Best Practices and Suggestions for Read Alignment

Sonal Singhal
Museum of Vertebrate Zoology
University of California, Berkeley

10 December 2012

1 Introduction

Read alignment is the first step for almost all downstream genomic analyses, whether you are working with an assembly you generated yourself or a well-annotated and high-quality reference. In this workshop, we will go over some best practices, introduce you to the software at hand, and briefly touch on downstream analyses.

In many papers, alignment is often treated as a throw-away step to get to the real work in genomics analyses, whether that is estimating expression, calling genotypes, or looking at copy number variation. Don't treat it like a simple step. As shown by a recent string of papers (Li, 2011; Lin *et al*., 2012; Kleinman and Majewski, 2012), alignment matters. A lot.

2 Key Programs to Know

I highlight these programs, because in this world of infinite number of programs and ever-multiplying options, they are continuously being improved and their authors are active on `SeqAnswers.com`. Other programs (like the `SOAP` suite of programs) are great, but they are black boxes and the programmers rarely interact with users.

- Cleaning reads: `trimmomatic`, `cutadapt`
- Merging reads: `Flash`
- Aligning reads: `bowtie2`, `bwa`, `NovoAlign`
- Working with SAM/BAM files: `SAMtools`, `PICARD`, `GATK`, `BamTools`

3 Useful links

- `http://www.usadellab.org/cms/index.php?page=trimmomatic`
- `http://code.google.com/p/cutadapt/`
- `http://genomics.jhu.edu/software/FLASH/index.shtml`
- `http://bowtie-bio.sourceforge.net/bowtie2/manual.shtml`
- `http://bio-bwa.sourceforge.net/`

- http://www.novocraft.com/main/index.php
- http://www.broadinstitute.org/gatk/guide/topic?name=best-practices
- http://picard.sourceforge.net/index.shtml
- http://samtools.sourceforge.net/samtools.shtml
- http://samtools.sourceforge.net/SAM1.pdf
- https://github.com/pezmaster31/bamtools/wiki/Using-the-toolkit
- https://sites.google.com/site/mvzseq/
- http://picard.sourceforge.net/explain-flags.html

4 Some Points of Consideration

- **phred+33 versus phred+64**: make sure you know your quality score encoding and that it is specified correctly in all programs
- **unique vs. non-unique**: most alignment programs allow reads to be aligned multiple times. Should you use these multiple alignments? This is a bigger problem with low-quality reference genomes.
- **local vs. global alignments**: `Bowtie2` (probably the most popular aligner currently) allows you to align reads globally (*i.e.*, the entire read from head to tail in one go) or locally (*i.e.*, like `BLAST` works as it can align just part of your query sequence to the reference) – which one do you want to use? You have to decide that for your data.
- **the importance of trimming**: very important!
- **quality control**: how much filtering will you do of the alignments themselves?
- **removing duplicates**: You definitely want to do this, but how will you?
- **validation with external data sets**: great if you can do it – will want use `GATK` to do so

5 Clean Reads

Cleaning reads prior to analysis is important for several reasons; namely, it increases the likelihood that a read will align and it reduces the likelihood that sequencing errors are mistaken for SNPs. There are numerous steps to clean data, some more arguably important than others. Some steps that folks are taking include:

- trimming adapters
- trimming low quality bases
- removing low complexity reads
- removing reads from potential contaminant sources

- removing reads that appear to be exact duplicates (*i.e.*, those due to over-amplification of libraries during PCR)
- merging paired reads whose sequences overlap

All of this cleaning is done in one master Perl script that I wrote and that I share with you: `1scrubReads.pl`. Here is what the script does. The script is not perfect (nor will it ever be) but there is one big hiccup that I am working actively to fix. If you ask, I will keep you looped in on updated versions.

1. Trim adaptors. Other researchers and I have found that most adaptor removal programs fail to remove all adaptors, especially in just one go. So, I remove adaptors through multiple rounds of trimming, using the programs `CutAdapt`, `trimmomatic`, and `bowtie2`. This is almost certainly overkill, but we have found it improves alignment quality and reduces error in genotype calling. The script I have to do this is still a work in progress, but I share it with you here.
2. Remove low-quality bases – `trimmomatic` does a great job with this, and I implement it in the script
3. Remove low-complexity bases – wrote a simple Perl subroutine to identify and ditch these reads
4. Remove reads from potential contaminant sources – to do this, I identify reads that align to the human genome using `bowtie2`. If you are working with human data, oops!
5. Remove exact duplicates – most people use `Picard` for this, but I found it over-zealous and wrote my own.
6. Merging paired reads – typically, if you sequenced paired-end data, some of the reads will overlap each other. Not accounting for this can mess up how certain you think SNP calls are or (more importantly) expression counts. Merging reads also improves sequence quality. We use `FLASH` to do it here, but we know it does not catch all reads. We are working to find a better program(s).
7. If you have paired end data, all this trimming typically leads to orphan reads, where one half of a paired end gets lost. I could not find a script to re-pair all the paired reads, so I wrote my own (memory-intensive) sub-routine, which is implemented in the script.

To evaluate data quality before and/or after trimming, I recommend using `FastQC`. This basic approach works. We found it reduced error rates six-fold.

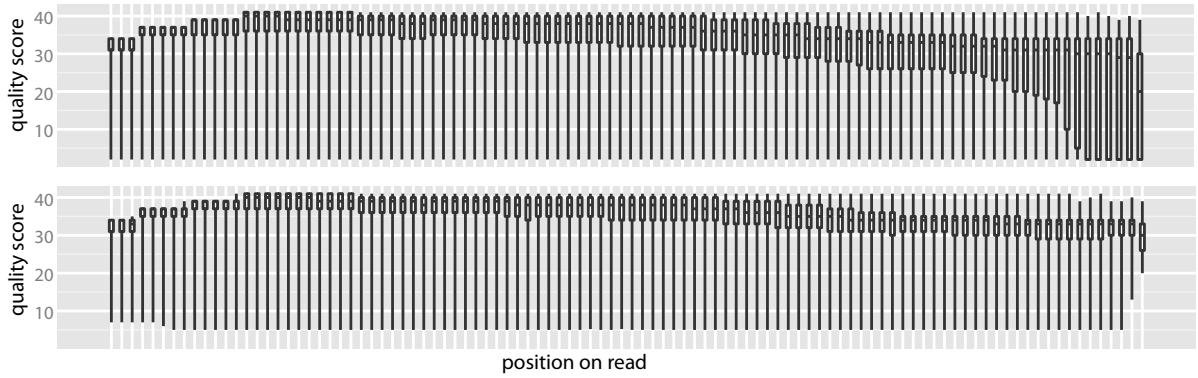

Figure 1: Quality scores in Phred along a read; top graph shows quality prior to cleaning and filtering, bottom shows quality after cleaning.

6 Read alignment

Next, you will want to align your cleaned reads to your reference genome. To do so, you must first chose an aligner. Folks are mainly using either `Bowtie2`, `bwa` or `Novoalign`, but there are many others including `SoapAligner` and `Stampy`. Simulated data suggest `bwa`, `Novoalign` and `Bowtie2` perform the best; you might want to test both with your data to see which one (and under which parameters) you can recover alignments that look reasonable. In particular, if you are aligning reads that you expect to be quite divergent from your reference genome (*i.e.*, perhaps the individuals are 15% divergent at the mitochondrial genome), you might need to sacrifice one of the speedy aligners (`Bowtie2`) for one of the more sensitive aligners (`Novoalign`).

7 Quality Control

- **Drop low-quality alignments**: Each alignment is given a quality score (MAPQ) which is equivalent to Phred scores (with respect to determining relative quality of different alignments). You can screen for low-quality alignments when you convert from SAM to BAM files using `samtools view`.
- **Drop repeat alignments**: Most alignment programs will report multiple hits for each read, if they exist. You can either turn off this option in the program itself, or parse through these later using (again) `samtools view`.
- **Drop duplicated reads**: If you want to be extra sure you do not have PCR duplicates in your library, you can use `Picard` and `MarkDuplicates` to identify any PCR duplicates and remove.
- **Re-align around indels**: `GATK` allows re-alignment around indels (especially known indels if you have an idea of this) and re-calibration of quality scores with known variants.
- **Variant distribution along read**: In theory, if you trimmed your data well and the alignments went well, the distribution of mismatches should be even across a read. You can check this using the SAM output and the scripts `snpReadDistribution.pl` and `snpReadDistribution_unpaired.p` both available at `https://sites.google.com/site/mvzseq/original-scripts-and-pipelines/pipelines`.

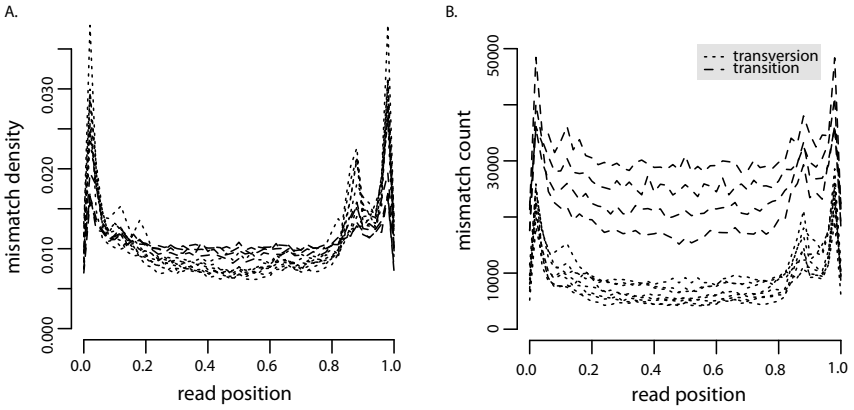

Figure 2: Identified mismatches between reads from a randomly-selected individual and the reference sequence, A. expressed in raw numbers and B. as a density distribution. This is how you DO NOT want your data to look!

References

Kleinman, C. and J. Majewski, 2012. Comment on "Widespread RNA and DNA sequence differences in the human transcriptome". Science 335:1302.

Li, H., 2011. A statistical framework for SNP calling, mutation discovery,association mapping and population genetical parameter estimation from sequencing data. Bioinformatics 27:2987–2993.

Lin, W., R. Piskol, M. Tan, and J. Li, 2012. Comment on "Widespread RNA and DNA sequence differences in the human transcriptome". Science 335:1302.