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

Han Fang[1,2,3], Yiyang Wu[1,2], Giuseppe Narzisi[3,4], Jason A. O'Rawe[1,2], Laura T. Jimenez Barrón[1,5], Julie Rosenbaum[3], Michael Ronemus[3], Ivan Iossifov[3], Michael C. Schatz[3§], Gholson J. Lyon[1,2§]

[1] Stanley Institute for Cognitive Genomics, One Bungtown Road, Cold Spring Harbor Laboratory, NY, USA; [2] Stony Brook University, 100 Nicolls Rd, Stony Brook, NY, USA; [3] Simons Center for Quantitative Biology, One Bungtown Road, Cold Spring Harbor Laboratory, NY, USA; [4] New York Genome Center, New York, NY; [5] Centro de Ciencias Genomicas, Universidad Nacional Autonoma de Mexico, Cuernavaca, Morelos, MX;

INDELs, especially those disrupting protein-coding regions of the genome, have been strongly associated with human diseases. However, there are still many errors with INDEL variant calling, driven by library preparation, sequencing biases, and algorithm artifacts. We characterized whole genome sequencing (WGS), whole exome sequencing (WES), and PCR-free sequencing data from the same samples to investigate the sources of INDEL errors. We also developed a classification scheme based on the coverage and composition to rank high and low quality INDEL calls. We performed a large-scale validation experiment on 600 loci, and find high-quality INDELs to have a substantially lower error rate than low quality INDELs (7% vs. 51%).

Simulation and experimental data show that assembly based callers are significantly more sensitive and robust for detecting large INDELs (>5bp) than alignment based callers, consistent with published data. The concordance of INDEL detection between WGS and WES is low (52%), and WGS data uniquely identifies 10.8-fold more high-quality INDELs. The validation rate for WGS-specific INDELs is also much higher than that for WES-specific INDELs (85% vs. 54%), and WES misses many large INDELs. In addition, the concordance for INDEL detection between standard WGS and PCR-free sequencing is 71%, and standard WGS data uniquely identifies 6.3-fold more low-quality INDELs. Furthermore, accurate detection with Scalpel of heterozygous INDELs requires 1.2-fold higher coverage than that for homozygous INDELs.

Lastly, homopolymer A/T INDELs are a major source of low-quality INDEL calls, and they are highly enriched in the WES data. Overall, we show that accuracy of INDEL detection with WGS is much greater than WES even in the targeted region. We calculated that 60X WGS depth of coverage from the HiSeq platform is needed to recover 95% of INDELs detected by Scalpel. While this is higher than current sequencing practice, the deeper coverage may save total project costs because of the greater accuracy and sensitivity. Finally, we investigate sources of INDEL errors (e.g. capture deficiency, PCR amplification, homopolymers) with various data that will serve as a guideline to effectively reduce INDEL errors in genome sequencing.

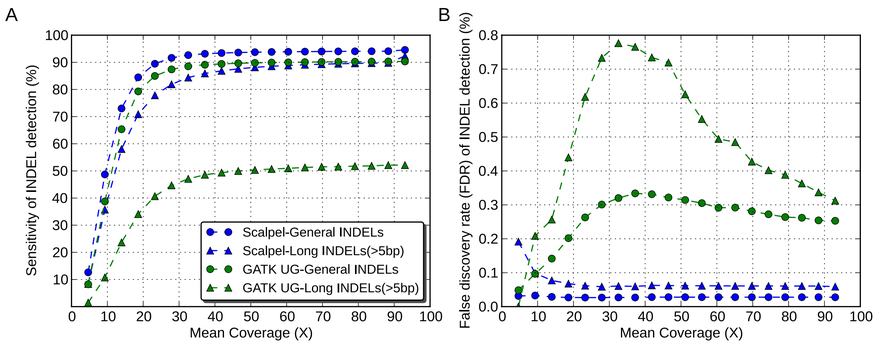

Figure 1. Performance comparison between the Scalpel and GATK-UnifiedGenotyper in terms of sensitivity (A) and false discovery rate (B) at different coverage (simulation data).

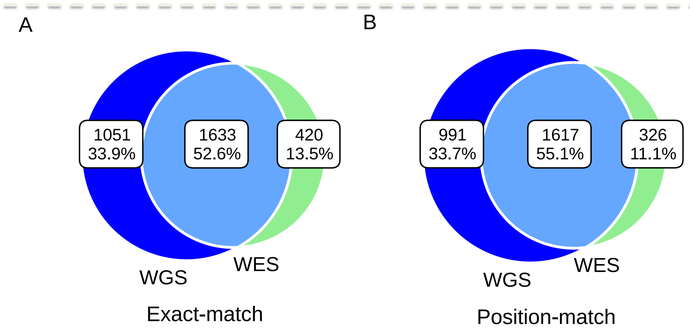

Figure 2. Mean concordance of INDELs over eight samples between WGS (blue) and WES (green) data.

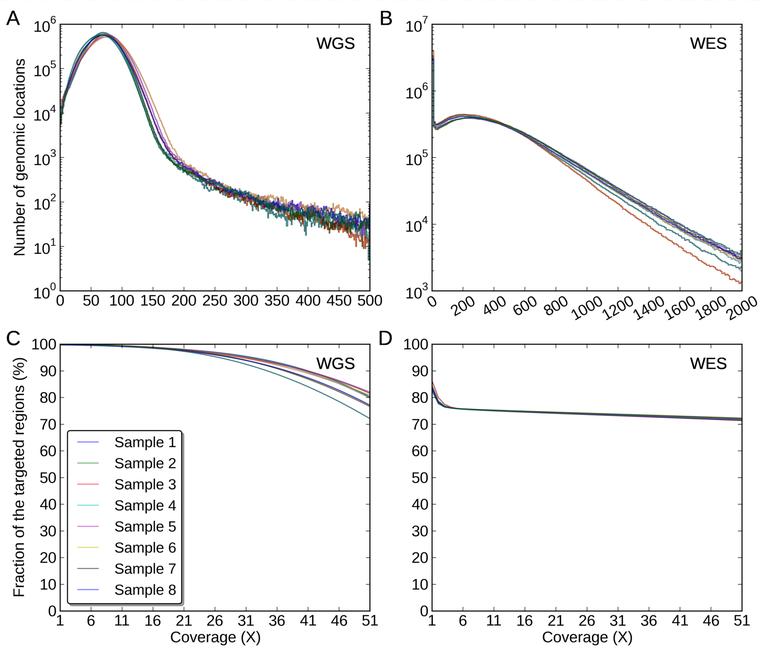

Figure 3. Coverage distributions of the exonic targeted regions in (A) the WGS data, (B) the WES data.

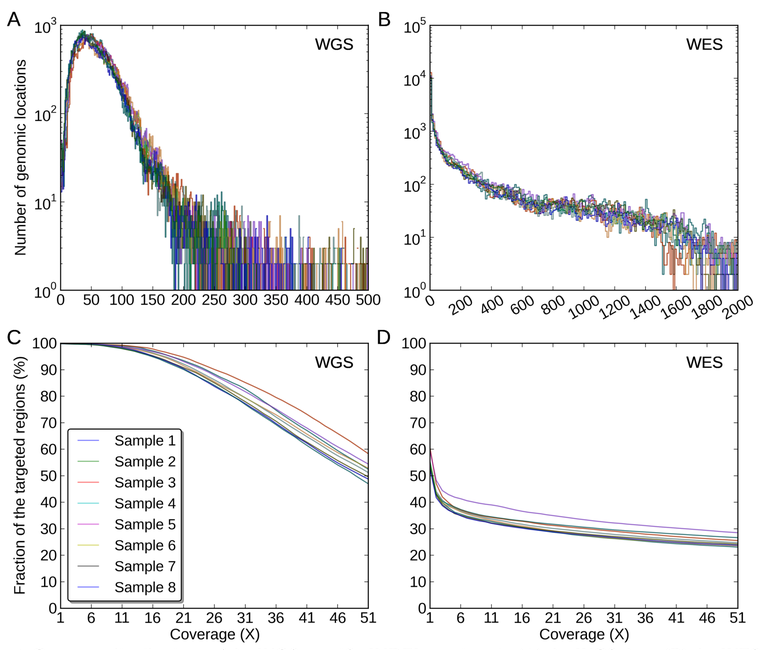

Figure 4. Coverage distributions of the WGS-specific INDELs regions in (A) the WGS data, (B) the WES data.

Table 1. Mean coefficients of variation of coverage with respects to the following regions: WGS-WES intersection INDELs, WGS-specfic INDELs, and WES-specific INDELs.

	Exonic targeted regions	WGS-WES intersection INDEL regions	WGS-specific INDEL regions	WES-specific INDEL regions
WGS	39.4% (1.9%)	47.2% (3.0%)	75.3% (5.7%)	56.1% (9.6%)
WES	109.3% (1.5%)	96.8% (3.2%)	281.5% (13.3%)	117.4% (22.8%)

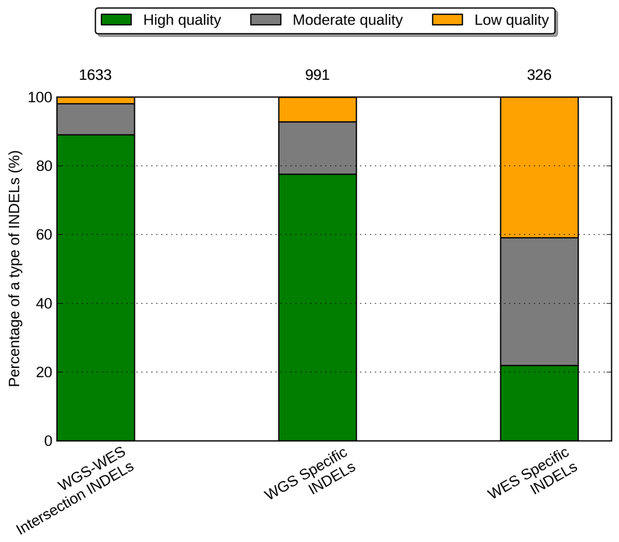

Figure 5. Percentage of high quality, moderate quality and low quality INDELs in three call set.

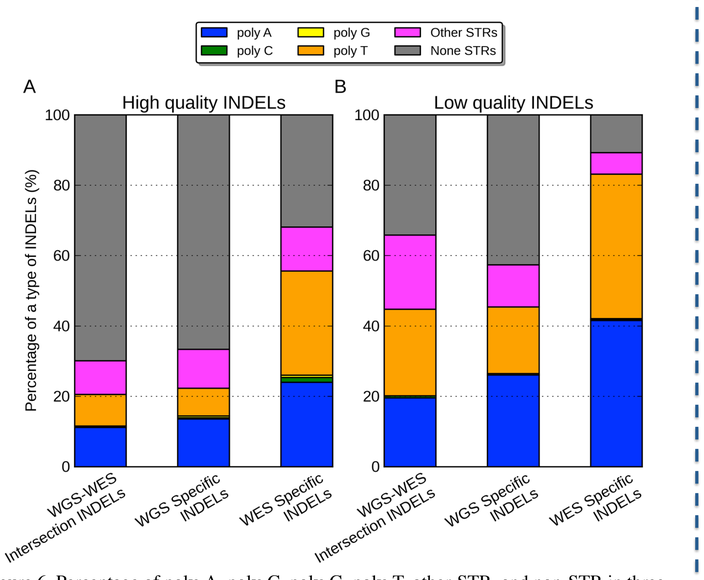

Figure 6. Percentage of poly-A, poly-C, poly-G, poly-T, other-STR, and non-STR in three call set.

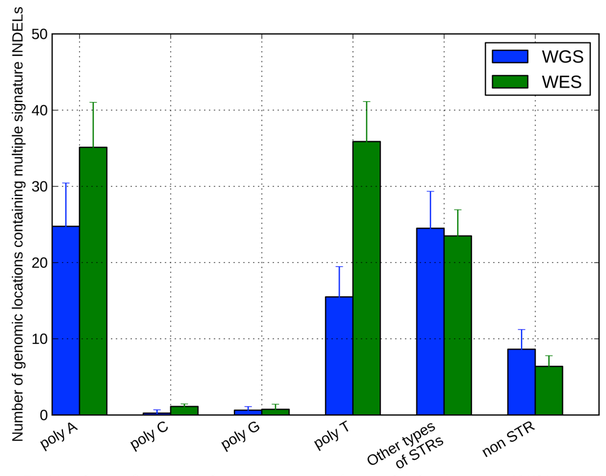

Figure 7. Numbers of genomic locations containing multiple signature INDELs in WGS (blue) and WES data (green).

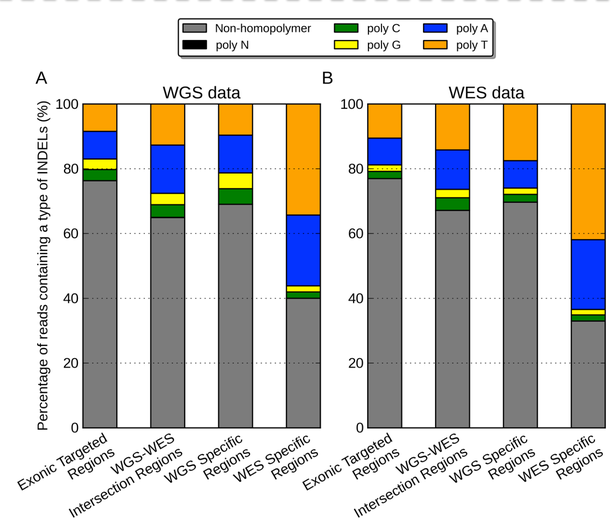

Figure 8. Percentage of reads near regions of Non-homopolymer, poly-N, poly-A, poly-C, poly-G, poly-T in (A) WGS data, (B) WES data.

Table 2. Validation rates of WGS-WES intersection INDELs, WGS-specfic, and WES-specific INDELs. We also calculated the validation rates of large INDELs (>5 bp) in each category. The validation rate, positive predictive value (PPV), is computed by the following: PPV=#TP/(#TP+#FP), where #TP is the number of true-positive calls and #FP is the number of false-positive calls.

	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%

Table 3. Number and fraction of large INDELs in the following INDEL categories: 1) WGS-WES intersection INDELs, 2) WGS-specific, and WES-specific.

	All INDELs	Large INDELs (>5bps)	Fraction of large INDELs (>5bp)
WGS-WES intersection	2009	176	8.8%
WGS-specific	494	104	21.1%
WES-specific	674	10	1.5%

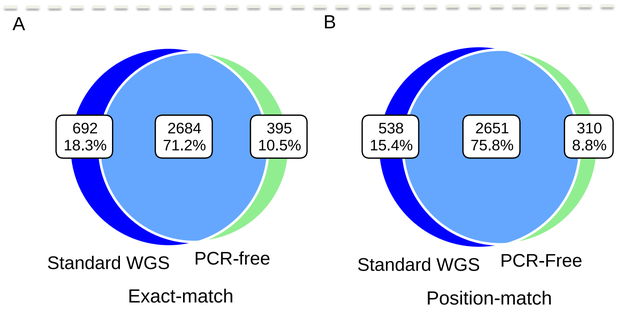

Figure 9. Concordance of INDEL detection between PCR-free and standard WGS data on NA12878.

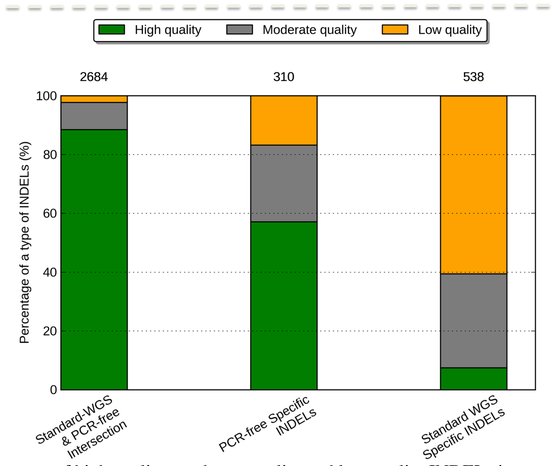

Figure 10. Percentage of high quality, moderate quality and low quality INDELs in two datasets.

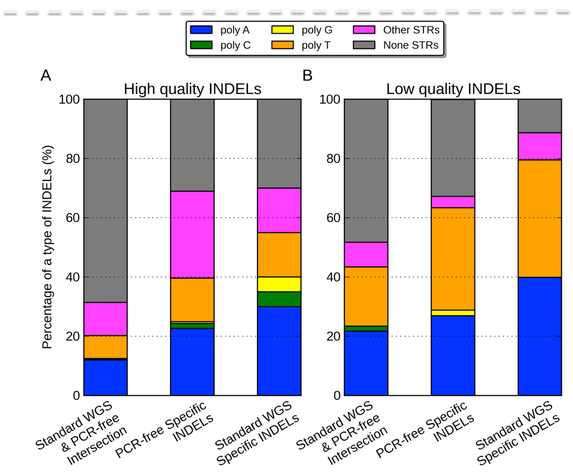

Figure 11. Percentage of poly-A, poly-C, poly-G, poly-T, other-STR, and non-STR in (A) high quality INDELs, (B) low quality INDELs.

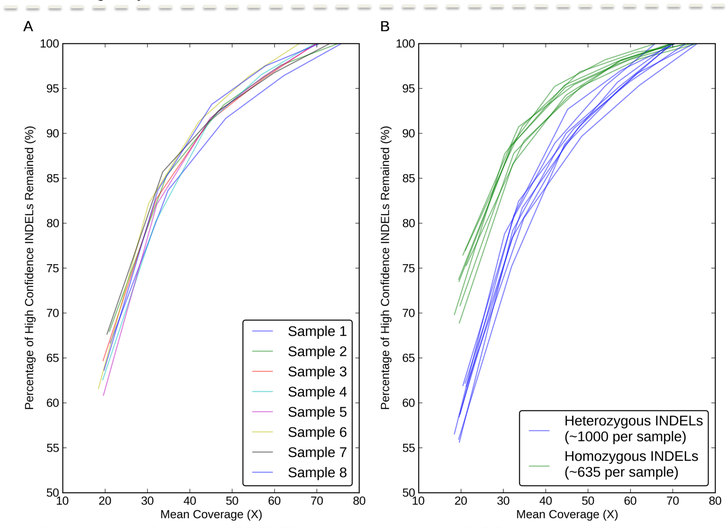

Figure 12. Sensitivity performance of INDEL detection with eight WGS datasets at different mean coverages on Illumina HiSeq2000 platform.