Growth Dynamics of Human Microbiota Pipeline for Parallelizing Computation

Jerry Pan, Sur Herrera Paredes
jerrypan@stanford.edu, surh@stanford.edu

Stanford | Biology
SCHOOL OF HUMANITIES & SCIENCES
Stanford MEDICINE

Introduction

As indicated by David Zeevi et al.[1], the ratio of sequencing coverage between the origin of replication and the replication terminus, defined as peak-to-trough ratio (PTR), serves a highly correlated index with the growth rate of the bacteria population. Hence, this indicator can be applied for the characterization of microbiome dynamics [2], framing the growth dynamics of microbiota out of static measurements. We set up a pipeline, given input files of reference genome and metagenomic reads, to examine the replication origin and terminus, calculate PTR, and make statistical inferences on bacterial replication rate.

Data and Methods

Data 1 Metagenomic samples *microbiome proliferation* of Citrobacter rodentium in FASTQ format

Data 2 Genome of Citrobacter rodentium

Preprocessing Improve genome accuracy

- Input: FASTA and FASTQ file
- Output: BAM file

1. Glimmer 3.02 for genome annotation
2. Sickle 1.33 for quality control in terms of the length and quality threshold
3. Bowtie 2 for genome alignment and reads mapping

Graph A sliding window of size 10Kbp traverses the entire bacterial chromosome at the step size of 100bp to generate the coverage reads graph. [3] A smoothing filter, LOESS regression, is employed to obtain a curve differentiable everywhere. To get global extrema, we extract all local extrema whose first derivative is equal to 0, and find the minimum and maximum ones out of all local extrema. PTR is calculated by the ratio of maximum y value and minimum y value.

Workflow

We set up an automated Nextflow pipeline to carry out the steps defined in *Data and Methods*. Our pipeline allowed us to parallelize the computation process on the server to obtain a higher efficiency. In concerns of the number of metagenomics data samples to be analyzed, our pipeline can greatly save manual input process.

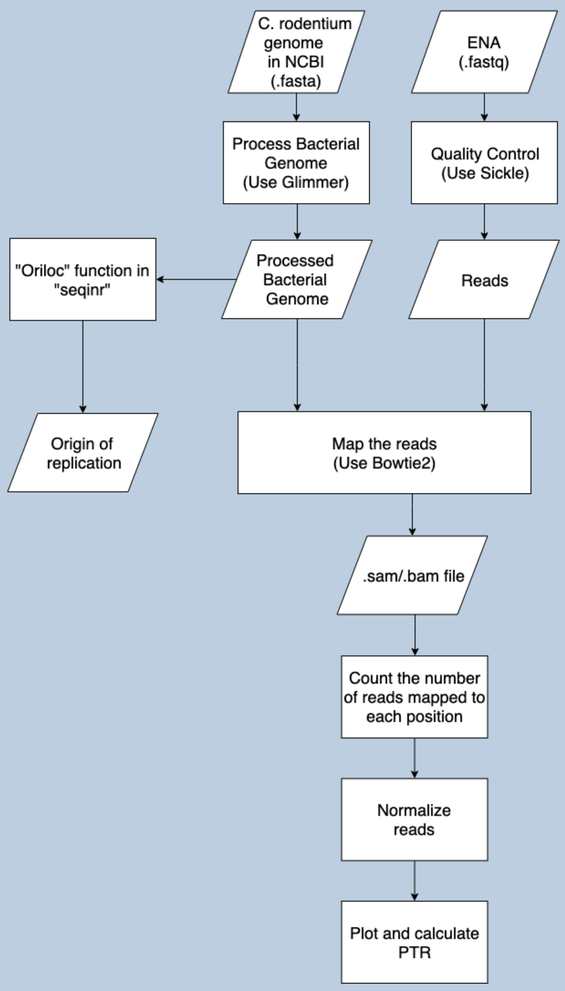

Figure 1: Pipeline flowchart

Result (Individual Sample)

Each sample in the format of FASTQ file undergoes the preprocessing through Glimmer 3, Sickle 1, Bowtie 2, and R graphing algorithms in a pipeline to calculate PTR and plot the coverage reads graph. In data filtration, LOESS regression successfully erases visually detectable outliers in the plot and obtain a curve differentiable everywhere. In most samples, expected graphical shapes are derived. Yet, over 3 graphs out of 54 have significant anomaly: first 10^6 base pairs presenting an interval of decreasing function with consistently negative second derivative. Generally, it poses no difficulty for PTR calculation, since LOESS regression can minimize the influence of outliers. For other samples, the graph shows a relatively flat curve with noises, revealing that these bacteria are not replicating.

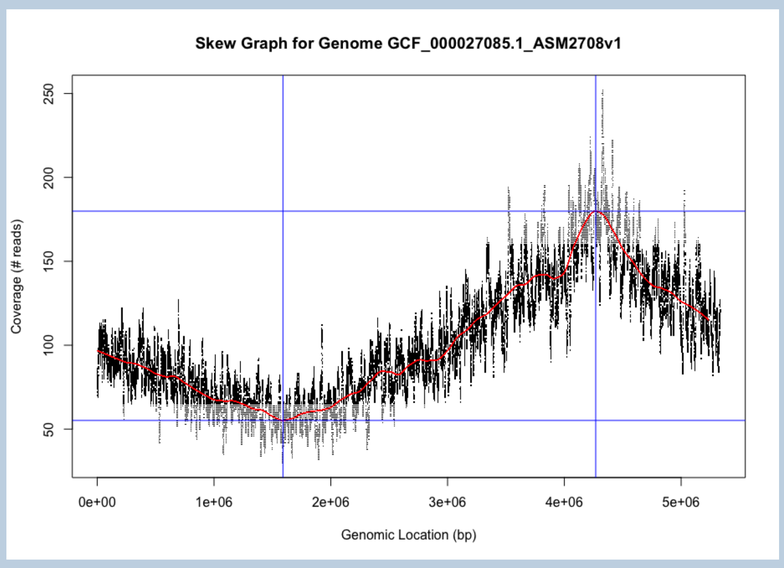

Figure 2: Skew Graph of GCF_000027085.1

	x-coordinate	y-coordinate
Peak	$4.227 * 10^6$	178
Trough	$1.685 * 10^6$	56
PTR	2.508	

Table 1: Skew Graph Peak, Trough, and PTR

Result (Metagenomic Sample)

In total, 54 C. rodentium metagenomic samples from *PRJEB9718 Microbiome proliferation*, 6 groups of 9 samples, are processed to calculate PTRs and graph box plot to compare the distribution of wild-type (WT) and mutant (MT) sample groups' PTRs. In the comparison of WT1, MT1, WT2, and MT2, no significant differences are detected.. Thus, both wild type group (WT1 and WT2) and mutant type group (MT1 and MT2) were combined as WT and MT to improve statistical power by increasing sample size. Consequently, the conclusion that wild type is replicating at a faster rate than mutant type can be drawn from the combined group, proving the validity of using PTR as an indicator of growth rate.

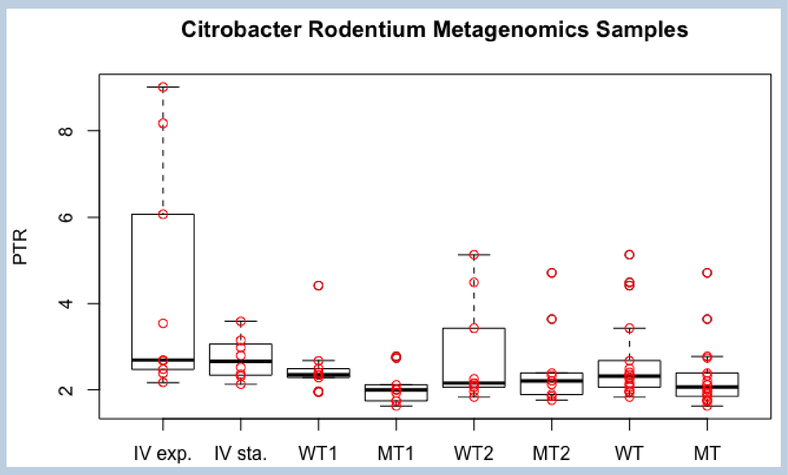

Figure 3: PTR of Metagenomic Samples
Shown are PTRs of virulent (Vcit_B3/C4; wild type; N=9) and non-virulent (Vcit_C1/D3; mutant; N=9) PTR of stationary and exponential in-vitro Citrobacter rodentium cultures are shown for reference. *P*-values are Mann-Whitney *U*-test.

Samples	Wild Type	Mutant
In-vitro exp.	**0.01992**	**0.003544**
In-vitro sta.	0.1961	**0.001873**
WT (period 1)	0.8167	0.1311
Mutant (period 1)	0.05251	0.4552
WT (period 2)	0.8167	0.2746
Mutant (period 2)	0.4948	0.4552

Table 2: Mann-Whitney *U*-test P value among samples (Mann-Whitney significant threshold = 0.05)

Conclusion

Both in individual samples and multiple group samples, PTR is considered as a valid indicator highly correlated with growth rate. PTRs of 54 samples show statistically significant differences between wild-type and mutant C. rodentium strains matching their known virulence.
The pipeline performs its intended function satisfactorily, showing adequate robustness to noises of input sample genome and reference genome. The pipeline is also optimized in space and time complexity at both algorithmic and organizational level. The pipeline takes space 19.4 MB and time under 1 hour for each sample considered plausible computational work with reasonable space and time.

Acknowledgements

I would like to express my deepest appreciation to my mentor Dr. Herrera Paredes, who continually and convincingly conveyed a spirit of adventure in regard to the overall picture and meticulous guidance throughout my project. Without his tutor, I would not be able to accomplish my project with all accurate information.
A special thank of mine goes to my Professor Hunter Fraser, colleagues Thomas Silvers, Roy Ang, Shi-An Chen, Alex Kern, Kate Lawrence, and etc., for both friendship and their exchange of valuable ideas. Thanks for their helping with technical difficulties, including cables for connection to the display and connection to fraser-server, as well as their care for numerous trivial matters.

References

[1] David Zeevi et al. Tal Korem. Growth dynamics of gut microbiota in health and disease inferred from single metagenomic samples. *doi: 10.1126/science.aac4812*, pages 1–9, July 30th 2015.

[2] Curtis Huttenhower et al. Structure, function and diversity of the healthy human microbiome. *doi:10.1038/nature11234*, 486(52):207–214, June 14th 2012.

[3] David Zeevi et al. Tal Korem. Supplementary materials for growth dynamics of gut microbiota in health and disease inferred from single metagenomic samples. *doi: 10.1126/science.aac4812*, pages 1–26, July 30th 2015.