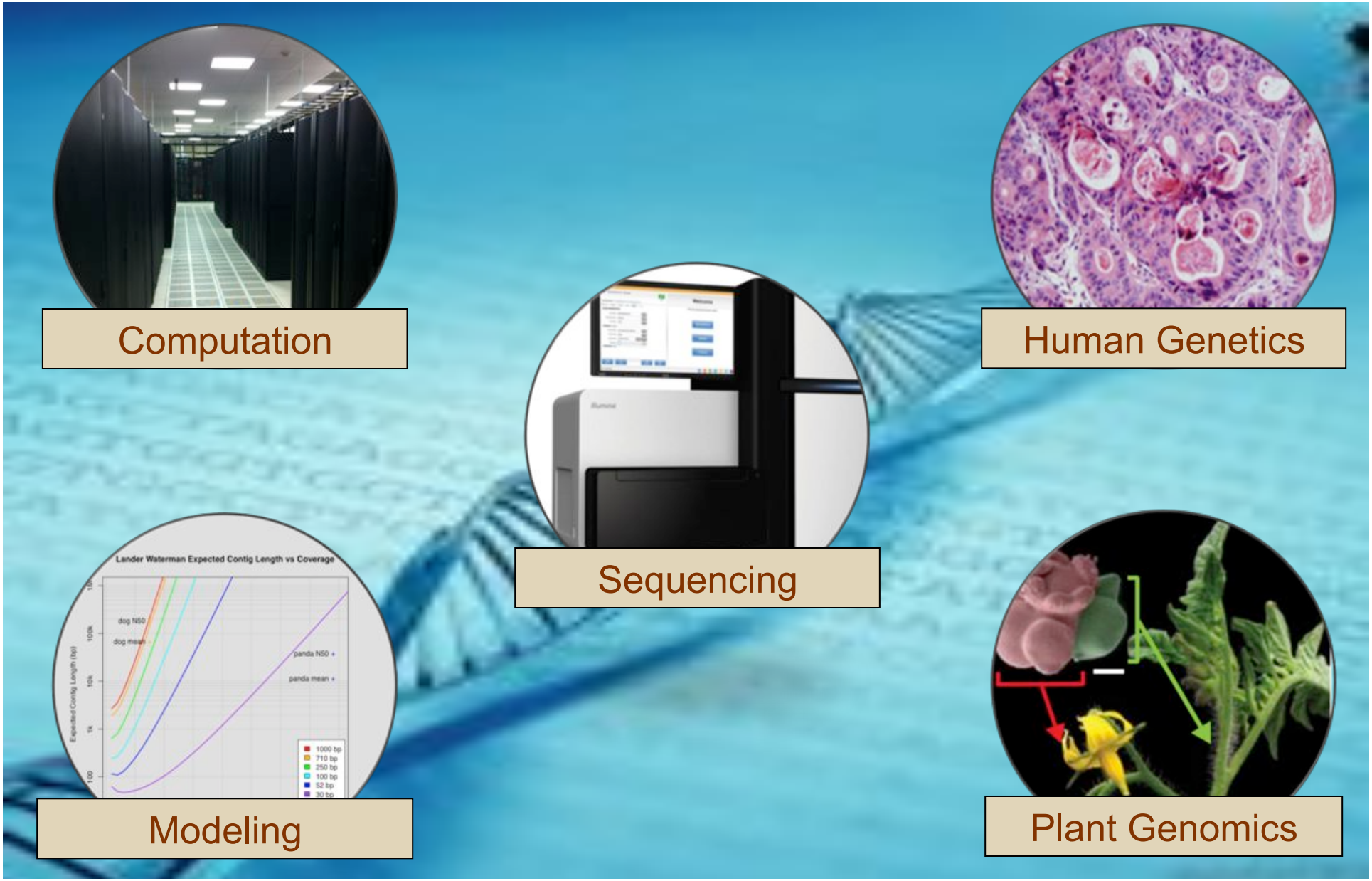
Next-gen sequence analysis

Michael Schatz

Introduction to Computational Biology
Oct 24, 2013

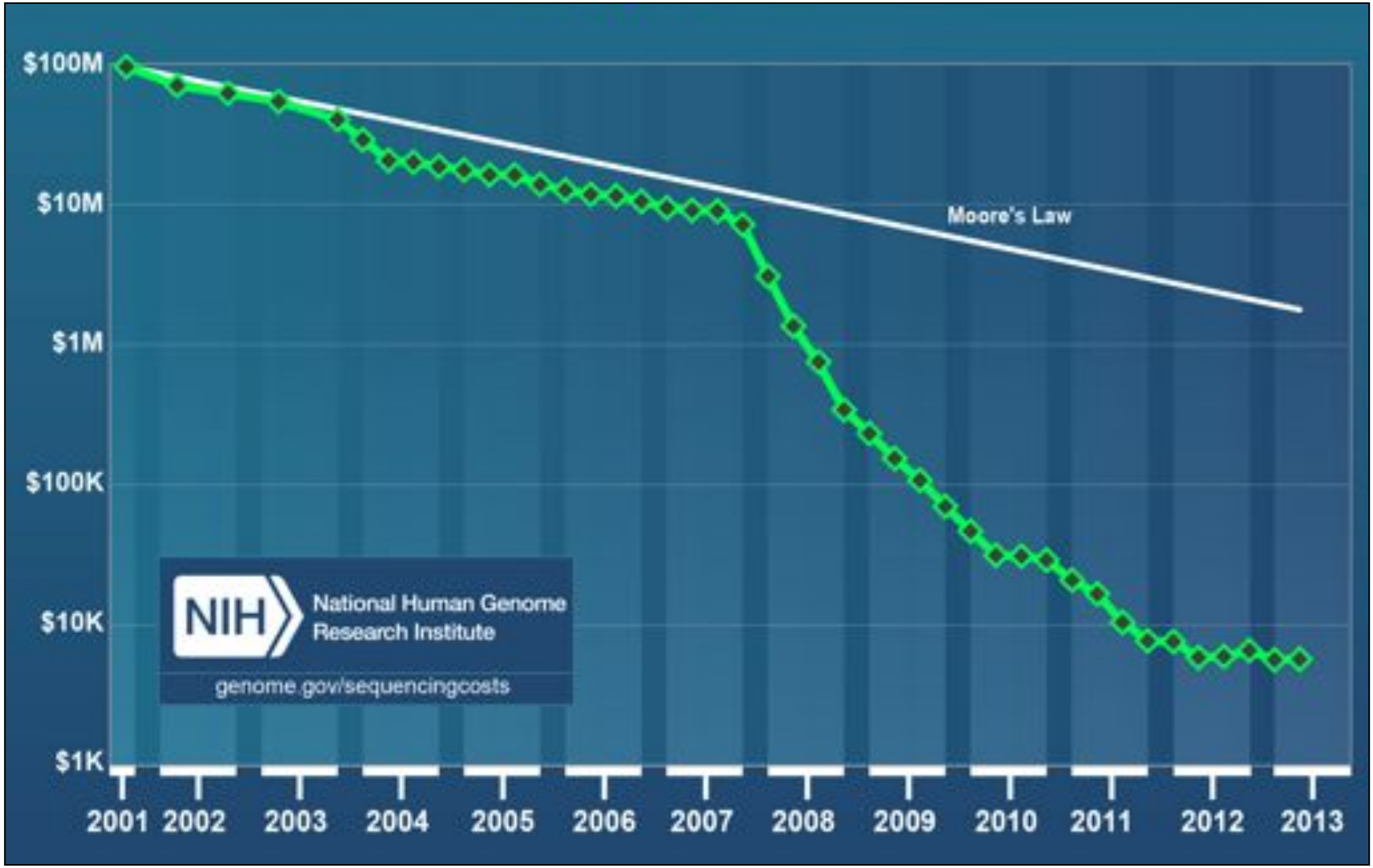
Schatz Lab Overview



Outline

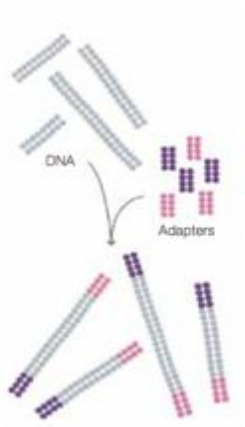
1. Rise of DNA Sequencing
2. Alignment and the BWT
3. Genetics of Autism

Cost per Genome

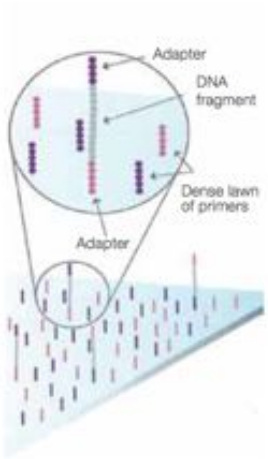


http://www.genome.gov/sequencingcosts/

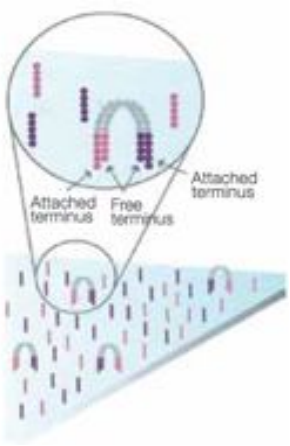
Illumina Sequencing by Synthesis



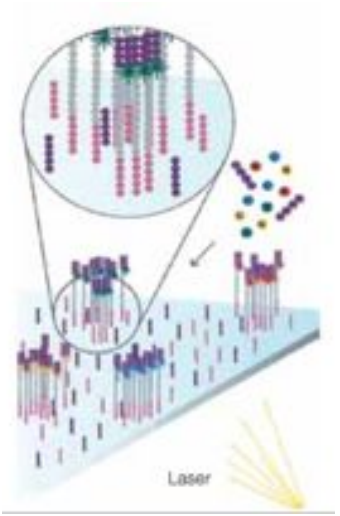
1. Prepare



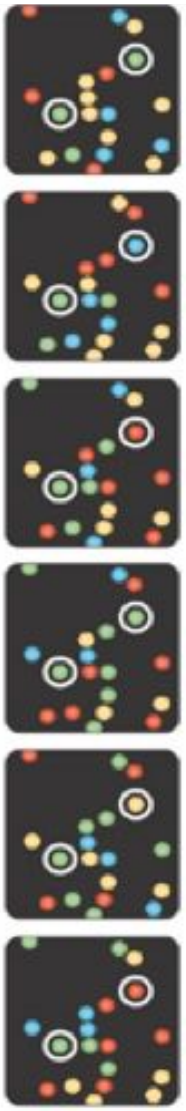
2. Attach



3. Amplify



4. Image



5. Basecall

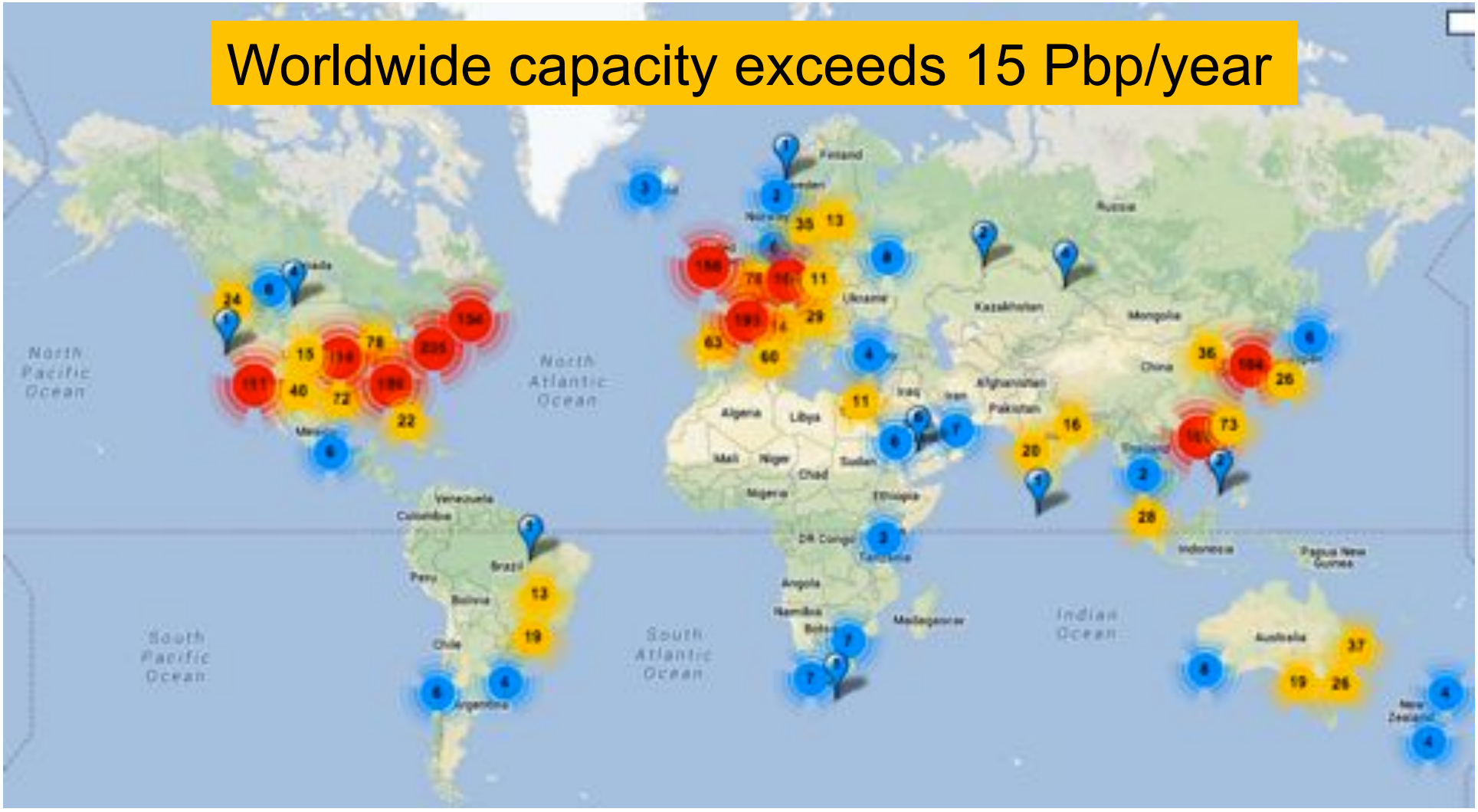
Metzker (2010) Nature Reviews Genetics 11:31-46

http://www.youtube.com/watch?v=l99aKKHcxC4

Inside the NY Genome Center

Sequencing Capacity: 16 HiSeq 2500 @ 600 Gbp / 11 day = 872 Gbp / day

Sequencing Centers



Next Generation Genomics: World Map of High-throughput Sequencers
http://omicsmaps.com

Milestones in Molecular Biology

There is tremendous interest to sequence:

- What is your genome sequence?
- How does your genome compare to my genome?

- Where are the genes and how active are they?
- How does gene activity change during development?
- How does splicing change during development?

- How does methylation change during development?
- How does chromatin change during development?
- How does is your genome folded in the cell?
- Where do proteins bind and regulate genes?

- What virus and microbes are living inside you?
- How has the disease mutated your genome?
- What drugs should we give you?

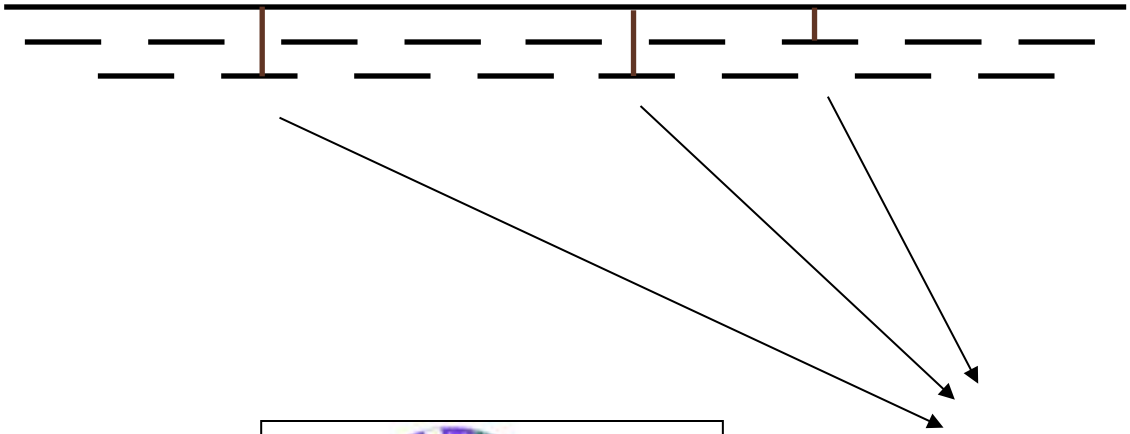
- …

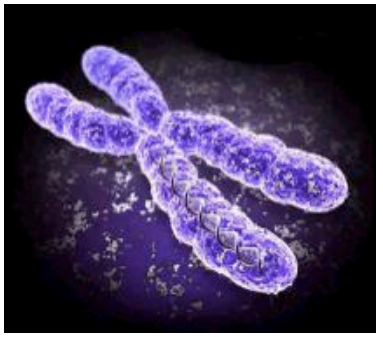
Outline

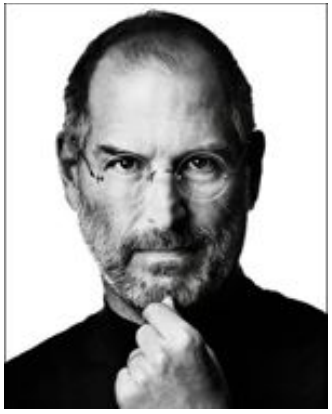
1. Rise of DNA Sequencing
2. Alignment and the BWT
3. Genetics of Autism

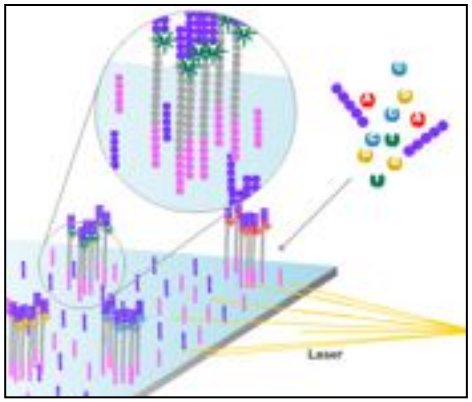
Personal Genomics

How does your genome compare to the reference?









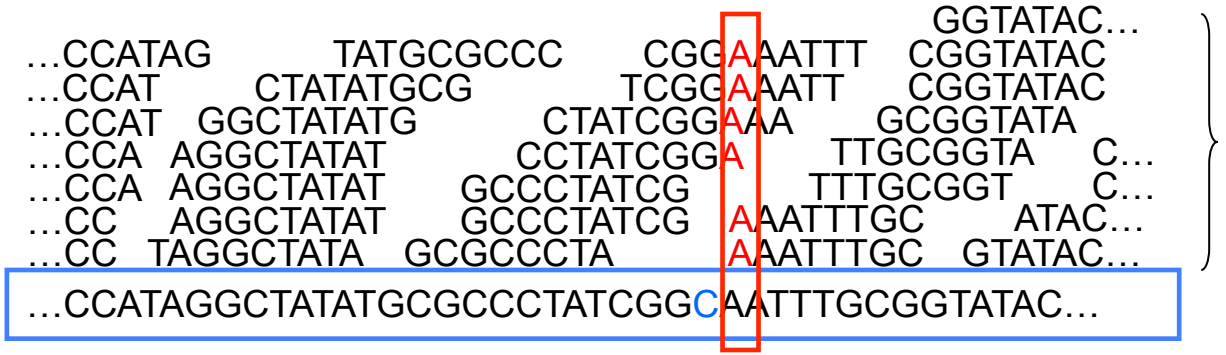
Heart Disease

Cancer

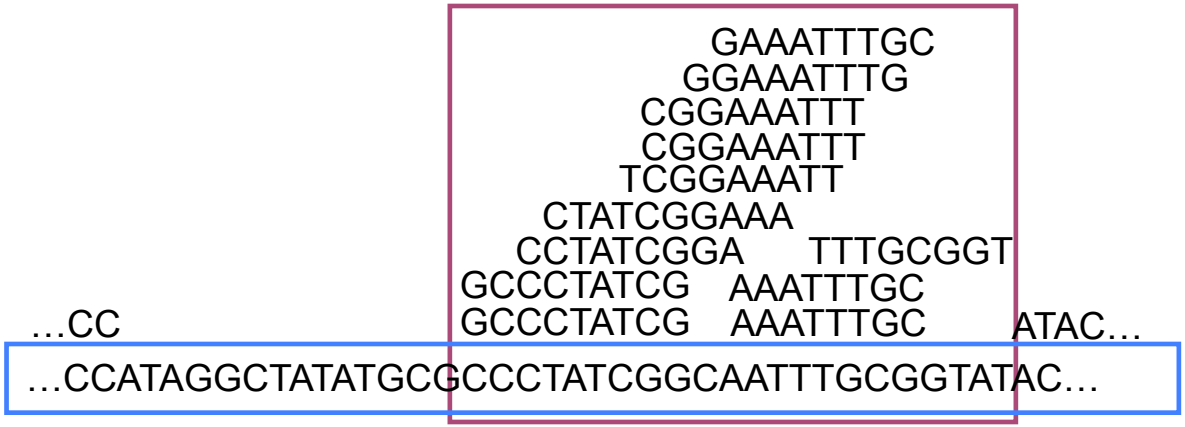
Creates magical technology

Short Read Applications

- Genotyping: Identify Variations

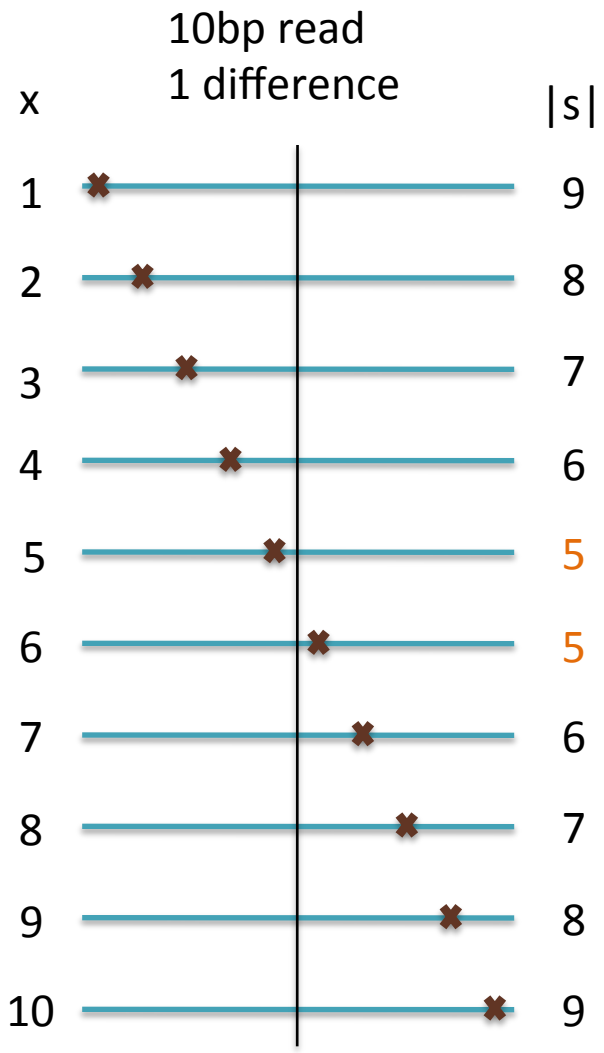


- *-seq: Classify & measure significant peaks



Seed-and-Extend Alignment

Theorem: An alignment of a sequence of length *m* with at most *k* differences ***must*** contain an exact match at least $s=m/(k+1)$ bp long
(*Baeza-Yates* and Perleberg, 1996)



- Proof: Pigeonhole principle
  - 1 pigeon can't fill 2 holes

- Seed-and-extend search
  - Use an index to rapidly find short exact alignments to seed longer in-exact alignments
    - BLAST, MUMmer, Bowtie, BWA, SOAP, …
  - Specificity of the depends on seed length
    - Guaranteed sensitivity for k differences
    - Also finds some (but not all) lower quality alignments <- heuristic

Exact Matching Review & Overview

Where is GATTACA in the human genome?

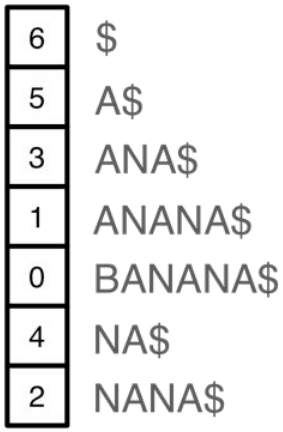
Brute Force (3 GB)

```
BANANA
BAN
  ANA
    NAN
      ANA
```

Naive

Slow & Easy

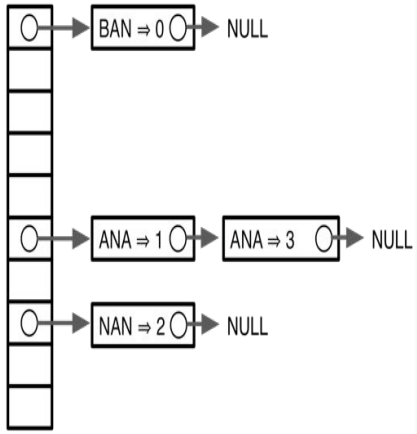
Suffix Array (>15 GB)



Vmatch, PacBio Aligner

Binary Search

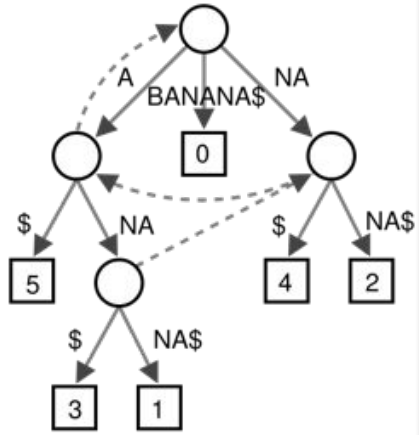
Hash Table (>15 GB)



BLAST, MAQ, ZOOM, RMAP, CloudBurst

Seed-and-extend

Suffix Tree (>51 GB)



MUMmer, MUMmerGPU

Tree Searching

*** These are general techniques applicable to any search problem ***

Algorithmic challenge

How can we combine the speed of a suffix tree (O(|q|) exact match) with the size of a brute force analysis (n bytes)?

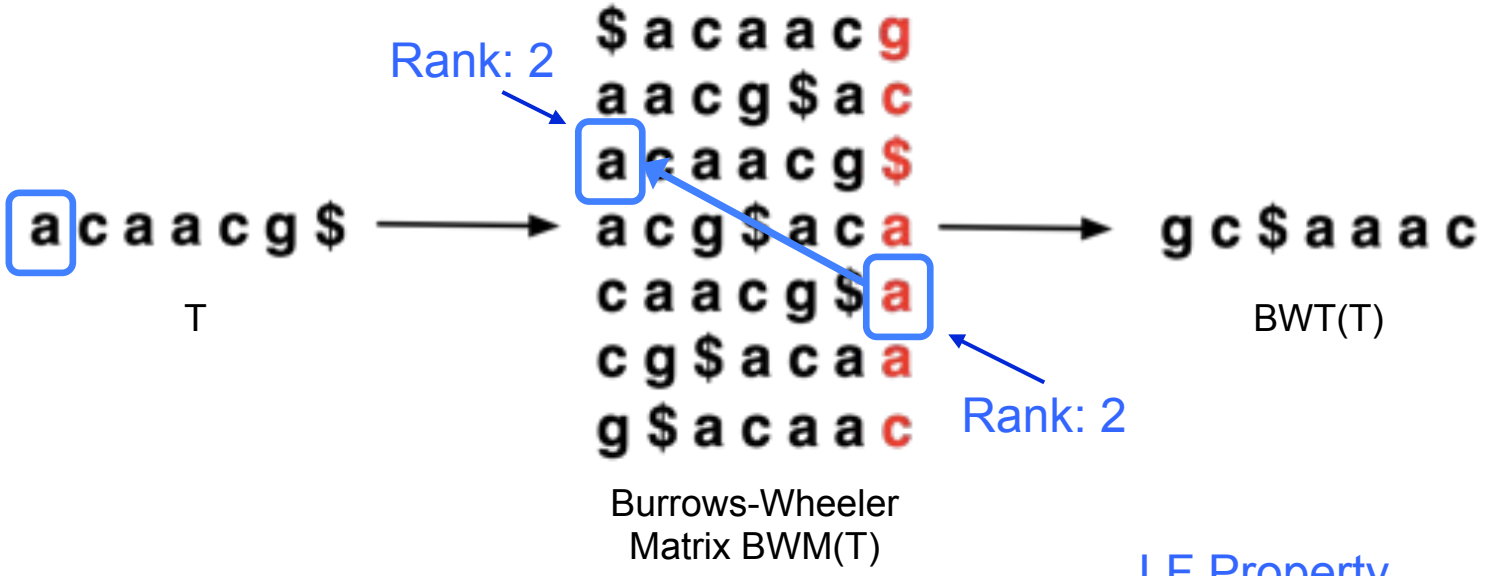
What would such an index look like?

Fast gapped-read alignment with Bowtie 2

Ben Langmead and Steven Salzberg (2012) Nature Methods. 9, 357–359

Burrows-Wheeler Transform

- Reversible permutation of the characters in a text



LF Property implicitly encodes Suffix Array

- BWT(T) is the index for T

A block sorting lossless data compression algorithm.
Burrows M, Wheeler DJ (1994) *Digital Equipment Corporation.* Technical Report 124

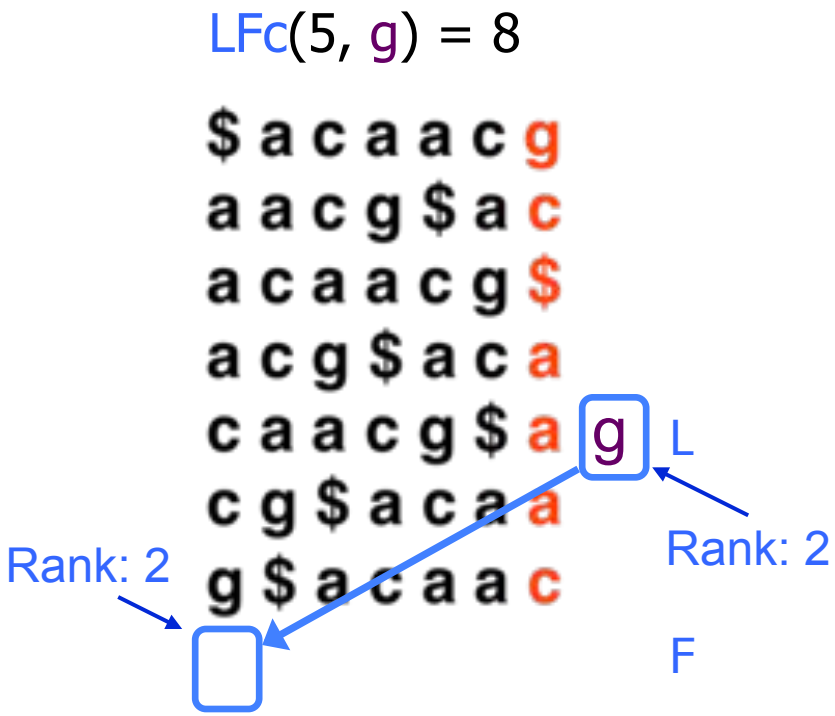
Burrows-Wheeler Transform

- Recreating T from BWT(T)
  - Start in the first row and apply **LF** repeatedly, accumulating predecessors along the way

[Decode this BWT string: ACTGA$TTA]

BWT Exact Matching

- **LFc**(r, c) does the same thing as **LF**(r) but it ignores r's actual final character and "pretends" it's c:



BWT Exact Matching

- Start with a range, (**top**, **bot**) encompassing all rows and repeatedly apply **LFc**:

  top = **LFc**(**top**, **qc**); **bot** = **LFc**(**bot**, **qc**)

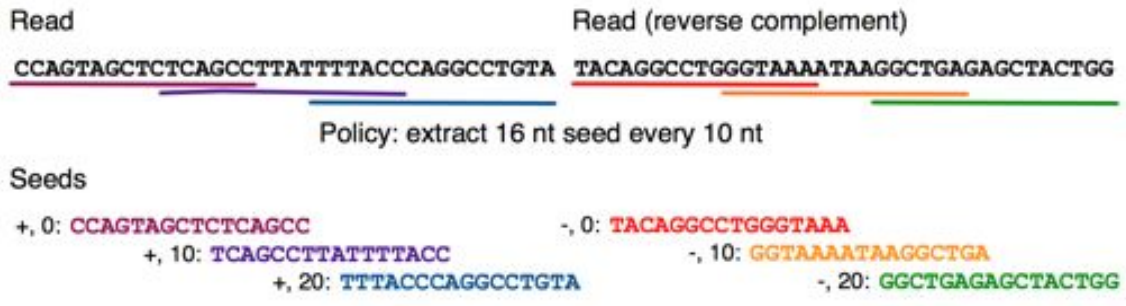
  qc = the next character to the left in the query

Ferragina P, Manzini G: Opportunistic data structures with applications. *FOCS. IEEE Computer Society; 2000.*

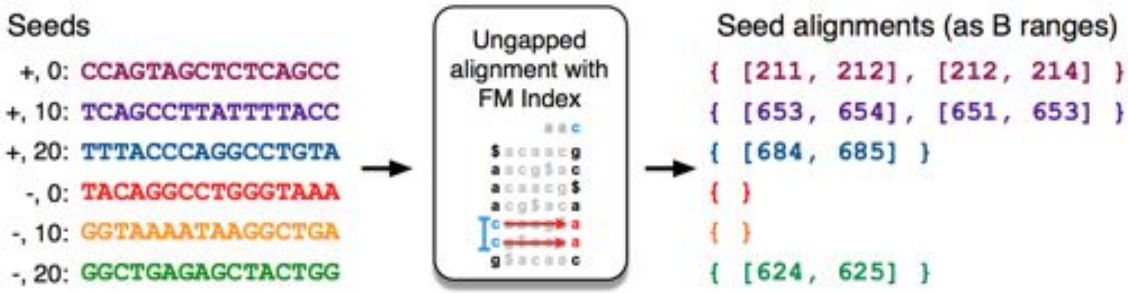
[Search for TTA this BWT string: ACTGA$TTA]

Algorithm Overview

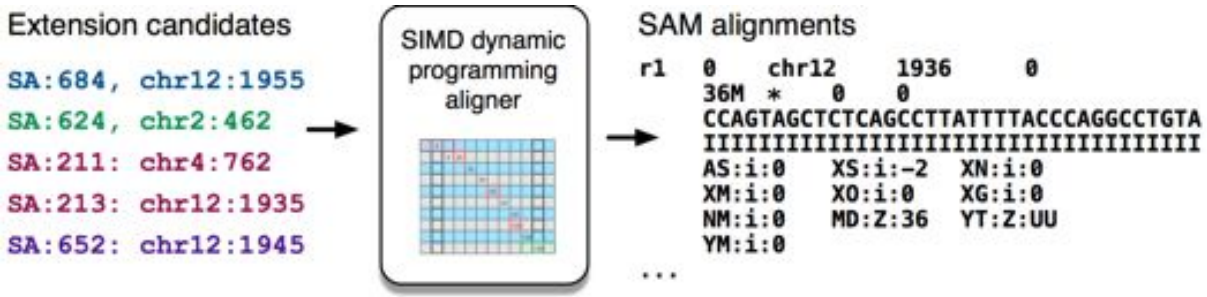
1. Split read into segments



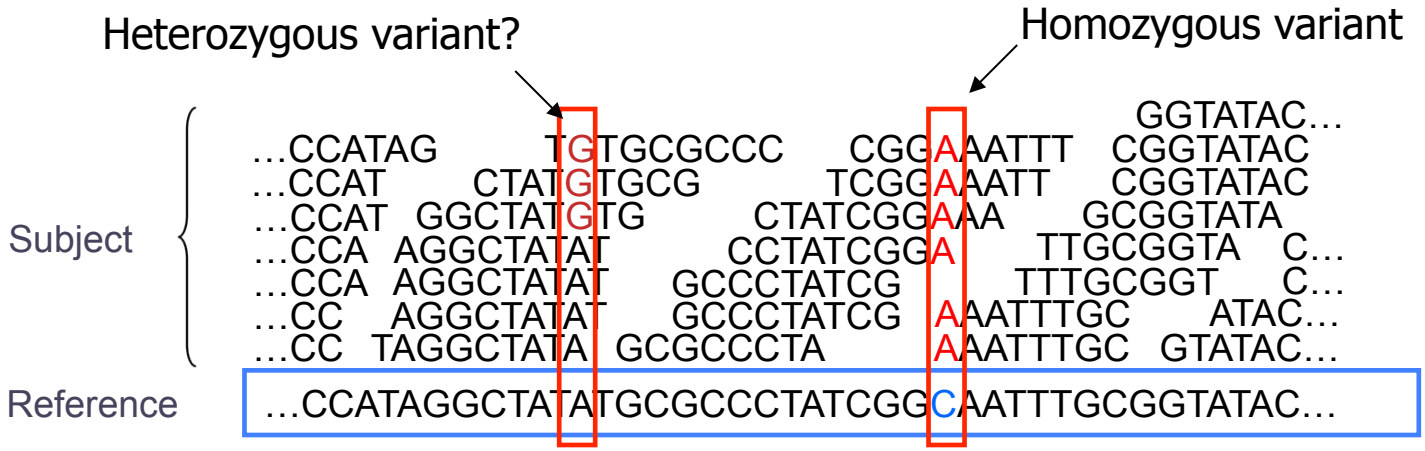
2. Lookup each segment and prioritize



3. Evaluate end-to-end match

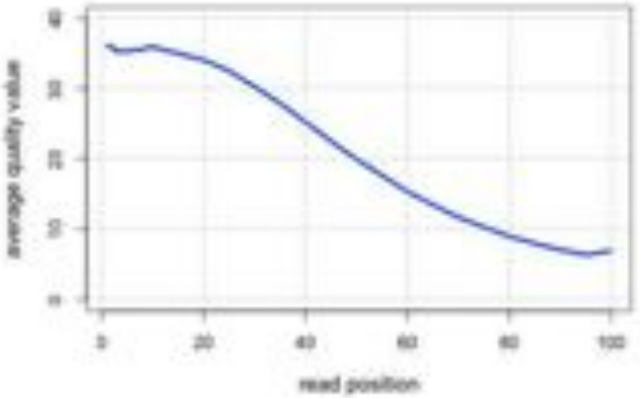


Genotyping



- Sequencing instruments make mistakes
  - Quality of read decreases over the read length

$$Q_{\text{sanger}} = -10 \log_{10} p$$



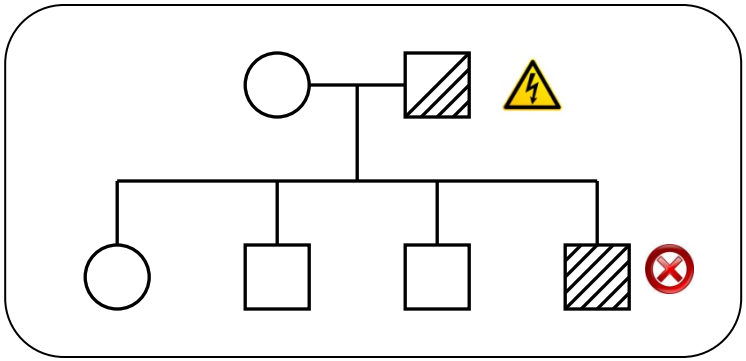
- A single read differing from the reference is probably just an error, but it becomes more likely to be real as we see it multiple times
  - Often framed as a Bayesian problem of more likely to be a real variant or chance occurrence of N errors
  - Accuracy improves with deeper coverage

Outline

1. Rise of DNA Sequencing
2. Alignment and the BWT
3. Genetics of Autism

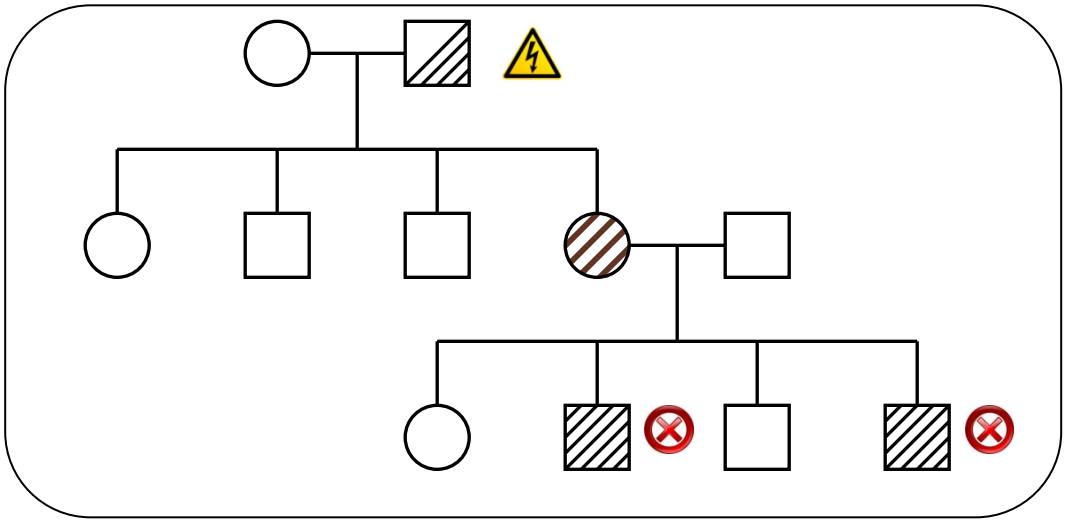
Unified Model of Autism

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



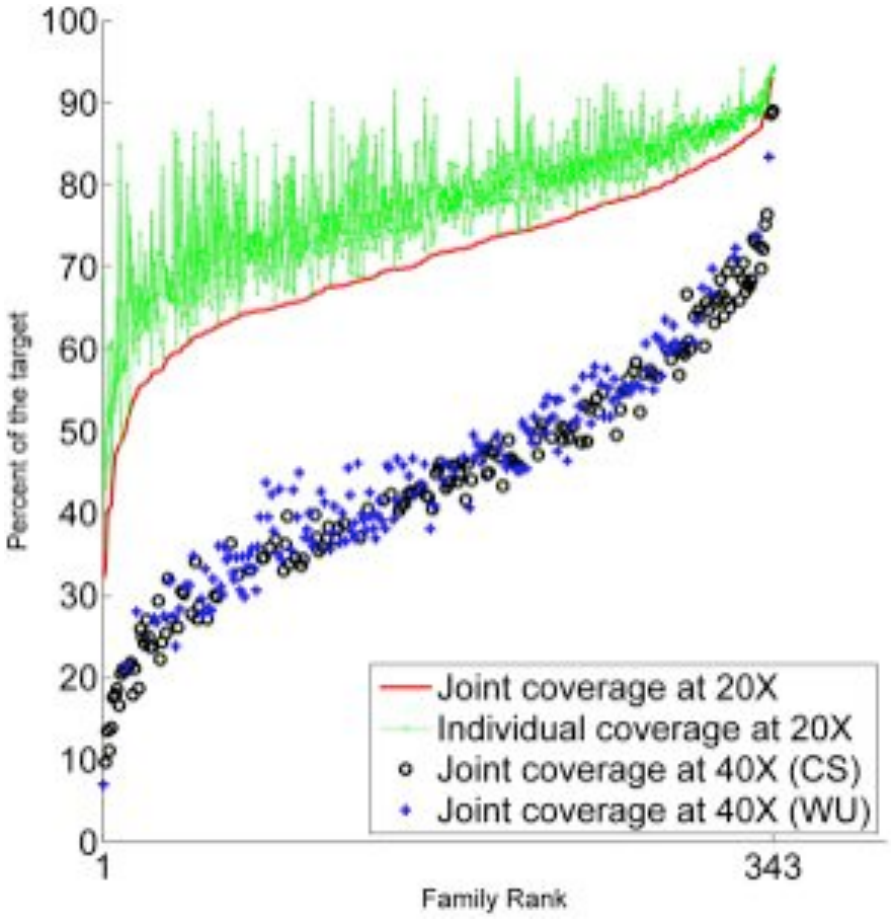
Legend

Sporadic mutation

Fails to procreate

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

Exome-Capture and Sequencing



Sequencing of 343 families from the Simons Simplex Collection

- Parents plus one child with autism and one non-autistic sibling
- Enriched for higher-functioning individuals

Families prepared and captured together to minimize batch effects

- Exome-capture performed with NimbleGen SeqCap EZ Exome v2.0 targeting 36 Mb of the genome.
- ~80% of the target at >20x coverage with ~93bp reads

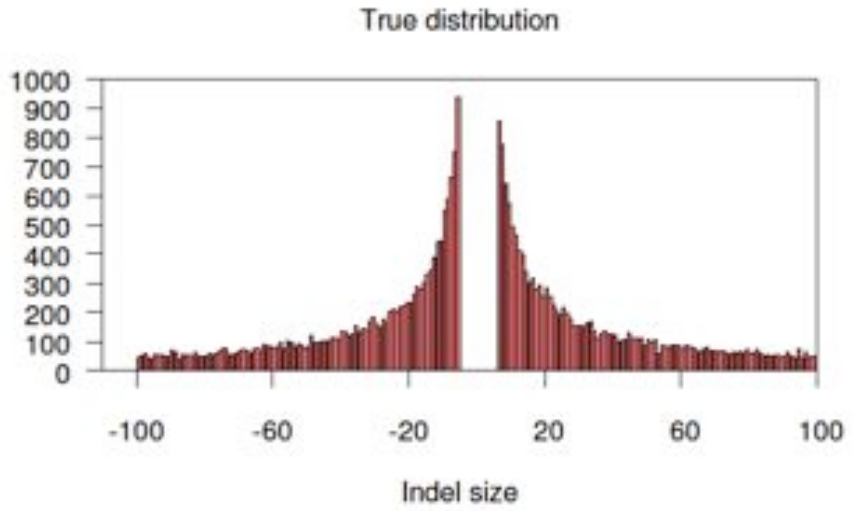
De novo gene disruptions in children on the autism spectrum
Iossifov *et al.* (2012) *Neuron*. 74:2 285-299

Variation Detection Complexity

SNPs + Short Indels

High precision and sensitivity

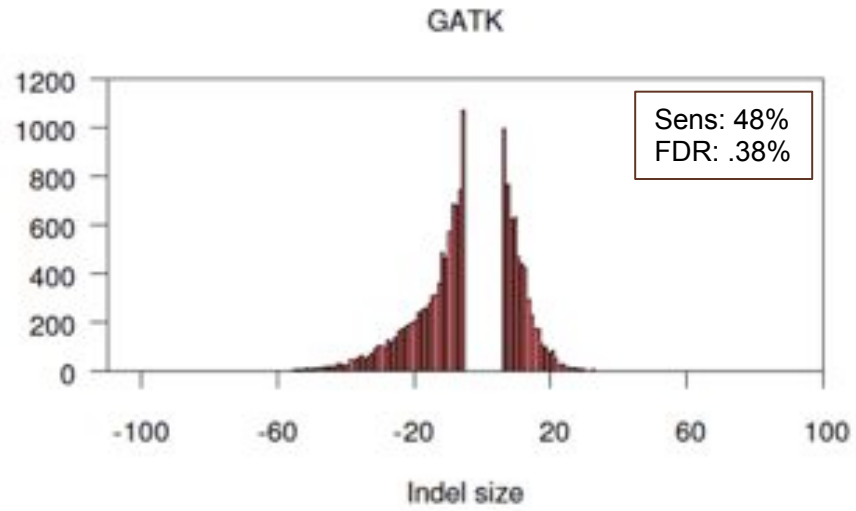
```
..TTTAGAATAG-CGAGTGC...
     ||||||| ||||
     AGAATAGGCGAG
```



"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

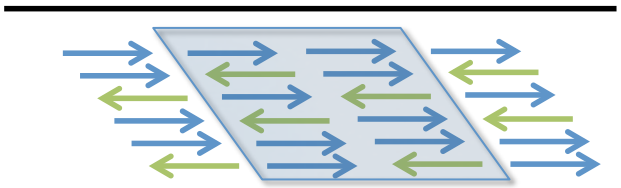
NRXN1 *de novo* SNP
(auSSC12501 chr2:50724605)

SCALPEL: Micro-assembly approach to accurately detect de novo and transmitted indel mutations within exome-Capture data

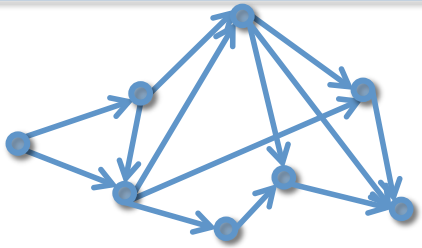
Narzisi, G, O'Rawe, J, Iossifov, I, Lee, Y, Wang, Z, Wu, Y, Lyon, G, Wigler, M, Schatz, MC (2013) *In preparation*

Scalpel Pipeline

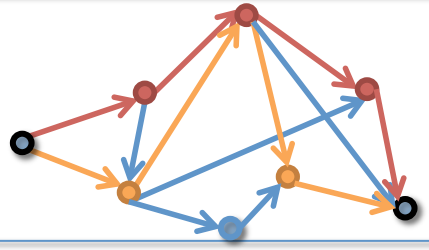
Extract reads mapping within the exon including (1) well-mapped reads, (2) soft-clipped reads, and (3) anchored pairs



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

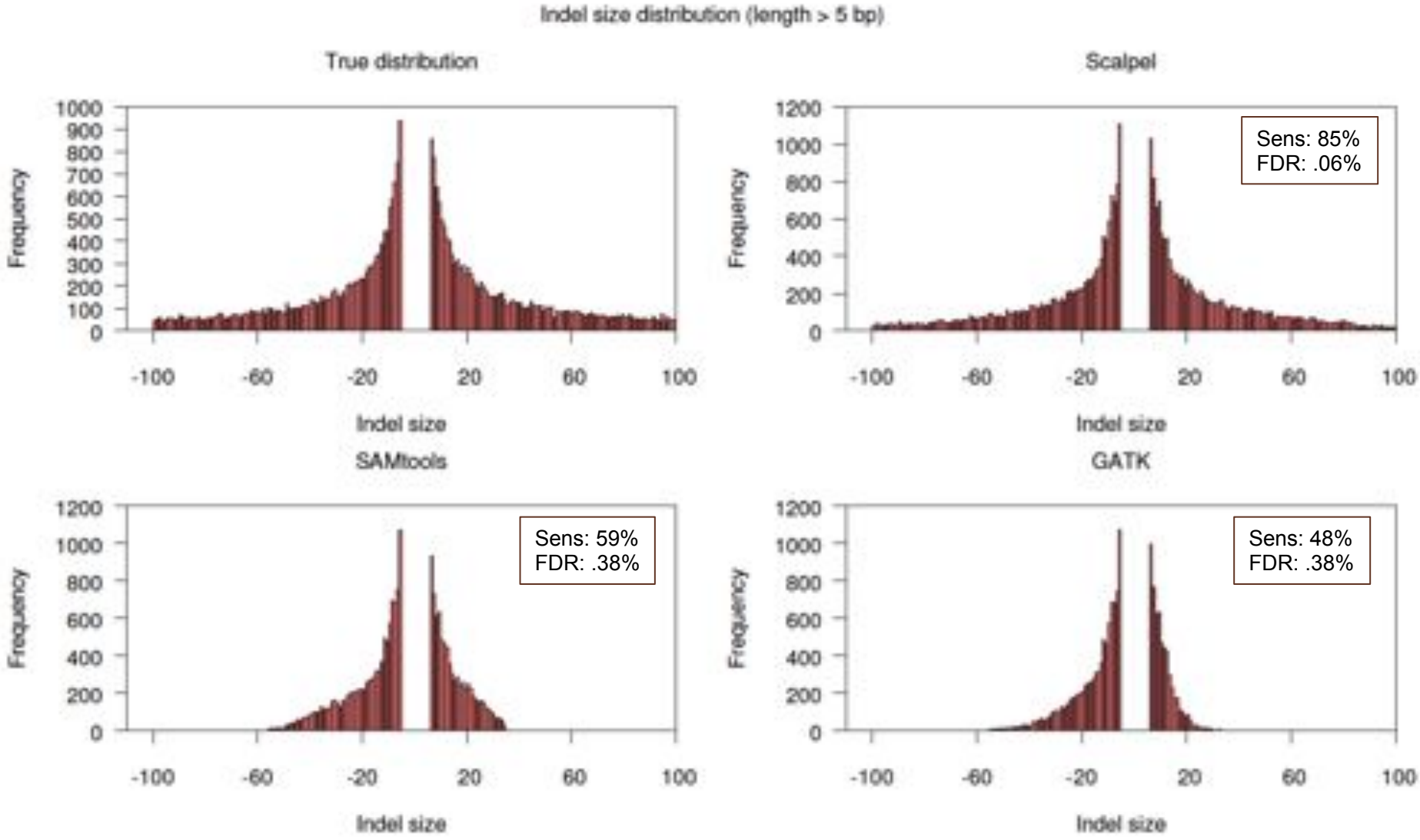


Find end-to-end haplotype paths spanning the region



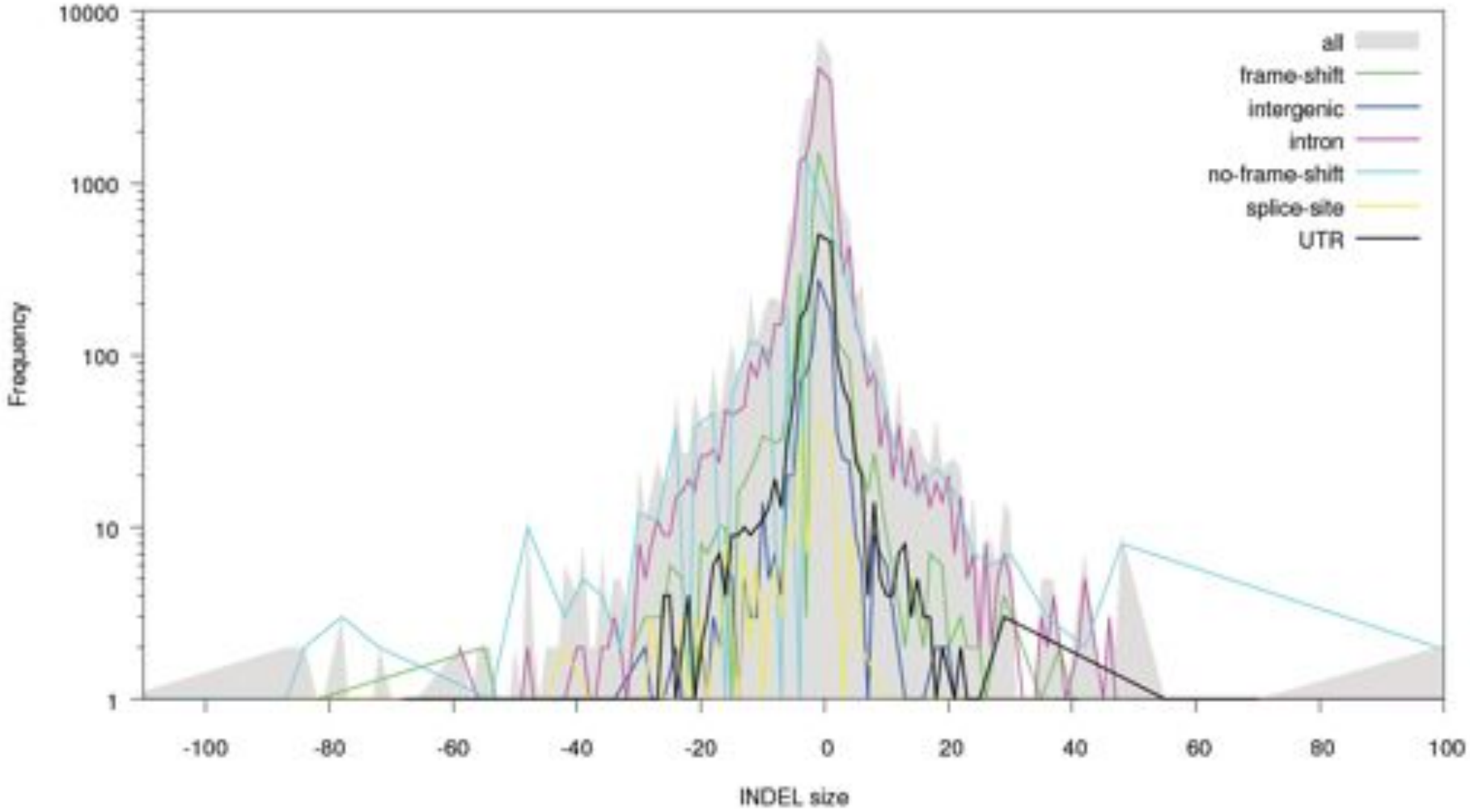
Align assembled sequences to reference to detect mutations

Simulation Analysis



Simulated 10,000 indels in a exome from a known log-normal distribution

Revised Analysis of the SSC

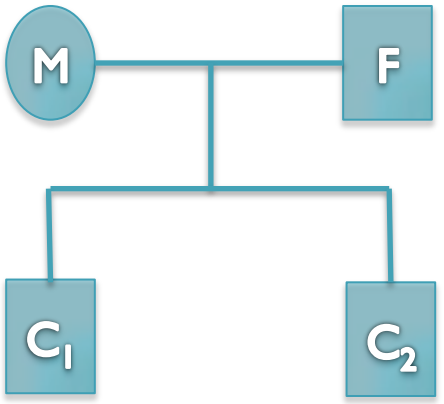


Constructed database of >1M transmitted and de novo indels
Many new gene candidates identified, population analysis underway

De novo mutation discovery and validation

Concept: Identify mutations not present in parents.

Challenge: Sequencing errors in the child or low coverage in parents lead to false positive de novos




```
Ref:     ...TCAGAACAGCTGGATGAGATCTTAGCCAACTACCAGGAGATTGTCTTTGCCCGGA...

Father:  ...TCAGAACAGCTGGATGAGATCTTAGCCAACTACCAGGAGATTGTCTTTGCCCGGA...
Mother:  ...TCAGAACAGCTGGATGAGATCTTAGCCAACTACCAGGAGATTGTCTTTGCCCGGA...
Sib:     ...TCAGAACAGCTGGATGAGATCTTAGCCAACTACCAGGAGATTGTCTTTGCCCGGA...
Aut(1):  ...TCAGAACAGCTGGATGAGATCTTAGCCAACTACCAGGAGATTGTCTTTGCCCGGA...
Aut(2):  ...TCAGAACAGCTGGATGAGATCTTACC------CCGGAGATTGTCTTTGCCCGGA...
```

6bp heterozygous deletion at chr13:25280526 ATP12A

De novo Genetics of Autism

- In 343 family quads so far, we see significant enrichment in de novo ***likely gene killers*** in the autistic kids
  - Overall rate basically 1:1 (432:396)
  - 2:1 enrichment in nonsense mutations
  - 2:1 enrichment in frameshift indels
  - 4:1 enrichment in splice-site mutations
  - Most de novo originate in the paternal line in an age-dependent manner (56:18 of the mutations that we could determine)
- Observe strong overlap with the 842 genes known to be associated with fragile X protein FMPR
  - Related to neuron development and synaptic plasticity

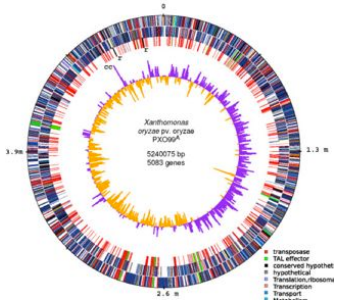
De novo gene disruptions in children on the autism spectrum
Iossifov *et al.* (2012) *Neuron*. 74:2 285-299

Summary

I'm interested in answering biological questions by developing and applying novel algorithms and computational systems

- Interesting biological systems: human diseases, foods, biofuels
- Interesting biotechnology: new sequencing technologies
- Interesting computational systems: parallel & cloud technology
- Interesting algorithms: assembly, alignment, interpretation

Also extremely excited to teach the next generation of scientists in the WSBS, URP, and high school programs



Acknowledgements

Schatz Lab
Giuseppe Narzisi
Shoshana Marcus
James Gurtowski
Srividya Ramakrishnan
Hayan Lee
Rob Aboukhalil
Mitch Bekritsky
Charles Underwood
Tyler Gavin
Alejandro Wences
Greg Vurture
Eric Biggers
Aspyn Palatnick

CSHL
Hannon Lab
Gingeras Lab
Jackson Lab
Iossifov Lab
Levy Lab
Lippman Lab
Lyon Lab
Martienssen Lab
McCombie Lab
Ware Lab
Wigler Lab

IT Department

SFARI
SIMONS FOUNDATION AUTISM RESEARCH INITIATIVE

National Human Genome Research Institute

U.S. DEPARTMENT OF ENERGY

NSF

See you at

Genome Informatics

Oct 30 – Nov 2

http://schatzlab.cshl.edu
@mike_schatz