Gene 245 Final Project
Matthew Durrant, Eli Moss, Ryan McKinney

Abstract

We have previously developed a read filtering approach to select read pairs within metagenomic sequencing datasets such that one read maps to a bacterial insertion sequence and the other to a bacterial genome (IS-genome pairs), indicating the presence of an insertion sequence absent from the reference genome sequence. In the present project we present an alignment correction, peak-calling, and differential abundance workflow to analyze the occurrence and abundance fluctuations of these insertion sites within longitudinal clinical microbiome sequencing data. We validate the approach on simulated longitudinal IS-genome pair data, then apply it to a longitudinal clinical gut microbiome dataset obtained from a patient at Stanford Hospital.

Background

Mobile sequence elements are one common instance of the widespread structural variation seen in bacterial genomes. These elements, collectively referred to as the mobilome, often encode proteins mediating their own mobilization, carry a 'payload' sequence encoding antibiotic resistance or other clinically significant phenotype, or modulate the regulation of genes adjacent to insertion sites. Because of their repetitive nature, these mobile sequences are difficult to analyze with conventional short read sequencing and assembly methods. Better understanding of these mobile elements provides insight into important bacterial genomic processes of sequence acquisition, transcriptional regulation and structural change.

Our focus is on the impact of these genomic processes on the human gut microbiome. Our clinical dataset captures an instance of severe domination of the gut bacterial composition by *Bacteroides caccae* in a patient at Stanford hospital. This type of takeover of the gut microbial community by a single organism presages poor outcomes in the immunocompromised patient population[1]. We focus on insertion sequence IS614, previously seen to mediate clinically significant regulatory effects in *Bacteroides fragilis*[2] and in strains of *B. caccae* in the present dataset (Moss et al., in preparation). This insertion sequence carries an outward-facing promoter capable of upregulating adjacent coding sequences. Locating the instances of this sequence as they fluctuate in relative abundance over time will enable us to test regulatory changes in nearby genes for association with this insertion sequence.

Due to their repetitive nature, insertion sequences pose an intractable problem for short-read sequence assembly. This is due to the fragment size of the short-read library; when fragments do not exceed the size of the repeated insertion sequence, read pairs cannot span its length. In the absence of pairs of reads linking left and right genomic flanks of instances of the repeated sequence, contiguous assembly of the insertion with both flanks cannot be performed. Instead, alignment-based approaches can be used to locate these sequences within existing reference sequences.

We previously created an alignment-based filtering approach modified from previous approaches[3] which obtains genome-mapping reads from IS-genome pairs in metagenomic data. The read profiles resulting from this filtration approach bear many similarities to ChIP-seq data, suggesting a new opportunity to apply analytical approaches developed in that setting. In the present project, we have designed and implemented an analytical workflow incorporating

alignment shifting, peak calling, and differential analysis for IS-genome pair data targeting bacterial insertion sequences.

Methods

Peak shifting

Overlap of library fragments with the insertion sequence create a distinctive pattern of read orientation: the first read of all library fragments overlapping the 5' end of the insertion maps to the genome upstream of the insertion site. Likewise, the second read of fragments overlapping the 3' end maps downstream. As a result, all genome-mapping reads from IS-genome pairs align oriented toward the insertion site at a distance determined by fragment length and overlap with the insertion sequence. This creates a distinctive bimodal coverage depth profile of filtered reads at insertion sites, similar to ChIP-Seq data[4] (Fig. 1).

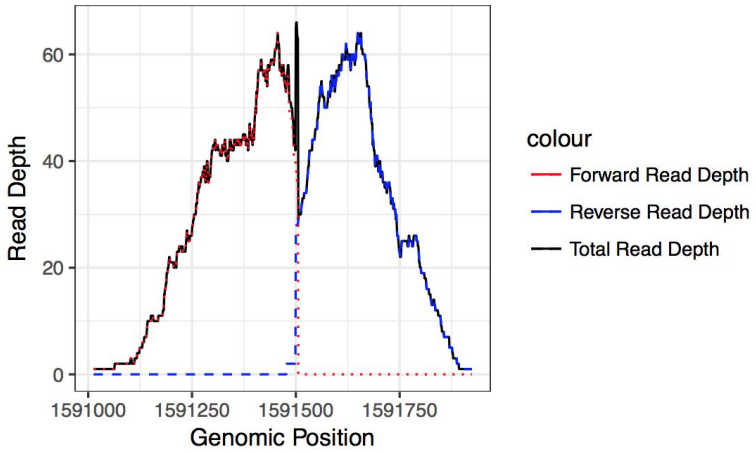

Fig. 1: Coverage profile of filtered reads at an insertion site. Coverage depths are colored by mapping orientation. The narrow central peak is caused by a short tandem repeat created during insertion sequence integration.

In order to call a single peak centered on the insertion site, we map filtered reads to a reference sequence, then shift forward-mapping reads downstream and reverse-mapping reads upstream. We expect the optimal offset to approximate half the mean library fragment size, approximately 500bp in the present data. In order to determine the optimal offset, cross-correlation is calculated on depth profiles for forward and reverse mapped reads (fig. 2). The offset maximizing forward and reverse coverage cross-correlation is

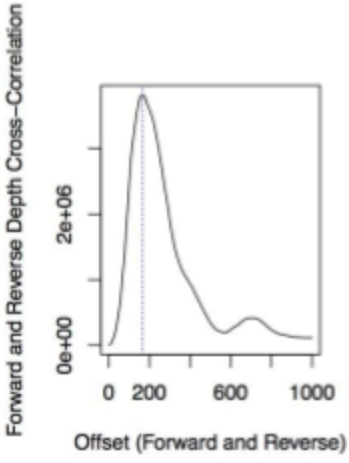

Fig. 2: Cross-correlations calculated at a range of alignment shift offsets in filtered IS-genome read pairs. Forward and reverse mapping reads are shifted symmetrically in order to obtain a single alignment peak centered on the insertion sequence. Plots for all samples are found in supplementary 1.

selected as the optimal correction and an adjusted dataset is output for the next stage of analysis.

Peak calling

We followed a peak calling approach that was very similar to the MACS procedure. The approach is based on a Poisson distribution, with the expected count of reads per window calculated as $\lambda_{BG} = L_W \times (N / L_C)$, where L_W is the length of the sliding window, N is the total number of reads aligned to the contig, and L_C is the length of the contig. The contig length L_w the estimated average fragment length, calculated from previous experimental knowledge or through the peak shift estimation. We then move a sliding window of length L_w across the genome, and calculate a *p*-value as $P(X > x)$ using the `scipy.stats.poisson` python module. We use a *p*-value threshold of 5×10^{-6} to correct for multiple hypothesis testing. This threshold was found to be a good choice in tests on simulated data. To improve runtime, we decided to only look at the windows where the read depth was at least 1, ignoring windows without any reads. Overlapping significant windows were merged together to produce the final peaks file.

We simulated random insertion events in twenty different different genomes, and then simulated paired-end read data using the `wgsim` command line tool. The insertion sequence was 1500 bp in length and was inserted at random 50 times throughout the *Bacteroides fragilis* genome.

Differential analysis

In order to identify peaks with significantly different levels of coverage across longitudinal series, we applied a simple two-step normalization and test workflow, and compared results to a popular existing tool, DESeq2[5]. In the place of true technical replicates, we generated bootstrapped replicates from the true and simulated data as a proof of concept. In addition, we simulated an increase in a subset of called peaks by doubling read coverage within the first 2Mb of reference sequence in a simulated dataset.

Results

Fragment size inference during peak shifting closely approximated (within 10%) the target fragment size selected during the Illumina Truseq library protocol used in clinical sample preparation. For simulated data, all inferred fragment sizes were within 10% of the specified fragment size.

Out of 19 simulated *B. caccae* samples (one sample failed due to unresolvable error), the average percentage of recovered insertion sites was 94% (+/- 3%). We also observed a high false positive rate (17% +/- 7%), perhaps due to an excessively lenient *p*-value threshold.

In simulated longitudinal data, we observed elevated p-values corresponding to those peaks within the first 2Mb with inflated coverage in the later time point, validating our approach.

In clinical data, we found a small number of peaks with qualitatively higher differential significance than background, although we observed overall very high p-values from both our

differential analysis and DESeq2 greatly exceeding even conservative multiple test corrected significance thresholds. This is likely due to low variance in the bootstrapping process used to simulate technical replicates.

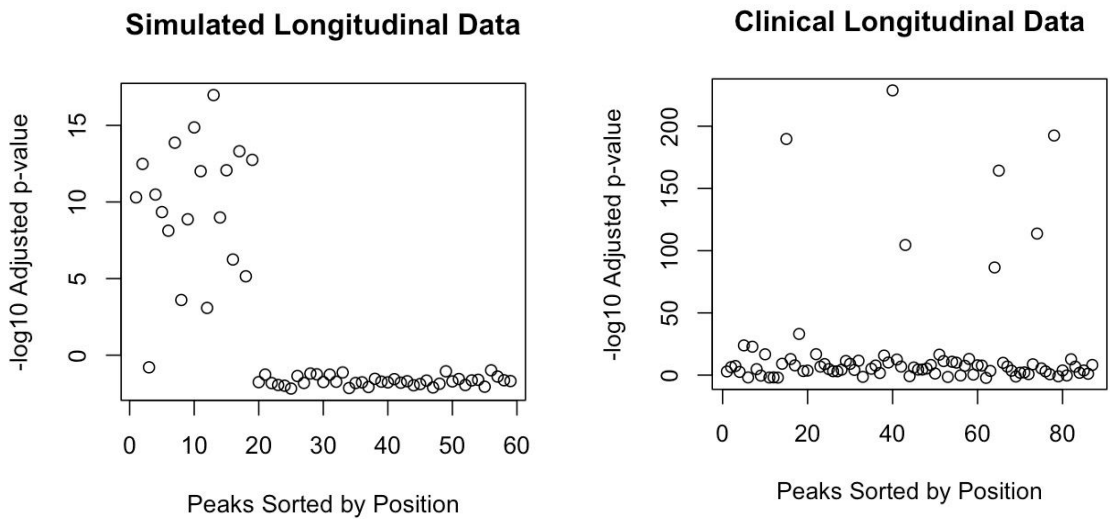

Fig 3 - **(Left)** Simulated data set successfully identifies 18 of 19 differential peaks. **(Right)** Differential Analysis of *B. caccae* IS614 insertion events indicates that 7 out of the 87 peaks is significantly differentially abundant between time points, suggesting strain-level changes in specific insertion event abundance.

Conclusions and Future Directions

We have implemented and validated a workflow based on described ChIP-seq methodology for the analysis of IS-genome read pair data originating from metagenomic shotgun sequencing. The workflow accepts filtered IS-genome read pairs and shifts forward and reverse peaks to correct for library fragment size, then calls significant peaks and performs differential analysis on longitudinal coverage data. We validated the workflow by simulating longitudinal IS-genome pair data with artificial differential peak coverage. We recovered the simulated peaks and correctly identified the peaks with simulated differential coverage between samples. Next, we applied the approach to clinical data and observed a small number of markedly differentially covered peaks between time points.

These results provide a useful underpinning for the targeted detection and tracking of mobile sequence elements in the metagenome. Previous work using a specialized sequencing approach has implicated these mobile elements in opportunistic pathogenicity of typically innocuous gut commensals (Moss et al., in preparation), highlighting the importance of an approach for mobile element detection using conventional short-read sequencing data. We plan to apply this approach to longitudinal metagenomic short-read sequencing data from patients at Stanford Hospital in order to characterize the incidence, genomic location and movement of IS612 and similar elements in bacterial genomes.

Bibliography

1. Taur, Y. *et al.* The effects of intestinal tract bacterial diversity on mortality following allogeneic hematopoietic stem cell transplantation. *Blood* **124,** 1174–1182 (2014).
2. Kato, N., Yamazoe, K., Han, C.-G. & Ohtsubo, E. New insertion sequence elements in the upstream region of cfiA in imipenem-resistant Bacteroides fragilis strains. *Antimicrob. Agents Chemother.* **47,** 979–985 (2003).
3. Hawkey, J. *et al.* ISMapper: identifying transposase insertion sites in bacterial genomes from short read sequence data. *BMC Genomics* **16,** 667 (2015).
4. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9,** R137 (2008).
5. Love, M., Anders, S. & Huber, W. Differential analysis of count data--the DESeq2 package. *Genome Biol.* **15,** 550 (2014).