Midterm exam: Friday, April 30

Exam covers topics through the end of this discussion of '-Omics'

See quizzes 1-5, and assignments 1 and 2 for examples of the kinds of questions to expect

Parallel measurements of gene expression/activity: RNA and protein " -omics"

1. The 'transcriptome'
   - Detecting expression of many genes: arrays of inverted northern blots
   - RNA-seq: 'next gen' sequencing provides an alternative to arrays

2. The 'proteome'
   - ID of all proteins in a mixture: 2-D gels and mass spectrometry
   - ID of DNA-binding protein locations: Chromatin Immunoprecipitation
   - Protein activity measurements

Guide to readings:

1)21 MC4 ChIP. Introduction to Chromatin Immunoprecipitation for DNA binding proteins

2) RNA seq 2009. The RNA-seq approach to transcriptome analysis.
3) Chip seq 2010. Chromatin immunoprecipitation followed by sequencing.
4) Mass spec proteomics 2007. Using mass spectrometry for identifying proteins in a complex mixture.
5) The ENCODE project: Encyclopedia of human DNA elements
6) Mass data false positives 2012. Big data sets can lead to false conclusions

What is an –ome?

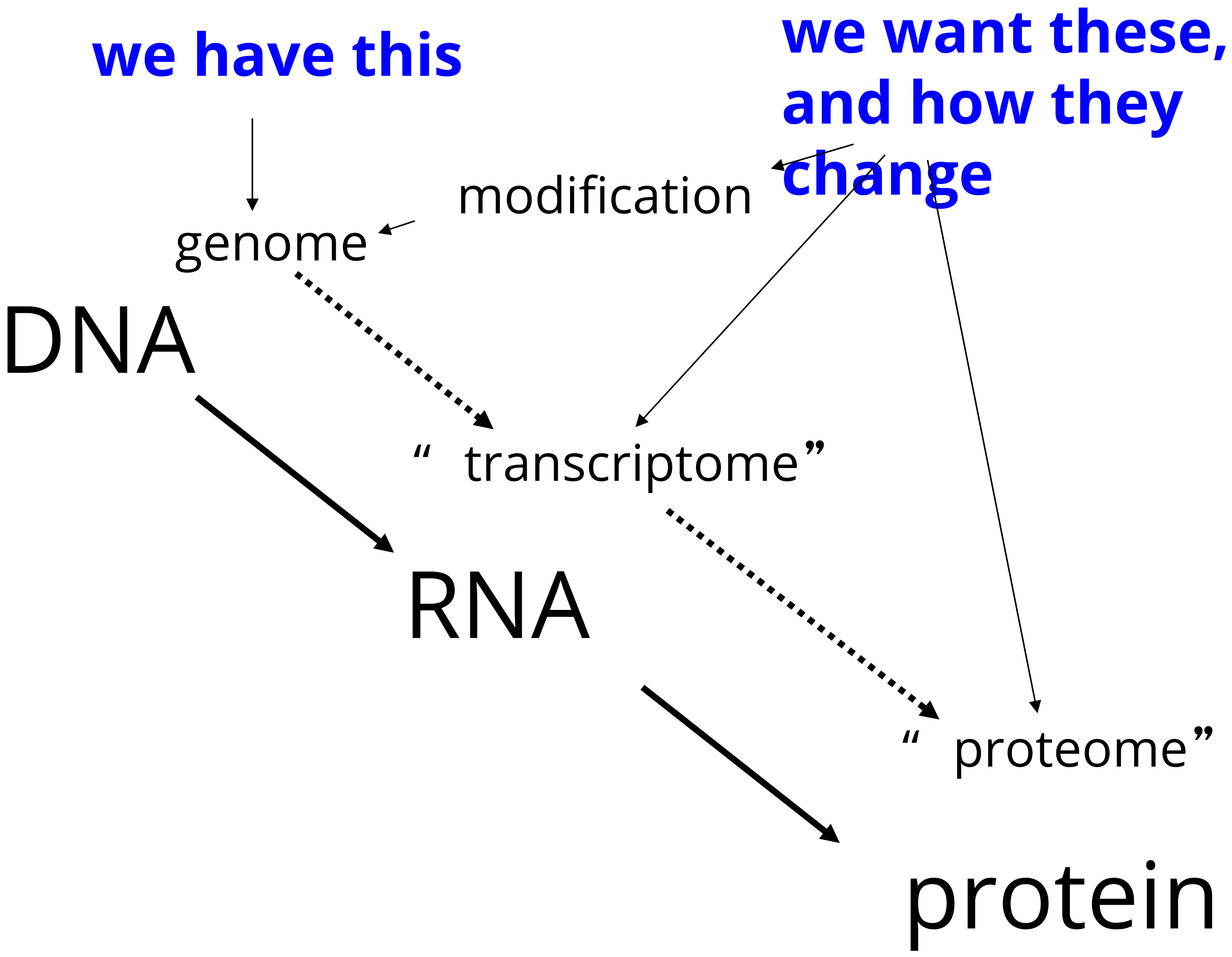
The totality of ____.

Genome (all the genes)
Transcriptome (all the RNAs)
Proteome (all the proteins)
Methylome (all the sites of DNA methylation, epigenetic modifications)

Omeome (all the –omes)
http://www.genomicglossaries.com/content/omes.asp

Why study the –omes?

To understand a complex system, we need to know what all the parts are, and what they are doing



we have this

we want these, and how they change

modification

genome

DNA

“ transcriptome”

RNA

“ proteome”

protein

Detection of mRNA transcripts

- Northern Blot – immobilize mRNA on membrane, detect specific sequence by hybridization with one labeled probe--requires a separate blotting for each probe

- DNA microarray – immobilize many probes (thousands) in an ordered array, hybridize (base pair) with labelled DNA

- RNA-seq – isolate RNA, reverse transcribe to make DNA, ‘next-gen’ sequencing

The value of DNA microarray/RNA-seq for studying gene expression

1) Measure the levels of all RNA transcripts at same time

2) RNA abundance usually determines the level of gene expression – a lot of gene control occurs at the level of transcription

3) Changes in transcription patterns correlate with changing environment – overarching patterns can be detected by microarray/RNA-seq, and may suggest new biological mechanisms

DNA microarray: an array of probes

Identify protein coding genes (from open reading frames in the genome), then…

- PCR each gene, attach each PCR product to a solid support in a specific order

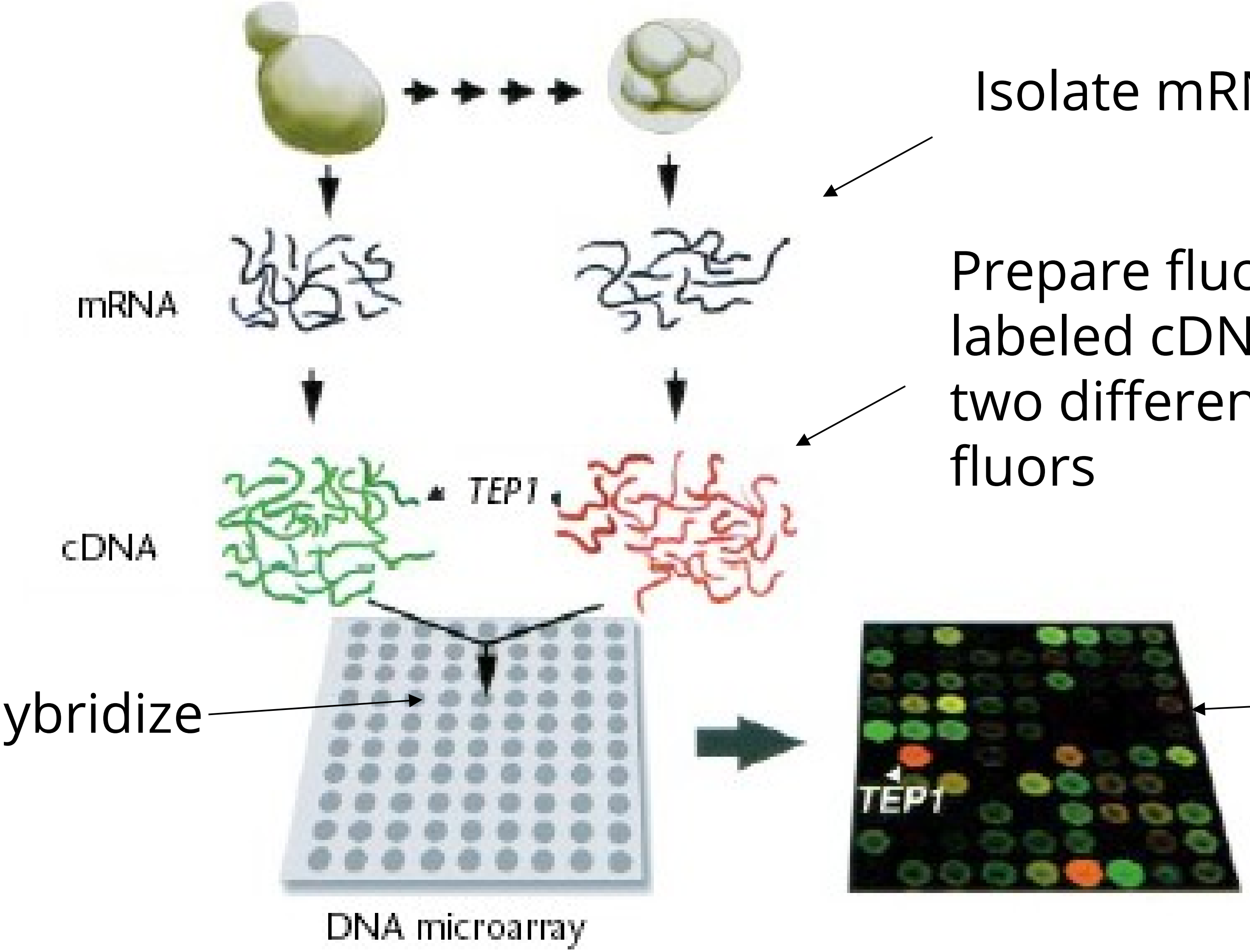
or

- Chemically synthesize gene-specific oligonucleotide probes directly on microchip

then

- Hybridize **labelled** RNA to the chip. More hyb. signal at a specific spot means more of that RNA

Microarray: genes up-regulated during meiosis?



vegetative

sporulating

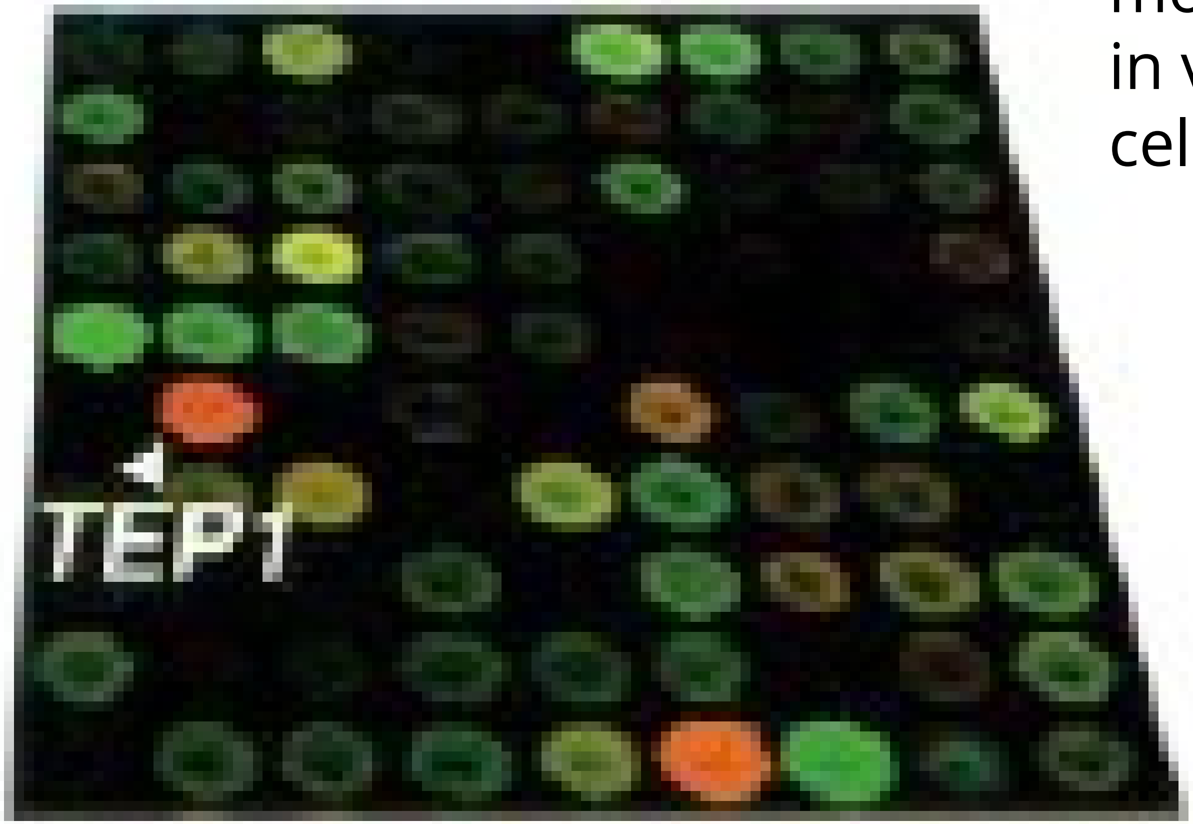
Isolate mRNA

Prepare fluorescently labeled cDNA with two different-colored fluors

hybridize

read-out

Example microarray data



Green: mRNA more abundant in vegetative cells

Yellow: equivalent mRNA abundance in vegetative and sporulating cells

Red: mRNA more abundant in sporulating cells

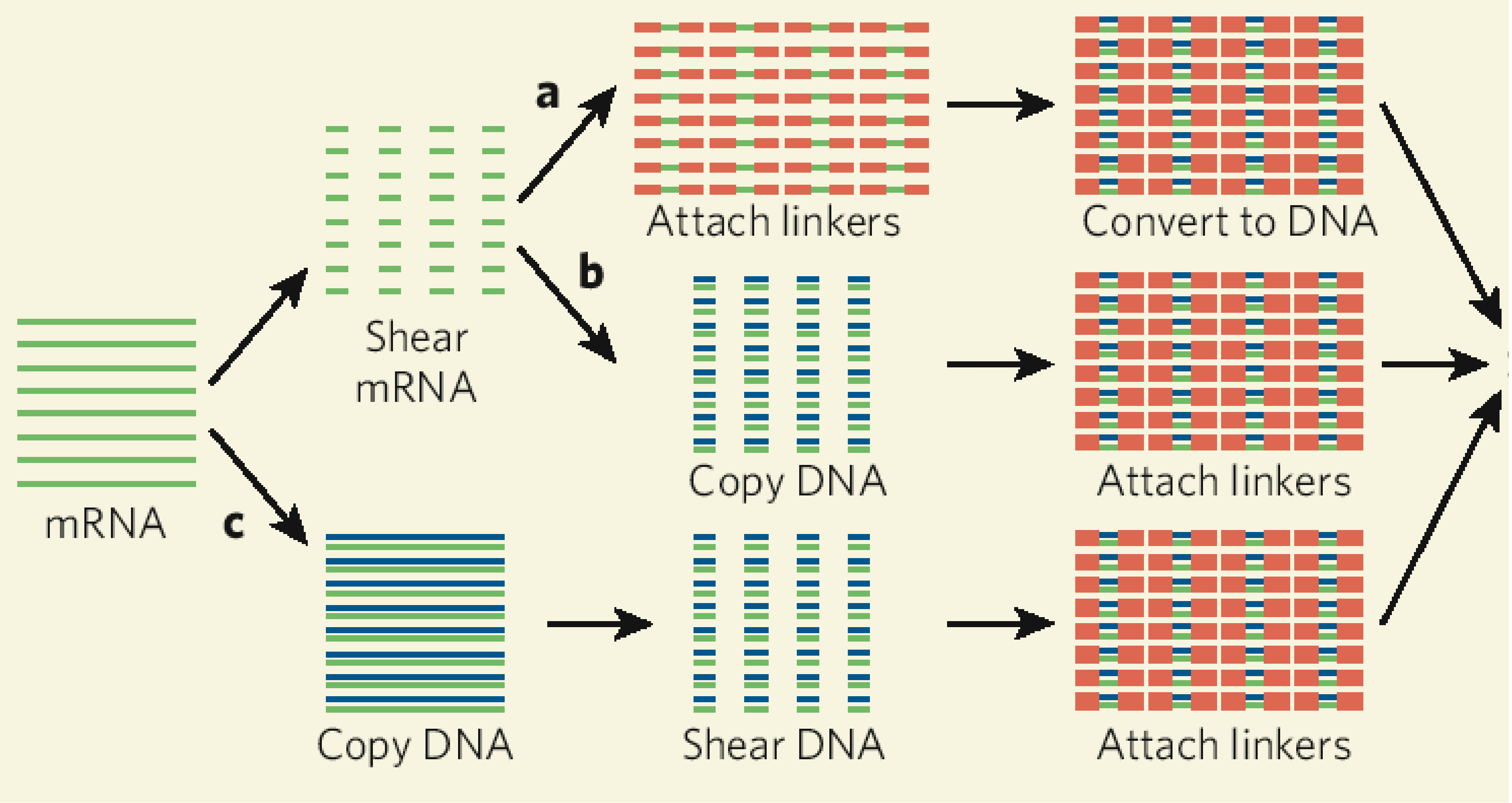
RNA-seq: sequence cDNA directly

1) Isolate RNA
2) Make cDNA
3) Sequence the cDNA with next generation methods
4) The more abundant the mRNA, the more sequence data

Advantages:

- Quantitative, high sensitivity, very low background
- No need to make an array
- Direct identification of the RNA being made
  - Gives info about splicing variants, 5’ and 3’ ends
  - (no need for hybridization, which doesn’t give RNA sequence information)
- Sequencing costs continue to drop

RNA-seq: which RNAs are expressed, and how much?



Graveley, “Power Sequencing”, p 1197 NATURE|Vol 453|26 June 2008

What to do with data for 1000s of genes?

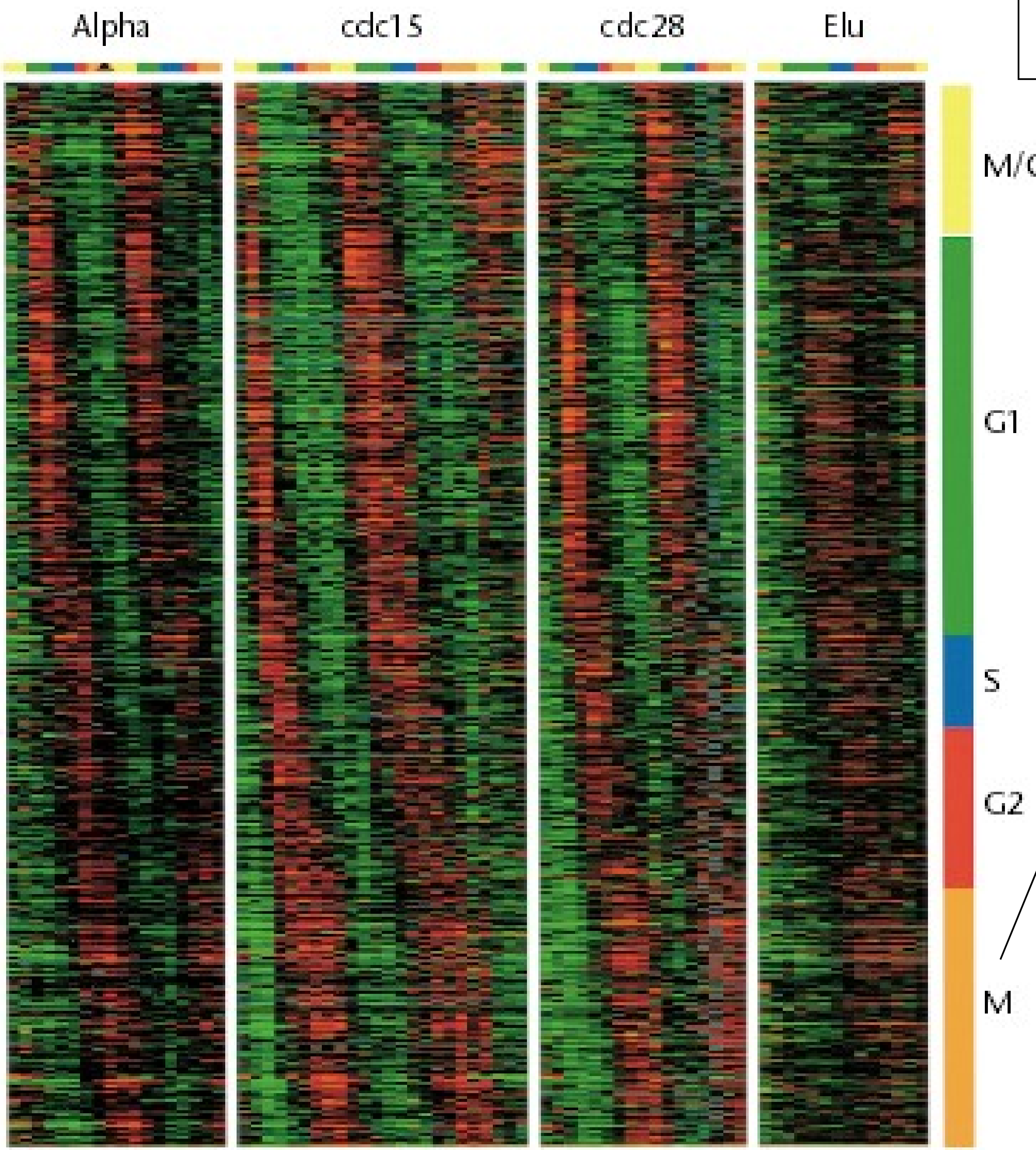
1)Organize data by clustering to see if patterns emerge

2)Display data graphically to assist in understanding and hypothesis generation

All cell cycle-regulated genes

Cell synchronization method

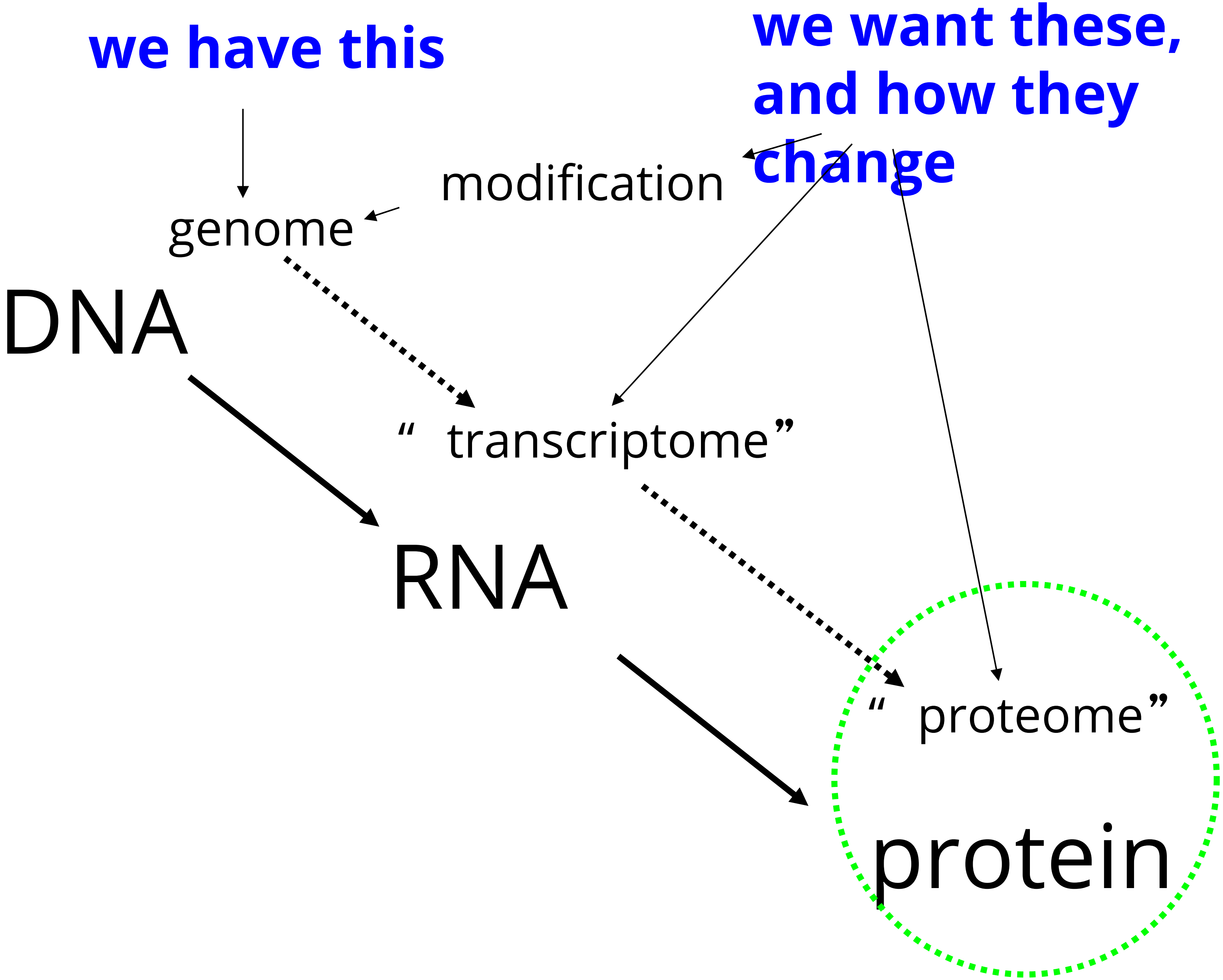
Each individual gene that is cell-cycle regulated →



phase in which each gene was expressed at hig levels

High mRNA levels 🟥

low mRNA levels 🟩



we have this

we want these, and how they change

modification

genome

DNA

" transcriptome"

RNA

" proteome"

protein

Analysis of the proteome: “ proteomics”

- Which proteins are present and when?
- What are the proteins doing?
  - What interacts with what?
    - Protein-DNA interactions (chromatin immunoprecipitation)
    - Protein-protein interactions
  - What is the function of each protein?

Phizicky et al. (2003) “ Protein analysis on a proteomic scale” *Nature* **422**, p. 208-215

How to detect protein expression

- Antibodies to specific proteins (those antibodies need to be available first)

- Specific label on protein *in vivo*, for example GFP to reveal expression in different tissues or subcellular locations

- Specific assay for activity: assuming you have a simple assay already designed

Above methods arduous for whole proteome

- Mass spectrometry for direct ID and quantitation

Defining the human proteome by immunodetection
http://www.proteinatlas.org/

- ~26,000 antibodies to human proteins, targeting ~17,000 different proteins

- 44 major tissues and organs, >13 million tissue-based immunohistochemistry images

- Complemented with RNA-seq analysis of the tissues and organs (to confirm RNAs for detected proteins)

- Cellular proteomes: https://www.proteinatlas.org/humanproteome/cell

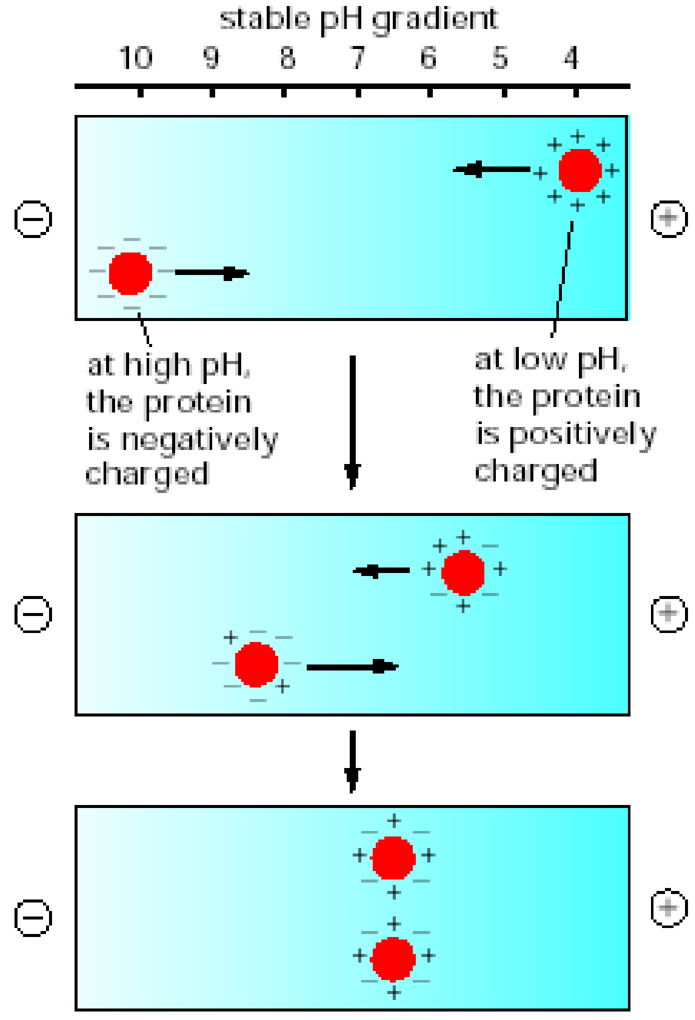
- Cancer proteomes: https://www.proteinatlas.org/humanproteome/pathology

Simultaneous detection and identification of all (or most) proteins

- 2D gel electrophoresis
  - Separate proteins in a given organism or tissue type by migration in gel electrophoresis
  - Identify protein (cut out of gel, sequence or mass-spec)
  - Pattern of spots like a barcode for hi-throughput studies
- Mass spectrometry
  - Separate individual proteins from cell by charge and mass, individual proteins can be identified (need genome sequence information for this)
- Microarray/seq analysis: identify all the DNA or RNA that is bound by a protein

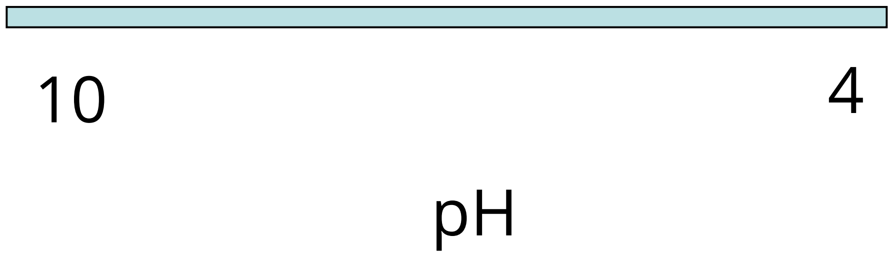
2D gel electrophoresis

1) Separate proteins by isoelectric point


stable pH gradient
10
9
8
7
6
5
4
at high pH, the protein is negatively charged
at low pH, the protein is positively charged

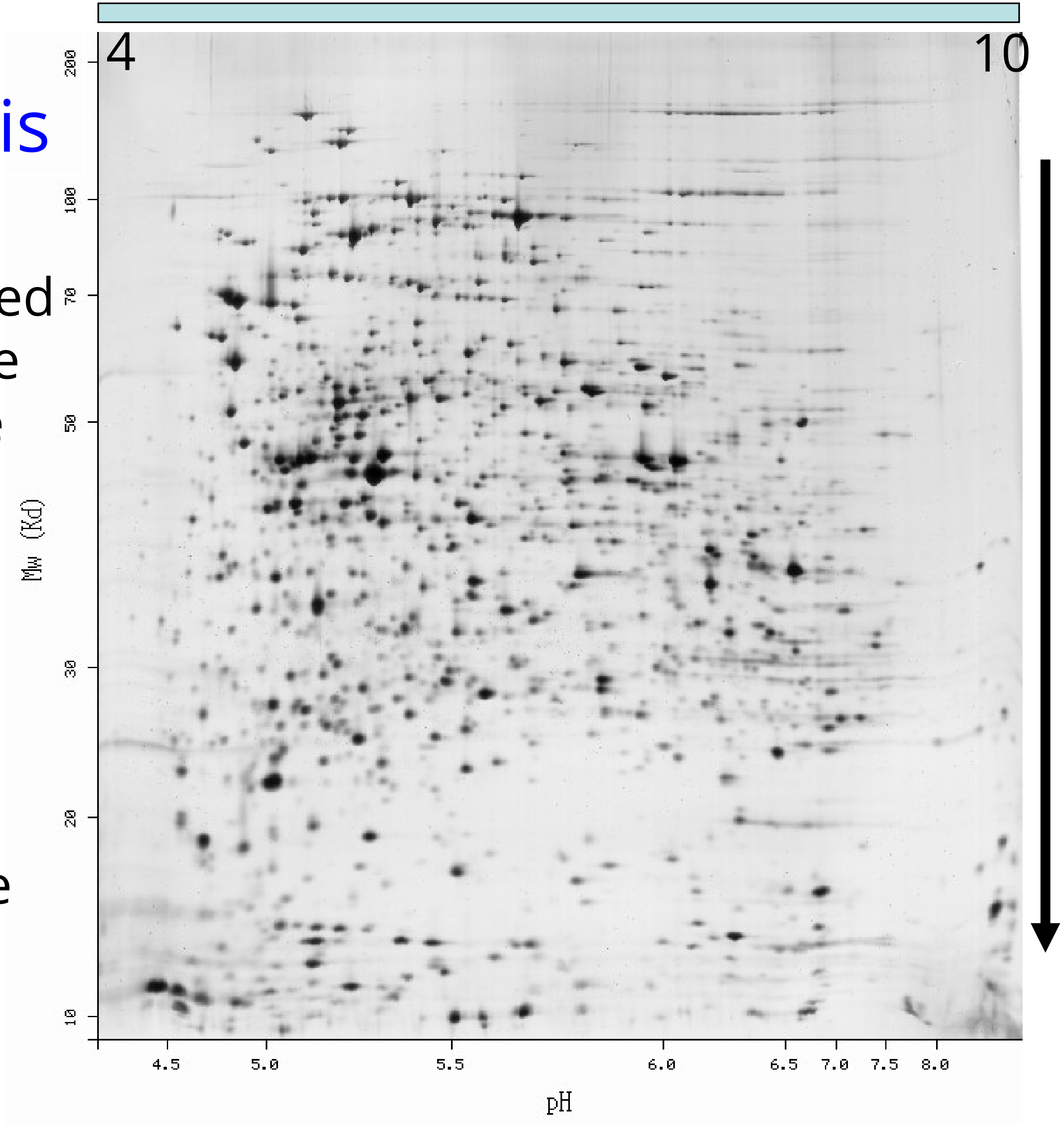
The protein shown here has an isoelectric pH of 6.5.

Use a long, narrow gel



2-D gel electrophoresis

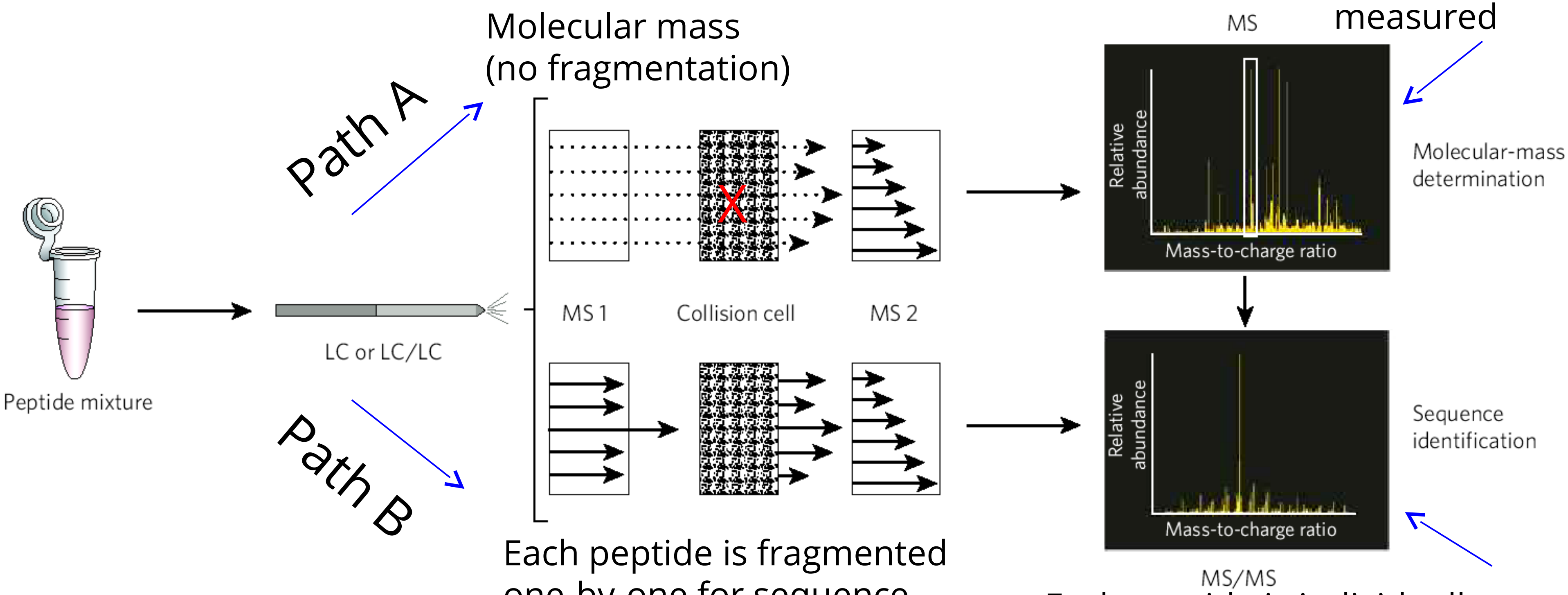
Lay gel containing isoelectrically focused protein on SDS page gel, separate on the basis of size



E.coli protein profile
From swissprot database,
www.expasy.ch

Mass spectrometry: identify the proteins from a complex mixture

Liquid chromatography followed by tandem mass spectrometry



Path A

Molecular mass (no fragmentation)

The total mass of each peptide is measured

Path B

Each peptide is fragmented one-by-one for sequence information

Each peptide is individually fragmented and sub-fragment masses are measured

From Cravatt *et al.* (2007) “ The biological impact of mass-spectrometry-based proteomics.” *Nature* **450**, p. 991.

How protein function gets defined

Classical methods define activity of protein, develop an assay for activity

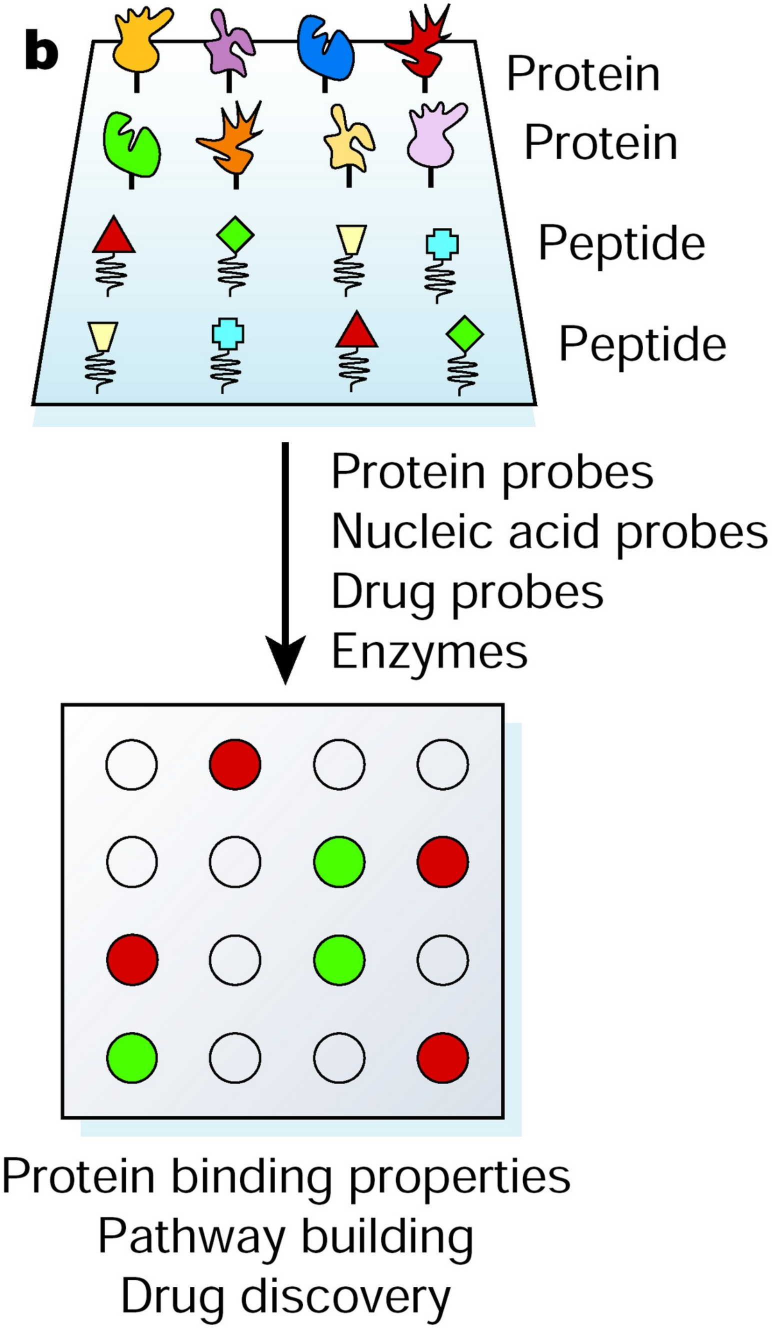
- Biochemistry: use a specific assay to purify a protein or protein complex from a cell, find out the structure and function of the protein *in vitro*

- Genetics: find mutant versions of a protein that have altered or lost activity, observe the phenotype of the organism with that mutation, obtain additional mutant genes that may interact with protein of interest, etc.

Protein structure and function analysis from a proteomic approach

- Massively parallel screens for activity--protein arrays
- Protein-DNA interactions: identifying binding sites for DNA-binding proteins, study regulation of gene expression
- (Structural genomics: solve structures of as many open reading frame peptides as possible)

Protein arrays for function

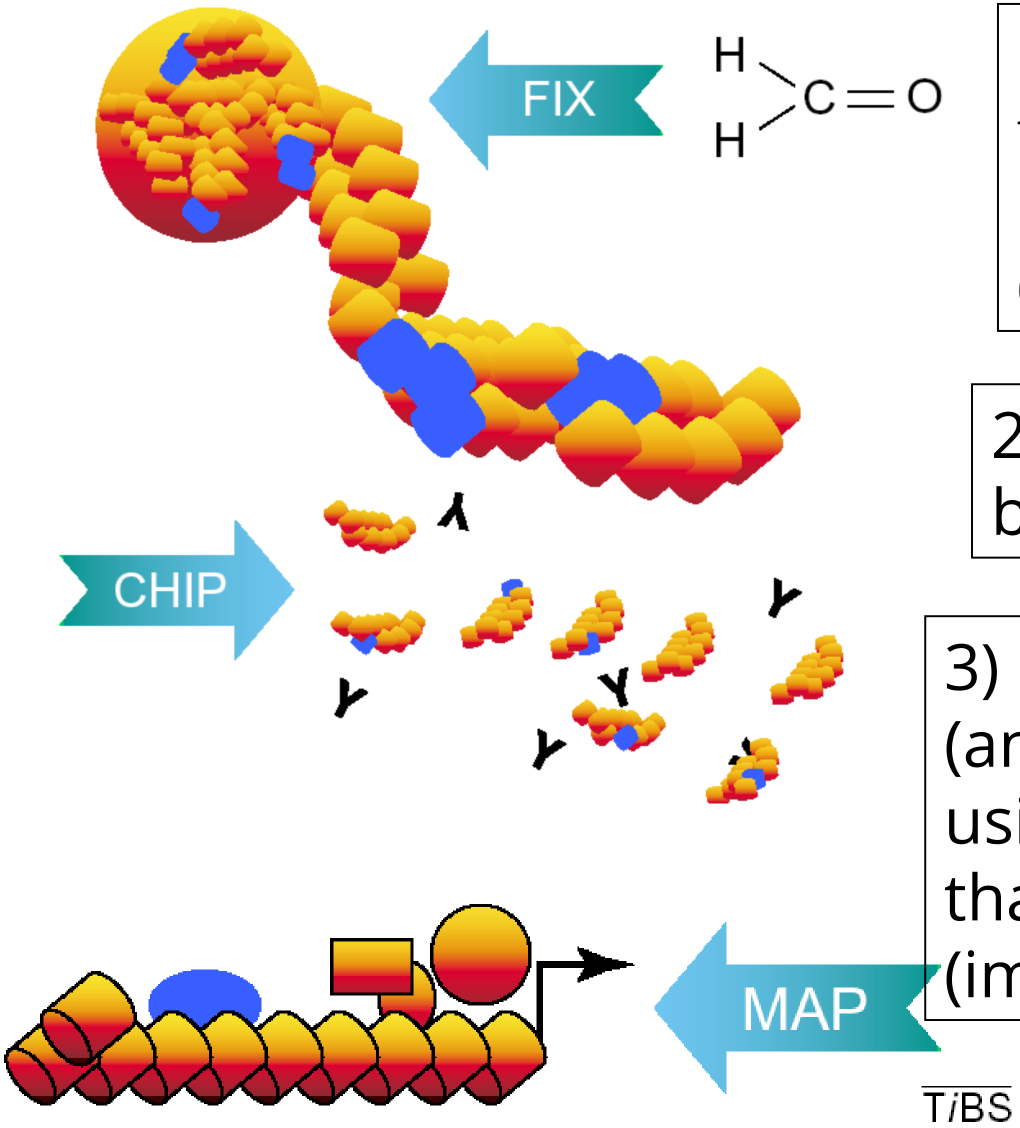


Proteins immobilized, usually by virtue of a tag sequence (6 x histidine tag, biotin, etc.)

Probe all proteins at once for a specific activity

Structural diversity and complexity of proteins means not all proteins are active in this form

" Chromatin ImmunoPrecipitation"   (ChIP)



1) Grow cells, add formaldehyde to cross-link everything to everything (including DNA to protein)

2) Lyse cells, break up DNA by shearing

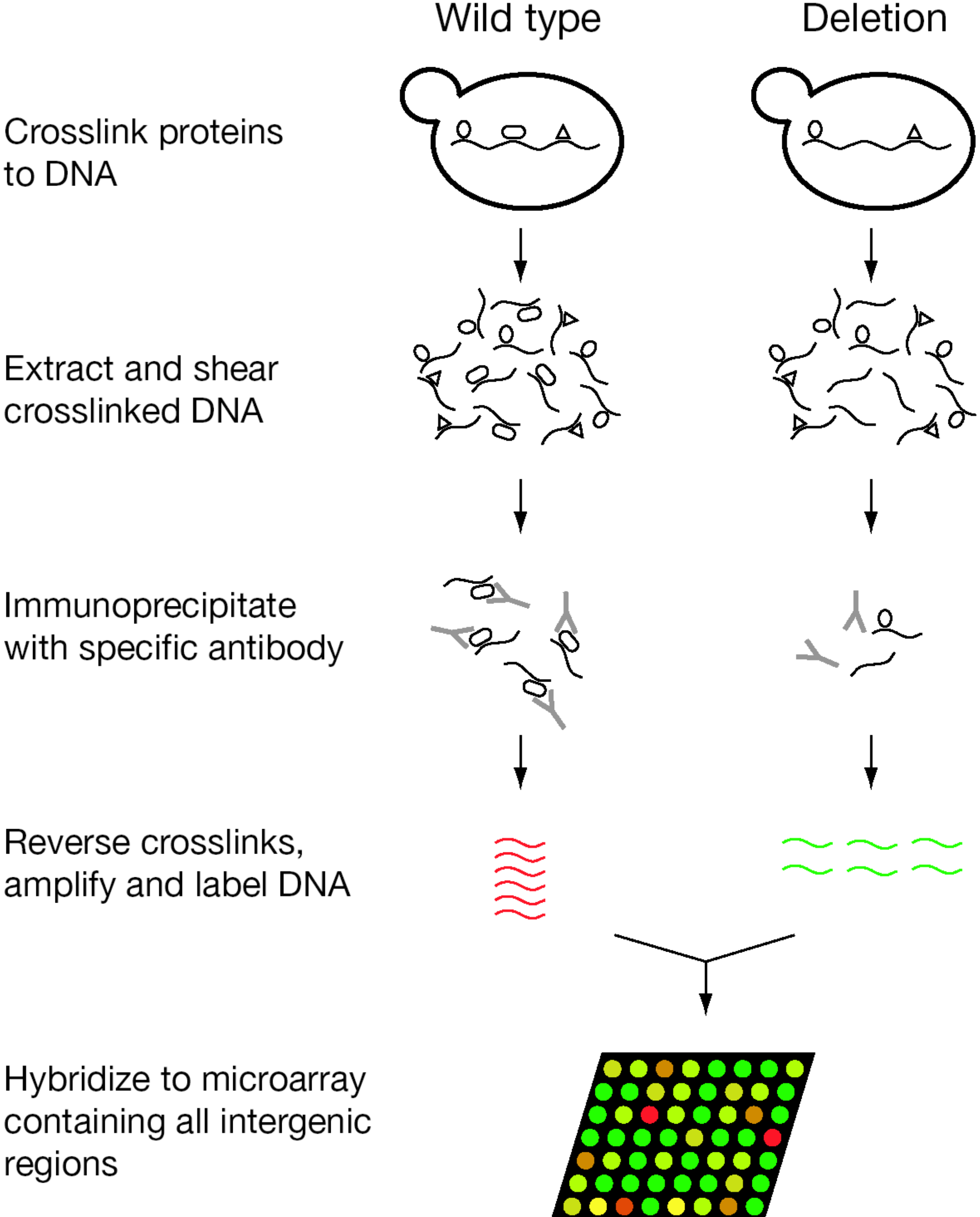
3) Retrieve protein of interest (and the DNA it is bound to) using specific antibody to that protein (immunoprecipitation)

4) Determine presence of DNA by quantitative PCR

V. Orlando (2000) *TIBS* **25**, p. 99

Genome-wide ChIP

PCR, label with fluorescent dyes



Crosslink proteins to DNA

Extract and shear crosslinked DNA

Immunoprecipitate with specific antibody

Reverse crosslinks, amplify and label DNA

Hybridize to microarray containing all intergenic regions

ChIP-seq

Sequence the IP DNA using " next generation" sequencing techniques

- High resolution (single base)
- Low noise
- Coverage not limited by array

APPLICATIONS OF NEXT-GENERATION SEQUENCING

ChIP–seq: advantages and challenges of a maturing technology

Peter J. Park

NATURE REVIEWS | GENETICS VOLUME 10 | OCTOBER 2009 | 669

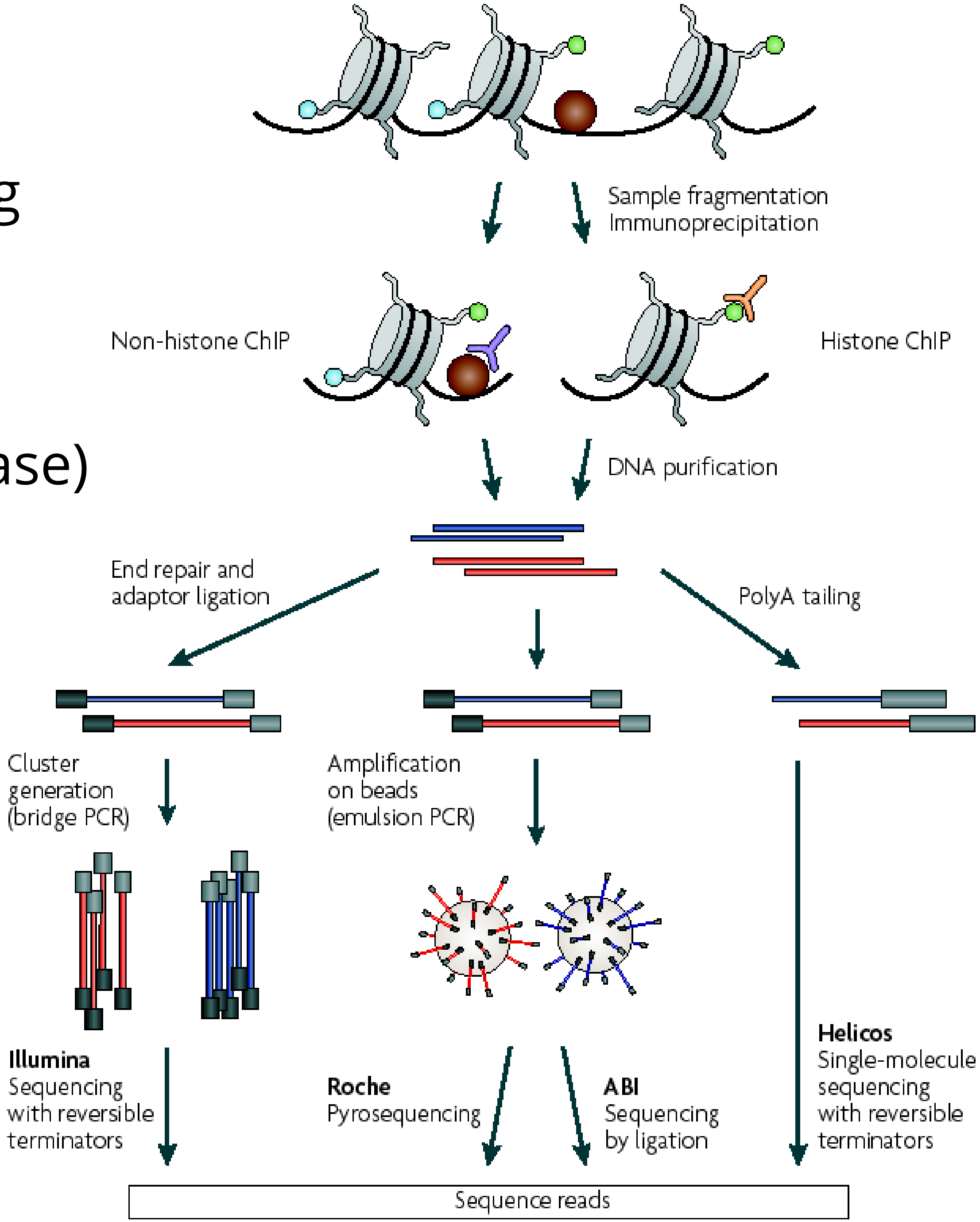
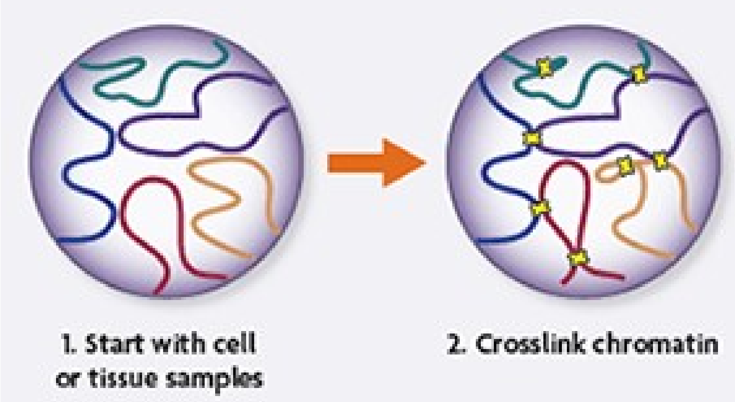
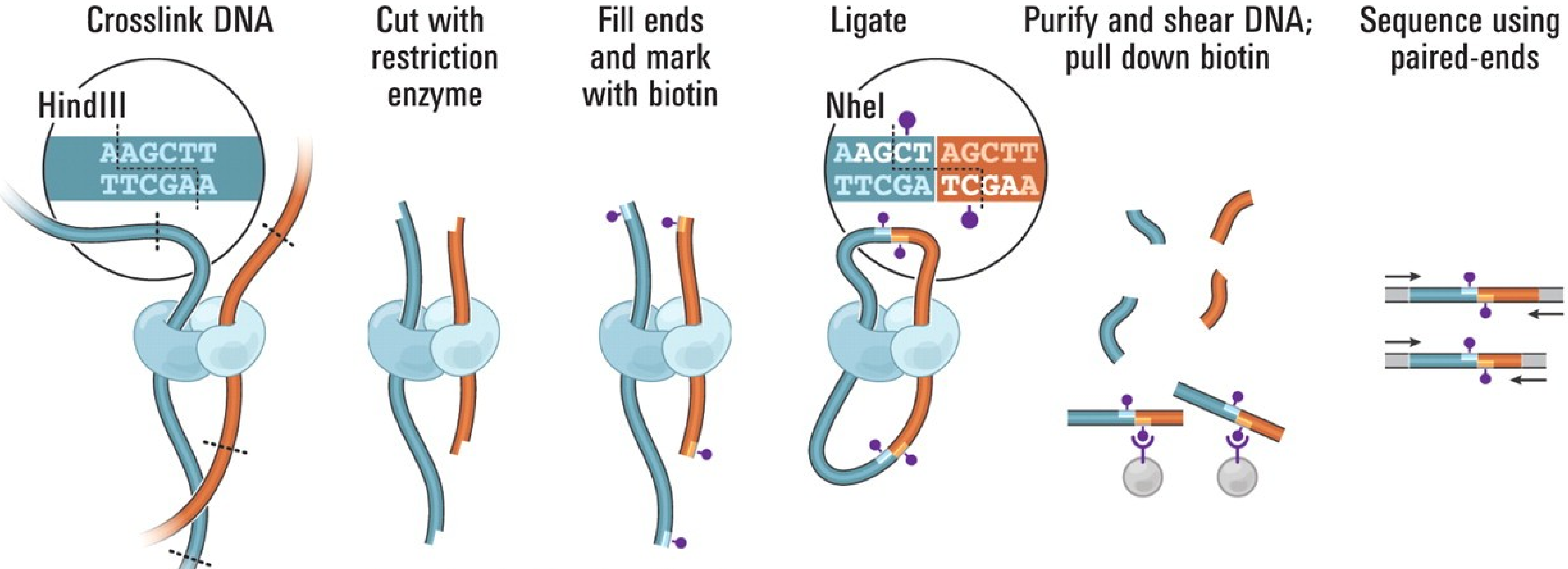


Figure 1 | **Overview of a ChIP–seq experiment.** Using chromatin immunoprecipitation

Chromatin conformation: which regions of DNA are close to each other

Hi-C: proximity based ligation, followed by massively parallel sequencing




Crosslink DNA
HindIII
AAGCTT
TTCGAA
Cut with restriction enzyme
Fill ends and mark with biotin
Ligate
NheI
AAGCT AGCTT
TTCGA TCGAA
Purify and shear DNA; pull down biotin
Sequence using paired-ends

Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome

Erez Lieberman-Aiden,[1,2,3,4]* Nynke L. van Berkum,[5]* Louise Williams,[1] Maxim Imakaev,[2] Tobias Ragoczy,[6,7] Agnes Telling,[6,7] Ido Amit,[1] Bryan R. Lajoie,[5] Peter J. Sabo,[8] Michael O. Dorschner,[8] Richard Sandstrom,[8] Bradley Bernstein,[1,9] M. A. Bender,[10] Mark Groudine,[6,7] Andreas Gnirke,[1] John Stamatoyannopoulos,[8] Leonid A. Mirny,[2,11] Eric S. Lander,[1,12,13]† Job Dekker[5]†

www.sciencemag.org SCIENCE VOL 326 9 OCTOBER 2009

The ENCODE project: putting it all together

ENCyclopedia Of functional DNA Elements in the human genome

https://www.encodeproject.org/

http://genome.ucsc.edu/ENCODE/

Systematic mapping of

- Regions that are transcribed
- Transcription factor binding sites in DNA
- RNA binding proteins
- Chromatin structure
- DNA modifications
- Histone modifications
- …and much much more

In 2019: >14,000 datasets

The ENCODE project

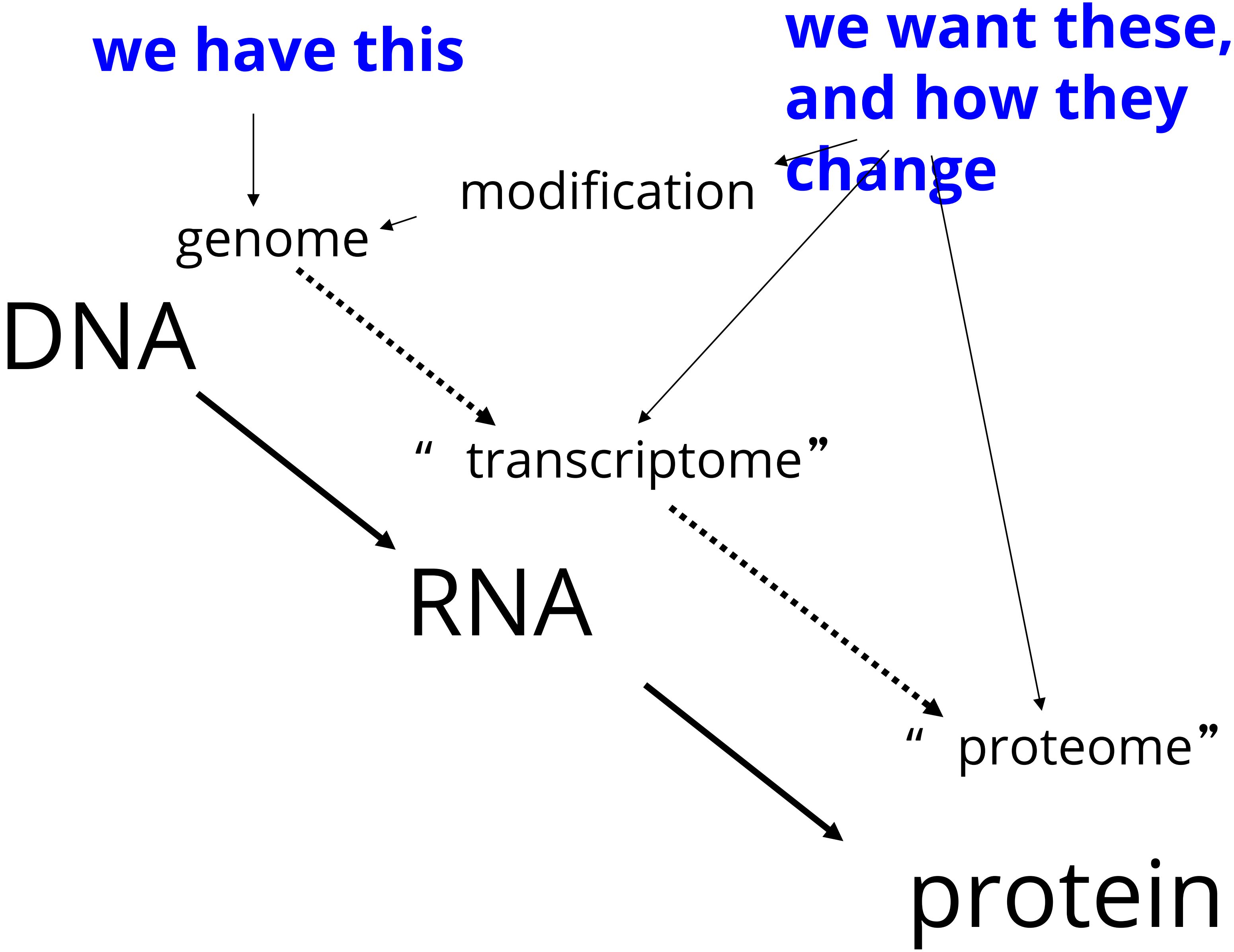
- ~80% of the human genome participates in an RNA or chromatin associated event, in at least one cell type

- RNA synthesis correlates with chromatin and transcription factor binding: promoters account for most RNA variation

- More disease associated Single Nucleotide Polymorphisms (SNPs) are in non-coding functional elements than in protein coding genes, e.g. affecting transcription factor binding

Not all cell types have been assayed
Not all transcription factors have been assayed

So there is <u>more to be done</u>



we have this

we want these, and how they change

modification

genome

DNA

“ transcriptome”

RNA

“ proteome”

protein

Face up to false positives

Macarthur (2012)
Nature **487** p.427

Dealing with very large data sets can suggest erroneous conclusions

1) Large data sets mean unusual (statistically rare, insignificant) events crop up often. Statistical analysis helps to assign significance (or lack of it)

2) Error/system bias often occurs in high-throughput methods – so the novice gets burned

Stringent quality control and standards are essential!