Alignment post-processing and variant calling

Simon Rasmussen
DTU Bioinformatics

27626 – Next Generation Sequencing Analysis
Outline

① Introduction

② Alignment post-processing

③ Aligned reads (BAM) ➔ genomic variation (VCF)

④ Variant effect
Generalized NGS analysis

Data size

Question  Raw reads  Pre-processing  Assembly: Alignment / de novo  Application specific: Variant calling, count matrix, ...

Compare samples / methods  Answer?
In principle it is very simple...
But reality is slightly more complex…

How much of what we observe is actually due to errors (noise) and how much represent real genomic variation?
Recommended workflow

1. **Alignment post-processing**
   - Mapping & alignment (.bam)
   - Alignment sorting, filtering and indexing (.sort.flt.bam)
   - Local Realignment (.realign.bam)
   - Duplicate marking/removal (.rmdup.bam)
   - Base Quality Score Recalibration (recalibrated.bam)

2. **Alignment statistics**
   - Mapping & alignment (.bam)

3. **Variant calling & genotyping (.vcf)**
   - Variant calling & genotyping (.vcf)
   - Variant filtering
   - Variant effect
ALIGNMENT POST-PROCESSING
Recommended workflow

1. Alignment sorting, filtering and indexing (.sort.flt.bam)
2. Local Realignment (.realign.bam)
3. Base Quality Score Recalibration (recalibrated.bam)
4. Duplicate marking/removal (.rmdup.bam)
5. Alignment statistics

Alignment post-processing

Mapping & alignment (.bam)

Variant calling & genotyping (.vcf)

Variant filtering

Variant effect
**Recommended workflow**

1. **Alignment sorting, filtering and indexing**
   - `.sort.flt.bam`

2. **Alignment post-processing**
   - **Local Realignment**
     - `.realign.bam`
   - **Duplicate marking/removal**
     - `.rmdup.bam`
   - **Base Quality Score Recalibration**
     - (recalibrated.bam)

3. **Variant post-processing**
   - **Variant calling & genotyping**
     - `.vcf`
   - **Variant filtering**
   - **Variant effect**
Local realignment\textsuperscript{19,20}

- Each read is independently aligned to the reference genome to cut down on the computational cost.
- But this means that many reads spanning indels are likely to be misaligned.
- We see this as regions containing indels as well as clusters of mismatching bases.
Local realignment\(^{19,20}\)

✧ How does it work?

✧ Step 1: Identify regions likely in need of realignment.
✧ Step 2: Perform a multiple sequence realignment in this region, such that the number of mismatching bases is minimized across all reads.

1 – At least one read contains an indel.
2 – A cluster of mismatching bases exists.
3 – An already known indel segregates at the site.
Recommended workflow

1. **Alignment sorting, filtering and indexing (.sort.flt.bam)**
2. **Local Realignment (.realign.bam)**
3. **Duplicate marking/removal (.rmdup.bam)**
4. **Base Quality Score (recalibrated.bam)**
5. **Alignment statistics**
6. **Mapping & alignment (.bam)**
7. **Variant calling & genotyping (.vcf)**
8. **Variant filtering**
9. **Variant effect**
Duplicate marking/removal

Why do we have duplicates in our data?
• The PCR amplification step included in the majority of NGS library construction techniques can introduce duplicates in the data.

Why do we need to remove these?
• It will bias our variant calls.
• PCR errors are propagated and sampled many times giving rise to FP variant calls.

Basic concepts of duplicate marking algorithm:
1. Identify genomic position and strand for 5’-most bases.
2. Mark reads that are duplicates of each other.
3. Within a group of duplicate reads the read with the highest sum of base quality scores is retained.

But it is not perfect...
❖ Does not account for sequencing errors.
❖ Does not account for natural duplicates.
❖ Does not account for duplicate reads with different mapping locations.

http://picard.sourceforge.net/
Recommended workflow

1. **Alignment sorting, filtering and indexing** (.sort.flt.bam)
2. **Local Realignment** (.realign.bam)
3. **Duplicate marking/removal** (.rmdup.bam)
4. **Base Quality Score** (recalibrated.bam)

**Alignment post-processing**

- **Variant calling & genotyping** (.vcf)
- **Variant filtering**
- **Variant effect**

**Alignment statistics**

- **Mapping & alignment** (.bam)
Base quality score recalibration$^{2,3}$

◆ Base-calling algorithms produce per-base quality scores by using noise estimates from image analysis.

◆ The raw Phred-scaled quality scores produced by base-calling algorithms may not accurately reflect the true base-calling error rates.

◆ Obtaining well-calibrated base quality scores is important as SNP and genotype calling depends on both the base calls and the per-base quality scores.
Base quality score recalibration

Base quality scores are systematically biased.

underestimated

overestimated
Base quality score recalibration\(^{2,3}\)

• How it works!

1. Collect information regarding the following features of the bases:
   - Reported quality score
   - Position within the read (machine cycle)
   - Dinucleotide context (current base plus previous base)

2. **Count the number of times a site was a mismatch to the reference (excluding known polymorphic sites).**

3. Estimate new quality score as:

   \[
   \text{Phred-scaled quality score} = \frac{\# \text{ of reference mismatches} + 1}{\# \text{ of observed bases} + 2}
   \]

Example:

We observed A [AA, pos. 35, Q20] a 1,000,000 times.
A in this context mismatches the reference a 1,000 times.
This gives us: \( Q \text{ value} = -10 \log_{10} \left( \frac{(1,000+1)}{(1,000,000+2)} \right) \sim Q_{30} \)
FROM BAM TO VCF

Aligned reads ➔ genomic variation
Recommended workflow¹

Alignment statistics

- Mapping & alignment (.bam)
- Alignment sorting, filtering and indexing (.sort.flt.bam)
- Local Realignment (.realign.bam)
- Duplicate marking/removal (.rmdup.bam)
- Base Quality Score (recalibrated.bam)

Alignment post-processing

- Variant calling & genotyping (.vcf)
  - Variant filtering
  - Variant effect
**Variant calling & genotyping**

- **Aims**
  - **Variant calling:** Identify polymorphic sites => sites that differs from the reference.
  - **Genotyping:** Determine the genotype for a certain individual at such sites.

- **Early methods**
  - Works by simply counting the alleles at each site, and then identifying a variant by use of simple cutoff rules.
Identify variants

What if it is human DNA?
What if it is bacterial DNA?
Variant calling & genotyping

- Several Bayesian genotyping methods available:
  - Use the information on base counts, base qualities, mapping quality
  - Calculate genotype likelihoods

- UnifiedGenotyper (GATK)
- HaplotypeCaller (GATK)
- FreeBayes
- Samtools
- Angsd
- HaplotypeCaller (GATK)
Variant calling & genotyping

- Probabilistic methods – a simple Bayesian genotyper\(^5\)
- The posterior probability of each genotype given the pileup of sequence reads is given by Bayes theorem:

\[
p(G|D) = \frac{p(G)p(D|G)}{p(D)}
\]

\[
p(D|G) = \prod_{b \in \text{pileup}} p(b|G)
\]

\[
p(b|G) = p(b|\{A_1, A_2\}) = \frac{1}{2} p(b|A_1) + \frac{1}{2} p(b|A_2)
\]

\[
p(b|A) = \begin{cases} 
\frac{e}{3} : b \neq A \\
1 - e : b = A
\end{cases}
\]

- \(p(G|D)\): Prob. of genotype given data (posterior probability)
- \(p(G)\): Prior prob. based on allele freqs. from HW
- \(p(D|G)\): Prob. of data given genotype (genotype likelihood)
- \(p(D)\): Prob. of data
- \(p(b|G)\): Prob. of genotype given observed base at a site
- \(e\) is the Phred scaled quality score of the base
We see 3 reads: T (Q20), T (Q20), G (Q10)

1. Calculate Probability of each base given the observed allele $p(b|A)$:

For read 1 and 2: T (Q20)
- $p(A|A) = 0.01/3$
- $p(C|A) = 0.01/3$
- $p(G|A) = 0.01/3$
- $p(T|A) = 0.99$
- Q20 = 0.01

For read 3: G (Q10):
- $p(A|A) = 0.1/3$
- $p(C|A) = 0.1/3$
- $p(G|A) = 0.9$
- $p(T|A) = 0.1/3$
- Q10 = 0.1

Determine for each possible genotype

\[
p(G|D) = \frac{p(G)p(D|G)}{p(D)}
\]

\[
p(D|G) = \prod_{b \in \text{pileup}} p(b|G)
\]

\[
p(b|G) = p(b|\{A_1, A_2\}) = \frac{1}{2}p(b|A_1) + \frac{1}{2}p(b|A_2).
\]

\[
p(b|A) = \begin{cases} 
\frac{e}{3} : b \neq A \\
1 - e : b = A 
\end{cases}
\]
We see 3 reads: T (Q20), T (Q20), G (Q10)

2. Calculate prob. of each base given genotype \( p(b|G) \):

For reads 1 and 2:
TT: \( 0.99/2 + 0.99/2 = 0.99 \)
TG: \( 0.99/2 + 0.01/(3*2) = 0.49 \)
GG: \( 0.01/(3*2) + 0.01/(3*2) = 0.003 \)

For read 3:
TT: \( 0.1/(3*2) + 0.1/(3*2) = 0.03 \)
TG: \( 0.1/(3*2) + 0.9/2 = 0.47 \)
GG: \( 0.9/2 + 0.9/2 = 0.9 \)
We see 3 reads: T (Q20), T (Q20), G (Q10)

3. Sum over all bases to calculate prob. of genotype $p(D\mid G)$:
   - TT: $0.99 \times 0.99 \times 0.03 = 0.029$
   - TG: $0.49 \times 0.49 \times 0.47 = 0.11$
   - GG: $0.003 \times 0.003 \times 0.9 = 8 \times 10^{-6}$

4. Calculate posterior prob. $p(G\mid D)$:
   $p(D) = 0.029 + 0.11 + 8 \times 10^{-6} = 0.139$
   $p(G)$: for simplicity we assume uniform prior
   Posterior probability $p_G(D)$:
   - TT: $0.029/p(D) = 0.21$
   - TG: $0.11/p(D) = 0.79$
   - GG: $8 \times 10^{-6}/p(D) = 5.8 \times 10^{-5}$
   This is “GQ” in VCF
Recommended workflow

Alignment statistics

- Mapping & alignment (.bam)

Alignment post-processing

- Alignment sorting, filtering and indexing (.sort.flt.bam)
- Local Realignment (.realign.bam)
- Duplicate marking/removal (.rmdup.bam)
- Base Quality Score (recalibrated.bam)

Variant calling & genotyping (.vcf)

- Variant filtering
- Variant effect
Variant filtration (soft)\textsuperscript{18}

- **Variant quality score recalibration (GATK)**
  - How do we remove false positive calls?
  - Use known polymorphic sites to estimate what a real variant and a false variant “looks like”
  - Learn how does the known sites (=truth set) look like in our data
  - Evaluate on all our data, filter sites that look different!
Train a predictor & Test:

Known polymorphic sites

Our called sites

Take out overlap

How does these sites look in our data?

Pos 123
Pos 345
Pos 778

Pos 55
Pos 123
Pos 345
Pos 400
Pos 665
Pos 988

Pos 123
Pos 345

Pos 123: GQ, Depth, Strand bias, …
Pos 345: GQ, Depth, Strand bias, …

Apply models to all our sites

HiSeq: evaluating new variants
- Filtered novel SNP
- Retained novel SNP

HiSeq: training on HapMap
- SNP found in dbSNP
- SNP found at HapMap3 site

Gaussian mixture model fits

Likely dbSNP errors

Heterozygous variants

Homzygous variants

More bias

Evidence for strand bias

Variant quality score / depth

More bias

Evidence for strand bias

Variant quality score / depth

Less bias

Less bias

Variant quality score / depth
Variant filtration (soft)\textsuperscript{18}

The raw SNP calls are partitioned into quality trances based on truth sensitivity.

Truth: Accept all sites until X\% of the truth sites have been found
Variant filtration (hard)

• Hard filtering\textsuperscript{16} (when VQSR is not possible*)
  
  – Variant quality score /depth
  – Mapping quality
  – Strand bias (the variant being seen only on the forward strand or only on the reverse strand)
  – Depth

• Some recommendations\textsuperscript{17}

For SNPs:
- QD < 2.0
- MQ < 40.0
- FS > 60.0
- HaplotypeScore > 13.0
- MQRankSum < -12.5
- ReadPosRankSum < -8.0

For indels:
- QD < 2.0
- ReadPosRankSum < -20.0
- InbreedingCoeff < -0.8
- FS > 200.0

10 or more samples!

*number of samples < 30 or if your doing targeted resequencing of a small region, non-model organism.
Recommended workflow¹

Alignment post-processing

- Alignment sorting, filtering and indexing (.sort.flt.bam)
- Local Realignment (.realign.bam)
- Duplicate marking/removal (.rmdup.bam)
- Base Quality Score (recalibrated.bam)

Alignment statistics

- Mapping & alignment (.bam)

Variant calling & genotyping (.vcf)
- Variant filtering
  - Variant effect
Variant annotation

Variant Effect Predictor script

About | Web version | Perl script | Data formats | Frequently asked questions

Download

The Variant Effect Predictor script can be downloaded as a tarball from the Ensembl CVS server:

Download latest version (71)

NOTE: VEP version numbers have from release 71 changed to match Ensembl release numbers.

Previous versions: Show

Requirements

Version 71 of the script requires at least version 71 of the Ensembl Core and Variation APIs and their
the supplied installer script.

Link: http://www.ensembl.org/info/docs/variation/vep/vep_script.html
References

2) http://gatkforums.broadinstitute.org/discussion/44/base-quality-score-recalibration-bqsr
3) http://www.youtube.com/watch?v=L4D1dwES9s8
7) http://www.1000genomes.org/wiki/Analysis/Variant%20Call%20Format/vcf-variant-call-format-version-41
8) http://www.ensembl.org/info/docs/variation/vep/index.html
References

9) http://www.broadinstitute.org/gatk/gatkdocs/org_broadinstitute_sting_gatk_walk
ers_genotyper_UnifiedGenotyper.html
10) http://www.broadinstitute.org/gatk/gatkdocs/org_broadinstitute_sting_gatk_walk
ers_haplotypecaller_HaplotypeCaller.html
12) http://www.broadinstitute.org/gatk/gatkdocs/org_broadinstitute_sting_gatk_walk
ers_variantrecalibration_VariantRecalibrator.html
13) http://www.broadinstitute.org/gatk/gatkdocs/org_broadinstitute_sting_gatk_walk
ers_variantrecalibration_ApplyRecalibration.html
14) http://www.broadinstitute.org/gatk/gatkdocs/org_broadinstitute_sting_gatk_walk
ers_varianteval_VariantEval.html
15) http://www.broadinstitute.org/gatk/guide/article?id=1247
16) http://www.broadinstitute.org/gatk/gatkdocs/org_broadinstitute_sting_gatk_walk
ers_filters_VariantFiltration.html
17) http://www.broadinstitute.org/gatk/guide/article?id=1186
18) http://www.broadinstitute.org/gatk/guide/article?id=39
19) http://www.broadinstitute.org/gatk/gatkdocs/org_broadinstitute_sting_gatk_walk
ers_indels_RealignerTargetCreator.html
20) http://www.broadinstitute.org/gatk/gatkdocs/org_broadinstitute_sting_gatk_walk
ers_indels_IndelRealigner.html

21) Heng Li, Jue Ruan, Richard Durbin. Mapping short DNA sequencing reads and
calling variants using mapping quality scores, Genome Research, 2008