SCIENTIFIC REPORTS
natureresearch

OPEN

Deep whole-genome sequencing of 3 cancer cell lines on 2 sequencing platforms

Kanika Arora, Minita Shah, Molly Johnson, Rashesh Sanghvi, Jennifer Shelton, Kshithija Nagulapalli, Dayna M. Oschwald, Michael C. Zody, Soren Germer, Vaidehi Jobanputra, Jade Carter & Nicolas Robine*

To test the performance of a new sequencing platform, develop an updated somatic calling pipeline and establish a reference for future benchmarking experiments, we performed whole-genome sequencing of 3 common cancer cell lines (COLO-829, HCC-1143 and HCC-1187) along with their matched normal cell lines to great sequencing depths (up to 278x coverage) on both Illumina HiSeqX and NovaSeq sequencing instruments. Somatic calling was generally consistent between the two platforms despite minor differences at the read level. We designed and implemented a novel pipeline for the analysis of tumor-normal samples, using multiple variant callers. We show that coupled with a high-confidence filtering strategy, the use of combination of tools improves the accuracy of somatic variant calling. We also demonstrate the utility of the dataset by creating an artificial purity ladder to evaluate the somatic pipeline and benchmark methods for estimating purity and ploidy from tumor-normal pairs. The data and results of the pipeline are made accessible to the cancer genomics community.

The field of cancer genomics has exploded with the development of high-throughput sequencing, largely driven by Illumina's short read sequencing technology. Thousands of tumors have been sequenced in the last decade, with strategies varying from variant hotspot panels[1], cancer gene panels[2], whole-exome (such as used in The Cancer Genome Atlas project[3]) or whole-genome sequencing (WGS)[4]. In 2014, Illumina introduced the HiSeq X Ten (HiSeqX) as their main sequencing instrument dedicated to human whole-genome sequencing. In 2017, they released the NovaSeq 6000 Sequencing System (NovaSeq), which is currently the latest generation of Illumina sequencing instruments. The primary difference between these platforms is the adoption of 2-channel Sequencing-by-Synthesis in the NovaSeq where clusters detected in the red wavelength filter correspond to a C nucleotide, clusters detected in the green wavelength filter correspond to a T, clusters detected by both colors correspond to A, and unlabeled clusters are G bases. For the NovaSeq, Illumina introduced a new base calling algorithm and method for estimating quality scores, with 4 quality bins (as opposed to 8 bins for HiSeqX).

With the introduction of any new sequencing technology, it is important to investigate the error profiles and biases of the technology, and to understand the subsequent impact of those on downstream analyses. This is especially important for cancer data analysis where varying tumor purity and intra-tumor heterogeneity make distinguishing low frequency somatic variants from sequencing noise challenging. Here, we have created a whole genome reference dataset of 3 matched tumor-normal cell lines sequenced deeply on both HiSeqX and NovaSeq, employed it to evaluate our somatic pipeline, and released it to the genomics community. The 3 cancer cell lines selected are common and represent the range of mutations profiles that a somatic pipeline is expected to identify correctly. COLO-829 was derived from a metastatic melanoma male patient and presents a pseudo-tetraploid karyotype[5,6]. It was the first cancer genome to be comprehensively characterized by whole-genome sequencing[7], has previously been characterized as hypermutated and was used to establish a reference for benchmarking somatic mutation pipelines[8]. HCC-1143 and HCC-1187 were isolated from patients with ductal carcinoma breast cancer[9]. HCC-1143 is near tetraploid[10] and heavily rearranged, and its matched normal cell line HCC-1143BL has a chromosome 2 amplification. HCC-1187 is hypotriploid[11]. We decided to share with the scientific community the data we generated and believe that it can be used as reference dataset, together with other similar dataset of real tumors[12,13] or cancer cell lines[8,14].

New York Genome Center, New York, NY, 10013, USA. *email: nrobine@nygenome.org

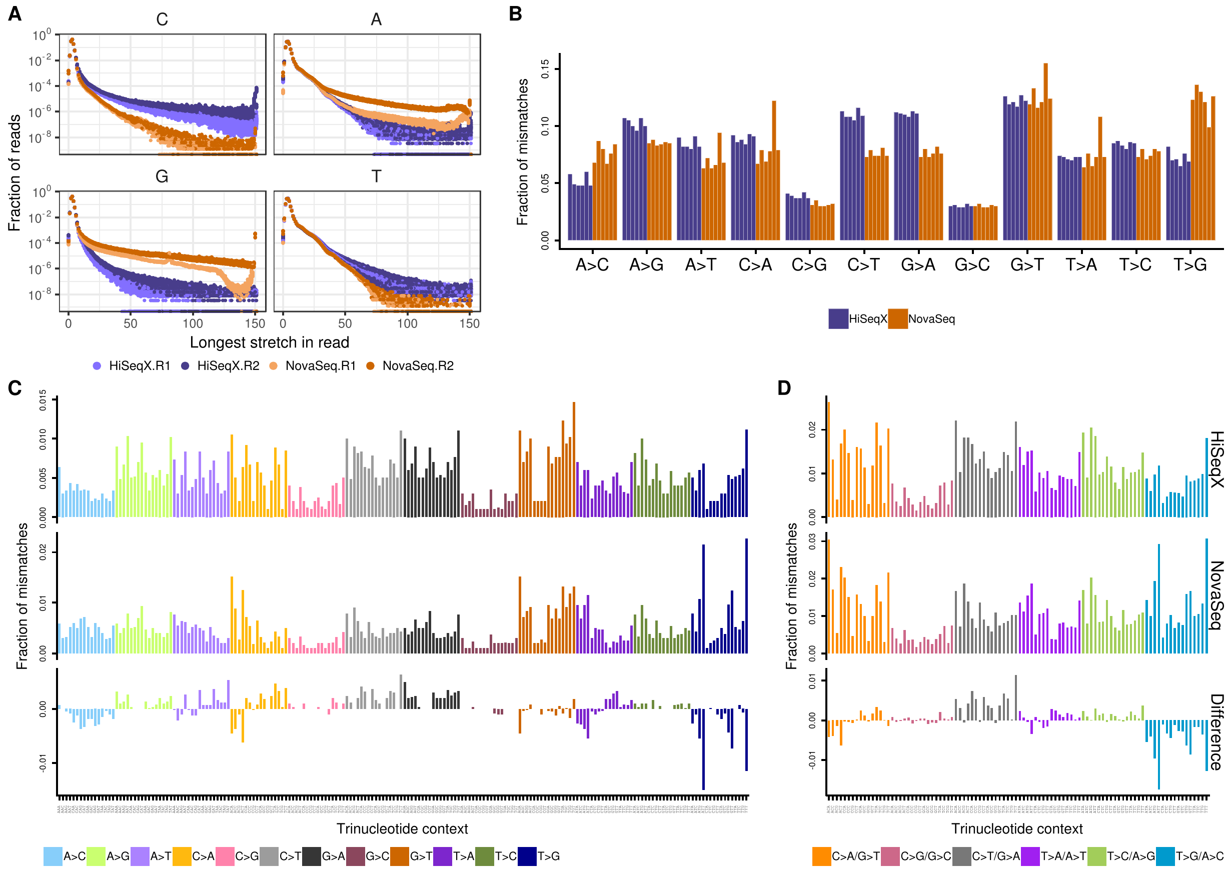

Figure 1. Homopolymer length and base mismatch comparisons between HiSeqX and NovaSeq. (**A**) Distribution of length of longest stretches of a nucleotide in HiSeqX and NovaSeq, Read 1 and Read 2 FASTQ files. Each dot represents fraction of reads in a single FASTQ file. Fraction of the total number of reads is represented in log-scale. (**B**) Single nucleotide mismatches by type in samples sequenced on NovaSeq and HiSeqX, with mapping quality (MQ) ≥10 and base quality (BQ) ≥10 cut-offs. Each bar represents a single sample and is colored based on sequencing platform. (**C**) Average mismatch rates for bases with MQ≥10 and BQ≥10 across the 6 cell line samples for each mismatch type per trinucleotide for HiSeqX (top row), NovaSeq (middle row) and difference between HiSeqX and NovaSeq (bottom row). (D) Same as (**C**), but with mismatch types categories collapsed with their respective reverse complements.

Results

Read level comparison. We sequenced between 2 and 6.3 billion reads for the 3 tumor cell lines and between 1 and 4 billion reads for the normal cell lines (Supplemental Table 2). Before applying our alignment and variant calling pipeline to the samples, we observed a few noticeable differences between the sequencing platforms. As expected, the quality score profiles along the reads differ, reflecting differences in the base calling and quality score estimation between the instruments. However, GATK's Base Quality Score Recalibration was effective in minimizing these differences (Supplemental Fig. S2). Still, the drop in quality score at the end of Read 2 on HiSeqX was more pronounced than on the NovaSeq. We noticed a slightly larger number of long homopolymers on the NovaSeq instruments (Supplemental Fig. S3). We computed for each read the longest stretch of each possible base and summarized the results in Fig. 1A. In both Read 1 and Read 2, NovaSeq instruments produced more stretches of Gs than HiSeqX, which we attributed to an artifact resulting from the fact that G is detected as the absence of signal in the 2-color chemistry of the NovaSeq platform. Although less pronounced, we also detected this effect for stretches of As (detected by the joint signal of both colors), especially in Read 2 and the inverse effect for C and T bases (each detected by the red and green signal respectively). We believe that some of the patterns, illustrated in Fig. 1A (such as the bump of stretches of 81 Gs in Read 1) are also due to differences in the base calling algorithms.

Alignment-level comparison. The mean coverage ranged from 80X to 278X for the tumor cell lines, and from 42X to 180X for the normal cell lines. The alignment rate was very comparable between the two platforms and always superior to 99.5%. The percentage of reads marked as duplicates was higher on HiSeqX (mean 11.25%) than NovaSeq (mean 6.6%), despite the deeper coverage of the NovaSeq samples. This was unexpected, because we usually see slightly higher duplication rates on NovaSeq as compared to HiSeqX. We attributed this observation to differences in loading concentration between the HiSeqX and the NovaSeq flow cells as this generally correlates strongly with the observed duplication rate for PCR-free libraries.

We observed differences between the sequencing platform in the single nucleotide mismatch profiles, where samples sequenced on NovaSeq contained more C > A and T > A mismatches compared to samples sequenced on HiSeqX which had more A > G and T > C (Supplemental Fig. S6). HiSeqX data had an average mismatch rate of 0.75% compared to 0.6% in NovaSeq data. Filtering out low mapping quality reads and low quality bases reduced most of the sequencing platform based differences, leaving higher T > G and lower C > T and G > A mismatches in NovaSeq samples (Fig. 1B). The overall mismatch rates in the two platforms after quality filtering were very similar, 0.24% in NovaSeq and 0.23% in HiSeqX.

The mismatches were further split based on their trinucleotide context. Figure 1 shows the fraction of mismatches averaged across all samples. We noticed that T > G mismatches in NovaSeq samples were predominantly found in the trinucleotide contexts of A[T > G]T, G[T > G]T and T[T > G]T. In order to compare with the somatic mutational profile, all mismatches were collapsed to the 6 mismatch types (C > A, C > G, C > T, T > A, T > C, T > G) (Fig. 1D). Even after collapsing the mismatch types, we saw more T > G/A > C mismatches in the NovaSeq data. The difference in mismatches between HiSeqX and NovaSeq, for all samples is shown in Supplemental Fig. S7, and for the collapsed mismatch types in Supplemental Fig. S8.

Variant-level comparison. In order to compare the variants called in HiSeqX and NovaSeq data for the same tumor-normal pairs, we first downsampled all the tumor samples to 80X coverage and the normal samples to 40X coverage and ran our variant calling pipeline. For all PASS-filtered somatic variants called by the pipeline, the concordance between the platforms ranged from 81 to 92% for SNVs, 45 to 58% for indels, and 50 to 77% for SVs (Fig. 2). When comparing the variants in our high confidence callset (obtained by intersecting the results of multiple callers, see Methods), the concordance was much higher: 90 to 94% for SNVs, 87 to 94% for indels and 81 to 88% for SVs. This clearly illustrates the advantages of using multiple callers and evidence from orthogonal strategies to reduce false positive calls, particularly for indel calling. The concordance of variants called by the individual caller is shown in Supplemental Fig. S4. We also compared two NovaSeq replicates of COLO-829 and COLO-829BL sequenced at 80X and 40X respectively, on distinct lanes of the same sequencing run, and found that the intra-run variability is comparable to the inter-platform variability (93% for SNVs, 91% for indels and 83% for SVs).

Focusing on the discordant calls, we observed that most of the somatic SNVs identified by one platform but not the other are observed at low allele frequency (<10%), indicating that they might have been missed due to insufficient coverage and sampling differences, or be false positives introduced by sequencing artifacts (Fig. 2B,D). We also observed that the mutation spectrum for discordant calls was very different from the concordant calls Fig. 2 (Supplemental Fig. S5), with a relatively large number of T > G mutations among the variants unique to the NovaSeq instrument (Fig. 2). This is in agreement with the higher T > G mismatches, especially in A[T > G]T, G[T > G]T and T[T > G]T context, seen in NovaSeq data (Fig. 1; Supplementary Fig. S5). We did not see the same trend when comparing high confidence variants (Supplemental Fig. S10). We therefore think that a lot of these T > G variants called only in NovaSeq data may in fact be artifactual calls that could be filtered out if we include more NovaSeq samples in our panel of normals. The allele frequency and mutational spectrum of discordant SNVs between HiSeqX and NovaSeq without Panel of Normal filtering is shown in Supplemental Fig. S9. Our PON predominantly consisted of HiSeqX samples and may therefore be better at removing HiSeqX-specific artifacts. For example, we saw that a lot of T[C > A]A SNV calls unique to HiSeqX (Supplemental Fig. S10) were filtered because of the panel of normal filtering step, and therefore were not seen in our final callsets.

Comparison to a reference callset and an alternative somatic pipeline. The three cell lines we sequenced have already been extensively studied and sequenced by other groups. In particular, Craig *et al.* sequenced different passages of COLO-829 in three different centers (TGen, BCSGSC and Illumina) and established a somatic reference dataset for SNV and indels from the consensus of their pipelines. This dataset also provided copy number gain/loss information for 6,586 genes. In Pan-Cancer Analysis of Whole Genomes (PCAWG) project, three best-practice pipelines were developed from the participating institutions and made accessible in Docker containers[15]. Of the three pipelines, we ran the pipeline from Sanger Institute on COLO-829 samples sequenced on HiSeqX and NovaSeq at NYGC. The other two pipelines from DKFZ and Broad Institute were not readily implementable, due to assumed dependencies in those workflows. The PASS variants from the PCAWG Sanger pipeline were compared against the Craig *et al.* reference dataset.

Overall, we saw that our pipeline called fewer SNVs than the Sanger pipeline, and yet we called over 98% of the Craig *et al.* SNVs compared to around 96.4% called by the Sanger pipeline (Fig. 3). Our pipeline called more indels compared to the Sanger pipeline, but we had far fewer indel calls in the high confidence callset. We called around 86% and 84% of the Craig *et al.* indels in our All Somatic and High Confidence callsets, compared to 80% called by the Sanger pipeline, suggesting that our pipeline may be more sensitive.

We explored the sources of discrepancies between our callset and the reference dataset established in Craig *et al.* and represented the different categories in Supplemental Fig. S11. Almost half of the variants absent from our final callset were in fact called as PASS-filtered by at least one of the callers, but were removed from our AllSomatic list due to PON filtering. Some of the discordant variants were in the raw callsets of one or more variant callers, but did not pass the caller-specific filtering. Other sources of discrepancies were evidence of the variant allele in the normal sample, low coverage or low variant allele frequency in the tumor sample. While there were some differences between SNV and indel calls between the two pipelines, we found that the CNV recall was very similar between the two pipelines based on a gene-level comparison (99.8% recall for both our pipeline and the Sanger pipeline).

Overall, despite some discrepancies, we are confident about our callset and believe that the PON filtering is a powerful method to remove technical artifacts. It is also entirely possible that some somatic variants were different between the cells used in the reference dataset and the ones used in this work.

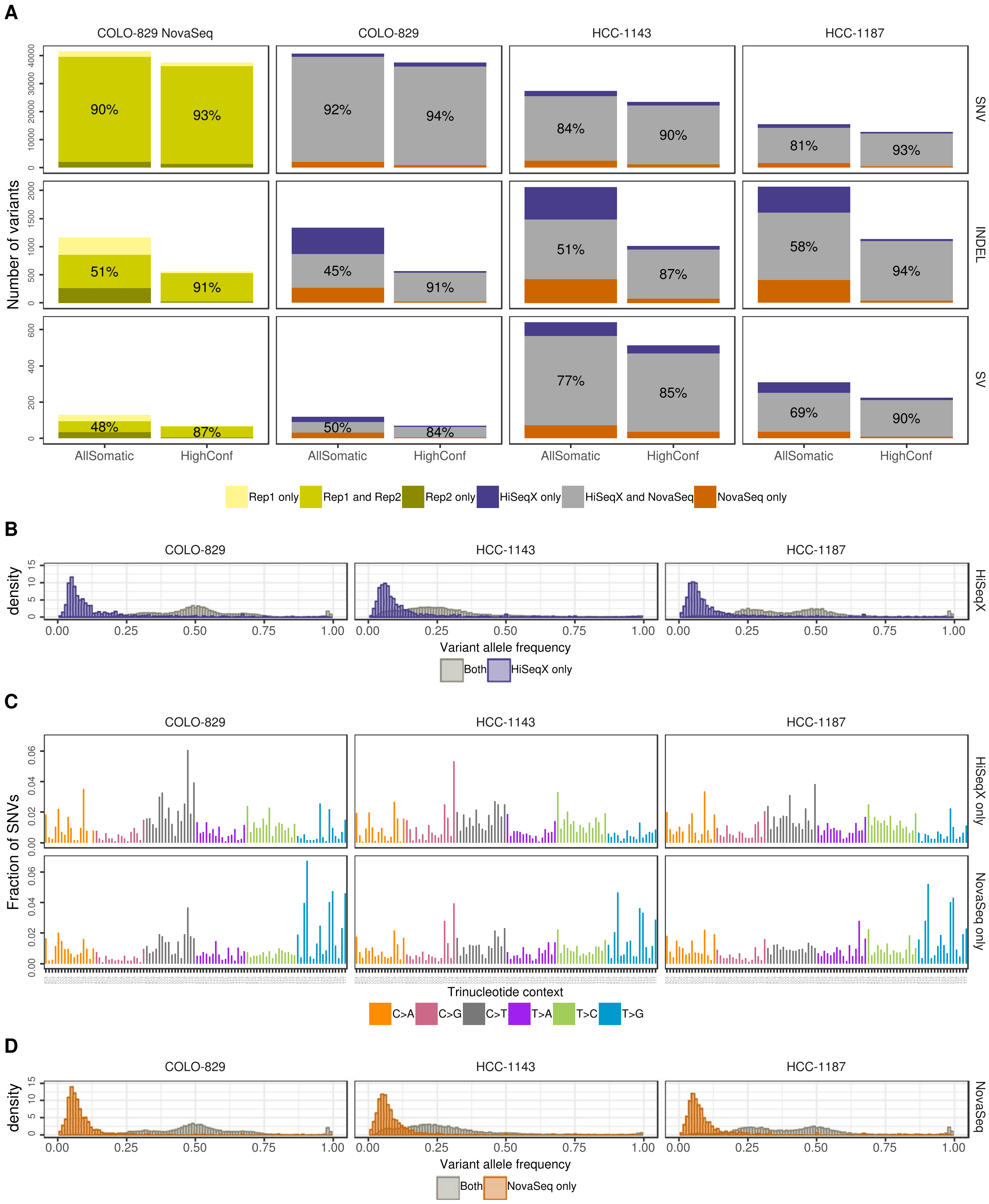

Figure 2. Intra- and inter-platform comparison of somatic variants. (**A**) Comparison of SNVs, Indels and structural variants between two replicates COLO-829 NovaSeq data (created using reads from mutually exclusive lanes) and between HiSeqX and NovaSeq data for the three cell lines. Orange bars (resp. purple) represent the number of variants called uniquely in the NovaSeq runs (resp. HiSeqX) and the grey bars correspond to the variants called in both samples. The numbers in the grey bars represent the concordance between the two samples, calculated as percentage of the (number of variants in the intersect)/(number of variants in the union). (**B**) Allele frequency of the variants called only in HiSeqX in purple, and for reference the allele frequency of variants called in both platforms in grey. (**C**) The decomposition in trinucleotide contexts of the SNVs called uniquely by each platform. Substitutions are represented by the pyrimidine of the mutated Watson-Crick base pair. (**D**) Similar to panel B but for variants uniquely called in NovaSeq. The AllSomatic callsets were used for panels B, C and D.

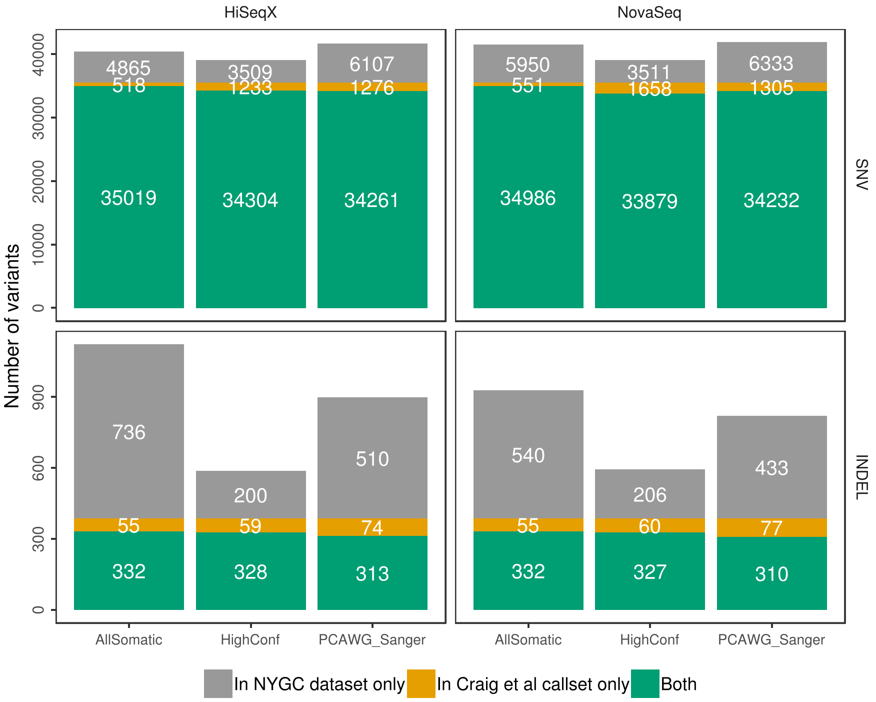

Figure 3. Comparison of somatic variants called on HiSeqX and NovaSeq COLO-829 tumor/normal data downsampled to 80X/40X to the Craig *et al.* reference dataset.

Performance at different coverages and purities. To assess the impact of our post-calling filtering steps, we ran our pipeline on 90X COLO-829BL (treated as tumor) against 40X COLO-829BL (treated as normal). Since the data is from the same normal cell line, any variant called on this pairing is a false positive. Amongst the individual callers, MuTect2 called the highest number of false positive SNVs and indels, whereas SvABA called the highest number of false positive SVs on this dataset. We saw a remarkable reduction in false positive calls due to the NYGC filtering steps, most of which were filtered using the Panel of Normal (Supplemental Fig. S18).

We wanted to evaluate the performance of our pipeline at different tumor and normal coverages. For this we paired tumor and normal samples for COLO-829 and HCC-1143 downsampled to different depths, and computed precision, recall and F1 scores for the different pairings by comparing to the high coverage data (see Methods). We found that the recall of SV callsets was not affected by the coverage of normal sample (Supplemental Fig. S15). Recall of SNV and Indel callsets was much lower for comparisons against 10X normal, however, it was very comparable for higher normal coverages. The precision of Indel and SV callsets was especially affected by the coverage of the normal sample.

Moreover, recall of the callsets improved, albeit slightly, as the coverage of tumor sample increased. Using the NYGC AllSomatic callset, we could achieve sensitivity of >92% for SNVs and Indels, and >82% for SVs with average coverages as low as 20X for the normal sample and 60X for the tumor sample.

Since the cell lines are very pure and mostly clonal, most of the variants are seen at high VAF and therefore easy to call even at low tumor coverages. We therefore additionally calculated recall for SNVs and Indels in different VAF bins (as computed from the high coverage data), and found a noticeable increase in recall with increase in tumor coverage for variants <=20% VAF (Supplemental Fig. S16).

One frequent concern in cancer genomics is that tumor samples are always heterogeneous, composed of tumor cells, stromal contamination and normal cells. Since one goal of a somatic pipeline is to establish the catalog of the somatic mutations occurring in the tumor cells, it is important to take into consideration the composition of the sample, usually summarized as the “tumor purity”. Using the two most deeply sequenced cell lines, we simulated lower purity tumor data and evaluated the performance of our SNV, indel and SV pipelines at different purities by comparing to the results obtained with high coverage (see Methods). We observed that precision remained good for high-confidence SNVs, indels and SVs even at low purity, but that recall decreased rapidly with purity lower than 50% for the highly rearranged HCC-1143, and below 25% for COLO-829, which had comparatively fewer chromosomal abnormalities (Fig. 4A, Supplemental Fig. S13). Not surprisingly, we found that the AllSomatic callset was more sensitive than any single caller at different simulated purities (Supplemental Fig. S17). We conducted a similar type of precision/recall analysis for CNVs (Fig. 4B, Supplemental Fig. S14), comparing calls across samples at the base-pair level and found that for amplifications, the recall tended to decrease more gradually as purity decreased but for deletions there was a sharp drop-off below 50% purity. This is mainly because deletions occupy few copy number states (e.g. 0 or 1 copy for a diploid genome), whereas amplifications can have higher copy number states that will still be captured at lower purity, as is the case with HCC-1143. Precision remained relatively high at different purities for both amplifications and deletions; however, for deletions it dropped off at around 25%, as very few calls were being made at this level of purity.

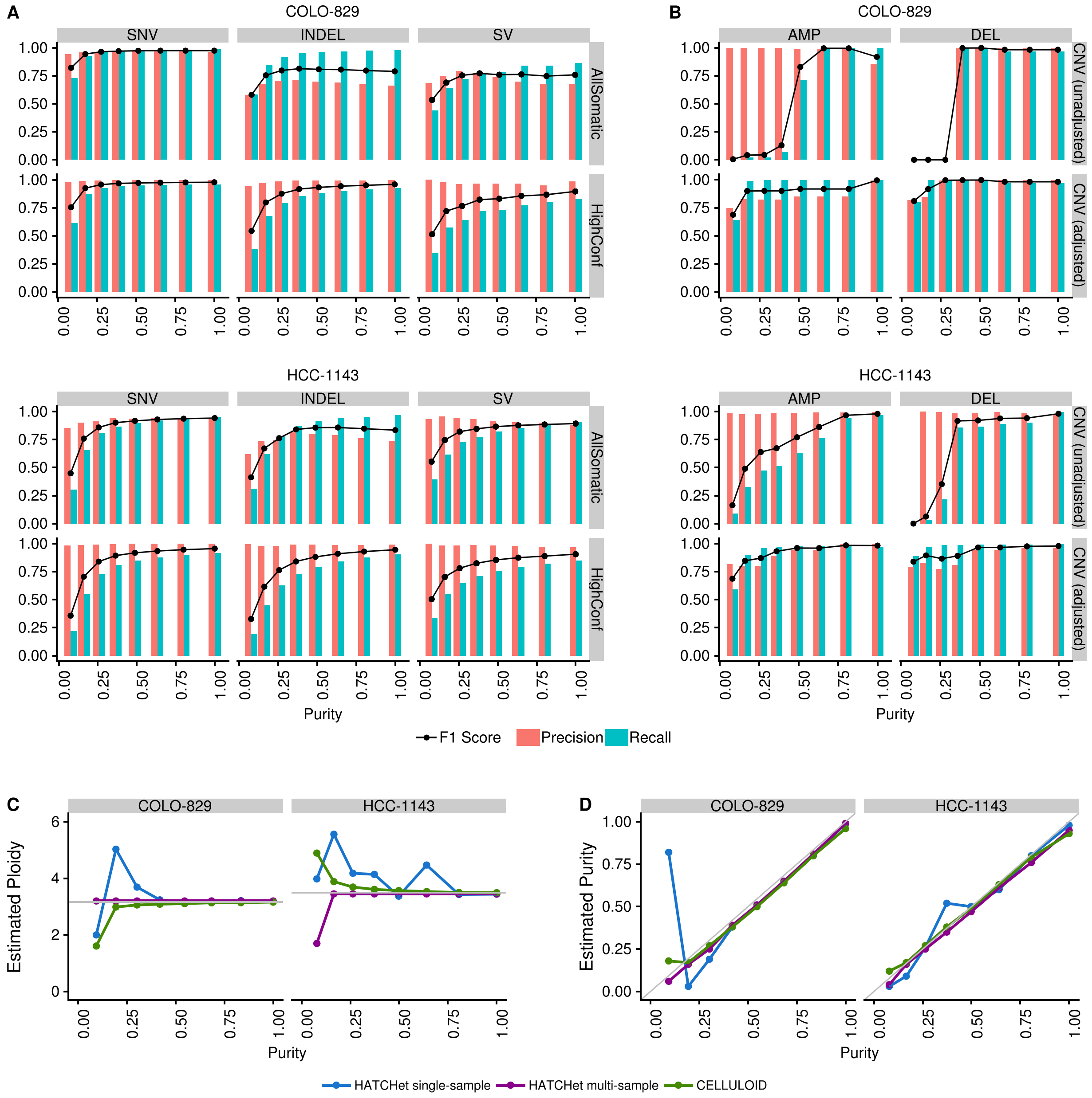

Figure 4. Precision, recall and F1 scores at different simulated purities for (**A**) SNVs (left),Indels (center) and SVs (right), and (**B**) CNVs without (left) and with (right) adjustments of log2 values for purity and ploidy. (**C**) Ploidy and (**D**) purity estimation for the purity ladder samples using CELLULOID and HATCHet in single-sample and multi-sample mode.

For low purity samples, some of the false negative CNV calls could be attributed to the chosen thresholds for categorizing amplifications and deletions, and the fact that BIC-seq2 outputs log2 ratios that are not adjusted for purity and ploidy. We wanted to investigate how purity/ploidy adjustment could rescue true events missed in low purity samples. For this, we based our purity estimates off of the fraction of reads we mixed from the tumor and normal sample to create these low purity samples, which should represent the true purity (see Methods). Using these estimates, the CNV log2 values were adjusted. We found that the recall was much higher than the original unadjusted data, but precision gradually decreased as the purity dropped. At the lowest purity level, precision and recall dropped, which was likely a result of the different segmentation at this purity level. The ability to capture CNV calls in lower purity samples by adjusting the log2 values based on purity/ploidy is very useful, but requires correct estimation of purity and ploidy for the tumor sample.

Therefore, we evaluated CELLULOID[16] and HATCHet[17] for their ability to correctly estimate the purity and ploidy values for our purity ladder samples (Fig. 4). Both tools use read depth information at germline heterozygous sites to infer the tumor purity/ploidy. In particular, HATCHet can be run in multisample mode, which can leverage information from high purity samples to infer the purity/ploidy of low purity samples from the same individual. We ran HATCHet in both single sample and multisample mode. We found that both tools tend to perform well above 50% purity. Furthermore, HATCHet (in multisample mode) and CELLULOID can both give close estimates of the purity/ploidy for much lower purity samples. CELLULOID estimates only seemed to drop-off for samples with a purity lower than 12.5%, while HATCHet in multisample mode produced a good estimate at this low purity level.

Further, we looked into the log2 values of the events that were captured at lower purities and their adjusted log2 values based on CELLULOID and HATCHet single sample estimates of purity and ploidy (Supplemental Fig. S12). In the low purity samples, we were able to identify many of the events that were originally lost at this purity level. However, it was also apparent that the CNVs being captured at the lowest purity level did not necessarily resemble the high purity CNVs. This is another example of how differences in segmentation in lower purity samples can have an impact on recapturing CNV calls.

Discussion

Cancer cell lines are useful models for studying cancer biology. They are widely available, easy to propagate and composed of a relatively homogeneous population of cells, making them extremely valuable for advancement of tools and methods for cancer genomics. However, they are imperfect models and do not represent the entire complexity of real tumor samples. They may also contain unique genomic features needed for immortalization and *in vitro* growth. Here, we used 3 cancer cell lines for benchmarking purposes and share our high-quality callset with the genomics community. We plan to keep using this data to test novel variant callers and may resequence these cell lines with novel sequencing technologies (such as long read technologies). Here, we demonstrated using these cell lines the existence of systematic differences between the reads produced by HiSeqX and by NovaSeq. The patterns we identified will need to be taken into account to fully exploit the signal produced by the sequencers. For instance, we believe that a deep learning model designed to filter out sequencing artifacts and detect real mutations at very low frequency (such as would be needed for early detection of cancer in liquid biopsy samples) would need to be trained independently for HiSeqX and for NovaSeq, depending on the instrument used for the real-life application of the model. We designed a pipeline for somatic variant calling, composed of multiple softwares for SNV, indel and structural variants. We showed the advantage of using multiple tools to obtain high confidence calls. We showed that with the standard coverage of tumor-normal whole genome sequencing (80X/40X) and our somatic pipeline, the pattern of homopolymer frequency does not translate into systematic biases once multiple somatic callers are applied. We noted a mild enrichment of T > G mutations in the variants called uniquely in NovaSeq and not in HiSeqX data. However, that was not the case when we compared our high confidence variants (those that are supported by multiple callers). Overall, this gives us the confidence to upgrade our sequencing platform to NovaSeq, without any loss of quality (and with a substantial gain in the cost of sequencing and a higher throughput). We demonstrated the importance of filtering recurrent artifacts with a Panel of Normals, ideally composed of a large number of samples from the platforms used to sequence the samples of interest and preferably using the same sequencing protocols. We expect to increase the number of normal samples included in our PON, especially from NovaSeq, as we keep sequencing properly consented samples. We plan to explore refined strategies to filter artifacts based on the allele detected in normal samples rather than, as currently, based on the location in the reference genome. Finally, we used the deeply sequenced libraries to test tools designed to estimate purity and ploidy of tumor samples and showed the importance of incorporating these estimates to improve copy-number detection.

Conclusion

We present high-quality, deeply sequenced whole-genome data for 3 common cancer lines. We used these samples to study in detail the differences between the two most recent high-throughput sequencers from Illumina, HiSeq X Ten and NovaSeq 6000. We ran these tumor-normal pairs through our somatic pipeline and demonstrated that the inter-platform variability was very similar to the intra-run variability, indicating that the systematic differences between the platforms at the read level are well-handled by the base calling algorithm and by our somatic pipeline. We demonstrated the advantages of combining multiple algorithms to detect SNV, indels and structural variants. We used the samples to study in details the effect of tumor purity on performance and tested tools aiming at estimating purity from WGS data. We show how to use these estimations to recalibrate copy-number events and re-categorize amplifications and deletions.

Material and Methods

Cell culture and DNA isolation.

The cancer cell lines (COLO-829 ATCC CRL-1974, COLO-829BL ATCC CRL-1980, HCC-1143 ATCC CRL-2321, HCC-1143 BL ATCC CRL-2362, HCC-1187 ATCC CRL-2322 and HCC-1187BL ATCC CRL-2323) were obtained from ATCC[18]. The cell lines were cultured using the recommendations from ATCC. Cultured cells were split into two aliquots for metaphase chromosome preparation and karyotype analysis. Representative images and karyotypes are reported in Supplemental Fig. S1. The number of passages is indicated in Supplemental Table 1.

Library preparation and sequencing.

Libraries were prepared using the TruSeq DNA PCR-free Library Preparation Kit (Illumina) with 1 μg DNA input following Illumina's recommended protocol[19], with minor modifications as described. Intact genomic DNA was concentration normalized and sheared using the Covaris LE220 sonicator to a target size of 450 bp. After cleanup and end-repair, an additional double-sided bead-based size selection was added to produce sequencing libraries with highly consistent insert sizes. This was followed by A-tailing, ligation of Illumina DNA Adapter Plate (DAP) adapters and two post-ligation bead-based library cleanups. These stringent cleanups resulted in a narrow library size distribution and the removal of remaining unligated adapters. Final libraries were run on the Fragment Analyzer to assess their size distribution and quantified by qPCR with adapter specific primers (Kapa Biosystems). The libraries were pooled together based on expected

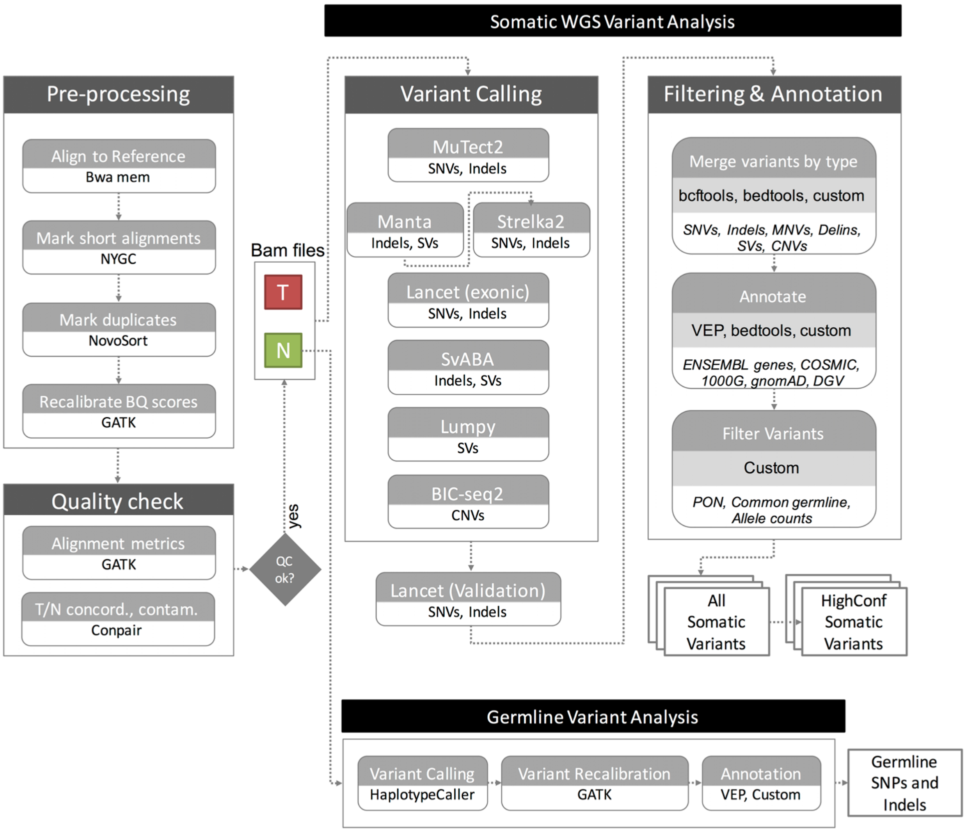

Figure 5. NYGC Somatic Pipeline for tumor-normal whole-genome sequencing samples.

final coverage and sequenced across multiple flow cell lanes to reduce impact of lane-to-lane variations in yield. Whole genome sequencing was performed on the NovaSeq (NSCS 1.3.1; RTA v3.3.3) and the HiSeqX (HCS HD 3.5.0.7; RTA v2.7.7) at 2 × 150 bp read length, using v1 S2/S4 300-cycle and SBS v3 reagents, respectively.

Pre-processing. The sequencing data for all the cell lines was demultiplexed using bcl2fastq (Illumina) v2.20.0.422. FASTQ files were then processed through NYGC's high-performance computing cluster using the NYGC automated pipeline (Fig. 5). Sequencing reads were aligned to the GRCh38 reference genome (1000 Genomes version) using BWA-MEM (v0.7.15)[20]. NYGC's ShortAlignmentMarking (v2.1)[21] was used to mark short reads as unaligned. This tool is intended to remove spurious alignments resulting from contamination (e.g. saliva sample bacterial content) or from too aggressive alignments of short reads the size of BWA-MEM's 19 bp minimum seed length. These spurious alignments result in pileups in certain locations of the genome and can lead to erroneous variant calling.

GATK (v4.1.0)[22] FixMateInformation was run to verify and fix mate-pair information, followed by Novosort (v1.03.01) markDuplicates to merge individual lane BAM files into a single BAM file per sample. Duplicates were then sorted and marked, and GATK's base quality score recalibration (BQSR) was performed. The final result of the pre-processing pipeline was a coordinate sorted BAM file for each sample.

Once preprocessing was complete, we computed a number of alignment quality metrics such as average coverage, %mapped reads and %duplicate reads using GATK (v4.1.0) and an autocorrelation metric (adapted for WGS from[23]) to check for unevenness of coverage. We also ran Conpair[24], a tool developed at NYGC to check the genetic concordance between the normal and the tumor sample and to estimate any inter-individual contamination in the samples.

Somatic variant calling and annotation. The tumor and normal bam files were processed through NYGC's variant calling pipeline which consists of MuTect2 (GATK v4.0.5.1)[25], Strelka2 (v2.9.3)[26] and Lancet (v1.0.7)[27] for calling Single Nucleotide Variants (SNVs) and short Insertion-or-Deletion (Indels), SvABA (v0.2.1)[28] for calling Indels and Structural variants (SVs), Manta (v1.4.0)[29] and Lumpy (v0.2.13)[30] for calling SVs and BIC-Seq2 (v0.2.6)[31] for calling Copy-number variants (CNVs). Manta also outputs a candidate set of Indels which is provided as input to Strelka2 (following the developers recommendation, as it improves Strelka2's sensitivity for calling indels >20nt). Due to its computing requirements, in this pipeline Lancet is only run on the exonic part of the genome. It is also run on the +/− 250nt regions around non-exonic variants that are called by only one of the other callers, to add confidence to such variants. Small SVs called by Manta are also used to add confidence to the indel calls.

Next, the calls were merged by variant type (SNVs, Multi Nucleotide Variants (MNVs), Indels and SVs). MuTect2 and Lancet call MNVs, however Strelka2 does not and it also does not provide any phasing information. So to merge such variants across callers, we first split the MNVs called by MuTect2 and Lancet to SNVs, and then merged the SNV callsets across the different callers. If the caller support for each SNV in a MNV is the same, we merged them back to MNVs. Otherwise those were represented as individual SNVs in the final callset. The SVs were converted to bedpe format, all SVs below 500 bp were excluded and the rest were merged across callers using bedtools pairtopair (slop of 300 bp, same strand orientation, and 50% reciprocal overlap). For CNVs, segments with $\log2 > 0.2$ were categorized as amplifications, and segments with $\log2 < -0.235$ were categorized as deletions (corresponding to a single copy change at 30% purity in a diploid genome, or a 15% Variant Allele Fraction). The resulting variants were annotated with Ensembl as well as databases such as COSMIC (v86)[32], 1000Genomes (Phase3)[33], gnomAD (r2.0.1)[34], dbSNP (v150)[35] and Database of Genomic Variants (DGV)[36] using Variant Effect Predictor (v93.2)[37] for SNVs and Indels, and bedtools[38] for SVs and CNVs.

Somatic variant filtering.

Panel of normals. The Panel Of Normals (PON) filtering removes recurrent technical artifacts from the somatic variant callset[25]. The Panel of Normals for SNVs, indels and SVs was created with whole-genome sequencing data from normal samples from 242 unrelated individuals. Of these, sequencing data for 148 individuals was obtained from the Illumina Polaris project[39] which was sequenced on the HiSeqX platform with PCR-free sample preparation. The remaining samples were sequenced by the NYGC. Of these, 73 individuals were sequenced on HiSeqX, 11 on NovaSeq, and 10 were sequenced on both.

We ran MuTect2 in artifact detection mode and Lumpy in single sample mode on these samples. For SNVs and indels, we created a PON list file with sites that were seen in two or more individuals and we used this list to filter the somatic variants in the merged SNV and indel files.

For SVs, we used SURVIVOR (v1.0.3)[40] to merge Lumpy calls. Variants were merged if they were of the same type, had the same strand orientation, and were within 300 bp of each other (maximum distance). We did not specify a minimum size. After merging SVs, we used these calls as a PON list. To filter our somatic SV callset, we merged our PON list with our callset using bedtools pairtopair (slop of 300 bp, same strand orientation, and 50% reciprocal overlap), and filtered those SVs found in two or more individuals in our PON.

Common germline variants. In addition to the PON filtering, we removed SNVs and Indels that had minor allele frequency (MAF) of 1% or higher in either 1000Genomes (phase 3) or gnomAD (r2.0.1)[34], and SVs that overlapped DGV and 1000Genomes (phase3). CNVs were annotated with DGV and 1000 Genomes but not filtered.

Allele counts. Since our variant callsets were generated by merging calls across callers, and each of them reported different allele counts, we reported final chosen allele counts for SNVs and indels. For SNVs, and for indels less than 10nt in length, these were computed as the number of unique read-pairs supporting each allele using the pileup method, with minimum mapping quality and base quality thresholds of 10 each.

For larger indels and complex events, we chose the final allele counts reported by the individual callers Strelka2, MuTect2, Lancet, in that order. For indels larger than 10nt that were only called by SvABA, we did not report final allele counts and allele frequencies because SvABA does not report the reference allele count, making it difficult to estimate the variant allele frequency.

We then used these final chosen allele counts and frequencies to filter the somatic callset. Specifically, we filtered any variant for which the variant allele frequency (VAF) in the tumor sample was less than 0.0001, or if the VAF in the normal sample was greater than 0.2, or if the depth at the position was less than 2 in either the tumor sample or the normal sample. We also filtered variants for which the VAF in normal sample was greater than the VAF in tumor sample. Variants that passed all of the above-mentioned filters were included in our final somatic callset (hereby referred to as AllSomatic).

High-confidence variants. For SNVs, indels and SVs, we also annotated a subset of the somatic callset as high confidence.

For SNVs and indels, high confidence calls were defined as those that were either called by two or more variant callers, or called by one caller and also seen in the Lancet validation calls or in the Manta SV calls.

For structural variants, high confidence calls were taken from the somatic callset if they met the following criteria: a SV was called by 2 or more variant callers, or called by Manta or Lumpy with either additional support from nearby CNV changepoint or split-read support from SplazerS[41], an independent tool used to calculate the number of split-reads supporting SV breakpoints. An SV was considered supported by SplazerS if it found at least 3 split-reads in the tumor only. Nearby CNV changepoints were determined by overlapping BIC-Seq2 calls with the SV callset using bedtools closest. An SV was considered to be supported by a CNV changepoint if the breakpoint of the CNV was within 1000 bp of an SV breakpoint.

Germline variant analysis.

We also called germline SNPs and indels using GATK HaplotypeCaller (v3.5), which generated a single-sample GVCF. We then ran GATK's GenotypeGVCF to perform single sample genotype refinement and output a VCF, followed by variant quality score recalibration (VQSR) for variant filtering (at tranche 99.6%). Next, we ran Variant Effect Predictor (v93.2) to annotate the variants with Ensembl as well as databases such as COSMIC (v86), 1000Genomes (Phase3), gnomAD (r2.0.1), dbSNP (v150), ClinVar (201805), Polyphen2 (v2.2.2) and SIFT (v5.2.2).

Mismatch analysis.

The whole genome data from both platforms was downsampled to 8X coverage for all samples, and the number of single nucleotide mismatches to the reference in each sample was computed in order to evaluate the technical error profiles for each sequencing platform. Duplicate, unmapped, supplementary and vendor/platform QC-failed reads were excluded from the calculations. Mismatches were represented with respect to the read strand. For downstream analyses, mapping quality $\geq$10 and base quality $\geq$10 cut-offs were applied. The mismatch types were classified by the trinucleotide context in which they occur in the read strand. Mismatches were also summarized for the 6 reduced categories: C > A, C > G, C > T, T > A, T > C, T > G, by reverse complementing the other categories.

PCAWG sanger pipeline. The CWL workflow definition file and the pre-compiled resource files were downloaded according to the instructions[42]. The pipeline was run using cwltool (v1.0.52) and only supports the GRCh37 reference.

Comparison to COLO-829 reference dataset. We downloaded the SNVs and Indels VCF file from EGA (Data accession ID: EGAD00001002142). This reference dataset was only provided for GRCh37. We therefore ran our pipeline and the PCAWG Sanger pipeline on the 80X/40X COLO-829/COLO-829BL HiSeqX and NovaSeq data aligned to GRCh37 with decoys reference (from GATK bundle for b37) and compared the somatic SNVs and Indels called on our data to the PASS-filtered variants in the Craig *et al*. VCF file. MNVs called by our pipeline were converted to SNVs for this comparison. CNV gene-level information was taken from Supplementary Table 2 of Craig *et al*. and compared to the CNV calls from our pipeline and the PCAWG Sanger pipeline after converting the data to gene-level results.

Data downsampling and purity ladder generation. Since the samples were sequenced to different depths between the two platforms, for the HiSeqX vs NovaSeq comparisons, we downsampled the tumor samples to 80X and the normal samples to 40X for all inter-platform variant comparisons. For this we used samtools view to randomly subsample the reads from the high coverage data.

For within platform, intra-run comparisons, we split the high coverage NovaSeq data for COLO-829, COLO-829BL, HCC-1143 and HCC-1143BL by readgroups using samtools split. Each readgroup corresponded to a different lane on the sequencer for that sample. Different sets of readgroups were used to create two replicates (Rep1 and Rep2) for each sample, thereby ensuring that the two replicate samples consisted of mutually exclusive set of reads. The tumor samples were then downsampled to 80X and normal samples to 40X.

We also used two replicates of normal cell line COLO-829BL, one with mean coverage of 90X and another of mean coverage of 40X, to assess false positive calls. For this we ran our pipeline treating the 90X coverage sample as tumor and the 40X sample as normal.

Additionally, we created what we refer to as "coverage ladder" samples for NovaSeq data of COLO-829 and HCC-1143 by downsampling the tumor samples to 10X, 20X, 30X, 40X, 60X, 80X and 90X and the matched normal samples to 10X, 20X, 30X and 40X. We ran our somatic pipeline on the different tumor-normal pairings from this ladder.

To simulate low tumor purity samples, we downsampled the Rep1 NovaSeq data for the tumors, COLO-829 and HCC-1143, to 10X, 20X, 30X, 40X, 50X, 60X and 70X, and mixed that with data from their matched normal samples at 70X, 60X, 50X, 40X, 30X, 20X and 10X respectively. We refer to these as the "purity ladder" samples. These mixed-in datasets were then analyzed against the 40X Rep1 matched normal sample data. Again, we used different readgroups for the mix-in than those that went into the Rep1 normal sample data.

To estimate the purity of these mixed-in samples, we had to take into consideration the average ploidy of the tumor. We used the ploidy that was estimated by CELLULOID[16] for the 100% purity samples, since they seemed to be very accurate upon manual review.

The tumor purity of the mixed-in samples was estimated using the following equation:

$$T_{purity} = \frac{T_{frac} \ * \ N_{ploidy}}{(1 - T_{frac}) \ * \ T_{ploidy} + T_{frac} \ * \ N_{ploidy}}$$

Where,

N_{ploidy} = Average ploidy of the normal sample, which we assumed to be 2.

T_{ploidy} = Average ploidy of the tumor sample, for which we used CELLULOID's estimate of ploidy on the 100% purity data.

T_{frac} = Fraction of the reads in the mixed-in sample that came from the tumor. (For a 10X tumor +70X normal mix-in sample, this was $10/(10 + 70) = 0.125$).

Recall and precision calculation for coverage ladder and purity ladder samples. SNVs and Indels called on the coverage and purity ladder samples were compared to the variants called in the high coverage data. True positives (TP) were considered to be variants that were also seen in the high confidence callset of the high coverage data, whereas false positives (FP) were considered to be variants that were called in the ladder sample but not seen in the AllSomatic callset of the high coverage data. Those variants that were called in the high confidence callset of the high coverage but not called in the ladder sample callset were classified as False Negatives (FN). Variants that were in the AllSomatic callset of the high coverage data but not in the HighConfidence callset were ignored for this analysis because they could not be confidently assigned as true variants.

For CNVs, events called in low purity samples were compared at the base level to the CNVs called in the high coverage 100% purity data. If a deletion or amplification was found in the low purity cell line, but not in the high coverage 100% purity cell line, this was classified as a FP. If a deletion/amplification was not found in the low purity cell line, but called in the high coverage 100% purity cell line, this was classified as a FN. TP were considered to be any deletions or amplifications that were found at the same position in both the low purity cell line and high coverage 100% purity cell line.

Precision, recall and F1 scores were calculated as:

$$\begin{aligned} Precision &= \frac{TP}{(TP + FP)} \\ Recall &= \frac{TP}{(TP + FN)} \\ F1\ score &= 2 * \frac{Precision\ *\ Recall}{Precision + Recall} \end{aligned}$$

Purity-ploidy adjustment of CNV log2 values:

$$\begin{aligned} T_{CN} &= \frac{T_{ploidy}\ *\ N_{CN}\ *\ (2^{Obs_{log2}} - (1 - T_{purity}))}{N_{ploidy}\ *\ T_{purity}} \\ Adj_{log2} &= log_2\left(\frac{T_{CN}}{T_{ploidy}}\ *\ \frac{N_{ploidy}}{N_{CN}}\right) \end{aligned}$$

Where,
T_{CN} = Absolute copy number for that segment in the tumor sample
N_{CN} = Absolute copy number for that segment in the normal sample (2 for autosomes, 1 for sex chromosomes in male, 2 for X chromosome in female)
N_{ploidy} = Average ploidy of the normal sample (which we assumed to be 2)
T_{ploidy} = Average ploidy of the tumor sample
T_{purity} = Tumor purity
Obs_{log2} = Observed log2 value for the segment (from BIC-seq2)
Adj_{log2} = Adjusted log2 value.

Purity-ploidy estimation using CELLULOID and HATCHet. For purity and ploidy estimation, we used CELLULOID (v0.11) and HATCHet[43]. CELLULOID was run in single-clone mode with default parameters and segment-based optimization. HATCHet was also run with default parameters.

Accession number. The raw genomic data are accessible on dbGAP phs001839.

Data availability

The raw genomic data are deposited on dbGAP (phs001839). The somatic variants for HighCoverage and downsampled 40X/80X are directly accessible on our website[44]. The website also contains the sample reports generated by the pipeline and a link to the Outpost QC interface.

Received: 4 June 2019; Accepted: 30 November 2019;
Published online: 13 December 2019

References

1. Simen, B. B. *et al*. Validation of a next-generation-sequencing cancer panel for use in the clinical laboratory. *Arch. Pathol. Lab. Med.* **139**, 508–517 (2015).
2. Cheng, D. T. *et al*. Memorial Sloan Kettering-Integrated Mutation Profiling of Actionable Cancer Targets (MSK-IMPACT): A Hybridization Capture-Based Next-Generation Sequencing Clinical Assay for Solid Tumor Molecular Oncology. *J. Mol. Diagn.* **17**, 251–264 (2015).
3. Cancer Genome Atlas Research Network. *et al*. The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* **45**, 1113–1120 (2013).
4. Campbell, P. J., Getz, G., Stuart, J. M., Korbel, J. O. & Stein, L. D. Pan-cancer analysis of whole genomes. *bioRxiv* 162784, https://doi.org/10.1101/162784 (2017).
5. Morse, H. G. & Moore, G. E. Cytogenetic homogeneity in eight independent sites in a case of malignant melanoma. *Cancer Genet. Cytogenet.* **69**, 108–112 (1993).
6. Bignell, G. R. *et al*. High-resolution analysis of DNA copy number using oligonucleotide microarrays. *Genome Res.* **14**, 287–295 (2004).
7. Pleasance, E. D. *et al*. A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature* **463**, 191–196 (2010).
8. Craig, D. W. *et al*. A somatic reference standard for cancer genome sequencing. *Scientific Reports* **6** (2016).
9. Gazdar, A. F. *et al*. Characterization of paired tumor and non-tumor cell lines established from patients with breast cancer. *Int. J. Cancer* **78**, 766–774 (1998).
10. Chen, W., Robertson, A. J., Ganesamoorthy, D. & Coin, L. J. M. sCNAphase: using haplotype resolved read depth to genotype somatic copy number alterations from low cellularity aneuploid tumors. *Nucleic Acids Res.* **45**, e34 (2017).
11. Newman, S. *et al*. The relative timing of mutations in a breast cancer genome. *PLoS One* **8**, e64991 (2013).
12. Alioto, T. S. *et al*. A comprehensive assessment of somatic mutation detection in cancer using whole-genome sequencing. *Nat. Commun.* **6**, 10001 (2015).
13. Griffith, M. *et al*. Optimizing cancer genome sequencing and analysis. *Cell Syst* **1**, 210–223 (2015).
14. Barretina, J. *et al*. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603–607 (2012).
15. Yung, C. K. *et al*. Large-Scale Uniform Analysis of Cancer Whole Genomes in Multiple Computing Environments. *bioRxiv* 161638, https://doi.org/10.1101/161638 (2017).
16. Notta, F. *et al*. A renewed model of pancreatic cancer evolution based on genomic rearrangement patterns. *Nature* **538**, 378–382 (2016).
17. Zaccaria, S. & Raphael, B. J. Accurate quantification of copy-number aberrations and whole-genome duplications in multi-sample tumor sequencing data. *bioRxiv* 496174, https://doi.org/10.1101/496174 (2018).

18. ATCC, https://www.atcc.org.
19. Illumina TruSeq DNA PCR-Free. https://support.illumina.com/content/dam/illumina-support/documents/documentation/chemistry_documentation/samplepreps_truseq/truseq-dna-pcr-free-workflow/truseq-dna-pcr-free-workflow-reference-1000000039279-00.pdf.
20. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv [q-bio.GN]* (2013).
21. *nygc-short-alignment-marking*. (Github), https://github.com/nygenome/nygc-short-alignment-marking.
22. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
23. Zhang, L. & Zhang, L. Use of autocorrelation scanning in DNA copy number analysis. *Bioinformatics* **29**, 2678–2682 (2013).
24. Bergmann, E. A., Chen, B.-J., Arora, K., Vacic, V. & Zody, M. C. Conpair: concordance and contamination estimator for matched tumor-normal pairs. *Bioinformatics* **32**, 3196–3198 (2016).
25. Cibulskis, K. *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.* **31**, 213–219 (2013).
26. Kim, S. *et al.* Strelka2: fast and accurate calling of germline and somatic variants. *Nat. Methods* **15**, 591–594 (2018).
27. Narzisi, G. *et al.* Genome-wide somatic variant calling using localized colored de Bruijn graphs. *Commun Biol* **1**, 20 (2018).
28. Wala, J. A. *et al.* SvABA: genome-wide detection of structural variants and indels by local assembly. *Genome Res.* **28**, 581–591 (2018).
29. Chen, X. *et al.* Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics* **32**, 1220–1222 (2016).
30. Layer, R. M., Chiang, C., Quinlan, A. R. & Hall, I. M. LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol.* **15**, R84 (2014).
31. Xi, R., Lee, S., Xia, Y., Kim, T.-M. & Park, P. J. Copy number analysis of whole-genome data using BIC-seq2 and its application to detection of cancer susceptibility variants. *Nucleic Acids Res.* **44**, 6274–6286 (2016).
32. Tate, J. G. *et al.* COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res.* **47**, D941–D947 (2019).
33. 1000 Genomes Project Consortium. *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
34. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).
35. Sherry, S. T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **29**, 308–311 (2001).
36. MacDonald, J. R., Ziman, R., Yuen, R. K. C., Feuk, L. & Scherer, S. W. The Database of Genomic Variants: a curated collection of structural variation in the human genome. *Nucleic Acids Res.* **42**, D986–92 (2014).
37. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol.* **17**, 122 (2016).
38. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
39. *Polaris*. (Github), https://github.com/Illumina/Polaris.
40. Jeffares, D. C. *et al.* Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. *Nat. Commun.* **8**, 14061 (2017).
41. Emde, A.-K. *et al.* Detecting genomic indel variants with exact breakpoints in single- and paired-end sequencing data using SplazerS. *Bioinformatics* **28**, 619–627 (2012).
42. Dockstore. Available at, https://dockstore.org/containers/quay.io/pancancer/pcawg-sanger-cgp-workflow:develop. (Accessed: 27th May 2019).
43. HATCHet version used in this study, https://github.com/raphael-group/hatchet commit 0e626b0.
44. NYGC companion website, https://www.nygenome.org/bioinformatics/3-cancer-cell-lines-on-2-sequencers/.

Acknowledgements

We thank Marcin Imielinski (NYGC-Cornell) for the HCC-1143 and HCC-1143BL cell lines, Simone Zaccaria and Ben Raphael (Princeton University) for their help with HATCHet. We also thank the Computational Biology team at the New York Genome Center for their support and suggestions, Stephen Bavier for his help with the companion website, Kevin McGhee for his help with data repository, Rajeeva Musunuri for the Outpost QC interface, the NYGC production lab for library preparation and sequencing, the NYGC Research Computing group for their help with data processing and the NYGC Software Engineering team for the automation of the pipeline. This study was funded internally by the New York Genome Center and partially supported by the Informatics Technology for Cancer Research (ITCR) program of the National Institutes of Health (NIH) [1R21CA220411-01A1].

Author contributions

K.A., M.S., D.M.O., M.C.Z., S.G. and N.R. conceived the project and designed the experiments. K.A., M.S., M.J., R.S., J.S., K.N., J.C. and N.R. analyzed the data. V.J. cultured and karyotyped the cell lines. D.M.O., S.G., M.C.Z. and N.R. supervised the work. K.A., M.S., M.J., R.S. and N.R. drafted the manuscript. All authors contributed to the manuscript and approved the submitted version.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41598-019-55636-3.

Correspondence and requests for materials should be addressed to N.R.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.