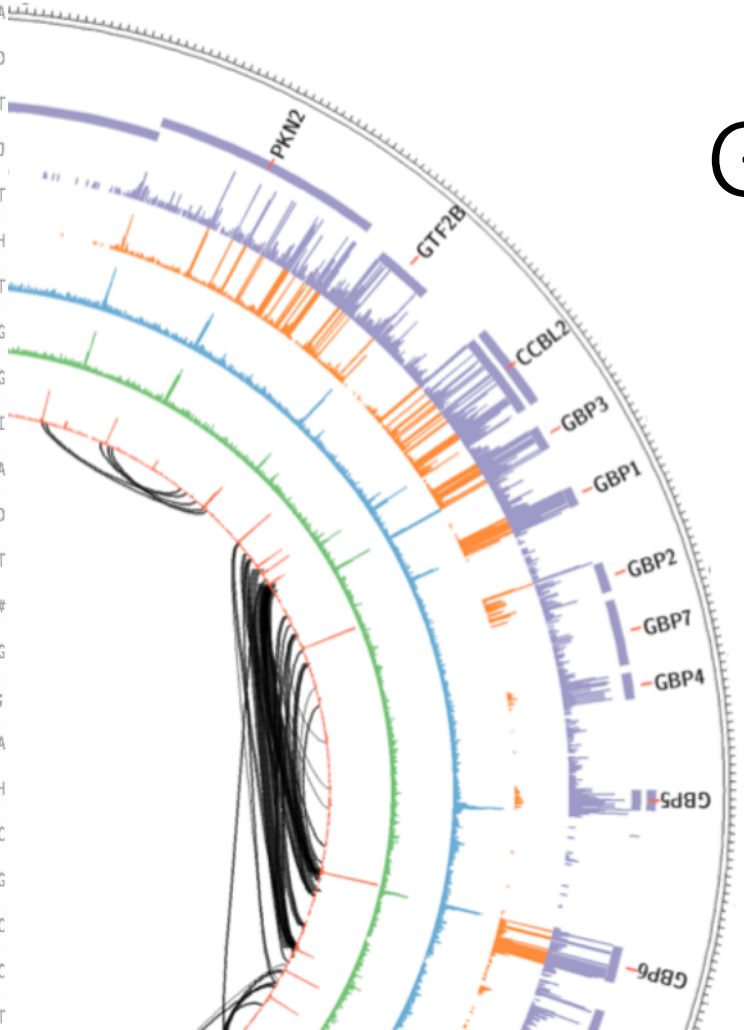
Understanding transcriptional regulation using advanced NGS methods: GRO-seq & Hi-C



Chris Benner

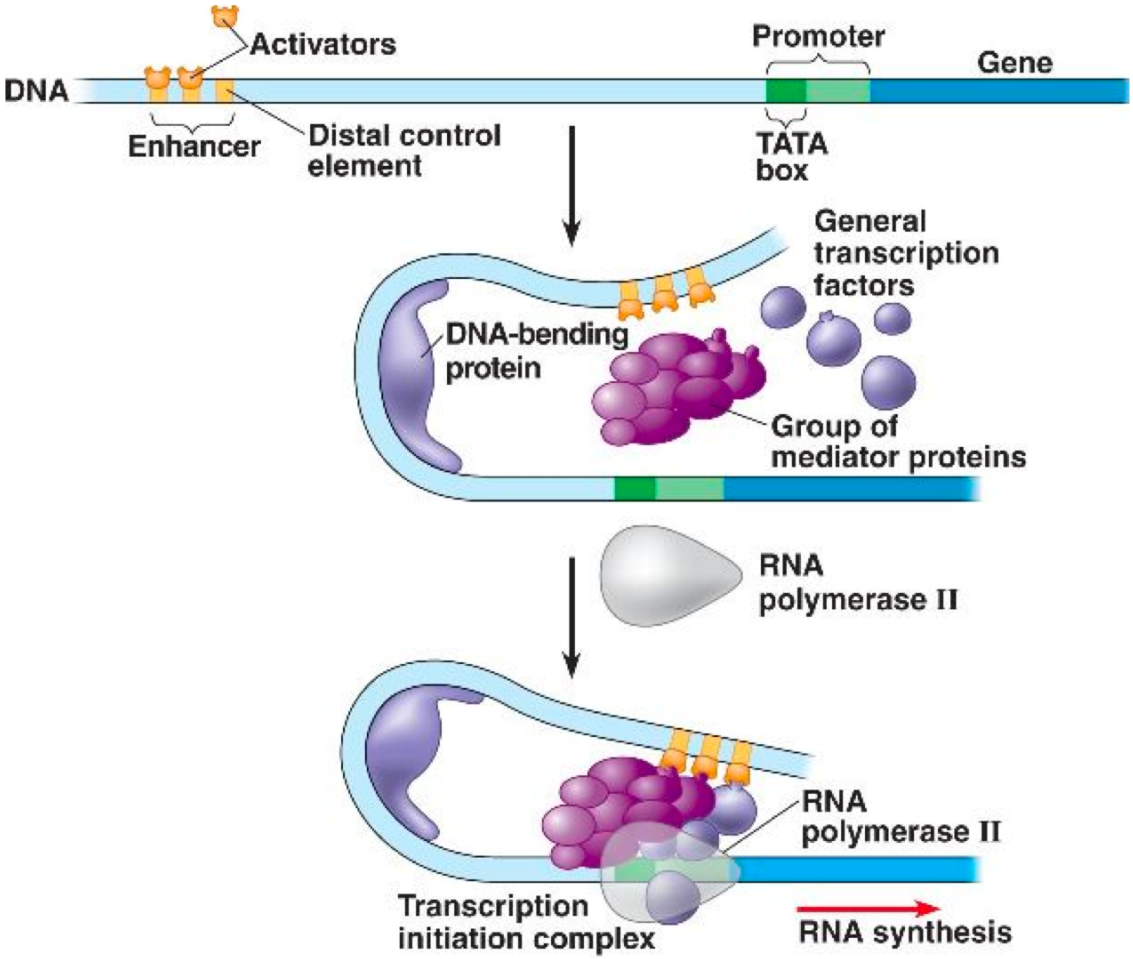
Assistant Professor

UCSD Department of Medicine

MED263

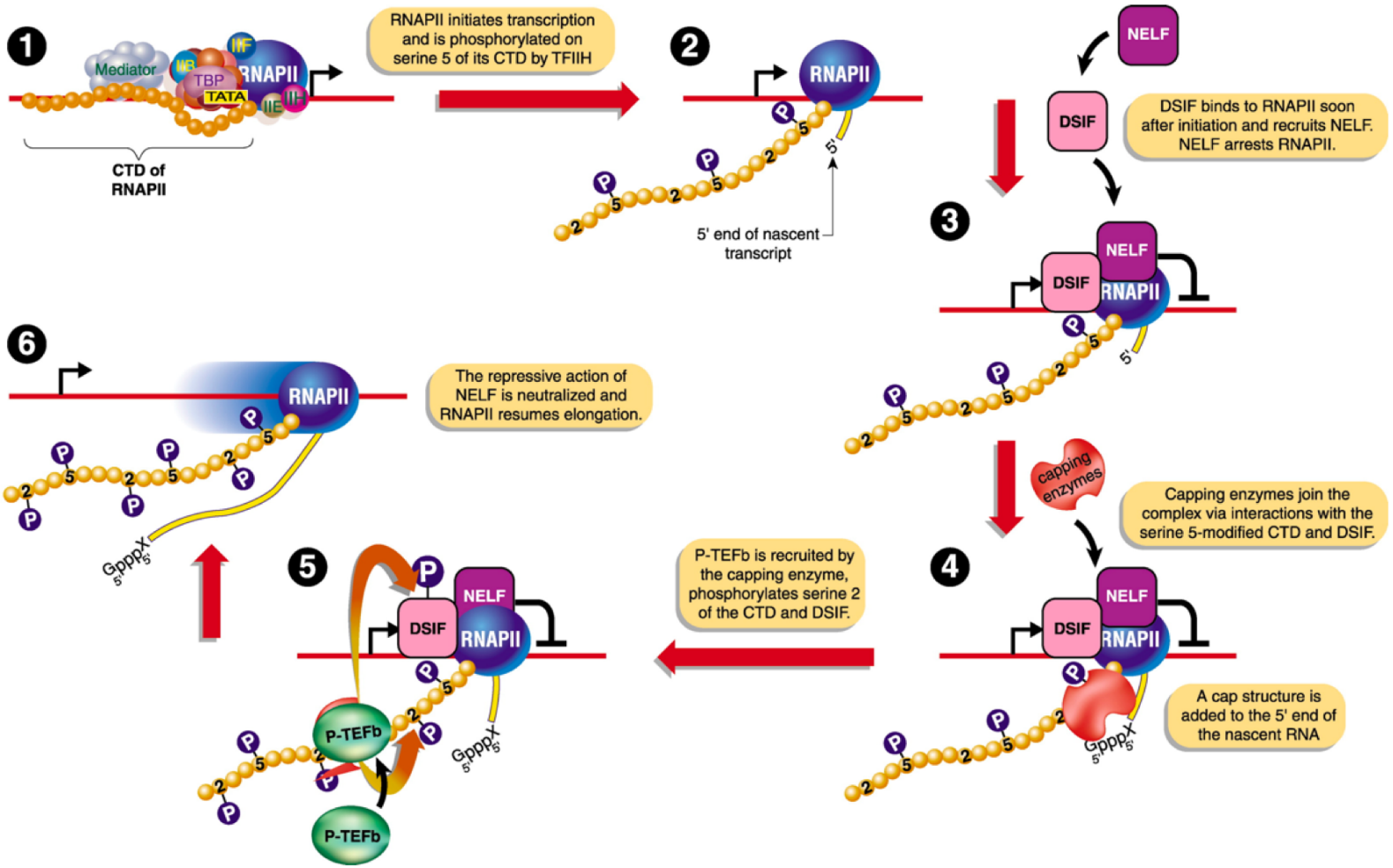
02/27/2018

Commonly used model of transcription

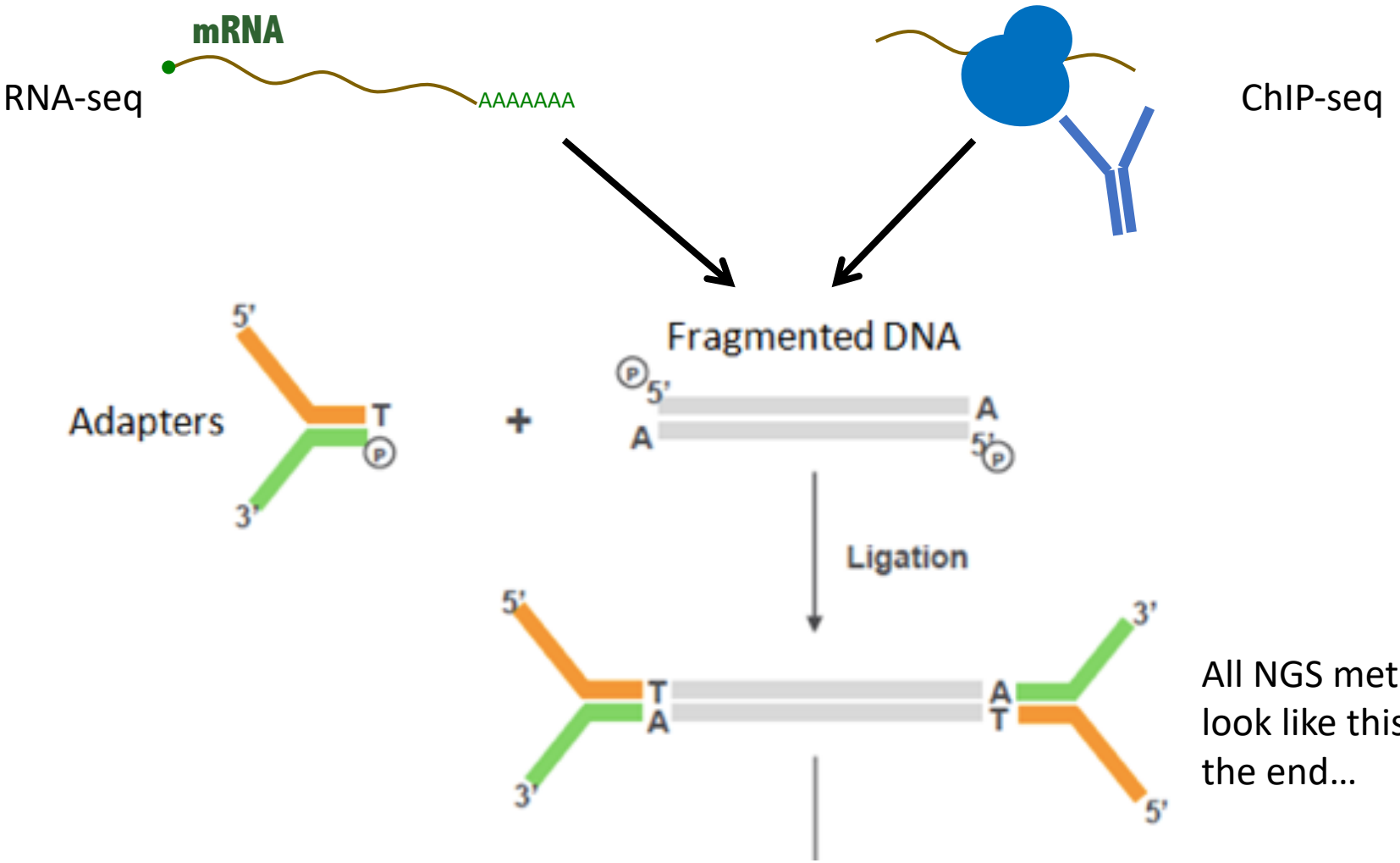


From Simone Silva http://flipper.diff.org/app/items/info/6731

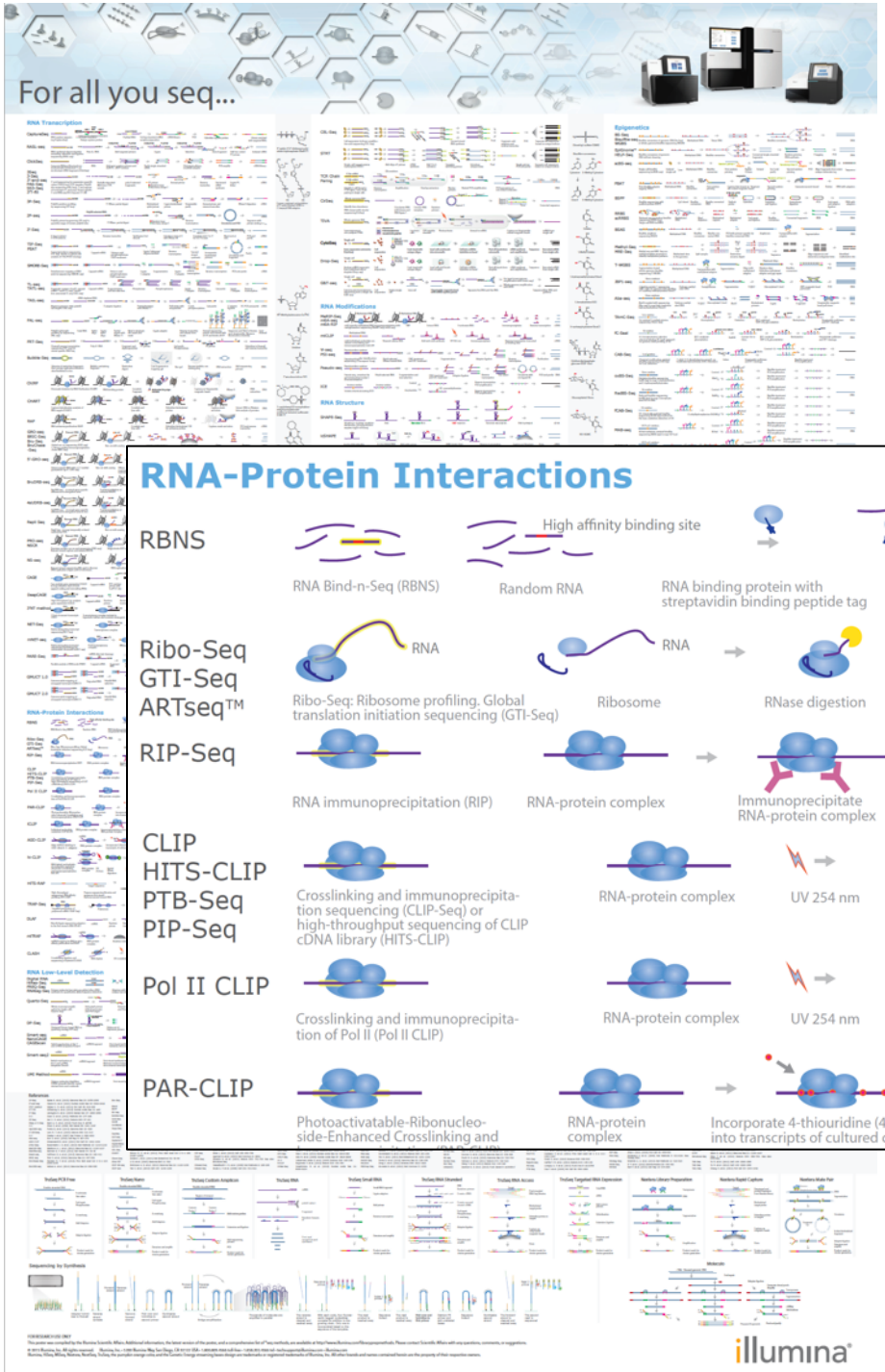
Transcription is more complicated that it seems at first…



Flexibility of the platform: Every NGS method ultimately becomes a dsDNA sequencing library



All NGS methods look like this in the end…



Golden age of NGS method development

Why "infer" measurements when you can design an assay to directly measure different aspects of cells?

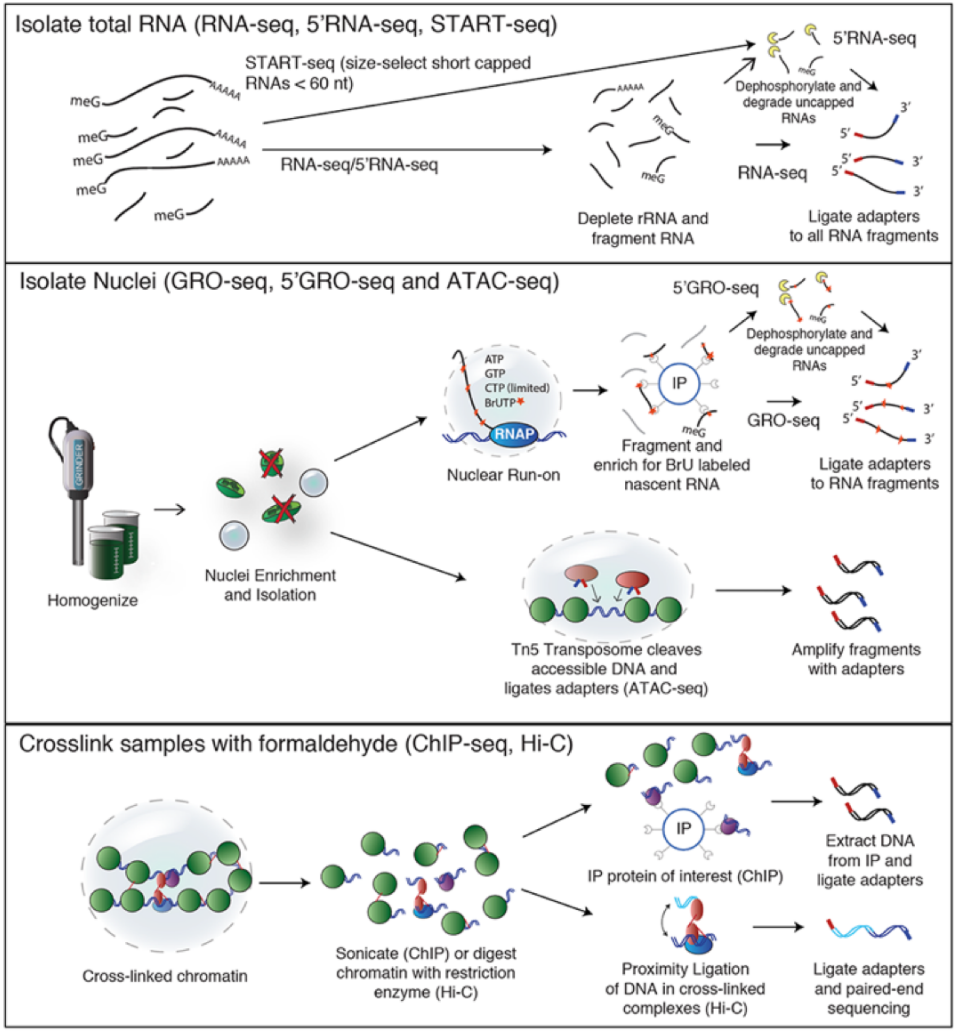
For All you Seq (Illumina website)

Key Impact of NGS Technology

- The methods and science are not new – most of the techniques, enzymes, strategies used in NGS methods predate sequencing
  - biochemistry and molecular biology are still the same!
- What's different: Measurements are now ***unbiased***, which impacts data quality, discovery, and interpretation.
- Using RNA-seq as an example:
  - Data Quality: Sample normalization and quality control much more robust with 10,000+ measurements per sample (i.e. internal control)
  - Discovery: Allows identification of novel RNAs, or foreign contaminants.
- ***Change in scientific approach:***
  - Ability to use genome-wide profiling enables *unbiased* analysis. Subtle but important change to which questions can be asked of data:
    - Before (i.e. qPCR): Is gene X regulated by stimulus Y?
    - Now (i.e. RNAseq): What genes are most strongly regulated by stimulus Y?

Two transformative methods to study transcriptional regulation:

- Nascent RNA methods
  - Used to measure transcriptional activity by sequencing nascent RNA
  - RNA-seq is a poor substitute because it only measures steady-state levels of stable transcripts)
- 3D Genome Structure*
  - Used to understand how the genome folds in the nucleus using proximity-ligation technology
  - Produces a contact matrix - not the same as actual 3D imaging
- **NOTE: the term "***genome structure***" (without the 3D) usually refers to linear genome organization, i.e. rearrangements, inversions, etc.*



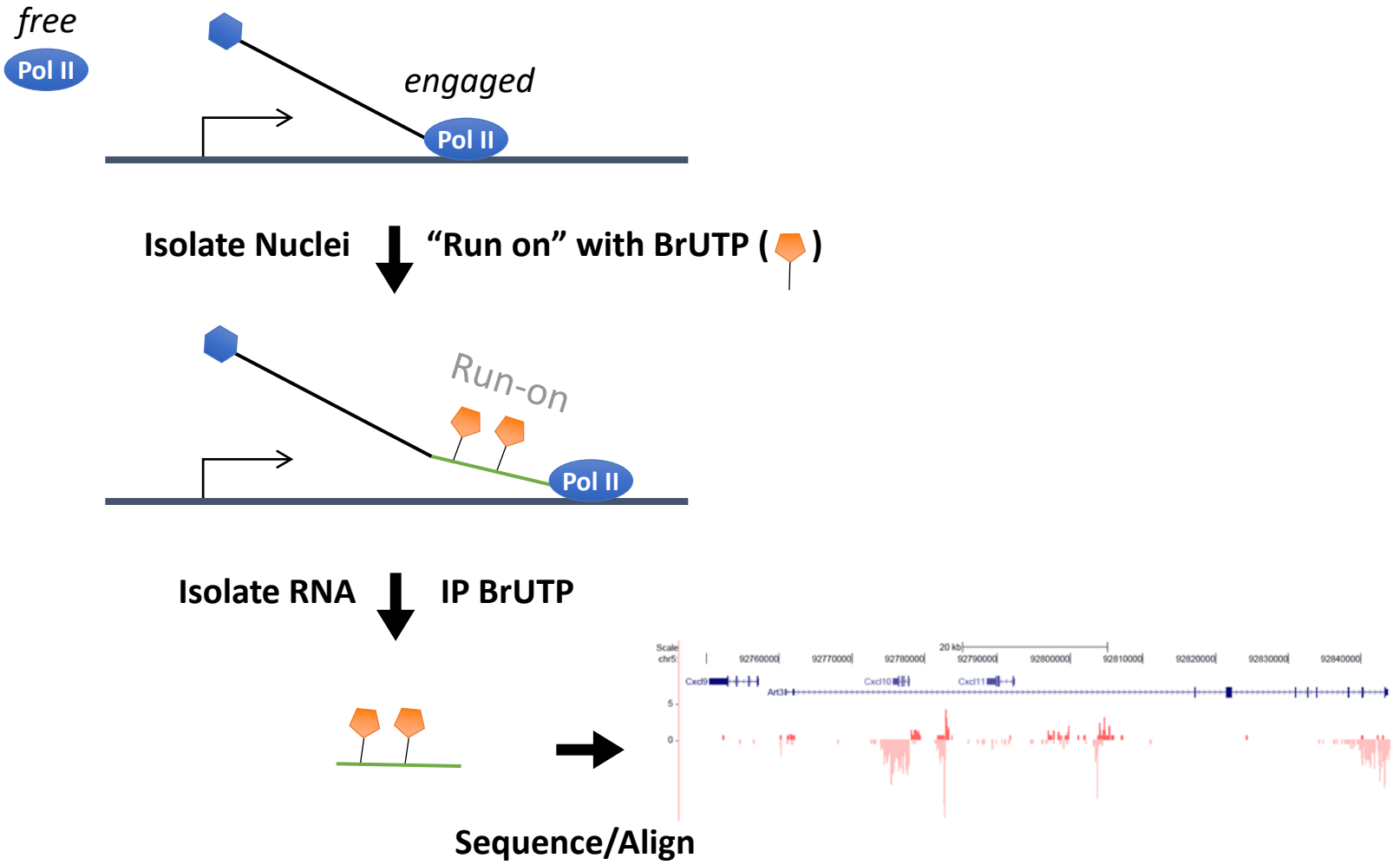
Overarching concepts for today

- Exciting time to study gene regulation – clash of unbiased genome-wide observations with long engrained concepts (that may or may not be true)
- Innovation in genomics/NGS methods and analysis are rapidly accelerating the field
- Nascent RNA != RNA-seq
- Current thinking about 3D genome structure is that it is organized at two different levels:
  - Global: Transcriptional activity
  - Local: Cohesin/condensin loops
- How important is 3D genome structure for gene expression (i.e. promoter enhancer loops) – do we really know??

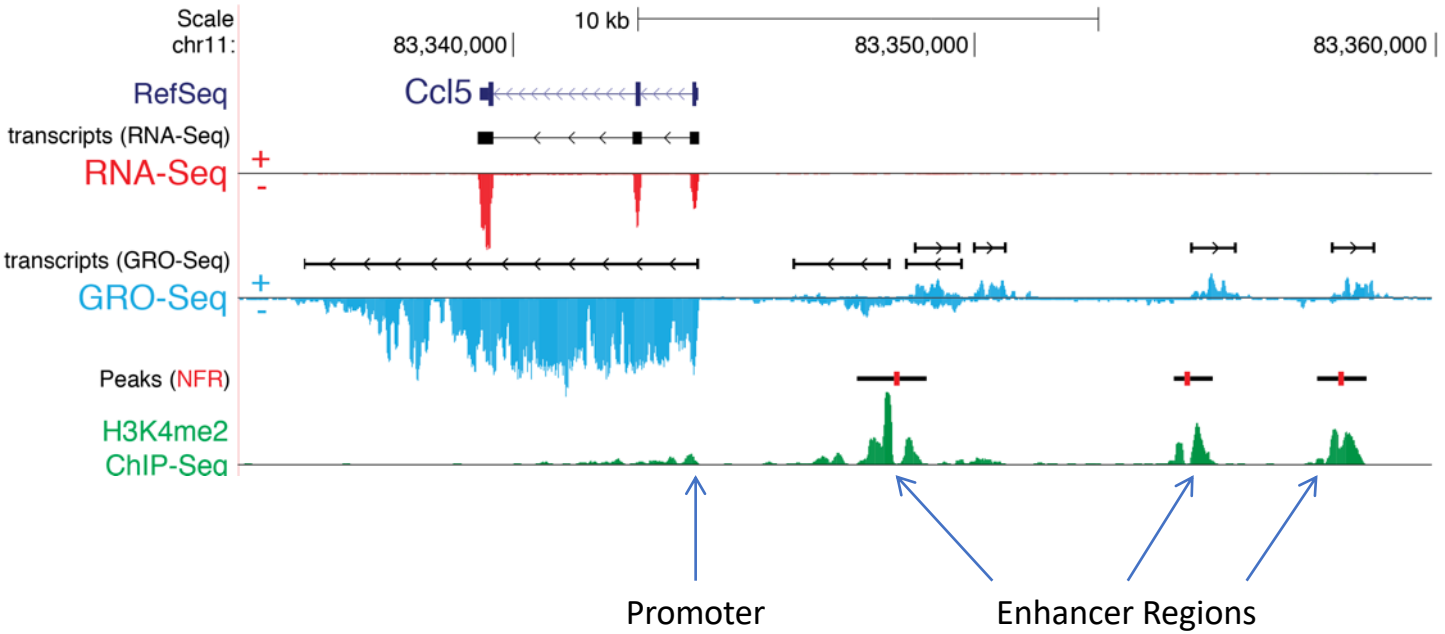
Part I: Using methods to measure nascent RNA

- Global nuclear Run-On (GRO) sequencing.
  - Core, L; Waterfall, J; Lis, JT (2008). "Nascent RNA Sequencing Reveals Widespread Pausing and Divergent Initiation at Human Promoters". Science.
  - Uses labelled nucleotides to isolate newly synthesized RNA molecules *before they have a chance to be degraded*
- Variants of the Method:
  - PRO-seq, NET-seq (similar)
  - TSS mapping/5' cap-enriched versions: GRO-cap, PRO-cap, Start-seq, 5'GRO-seq
  - Methods that are similar to GRO-seq but not the same:
    - 4sU labelled sequencing
    - Chromatin-associated RNA-seq

Global Run-on Sequencing (GRO-Seq, Core et al. *Science* 2008)

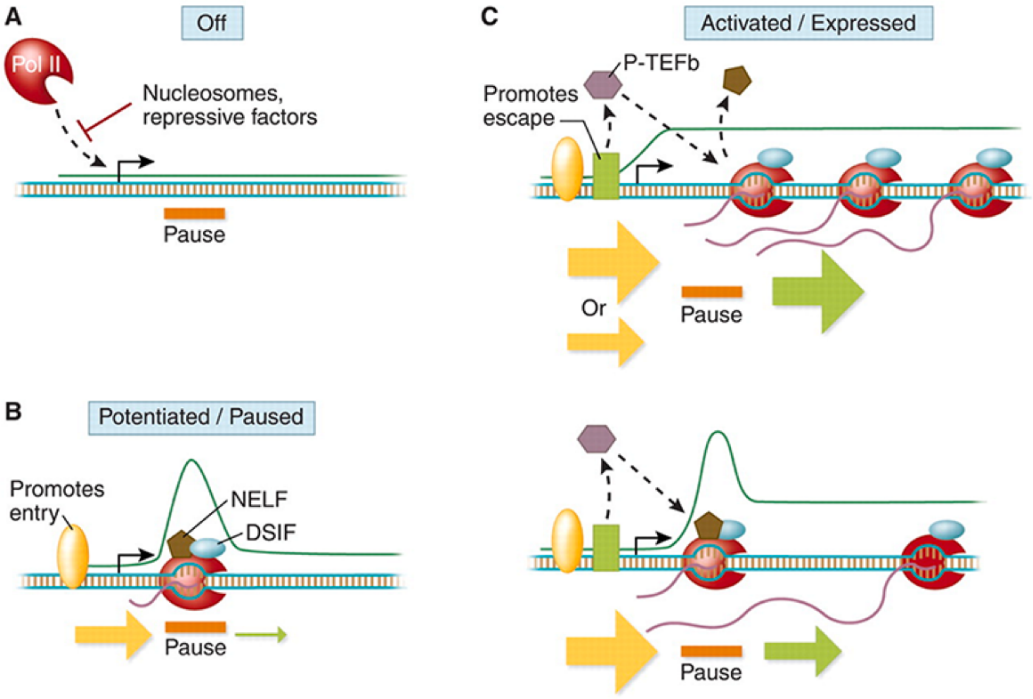


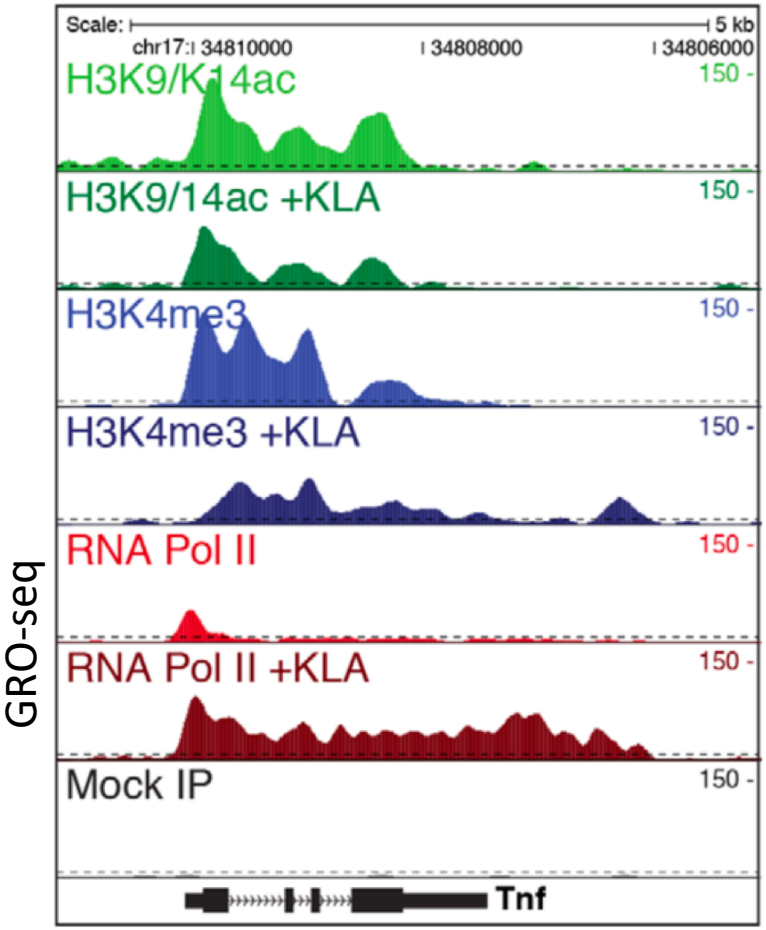
Example comparison of RNA-seq, GRO-seq, and ChIP-Seq



- Features of GRO-seq that can be studied:
  - Gene expression levels (compare to RNA-seq to estimate transcript stability)
  - Promoter-proximal pausing
  - Enhancer activity (eRNA)
  - Boundaries of transcription (i.e. extension transcription past the 3’ end of genes)
  - Transcription speed
  - RNA production at non-Pol II transcripts (i.e. rRNA, tRNA, miRNA primary transcripts, endogenous retroviruses/repeats, etc.)

Promoter Proximal Pausing

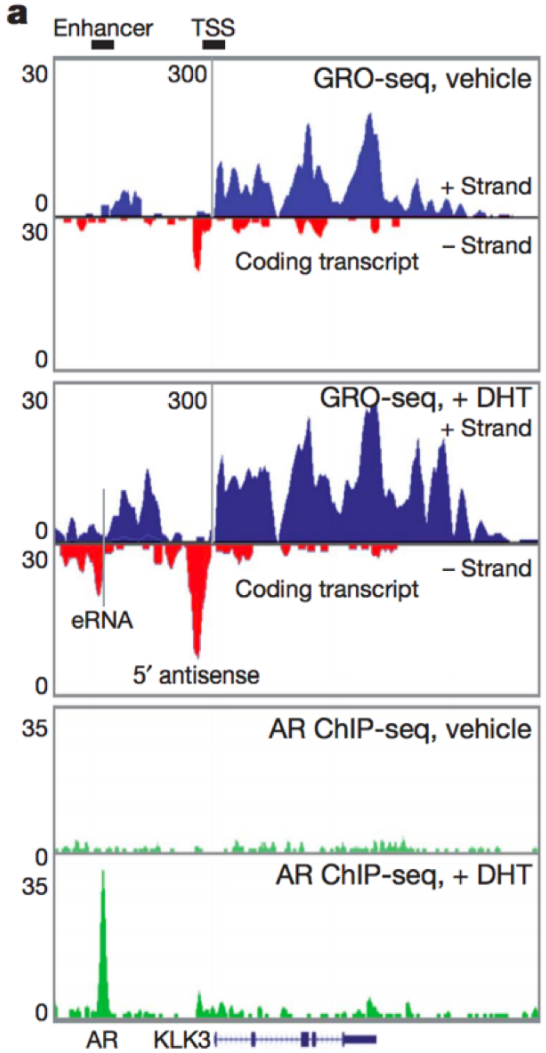




Model from Core & Lis Science 2008

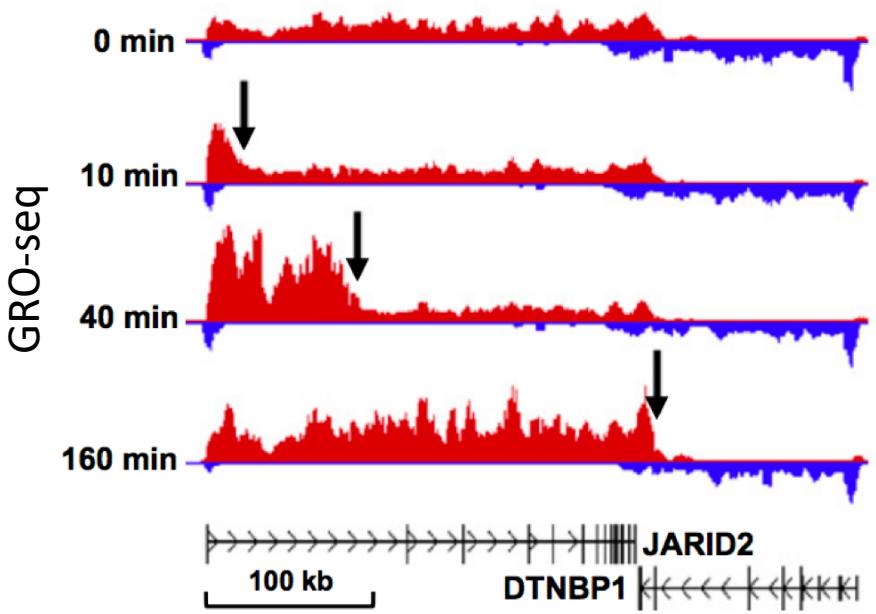
Data from Escoubet-Lozach et al. PloS Genetics 2011

Measuring enhancer RNA (eRNA) induction



Dong et al. Nature (2011)

Measuring speed of transcriptional elongation

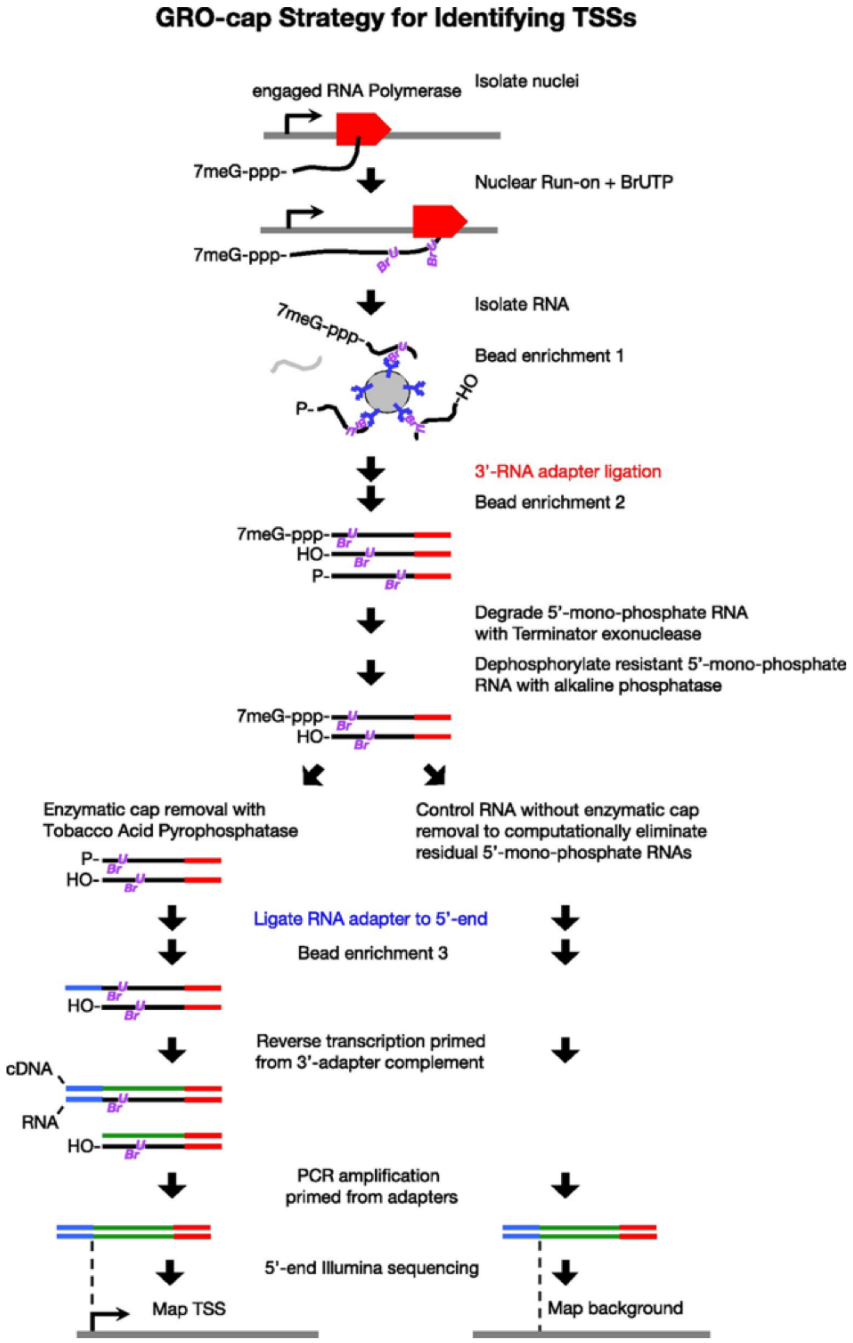


Breast cancer cells (MCF7) responding to estradiol

Nah et al. Cell (2011)

Combining GRO-seq with Transcription initiation mapping

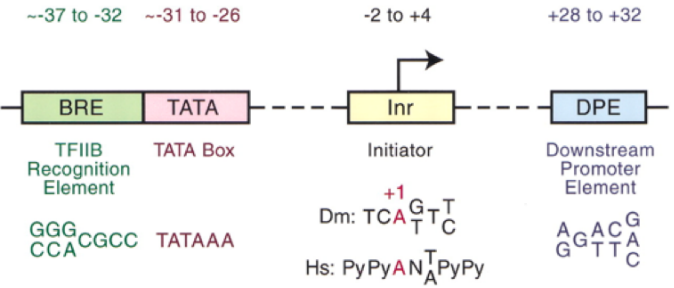
- Maps the exact nucleotide where transcription starts independent of transcript stability
- Methods:
  - 5'GRO-seq
  - GRO-cap
  - PRO-cap
  - Start-seq (similar data but method: works by simply isolating short [<65bp] capped nuclear RNAs)

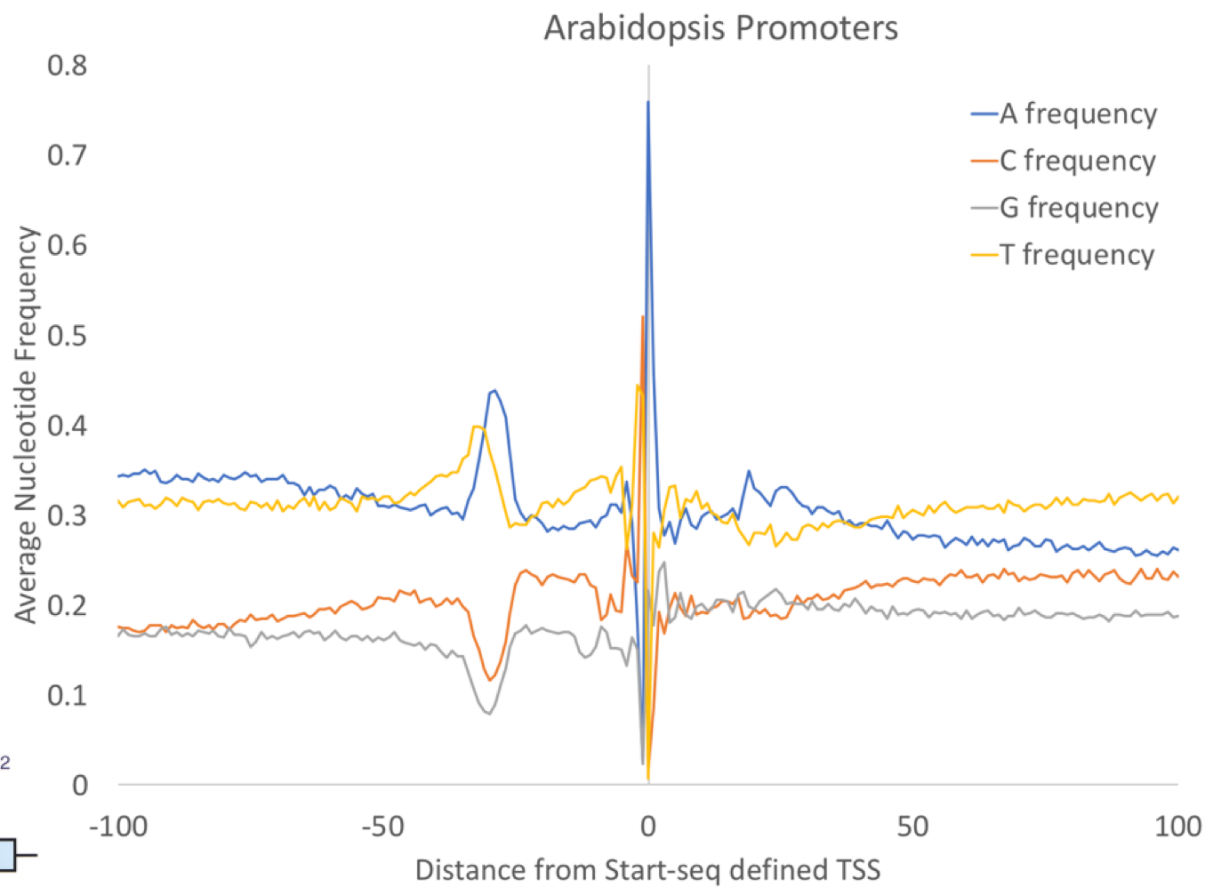


From Kruesi et al. Elife 2013

Unbiased discovery from NGS

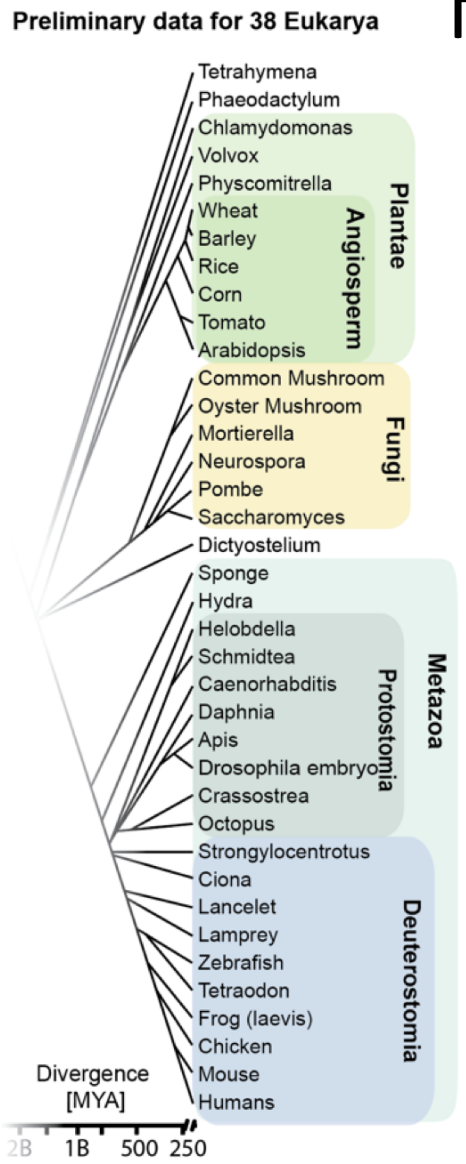
- Find TSS from Start-seq data (i.e. peak finding)
- Calculate the average nucleotide frequencies near TSS
- Can independently verify evidence for Initiator and TATA box

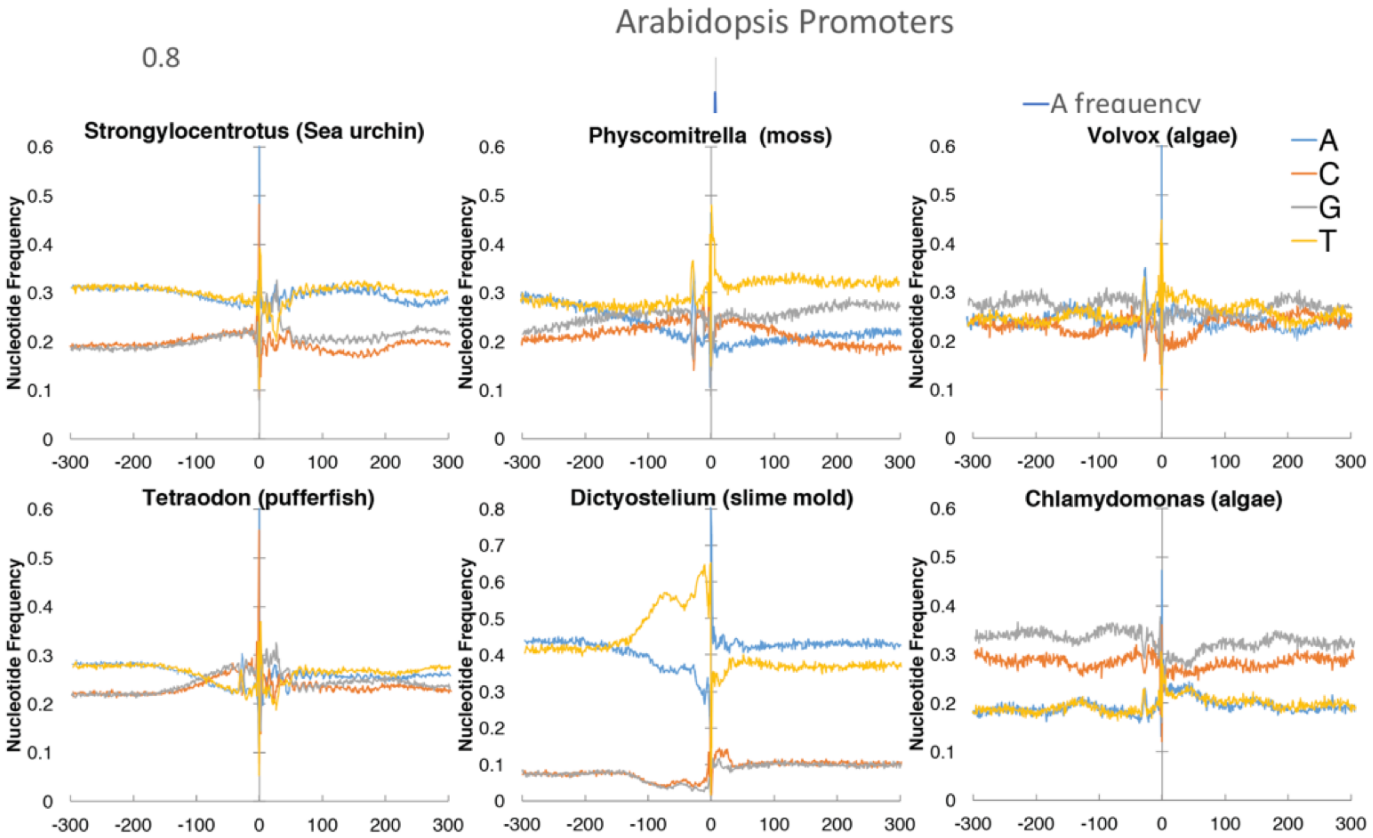




Data by Sascha Duttke

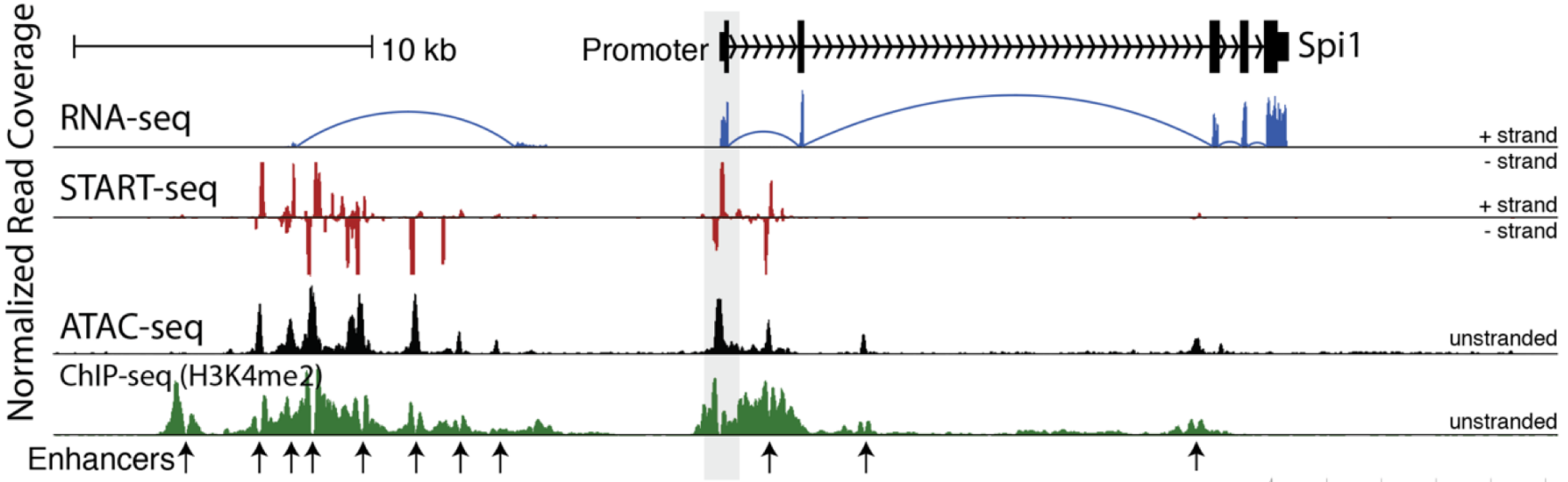
Having fun with evolution…



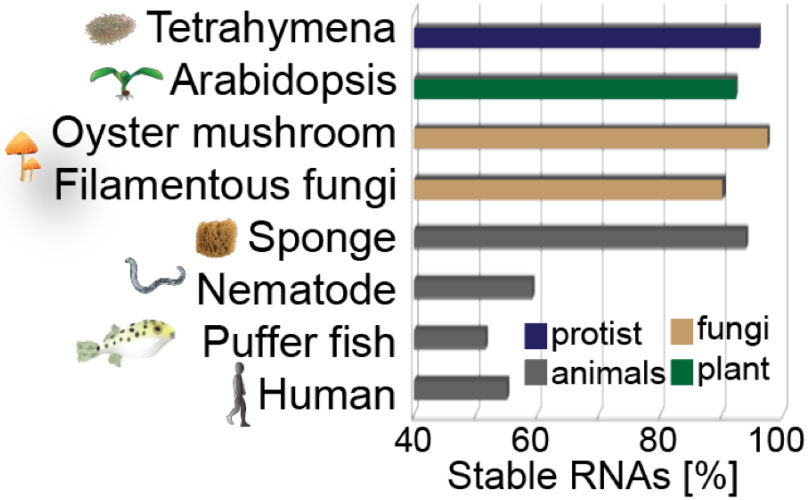


Data by Sascha Duttke

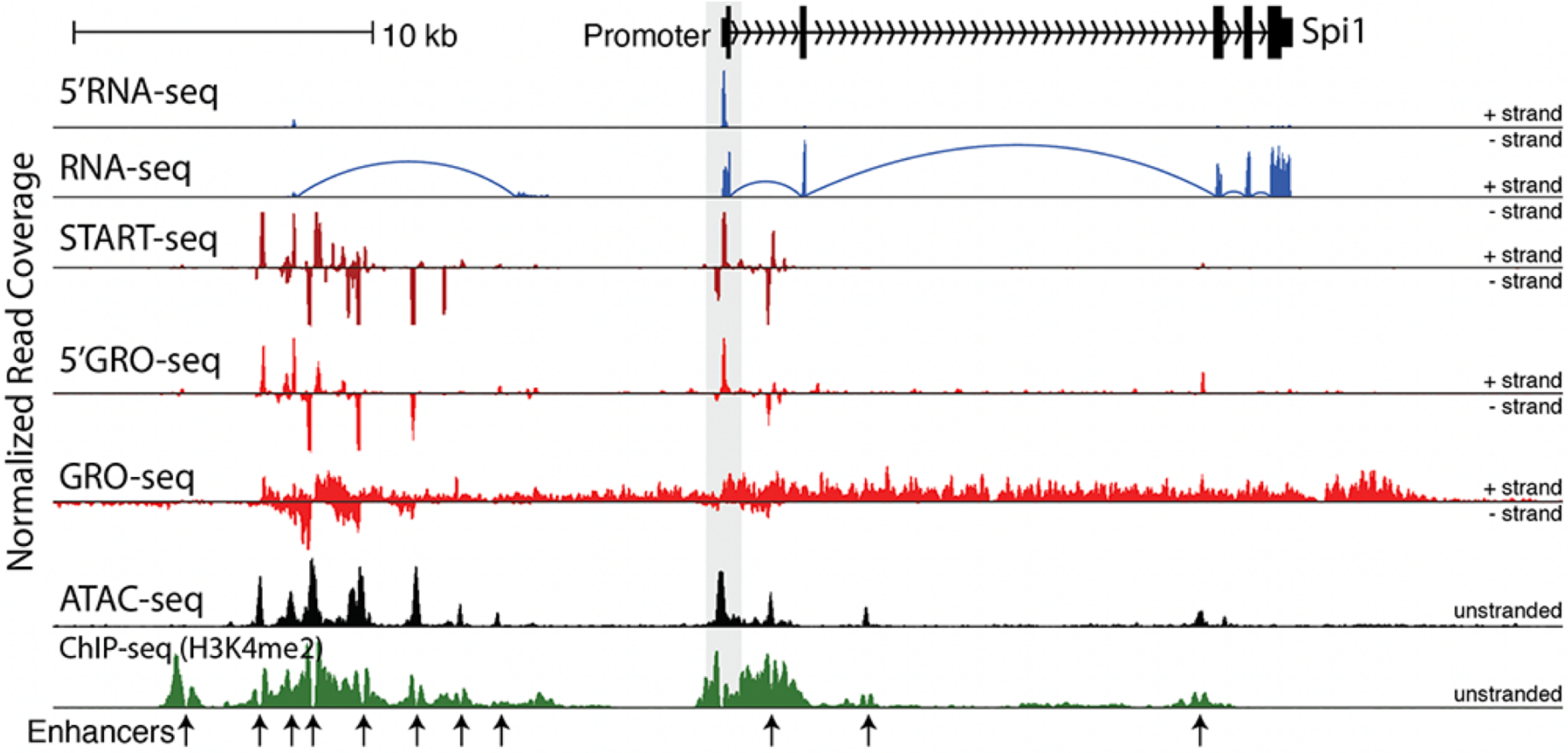
Start-seq data in mammalian cells



- Transcription initiation is pervasive throughout mammalian genomes at distal regulatory regions (i.e. enhancers)
- These sites [typically] do not create stable RNAs

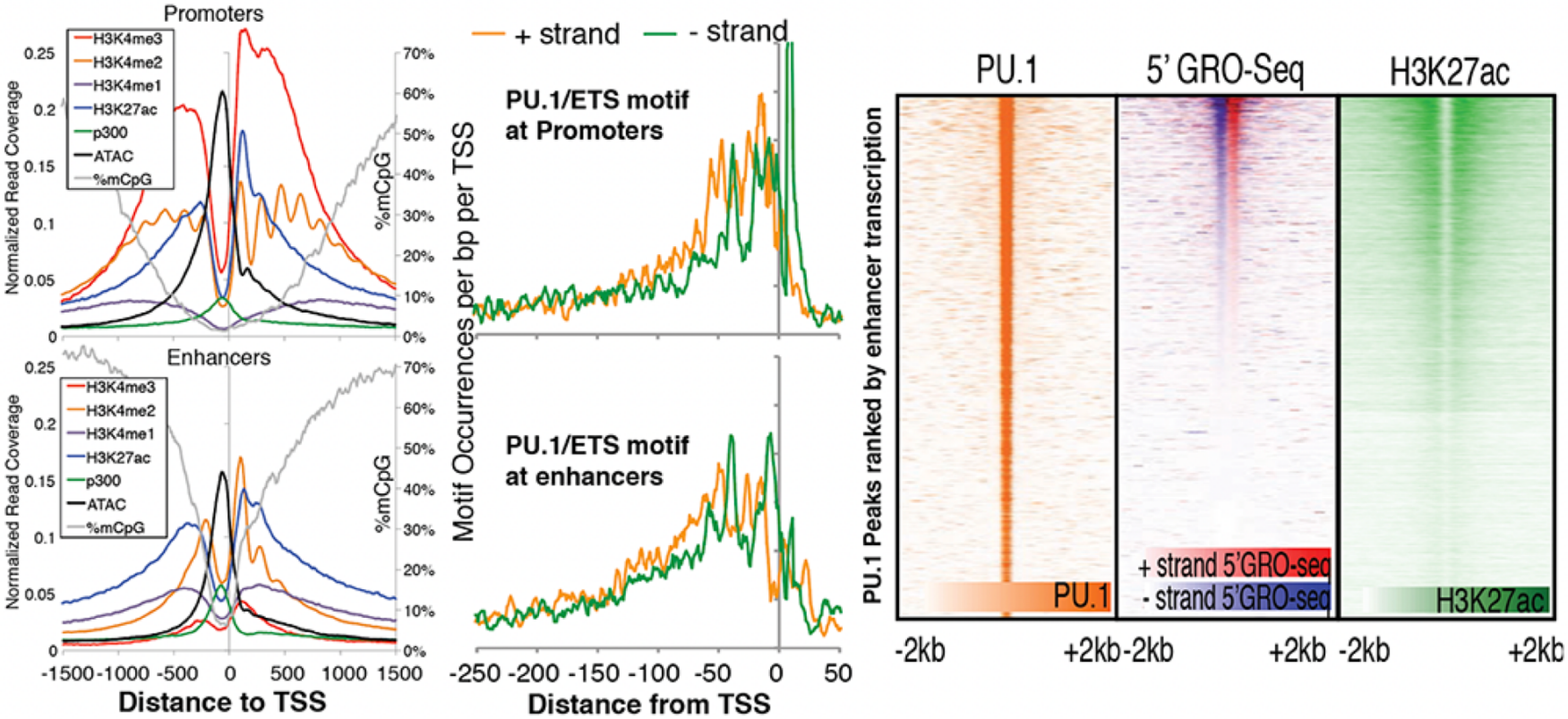


Comparison of GRO-seq and 5’GRO-seq



- Advantages of 5’GRO-seq/START-seq vs. GRO-seq
  - Reads are concentrated at enhancers and promoters instead of gene bodies, providing quantitative activity of regulatory elements with relatively little sequencing
  - Maps initiation sites at single nucleotide resolution
  - Allows the identification and quantification of intragenic enhancers

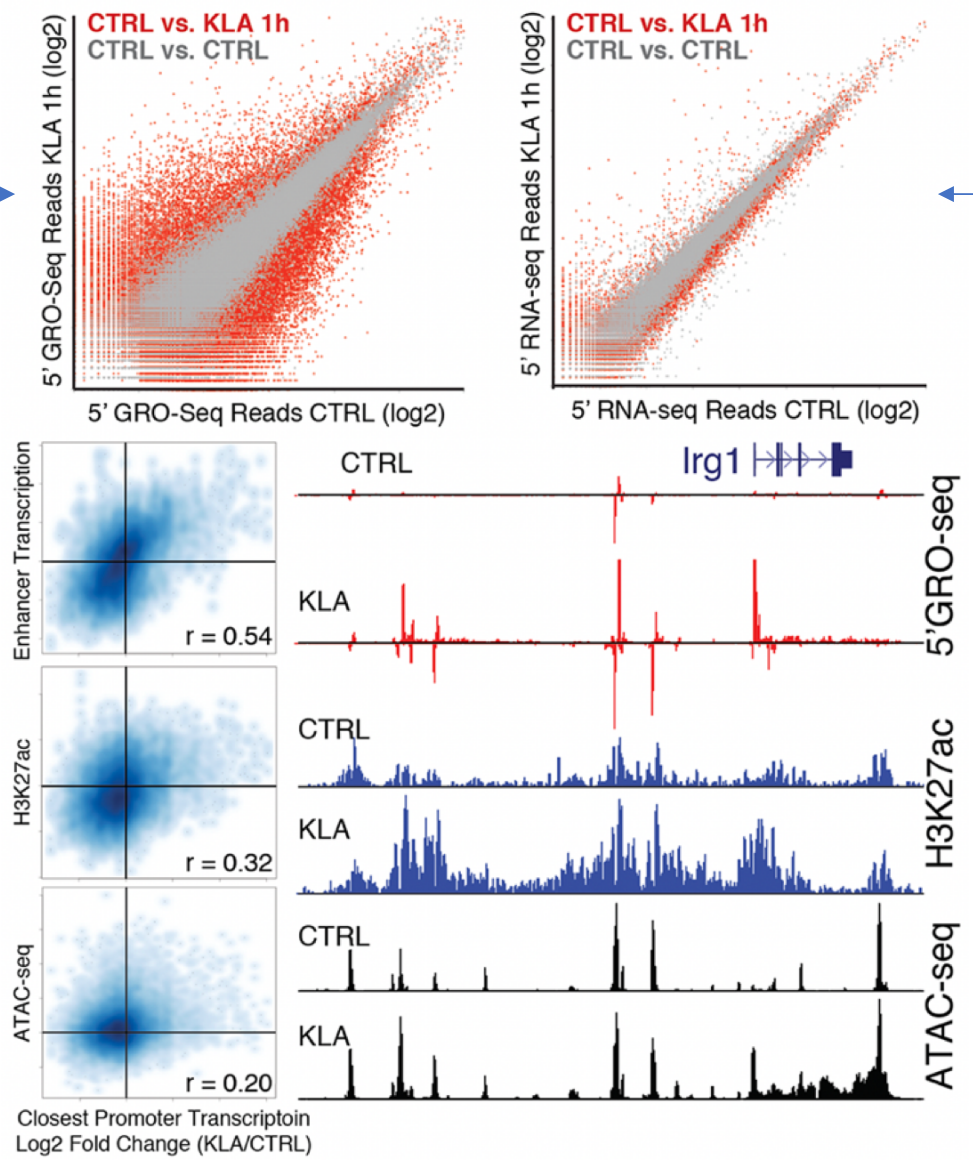
5’GRO-seq maps transcription start sites at enhancers for the first time



Turns out enhancer and promoters… are essentially the same thing. Noticeable differences: Enhancers lack H3K4me3 (epigenetic) and are depleted for splice donor sites downstream of the TSS (genetic)

5’GRO-seq reveals sensitivity of enhancers to perturbation

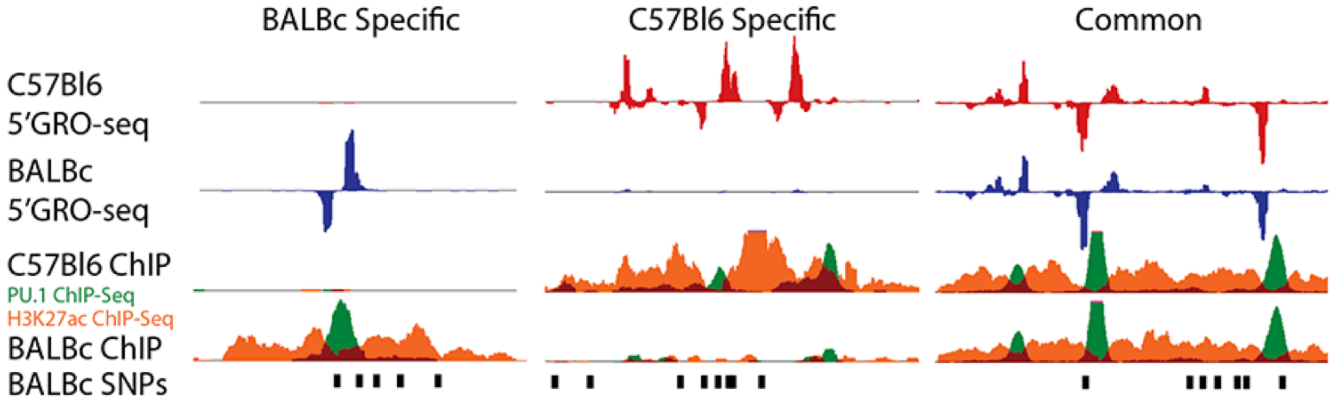
Changes in nascent RNA initiation levels after 1h of TLR4 activation



Changes in stable RNA initiation levels after 1h of TLR4 activation

Use of GRO-seq in medical genomics

- Just starting to be used – very new technology!
- Annotation of non-coding variants
  - GRO-seq is very sensitive at measuring activity at enhancers
  - Helps to identify variants with roles in transcriptional initiation, pausing, enhancer function, or elongation defects



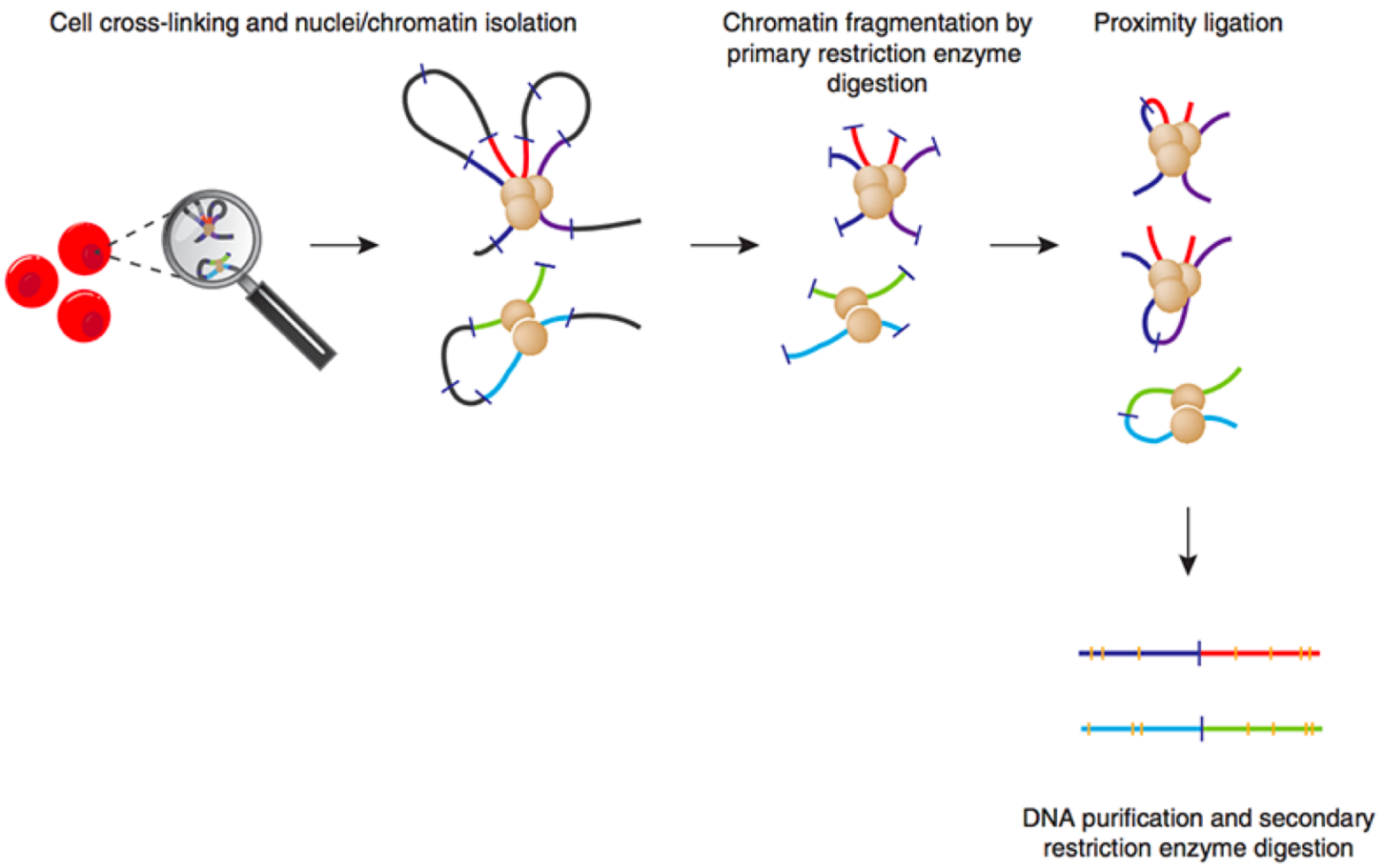
Analysis of GRO-seq Analysis

- More details to come on Thursday's Class
- Analysis for GRO-seq is specialized and represents a hybrid of analysis strategies between RNA-seq and ChIP-seq
  - Reads mapped like ChIP-seq (shouldn't have splicing in nascent RNA)
  - *De novo* transcript identifiction
  - *De novo* TSS identification for 5'GRO-seq
  - Quantification of expression levels at genes, enhancers, etc.
- Some tools for GRO-seq analysis:
  - NOTE: Most of GRO-seq analysis has been done with custom tools or patchworks of general NGS analysis tools (i.e. bedTools)
  - HOMER (http://homer.ucsd.edu/homer/)
  - groHMM (http://www.bioconductor.org/packages/release/bioc/html/groHMM.html
  - dREG (https://github.com/Danko-Lab/dREG)

Part II: Measuring 3D Chromatin Conformation in the Genome with Hi-C

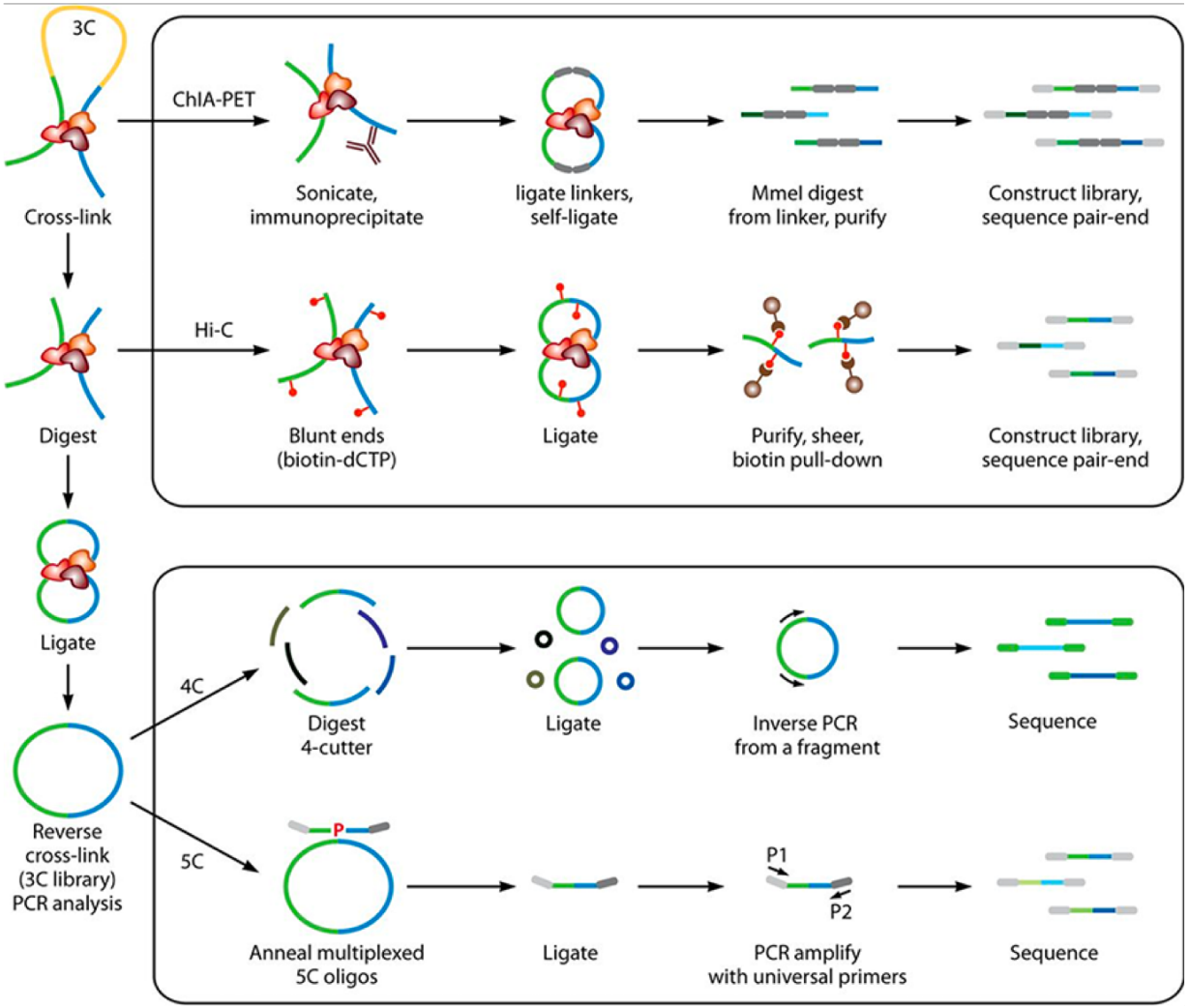
- Proximity ligation based technology overview
  - 3C, 4C, 5C, Hi-C, capture Hi-C, ChIA-PET
  - Basics of Analysis
- Application to measuring *in vivo* genome structure and function
  - Compartmentalization of the genome (active vs. inactive)
  - Topological associated domains (TADs)
  - Chromatin loops and CTCF
- Additional potential medical applications:
  - Haplotype phasing
  - Genome scaffold assembly
  - Detecting structural rearrangements in genomes

3C: Chromosome Conformation Capture



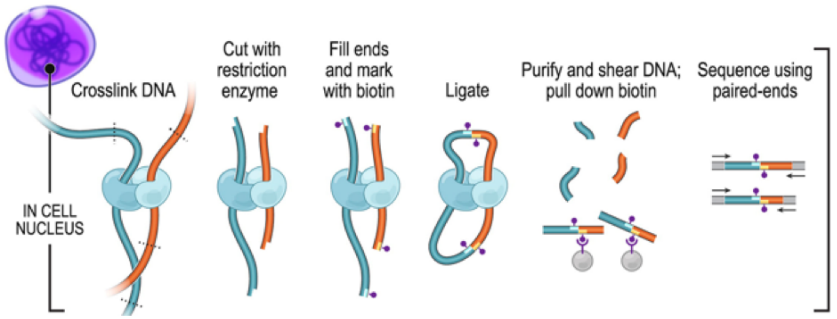
Original method Dekker et al. Science 2002
Picture from Stadhouders et al. Nat Methods 2013

3C, 4C, 5C, Hi-C, and ChIA-PET

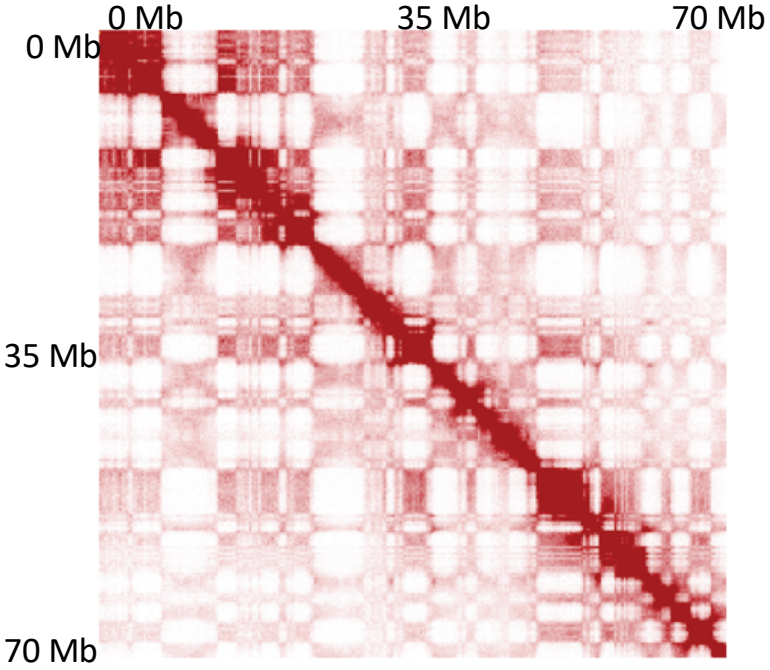


From Fraser et al. Micro and Molecular Biology Reviews 2015

Determining 3D genome structure (*in situ* Hi-C)



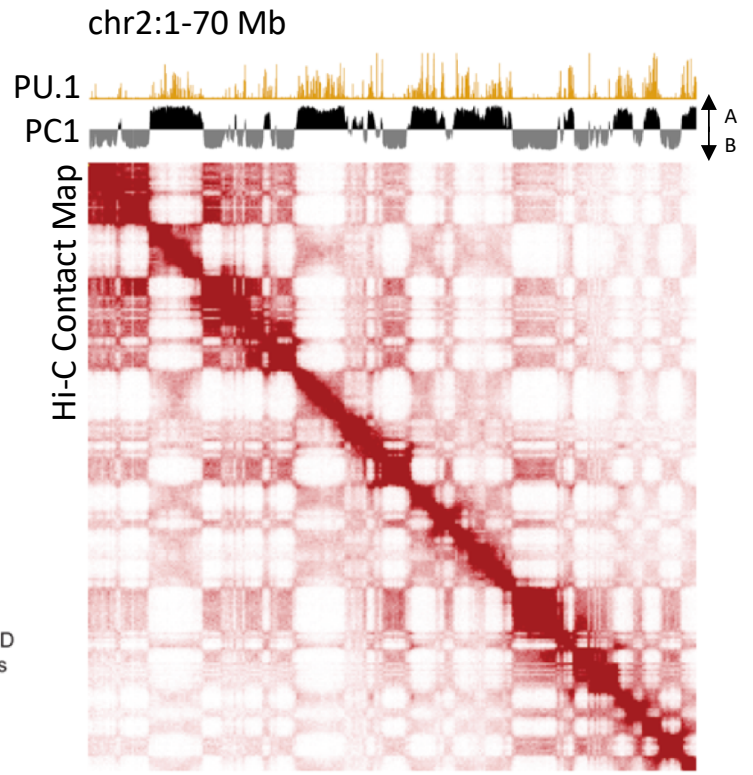
Hi-C Contact Map
chr2

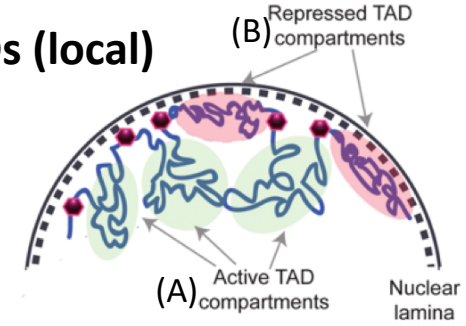


Hi-C method developed by Lieberman-Aiden *Science* 2009

Two major structural features of the genome:

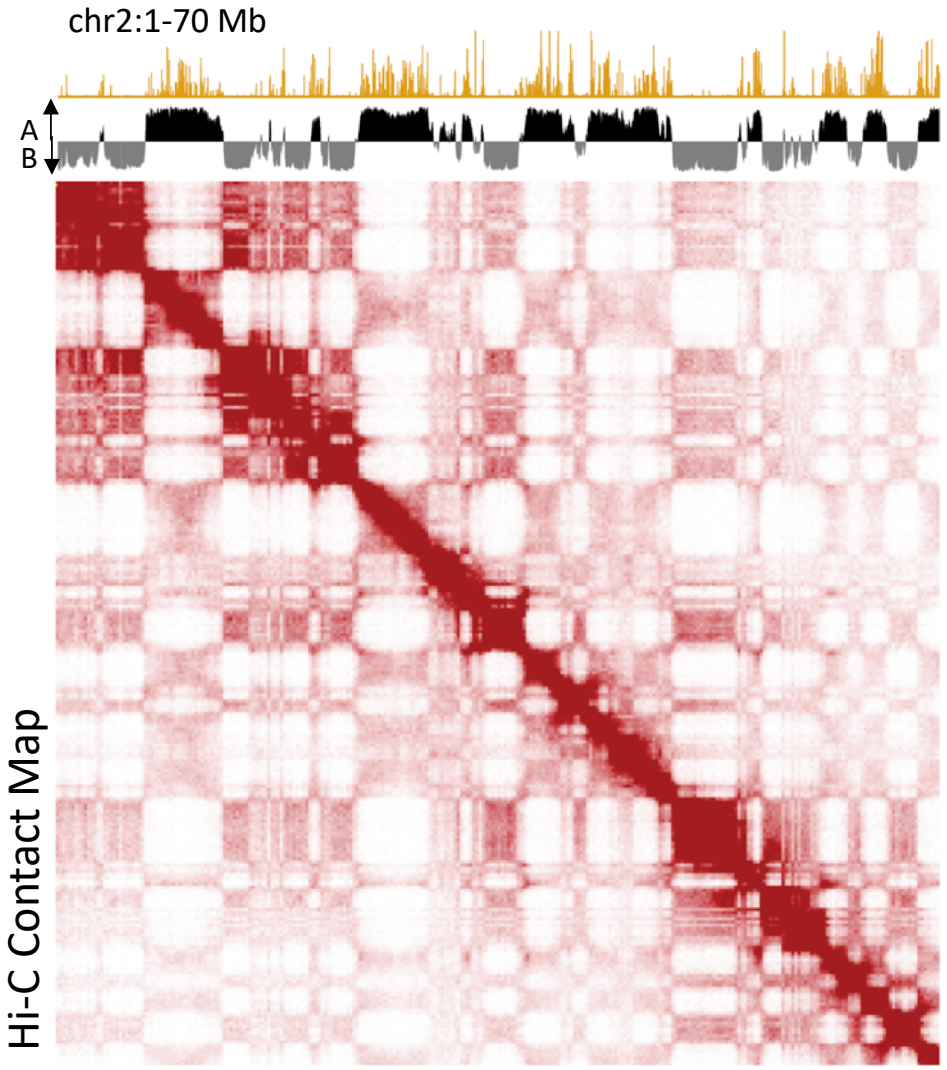
- **(1) Compartments (global)**
  - Chromatin is generally separated into A (active) and B (inactive) compartments
    - A compartment: permissive chromatin, active, luminal, bound by TFs
    - B compartment: inert chromatin, peripheral/nuclear lamina, minimal TF binding
  - Highly correlated with transcription
  - Responsible for the “checkerboard” pattern seen in Hi-C data
  - PCA analysis often used to determine compartments
- **(2) Loops/Domains/TADs (local)**





Compartment concept introduced by Lieberman-Aiden *Science* 2009

TF binding is largely restricted to the active compartment



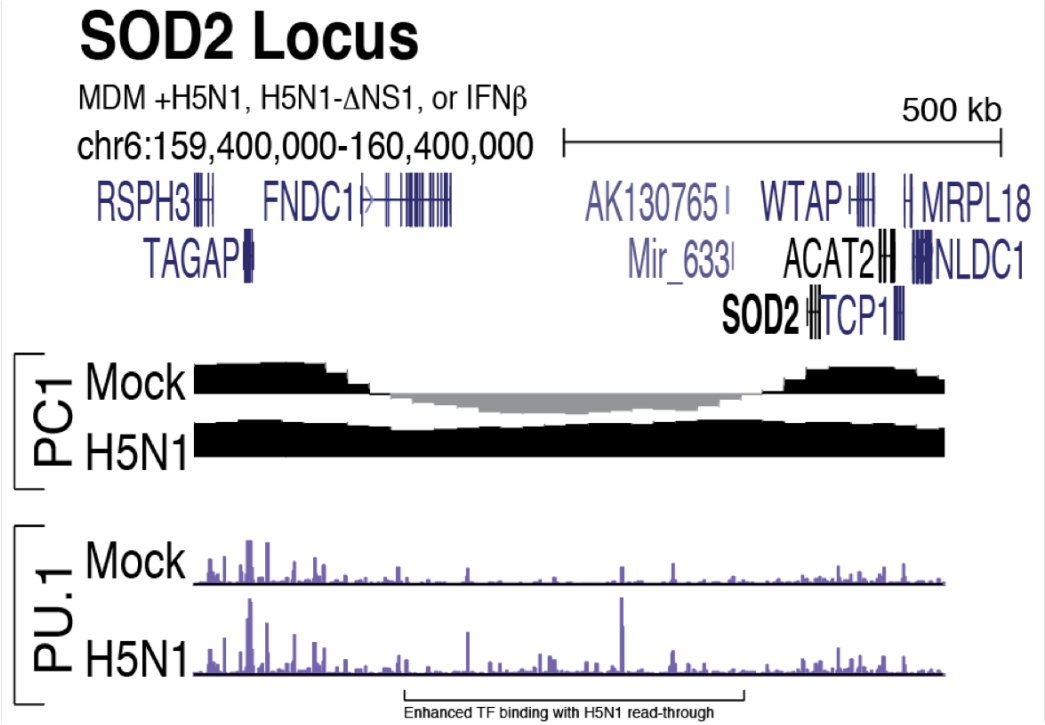
PU.1 Binding (TF)

PC1 (compartment analysis)

TF binding, transcription, Open chromatin, activation Histone marks, etc. are all preferentially located in the A compartment

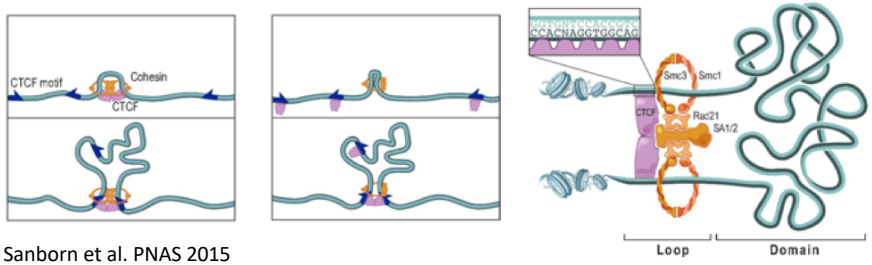
The B compartment is usually associated with the nuclear lamin and is refactory to TF binding

When compartments change, TF binding is often enhanced.

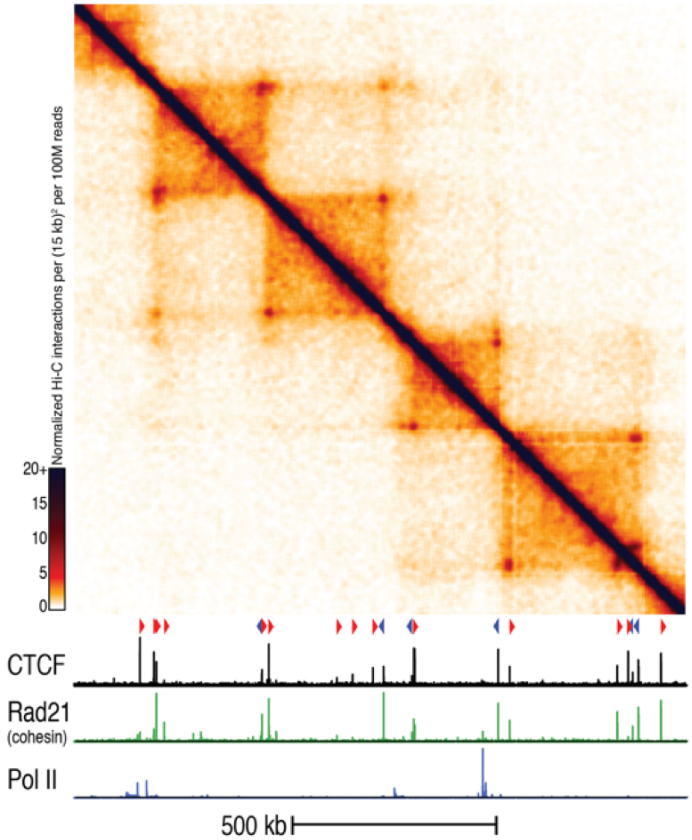


Two major structural features of the genome:

- **(1) Compartments (global)**
- **(2) Loops/Domains/TADs (local)**
  - The transcription factor CTCF creates anchor points for cohesin to create loop domains
  - The DNA motif bound by CTCF is directional: loops are preferentially formed between CTCF sites with convergent CTCF motifs
  - Actively transcribed regions and promoter/enhancers can also create loops, but CTCF mediated loops are the dominate features in the Hi-C data
  - Domains/TADs are thought to encapsulate groups of co-regulated transcripts/regulatory elements
  - Relatively stable between cell types
  - Loops are most likely formed by cohesin extrusion:

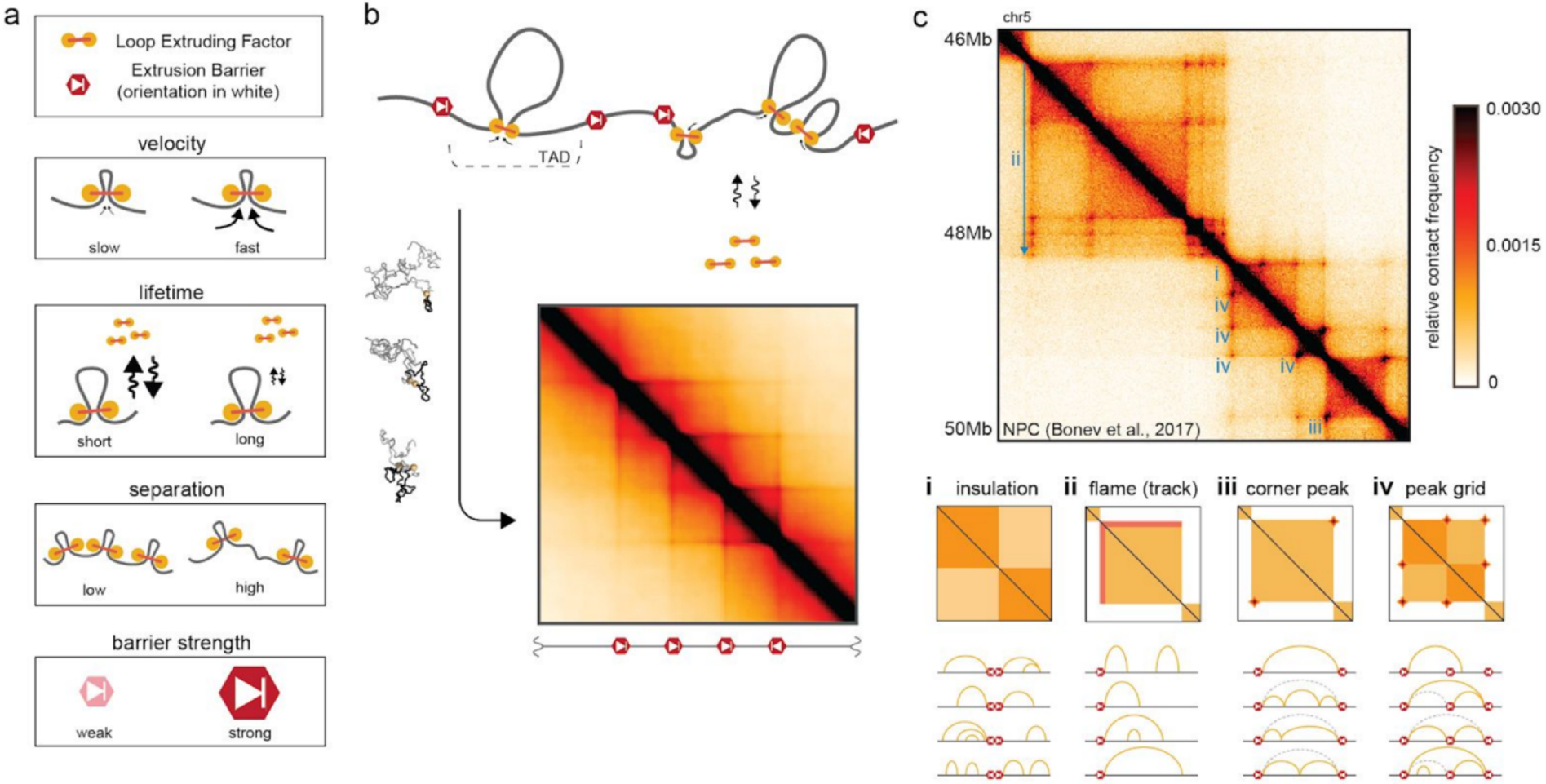


Sanborn et al. PNAS 2015



TAD concept introduced by Dixon et al. *Nature* 2012
CTCF motif directionality by Rao et al. *Cell* 2014

Cohesin loop extrusion model

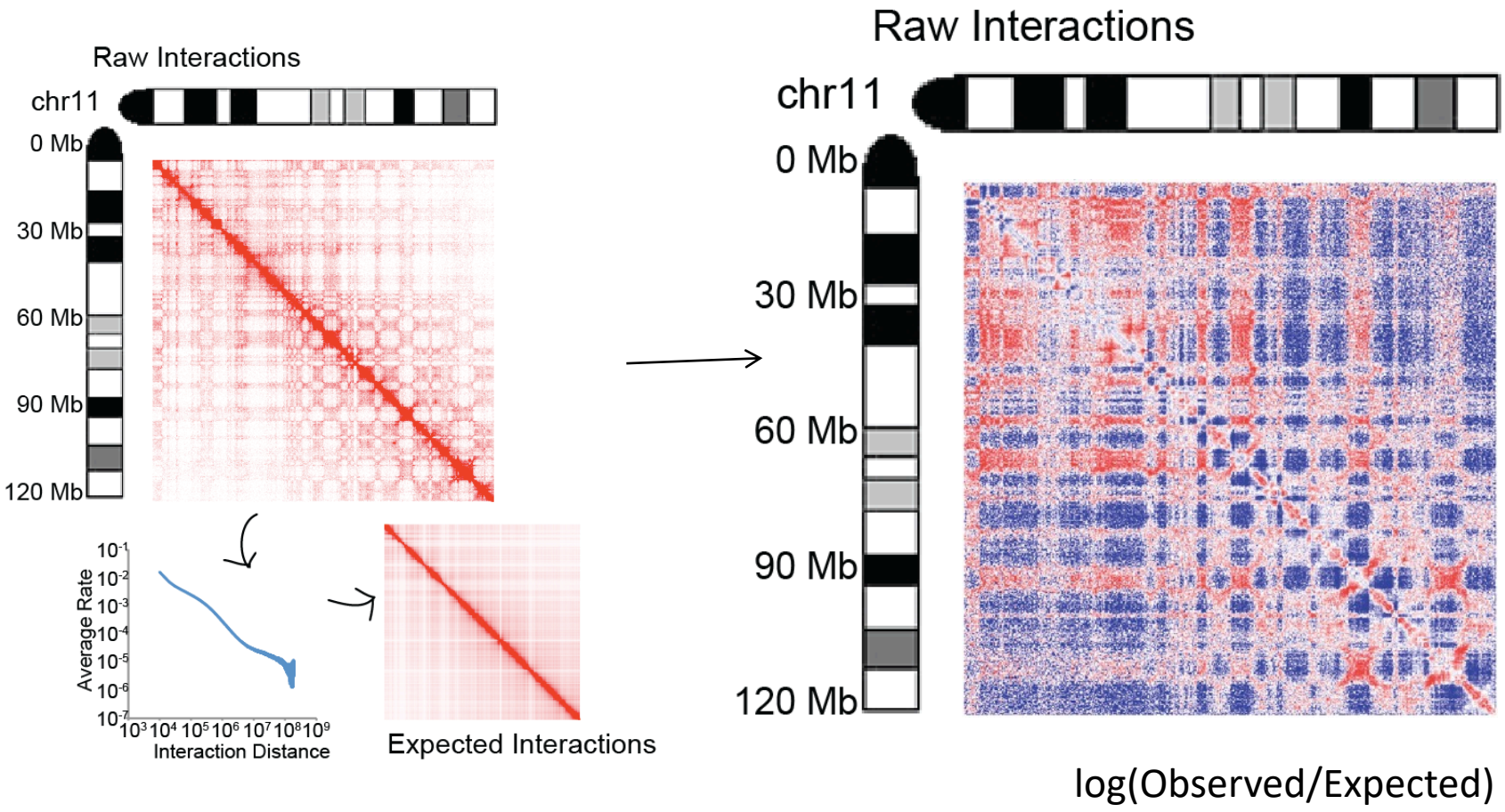


Fudenburg et al. Biorxiv 2018

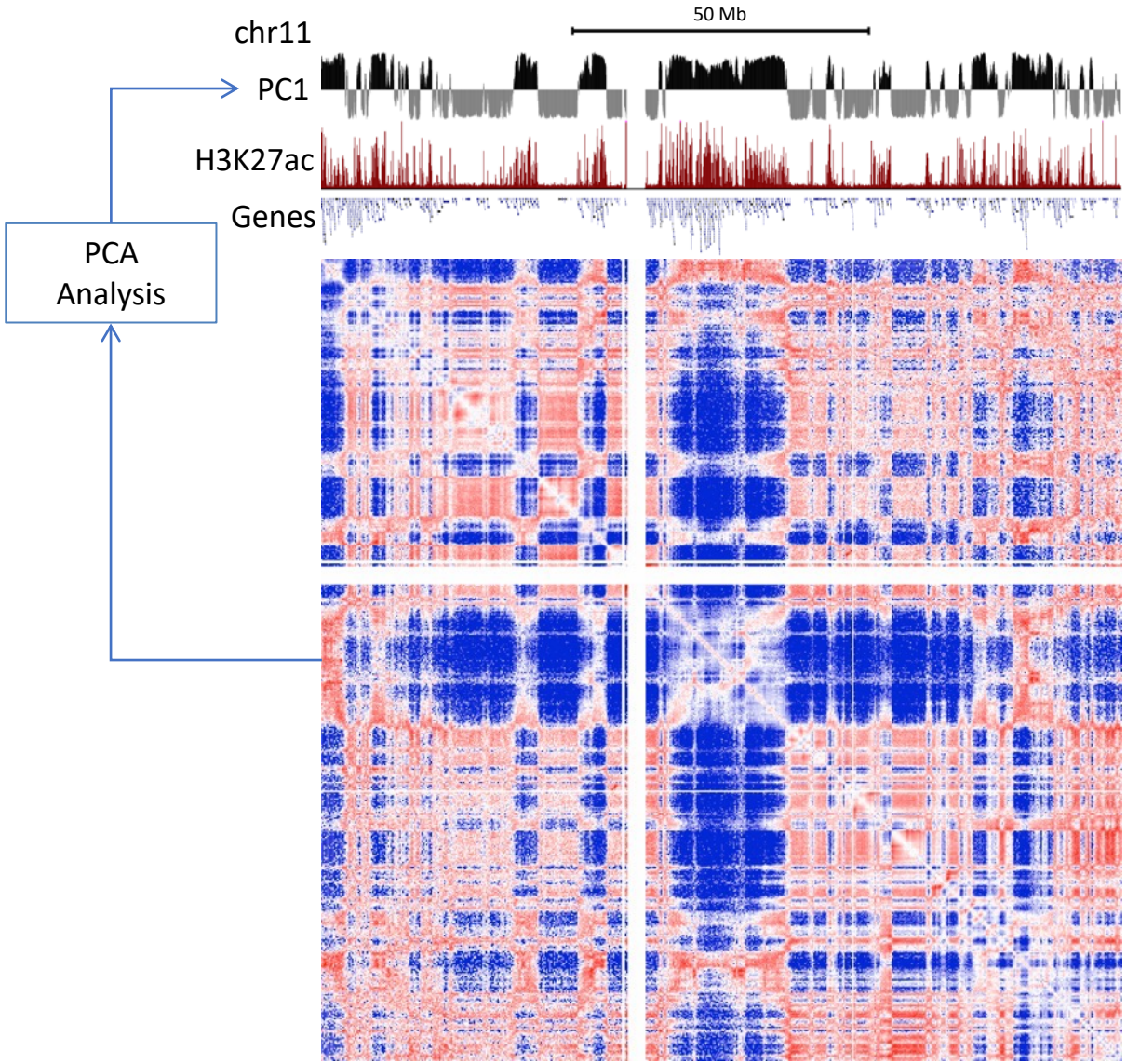
Challenges for Hi-C data analysis

- Normalizing interaction counts
  - Sequence coverage
  - Bias from restriction sites
  - GC-content bias
  - Interactions scale with distance between loci
- Specialized types of analysis:
  - Clustering/PCA analysis to partition genomic compartments
  - Identification of topological domains (TAD)
  - Finding anchor points/significant Interactions

Normalization of Hi-C Data

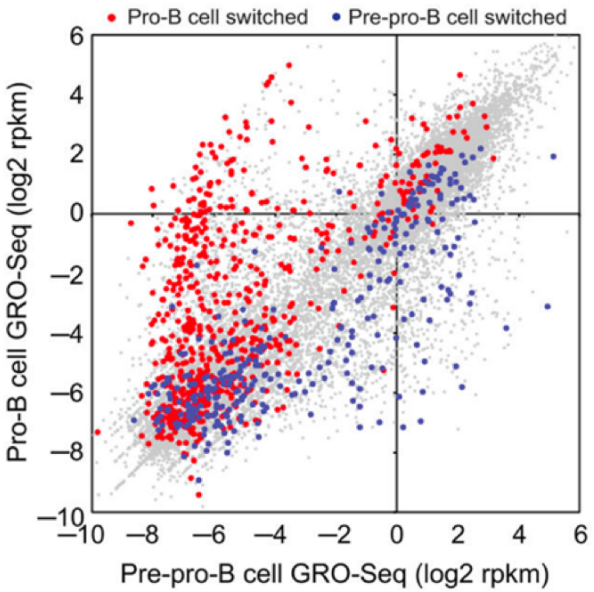


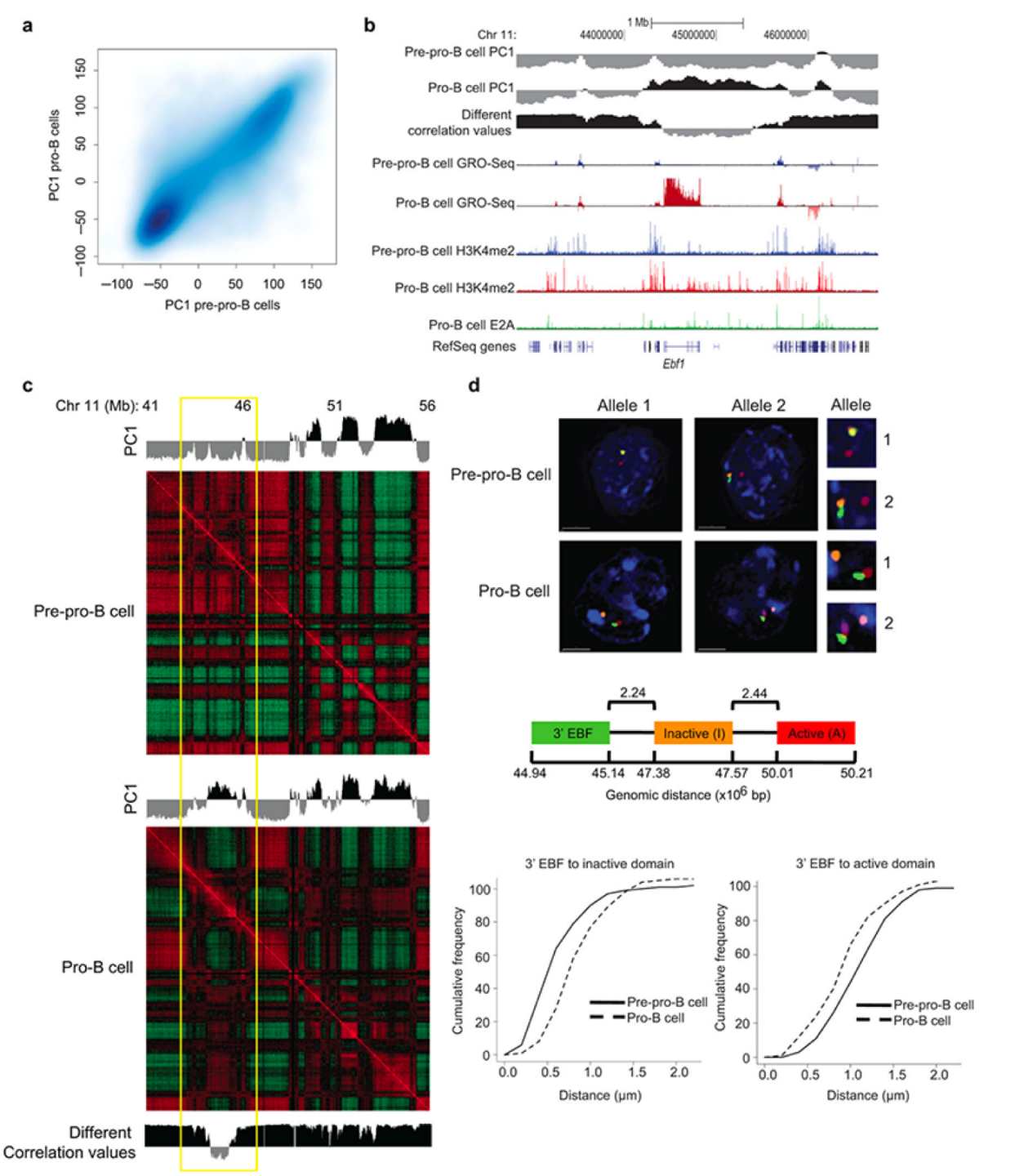
Identification of Genomic Compartments with Principal Component Analysis



- Mammalian genomes naturally partition into two major compartment (A & B), segregating active from inactive regions of the genome.
- Potentially more sub-compartments with additional functions (usually requires deeper sequencing to easily identify)

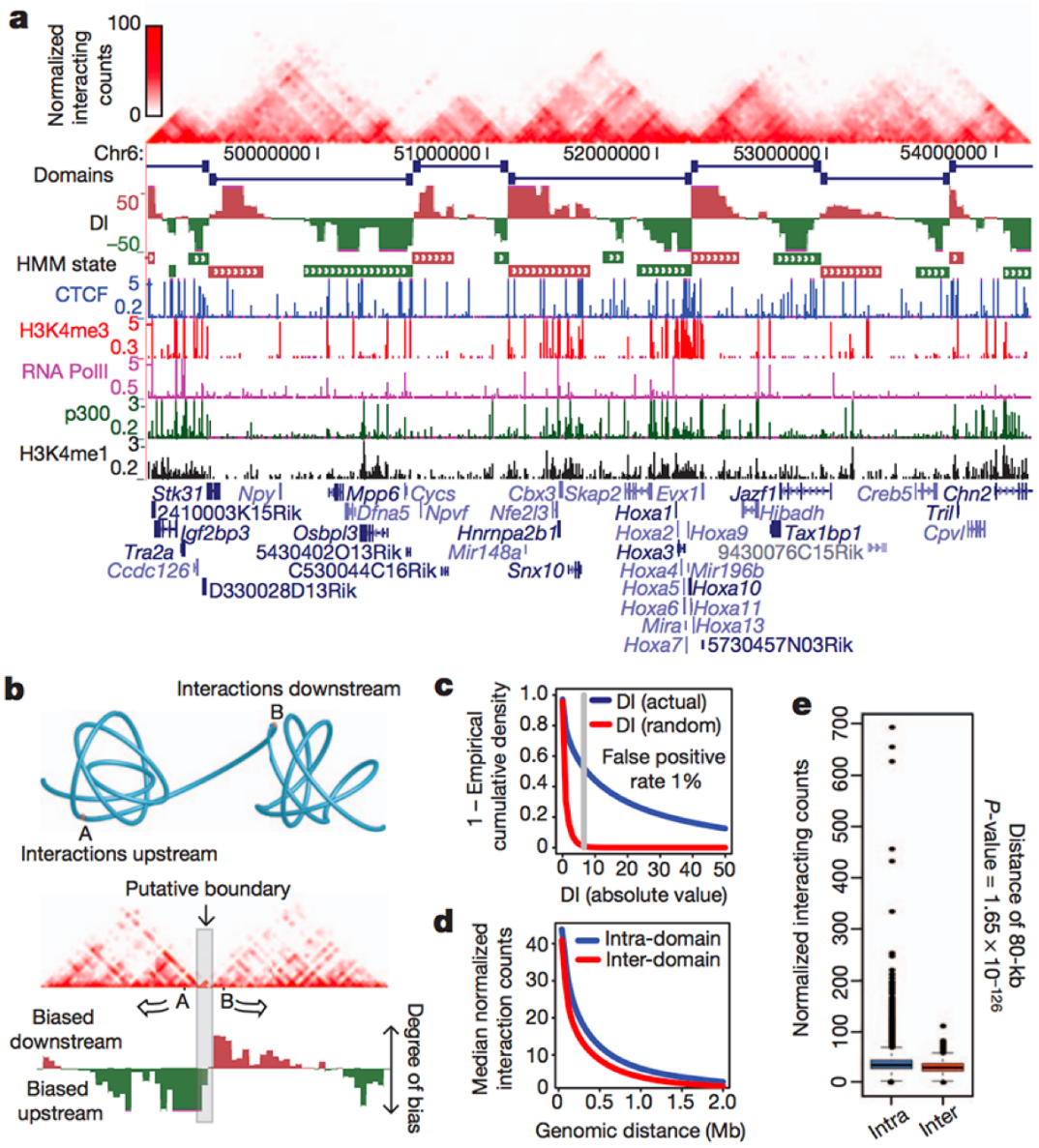
Compartments are regulated during differentiation





Lin et al. Nat. Immunology 2013

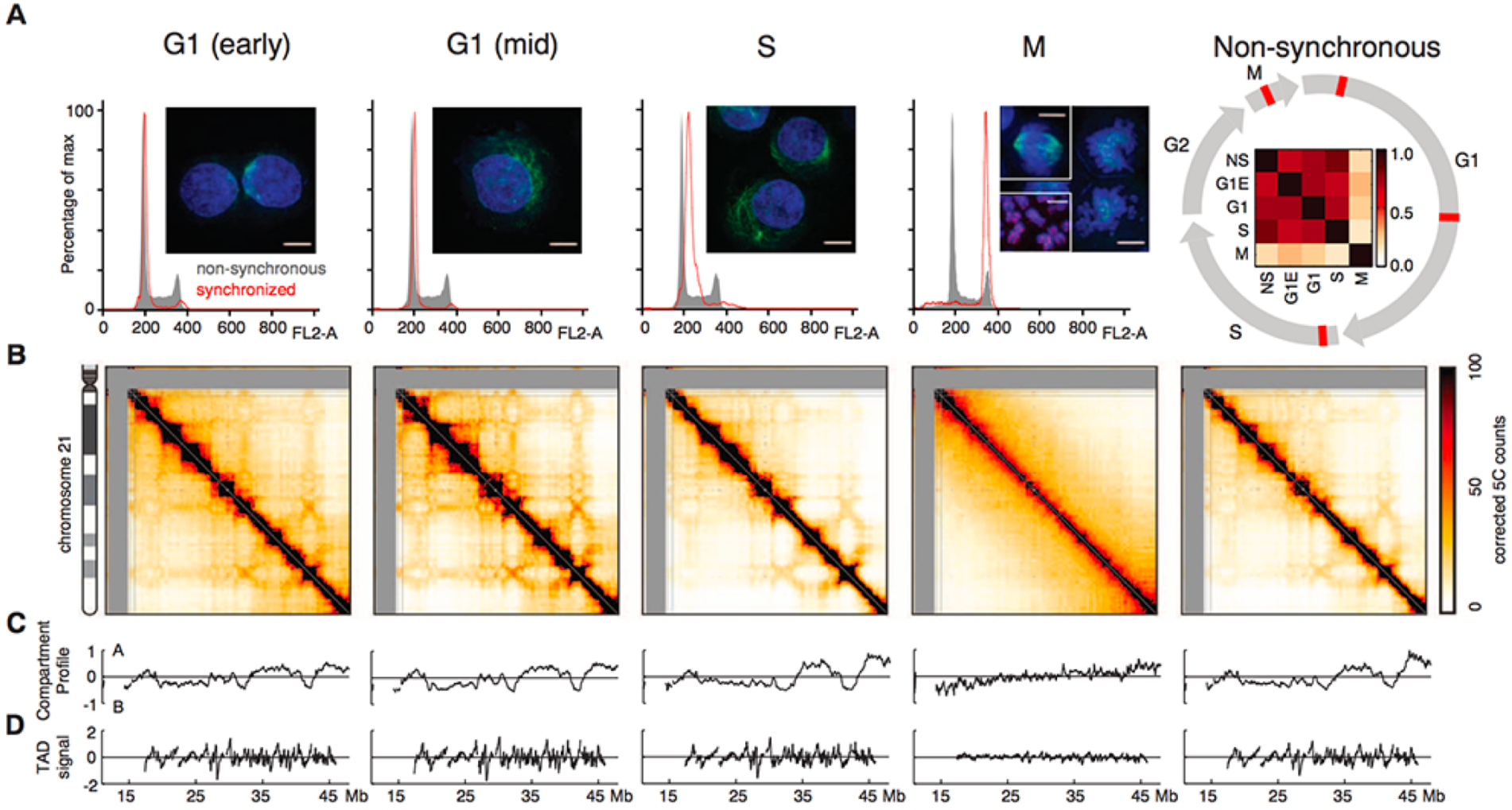
Topological Associated Domains (TADs)



- Identification of tightly interacting domains along the chromosome
- Genes within these regions are often coordinately regulated

Dixon et al. Nature 2012

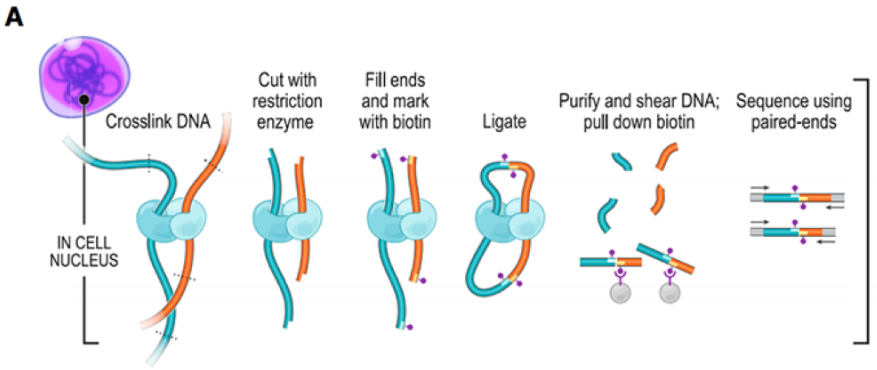
Compartments/TADs are lost during mitosis

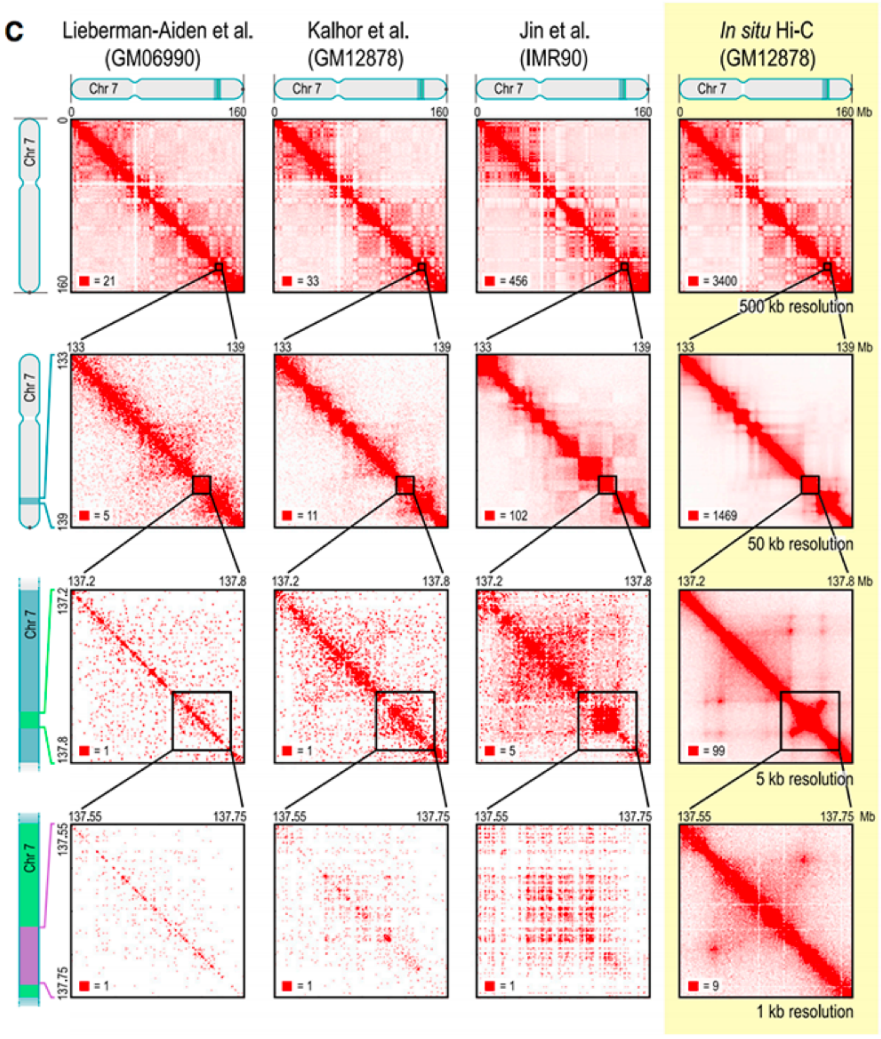


Naumova et al. Science 2013

Identifying Significant Interactions

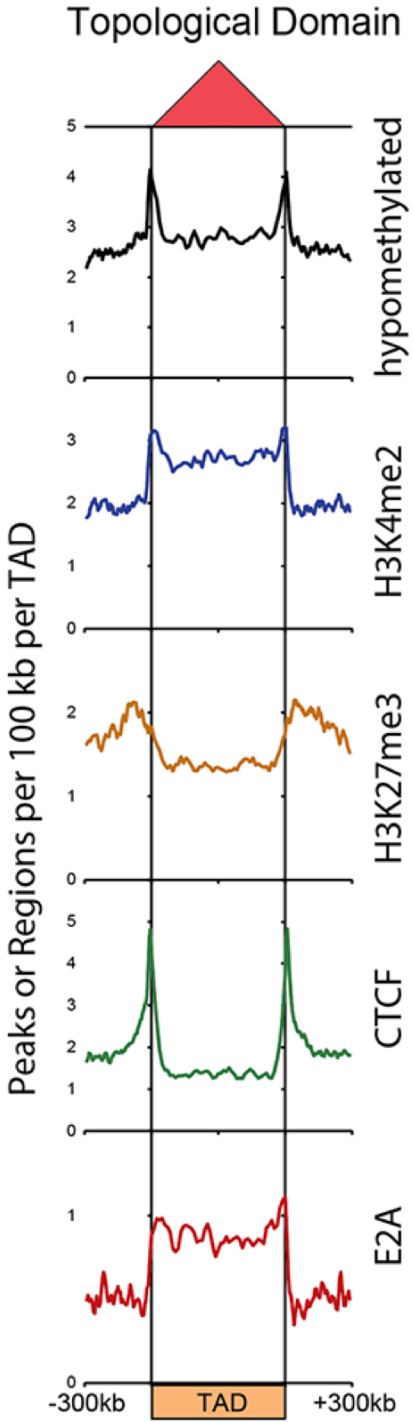
- Identify single contact points that have many more interactions than on average
  - Requires much deeper sequencing
  - Most are at TAD boundaries, not enhancer-promoter contacts

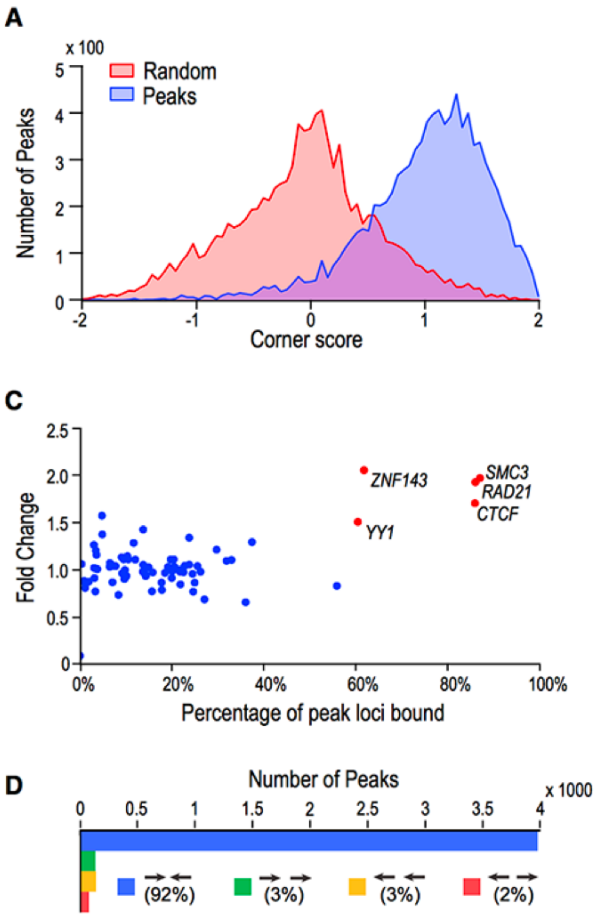


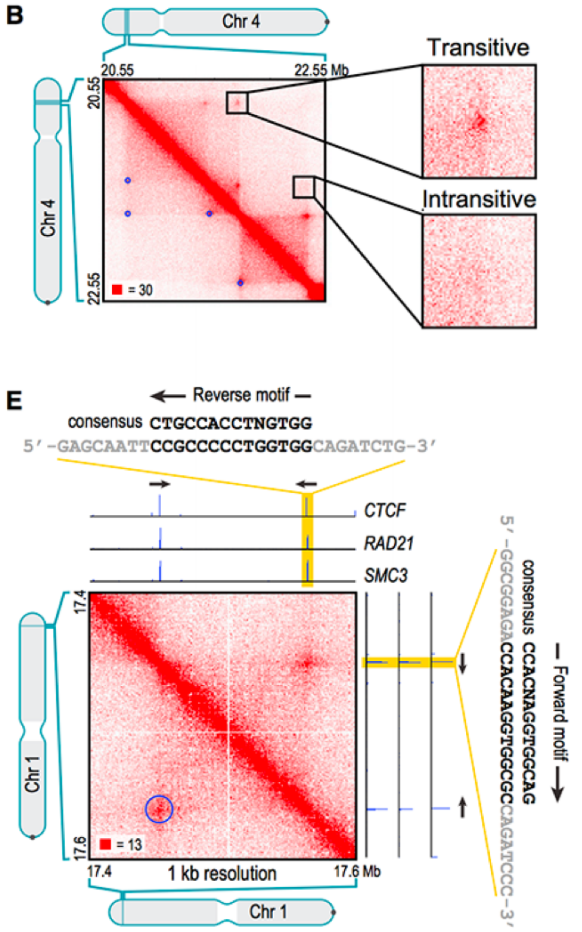


Rao et al. Cell 2014

CTCF – key transcription factor found at TAD boundaries at at significant Hi-C Interactions

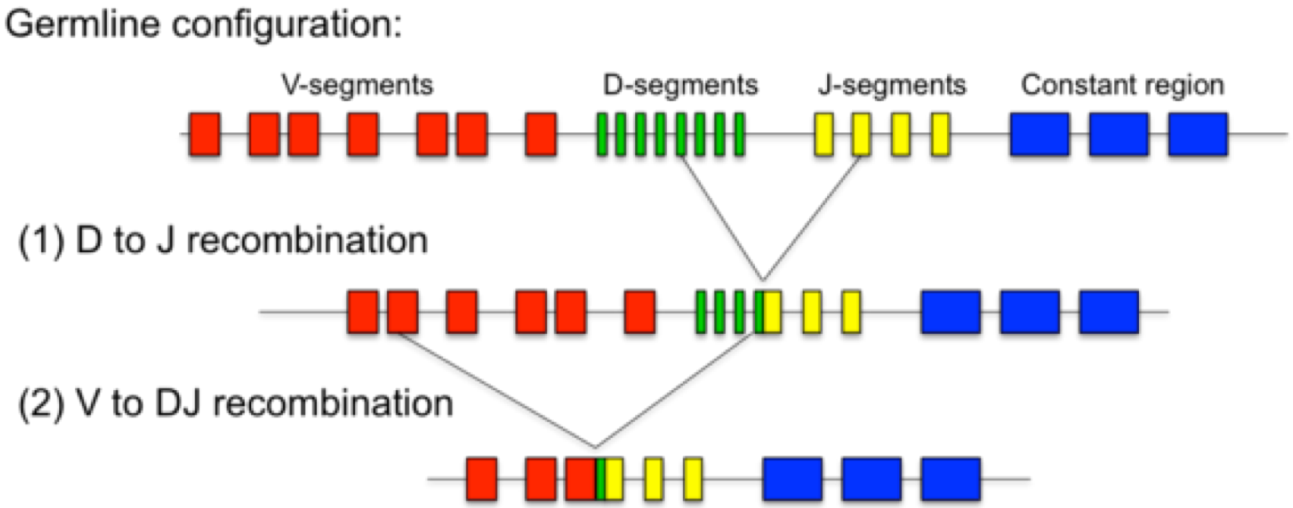


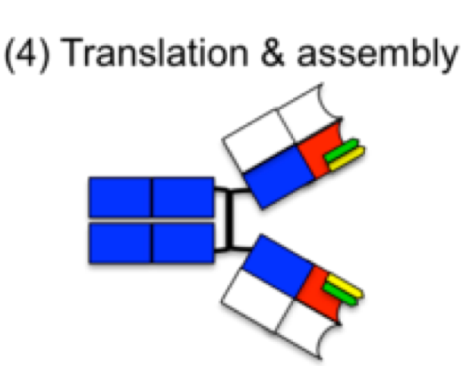




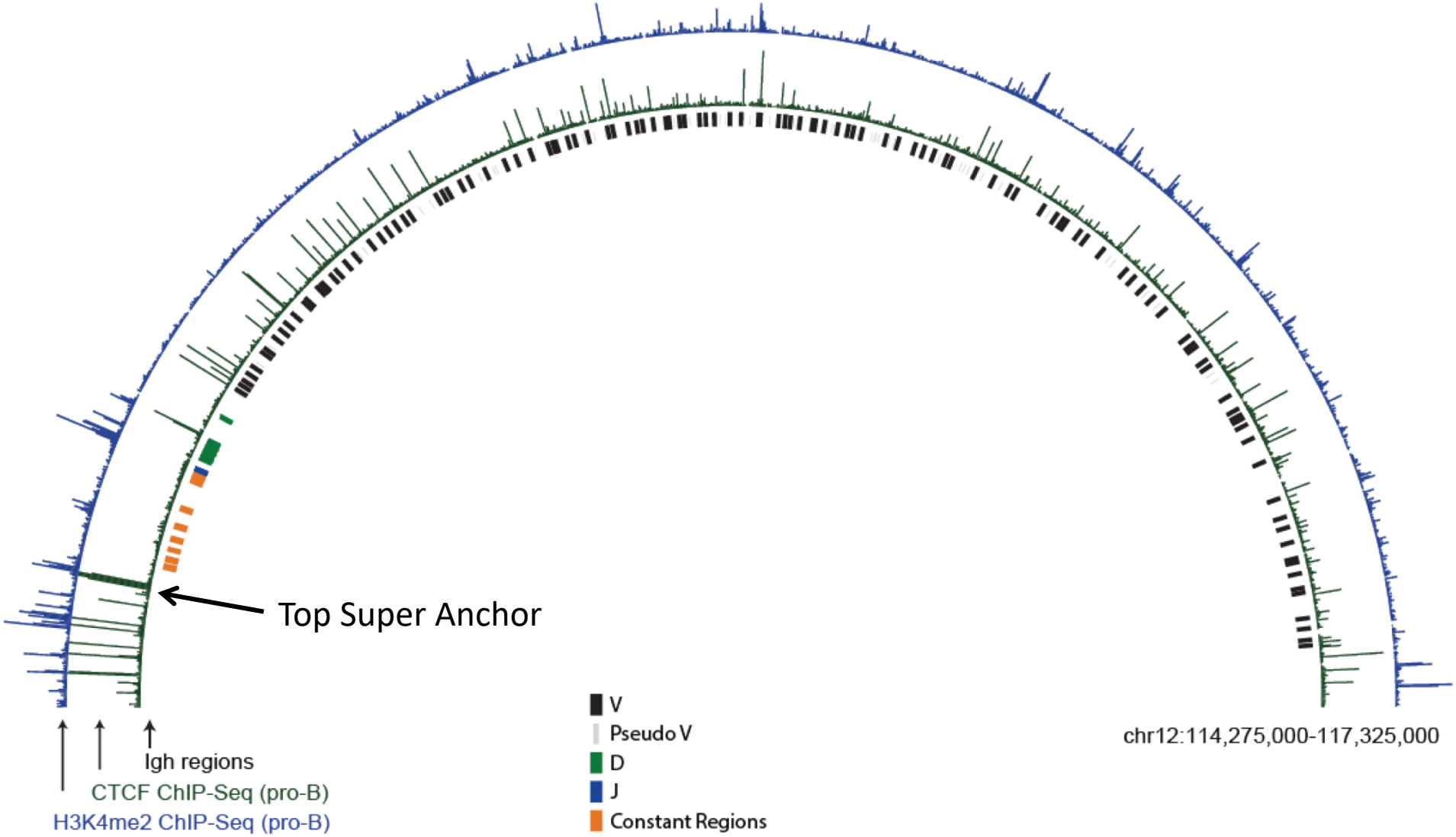
Benner et al. PNAS 2015, Rao et al. Cell 2014

Overview of Immunoglobulin Heavy Chain Locus

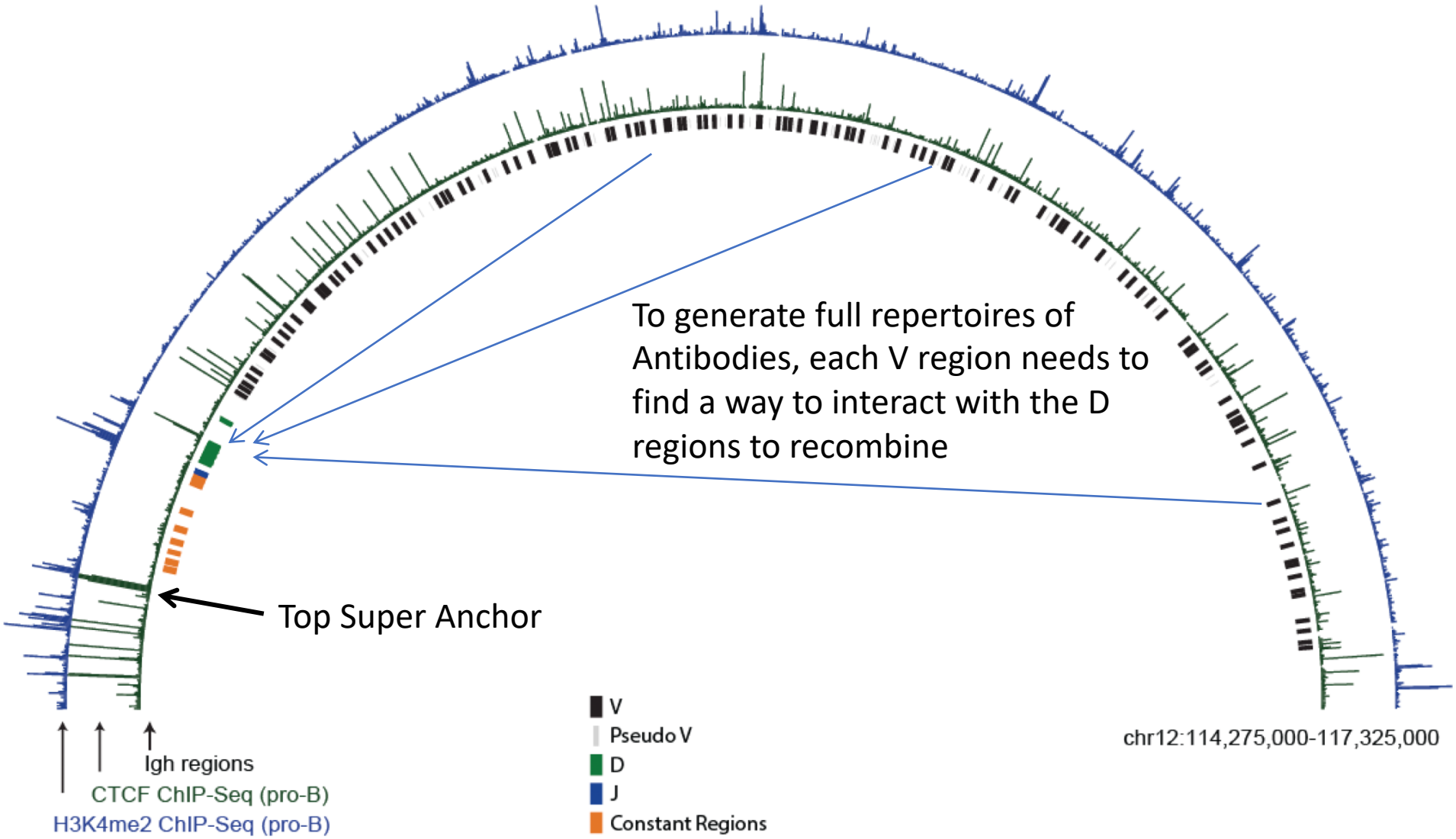




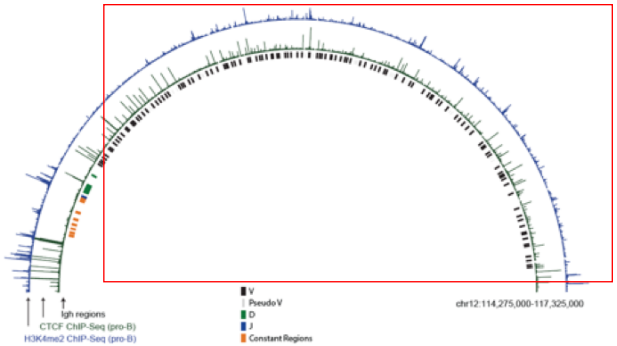
Igh Locus in the Genome (~3 Mb)

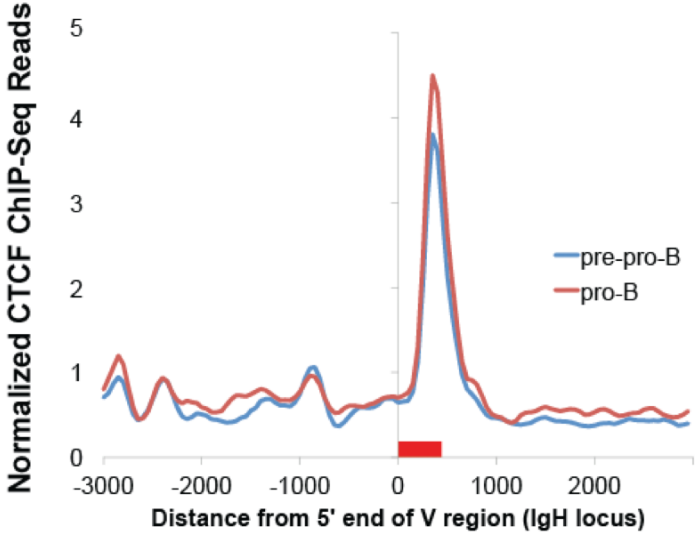


Igh Locus in the Genome (~3 Mb)



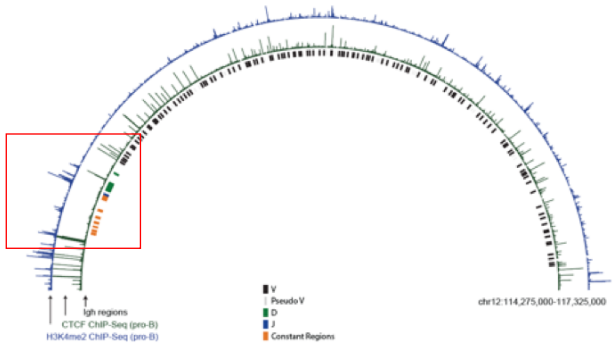
V regions in Igh locus are associated with CTCF sites

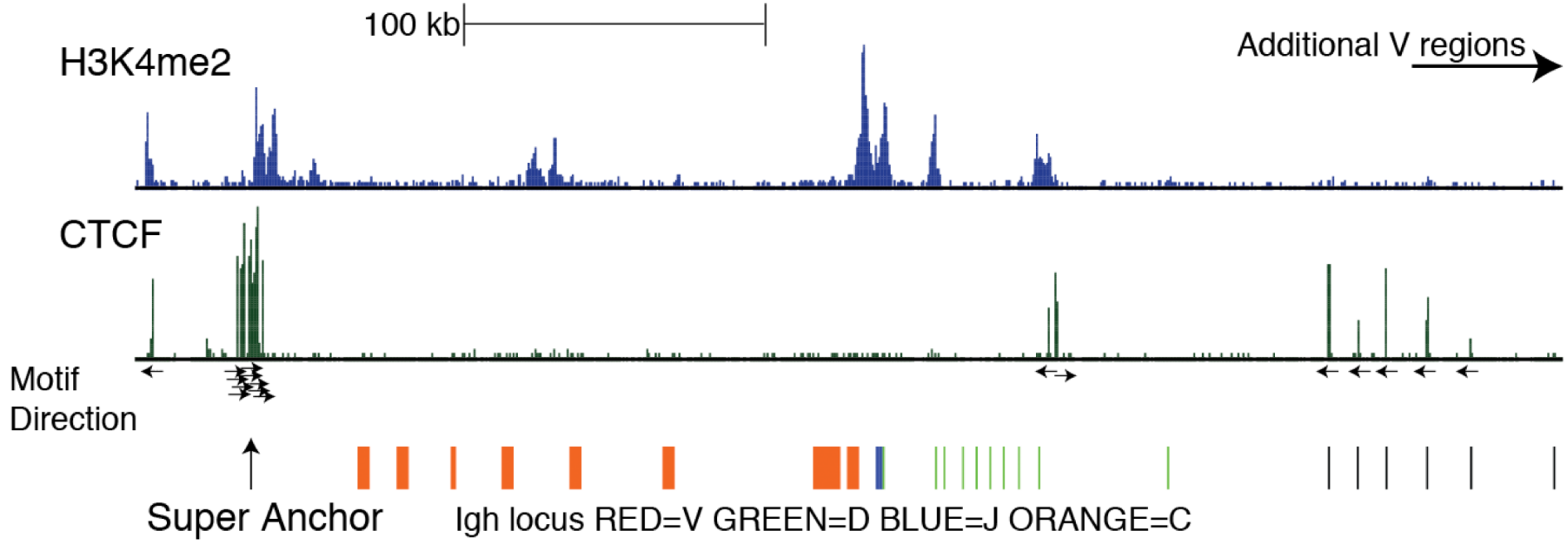




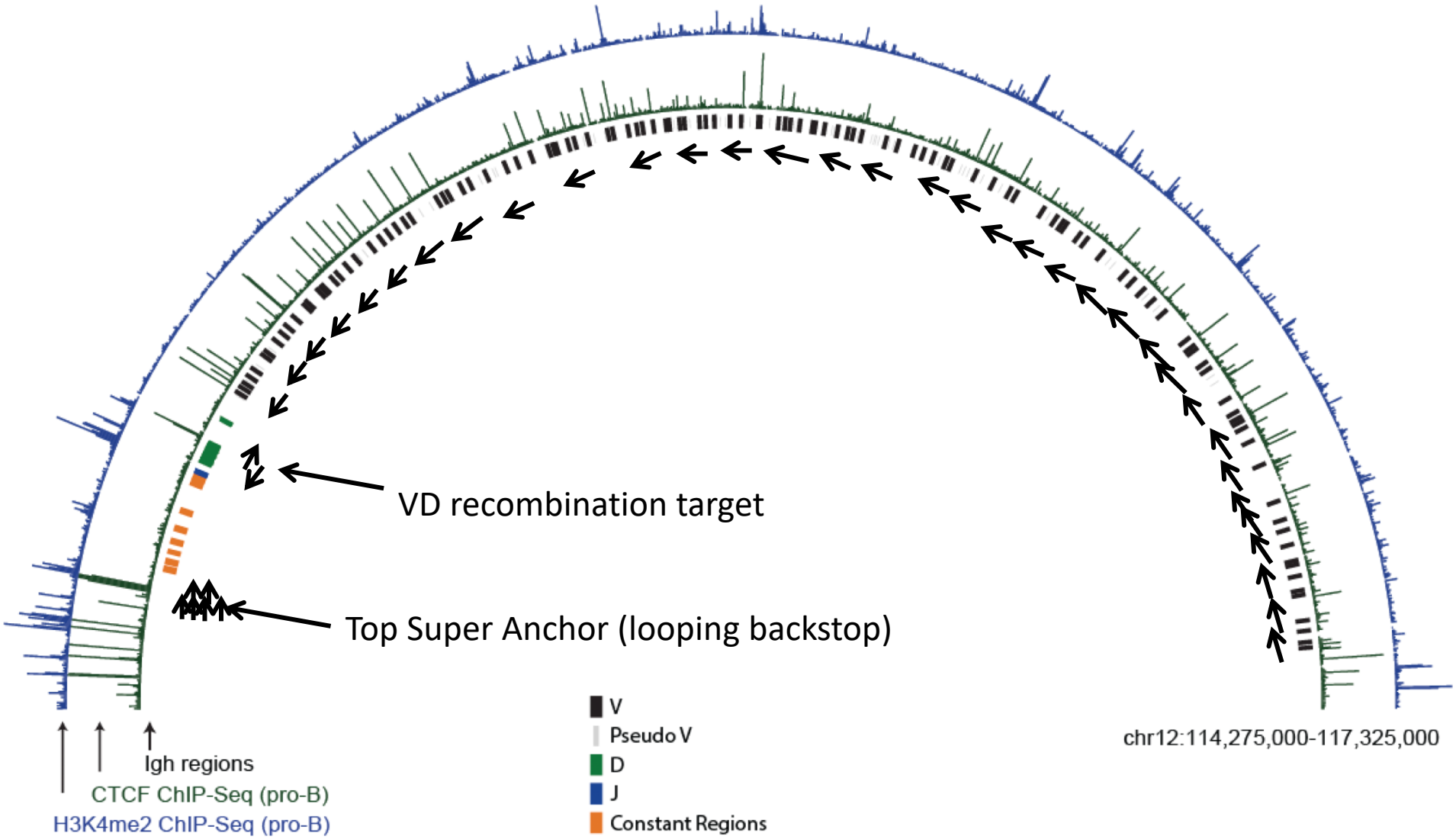
In addition, each CTCF site associated with V regions is in a consistent orientation

CTCF Orientation at D/J regions

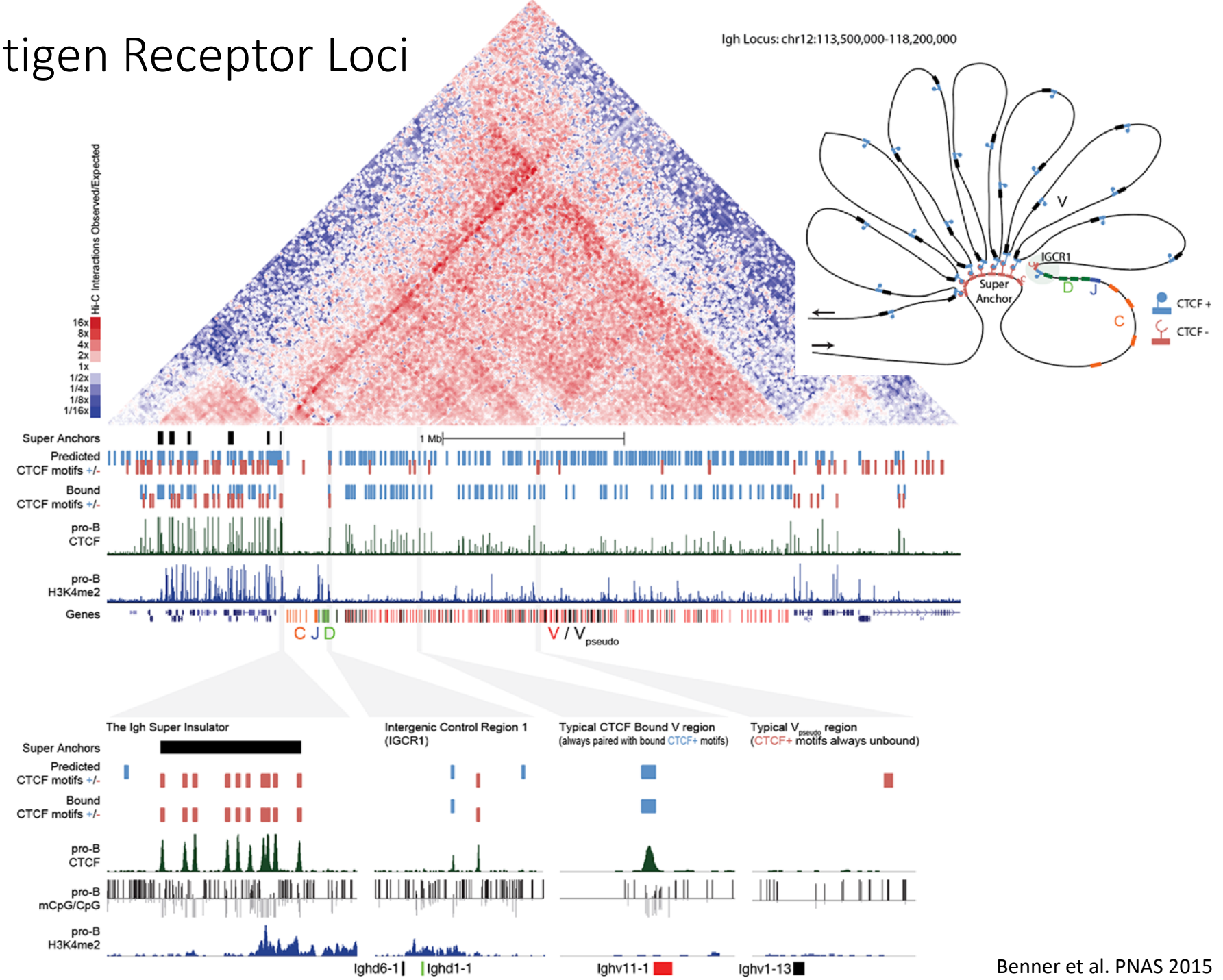




Igh Locus Model



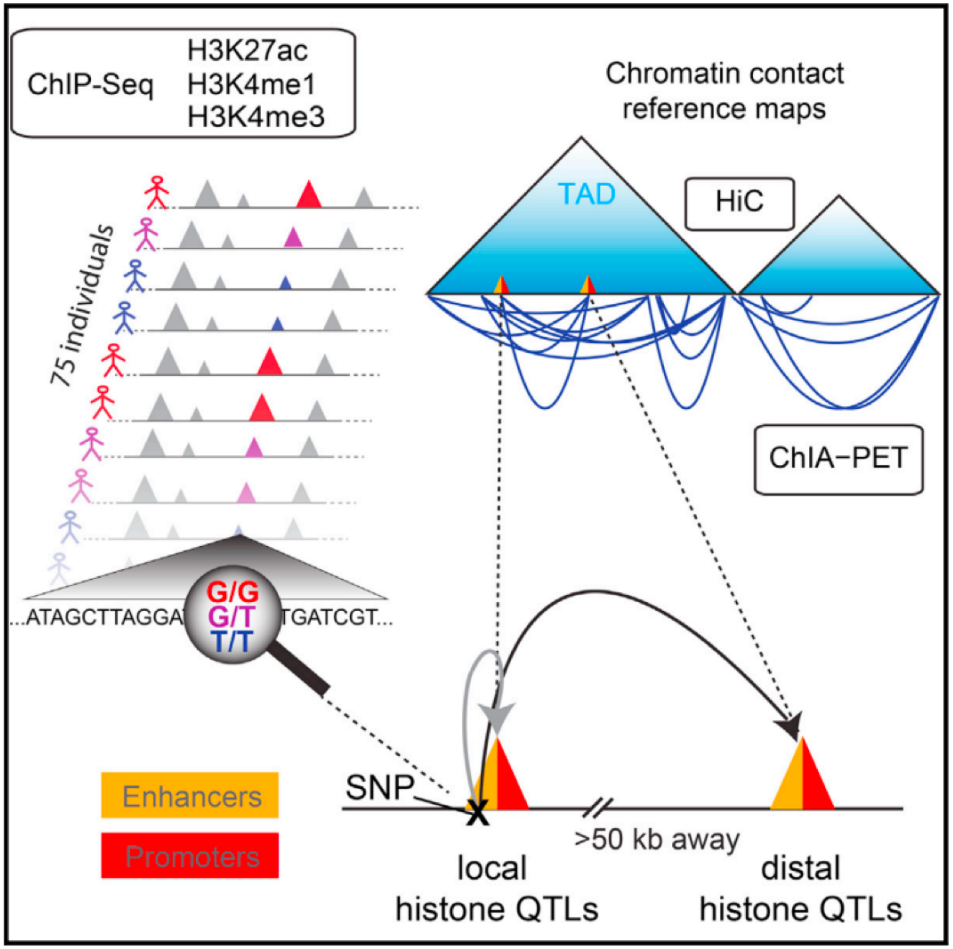
Antigen Receptor Loci



Benner et al. PNAS 2015

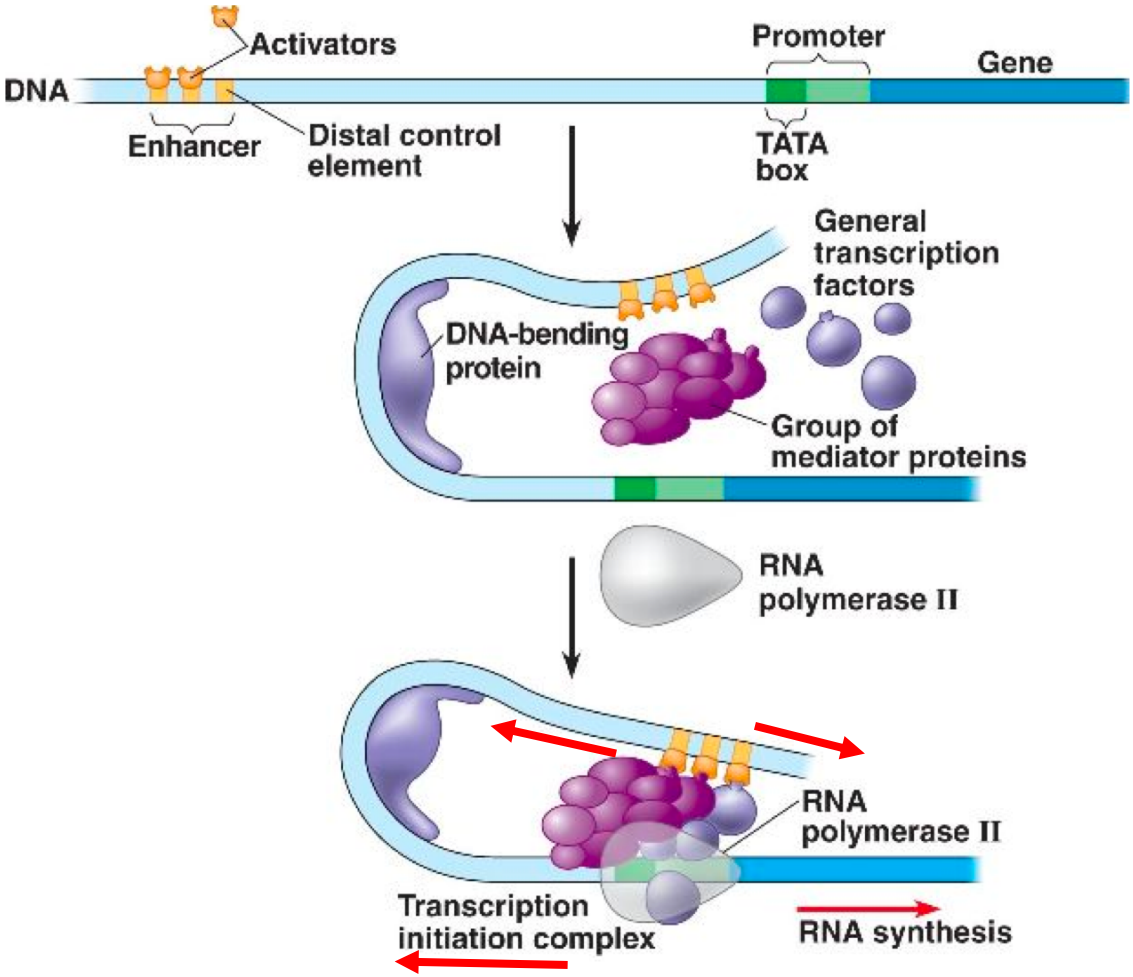
Use of chromatin structure in Medical applications

- Mostly used for basic research/mechanism studies so far
- Non-coding variant annotation (assignment to nearby regulated genes, etc.)
- Might be useful as a diagnostic, but that hasn't been done yet.



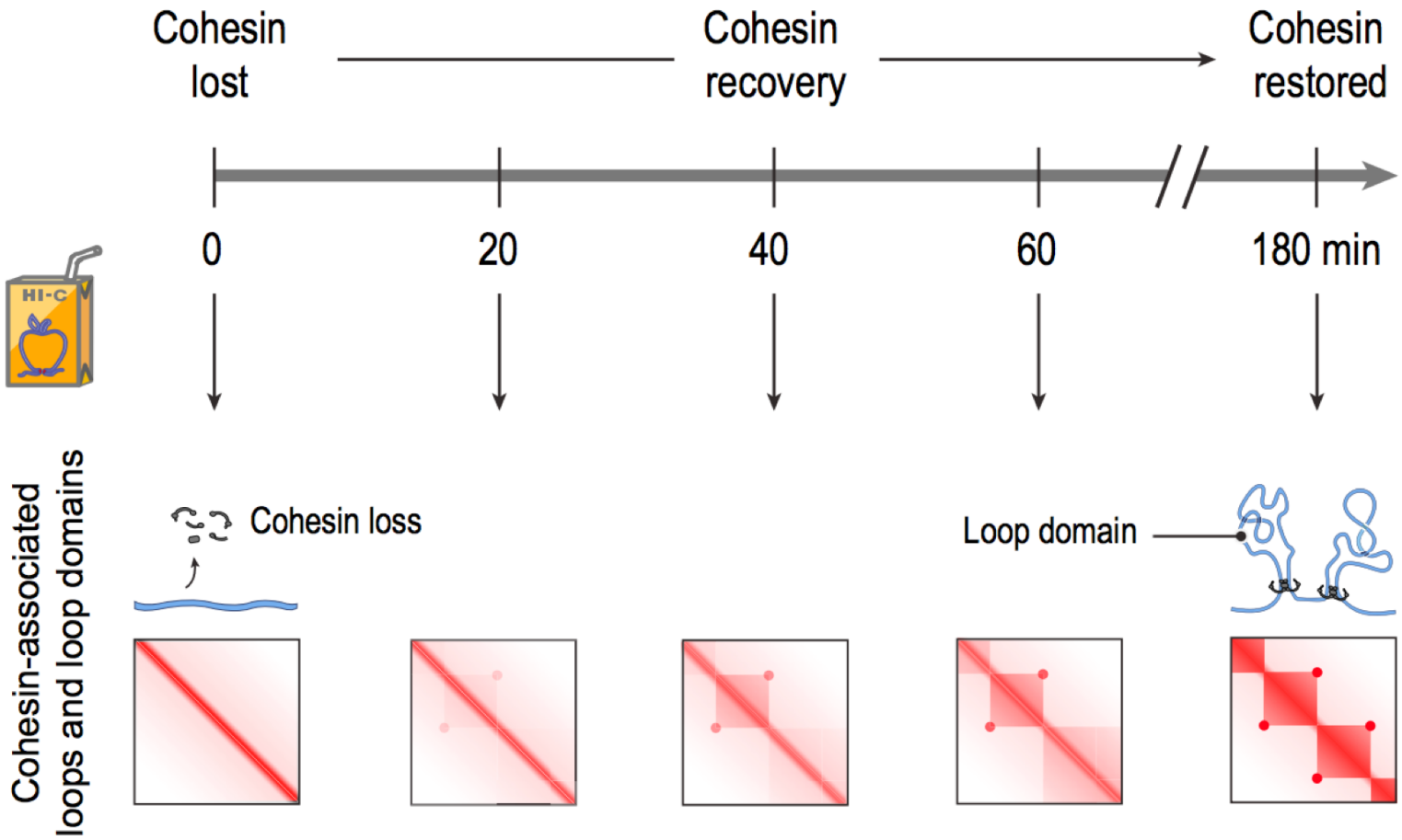
Grubert et al. Cell 2015

Commonly used model of transcription: Promoter-Enhancer looping



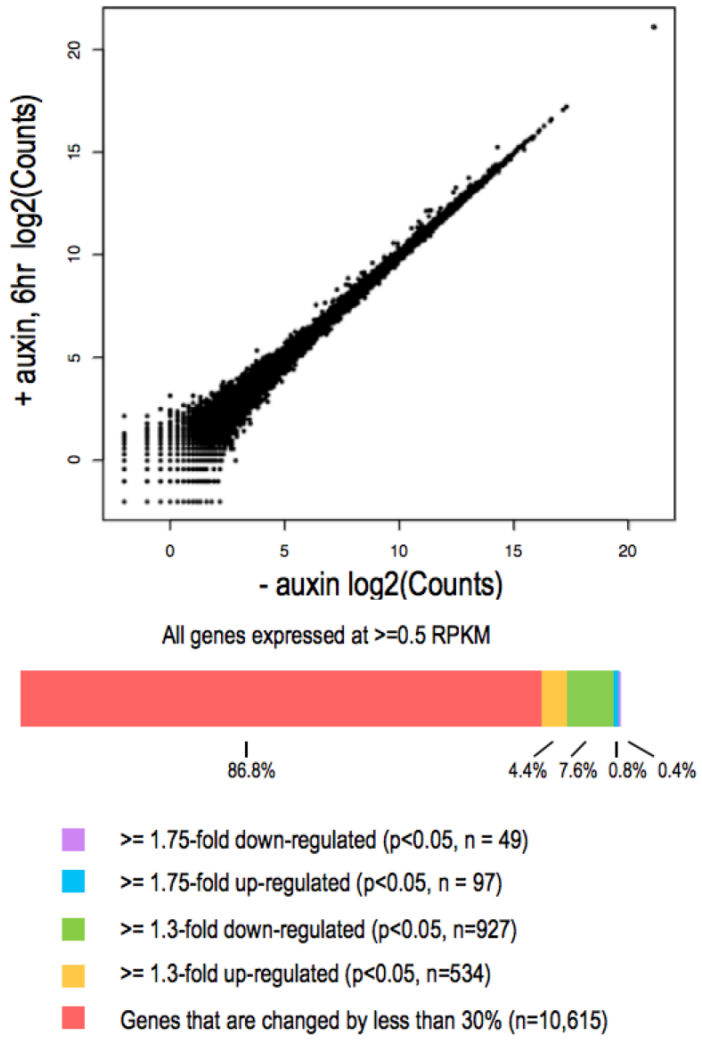
From Simone Silva http://flipper.diff.org/app/items/info/6731

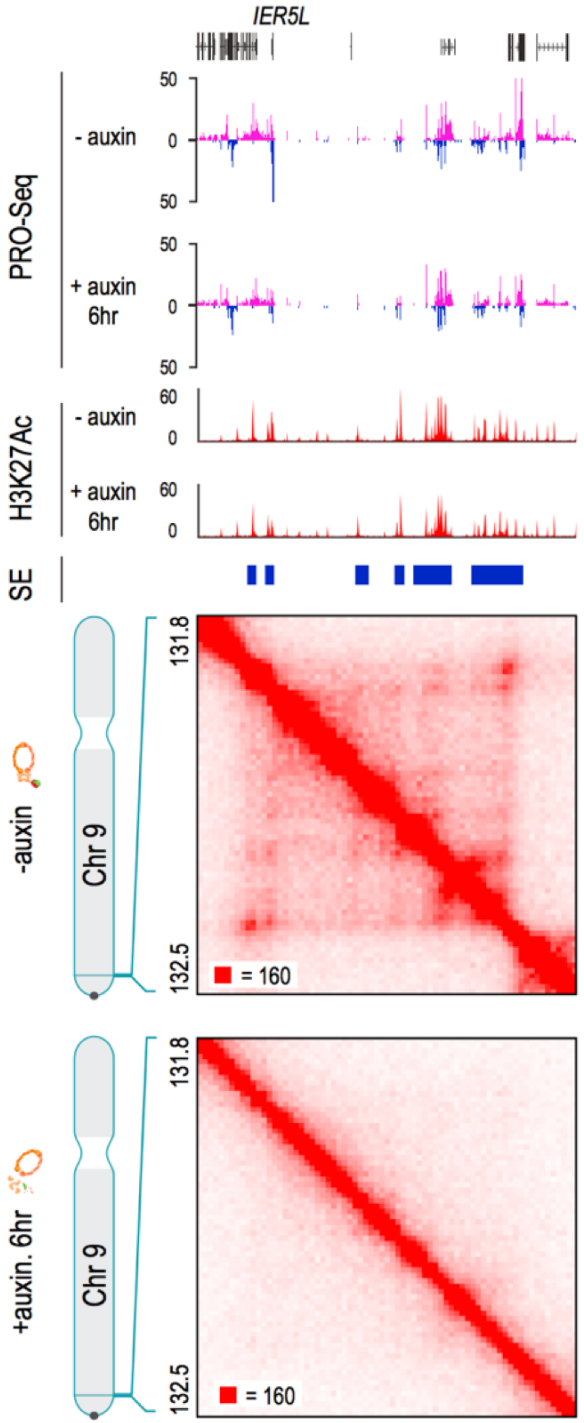
What happens when you eliminate all loops?



Rao et al. Cell 2017

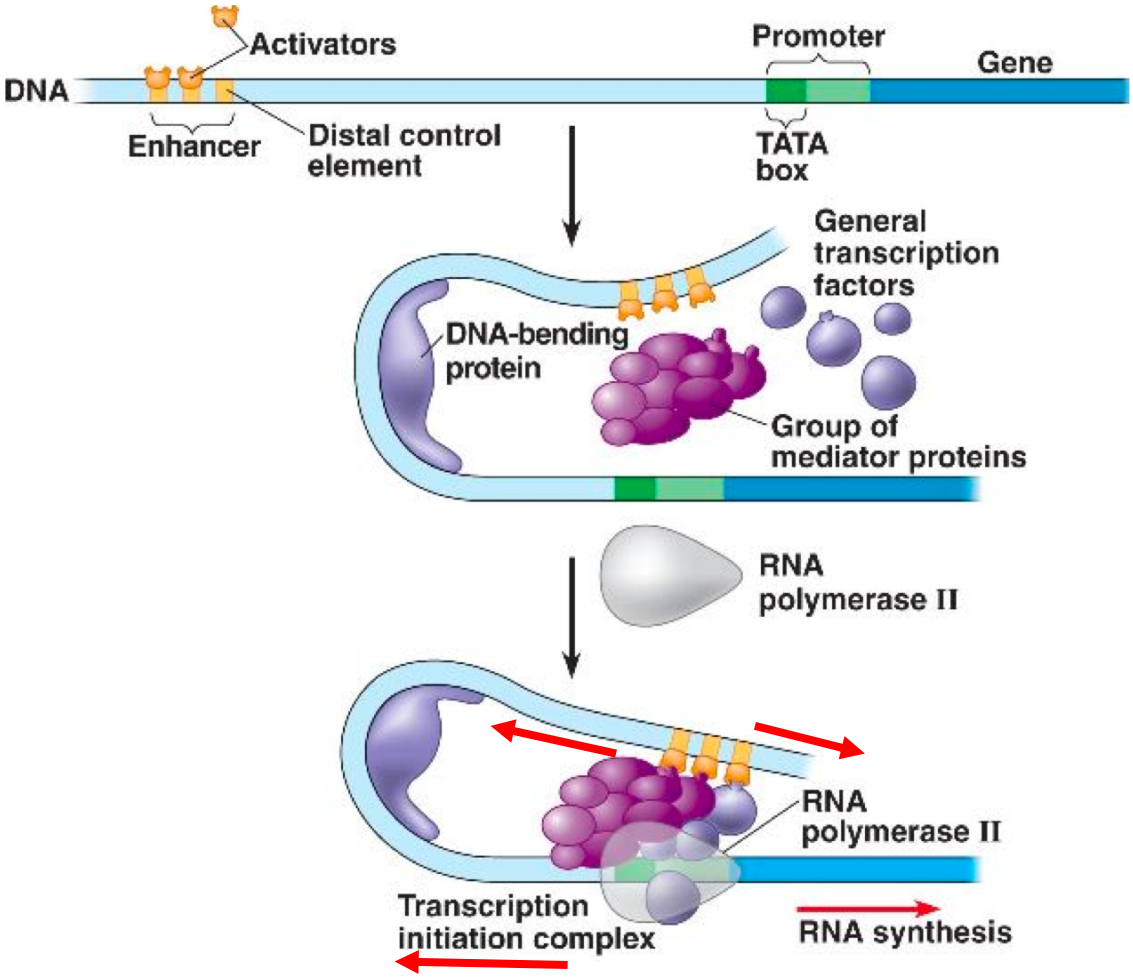
Cohesin loops don't seem to play a large role in gene expression levels





Rao et al. Cell 2017

So… is this right?



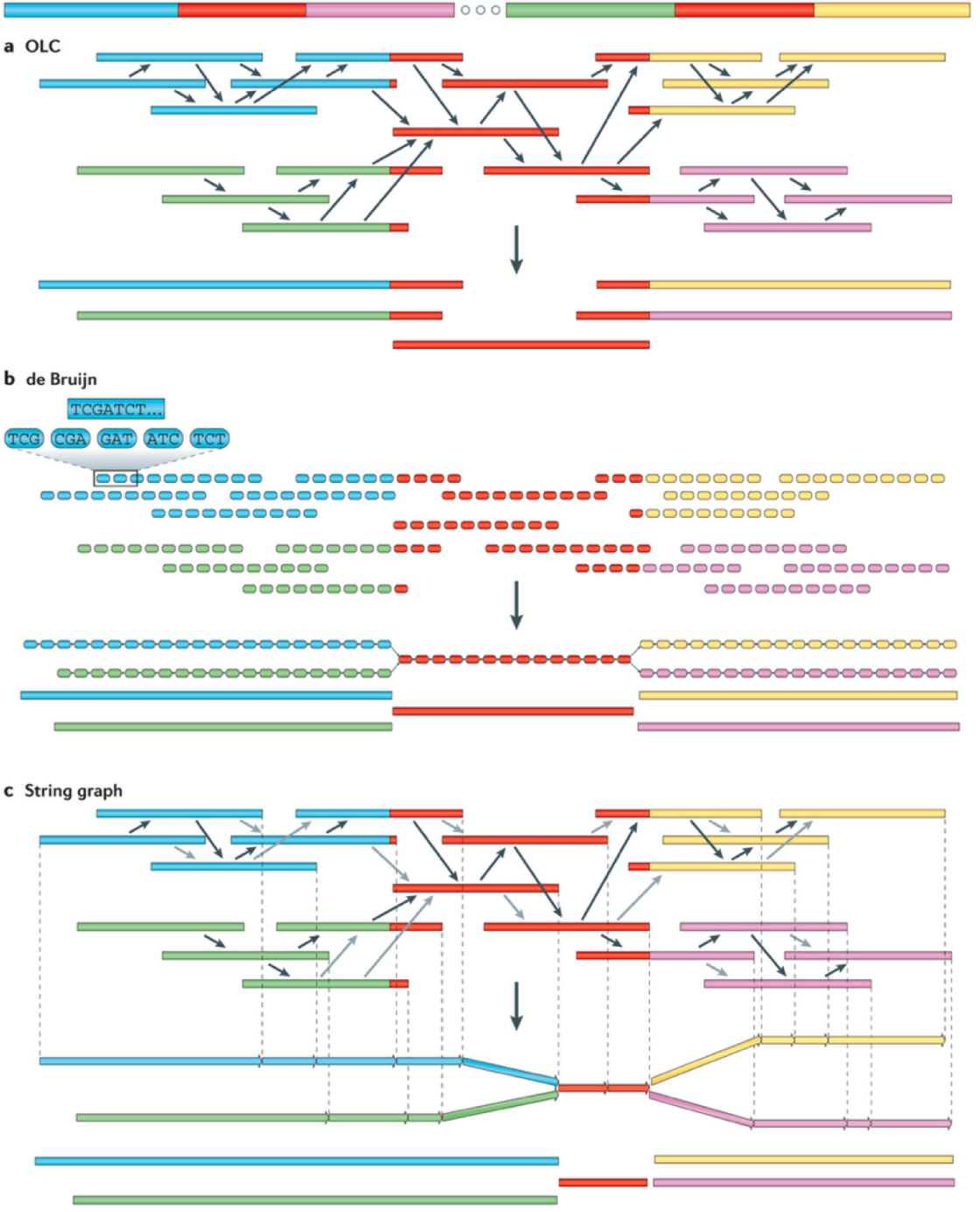
From Simone Silva http://flipper.diff.org/app/items/info/6731

Part III: Alternative Uses of Proximity Ligation (i.e. Hi-C) Technology

- Many new techniques have been developed around the properties of proximity ligation technologies, which provide information about which regions of DNA are close to one another.
- De novo genome assembly of scaffolds using Hi-C (making it much easier to finish genome assembly)
- Phasing complete chromosomes using Hi-C (eliminating the need for expensive microfluidic methods)
- Identification of large structural variations in genomes (i.e. translocations in cancer)

Genome Assembly

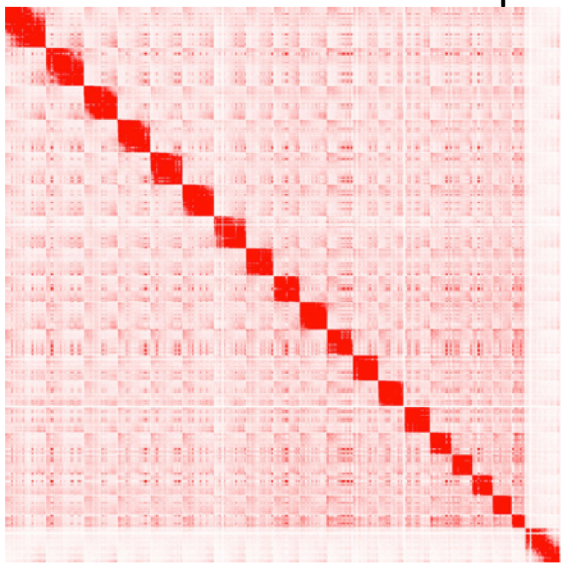
- Short read shotgun sequencing is cheap and easy – however, it becomes very difficult (if not impossible) to assemble the sequence across repetitive elements or duplications in the genome
- As a result, short read sequencing yields tens of thousands of “scaffolds” of varying sizes, the edges of which cannot be confidently connected to other scaffolds due to ambiguous sequence overlap.



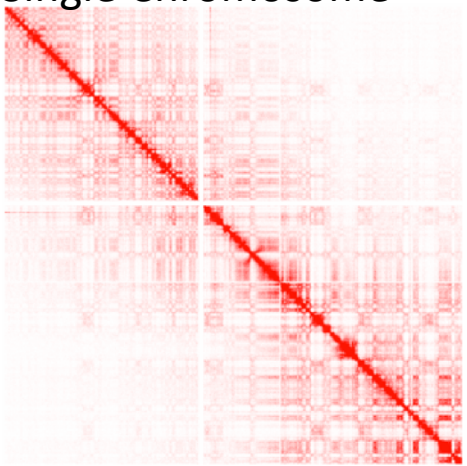
Assembling Scaffolds using Hi-C

- Based on two simple observations:
  - Interactions within chromosomes are MUCH more common then between chromosomes
  - Interactions between two loci on the same chromosome are much more common if they are near one another (linear) than far apart.
- NOTE: also popular right now is a company called Dovetail that has an in vitro version of Hi-C to accomplish pretty much the same thing.

Full Genome Contact Map



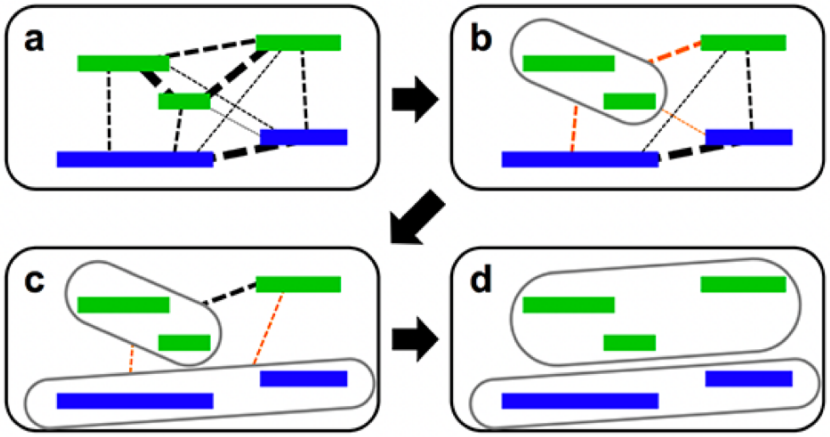
Single Chromosome



Burton et al. 2013

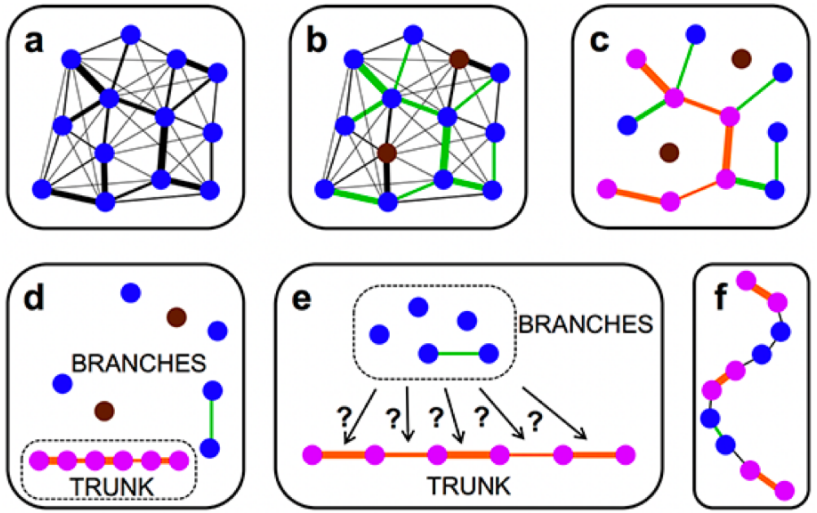
Assembling Scaffolds using Hi-C

First group scaffolds by chromosome…



Supplementary Figure 1 | An illustrated overview of the *LACHESIS* clustering algorithm. a. An assembly consisting of five contigs, which in truth belong to two chromosomes (green and blue). Hi-C links between the contigs are shown as black dotted lines, with thicker lines indicating higher normalized link density. **b.** The agglomerative hierarchical clustering algorithm begins. The two contigs sharing the highest normalized link density are merged together to create a cluster (gray oval). The new link densities between this cluster and each other contig (orange dotted lines) are calculated as the average (normalized) linkage between the two contigs in this cluster and the other contig. **c.** Again, the two contigs sharing the highest normalized link density are merged to create a cluster. New average link densities are calculated (orange dotted lines); note that the link density between the two multi-contig clusters is the average of four original link densities. **d.** Another merge. The user-specified limit of two clusters has been reached, so the algorithm is complete. It has correctly found groups for each chromosome.

… then order them



Supplementary Figure 2 | An illustrated overview of the *LACHESIS* ordering algorithm. a. A group of contigs depicted as a graph. Each blue vertex indicates a contig, and the edges between vertices indicate normalized Hi-C link densities (for clarity, edges are not shown between all pairs of contigs). **b.** A spanning tree (a set of edges that connects all vertices with no loops) is found (green edges). The edges of the spanning tree are chosen to have the maximum possible link densities. Short contigs (dark brown dots) are not included in the spanning tree. **c.** The longest path in the spanning tree (magenta dots, orange edges) is found. This path constitutes the "trunk", an initial contig ordering with high accuracy but low completeness. **d.** The trunk is removed from the spanning tree, leaving a set of vertices and edges called "branches", many of which consist of a single isolated vertex. **e.** Lastly, the branches are considered for reinsertion into the trunk at all possible positions and orientations. Each possible reinsertion site is given a "score" equal to the sum of the reciprocals of all link distances. Very short branches are not reinserted. **f.** The final contig ordering.

Burton et al. 2013

Haplotype Phasing – knowing which variants are on which chromosome

Haplotype Phasing

Haplotypes	Genotype
ATCCGA	A{T/G}{C/A}CG{C/A}
AGACGC	

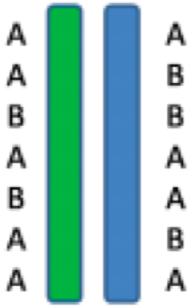
- High throughput cost effective sequencing technology gives genotypes and not haplotypes.

Possible phases: ATACGA AGACGA AGCCGC ATCCGC ….

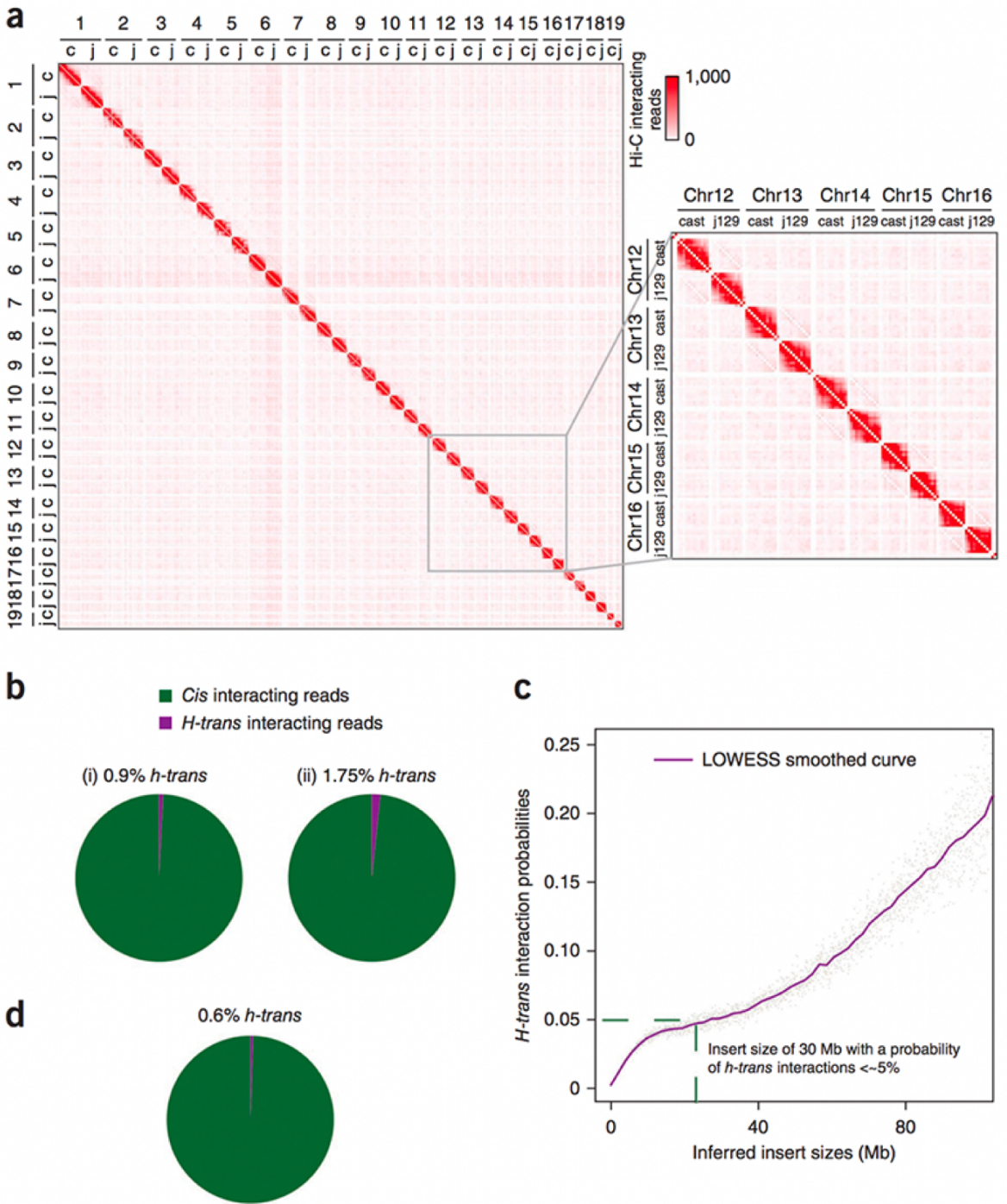
Phased v. Unphased Data

a. Unphased Data (Genotype)

b. Phased Data (Haplotypes)

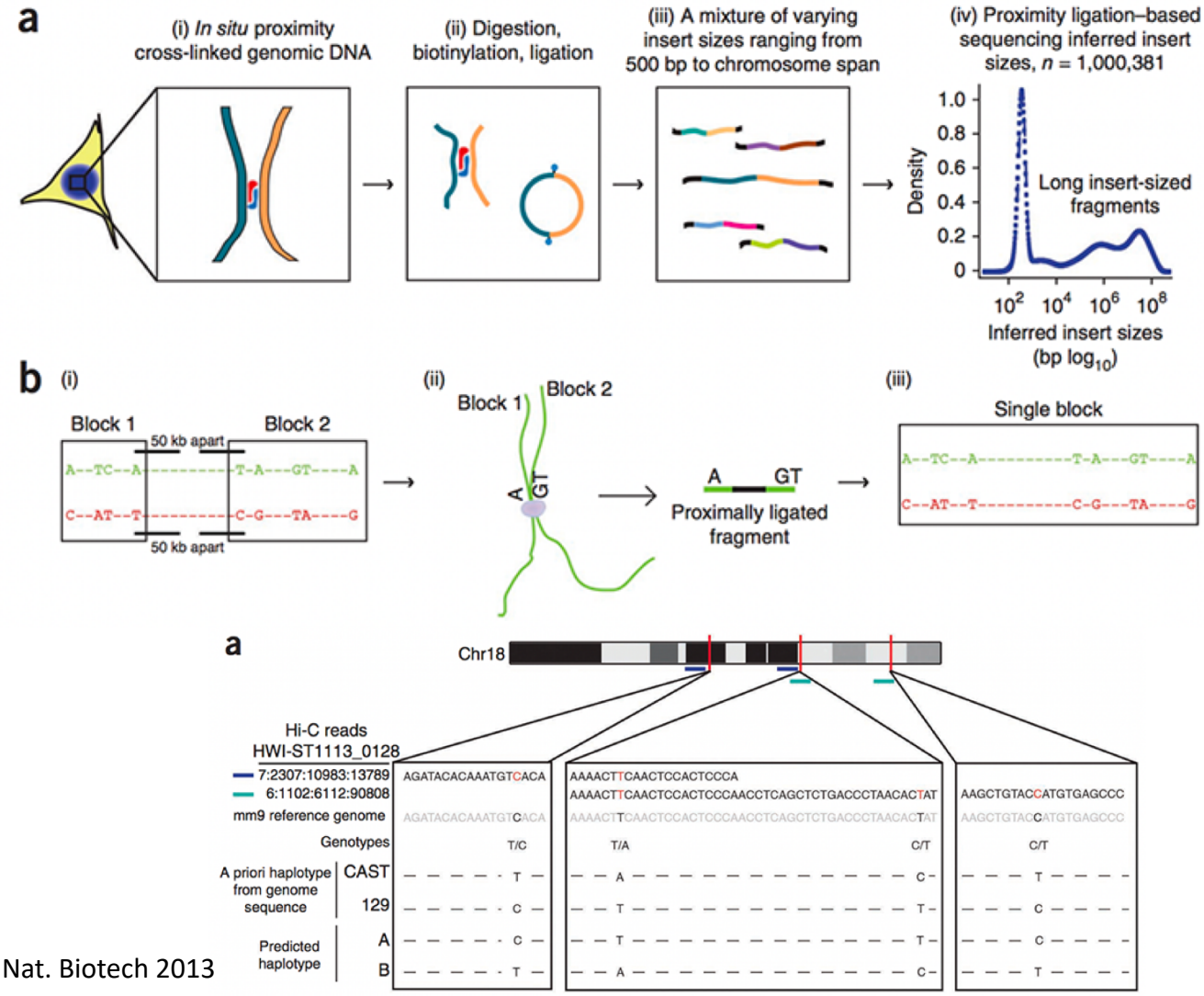


Haplotype Phasing



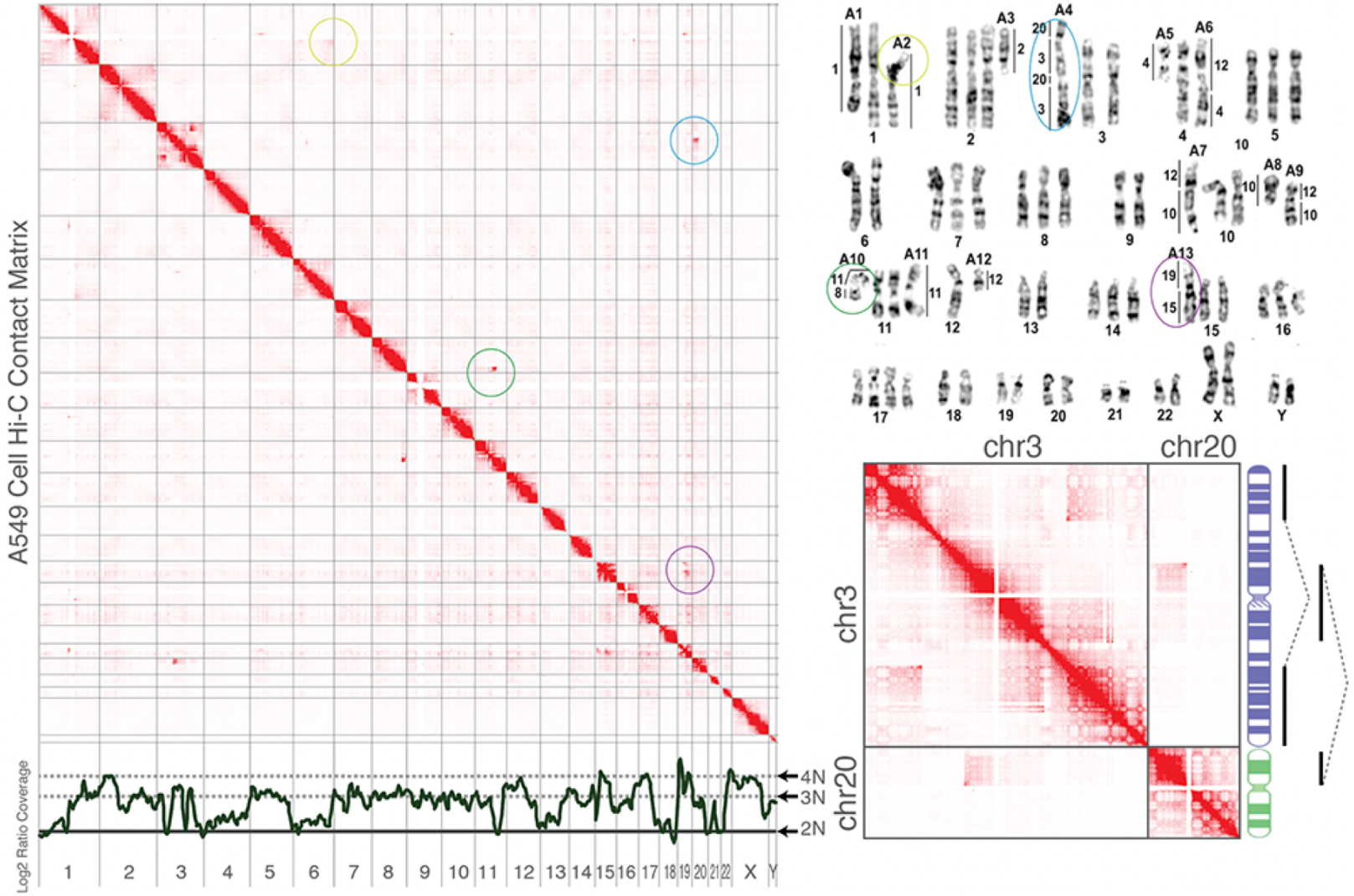
Selvaraj et al. Nat. Biotech 2013

Haplotype Phasing



Selvaraj et al. Nat. Biotech 2013

Identifying translocations/structural variants



Overarching concepts for today

- Exciting time to study gene regulation – clash of unbiased genome-wide observations with long engrained concepts (that may or may not be true)
- Innovation in genomics/NGS methods and analysis are rapidly accelerating the field
- Nascent RNA != RNA-seq
- Current thinking about 3D genome structure is that it is organized at two different levels:
  - Global: Transcriptional activity
  - Local: Cohesin/condensin loops
- How important is 3D genome structure for gene expression (i.e. promoter enhancer loops) – do we really know??