Data Science Minor Project

Area Code: 03, 05

Modelling the impact of DNA sequence repeats on gene inversions and gene-strand bias in bacteria

A Dissertation Submitted
in Partial Fulfilment of the Requirements
for the degree of

MINOR DEGREE

in

School of Data Science

by

Siva Subramanian A
(Roll No. IMS19217)

to

SCHOOL OF DATA SCIENCE
INDIAN INSTITUTE OF SCIENCE EDUCATION AND RESEARCH
THIRUVANANTHAPURAM - 695 551, INDIA
April 2023

DECLARATION

I, **Siva Subramanian A (Roll No: IMS19217)** hereby declare that, this report entitled “**Modelling the impact of DNA sequence repeats on gene inversions and gene-strand bias in bacteria**” submitted to Indian Institute of Science Education and Research Thiruvananthapuram towards the partial requirement of **Minor Degree** in **Data Science**, is an original work carried out by me under the supervision of **Dr. Sabari Sankar Thirupathy** and has not formed the basis for the award of any degree or diploma, in this or any other institution or university. I have sincerely tried to uphold academic ethics and honesty. Whenever a piece of external information or statement or result is used then, that has been duly acknowledged and cited.

Thiruvananthapuram - 695 551 Siva Subramanian A

April 2023

CERTIFICATE

This is to certify that the work contained in this project report entitled "**Modelling the impact of DNA sequence repeats on gene inversions and gene-strand bias in bacteria**" submitted by **Siva Subramanian A (Roll No: IMS19217)** to Indian Institute of Science Education and Research, Thiruvananthapuram towards the partial requirement of **Minor Degree** in **Data Science** has been carried out by him under my supervision and that it has not been submitted elsewhere for the award of any degree.

Thiruvananthapuram - 695 551 Dr Sabari Sankar Thirupathy

April 2023 Project Supervisor

ACKNOWLEDGEMENT

I want to extend a sincere and heartfelt obligation towards all the personages without whom the completion of the project was not possible. I express my profound gratitude and deep regard to Dr Sabari Sankar Thirupathy, IISER Thiruvananthapuram for his guidance, valuable feedback, and constant encouragement throughout the project. His valuable suggestions were of immense help. I sincerely acknowledge his constant support and guidance during the project.

I am immensely grateful to Atre Malhar Vivek for his constant support, guidance, and encouragement during the entire the span of this project. I am also grateful to the Indian Institute of Science Education and Research Thiruvananthapuram, for allowing me to do this project and providing all the required facilities.

Thiruvananthapuram - 695551 Siva Subramanian A

April 2023

ABSTRACT

Name of the student: **Siva Subramanian A** Roll No: **IMS19217**
Degree for which submitted: **Minor Degree** Department: **School of Data Science**
Thesis title: **Modelling the impact of DNA sequence repeats on gene inversions and gene-strand bias in bacteria**
Thesis supervisor: **Dr Sabari Sankar Thirupathy**
Date of thesis submission: **April 2023**

In bacteria, gene inversions that switch genes from leading to the lagging strands of DNA replication and vice versa could influence the degree of gene-strand bias. Here, we modeled bacterial genome in Python that examines how frequency and distribution of DNA sequence repeat (direct, inter-inverted, and intra-inverted repeats) affect gene inversions and, consequently, gene-strand bias. A null model with no selection pressure that allowed inversions without constraints greatly impacted gene-strand bias. The gene-strand bias stabilized at 50-50 (leading and lagging) irrespective of the initial parameters set. Imposing selection pressure onto the null model stabilizes the extant gene-strand bias.

Keywords:

Inversions, gene-strand bias

CONTENTS

List of Figures

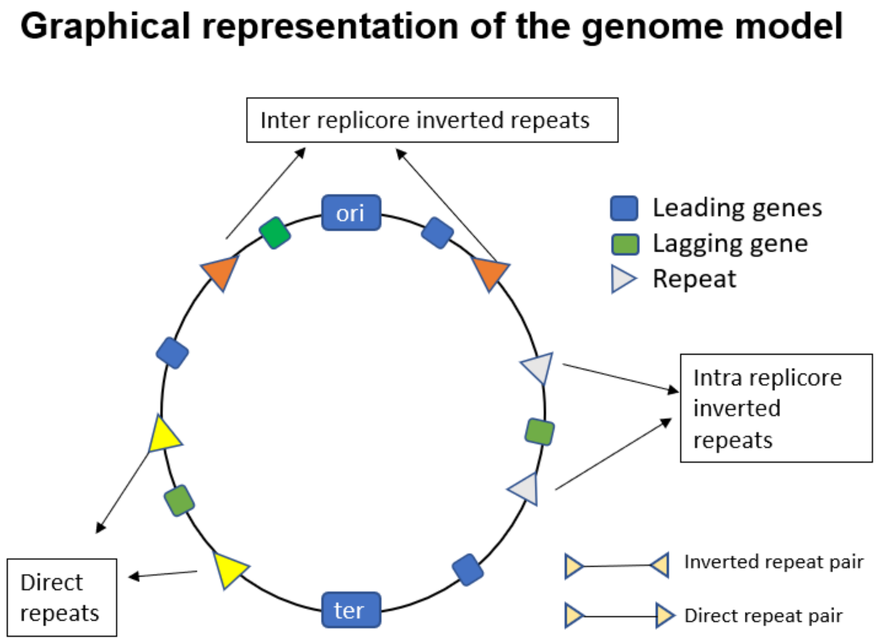
LIST OF TABLES

Chapter 1

Introduction

1.1 Background

In bacterial genomes, repetitive DNA sequences are present across the genome, and two main types are direct and inverted repeats. Direct repeats are identical sequences repeated in a head-to-tail fashion, while inverted repeats are identical sequences oriented in opposite directions. Among those repeats, inverted repeats are capable of inversions, defined as a type of chromosomal rearrangement where a segment of DNA is reversed and reinserted back into the same chromosome. These inversions impact gene-strand bias, it refers to the uneven distribution of genes between the leading and lagging strands of DNA. Inverted repeats can be classified into two main categories based on their location and orientation within the genome: intra-inverted located within the same replicore and inter-inverted repeats located between two replicore. A replicore is a set of DNA sequences from the ori to ter site. Genes can be located on either the leading or lagging strand. Studying the relationship between inversions and gene strand bias is important for understanding how bacteria adapt to changing environments and evolve new traits. The model is similar to the natural bacterial genome, considering the bacterial genome as a circular genome with direct repeats, inverted repeats, leading strand genes, lagging strand genes, ori, and ter sites. Studying the distribution and characteristics of inverted repeats and the impact of inversions on gene-strand bias in bacterial genomes can provide the evolution of bacterial genomes, adaptation, and selection of the stable gene-strand bias. The constructed Python model can provide a valuable tool for exploring these relationships and testing hypotheses about the role of inverted repeats in shaping bacterial genomes.


Graphical representation of the genome model
Inter replicore inverted repeats
ori
Leading genes
Lagging gene
Repeat
Intra replicore inverted repeats
Inverted repeat pair
Direct repeat pair
Direct repeats
ter

(Fig 1: Graphical representation of the built bacterial genome model)

(a) Repeats oriented in the same direction are direct repeats - no inversions

(b) Repeats oriented in opposite directions are inverted repeats - undergo inversions

1.2 Model Design

Two models were built

(i) Null model: The null model with no selection pressure on inversions and a model with selection pressure constraints. The null model was built with repeats and genes stochastically distributed across the two replicores to mimic the bacterial genome, which didn't have constraints on inversions.

A variant of the null model with no selection constraints where the null model + the number of genes within the inverted repeats were sampled from a Gaussian distribution N (μ, σ) with mean μ and standard deviation σ. Specifically, the focus was on the number of genes between the intra-inverted repeats sampled from a normal distribution. This distribution reflects the natural variation in gene distribution that occurs in bacterial genomes and can be used to simulate different scenarios in which an inversion has occurred. Additionally, by manipulating the normal distribution parameters, such as changing the mean or standard deviation, it is possible to explore how different gene distributions affect gene strand bias and other factors related to bacterial evolution and adaptation.

(ii) Model with selection pressure constraint:

(a) Inversions causing drastic gene imbalance between ori-ter and ter-ori gene counts are omitted:

Ori-ter gene counts refer to the number of genes from ori to ter sites in the first replicore. Ter-ori gene counts refer to the number of genes from ter-ori sites in the second replicore. Inversions can disrupt the regular gene distribution pattern, leading to gene count imbalance between ori-ter and ter-ori regions. This can significantly impact bacterial evolution and adaptation, as genes in specific regions may be essential for survival or confer advantageous traits. Therefore, it is important to consider how inversions affect gene counts when studying gene-strand bias. Thus, drastic imbalances in gene counts caused by inversions are omitted from the analysis. This was considered to avoid confounding factors that may arise from extreme imbalances in gene counts.

(b) Selection pressure with inversion disparity limit on inversions:

Inversion disparity score for an inversion based on the leading and lagging genes between an intra-inverted repeat pair. The inversion disparity score is a measure used to assess an inversion event's potential impact on a bacterial species' genome. It is calculated based on the number of leading and lagging genes between an intra-inverted repeat pair. Leading genes are located on the same strand as the replication origin (ori) while lagging genes are located on the opposite strand.

The Inversion disparity score (IDS) is defined as the resulting sum of the genes determined by assigning a positive value (+1) to each leading gene and a negative value (-1) to each lagging gene between the intra-inverted repeated pair. The score can be positive or negative depending on the relative number of leading and lagging genes.

An inversion disparity limit (IDL) is set to determine if an inversion event can occur. If the IDS falls within the range of IDL (e.g., IDL - 5, then the range is -5 to +5), the inversion could happen. If the score exceeds the inversion disparity limit, the inversion is considered too disruptive and unlikely to occur. This is because it reflects the gene imbalance caused by the inversion, and it is detrimental to the overall genome structure and function if it exceeds the established limit. Therefore, the inversion disparity limit was set as the acceptable disruption amount. If the score of an inversion falls within this acceptable range, it is more likely to be a natural and tolerable occurrence, and it also ensures that the genome remains functional and stable

In summary, this inversion disparity score is a valuable tool for predicting the potential impact of an inversion event on a bacterial genome. Expanding the inversion disparity limit for gene inversion scores allows for more inversions and provides a valuable tool for studying the effects of inversions on the gene-strand bias. It can help researchers to understand the evolutionary dynamics of these species better.

Chapter 2

Model building - Parameters, pipeline, graphical abstract

2.1 Model construction

Inversions in the model are considered chromosomal rearrangements where a segment of DNA is reversed and reinserted back into the same chromosome. The inversions for one generation are detailed in the graphical representation.

Before constructing the model, prior literature reviews were done to glimpse the number, location, and distribution of repeats across the bacterial genomes. The percentage of the genes on leading and lagging strands was also considered. To model the impact of inversions on the gene-strand bias, bacterial genome representation with variables and parameters was defined using Python, and the conditions were detailed.

Parameters are taken for a null model:

(a) Direct and inverted repeats pair percent,

(b) Inter and intra-inverted repeat,

(c) Leading and lagging strand gene percent in both replicores,

(d) Gene distribution between intra-inverted repeat pairs in replicores - N (μ, σ)

(e) Stochastic size and location of the inverted repeat, the distance between the repeat and nearby genes, and the orientation of the genes relative to the repeat.

(f) Ori and ter sites are dynamic - these elements are considered for inversion.

These parameters simulate different scenarios in which an inversion has occurred and observe the resulting changes in gene strand bias. When constructing the model, different types of inversions (inter and intra) were considered, such as those that occur within the same replicore (intra) or between different replicons (inter), and how this impact on the gene-strand bias was studied. Selection pressures were imposed on the inversions to study how this affects gene-strand bias.

Additional parameters for the model with selection pressure:

(a) Inversions causing drastic gene imbalance between ori-ter and ter-ori gene counts are not included.

(b) Inversion disparity score for an inversion based on the leading and lagging genes between an intra-inverted repeat pair

Pipeline for the Null model building:

(a) Initializing parameters:

(i) Inverted repeat pairs were denoted as swapped string elements such as A1 and a1, A2 and a2, and so on. Directed repeats were denoted as similar string elements such as A3 and A3, a4 and a4, and so on. The percentage of intra-inverted repeats in both replicore was equal to the spread of the stochastic distribution of repeat pairs across two replicores.

(ii) Genes were denoted as integers, with leading strand genes as negative integers and lagging strand genes as positive integers. The percentage of leading and lagging genes in both replicore was equal and the distribution of genes is stochastic across two replicores.

(b) Distribution of genes in repeats:

(i) Genes were distributed stochastically across the intra-inverted repeat pairs.

(ii) Genes were distributed normally with mean μ and standard deviation σ in between the intra-inverted repeat pairs N (μ, σ).

(c) The null model was conditioned with no selection pressure on inversions:

(i) Genomic elements and repeats are all randomly distributed across the genome.

(ii) Ori & ter sites are fixed and dynamic.

(d) Conditions for inversions:

(i) Selection of a repeat for inversion is randomized. If the repeat selected is a direct repeat, then no inversion happens; if it is an inverted repeat, inversions happen.

(ii) Short-fragment is always considered for the inversion, as it is highly unlikely for the large fragment to undergo inversion.

Pipeline for the model with selection pressure:

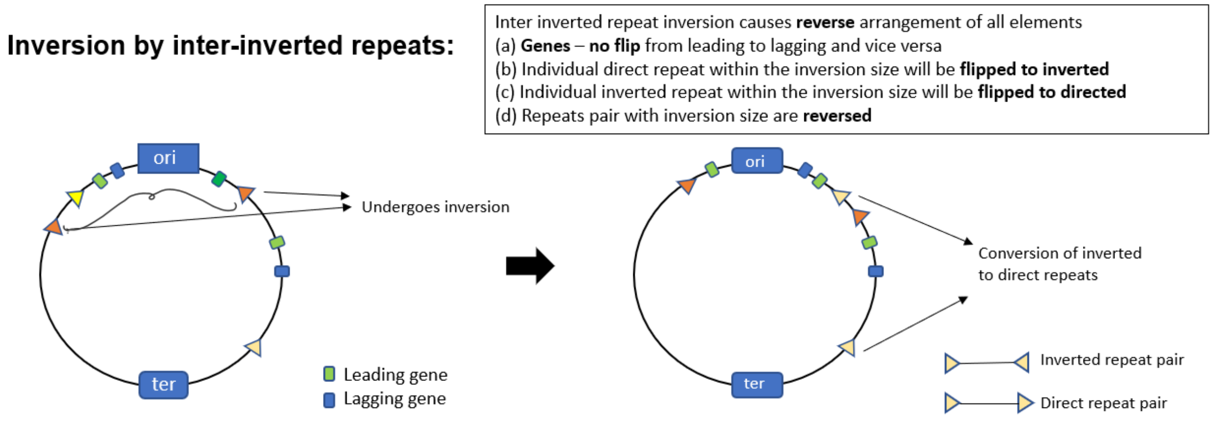
Selection pressure constraints are added additionally to the null model:

(a) **Inversions causing drastic gene imbalance between ori-ter and ter-ori gene counts are omitted:** The ideal threshold ratio for the extreme gene count imbalance was kept at 0.25 if the ratio of ori-ter and ter-ori gene counts is within 0.25 then that inversion is accepted. (This value can be manipulated)

(b) **Selection pressure on inversion disparity score:** Inversion disparity score for an inversion based on the leading and lagging genes between an intra-inverted repeat pair

Graphical abstract of an inversion:

(a) Example of how inversion by inter-inverted repeat pair:

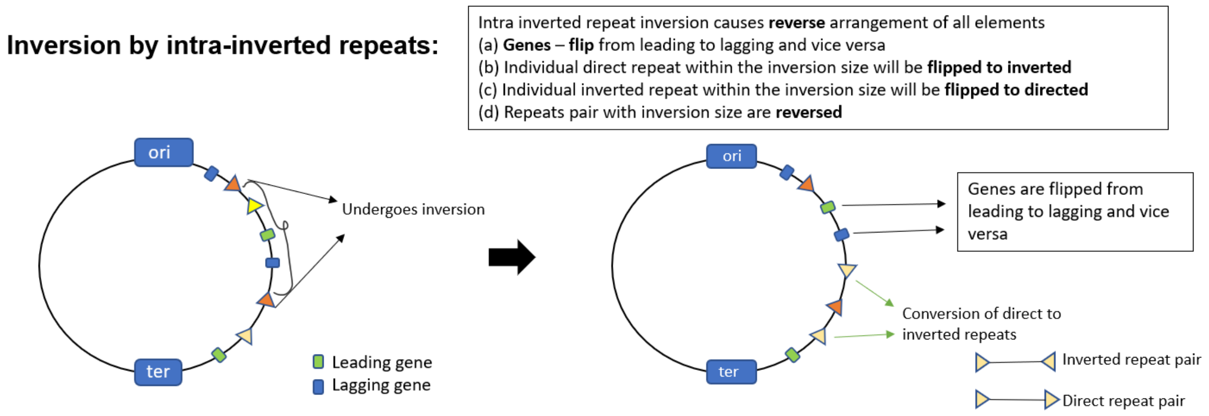


Inversion by inter-inverted repeats:

Inter inverted repeat inversion causes **reverse** arrangement of all elements
(a) **Genes – no flip** from leading to lagging and vice versa
(b) Individual direct repeat within the inversion size will be **flipped to inverted**
(c) Individual inverted repeat within the inversion size will be **flipped to directed**
(d) Repeats pair with inversion size are **reversed**

(Fig 2: Inversion by inter-inverted repeats)

(b) Example of how inversion by intra-inverted repeat pair:



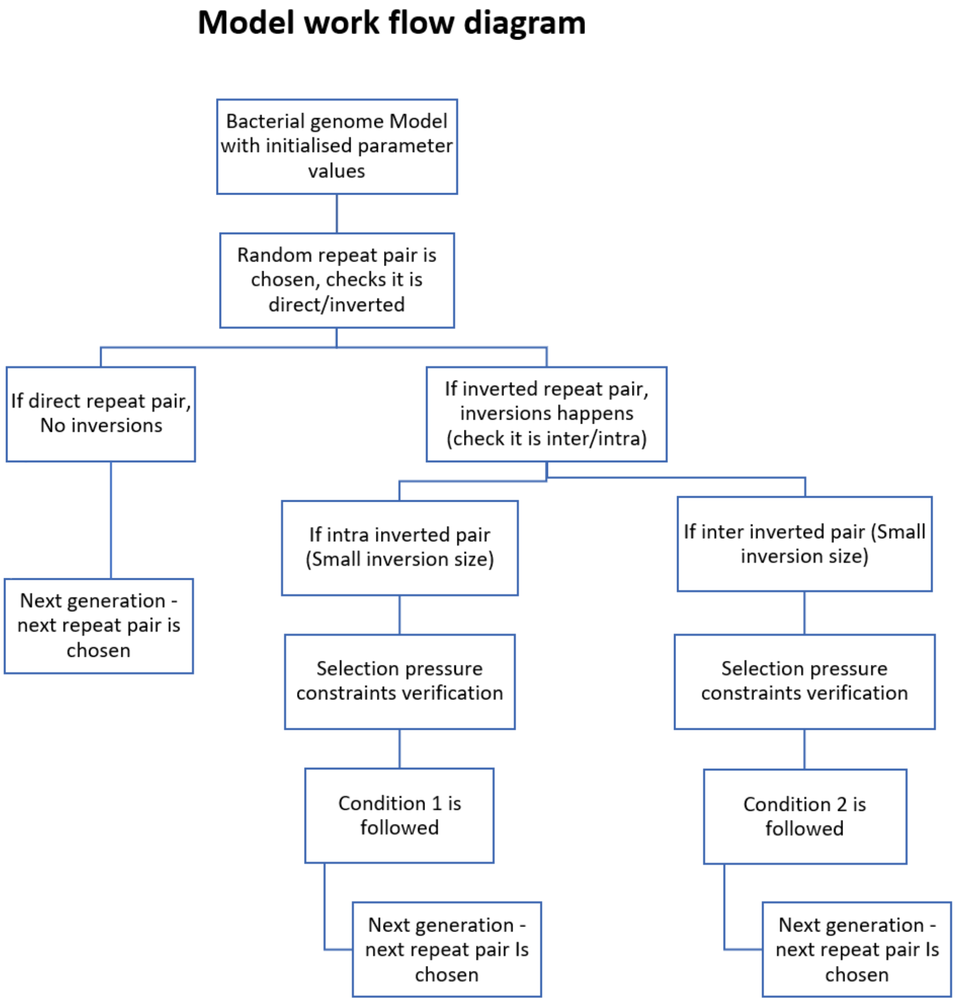
Inversion by intra-inverted repeats:

Intra inverted repeat inversion causes **reverse** arrangement of all elements
(a) **Genes** – **flip** from leading to lagging and vice versa
(b) Individual direct repeat within the inversion size will be **flipped to inverted**
(c) Individual inverted repeat within the inversion size will be **flipped to directed**
(d) Repeats pair with inversion size are **reversed**

Genes are flipped from leading to lagging and vice versa

(Fig 3: Inversion by intra-inverted repeats)

2.2 Graphical abstract of the model


Model work flow diagram
Bacterial genome Model with initialised parameter values
Random repeat pair is chosen, checks it is direct/inverted
If direct repeat pair, No inversions
If inverted repeat pair, inversions happens (check it is inter/intra)
Next generation - next repeat pair is chosen
If intra inverted pair (Small inversion size)
If inter inverted pair (Small inversion size)
Selection pressure constraints verification
Selection pressure constraints verification
Condition 1 is followed
Condition 2 is followed
Next generation - next repeat pair Is chosen
Next generation - next repeat pair Is chosen

(Fig 4: Work flow diagram)

Condition 1:

Intra-inverted repeat inversion causes the **reverse** arrangement of all elements

(a) **Genes** – **flip** from leading to lagging and vice versa

(b) Individual direct repeat within the inversion size will be **flipped to inverted**

(c) Individual inverted repeat within the inversion size will be **flipped to directed**

(d) Repeats pair within inversion size are **reversed**

Condition 2:

Inter-inverted repeat inversion causes the **reverse** arrangement of all elements

(a) Genes – **no flip** from leading to lagging and vice versa

(b) Individual direct repeat within the inversion size will be **flipped to inverted**

(c) Individual inverted repeat within the inversion size will be **flipped to directed**

(d) Repeats pair within inversion size are **reversed**

This is the working flow of the null model.

To build the second model, selection pressure constraints on inversions were added to the null model. They are explained below:

Selection pressure constraints verification:

(a) Inversions causing drastic gene imbalance between ori-ter and ter-ori gene counts are omitted: The ideal threshold ratio for the extreme gene count imbalance was kept at 0.25 if the ratio of ori-ter and ter-ori gene counts is within 0.25 then that inversion is accepted. (This value can be manipulated)

(b) Selection pressure on inversion disparity score: Inversion disparity score for an inversion based on the leading and lagging genes between an intra-inverted repeat pair. Inversion disparity limit was set at 10, 25, and 50

Chapter 3

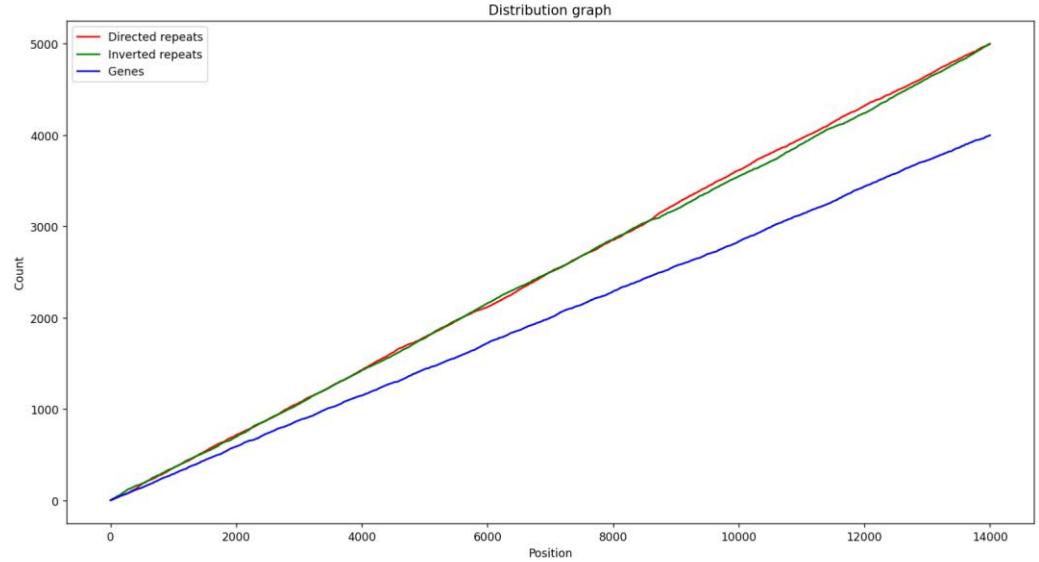
Validation – repeats and gene distribution

3.1 Stochastic distribution genome model

(a) Bacterial genome null model:

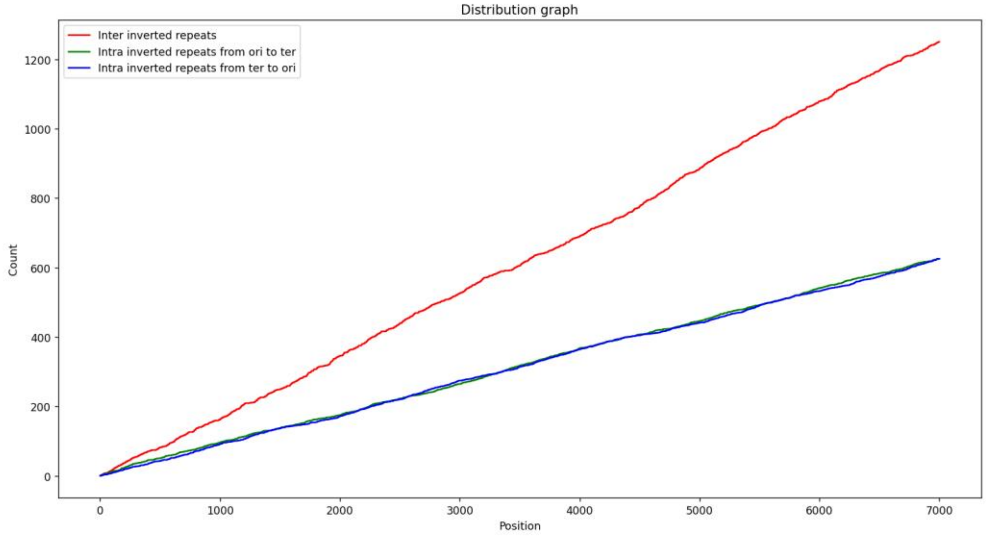
(i) Repeat pairs - 5000 (Direct repeat - 2500 pairs, Inverted repeat - 2500 pairs)

(ii) Genes - 4000



(Fig 5: Stochastic distribution of repeats and the genes across the genome)

(iii) Inverted repeat pairs - 2500 (Inter-inverted 60%, Intra-inverted in first replicore 20%. Intra-inverted in second replicore 20%)



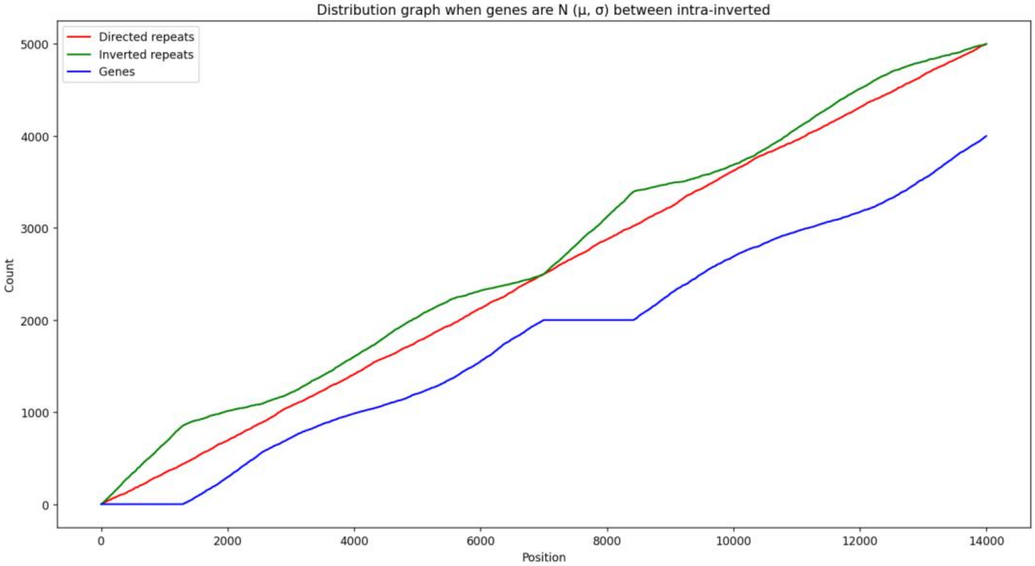
(Fig 6: Stochastic distribution of inter and intra-inverted repeats across the genome)

This validates that the repeat and genes are stochastically distributed across the genome in this specific parametric null model, with 14000 elements (5000 repeat pairs + 4000 genes). This is stochastic distribution as the graph slope increases along the genome, indicating the distribution of elements across the genome.

3.2 Normal distribution genome model

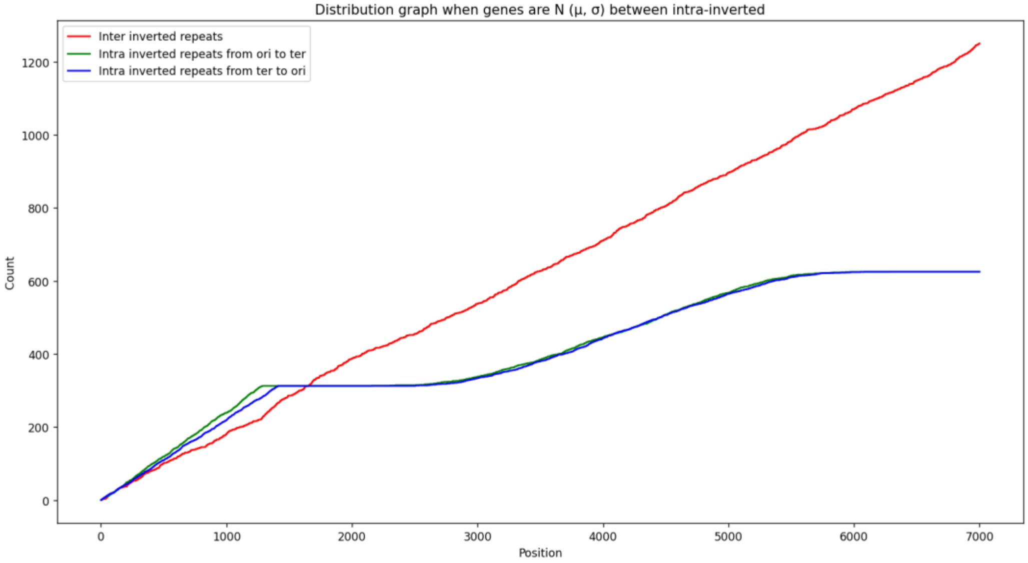
Bacterial genome model with normal distribution of genes within intra-inverted repeat pairs:

- **(i)** Repeat pairs - 5000 (Direct repeat - 2500 pairs, Inverted repeat - 2500 pairs)
- **(ii)** Genes – 4000



(Fig 7: Distribution of repeats and the genes when the genes are normally sampled between the intra-inverted repeat pairs across the genome)

(iii) Inverted repeat pairs - 2500 (Inter-inverted 60%, Intra-inverted in first replicore 20%. Intra-inverted in second replicore 20%)



(Fig 8: Distribution of inter and intra-inverted repeats when the genes are normally sampled between the intra-inverted repeat pairs across the genome)

This validates that the genes are normally distributed N (μ, σ) = N (0.25, 0.05) across the genome in this specific parametric model, with 14000 elements (5000 repeat pairs + 4000 genes). Inter-inverted are stochastically distributed; only the genes within the intra-inverted are normally distributed. This is the validation that the distribution of elements is all across our required distribution of genes within repeat pairs.

Chapter 4:

Results of simulations - Working flow, simulation graphs and inferences

4.1 Null model simulations

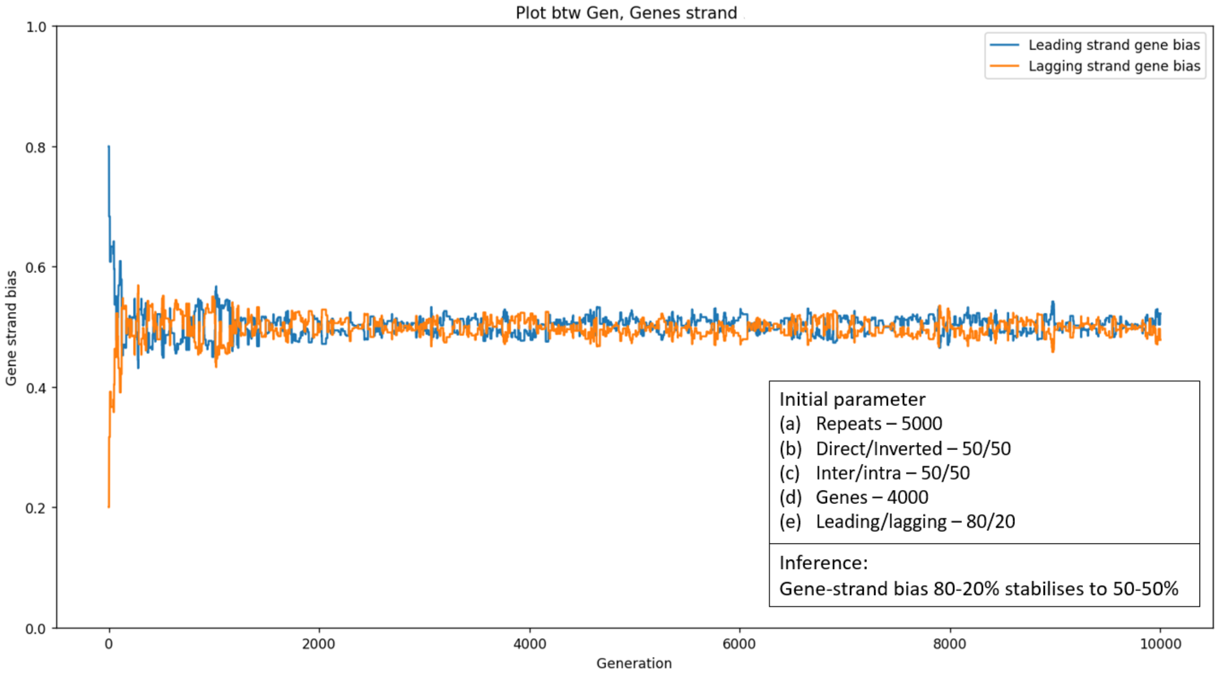
Different variations in initial gene-strand bias with standard repeat pair number and genes stabilizes in 50% - 50% gene-strand bias. This holds true according to the evolutionary model of inversions caused by repeats, as the null model has no selection pressure on the inversions and allows all random inversions to be accessible.

Most of the simulations were run with the following parameters:

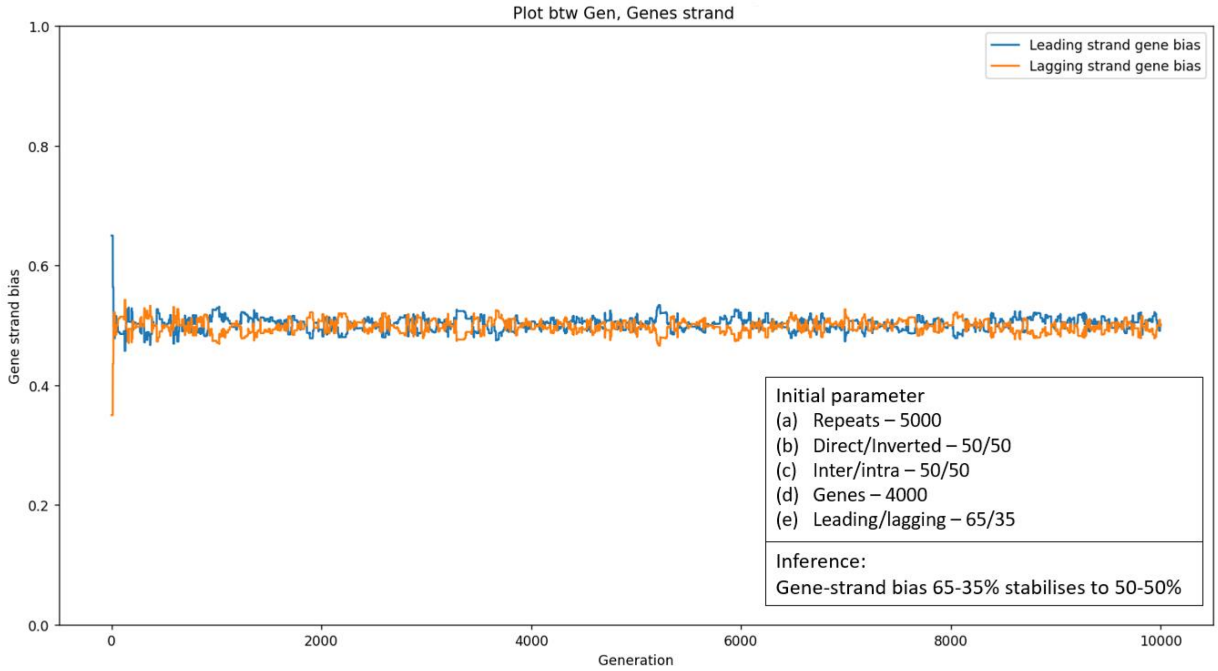
Simulations entries	**Parameter's values**
Repeats pairs	5000
Inverted repeat percent	50%, 60%, 67% (inter/intra repeat= 1:1, 1.5:1, 2:1)
Genes	4000
Initial gene-strand bias: Leading strand genes	20-80% range by ± 5%
Generations	10,000

(i) Working flow: There are 14000 elements inside the constructed circular bacterial genome model for the above tabular column parameter's values. The elements obey the inverted repeat percent, leading strand gene bias among them. The model runs for 10000 generations where each generation, a randomized repeat pair irrespective of direct/inverted, will be chosen. Conditions for inversions will be crosschecked (it will undergo inversions only when it is an inter/intra inverted repeat) by the model and inversions (rearrangements happens according to the inversion conditions).

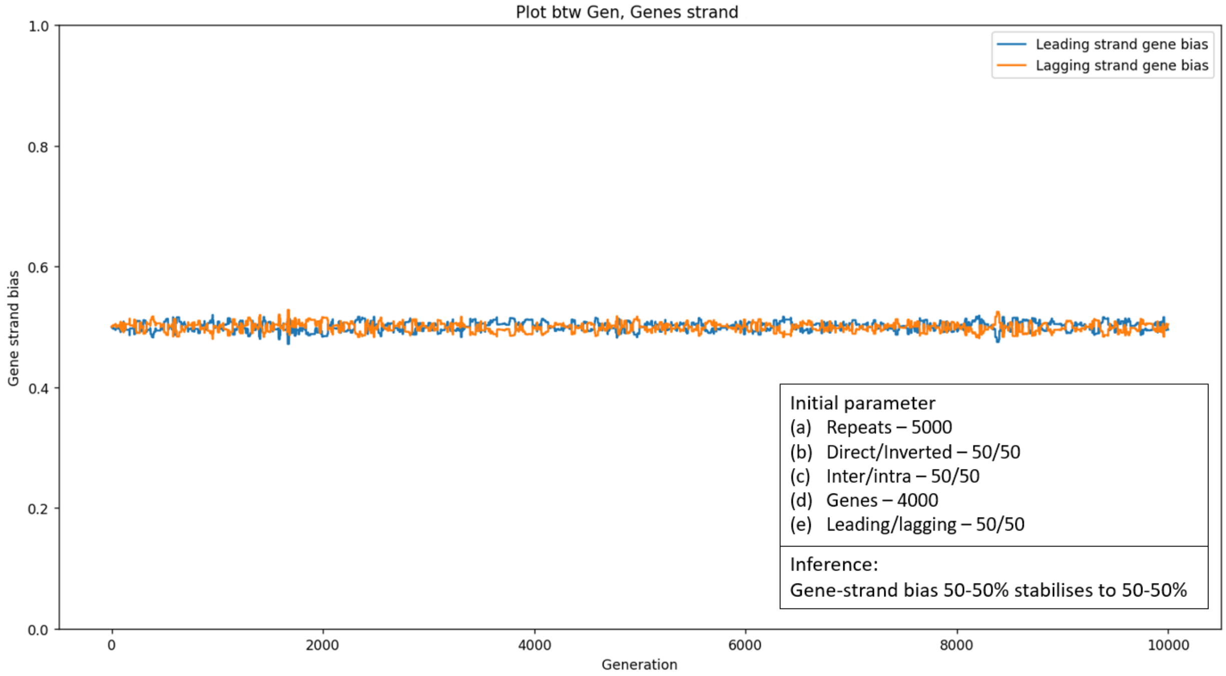
(ii) Simulations of the null model:



(Fig 9: Null model results with specified parameter set and inference)



(Fig 10: Null model results with specified parameter set and inference)



(Fig 11: Null model results with specified parameter set and inference)

These figures are also the representative results of the null model without selection pressure for the different permutations and combinations of the parameters from the below table.

Simulations entries	**Parameter's values**
Repeats pairs	500, 1000, 2000, 2500, 5000
Inverted repeat percent	50%, 60%, 67% (inter/intra repeat= 1:1, 1.5:1, 2:1)
Genes	400, 800, 1600, 2000, 4000
Initial gene-strand bias: Leading strand genes	20-80% range by ± 5%
Generations	10,000

(iii) Model results: Since there is no selection pressure on the inversions, no matter the initial gene-strand bias, it will stabilize to 50%-50% gene-strand bias after 10000 generations of inversions. It is found that even when the variables of the repeat number, direct/inverted repeats, inter/intra inverted repeats/ gene number, and initial leading/lagging strand genes are changed, the null model stabilizes the initial gene-strand bias to 50-50%, as this is valid to the predicted results from the evolutionary aspects. This holds true according to the evolutionary model of inversions caused by repeats, as the null model has no selection pressure on the inversions and allows all random inversions to be accessible

4.2 Model with selection pressure simulations

Variations in gene-strand bias stabilize to the initial bias with standard repeat pair number and genes. No change is observed, but a 50-50% balance is achieved when the penalty limit decreases gradually from 10 to 50.

But for the standardization with the real-world bacterial genome, most of the simulations were run with the following parameters:

Simulations entries	**Parameter's values**
Repeats pairs	5000
Inverted repeat percent	50%, 60%, 67% (inter/intra repeat= 1:1, 1.5:1, 2:1)
Genes	4000
Initial gene-strand bias: Leading strand genes	20-80% range by ± 5%
Generations	10,000

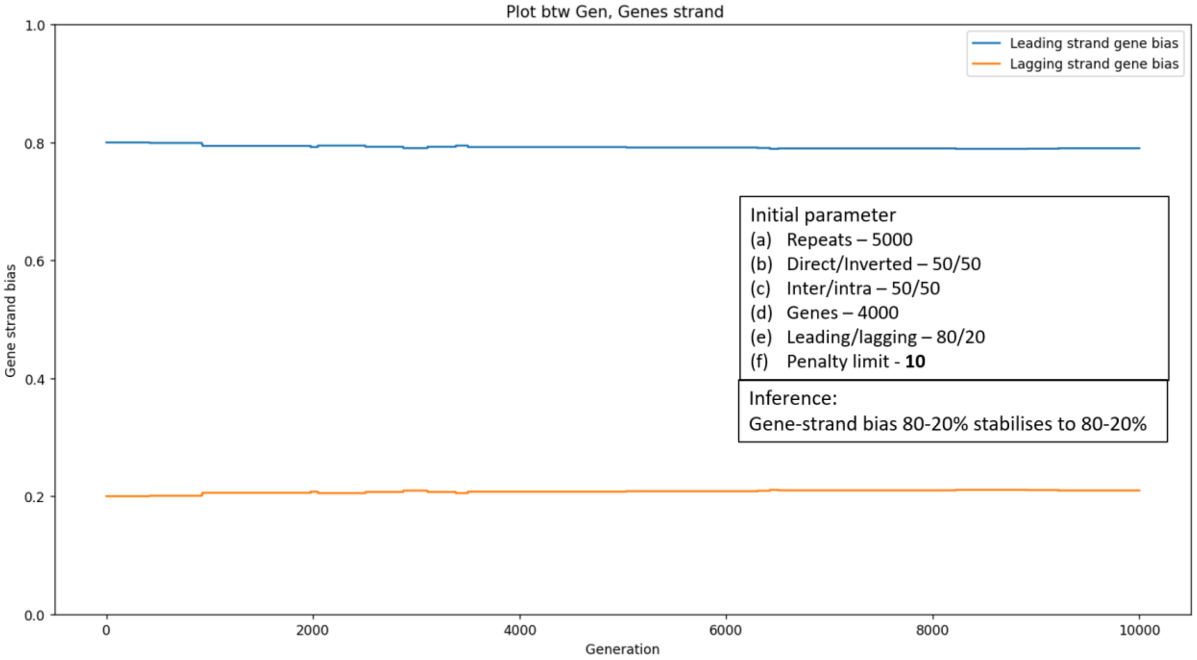
Selection pressure constraints verification:

(a) Inversions causing drastic gene imbalance between ori-ter and ter-ori gene counts are omitted: The ideal threshold ratio for the extreme gene count imbalance was kept at 0.25 if the ratio of ori-ter and ter-ori gene counts is within 0.25 then that inversion is accepted. (This value can be manipulated)

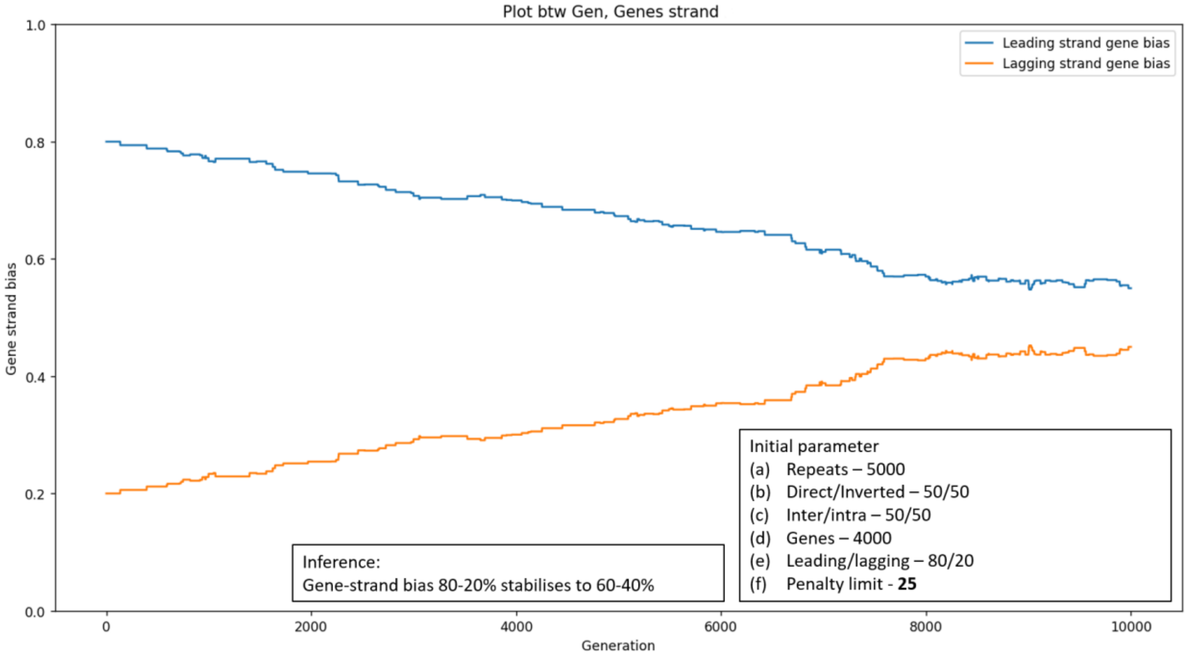
(b) Selection pressure on penalty score: Penalty score for an inversion based on the leading and lagging genes between an intra-inverted repeat pair. Penalty limit was set at 10, 25, and 50

(i) Working flow: 14000 elements are inside the constructed circular bacterial genome model for the above tabular column parameter's values. The elements obey the inverted repeat percent, leading strand gene bias among them. The model runs for 10000 generations where each generation, a randomized repeat pair irrespective of direct/inverted, will be chosen. Selection pressure verification is performed, followed by the conditions for inversions will be crosschecked (it will undergo inversions only when it is an inter/intra inverted repeat) by the model and inversions (rearrangements happens according to the inversion conditions).

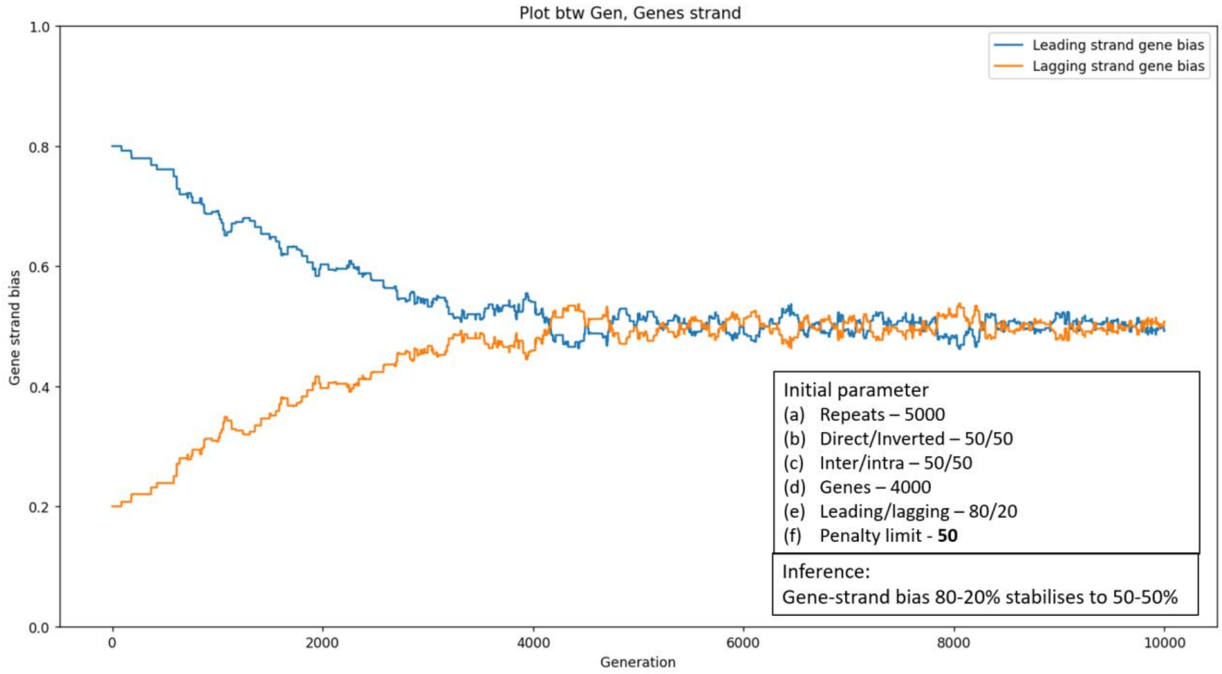
(ii) Simulations of the model:



(Fig 12: Model results with specified parameter set (Penalty limit - 10) and inference)



(Fig 13: Model results with specified parameter set (Penalty limit - 25) and inference)



(Fig 14: Model results with specified parameter set (Penalty limit - 50) and inference)

(iii) Model results:

Inferences from the simulations: When the inversion disparity limit is at 10, the initial gene-strand bias 80-20% stays at 80-20%. When the inversion disparity limit is increased to 25, the initial gene-strand bias of 80-20% comes to 60-40%, and when the inversion disparity limit is further increased to 50, the initial gene-strand bias of 80-20% comes to 50- 50%

When the inversion disparity limit is gradually decreased from 10 to 25 and then to 50, this allows more inversions with the inversion score to happen; thus, this decreases the gene-strand bias as this is valid to the predicted results from the evolutionary aspects. This holds true according to the evolutionary model of inversions caused by repeats, as the model has selection pressure on the inversions as it takes selection constraints into the model.

Model showed similar results when different permutations and combinations of the parameters was used from the below table:

Simulations entries	**Parameter's values**
Repeats pairs	500, 1000, 2000, 2500, 5000
Inverted repeat percent	50%, 60%, 67% (inter/intra repeat= 1:1, 1.5:1, 2:1)
Genes	400, 800, 1600, 2000, 4000
Initial gene-strand bias: Leading strand genes	20-80% range by ± 5%
Generations	10,000

Gene inversions can be restricted by selection pressure. Inversions that cause significant differences in gene counts between ori-ter and ter-ori are not allowed. A penalty score is used to evaluate the suitability of inversions based on the orientation of genes. This score is limited by a set penalty limit of either 10, 25, or 50.

Chapter 5:

Discussion

5.1 Interpretation of the results

(a) Null model with no selection stabilizes the gene-strand bias to 50-50% and this predicts the evolutionary aspects model as no selection is imposed on the inversions, thus leading and lagging strand genes stabilizes to 50-50%.

(b) The model with selection pressure, imposed selection on the inversion based on gene imbalance and inversion score, thus the gene-strand bias is not affected. When the stringent inversion disparity score is low, gene strand bias remains constant and increasing inversion disparity limit, gene-strand bias decreases.

5.2 Strengths of the model

(a) The model was built with less complex features from the real-world bacterial genomic model, considering only the repeat pairs and the distribution of genes within for consideration - thus, a simplistic model studying the inversion by repeat pairs affects the gene-strand bias.

(b) Evolutionary constraints on how bacterial gene-strand bias has evolved can be studied with the help of imposed selection pressure onto the constraints.

(c) By playing around with the parameter's value, one can find how the repeats - number, size, location, and genes - number and distribution affect the gene-strand bias.

(d) N number of models with specific repeats and genes can be made, and their gene-strand bias can be analyzed.

5.3 Limitations of the model

(a) Bacterial genomic sequences (A, T, C, G) were not considered to study the gene-strand bias - instead, the repeat pairs, genes, ori, and ter were considered DNA blocks subjected to inversions.

(b) The model does not consider GC skewness, gene gain-loss, and transcription-replication machinery collisions, which also impact the gene-strand bias.

5.4 Suggestions for future improvements

(a) To find the optimized penalty limit for the bacterial genome with specific repeats and genes, model can be extended using maximum likelihood estimator but since this is computationally exhaustive (more optimization on the code should be performed)

(b) Population genetics - fitness values for individual leading and lagging genes can be introduced to analyze how this affects the gene-strand bias.

(c) Real life bacterial genome can be comparatively analyzed along with the model, DNA sequences and transcription-replication machinery can be introduced into this model to study how these additional features modulate the gene-strand bias.

Bibliography

1) van Belkum A, Scherer S, van Alphen L, Verbrugh H. Short-sequence DNA repeats in prokaryotic genomes. Microbiol Mol Biol Rev. 1998 Jun;62(2):275-93. doi: 10.1128/MMBR.62.2.275-293.1998. PMID: 9618442; PMCID: PMC98915.
2) Achaz G, Coissac E, Netter P, Rocha EP. Associations between inverted repeats and the structural evolution of bacterial genomes. Genetics. 2003 Aug;164(4):1279-89. doi: 10.1093/genetics/164.4.1279. PMID: 12930739; PMCID: PMC1462642.
3) Ussery, David W.; Wassenaar, Trudy; Borini, Stefano (2008-12-22). "Word Frequencies, Repeats, and Repeat-related Structures in Bacterial Genomes". Computing for Comparative Microbial Genomics: Bioinformatics for Microbiologists. Computational Biology. Vol. 8 (1 ed.). Springer. pp. 133–144. ISBN 978-1-84800-254-8.
4) "Inverted Repeats and Microsatellites: Genes in Pursuit of Transposable Elements" by Wenliang Wang and Anna M. Chiara, Genes 2018, 9, 558; doi:10.3390/genes9110558
5) "Inversions and their Consequences: Reversing our Thinking" by Gabriel E. Zentner and Laura A. Carrel, Trends in Genetics, Volume 36, Issue 11, 817 - 827, doi:10.1016/j.tig.2020.07.006
6) "Genomic Rearrangements: Cause and Consequence" by Eugene W. Myers and Haixu Tang, Annual Review of Genetics, Vol. 49: 397-427, doi:10.1146/annurev-genet-112618-043614
7) "Selection against deleterious inversions in small populations: balancing selection and genetic drift" by Nicolas Galtier and Nicolas Bierne, Genetics Research, Volume 82, Issue 3, 201-211, doi:10.1017/S0016672305007981