Next Generation Human Genetics

Jay Shendure, MD, PhD
Dept. of Genome Sciences
University of Washington

Shendure Lab

- Goal = push the envelope on technology for genomics
- No defining biological theme or model organism
- Coherence & synergies on methodological themes

Genetics

Function

Synthetics

Tagging

Contiguity

Translation

Next generation…

Genetics

Function

Synthetics

Tagging

Contiguity

Translation

Next generation…

Genetics

Function

Synthetics

Tagging

Contiguity

Translation

Next generation…

Genetics

Function

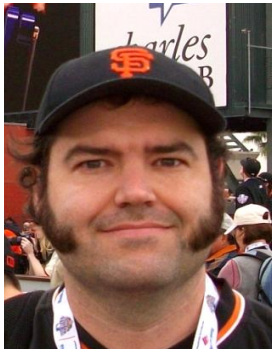
Synthetics

Tagging

Contiguity

Translation

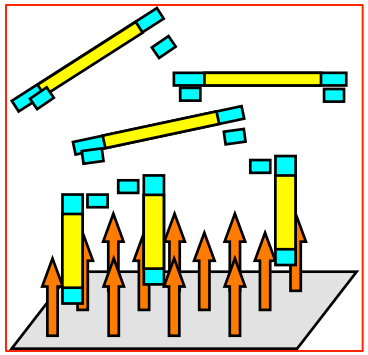
Sarah Ng

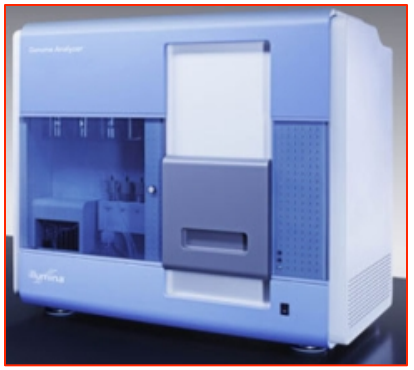


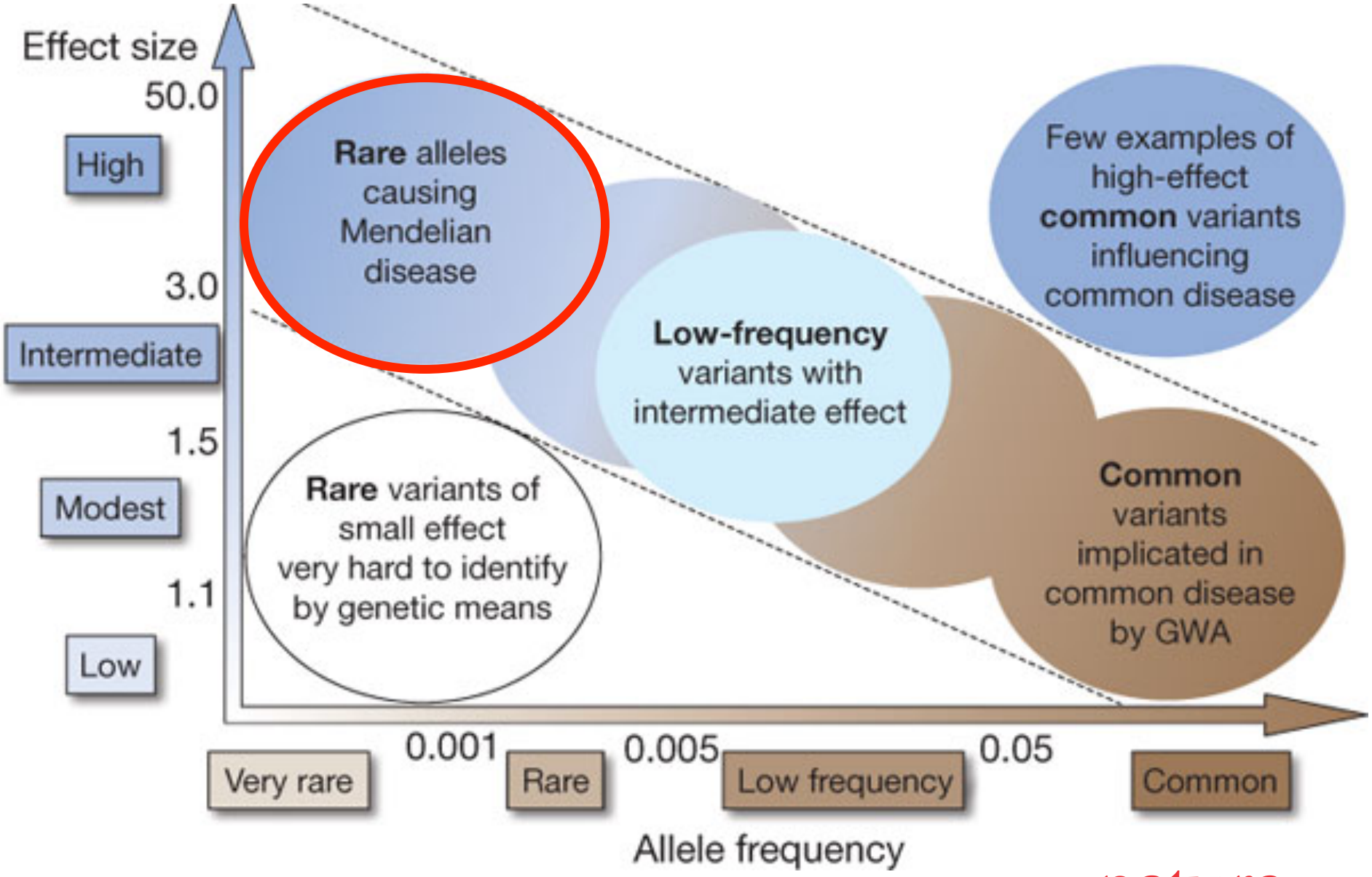
Brian O'Roak

Rewind to 2008…

- Cost of human genome = **~$250,000**
- Technical goal = **exome sequencing** (~1%)
- Still not entirely clear what human genome (or exome) sequencing would be useful for.
- How do you learn anything from a small sample size?







Manolio *et al.* (2009)

nature

Mendelian disorders

- Rare, single gene diseases
- Monogenic subsets of common diseases

- **>2,000 solved**
- **>2,000 unsolved**

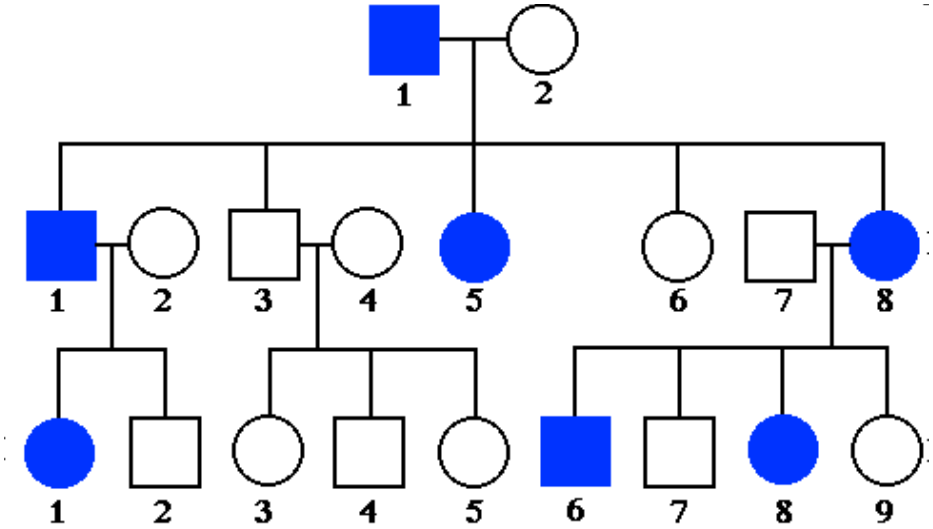


Table 2. SNPs Identified through Whole-Genome Sequencing of DNA from the Proband.*

SNP Type	No. of SNPs
Nongene	2,255,102
Gene	1,165,204
Intron	1,064,655
Promoter	60,075
3′ UTR	16,350
5′ UTR	3,517
Splice regulatory site	2,089
Splice site	112
Synonymous	9,337
Stop→stop	17
Nonsynonymous	9,069
Stop→gain	121
Stop→loss	27
Total	3,420,306

How does one pinpoint **causal** variants in human genome(s)?

Lupski et al.
NEJM (2010)

nature Vol 461|10 September 2009|doi:10.1038/nature08250

Targeted capture and massively parallel sequencing of 12 human exomes

Sarah B. Ng[1], Emily H. Turner[1], Peggy D. Robertson[1], Steven D. Flygare[1], Abigail W. Bigham[2], Choli Lee[1], Tristan Shaffer[1], Michelle Wong[1], Arindam Bhattacharjee[4], Evan E. Eichler[1,3], Michael Bamshad[2], Deborah A. Nickerson[1] & Jay Shendure[1]

1. Sequence exomes of four unrelated probands affected with a dominant Mendelian disorder

2. Remove "common" variants

- 8 HapMap 'control' exomes
- public SNP databases

3. Find genes that contain "uncommon" variants in *all* affected individuals

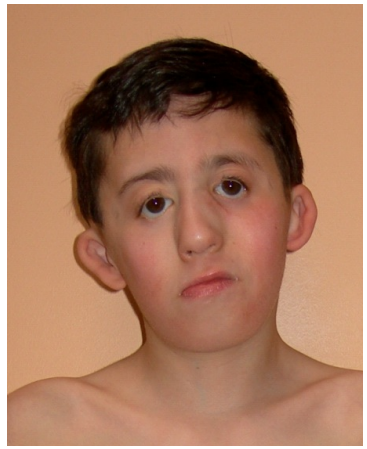
Ng *et al.* Nature (2009)

Can exome sequencing be applied to identify the cause of a Mendelian disorder?

How many genes in genome with….	Freeman-Sheldon probands			
	1/1	2/2	3/3	4/4
…any nsSNP, splice-site, or indel	4,510	3,284	2,765	2,479
…not in dbSNP	513	128	71	53
…not in 8 HapMap exomes	799	168	53	21
…not in either dbSNP or 8 HapMap exomes	360	38	8	***1 MYH3***

Ng *et al.* Nature (2009)

Miller syndrome



Postaxial acrofacial dysostosis

Presumed **autosomal recessive**

Sequenced **4 exomes** from **3 kindreds**

2 siblings from kindred #1 also had CF-like lung phenotype (but not in other Miller cases)

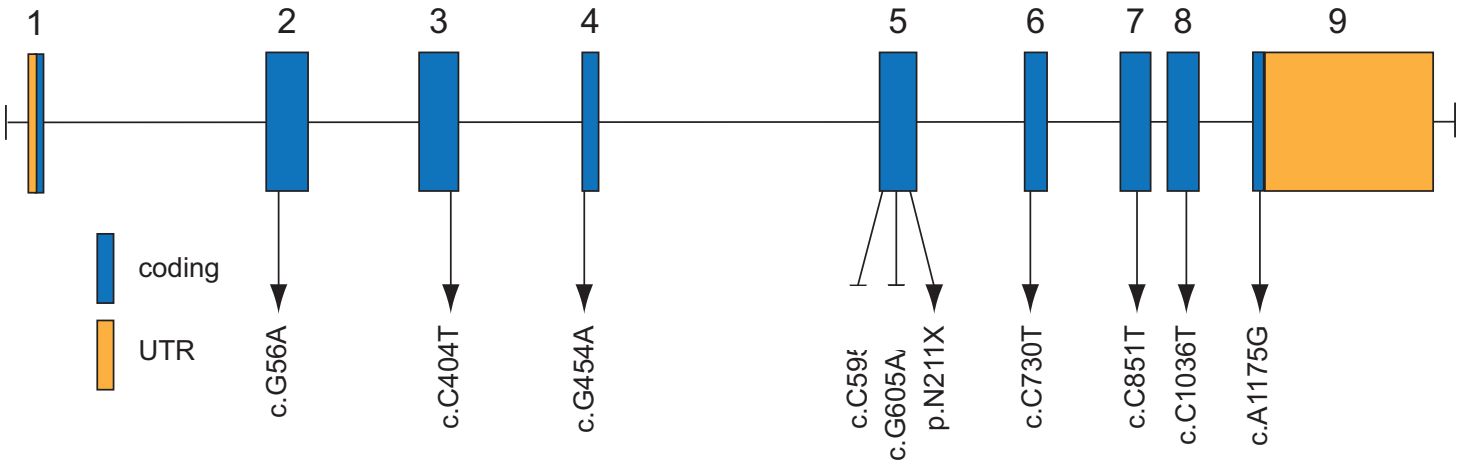
Ng *et al.* Nature Genetics (2010)

Exome analysis of Miller syndrome

How many genes in genome with…	1-A	1-B	1-A+B	2/2	3/3
…any **2** nsSNP, splice-site, or indel	2,789	2,777	2,278	1,740	1,461
…not in dbSNP	93	80	40	16	12
…not in 8 HapMap exomes	127	120	47	8	3
…not in either dbSNP or 8 HapMap exomes	31	26	8	***1 DHODH***	***1 DHODH***

Ng *et al.* Nature Genetics (2010)

DHODH = dihydroorotate dehydrogenase



de novo pyrimidine biosynthesis

inborn error of metabolism

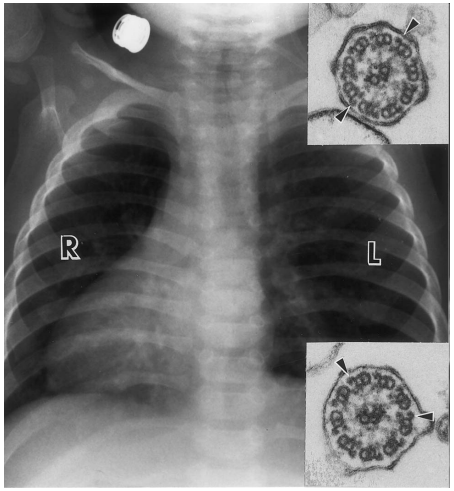
same pathway as *rudimentary*

What about the lung phenotype?

Sibling kindred (only)

Chronic sinopulmonary infections (CF-like)

Compound heterozygotes for rare, damaging mutations in **DNAH5**

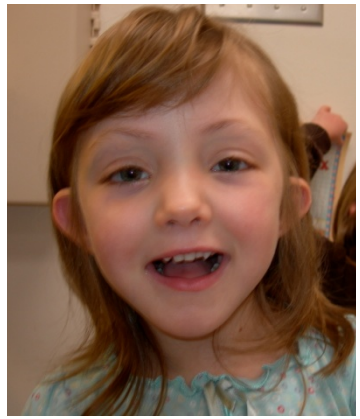


Primary cilliary dyskinesia

Ng *et al.* Nature Genetics (2010)

It’s not always that easy…

- **Kabuki syndrome**
- Multiple malformation syndrome
- Suspected **dominant**
- Originally described by Niikawa (1981)
- Sequenced exomes of 10 probands



Ng *et al.* Nature Genetics (2010)

Exome analysis of Kabuki syndrome

How many genes in genome with…	5/5	6/6	7/7	8/8	9/9	10/10
…any nsSNP, splice-site, or indel	2,028	1,899	1,772	1,596	1,524	1,459
…not in dbSNP or 1000 Genomes	63	54	42	36	35	34
…not in 26 control exomes	44	41	35	5	5	4
…not in either dbSNP, 1000G, or controls	14	11	6	1	1	1

MUC16 = artifact!

Why didn’t this work?

Genetic **heterogeneity**?

Undercalling of coding variants?

Causal **non-coding** or **structural** variants?

More than one of the above?

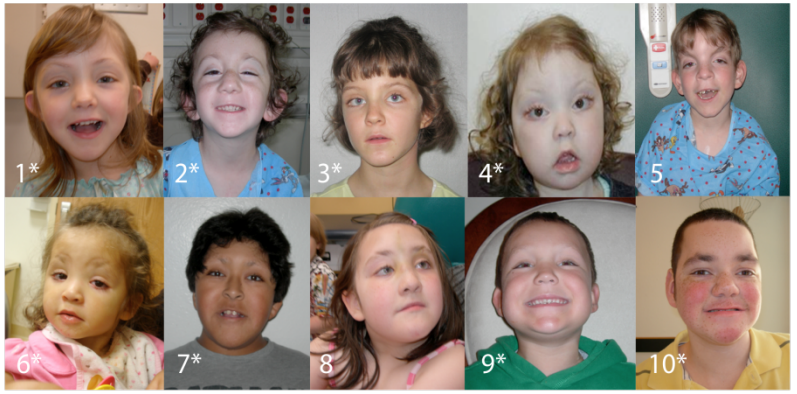
Exome analysis of Kabuki syndrome

How many genes in genome with…	5/10	6/10	7/10	8/10	9/10	10/10
…any nsSNP, splice-site, or indel	5,289	4,581	3,940	3,244	2,486	1,459
…not in dbSNP or 1000 Genomes	288	192	128	88	60	34
…not in 26 control exomes	311	192	124	77	36	4
…not in either dbSNP, 1000G, or controls	140	69	31	14	5	1

Ng *et al.* Nature Genetics (2010)

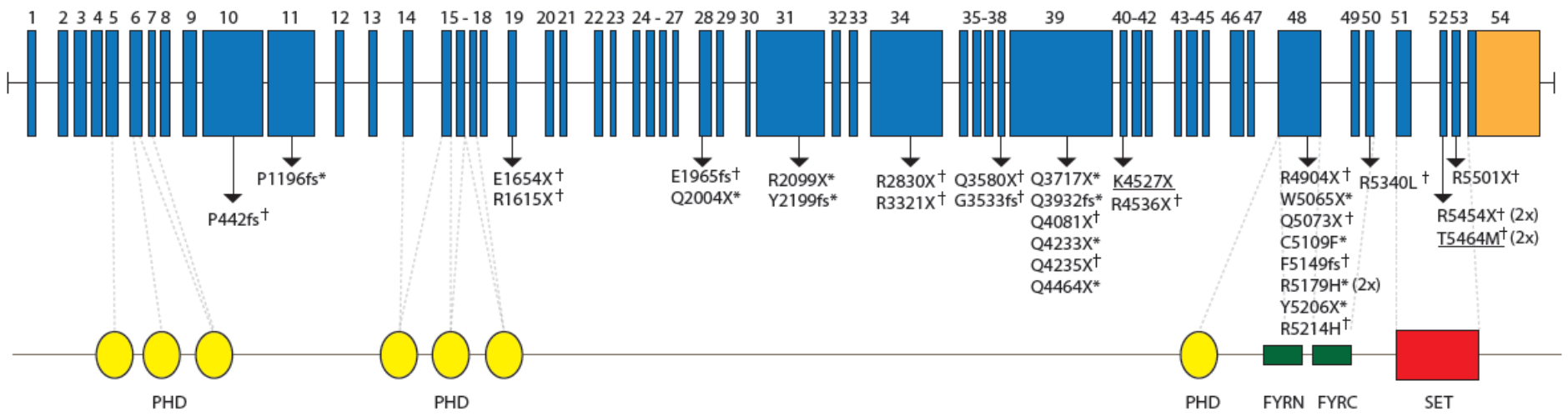
Exome analysis of Kabuki syndrome

phenotypic stratification



genotypic stratification

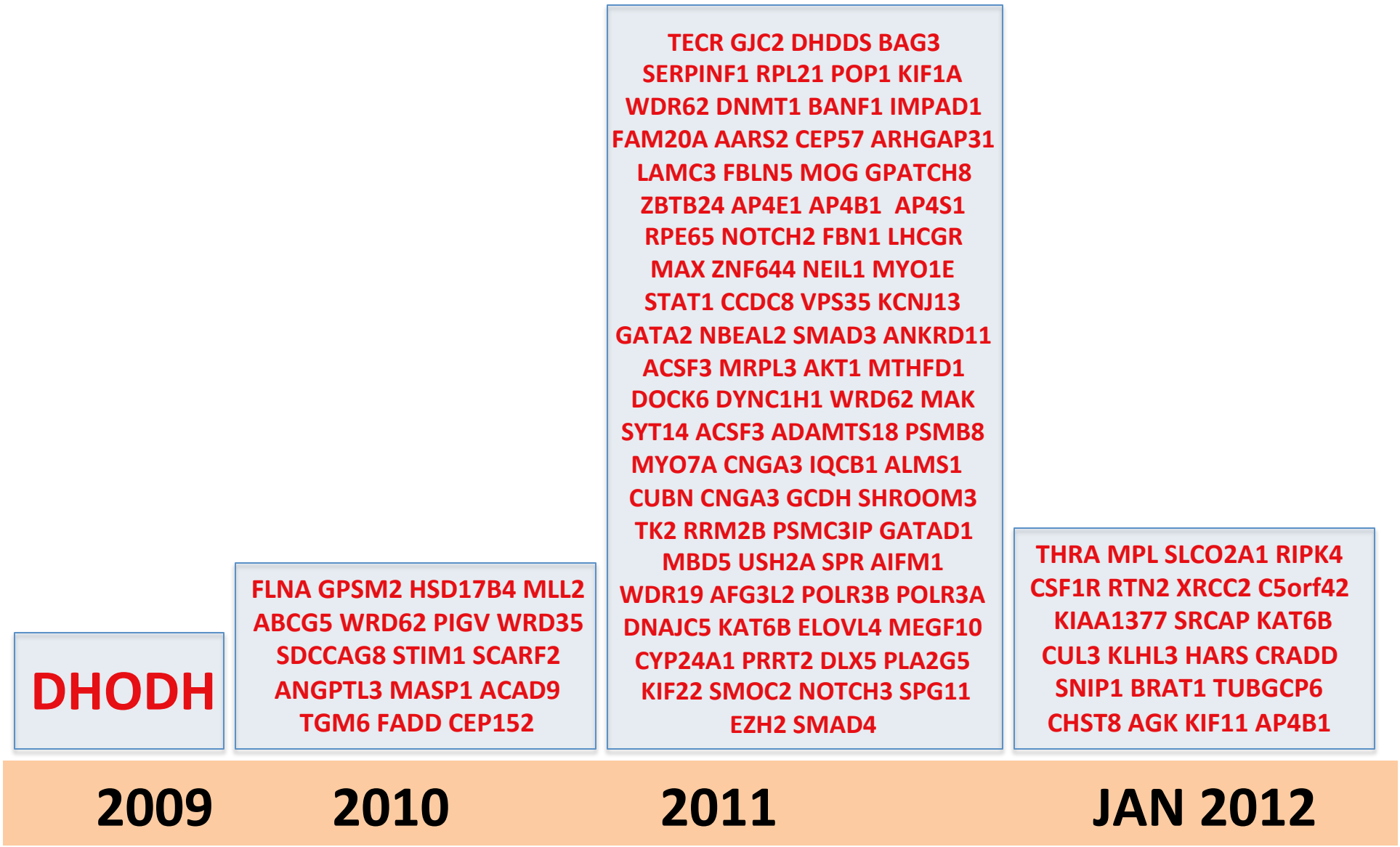
nonsense > missense
conserved > nonconserved



- **MLL2** = trithorax-group histone methyltransferase
- 7 of 10 probands w/ loss-of-function mutation
- 81 mutations in 110 kindreds (74%)
- *de novo*, mostly trunctuating
- **Couldn't have been solved by mapping!**

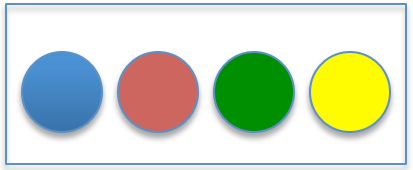
Ng *et al.* Nature Genetics (2010)

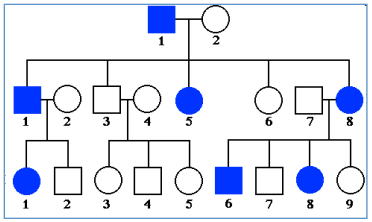
>100 Mendelian disease genes by exome

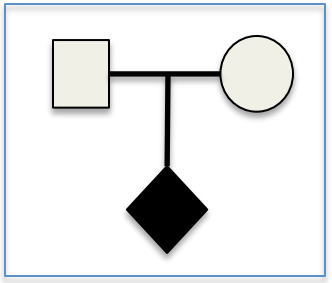


What generalizations can we make?

1) Solving Mendelian disorders is a powerful application of exome/genome sequencing







2) Only the successes get published!

Incomplete Information

Genetic Heterogeneity

Centers for Mendelian Genomics

Goals are to:

1) Identify the genetic basis for as many Mendelian disorders as possible
2) Develop and disseminate improved methods
3) Create public resources that facilitate coordinated discovery activities

University of Washington
Center for Mendelian Genomics
(coordinating center)

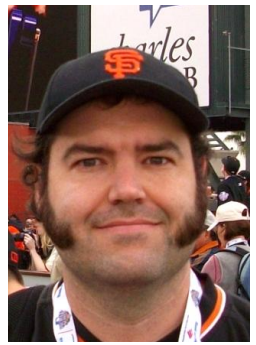
Yale Center for Mendelian Disorders

Baylor-Johns Hopkins Center for Mendelian Genetics

http://www.mendelian.org gmendel@mendelian.org

Autism Spectrum Disorders (ASD)

Complex phenotype w/ complex genetics

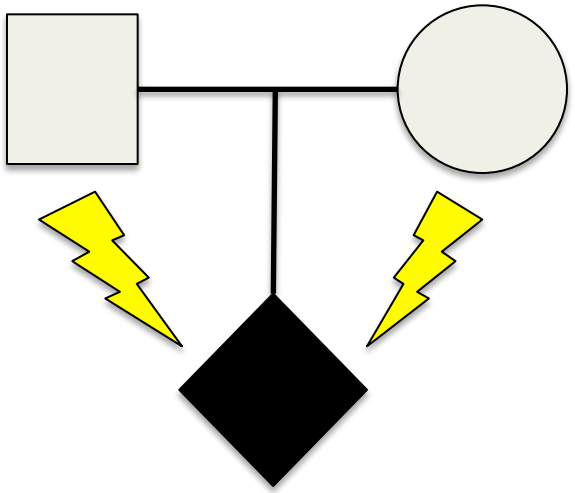


<u>Simplifying assumptions</u>

1) Coding mutations w/ large effects likely contribute to autism
2) Some % of these are expected to be ***de novo*** point mutations
3) But probably 100's of different genes

Trio Sequencing in Simplex Autism

De novo mutations in parent-child trios



Simplex pedigrees (Simons)

209 trios (627 exomes)

Solution hybridization capture (NimbleGenEZ ExomeV2.0) & Illumina sequencing (~90% of ~30 Mb target @ >=8x)

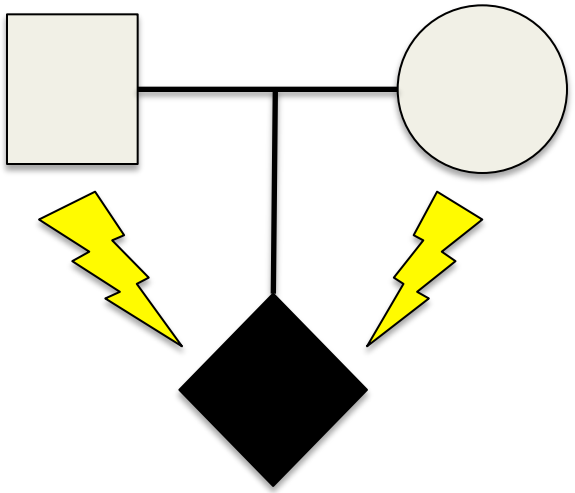
O'Roak *et al.* Nature (2012)

de novo SNV analysis: "Haystack"

1. Examine bases called in full trio
2. Find Mendelian errors with discordant proband
3. Filter against 2,000+ other exomes
4. Annotate
5. Manual review
6. Sanger confirmation

Trio Sequencing in Simplex Autism

242 new *de novo* variants confirmed

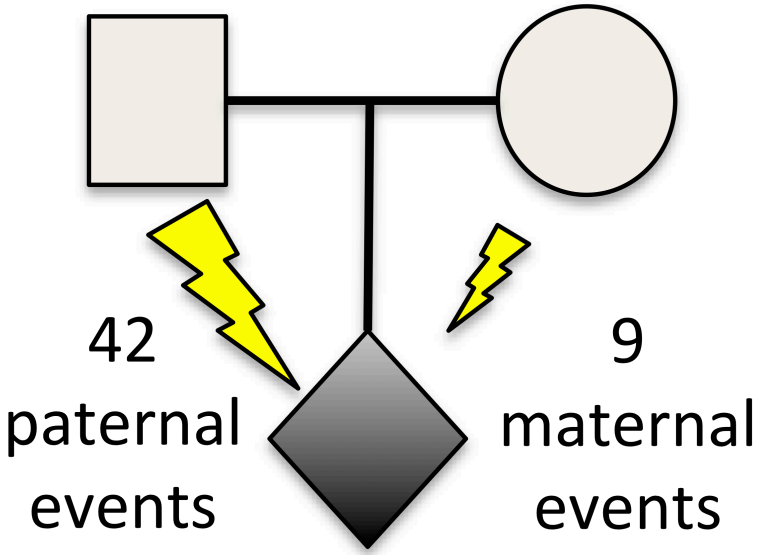


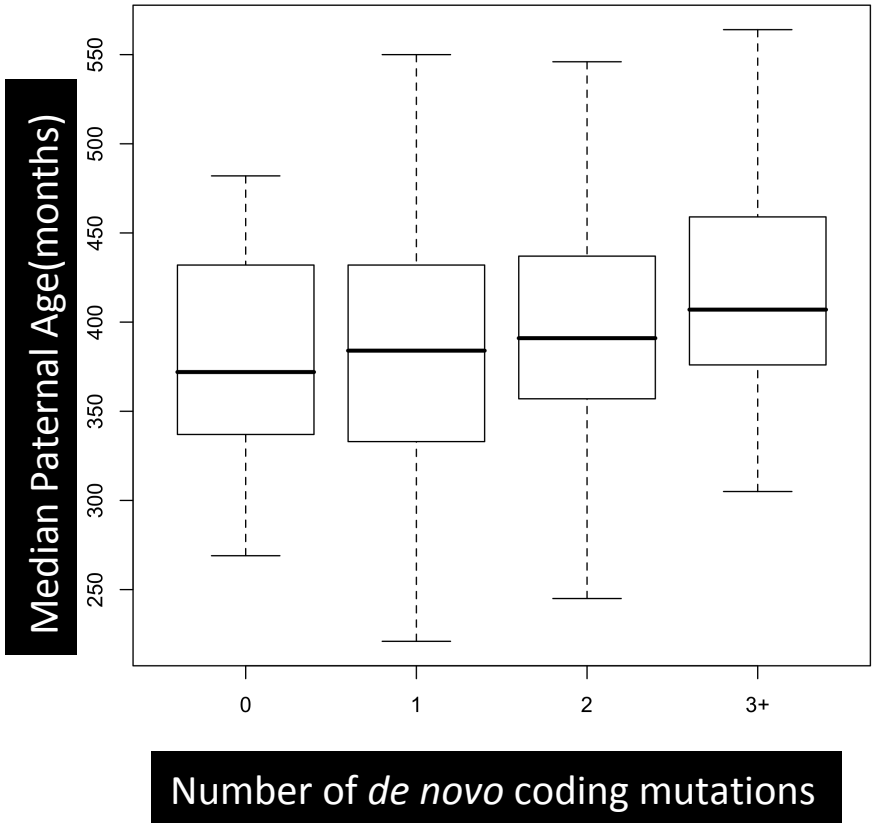
Exome $\mu = 2.17 \times 10^{-8}$ per generation

225 substitutions (Ti/Tv= 2.6; 36% CpG)
181 non-synonymous (73%)
17 indels (15 truncating)

11 examples of parental or somatic **mosaicism**

Strong parent-of-origin bias



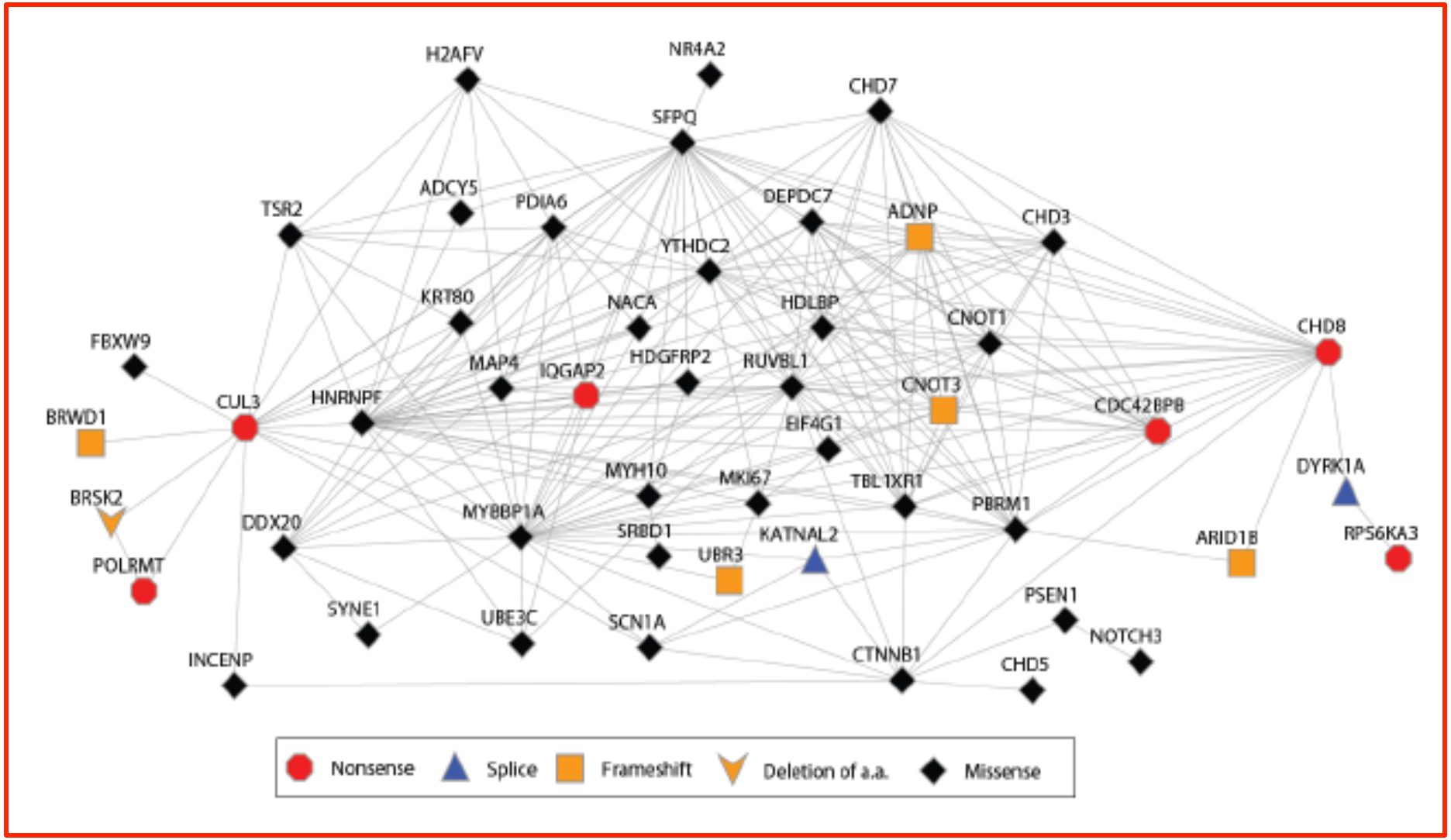


Haldane (1935) was right…

Extreme Genetic Heterogeneity

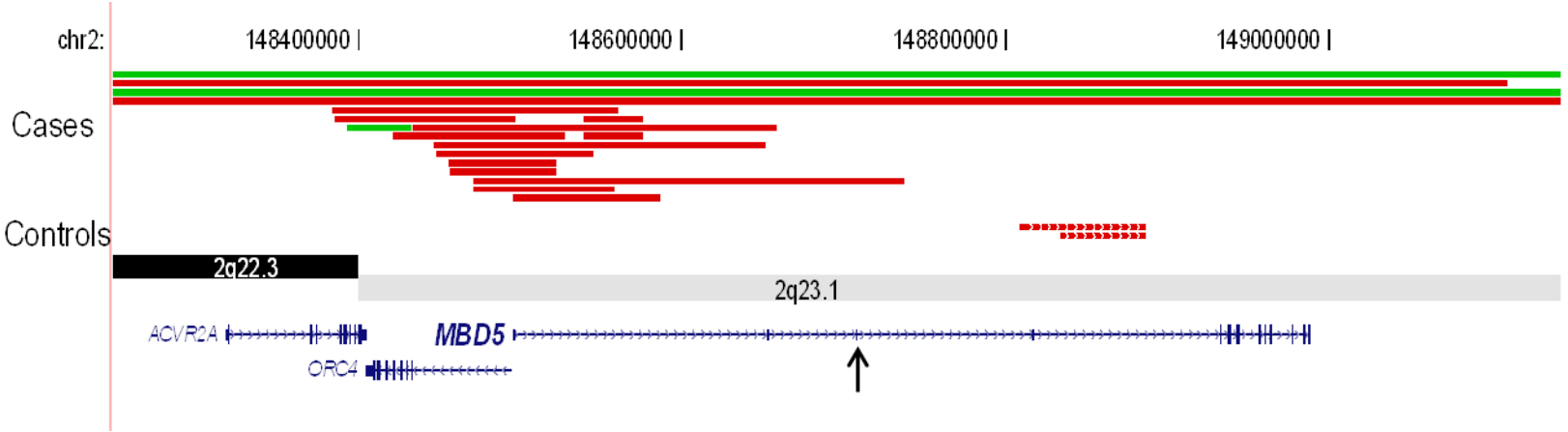
- 209 trios (627 exomes)
- ~200 nonsynonymous *de novo* mutations
- Many intriguing candidate genes for autism
  (e.g. GRIN2B, LAMC3, FOXP1, SCN1A)
- **But only 2 recurrently disrupted genes!**
- Where do we go from here?

A protein-protein network



A subset of mutations pinpoint CNV candidate genes

- 2q23.1 Microdeletion Syndrome-*MBD5*

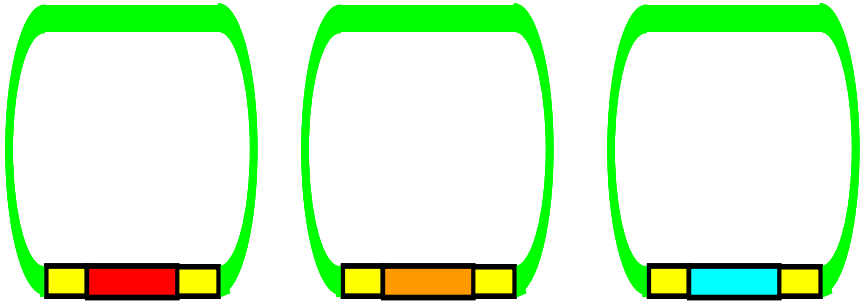


Williams et al. 2009, EJHG; Talkowski et al. 2011 AJHG

- del(18)(q12.2q21.1) syndrome-*SETBP1*
- Down Syndrome critical region-*DYRK1A*

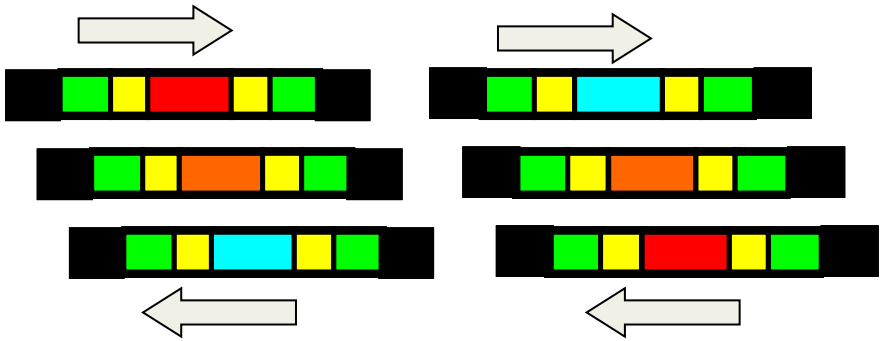
Molecular Inversion Probes

Multiplex sequence capture



Inverse PCR with common primers that add Illumina adaptors

Direct sequencing with 76 base-pair reads from both directions

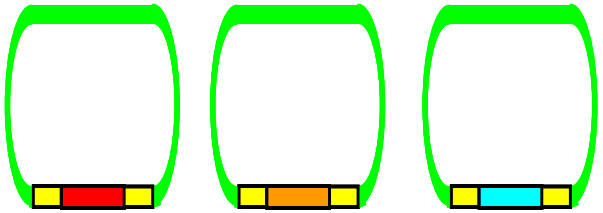


Turner *et al.* Nature Methods (2009)

Cheap Deep Resequencing

Target = **6 genes** (autism candidates)

Low input (**50 nanograms**)



1 capture + PCR per sample → pool

192-plex or **384-plex** per HiSeq2000 lane

1703 probands and 733 controls (**n=2,436**)

Cheap Deep Resequencing

- >95% of target @ >50x coverage (**192-plex**)
- Both sensitive (99%) & specific (1% FP)
- 2+ *de novo* in 3 genes (in 1,703 probands)

Gene	Proband	Type	Chrom	Pos (hg19)	Genotype
GRIN2B	12681	splice	12	13722953	**Y**
GRIN2B	12547	nonsense	12	13764762	**Y**
GRIN2B	11691	frameshift	12	14019043	**+G/***
LAMC3	11666	missense	9	133914290	**R**
LAMC3	12432	missense	9	133914556	**R**
LAMC3	11704	missense	9	133952690	**R**
SCN1A	12499	missense	2	166848071	**R**
SCN1A	12340	missense	2	166848006	**Y**

Trio Sequencing in Simplex Autism

- Extreme Locus Heterogeneity - No single gene causes >0.5% of simplex autism
- Small & large *de novo* mutations clearly relevant
- **Identifying *de novo* mutations in simplex parent-child trios will yield candidate genes**
- **However, replication in large numbers of cases and controls will be key**

Next generation…

Genetics

Function

Synthetics

Tagging

Contiguity

Translation

Jacob Kitzman

Incomplete Genomics

- Short reads align ambiguously to ~5% of genome
- Short reads poorly detect many structural variants
- Gaps and errors in the reference genome
- Next-generation ‘whole genomes’ are essentially blind with respect to phase / haplotype

Incomplete Genomics

- Short reads align ambiguously to ~5% of genome
- Short reads poorly detect many structural variants
- Gaps and errors in the reference genome
- Next-generation ‘whole genomes’ are essentially blind with respect to **phase / haplotype**

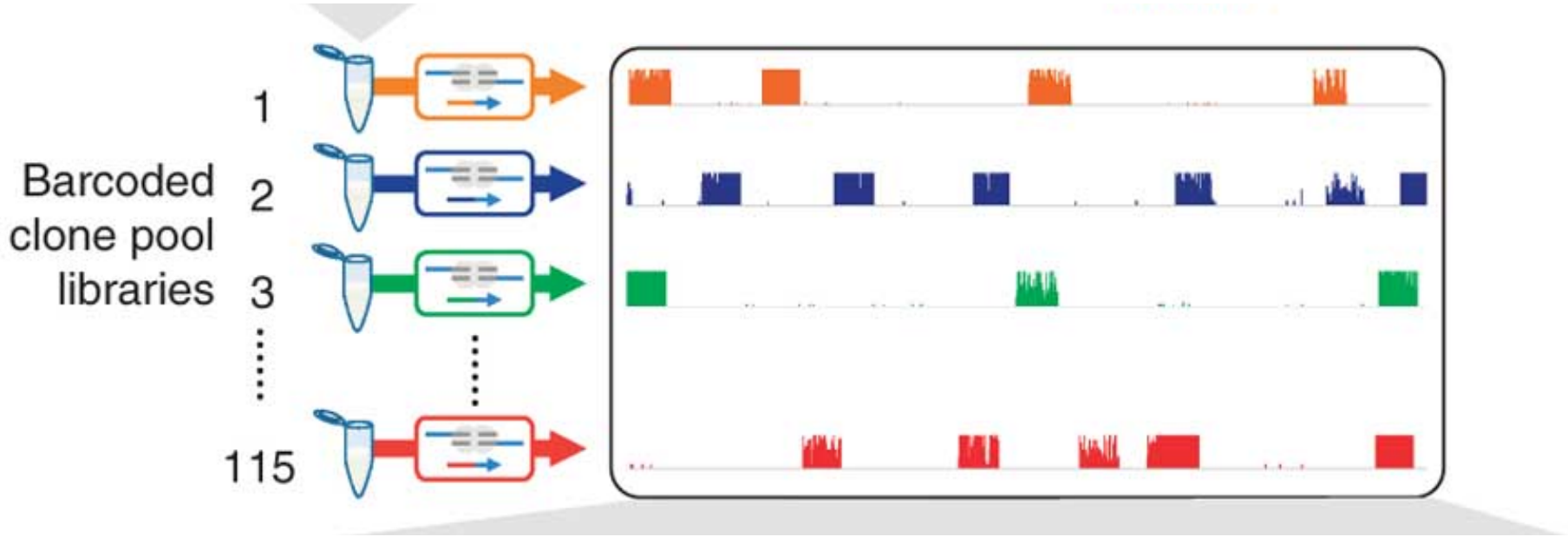
Kitzman *et al.* **Nature Biotechnology** (2011)

Haplotype Resolved Genomes

1. Unphased genome sequencing (NA20847 → 15×)

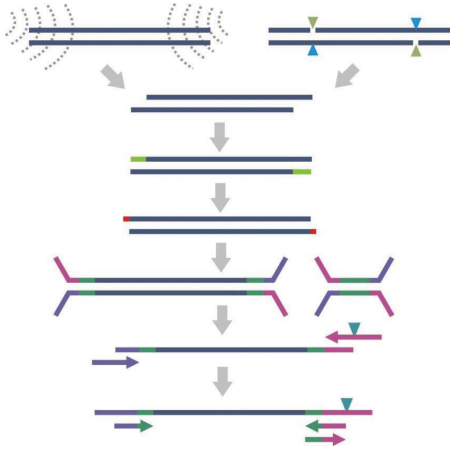
2. Clone gDNA to a single **fosmid** library

3. Split fosmid library into ~100 **subpools**, each w/ ~3% of diploid genome, and sequence separately

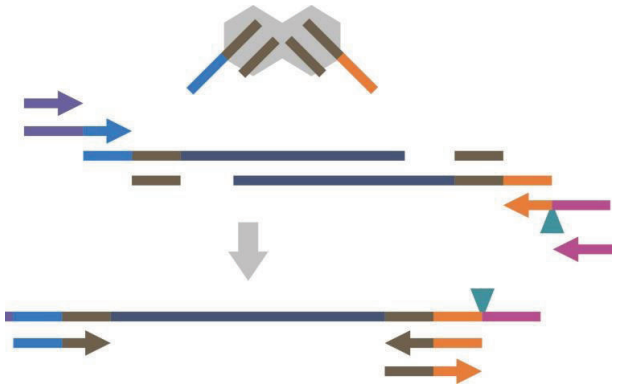


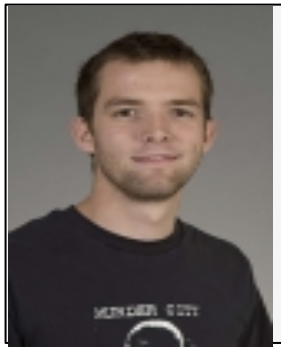
Making 96, 192, … indexed libraries quickly

Standard
(2 days)



Transposase
(2 hours)



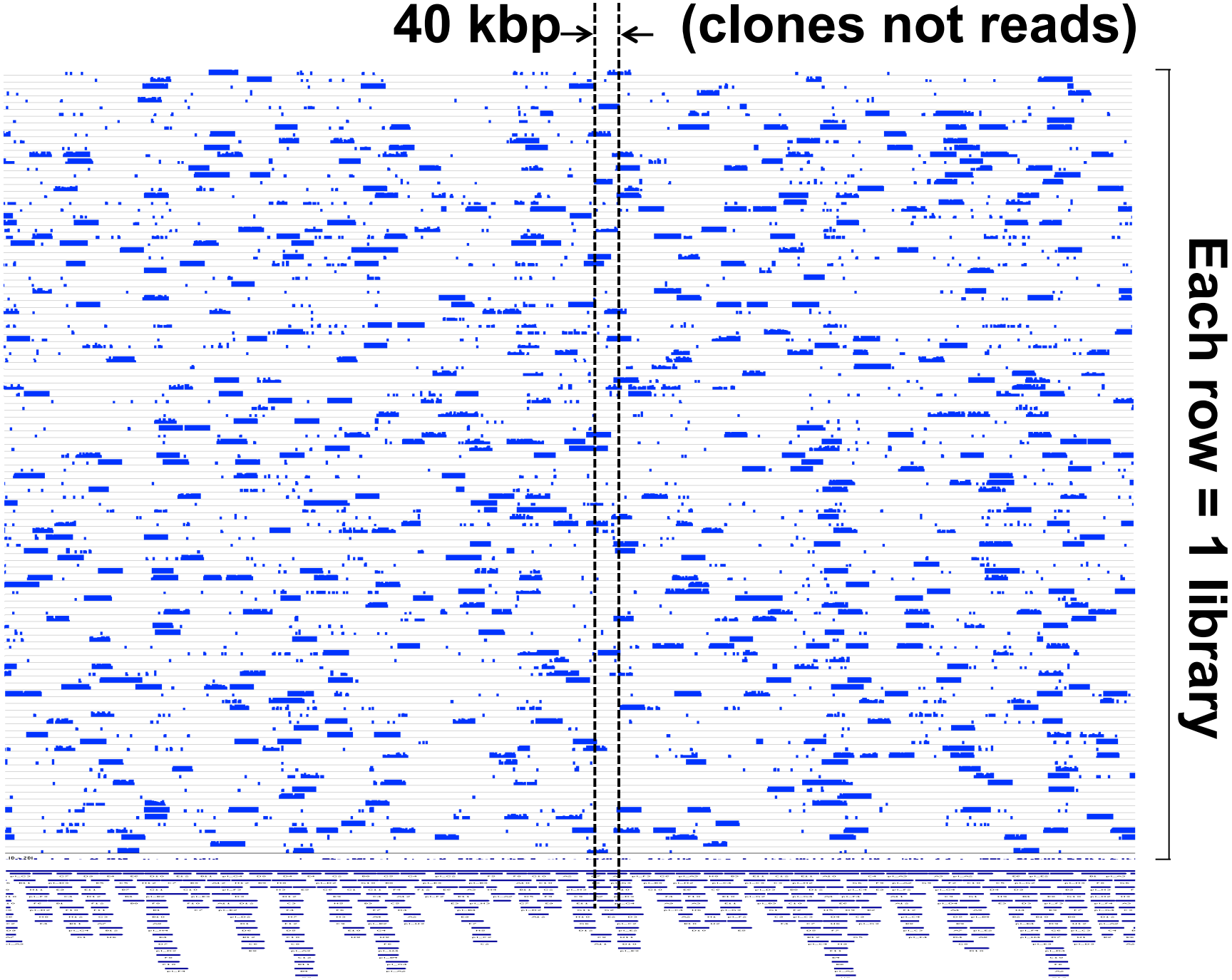


Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition

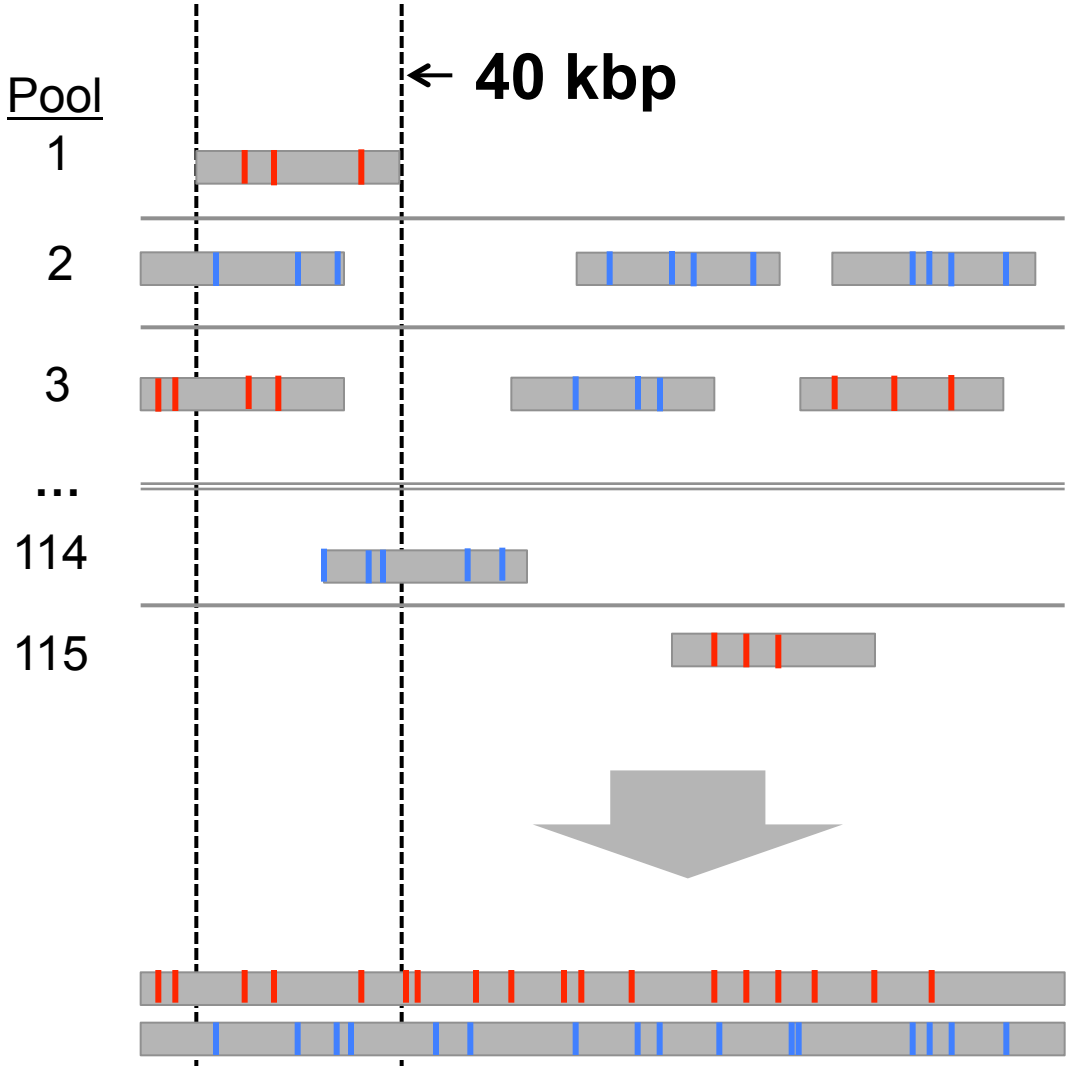
Genome Biology 2010, **11**:R119 doi:10.1186/gb-2010-11-12-r119

Andrew Adey (acadey@uw.edu)

Genome Biology

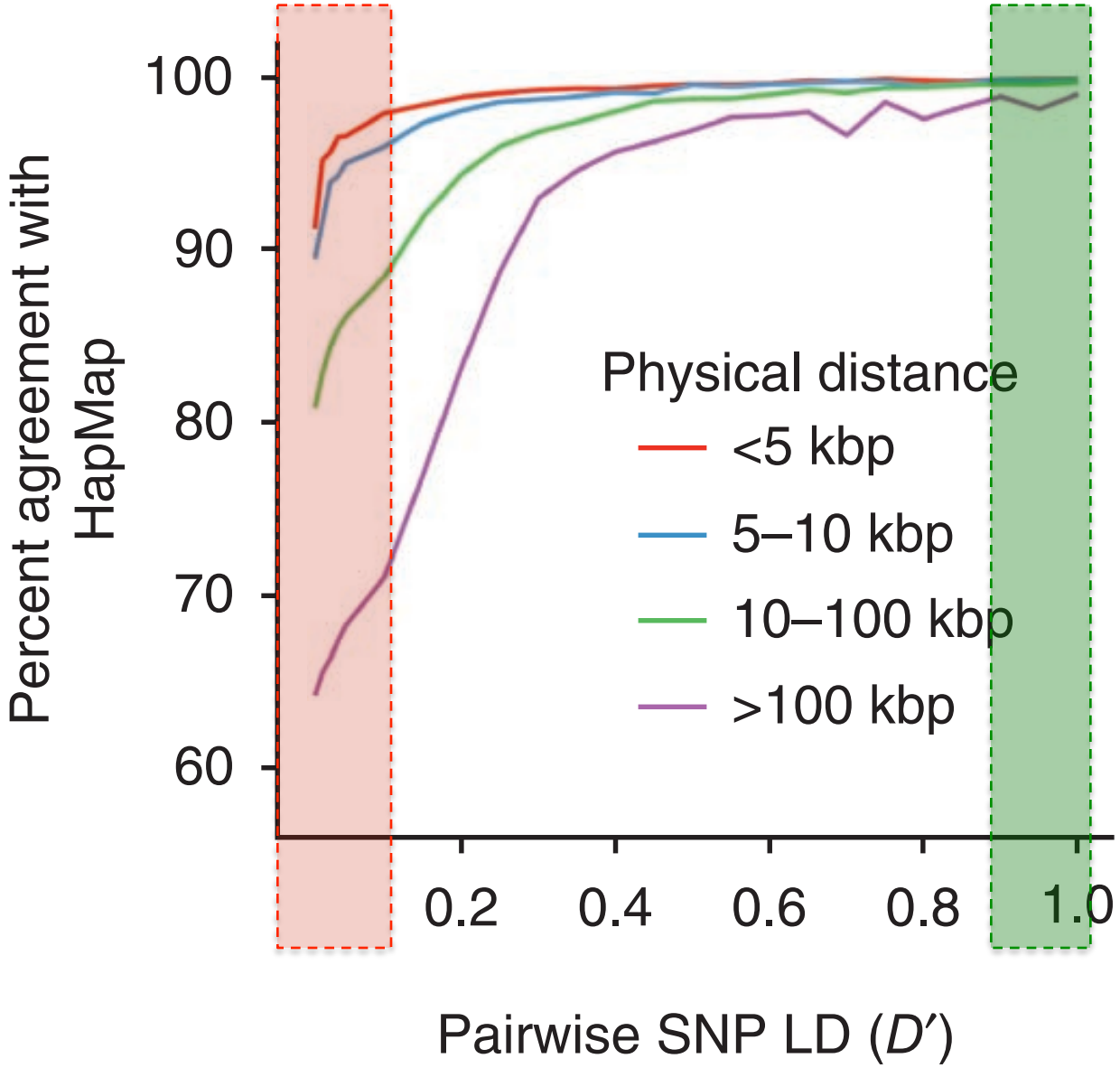

40 kbp
(clones not reads)
Each row = 1 library

Haplotype Assembly



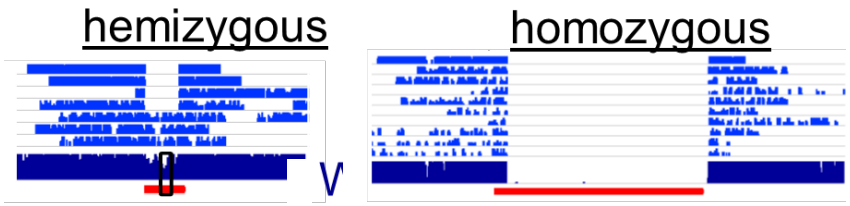
- Genotype the clone pools
- Max-parsimony assembly (Bansal *et al.*)
- **N50 = 386 Kb**

Strong concordance with HapMap calls

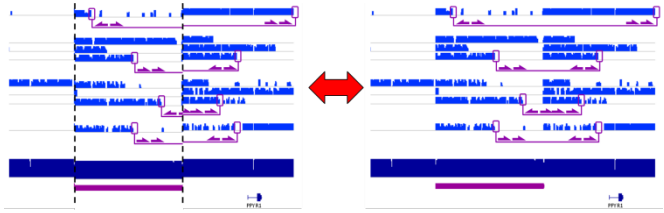


Resolving structural variation

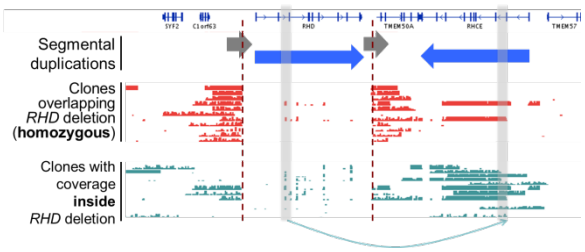
- Deletions



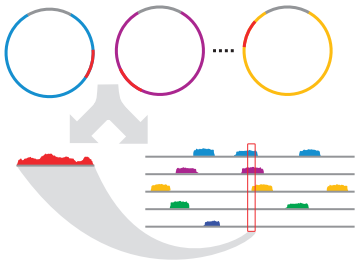
- Inversions



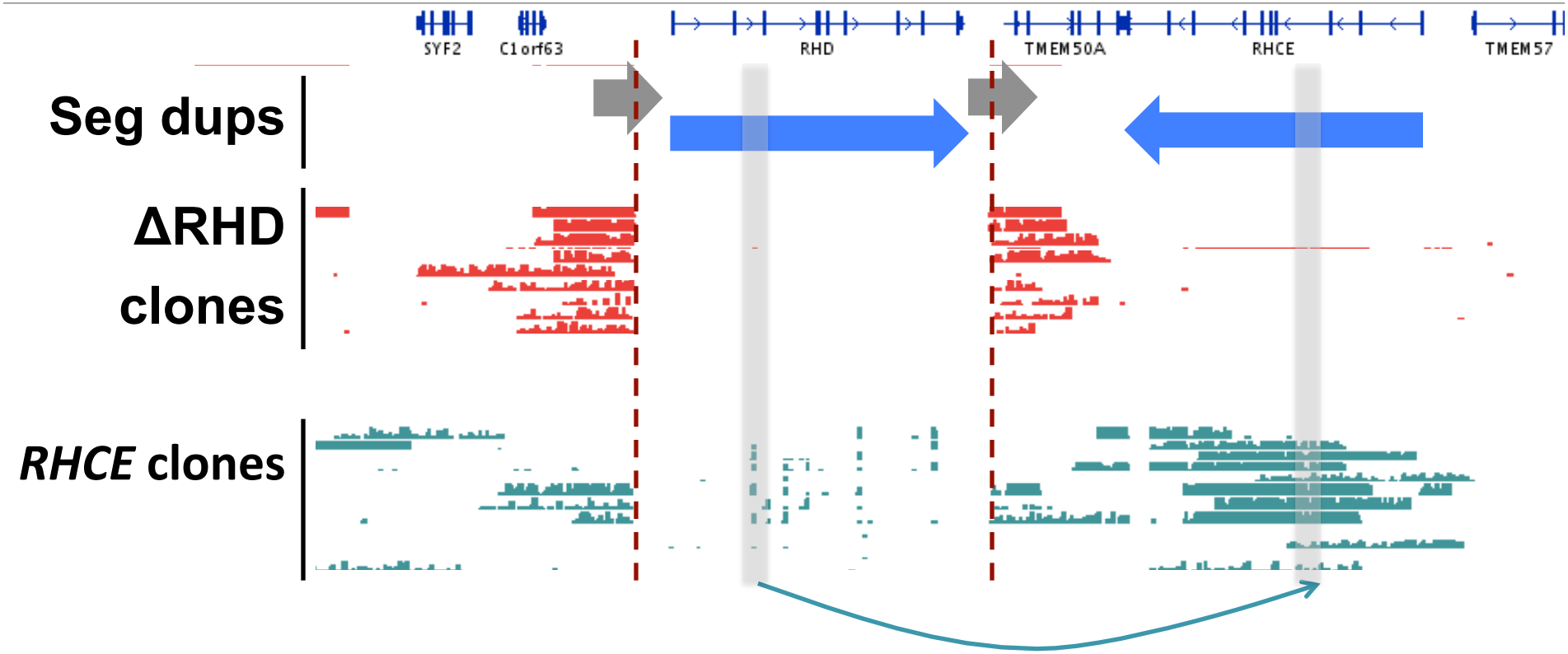
- Gene conversion



- Anchoring insertions

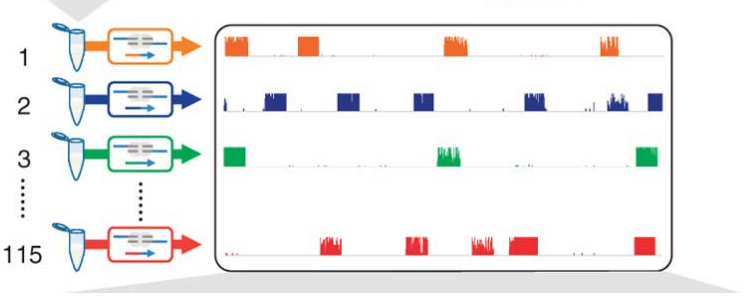


Deletion and gene conversion at *RHD*



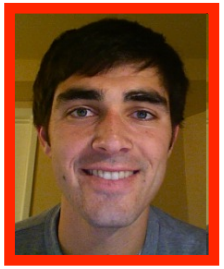
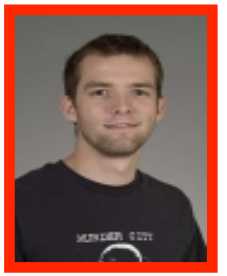
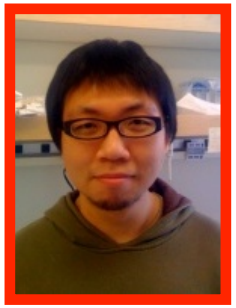
Kitzman *et al.* **Nature Biotechnology** (2011)

Technology for contiguity

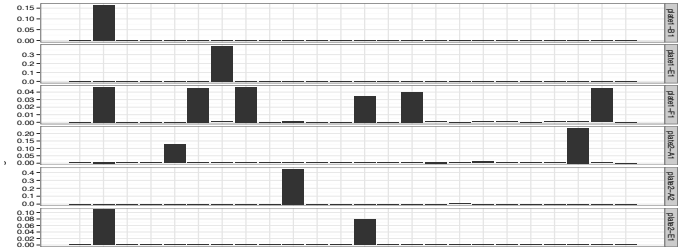


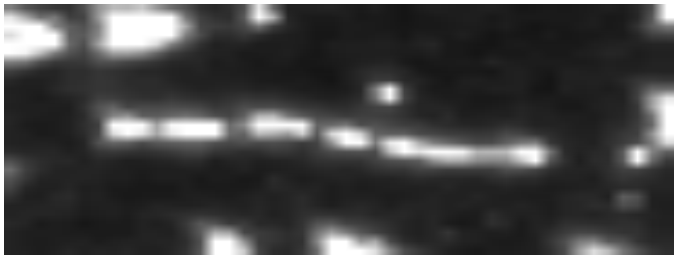
Clone Dilution
Phasing



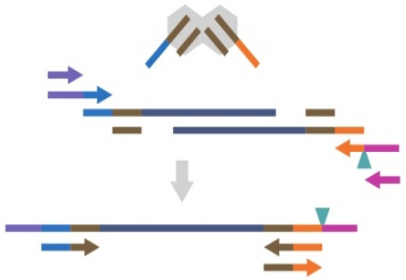

Chromosome
Sorting



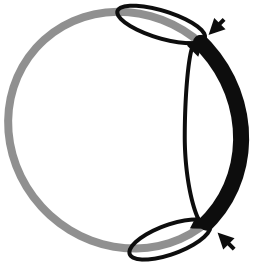


Optical
Sequencing

Transposase
Tagging



Fosmids +
Subassembly



Next generation…

Genetics

Function

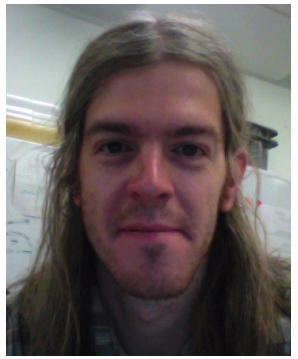
Synthetics

Tagging

Contiguity

Translation

Jacob Kitzman



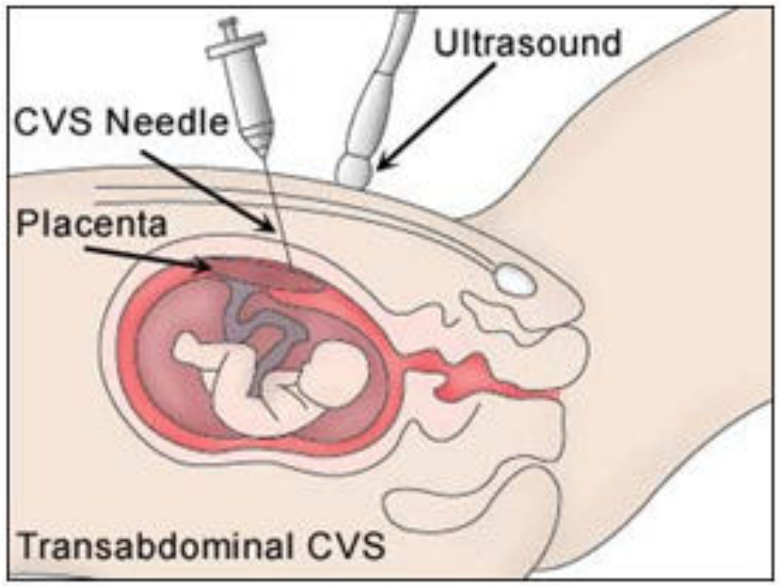
Matthew Snyder

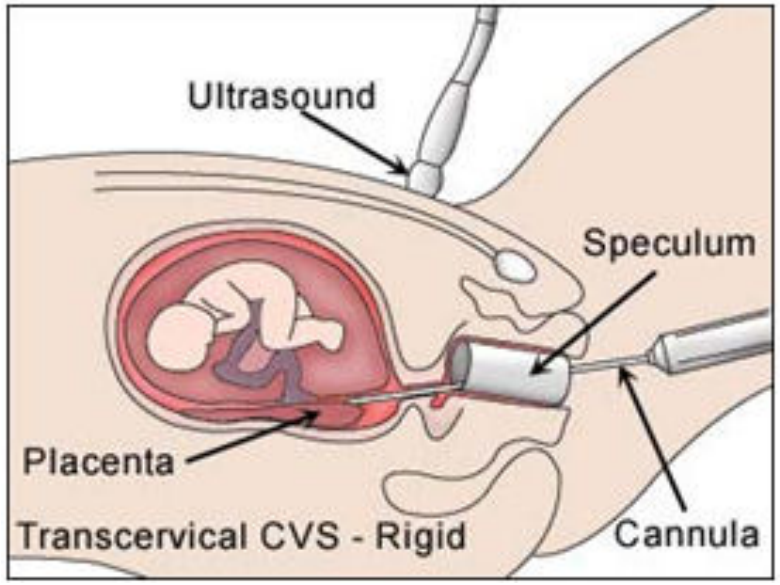
Prenatal diagnostics

Non-invasive methods:
- Ultrasound
- Maternal Serum Screening

Invasive methods:
- Amniocentesis
- Chorionic Villus Sampling

Risk vs. Sensitivity





Cell-free fetal DNA in maternal plasma

THE LANCET

Early report

Vol 350 • August 16, 1997

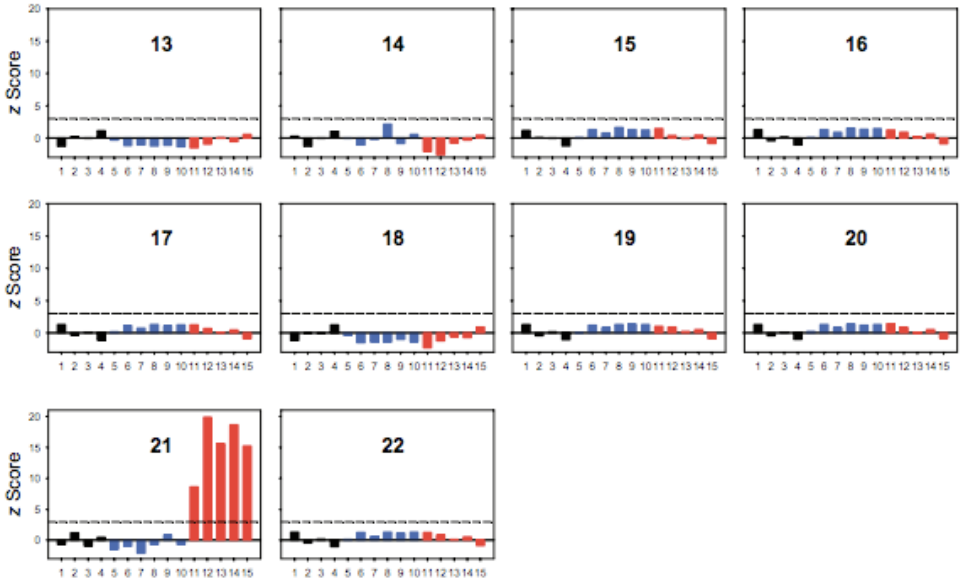
Presence of fetal DNA in maternal plasma and serum

Y M Dennis Lo, Noemi Corbetta, Paul F Chamberlain, Vik Rai, Ian L Sargent, Christopher W G Redman, James S Wainscoat

~5-15% of all cell-free DNA in plasma

Most/all of fetal genome represented

Short fragments



Maternal Plasma DNA Analysis with Massively Parallel Sequencing by Ligation for Noninvasive Prenatal Diagnosis of Trisomy 21

Rossa W.K. Chiu,[1,2] Hao Sun,[1,2] Ranjit Akolekar,[3] Christopher Clouser,[4] Clarence Lee,[4] Kevin McKernan,[4] Daixing Zhou,[4] Kypros H. Nicolaides,[3] and Y.M. Dennis Lo[1,2*]

Clinical Chemistry

But can we use this cell-free DNA to infer the complete fetal genome?

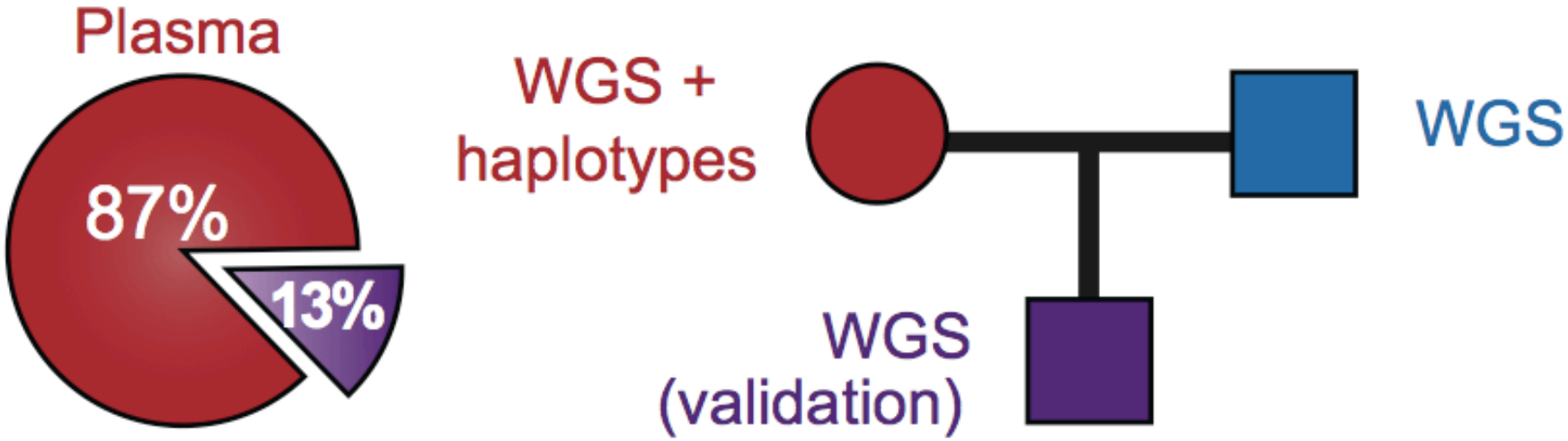
Inference of inherited alleles

- at sites where mom and dad are homozygous
- at sites where mom is heterozygous
- at sites where dad is heterozygous

Identification of candidate *de novo* mutations

Kitzman, Snyder *et al.* **Science Translational Medicine** (2012)

Experimental overview

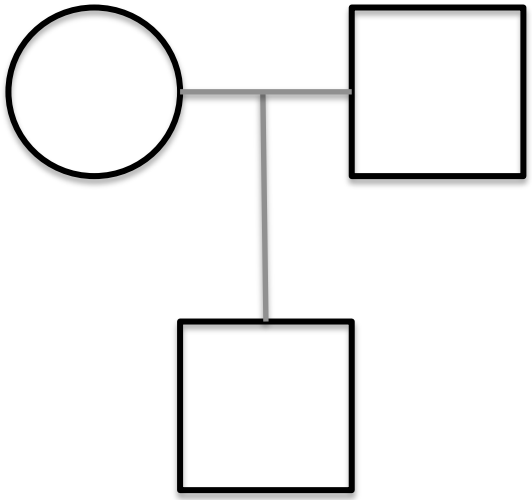


Individual	Sample	Depth of coverage
Mother (I1-M)	Plasma (5 ml, GA 18.5 wk)	78
	Whole blood (<1 mL)	32
Father (I1-P)	Saliva	39
Offspring (I1-C)	Cord blood at delivery	40

Inference is trivial at hom/hom sites

Parental Genotypes:

A / A



A / A

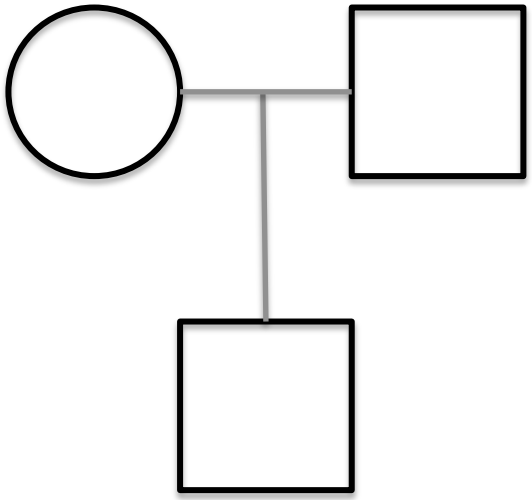
Fetal Inheritance:

A / A

Inference is trivial at ANY hom/hom site

Parental Genotypes:

A / A



B / B

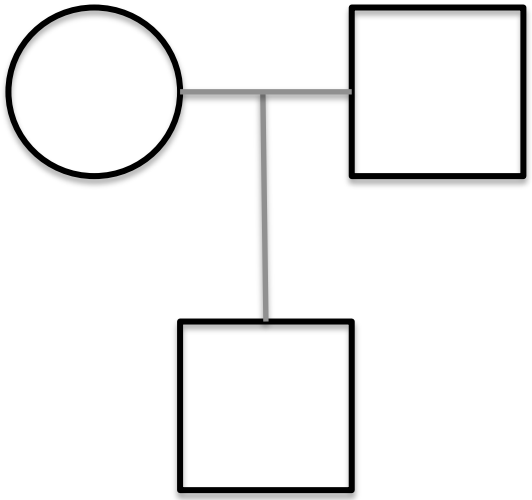
Fetal Inheritance:

A / B

Inference is trickier at heterozygous sites

Parental Genotypes:

A / B



A / A

Fetal Inheritance:

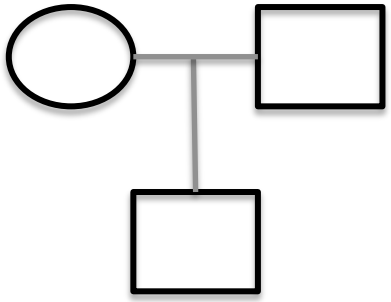
? / A

Could use allelic imbalance in plasma

Inference is trickier at heterozygous sites

Parental Genotypes:

A / B



A / A

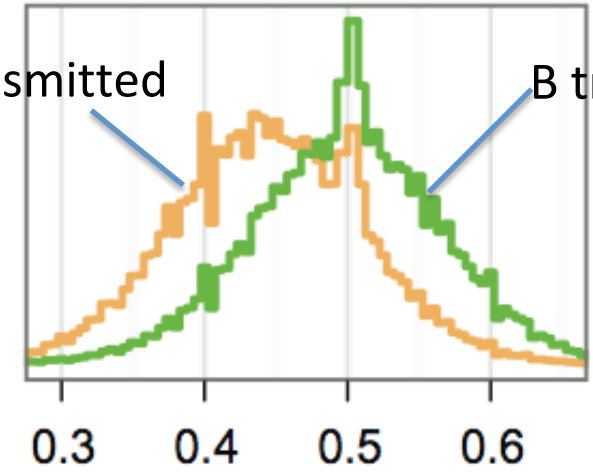
Expected # B reads:

- 43.5 *if A transmitted*
- 50 *if B transmitted*

Fetal Inheritance:

? / A

100 pla
read



Transmitted allele fraction in plasma

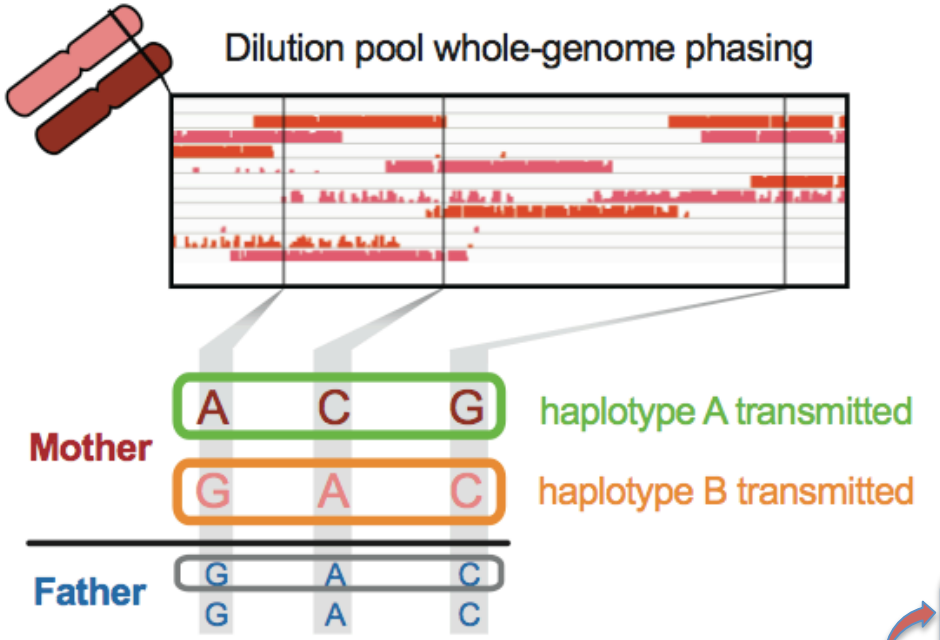
43.5 A

43.5 B

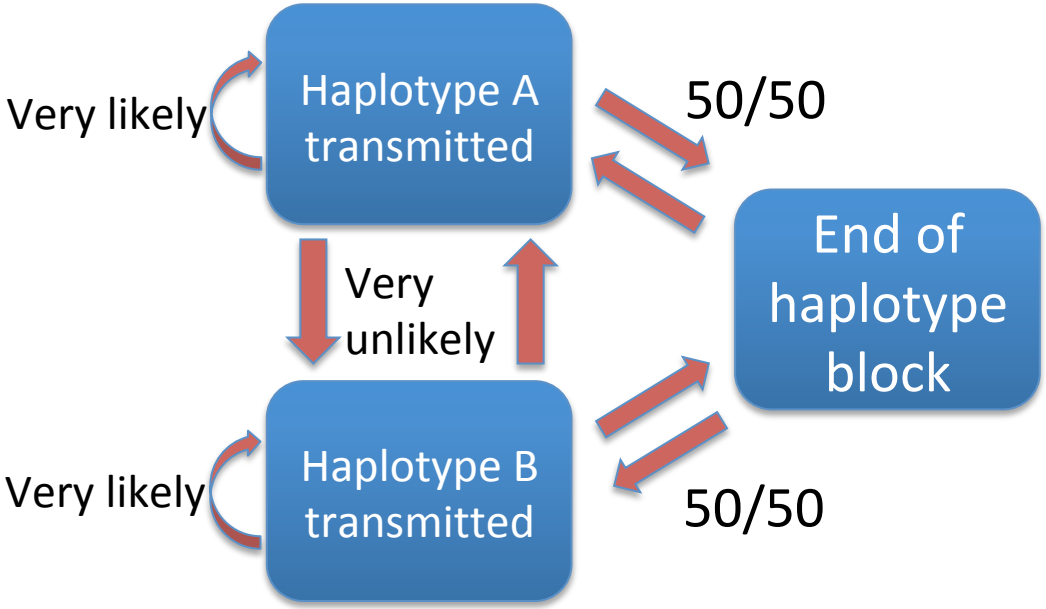
6.5 A (from dad)

6.5 A or B (from mom)

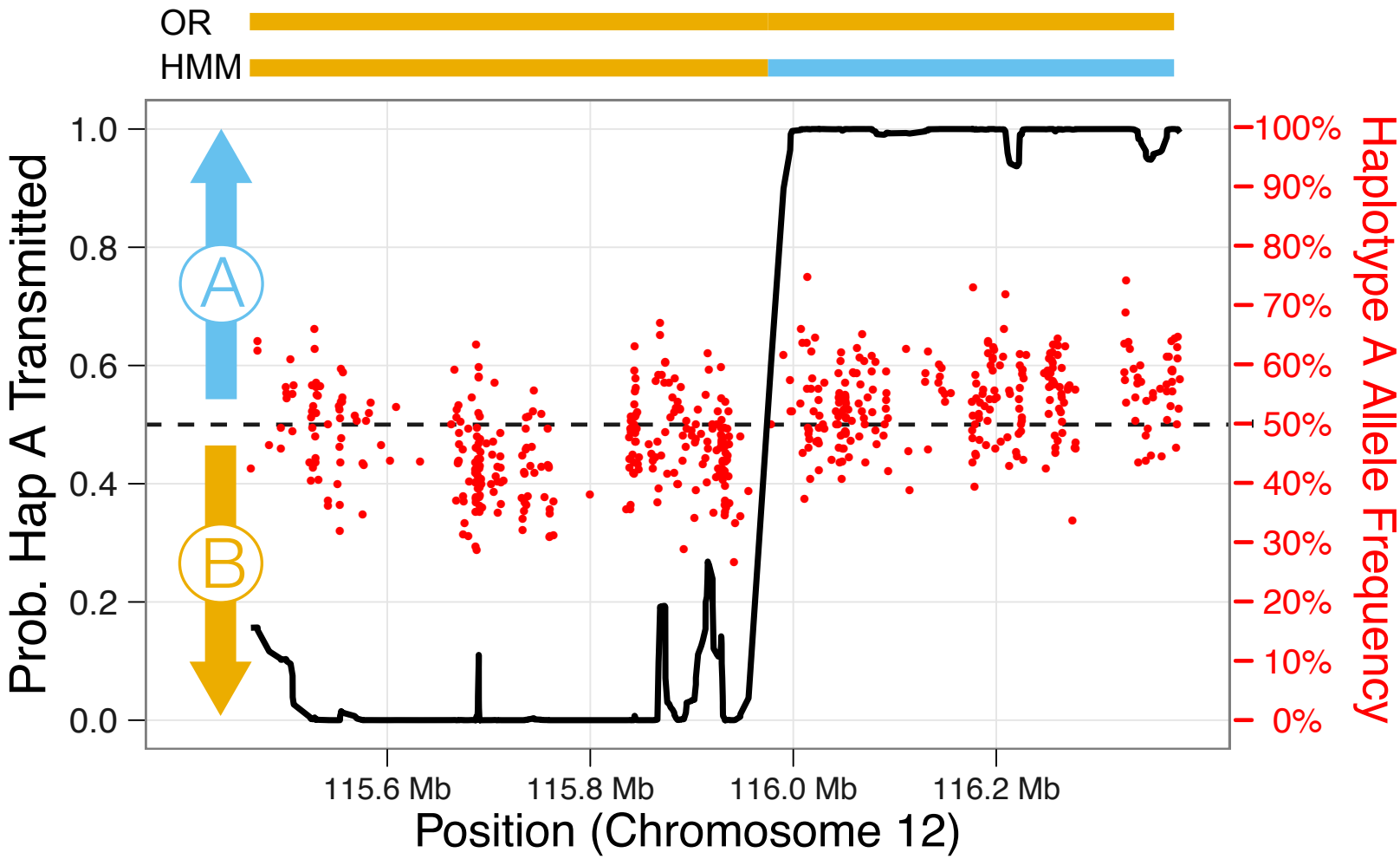
Haplotypes to the rescue!



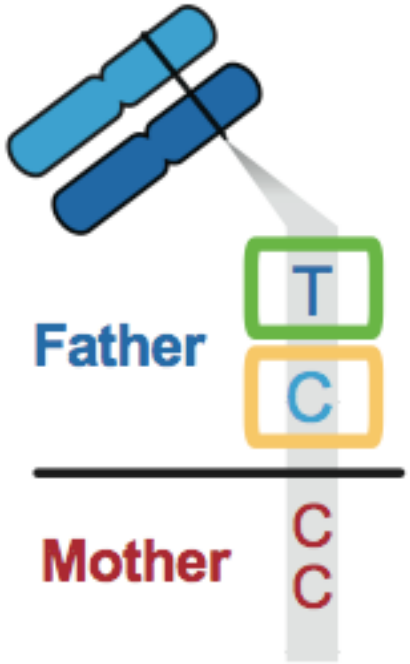
Maternal haplotype
N50 = 326 Kbp

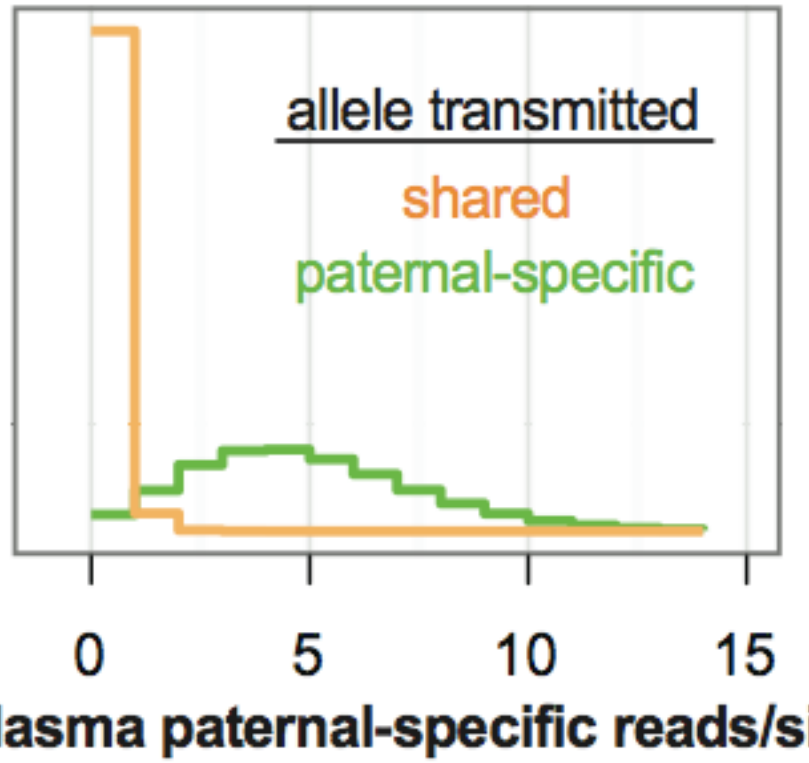


HMM allows identification of switch errors and recombination events



Paternal transmission: site-by-site

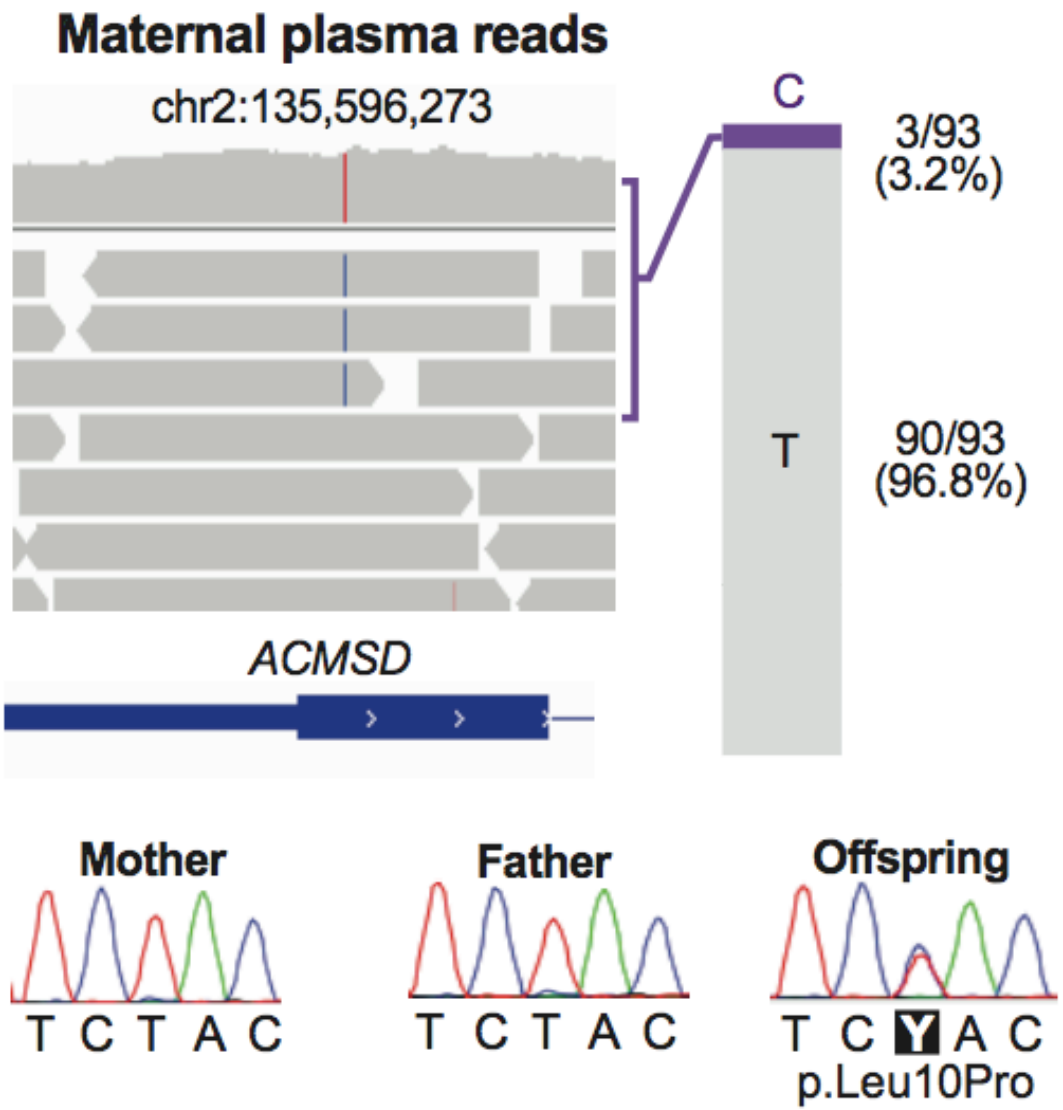




Accuracy for predicting transmission

Category	Number of Sites	Accuracy
Maternal het	1.06×10^6	99.3%
Paternal het	1.13×10^6	96.8%
Het/het	5.76×10^5	98.7%

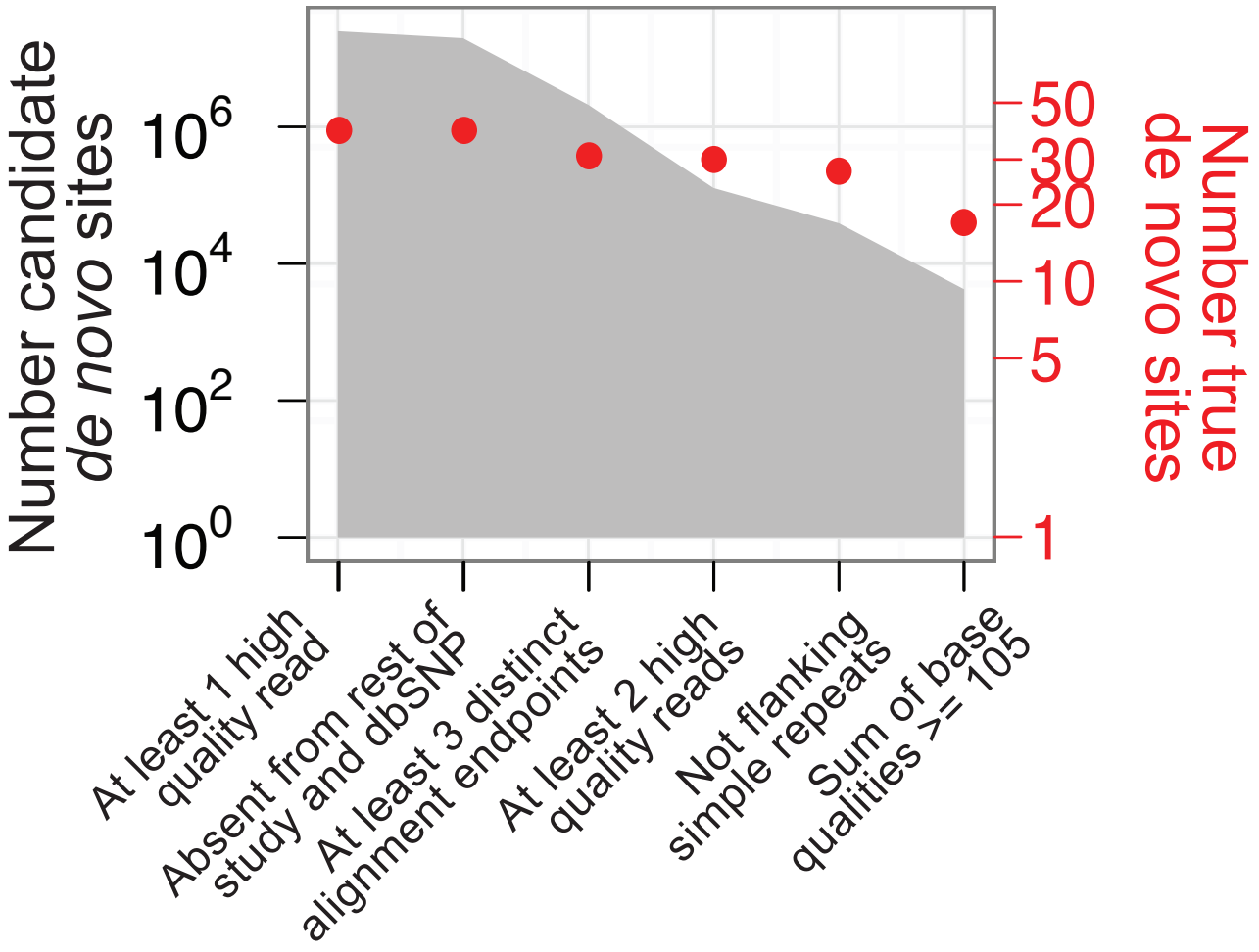
Identifying candidate *de novo* mutations



Identified 39/44 (89%) "true" *de novo* sites

Also identified 25 million other sites…

Filtering candidate *de novo* sites greatly improves signal-to-noise ratio

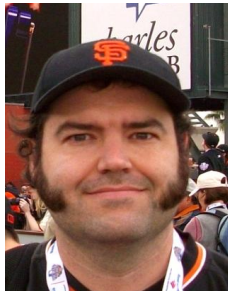


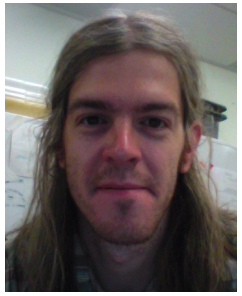
Where is this going?

- Replace all current genetic tests that are done invasively w/ non-invasive alternative
- Non-invasive pre-natal diagnosis of Mendelian disorders more broadly?
- Important to note that many technical, analytical & practical challenges remain

SHENDURE LAB

Andrew Adey
Evan Boyle
Josh Burton
Meara Davies
Joseph Hiatt
Martin Kircher
Jacob Kitzman
Akash Kumar
Choli (Charlie) Lee
Alexandra Lewis
Beth Martin
Sarah Ng
Brian O'Roak
Rupali Patwardhan
Ruolan Qiu
David Roach
Steve Salipante
Matthew Sndyer
Jerrod Schwartz
Bethany Stackhouse
Emily Turner



Mario Ventura
Holly Tabor
Jeffrey Murray
Donna Santillan
Craig Rubens
Hilary Gammill
LaVone Simmons

Michael Bamshad & Lab
Evan Eichler & Lab
Deborah Nickerson,
Mark Rieder & Lab

UNIVERSITY of WASHINGTON