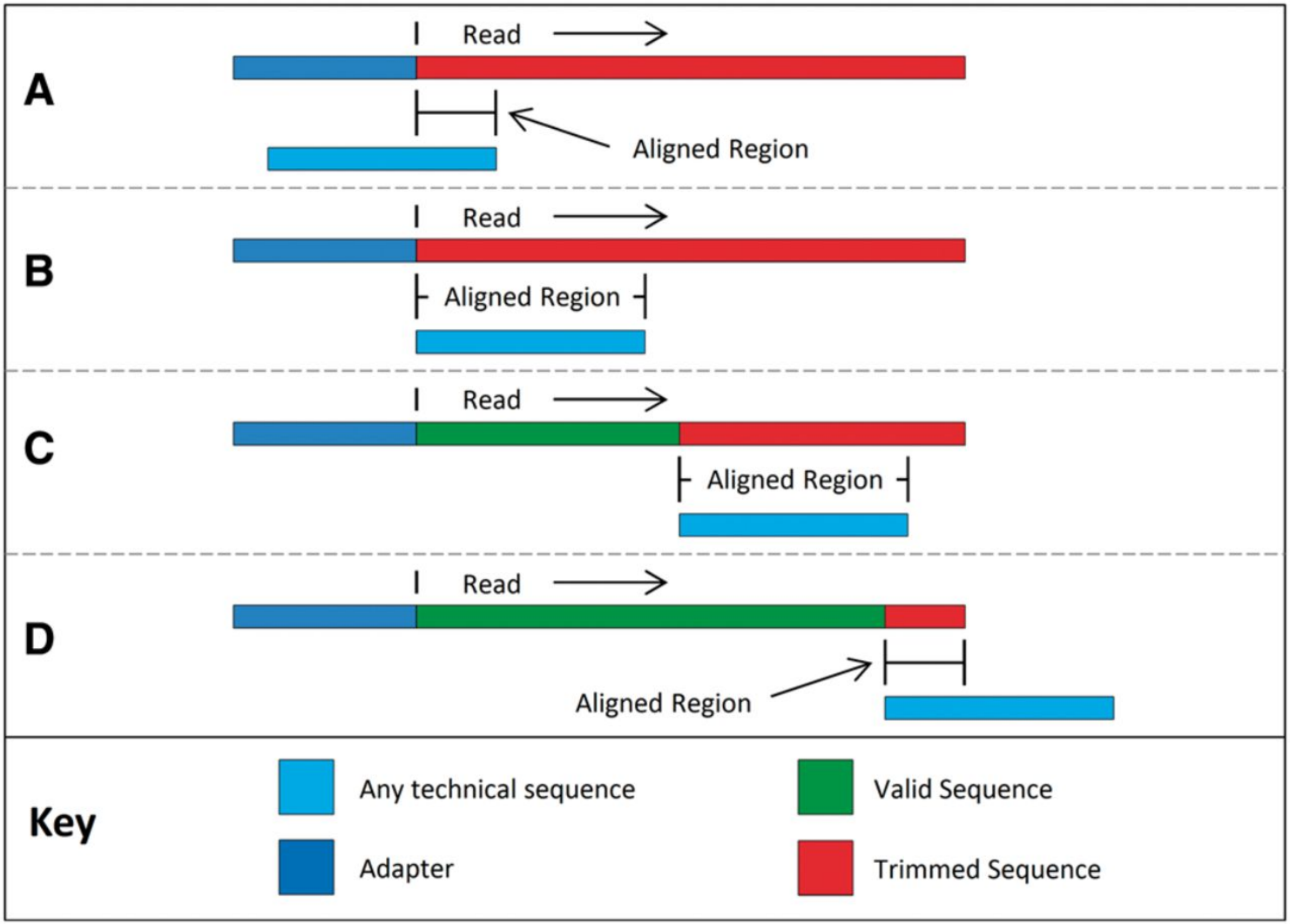
Per base sequence quality



(fastqc webpage)

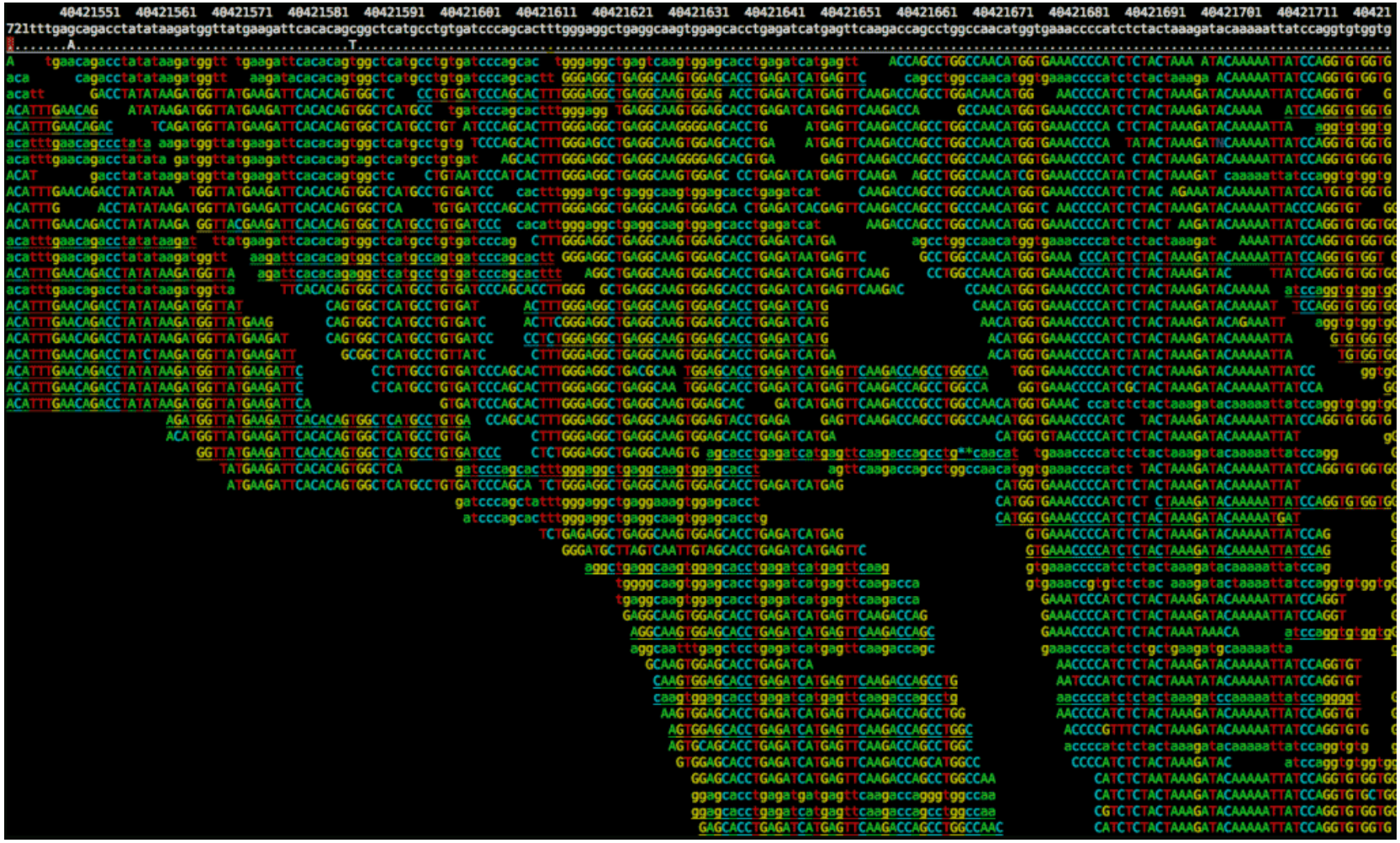
Trimming/clipping

(fastxtools, cutadapt, trimmomatic)



Anthony M. Bolger et al. Bioinformatics 2014;30:2114-2120

Alignment



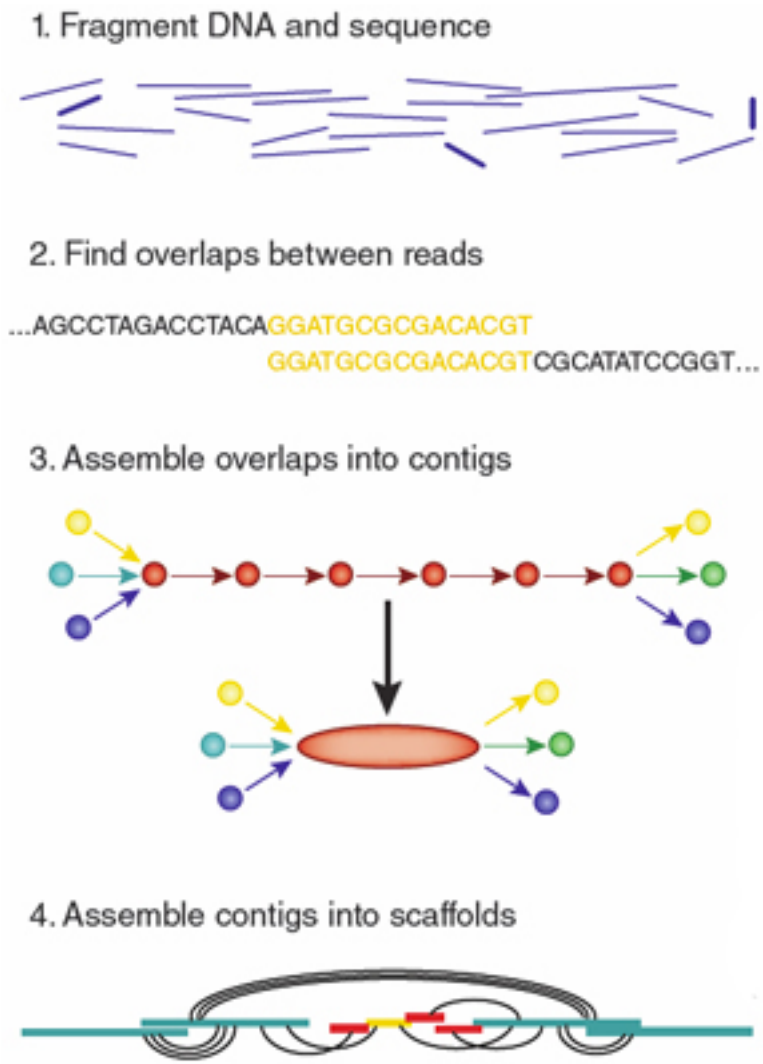
Examples: Bowtie, BWA, SOAP, STAR, HiSat, …

(image from labtimes.org)

(or assembly…)

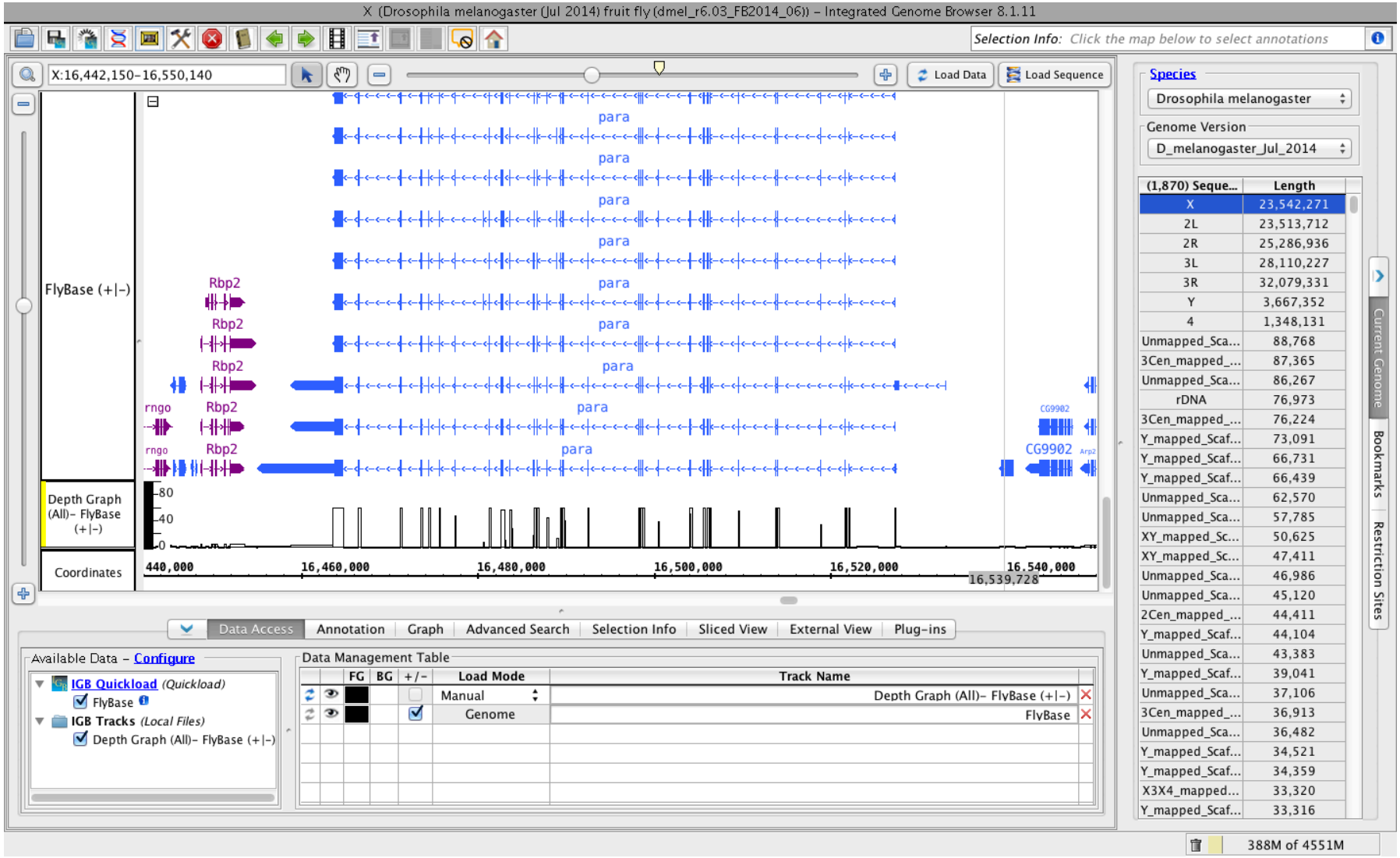
Examples:
ABySS, MIRA, SSAKE
(genome)

Cufflinks, Stringtie,
Trinity
(transcriptome)



(image from Michael Schatz)

Analysis



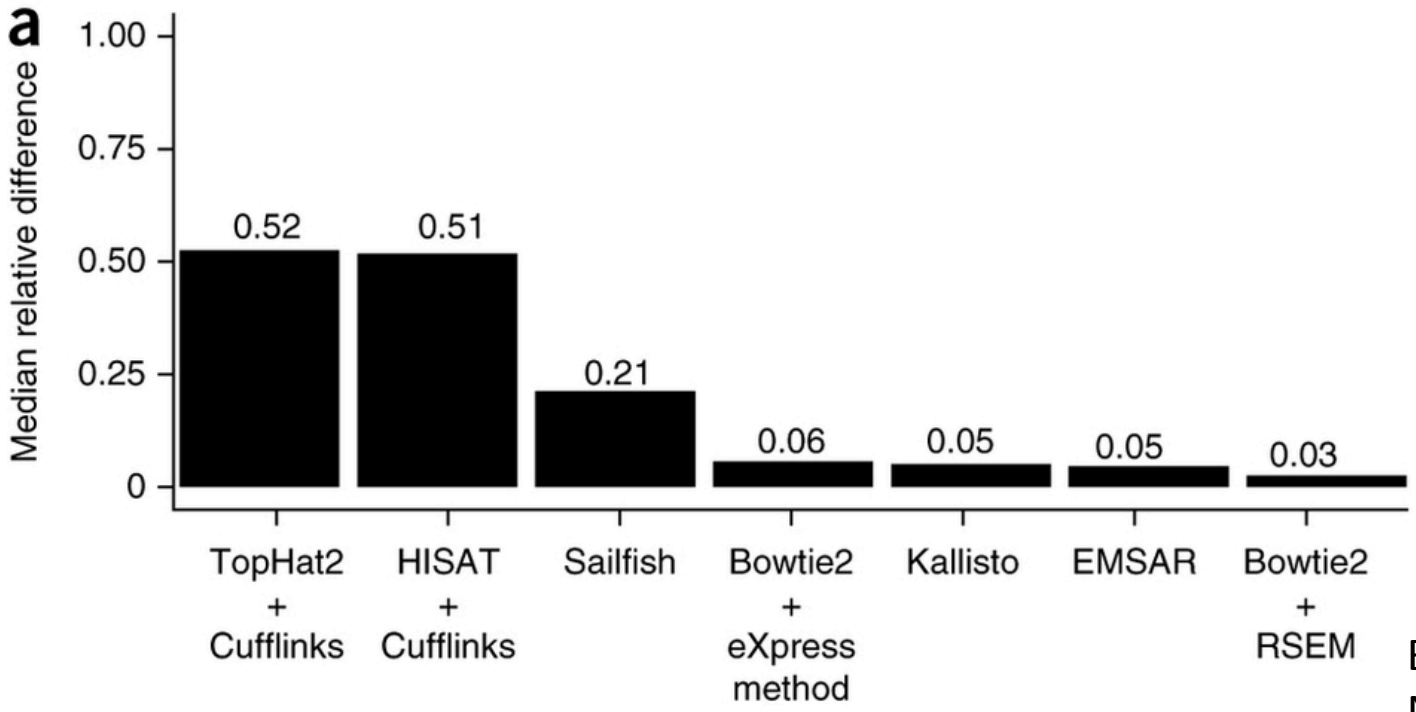
(IGB screenshot by Ann Loraine)

Common tasks

- Transcript quantitation
- Isoform calling
- Peak calling
- Gene set enrichment analysis
- Motif analysis
- Clustering/network inference

Pseudoalignment and fast RNA-seq workflows

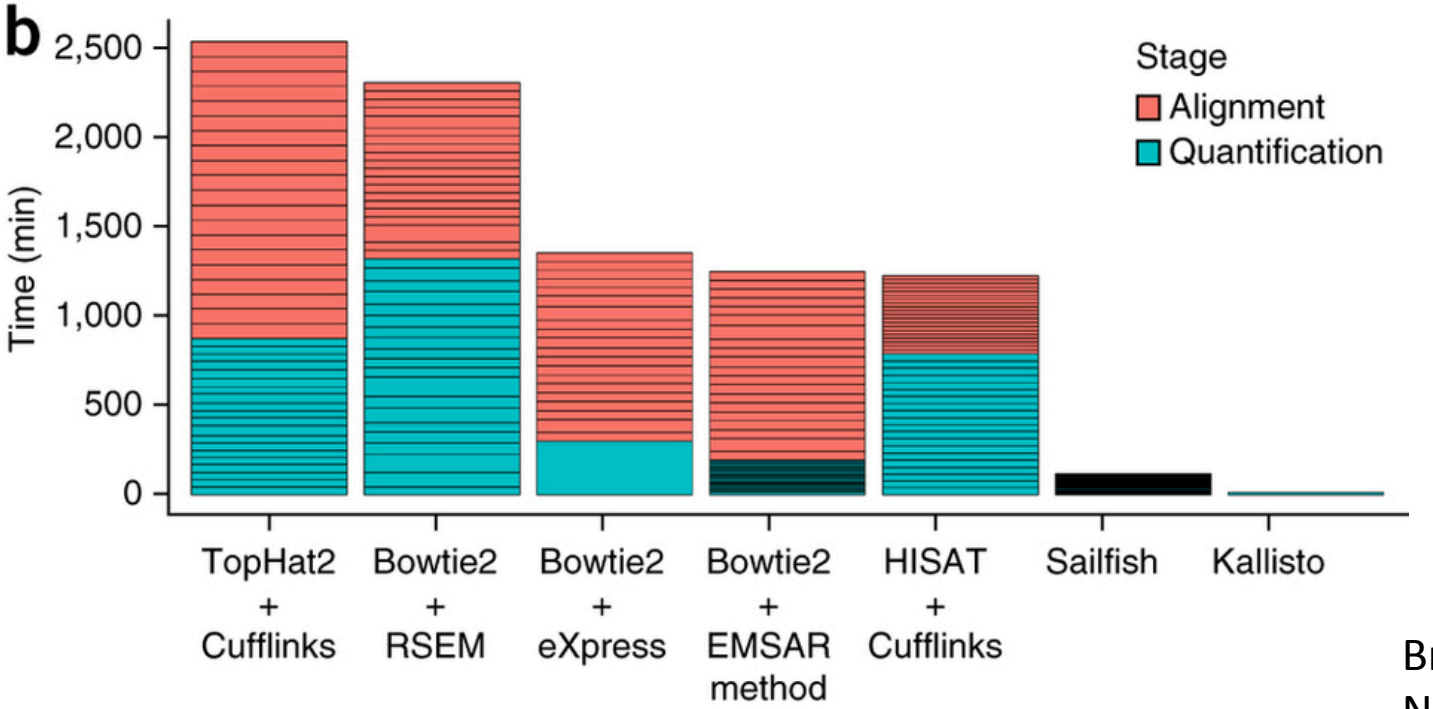
- Don't get exact alignments, just find transcripts compatible with each read
- Examples: kallisto, sailfish, salmon



Bray et al.,
Nat. Biotech. 2016

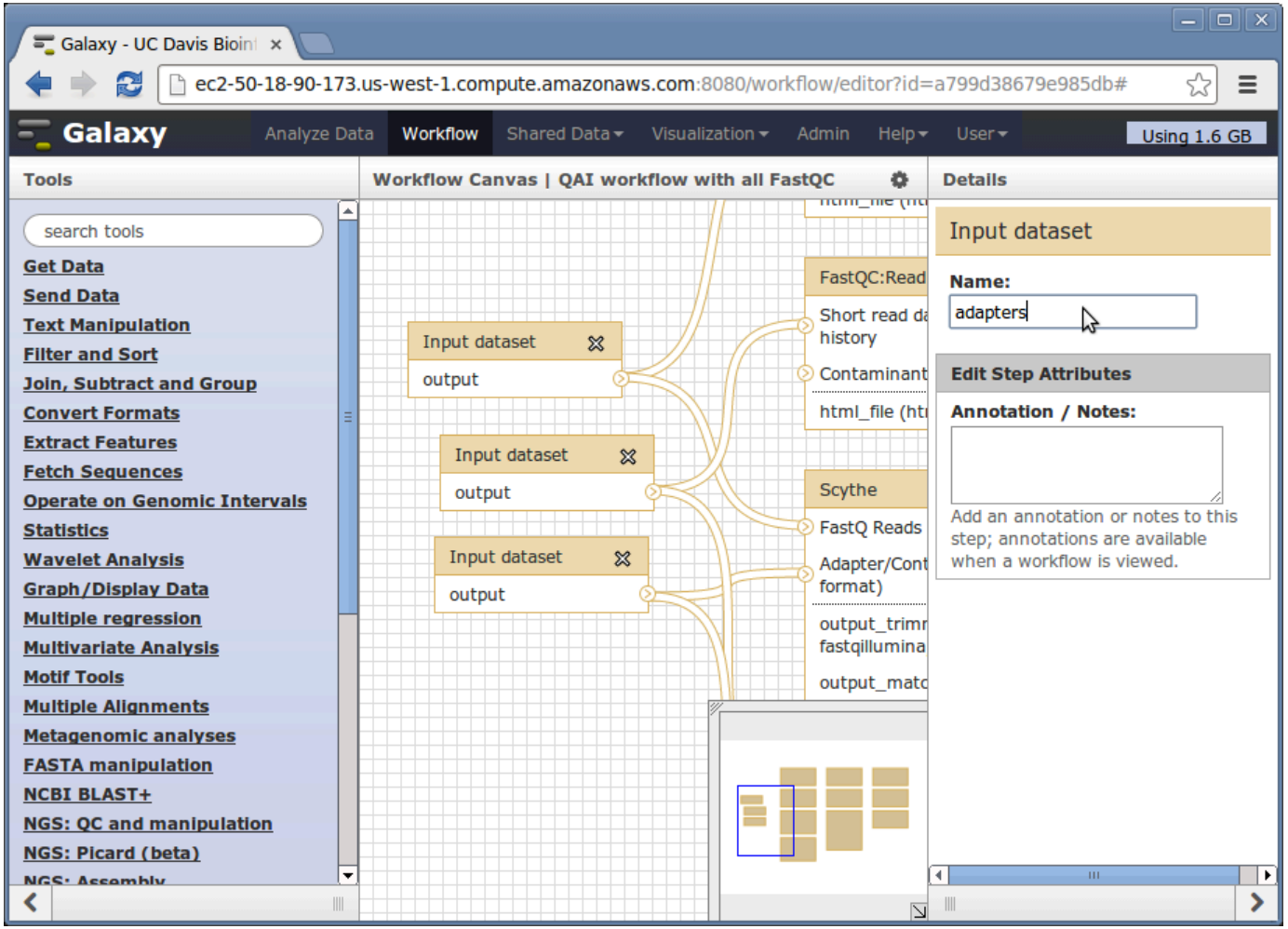
Pseudoalignment and fast RNA-seq workflows

- Don't get exact alignments, just find transcripts compatible with each read
- Examples: kallisto, sailfish, salmon



Bray et al.,
Nat. Biotech. 2016

A unified graphical interface for NGS analysis



Galaxy: usegalaxy.org; image from UC Davis Bioinformatics Core

Outline

- Summary of NGS technologies (sequencing and applications)
- Introduction to NGS data analysis
- **Commonly available databases**
- Workflow integration and making use of existing NGS data

GEO – the gene expression omnibus

www.ncbi.nlm.nih.gov/geo

NCBI Resources How To petefred My NCBI Sign Out

GEO Home | Documentation | Query & Browse | Email GEO | My GEO Submissions

Gene Expression Omnibus

GEO Gene Expression Omnibus

GEO is a public functional genomics data repository supporting MIAME-compliant data submissions. Array- and sequence-based data are accepted. Tools are provided to help users query and download experiments and curated gene expression profiles.

Keyword or GEO Accession | Search

Getting Started

- Overview
- FAQ
- About GEO DataSets
- About GEO Profiles
- About GEO2R Analysis
- How to Construct a Query
- How to Download Data

Tools

- Search for Studies at GEO DataSets
- Search for Gene Expression at GEO Profiles
- Search GEO Documentation
- Analyze a Study with GEO2R
- GEO BLAST
- Programmatic Access
- FTP Site

Browse Content

- Repository Browser

DataSets:	4348
Series:	82875
Platforms:	17052
Samples:	2018686

Information for Submitters

- My GEO Submissions
- My GEO Profile
- Submission Guidelines
- Update Guidelines
- MIAME Standards
- Citing and Linking to GEO
- Guidelines for Reviewers
- GEO Publications

| NLM | NIH | Email GEO | Disclaimer | Accessibility |

Example GEO dataset

NCBI

GEO
Gene Expression Omnibus

HOME | SEARCH | SITE MAP | GEO Publications | FAQ | MIAME | Email GEO

NCBI > GEO > **Accession Display**

Contact: petefred | My submissions | Logout

Scope: Self Format: HTML Amount: Quick GEO accession: GSE32022 GO

Series GSE32022 UPDATE Query DataSets for GSE32022

Status	Public on Jun 01, 2012
Title	Characterization of transcriptional and fitness effects of a loss of function mutation in rho
Platform organism	Escherichia coli
Sample organism	Escherichia coli str. K-12 substr. MG1655
Experiment type	Expression profiling by genome tiling array Genome variation profiling by genome tiling array
Summary	In order to study the effects of a mutation to the transcriptional termination regulator Rho (referred to as rho*), we made use of expression microarrays to observe the direct and indirect effects of rho* on gene expression. In addition, we used arrays to map the fitness of strains from transposon mutagenized libraries under four conditions, showing that in each case the majority of genes with significant fitness effects were dependent on the genotype at rho.
Overall design	For expression arrays, we performed two-color microarrays comparing transcript levels in rho* and wild type cells during exponential growth in glucose minimal media. For selection experiments, transposon insertions were mapped through selective amplification of genomic regions adjacent to them. We then measured the fitness effects of insertions throughout the genome using two-color microarrays, comparing amplified DNA from a population grown under a selective condition of interest to an isogenic control population grown under a reference condition (glucose minimal media). All arrays were performed in duplicate, and the source material for the duplicates came from separate biological replicates.
Contributor(s)	Freddolino PL, Goodarzi H
Citation(s)	Freddolino PL, Goodarzi H, Tavazoie S. Fitness landscape transformation through a single amino acid change in the rho terminator. *PLoS Genet* 2012 May;8(5):e1002744. PMID: 22693458

Example GEO dataset

Platforms (1) GPL10286 Escherichia coli whole-genome tiling array (2 x 105K)

Samples (9)
Less...

GSM794088 RNA
GSM794089 WT_AKG
GSM794090 WT_CML
GSM794091 WT_NADM
GSM794092 WT_STP
GSM794093 MUT_AKG
GSM794094 MUT_CML
GSM794095 MUT_NADM
GSM794096 MUT_STP

Relations

BioProject PRJNA147515

Download family	Format
SOFT formatted family file(s)	SOFT
MINiML formatted family file(s)	MINiML
Series Matrix File(s)	TXT

Supplementary file	Size	Download	File type/resource
GSE32022_RAW.tar	604.3 Mb	(http)(custom)	TAR (of TXT)

Raw data provided as supplementary file

Processed data provided as supplementary file

Example GEO dataset

Platforms (1) GPL10286 Escherichia coli whole-genome tiling array (2 x 105K)

Samples (9)
Less...

GSM794088 RNA
GSM794089 WT_AKG
GSM794090 WT_CML
GSM794091 WT_NADM
GSM794092 WT_STP
GSM794093 MUT_AKG
GSM794094 MUT_CML
GSM794095 MUT_NADM
GSM794096 MUT_STP

Relations

BioProject PRJNA147515

Download family	Format
SOFT formatted family file(s)	SOFT
MINiML formatted family file(s)	MINiML
Series Matrix File(s)	TXT

Supplementary file	Size	Download	File type/resource
GSE32022_RAW.tar	604.3 Mb	(http)(custom)	TAR (of TXT)

Raw data provided as supplementary file

Processed data provided as supplementary file

Example GEO dataset

NCBI

GEO Gene Expression Omnibus

HOME | SEARCH | SITE MAP | GEO Publications | FAQ | MIAME | Email GEO

NCBI > GEO > **Accession Display**

Contact: petefred | My submissions | Logout

Scope: Self Format: HTML Amount: Quick GEO accession: GSM794088 GO

Sample GSM794088 UPDATE Query DataSets for GSM794088

Status	Public on Jun 01, 2012
Title	RNA
Sample type	RNA

Channel 1	
Source name	WT cells_mid-log-phase_M9t/glucose
Organism	Escherichia coli str. K-12 substr. MG1655
Characteristics	genotype/variation: WT growth media: M9t/glucose growth phase: mid-log
Growth protocol	For RNA samples, WT or rho* cells were grown to mid-log phase in M9t/glucose. WT or rho* transposon mutagenized libraries were grown overnight in the media indicated.
Extracted molecule	total RNA
Extraction protocol	Total RNA was extracted using total RNA purification kit (Norgen Biotek, Cat 17200).
Label	Cy5
Label protocol	A poly-A tail was added to the RNA samples using E. coli Poly(A) polymerase (NEB, M0276) for 15 minutes. Using an Agilent low input quick amp labeling kit, the rho* and WT samples were then labeled with Cy3 and Cy5, respectively.

Channel 2	
Source name	rho* cells_mid-log-phase_M9t/glucose
Organism	Escherichia coli str. K-12 substr. MG1655
Characteristics	genotype/variation: rho* growth media: M9t/glucose growth phase: mid-log
Growth protocol	For RNA samples, WT or rho* cells were grown to mid-log phase in M9t/glucose. WT or rho* transposon mutagenized libraries were grown overnight in the media indicated.
Extracted molecule	total RNA
Extraction protocol	Total RNA was extracted using total RNA purification kit (Norgen Biotek, Cat 17200).

Example GEO dataset

Supplementary file	Size	Download	File type/resource	
GSM794088_TAHG20110218_252456810085_S01_GE2-v5_95_Feb07_1_1.txt.gz	32.9 Mb	(ftp)(http)	TXT	Raw
GSM794088_TAHG20110218_252456810085_S01_GE2-v5_95_Feb07_1_2.txt.gz	32.8 Mb	(ftp)(http)	TXT	Raw
GSM794088_rna_lograt_zscore.txt.gz	1.3 Mb	(ftp)(http)	TXT	Processed

Raw data provided as supplementary file

Processed data provided as supplementary file

Navigating GEO datasets

- GPLXXXXX – Platform identifier
- GSEXXXXX – series of data sets (e.g., one paper)
- GSMXXXXX – One sample (may be one or more replicates, but should be same condition)
- GDSXXXXX – Curated data set with additional options available

Getting GEO expression data

Easiest: Use curated data sets
https://www.ncbi.nlm.nih.gov/sites/GDSbrowser/

NCBI

DATASET BROWSER (CURATED)

GEO Gene Expression Omnibus

Search for [ ] Search Clear Show All Advanced Search

Page size 20

4348 DataSet records

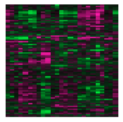
Page 1 of 218 > >>

DataSet	Title	Organism(s)	Platform	Series	Samples
GDS6082	Sendai virus infection effect on monocytic cell line: dose response	*Homo sapiens*	GPL10558	GSE67198	11
GDS6064	Arthritic tarsal joints induced by collagen: time course	*Mus musculus*	GPL6246	GSE61140	15
GDS6063	Influenza A effect on plasmacytoid dendritic cells	*Homo sapiens*	GPL10558	GSE68849	10
GDS6016	Transcription factor engrailed-2 loss-of-function model of autism spectrum disorder: hippocam...	*Mus musculus*	GPL7202	GSE51612	6
GDS6010	Influenza virus H5N1 infection of U251 astrocyte cell line: time course	*Homo sapiens*	GPL6480	GSE66597	18
GDS6000	High-fat diet effect on brown adipose tissue development	*Mus musculus*	GPL6887	GSE64718	33
GDS5948	Zipper-interacting protein kinase deficiency effect on coronary artery smooth muscle cells in vi...	*Homo sapiens*	GPL6244	GSE56819	6
GDS5914	YAP transcriptional regulator depletion effect on endothelial cells	*Homo sapiens*	GPL6244	GSE61989	12
GDS5913	SRPIN803 small molecule inhibitor of SRPK1 effect on retinal pigment epithelial cell line	*Homo sapiens*	GPL570	GSE62947	6
GDS5881	Nebulin deficiency effect on the soleus	*Mus musculus*	GPL6246	GSE70213	12

DataSet Record GDS6177: Expression Profiles | Data Analysis Tools | Sample Subsets

Title:	Acute alcohol consumption effect on whole blood (control group): time course		
Summary:	Analysis of blood from subjects administered orange juice w/o alcohol. Blood collected at time points corresponding to collection times for the alcohol group in GDS4938. These results, together with those from GDS4938, provide insight into molecular response of blood during acute ethanol exposure.		
Organism:	*Homo sapiens*		
Platform:	GPL570: [HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array		
Citation:	Kupfer DM, White VL, Strayer DL, Crouch DJ et al. Microarray characterization of gene expression changes in blood during acute ethanol exposure. *BMC Med Genomics* 2013 Jul 25;6:26. PMID: 23883607		
Reference Series:	GSE20489	**Sample count:**	25
Value type:	transformed count	**Series published:**	2013/07/25

Cluster Analysis



Download

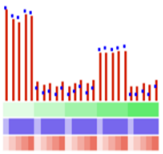
- DataSet full SOFT file
- DataSet SOFT file
- Series family SOFT file
- Series family MINiML file
- Annotation SOFT file

NLM NIH GEO Help Disclaimer Accessibility

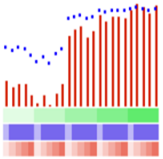
Getting GEO expression data

Easiest: Use curated data sets
https://www.ncbi.nlm.nih.gov/sites/GDSbrowser/

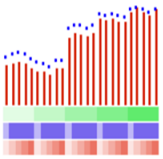
1. HLA-DRB4 - Acute alcohol consumption effect on whole blood (control group): time course
   Annotation: HLA-DRB4, major histocompatibility complex, class II, DR beta 4
   Organism: Homo sapiens
   Reporter: GPL570, 215666_at (ID_REF), **GDS6177**, 3126 (Gene ID), U70544
   DataSet type: Expression profiling by array, transformed count, 25 samples
   ID: 132643760
   GEO DataSets Gene UniGene Profile neighbors



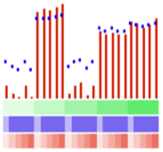
2. TMEM176B - Acute alcohol consumption effect on whole blood (control group): time course
   Annotation: TMEM176B, transmembrane protein 176B
   Organism: Homo sapiens
   Reporter: GPL570, 220532_s_at (ID_REF), **GDS6177**, 28959 (Gene ID), NM_014020
   DataSet type: Expression profiling by array, transformed count, 25 samples
   ID: 132648617
   GEO DataSets Gene UniGene Profile neighbors Chromosome neighbors Homologene neighbors



3. TMEM176A - Acute alcohol consumption effect on whole blood (control group): time course
   Annotation: TMEM176A, transmembrane protein 176A
   Organism: Homo sapiens
   Reporter: GPL570, 218345_at (ID_REF), **GDS6177**, 55365 (Gene ID), NM_018487
   DataSet type: Expression profiling by array, transformed count, 25 samples
   ID: 132646431
   GEO DataSets Gene UniGene Profile neighbors Chromosome neighbors Homologene neighbors



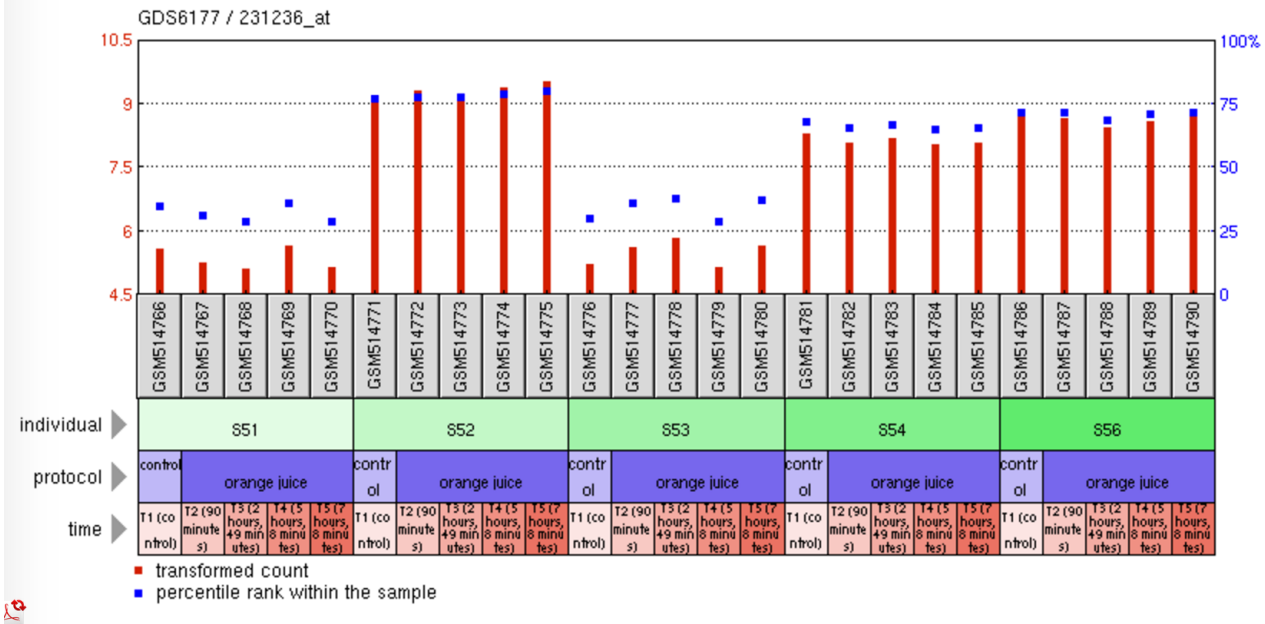
4. ZFP57 - Acute alcohol consumption effect on whole blood (control group): time course
   Annotation: ZFP57, ZFP57 zinc finger protein
   Organism: Homo sapiens
   Reporter: GPL570, 231236_at (ID_REF), **GDS6177**, 346171 (Gene ID)
   DataSet type: Expression profiling by array, transformed count, 25 samples
   ID: 132659291
   GEO DataSets Gene UniGene Profile neighbors Chromosome neighbors Homologene neighbors



Getting GEO expression data

Easiest: Use curated data sets
https://www.ncbi.nlm.nih.gov/sites/GDSbrowser/

Profile GDS6177 / 231236_at
Title Acute alcohol consumption effect on whole blood (control group): time course
Organism Homo sapiens



Graph caption help

Sample	Title	Value	Rank
GSM514766	Blood_OJcontrol_T1_S51	5.61881	35
GSM514767	Blood_OJcontrol_T2_S51	5.28791	31
GSM514768	Blood_OJcontrol_T3_S51	5.13084	29

Getting GEO expression data

Otherwise: Get processed data from GSM pages…

Supplementary file	Size	Download	File type/resource	
GSM794088_TAHG20110218_252456810085_S01_GE2-v5_95_Feb07_1_1.txt.gz	32.9 Mb	(ftp)(http)	TXT	Raw
GSM794088_TAHG20110218_252456810085_S01_GE2-v5_95_Feb07_1_2.txt.gz	32.8 Mb	(ftp)(http)	TXT	Raw
GSM794088_rna_lograt_zscore.txt.gz	1.3 Mb	(ftp)(http)	TXT	Processed

Raw data provided as supplementary file

Processed data provided as supplementary file

| NLM | NIH | GEO Help | Disclaimer | Accessibility |

GEO isn't JUST about gene expression…

Platforms (1) GPL17021 Illumina HiSeq 2500 (Mus musculus)

Samples (39)
More…

GSM2143252 114_FCX_120m_Control_input
GSM2143253 115_FCX_120m_Treatment_input
GSM2143254 117_FCX_120m_Control_H3K27ac

This SubSeries is part of SuperSeries:
GSE80345 Transcriptional regulatory dynamics set the stage for a coordinated metabolic and neural response to social threat in mice

Relations

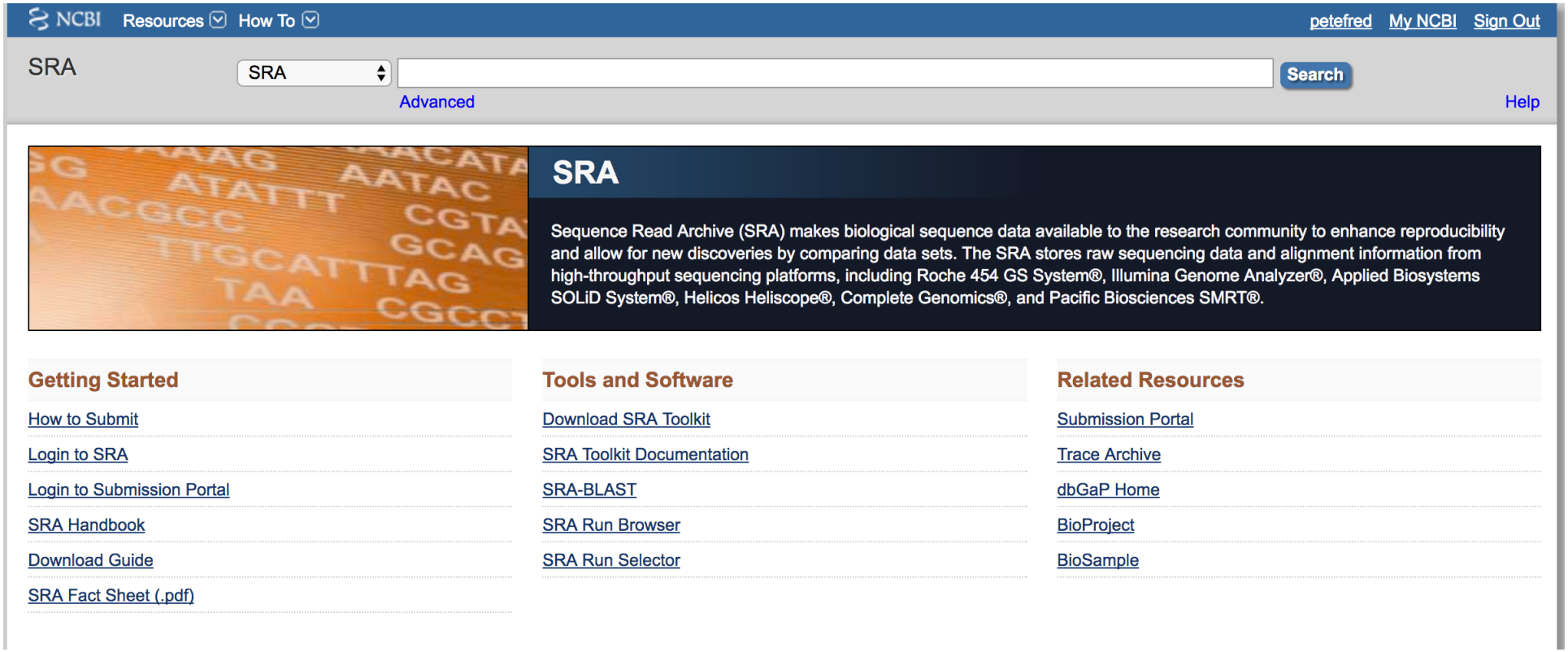
BioProject PRJNA320640
SRA SRP074385

Download family	Format
SOFT formatted family file(s)	SOFT
MINiML formatted family file(s)	MINiML
Series Matrix File(s)	TXT

Supplementary file	Size	Download	File type/resource
GSE81122_114_FCX_120min_CK_input-117+118.ucsc.bigWig	133.5 Mb	(ftp)(http)	BIGWIG
GSE81122_115_FCX_120min_EX_input-119+120.ucsc.bigWig	163.6 Mb	(ftp)(http)	BIGWIG
GSE81122_117+118_FCX_120min_CK_1M_H3K27ac.ucsc.bigWig	317.5 Mb	(ftp)(http)	BIGWIG
GSE81122_119+120_FCX_120min_EX_1M_H3K27ac.ucsc.bigWig	242.3 Mb	(ftp)(http)	BIGWIG
GSE81122_121+122_Amy_CK_120min_1M_H3K27ac.ucsc.bigWig	351.1 Mb	(ftp)(http)	BIGWIG
GSE81122_125_Amy_CK_120min_0.1M_input.ucsc.bigWig	173.9 Mb	(ftp)(http)	BIGWIG
GSE81122_126+127_Amy_Exp_120min_1M_H3K27ac.ucsc.bigWig	352.7 Mb	(ftp)(http)	BIGWIG

Raw data available via the SRA

https://www.ncbi.nlm.nih.gov/sra/



Raw data available via the SRA

Sequence Read Archive

Main | Browse | Search | Download | Submit | Documentation | Software | Trace Archive | Trace Assembly | Trace BLAST

Studies | Samples | Analyses | Run Browser | Run Selector | Provisional SRA

Massive transcriptional start site mapping of human fetal brain cells.

Identifiers:	SRA:	DRP000023
	BioProject:	PRJDA34559
	UT-MGS:	DRP000023
Study Type:	Transcriptome Analysis	
Submission:	DRA000023	
Abstract:	Comprehensive identification and characterization of the transcriptional start sites of human genes were carried out. For this purpose, we used our TSS-Seq method, in which next gene sequencing technology and our full-length cDNA library technology, oligo-capping were combined.	
Description:	Although recent studies have revealed that the majority of human genes are subjected to regulation of alternative promoters (APs), the biological relevance of this phenomenon remains unclear. To enable more comprehensive TSS analysis in the respective cell types, we recently developed a method, combining oligo-capping with the massively paralleled sequencing technology, Illumina GA. In this method, which we named TSS Seq, sequence adaptor which is necessary for Illumina GA sequencing is directly introduced to the cap site of the mRNA. By sequencing 36-48 sequence immediately downstream of the TSSs (TSS tags), it is possible to obtain precise positional information of transcriptional start sites (TSSs). In this paper, we used the TSS tag data accumulated from twelve different cell types and normal tissues in humans for the identification and characterization of the APs in human genes.	
Center Project:	Integrateve Transcriptome Analysis	
External Link:	DBTSS	

Related SRA data

Experiments: 1
Runs: 4 (880.5Mbp; 2.5Gb)

Raw data available via the SRA

Processed data often available through GEO link

Sequence Read Archive

Main | Browse | Search | Download | Submit | Documentation | Software | Trace Archive | Trace Assembly | Trace BLAST

Studies | Samples | Analyses | Run Browser | Run Selector | Provisional SRA

Histone modification (H3K4me3 and H3K27me3) during vascular endothelial cell differentiation from mouse embryonic stem cells

Identifiers:
SRA: SRP099437
BioProject: PRJNA374538
GEO: GSE94828

Study Type: Other

Submission: SRA537391

Abstract: Although studies of the differentiation from mouse embryonic stem (ES) cells to vascular endothelial cells (ECs) provide an excellent model for investigating the molecular mechanisms underlying vascular development, temporal dynamics of gene expression and chromatin modifications have not been well studied. Herein, using transcriptomic and epigenomic analyses based on the H3K4me3 and H3K27me3 modifications at a genome-wide scale, we analyzed the EC differentiation steps from ES cells and crucial epigenetic modifications unique to ECs. We determined that Gata2, Fli1, Sox7, and Sox18 are master regulators of EC induced following expression of the hemangioblast commitment pioneer factor, Etv2. These master regulator gene loci were repressed by H3K27me3 under the mesoderm period, but rapidly transitioned to the histone modification switching from H3K27me3 to H3K4me3 after treatment with vascular

Related SRA data

Experiments: 20
Runs: 20 (22.1Gbp; 12.6Gb)

Outline

- Summary of NGS technologies (sequencing and applications)
- Introduction to NGS data analysis
- Commonly available databases
- **Workflow integration and making use of existing NGS data**

Simplest: Download processed data and view in spreadsheet

Simplest: Download processed data and view in spreadsheet

Example: gene_exp.diff from cufflinks

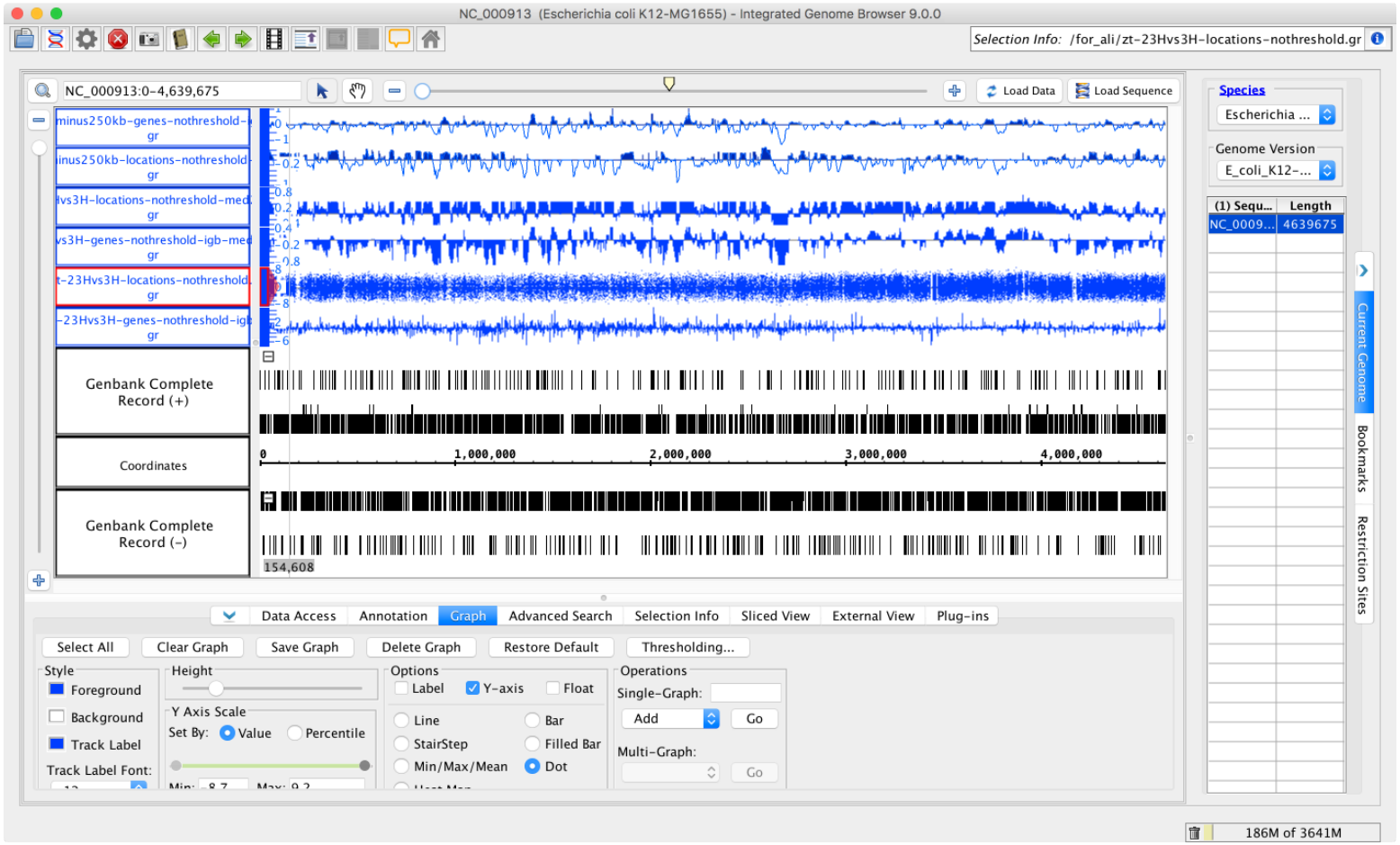
test_id	gene_id	gene	locus	sample_1	sample_2	status	value_1	value_2	log2(fold_change)	test_stat	p_value	q_value	sig
XLOC_000001	XLOC_000001	CG11023	2L: 7528-9484	cupcake_sated	cupcake_hungry	OK	5.49313	0.206789	-4.7314	-1.90088	0.42235	0.593877	no
XLOC_000002	XLOC_000002	Ir21a	2L: 21918-25163	cupcake_sated	cupcake_hungry	OK	2106.08	2913.7	0.468291	2.53576	0.85965	0.903137	no
XLOC_000031	XLOC_000031	dbr	2L: 67043-71390	cupcake_sated	cupcake_hungry	OK	14.9389	16.8551	0.174119	0.230548	0.67695	0.781675	no
XLOC_000032	XLOC_000032	galectin	2L: 72387-76211	cupcake_sated	cupcake_hungry	OK	112.48	85.6742	-0.39273	-0.67515	0.24695	0.423915	no
XLOC_000033	XLOC_000033	CG11374	2L: 76445-77639	cupcake_sated	cupcake_hungry	OK	6.13796	14.0256	1.19223	0.705069	0.22855	0.416548	no
XLOC_000034	XLOC_000034	-	2L: 80193-80263	cupcake_sated	cupcake_hungry	OK	0	402.552	inf	#NAME?	0.01965	0.158954	no

Simplest: Download processed data and view in spreadsheet

Example: gene_exp.diff from cufflinks

test_id	gene_id	gene	locus	sample_1	sample_2	status	value_1	value_2	log2(fold_change)	test_stat	p_value	q_value	sig
XLOC_002456	XLOC_002456	TepIV	2L:19549792-19556455	cupcake_sated	cupcake_hungry	OK	27.2788	0.762464	-5.16097	-3.36415	5.00E-05	0.004022	yes
XLOC_004017	XLOC_004017	dp	2L:4479470-4591963	cupcake_sated	cupcake_hungry	OK	1.66149	0.451918	-1.87835	-2.1873	0.00045	0.024242	yes
XLOC_008185	XLOC_008185	Cam	2R:8146912-8166208	cupcake_sated	cupcake_hungry	OK	2811.67	2026.02	-0.47278	-1.67712	0.0034	0.034841	yes

Genome Browsers



Allow loading and comparisons of various data sets with genomic features

Main candidates: IGB, IGV, UCSC Genome Browser

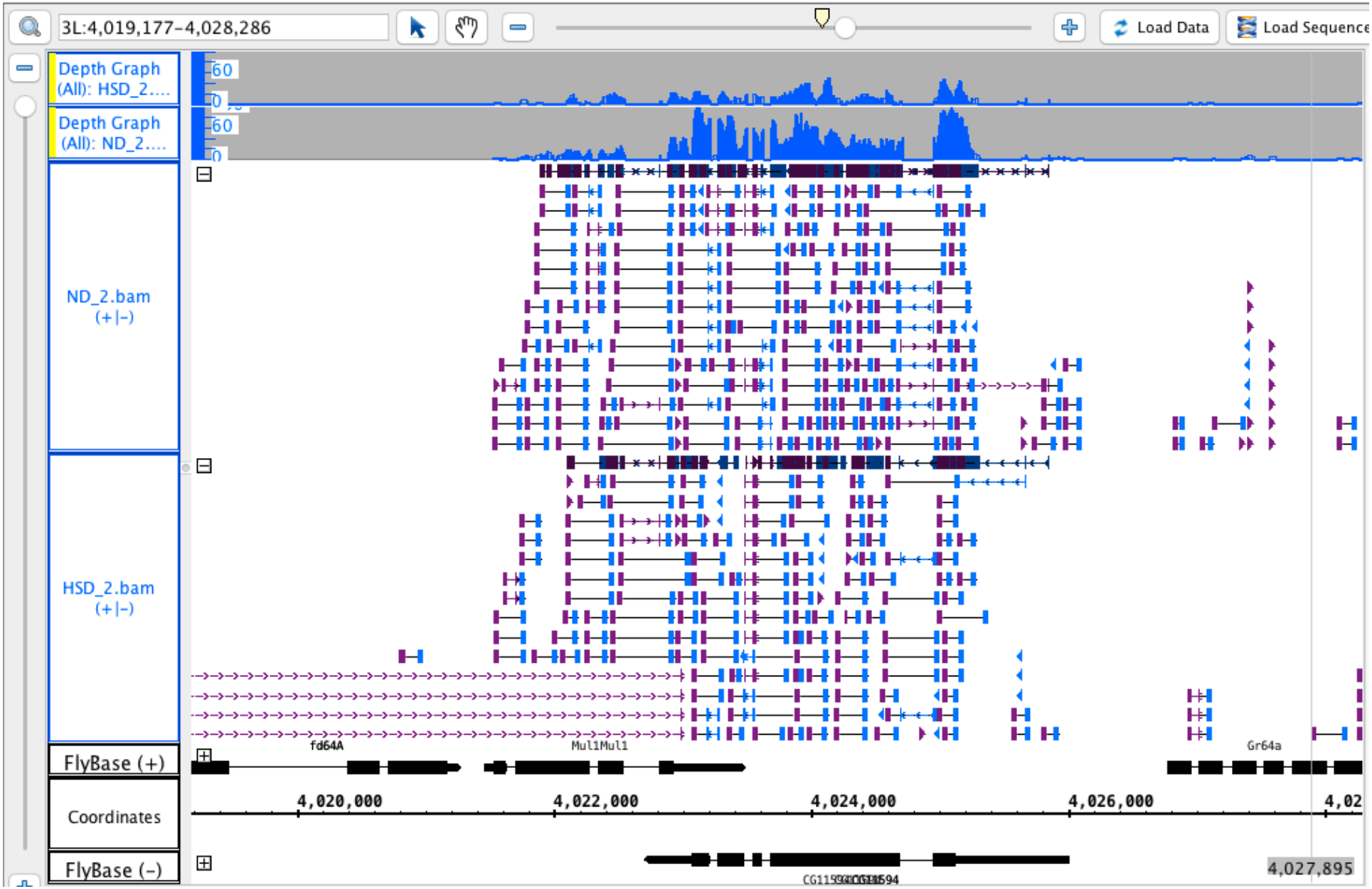
Genome Browsers

Analysis can scale from raw reads to highly processed functions of multiple data sets

Genome Browsers

Analysis can scale from raw reads to highly processed functions of multiple data sets

Genome Browsers



Analysis can scale from raw reads to highly processed functions of multiple data sets

Genome Browsers

NCBI

GEO Gene Expression Omnibus

HOME | SEARCH | SITE MAP | GEO Publications | FAQ | MIAME | Email GEO

NCBI > GEO > **Accession Display**

Contact: petefred | My submissions | Logout

Scope: Self Format: HTML Amount: Quick GEO accession: GSE87509 GO

Series GSE87509 Query DataSets for GSE87509

Status	Public on Mar 16, 2017
Title	ChIP-seq of Atrophin in Drosophila S2 cells
Organism	Drosophila melanogaster
Experiment type	Genome binding/occupancy profiling by high throughput sequencing
Summary	Drosophila Atro mutants have a large range of phenotypes, including neurodegeneration, segmentation, patterning and planar polarity defects. Although Atro mutants have diverse phenotypes, little is known about Atro's binding partners and downstream targets. We present the first genomic analysis of Atro using ChIP-seq against endogenous Atro. These data sets will serve as a valuable resource for future studies on Atro.
Overall design	We performed three independent biological replicates of Atro ChIP-seq experiments in untreated S2 cells. A corresponding non-specific IgG control ChIP was performed with each Atro ChIP-seq and was used as a control.

Many GEO files are directly loadable

Genome Browsers

...

Download family	Format
SOFT formatted family file(s)	SOFT
MINiML formatted family file(s)	MINiML
Series Matrix File(s)	TXT

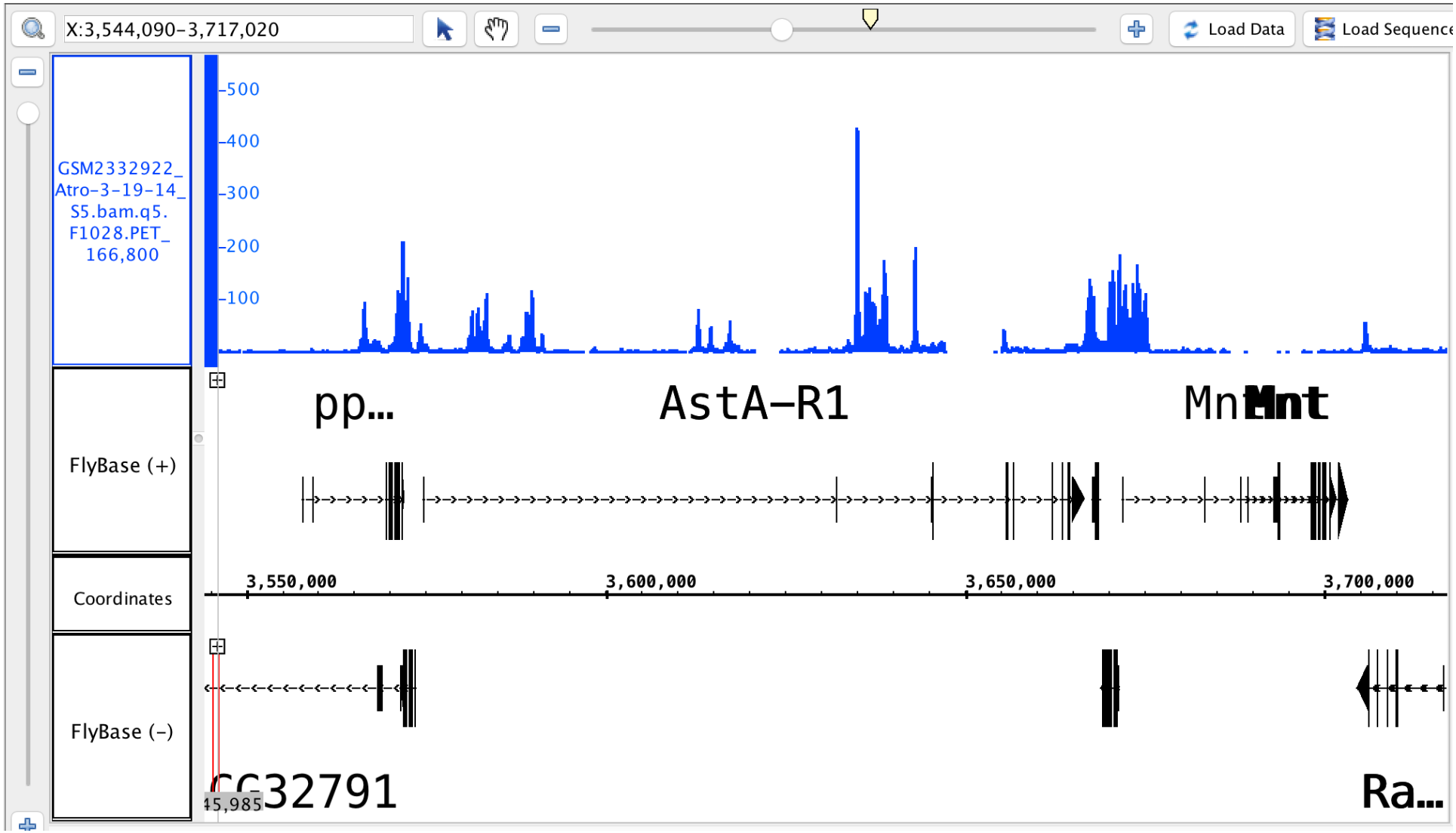
Supplementary file	Size	Download	File type/resource
SRP/SRP090/SRP090681		(ftp)	SRA Study
GSE87509_RAW.tar	34.5 Mb	(http)(custom)	TAR (of WIG)

Raw data provided as supplementary file

Processed data provided as supplementary file

Many GEO files are directly loadable

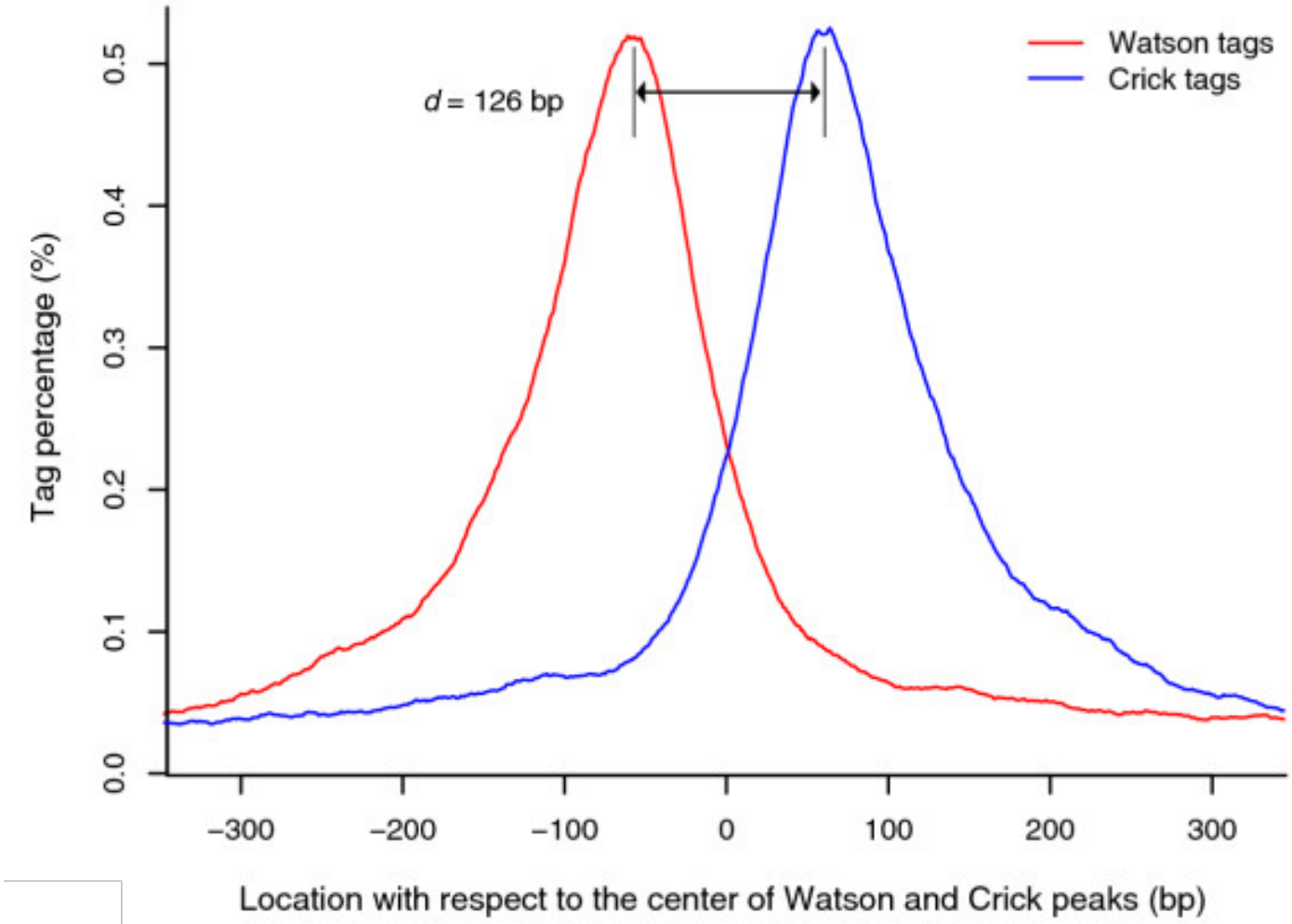
Genome Browsers



Many GEO files are directly loadable

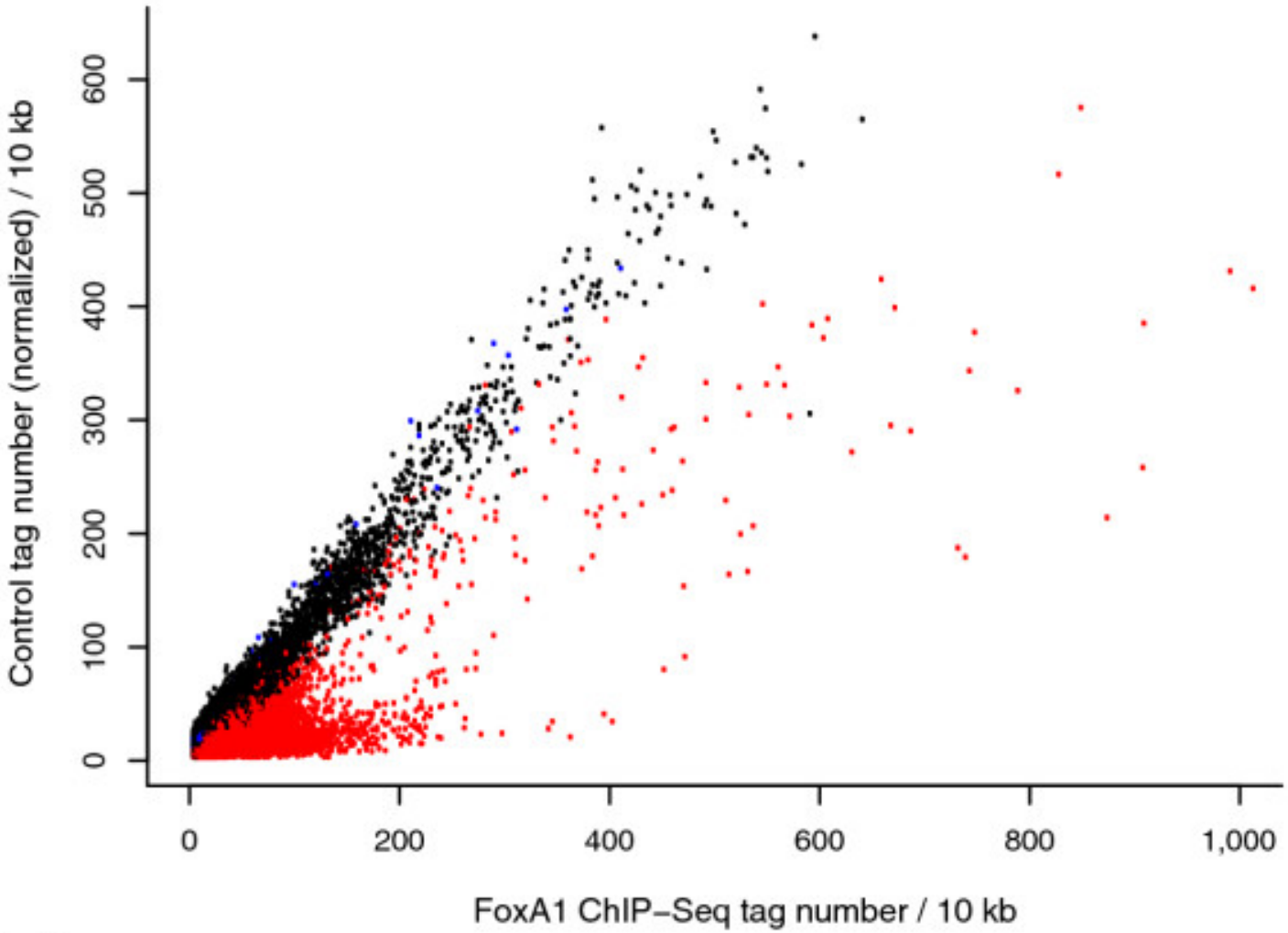
Peak calling or differential calling are common tasks

Peak calling or differential calling are common tasks



(Image from Zhang et al., Genome Biol., 2008)

Peak calling or differential calling are common tasks



(Image from Zhang et al., Genome Biol., 2008)

So what do you do once you have peaks/expression calls/etc.?

- Direct inspection of known biological targets
- Literature-driven inference and hypothesis generation
- Gene set enrichment analysis
- Motif analysis
- Network inference

Gene set enrichment analysis

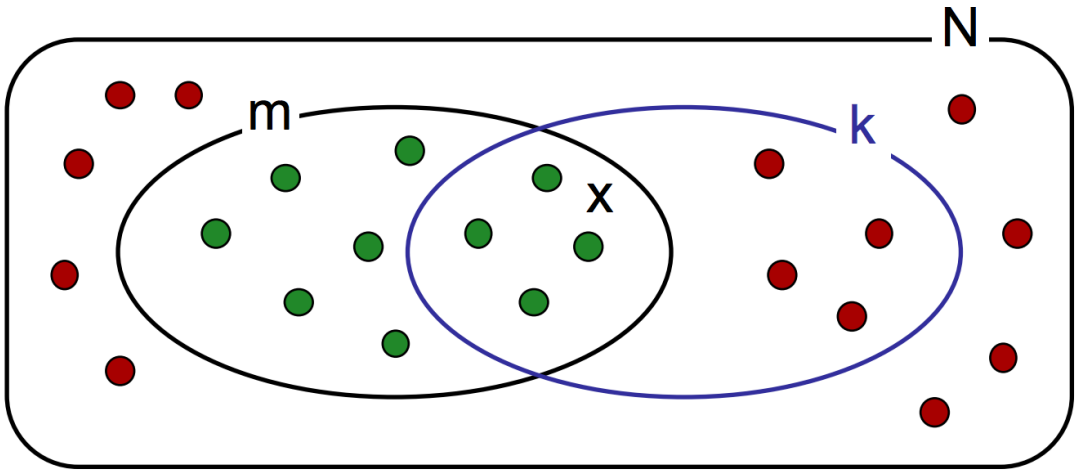
Identification of gene categories (e.g., GO terms)
that are correlated with another data set

Common Tools: GSEA, DAVID, iPAGE

Gene set enrichment analysis

Identification of gene categories (e.g., GO terms) that are correlated with another data set

Common Tools: GSEA, DAVID, iPAGE



N = total number of elements

m = number of marked elements

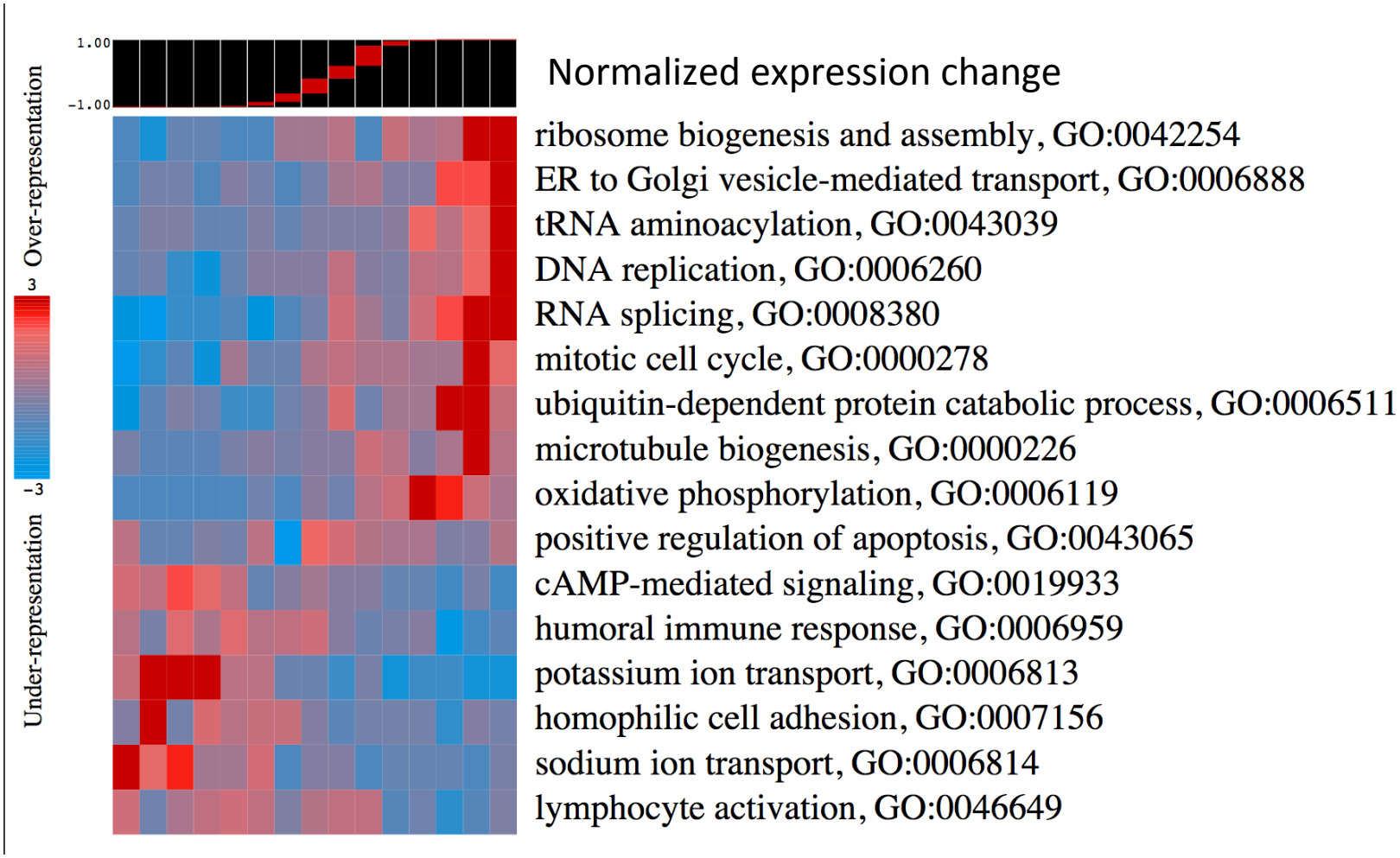
k = number of sampled elements

x = number of marked sampled elements

Gene set enrichment analysis

Identification of gene categories (e.g., GO terms) that are correlated with another data set

Example: Gene expression

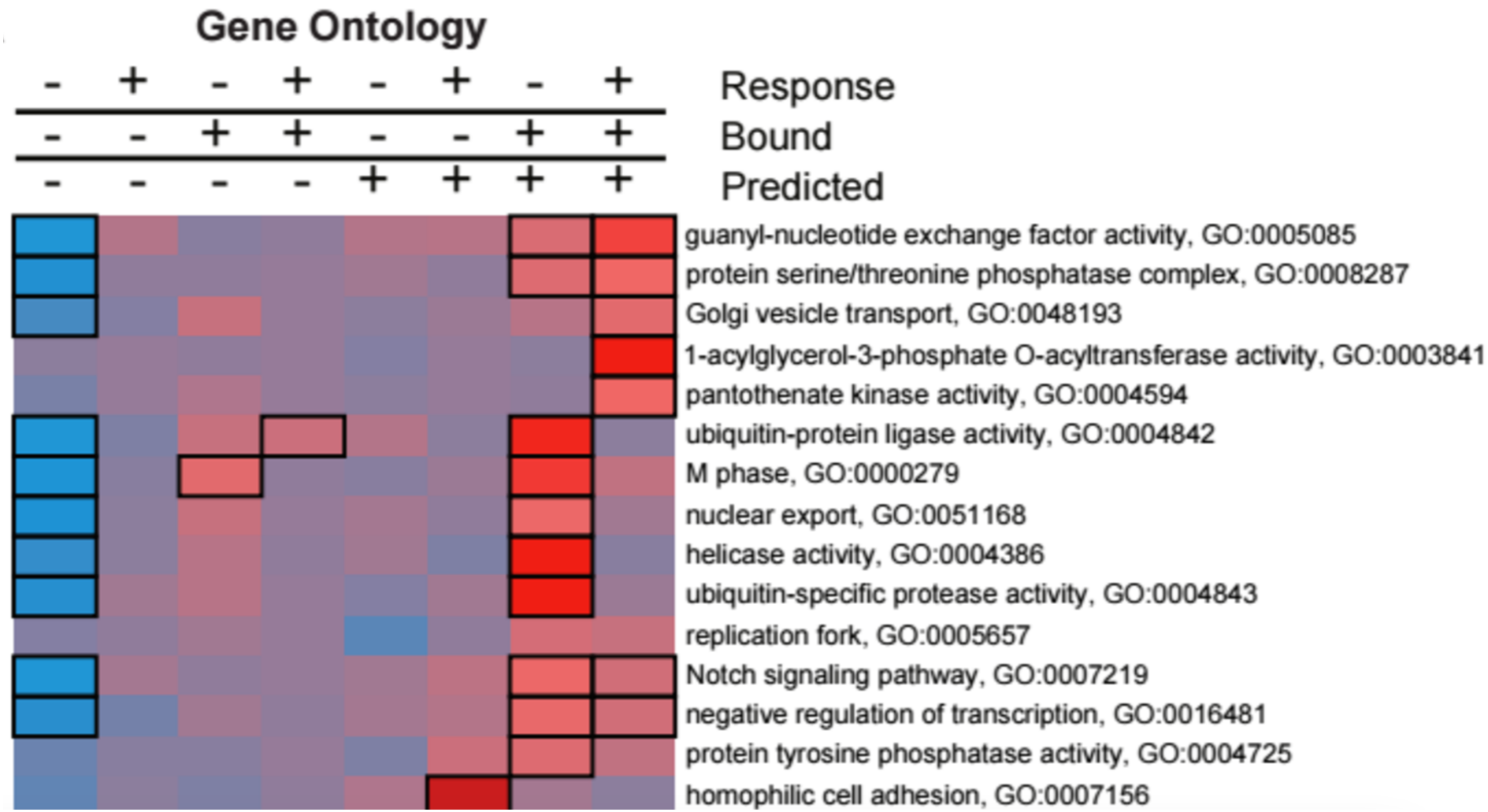


(From Goodarzi et al., Mol. Cell, 2009)

Gene set enrichment analysis

Identification of gene categories (e.g., GO terms) that are correlated with another data set

Example: Integration of data sets



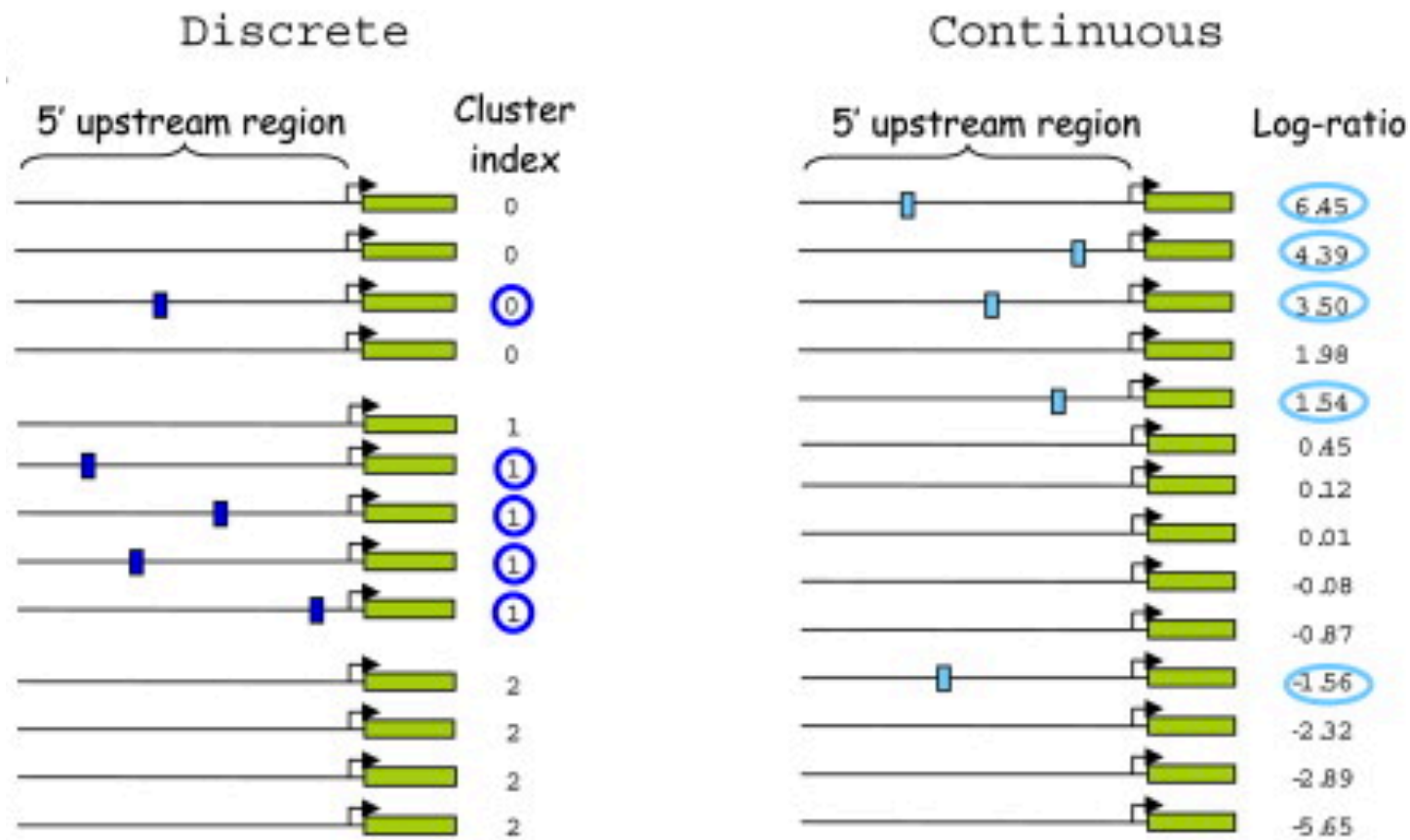
Motif analysis

Identify motifs (typically nucleic acid sequences) correlated with a data set of interest

Used in a variety of applications (RNA-seq, ChIP-seq, ribosome profiling, etc.)

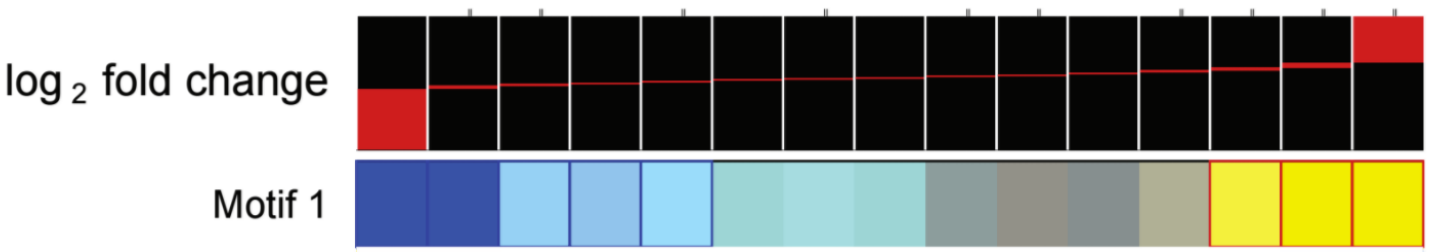
Example tools: MEME suite, FIRE/TEISER, kmersvm

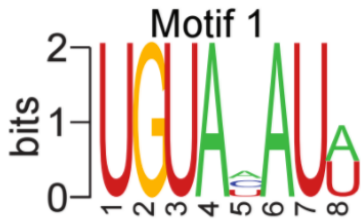
Motif analysis

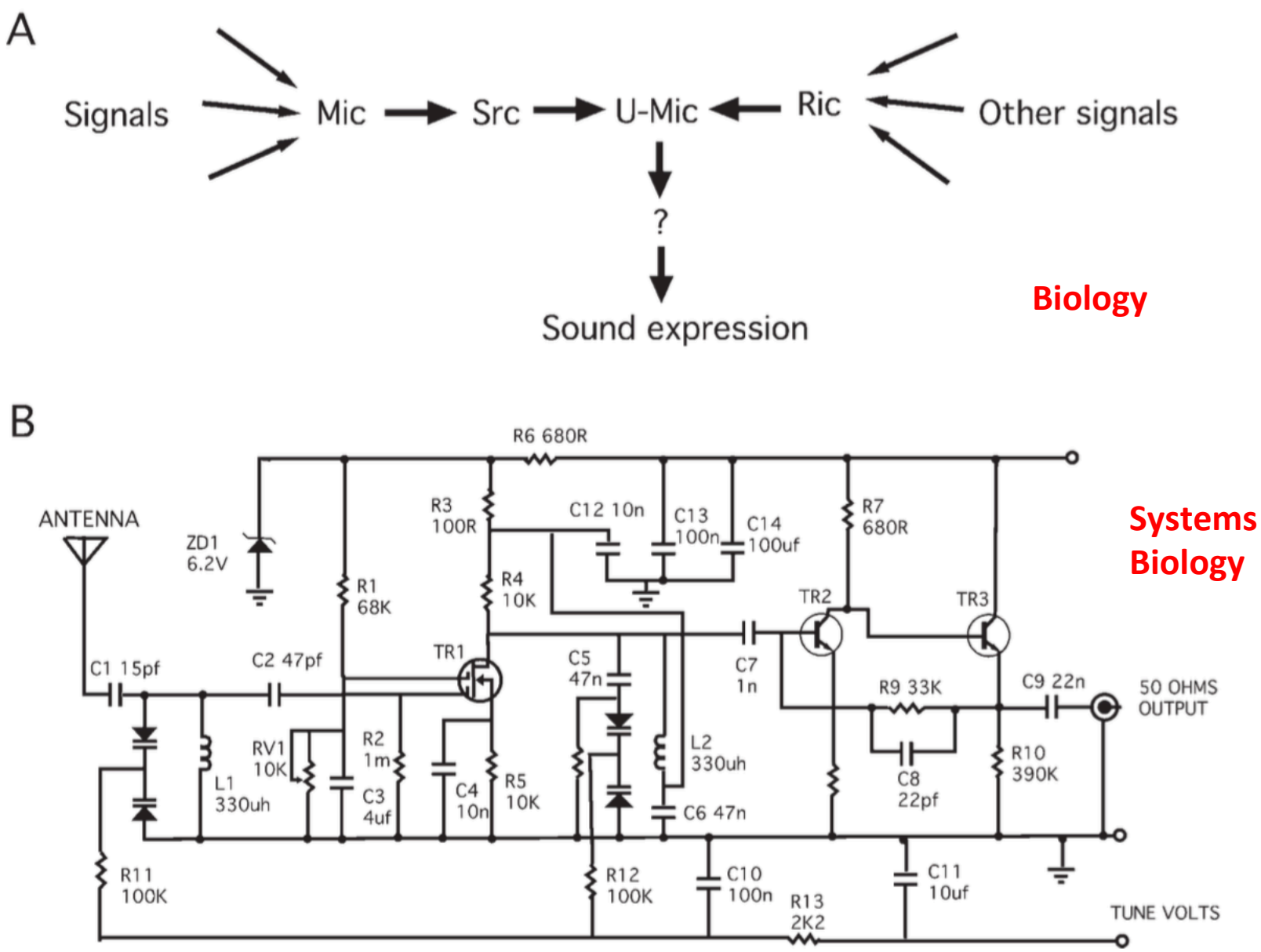


(Image from Elemento *et al.*, Mol. Cell 2007)

Motif analysis







Lazebnik, Y. Cancer Cell 2002
(slides via Michael Wolfe)