Chang et al., Next Generat Sequenc & Applic 2014, 1:1
DOI: 10.4172/2469-9853.1000109

Journal of
Next Generation Sequencing & Applications

Research Article Open Access

Validation of Next Generation Sequencing Cancer Panels for Clinical Somatic Mutation Profiling- Identification of Source of Variations and Artifacts using FFPE Tissues

Ken CN Chang*, Yun Zhao, John Kang, Saumya Pant, Ping Qiu, Bo Wei, Russell Weiner and Matthew J Marton

Molecular Biomarkers and Diagnostics, Merck and Co Inc, Rahway, New Jersey, USA

Abstract

Use of a Next Generation Sequencing (NGS)-based test as a clinical trial assay requires that the test be analytically validated to CLIA/CAP regulations. During the course of analytically validating NGS cancer panels for use as patient enrollment assays, we assessed the repeatability, reproducibility and accuracy of commercially available Cancer Panels (Illumina TruSeq Cancer Panel on MiSeq and Life Technologies AmpliSeq v2 Cancer Panel on Ion Torrent PGM). We measured the repeatability and reproducibility by evaluating all variant calls among technical replicates and found in both platforms that variants with higher variant frequency (VF >30%) were called with much higher repeatability and reproducibility than those with lower VF (between 5 and 25%), a level at which many somatic mutations are found. Also, Illumina MiSeq run-to-run reproducibility was significantly higher than that of Ion Torrent PGM. However, Illumina TruSeq library preparation protocol resulted in much lower repeatability than those obtained from Ion Torrent AmpliSeq protocol at the low VF range. To determine the optimal variant call settings, we used different sets of more stringent filters (lower false positive rate, but higher false negative rate), each platform could achieve close to 95% reproducibility and repeatability. Sequenom MassArray was used as a tie-breaker assay for discordant calls between the two NGS platforms to establish the "truth". Our data provide insight into the steps that contribute most to variability, such as the procedure of library preparation, the sequencing-by-synthesis chemistry, the factors that impact mutation calls and sampling variation. We also found very high C to T mutation calls associated with Illumina Cancer Panel using TruSeq library preparation protocol (but not with Ion Torrent using AmpliSeq protocol) when a less stringent filter set was used. The C to T artifact mutation calls from formalin-fixed paraffin-embedded (FFPE) tissue samples observed in this study together with high C to A artifact mutation calls from acoustic shearing of intact DNA observed by others have the potential to negatively impact mutation profiling and mutation signature identification if not carefully addressed. Based on these results we conclude library preparation protocols that start with PCR amplification, such as the AmpliSeq protocol, provide higher repeatability on variant calls with low VF and therefore, are more suitable for mutation profiling and mutation signature studies where somatic mutations (including unknown mutations) are the focus and balanced false positive and false negative rates are critical to success.

Keywords: Next generation sequencing; Somatic mutation; FFPE tissues

Introduction

Technologies come and go; some leave a dramatic and lasting impact, while others are quickly replaced by newer ones. However, never has the field of clinical genomics been so profoundly impacted as it has been by next generation sequencing (NGS) technology. The cluster of new instruments have impacted nearly every front of genomic research from sequencing an entire genome in couple days, to discovering novel aspects of the genome, to identifying large number of rare or unknown mutations that could be used to guide clinical treatment. Whether or not another round of new sequencing technologies comes soon to replace current "next generation" sequencing technologies, their revolutionary impact is here to stay. Recently a huge impulse of NGS publications flooded into the scientific journals [1-12]. This is understandable as scientists try to make their mark in the NGS field. On one hand, they recognize that such innovative technology may fundamentally change the practice of medicine, while on the other hand they know newer technologies may replace the current technology overnight. As the famous quote stated "With great power there must come great responsibility", clinical genomic scientists have the responsibility to ensure the accuracy of any rare or unknown mutations identified through using NGS. They must make certain that mutation calls are reliable and must identify potential data artifacts by carefully validating assays before using them to support patient treatment or to draw important novel conclusions.

With this in mind, two of the most popular bench-top NGS instruments that show great potential for clinical research, the Illumina MiSeq and Ion Torrent Ion Personal Genome Machine (PGM), were evaluated using their corresponding commercially available cancer panels. The goal of this study was to thoroughly understand the two platforms by discovering their pros and cons, to evaluate their performance in terms of library-to-library repeatability and run-to-run reproducibility, and to resolve any discordant variant calls between the two platforms using tie-breaker assays. These efforts will contribute to the validation and decision of which clinical NGS applications could be carried out by either or both platforms. One of the study designs presented here involves six replicates of separate library preparations from the same genomic DNA isolated from FFPE tissue. Each of these library preps were run in the same chip or flow cell to observe library-

***Corresponding author:** Ken CN Chang, Molecular Biomarkers and Diagnostics, Merck & Co Inc, Rahway, New Jersey, USA, Tel: +08889-0100; E-mail: ken.chang@merck.com

Received February 21, 2014; **Accepted** August 22, 2014; **Published** August 24, 2014

Citation: Chang KCN, Zhao Y, Kang J, Pant S, Qiu P, et al. (2014) Validation of Next Generation Sequencing Cancer Panels for Clinical Somatic Mutation Profiling- Identification of Source of Variations and Artifacts using FFPE Tissues. Next Generat Sequenc & Applic 1: 109. doi:10.4172/2469-9853.1000109

to-library variation. Another study design involved six different FFPE tissue samples run in the same chip or flow cell and then the entire experiment was repeated two more times on different days using the same library preparations. This allowed for the understanding of run-to-run variation. Since one of our interests was to evaluate and, when appropriate, validate these two platforms for performing mutation profiling in clinical studies, we focused our attention on the low frequency variant calls, which are common among somatic mutations (including rare and unknown mutations). In order to test the limit of each platform, the manufacturer recommended minimum genomic DNA input was used. The quality of the run and all run-related QC metrics were carefully analyzed to make sure data generated were at least above average based on the manufacturer's specifications.

Various filters including Q scores, minimum read coverage and variant frequencies were used to compare the two platforms, and specific definitions of acceptable range of variant frequencies were applied to determine replicate repeatability/reproducibility. Step-by-step analysis of each protocol was carefully investigated and potential sources of variation and impact factors that could lead to inconsistent low frequency variant calls were identified and are discussed in this manuscript. Recommendations to circumvent the challenges and improve the protocol for future mutation profiling research will also be made in the Discussion section, which we hope will aid efforts to identify mutation signatures [13,14] that could be used to guide patient treatments and predict patient treatment responses.

Materials and Methods

FFPE tissue source and sample preparation

Six colorectal cancer (CRC) FFPE tissue blocks that had been profiled with a NGS-based cancer panel and their corresponding 5 µm sectioned slides were purchased from BioChain Institute (Newark, CA). Sample data generated by BioChain using MiSeq TruSeq Cancer Panel are included in the Supplemental Data section Table 1. Genomic DNA (gDNA) was isolated using Qiagen DNA FFPE Tissue Extraction Kit (Qiagen, Germantown, MD). Qubit (Life Technologies, Carlsbad, CA) and Nanodrop (Thermo Scientific, Waltham, MA) quantification as well as Bioanalyzer (Agilent, Santa Clara, CA) DNA quality analyses were done according to the standard protocol provided by the manufacturers.

Illumina and ion torrent cancer panel comparison

Specifications for each cancer panel, Illumina (San Diego, CA) and Ion Torrent (Life Technologies, Carlsbad, CA), can be found through the corresponding platform websites [15,16]. A comparison table showing platform-specific information including average amplicon sizes and minimum input DNA is included in the Supplemental Data Table 2.

Illumina TruSeq and Ion Torrent AmpliSeq library preparation procedures

Ten 5 µm sections of FFPE tissue slides each were used to prepare gDNA stocks for the six CRC FFPE tissue samples. A standard Qubit-quantified gDNA input amount of 250 ng and 10 ng for Illumina and Ion Torrent Cancer Panels, respectively, was used for all the experiments unless otherwise specified. The TruSeq Amplicon cancer panel library preparation procedures were used for Illumina MiSeq platform and the AmpliSeq cancer panel library preparation procedures were used for Ion Torrent PGM platform. A direct comparison of library preparation protocols [17,18] is included in the Supplemental Data Figure 1. The major differences between the two protocols are 1) the template for the initial PCR product and 2) the number and timing of the PCR. In the

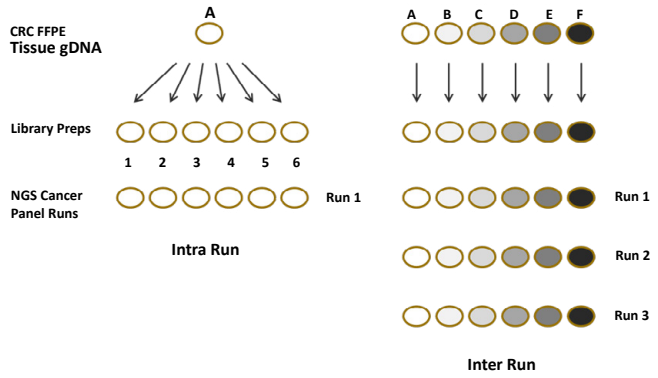

Figure 1: Experimental design Circles with different shades represent gDNA isolated from different colorectal cancer (CRC) FFPE tissue samples. Each run was done by combining libraries with different barcodes either from same sample different library preparations (Intra Run) or from different samples (Inter Run).

Sample Id	Clusters PF	% Reads Identified	% Aligned R1	% Aligned R2	Mean Depth	Het SNPs
Illumina_intra_94_01	1724900	10.8179883	86.4	84.7	7414.67	10658
Illumina_intra_94_02	2157776	13.53283989	85.5	83.7	9169.15	9730.8
Illumina_intra_94_03	2483872	15.5780035	87.2	85.4	10759.37	11699
Illumina_intra_94_04	2332000	14.062551378	86.2	84.5	9999.39	9999.4
Illumina_intra_94_05	2740362	17.18662187	86.2	83.7	11699.01	10759.4
Illumina_intra_94_06	2233597	14.00836351	87.6	85.9	9730.75	9169.1

Table 1: Illumina Intra-Run QC (same sample different library preparations, additional QC data could be found in the Supplemental Data Table S4).

Sample ID	Mapped Reads	Reads On Target	Based On Target	Reads Depth
Ion Torrent_94_01	870,390	95.04%	88.65	3,630.34
Ion Torrent_94_02	1,001,218	95.89%	89.28	4,205.33
Ion Torrent_94_03	917,007	94.20%	88.09	3,771.41
Ion Torrent_94_04	827,560	94.29%	87.97	3,414.00
Ion Torrent_94_05	874,856	94.64%	88.45	3,603.98
Ion Torrent_94_06	979,857	95.77%	89.40	4,099.20

Table 2: Ion Torrent Intra Run QC (same sample different library preparations, additional QC data could be found in the Supplemental Data Table S4).

initial steps, the TruSeq method copies genomic DNA from one strand while the AmpliSeq method copies from both strands. AmpliSeq's first step is a 20-cycle PCR (designated as "Amplify Now") and TruSeq's final step is a 27-cycle PCR (designated as "Amplify Later").

Illumina MiSeq and ion torrent ion Personal Genome Machine (PGM) sequencing procedures

MiSeq Reagent Kit v2 and Ion 318 Chip Kit were used for all the NGS runs with MiSeq and Ion PGM, respectively. Standard protocols were followed for all experiments and instrument runs [19,20], and the reagents were freshly prepared for each run. Seven samples (6 target samples plus a control sample) were multiplexed per chip or flow cell using unique index/barcodes (sample ID and barcode corresponding tables are included in the Supplemental Data, Table 2).

Data analysis process and sequence analysis

Data analysis presented in this manuscript was performed using either MiSeq Reporter (Illumina), Torrent Suite (Ion Torrent, Life Technologies), or OmicSoft (OmicSoft Corporation, Cary, NC) (for both platforms) as mentioned in each result sub-sections or Table legends. The default setting of Q score cutoff for OmicSoft is Q13 unless otherwise specified. Various quality scores and coverage cutoffs were also used in different Tables and Figures and are specified under legends.

Definition of concordant calls among replicates (The Concordance Calculator Excel tool)

An Excel tool (The Concordance Calculator) was designed and an equation was empirically developed to identify the range of variant frequency (VF) for which variant calls between replicates was concordant for a particular data set. This was accomplished individually for each variant called by searching the other replicates for the same variant call. Variables in the equation were the acceptable % CV (coefficient of variation) and the % background variation, which can be altered to generate acceptable VF ranges for a list of variant calls with different VF. Although all variables can be easily changed by entering the desired numbers on the top of the Excel sheet, the proposed equation with starting variables was empirically determined initially based on the observed average degree of variation among all data sets. Since the variables (acceptable % CV and background variation) can be dialed up or down in The Concordance Calculator for other data sets, these variables can be provided in the final clinical sample report for data interpretation. The tool can be applied to any particular data set containing replicate data by 1) starting with a list of minimally filtered variant calls (remove all calls with Q13 or below and all variant calls at VF ≤ 1% and all VF ≤ 100x coverage for Ion Torrent data sets and ≤ 500x coverage for Illumina data sets) from one replicate variant call, 2) calculating the acceptable VF range using the equation of VF +/- (%VF x acceptable %CV + % background variation), then 3) looking for the same ID of the variant on the other replicate data sets to determine if such variant call can be found, and if yes, 4) determining whether the corresponding VF is within the calculated acceptable range. If both answers are yes, this variant call is considered repeatable or reproducible. With the equation one can calculate how many variant calls are repeatable or reproducible, from which % repeatability can be derived. If % repeatability is low, more stringent criteria can be applied, such as higher coverage (e.g., increase from 100x to 300x) or greater VF range (e.g., higher acceptable % VF in the equation) can be easily changed. For example, the acceptable repeatability (such as higher than 95%) might be achievable with higher coverage cutoff. Thus, the tool permits one to report % repeatability in the context of specific filters, which then should be provided with the validation report or clinical sample testing report to specify the limitation of the data.

Experimental design

Since each platform has its own limitations, the manufacturer recommended optimal conditions were used for each cancer panel (including 250 ng input gDNA for Illumina and 10 ng for Ion Torrent). The basic design of the experiments comprised the evaluation of 6 replicates starting from gDNA isolated from a pooled sample in which each replicate underwent independent library preparation and was analyzed in a multiplexed run in the same flow cell or chip (repeatability for library preparation). We also evaluated 6 different CRC FFPE tissue samples through a similar multiplexing design and repeated the same experiment for three times on different days (run-to-run reproducibility). A design diagram depicts exactly how the experiments were conducted is shown in Figure 1.

Sequenom MassArray mutation confirmation assay design

Custom Sequenom MassArray assays were designed to determine whether variant calls could be confirmed on an orthogonal platform [21]. Analysis was performed at Sequenom (San Diego, CA) using standard assay design with primer sequences and the target nucleotides listed under the Supplemental Data Table 3.

Results

Experimental design and library QC and Data QC analysis for both Illumina and Ion Torrent cancer panel comparisons

Due to the frequently encountered challenge of limited availability of clinical samples, the minimum recommended input DNA was used for the reproducibility/repeatability study. A design diagram depicts how the experiments were executed is shown in Figure 1. To make sure our conclusions would be representative of data generated in a clinical study, we ensured that all NGS run passed QC thresholds established by the vendors.

QC data are shown in Table 1. For comparison purpose, Ion Torrent Intra Run QC data is also summarized in Table 2. Additional QC data can be found in Supplemental Data section Table 4.

One key piece of supporting evidence showing high quality data was obtained from these samples is presented in Table 3 (data shown is from the representative sample 94, similar data were obtained from the other 5 samples) as the mutations with VF 30% or higher identified from the two different NGS platforms using two different library preparation protocols and from two different labs (our internal lab and from BioChain) using same FFPE tissue samples were more than 95% concordant for those sequences shared by both cancer panels.

Comparison of library-to-library preparation repeatability for the two cancer panels

Upon evaluation of variant call data, our first observation was that variant calls with high VF were highly concordant among different library preparation replicates. However variant calls with VF between 5% and 25% were not as repeatable and most variant calls with VF below 5% were not repeatable under these conditions. Variants called at <1% VF were excluded from the analysis. Tables 4 and 5 are examples of these variant calls generated from Illumina and Ion Torrent cancer panels, respectively, using same gDNA stock solution and different library preparations. As indicated in the lower three sections of Table

Mutation Frequency	Coverage	Chromosome	Position	Reference	Mutation	Platform
0.5063	399	10	43613843	G	T	Illumina MiSeq Ion Torrent PGM Mi Seq (Biochain)
0.4906	638	10	43613843	G	T	Illumina MiSeq Ion Torrent PGM Mi Seq (Biochain)
0.2	n/a	10	43613843	G	T	Illumina MiSeq Ion Torrent PGM Mi Seq (Biochain)
0.535	1641	10	89717672	C	T	Illumina MiSeq Ion Torrent PGM Mi Seq (Biochain)
0.5781	365	10	89717672	C	T	Illumina MiSeq Ion Torrent PGM Mi Seq (Biochain)
0.16	n/a	10	89717672	C	T	Illumina MiSeq Ion Torrent PGM Mi Seq (Biochain)
0.9976	1698	13	28610183	A	G	Illumina MiSeq Ion Torrent PGM Mi Seq (Biochain)
0.9975	405	13	28610183	A	G	Illumina MiSeq Ion Torrent PGM Mi Seq (Biochain)
1	n/a	13	28610183	A	G	Illumina MiSeq Ion Torrent PGM Mi Seq (Biochain)
0.9976	1695	17	7579472	G	C	Illumina MiSeq Ion Torrent PGM Mi Seq (Biochain)
0.9597	248	17	7579472	G	C	Illumina MiSeq Ion Torrent PGM Mi Seq (Biochain)
1	47	17	7579472	G	C	Illumina MiSeq Ion Torrent PGM Mi Seq (Biochain)
0.9908	1305	4	1807894	G	A	Illumina MiSeq Ion Torrent PGM Mi Seq (Biochain)
0.9933	445	4	1807894	G	A	Illumina MiSeq Ion Torrent PGM Mi Seq (Biochain)
1	24	4	1807894	G	A	Illumina MiSeq Ion Torrent PGM Mi Seq (Biochain)
0.3494	5045	4	55141026	C	T	Illumina MiSeq Ion Torrent PGM Mi Seq (Biochain)
0.3003	323	4	55141026	C	T	Illumina MiSeq Ion Torrent PGM Mi Seq (Biochain)
0.4554	101	4	55141026	C	T	Illumina MiSeq Ion Torrent PGM Mi Seq (Biochain)
0.9982	5046	4	55141055	A	G	Illumina MiSeq Ion Torrent PGM Mi Seq (Biochain)
1	327	4	55141055	A	G	Illumina MiSeq Ion Torrent PGM Mi Seq (Biochain)
0.9915	118	4	55141055	A	G	Illumina MiSeq Ion Torrent PGM Mi Seq (Biochain)
0.9873	10988	5	112175770	G	A	Illumina MiSeq Ion Torrent PGM Mi Seq (Biochain)
0.997	331	5	112175770	G	A	Illumina MiSeq Ion Torrent PGM Mi Seq (Biochain)
0.9844	257	5	112175770	G	A	Illumina MiSeq Ion Torrent PGM Mi Seq (Biochain)
0.6609	4064	7	55249063	G	A	Illumina MiSeq Ion Torrent PGM Mi Seq (Biochain)
0.6438	306	7	55249063	G	A	Illumina MiSeq Ion Torrent PGM Mi Seq (Biochain)
0.5737	190	7	55249063	G	A	Illumina MiSeq Ion Torrent PGM Mi Seq (Biochain)

Table 3: Sequences shared by the two cancer panels showed excellent concordance between platforms and labs for variants with high variant frequency.

Chromosome	Position	Reference	Mutation	Replicate 1 VF	Coverage	Replicate 2 VF	Coverage	Replicate 3 VF	Coverage
Examples of Consistent Variant Calls									
1	43815071	T	G	0.1699	5126	0.1504	6973	0.1429	5075
1	43815081	G	A	0.0772	5003	0.065	6903	0.069	5017
1	43815149	A	C	0.1395	2151	0.0704	2927	0.0551	1977
1	115258743	A	C	0.0971	3843	0.1238	1761	0.0799	5331
1	115258746	A	C	0.0755	3922	0.0909	1286	0.0663	5418
2	212578396	T	A	0.2816	522	0.2754	552	0.2921	736
2	212578412	G	T	0.1073	522	0.125	552	0.0924	736
Examples of Inconsistent Variant Calls for Replicate 1									
3	10183861	C	T	0.3133	466				
3	10191550	C	T	0.1421	739				
3	41266050	C	T	0.1081	2913				
3	178928035	C	T	0.13	1846				
3	178951951	C	T	0.1149	3654				
4	1806164	C	T	0.1203	1663				
4	1807877	G	T	0.1173	1671				
Examples of Inconsistent Variant Calls for Replicate 2									
3	10183871	G	A			0.145	607		
3	10183893	C	T			0.3424	625		
3	10191559	C	T			0.1485	1192		
3	10191562	C	T			0.104	1192		
4	1803621	C	T			0.1311	1556		
4	1807873	C	T			0.1344	774		
4	55144250	C	T			0.1311	2105		
Examples of Inconsistent Variant Calls for Replicate 3									
3	10183890	C	T					0.2679	530
3	10191581	C	T					0.2537	816
3	178916899	C	T					0.1729	931
3	178916929	G	T					0.1697	931
4	1807876	C	T					0.1754	1283
4	1807893	C	T					0.2448	1307
4	153247288	C	T					0.1654	1475

Table 4: Representative variant calls of Illumina Intra Run repeatability for low frequency variants.

4, some variants called for an individual replicate were not observed (or VF below 1%), indicating inconsistent variant calls between replicates. Interestingly in Table 4, several low coverage variant calls with VF as high as 30% (chromosome positions 3-10183861 and 3-10183893) were not observed in the other replicates. Also, some variant calls with as high as 2-3000x coverage (chromosome positions 3-41266050 and 3-178951951) were also not called in all three replicates. Table 5 shows very high consistent variant calls among different library replicates for Ion Torrent AmpliSeq cancer panel even at the level of 3-5% VF range or at below 200x coverage level. The complete data set for all six replicates can be found in the Supplemental Data Table 5.

Comparison of run-to-run reproducibility between Illumina MiSeq and Ion Torrent PGM platforms

Tables 6 and 7 are examples of variant calls generated from Illumina and Ion Torrent cancer panels respectively using the same gDNA, same library preparation, and run on three different days. In general, the run-to-run reproducibility was very high for Illumina runs Table 6; however, the reproducibility for the variant calls for Ion Torrent runs was not as high. As shown in Table 7 (Ion Torrent data set), variant calls at chromosome position 20-36031767 have a range of VF between 15% and 35% among the replicates. Also, variant calls at chromosome position 20-57484566 have a range of VF between 4% and 11% among the replicates, showing good reproducibility but with a higher %CV. The complete data set can be found in the Supplemental Data (Table 6).

Quantification of repeatability and reproducibility

From the examples shown in Table 3, variant calls with high VF (>30%) were very repeatable and could be easily identified by using a relatively low stringency QC filter. However, variant calls with low VF (5-25%) were much less repeatable depending on the library preparation protocol or platform used. In order to quantify the observations shown in Tables 4-7, well-defined acceptance criteria will need to be specified first. A variant call with VF 99% is clearly different with different implications than the identical variant call with VF 25%. Thus, it is necessary to establish a range of VF within which the variant call will be designated as repeatable or reproducible. An Excel tool called The Concordance Calculator was designed and an equation was empirically developed to define the acceptable range of VF when one variant call with a specific VF was used to search against the same variant call from the other replicates (see Materials and Methods). Although the equation was empirically developed, two variables (acceptable % CV and background variation) can be dialed up or down in The Concordance Calculator for other data sets, and if used for assay validation these variables will be provided in the final clinical sample report for data interpretation. For example, if a specific mutation is detected at VF of 12%, using the equation with %CV= 40% and background variation= 2%, a range of 3 and 21% will be implicated as when the test is repeated greater than 95% of time the same mutation will be detected within this range. This is necessary since a recent study showed that cetuximab is more effective in TP53 wild-type rectal cancer. Five-year survival rate increases from 67% to 92% for TP53 wild-type compared to mutant tumors [22]. This study was based on the Sanger sequencing data, which is well known to have sensitivity around 20%. For example, a patient whose tumor has a TP53 mutation detected at a VF of 12% by NGS (which most likely not be detected by Sanger) may respond to the cetuximab treatment significantly better than a patient with same mutation detected at the VF of 25%. Tables 8 and 9 show the example of how The Concordance Calculator and the repeatability/reproducibility equation work. Table 8 shows if Q13 and 100x coverage are used as the default filters for Ion Torrent cancer

Chromosome	Position	Reference	Mutation	Replicate1 VF	Coverage	Replicate 2 VF	Coverage	Replicate 3 VF	Coverage
10	43617393	G	T	0.0769	325	0.0665	316	0.0541	333
10	89624217	G	A	0.0728	261	0.0705	227	0.0498	261
10	89717671	A	G	0.0313	352	0.0487	390	0.0275	363
11	108236191	A	C	0.1244	217	0.1535	241	0.1392	273
13	28608229	A	T	0.1698	212	0.1298	208	0.1216	222
13	28608233	T	C	0.099	202	0.0543	184	0.0686	204
13	48941615	C	A	0.0702	171	0.1173	162	0.0828	157
19	1207082	A	G	0.1206	315	0.0932	322	0.0724	304
19	1207084	G	A	0.1711	304	0.1395	301	0.0807	285
19	1220312	G	A	0.0766	235	0.0984	254	0.08	250
19	1220315	G	A	0.0652	276	0.0532	282	0.0802	324
19	1220316	A	G	0.0496	282	0.0439	296	0.0485	330
19	1220317	C	A	0.0417	264	0.0373	268	0.0508	295
19	1220321	T	C	0.1581	291	0.1845	309	0.2045	352
19	1220519	G	A	0.0915	142	0.094	117	0.0984	122
19	1221246	G	A	0.0343	437	0.0401	399	0.0531	377
19	1221249	G	A	0.0338	414	0.0645	403	0.0565	372
19	1220321	T	C	0.1581	291	0.1845	309	0.2045	352
20	36031767	C	G	0.1077	808	0.135	889	0.1138	800
22	24134064	C	A	0.0343	495	0.0383	470	0.0259	501
4	55141026	C	T	0.3435	262	0.3184	223	0.3003	323
7	116339676	C	T	0.0376	479	0.0343	525	0.0405	518
7	116417428	C	T	0.0749	387	0.0678	398	0.0911	406
7	128845110	A	C	0.2491	273	0.1624	271	0.2305	295
7	128845112	C	A	0.3519	216	0.2404	208	0.3373	252
9	5073849	A	T	0.0411	316	0.0565	354	0.0534	281

Table 5: Representative variant calls of Ion Torrent Intra Run repeatability for low frequency variants.

Chromosome	Positiontable	Reference	Mutation	Replicate 1 VF	Coverage	Replicate 2 VF	Coverage	Replicate 3 VF	Coverage
Examples of Consistent Variant Calls									
1	43815028	C	T	0.067	7251	0.0631	6942	0.0635	6941
1	43815081	G	A	0.0581	7234	0.0697	6889	0.0879	6915
1	43815108	C	T	0.1213	6618	0.1201	6294	0.1196	6080
1	115258766	G	T	0.0822	5197	0.0785	4876	0.0848	4835
1	115258746	A	C	0.0842	5181	0.073	4863	0.0878	4817
1	115258743	A	C	0.1119	5033	0.0974	4756	0.1342	4643
2	212587216	C	T	0.0537	8506	0.0515	7779	0.0522	7229
2	209113154	G	T	0.0655	6241	0.0622	5881	0.0534	5333
2	212652844	G	T	0.1023	2912	0.1226	2602	0.1019	2570
2	212652818	C	T	0.1019	2906	0.1222	2602	0.0988	2570
2	212289046	C	T	0.118	2804	0.0972	2563	0.115	2460
3	10183826	C	T	0.0707	2191	0.0763	2150	0.0974	2105
3	10183832	C	T	0.1134	1967	0.1156	1929	0.1046	1855
3	10183835	C	T	0.0975	1958	0.0727	1966	0.0785	1872
3	10183859	C	T	0.0851	2256	0.0697	2211	0.0814	2187
3	10183867	C	T	0.1571	2254	0.1421	2209	0.1318	2192
3	10183884	C	T	0.1206	2189	0.0947	2164	0.1046	2131
3	10183893	C	T	0.1597	2254	0.1452	2210	0.1318	2193
4	153249445	C	T	0.2599	1358	0.2621	1244	0.2826	1175
4	153249463	T	A	0.1338	1360	0.1342	1244	0.143	1175
4	153249426	C	T	0.0523	1358	0.0603	1244	0.074	1175
5	149453185	C	A	0.0606	3351	0.0678	2937	0.0605	2778
5	170837532	C	T	0.0931	2203	0.0916	2097	0.0839	1920
5	170837625	C	T	0.0729	2208	0.0758	2097	0.0801	1922
5	170837676	C	T	0.0622	1398	0.0558	1291	0.0528	1211
17	7578328	C	T	0.065	3432	0.0739	3344	0.0734	3147
17	7578380	C	T	0.0546	12735	0.057	11829	0.0539	11163
17	7579343	T	A	0.0788	1472	0.0675	1526	0.069	1377
17	7579352	C	T	0.0944	1472	0.0996	1526	0.0922	1377
17	7579374	C	T	0.0703	3430	0.0763	3434	0.0693	3277

Table 6: Representative variant calls of Illumina Inter Run reproducibility for low frequency variants.

panel data set from same sample different library preps, the measured repeatability is around 78% for those variants with VF between 5 and 25%. However, if 300x coverage is used as the cutoff under same condition the repeatability increased to around 95% Table 9. Therefore one should be able to do such a validation experiment for any given validation sample set and define the cutoff requirements through dialing up or down with these variables to obtain a set of variant calls with >95% repeatability, then use this set of filters to do clinical sample testing data analysis. The important thing is to define filters used and the meaning of VF such as 15% means if repeat the same test it is likely (>95% of chance) going to generate same mutation within the VF range of 7 and 23%. For Illumina Intra Run data (same sample different library preparations) using Q13 as the cutoff, if variant calls with coverage below 500x and VF below 5% are first removed before sorting the data by VF it is possible to obtain an identical list of variant calls with VF above 30% as those generated from Illumina's MiSeq Reporter with quality score of 100 (Table 10 and Supplemental data, Venn diagram Table 7). However, using this tool it is also possible to sort the data for those variant calls with VF below 30% then filter out variant calls with coverage above 1900x and VF above 10% to generate a list of variant calls with repeatability of 43.8% which means additional 7 variant calls that are repeatable could be obtained One very important observation is that the repeatability/reproducibility of a variant call is a function of VF and coverage after the most basic default filter is applied, and variant calls with higher VF appeared to be much more repeatable or reproducible than those with lower VF. One more interesting observation is that for those variant calls with VF higher than 30%, not too many of them have C to T calls. However, many C to T variant calls could be found among those with VF between 10 and 20%. And, interestingly, most of these C to T calls were not repeatable, suggesting that most of them might be due to Post Tissue Collection Modifications (PTCM, such as deamination). A complete data set used in these calculations is provided under the Supplemental Data (Tables 8-10).

Resolving discordant calls with Sequenom mass array assay

Sequenom MassArray technology was used to serve as a tie-breaker assay for the discordant calls identified between the two platforms or within the same platform. As Sequenom MassArray is reported to have an average detection sensitivity of around 8% a list of discordant variant calls with VF above 6% were selected for this tie-breaker assay (Table 11). Results of the Sequenom data are shown in Table 11. Some variant calls were consistent with those obtained from Ion Torrent cancer panel while others were consistent with data obtained from Illumina cancer panel. Several discordant variant calls from different samples were not detected by the Sequenom assay and most of them were indicated as below the detection sensitivity and therefore excluded from this table.

Discovery of high number of C to T variants with low variant frequency mainly below 20%

As shown in Tables 4-7, a very large portion of discordant calls in the VF range below 20% within Illumina platform involved C to T mutation, while no such bias was found in Ion Torrent platform. Figure 2 shows such analysis using complete data set. Even with those that showed

Chromosome	Position	Reference	Mutation	Replicate 1 VF	Coverage	Replicate 2 VF	Coverage	Replicate 3 VF	Coverage
Examples of Consistent & Not So Consistent Variant Calls									
10	89685263	T	G	0.057	316	0.0379	264	0.0556	198
10	89717775	A	G	0.1218	238	0.0798	213	0.1348	141
10	89717778	T	G	0.0583	343	0.0769	325	0.0616	211
11	108172364	G	T	0.1824	329	0.2424	363	0.197	198
11	108173701	T	G	0.1498	207	0.249	253	0.1882	170
13	28608229	A	T	0.1366	227	0.0663	196	0.1319	91
14	105241506	C	T	0.0958	428	0.0902	410	0.0569	281
17	7578183	C	G	0.2733	344	0.4144	263	0.3564	202
17	7578253	C	A	0.3196	388	0.3272	327	0.3156	244
5	170837513	C	T	0.1281	203	0.115	200	0.1552	116
7	55221793	G	C	0.1644	590	0.1473	516	0.2708	240
7	55221802	G	A	0.0979	572	0.1931	492	0.0989	263
20	36031767	C	G	0.1525	1167	0.27	1185	0.3556	689
20	57484563	T	C	0.0812	579	0.112	607	0.0439	387
20	57484566	T	C	0.0897	591	0.1086	617	0.0388	387
22	24176307	C	G	0.0483	290	0.0383	261	0.0764	157
4	55972974	T	A	0.5677	421	0.5861	447	0.6202	258
4	153249356	T	C	0.1333	360	0.2216	370	0.145	269
4	153249359	C	A	0.152	375	0.2289	380	0.1786	280
4	153249361	T	C	0.2829	205	0.3636	231	0.2911	158

Table 7: Representative variant calls of Ion Torrent Inter Run reproducibility for low frequency variants.

Replicate 1 VF	Coverage	Chromosome	Position	Reference	Mutation	Replicate 2 VF	Yes=2nd file within the range	Replicate 3 VF	Yes=3rd file within the range	*Yes in Replicates 2 & 3
0.0105	956	18	48603079	C	T					N/A
0.0131	915	18	48603083	G	A					N/A
0.0118	847	7	55221909	G	A					N/A
0.0543	847	7	55221915	T	G	4.65%	Y	5.34%	Y	Y
0.0201	845	7	55221903	C	T					N/A
0.0195	822	7	55221878	C	T					N/A
0.0122	817	20	36031726	C	T					N/A
0.1077	808	20	36031767	C	G	13.50%	Y	11.38%	Y	Y
0.0137	805	20	36031725	C	T					N/A
0.015	802	7	55221852	C	T					N/A
0.055	782	7	55221914	C	T	6.43%	Y	6.81%	Y	Y
(Data with coverage between 782 and 227 were hidden to condense the table)										
0.0529	227	22	24176357	C	G	4.39%	Y	#N/A	#N/A	#N/A
0.5286	227	4	55980239	C	T					N/A
0.0724	221	13	28608234	C	T	5.07%	Y	8.19%	Y	Y
0.1244	217	11	108236191	A	C	15.35%	Y	13.92%	Y	Y
0.3704	216	11	108172365	T	G					N/A
0.3519	216	7	128845112	C	A					N/A
0.0605	215	19	17945623	G	T	3.85%	Y	5.24%	Y	Y
0.514	214	4	153249361	T	C					N/A
0.1698	212	13	28608229	A	T	12.98%	Y	12.16%	Y	Y
0.0825	206	11	108173701	T	G	9.39%	Y	10.33%	Y	Y
0.099	202	13	28608233	T	C	5.43%	Y	6.86%	Y	Y
0.0714	196	2	29443612	C	T	#N/A	#N/A	12.75%	N	#N/A
0.228	193	3	10188301	T	G	23.05%	Y	24.41%	Y	Y
0.0702	171	13	48941615	C	A	11.73%	Y	8.28%	Y	Y
0.0625	160	17	7579466	G	C	#N/A	#N/A	12.93%	N	#N/A
0.0915	142	19	1220519	G	A	9.40%	Y	9.84%	Y	Y
0.1111	99	13	48941616	T	C	20.00%	N	13.13%	Y	N
*N/A= replicate 1 VF > 25% (0.25)										
N= one N and one Y, or two Ns for the two replicates										
#N/A= one or more #N/A were identified from the two replicates										

Table 8: Calculation of percentage repeatability using empirically derived equation and a custom-designed The Concordance Calculator at low stringency (Ion Torrent Intra Run data).

Replicate 1 VF	Coverage	Chromosome	Position	Reference	Mutation	Replicate 2 VF	Yes=2nd file within the range	Replicate 3 VF	Yes=3rd file within the range	*Yes in Replicates 2 & 3
0.0105	956	18	48603079	C	T					N/A
0.0131	915	18	48603083	G	A					N/A
0.0118	847	7	55221909	G	A					N/A
0.0543	847	7	55221915	T	G	4.65%	Y	5.34%	Y	Y
0.0201	845	7	55221903	C	T					N/A
0.0195	822	7	55221878	C	T					N/A
0.0122	817	20	36031726	C	T					N/A
0.1077	808	20	36031767	C	G	13.50%	Y	11.38%	Y	Y
0.0137	805	20	36031725	C	T					N/A
0.015	802	7	55221852	C	T					N/A
0.055	782	7	55221914	C	T	6.43%	Y	6.81%	Y	Y
Data with coverage between 782 and 322 were hidden to condense the table										
0.0373	322	19	17945621	G	A					N/A
0.0374	321	2	212652721	G	A					N/A
0.0596	319	19	1207074	G	A	#N/A	#N/A	4.22%	Y	#N/A
0.0314	318	4	55979610	G	A					N/A
0.0411	316	9	5073849	A	T					N/A
0.1206	315	19	1207082	A	G	9.32%	Y	7.24%	Y	Y
0.0351	313	7	116411973	C	T					N/A
0.0321	312	13	48942631	G	A					N/A
0.0452	310	15	90631878	G	A					N/A
0.0356	309	7	128851504	C	T					N/A
0.0552	308	7	128851501	A	G	3.28%	Y	4.55%	Y	Y
0.0525	305	7	128851502	G	A	2.70%	Y	4.86%	Y	Y
0.1711	304	19	1207084	G	A	13.95%	Y	8.07%	N	N
0.1419	303	20	57484567	C	T	12.63%	Y	12.00%	Y	Y
0.3709	302	4	153249358	A	T					N/A
0.0464	302	7	128851500	G	T					N/A
0.04	300	17	7579879	G	A					N/A
*N/A= replicate 1 VF > 25% (0.25)										
N= one N and one Y, or two Ns for the two replicates										
#N/A= one or more #N/A were identified from the two replicates										

Table 9: Effect of applying a more stringent filter to same data set as in Table 8 (Ion Torrent Intra Run data).

consistent calls across all 6 replicates the C to T bias still exist within Illumina platform likely due to the small random chances as those showed in Figure 2 represent only a very small fraction of all C to T calls. Additional data from the other 5 FFPE tissue samples used in this study also showed high C to T bias in Illumina platform yet no such bias in Ion Torrent can be found in the Supplemental Data (Table 11).

Discussion

As stated in the Introduction, the most important goals for this study were to determine the repeatability and reproducibility for the two NGS platforms, Illumina MiSeq and Ion Torrent Ion PGM, using FFPE tissue samples and commercially available cancer panels. Throughout the process of the study, many questions were raised:

1) Why is the recommended genomic DNA input for FFPE tissue is 25-fold higher for the Illumina cancer panel compared to the Ion Torrent

2) What is the impact of the different average amplicon sizes for the two cancer panels (184 bp for Illumina and 119 bp for Ion Torrent)

3) What is the impact of the fundamentally different library preparation methods (i.e., the "Amplify Now" and "Amplify Later" protocols, and one strand versus two strand copy strategies.

4) What is the impact of sequencing chemistry-specific quality scores on the repeatability/reproducibility of low frequency variant calls? These fundamental questions and issues will be discussed in detail in the following paragraphs.

Quality of data

Efforts to improve clinical assays are constantly moving toward higher sensitivity and higher accuracy assays. Higher throughput is rarely a major concern as clinical sample testing is usually done in smaller batches, especially for diagnostic purposes, since turn-around time is critical for patient stratification and/or treatment guidance. In a clinical trial setting, sample availability is always a concern and clinical objectives can be compromised by limiting sample. When only a few FFPE tissue slides are available for the study, whether or not it is possible to perform all the assays necessary to gather all important information depends on the sample input requirements. Tests with similar performance that require more starting material might be deprioritized from the long list of assays need to be done. Therefore, the recommended minimum input DNA/RNA, in the case of genomic assays, is likely going to be used as a starting condition for assay optimization and validation. Nevertheless, minimum sample quality and data QC resulting in meeting the minimum assay specifications is still required to achieving a valid test result. As shown in the Results section and based on the supplemental information provided in this manuscript all minimum QC criteria were met for the data sets generated.

Replicate 1 VF	Coverage	Chromosome	Position	Reference	Mutation	Replicate 2 VF	Yes=2nd file within the range	Replicate 3 VF	Yes=3rd file within the range	*Yes in Replicates 2 & 3
0.9992	1303	11	108218196	T	C					N/A
0.9982	5046	4	55141055	A	G					N/A
0.9981	2152	11	108225661	A	G					N/A
0.998	7074	4	55946081	A	G					N/A
0.9976	1698	13	28610183	A	G					N/A
0.9976	1695	17	7579472	G	C					N/A
0.9975	5146	2	132181460	T	C					N/A
0.9963	8151	17	7578115	T	C					N/A
0.9957	5146	2	132181483	G	T					N/A
0.9908	1305	4	1807894	G	A					N/A
0.9873	10988	5	112175770	G	A					N/A
0.7693	4834	10	89720907	T	G					N/A
0.6609	4064	7	55249063	G	A					N/A
0.535	1641	10	89717672	C	T					N/A
0.4695	4714	22	24145675	G	C					N/A
0.3613	1323	19	1223125	C	T					N/A
0.3569	10051	9	80409345	A	G					N/A
0.3494	5046	4	55141026	C	T					N/A
0.3268	1285	4	1803652	T	G					N/A
0.1	5081	5	112175974	G	A	9.43%	Y	4.86%	Y	Y
0.1429	5075	1	43815071	T	G	16.99%	Y	15.04%	Y	Y
0.1105	4867	2	29432781	C	T	7.66%	Y	#N/A	#N/A	#N/A
0.1381	4033	12	121431533	T	G	13.64%	Y	13.99%	Y	Y
0.1135	3877	4	55597594	C	T	#N/A	#N/A	#N/A	#N/A	#N/A
0.1029	3701	22	24176357	C	T	14.71%	Y	#N/A	#N/A	#N/A
0.13	2946	7	128850423	C	T	#N/A	#N/A	#N/A	#N/A	#N/A
0.1345	2759	17	7577114	C	T	#N/A	#N/A	#N/A	#N/A	#N/A
0.1265	2752	2	132181331	C	T	#N/A	#N/A	#N/A	#N/A	#N/A
0.168	2703	7	128850379	A	C	17.63%	Y	17.12%	Y	Y
0.1282	2542	12	112926959	C	T	#N/A	#N/A	#N/A	#N/A	#N/A
0.1009	2161	20	36031817	A	G	11.91%	Y	9.00%	Y	Y
0.1094	2149	20	36031797	A	C	11.46%	Y	10.62%	Y	Y
0.1072	2033	11	108204781	C	T	#N/A	#N/A	#N/A	#N/A	#N/A
0.1218	1995	7	55211154	A	T	#N/A	#N/A	#N/A	#N/A	#N/A
0.2007	1963	20	36031830	T	G	21.28%	Y	23.54%	Y	Y
*N/A= replicate 1 VF > 25% (0.25)										
N= one N and one Y, or two Ns for the two replicates										
#N/A= one or more #N/A were identified from the two replicates										

Table 10: Effect of applying a more stringent filter to the Illumina Intra Run data set.

Sample ID	Chromosome	Position	Reference	Mutation	Variant Frequency	Coverage	Platform	Sequnome Variant Call	Sequnome VF
CRC 80	4	55152040	C	T	0.143	5138	Miseq Ion Torrent	T	0.35
	4	55152040	C	T	0.306	421			
CRC 83	10	43613843	G	T	0.090	222	Miseq Ion Torrent	T	0.42
	10	43613843	G	T	0.239	654			
CRC 86	22	24134064	C	T	0.087	2705	Miseq Ion Torrent	T	0.05
	22	24134064	C	A	0.074	558			
CRC 88	5	112175571	C	A	0.0866	3073	Miseq Ion Torrent	T	0.07
	5	112175571	C	T	0.0939	181			
CRC 89	17	7578253	C	A	0.1201	6204	Miseq Ion Torrent	A	0.35
	17	7578253	C	A	0.3196	388			
CRC 89	18	48581234	C	T	0.1706	3604	Miseq Ion Torrent	T	0.15
	18	48581234	C	T	0.0823	486			
CRC 94	10	89717672	C	T	0.897	1563	Miseq Ion Torrent	T	0.61
	10	89717672	C	T	0.5513	390			

Table 11: List of discordant calls with tie-breaker Sequenom assay results.

Illumina and Ion Torrent chemistry

Although both Illumina and Ion Torrent sequencing belong to the category of "sequencing-by-synthesis" technology, the biochemistry methodologies involved in the two technologies are quite different. Illumina utilizes the reversible-dye-terminator chemistry and provides all four nucleotides at the same time during each cycle of synthesis. It relies on the % yield (or % completion) of de-allylation for all labelled dyes for each nucleotide and the de-allylation of 3' OH protecting group [23]. Assuming each step of de-allylation is 99% completion, after 100 cycles only approximately 36% of the initial primers will have the targeted sequences. Since certain dyes linked to the nucleotides may result in different de-allylation rates (such as 98% vs. 99%), this results in sequence-dependent yield and a phasing phenomenon called T-accumulation, which also may interfere with the fluorescent detection. According to a recent publication, after 50 synthesis cycles, 59.5% of the fluorophores reflect the current cycle, 17.4% are exactly one cycle behind, 17.4% one cycle ahead and 33.9% of the measured cluster intensity is caused by T-accumulation [24]. Chemistry specific algorithms were developed to correct such phasing and pre-phasing, however, any small deviations that could not be accurately corrected by the algorithm will result in the inaccurate detection of base calls. This process is very similar to the oligonucleotide synthesis chemistry for which the longer the oligonucleotide the lower the percentage of the final product will be for the full-length target sequence, which creates the need for gel electrophoresis or HPLC to purify the product containing intended sequence. For many long oligonucleotides the final yields are less than 10% after all the synthesis cycles are completed. Many review articles have discussed these chemistry and instrumentation fluorescent detection issues in great detail [25,26].

Ion Torrent utilizes the Ion conductor chemistry which detects the release of proton after each base is added to the extending strand of DNA. The proton release causes a p^H change that is detected by the millions of pH meters at the bottom of each well inside of the chip that will result in the voltage change and therefore convert into electronic signal for the detector [27]. One major difference between the Ion Torrent chemistry and Illumina chemistry is that Ion Torrent's sequencing-by-synthesis approach involves the incorporation of one dNTP at a time and detects all incorporation of same nucleotide no matter how long the stretch for that particular nucleotide is. That means if you have 10 Ts in a row the synthesis will incorporate 10 Ts to the extending strand and release 10 protons and therefore proportionally the pH changes. However, it is challenging to be able to resolve these multiple proton releases and accurately convert them into correct number of base calls. Ion Torrent claims that their algorithm is now capable of resolving 7-8 consecutive same nucleotide base calling [28]; unfortunately imperfect conditions will throw these algorithms off and issues with inaccurate calls of lower or higher number of consecutive bases could be encountered.

In our study, we likely observed all of these above discussed issues from both platforms after looking at many of the low frequency rare and/or unknown variants. Here, we propose that after all the recommended filters are used, all variants that are in the final variant call list should be individually inspected for their flanking sequences and potential non-specific variant calls as well as deamination or oxidation artifacts.

The impact of Q scores

The original Q scores came from Phred base-calling algorithm for slab gel-based Sanger sequencing [29]. Phred algorithm was based on the analysis and training of a large set of accurate electropherograms of Sanger sequencing data by calculating parameters related to the peak shape, peak height, and peak resolution for each base. Applied Biosystems adapted this algorithm and further improved base calling accuracy for their capillary electrophoresis with great success [29]. Both Illumina and Ion Torrent developed their own Q score algorithms since Q scores are chemistry specific and NGS data outputs do not generate color trace files like that used for developing the original Phred algorithm. According to the Illumina website, the Q scores for Illumina were developed using parameters relevant to a particular sequencing chemistry and analyzed for a large empirical data set of known accuracy. Since Ion Torrent ion conductor chemistry is completely different from Illumina's reversible dye-terminator chemistry the definition of Q score is quite different, and therefore there is no direct correlation between the Illumina and Ion Torrent generated Q scores. One more important distinction between Q scores for Sanger sequencing and Q scores for NGS is that Sanger Q scores are for each final base call while NGS Q scores are for each base read. Since a final base call for NGS consists multiple base reads (often thousands of reads for one base call), we need to treat the Q scores for NGS differently apart from the conventional concept of the relationship of base calling accuracy and the Q scores for that base. As a result of this distinction, the impact of Q scores for the accuracy of that particular variant allele base call is only loosely correlated. However, analysis ascertains minimum Q scores and includes the average of multiple reads could be an advantage over traditional Phred values. One major factor that impacts the accuracy of the variant allele base call is the total variant hits, and our recommendation is to have minimum of 10 variant hits for reliable variant calls. Each hit carries a Q score with Q13 (the default setting of OmicSoft) more than enough to be confident of that particular read. However, there are many other factors that could cause a read not to be reliable. These include fine crystals or dust and lint particles that block or reflect the fluorescent reading on the flow cell [24]. Therefore, multiple variant hit is critical to have high confidence for the reproducibility or accuracy of that particular variant call.

Source of variation and the list of critical factors that impact the variant calls

In order to understand the source of variation, one important question to ask is why the recommended input genomic DNA (gDNA) amount is 25-fold different between Illumina's 250 ng and Ion Torrent's 10 ng. Based on our prior experience developing Single Nucleotide Primer Extension (SNPE) multiplexing mutation assay for detecting 33 KRAS/BRAF/NRAS hotspot mutations [30], one likely reason is the amplicon size. The average amplicon size for Illumina Cancer Panel is 184 bp and for Ion Torrent Cancer Panel is 119 bp. In the above mentioned *SNPE* mutation detection study, the original amplicon sizes of 150 to 200 bp were designed for *KRAS* exon 2 and 3 to cover hotspots located in various codons (Chang et. al., unpublished study), while amplicon sizes designed for *NRAS* and *BRAF* were in the range of 100-110 bp. When intact human healthy volunteer blood gDNA was used, high quality data were generated with clear bands identified using Bioanalyzer from the PCR amplification reactions. Once the study was moved to using FFPE tissue gDNA (using 15 to 20 ng), several PCR amplicons from *KRAS* design failed to generate visible PCR amplified products while all amplicons from *NRAS* and *BRAF* were successfully amplified. Since the average FFPE tissue gDNA fragment sizes were around 200 bp (Chang et.al. unpublished study), few gDNA fragments containing complete target sequences for *KRAS* were available for amplification. Based on this result, the estimated copy number of amplifiable gDNA was likely to be approximately 2-300 for Illumina using 250 ng input gDNA. Similar number of amplifiable copies

was probably available for the Ion Torrent Cancer Panel using 10 ng of input gDNA due to the smaller average amplicon size (Figure 2). However, the Illumina TruSeq library preparation protocol starts with hybridization of primers on the original input gDNA followed by stringent washing of unbound primers and the extension-ligation procedures of 2-300 copies of amplifiable gDNA, which understandably is likely going to introduce high variation for low frequency variants. Once the manipulations are done and variations introduced, additional PCR amplification (defined as "Amplify Later" protocol) will not change that pool of distribution (Figure 1). In contrast, the Ion Torrent AmpliSeq protocol starts with 20 cycles of PCR amplification (defined as "Amplify Now" protocol) using target specific primer pairs to increase the couple hundred copies of amplifiable gDNA to at least several million copies of PCR products. Although specificity could potentially be affected by the initial amplification (depending on how many cycles and how specific the primer sets are used in the protocol), copy number of low frequency variants are likely going to be proportionally increased and therefore variation will be reduced (compared to working on a limited copies of initial input DNA).

For run-to-run variation it is understandable that Illumina should have lower variability among replicates since much less hands-on procedures are involved with standard cassette-styled reagent packs plug-and-play walk-away operation after library preparations are done. In contrast, the Ion PGM involves a number of potentially variable steps: One Touch emulsion PCR and Enrichment Station, sample chip loading (especially for 314 chips), preparation of buffers with narrow pH range, and preparation of individual nucleotide dilutions for initial set up before each run. These steps likely contribute to the higher run-to-run variation observed in our data sets, e.g., for those data generated from same sample and same library preparation, but different runs. Recent introduction of "Ion Chef" robotic workstation could potentially reduce Ion Torrent's instrument run-to-run variation [31].

High frequency variant calls vs. low frequency variant calls (sequence specificity of DNA polymerase)

According to our data, high frequency variants are much more repeatable or repeatable than their low frequency counterparts. This is an important observation since many clinical NGS data analysis practices use one pre-set fold-coverage, such as 1000x coverage as a cutoff, in an effort to sort out mutations or variants that are not repeatable or reproducible. Our data suggest that a more progressive approach should be implemented to reduce false negative rate, as in order to reduce the false positive rate, the false negative rate is always compromised. The best way to minimize these compromises is to establish an equation based on your training data set to establish each variant frequency bracket as described in the result section. Not surprisingly, hot spot mutations and well characterized SNPs in our data sets are much more repeatable or reproducible than those poorly characterized rare or unknown mutations. That might be the reason why hotspot mutations became hotspot in the first place since they kept on showing up reproducibly. According to our previous experience developing and validating different mutation assays and based on this current NGS data set, both detection sensitivity, repeatability and reproducibility for each mutation is different [30]. Therefore, validating a subset of mutations may not justify for the validity of the rest of mutations. Especially validating high frequency mutations and assuming the same confirmation rate can be applied to the rest of un-validated low frequency mutations is unjustified and is a fundamentally flawed approach. This issue might have something to do with the sequence specificity of the DNA polymerase. Previous publications showed that DNA polymerase paused or stalled reproducibly at many specific nucleotide sequence locations when replication was slowed down using lower temperature, but then continued to proceed further to completion with time [32]. Those specifically stalled nucleotide locations during DNA polymerase replication might have higher error rate and could be the source of where the original mutations came from. Alternatively, during DNA sequencing reads, if specific sequence locations cause difficulty for the DNA polymerase to read through, a higher rate of mistakes could occur. It is a challenge to distinguish these two processes since they both resulted in the same variant calls. When estimating fidelity for DNA polymerase, an average error rate was derived based on large number of sequencing reads mainly on the normal sequences. When millions of reads are done through NGS, those small percentage of mistakes will always be captured through sorting out a list of variant calls, and if those mistakes are with low coverage and low variant hits they could become false positives and could be included in the final selection of variants. Of course if these mistakes are random the impact of the final base calling is limited. However, if these mistakes are sequence specific and only specific locations have higher error rate, then false positive variant calls will be obtained. One way to minimize these issues is to only select variants with higher number of total variant hits (such as 10 hits or higher).

The impact of low confirmation rate and high variation on low frequency variant calls on the interpretation of tumor heterogeneity

As discussed in the previous paragraph, high frequency variants are much more repeatable or reproducible than their low frequency counterparts. Unfortunately only Sanger sequencing is currently considered as "Gold Standard" for NGS mutation call confirmation. Since the detection sensitivity for Sanger sequencing is between 15 and 20% those selected mutation calls subjected to the Sanger sequencing confirmation are likely to have mutation frequencies higher than 25% to avoid controversial results. If a set of somatic mutations with frequencies between 25 and 50% were used for Sanger sequencing confirmation and the confirmation rate is 90% one should not assume that those low frequency mutations with mutation frequencies between 1 and 25% will have similar confirmation rate. As shown in our Illumina data set using replicates from same sample different library preparations, many variant calls with extremely high quality (high coverage or high Q score) were not repeatable. Of course, if a different library preparation protocol were used (such as "Amplify Now" protocol defined earlier such as AmpliSeq) the degree of variation might not be so high. Nevertheless, any such low frequency mutations, if not directly validated, should not be used for interpreting important biological effects, such as degree of tumor heterogeneity [33], as the same sample, different library preparations of technical replicates could potentially generate similar pattern of mutations including so-called "private mutations" and "shared mutations" as described in the publication from those different regions of the tumor tissue. Selected samples with a list of "private mutations" could be easily verified using same sample and two other different library preparations. "Not enough sample" should not be used as a reason to justify not doing such important repeatability/reproducibility study if the data are used to justify novel conclusions.

The impact of library preparation protocols on the "Post Tissue Collection Modifications" (PTCM)

As mentioned earlier in the Discussion section, copying from one

strand is quite different than copying from both strands during the initial steps. If DNA sequence changes are generated after the tissues are collected, such as deamination caused by FFPE tissue preparation process or oxidation caused by mechanical fragmentation [1], the initial steps of copying from one strand, such as Illumina's TruSeq Cancer Panel protocol, will result in detecting approximately the same percentage frequency for the impacted post tissue collection modifications (PTCM) at the end of the sequencing process. However, copying from both strands during the initial steps (such as the PCR amplification at the beginning of the AmpliSeq Cancer Panel library preparation procedure) will dilute the post tissue collection modification variant allele frequency into half, since the sequence information on the opposite strand will remain the same throughout the amplification as those from pre-tissue collection stage. With this fundamental difference, FFPE tissue sample libraries prepared through the TruSeq protocol will likely result in two-fold higher frequency of each deamination than those processed through AmpliSeq library preparation protocol. This issue could be more profound with FFPE tissue slides pre-sectioned and stored at the room temperature for long periods of time as those FFPE tissue samples received from clinical sites. Higher C to T variant call frequencies could be derived through the TruSeq Cancer Panel protocol if input FFPE tissue gDNA is limited and highly fragmented (average fragment size less than 200 bp) as explained earlier in the Discussion section. As observed in our Illumina data set using same sample with different library preparation, several C to T mutations with high quality reads and high coverage (greater than 1500x) and with as high as 15% variant frequencies were identified to be not repeatable among the replicates. In order to minimize this problem for the mutation profiling study, we recommend using what we call the "Amplify Now" library preparation protocol to detect unknown variant calls with low variant allele frequency.

The impact of artifact C to T and C to A mutations on mutation profiling and mutation signature identification

As shown in Figure 2, very high percentage of C to T variant calls were found in the Illumina data sets, but which was not obvious in Ion Torrent data sets. This C to T bias is likely due to the deamination from FFPE sample preparation and the so-called copy from one strand Illumina TruSeq protocol (Figure 1). Whether T-accumulation contributed to this high C to T variant calls is unclear. Interestingly, a recent publication from Broad Institute showed high C to A variant calls was associated with the degree of mechanical fragmentation

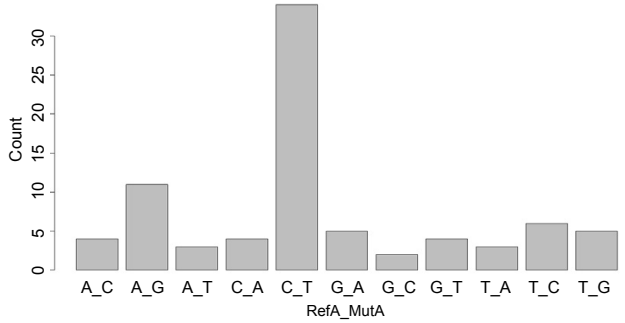

Figure 2A: High C to T variant calls in Illumina data but not in Ion Torrent data (a) Distribution of the unique reference allele_mutation allele (RefA_MutA) combinations in the mutations found by Illumina TruSeq Cancer Panel across all 6 library preparation replicates for FFPE tissue sample ID 94.

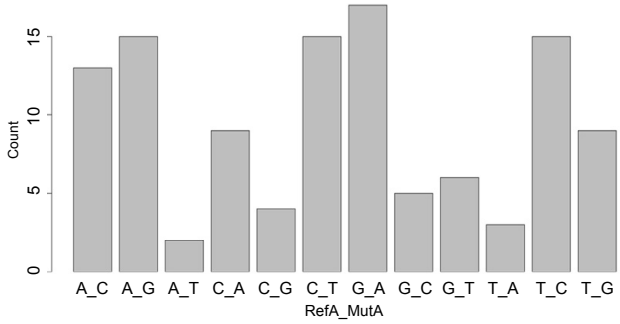

Figure 2B: Distribution of the unique reference allele_mutation allele combinations in the mutations found by IonTorrent AmpliSeq Cancer Panel across all 6 library preparation replicates for FFPE tissue sample ID 94. Variant calls with VF less than 1% were not included in this analysis.

during the library preparation [1]. The fact that they did not find any C to T bias and we did not find any C to A bias indicates that C to T bias is mainly FFPE tissue related (no DNA fragmentation step required) and C to A bias is mainly coming from tissues with intact DNA that require fragmentation procedure. Since both of these biased variant calls are PTCM (post tissue collection modification) errors with relatively low variant allele frequency (mainly below 10%) it raises a concern regarding mutation signature identification, especially for those large numbers of mutation involved in the signature data set where individual mutation validation is impossible. A recent publication titled "Signatures of Mutational Processes in Human Cancer" [13] showed among 21 mutation signatures identified C to T and C to A mutations appeared to be dominated across most signatures. Although reasonable interpretations based on the existing knowledge were provided to explain why some mutation signatures among specific cancers might have higher than usual C to T mutations or C to A mutations, it is not clear how many of these C to T and C to A mutations could be due to the FFPE tissue and tumors with intact DNA (such as blood DNA, or fresh frozen tissue DNA) used in deriving that particular signature set. Nevertheless, any future mutation signatures identified in data sets with high C to T mutations or C to A mutations should be carefully examined for evidence of these types of artifacts. One approach to confirm these mutations is to select some representative sample sets and perform replicates using different library preparations and see if identical mutations are obtained, since these randomly modified PTCM variant calls are likely not going to be repeatable from library preparation to library preparation as shown in Table 10.

Conclusion

In this Discussion, many specific recommendations were made to address the issues we encountered throughout this NGS platform evaluation and validation with a focus on the future application of these platforms for mutation profiling in clinical studies using FFPE tissues. A recommendation of how to circumvent the challenges was provided for each issue discussed. One unique feature of mutation profiling studies compared to other clinical NGS applications is the balance of false positives and false negatives as the value of mutation profiling is to include as many true mutations as possible without including too many false positives. Low allele frequency somatic mutation detection (5-25%) represents the most valuable information for mutation profiling studies. Since most of the detected somatic mutations have mutation frequencies of 50% or lower and the average percentage of tumor

cells in the clinical FFPE tissue slides generally below 50%, if macro-dissection procedures were not applied most of these valuable somatic mutations is likely going to be below 25%. Furthermore, based on the factors discussed, reliable mutation detection below 5% is a major challenge based on the limited input of gDNA from highly fragmented FFPE tissue samples. Therefore, the most impactful frequency range of somatic mutation detection for mutation profiling effort is likely between 5 and 25%. With these in mind, we recommend using a library preparation protocol taking the "Amplify Now" approach, and if possible a triplicate library preparation approach should be considered. After the data passed the minimum QC, a progressive approach, including a combination of sequencing coverage and variant frequency, should be used to identify a list of high confidence mutations. This is to sort out a list of repeatable mutations with a good balance between low false positive and low false negative rates depending on which platform is used. The list of high confidence mutations should then be analyzed for their flanking sequences and strand biases to further eliminate miss-alignment errors and non-specific mutations including PTCM mutation artifacts.

Acknowledgment

The authors would like to thank Dr. Helen Fernandez of Cornell Medical College for helpful discussions, Jason Hughes of Merck Research Labs in helping to set up data processing pipeline, and John Thompson of Merck Research Labs for training on OmicSoft applications.

References

1. Costello M, Pugh TJ, Fennell TJ, Stewart C, Lichtenstein L, et al. (2013) Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. Nucleic Acids Res 41: e67.
2. O'Rawe J, Jiang T, Sun G, Wu Y, Wang W, et al. (2013) Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. Genome Med 5: 28.
3. Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, et al. (2013) Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature 499: 214-218.
4. Roberts RJ, Carneiro MO, Schatz MC (2013) The advantages of SMRT sequencing. Genome Biol 14: 405.
5. Jung H, Bleazard T, Lee J, Hong D (2013) Systematic investigation of cancer-associated somatic point mutations in SNP databases. Nature Biotech 31: 787–789.
6. Genovese G, Handsaker RE, Li H, Altemose N, Lindgren AM, et al. (2013) Using population admixture to help complete maps of the human genome. Nature Genet 45: 406–414.
7. Ratan A, Miller W, Guillory J, Stinson J, Seshagiri S, et al. (2013) Comparison of sequencing platforms for single nucleotide variant calls in a human sample. PLoS ONE 8: 55089.
8. Meynert AM, Bicknell LS, Hurles ME, Jackson AP, Taylor MS (2013) Quantifying single nucleotide variant detection sensitivity in exome sequencing. BMC Bioinformatics 14: 195.
9. Robins WP, Faruque SM, Mekalanos JJ (2013) Coupling mutagenesis and parallel deep sequencing to probe essential residues in a genome or gene. Proc Natl Acad Sci USA 110: 848–857.
10. Barrick JE, Lenski RE (2013) Genome dynamics during experimental evolution. Nature Rev Genet 14: 827–839.
11. Kildegaard HF, Baycin-Hizal D, Lewis NE, Betenbaugh MJ (2013) The emerging CHO systems biology era: harnessing the 'omics revolution for biotechnology. Curr Opin Biotechnol 24: 1102–1107.
12. Gonzalez-Perez A, Mustonen V, Reva B, Ritchie GR, Creixell P, et al. (2013) Computational approaches to identify functional genetic variants in cancer genomes. Nature Methods 10: 723–729.
13. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SA, Behjati S, et al (2013) Signatures of mutational processes in human cancer. Nature 500: 415-421.
14. Kilpivaara O, Aaltonen LA (2013) Diagnostic Cancer Genome Sequencing and the Contribution of Germline Variants. Science 339: 1559-1562.
15. http://res.illumina.com/documents/products/datasheets/datasheet_truseq_amplicon_cancer_panel.pdf
16. http://www.lifetechnologies.com/order/catalog/product/4475346
17. http://supportres.illumina.com/documents/myillumina/02fe2a31-7867-495f-9783-de30d3ccc919/truseq_amplicon_cancer_panel_guide_15031875_a.pdf
18. http://www3.appliedbiosystems.com/cms/groups/applied_markets_marketing/documents/generaldocuments/cms_098568.pdf
19. http://supportres.illumina.com/documents/documentation/system_documentation/miseq/miseq-system-user-guide-15027617-l.pdf
20. http://download.bioon.com.cn/view/upload/201312/14155016_8616.pdf
21. http://bioscience.sequenom.com/sites/bioscience.sequenom.com/files/iPLEX%20Brochure%20(1).pdf
22. http://www.ascopost.com/issues/december-1,-2013/tp53-status-may-predict-benefit-from-cetuximab-in-locally-advanced-rectal-cancer.aspx
23. Ju J, Kim DH, Bi L, Meng Q, Bai X, et al. (2006) Four-color DNA sequencing by synthesis using cleavable fluorescent nucleotide reversible terminators. Proc Natl Acad Sci USA 103: 19635-19640.
24. Kircher M, Stenzel U, Kelso J (2009) Improved base calling for the Illumina Genome Analyzer using machine learning strategies. Genome Biology 10: 83.
25. Nakamura K, Oshima T, Morimoto T, Ikeda S, Yoshikawa H, et al (2011) Sequence-specific error profile of Illumina sequencers. Nucleic Acids Res 39: 90.
26. Loman NJ, Misra RV, Dallman TJ, Constantinidou C, Gharbia SE, et al (2012) Performance comparison of benchtop high-throughput sequencing platforms. Nat Biotechnol 30: 434-439.
27. Quail M, Smith ME, Coupland P, Otto TD, Harris SR, et al (2012) A tale of three next generation sequencing platforms: comparison of Ion torrent, pacific biosciences and illumine MiSeq sequencers. BMC Genomics 13: 341.
28. http://mendel.iontorrent.com/ion-docs/Technical-Note-Base-Recalibration_33325761.html
29. http://en.wikipedia.org/wiki/Phred_quality_score
30. Chang KCN, Galuska S, Weiner R, Marton MJ (2013) Development and validation of a clinical trial patient stratification assay that interrogates 27 mutation sites in MAPK pathway genes. PLOS ONE 8: 72239.
31. http://www.lifetechnologies.com/us/en/home/life-science/sequencing/next-generation-sequencing/ion-torrent-next-generation-sequencing-workflow/prepare-template/ion-torrent-next-generation-sequencing-ion-chef-system.html
32. Takeshita M., Chang CN, Johnson F, Will S, Grollman AP (1987) Oligodeoxynucleotides containing synthetic abasic sites: Model substrates for DNA polymerases and AP endonucleases. J Biol Chem 262: 10171-10179.
33. Gerlinger M, Rowan AJ, Horswell S, Larkin J, Endesfelder D, et al (2012) Intratumor Heterogeneity and Branched Evolution Revealed by Multiregion Sequencing. New England J Med 366: 883-892