Published online 15 May 2014 *Nucleic Acids Research, 2014, Vol. 42, No. 12* ***e101***
doi: 10.1093/nar/gku392

Performance comparison of SNP detection tools with illumina exome sequencing data—an assessment using both family pedigree information and sample-matched SNP array data

Ming Yi[1,2,*], Yongmei Zhao[1], Li Jia[1], Mei He[3], Electron Kebebew[3] and Robert M. Stephens[1,2,*]

[1]Advanced Biomedical Computing Center, SAIC-Frederick, Inc., Frederick National Laboratory for Cancer Research, Frederick, MD 21702, USA, [2]Current address: Cancer Research and Technology Program, Frederick National Laboratory for Cancer Research, Leidos Biomedical Research, Inc. PO Box B, Frederick, MD, 21702. and [3]Endocrine Oncology Branch, Center for Cancer Research, National Cancer Institute, Bethesda, MD 20892, USA

Received December 6, 2013; Revised March 27, 2014; Accepted April 22, 2014

ABSTRACT

To apply exome-seq-derived variants in the clinical setting, there is an urgent need to identify the best variant caller(s) from a large collection of available options. We have used an Illumina exome-seq dataset as a benchmark, with two validation scenarios—family pedigree information and SNP array data for the same samples, permitting global high-throughput cross-validation, to evaluate the quality of SNP calls derived from several popular variant discovery tools from both the open-source and commercial communities using a set of designated quality metrics. To the best of our knowledge, this is the first large-scale performance comparison of exome-seq variant discovery tools using high-throughput validation with both Mendelian inheritance checking and SNP array data, which allows us to gain insights into the accuracy of SNP calling through such high-throughput validation in an unprecedented way, whereas the previously reported comparison studies have only assessed concordance of these tools without directly assessing the quality of the derived SNPs. More importantly, the main purpose of our study was to establish a reusable procedure that applies high-throughput validation to compare the quality of SNP discovery tools with a focus on exome-seq, which can be used to compare any forthcoming tool(s) of interest.

INTRODUCTION

As observed in the past decade, it has been a great challenge to globally map the enormous number of human genetic variations (1–3) to direct relationships between genotype and phenotype identifying the causal variants among thousands of candidates. Evident from many landmark results already made by their application (4–6), the emergence of Next-generation sequencing (NGS) methods has provided an opportunity to enable large-scale sequencing projects ranging from a global study characterizing genetic diversity on a population level (3) to a specific clinical application of identifying a causal variant in a single patient as a definitive diagnosis, with a great potential to guide therapy (6–11). A combination of factors has resulted in the fact that most of these studies have focused on exome sequencing (exome-seq) technology (6,7,9,11).

It has been shown that both the target enrichment strategy (12,13) and the NGS platform used with exome-seq technologies (4,14) have a large impact on the quality and coverage efficiency of sequencing reads and the resulting variant calls. Similarly, the alignment method used impacts the results obtained (15) with the most popular tools being MAQ (16) and BWA (17) as well as the CASAVA package for the Illumina platform (15). There continues to be an emergence of many new choices of alignment tools (18). Once the reads are aligned and based on a unified approach relying on Bayesian posterior probabilities that can be calculated for each potential genotype (19,20), two popular NGS variant-calling tools have emerged from the academic community: SAMtools (21) and the Genome Analysis Toolkit or GATK (22,23). Other tools have been developed to exploit aspects of specific types of NGS technologies or platforms including Pyrobayes for data from 454 platform (24), SOAP-

[*]To whom correspondence should be addressed. Tel: 301-836-5787; Fax: 301-846-5762; Email: stephensr@mail.nih.gov
Correspondence may also be addressed to Ming Yi. Tel: 301-846-5764; Fax: 301-846-5762; Email: yiming@mail.nih.gov

Published by Oxford University Press on behalf of Nucleic Acids Research 2014.

snp (25) for Illumina platform in addition to GATK and SAMtools, SAMtools-based Germ Line Variant Calling Plugin for Ion Torrent platform (http://www.edgebio.com/variant-calling-ion-torrent-data), and coming new variant-calling tool for both Ion Torrent and 454 sequencing data (26) and Quiver and modified GATK procedure for Pacbio platform data (http://www.pacificbiosciences.com/products/software/algorithms/, 27). In addition, vendor and commercial products have also joined the pack including Illumina's CASAVA (http://www.illumina.com/), CLCbio Genomics Bench (http://www.clcbio.com), Partek genomic suite and Partek Flow (http://www.partek.com). NGS-based single nucleotide polymorphism (SNP) identification tools have been compared in general for their features and strengths (28) and there are many others available from the open-source community for users to choose from (http://seqanswers.com/wiki/Software). In our assessment, while general guidance and descriptions of NGS SNP calling tools have been provided to the community, the question of which one to choose from a large collection (29) and how to apply it remains an open question.

An earlier performance comparison of whole exome sequencing (WES) technologies (14) and whole genome sequencing (WGS) platforms (30), which assessed the impacts of these platforms on the quality of variant detection, either applied SNP array data from matched samples of an exome-seq dataset (14), or used a series of evaluation criteria (30). Others have applied Mendelian inheritance error checking (MIEC) to assess the quality of variant calls (31). These studies mainly highlighted sequencing platforms and/or mapping-based differences but did not focus on the variant-calling tools themselves. A recent survey of variant analysis tools did report general comparison and concordance for these tools (32). In addition, a very recent study (33) comprehensively evaluated the variant-calling tools themselves and reported low concordance of multiple variant-calling pipelines in variant calling, which demonstrated fundamental methodological variation between these commonly used pipelines. However, unfortunately, probably due to lack of a full truth set of variants for validation, this study did not provide any assessment of the tools in the context of of the quality of the detected variants (33). This type of comparison is more needed by the community in order to make the right choice of the variant discovery tools. In general, the comparisons focused on the differences observed between methods rather than on the method producing the highest accuracy relative to the validation information. Much of these previous reports and comparison studies neither directly and comprehensively evaluated the performance of variant-calling tools themselves, nor tried to establish reusable and reliable comparison metrics and strategies that can be applied to any forthcoming or existing tools.

As exome-seq has become the most cost-effective NGS technology and Illumina exome-seq has emerged as the most popular platform that has been widely used in the NGS community for academic and clinical applications, in this study, we used a benchmark dataset that is composed of Illumina exome-seq data from family members and a SNP array dataset from the same family members. Thus, in our study, we sought to focus on the most popular exome-seq platform and also hold other processing elements such as mapping components constant so that the variant-calling tool evaluation would be the most informative. The SNP array has been widely used in many early genome-wide association studies and is well known to be able to detect the variants or single nucleotide polymorphisms in a population with very high accuracy primarily at well-annotated genomic locations based on numerous previous studies, which prompted us to use the SNP array data of same samples as a way of high-throughput validation on variants derived from NGS data. Our selected dataset allows us to evaluate the quality of single nucleotide variant (SNV) calls for selected SNP discovery tools through both MIEC across family members and validation using the SNP array data in a high-throughput fashion.

Due to the rapidly evolving nature of SNP detection tools and the relatively small fraction of these tools that can be effectively evaluated in a single study, the ranking of our selected tools is intended to demonstrate a framework under which different tools can be compared and evaluated. Our scan of the tools represents only a snapshot in time of each tools performance and by no means is static as all of these tools continue to improve over time. Although there are many other platforms such as WGS and many other vendors, by focusing on the most popular exome-seq platform from Illumina, we can concentrate on comparing the variant detection tools for the same source of the most common type of NGS data. Therefore, although the main purpose of this study is to report a method that establishes a reusable and replicable procedure for comparison of existing or any forthcoming SNP discovery tools with Illumina exome-seq data, such an approach and strategy could be easily applied to compare variant-calling tools for other platforms and/or other types of NGS data.

MATERIALS AND METHODS

Sample collection

The benchmark exome-seq data was generated from germline DNA extracted from lymphocytes of 19 human whole blood samples from two families with known pedigrees. Briefly, genomic DNA was isolated by density gradient centrifugation at 400 × g for 25 min using the lymphocyte separation media and was isolated using a PAXgene Blood DNA Kit (A Qiagen/BD Company Cat. No 761133). After quality control with Agilent Nano kit, the library was hybridized to biotinylated cRNA oligonucleotide baits from the SureSelect Human All Exon 50MB kit (Agilent, CA) and paired-end (108 × 108 bp) sequencing was performed using the Illumina Genome Analyzer IIx (Illumina Inc., San Diego, CA).

SNP array

DNA derived from the same blood samples used for exome-seq was prepared using Qiagen preparation kit to be run on the Illumina Human Omni-Quad BeadChip (Illumina Inc., San Diego, CA). SNP genotype calls were generated using the Genome Studio program from Illumina with default settings on Gencall at a threshold of 0.15.

WES

Ten micrograms of genomic DNA isolated from blood samples were used. One hundred and eight base pair paired-end reads were generated using the Illumina Genome Analyzer IIx (Illumina Inc., San Diego, CA). Illumina sequence reads were mapped to the human reference genome (hg19, NCBI37) with Illumina Eland (Elandv2) method or Burrows-Wheeler Alignment (BWA) method at their default settings.

Agilent SureSelect target enrichment and sequencing

The Agilent SureSelect Human All Exon 50Mb kit was used and target interval file was downloaded from Agilent eArray service (Agilent, CA). Paired-end Illumina libraries were captured in solution according to the Agilent SureSelect protocol. One hundred and eight base pair paired-end reads were generated using the Illumina GAIIx sequencing platform.

SNV detection

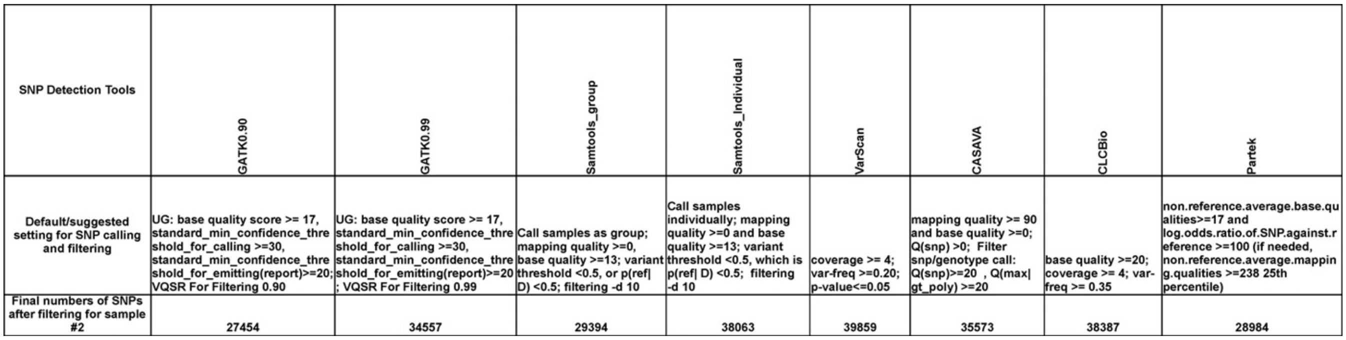
For all NGS data, SNVs were detected using the selected tools at their default settings as much as possible and all filtering procedures used were based on suggestions from tool providers or following default settings to the best of our knowledge (summarized in Table 1). Both default thresholds at 0.99 (annotated as GATK0.99) and the more stringent threshold at 0.90 (annotated as GATK 0.90) for variant quality score recalibration (VQSR) steps were used for the GATK procedure. Each tool was run using their basic and default implemented features for SNP calling without mixing with features or modules provided by any other tool. Especially for the GATK pipeline (23), although many of its phase I procedures including local realignment around indels and base quality score recalibration that are not directly involved in SNP calling but are all part of the GATK pipeline, they are included in the GATK SNP calling procedure as a default. Since BWA was suggested by GATK development group as the favored mapper for Illumina exome-seq data, to avoid comparison bias towards the GATK tool, we chose to use Eland as the major mapper to generate mapped data for SNP detection of all selected SNP detection tools. As a complementary approach, BWA mapped data in combination with Eland mapped data was indeed used to assess the impact on quality of derived SNP variant (SNV) by different versions of GATK (v1 versus v2), sample size, reference contents (with or without chrUn contigs) and mappers (BWA versus Eland) (Table 3; Table 5; Supplementary Tables S8 and S9). To provide a 'fair' comparison, the versions of each tool were fixed approximately at the same time period of the study, except for the new version of GATK (v2.0 or above, prior to v2–5.2) that was later to show the significant improvement by HyplotypeCaller (HTC) within the more recent new version of GATK (version 2.6–4, see below and Results section). GATK: different versions were used due to its dynamic evolving nature including the new version (V2.0 up to V2.2.4, V2.6–4 for improved HTC) or old version (up to V1.6.7); VarScan (V2.0); CASAVA (v1.8.2); CLCBio (v4.9). It should be noted that since the new raw SNP caller HTC was available in GATK v2.0 or above as the Unified genotyper counterpart has still undergone dynamic evolving with many experimental features and often encounters much longer run times (up to 5–20 times longer) compared to the Unified genotyper (http://gatkforums.broadinstitute.org/) and other technical issues. Recently the performance has been largely improved, our variants derived from GATK either old or new versions are all using the Unified genotyper as the raw caller before being subjected to VQSR module, although our very recent data indeed showed that with the help of VQSR module, Unified genotyper and HTC achieved a similar quality of SNVs for the final call sets until version 2.6–4 (compared to v2–5.2 and older versions), where HTC's performance has been significantly improved (http://gatkforums.broadinstitute.org/) (Supplementary Tables S4 and S5).

Comparison of SNVs derived from selected tools.

SNV results from each tool were parsed, edited, combined, compared and analyzed using custom R scripts (www.r-project.org). The GATK SelectVariants module was used to select out the SNVs if the original result files were in vcf format (i.e. GATK, samtools, VarScan results) or customized R scripts were composed and used to parse them out if not in standard vcf format (e.g. CLCBio, Partek, CASAVA results). The GATK VariantEval module was used to assess the Ti/Tv ratios of SNV call sets. MIEC is a way to determine whether SNP calls for the same chromosomal positions from members of a family trio set (father, mother and their child) pass or fail simple Mendelian inheritance rules across the family (see Supplementary Table S3 for rules). This checking was done using customized R scripts as a way to validate the derived SNPs in a high-throughput manner. Briefly, for each chosen SNP tool used for our comparison and for each family trio set (father, mother and child), the position of each derived SNP (i.e. heterozygote or homozygote variants or non-reference genotype) for each trio member was assessed for Mendelian inheritance consistency using the genotypes of the other two trio members at the same position. The positions of SNPs that passed or failed Mendelian inheritance were retrieved for each family trio member and were subjected to further analysis using Venn diagrams and testing against the SNP array data for the same SNP positions. The Medelian inheritance error rates were calculated for each member of a family trio set by dividing the numbers of the SNPs that passed the Mendelian inheritance rules by the numbers of that failed for each member. Then the average error rates were simply calculated as the means of the error rates of each member of the family trio set.

Using array data as the standard truth set for high-throughput validation and categorizing the SNP calls by genotypes, the specificity and sensitivity were calculated based on sensitivity = 100*TP/(TP+FN) and specificity = 100*TN/(TN+FP), where TP is the number of truth positives that were assessed as the number of SNPs (i.e. homozygote and heterozygote variants) in each NGS data call set from a specific tool that were also detected as the same genotype in the SNP array data; FN is the number of false negatives that were assessed as the number of non-SNPs (i.e. homozygote reference) in each NGS data call set but were

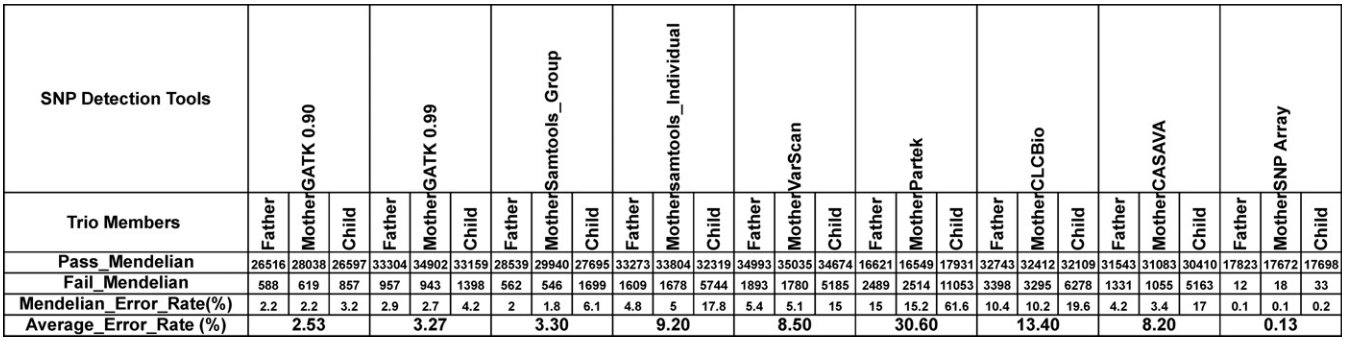
Table 1. Summary of parameter settings used for SNP calling and filtering for the selected SNP tools



SNP Detection Tools	GATK0.90	GATK0.99	Samtools_group	Samtools_Individual	VarScan	CASAVA	CLCBio	Partek
Default/suggested setting for SNP calling and filtering	UG: base quality score >= 17, standard_min_confidence_threshold_for_calling >=30, standard_min_confidence_threshold_for_emitting(report)>=20; VQSR For Filtering 0.90	UG: base quality score >= 17, standard_min_confidence_threshold_for_calling >=30, standard_min_confidence_threshold_for_emitting(report)>=20 ; VQSR For Filtering 0.99	Call samples as group; mapping quality >=0, base quality >=13; variant threshold <0.5, or p(ref| D) <0.5; filtering -d 10	Call samples individually; mapping quality >=0 and base quality >=13; variant threshold <0.5, which is p(ref| D) <0.5; filtering -d 10	coverage >= 4; var-freq >=0.20; p-value<=0.05	mapping quality >= 90 and base quality >=0; Q(snp) >0; Filter snp/genotype call: Q(snp)>=20 , Q(max|gt_poly) >=20	base quality >=20; coverage >= 4; var-freq >= 0.35	non.reference.average.base.qualities>=17 and log.odds.ratio.of.SNP.against.reference >=100 (if needed, non.reference.average.mapping.qualities >=238 25th percentile)
Final numbers of SNPs after filtering for sample #2	27454	34557	29394	38063	39859	35573	38387	28984

All tools used default settings or suggested settings based on either direct communication with its author, or technical support, or forum communication. Although 0.99 is the default setting for VQSR, two thresholds of the VQSR step of GATK (0.99 and 0.90) were used to assess the robustness of the tools and were designated as GATK0.99 and GATK0.90; samtools_group was designated for SAMtools calls using pooled samples simultaneously, whereas samtools_individual designated as the SAMtools calls using individual samples one at a time. UG: Unified genotyper from GATK.

Table 2. Summary of Mendelian inheritance error rates on SNPs detected from selected tools amongst the chosen family trio members



SNP Detection Tools	GATK 0.90			GATK 0.99			Samtools_Group			samtools_Individual			VarScan			Partek			CLCBio			CASAVA			SNP Array		
Trio Members	Father	Mother	Child	Father	Mother	Child	Father	Mother	Child	Father	Mother	Child	Father	Mother	Child	Father	Mother	Child	Father	Mother	Child	Father	Mother	Child	Father	Mother	Child
Pass_Mendelian	26516	28038	26597	33304	34902	33159	28539	29940	27695	33273	33804	32319	34993	35035	34674	16621	16549	17931	32743	32412	32109	31543	31083	30410	17823	17672	17698
Fail_Mendelian	588	619	857	957	943	1398	562	546	1699	1609	1678	5744	1893	1780	5185	2489	2514	11053	3398	3295	6278	1331	1055	5163	12	18	33
Mendelian_Error_Rate(%)	2.2	2.2	3.2	2.9	2.7	4.2	2	1.8	6.1	4.8	5	17.8	5.4	5.1	15	15	15.2	61.6	10.4	10.2	19.6	4.2	3.4	17	0.1	0.1	0.2
Average_Error_Rate (%)	2.53			3.27			3.30			9.20			8.50			30.60			13.40			8.20			0.13		

This table was assessing the SNPs detected with the selected SNP detection tools for samples #9, #10 and #2 from a family trio (Supplementary Figure S1). GATK0.90 and GATK0.99: GATK with two different VQSR thresholds at 0.90 and 0.99, respectively. All other SNP tools called SNPs at their default settings or suggested by authors as described at Table 1. Although the result reported here considered only SNPs on the array within the target interval regions of the genome defined by the exome enrichment kit (exome-subset), a similar result was obtained when all SNPs on the array were used (data not shown). Samtools_group: SAMtools calls with all available samples together. Samtools_Individual: SAMtools calls with each sample assessed individually. We included both, since there was some debate within the community of SAMtools users whether multi-sample SNP calling enhances the power for calling SNPs shared between samples and reduces the power for singleton SNPs (communication from the SAMtools developer in samtools-help forum discussion). Similar results were obtained for the other two trio sets (with sample #3 or #4 as child) available in the family (data for all of these family trio sets were shown at the top panels of multiple tables for Supplementary Table S4).

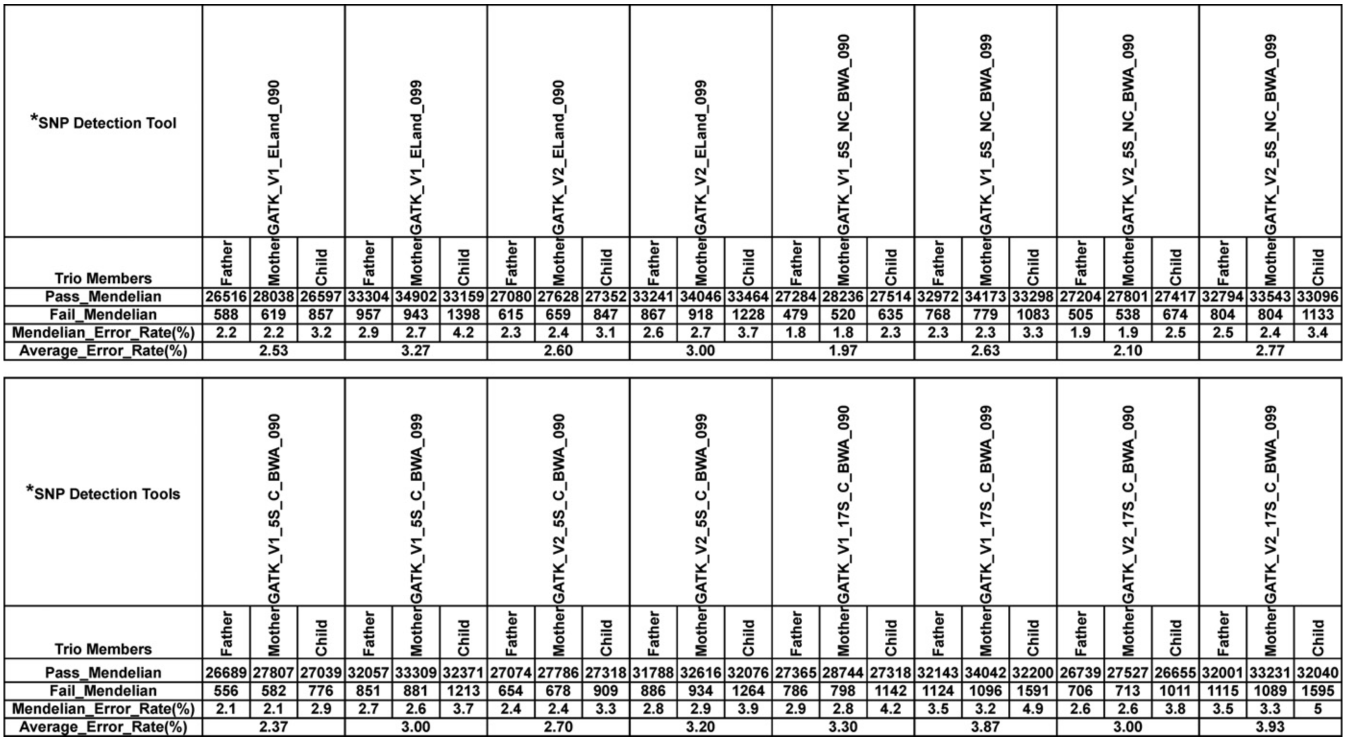
detected as SNPs in the SNP array data; TN is the number of true negatives that were assessed as the number of non-SNPs that were also detected as non-SNPs in the SNP array data and FP is the number of false positives that were assessed as the number of SNPs detected in NGS data but detected as non-SNPs in the SNP array data. Similarly, SNP concordance between exome-seq data and array data was also assessed for all SNPs, heterozygote, homozygote variants as well as for categorized genotypes, respectively.

RESULTS

Initial SNV calling for the benchmark data using selected tools

As described in more detail in the Materials and Methods section, the benchmark exome-seq data was generated from germline DNA extracted from lymphocytes of 19 human samples from two families with known pedigrees (Supplementary Figure S1), in which there are three complete trio family sets (father, mother and child) that are available and applicable for MIEC-based comparison strategy described below. BWA is the preferred alignment tool for Illumina data by the GATK development group (17,23). However, Eland-v2 is the mapper from Illumina with the attribute of local realignment that GATK phase I provides and was also claimed to perform even better than BWA in a previous comparison study (34). Therefore, to avoid comparison bias towards either GATK or the CASAVA tools, we chose to use the standard Illumina Eland-v2 method as the major mapping application to generate the mapped data for the main body of performance comparison of derived SNP quality for all selected SNP detection tools. As a complimentary approach, we also used BWA mapped data to help understand the impact of mapping methods on the quality of the downstream variant calling with GATK as the constant SNP caller. Since our main focus was to compare the performance of the SNP detection tools by developing reusable and reliable comparison metrics and strategies/methods for the community, we chose not to include a thorough evaluation of the mappers in this report.

Table 3. Summary of Mendelian inheritance error rates on SNPs derived from the new and old versions of GATK and different mappers



*SNP Detection Tool	GATK_V1_ELand_090			GATK_V1_ELand_099			GATK_V2_ELand_090			GATK_V2_ELand_099			GATK_V1_5S_NC_BWA_090			GATK_V1_5S_NC_BWA_099			GATK_V2_5S_NC_BWA_090			GATK_V2_5S_NC_BWA_099		
Trio Members	Father	Mother	Child	Father	Mother	Child	Father	Mother	Child	Father	Mother	Child	Father	Mother	Child	Father	Mother	Child	Father	Mother	Child	Father	Mother	Child
Pass_Mendelian	26516	28038	26597	33304	34902	33159	27080	27628	27352	33241	34046	33464	27284	28236	27514	32972	34173	33298	27204	27801	27417	32794	33543	33096
Fail_Mendelian	588	619	857	957	943	1398	615	659	847	867	918	1228	479	520	635	768	779	1083	505	538	674	804	804	1133
Mendelian_Error_Rate(%)	2.2	2.2	3.2	2.9	2.7	4.2	2.3	2.4	3.1	2.6	2.7	3.7	1.8	1.8	2.3	2.3	2.3	3.3	1.9	1.9	2.5	2.5	2.4	3.4
Average_Error_Rate(%)	2.53			3.27			2.60			3.00			1.97			2.63			2.10			2.77		

*SNP Detection Tools	GATK_V1_5S_C_BWA_090			GATK_V1_5S_C_BWA_099			GATK_V2_5S_C_BWA_090			GATK_V2_5S_C_BWA_099			GATK_V1_17S_C_BWA_090			GATK_V1_17S_C_BWA_099			GATK_V2_17S_C_BWA_090			GATK_V2_17S_C_BWA_099		
Trio Members	Father	Mother	Child	Father	Mother	Child	Father	Mother	Child	Father	Mother	Child	Father	Mother	Child	Father	Mother	Child	Father	Mother	Child	Father	Mother	Child
Pass_Mendelian	26689	27807	27039	32057	33309	32371	27074	27786	27318	31788	32616	32076	27365	28744	27318	32143	34042	32200	26739	27527	26655	32001	33231	32040
Fail_Mendelian	556	582	776	851	881	1213	654	678	909	886	934	1264	786	798	1142	1124	1096	1591	706	713	1011	1115	1089	1595
Mendelian_Error_Rate(%)	2.1	2.1	2.9	2.7	2.6	3.7	2.4	2.4	3.3	2.8	2.9	3.9	2.9	2.8	4.2	3.5	3.2	4.9	2.6	2.6	3.8	3.5	3.3	5
Average_Error_Rate(%)	2.37			3.00			2.70			3.20			3.30			3.87			3.00			3.93		

Summary of Mendelian inheritance error rates amongst the chosen family trio members (samples #9, #10 and #2; Supplementary Figure S1) for the SNPs detected with the new (V2.0 up to V2.2.4) or old (up to V1.6.7) versions of GATK and different mappers including Eland and BWA.

*Name designation for variations of GATK versions and mappers in the following format:

GATK_version_NumSample(option)_MapperContigOptionsForBWA(Option)_Mapper_VQSRThreshold. Version: V1 or V2 of GATK; NumSample: optional, if 5S, 5 samples with trio relations; if 17S, 17 samples in larger family; MapperContigOptionsForBWA: optional, if NC; no contigs in genome reference; C: with contigs in genome reference;

Mapper: BWA or Eland methods; VQSRThreshold: VQSR thresholds as 099 for 0.99 or 090 for 0.90. Similar results were obtained for the other two trio sets (with sample #3 or #4 as child) available in the family (data not shown).

As stated above, it has been demonstrated that the mapper will affect the identification of variants, and our approach is to separate these two important and interacting components and treat them independently. Although we did compare Eland and BWA in a limited effort (see below in the Results section), a more thorough and sophisticated performance comparison study with exhaustive permutation and combination of available mappers and SNP detection tools, which can evaluate the dependence of SNP detection tools on the choice of mappers, would be studied in the future as it would directly benefit from the outcome of this study. The statistical details of our benchmark exome-seq data are listed in a supplementary table (Supplementary Table S1) including numbers of reads, mapping rates as well as depth of coverage.

Currently, the choice of the variant-calling tools for each project remains subjective and arbitrary and is up to users' preference in spite of the fact that GATK has been well accepted and is regarded as 'Gold Standard' tool by the community. However, this status lacks solid objective evidence from a third party study. Therefore, we chose to focus on establishing the metrics of evaluating performance for several popular tools rather than perform a less detailed examination of all of the available. For our assessment, we selected some of the most popular tools from both academic and industrial settings including the pre-eminent GATK, SAMtools, VarScan (35), Illumina CASAVA, CLCBio Genomics Workbench and Partek Genomic Suite, which covered the three major SNP calling methods, i.e the heuristic filtering method (e.g. VarScan), Bayesian method (e.g. GATK Unified genotyper, samtools), and hyplotype-based method (e.g. GATK HyplotypeCaller) (28,32). It is likely that the some researchers' favorite tools are not present in this list of tools; however, our primary goal was to establish a method and benchmark dataset that other users would then be able to extend to include their own favorite tools.

Although it is possible in theory to combine modular components of different tools to produce overall performance improvements, for this study we chose to examine the capabilities of each application as a stand-alone entity. Our judgment was that this was the best way to provide a fair assessment, leaving the assessment of performance improvements from combining the best performing module for each step as a subsequent analysis. After the raw variant call set was obtained for each tool, either the default settings or the thresholds suggested by the tool producers were used to filter and optimize the raw call set to produce the final call set used for the performance comparison. Our entire workflow for this comparison study from data generation to SNP

Table 4. Summary of Ti/Tv ratios of SNP call sets derived from the selected SNP detection tools

SNP Detection Tools	GATK_090_Eland			GATK_099_Eland			GATK_NoFilter			samtools_NoFilter			samtools_group			samtools_individual		
Novelty	all	known	novel	all	known	novel	all	known	novel	all	known	novel	all	known	novel	all	known	novel
nTi	20481	20198	283	25230	24589	641	37244	30463	6781	32282	28777	3505	20976	18842	2134	27479	25449	2030
nTv	6973	6866	107	9327	8980	347	17915	12228	5687	13031	11117	1914	7766	6712	1054	10357	9348	1009
TiTvRatio	2.94	2.94	2.64	2.71	2.74	1.85	2.08	2.49	1.19	2.48	2.59	1.83	2.7	2.81	2.02	2.65	2.72	2.01

SNP Detection Tools	VarScan			CASAVA			CLCBio			Partek			SNP_Array		
Novelty	all	known	novel	all	known	novel	all	known	novel	all	known	novel	all	known	novel
nTi	28631	26498	2133	25869	24875	994	27689	25898	1791	20549	19054	1495	13724	13722	2
nTv	11153	9966	1187	9690	9176	514	10698	9797	901	8308	7306	1002	3927	3923	4
TiTvRatio	2.57	2.66	1.8	2.67	2.71	1.93	2.59	2.64	1.99	2.47	2.61	1.49	3.49	3.5	0.5

This table was assessing the SNPs detected with the selected SNP detection tools for one of the family members #2 (similar results are obtained for #9 and #10). GATK0.90 and GATK0.99: GATK with two different VQSR thresholds at 0.90 and 0.99, respectively. All other SNP tools called SNPs at their default settings or suggested by authors as described in Table 1. GATK_NoFilter: raw SNP call set (using Unified genotyper of GATK); samtools_NoFilter: raw SNP call set from Samtools_Group (using SAMtools without filtering); Other SAMtools results used –d 10 option to filter raw SNP call set (similar results obtained using –D option for filtering). samtools_group: SAMtools calls using all samples together; samtools_individual: SAMtools calls using each sample individually. (Note: the six novel SNPs of the SNP array were probably caused by the difference between the SNP array and dbSNP in SNP annotations).

Table 5. Summary of Ti/Tv ratios of SNP call sets derived from new and old versions of GATK and different mappers

*SNP Detection Tools	GATK_V1_Eland_090			GATK_V1_Eland_099			GATK_V2_Eland_090			GATK_V2_Eland_099			GATK_V1_5S_NC_BWA_090			GATK_V1_5S_NC_BWA_099			GATK_V2_5S_NC_BWA_090			GATK_V2_5S_NC_BWA_099		
Novelty	all	known	novel	all	known	novel	all	known	novel	all	known	novel	all	known	novel	all	known	novel	all	known	novel	all	known	novel
nTi	20481	20198	283	25230	24589	641	20976	20579	397	25273	24602	671	20792	20482	310	25174	24632	542	20773	20404	369	25061	24512	549
nTv	6973	6866	107	9327	8980	347	7218	7051	167	9407	8972	435	7351	7235	116	9200	8976	224	7314	7184	130	9163	8941	222
TiTvRatio	2.94	2.94	2.64	2.71	2.74	1.85	2.91	2.92	2.38	2.69	2.74	1.54	2.83	2.83	2.67	2.74	2.74	2.42	2.84	2.84	2.84	2.74	2.74	2.47

*SNP Detection Tools	GATK_V1_5S_C_BWA_090			GATK_V1_5S_C_BWA_099			GATK_V2_5S_C_BWA_090			GATK_V2_5S_C_BWA_099			GATK_V1_17S_C_BWA_090			GATK_V1_17S_C_BWA_099			GATK_V2_17S_C_BWA_090			GATK_V2_17S_C_BWA_099		
Novelty	all	known	novel	all	known	novel	all	known	novel	all	known	novel	all	known	novel	all	known	novel	all	known	novel	all	known	novel
nTi	20591	20221	370	24638	24084	554	20928	20578	350	24478	23953	525	21085	20711	374	24778	24230	548	20590	20256	334	24687	24135	552
nTv	7219	7092	127	8941	8735	206	7294	7177	117	8857	8659	198	7365	7247	118	8999	8791	208	7071	6951	120	8936	8726	210
TiTvRatio	2.85	2.85	2.91	2.76	2.76	2.69	2.87	2.87	2.99	2.76	2.77	2.65	2.86	2.86	3.17	2.75	2.76	2.63	2.91	2.91	2.78	2.76	2.77	2.63

Summary of Ti/Tv ratios of SNP call sets derived from the new (V2.0 up to V2.2.4) or old (up to V1.6.7) versions of GATK and different mappers including Eland and BWA. *Name designation for variations of GATK versions and mappers in the following format:
GATK_version_NumSample(option)_MapperContigOptionsForBWA(Option)_Mapper_VQSRThreshold. Version: V1 or V2 of GATK; NumSample: optional, if 5S, 5 samples with trio relations; if 17S,17 samples in larger family; MapperContigOptionsForBWA: optional, if NC; no contigs in genome reference; C: with contigs in genome reference. Mapper: BWA or Eland methods.; VQSRThreshold: VQSR threshold as 099 for 0.99 or 090 for 0.90.

calling and to validation are illustrated schematically in a supplementary table (Supplementary Table S2).

Interestingly, although the parameters and thresholds used for filtering are quite diverse amongst the tools, the final SNP call sets they produced all have rather similar numbers of SNVs as the starting point for the comparisons, as shown in Table 1. This is coincidently very useful in minimizing the impact of the size of call sets on the quality of the variants.

To simplify the performance evaluation and comparison, only the SNVs were selected from the original variant call sets of these tools for this report. The decision to focus on SNVs is not to diminish the importance of indels, but rather to assess the performance of the data in a systematic manner that maximizes the opportunity for validation, which is one of the main strengths of this study: being able to leverage the power of both SNP array and family information for MIEC for high-throughput validation. Furthermore, indels are much harder to evaluate and compare due to their complexity in nature, e.g. associated mapping/alignment issues and a lack of standardization on their discovery and reporting causing an obstacle in accurate comparison of indels derived from different tools as has been previously reported (33). Therefore, we chose to focus on the SNVs in this report and present an in-depth comparison of indel calling performance for later.

Since we mentioned that the evaluation of the performance of variant callers will be important in the context of clinical applications, it is important to mention that both quality of the calls and the performance of the caller will be critical. It is our view that these separate dimensions of performance should be evaluated independently. It seems likely that once the tool has been identified that provides the highest quality calls, the motivation to improve its computational performance will become self evident. Therefore, for this study, we primarily focused on directly assessing the quality of derived variants rather than the operational aspects of the performance of these tools.

MIEC

The final SNV call sets derived from the selected tools were first subjected to MIEC within each available family trio. The simple scoring rules we used are listed in a supplementary table (Supplementary Table S3) along with the methods applied for calculating these error rates. The quality of SNVs was assessed based on the calculated Mendelian inheritance error rates as an indirect indicator from each member of the trio (Supplementary Figure S1; Table 2) and the distribution of SNP positions across the family members of each trio was also assessed (Figure 1).

Since we intended to use this information as an additional basis for validation, in addition to the NGS-based call sets, Illumina SNP array data and the corresponding SNP call set was also generated from the same set of samples as described in the Materials and Methods section. As shown in Table 2, The SNP call set produced by the SNP array had the lowest Mendelian inheritance error rate (average 0.13%) compared to those derived from exome-seq data using any of the selected SNP detection tools (ranging from 2.53% to 30.60%). This result was consistent with the notion that SNP arrays were designed with well-behaved and well-known common variants, and also supported our assumption that the SNP call set generated from the array can be used as a reasonable ‘standard call set’ for high-throughput validation and evaluation of the quality of NGS SNP call sets generated by the selected SNP detection tools from the corresponding exome-seq data from the same biological subjects.

Amongst the tools for SNP detection using exome-seq data, the GATK call sets GATK0.99 and GATK0.90 had the lowest average error rates (3.27% and 2.53%, respectively; Table 2). While maintaining similar or even higher detection power especially for GATK0.99 (Table 2, Figure 1), the errors were distributed between common and unique SNPs across the trio members (Supplementary Figure S2).

To further study the details of the errors derived from the shared SNVs common to all family trio members, the corresponding SNVs detected by GATK0.99, samtools_individual and CASAVA were taken out for further analysis (Figure 2). Amongst the selected three groups, GATK0.99 detected more unique SNVs that passed the MIEC (4479, Figure 2a) than the other two tools. In addition, a larger portion of the unique SNVs detected by GATK0.99 were also identified as SNPs designed for detection on the array (42.53%) compared to CASAVA (2.38%) and samtools_individual (10.42%) (Figure 2a and b). Furthermore, for the SNVs that were uniquely detected by each NGS tool that also passed the MIEC (Figure 2c), a large portion of them were identified as SNPs on the array that also passed the MIEC over the SNPs detected by the array (Figure 2d). Within this class, GATK0.99 overwhelmingly outnumbered the other two (1424 from GATK versus 44 from CASAVA and 33 from samtools_individual) (Figure 2c). Finally, GATK0.99 detected many more ‘good’ unique SNVs that passed MIEC (4479, Figure 2e) but with a relatively low error rate (1.18%, Figure 2f), compared to the other two methods. These observations all suggest that GATK may have the best capacity to detect more unique SNPs that are of a higher quality. Similar observations were made when comparing GATK0.99, with VarScan and CLCBio (Supplementary Figure S3).

As we were preparing this manuscript, a new version of GATK (V2.0) was released that claimed to have made improvements in multiple phases of the procedure. This prompted us to perform an evaluation of how the new version GATK would impact our initial observations described above. In addition, it is also of great interest to assess whether the choice of aligners/mappers would impact the quality of variant calling. To accommodate both questions, we used the exome-seq data of the same samples mapped with either Eland or BWA and then processed with the GATK new and old versions to derive the SNP variants and then use a similar strategy to assess the MIEC as we did in Table 2. Interestingly, with the combination of different mappers and versions of GATK, we observed MIEC error rates in a very similar range between many new and old versions of GATK and between the different mappers (Table 3). After our initial submission of this manuscript, due to the dynamic and evolving nature of the GATK tool, there were a few more recent new versions of GATK that also were released with the announcement of the Hyplo-

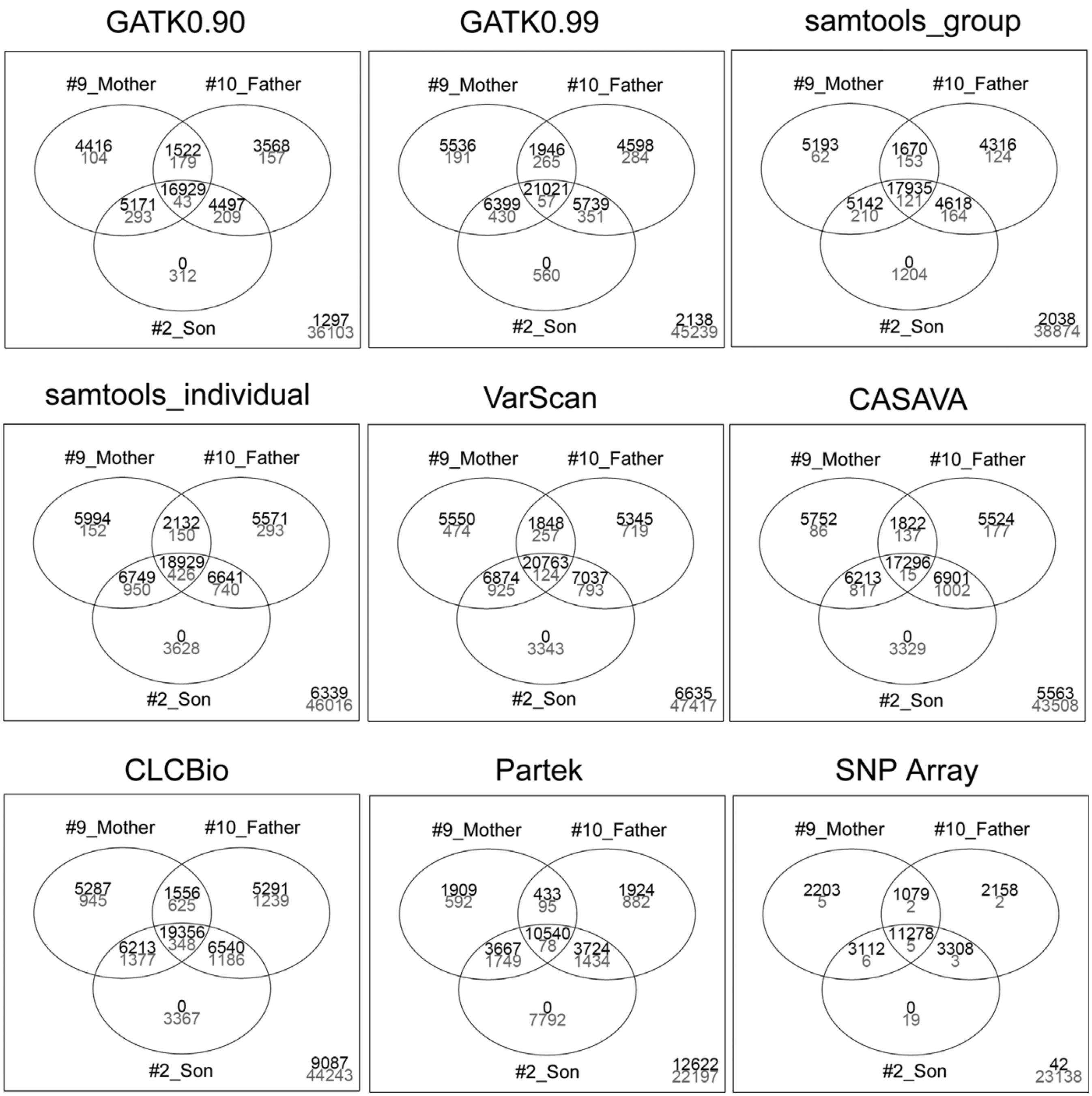


Figure 1. Distribution of SNP positions across family trio in selected SNP callers. SNPs of family trio composed of samples #9 (mother), #10 (father) and #2 (son) (Supplementary Figure S1), which were generated from GATK0.90 and GATK0.99 (VQSR at 0.90 and 0.99 threshold levels), samtools (call SNPs from samples either as a group or as individuals), VarScan, Partek, CLCBio, Illumina CASAVA and SNP array, were subjected to Venn diagram analysis for their positions. The numbers shown in the overlap indicate shared SNVs between the trio members and those in unique areas indicate unique SNVs for those members. Numbers in black are the number of SNV positions passing MIEC, whereas numbers in gray are the number of SNV positions failing MIEC. Similar results were obtained for the other two trio sets (with sample #3 or # 4 as child) available in the family (data not shown).

typeCaller improvements in both running time and performance. This step had been relatively time consuming and not practical to use in the version we initially evaluated. Interestingly, using our comparison methods, we were able to show that after version 2.6–4, the HTC showed a great improvement (compared to old version HTC) in quality of SNP calls given the fact that other steps in GATK did not change much including the VQSR step (Supplementary Table S4). We observed a similar trend for GATK HTC improvement (compared to GATK Unified genotyper of the same new version) even for another independent exome-seq dataset using our MIEC comparison scheme (Supplementary Table S5), which was derived using a newer Illumina platform (Hiseq) and updated reagents. Interestingly, we

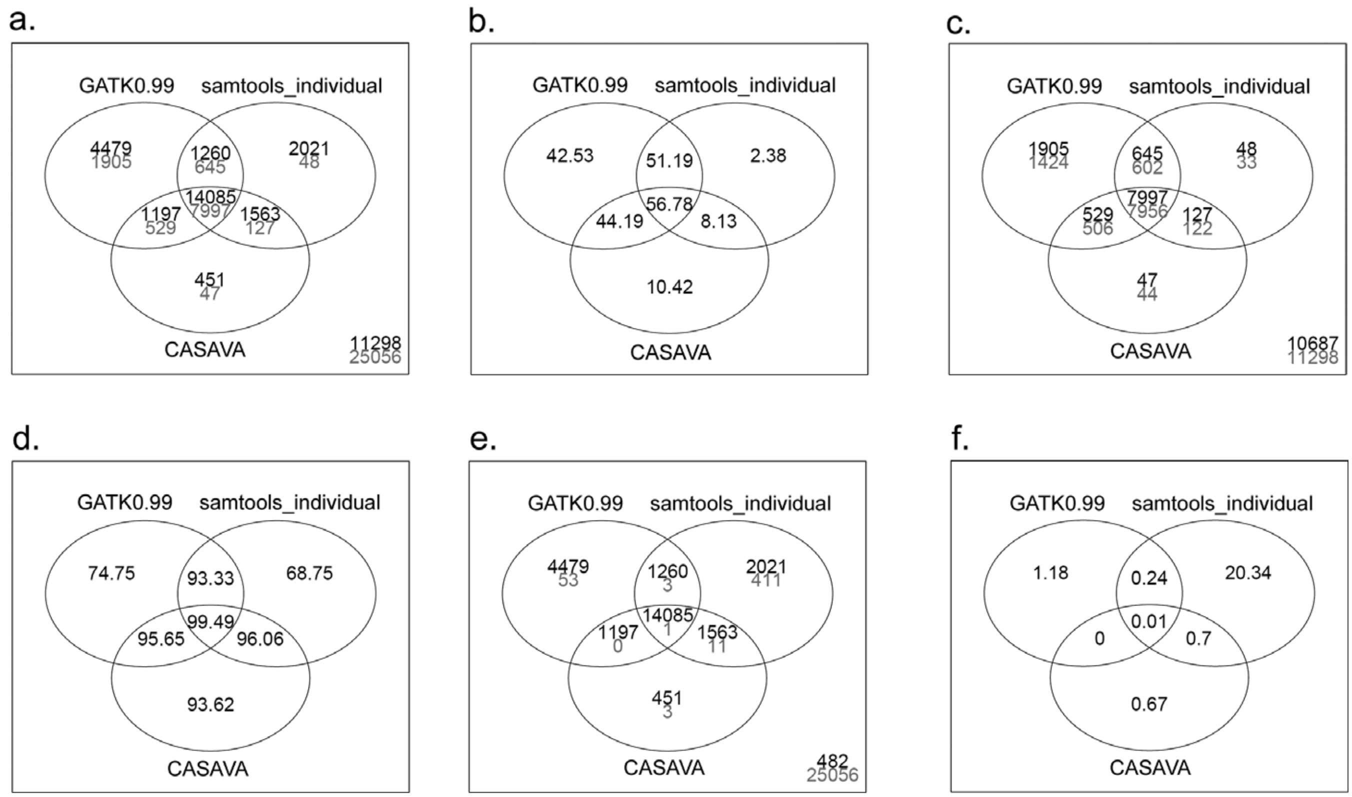


Figure 2. Distribution of SNP positions of the common SNPs of all members of family trio detected by GATK 0.99, samtools Individuals and Illumina CASAVA. Common SNPs of all members from family trio including #9, #10 and #2, which were generated from GATK 0.99 (0.99 threshold levels), samtools (call SNPs from samples as individuals) and Illumina CASAVA (Figure 1), were subjected to further Venn diagram analysis for their positions in details. The numbers shown in the overlapping areas indicate shared variants between the tools and those in unique areas indicate unique variants for each tool. (**a**) Numbers in black are the number of SNV positions passing MIEC, whereas numbers in gray are the number of variants designed for detection on the SNP array. (**b**) Numbers are percentage of SNVs passing MIEC that are also SNPs designed for detection on the SNP array (gray number divided by black number in each corresponding section of (a). (**c**) Numbers in black are the numbers of NGS-detected SNVs passing MIEC that are also designed for detection on the array, whereas numbers in gray are the number of SNPs designed for detection on SNP array for the same positions of NGS-detected SNVs passing MIEC, which has also passed MIEC within array data. (**d**) Percentage of NGS-detected SNVs passing MIEC that were also array-detected SNPs passing MIEC (gray number divided by black number in each corresponding section of (c). (**e**) Numbers in black indicate number of SNVs passing MIEC, whereas numbers in gray indicate the number of SNVs failing MIEC. (**f**) Error rate of MIEC based on (e) (gray number divided by black number in each corresponding section of (e).

only observed minor impacts from the choice of the mapper between BWA and Eland and also with the same mapper but with different genome references (e.g. with or without including the chrUn contigs) (Table 3; Supplementary Table S4).

Transition/transversion ratio (Ti/Tv) assessment

Beyond data validation with MIEC, the Ti/Tv ratio of variants has also been used as a quality metric for variant calls. The final SNP call sets derived from each selected tools were also subjected to Ti/Tv ratio assessment for known and novel SNVs using the GATK utility tool VariantEval. Table 4 lists the assessed Ti/Tv ratios of the SNP call sets derived from each selected NGS tool and the SNP array for a single sample. As expected, the SNP array has the highest Ti/Tv ratio of 3.5 for the known SNP set (Table 4). It is well known that the previously identified and more likely true SNVs will have a relatively high Ti/Tv ratio (22), although this is clearly only one dimension of the assessment. By definition, all of the SNPs placed on the SNP array are known and documented SNPs. This observation further substantiates the claim that the SNP call set generated from the array can be used as a reasonable 'standard' for high-throughput validation and evaluation of the quality of the exome-seq SNP call sets derived from the same sample. Amongst all exome-seq-based SNP call sets, the ones detected by GATK0.99 and GATK0.90 were generally higher in their Ti/Tv ratios of all SNVs (2.94 and 2.71, respectively) than those of other NGS tools (Table 4). These ratios were similar to the estimates of ~2.8 from 1000 Genomes data (3,23), but slightly higher than the estimates of 2.53–2.67 of human exome-seq data in a recent study (14).

To evaluate how the versions of GATK and the choice of aligners/mappers would impact our initial observations of the Ti/Tv ratios, we also assessed the Ti/Tv ratios of the variant call sets derived from the new version of GATK (V2.0) with the exome-seq data of these same samples but mapped with either Eland or BWA. In good agreement with MIEC results (Table 3), with the combination of different mappers and versions of GATK, we again observed a very similar range of Ti/Tv ratios between the new and old version GATK as well as with the different mappers (Table 5).

High-throughput validation by SNP array data

As described above, based on the lowest Mendelian inheritance error rate and highest Ti/Tv ratio for the SNPs detected on the array, it is reasonable to assume that the array SNPs have high overall accuracy and represent a good SNP set for high-throughput validation of the final SNV call sets generated with each of the selected tools. By design, the SNVs used for comparison are restricted to bases within the target interval regions that were defined by the SureSelect capture kit used for the exome-seq. For SNVs detected with the exome-seq data using the selected tools, based on whether they are consistent with the SNPs detected on the array data or not, various SNP call error rates were calculated (Supplementary Table S6). GATK-0.99, GATK_Nofilter and samtools_NoFilter have the lowest overall error rates of 1.79%,1.6% and 1.44%, respectively, although CASAVA, samtools_indiv and GATK0.90 have the lowest SNP call error rates of 1.36%, 1.39% and 1.73%, respectively (Supplementary Table S6), when considering only heterozygous and homozygous SNVs.

When considering one sample at a time and categorizing SNP calls by genotypes between each of the detection tools using the array data as the 'standard' (Supplementary Table S7a), GATK0.99 and samtools_NoFilter, obtained the best combination of Sensitivity and Specificity as well as SNP concordance rates amongst all the selected tools, with CASAVA performing the best amongst the commercial tools. As expected for GATK, the higher threshold of 0.99 for the VQSR step has higher sensitivity than that at more stringent threshold of 0.90 of VQSR, but the specificity at 0.99 was lower than that at 0.90. In addition, as generally recommended by the GATK team for the VQSR step, using the 0.99 threshold as the default setting appears to have a better combination of specificity and sensitivity than the 0.90 threshold. The relatively high calculated specificity for all of the selected tools may be caused by the relatively high number of homozygote reference genotypes in the SNP array data for those samples. Interestingly, heterozygous SNVs had the lowest percentages of consistency in general compared to homozygous references and homozygous SNV calls (Supplementary Table S7b).

Consistent with the observations described above, with combinations of different mappers and versions of GATK, we again observed very similar ranges of SNP array based error rates, or sensitivity, specificity and concordance rates, between the new and old version GATK and between the different mappers (see Supplementary Table S8 or Supplementary Table S9, respectively).

Side-by-side comparison of tools using only the subset of SNPs passing MIEC

In order to limit the comparison to only the best sets of SNPs passing MIEC, the subsets of SNPs that were derived from the chosen SNP tools to be compared and met one of the three best scenarios were used for a side-by-side comparison for each variant at the same position (Figure 3). In light of initial efforts of comparing GATK and CASAVA (36), side-by-side comparison between GATK and CASAVA was done along with the raw GATK call set (without the VQSR step) as a control set. As shown in Figure 3a, the majority of qualified SNPs (those that met the scenario) were common amongst the raw GATK call set, GATK0.99 and CASAVA. Interestingly, GATK0.99 had many unique SNVs not seen by CASAVA, but these were derived from the GATK VQSR step filtering from the raw GATK call set (2050, Figure 3a). Many of these were also present on the SNP array (823, Figure 3a). Not only were more SNVs detected by GATK and retained after GATK0.99 VQSR filtering and missed by CASAVA (2050 from GATK0.99 versus 91 from CASAVA) (Figure 3a), but also a higher proportion of them were detected on the SNP array (40.15% from GATK0.99 versus 23.08% from CASAVA) (Figure 3b). In addition, a large proportion (94.9%, 781 out of 823) of these SNVs that were only detected by GATK0.99 also met the same scenario over the SNPs detected on the array (Figure 3c and d). These observations indicate that these unique SNVs that were only detected by GATK0.99 (filtered from raw GATK calls with VQSR) had a higher probability to be true SNPs and they outnumbered the CASAVA (2050 versus 91) with criteria indicative of high quality including passing MIEC, a larger proportion on SNP array and a larger portion passing MIEC on the SNP array data. This observation suggested that the major driver for the higher accuracy in GATK's final call set might be the VQSR step.

Sanger sequencing validation

In order to have one additional look at the obtained SNP call sets, Sanger sequencing was applied to validate a selected handful of the identified SNPs. Due to the consistent high quality of SNPs and performance of the tools in most, if not all of evaluation metrics described above, GATK0.99's SNP call set was used as the input set for possible validation. Amongst the set of tested SNP calls from GATK0.99, only 5.26% (10 out of a total 190 SNVs from 10 selected validated SNVs from 19 samples of the dataset) SNVs failed to be validated (Supplementary Table S10), which is within the general range for acceptable validation rates used with the 1000 Genomes project (>95%) (3,37) and others (38). It is interesting to note that the major failures were heterozygous SNVs (Supplementary Table S10). With a close-up look at examples, SNVs that failed to be validated were simply caused by a largely unbalanced occurrence of reads matching the variant and reference base at the SNP site (Supplementary Figures S4 and S5). This observation suggests that allelic frequency could be used as an additional filtering criterion to help weed out false positive SNPs even for the best SNP detection tools. Of course, this assumption would only be valid for detection of SNVs from germline DNAs where diploid content can be assumed. Interestingly, the SNPs that passed the validation from Sanger sequencing result were primarily common SNVs that were detected by most of the NGS tools as well (data not shown), which is consistent with other studies suggesting a voting approach to SNV calling (30).

DISCUSSION

The ability to interpret NGS data in a clinical setting to guide both diagnosis and therapeutic course demands that we are able to derive an accurate list of the variations harbored in an individual's genome (39,40). Clearly we need to

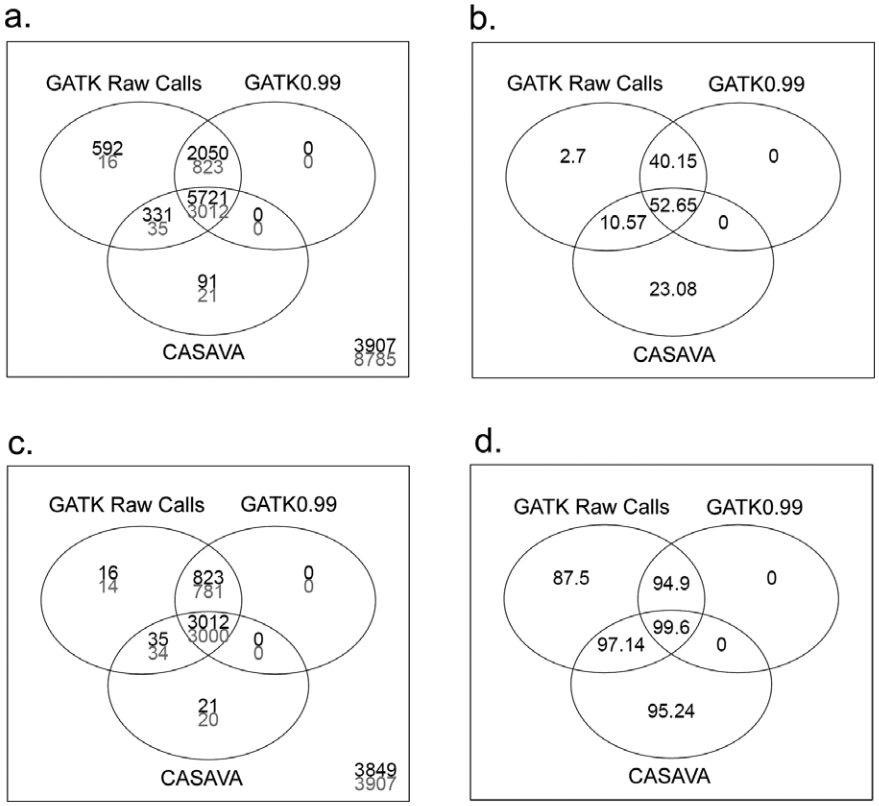


Figure 3. Heads-up comparison of GATK and CASAVA on the subset of SNPs passing MIEC under one of the three best MIEC scenarios. The three best MIEC scenarios are as follows. (i) Both parents are homozygous variant and child has to be homozygous variant. (ii) Both parents are homozygous reference and child has to be homozygous reference. (iii) One of parents is homozygous variant and the other parent is homozygous reference and the child has to be heterozygous variant. SNVs derived from GATK raw calls (No VQSR Filtering), GATK0.99 or CASAVA that meets the MIEC scenario (i) were subjected to Venn diagram analysis. (**a**) Numbers in black indicate number of SNVs derived from NGS data that passed the above MIEC scenario. Numbers in gray indicate the number of SNVs derived from NGS data that passed the above MIEC scenario and also were designated SNVs for detection on SNP array. (**b**) Percentage of numbers in gray over the numbers in black in each area of (a). (**c**) numbers in black indicate numbers of SNVs derived from NGS data that passed the above MIEC scenario and also were designated SNPs for detection on SNP array. Numbers in gray are the numbers of SNPs designed for detection on SNP array for the same positions of NGS-detected SNVs passing MIEC, which has also passed MIEC within array data. (**d**) Percentage of numbers in gray over the numbers in black in each area of (c). A similar observation was made for other two scenarios (data not shown).

ensure that we are seeing all of the relevant mutations that might inform treatment without missing any potential leads and also avoiding the many pitfalls that would arise from false positives. This is also especially critical for the studies on rare coding variation (41). As a result, it is important to study the behavior of the tools applied at each of these critical steps in the variant identification process in a rigorous manner using the best possible dataset that possesses internal validation capabilities.

In this study, we have used a benchmark Illumina exome-seq dataset to evaluate the quality of SNP calls derived from a set of selected SNP discovery tools. Family pedigree information and SNP array data for the same samples offers great power to assess the quality of SNP calls from these selected tools side by side in a high-throughput fashion by checking for Mendelian inheritance inconsistency errors within the SNP calls of family members and by ‘high-throughput validation’ using the SNP array data. There were many concerns in previous studies (31) that only applied MIEC to assess the quality of variant calls with potential bias of calls on parents not being truth calls but rather consistent calls amongst family trio members, as well as studies (32,33) that only reported general concordance of multiple variant-calling pipelines lacking of a full truth set of variants for high-throughput validation. In spite of emerging potential truth datasets derived from some new technologies such as polymerase chain reaction free whole genome data and call sets of multiple families with known genotype from 1000 Genomes Project, to what level especially at high-throughput scale they are comparable to well-known high-quality SNP array dataset will have to wait to see with much detailed characterization and careful assessment. We believe that our strategy combining both concordance checking with MIEC and high-throughput validation with SNP array data provided maximum benefit for current technology being lack of high-throughput validation datasets in the field. In addition, our analysis compared several different metrics of performance so that the relative strengths of each tool were objectively assessed from different perspectives and this helped to identify the best performing tool. With the help of the SNP array, the proportion of known or documented SNPs can be assessed and used as an indirect indicator for the quality of the SNV set, since known and documented SNPs would more likely represent true SNPs. To the best of our knowledge, none of the previously reported studies have applied the strategy of 'high-throughput validation' as we did using both family pedigree information and SNP array data. As a consequence, only our study can uniquely report the relative performance and strengths of each tool using objective comparison metrics, rather than simply the concordance rate amongst the tools in comparison that would not inform any relative strength and performance in quality of derived variants from each tool as previous studies did (32,33).

Although GATK has been widely accepted and used by the field as a 'gold standard' for SNP detection, at least as applied to germline NGS data (37,42,43), solid evidence was needed to substantiate the claim. Our evaluation and comparison results on Illumina exome-seq data show that, amongst the selected tools at the time of testing, GATK performed either the best or in the top tier for most, if not all, of our comparison metrics and schemes and also was the most consistent across different evaluative tests. Our study is consisted with the current view in the field that GATK represents the gold standard as a SNP detection tool in the field, and much of its power is derived from the VQSR module. The GATK development team did show the benefit of the machine learning VQSR module of GATK comparing to manual hand filtering primarily using concordance with Hapmap and 1KG data, and TiTv ratios (23). However, in our study, we not only observed many lines of evidence (e.g. side-by-side comparison of GATK with CASAVA with validation by SNP array, using the best subset of MIEC, Ti/Tv ratio, etc.) that clearly showed that it is the VQSR step of GATK that contributes the most to the high quality of its SNV call set, but also specifically showed that it is the VQSR module not the Unified genotyper that outperformed CASAVA, which was not described in the previous comparison study (36). In addition, our results provide insights for the potential for further improvement with filtering by read allelic frequency even for the current best-performer, GATK (not the population-based allelic frequency, e.g. minor allele frequency (MAF) for SNPs).

However, it shall be noted that due to the rapidly evolving nature of SNP detection tools and the relatively small number of these tools we evaluated, the ranking of our selected tools is intended to demonstrate our comprehensive framework under which different tools can be compared and evaluated. Our comparison of these selected tools represents only a snapshot in time of each tools performance and by no means is static as all of these tools are expected to continue to improve over time. These examples also suggest that our comparison study and the comparison methods and metrics we developed would also be able to reveal mechanistic details as to why one tool performs better than others. Finally, a more detailed assessment of combination of modules or steps of multiple-step or multiple-module tools such as GATK using a similar approach described in this report would help understand why some SNP callers or their steps/modules perform better than the others.

Consistent with the general notion in the community and a previous study (37), we also observed that the common SNVs detected by multiple tools may have the best quality and are the most reliable call sets for validation. It should be noted that although some previous studies did reach a similar conclusion based on the simple concordance between the tools (32,33), our study indeed specifically showed a big portion of such shared variants have much lower MIEC rate but higher percentage of those that are annotated to be detected in SNP array compared to unique variants detected by other tools, which consequently tend to be variants of higher quality (e.g. Figure 2b and f, Supplementary Figure S3b and S3f, data not shown). Such observations revealed that our comparison strategies empowered by the MIEC and SNP array validation can allow fine-tuned comparison of variants and flexibility in addressing different aspects of variant tool performance. In addition, our results strongly suggested that GATK0.99 has the greatest capacity to uncover more unique SNVs but with a lower false discovery rate than any other selected tools at the time of testing,

One of the commonly used metrics in quality comparisons of SNP call sets is the Ti/Tv ratio. Similar to what has been used by the GATK development team (23), Ti/Tv ratio has also been used as one of metrics for performance comparison in our study. However, precaution should be taken to overinterpret the Ti/Tv ratios versus the quality of variant call sets, since it is not always true that Ti/Tv ratios would necessarily mean more accurate (i.e. more specific) SNP call sets. Sometimes, collections of low-frequency (rare) variants often have higher Ti/Tv ratio than moderate-frequency SNPs (reason not completely understood) and specific values of the Ti/Tv ratio sometimes may just point to characteristics of call sets having nothing to do with their quality. However, Ti/Tv ratio is just one of many metrics that we used for performance comparison, and many of others are primarily based on the high-throughput validation by family-pedigree based MIEC and annotation of SNP array to ensure that conclusion was not derived from single aspect of the comparison but a series of metrics considering different aspects of the variants. As such, a community forum type of setting should be fostered to allow collection and collation of comparison data, strategies to design comparison metrics that are contributed from the community, which in turn would be beneficial to the entire research community. As expected, the comparison metrics and strategy used and presented in this study would be a very good starting point for such activities.

Our intention with this study was not only to provide an example cases study to illustrate our comparison strategies and metrics using our benchmark data, but also to foster and promote such efforts and support from the community for providing resources with more benchmark datasets and/or even better comparison strategies and metrics. With multiple independent benchmark datasets available contributed from the community, we can certainly minimize bias from individual datasets, platfoms as well as choice of tools used in comparisons.

It is also important to note that, as one limitation of our study, we only compared variant detection tools for germline exom-seq data and as a result our conclusions shall focus only on the germline variants from diploid genomes but not directly on the somatic mutations. However, many of the principles and strategies employing high-throughput validation with family pedigree information and SNP array and lesson learnt from this study could be very helpful in designing comparison metrics and strategies for comparison of somatic mutation detection tools. We believe similar strategies could be employed if we can collect samples for both germline and somatic mutations of the same patients and use the family pedigree information and available sample-matched SNP array data to validate the background mutations/variants, which would make the somatic mutation much easier to be assessed for the quality of the detection tools. Therefore, we believe our strategy and method used in this study can be generalized and reusable in a broad range of performance comparison studies.

To the best of our knowledge, this is the first report of a large-scale performance comparison and evaluation of SNP discovery tools on the most popular Illumina exome-seq data employing both MIEC across family members within family trios and SNP array data from the same samples for high-throughput validation. The comparison results immediately provide information relating to the accuracy of the selected tools and insights for the selection of a SNP discovery tool. More importantly, our comparison approach using the tested benchmark dataset and well-designed evaluation metrics and schemes provides a template for objectively comparing SNP discovery tools, which can be applied to any forthcoming SNP detection tools of interest as they become available. It should be also noted that there are many other platforms such as WGS, more narrow-spectrum capture assays and many other vendors such as Pacific Biosciences and Ion Torrent. Due to the complexity of this dynamically evolving field, we chose to focus on the most popular exome-seq platform from Illumina in the hope that we can not only establish a solid and reusable comparison metrics and method for existing or any forthcoming SNP discovery tools on Illumina exome-seq data, but also provided a generic approach and strategy that can be potentially replicated to compare variant-calling tools for other platforms and/or other types of NGS data.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Jigui Shan and Xiongfeng Chen for assistance with installation and maintenance of tools from academic and commercial sides, respectively. We also thank representatives or developers from Illumina, CLCBio, Partek, GATK, VarScan and SAMtools communication community for their support, and helpful discussions as we performed these tests. We would like to thank the Editor and two referees for helpful comments that led to improvements in this paper. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products or organizations imply endorsement by the U.S. Government.

FUNDING

National Cancer Institute; National Institutes of Health (Contract No. HHSN261200800001E, in whole or in part). Funding for open access charge: National Cancer Institute, National Institutes of Health (Contract No. HHSN261200800001E).

Conflict of interest statement. None declared.

REFERENCES

1. Hamosh,A., Scott,A.F., Amberger,J.S., Bocchini,C.A and Mckusick,VA. (2005) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.* **33**, D514–D517.
2. Hindorff,L.A., Sethupathy,P., Junkins,H.A., Ramos,E.M., Mehta,J.P., Collins,F.S. and Manolio,T.A. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. U.S.A.*, **106**, 9362–9367.
3. The 1000 Genomes Project Consortium (2010) A map of human genome variation from population-scale sequencing. *Nature*, **467**, 1061–1073.
4. Shendure,J. and Ji,H. (2008) Next-generation DNA sequencing. *Nat. Biotechnol.*, **26**, 1135–1145.
5. Metzker,M.L. (2010) Sequencing technologies - the next generation. *Nat. Rev. Genet.*, **11**, 31–46.
6. Gilissen,C., Hoischen,A., Brunner,H.G. and Veltman,J.A. (2011) Unlocking Mendelian disease using exome sequencing. *Genome Biol.*, **12**, 228.
7. Worthey,E.A., Mayer,A.N., Syverson,G.D., Helbling,D., Bonacci,B.B., Decker,B., Serpe,J.M., Dasu,T., Tschannen,M.R., Veith,R.L. *et al.* (2011) Making a definitive diagnosis: successful clinical application of whole exome sequencing in a child with intractable inflammatory bowel disease. *Genet. Med.*, **13**, 255–262.
8. Meyerson,M., Gabriel,S. and Getz,G. (2010) Advances in understanding cancer genomics through second-generation sequencing. *Nat. Rev. Genet.*, **11**, 685–696.
9. Bamshad,M.J., Ng,S.B., Bigham,A.W., Tabor,H.K., Emond,M.J., Nickerson,D.A. and Shendure,J. (2011) Exome sequencing as a tool for Mendelian disease gene discovery. *Nat. Rev. Genet.* **12**, 745–755.
10. Cirulli,E.T. and Goldstein,D.B. (2010) Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nat. Rev. Genet.*, **11**, 415–425.
11. Stitziel,N.O., Kiezun,A. and Sunyaev,S. (2011) Computational and statistical approaches to analyzing variants identified by exome sequencing. *Genome Biol.*, **12**, 227.
12. Okou,D.T., Steinberg,K.M., Middle,C., Cutler,D.J., Albert,T.J. and Zwick,M.E. (2007) Microarray-based genomic selection for high-throughput resequencing. *Nat. Methods*, **4**, 907–909.
13. Gnirke,A., Melnikov,A., Maguire,J., Rogov,P., LeProust,E.M., Brockman,W., Fennell,T., Giannoukos,G., Fisher,S., Russ,C. *et al.* (2009) Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat. Biotechnol.*, **27**, 182–189.
14. Clark,M.J., Chen,R., Lam,H.Y., Karczewski,K.J., Chen,R., Euskirchen,G., Butte,A.J. and Snyder,M. (2011) Performance comparison of exome DNA sequencing technologies. *Nat. Biotechnol.*, **29**, 908–914.
15. Li,H. and Homer,N. (2010) A survey of sequence alignment algorithms for next-generation sequencing. *Brief Bioinform.*, **11**, 473–483.
16. Li,H., Ruan,J. and Durbin,R. (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.*, **18**, 1851–1858.
17. Li,H. and Durbin,R. (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*, **26**, 589–595.
18. Fonseca,N.A., Rung,J., Brazma,A. and Marioni,J.C. (2012) Tools for mapping high-throughput sequencing data. *Bioinformatics*, **28**, 3169–3177.
19. Marth,G.T., Korf,I., Yandell,M.D., Yeh,R.T., Gu,Z., Zakeri,H., Stitziel,N.O., Hillier,L., Kwok,P.Y. and Gish,W.R. (1999) A general approach to single-nucleotide polymorphism discovery. *Nat. Genet.*, **23**, 452–456.
20. Li,H., Ruan,J. and Durbin,R. (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.*, **18**, 1851–1858.

21. Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G. and Durbin,R. (2009) 1000 Genome Project Data Processing Subgroup: The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.
22. McKenna,A., Hanna,M., Banks,E., Sivachenko,A., Cibulskis,K., Kernytsky,A., Garimella,K., Altshuler,D., Gabriel,S., Daly,M. and DePristo,M.A. (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.*, **20**, 1297–1303.
23. DePristo,M.A., Banks,E., Poplin,R., Garimella,K.V., Maguire,J.R., Hartl,C., Philippakis,A.A., del Angel,G., Rivas,M.A., Hanna,M. *et al.* (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.*, **43**, 491–498.
24. Quinlan,A.R., Stewart,D.A., Strömberg,M.P. and Marth,G.T. (2008) Pyrobayes: an improved base caller for SNP discovery in pyrosequences. *Nat. Methods*, **5**, 179–181.
25. Li,R., Li,Y., Fang,X., Yang,H., Wang,J., Kristiansen,K. and Wang,J. (2009) SNP detection for massively parallel whole-genome resequencing. *Genome Res.*, **19**, 1124–1132.
26. Zeng,F., Jiang,R. and Chen,T. (2013) PyroHMMsnp: an SNP caller for Ion Torrent and 454 sequencing data. *Nucleic Acids Res.*, **41**, e136.
27. Carneiro,M.O., Russ,C., Ross,M.G., Gabriel,S.B., Nusbaum,C. and DePristo,M.A. (2013) Pacific biosciences sequencing technology for genotyping and variation discovery in human data. *BMC Genomics*, **13**, 375.
28. Nielsen,R., Paul,S.J., Albrechtsen,A. and Song,S.Y. (2011) Genotype and SNP calling from next-generation sequencing data. *Nat. Rev. Genet.*, **12**, 443–451.
29. Altmann,A., Weber,P., Bader,D., Preuss,M., Binder,E.B. and Müller-Myhsok,B. (2012) A beginners guide to SNP calling from high-throughput DNA-sequencing data. *Hum. Genet.*, **131**, 1541–1554.
30. Lam,H.Y., Clark,M.J., Chen,R., Chen,R., Natsoulis,G., O'Huallachain,M., Dewey,F.E., Habegger,L., Ashley,E.A., Gerstein,M.B. *et al.* (2012) Performance comparison of whole-genome sequencing platforms. *Nature Biotechnol.*, **30**, 78–82.
31. Deng,X (2011) SeqGene: a comprehensive software solution for mining exome- and transcriptome-sequencing data. *BMC Bioinformatics*, **12**, 267.
32. Pabinger,S., Dander,A., Fischer,M., Snajder,R., Sperk,M., Efremova,M., Krabichler,B., Speicher,M.R., Zschocke,J. and Trajanoski,Z. (2014) A survey of tools for variant analysis of next-generation genome sequencing data. *Briefs Bioinform.*, **15**, 256–278.
33. O'Rawe,J., Jiang,T., Sun,G., Wu,Y., Wang,W., Hu,J., Bodily,P., Tian,L., Hakonarson,H., Johnson,W.E. *et al.* (2013) Low concordance of multiple variant-calling pipelines, practical implications for exome and genome sequencing. *Genome Med.*, **5**, 28.
34. Bauer,M.J., Cox,A.J., Dirk,J. and Evers,D.J. (2010) ELANDv2 - Fast gapped read mapping for Illumina reads. ISCB 2010 poster collection category *'J'-Genomics*
35. Koboldt,D.C., Zhang,Q., Larson,D.E., Shen,D., McLellan,M.D., Lin,L., Miller,C.A., Mardis,E.R., Ding,L. and Wilson,R.K. (2012) VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.*, **22**, 568–576.
36. Bauer,C.D. (2011) Variant calling comparison CASAVA1.8 and GATK. *Nat. Proc.*, doi:10.1038/npre.2011.6107.1.
37. Marth,G.T., Yu,F., Indap,A.R., Garimella,K., Gravel,S., Leong,W.F., Tyler-Smith,C., Bainbridge,M., Blackwell,T., Zheng-Bradley,X. *et al.* (2011) The functional spectrum of low-frequency coding variation. *Genome Biol.*, **12**, R84.
38. Barbieri,C.E., Baca,S.C., Lawrence,M.S., Demichelis,F., Blattner,M., Theurillat,J.P., White,T.A., Stojanov,P., Van Allen,E., Stransky,N. *et al.* , (2012) Exome sequencing identifies recurrent SPOP, FOXA1 and MED12 mutations in prostate cancer. *Nat. Genet.*, **44**, 685–689.
39. Desai,A.N. and Jere,A. (2012) Next-generation sequencing: ready for the clinics? *Clin. Genet*., **81**, 503–510.
40. Cooper,G.M. and Shendure,J. (2011) Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. *Nat. Rev. Genet.*, **12**, 628–640.
41. Tennessen,J.A., Bigham,A.W., O'Connor,T.D., Fu,W., Kenny,E.E., Gravel,S., McGee,S., Do,R., Liu,X., Jun,G. *et al.* (2012) Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science*, **337**, 64–69.
42. Zang,Z.J., Cutcutache,I., Poon,S.L., Zhang,S.L., McPherson,J.R., Tao,J., Rajasegaran,V., Heng,H.L., Deng,N., Gan,A. *et al.* (2012) Exome sequencing of gastric adenocarcinoma identifies recurrent somatic mutations in cell adhesion and chromatin remodeling genes. *Nat. Genet.*, **44**, 570–574.
43. Banerji,S., Cibulskis,K., Rangel-Escareno,C., Brown,K.K., Scott,L., Carter,S.L., Frederick,A.M., Lawrence,M.S., Sivachenko,A.Y., Sougnez,C. *et al.* (2012) Sequence analysis of mutations and translocations across breast cancer subtypes. *Nature*, **486**, 405–409.