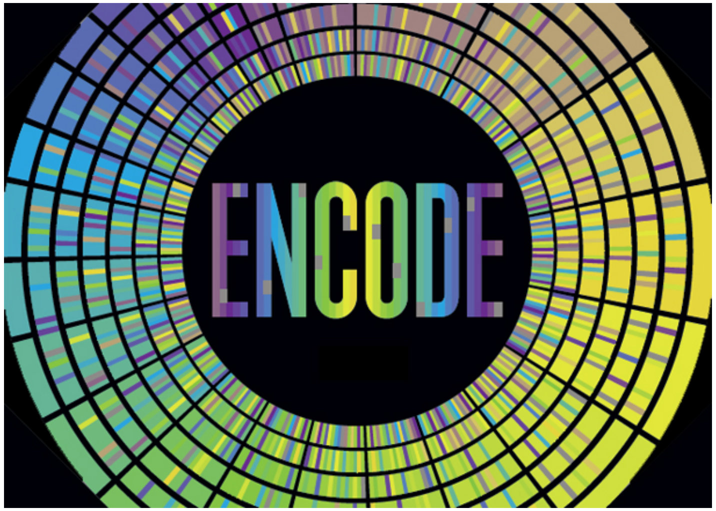
Module 4

Working with ENCODE data



Emily Perry

Ensembl Outreach Team

EMBL-EBI

http://www.sanger.ac.uk/resources/talksandtraining/opendoor/malawi.html

This session

- Introduction to ENCODE
- The ENCODE portal
- ENCODE data in UCSC (Jane)
- ENCODE data in Ensembl

Course materials

http://www.sanger.ac.uk/resources/talksandtraining/opendoor/malawi.html

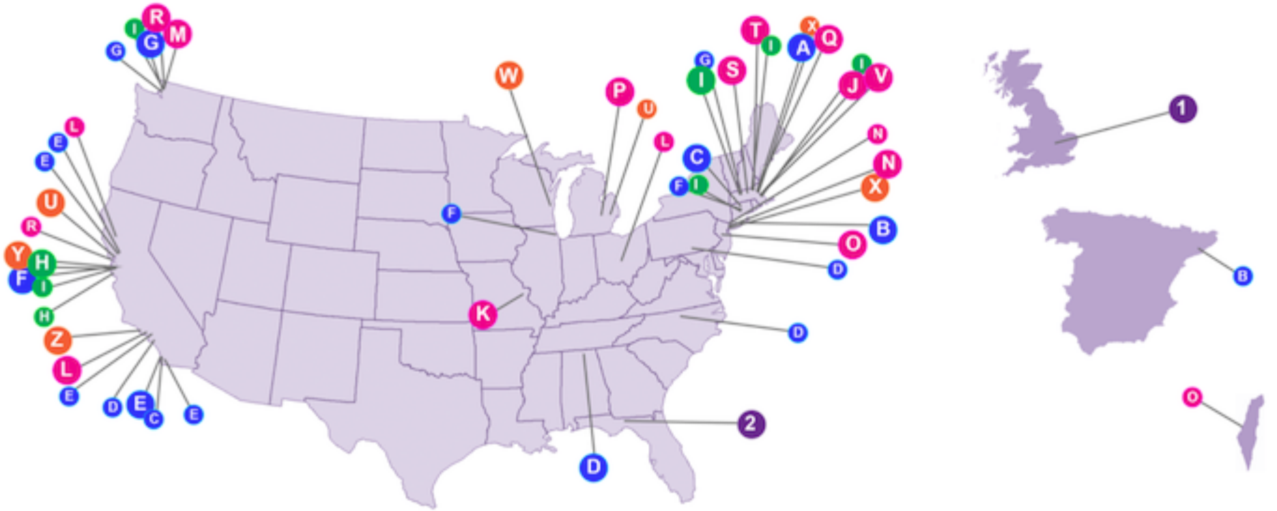
- Presentations
- Coursebook

- Paper coursebook page 107-134
- Exercise answers page 131-134

ENCODE project

- NHGRI launched a public research consortium named ENCODE, the Encyclopaedia Of DNA Elements, in September 2003.
  - Aim: “to identify all functional elements in the human genome sequence”.
- Implementation:
  - Pilot phase (Sept 2003- Sept 2007)
  - Technology development phase (Sept 2003- Sept 2007)
  - Scale up (Production) phase (Oct 2007 -)

Who are ENCODE?



Production Groups

- A Broad Institute
- B Cold Spring Harbor; Centre for Genomic Regulation (CRG);
- C University of Connecticut Health Center; UCSD
- D HudsonAlpha; Pennsylvania State; UC Irvine; Duke; Caltech
- E UCSD; Salk Institute ; Joint Genome Institute; Lawrence Berkeley National Laboratory; UCSD
- F Stanford; University of Chicago; Yale
- G University of Washington; Fred Hutchinson Cancer Research Center; University of Massachusetts Medical School

Data Coordination Center

- H Stanford; UCSC

Data Analysis Center

- I University of Massachusetts Medical School; Yale; MIT; Stanford; Harvard; University of Washington

Technology Development Groups

- J MIT
- K Washington University, St. Louis
- L USC; Ohio State University; UC, Davis
- M University of Washington
- N Sloan-Kettering; Weill Cornell Medical College
- O Princeton; Weizmann
- P University of Michigan
- Q Broad Institute
- R University of Washington; UCSF
- S Advanced RNA Technologies, LLC
- T Harvard

Computational Analysis Groups

- U Berkeley; Wayne State University
- V MIT
- W University of Wisconsin
- X Sloan-Kettering; Broad Institute
- Y Stanford
- Z UCLA

Affiliated Groups

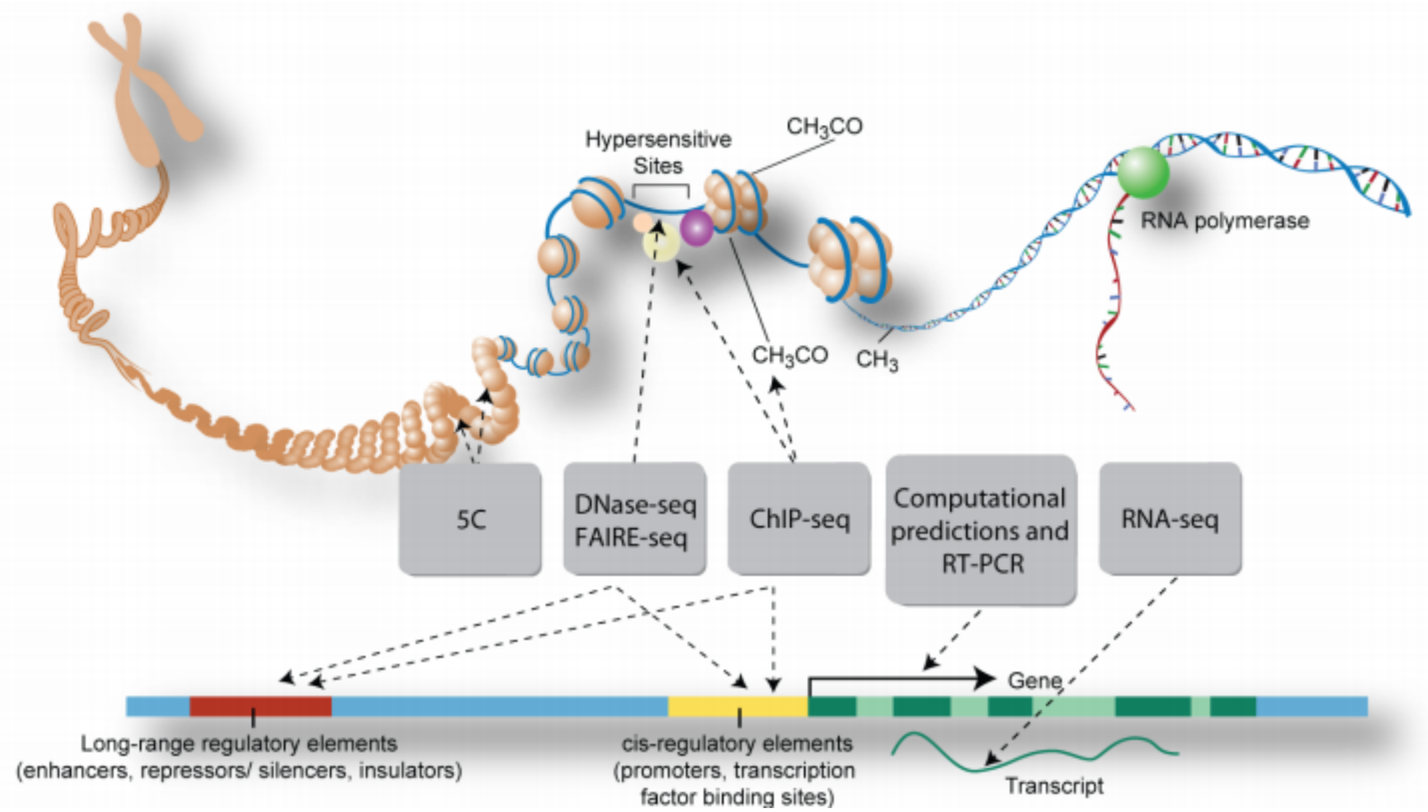
- 1 Wellcome Trust Sanger Institute
- 2 Florida State University

http://www.sanger.ac.uk/resources/talksandtraining/opendoor/malawi.html

Main cell types

Cell Type	Tier	Description	Source
GM12878	1	B-Lymphoblastoid cell line	Coriell GM12878
K562	1	Chronic Myelogenous / Erythroleukemia cell line	ATCC CCL-243
H1-hESC	1	Human Embryonic Stem Cells, line H1	Cellular Dynamics International
HepG2	2	Hepatoblastoma cell line	ATCC HB-8065
HeLa-S3	2	Cervical carcinoma cell line	ATCC CCL-2.2
HUVEC	2	Human Umbilical Vein Endothelial Cells	Lonza CC-2517
Various (Tier 3)	3	Various cell lines, cultured primary cells, and primary tissues	Various

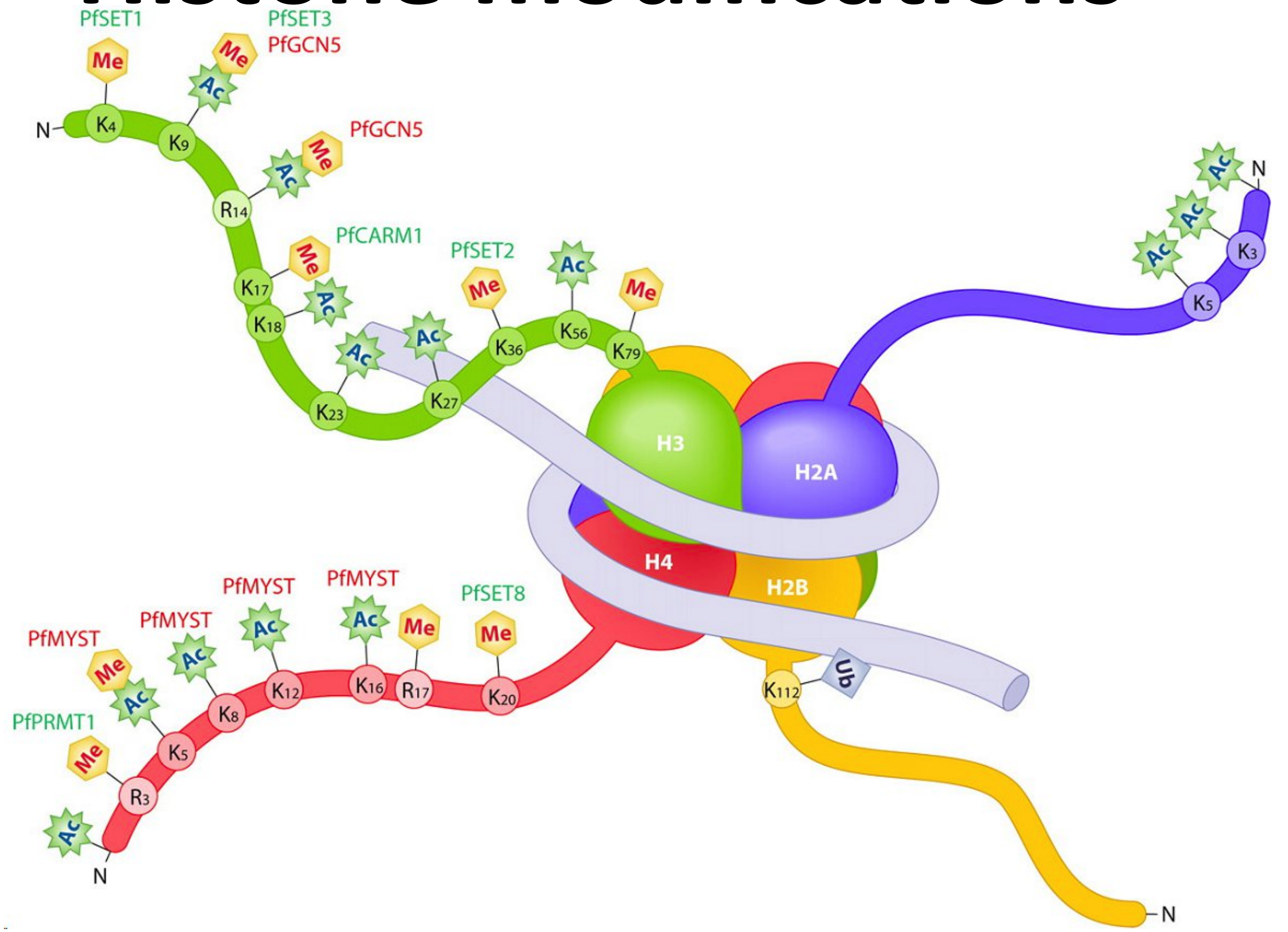
Experiments


Hypersensitive Sites
CH_3CO
RNA polymerase
CH_3CO
CH_3
5C
DNase-seq
FAIRE-seq
ChIP-seq
Computational predictions and RT-PCR
RNA-seq
Gene
Long-range regulatory elements
(enhancers, repressors/ silencers, insulators)
cis-regulatory elements
(promoters, transcription factor binding sites)
Transcript

http://www.sanger.ac.uk/resources/talksandtraining/opendoor/malawi.html

The Open Door Workshop

Histone modifications



We describe histone modifications using the form Subunit, Amino acid, Position, Modification, eg **H3K36me3**.

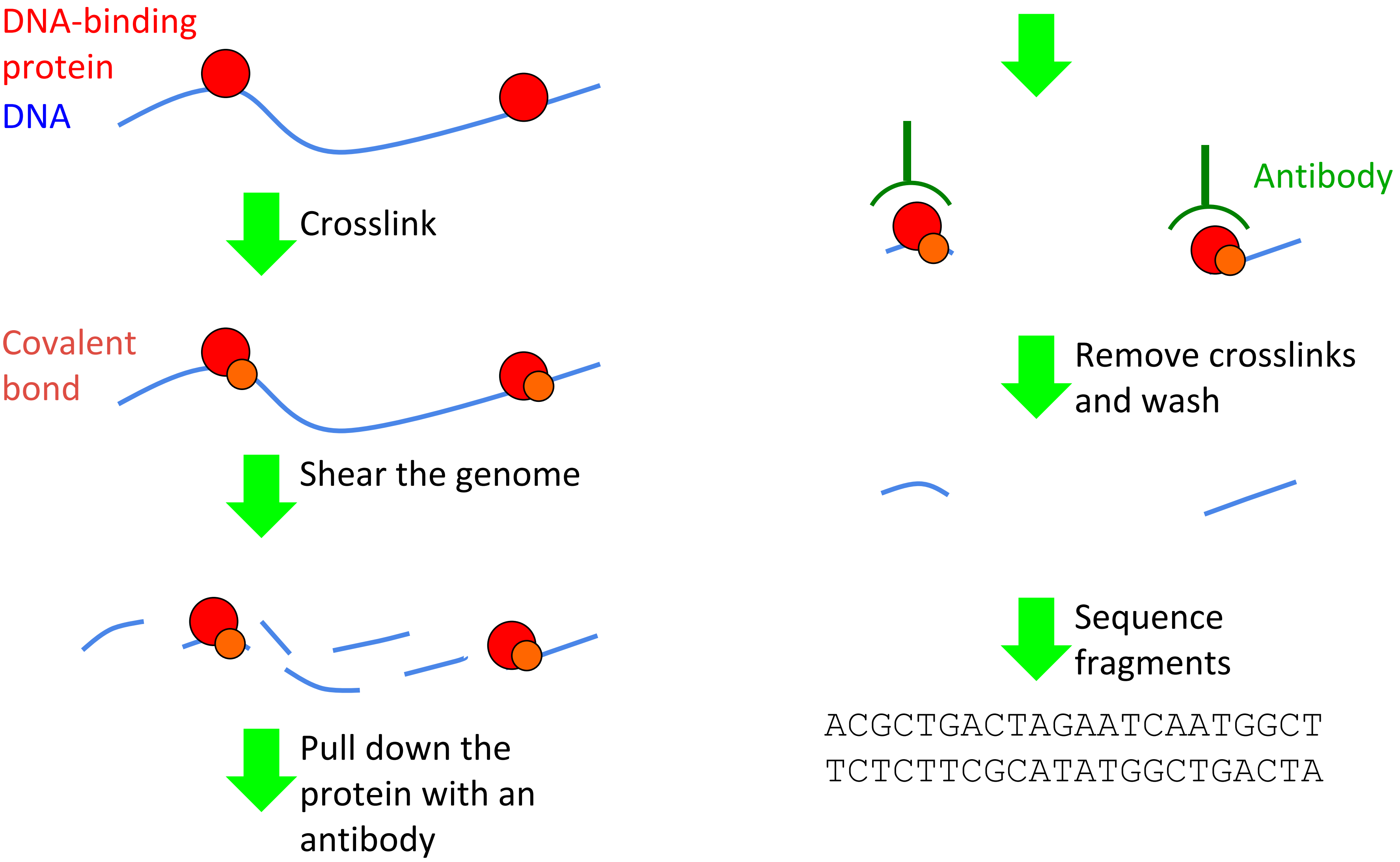
http://www.sanger.ac.uk/resources/talksandtraining/opendoor/malawi.html

Histone code

Modification	Histone						
	H3K4	H3K9	H3K14	H3K27	H3K79	H4K20	H2BK5
me1	🟩	🟩		🟩	🟩	🟩	🟩
me2	🟩	🟥		🟥	🟩		
me3	🟩	🟥		🟥	🟨		🟥
ac		🟩	🟩				

http://www.sanger.ac.uk/resources/talksandtraining/opendoor/malawi.html

ChIP-seq



Crosslink

Shear the genome

Pull down the protein with an antibody

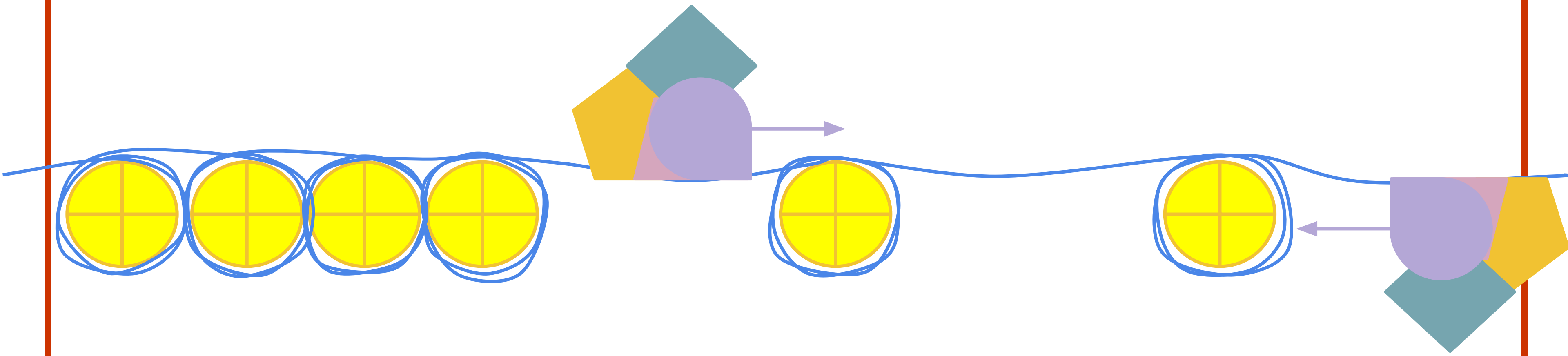
Remove crosslinks and wash

Sequence fragments

ACGCTGACTAGAATCAATGGCT
TCTCTTCGCATATGGCTGACTA

http://www.sanger.ac.uk/resources/talksandtraining/opendoor/malawi.html

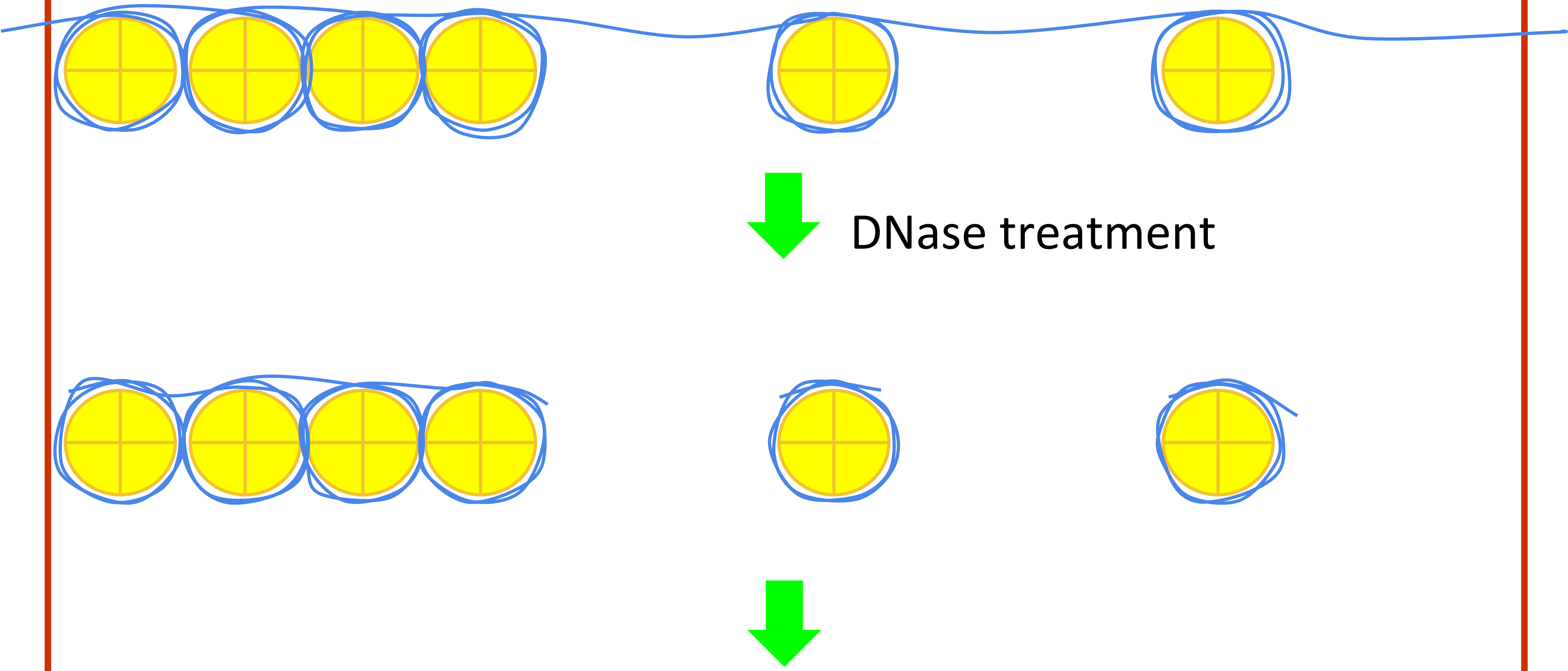
Open/closed chromatin



Open chromatin is transcriptionally active.
Closed chromatin is inactive.

http://www.sanger.ac.uk/resources/talksandtraining/opendoor/malawi.html

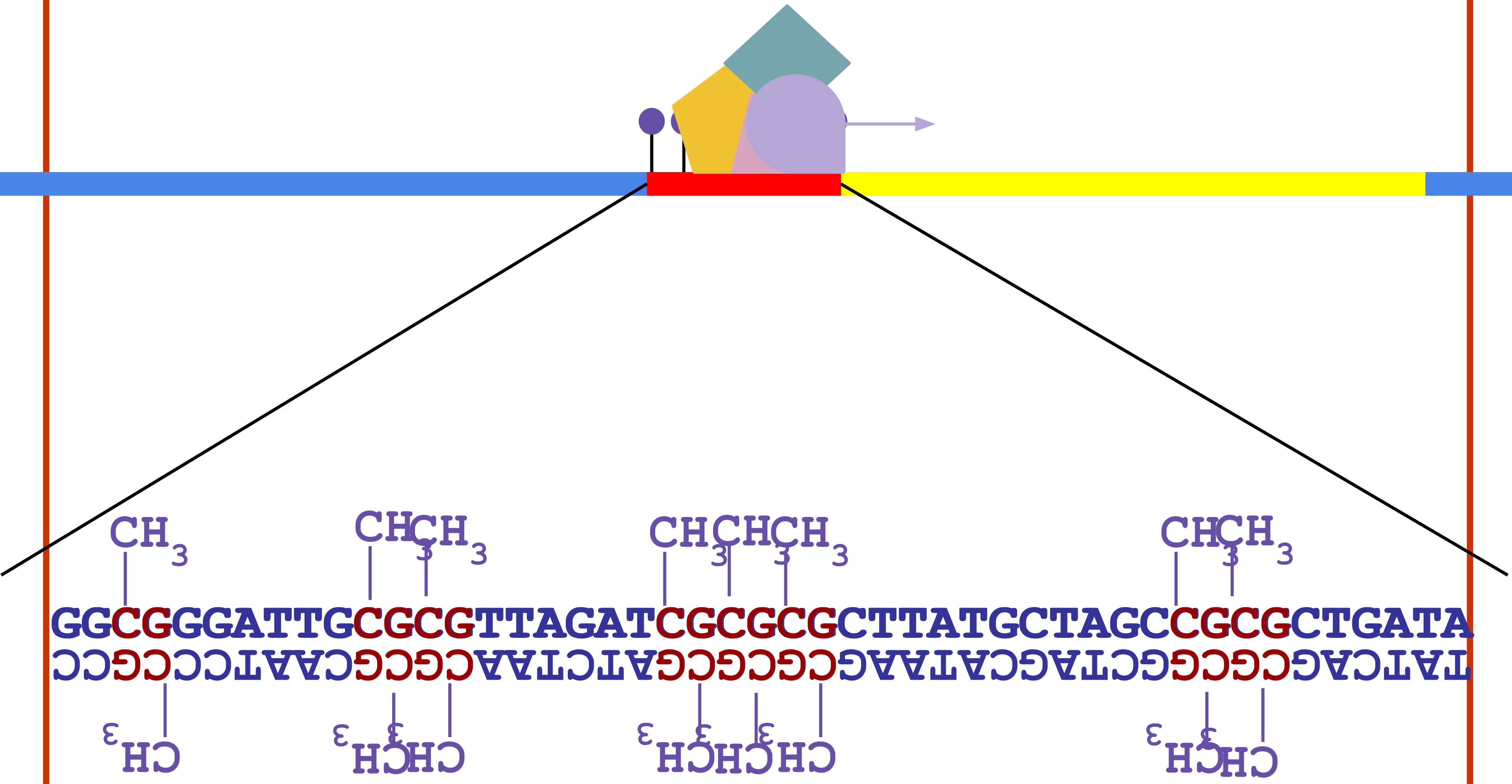
DNase hypersensitivity



DNase treatment

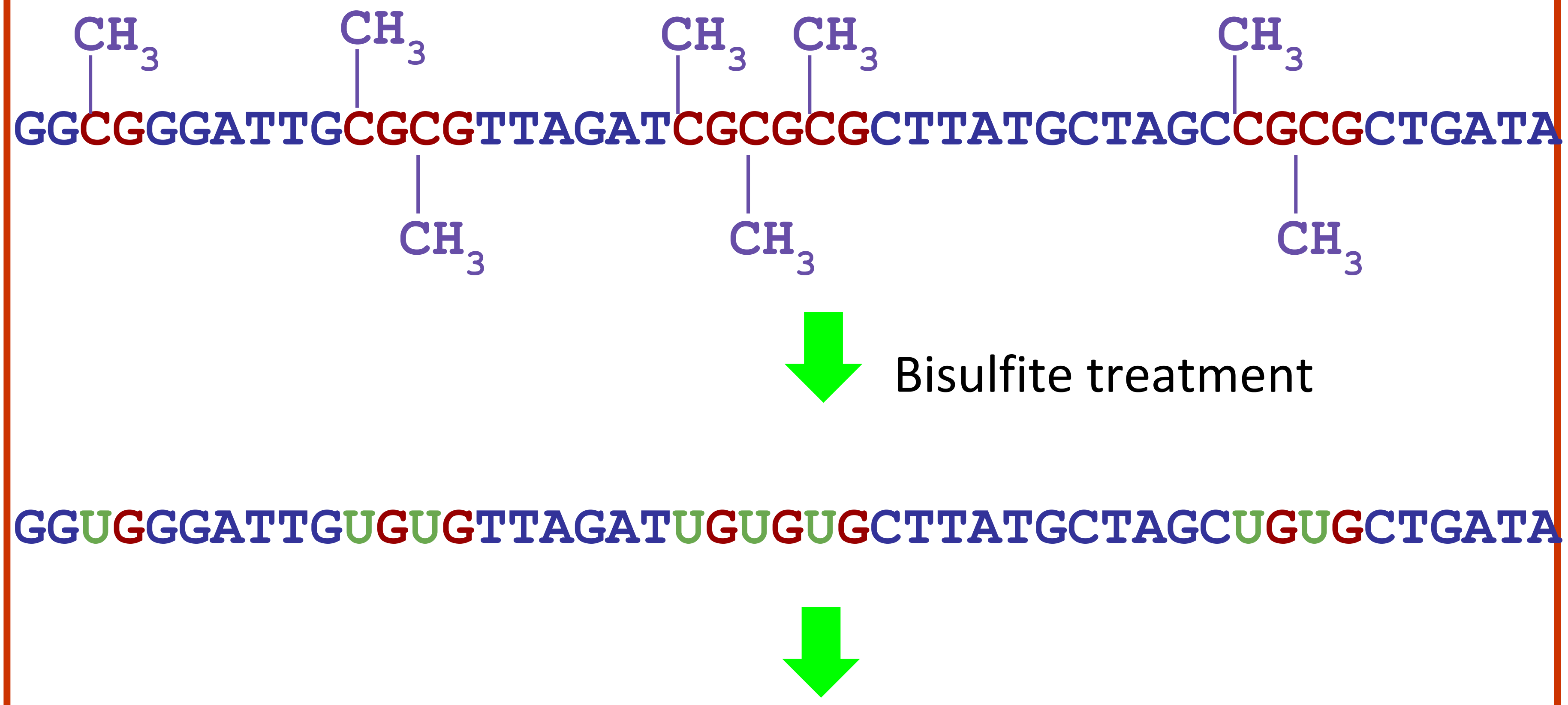
Sequence and compare to reference

DNA methylation -> inactive


CH3
CH3 CH3
CH3 CH3 CH3
CH3 CH3
GGCGGGATTGCGCGTTAGATCGCGCGCTTATGCTAGCCGCGCTGATA
TATCAGCGCGGCTAGCATAAGCGCGCGATCTAACGCGCAATCCCGCC
CH3
CH3 CH3
CH3 CH3 CH3
CH3 CH3

http://www.sanger.ac.uk/resources/talksandtraining/opendoor/malawi.html

Bisulfite sequencing



GGCGGGATTGCGCGTTAGATCGCGCGCTTATGCTAGCCGCGCTGATA

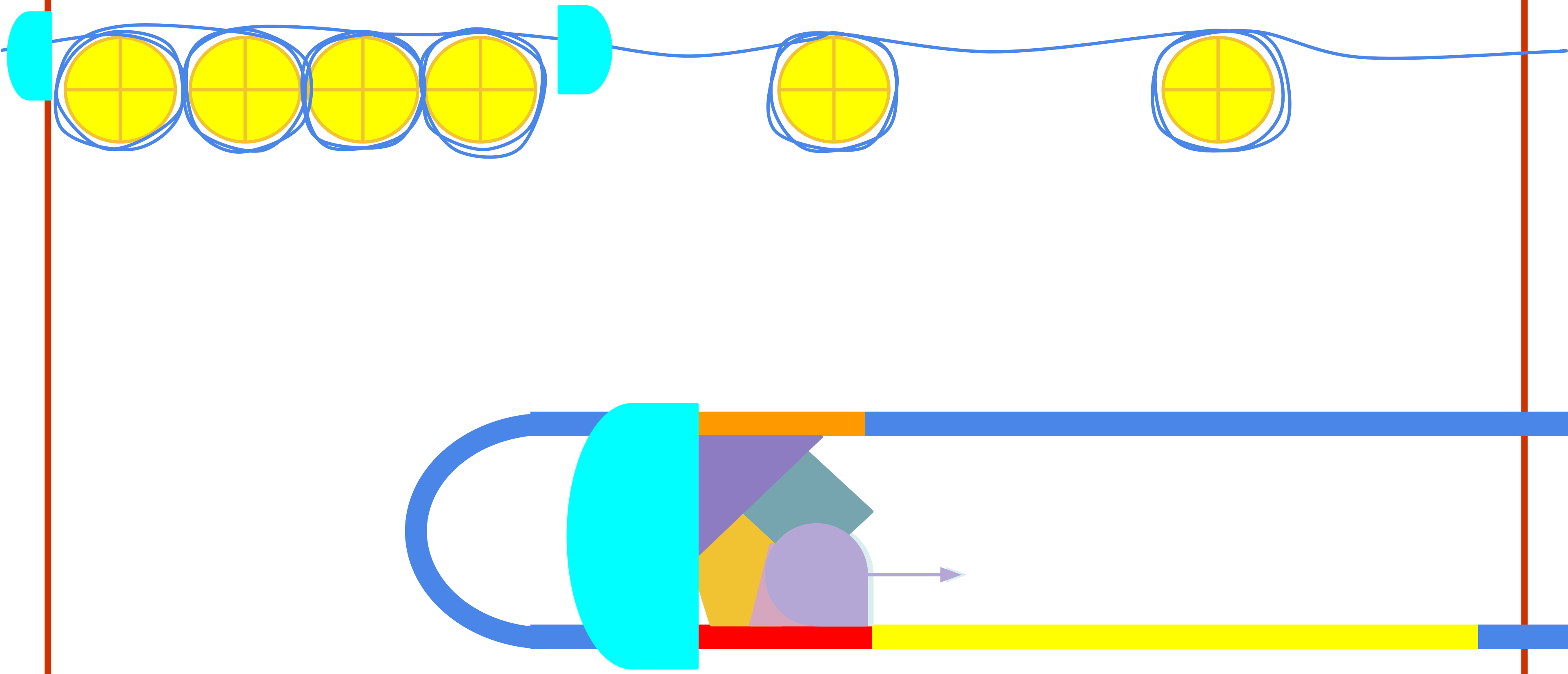
Bisulfite treatment

GGUGGGATTGUGUGTTAGATUGUGUGCTTATGCTAGCUGUGCTGATA

Sequence and compare to reference

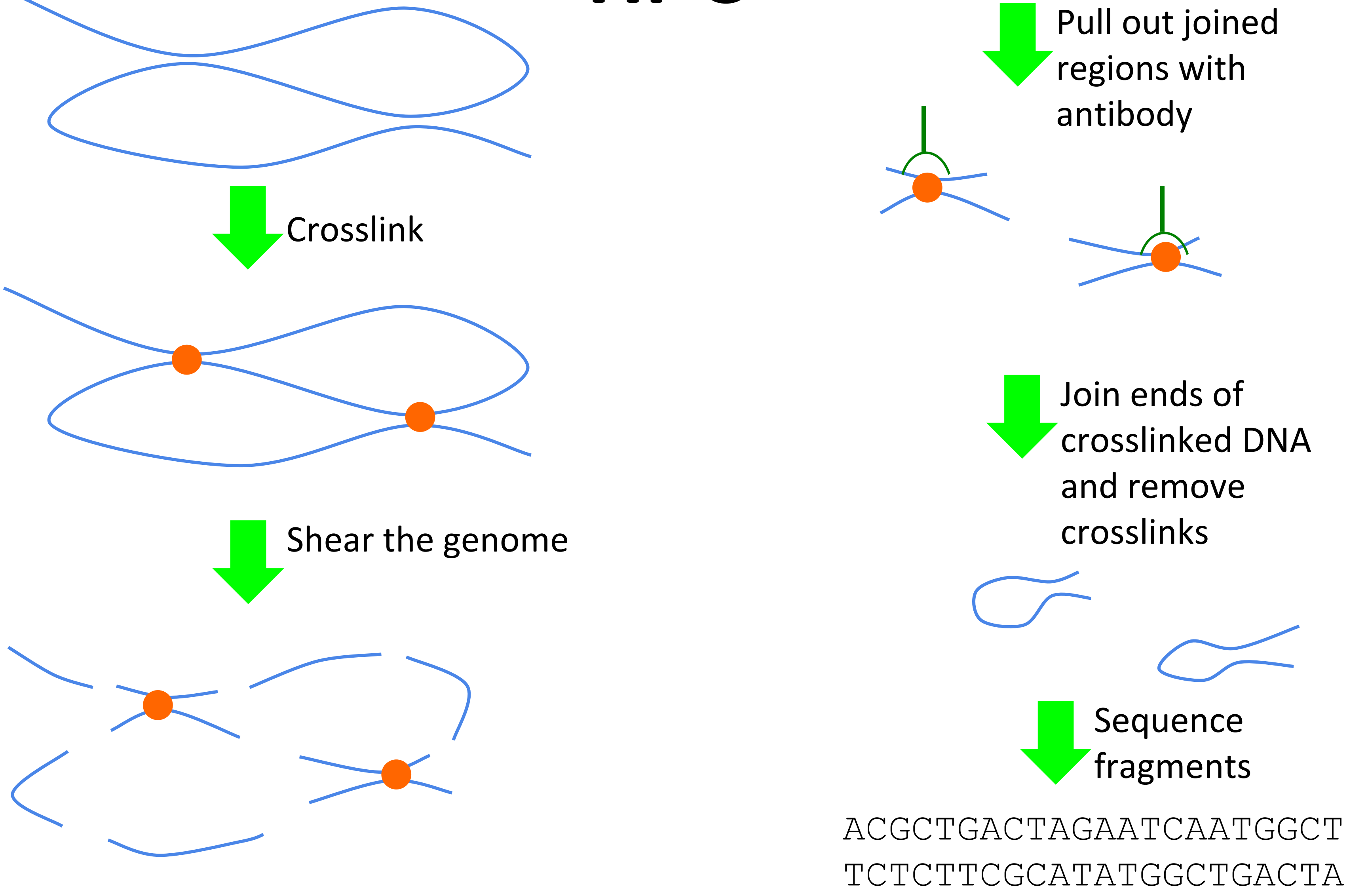
http://www.sanger.ac.uk/resources/talksandtraining/opendoor/malawi.html

3D interactions



http://www.sanger.ac.uk/resources/talksandtraining/opendoor/malawi.html

Hi-C



Crosslink

Shear the genome

Pull out joined regions with antibody

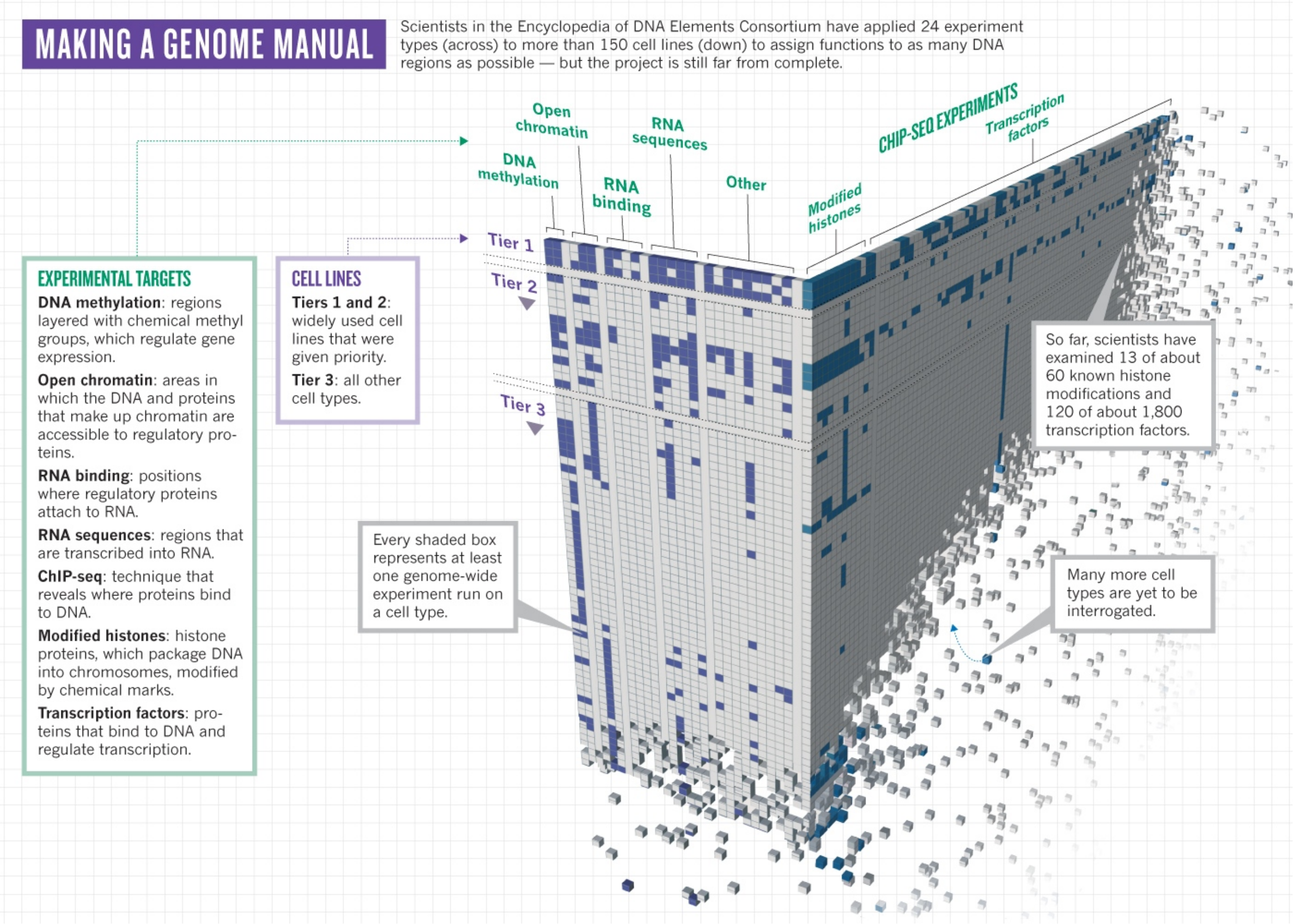
Join ends of crosslinked DNA and remove crosslinks

Sequence fragments

ACGCTGACTAGAATCAATGGCT
TCTCTTCGCATATGGCTGACTA

http://www.sanger.ac.uk/resources/talksandtraining/opendoor/malawi.html

What data do we have?



MAKING A GENOME MANUAL

Scientists in the Encyclopedia of DNA Elements Consortium have applied 24 experiment types (across) to more than 150 cell lines (down) to assign functions to as many DNA regions as possible — but the project is still far from complete.

EXPERIMENTAL TARGETS

DNA methylation: regions layered with chemical methyl groups, which regulate gene expression.

Open chromatin: areas in which the DNA and proteins that make up chromatin are accessible to regulatory proteins.

RNA binding: positions where regulatory proteins attach to RNA.

RNA sequences: regions that are transcribed into RNA.

ChIP-seq: technique that reveals where proteins bind to DNA.

Modified histones: histone proteins, which package DNA into chromosomes, modified by chemical marks.

Transcription factors: proteins that bind to DNA and regulate transcription.

CELL LINES

Tiers 1 and 2: widely used cell lines that were given priority.

Tier 3: all other cell types.

Every shaded box represents at least one genome-wide experiment run on a cell type.

So far, scientists have examined 13 of about 60 known histone modifications and 120 of about 1,800 transcription factors.

Many more cell types are yet to be interrogated.

http://www.sanger.ac.uk/resources/talksandtraining/opendoor/malawi.html

Accessing ENCODE

- ENCODE papers http://www.nature.com/encode/
- The ENCODE portal https://www.encodeproject.org/ Demo 1
- UCSC Genome Browser http://genome.ucsc.edu/ Demo 2
- Ensembl Genome Browser http://www.ensembl.org/index.html Demo 3
- ENCODE/Roadmap browser http://www.encode-roadmap.org/
- IHEC portal http://epigenomesportal.ca/ihec/index.html

Hands on

- We’re going to look at the ENCODE portal to see if we can find any ChIP-seq data for human kidney tissue.
- We will take a brief glance at the ENCODE/Roadmap browser and IHEC.
- Demo: page 108-113

Ensembl Regulation – ENCODE and more!

Ensembl Regulation

The goal of Ensembl Regulation team is to annotate the genome with features that may play a role in the transcriptional regulation of genes.

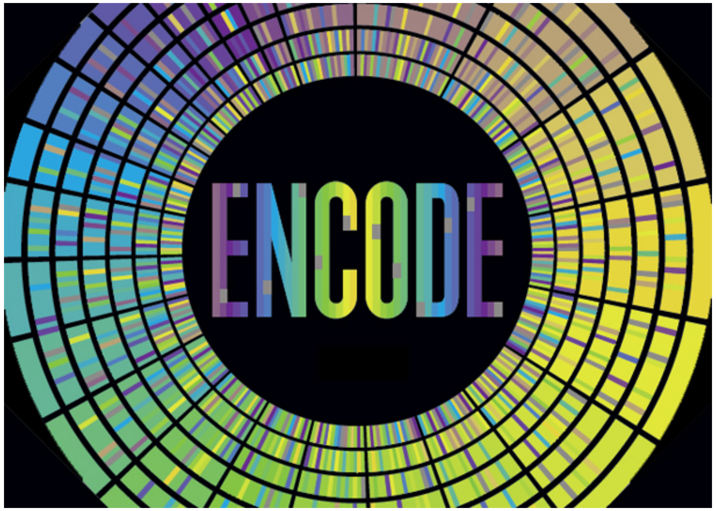
- Predicted open/closed chromatin
  - DNase I sensitivity
  - FAIRE
- Transcription factor binding sites
- Epigenetic marks
  - Histone modifications
  - DNA methylation
- RNA Pol binding

We do not…

- …link promoters/enhancers/insulators or any other regulatory features to genes. We allow you see what is where and make your own inferences.
- …link regulatory features to gene expression. We have cell-line specific regulation data and tissue specific expression data – make of it what you will.

Regulatory data is incredibly complex and still in relative infancy. There is no comprehensive database of regulation data.

Data sources



Ex vivo primary cells and stem cells – involved in human disease.
http://www.roadmapepigenomics.org/

Haematopoietic cell lineage.
http://www.blueprint-epigenome.eu/

A subset of cell types

- Only a subset of available data is displayed in Ensembl.
- We display cell types that have, at a minimum:
  - CTCF binding
  - DNase or FAIRE data
  - H3K4me3, H3K27me3, H3K36me3 data
- We display all TFBS and histone modification data known in these cell types.
- We process these data to predict activity.

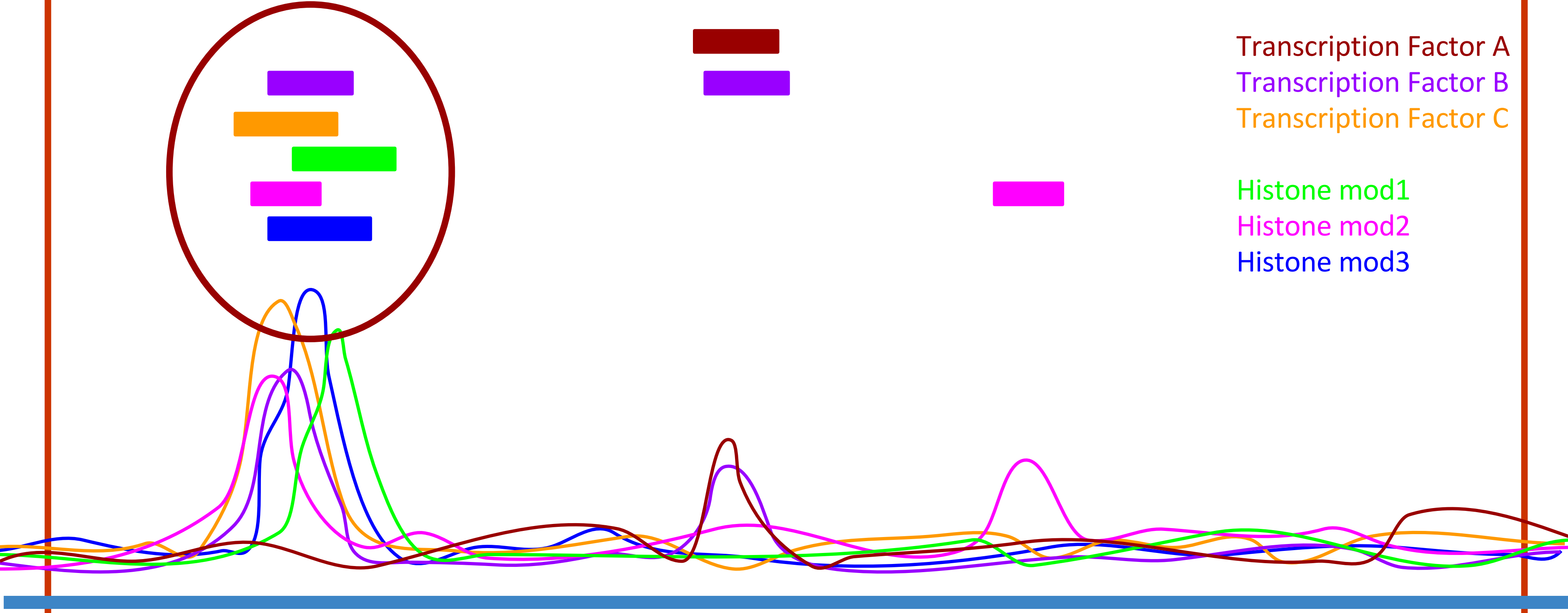
- Further data can be added using track hubs.

Processing the data

- The raw data is taken from the various sources.
- This is processed to predict the positions of regulatory features, such as promoters, enhancers and insulators.
- The activity of these features is predicted in the different cell types.

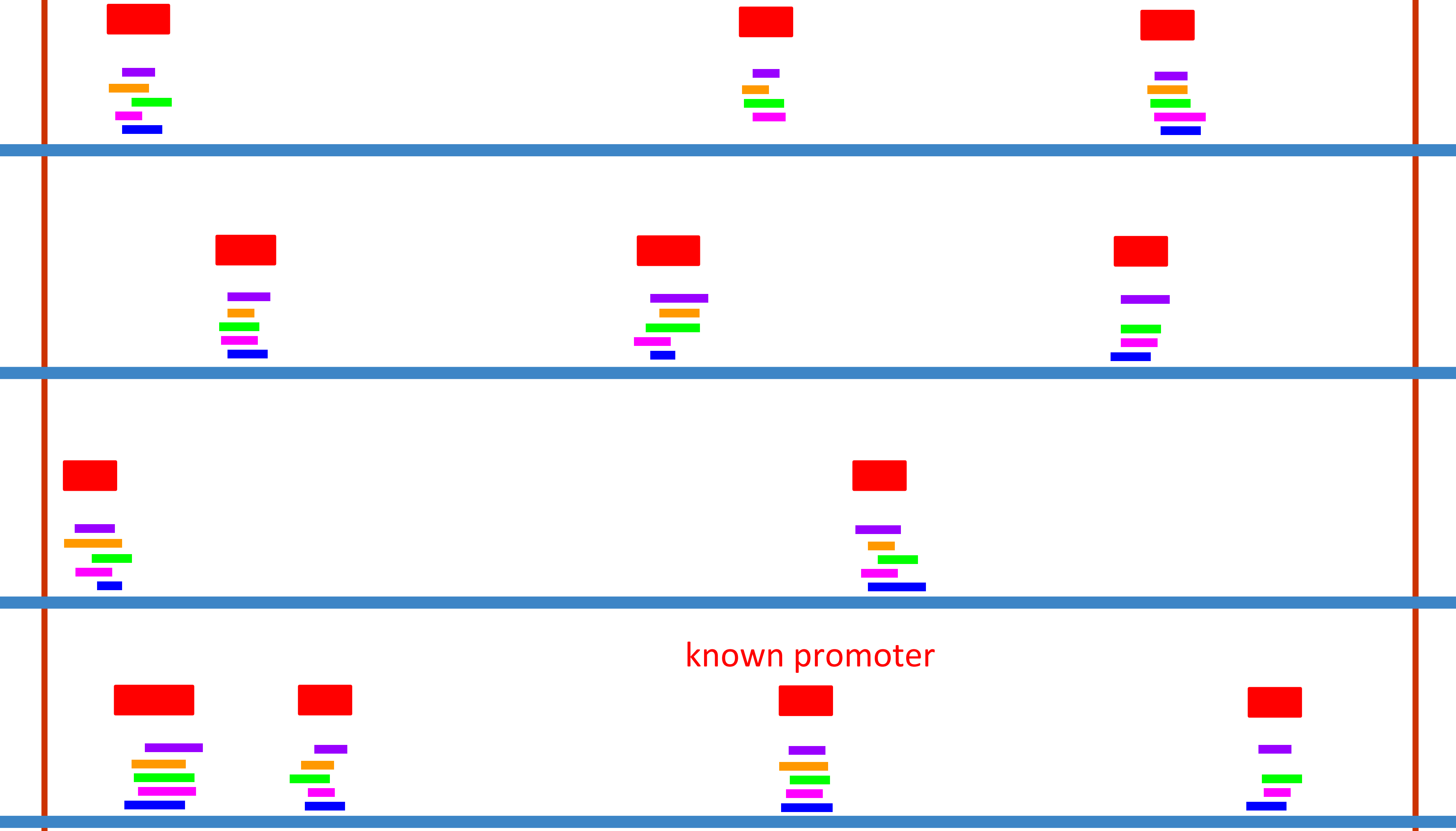
- All of this can be viewed in the genome browser.

Raw data



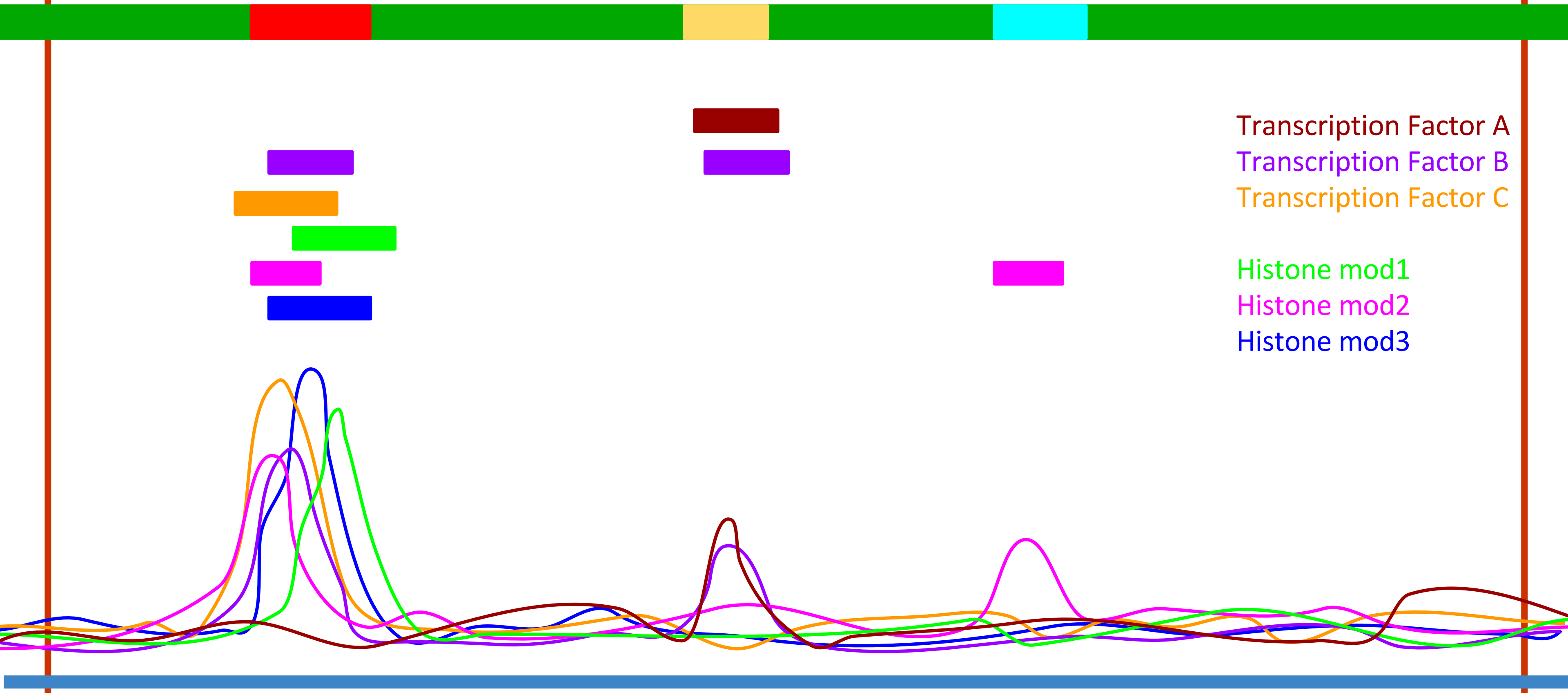
http://www.sanger.ac.uk/resources/talksandtraining/opendoor/malawi.html

Searching for patterns



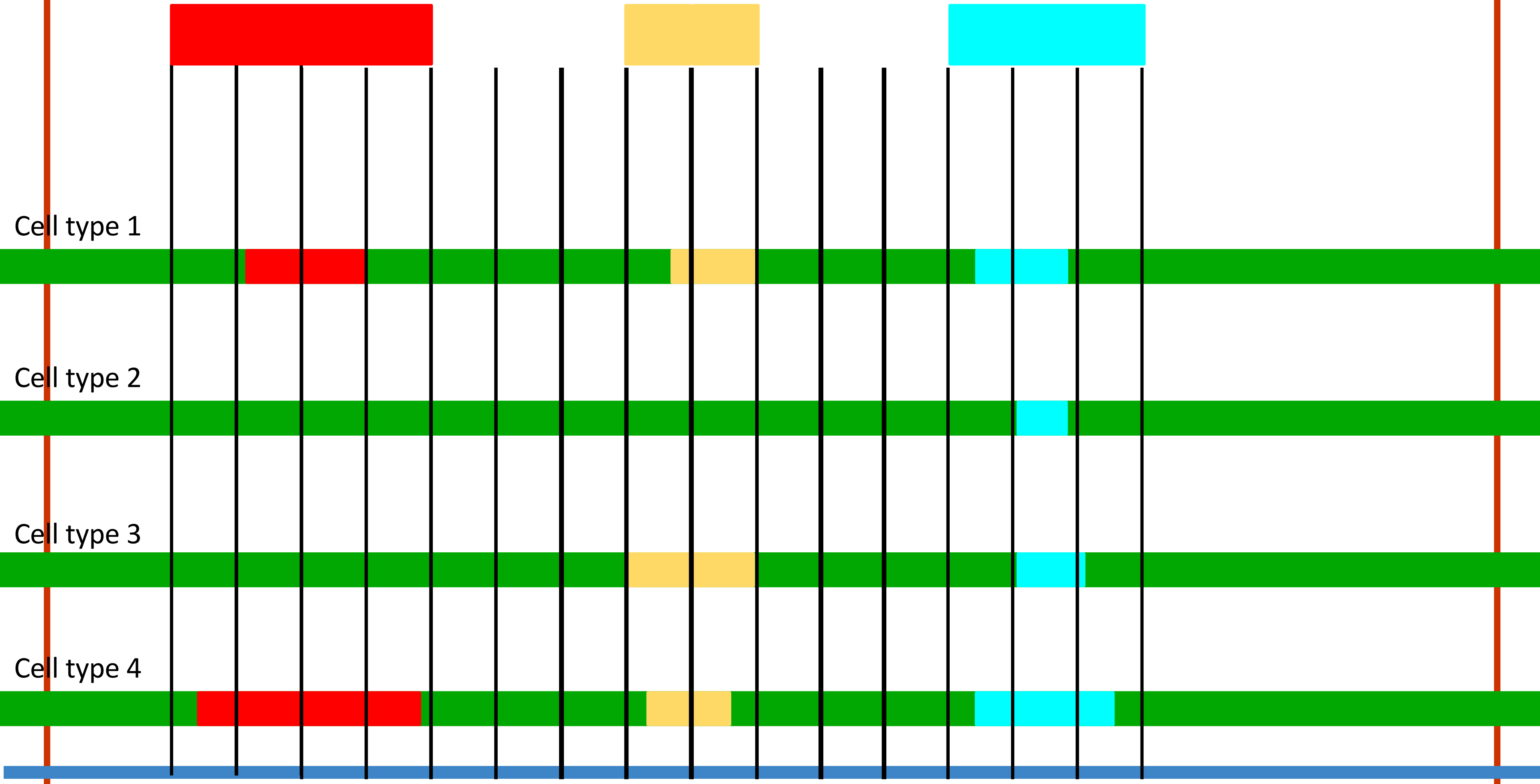
http://www.sanger.ac.uk/resources/talksandtraining/opendoor/malawi.html

Segmentation



http://www.sanger.ac.uk/resources/talksandtraining/opendoor/malawi.html

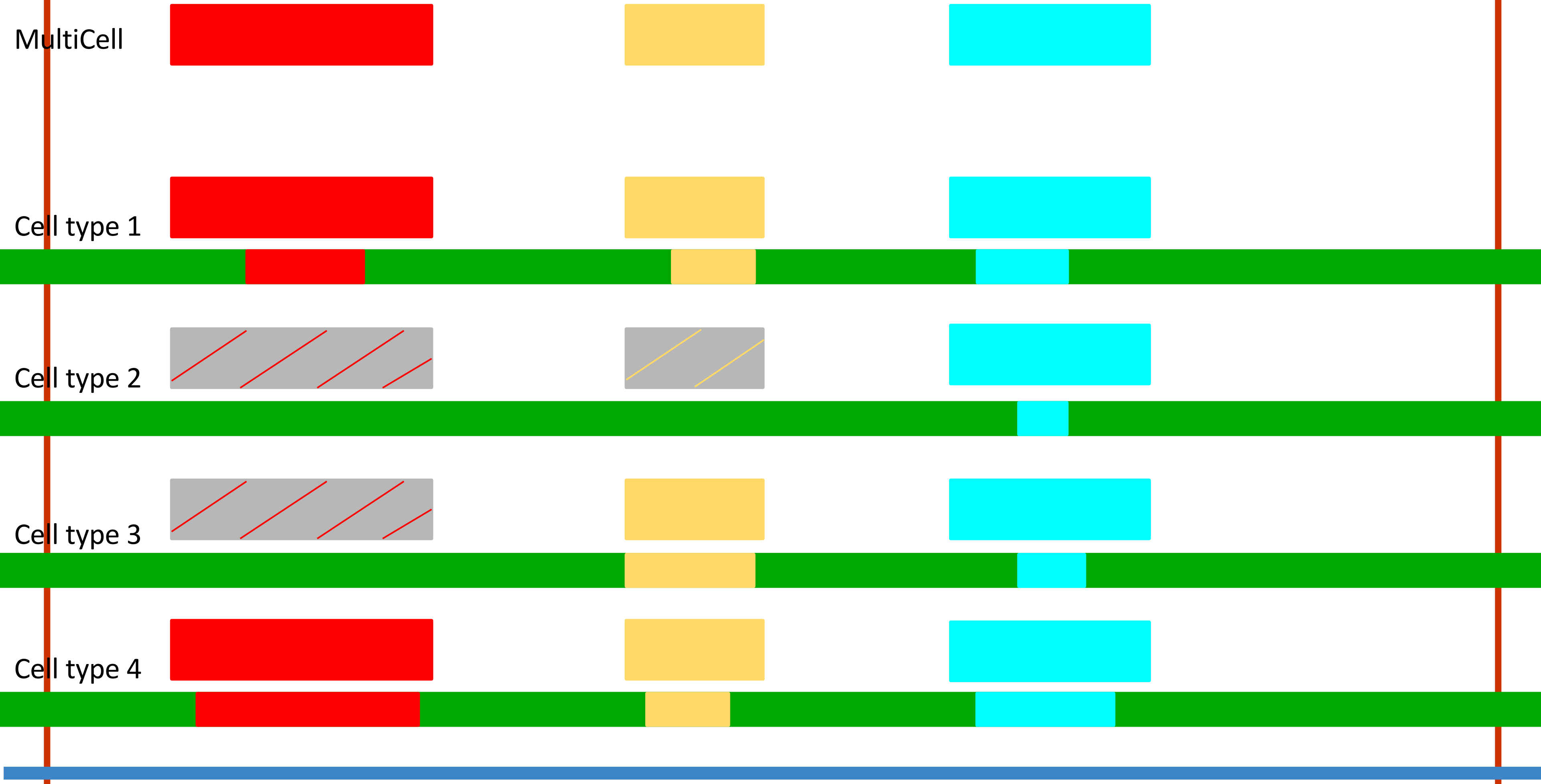
MultiCell features



http://www.sanger.ac.uk/resources/talksandtraining/opendoor/malawi.html

The Open Door Workshop

Cell-specific features



http://www.sanger.ac.uk/resources/talksandtraining/opendoor/malawi.html

Coverage

Label	Count	Mean length (bp)	Max length (bp)	Total length (Mbp)
TSS	40,249	973.2	11,400	39.2
Proximal Reg.	101,206	1005.5	15,000	101.8
Distal Reg.	209,081	526.1	8,400	110.0
CTCF	108,284	550.1	5,200	59.6
Unannotated TFBS	163,528	155.8	1,630	25.5
Union				299.2

Hands on

- We're going to look at the region of a gene *LIMD2* to find regulatory features and explore what cells types they are active in and what evidence there is to show this.
- Demo: page 118-129
- Exercises: page 130-131
  - Answers: page 131-134