Illuminating the genetics of complex human diseases

Michael Schatz

Sept 27, 2012
Beyond the Genome

Outline

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

2. Illuminating the Genomic Dark Matter
    1. Genome Mappability Score
    2. Long read single molecule sequencing

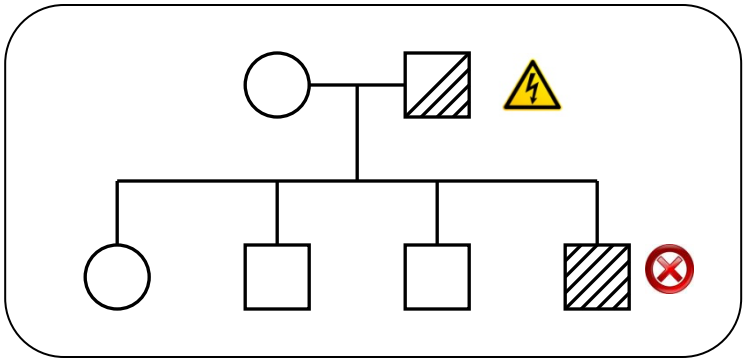
Outline

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

2. Illuminating the Genomic Dark Matter
    1. Genome Mappability Score
    2. Long read single molecule sequencing

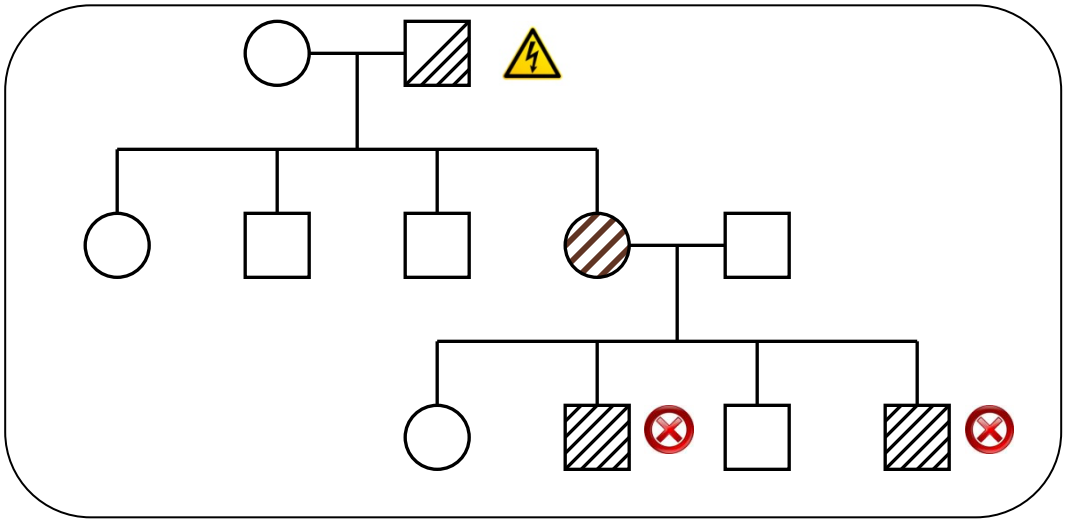
Unified Model of Autism

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



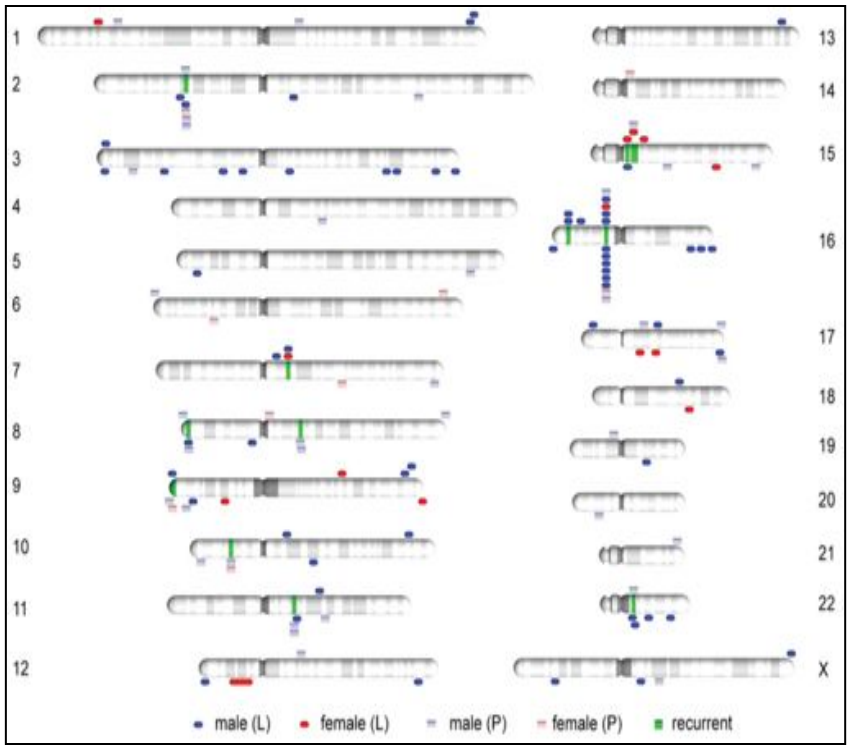
Legend

Sporadic mutation

Fails to procreate

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

Autism and de novo CNVs



Analysis of Simons Simplex Collection

- CGH arrays of 510 family quads
- 94 total de novo CNVs discovered

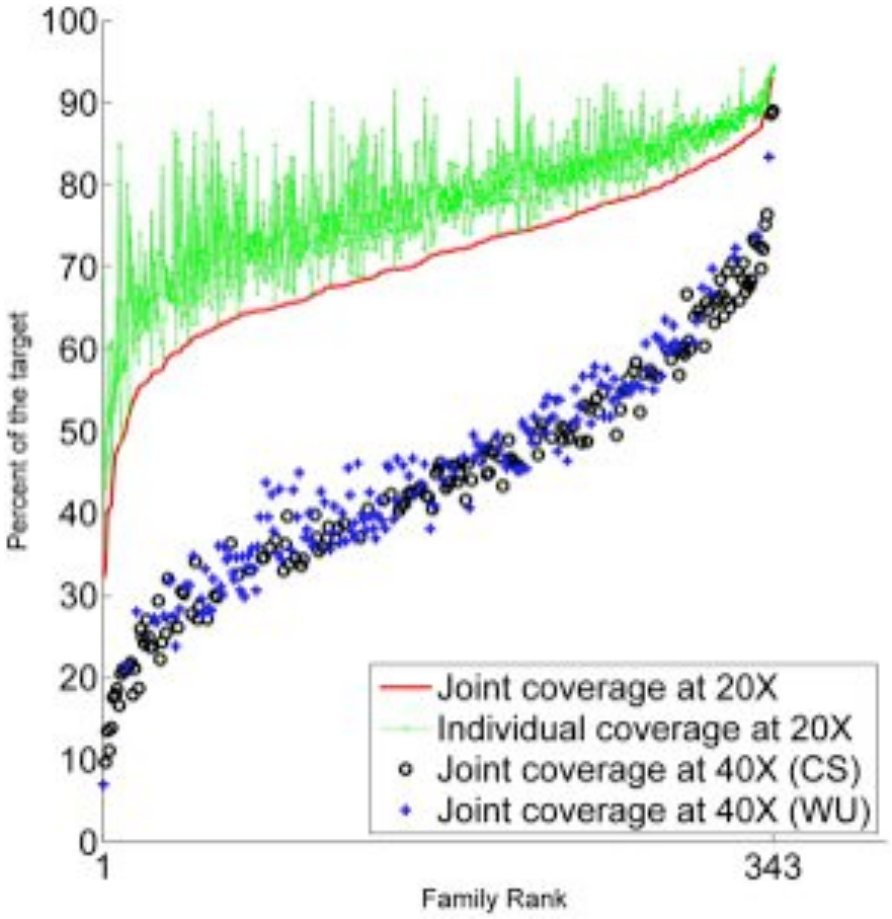
De novo CNVs are more common in autistic children

- 4:1 ratio in autistic kids relative to their non-autistic siblings
- Some recurrence at genes related to other psychiatric conditions

	Counts of De Novo Events			Children with De Novo Events			Frequency in Children		
	Combined	Del	Dup	Combined	Del	Dup	Combined	Del	Dup
aut	75	46	29	68	44	27	7.9%	5.1%	3.1%
sib	19	9	10	17	8	9	2.0%	0.9%	1.0%

Rare de novo and transmitted copy-number variation in autism spectrum disorders.
Levy *et al.* (2011) *Neuron.* 70:886-897.

Exome-Capture and Sequencing



Sequencing of 343 families from the Simons Simplex Collection

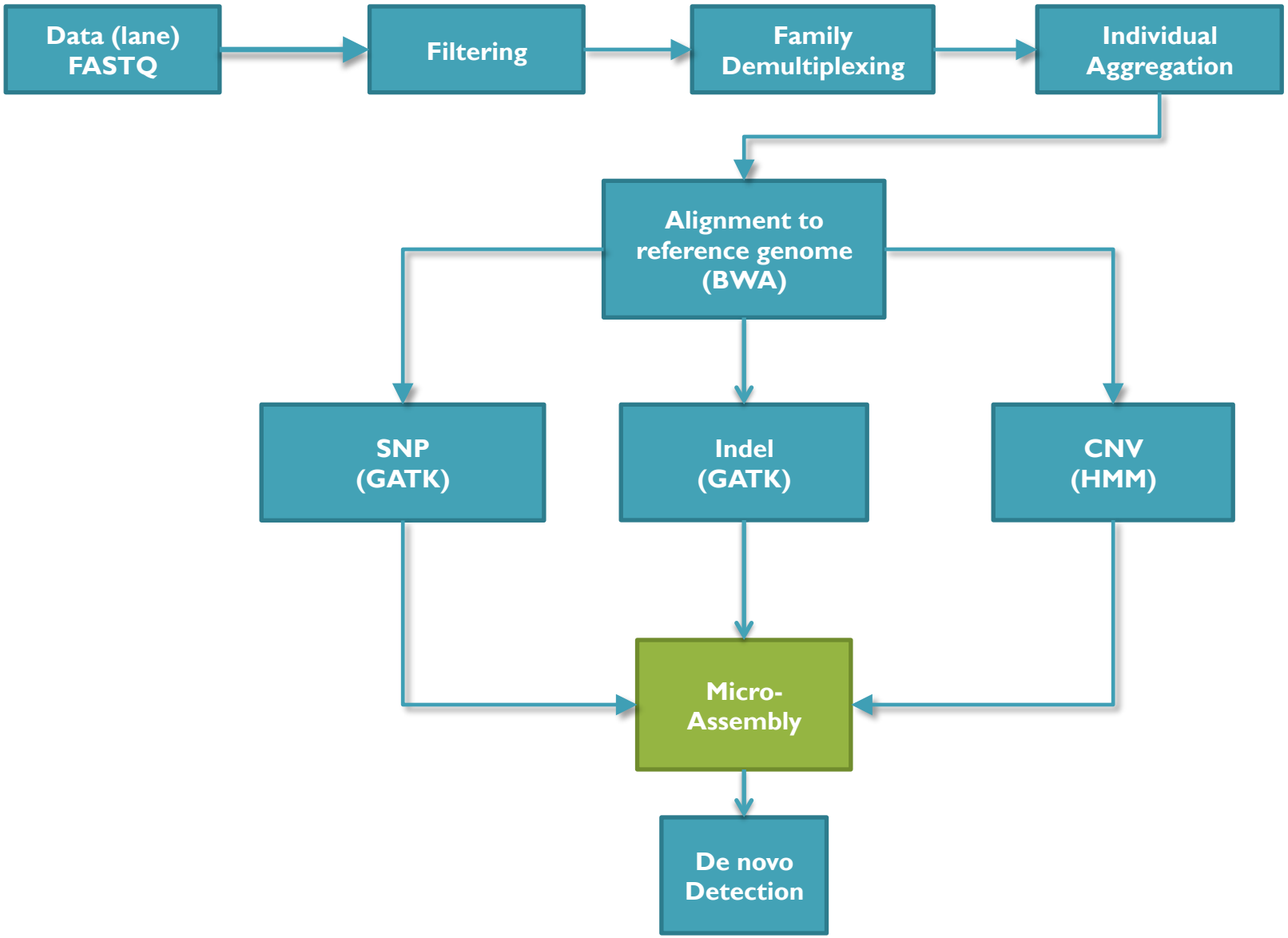
- Parents plus one child with autism and one non-autistic sibling
- Enriched for higher-functioning individuals

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

Exome Sequencing Pipeline



Variation Detection Complexity

SNPs

Seed-and-extend + scan/permute

Short indels (<3bp)

Seed-and-extend + dynamic programming

Medium indels (<15bp)

Split-read mapping w/short seeds

Long indels (15bp+)

Split-read / soft-clipped / failed map

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

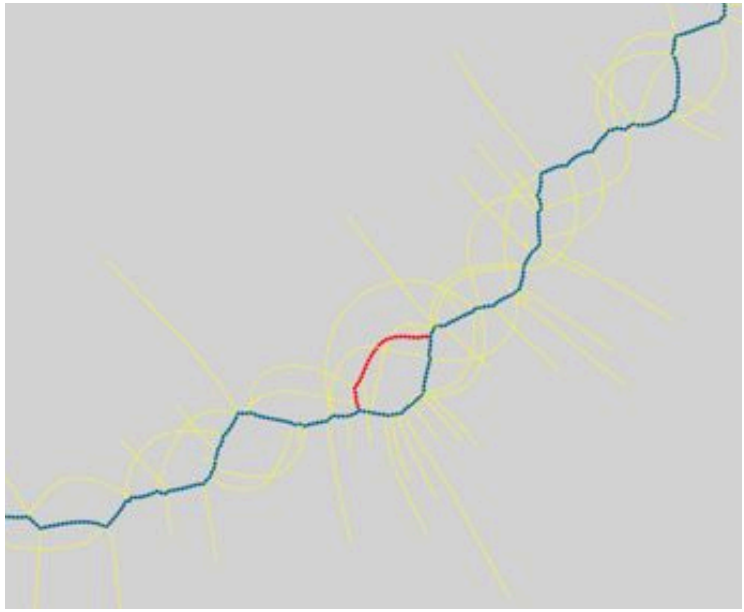
Scalpel: Haplotype Microassembly

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

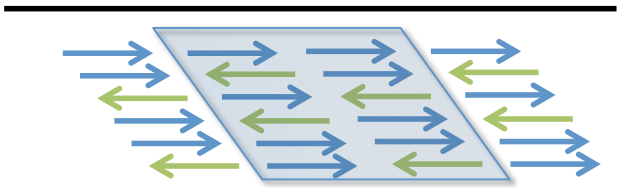
1. Combine mapping and assembly
2. Exhaustive search of haplotypes
3. De novo mutations



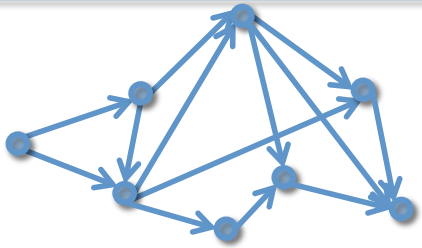
NRXN1 *de novo* SNP
(auSSC12501 chr2:50724605)

Scalpel Pipeline

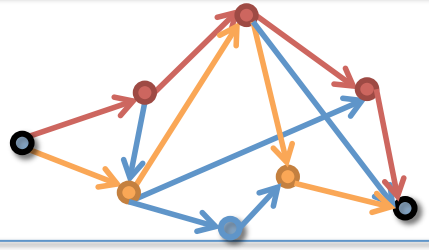
Extract reads mapping within the exon including (1) well-mapped reads, (2) soft-clipped reads, and (3) anchored pairs



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



Find end-to-end haplotype paths spanning the region

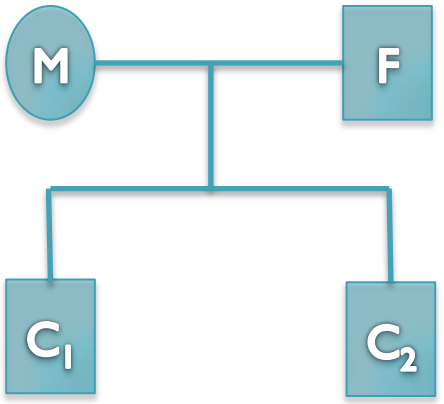


Align assembled sequences to reference to detect mutations

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------CCGGAGATTGTCTTTGCCCGGA...
```

6bp heterozygous deletion at chr13:25280526 ATP12A

De novo Genetics of Autism

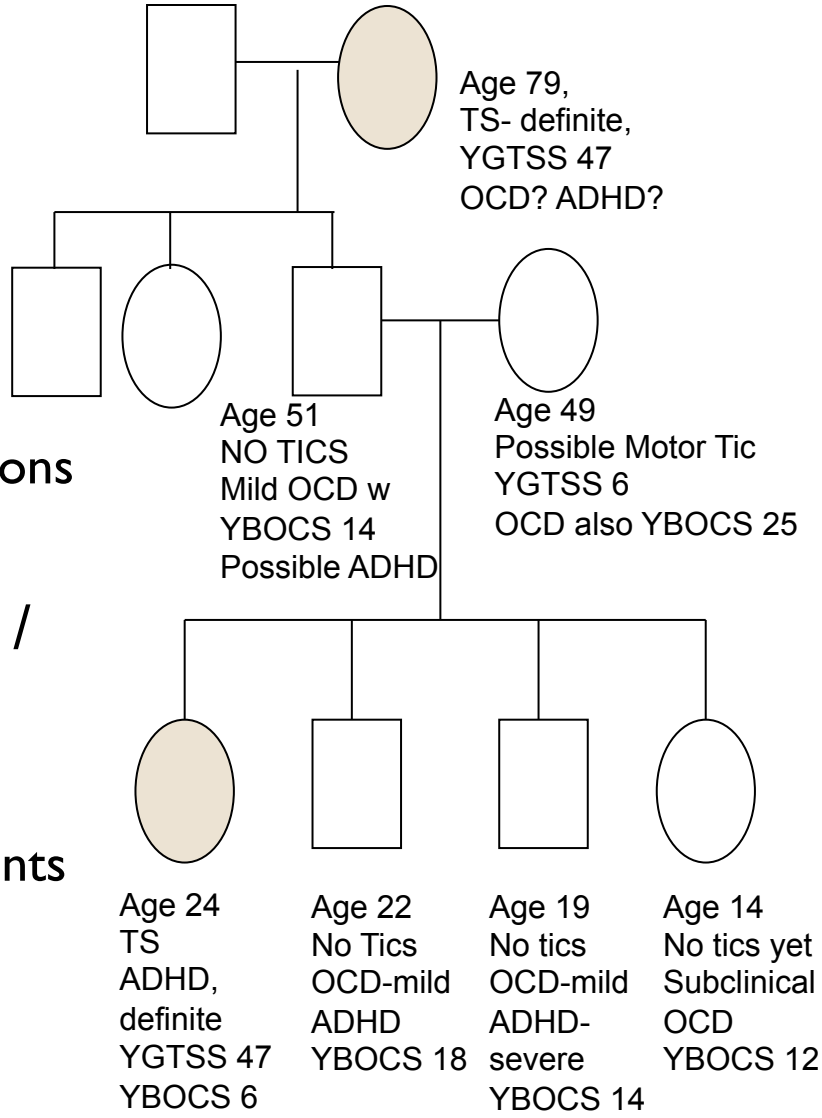
- In 343 family quads so far, we see significant enrichment in de novo ***likely gene killers*** in the autistic kids
  - Overall rate basically 1:1 (432:396)
  - 2:1 enrichment in nonsense mutations
  - 2:1 enrichment in frameshift indels
  - 4:1 enrichment in splice-site mutations
  - Most de novo 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

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



Outline

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

2. Illuminating the Genomic Dark Matter
    1. Genome Mappability Score
    2. Long read single molecule sequencing

Genomic Dark Matter

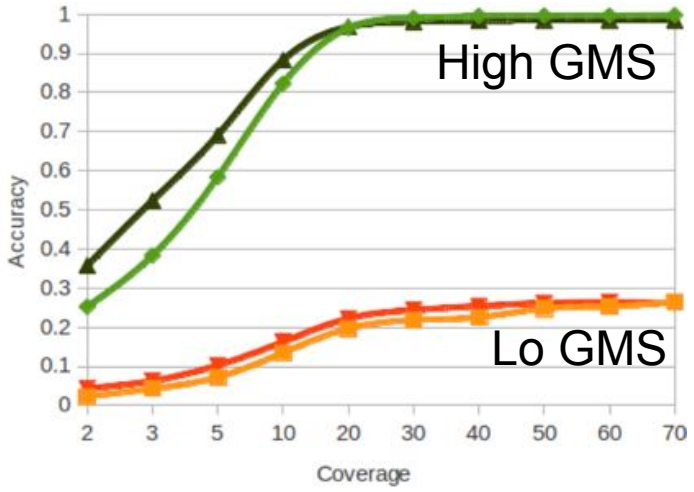
Short read mapping is a widely used for identifying mutations in the genome

- Not every base of the genome can mapped equally well, because repeats may obscure where the reads originated

Introduced a new probabilistic metric - the Genome Mappability Score - that quantifies how reliably reads can be mapped to every position in the genome

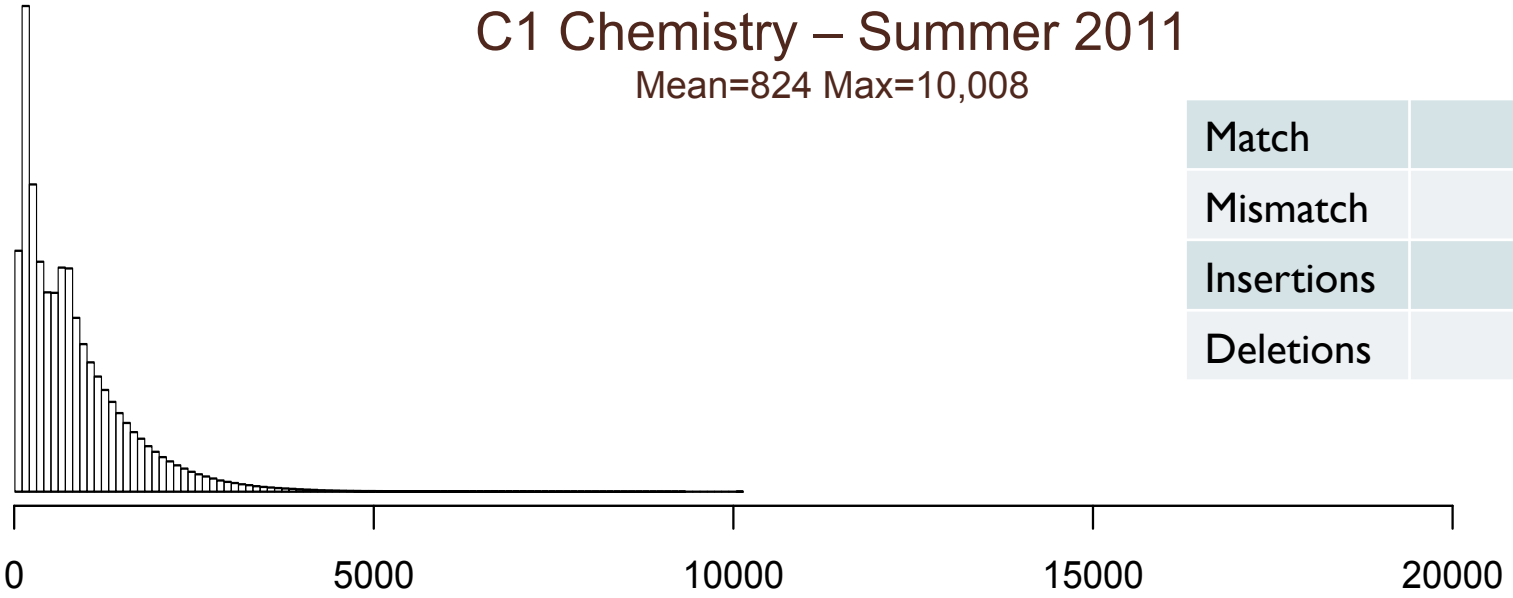
- We have little power to measure 11-13% of the human genome, including of known clinically relevant variations
- Errors in variation discovery are dominated by false negatives in low GMS regions

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

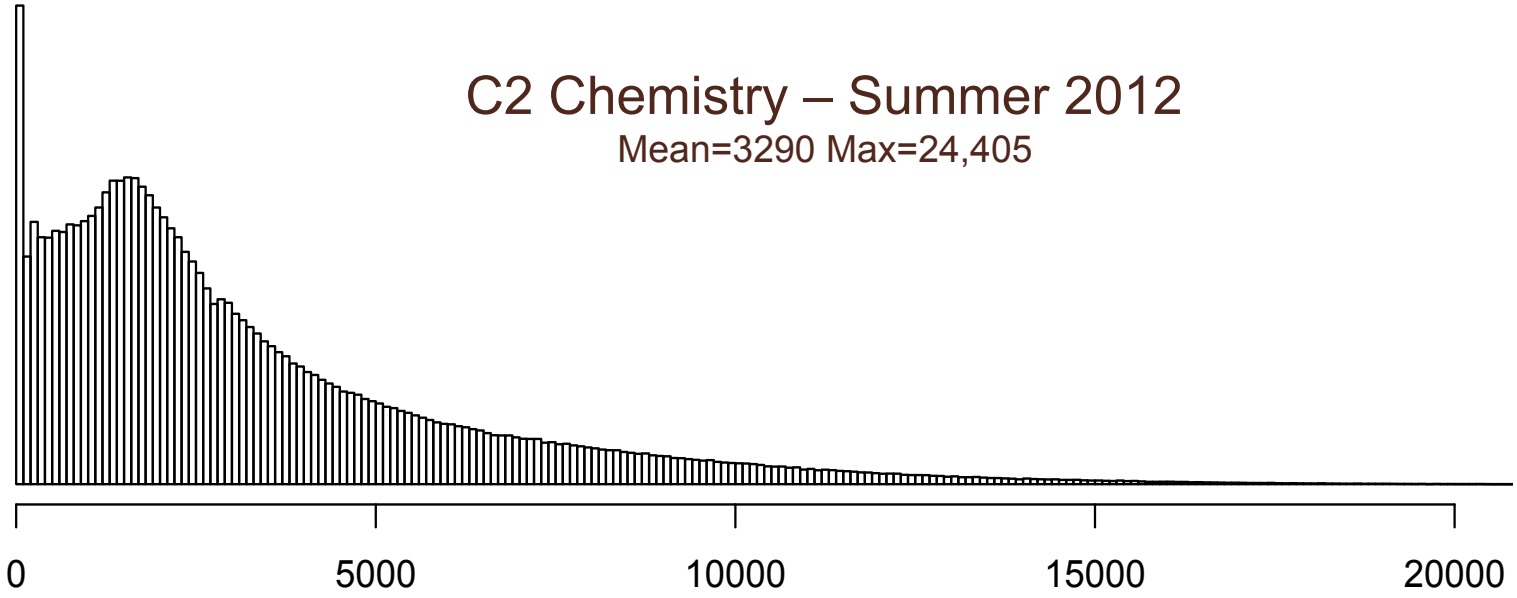


Genomic Dark Matter: The reliability of short read mapping illustrated by the GMS.
Lee, H., Schatz, M.C. (2012) *Bioinformatics*. 10.1093/bioinformatics/bts330

PacBio Long Read Sequencing



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

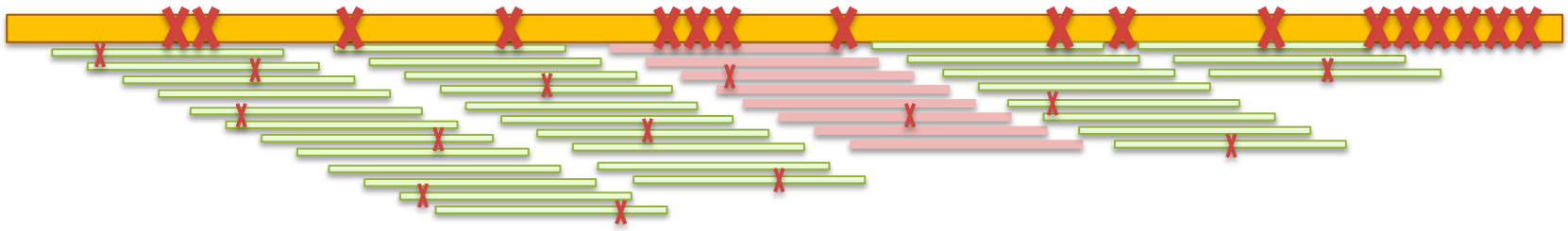


PacBio Error Correction

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

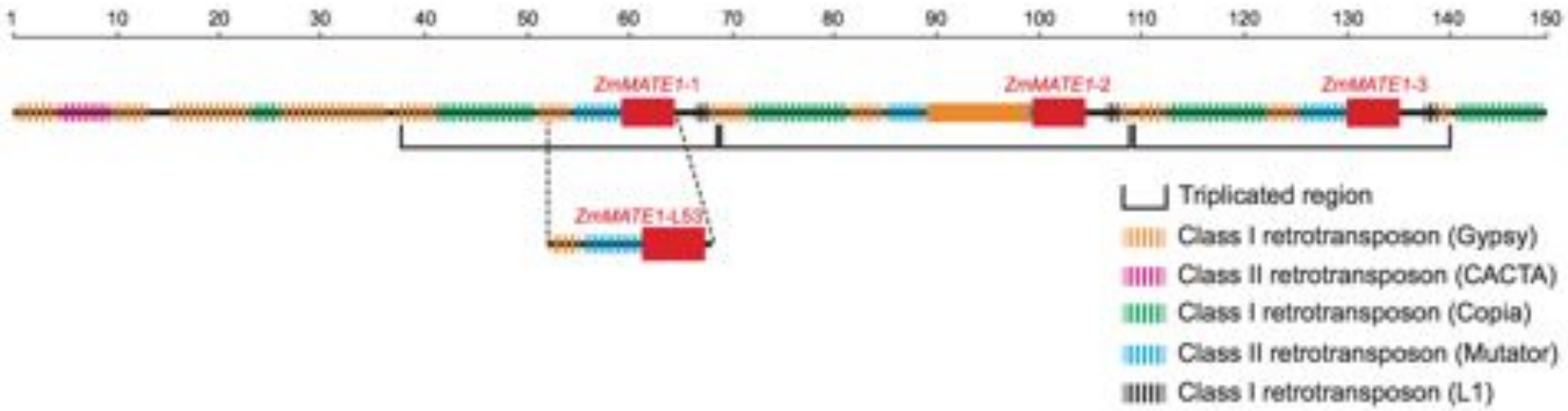
Organism	Technology	Reference bp	Assembly bp	# Contigs	Max Contig Length	N50
Lambda NEB3011	Illumina 100X 200bp	48 502	48 492	1	48 492 / 48 492	48 492 / 48 492 (100%) *
(median: 727 max: 3 280)	PacBio PBcR 25X		48 440	1	48 444 / 48 444	48 444 / 48 440 (100%) *
E. coli K12	Illumina 100X 500bp	4 639 675	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[a]		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

Long Read CNV Analysis

Aluminum tolerance in maize is important for drought resistance and protecting against nutrient deficiencies

- Segregating population localized a QTL on a BAC, but unable to genotype with Illumina sequencing because of high repeat content
- Long read PacBio sequencing revealed an additional copy of the ZnMATE1 membrane transporter and enabled assembly of the entire gene cluster



A rare gene copy-number variant that contributes to maize aluminum tolerance and adaptation to acid soils

Maron, LG *et al.* (2012) *Under review.*

Summary

Likely gene-killing de novo mutations in FMRP-related genes have a significant role in autism spectrum disorders

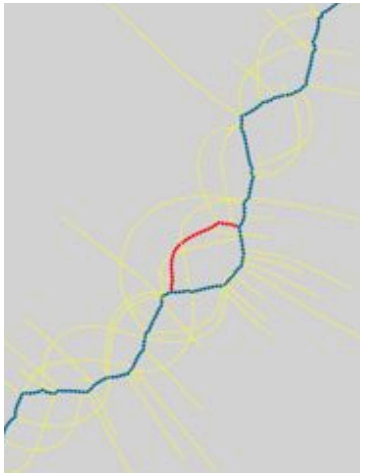
- Explains how the disorder can appear in otherwise low risk families, explains the recurrence in families, explains how development can be impaired
- Lends itself to early diagnosis and early intervention
- We suspect similar mechanisms at work in other neurological disorders
- Discovering de novo mutations requires great care – must be both highly sensitive and highly specific to overcome the noise without missing the rare events

Beware of the dark matter

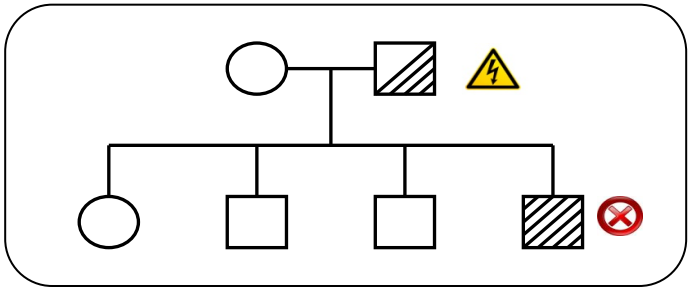
- Use the GMS to pinpoint the blind spots in your study

Exciting developments on the horizon

- Longer reads, higher throughput PacBio & Nanopore
- Cloud resources for disease and bioenergy research



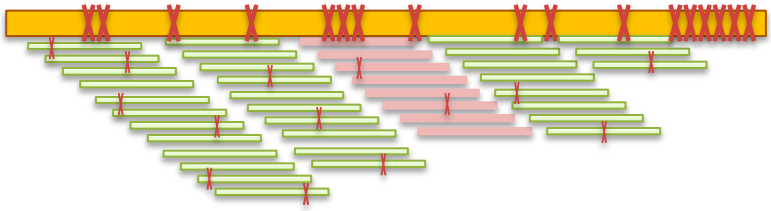
Acknowledgements



Giuseppe Narzisi

Wigler Lab
Iossifov Lab
Levy Lab
Lyon Lab

SFARI
SIMONS FOUNDATION
AUTISM RESEARCH INITIATIVE



Hayan Lee

McCombie Lab
Adam Phillippy (NBACC)
Sergey Koren (NBACC)
Lyza Maron (Cornell)

National Human
Genome Research
Institute

Thank You!

Want to push the frontier of bioinformatics, biotechnology, & genetics?

http://schatzlab.cshl.edu/apply