

Next Generation Sequence Alignment on the BRC Cluster

Steve Newhouse
22 July 2010

Overview

- Practical guide to processing next generation sequencing data on the cluster
- No details on the inner workings of the software/code & theory
- Taken from the 1000 genomes project pipeline developed by the Broad & Wellcome Trust
- Focus on Illumina paired-end sequence data
- Raw data format: FASTQ files
- Quality checking
- Quality filtering
- Short read alignment to the genome
- SNP calling
- *This is one way processing the data that works well*

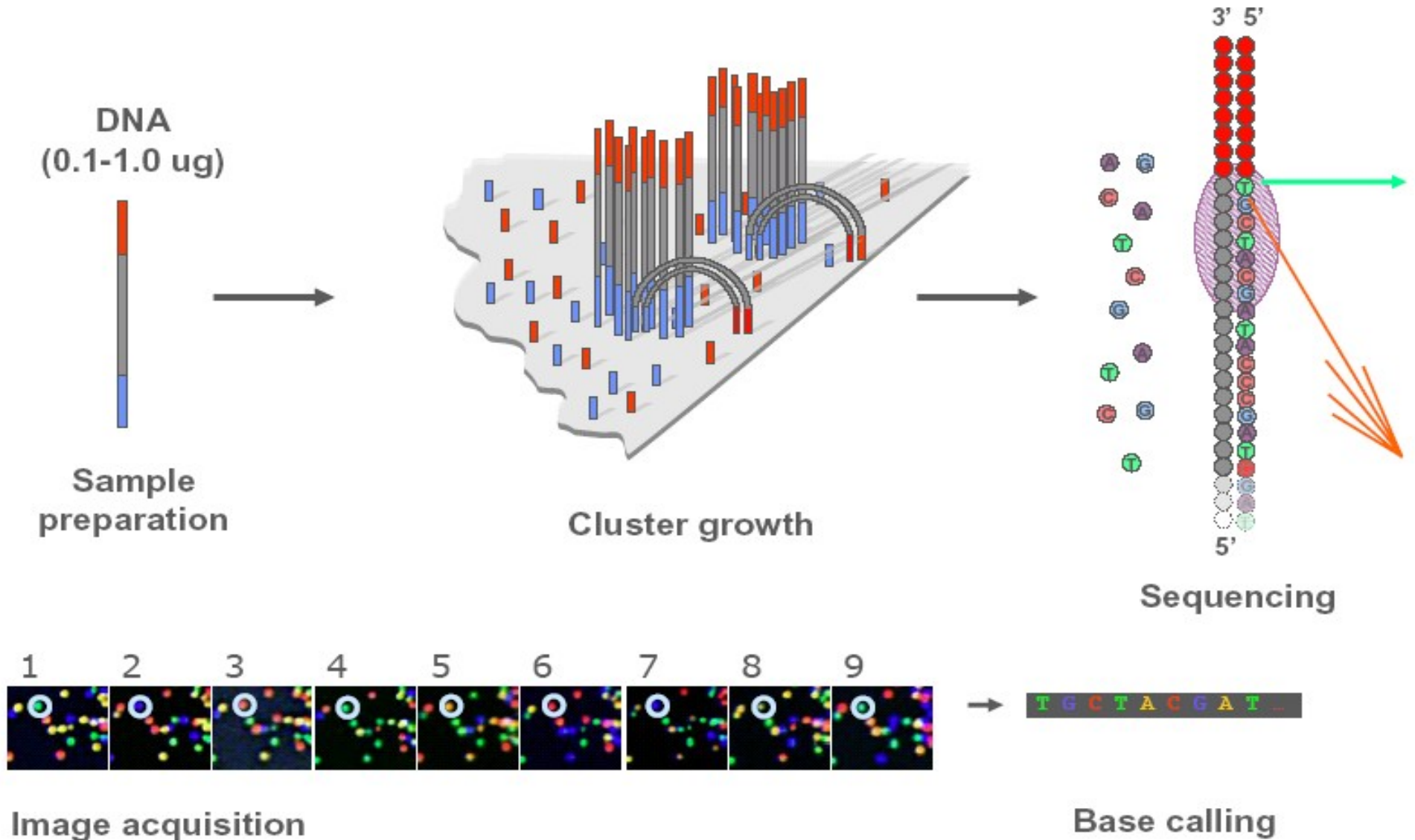
Main Tools/Resources

- BRC Cluster Software : <http://compbio.brc.iop.kcl.ac.uk/cluster/software.php>
- Maq: <http://maq.sourceforge.net/>
- Fastqc : <http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/>
- Fastx: http://hannonlab.cshl.edu/fastx_toolkit/
- cmpfastq.pl : <http://compbio.brc.iop.kcl.ac.uk/compbio-dev/software/cmpfastq.php>
- BWA: <http://bio-bwa.sourceforge.net/bwa.shtml>
- Genome Analysis Toolkit:
http://www.broadinstitute.org/gsa/wiki/index.php/The_Genome_Analysis_Toolkit
- PICARD TOOLS: <http://picard.sourceforge.net/>
- SAMTOOLS: <http://samtools.sourceforge.net/>
- FASTQ Files : http://en.wikipedia.org/wiki/FASTQ_format
- SAM/BAM Format : <http://samtools.sourceforge.net/SAM1.pdf>
- PHRED Scores: http://en.wikipedia.org/wiki/Phred_quality_score
- Next Generation Sequencing Library: <http://ngslib.genome.tugraz.at/>

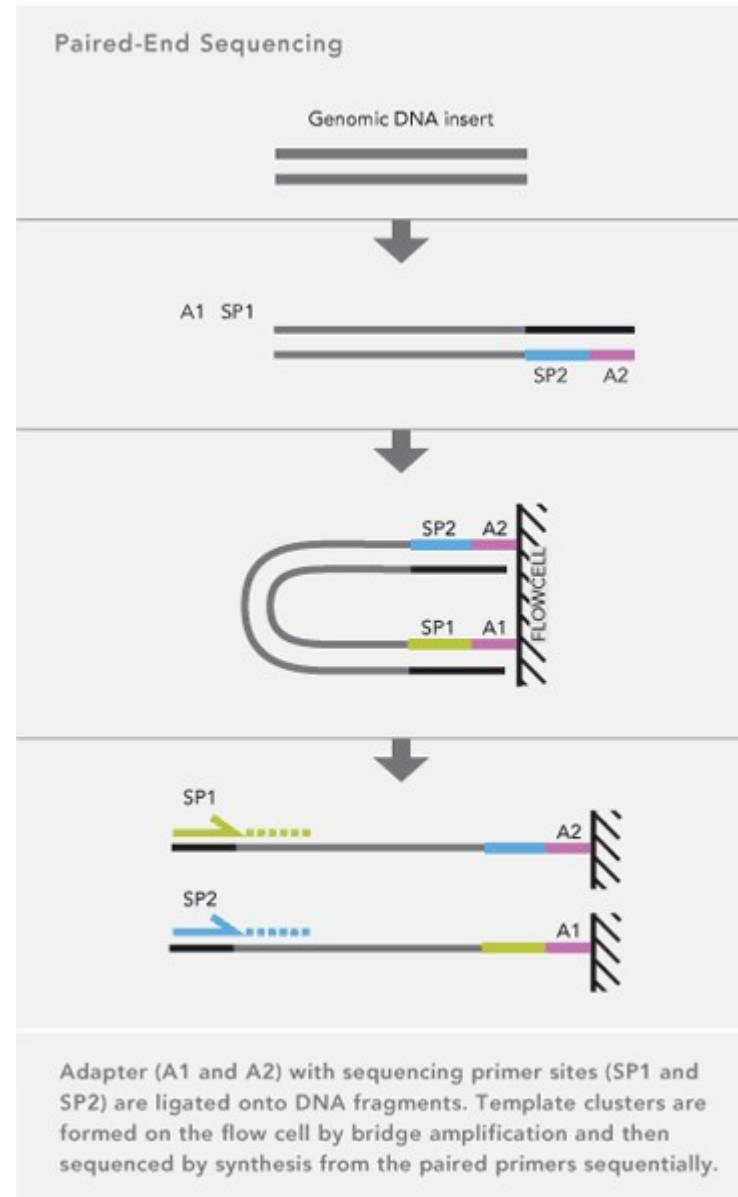
Main Tools/Resources

- BRC Cluster Software : <http://compbio.brc.iop.kcl.ac.uk/cluster/software.php>
- Maq: <http://maq.sourceforge.net/>
- Fastqc : <http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/>
- Fastx: http://hannonlab.cshl.edu/fastx_toolkit/
- cmpfastq.pl : <http://compbio.brc.iop.kcl.ac.uk/compbio-dev/software/cmpfastq.php>
- BWA: <http://bio-bwa.sourceforge.net/bwa.shtml>
- Genome Analysis Toolkit:
http://www.broadinstitute.org/gsa/wiki/index.php/The_Genome_Analysis_Toolkit
- PICARD TOOLS: <http://picard.sourceforge.net/>
- SAMTOOLS: <http://samtools.sourceforge.net/>
- FASTQ Files : http://en.wikipedia.org/wiki/FASTQ_format
- SAM/BAM Format : <http://samtools.sourceforge.net/SAM1.pdf>
- PHRED Scores: http://en.wikipedia.org/wiki/Phred_quality_score
- Next Generation Sequencing Library: <http://ngslib.genome.tugraz.at/>

Illumina sequencing technology



Paired-end Sequencing Assay



What does the raw data look like?

- Fastq Format : *_sequence.txt
- Text file storing both nucleotide sequence and quality scores.
- Both the sequence letter and quality score are encoded with a single ASCII character for brevity.
- Standard for storing the output of high throughput sequencing instruments such as the Illumina Genome Analyzer
- Details : http://en.wikipedia.org/wiki/FASTQ_format

FASTQ Format

```
@315ARAAXX090414:8:1:567:552#0
TGTTTCTTTAAAAAGGTAAGAATGTTGTTGCTGGGCTTAGAAATATGAATAACCATATGCCAGATAGATAGATGGA
+
;<<=<=====::==>====<<<;;;:::999999888877666555444333222211111000//
@315ARAAXX090414:8:28:131:1063#0
TGTTTCTTTAAAAAGGTAAGAATGTTGTTGCTGGGCTTAGAAATATGAATAACCATATGCCAGATAGATAGATGGA
+
;<<=====<=<<==8<=<====<<<;;;::8::99988558787666535544443332202.101/0.00.-
@315ARAAXX090414:8:67:1150:1651#0
TGTTTCTTTAAAAAGGTAAGAATGTTGTTGCTGGGCTTAGAAATATGAATAACCATATGCCAGATAGATAGATGGA
+
;<<<====<>=<69==>6==;9==<<<8:;:::697998936577774464635441332222/11110.0/000.
```

@315ARAAXX090414:8:1:567:552#0

- **@315ARAAXX090414**: the unique instrument name
- **8**: flowcell lane
- **1**: tile number within the flowcell lane
- **567**: 'x'-coordinate of the cluster within the tile
- **552**: 'y'-coordinate of the cluster within the tile
- **#** :index number for a multiplexed sample (0 for no indexing)
- **0** :the member of a pair, /1 or /2 (paired-end or mate-pair reads only)

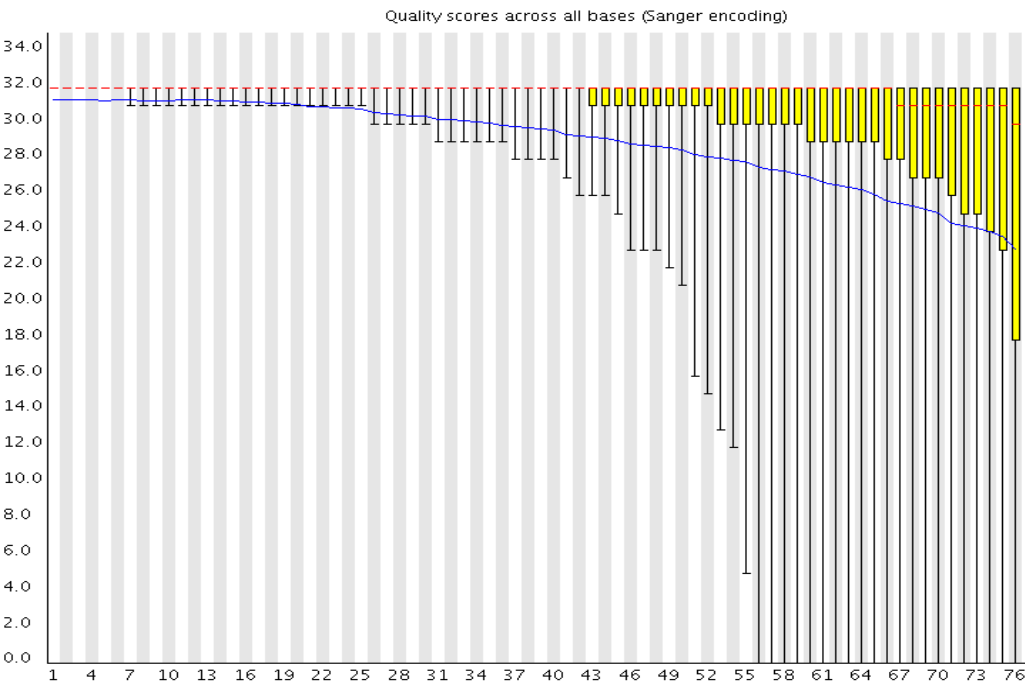
Quality Control & Pre-processing

Quality Control/Pre-processing 1

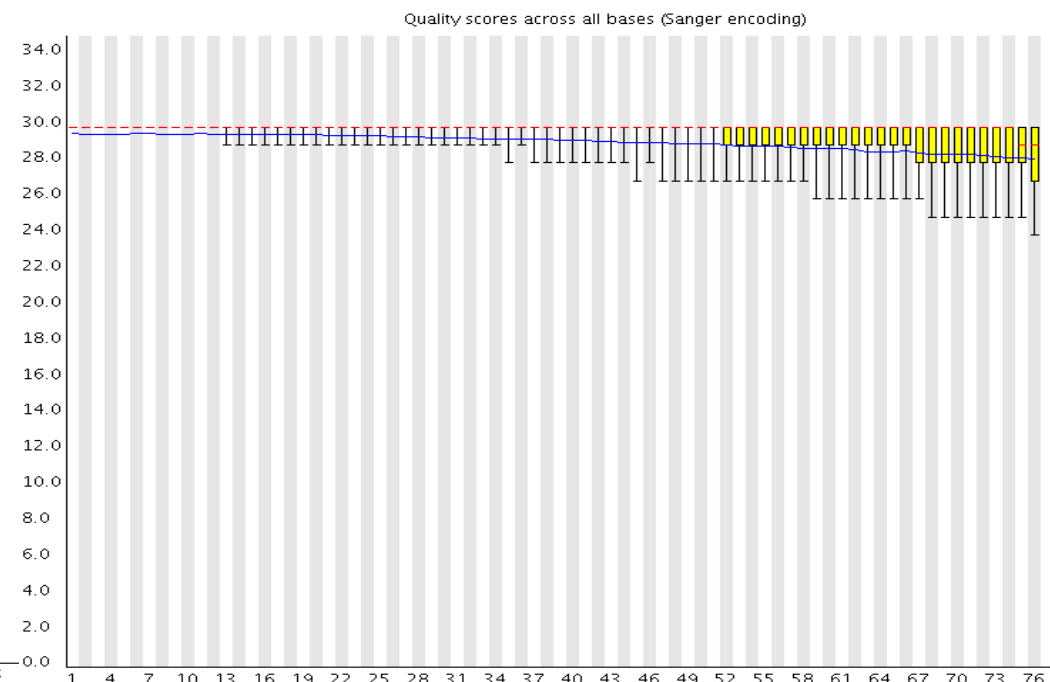
- **Input:** Raw *.sequence.txt (fastq) files
- **Step 1:** convert Illumina quality scores to Sanger Phred quality score
 - `maq ill2sagner s_1_1_sequence.txt s_1_1_sequence.fastq`
 - `maq ill2sagner s_1_2_sequence.txt s_1_2_sequence.fastq`
- **Step 2:** quality control check on raw sequence data
 - `fastqc s_1_1_sequence.fastq`
 - `fastqc s_1_2_sequence.fastq`
- **Step 3:** quality control filter raw sequence data [-Q 33, -Q 64 undocumented options]
 - `cat s_1_1_sequence.fastq | \`
 - `fastx_clipper -Q 33 -l 20 -v -a ACACTCTTTCCCTACACGACGCTCTTCCGATCT | \`
 - `fastx_clipper -Q 33 -l 20 -v -a CGGTCTCGGCATTCTACTGAACCGCTCTTCCGATCT | \`
 - `fastx_clipper -Q 33 -l 20 -v -a`
`AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATC | \`
 - `fastx_clipper -Q 33 -l 20 -v -a`
`CAAGCAGAAGACGGCATAACGAGATCGGTCTCGGCATTCTGCTGAACCGCTCTTCCGATC | \`
 - `fastq_quality_trimmer -Q 33 -t 20 -l 20 -v | \`
 - `fastx_artifacts_filter -Q 33 -v | \`
 - `fastq_quality_filter -Q 33 -q 20 -p 50 -v -o s_1_1_QC.fastq`

Quality Control/Pre-processing 2

- **Step 4:** quality control check on QC'd data
 - `fastqc s_1_1_QC.fastq`
 - `fastqc s_1_2_QC.fastq`



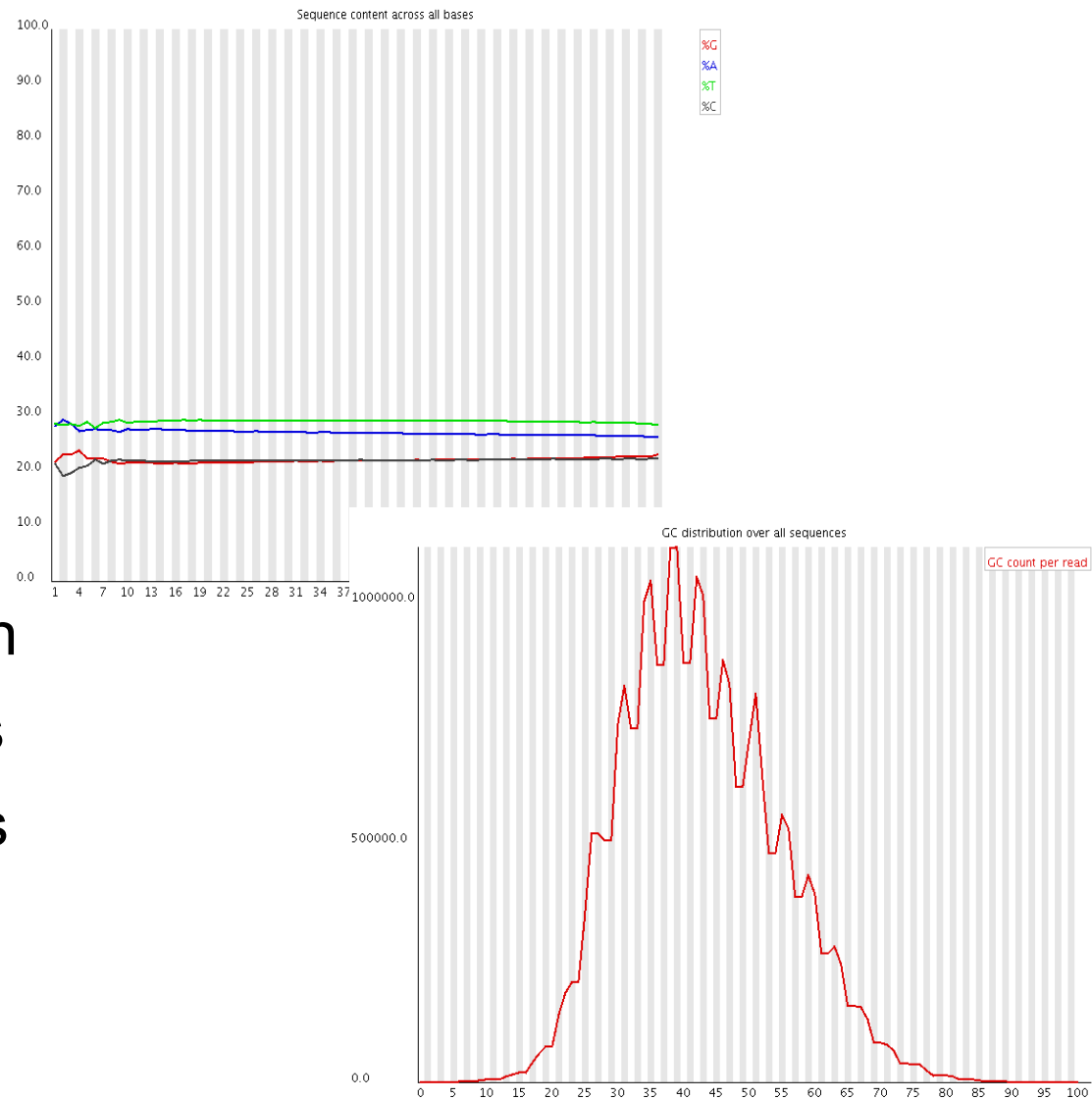
Total Sequences 28900871



Total Sequences 17949113

FASTQC OUTPUT : QC Checks

- Basic Statistics, Total Sequences, Sequence length, %GC
- Per base sequence quality
- Per sequence quality scores
- Per base sequence content
- Per base GC content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Over-represented sequences



FASTX OUTPUT : QC filtering

Clipping Adapter: GATCGGAAGAGCGGTTCAGCAGGAATGCCGAG

Min. Length: 20

Input: 28900871 reads.

Output: 26932037 reads.

discarded 1346166 too-short reads.

discarded 83269 adapter-only reads. # NB!

discarded 539399 N reads # NB!

Input: 18234157 reads.

Output: 17977009 reads.

discarded 257148 (1%) too-short reads.

Input: 17977009 reads.

Output: 17972702 reads.

discarded 4307 (0%) artifact reads.

Quality cut-off: 20

Minimum percentage: 50

Input: 17972702 reads.

Output: 17949113 reads.

discarded 23589 (0%) low-quality reads.

Quality Control/Pre-processing 3

- **Step 5:** compare QCd fastq files
 - One end of each read could be filtered out in QC
 - BWA cant deal with mixed PE & SE data
 - Need to id reads that are still paired after QC
 - Need to id reads that are no longer paired after QC
- `perl cmpfastq.pl s_1_1_QC.fastq s_1_2_QC.fastq`
 - Reads matched on presence/absence of id's in each file :
 - `s_1_1_*: @315ARAAXX090414:8:1:567:552#1`
 - `s_1_2_*: @315ARAAXX090414:8:1:567:552#2`
- **Output:** 2 files for each *QC.fastq file
 - `s_1_1_QC.fastq-common-out` (reads in `s_1_1_*` == reads in `s_1_2_*`)
 - `s_1_1_QC.fastq-unique-out` (reads in `s_1_1_*` not in `s_1_2_*`)
 - `s_1_2_QC.fastq-commont-out` (reads in `s_1_1_*` == reads in `s_1_2_*`)
 - `s_1_2_QC.fastq-unique-out` (reads in `s_1_2_*` not in `s_1_1_*`)

Next Generation sequence Alignment

What does sequence mapping do?

Find all the matches for a read in the genome

A DNA Sequence: ACATAGGATCATGAAGTACCCATATCTAGTGGG
reads: AGGA, CATC, ATAT, TTTG, ATGT

Matched Results: ACATAGGATCATGAAGTACCCATATCTAGTGGG
Perfect Match: AGGA
1bp Mismatch: CATC CATC CATC
1bp Indel: A-TGT

Mismatch: Potential variants or quality issues

BWA: Burrows-Wheeler Aligner requirements

- **Raw data** : QC filtered *.fastq files
- **Reference genome** : BWA indexed
- **Location**:
/scratch/data/reference_genomes/human
- **Available genomes**
 - Homo_sapiens_assembly18.fasta
 - human_b36_both.fasta
 - **human_g1k_v37.fasta (1000 genomes)**
 - Bowtie index available

Running BWA on the BRC cluster

- **Step 1:** Index the genome
 - `bwa index -a bwtsv human_g1k_v37.fasta` (already done for you)
- **Step 2:** Generate alignments:
 - `REF=/scratch/data/reference_genomes/human/human_g1k_v37.fasta`
 - `bwa aln -t 8 $REF read.1.fastq > read.1.sai`
 - `bwa aln -t 8 $REF read.2.fastq > read.2.sai`
 - No QC filtering prior to alignment: option -q15 if the quality is poor at the 3' end of reads
 - Multi threading option : -t N (~ 20 mins to align 15 mill reads with -t option)
- **Step 3:** Generate alignments in the SAM format:
 - `bwa sampe $REF read.1.sai read.2.sai read.1.fastq read.2.fastq > aln.sam`
- **Details SAM/BAM Format :** <http://samtools.sourceforge.net/SAM1.pdf>

Processing SAM/BAM files

- **samtools/PICARD/GATK**
 - 1. Remove duplicates
 - 2. Base Quality Score Recalibration
 - 3. Indel realignment
 - 4. Call SNPs/INDELS
- Many other quality stats/options for processing files using these tools : see web documentation
- **Genome Analysis Toolkit:**
http://www.broadinstitute.org/gsa/wiki/index.php/The_Genome_Analysis_Toolkit
- **PICARD TOOLS:** <http://picard.sourceforge.net/>
- **SAMTOOLS:** <http://samtools.sourceforge.net/>

Remove duplicates

- **Picard tools : MarkDuplicates**

- Examines aligned records in the supplied SAM or BAM file to locate duplicate molecules.
- `PICARD=/home/snewhousebrc/picard-tools-1.24`
- `java -Xmx4g -jar $PICARD/MarkDuplicates.jar \`
- `INPUT=$RUN.bam \`
- `OUTPUT=$RUN.dupermv.bam \`
- `METRICS_FILE=$RUN.METRICS.FILE \`
- `REMOVE_DUPLICATES=true \` # if false : all records are written to output file with duplicate records flagged
- `TMP_DIR=./tmp \` # NB! Important option as tool writes many files to local disk
- `VALIDATION_STRINGENCY=SILENT;`

Base Quality Score Recalibration

- Correct for variation in quality with machine cycle and sequence context
- Recalibrated quality scores are more accurate
- Closer to the actual probability of mismatching the reference genome

- Done by analysing the covariation among several features of a base
 - Reported quality score
 - The position within the read
 - The preceding and current nucleotide (sequencing chemistry effect) observed by the sequencing machine
 - Probability of mismatching the reference genome &
 - Known SNPs taken into account (dbSNP 131)

- Covariates are then used to recalibrate the quality scores of all reads in a BAM file

Base Quality Score Recalibration

1: CountCovariates

- **Step 1: CountCovariates**

GATK=/share/apps/GenomeAnalysisTK_1.0.3471/GenomeAnalysisTK.jar

REF=/scratch/data/reference_genomes/human/human_g1k_v37.fasta

ROD=/scratch/data/reference_genomes/human/dbsnp_131_g1k_b37.sjn.rod # dbSNP131

PICARD=/home/snewhousebrc/picard-tools-1.24

TMPDIR=./tmp

RUN=/home/snewhousebrc/NGS/read.1

```
java -Xmx10g -jar $GATK -T CountCovariates -R $REF --DBSNP $ROD \  
-I $RUN.dupermv.bam -recalFile $RUN.recal_data.csv --default_platform Illumina \  
--max_reads_at_locus 20000 \  
-cov ReadGroupCovariate \  
-cov QualityScoreCovariate \  
-cov CycleCovariate \  
-cov DinucCovariate;
```

Base Quality Score Recalibration

2: TableRecalibration

- **Step 2: TableRecalibration**

- After counting covariates rewrite quality scores using the data in the recal_data.csv file, into a new BAM file.

- `java -Xmx10g -jar $GATK -T TableRecalibration -R $REF \
-I $RUN.clean.bam \
-recalFile $RUN.recal_data.csv \
-outputBam $RUN.recalibrated.bam \
--default_platform Illumina; # NB`

- Index BAM files for GATK

- `samtools index $RUN.recalibrated.bam`

Local realignment around Indels

- Sequence aligners are unable to perfectly map reads containing insertions or deletions
 - Alignment artefacts
 - False positives SNPs
 - Unmapped reads : missing data!
- Steps to the realignment process:
 - Determining (small) suspicious intervals which are likely in need of realignment
 - Running the re-aligner over the intervals

Local realignment around Indels

Step 1 : Creating Intervals

- `java -Xmx10g -jar $GATK -T RealignerTargetCreator -R $REF -D $ROD \`
`-I $RUN.recalibrated.bam \`
`-o $RUN.forRealigner.intervals;`

Step 2 : Realigning

- `java -Djava.io.tmpdir=$TMPDIR -Xmx10g -jar $GATK -T IndelRealigner \`
`-R $REF -D $ROD \`
`-I $RUN.recalibrated.bam \`
`-targetIntervals $RUN.forRealigner.intervals \`
`--output $RUN.RecalRealn.bam ;`

Step 3 : Index

- `samtools index $RUN.RecalRealn.bam;`

SNP Calling : UnifiedGenotyper

GATK=/share/apps/GenomeAnalysisTK_1.0.3471/GenomeAnalysisTK.jar

REF=/scratch/data/reference_genomes/human/human_g1k_v37.fasta

ROD=/scratch/data/reference_genomes/human/dbsnp_131_g1k_b37.sjn.rod # dbSNP131

PICARD=/home/snewhousebrc/picard-tools-1.24

TMPDIR=./tmp

RUN=/home/snewhousebrc/NGS/read.1

- **Run UnifiedGenotyper**
- `java -Xmx10g -jar $GATK -T UnifiedGenotyper -R $REF -D $ROD \`
- `-I $RUN.RecalRealn.bam \`
- `-varout $RUN.snps.raw.vcf \`
- `-stand_call_conf 30.0;` # minimum phred-scaled Qscore threshold to separate high confidence from low confidence calls (that aren't at 'trigger' sites). Only genotypes with confidence \geq this threshold are emitted as called sites. A reasonable threshold is 30 (this is the default)

VCF FORMAT: http://1000genomes.org/wiki/doku.php?id=1000_genomes:analysis:vcfv3.2

Details : http://www.broadinstitute.org/gsa/wiki/index.php/Unified_genotyper

Inside VCF Files

1. CHROM chromosome: an identifier from the reference fasta file
2. POS position (1st base has position 1). Positions are sorted numerically, in increasing order, within each reference sequence CHROM.
3. ID A unique identifier where available, else '!'. If this is a dbSNP variant it is encouraged to use the rs number.
4. REF reference base. One of A,C,G,T,N
5. ALT comma separated list of alternate non-reference alleles
6. QUAL phred-scaled quality score for the assertion made in ALT.al
7. FILTER QC filter
8. INFO additional information, eg:
 - * AA ancestral allele, encoded as REF and ALT
 - * AC allele count in genotypes, for each ALT allele, in the same order as listed
 - * AN total number of alleles in called genotypes
 - * AF allele frequency for each ALT allele
 - * DP depth, e.g. D=154
 - * MQ RMS mapping quality, e.g. MQ=52
 - * NS Number of samples with data
 - * DB dbSNP membership
9. GT genotype & genotype information eg :
 - * GQ genotype quality, encoded as a phred quality $-10\log_{10}p(\text{genotype call is wrong})$, max quality 99
 - * DP read depth at this position for this sample
 - * FT sample genotype filter indicating if this genotype was “called”

INDEL Calling:IndelGenotyperV2

GATK=/share/apps/GenomeAnalysisTK_1.0.3471/GenomeAnalysisTK.jar

REF=/scratch/data/reference_genomes/human/human_g1k_v37.fasta

ROD=/scratch/data/reference_genomes/human/dbsnp_131_g1k_b37.sjn.rod # dbSNP131

PICARD=/home/snewhousebrc/picard-tools-1.24

TMPDIR=./tmp

RUN=/home/snewhousebrc/NGS/read.1

- **java -Xmx10g -jar \$GATK -T IndelGenotyperV2 -R \$REF -D \$ROD **
- **-I \$RUN.RecalRealn.bam **
- **-O \$RUN.indels.raw.bed **
- **-o \$RUN.detailed.output.bed **

• Output

- chr1 556817 556817 +G:3/7
- chr1 3535035 3535054 -TTCTGGGAGCTCCTCCCCC:9/21
- chr1 3778838 3778838 +A:15/48

Details : http://www.broadinstitute.org/gsa/wiki/index.php/Indel_Genotyper_V2.0

