Omics Bootcamp

Michael Schatz

Sept 19, 2014
WSBS Genomics

Outline

1. Applications of DNA Sequencing
   - Basic Concepts
   - Applications to Autism Genetics

2. "Functional" Assays
   - RNA-seq
   - Methyl-seq
   - ChIP-seq
   - Single Cell Sequencing

Milestones in DNA Sequencing

(TIGR/Celera, 1995-2001)

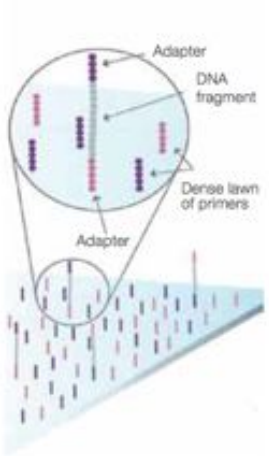
Inside the NY Genome Center

Sequencing Capacity Exceeds 2 Pbp/year (18,000 genomes / year)

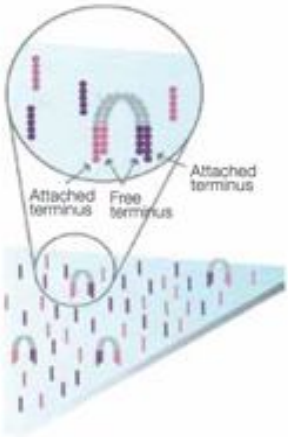
Massively Parallel Sequencing

Illumina HiSeq 2000
Sequencing by Synthesis

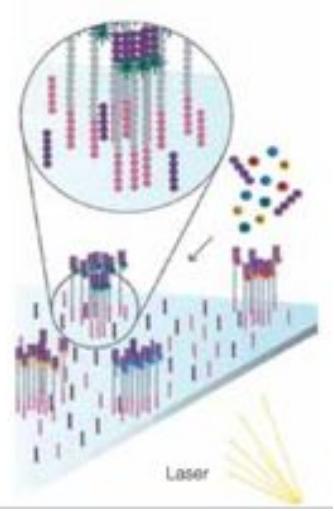
>60Gbp / day



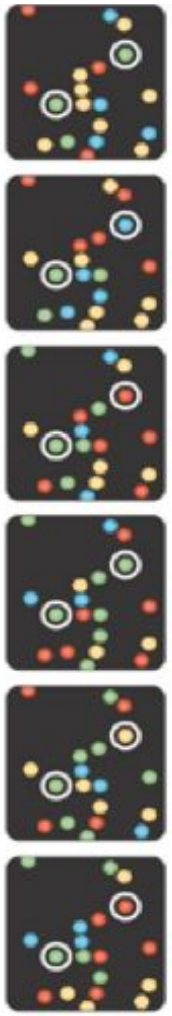
1. Attach



2. Amplify



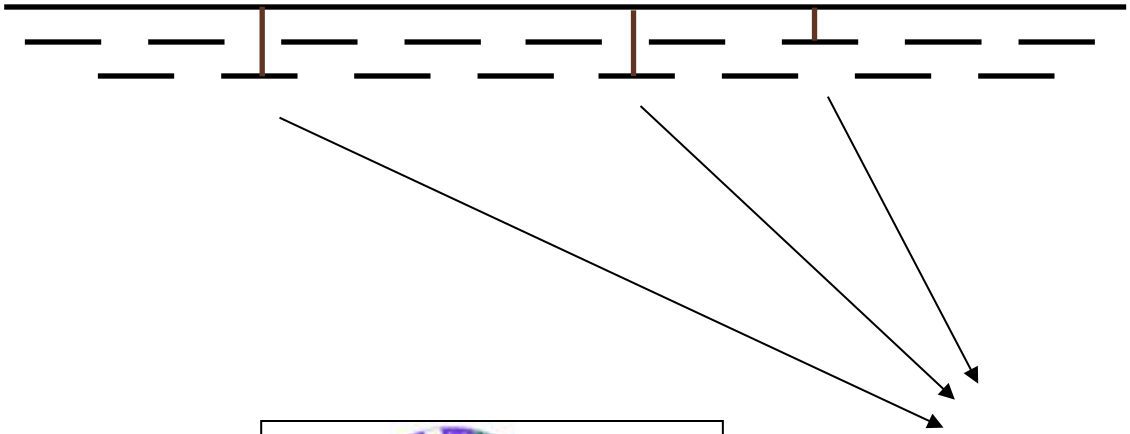
3. Image

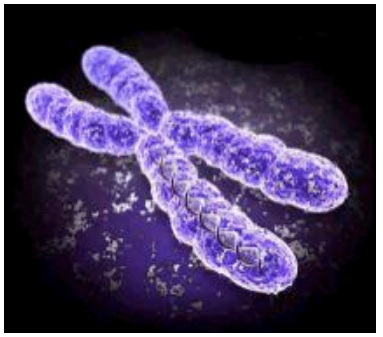


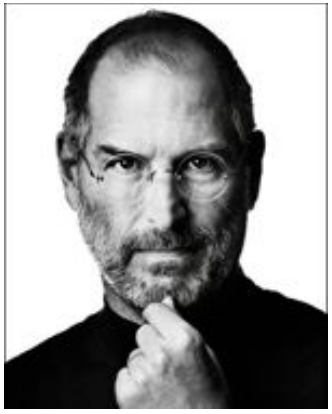
Metzker (2010) Nature Reviews Genetics 11:31-46
http://www.youtube.com/watch?v=l99aKKHcxC4

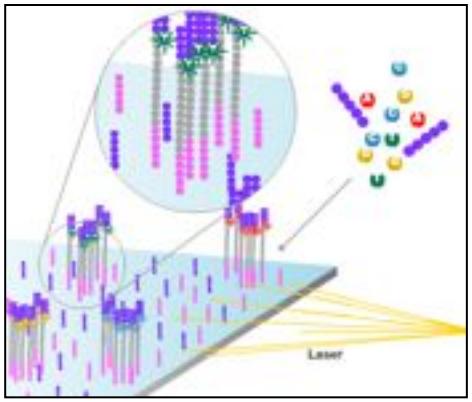
Personal Genomics

How does your genome compare to the reference?







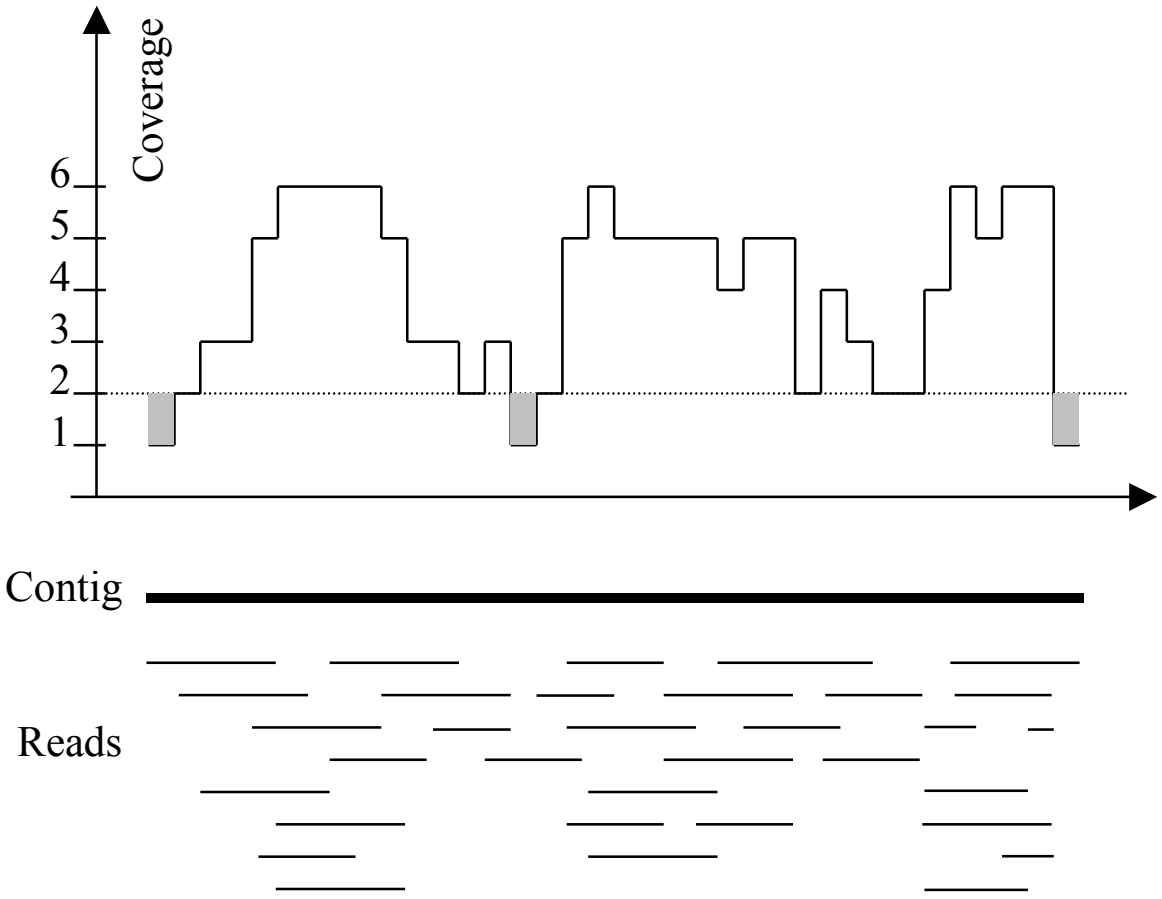


Heart Disease

Cancer

Creates magical technology

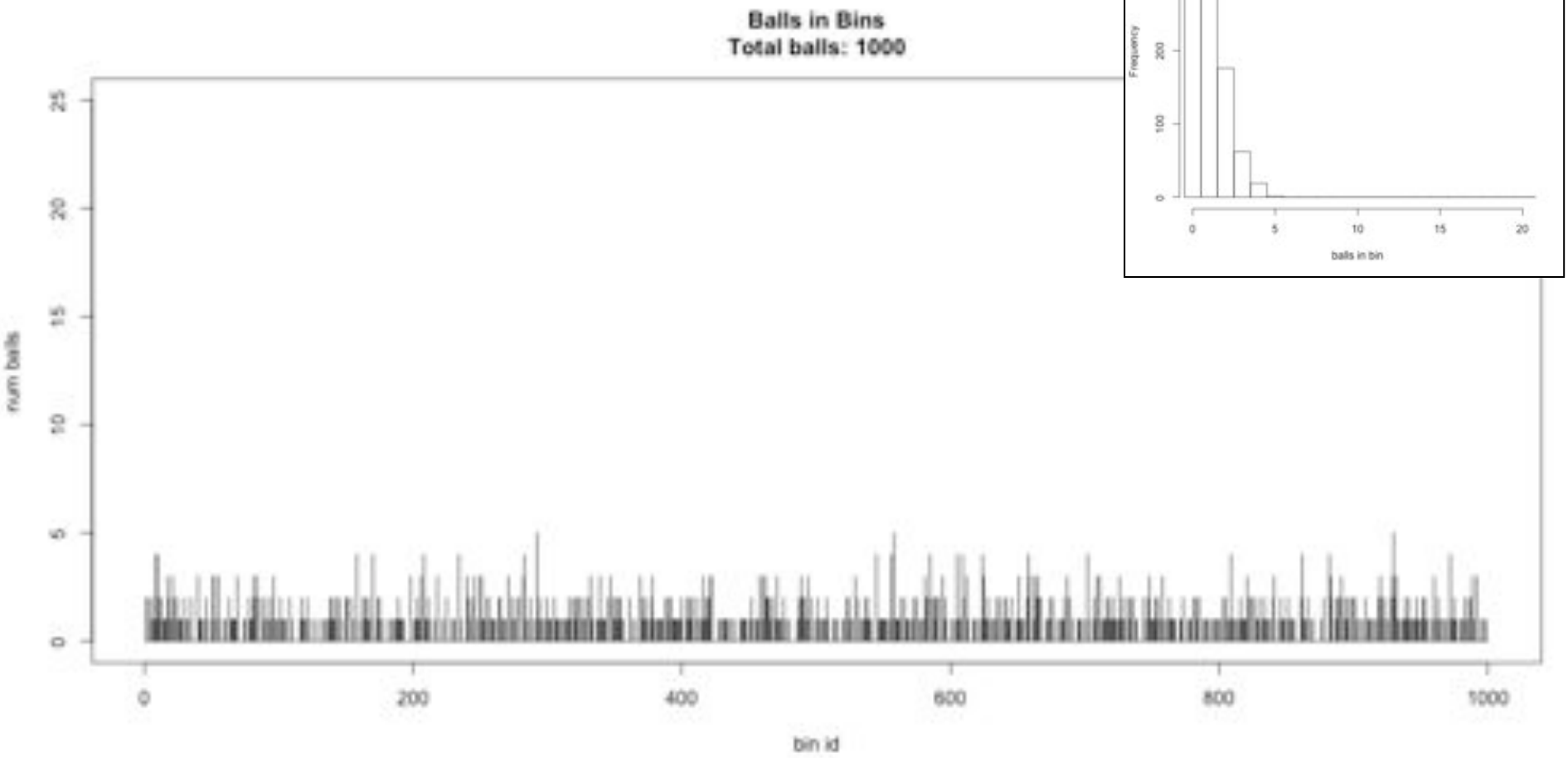
Typical sequencing coverage



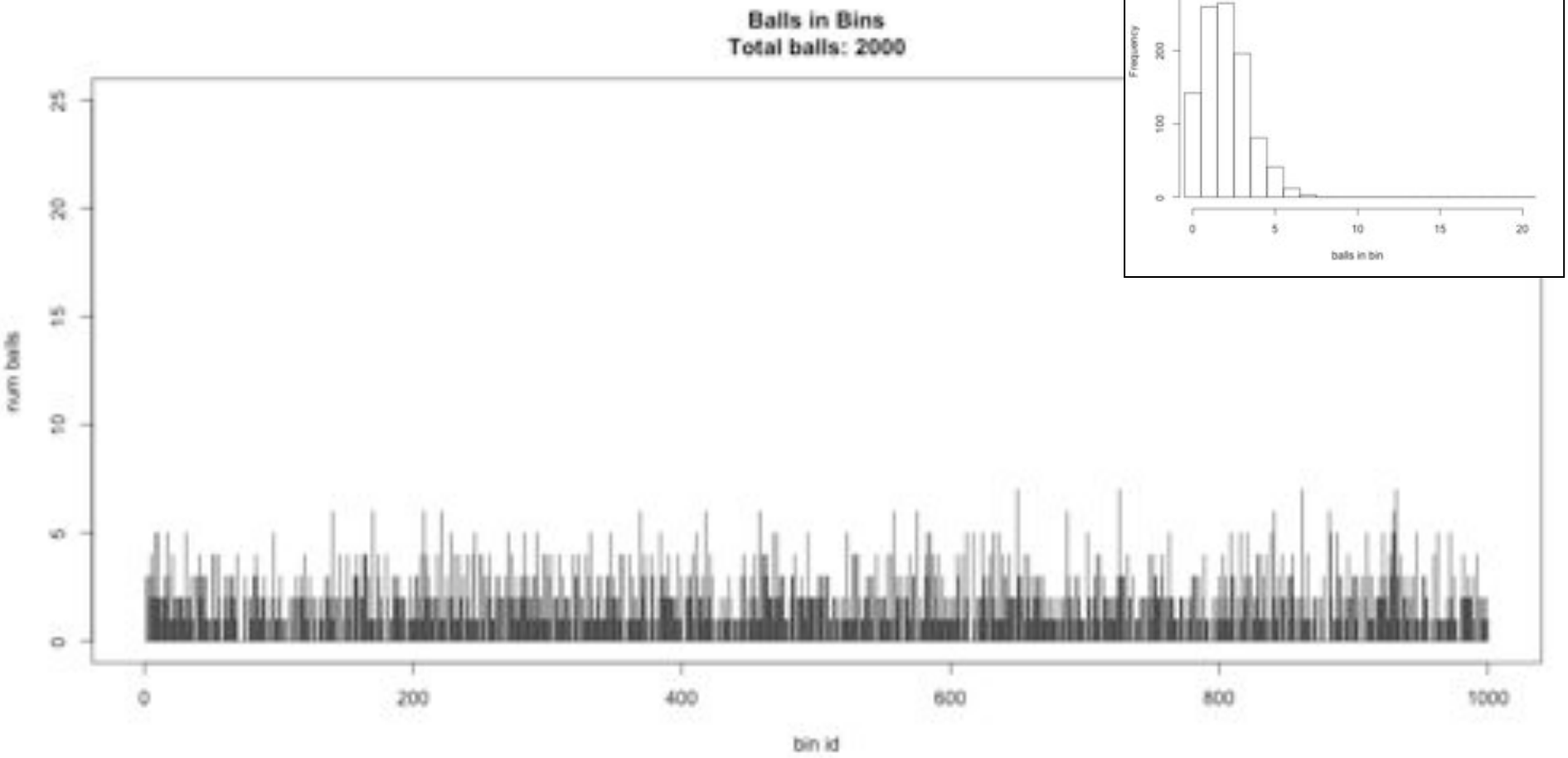
Imagine raindrops on a sidewalk

We want to cover the entire sidewalk but each drop costs $1

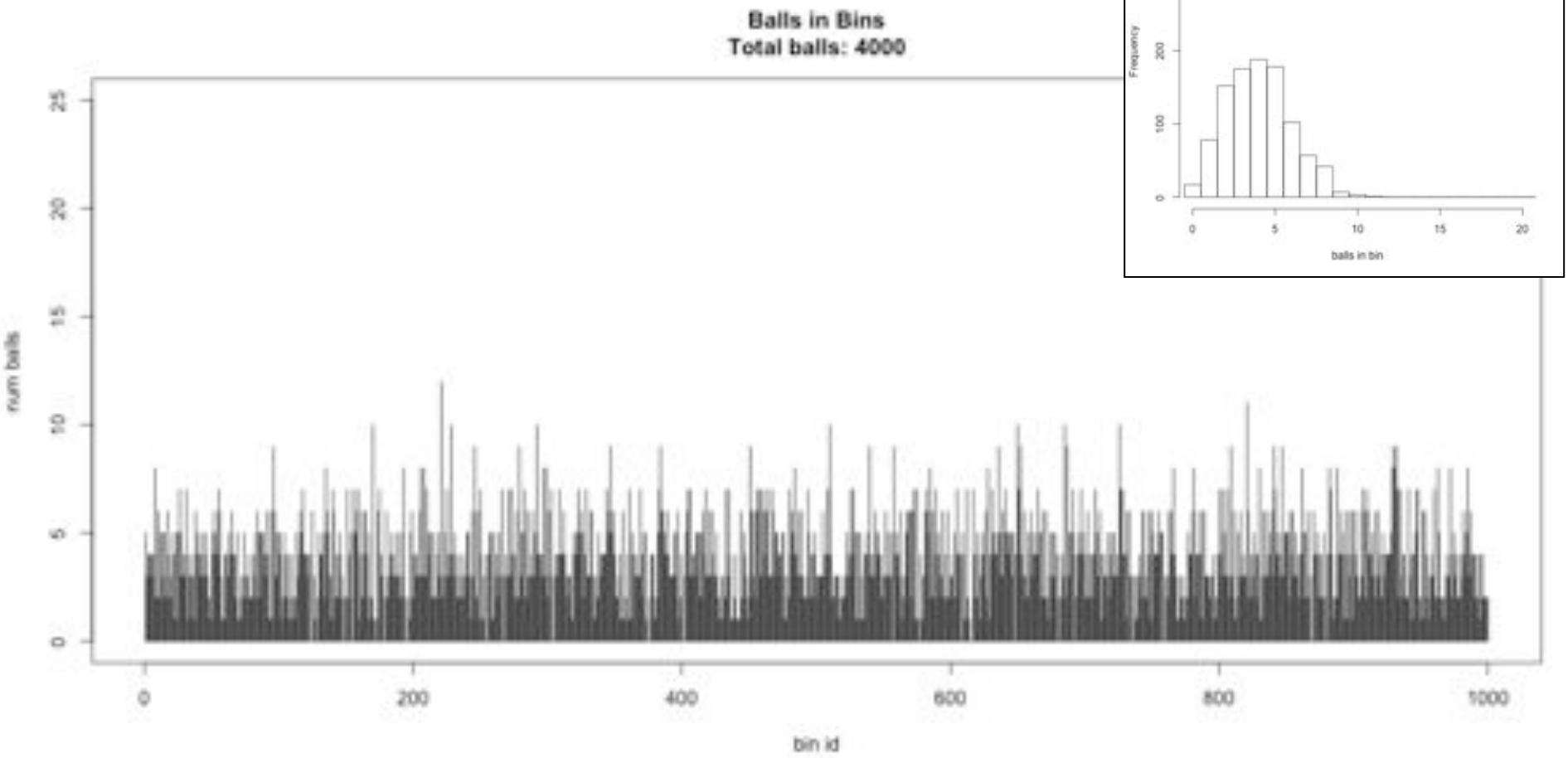
1x sequencing



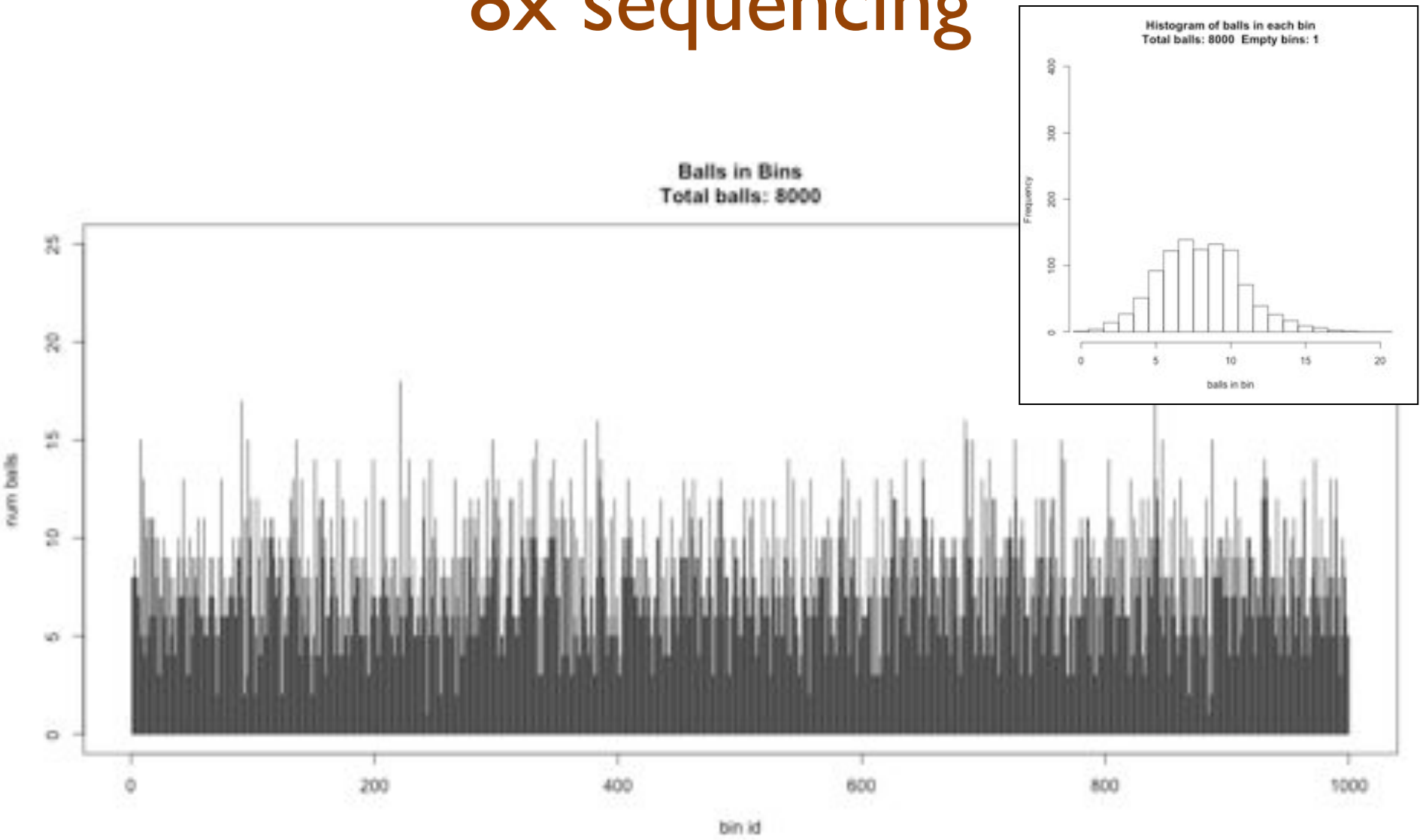
2x sequencing



4x sequencing



8x sequencing



Poisson Distribution

The probability of a given number of events occurring in a fixed interval of time and/or space if these events occur with a known average rate and independently of the time since the last event.

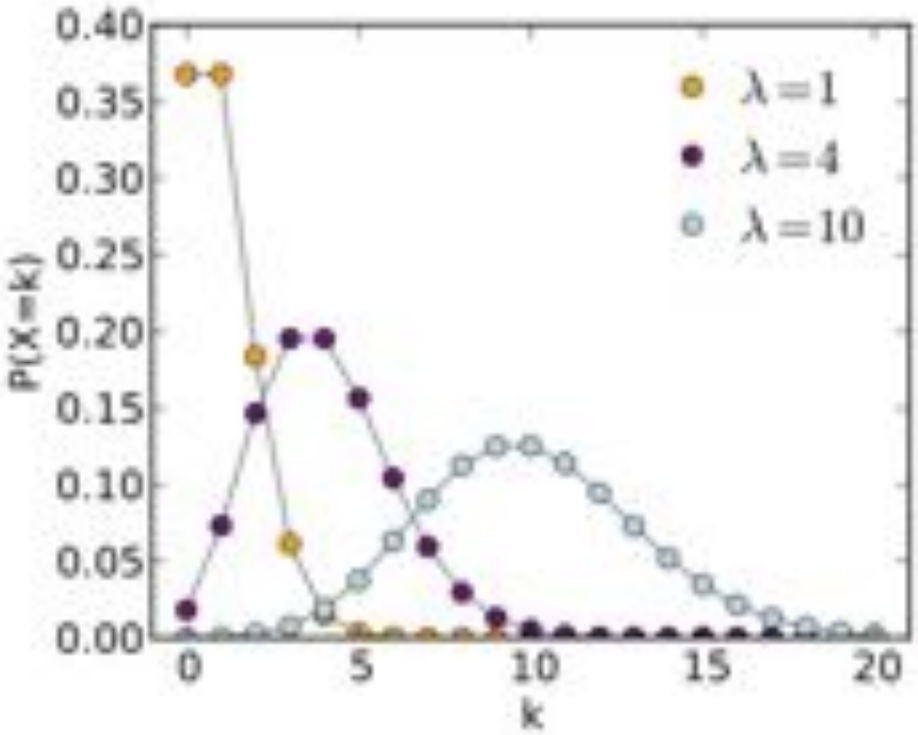
$$P(k) = \frac{\lambda^k}{k!} e^{-\lambda}$$

Formulation comes from the limit of the binomial equation

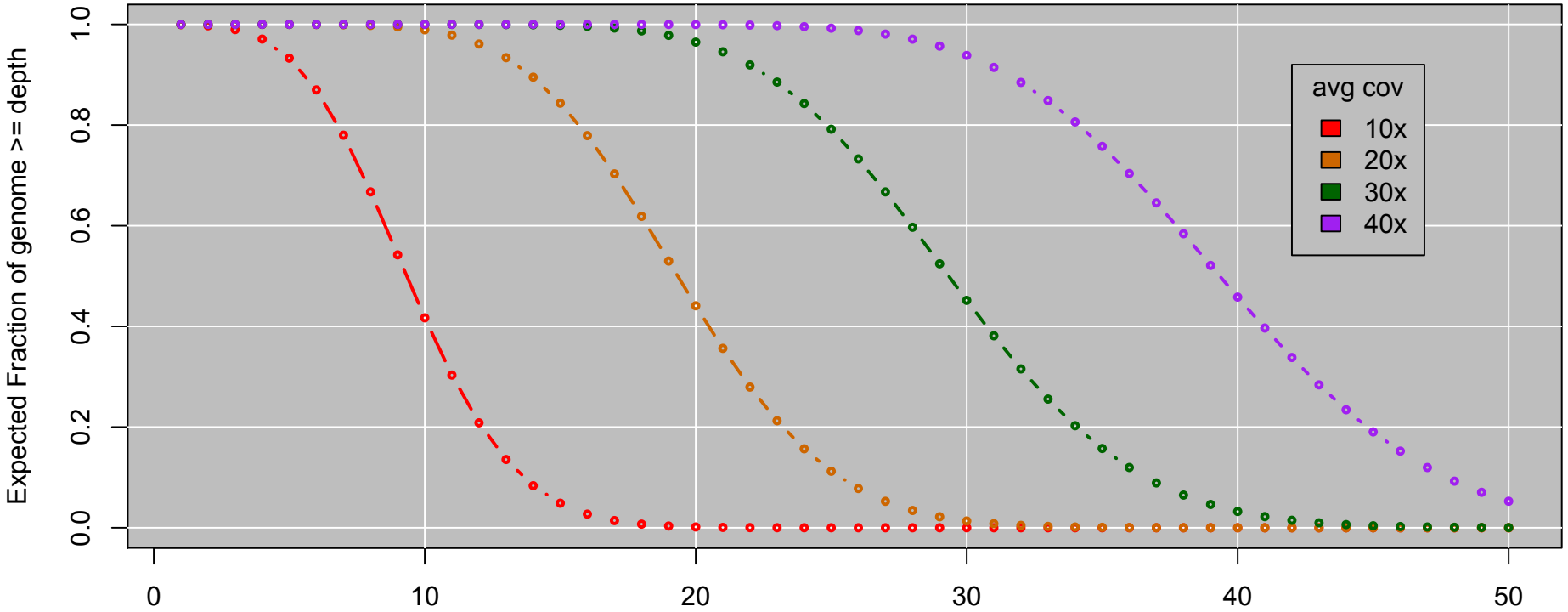
Resembles a normal distribution, but over the positive values, and with only a single parameter.

Key property:

- ***The standard deviation is the square root of the mean.***



Genome Coverage Distribution



Expect Poisson distribution on depth

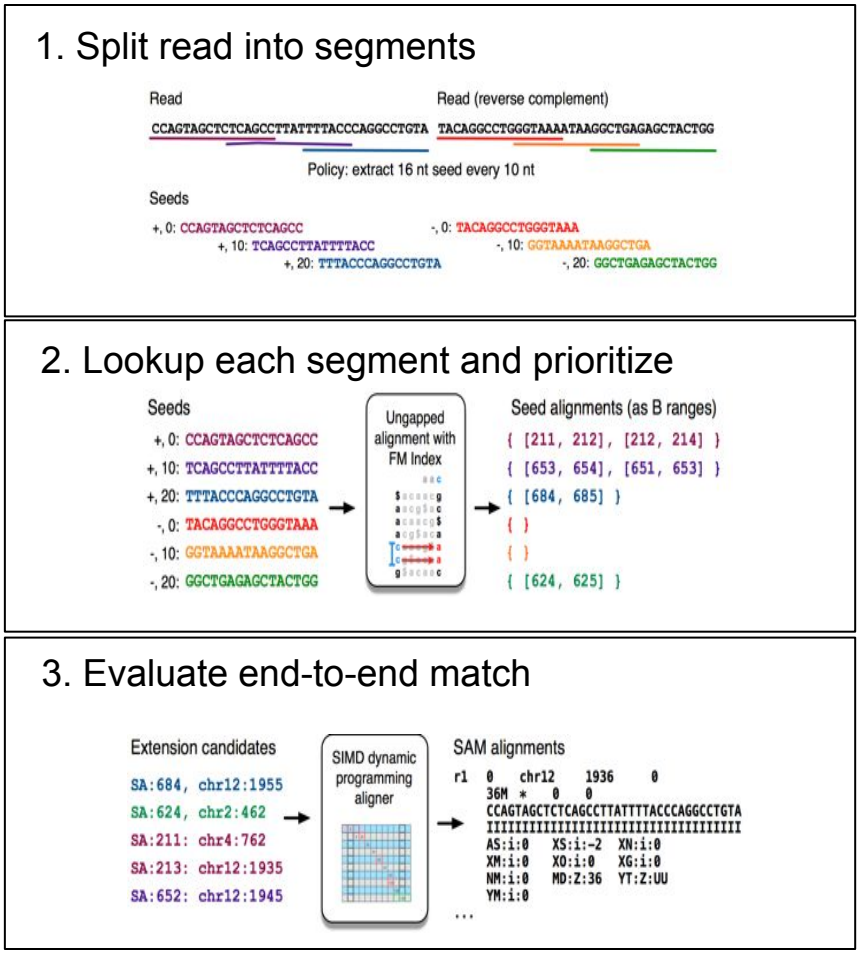
- Standard Deviation = sqrt(cov)

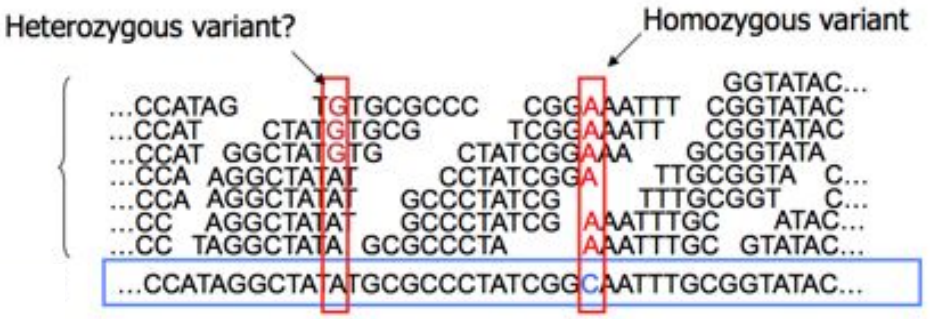
This is the mathematically model => reality may be much worse

- Double your coverage for diploid genomes
- Can use somewhat lower coverage in a population to find common variants

Algorithms for Mapping & Genotyping

QB Week 1: Sept 29





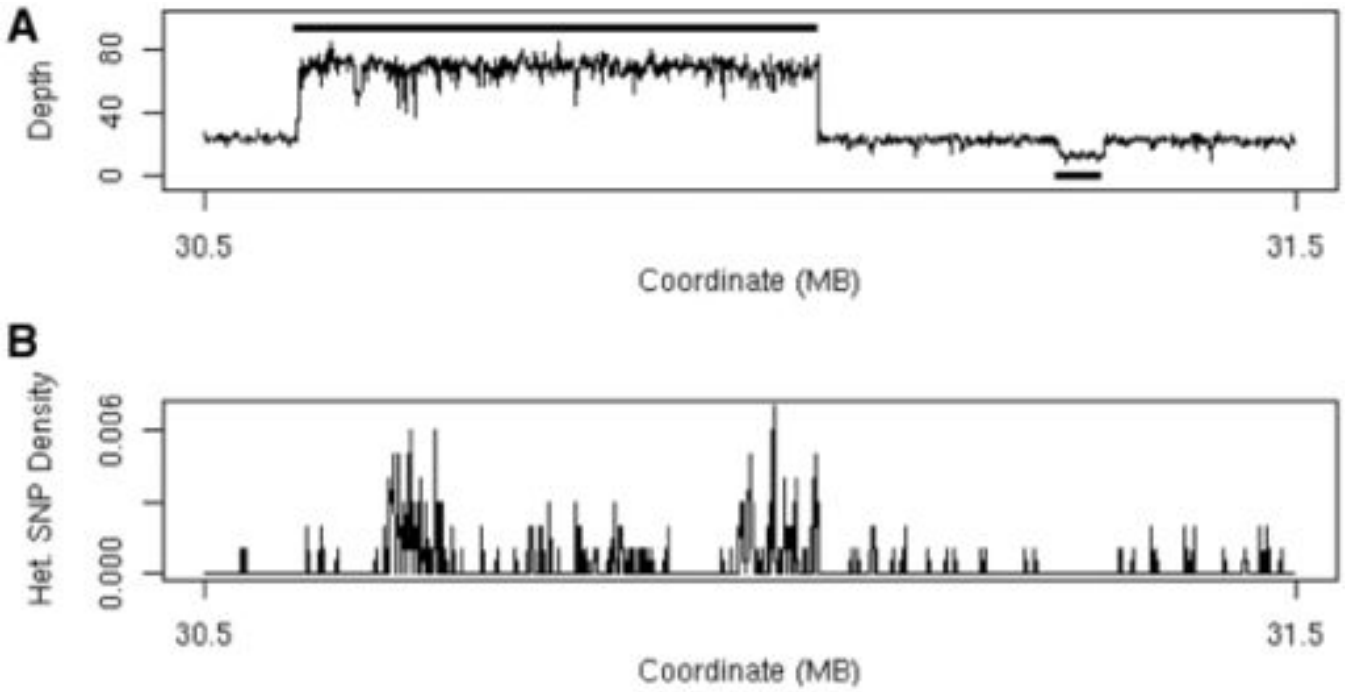
- Distinguishing SNPs from sequencing error typically a likelihood test of the coverage
  - Hardest to distinguish between errors and heterozygous SNP.
  - Coverage is the most important factor!
    - Target at least 10x, 30x more reliable

Fast gapped-read alignment with Bowtie 2
Langmead & Salzberg. (2012) *Nature Methods*. 9:357-359.

The Sequence Alignment/Map format and SAMtools
Li H et al. (2009) *Bioinformatics*. 25:16 2078-9

CNV calling

Beware of (Systematic) Errors

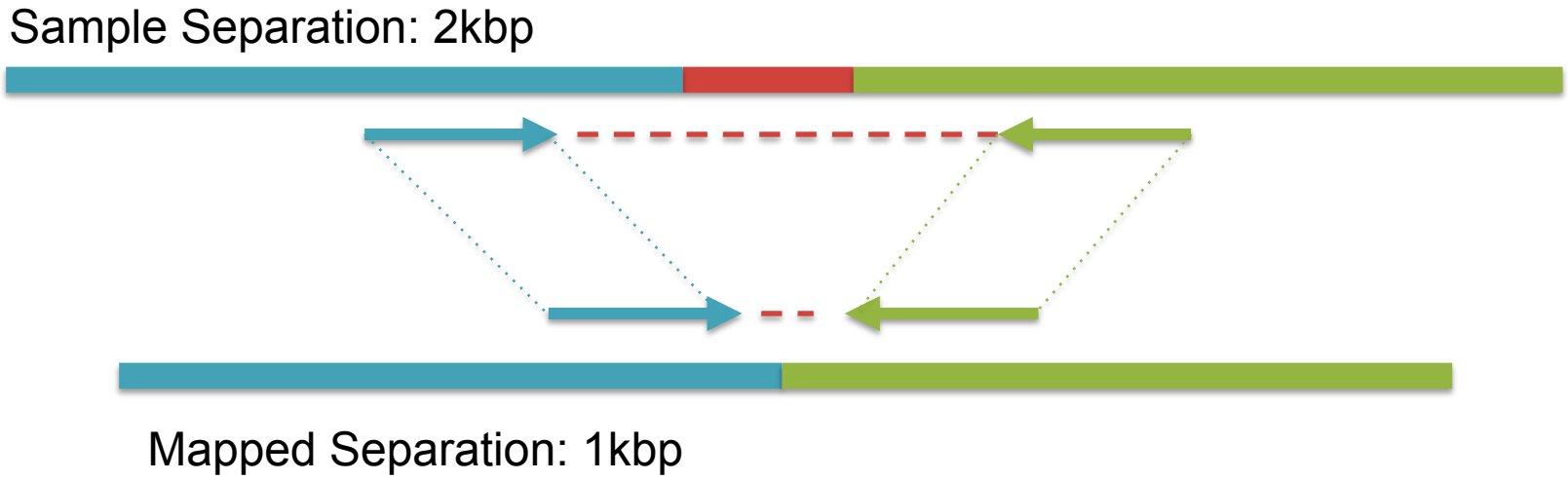


(A) Plot of sequencing depth across a one megabase region of A/J chromosome 17 clearly shows both a region of 3-fold increased copy number (30.6–31.1 Mb) and a region of decreased copy number (at 31.3 Mb).

Simpson J T et al. Bioinformatics 2010;26:565-567

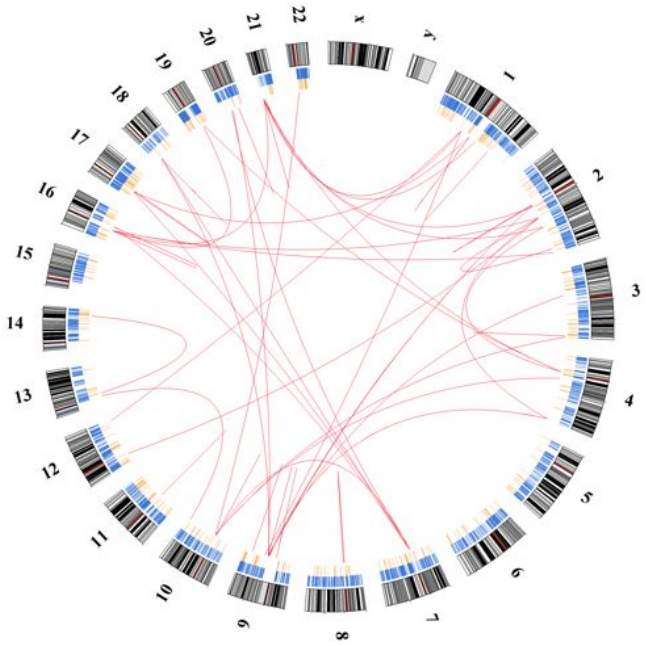
- Identify CNVs through increased depth of coverage & increased heterozygosity
  - Segment coverage levels into discrete steps
  - Be careful of GC biases and mapping biases of repeats

Structural Variations



SVs tend to be flanked by repeats, making it hard to localize

- Cannot trust results from a single compress/expanded mate, look for a cluster of them
- Longer reads are the key to resolving them



Circos plot of high confidence SVs specific to esophageal cancer sample

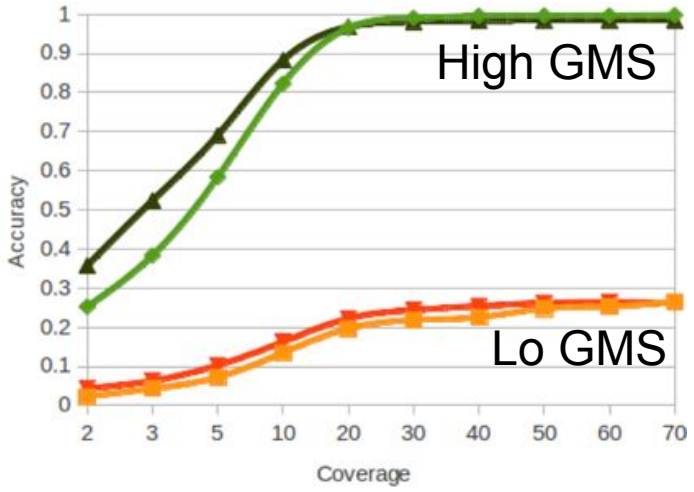
- Red: SV links
- Orange: 375 cancer genes
- Blue: 4950 disease genes

Beware of Mapping Errors

- Short read mapping is a essential for identifying mutations in the genome
  - Not every base of the genome can mapped equally well, especially because of repeats

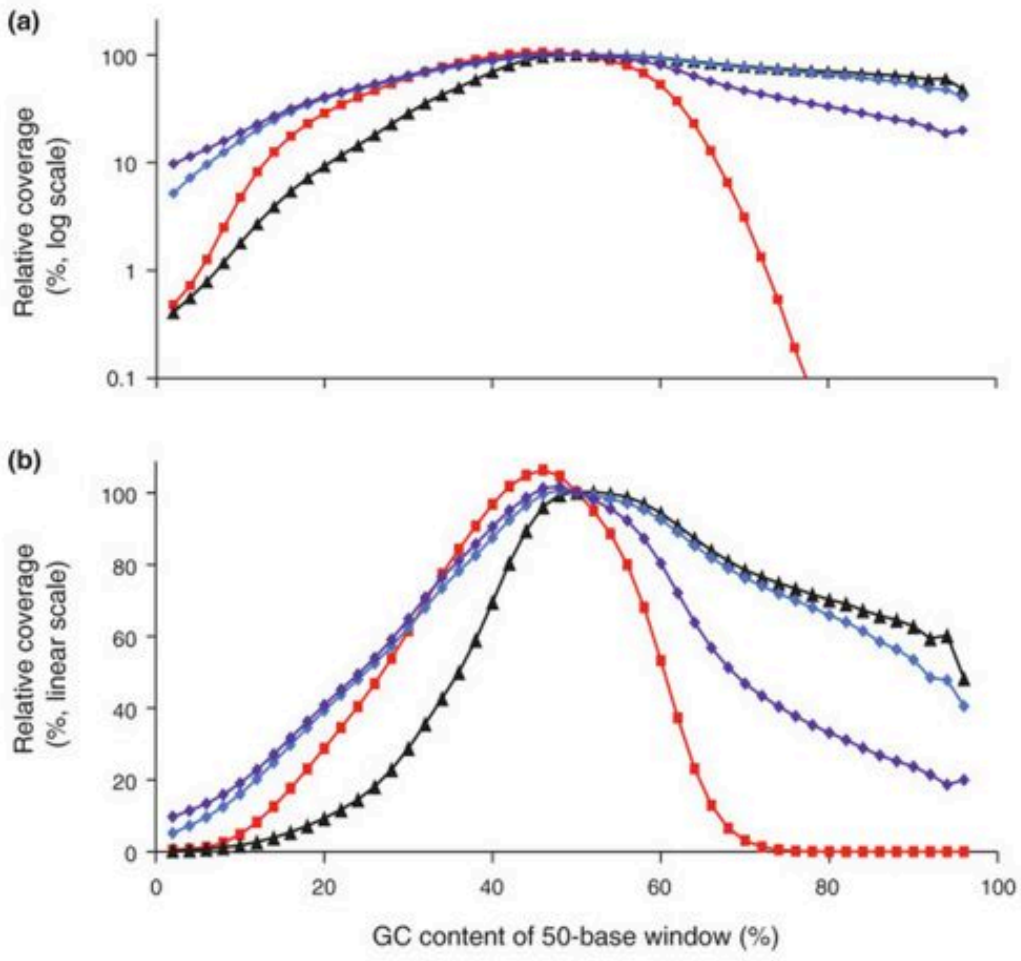
- Introduced a new probabilistic metric - the Genome Mappability Score - that quantifies how reliably reads can be mapped to every position in the genome
  - We have little power to measure 11-13% of the human genome, including of known clinically relevant variations
  - Errors in variation discovery are dominated by errors in low GMS regions

Species (build)	size	paired/single	whole (%)	transcription (%)
yeast (sc2)	12 Mbp	paired	94.85	95.04
		single	94.25	94.62
fly (dm3)	130 Mbp	paired	90.52	96.14
		single	89.70	95.94
mouse (mm9)	2.7 Gbp	paired	89.39	96.03
		single	87.47	94.75
human (hg19)	3.0 Gbp	paired	89.02	97.40
		single	87.79	96.38



Genomic Dark Matter: The reliability of short read mapping illustrated by the GMS.
Lee and Schatz (2012) *Bioinformatics*. doi: 10.1093/bioinformatics/bts330

Beware of GC Biases

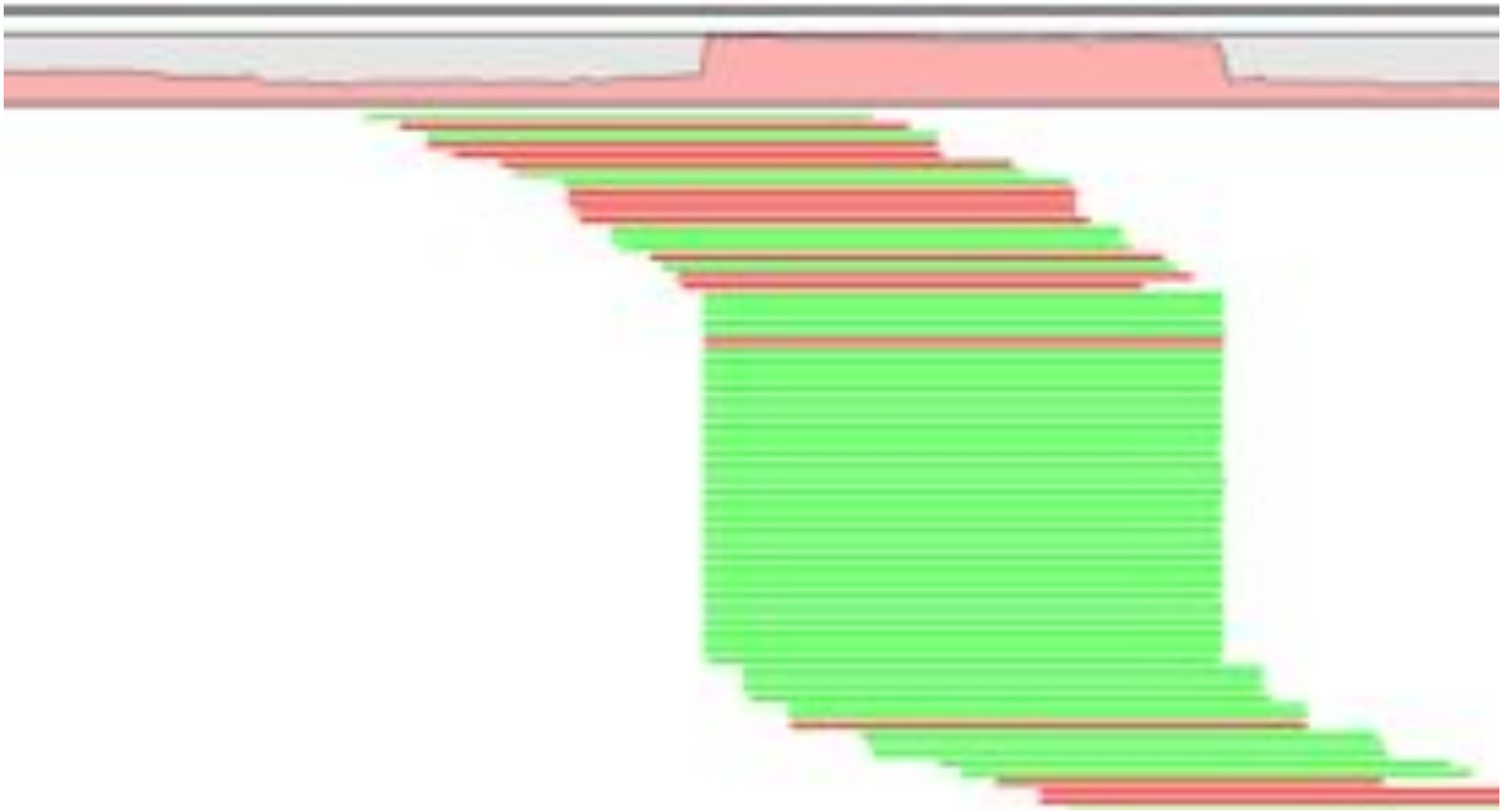


Illumina sequencing does not produce uniform coverage over the genome

- Coverage of extremely high or extremely low GC content will have reduced coverage in Illumina sequencing
- Biases primarily introduced during PCR; lower temperatures, slower heating, and fewer rounds minimize biases
- This makes it very difficult to identify variants (SNPs, CNVs, etc) in certain regions of the genome

Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries.
Aird et al. (2011) *Genome Biology.* 12:R18.

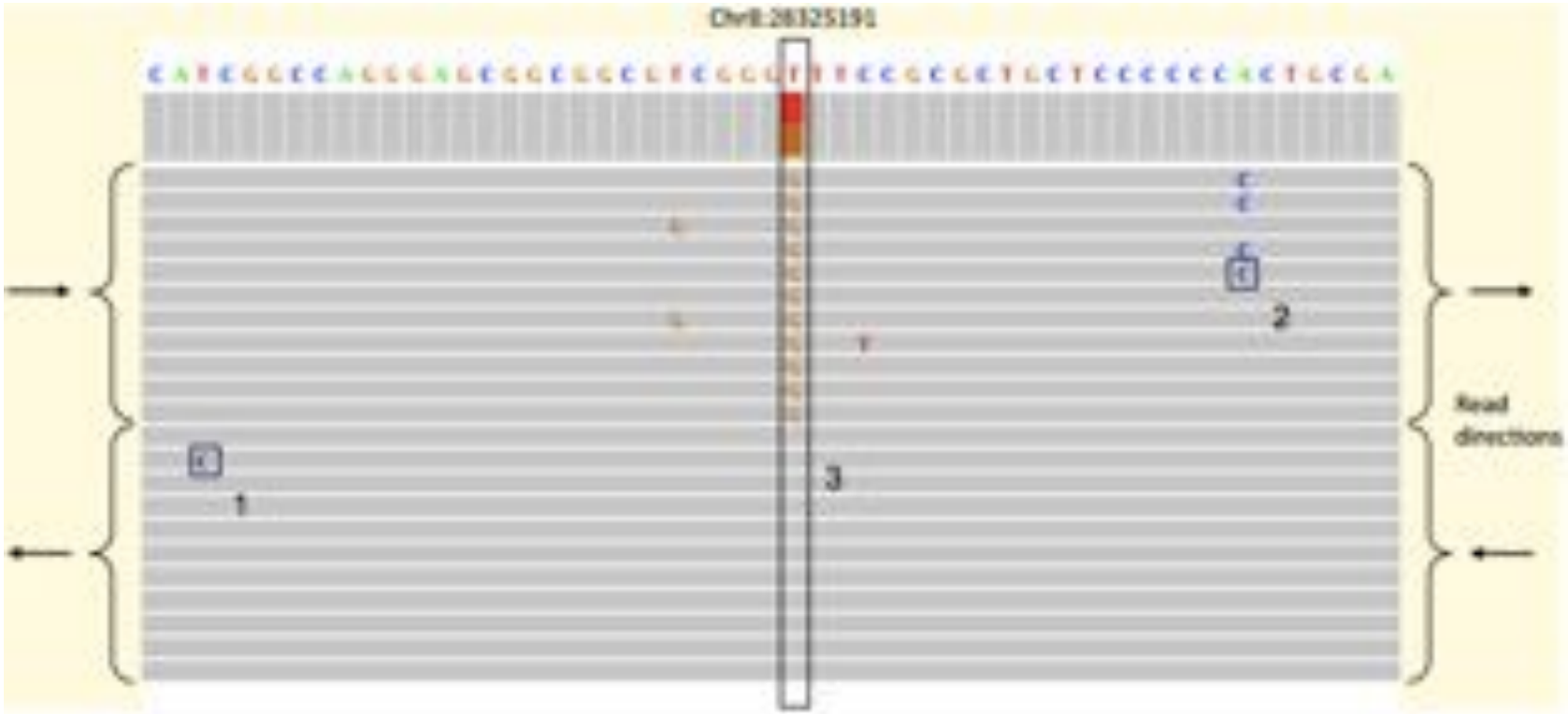
Beware of Duplicate Reads



The Sequence alignment/map (SAM) format and SAMtools.
Li et al. (2009) *Bioinformatics.* 25:2078-9

Picard: http://picard.sourceforge.net

Beware of (Systematic) Errors



Identification and correction of systematic error in high-throughput sequence data
Meacham et al. (2011) *BMC Bioinformatics.* 12:451

A closer look at RNA editing.
Lior Pachter (2012) *Nature Biotechnology.* 30:246-247

Genetic Basis of Autism Spectrum Disorders

Complex disorders of brain development

- Characterized by difficulties in social interaction, verbal and nonverbal communication and repetitive behaviors.
- Have their roots in very early brain development, and the most obvious signs of autism and symptoms of autism tend to emerge between 2 and 3 years of age.

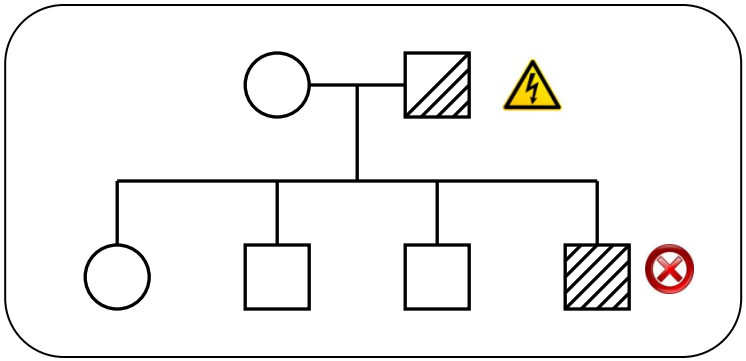
U.S. CDC identify around 1 in 68 American children as on the autism spectrum

- Ten-fold increase in prevalence in 40 years, only partly explained by improved diagnosis and awareness.
- Studies also show that autism is four to five times more common among boys than girls.
- Specific causes remain elusive

What is Autism?
http://www.autismspeaks.org/what-autism

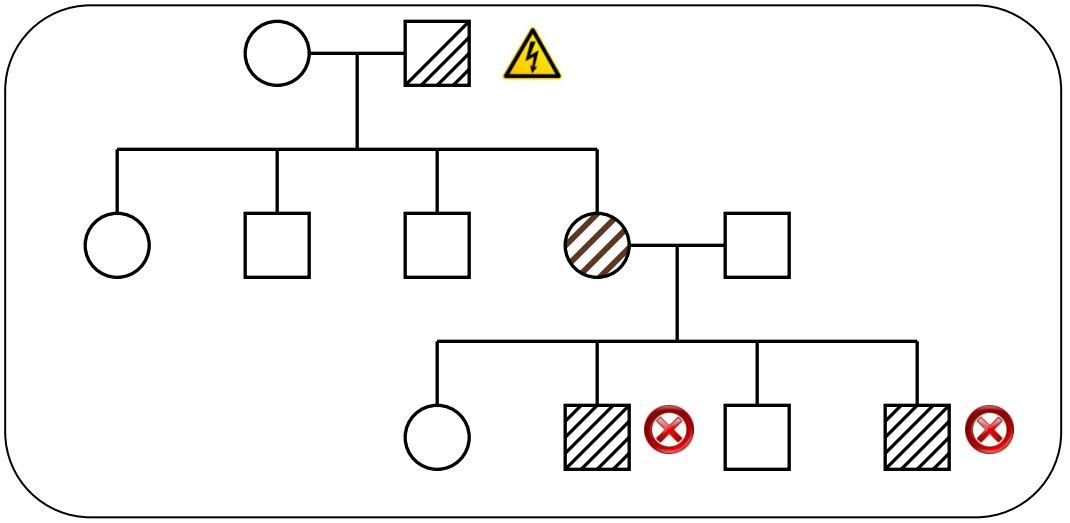
Unified Model of Autism

Sporadic Autism: 1 in 100



Prediction: De novo mutations of high penetrance contributes to autism, especially in low risk families with no history of autism.

Familial Autism: 90% concordance in twins



Legend

Sporadic mutation

Fails to procreate

A unified genetic theory for sporadic and inherited autism
Zhao *et al.* (2007) *PNAS. 104(31)12831-12836.*

Searching for the genetics behind human disorders and plant phenotypes

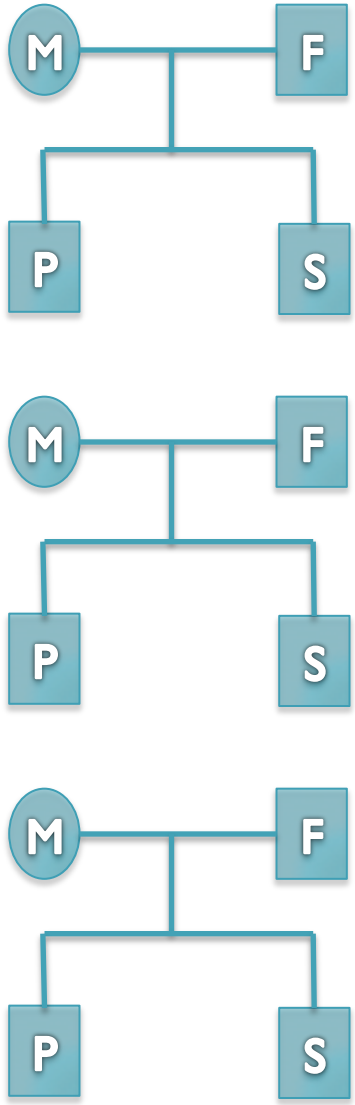
Search Strategy

- Currently uses whole exome short read resequencing for economic reasons

- Collaborate with Lyon, McCombie, Tuveson, and Wigler labs to examine the genetic basis of cancer, ASD, and other psychiatric disorders

- Also collaborating with the Lippman, Ware, and Gingeras labs to study high value crops

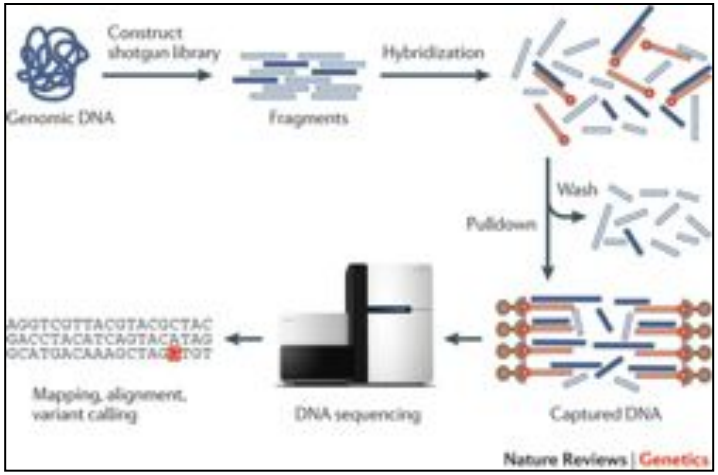
Are there any genetic variants present in affected individuals, that are not present or are present at a substantially reduced rate in their relatives?

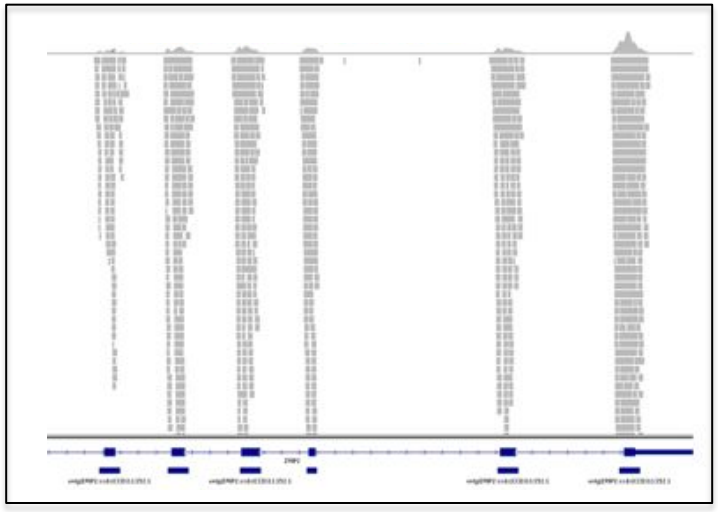


Exome-Capture Sequencing

Exome-capture reduces the costs of sequencing

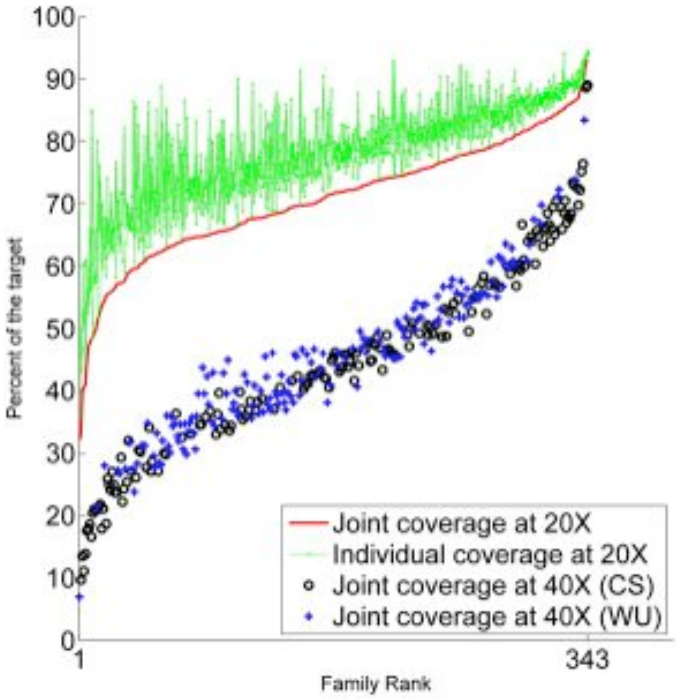
- Currently targets around 50Mbp of sequence: all exons plus flanking regions
- WGS currently costs ~$2000 per sample, while WES currently costs ~$400 per sample
- Coverage is highly localized around genes, although will get sparse coverage throughout rest of genome





Exome sequencing as a tool for Mendelian disease gene discovery
Bamshad et al. (2011) *Nature Reviews Genetics*. 12, 745-755

Exome sequencing of the SSC



The year 2012 was an exciting year for autism genetics

- 3 reports of >593 families from the Simons Simplex Collection (parents plus one child with autism and one non-autistic sibling)
- All attempted to find mutations enriched in the autistic children
- ***All used poor or no tools for indels:***
  - Iossifov (343 families) and O'Roak (50 families) used GATK UnifiedGenotype
  - Sanders (200 families) didn't attempt

De novo gene disruptions in children on the autism spectrum
Iossifov *et al.* (2012) *Neuron*. 74:2 285-299

De novo mutations revealed by whole-exome sequencing are strongly associated with autism
Sanders *et al.* (2012) *Nature*. 485, 237–241.

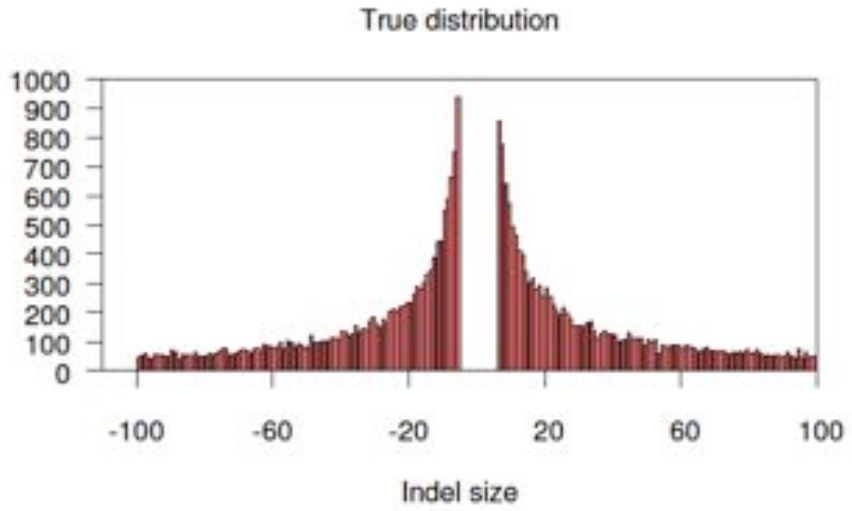
Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations
O'Roak *et al.* (2012) *Nature*. 485, 246–250.

Variation Detection Complexity

SNPs + Short Indels

High precision and sensitivity

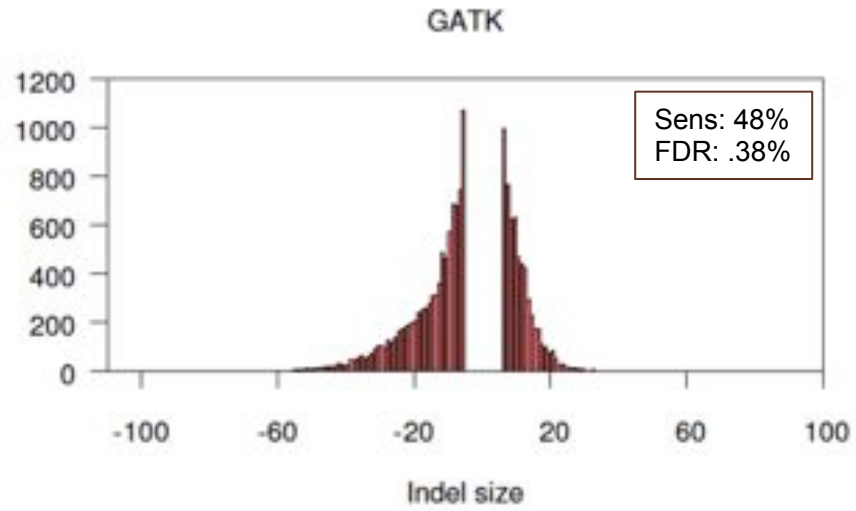
```
..TTTAGAATAG-CGAGTGC...
     ||||||| ||||
     AGAATAGGCGAG
```



"Long" Indels (>5bp)

Reduced precision and sensitivity

```
..TTTAG--------AGTGC...
  |||||         |||||
  TTTAGAATAGGC  |||||
        ATAGGCGAGTGC
```



Analysis confounded by sequencing errors, localized repeats, allele biases, and mismapped reads

Scalpel: Haplotype Microassembly

DNA sequence **micro-assembly** pipeline for accurate detection and validation of *de novo* mutations (SNPs, indels) within exome-capture data.

Features

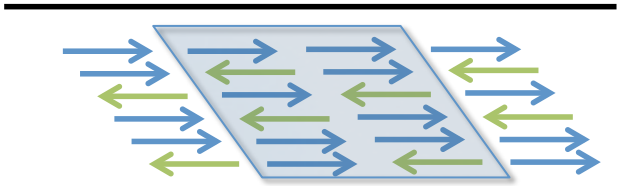
1. Combine mapping and assembly
2. Exhaustive search of haplotypes
3. De novo mutations

NRXN1 *de novo* SNP
(auSSC12501 chr2:50724605)

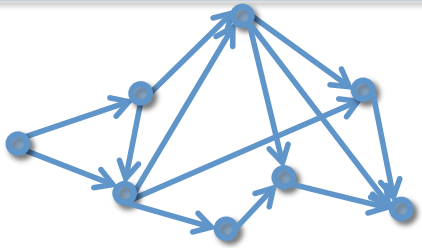
Accurate de novo and transmitted indel detection in exome-capture data using microassembly.
Narzisi et al. (2014) *Nature Methods*. doi:10.1038/nmeth.3069

Scalpel Algorithm

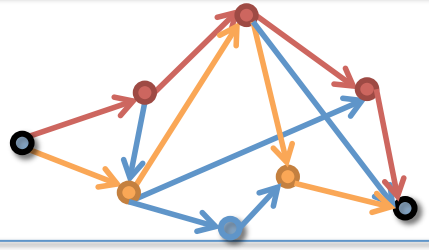
Extract reads mapping within the exon including (1) well-mapped reads, (2) soft-clipped reads, and (3) anchored pairs



Decompose reads into overlapping *k*-mers and construct de Bruijn graph from the reads



Find end-to-end haplotype paths spanning the region



Align assembled sequences to reference to detect mutations

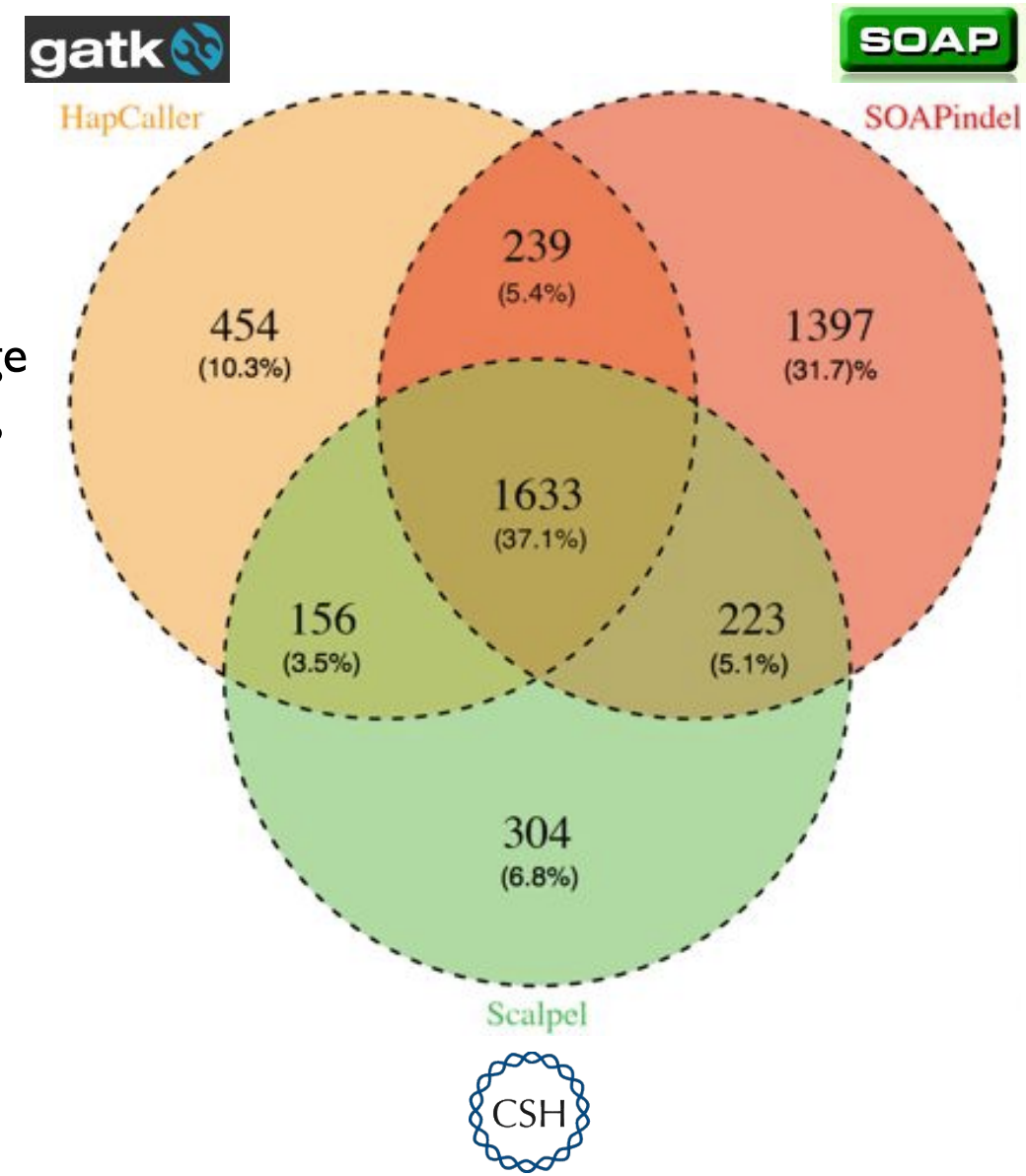
Experimental Analysis & Validation

Selected one deep coverage exome for deep analysis

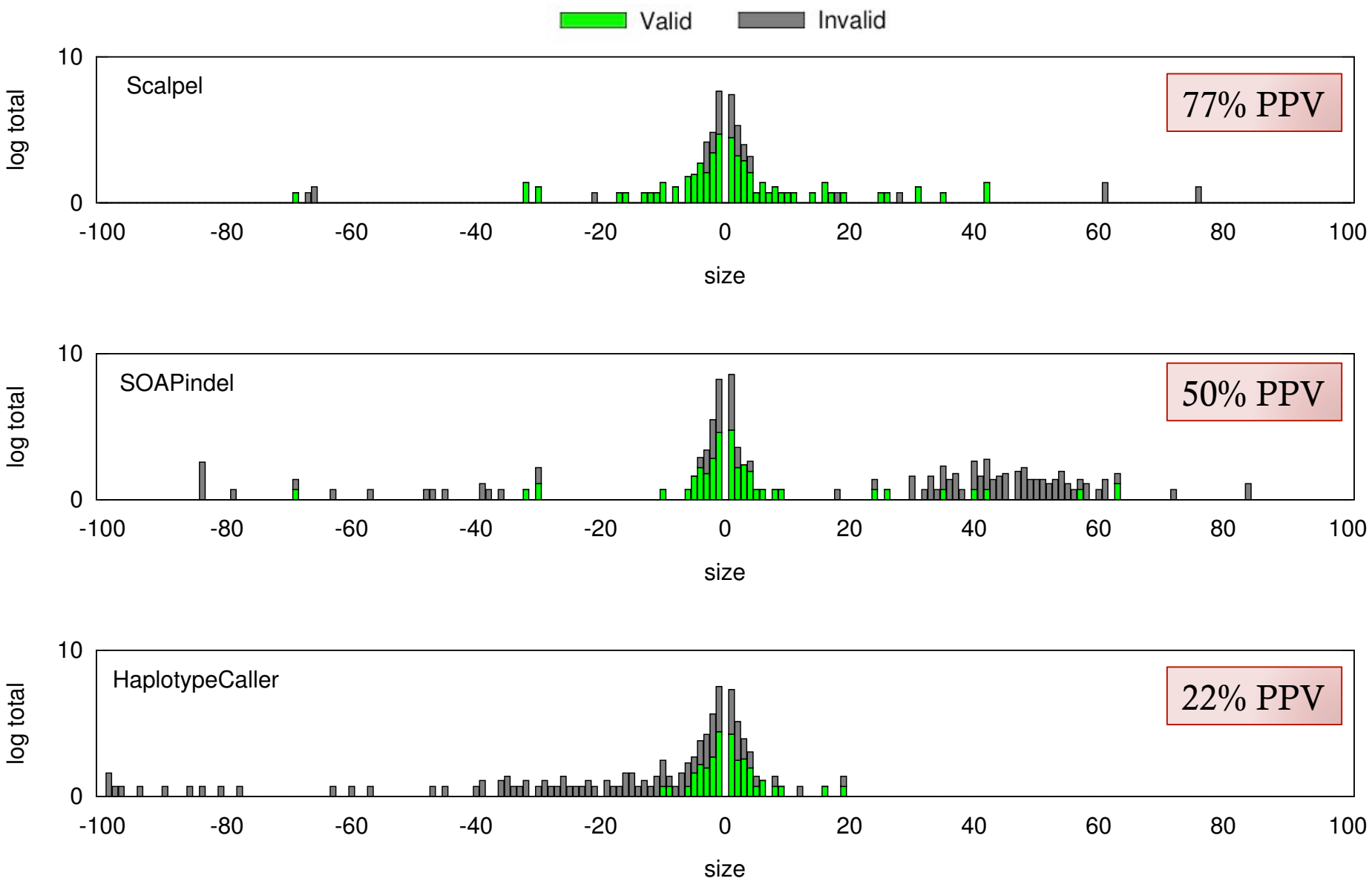
- Individual was diagnosed with ADHD and turrets syndrome
- 80% of the target at >20x coverage
- Evaluated with Scalpel, SOAPindel, and GATK Haplotype Caller

1000 indels selected for validation

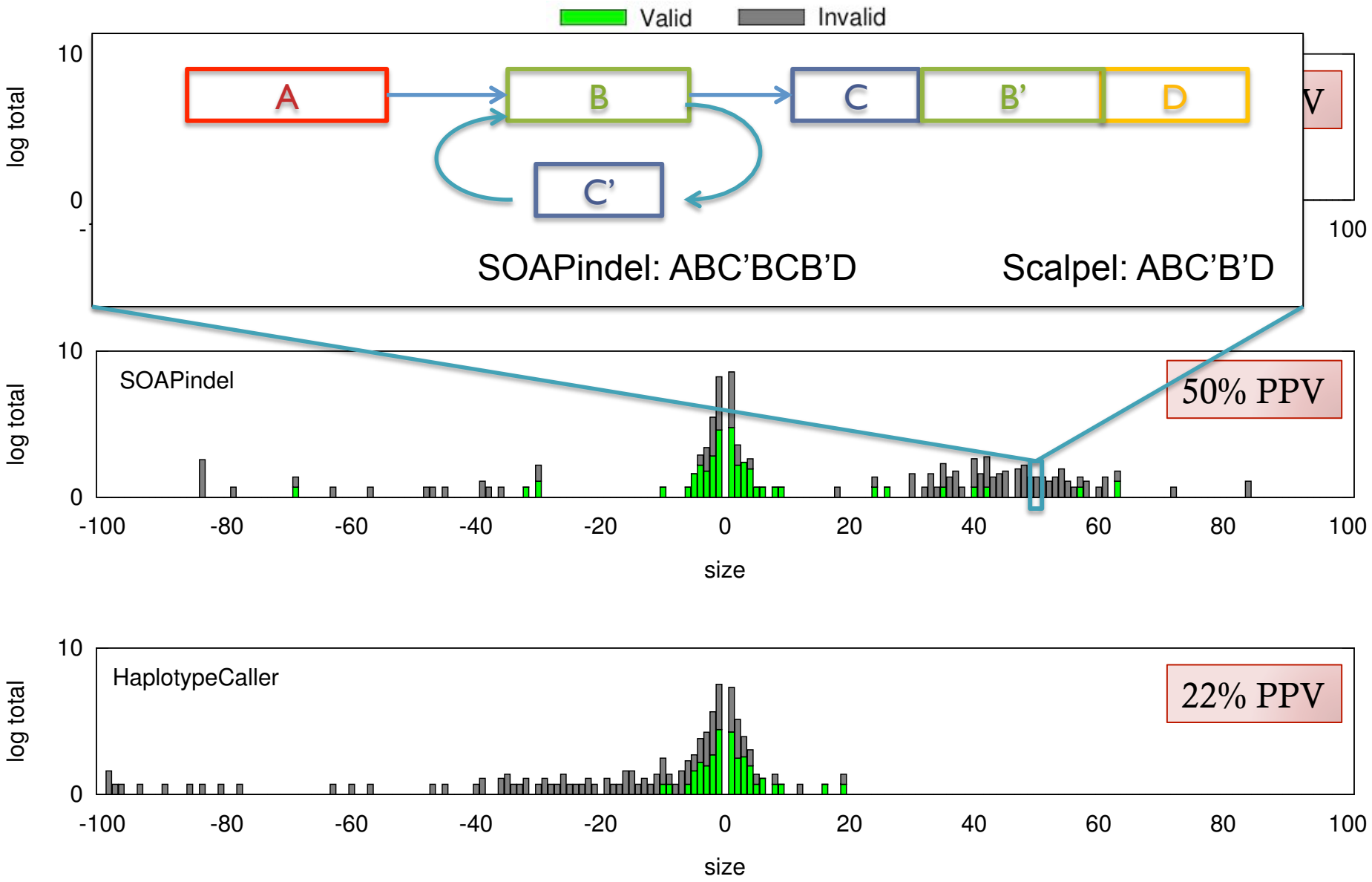
- 200 Scalpel
- 200 GATK Haplotype Caller
- 200 SOAPindel
- 200 within the intersection
- 200 long indels (>30bp)



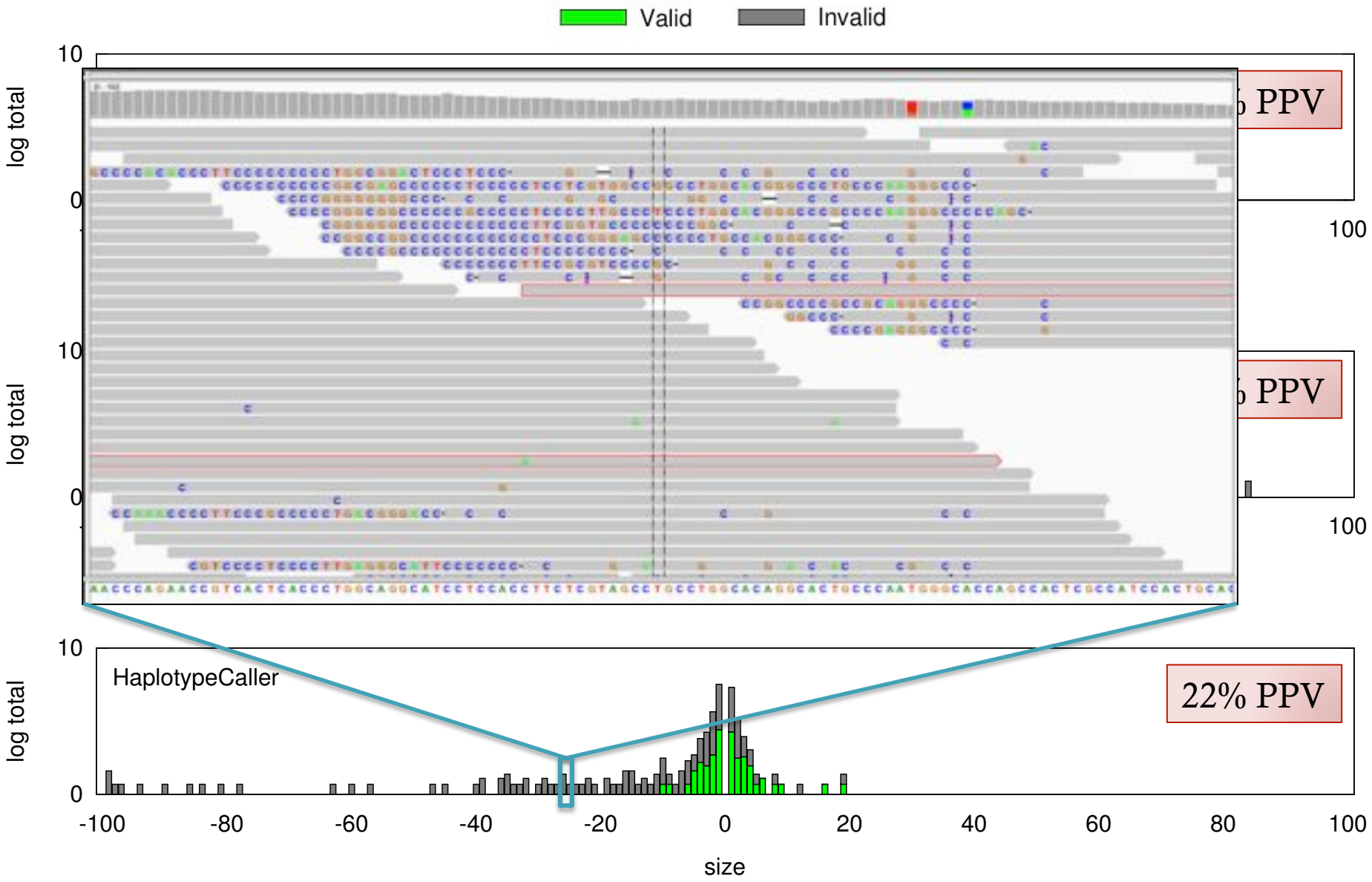
Scalpel Indel Validation



Scalpel Indel Validation



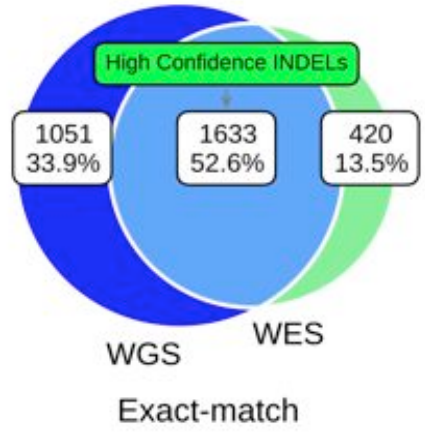
Scalpel Indel Validation



Refined indel analysis

Examine sources of indel errors

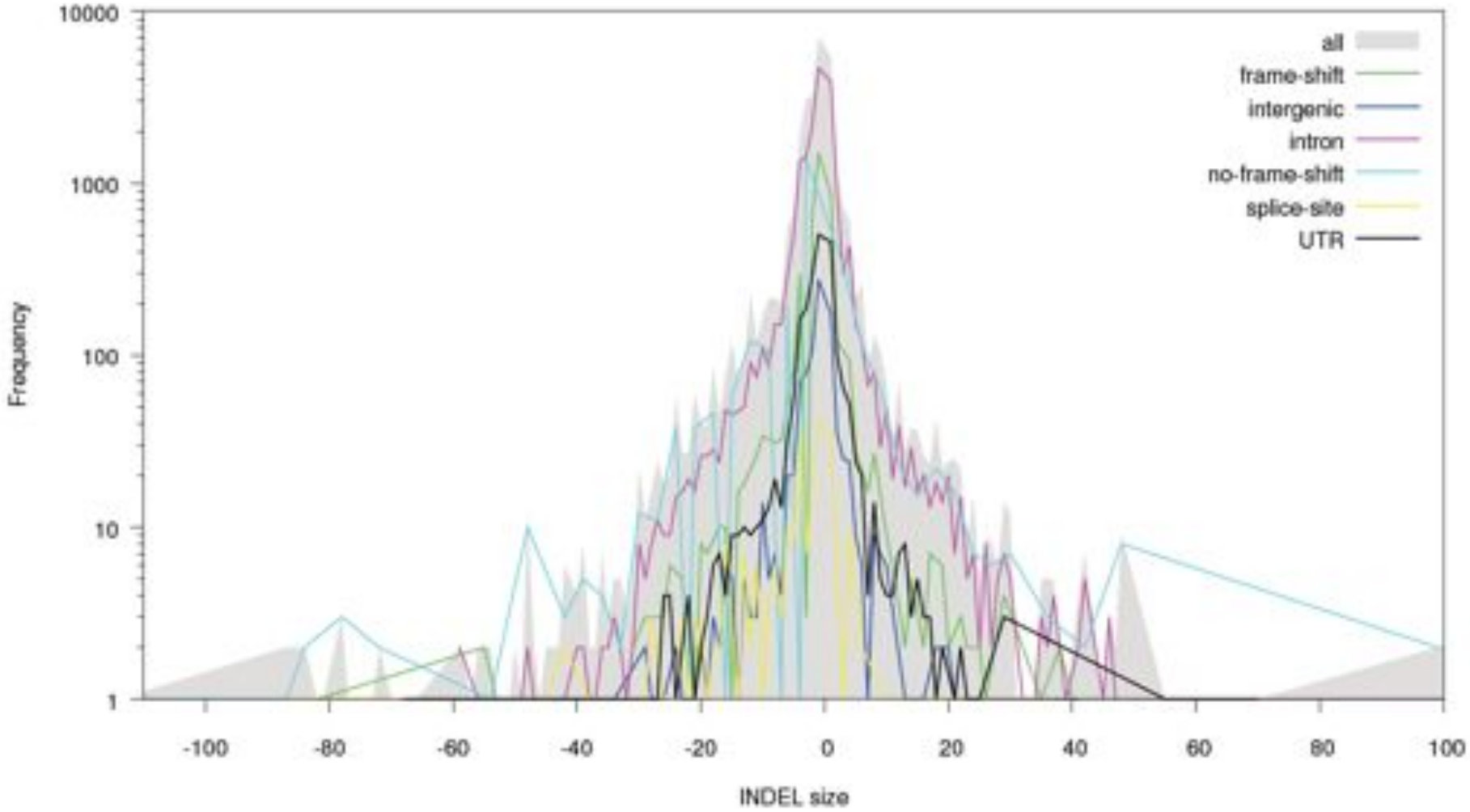
- Experimental validation of indels called from 30x whole genome vs. 110x whole exome of the same sample
- Most of the errors due to short microsatellite errors introduced during exome capture, also misses most long indels
- Recommend WGS for indel analysis instead



	All INDELs	Valid	PPV	INDELs >5bp	Valid (>5bp)	PPV (>5bp)
Intersection	160	152	95.0%	18	18	100%
WGS	145	122	84.1%	33	25	75.8%
WES	161	91	56.5%	1	1	100%

Reducing INDEL calling errors in whole-genome and exome sequencing data
Fang, H, Wu, Y, Narzisi, G, O'Rawe, JA, Jimenez Barrón LT, Rosenbaum, J, Ronemus, M, Iossifov I, Schatz, MC§, Lyon, GL§
http://www.biorxiv.org/content/early/2014/06/10/006148

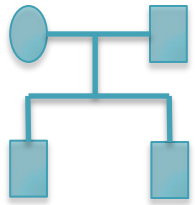
Revised Analysis of the SSC



Constructed database of >1M transmitted and de novo indels
Many new gene candidates identified, population analysis underway

De novo mutation discovery and validation

De novo mutations:
Sequences not inherited from your parents.



```
Reference:    ...TCAAATCCTTTTAATAAAGAAGAGCTGACA...

Father(1):    ...TCAAATCCTTTTAATAAAGAAGAGCTGACA...
Father(2):    ...TCAAATCCTTTTAATAAAGAAGAGCTGACA...

Mother(1):    ...TCAAATCCTTTTAATAAAGAAGAGCTGACA...
Mother(2):    ...TCAAATCCTTTTAATAAAGAAGAGCTGACA...

Sibling(1):   ...TCAAATCCTTTTAATAAAGAAGAGCTGACA...
Sibling(2):   ...TCAAATCCTTTTAATAAAGAAGAGCTGACA...

Proband(1):   ...TCAAATCCTTTTAATAAAGAAGAGCTGACA...
Proband(2):   ...TCAAATCCTTTTAAT****AAGAGCTGACA...
```

4bp heterozygous deletion at chr15:93524061 CHD2

De novo Genetics of Autism

- In 593 family quads so far, we see significant enrichment in de novo ***likely gene disruptions (LGDs)*** in the autistic kids
  - Overall rate basically 1:1
  - 2:1 enrichment in frameshift indels (35:16)
- Confirmed trends observed in previous studies, contributed dozens of new autism candidate genes.
  - 8 out of 35 indel LGDs in autistic children overlapped with the 842 FMRP-associated genes
  - Trends further confirmed in larger study over the entire collection that is currently under review

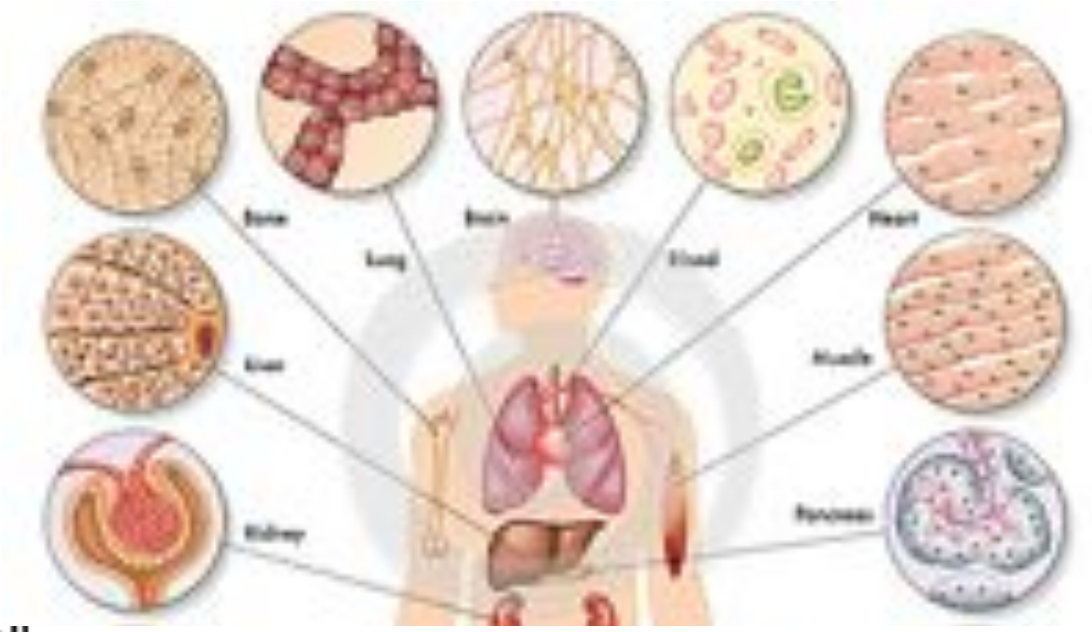
Accurate de novo and transmitted indel detection in exome-capture data using microassembly.
Narzisi et al. (2014) *Nature Methods* doi:10.1038/nmeth.3069

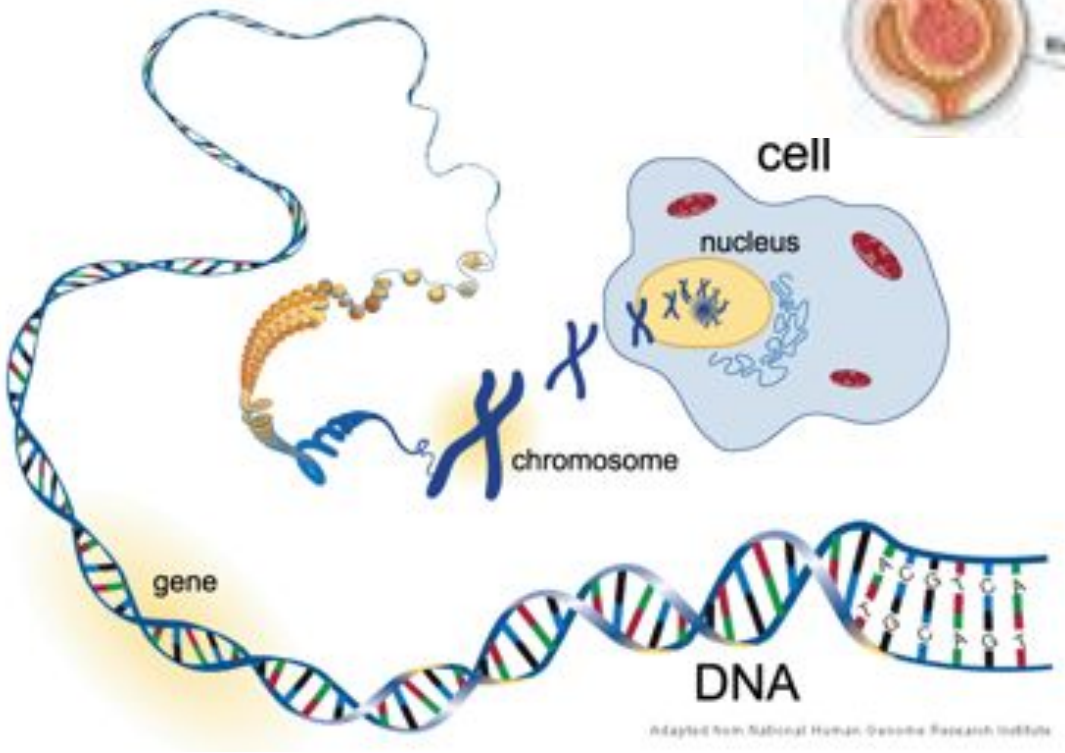
The burden of de novo coding mutations in autism spectrum disorders.
Iossifov *et al* (2014) *Under review.*

Break

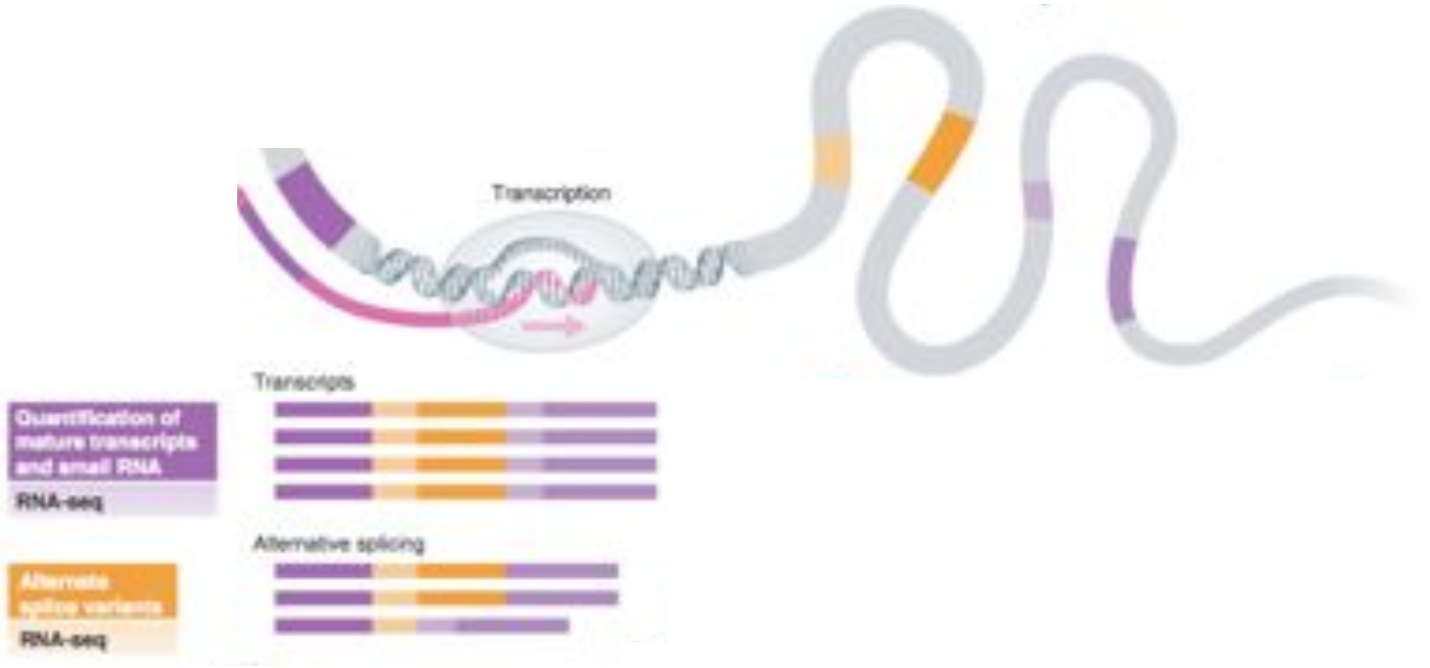
Cells & DNA

Each cell of your body contains an exact copy of your 3 billion base pair genome.





Your specific nucleotide sequence encodes the genetic program for your cells and ultimately your traits



Soon et al., Molecular Systems Biology, 2013

Sequencing Assays

The *Seq List (in chronological order)

1. Gregory E. Crawford et al., "Genome-wide Mapping of DNase Hypersensitive Sites Using Massively Parallel Signature Sequencing (MPSS)," Genome Research 16, no. 1 (January 1, 2006): 123–131, doi:10.1101/gr.4074106.
2. David S. Johnson et al., "Genome-Wide Mapping of in Vivo Protein-DNA Interactions," Science 316, no. 5830 (June 8, 2007): 1497–1502, doi:10.1126/science.1141319.
3. Tarjei S. Mikkelsen et al., "Genome-wide Maps of Chromatin State in Pluripotent and Lineage-committed Cells," Nature 448, no. 7153 (August 2, 2007): 553–560, doi:10.1038/nature06008.
4. Thomas A. Down et al., "A Bayesian Deconvolution Strategy for Immunoprecipitation-based DNA Methylome Analysis," Nature Biotechnology 26, no. 7 (July 2008): 779–785, doi:10.1038/nbt1414.
5. Ali Mortazavi et al., "Mapping and Quantifying Mammalian Transcriptomes by RNA-Seq," Nature Methods 5, no. 7 (July 2008): 621–628, doi:10.1038/nmeth.1226.
6. Nathan A. Baird et al., "Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers," PLoS ONE 3, no. 10 (October 13, 2008): e3376, doi:10.1371/journal.pone.0003376.
7. Leighton J. Core, Joshua J. Waterfall, and John T. Lis, "Nascent RNA Sequencing Reveals Widespread Pausing and Divergent Initiation at Human Promoters," Science 322, no. 5909 (December 19, 2008): 1845–1848, doi:10.1126/science.1162228.
8. Chao Xie and Martti T. Tammi, "CNV-seq, a New Method to Detect Copy Number Variation Using High-throughput Sequencing," BMC Bioinformatics 10, no. 1 (March 6, 2009): 80, doi:10.1186/1471-2105-10-80.
9. Jay R. Hesselberth et al., "Global Mapping of protein-DNA Interactions in Vivo by Digital Genomic Footprinting," Nature Methods 6, no. 4 (April 2009): 283–289, doi:10.1038/nmeth.1313.
10. Nicholas T. Ingolia et al., "Genome-Wide Analysis in Vivo of Translation with Nucleotide Resolution Using Ribosome Profiling," Science 324, no. 5924 (April 10, 2009): 218–223, doi:10.1126/science.1168978.
11. Alayne L. Brunner et al., "Distinct DNA Methylation Patterns Characterize Differentiated Human Embryonic Stem Cells and Developing Human Fetal Liver," Genome Research 19, no. 6 (June 1, 2009): 1044–1056, doi:10.1101/gr.088773.108.
12. Mayumi Oda et al., "High-resolution Genome-wide Cytosine Methylation Profiling with Simultaneous Copy Number Analysis and Optimization for Limited Cell Numbers," Nucleic Acids Research 37, no. 12 (July 1, 2009): 3829–3839, doi:10.1093/nar/gkp260.
13. Zachary D. Smith et al., "High-throughput Bisulfite Sequencing in Mammalian Genomes," Methods 48, no. 3 (July 2009): 226–232, doi: 10.1016/j.ymeth.2009.05.003.
14. Andrew M. Smith et al., "Quantitative Phenotyping via Deep Barcode Sequencing," Genome Research (July 21, 2009), doi:10.1101/gr.

What is a *Seq assay?

Molecular Biology

Mathematics & Computer Science

Computational Biology

Desired measurement → **reduce to sequencing** → Sequence → Solve inverse problem → Analyze

Creativity

Chemistry & Physics

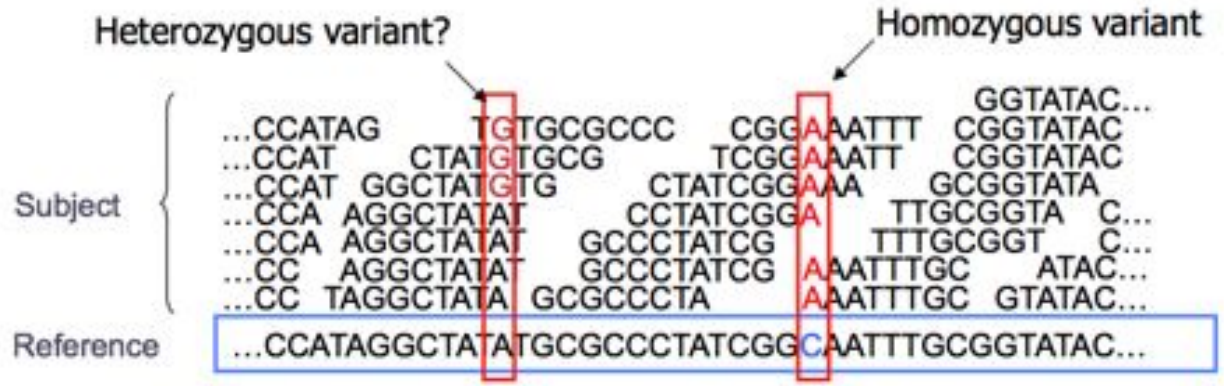
Statistics

Biology

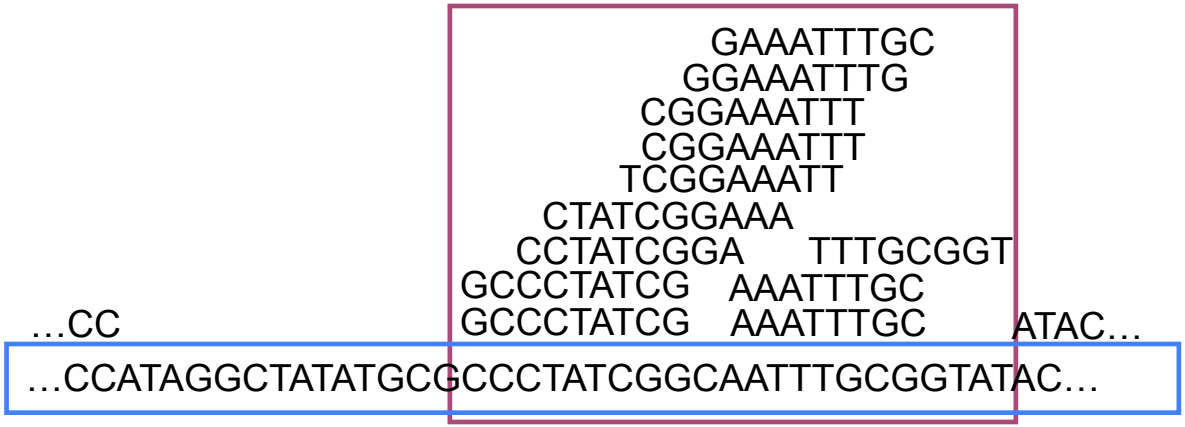
$$\mathbb{P}(f=(p,t,l)) \approx \frac{\lambda_l \cdot \frac{\tau_t}{l(t)} \cdot w_{p|t,l} \cdot \phi_{f|p,t,l}}{\sum_{(q,r,m)\in\mathcal{A}(f)} \lambda_m \cdot \frac{\tau_r}{l(r)} \cdot w_{q|r,m} \cdot \phi_{f|q,r,m}}$$

Short Read Applications

- Genotyping: Identify Variations

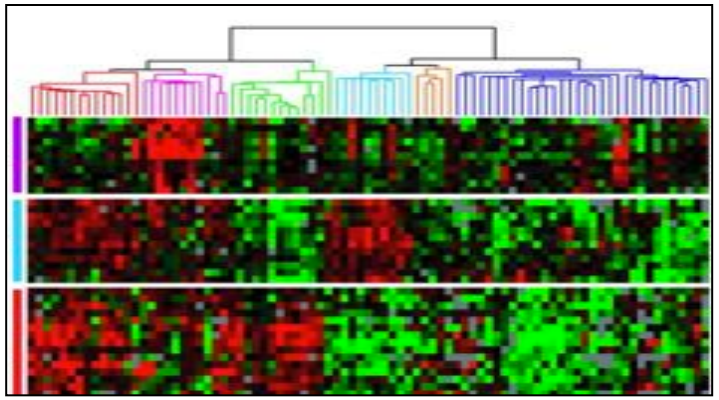


- *-seq: Classify & measure significant peaks

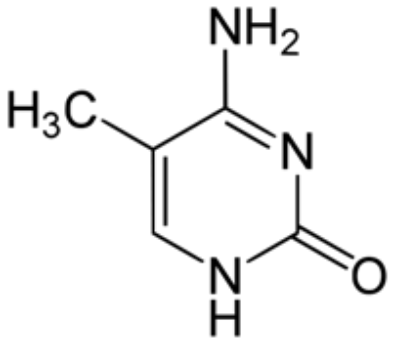


*-seq in 4 short vignettes

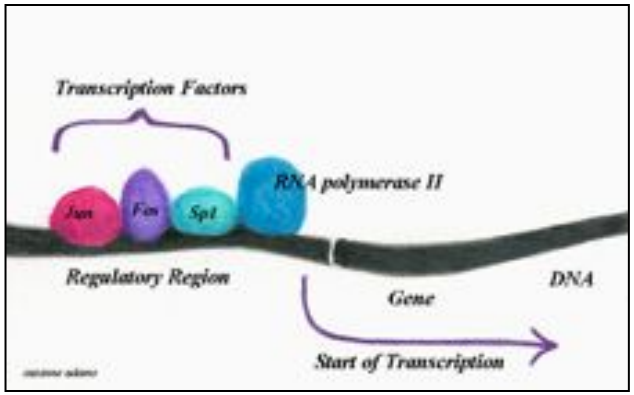
RNA-seq



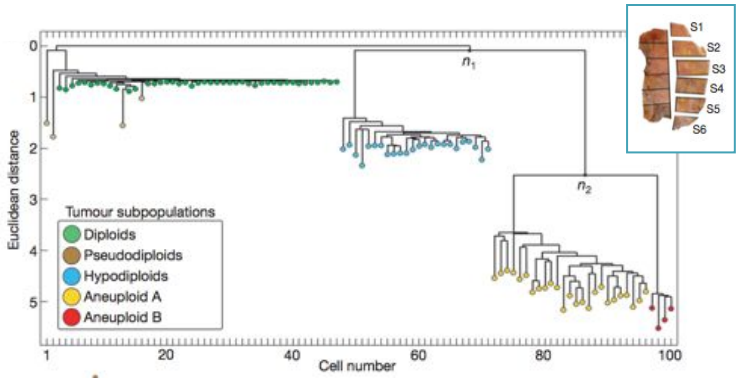
Methyl-seq



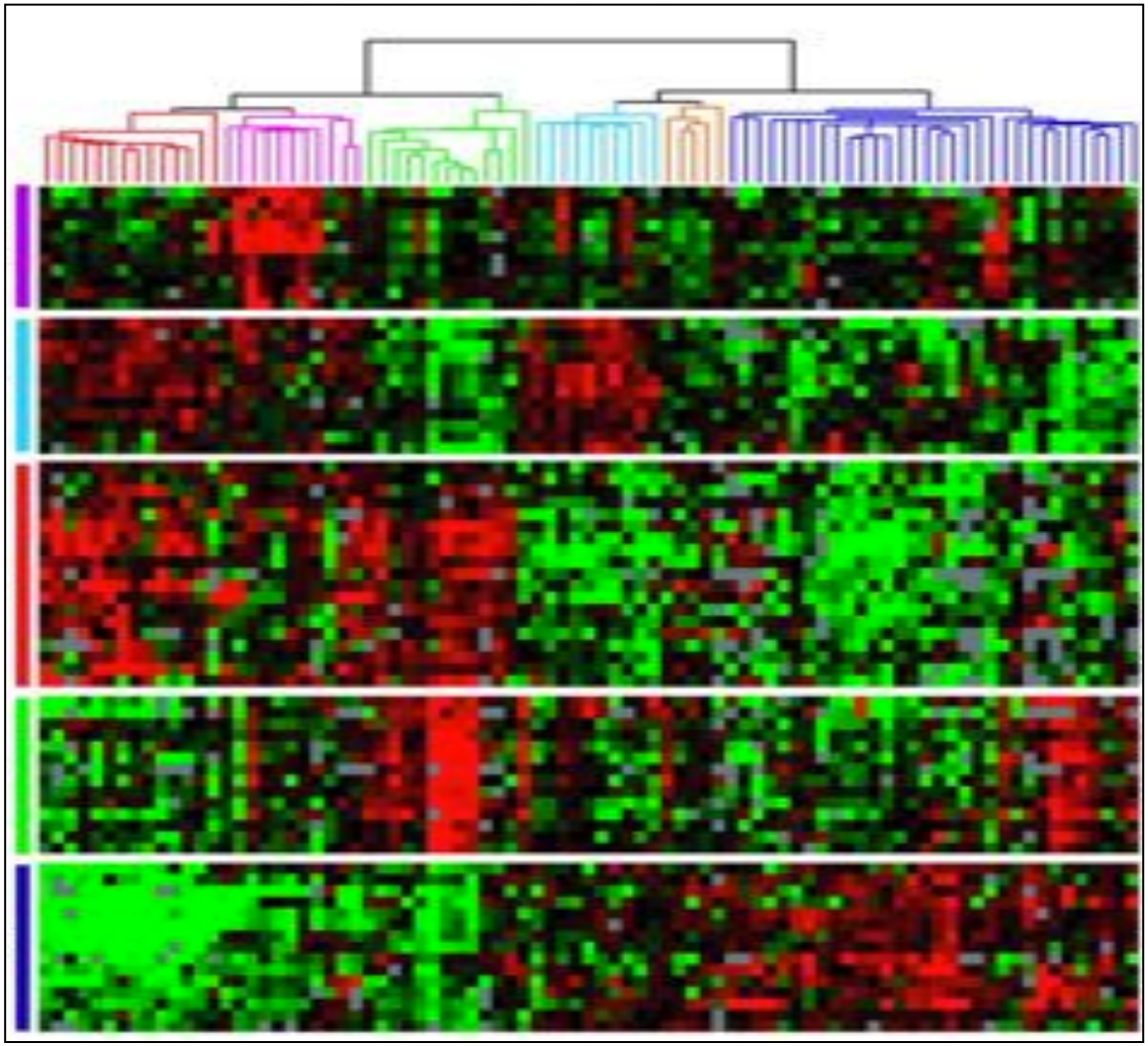
ChIP-seq



Single Cell-seq

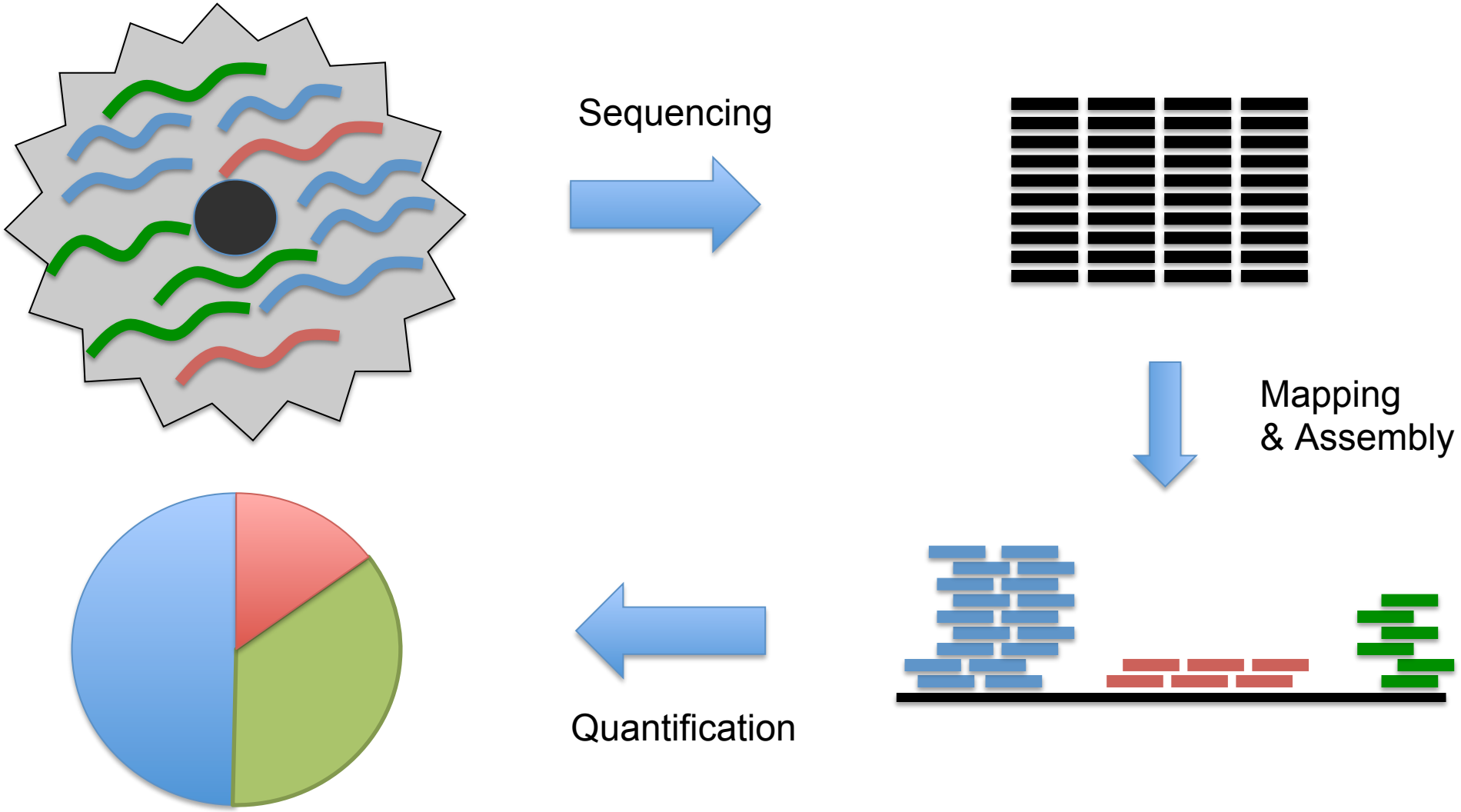


RNA-seq

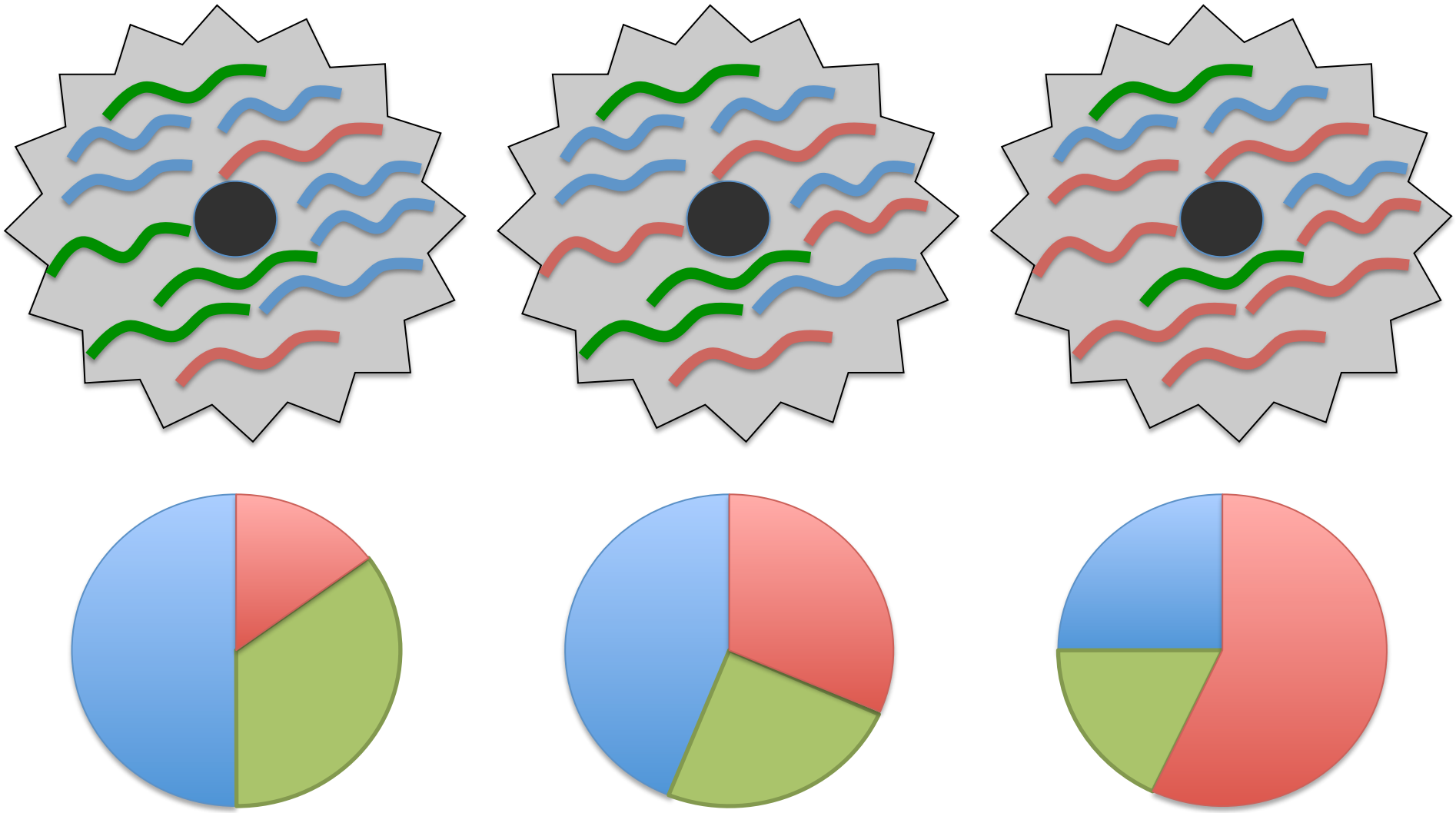


Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications.
Sørlie et al (2001) *PNAS*. 98(19):10869-74.

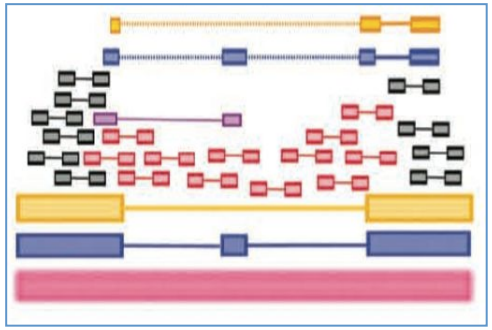
RNA-seq Overview



RNA-seq Overview



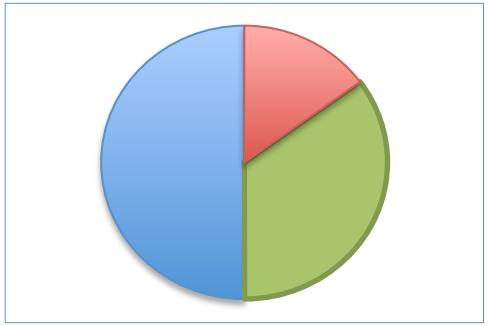
RNA-seq Challenges



Challenge 1: Eukaryotic genes are spliced

Solution: Use a spliced aligner, and assemble isoforms

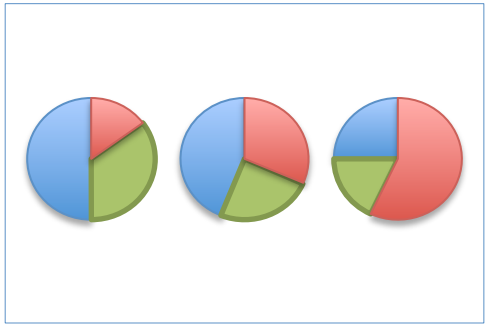
TopHat: discovering spliced junctions with RNA-Seq.
Trapnell et al (2009) *Bioinformatics*. 25:0 1105-1111



Challenge 2: Read Count != Transcript abundance

Solution: Infer underlying abundances (e.g. FPKM)

Transcript assembly and quantification by RNA-seq
Trapnell et al (2010) *Nat. Biotech.* 25(5): 511-515



Challenge 3: Transcript abundances are stochastic

Solution: Replicates, replicates, and more replicates

RNA-seq differential expression studies: more sequence or more replication?
Liu et al (2013) *Bioinformatics.* doi:10.1093/bioinformatics/btt688

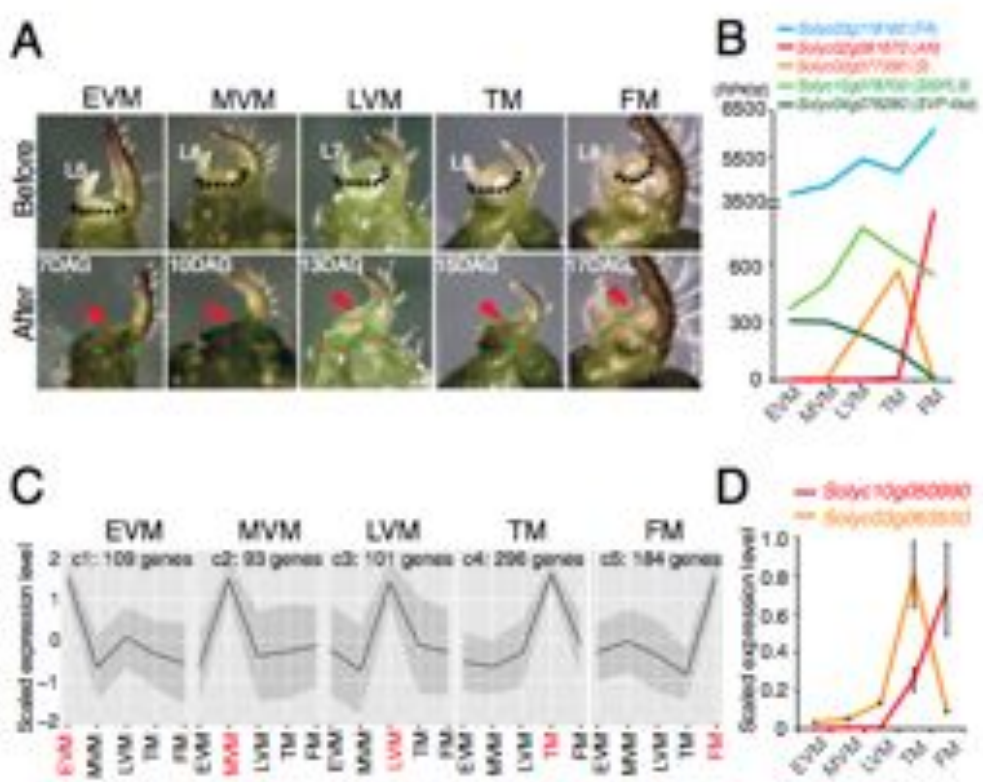
PNAS

Rate of meristem maturation determines inflorescence architecture in tomato

Soon Ju Park[1], Ke Jiang[1], Michael C. Schatz, and Zachary B. Lippman[2]

Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724

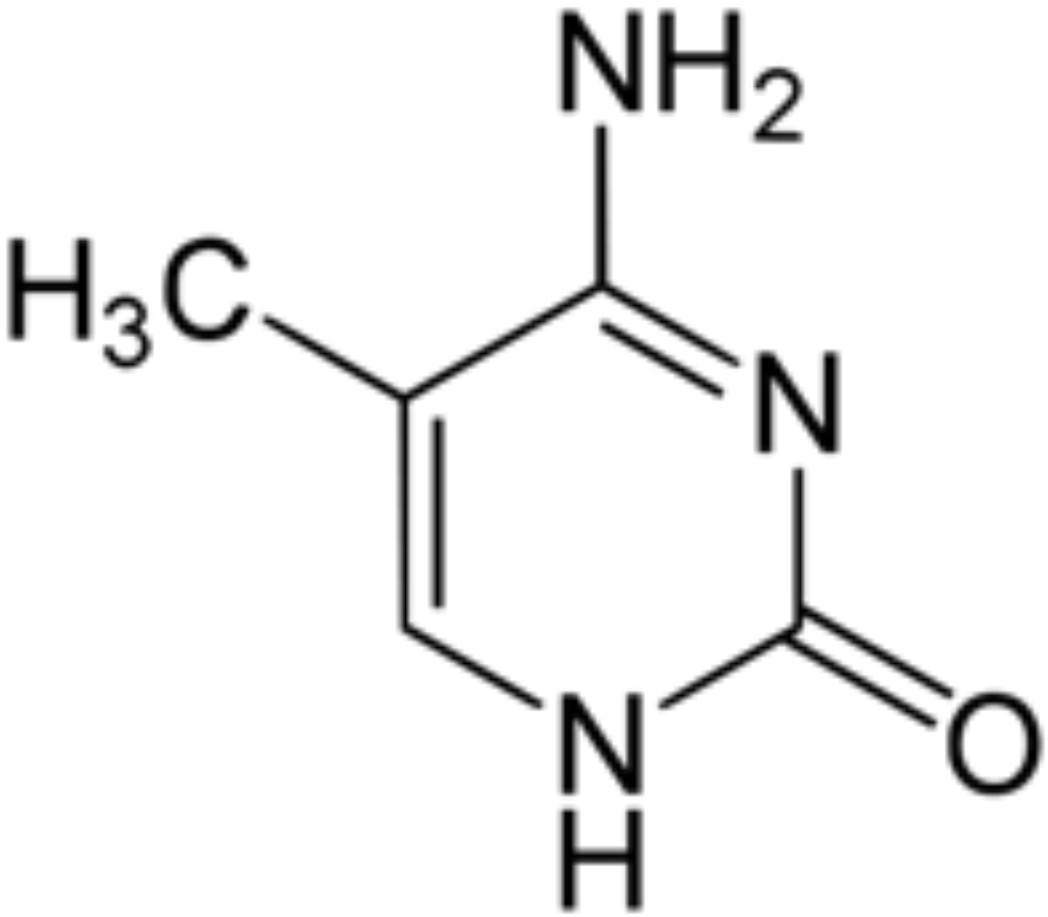
Edited by Maarten Koornneef, Wageningen University and Research Centre, Cologne, Germany, and approved November 28, 2011 (received for review September 12, 2011)



RNA-seq to determine the expression dynamics during development

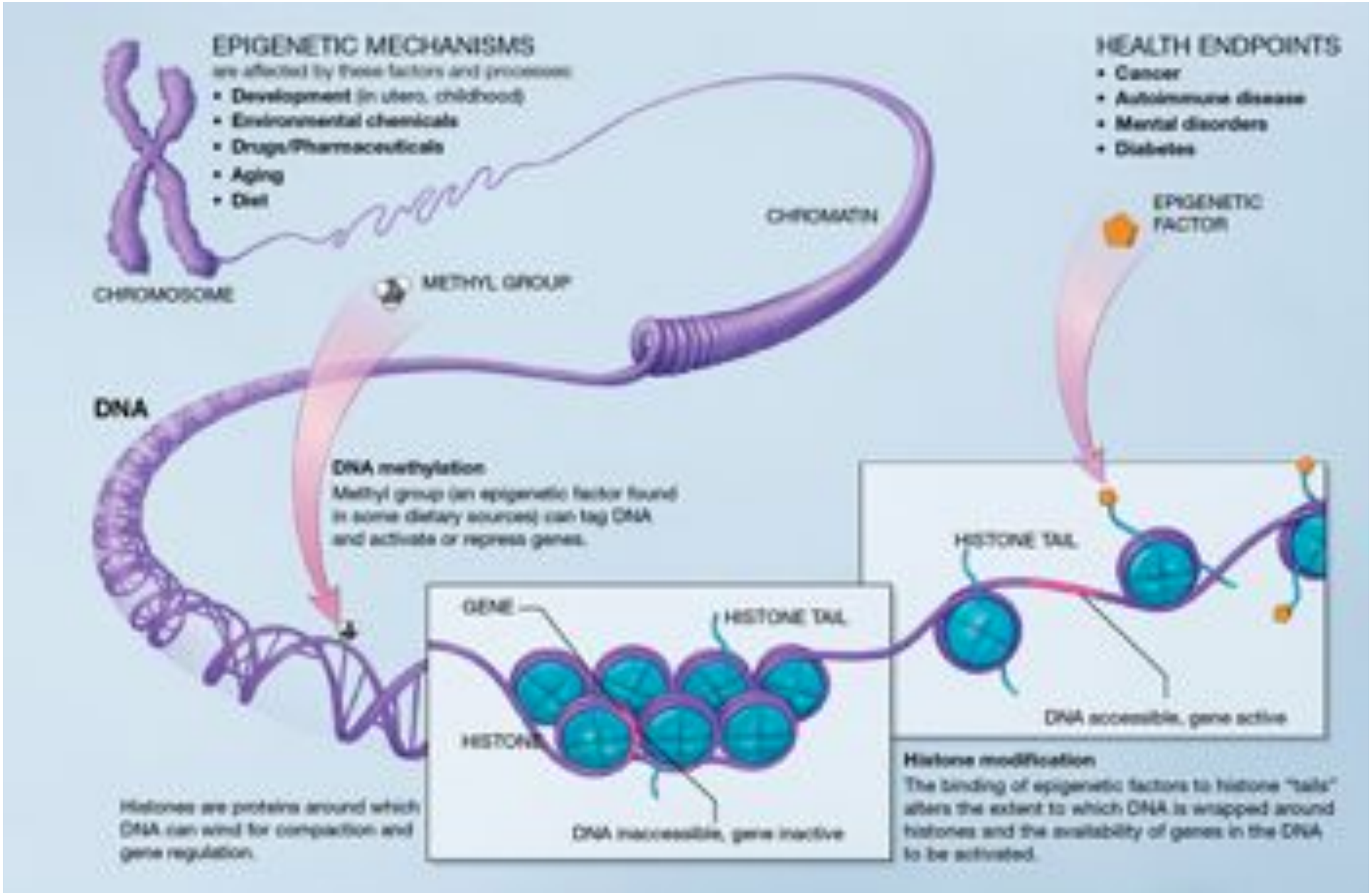
- Laser microdissection to precisely extract tissue from developing organs
- Use RNA-seq to watch different classes of genes become activated at different stages of development
- When those genes are delayed or interupted, tomato mutants take on very different branching patterns.

Methyl-seq



Finding the fifth base: Genome-wide sequencing of cytosine methylation
Lister and Ecker (2009) *Genome Research.* 19: 959-966

Methylation & Epigenetics



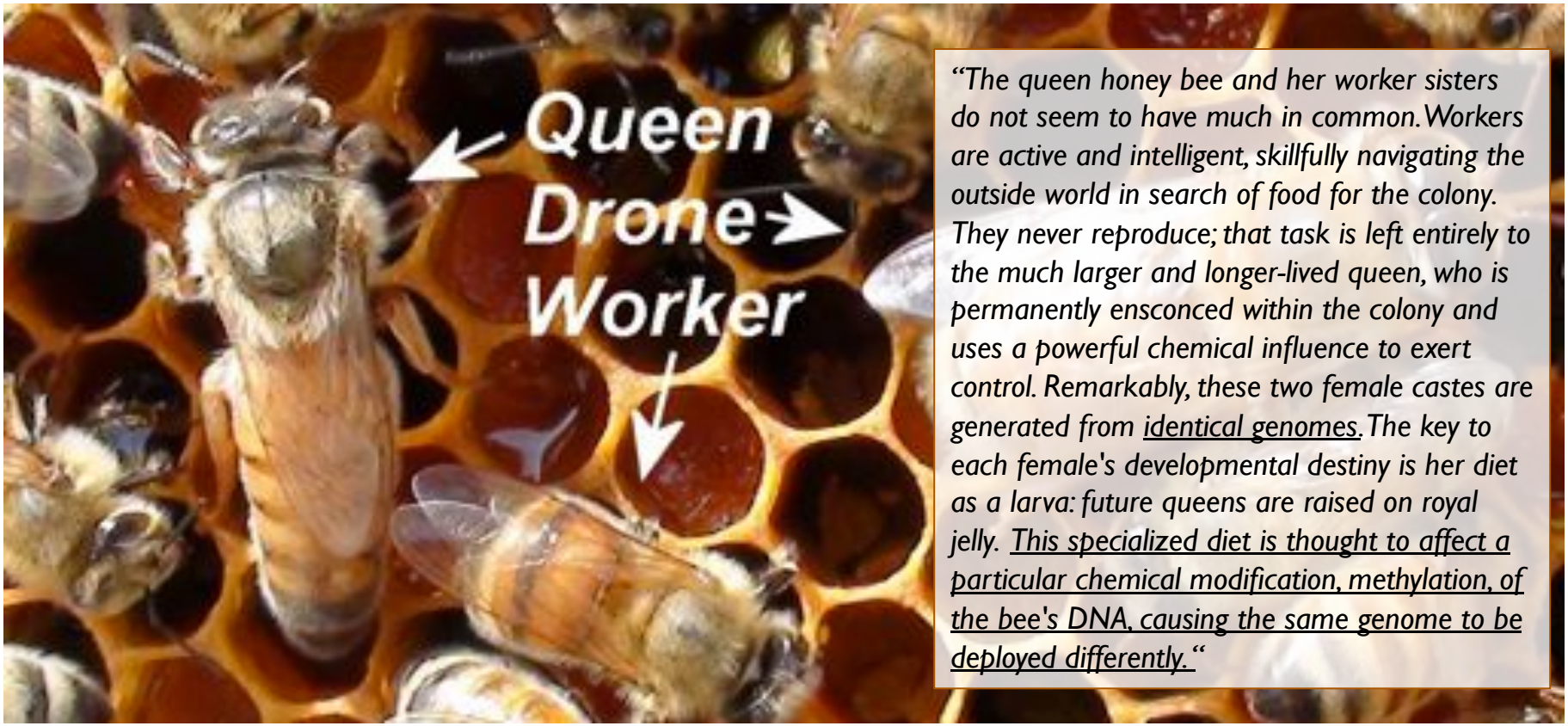
OPEN ACCESS Freely available online

PLoS BIOLOGY

The Honey Bee Epigenomes: Differential Methylation of Brain DNA in Queens and Workers

Frank Lyko[1☯], Sylvain Foret[2☯], Robert Kucharski[3], Stephan Wolf[4], Cassandra Falckenhayn[1], Ryszard Maleszka[3*]

1 Division of Epigenetics, DKFZ-ZMBH Alliance, German Cancer Research Center, Heidelberg, Germany, 2 ARC Centre of Excellence for Coral Reef Studies, James Cook University, Townsville, Australia, 3 Research School of Biology, the Australian National University, Canberra, Australia, 4 Genomics and Proteomics Core Facility, German Cancer Research Center, Heidelberg, Germany

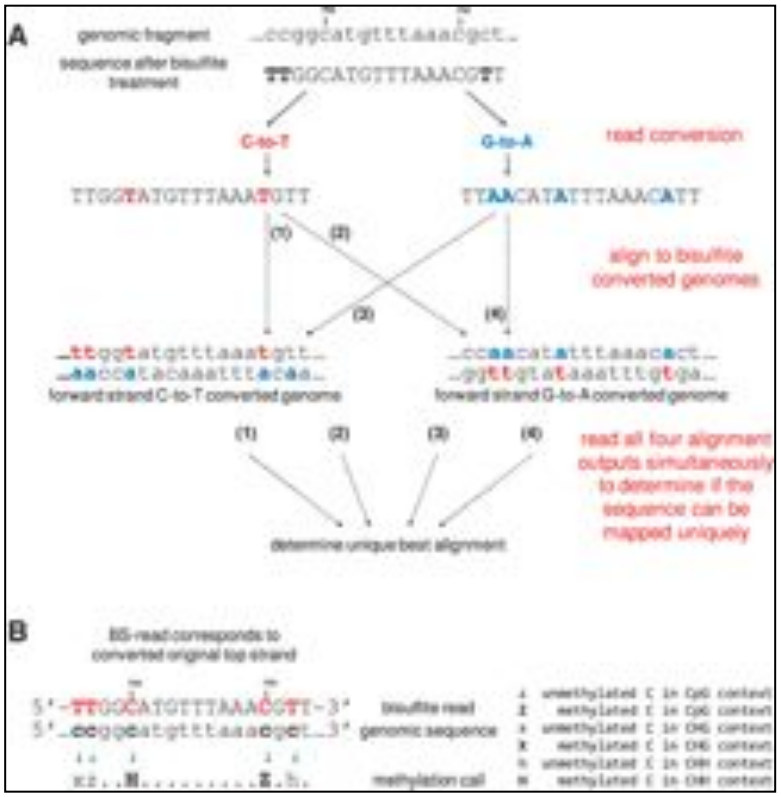


"The queen honey bee and her worker sisters do not seem to have much in common. Workers are active and intelligent, skillfully navigating the outside world in search of food for the colony. They never reproduce; that task is left entirely to the much larger and longer-lived queen, who is permanently ensconced within the colony and uses a powerful chemical influence to exert control. Remarkably, these two female castes are generated from identical genomes. The key to each female's developmental destiny is her diet as a larva: future queens are raised on royal jelly. This specialized diet is thought to affect a particular chemical modification, methylation, of the bee's DNA, causing the same genome to be deployed differently."

Bisulfite Conversion

Treating DNA with sodium bisulfite will convert unmethylated C to T

- 5-MethyC will be protected and not change, so can look for differences when mapping
- Requires great care when analyzing reads, since the complementary strand will also be converted (G to A)
- Typically analyzed by mapping to a "reduced alphabet" where we assume all Cs are converted to Ts once on the forward strand and once on the reverse



Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications
Krueger and Andrews (2010) *Bioinformatics*. 27 (11): 1571-1572.

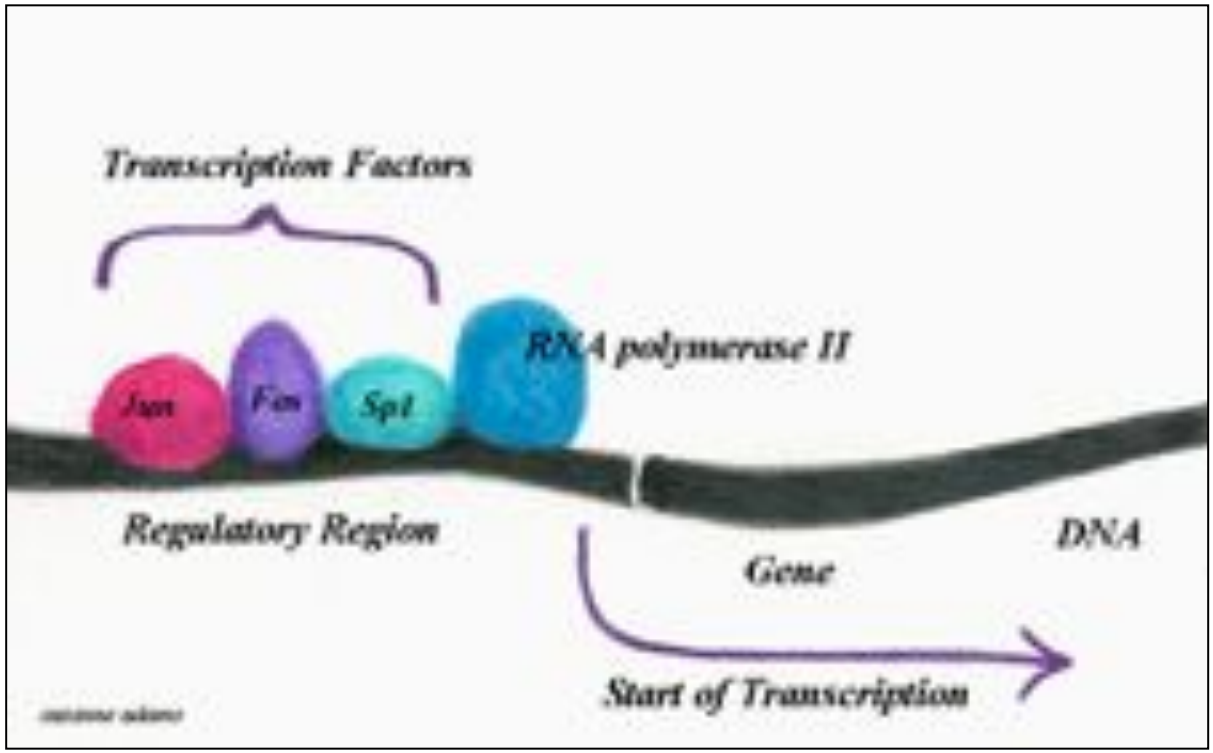
Bisulfite Conversion

Tr
w

-
-
-

Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications
Krueger and Andrews (2010) *Bioinformatics*. 27 (11): 1571-1572.

ChIP-seq

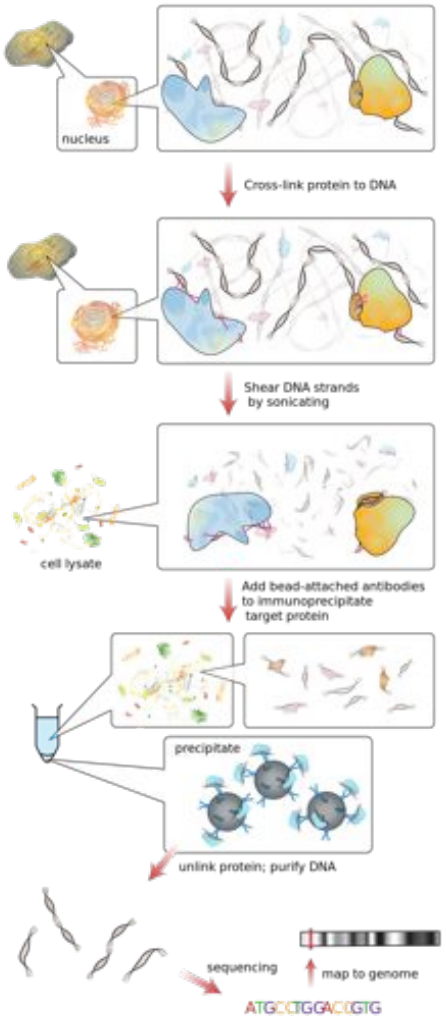


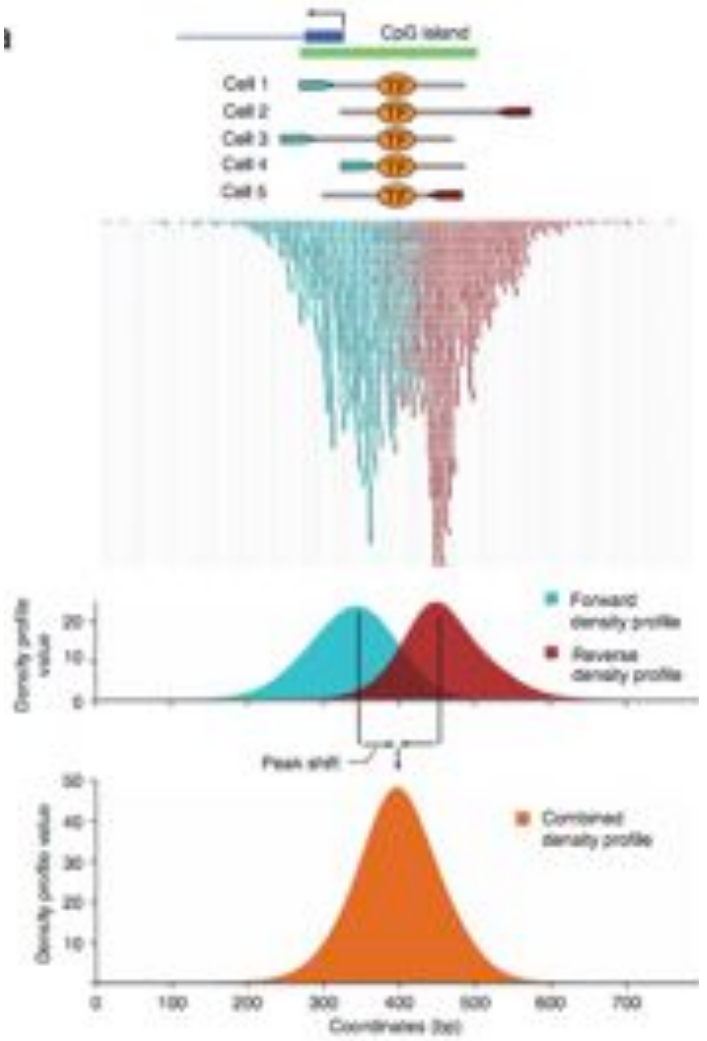
Genome-wide mapping of in vivo protein-DNA interactions.
Johnson *et al* (2007) *Science*. 316(5830):1497-502

ChIP-seq

Goals:

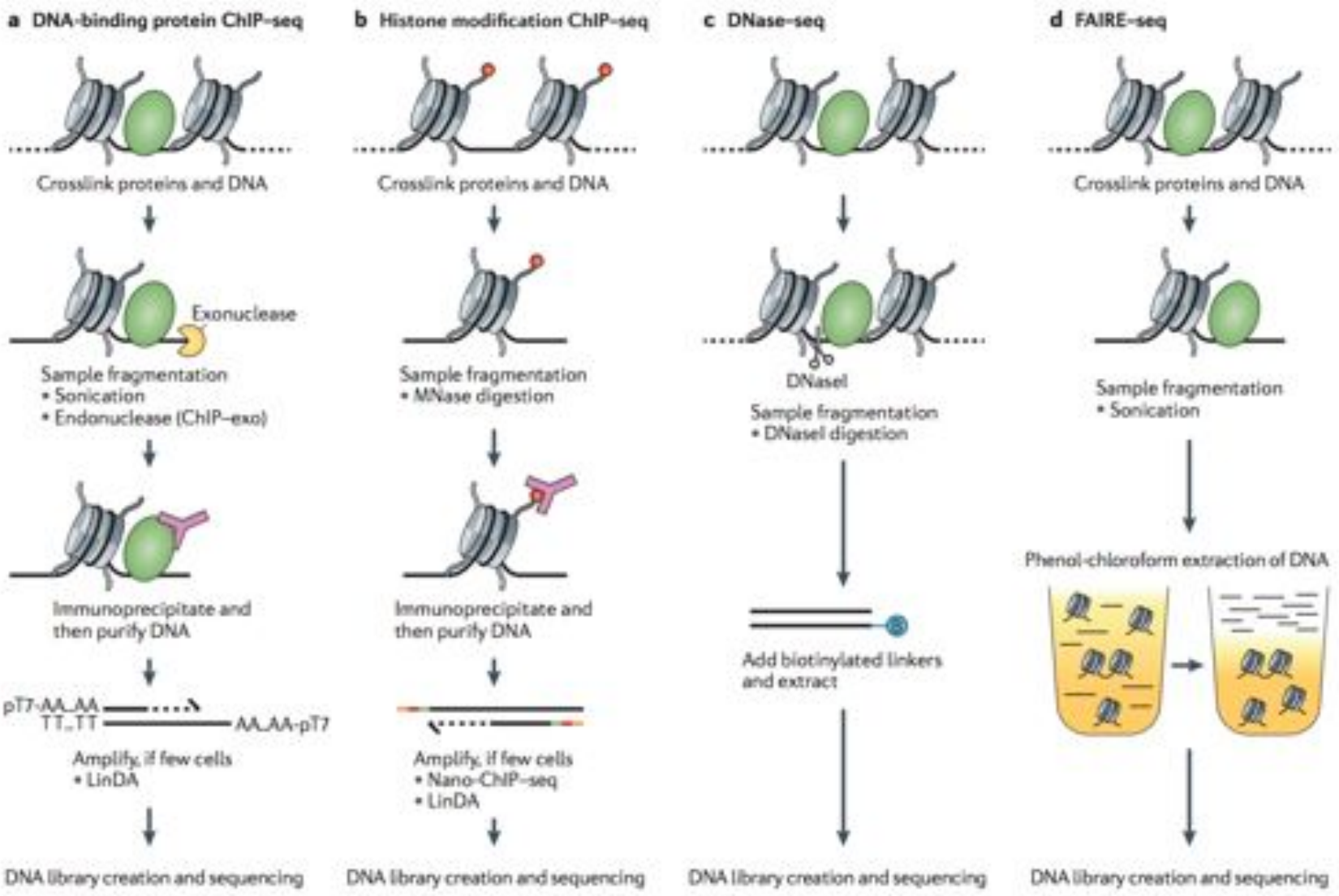
- Where are transcription factors and other proteins binding to the DNA?
- How strongly are they binding?
- Do the protein binding patterns change over developmental stages or when the cells are stressed?





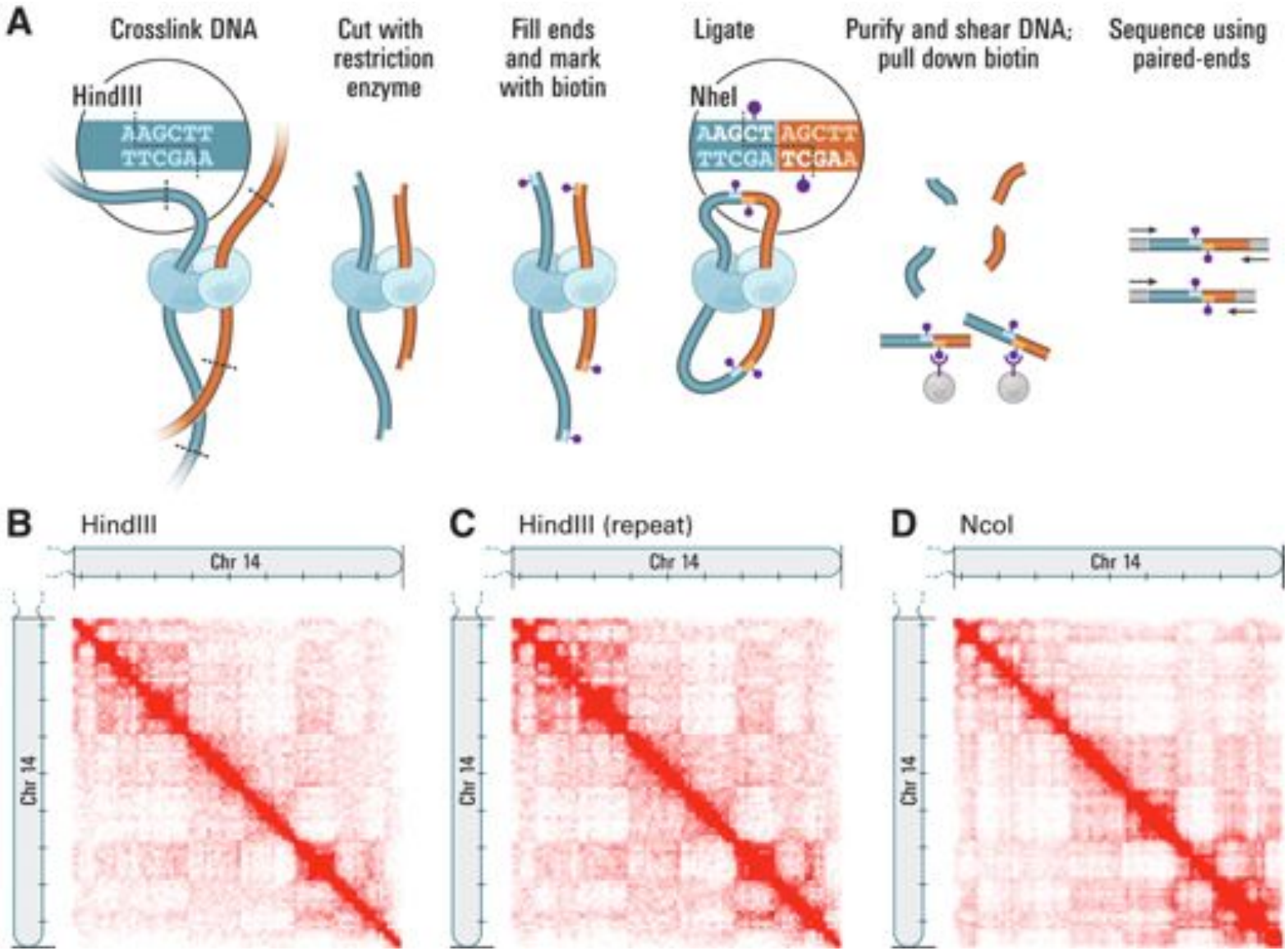
Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data
Valouev et al (2008) *Nature Methods*. 5, 829 - 834

Related Assays



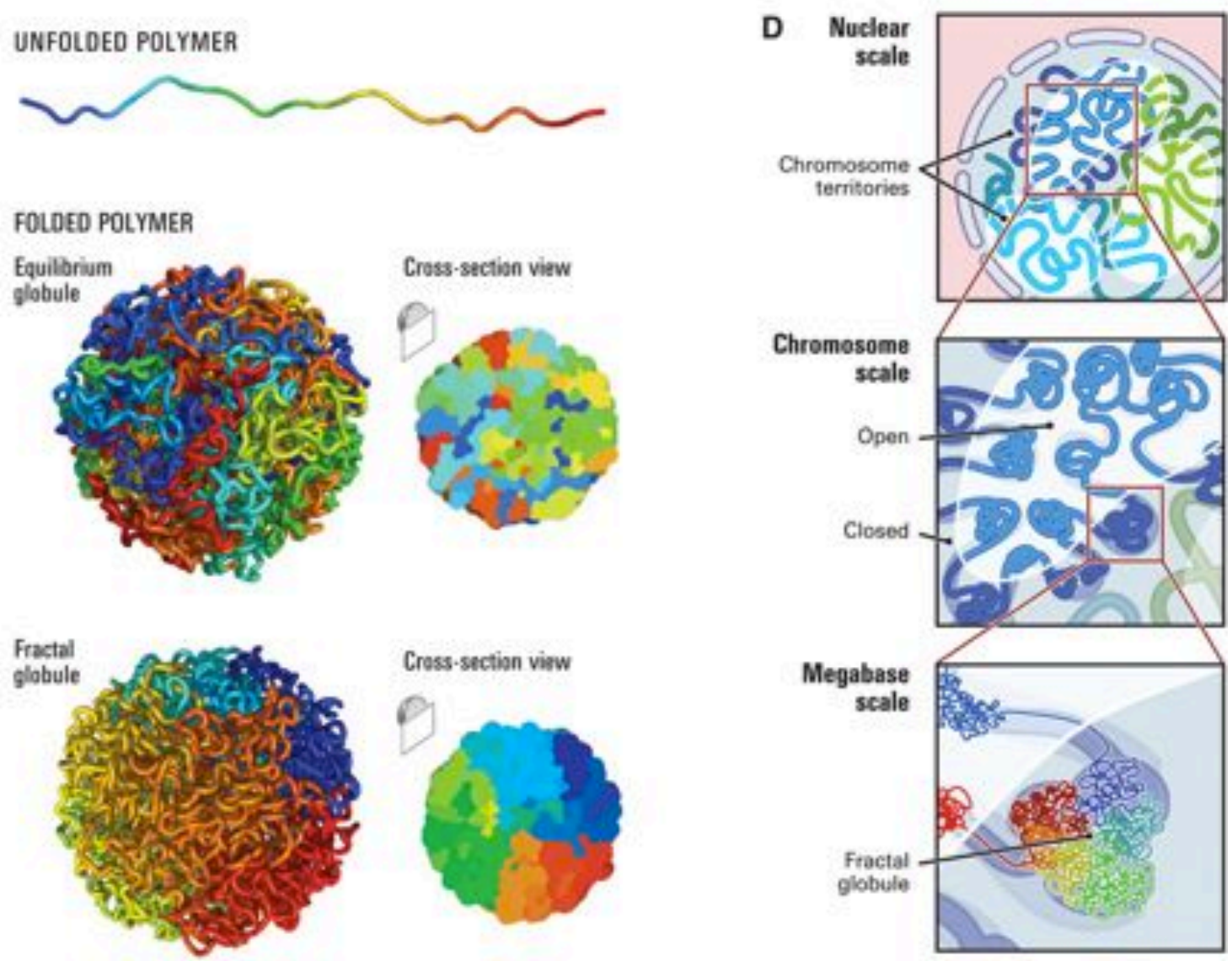
ChIP–seq and beyond: new and improved methodologies to detect and characterize protein–DNA interactions
Furey (2012) *Nature Reviews Genetics*. 13, 840-852

HI-C: Mapping the folding of DNA



Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome
Liberman-Aiden et al. (2009) *Science*. 326 (5950): 289-293

HI-C: Mapping the folding of DNA



Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome
Liberman-Aiden et al. (2009) *Science*. 326 (5950): 289-293

Gene Regulation in 3-dimensions

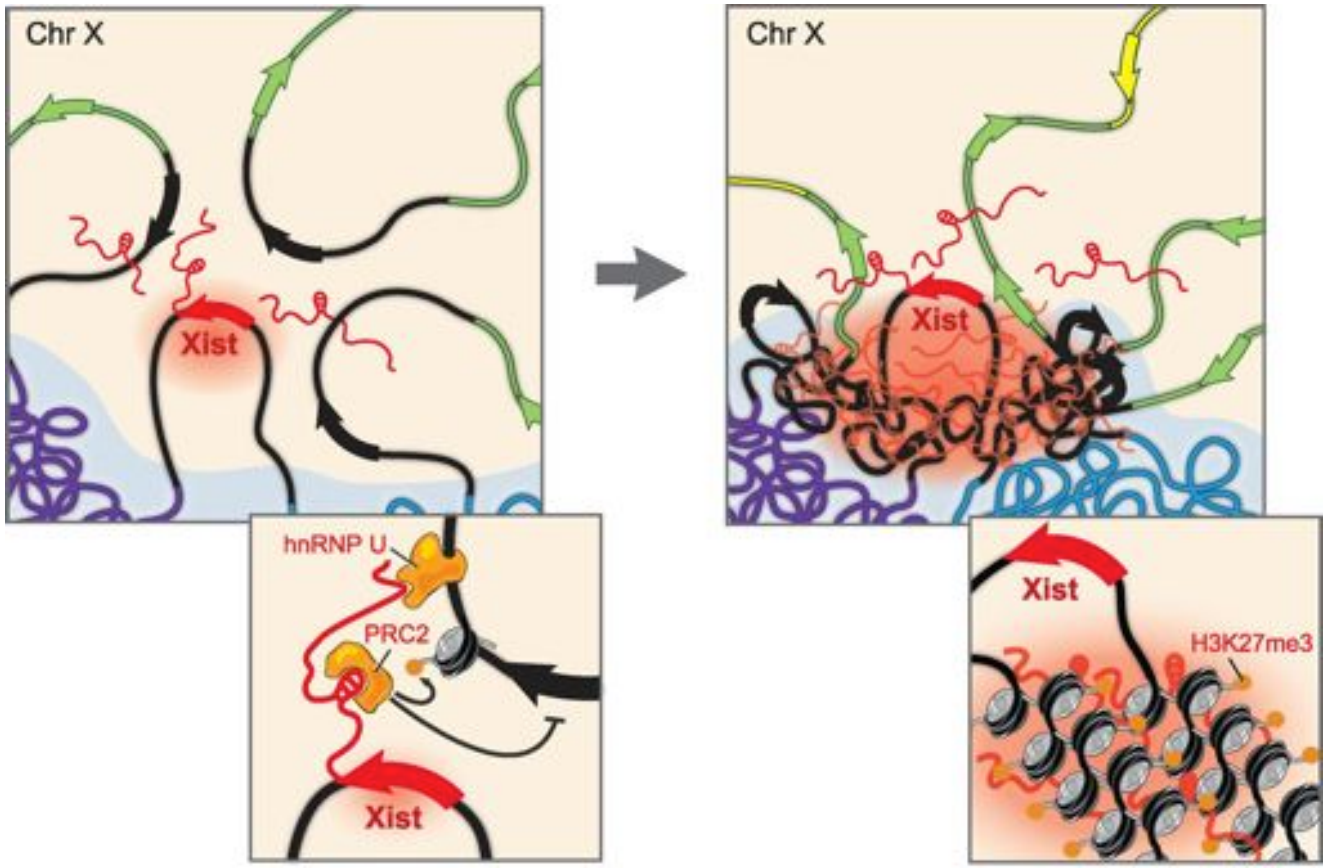
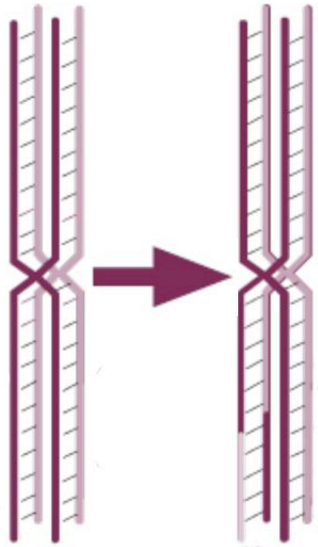
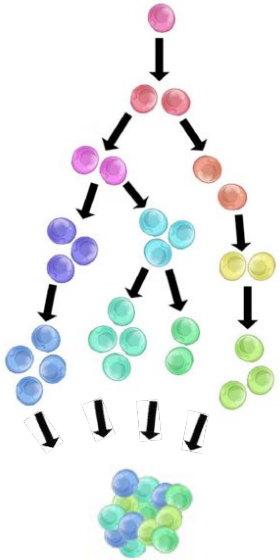
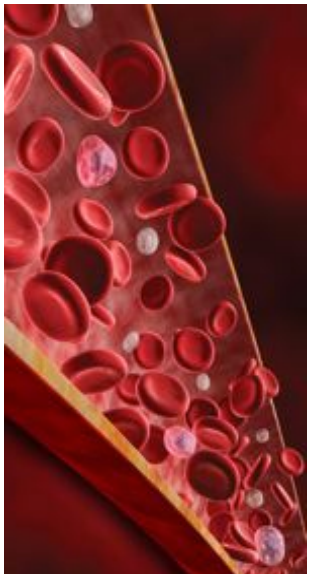


Fig 6. A model for how Xist exploits and alters three-dimensional genome architecture to spread across the X chromosome.

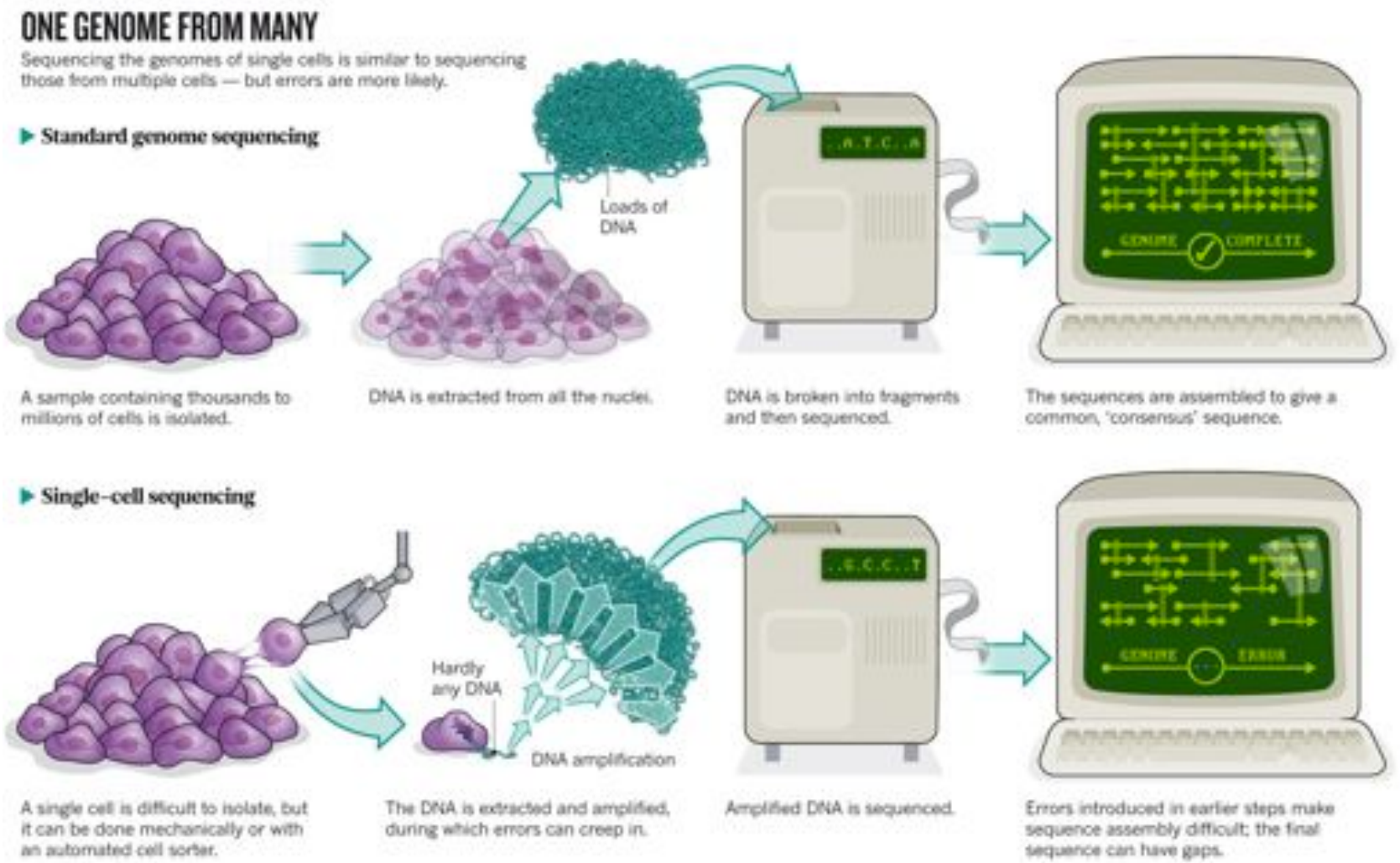
The Xist lncRNA Exploits Three-Dimensional Genome Architecture to Spread Across the X Chromosome
Engreitz et al. (2013) *Science*. 341 (6147)

Single Cell Sequencing

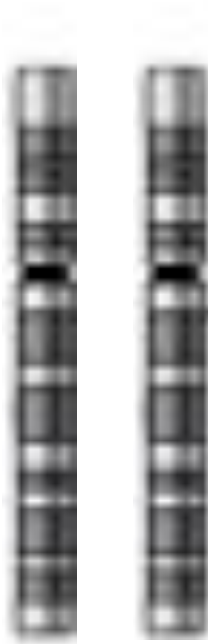
Germ Cells	Heterogeneous Tumors	Heterogeneous Tissues: blood and lymph
		
Recombination & crossover events	Clonal expansion	Isolating unique cell types

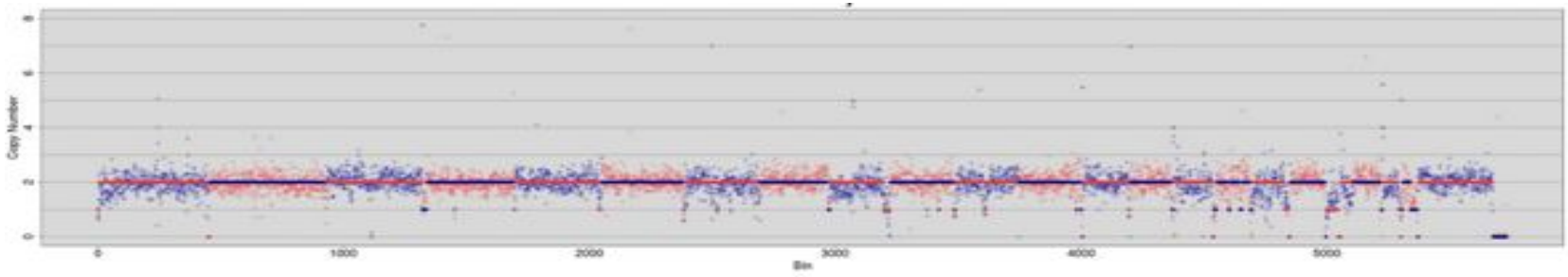
Cancer genomics: one cell at a time
Navin *et al* (2014) *Genome Biology.* 15:452

Bulk vs Single Cell

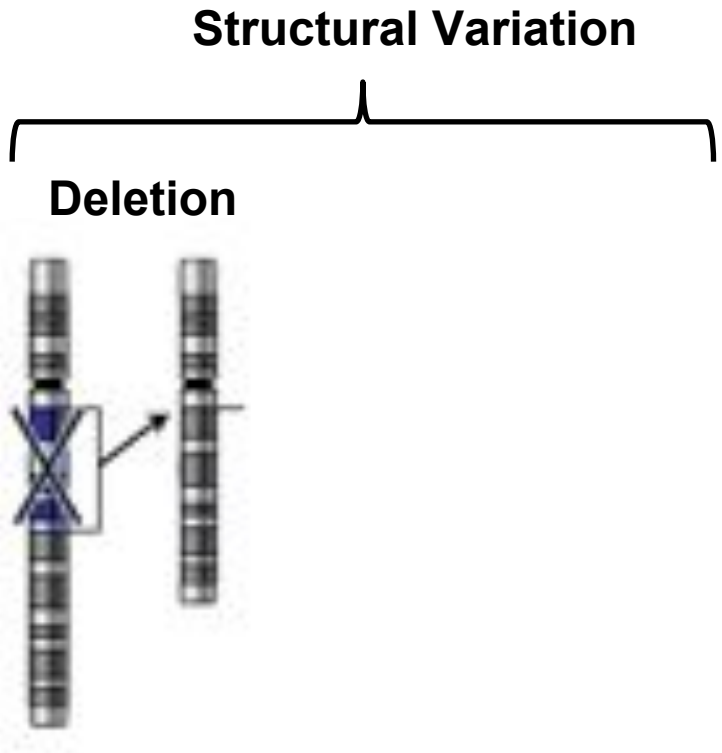


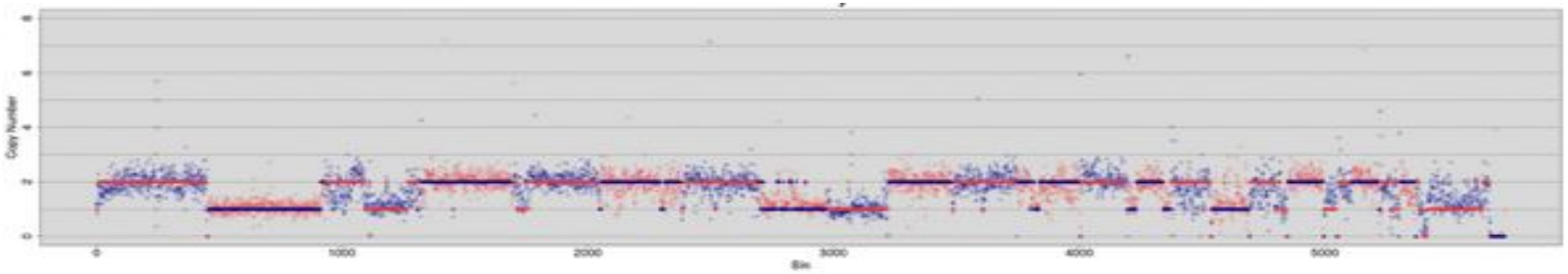
Copy Number Variants





Copy Number Variants

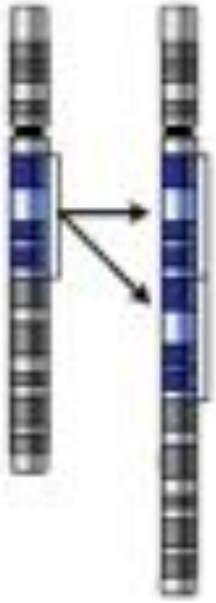


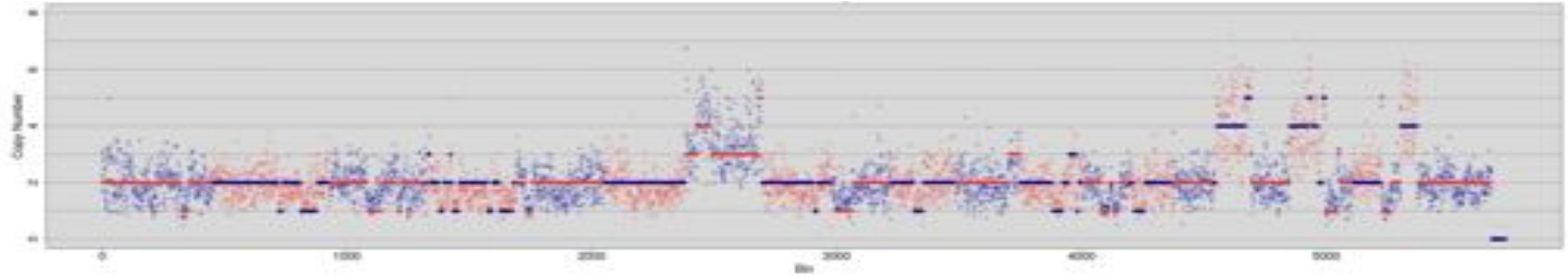


Copy Number Variants

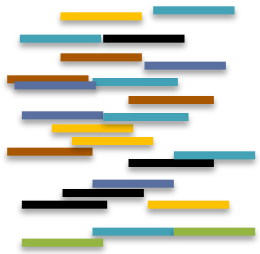
Structural Variation

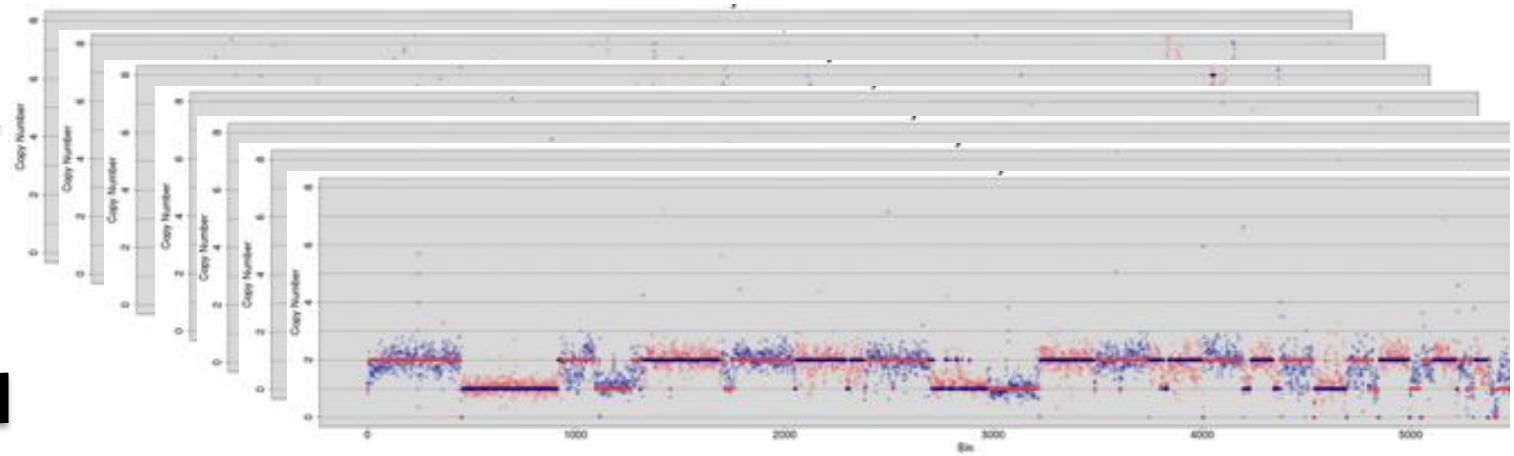
Amplification

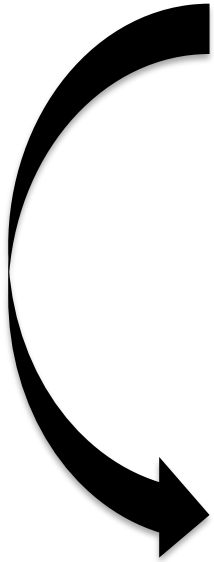


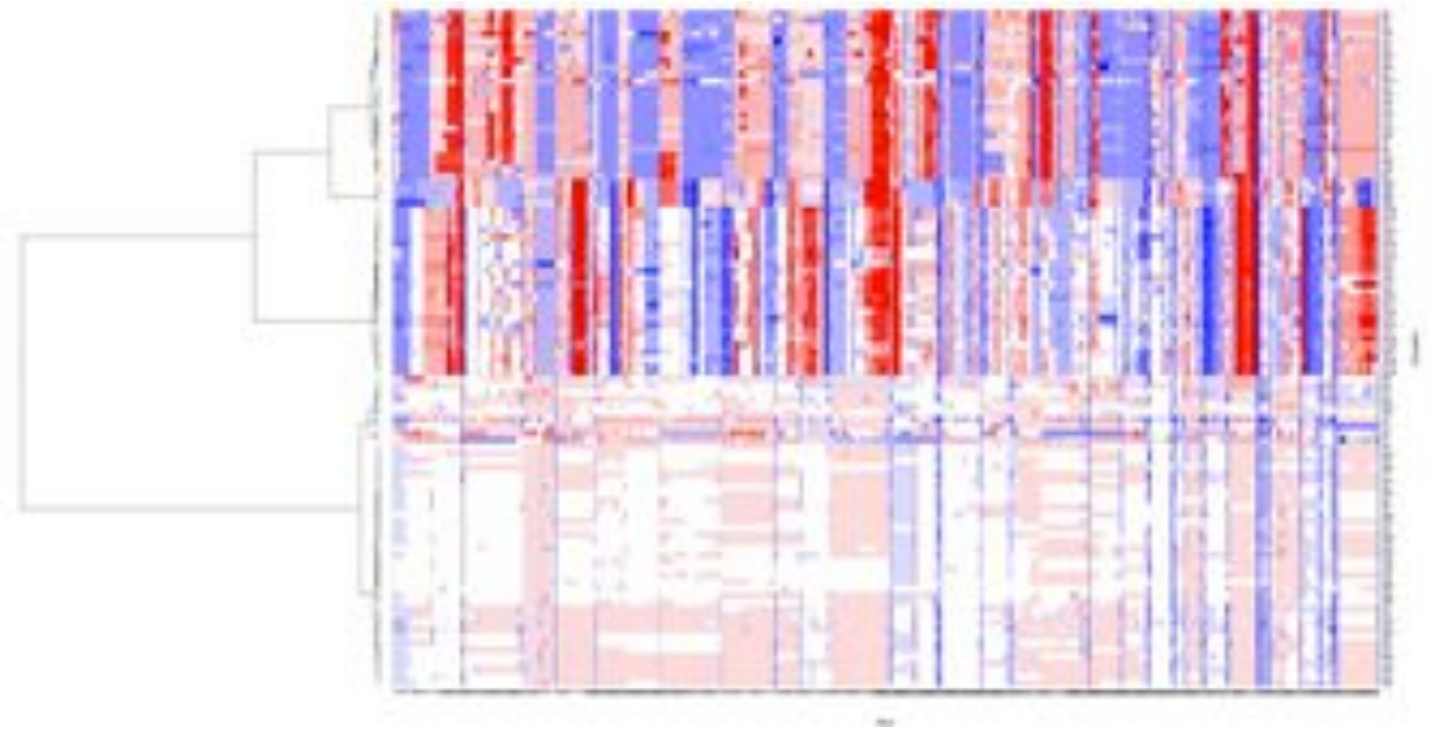


CNV Analysis Overview







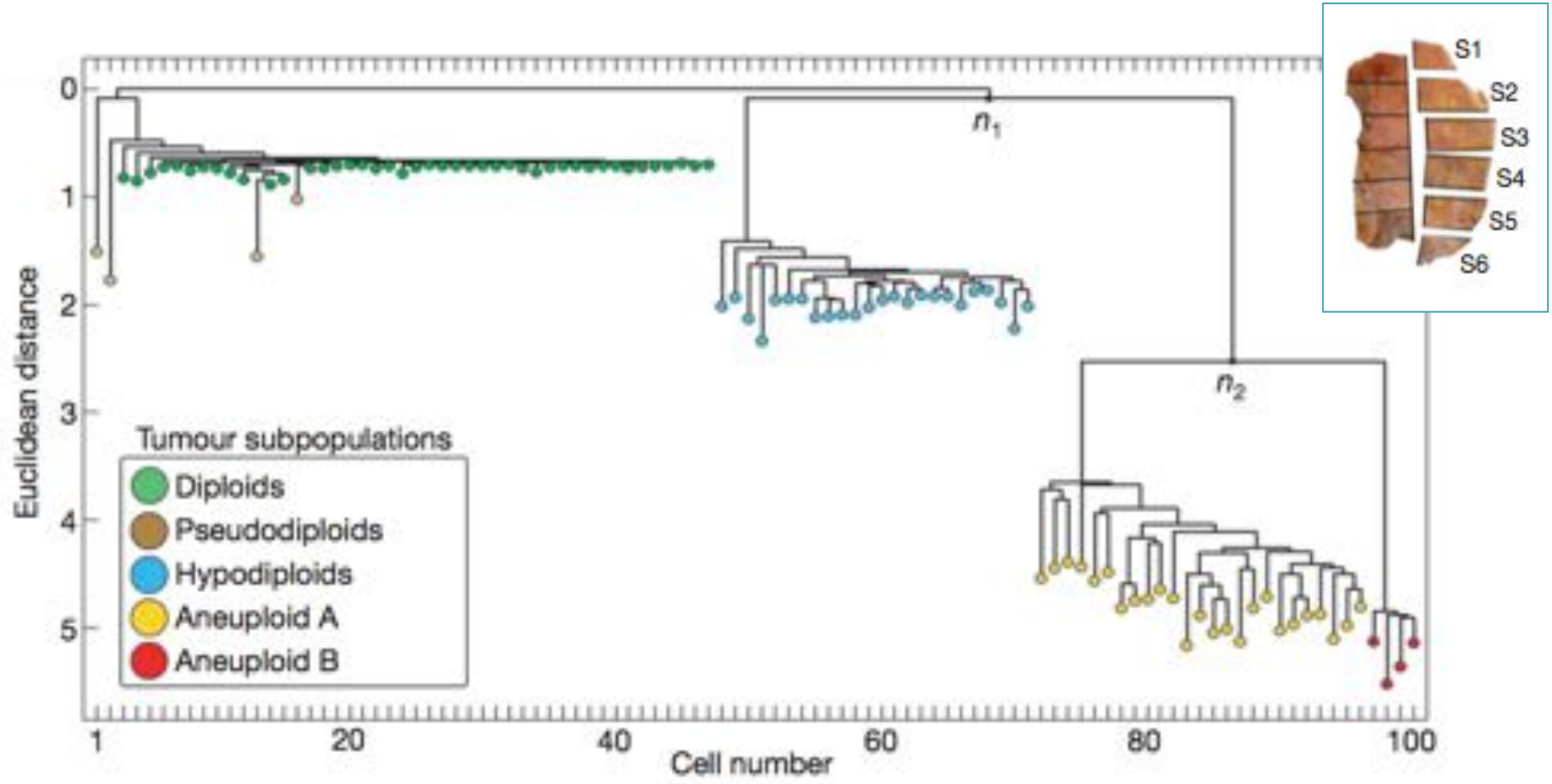


LETTER

doi:10.1038/nature09807

Tumour evolution inferred by single-cell sequencing

Nicholas Navin[1,2], Jude Kendall[1], Jennifer Troge[1], Peter Andrews[1], Linda Rodgers[1], Jeanne McIndoo[1], Kerry Cook[1], Asya Stepansky[1], Dan Levy[1], Diane Esposito[1], Lakshmi Muthuswamy[3], Alex Krasnitz[1], W. Richard McCombie[1], James Hicks[1] & Michael Wigler[1]



Other Examples: CNV + RNA

PNAS

Reproducible copy number variation patterns among single circulating tumor cells of lung cancer patients

Xiaohui Ni[a,b,1], Minglei Zhuo[c,1], Zhe Su[a,1], Jianchun Duan[c,1], Yan Gao[a,1], Zhijie Wang[c,1], Chenghang Zong[b,1,2], Hua Bai[c], Alec R. Chapman[b,d], Jun Zhao[c], Liya Xu[a], Tongtong An[c], Qi Ma[a], Yuyan Wang[c], Meina Wu[c], Yu Sun[e], Shuhang Wang[c], Zhenxiang Li[c], Xiaodan Yang[c], Jun Yong[b], Xiao-Dong Su[a], Youyong Lu[f], Fan Bai[a,3], X. Sunney Xie[a,b,3], and Jie Wang[c,3]

Science
AAAS

Probing Meiotic Recombination and Aneuploidy of Single Sperm Cells by Whole-Genome Sequencing

Sijia Lu *et al.*

Science **338**, 1627 (2012);
DOI: 10.1126/science.1229112

Science
AAAS

Mosaic Copy Number Variation in Human Neurons

Michael J. McConnell *et al.*

Science **342**, 632 (2013);
DOI: 10.1126/science.1243472

Cell

Genome Analyses of Single Human Oocytes

Yu Hou,[1,6] Wei Fan,[1,4,6] Liying Yan,[1,6] Rong Li,[1] Ying Lian,[1] Jin Huang,[1] Jinsen Li,[1] Liya Xu,[1] Fuchou Tang,[1,5,*] X. Sunney Xie,[1,2,*] and Jie Qiao[1,3,*]

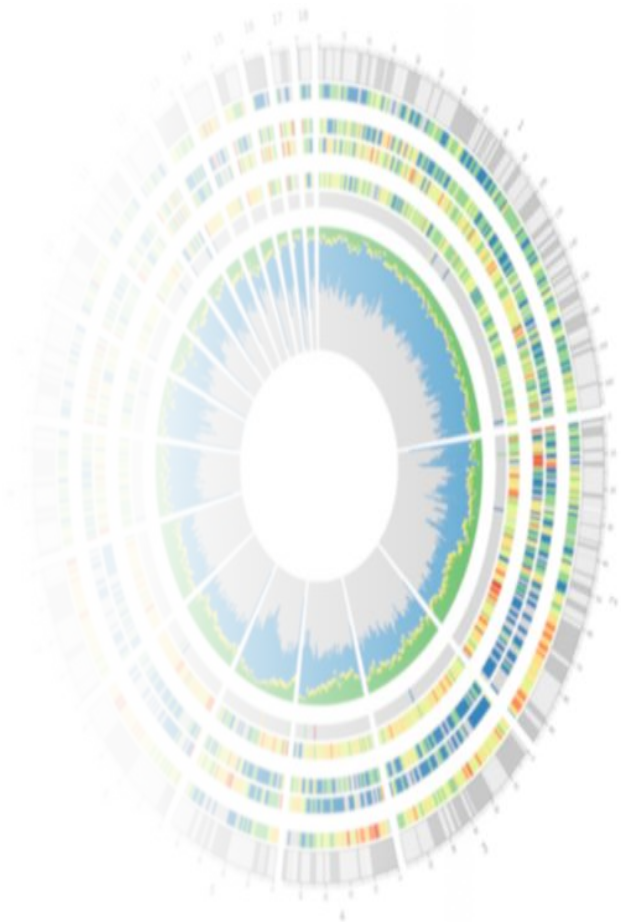
nature biotechnology

The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells

Cole Trapnell[1,2,6], Davide Cacchiarelli[1–3,6], Jonna Grimsby[2], Prapti Pokharel[2], Shuqiang Li[4], Michael Morse[1,2], Niall J Lennon[2], Kenneth J Livak[4], Tarjei S Mikkelsen[1–3] & John L Rinn[1,2,5]

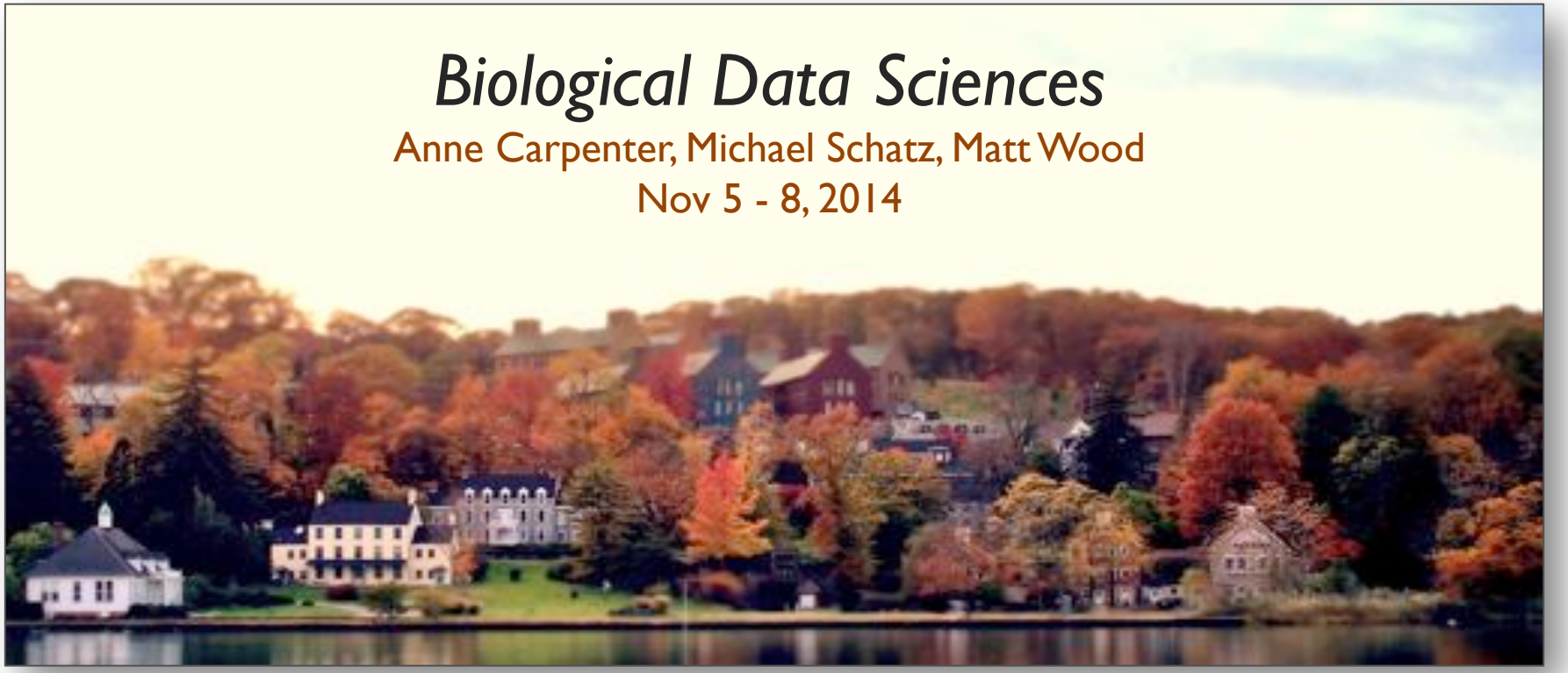
OMICS Summary

- DNA sequencing is extremely powerful and widespread to genotype large populations
  - The types of questions we ask have fundamentally changed over the last 10 years
  - Expect millions of human genomes over your PhD



- DNA sequencing is used for much more than sequencing DNA!
  - Flexible technology to observe the dynamics inside cells
  - Count the frequency of different molecules
  - See the "shadow" of chemical modification
  - See the "shadow" of molecules binding

- Coming up
  - Human Medical Genetics (Lyon)
  - Expression analysis (Gillis)
  - Genetics of modern and ancient humans (Schatz)
  - Group Discussion on ENCODE



Thank you!

http://schatzlab.cshl.edu
@mike_schatz