Reducing INDEL calling errors in whole genome and exome sequencing data.

Han Fang

November 8, 2014
CSHL Biological Data Science Meeting

Acknowledgments

Lyon Lab

Yiyang Wu
Jason O'Rawe
Laura J Barron
Max Doerfel
Constantine Hartofilis

Gholson Lyon

Michael Schatz

Schatz Lab

Giuseppe Narzisi
Hayan Lee
James Gurtowski
Tyler Garvin
Maria Nattestad
Srividya Ramakrishnan

Colleagues from Cold Spring Harbor Laboratory

Sara Ballouz
Wim Verleyen
Jesse Gillis
Shane McCarthy

Eric Antoniou
Elena Ghiban
Melissa Kramer
Stephanie Muller
Senem M Eskipehlivan

Michael Wigler
Ivan Iossifov
Michael Ronemus
Julie Rosenbaum
Rob Aboukhalil

IT department

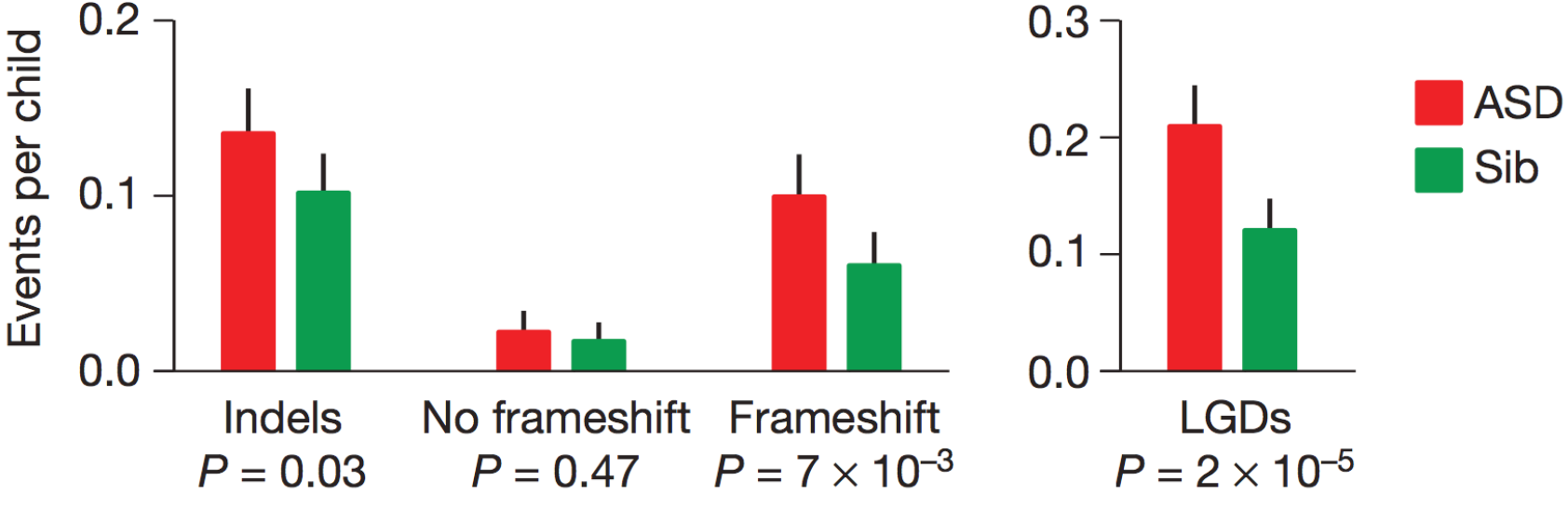
CSH Stanley Institute for Cognitive Genomics
Cold Spring Harbor Laboratory

SFARI
Simons Foundation Autism Research Initiative

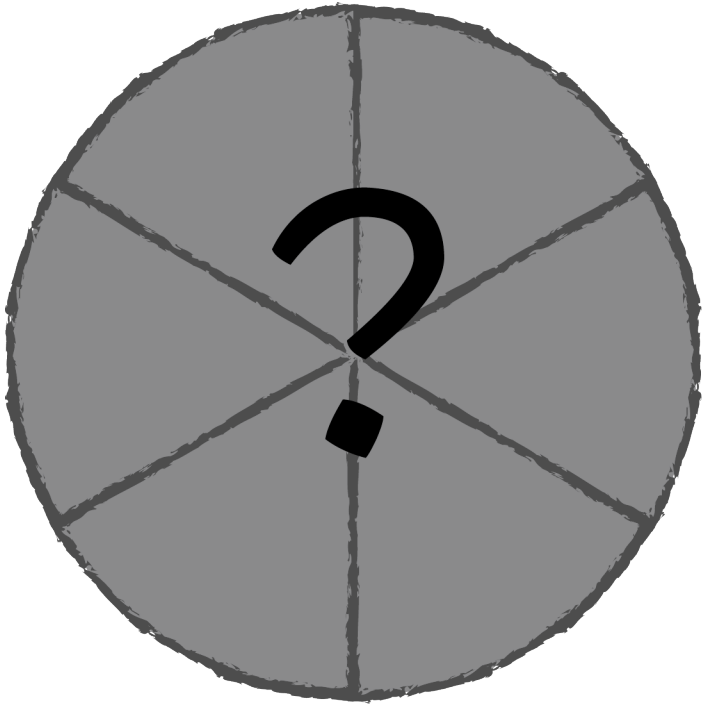
NSF

Stony Brook University

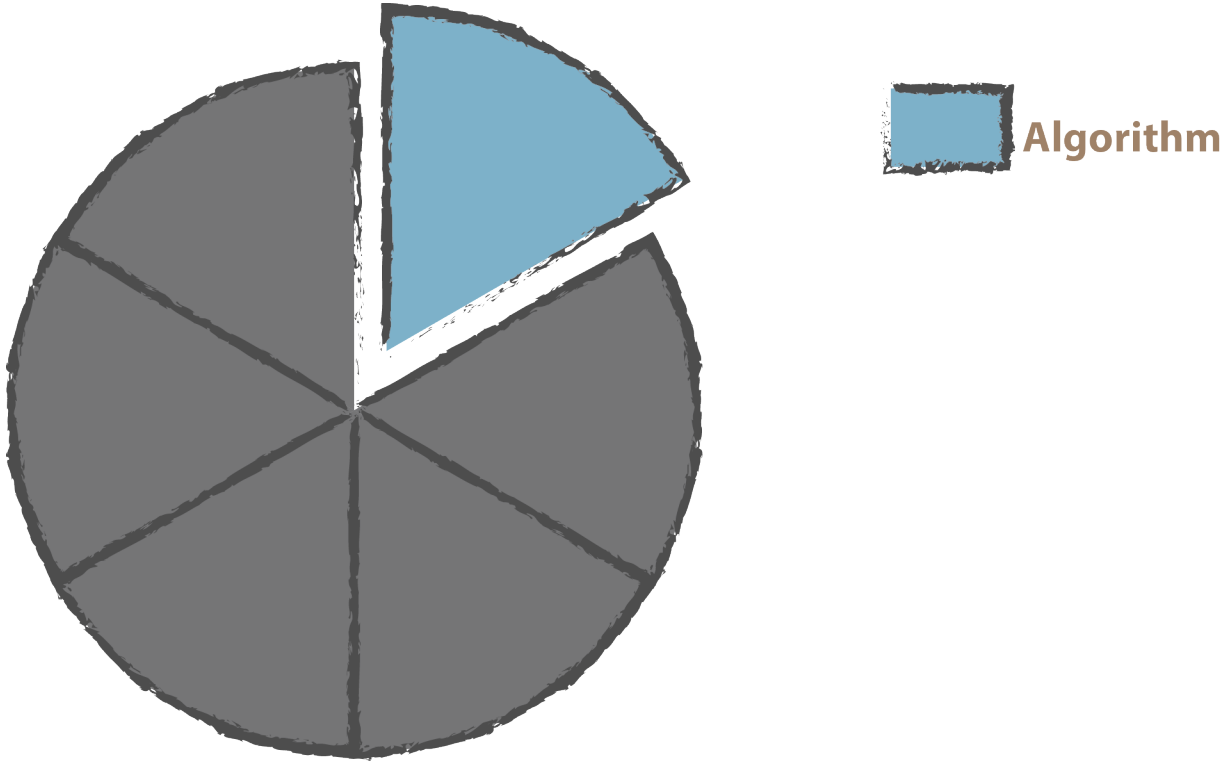
Significantly higher rates of *de novo* frame-shifts & LGDs in the affected vs. unaffected siblings



The contribution of de novo coding mutations to autism spectrum disorder.
Iossifov I, O'Roak BJ, Sanders SJ, Ronemus M, et al. (2014) *Nature.* doi:10.1038/nature13908

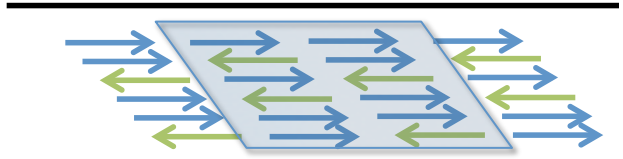


Sources of INDEL calling errors?

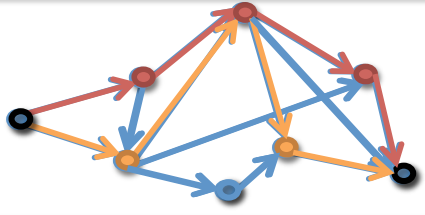

Algorithm

Scalpel: Haplotype Microassembly

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



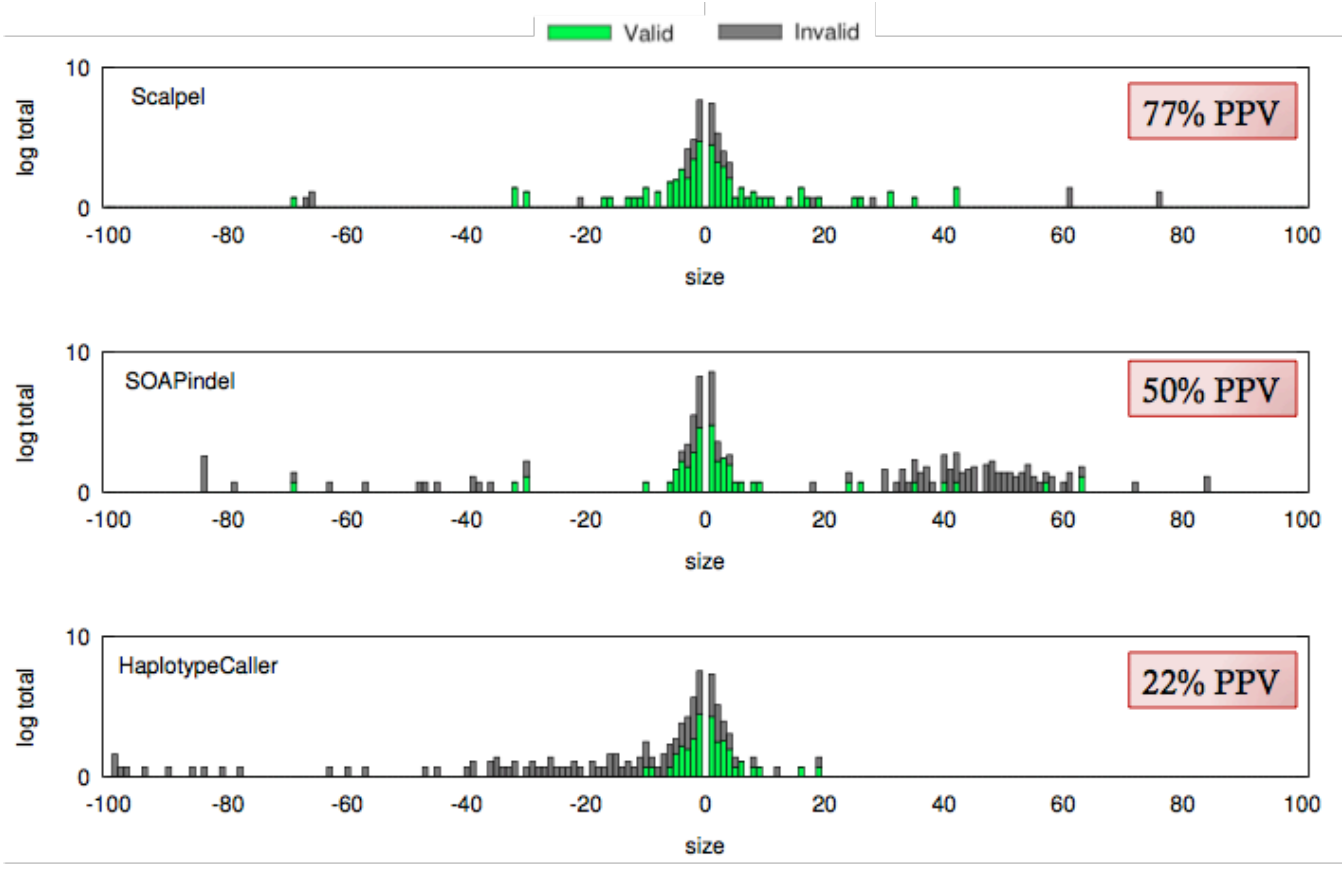
- Decompose reads into overlapping *k*-mers and construct de Bruijn graph from the reads.
- Find end-to-end haplotype paths spanning the region.

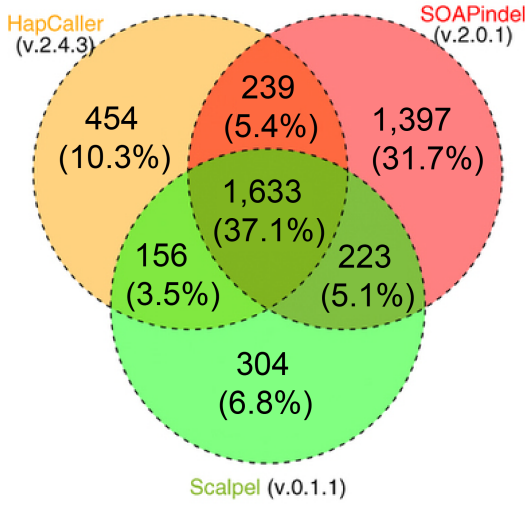


- Align assembled sequences to reference to detect mutations.

Accurate de novo and transmitted indel detection in exome-capture data using microassembly.
Narzisi G, O'Rawe JA, Iossifov I, Fang H, Lee YH, Wang Z, Wu Y, Lyon GJ, Wigler M, Schatz MC (2014)
Nature Methods. doi: 10.1038/nmeth.3069

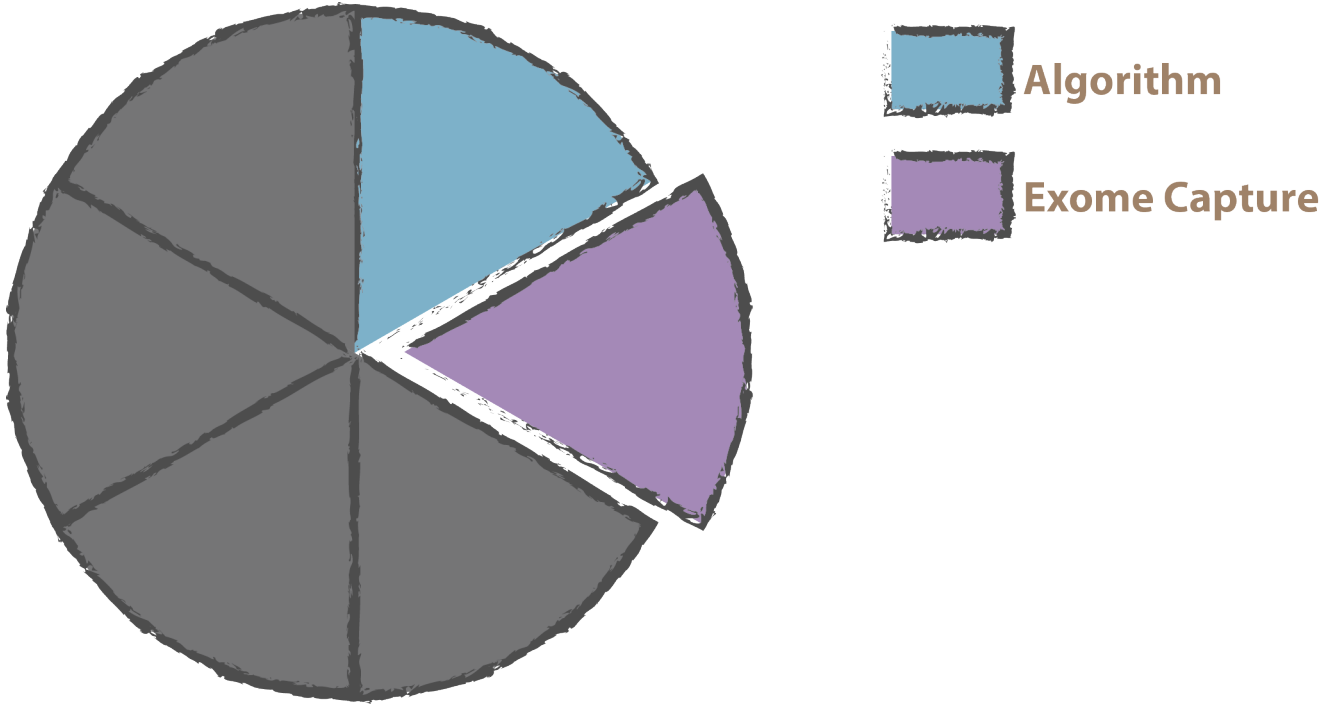
Scalpel INDEL Validation



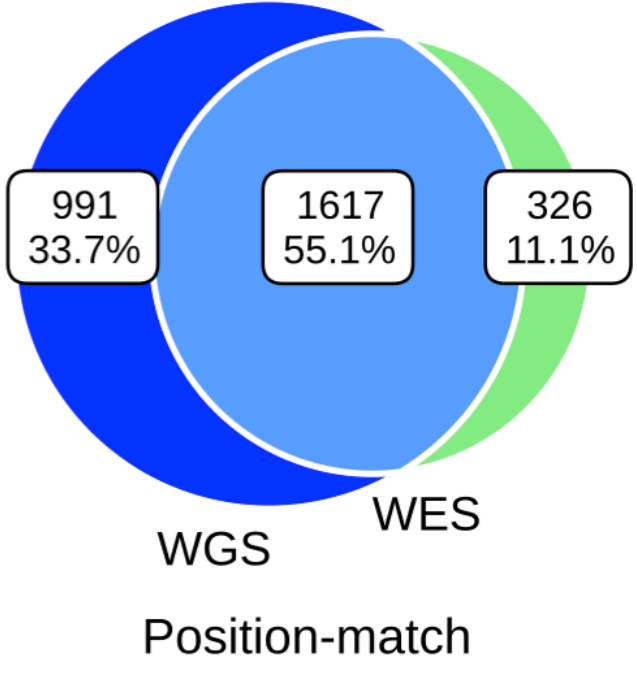


1000 INDELs selected for validation

- 200 Scalpel-specific
- 200 GATK HapCaller-specific
- 200 SOAPindel-specific
- 200 within the intersection
- 200 long indels (>30bp)


Algorithm
Exome Capture

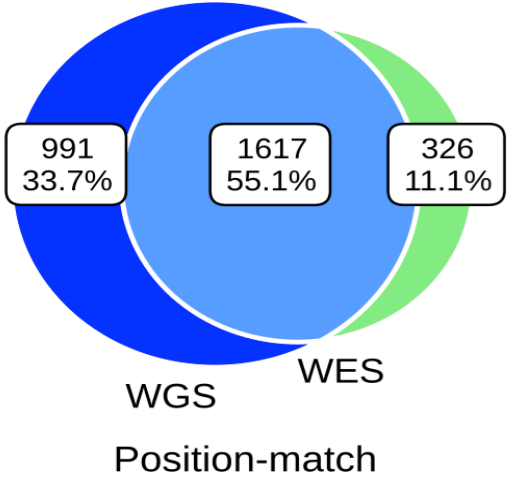
Concordance between WGS and WES data.



Reducing INDEL errors in whole genome and exome sequencing data.
Fang H, Wu Y, Narzisi G, O'Rawe JA, Jimenez Barrón LT, Rosenbaum J, Ronemus M, Iossifov I, Schatz MC*, Lyon, GJ* (2014) *Genome Medicine.* doi: 10.1186/s13073-014-0089-z

Validation results

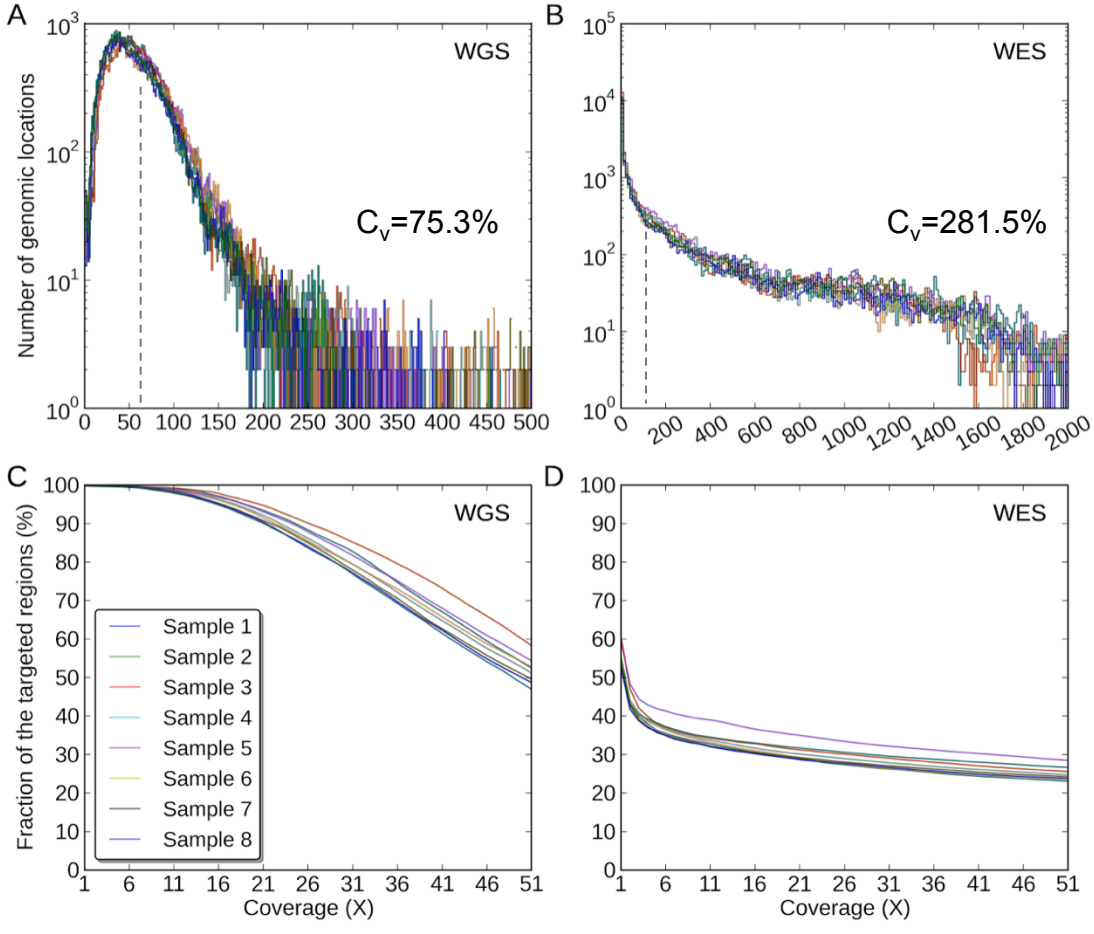
- The validation rate of WGS-WES intersection INDELs was in fact very high (95%).
- Accuracy of INDEL detection with WES is much lower than that with WGS.
- The WES-specific set had a much smaller fraction of large INDELs.

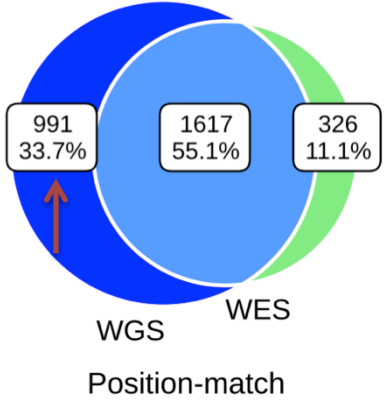


	INDELs	Valid	PPV	INDELs (>5bp)	Valid (>5bp)	PPV (>5bp)
WGS-WES intersection	160	152	95.0%	18	18	100%
WGS-specific	145	122	84.1%	33	25	75.8%
WES-specific	161	91	56.5%	1	1	100%

Example of WES missing a large INDEL

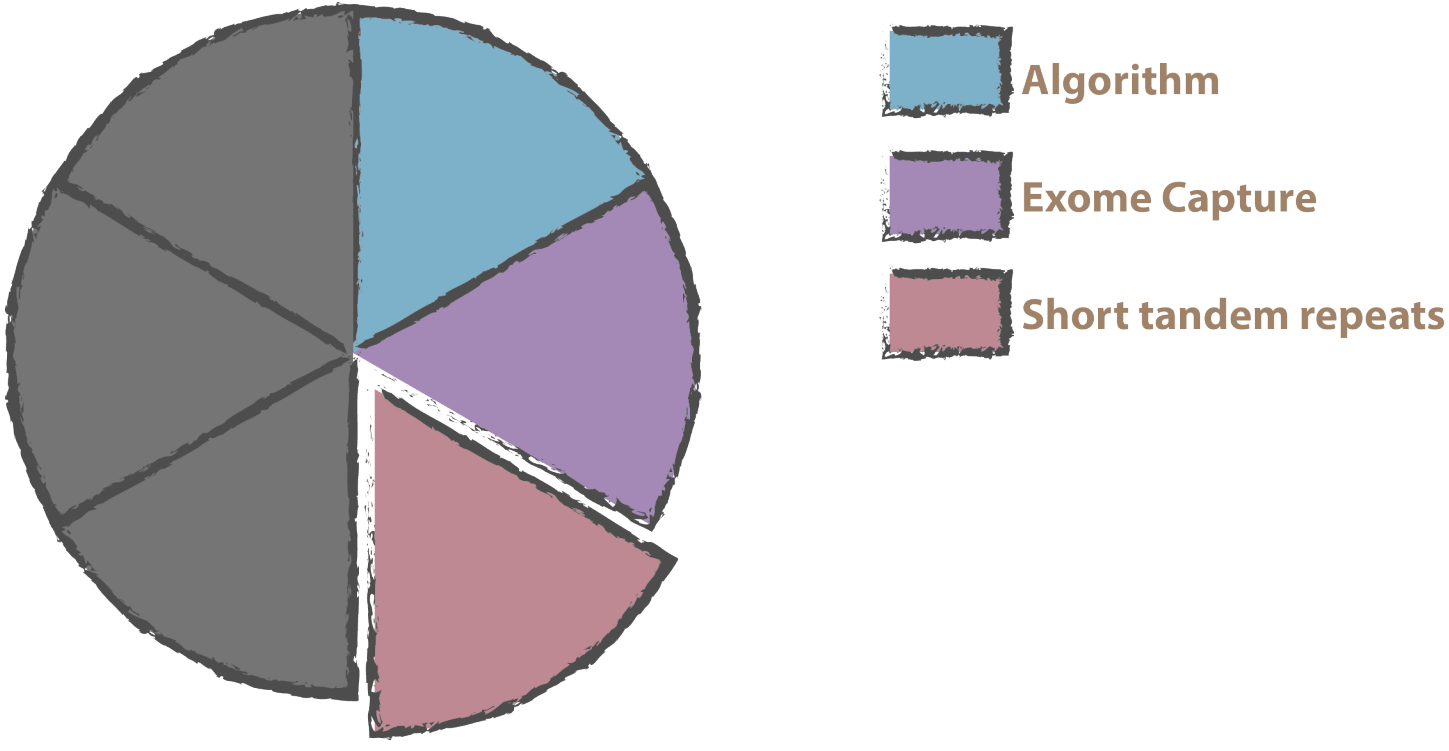
Coverage distributions (WGS-specific INDELs regions)





Coefficient of variation (C_v)

$$\widetilde{C_v}^{*} = \left(1 + \frac{1}{4n}\right) * \left(\frac{s}{\bar{x}}\right)$$


Algorithm
Exome Capture
Short tandem repeats

Introducing the k-mer Chi-Square scores in Scalpel

The k-mer Chi-Square scores $\chi^2 = \frac{(C_o^{Ref} - C_e^{Ref})^2}{C_e^{Ref}} + \frac{(C_o^{Alt} - C_e^{Alt})^2}{C_e^{Alt}}$, where C_o^{Ref} and C_o^{Alt} are the observed k-mer coverage for the reference and alternative alleles, C_e^{Ref} and C_e^{Alt} are the expected k-mer coverage, i.e. $C_e^{Ref} = C_e^{Alt} = \frac{C_o^{Ref} + C_o^{Alt}}{2}$.

a)

In a) , $C_o^{Ref} = 52, C_o^{Alt} = 48$,

so $\chi^2 = \frac{(52-50)^2}{50} + \frac{(48-50)^2}{50} = 0.16$

b)

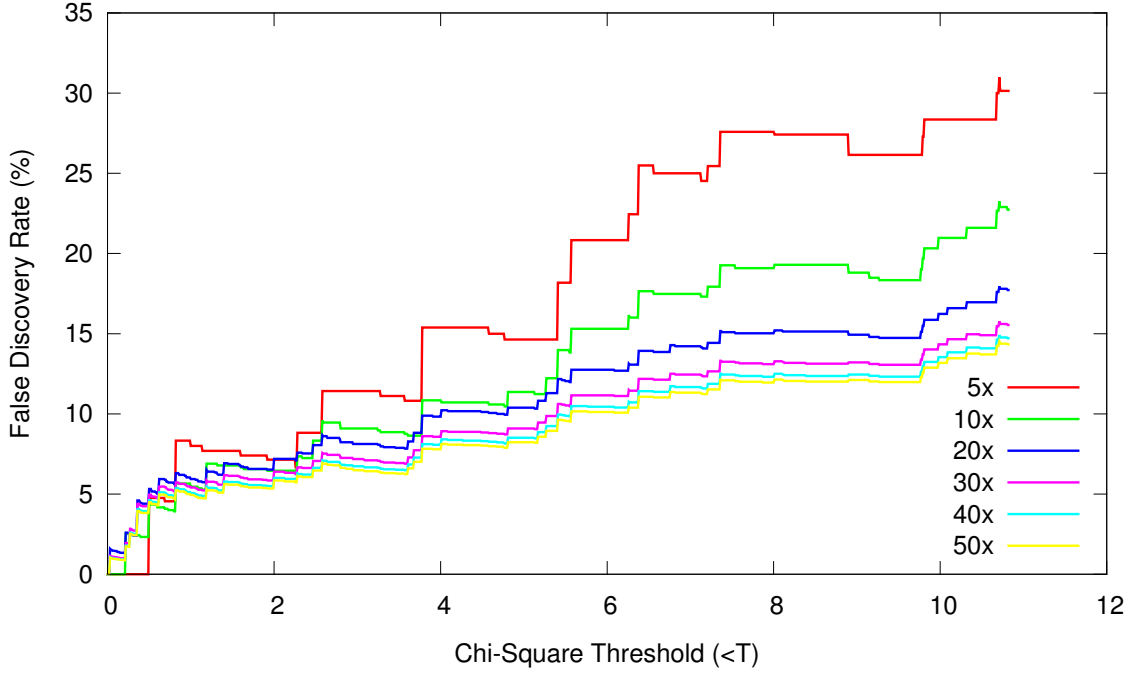
In b) , $C_o^{Ref} = 90, C_o^{Alt} = 10$,

so $\chi^2 = \frac{(90-50)^2}{50} + \frac{(10-50)^2}{50} = 64$

Figures are customized from http://cdn.vanillaforums.com/gatk.vanillaforums.com/FileUpload/a4/5ac06fc8af4b1b0c474f03e45f9017.png

Benchmarking

Effectively distinguish behaviours of problematic INDEL calls from likely true-positives.
Can be easily applied to screen INDEL calls and understand their characteristics.



High quality INDELs (low error-rate - 7%):

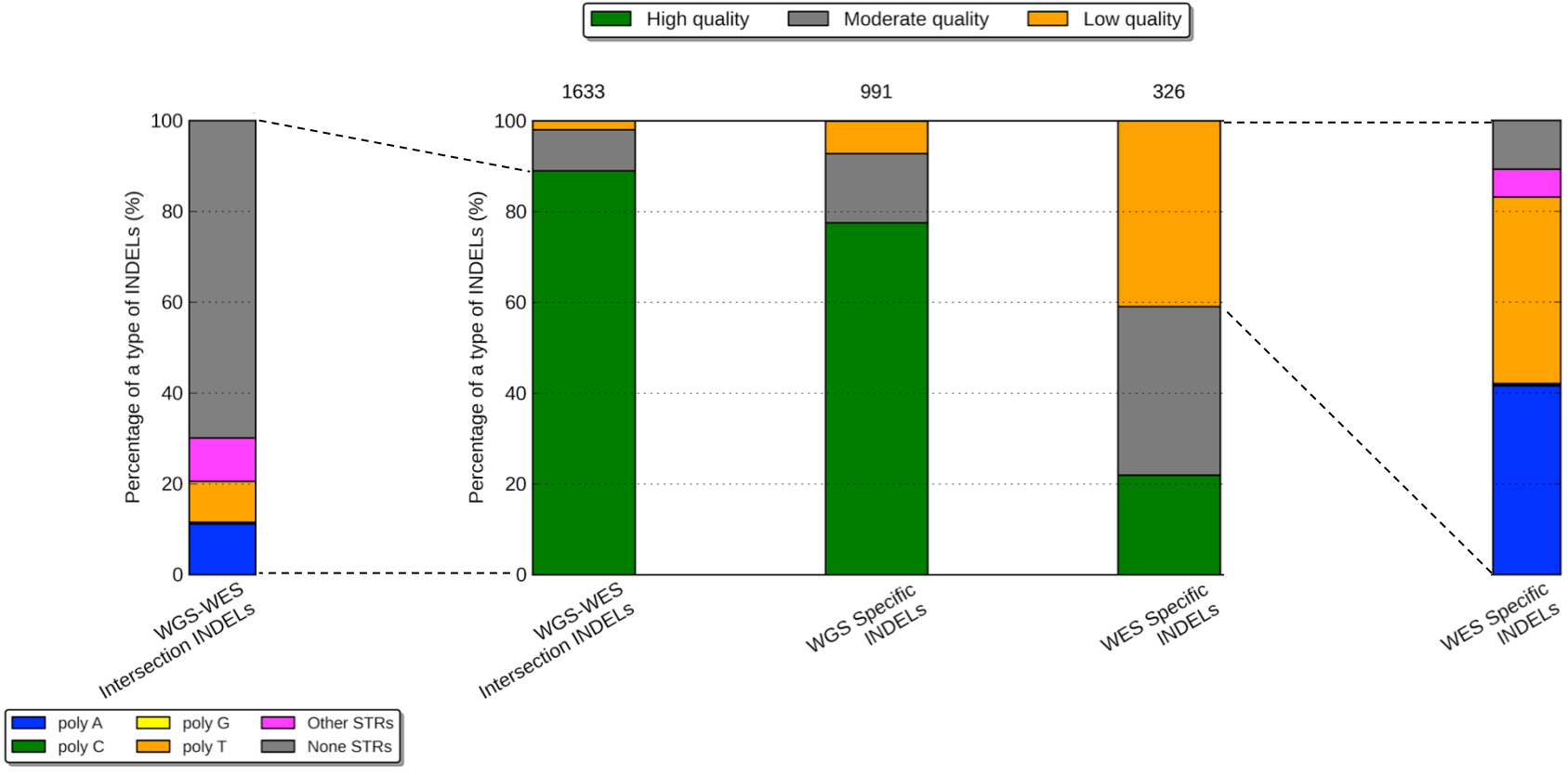
$$\begin{cases} \chi^2 \leq 2.0 & \text{if } C_o^{Alt} \leq 5 \\ \chi^2 \leq 4.5 & \text{if } C_o^{Alt} \leq 10 \\ \chi^2 \leq 10.8 & \text{if } C_o^{Alt} > 10 \end{cases}$$

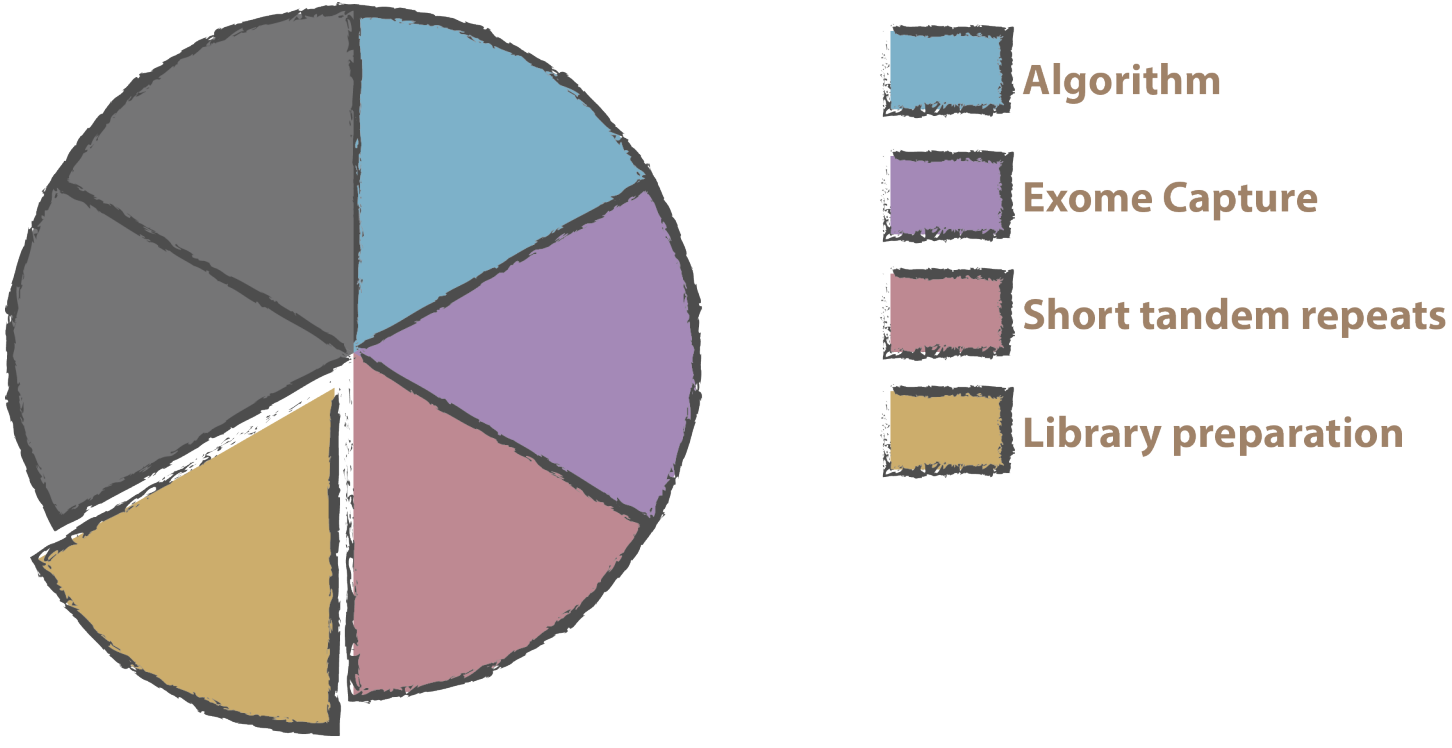
Low quality INDELs (high error-rate - 51%):

$$\chi^2 \geq 10.8 \qquad \text{if } C_o^{Alt} \leq 10$$

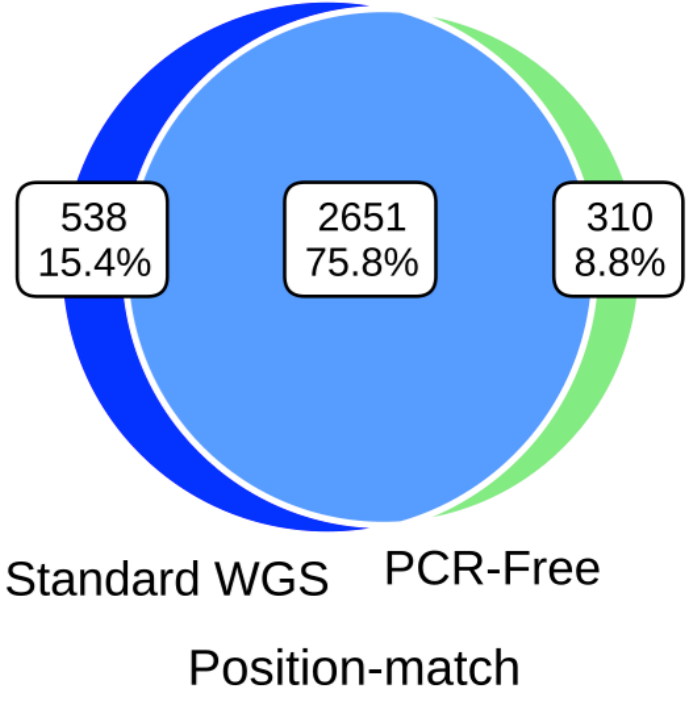
WGS yielded more high-quality INDELs than WES.

Poly-A/T is a major contributor to the low quality INDELs, which gives rise to much more errors in the WES-specific set.



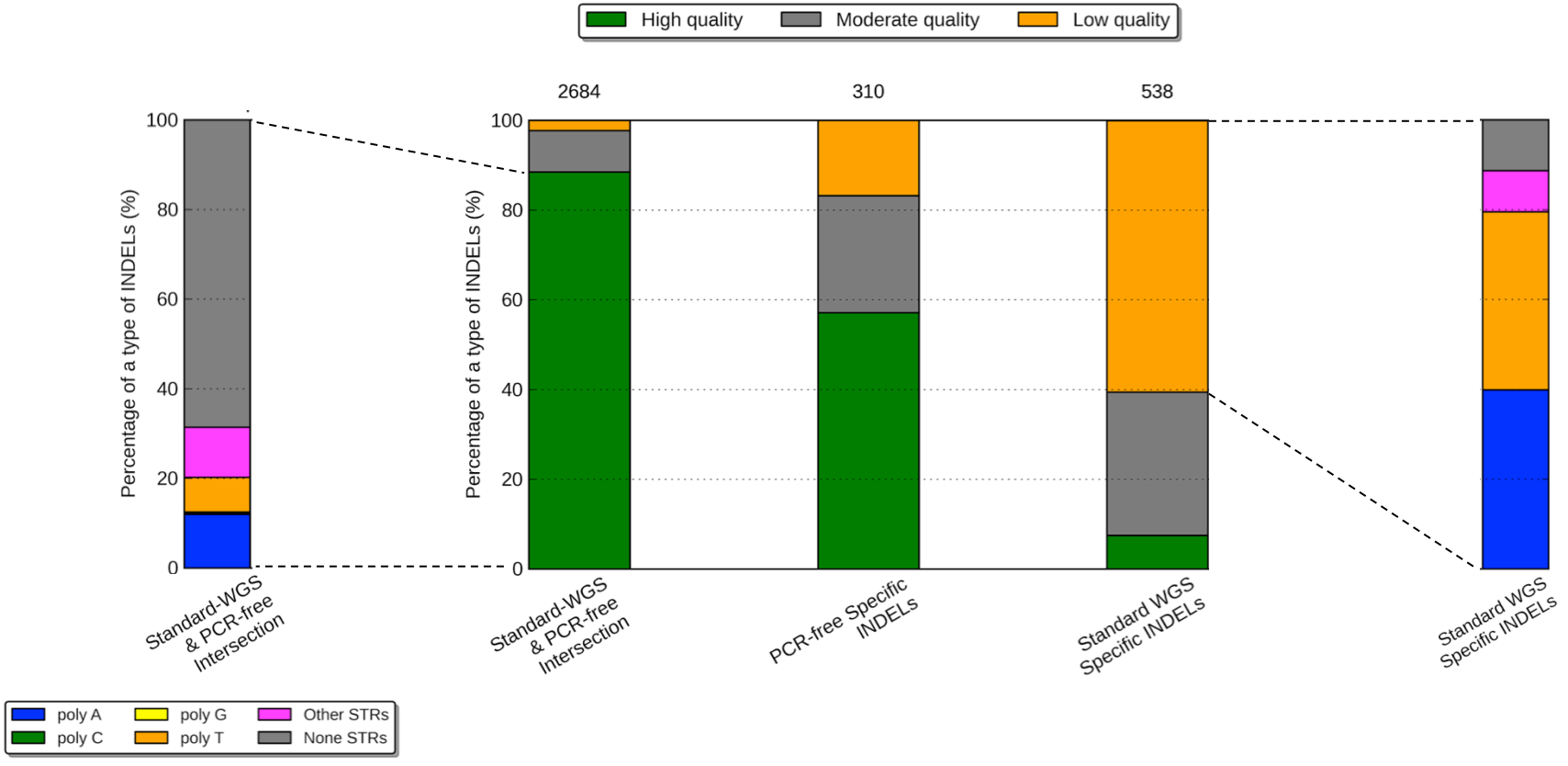

Algorithm
Exome Capture
Short tandem repeats
Library preparation

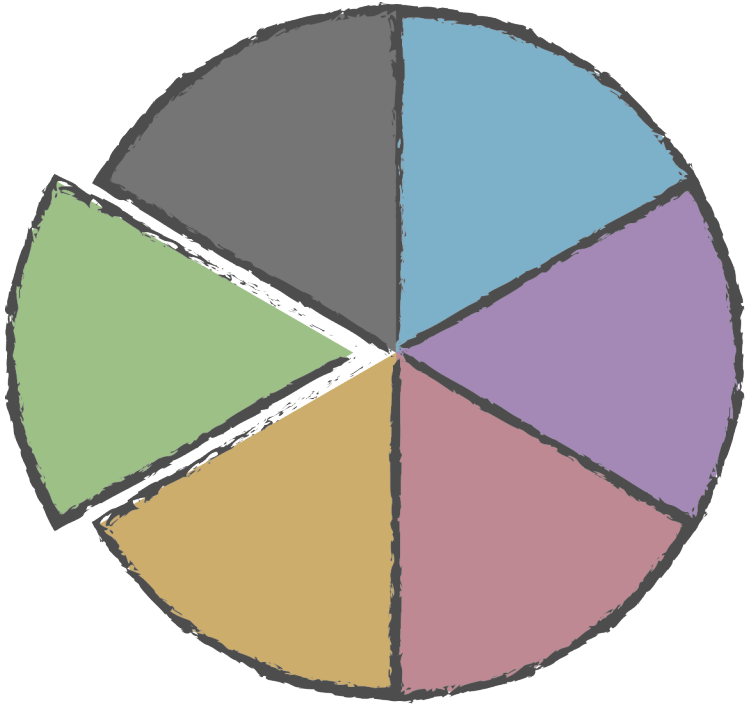
Concordance between standard WGS & PCR-free data



PCR-free data yielded more high-quality INDELs.

PCR amplification induced many error-prone poly-A/T INDELs to the library; reducing the rate of amplification could effectively increase calling quality.





Algorithm

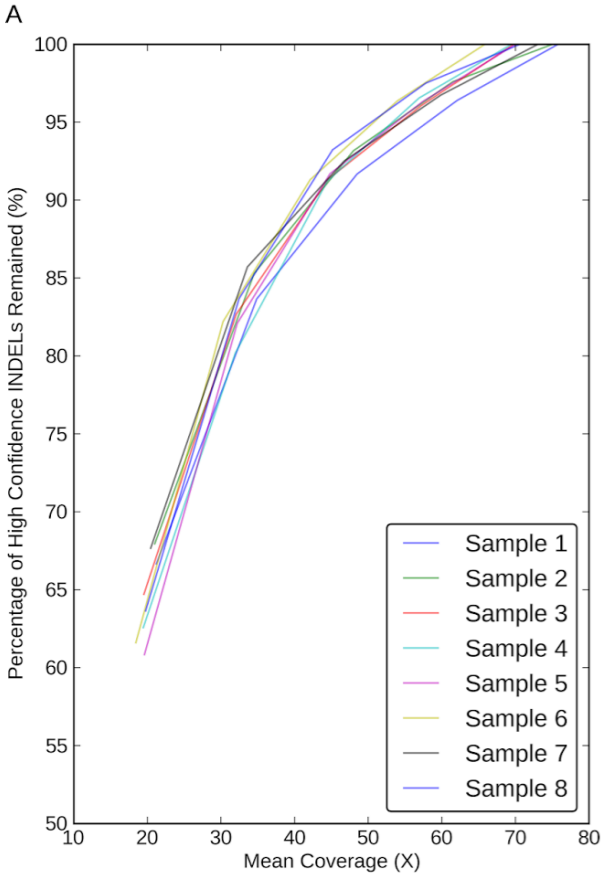
Exome Capture

Short tandem repeats

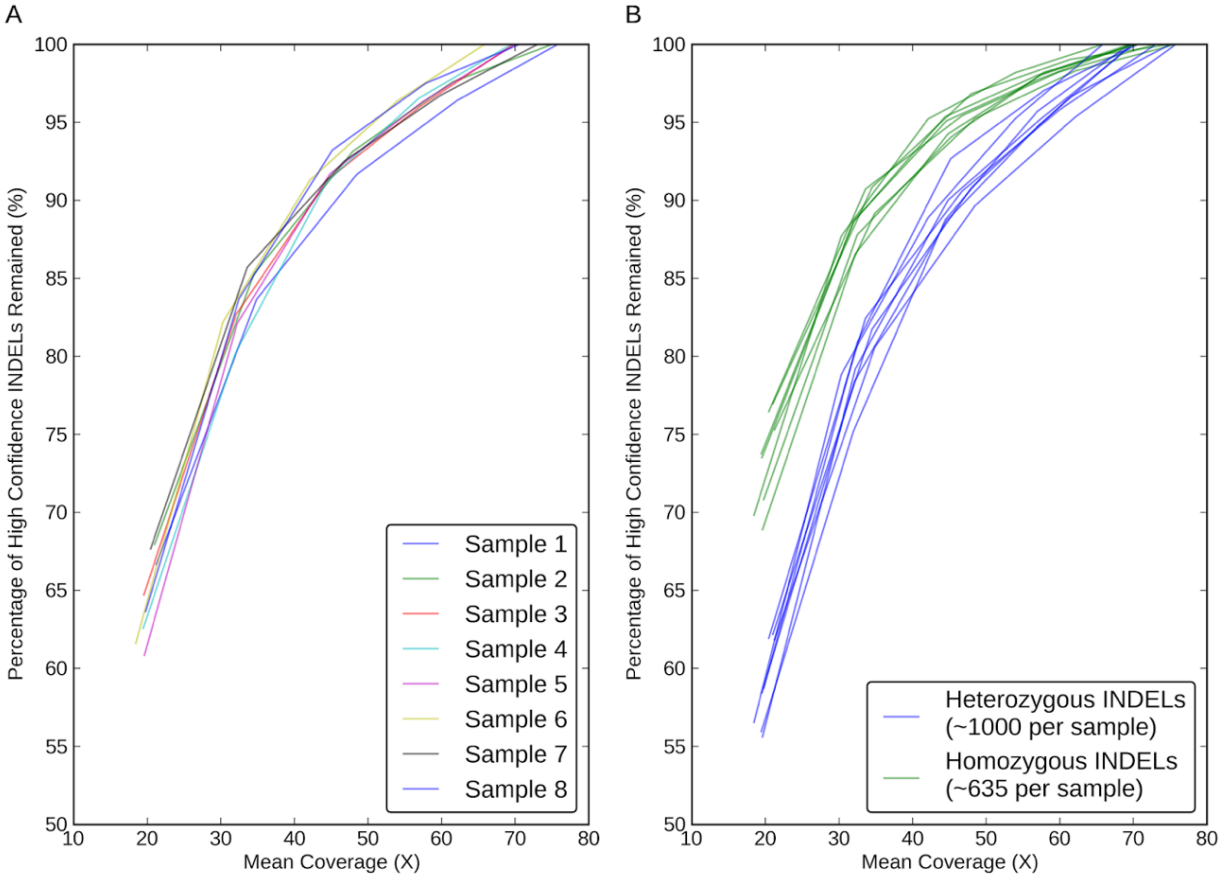
Library preparation

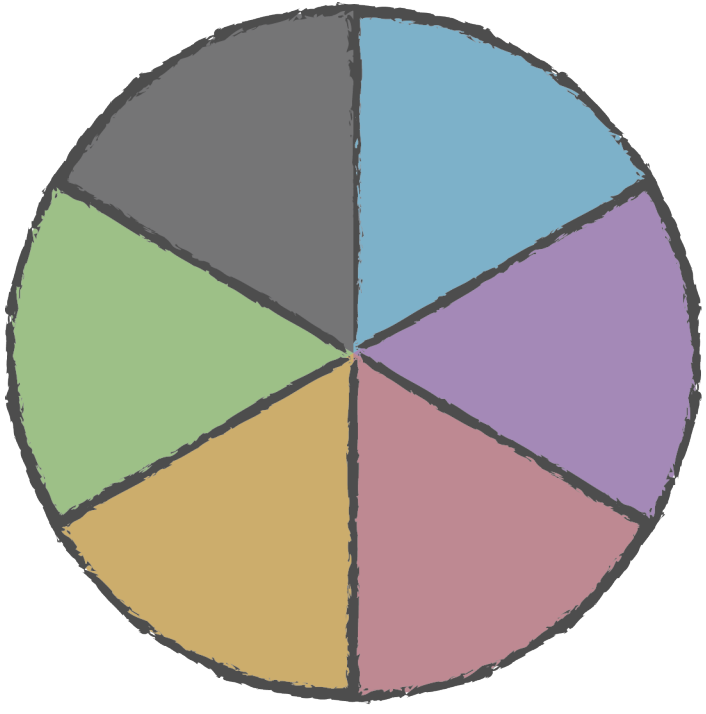
Coverage

60X WGS is needed to recover 95% of INDEL. Detection of het INDELs requires higher coverage.



60X WGS is needed to recover 95% of INDEL. Detection of het INDELs requires higher coverage.





Algorithm

Exome Capture

Short tandem repeats

Library preparation

Coverage

Other factors

Summary

- **Discussed:**

1) Introducing a highly accurate & open-source algorithm, Scalpel (http://scalpel.sourceforge.net/)
2) Higher accuracy of INDEL detection with WGS data than that with WES data.
3) WES data has more false-positives, and misses a lot of large INDELs.
4) STR regions: major sources of INDEL errors, especially near A/T homopolymers.
5) Identify the errors introduced by PCR amplifications and caution about them.

- **Implications:**

1) Recommend WGS data for INDEL analysis (60X PCR-free).
2) Classification scheme of INDEL calls based off of Chi-Square scores and alternative allele coverage.