Initial Processing of Sequence Data: Read Mapping

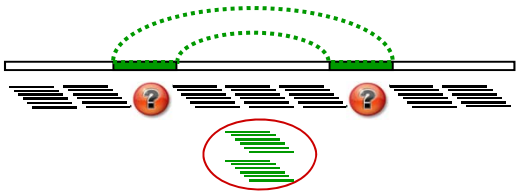
Gonçalo Abecasis, Hyun Min Kang

Massive Throughput Sequencing

- Tools to generate sequence data evolving rapidly
- Commercial platforms produce gigabases of sequence rapidly and inexpensively
  - ABI SOLiD, Illumina Solexa, Roche 454, Complete Genomics, and others…
- Sequence data consist of thousands or millions of short sequence reads with moderate accuracy
  - 0.5 – 1.0% error rates per base may be typical

Read mapping

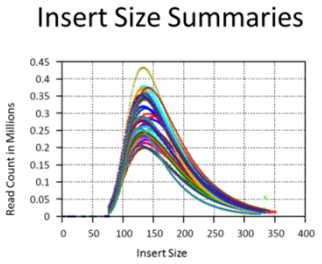
- Read mapping
- Quality recalibration
- Duplicate removal
- Indel realignment



FASTQ

Data QC

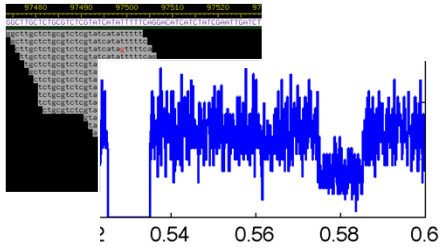
- Descriptive statistics
- Sequence quality checking
- Sample mix-ups



BAM/SAM

Variant calling

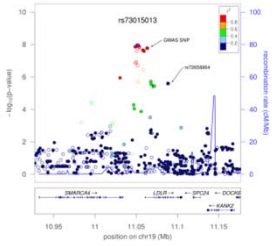
- SNPs, short indels
- Structural variants
- Genotype refinement



VCF

Variant interpretation

- Annotations: SIFT, PolyPhen
- Annotation with genome features
- Statistical association analysis



Knowledge

Shotgun Sequence Reads

- Typical short read might be <25-100 bp long and not very informative on its own

- Reads must be arranged (*aligned*) relative to each other to reconstruct longer sequences

Read Alignment

GCTAGCTGATAGCTAGCTAGCTGATGAGCCCGA

Short Read (30-100 bp)

5’-ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3’

Reference Genome (3,000,000,000 bp)

- The first step in analysis of human short read data is to align each read to genome, typically using a hash table based indexing procedure
- This process now takes no more than a few hours per million reads …
- Analyzing these data without a reference human genome would require much longer reads or result in very fragmented assemblies

Read Alignment – Food for Thought

- Typically, all the words present in the genome are indexed to facilitate read mapping …
  - What are the benefits of using short words?
  - What are the benefits of using long words?

- How matches do you expect, on average, for a 10-base word?
  - Do you expect large deviations from this average?

Shotgun Sequence Data

TAGCTGATAGCTAG**A**TAGCTGATGAGCCCGAT

ATAGCTAG**A**TAGCTGATGAGCCCGATCGCTGCTAGCTC

ATGCTAGCTGATAGCTAG**C**TAGCTGATGAGCC

AGCTGATAGCTAG**C**TAGCTGATGAGCCCGATCGCTG

GCTAGCTGATAGCTAG**C**TAGCTGATGAGCCCGA

Sequence Reads

5’-ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3’

Reference Genome

A/C

Predicted Genotype

From Mapped Reads to Genotype

- Each aligned read provides a small amount of evidence about the underlying genotype
  - Read may be consistent with a particular genotype …
  - Read may be less consistent with other genotypes …
  - A single read is never definitive
- For read evidence for (or against) a variant to be interpreted correctly, it is important that we are very confident of the alignment
- Common to apply multiple refinement steps to initial alignment

FASTQ

- Sequences with base qualities

```
@IL11_193:4:1:878:501
TATTTTGACTTTGAGCGTATCGAGGCTCTTTAACCTGAACGTCAG
+
IIIIIIIIIIIIII1IDII<IIIIIIIIIIIII(I&/97.,8&
```

- Sequences are stored together with associated base qualities
- Base qualities are integers (typically, ranging up to 40) and stored as characters for compactness

SAM/BAM

- Aligned sequence reads

```
@HD VN:1.0
@SQ SN:chr20 LN:62435964
@RG ID:L1 PU:SC_1_10 LB:SC_1 SM:NA12891
@RG ID:L2 PU:SC_2_12 LB:SC_2 SM:NA12891

read_29006_6945 99 chr20 28833 20 3M1D25M = 28993 195 \
AGCTTATCTTGGTCTTGGCCG <<<<<<<<<:<9/,&,22;;<<< RG:Z:L1

read_28881_323b 147 chr20 28934 30 35M = 28701 -268 \
ACCTATATCTGCGGCCTTGCA <<<<<<<<7;:<<<6;<<<<7<< RG:Z:L2
```

- Header describes alignment and sample information
- Each row describes alignment for a single short read
- Can be viewed more conveniently with **samtools tview** or **IGV**

- http://samtools.sourceforge.net/

VCF

- Variants and Genotypes

```
##fileformat=VCFv4.0
##fileDate=20100721
##reference=NCBI36
##INFO= <ID=AA, Number=1, Type=String, Description="Ancestral Allele">
##INFO= <ID=H2, Number=0, Type=Flag, Description="HapMap2 membership">
##FORMAT=<ID=GT, Number=1, Type=String, Description="Genotype">
##FORMAT=<ID=GQ, Number=1, Type=Integer, Description="Genotype Quality">
##FORMAT=<ID=DP, Number=1, Type=Integer, Description="Read Depth">

#CHROM POS ID   REF ALT  QUAL FILTER INFO    FORMAT SAMPLE1 SAMPLE2

1      50  rs12 A   G    100  PASS   H2,AA=G GT:DP   1|0:16 2|2:20
1      77  rs3  C   T    100  PASS   H2;AA=T GT:DP   0|1|5  2|2|20
1      90  .    ACG A,AC 40   PASS   .       GT:DP   1|1:13 2|2:29
```

- Header describes file contents and free form fields
- Each row describes information on a single variant site (and optionally, genotypes)

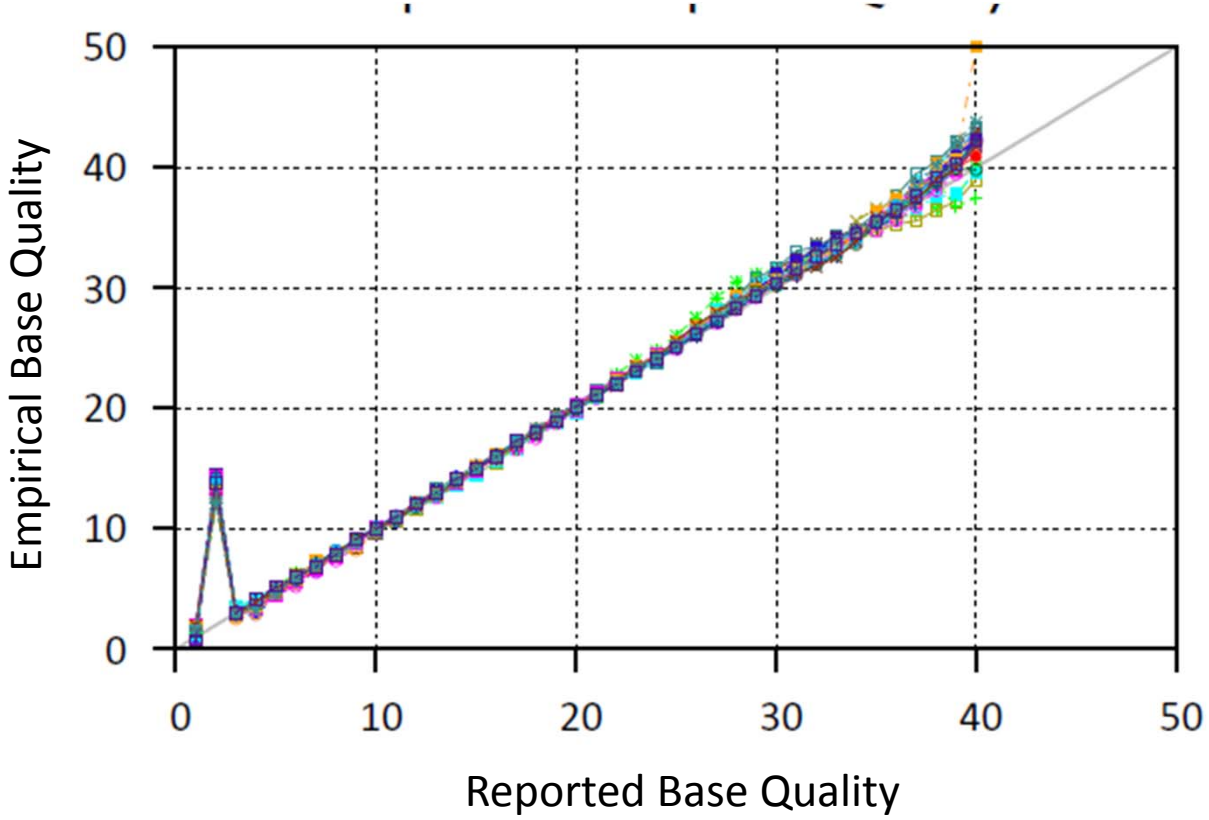
Mapping Quality Score

- Are we confident the read is placed correctly?
- Compare placement to alternative placements.

$$MapQ = -10log_{10}(\left(1 - \frac{P(read|best\ placement)}{\sum_i P(read|placement\ i)}\right)$$

- Reads with low mapping quality are typically excluded from variant calling.

Base Quality Calibration



Compare reported base qualities with actual mismatch rates.
Adjust reported base qualities to ensure they are nicely calibrated, as in the slide.

Per Base Alignment Qualities

Short Read

GATAGCTAGCTAGCTGATGA GCCG

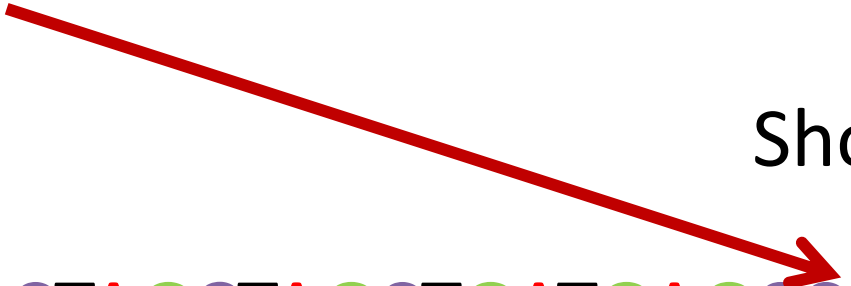
5’-AGCTGATAGCTAGCTAGCTGATGAGCCCGATC-3’

Reference Genome

Heng Li

Per Base Alignment Qualities

Should we insert a gap?



Short Read

GATAGCTAGCTAGCTGATGAGCC-G

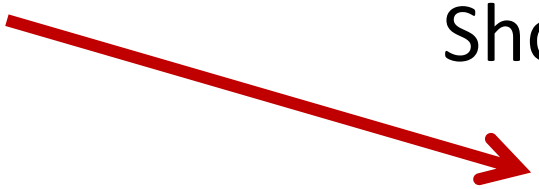
5’-AGCTGATAGCTAGCTAGCTGATGAGCCCGATC-3’

Reference Genome

Heng Li

Per Base Alignment Qualities

Compensate for Alignment Uncertainty With Lower Base Quality



GATAGCTAGCTAGCTGATGAGCCG

5’-AGCTGATAGCTAGCTAGCTGATGAGCCCGATC-3’

Reference Genome

Heng Li

Indel Realignment

- Evaluate groups of overlapping reads, so that…

```
          AGCTAGCTAGCTGATC
           GCTAGCTAGCTGATCC
               GCTAGCTGAT---CCGCTA
REF  GATAGCTAGCTAGCTGATGAGCCGCTA
```

- Becomes …

```
          AGCTAGCTAGCTGAT---C
           GCTAGCTAGCTGAT---CC
            CTAGCTAGCTGAT---CCG
REF  GATAGCTAGCTAGCTGATGAGCCG
```

Sample Contamination Checks

- Did we sequence the right sample?
- Are there contaminants in the sequence data?

- It is very useful to:
  - Check variant sites for evidence of contamination.

  (more “heterozygotes”, fewer “homozygotes” than expected)
  - Compare reads to known genotypes

  (more mismatches than expected would be troubling)

Recommended Reading

- The 1000 Genomes Project (2010) A map of human genome variation from population-scale sequencing. *Nature* **467**:1061-73