Tools and statistical approaches for integrating DNA sequencing into clinical care

Doctorate of Philosophy Dissertation Defense

Dayne L Filer

Monday, November 16 2020

Department of Genetics, Curriculum in Bioinformatics & Computational Biology
Renaissance Computing Institute

Talk Outline

One-minute primer on human genetics

- Genetic material organized into chromosomes
- Central dogma of molecular biology

Copy number variation in exome sequencing

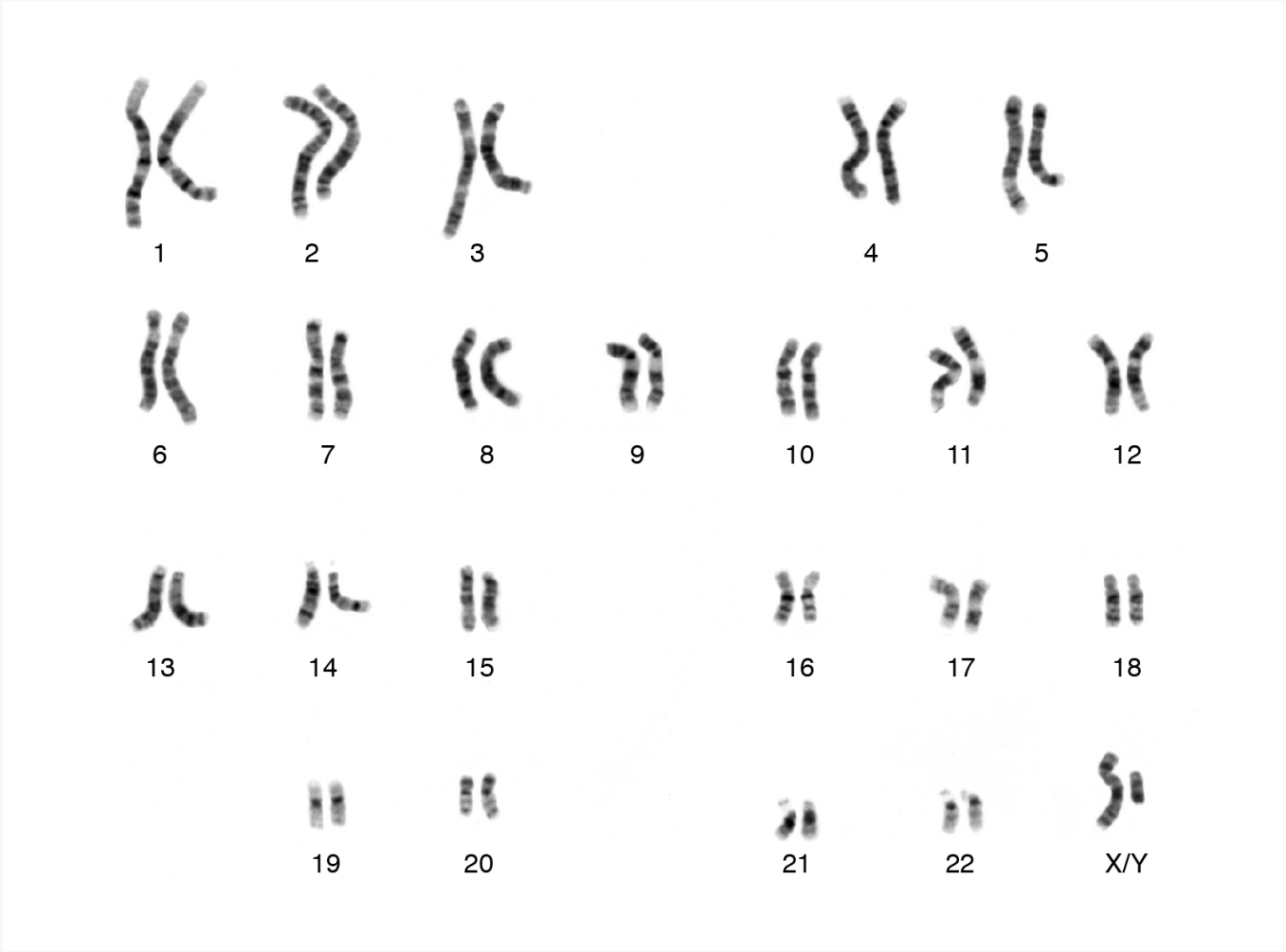
- Background and motivation
- Better capturing the exome
- Novel copy number variant algorithm

Detecting fetal variation from cell-free DNA

- Background
- Novel maternal-fetal genotyping algorithm
- Shortcomings of noninvasive exome sequencing

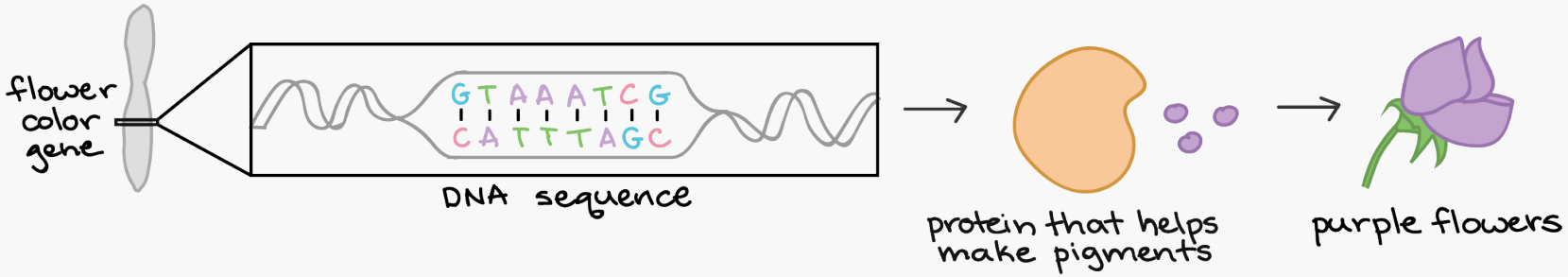
One-minute primer on human genetics

Genetic material organized into chromosomes



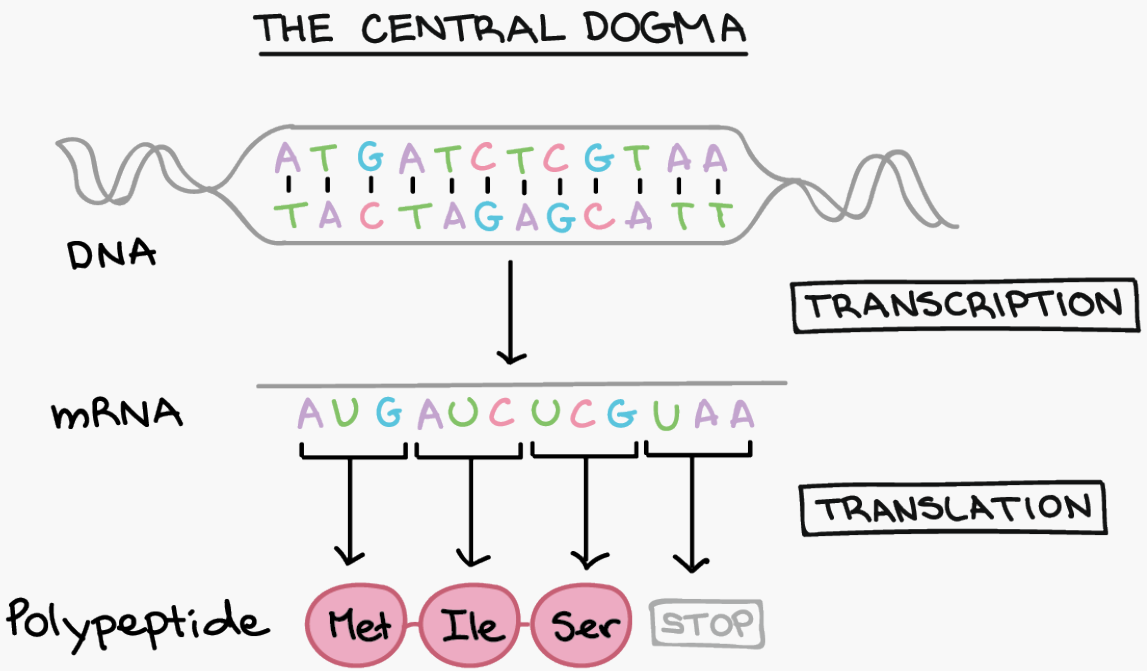
(Image taken from NHGRI)

Central dogma of molecular biology



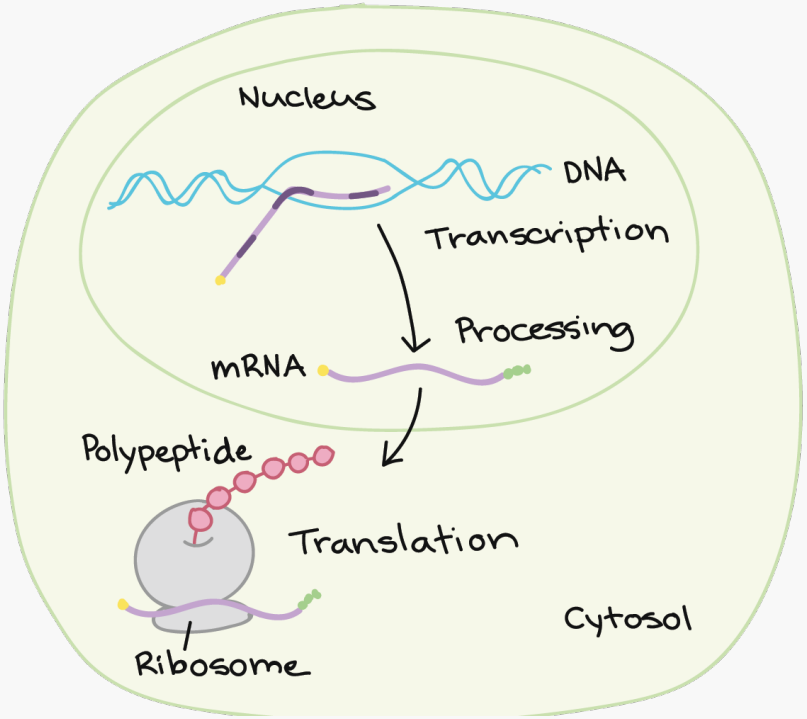
(Image taken from Khan Academy)

Central dogma of molecular biology



(Image taken from Khan Academy)

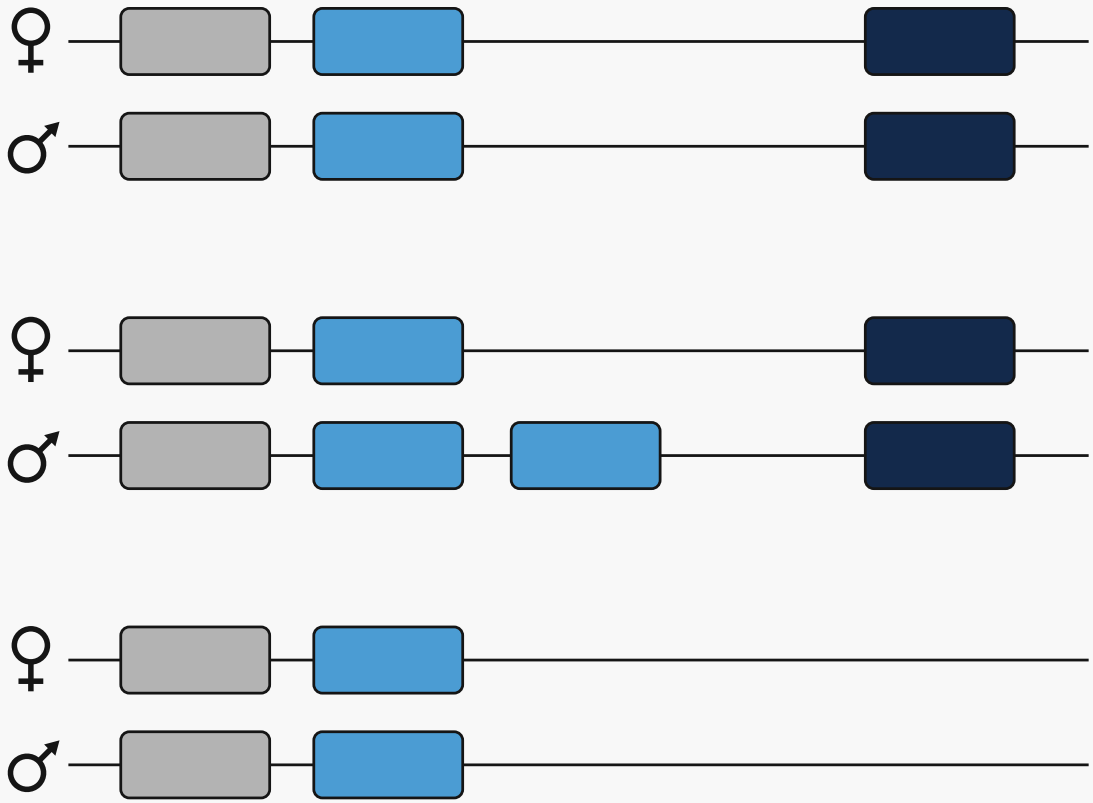
Central dogma of molecular biology



(Image taken from Khan Academy)

Copy number variation in exome sequencing

What is a copy number variant?



Motivation

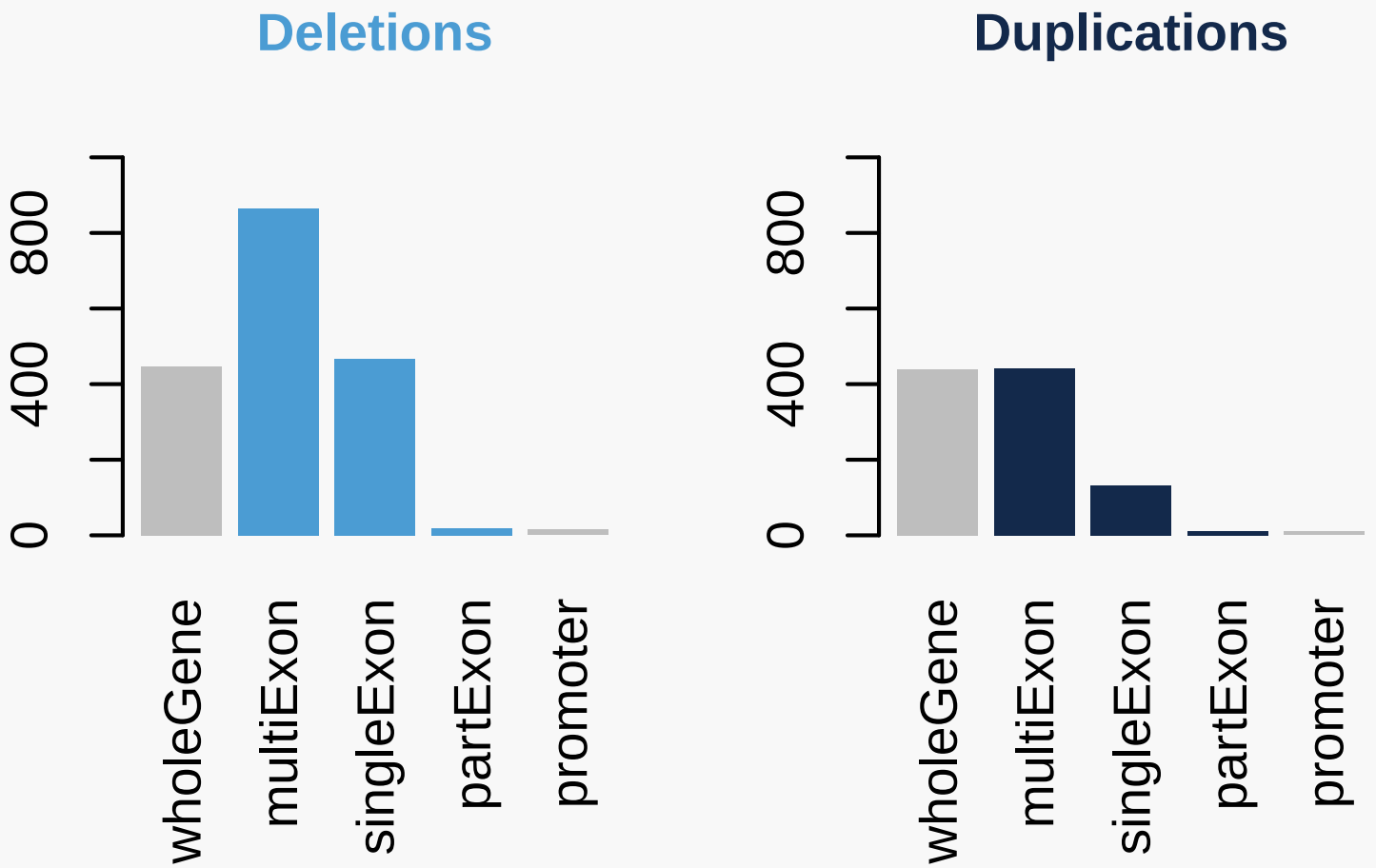
Genetics inMedicine | ARTICLE

Prevalence and properties of intragenic copy-number variation in Mendelian disease genes

Rebecca Truty, PhD[1], Joshua Paul, PhD[1], Michael Kennemer, MS[1], Stephen E. Lincoln, BS[1], Eric Olivares, PhD[1], Robert L. Nussbaum, MD, FACMG[1,2] and Swaroop Aradhya, PhD, FACMG[1]

Our analysis identified 2844 intragenic CNVs in 384 clinically tested genes. CNVs were observed in 1.9% of the entire cohort but in a disproportionately high fraction (9.8%) of individuals with a clinically significant result.

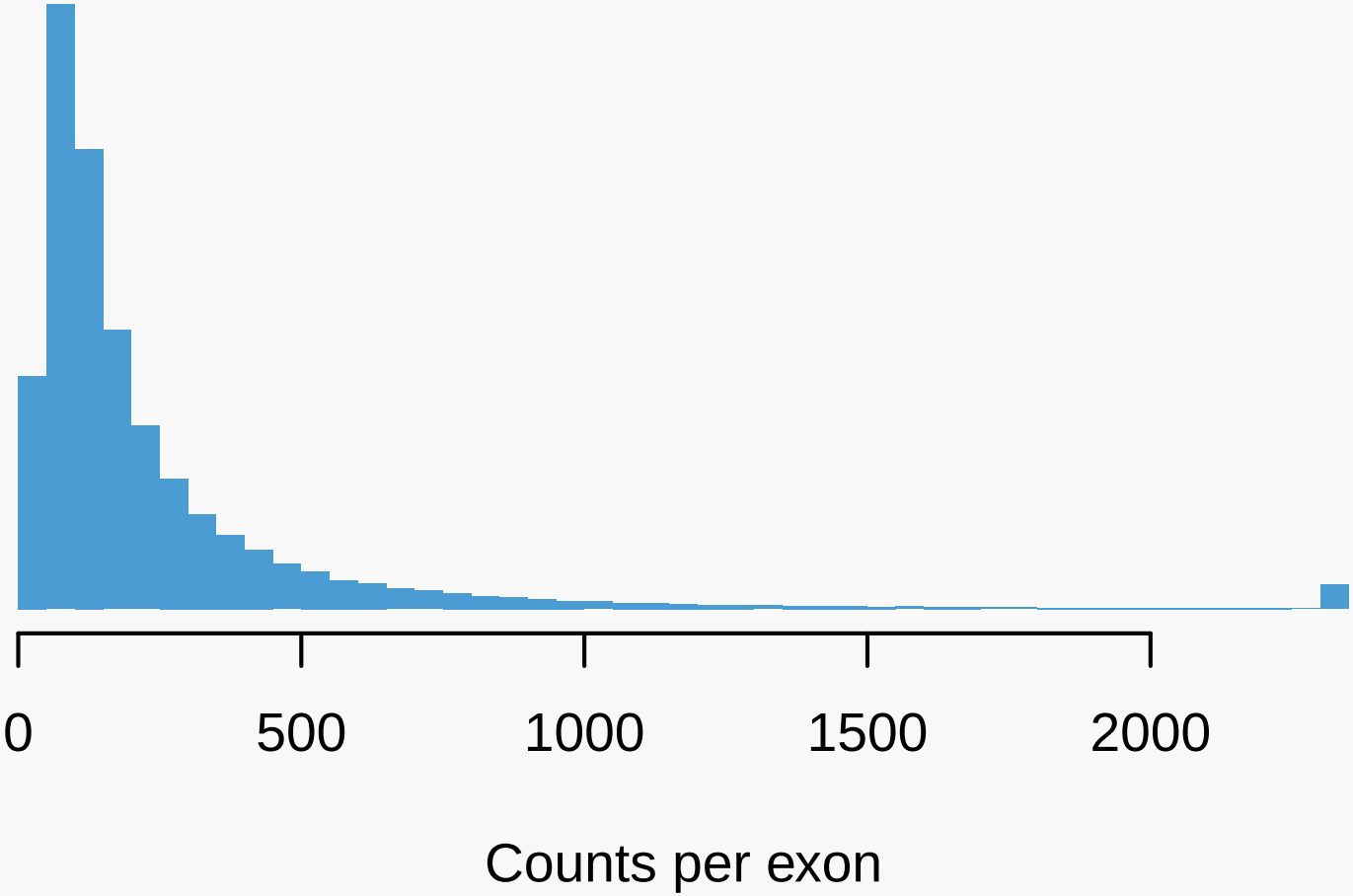
Motivation


Deletions
0
400
800
wholeGene
multiExon
singleExon
partExon
promoter
Duplications
0
400
800
wholeGene
multiExon
singleExon
partExon
promoter

Why exome sequencing?

- Only 1-2% of our DNA codes for protein (exome)
- Sequencing only the exome is cost-efficient and easier to interpret

Exon-to-exon capture variation



The current state-of-the-art: ExomeDepth

BIOINFORMATICS **ORIGINAL PAPER**

Vol. 28 no. 21 2012, pages 2747–2754
doi:10.1093/bioinformatics/bts526

Sequence analysis

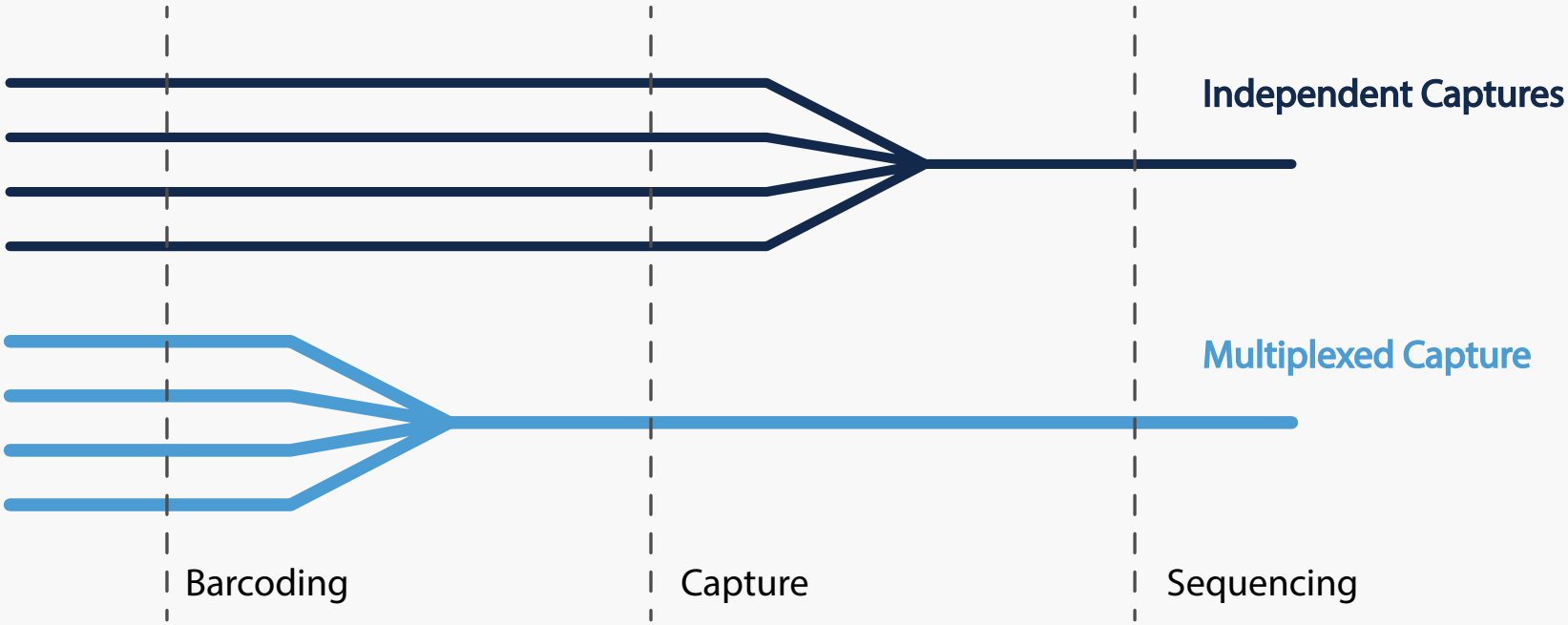
Advance Access publication August 31, 2012

A robust model for read count data in exome sequencing experiments and implications for copy number variant calling

Vincent Plagnol[1,*], James Curtis[2], Michael Epstein[1,3], Kin Y. Mok[4], Emma Stebbings[2], Sofia Grigoriadou[5], Nicholas W. Wood[4], Sophie Hambleton[6], Siobhan O. Burns[7], Adrian J. Thrasher[7], Dinakantha Kumararatne[8], Rainer Doffinger[8] and Sergey Nejentsev[2]

- ExomeDepth builds a comparative model based on the beta-binomial distribution
- Begins by finding highly-correlated samples to build a control vector
- Defaults to an expected CNV size of 50kb – not expected to perform well for small (exon level) variation

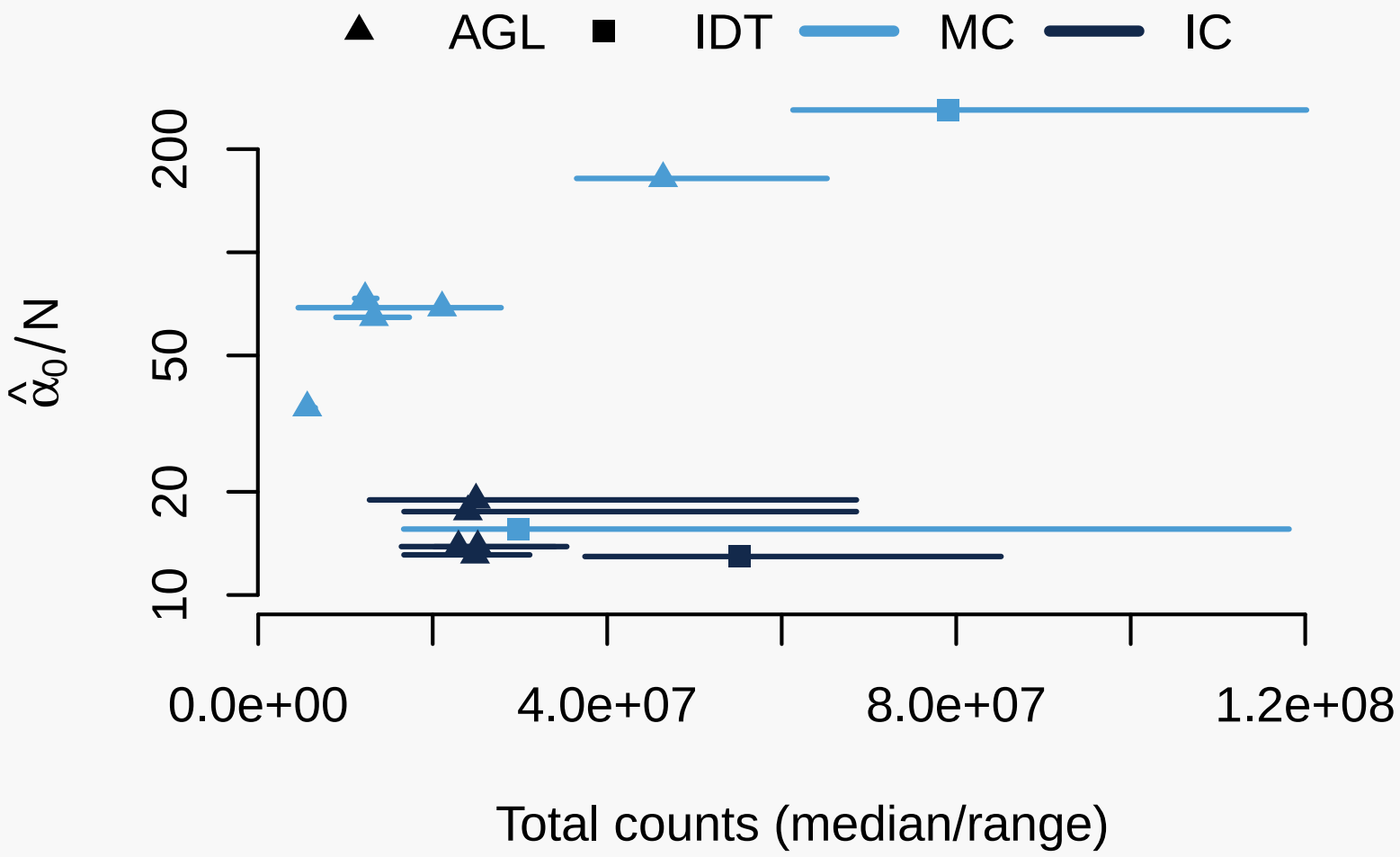
Can we better capture exomes?



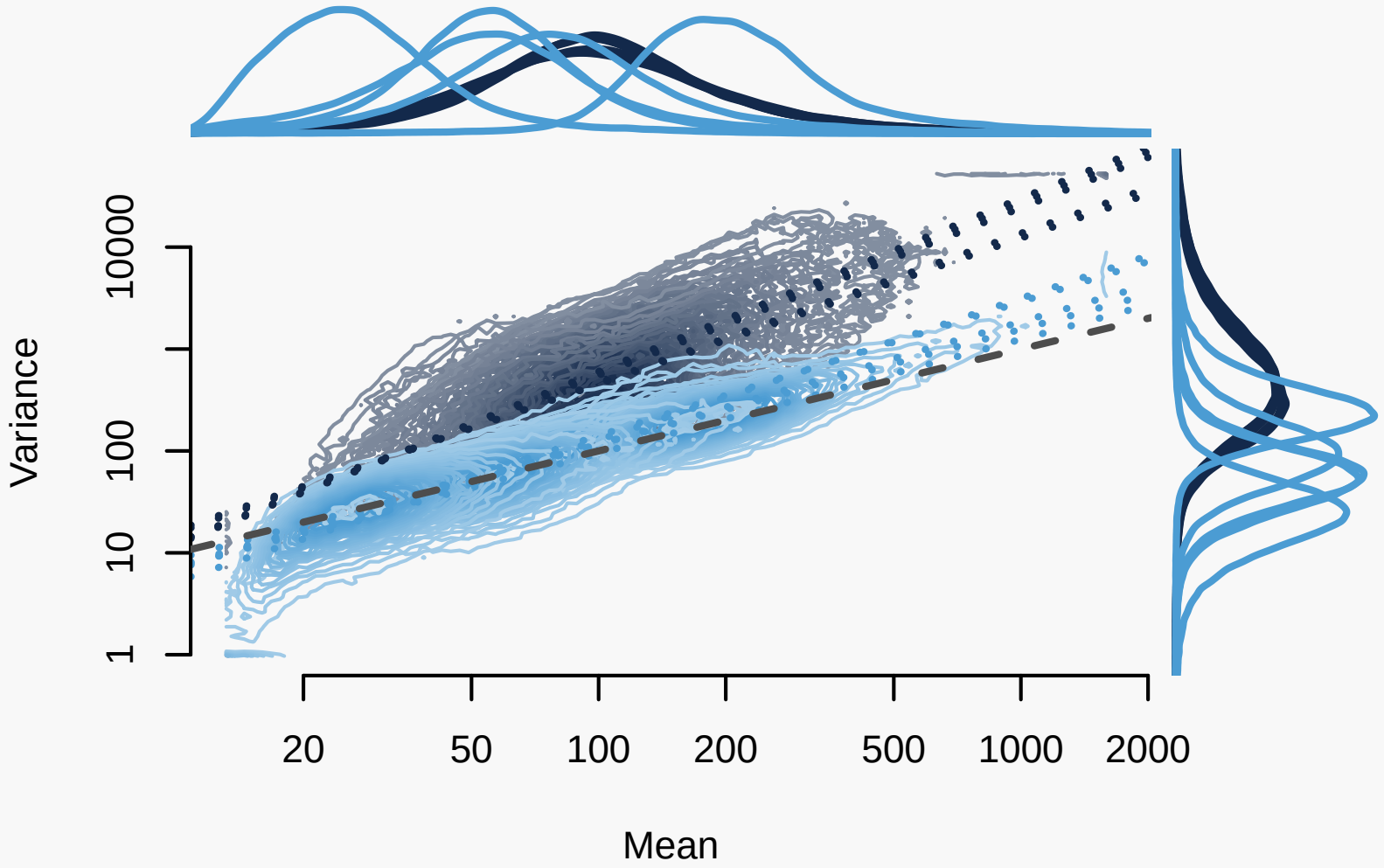
Can we better capture exomes?

pool	capture	N	medExon	medTotal	rsdTotal
IDT-IC†	IDT	16	143	55,149,058	22.4
IDT-MC	IDT	16	93	29,772,684	64.2
IDT-RR	IDT	16	272	79,079,629	22.9
NCGENES†	Agilent	112	93	24,451,245	27.6
Pool1	Agilent	16	56	13,265,614	18.5
Pool2	Agilent	16	86	21,076,056	27.6
SMA1	Agilent	8	56	12,256,002	6.2
SMA2	Agilent	8	25	5,622,040	10.4
WGS	Agilent	16	196	46,406,224	16.4

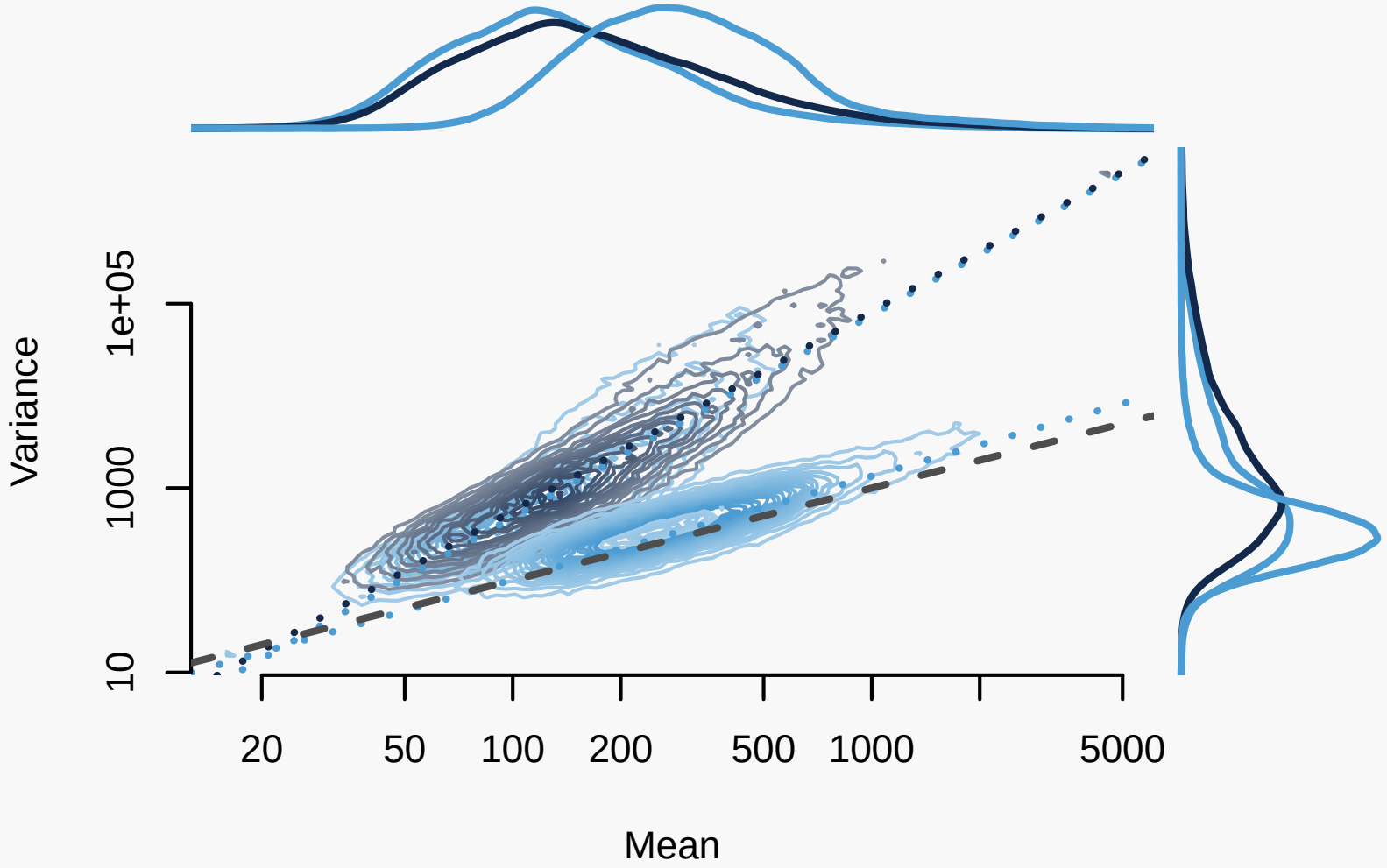
Can we better capture exomes?



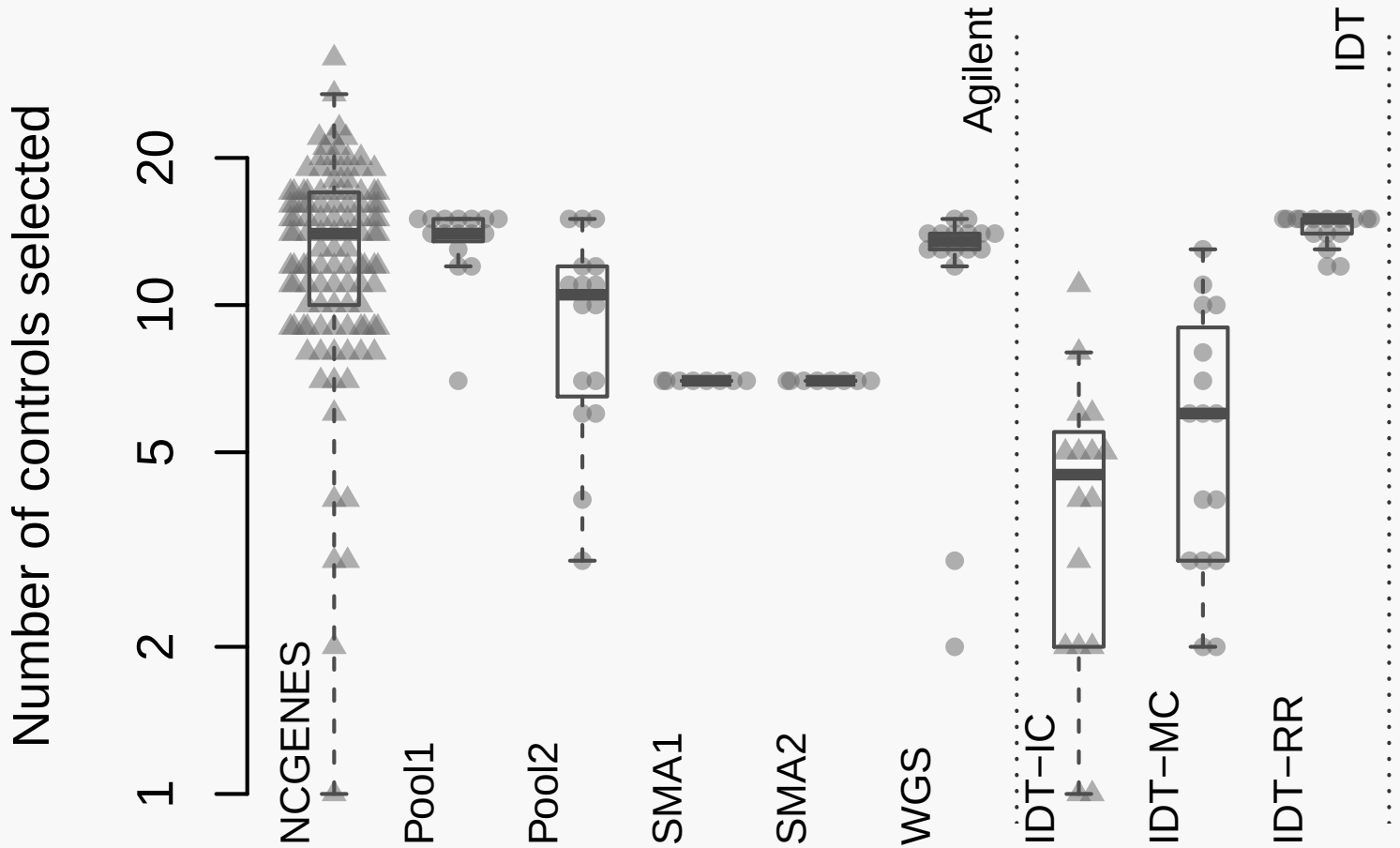
Can we better capture exomes?

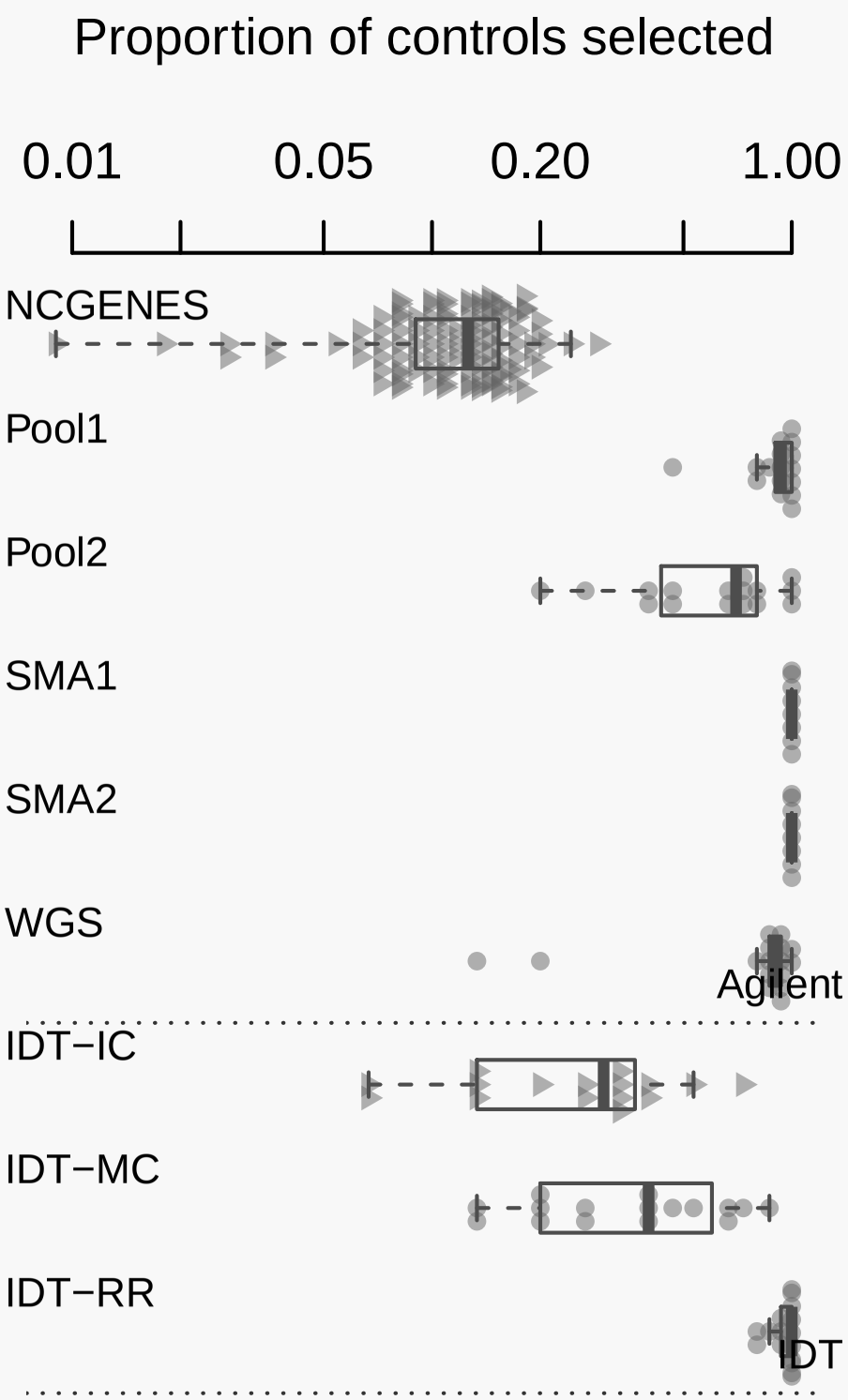


Can we better capture exomes?

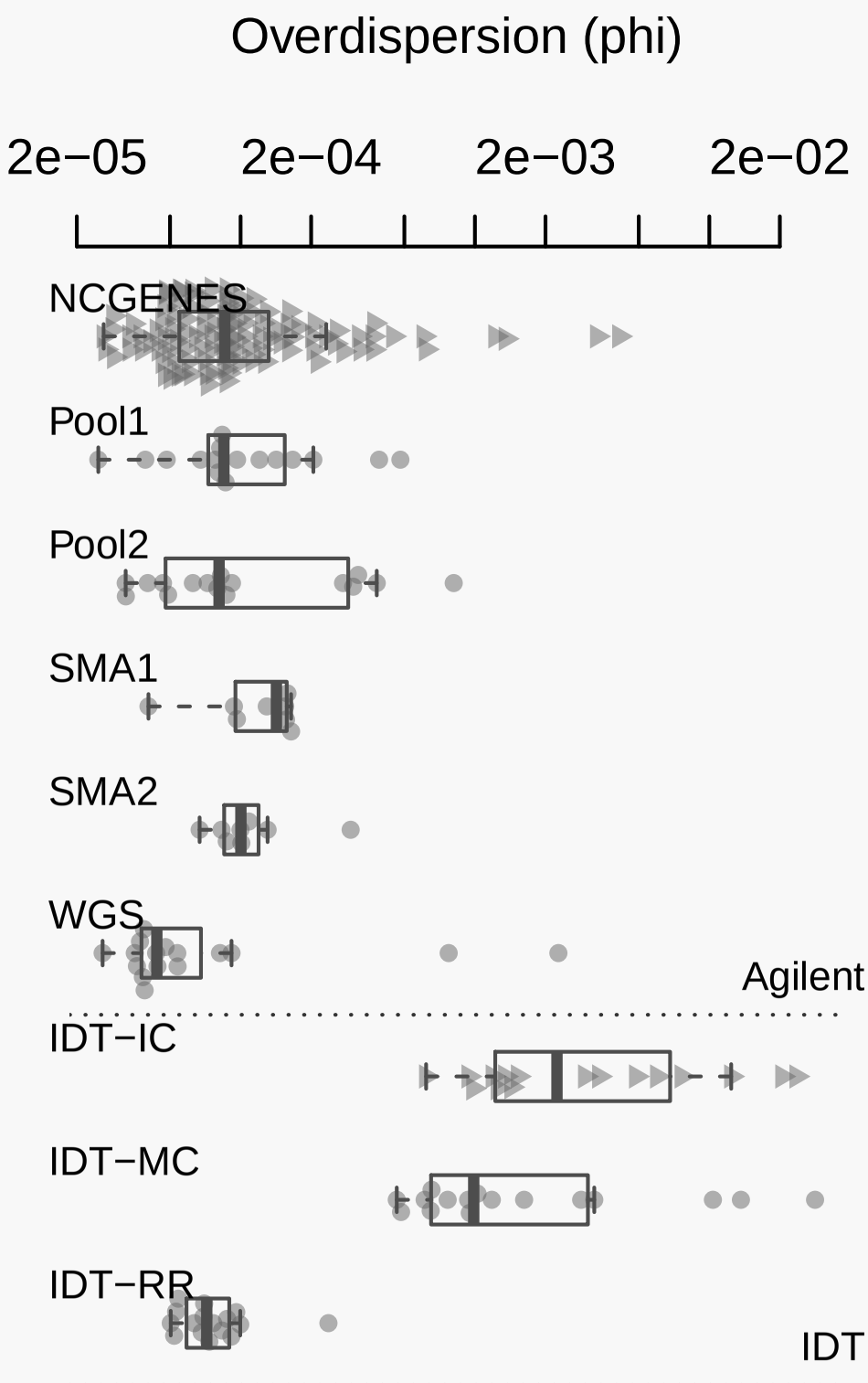


Can we better capture exomes?




Proportion of controls selected
0.01
0.05
0.20
1.00
NCGENES
Pool1
Pool2
SMA1
SMA2
WGS
Agilent
IDT−IC
IDT−MC
IDT−RR
IDT

Can we better capture exomes?



Can we better capture exomes?

Yes!

We recommend

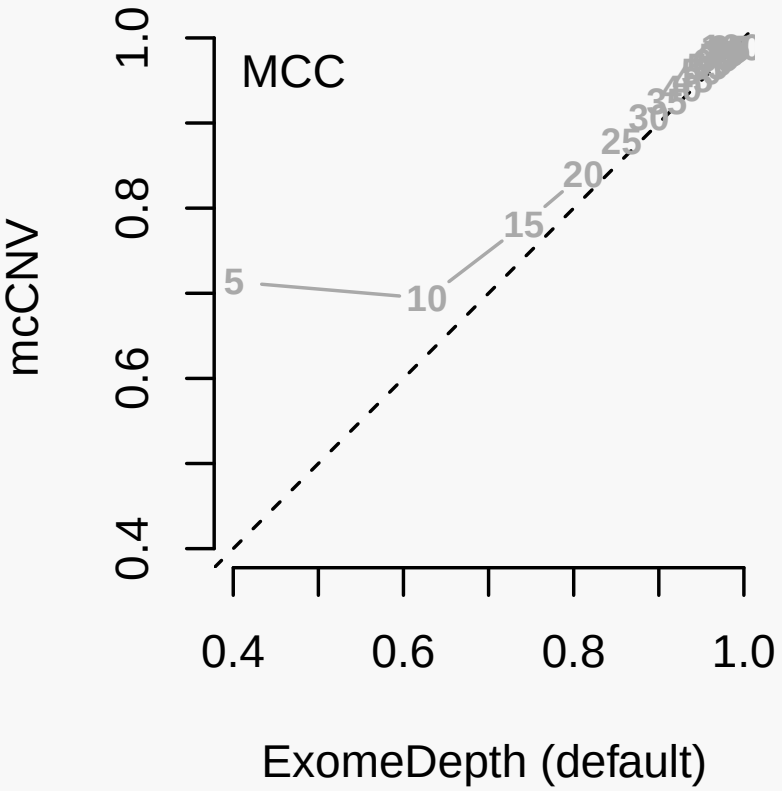
1. Multiplexing roughly 16 samples prior to exome capture
2. Checking for library balance prior to exome capture (from our limited data, we suggest RSD 30%)

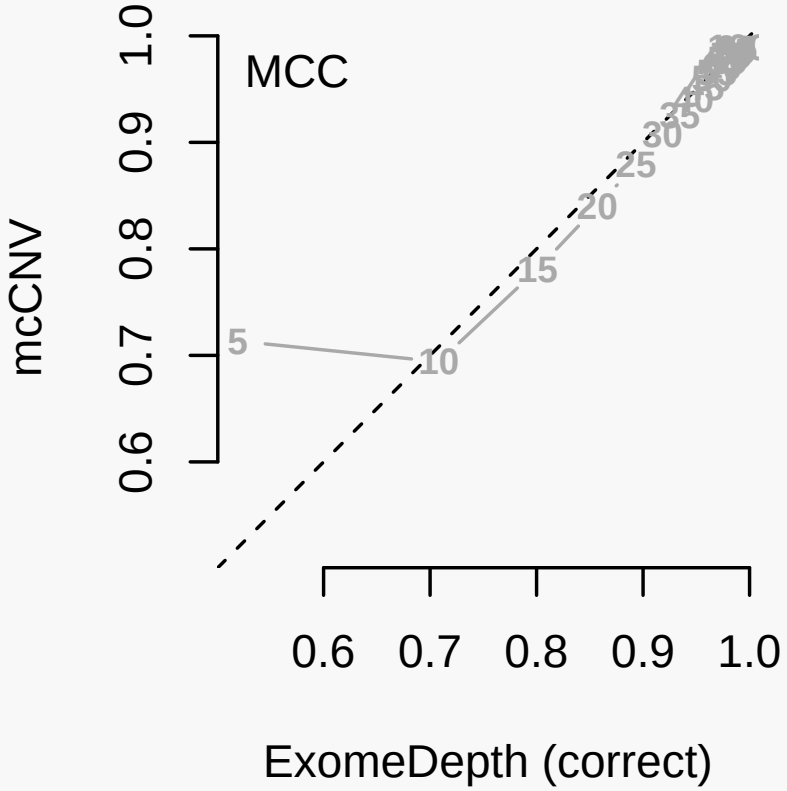
Can we better analyze exomes?

mcCNV algorithm

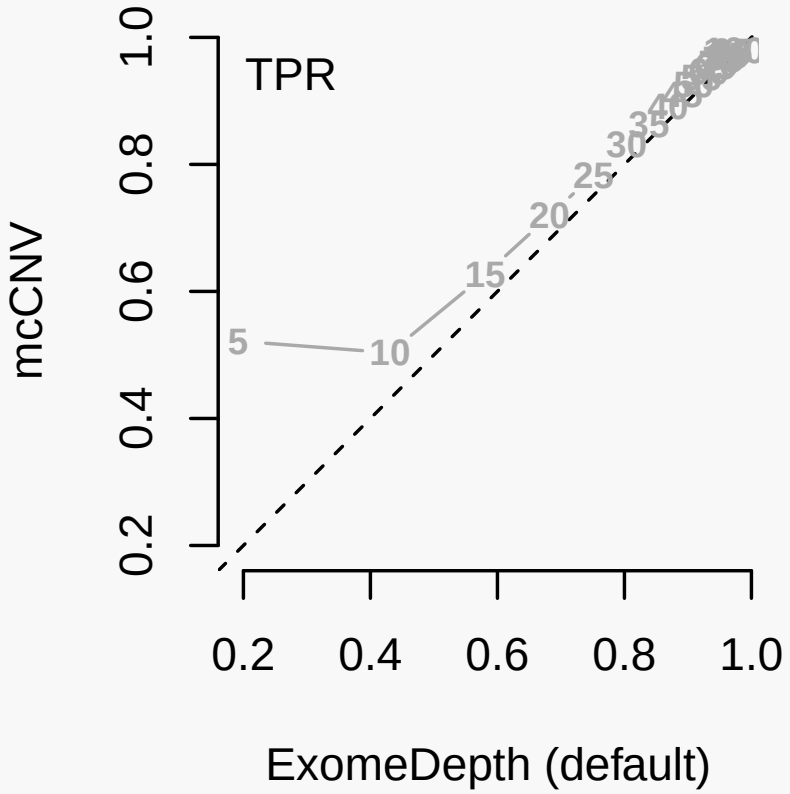
1. Assumes multiplexed capture
2. Models observed counts using the negative binomial model
3. Uses a shrinkage estimator for the dispersion

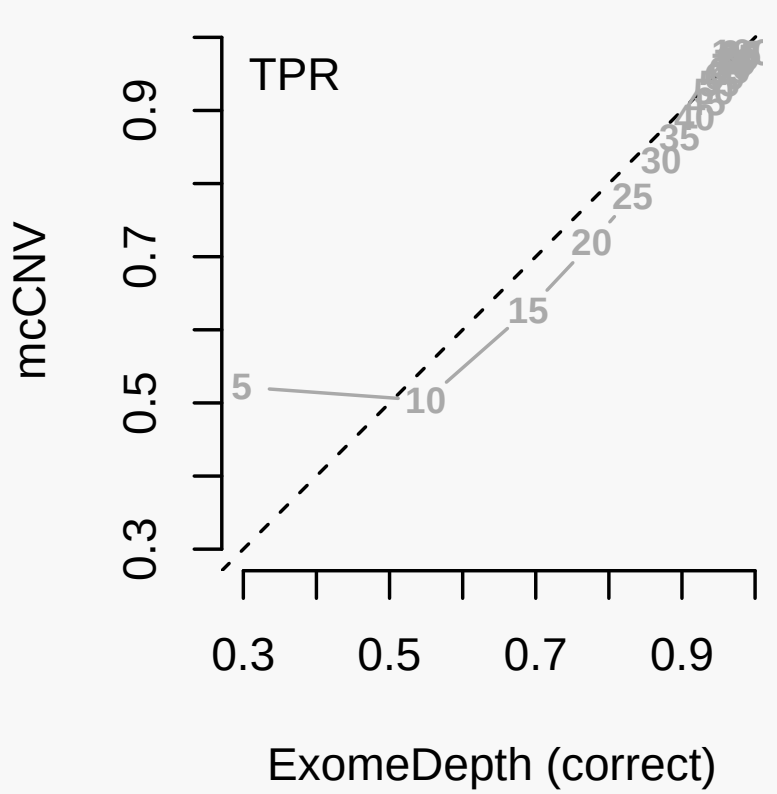
Can we better analyze exomes?



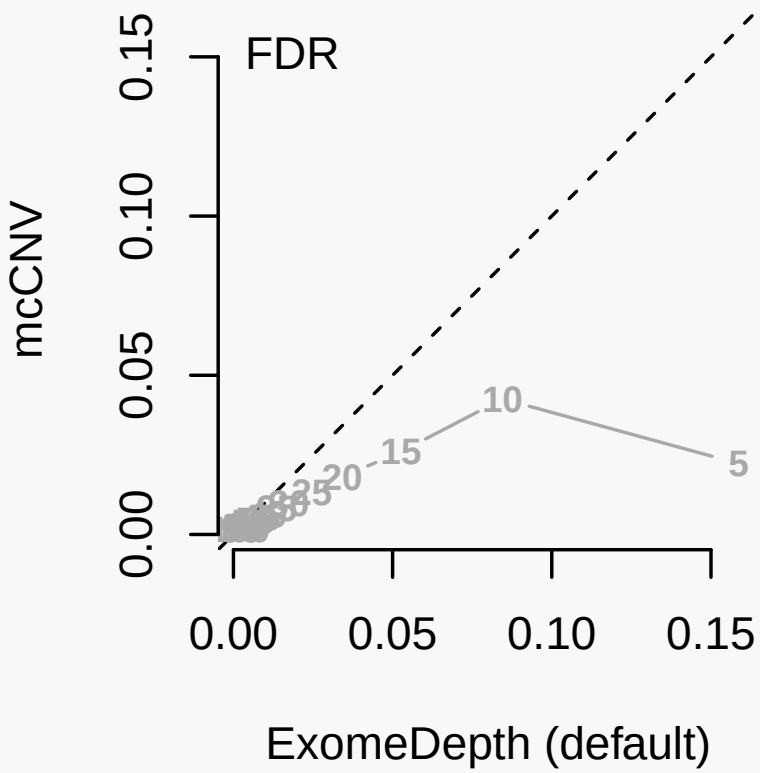


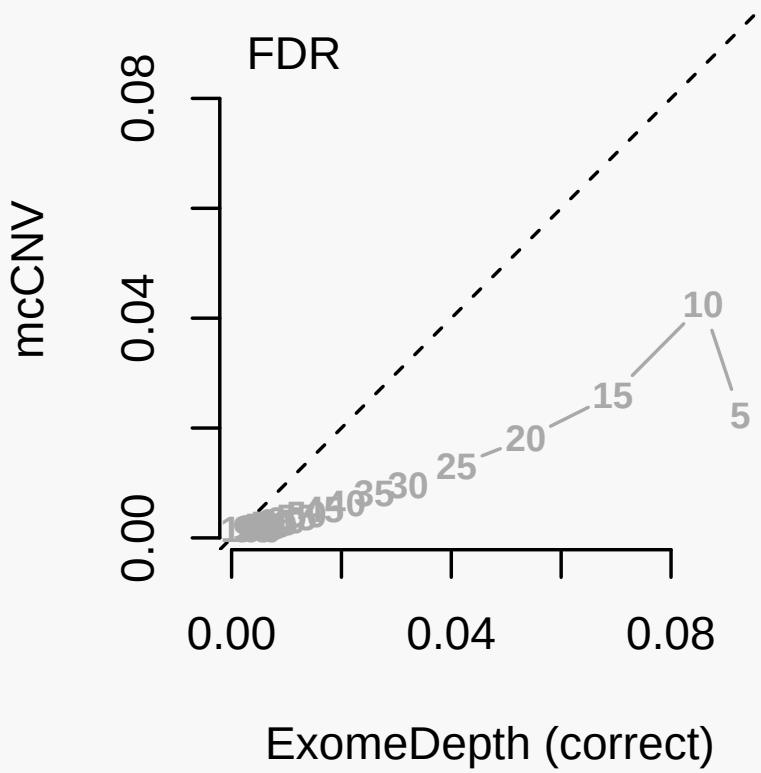
Can we better analyze exomes?





Can we better analyze exomes?





Can we better analyze exomes?

ARTICLE

A Comprehensive Workflow for Read Depth-Based Identification of Copy-Number Variation from Whole-Genome Sequence Data

Brett Trost,[1] Susan Walker,[1] Zhuozhi Wang,[1] Bhooma Thiruvahindrapuram,[1] Jeffrey R. MacDonald,[1] Wilson W.L. Sung,[1] Sergio L. Pereira,[1] Joe Whitney,[1] Ada J.S. Chan,[1,2] Giovanna Pellecchia,[1] Miriam S. Reuter,[1] Si Lok,[1] Ryan K.C. Yuen,[1] Christian R. Marshall,[1,3] Daniele Merico,[1,4,6] and Stephen W. Scherer[1,2,5,6,*]

- Combines calls from two algorithms: ERDS and CNVnator
- Excludes repetitive and low-complexity regions (RCLRs)
- Followed their recommendations, and removed all exons overlapping their defined RLCRs

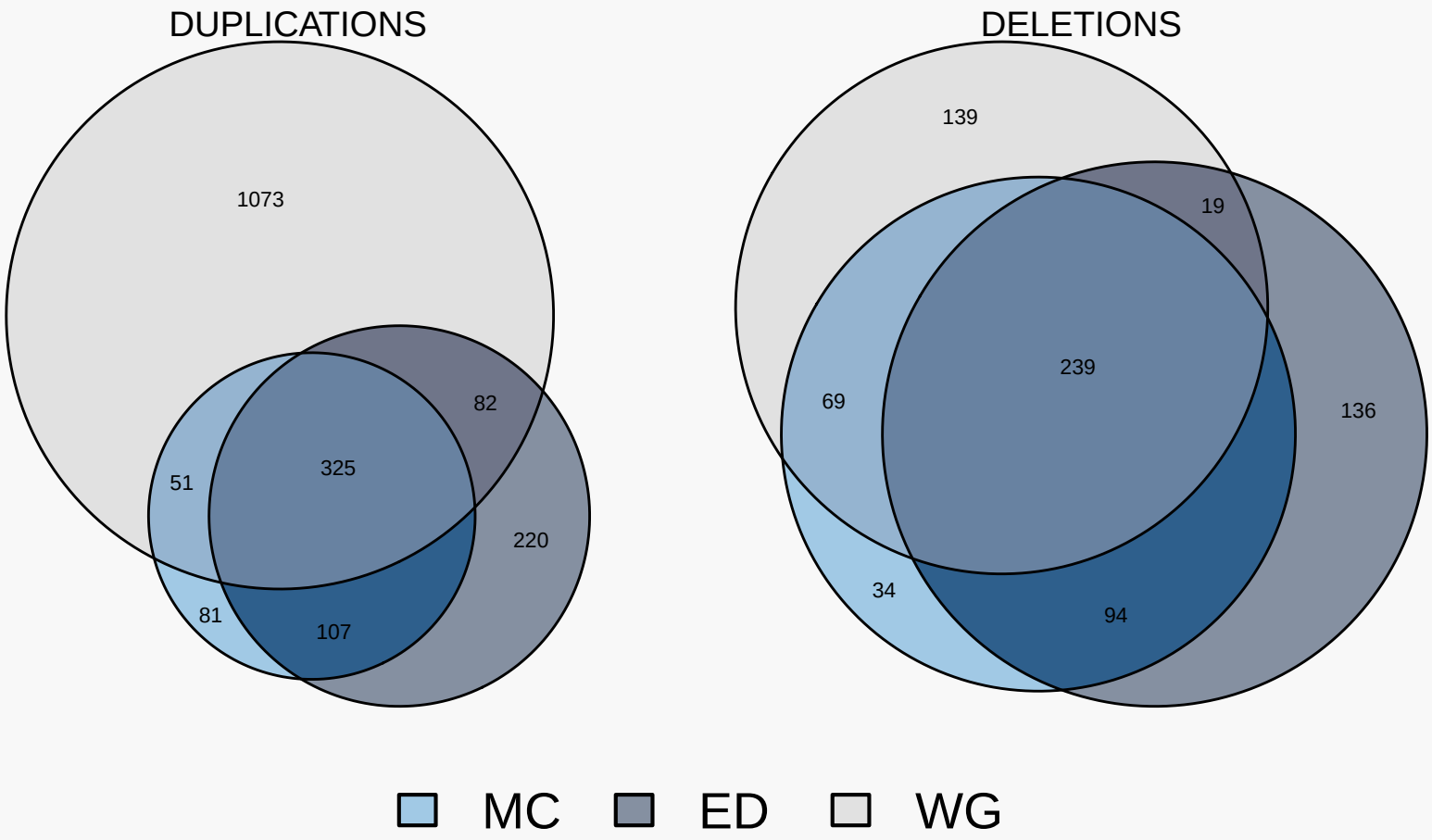
Can we better analyze exomes?

	Total			Duplications			Deletions		
subject	MC	ED	WG	MC	ED	WG	MC	ED	WG
NCG_00012	90	106	143	61	73	121	29	33	22
NCG_00237	82	101	165	50	64	129	32	37	36
NCG_00525	68	74	151	30	33	110	38	41	41
NCG_00593	45	58	142	22	28	81	23	30	61
NCG_00676	66	78	112	38	46	92	28	32	20
NCG_00790	5,156	2,204	121	19	37	92	5,137	2,167	29
NCG_00819	68	76	134	30	41	100	38	35	34
NCG_00840	78	92	157	44	52	115	34	40	42
NCG_00851	1,151	859	141	28	51	102	1,123	808	39
NCG_00857	59	75	119	10	15	81	49	60	38
NCG_00976	46	58	114	25	37	93	21	21	21
NCG_01023	59	95	143	32	60	113	27	35	30
NCG_01043	73	94	128	40	64	105	33	30	23
NCG_01076	36	57	105	7	22	78	29	35	27
NCG_01077	135	157	230	103	121	184	32	36	46
NCG_01117	95	101	154	72	78	129	23	23	25

Can we better analyze exomes?

			MCC	TPR	FDR	PPV
ALL	Total	MC	0.185	0.335	0.897	0.1030
		ED	0.263	0.363	0.809	0.1910
	Sub	MC	0.487	0.345	0.311	0.6890
		ED	0.482	0.378	0.383	0.6170
DUP	Total	MC	0.396	0.236	0.334	0.6660
		ED	0.347	0.240	0.496	0.5040
	Sub	MC	0.404	0.246	0.333	0.6670
		ED	0.384	0.266	0.446	0.5540
DEL	Total	MC	0.180	0.639	0.949	0.0509
		ED	0.219	0.558	0.914	0.0861
	Sub	MC	0.683	0.661	0.294	0.7060
		ED	0.541	0.554	0.471	0.5290

Can we better analyze exomes?



Can we better analyze exomes?

Possibly...

- Found comparable performance overall to ExomeDepth
- However! We do not require prior information

Detecting fetal variation from cell-free DNA

Cell-free DNA used for fetal genetic testing

THE LANCET

Early report

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

Noninvasive testing can now detect:

- aneuploidy and large chromosomal deletions (>5Mb)
- autosomal dominant single-gene disorders
- autosomal recessive single-gene disorders using relative haplotype dosing (RHDO)

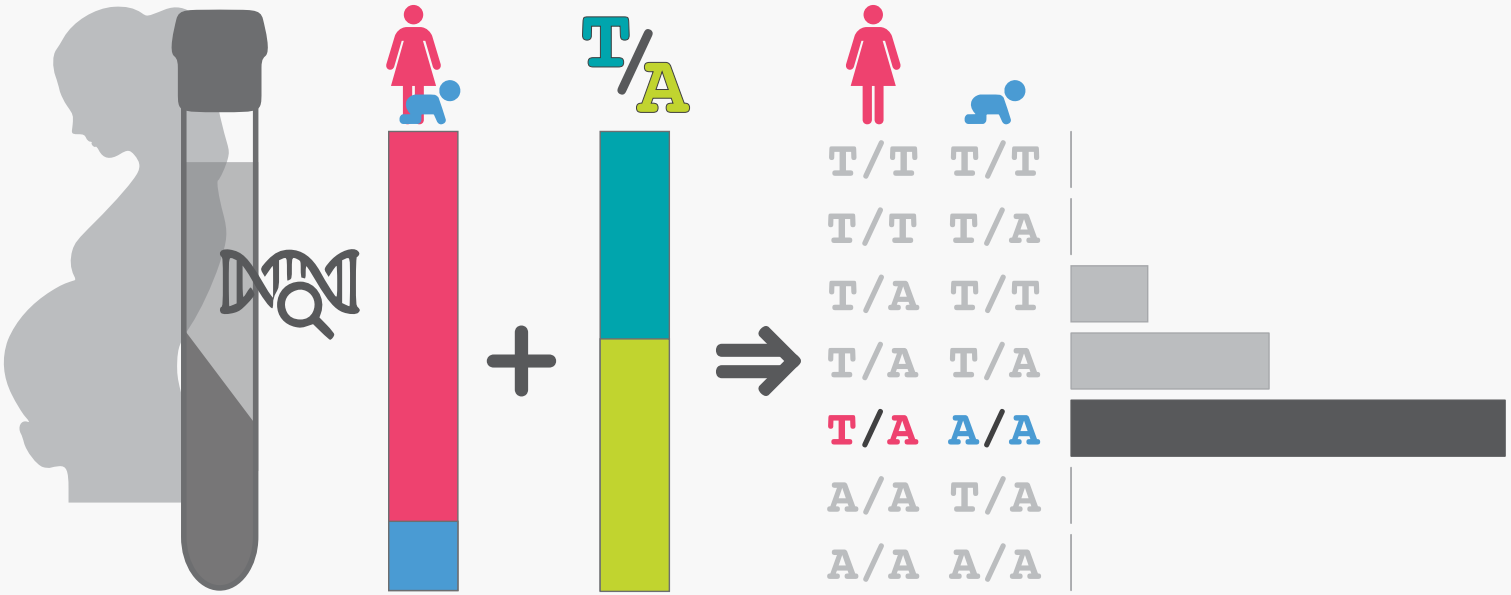
Motivation

No one has demonstrated accurate fetal genotyping from cell-free DNA without additional parental sequencing

Fetal genotyping using only cell-free DNA would

1. enable population-level screening for Mendelian disorders, allowing immediate neonatal management;
2. provide more targets for developing *in utero* therapies.

The two estimation problems



Estimating fetal fraction

Represent maternal and fetal genotype pairs with capital and lowercase letters, where 'A' and 'B' represent the major and minor alleles (e.g. 'AAab' represents the fetus uniquely heterozygous for the minor allele). Define the fetal fraction and PMAR as the random variables F and M.

$$\mathsf{E}[M|G = \mathsf{AAab}, F = f] = \frac{f}{2} \tag{1}$$

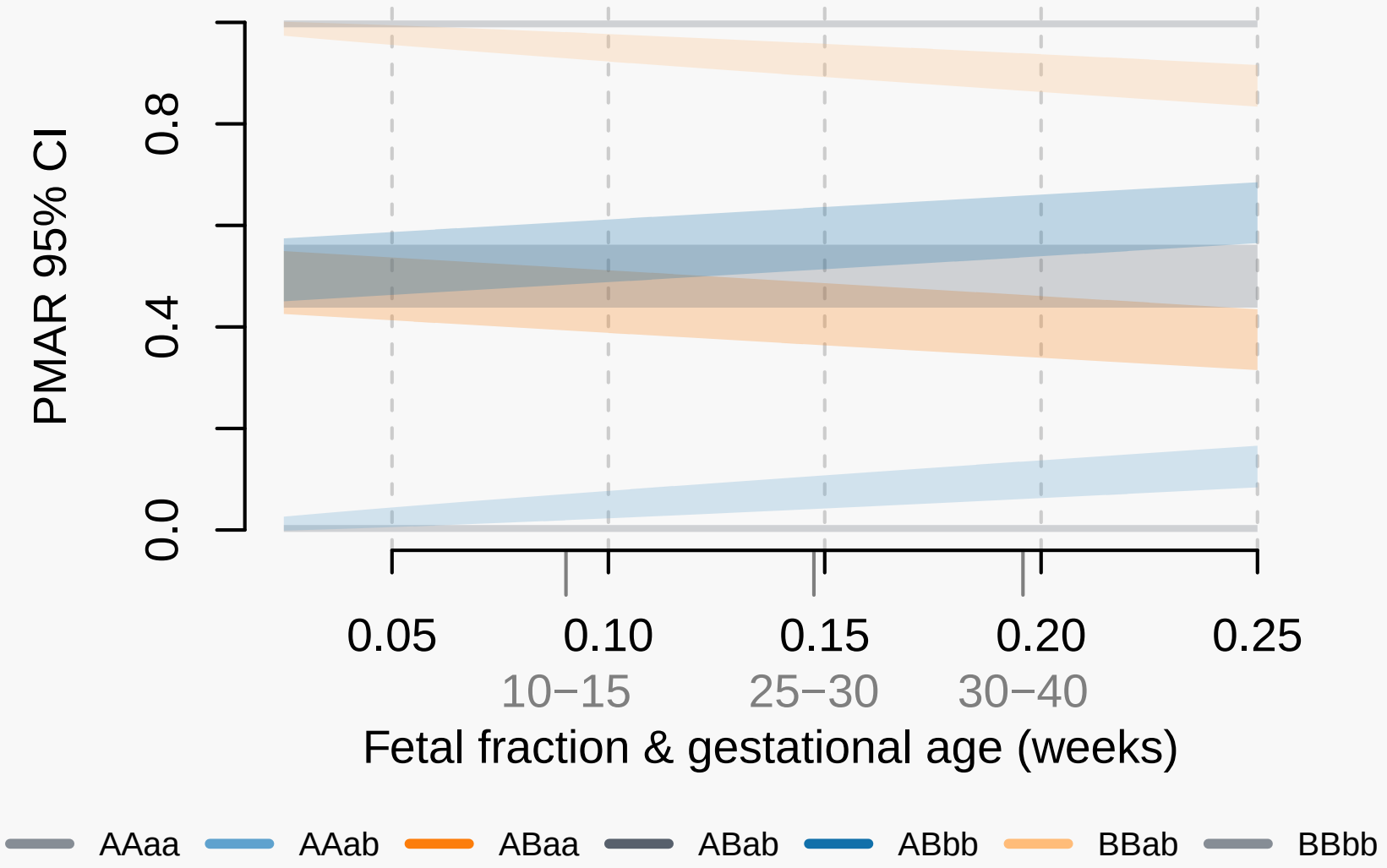
$$\mathsf{E}[M|G = \mathsf{ABaa}, F = f] = \frac{1-f}{2} \tag{2}$$

$$\mathsf{E}[M|G = \mathsf{ABab}, F = f] = \frac{1}{2} \tag{3}$$

$$\mathsf{E}[M|G = \mathsf{ABbb}, F = f] = \frac{1+f}{2} \tag{4}$$

$$\mathsf{E}[M|G = \mathsf{BBab}, F = f] = 1 - \frac{f}{2} \tag{5}$$

Estimating fetal fraction



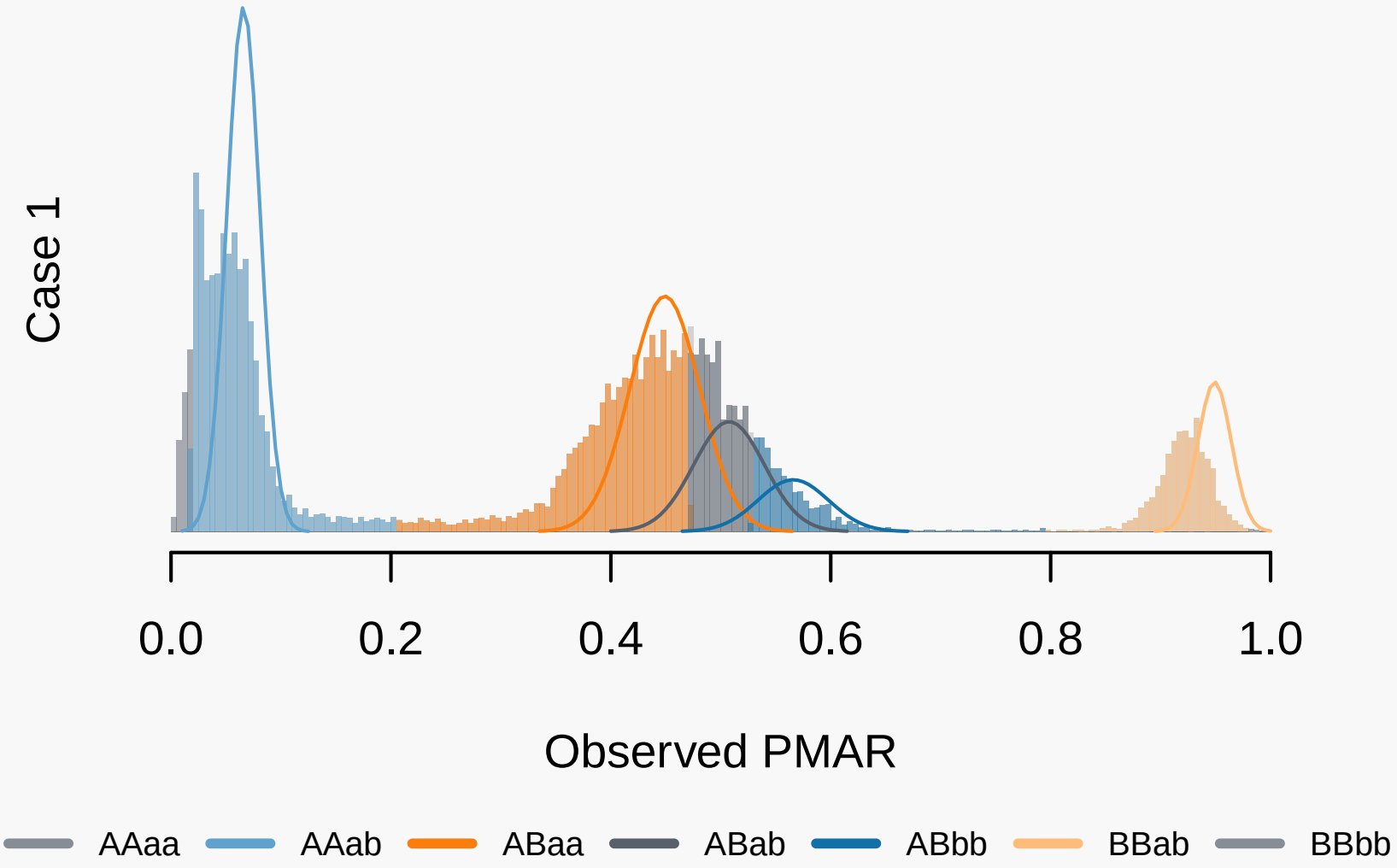
Novel maternal-fetal genotyping algorithm

- Perform empirical Bayes EM routine to identify unique fetal heterozygosity
- Estimate fetal fraction as the median across all sites with fetal heterozygosity
- Estimate maternal-fetal genotypes as the maximal-likelihood given the fetal fraction estimate

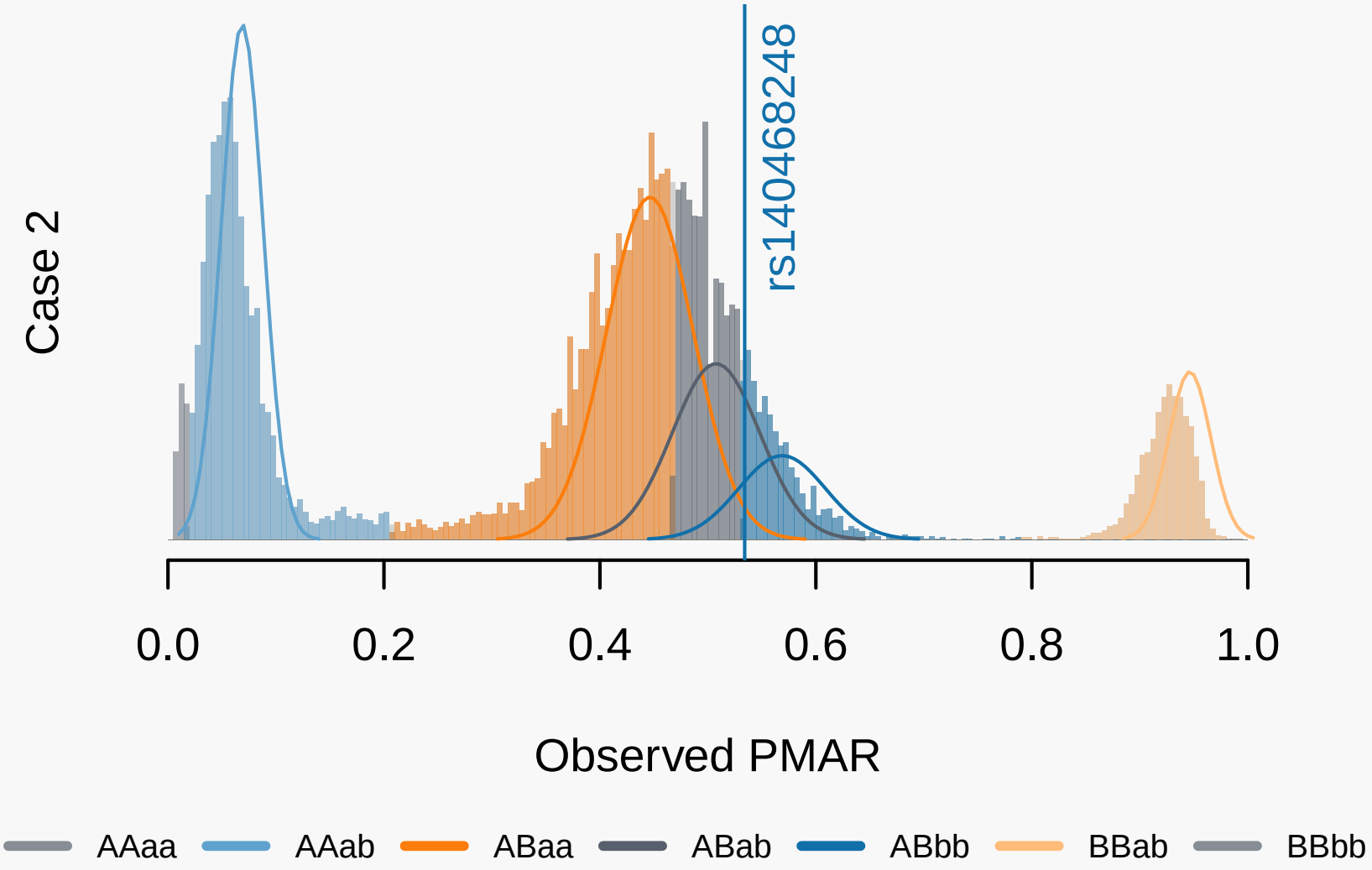
Applying genotyping algorithm to cell-free exome sequencing

	GA	Clinical findings	Genetic diagnosis	FF	Dep	%Dup	%Filt
1	32w2d	5 prior pregnancies affected with X-linked recessive Menke's syndrome	Menke's syndrome; del. ATP7A exon 1	0.117	241	42.80	21.96
2	24w5d	Fetal sonogram at 21w5d showed femoral bowing with shortened length ($<$3% for GA) bilaterally	Osteogenesis imperfecta type VIII; P3H1 c.1120G$>$T (rs140468248)	0.122	152	33.32	22.09
3	34w0d	Fetal sonogram at 19w0d showed bilateral club foot with bilateral upper limb arthrogryposis	None, to date, despite exome and genome sequencing of newborn	0.169	330	53.67	32.65

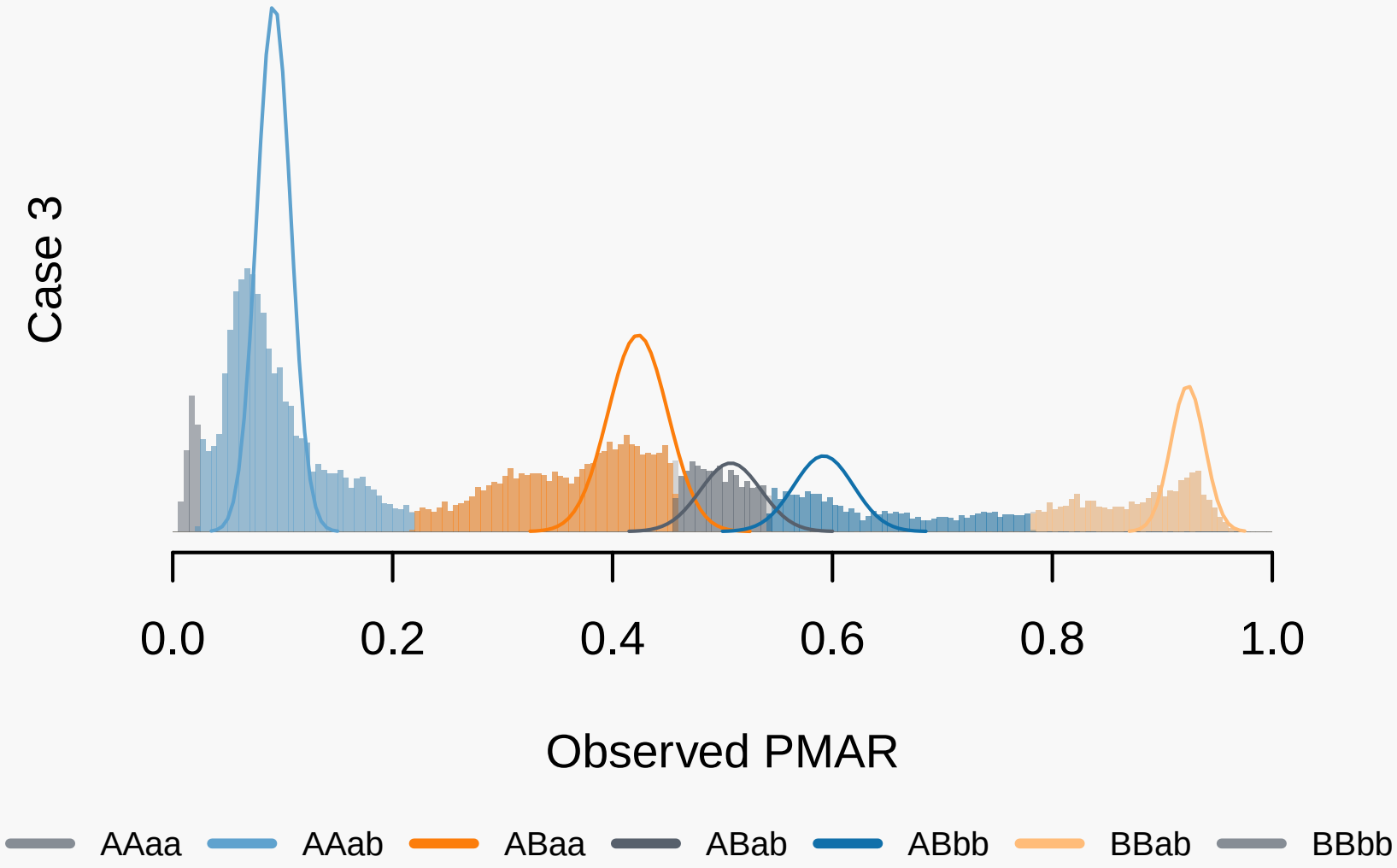
Applying genotyping algorithm to cell-free exome sequencing



Applying genotyping algorithm to cell-free exome sequencing



Applying genotyping algorithm to cell-free exome sequencing

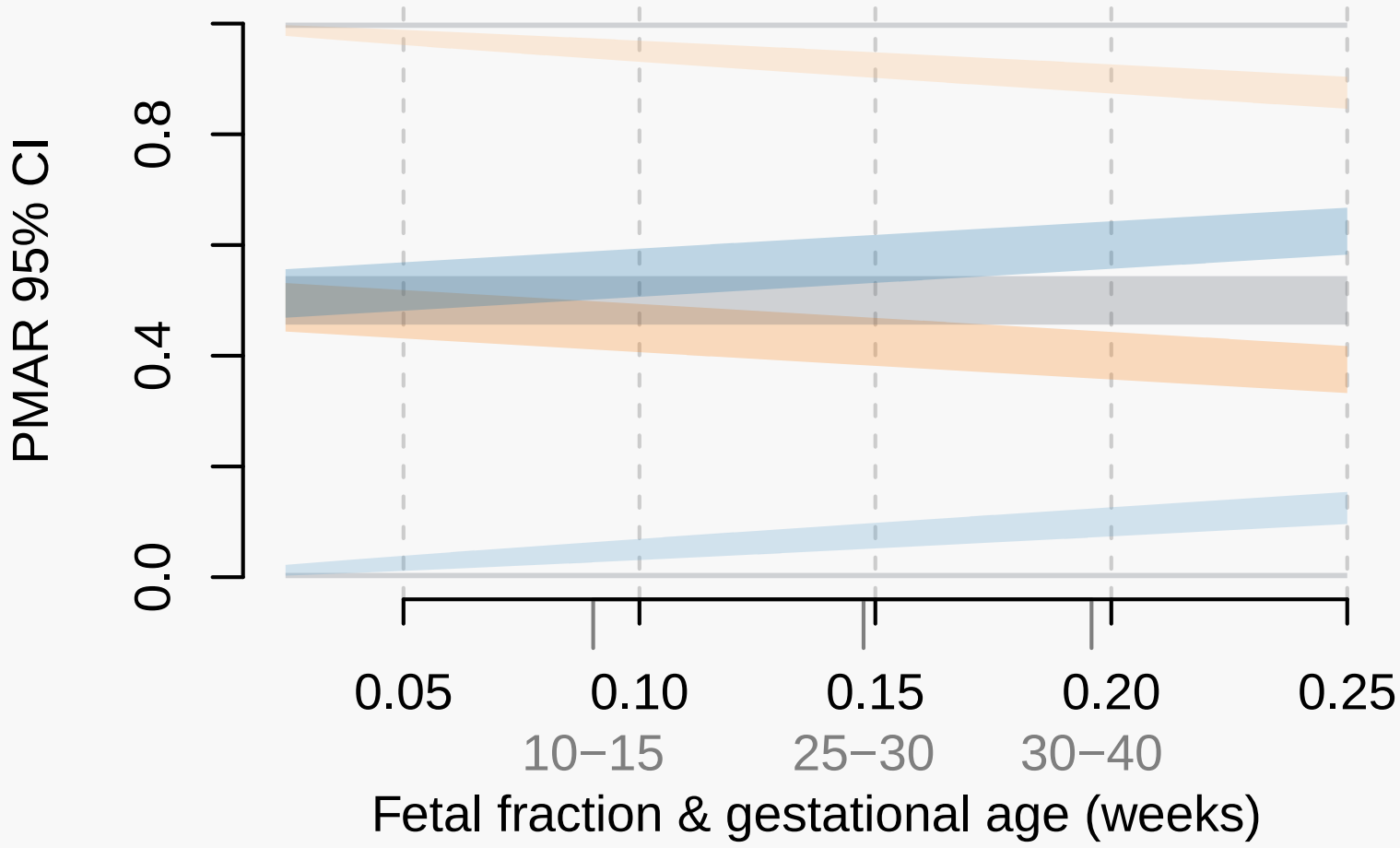


Applying genotyping algorithm to cell-free exome sequencing

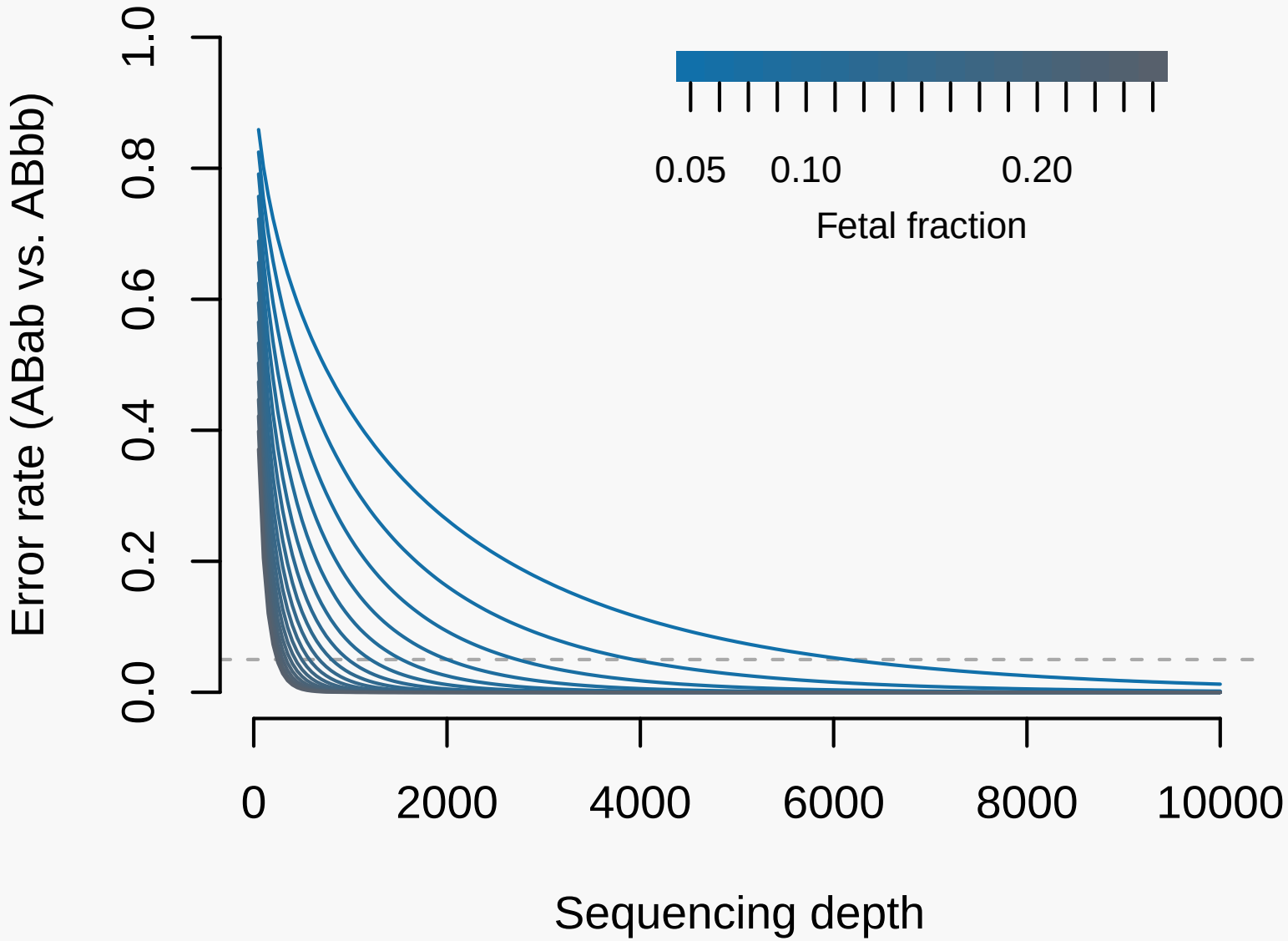
- In Case 3, we also have individual exome sequencing for the mother, father, and fetus
- Overall, we found a 50.91% genotyping accuracy

		Cell-free		
		0/0	0/1	1/1
Fetal	0/0	1,063	1,857	9
	0/1	3,598	7,079	1,454
	1/1	76	2,197	1,391

Explaining the poor performance



Explaining the poor performance



Does fragment length matter?

Clinical Chemistry 50:1
88–92 (2004)

Molecular Diagnostics
and Genetics

Size Distributions of Maternal and Fetal DNA in Maternal Plasma

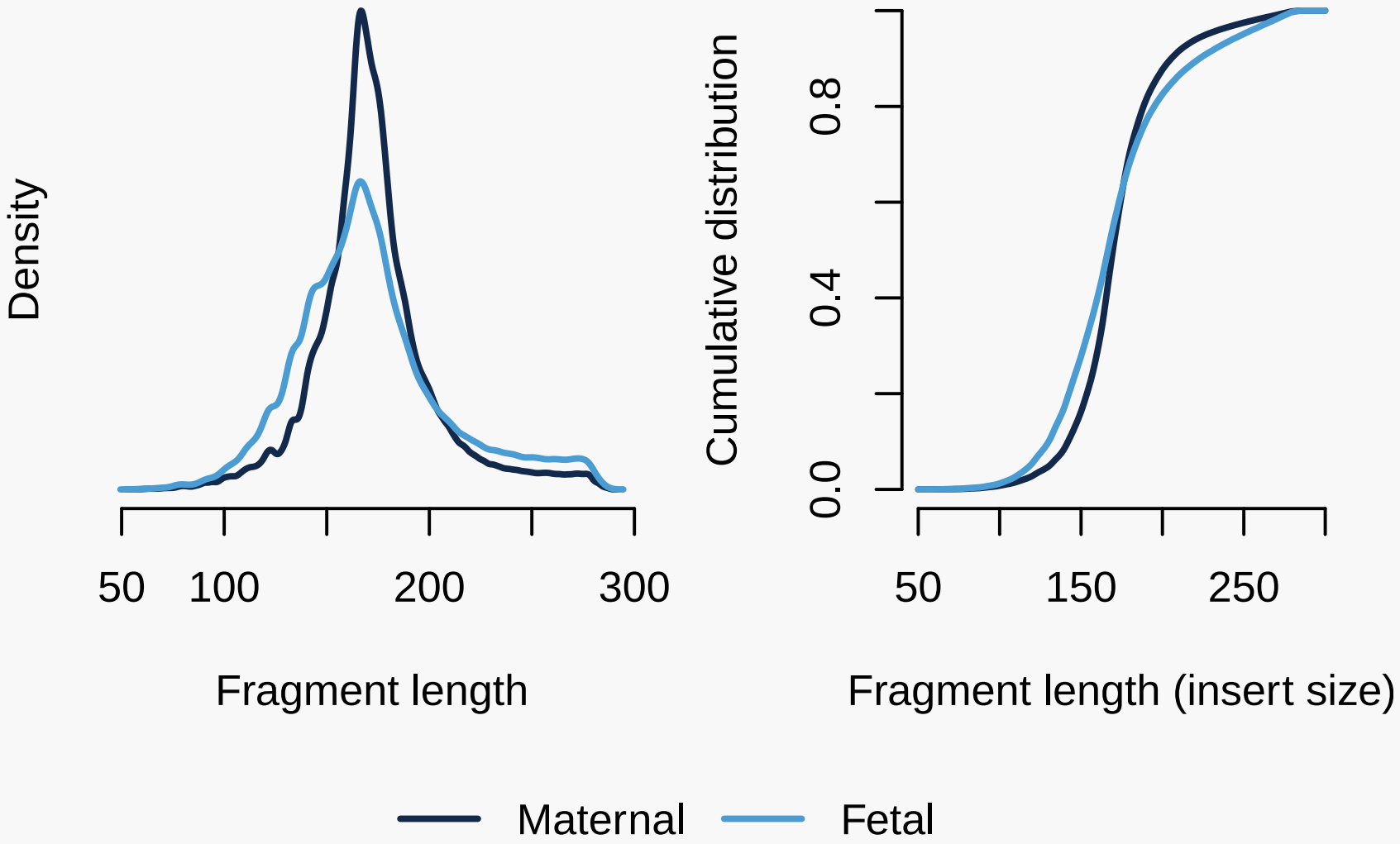
K.C. Allen Chan,[1†] Jun Zhang,[1†] Angela B.Y. Hui,[2] Nathalie Wong,[3] Tze K. Lau,[4] Tse N. Leung,[4] Kwok-Wai Lo,[2] Dolly W.S. Huang,[3] and Y.M. Dennis Lo[1*]

Method

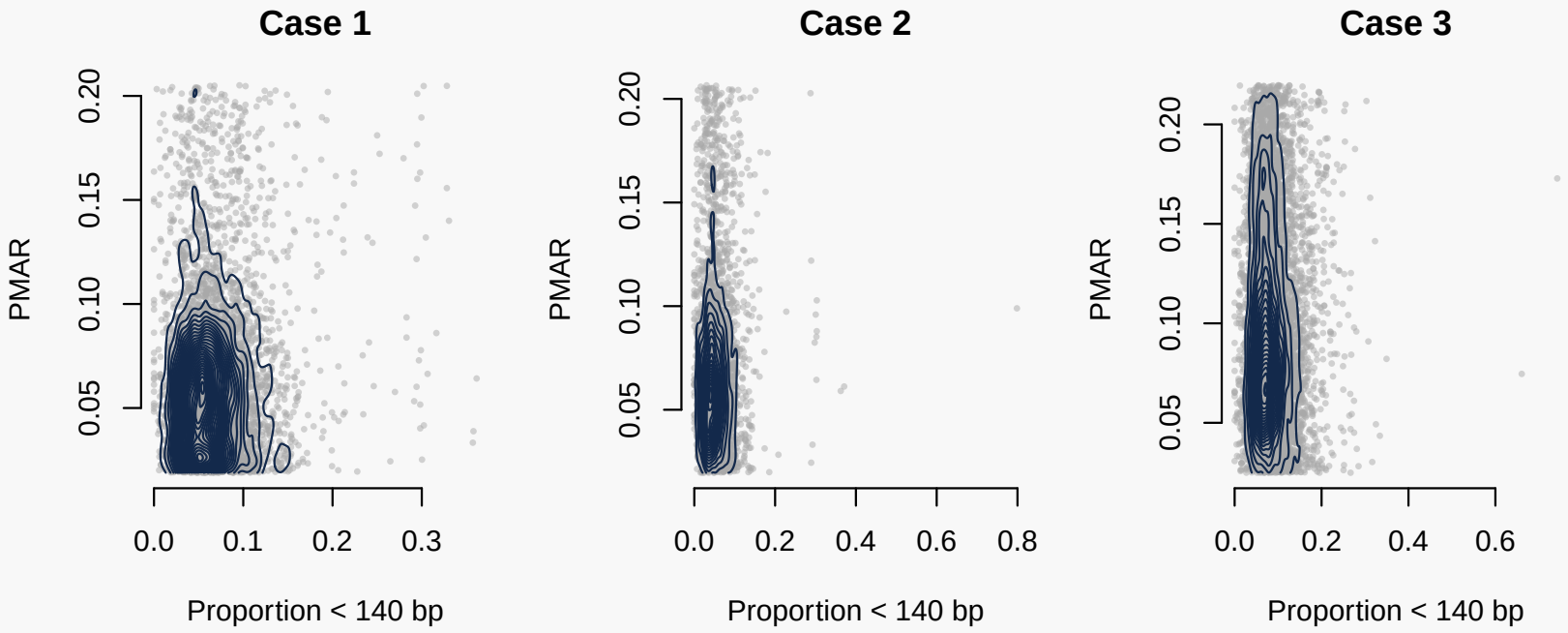
Bayesian-based noninvasive prenatal diagnosis of single-gene disorders

Tom Rabinowitz,[1] Avital Polsky,[1] David Golan,[2] Artem Danilevsky,[1] Guy Shapira,[1] Chen Raff,[1] Lina Basel-Salmon,[1,3] Reut Tomashov Matar,[3] and Noam Shomron[1]

Does fragment length matter?



Does fragment length matter?



- Rabinowitz et al. found the difference in accuracy varied from -0.25% to 1.89% when using versus not using fragment length in their exome analyses.
- Correcting for fragment length will not overcome the bounds of the binomial distribution

Summary

- Noninvasive exome sequencing would require cost-prohibitive sequencing depths
- Despite suggestions by others, simply correcting for fragment length will not facilitate noninvasive genome/exome sequencing in the clinic
- We recommend a more targeted approach

Acknowledgments

- Kirk Wilhelmsen
  - Fenshen Kuo
  - Chris Bizon
  - Jeff Tilson
  - Darius Bost
  - Kimberly Robasky
  - Phil Owen
- Thesis Committee
  - Bradford Powell (Chair)
  - Stan Ahalt
  - Yun Li
  - Neeta Vora
- Jonathan Berg
  - Christian Tilley
  - Alicia Brandt
- UNC Genetics/BCB
  - Will Valdar
  - Tim Elston
  - Jonathan Cornett
  - Cara Marlow
  - Fernando Pardo-Manuel de Villena
- MD/PhD Program
  - Toni Darville
  - Mohanish Deshmuk
  - Alison Reagan
  - Carol Herrion
  - Amber Brosius

renci

UNC | SCHOOL OF MEDICINE

Appendix

Appendix

Median count per exon

Produce simulated exomes

mcCNV algorithm

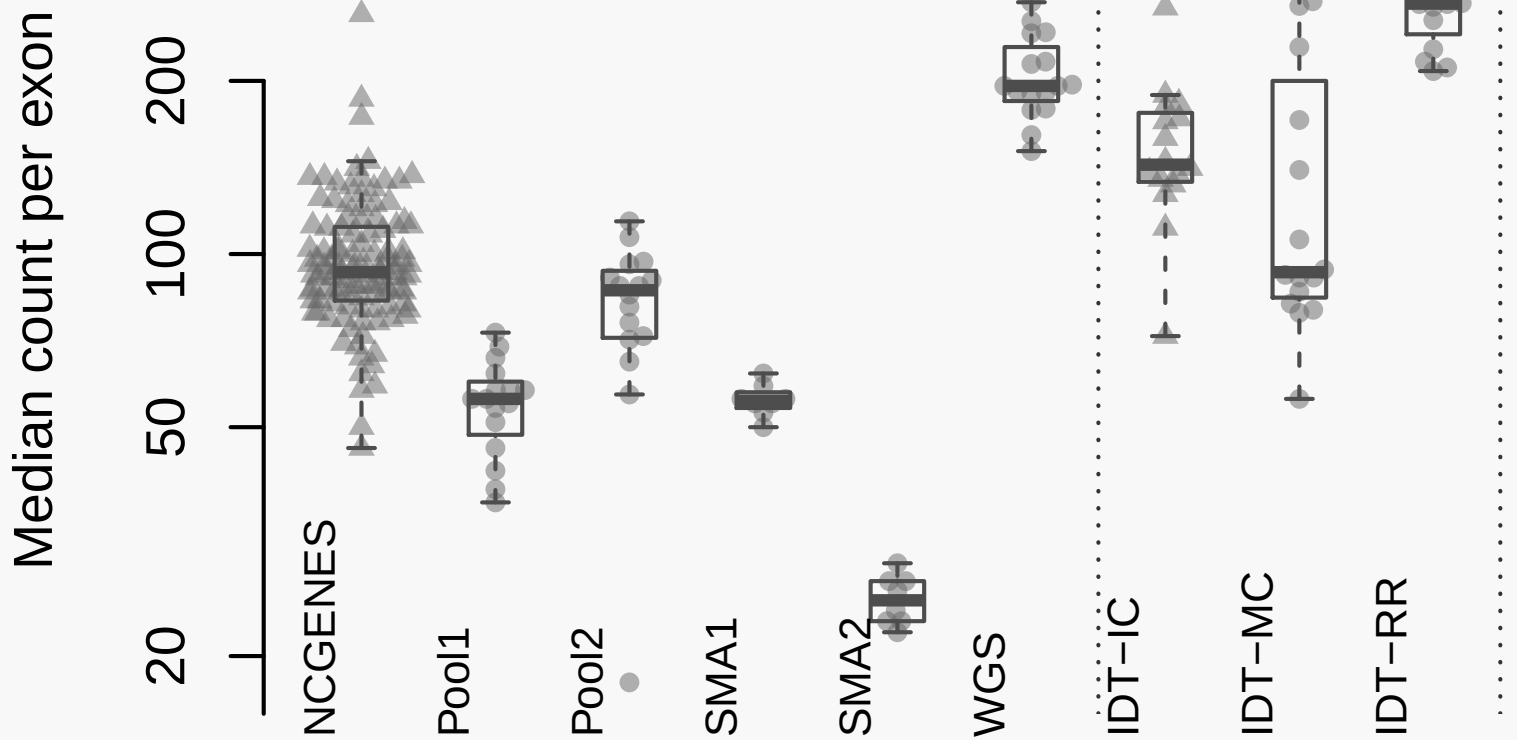
Venn diagrams with all samples

Aglient capture versus IDT capture

Maternal-fetal genotyping algorithm

All case 3 calls

Median count per exon



Produce simulated exomes

We use an alternate definition of copy state, such that 1 represents the normal diploid state. Let N represent the total number of molecules (read pairs) and $e_j \in \mathbb{E}$ represent the probability of capturing target j, then for each subject, i:

1. Randomly select $s_{ij} \in \mathbb{S}_i$ from $S = \{0.0, 0.5, 1, 1.5, 2\}$ as the copy number at target j
2. Adjust the subject-specific capture probabilities by the copy number, $\mathbb{E}_i = \frac{\mathbb{E} \odot \mathbb{S}_i}{\sum_j \mathbb{E} \odot \mathbb{S}_i}$
3. Draw N times from Multinomial($\mathbb{E}_i$), giving the molecule counts at each target j for sample i, $c_{ij} \in \mathbb{C}_i$

mcCNV algorithm

The mcCNV algorithm adjusts the sSEQ probability model by adding a multiplier for the copy state:

$$C_{ij} \sim \mathcal{NB}(f_i s_{ij} \hat{\mu}_j, \tilde{\phi}_j / f_i)$$

where the random variable C_{ij} represents observed molecule counts for subject i at target j, f_i is the size factor for subject i, s_{ij} is the copy state, μ_j is the expected mean under the diploid state at target j, and $\tilde{\phi}_j$ is the shrunken phi at target j. We observe c_{ij} and wish to estimate s_{ij}, $\hat{s}_{ij}$.

mcCNV algorithm

Initialize by setting $\hat{s}_{ij} = 1$ for all i, j. Then,

1. Adjust the observed values for the estimated copy-state,

$$c'_{ij} = \frac{c_{ij}}{\hat{s}_{ij}}.$$

2. Subset c'_{ij} such that $c'_{ij} > 10,\ \hat{s}_{ij} > 0$
3. Calculate the size-factor for each subject

$$f_i = \text{median}\left(\frac{c'_{ij}}{g_j}\right),$$

where g_j is the geometric mean at each exon.

4. Use method of moments to calculate the expected dispersion

$$\hat{\phi}_j = \max\left(0, \frac{\hat{\sigma}_j^2 - \hat{\mu}_j}{\hat{\mu}_j^2}\right)$$

where $\hat{\mu}_j$ and $\hat{\sigma}_j^2$ are the sample mean and variance of c'_{ij}/f_i.

mcCNV algorithm

5. Let J represent the number of targets. Shrink the phi values to

$$\tilde{\phi}_j = (1-\delta)\hat{\phi}_j + \delta\hat{\xi}$$

such that

$$\delta = \frac{\sum_j \left(\hat{\phi}_j - \frac{1}{n_j}\sum_j \hat{\phi}_j\right)^2 /(J-1)}{\sum_j \left(\hat{\phi}_j - \hat{\xi}\right)^2 /(n_j - 2)}$$

and

$$\hat{\xi} = \underset{\xi}{\text{argmin}} \left\{ \frac{d}{d\xi} \frac{1}{\sum_j \left(\hat{\phi}_j - \xi\right)^2} \right\}.$$

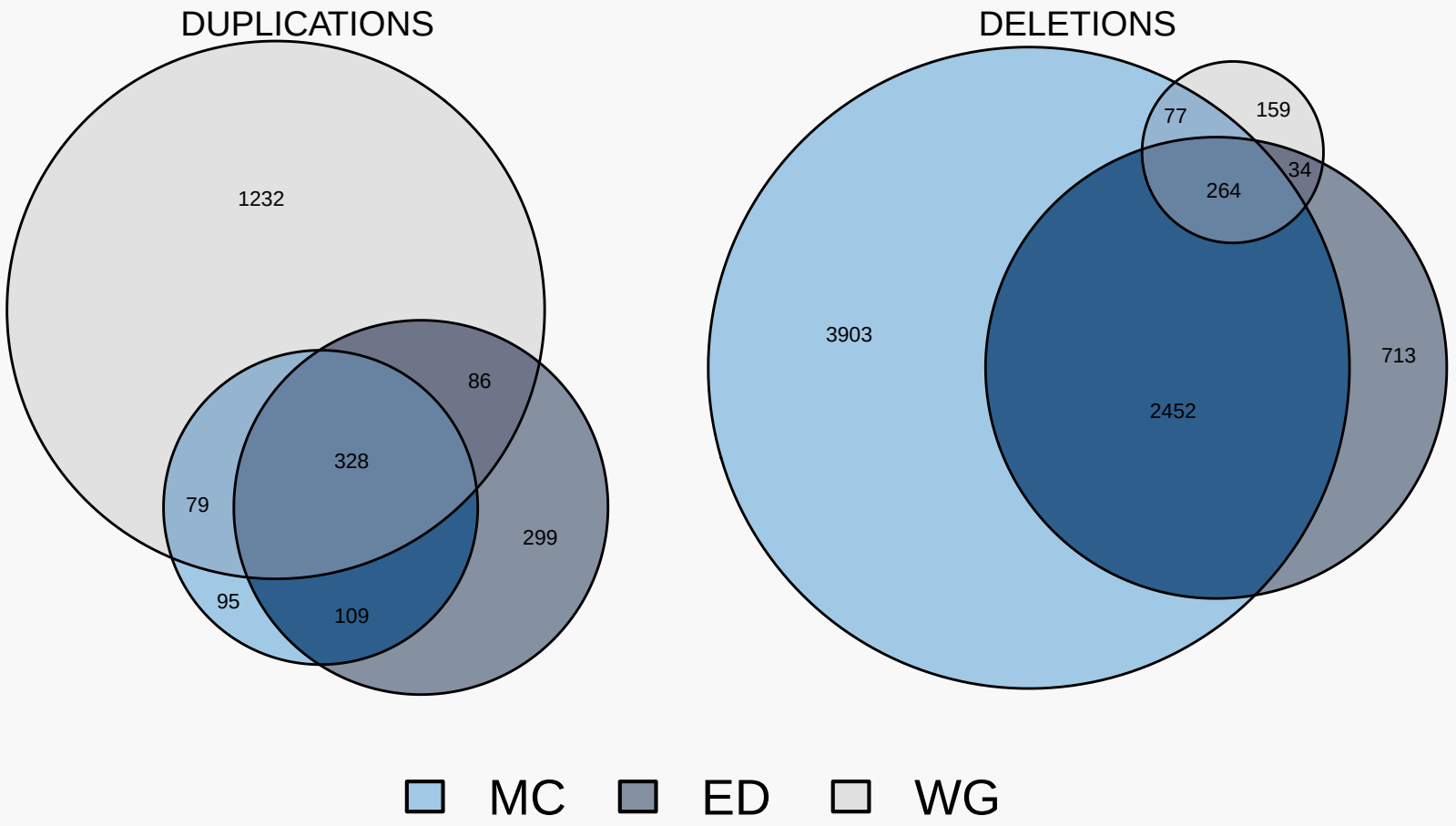
mcCNV algorithm

6. Update $\hat{s}_{ij}$,

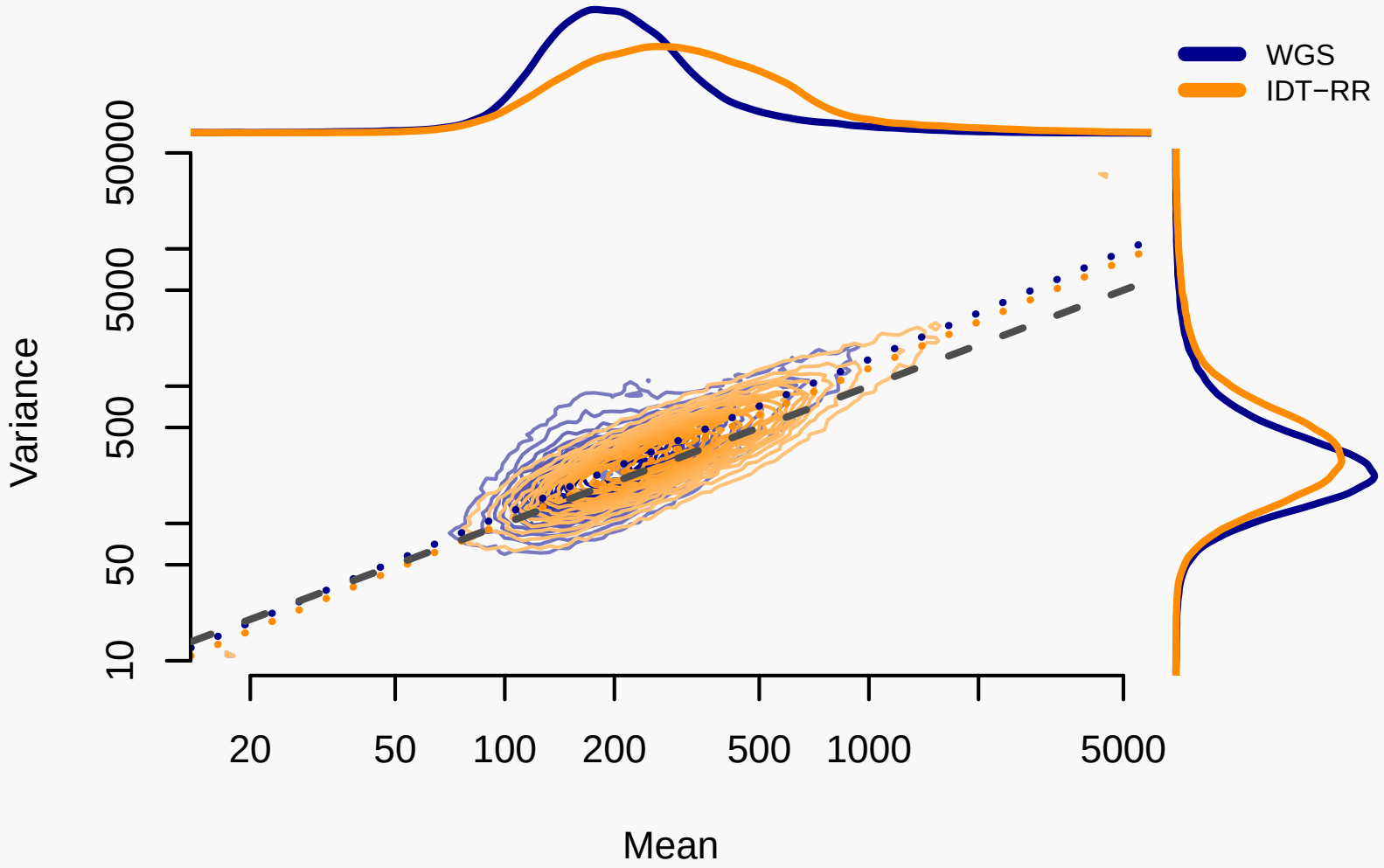
$$\underset{s \in S}{\text{argmax}} \left\{ \mathcal{L}(s | c_{ij}, f_i, \hat{\mu}_j, \tilde{\phi}_j) \right\}$$

   where $S = \{0.001, 0.5, 1, 1.5, 2\}$.
7. Repeat until the number of changed states falls below a threshold or a maximum number of iterations is reached.
8. After convergence, calculate p-values for the diploid state, $\pi_{ij} = \text{Pr}(s_{ij} = 1)$.
9. Adjust p-values using the Benjamini–Hochberg procedure and filter to a final call-set such that adjusted p-values fall below some threshold, α.

Venn diagrams with all samples



Aglient capture versus IDT capture



Maternal-fetal genotyping algorithm

Represent maternal and fetal genotype pairs, given by the random variable G, with capital and lowercase letters, where 'A' and 'B' represent the major and minor alleles (e.g. 'AAab' represents the fetus uniquely heterozygous for the minor allele).
Let X, Y be random variables for major and minor allele read counts. Define the fetal fraction and PMAR as the random variables F and M. We employ an empirical Bayesian expectation-maximization algorithm to identify loci with unique fetal heterozygosity, i.e. $g \in \{\text{AAab}, \text{BBab}\}$. We pick reasonable starting values for the fetal fraction, $F = f$, and the average minor allele frequency, then iteratively update the expected allele distribution and expected PMAR values until some convergence:

Maternal-fetal genotyping algorithm

1. Initialize the genotype probabilities, $p_g^* = \text{Pr}\{G = g\}$, and the expected PMAR, $m_g^* = m_g$, based on reasonable estimates for the average minor allele frequency and fetal fraction
2. Update $\hat{\mathbb{G}}$:

$$\hat{g}_i = \underset{g \in G}{\text{argmax}} \left\{ p_g^* \mathcal{L}(g | m_g^*, x_i, y_i) \right\}, Y_i \sim \text{Bin}(x_i + y_i, m_g^*)$$

3. Update the genotype probabilities:

$$p_g^* = \frac{\sum_i \text{I}(\hat{g} = g) + N\text{Pr}\{G = g\} - 1}{\sum_g \left\{ \sum_i \text{I}(\hat{g} = g) + N\text{Pr}\{G = g\} - 2 \right\}}$$

where N is the weight given to the initial estimate of the genotype probability, $\text{Pr}\{G = g\}$.

Maternal-fetal genotyping algorithm

4. Update the expected PMAR:

$$m_g^* = \frac{\sum_i y_i \mathsf{I}(\hat{g} = g) + N m_g - 1}{\sum_i (x_i + y_i)\mathsf{I}(\hat{g} = g) + N - 2}$$

   where N is the weight given to the initial estimate of the PMAR, m_g.
5. Continue updating $\hat{\mathbb{G}}$, p_g^*, and m_g^* until $\hat{\mathbb{G}}$ converges.
6. For all loci j, such that $\hat{g} \in \{\mathsf{AAab}, \mathsf{BBab}\}$, calculate $\hat{f}_j$:

$$\hat{f}_j = \begin{cases} \dfrac{2y_j}{x_j + y_j}, & \hat{g} = AAab \\ 2 - \dfrac{2y_j}{x_j + y_j}, & \hat{g} = BBab \end{cases}$$

Maternal-fetal genotyping algorithm

7. Let
$$\hat{f} = \text{median}\left(\hat{f}_j\right)$$
8. Calculate the expected PMAR using the fetal fraction estimate,
$$m_g = \mathsf{E}[M|\hat{f}, g]$$
9. Finally, for all loci, i, estimate $\hat{g}_i \in \hat{\mathbb{G}}$,
$$\hat{g}_i = \underset{g \in G}{\text{argmax}} \left\{ \mathcal{L}(g|m_g, x_i, y_i) \right\}, Y_i \sim \text{Bin}(x_i + y_i, m_g)$$

All case 3 calls

Mat	Fet	Cff	N
0/0	0/0	0/0	468
		0/1	1,159
	0/1	0/0	64
		0/1	352
	1/1	0/0	1
		0/1	2
1/1	0/0	0/1	1
		1/1	2
	0/1	0/0	3
		0/1	1,308
		1/1	648
	1/1	0/1	1,601
		1/1	23

Mat	Fet	Cff	N
0/1	0/0	0/0	387
		0/1	107
		1/1	6
	0/1	0/0	3,072
		0/1	1,967
		1/1	713
	1/1	0/0	68
		0/1	458
		1/1	1,291