

The RTG Pipeline for High Performance Variant Detection in Whole Genome and Exome Sequencing Applications

A comparison to the BWA/GATK Framework

Brian Hilbush, Ph.D.
Technical Director
Real Time Genomics, Inc.



Real Time Genomics, Inc.
576 Folsom Street, 2nd Floor
San Francisco, CA 94105
info@realtimegenomics.com
<http://www.realtimegenomics.com>

Abstract

After a decade of progress since the first human genome was sequenced¹⁻², there remain significant hurdles to enable rapid, highly sensitive and accurate analysis of complex genomes. The current procedures utilize a collection of tools and require substantial computational costs and bioinformatics expertise for successful implementation. Here, we present an integrated, robust analysis pipeline that outputs high quality variants and processes next-generation sequencing data on time scales that are an order of magnitude faster than comparable methods established with a repertoire of techniques and stand-alone steps. The RTG pipeline is anchored by a gapped alignment algorithm that performs search and alignment, pairing, and recalibration in a single unified step, and outputs sorted alignment files in SAM or BAM format for variant calling. The RTG variant caller utilizes a Bayesian technique for SNP and indel detection, integrates a probabilistic aligner for local realignment of read data around complex regions and performs analysis of genome locus ambiguity in a single processing operation. The efficient, two-stage RTG pipeline identifies the vast majority of variants detected by a BWA/GATK/variant recalibrator process run on whole genome sequencing or exome data, attaining 99.76% concordance of overlapping calls, while achieving the end result in a fraction of the time.

Introduction

Whole genome sequencing approaches have produced an immense wealth of data on human population-level sequence diversity and revealed the unexpected complexity of an individual's normal and disease-state genomes³⁻⁷. Exponential expansion of the number of human genomes sequenced is ushering in a new era of biology and a framework for genomics-based medicine. Although sequencing chemistries and instrumentation continue to advance, discriminating true variants from noise will continue to require extremely robust data processing and software analytical techniques. The accurate identification of known genetic variants and the discovery of 'private' variants in previously unsequenced individuals are prerequisites for personalized medicine and genome-based diagnostics⁸⁻⁹. Thus, the major goals toward this aim in downstream data analysis are to identify medically relevant and actionable disease-associated mutations such as single nucleotide polymorphisms (SNPs), copy number variants (CNVs) and chromosomal aberrations, and to ultimately produce an accurate representation of an individual's genome structure.

The analysis of whole genome sequencing (WGS) data for variant detection has required the development of an array of bioinformatics techniques that function together in a serial pipeline. Examples include pipelines developed around open source software by large genome centers¹⁰⁻¹¹ and those with proprietary algorithms and software from commercial sequencing technology vendors^{3,7}. Although the primary aim of these pipelines has been to discriminate machine or algorithm-induced artifacts from true sequence variants, most of

the software tools have been built *ad hoc* and without the ability to compensate for noise generated in previous computational steps. As a result, both systematic and random errors introduced at early stages of the pipeline often accumulate and are transmitted to the downstream variant calling algorithms. At present, the standard experimental design for WGS studies (typically >30X coverage) combined with Bayesian statistical approaches greatly reduce the probability of assigning a random machine-generated sequencing error as a high quality variant. However, more challenging are alignment-induced errors caused by read mapping algorithms that attempt to match each read to a genomic position without the benefit of consensus information. In addition, several of the mapping algorithms lack adequate gap and mismatch tolerance during the search or alignment phases. Errors at the read mapping stage are more likely to be systematic and are responsible for the majority of incorrect genotypes and variant calls seen prior to extensive filtering and validation.

The current pipelines based on open source tools invoke a labyrinth of computational steps tailored to the current sequencing technologies employed in the researcher's laboratory. As a result, these pipelines are difficult to establish in computing environments designed for other purposes or analytical software, and they may not function end-to-end with alternative algorithms that are preferred by the researcher or in current use. What is needed is an efficient analysis pipeline that delivers high quality variant data, reduces the number of data processing operations and maintains modularity. Here, we present a compact, integrated analysis pipeline built for highly sensitive and accurate detection of sequence variants following reference genome mapping and alignment of next-generation sequencing data. We describe innovative computational techniques to overcome the inherent noise and uncertainty in data generated from multiple platforms and experimental methodologies. The robustness of the pipeline is demonstrated in comparison to whole genome and exome sequencing data analyzed in a recent publication on the 1000 Genomes Project sample NA12878.

RTG's Variant Detection Pipeline

The RTG pipeline is built to process data from Illumina, 454 or Complete Genomics sequencing platforms and adjusts for their respective error models. The variant pipeline is sequence technology-aware, as read data can be analyzed separately or combined for multi-platform variant calling on the same sample. The RTG software utilizes a novel Bayesian technique with realignment and ambiguity filtering integrated into the variant caller. We designed the software to work with standard formats to enable interoperability with open source tools. The read mapping module produces sorted SAM output and the variant caller takes sorted SAM or BAM files as input. Multiple analysis pipelines built on a single RTG platform or hybrids with user-defined tools can be run in several modes, including operations with RTG tools only as in a RTG map → SAM → RTG snp configuration, or with alternative aligners or variant callers at either end of the pipeline.

The RTG variant detection toolset also has utilities for calculating genome depth of coverage, deriving intersections between datasets and for post-filtering and annotation of snp files. A variety of utilities are also available for evaluating SAM alignment files. The variant detection software includes a CNV module for analysis of copy number variants in sample pairs (not described here).

The schematic in Figure 1 illustrates the RTG algorithms and processing steps (light blue) that produce variant data for human genome analyses. Files containing raw, primary sequence data from a genome are mapped to an assembled human reference genome. Alignment data (orange) in the SAM reference file format serve as the basis for detecting sequence variants and structural variants, such as CNVs.

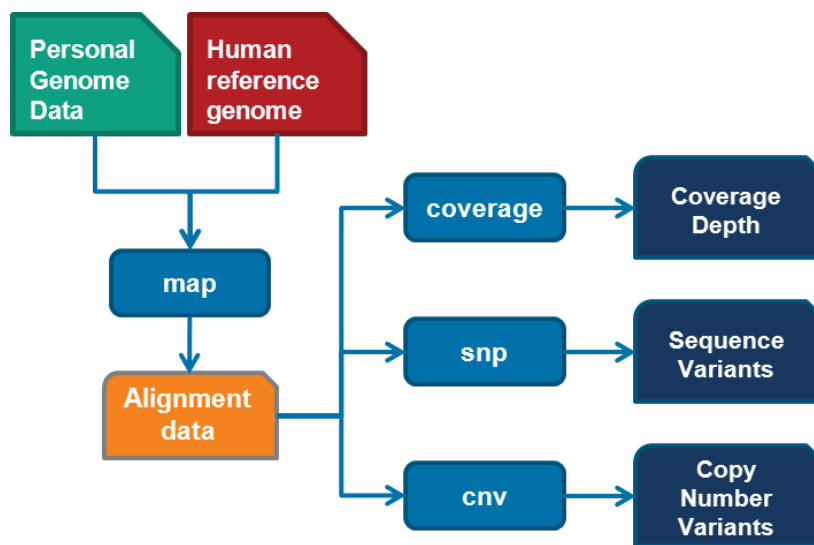


Figure 1. RTG Human Genome Analysis Pipeline

Features. The variant caller (command in RTG V2.2 is named ‘snp’) accepts alignment data from multiple sequencing platforms to be used during variant detection. Variant calling options include the ability to exclude unpaired reads and thresholds can be set on genome coverage levels and alignment scores for both mated and unmated read data. RTG snp produces calls for SNPs, MNPs, and short indels and outputs data in standard VCF file format. In addition to explicit identification of variants, complex regions encountered after comparing read data to the reference genome are delineated. The Bayesian variant caller integrates analysis of locus complexity (potential indel site, MNP, or other genomic feature) and a probabilistic alignment mechanism to produce the final variant calls and posterior scores. In the first stage, variant sites are located and assigned to either simple or complex classes. Variants in the simple class, typically isolated and not found in close proximity to other variants, are called directly without passing through a realignment stage. An ambiguity ratio is determined at every locus, which is computed as the ratio of ambiguous

to unique overlapping read mappings. For loci below the ambiguity threshold setting (default setting is 0.1), SNPs are output in VCF format. In the current RTG VCF output, variants clearing the user-defined filtering or threshold settings are denoted with 'PASS', those failing ambiguity or coverage thresholds (parameters set in snp command) are indicated with tags such as 'a10' or 'c100', respectively, and those in complex regions reported with either a "RS" or 'RX' designation.

Simple SNPs. The RTG SNP caller uses a Bayesian model to describe the read data mapped to the reference model in terms of "hypotheses" about the genotype at each position. There are ten different possible hypotheses, 4 homozygous hypotheses (AA, CC, GG and TT) and six heterozygous hypotheses (AC, AG, AT, CG, CT, and GT). The RTG method differs from some Bayesian SNP callers that consider only the 4 homozygous hypotheses and produce a heterozygous call from the best two hypotheses. The standard Bayesian calculation is as follows:

$$P(H|D) = \frac{P(D|H)P(H)}{\sum_{G \neq H} P(D|G)P(G)}$$

with $P(H)$ the prior probability of hypothesis H (one of the 10 possible genotype hypotheses), $P(D|H)$ the posterior probability that the data (D - the set of nucleotides mapped to this position) will occur given that the hypothesis H is true, and $P(H|D)$ the posterior probability that the hypothesis H is true given the data D .

The posterior likelihoods are normalized to a probability that the hypothesis is correct and thus the score reported by the caller is $\log_{10}(p/(1-p))$. In this scoring scheme, if the probability that the hypothesis is incorrect equals 1/2, then the posterior score calculated is 0.0. Similarly, if the probability that the call is incorrect equals 1/10 then the computed posterior score will be approximately 1.0. Post-filtering is achieved with thresholds on posterior scores, SNP density, coverage or other criteria. The RTG approach utilizes prior probabilities for each genotype as derived from experimental observations. $P(D|H)$ is computed by first assuming that the contributions of the different nucleotides mapped to a location are independent of each other:

$$P(D|H) = \prod_d P(d|H)$$

where d is the individual nucleotides. The calculation of $P(d|H)$ takes account of the manufacturer's estimates of the quality of the called base in the read. These quality estimates have been recalibrated from a set of mapped reads. Let q be the probability that the called base is incorrect, then if H is a homozygous call:

$$P(d|H) = \begin{cases} 1 - q & d = \text{reference} \\ q/3 & d \neq \text{reference} \end{cases}$$

and if H is a heterozygous call of H₁ or H₂:

$$P(d|H) = \frac{P(d|H_1) + P(d|H_2)}{2}$$

$P(H)$ the prior for the hypothesis H, is computed from observed rates of the various types of SNPs. These rates take note of the reference nucleotide as well as the rates observed SNP given that reference.

Complex SNPs. The SNP caller detects regions as “complex” in a two-step process. It first notes locations that are “interesting” when

- there is a simple SNP called
- there is a low confidence call that it is equal to the reference (using a default threshold of 1.0)

Complex regions are detected when:

- two interesting regions occur close to each other (by default if they are 3 bp apart or less)
- two or more insertions occur at a location
- complex regions that overlap are merged into a larger complex region

If a complex region is too large (by default more than 20 nt long) then it is classified as a “hyper-complex” region which is noted in the VCF output but no attempt is made to actually call it.

Example:

<u>Location</u>	<u>Observation</u>
42	10 insertions
52	SNP
55	5 insertions
57	low posterior reference call
65	SNP

During the call process, this region is found and split into two complex regions, with the indel at start position 42, a complex region at position 52 to 57, and a SNP at position 65 remaining as a simple SNP call.

The remaining complex regions are called by a Bayesian technique, which realigns the reads. This is done by first extracting a set of hypotheses for the complex region from reads which completely overlap the region. These hypotheses are taken from the original alignments. Then each hypothesis is used to replace the part of the genome covered by the complex region and all the reads are realigned against this modified template. The realignment is a probabilistic one that considers the sum of probabilities of all ways that the read could realign against the modified template. The total probability is taken as the probability $P(d|H)$ where d is the read and H is the hypothesized replacement for the complex region. These probabilities are then combined via Bayes theorem to give a posterior probability for which hypothesis or pair of hypotheses is the most likely diploid call at that location.

RTG analysis pipeline for whole genome sequencing data from NA12878

Read Mapping. The RTG pipeline was applied to the analysis of whole genome sequencing data on the 1000 Genomes project sample NA12878. Deep coverage (>60X) was attained from 16 lanes of Illumina HiSeq paired-end reads (101bp read length). These data were generated at the Broad Institute (ref. 10) and available [here](#). A total of 2.41 billion reads (1.205×10^9 pairs) were mapped to the human reference (hg18) using RTG's alignment algorithm (map command). The mapping parameters were set to default settings for alignment score threshold (-e 10%), word (-w 18) and step (-s 18), with a permissive mating distance window set at 1000 (-m 0 and -M 1000). With RTG's map, read data indexing, mate pair detection, collection of data for recalibration, and alignment are all performed in a single processing step for each set of paired fastq files (up to 80 million reads/file/run in the computing environment described below). The alignment data were reported in SAM format, with a maximum of 5 reported results (-n 5) per read.

Read mapping statistics

The RTG map command outputs summary statistics for each job by assigning reads to several categories. In the table below, a summary is given for reads that were either properly mated (paired) or unmated and mapped either uniquely or ambiguously. Reads classified as unmapped or that fail QC measures during various mapping stages fall into several categories. Examples include reads with mappings across too many locations ($n > 5$ top positions), reads that map to reference but whose mismatch total exceeds the alignment score threshold (10 across the 101bp read length), reads with poor matings (outside of pairing window) and reads with no hits. The edit distance calculation used for determining the alignment score for each read to the reference position is the sum of mismatches (+1), gaps (+1) and gap extensions (+1), where a one-base indel is scored +2.

Table 1. RTG Aligner Mapping Summary for NA12878

Mapping Category	Read count	% of total
Mated		
uniquely mapped	2,144,438,186	89.12%
ambiguously mapped	32,583,846	1.35%
total mated	2,177,022,032	90.48%
Unmated		
uniquely mapped	74,832,335	3.11%
ambiguously mapped	2,490,426	0.10%
total unmated	77,322,761	3.21%
Unmapped or QC failure	151,757,343	6.31%
total	2,406,102,136	100.00%

Genome Coverage Statistics

Coverage depth was computed after mapping against hg18 (regions with N bases excluded) for both total and uniquely mapped reads. The data are shown in Table 2.

Table 2. Illumina HiSeq Coverage of NA12878 with RTG Mapped Reads

	Mapped Reads	Mapped bases (Gb)	Coverage depth (fold)
Total	2.25E+09	228	79.7
Unique only	2.22E+09	224	78.5

Performance. The WGS dataset was aligned to the human reference genome followed by variant calling in approximately 24 hours (total wall clock time) on a single node in a compute cluster. Each cluster node has a dual Intel Xeon E5520 quad-core processor (hyper-threaded) running at 2.3GHz and with 48GB RAM. Data was read and written on shared NFS networked drives. The read mapping throughput of RTG's gapped aligner on the 101bp HiSeq data is determined by averaging each of the processing jobs (32 total, each lane split into approximately half). The processing times are given from the start of the mapping job to the completion of the writing of SAM output to disk. Times are given per lane and also normalized per core (reads/core/hr). The total compute time for the alignment step was 17 node hours, while variant calling, which utilized the direct output of the RTG aligner (sorted SAM files), had a compute time of 7 node hours (snp calls were performed on a per chromosome basis with up to 8 runs performed in parallel on a compute node).

Pipeline Processing Statistics

Table 3A. Read Mapping Throughput

Data	Lanes	Reads per lane	Processing time (hr/lane)	Throughput (reads/hr)	Per core processing rate (reads/core/hr)
HiSeq, 101bp, paired-end	16	1.50E+08	1.1	135,542,169	8,471,386

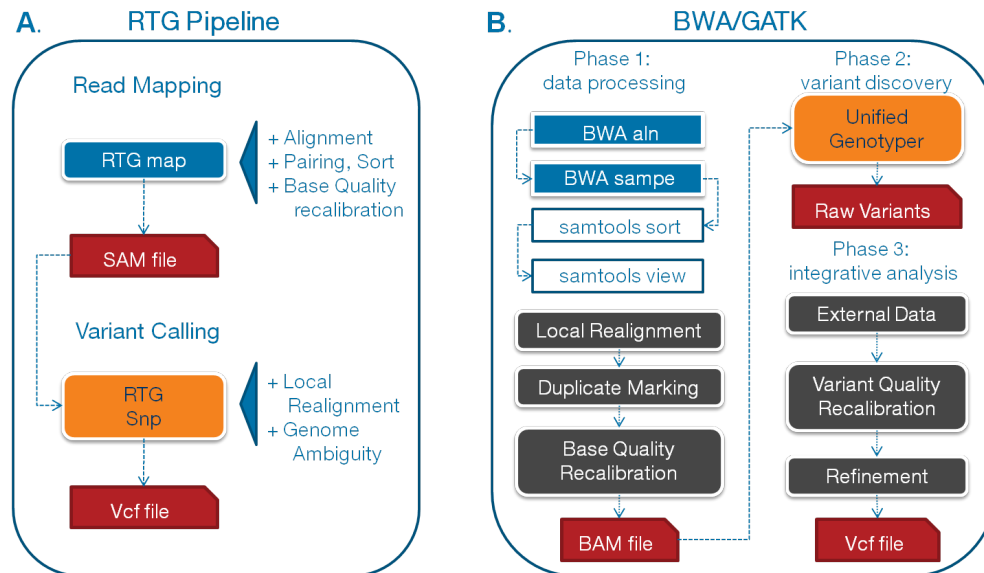
Table 3B. SNP Processing Rate

Data (GB)	Chromosome set	Processing time (hr/chr)	Throughput (nt/hr)	Total compute time (hr)
SAM (~200GB)	chr 1-22, X	0.305	4.62E+08	56.16

The RTG Variant Detection Pipeline

The RTG analysis pipeline comprises two main steps: read mapping and variant calling (Fig2A). The compact, computationally efficient pipeline differs substantially from those that impose a large number of independent steps, such as found in the recently described framework built around the BWA aligner and Unified Genotyper from the Broad Institute¹⁰ (Fig2B). RTG's gapped alignment algorithm and paired-end processing mode produces sorted SAM output and calibration files that are used by the variant caller. In contrast to the Broad Institute's pipeline, the evaluation of complex genomic loci and local realignment steps are performed as an integral part of the variant caller. Variant calling options allow unpaired reads to be excluded and users can set thresholds on genome coverage levels and alignment scores for both mated and unmated read data.

Figure 2. Variant Detection Pipelines



RTG Variant Calling on NA12878- Whole Genome Sequencing Data

Variant calls were generated from RTG alignment data on NA12878 using the collection of mated and unmated reads passing the default alignment threshold (see above). SNPs and MNPs were detected and output in standard VCF format. During variant calling, a coverage depth cutoff of 150 was used along with an ambiguity ratio limit of 0.1. A posterior score cutoff of 1.0 was applied to create the RTG raw call set listed in Table 4. These initial calls showed substantial overlap with the calls output and reported by GATK, but with many fewer apparent false positives based on QC metrics (Table S1). We applied an allele balance filter ($AB \geq 0.75$) to reduce erroneous heterozygous SNP calls in a single 'hard' filtering step. The AB filter effectively eliminated a substantial fraction (139,177 out of 331,427) of the novel calls with poor QC metrics in the RTG unique call set. The two pipelines produced 3,284,953 at the same position, with a 99.76% concordance in the overlap. The quality of the non-reference sites detected in both GATK and RTG unique call sets was substantiated by level of novelty in dbSNP (build130). Although a somewhat smaller fraction of RTG unique calls are present in dbSNP (60% vs. 70% for GATK), the filtering breakdown (Table S1) suggests that the RTG variant caller has provided a higher degree of confidence (i.e. calculated better posterior scores) for detection and inclusion of SNPs in or near clusters and around indels compared to GATK and the variant recalibrator.

Table 4. Variant Call Analysis - NA12878 WGS Data

NA12878 HiSeq Variant Sets	SNPs	vs. GATK (%)	dbSNP 130 (%)
GATK raw calls	4,312,621		89.5
RTG raw calls ($p \geq 1$)	3,634,285		93.5
Overlap		92.2	91.3
GATK 'PASS' (variant recalibration)	3,581,926		97.0
RTG 'PASS' ($p \geq 1$, $AB \geq 0.75$)	3,477,202		96.3
Overlap	3,284,953	94.0	95.8
concordant calls	3,276,977	99.76	
discordant calls	7,976	0.24	
GATK unique calls	296,973		70.3
RTG unique calls	192,250		60.5

Variant Calling on NA12878- Exome Data

RTG's analysis pipeline was also applied to detect variants from exome capture data (available [here](#)). Read data were re-mapped with the RTG alignment tool and variant calling was performed as described above for WGS data. Following variant calling, SNPs were identified in exome regions (as defined [here](#)) along with the full call set (target plus non-target SNPs). Comparisons were then performed on SNPs found in both regions from the

Illumina GAI data between RTG and GATK (Table 5). Additional intersections were derived for SNP data taken from the exome GAI set and the HiSeq runs for the WGS study. RTG’s pipeline found 97.1% of the GATK calls on GAI exome data. The concordant rate was extremely high, with 99.76% agreement of genotypes called between the GATK GAI ‘PASS’ SNPs and the RTG target region SNPs.

Table 5. Variant Call Analysis- NA12878 Exome Data

exome call set 1	exome call set 2	overlap (%)	intersect	GATK only	RTG only
gatk_ga2_PASS	rtg_ga2_target_p1	97.1	16,473	461	2,990
gatk_ga2_PASS	rtg_ga2_full_p1	96.3	16,350	536	3,002
gatk_ga2_PASS	rtg_hiseq_p1	95.2	16,194	776	2,711
gatk_hiseq_PASS	rtg_ga2_target_p1	92.9	16,495	1,178	2,923
gatk_hiseq_PASS	rtg_ga2_full_p1	97.9	17,388	308	1,990
gatk_hiseq_PASS	rtg_hiseq_p1	96.8	17,151	565	1,716

Discussion

We demonstrate that RTG’s analysis pipeline and underlying algorithms function in a highly integrated fashion to produce high quality variant data for whole genome and exome sequencing experiments. The computational efficiencies built into the software overcome the noise and uncertainty inherent in NGS data from current instrumentation and deliver genomic variants with high sensitivity and precision. The speed of the mapping and SNP calling algorithms combine with an efficient, parallel processing design to circumvent the expensive computational steps and requisite expertise needed to operate current bioinformatics analysis pipelines.

In our measurements of data processing performance on identical data sets, we demonstrate that RTG is an order of magnitude faster than the most recent implementation of the Broad Institute’s variant discovery framework. The speed of the RTG analysis pipeline does not lead to sacrifice in sensitivity in variant detection, as evidenced by the high percentage overlap of SNPs reported from both RTG and GATK. Variant call sets from RTG detected comparable levels of SNPs with extremely high concordance, both before and after GATK variant recalibration or filtering. The quality metrics associated with the concordant set and the novel call set from RTG provide strong evidence that the integrated approach performs at quality levels attained only after employing steps of local realignment prior to SNP calling and post-processing with the variant recalibrator in GATK. The application of these steps in the BWA/GATK process identified 730,695 SNPs in low quality categories or with low variant scores (i.e. 4,312,621 ‘raw’ minus 3,581,926 recalibrated SNPs). The RTG pipeline removed 84% of these during the single variant

calling step. Interestingly, in the remaining 16%, many of these candidate SNPs reside in the lower scoring, 'permissive' tranches after variant recalibration and scoring in the Broad framework (Table S1). RTG's variant caller retained approximately 55% of these in total. Most of these would be relatively easy to identify and classify into posterior score 'tranches'. We also determined that fewer than 5.5% (RTG) or 8.2% (GATK) of the final calls were found to be unique to either pipeline. These data indicate that both pipelines equally detect a vast majority of candidate variants, while providing complementary sets for novel discovery. The variant calls from the RTG pipeline may be essential for projects that require the greatest latitude for discovery of single mutations or rare variants, such as in Mendelian disease studies, where high sensitivity and good scoring mechanisms are required to evaluate candidate SNPs from NGS data.

In summary, any deployment of RTG's pipeline will greatly reduce the number of computational steps and associated costs, eliminate the need for time-consuming optimization of filtering protocols, and will perform as well as the multi-phase pipeline without need for a heterogeneous collection of software components.

References

1. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*. 2001 Feb 15;409(6822):860-921.
2. Venter, et al., The sequence of the human genome. *Science*. 2001 Feb 16;291(5507):1304-51.
3. Bentley, DR et al. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*. 2008 Nov 6;456(7218):53-9.
4. Wheeler, DA et al. The complete genome of an individual by massively parallel DNA sequencing. *Nature*. 2008 Apr 17;452(7189):872-6.
5. Ley TJ, et al. DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature*. 2008;456(7218):66-72.
6. Mardis ER, et al. Recurring mutations found by sequencing an acute myeloid leukemia genome. *N Engl J Med*. 2009;361(11):1058-1066.
7. Drmanac R, et al. Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science*. 2010 Jan 1;327(5961):78-81.
8. Welch JS, et al. Use of whole-genome sequencing to diagnose a cryptic fusion oncogene. *JAMA*. 2011 Apr 20;305(15):1577-84.
9. Link DC, et al. Identification of a novel TP53 cancer susceptibility mutation through whole-genome sequencing of a patient with therapy-related AML. *JAMA*. 2011 Apr 20;305(15):1568-76.
10. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, Del Angel G, Rivas MA, Hanna M, McKenna A, Fennell TJ, Kernytzky AM, Sivachenko AY, Cibulskis K, Gabriel SB, Altshuler D, Daly MJ. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet*. 2011 May;43(5):491-8.
11. Koboldt, DC., Chen, K., Wylie, T. Larson DE, McLellan MD, Mardis ER, Weinstock GM, Wilson RK, Ding L. VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics*. 2009 Sep 1;25(17): 2283-5.
12. 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature*. 2010 Oct 28;467(7319):1061-73.

Supplemental Data

Table S1. GATK Filtered SNP Analysis

GATK filter field description	SNPs retained by RTG				
	GATK total	same	diff	RTG total	fraction
DPFilter;HARD_TO_VALIDATE;Indel	1085	35	9	44	0.0004
DPFilter;HARD_TO_VALIDATE;Indel;LowQual	328	6	1	7	0.0001
DPFilter;HARD_TO_VALIDATE;Indel;LowQual;SnpCluster	71	1	0	1	0.0000
DPFilter;HARD_TO_VALIDATE;Indel;SnpCluster	320	4	139	143	0.0012
DPFilter;HARD_TO_VALIDATE;LowQual	20963	679	0	679	0.0058
DPFilter;HARD_TO_VALIDATE;LowQual;SnpCluster	2203	28	2	30	0.0003
DPFilter;HARD_TO_VALIDATE;SnpCluster	12067	133	20	153	0.0013
DPFilter;Indel	10325	568	168	736	0.0063
DPFilter;Indel;LowQual	4021	70	26	96	0.0008
DPFilter;Indel;LowQual;SnpCluster	867	25	11	36	0.0003
DPFilter;Indel;SnpCluster	2884	102	26	128	0.0011
DPFilter;LowQual	72231	5566	106	5672	0.0484
DPFilter;LowQual;SnpCluster	7852	223	26	249	0.0021
DPFilter;SnpCluster	47925	1867	142	2009	0.0172
FDRtranche1.00to10.00	89914	31180	606	31786	0.2715
FDRtranche1.00to10.00+	313615	43924	3140	47064	0.4020
HARD_TO_VALIDATE;Indel	1197	135	28	163	0.0014
HARD_TO_VALIDATE;Indel;LowQual	174	5	1	6	0.0001
HARD_TO_VALIDATE;Indel;LowQual;SnpCluster	52	0	0	0	0.0000
HARD_TO_VALIDATE;Indel;SnpCluster	332	20	7	27	0.0002
HARD_TO_VALIDATE;LowQual	6467	350	26	376	0.0032
HARD_TO_VALIDATE;LowQual;SnpCluster	672	30	3	33	0.0003
HARD_TO_VALIDATE;SnpCluster	4911	517	86	603	0.0052
Indel	48442	11466	3040	14506	0.1239
Indel;LowQual	1973	72	25	97	0.0008
Indel;LowQual;SnpCluster	589	25	3	28	0.0002
Indel;SnpCluster	8165	457	108	565	0.0048
LowQual	14341	1248	35	1283	0.0110
LowQual;SnpCluster	3300	214	20	234	0.0020
SnpCluster	53409	9528	791	10319	0.0881
totals	730695	108478	8595	117073	1