Sequence Alignment with BWA

Shamith Samarajiwa

July 27, 2015

Introduction

I'll be using a ChIP-seq dataset to show sequence alignment. Here are the dataset details:

SRR1186252
Sample: GM00011_ChIP_p53
Study summary: SRP039598
Integrative Genomic Analysis Reveals Widespread Enhancer Regulation by p53 in Response to DNA Damage
PMID: 25883152
Title: Integrative genomic analysis reveals widespread enhancer regulation by p53 in response to DNA damage.
Authors: Younger ST, Kenzelmann-Broz D, Jung H, Attardi LD, Rinn JL.
Publication date: May 2015

Download the dataset (fastq file) from Sequence Read Archive **(SRA)**. There are multiple ways of doing this.

1. Browse to the SRA database [1]
2. Use the SRA toolkit. You need to install and configure this on your computer first. Detailed instructions are here [2]
3. Use the bioconductor package SRAdb to download SRR files [3].

[1] http://www.ncbi.nlm.nih.gov/sra

[2] http://www.ncbi.nlm.nih.gov/Traces/sra/?view=software

[3] http://www.bioconductor.org/packages/release/bioc/html/SRAdb.html

We did a quality analysis of the file using **FastQC** and **Trim_galore (cutadapt)** was used to remove adaptor/primer/artefact contamination. I then extracted chromosome 6 reads from this file. This chromosome 6 *fastq* file will be used in this BWA alignment demonstration.

```
# cd ~/Course_Materials/Day1/alignment-demo
```

Create BWA hg19chr6 index

```
# bwa index -p hg19chr6bwaidx -a bwtsw hg19chr6.fa
```

Align to hg19 reference using BWA aln algorithm

```
# bwa aln -t 4 hg19chr6bwaidx SRR1186252_trimmed.fq.chr6.fq > SRR1186252_trimmed.chr6.fq.bwa
```

Generate the single end alignment SAM file

```
# bwa samse hg19chr6bwaidx SRR1186252_trimmed.chr6.fq.bwa SRR1186252_trimmed.fq.chr6.fq>
# SRR1186252_trimmed.chr6.fq.sam
# head SRR1186252_trimmed.chr6.fq.sam
```

Generate BAM file

```
# samtools view -bS SRR1186252_trimmed.chr6.fq.sam > SRR1186252_trimmed.chr6.fq.bam
```

Sort BAM file

```
# samtools sort -O bam -o SRR1186252_trimmed.chr6.fq.sorted.bam -T temp SRR1186252_trimmed.chr6.fq.bam
```

Generate BAM index

```
# samtools index SRR1186252_trimmed.chr6.fq.sorted.bam
```

Run SAMstat

SAMstat can be downloaded and installed from the given gitgub repository [4]. It's a C program that helps to quality control alignment files (SAM and BAM) and can identify errors, and provide detailed alignment statistics.

[4] https://github.com/cidvbi/PathogenPortal/tree/master/portal_rnaseq_galaxy/pathogen_portal/samstat

```
# samstat SRR1186252_trimmed.chr6.fq.sorted.bam
```

Generate a tdf (tile data format) file for viewing in IGV

```
# igvtools count -z 5 -w 25 -e 250  SRR1186252_trimmed.chr6.fq.sorted.bam chr6.tdf hg19
```