Aimusa Arsu (Beer) SIMB/M 6137762

Term Paper

Title: Genome-wide detection of conservative site- specific recombination in bacteria

Ognjen Sekulovic, et al.

PLOS Genetics | https://doi.org/10.1371/journal.pgen.1007332 April 5, 2018

Section 1: What is already know on this paper?

- Bacteria can survive and successfully proliferate by adaptation to challenging conditions. This ability, clonal bacterial populations to generate genomic and phenotypic heterogeneity, is thought to be of importance for many pathogenic bacteria.
- Gene expression through “DNA inversion” is a common mechanism adopted by many bacterial species that allows quick generation of distinct subpopulations with altered fitness.
- It is lagging due to the difficulties to accurately detect such inversion on a population level. This study implements an easy-to-use method for detecting small genomic inversions in bacterial genomes.

Section 2: What this study adds?

- The method can be applied for detection of small genomic inversions.
- Several species from bacteria presented multiple DNA inversions spanning various cell surface components such as capsule and putative carbohydrate-binding receptors.
- The association of identified invertible sites with genes encoded in their vicinity hints at possible regulation of a diverse set of bacterial functions.

Section 3: How the data were analyzed?

Pick 3 Figures and describe the followings

1. What is the question that this figure tried to answer?
2. What data have been generated to answer the question?
3. What analysis have been done to get the conclusion?

Aimusa Arsu (Beer) SIMB/M 6137762

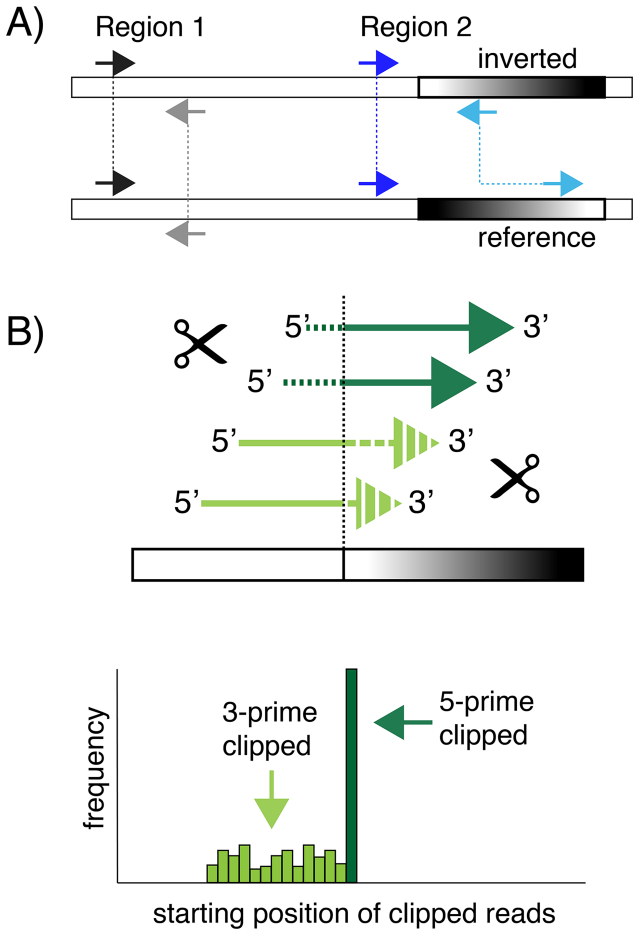

Fig 1. Genomic inversions are detected by clustering unusual read pairs.
(A) DNA inversion results in read pairs with same relative orientation and increased inner-mate distance. The invariable DNA segment is illustrated by a white box and the invertible segment by a black-and-white gradient box. Individual reads are illustrated for two distinct regions by colored arrows.
(B) Top cartoon: reads that overlap the junction of invertible sites are trimmed either on the 5-prime end (dark green) or 3-prime end (light green) during the alignment step. The trimmed portion of the reads are shown as dashed. Only the left side of the inversion site is illustrated. Bottom cartoon: histogram illustration of the frequency of starting position of all trimmed reads. Read trimmed at 5-prime end all start at the same genomic position and present high enrichment. Starting position of 3-prime end trimmed read is variable and present low enrichment.

1. Did the genomic inversions can be detected by looking for specific signatures following high-throughput deep sequencing of microbial genomes?
2. An atypical orientation of paired reads is produced when an inverted segment is sequenced and mapped back to the unchanged, reference sequence. The resulting read pairs have the same orientation. Additionally, the read pair has an increase in the inner-mate distance matching approximately the size of the inversion. Therefore, the presence of read pairs having the same orientation and higher mean inner-mate distance can be used as specific signature to detect genomic inversions. Therefore, small genomic inversions create specific and easily identifiable signatures in sequencing datasets.
3. Identifying invertible regions, recombination sites at single-nucleotide level can also be inferred from genomic next generation sequencing on Illumina's platforms.

Aimusa Arsu (Beer) SIMB/M 6137762

Method of next generation sequencing

- Random DNA sequence was generated using Bioinformatics Toolbox from Matlab R2017b (MathWorks)
- Paired-end Illumina high-throughput sequencing reads were simulated using ART-MountRainier-2016-06-05 using Illumina HiSeq 2500 and HiSeq 2000
- bwa version 0.7.13-r1126 used for reads allignment
- SAM file manipulation was done using SAMtools version 1.5
- Sequencing coverage was calculated using bedtools v2.26.0 and HTSeq version 0.9.1

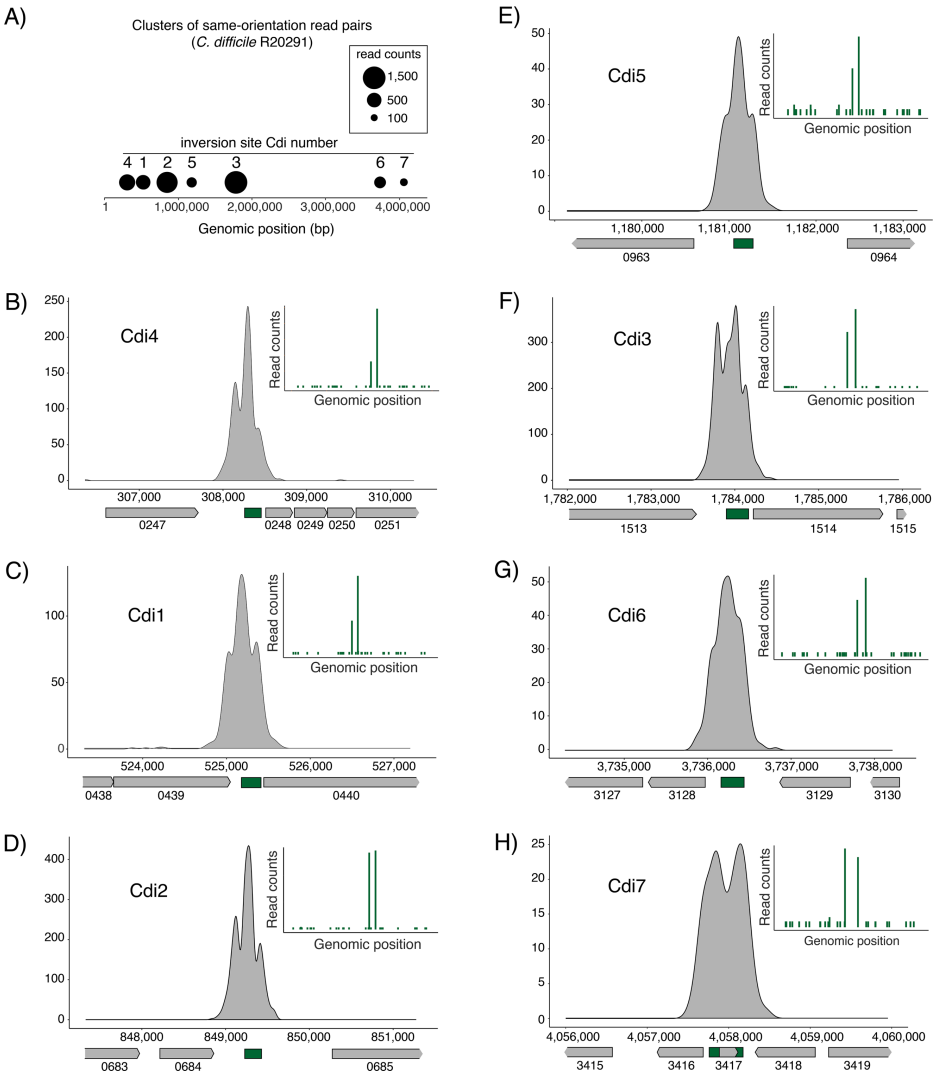

Fig 2. *C. difficile* R20291 genome harbors seven inversion sites.

(A) Seven distinct clusters (black circles) composed of unidirectional read pairs are detected in *C. difficile* R20291 genome. Each cluster is numbered following a previously described convention.

(B) (B-H) Illustration of unidirectional reads enrichment relative to the genomic on-scale context. The y axis represents read number and the x axis represent genomic position. Associated open reading frames are shown with arrowed boxes and labeled by the numeric value from their respective locus_tag number. The position of the invertible segment is illustrated by the dark green box relative to the gene content. The enrichment of 5-prime end clipped reads is shown in the separate plot at the top right corner. The two-major dark green bands correspond to left and right boundary of the invertible site.

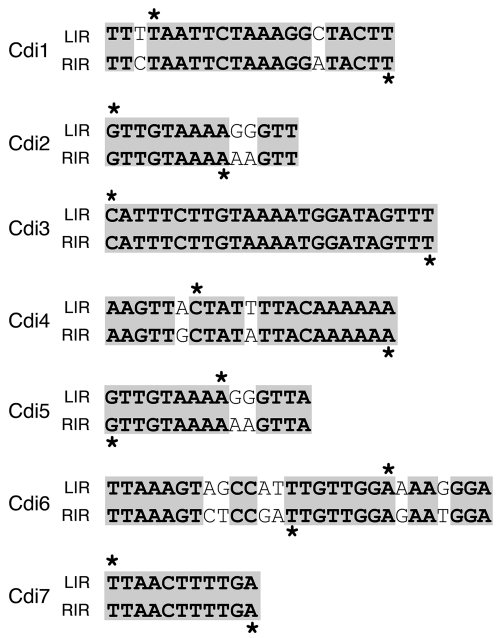

Fig 3. Terminal inverted DNA repeats border inversion sites in *C. difficile*. LIR stands for "left inverted repeat" and RIR stands for "right inverted repeat". Gray shading illustrates sequence homology. The asterisk (*) points to the nucleotide where recombination is taking place as identified by enrichment of 5-prime end clipped reads.

1. How inversion signatures are present in the *Clostridium difficile* genome?
   Did inversion signatures effect gene or function of gene in the *Clostridium difficile*?
2. Identified four short invertible regions designated "Cdi" (for *Clostridium difficile* inversion)
   Results

- Fig 2A revealed seven distinct groups with inversion sizes ≤5 kbp and variable read counts.
- Fig 2B–2G revealed strong enrichment of unidirectional read pairs in intergenic regions.
    - Cdi4 and Cdi1 correspond to the previously described flagella and cwpV switches respectively.
    - Cdi2 located in vicinity of a signaling protein (locus_tag CDR20291_0685) with conserved diguanylate cyclase (GGDEF), diguanylate phosphodiesterase (EAL), and sensor PAS domains.
    - Cdi5, CDR20291_0963 is the first gene in an apparent operon of three encoding a putative membrane bound O-acetyl transferase.
    - Cdi3 is located upstream of another putative signaling protein (locus_tag CDR20291_1514) containing both GGDEF and EAL domains.
    - Cdi6 is located upstream of an apparent operon of three genes encoding a classical two-component regulatory system. CDR20291_3127 encodes a putative signal transduction histidine kinase.

- Fig 2H spanned a small open reading frame.
    - Cdi7, spans a small open reading frame of 228 nucleotides encoding a hypothetical protein (locus_tag CDR20291_3417) containing a domain of unknown function DUF1413.
- Fig 3 show that since inverted repeats border the putative inversion segments, these regions may be recombining to generate an inversion event.

Collectively, these results suggest that there are more invertible sites in the *C. difficile* R20291 genome than previously thought; these sites may influence a variety of cellular functions based on the functional diversity of genes encoded in vicinity.

3. *In silico* random sequence generation for validation of inversion detection and cluster analysis.

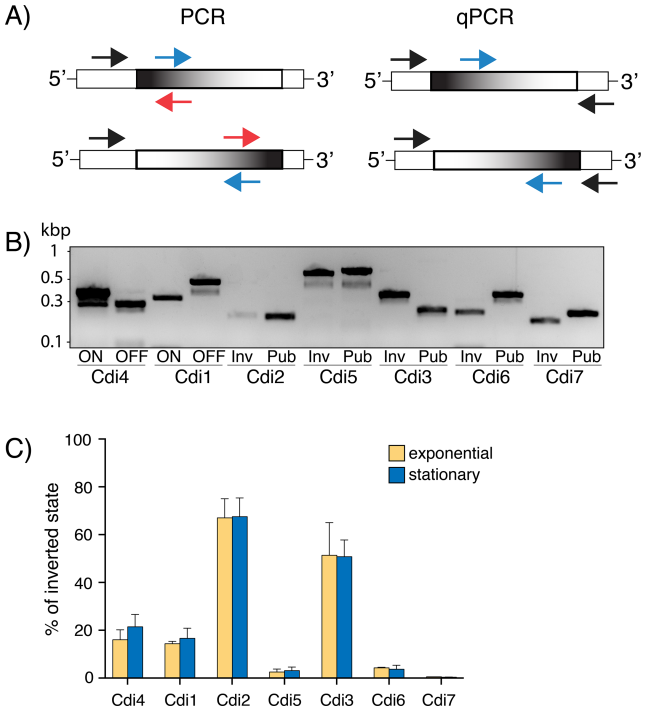

Fig 4. Detection and quantification of inversion events in *C. difficile*.
(A) Schematic representation of orientation-specific PCR and qPCR strategies. Invariable DNA region is illustrated by white and invertible region by black-and-white gradient box. Primers used for PCR amplification are shown as arrows.
(B) The agarose gel showing PCR products for either published (Pub) or inverted (Inv) orientation. For Cdi4 and Cdi1, permissive (ON) and restrictive (OFF) orientations for downstream gene expression have been determined elsewhere. DNA from 3 biological replicates was tested and one representative example is shown.
(C) Quantification of inverted orientation relative to the published orientation by qPCR. Presented are means from 3 biological replicates. Error bars correspond to 2 standard deviations.

1. If identified regions are subject to inversion during bacterial growth, can the method detect phase variable-like genomic inversions with wide range of input frequencies?
2. Fig 4B showed detection of two bands with the expected molecular weight corresponding to both orientations for all seven identified regions. This result confirmed that all invertible segments exist in both orientations in actively growing cells under normal lab conditions.
   Fig 4C revealed the percentage of the inverted state relative to the published orientation for *C. difficile* R20291 was determined during mid-exponential and early-stationary phase growth rate.

- Cdi4, flagellar switch, had % of inverted state = 16.0%, 12.6–20.2% (2SD or 95.45%CI) in actively growing cells.
- Cdi1, cwpVb switch, had % of inverted state = 14.3%, 13.4–15.4%.
- Cdi2 and Cdi3, switches detected upstream of both cyclic-di-GMP signaling proteins, had % of inverted state = 67.0%, 57.8–75.1% and 51.4%, 37.7–65.0%, respectively.
- Cdi7 had % of inverted state = 0.5%, 0.4–0.6%
- Cdi5 had % of inverted state = 2.6%, 1.7–3.8%
- Cdi6 had % of inverted state = 4.3%, 4.1– 4.6%

Collectively, these results demonstrated that method allows for accurate and genome-wide detection of phase variable-like genomic inversions with wide range of input frequencies.

3. Orientation-specific PCR on genomic DNA extracted from exponentially growing cells.
   Orientations were confirmed by DNA sequencing of the PCR products. The ΔΔCT method (used new primers with high and very similar amplification efficiencies to allow for relative quantification) was used because end-point PCR approach did not allow for quantification of the frequency at which different orientations exist within bacterial population.

Conclusion, this study shows that small genomic inversions, often associated with regulation of expression of neighboring genes, are prevalent in the bacterial world. Some species, including the human opportunistic pathogen *C. difficile* seems to have adopted this mode of regulation for a wide variety of functions including cell surface modification, intracellular signaling and environment sensing. As new bacterial strains are sequenced, this method will enable detection of novel inversions potentially providing valuable insights about bacterial evolution and lifestyle.