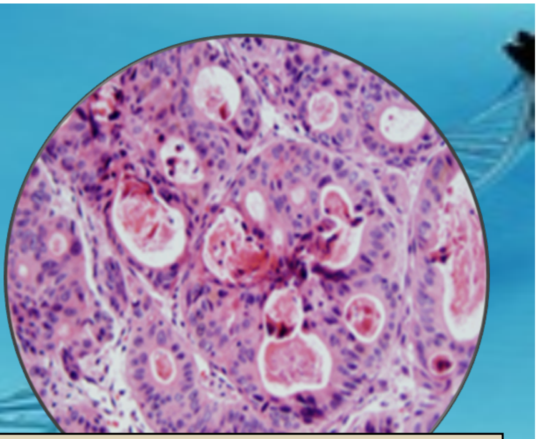
Human Genetics and Plant Genomics: The long and the short of it

Michael Schatz
Simons Center for Quantitative Biology

CSHL In-House Symposium XXVI
November 20, 2012

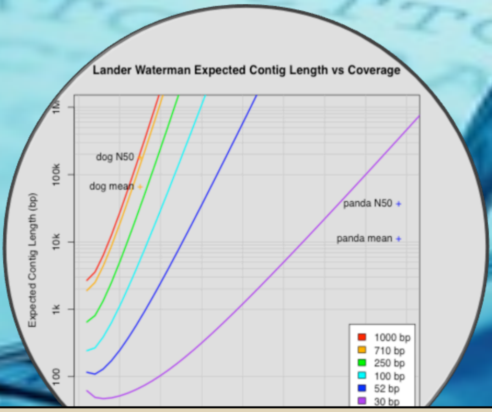
Schatz Lab Overview

Computation

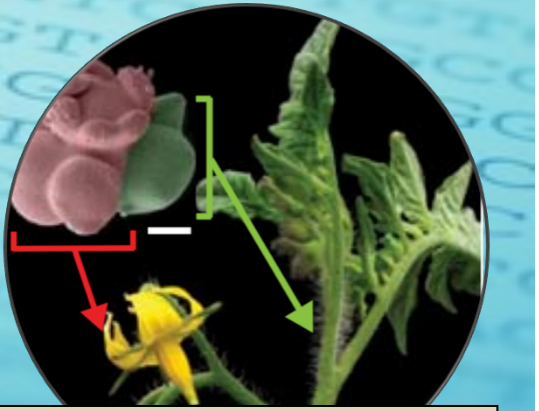


Human Genetics

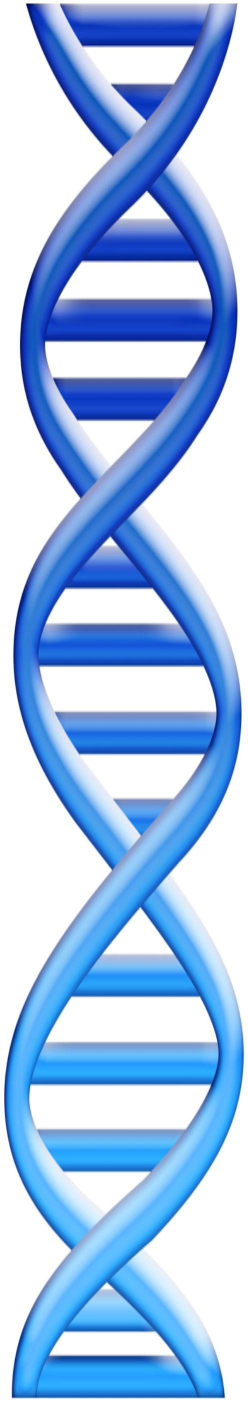
Sequencing



Modeling



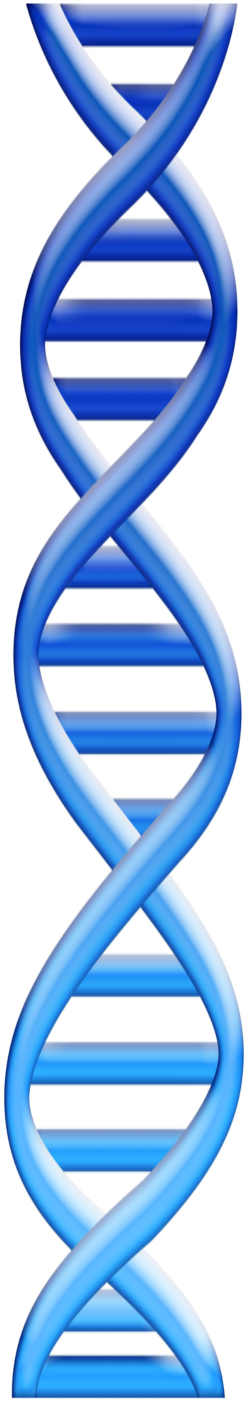
Plant Genomics



Outline

1. De novo mutations in human diseases
    1. Autism Spectrum Disorder
    2. Applications to ADHD & Tourette's

2. Plant Genome Assembly
    1. Long read single molecule sequencing
    2. Other applications



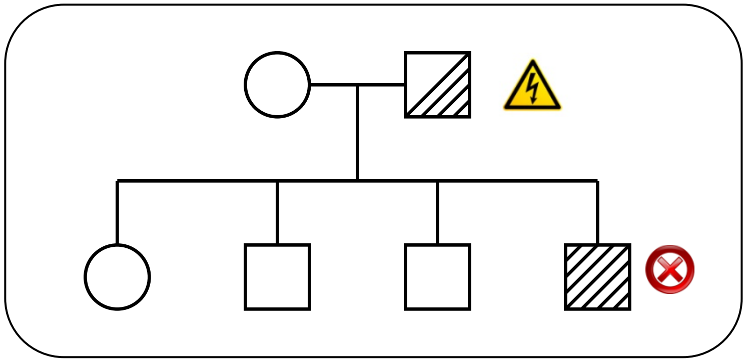
Outline

1. De novo mutations in human diseases
   1. Autism Spectrum Disorder
   2. Applications to ADHD & Tourette's

2. Plant Genome Assembly
   1. Long read single molecule sequencing
   2. Other applications

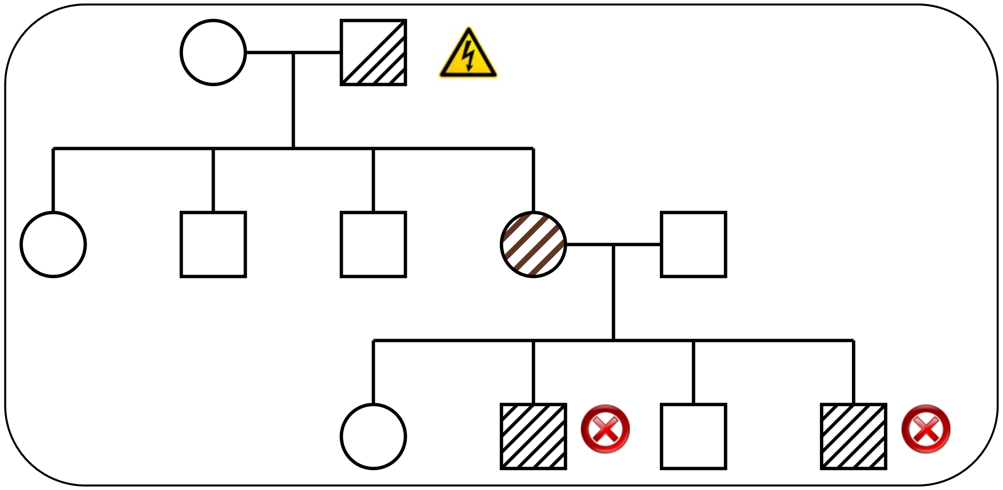
Unified Model of Autism

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



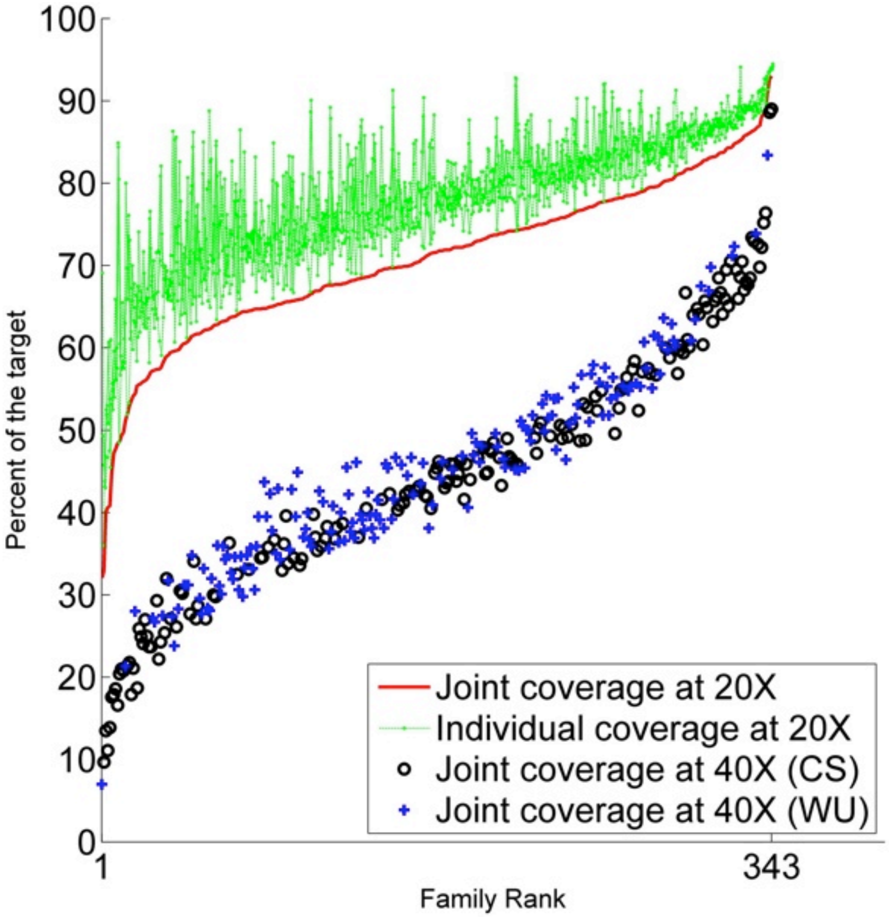
Legend

⚠ Sporadic mutation

⊗ Fails to procreate

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

Exome sequencing of the SSC



Sequencing of 343 families from the Simons Simplex Collection

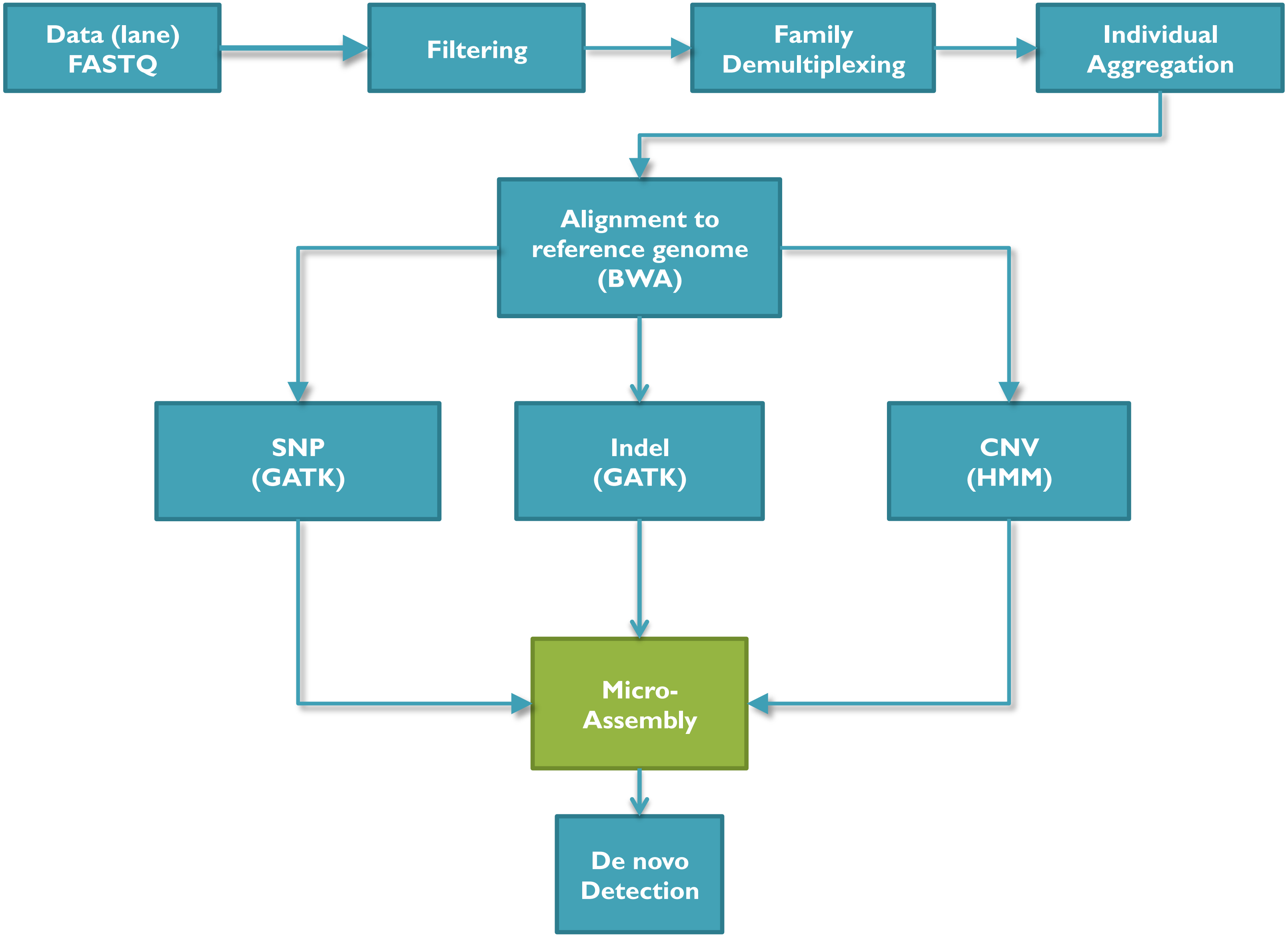
- Parents plus one child with autism and one non-autistic sibling
- Enriched for higher-functioning individuals

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

Exome Sequencing Pipeline

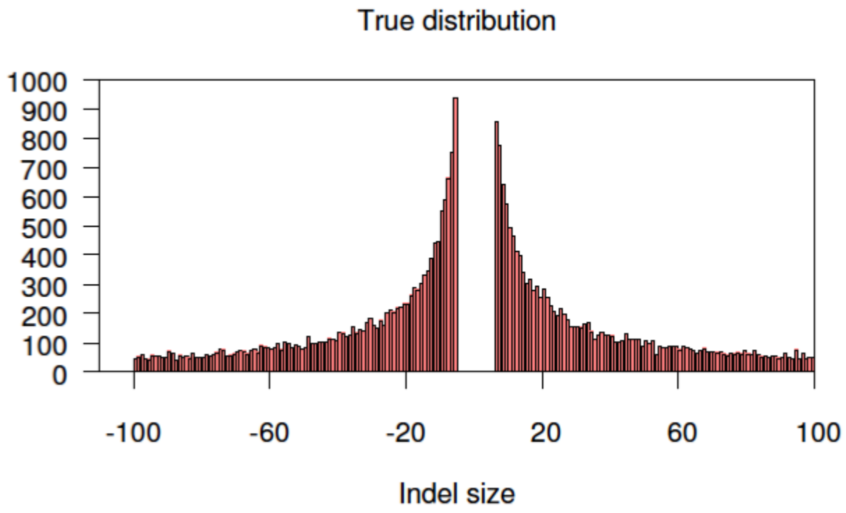


Variation Detection Complexity

SNPs + Short Indels

High precision and sensitivity

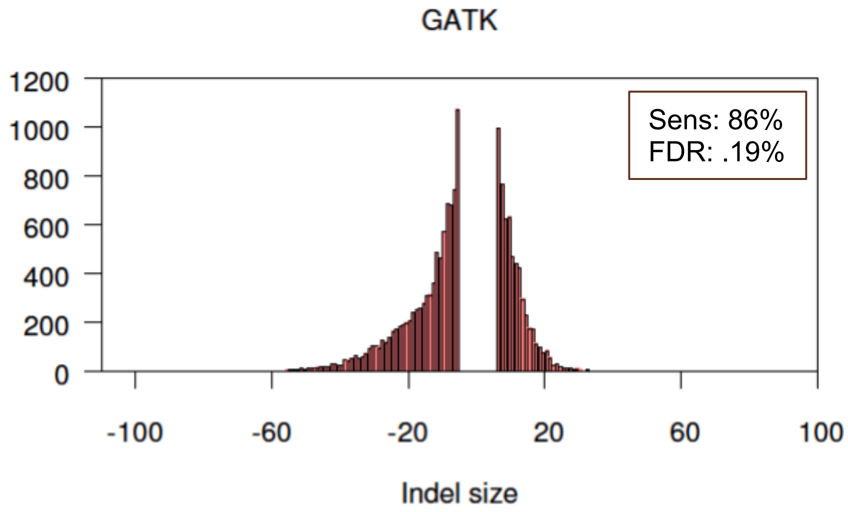
```
..TTTAGAATAG-CGAGTGC...
     ||||||| ||||
     AGAATAGGCGAG
```



"Long" Indels (>5bp)

Reduced precision and sensitivity

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



Analysis confounded by localized repeats: 30% of exons have at least a 10bp repeat

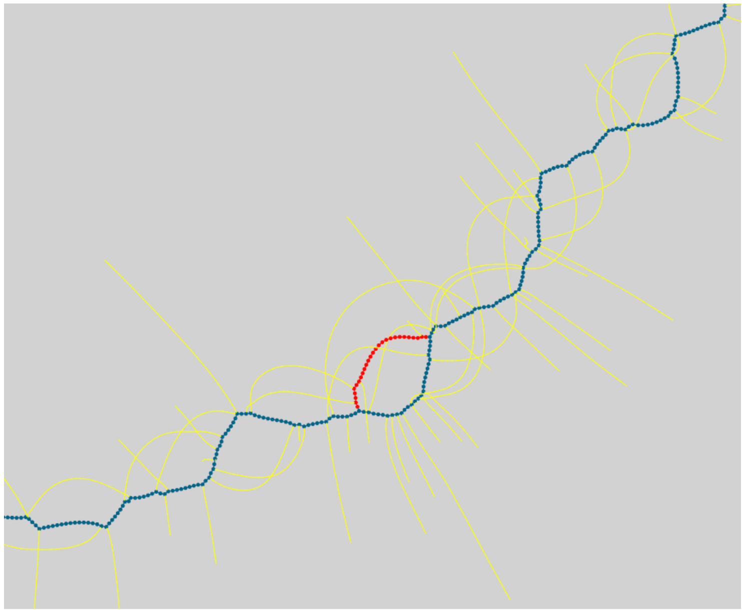
Scalpel: Haplotype Microassembly

G. Narzisi, D. Levy, I. Iossifov, J. Kendall, M. Wigler, M. Schatz

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

Features

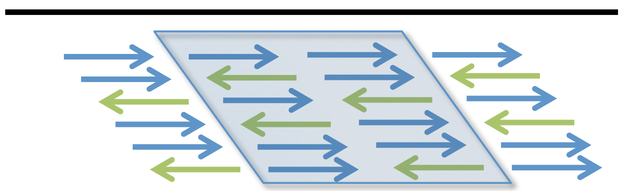
1. Combine mapping and assembly
2. Exhaustive search of haplotypes
3. De novo mutations



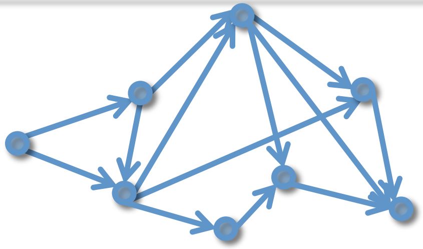
NRXN1 *de novo* SNP
(auSSC12501 chr2:50724605)

Scalpel Pipeline

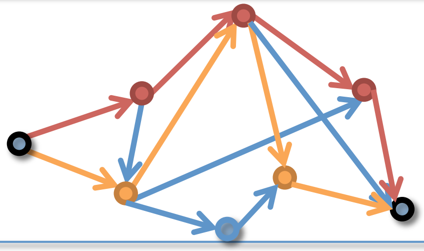
Extract reads mapping within the exon including (1) well-mapped reads, (2) soft-clipped reads, and (3) anchored pairs



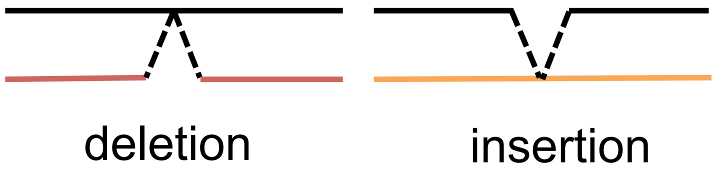
Decompose reads into overlapping *k*-mers and construct de Bruijn graph from the reads



Find end-to-end haplotype paths spanning the region



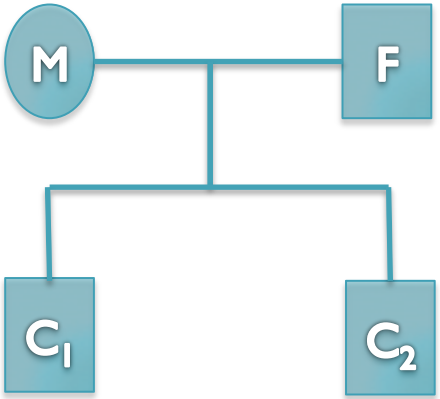
Align assembled sequences to reference to detect mutations



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------CCGGGAGATTGTCTTTGCCCGGA...
```

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)

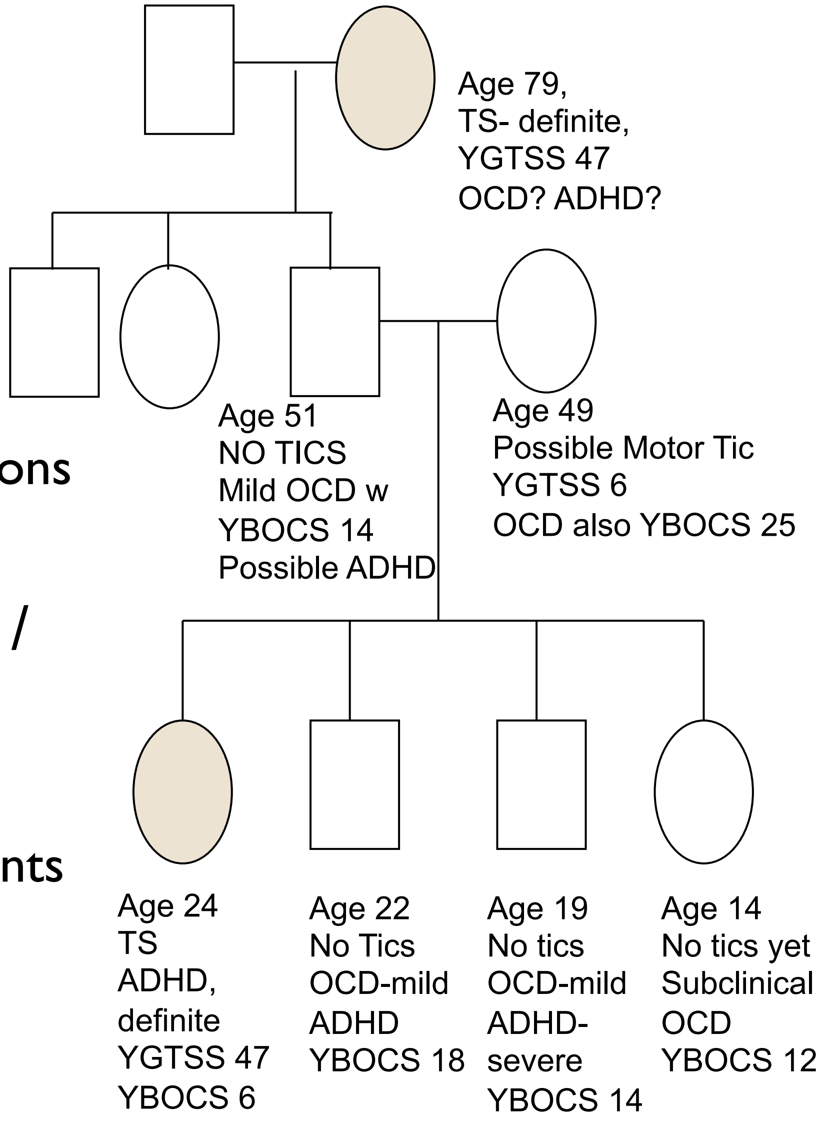
- Observe strong overlap with the 842 genes known to be associated with fragile X protein FMPR
  - Related to neuron development and synaptic plasticity
  - Also strong overlap with chromatin remodelers

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

Applications to ADHD & Tourette's

J. O'Rawe, G. Narzisi, M. Schatz, G. Lyon

- We believe similar mechanisms are involved in ADHD and Tourette's syndrome
  - Begun sequencing of families
  - Identify de novo and segregating mutations
- Cross analysis of GATK / SAMTools / SOAPindel / Scapel
  - High concordance on small events
  - Scalpel tends to identify more large events
  - Extensive wetlab validation in progress


Age 79,
TS- definite,
YGTSS 47
OCD? ADHD?

Age 51
NO TICS
Mild OCD w
YBOCS 14
Possible ADHD

Age 49
Possible Motor Tic
YGTSS 6
OCD also YBOCS 25

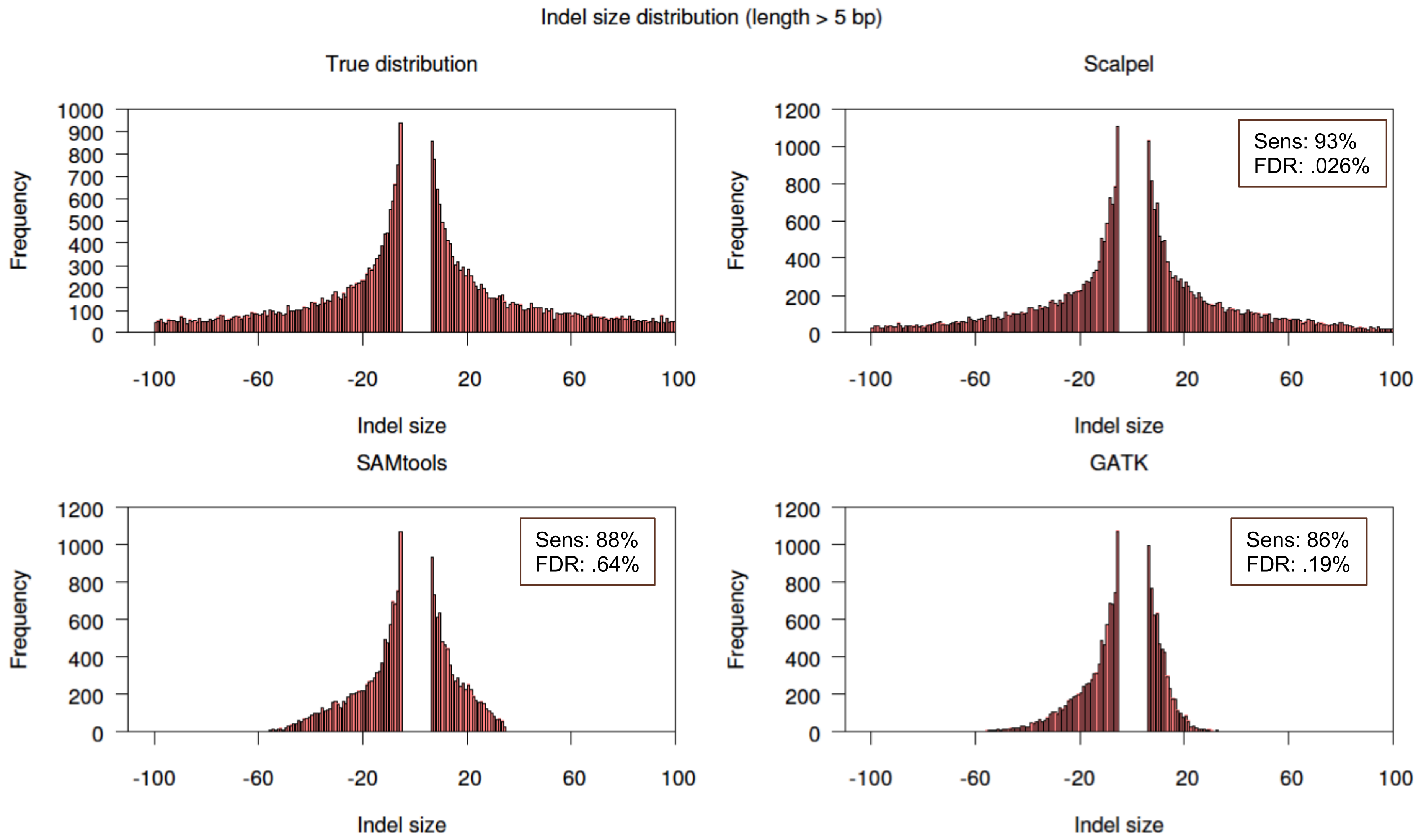
Age 24
TS
ADHD,
definite
YGTSS 47
YBOCS 6

Age 22
No Tics
OCD-mild
ADHD
YBOCS 18

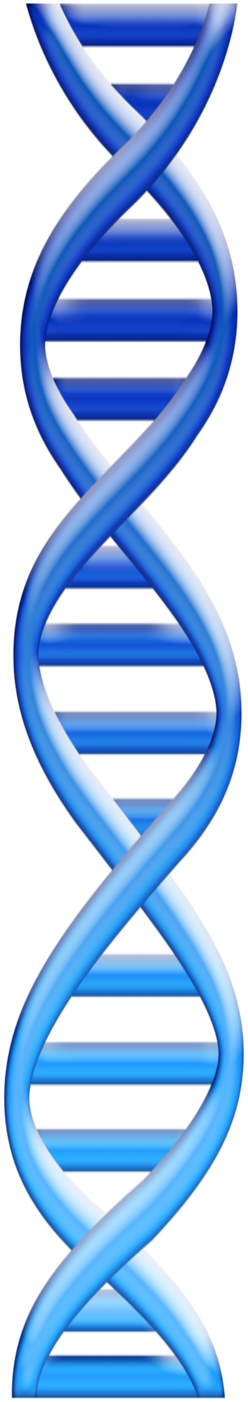
Age 19
No tics
OCD-mild
ADHD-
severe
YBOCS 14

Age 14
No tics yet
Subclinical
OCD
YBOCS 12

Scapel Indel Discovery



Detection of de novo mutations in exome-capture data using micro-assembly
Narzisi *et al.* (2012) *In preparation*



Outline

1. De novo mutations in human diseases
    1. Autism Spectrum Disorder
    2. Applications to ADHD & Tourette's

2. Plant Genome Assembly
    1. Long read single molecule sequencing
    2. Other applications

Genome Assembly Projects

Sacred lotus

***Nelumbo nucifera* Gaertn.**

Ming, R, *et al.* (2012) *Under Review*

Known for religious significance, herbal medicines, seed longevity, and water repellency

Illumina + 454 sequencing

- 900 Mbp Genome Size
- Low Heterozygosity

=> Excellent assembly

Red Raspberry

Rubus ideaus L.

Price, J, *et al.* (2012) *In prep*

Member of the Rosacea family along with apple, pear, peach, strawberry.

Illumina + 454 sequencing

- 300 Mbp Genome Size
- High Heterozygosity

=> Good assembly

Wheat DD

Aegilops tauschii

Schatz/Ware/McCombie collab.

One of the most important cereal crops in the world, one of three ancestral species of allohexaploid bread wheat

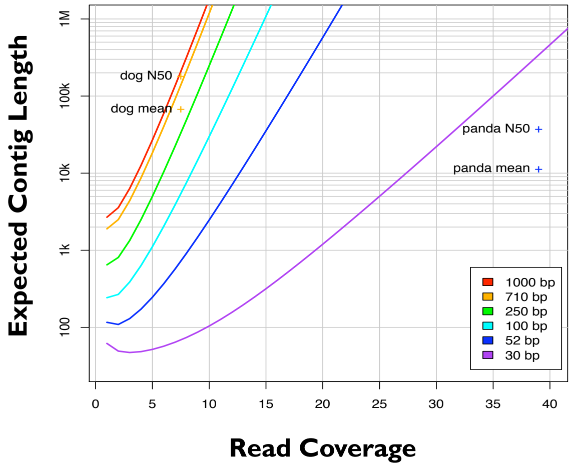
Illumina sequencing

- 4.5 Gbp Genome Size
- High repeat content

=> Challenged assembly

Ingredients for a good assembly

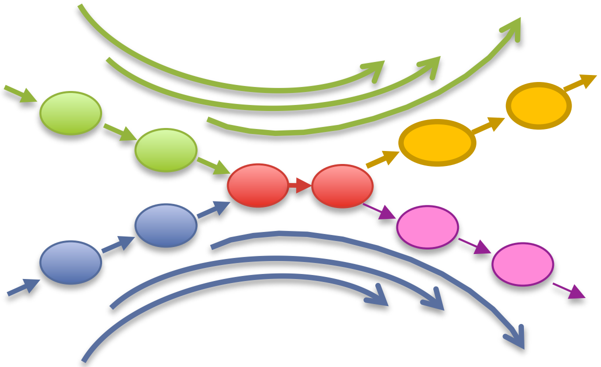
Coverage



High coverage is required

- Oversample the genome to ensure every base is sequenced with long overlaps between reads
- Biased coverage will also fragment assembly

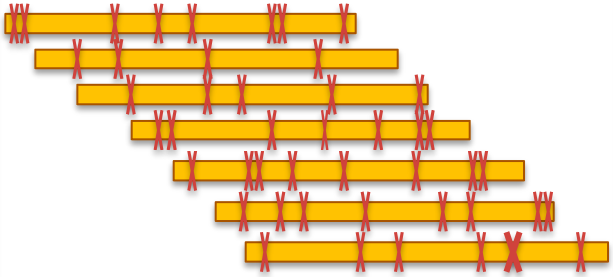
Read Length



Reads & mates must be longer than the repeats

- Short reads will have ***false overlaps*** forming hairball assembly graphs
- With long enough reads, assemble entire chromosomes into contigs

Quality



Errors obscure overlaps

- Reads are assembled by finding kmers shared in pair of reads
- High error rate requires very short seeds, increasing complexity and forming assembly hairballs

Current challenges in *de novo* plant genome sequencing and assembly
Schatz MC, Witkowski, McCombie, WR (2012) *Genome Biology*. 12:243

Hybrid Sequencing

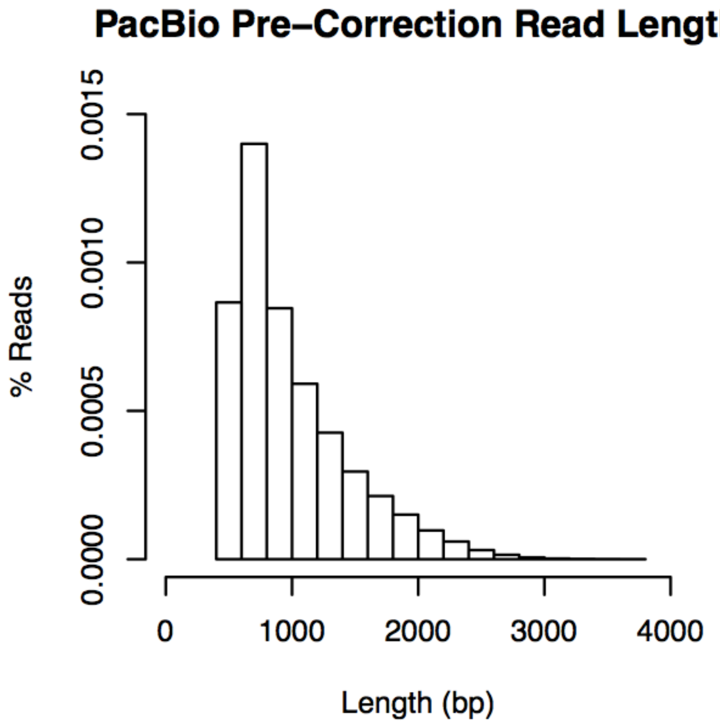
Illumina
Sequencing by Synthesis

High throughput (60Gbp/day)
High accuracy (~99%)
Short reads (~100bp)

Pacific Biosciences
SMRT Sequencing

Lower throughput (600Mbp/day)
Lower accuracy (~85%)
Long reads (2-5kbp+)

SMRT Sequencing Data



Match	83.7%
Insertions	11.5%
Deletions	3.4%
Mismatch	1.4%

```
TTGTAAGCAGTTGAAAACTATGTGTGGATTTAGAATAAAGAACATGAAAG
||||||||||||||||||||||||| ||||||| |||||||||||| |||
TTGTAAGCAGTTGAAAACTATGTGT-GATTTAG-ATAAAGAACATGGAAG

ATTATAAA-CAGTTGATCCATT-AGAAGA-AAACGCAAAAGGCGGCTAGG
| |||||| ||||||||||||| |||| | |||||| |||||| ||||||
A-TATAAATCAGTTGATCCATTAAGAA-AGAAACGC-AAAGGC-GCTAGG

CAACCTTGAATGTAATCGCACTTGAAGAACAAGATTTTATTCCGCGCCCG
| |||||| |||| ||  ||||||||||||||||||||||||||||||||
C-ACCTTG-ATGT-AT--CACTTGAAGAACAAGATTTTATTCCGCGCCCG

TAACGAATCAAGATTCTGAAAACACAT-ATAACAACCTCCAAAA-CACAA
| ||||||| |||||||||||||| || ||    |||||||||| |||||
T-ACGAATC-AGATTCTGAAAACA-ATGAT----ACCTCCAAAAGCACAA

-AGGAGGGGAAAGGGGGGAATATCT-ATAAAAGATTACAAATTAGA-TGA
 ||||||   ||     |||||||| || |||||||||||||| || |||
GAGGAGG---AA-----GAATATCTGAT-AAAGATTACAAATT-GAGTGA

ACT-AATTCACAATA-AATAACACTTTTA-ACAGAATTGAT-GGAA-GTT
||| ||||||||| | ||||||||||||| ||| ||||||| |||| |||
ACTAAATTCACAA-ATAATAACACTTTTAGACAAAATTGATGGGAAGGTT

TCGGAGAGATCCAAAACAATGGGC-ATCGCCTTTGA-GTTAC-AATCAAA
|| ||||||||| ||||||| ||| |||| |||||| ||||| |||||||
TC-GAGAGATCC-AAACAAT-GGCGATCG-CTTTGACGTTACAAATCAAA

ATCCAGTGGAAAATATAATTTATGCAATCCAGGAACTTATTCACAATTAG
||||||| |||||||||  |||||| ||||| ||||||||||||||||||
ATCCAGT-GAAAATATA--TTATGC-ATCCA-GAACTTATTCACAATTAG
```

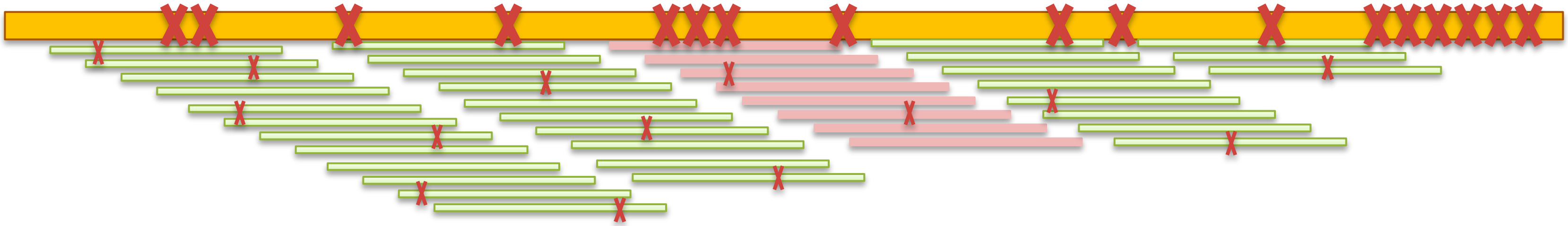
Sample of 100k reads aligned with BLASR requiring >100bp alignment

PacBio Error Correction

http://wgs-assembler.sf.net

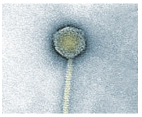
1. Correction Pipeline
    1. Map short reads (SR) to long reads (LR)
    2. Trim LRs at coverage gaps
    3. Compute consensus for each LR

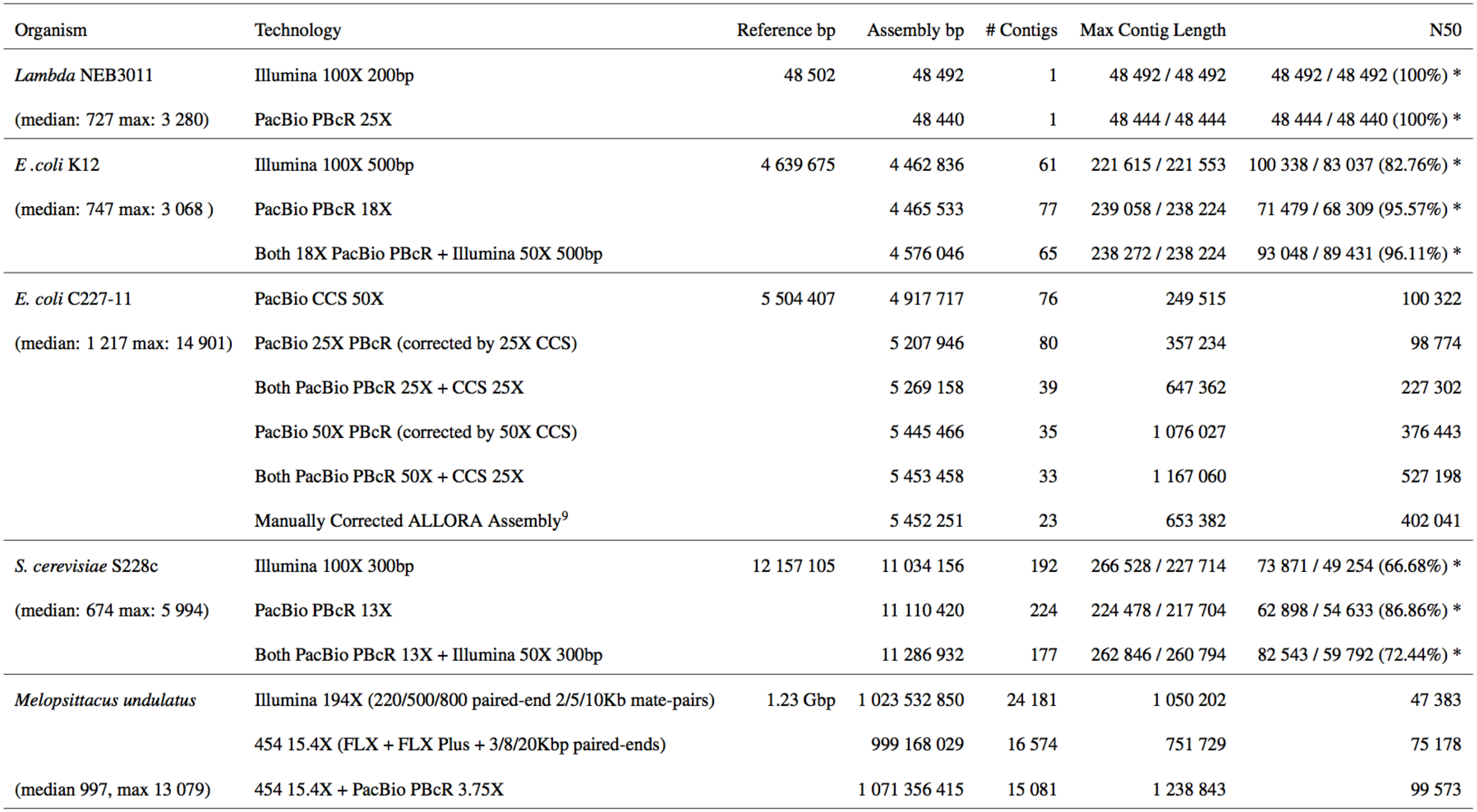
2. Error corrected reads can be easily assembled, aligned
    1. Improves accuracy from ~85% to ~99%



Hybrid error correction and de novo assembly of single-molecule sequencing reads.
Koren, S, Schatz, MC, *et al.* (2012) *Nature Biotechnology.* doi:10.1038/nbt.2280

SMRT-Assembly Results

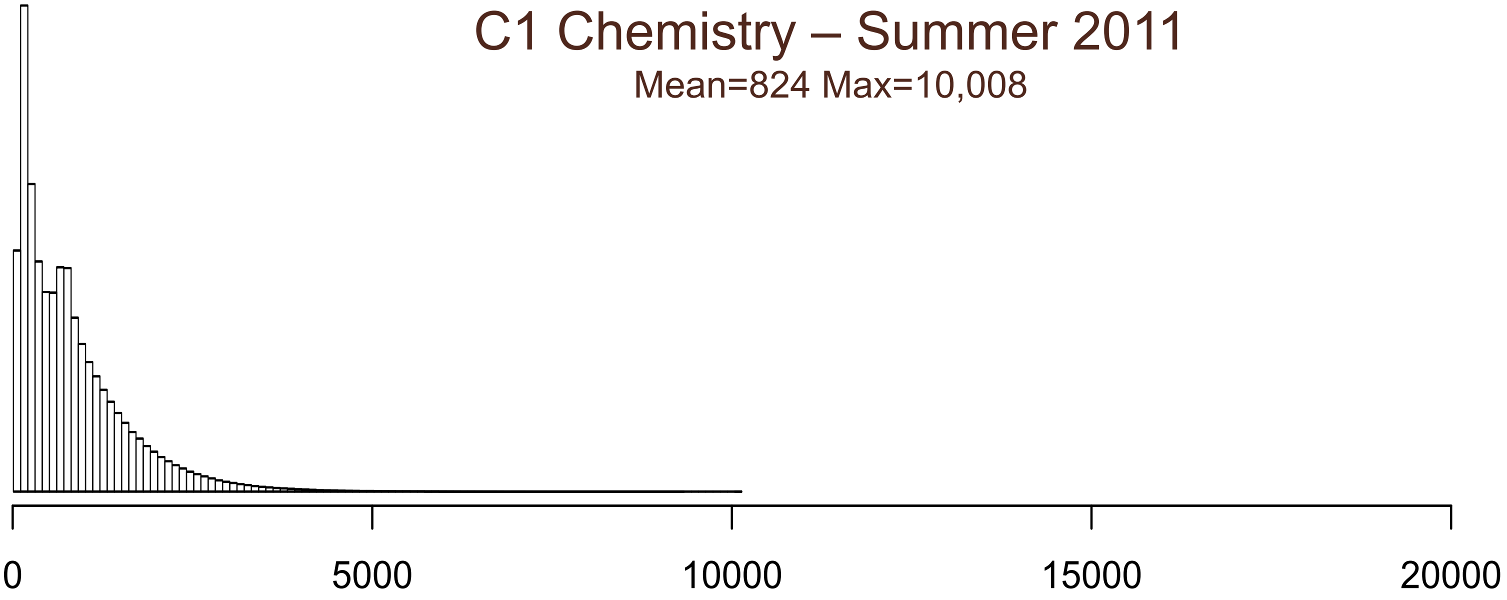




Organism	Technology	Reference bp	Assembly bp	# Contigs	Max Contig Length	N50
Lambda NEB3011	Illumina 100X 200bp	48 502	48 492	1	48 492 / 48 492	48 492 / 48 492 (100%) *
(median: 727 max: 3 280)	PacBio PBcR 25X		48 440	1	48 444 / 48 444	48 444 / 48 440 (100%) *
E .coli K12	Illumina 100X 500bp	4 639 675	4 462 836	61	221 615 / 221 553	100 338 / 83 037 (82.76%) *
(median: 747 max: 3 068)	PacBio PBcR 18X		4 465 533	77	239 058 / 238 224	71 479 / 68 309 (95.57%) *
	Both 18X PacBio PBcR + Illumina 50X 500bp		4 576 046	65	238 272 / 238 224	93 048 / 89 431 (96.11%) *
E. coli C227-11	PacBio CCS 50X	5 504 407	4 917 717	76	249 515	100 322
(median: 1 217 max: 14 901)	PacBio 25X PBcR (corrected by 25X CCS)		5 207 946	80	357 234	98 774
	Both PacBio PBcR 25X + CCS 25X		5 269 158	39	647 362	227 302
	PacBio 50X PBcR (corrected by 50X CCS)		5 445 466	35	1 076 027	376 443
	Both PacBio PBcR 50X + CCS 25X		5 453 458	33	1 167 060	527 198
	Manually Corrected ALLORA Assembly[9]		5 452 251	23	653 382	402 041
S. cerevisiae S228c	Illumina 100X 300bp	12 157 105	11 034 156	192	266 528 / 227 714	73 871 / 49 254 (66.68%) *
(median: 674 max: 5 994)	PacBio PBcR 13X		11 110 420	224	224 478 / 217 704	62 898 / 54 633 (86.86%) *
	Both PacBio PBcR 13X + Illumina 50X 300bp		11 286 932	177	262 846 / 260 794	82 543 / 59 792 (72.44%) *
Melopsittacus undulatus	Illumina 194X (220/500/800 paired-end 2/5/10Kb mate-pairs)	1.23 Gbp	1 023 532 850	24 181	1 050 202	47 383
	454 15.4X (FLX + FLX Plus + 3/8/20Kbp paired-ends)		999 168 029	16 574	751 729	75 178
(median 997, max 13 079)	454 15.4X + PacBio PBcR 3.75X		1 071 356 415	15 081	1 238 843	99 573

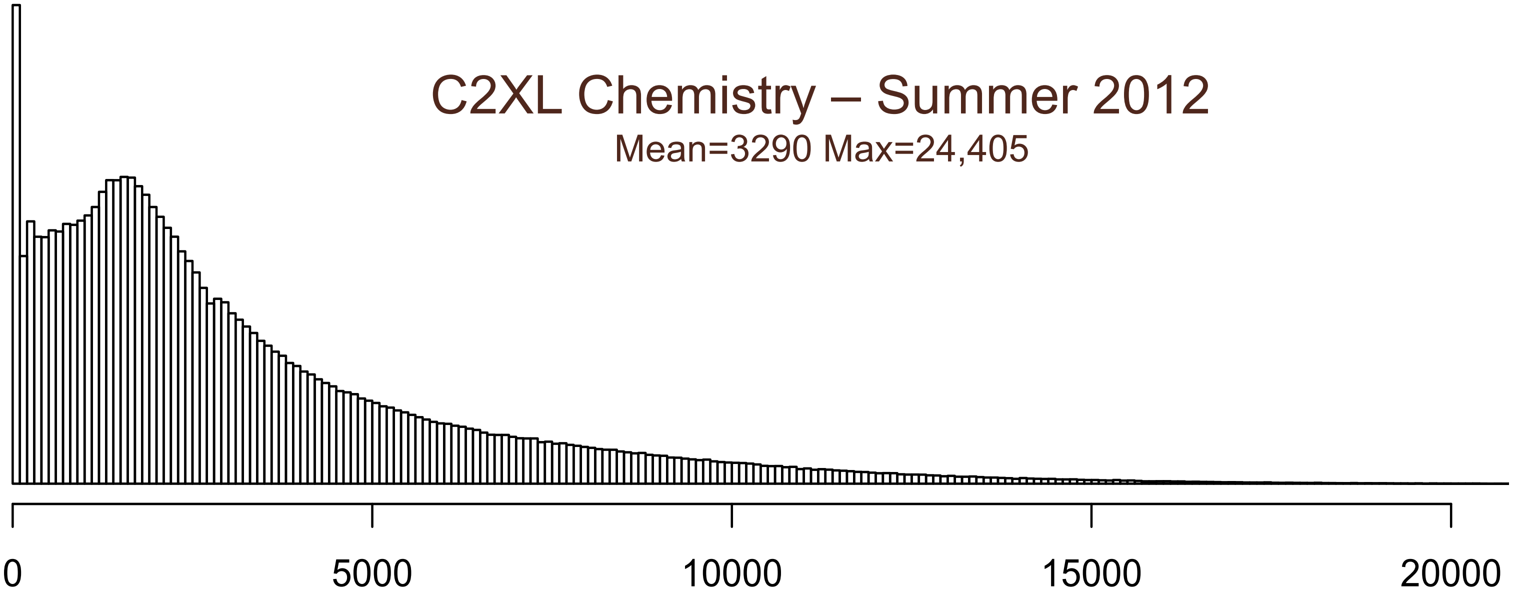
Hybrid assembly results using error corrected PacBio reads
Meets or beats Illumina-only or 454-only assembly in every case
*** Also useful for transcriptome and CNV analysis ***

PacBio Long Read Sequencing



C1 Chemistry – Summer 2011

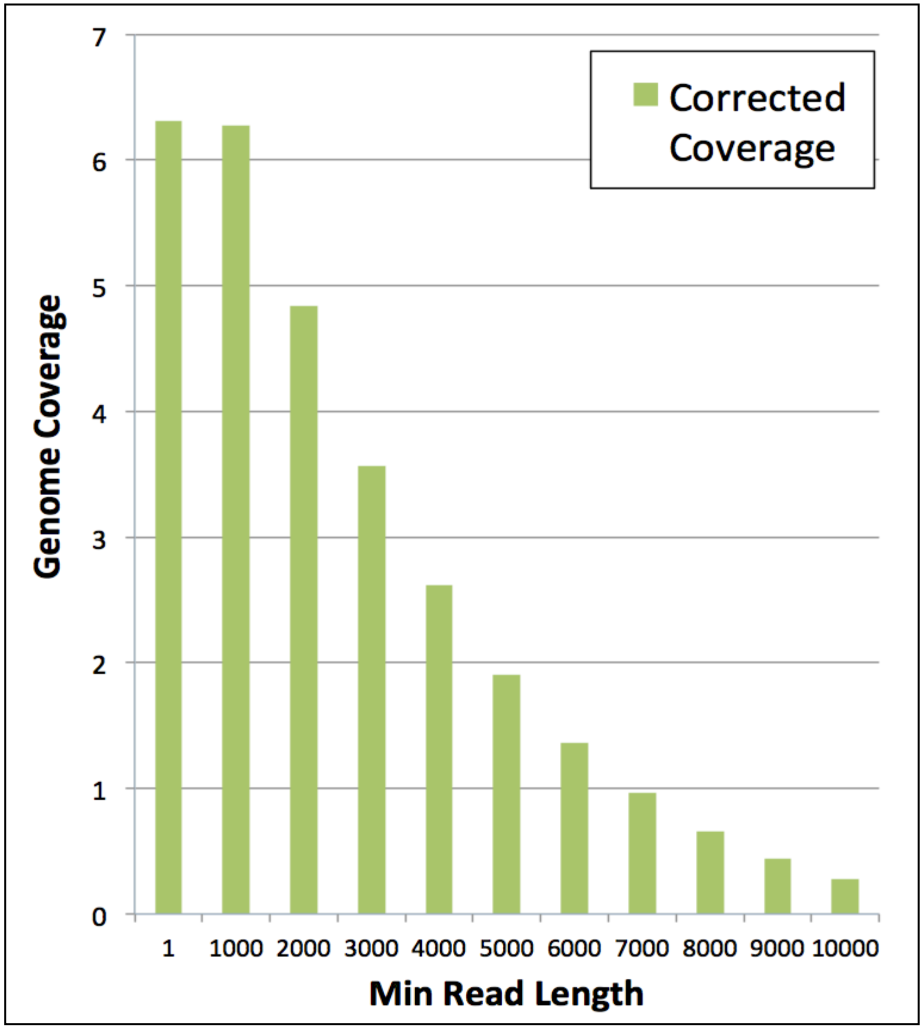
Mean=824 Max=10,008



C2XL Chemistry – Summer 2012

Mean=3290 Max=24,405

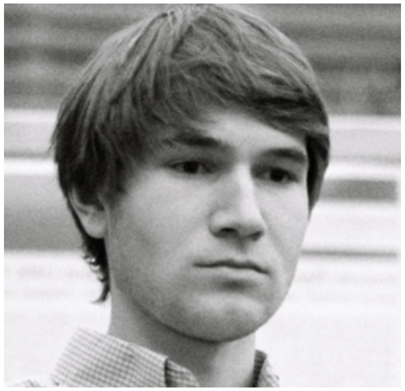
Preliminary Rice Assemblies



Assembly	Contig N50
Illumina Fragments 50x 2x100bp @ 180	3,925
MiSeq Fragments 23x 459bp 8x 2x251bp @ 450	6,444
PBeCR Reads 6.3x 2146bp ** MiSeq for correction	13,600
Illumina Mates 50x 2x100bp @ 180 36x 2x50bp @ 2100 51x 2x50bp @ 4800	13,696
PBeCR + Illumina Shred 6.3x 2146bp ** MiSeq for correction 51x 2x50bp @ 4800	25,108

In collaboration with McCombie & Ware labs @ CSHL

Other Research Projects



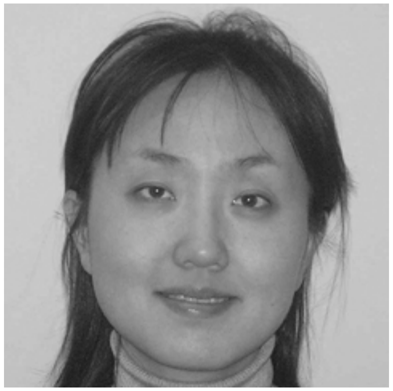
High Performance
Variant Detection
And Interpretation

>168-fold speed up
genotyping maize

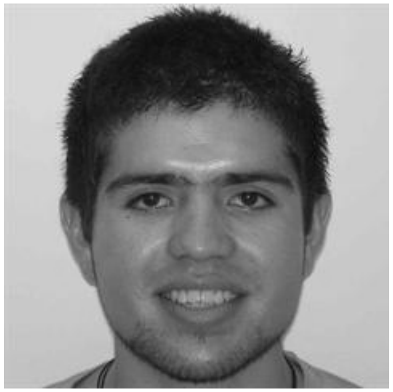
Answering the demands of digital genomics
Titmus, MA, Gurtowski, J, Schatz, MC (2012)
Concurrency and Computation: Practice and Experience

Analyzing
Genomic Repeats and
Sequencing Libraries

Pinpoint the regions
we cant sequence with
today's tech



Genomic Dark Matter
Lee, H., Schatz, M.C. (2012)
Bioinformatics. 28 (16): 2097-2105.



Merge different
assemblies into a high-accuracy consensus

Fix mistakes and capture
all the information

Improving Genome Assembly with Meta-assembly
Wences, A, Schatz, M.C. (2012)
In preparation

Evaluate the limits of
assembling human, wheat
and other genomes

How long is long
enough?

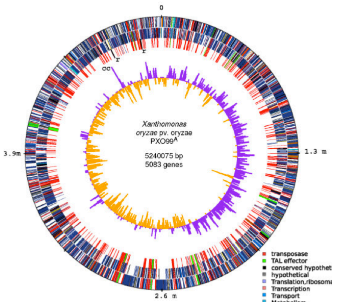
Assembly Complexity of Long Sequencing Reads
Marcus S, Lee, H., Schatz, M.C. (2012)
In preparation

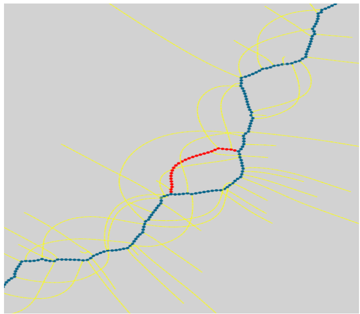
Summary

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
Alejandro Wences
Hayan Lee
Rob Aboukhalil
Mitch Bekritsky
Charles Underwood
Rushil Gupta
Avijit Gupta
Shishir Horane
Deepak Nettem
Varrun Ramani
Piyush Kansal
Eric Biggers
Aspyn Palatnick

CSHL

Hannon Lab
Iossifov Lab
Levy Lab
Lippman Lab
Lyon Lab
Martienssen Lab
McCombie Lab
Ware Lab
Wigler Lab

IT Department

NBACC

Adam Phillippy
Sergey Koren

SFARI
SIMONS FOUNDATION
AUTISM RESEARCH INITIATIVE

National Human
Genome Research
Institute

U.S. DEPARTMENT OF
ENERGY

NSF

Thank You!

http://schatzlab.cshl.edu/

@mike_schatz