Studies on the distribution and mutation of microsatellites in pathogenic bacterial genomes

Thesis submitted to

प्रज्ञानं ब्रह्म

Manipal
INSPIRED BY LIFE

Manipal University
Manipal, INDIA

For the Degree of

Doctor of Philosophy

Registration Number: 050100005

DECLARATION

The research work embodied in this thesis entitled "**Studies on the distribution and mutation of microsatellites in pathogenic bacterial genomes**", has been carried out by me at the Centre for DNA Fingerprinting and Diagnostics, Hyderabad, under the guidance of **Dr. H. A. Nagarajaram**. I hereby declare that this work is original and has not been submitted in part or full for any other degree or diploma of any other university or institution.

Pankaj Kumar
Laboratory of Computational Biology
Centre for DNA Fingerprinting and Diagnostics
Hyderabad 500 001.

CERTIFICATE

This is to certify that the thesis entitled, "**Studies on the distribution and mutation of microsatellites in pathogenic bacterial genomes**" submitted by **Mr. Pankaj Kumar** for the Degree of **Doctor of Philosophy** to **Manipal University** is based on the work carried out by him at the Centre for DNA Fingerprinting and Diagnostics, Hyderabad. This work is original and has not been submitted in part or full for any degree or diploma of any other university or institution.

Dr. H. A. Nagarajaram	**Dr. Shekhar C. Mande**
Thesis Supervisor	Dean, Academics
CDFD, Hyderabad.	CDFD, Hyderabad.

Dedicated to My Parents, wife and jigisha

Acknowledgements

It is my immense pleasure to utilize this space of my work to acknowledge all those who directly or indirectly brought light to this work. It is my pleasure to express my sincere gratitude to all the people who have helped and inspired me directly or indirectly in their various capacities during my doctoral study.

First and foremost I thank my mentor Dr. H. A. Nagarajaram for allowing me to work in his laboratory, for his full fledged cooperation, endless patience, affection and constructive suggestions. He trained me in the ethics of science. His critical comments always worked as a booster dose for me to work in a better fashion. I feel fortunate to have him as my mentor who has treated me like a child and forgiven my flaws.

I gratefully acknowledge Dr. J. Gowrishankar, Director CDFD and Prof. Seyed E. Hasnain, former Director CDFD, for providing all the research facilities and for being a constant source of encouragement. I pay my regards to Dr. Shekhar C. Mande for his guidelines to complete this highest academic degree. I am also grateful to the members of my doctoral committee, Dr. Akash Ranjan and Dr Sanjeev Khosla, for their constructive suggestions. Deepest gratitude is also due to Dr. Gayatri Ramakrishna, Staff Scientist and Warden CDFD Hostel, for being extremely kind and supportive throughout my stay in CDFD. She has been a special friend to me and had extended her warm encouragement and support throughout my PhD tenure.

I would like to thank my respected seniors Bara saheb (sreenu) and Chota saheb (sridhar) for all their help. I had learned a lot from them and they were always there whenever I was in need. Thank you very much. My special thanks to two of my best friends Anwar and Gyan who had extended their help by all means. I also thank Suresh babu and PSN Suresh for their nice company.

I also thank my senior N. Mrinal for making my stay memorable in CDFD nacharam hostel. I was very fortunate to have a roommate like him. Thanks a lot for all the valuable discussions we had.

My special thanks to all my batchmates (JRF 2003) Arun, Kitty (Bibhusita), Geetha, Grakul (Gokul), Chubbu (Mrk), Mama (Madhav), Begum (Nasreena), Rahul (Noor), Santoo (Santosh), Raku (Rakesh) Sman (Smanla) and off course Uma. I really enjoyed having them as my batchmate and friends. I can never forget the time spent together by us during our coursework, gatherings and stage performances from time to time.

I am very thankful to Tabrez for his support. He had always extended his support in all means. "A friend in need is a friend indeed" saying, holds good for him. He is the person to whom I make the first call for any help and he is always there for me.

I would like to thank my wonderful and bright juniors Vishal, Anupam, Rachita and Manjari for creating a healthy atmosphere in the lab and extending necessary help. I will miss you all. Vishal I wish we could have played shuttle once more before I leave but… Wish you all the best. Thanks a lot for all the personal help.

I would like to thank present members of the Bioinformatics team - Prashanthi, Pavan, Chandra Mohan, Rajeswari, Vinod and Sarvesh, as well as to former members – Kavita, Geetha, Ranjit, Phani and Azhar Imam, for their kind help and support.

I would like to thank my seniors Dharma, Sailu, Ramna, Sharmishta, Rohini, Prachi, Satish, Subboo, VLN reddy, Akif, Sudish, Senthil, Kshama, Yasin, Rajendra, Prabhat, Anil, Arvind, Anoop, Siba, and Aisha for helping and being very kind to me.

My interaction with students from other labs and hostel mates has been also very helpful and rewarding. I am highly thankful to Adeel, Aisha, Akhilesh, Amit, Amitava, Anujeet, Aravind, Atul, Babul, Bhavik, Charitha, Debashree, Deepa, Devi, Dhananjay, Garima, Gaurav, Ghazala, Gita, Jamshaid, Jisha, Jyoti, Kaiser, Khalid, Khursheed, Nandan, Nazia, Nora, Pramod, Ratheesh, Ritu, Sandeep, Sapan, Saurabh, Shanti, Shashi, Shiny, Shivalika, Shubhada, Sreejeet, Swarna, Tej, Thoh, Vijay and Yusuf for being extremely nice and kind to me.

During my PhD tenure I had the kind pleasure to work with few bright and lovely dissertation trainees - Anirban, Chaitanya, Deeksha, Kirti, Purna,, Ramya, Shankar, Shyam, Venkat and many more. The interactions with them and numerous other trainees were quite refreshing and always reminds me of my pre PhD days.

I convey my special acknowledgment to all the staff of CDFD for their indispensable help and support and for making my stay in CDFD smooth and enjoyable. Special thanks are due to the staff of the Academic section, Mr. Siddiqui, Ms. Punitha, Mr. Santosh and Mr. Sankar, for taking care of all the administrative formalities. I am also thankful to hostel care-takers - Khursheed bhai, Abdul Rasheed, M. Rasheed, Masood and Baljeet; old canteen staff – Mantri, Zafar, Ravi, Sreenu, Srisailam, Kaiser, and Hameed; electricians – Reddy and Javed; drivers – Sajid, Yusuf, Asghar, Suresh, Shyam, Saleem and Venkatesh; and all other staff for making my stay in CDFD comfortable and enjoyable.

I would like to thanks some of my special friend, Rohini, Archana, Manju and swarna.

Deepest sense of gratitude is also due to Mr Sanjeev Rao, Mr. Zaman Khan, Mr. Anand Rao, Mr. Arun, Mr. Punnaiah, Mr. Purnachandra Rao, Mr. Murthy, Mr. Kasipathi, Mr. Pasha, Mr. PP Rao, Mr. Aziz, Mr. SR Rao and other security staffs, for their never say no attitude and all the help throughout my stay in CDFD.

I am happy to have developed friendly and courteous association with Krishnaji, Baljeet, Moiz, Amjad, Balaraj, Nagaraj, Kishore, Hussain bhai, Waris bhai and Mishraji, who have come to my rescue during several difficult situations. I am humbled by their gesture of love and affection.

My deepest gratitude goes to my family especially my mother, uncle, and brothers for their unflagging love and support throughout my life. My special love and thanks to khushboo.

Last but not the least I would like to thank my mother and father in law, Akka, Sam bhai, bhaiya, bhabhi and naughty monu and lovely chotu. Thanking would not be sufficient for Uma and my sweet Jigisha who brought a smile in my face in the most stressful situations. I simply say this thesis would have not been possible without your support.

Finally, I owe thanks to everybody who have been important for whom i might not have mentioned, Thank you very much.

I am grateful to the Manipal University, Manipal, for allowing me to register for the PhD degree. The financial assistance from CSIR is also gratefully acknowledged.

Pankaj Kumar

Contents

Declaration **i**

Certificate **ii**

Dedication **iii**

Acknowledgements **iv**

List of Figures **x**

List of Tables **xii**

List of Abbreviations **xiv**

Preface **xvi**

List of Publications **xxiii**

1. Microsatellites (Simple Sequence Repeats): An Introductory Review 1
1.1. Introduction 2
1.2. Microsatellites (Simple sequence repeats) discovery and nomenclature 3
1.3. Classification of simple sequence repeats 5
1.4. Origin of simple sequence repeats 5
1.5. Extraction of SSRs 6
1.5.1. Experimental methods 6
1.5.2. Computational methods 7
1.5.3. Influence of parameters 12
1.6. Distribution of SSRs 12
1.6.1. Distribution of SSR in eukaryotes 13
1.6.2. Distribution of SSR in prokaryotes 16
1.7. Origin of SSR length polymorphism 18
1.7.1. SSR polymorphism 20
1.7.2. SSR and Mismatch repair system 21
1.7.3. Directionality in SSR evolution 23
1.7.3.1. Directionality in SSR evolution in eukaryotes 23
1.7.3.2. Directionality in SSR evolution in prokaryotes 25
1.7.4. SSR mutation models 25
1.7.5. Polymorphic SSRs and their effects 28
1.7.6. SSR polymorphism in Eukaryotes 28
1.7.6.1. Beneficial effect of SSRs polymorphism in eukaryotes 30
1.7.6.2. Harmful effect of SSR polymorphism in eukaryotes 31

1.7.7. SSR polymorphism in Prokaryotes 37
1.7.7.1. SSR polymorphism in coding regions 39
1.7.7.2. SSR polymorphism in non-coding regions 44
1.8. The present work 47
1.9. Summary 50
2. A Study on Distribution and Enrichment of Simple Sequence Repeats in Yersinia Genomes 52
2.1. Introduction 53
2.2. Methods 54
2.2.1. Identification of perfect SSRs from the whole genome sequences of Yersinia 54
2.2.2. Enrichment and its significance 56
2.3. Results 56
2.3.1. Genome-wide distribution and abundance of SSRs 56
2.3.2. Enrichment of SSRs 58
2.3.3. Propensities of SSRs in different protein functions 64
2.3.4. Intragenic position-specific distribution of SSR 66
2.4. Discussion 68
2.5. Summary 70
3. PSSRFinder (Polymorphic Simple Sequence repeat Finder): A tool for comparing Equivalent Simple sequence Repeats across Multiple Related Genomes 71
3.1. Introduction 72
3.2. Polymorphic Simple sequence repeats Finder (PSSRFinder) 73
3.3. The output of PSSRFinder 74
3.4. Summary 75
4. Investigations on mutational status of SSRs in *Yersinia pestis* and its ancestor 79
4.1. Introduction 80
4.2. Methods 81
4.3. Results 81
4.3.1. Polymorphic SSRs 81
4.3.2. Distribution of PSSRs in various COG groups 82
4.3.3. Position-specific distribution of PSSRs in the intra-genic regions 85
4.3.4. SSRs showing length variation between *Yersinia pestis* and *Yersinia pseudotuberculosis* 89
4.3.5. SSRs showing length variation among the different strains of *Yersinia pestis* 92
4.4. Discussion 94
4.5. Summary 96
5. Mutational Dynamics of Simple Sequence Repeats in the Intragenic Regions of Prokaryotic Genomes 97
5.1. Introduction 98

5.2. Methods 99
5.2.1. Extraction of SSRs and comparison of equivalent SSRs 99
5.2.2. Prediction of origin and termination site of replication and calculation of SSR and PSSRs in head on and co-oriented directions 99
5.3. Results 100
5.3.1. Polymorphic SSRs in various prokaryotic genomes 100
5.3.2. PSSRs in co- and anti-oriented direction of replication and transcription 107
5.3.3. Intra-genic distribution of PSSRs 109
5.3.4. PSSRs in host adapted pathogens 112
5.4. Discussion 113
5.5. Summary 114

6. A Study on Mutational Dynamics of Simple Sequence Repeats in the intergenic regions of Prokaryotic genomes 116
6.1. Introduction 117
6.2. Methods 117
6.2.1. Systematic search for MMR system in the 43 species of prokaryotes 117
6.2.2. Extraction of non-coding PSSRs 118
6.2.3. Identification of contractions/expansions in PSSRs in a species 119
6.3. Results and Discussions 120
6.3.1. MMRS in bacterial and archaeal genomes 120
6.3.2. Non-coding PSSRs 124
6.3.3. MMR deficiency leads to destabilization of SSR tracts 125
6.3.4. Mismatch repair is biased 129
6.3.5. Primary SSR mutations are biased towards expansion 131
6.3.6. PSSRs are contraction biased in MMRS proficient species 133
6.3.7. PSSR contraction bias is independent of repeat count 135
6.3.8. Directionality of SSR mutation is independent of sequence composition of genome 136
6.4. Summary 137
7. Conclusions 147
8. References 158

List of Figures

1.1 An Illustration of a slippage event that happens at a SSR locus during replication of DNA. 19

1.2 An Illustration of the four sites, relative to a gene, at which mutations in SSRs affect that gene. 29

1.3 A schematic representation of the various effects caused by INDELs of repeat units in SSRs in the coding regions (shown as arrows) of the three mycobacterial genomes, *M. bovis, M. tuberculosis* (CDC1551) and *M. tuberculosis* (H37Rv) 29

2.1 SSR density profile in Yersinia genomes. 59

2.2 Average propensities for mono, di, tri and tetra SSRs in the ORFs coding for proteins belonging to different COG functional groups. 65

2.3 The number of SSRs found in the 5', middle and 3' parts of genes. 67

3.1 Schematic representation of PSSRFinder. 76

3.2 Example showing the output of coding PSSRs by PSSRFinder in 22 strains of *E. coli* genomes. 77

3.3 Example showing the output of non-coding PSSRs by PSSRFinder in 22 strains of *E. coli* genomes. 78

4.1 Ratio of percentages of PSSRs and SSRs found in ORFs encoding proteins belonging to different COG functional groups. 86

4.2 Distribution of PSSRs in the 5' (a), middle (b) and 3' (c) parts of genes. 87

5.1 Tract densities of PSSRs in the coding (blue bar) and the non-coding (magenta bar) regions of various prokaryotic species. 102

5.2 Ratio of % of PSSRs to % of SSRs in the coding (blue) and the non-coding (magenta) regions of various prokaryotic species. 103

5.3 Number of FS-PSSRs (blue) and IF-PSSRs (magenta) found in various prokaryotic species. 104

5.4 Ratio of % of PSSRs to % of SSRs in the co- (blue) and the head-on (magenta) directions of replication and transcription in various prokaryotic species. 108

5.5 Deviations of observed percentage of FS-PSSRs from the expected percentages of PSSRs found at 5’ (blue), middle (magenta) and 3’ (yellow) parts of genes. 111

5.6 Tract density of coding PSSRs in various prokaryotic species. 112

6.1 Phylogenetic relationship among all the species considered in this work. 123

6.2 Number of PSSRs/Mb/strain found in the non-coding regions of different prokaryotic species. 124

6.3 The number of PSSRs (Obs) found in each strain of (A) *Acinetobacter baumannii* and (B) *Haemophilus influenzae* (C) *Staphylococcus aureus*. 126

6.4 Number of PSSR per Mb per strain in various species of prokaryotes. Magenta = MMR deficient species. Blue = MMR proficient species. 128

6.5 Tract densities (the number of SSRs per Mb per strain) of mononucleotide SSRs longer than 8bp in different prokaryotic species. 132

6.6 An Illustration of contraction bias of SSR mutations is not related to repeat count. 135

List of Tables

1.1 List of SSR extraction tools and their important features. 09

1.2 Simple sequence repeats polymorphism in human. 34

1.3 SSR polymorphism in bacteria leading to frame-shift mutations in ORFs. 42

1.4 SSR polymorphism in the non-coding regulatory regions leading to phase and antigenic variation in bacteria. 46

2.1 Yersinia genomes considered for simple sequence repeat analysis. 55

2.2 Number of different types of SSRs found in Yersinia genomes. 58

2.3 Observed and expected numbers of SSRs found in Yersinia genomes. 60

2.4a Observed and expected numbers of SSRs of different copy numbers found in plague causing (YPO), non plague causing (YP) and ancestral (YPTB) Yersinia genomes. 62

2.4b Observed and expected numbers of SSRs of different repeat copy numbers found in coding and non-coding regions of *Yersinia pestis CO92* genome. 63

4.1 Distribution of PSSRs in coding and non-coding regions of Yersinia genomes. 83

4.2 Number of PSSRs of various repeat types (mono to hexa) of various repeat copy numbers. 83

4.3 The number of PSSRs showing INDELs of more than one repeat unit. 83

4.4 PSSRs had shown INDELs of more than one repeat unit. 84

5.1 The number of coding PSSRs, repeat unit size-wise, found in various prokaryotic species. 105

6.1 Species where *MutS* and *MutL* genes are present ('+') and absent. 120

6.2 PSSRs found in MMR proficient and deficient strains of *Acinetobactor baumannii, Haemophilus influenza* and *Staphylococcus aureans.* 130

6.3 Number of PSSRs found in MMR deficient species. 132

6.4 The number of PSSRs found in MMR proficient species. Cont = contraction, Exp = expansion. 134

6.5 Comparison between MMR proficient and deficient species with similar GC percentage. 136

List of Abbreviations

A	Adenine
ATP	Adenosine-triphosphate
BLAST	Basic local alignment search tool
C	Cytosine
COG	Cluster of orthologous genes
Cont	Contraction
DF	Degree of freedom
DM	Dystrophia myotonica
DNA	Deoxyribonucleic acid
DRPLA	Dentatorubropallidoluysian atrophy
EGFR	Epidermal growth factor receptor
Exp	Expansion
FRDA	Friedreichs ataxia
FS-PSSR	Frame-shift polymorphic simple sequence repeats
G	Guanine
HD	Huntingtons disease
IF-PSSR	In-frame polymorphic simple sequence repeats
IMEx	Imperfect microsatellite extractor
INDELS	Insertions/deletions
LOS	Lipooligosaccharide
LPS	Lipopolysaccarides
MDR	Multi-drug resistant
MMR	Mismatch repair

MMRS	Mismatch repair system
MSI	Microsatellite instability
NCBI	National center for biotechnology information
NS	Not significant
ORF	Open reading frame
PCR	Polymerase chain reaction
PSSRFinder	Polymorphic Simple Sequence Repeat Finder
SBMA	Spinobulbar muscular atrophy
SCA	Spinocerebellar ataxia
SD	Standard deviation
SMM	Stepwise mutation model
SNP	Single nucleotide polymorphism
SSM	Slipped strand mispairing
SSRs	Simple sequence repeats/Microsatellites
STR	Short tandem repeat
SSRF	Simple sequence repeat finder
T	Thymine
TH	Tyrosine hydroxylase
UTRs	Untranslated regions
VNTR	Variable number of tandem repeat
YP	Yersinia pestis
YPTB	Yersinia pseudotuberculosis
ZNF9	Zinc finger protein 9
aa	Amino acid
bp	Base pair

Preface

All organisms face the challenge of maintaining their fitness in diverse and changing environments. The key to fitness resides in their capacity to change their biological characteristics. Variation within a population is one of the ways of nature creating biodiversity to ensure survival of the fittest. Phenotypic variation can be achieved by regulating gene expression or through genetic change. In pathogenic bacteria, a host environment imposes stringent tests for their adaptive potential. Two features of their lifestyle give pathogens an advantage in the production of genetic variation that can contribute positively to the host-microbe "arms race": (a) potentially large size populations and (b) rapid rates of replication. The low generation time gives the ability to quickly generate large numbers of progeny. Generating large number of progeny increase the probability of at least one cell in any population will have a mutation that will allow it to adapt to any new environment.

In many bacterial species, this mutational flexibility results from contingency loci in the genome. Contingency loci are hyper-mutable and provide a repertoire of variation thereby allowing populations to adapt rapidly in the face of unpredictable contingencies, such as changes in host environments. These loci generate large amounts of genetic variation that have a disproportionately large impact on microbial fitness, and this process operates in a stochastic manner. Although this hyper-mutability can result from several mechanisms, there is growing interest in those loci containing simple sequence repeats (SSRs).

SSRs are nucleotide sequences of repeating motifs of length 1-6bp long. They are ubiquitously present in all the genomes ranging from viruses and prokaryotes to eukaryotes. One of the unique properties of these repetitive elements is their length polymorphism. They undergo mutations in the form of insertions and deletions (INDELs) of their repeat units, which are in the range of 10^{-6} to 10^{-2} per generation- several fold higher than random mutations (10^{-8}). The basic mechanism behind INDELs of SSRs is thought to be strand slippage during DNA replication. A loop generated on the nascent strand leads to insertion of repeat units where as a loop generated on the template strand manifests in deletion of repeat units. These primary replication errors are usually corrected in the cell by the post replicative mismatch repair system which consists of genes *mutS*, *mutL* and *mutH*.

The repeat unit INDEL mutational property of SSRs is a key source of genetic variation in bacteria. Variation in the length of SSRs tracts often causes changes in the structure and function of nearby genes leading to certain phenotypic changes, such as changes in motility and colony morphology. Furthermore, these mutations are reversible and can result in multiple possible “ON” and “OFF” states for genes.

The work presented in this thesis focuses on these interesting SSRs and essentially forms a part of an on-going project on systematic studies on SSRs in prokaryotic genomes using computational approaches. Our group have been studying the distribution, abundance, enrichment and polymorphism of SSRs in order to understand their roles in the adaptability, pathogenicity and evolution of pathogens of special interest. In the present thesis work we have analyzed known Yersinia genomes. In addition we have also addressed one of the important aspects of SSRs in bacterial genomes viz., their mutational dynamics in coding and non-coding regions and have carried out studies to answer questions such as: Are SSRs in bacteria prone for expansions or contractions? If so, what inherent features of genomes responsible for such biases? Why do bacterial genomes rarely harbor long tracts of SSRs similar to eukaryotic genomes? What is the role of mis-match repair system in the expansion/contraction bias of SSRs?" The entire work has been organized and presented as five chapters where Chapters 2 and 4 focus on SSRs in Yersinia genomes and Chapters 5 and 6 focus on mutational dynamics of SSRs in bacterial genomes. The third chapter is devoted to the development of novel algorithm. These five chapters precede and succeed, respectively, by an introductory review and a conclusion chapter. The contents of each chapter are given below.

Chapter 1 forms the introductory review of all the relevant literature pertaining to the work presented in the thesis. It begins with details about the discovery, nomenclature and origin of SSRs. This is followed by an overview of different methods and tools available for the extraction of SSRs from DNA sequences. We also give an overview of the distribution of SSRs in different eukaryotic and prokaryotic genomes. The chapter also gives a concise summary of the origin of SSR length polymorphism, mechanisms involved in SSR length polymorphism, the effect of mismatch repair systems (MMRS) on SSR length variation, directionality in SSR mutations and different mutation models of SSRs. The review of the literature about SSR ends with known experimental evidence of different effects of polymorphic SSRs in eukaryotic and prokaryotic genomes. Finally the chapter introduces the objectives of the current work.

Although a number of complete genomes of prokaryotic species have been studied for the distribution of SSRs, detailed systemic studies of all the known whole genomes of *Yersinia pestis* have not been reported. *Yersinia pestis* is a well known causative agent of plague and is presently dreaded for its potential use in bio-terrorism. It is believed that this pathogen has evolved from *Yersinia pseudotuberculosis,* which mainly causes gastritis. Availability of high quality complete genome sequences of a number of stains of *Yersinia pestis* as well as of *Yersinia pseudotuberculosis* prompted me to analyze them for distribution, abundance and enrichment of SSRs. The details of this analysis are given in chapter 2. All the strains of *Yersinia pestis* and *Yersinia pseudotuberculosis* are nearly 98% identical at nucleotide level and show very similar distribution, abundance and enrichment profiles of SSRs. As is typical to bacterial genomes, long tracts of SSRs are hardly seen. In general SSRs are not enriched in coding regions of Yersinia genomes. However ORFs encoding proteins believed to interact with hosts show high propensities for SSRs whereas the ORFs which encode proteins involved in general metabolism show low propensities for SSRs. When SSRs were analyzed with regard to their positional preferences in ORFs, it was found that they mostly prefer 5' and 3' ends as compared to the 'middle' region.

After studying the distribution and enrichment of SSRs in Yersinia genomes our interest turned towards polymorphic SSRs (PSSRs) in these genomes. To carry out the cross genome comparison a novel toll was developed. Chapter 3 gives the details of the in-house program PSSRFinder developed for discovery of PSSRs within a pool of related genomes. Further the cross genome comparisons of all equivalent SSRs was carried out and discovered 805 PSSRs of

which 50% were in coding regions. When we looked into the coding regions harboring PSSRs, in detail, we found interesting consequences of SSR INDEL mutations such as frame-shift mutations leading to premature terminations, length variations and splitting of ORFs. Further in-depth studies revealed that such effects of PSSRs underpin some of the observed intra-species variations in pathogenesis and adaptations. In essence, these studies suggest a role for SSR mutations in the evolution of Yersinia.

Further a global analysis of SSR length polymorphism in 43 species of prokaryotes was carried. The equivalent SSRs across different strains within every species were compared and their mutational dynamics were studied. Most of the PSSRs that were discovered have undergone INDEL mutations of single repeat units. Intragenic regions were far less dense within PSSRs as compared to intergenic regions, indicating restrained SSR polymorphism in coding regions as compared to non-coding regions. Within the intragenic regions, examples of SSR polymorphism were more frequently observed in the regions where replication and transcription directions are opposite to each other as compared to the regions where replication and transcription happen in the same direction. We also found that PSSRs show position specific preferences along the genes. Furthermore, our studies also revealed that host adapted pathogens (animal pathogen, mostly human), in general, have a higher density of PSSRs, compared to free-living and plant pathogens, in intragenic regions of the genomes. All the above results are presented in Chapter 5.

Mutational bias towards expansion or contraction of SSRs, also referred to as the directionality of SSR evolution, in relation to the presence or absence of mismatch repair (MMR) system has been studied in some eukaryotes as well as prokaryotes. However, the results have not been in unison to state clearly the relationship between the mutational bias of SSRs and the presence/absence of MMR systems. Therefore known prokaryotic genomes were investigated for intra-species mutational patterns of SSRs in non-coding regions (Chapter 6). These studies were carried out on three groups of species: (1) species having at least one strain lacking the MMR system; (2) species where the MMR system is absent in all the known sequenced strains and (3) species where the MMR system is present in all the known sequenced strains. It was found that for a species, the MMR system deficient strains harbored significantly more polymorphic simple sequence repeats (PSSRs) as compared to MMR system proficient strains. MMR system-deficient species, in general, showed a bias towards SSR expansions, indicating a preference for insertions over deletions amongst the primary mutations. However, those MMR system-proficient species, where the observed mutations are secondary mutations (mutations escaping repair by the MMR system) showed a bias towards the contraction of SSRs, perhaps indicating the low efficiency of MMR systems to repair SSR-induced slippage on template strands. We reasoned that this bias towards deletion in the SSR tracts is the root cause of scarcity for long SSR tracts in prokaryotic systems which are mostly MMR proficient. In conclusion, our study clearly demonstrates the mutational dynamics of SSRs in relation to presence/absence of MMR system in the prokaryotic system.

The thesis ends with Chapter 7 which essentially recapitulates all the major aspects of the thesis work including the goals of the research work undertaken, the investigations carried out, a summary of the results obtained and conclusions reached.

Most of the work presented in this thesis has either been published or written in the form of manuscripts for communication.

Publications:

1. Sreenu VB*, **Kumar P***, Nagaraju J, Nagarajam HA. Microsatellite polymorphism across the M. tuberculosis and M. bovis genomes: implications on genome evolution and plasticity. BMC Genomics. 2006,10;7:78. ***equal contribution**
2. Sreenu VB, **Kumar P**, Nagaraju J, Nagarajam HA. Simple sequence repeats in mycobacterial genomes. J Biosci. 2007, 32:3-15.

   *Candidate has done the randomization (thousand times) of genome considered in this study, followed by extraction of SSRs from randomized genomes and comparison of the average number of SSRs extracted from randomized genome with the observed number of SSRs in each strain. The statistical significance of abundance was also performed.
3. Maikho Thoh, **Kumar P,** Nagarajam HA and Sunil K Manna. Azadirachtin interacts with TNF receptors and inhibits TNF-induced biological responses J. Biol. Chem. (in Press; published December 14, 2009 as doi:10.1074/jbc.M109.065847)

   *Candidate has done the Molecular docking studies on human TNF receptor protein with Azadirachtin (an extract of neem plant). The docking study showed the possible inhibitory role of Azadirachtin on TNF receptor binding with its ligand.
4. **Kumar P** and Nagarajam HA. Survey and analysis of Simple Sequence Repeats in *Yersinia pestis* genomes (*Manuscript under preparation*)
5. **Kumar P** and Nagarajam HA. Mutational Dynamics of Simple Sequence Repeats in the Intragenic Regions of Prokaryotic Genomes (*Manuscript under preparation*)
6. **Kumar P** and Nagarajam HA. A Study on Mutational Dynamics of Simple Sequence Repeats in the intergenic regions of Prokaryotic genomes (*Manuscript under preparation*)

Posters presented in Symposia:

1. **Kumar P** and Nagarajam HA "Survey and analysis of simple sequence repeat in Yersinia genome" 13th Human Genome Meeting, HUGO" held in Hyderabad, India from 27th-30th September 2008.
2. **Kumar P** and Nagarajam HA Simple sequence repeat distribution, abundance, enrichment and polymorphism in Yersinia genome "16th Annual International Conference on Intelligent Systems for Molecular Biology", ISMB-2008 held in Toronto, Canada, 19th-23rd July 2008
3. Simple sequence repeat in Yersinia genome "International Conference in Bio-informatics, INCOB-2006", organized by the Indian Institute of Technology, New Delhi.
4. Simple sequence repeat distribution, abundance, enrichment and polymorphism in Mycobacterium genome "International Symposium on New Frontiers in Tuberculosis Research" from 4-6th Dec 2006, organized by International Centre for Genetic Engineering and Biotechnology, New Delhi, INDIA

Chapter 1

Microsatellites (Simple Sequence Repeats): An Introductory Review

1.1 Introduction

Nucleotides (adenine (A), guanine (G), thymine (T), and cytosine (C)) are the building blocks of any genome. These nucleotides are linearly arranged on a chromosome. However, nucleotides are not scattered randomly in the genome. The genomic global and local compositional heterogeneity is widely recognized. The regions of the chromosomes whjch code for proteins or functional RNAs are known as coding regions, whereas, those regions which do not code for proteins or functional RNAs are known as non-coding regions. Other than these there are many facets of DNA heterogeneity which include isochore compartments in vertebrate species (Bernardi, Mouchiroud, and Gautier 1988), the G+C and A+T-rich halves of the bacteriophage lamda genome (Inman 1966), characteristics telomeric sequences (hexa nucleotide AGGGTT), GNN periodicity in coding sequences (Fickett 1982) and methyl transferase modifications (Nelson and McClelland 1991). Studies including pattern matching (Galas, Eggert, and Waterman 1985), word frequency counting (Karlin and Burge 1995) and basic linguistic studies (Pevzner, Borodovsky, and Mironov 1989a; Pevzner, Borodovsky, and Mironov 1989b) have further reinforced the compositional heterogeneity of genomes. Genomes are also characterized by overrepresentation of some motifs such as restriction enzyme recognition sites, regulatory motifs, repetitive sequences etc and underrepresentation of some motifs such as TA dinucleotide motif (Burge, Campbell, and Karlin 1992). Assuming chance association of nucleotides, the probability of finding a GAATTC motif of repeat count six such as $(GAATTC)_6$ more than once in the human genome is negligible. However, perfect or near-perfect tandem iterations of short sequence motifs of this kind are extremely common in eukaryotic genomes and, in the case of the human genome; they are found at hundreds of thousands of places along the chromosomes (Lander et al. 2001; Ellegren 2004). Prokaryotic genomes too harbor such tandem repeats of sequences, however as short tracts (Mrazek, Guo, and Shah 2007). Repetitive sequences have almost all the features required for tracking regions of the genomes and hence offer as simple molecular ways in the current genomics and high throughput sequencing era. Repetitive sequences in addition to being abundant are co-dominant and distributed over the euchromatic part of the genome. They are highly polymorphic (Schlotterer 1998) and hence have significantly aided genome mapping efforts. In this chapter an overview has been presented on the distribution, abundance and polymorphism of one of the special classes of repetitive sequences called simple sequence repeats (SSRs) which are also referred to as microsatellites.

1.2 Microsatellites (Simple sequence repeats) discovery and nomenclature

Repetitive DNA which is often referred to as satellite DNA has a characteristic physical feature due to its specific nucleotide composition. The term satellite might appear unusual for describing a nucleotide repeat tract, but it carries a little history behind it. When subjected to density gradient centrifugation, repetitive DNA lagged behind the bulk of DNA and was presented as satellite fractions due to differences in thermodynamic stability and reassociation kinetics (Britten and Kohne 1968). Later, the satellite DNA was shown to be constituted of repetitive sequences. Since then, researchers started using the term satellite DNA rather than finding a new term to represent repetitive sequences. A good general rule of thumb is that the bigger the repeat unit is, the longer the arrays tend to be. Human genome contains major satellite DNAs which are long tracts, up to several Mb in length, of well known families of repeat sequence elements (LINES, SINES etc.). Apart from this, direct tandem repeat sequences of motifs of 10-30bp size classified as minisatellites are also present (Jeffreys, Wilson, and Thein 1985). Further tandem repeats with shorter repeat motifs referred to as microsatellites have been reported (Litt and Luty 1989; Weber and May 1989). Microsatellites include simple sequence repeat elements in tracts less, or much less than 1 kb. However, there have been some differences in the exact definition of microsatellites or SSRs. For example, Armour et al. define microsatellites as the tracts of repeats of 2–8 bp (Armour 1999) whereas others, e.g. Goldstein and Pollock (Goldstein and Pollock 1997) and our group define repeats of 1–6 bp (Goldstein and Pollock 1997; Sreenu et al. 2003a) as microsatellites.

Initial findings of different forms of repetitive sequences in many genomes created an uncertainty in defining the term microsatellite. Although attempts have been made to standardize the nomenclature of microsatellites, there seems to be no consensus among the researchers (Tautz 1993; Chambers and MacAvoy 2000). Nomenclature for use with these systems was first presented by their discoverers (Litt and Luty 1989; Weber and May 1989; Tautz 1993) and has been updated periodically since then (Tautz 1993; Gur-Arie et al. 2000; Schlotterer 2000; Buschiazzo and Gemmell 2006; Kashi and King 2006; Moxon, Bayliss, and Hood 2006). However, some ambiguities do still remain (e.g. the apparent size gap between mini- and microsatellites). Keeping in view all these I adhere to the definition given by Tautz (Tautz 1993) according to which a microsatellite is the one characterized by a repeating motif of size within the range 1-6bp (arranged in a head-to-tail manner) iterating at least two times at a given locus. Since the last decade researchers have also started using simple sequence repeats (SSRs) as a synonym for microsatellites. Though there is a good body of evidence showing that the microsatellite evolution is not as simple as the synonymous name suggests, researchers still prefer to use the term simple sequence repeats in place of microsatellites because they make simple structure. While the other sequences generally produce a scrambled pattern on a sequencing gel, microsatellites can be quickly identified by their simple structure (pattern of nucleotide) (Schlotterer 1998). Through out this thesis the term simple sequence repeats (SSRs) has been used for repeats of 1-6 bp in place of microsatellites.

1.3 Classification of simple sequence repeats

According to the size of repeat units, SSRs are classified into mono-, di-, tri-, tetra-, penta-, and hexa nucleotide repeats. On the basis of repetitive architecture, SSRs are classified as perfect, imperfect or compound. In a perfect SSR the repeat motifs are arranged in head to tail manner without any interruption (*e.g.* TATATATATATATATA) while in an imperfect SSR there could be insertion, deletion or substitution of bases in the repeated motifs (*e.g.* TATATATA<u>C</u>TATATA). In the case of a compound SSR the sequence contains two adjacent distinct SSRs separated by none to any number of base pairs (*e.g.* TATATATATAGTGTGTGTGT). I woud like to mention here that the work presented in this thesis focuses entirely on perfect SSRs.

1.4 Origin of simple sequence repeats

It is widely assumed that short “proto”-microsatellites or SSRs are generated by chance (Levinson and Gutman 1987b). These “proto”-SSRs further expand into long SSRs by a mechanism called slipped strand mispairing during replication (details are given in section

1.7). It has been suggested that a minimum number of repeats is required before DNA slippage can extend the proto-SSR (threshold model) (Rose and Falush 1998). Another report on yeast, however, showed that DNA slippage does not have such a requirement (Pupko and Graur 1999). Support for the latter result comes from cross species comparisons, which showed that SSRs as short as two repeats also change in length (Schlotterer 2000; Sreenu et al. 2006). Hence, SSRs can be generated by random point mutations or certain combinations of codons (Kruglyak et al. 1998) that create short proto-SSRs which are expanded further by strand slippage events during the replication process.

1.5 Extraction of SSRs

1.5.1 Experimental methods

Simple sequence repeats (SSRs), have been the most widely applied class of molecular markers in genetic studies, with applications in many fields of genetics including genetic conservation, population genetics, molecular breeding, and paternity testing. This range of applications is due to the fact that SSR markers are co-dominant and multi-allelic, are highly reproducible, have high-resolution and are based on the polymerase chain reaction (PCR). The great popularity of SSRs is also demonstrated by the growing number of reports describing the isolation of these markers in many organisms. For most of the classically trained biologists, establishment of SSR analysis systems from scratch for a new species presents a considerable technical challenge. Despite this technical challenge SSRs have become very popular tools. Their popularity stands as a testimony to their wide utility, resolving power and convenience. In some instances, SSR systems originally developed in a single focal species can be adapted for use in other closely related species (Gemmell et al. 1997; Primmer and Ellegren 1998).

Traditionally, SSR loci have been isolated from partial genomic libraries (selected for small insert size) of species of interest, screening several thousands of clones through colony hybridization with repeat containing probes (Rassmann, Schlotterer, and Tautz 1991). High quality genomic DNA is fragmented using restriction enzyme. Although relatively simple for SSR rich genomes, this approach can become extremely tedious and inefficient for species with low SSR frequencies. In order to reduce the time invested in SSR isolation and significantly increase yields several alternative strategies which do not require complete genome sequence information have been devised (reviewed in (Zane, Bargelloni, and Patarnello 2002)). The major drawback of experimental approach is that SSRs have to be isolated *de novo* for most species being examined for the first time.

1.5.2 Computational methods

In the post-genomic era, availability of complete genome sequences has simplified and eased the task of screening genomes for SSRs. A computer program specially developed for SSR screening is sufficient to scan the whole genome for sequence composition of motifs, location and frequency of SSR tracts. There are many computational tools available for extraction of SSRs from genomic regions. A complete list of these tools with their important features is given in **Table 1.1**. In the following section some of the widely used tools have been decribed.

Tandem repeat finder

Tandem Repeat Finder (TRF) (Benson 1999) is the most popular tool. This can be used to find tandem repeats of sizes as large as 2000 bp. TRF uses a heuristic procedure and identifies regions where a motif is periodically repeated and aligns the sequence with its perfect counter-part using a wraparound dynamic programming algorithm. An alignment score is generated based on the weights for matches (default match Weight: 2) and mismatches (substitutions and indels) (default mismatch weights: 7, 7) and if the score crosses a certain threshold (30 for smaller repeats), the repeat tract is reported. TRF has been designed for detecting large tandem repeats and therefore might miss many potential SSRs. No specification of pattern or pattern size is required for this program. TRF models tandem repeats by percent identity and frequency of indels between adjacent pattern copies using statistically based recognition criteria. A World WideWeb (WWW) server interface http://atc3.biomath.mssm.edu/trf.html has been established for the automated use of this program. The program takes sequence input in the FASTA format. The result of the analysis is sent back to the web browser of the user as two files, a summary table file and an alignment file. The summary table contains information about each repeat, including its location, size, number of copies and nucleotide content. Clicking the location indices of the table entries, results in the opening of a second web browser that provides the alignment of the copies against a consensus pattern. The program is extremely fast, analyzing sequences in the order of 0.5Mb in just a few seconds. The length of submitted sequences can be up to a maximum of 5Mb.

Table 1.1: List of SSR extraction tools and their important features. Presence or absence of a given feature in a given tool is indicated by 'Yes' or 'No'.

Program name	Web-interface	Availability	Perfect repeat	Imperfect repeat	Coding/ Noncoding	References
ATR Hunter	Yes	http://bioinfo.cs.technion.ac.il/atrhunter/ATRinformation.htm	No	Yes	Yes	(Wexler et al. 2005)
IMEx	Yes	http://210.212.215.200/IMEX/index.html	Yes	Yes	Yes	(Mudunuri and Nagarajaram 2007)
Misa	No	-	Yes	No	No	-
Mreps	Yes	http://bioinfo.lifl.fr/mreps/	Yes	Yes	No	(Kolpakov, Bana, and Kucherov 2003)
Msatfinder	Yes	http://www.genomics.ceh.ac.uk/msatfinder/	Yes	Yes	No	(Thurston 2005)
poly	No	-	Yes	No	No	(Bizzaro and Marx 2003)
Repeat Masker	Yes	http://www.repeatmasker.org/	Yes	Yes	Yes	(Smit 1996)
SciRoKoCo	No	-	Yes	Yes	No	(Kofler, Schlotterer, and Lelley 2007)
Sputnik	Yes	http://espressosoftware.com/sputnik/index.html	Yes	Yes	No	-
SSRIT	Yes	http://gramene.agrinome.org/db/searches/ssrtool	Yes	No	No	(Temnykh et al. 2001)
STAR	Yes	http://atgc.lirmm.fr/star/	No	Yes	No	(Delgrange and Rivals 2004)
STRING	No	-	No	Yes	No	(Parisi, De Fonzo, and Aluffi-Pentini 2003)
TandemSWAN	Yes	http://favorov.imb.ac.ru/swan/	No	Yes	No	-
TRF	Yes	http://tandem.bu.edu/trf/trf.html	No	Yes	No	(Benson 1999)
TROLL	No	-	Yes	No	No	(Castelo, Martins, and Gao 2002)

Sputnik

Sputnik uses a recursive procedure to scan the sequence with adjacent copies of the same repeat and the score, based on the weights, is increased for every match (default match weight: 1) and decreased for every mismatch (substitutions and indels) (default mismatch weight: -6) and if the score crosses a certain threshold (default cut-off score: 8), the tract is reported as valid. Sputnik extracts perfect as well as imperfect SSRs. The outputs from Sputnik are amenable for use with other programs for analysis. This program is available for free download on WWW (http://espressosoftware.com/ pages/ sputnik.jsp).

Simple Sequence Repeat Finder (SSRF)

SSRF (Sreenu et al. 2003b) is a very fast method of extracting perfect repeats of 1-6 bp length from DNA sequences of any length. SSRF employs a brute-force approach for finding perfect microsatellites. This program does not require any apriori knowledge of repeat motif's sequence or repeat number and does not use pre-compiled list of motifs to check for repetitions. All the repeat sequences are generated on the fly without any limit to the repeat number. The program takes genome sequence file accession number as a command line argument and searches for the .fna file (sequence file) and the .ffn file (ORFs file) in the local directory. This program reports all the perfect SSRs of repeat motifs of 1-6 bp along with information such as repeat motifs, their co-ordinates in genome (start and end), the number of repeats and location with respect to the coding region in their neighbourhood.

SciRoKoCo

SciRoKoCo (Kofler, Schlotterer, and Lelley 2007) considers the length of the SSR tract as the main parameter. It includes 5 different search modules; three for perfect and two for imperfect SSRs. For imperfect repeat search, SciRoKoCo looks for perfect SSRs that act as a seed for imperfect SSR and are expanded on both sides by calculating the score using the match (default match weight: 1) and mismatch penalties (default penalty: 5) and reports the SSR if the score crosses a cut-off value (default cut-off score: 15). The combination of an extremely fast search algorithm with a built-in summary statistic tool makes this program a very useful tool for full genome analysis. Compared to the other already existing tools, SciRoKo also allows an analysis of compound SSRs.

IMEx

IMEx (Mudunuri and Nagarajaram 2007) is a fast and the most sensitive tool which uses a simple string search algorithm to look for two adjacent perfect repeats with or without an intervening analogous motif with some imperfections. Once such a tract is found then it is expanded on both sides in steps of the repeat motif with or without some imperfections. The level of imperfection allowed per repeat unit and also for the entire tract is monitored by user defined parameters. For every match (default weight: 1), the score is increased and for every mismatch (default weight: 1), the score is decreased. IMEx uses ‘edit distance’ rather than ‘hamming distance’ (which does not consider indels) to calculate the score. Unlike the other 3 programs that calculate the score, IMEx calculates the imperfection percentage (% of mismatches of the entire tract) and if the imperfection % crosses the threshold (p%) (Default: 10% means 1 imperfection for a tract length of 10), checks with the minimum number of repetitions and then reports it in the output. Its web version is available at http://www.cdfd.org.in/imex.

1.5.3 Influence of parameters

Over the last 10 years, a large number of studies on SSR distribution in eukaryotic genomes have been published. Unfortunately, it is not always an easy task to compare the published results, since different authors use different algorithms (sometimes developed in-house and not necessarily published), and they need not always agree on the definition of a SSR and therefore use different settings and thresholds to detect what they think should be defined as a SSR (Leclercq, Rivals, and Jarne 2007; Richard, Kerrest, and Dujon 2008; Mudunuri et al. 2010). Since the algorithm used strongly influences detection of SSRs, data pertaining to their frequencies, distribution etc. may vary from program to program. It is therefore strongly suggested to exercise caution while, studying SSRs using different tools (Leclercq, Rivals, and Jarne 2007). The variation in the number of trinucleotide repeats in *S. cerevisiae* by using different parameter and program by different group was highlighted by Richard et al. (Richard, Kerrest, and Dujon 2008). The absolute numbers of such repeats in the yeast genome vary from 92 to 1,769 depending on the parameter chosen by authors (Richard, Kerrest, and Dujon 2008).

1.6 Distribution of SSRs

SSRs are ubiquitous in prokaryotes and eukaryotes, present even in the smallest bacterial genome (Hancock 1996; Field and Wills 1998). SSRs can be found anywhere in a genome, both in protein coding and non-coding regions. Because of their high mutability, SSRs are thought to play a significant role in genome evolution by creating and maintaining genetic variation (Tautz and Schlotterer 1994; Kashi, King, and Soller 1997). However, with regard to their distribution there is a clear distinction between prokaryotic and eukaryotic genomes and this aspect has been expounded in the following sections.

1.6.1 Distribution of SSRs in eukaryotes

The frequency of SSRs in eukaryotic genomes is much higher than that could be expected by chance alone (Toth, Gaspari, and Jurka 2000; Ellegren 2004). In the human and mouse genomes, SSR density that is the number of SSR tracts in a unit genome is nearly two-fold higher near the ends of the chromosome arms than the other parts of the chromosome (Waterston et al. 2002). It has also been shown that the distribution of SSRs on the allosomes differs from that on the autosomes (Bachtrog et al. 2000).

Tri- and hexanucleotide repeats prevail in protein-coding exons of all taxa, whereas the dependence of repeat abundance on the length of the repeated unit shows a very different pattern as well as taxon-specific variation in intergenic regions and introns (Toth, Gaspari, and Jurka 2000). The frequency of genomic SSRs also varies per taxon, in terms of absolute numbers of SSR loci and preferential repeats (Toth, Gaspari, and Jurka 2000).

Among the fully sequenced eukaryotic genomes, SSR density is the highest in mammals, while bird and plant genomes show lower densities (Primmer et al. 1997; Primmer and Ellegren 1998). A comparison of available eukaryotes revealed that SSRs are most abundantly found in *Homo sapiens*, followed by *Drosophila melanogaster*, *Arabidopsis thaliana*, *Saccharomyces cerevisiae* and *Cenorhabditis elegans* (Katti, Ranjekar, and Gupta 2001). The complete sequence of the human genome harbours more than one million SSRs with the frequency of about one tract per every 2-3 kb and thus constituting about 3% of the genome (Lander et al. 2001). Among these, dinucleotide repeats are more common than mono and tetranucleotide repeats, while trinucleotide repeats are the least abundant. These observations are however, a matter of how SSRs are defined in these studies. Among repeats that are at least 12 bp long, mononucleotide repeats outnumber dinucleotide repeats; the reverse situation is not valid until a higher threshold is used (Ellegren 2004).

The mouse genome consists of two to three times more number of SSRs than the human genome (Waterston et al. 2002). In addition, SSRs are longer in mouse than in human and the same is true when the rat and human genomes are compared (Gibbs et al. 2004). *Drosophila melanogaster* has shorter SSRs compared to human (Bachtrog et al. 2000).

It has been shown that SSR density is positively correlated to genome size in animals (Hancock 1996; Toth, Gaspari, and Jurka 2000; Katti, Ranjekar, and Gupta 2001) and in plants, it is negatively correlated (Morgante, Hanafey, and Powell 2002). The most striking difference between the plant genomes and the other genomes is, the “AT” repeat is the most commonly found dinucleotide motif in the plant genomes (Morgante, Hanafey, and Powell 2002), whereas the same motif is considered as “universally under represented" in all the other genomes (Karlin, Mrazek, and Campbell 1997).

The distribution of SSRs differs substantially among exons, introns, 5’ and 3’ regions of genes (Morgante, Hanafey, and Powell 2002). In primate genes it was found that only trinucleotide repeats show repeat enrichment well above the threshold of statistical significance and cover slightly more than 0.3% of the protein coding regions of genes (Borstnik and Pumpernik 2002). Moreover the pattern of SSR over- and underrepresentation differs among repeat types, SSR length classes, and species (Dieringer and Schlotterer 2003). The long SSRs are overrepresented. Interestingly, the length at which significant overrepresentation occurs differs among repeat types and species (Dieringer and Schlotterer 2003).

SSR frequency also differs amongst plant species *i.e.* Arabidopsis, maize, soybean, wheat and rice, and is high in Arabidopsis and low in species with comparatively larger genomes such as maize and wheat (Morgante, Hanafey, and Powell 2002).

Due to their high mutation rates SSRs are expected to be present in low numbers in coding regions because a large presence of SSRs means high chance for SSR mutation associated functional loss. Comparative studies in both coding and non-coding regions of different species of eukaryotes have confirmed this. Only tri- and hexa-nucleotides are found in excess and quite contrastingly, other types of repeats are less frequently found (Toth, Gaspari, and Jurka 2000). Enrichment which is measured as the ratio of observed to expected frequencies of SSRs differs in coding and non-coding regions. Furthermore, all types of non-coding repeats are subject to similar mutational and selective processes. Coding repeats, however, appear to be under much stronger and more specific constraints mainly arising out of the selection against frame shift mutations in coding regions resulting from length changes in non triplet repeats (Metzgar, Bytof, and Wills 2000).

The SSR density in genomic regions and expressed sequence tags (ESTs) of *Arabidopsis thaliana, Oryza sativa, Glycine max, Zea mays* and *Triticum aestivum* indicates action of different selective pressures on 5’ and 3’ untranslated regions (UTRs) and open reading frames (ORFs) of transcription units (Metzgar, Bytof, and Wills 2000; Morgante, Hanafey, and Powell 2002). SSR frequency at the 3’ UTR regions was found to be higher than that expected for the whole genome, with tri- and tetra-nucleotides contributing markedly to this increase. Moreover, the 5' UTR regions showed a much higher SSR frequency than other genomic fractions, and this was due to the presence of di- and tri-nucleotides, principally AG/CT and AAG/CTT repeats.

1.6.2 Distribution of SSRs in prokaryotes

SSRs, in prokaryotic genomes are uniformly distributed (Field and Wills 1998; Gur-Arie et al. 2000; Yang et al. 2003; Coenye and Vandamme 2005). Compared to eukaryotes, microsatellite tracts are shorter in prokaryotes and they are even shorter in plasmids (Yang et al. 2003; Coenye and Vandamme 2005). In general, short repeats are more prevalent than long tracts (Hood et al. 1996; Field and Wills 1998; Gur-Arie et al. 2000; Yang et al. 2003; Coenye and Vandamme 2005). SSRs in prokaryotic genomes have been analysed in several studies. Bacterial genomes with an underrepresentation of mononucleotide SSRs are generally larger and GC-rich, while those with an overrepresentation of mononucleotide SSRs are in general smaller and AT-rich (Coenye and Vandamme 2005). Field and Wills (Field and Wills 1998) analyzed SSR occurrences in several complete genomes and found an overrepresentation of short SSRs (up to the length of 7–8 bp for mononucleotide) and underrepresentation of long SSRs. This they reasoned as due to active selection against long SSRs. An exception to this has been seen in some Mycoplasma genomes which exhibit an overrepresentation of mononucleotide SSRs of lengths 4–7 bp (Mrazek 2006). SSR representations in most prokaryotic genomes exhibit few deviations from random models. One general exception is the sharp decline in mononucleotide SSRs beyond the length of 8 bp, which is common among prokaryotes and applies to both intragenic and intergenic regions (Mrazek 2006; Mrazek, Guo, and Shah 2007; Sreenu et al. 2007). Prokaryotic genomes vary significantly in terms of the number of long SSRs they harbor. Specifically, as reported by Mrazek et al (Mrazek, Guo, and Shah 2007) 50% of the 378 prokaryotic chromosomes analyzed contain <7 long SSRs (SSR length >7bp), whereas the *Anabaena variabilis* ATCC 29413 chromosome contains 502 long SSRs (>7bp), and >700 are present in the *B. pseudomallei* genome (both chromosomes) (Mrazek, Guo, and Shah 2007).

Previously it was reported that coding and non-coding regions differ in terms of the number of SSRs they harbor (Yang et al. 2003). The tri- and hexanucleotide SSRs prevail in the coding regions while the mono- and dinucleotide SSRs are more common in the non-coding regions (Yang et al. 2003). Mononucleotide SSRs are over-represented in non-coding regions and under-represented in open reading frames (ORFs) (Coenye and Vandamme 2005). Nucleotide composition of mono- and dinucleotide SSRs, both in ORFs and in noncoding regions, differ from that of the genomic region in which they occur, with 93% of all mononucleotide SSRs proving to be of A or T type (Gur-Arie et al. 2000; Coenye and Vandamme 2005). The differential distribution of SSRs has been attributed to selection of DNA sequence for stability (Ackermann and Chao 2006). It has been shown that codons are used to encode proteins in a way that avoids the emergence of mononucleotide repeats (Ackermann and Chao 2006).

In *E. coli*, di, tetra and pentanucleotide repeats are less than expected and largely comprise of GC/CG repeats. The AT/TA repeats are over-represented in the non-coding regions, while tri and hexanucleotide repeats are over-represented in the coding regions (Gur-Arie et al. 2000). In *Shigella flexneri*, of the total number of SSRs found a higher percentage is found in the coding regions than non-coding regions (Yang et al. 2003).

Furthermore, the distribution of SSRs along a gene also varies. 5′ bias in the intragenic location of SSRs (more than 5 bp) which increases as SSR length increases was observed across prokaryotic genomes, indicating a general trend towards SSR accumulation in that part of genes (van Passel and Ochman 2007). By contrast, the severely reduced genome of *Carsonella ruddii* which is extremely A + T-rich showed neither the 5′ bias of SSRs nor the 5′ shift with increased repeat length, despite presence of a large number of SSRs in its coding sequences (van Passel and Ochman 2007).

1.7 Origin of SSR length polymorphism

Length variations among SSRs arise due to insertions and deletions (INDELs) of repeat units, a consequence of slipped strand mispairing during replication (Streisinger et al. 1966; Levinson and Gutman 1987a; Schlotterer and Tautz 1992) (please see **Figure 1.1**). The basis of this mechanism was established in the mid 1960s (Fresco and Alberts 1960; Kornberg et al. 1964), and this has been tested and confirmed by *in vitro* experiments (Schlotterer and Tautz 1992). Slippage on the template strand leads to contraction (deletion of repeat units) of SSRs whereas slippage on the nascent strand manifests as an expansion (insertion of repeat units) of SSRs (Streisinger et al. 1966; Levinson and Gutman 1987a; Harr, Todorova, and Schlotterer 2002; Mirkin 2005; Garcia-Diaz and Kunkel 2006). The INDELs of repeat units generated during replication are referred to as primary mutations and most of these are corrected by a mismatch repair system (discussed below) and only a small fraction that is not repaired ends up as SSR mutations (Strand et al. 1993).

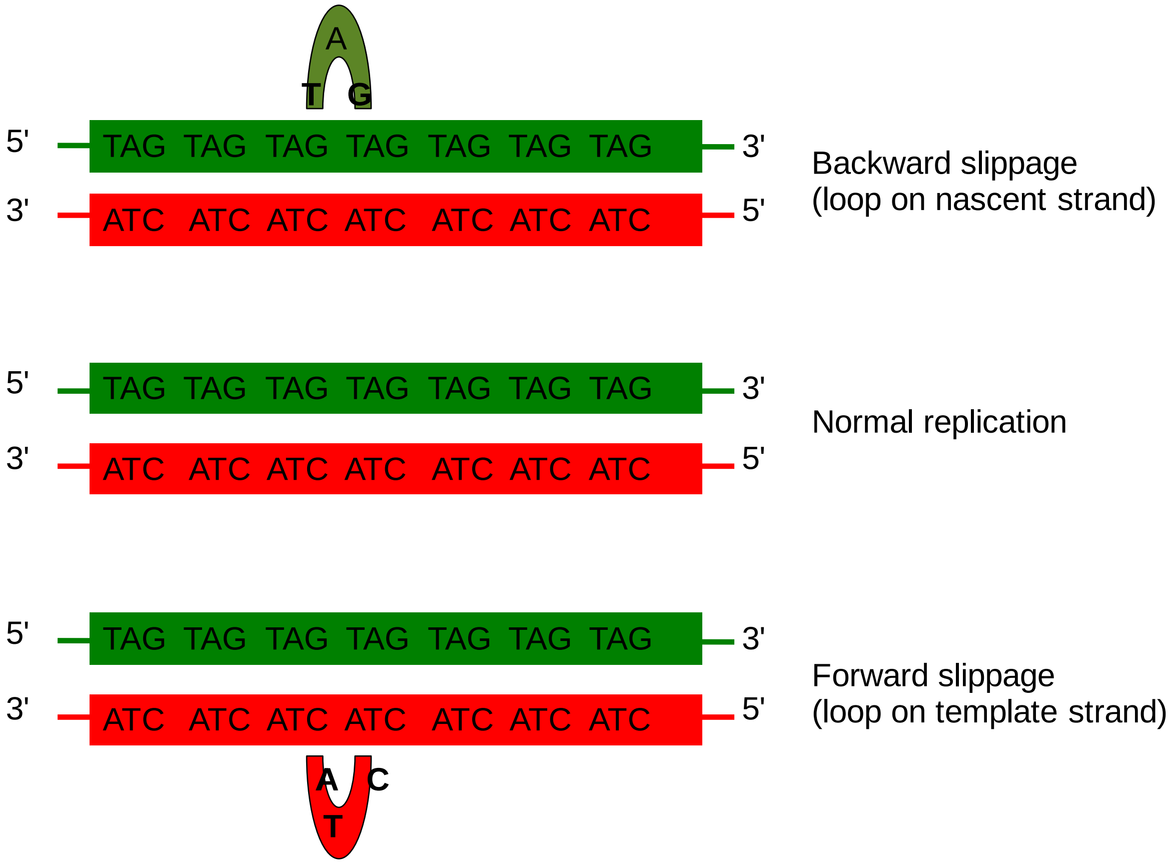

A
T G
5'
TAG TAG TAG TAG TAG TAG TAG
3'
3'
ATC ATC ATC ATC ATC ATC ATC
5'
Backward slippage
(loop on nascent strand)
5'
TAG TAG TAG TAG TAG TAG TAG
3'
3'
ATC ATC ATC ATC ATC ATC ATC
5'
Normal replication
5'
TAG TAG TAG TAG TAG TAG TAG
3'
3'
ATC ATC ATC ATC ATC ATC ATC
5'
Forward slippage
(loop on template strand)
A C
T

Figure 1.1: An Illustration of a slippage event that happens at a SSR locus during replication of DNA. Two DNA strands dissociate after the replication is initiated. While realigning either the nascent strand or the template strand may realign to the other strand out of register. These results in a loop formation either on the nascent strand called backward slippage or template strand called forward slippage. As a consequence DNA polymerase either adds extra or lesser bps to the nasent strand as compared to the template strand leading to expansion or contraction of SSR.

1.7.1 SSR polymorphism

The INDEL mutations of repeat units in SSRs occur at high frequencies ranging from 10^{-6} to 10^{-2} per generation, which is much higher than base substitution rates (van Belkum et al. 1998; Kunkel 2004; Garcia-Diaz and Kunkel 2006; Sreenu et al. 2006).

Earlier studies on SSR mutations led researchers to propose a length threshold for SSRs to exhibit polymorphism (Rose and Falush 1998). However, studies by Pupko and Gruar (Pupko and Graur 1999) showed non-existence of length limit for microsatellite tracts to show variation, except a positive correlation between the repeat number and the mutation rate.

The mutation rate at a SSR in part depends on its intrinsic features, e.g., the number of repeated units, length (in base pairs), and the repeated motif (Schlotterer and Tautz 1992; Schlotterer 2000; Kelkar et al. 2008). SSRs with a large number of repeats are highly mutable due to the increased probability of slippage (Pearson, Nichol Edamura, and Cleary 2005; Kelkar et al. 2008). A negative correlation was suggested between the length of the repeat unit and the rate of slippage (Schlotterer and Tautz 1992; Schlotterer 2000; Kelkar et al. 2008). A Study by Weber and Wong (Weber and Wong 1993) showed that the average mutation rate of tetranucleotide repeat tracts is nearly four times higher than that of dinucleotide repeats. However, in a later study, different results were obtained by Chakraborty et al. (1997) and Kelkar et al. (2008) (Chakraborty et al. 1997; Kelkar et al. 2008).

The sequences of repeat motifs of SSR tract also influence its mutability. Among mononucleotides the mutability is significantly higher for $(A)_n$ than $(C)_n$ SSRs, at least at low repeat numbers (n < 17) (Boyer et al. 2002; Kelkar et al. 2008). Among the dinucleotides, $(AT)_n$ runs exhibit the highest mutability (Chakraborty et al. 1997; Kelkar et al. 2008). For n < 12, the mutability is significantly higher for $(AC)_n$ than $(AG)_n$, while for n > 15, the pattern is reversed (Kelkar et al. 2008). Among trinucleotides the $(AAG)_n$ are highly mutatable and among the tetra nucleotide tracts the highly mutatable SSRs are $(AAAG)_n$ and $(AAGG)_n$ (Kelkar et al. 2008).

Furthermore, the mutability also differs on sex chromosomes and autosomes of human beings. For mononucleotides, mutability is significantly higher on Y than on the other chromosomes, and higher on autosomes than on X for most repeat numbers, although often non significantly (Kelkar et al. 2008). This is consistent with replication-related mechanisms and male origin of many mononucleotide SSR mutations because of the higher number of replications in the male than the female germline (Ellegren 2004). In agreement with this, several pedigree analyses also indicated higher SSR mutation rates in human males than females (Brinkmann et al. 1998; Ellegren 2000). In addition the degree of polymorphism is also found to be both species and locus specifc (Amos et al. 1996; Harr et al. 1998; Ellegren 2000).

1.7.2 SSR and Mismatch repair system

In vitro experiments have demonstrated that SSR mutations occur at very high rates (Schlotterer and Tautz 1992; Schlotterer 2000). However, *in vivo* they occur at lower rates. This discrepancy between *in vitro* and *in vivo* mutation rates can be explained as due to the presence of a mismatch repair system (Schlotterer 2000). It has been shown that a functional mismatch repair (MMR) system reduces the mutation rate by about 100 to 1000-fold (Strand et al. 1993).

The MMR system (MMRS) in *E. coli* is constituted by three enzymes MutS, MutL and MutH. Homologues of MutS and MutL are present in most of the prokaryotes as well as in eukaryotes (Eisen 1998; Lin, Nei, and Ma 2007) whereas such a universal presence of MutH has been questioned (Claverys and Lacks 1986). Structure and function of MutS have been studied in *E. coli* and *Bacillus subtilis* (Yang 2000; Sixma 2001; Smith, Grossman, and Walker 2001). It is now known that the functional form of MutS is a homo dimer (structurally hetero-dimer) that recognizes the mismatched nucleotide and INDELs, and forms a complex with DNA (Lamers et al. 2000; Obmolova et al. 2000; Hsieh 2001; Iyer et al. 2006; Li 2008). The MutS-DNA complex further interacts with the MutL homo dimer protein in an energy dependent manner (Acharya et al. 2003; Selmane et al. 2003). The interaction of these two complexes activates MutH which is an endonuclease, which cleaves the newly synthesized strand by recognizing the methylation pattern of "G" in the motif GATC on template strand and removes the mismatched bases by the excision of incorrect nucleotides on nascent strand (primer strand) (Lee et al. 2005). MMR system not only repairs mismatches but also the small INDELs of bases less than 5bp caused by mutation of simple sequence repeats (Parker and Marinus 1992; Jaworski et al. 1995; Modrich and Lahue 1996; Schofield and Hsieh 2003). Mutation in the MMR system has been shown to destabilize simple sequence repeats (SSRs) in eukaryotes (Strand et al. 1993; Acharya et al. 1996) as well as in prokaryotes (Jaworski et al. 1995; Vogler et al. 2006).

1.7.3 Directionality in SSR evolution

As already mentioned SSR INDEL mutations happen as a consequence of slipped strand mispairing during replication (Streisinger et al. 1966; Levinson and Gutman 1987a; Schlotterer and Tautz 1992). Slippage on the template strand leads to contraction (deletion of repeat units) whereas slippage on the nascent strain manifests as an expansion (insertion of repeat units) of SSRs (Streisinger et al. 1966; Levinson and Gutman 1987a; Harr, Todorova, and Schlotterer 2002; Mirkin 2005; Garcia-Diaz and Kunkel 2006). If SSR mutations are evolutionally neutral then one should observe equal number of contraction and expansion events. However, it has been found that these two events do not occur with equal rates and certain biasness towards one of the events is always observed. This mutational bias in SSRs is referred to as the directionality of SSR evolution (Ellegren, Primmer, and Sheldon 1995; Rubinsztein et al. 1995; Rubinsztein, Leggo, and Amos 1995; Amos et al. 1996; Primmer et al. 1996).

1.7.3.1 Directionality in SSR evolution in eukaryotes

An interest to find the mechanism behind directionality of SSR evolution has increased greatly in the last decade with the observation that several hereditary diseases are associated with expansion of triplet repeats (Ashley and Warren 1995). In addition to this, mono and di nucleotide repeats are also unstable in colon cancer cells (Aaltonen et al. 1993; Ionov et al. 1993; Thibodeau, Bren, and Schaid 1993; Marra and Boland 1995). To understand the origin of these disease-associated instabilities and generation of directional biasness, one needs to know how SSR mutations are created and repaired. Though directional evolution of SSRs has been debated and commented in the past (Ellegren, Primmer, and Sheldon 1995; Rubinsztein et al. 1995; Rubinsztein, Leggo, and Amos 1995; Amos and Rubinstzein 1996; Amos et al. 1996; Primmer et al. 1996; Ellegren 2002; Harr, Todorova, and Schlotterer 2002; Webster, Smith, and Ellegren 2002; Mirkin 2005), sufficient understanding of factors influencing directionality of SSR mutations has not been gained. There have been conflicting reports with regard to the directionality of SSR evolution in prokaryotes and eukaryotes (Henderson and Petes 1992; Metzgar et al. 2002). In most of the cases dinucleotide repeats have been the subjects of study which reported that mutations are biased towards expansion in human as well as in swallow (Ellegren et al. 1997). Twerdi *et al* (Twerdi, Boyer, and Farber 1999) reported similar expansion bias in SSR mutation in both MMR deficient as well as MMR proficient mammalian cell lines. Similar results were reported by Yamada *et al.* in human MMR deficient cell lines (Yamada et al. 2002). Webster *et al* also reported similar di-nucleotide expansion bias by comparing the equivalent SSRs between the human and chimpanzee (Webster, Smith, and Ellegren 2002). Xu *et al* reported that long alleles of a tetra nucleotide SSR tracts show contraction bias (Xu, Peng, and Fang 2000). Huang *et al* (2002) reported that the long alleles of a dinucleotide repeat show bias towards contraction whereas short alleles are biased towards expansion. Even mononucleotides were also shown to be biased towards expansion in MMR proficient and deficient mammalian cell lines and has also been argued that MMR system does not influence the direction of SSR evolution (Boyer et al. 2002). However, in-depth study by Harr et al. in drosophila showed the role of mismatch repair in the directionality of SSR evolution (Harr, Todorova, and Schlotterer 2002). In wild type cell lines SSR mutations were significantly biased towards contraction whereas in the *spellchecker* mutation accumulation cell lines which were deficient for mismatch repair, SSR mutations were slightly biased towards expansion (Ellegren 2002; Harr, Todorova, and Schlotterer 2002).

1.7.3.2 Directionality in SSR evolution in prokaryotes

Very few prokaryotic systems have been studied from the stand point of directionality of SSR evolution. Studies on *Haemophilus influenza* and *E. coli* genomes (Morel et al. 1998; De Bolle et al. 2000) suggested that bacterial SSRs do show directionality during evolution. The length of SSR tracts were, $(AGTC)_{17\text{-}38}$ and $(AC)_{51}$ in *Haemophilus influenzae* and *E. coli* respectively. It is to be noted that *E. coli* has functional mismatch repair system (Levy and Cebula 2001) and its homologue is present in *Haemophilus influenza* as well. In *M. gallisepticum* where MMR appears to be absent (Himmelreich et al. 1996; Carvalho et al. 2005) it was observed that a trinucleotide SSR tract $(GAA)_{12}$ shows mutational bias towards deletion (Metzgar et al. 2002). Furthermore, a study in *E. coli* on the tetra nucleotide tract $(AAGG)_9$ had shown significant expansion bias (Eckert and Yan 2000). It should be noted that the length of SSRs considered in the aforementioned studies are much longer than the average length of SSRs in prokaryotic genomes (Field and Wills 1998; Mrazek, Guo, and Shah 2007; Sreenu et al. 2007).

1.7.4 SSR mutation models

There are three main theoretical models available in literature, which have been proposed for SSR mutations and these are given below.

Infinite alleles (IA) model

In this model, each mutation randomly creates a new allele. Applying this model to SSR loci, mutations alter the number of repeats. For example, an allele with 10 repeats is considered to be as closely related genetically to an allele with 15 repeats as to one with 16 repeats, *i.e.* proximity in terms of the number of repeats does not indicate a greater phylogenetic relationship. This is Wrights (Wright 1931) classical model in which he uses *F*-statistics.

Step-wise mutation model

Several models have been proposed to explain the dynamic nature of mutations in SSRs. Most of these models have been derived from the original step-wise mutation model (SMM) (Ota and Kimura 1973). The original SMM suggests that a mutation changes the length of a repetitive array, via the addition or removal of one repeat unit at a fixed rate (Shriver et al. 1993; Valdes, Slatkin, and Freimer 1993; Kimmel and Chakraborty 1996; Kimmel et al. 1996). When a SSR locus mutates, it gains or loses a repeat. This implies that two alleles differing by only one motif are more related (*i.e.* share a more recent common ancestor) than alleles differing by several repeats. Later, Di Rienzo et al. (Di Rienzo et al. 1994) tested the step-wise mutation model for a population with a known demographic history and postulated a modified version of the model called a two phase model that provided a better fit than the original step-wise mutation model. In this model, the vast majority of the mutations are single step mutations but multi step mutations are also allowed at a small rate. Following that, several more complex variants of the step-wise mutation model have also been proposed (Garza, Slatkin, and Freimer 1995; Kimmel et al. 1996; Nauta and Weissing 1996; Feldman et al. 1997). The SMM model is usually preferred when estimating relations between individuals and population structure, except in the presence of homoplasy (*i.e.* when two alleles are identical by state but not by descent). Homoplasy may seriously influence population studies involving high mutation rates and large population sizes together with strong allele size constraints (Estoup, Jarne, and Cornuet 2002).

Equilibrium model

Since a simple SMM does not explain SSR-length distributions (Di Rienzo et al. 1994), other models were proposed to explain the observed variations in the SSR tracts. Among these, the equilibrium model is an improved model postulated to explain SSR evolution. According to this model, a genome-wide distribution of SSR repeat lengths that rests at equilibrium, results from a balance between length and point mutations (Bell and Jurka 1997; Kruglyak et al. 1998; Calabrese, Durrett, and Aquadro 2001). This model proposes three mutational forces that operate on SSR sequences. According to this, DNA slippage mutations increase with the increasing repeat count to attain arbitrary high values and random point mutations break these long tracts into smaller units and make them immune to slippage. The random point mutations also create sufficiently long SSRs which can undergo slippage mutations. In a genome, there is a constant distribution of repeat lengths, governed by the rates of length and point mutations. Later, evidence supporting this model came from a study of homologous SSR loci in rat and mouse where long SSRs were preferentially found in the regions with low substitution rate (Santibanez-Koref, Gangeswaran, and Hancock 2001). This model has been well received in recent years because it explains the differences in SSR distribution among species and provides a well-dressed solution to the problem of why SSRs do not expand into enormous arrays.

1.7.5 Polymorphic SSRs and their effects

Polymorphism in SSRs can have different effects depending on the location of SSRs relative to the organization of genes. SSRs that are located far from coding regions may evolve neutrally and have very little effect on structure and function of genes. However, polymorphic SSRs residing in the open reading frames (ORFs) and upstream or downstream of regulatory elements produce a considerable effect on the mechanisms of gene transcription as well as translation. Different effects of SSR polymorphism relative to a gene is schematically shown in **Figure 1.2**. Furthermore, the severity of the effect depends on the repeat type and the repeat location. Polymorphic SSRs of repeating motif length three nucleotides (triplet) or multiples of three always bring out in-frame mutations whereas that of non-triplet repeats bring out either frame-shifts (repeat number variation is not multiple of three) or in-frame mutations (repeat number variation is multiple of three). A schematic representation of various effect caused by INDELs of repeat units in coding regions is shown in **Figure 1.3.** In the following sections a review on SSR polymorphism reported so far and their effects on the various organisms, is given.

1.7.6 SSR polymorphism in Eukaryotes

Mutations in SSRs can be beneficial, deleterious or neutral to organisms. Some of the known beneficial as well as deleterious effects of SSR mutations in eukaryotes are described in the following section.

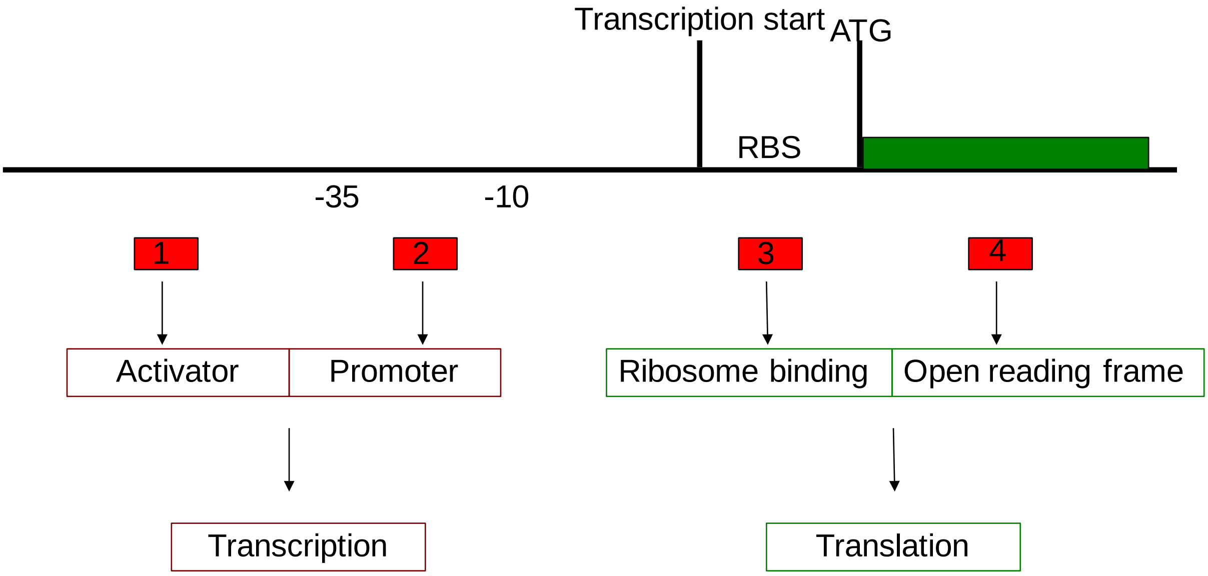


Figure 1.2: An Illustration of the four sites, relative to a gene, at which mutations in SSRs affect that gene. SSR mutations at regions 1 through 4 can affect transcription initiation (regions 1 and 2), translation (region 3 and 4) (see the text for the example from each regions).

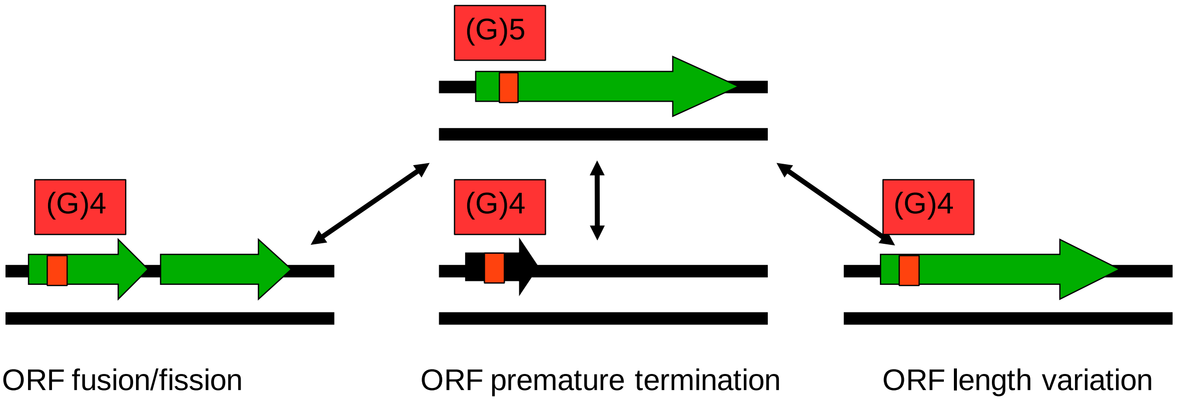


Figure 1.3: A schematic representation of the various effects caused by INDELs of repeat units in SSRs in the coding regions (shown as arrows) of the three mycobacterial genomes, *M. bovis, M. tuberculosis* (CDC1551) and *M. tuberculosis* (H37Rv) (Sreenu 2006).

1.7.6.1 Beneficial effect of SSRs polymorphism in eukaryotes

Temperature compensation of circadian rhythm in Drosophila:

The *Drosophila melanogaster* clock gene contains a polymorphic hexanucleotide SSR which encodes for poly (Thr-Gly) (Sawyer et al. 1997). Two of the alleles of this gene are more common than the others: the shorter (Thr-Gly)$_{17}$ allele yields a circadian period closer to 24 hours, whereas the longer (Thr-Gly)$_{20}$ variant yields better temperature compensation so that temperature fluctuations have a lesser impact on circadian cycle. Additional evidence has recently come from the 'Evolution Canyon' ecological study site at Mount Carmel in Israel (Kashi and King 2006). This canyon presents a dramatic microclimatic contrast, with the sunny, south-facing slope experiencing greater temperature and drought stress than the north-facing slope. The longer, cold-climate allele of the Drosophila *per* gene was more than twice as abundant on the cooler, north-facing slope than it was on the warmer, sunny slope, supporting the conclusion that natural selection of these SSR allele 'fine-tunes' the Drosophila circadian clock to differing environmental conditions (Zamorzaeva et al. 2005).

Social behavior in voles:

The experimental evidence of the role of SSR polymorphism intimately involved in phenotypic variation at the interspecies and at the individual level has recently been provided by Hammock and Young's study of social behavior in voles (Hammock and Young 2004; Hammock and Young 2005). Prairie and pine voles (*Microtus ochrogaster* and *Microtus pinetorum*) are highly social, monogamous rodent species, whereas montane and meadow voles (*Microtus montanus* and *Microtus pennsylvanicus*) are asocial and not monogamous. This difference in behavior has been suggested due to the expression of vasopressin receptor gene expression. Although the protein-coding region of vasopressin receptor (*avpr1a*) is highly conserved among voles, the $(GA)_n$ dinucleotide SSR in the 5' UTR of this vasopressin receptor gene differs in the repeat count. The two social species have a long, SSR in the 5' regulatory region of this gene, much of which is absent in the two asocial species (Hammock and Young 2005). These differing social behaviors depend on the pattern of expression of the vasopressin receptor (encoded by *avpr1a*), with greater levels of expression in the ventral forebrain of the social voles (Hammock and Young 2005; Kashi and King 2006). Remarkably, comparison of the *avpr1a* orthologues in humans, chimpanzees (*Pan troglodytes*), and bonobos (*Pan paniscus*) shows that the SSR is conserved in humans and bonobos, both species sharing similar sociosexual behaviors, whereas in chimpanzees, a 360-bp sequence encompassing the SSR is deleted (Hammock and Young 2005). Experiments that transfected two versions of the SSR locus from social and asocial species into cultured rat cells showed that the species divergence in SSR lengths at this locus is sufficient to alter expression of *avpr1a* in a manner that is dependent on cell type. Such effects of SSR repeat number on cell-type-specific gene expression in culture together with the correlation of SSR repeat length with social behavior and gene expression in intact animals support a strong presumption that SSR variation, mediated through expression of the gene encoding the vasopressin receptor is at least partially responsible for both individual and interspecies variation in social behavior phenotypes in voles.

1.7.6.2 Harmful effect of SSR polymorphism in eukaryotes

Most of the effects of SSR polymorphism studied in humans highlight harmful effects of SSR length variation in eukaryotes. Some of the well studied SSR polymorphisms have been described as below, with an overview on the molecular basis of functional effects of SSRs in both coding and non-coding domains.

Disease-causing repeat instability is an important and unique form of mutation that is linked to more than 40 neurological, neurodegenerative and neuromuscular disorders. The locations of these polymorphic SSRs relative to the genes concerned are schematically given in **Table 1.2**. It can be seen in figure that most of the repeats are trinucleotide repeats.

Polymorphic SSRs in coding regions:

Phenotypic effects of SSRs in coding regions have been extensively studied in relation to human diseases, revealing abundant evidence on human neuronal disorders and cancers. In human cDNA database, more than 92% of the predicted microsatellite variations are caused by SSRs of length three and multiples of three (Wren et al. 2000). Though, trinucleotide repeat variations do not shift the reading frame of ORFs, expansion of the tract above a certain threshold has been observed to cause disorders, some of which are discussed below. The list of experimentally characterized polymorphic SSRs is given in **Table 1.2**. On reaching its threshold level, the size of the repeated array tends to increase with subsequent generations. The largest class of diseases results from the expansion of coding CAG repeats that are translated into extended $(Gln)_n$ tracts within the corresponding proteins. These dominantly inherited diseases include Huntington’s disease (HD), dentatorubro-pallidoluysian atrophy (DRPLA), spinobulbar muscular atrophy (SBMA), and spinocerebellar ataxia (SCA1, SCA2, SCA3, SCA6, and SCA7; table 3) (Li et al. 2004). All eight disorders are progressive, typically striking in midlife, and cause increased neuronal dysfunction and eventual neuronal loss, 10–20 years after the onset of symptoms (Li et al. 2004). Expansion of the CAG repeat in the androgen receptor gene leads to various other disorders including abnormal activity of androgen receptor, risk of prostate cancer (Coetzee and Irvine 2002), SBMA with partial androgen insensitivity and is also related to Kennedy's disease (Dejager et al. 2002). The only reported disorder caused by non-CAG repeat expansions in human coding regions is oculopharyngeal muscular dystrophy. This is a result of CGC repeat expansion in the PABP2 gene (Brais et al. 1998).

On the other hand SSR length variations in human genes that are leading to frame-shifts are associated with several cancers and other neuromuscular disorders (**Table 1.2**). It can be seen in Table 1.2 that the mutational inactivation is caused mainly by frame shift occurring within the $(A)_n$ tract. The Association of microsatellite mutation with cancer was first reported independently by two different research groups in 1993 (Aaltonen et al. 1993; Thibodeau, Bren, and Schaid 1993). Since then, a wide variety of cancer types have been found that are associated with elevated levels of microsatellite mutations (Li et al. 2004). Most of the cancer causing frame-shift mutations due to SSR length variation is observed in genes involved in the repair mechanism, tumor suppression, cell signaling and cell cycle (Li et al. 2004).

Table 1.2: Simple sequence repeats polymorphism in human

Organism	Repeat	Location	Gene	Phenotype	Reference
Human	TCAT	intron	TH gene	Acts as transcription regulatory element	(Meloni et al. 1998)
Human	CA	intron	egfr	Enhances egfr transcription and involved in breast carcinogenesis	(Tidow et al. 2003)
Human	T	intron	ATM gene	Aberrant splicing and abnormal transcription in colon tumor cells	(Ejima, Yang, and Sasaki 2000)
Human	CCTG	intron	ZNF9	Leads to DM2 disease	(Liquori et al. 2001)
Human	ATTCT	intron	SCA10 gene	SCA10 disease	(Matsuura et al. 2000)
Tilapia (fish)	GT	5'-UTR	prl 1	prl 1 Influence gene expression and growth response of salt-challenged fishes	(Streelman and Kocher 2002)
Human	CGG	5'-UTR	FMR-1	(CGG)>200 cause human mental retardation (CGG)40to200 related in fragile-X-like cognitive/psychosocial impairment (CGG)40to60 associated in woman ovarian dysfunction	(Kenneson et al. 2001) (Franke et al. 1998) (Youings et al. 2000)
Human	GCC	5'-UTR	FMR-2	Reduced FMR2 causing abnormal neuronal gene regulation	(Cummings and Zoghbi 2000)
Human	CAG	5'-UTR	PPP2R2B	(CAG)5578 causes SCA12 disease	(O'Hearn et al. 2001)
Human	CTG	3'-UTR	DMPK	Expansion causes DM1 disease	(Aslanidis et al. 1992)
Human	CTG	3'-UTR	SCA8	Expansion causes SCA8 disease	(Koob et al. 1999)
Human	GAA	intron	FRDA	Leads to FRDA disease	(Campuzano et al. 1996)
Human	A	Coding	hMSH2, hMLH1, hMSH6, hPMS1, hPMS2	Causes human cancers	(Vassileva et al. 2002)
Human	A	Coding	MBD4/MED1	Causes human cancers	(Yamada et al. 2002)
Human	A	Coding	TGFRII, IGFIIR, WISP,GRB-14, AXIN-2	Tumor-suppressive	(Markowitz et al. 1995)
Human	G	Coding	BAX, caspase 5, APAF-1, BCL-10, FAS	Tumor-suppressive	(Rampino et al. 1997)
Human	A	Coding	TCF-4, CDX2	Tumor-suppressive	(Duval et al. 1999)
Human	A	Coding	2M	Tumor-suppressive	(Bicknell et al. 1996)
Human	A	Coding	BLM, CHK1, RAD-50	Tumor-suppressive	(Duval and Hamelin 2002)

Polymorphic SSRs in non-coding regions:

There are other triplet repeat expansions in untranslated regions (UTRs) of the genes, which also cause disorders in humans. Most of these repeats are located in the 5'-UTR of the ORFs (please see the **Table 1.2**). The first genetic disease reported in this category was the fragile X syndrome, the most common form of familial mental retardation (Fu et al. 1991). This syndrome is a result of the expansion of a CGG trinucleotide repeat in the 5'-UTR of the FMRI gene (Fu et al. 1991; Kremer et al. 1991; Verkerk et al. 1991). Soon afterwards, the same loci with different repeat numbers were shown to cause different symptoms. Repeat numbers 40-60 of the CGG microsatellite are associated with other fragile-X-like phenotypes and woman ovarian dysfunction (Youings et al. 2000). While repeat numbers 40-200 are related to fragile-X-like cognitive/psychosocial impairment (Franke et al. 1998); a repeat number greater than 200 has been reported to result in the loss of FMR-1 function, thus causing mental retardation (Kenneson et al. 2001). A similar repeat expansion observed in the 5'-UTR of the FMR-2 gene causes abnormal neuronal gene regulation (Cummings and Zoghbi 2000). Other ailments arising from triplet repeat expansions in the UTR regions include Myotonic dystrophy (DM) (CTG repeat expansion in 3'-UTR of DMPK (Aslanidis et al. 1992)), Friedreichs ataxia (GAA repeat expansion in the first intron of FRDA (Campuzano et al. 1996)), and spinocerebellar ataxia type 12 (CAG repeat expansion in 5'-UTR of PPP2R2B (O'Hearn et al. 2001)). SSR polymorphism in the UTRs especially in the first intron of many genes is reported to cause several diseases (figure 1.4). Variation of the tetrameric repeat of TCAT located in the first intron of the tyrosine hydroxylase (TH) gene is shown to be a regulatory sequence *in vitro* (Meloni et al. 1998). Similarly, the CCTG expansion in the first intron sequence of the zinc finger protein 9 (ZNF9) is involved in the manifestation of Myotonic dystrophy (DM) (Liquori et al. 2001), and polymorphism of the CA repeat located in the first intron of the epidermal growth factor receptor (egfr) is associated with breast cancer (Tidow et al. 2003). Intronic changes of poly “T” in the ATM gene is shown to interfere in splicing and cause colon cancer (Ejima, Yang, and Sasaki 2000). In spinocerebellar ataxia type 10, large expansions of the ATTCT pentanucleotide repeat is observed in the 9th intron of the SCA10 gene (Matsuura et al. 2000).

To date, SSR repeat linked disorders have been observed only in humans and cannot be remedied after the onset. A study of all these disorders is based only on genetic correlation. A clear involvement of these repeats in the actual disease process is yet to be established. The common feature of all trinucleotide repeat expansion disorders is the existence of a threshold length below which the repeats are normal and beyond which they become “pathogenic". Moreover, this threshold level for the onset of a disease varies among individuals. In all the known disorders, the severity of the disease correlates with the length of the repeat array. The molecular mechanisms behind the SSR expansion process still remain obscure, although the most common explanation is that these motifs create a form of internal hairpin structure when single stranded (Usdin and Grabczyk 2000). So far, only three triplets have been known to be associated with expansion diseases. These are the triplets CGG|CCG, CTG|CAG, and GAA|TTC. The repeats of these triplets can form a variety of non-canonical secondary structures depending on the sequence of the repeat, the number of repeats, the pH, ionic strength, DNA concentration, superhelical density and whether or not the repeat is single stranded (Usdin and Grabczyk 2000). The fact that all hypervariable sequences form secondary structures led to the suggestion that these structures may play a role in the instability of the repeat sequences and in some cases, in disease pathology (Usdin and Grabczyk 2000).

1.7.7 SSR polymorphism in Prokaryotes

SSRs are less abundantly found in prokaryotes than eukaryotes (with respect to tract densities as well as numbers). However, compared to eukaryotes, a lot more information is available on SSR repeat length variation in prokaryotes. Unlike in eukaryotes, SSR repeat length variations seemed to have advantageous effects on prokaryotes (Moxon et al. 1994). SSR variations enable prokaryotes, particularly, bacteria to respond to diverse environmental factors, and many of them are clearly related to bacterial pathogenesis and virulence. One major set of selective forces which operates to shape the phenotype of the bacterial pathogen is the host immune system. The host immune response following contact with the pathogen is itself adaptive and seeks to eliminate or restrict bacterial replication. Thus, successful bacterial pathogens must be able to avoid or adapt to evolving host defenses (Brunham, Plummer, and Stephens 1993). Many pathogens have evolved the ability to alter surface-exposed molecules, most often in response to selective pressures associated with the host immune system (Brunham, Plummer, and Stephens 1993). Pathogenic bacteria exhibit numerous examples of this adaptive strategy, and a range of molecular mechanisms have evolved in these bacteria for generating genetic variation at individual loci termed “contingency loci” (Moxon et al. 1994). Many contingency genes are controlled by simple sequence DNA repeats that accumulate reversible, *rec*-independent mutations at high frequencies (Bayliss, Field, and Moxon 2001). These reversible changes alternate phenotypes in a heritable and reversible manner which can be classified as phase variation or antigenic variation (van der Woude and Baumler 2004). These terms, phase variation and antigenic variation, however, have been used in various ways and hence a description of these terms has been provided below, followed by an overview of many bacterial proteins and structures that are under the control of phase or antigenic variation due to SSR polymorphism in either the coding or non-coding regions.

Phase variation:

Phase variation in general refers to a reversible switch between an "all” and a “none" (on/off) expressing phase, and the frequency of this reversion should exceed that of a random mutation, resulting in variation in the level of expression of one or more proteins between individual cells of a clonal population. Thus, in a clonal population after cell division, the majority of daughter cells will retain the expression phase of the parent but a minority will have switched expression phase. The switch is a stochastic event, even though the chance that it occurs in some cases can be influenced by external factors; in other words, the switching frequency can be modulated. The frequency with which this occurs is characteristic of the gene, the bacterial species, and the regulatory mechanism. This can be as high as a change in 1 cell per 10 per generation but more often is of the order of 1 change per 10^3 cells per generation. The term "phase variation" is used in other contexts as well, describing phenotypic "phase variants" of a species in which the change is irreversible or variants that are a result of environmental regulation, selection, or unidirectional mutation. In this review, however, I adhere to the definition in the sense that the expression phase must be inherited by a genetic mechanism and that this change must be reversible. The actual switching frequencies are not reported here, because the methods that are used to determine them vary significantly and because they can be modulated by growth conditions. Differences are therefore difficult to interpret.

Antigenic variation:

Related to phase variation, antigenic variation refers to the expression of a number of alternative forms of an antigen on the cell surface (such as lipoproteins, polysaccharides, type IV pili); this generates within a clonal population of individual cells that are antigenically distinct, allowing bacterial pathogens to escape the host immune system. At the molecular level, some cases of antigenic variation share common features with phase variation mechanisms.

1.7.7.1 SSR polymorphism in coding regions:

The list of coding SSRs which show the length variation in different genera of bacteria and which are also experimentally characterized is given in **Table 1.3**. In the coding regions functional changes are mostly due to changes in reading frames owing to SSR length variations however radical functional changes have also been reported in a few cases by variation in triptet repeats. A description of the effect of SSR length variation in some selected species is given below.

Phase variations as a result of SSR polymorphism in genes have been extensively studied in *H. infuenzae* and Neisseria species. The first report of SSR polymorphism affecting expression of a bacterial virulence factor was that of the pentanucleotide repeat CTCTT in a gene coding for an outer membrane protein (Opa) in *Neisseria gonorrhoeae* (Stern et al. 1986; Stern and Meyer 1987). The relationship between the superficial Opa protein composition of bacterial isolates and invasiveness into the human epithelium has been demonstrated experimentally (Makino, van Putten, and Meyer 1991). In *H. infuenzae*, the first genetic loci exhibiting SSR polymorphism was *lic1* that is involved in the synthesis of carbohydrate structures on the outer-membrane lipopolysaccharide (LPS) (Weiser, Love, and Moxon 1989). Subsequently, the role of tetrameric repeats in lipopolysaccharide (LPS) genes was reported in *H. influenzae* (Weiser, Love, and Moxon 1989; High, Jennings, and Moxon 1996). In the 5’ end of this gene, within the open reading frame, variation of the tetramer repeat CAAT was shown to shift the upstream initiation codons in or out of frame and by such a shift, these 4 bp units create a translational switch (Weiser, Love, and Moxon 1989). Subsequently, other genes (lic2 and lic3) consisting of the CAAT repeats and involved in the same pathway were also reported (Weiser et al. 1990; Robertson and Meyer 1992; Roche and Moxon 1995). Consequently, similar tetrameric repeats located in genes and displaying phase variations were also identified in other pathogens, like Neisseria species, *Moraxella catarrhalis* and *Haemophilus somnus* (Peak et al. 1996; Inzana et al. 1997).

In *N. gonorrhoeae*, expression of the outer membrane protein P.II is controlled by the pentamer repeat `CTCTT` in the leader sequence (Murphy et al. 1989). The signal peptide coding regions of P.II genes contain variable numbers of tandem repeats of the sequence CTCTT. Changes in the number of CTCTT units, leading to frameshifts within the gene, are responsible for changes in P.II expression.

Phase variation by mononucleotide SSRs in genes of *N. gonorrhoeae* and *Chlamydia pneumoniae* were also reported. In *N. gonorrhoeae*, poly “G” tract variation in one of the lipooligosaccharide (LOS) synthesis genes, namely, *lsi2* was shown to be responsible for LOS-specific phenotypic variation (Burch, Danaher, and Stein 1997). This study has demonstrated that antigenic variation in *Neisseria gonorrhoeae* is due to SSR mutation which manifests into production of multiple lipooligosaccharides due to shifting of the reading frame. In addition to that localized frameshift mutation generates selective, high-frequency phase variation of a surface lipoprotein encoded by a mycoplasma ABC transporter operon (Theiss and Wise 1997). SSR mutations in the antigenic membrane protein (P78) are regulated by the variation of ploy "A" repeats in *Mycoplasma fermentans* (Theiss and Wise 1997). By alternations in the number of adenosines, strains were either capable of or prevented from producing a substrate binding lipoprotein.

Phase variation of the polymorphic membrane proteins (Pmp) in obligate intracellular bacterium *C. pneumoniae* was also due to polymorphism of poly "G" tracts, and has been shown to be involved in virulence and pathogenesis (Grimwood, Olinger, and Stephens 2001).

Most of the SSR length variations discussed above are of non- triplet repeat types and hence their polymorphism were associated with drastic effects as a consequence of changes in reading frames of the concerned coding regions. Very rarely polymorphism in triplet repeat has been shown to be associated with a change in gene function. For example, in *E. coli*, the variation of length from (TCT)4 to (TCT)5 in ahpC was observed to bring in a conversion in the function of the protein, from a peroxiredoxin to a disulfide reductase (Ritz et al. 2001).

Table1.3: SSR polymorphism in bacteria leading to frame-shift mutations in ORFs.

Organism name	Repeat motif	Ecological niche	Gene (function)	Reference
Neisseria gonorrhoeae	CTCTT	Human	Opa (opacity protein)	(Stern and Meyer 1987)
Neisseria gonorrhoeae	CTCTT	Human	P.II (Membrane protein)	(Murphy et al. 1989)
Haemophilus influenzae	CAAT	Human	lic-1 (LPS)	(Weiser, Love, and Moxon 1989)
Neisseria gonorrhoeae	G	Human	pilC (pili)	(Jonsson, Nyberg, and Normark 1991)
Haemophilus influenzae	CAAT	Human	lic2A (LPS)	(High, Deadman, and Moxon 1993)
Neisseria gonorrhoeae	G	Human	lsi-2 (LPS)	(Danaher et al. 1995)
Neisseria meningitidis	C	Human	siaD (polysialyltransferase)	(Hammerschmidt et al. 1996)
Haemophilus influenzae	CAAT	Human	IgtC (glycosyltransferase)	(Hood et al. 1996)
Neisseria gonorrhoeae	G	Human	lgtC, and lgtD (LPS)	(Yang and Gotschlich 1996)
Vibrio cholerae	G	Human	tcpH (regulatory protein)	(Carroll et al. 1997)
Haemophilus somnus	CAAT	Bovine	lex-2B (LPS)	(Inzana et al. 1997)
Mycoplasma fermentans	A	Human	P78 (substrate binding lipoprotein)	(Theiss and Wise 1997)
Neisseria gonorrhoeae	G	Human	hpuA (hemoglobin-utilization)	(Chen, Elkins, and Sparling 1998)
Haemophilus influenzae	G	Human	IgtA (LPS)	(Jennings et al. 1999)
Neisseria meningitidis	G	Human	HmbR (hemoglobin utilization protein's receptor)	(Lewis et al. 1999)
Haemophilus influenzae	CCAA	Human	hgbA and hgbB (hemoglobin-haptoglobin binding)	(Cope, Hrkal, and Hansen 2000)
Haemophilus influenzae	AGTC	Human	Mod (DNA methyltransferases)	(De Bolle et al. 2000)
Helicobactor pylori	C	Human	flip (flagellar basal body)	(Josenhans et al. 2000)
Campylobacter jejuni	G	Human	wlaN (beta-1,3 galactosyltransferase)	(Linton et al. 2000)
Campylobacter coli	T	Human	flhA (flagellin)	(Park, Purdy, and Leach 2000)
Helicobacter pylori	A & C	Human	Lewis antigen	(Wang et al. 2000)
Haemophilus somnus	GA	Cattle pathogen	lob-2A	(Wu et al. 2000)
Streptococcus pyogenes	AACAA	Human	SclB (surface protein)	(Rasmussen and Bjorck 2001)
Helicobacter pylori	C	Human	pldA (phospholipase A)	(Tannaes et al. 2001)
Neisseria gonorrhoeae	G	Human	pgtA (pilin galactosyl transferase)	(Banerjee et al. 2002)

Helicobacter pylori	G	Human	Res (type III DNA methyltransferase)	(de Vries et al. 2002)
Campylobacter jejuni	G	Human	maf1 (flagella)	(Karlyshev et al. 2002)
Streptococcus pneumoniae	A	Human	*pspA*	(Pericone et al. 2002)
Streptococcus pneumoniae	A	Human	*spxB*	(Pericone et al. 2002)
Streptococcus pneumoniae	T	Human	*Xba*	(Pericone et al. 2002)
Neisseria meningitidis	G	Human	pptA (pili)	(Warren and Jennings 2003)
H. pylori	CT	Human	BabB (blood group antigen-binding adhesin)	(Backstrom et al. 2004)
Bacillus subtilis	A	Human	swrA (swarming motility)	(Kearns et al. 2004)
Pseudomonas putida	G	Soil	flhB (flagellar export system)	(Segura et al. 2004)
Staphylococcus aureus	A	Human	mapW (major histocompatibility complex analogue)	(Buckling et al. 2005)
Campylobacter jejuni	A	Human	FlgR (response regulator)	(Hendrixson 2006)
Neisseria meningitidis	C	Human	AusI (autotransporters)	(van Ulsen et al. 2006)

In contrast to the above mentioned individual locus specific SSR polymorphism studies in different genera, a whole genome comparison of SSRs has been done in Mycobacteria. Cross-genome comparative studies of equivalent SSRs in mycobacteria revealed hundreds of examples of SSR polymorphism (Sreenu et al. 2006). It was observed that length variation in SSRs is associated with changes in ORFs such as splitting of ORFs, fusion of ORF, premature terminations etc. (please see Fig. 1.3). These observations threw light on possible roles of SSR length polymorphisms in imparting novel functions, thus bringing out plasticity in mycobacterial genomes, perhaps associated with molecular mechanisms that are involved in their adaptability and evolution.

1.7.7.2 SSR polymorphism in non-coding regions

As mentioned in section 1.7.5 that SSR polymorphism in non-coding regions can influence expression of genes, provided, they occur: (1) in between the -10 and -35 site of promoter; (2) In the activator region (just upstream of -35 site) and (3) down streams of -10 site but upstream of translation start site.

Regulation at the level of transcription occurs when the repeats are located in the promoter region between the -10 and -35 sites for RNA polymerase binding. **Table 1.4** lists the polymorphic SSR in non-coding regions which are involved in the regulation of gene expression. The spacing of these sites is critical for the level of transcription, and even a single-nucleotide deviation from the optimal 17-nucleotide (nt) spacing has an effect. *Haemophilus influenzae* infections are preceded by airway colonization, a process facilitated by fimbriae. In *H. influenza*, variation of the TA repeat (9, 10 and 11 repeat unit) in the promoter region of hifA/hifB genes that encode fimbrial subunit proteins is a classical example of SSR involvement in gene regulation (Van Ham et al. 1993; Van Ham et al. 1994). The expression of *H. influenzae* fimbriae is subject to reversible phase variation between three expression levels. This phenomenon is controlled at the transcriptional level of two divergently orientated genes, *hifA* and *hifB*, encoding the major fimbrial subunit and the fimbrial chaperone, respectively (van Ham et al. 1993). The *hifA* and *hifB* promoter regions were found to be clustered through an almost complete divergent overlap with a variable DNA backbone of repetitive TA units (van Ham et al. 1993; van Ham et al. 1994). Variation in the number of units changes the normally strictly constrained spacing between the -35 and -10 sequences and controls the bidirectional transcription initiation, thus forming a novel mechanism directing multiple gene transcription (van Ham et al. 1993).

Antigenic diversity is generated in the wall-less pathogen *Mycoplasma hyorhinis* by combinatorial expression and phase variation of multiple, size-variant membrane surface lipoproteins (Vlps). A contiguous strand of adenine residues to the upstream of the -10 box that is subjected to frequent mutations, has been shown to regulate phase variation of size-variant membrane surface lipoproteins (Vlps) in *Mycoplasma hyorhinis* (Yogev et al. 1991). Furthermore, in the other species of Neisseria like *N. meningitides*, transcription of the outer membrane protein (Opc) is under the regulation of the poly "C" repeat tract (Sarkari et al. 1994). Efficient expression of Opc occurred in strains with 12 to 13 cytidine residues, intermediate expression in strains with 11 or 14 residues and no expression with < or = 10 or > or = 15 residues. This unusual regulation may have evolved because the Opc protein enables meningococcal invasion and is immunogenic (Sarkari et al. 1994).

Table 1.4: SSR polymorphism in the non-coding regulatory regions leading to phase and antigenic variation in bacteria.

Organism name	**Repeat motif**	**Ecological niche**	**Phenotype(s) affected**	**Location**	**Reference**
Bordetella pertussis	C	Human	Fim (Fimbriae)	Up stream of -35	(Willems et al. 1990)
Mycoplasma hyorhinis	A	Pig	Vlp (lipoprotein)	Between -10 and -35	(Yogev et al. 1991)
H. influenzae	TA	Human	hifA and hifB (Fimbriae)	Between -10 and -35	(van Ham et al. 1993)
Neisseria meningitidis	C	Human	Opc (Membrane protein)	Between -10 and -35	(Sarkari et al. 1994)
Mycoplasma gallisepticum	GAA	Avian pathogen	pMGA	Upstream of promoter	(Glew et al. 1998)
Neisseria gonorrhoeae	G	Human	FetA (siderophore receptor)	Between -10 and -35	(Carson et al. 2000)
Moraxella catarrhalis	G	Human	UspA1	-10 and SD	(Lafontaine, Wagner, and Hansen 2001)
Mycoplasma mycoides	TA	Bovine	Vmm (surface protein)	Between -10 and -35	(Persson et al. 2002)

Phase variation of individual fimbrial genes in *B. pertussis* is proposed to occur as a result of a change in a poly “C” tract that alters the distance between the binding sites of an activator and RNA polymerase (Willems et al. 1990). Similarly, in *N. meningitidis* strain MC58, a change in the unit repeat number of TAAA repeat motif in the sequence upstream of the -35 sequence of the promoter of the adhesion encoding *nadA* gene affects its promoter strength (Martin et al. 2005).

Transcription can also be affected by changes in repeat sequences located outside of the promoter. In certain isolates of *Moraxella catarrhalis*, the length of a poly (G) tract that is located downstream of the promoter for the adhesin gene *uspA* but upstream of the translation initiation site correlated with the level of gene expression (Lafontaine, Wagner, and Hansen 2001; van der Woude and Baumler 2004). This is an example where SSR polymorphism affects the ribosome binding efficiency at SD region. The UspA1 protein of *Moraxella catarrhalis* has been shown to function as an adhesin that mediates adherence to human epithelial cell lines *in vitro* (Lafontaine et al. 2000). Nucleotide sequence analysis revealed that isolates expressing relatively high levels of UspA1 had 10 G residues in their uspA1 poly (G) tracts, whereas isolates that expressed much lower levels of UspA1 had 9 G residues. This poly (G) tract was located 30 nucleotides (nt) upstream of the uspA1 ORF and 168 nt downstream of the uspA1 transcriptional start site (Lafontaine, Wagner, and Hansen 2001).

1.8 The present work

A review of literature as presented in this chapter demonstrates that one of the most important properties of the SSR tracts is its repeat unit length variation. The SSRs are an asset for the bacteria and their polymorphic nature has been used by many pathogens for generating genetic variation in a heritable and reversible manner. Pathogenic bacteria exhibit numerous examples of this adaptive strategy at individual loci termed “contingency loci”. From the examples given in **Table 1.3** and **1.4** it is clear that the use of SSR polymorphism as a phenotypic switch is a widespread strategy and this has been used by various bacterial genera belonging to different bacterial class. However from these two tables it can also be inferred that majority of the reported SSR polymorphisms are seen in some specific genera (Neisseria, Haemophilus, Mycoplasma, Helicobactor and Campylobacter). In addition to that most of the SSR polymorphisms are reported in the genes coding for surface proteins or in those genes which are involved in the biosynthesis of extracellular structure of an organism either directly or indirectly. Do these groups of bacteria lack some system which is essential for the stability of SSRs? Are they more exposed to the host immune machinery? Are these the most studied class of bacteria? What are the factors which can modulate the dynamics of this SSR polymorphism?

In this genomic era nearly thousands of complete bacterial genome sequences are available and for most of the species complete genome sequences are available for more than one strain. The availability of complete genome sequences of multiple strains within species of various prokaryotes is a very important asset to study the mutational dynamics of SSRs in prokaryotic genomes. The computer-based approach has the advantage that it is rapid and the SSR length variation is identified without a bias to a certain specie or region of a genome or functional class of proteins, allowing phase variation to be discovered for novel classes of protein or organism.

Though the SSR loci have been used by various groups of bacteria as a munitions store for generating diversity to cope with the host immune response or to save their resources if a particular phenotype is not required, to my surprise, not a single report is available for any of the SSR polymorphism in the deadly pathogen *Yersinia pestis*, which has been a major cause for millions of deaths in the past. In order to study the SSRs in this pathogen I analyzed the SSR distribution, enrichment and polymorphism in Yersinia (*Yersinia pestis* and *Yersinia pseudotuberculosis*) genome. Because the complete genome sequence of a number of strains of ancestor and descendant are available, it is possible to computationally analyze individual genome as well as compare genomes to infer evolution of SSRs across its different strains and species. The Yersinia genome was selected due to the following reasons: (a) Sixteen complete genome sequences of *Yersinia pestis* and four complete genome sequences of *Yersinia pseudotuberculosis* are available, (b) *Yersinia pseudotuberculosis* is an ancestor of *Yersinia pestis,* (c) Ancestor *Yersinia pseudotuberculosis* which causes gastritis is far less virulent than descendent *Yersinia pestis* which is the agent of bubonic plague and (d) Of the sixteen strains of *Yersinia pestis* one of the strains is non pathogenic to human (non plague causing). Hence the study was aimed to understand any possible role of SSRs in the evolution of the deadly pathogen from its mildly pathogenic ancestor.

In addition to that from a review on the known cases of SSR polymorphism in prokaryotes it could be inferred that some groups of bacteria (Neisseria, Haemophilus, and Mycoplasma etc.) are well studied for the role of SSR polymorphism while for others limited information is available. Furthermore in most of the cases the major focus was on the surface proteins, LPS, extracellular proteins etc. Hence in the present work a global study on the mutational dynamics of SSR length polymorphism in the coding and non-coding regions of forty-three prokaryotic species was done. In our global analysis, I focused on some very basic but unknown facts about SSRs in prokaryotic genomes. For example; are the coding and non-coding regions of genomes equally dispensable for SSR length polymorphism? Are the SSR length polymorphisms a random process or are they modulated by selection? Do the host adapted pathogens and non pathogenic bacteria have different densities of PSSRs? What is the directionality of SSR evolution in prokaryotes? What are the factors which modulate the directionality of SSR evolution? What could be the reason for the prokaryotic genomes to harbor small tracts of SSRs compared to their eukaryotic counterparts? This thesis is an attempt to answer these questions through the global analysis of SSR length polymorphism in various species of prokaryotes.

1.9 Summary

This chapter is an introductory review on simple sequence repeats in both eukaryotes and prokaryotes covering details of their discovery, classification, nomenclature, methods for extraction, origin, distribution, abundance, polymorphisms and evolution.

The review of literature about SSR, as presented in the current chapter, is a reflection of current understanding about SSRs. In order to begin, a historical view which led to its discovery and nomenclature is presented. Later discovery of more and more number of SSR tracts led to the classification of SSRs. Further an overview of different tools available for the extraction of the SSRs is given. Futhermore an accepted view of the origin of SSRs is also provided. Since different SSR extraction programs have different settings and the default setting of different programs output different number of SSRs, therefore, with a cautionary note an overview on the studies conducted on the distribution of SSR in different species of eukaryotes and prokaryotes is presented. In addition to that a description of some accepted SSR mutation models explaining the observed length distribution of SSR in known literature is also given. Rests of the chapters focus on one of the very important property of the SSR i.e. its polymorphic ability. Hence the mechanism of length polymorphism and the factor which influences this was described. Later some examples of experimentally known effects of SSR polymorphism in eukaryotes and prokaryotes were provided. At the end, an outlook of the work presented in this thesis was also given.

Chapter 2

A Study on Distribution and Enrichment of Simple Sequence Repeats in Yersinia Genomes

2.1 Introduction

As mentioned in the previous chapter simple sequence repeats (SSRs) are highly polymorphic and therefore offer certain amount of plasticity to genomes. They are one of the key sources of genetic variation in bacteria. Variations in the length of SSR tracts often cause changes in structure and function of the nearby genes leading to certain phenotypic changes such as changes in motility, colony morphology etc (van der Woude and Baumler 2004). Hence it is of importance to study pathogens with regard to distribution, abundance and enrichment of SSRs. This chapter presents the details of my study carried out on SSRs in Y*ersinia Pestis* and *Yersinia pseudotuberculosis* genomes. *Yersinia pestis* is a well known causative agent of plague and is presently dreaded for its potential use in Bio-terrorism (Inglesby et al. 2000). It is believed that this pathogen has evolved from *Yersinia pseudotuberculosis* which mainly causes gastritis (Achtman et al. 1999). Both these pathogens are gram-negative bacilli belonging to the family enterobacteriaceae. *Y. pestis* is highly monomorphic species, and thus a number of studies have been done by identifying variable number of tandem repeats (VNTRs) of Yersinia for finding a marker to classify the different strains (Adair et al. 2000; Klevytska et al. 2001; Le Fleche et al. 2001; Pourcel et al. 2004). In most of these studies families of long repeats which are minisatellites were considered and surprisingly SSRs have not been analysed in detail. Availability of the complete genome sequences of different strains of *Yersinia pestis* (sixteen in numbers) and of *Yersinia pseudotuberculosis* (four in numbers) gave an opportunity to analyze genome-wide distribution, abundance and enrichment of perfect SSRs in these genomes, the results of which are reported in this chapter.

2.2 Methods

2.2.1 Identification of perfect SSRs from the whole genome sequences of Yersinia:

The complete list of each strain, accession number and their source of availability are given in **Table 2.1**. For every genome its randomized equivalents were also considered for calculating expected numbers of SSRs in order to investigate enrichment of SSRs. The randomized genome sequences were computed using a stochastic model as implemented in Genrand program (Mrazek 2006). This stochastic model generates non-coding regions on the basis of the first order Markov model (16 types of dinucleotide probabilities) and coding regions on the basis of periodic first order Markov model.

The genome sequences and their stochastically modelled equivalents were surveyed for perfect SSRs using SSRF (Sreenu et al. 2003b; Sreenu et al. 2007). SSRF, as described in the previous chapter, scans a given nucleotide sequence and extracts SSRs with motif sizes between 1 to 6bp having at least two repeat copy numbers. However, for the purpose of the present study only those SSR tracts which were longer than 5bp was considered as a means to consider only those having some potential to undergo mutations due to slipped strand mis-pairing and filtered out very small tracts which are essentially the “noise”. The program uses GenBank annotation files, “xxx.ffn” and “xxx.fna” (xxx = genome name) that contain complete genome DNA and exon boundary information respectively and records location of SSRs relative to protein coding regions. It also takes care of internal motif redundancy, i.e., a sequence of the type (AAAAGCAAAAGCAAAAGC) is represented as $(AAAAGC)_3$ and its internal AAAA repeat is not reported separately as $(A)_4$ tract.

Table 2.1: Yersinia genomes considered for simple sequence repeat analysis. YP = *Yersinia pestis* and YPTB = *Yersinia pseudotuberculosis*.

Strain name	RefSeq/Accession Number	ORF code	Reference/Sequencing center
YP_Angola	NC_010159	YpAngola_A	TIGR
YP_Antiqua	NC_008150	YPA	(Chain et al. 2006)
YP_biovar_Antiqua_str_B42003004	NZ_AAYU0000000	YpB4200300	TIGR
YP_biovar_Antiqua_str_E1979001	NC_005810	YpE1979001	TIGR
YP_biovar_Antiqua_str_UG05_045	NZ_AAYR0000000	YpUG05045	TIGR
YP_biovar_Mediaevails_str_91001	NC_005810	YP	(Song et al. 2004)
YP_biovar_Mediaevalis_str_K19730	NZ_AAYT0000000	YpK1973002	TIGR
YP_biovar_Orientalis_str_F1991016	NZ_ABAT0000000	YpF1991016	TIGR
YP_biovar_Orientalis_str_IP275	NZ_AAOS0000000	YPIP275	TIGR
YP_biovar_Orientalis_str_MG05	NZ_AAYS0000000	YpMG05102	TIGR
YP_CA88_4125	NZ_ABCD0000000	YPE	ERIC
YP_CO92	NC_003143	YPO	(Parkhill et al. 2001)
YP_FV_1	NZ_AAUB0000000	YpesF	TGen
YP_KIM	NC_004088	Y	(Deng et al. 2002)
YP_Nepal516	NC_008149	YPN	(Chain et al. 2006)
YP_Pestoides_F	NC_009381	YPDSF	DOE
YPTB_IP_31758	NC_009708	YpsIP31758	(Eppinger et al. 2007)
YPTB_IP32953	NC_006155	YPTB	(Chain et al. 2004)
YPTB_PB1	NC_010634	YPTS	DOE
YPTB_YPIII	NC_010465	YPK	DOE

2.2.2. Enrichment and its significance

Enrichment indicates over/under representation of SSRs of a given type and is calculated as the ratio of observed (Obs) to the expected (Exps) number of SSRs found in a genome. The number of a given SSR found in a genome sequence and the number of the same SSR found in a randomized sample are referred to as observed number and expected number (average of hundred randomizations) respectively. When enrichment of SSRs of a particular type is >1.0 (when Obs>Exps) then that type of SSR is referred to as over-represented and when enrichment is <1.0 (when Obs<Exps) the SSR is referred to as under-represented. The statistical significance of over-representation / under-representation is calculated using student's t-test (Press et al. 1992) and also by calculating Z-score which is as given below.

$$Z = \frac{Observed - Expected}{\sqrt{Expected}}$$

2.3 Results

2.3.1 Genome-wide distribution and abundance of SSRs

All the yersinia genomes show very similar genome-wide distribution and abundance of perfect SSRs (here after referred to as SSRs unless otherwise mentioned). Each genome harbours as many as 90,000 SSRs (as mentioned SSRs >5bp was considered) occurring with a mean frequency of one SSR per 50bp. **Table 2.2** gives the frequencies of SSRs found in some of the representative strains of *Yersinia pestis* and *Y. pseudotuberculosis*. Among the SSRs those formed by tri nucleotide repeats are the most abundant types with a mean frequency of 14 tracts per kb and those formed by hexa nucleotide repeats are the least abundant types with a mean frequency of 0.4 SSR tracts per kb **(Table 2.2)**. In any genome all SSRs put together constitute to about 14% of its total number of base pairs. A large majority (~90%) of the SSRs are short tracts (<8bp) indicating that tract expansions have not been favored in these genomes. **Figure 2.1** shows tract density profiles (the number of SSRs per ten thousand bases) of five genomes of *Y. pestis* and one genome of *Y. pseudotuberculosis*. The tract densities vary in the range of 150-300 tracts/10kb with mean values of ~210 bp/10kb. The profiles indicate non-uniform distribution of the SSRs along the genomes resulting in an arrangement of regions abundant in SSRs followed by the regions with poor presence of SSRs. Regions abundant in SSRs (> mean + 2SD) are expected to be more prone to mutations involving SSRs as compared to SSR poor regions. Between coding and non-coding regions of the genomes non-coding regions (average density = 40 SSRs/10kbp) were found to be much denser with SSRs than coding regions (average density= 14 SSRs/10kbp).

Table 2.2: Number of different types of SSRs found in Yersinia genomes.

Repeat type	YPA	YP	YPO	Y	YPN	YPTB	RPK
Mono	7119	7012	7080	7012	6905	7363	1.5
Di	6729	6583	6651	6584	6443	6790	1.4
Tri	65533	64208	64911	64150	63078	66815	14.0
Tetra	13797	13445	13653	13471	13202	13997	3.0
Penta	3697	3518	3596	3532	3443	3677	0.7
Hexa	1735	1690	1702	1685	1678	1795	0.4
Total	98610	96456	97593	96434	94749	100437	21.0

Footnotes: (RPK = repeat per kb. YP = *Yersinia pestis* 91001; YPO= *Yersinia pestis* CO92; YPN = *Yersinia pestis* Nepal516; Y=*Yersinia pestis* KIM; YPTB=*Yersinia pseudotuberculosis*IP32953*;* YPA=*Yersinia pestis* Antiqua)

2.3.2 Enrichment of SSRs

From **Table 2.3** it can be seen that SSRs of mono, tri, tetra and hexa nucleotide motifs are overrepresented (P-value 0.0001 and Z score >4) whereas SSRs of di and penta nucleotide motifs are underrepresented in all the genomes. Enrichment of SSRs of tri and hexa motifs is well known in genomes. However, enrichment of mono and tetra SSRs may appear surprising at the first glance given the fact that enrichment of these tracts increases chances of frame-shift mutations. When their distribution in coding and non-coding regions was analysed it was found that these tracts are enriched in non-coding regions (please see **Table 2.4 b**) and not in coding regions.

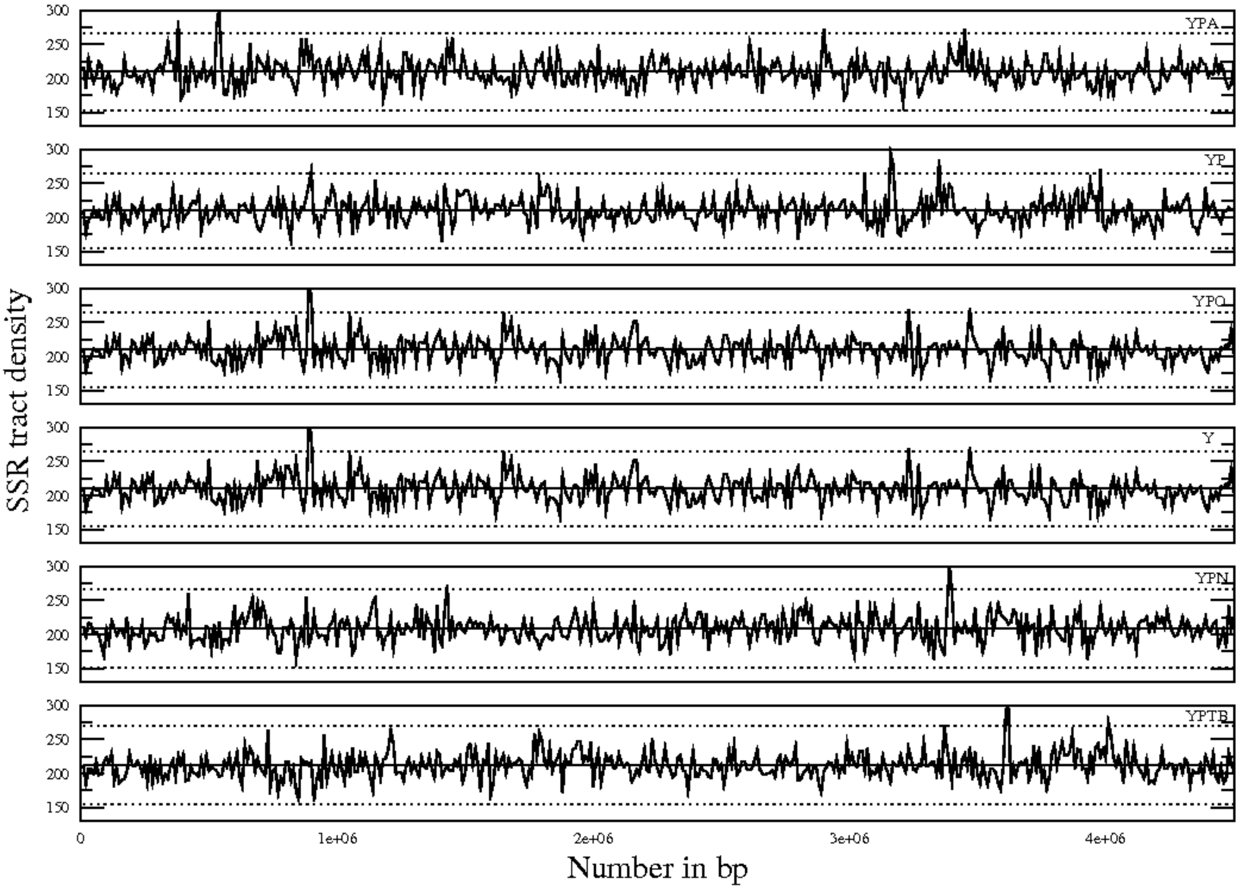


Figure 2.1: SSR density profile in Yersinia genomes. The SSR tract density profiles of YPA, YP, YPO, Y, YPN and YPTB. The average of the observed SSR tracts per 10 kb is shown as a continuous thin line. The two dotted lines shown above and below the thin line correspond to the numbers at the levels of two standard deviations over and below the mean number. The number of peaks having heights greater and less than two standard deviations from the mean number are considered as SSRs abundant and poor regions respectively.

Table 2.3: The observed and expected numbers of SSRs found in Yersinia genomes. Overrepresentation/underrepresentation of different types of SSRs in Yersinia genomes is also given. Part “A” gives the observed (the first value) and the expected numbers (second value in parenthesis) of SSRs. The “+” sign and “-” represent statistical significance of overrepresentation and underrepresentation respectively. In part “B” enrichment which is the ratio of the observed to the expected number of SSR’s is given. The significance is calculated as a Z score given within parentheses. The Z score more than “+4” and less than “-4” signifies the overrepresentation and underrepresentation respectively.

(A)						
Repeat Type	**YPA**	**YP**	**YPO**	**Y**	**YPN**	**YPTB**
Mono	7119 + (6819)	7012 + (6687)	7080 + (6740)	7012 + (6660)	6905 + (6689)	7363 + (6926)
Di	6729 - (8280)	6583 - (8154)	6651 - (8219)	6584 - (8132)	6443 - (7950)	6790 - (8356)
Tri	65533 + (63774)	64208 + (61698)	64911 + (62567)	64150 + (61990)	63078 + (61467)	66815 + (64181)
Tetra	13797 + (13238)	13445 + (12983)	13653 + (13151)	13471 + (12933)	13202 + (12710)	13997 + (13350)
Penta	3697 (3690)	3518 (3610)	3596 - (3687)	3532 (3609)	3443 - (3565)	3677 (3736)
Hexa	1735 + (1636)	1690 + (1574)	1702 + (1620)	1685 + (1596)	1678 + (1582)	1795 + (1655)
(B)						
Mono	1.04 (3.64)	1.05 (3.97)	1.05 (4.14)	1.05 (4.31)	1.03 (2.64)	1.06 (5.25)
Di	0.81 (-17.05)	0.81 (-17.40)	0.81 (-17.29)	0.81 (-17.17)	0.81 (-16.91)	0.81 (-17.13)
Tri	1.03 (6.96)	1.04 (10.11)	1.04 (9.37)	1.03 (8.68)	1.03 (6.50)	1.04 (10.40)
Tetra	1.04 (4.86)	1.04 (4.05)	1.04 (4.38)	1.04 (4.73)	1.04 (4.36)	1.05 (5.60)
Penta	1.00 (0.11)	0.97 (-1.53)	0.98 (-1.50)	0.98 (-1.29)	0.97 (-2.04)	0.98 (-0.97)
Hexa	1.06 (2.44)	1.07 (2.91)	1.05 (2.05)	1.06 (2.24)	1.06 (2.42)	1.08 (3.44)

The relationship between repeat copy number in SSRs and their enrichment values was also examined and the results obtained are given in (**Table 2.4 (a)**). Among the mono tracts the repeats of six and seven times are significantly over-represented in all the genomes. The longer repeats of mono (>7bp) are significantly under-represented. Among the remaining types of SSRs two and three times repeats of tri SSRs and two times repeats of tetra SSRs are overrepresented whereas the others are under-represented.

Further investigations revealed that among the mono SSRs of six and seven times repeats, A|T type repeats are significantly enriched as compared to G|C repeats (data shown only for one of the strains in **Table 2.4b**). The A|T tracts of length 6 bp are overrepresented in coding as well as in non-coding regions but those of 7 bp in length are over-represented only in the non-coding regions.

The over-representation of A|T repeats is perhaps a consequence of their importance in transcription termination. It is well known that in the bacterial genomes 50% of the genes use A|T repeat tracts which can form hairpin structure in rho-independent transcription termination (Yeung et al. 1998; Lesnik et al. 2001). The highly significant under-representation of G|C repeats in coding regions is, perhaps, due to selection against these tracts which are highly polymorphic in nature (Harfe and Jinks-Robertson 2000; Gragg, Harfe, and Jinks-Robertson 2002)

Table 2.4 (a): Observed and expected numbers of SSRs of different copy numbers found in plague causing (YPO), non plague causing (YP) and ancestral (YPTB) Yersinia genomes.

CNRT	YPO		YP		YPTB	
	Obs (Exp)	E=(Obs/Exp)	Obs (Exp)	E=(Obs/Exp)	Obs (Exp)	E=(Obs/Exp)
Mono6	5123 + (4554)	1.13 (8.44)	5054 + (4515)	1.12 (8.02)	5280 + (4679)	1.13 (8.79)
Mono7	1536 + (1425)	1.08 (2.93)	1524 + (1419)	1.07 (2.78)	1611 + (1453)	1.11 (4.14)
Mono8	355 - (491)	0.72 (-6.13)	368 - (485)	0.76 (-5.32)	400 - (520)	0.77 (-5.24)
Mono9	56 - (168)	0.33 (-8.65)	58 - (165)	0.35 (-8.35)	63 - (166)	0.38 (-7.98)
Mono≥10	10 - (102)	0.10 (-9.11)	8 - (103)	0.08 (-9.35)	9 - (110)	0.08 (-9.60)
Di3	6259 - (7621)	0.82 (-15.60)	6201 - (7561)	0.82 (-5.64)	6415 - (7750)	0.83 (-15.17)
Di4	369 - (550)	0.67 (-7.70)	359 - (544)	0.66 (-7.94)	348 - (556)	0.63 (-8.84)
Di≥5	23 - (48)	0.48 (-3.61)	23 - (49)	0.47 (-3.68)	27 - (49)	0.55 (-3.19)
Tri2	62835 + (60821)	1.03 (8.17)	62155 + (59970)	1.04 (8.92)	64638 + (62351)	1.04 (9.16)
Tri3	2011 + (1679)	1.20 (8.11)	1986 + (1665)	1.19 (7.88)	2106 + (1760)	1.20 (8.25)
Tri≥4	65 (67, 7.11))	0.97 (-0.28)	67 (63)	1.06 (0.48)	71 (70)	1.02 (0.13)
Tetra2	13598 + (13095)	1.04 (4.39)	13390 + (12925)	1.04 (4.09)	13936 + (13294)	1.05 (5.56)
Tetra≥3	55 (56, 5.28))	0.99 (-0.08)	55 (58)	0.94 (-0.44)	61 (56)	1.09 (0.67)
Penta2	3591 - (3682)	0.98 (-1.50)	3515 (3606)	0.97 (-1.51)	3667 (3731)	0.98 (-1.05)
Penta>2	5 (5, 1.40))	0.96 (-0.09)	3 (5)	0.65 (-0.75)	10 + (5)	1.96 (2.17)
Hexa2	1691 + (1618)	1.05 (1.82)	1680 + (1573)	1.07 (2.70)	1781 + (1654)	1.08 (3.13)
Hexa>2	11 + (2)	6.11 (6.86)	10 + (2)	6.67 (6.94)	14 + (1)	10.77 (11.14)

Footnotes: "+" and "-" indicate significance of overrepresentation and underrepresentation of the tracts according to student's t-test. The enrichment value E (Obs/Exp) along with the Z score (in parentheses) is also given (CNRT = copy number of different repeat types).

Table 2.4 (b): Observed and expected numbers of SSRs of different repeat copy numbers found in coding and non-coding regions of *Yersinia pestis CO92* genome.

Yersinia pestis CO92				
Repeat	Coding regions		Non-coding regions	
type	Obs (Exp) t-test	E (Z score)	Obs (Exp) t-test	E (Z score)
Monoat6	2668 + (2271)	1.17 (8.33)	1422 + (1035)	1.37 (12.04)
Monoat7	669 - (722)	0.93 (-1.98)	552 + (370)	1.49 (9.46)
Monoat8	133 - (246)	0.54 (-7.19)	155 + (140)	1.10 (1.22)
Monoat9	13 - (81)	0.16 (-7.54)	26 - (54)	0.49 (-3.77)
Monoat≥1	0 - (44)	0.00 (-6.62)	3 - (37)	0.08 (-5.55)
Monogc6	787 - (980)	0.80 (-6.15)	212 (225)	0.94 (-0.88)
Monogc7	237 - (252)	0.94 (-0.91)	71 + (62)	1.14 (1.13)
Monogc8	56 - (78)	0.72 (-2.46)	10 - (18)	0.56 (-1.87)
Monogc9	9 - (24)	0.38 (-3.02)	7 (6)	1.21 (0.50)
Mgc≥10	2 - (16)	0.13 (-3.49)	5 + (2)	2.38 (2.00)
Di3	4505 - (5702)	0.79 (-15.85)	1712 - (1851)	0.92 (-3.24)
Di4	243 - (392)	0.62 (-7.52)	124 - (149)	0.83 (-2.05)
Di≥5	7 - (33)	0.21 (-4.57)	16 (14)	1.18 (0.65)
Tri2	50492 + (49932)	1.01 (2.51)	11843 + (10450)	1.13 (13.63)
Tri3	1655 + (1424)	1.16 (6.13)	336 + (232)	1.45 (6.80)
Tri≥4	60 (59)	1.02 (0.16)	5 (8, 3.26))	0.67 (-0.91)
Tetra2	9930 (9946)	1.00 (-0.16)	3492 + (3000)	1.16 (8.99)
Tetra≥3	29 (34)	0.85 (-0.87)	23 (19)	1.19 (0.82)
Penta2	2685 (2719)	0.99 (-0.65)	855 - (903)	0.95 (-1.61)
Penta>2	2 (3)	0.65 (-0.62)	3 (2)	1.67 (0.89)
Hexa2	1378 + (1338)	1.03 (1.09)	300 + (250)	1.20 (3.15)
Hexa>2	4 + (1)	2.86 (2.20)	6 + (0)	20.00 (10.41)

2.3.3 Propensities of SSRs in different protein functions

As revealed in the aforementioned section coding regions are not enriched by SSRs, however, it is still interesting to find propensity (relative preferences) of SSRs in various protein functions, particularly for mono, di, tetra and penta repeats as these can induce frameshift mutations. The ORFs were organized based on COG functions and propensities of SSRs were calculated in these groups as follows.

Propensity of SSRs in a COG group:

$$PSi = \frac{\frac{ni}{N} * 100}{\frac{mi}{M} * 100}$$

Here, ni = Number of SSRs in bps found in the ith COG functional group, N = Number of SSRs in bps found in all the ORFs assigned to COG functional groups, m = number of bps found in the ith COG functional group, M = Total number of nucleotides in all the ORFs assigned to COG functional groups.

Propensity values greater than 1.0 (high) and less than 1.0 (low) respectively indicate preference and non-preference of SSRs for a particular COG function (propensity equal to 1.0 indicates indifference). **Figure 2.2** shows the propensity values of SSRs for various COG functions. It is interesting to note that all the SSRs of 6bp in length although show propensities > 1.0 for certain COG functions; their values are not as pronounced as that of SSRs longer than 6bp. The long SSRs clearly show high propensities for the proteins involved in replication, recombination and repair, defense mechanism, cell wall/membrane/envelope

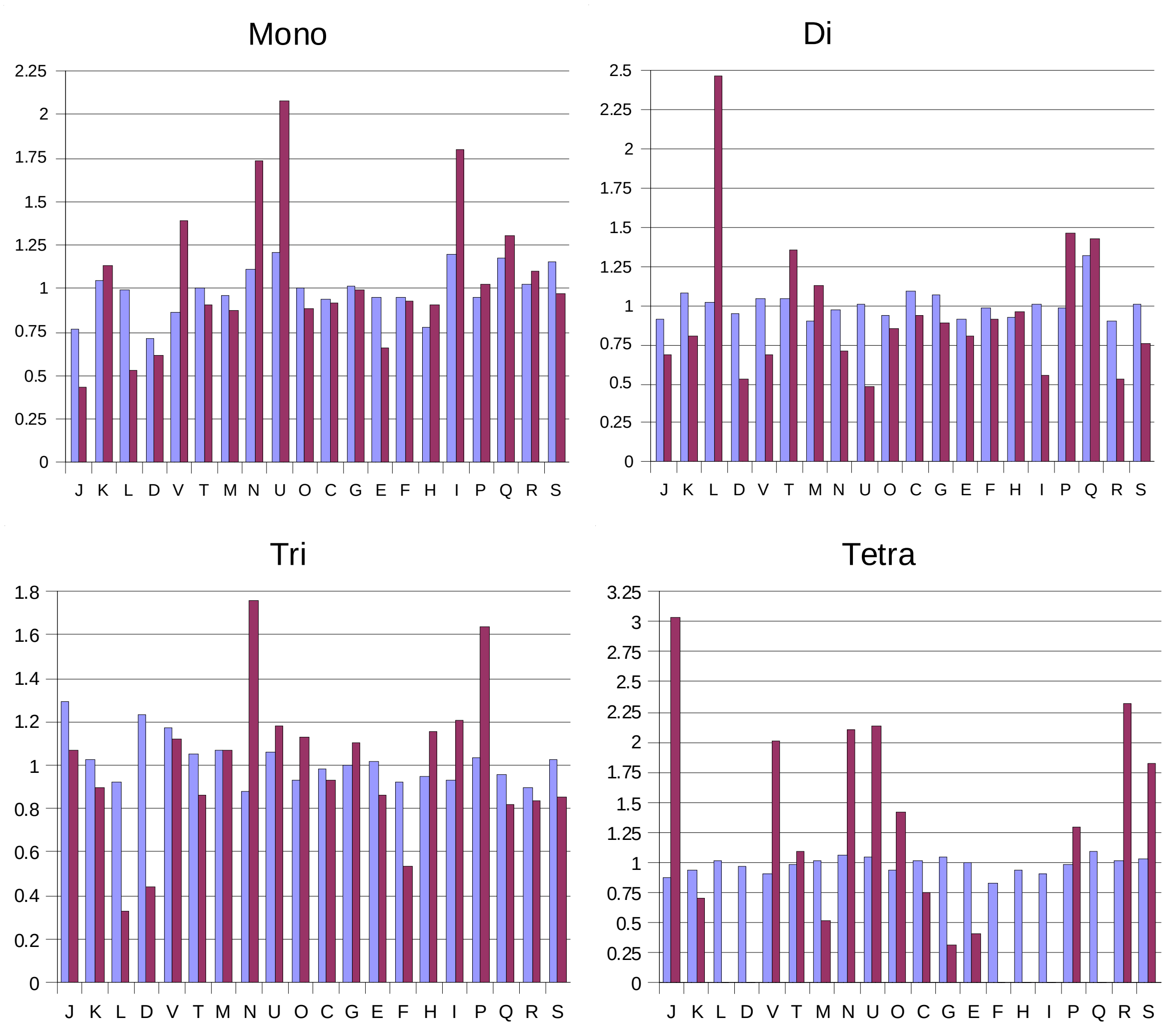


Figure 2.2: Average propensities for mono, di, tri and tetra SSRs in the ORFs coding for proteins belonging to different COG functional groups. SSRs of 6bp and more than 6bp are shown as blue and magenta bars respectively. In case of tetra blue bar is <=8bp and magenta >8bp.

Key for COG functional groups (Tatusov et al. 2001): J = Translation, ribosomal structure and biogenesis; K = Transcription; L = Replication, recombination and repair; B = Chromatin structure and dynamics; D = Cell cycle control, cell division, chromosome partitioning; Y = Nuclear structure; V = Defense mechanisms; T = Signal transduction mechanisms; M = Cell wall/membrane/envelope biogenesis; N = Cell motility; Z = Cytoskeleton; W = Extracellular structures; U = Intracellular trafficking, secretion, and vesicular transport; O = Posttranslational modification, protein turnover, chaperones; C = Energy production and conversion; G = Carbohydrate transport and metabolism; E = Amino acid transport and metabolism; F = Nucleotide transport and metabolism; H = Coenzyme transport and metabolism; I = Lipid transport and metabolism; P = Inorganic ion transport and metabolism; Q = Secondary metabolites biosynthesis, transport and catabolism; R = General function prediction only; and S = hypothetical protein.

biogenesis, cell motility, intracellular trafficking, secretion and vesicular transport, lipid transport and metabolism, inorganic ion transport and metabolism, and low propensities for ORFs encoding proteins involved in cell division, energy production and conversion, carbohydrate transport and metabolism, amino acid transport and metabolism, nucleotide transport and metabolism, coenzyme transport and metabolism, and secondary metabolites biosynthesis, transport and catabolism. This shows that SSR distribution in the coding regions has been skewed towards certain specific functions.

2.3.4 Intragenic position-specific distribution of SSRs

It is interesting to analyze position specific distribution of SSRs within genes as this gives us an indication of potential vulnerability of genes due to SSR mutations. Over-representation of SSRs in the beginning, middle or at the end of a gene would indicate varying extents of vulnerabilities with regard to structure and function of the genes. Positional distribution of SSRs along the length of genes was therefore analyzed (viz., ORFs). For the purpose of this study a gene was visualized as composed of three parts: 5’-part (0-33%), middle part (34-66%) and 3’-part (67-100%) and the number of SSRs found in each of these parts in all the ORFs was calculated. Among the SSRs the most abundantly found trinucleotide SSRs show more or less equal distribution whereas the other SSRs show uneven distribution in the three parts of the ORFs which again varies from function to function (**Figure 2.3**). For example, SSRs are mostly harbored at the 3’ side (i.e., III part) of the ORFs encoding proteins belonging to genes involved in replication, recombination and repair. As a consequence SSR mutations potentially have least effects on functions of these groups of genes which are mainly responsible for the genome stability.

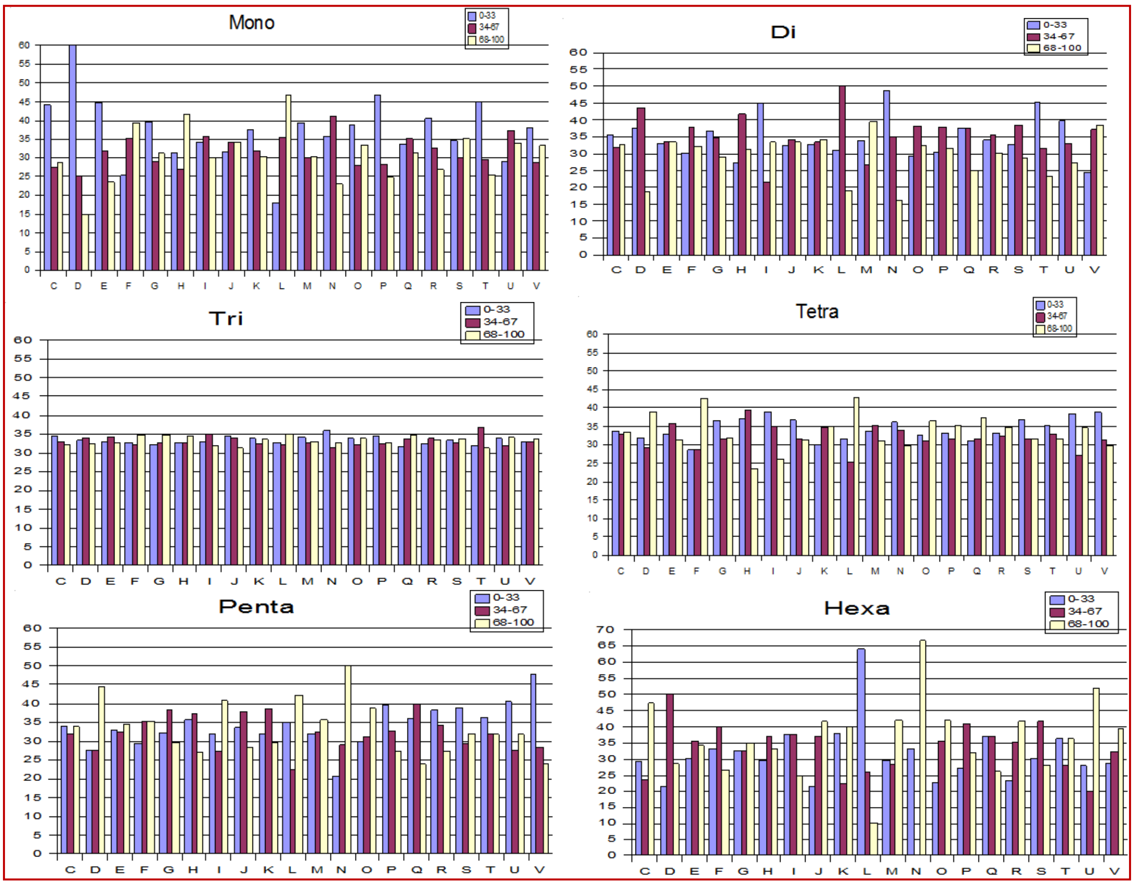


Figure 2.3: The number of SSRs found in the 5', middle and 3' parts of genes. The tri SSRs are equally distributed in the three parts whereas the other types of SSRs show biases towards 5' or the 3' parts of genes. Key for COG functional group please see the legend of Figure 2.2.

It is interesting to see relatively high abundant occurrence of SSRs in 5’ part of most of the protein function groups. SSR mutations in these ORFs can lead to shifts in their reading frames which in turn can lead to coding of different proteins with different functions or premature terminations. The first possibility viz., encoding of a different protein has not been reported in any organism. The second possibility viz., premature terminations have been reported in Neisseria and Mycobateria. It is in fact a well known mechanism of switching off genes by SSRs. Since SSR mutations are reversible, genes also switch ‘off’ and ‘on’ (van Belkum et al. 1998; Moxon, Bayliss, and Hood 2006). A 5' SSR mutation can also generate small sized peptides which may be less deleterious for the organism compared to larger peptides. It has been argued that the placement of the mononucleotide repeats towards the 5' end of the gene is for the active removal of the gene (van Passel and Ochman 2007).The SSR mutations at 3' side may help in the fusion of the genes (Sreenu et al. 2006).

2.4 Discussion

The study as reported in this chapter has revealed that perfect SSRs are found in plenty in Yersinia genomes. Interestingly any noticeable difference in SSRs with regard to their distribution, abundance and enrichment among the various strains of Yersinia was not observed. All the genomes harbor short SSR tracts and long SSR tracts are sparingly seen indicating a strong disfavor for SSR tract expansions in these genomes. All the genomes are characterized by the presence of several SSR abundant regions as well as SSR deficient regions. The regions dense with SSRs have potential to undergo slippage during replication leading to mutated SSRs and hence such regions act as resources of genetic variability. SSRs having potential to shift reading frames are sparser in coding regions compared to non-coding regions. In general, SSRs are not enriched in coding regions of Yersinia genomes. Nonetheless available SSRs in particular those having potential to cause frame-shift mutations show preferential distribution towards certain protein functions. High preferences are shown for ORFs encoding proteins which are either directly or indirectly involved in defense mechanisms, interactions with host environments (extracellular structure), secretion and cell motility. These proteins face most of the insults of the host immune defence mechanism and the SSRs offers the required functional plasticity required for the survival of the pathogens. On the other hand, SSRs show low preferences for the ORFs which code for protein function in general metabolism. This seems to suggest how SSRs have evolved in a biased manner so as to help pathogens to adapt to different hosts.

Furthermore abundance of SSRs in the 5’ regions of some of the ORFs is interesting. This indicates potential vulnerability of these genes for undergoing on-off switching. In addition to this the mutation generated towards the 5’ side of a gene may translate into smaller non functional peptides which are less deleterious than the longer peptides.

Although the genomes analysed in the present study are highly similar to each other (≥98% identity in the homologous regions) they vary in their disease causing abilities. Earlier it was believed that the extent of virulence is related to the number of pathogenic plasmids in them. However, after sequencing the non-pathogenic *Yersinia pestis* 91001 (Song et al. 2004) strain which contains all the pathogenic plasmids, it is assumed that some other factors may perhaps be responsible for their virulence (Revell and Miller 2001; Song et al. 2004). As indicated in the present study SSRs may play some role in rendering virulence to the pathogens.

The rise and fall of plague epidemic remains a mystery (Wren 2003) and the factors responsible for the onset and subsiding of plague epidemic is unknown. In this light, it can be argued that SSRs due to their reversible mutations have some role towards the rise of virulent forms causing rapid spread of disease followed by rise of less virulent forms causing disease to disappear. In general, SSRs have an important role in generating the required genetic variation in pathogens for selection to act upon as a means to increase their potential to adapt to changing environments; to change their life style as well as to increase/decrease their virulent nature.

2.5 Summary

The work reported in this chapter pertains to a detailed analysis of SSRs in the whole genome sequences of different strains of *Y. pestis* and *Y. pseudotuberculosis.* The SSRs were extracted from the genomes and were analysed with regard to their genome-wide and gene-wise distribution, abundance and enrichment. Our studies reveal similar SSR distributions in both Y*. pestis* and *Y. pseudotuberculosis* strains. In general, coding regions of all the Yersinia genomes are underrepresented in SSRs except the A|T motifs of repeat count six and trinucleotide repeat. However within coding regions SSRs show preferences for certain protein functions. Further analysis on position-specific distribution of SSRs within genes indicated higher abundance of SSRs in termini parts viz., 5’ and 3’ parts than the middle part.

Chapter 3

PSSRFinder (**P**olymorphic **S**imple **S**equence **R**epeat **Finder**): A tool for comparing Equivalent Simple Sequence Repeats across Multiple Related Genomes

3.1 Introduction

A review of literature as presented in the introductory chapter reveals that simple sequence repeats form an interesting group of genomic features owing to their property of length polymorphism arising out of insertions and deletions of their repeat units which occur mostly as a consequence of slipped-strand mispairing during DNA replication. The usefulness of SSR lies in their polymorphic ability and hence it is interesting to first identify polymorphic SSRs in the genome in order to decipher any role played by them in the fitness of the organism. The experimental method for finding polymorphic simple sequence repeats is time consuming and cumbersome and involves a number of steps such as, identification of SSRs in genome, designing primer, PCR amplification from genomic DNA of number of strains of particular species and finally running on a gel or DNA sequencing to know the tract length of amplicon. As compared to the experimental approach, in a computer based approach which is much easier and faster SSR length variation could be identified by comparing equivalent SSRs across different strains of species. Availability of complete genome sequences of multiple strains of almost all the species in the public domain has created an opportunity to take the computational approach. As discussed in the introductory chapter, a number of computational tools are available for the extraction of SSRs from complete genome DNA sequences. All these tools focus on extraction of repeats from individual genomes. Though a number of tools are available for the extraction of SSRs from the complete genomes, none of them can automatically compare equivalent SSRs and report SSR length variation across the genomes compared. Therefore a tool called PSSRFinder (polymorphic simple sequence repeat finder) which compares the equivalent SSRs across multiple genomes was developed, the details of which are reported in this chapter.

3.2 Polymorphic Simple sequence repeats Finder (PSSRFinder)

PSSRFinder is a fully automated tool for comparing the equivalent simple sequence repeats across multiple identical genomes. It reports the polymorphic perfect simple sequence repeats of 1-6 base pair in length across any number of strains in a species at the whole genome level. The schematic representation for the extraction of polymorphic SSRs (PSSRs) is shown in **Figure 3.1**. Internally the PSSRfinder uses a number of programs (including those developed by the author), for example, comparing the two equivalent regions in different genomes using BLAST, extraction of SSRs in complete genomes is done by SSRF, and parsing of coding and non-coding BLAST output file by C_PSSRF and N_PSSRF respectively. The details of this program are given below.

1. Identification of SSRs from the given genomes using SSRF (Sreenu et al. 2003b) which reports SSR motif, motif repeat counts, co-ordinate of SSR tract in the genome and its location relative to coding and non-coding regions.

2. Identification of PSSRs is done by comparing tract lengths of equivalent SSRs found in all the given genomes. Here equivalent SSRs means the SSRs with their flanking DNA sequences are identical in all the given genomes. The conserved segments harboring equivalent SSRs can be part of coding or non-coding regions in all the genomes or coding regions in some genomes and non-coding regions in the other genomes. If in all the genomes the conserved segments are found in non-coding regions then the corresponding SSR is referred to as non-coding SSR. If it is found as a part of a coding region even in one of the genomes then the

SSR is referred to as coding SSR.

3. Equivalent SSRs along with their conserved flanking segments are found using BLAST searches which are carried out pair-wise with the following set of parameters: E-value ≤ 0.001; X drop off value for final gapped alignment = 200; and repeat masking filter = off. In a BLAST search every ORF or inter-genic region from one genome is taken as the query and searched against the whole genome sequence of another genome taken as the database. The BLAST output files from the searches of ORFs and inter-genic regions are separately parsed using the programs C_PSSRF (Coding polymorphic simple sequence finder) and NC_PDDRF (Non-coding simple sequence finder) respectively.

From each pair-wise BLAST output these programs pick up the first hit provided the query and subject are ~100% identical to each other but with aligned SSRs they have different lengths. Such SSR is annotated as PSSR. This is repeated for all pair-wise BLAST outputs and a concise report is generated for every PSSR as mentioned below.

3.3. The output of PSSRFinder:

As already mentioned PSSRFinder reports a PSSR as non-coding PSSR if it is present in the non-coding region of all the compared genomes; otherwise it is reported as coding PSSR. The sample output of coding and non-coding PSSRs extracted from the 22 strains of E. coli are shown in **Figure 3.2** and **3.3**. In both the cases (coding as well as non-coding) PSSRFinder reports PSSR motif, repeat count, co-ordinate, mutation point, and alignment of a 15bp region upstream and downstream from the mutation point in each genome. In case of coding PSSRs it further reports the information of the gene getting affected by the polymorphism of SSR. In the case of non-coding PSSRs it gives information about the upstream and downstream genes.

PSSRFinder outputs different information in case of coding and non-coding PSSRs. In the case of coding PSSRs it reports SSR co-ordinate, mutation point 15bp upstream and downstream DNA sequence from the point of mutation, start and end of ORF harboring the PSSR, strand location of ORF, protein length, protein ID, ORF name and its function in each genome. Whereas in case of non-coding PSSRs it reports the SSR co-ordinate, mutation point, 15bp upstream and downstream of the DNA sequence from the point of mutation, upstream and downstream ORF information (start and end of ORF, strand location of ORF, protein length, protein ID and ORF name) from the point of mutation in each genome.

3.4. Summary

In this chapter I gave the details of the development of a tool called PSSRFinder. This tool compares equivalent SSR across multiple strains of species. As polymorphic SSRs have been considered very important for bacteria this tool would be very helpful to study the evolution of SSRs.

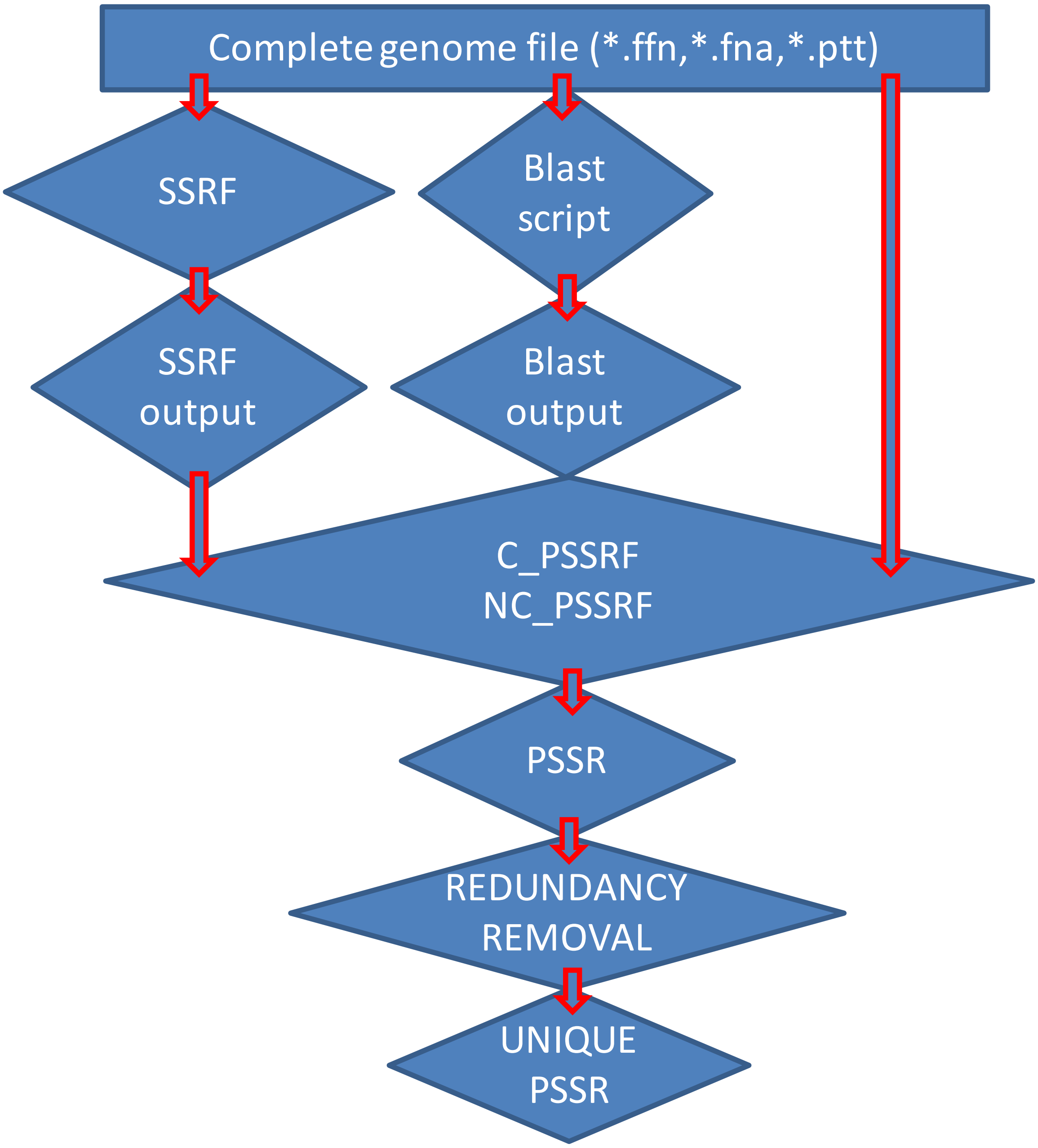
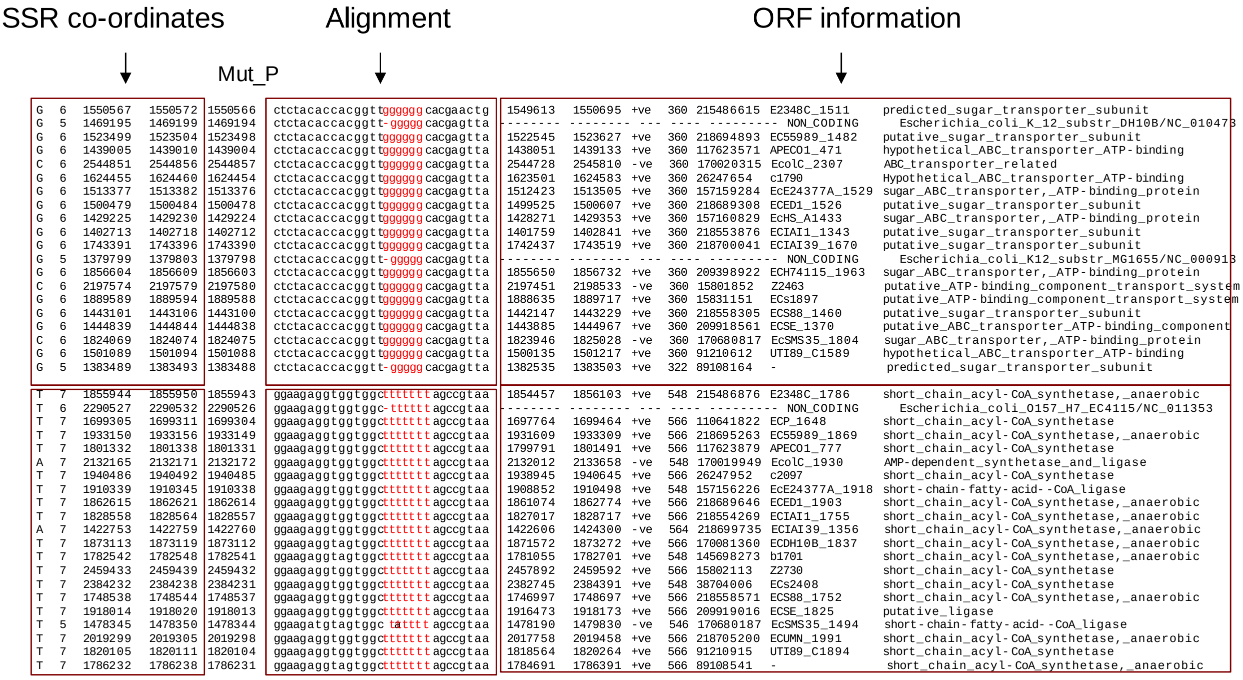


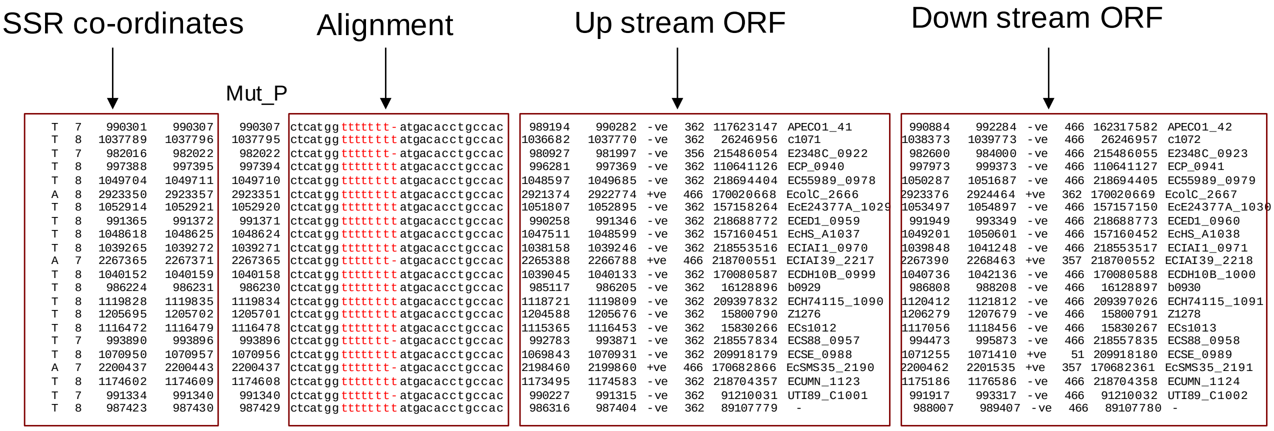
Figure 3.1: Schematic representation of PSSRFinder.



```
G  6  1550567  1550572  1550566  ctctacaccacggttgggggg cacgaactg  1549613  1550695  +ve  360  215486615  E2348C_1511       predicted_sugar_transporter_subunit
G  5  1469195  1469199  1469194  ctctacaccacggtt-ggggg cacgagtta  --------  --------  ---  ----  ---------  NON_CODING   Escherichia_coli_K_12_substr_DH10B/NC_010473
G  6  1523499  1523504  1523498  ctctacaccacggttgggggg cacgagtta  1522545  1523627  +ve  360  218694893  EC55989_1482      putative_sugar_transporter_subunit
G  6  1439005  1439010  1439004  ctctacaccacggttgggggg cacgagtta  1438051  1439133  +ve  360  117623571  APECO1_471        hypothetical_ABC_transporter_ATP-binding
C  6  2544851  2544856  2544857  ctctacaccacggttgggggg cacgagtta  2544728  2545810  -ve  360  170020315  EcolC_2307        ABC_transporter_related
G  6  1624455  1624460  1624454  ctctacaccacggttgggggg cacgagtta  1623501  1624583  +ve  360  26247654   c1790             Hypothetical_ABC_transporter_ATP-binding
G  6  1513377  1513382  1513376  ctctacaccacggttgggggg cacgagtta  1512423  1513505  +ve  360  157159284  EcE24377A_1529    sugar_ABC_transporter,_ATP-binding_protein
G  6  1500479  1500484  1500478  ctctacaccacggttgggggg cacgagtta  1499525  1500607  +ve  360  218689308  ECED1_1526        putative_sugar_transporter_subunit
G  6  1429225  1429230  1429224  ctctacaccacggttgggggg cacgagtta  1428271  1429353  +ve  360  157160829  EcHS_A1433        sugar_ABC_transporter,_ATP-binding_protein
G  6  1402713  1402718  1402712  ctctacaccacggttgggggg cacgagtta  1401759  1402841  +ve  360  218553876  ECIAI1_1343       putative_sugar_transporter_subunit
G  6  1743391  1743396  1743390  ctctacaccacggttgggggg cacgagtta  1742437  1743519  +ve  360  218700041  ECIAI39_1670      putative_sugar_transporter_subunit
G  5  1379799  1379803  1379798  ctctacaccacggtt-ggggg cacgagtta  --------  --------  ---  ----  ---------  NON_CODING   Escherichia_coli_K12_substr_MG1655/NC_000913
G  6  1856604  1856609  1856603  ctctacaccacggttgggggg cacgagtta  1855650  1856732  +ve  360  209398922  ECH74115_1963     sugar_ABC_transporter,_ATP-binding_protein
C  6  2197574  2197579  2197580  ctctacaccacggttgggggg cacgagtta  2197451  2198533  -ve  360  15801852   Z2463             putative_ATP-binding_component_transport_system
G  6  1889589  1889594  1889588  ctctacaccacggttgggggg cacgagtta  1888635  1889717  +ve  360  15831151   ECs1897           putative_ATP-binding_component_transport_system
G  6  1443101  1443106  1443100  ctctacaccacggttgggggg cacgagtta  1442147  1443229  +ve  360  218558305  ECS88_1460        putative_sugar_transporter_subunit
G  6  1444839  1444844  1444838  ctctacaccacggttgggggg cacgagtta  1443885  1444967  +ve  360  209918561  ECSE_1370         putative_ABC_transporter_ATP-binding_component
C  6  1824069  1824074  1824075  ctctacaccacggttgggggg cacgagtta  1823946  1825028  -ve  360  170680817  EcSMS35_1804      sugar_ABC_transporter,_ATP-binding_protein
G  6  1501089  1501094  1501088  ctctacaccacggttgggggg cacgagtta  1500135  1501217  +ve  360  91210612   UTI89_C1589       hypothetical_ABC_transporter_ATP-binding
G  5  1383489  1383493  1383488  ctctacaccacggtt-ggggg cacgagtta  1382535  1383503  +ve  322  89108164   -                 predicted_sugar_transporter_subunit

T  7  1855944  1855950  1855943  ggaagaggtggtggcttttttt agccgtaa  1854457  1856103  +ve  548  215486876  E2348C_1786       short_chain_acyl-CoA_synthetase,_anaerobic
T  6  2290527  2290532  2290526  ggaagaggtggtggc-tttttt agccgtaa  --------  --------  ---  ----  ---------  NON_CODING   Escherichia_coli_O157_H7_EC4115/NC_011353
T  7  1699305  1699311  1699304  ggaagaggtggtggcttttttt agccgtaa  1697764  1699464  +ve  566  110641822  ECP_1648          short_chain_acyl-CoA_synthetase
T  7  1933150  1933156  1933149  ggaagaggtggtggcttttttt agccgtaa  1931609  1933309  +ve  566  218695263  EC55989_1869      short_chain_acyl-CoA_synthetase,_anaerobic
T  7  1801332  1801338  1801331  ggaagaggtggtggcttttttt agccgtaa  1799791  1801491  +ve  566  117623879  APECO1_777        short_chain_acyl-CoA_synthetase
A  7  2132165  2132171  2132172  ggaagaggtggtggcttttttt agccgtaa  2132012  2133658  -ve  548  170019949  EcolC_1930        AMP-dependent_synthetase_and_ligase
T  7  1940486  1940492  1940485  ggaagaggtggtggcttttttt agccgtaa  1938945  1940645  +ve  566  26247952   c2097             short_chain_acyl-CoA_synthetase
T  7  1910339  1910345  1910338  ggaagaggtggtggcttttttt agccgtaa  1908852  1910498  +ve  548  157156226  EcE24377A_1918    short-chain-fatty-acid--CoA_ligase
T  7  1862615  1862621  1862614  ggaagaggtggtggcttttttt agccgtaa  1861074  1862774  +ve  566  218689646  ECED1_1903        short_chain_acyl-CoA_synthetase,_anaerobic
T  7  1828558  1828564  1828557  ggaagaggtggtggcttttttt agccgtaa  1827017  1828717  +ve  566  218554269  ECIAI1_1755       short_chain_acyl-CoA_synthetase,_anaerobic
A  7  1422753  1422759  1422760  ggaagaggtggtggcttttttt agccgtaa  1422606  1424300  -ve  564  218699735  ECIAI39_1356      short_chain_acyl-CoA_synthetase,_anaerobic
T  7  1873113  1873119  1873112  ggaagaggtagtggcttttttt agccgtaa  1871572  1873272  +ve  566  170081360  ECDH10B_1837      short_chain_acyl-CoA_synthetase,_anaerobic
T  7  1782542  1782548  1782541  ggaagaggtagtggcttttttt agccgtaa  1781055  1782701  +ve  548  145698273  b1701             short_chain_acyl-CoA_synthetase,_anaerobic
T  7  2459433  2459439  2459432  ggaagaggtggtggcttttttt agccgtaa  2457892  2459592  +ve  566  15802113   Z2730             short_chain_acyl-CoA_synthetase
T  7  2384232  2384238  2384231  ggaagaggtggtggcttttttt agccgtaa  2382745  2384391  +ve  548  38704006   ECs2408           short_chain_acyl-CoA_synthetase
T  7  1748538  1748544  1748537  ggaagaggtggtggcttttttt agccgtaa  1746997  1748697  +ve  566  218558571  ECS88_1752        short_chain_acyl-CoA_synthetase,_anaerobic
T  7  1918014  1918020  1918013  ggaagaggtggtggcttttttt agccgtaa  1916473  1918173  +ve  566  209919016  ECSE_1825         putative_ligase
T  5  1478345  1478350  1478344  ggaagatgtagtggc tatttt agccgtaa  1478190  1479830  -ve  546  170680187  EcSMS35_1494      short-chain-fatty-acid--CoA_ligase
T  7  2019299  2019305  2019298  ggaagaggtggtggcttttttt agccgtaa  2017758  2019458  +ve  566  218705200  ECUMN_1991        short_chain_acyl-CoA_synthetase,_anaerobic
T  7  1820105  1820111  1820104  ggaagaggtggtggcttttttt agccgtaa  1818564  1820264  +ve  566  91210915   UTI89_C1894       short_chain_acyl-CoA_synthetase
T  7  1786232  1786238  1786231  ggaagaggtagtggcttttttt agccgtaa  1784691  1786391  +ve  566  89108541   -                 short_chain_acyl-CoA_synthetase,_anaerobic
```

Figure 3.2: Example showing the output of coding PSSRs by PSSRFinder in 22 strains of *E. coli* genomes.



```
T  7   990301   990307   990307 ctcatggttttttt-atgacacctgccac   989194   990282 -ve 362 117623147 APECO1_41          990884   992284 -ve 466 162317582 APECO1_42
T  8  1037789  1037796  1037795 ctcatggttttttttatgacacctgccac  1036682  1037770 -ve 362  26246956 c1071             1038373  1039773 -ve 466  26246957 c1072
T  7   982016   982022   982022 ctcatggttttttt-atgacacctgccac   980927   981997 -ve 356 215486054 E2348C_0922         982600   984000 -ve 466 215486055 E2348C_0923
T  8   997388   997395   997394 ctcatggttttttttatgacacctgccac   996281   997369 -ve 362 110641126 ECP_0940            997973   999373 -ve 466 110641127 ECP_0941
T  8  1049704  1049711  1049710 ctcatggttttttttatgacacctgccac  1048597  1049685 -ve 362 218694404 EC55989_0978      1050287  1051687 -ve 466 218694405 EC55989_0979
A  8  2923350  2923357  2923351 ctcatggttttttttatgacacctgccac  2921374  2922774 +ve 466 170020668 EcolC_2666        2923376  2924464 +ve 362 170020669 EcolC_2667
T  8  1052914  1052921  1052920 ctcatggttttttttatgacacctgccac  1051807  1052895 -ve 362 157158264 EcE24377A_1029    1053497  1054897 -ve 466 157157150 EcE24377A_1030
T  8   991365   991372   991371 ctcatggttttttttatgacacctgccac   990258   991346 -ve 362 218688772 ECED1_0959          991949   993349 -ve 466 218688773 ECED1_0960
T  8  1048618  1048625  1048624 ctcatggttttttttatgacacctgccac  1047511  1048599 -ve 362 157160451 EcHS_A1037        1049201  1050601 -ve 466 157160452 EcHS_A1038
T  8  1039265  1039272  1039271 ctcatggttttttttatgacacctgccac  1038158  1039246 -ve 362 218553516 ECIAI1_0970       1039848  1041248 -ve 466 218553517 ECIAI1_0971
A  7  2267365  2267371  2267365 ctcatggttttttt-atgacacctgccac  2265388  2266788 +ve 466 218700551 ECIAI39_2217      2267390  2268463 +ve 357 218700552 ECIAI39_2218
T  8  1040152  1040159  1040158 ctcatggttttttttatgacacctgccac  1039045  1040133 -ve 362 170080587 ECDH10B_0999      1040736  1042136 -ve 466 170080588 ECDH10B_1000
T  8   986224   986231   986230 ctcatggttttttttatgacacctgccac   985117   986205 -ve 362  16128896 b0929               986808   988208 -ve 466  16128897 b0930
T  8  1119828  1119835  1119834 ctcatggttttttttatgacacctgccac  1118721  1119809 -ve 362 209397832 ECH74115_1090     1120412  1121812 -ve 466 209397026 ECH74115_1091
T  8  1205695  1205702  1205701 ctcatggttttttttatgacacctgccac  1204588  1205676 -ve 362  15800790 Z1276             1206279  1207679 -ve 466  15800791 Z1278
T  8  1116472  1116479  1116478 ctcatggttttttttatgacacctgccac  1115365  1116453 -ve 362  15830266 ECs1012           1117056  1118456 -ve 466  15830267 ECs1013
T  7   993890   993896   993896 ctcatggttttttt-atgacacctgccac   992783   993871 -ve 362 218557834 ECS88_0957          994473   995873 -ve 466 218557835 ECS88_0958
T  8  1070950  1070957  1070956 ctcatggttttttttatgacacctgccac  1069843  1070931 -ve 362 209918179 ECSE_0988         1071255  1071410 +ve  51 209918180 ECSE_0989
A  7  2200437  2200443  2200437 ctcatggttttttt-atgacacctgccac  2198460  2199860 +ve 466 170682866 EcSMS35_2190      2200462  2201535 +ve 357 170682361 EcSMS35_2191
T  8  1174602  1174609  1174608 ctcatggttttttttatgacacctgccac  1173495  1174583 -ve 362 218704357 ECUMN_1123        1175186  1176586 -ve 466 218704358 ECUMN_1124
T  7   991334   991340   991340 ctcatggttttttt-atgacacctgccac   990227   991315 -ve 362  91210031 UTI89_C1001         991917   993317 -ve 466  91210032 UTI89_C1002
T  8   987423   987430   987429 ctcatggttttttttatgacacctgccac   986316   987404 -ve 362  89107779  -                  988007   989407 -ve 466  89107780 -
```

Figure 3.3: Example showing the output of non-coding PSSRs by PSSRFinder in 22 strains of *E. coli* genomes.

Chapter 4

Investigations on mutational status of SSRs in *Yersinia pestis* and its ancestor

4.1 Introduction

In Chapter 2 a detailed genome-wide and gene-wise analyses of SSRs with regard to their distribution, abundance and enrichment in the known Yersinia genomes was presented. As a logical extension, this chapter reports further studies on the mutational status of SSRs and their effects on the nearby coding regions. In order to discover mutational status of the SSRs and their effects on gene functions cross-genome comparisons of SSRs were carried out as detailed in this chapter. These comparative studies helped to discover a number of polymorphic SSR tracts which seemed to be involved in the evolution of Yersinia from its ancestor to all its different known forms of strains.

It is worthwhile to mention here that a number of whole genome comparative studies have been carried out between some strains of *Y. pestis* and *Y. pseudotuberculosis* to infer changes involved in molecular evolution. For example, suppression subtractive hybridization studies (Wang et al. 2006), DNA microarray studies (Hinchliffe et al. 2003; Wang et al. 2006) and whole genome comparisons (Chain et al. 2004; Pouillot, Fayolle, and Carniel 2008). However, these studies considered only the large rearrangements and large (in kb range) insertions and deletions. Some other studies considered long repetitive sequences for finding markers for strain differentiation (Adair et al. 2000; Klevytska et al. 2001; Le Fleche et al. 2001; Pourcel et al. 2004). Girard et al. (2004) and Vogler et al. (2007) analyzed mutation rates and mutational mechanism of variable number of tandem repeats (VNTR) identified from various foci. To the best of my knowledge I have not come across any report on systematic cross-genome comparative study of polymorphic SSRs and their effects on coding regions as detailed in this chapter.

4.2 Methods

In order to identify the polymorphic SSRs in Yersinia genomes I used the program PSSRFinder whose details have already been mentioned in the last chapter.

4.3 Results

4.3.1 Polymorphic SSRs

PSSRFinder identified 805 PSSRs which have undergone INDEL mutations of their repeat units in Yersinia genomes which include various strains of both *Yersinia pestis* and *Yersinia pertuberculosis*. In order to sort PSSRs as belonging to coding or non-coding regions the regions harboring them were examined. If all the equivalent SSRs of a PSSR occured within intergenic regions in all the genomes then it was referred to as non-coding PSSR otherwise it was referred to as coding PSSR. **Tables 4.1, 4.2 and 4.3** give some details of these PSSRs from which the following salient features of PSSRs can be noticed:

(a) The non-coding regions are denser with PSSRs than the coding regions

(b) a large majority (695 out of 805) of the PSSRs are mono-nucleotide tracts

(c) 95% of PSSRs are formed by small tracts (≤8bp)

(d) Only 10% of SSRs have undergone tract expansion/contraction of more than one repeat unit regardless of the tract length (please see **Table 4.3**). In other words, in a large majority amounting to 90% of PSSRs length polymorphism is due to INDEL of one repeat unit.

The complete list of PSSRs showing more than one repeat unit increment or decrement for each strain is given in **Table 4.4**. Since *Yersinia pestis* is a highly monomorphic species, these polymorphic SSR loci can offer potential markers for strain differentiation.

4.3.2 Distribution of PSSRs in various COG groups

In Chapter **2**, it was observed that SSR distribution is biased towards coding regions belonging to certain COG groups. In the light of this it was interesting to compare percentages of PSSRs to the percentages of SSRs found in various COG groups and therefore ratio of percentage of PSSRs to percentage of SSRs were computed in various COG groups. The values of the ratio are shown in **Figure 4.1.** It is interesting to note that ORFs encoding proteins with certain functions such as replication, recombination and repair (L), signal transduction (T), Cell motility (N), Intracellular trafficking, secretion, and vesicular transport (U) etc. show the ratio values >1.0 indicating preferential distribution of PSSRs in these regions. On the other hand ORFs having roles in general metabolism show the ratio values <1.0. It would be interesting to investigate further why SSR polymorphism should show such a bias.

Table: 4.1: Distribution of PSSRs in coding and non-coding regions of Yersinia genomes.

Repeat type	Coding regions		Non-coding regions	
	Number of PSSRs/10kb	Number of PSSRs	Number of PSSRs/10kb	Number of PSSRs
Mono nucleotide	10.31	382	33	313
Di nucleotide	0.08	3	0.9	8
Tri nucleotide	0.7	26	0.3	3
Tetra nucleotide	0.4	13	1.0	10
Penta nucleotide	0.05	2	1.0	9
Hexa nucleotide	0.4	15	2.3	21
Total (805)	11.92	441	38.5	364

Table 4.2: The number of PSSRs of different repeat types (mono to hexa) of different repeat copy numbers.

	2	3	4	5	6	7	8	9	10
Mono	52	84	103	76	98	133	102	26	4
Di	3	0	3	3	2	0	0	0	0
Tri	5	16	4	0	0	0	0	0	0
Tetra	7	14	5	0	0	0	0	0	0
Penta	5	3	0	0	0	0	0	0	0
Hexa	11	3	1	0	0	0	0	0	0

Table 4.3: The number of PSSRs showing INDELs of more than one repeat unit.

	Mono	Di	Tri	Tetra	Penta	Hexa
Coding regions (34)	23	0	1	2	0	8
Non-coding regions (46)	27	0	0	5	6	8

Table 4.4: The list of PSSRs showing INDEL mutations of more than one repeat unit. The start and end positions of the PSSRs are given with reference to YPTB (*Yersinia pseudotuberculosis* IP32953) strain and where the homologous region is absent in YPTB then the start and end positions are given with reference to *Y. pestis* CO92. "Range" indicates the range of repeat count variation observed in all the Yersinia genomes.

Repeat Motif	Repeat count	Start position	End position	Genome Name	Range
GCA	6	3063319	3063336	YPTB	5−11
C	7	3896530	3896536	YPTB	7−16
G	9	1177861	1177869	YPTB	7−10
A	8	1986119	1986126	YPTB	6−8
G	10	1651010	1651019	YPTB	8−11
G	6	925331	925336	YPTB	6−11
C	11	4281415	4281425	YPTB	9−14
TCACT	9	214821	214865	YPTB	2−9
TCAAC	2	3169200	3169209	YPTB	2−5
AAATA	3	3407101	3407115	YPTB	2−5
TGAAC	2	979946	979955	YPTB	2−4
ACTC	5	1170842	1170861	YPTB	2−5
GATA	7	2817206	2817233	YPTB	3−7
TATC	6	3833531	3833554	YPTB	3−8
ATTCA	3	807520	807534	YPTB	2−5
AACA	11	2494167	2494210	CO92	3−15
TCATGT	2	4432550	4432561	CO92	2−4
TACACC	3	2570948	2570965	CO92	2−4
G	9	1921508	1921516	CO92	8−14

4.3.3 Position-specific distribution of PSSRs in the intra-genic regions

In Chapter 2 it was shown that SSRs to occur in a biased manner in the 5’ and 3’ parts of the genes as compared to their ‘middle’ parts. PSSRs also show a similar bias (38% in 5’ as well as in 3’ parts; chi-square = 16.32, DF = 3, P value = 0.001). These PSSRs are associated with certain changes in coding regions as illustrated and summarized in **Figure 4.2**. It should be noted that the homologus coding regions harboring each one of the PSSRs are identical except for the change as indicated by the repeat copy number variation of the SSR. Therefore any change seen in a coding region is entirely the consequence of expansion or contraction of the associated PSSR tract. In the following paragraphs various effects caused by SSR mutations have been described.

N-terminal extension: This is observed in 18 ORFs. In these cases PSSRs cause N-terminal side extensions of proteins. For example: the competence/damage-inducible gene (y0880) codes for a 162 amino acids long protein in all the Yersinia genomes except the *Y. pestis* Angola strain where it codes for a 184 amino acids long protein as a consequence of frame-shift mutation induced by one G deletion in $(G)_5$ tract.

N-terminal contraction: This is observed in 54 ORFs. In these cases PSSRs are associated with deletions of peptide segments from the N-terminal sides. For example: ORFs YPA_4156 and YpUG050454_2819 encode for 253 aa long periplasmic solute binding protein in *Y. pestis* Antiqua and *Y. pestis* UG05_0454 respectively whereas the equivalent ORFs in the other strains encode for a protein of length 284 amino acids. The shortened protein variant in *Y. pestis* Antiqua and *Y. pestis* UG05_0454 strains lack the N-termial secretory signal peptide as compared to its longer variant in the other strains. This N-terminal deletion is due to a deletion mutation in a tract of $(T)_8$ which has frame-shifted the start site by 31 amino acids downstream.

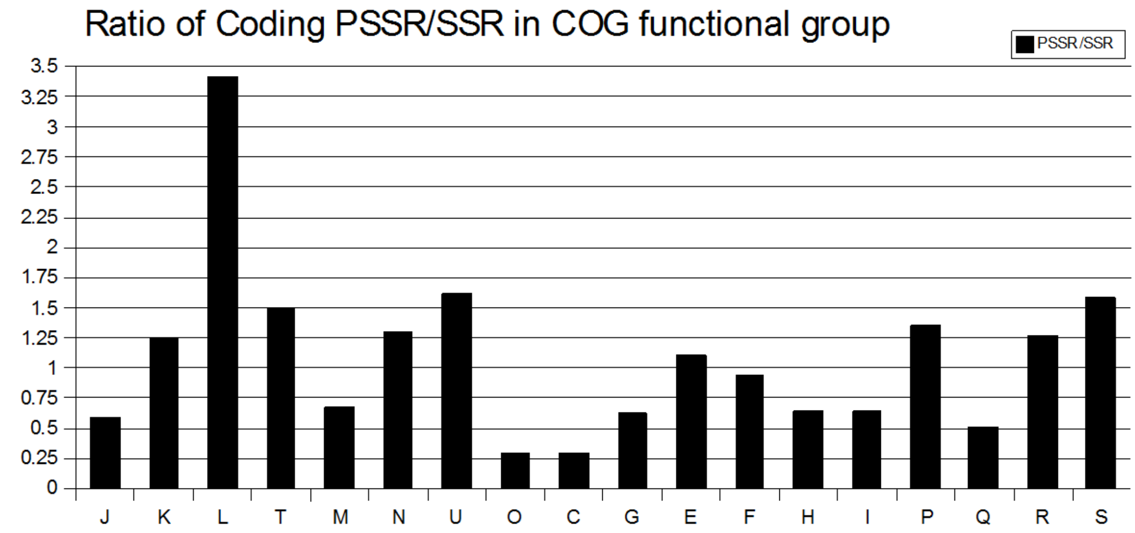


Figure 4.1: Ratio of percentages of PSSRs and SSRs found in ORFs encoding proteins belonging to different COG functional groups. The ratio of coding PSSR/SSR is comparatively more than one in the COG functional groups; L, T, M, N, U & O whereas the proteins belonging to general metabolism the values are always <1. (Key for various COG functional groups is given on page number 67 in the last chapter).

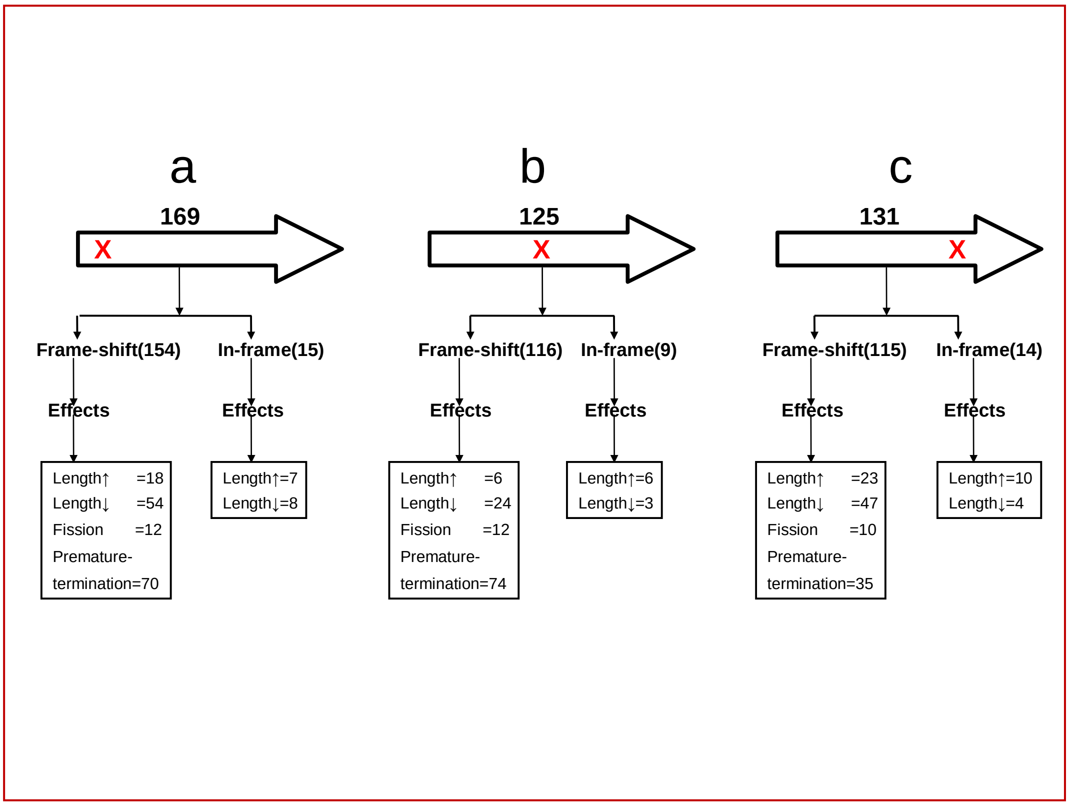


Figure 4.2: Distribution of PSSRs in the 5' (a), middle (b) and 3' (c) parts of genes. The observed effects (show below (a), (b) and (c)) (length↑ and length↓ indicate increment and decrement respectively of translated protein product due to PSSR).

C-terminal extension: In this case as observed in 23 ORFs SSR mutations at 3'end have lead to extensions of C-terminal sides in the corresponding protein products. For example disulfide oxidoreductase gene (DsbE), encodes for a 188 amino acids long protein in all the Yersinia genomes except in *Y. pestis* Nepal (YPN_2207) where it encodes for a 239 amino acid long protein. This reading frame extension has happened due to the frameshift mutation caused by deletion of one "G" in $(G)_6$ tract. .

C-terminal contraction: This is observed in 47 ORFs. In these cases protein products have shorter C-terminii due to deletion of a few amino acids as compared to their wild-type counterparts. For example, in the *Y. pestis* FV_1 strain due to insertion of one "G" in the $(G)_4$ tract, the ORF has terminated at 354th amino acid (YpesF_010100018914) whereas the homologous region in other Yersinia genomes code for a 744 amino acid long GTP pyrophosphokinase which has three domains (Rel-Spo_like (250-350) TGS_RelA_Spo (410-470) and ACT_RelA-SpoT (660-744)).

Gene fission: This is observed in 22 genes. In all these cases PSSRs have caused splitting of ORFs into two smaller ORFs. For example ORF YPTB1505 in *Y. pseudotuberculosis* IP32953 strain codes for a 150 amino acid long protein which is a type VI secretion system lysozyme related protein. Due to the $(C)_2$ to $(C)_3$ expansion the ORF has split into two ORFs encoding peptides of 63 (y2681) and 89 (y2680) amino acids in *Yersinia pestis* CO92 strain.

Psuedogene formation: This is observed in 179 genes. In these cases PSSRs have converted genes into pseudo genes. For example ORF y1081 encodes a 468 amino acid protein mannose-1phosphate guanylyltransferase in *Y. pestis* CO92 strain whereas the homologous region in the strain *Y. pestis* media 91000 is annotated as non-coding region because it is prematurely terminated due to the $(C)_4$ to $(C)_3$ mutation.

4.3.4 SSRs showing length variation between *Yersinia pestis* and *Yersinia pseudotuberculosis*

As mentioned earlier *Y. pseudocuberculosis is* the ancestor of *Y. pestis* and it is therefore interesting to investigate the nature of SSR polymorphism between these two different strains. Among the PSSRs found in coding regions 96 showed length variation between *Y. pestis* and *Y. pseudotuberculosis* genomes. Previously, it was repoted that several genes of YPTB have become non-functional in *Y. pestis* and this gene-loss was attributed as a cause for increased virulence in the later (Parkhill et al. 2001; Song et al. 2004)**.** About 31 genes of YPTB were found terminated prematurely in all the strains of *Y. pestis* due to PSSRs. Detailed analyses of all the genes which have been affected by SSR polymorphism was carried out. Of these genes following examples were found to be interesting owing to their role in adaptability and pathogenesis of *Y. pestis* as compared to YPTB.

Urease enzyme: The functional UreD gene which encodes urease enzyme harbors $(G)_5$ tract in YPTB. In *Y. pestis* this tract has been expanded to $(G)_6$ resulting in a frame-shift mutation leading to premature termination of the gene. This correlates with the life style of these pathogens. YPTB, as it lives in soil and water, requires *urease* enzyme for the degradation of nitrogenous products for its nitrogen requirements whereas *Y. pestis* due to its different life style does not require *urease* activity.

Genes involved in O-antigen and Lipopolysaccharide (LPS) synthesis: In enteropathogenic bacteria, the O-antigen is an important component of LPS- a major component of the outer membrane of Gram-negative bacteria (Skurnik and Bengoechea

2003). In *Y. pestis* genomes, genes involved in O-antigen synthesis have been prematurely terminated as compared to YPTB due to SSR copy number variations in them. The genes in YPTB which have been prematurely terminated in *Y. pestis* are: YPTB1000 (CDP-D glucose 4-6-dehydratase) due to $(G)_9$ tract expansion; YPTB1004 (putative O-unit flippase) due to contraction of $(T)_8$; YPTB1007 (O-unit polymerase like protein) due to $(T)_8$ expansion; YPTB1009 (GDP-mannose 4,6-dehydratase) due to expansion of $(T)_2$; YPTB1010 (GDP-fucose synthase) due to contraction of $(G)_7$; and YPTB1011 (mannose-1-phosphate guanyltransferase) due to contraction of $(C)_4$. As a consequence, YP produces rough LPS whereas YPTB with intact O-antigen biosynthesis pathway genes produces smooth LPS. It has been demonstrated that YPTB which produces smooth LPS inhibits the activity of plasminogen *(Pla)* (Kukkonen et al. 2004; Pouillot et al. 2005; Lathem et al. 2007) which is central in the pathogenesis of the plague. Absence of O-antigen therefore leads to active plasminogen.

Inv* and *YopI* genes*: In YPTB the *Inv* gene (YPTB1572), a chromosomal gene and *YopI* (pYV0013), a plasmid pYV (pCD) gene, encode membrane proteins. In *Y. pestis* both these genes are non-functional whereas in YPTB both are functional. Studies have shown that non-functionality of these genes is essential for the pathogenicity of *Y. pestis* (Rosqvist, Skurnik, and Wolf-Watz 1988). In *Y. pestis* the *Inv* gene has become non-functional due to an insertional sequence whereas *YopI* gene has prematurely terminated due to the frame-shift mutation caused by the contraction of poly "A" tract from $(A)_9$ in YPTB to $(A)_8$. It appears that *Y. pestis* after acquiring insertional sequence in *Inv* has selected for non-functional *YopI* which has been made possible by the contraction of poly "A" tract. In *Y.*

pseudotuberculosis genome one more ORF function is annotated as YadA domain containing protein (YPTS_3309) of size 716 amino acids. The homolog of this ORF is a pseudogene in all the *Y. pestis* genomes. This is due to the deletion of ATTT tract in *Y. pestis* whereas it is repeated three times in *Y. pseudotuberculosis.* This is an example where SSR polymorphism has led to the switching off of a gene for sustenance of virulence by the pathogen.

Genes involved in flagella and pili biosynthesis pathway: YPTB has *FlhA* an important gene for flagella biosynthesis which encodes for a protein of 698 amino acids (YPTB3357) in length. In most of the *Y. pestis* genomes *FlhA* genes have been prematurely terminated as compared to YPTB due to the insertion of one "G" in $(G)_8$ SSR tract. In addition to this, the upstream regions of some of the genes predicted to have role in flagella assembly also harbor five PSSRs which might have some regulatory effect on the expression of the flagellar gene. These observations correlate with the fact that YPTB has Mot+ phenotype at 30° C and Mot- at 37° C whereas the YP is a known non-motile organism. Furthermore the genes involved in pili synthesis which translate into an extra cellular structure also harbor PSSRs. In YPTB strains the pili assembly chaperone is intact and code for a protein of length equal to 246 amino acids (YPTB1917) whereas in *Yersinia pestis* the homologous ORFs are frame shifted due to the deletion of one "T" nucleotide from the tract of $(T)_4$.

Genes involved in Cysteine synthesis: Cystathionine gamma-synthase, is central to cysteine biosynthesis, is intact in YPTB (YPTB0105), but is prematurely terminated in most of the *Y. pestis* genomes due to $(G)_8$ tract contraction. In addition to this, one more gene of cysteine biosynthesis pathway, cysteine synthase B, has also undergone N-terminal contraction in all the YP strains as compared to its homolog in YPTB (YPTB2731) due to contraction of $(A)_7$ tract. These mutations in the cysteine synthesis pathway in *Y. pestis* are one of the reasons that the pathogen is an auxotroph whereas its ancestor *Y. pseudotuberculosis* is a chemoheterotroph.

Glycosyl hydrolase: Glycosyl hydrolase cleaves and disturbs the initial polymerization of poly- β-1, 6-GlcNAc in *Y. pestis* as well as in the other bacteria (Kaplan et al. 2003; Itoh et al. 2005). The poly-β-1, 6-GlcNAc is a major constituent of biofilm in *Y. pestis*. The biofilm formation in flea is essential for the blockage in flea gut (digestive tract) which is vital to the flea borne transmission of *Y. pestis* to host.

In YPTB, glycosyl hydrolase (YPTB3837) which encodes for a 371 amino acids long protein has an N-terminal signal peptide (position 1 to 20) indicating that it is a secretory protein. The homologs in YP strains have undergone N-terminal contraction by about 60 amino acid residues that include signal peptide as a consequence of contraction of $(A)_8$ tract. This affects secretion of the protein into the periplasm. Absence of this protein in the periplasm may stabilize initial polymerization of poly- β-1, 6-GlcNAc in *Y. pestis*. This difference in glycosyl hydrolase might be one of the reasons for the difference in the ability of *Y. pestis* and YPTB to form biofilm in flea gut eventhough they share identical loci (*hms*) for the formation of biofilm.

4.3.5 SSRs showing length variation among the different strains of *Yersinia pestis*

Of the PSSRs discovered, 208 PSSRs were found across different strains of *Yersinia pestis*. In these examples the repeat copy number remains the same between YPTB and one of the *Y.*

pestis genomes but varies among the different strains of *Y. pestis*. Length variation in these PSSRs may have played some role during strain differentiation and development of *Y. pestis.* Of these PSSRs 14 tracts show length conservation between *Y. pseudotuberculosis* and *Y. pestis* 91001 (non-plague causing Yersinia strain) but vary in all the other genomes. These are interesting tracts as YPTB and YP are non-pathogenic (non plague causing) to humans and hence it would be interesting to explore the role of these SSR mutations in pathogenicity. In plague causing strains the coding regions harboring these PSSRs have undergone changes such as premature termination (cystathione gamma-synthase, histidine kinase, guanosine-5'-triphosphate,3'-diphosphate diphosphatase and ABC putative drug efflux transporter), length variation (repressor of aceBA operon, gamma glutamyl trans-peptidase precursor, general secretory protein and two hypothetical proteins) and domain splitting (ATPase component of ABC transporters with duplicated ATPase domains and a conserved hypothetical protein).

16 tracts were also found whose lengths are conserved in all the plague causing strains but vary in the *Y. pestis* 91001 genome which is non-pathogenic (non-plague causing) to humans. These 16 loci are found in the ORFs encoding proton glutamate symport, flagellar motor transmembrane protein, putative aliphatic sulfonates binding protein (solute binding periplasmic protein), putative invasion protein, putative insecticidal toxin, putative O-unit flippase, mannose-1-phosphate guanylyltransferase, outer membrane usher protein, ABC transporter protein and hypothetical proteins. It is interesting to note here that some of the O-antigen cluster genes which are mutated in all other *Y. pestis* are intact in this strain. These genes need to be further studied in order to reveal their role in plague related

pathogenicity in *Y. pestis*.

4.4 Discussion

The cross-genome comparative studies have revealed several polymorphic SSRs either within the ORFs which encode surface proteins or within ORFs which are involved in the regulation of genes that interact with the host environment. These structural proteins face most of the insults of the host immune defence mechanism and thus drastic changes brought out by SSR mutations such as length variations, domain fission/fusion and even premature terminations are beneficial for the survival of the pathogens.

Between Y. *pestis* and its ancestor a number of exclusive PSSRs have been discovered. Among these, 38 of them have caused premature termination in several genes suggesting SSR mutation as one of the reasons for gene loss in *Y. pestis*.

It is to be noted that the number of PSSRs is negligibly small as compared to the total number of available SSRs in the genomes. However, these tracts are noteworthy as they are the evidences of polymorphism selected during evolution. Perhaps it is apt to call them as “SSR fossils”. Although the genomes harbor many SSRs and all of which have potential to undergo length variation due to slippage, the available data suggest that mis-match repair perhaps compounded by selection have allowed only a small number of SSRs to undergo mutations and exhibit polymorphism.

PSSR analysis corroborated with the available experimental data clearly postulates the important role of some of the SSRs helping to bring out intra-species variations. In one of the examples, in Yersinia genomes, the LPS synthesis constitutes seventeen genes which are located between the *hemH* and *gsk* loci. About six of these genes correspond to the O-antigen gene cluster that are prematurely terminated in all the plague causing strains of *Y. pestis* genomes as compared to YPTB due to mutations in SSRs. The heterologous replacement of these loci in *Y. pestis* decreased its virulence (Kukkonen et al. 2004; Lathem et al. 2007). On the other hand non-functional genes of flagella and pili operon biosynthesis pathway of *Y. pestis* due to SSR mutation may be viewed as an adaptation due to the change in the life style of the pathogen. Furthermore, SSR mutations in the cysteine synthesis pathway are one of the reasons that *Y. pestis* is an auxotroph whereas its ancestor *Y. pseudotuberculosis* is a chemoheterotroph.

The unique SSR mutations found in the non-pathogenic YP is noteworthy. Genes showing SSR mutation induced changes in all the plague causing *Y. pestis* but are intact in the non-pathogenic YP (non-plague causing *Y. pestis*) suggesting that the corresponding gene products probably lower human-specific virulence properties in YP. On the other hand, there are some other genes which are intact in all the plague causing Yersinia but are mutated due to SSR mutations in *Y. pestis* 91001. These gene products may also have a role in causing virulence.

Though, all the 20 (16 YP and 4 YPTB) genomes analysed in the present study are highly similar to each other (≥98% identity in the homologous regions), they vary in their disease causing abilities. Earlier it was believed that the extent of virulence is related to the number of pathogenic plasmids in them. However, after sequencing the non-pathogenic *Y. pestis* 91001 (Song et al. 2004) strain which contains all the pathogenic plasmids, it is assumed that other factors may perhaps be responsible for their virulence (Revell and Miller 2001; Song et al. 2004). As revealed from this study mutations in PSSRs seem to play some role in rendering virulence to the pathogen.

4.5 Summary

Yersinia genomes harbor a number of polymorphic SSRs in coding as well as non-coding regions. The non-coding regions are three times denser in PSSRs than the coding regions indicating lesser evolutionary restraints on SSR polymorphism in non-coding regions than coding regions. Most of the PSSRs (90%) in coding regions are of mononucleotide repeats and hence the majority of PSSRs have caused shifts in the reading frames manifesting in premature terminations, length variations and splitting of ORFs in the mutated genomes. Corroborating with the available experimental data our PSSR analysis clearly postulates the important role of some of the SSRs helping to bring out intra-species variations. The observation made in this study throws light on possible roles of SSR length variation in the evolution of *Yersinia pestis* from its ancestor *Yersinia pseudotuberculosis*.

Chapter 5

Mutational Dynamics of Simple Sequence Repeats in the Intragenic Regions of Prokaryotic Genomes

5.1 Introduction

In the last chapter cross-Yersinia genome comparative studies of SSRs was presented with several SSRs showing length polymorphism that were discovered in this study. A further investigation revealed an interesting facet of SSR polymorphism viz., its involvement in the evolution of the pathogen. Such cross-genome comparative studies on polymorphic status of SSRs have rarely been reported in the literature despite the fact that complete genome sequences pertaining to several strains and isolates are available for a number of prokaryotes. It is now possible to compare equivalent SSRs across different strains of a species and address some interesting questions concerning intra-species (or inter-strain) polymorphism and evolution of SSRs. In the present study I havecarried out a number of cross-genome comparisons involving forty-three species of prokaryotes and have discovered a large number of intra-species PSSRs. These PSSRs were studied for their distribution tendencies in intragenic and intergenic regions and in different positions (5'-end, middle and 3'-end) within intragenic regions. Since our global study on PSSRs considers a number of prokaryotic species belonging to animal pathogens, plant pathogens and free living organisms I also compared the PSSR distribution in these groups. In prokaryotes, DNA replication and transcription occurs simultaneously and continueously throughout their life cycle and hence collision between the DNA and RNA polymerases are deemed inevitable. The intragenic regions where replication and transcription directions are opposite to each other (collision is head-on and it is called as head-on direction) have been reported to be more prone for mutations than the intragenic regions where the replication and transcription directions are same (collision is co-oriented and it is called as co-oriented direction). Therefore, it was considered important to investigate the distribution pattern-of PSSRs in head-on as well as co-oriented directions of transcription and replication.

5.2 Methods

5.2.1 Extraction of SSRs and comparison of equivalent SSRs

Annotated whole genome sequences of prokaryotic genomes available on public domain were downloaded from the ncbi (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/). In addition to these genomes partially annotated whole genome assembly (WGS) sequences of some of the prokaryotes were also downloaded from http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi. However only those species were considered where complete genome sequences are available for at least three strains. The whole genome sequences were screened for extracting SSRs and PSSRs using SSRF and PSSRFinder respectively, as already detailed in the previous chapters. PSSRs only if harbored within non-coding regions in all the compared strains were categorized as non-coding PSSRs otherwise they were referred to as coding PSSRs.

5.2.2 Prediction of origin and termination site of replication and calculation of SSR and PSSRs in head-on and co-oriented directions

The prediction of *ori* and *ter* sites in various prokaryotic genomes was done on the basis of their GC-skewness. Using these sites as reference positions of SSRs and PSSRs relative to co-oriented and anti-oriented directions of replication and transcription were determined. Only those species were considered in which the size of replichore was similar in all the strains. In other words, species where an exchange of DNA segments (atleast in kb) has taken place between their replichores was not considered for analysis. The statistical significance of distribution of PSSRs in different regions (coding verses non-coding, head-on PSSR verses co-oriented PSSR) was tested using Chi-square test. If SSR polymorphism is neutral then one would expect distribution pattern of PSSRs to be very similar to that of SSRs. As it is a case of one degree of freedom, any chi-square value greater than 3.84 is significant. Chi-square value was calculated as given below.

$$\chi^2 = \sum_{i=1}^{k} (O_i - E_i)^2 / E_i$$

Where, O_i = number of PSSRs observed in ith region, E_i = number of PSSRs expected under the assumption that the number of of PSSRs is same as the number of SSRs observed in ith region. Hence, Ei=N*Mi where N=Total number of PSSRs and Mi=fraction of SSRs observed in the ith region.

i =1, *k* = 2 (only two places (coding or non-coding, head-on or co-oriented)

5.3 Results

5.3.1 Polymorphic SSRs in various prokaryotic genomes

Cross-genome comparisons revealed 18351 PSSRs from 43 species of prokaryotes with an average of 426 PSSRs per species. Of these PSSRs, 7075 were found in the coding regions and 11276 in the non-coding regions. **Figure 5.1** shows the number of PSSRs found in each species as well as their densities in the coding and the non-coding regions. In all the species the tract density in the non-coding regions is strikingly higher than (on average ~ 14 times) that in the coding regions. Higher incidence of SSR polymorphism in the non-coding regions as compared to the coding regions indicates relatively unrestrained polymorphism in the non-coding regions. Restraint on SSR polymorphism in the coding regions can be attributed to selection pressures. If SSR mutation in the coding region is not under selection pressure then one would expect distribution of PSSRs to be very similar to the distribution of SSRs in both the coding as well as the non-coding regions. In order to find if such a distribution pattern exists, a ratio of % of PSSRs to % of SSR in the coding and non-coding regions for every species was calculated. The value of the ratio equal to one indicates that distribution of PSSRs is same as that of SSRs in coding and non-coding regions. It is evident from the **Figure 5.2** that the % PSSRs is lower than % SSRs in the coding regions whereas they are higher in the non-coding regions. This clearly demonstrates that SSR polymorphism in the coding regions is restrained in comparison to the non-coding regions and the underlying factor responsible for this is the selection against frame shift mutations in coding regions (Metzgar, Bytof, and Wills 2000). The significance of this biased distribution of PSSRs in non-coding regions was calculated and it was highly significant in all most all the species except Clostridium *perfringens* and *Mycobacterium tuberculosis* (Chi-square value was very high and P-value <<<0.00001).

Mutations in mono, di, tetra, and penta nucleotide SSRs shift reading frames in coding regions therefore these SSRs are refered as frame-shift PSSRs. On the other hand mutations in tri and hexa nucleotide SSRs would only create an in-frame mutation and hence these SSRs are refered as in-frame PSSRs. The number of PSSRs belonging to these two groups in various species is shown in **Figure 5.3**. It is interesting to note that about 90% of the coding

PSSRs are constituted by the frame-shift PSSRs in coding regions of genomes and hence it can be argued that selection has allowed frame shift mutations in coding regions. The complete list of coding PSSRs of different repeat types in various species of prokaryotes is shown in **Table 5.1**. From the table it can be seen that the mono nucleotide tracts are the most abundant among the PSSRs followed by tri and hexa nucleotide tracts.

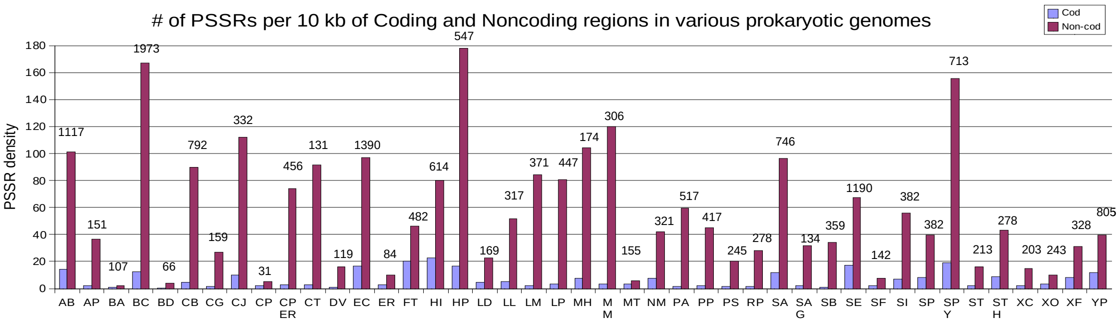


Figure 5.1: Tract densities of PSSRs in the coding (blue bar) and the non-coding (magenta bar) regions of various prokaryotic species. The number of PSSRs found in each species is shown on each magenta colored bar.

(Foot note) AB = *Acinetobacter baumannii*, AP = *Actinobacillus pleuropneumoniae*, BA = *Bacillus anthracis*, BC = *Bacillus cereus*, CB = *Clostridium botulinum*, CG = *Corynebacterium glutamicum*, CJ = *Campylobacter jejuni*, CP = *Chlamydophila pneumoniae*, CPER = *Clostridium perfringens*, CT = *Chlamydia trachomatis*, DV = *Desulfovibrio vulgaris*, EC = *Escherichia coli*, ER = *Ehrlichia ruminantium*, FT = *Francisella tularensis*, HI = *Haemophilus influenzae*, HP = *Helicobacter pylori*, LL = *Lactococcus lactis*, LM = *Listeria monocytogenes*, LP = *Legionella pneumophila*, MH = *Mycoplasma hyopneumoniae*, MM = *Methanococcus maripaludis*, MT = *Mycobacterium tuberculosis and Mycobacterium bovis*, NM = *Neisseria meningitidis*, PA = *Pseudomonas aeruginosa*, PP = *Pseudomonas putida*, PS = *Pseudomonas syringae*, RP = *Rhodopseudomonas palustris*, SA = *Staphylococcus aureus*, SAG = *Streptococcus agalactiae*, SB = *Shewanella baltica*, SE = *Salmonella enterica*, SF = *Shigella flexneri*, SI = *Sulpholobus islandicus*, SP = *Streptococcus pneumoniae*, SPY = *Streptococcus pyogenes*, ST = *Salmonella typhi*, STH = *Streptococcus thermophilus*, XC = *Xanthomonas campestris*, XO = *Xanthomonas oryzae*, XF = *Xylella fastidiosa*, YP = *Yersinia pestis.*

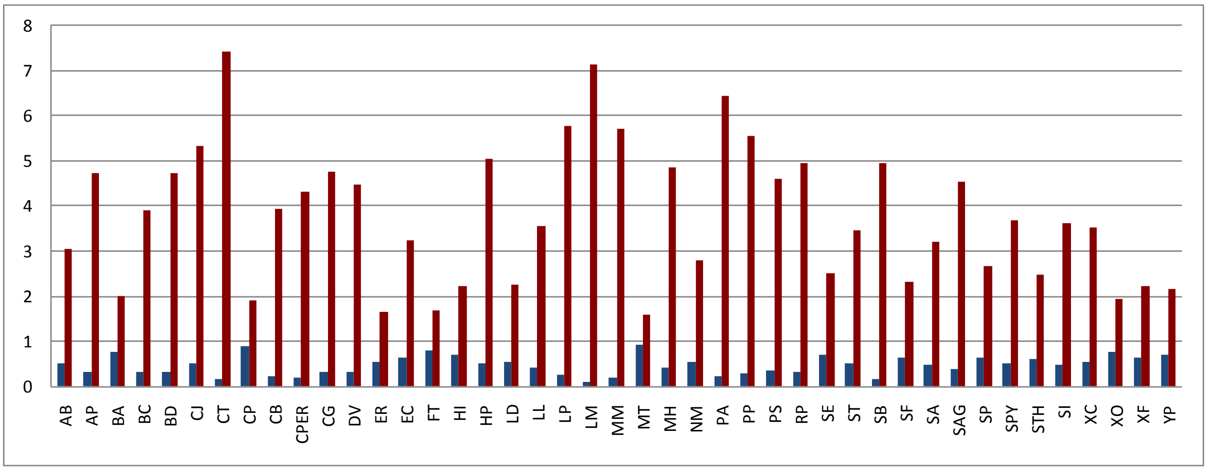


Figure 5.2: Ratio of % of PSSRs to % of SSRs in the coding (blue) and the the non-coding (magenta) regions of various prokaryotic species. In most of the species (except CP and MT) the chi-square value is very high and P-value is <<< 0.0001.

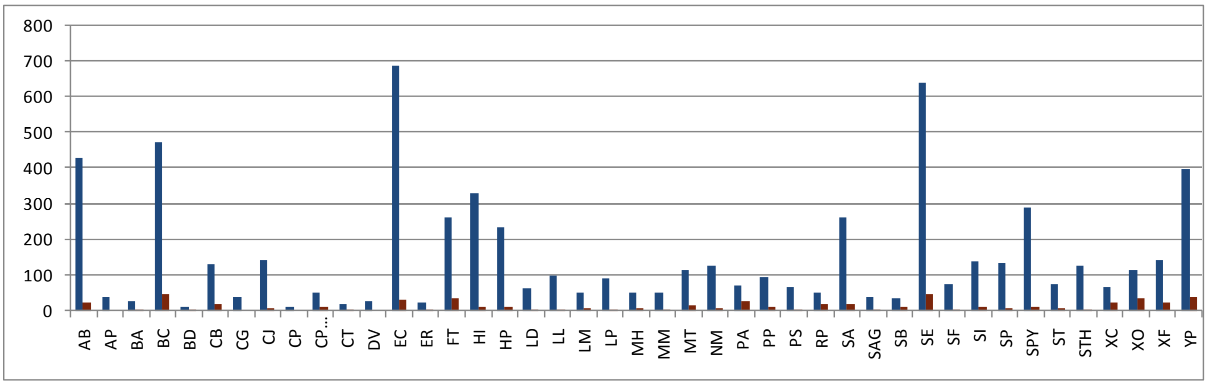


Figure 5.3: Number of FS-PSSRs (blue) and IF-PSSRs (magenta) found in various prokaryotic species.

Table 5.1: The number of coding PSSRs, repeat unit size-wise, found in various prokaryotic species.

Species name followed by number of strains	Total	Mono	Di	Tri	Tetra	Penta	Hexa
Acinetobacter_baumannii_6	**452**	422	6	17	0	1	6
Actinobacillus_pleuropneumoniae_3	**41**	38	0	1	1	0	1
Bacillus_anthracis_3	**31**	25	1	5	0	0	0
Bacillus_cereus_8	**524**	458	14	41	2	0	9
Bradyrhizobium_3	**18**	12	1	5	0	0	0
Clostridium_botulinum_a_8	**149**	125	5	16	0	0	3
Corynebacterium_glutamicum_3	**41**	39	0	1	0	0	1
Campylobacter_jejuni_5	**155**	139	5	9	1	0	1
Chlamydophila_pneumoniae_4	**24**	10	0	0	0	0	0
Clostridium_perfringens_3	**66**	49	3	14	0	0	0
Chlamydia_trachomatis_4	**24**	20	1	3	0	0	0
Desulfovibrio_vulgaris_3	**30**	28	1	1	0	0	0
Escherichia_coli_22	**736**	668	15	25	5	1	8
Ehrlichia_ruminantium_3	**27**	21	1	5	0	0	0
Francisella_tularensis_8	**300**	238	22	32	1	2	5
Haemophilus_influenzae_4	**342**	323	0	10	7	1	1
Helicobacter_pylori_6	**247**	228	4	14	1	0	0
Lactobacillus_delbrueckii_2	**65**	61	1	3	0	0	0
Lactococcus_lactis_3	**104**	101	0	3	0	0	0
Listeria_monocytogenes_3	**58**	50	1	7	0	0	0
Legionella_pneumophila_4	**96**	88	1	5	1	0	1
Mycoplasma_hyopneumoniae_3	**60**	49	3	6	0	1	1
Methanococcus_maripaludis_4	**53**	48	1	3	0	1	0
Mycobacterium_tuberculosis_6	**131**	107	7	15	0	1	1
Neisseria_meningitidis_4	**135**	125	0	6	2	0	2
Pseudomonas_aeruginosa_4	**103**	72	1	17	0	0	13
Pseudomonas_putida_4	**106**	92	2	10	0	0	2

Pseudomonas_syringae_3	**69**	63	2	3	0	1	0
Rhodopseudomonas_palustris_6	**72**	49	4	19	0	0	0
Staphylococcus_aureus_14	**280**	248	10	18	3	0	1
Streptococcus_agalactiae_3	**44**	38	1	3	0	0	2
Shewanella_baltica_3	**47**	34	0	8	0	0	5
Salmonella_enterica_13	**690**	616	20	40	6	0	8
Shigella_flexneri_3	**70**	70	0	1	3	1	4
Sulpholobus_islandicus_6	**154**	133	8	13	0	0	0
Streptococcus_pneumoniae_5	**144**	135	1	6	1	0	1
Streptococcus_pyogenes_13	**304**	290	1	11	1	0	1
Salmonella_typhi_3	**87**	72	4	6	1	0	4
Streptococcus_thermophilus_3	**133**	125	1	6	0	1	0
Xanthomonas_campestris_4	**94**	63	4	11	0	1	15
Xanthomonas_oryzae_3	**150**	105	3	11	5	2	24
Xylella_fastidiosa_4	**165**	134	5	11	3	0	12
Yersinia_pestis_20	**441**	382	3	26	13	2	15

5.3.2 PSSRs in the co- and anti-oriented directions of replication and transcription

In bacteria, DNA replication and transcription occur throughout their life cycle. In *E. coli* the replication fork and the RNA polymerase progress at 600-1000 nt/s and 30-80 nt/s, respectively (Rocha 2008). Since both polymerases are bound to the same template and replication and transcription occur simultaneously in dividing cells, collisions between them are inevitable. These collisions can happen head-on, if the transcribing gene is in the lagging strand and, be co-oriented, if the transcribing gene is on the leading strand (Mirkin and Mirkin 2005; Rocha 2008). Collisions between DNA and RNA polymerases happen both in head-on and co-oriented directions. In the co-oriented direction collision happens due to differences in the speed of the two polymerases (Rocha 2008). It has been suggested that head-on collisions more frequently lead to arrest of replication fork and they may also result in higher local mutagenesis (Mirkin and Mirkin 2005; Rocha 2008). In order to understand whether SSR mutations too are affected by the head-on/co-oriented direction of transcription and replication, I calculated the % of PSSRs to % of SSRs (Fig. 5.4) in both the directions. It can be seen from **Figure 5.4** that in most of the species the ratio of percentages of PSSRs to SSRs is more than one in the head-on direction whereas it is less than one in the co-oriented direction. The significance of this enrichment was calculated and it was found that in six of these species it is highly significant. The non-uniformity in significance in different species can be attributed to the role of selection in the coding regions of a genome. This differential tendency of SSR mutations can be attributed to the higher gene density on leading strand of DNA replication as compared to the lagging strand

(Nilsson et al. 2005; Wang, Berkmen, and Grossman 2007; Rocha 2008). It has also been shown in the case of *E. coli* that more than 80% of essential genes are on the leading strand.

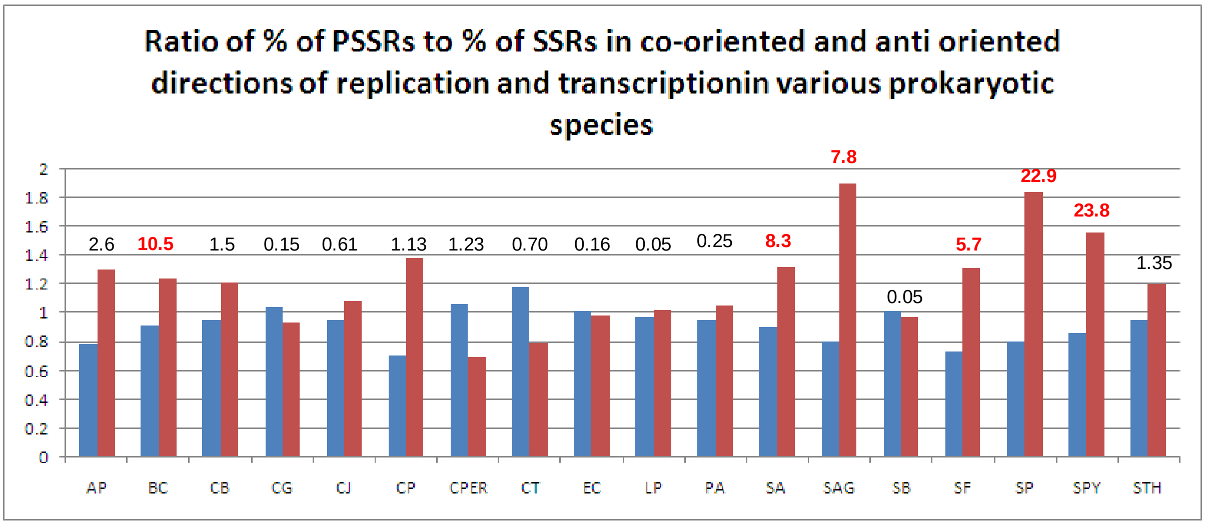


Figure 5.4: Ratio of % of PSSRs to % of SSRs in the co- (blue) and the head-on (magenta) directions of replication and transcription in various prokaryotic species. The chi-square values are given on top of magenta bars. Significant chi-square values are shown in red color.

5.3.3 Intra-genic distribution of PSSRs

Frame shift mutation due to SSR in a gene can affect the structure and function of its protein product. However these effects vary depending on the position of mutation relative to the start of ORF. Mutation in ORF at different positions can affect the fitness of an organism differently depending on the toxicity of the mutated product, which is in general harmful to an organism. If a mutation is random in the coding region, one would expect an unbiased distribution of PSSRs along the length of ORFs. To find, the distribution of PSSRs (which are able to shift the reading frame of ORFs) along the intragenic location, the ORF was divided into three equal parts (5’ (0-33%), middle (34-66%) and 3’ (67-100%)) and the number of PSSRs found in these regions were calculated and compared with the numbers expected under the assumption that the three parts are equally populated by PSSRs (**Figure 5.5)**. Interestingly the middle part harbors lower number of PSSRs than expected and both the 5’ and 3’ parts harbor more number of PSSRs than expected. Between 5’ and 3’ parts it is the former which is more significantly populated with PSSRs (P-value = 0.0001) than the latter part.

I did not investigate further into the effects of the PSSRs. As presented in Chapter 4, in case of Yersinia genomes majority of PSSRs in the N-terminal and in the middle of ORFs are involved in premature termination of genes whereas majority of C-terminal PSSRs are involved in length variation of ORFs. This observation hints at a possibility that similar effects could be seen on genes in the 5’, middle and 3’ parts. However, I would like to make a special mention with regard to genes harboring PSSRs in their 5’ parts. It is to be noted that an ORF in a genome has been tuned in such a way that frame shift mutations result in an early stop codon (Itzkovitz and Alon 2007). In real genes a stop is encountered 15 codons after a frame shift event, however in a randomized code a stop signal is encountered after 23 codons (Itzkovitz and Alon 2007). This apart, an abundance of PSSRs in 5’ part as compared to the middle part can arise because of costs associated with translating useless sequences. One might expect that selection favors mutations that diminish translational costs by producing maximally truncated proteins. In addition to this the reduction in the length of the frame shifted peptide can also have beneficial effects, such as reducing the potential toxicity of the truncated proteins and reducing the chances of misfolded protein aggregates.

However, some exceptions to premature terminations have been found as a consequence of multiple start codons at the 5’ side of the corresponding genes (Dixon et al. 2007). For example, Mycobacterial genomes harbor 14 secretory proteins which only lack their N-terminal signal peptides due to PSSRs in some of the strains (Sreenu et al. 2006). As detailed in the previous chapter Yersinia genomes too harbor such genes. To summarize it can be concluded that PSSR at the 5’ side can serve two purposes; (a) converting a gene into a pseudogene with least toxic effect of mutated product and (2) converting secretory protein into non-secretory protein thereby affecting its sorting. This, in particular, may be beneficial to pathogens because the protein is no longer secreted out and therefore becomes hidden from the host-immune system.

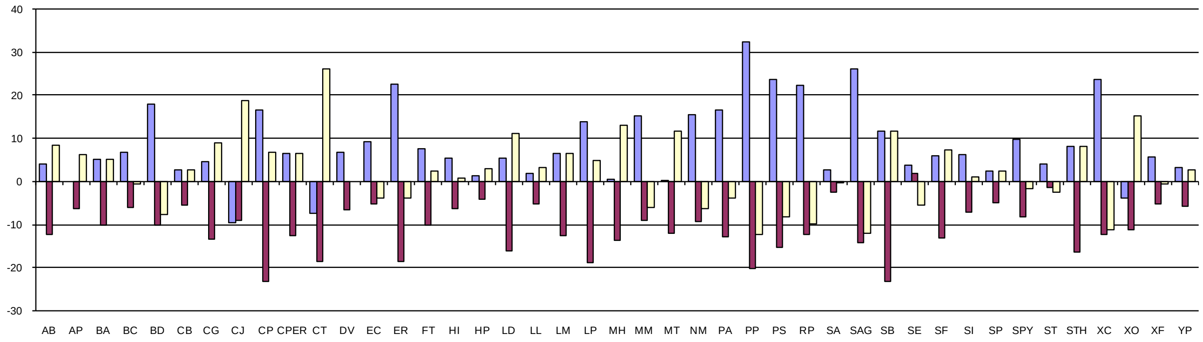


Figure 5.5: Deviations of observed percentage of FS-PSSRs from the expected percentages of PSSRs found at 5’ (blue), middle (magenta) and 3’ (yellow) parts of genes.

5.3.4 PSSRs in the host adapted pathogens

The tract density profile of PSSRs (Number of PSSRs per Mb per strain) in various prokaryotic species is shown in **Figure 5.6**. It can be seen that non-pathogenic species (shown with blue bar) have low tract densities of PSSRs where as the host-adapted pathogenic bacteria (shown with 'magenta bar') have high tract densities of PSSRs. High incidences of SSR polymorphism in the pathogenic bacteria is interesting and perhaps indicate a need-based variability related to their survival strategies in host-environments.

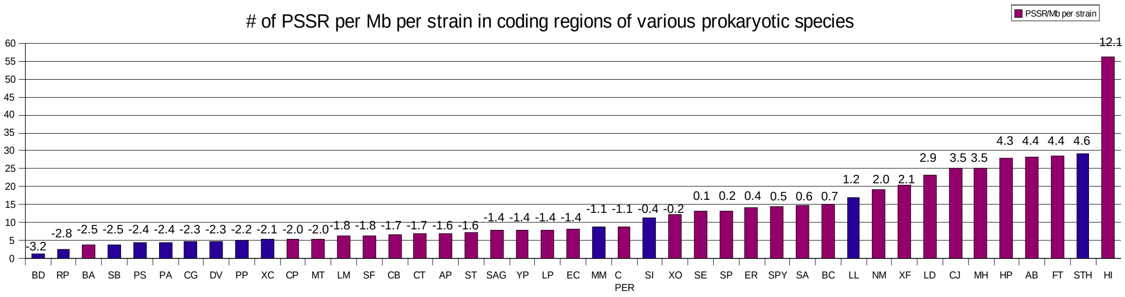


Figure 5.6 Tract density of coding PSSRs in various prokaryotic species. The non-pathogen or plant pathogenic species harbor less number of PSSRs as compared to pathogenic species. Z-score is given on the top of each bar. Pathogens and non-pathogens are shown in magenta and blue bar respectively.

5.4 Discussion

The discovery of PSSRs in all the 43 species considered in the present study supports that replication slippage occurs in all species of prokaryotes and contributes to the intraspecies genetic diversity. Mono nucleotide SSRs despite their frame shifting ability are the most abundant PSSRs in the coding regions although tracts longer than 8bp are rarely found. Different selective constraints operate in the coding and the non-coding regions which are well reflected in the number of PSSRs found in those regions. One would expect avoidance of frame-shift PSSRs in protein coding regions however abundance of mononucleotide tracts seem to reveal a contrasting picture that bacterial genomes, in general, prefer frame-shift mutations.

I would like to mention here that in all the cases of inter-strain comparisons, other than INDELs in SSRs, the compared ORFs are ~100% identical and therefore any effect that is seen is purely a consequence of the SSR mutations. However, SSR mutations are often reversible. Hence it is quite possible that premature termination of coding regions, can either become on or off after gaps of some generations. This could be said because in general, pseudo genes are non functional regions of genomes, and get actively removed to save resources (Mira, Ochman, and Moran 2001; Nilsson et al. 2005). However the cross-genome comparisons carried out in this study reveal that pseudo genes arising out of frame shift mutation of SSRs are conserved across the different strains indicating the possible utility of these genes in a dynamic manner.

PSSRs are relatively more abundant in head on orientation than co-oriented direction of replication and transcription. This differential abundance can be attributed to higher gene-density on the leading strand as compared to the lagging strand.

Within intra-genic regions PSSRs show positon-specific biases. 5’ and 3’ parts are more abundant with PSSRs than middle parts. Abundance of PSSRs, at 5’ parts which leads to premature terminations seems to be the direct consequence of reduced translation costs involved in translating a useless gene. An exception to this is found in one of the special cases of some secretary proteins having alternative RBS start sites which give rise to protein products without their signal peptides. Low abundance of PSSRs in the middle of ORFs can be argued as a consequence of avoidance of the production of toxic truncated non-functional proteins. PSSRs in 3’ parts are less deleterious as they lead to only small length variations in protein products.

The high abundance of PSSRs in host adapted pathogens such as *Haemophilus influenzae*, *Acinetobacter baumannii* etc. which are believed to be extracellular and therefore exposed to host-immune systems is the consequence of need-based variability required in pathogen-host interactions. Benefits of PSSRs in pathogens depend on the degree to which they are exposed to host immune systems. All the above observations and arguments lead to the assumption that SSR evolution in the coding regions of bacterial genomes is non-neutral.

5.5 Summary

In this chapter, details of cross-genome comparisons of all available strains belonging to 43 species were reported. Our studies reveal that on an average each species harbours a few hundreds of PSSRs. Most of the PSSRs have shown an increment or decrement of one repeat unit. Non-coding regions are more abundant with PSSRs than coding regions. Less abundance in intragenic regions as compared to the intergenic regions reconfirms our earlier observation in Yersinia genomes and further suggests that “selection against frame shift mutations” is a global phenomenon in prokaryotic genomes. Furthermore, it was found that the intragenic regions where replication and transcription directions are opposite to each other (collision is head on) harbor relatively more number of PSSRs than the intragenic regions where replication and transcription directions are same (collision is co-oriented). I further found that PSSRs are mostly found in the 5’- and 3’- parts of the genes as compared to their middle parts. The intragenic regions in the genomes of host adapted pathogens (animal pathogen (mostly in human)), are denser with PSSRs as compared to free living and plant pathogens. All these observations lead us to assume that SSR evolution in the coding regions is non-neutral as these play important roles in adaptation and pathogenesis of bacteria.

Chapter 6

A Study on Mutational Dynamics of Simple Sequence Repeats in the intergenic regions of Prokaryotic genomes

6.1 Introduction

Replication slippage is the main mechanism of SSR mutation where slippage on the template strand leads to a contraction (deletion of repeat units) of SSRs whereas slippage on the nascent strand manifests as an expansion (insertion of repeat units) of SSRs (Streisinger et al. 1966; Levinson and Gutman 1987a; Harr, Todorova, and Schlotterer 2002; Mirkin 2005; Garcia-Diaz and Kunkel 2006). A survery of the published literature (details are given in Chapter 1) reveals that SSRs show mutational bias towards contraction or expansion, which is related to the presence or absence of mis-match repair system (MMRS) in both eukaryotes as well as prokaryotes. However, the published reports have not been in unison to state clearly the exact relationship between the mutational bias of SSRs and presence/absence of MMRS (Henderson and Petes 1992; Metzgar et al. 2002). Therefore, this aspect was investigated in the cases of non-coding PSSRs discovered in the 43 species mentioned in the last chapter. The non-coding PSSRs were considered because they are assumed to be selectionally neutral as compared to the coding PSSRs.

6.2 Methods

6.2.1 Systematic search for MMR system in the 43 species of prokaryotes

The details of the acquisition of annotated whole genome sequences of prokaryotic genomes have already been given in the last chapter. Hence, only the other details relevant to the present work are given here. As the main aim of this work was to find a relationship between the presence/absence of MMRS and the mutational bias in the non-coding PSSRs, the annotations of the genome sequences for the presence of MutS and MutL were searched. MutH was not considered because its universal presence as a part of the MMR system in the prokaryotic genomes has been doubted (Claverys and Lacks 1986). Furthermore, the search was limited to the presence of MutS1 amongst the MutS homologues as it is the one that is involved in MMR. Therefore here MutS1 has been refered as MutS unless otherwise mentioned specifically.

After searching for the annotation, the translated protein sequences of the ORFs annotated as MutS and MutL were searched for the presence of functionally conserved domains using Pfam (Finn et al. 2008) and SMART databases (Schultz et al. 1998). Only those sequences containing all the characteristic functional conserved domains were considered as MutS and MutL homologues. Genomes where MutS and MutL homologues were absent and those in which the homologues were found but missing one of the characteristic functional domains were subjected to tblastn searches using full length MutS and MutL ORFs as queries. This enabled us to discover in some genomes, DNA sequences very similar to the full length MutS/MutL ORFs but containing frameshift mutations causing premature terminations. Those genomes which failed to yield sequences similar to MutS/MutL were further subjected to profile based searches using PROFILE-SS to reconfirm the complete absence of the two genes. The profiles for the two proteins were constructed from their respective multiple sequence alignments of all the homologues gathered by us. All these results pertaining to presence/absence of MutS/MutL homologues were also reconfirmed by comparing with the published literature.

6.2.2 Extraction of non-coding PSSRs

The whole genome sequences were screened for extracting PSSRs using PSSRFinder as already detailed in the previous chapter. As stated in the previous chapter, PSSRs only if harbored within the non-coding regions in all the compared strains were categorized as the non-coding PSSRs. Of these PSSRs, only those showing two polymorphic states (i.e., two alleles) were considered and among these equally populated allelic cases were discarded. In cases (species) where multiple sub-strains were available only one sub-strain was considered for analysis.

6.2.3 Identification of contractions/expansions in PSSRs in a species

In order to decide whether a given PSSR in a species is a case of expansion or contraction the length distribution of PSSRs across strains was examined. If the most abundant allele is longer than the other allele then the PSSR was considered as a case of contraction whereas if the most abundant allele is shorter than the other allele then PSSR was considered as expanded. However in the case of Yersinia species the status of PSSRs as expansion or contraction was decided with reference to its ancestor *Yersinia pseudotuberculosis*. In addition to this only those PSSRs were considered where the reference alleles (the most abundant or ancestor allele) were at least three repeat units long.

The statistical significance of SSRs expansion/contraction was tested using Chi-square test. As it is a case of one degree of freedom, any chi-square value greater than 3.84 is significant. Chi-square value is calculated as given below.

$$\chi^2 = \sum_{i=1}^{k}(O_i - E_i)^2/E_i$$

Where, O_i = observerd number of events, E_i = expected number of events (total number of contraction and expansion divided by 2), i =1, k = 2 (only two events are possible (expansion or contraction)).

6.3 Results and Discussions

6.3.1 MMRS in bacterial and archaeal genomes

Our search for the MutS and MutL gene in the bacterial and archaeal genomes yielded three groups: (a) species where one of the strains lacks MMRS (**Table 6.1);** (b) species where all the strains lack MMRS (**Table 6.1)** and (c) species where all the strains have MMRS**.** Additional information such as protein IDs of MutS and MutL of each strain are given as **Supplementary Table 6.1** at the end of the chapter.

Table 6.1: Species where *MutS* and *MutL* genes are present ('+') and absent ('-'). In Acinetobacter *baumannii Haemophilus influenzae* and *Staphylococcus aureus* one of the strains lacks functional MMRS because of a frameshift mutation and therefore the corresponding strains are MMRS deficient.

A				B			
Species Name	Name of strain(n)	MutS	MutL	Species Name	#strains	MutS	MutL
Acinetobacter baumannii	AB0057, AB307_0294, ACICU, and SDF	+	+	*Campylobacter jejuni*	5	-	-
	ATCC_17978	-	-	*Helicobacter pylori*	6	-	-
Haemophilus influenzae	RD_KW20, 86_028NP and PittEE	+	+	*Mycobacterium tuberculosis*	3	-	-
	PittGG	-	+	*Mycobacterium bovis*	3	-	-
Staphylococcus aureans	MRSA, MSSA476, MW2, JH1, Mu3, Mu50, N315, COL, Newmann, NCTCT, USA300	+	+	*Mycoplasma hyopneumoniae*	3	-	-
	RF122	+	-	*Methanococcus maripaludis*	4	-	-
				Sulpholobus islandicus	6	-	-

(a) Species where one of the strains lacks MMRS

Of all the species that were screened it was found that one of the strains of *Haemophilus influenza* and *Acinetobacter baumannii* (pittGG strain in *Haemophilus influenzae*, and ATCC 17978 strain in *Acinetobacter baumannii*) lacked functional MutS due to a frameshift mutation. In *Haemophilus influenzae* pittGG strain, MutS is prematurely terminated due to the deletion of one repeat unit in a seven bp mono nucleotide, a poly “A” SSR tract positioned at 2155 bp as compared to its homologs in the other strains of *Haemophilus influenzae*. Similarly in *Acinetobacter baumannii* ATCC 17978 strain MutS is frame shifted at position 1989 due to the insertion of one nucleotide in the mono nucleotide tract of a poly “A” tract, which is 6bp in length in all other strains of this species. Due to this frame shift the ORF has split into two parts, A1S_1251 (573) and A1S_1252 (206). In the ORF A1S_1251, the ATPase and MutSIII domains are intact while in the second ORF the MutSII domain which is believed to be a connector domain between mismatch recognition domain and ATPase domain (Lamers et al. 2000; Obmolova et al. 2000), has been deleted. Although it is annotated as a coding region in the genome annotation file (Smith et al. 2007), splitting of domain suggests that the MutS protein may not be functional in this strain. In addition to this it was also found that in *Acinetobacter baumannii* and *Staphylococcus aureans*, one of the strains (ATCC 17978 strain in *Acinetobacter baumannii* and RF122 strains of *Staphylococcus aureans*) lacked functional MutL due to a frameshift mutation. In *Acinetobacter baumannii*, ATCC 17978 *MutL* gene is frame shifted due to the deletion of one “A” in a poly “A” tract of 7bp. In *Acinetobacter baumannii* wild type gene of *MutL* codes for a 650 amino acid long protein and this due to frame shift mutation in strain ATCC 17978 has resulted in two ORFs of which the second ORF that translates into a 369 amino acid sequence corresponds to a truncated TopoII domain which is a mismatch repair domain. On the other hand in *Staphylococcus aureans* RF122 strain *MutL* gene is frame shifted due to the deletion of 1bp at position 1450 in a non SSR tract and hence the homologous region in the genome is annotated as a non-coding region.

(b) Species where all the strains lack MMRS

In addition to the above mentioned species, seven other species were also found (two belonging to archaea and five belonging to bacteria) where both *MutS* and *MutL* genes (the complete DNA sequence) were absent in all their known strains (**Table 6.1**) thus contributing as suitable systems to study SSR mutations in the light of MMRS deficiency. Among these, MMR deficiency in five of the species viz., *Mycobacterium tuberculosis, Mycobacterium bovis*, *Mycoplasma hyopneumoniae*, *Helicobactor pylori* and *Campylobactor jejuni* has already been reported (Eisen 1998; Mizrahi and Andersen 1998; Carvalho et al. 2005; Lin, Nei, and Ma 2007; Dos Vultos et al. 2009).

The phylogenetic relationships among the MMRS proficient and deficient species is shown in **Figure 6.1**. A phylogenetic tree was constructed based on their 16S rRNA sequences which showed clustering of MMRS deficient species into four distinct groups indicating that the loss of MMRS is not an individual character but inherited from a common ancestor that got selected for a loss of MMRS.

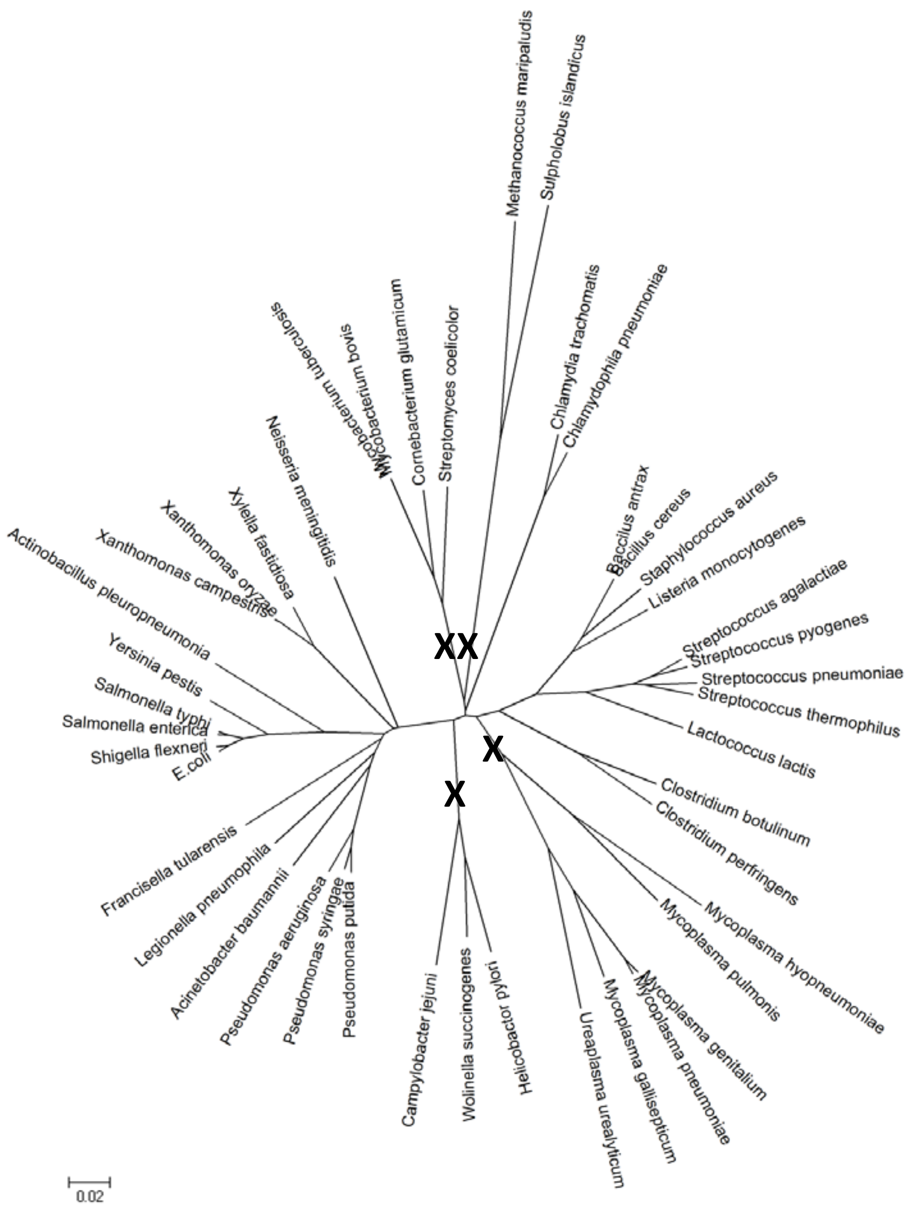


Figure 6.1: Phylogenetic relationship among all the species considered in this work. It is to be noted that the tree also includes some species which were not included in the present analysis because they were not meeting the required criteria as mentioned in methods section. These extra species were included in the tree because they are MMRS deficient and cluster together with other MMRS deficient species. It can be seen that the MMR deficient species are clustered together (marked by "X"). The tree was constructed using the 16S rRNA sequences. Multiple sequence alignment was generated by CLUSTALX. Phylogeny was constructed by using the PHYLIP package. The unrooted tree was draw using drawtree. Boot strap value was 1000.

6.3.2 Non-coding PSSRs

About 10543 non-coding PSSRs were found from 36 species (each species with a minimum of three strains) of which a large majority was constituted by mono nucleotide tracts (**Fig. 6.2**). The number of PSSRs of each repeat type in different species is given in **Supplementary Table 6.2** (at the end of the chapter). Most of the PSSRs (99%) have undergone expansion or contraction of one repeat unit. Since 95% of the PSSRs are mononucleotide tracts only these tracts were considered for further studies.

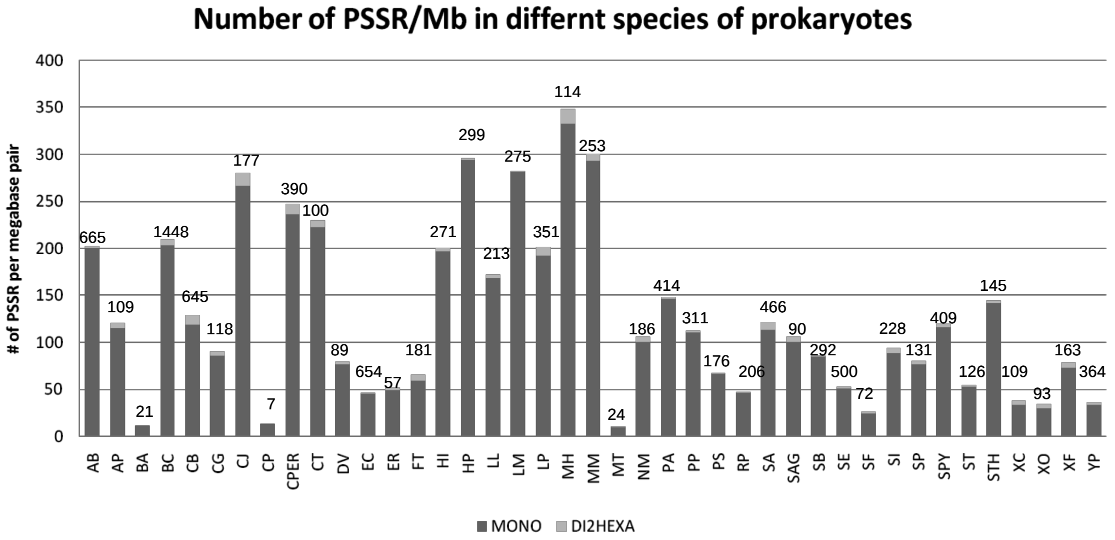


Figure 6.2: Number of PSSRs/Mb/strain found in the non-coding regions of different prokaryotic species. In each bar the darkly shaded portion represents mono nucleotide SSRs and the lightly shaded portion represents the other SSRs (di to hexa).

(Foot note) AB = *Acinetobacter baumannii*, AP = *Actinobacillus pleuropneumoniae*, BA = *Bacillus anthracis*, BC = *Bacillus cereus*, CB = *Clostridium botulinum*, CG = *Corynebacterium glutamicum*, CJ = *Campylobacter jejuni*, CP = *Chlamydophila pneumoniae*, CPER = *Clostridium perfringens*, CT = *Chlamydia trachomatis*, DV = *Desulfovibrio vulgaris*, EC = *Escherichia coli*, ER = *Ehrlichia ruminantium*, FT = *Francisella tularensis*, HI = *Haemophilus influenzae*, HP = *Helicobacter pylori*, LL = *Lactococcus lactis*, LM = *Listeria monocytogenes*, LP = *Legionella pneumophila*, MH = *Mycoplasma hyopneumoniae*, MM = *Methanococcus maripaludis*, MT = *Mycobacterium tuberculosis and Mycobacterium bovis*, NM = *Neisseria meningitidis*, PA = *Pseudomonas aeruginosa*, PP = *Pseudomonas putida*, PS = *Pseudomonas syringae*, RP = *Rhodopseudomonas palustris*, SA = *Staphylococcus aureus*, SAG = *Streptococcus agalactiae*, SB = *Shewanella baltica*, SE = *Salmonella enterica*, SF = *Shigella flexneri*, SI = *Sulpholobus islandicus*, SP = *Streptococcus pneumoniae*, SPY = *Streptococcus pyogenes*, ST = *Salmonella typhi*, STH = *Streptococcus thermophilus*, XC = *Xanthomonas campestris*, XO = *Xanthomonas oryzae*, XF = *Xylella fastidiosa*, YP = *Yersinia pestis.*

6.3.3 MMR deficiency leads to destabilization of SSR tracts

The number of PSSRs found in *Haemophilus influenza*, *Acinetobactor baumannii* and *Staphylococcus aureans* are shown in **Figure 6.3**. The MMRS deficient strains show significantly higher (Z>4) number of PSSRs as compared to the MMRS proficient strains indicating MMRS deficiency has led to destabilization of SSR tracts. This is in agreement with earlier reports (Levy and Cebula 2001; Watson, Burns, and Smith 2004). It is pertinent to note that MutS/MutL genes themselves harbor SSR tracts and as already noted the MMRS deficiency has arisen because of the frameshift mutations caused by SSR tracts and therefore in these species MMRS deficiency/proficiency is a matter under the control of SSR mutations. Unless fixed, it is quite probable that mutating SSRs yield functional MMRS for some generations and non-functional MMRS for some other generations. Evidence to this argument can be seen from **Figure 6.3** where some of the strains like ACICU of *Acinetobacter baumannii* and pittEE strain of *Haemophilus influenzae* also show high levels of SSR instability (however with low significance, with a Z-score<4.0) despite being MMRS proficient.

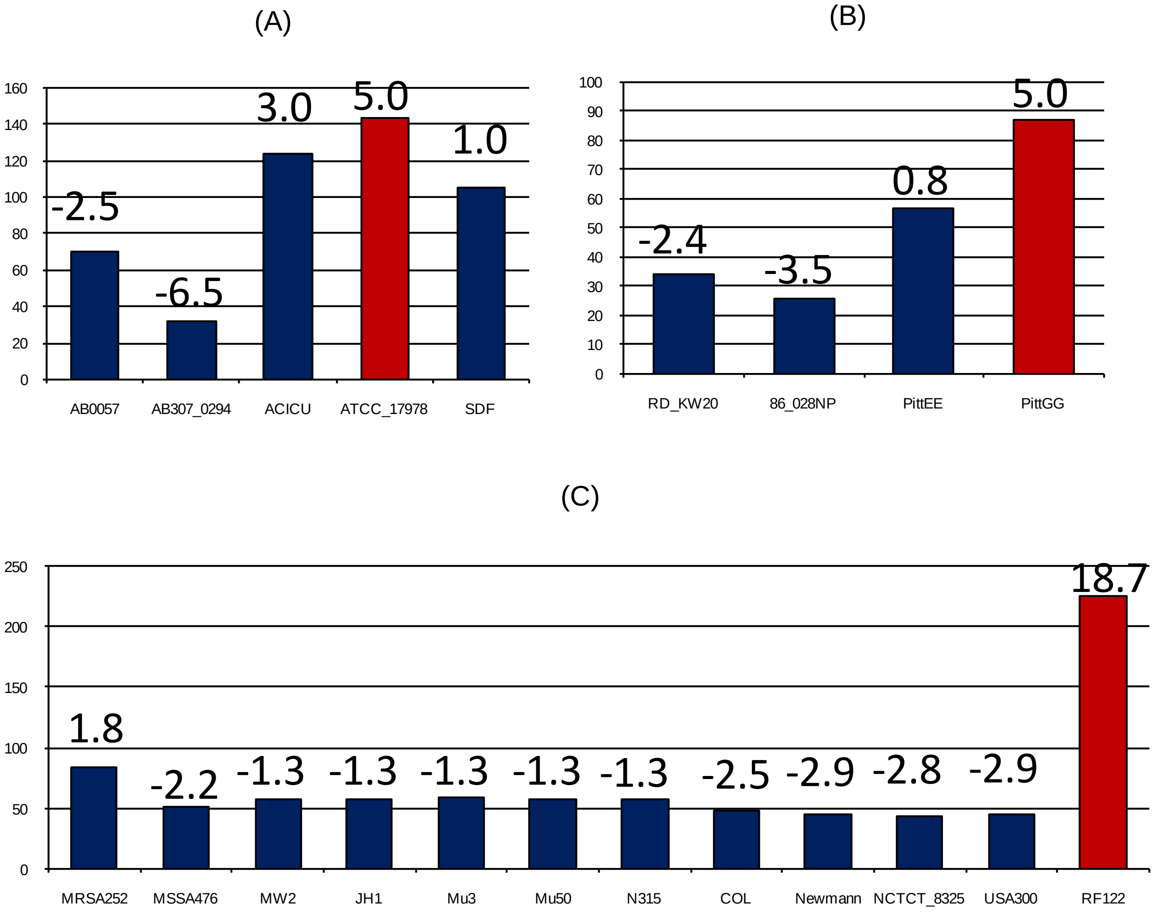


Figure 6.3: The number of PSSRs (Obs) found in each strain of (A) *Acinetobacter baumannii* and (B) *Haemophilus influenzae* (C) *Staphylococcus aureus*. The MMRS deficient and proficient strains are shown in red and blue bars respectively. Z-scores $(Z - Score = \frac{(Obs-Exp)^2}{Exp})$ indicated above the bars. Average number of PSSRs in each species was considered as expected number (Exp) of PSSRs in each species.

The PSSRdensities (the number of PSSRs per Mb per strain) found in the MMRS deficient species are shown in **Figure 6.4**. It is pertinent to note that a threshold has not been defined for PSSR density to discriminate species as ‘high’ or ‘low’ in PSSRs. However, as the PSSR density relates to SSR stability it can be said that higher the PSSR density lower is the SSR stability. From the figure it can be seen that four out of the six MMR deficient species (*Mycoplasma hypopneumoniae, Methanococcus maripalusis, Helicobacter pylori* and *Campylobacter jejuni*) are positioned at the high end of the PSSR tract density chart. The other deficient species viz., Mycobacterium (includes both *Mycobacterium tuberculosis* and *Mycobacterium bovis*) and *Sulpholobus islandicus* harbor relatively low density of PSSRs and seem ‘outliers’ amongst the MMRS deficient species. Of these outliers, Mycobacteria are known for low mutation rates which have been linked to the slow rate of DNA synthesis in these bacteria (Hiriyanna and Ramakrishnan 1986; Dos Vultos et al. 2009). Slow DNA synthesis may promote increased fidelity of the DNA polymerases even in the MMR-deficient background (Radman 1998). The other outlier *Sulpholobus islandicus* is also a slow growing bacterium with a doubling time of 7-8 hours. From the above observations it can be concluded that MMRS deficiency, except in the case of slow growing organisms, is associated to the high incidences of PSSRs. In other words MMR deficiency leads to the destabilization of SSRs.

However, I would like to point out an observation made in *E. coli* which is MMRS proficient species that F-plasmid can destabilize SSR tracts (Schlotterer et al. 2006). Given this, it can be assumed that MMR proficient species too can harbor high number of PSSRs as a consequence of yet to be discovered mechanisms some of which may be species specific.

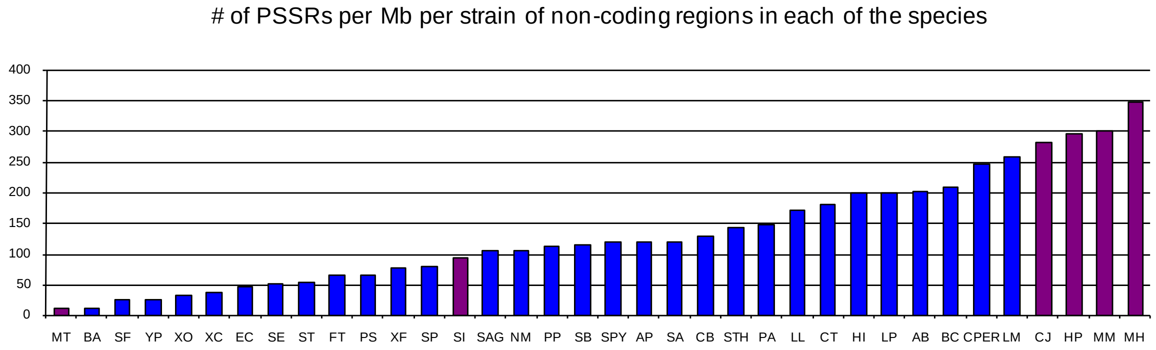


Figure 6.4: Number of PSSR per Mb per strain in various species of prokaryotes. Magenta = MMR deficient species. Blue = MMR proficient species.

6.3.4 Mismatch repair is biased

During replication, there is an equal probability for slippage error to occur on both nascent as well as template strands and therefore one can expect equal number of contraction and expansion events in PSSRs. Results shown in **Table 6.2** reveal that PSSRs are contraction biased in MMR proficient strains (*Haemophilus influenzae*, *Acinetobactor baumannii* and *Staphylococcus aureus)*. On the other hand MMR deficient strains do not show such a significant bias. These results suggest that mismatch repair system (MMRS) is less efficient in repairing slippage mutation on the template strand as compared to nascent strand. It can also be argued that this contraction bias might also arise due to high frequency of primary mutations (slip out) in the template strand as compared to nascent strand. In order to test which one of the aforementioned arguments is correct, the contraction/expansion bias of PSSRs in the species lacking MMRS was examined.

Table 6.2: PSSRs found in MMR proficient and deficient strains of *Acinetobactor baumannii, Haemophilus influenza* and *Staphylococcus aureans.* Number of instances of PSSR expansions and contractions are also given. The significant one is indicated as bold. "NS" means not significant.

Species Name	Name of strain(s)	#PSSR	Cont	Exp	Cont/Exp	χ2 1:1	P-value
Acinetobactor baumannii	AB0057, SDF, AB307_0294, ACICU and SDF	331	180	132	1.36	7.38	**<0.006**
	ATCC_17978	139	76	63	1.20	1.22	NS
Haemophilus influenzae	RD_KW20, 86_028NP and	103	71	32	2.21	14.17	**0.0002**
	PittGG	79	43	36	1.2	0.62	NS
Staphylococcus aureus	MRSA252, MSSA476, MW2, JH1, Mu3, Mu50, N315, COL, Newmann, NCTCT, USA300	312	183	129	1.41	9.35	**0.002**
	RF122	180	89	91	1.0	0.02	NS

6.3.5 Primary SSR mutations are biased towards expansion

The mutations observed in the absence of MMRS are referred to as the primary SSR mutations. **Table 6.1** gives the list of species lacking MMR system. The number of PSSRs in all the five species (*Campylobacter jejuni, Helicobacter pylori, Mycoplasma hypopneumoniae, Methanococcus maripalusis* and *Sulpholobus islandicus*) along with the number of times they were found contracted and expanded are given in **Table 6.3**. It can be seen from the table that most of the MMR deficient species show more number of expanded PSSRs than contracted PSSRs indicating that primary mutations are biased towards expansions. However the significance of this bias varies between the species. It can be seen in the case of *Mycoplasma hypopneumoniae* that the SSR expansion is significantly biased. As a consequence of bias towards expansions one can expect presence of long SSR tracts in the MMR deficient species as compared to MMR proficient species. In order to find this, tract density of mono SSRs >8bp per Mb per strain in all the genomes was calculated (**Figure 6.5**). In fact in three out of five species (*Campylobacter jejuni, Helicobacter pylori* and *Mycoplasma hypopneumoniae*) which lack MMRS the genomes harbor very large number of long SSR tracts as compared to the MMRS proficient species (magenta color).

Table 6.3: Number of PSSRs found in MMR deficient species. Table also show the number of expansion and contraction events in each species. "*" = Archaea, Cont = contraction, Exp = expansion. Significant one is shown in bold. "NS" means not significant.

Name of species	#strain	GC%	Size (10^6)	#PSSR	Cont	Exp	Cont/Exp	χ2 1:1	P-value
Campylobacter jejuni	5	31	1.7	153	67	86	0.78	2.36	NS
Helicobacter pylori	6	39	1.6	189	96	93	1.01	0.05	NS
**Methanococcus maripaludi*	4	33	1.7	174	84	90	0.93	0.21	NS
Mycoplasma hypopneumoniae	3	29	0.9	80	28	52	0.53	7.20	**0.007**
**Sulpholobus islandicus*	6	35	2.7	141	61	80	0.76	2.56	NS

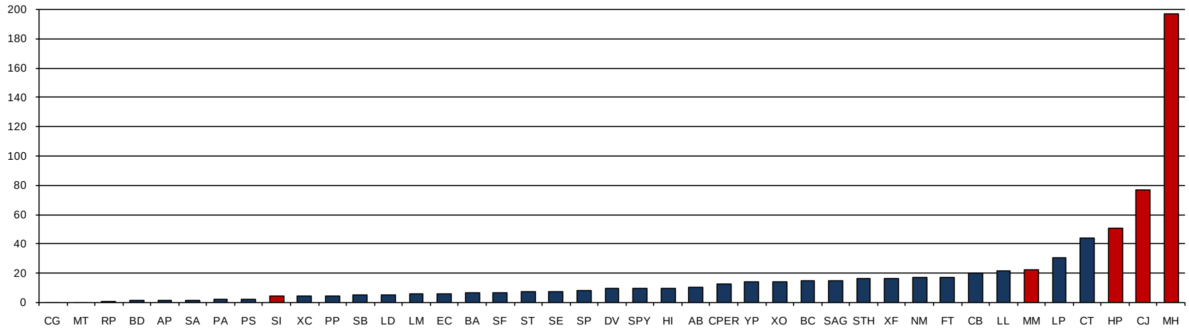


Figure 6.5: Tract densities (the number of SSRs per Mb per strain) of mononucleotide SSRs longer than 8bp in different prokaryotic species. MMRS deficient are shown in magenta color bar whereas MMRS proficients are shown in blue bar.

6.3.6 PSSRs are contraction biased in MMRS proficient species

Mutations in SSRs which escape MMR system are referred to as secondary mutations. The direction of these secondary mutations (PSSRs) in MMRS proficient species was also examined in order to check whether they are contraction oriented or expansion oriented. The number of PSSRs found in MMR proficient species examined in this study is given in **Table 6.4**. It can be seen that in most of the species PSSRs are significantly contraction biased.

Table 6.4: The number of PSSRs found in MMR proficient species. Cont = contraction, Exp = expansion.

Name of species	#strain	GC%	Size (10^6)	#PSSR	Cont	Exp	Con/Exp	χ2 1:1	P-value
Actinobacillus	3	41	2.2	94	53	41	1.30	1.53	0.21
Bacillus cereus	8	35	5.1	962	526	436	1.20	8.42	0.0037
Clostridium botulinum	7	28	3.8	349	201	148	1.36	8.05	0.0046
Clostridium perfringens	3	28	3.0	275	156	119	1.31	4.98	0.025
Chlamydia trachomatis	6	41	1.04	87	50	37	1.35	1.94	0.16
Escherichia coli	21	51	5.0	499	297	202	1.47	18.09	0.0001
Francisella tularensis	7	32	1.9	143	88	55	1.6	7.62	0.0058
Lactococcus lactis	3	36	2.4	95	53	42	1.26	1.27	0.26
Listeria monocytogenes	4	38	2.9	204	116	88	1.31	3.84	0.05
Legionella pneumophila	4	38	3.5	199	120	79	1.52	8.45	0.0037
Neisseria meningitidis	4	52	2.2	99	55	44	1.25	1.22	0.27
Pseudomonas aeruginosa	4	66	6.5	288	178	110	1.6	16.06	0.0001
Pseudomonas putida	4	62	6.0	150	90	60	1.5	6.00	0.014
Pseudomonas syringae	4	66	6.5	106	74	32	2.31	16.64	0.0001
Streptococcus agalactiae	3	36	2.1	74	41	33	1.24	0.86	0.35
Shewanella baltica	4	46	5.2	231	151	80	1.9	21.82	0.0001
Salmonella enteric	13	52	4.7	463	269	194	1.39	12.15	0.0005
Shigella flexneri	3	51	4.6	51	25	26	1.00	0.02	0.88
Streptococcus pneumoniae	6	40	2.1	106	62	44	1.40	3.06	0.08
Streptococcus pyogenes	13	39	1.9	397	214	182	1.18	2.59	0.10
Salmonella typhi	3	52	4.8	113	48	65	0.73	2.56	0.10
Streptococcus	3	39	1.8	109	60	49	1.22	1.11	0.29
Xanthomonas campestris	4	65	5.1	72	53	19	2.7	16.06	0.0001
Xanthomonas oryzae	3	64	5.0	57	25	32	0.78	0.86	0.35
Xylella fastidiosa	4	52	2.6	92	49	43	1.13	0.39	0.53
Yersinia pestis	20	48	4.6	282	164	118	1.40	7.50	0.006

6.3.7 PSSR contraction bias is independent of repeat count

In several studies it has been reported that the SSRs with high repeat counts experience a downward mutation bias whereas those with low repeat counts are prone to expansions (Xu, Peng, and Fang 2000; Huang et al. 2002**.** However, the present data (**Figure 6.6**) does not support this observation. It can be seen that in general the SSR mutation is biased towards contraction irrespective of repeat count suggesting that the mutational bias of a SSR tract is related to MMRS rather than its repeat count.

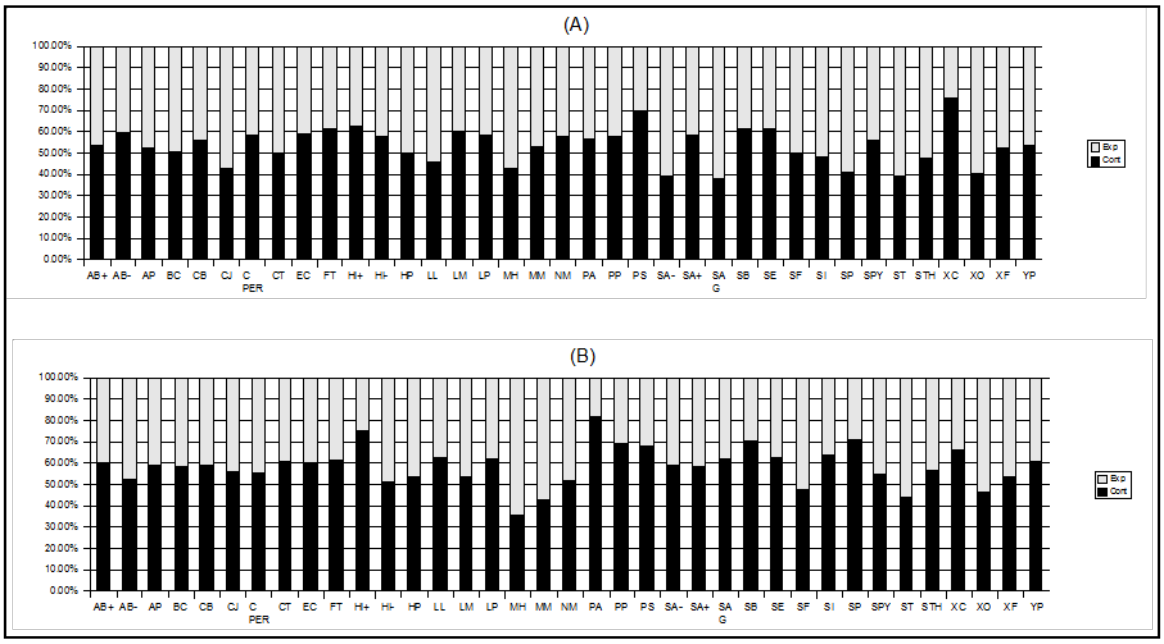


Figure 6.6: An Illustration of contraction bias of SSR mutations is not related to repeat count. Count less than equal to 5bp (A) and count >=6bp (B).

6.3.8 Directionality of SSR mutation is independent of sequence composition of genome

Table 6.5 gives a comparison of the directionality of SSR mutations between the MMR proficient and deficient species having similar GC content of the genome. It can be seen that the genomes having similar genome compositions have different directionality of SSR mutations and hence it can be argued that the observed differences in the directionality of these mutations between the MMR proficient and deficient species is not dependent on composition bias of the genomes but is due to the presence or absence of the MMRS. The general scarcity of long tracts in MMR proficient species indicates the role of MMR system in restricting SSR tract expansions in prokaryotes.

Table 6.5: Comparison between MMR proficient and deficient species with similar GC percentage.

Species Name	GC%	Cont/Exp	χ2 1:1	Species Name	GC%	Cont/Exp	χ2 1:1
HP	39	1.01	0.05	LM	38	1.31	3.84
CJ	36	0.78	2.36	SAG	36	1.6	16.06
MM	33	0.93	0.21	SA	33	1.41	9.35
MH	29	0.53	7.20	CPER	28	1.31	4.98
SI	35	0.76	2.56	BC	35	1.20	8.42

6.4 Summary

In this chapter relationship between mutational bias of SSRs towards expansion or contraction in relation to presence and absence of MMR system in prokaryotes was investigated. These investigations revealed that the MMR deficiency as a consequence of frameshift mutation in MutS has led to an increase in the number of SSR tracts undergoing slippage related INDEL mutations. It was also found that species lacking MutS have significant mutational bias towards expansion of SSRs. On the other hand species where MMRS is present the SSR mutations showed significant bias towards deletions. Furthermore it was also found that contraction bias is independent of sequence and length of SSRs. The presence of large number of long tracts in MMRS deficient species as compared to MMRS proficient strains and general scarcity of long tracts in MMRS proficient species indicates the role of MMRS in restricting SSR tract expansions in prokaryotes.

Supplementary Table 6.1: Protein ID (PID) of MutS and MutL in various strains of prokaryotes analyzed in Chapter 5.

Name of the Strains	MutS (PID)	MutL (PID)
Acinetobacter_baumannii_AB0057	213156380	213158497
Acinetobacter_baumannii_AB307_0294	215483970	215483049
Acinetobacter_baumannii_ACICU	184157560	184158632
Acinetobacter_baumannii_ATCC_17978	126641297 126641298	126642159 126642160
Acinetobacter_baumannii_AYE	169796507	169795562
Acinetobacter_baumannii_SDF	169632885	169633284
Actinobacillus_pleuropneumoniae_L20	126209066	126209412
Actinobacillus_pleuropneumoniae_serovar_3_JL03	165977038	165977401
Actinobacillus_pleuropneumoniae_serovar_7_AP76	190150933	190151315
Bacillus_anthracis_Ames_0581	47529195	47529193
Bacillus_anthracis_Ames	30263775	30263774
Bacillus_anthracis_str_Sterne	49186619	49186618
Bacillus_cereus_AH187	217961190	217961189
Bacillus_cereus_AH820	218904897	218904896
Bacillus_cereus_ATCC_10987	42782853	42782852
Bacillus_cereus_ATCC14579	30021861	30021860
Bacillus_cereus_B4264	218232426	218231496
Bacillus_cereus_cytotoxis_NVH_391_98	152976135	152976134
Bacillus_cereus_G9842	218898869	218898868
Bacillus_cereus_ZK	52141719	52141721
Campylobacter_jejuni_81116	-	-
Campylobacter_jejuni_81_176	-	-
Campylobacter_jejuni_doylei_269_97	-	-
Campylobacter_jejuni	-	-
Campylobacter_jejuni_RM1221	-	-
Chlamydia_trachomatis_434_Bu	166154134	166154790
Chlamydia_trachomatis_A_HAR_13	76789534	76789312

Chlamydia_trachomatis_B_TZ1A828_OT	237805143	237804926
Chlamydia_trachomatis_L2b_UCH_1_proctitis	166155009	166155665
Chlamydia_trachomatis_D_UW_3_CX	15605525	15605304
Chlamydia_trachomatis_Jali20	237803222	237803004
Chlamydophila_pneumoniae_AR39	16752091	16752228
Chlamydophila_pneumoniae_CWL029	15618850	15618721
Chlamydophila_pneumoniae_J138	15836474	15836345
Chlamydophila_pneumoniae_TW_183	33242306	33242172
Clostridium_botulinum_A3_Loch_Maree	170758982	170761162
Clostridium_botulinum_A_ATCC_19397	153930881	153933845
Clostridium_botulinum_A	148379759	148379758
Clostridium_botulinum_B1_Okra	170755080	170757624
Clostridium_botulinum_B_Eklund_17B	187935238	187932933
Clostridium_botulinum_E3_Alaska_E43	188590144	188588377
Clostridium_botulinum_F_Langeland	153940053	153938478
Clostridium_perfringens_ATCC_13124	110799241	110800749
Clostridium_perfringens	18310137	18310138
Clostridium_perfringens_SM101	110803440	110801747
Escherichia_coli_536	215488050	215489514
Escherichia_coli_55989	218696327	218697919
Escherichia_coli_APEC_O1	117624966	117626517
Escherichia_coli_C_ATCC_8739	170019021	170021820
Escherichia_coli_CFT073	26249133	26251062
Escherichia_coli_E24377A	157156430	157158119
Escherichia_coli_ED1a	218690858	218692504
Escherichia_coli_HS	157162181	157163633
Escherichia_coli_IAI1	218555277	218556722
Escherichia_coli_IAI39	218701226	218702867
Escherichia_coli_K_12_substr_DH10B	170082309	170083616
Escherichia_coli_K12_substr_MG1655	16130640	16131992
Escherichia_coli_O157_H7_EC4115	209399182	209398397
Escherichia_coli_O157H7_EDL933	15803252	15804759

Escherichia_coli_O157H7	15832843	15834400
Escherichia_coli_S88	218559724	218561329
Escherichia_coli_SE11	209920172	209921658
Escherichia_coli_SMS_3_5	170680116	170681264
Escherichia_coli_UMN026	218706228	218707781
Escherichia_coli_UTI89	91212095	91213719
Escherichia_coli_W3110	89109520	89110890
Francisella_tularensis_FSC_198	110671009	110670099
Francisella_tularensis_holarctica_FTA	156501680	156503027
Francisella_tularensis_holarctica	89255725	89256853
Francisella_tularensis_holarctica_OSU18	115314221	115315236
Francisella_tularensis_mediasiatica_FSC147	187931226	187932042
Francisella_tularensis_novicida_U112	118498078	118497176
Francisella_tularensis_tularensis	56708538	56707628
Francisella_tularensis_WY96-3418	134301816	134302436
Haemophilus_influenzae_86_028NP	68249283	68248618
Haemophilus_influenzae	16272647	16272041
Haemophilus_influenzae_PittEE	148826651	148825652
Haemophilus_influenzae_PittGG	-	148827220
Helicobacter_pylori_26695	-	-
Helicobacter_pylori_G27	-	-
Helicobacter_pylori_HPAG1	-	-
Helicobacter_pylori_J99	-	-
Helicobacter_pylori_P12	-	-
Helicobacter_pylori_Shi470	-	-
Lactococcus_lactis_cremoris_MG1363	125625247	125625245
Lactococcus_lactis_cremoris_SK11	116513155	116513153
Lactococcus_lactis	15674192	15674190
Legionella_pneumophila_Corby	148359351	148358566
Legionella_pneumophila_Lens	54294692	54295538
Legionella_pneumophila_Paris	54297717	54298688
Legionella_pneumophila_Philadelphia_1	52842032	52842903

Listeria_monocytogenes_4b_F2365	46907631	46907632
Listeria_monocytogenes_HCC23	217964450	217964449
Listeria_monocytogenes	16803443	16803444
Methanococcus_maripaludis_C5	-	-
Methanococcus_maripaludis_C6	-	-
Methanococcus_maripaludis_C7	-	-
Methanococcus_maripaludis_S2	-	-
Mycobacterium_bovis_BCG_Pasteur_1173P2	-	-
Mycobacterium_bovis	-	
Mycobacterium_tuberculosis_CDC1551	-	-
Mycobacterium_tuberculosis_F11	-	-
Mycobacterium_tuberculosis_H37Ra	-	-
Mycobacterium_tuberculosis_H37Rv	-	-
Mycoplasma_hyopneumoniae_232	-	-
Mycoplasma_hyopneumoniae_7448	-	-
Mycoplasma_hyopneumoniae_J	-	-
Neisseria_meningitidis_053442	161871024	161870311
Neisseria_meningitidis_FAM18	121635803	121635130
Neisseria_meningitidis_MC58	15677973	15677300
Neisseria_meningitidis_Z2491	15793265	15794549
Pseudomonas_putida_F1	148549358	148549974
Pseudomonas_putida_GB_1	167032192	167035937
Pseudomonas_putida_KT2440	26988358	26991574
Pseudomonas_putida_W619	170723211	170723845
Pseudomonas_syringae_phaseolicola_1448A	71738021	71738065
Pseudomonas_syringae_pv_B728a	66044624	66043837
Pseudomonas_syringae_tomato_DC3000	28871201	28872058
Rhodopseudomonas_palustris_BisA53	115522274	115523441
Rhodopseudomonas_palustris_BisB18	90421887	90422900
Rhodopseudomonas_palustris_BisB5	91974801	91978493
Rhodopseudomonas_palustris_CGA009	39933589	39937431
Rhodopseudomonas_palustris_HaA2	86747654	86751283

Rhodopseudomonas_palustris_TIE_1	192288943	192293214
Salmonella_enterica_arizonae_serovar_62_z4_z23	161502025	161505138
Salmonella_enterica_Choleraesuis	62181411	62182805
Salmonella_enterica_Paratypi_ATCC_9150	56414858	56416150
Salmonella_enterica_serovar_Agona_SL483	197249260	197247369
Salmonella_enterica_serovar_Dublin_CT_02021853	198242699	198245330
Salmonella_enterica_serovar_Enteritidis_P125109	207858169	207859505
Salmonella_enterica_serovar_Gallinarum_287_91	205353850	205355117
Salmonella_enterica_serovar_Heidelberg_SL476	194449728	194450332
Salmonella_enterica_serovar_Newport_SL254	194445735	194445394
Salmonella_enterica_serovar_Paratyphi_A_AKU_12601	197363786	197365076
Salmonella_enterica_serovar_Paratyphi_B_SPB7	161615830	161617629
Salmonella_enterica_serovar_Schwarzengrund_CVM19633	194737683	194734797
Salmonella_enterica_serovar_Typhi_Ty2	29143167	29144657
Salmonella_typhimurium_LT2	16766215	16767605
Salmonella_typhi	16761683	16763178
Salmonella_typhi_Ty2	29143167	29144657
Shewanella_baltica_OS155	126175319	126172805
Shewanella_baltica_OS185	153001640	153002275
Shewanella_baltica_OS195	160876376	160876999
Shigella_flexneri_2a_2457T	30064089	30065542
Shigella_flexneri_2a	24114026	56480585
Shigella_flexneri_5_8401	110806640	110808088
Staphylococcus_aureus_COL	57651865	57651866
Staphylococcus_aureus_JH1	150393844	150393845
Staphylococcus_aureus_JH9	148267785	-
Staphylococcus_aureus_Mu3	156979616	156979617
Staphylococcus_aureus_Mu50	57634628	15924287
Staphylococcus_aureus_MW2	21282907	21282908
Staphylococcus_aureus_N315	15926878	15926879
Staphylococcus_aureus_Newman	151221416	151221417
Staphylococcus_aureus_NMCTC_8325	88195005	88195006

Staphylococcus_aureus_RF122	82750895	-
Staphylococcus_aureus_USA300	87160655	87161404
Staphylococcus_aureus_USA300_TCH1516	161509461	161509462
Streptococcus_agalactiae_2603	22538235	22538232
Streptococcus_agalactiae_A909	76786970	76787614
Streptococcus_agalactiae_NEM316	25012092	25012090
Streptococcus_pneumoniae_CGSP14	182685013	182683158
Streptococcus_pneumoniae_D39	116515831	116516396
Streptococcus_pneumoniae_G54	194397405	194398020
Streptococcus_pneumoniae_Hungary19A_6	169832918	169833770
Streptococcus_pneumoniae_R6	15903929	15902204
Streptococcus_pyogenes_M1_GAS	15675890	15675871
Streptococcus_pyogenes_Manfredo	139474579	139474561
Streptococcus_pyogenes_MGAS10270	94991407	94991382
Streptococcus_pyogenes_MGAS10394	50915171	50915149
Streptococcus_pyogenes_MGAS10750	94995318	94995292
Streptococcus_pyogenes_MGAS2096	94993335	94993334
Streptococcus_pyogenes_MGAS315	21911342	21911341
Streptococcus_pyogenes_MGAS5005	71911618	71911617
Streptococcus_pyogenes_MGAS6180	71904476	71904450
Streptococcus_pyogenes_MGAS8232	19746989	19746988
Streptococcus_pyogenes_MGAS9429	94989446	94989445
Streptococcus_pyogenes_NZ131	209560232	209560231
Streptococcus_pyogenes_SSI-1	28896716	28896715
Streptococcus_thermophilus_CNRZ1066	55822041	55822045
Streptococcus_thermophilus_LMD-9	116627015	116627019
Streptococcus_thermophilus_LMG_18311	55820157	55820153
Sulfolobus_islandicus_L_S_2_15	-	-
Sulfolobus_islandicus_M_14_25	-	-
Sulfolobus_islandicus_M_16_27	-	-
Sulfolobus_islandicus_M_16_4	-	-
Sulfolobus_islandicus_Y_G_57_14	-	-

Sulfolobus_islandicus_Y_N_15_51	-	-
Xanthomonas_campestris_8004	66769341	66768138
Xanthomonas_campestris_ATCC_33913	21230664	21231736
Xanthomonas_campestris_B100	188992527	188991275
Xanthomonas_campestris_vesicatoria_85_10	78046908	78048158
Xanthomonas_oryzae_KACC10331	58581265	58582355
Xanthomonas_oryzae_MAFF_311018	84623184	84624234
Xanthomonas_oryzae_PXO99A	188577456	188576161
Xylella_fastidiosa_M12	170730376	170731141
Xylella_fastidiosa_M23	182681688	182682515
Xylella_fastidiosa	15838317	15837362
Xylella_fastidiosa_Temecula1	28198974	28199765
Yersinia_pestis_Angola	162418943	162418239
Yersinia_pestis_Antiqua	108808780	108809903
Yersinia_pestis_biovar_Antiqua_str_B42003004	166212949	166213990
Yersinia_pestis_biovar_Antiqua_str_E1979001	166010151	166011848
Yersinia_pestis_biovar_Antiqua_str_UG05_0454	167398361	167400705
Yersinia_pestis_biovar_Mediaevails	45440190	45440381
Yersinia_pestis_biovar_Mediaevalis_str_K1973002	167423797	167423238
Yersinia_pestis_biovar_Orientalis_str_F1991016	165925876	165926767
Yersinia_pestis_biovar_Orientalis_str_IP275	165936708	165936468
Yersinia_pestis_biovar_Orientalis_str_MG05	167421538	167418954
Yersinia_pestis_CA88_4125	153997602	150260585
Yersinia_pestis_CO92	16123504	16120706
Yersinia_pestis_FV_1	167469363	167470469
Yersinia_pestis_KIM	22124746	22124542
Yersinia_pestis_Nepal516	108810904	108813460
Yersinia_pestis_Pestoides_F	145600263	145600850
Yersinia_pseudotuberculosis_IP_31758	153948692	153947725
Yersinia_pseudotuberculosis_IP32953	51595128	51594775
Yersinia_pseudotuberculosis_PB1	186894140	186893782
Yersinia_pseudotuberculosis_YPIII	170025639	170026016

Supplementary Table 6.2: The number of non-coding PSSRs, repeat unit size-wise, found in various prokaryotic species.

Name of species followed by number of strains	Total	Mono	Di	Tri	Tetra	Penta	Hexa
Acinetobacter_baumannii_6	**665**	658	4	3	0	0	0
Actinobacillus_pleuropneumoniae_3	**110**	105	2	2	0	1	0
Bacillus_anthracis_3	**21**	21	0	0	0	0	0
Bacillus_cereus_8	**1449**	1402	39	5	3	0	0
Bradyrhizobium_3	**48**	48	0	0	0	0	0
Clostridium_botulinum_a_8	**643**	587	43	11	2	0	0
Corynebacterium_glutamicum_3	**118**	111	6	1	0	0	0
Campylobacter_jejuni_5	**177**	168	4	4	1	0	0
Chlamydophila_pneumoniae_4	**7**	7	0	0	0	0	0
Clostridium_perfringens_3	**390**	370	11	6	2	1	0
Chlamydia_trachomatis_4	**100**	97	1	1	1	0	0
Desulfovibrio_vulgaris_3	**89**	86	1	1	1	0	0
Escherichia_coli_22	**654**	640	7	6	1	0	0
Ehrlichia_ruminantium_3	**57**	54	1	1	1	0	0
Francisella_tularensis_8	**182**	162	8	8	2	0	2
Haemophilus_influenzae_4	**272**	265	4	1	0	1	1
Helicobacter_pylori_6	**300**	298	2	0	0	0	0
Lactobacillus_delbrueckii_2	**104**	101	3	0	0	0	0
Lactococcus_lactis_3	**213**	208	3	1	0	1	0
Listeria_monocytogenes_3	**275**	272	3	0	0	0	0
Legionella_pneumophila_4	**351**	334	8	7	1	1	0
Mycoplasma_hyopneumoniae_3	**114**	109	0	2	0	0	3
Methanococcus_maripaludis_4	**253**	246	4	2	1	0	0
Mycobacterium_tuberculosis_6	**24**	23	0	1	0	0	0
Neisseria_meningitidis_4	**186**	178	4	2	2	0	0
Pseudomonas_aeruginosa_4	**414**	408	3	1	0	2	0
Pseudomonas_putida_4	**311**	305	5	0	1	0	0
Pseudomonas_syringae_3	**176**	173	3	0	0	0	0
Rhodopseudomonas_palustris_6	**206**	199	5	1	1	0	0

Staphylococcus_aureus_14	**466**	435	20	9	1	0	1
Streptococcus_agalactiae_3	**90**	85	3	0	1	1	0
Shewanella_baltica_3	**292**	286	5	1	0	0	0
Salmonella_enterica_13	**500**	471	9	3	1	0	16
Shigella_flexneri_3	**72**	67	2	1	1	0	1
Sulpholobus_islandicus_6	**228**	215	7	6	0	0	0
Streptococcus_pneumoniae_5	**131**	125	5	0	1	0	0
Streptococcus_pyogenes_13	**409**	397	4	6	0	2	0
Salmonella_typhi_3	**126**	121	3	1	1	0	0
Streptococcus_thermophilus_3	**145**	143	2	0	0	0	0
Xanthomonas_campestris_4	**109**	93	8	0	0	0	8
Xanthomonas_oryzae_3	**93**	81	4	3	2	1	2
Xylella_fastidiosa_4	**163**	148	7	3	0	2	3
Yersinia_pestis_20	**364**	313	8	3	10	9	21

Chapter 7

Conclusions

Chapter 7 *Conclusions*

This chapter summarises all the major findings and conclusions of my investigations on simple sequence repeats (SSRs) in prokaryotic genomes presented in this thesis. In the begining, an in-depth analysis of SSRs with regard to the distribution, enrichment and polymorphism of SSRs in all the known genomes of the well known pathogen *Yersinia pestis* and its ancestor *Yersinia pseudotuberculosis* was done. The details of these studies are presented in Chapter 2. Further I carried out cross-genome comparisons of the known Yersinia genomes to discover the SSRs that have undergone length variations (i.e., repeat copy number variations). The effects of these polymorphic SSRs on protein function were later investigated. It was found that the coding regions have undergone changes such as premature termination; length variations etc and such changes seem to have formed the basis for evolution of pathogens to adapt to new environments, foster new life styles and even become virulent or non-virulent. The details of these findings have been given in Chapter 4. These studies have given some insights into how SSRs, in general, are involved in the evolution of pathogens from their ancestor to various strains. Some of the important findings of these studies are given below.

- All the strains of Yersinia are very similar to each other in terms of SSR distribution, abundance of motifs and their enrichment. Majority of SSRs are short and long tracts are nearly absent indicating that SSR tract expansions have not been favoured in these genomes.
- SSRs are found both in coding as well as non-coding regions; however, their tract densities are higher in non-coding regions than coding regions.

- ORFs encoding proteins which interact with host show high propensities for SSRs whereas the ORFs which code for proteins involved in general metabolism show low propensities.
- The 5’ and 3’ parts of genes are denser with SSRs than their middle parts.
- Analysis of polymorphic SSRs corroborated with the available experimental data clearly postulates the important role of some of the SSRs helping to bring out intra-species variations. For example, the LPS synthesis constitutes seventeen genes which are located between the *hemH* and *gsk* loci. Six of these genes corresponding to O-antigen gene cluster are prematurely terminated in all the plague causing strains of *Y. pestis* genomes as compared to YPTB due to mutations in SSRs.
- A number of ORFs in YPTB have undergone premature terminations in *Y. pestis* as a consequence of frame-shift mutations caused by polymorphic SSRs in these ORFs suggesting SSR mutation as one of the reasons for gene loss in *Y. pestis*.
- An investigation of PSSRs indicated that Yesinia genomes have evolved such that there is a preference for the frame-shift mutation in the genes which are either directly or indirectly involved in defense mechanism, interaction with host or environment (extracellular structure), secretion and cell motility. For example, the non functional genes of flagella and pili operon biosynthesis pathway of *Y. pestis* that are a result of SSR mutations is an example of the role played by SSR mutation in adaptation, to best suite the change in life style of the pathogen. SSR mutations in the cysteine synthesis pathway are one of the reasons that *Y. pestis* is an auxotroph where as its ancestor *Y. pseudotuberculosis* is a chemoheterotroph.

- The unique pattern of SSR mutation in the case of non-plague causing YP strain seems very interesting. Some genes, showing SSR mutation induced changes which lead to premature termination in all the plague causing *Y. pestis*, are intact in the non-pathogenic YP (non plague causing *Y. pestis*). For example, six genes of O-antigen cluster are mutated in *Y. pestis;* however three of these mutated genes are intact in non plague causing *Y. pestis* (*Y. pestis* 91001). This suggests that the corresponding gene products are probably lowering human-specific virulence properties in YP. On the other hand, some other genes which are intact in all the plague causing Yersinia have been mutated due to premature terminations in *Y. pestis* 91001 (non plague causing *Y. pestis*). These gene products may thus have a role in causing virulence. Our studies, therefore, support the view that factors other than pathogenic plasmids are also responsible for the virulence of *Yersinia pestis* (Revell and Miller 2001; Song et al. 2004).

Chapters 5 and 6 focus on mutational tendencies of SSRs in prokaryotic genomes. While Chapter 5 focuses on the coding PSSRs, Chapter 5 focuses on the non-coding PSSRs. Repeat copy number variations in SSRs arise as a consequence of mispairing of nascent and template strands due to slippage during replication and hence can be affected by various factors inherent to the genomes, perhaps further compounded by selection. SSRs are considered neutral in eukaryotes; however, in prokaryotes they come under selection pressures especially when they are in the coding regions. The distribution tendencies of polymorphic SSRs in the coding regions was analysed in a large set of prokaryotic genomes in the light of null hypothesis that PSSRs do not show region specific distributions. The major findings in Chapter 5 are given below.

- The repeat copy number variations in SSRs are not limited to pathogenic bacteria alone and hence it seems that replication slippage occurs in all species of prokaryotes and contributes to their intra-species genetic diversity.
- In all the species the tract density of PSSRs in the non-coding regions is strikingly higher than the coding regions. Higher incidences of SSR polymorphism in non-coding regions as compared to the coding regions indicates relatively unrestrained polymorphism in non-coding regions and restraint on SSR polymorphism in coding regions can be attributed to selection pressures to avoid frame shift mutations (Metzgar, Bytof, and Wills 2000).
- Mono nucleotide SSRs despite their frame shifting ability are the most abundant PSSRs in the coding regions although tracts longer than 8bp are rarely found.
- In all the cases of inter-strain comparisons, other than INDELs in SSRs, the compared ORFs are ~100% identical and therefore any effect that is seen is purely a consequence of the SSR mutations.
- It was found that some pseudo genes harboring PSSRs are conserved in some species. Given the fact that SSR mutations are often reversible, it is quite possible that these pseudo genes are only the transitory phases of functional genes where SSRs act as on-off switches.
- Though the mutation in SSR is thought to be a random process but the global analysis clearly demonstrates that the selection has a big role to play in this case,

which was evident from the constrained polymorphism in the coding regions as compared to non-coding regions of a genome.

- The regions where replication and transcription happen in the head-on directions harbor higher densities of PSSRs than the regions where the two processes happen in co-oriented directions. This also explains why in most of the prokaryotic genomes the gene density on the leading strand is high.
- Positional distribution bias of PSSRs in intragenic region further indicates non-neutrality of SSRs in the coding regions of prokaryotic genomes.
- Low presence of PSSRs in the middle of ORF is note worthy as it could produce less number of truncated non-functional proteins which are toxic to the organism and decrease fitness.
- 5’ and 3’ parts of genes show high density of PSSRs. PSSR in the 5’ part of a gene can lead to premature termination of that gene or may lead to loss of its signal peptide. The former saves the translational costs of the organism and the latter affects the sorting of the protein.
- PSSR at the 3’ part can have the least effect on the structure and function of a gene. Hence high incidences of PSSRs on 5’ and 3’ parts of genes seem to be the consequence of selection pressures.
- The high abundance of PSSRs in host adapted pathogen is consistent with a possible role of these PSSRs in pathogen host interactions.
- The high abundance of PSSRs in *Haemophilus influenzae*, *Acinetobacter baumannii*, *Helicobacter pylori*, *Mycoplasma hyopneumoniae*, *Streptococcus pyogenes*, *Neisseria*

meningitides and *Bacillus cereus* is an indication of the involvement of SSR in phase variation in these species because these pathogens are believed to be extracellular and therefore exposed to the host immune system.

Further, non-coding PSSRs were investigated for their tendencies of expansion and contraction and details were presented in Chapter 6. Earlier studies had showed that mutational tendency of SSRs towards expansion or contraction is related to the presence or absence of mismatch repair system (MMRS). However, the observations varied from system to system thus giving a confusing picture about the exact relationship between the two viz., mutational tendency of SSRs and MMRS. Therefore PSSRs in the non-coding regions were analysed for their mutational tendency towards expansion or contraction in MMRS proficient and deficient species. The important findings from Chapter 5 are given below.

- Strains lacking MMR show significantly more number of PSSRs as compared to the MMR proficient strains indicating MMR deficiency has led to destabilization of SSR tracts.
- PSSRs are contraction biased in MMR proficient strains (*Haemophilus influenzae, Acinetobacter baumannii* and *Staphylococcus aureans)*. On the other hand MMR deficient strains do not show such a bias. These results suggest that mismatch repair system (MMRS) is less efficient in repairing slippage mutation on the template strand as compared to nascent strand.
- For species that lack MMR, PSSRs show slight bias towards expansion which indicates that primary SSR mutations are expansion biased and as a consequence, high numbers of long SSR tracts (>8bp) are present in those MMR deficient species which

show a significant expansion bias (*Mycoplasma hyopneumoniae*) as compared to MMR proficient species.

- Mutations in SSRs which escape MMR system are referred to as secondary mutations. In most of the species these are significantly contraction biased.
- The expansion or contraction of SSRs does not depend on the length of SSR tracts. Contractions in small as well as long SSR tracts in the MMRS proficient species are equally biased.
- The directionality of SSR evolution does not depend on the genome composition of an organism i.e., GC composition of genome.
- The general scarcity of long tracts of SSRs in prokaryotic species, which are mainly MMRS proficient, is mainly due to the functional MMR system.

Significance and innovation in the research work presented:

My study on Yersinia has disovered that evolution of yersinia is indeed assisted by SSRs. For example, LPS synthesis in Yersinia constitutes seventeen genes of which about six of these genes correspond to the O-antigen gene cluster are prematurely terminated in all the plague causing strains of *Y. pestis* genomes as compared to YPTB due to mutations in SSRs. The heterologous replacement of these loci in *Y. pestis* decreased its virulence (Kukkonen et al. 2004; Lathem et al. 2007). On the other hand non-functional genes of flagella and pili operon biosynthesis pathway of *Y. pestis* due to SSR mutation may be viewed as an adaptation due to the change in the life style of the pathogen. Furthermore the functional UreD gene which encodes urease enzyme in YPTB is pseudogene due to SSR polymorphism in *Y. pestis*. This too correlates with the life style of these pathogens. YPTB, as it lives in soil and water, requires *urease* enzyme for the degradation of nitrogenous products for its nitrogen requirements whereas *Y. pestis* due to its different life style does not require *urease* activity.

The identification of PSSRs in a species has a very good advantage. Depending upon the region of occurrence it could have different potential application. The strain specific PSSR (SSR length varies only in one species) could be used for the identification of that strain and is of importance in making diagnostic kits.

In addition to that SSR loci have been used by various groups of bacteria as a munitions store for generating diversity to either cope with the host immune response or to save their resources. Hence a large number of PSSRs found in coding regions in this study could be good candidates to study the functional role of genes in pathogenesis and virulence. The presence of PSSRs in genes indicates structural and functional variable nature of genes. In the past attempts to develop vaccine against group A Streptococcus (*S. pyogens*) and *H. pylori* have encountered major hurdles due to the variability of the proteins between strains of a species which were being considered for vaccine design (Telford 2008). Hence the information of genes showing structural and functional variability due to PSSR could be a very important asset for therapeutic and vaccine design, for which, a conserved regions of a gene or gene product are considered that have important roles in pathogenesis and virulence.

Further it was also found that species lacking MMR have mutational bias towards expansion of SSRs. On the other hand species where MMRs is present the SSR mutations showed significant bias towards deletions. Furthermore it was also found that contraction bias is independent of sequence and length of SSRs. These finding shed some light on the smaller tract of SSR in bacterial genomes which are mostly MMR proficient.

Outcomes of the thesis

The main objective of this thesis was to study the evolution of SSRs in prokaryotic genomes. I, therefore, initially considered Yersinia genomes and analyzed them for SSR distribution and polymorphism. Comparison of equivalent SSRs between the plague causing *Y. pestis* and its ancestor *Y. pseudotuberculosis* (non plague causing) revealed a large number of polymorphic SSRs. Some of the PSSR's effects on the genes were corroborated with the life style of the pathogenic and non-pathogenic strain indicating the possible role of SSRs in the evolution of the deadly pathogen.

The global analysis of PSSRs in forty three prokaryotic species supported the view of non random distribution of PSSRs in genomes, which was evident from the constrained polymorphism in the coding regions as compared to non-coding regions of a genome. Positional distribution bias of PSSRs in intragenic region further indicated non-neutrality of SSRs in the coding regions of prokaryotic genomes. These studies meet the second objective of this thesis which was to study the distribution of PSSR in prokaryotic genomes to find the random or non random nature of PSSR mutation.

The high abundance of PSSRs in *Haemophilus influenzae*, *Acinetobacter baumannii*, *Helicobacter pylori*, *Mycoplasma hyopneumoniae*, *Streptococcus pyogenes*, *Neisseria meningitides* and *Bacillus cereus* were an indication of the involvement of SSRs in phase variation in these species because these pathogens are believed to be extracellular and are therefore exposed to the host immune system. The high abundance of PSSRs in these host adapted pathogens was consistent with a possible role of these PSSRs in pathogen host interactions. These results meet the third objective of my study where I wished to find out the differences in PSSR density between the host adapted pathogens and the non pathogenic bacteria.

Strains/species deficient in MMR showed significantly more number of PSSRs as compared to the MMR proficient strains/species indicating that the MMR deficiency has led to destabilization of SSR tracts. These results counter the fourth objective of my study. In most of the MMR proficient species SSR mutations were significantly contraction biased whereas in species deficient in MMR, SSR mutations were slightly biased towards expansion. The general scarcity of long tracts of SSRs in prokaryotic species, which are mainly MMRs proficient, could be due to deletion bias of SSR evolution. These results meet the last objective of this thesis which was to find out the direction of evolution of SSRs which could give an insight into the evolution of SSRs and rarity of longer SSRs in prokaryotic genomes.

Future directions

In this study I have found more than eighteen thousand PSSRs across forty three species of prokaryotes. The availability of all these PSSRs on World Wide Web would be a great resource for the scientific community. I also see a possibility of functional study of the genes harboring these PSSRs as a means to find out their importance in host pathogen interactions.

References

References

Aaltonen, L. A., P. Peltomaki, F. S. Leach, P. Sistonen, L. Pylkkanen, J. P. Mecklin, H. Jarvinen, S. M. Powell, J. Jen, S. R. Hamilton, and et al. 1993. Clues to the pathogenesis of familial colorectal cancer. Science 260:812-816.

Acharya, S., P. L. Foster, P. Brooks, and R. Fishel. 2003. The coordinated functions of the E. coli MutS and MutL proteins in mismatch repair. Mol Cell 12:233-246.

Acharya, S., T. Wilson, S. Gradia, M. F. Kane, S. Guerrette, G. T. Marsischky, R. Kolodner, and R. Fishel. 1996. hMSH2 forms specific mispair-binding complexes with hMSH3 and hMSH6. Proc Natl Acad Sci U S A 93:13629-13634.

Achtman, M., K. Zurth, G. Morelli, G. Torrea, A. Guiyoule, and E. Carniel. 1999. Yersinia pestis, the cause of plague, is a recently emerged clone of Yersinia pseudotuberculosis. Proc Natl Acad Sci U S A 96:14043-14048.

Ackermann, M., and L. Chao. 2006. DNA sequences shaped by selection for stability. PLoS Genet 2:e22.

Adair, D. M., P. L. Worsham, K. K. Hill, A. M. Klevytska, P. J. Jackson, A. M. Friedlander, and P. Keim. 2000. Diversity in a variable-number tandem repeat from Yersinia pestis. J Clin Microbiol 38:1516-1519.

Amos, W., and D. C. Rubinstzein. 1996. Microsatellites are subject to directional evolution. Nat Genet 12:13-14.

Amos, W., S. J. Sawcer, R. W. Feakes, and D. C. Rubinsztein. 1996. Microsatellites show mutational bias and heterozygote instability. Nat Genet 13:390-391.

Armour, J. A. L., Alegre, S.A., Miles, S., Williams, L.J., Badge, R.M.,. 1999. Minisatellites and mutation processes in tandemly repetitive DNA. Pp. 24-33 *in* D. B. Goldstein, Schlötterer, C. (Eds.), ed. Microsatellites: Evolution and Applications. Oxford University Press, Oxford.

Ashley, C. T., Jr., and S. T. Warren. 1995. Trinucleotide repeat expansion and human disease. Annu Rev Genet 29:703-728.

Aslanidis, C., G. Jansen, C. Amemiya, G. Shutler, M. Mahadevan, C. Tsilfidis, C. Chen, J. Alleman, N. G. Wormskamp, M. Vooijs, and et al. 1992. Cloning of the essential myotonic dystrophy region and mapping of the putative defect. Nature 355:548-551.

Bachtrog, D., M. Agis, M. Imhof, and C. Schlotterer. 2000. Microsatellite variability differs between dinucleotide repeat motifs-evidence from Drosophila melanogaster. Mol Biol Evol 17:1277-1285.

Backstrom, A., C. Lundberg, D. Kersulyte, D. E. Berg, T. Boren, and A. Arnqvist. 2004. Metastability of Helicobacter pylori bab adhesin genes and dynamics in Lewis b antigen binding. Proc Natl Acad Sci U S A 101:16923-16928.

Banerjee, A., R. Wang, S. L. Supernavage, S. K. Ghosh, J. Parker, N. F. Ganesh, P. G. Wang, S. Gulati, and P. A. Rice. 2002. Implications of phase variation of a gene (pgtA) encoding a pilin galactosyl transferase in gonococcal pathogenesis. J Exp Med 196:147-162.

Bayliss, C. D., D. Field, and E. R. Moxon. 2001. The simple sequence contingency loci of Haemophilus influenzae and Neisseria meningitidis. J Clin Invest 107:657-662.

Bell, G. I., and J. Jurka. 1997. The length distribution of perfect dimer repetitive DNA is consistent with its evolution by an unbiased single-step mutation process. J Mol Evol 44:414-421.

Benson, G. 1999. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res 27:573-580.

Bernardi, G., D. Mouchiroud, and C. Gautier. 1988. Compositional patterns in vertebrate genomes: conservation and change in evolution. J Mol Evol 28:7-18.

Bicknell, D. C., L. Kaklamanis, R. Hampson, W. F. Bodmer, and P. Karran. 1996. Selection for beta 2-microglobulin mutation in mismatch repair-defective colorectal carcinomas. Curr Biol 6:1695-1697.

Bizzaro, J. W., and K. A. Marx. 2003. Poly: a quantitative analysis tool for simple sequence repeat (SSR) tracts in DNA. BMC Bioinformatics 4:22.

Borstnik, B., and D. Pumpernik. 2002. Tandem repeats in protein coding regions of primate genes. Genome Res 12:909-915.

Boyer, J. C., N. A. Yamada, C. N. Roques, S. B. Hatch, K. Riess, and R. A. Farber. 2002. Sequence dependent instability of mononucleotide microsatellites in cultured mismatch repair proficient and deficient mammalian cells. Hum Mol Genet 11:707-713.

Brais, B., J. P. Bouchard, Y. G. Xie, D. L. Rochefort, N. Chretien, F. M. Tome, R. G. Lafreniere, J. M. Rommens, E. Uyama, O. Nohira, S. Blumen, A. D. Korczyn, P. Heutink, J. Mathieu, A. Duranceau, F. Codere, M. Fardeau, and G. A. Rouleau. 1998. Short GCG expansions in the PABP2 gene cause oculopharyngeal muscular dystrophy. Nat Genet 18:164-167.

Brinkmann, B., M. Klintschar, F. Neuhuber, J. Huhne, and B. Rolf. 1998. Mutation rate in human microsatellites: influence of the structure and length of the tandem repeat. Am J Hum Genet 62:1408-1415.

Britten, R. J., and D. E. Kohne. 1968. Repeated sequences in DNA. Hundreds of thousands of copies of DNA sequences have been incorporated into the genomes of higher organisms. Science 161:529-540.

Brunham, R. C., F. A. Plummer, and R. S. Stephens. 1993. Bacterial antigenic variation, host immune response, and pathogen-host coevolution. Infect Immun 61:2273-2276.

Buckling, A., J. Neilson, J. Lindsay, R. ffrench-Constant, M. Enright, N. Day, and R. C. Massey. 2005. Clonal distribution and phase-variable expression of a major histocompatibility complex analogue protein in Staphylococcus aureus. J Bacteriol 187:2917-2919.

Burch, C. L., R. J. Danaher, and D. C. Stein. 1997. Antigenic variation in Neisseria gonorrhoeae: production of multiple lipooligosaccharides. J Bacteriol 179:982-986.

Burge, C., A. M. Campbell, and S. Karlin. 1992. Over- and under-representation of short oligonucleotides in DNA sequences. Proc Natl Acad Sci U S A 89:1358-1362.

Buschiazzo, E., and N. J. Gemmell. 2006. The rise, fall and renaissance of microsatellites in eukaryotic genomes. Bioessays 28:1040-1050.

Calabrese, P. P., R. T. Durrett, and C. F. Aquadro. 2001. Dynamics of microsatellite divergence under stepwise mutation and proportional slippage/point mutation models. Genetics 159:839-852.

Campuzano, V., L. Montermini, M. D. Molto, L. Pianese, M. Cossee, F. Cavalcanti, E. Monros, F. Rodius, F. Duclos, A. Monticelli, F. Zara, J. Canizares, H. Koutnikova, S. I. Bidichandani, C. Gellera, A. Brice, P. Trouillas, G. De Michele, A. Filla, R. De Frutos, F. Palau, P. I. Patel, S. Di Donato, J. L. Mandel, S. Cocozza, M. Koenig, and M. Pandolfo. 1996. Friedreich's ataxia: autosomal recessive disease caused by an intronic GAA triplet repeat expansion. Science 271:1423-1427.

Carroll, P. A., K. T. Tashima, M. B. Rogers, V. J. DiRita, and S. B. Calderwood. 1997. Phase variation in tcpH modulates expression of the ToxR regulon in Vibrio cholerae. Mol Microbiol 25:1099-1111.

Carson, S. D., B. Stone, M. Beucher, J. Fu, and P. F. Sparling. 2000. Phase variation of the gonococcal siderophore receptor FetA. Mol Microbiol 36:585-593.

Carvalho, F. M., M. M. Fonseca, S. Batistuzzo De Medeiros, K. C. Scortecci, C. A. Blaha, and L. F. Agnez-Lima. 2005. DNA repair in reduced genome: the Mycoplasma model. Gene 360:111-119.

Castelo, A. T., W. Martins, and G. R. Gao. 2002. TROLL--tandem repeat occurrence locator. Bioinformatics 18:634-636.

Chain, P. S., E. Carniel, F. W. Larimer, J. Lamerdin, P. O. Stoutland, W. M. Regala, A. M. Georgescu, L. M. Vergez, M. L. Land, V. L. Motin, R. R. Brubaker, J. Fowler, J. Hinnebusch, M. Marceau, C. Medigue, M. Simonet, V. Chenal-Francisque, B. Souza, D. Dacheux, J. M. Elliott, A. Derbise, L. J. Hauser, and E. Garcia. 2004. Insights into the evolution of Yersinia pestis through whole-genome comparison with Yersinia pseudotuberculosis. Proc Natl Acad Sci U S A 101:13826-13831.

Chain, P. S., P. Hu, S. A. Malfatti, L. Radnedge, F. Larimer, L. M. Vergez, P. Worsham, M. C. Chu, and G. L. Andersen. 2006. Complete genome sequence of Yersinia pestis strains Antiqua and Nepal516: evidence of gene reduction in an emerging pathogen. J Bacteriol 188:4453-4463.

Chakraborty, R., M. Kimmel, D. N. Stivers, L. J. Davison, and R. Deka. 1997. Relative mutation rates at di-, tri-, and tetranucleotide microsatellite loci. Proc Natl Acad Sci U S A 94:1041-1046.

Chambers, G. K., and E. S. MacAvoy. 2000. Microsatellites: consensus and controversy. Comp Biochem Physiol B Biochem Mol Biol 126:455-476.

Chen, C. J., C. Elkins, and P. F. Sparling. 1998. Phase variation of hemoglobin utilization in Neisseria gonorrhoeae. Infect Immun 66:987-993.

Claverys, J. P., and S. A. Lacks. 1986. Heteroduplex deoxyribonucleic acid base mismatch repair in bacteria. Microbiol Rev 50:133-165.

Coenye, T., and P. Vandamme. 2005. Characterization of mononucleotide repeats in sequenced prokaryotic genomes. DNA Res 12:221-233.

Coetzee, G., and R. Irvine. 2002. Size of the androgen receptor CAG repeat and prostate cancer: does it matter? J Clin Oncol 20:3572-3573.

Cope, L. D., Z. Hrkal, and E. J. Hansen. 2000. Detection of phase variation in expression of proteins involved in hemoglobin and hemoglobin-haptoglobin binding by nontypeable Haemophilus influenzae. Infect Immun 68:4092-4101.

Cummings, C. J., and H. Y. Zoghbi. 2000. Trinucleotide repeats: mechanisms and pathophysiology. Annu Rev Genomics Hum Genet 1:281-328.

Danaher, R. J., J. C. Levin, D. Arking, C. L. Burch, R. Sandlin, and D. C. Stein. 1995. Genetic basis of Neisseria gonorrhoeae lipooligosaccharide antigenic variation. J Bacteriol 177:7275-7279.

De Bolle, X., C. D. Bayliss, D. Field, T. van de Ven, N. J. Saunders, D. W. Hood, and E. R. Moxon. 2000. The length of a tetranucleotide repeat tract in Haemophilus influenzae determines the phase variation rate of a gene with homology to type

III DNA methyltransferases. Mol Microbiol 35:211-222.

de Vries, N., D. Duinsbergen, E. J. Kuipers, R. G. Pot, P. Wiesenekker, C. W. Penn, A. H. van Vliet, C. M. Vandenbroucke-Grauls, and J. G. Kusters. 2002. Transcriptional phase variation of a type III restriction-modification system in Helicobacter pylori. J Bacteriol 184:6615-6623.

Dejager, S., H. Bry-Gauillard, E. Bruckert, B. Eymard, F. Salachas, E. LeGuern, S. Tardieu, R. Chadarevian, P. Giral, and G. Turpin. 2002. A comprehensive endocrine description of Kennedy's disease revealing androgen insensitivity linked to CAG repeat length. J Clin Endocrinol Metab 87:3893-3901.

Delgrange, O., and E. Rivals. 2004. STAR: an algorithm to Search for Tandem Approximate Repeats. Bioinformatics 20:2812-2820.

Deng, W., V. Burland, G. Plunkett, 3rd, A. Boutin, G. F. Mayhew, P. Liss, N. T. Perna, D. J. Rose, B. Mau, S. Zhou, D. C. Schwartz, J. D. Fetherston, L. E. Lindler, R. R. Brubaker, G. V. Plano, S. C. Straley, K. A. McDonough, M. L. Nilles, J. S. Matson, F. R. Blattner, and R. D. Perry. 2002. Genome sequence of Yersinia pestis KIM. J Bacteriol 184:4601-4611.

Di Rienzo, A., A. C. Peterson, J. C. Garza, A. M. Valdes, M. Slatkin, and N. B. Freimer. 1994. Mutational processes of simple-sequence repeat loci in human populations. Proc Natl Acad Sci U S A 91:3166-3170.

Dieringer, D., and C. Schlotterer. 2003. Two distinct modes of microsatellite mutation processes: evidence from the complete genomic sequences of nine species. Genome Res 13:2242-2251.

Dixon, K., C. D. Bayliss, K. Makepeace, E. R. Moxon, and D. W. Hood. 2007. Identification of the functional initiation codons of a phase-variable gene of Haemophilus influenzae, lic2A, with the potential for differential expression. J Bacteriol 189:511-521.

Dos Vultos, T., O. Mestre, T. Tonjum, and B. Gicquel. 2009. DNA repair in Mycobacterium tuberculosis revisited. FEMS Microbiol Rev 33:471-487.

Duval, A., J. Gayet, X. P. Zhou, B. Iacopetta, G. Thomas, and R. Hamelin. 1999. Frequent frameshift mutations of the TCF-4 gene in colorectal cancers with microsatellite instability. Cancer Res 59:4213-4215.

Duval, A., and R. Hamelin. 2002. Mutations at coding repeat sequences in mismatch repair-deficient human cancers: toward a new concept of target genes for instability. Cancer Res 62:2447-2454.

Eckert, K. A., and G. Yan. 2000. Mutational analyses of dinucleotide and tetranucleotide microsatellites in Escherichia coli: influence of sequence on expansion mutagenesis. Nucleic Acids Res 28:2831-2838.

Eisen, J. A. 1998. A phylogenomic study of the MutS family of proteins. Nucleic Acids Res 26:4291-4300.

Ejima, Y., L. Yang, and M. S. Sasaki. 2000. Aberrant splicing of the ATM gene associated with shortening of the intronic mononucleotide tract in human colon tumor cell lines: a novel mutation target of microsatellite instability. Int J Cancer 86:262-268.

Ellegren, H. 2004. Microsatellites: simple sequences with complex evolution. Nat Rev Genet 5:435-445.

Ellegren, H. 2002. Mismatch repair and mutational bias in microsatellite DNA. Trends Genet 18:552.

Ellegren, H. 2000. Microsatellite mutations in the germline: implications for evolutionary inference. Trends Genet 16:551-558.

Ellegren, H., G. Lindgren, C. R. Primmer, and A. P. Moller. 1997. Fitness loss and germline mutations in barn swallows breeding in Chernobyl. Nature 389:593-596.

Ellegren, H., C. R. Primmer, and B. C. Sheldon. 1995. Microsatellite 'evolution': directionality or bias? Nat Genet 11:360-362.

Eppinger, M., M. J. Rosovitz, W. F. Fricke, D. A. Rasko, G. Kokorina, C. Fayolle, L. E. Lindler, E. Carniel, and J. Ravel. 2007. The complete genome sequence of Yersinia pseudotuberculosis IP31758, the causative agent of Far East scarlet-like fever. PLoS Genet 3:e142.

Estoup, A., P. Jarne, and J. M. Cornuet. 2002. Homoplasy and mutation model at microsatellite loci and their consequences for population genetics analysis. Mol Ecol 11:1591-1604.

Feldman, M. W., A. Bergman, D. D. Pollock, and D. B. Goldstein. 1997. Microsatellite genetic distances with range constraints: analytic description and problems of estimation. Genetics 145:207-216.

Fickett, J. W. 1982. Recognition of protein coding regions in DNA sequences. Nucleic Acids Res 10:5303-5318.

Field, D., and C. Wills. 1998. Abundant microsatellite polymorphism in Saccharomyces cerevisiae, and the different distributions of microsatellites in eight prokaryotes and S. cerevisiae, result from strong mutation pressures and a variety of selective forces. Proc Natl Acad Sci U S A 95:1647-1652.

Finn, R. D., J. Tate, J. Mistry, P. C. Coggill, S. J. Sammut, H. R. Hotz, G. Ceric, K. Forslund, S. R. Eddy, E. L. Sonnhammer, and A. Bateman. 2008. The Pfam protein families database. Nucleic Acids Res 36:D281-288.

Franke, P., M. Leboyer, M. Gansicke, O. Weiffenbach, V. Biancalana, P. Cornillet-Lefebre, M. F. Croquette, U. Froster, S. G. Schwab, F. Poustka, M. Hautzinger, and W. Maier. 1998. Genotype-phenotype relationship in female carriers of the premutation and full mutation of FMR-1. Psychiatry Res 80:113-127.

Fresco, J. R., and B. M. Alberts. 1960. The Accommodation of Noncomplementary Bases in Helical Polyribonucleotides and Deoxyribonucleic Acids. Proc Natl Acad Sci U S A 46:311-321.

Fu, Y. H., D. P. Kuhl, A. Pizzuti, M. Pieretti, J. S. Sutcliffe, S. Richards, A. J. Verkerk, J. J. Holden, R. G. Fenwick, Jr., S. T. Warren, and et al. 1991. Variation of the CGG repeat at the fragile X site results in genetic instability: resolution of the Sherman paradox. Cell 67:1047-1058.

Galas, D. J., M. Eggert, and M. S. Waterman. 1985. Rigorous pattern-recognition methods for DNA sequences. Analysis of promoter sequences from Escherichia coli. J Mol Biol 186:117-128.

Garcia-Diaz, M., and T. A. Kunkel. 2006. Mechanism of a genetic glissando: structural biology of indel mutations. Trends Biochem Sci 31:206-214.

Garza, J. C., M. Slatkin, and N. B. Freimer. 1995. Microsatellite allele frequencies in humans and chimpanzees, with implications for constraints on allele size. Mol Biol Evol 12:594-603.

Gemmell, N. J., P. J. Allen, S. J. Goodman, and J. Z. Reed. 1997. Interspecific microsatellite markers for the study of pinniped populations. Mol Ecol 6:661-

666.

Gibbs, R. A.G. M. WeinstockM. L. MetzkerD. M. MuznyE. J. SodergrenS. SchererG. ScottD. SteffenK. C. WorleyP. E. BurchG. OkwuonuS. HinesL. LewisC. DeRamoO. DelgadoS. Dugan-RochaG. MinerM. MorganA. HawesR. GillCeleraR. A. HoltM. D. AdamsP. G. AmanatidesH. Baden-TillsonM. BarnsteadS. ChinC. A. EvansS. FerrieraC. FoslerA. GlodekZ. GuD. JenningsC. L. KraftT. NguyenC. M. PfannkochC. SitterG. G. SuttonJ. C. VenterT. WoodageD. SmithH. M. LeeE. GustafsonP. CahillA. KanaL. Doucette-StammK. WeinstockK. FechtelR. B. WeissD. M. DunnE. D. GreenR. W. BlakesleyG. G. BouffardP. J. De JongK. OsoegawaB. ZhuM. MarraJ. ScheinI. BosdetC. FjellS. JonesM. KrzywinskiC. MathewsonA. SiddiquiN. WyeJ. McPhersonS. ZhaoC. M. FraserJ. ShettyS. ShatsmanK. GeerY. ChenS. AbramzonW. C. NiermanP. H. HavlakR. ChenK. J. DurbinA. EganY. RenX. Z. SongB. LiY. LiuX. QinS. CawleyA. J. CooneyL. M. D'SouzaK. MartinJ. Q. WuM. L. Gonzalez-GarayA. R. JacksonK. J. KalafusM. P. McLeodA. MilosavljevicD. VirkA. VolkovD. A. WheelerZ. ZhangJ. A. BaileyE. E. EichlerE. TuzunE. BirneyE. MonginA. Ureta-VidalC. WoodwarkE. ZdobnovP. BorkM. SuyamaD. TorrentsM. AlexanderssonB. J. TraskJ. M. YoungH. HuangH. WangH. XingS. DanielsD. GietzenJ. SchmidtK. StevensU. VittJ. WingroveF. CamaraM. Mar AlbaJ. F. AbrilR. GuigoA. SmitI. DubchakE. M. RubinO. CouronneA. PoliakovN. HubnerD. GantenC. GoeseleO. HummelT. KreitlerY. A. LeeJ. MontiH. SchulzH. ZimdahlH. HimmelbauerH. LehrachH. J. JacobS. BrombergJ. Gullings-HandleyM. I. Jensen-SeamanA. E. KwitekJ. LazarD. PaskoP. J. TonellatoS. TwiggerC. P. PontingJ. M. DuarteS. RiceL. GoodstadtS. A. BeatsonR. D. EmesE. E. WinterC. WebberP. BrandtG. NyakaturaM. AdetobiF. ChiaromonteL. ElnitskiP. EswaraR. C. HardisonM. HouD. KolbeK. MakovaW. MillerA. NekrutenkoC. RiemerS. SchwartzJ. TaylorS. YangY. ZhangK. LindpaintnerT. D. AndrewsM. CaccamoM. ClampL. ClarkeV. CurwenR. DurbinE. EyrasS. M. SearleG. M. CooperS. BatzoglouM. BrudnoA. SidowE. A. StoneB. A. PayseurG. BourqueC. Lopez-OtinX. S. PuenteK. ChakrabartiS. ChatterjiC. DeweyL. PachterN. BrayV. B. YapA. CaspiG. TeslerP. A. PevznerD. HausslerK. M. RoskinR. BaertschH. ClawsonT. S. FureyA. S. HinrichsD. KarolchikW. J. KentK. R. RosenbloomH. TrumbowerM. WeirauchD. N. CooperP. D. StensonB. MaM. BrentM. ArumugamD. ShteynbergR. R. CopleyM. S. TaylorH. RiethmanU. MudunuriJ. PetersonM. GuyerA. FelsenfeldS. OldS. Mockrin, and F. Collins. 2004. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. Nature 428:493-521.

Glew, M. D., N. Baseggio, P. F. Markham, G. F. Browning, and I. D. Walker. 1998. Expression of the pMGA genes of Mycoplasma gallisepticum is controlled by variation in the GAA trinucleotide repeat lengths within the 5' noncoding regions. Infect Immun 66:5833-5841.

Goldstein, D. B., and D. D. Pollock. 1997. Launching microsatellites: a review of mutation processes and methods of phylogenetic interference. J Hered 88:335-342.

Gragg, H., B. D. Harfe, and S. Jinks-Robertson. 2002. Base composition of mononucleotide runs affects DNA polymerase slippage and removal of frameshift intermediates by mismatch repair in Saccharomyces cerevisiae. Mol

Cell Biol 22:8756-8762.

Grimwood, J., L. Olinger, and R. S. Stephens. 2001. Expression of Chlamydia pneumoniae polymorphic membrane protein family genes. Infect Immun 69:2383-2389.

Gur-Arie, R., C. J. Cohen, Y. Eitan, L. Shelef, E. M. Hallerman, and Y. Kashi. 2000. Simple sequence repeats in Escherichia coli: abundance, distribution, composition, and polymorphism. Genome Res 10:62-71.

Hammerschmidt, S., A. Muller, H. Sillmann, M. Muhlenhoff, R. Borrow, A. Fox, J. van Putten, W. D. Zollinger, R. Gerardy-Schahn, and M. Frosch. 1996. Capsule phase variation in Neisseria meningitidis serogroup B by slipped-strand mispairing in the polysialyltransferase gene (siaD): correlation with bacterial invasion and the outbreak of meningococcal disease. Mol Microbiol 20:1211-1220.

Hammock, E. A., and L. J. Young. 2005. Microsatellite instability generates diversity in brain and sociobehavioral traits. Science 308:1630-1634.

Hammock, E. A., and L. J. Young. 2004. Functional microsatellite polymorphism associated with divergent social structure in vole species. Mol Biol Evol 21:1057-1063.

Hancock, J. M. 1996. Simple sequences in a "minimal' genome. Nat Genet 14:14-15.

Harfe, B. D., and S. Jinks-Robertson. 2000. Sequence composition and context effects on the generation and repair of frameshift intermediates in mononucleotide runs in Saccharomyces cerevisiae. Genetics 156:571-578.

Harr, B., J. Todorova, and C. Schlotterer. 2002. Mismatch repair-driven mutational bias in D. melanogaster. Mol Cell 10:199-205.

Harr, B., B. Zangerl, G. Brem, and C. Schlotterer. 1998. Conservation of locus-specific microsatellite variability across species: a comparison of two Drosophila sibling species, D. melanogaster and D. simulans. Mol Biol Evol 15:176-184.

Henderson, S. T., and T. D. Petes. 1992. Instability of simple sequence DNA in Saccharomyces cerevisiae. Mol Cell Biol 12:2749-2757.

Hendrixson, D. R. 2006. A phase-variable mechanism controlling the Campylobacter jejuni FlgR response regulator influences commensalism. Mol Microbiol 61:1646-1659.

High, N. J., M. E. Deadman, and E. R. Moxon. 1993. The role of a repetitive DNA motif (5'-CAAT-3') in the variable expression of the Haemophilus influenzae lipopolysaccharide epitope alpha Gal(1-4)beta Gal. Mol Microbiol 9:1275-1282.

High, N. J., M. P. Jennings, and E. R. Moxon. 1996. Tandem repeats of the tetramer 5'-CAAT-3' present in lic2A are required for phase variation but not lipopolysaccharide biosynthesis in Haemophilus influenzae. Mol Microbiol 20:165-174.

Himmelreich, R., H. Hilbert, H. Plagens, E. Pirkl, B. C. Li, and R. Herrmann. 1996. Complete sequence analysis of the genome of the bacterium Mycoplasma pneumoniae. Nucleic Acids Res 24:4420-4449.

Hinchliffe, S. J., K. E. Isherwood, R. A. Stabler, M. B. Prentice, A. Rakin, R. A. Nichols, P. C. Oyston, J. Hinds, R. W. Titball, and B. W. Wren. 2003. Application of DNA microarrays to study the evolutionary genomics of Yersinia pestis and Yersinia pseudotuberculosis. Genome Res 13:2018-2029.

Hiriyanna, K. T., and T. Ramakrishnan. 1986. Deoxyribonucleic acid replication time in

Mycobacterium tuberculosis H37 Rv. Arch Microbiol 144:105-109.

Hood, D. W., M. E. Deadman, M. P. Jennings, M. Bisercic, R. D. Fleischmann, J. C. Venter, and E. R. Moxon. 1996. DNA repeats identify novel virulence genes in Haemophilus influenzae. Proc Natl Acad Sci U S A 93:11121-11125.

Hsieh, P. 2001. Molecular mechanisms of DNA mismatch repair. Mutat Res 486:71-87.

Huang, Q. Y., F. H. Xu, H. Shen, H. Y. Deng, Y. J. Liu, Y. Z. Liu, J. L. Li, R. R. Recker, and H. W. Deng. 2002. Mutation patterns at dinucleotide microsatellite loci in humans. Am J Hum Genet 70:625-634.

Inglesby, T. V., D. T. Dennis, D. A. Henderson, J. G. Bartlett, M. S. Ascher, E. Eitzen, A. D. Fine, A. M. Friedlander, J. Hauer, J. F. Koerner, M. Layton, J. McDade, M. T. Osterholm, T. O'Toole, G. Parker, T. M. Perl, P. K. Russell, M. Schoch-Spana, and K. Tonat. 2000. Plague as a biological weapon: medical and public health management. Working Group on Civilian Biodefense. Jama 283:2281-2290.

Inman, R. B. 1966. A denaturation map of the lambda phage DNA molecule determined by electron microscopy. J Mol Biol 18:464-476.

Inzana, T. J., J. Hensley, J. McQuiston, A. J. Lesse, A. A. Campagnari, S. M. Boyle, and M. A. Apicella. 1997. Phase variation and conservation of lipooligosaccharide epitopes in Haemophilus somnus. Infect Immun 65:4675-4681.

Ionov, Y., M. A. Peinado, S. Malkhosyan, D. Shibata, and M. Perucho. 1993. Ubiquitous somatic mutations in simple repeated sequences reveal a new mechanism for colonic carcinogenesis. Nature 363:558-561.

Itoh, Y., X. Wang, B. J. Hinnebusch, J. F. Preston, 3rd, and T. Romeo. 2005. Depolymerization of beta-1,6-N-acetyl-D-glucosamine disrupts the integrity of diverse bacterial biofilms. J Bacteriol 187:382-387.

Itzkovitz, S., and U. Alon. 2007. The genetic code is nearly optimal for allowing additional information within protein-coding sequences. Genome Res 17:405-412.

Iyer, R. R., A. Pluciennik, V. Burdett, and P. L. Modrich. 2006. DNA mismatch repair: functions and mechanisms. Chem Rev 106:302-323.

Jaworski, A., W. A. Rosche, R. Gellibolian, S. Kang, M. Shimizu, R. P. Bowater, R. R. Sinden, and R. D. Wells. 1995. Mismatch repair in Escherichia coli enhances instability of (CTG)n triplet repeats from human hereditary diseases. Proc Natl Acad Sci U S A 92:11019-11023.

Jeffreys, A. J., V. Wilson, and S. L. Thein. 1985. Hypervariable 'minisatellite' regions in human DNA. Nature 314:67-73.

Jennings, M. P., Y. N. Srikhanta, E. R. Moxon, M. Kramer, J. T. Poolman, B. Kuipers, and P. van der Ley. 1999. The genetic basis of the phase variation repertoire of lipopolysaccharide immunotypes in Neisseria meningitidis. Microbiology 145 (Pt 11):3013-3021.

Jonsson, A. B., G. Nyberg, and S. Normark. 1991. Phase variation of gonococcal pili by frameshift mutation in pilC, a novel gene for pilus assembly. EMBO J 10:477-488.

Josenhans, C., K. A. Eaton, T. Thevenot, and S. Suerbaum. 2000. Switching of flagellar motility in Helicobacter pylori by reversible length variation of a short homopolymeric sequence repeat in fliP, a gene encoding a basal body protein. Infect Immun 68:4598-4603.

Kaplan, J. B., C. Ragunath, N. Ramasubbu, and D. H. Fine. 2003. Detachment of Actinobacillus actinomycetemcomitans biofilm cells by an endogenous beta-hexosaminidase activity. J Bacteriol 185:4693-4698.

Karlin, S., and C. Burge. 1995. Dinucleotide relative abundance extremes: a genomic signature. Trends Genet 11:283-290.

Karlin, S., J. Mrazek, and A. M. Campbell. 1997. Compositional biases of bacterial genomes and evolutionary implications. J Bacteriol 179:3899-3913.

Karlyshev, A. V., D. Linton, N. A. Gregson, and B. W. Wren. 2002. A novel paralogous gene family involved in phase-variable flagella-mediated motility in Campylobacter jejuni. Microbiology 148:473-480.

Kashi, Y., D. King, and M. Soller. 1997. Simple sequence repeats as a source of quantitative genetic variation. Trends Genet 13:74-78.

Kashi, Y., and D. G. King. 2006. Simple sequence repeats as advantageous mutators in evolution. Trends Genet 22:253-259.

Katti, M. V., P. K. Ranjekar, and V. S. Gupta. 2001. Differential distribution of simple sequence repeats in eukaryotic genome sequences. Mol Biol Evol 18:1161-1167.

Kearns, D. B., F. Chu, R. Rudner, and R. Losick. 2004. Genes governing swarming in Bacillus subtilis and evidence for a phase variation mechanism controlling surface motility. Mol Microbiol 52:357-369.

Kelkar, Y. D., S. Tyekucheva, F. Chiaromonte, and K. D. Makova. 2008. The genome-wide determinants of human and chimpanzee microsatellite evolution. Genome Res 18:30-38.

Kenneson, A., F. Zhang, C. H. Hagedorn, and S. T. Warren. 2001. Reduced FMRP and increased FMR1 transcription is proportionally associated with CGG repeat number in intermediate-length and premutation carriers. Hum Mol Genet 10:1449-1454.

Kimmel, M., and R. Chakraborty. 1996. Measures of variation at DNA repeat loci under a general stepwise mutation model. Theor Popul Biol 50:345-367.

Kimmel, M., R. Chakraborty, D. N. Stivers, and R. Deka. 1996. Dynamics of repeat polymorphisms under a forward-backward mutation model: within- and between-population variability at microsatellite loci. Genetics 143:549-555.

Klevytska, A. M., L. B. Price, J. M. Schupp, P. L. Worsham, J. Wong, and P. Keim. 2001. Identification and characterization of variable-number tandem repeats in the Yersinia pestis genome. J Clin Microbiol 39:3179-3185.

Kofler, R., C. Schlotterer, and T. Lelley. 2007. SciRoKo: a new tool for whole genome microsatellite search and investigation. Bioinformatics 23:1683-1685.

Kolpakov, R., G. Bana, and G. Kucherov. 2003. mreps: Efficient and flexible detection of tandem repeats in DNA. Nucleic Acids Res 31:3672-3678.

Koob, M. D., M. L. Moseley, L. J. Schut, K. A. Benzow, T. D. Bird, J. W. Day, and L. P. Ranum. 1999. An untranslated CTG expansion causes a novel form of spinocerebellar ataxia (SCA8). Nat Genet 21:379-384.

Kornberg, A., L. L. Bertsch, J. F. Jackson, and H. G. Khorana. 1964. Enzymatic Synthesis of Deoxyribonucleic Acid, Xvi. Oligonucleotides as Templates and the Mechanism of Their Replication. Proc Natl Acad Sci U S A 51:315-323.

Kremer, E. J., M. Pritchard, M. Lynch, S. Yu, K. Holman, E. Baker, S. T. Warren, D. Schlessinger, G. R. Sutherland, and R. I. Richards. 1991. Mapping of DNA instability at the fragile X to a trinucleotide repeat sequence p(CCG)n. Science

252:1711-1714.

Kruglyak, S., R. T. Durrett, M. D. Schug, and C. F. Aquadro. 1998. Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations. Proc Natl Acad Sci U S A 95:10774-10778.

Kukkonen, M., M. Suomalainen, P. Kyllonen, K. Lahteenmaki, H. Lang, R. Virkola, I. M. Helander, O. Holst, and T. K. Korhonen. 2004. Lack of O-antigen is essential for plasminogen activation by Yersinia pestis and Salmonella enterica. Mol Microbiol 51:215-225.

Kunkel, T. A. 2004. DNA replication fidelity. J Biol Chem 279:16895-16898.

Lafontaine, E. R., L. D. Cope, C. Aebi, J. L. Latimer, G. H. McCracken, Jr., and E. J. Hansen. 2000. The UspA1 protein and a second type of UspA2 protein mediate adherence of Moraxella catarrhalis to human epithelial cells in vitro. J Bacteriol 182:1364-1373.

Lafontaine, E. R., N. J. Wagner, and E. J. Hansen. 2001. Expression of the Moraxella catarrhalis UspA1 protein undergoes phase variation and is regulated at the transcriptional level. J Bacteriol 183:1540-1551.

Lamers, M. H., A. Perrakis, J. H. Enzlin, H. H. Winterwerp, N. de Wind, and T. K. Sixma. 2000. The crystal structure of DNA mismatch repair protein MutS binding to a G x T mismatch. Nature 407:711-717.

Lander, E. S.L. M. LintonB. BirrenC. NusbaumM. C. ZodyJ. BaldwinK. DevonK. DewarM. DoyleW. FitzHughR. FunkeD. GageK. HarrisA. HeafordJ. HowlandL. KannJ. LehoczkyR. LeVineP. McEwanK. McKernanJ. MeldrimJ. P. MesirovC. MirandaW. MorrisJ. NaylorC. RaymondM. RosettiR. SantosA. SheridanC. SougnezN. Stange-ThomannN. StojanovicA. SubramanianD. WymanJ. RogersJ. SulstonR. AinscoughS. BeckD. BentleyJ. BurtonC. CleeN. CarterA. CoulsonR. DeadmanP. DeloukasA. DunhamI. DunhamR. DurbinL. FrenchD. GrafhamS. GregoryT. HubbardS. HumphrayA. HuntM. JonesC. LloydA. McMurrayL. MatthewsS. MercerS. MilneJ. C. MullikinA. MungallR. PlumbM. RossR. ShownkeenS. SimsR. H. WaterstonR. K. WilsonL. W. HillierJ. D. McPhersonM. A. MarraE. R. MardisL. A. FultonA. T. ChinwallaK. H. PepinW. R. GishS. L. ChissoeM. C. WendlK. D. DelehauntyT. L. MinerA. DelehauntyJ. B. KramerL. L. CookR. S. FultonD. L. JohnsonP. J. MinxS. W. CliftonT. HawkinsE. BranscombP. PredkiP. RichardsonS. WenningT. SlezakN. DoggettJ. F. ChengA. OlsenS. LucasC. ElkinE. UberbacherM. FrazierR. A. GibbsD. M. MuznyS. E. SchererJ. B. BouckE. J. SodergrenK. C. WorleyC. M. RivesJ. H. GorrellM. L. MetzkerS. L. NaylorR. S. KucherlapatiD. L. NelsonG. M. WeinstockY. SakakiA. FujiyamaM. HattoriT. YadaA. ToyodaT. ItohC. KawagoeH. WatanabeY. TotokiT. TaylorJ. WeissenbachR. HeiligW. SaurinF. ArtiguenaveP. BrottierT. BrulsE. PelletierC. RobertP. WinckerD. R. SmithL. Doucette-StammM. RubenfieldK. WeinstockH. M. LeeJ. DuboisA. RosenthalM. PlatzerG. NyakaturaS. TaudienA. RumpH. YangJ. YuJ. WangG. HuangJ. GuL. HoodL. RowenA. MadanS. QinR. W. DavisN. A. FederspielA. P. AbolaM. J. ProctorR. M. MyersJ. SchmutzM. DicksonJ. GrimwoodD. R. CoxM. V. OlsonR. KaulN. ShimizuK. KawasakiS. MinoshimaG. A. EvansM. AthanasiouR. SchultzB. A. RoeF. ChenH. PanJ. RamserH. LehrachR. ReinhardtW. R. McCombieM. de la BastideN. DedhiaH. BlockerK. HornischerG. NordsiekR. AgarwalaL. AravindJ. A. BaileyA. BatemanS. BatzoglouE. BirneyP. BorkD. G. BrownC. B. BurgeL.

CeruttiH. C. ChenD. ChurchM. ClampR. R. CopleyT. DoerksS. R. EddyE. E. EichlerT. S. FureyJ. GalaganJ. G. GilbertC. HarmonY. HayashizakiD. HausslerH. HermjakobK. HokampW. JangL. S. JohnsonT. A. JonesS. KasifA. KaspryzkS. KennedyW. J. KentP. KittsE. V. KooninI. KorfD. KulpD. LancetT. M. LoweA. McLysaghtT. MikkelsenJ. V. MoranN. MulderV. J. PollaraC. P. PontingG. SchulerJ. SchultzG. SlaterA. F. SmitE. StupkaJ. SzustakowskiD. Thierry-MiegJ. Thierry-MiegL. WagnerJ. WallisR. WheelerA. WilliamsY. I. WolfK. H. WolfeS. P. YangR. F. YehF. CollinsM. S. GuyerJ. PetersonA. FelsenfeldK. A. WetterstrandA. PatrinosM. J. MorganP. de JongJ. J. CataneseK. OsoegawaH. ShizuyaS. Choi, and Y. J. Chen. 2001. Initial sequencing and analysis of the human genome. Nature 409:860-921.

Lathem, W. W., P. A. Price, V. L. Miller, and W. E. Goldman. 2007. A plasminogen-activating protease specifically controls the development of primary pneumonic plague. Science 315:509-513.

Le Fleche, P., Y. Hauck, L. Onteniente, A. Prieur, F. Denoeud, V. Ramisse, P. Sylvestre, G. Benson, F. Ramisse, and G. Vergnaud. 2001. A tandem repeats database for bacterial genomes: application to the genotyping of Yersinia pestis and Bacillus anthracis. BMC Microbiol 1:2.

Leclercq, S., E. Rivals, and P. Jarne. 2007. Detecting microsatellites within genomes: significant variation among algorithms. BMC Bioinformatics 8:125.

Lee, J. Y., J. Chang, N. Joseph, R. Ghirlando, D. N. Rao, and W. Yang. 2005. MutH complexed with hemi- and unmethylated DNAs: coupling base recognition and DNA cleavage. Mol Cell 20:155-166.

Lesnik, E. A., R. Sampath, H. B. Levene, T. J. Henderson, J. A. McNeil, and D. J. Ecker. 2001. Prediction of rho-independent transcriptional terminators in Escherichia coli. Nucleic Acids Res 29:3583-3594.

Levinson, G., and G. A. Gutman. 1987a. Slipped-strand mispairing: a major mechanism for DNA sequence evolution. Mol Biol Evol 4:203-221.

Levinson, G., and G. A. Gutman. 1987b. High frequencies of short frameshifts in poly-CA/TG tandem repeats borne by bacteriophage M13 in Escherichia coli K-12. Nucleic Acids Res 15:5323-5338.

Levy, D. D., and T. A. Cebula. 2001. Fidelity of replication of repetitive DNA in mutS and repair proficient Escherichia coli. Mutat Res 474:1-14.

Lewis, L. A., M. Gipson, K. Hartman, T. Ownbey, J. Vaughn, and D. W. Dyer. 1999. Phase variation of HpuAB and HmbR, two distinct haemoglobin receptors of Neisseria meningitidis DNM2. Mol Microbiol 32:977-989.

Li, G. M. 2008. Mechanisms and functions of DNA mismatch repair. Cell Res 18:85-98.

Li, Y. C., A. B. Korol, T. Fahima, and E. Nevo. 2004. Microsatellites within genes: structure, function, and evolution. Mol Biol Evol 21:991-1007.

Lin, Z., M. Nei, and H. Ma. 2007. The origins and early evolution of DNA mismatch repair genes--multiple horizontal gene transfers and co-evolution. Nucleic Acids Res 35:7591-7603.

Linton, D., M. Gilbert, P. G. Hitchen, A. Dell, H. R. Morris, W. W. Wakarchuk, N. A. Gregson, and B. W. Wren. 2000. Phase variation of a beta-1,3 galactosyltransferase involved in generation of the ganglioside GM1-like lipo-oligosaccharide of Campylobacter jejuni. Mol Microbiol 37:501-514.

Liquori, C. L., K. Ricker, M. L. Moseley, J. F. Jacobsen, W. Kress, S. L. Naylor, J. W.

Day, and L. P. Ranum. 2001. Myotonic dystrophy type 2 caused by a CCTG expansion in intron 1 of ZNF9. Science 293:864-867.

Litt, M., and J. A. Luty. 1989. A hypervariable microsatellite revealed by in vitro amplification of a dinucleotide repeat within the cardiac muscle actin gene. Am J Hum Genet 44:397-401.

Makino, S., J. P. van Putten, and T. F. Meyer. 1991. Phase variation of the opacity outer membrane protein controls invasion by Neisseria gonorrhoeae into human epithelial cells. Embo J 10:1307-1315.

Markowitz, S., J. Wang, L. Myeroff, R. Parsons, L. Sun, J. Lutterbaugh, R. S. Fan, E. Zborowska, K. W. Kinzler, B. Vogelstein, and et al. 1995. Inactivation of the type II TGF-beta receptor in colon cancer cells with microsatellite instability. Science 268:1336-1338.

Marra, G., and C. R. Boland. 1995. Hereditary nonpolyposis colorectal cancer: the syndrome, the genes, and historical perspectives. J Natl Cancer Inst 87:1114-1125.

Martin, P., K. Makepeace, S. A. Hill, D. W. Hood, and E. R. Moxon. 2005. Microsatellite instability regulates transcription factor binding and gene expression. Proc Natl Acad Sci U S A 102:3800-3804.

Matsuura, T., T. Yamagata, D. L. Burgess, A. Rasmussen, R. P. Grewal, K. Watase, M. Khajavi, A. E. McCall, C. F. Davis, L. Zu, M. Achari, S. M. Pulst, E. Alonso, J. L. Noebels, D. L. Nelson, H. Y. Zoghbi, and T. Ashizawa. 2000. Large expansion of the ATTCT pentanucleotide repeat in spinocerebellar ataxia type 10. Nat Genet 26:191-194.

Meloni, R., V. Albanese, P. Ravassard, F. Treilhou, and J. Mallet. 1998. A tetranucleotide polymorphic microsatellite, located in the first intron of the tyrosine hydroxylase gene, acts as a transcription regulatory element in vitro. Hum Mol Genet 7:423-428.

Metzgar, D., J. Bytof, and C. Wills. 2000. Selection against frameshift mutations limits microsatellite expansion in coding DNA. Genome Res 10:72-80.

Metzgar, D., L. Liu, C. Hansen, K. Dybvig, and C. Wills. 2002. Domain-level differences in microsatellite distribution and content result from different relative rates of insertion and deletion mutations. Genome Res 12:408-413.

Mira, A., H. Ochman, and N. A. Moran. 2001. Deletional bias and the evolution of bacterial genomes. Trends Genet 17:589-596.

Mirkin, E. V., and S. M. Mirkin. 2005. Mechanisms of transcription-replication collisions in bacteria. Mol Cell Biol 25:888-895.

Mirkin, S. M. 2005. Toward a unified theory for repeat expansions. Nat Struct Mol Biol 12:635-637.

Mizrahi, V., and S. J. Andersen. 1998. DNA repair in Mycobacterium tuberculosis. What have we learnt from the genome sequence? Mol Microbiol 29:1331-1339.

Modrich, P., and R. Lahue. 1996. Mismatch repair in replication fidelity, genetic recombination, and cancer biology. Annu Rev Biochem 65:101-133.

Morel, P., C. Reverdy, B. Michel, S. D. Ehrlich, and E. Cassuto. 1998. The role of SOS and flap processing in microsatellite instability in Escherichia coli. Proc Natl Acad Sci U S A 95:10003-10008.

Morgante, M., M. Hanafey, and W. Powell. 2002. Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes. Nat Genet 30:194-200.

Moxon, E. R., P. B. Rainey, M. A. Nowak, and R. E. Lenski. 1994. Adaptive evolution of highly mutable loci in pathogenic bacteria. Curr Biol 4:24-33.

Moxon, R., C. Bayliss, and D. Hood. 2006. Bacterial contingency Loci: the role of simple sequence DNA repeats in bacterial adaptation. Annu Rev Genet 40:307-333.

Mrazek, J. 2006. Analysis of distribution indicates diverse functions of simple sequence repeats in Mycoplasma genomes. Mol Biol Evol 23:1370-1385.

Mrazek, J., X. Guo, and A. Shah. 2007. Simple sequence repeats in prokaryotic genomes. Proc Natl Acad Sci U S A 104:8472-8477.

Mudunuri, S. B., and H. A. Nagarajaram. 2007. IMEx: Imperfect Microsatellite Extractor. Bioinformatics 23:1181-1187.

Mudunuri, S. B., A. A. Rao, P. Mishra, and H. A. Nagarajaram. 2010. Comparative Analysis of Microsatellite Detecting Software: A Significant Variation in Results and Influence of Parameters. International Symposium on Biocomputing, 15-17 February 2010.

Murphy, G. L., T. D. Connell, D. S. Barritt, M. Koomey, and J. G. Cannon. 1989. Phase variation of gonococcal protein II: regulation of gene expression by slipped-strand mispairing of a repetitive DNA sequence. Cell 56:539-547.

Nauta, M. J., and F. J. Weissing. 1996. Constraints on allele size at microsatellite loci: implications for genetic differentiation. Genetics 143:1021-1032.

Nelson, M., and M. McClelland. 1991. Site-specific methylation: effect on DNA modification methyltransferases and restriction endonucleases. Nucleic Acids Res 19 Suppl:2045-2071.

Nilsson, A. I., S. Koskiniemi, S. Eriksson, E. Kugelberg, J. C. Hinton, and D. I. Andersson. 2005. Bacterial genome size reduction by experimental evolution. Proc Natl Acad Sci U S A 102:12112-12116.

O'Hearn, E., S. E. Holmes, P. C. Calvert, C. A. Ross, and R. L. Margolis. 2001. SCA-12: Tremor with cerebellar and cortical atrophy is associated with a CAG repeat expansion. Neurology 56:299-303.

Obmolova, G., C. Ban, P. Hsieh, and W. Yang. 2000. Crystal structures of mismatch repair protein MutS and its complex with a substrate DNA. Nature 407:703-710.

Ota, T., and M. Kimura. 1973. A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. Genet Res 22:201-204.

Parisi, V., V. De Fonzo, and F. Aluffi-Pentini. 2003. STRING: finding tandem repeats in DNA sequences. Bioinformatics 19:1733-1738.

Park, S. F., D. Purdy, and S. Leach. 2000. Localized reversible frameshift mutation in the flhA gene confers phase variability to flagellin gene expression in Campylobacter coli. J Bacteriol 182:207-210.

Parker, B. O., and M. G. Marinus. 1992. Repair of DNA heteroduplexes containing small heterologous sequences in Escherichia coli. Proc Natl Acad Sci U S A 89:1730-1734.

Parkhill, J., B. W. Wren, N. R. Thomson, R. W. Titball, M. T. Holden, M. B. Prentice, M. Sebaihia, K. D. James, C. Churcher, K. L. Mungall, S. Baker, D. Basham, S. D. Bentley, K. Brooks, A. M. Cerdeno-Tarraga, T. Chillingworth, A. Cronin, R. M. Davies, P. Davis, G. Dougan, T. Feltwell, N. Hamlin, S. Holroyd, K. Jagels, A. V. Karlyshev, S. Leather, S. Moule, P. C. Oyston, M. Quail, K. Rutherford, M. Simmonds, J. Skelton, K. Stevens, S. Whitehead, and B. G. Barrell. 2001.

Genome sequence of Yersinia pestis, the causative agent of plague. Nature 413:523-527.

Peak, I. R., M. P. Jennings, D. W. Hood, M. Bisercic, and E. R. Moxon. 1996. Tetrameric repeat units associated with virulence factor phase variation in Haemophilus also occur in Neisseria spp. and Moraxella catarrhalis. FEMS Microbiol Lett 137:109-114.

Pearson, C. E., K. Nichol Edamura, and J. D. Cleary. 2005. Repeat instability: mechanisms of dynamic mutations. Nat Rev Genet 6:729-742.

Pericone, C. D., D. Bae, M. Shchepetov, T. McCool, and J. N. Weiser. 2002. Short-sequence tandem and nontandem DNA repeats and endogenous hydrogen peroxide production contribute to genetic instability of Streptococcus pneumoniae. J Bacteriol 184:4392-4399.

Persson, A., K. Jacobsson, L. Frykberg, K. E. Johansson, and F. Poumarat. 2002. Variable surface protein Vmm of Mycoplasma mycoides subsp. mycoides small colony type. J Bacteriol 184:3712-3722.

Pevzner, P. A., M. Borodovsky, and A. A. Mironov. 1989a. Linguistics of nucleotide sequences. II: Stationary words in genetic texts and the zonal structure of DNA. J Biomol Struct Dyn 6:1027-1038.

Pevzner, P. A., M. Borodovsky, and A. A. Mironov. 1989b. Linguistics of nucleotide sequences. I: The significance of deviations from mean statistical characteristics and prediction of the frequencies of occurrence of words. J Biomol Struct Dyn 6:1013-1026.

Pouillot, F., A. Derbise, M. Kukkonen, J. Foulon, T. K. Korhonen, and E. Carniel. 2005. Evaluation of O-antigen inactivation on Pla activity and virulence of Yersinia pseudotuberculosis harbouring the pPla plasmid. Microbiology 151:3759-3768.

Pouillot, F., C. Fayolle, and E. Carniel. 2008. Characterization of chromosomal regions conserved in Yersinia pseudotuberculosis and lost by Yersinia pestis. Infect Immun 76:4592-4599.

Pourcel, C., F. Andre-Mazeaud, H. Neubauer, F. Ramisse, and G. Vergnaud. 2004. Tandem repeats analysis for the high resolution phylogenetic analysis of Yersinia pestis. BMC Microbiol 4:22.

Press, W. H., B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling. 1992. Numerical recipes in C (Cambridge: Cambridge University Press).

Primmer, C. R., and H. Ellegren. 1998. Patterns of molecular evolution in avian microsatellites. Mol Biol Evol 15:997-1008.

Primmer, C. R., T. Raudsepp, B. P. Chowdhary, A. P. Moller, and H. Ellegren. 1997. Low frequency of microsatellites in the avian genome. Genome Res 7:471-482.

Primmer, C. R., N. Saino, A. P. Moller, and H. Ellegren. 1996. Directional evolution in germline microsatellite mutations. Nat Genet 13:391-393.

Pupko, T., and D. Graur. 1999. Evolution of microsatellites in the yeast Saccharomyces cerevisiae: role of length and number of repeated units. J Mol Evol 48:313-316.

Radman, M. 1998. DNA replication: one strand may be more equal. Proc Natl Acad Sci U S A 95:9718-9719.

Rampino, N., H. Yamamoto, Y. Ionov, Y. Li, H. Sawai, J. C. Reed, and M. Perucho. 1997. Somatic frameshift mutations in the BAX gene in colon cancers of the microsatellite mutator phenotype. Science 275:967-969.

Rasmussen, M., and L. Bjorck. 2001. Unique regulation of SclB - a novel collagen-like surface protein of Streptococcus pyogenes. Mol Microbiol 40:1427-1438.

Rassmann, K., C. Schlotterer, and D. Tautz. 1991. Isolation of simple-sequence loci for use in polymerase chain reaction-based DNA fingerprinting. Electrophoresis 12:113-118.

Revell, P. A., and V. L. Miller. 2001. Yersinia virulence: more than a plasmid. FEMS Microbiol Lett 205:159-164.

Richard, G. F., A. Kerrest, and B. Dujon. 2008. Comparative genomics and molecular dynamics of DNA repeats in eukaryotes. Microbiol Mol Biol Rev 72:686-727.

Ritz, D., J. Lim, C. M. Reynolds, L. B. Poole, and J. Beckwith. 2001. Conversion of a peroxiredoxin into a disulfide reductase by a triplet repeat expansion. Science 294:158-160.

Robertson, B. D., and T. F. Meyer. 1992. Genetic variation in pathogenic bacteria. Trends Genet 8:422-427.

Rocha, E. P. 2008. The organization of the bacterial genome. Annu Rev Genet 42:211-233.

Roche, R. J., and E. R. Moxon. 1995. Phenotypic variation in Haemophilus influenzae: the interrelationship of colony opacity, capsule and lipopolysaccharide. Microb Pathog 18:129-140.

Rose, O., and D. Falush. 1998. A threshold size for microsatellite expansion. Mol Biol Evol 15:613-615.

Rosqvist, R., M. Skurnik, and H. Wolf-Watz. 1988. Increased virulence of Yersinia pseudotuberculosis by two independent mutations. Nature 334:522-524.

Rubinsztein, D. C., W. Amos, J. Leggo, S. Goodburn, S. Jain, S. H. Li, R. L. Margolis, C. A. Ross, and M. A. Ferguson-Smith. 1995. Microsatellite evolution--evidence for directionality and variation in rate between species. Nat Genet 10:337-343.

Rubinsztein, D. C., J. Leggo, and W. Amos. 1995. Microsatellites evolve more rapidly in humans than in chimpanzees. Genomics 30:610-612.

Santibanez-Koref, M. F., R. Gangeswaran, and J. M. Hancock. 2001. A relationship between lengths of microsatellites and nearby substitution rates in mammalian genomes. Mol Biol Evol 18:2119-2123.

Sarkari, J., N. Pandit, E. R. Moxon, and M. Achtman. 1994. Variable expression of the Opc outer membrane protein in Neisseria meningitidis is caused by size variation of a promoter containing poly-cytidine. Mol Microbiol 13:207-217.

Sawyer, L. A., J. M. Hennessy, A. A. Peixoto, E. Rosato, H. Parkinson, R. Costa, and C. P. Kyriacou. 1997. Natural variation in a Drosophila clock gene and temperature compensation. Science 278:2117-2120.

Schlotterer, C. 2000. Evolutionary dynamics of microsatellite DNA. Chromosoma 109:365-371.

Schlotterer, C. 1998. Genome evolution: are microsatellites really simple sequences? Curr Biol 8:R132-134.

Schlotterer, C., M. Imhof, H. Wang, V. Nolte, and B. Harr. 2006. Low abundance of Escherichia coli microsatellites is associated with an extremely low mutation rate. J Evol Biol 19:1671-1676.

Schlotterer, C., and D. Tautz. 1992. Slippage synthesis of simple sequence DNA. Nucleic Acids Res 20:211-215.

Schofield, M. J., and P. Hsieh. 2003. DNA mismatch repair: molecular mechanisms and biological function. Annu Rev Microbiol 57:579-608.

Schultz, J., F. Milpetz, P. Bork, and C. P. Ponting. 1998. SMART, a simple modular architecture research tool: identification of signaling domains. Proc Natl Acad Sci U S A 95:5857-5864.

Segura, A., A. Hurtado, E. Duque, and J. L. Ramos. 2004. Transcriptional phase variation at the flhB gene of Pseudomonas putida DOT-T1E is involved in response to environmental changes and suggests the participation of the flagellar export system in solvent tolerance. J Bacteriol 186:1905-1909.

Selmane, T., M. J. Schofield, S. Nayak, C. Du, and P. Hsieh. 2003. Formation of a DNA mismatch repair complex mediated by ATP. J Mol Biol 334:949-965.

Shriver, M. D., L. Jin, R. Chakraborty, and E. Boerwinkle. 1993. VNTR allele frequency distributions under the stepwise mutation model: a computer simulation approach. Genetics 134:983-993.

Sixma, T. K. 2001. DNA mismatch repair: MutS structures bound to mismatches. Curr Opin Struct Biol 11:47-52.

Skurnik, M., and J. A. Bengoechea. 2003. The biosynthesis and biological role of lipopolysaccharide O-antigens of pathogenic Yersiniae. Carbohydr Res 338:2521-2529.

Smit, A., Hubley, R., and Green, P. . 1996. RepeatMasker {Error! Hyperlink reference not valid..

Smith, B. T., A. D. Grossman, and G. C. Walker. 2001. Visualization of mismatch repair in bacterial cells. Mol Cell 8:1197-1206.

Smith, M. G., T. A. Gianoulis, S. Pukatzki, J. J. Mekalanos, L. N. Ornston, M. Gerstein, and M. Snyder. 2007. New insights into Acinetobacter baumannii pathogenesis revealed by high-density pyrosequencing and transposon mutagenesis. Genes Dev 21:601-614.

Song, Y., Z. Tong, J. Wang, L. Wang, Z. Guo, Y. Han, J. Zhang, D. Pei, D. Zhou, H. Qin, X. Pang, Y. Han, J. Zhai, M. Li, B. Cui, Z. Qi, L. Jin, R. Dai, F. Chen, S. Li, C. Ye, Z. Du, W. Lin, J. Wang, J. Yu, H. Yang, J. Wang, P. Huang, and R. Yang. 2004. Complete genome sequence of Yersinia pestis strain 91001, an isolate avirulent to humans. DNA Res 11:179-197.

Sreenu, V. B. 2006. Computational Studies on Microsatellites in Prokaryotic Genomes with Special Reference to Mycobacterial Genomes. PhD Thesis.

Sreenu, V. B., V. Alevoor, J. Nagaraju, and H. A. Nagarajaram. 2003a. MICdb: database of prokaryotic microsatellites. Nucleic Acids Res 31:106-108.

Sreenu, V. B., P. Kumar, J. Nagaraju, and H. A. Nagarajam. 2007. Simple sequence repeats in mycobacterial genomes. J Biosci 32:3-15.

Sreenu, V. B., P. Kumar, J. Nagaraju, and H. A. Nagarajaram. 2006. Microsatellite polymorphism across the M. tuberculosis and M. bovis genomes: implications on genome evolution and plasticity. BMC Genomics 7:78.

Sreenu, V. B., G. Ranjitkumar, S. Swaminathan, S. Priya, B. Bose, M. N. Pavan, G. Thanu, J. Nagaraju, and H. A. Nagarajaram. 2003b. MICAS: a fully automated web server for microsatellite extraction and analysis from prokaryote and viral genomic sequences. Appl Bioinformatics 2:165-168.

Stern, A., M. Brown, P. Nickel, and T. F. Meyer. 1986. Opacity genes in Neisseria gonorrhoeae: control of phase and antigenic variation. Cell 47:61-71.

Stern, A., and T. F. Meyer. 1987. Common mechanism controlling phase and antigenic variation in pathogenic neisseriae. Mol Microbiol 1:5-12.

Strand, M., T. A. Prolla, R. M. Liskay, and T. D. Petes. 1993. Destabilization of tracts of simple repetitive DNA in yeast by mutations affecting DNA mismatch repair. Nature 365:274-276.

Streelman, J. T., and T. D. Kocher. 2002. Microsatellite variation associated with prolactin expression and growth of salt-challenged tilapia. Physiol Genomics 9:1-4.

Streisinger, G., Y. Okada, J. Emrich, J. Newton, A. Tsugita, E. Terzaghi, and M. Inouye. 1966. Frameshift mutations and the genetic code. This paper is dedicated to Professor Theodosius Dobzhansky on the occasion of his 66th birthday. Cold Spring Harb Symp Quant Biol 31:77-84.

Tannaes, T., N. Dekker, G. Bukholm, J. J. Bijlsma, and B. J. Appelmelk. 2001. Phase variation in the Helicobacter pylori phospholipase A gene and its role in acid adaptation. Infect Immun 69:7334-7340.

Tautz, D. 1993. Notes on the definition and nomenclature of tandemly repetitive DNA sequences. EXS 67:21-28.

Tautz, D., and Schlotterer. 1994. Simple sequences. Curr Opin Genet Dev 4:832-837.

Telford, J. L. 2008. Bacterial genome variability and its impact on vaccine design. Cell Host Microbe 3:408-416.

Temnykh, S., G. DeClerck, A. Lukashova, L. Lipovich, S. Cartinhour, and S. McCouch. 2001. Computational and experimental analysis of microsatellites in rice (Oryza sativa L.): frequency, length variation, transposon associations, and genetic marker potential. Genome Res 11:1441-1452.

Theiss, P., and K. S. Wise. 1997. Localized frameshift mutation generates selective, high-frequency phase variation of a surface lipoprotein encoded by a mycoplasma ABC transporter operon. J Bacteriol 179:4013-4022.

Thibodeau, S. N., G. Bren, and D. Schaid. 1993. Microsatellite instability in cancer of the proximal colon. Science 260:816-819.

Thurston, M. I., Field, D. 2005. Msatfinder: Detection and Characterization of Microsatellites. Oxford: Centre for Ecology and Hydrology. 2005 Available at: http://www.genomics.ceh.ac.uk/msatfinder.

Tidow, N., A. Boecker, H. Schmidt, K. Agelopoulos, W. Boecker, H. Buerger, and B. Brandt. 2003. Distinct amplification of an untranslated regulatory sequence in the egfr gene contributes to early steps in breast cancer development. Cancer Res 63:1172-1178.

Toth, G., Z. Gaspari, and J. Jurka. 2000. Microsatellites in different eukaryotic genomes: survey and analysis. Genome Res 10:967-981.

Twerdi, C. D., J. C. Boyer, and R. A. Farber. 1999. Relative rates of insertion and deletion mutations in a microsatellite sequence in cultured cells. Proc Natl Acad Sci U S A 96:2875-2879.

Usdin, K., and E. Grabczyk. 2000. DNA repeat expansions and human disease. Cell Mol Life Sci 57:914-931.

Valdes, A. M., M. Slatkin, and N. B. Freimer. 1993. Allele frequencies at microsatellite loci: the stepwise mutation model revisited. Genetics 133:737-749.

van Belkum, A., S. Scherer, L. van Alphen, and H. Verbrugh. 1998. Short-sequence DNA repeats in prokaryotic genomes. Microbiol Mol Biol Rev 62:275-293.

van der Woude, M. W., and A. J. Baumler. 2004. Phase and antigenic variation in bacteria. Clin Microbiol Rev 17:581-611, table of contents.

van Ham, S. M., L. van Alphen, F. R. Mooi, and J. P. van Putten. 1993. Phase variation of H. influenzae fimbriae: transcriptional control of two divergent genes through a variable combined promoter region. Cell 73:1187-1196.

van Ham, S. M., L. van Alphen, F. R. Mooi, and J. P. van Putten. 1994. The fimbrial gene cluster of Haemophilus influenzae type b. Mol Microbiol 13:673-684.

van Passel, M. W., and H. Ochman. 2007. Selection on the genic location of disruptive elements. Trends Genet 23:601-604.

van Ulsen, P., B. Adler, P. Fassler, M. Gilbert, M. van Schilfgaarde, P. van der Ley, L. van Alphen, and J. Tommassen. 2006. A novel phase-variable autotransporter serine protease, AusI, of Neisseria meningitidis. Microbes Infect 8:2088-2097.

Vassileva, V., A. Millar, L. Briollais, W. Chapman, and B. Bapat. 2002. Genes involved in DNA repair are mutational targets in endometrial cancers with microsatellite instability. Cancer Res 62:4095-4099.

Verkerk, A. J., M. Pieretti, J. S. Sutcliffe, Y. H. Fu, D. P. Kuhl, A. Pizzuti, O. Reiner, S. Richards, M. F. Victoria, F. P. Zhang, and et al. 1991. Identification of a gene (FMR-1) containing a CGG repeat coincident with a breakpoint cluster region exhibiting length variation in fragile X syndrome. Cell 65:905-914.

Vogler, A. J., C. Keys, Y. Nemoto, R. E. Colman, Z. Jay, and P. Keim. 2006. Effect of repeat copy number on variable-number tandem repeat mutations in Escherichia coli O157:H7. J Bacteriol 188:4253-4263.

Wang, G., Z. Ge, D. A. Rasko, and D. E. Taylor. 2000. Lewis antigens in Helicobacter pylori: biosynthesis and phase variation. Mol Microbiol 36:1187-1196.

Wang, J. D., M. B. Berkmen, and A. D. Grossman. 2007. Genome-wide coorientation of replication and transcription reduces adverse effects on replication in Bacillus subtilis. Proc Natl Acad Sci U S A 104:5608-5613.

Wang, X., D. Zhou, L. Qin, E. Dai, J. Zhang, Y. Han, Z. Guo, Y. Song, Z. Du, J. Wang, J. Wang, and R. Yang. 2006. Genomic comparison of Yersinia pestis and Yersinia pseudotuberculosis by combination of suppression subtractive hybridization and DNA microarray. Arch Microbiol 186:151-159.

Warren, M. J., and M. P. Jennings. 2003. Identification and characterization of pptA: a gene involved in the phase-variable expression of phosphorylcholine on pili of Neisseria meningitidis. Infect Immun 71:6892-6898.

Waterston, R. H.K. Lindblad-TohE. BirneyJ. RogersJ. F. AbrilP. AgarwalR. AgarwalaR. AinscoughM. AlexanderssonP. AnS. E. AntonarakisJ. AttwoodR. BaertschJ. BaileyK. BarlowS. BeckE. BerryB. BirrenT. BloomP. BorkM. BotcherbyN. BrayM. R. BrentD. G. BrownS. D. BrownC. BultJ. BurtonJ. ButlerR. D. CampbellP. CarninciS. CawleyF. ChiaromonteA. T. ChinwallaD. M. ChurchM. ClampC. CleeF. S. CollinsL. L. CookR. R. CopleyA. CoulsonO. CouronneJ. CuffV. CurwenT. CuttsM. DalyR. DavidJ. DaviesK. D. DelehauntyJ. DeriE. T. DermitzakisC. DeweyN. J. DickensM. DiekhansS. DodgeI. DubchakD. M. DunnS. R. EddyL. ElnitskiR. D. EmesP. EswaraE. EyrasA. FelsenfeldG. A. FewellP. FlicekK. FoleyW. N. FrankelL. A. FultonR. S. FultonT. S. FureyD. GageR. A. GibbsG. GlusmanS. GnerreN. GoldmanL. GoodstadtD. GrafhamT. A. GravesE. D. GreenS. GregoryR. GuigoM. GuyerR. C. HardisonD. HausslerY. HayashizakiL. W. HillierA. HinrichsW. HlavinaT. HolzerF. HsuA. HuaT. HubbardA. HuntI. JacksonD. B. JaffeL. S. JohnsonM. JonesT. A. JonesA. JoyM. KamalE. K. KarlssonD. KarolchikA. KasprzykJ. KawaiE. KeiblerC. KellsW. J.

KentA. KirbyD. L. KolbeI. KorfR. S. KucherlapatiE. J. KulbokasD. KulpT. LandersJ. P. LegerS. LeonardI. LetunicR. LevineJ. LiM. LiC. LloydS. LucasB. MaD. R. MaglottE. R. MardisL. MatthewsE. MauceliJ. H. MayerM. McCarthyW. R. McCombieS. McLarenK. McLayJ. D. McPhersonJ. MeldrimB. MeredithJ. P. MesirovW. MillerT. L. MinerE. MonginK. T. MontgomeryM. MorganR. MottJ. C. MullikinD. M. MuznyW. E. NashJ. O. NelsonM. N. NhanR. NicolZ. NingC. NusbaumM. J. O'ConnorY. OkazakiK. OliverE. Overton-LartyL. PachterG. ParraK. H. PepinJ. PetersonP. PevznerR. PlumbC. S. PohlA. PoliakovT. C. PonceC. P. PontingS. PotterM. QuailA. ReymondB. A. RoeK. M. RoskinE. M. RubinA. G. RustR. SantosV. SapojnikovB. SchultzJ. SchultzM. S. SchwartzS. SchwartzC. ScottS. SeamanS. SearleT. SharpeA. SheridanR. ShownkeenS. SimsJ. B. SingerG. SlaterA. SmitD. R. SmithB. SpencerA. StabenauN. Stange-ThomannC. SugnetM. SuyamaG. TeslerJ. ThompsonD. TorrentsE. TrevaskisJ. TrompC. UclaA. Ureta-VidalJ. P. VinsonA. C. Von NiederhausernC. M. WadeM. WallR. J. WeberR. B. WeissM. C. WendlA. P. WestK. WetterstrandR. WheelerS. WhelanJ. WierzbowskiD. WilleyS. WilliamsR. K. WilsonE. WinterK. C. WorleyD. WymanS. YangS. P. YangE. M. ZdobnovM. C. Zody, and E. S. Lander. 2002. Initial sequencing and comparative analysis of the mouse genome. Nature 420:520-562.

Watson, M. E., Jr., J. L. Burns, and A. L. Smith. 2004. Hypermutable Haemophilus influenzae with mutations in mutS are found in cystic fibrosis sputum. Microbiology 150:2947-2958.

Weber, J. L., and P. E. May. 1989. Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. Am J Hum Genet 44:388-396.

Weber, J. L., and C. Wong. 1993. Mutation of human short tandem repeats. Hum Mol Genet 2:1123-1128.

Webster, M. T., N. G. Smith, and H. Ellegren. 2002. Microsatellite evolution inferred from human-chimpanzee genomic sequence alignments. Proc Natl Acad Sci U S A 99:8748-8753.

Weiser, J. N., J. M. Love, and E. R. Moxon. 1989. The molecular mechanism of phase variation of H. influenzae lipopolysaccharide. Cell 59:657-665.

Weiser, J. N., D. J. Maskell, P. D. Butler, A. A. Lindberg, and E. R. Moxon. 1990. Characterization of repetitive sequences controlling phase variation of Haemophilus influenzae lipopolysaccharide. J Bacteriol 172:3304-3309.

Wexler, Y., Z. Yakhini, Y. Kashi, and D. Geiger. 2005. Finding approximate tandem repeats in genomic sequences. J Comput Biol 12:928-942.

Willems, R., A. Paul, H. G. van der Heide, A. R. ter Avest, and F. R. Mooi. 1990. Fimbrial phase variation in Bordetella pertussis: a novel mechanism for transcriptional regulation. Embo J 9:2803-2809.

Wren, B. W. 2003. The yersiniae--a model genus to study the rapid evolution of bacterial pathogens. Nat Rev Microbiol 1:55-64.

Wren, J. D., E. Forgacs, J. W. Fondon, 3rd, A. Pertsemlidis, S. Y. Cheng, T. Gallardo, R. S. Williams, R. V. Shohet, J. D. Minna, and H. R. Garner. 2000. Repeat polymorphisms within gene regions: phenotypic and evolutionary implications. Am J Hum Genet 67:345-356.

Wright, S. 1931. Evolution in Mendelian Populations. Genetics 16:97-159.

Wu, Y., J. H. McQuiston, A. Cox, T. D. Pack, and T. J. Inzana. 2000. Molecular cloning and mutagenesis of a DNA locus involved in lipooligosaccharide biosynthesis in Haemophilus somnus. Infect Immun 68:310-319.

Xu, X., M. Peng, and Z. Fang. 2000. The direction of microsatellite mutations is dependent upon allele length. Nat Genet 24:396-399.

Yamada, N. A., G. A. Smith, A. Castro, C. N. Roques, J. C. Boyer, and R. A. Farber. 2002. Relative rates of insertion and deletion mutations in dinucleotide repeats of various lengths in mismatch repair proficient mouse and mismatch repair deficient human cells. Mutat Res 499:213-225.

Yang, J., J. Wang, L. Chen, J. Yu, J. Dong, Z. J. Yao, Y. Shen, Q. Jin, and R. Chen. 2003. Identification and characterization of simple sequence repeats in the genomes of Shigella species. Gene 322:85-92.

Yang, Q. L., and E. C. Gotschlich. 1996. Variation of gonococcal lipooligosaccharide structure is due to alterations in poly-G tracts in lgt genes encoding glycosyl transferases. J Exp Med 183:323-327.

Yang, W. 2000. Structure and function of mismatch repair proteins. Mutat Res 460:245-256.

Yeung, G., L. M. Choi, L. C. Chao, N. J. Park, D. Liu, A. Jamil, and H. G. Martinson. 1998. Poly(A)-driven and poly(A)-assisted termination: two different modes of poly(A)-dependent transcription termination. Mol Cell Biol 18:276-289.

Yogev, D., R. Rosengarten, R. Watson-McKown, and K. S. Wise. 1991. Molecular basis of Mycoplasma surface antigenic variation: a novel set of divergent genes undergo spontaneous mutation of periodic coding regions and 5' regulatory sequences. EMBO J 10:4069-4079.

Youings, S. A., A. Murray, N. Dennis, S. Ennis, C. Lewis, N. McKechnie, M. Pound, A. Sharrock, and P. Jacobs. 2000. FRAXA and FRAXE: the results of a five year survey. J Med Genet 37:415-421.

Zamorzaeva, I., E. Rashkovetsky, E. Nevo, and A. Korol. 2005. Sequence polymorphism of candidate behavioural genes in Drosophila melanogaster flies from 'Evolution canyon'. Mol Ecol 14:3235-3245.

Zane, L., L. Bargelloni, and T. Patarnello. 2002. Strategies for microsatellite isolation: a review. Mol Ecol 11:1-16.