Humanizing Genomics
macrogen

Homo sapiens
Whole Genome Resequencing

Report

January 2018

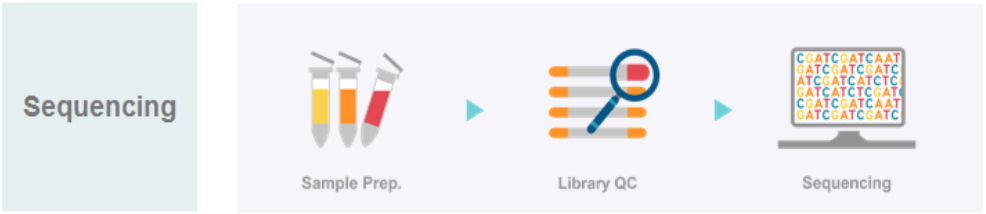
Basic Information

SampleID	NA12878
Project	0000KHX-0000
Institute	macrogen
Customer	macrogen

Table of Contents

1. HiSeq X Experiment

1. 1. Experiment Overview



The samples were prepared according to the Illumina TruSeq Nano DNA library preparation guide or TruSeq DNA PCR-free library preparation guide. The libraries were sequenced using Illumina HiSeq X sequencer.

1. 2. Experiment Procedure

1. 2. 1. Library Construction

- **DNA Fragmentation**

  Each sequenced sample is prepared according to the Illumina TruSeq DNA sample preparation guide to obtain a final library of 300-400 bp average insert size. One microgram (TruSeq DNA PCR-free library) or 100 nanogram (TruSeq Nano DNA library) of genomic DNA is fragmented by covaris systems, which generates dsDNA fragments with 3' or 5' overhangs.

- **End Repair and Size Selection**

  The double-strand DNA fragments with 3' or 5' overhangs are converted into blunt ends using an End Repair Mix. The 3' to 5' exonuclease removes the 3' overhangs, and the polymerase fills in the 5' overhangs. Following the end repair, the appropriate library size is selected using different ratios of the Sample Purification Beads.

- **Adenylation of 3' End**

  A single 'A' nucleotide is added to the 3' ends of the blunted fragments to prevent them from ligating to one another during the adapter ligation reaction. A corresponding single 'T' nucleotide on the 3' end of the adapter provides a complementary overhang for ligating the adapter to the fragment.

- **Adapters Ligation**

  Multiple indexing adapters are ligated to the ends of the DNA fragments to prepare them for hybridization onto a flow cell.

- **DNA Fragments Enrichment (TruSeq Nano DNA library only)**

  PCR is used to amplify the enriched DNA library for sequencing. The PCR is performed with a PCR primer solution that anneals to the ends of each adapters.

- **Library Validation**

  Macrogen performs quality control analysis on the sample library and quantification of the DNA library templates.

1. 2. 2. Clustering & Sequencing

Illumina utilizes a unique "bridged" amplification reaction that occurs on the surface of the flow cell. A flow cell containing millions of unique clusters is loaded into the HiSeq X for automated cycles of extension and imaging.

Sequencing-by-Synthesis chemistry utilizes four proprietary nucleotides possessing reversible fluorophore and termination properties. Each sequencing cycle occurs in the presence of all four nucleotides leading to higher accuracy than methods where only one nucleotide is present in the reaction mix at a time. This cycle is repeated, one base at a time, generating a series of images each representing a single base extension at a specific cluster.

1. 2. 3. Generation of Raw Data

The Illumina HiSeq X Ten generates raw images and base calling through an integrated primary analysis software called RTA 2(Real Time Analysis 2).

The BCL (base calls) binary is converted into FASTQ using illumina package bcl2fastq2-v2.20.0.

The demultiplexing option (--barcode-mismatches) was set to default (value : 1).

2. Data Handling Procedure

2. 1. Analysis Overview

2. 2. Analysis Software

2. 2. 1. Isaac Aligner

The Isaac aligner is an ultrafast DNA sequence aligner, designed to align next-generation sequencing data with low-error rates (single or paired-ends). It is four to five times faster than BWA + GATK on equivalent hardware, with comparable accuracy. The Isaac aligner was developed by illumina, Inc.

Please refer to the below paper for more information.

Raczy C, Petrovski R, Saunders CT, Chorny I, Kruglyak S, Margulies EH, Chuang HY, Kallberg M, Kumar SA, Liao A, Little KM, Stromberg MP, Tanner SW. **Isaac: ultra-fast whole-genome secondary analysis on Illumina sequencing platforms.** Bioinformatics 2013, 29(16), 2041-2043.

2. 2. 2. Isaac Variant Caller (IVC)

The Isaac Variant Caller identifies and genotypes single-nucleotide variants (SNVs) and small indels in the diploid genome case. The produced VCF file captures the genotype at each position, the probability that the consensus call differs from reference, and the probability of the called genotype.

More information can be found here:
LINK whitepaper_isaac_workflow.pdf

2. 2. 3. SnpEff - Annotation Tool

SnpEff is a variant annotation and effect prediction tool. It annotates and predicts the effects of variants on genes (such as amino acid changes). Using this tool, we follow the annotation cascade shown below.

(1) Gene annotation based on hg19 coordinates
(2) dbSNP138 ID mapping
(3) dbSNP142 ID mapping
(4) 1000 Genomes phase I release v3 mapping
(5) ESP6500 data mapping

More information can be found here:
LINK http://snpeff.sourceforge.net/SnpEff.html

2. 2. 4. Control-FREEC - Copy Number Variant Caller

Control-FREEC is a tool which enables automatic calculation of copy number and allelic content profiles, and consequently predicts regions of genomic alterations such as gains and losses. It accurately calls genotype status even when no control experiment is available. It also corrects for GC-content mappability biases of the polyploid genomes.

More information can be found here:

Boeva, V.; Popova, T.; Bleakley, K.; Chiche, P.; Cappo, J.; Schleiermacher, G.; Janoueix-Lerosey, I.; Delattre, O.; Barillot, E. **Control-freec: A tool for assessing copy number and allelic content using next-generation sequencing data.** Bioinformatics 2012, 28, 423-425.

2. 2. 5. Manta - Structural Variant Caller

Manta is a tool to call structural variants and indels from short paired-end sequencing reads. It combines paired-end and split read evidence during SV discovery and scoring to improve performance.

However, it does not require split reads or successful breakpoint assemblies to report a variant in cases where there is strong evidence of an imprecise variant. It provides genotypes and quality scores for variants in single diploid samples, and will also call somatic variants when a matched tumor sample is specified. Manta can detect all classes of structural variants which can be identified in the absence of copy number analysis and large-scale assembly.

This tool was developed specifically to work with Isaac alignment and its performance was verified in the recent ICGC-TCGA DREAM Mutation Calling Challenge.

LINK https://www.synapse.org/#!Synapse:syn312572

More information can be found here:

LINK https://github.com/StructuralVariants/manta

2. 3. Reference, Software and Tuned Parameters

2. 3. 1. Mapping Reference

hg19 from UCSC (original GRCh37 from NCBI, Feb. 2009)

2. 3. 2. Software Versions

Software	Version
Isaac aligner	01.15.02.08
Isaac variant caller	2.0.13
SnpEff	3.3
Manta	0.20.2
Control-FREEC	6.4

2. 3. 3. Tuned Parameters

Software	Parameter	Value	Remark
Isaac aligner	--base-quality-cutoff	15	3' end quality soft-clipping cutoff
	--keep-duplicates	1	Does not remove duplicated reads
	--default-adapters	AGATCGGAAGAGC*, *GCTCTTCCGATCT	
SnpEff	Source	hg19	
		dbSNP138, dbSNP142	
		1000 Genomes Phase 1 release v3	
		ESP6500	
Control -FREEC	forceGCcontent Normalization	1	Corrects the Read Count (RC) for GC-content bias
	ploidy	2	Genome ploidy
	sex	XY	Sample sex
	window	10000	Calculation window size
	mateOrientation	FR	FR: illumina paired-ends

- Software not listed in the table uses all default settings

3. Analysis Result

3. 1. Sample & Run Information

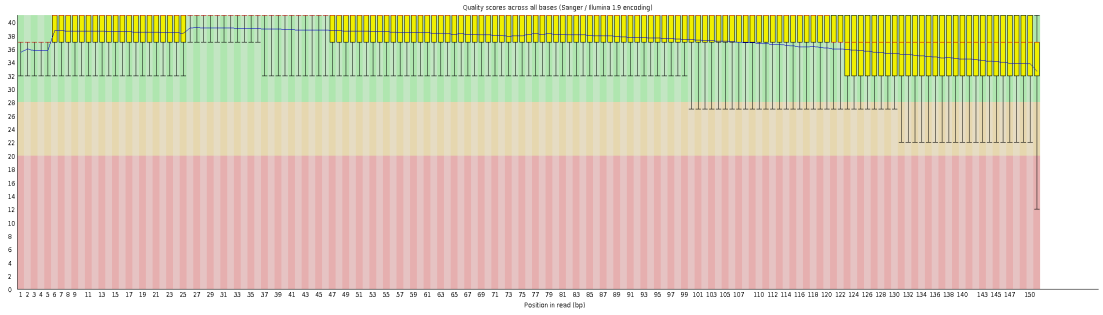
SampleID	NA12878
Project	0000KHX-0000
Instrument	HiSeq X
Read length	151

3. 2. Fastq

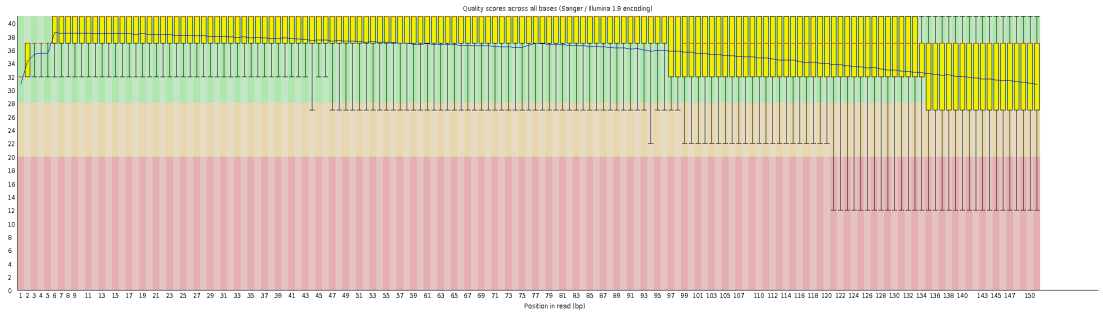
3. 2. 1. Statistics

TotalBases	ReadCount	GC(%)	Q20(%)	Q30(%)
131,587,455,402	871,440,102	41.70	94.55	88.05

3. 2. 2. Read1 Quality by Cycle



3. 2. 3. Read2 Quality by Cycle



3. 3. Pre-alignment Statistics

Total number of reads	871,440,102
Read length (bp)	150.00
Total yield (Mbp)	130,716
Reference size (Mbp)	2,858
Throughput mean depth (X)	45.70

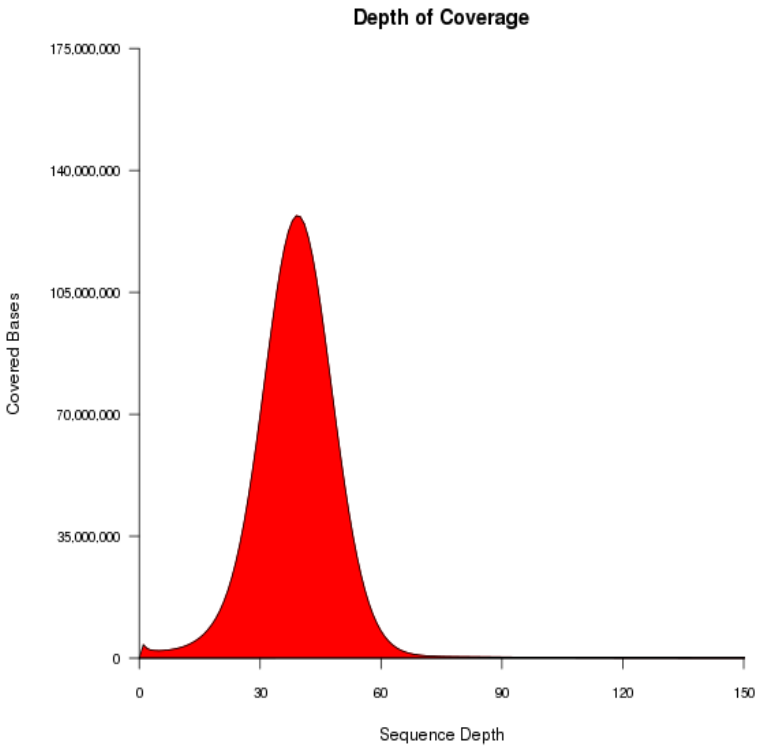
- Total yield: {total number of reads} * {read length}
- Reference size : Non-N human genome reference size
- Throughput mean depth: {total yield} / {reference size}

3. 4. Post-alignment Statistics

De-duplicated reads	817,524,016
De-duplicated reads %	93.80
Mappable reads (reads mapped to human genome)	743,439,268
Mappable reads % (out of de-duplicated reads)	90.90
Mappable yield (Mbp)	111,515
Mappable mean depth (X)	39.00

- Non-N human genome reference size : 2,858Mbp
- De-duplicated reads %: 100 * {number of de-duplicated reads} / {total number of reads}
- Mappable reads %: 100 * {number of mappable reads} / {number of de-duplicated reads}
- Mappable yield: {number of mappable reads} * {read length}
- Mappable mean depth (X): {mappable yield} / {reference size}

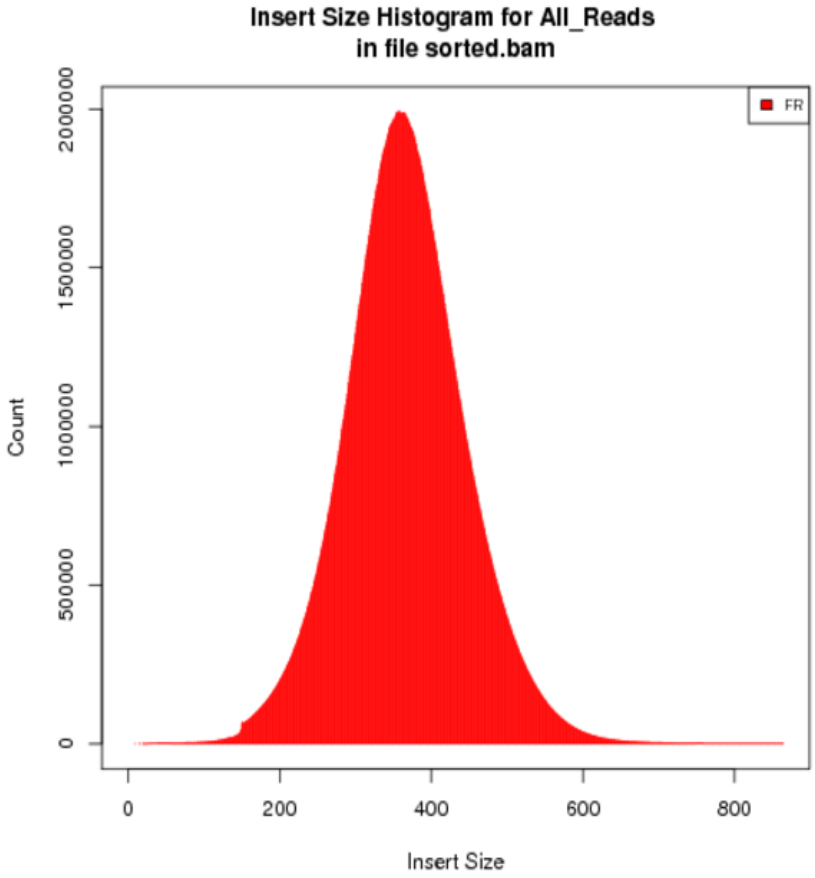
3. 5. Alignment Coverage



% Coverage	%>1X	%>5X	%>10X	%>15X	%>20X	%>30X
Value	98.7	98.4	97.9	97.3	95.8	84.2

- % Coverage : The percentage of bases in non-N reference regions with specific depth of coverage or greater

3. 6. Insert Statistics



Fragment length median	Standard deviation
364 bp	79.3 bp

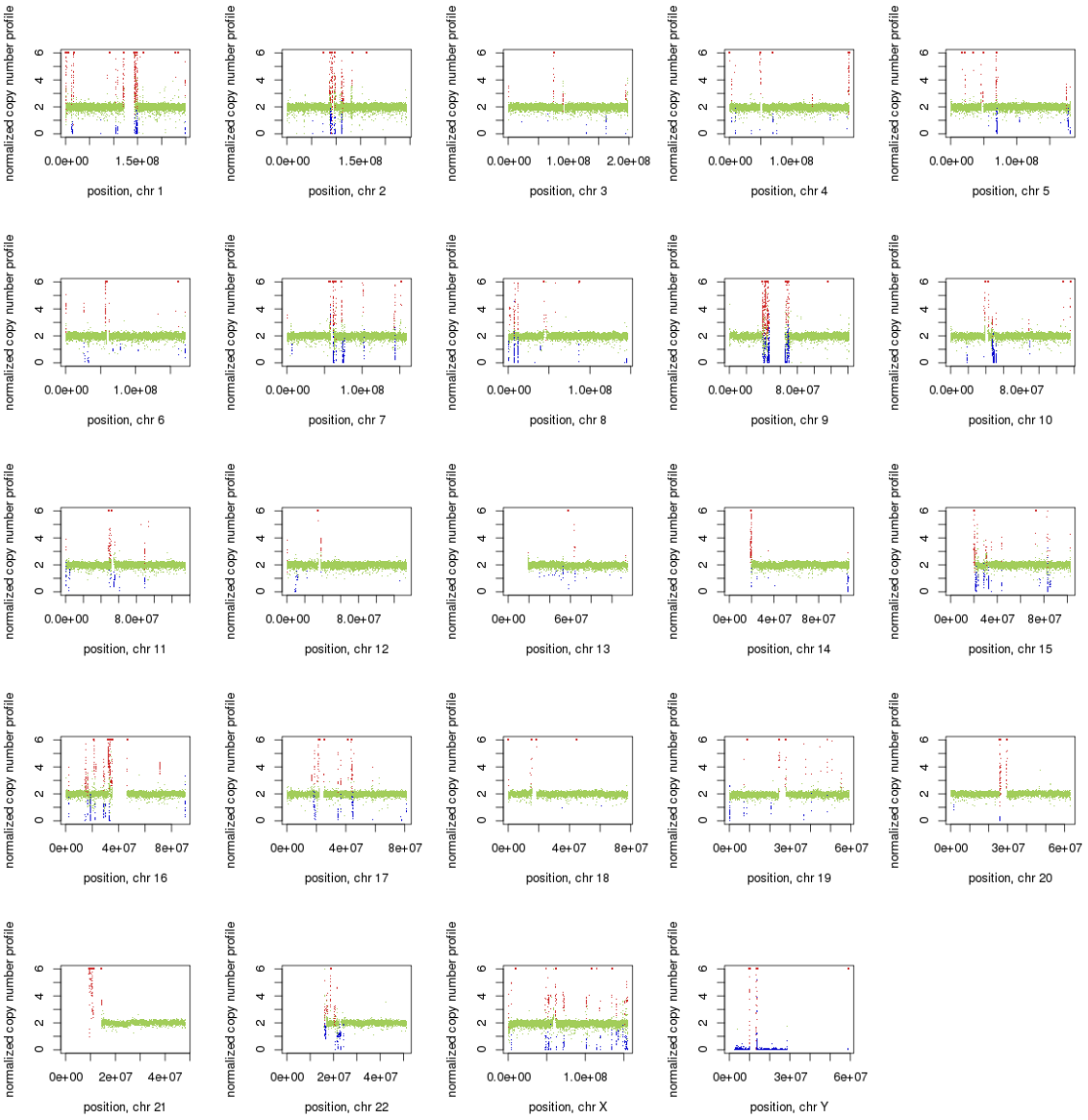
4. SNP & INDEL

	SNPs	Small insertions	Small deletions
# of variants	3,510,157	229,643	243,886
# of synonymous variants	11,241	-	-
# of non-synonymous variants	10,115	-	-
# of splicing variants	258		
# of stop gained	65		
# of stop loss	35		
# of frame shift	366		
% found in dbSNP138	98.1		
% found in dbSNP142	98.5		
Het/Hom ratio	1.69		
Ts/Tv ratio	2.0899		

- Het/Hom ratio : Ratio of Number of heterozygous variants to Number of homozygous variants.
- Ts/Tv ratio : Ratio of Transition rate of SNVs that pass the quality filters divided by transversion rate of SNVs that pass the quality filters. Transition rate of SNVs that pass the quality filters divided by transversion rate of SNVs that pass the quality filters. Transitions are interchanges of purines (A , G) or of pyrimidines (C, T). Transversions are interchanges between purine and pyrimidine bases (for example, A to T).

5. Copy Number Variant (CNV)

Copy number gains (>2)	564
Copy number losses (<2)	263



6. Structural Variant (SV)

SV type	# of variants
Duplications	84
Insertions	1,870
Deletions	5,502
Inversions	101
Translocations	98

- Duplication: a section of DNA is duplicated and both copies end up in the same chromosome
- Insertion: extra base pairs are inserted into DNA sequence
- Deletion: a section of DNA is lost, or deleted
- Inversion: a section of DNA is put in backwards
- Translocation: two non-homologous chromosomes exchange sections of DNA

7. Data Deliverables

7. 1. Deliverables List

File	Description
NA12878_R1.fastq.gz*	Raw read1 sequence data
NA12878_R2.fastq.gz*	Raw read2 sequence data
NA12878_sorted.bam	Isaac alignment file
NA12878_sorted.bam.bai	Isaac alignment index file
NA12878_[chr*].xlsx	Convert SNP_INDEL.vcf to excel
0000KHX-0000_NA12878_SNP_INDEL.vcf	SNP/INDEL result
0000KHX-0000_NA12878_CNVs.txt	Control-FREEC CNV result
0000KHX-0000_NA12878_SV.vcf	Manta SV result
0000KHX-0000_NA12878.pdf	Analysis report

- FASTQ files are saved compressed in the GNU zip format, an open source data compression program.

7. 2. Deliverables File Format

7. 2. 1. Fastq

7. 2. 1. 1. FASTQ Format Example:

```
@ST-E00104:157:H03N0ALXX:1:1101:2837:1309 1:N:0:3
AAAACAACTCCCCTGGTATGATATATATTTGAGCAGAATTTATAATTTCAC
+
AAFFFJJJJJJFJJJJJJJFJJJJJJJJJJFJAJJJJJJJJJJJJJJJJJ
```

7. 2. 1. 2. FASTQ File Consists of Four Lines:

- Line1: Sequence identifier
- Line2: Nucleotide sequences
- Line3: Quality score identifier line - character '+'
- Line4: Quality score

7. 2. 1. 3. Phred Scores

$Q = -10 \log_{10}(\text{error rate})$

PhredQualityScore	Probability of in-correct base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10000	99.99%

- Encoding: ASCII Character Code=Phred Quality Value + 33

7. 2. 1. 4. Quality Score Bins for Optimized 8-Level Mapping

Q score of HiSeq X Ten system : Q scores have been calibrated specifically to the HiSeq X Ten system and its consumables. It does use Q score binning. This is necessary for HiSeq X Ten runs due to the quantity of data being generated and since it cannot be turned off. Please refer to this table below, Q Scores for HiSeq X Ten are binned using the following criteria.

Q-Score Bins	Example of Empirically Mapped Q-Scores
N (no call)	N (no call)
2-9	7
10-19	11
20-24	22
25-29	27
30-34	32
35-39	37
40-45	42

- The quality score table above is typically updated when significant characteristics of the sequencing platform change, such as new hardware, software, or chemistry versions.

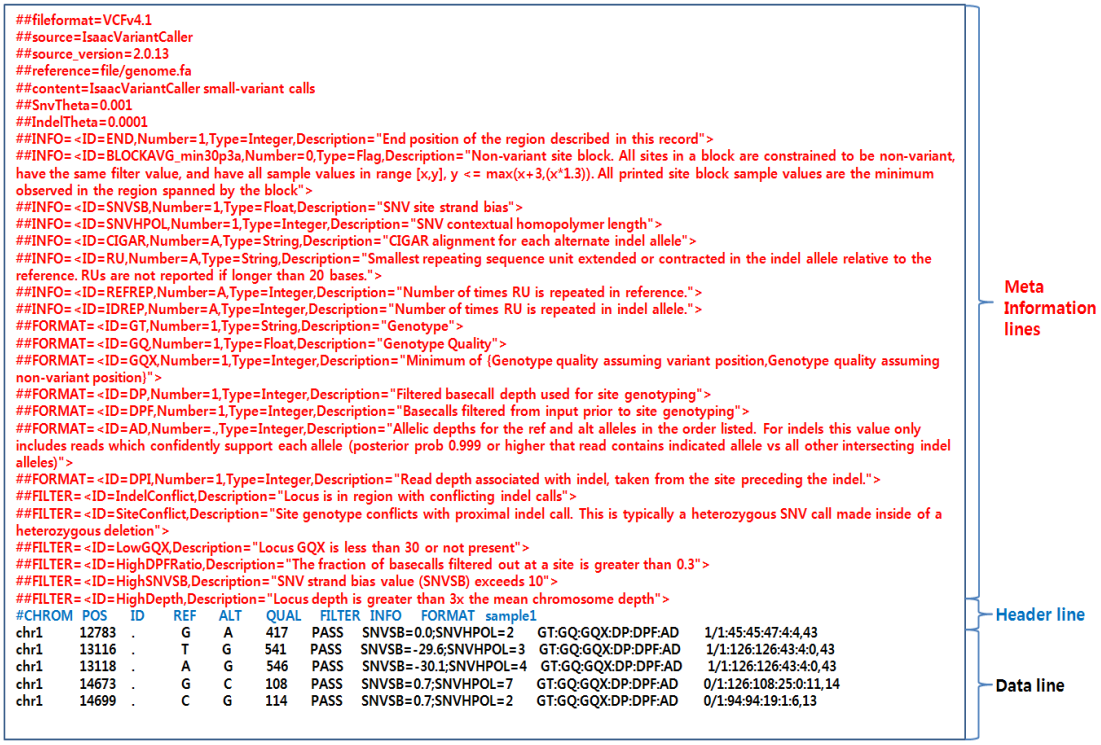
More information can be found here:

LINK http://support.illumina.com/help/SequencingAnalysisWorkflow/Content/Vault/Informatics/Sequencing_Analysis/CASAVA/swSEQ_mCA_FASTQFiles.htm

7. 2. 2. VCF

The Variant Call Format (VCF) is a text file format that contains information about variants found at specific positions in a reference genome. The file format consists of meta-information lines, a header line, and data lines. Each data line contains information about a single variant.

Example :



7. 2. 2. 1. Header Line

header	Description
#CHROM	Chromosome
POS	Position (with the 1st base having position 1)
ID	The dbSNP rs identifier of the SNP
REF	Reference base(s)
ALT	Comma separated list of alternate non-reference alleles called on at least one of the samples
QUAL	A Phred-scaled quality score assigned by the variant caller. Higher scores indicate higher confidence in the variant (and lower probability of errors).
FILTER	See FILTER tag table for possible entries.
INFO	See INFO tag table for possible entries.
FORMAT	See FORMAT tag table for possible entries.

7. 2. 2. 2. FILTER Tag

Tag	Description
IndelConflict	Locus is in region with conflicting indel calls
SiteConflict	Site genotype conflicts with proximal indel call, typically a heterozygous SNV call made inside of a heterozygous deletion
LowGQX	Locus GQX is less than 30 or not present
HighDPFRatio	The fraction of base calls filtered out at a site is greater than 0.4
HighSNVSB	SNV strand bias value (SNVSB) exceeds 10
HighDepth	Locus depth is greater than 3 times the mean chromosome depth

7. 2. 2. 3. INFO Tag

Tag	Description
SNVSB	SNV site strand bias
SNVHPOL	SNV contextual homopolymer length
CIGAR	CIGAR alignment for each alternate indel allele
RU	Smallest repeating sequence unit extended or contracted in the indel allele relative to the reference. RUs longer than 20 bases are not reported.
REFREP	Number of times RU is repeated in reference.
IDREP	Number of times RU is repeated in indel allele.
END	End position of the region described in this record
BLOCKAVG _min30p3a	Non-variant site block. All sites in a block are constrained to be non-variant, have the same filter value, and have all sample values in range [x,y], y <= max(x+3,(x*1.3)). All printed site block sample values are the minimum observed in the region spanned by the block

7. 2. 2. 4. FORMAT Tag

Tag	Description
GQX	Minimum of {Genotype quality assuming variant position, Genotype quality assuming non-variant position}
GT	Genotype 0/0 - the sample is homozygous reference 0/1 - the sample is heterozygous, carrying 1 copy of each of the REF and ALT alleles 1/1 - the sample is homozygous alternate
GQ	Genotype Quality
DP	Filtered base call depth used for site genotyping
DPF	Base calls filtered from input before site genotyping
AD	Allelic depths for the ref and alt alleles in the order listed. For indels, this value only includes reads that confidently support each allele (posterior probability 0.999 or higher that read contains indicated allele vs all other intersecting indel alleles)
DPI	Read depth associated with indel, taken from the position preceding the indel.

More information can be found here:
LINK www.broadinstitute.org/gatk/guide/article?id=1268

7. 2. 3. CNVs File

CNVs file is a tab delimited text file format that contains coordinates of predicted copy number alterations.

Information for each column:

- chromosome
- start position
- end position
- predicted copy number
- type of alteration
- gene

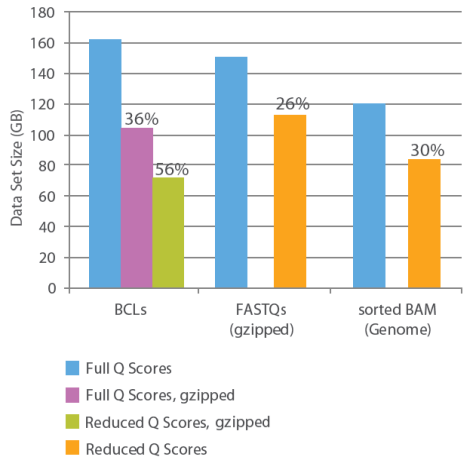
Appendix. Frequently Asked Questions (FAQs)

(Q1) Base qualities shown in HiSeq X FASTQ files look different from those generated by HiSeq 2000/2500.

(A1) This is necessary for HiSeq X runs due to the huge quantity of data being generated. HiSeq X system utilized Q score binning method to reduce result file sizes.

Quality Score Bins	Example of Empirically Mapped Quality Scores*
N (no call)	N (no call)
2–9	6
10–19	15
20–24	22
25–29	27
30–34	33
35–39	37
≥ 40	40

By replacing the quality scores between 19 and 25 with a new score of 22, data storage space is conserved.
*The mapped quality score of each bin (except "N") is subject to change depending on individual Q-tables.



Significant data size reductions in a 43× human genome data set can be accomplished with compression (gzip) and reduced quality scores. The percentages shown are in comparison to the data file sizes of full resolution genomes.

LINK whitepaper_datacompression.pdf

(Q2) HiSeq X system with Issac aligner + Isaac variant caller is a little bit new to me. Does it show compatible performance compared to HiSeq 2000/2500 with BWA aligner + GATK (or SAMtools)?

(A2) We benchmarked the Hiseq X system with Isaac+IVC by sequencing one HapMap sample (NA12878).

Platform	HiSeq 2000	HiSeq 2000	HiSeq X
Aligner + variant caller	BWA + SAMtools	BWA + GATK3.0-0	iSAAC + IVC
SNPs called (% dbSNP138 | dbSNP135)	3,759,251 (98.6% | 97%)	3,909,016 (98.2% | 96.4%)	3,915,275 (97% | 95.3%)
INDELs called (% dbSNP138 | dbSNP135)	602,778 (11.5% | 10.1%)	778,686 (82.1% | 69.8%)	576,128 (88.8% | 77.1%)
GIAB SNP sensitivity & precision	97% | 74.5%	98% | 72.4%	97.5% | 71.8%
GIAB INDEL sensitivity & precision	3.6% | 2.6%	74.7% | 42.3%	67.2% | 51.4%

(Q3) How good are base qualities produced by HiSeq X? Are they comparable with those produced by HiSeq 2000/2500?

(A3) Here are "% of ≥Q30 bases" values estimated by our real data together with Illumina specifications.

System	Read length	% of ≥ Q30 (illumina specification)	% of ≥ Q30 (real data)		
			Avg.	Median	Stdev
HiSeq 2000	2 x 101 bp	≥85%	87.4%	88.2%	2.5%
HiSeq 2500	2 x 151 bp	≥75%	89.2%	89.1%	0.6%
HiSeq X	2 x 151 bp	≥75%	89.2%	89.1%	0.6%
HiSeq X	2 x 101 bp	-	93.6%	93.6%	0.4%

According to our HiSeq X data, % of ≥Q30 bases from HiSeq X is comparable with HiSeq 2000/2500 even with the original length 151.

Research use only

Macrogen Korea

10F, 254 Beotkkot-ro,
Geumcheon-gu, Seoul
Rep. of Korea
Phone : +82-2113-7000

Contact

Web : www.macrogen.com
Lims : http://dna.macrogen.com