Reducing INDEL errors in whole-genome and exome sequencing

Han Fang[1,2], Giuseppe Narzisi[3], Jason A. O'Rawe[1,2], Yiyang Wu[1,2], 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, 11724;

[§] Co-corresponding author

Email addresses:

HF: hanfang.cshl@gmail.com

GN: gnarzisi@cshl.edu

JAO: jazon33y@gmail.com

YW: yiwu@cshl.edu

JR: jrosenba@cshl.edu

MR: ronemus@cshl.edu

II: iossifov@cshl.edu

MCS: mschatz@cshl.edu

GJL: gholsonjlyon@gmail.com

Abstract

Background

INDELs, especially those disrupting protein-coding regions of the genome, have been associated with human diseases. However, there are still many errors with INDEL variant calling, driven by library preparation, sequencing biases, and algorithm artifacts. We have recently developed a new INDEL-calling algorithm, Scalpel, with substantially improved accuracy.

Results

We characterized whole genome sequencing (WGS), whole exome sequencing (WES), and PCR-free sequencing data from the same samples to investigate false-positive and false-negative INDEL errors. We developed a classification scheme utilizing validation data to define a class of low-quality INDELs with ~2.7-fold higher error rates than high-quality INDELs. The mean concordance of INDEL detection between WGS and WES data was ~52%, while WGS data uniquely identified ~10.8-fold more high-quality INDELs. Concordance of INDEL detection between standard and PCR-free sequencing data was ~71%, while PCR-free data uniquely yielded ~6.3-fold fewer low-quality INDELs. We demonstrate that these INDEL errors are significantly reduced with a PCR-free library protocol, implying that these errors are introduced with PCR amplification. We calculated that 60X WGS data from the HiSeq 2000 platform are needed to recover ~95% of INDELs, much higher than that for SNP detection. Accurate detection of heterozygous INDELs requires ~1.2-fold higher coverage than that for homozygous INDELs.

Conclusions

Homopolymer A/T INDELs are a major source of low quality and/or uncertain INDEL calls, and these are highly enriched in the WES data. We recommend WGS for human genomes at 60X mean coverage with PCR-free protocols, which can substantially improve the quality of personal genomes.

Keywords: INDEL, whole genome sequencing, exome sequencing, homopolymers, short tandem repeats, PCR amplification, personal genomes.

Background

With the increasing use of next-generation sequencing (NGS), there is growing interest from physicians, patients and consumers to better understand the underlying genetic contributions to various conditions. For rare diseases and cancer studies, there has been increasing success with exome/genome sequencing in identifying mutations that have a large effect size for particular phenotypes [1, 2]. Some groups have been trying to implement genomic approaches to interpret disease status and inform preventive medicine [3-6]. However, we are still facing practical challenges for both analytic validity and clinical utility of genomic medicine [7-9]. In addition, the genetic architecture behind most human disease remains unresolved [10-15]. Some have argued that we should bring higher standards to human-genetics research in order to return results and/or reduce false-positive reports of "causality" without rigorous standards [16, 17]. Others have reported that analytic validity for WES and WGS is still a major issue, because the accuracy and reliability of sequencing and bioinformatics analysis is too low for a clinical setting [8, 9, 18-20].

There is also debate whether we should use whole genome sequencing (WGS) or whole exome sequencing (WES) for personal genomes. Some have argued that a first-tier cost-effective WES is a powerful way to dissect the genetic basis of diseases and to facilitate the accurate diagnosis of individuals with Mendelian disorders [21, 22]. Others argue that WGS could reveal structural variants (SVs), maintain a more uniform coverage, and is free of exome capture efficiency issues [23, 24]. One group directly compared WGS with WES, but investigation of INDEL errors was not the focus of this comparison [19, 20]. Substantial genetic variation involving INDELs in the human genome has been previously reported but accurate INDEL calling is still difficult [25-27]. There has been a dramatic decrease of sequencing cost over the past few years, and this cost might decrease further with the release of the Illumina HiSeq X Ten sequencers which have capacity for nearly twenty thousand whole human genomes per instrument per year. However, it is still unclear whether we can achieve a high-accuracy personal genome

with a mean coverage of 30X from the Illumina HiSeq X Ten sequencers. In addition, there has been controversy on the use of PCR amplification in the library preparations for NGS, although very few have characterized the PCR errors that might be complicating the detection of insertions and deletions (INDELs).

Previously our group reported low concordance rates of multiple variant calling pipelines, and of all INDELs detected by GATK Unified Genotyper (v1.5), SOAPindel (v1.0) and SAMtools (v0.1.18), only 26.8% were in the intersection [8]. Some groups also reported low concordance rates for INDELs between different sequencing platforms, further showing the difficulties of accurate INDEL calling [17]. Other efforts have been made to understand the sources of variant calling errors [28]. Common INDEL issues, such as realignment errors, errors near perfect repeat regions, and an incomplete reference genome have caused problems for alignment-based callers [29]. De novo assembly using de Brujin graphs has been reported to tackle some of these limitations [30]. Fortunately, with the optimizations of micro-assembly, these errors have been reduced with a novel algorithm, Scalpel, which outperformed any other INDEL callers available. Based on validation data, positive prediction rate (PPV) of algorithm specific INDELs was high for Scalpel (~80%), but much lower for GATK HaplotypeCaller (v3.0) (~45%) and SOAP-indel (v2.01) (~50%) [31].

Thus, we set out to investigate the complexities of INDEL detection on Illumina reads using this most accurate INDEL-calling algorithm. Firstly, we used simulation data to understand the limits of how coverage affects INDEL calling with perfect Illumina-like reads. Secondly, we analyzed a dataset including high coverage WGS and WES data from two quad families (mother, father and two children), in addition to extensive high-depth validation data on an in-house sample, K8101-49685s. In order to further understand the effects of PCR amplification on INDEL calling, we also downloaded and analyzed two WGS datasets prepared with and without PCR from the well-known HapMap sample NA12878. We characterized the data in terms of read depth, coverage uniformity, base-pair composition pattern, GC contents and other sequencing features, in

order to partition and quantify the INDEL errors. We were able to simultaneously identify both the false-positives and false-negatives of INDEL calling, which should be useful for population-scale experiments. We notice that homopolymer A/T INDELs are a major source of low quality INDELs and multiple signatures. As more and more groups start to use these new micro-assembly based algorithms, practical considerations for experimental design should be introduced to the community. Lastly, we explicitly address the question concerning the necessary depth of coverage for accurate INDEL calling for WGS. This provides important insights and guidelines to achieve a highly accurate INDEL call set and to improve the sequencing quality of personal genomes.

Results and discussion

Simulated data: Comparisons with sub-alignments

We started our study with the following question: how does depth of sequencing coverage affect INDEL calling? Thus, we simulated reads with known error rates across the genome to answer this question. Supplemental Figure S1A shows that better coverage improves sensitivity of detecting both general INDELs (i.e. any size starting from 1bp) and large INDELs (i.e. size greater than 5bp). For general INDEL detection, the increase of mean coverage effectively increased sensitivity although it seemed to saturate after a mean coverage of 28X. Furthermore, for large INDELs, the increase of sensitivity did not saturate until reaching a mean coverage of 42X. We also noticed that detection of large INDELs was uniformly more difficult than general INDELs. To achieve a sensitivity of 90%, a mean coverage of 30X was required for general INDEL detection while 90X was needed to detect large INDELs at a similar sensitivity. This showed that much higher coverage is needed for large INDEL detection, especially to maintain coverage across the INDEL and to have enough partially mapping or soft-clipped reads to use for the micro-assembly. However, the false discovery rates (FDR) remained almost the same as the mean coverage increased for general INDEL detection. Furthermore, for the detection of large INDELs, the FDRs marginally decreased as the mean coverage increased from 5X to 28X, and remained basically the same again from 33X to 93X (Supplemental Figure S1B). This indicates that for large INDELs, insufficient coverage results in more assembly errors, which results in a higher error rate for micro-assembly variant calling. Based on the simulation data, one would need at least a mean coverage of 30X to maintain a reasonable FDR for micro-assembly based INDEL callers. However, since these results were based on simulation data, which does not include the effects of any sequencing artifacts on INDEL calling, these values establish the upper bound of accuracy and performance compared to genuine sequence data.

WGS vs. WES: Low concordance on INDEL calling

We analyzed a dataset including high coverage WGS and WES data from eight samples. The mean INDEL concordance over eight samples between WGS and WES data using exact-match and position-match was 53%±0.8% and 55%±0.9%, respectively (Figure 1). Position-match means two INDELs have the same genomic coordinate, while exact-match additionally requires that two INDELs also have the same base-pair change(s) (Methods). When we excluded regions with less than one read in either dataset, the mean concordance rates based on exact match and position-match increased to 62%±1.1% and 66%±1.0%, respectively. If we excluded regions with base coverage in either dataset with less than 20, 40, 60, or 80 reads, the mean concordance rate based on exact-match and position-match both continued to increase until reaching a base coverage of 80 reads (Table 1). This showed that some INDELs were missing in either dataset because of low sequencing efficiency in those regions. Although WES data had higher mean coverage than WGS data, we were surprised to see that in regions requiring at least 80 reads, there was only 4%±1.5% of INDEL specific to WES but 21%±3.2% specific to WGS. Regions with excessive coverage might indicate problems of sequencing or library preparation and this highlights the importance of coverage uniformity in WGS. Based on exact-match, the proportion of the WGS-specific INDELs was 34%±1.4%, which was ~2.5-fold higher than the proportion of WES-specific INDELs (14%±1.2%). This ratio was even higher based on position-match (~3-fold). Reasons for this could be either high sensitivity of INDEL detection with WGS data or high specificity of INDEL detection with WES data. We will further illustrate this in the downstream analysis.

Coverage distributions of different regions in WGS and WES data

We define the proportion of a region covered with at least X reads to be the coverage fraction at X reads for this region. Comparisons of the coverage distributions are shown in the following order: 1) Exonic targeted regions, i.e. the exons that exome capture kit was designed to pull down and enrich; 2) High confidence INDEL regions, i.e. the regions where WGS and WES revealed the identical INDELs based on exact-match; 3) WGS-specific INDEL regions, i.e. the regions where only WGS revealed INDELs based on position-match; 4) WES-specific INDEL regions, i.e. the regions where only WES revealed INDELs based on position-match.

In the exonic targeted regions, the mean coverages across eight samples were 71X±3.3X and 337X±18.2X for WGS and WES data, respectively (Figure 2). The coverage fractions at 20X were 98%±0.2% and 75%±0.1% in WGS and WES data, respectively. We noticed that there was a capture efficiency issue with WES in some regions, as the coverage fraction at 1X was 99.9%±0.08% in WGS data but only 84%±1.1% in WES data, indicating that ~16% of the exonic targeted regions were not captured for sequencing. In the high confidence INDEL regions, the mean coverage across eight samples were 58X±3.4X and 252X±7.0X for WGS and WES data, respectively (Figure 3). The coverage fractions at 20X were 96%±1.1% and 97%±0.3% in WGS and WES data, respectively. We noticed that there was a significant increase of coverage uniformity for WES in the high confidence INDEL regions, relative to the exonic targeted regions. This increase was even more significant when we looked at the coverage fractions at 50X, which were 58%±6.0% and 86%±0.7% in WGS and WES data, respectively. This suggested that WGS was able to reveal high confidence INDELs at a much lower coverage and a better uniformity across the genome, relative to WES. Depth coverage distributions were skewed in the WES data, with some regions poorly covered and other regions over saturated with redudant reads.

However, in WGS-specific INDEL regions, the mean coverages across eight samples were 61X±2.9X and 137X±12.1X for WGS and WES data, respectively (Figure 4). Compared to the targeted regions, the mean coverage for WES data was significantly reduced in these regions. The coverage fractions at 1X were 99.9%±0.03% and 56%±0.3% in WGS and WES data, respectively. The difference became even larger for coverage fractions at 20X: 94%±1.4% for WGS and 31%±2.1% for WES data. The reason why WES data missed these INDELs could be: ~44% of these regions in WES data were not covered at all and ~69% were covered with fewer than 20 reads. Furthermore, in WES-specific INDELs regions, the mean coverages across eight samples were 41X±5.2X and 172X±10.0X for WGS and WES data, respectively (Figure 5). The coverage fractions at 20X were 87%±6.1% and 96%±1.0% in WGS and WES data,

respectively. We noticed in these regions, the depth coverage of WGS data was significantly lower than WES data because the coverage fractions at 50X decreased to 29%±9.4%, while it was 79%±3.3% for WES data. It was difficult to directly understand the issues with these regions, so we used the high confidence INDEL set as a positive control and proceeded to assess each call set with newly developed quality criteria.

Assessment of the INDEL calls sets from WGS and WES

Based on previous validation data, we selected three combinations of thresholds to define the calling quality of an INDEL call as either high, moderate or low quality based on the following two metrics: the coverage of the alternative allele and the k-mer Chi-Square score of an INDEL (Methods). Using these criteria, 89% of the high confidence INDELs were considered as high quality, 9% as moderate quality, and only 2% as low quality (Figure 6, Supplemental Table S1). We noticed a similar pattern for WGS-specific INDELs: 78% were of high quality, 15% as moderate quality, and only 7% as low quality. However, for WES-specific INDELs, there was a striking enrichment of low quality events (41%), additionally with a ~4.1-fold decrease of the high quality events (22%). Notably, among these 8 samples. there were 991 WGS-specific INDELs and 326 WES-specific INDELs, and from these, 769 of WGS-specific INDELs and 71 of the WES-specific INDELs were of high quality. This comparison suggested that WGS yielded ~10.8-fold more high quality INDELs than WES. Futhermore, WES produced 133 low quality INDELs per sample, while WGS only produced 71 low quality INDELs per sample. That being said, WES yielded ~1.9-fold more low quality INDELs. This indicates WES tends to produce a larger fraction of error-prone INDELs, while WGS reveals a more sensitive and specific set of INDELs.

In order to understand what was driving the error rates in different data sets, we partitioned the INDELs by the following six regions: homopolymer A (poly-A), homopolymer C (poly-C), homopolymer G (poly-G), homopolymer T (poly-T), short tandem repeats (STRs) except homopolymers (other STRs), and non-STRs. We noticed that for the high quality events, the majority of the high confidence INDELs (70%±1.2%) and WGS-specific INDELs (67%±1.1%) were within non-STRs regions (Figure 7).

Further, a majority of the high quality INDELs specific to WES were within poly-A (24%±3.0%) and poly-T regions (30%±3.5%). When we compared the low quality INDELs to the high quality INDELs, there were consistent enrichment of homopolymer A or T (poly-A/T) INDELs in all three call sets, ~2.3-fold for high confidence events, ~2.1-fold for WGS-specific events, and ~1.5-fold for WES-specific events. The WES-specific call set contained a much higher proportion (83%±4.8%) of Poly-A/T INDELs from the low-quality INDELs, relative to the high confidence call set (44%±16.8%), and the WGS-specific call set (45%±9.0%). This suggested that poly-A/T is a major contributor to the low quality INDELs, which gives rise to much more INDEL errors. We explored this further by a comparison of PCR-free and standard PCR-involved WGS data below.

Sources of multiple signatures in WGS and WES data

Another way of understanding INDEL errors is to look at multiple signatures at the same genomic location. Multiple signatures means that for the same genomic location, there are more than one INDELs called. If we assume only one signature can be the true INDEL in the genome, any additional signatures would represent false-positive calls. So if we have a higher number of multiple signatures, it means that these reads contained more INDEL errors or the algorithm tends to make more mistakes in these regions. We combined the call sets from both datasets and identified multiple signatures in the union for each sample. In order to understand the error behaviors in the above assessment, we also partitioned the signatures by the same regional criteria. We noticed that the poly-A/T INDELs are the major source of multiple signatures, which are enriched in WES data (72%±7.4% for WES vs. 54%±9.3% for WGS) (Figure 8). In particular, there are a higher number of both poly-A (35±5.9 vs. 25±5.7) and poly-T (36±5.3 vs. 16±4.0) INDEL errors in the WES data than in the WGS data.

We tried to understand the source of multiple signatures by the numbers of reads containing homopolymer INDELs inferred by the CIGAR code (Figure 9). Figure 9 showed that there is a much higher proportion of poly-A/T INDELs in the WES-specific regions from both WGS (56%±7.5%) and WES data (64%±5.8%), relative to other

regions. In addition, WES data has ~6.3-fold more reads than WGS data in the regions with INDELs specific to WES data (11251 vs. 1775, Supplemental Table S2). According to Qualimap, a large number of homopolymer indels might indicate a problem in sequencing for that region. Here we particularly identified the effects of these problematic sequencing reads on INDEL calling, which revealed more multiple signatures of poly-A/T INDELs.

Applications of using filtering criteria to reduce false positive de novo INDELs

Supplemental Table S3 shows the probabilities of seeing more than K INDELs from one of the 343 families reported in Iossifov *et al.* 2012 [32]. Scalpel has a de novo analysis mode; it could re-assemble each region associated with the candidate INDELs across the family members using a more sensitive parameter setting. This setting was indeed more sensitive for detecting de novo INDELs than single-sample calling. Due to this, we used the following more rigorous filtering criteria than the above assessment to exclude any spurious false-positive de novo INDELs: coverage of the alternative allele >10 and Chi-Square score <4. Supplemental Table S4 showed the number of putative de novo INDELs in two families before and after applying this filtering criteria. All of the putative events in the two families were successfully excluded, which was consistent with the validated results in the variant database reported by Iossifov *et al.* 2012 [32]. We noticed that, in both families, the majority of these false-positive de novo INDELs were poly-A/T relevant (91% for WGS, 78% for WES), which was consistent with the above assessment. This suggested that if we used very sensitive callers, we should control for poly-A/T false-positive de novo INDELs by applying a more rigorous filtering criteria, especially in population-scale sequencing projects, where there is substantial expense with experimental validation.

Validation of INDELs in WGS and WES data on the sample K8101-49685s

We previously selected ~1400 INDELs called with multiple variant callers from the WES data on the sample K8101-49685s. 453 of those were called by Scalpel with this WES data and 372 of them were validated [31]. In this study, we called the INDELs with Scalpel again but from the WGS data on this sample. The WGS data revealed 324 out of these 453 INDELs (72%). This sensitivity at 30X is consistent with what we predict in

the later discussion (75% at 30X). Among those, 286 INDELs were validated and 38 INDELs were invalid according to the validation experiment. This showed that 43 (53%) of these false-positive INDELs were not called from the WGS data and thus specific to the WES data. The error rate of INDELS from the WGS data based on the validation subset was 12%, which was the same as the high quality INDELs discussed above. Furthermore, 126 of those 453 INDELs were specific to the WES data and 43 failed to pass the validation. The error rate of INDELs specific to the WES data is 34%, which was almost the same as the low quality INDELs discussed above. This validation data showed that the high confidence INDELs called by both data-types were indeed of high quality and low error rate, even though this WGS data was with a mean coverage of 30X. There is ~2.8-fold higher error rate for INDELs specific to the WES data, suggesting that INDELs specific to the WES data in fact are of low quality.

PCR-involved vs. PCR-free: assessment of INDELs calling quality

The concordance rate between PCR-involved and PCR-free data on NA12878 using exact-match and position-match were 71% and 76%, respectively (Figure 10). These concordance rates were higher than those between WGS and WES, even for regions having at least one read in both datasets. Based on exact-match, the proportion of INDELs specific to PCR-involved data was 18%, which is ~1.6-fold higher than the proportion of INDELs specific to PCR-free data (11%). This ratio was similar based on position-match (~1.7-fold). Like previous assessments, we classified the three call sets with respects to calling quality. We again used the high confidence INDEL call set as a positive control. Figure 11 shows that 89% of the high confidence INDELs are considered as high quality, 9% as moderate quality, and only 2% as low quality. However, for INDELs specific to PCR-involved data, there is a large proportion of low quality events (61%), and a very limited proportion are of high quality (7%). There were on average 310 INDELs specific to PCR-free data and 538 INDELs specific to PCR-involved data. Notably, 177 of the PCR-free-specific INDELs and 40 of the PCR-involved-specific INDELs were of high quality, suggesting that in these specific regions, PCR-free data yielded ~4.4-fold more high quality INDELs than PCR-involved data. Furthermore, 326 of the PCR-involved-specific INDELs were of low quality, while in the

PCR-free-specific call set, 52 INDELs were of low quality. That being said, in regions specific to data types, PCR-involved data yielded ~6.3-fold more low quality INDELs. Consistent with the comparisons between WGS and WES data, this suggested PCR amplification induced a large number of error-prone INDELs to the library, and we could effectively increase INDEL calling quality by reducing the rate of PCR amplification.

To understand the behaviors of errors in the poly-A/T regions, we partitioned the INDEL call set by the same six regions again. We noticed that for the high quality events, a majority of the high confidence INDELs (68%) were within non-STRs regions (Figure 12). The proportion of poly-A/T INDELs was small for the high confidence call set (20%), larger for PCR-free-specific call set (35%), and even larger for PCR-involved-specific call set (51%). This was similar to the WGS and WES comparisons because there would be more poly-A/T INDELs when a higher rate of PCR amplification was performed. A majority of the high quality INDELs specific to PCR-involved data were within poly A (24%) and poly T regions (38%). When we compared the low quality INDELs to the high quality ones, there was consistent enrichment of poly-A/T INDELs in all three call sets, 2.3-fold for high confidence events, 2.3-fold for PCR-free-specific events, and 1.3-fold for PCR-involved-specific events. For INDELs specific to PCR-involved data and PCR-free data, poly-A/T INDELs represented a large proportion of the low quality INDELs: 80% and 62%, respectively. This again suggested that PCR amplification was a major source for low quality poly-A/T INDELs.

What coverage is required for accurate INDEL calling?

Ajay *et al.* 2011 reported that the number of SNVs detected exponentially increased until saturation at ~40-45X average coverage [28]. However, it was not clear what the coverage requirement should be for INDEL detection. To answer this question, we need a truth set to get accurate measurement. Fortunately, we were able to robustly measure sensitivity based on our high confidence INDELs. We down-sampled the reads, called INDELs again, and measured corresponding sensitivity for each sample (Methods). Figure 13 shows that we are missing 25% of the high confidence INDELs at a mean coverage of 30X. Even at 40X coverage recommended by Ajay *et al.* 2011 [28], we could

only discover 85% of the high confidence INDELs. At the level of a person's whole genome, in order to achieve a sensitivity of 95% for INDEL detection, we recommend that the sequencing requirement, at least on the Illumina HiSeq2000 platform, should be WGS at a mean coverage of 60X, after removing PCR duplicates (Figure 13). Of course, due to the above results, a WGS at 60X mean coverage with PCR-free library preparation is even more ideal.

Some groups previously reported that determining heterozygous SNPs requires higher coverage than homozygous ones [33]. The sensitivity of heterozygous SNP detection was limited by depth of coverage, which requires at least one read from each allele at any one site and in practice much more than one read to account for sequencing errors [34]. However, the read depth requirement of INDEL detection in terms of zygosity has not been well understood. To answer this question, we took the high confidence INDELs and partitioned them by zygosities. We first plotted the pair-wise coverage relationship between WGS and WES for each high confidence INDEL. Supplemental Figure S2 shows that the detection of homozygous INDELs starts with a lower coverage, which is consistent in both WGS and WES datasets, although the rest of the homozygotes and heterozygotes were highly overlapping. To further quantitatively understand this phenomenon, we measured the sensitivity again for heterozygous INDELs and homozygous INDELs separately. At a mean coverage of ~20X, the false negative rates of high confidence INDELs was ~45% for heterozygous INDELs and ~30% for homozygous INDELs (Figure 14). Figure 14 also shows that detection of heterozygous INDELs indeed requires higher coverage than homozygous ones (sensitivity of 95% at 60X vs. 50X). Notably, the number of heterozygous INDELs was ~1.6-fold higher than homozygous ones (~1600 vs. ~635 per sample). This re-affirms our recommendation of sequencing personal genome at 60X mean coverage to achieve a high accuracy INDEL call set.

Conclusions

Despite the fact that both WES and WGS have been widely used in biological studies and rare disease diagnosis, limitations of these techniques on INDEL calling are still not well characterized. One reason is that accurate INDEL calling is in general much more difficult than SNP calling. Another reason is that many groups tend to use WES, which we have determined is not ideal for INDEL calling for several reasons. We report here our characterization of calling errors for INDEL detection using Scalpel. As expected, higher coverage improves sensitivity of INDEL calling, and large INDEL detection is uniformly more difficult than detecting smaller INDELs. For large INDELs, insufficient coverage can result in more assembly errors, which yields a higher error rate of INDEL calling. In simulations, if we have Illumina-like reads with minimal sequencing errors, Scalpel can achieve a sensitivity of only 90% at 30X mean coverage, while maintaining a reasonable false discovery rate even for large INDELs.

There are several reasons for the low concordance for WGS and WES on INDEL detection. First, due to the low capture efficiency, WES failed to capture ~15% of candidate exons, subsequently missing possible INDELs in these regions. Second, even at sites that were successfully captured, there were more coverage biases in WES data, which led to only 30% of the regions being covered with 20 reads or more in regions with INDELs specific to WGS. Third, some microsatellite regions usually lacked coverage, which made it more difficult to assemble those regions. Fourth, PCR amplification introduces reads with higher INDEL error rate, especially in regions near homopolymer A/T's. Fifth, STR regions, especially homopolymer A/T regions were more likely to result in multiple candidates at the same locus. Sixth, WGS was able to detect the high confidence INDELs at a much lower coverage and a better uniformity across the genome, relative to WES. WGS data uniquely identifies ~10.8-fold more high-quality INDELs, and WES data uniquely identifies ~1.9-fold more low-quality INDELs. Seventh, WES is more likely to introduce a larger number of error-prone INDELs, while WGS is able to reveal a more sensitive and specific set of INDELs.

Our validation data showed that INDELs called by both WGS and WES data were indeed of high quality and with a low error rate. Even though the WGS data has much lower depth coverage in general, it is able to eliminate about one half of the false-positive INDEL calls. Homopolymer A/T INDELs are a major source of low quality INDELs and multiple signature events, and these are highly enriched in the WES data. Homopolymer A/T played an important role in giving rise to the low quality INDELs, which contributed much more INDEL errors. This was confirmed by the comparison of PCR-free and PCR-involved data on the sample NA12878. PCR-free data uniquely yields ~4.4-fold more high-quality INDELs and PCR-involved data uniquely yields ~6.3-fold more low-quality INDELs, implying that these errors are introduced with PCR amplification. Thus, one way to reduce the error rate of INDEL calling is to eliminate PCR amplification at the library preparation step. In terms of sensitivity, 95% of INDELs can be detected with 60X mean coverage WGS data from the Illumina HiSeq 2000 platform. We also quantitatively showed that accurate detection of heterozygous INDELs naturally requires ~1.2-fold higher coverage relative to that required for homozygous INDELs.

As more and more groups are moving on to these new micro-assembly based algorithms, practical considerations for experimental design should be introduced to the community. Here we present a novel classification scheme utilizing the validation data, and we encourage researchers to use this guideline for evaluating their call sets. The combination of alternative allele coverage and the k-mer Chi-Square score is an effective filter criterion for reducing INDEL calling errors. One could easily apply this criterion to identify both high-quality and low-quality INDELs, which could help to quickly understand the characteristics of an INDEL call. If we use INDEL callers with a very sensitive setting, we should control for homopolymer false-positive INDELs by applying a more rigorous filtering criteria. This is essential for population-scale sequencing projects, because the expense of experimental validation scales with the sample size. For consumer genome sequencing purposes, we recommend sequencing human genomes at 60X mean coverage with PCR-free protocols, which can substantially improve the quality of personal genomes. Although this recommendation might initially cost more than the current standard protocol of genome sequencing used by some facilities, we argue that

the significantly higher accuracy and decreased costs for validation with WGS at 60X with PCR-free protocols would ultimately be cost-effective as the sequencing costs continue to decrease, relative to either WES or WGS at a lower coverage.

Methods

Analysis of Simulated Data

We simulated Illumina-like 2*101 paired-end reads with randomly distributed INDELs, which ranged from 1 bp to 100 bp. The simulated reads were mapped to human reference genome hg19 using BWA-mem v0.7-6a using default parameters [35]. The alignment was sorted with SAMtools (v0.1.19-44428cd) [36] and the duplicates were marked with Picard using default parameters (v1.106), resulting in a mean coverage of 93X. We down-sampled the reads with Picard to generate 19 sub-alignments. The minimum mean coverage of the sub-alignments was 4.7X and increased by 4.7X each time, before it got to the original coverage (93X). Scalpel (v0.1.1) was used to assemble the reads and call INDELs from each alignment separately, resulting in 20 INDEL call-sets from these 20 alignments, using the following parameter setting: "--single --lowcov 1 --mincov 3 –outratio 0.1 --numprocs 10 --intarget". We then computed both the sensitivity and specificity of all and large (greater than 5bp) INDEL detection, respectively. The same versions and the same sets of parameter settings for BWA-mem, Picard, and Scalpel, were also used in the rest of the study, including the analysis of WGS/WES data, PCR-involved and PCR-free data.

Generation of WGS and WES data

Blood samples were collected from eight humans of two quartets from the Simons Simplex Collection. Both WGS and WES were performed on the same genomic DNA isolated from these eight blood samples. The exome capture kit used was NimbleGen SeqCap EZ Exome v2.0, which was designed to pull down 36Mb (approximately 300,000 exons) of the human genome hg19. The actual probe regions were much wider than these targeted regions, because probes also covered some flanking regions of genes, yielding a total size of 44.1Mb. All of the libraries were constructed with PCR amplification. We sequenced both sets of libraries on Illumina HiSeq2000 with average read length of 100 bp at the sequencing center of Cold Spring Harbor Laboratory (CSHL).

Analysis of the INDELs from WGS and WES data

We excluded all of the low quality raw reads, aligned the remaining high quality ones with BWA-mem, and mark-duplicated with Picard. We used Scalpel to assemble the reads and identify INDELs under both single mode and quad mode. The single mode outputs all of the putative INDELs per person, and the quad mode outputs only the putative de novo INDELs in the children in a family. We expanded each of the exons by 20bp upstream and 20bp downstream in order to cover the splicing sites and we called this set of expanded regions the "exonic targeted regions". The exonic targeted regions are fully covered by the exome capture probe regions. We excluded INDELs that were outside the exonic targeted regions in the downstream analysis.

We left-normalized the INDELs and compared the two call sets for the same person using two criteria: exact-match and position-match. Position-match means two INDELs have the same genomic coordinate, while exact-match additionally requires that two INDELs also have the same base-pair change(s). We called the INDELs in the intersection based on exact-match as high confidence INDELs because these INDELs were successfully sequenced, mapped, assembled, and called from two high coverage orthologous sequencing experiments. We named the INDELs only called from one dataset as "WGS-specific" and "WES-specific" INDELs, respectively. Regions of the above three categories of INDELs were partitioned and investigated separately. In particular, we focused on regions containing short tandem repeats (STRs) and homopolymers. We used BedTools (v2.18.1) with the region file from lobSTR (v2.04) to identify homopolymeric regions and other STRs (dual repeats, triplets and etc.) in the human genome [37, 38].

Generating summary statistics of alignment from WGS and WES

We used Qulimap (v0.8.1) to generate summary statistics of the alignment files of interest [39]. The reference genome used here is hg19. There were four region files that we used for this part of the analysis. The first one is the exon region bed file from NimbleGen. We generated the other three region files by expanding 25bp upstream and downstream around loci of high confidence INDELs, WGS-specific INDELs, and WES-specific INDELs, respectively. We followed all of the default settings except for requiring the homopolymer size to be at least five (-hm 5). Finally, we used Matplotlib to

generate the figures with the raw data from Qualimap under Python environment 2.7.2 [40].

Generation of MiSeq validation data

In this study, we used the MiSeq validation data on the sample K8101-49685s that we previously reported [31]. For more technical details, please refer to the original paper. In brief, library construction for the MiSeq Personal Sequencer platform (Illumina Inc.) was performed based on the TruSeq DNA Sample Prep LS protocol. We generated high quality 250 bp paired-end reads with an average coverage of ~47,000X over the selected INDELs. We aligned the reads with BWA-MEM (v0.7.5a) to hg19, sorted the alignment with SAMtools (v0.1.18) and marked PCR duplicates with Picard (v1.91). INDELs were realigned with the GATK IndelRealigner (v2.6-4), base quality scores were recalibrated, and finally variants were called with GATK UnifiedGenotyper (v2.4-3). All of the MiSeq data can be downloaded from the Sequence Read Archive (http://www.ncbi.nlm.nih.gov/sra) under project accession number SRX386284.

Classifications of INDEL with calling quality based on the validation data of sample K8101

We previously benchmarked Scalpel with respect to the coverage of the alternative allele (C_o^{Alt}) and the k-mer Chi-Square scores (χ^2). Scalpel applied the standard formula for the Chi-Square statistics and applied to the K-mer coverage of both alleles of an INDEL.

$$\chi^2 = \frac{(C_o^{Ref} - C_e^{Ref})^2}{C_e^{Ref}} + \frac{(C_o^{Alt} - C_e^{Alt})^2}{C_e^{Alt}}$$

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. $C_e^{Ref} = C_e^{Alt} = \frac{C_o^{Ref}+C_o^{Alt}}{2}$.

Among ~1400 INDELs that we selected for validation, 453 of them were called by Scalpel. Here we used all 453 INDELs to understand the relationship between the false discovery rate FDR and these two metrics (Supplemental Figure S3). Our validation data showed that with the same χ^2, INDELs with a lower C_o^{Alt} tend to have a higher FDR, especially for INDELs with C_o^{Alt} not greater than 10 (Supplemental Figure S3). For

INDELs with relatively the same C_o^{Alt}, a higher χ^2 also made them less likely to pass the validation. We noticed that the calling quality could be determined by the error rate inferred by these two metrics. To achieve a consistent accuracy for INDELs with different C_o^{Alt}, we classified INDELs from the WGS and WES data and determined the calling quality with the below criteria:

High quality INDELs: low error-rate (12%) INDELs meeting any of the three cutoffs: C_o^{Alt} >10 and χ^2 < 10.8, or 5 < C_o^{Alt} ≤ 10 and χ^2 ≤ 4.5, or C_o^{Alt} ≤ 5 and χ^2 ≤ 2;

Low quality INDELs: high error-rate (32%) INDELs meeting the following cutoff: C_o^{Alt} ≤ 10 and χ^2 >10.8;

Moderate quality: The remaining INDELs that do not fall into the above two categories.

Analysis of the effect of new filtering criteria on de novo INDEL calls

The two families in this study were previously reported in a population-scale autism study, with Sanger validation of de novo calls [32]. We used the de novo mode of Scalpel to identify de novo INDELs in these two families again, resulting in one de novo call set for WGS data and another de novo call set for WES data per family. We partitioned each call set by regions and filtered out the low quality INDELs. Iossifov *et.al* 2012 reported a total of N=85 de novo exonic INDELs in 343 families, i.e. there was ~0.1 de novo exonic INDEL per child [32]. If we assume a binomial distribution of the de novo exonic INDELs with an equal chance (p=1/343), the probability of seeing at least X de novo exonic INDELs in a given family in this study can be computed as below:

$$P(X \geq k \text{ INDELs}) = 1 - \sum_{0}^{k} P(X = k-1) = 1 - \sum_{0}^{k-1} \binom{N}{X} p^X q^{N-X}$$

where P(X ≥ k) is the probability of a given family having k or more de novo INDELs; N is the total number of exonic de novo INDELs reported, i.e. N=85; p is the probability of a hit on a given trial, i.e p=1/343; q=1-p.

Analysis of PCR-free and PCR-involved data of NA12878

We downloaded PCR-free WGS data of NA12878 (access Code: ERR194147), which is publicly available in the Illumina Platinum Genomes project. We also download a PCR-involved WGS data of NA12878 from the Sequence Read Archive (SRA) (access Code:

SRR533281, SRR533965, SRR539965, SRR539956, SRR539947, SRR539374, SRR539357). Both data were generated on Illumina HiSeq 2000 platform. Although the PCR-free data was not supposed to have any PCR duplicates, we observed a duplication rate of 2% as reported by Picard, and we excluded these reads, yielding ~50X mean coverage for both datasets after removing PCR duplicates. We used the same methods for alignment, INDEL calling, and downstream analysis as described above.

Analysis of INDEL detection sensitivity in WGS data

We were interested to know how depth of coverage affects the sensitivity of INDEL detection in WGS data. To accurately measure this sensitivity, one needs a robust call set as a truth set. Fortunately, we had exact-match INDELs concordant between high coverage WGS and high coverage WES data. We therefore measured sensitivity based on these high confidence INDELs, rather than on the whole set of INDELs, which might contain much more false positives. We down-sampled each WGS dataset to mean coverages of ~20X, ~32X, ~45X, and ~57X. We then used Scalpel to call INDELs from the resulting 4 sub-alignment files for each sample and computed the sensitivity at a certain mean coverage (X) for each sample by the equation:

$$\text{Sensitivity at X coverage} = \frac{\text{Number of high confidence INDELs called at X coverage}}{\text{Number of high confidence INDELs at the original coverage}}$$

This equation measures how many of the high confidence INDELs can be discovered as a function of read depth. We also analyzed the high confidence INDEL call set in terms of zygosity: high confidence heterozygous and homozygous INDEL, subsequently measuring the sensitivity with respect to different zygosities.

List of abbreviations used

Insertions and Deletions (INDELs), whole genome sequencing (WGS), whole exome sequencing (WES), next-generation sequencing (NGS), bp (base pair), PCR (polymerase chain reaction), short tandem repeats (STRs), homopolymer A (poly-A), homopolymer C (poly-C), homopolymer G (poly-G), homopolymer T (poly-T), short tandem repeats (STRs) except homopolymers (other STRs), homopolymer A or T (poly-A/T)

Competing Interest

The authors do not have any financial conflicts of interest to declare.

Authors' contributions

H.F. analyzed the data and wrote the manuscript. G.N. assisted in characterizing the simulation and validation data. J.A.O. designed the primers and analyzed the MiSeq data. Y.W. assisted in designing the primers and performed the MiSeq validation experiments. J.R. generated the WGS and WES data. M.R. supervised the generation of the WGS and WES data. I.I. developed the tool for the simulated data. H.F., M.C.S. and G.J.L. designed and analyzed the experiments. G.J.L. developed experimental design for INDEL validation, suggested, reviewed and supervised the data analysis, and wrote the manuscript. All of the authors have read and approved the final manuscript.

Authors' information

G.J.L.,M.C.S., M.R. and I.I. are faculty members at Cold Spring Harbor Laboratory (CSHL). G.N was a post-doc at CSHL and is currently employed at the New York Genome Center. J.R. is a lab technician at CSHL. H.F., J.A.O., and Y.W. are graduate students at Stony Brook University and CSHL.

Acknowledgements

The laboratory of G.J.L. is supported by funds from the Stanley Institute for Cognitive Genomics at Cold Spring Harbor Laboratory (CSHL). The laboratory of M.C.S. is

supported, in part, by National Institutes of Health award (R01-HG006677). The CSHL genome center is supported in part by a Cancer Center Support Grant (CA045508) from the NCI. This work was partially supported by a grant from the Simons Foundation (SF235988) to Michael Wigler. We are grateful to all of the families at the participating SFARI Simplex Collection (SSC) sites, as well as the principal investigators (A. Beaudet, R. Bernier, J. Constantino, E. Cook, E. Fombonne, D. Geschwind, D. Grice, A. Klin, R. Kochel, D. Ledbetter, C. Lord, C. Martin, D. Martin, R. Maxim, J. Miles, O. Ousley, B. Pelphrey, B. Peterson, J. Piggot, C. Saulnier, M. State, W. Stone, J. Sutcliffe, C. Walsh, and E. Wijsman). The DNA samples used in this work are included within SSC release 15. Approved researchers can obtain the SSC population dataset described in this study by applying at https://base.sfari.org. We thank S. Eskipehlivan for the technical assistance with the MiSeq validation experiments. We thank Sara Ballouz, Wim Verleyen, Jesse Gillis, Laura Jimenez Barron, Ruibang Luo, and Shane McCarthy for helpful discussions and comments on the paper.

References

1. Gudmundsson J, Sulem P, Gudbjartsson DF, Masson G, Agnarsson BA, Benediktsdottir KR, Sigurdsson A, Magnusson OT, Gudjonsson SA, Magnusdottir DN, et al: **A study based on whole-genome sequencing yields a rare variant at 8q24 associated with prostate cancer.** *Nature genetics* 2012, **44:**1326-1329.

2. Rope AF, Wang K, Evjenth R, Xing J, Johnston JJ, Swensen JJ, Johnson WE, Moore B, Huff CD, Bird LM, et al: **Using VAAST to identify an X-linked disorder resulting in lethality in male infants due to N-terminal acetyltransferase deficiency.** *American journal of human genetics* 2011, **89:**28-43.

3. Patel CJ, Sivadas A, Tabassum R, Preeprem T, Zhao J, Arafat D, Chen R, Morgan AA, Martin GS, Brigham KL, et al: **Whole genome sequencing in support of wellness and health maintenance.** *Genome Med* 2013, **5:**58.

4. O'Rawe JA, Fang H, Rynearson S, Robison R, Kiruluta ES, Higgins G, Eilbeck K, Reese MG, Lyon GJ: **Integrating precision medicine in the study and clinical treatment of a severely mentally ill person.** *PeerJ* 2013, **1:**e177.

5. Chen R, Mias GI, Li-Pook-Than J, Jiang L, Lam HYK, Chen R, Miriami E, Karczewski KJ, Hariharan M, Dewey FE, et al: **Personal omics profiling reveals dynamic molecular and medical phenotypes.** *Cell* 2012, **148:**1293-1307.

6. Hood L, Rowen L: **The human genome project: big science transforms biology and medicine.** *Genome Med* 2013, **5:**79.

7. Lyon GJ, Wang K: **Identifying disease mutations in genomic medicine settings: current challenges and how to accelerate progress.** *Genome medicine* 2012, **4:**58-58.

8. O'Rawe J, Jiang T, Sun G, Wu Y, Wang W, Hu J, Bodily P, Tian L, Hakonarson H, Johnson WE, et al: **Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing.** *Genome Med* 2013, **5:**28.

9. Dewey FE, Grove ME, Pan C, Goldstein BA, Bernstein JA, Chaib H, Merker JD, Goldfeder RL, Enns GM, David SP, et al: **Clinical interpretation and implications of whole-genome sequencing.** *JAMA* 2014, **311:**1035-1045.

10. Lupski JR, Belmont JW, Boerwinkle E, Gibbs RA: **Clan genomics and the complex architecture of human disease.** *Cell* 2011, **147:**32-43.

11. Lyon GJ, O'Rawe J: **Human genetics and clinical aspects of neurodevelopmental disorders.** In *The Genetics of Neurodevelopmental Disorders* (Mitchell K ed.: Cold Spring Harbor Labs Journals; 2014.

12. McClellan J, King M-C: **Genetic heterogeneity in human disease.** *Cell* 2010, **141:**210-217.

13. Ober C, Vercelli D: **Gene-environment interactions in human disease: nuisance or opportunity?** *Trends in genetics : TIG* 2011, **27:**107-115.

14. Clerget-Darpoux F, Elston RC: **Will formal genetics become dispensable?** *Hum Hered* 2013, **76:**47-52.

15. Weiss KM, Terwilliger JD: **How many diseases does it take to map a gene with SNPs?** *Nat Genet* 2000, **26:**151-157.

16. Lyon GJ: **Personalized medicine: Bring clinical standards to human-genetics research.** *Nature* 2012, **482:**300-301.

17. MacArthur DG, Manolio TA, Dimmock DP, Rehm HL, Shendure J, Abecasis GR, Adams DR, Altman RB, Antonarakis SE, Ashley EA, et al: **Guidelines for investigating causality of sequence variants in human disease.** *Nature* 2014, **508:**469-476.

18. Ross MG, Russ C, Costello M, Hollinger A, Lennon NJ, Hegarty R, Nusbaum C, Jaffe DB: **Characterizing and measuring bias in sequence data.** *Genome Biol* 2013, **14:**R51.

19. Clark MJ, Chen R, Lam HYK, Karczewski KJ, Chen R, Euskirchen G, Butte AJ, Snyder M: **Performance comparison of exome DNA sequencing technologies.** *Nature biotechnology* 2011, **29:**908-914.

20. Lam HY, Clark MJ, Chen R, Chen R, Natsoulis G, O'Huallachain M, Dewey FE, Habegger L, Ashley EA, Gerstein MB, et al: **Performance comparison of whole-genome sequencing platforms.** *Nat Biotechnol* 2012, **30:**78-82.

21. Bamshad MJ, Ng SB, Bigham AW, Tabor HK, Emond MJ, Nickerson DA, Shendure J: **Exome sequencing as a tool for Mendelian disease gene discovery.** *Nature reviews Genetics* 2011, **12:**745-755.

22. Bamshad MJ, Shendure JA, Valle D, Hamosh A, Lupski JR, Gibbs RA, Boerwinkle E, Lifton RP, Gerstein M, Gunel M, et al: **The Centers for Mendelian Genomics: a new large-scale initiative to identify the genes**

underlying rare Mendelian conditions. *Am J Med Genet A* 2012, **158A:**1523-1525.

23. Metzker ML: **Sequencing technologies - the next generation.** *Nature reviews Genetics* 2010, **11:**31-46.

24. Zhu M, Need AC, Han Y, Ge D, Maia JM, Zhu Q, Heinzen EL, Cirulli ET, Pelak K, He M, et al: **Using ERDS to infer copy-number variants in high-coverage genomes.** *Am J Hum Genet* 2012, **91:**408-421.

25. Mullaney JM, Mills RE, Pittard WS, Devine SE: **Small insertions and deletions (INDELs) in human genomes.** *Hum Mol Genet* 2010, **19:**R131-136.

26. Mills RE, Pittard WS, Mullaney JM, Farooq U, Creasy TH, Mahurkar AA, Kemeza DM, Strassler DS, Ponting CP, Webber C, Devine SE: **Natural genetic variation caused by small insertions and deletions in the human genome.** *Genome Res* 2011, **21:**830-839.

27. Mills RE, Luttig CT, Larkins CE, Beauchamp A, Tsui C, Pittard WS, Devine SE: **An initial map of insertion and deletion (INDEL) variation in the human genome.** *Genome Res* 2006, **16:**1182-1190.

28. Ajay SS, Parker SC, Abaan HO, Fajardo KV, Margulies EH: **Accurate and comprehensive sequencing of personal genomes.** *Genome Res* 2011, **21:**1498-1505.

29. Li H: **Towards Better Understanding of Artifacts in Variant Calling from High-Coverage Samples.** In *ArXiv e-prints*, vol. 1404. pp. 929; 2014:929.

30. Iqbal Z, Caccamo M, Turner I, Flicek P, McVean G: **De novo assembly and genotyping of variants using colored de Bruijn graphs.** *Nat Genet* 2012, **44:**226-232.

31. Narzisi G, O' Rawe JA, Iossifov I, Fang H, Lee Y-h, Wang Z, Wu Y, Lyon GJ, Wigler M, Schatz MC: **Accurate detection of de novo and transmitted INDELs within exome-capture data using micro-assembly.** *bioRxiv* 2014.

32. Iossifov I, Ronemus M, Levy D, Wang Z, Hakker I, Rosenbaum J, Yamrom B, Lee YH, Narzisi G, Leotta A, et al: **De novo gene disruptions in children on the autistic spectrum.** *Neuron* 2012, **74:**285-299.

33. McKernan KJ, Peckham HE, Costa GL, McLaughlin SF, Fu Y, Tsung EF, Clouser CR, Duncan C, Ichikawa JK, Lee CC, et al: **Sequence and structural variation in a human genome uncovered by short-read, massively parallel**

ligation sequencing using two-base encoding. *Genome research* 2009, **19:**1527-1541.

34. Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, McGuire A, He W, Chen Y-J, Makhijani V, Roth GT, et al: **The complete genome of an individual by massively parallel DNA sequencing.** *Nature* 2008, **452:**872-876.

35. Li H: **Aligning sequence reads , clone sequences and assembly contigs with BWA-MEM.** *arXiv Prepr* 2013.

36. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S: **The Sequence Alignment/Map format and SAMtools.** *Bioinformatics* 2009, **25:**2078-2079.

37. Quinlan AR, Hall IM: **BEDTools: a flexible suite of utilities for comparing genomic features.** *Bioinformatics* 2010, **26:**841-842.

38. Gymrek M, Golan D, Rosset S, Erlich Y: **lobSTR: A short tandem repeat profiler for personal genomes.** *Genome Research* 2012.

39. García-Alcalde F, Okonechnikov K, Carbonell J, Cruz LM, Götz S, Tarazona S, Dopazo J, Meyer TF, Conesa A: **Qualimap: evaluating next-generation sequencing alignment data.** *Bioinformatics* 2012, **28:**2678-2679.

40. Hunter JD: **Matplotlib: A 2D Graphics Environment.** *Computing in Science & Engineering* 2007, **9:**90-95.

Figures

Figure 1

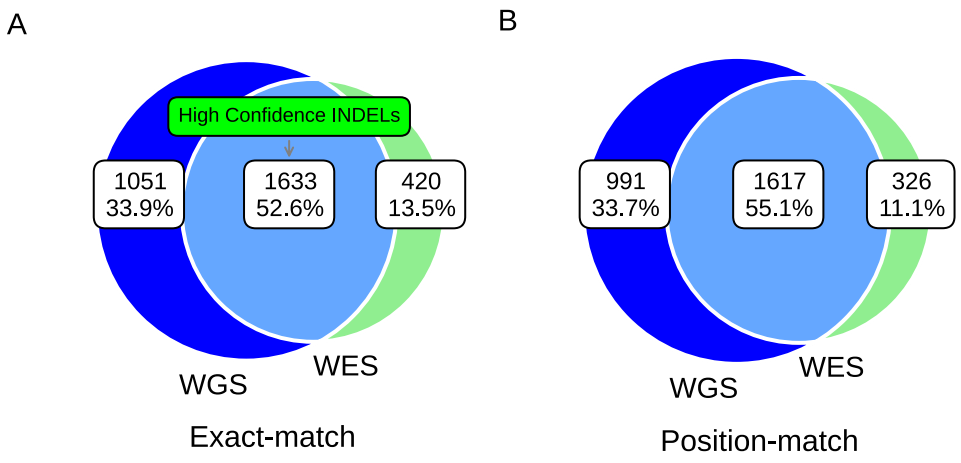


Figure 2

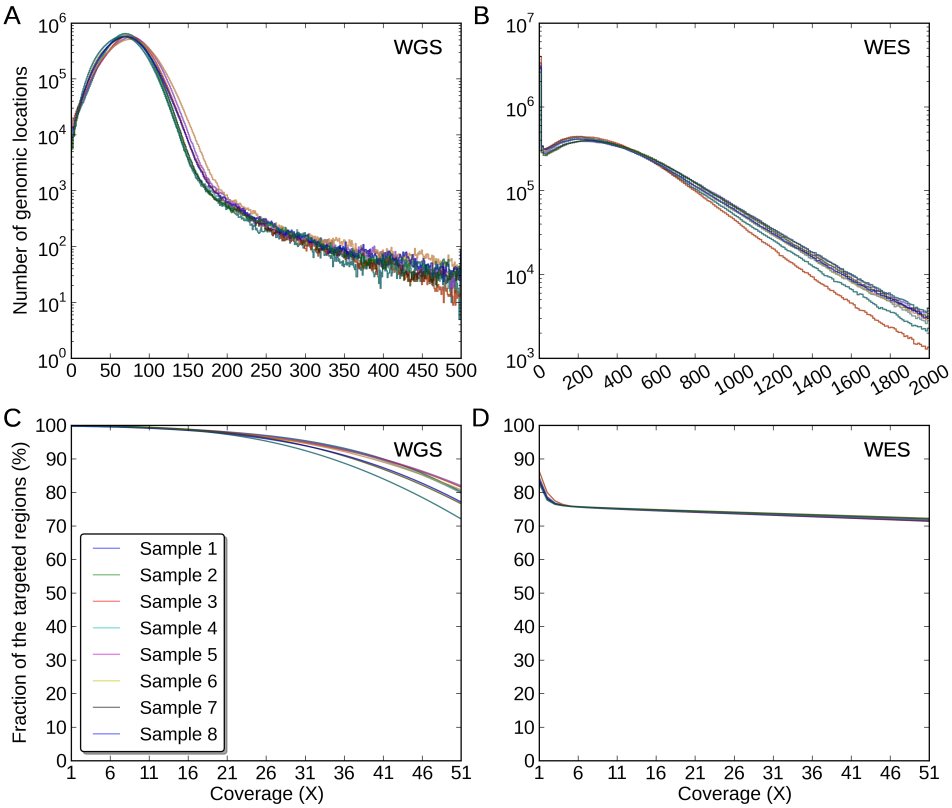

A
WGS
Number of genomic locations
B
WES
C
WGS
Fraction of the targeted regions (%)
Coverage (X)
Sample 1
Sample 2
Sample 3
Sample 4
Sample 5
Sample 6
Sample 7
Sample 8
D
WES
Coverage (X)

Figure 3

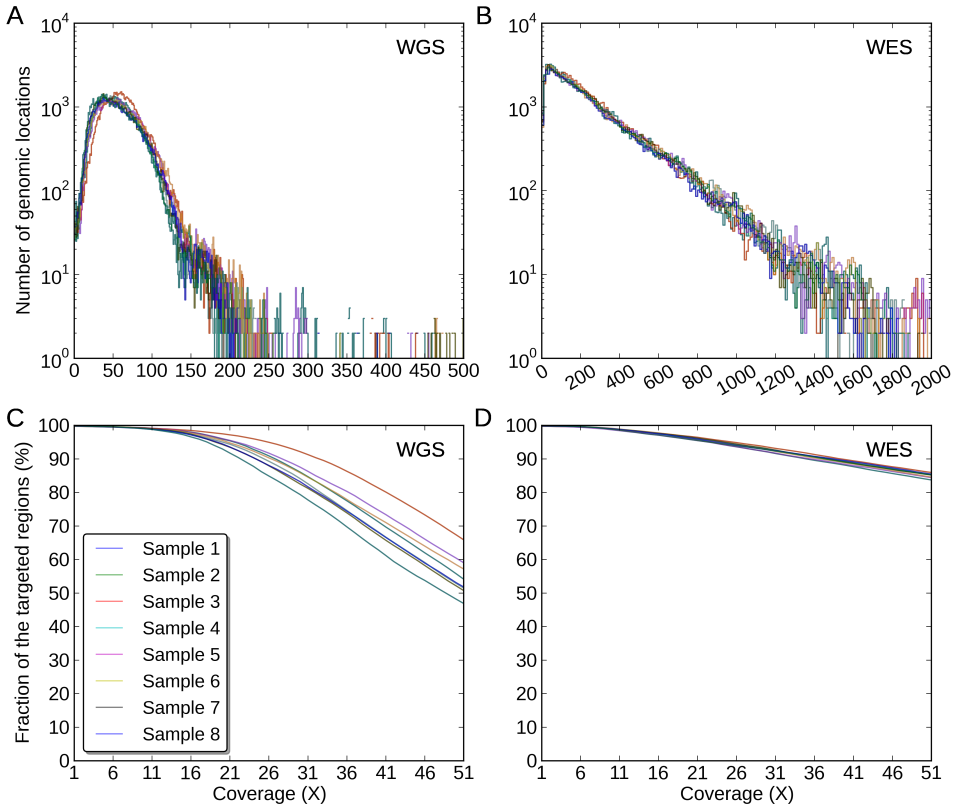

A
B
C
D
WGS
WES
WGS
WES
Number of genomic locations
Fraction of the targeted regions (%)
Coverage (X)
Coverage (X)
Sample 1
Sample 2
Sample 3
Sample 4
Sample 5
Sample 6
Sample 7
Sample 8

Figure 4

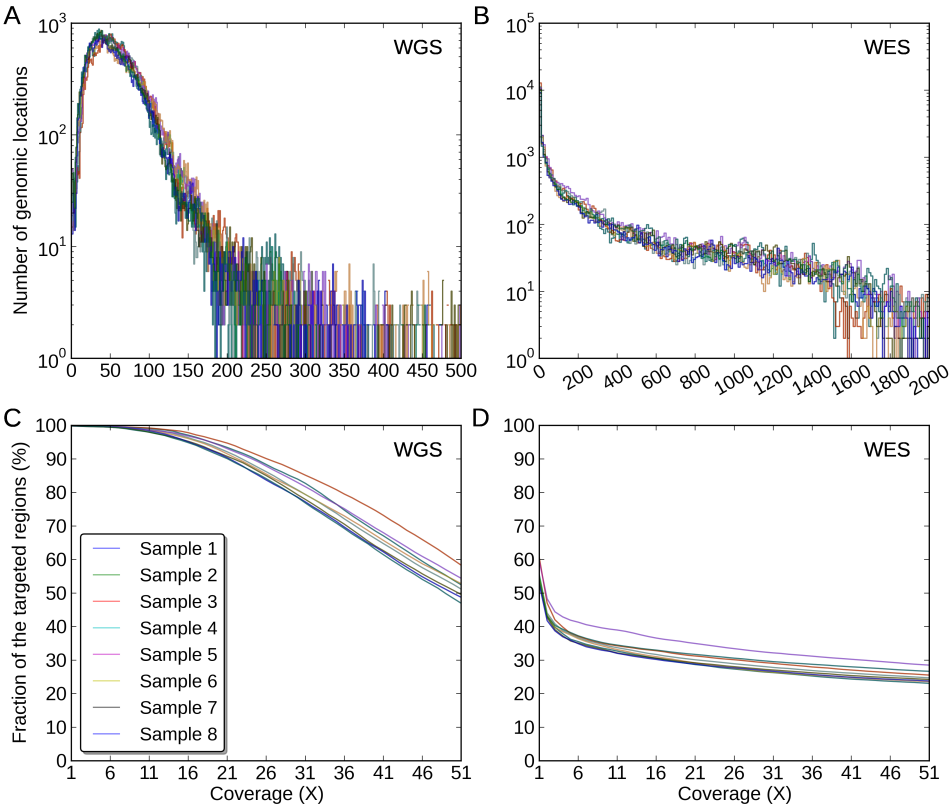

A
WGS
Number of genomic locations
B
WES
C
WGS
Fraction of the targeted regions (%)
Coverage (X)
Sample 1
Sample 2
Sample 3
Sample 4
Sample 5
Sample 6
Sample 7
Sample 8
D
WES
Coverage (X)

Figure 5

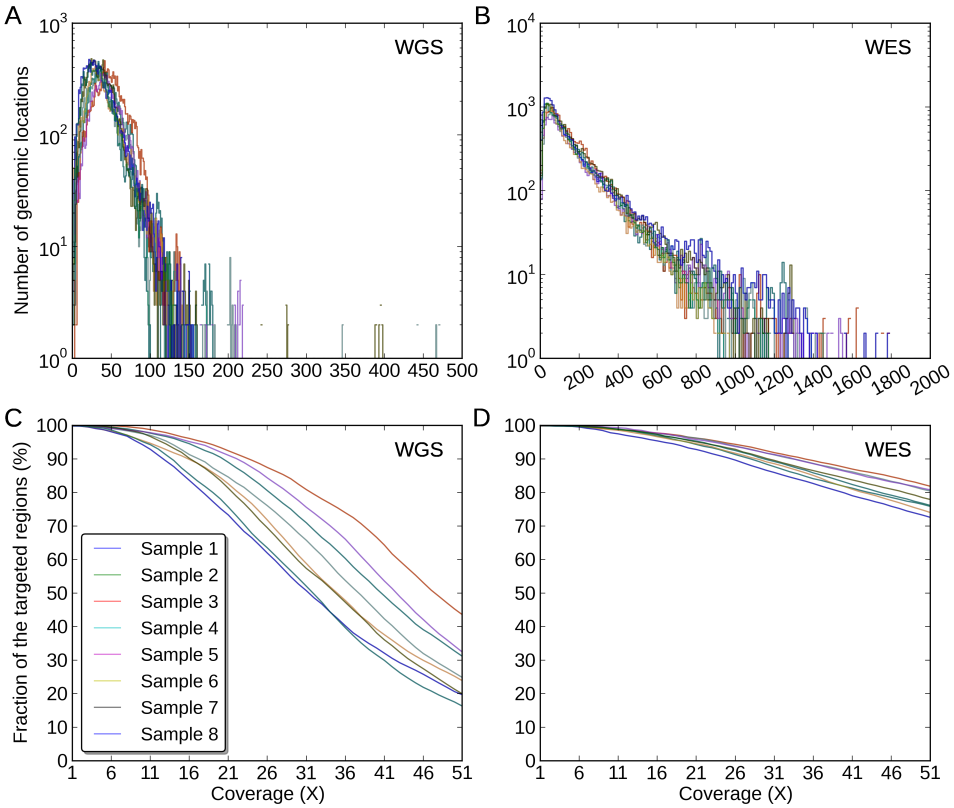

A
WGS
Number of genomic locations
B
WES
C
WGS
Fraction of the targeted regions (%)
Coverage (X)
Sample 1
Sample 2
Sample 3
Sample 4
Sample 5
Sample 6
Sample 7
Sample 8
D
WES
Coverage (X)

Figure 6

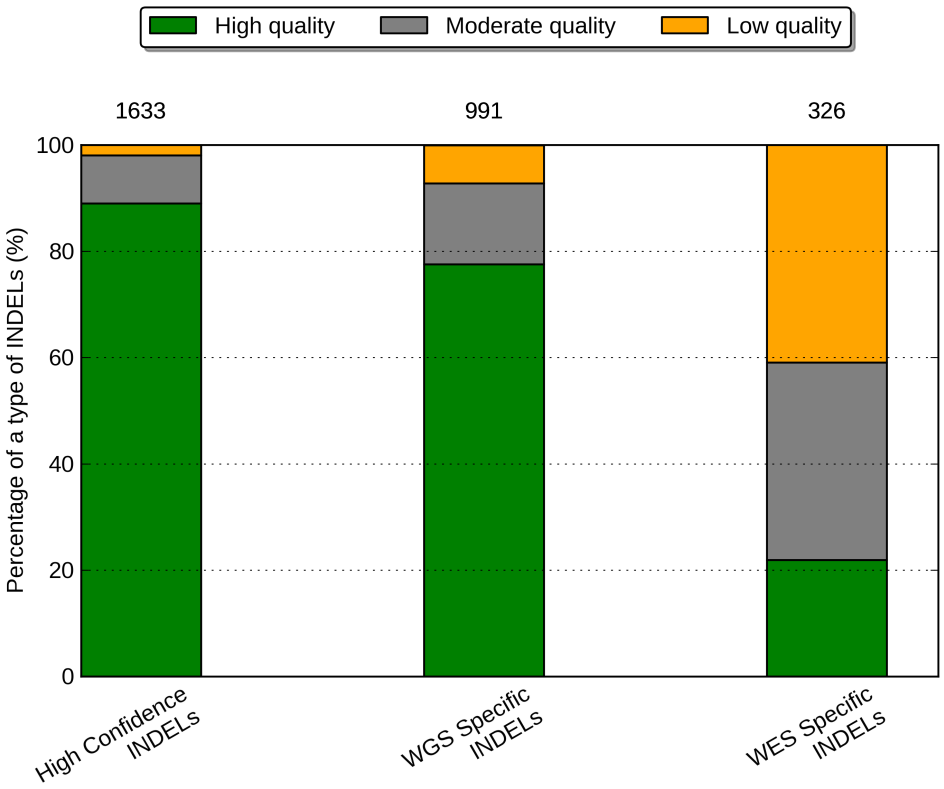

High quality
Moderate quality
Low quality
1633
991
326
100
80
60
40
20
0
Percentage of a type of INDELs (%)
High Confidence INDELs
WGS Specific INDELs
WES Specific INDELs

Figure 7

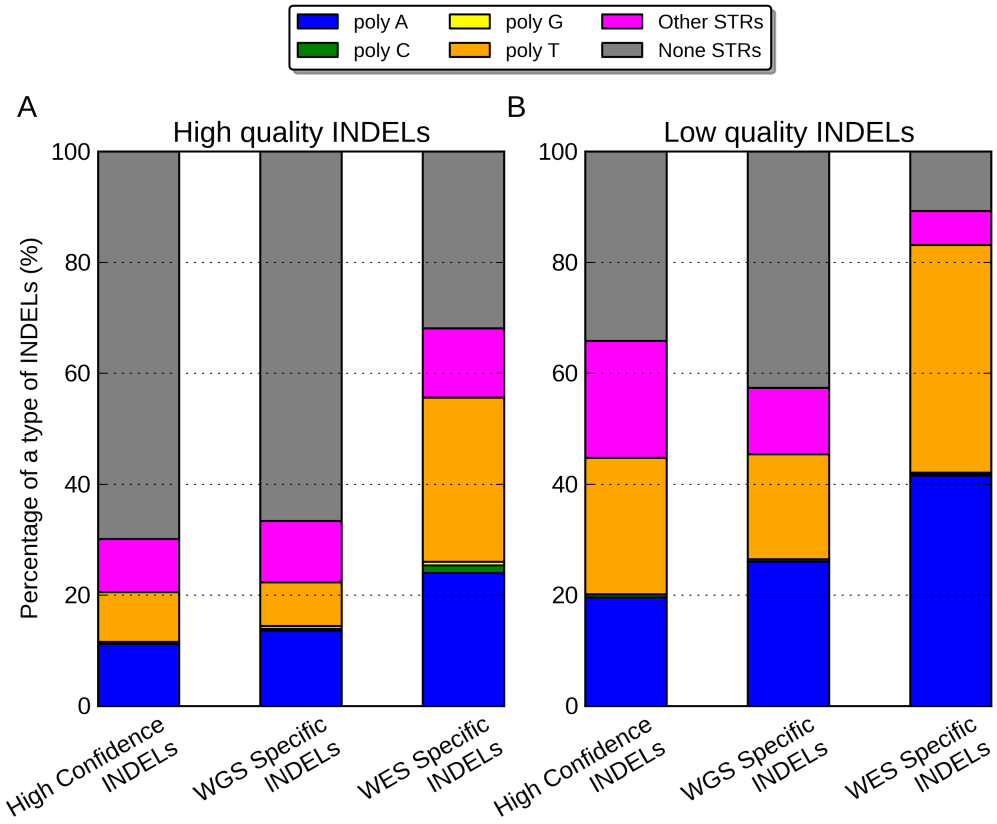

poly A
poly C
poly G
poly T
Other STRs
None STRs
A
High quality INDELs
B
Low quality INDELs
Percentage of a type of INDELs (%)
100
80
60
40
20
0
High Confidence INDELs
WGS Specific INDELs
WES Specific INDELs

Figure 8

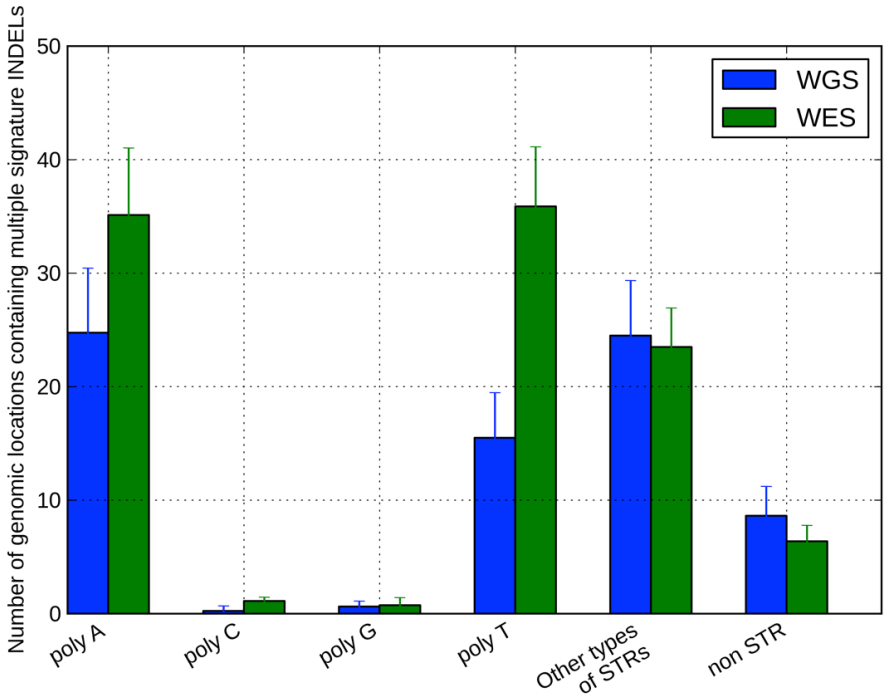

Number of genomic locations containing multiple signature INDELs
50
40
30
20
10
0
WGS
WES
poly A
poly C
poly G
poly T
Other types of STRs
non STR

Figure 9

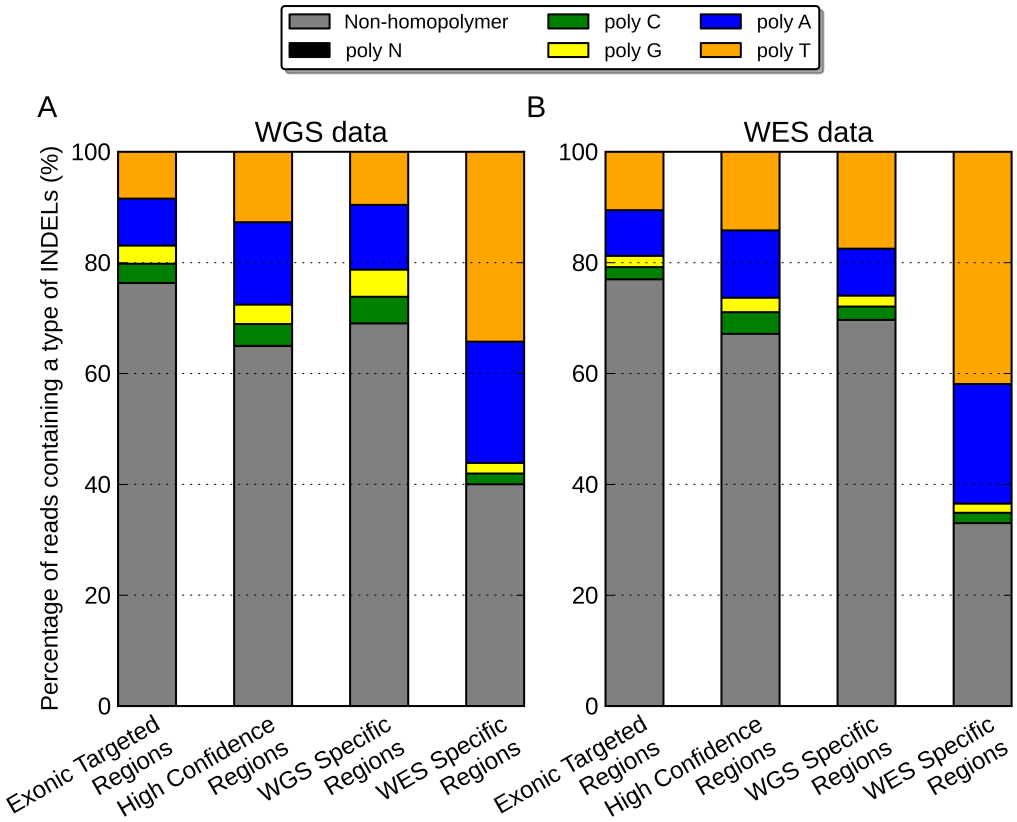

Non-homopolymer
poly N
poly C
poly G
poly A
poly T
A
WGS data
B
WES data
Percentage of reads containing a type of INDELs (%)
100
80
60
40
20
0
Exonic Targeted Regions
High Confidence Regions
WGS Specific Regions
WES Specific Regions

Figure 10

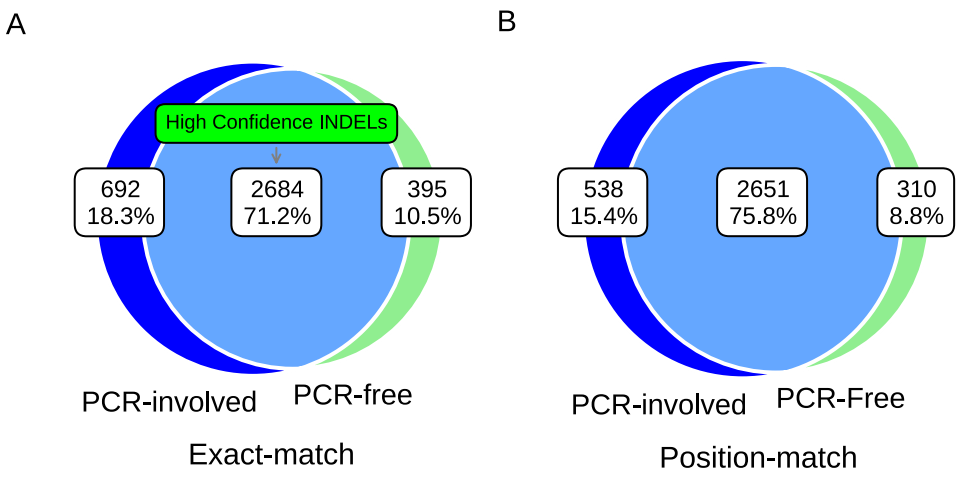

A
High Confidence INDELs
692
18.3%
2684
71.2%
395
10.5%
PCR-involved
PCR-free
Exact-match
B
538
15.4%
2651
75.8%
310
8.8%
PCR-involved
PCR-Free
Position-match

Figure 11

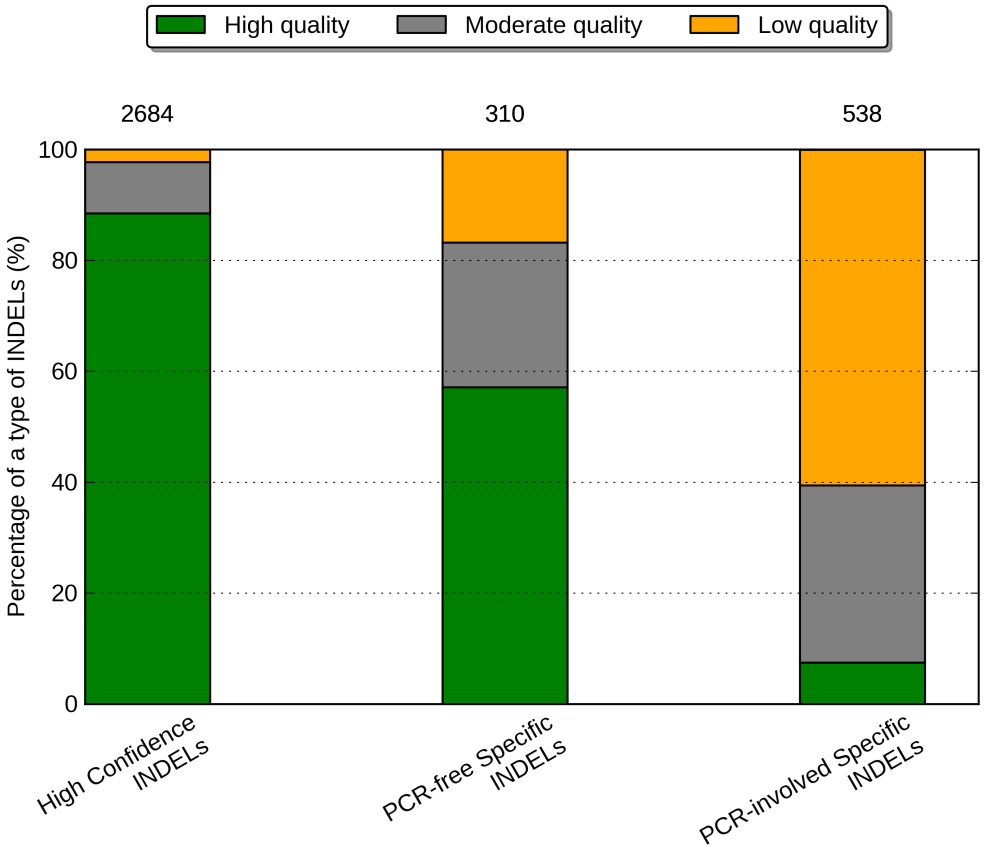

High quality
Moderate quality
Low quality
2684
310
538
100
80
60
40
20
0
Percentage of a type of INDELs (%)
High Confidence INDELs
PCR-free Specific INDELs
PCR-involved Specific INDELs

Figure 12

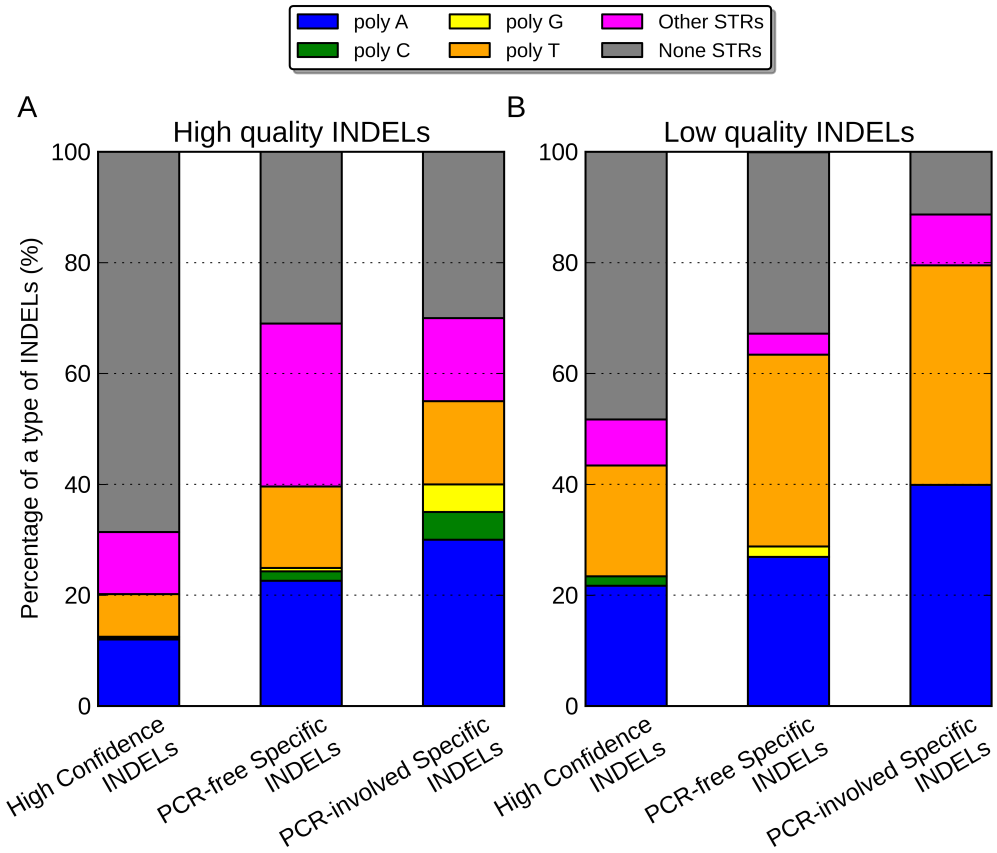

poly A
poly C
poly G
poly T
Other STRs
None STRs
A
High quality INDELs
B
Low quality INDELs
Percentage of a type of INDELs (%)
0
20
40
60
80
100
High Confidence INDELs
PCR-free Specific INDELs
PCR-involved Specific INDELs

Figure 13

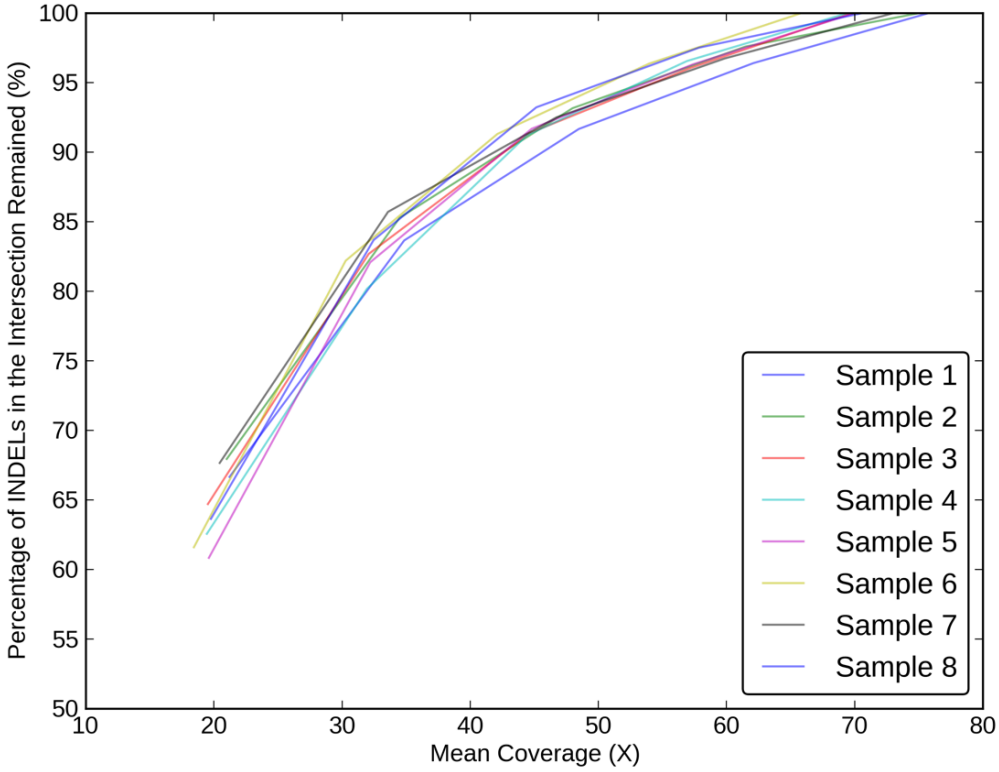
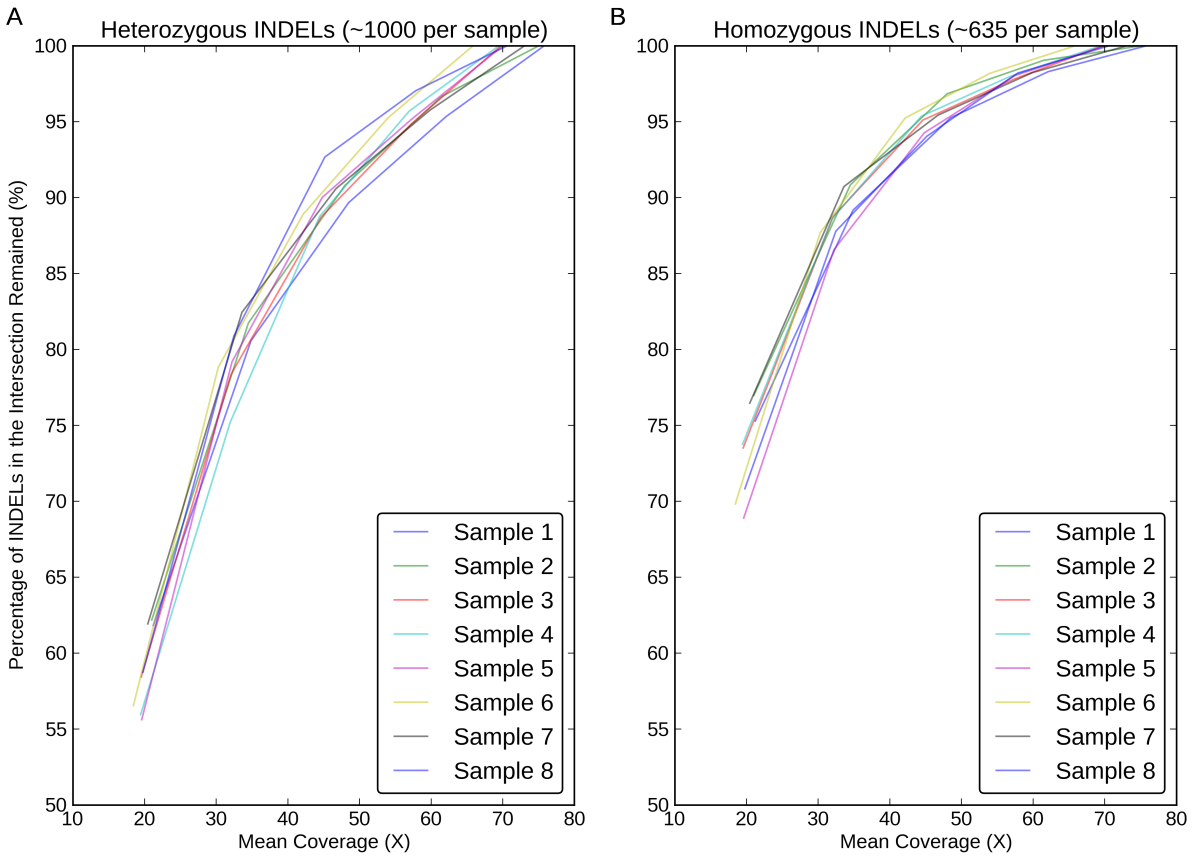

Percentage of INDELs in the Intersection Remained (%)
Mean Coverage (X)
100
95
90
85
80
75
70
65
60
55
50
10
20
30
40
50
60
70
80
Sample 1
Sample 2
Sample 3
Sample 4
Sample 5
Sample 6
Sample 7
Sample 8

Figure 14


A
Heterozygous INDELs (~1000 per sample)
Percentage of INDELs in the Intersection Remained (%)
Mean Coverage (X)
Sample 1
Sample 2
Sample 3
Sample 4
Sample 5
Sample 6
Sample 7
Sample 8
B
Homozygous INDELs (~635 per sample)
Mean Coverage (X)
Sample 1
Sample 2
Sample 3
Sample 4
Sample 5
Sample 6
Sample 7
Sample 8

Figures legends

Figure 1. **Mean concordance of INDELs over eight samples between WGS (blue) and WES (green) data.** Venn diagram showing the numbers and percentage of shared between data types based on (A) Exact-match (B) Position-match. The mean concordance rate increased when we required at least a certain number of reads in both data (Table 1).

Figure 2. **Coverage distributions of the exonic targeted regions in (A) the WGS data, (B) the WES data.** The Y-axis for A) and B) is of log10-scale. The coverage fractions of the exonic targeted regions from 1X to 51X in (C) the WGS data, (D) the WES data.

Figure 3. **Coverage distributions of the high confidence INDELs regions in (A) the WGS data, (B) the WES data**. The Y-axis for A) and B) is of log10-scale. The coverage fractions of the high confidence INDELs regions from 1X to 51X in (C) the WGS data, (D) the WES data.

Figure 4. **Coverage distributions of the WGS-specific INDELs regions in (A) the WGS data, (B) the WES data**. The Y-axis for A) and B) is of log10-scale. The coverage fractions of the WGS-specific INDELs regions from 1X to 51X in (C) the WGS data, (D) the WES data.

Figure 5. **Coverage distributions of the WES-specific INDELs regions in (A) the WGS data, (B) the WES data**. The Y-axis for A) and B) is of log10-scale. The coverage fractions of the WES-specific INDELs regions from 1X to 51X in (C) the WGS data, (D) the WES data.

Figure 6. **Percentage of high quality, moderate quality and low quality INDELs in three call set.** (A) the high confidence INDELs, (B) the WGS-specific INDELs, (C) the WES-specific INDELs. The numbers on top of a call set represent the mean number of INDELs in that call set over eight samples.

Figure 7. **Percentage of poly-A, poly-C, poly-G, poly-T, other-STRs, and none-STRs in three call set.** (A) high quality INDELs, (B) low quality INDELs. In both figures, from left to the right are high confidence INDELs, WGS-specific INDELs, and WES-specific INDELs.

Figure 8. **Numbers of genomic locations containing multiple signature INDELs in WGS (blue) and WES data (green)**. The height of the bar represents the mean across eight samples and the error bar represent the standard deviation across eight samples.

Figure 9. **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**. In both figures, from left to the right are exonic targeted regions, high confidence INDELs, WGS-specific INDELs, and WES-specific INDELs.

Figure 10. **Concordance of INDEL detection between PCR-free and PCR-involved data on NA12878.** Venn diagram showing the numbers and percentage of shared between data types based on (A) Exact-match (B) Position-match.

Figure 11. **Percentage of high quality, moderate quality and low quality INDELs in two datasets**. (A) the high confidence INDELs, (B) the PCR-free-specific INDELs, (C) the PCR-involved-specific INDELs. The numbers on top of a call set represent the number of INDELs in that call set.

Figure 12. **Percentage of poly-A, poly-C, poly-G, poly-T, other-STRs, and non-STRs in (A) high quality INDELs, (B) low quality INDELs**. In both figures, from left to the right are high confidence INDELs, INDELs specific to PCR-free data, and INDELs specific to PCR-involved data.

Figure 13. **Sensitivity performance of INDEL detection with eight WGS datasets at different mean coverages on Illumina HiSeq2000 platform**. The Y-axis represents the percentage of the high confidence INDELs revealed at a certain lower mean coverage.

Figure 14. **Sensitivity performance of (A) heterozygous and (B) homozygous INDEL detection with eight WGS datasets at different mean coverages on Illumina HiSeq2000 platform**. The Y-axis represent the percentage of the high confidence INDELs revealed at a certain lower mean coverage.

Tables

Table 1

Concordance Rate	Without filtering	≥ 1 read	≥ 20 reads	≥ 40 reads	≥ 60 reads	≥ 80 reads
Exact-match	53% (0.8%)	62% (1.1%)	69% (1.5%)	73% (2.3%)	76% (1.6%)	74% (1.3%)
Position-match	55% (0.8%)	66% (1.0%)	73% (1.1%)	77% (1.8%)	79% (1.1%)	76% (1.3%)
Discordance Rate	**Without filtering**	**≥ 1 read**	**≥ 20 reads**	**≥ 40 reads**	**≥ 60 reads**	**≥ 80 reads**
WGS-Specific	34% (1.4%)	20% (1.5%)	14% (1.6%)	14% (2.2%)	15% (2.5%)	20% (3.2%)
WES-Specific	11% (1.2%)	14% (1.4%)	13% (1.3%)	9% (2.6%)	6% (2.2%)	4% (1.5%)

Table legends

Table 1. **Mean concordance and discordance rates of INDEL detection between WGS and WES data in different regions**. The data is shown in the following order: 1) regions without filtering, and regions filtered by requiring base coverage to be at least 2) one read, 3) 20 reads, 4) 40 reads, 5) 60 reads, or 6) 80 reads in both data. The mean discordance rate is calculated based on position-match, which is the percentage of INDELs specific to either dataset. The standard deviation is shown in parenthesis.

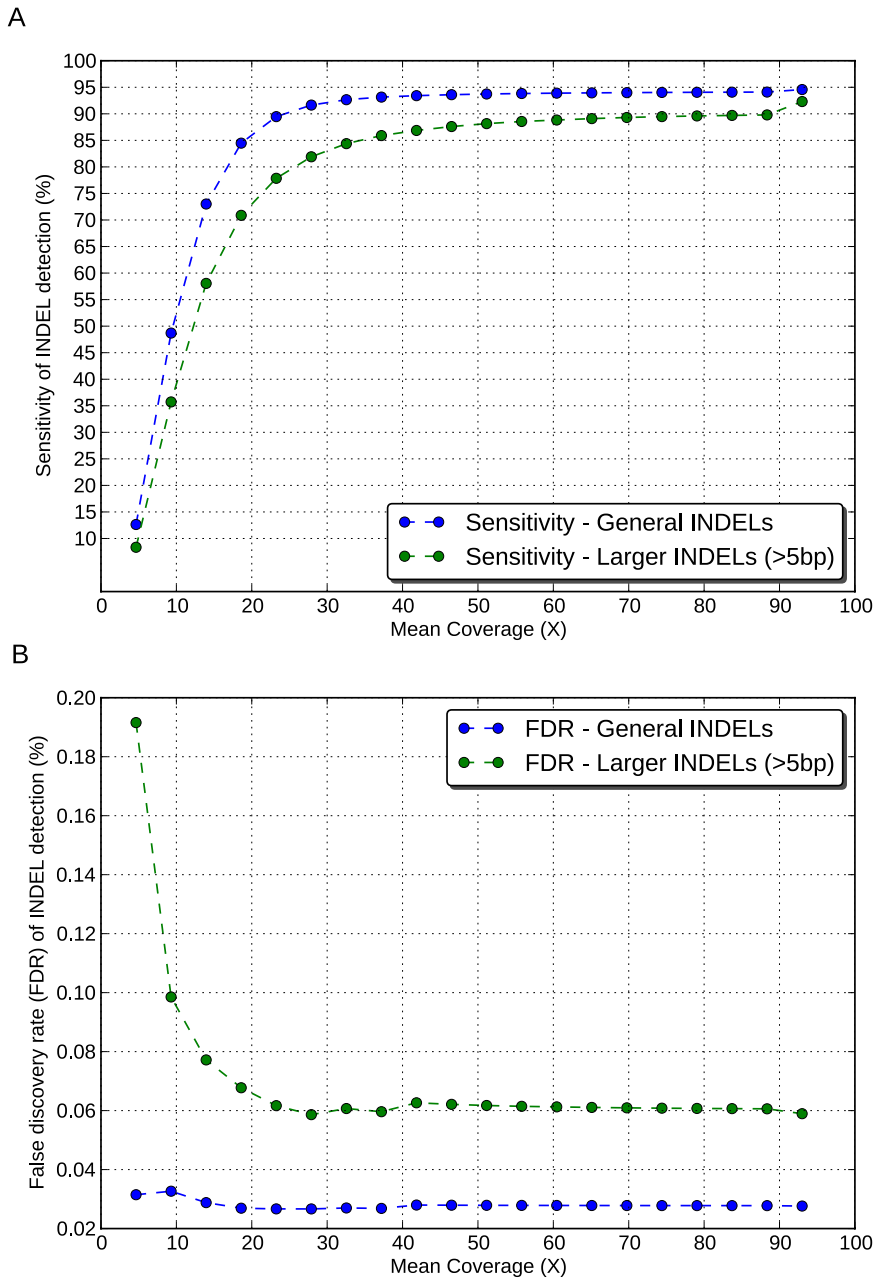
Additional files

Additional file 1 – Supplemental figures and tables
This file includes supplemental figure S1-S3, supplemental table S1-S4, and their corresponding figure/table legends.

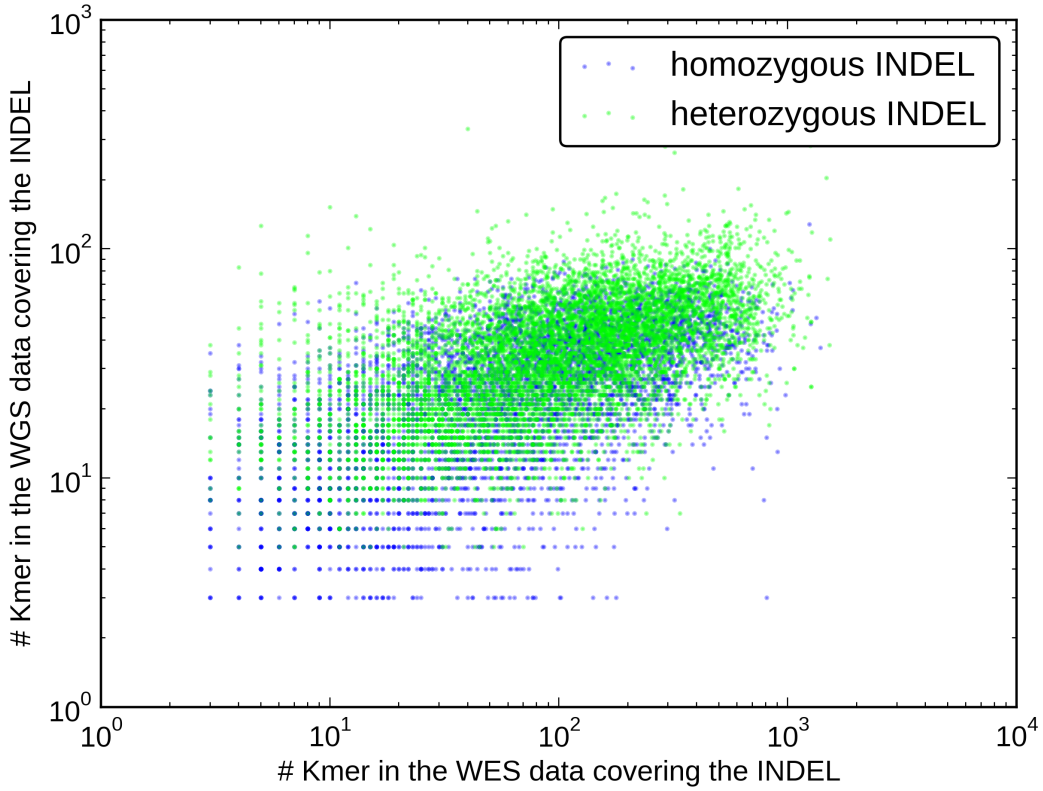
Additional Data File 1

Supplemental Figures

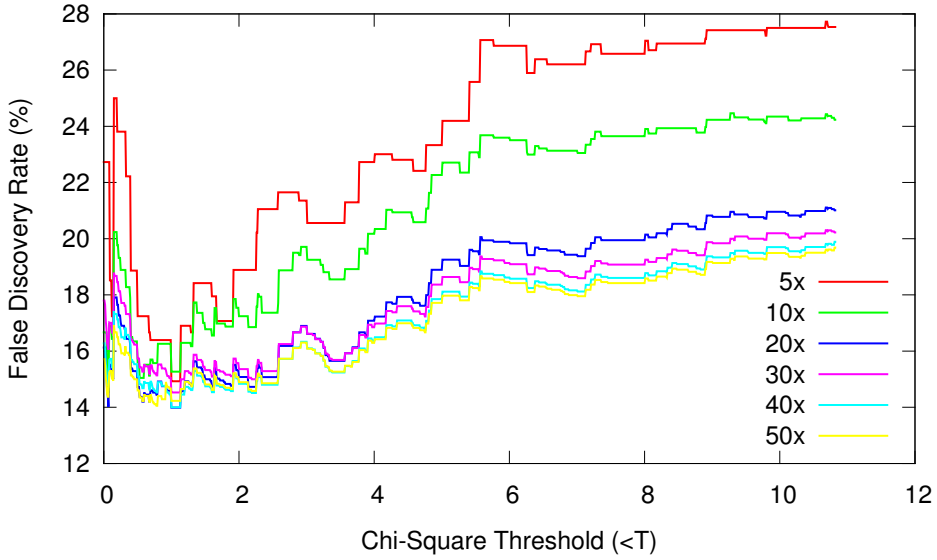
Supplemental Figure S1



Supplemental Figure S2



Supplemental Figure S3



Supplemental figure legends

Supplemental Figure S1. **Performance of sensitivity (A) and false discovery rate (B) at different coverage based on simulation data**. Each dot represent one down-sampled experiment. Blue dots represent performance of general INDELs (i.e. INDELs of size starting at 1bp) and green dots represent performance of large INDELs (i.e. INDELs of size greater than 5bp).

Supplemental Figure S2. **Pair-wise base coverage relationship of INDEL called by both WGS and WES data**. These INDELs were partitioned by zygosities: homozygous (blue) and heterozygous INDELs (green). The X-axis shows the number of k-mer covering an INDEL in the WES data, and the Y-axis shows the number for WGS data.

Supplemental Figure S3. **Characterization of the false discovery rate (FDR) based on validation data**. INDELs were partitioned based on k-mer coverage of the alternative allele and the INDEL Chi-Square scores. The X-axis shows Chi-Square scores of INDELs less than a certain threshold, and the Y-axis represents the FDR.

Supplemental Tables

Supplemental Table S1

	High quality	Moderate quality	Low quality
High confidence INDELs	89% (0.7%) 1454 (11.7)	9% (0.5%) 148 (7.3)	2% (0.5%) 31 (8.3)
WGS-specific INDELs	78% (1.4%) 769 (13.9)	15% (1.1%) 151 (10.7)	7% (1.6%) 71 (15.8)
WES-specific INDELs	22% (3.4%) 71 (11.2)	37% (3.7%) 121 (11.9)	41% (3.3%) 133 (10.9)

Supplemental Table S2

Number of reads	Exonic targeted regions	High confidence INDEL regions	WGS-specific INDEL regions	WES-specific INDEL regions
WGS	241984	49008	26417	1775
WES	815945	205698	44346	11251

Supplemental Table S3

Number of INDELs	**= 0**	**≥ 1**	**≥ 2**
Probability	0.78	0.22	0.03
Number of INDELs	**≥ 3**	**≥ 4**	**≥ 5**
Probability	0.0020	0.0001	0.000005

Supplemental Table S4

Putative *de novo*	WGS (poly-A, poly-T, ms)	WES (poly-A, poly-T, ms)	WGS (After filtering)	WES (After filtering)
Family 1	45 (27,14,4)	5 (3,1,1)	0	0
Family 2	49 (24,22,3)	17 (8,5,4)	0	0

Supplemental Table legends

Supplemental Table S1. **Mean percentage and mean number of high quality, moderate quality, low quality INDELs in each call set.** The mean percentage and the mean number over eight samples are shown in the upper and the lower of a cell, respectively. The standard deviation is shown in parenthesis.

Supplemental Table S2. **Number of reads in the following four regions: Exonic targeted regions, high confidence INDEL regions, WGS-specific INDEL regions, WES-specific INDEL regions**.

Supplemental Table S3. **Probabilities of seeing k or more INDELs in a given family assuming a binomial distribution**. Here we assumed a binomial distribution of the de novo exonic INDELs in the 343 SSC families.

Supplemental Table S4. **Putative de novo exonic INDELs in these two families before and after applying filtering critiria**. The number of INDELs within regions of homopolymer A (poly-A), homopolymer T (poly-T), and microsatellites (ms) are shown in the parenthesis.