Comparison of SureSelect and Nextera Exome Capture Performance in Single-Cell Sequencing

Wendy J. Huss^a Qiang Hu^b Sean T. Glenn^{c,d} Kalyan J. Gangavarapu^a
Jianmin Wang^b Jesse D. Luce^c Paul K. Quinn^c Elizabeth A. Brese^e
Fenglin Zhan^f Jeffrey M. Conroy^c Gyorgy Paragh^{g,h} Barbara A. Foster^a
Carl D. Morrison^c Song Liu^b Lei Wei^b

^aDepartment of Pharmacology and Therapeutics, Roswell Park Comprehensive Cancer Center, Buffalo, NY, USA;

^bDepartments of Biostatistics and Bioinformatics, Roswell Park Comprehensive Cancer Center, Buffalo, NY, USA;

^cCenter of Personalized Medicine, Roswell Park Comprehensive Cancer Center, Buffalo, NY, USA; ^dDepartment of Cancer Genetics, Roswell Park Comprehensive Cancer Center, Buffalo, NY, USA; ^eDepartment of Pathology, Roswell Park Comprehensive Cancer Center, Buffalo, NY, USA; ^fPET/CT Center, The First Affiliated Hospital of the University of Science and Technology of China, Hefei, China; ^gDepartment of Dermatology, Roswell Park Comprehensive Cancer Center, Buffalo, NY, USA; ^hDepartment of Cell Stress Biology, Roswell Park Comprehensive Cancer Center, Buffalo, NY, USA

Keywords

Single-cell sequencing · Exome capture · Agilent SureSelect · Illumina Nextera · Somatic mutations

Abstract

Background: Advances in single-cell sequencing provide unprecedented opportunities for clinical examination of circulating tumor cells, cancer stem cells, and other rare cells responsible for disease progression and drug resistance. On the genomic level, single-cell whole exome sequencing (scWES) started to gain popularity with its unique potentials in characterizing mutational landscapes at a single-cell level. Currently, there is little known about the performance of different exome capture kits in scWES. Nextera rapid capture (NXT; Illumina, Inc.) has been the only exome capture kit recommended for scWES by Fluidigm C1, a widely accessed system in single-cell preparation. **Results:** In this study, we com-

pared the performance of NXT following Fluidigm's protocol with Agilent SureSelectXT Target Enrichment System (AGL), another exome capture kit widely used for bulk sequencing. We created DNA libraries of 192 single cells isolated from spheres grown from a melanoma specimen using Fluidigm C1. Twelve high-yield cells were selected to perform dual-exome capture and sequencing using AGL and NXT in parallel. After mapping and coverage analysis, AGL outperformed NXT in coverage uniformity, mapping rates of reads, exome capture rates, and low PCR duplicate rates. For germline variant calling, AGL achieved better performance in overlap with known variants in dbSNP and transition-transversion ratios. Using calls from high coverage bulk sequencing from blood DNA as the golden standard, AGL-based scWES demonstrated high positive predictive values, and medium to high sen-

Wendy J. Huss, Qiang Hu, and Sean T. Glenn contributed equally to this work.

sitivity. Lastly, we evaluated somatic mutation calling by comparing single-cell data with the matched blood sequence as control. On average, 300 mutations were identified in each cell. In 10 of 12 cells, higher numbers of mutations were identified using AGL than NXT, probably caused by coverage depth. When mutations are adequately covered in both AGL and NXT data, the two methods showed very high concordance (93–100% per cell). **Conclusions:** Our results suggest that AGL can also be used for scWES when there is sufficient DNA, and it yields better data quality than the current Fluidigm's protocol using NXT.

© 2019 S. Karger AG, Basel

Background

Advances in single-cell sequencing (SCS) have greatly expanded our understanding of heterogeneous cellular systems [1–6]. Compared with traditional bulk sequencing which measures undifferentiated signals from a mixture of heterogeneous populations, SCS focuses on genetic mosaicism at a single-cell level. This unprecedented resolution allows us to study complex and dynamic biological processes such as tumor clonal evolution [5, 7], immune response [8], and the central nervous system [9].

Most early SCS studies were based on RNAseq. Recently single-cell whole exome sequencing (scWES) has started to gain popularity because of its capacity to identify genomic alterations. Compared with bulk WES, scWES faces unique challenges. A typical single cell only contains pictograms of DNA; therefore, scWES requires whole-genome amplification (WGA) followed by exome capture. While there are some studies comparing various types of WGA methods [10, 11], little is known about the performance of different exome capture kits in the single-cell environment. Previous evaluations of exome capture kits were limited to bulk samples [12]. In the first commercially available automated single-cell preparation system Fluidigm C1, the only manufacturer-recommended exome enrichment kit is Illumina Nextera Rapid Capture Exome (NXT; <https://www.fluidigm.com/reagents/single-cell-dna-sequencing>). Agilent SureSelectXT Reagent (AGL), a popular exome capture kit for bulk sequencing, has been used for scWES in recently published studies [13].

In this study, we compared the performance of Illumina NXT and AGL capture kits for scWES. To avoid varying DNA quality among individual single cells, we designed a single-cell dual-exome sequencing strategy where two aliquots of DNA from the same single cell were

extracted and then captured and sequenced using NXT and AGL separately. We also sequenced the matched blood sample for comparison purpose. The single-cell data were evaluated on three levels: mapping and coverage statistics, germline single nucleotide polymorphism (SNP) calling, and somatic mutation identification.

Methods

Melanoma Sphere Formation Assay

The sphere formation assay was performed with cells isolated from 8 enzymatically digested pieces of a metastatic melanoma specimen. The primary digestion was in a dispase solution for 16 h at 4°C as previously described [14]. The secondary digestion and sphere formation assay was performed as previously described [15] with the following reagents. Cells were plated at a density of 50,000 cells/well in ultra-low attachment 24-well plates (VWR, Radnor, PA, USA) in 12 replicates. Cells were suspended in 100 µL medium with 0.5% w/v methyl cellulose (Sigma Aldrich, St. Louis, MO, USA). The mixture was pipetted around the rim of the well, and a total of 900 µL of DMEM/F-12 (1:1) medium (Invitrogen, Carlsbad, CA, USA), supplemented with 40 ng/mL each of epidermal growth factor (Invitrogen), progesterone (Sigma Aldrich), basic fibroblast growth factor (Peprotech, Rocky Hill, NJ, USA), and 20 µg/mL of insulin (Sigma Aldrich), was added to each well. The plates were maintained in a 5% CO₂ incubator at 37°C for 10–14 days. The spheres in each well were visually quantitated after 10–14 days under a microscope at 10× magnification.

The spheres were dissociated from methyl cellulose with dispase and incubated at 37°C for 30 min. The methyl cellulose and dispase mixture was mixed thoroughly by pipetting several times. The spheres were collected with centrifugation at 240 g, supernatant was aspirated, and the sphere pellet was suspended in trypsin-EDTA (0.25 v/v; Invitrogen) and incubated at 37°C for 15–30 min. Digestion was inhibited by adding DMEM/F-12 (1:1) medium (Invitrogen), supplemented with 40 ng/mL each of epidermal growth factor, progesterone, basic fibroblast growth factor, and 20 µg/mL of insulin. The spheres were collected by centrifugation at 530 g. The sphere pellet was washed by resuspension in HBSS with 2% FBS, and the cells were collected by centrifugation at 530 g. The cells were resuspended in 1 mL DMEM/F-12 (1:1) medium and then passed approximately 10–15 times through a 20G needle to establish a single-cell suspension. The single cells were then washed, collected by centrifugation, resuspended in 1× PBS with HBSS and 2% FBS, and sent to Genomics Shared Resources facility at Roswell Park, where they were then used for extraction of DNA.

Single-Cell Capture and Library Preparation Using the Fluidigm C1

Whole genome-amplified DNA is generated using the Fluidigm C1 system and the C1 integrated fluidic circuit (IFC) from single cells. The IFC is first primed using the provided reagents in the Fluidigm DNA-sequencing kit. While the IFC is priming, cell size, count, and viability are determined using trypan blue and an automated cell counter. A cell suspension is prepared at approximately 400 cells/µL, and C1 suspension reagent is added to ensure proper buoyancy of the cells as they pass through the IFC for capture. The

cell suspension is loaded into the primed IFC, and capture is completed by the C1 system. After the capture is completed, the IFC is imaged on an EVOS FL Auto2 to determine the number of successful single-cell captures. The remaining reagents for lysis and WGA are added to the IFC and processed by the C1 system. Whole genome-amplified DNA is harvested from the IFC at completion and quantitated by PicoGreen for downstream sequencing applications.

Fluidigm NXT

Dual-indexed paired-end libraries followed by exome enrichment is prepared using the NXT kit (Illumina, Inc.). Following the Fluidigm's protocol (#PN 100-7135 H1) for using C1 to generate single-cell libraries for DNA sequencing, 10 ng of each DNA sample is fragmented to a size range of 150–1,000 bp with adapter sequences added by tagmentation as per the manufacturer's instructions. Following purification, the tagmented DNA is PCR amplified for 10 cycles with unique indices included in each set of samples. Libraries are purified and validated for appropriate size on a 2100 Bioanalyzer High Sensitivity DNA chip (Agilent Technologies, Inc.). The DNA library is quantitated using PicoGreen quantification (Qubit) prior to normalization and pooling. Whole exome capture of 250 ng of each indexed DNA library is performed by pooling groups of 6 samples and hybridizing to the capture probes for 2 h at 58 °C. The captured regions are then bound to streptavidin magnetic beads and washed to remove any non-specific bound products. The purified products are eluted from the beads and the hybridization. Wash and elution procedure is repeated to further enrich the targeted regions. The enriched DNA library is purified, PCR is amplified a second time, and purified again prior to quantitation by KAPA qPCR. Each pool is normalized to 10 pM, loaded, and clustered to individual lanes of a HiSeq Flow Cell using an Illumina cBot (TruSeq PE Cluster Kit v3), followed by 2 × 101 PE sequencing on a HiSeq2500 sequencer according to the manufacturer's recommended protocol (Illumina, Inc.).

SureSelectXT All Exons

Individual exome capture of each DNA sample followed by single-indexed library generation was carried out using AGL (Agilent, Inc.). 200 ng of DNA was fragmented to a size range of 150–200 bp, followed by end repair, adaptor ligation, and low-cycle (6–10 cycles) PCR. Libraries are purified and validated for appropriate size (225–275 bp) on a 2100 Bioanalyzer DNA1000 chip (Agilent Technologies, Inc.). 750 ng of purified library was then hybridized to the SureSelectXT All Exon V5 Capture library (Agilent, Inc.) for 18 h at 65 °C. The captured regions are then bound to streptavidin magnetic beads and washed to remove any non-specific bound products. The eluted library undergoes a second 12-cycle PCR amplification to add sample-specific barcodes necessary for multiplexing. The final libraries are purified, measured by Bioanalyzer (250–350 bp), and quantitated using KAPA qPCR. The individual libraries were pooled in equimolar 2 nM final concentration. Each pool is diluted to 10 pM, loaded, and clustered to individual lanes of a HiSeq Flow Cell using an Illumina cBot (TruSeq PE Cluster Kit v3), followed by 2 × 100 PE sequencing on a HiSeq2500 sequencer according to the manufacturer's recommended protocol (Illumina, Inc.).

Bioinformatics Data Analysis

The bioinformatics analysis began by utilizing high-quality paired-end reads passing Illumina RTA filter aligned to the NCBI

human reference genome (GRCh37) using Burrows-Wheeler Aligner [16]. PCR-duplicated reads were marked and removed using Picard (<http://picard.sourceforge.net/>). Mapping and coverage statistics were calculated using SAMtools [17] and summarized by a customized program. Germline variants, including SNPs and small insertions/deletions (Indels), were identified using GATK with default parameters [18]. The quality of variant calls was evaluated using a previously published method by the heterozygous/homozygous ratio, the percentage of overlap with dbSNP, and the transition-transversion ratio [19]. Putative somatic mutations were identified by running variation detection module of Bambino using the bulk blood data as control [20]. The raw mutation calls were filtered to exclude false calls by mapping and sequencing errors as previously described [21]. The identified somatic mutations were further compared to the public human germline databases including dbSNP [22], 1000 Genomes Project [23], and National Heart, Lung, and Blood Institute's Exome Sequencing Project [24] to exclude remaining germline polymorphisms. All mutations were annotated using ANNOVAR [25] with NCBI RefSeq database. To compare the somatic mutations' statuses across samples, all identified mutations were first revisited in every sample to extract the mutant read counts and coverage using the Mutation Reads Extractor (in submission). A minimum of two mutant reads was required for any somatic mutation to be considered as positive (mutation present) in a given sample. For non-positive samples, a minimum of 20× coverage was required to be considered as negative; otherwise it would be classified as unknown (due to low coverage). We also excluded any somatic calls where the matched germline coverage was less than 10× to avoid remaining germline variants.

Results

Single-Cell Preparation Using Fluidigm C1

Due to tumor heterogeneity, we digested the melanoma specimen and cultured as spheroids in order to enrich for a more homogeneous cell population. The spheroid formation assay enriched for epithelial cells and clonal expansion of progenitor cells. Spheroids were digested to a single-cell suspension to allow single-cell capture. We captured 192 cells on two plates (96 cells per plate). The amount of yielded DNA of each cell ranged from 8.6 to 414.8 ng (median = 174.2 ng) (online suppl. Table S1; see www.karger.com/doi/10.1159/000490506 for all online suppl. material). One tenth (21 out of 192) of the captured single cells were unsuccessful due to debris and excluded from further analyses.

Single-Cell Dual-Exome Sequencing and Coverage Statistics

We selected 12 cells with high DNA yield (≥250 ng) for comparison with the two exome capture methods. Each cell's DNA was captured using two types of exome kits, NXT and AGL, in parallel, then sequenced by a

Table 1. Summary of single-cell DNA yield and whole exome sequencing (WES) data statistics

Cell ID	DNA yield, ng	WES capture method	Total reads	Average sequencing depth	Bases covered by $\geq 20\times$, %	Mapping rate, %	PCR duplicate rate, %	Exome capture rate, %
C2-ES41	274.1	AGL NXT	53,565,804 16,693,326	41.44 6.80	61.41 9.20	98.36 97.93	5.66 49.84	38.59 24.95
C2-ES52	332.8	AGL NXT	51,521,762 189,716,350	38.84 91.37	63.51 74.00	98.19 97.83	6.72 43.27	37.60 29.55
C2-ES53	251.8	AGL NXT	54,951,830 788,274	39.83 0.34	50.85 0.00	98.28 97.61	6.74 44.33	36.15 26.55
C2-ES76	254.3	AGL NXT	55,470,896 3,444,166	41.11 1.90	40.91 1.61	98.24 97.68	8.15 36.01	36.97 33.89
C2-ES77	331.6	AGL NXT	58,879,580 96,290,974	44.96 41.72	56.41 46.02	98.05 97.33	8.03 47.47	38.09 26.55
C2-ES89	294.4	AGL NXT	49,516,628 19,342,198	36.59 6.00	29.95 8.46	98.25 97.77	7.92 61.16	36.86 19.01
D2-ES15	311.1	AGL NXT	60,049,500 171,272,932	42.02 55.91	43.02 46.50	98.03 98.05	9.05 57.78	34.90 20.06
D2-ES30	336.4	AGL NXT	54,827,196 79,052,692	36.45 31.14	55.55 45.55	98.04 98.32	13.57 52.36	33.16 24.20
D2-ES48	358.0	AGL NXT	48,950,294 53,949,680	33.63 21.94	23.97 17.70	97.85 97.78	12.41 52.49	34.27 24.97
D2-ES78	318.1	AGL NXT	58,009,390 49,801,180	39.55 19.53	59.21 31.90	98.13 98.19	10.00 53.37	34.01 24.08
D2-ES84	361.0	AGL NXT	49,593,482 26,526,156	34.05 12.07	34.57 15.02	98.13 97.24	9.38 47.99	34.24 27.90
D2-ES96	350.4	AGL NXT	59,139,102 21,932,956	39.79 9.46	30.22 10.97	97.97 97.67	12.34 48.56	33.56 26.45

HiSeq2500 with six samples per lane. The two types of kits led to dramatically different levels of reads uniformity: all AGL samples were within 49–60 million reads per sample; however, in the NXT samples, it varied from 0.8 to 189 million reads (Table 1).

Sample-level mean coverages followed a similar pattern with the total numbers of reads. Even with a slightly lower number of total reads (654 million for all 12 samples combined using AGL vs. 728 million using NXT), the AGL samples generated a higher mean coverage in 10 out of 12 cells. More specifically, the mean coverage of the AGL samples within the targeted region ranged from 34 $\times$ to 45 $\times$ with a median of 40 $\times$, compared with the mean coverage of the NXT samples from 0.34 $\times$ to 91 $\times$ with a median of 16 $\times$. Within each capture method, higher coverage values were correlated with the number of reads (Pearson $r = 0.78$ and 0.97 for AGL and NXT,

respectively), but the slope of this correlation in AGL (0.7 $\times$ per million reads) was higher than in NXT (0.4 $\times$ per million reads), suggesting the reads were more “efficiently” used to generate coverage in AGL than in NXT (Fig. 1). Further analyses revealed several factors that may contribute to this: (1) although both methods have good mapping rates (median: 98.13 and 97.78% for AGL and NXT, respectively), the mapping rate was slightly higher in AGL than in NXT runs ($p < 0.05$, Student’s t test); (2) probably the biggest factor was PCR duplicate rates: all NXT runs were associated with high PCR duplicate rates, ranging from 36.01 to 61.16% with a median of 49.20%; in comparison, none of the AGL runs’ PCR duplicate exceeded 14%, whereas in most (8 of 12) cells, it was less than 10%; (3) exome capture rates, which represents the fraction of reads that fall into the designated target region, were higher in AGL runs (median = 35.5%)

Fig. 1. Comparison of coverage statistics using the Agilent SureSelectXT Target Enrichment System (AGL) and Nextera rapid capture (NXT). **a** Cumulative coverage distribution within the targeted region for AGL (left) and NXT (right). **b** Correlation between mean coverage and total numbers of reads using AGL and NXT single-cell whole exome sequencing.

than the NXT ones (median = 25.7%) (Fig. 2a–c). Noticeably, the insert size was also different in the two methods: 199–235 bp and 114–287 bp for AGL and NXT runs, respectively (Fig. 2d).

Performance of SNP Calling Compared with Blood Sample

To assess the feasibility of identifying SNPs from sc-WES data and the quality of potential calls, germline SNP calls were generated from the current single-cell samples

Fig. 2. Comparison of mapping statistics using the Agilent SureSelectXT Target Enrichment System (AGL) and Nextera rapid capture (NXT). **a** Mapping rates: percentage of reads that were successfully mapped to the reference genome. **b** PCR duplicate rates: percentages of reads that were marked as PCR duplicate. **c** Exome capture rates: percentage of data within the targeted region. **d** Mean insert sizes: the average inferred insert sizes.

and a matched blood sample of the same patient. The blood sample was sequenced at a higher depth (mean 230×) to minimize the false-negative rate. A total of 43,000 SNPs were found from the blood sample. For single cells, the numbers of SNP calls from AGL samples were between 16,000 and 40,000 with a median of 31,000, but they fell into a much wider range for the NXT samples (3,000–53,000, median = 23,000) (Fig. 3a). The ratio of heterozygous/homozygous calls was highest (1.49) in the blood sample and much lower in the single-cell data (median 0.56 and 0.47 for AGL and NXT samples, respectively). One single-cell sample, C2-ES52 by NXT, had the

highest heterozygous/homozygous ratio among all single cells (1.42). This sample also had the highest numbers of SNPs (53,000), which can be explained by the nonproportionally high number of reads (406 million) and overall coverage in this sample (average 203×) (Fig. 3b).

We further evaluated the quality of SNP calls by comparing the identified SNPs with the known human SNP database of dbSNP [22]. All AGL and blood samples exhibited a high concordance with the dbSNP (99.47% in blood, 99.21–99.39% for the AGL samples), while the NXT samples had noticeably lower overlaps with the dbSNP (97.64–98.64%). A similar pattern was observed in

Fig. 3. Summary statistics of germline variant calls. Four common metrics of germline variant calls: total numbers of variant calls per cell (a); the percentage of variant calls that overlap with known variants in dbSNP (b); the ratios of heterozygous versus homozygous calls (c); the ratios of transition versus transversion calls (d) using blood bulk, Agilent SureSelectXT Target Enrichment System (AGL) and Nexera rapid capture (NXT) single-cell (SC) whole exome sequencing.

another quality metric of SNP calls, the transition/transversion ratio: highest for the blood sample (2.63), similar with a slight decrease in the AGL samples (range 2.57–2.62 with a median of 2.60), and much lower in the NXT samples (2.43–2.57 with a median of 2.49) (Fig. 3c, d).

To summarize the overall data quality, we calculated the sensitivity and positive predictive values (PPVs) for each single-cell sample with AGL and NXT, using the blood sample as the golden standard. Indels and SNPs were summarized separately. The AGL samples had a similar sensitivity for SNPs (35–88%) and Indels (31–80%), but higher

PPVs for SNPs (median 96%) than Indels (median 72%). On the other hand, the NXT samples showed a much poorer performance: the sensitivities for SNPs ranged from 3 to 45%, and the PPVs were around 40% (Fig. 4, top); the Indels had even worse performance: neither sensitivity nor PPVs reached 15% in any cell (Fig. 4, bottom).

Concordance of Somatic Mutations between AGL and NXT

To evaluate the performance of somatic mutation calling, we generated somatic mutation calls from each sin-

Fig. 4. Concordance of single-cell germline variant calls compared with blood. Sensitivity and positive predictive values of germline variant calls using high coverage bulk sequencing from blood as the golden standard using blood bulk, Agilent SureSelectXT Target Enrichment System (AGL) and Nextra rapid capture (NXT) single-cell (SC) whole exome sequencing. Top: SNPs, bottom: small insertions/deletions (Indels).

gle-cell sample by comparing it with the matched blood sequences. We found an average of 297 mutations per cell with a range of 171 to 530 mutations. The complete list of identified somatic mutations is available as supplementary information (see online suppl. Table S2). Most were single nucleotide variants, which account for over 99% of all mutations. The median variant allele fractions in most cells were around 20%. The missense-to-silent ratios ranged from 1.4 to 2.4. In 10 of 12 cells, the AGL method found more mutations than NXT, which was consistent with sequencing depth. Overall, the total number of mutations identified by AGL was much higher than the number of mutations identified by NXT (ratio = 1.59). There were only two cells (C2-ES52 and D2-ES15) where NXT found more mutations, whereas AGL still identified 79% and 85% of all mutations in those two cells (Table 2).

When adequately covered in both platforms, the mutations' statuses measured by AGL and NXT were highly consistent. The concordance between mutation statuses, defined as the percentage of mutations where AGL and NXT data had the same status ("present" or "absent"), ranged from 93.0 to 100% in the current 12 cells (Table 2). This high concordance between two independent runs of exome capture and sequencing confirmed that these mutations do exist in the single-cell library and that they were not caused by artefacts during exome capture or sequencing. However, the majority of these mutations (>95% in all cells) were only observed in one but not in any of the other single cells, suggesting these mutations were unlikely to be present in the bulk tissue before cell isolation and growing of spheres.

Discussion

SCS has been emerging as a new technology with great potentials in providing high-resolution insights into heterogeneous cell populations. Using SCS to track the dynamic mutational profile on the DNA level, as one of its most potentially useful applications, has been slow to develop due to technical difficulties such as allelic dropout, non-uniform coverage, and WGA artefacts [26, 27]. There is an urgent need for developing protocols that overcome these SCS-specific challenges.

The current results suggest that the AGL kit outperformed the Fluidigm-recommended NXT protocol by a higher mapping rate, a lower duplicate rate, and a better uniformity of coverage. Furthermore, the AGL kit outperformed NXT in germline SNP calling in the numbers of identified variants, the quality of variant calls as reflected by the percentage of overlap with dbSNP, and the transition-transversion ratio. Remarkably, the quality of variant calls by AGL single cells was comparable to that of the blood bulk sequencing, when measured by these two quality metrics. On the other hand, when the available amount of DNA is low, NXT might be the only choice as it only requires 10 ng of DNA instead of 200 ng required by AGL. In the current study of the 192 cells captured using C1, only 73 (38%) yielded more than 200 ng as required by AGL, whereas almost all cells (191/192) reached the input DNA requirement of NXT. Equally importantly, however, the DNA input amount could play an important role in determining data quality. Although we previously evaluated and found little difference in data quality with over ten times difference in DNA input amounts (200 to 3,000 ng) [19], it is possible that the data

Table 2. Somatic mutations identified in 12 single cells

Cell	All mutations	Indels	VAF median	Missense/silent ratio	Percent of private mutations*	Found by		Concordance** (AGL vs. NXT), %
						AGL	NXT	
C2-ES41	257	0	0.22	2.1	98.1	256	58	98.3
C2-ES52	390	3	0.16	1.9	97.9	309	347	93.0
C2-ES53	338	1	0.23	1.7	99.1	338	3	100.0
C2-ES76	269	1	0.24	2.3	94.8	269	31	100.0
C2-ES77	413	1	0.18	1.7	98.1	368	326	94.9
C2-ES89	201	1	0.19	1.4	96.5	198	76	100.0
D2-ES15	530	1	0.19	1.8	97.2	451	481	96.9
D2-ES30	281	1	0.18	1.5	96.1	252	218	95.0
D2-ES48	171	2	0.21	2.4	94.7	155	123	94.7
D2-ES78	230	1	0.23	2.4	97.4	218	151	95.9
D2-ES84	266	1	0.20	1.7	97.7	258	144	97.8
D2-ES96	222	2	0.20	2.0	95.5	214	113	95.5

Indel, small insertions/deletions; VAF, variant allele fractions.

* Private mutations: mutations that are only found in one cell but not in any of the other cells in the current study. ** Concordance defined as the fraction of comparisons where the two methods generated the same status for a mutation in a given sample, excluding low coverage samples (either AGL or NXT or both) and the samples where the mutation was not detected by any method.

quality is more sensitive to DNA amount under the current SCS settings with significantly less DNA.

In this study, we focused on comparing existing protocols as-is and did not evaluate the effects of the amount of input DNA. Future studies using modified protocols to compare the two capture methods using the same amounts of DNA, at both high and low levels, would be more informative in optimizing capture method selection for different conditions. Also, dual-exome sequencing was only performed on 12 high DNA yield cells. Future evaluations on larger numbers of single cells, especially low DNA yield cells, will be necessary to provide a more systematic comparison between the two capture methods.

Some of the quality metrics we measured in this study are related to coverage. For example, the number of SNPs can be improved with higher sequencing depth. Another example is the heterozygous/homozygous ratio: since a low ratio in single-cell samples is likely caused by the allelic dropout during WGA, higher sequencing depth can help detect the allele that was not well-amplified. One particular NXT sample (C2-ES52) with the highest coverage also had the highest number of variants and best heterozygous/homozygous ratio. In future sequencing, these metrics can be improved by simply increasing coverage. Other metrics, such as mapping rates and overlap with dbSNP, are unlikely to be caused by coverage, but reflect the quality of reads.

Somatic mutation calling from scWES remains a challenge. In the current study, several hundreds of mutations were detected in every cell. Most mutations were found by sequencing runs of both exome capture kits, suggesting they are unlikely to be sequencing errors. Therefore, these mutations were either real biological mutations acquired during sphere culture or potential artefacts introduced by the WGA process as discussed by previous studies [26–28]. The median variant allele fractions in most cells were low, which could be caused by either WGA artefacts or additional copy number variations. In this study, we did not retrospectively bank the bulk spheres DNA and, therefore, cannot further distinguish these two possibilities. Furthermore, the performance of mutation calling can also be improved by combining multiple callers especially the ones optimized for SCS.

Conclusions

In summary, our results suggest that AGL is suitable for exome capture in SCS and can achieve better quality data than the default NXT kit in several aspects such as coverage uniformity, low PCR duplicate rates, and quality of reads. As a result, AGL outperformed NXT in overall germline and somatic variant calling.

Availability of Data and Materials

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Statements of Ethics

The Roswell Park Comprehensive Cancer Center's Institutional Review Board (IRB) gave approval for this study. The patient was consented for next-generation sequencing, remnant tissue procurement, and collection of blood through RPCI IRB-approved protocols.

Disclosure Statement

The authors report no conflicts of interest.

Funding Sources

This work was supported by the Roswell Park Comprehensive Cancer Center and National Cancer Institute (NCI) grant P30CA016056 and the Roswell Park Alliance Foundation.

References

- 1 Satija R, Farrell JA, Gennert D, Schier AF, Regev A: Spatial reconstruction of single-cell gene expression data. *Nat Biotechnol* 2015;33:495–502.
- 2 Eberwine J, Sul JY, Bartfai T, Kim J: The promise of single-cell sequencing. *Nat Methods* 2014;11:25–27.
- 3 Hou Y, Song L, Zhu P, Zhang B, Tao Y, Xu X, et al: Single-cell exome sequencing and monoclonal evolution of a JAK2-negative myeloproliferative neoplasm. *Cell* 2012;148:873–885.
- 4 Xu X, Hou Y, Yin X, Bao L, Tang A, Song L, et al: Single-cell exome sequencing reveals single-nucleotide mutation characteristics of a kidney tumor. *Cell* 2012;148:886–895.
- 5 Navin NE: The first five years of single-cell cancer genomics and beyond. *Genome Res* 2015;25:1499–1507.
- 6 Wills QF, Mead AJ: Application of single-cell genomics in cancer: promise and challenges. *Hum Mol Genet* 2015;24:R74–R84.
- 7 Eirew P, Steif A, Khattra J, Ha G, Yap D, Farahani H, et al: Dynamics of genomic clones in breast cancer patient xenografts at single-cell resolution. *Nature* 2015;518:422–426.
- 8 Shalek AK, Satija R, Adiconis X, Gertner RS, Gaublotte JT, Raychowdhury R, Schwartz S, Yosef N, Malboeuf C, Lu D, Trombetta JJ, Gennert D, Gnirke A, Goren A, Hacohen N, Levin JZ, Park H, Regev A: Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature* 2013;498:236–240.
- 9 Molinaro AM, Pearson BJ: In silico lineage tracing through single cell transcriptomics identifies a neural stem cell population in planarians. *Genome Biol* 2016;17:87.
- 10 Babayan A, Alawi M, Gormley M, Müller V, Wilkman H, McMullin RP, Smirnov DA, Li W, Geffken M, Pantel K, Joosse SA: Comparative study of whole genome amplification and next generation sequencing performance of single cancer cells. *Oncotarget* 2017;8:56066–56080.
- 11 Borgstrom E, Paterlini M, Mold JE, Frisen J, Lundeberg J: Comparison of whole genome amplification techniques for human single cell exome sequencing. *PLoS One* 2017;12:e0171566.
- 12 Chilamakuri CS, Lorenz S, Madoui MA, Vodák D, Sun J, Hovig E, Myklebost O, Meza-Zepeda LA: Performance comparison of four exome capture systems for deep sequencing. *BMC Genomics* 2014;15:449.
- 13 Walter C, Pozzorini C, Reinhardt K, Geffers R, Xu Z, Reinhardt D, von Neuhoff N, Hanenberg H: Single-cell whole exome and targeted sequencing in NPM1/FLT3 positive pediatric acute myeloid leukemia. *Pediatr Blood Cancer* 2018;65.
- 14 Gangavarpu KJ, Huss WJ: Isolation and applications of prostate side population cells based on dye cycle violet efflux. *Curr Protoc Toxicol* Chapter 2011;22(Unit 22):22.
- 15 Boiko AD: Isolation of melanoma tumor-initiating cells from surgical tissues. *Methods Mol Biol* 2013;961:253–259.
- 16 Li H, Durbin R: Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 2009;25:1754–1760.
- 17 Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R; 1000 Genome Project Data Processing Subgroup: The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 2009;25:2078–2079.
- 18 McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernysky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA: The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 2010;20:1297–1303.
- 19 Zhu Q, Hu Q, Shepherd L, Wang J, Wei L, Morrison CD, Conroy JM, Glenn ST, Davis W, Kwan ML, Ergas JJ, Roh JM, Kushi LH, Ambrosone CB, Liu S, Yao S: The impact of DNA input amount and DNA source on the performance of whole-exome sequencing in cancer epidemiology. *Cancer Epidemiol Biomarkers Prev* 2015;24:1207–1213.
- 20 Edmonson MN, Zhang J, Yan C, Finney RP, Meerzaman DM, Buetow KH: Bambino: a variant detector and alignment viewer for next-generation sequencing data in the SAM/BAM format. *Bioinformatics* 2011;27:865–866.
- 21 Wei L, Liu S, Conroy J, Wang J, Papanicolaou-Sengos A, Glenn ST: Whole-genome sequencing of a malignant granular cell tumor with metabolic response to pazopanib. *Cold Spring Harb Mol Case Stud* 2015;1:a000380.
- 22 Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K: dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* 2001;29:308–311.
- 23 1000 Genomes Project Consortium, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA: An integrated map of genetic variation from 1,092 human genomes. *Nature* 2012;491:56–65.
- 24 NHLBI E: Exome Variant Server, NHLBI GO Exome Sequencing Project (ESP), Seattle, WA. (accessed: June 04, 2013).
- 25 Wang K, Li M, Hakonarson H: ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* 2010;38:e164.
- 26 Dong X, Zhang L, Milholland B, Lee M, Maslov AY, Wang T, Vijj J: Accurate identification of single-nucleotide variants in whole-genome-amplified single cells. *Nat Methods* 2017;14:491–493.
- 27 Hou Y, Wu K, Shi X, Li F, Song L, Wu H: Comparison of variations detection between whole-genome amplification methods used in single-cell resequencing. *GigaScience* 2015;4:37.
- 28 Zafar H, Wang Y, Nakhleh L, Navin N, Chen K: Monovar: single-nucleotide variant detection in single cells. *Nat Methods* 2016;13:505–507.