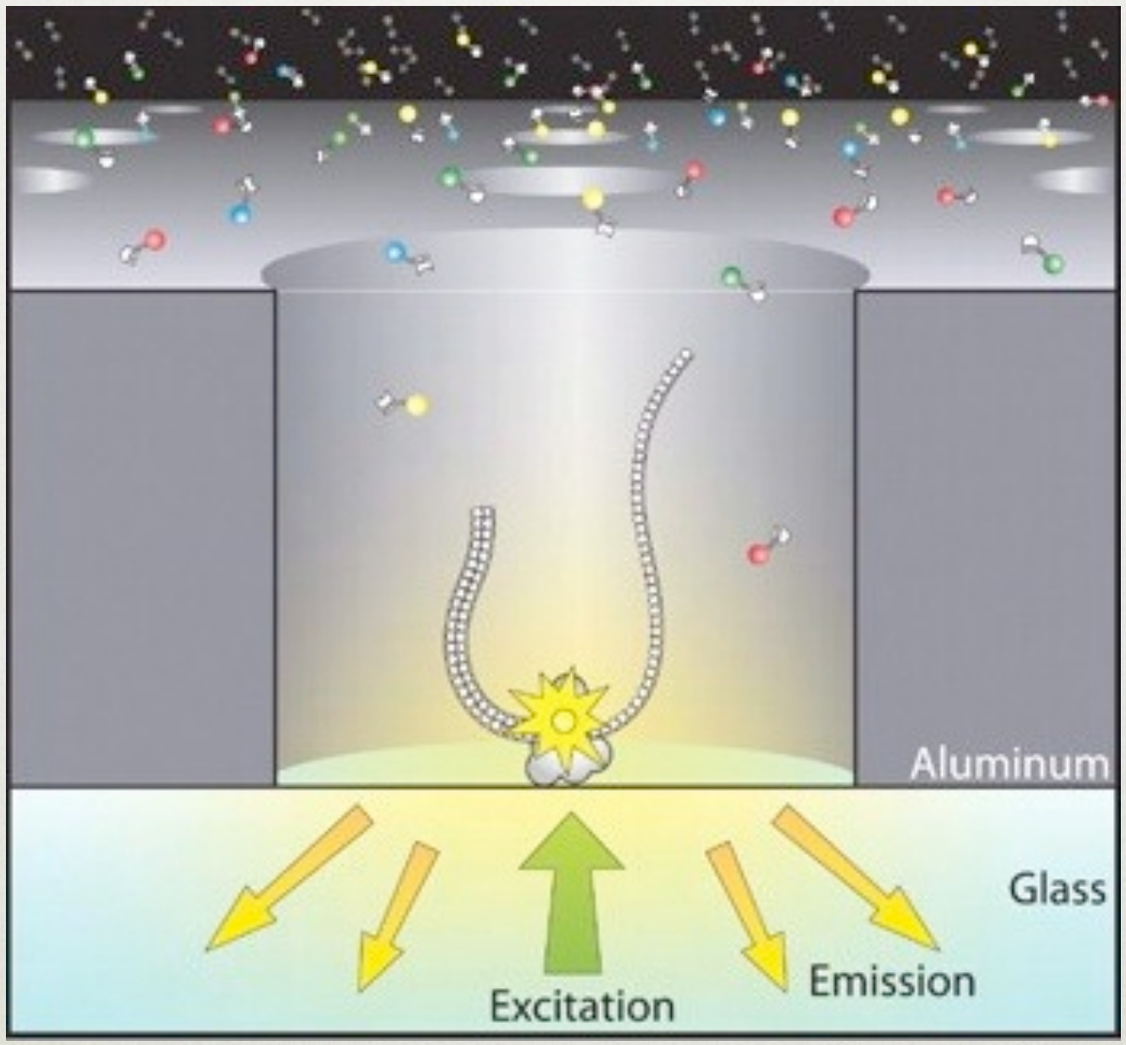
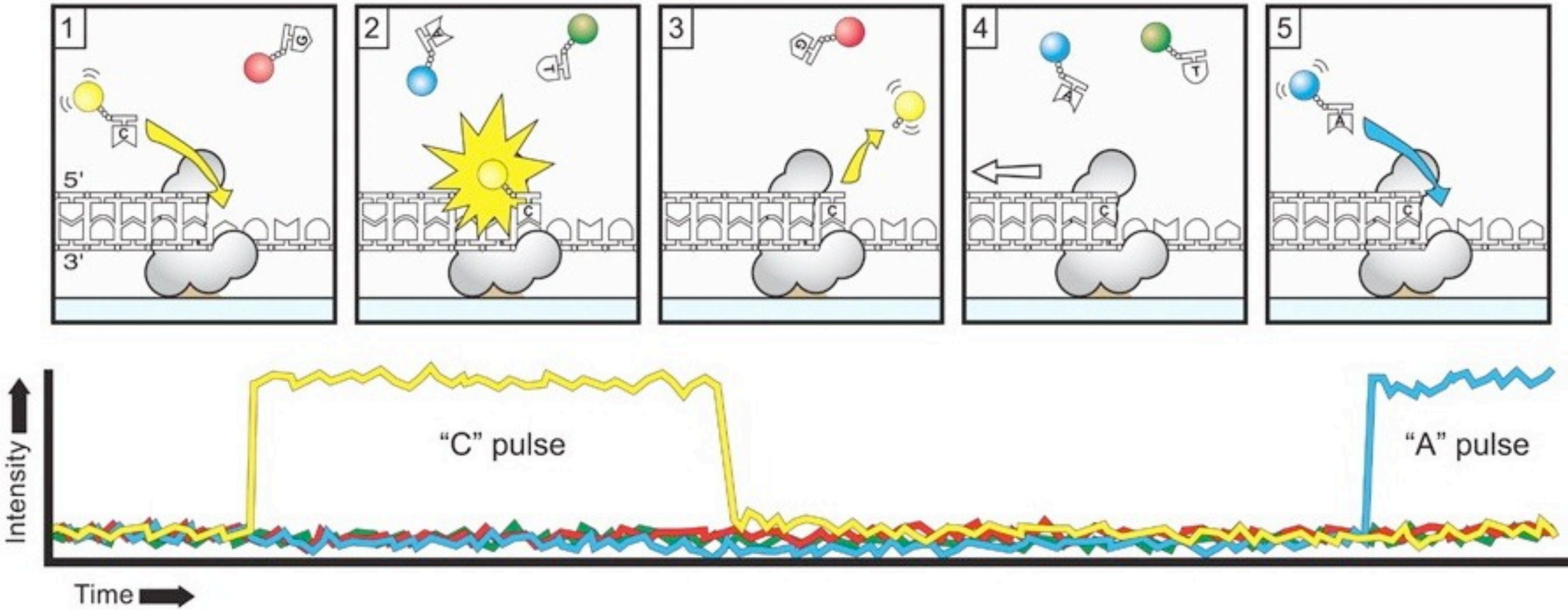
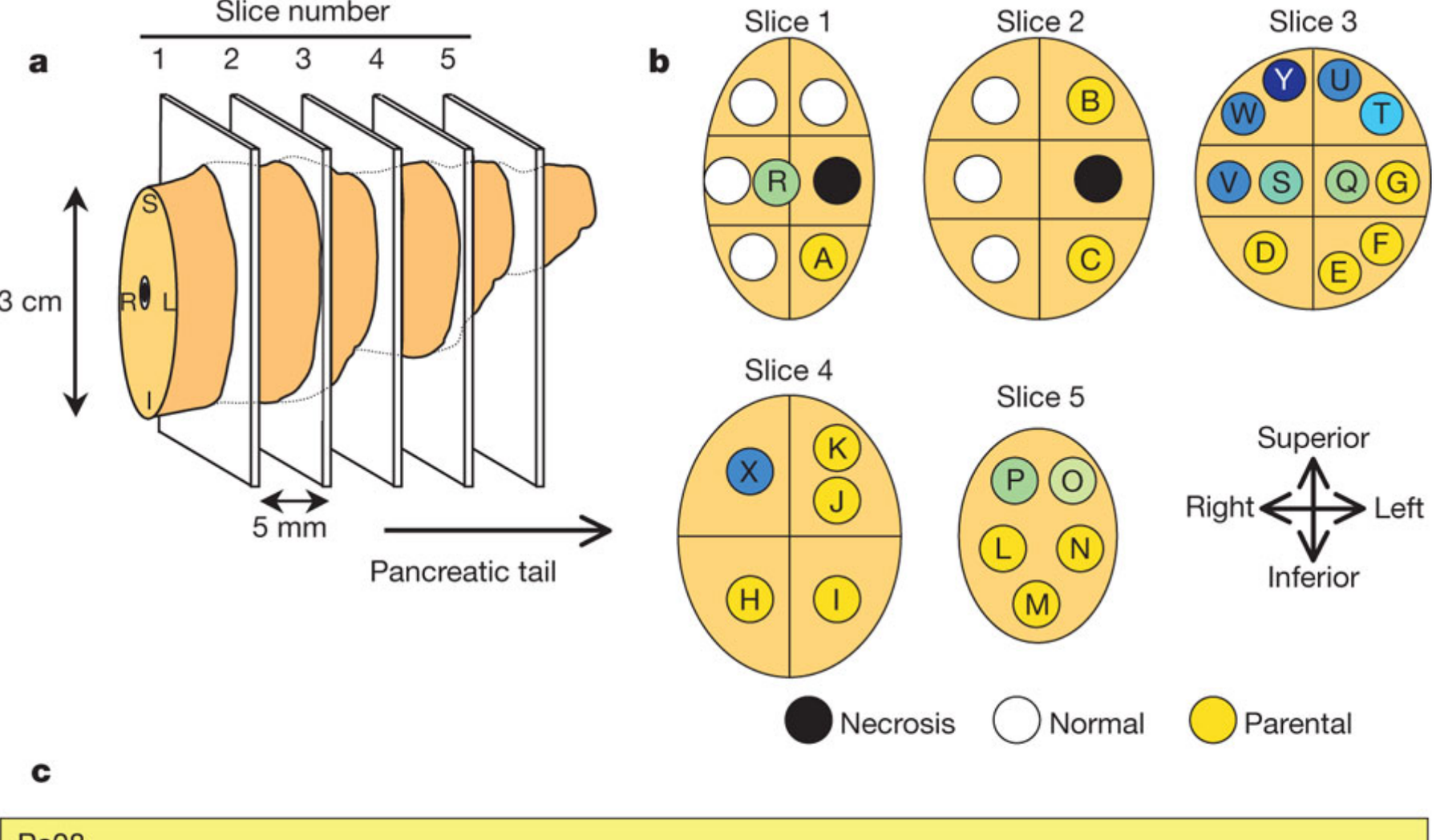
Pacific Biosciences

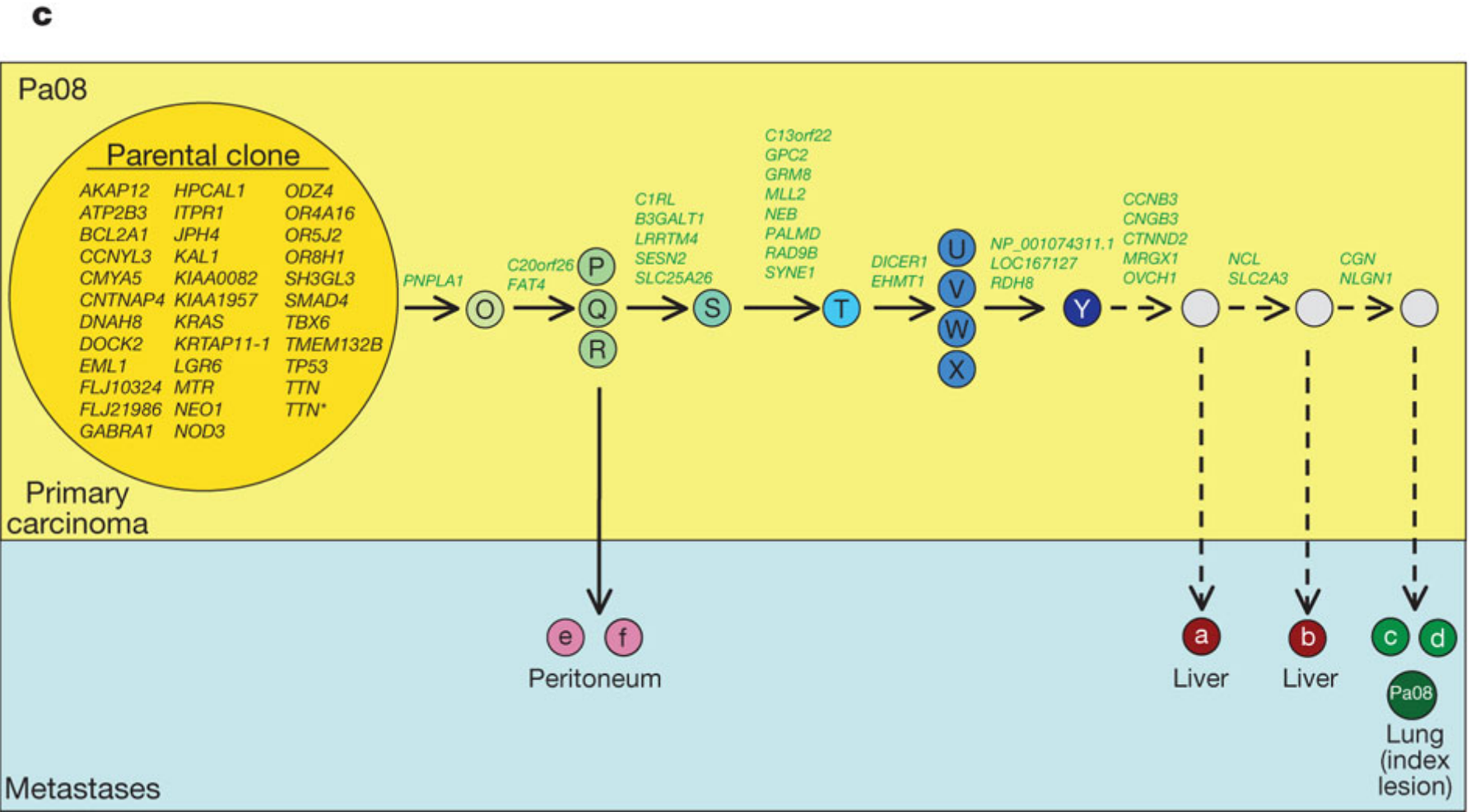

Aluminum
Glass
Excitation
Emission


1
2
3
4
5
5'
3'
C
A
T
G
"C" pulse
"A" pulse
Intensity
Time

Unlike Sanger sequencing, which average over many molecules, in nextGen sequencing PCR errors do not average away

Application: tumorigenesis



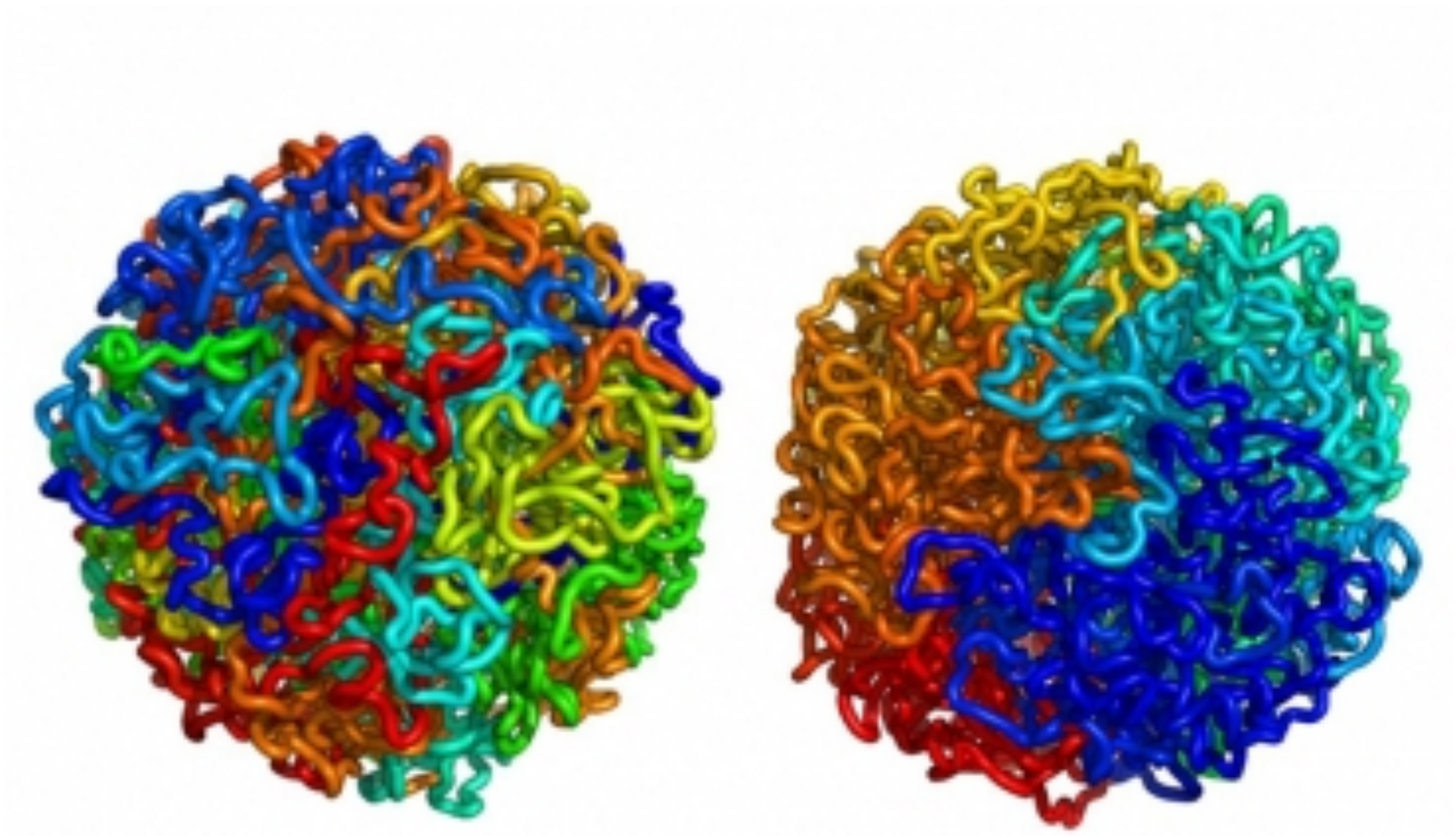


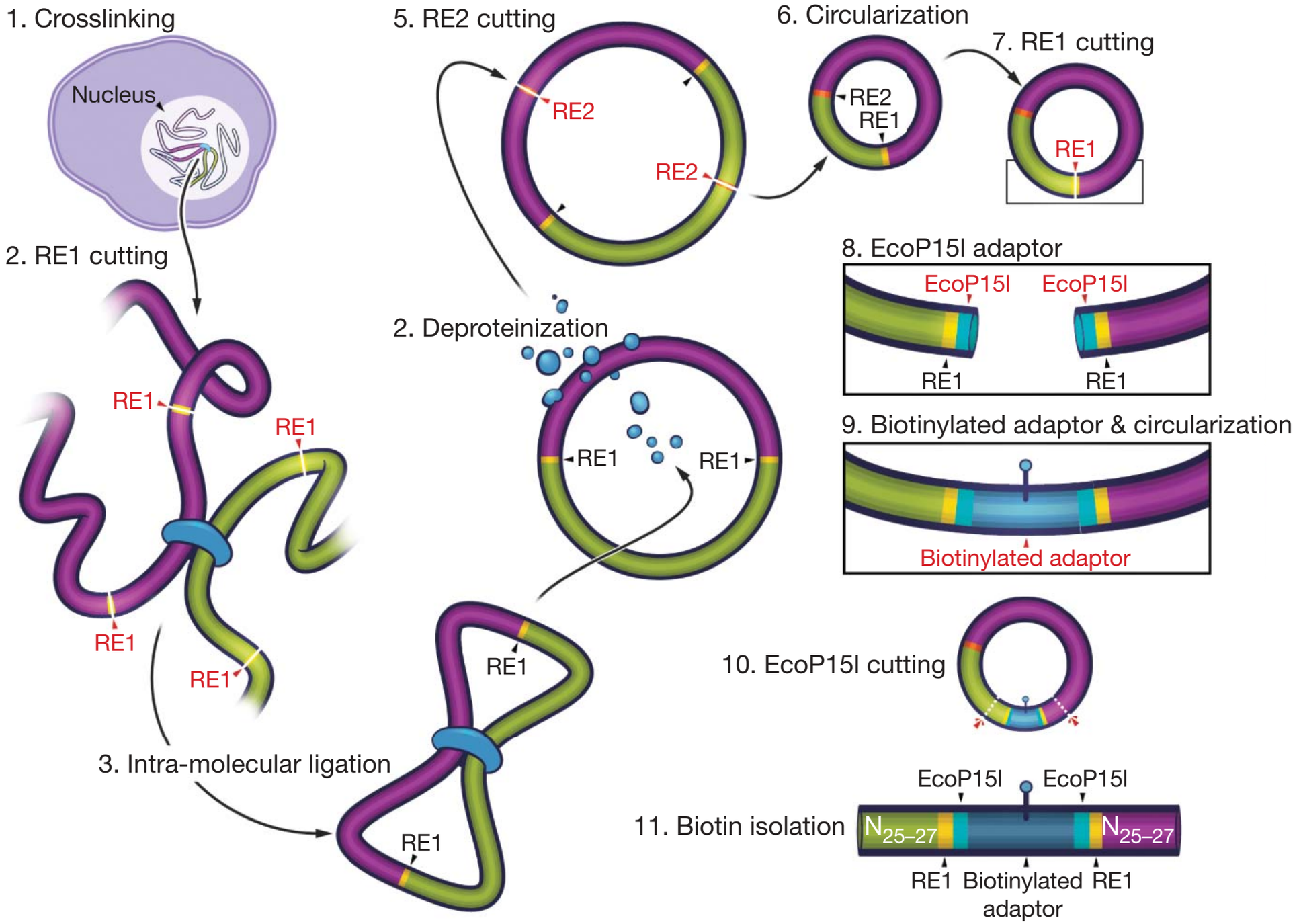
Geographic mapping of metastatic clones within the primary carcinoma and proposed clonal evolution of Pa08.

S Yachida *et al. Nature* **467**, 1114-1117 (2010) doi:10.1038/nature09515

Application: 3D genome

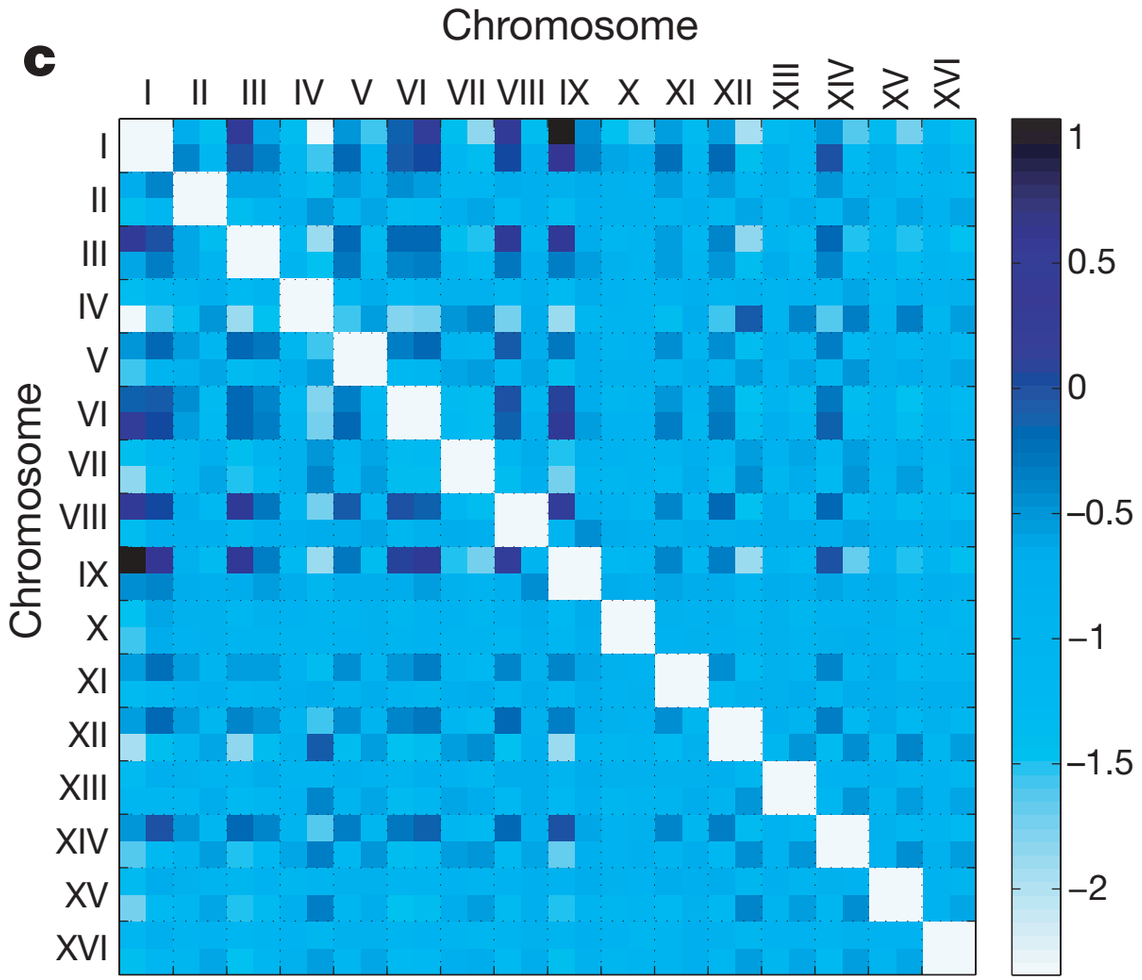
3D Genome?

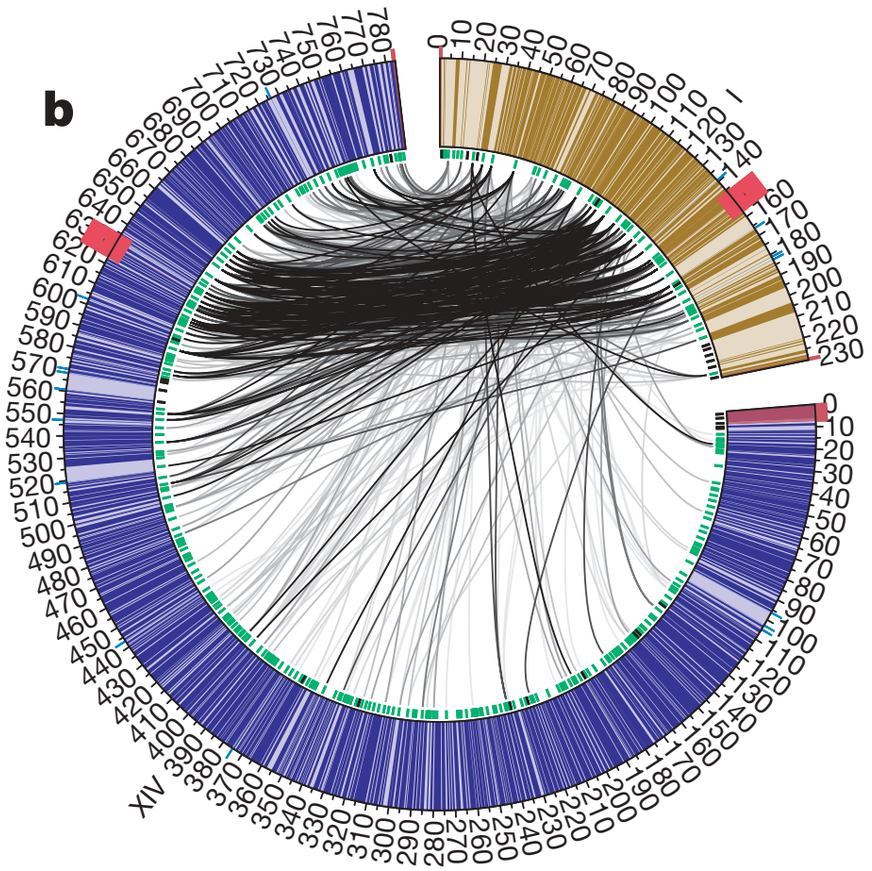




Nature **465** 363 (2010)

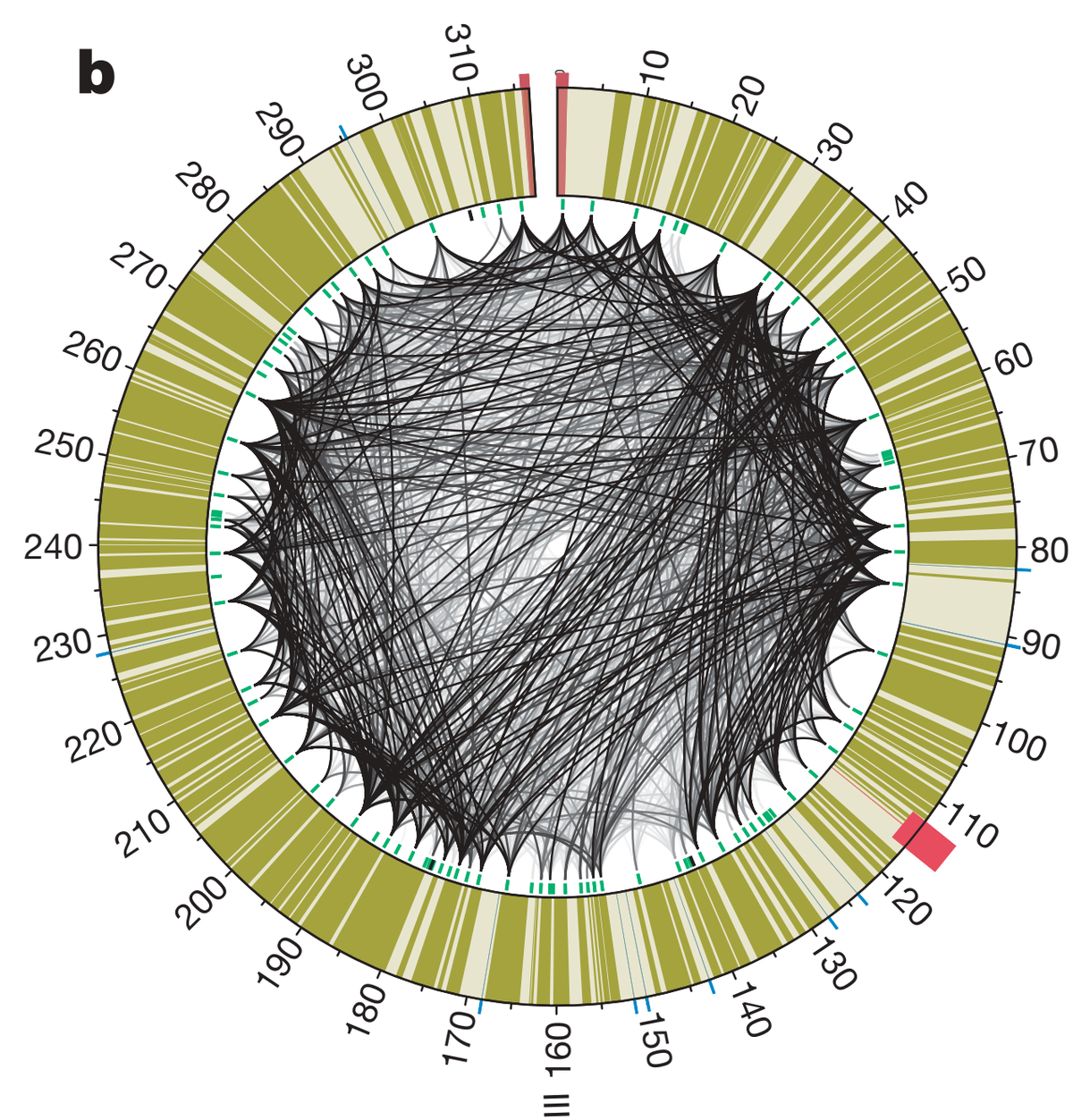
3D Genome

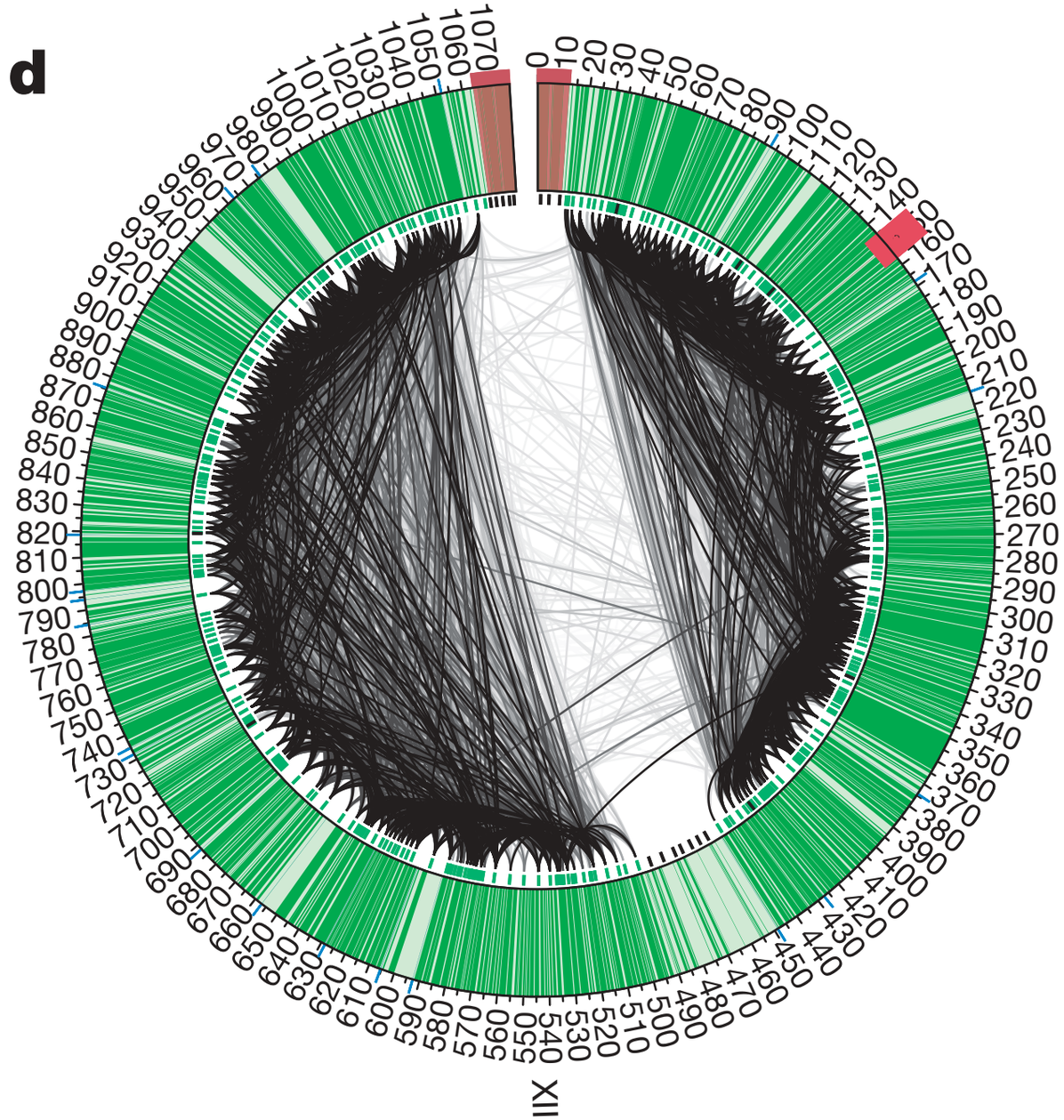




Nature **465** (2010) 363

3D Genome





Nature **465** (2010) 363

Illumina pipeline

image processing

Illumina Firecrest makes mistakes: optical ghost.

50% --70%

base-calling

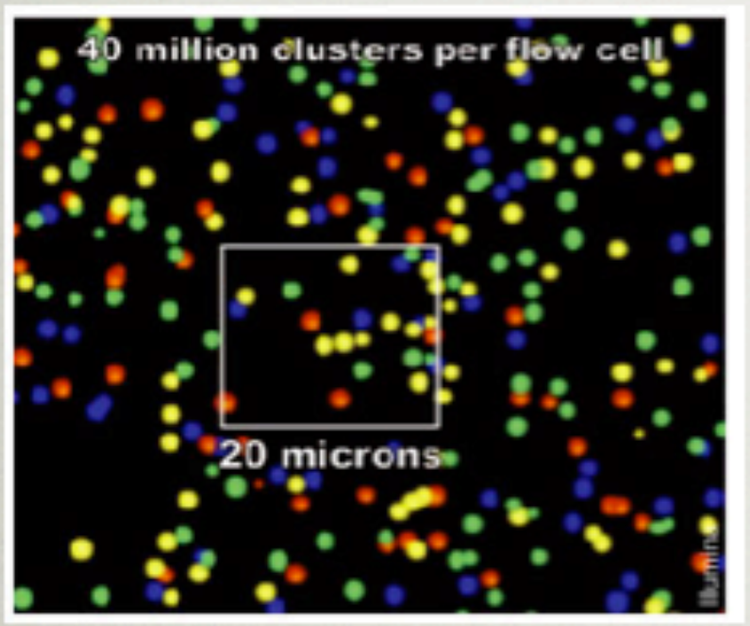
Illumina Bustard

genome alignment

Illumina eland: 2 mismatches
MAQ: similar to eland
bowtie, BWA: very fast

40% --80%

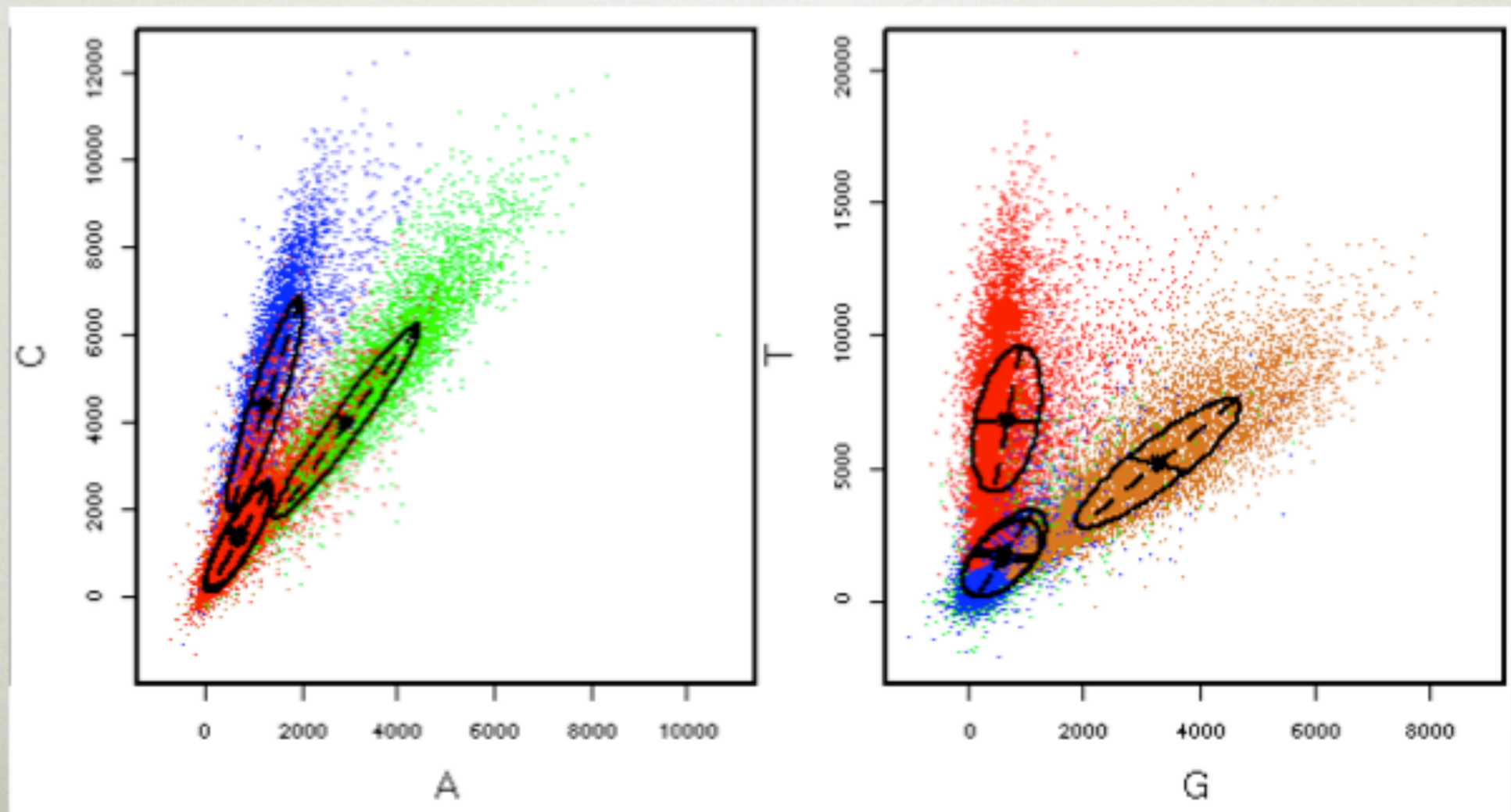
IMAGE PROCESSING



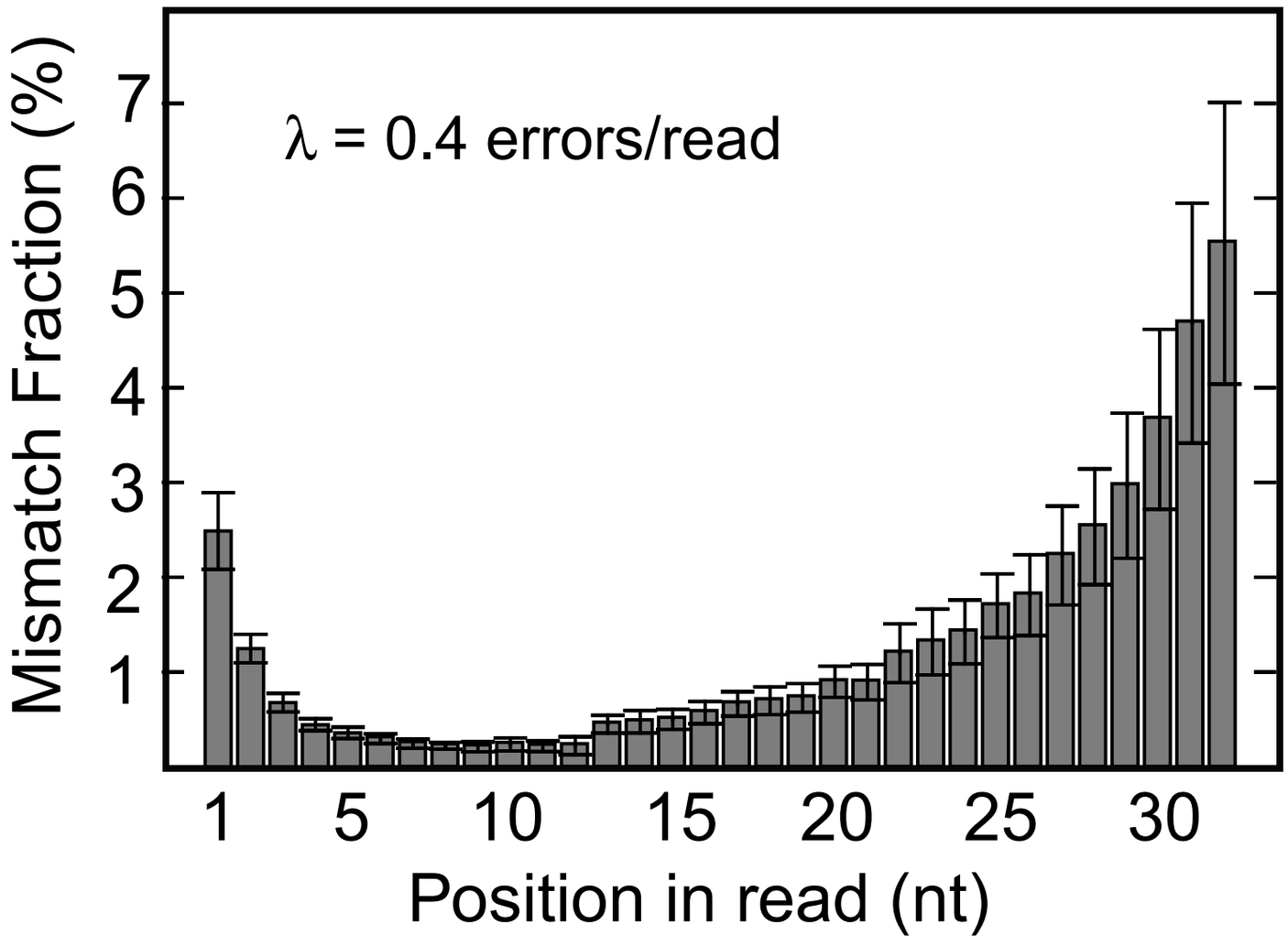
4 by 36 matrices for each colony: intensity of four nucleotides reading for each position

BASE-CALLING

4 by 36 matrix → 36-mer DNA sequence



error rate depends on sequencing position



Wang & Sandberg et al, Nature 2008

Mapper: match tags to the reference genome

- Generally allows 2 mismatches to the reference genome.
- complexity must be linear in N, size of the genome
  - N ~ 10^9
  - CPU clock is ns.
- How to find exact match without mismatch?
  - sorting all 36mers in reference genome.
  - search a sorted list in log(N) steps.

Mapper: match tags to the reference genome

- Method I, sorted list of genomic oligomers or hash table
  - Lam *et al*, Bioinformatics 24 2008, 791.
  - divide 36 bp into six sections
  - matching 15 times (6*5/2)
  - ELAND (Illumina)
  - MAQ (Li et al 2008 Genome Res., 18, 1851–1858)

Mapper: match tags to the reference genome

- Method II, Burrows-Wheeler transformation
  - Burrows, M. and Wheeler, D. J. (1994) Technical report 124, Digital Equipment Corporation, Palo Alto CA
  - Bowtie (Langmead, B. et al. (2009) Genome Biol, 10:R25)
  - BWA (Li & Durbin Bioinformatics **25** 2009 1754)

Mapper: match tags to the reference genome

- Other methods
  - SeqMap (Jiang & Wong, Bioinformatics 24 2008, 2395)
  - SOAP (Li et al. Bioinformatics 2008, 24(5):713)
  - BLAT (Kent, UCSC genome browser)
  - Mosaik-Aligner (http://bioinformatics.bc.edu/marthlab/Mosaik) from Boston College

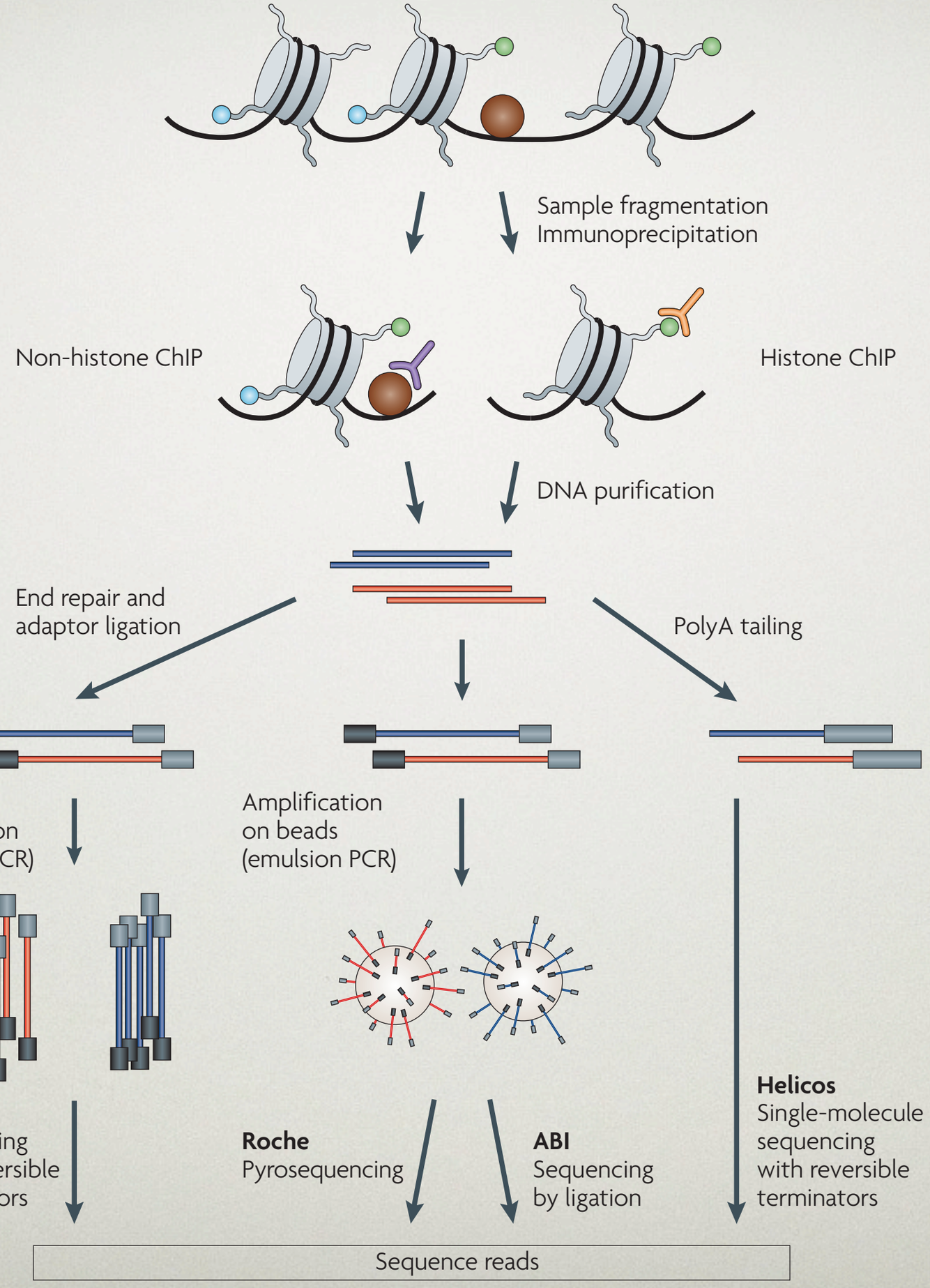
Illumina Genome Analyzer Output

- three types of files
  - s_7_sequence.txt
    - HWUSI-EAS230-R_0023:7:1:1406:20572#0/2:CCGCGAGAGCCATCGCGCGGCTCCGGTCCCTGTTCC:TYdTdbLLTY\Z\\MXZUZ]`^LK`bMM_\Y`^K`B
  - s_7_export.txt
    - HWUSI-EAS230-R 0023 7 1 1406 20572 0 2 CCGCGAGAGCCATCGCGCGGCTCCGGTCCCTGTTCC TYdTdbLLTY\Z\\MXZUZ]`^LK`bMM_\Y`^K`B NM Y
    - HWUSI-EAS230-R 0023 7 1 1245 18361 0 2 CTCTTCCTCAACACAGAGGGGGTTAACAAGCCATGC d \ddddTddacTbdcTT]Y`Z[]L``cTYbd\cYcb c6.fa 110171719 F 36 118 236 69 R Y
  - s_7_sorted.txt
    - HWUSI-EAS230-R 0023 7 1 1245 18361 0 2 CTCTTCCTCAACACAGAGGGGGTTAACAAGCCATGC d \ddddTddacTbdcTT]Y`Z[]L``cTYbd\cYcb c6.fa 110171719 F 36 118 236 69 R
- Solexa manual: http://watson.nci.nih.gov/solexa/

Illumina Data

- UCSC genome browser ENCODE
  - http://genome.ucsc.edu/ENCODE/
- GEO short read archive
  - http://www.ncbi.nlm.nih.gov/geo/
- NIH epigenomics roadmap
  - http://www.ncbi.nlm.nih.gov/geo/roadmap/epigenomics/

ChIP-seq

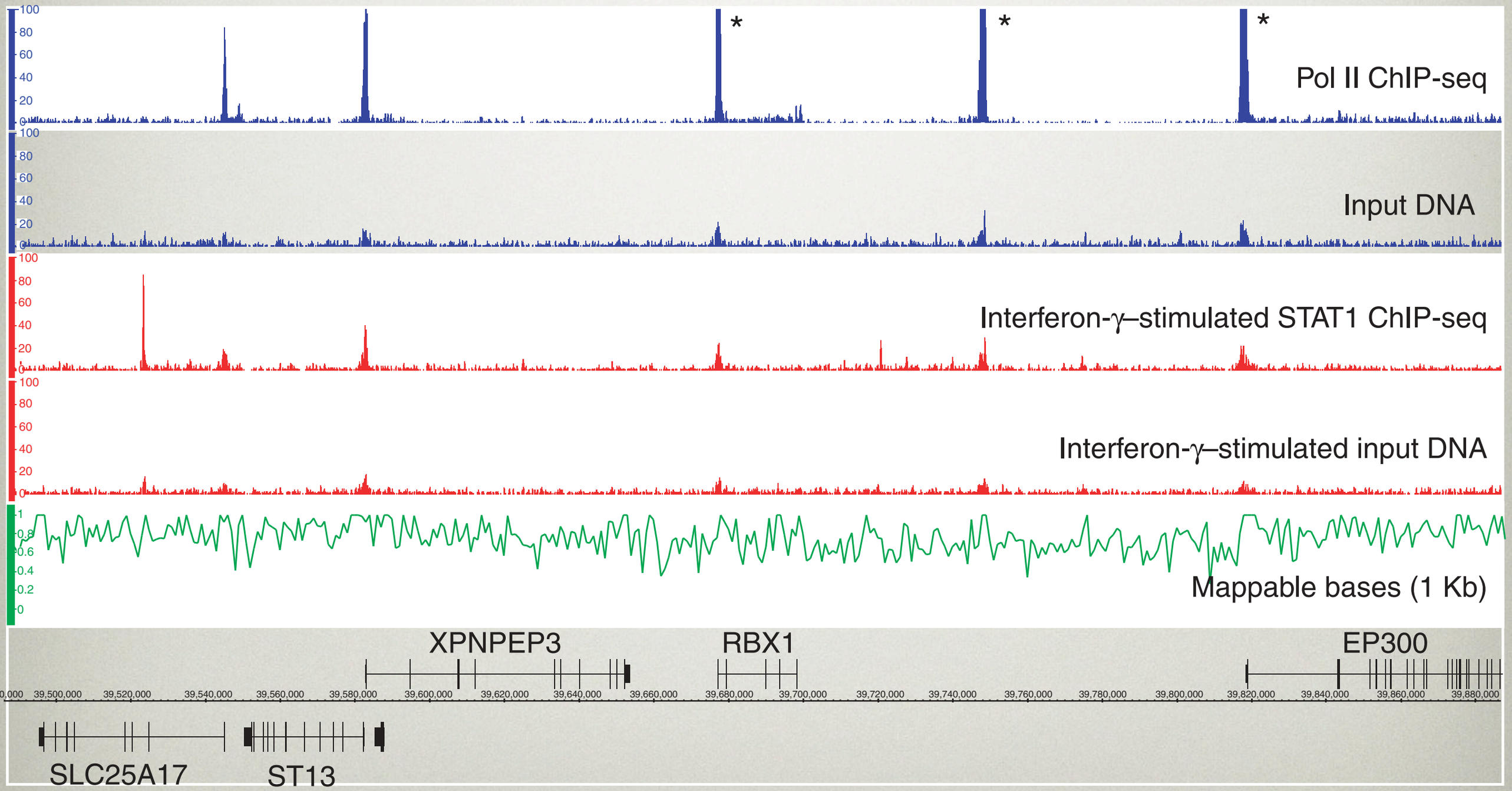


Nat Rev Gen 10
669 (2009)

Types of binders

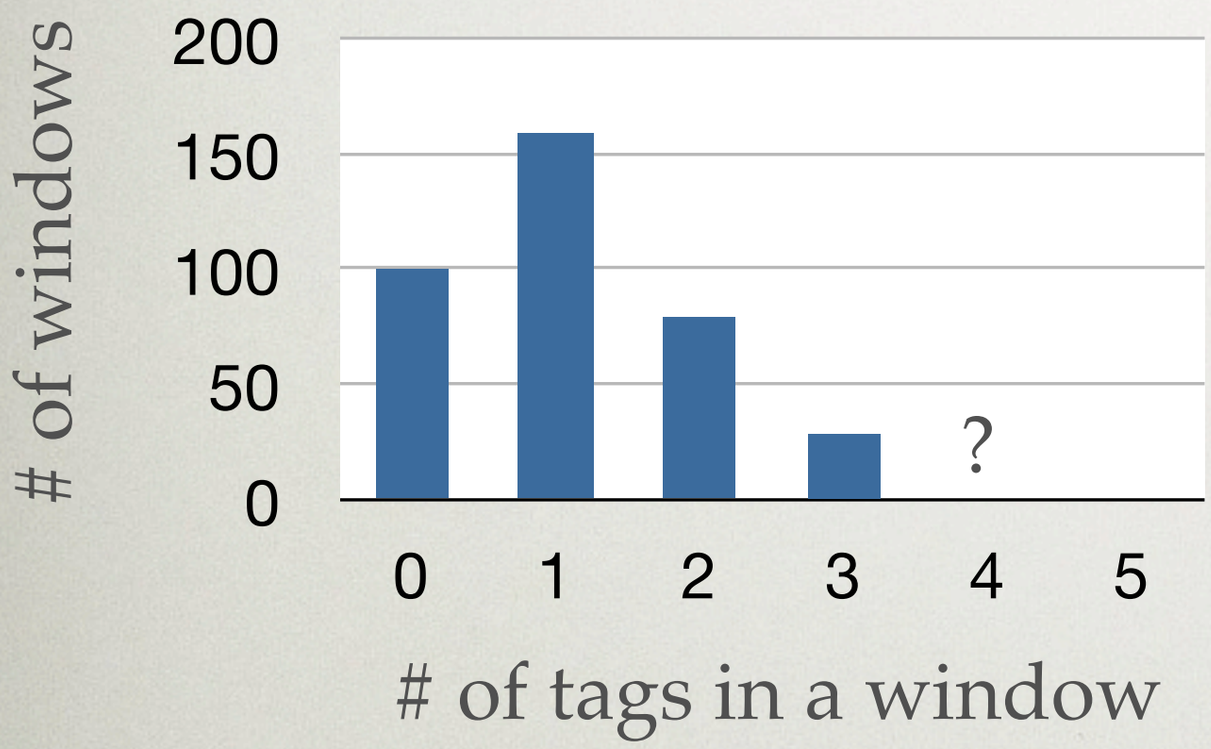
1. point-like binding such as transcription factors, or CTCF.
2. extended binding region: histone modifications. H3K27 has larger domain than H3K4.
3. PolII: point-like in promoter, and extended in gene body.

CONTROL EXPERIMENT



Rozowsky *et al.* Nat Biotechnol 2009, **27:**66.

Calibrating background



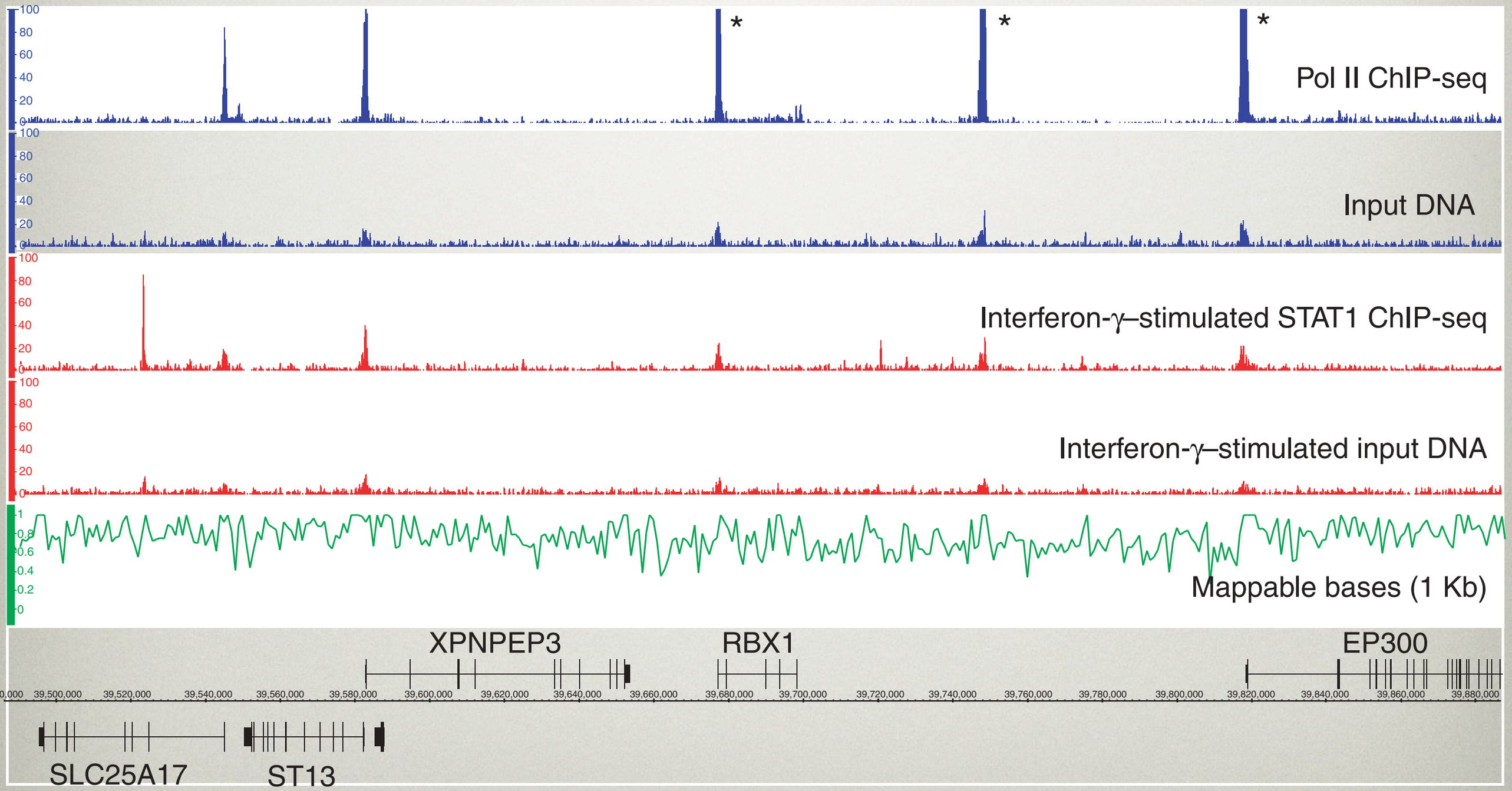
What is the likelihood of finding a window with large tag count?

Extrapolate!

Poisson statistics

Negative Binomial

CONTROL EXPERIMENT



Rozowsky *et al.* Nat Biotechnol 2009, **27:**66.

Program	Special feature	How to shift tag	Peaks ranked by	artifact filtering strand/duplicate
MACS	shift tags	high quality	poisson p-value	No/Yes
SICER	histone domain	input	poisson q-value	No/Yes
cisGenome	to background	high quality	negative binomial	Yes/Yes
QuEST	shape kernel	corelation	fold of enrichme	Yes/Yes
PeakSeq	mappability	tag extension	Poisson q-value	No/No
spp	find summit	correlation	MC p-value	Yes/No
GLITR	FDR from background	tag extension	peak height	No/No

Modified from Pepke, S *Nature Methods Supplement* **6**, S22 - S32 (2009)

Genome Browser is your friend

- UCSC Genome Browser
- http://genome.ucsc.edu
- Tutorial: Zweig et al. Genomics 92 (2008) 75–84
- custom track
  - bed file: block data
  - wiggle file: continuous data