CSE 291 - PERSONAL GENOMICS FOR BIOINFORMATICIANS

Problem Set 4 - Next-generation sequencing

This homework should be sent to mgymrek@ucsd.edu with subject line [CSE291 PS4:LASTNAME] by the beginning of class on **Tuesday, February 28**. The assignment is worth 10 points total.

As in previous problem sets, template code is provided for some problems. Using the template code is optional, it is simply there to guide you.

Objectives

- Gain experience using standard short read alignment, genotyping, and visualization techniques.
- Learn how to detect and filter artifacts in alignment or variant calling that can arise from next generation sequencing analysis.
- Explore new long read sequencing technologies.

A description of data files for this problem set and several setup steps can be found at PS4 resources.

Part 1: Sequence alignment and visualization (4 points)

Overview

Exercises

1. **(0.5 points)** Describe and draw the difference between "single-end" and "paired-end" sequencing reads.
2. **(1 point)** Align the reads from the fastq files using bwa-mem and convert the resulting alignment to a BAM file. What percent of reads were aligned?
3. **(1.5 points)** Now we will do some quality controls on the resulting alignment:
   - Plot a histogram of the coverage per base for the autosomes, chrX, and chrY. Is this a male or female sample? How do you know?
   - Plot a histogram of the distance between mate pairs. What is the average? Standard deviation?
4. **(1 point)** Use samtools tview to visualize the resulting alignment. Navigate to positions XX, XX, and XX. Do these look like homozygous reference, heterozygous, or homozygous non-reference positions? Do any positions look like they are affected by sequencing or alignment artifacts? Explain.

Part 2: Writing a simple SNP caller (4 points)

Overview

Exercises

1. **(0.5 point)** Convert the BAM file of your sequence alignment from the last problem to samtools "pileup" format. Describe the resulting format.

2. **(1.5 points)** Write a SNP caller that takes in the pileup format and outputs a list of putative SNPs. To simplify things, you can restrict to sites with at most 2 alleles present and only run your tool on chromosome 22. Briefly describe your method

3. **(1 point)** Compare your results to those obtained by the Affy genotype array. Based on this comparison, what is your accuracy rate? False positive rate? False negative rate? How does your accuracy change if you restrict to loci covered by at least 5 reads? 10 reads? 20?

4. **(1 point)** Examine, for instance using samtools tview, 2 cases each of false negative calls or false positive calls. Hypothesize where your SNP caller went wrong.

Part 3: Long read technologies (2 points)

Overview

Exercises