Tandem Duplications in the Human Genome

Philipp W. Messer

Januar 2008

Dissertation zur Erlangung des Grades
eines Doktors der Naturwissenschaften (Dr. rer. nat.)
am Fachbereich Mathematik und Informatik
der Freien Universität Berlin

Gutachter:
Prof. Dr. Martin Vingron
Prof. Dr. Nikolaus Rajewsky

1. Referent: Prof. Dr. Martin Vingron
2. Referent: Prof. Dr. Nikolaus Rajewsky

Tag der Promotion: 20. März 2008

Preface

Acknowledgments This work was carried out at the *Evolutionary Genomics* group of the Department of Computational Molecular Biology at the Max Planck Institute for Molecular Genetics in Berlin. I thank all past and present colleagues for the good working atmosphere and inspiring discussions.

Especially, I am grateful to my supervisor *Peter Arndt* for suggesting the topic, his scientific support, and the opportunity to write this thesis under his guidance. I thank *Ralf Bundschuh* and *Martin Vingron*, who – with their profound expertise in sequence alignment statistics – substantially contributed to the analysis conducted in Chapter 4. Special thanks go to *Michael Lässig* for the fruitful discussions and ideas that contributed to Chapter 3 and *Nicole de la Chaux*, who carried out parts of the programming in the section on indels in protein-coding regions. I also thank the Kavli Institute for Theoretical Physics for hospitality and the International Max Planck Research School for Computational Biology and Scientific Computing. Last but not least, great thanks go to my family and friends. Without their continuous support this work would not have been possible.

Publications Most parts of this thesis subsume the content of seven publications; six of which were conducted during my PhD studies. The first part of Chapter 2, which describes the analysis of indels in the human lineage, appeared in *Molecular Biology and Evolution* [110]. The second part concerning indels in protein-coding regions was published in *BMC Evolutionary Biology* [46]. Chapter 3 was motivated by an analysis I carried out during my diploma work at the Institute for Theoretical Physics at Cologne University [106]. The original work was published in *Physical Review Letters* [111]. Some of its results are recalled in the beginning of Chapter 3 because they form the basis for the further analysis, which appeared in *Journal of Statistical Mechanics: Theory and Experiment* [114]. The web server CorGen was described in *Nucleic Acids Research* [109]. Chapter 4 resulted from a conference paper at the *10th Annual International Conference, RECOMB 2006* [112]. An extended version was later published in *Journal of Computational Biology* [113].

Philipp Messer Berlin, January 2008

Contents

Preface **i**

1 Introduction **1**
1.1 Molecular biology and evolution 1
1.2 Mutational processes 7
1.3 Inferring mutation characteristics 10
1.4 Background models of DNA sequences 18
1.5 Thesis organization 19

2 DNA insertions and deletions in the human lineage **21**
2.1 Identification of indels in the human lineage 21
2.2 Tandem duplications and molecular mechanisms 25
2.3 Indels in protein-coding regions 35

3 Tandem duplications and genomic correlations **43**
3.1 Dynamical model of sequence evolution 44
3.2 Sequence growth and average composition 46
3.3 Stationary two-point correlations 49
3.4 Finite-size distribution of the composition bias 55
3.5 Symmetry breaking and universality 58
3.6 Dynamical correlations 62
3.7 General four-letter model and web service CorGen 66
3.8 Origin of genomic correlations 71

4 Genomic correlations and sequence alignment statistics **75**
4.1 Sequence alignment and significance assessment 75
4.2 The Gaussian approximation 77
4.3 Numerical results 80
4.4 Consequences for genomic alignments 84

5 Summary **87**

Bibliography **89**

Notation and abbreviations **103**

Zusammenfassung **105**