University of Warsaw
College of Inter-Faculty Individual Studies in Mathematics
and Natural Sciences

Michał Startek

Modelling the evolution of mobile genetic elements

PhD dissertation

Supervisors

dr hab. Anna Gambin, prof. dr hab. Dariusz Grzebelus

University of Warsaw

October 2014

Author's declaration:
aware of legal responsibility I hereby declare that I have written this dissertation myself and all the contents of the dissertation have been obtained by legal means.

October 8, 2014 .....

date *Michał Startek*

Supervisors' declarations:
the dissertation is ready to be reviewed

October 8, 2014 .....

date *dr hab. Anna Gambin, prof. dr hab. Dariusz Grzebelus*

Thesis supervisors: dr hab. Anna Gambin
prof. dr hab. Dariusz Grzebelus Michał Startek

Modelling the evolution of mobile genetic elements

Abstract

This thesis studies the long-standing question about the role of transposable elements in host genomes. We put forward and study a hypothesis that transposable elements may act as evolutionary helpers, hastening the adaptation of the host organism to a new environment, while remaining inactive during periods of environmental stability.

This hypothesis is analyzed by a modelling approach: we construct and study a computational model of transposable element proliferation, in conditions of environmental stress, then we show that environmental stress causes a burst of transposable element activity, increasing mutability of the organisms, and assisting adaptation.

Next, we analyze a mathematical model of transposable element proliferation, deriving closed-form formulas for selection-mutation equilibrium in organisms with certain kinds of transposable elements.

The study is augmented by empirical experiments, tying the models to real-world phenomena: by analyzing the role of LINE elements in mediating Nonallelic Homologous Recombination we prove that even inactive transposable elements may exert a significant mutagenic effect on the genome.

Finally, we conclude by presenting two tools for mining transposable elements in genomic sequences which we have developed over the course of experiments: TRANScendence, and TIRfinder.

Keywords: Transposable elements, Fisher's geometric model, moving phenotypic optimum, gaussian mutations, Hidden Markov Model, mutation-selection equilibrium
ACM Classification: J.3

Promotorzy rozprawy: dr hab. Anna Gambin
prof. dr hab. Dariusz Grzebelus Michał Startek

Modelowanie ewolucji ruchomych elementów genetycznych

Streszczenie

W ramach tej rozprawy zostały przeprowadzone badania nad znanym pytaniem dotyczącym roli elementów transpozonowych w genomach organizmów. Przedkładamy tutaj, i analizujemy hipotezę mówiącą o tym, że elementy transpozonowe mogą działać jako czynnik wspomagający ewolucję organizmu gospodarza, pozwalając mu szybciej dostosować się do nowych warunków środowiskowych, pozostając jednocześnie nieaktywnymi w okresach stabilności środowiska.

Hipoteza ta jest analizowana poprzez stworzenie modeli aktywności elementów transpozonowych: w ramach tej pracy zaproponowany został stochastyczny, obliczeniowy model proliferacji elementów transpozonowych w warunkach stresu środowiskowego. Następnie przy jego pomocy pokazane zostało, że stres środowiskowy powoduje wybuch aktywności elementów transpozonowych, zwiększając szybkość mutacji organizmów, i wspomagając je w adaptacji.

Dodatkowo, sformułowany i przeanalizowany został matematyczny model proliferacji elementów transpozonowych, oraz wyprowadzone zostały zamknięte formuły opisujące stan równowagi pomiędzy mutacjami i selekcją w organizmach zawierających pewne typy elementów transpozonowych.

Badania te zostały poparte wynikami eksperymentalnymi, które wiążą nasze modele z rzeczywistymi zjawiskami: analizujemy rolę elementów LINE w mediowaniu nieallelicznej homologicznej rekombinacji, pokazując że nawet nieaktywne elementy transpozonowe mogą wywierać znaczący mutagenny wpływ na organizm gospodarza.

Wreszcie na końcu przedstawione są dwa narzędzia, służące do wykrywania elementów transpozonowych w zsekwencjonowanych genomach: TRANScendence oraz TIRfinder.

Słowa kluczowe: Elementy transpozonowe, model geometryczny Fischer'a, zmienne optimum fenotypowe, mutacje gaussowskie, Ukryty Model Markova, równowaga między selekcją a mutacjami
Klasyfikacja tematyczna ACM: J.3

Contents

1 Introduction **1**

1.1 Mobile genetic elements 1

1.2 Characteristics of transposable elements, classification . . . 2

1.2.1 Autonomous versus nonautonomous 2

1.2.2 Active versus inactive 2

1.2.3 Class-I versus class-II 3

1.2.4 Abundance, effects on the host genome 4

1.3 Motivation of this work 4

1.3.1 The computational model 5

1.3.2 Mathematical modelling 6

1.3.3 Cell suspension culture of *Medicago truncatula* . . . 7

1.3.4 Transposon detection tools: TRANScendence and TIRfinder 8

1.3.5 A study of mutations caused by inactive LINE elements in human genome 9

1.4 Collaboration 9

1.5 Papers, manuscripts and software tools 11

2 Transposons in asexual populations – a computational model **13**

2.1 Method and results 16

2.1.1 Phenotype and natural selection 17

2.1.2 Transposition 18

2.1.3 Mutations 18

2.1.4 Constant environmental pressure ('Global Warming' (GW) scenario) . . . 19
2.1.5 Rapid environmental change ('Meteor Impact' (MI) scenario) . . . 21
2.1.6 General properties of the model . . . 23
2.2 Discussion . . . 24
2.2.1 Model . . . 24
2.2.2 Impact of TEs on genome evolution . . . 26
2.2.3 Conclusions . . . 27

3 Mathematical modelling of TE-carrying populations in the presence of environmental stress **31**
3.1 Previous work . . . 32
3.2 Notation and basic definitions . . . 32
3.3 Transposition model for class I TEs . . . 34
3.3.1 Introduction – general setting . . . 34
3.3.2 Mathematical formulation . . . 36
3.4 Mutator model . . . 40
3.4.1 Introduction . . . 40
3.4.2 Simplification of model for class-I TEs . . . 40
3.4.3 Stationary state . . . 41
3.4.4 Proof of convergence . . . 42
3.4.5 Analysis of the equilibrium state . . . 56
3.4.6 A model for class II transposons . . . 60
3.4.7 Concluding remarks and interpretation . . . 68

4 Analysis of TEs in human genome as contributors to genomic disease through nonallelic homologous recombination **69**
4.1 Motivations and related research . . . 70
4.2 Materials and methods . . . 71
4.2.1 Identification of LINE pairs able to mediate NAHR 71
4.2.2 Clinical CMA . . . 73
4.2.3 Subjects . . . 74
4.2.4 Long Range PCR and DNA sequencing . . . 74

4.2.5 LR-PCR analysis of healthy subjects 74
4.2.6 Array CGH analysis of healthy subjects 75
4.2.7 DNA sequence analysis 76
4.2.8 Preparation of sequences for analysis 76
4.2.9 Hidden Markov model for breakpoints identification 77
4.2.10 Enrichment of CMA instability regions with LINEs 82
4.3 Results 84
4.3.1 Regions potentially prone to LINE-LINE-mediated NAHR 84
4.3.2 Analyses of NAHR breakpoints observed in individuals 84
4.3.3 Enrichment of predicted LINE-LINE/NAHR pairs in CNV uncertainty regions 86
4.3.4 NAHR hotspots in flanking LINE elements 86
4.3.5 aCGH Analysis of Healthy Individuals 88
4.4 Discussion 89

5 TIRfinder and TRANScendence: transposable element detection tools **91**
5.1 TIRfinder 91
5.1.1 Methods of TE detection and analysis 93
5.1.2 Usage 98
5.1.3 TIRfinder implementation 99
5.1.4 Case Studies 99
5.2 TRANScendece 101
5.2.1 Functionality of TRANScendence 103
5.2.2 The TE Landscape of the *Medicago truncatula* . . . 105
5.2.3 TEs interruption graph 108
5.2.4 Automatic TE mining 110
5.2.5 TE repository 111
5.2.6 Technologies used 111
5.2.7 Discussion 111
5.3 Conclusions 112

6 Conclusions and further research **113**
6.1 Spatial extension to computational model 113
6.2 Model for sexually-reproducing organisms 114
6.3 Mathematical modelling of class-I TE and sexually-reproducing organisms 114
6.4 Further development of TE-detection tools 115

Glossary **116**

Glossary **117**

References **118**

List of Figures

1.3.1 Overview of the plan of the project described in this thesis. Arrows represent information flow. 10

2.1.1 General outline of the model. (A): Life cycle of the population: (1) simulation starts with a population of 10,000 individuals; (2) both autonomous and non-autonomous transposable elements are mobilized by the transposition machinery produced by autonomous TEs; (3) each TE inserted causes a mutation of the host phenotype; (4) non transposition-related mutations also modify the phenotype (5) better adapted individuals (i.e. closer to the *optimal phenotype*) have greater probability of survival; (6) surviving individuals reproduce to fill the environment back to its capacity. (B): Evolution of transposable elements: an autonomous element can duplicate during the process of transposition or become non-autonomous; a non-autonomous element can still proliferate using the transposition machinery produced by other, autonomous elements. 16

2.1.2 (A) Number of runs (out of 100) in which autonomous TEs are eliminated by generation 5000 under the "Global Warming" scenario (no non-autonomous copies). In low stress levels all TEs are lost, and the phenotypic optimum is tracked purely through TE-unrelated mutations. With higher levels of environmental change, TEs are maintained more often, and assist in tracking the phenotypic optimum by providing extra mutations. (B) Distribution of the average number of autonomous TE copies at generation 5000, from 100 simulation runs at each change level. The TE copy number at generation 5000 increases linearly with the rate of environmental change above the minimal threshold allowing for TE persistence. 20

2.1.3 Effect of smooth environmental change ('Global Warming' scenario). (A) Reference simulation without TEs: transposition-unrelated mutations are enough to track the optimum in the long term. The distance of the population from the optimum is larger than with TEs (the average fitness is lower). The color scale is proportional to the density distribution of fitnesses in the population. (B) Non-autonomous copies disabled. TEs assist in tracking of the phenotypic optimum. (C) Non-autonomous copies enabled. 22

2.1.4 Effect of periodic dramatic environmental shifts ('Meteor Impact' scenario) (A) Without TEs the population diverges from the phenotypic optimum (B) With autonomous copies the population is able to track more closely the phenotypic optimum (C) With both autonomous and non-autonomous copies, the population follows the phenotypic optimum, until TEs are eliminated. 23

3.4.1 Construction of the dominating function f, as in proof of Lemma 6. The dominating function is in black, the Gaussian PDF with lowest variance is red, the one with highest variance is orange, the one with lowest mean is blue, and the one with highest is black; (some) other PDFs from the class have also been plotted, in gray. 52

3.4.2 Plots of survivability function with various parameters fixed. **(A)** Survivability as a function of random mutation rate. A clear optimal random mutation rate is seen. **(B)** Reducing the strength of selection (by increasing the selection radius σ) causes a decrease in survivability, however, the optimal mutation rate does not appreciably change **(C)** Increasing the speed of environmental change reduces survivability, and increases the optimal mutation rate. **(D)** The effects of both lower selection and faster environmental change **(E)** Survivability as a function of environmental change. Highest survivability occurs in a stable environment. **(F)** Survivability as a function of selection radius. Highest survivability occurs with most stringent selection. 58

3.4.3 Plot of survivability function with $\eta = 1$ and varying mutation rates and selection radii. 59

3.4.4 Plots of average fitness function with various parameters fixed. **(A)** Plot of average fitness as a function of random mutation rate. There is an evident optimal mutation rate, however, increasing mutation rate beyond it causes only a slight drop in average fitness. **(B)** Same plot, with less stringent selection (higher selection radius σ). Decreasing the strength of selection doesn't appreciably relocate the optimal mutation rate, however, it makes adaptation more difficult. **(C)** Increasing the speed of environment change increases the mutation rate needed to keep up with the environment, however, the average fitness does not significantly decrease. **(D)** Combined effects of high environment change and low selection. **(E)** Average fitness as a function of environment change, with fixed mutation rate: if we allow the speed of environmental change to increase without allowing the mutation rate to increase as well, the average fitness of population tends to zero. **(F)** Average fitness as a function of selection radius. With more stringent selection we can arbitrarily increase the population's fitness. 61

3.4.5 Plot of average fitness of the population as a function of mutation rate and selection radius, with $\eta = 1$. 62

3.4.6 Plot of an example survivability function with notations introduced in Remark 1 66

4.2.1 The schematic representation used to define the *uncertain regions* for breakpoint analyses. The proximal NAHR breakpoint maps within the left red area, the distal one in the right red area. A similar approach was used for duplications. 72

4.2.2 Scatterplot of patients with CNVs, where a pair of matching LINEs lies in *uncertain regions* for the ends of the CNV. Only cases where the alignment between LINEs is longer than 4 bp with over 96% identity are shown. Patients are sorted by the length of their shortest CNV; one patient may possess multiple CNVs. Cases selected for PCR confirmation are highlighted with arrows. 73

4.2.3 Artificial sequences computed for primer design for detection of chimeric LINE sequences. Shown on figure is the process for deletion, with duplications and inversions being handled in similar fashion. 75

4.2.4 Construction of input sequence for estimation of NAHR breakpoint location. In artificial sequence, the S and E are special markers, for beginning and end of the sequence, L means that the observed sequence seems to come from the left (first) LINE, R means it comes from the right (second) one, N means that the source LINE cannot be determined from this location. 76

4.2.5 Hidden Markov model used for estimation of breakpoint location. The NAHR site maps at the point of $S_1 \rightarrow S_2$ transition. The prior and posterior values of $\alpha, \beta, \gamma, \rho$ can be found in Table 4.2.1. 81

4.2.6 A plot of observed/expected ratio of matching LINE pairs lying within CNV breakpoint regions. For each point on X and Y axes the observed/expected ratio of LINE pairs with parameters equal or better is shown. It is evident that identity percentage of 96 or better is needed to mediate NAHR, while there is no sharp restriction on minimal length of the homology. 83

4.3.1 Ideogram showing the susceptibility of human genome to LINE-LINE-mediated NAHR. Each horizontal red line corresponds to one potentially NAHR-mediating LINE pair: the LINE elements map at the ends of the line, whereas the segment covers the potentially deleted or duplicated regions. For clarity of the figure, inversions and translocations are not shown. 85
4.3.2 Estimated NAHR breakpoint location probabilities from the Hidden Markov Model for duplications between LINEs on chromosome 20. Three distinct NAHR loci were identified among the tested patients. 87
4.3.3 Molecular validation of predicted LINE-LINE CNVs identified among healthy individuals by aCGH. A) Array CGH data indicates a CNV at 2q34 in subject 1 or 2. B) Schematic representation of the L1PA elements that mediate the CNV and LR-PCR primers testing for the CNV. C) LR-PCR identifies the presence of a deletion in subject 1. D) Array CGH data indicates a CNV at 8p23.3 in subject 1 or 2 and subject 5 or 6. E) Schematic representation of the L1PA elements that mediate the CNVs and LR-PCR primers testing for the CNVs. F) LR-PCR identifies the presence of homozygous duplications in subjects 1 and 5. 88
5.1.1 The control flow through different phases of the TIRfinder TEs detection method. 94
5.1.2 TIRfinder algorithm. The suffix trees are overlapping to ensure that sequences on the boundaries between two suffix trees are detected correctly. 96
5.1.3 TIRfinder – structural analysis. **(A)** Explanation of TIR and TSD mask concept. **(B)** Overview of TEs detection phase 97
5.1.4 Pogo-like TEs landscape detected by TIRfinder in *A. thaliana.* 100
5.1.5 The example of TIRfinder outcomes: search for *PIF/Harbinger* TEs family in chromosome 5 of *M. truncatula* genome. . . 101
5.2.1 Overview of the TRANScendence tool. 104

5.2.2 All found putative TE elements are classified into classes, orders and superfamilies. 106
5.2.3 Each TE family is annotated by BLASTing the consensus sequence against Repbase content. 107
5.2.4 Nesting structure for TE families detected in *Medicago truncatula*. 109

List of Tables

2.1.1 Parameters values used in simulations presented in: Fig. 3B (Global-warming without non-autonomous TEs), Fig. 3C (Global-warming with non-autonomous TEs), Fig. 3A (Global warming without TEs), Fig. 4A (Meteor impact without TEs, Fig. 4B (Meteor impact without non-autonomous TEs), and Fig. 4C (Meteor impact with non-autonomous TEs). 19

4.2.1 Table of HMM parameters used. Parameter names from Figure 4.2.5 81

5.1.1 Pogo-like TEs found by TIRfinder vs. annotated in Repbase. 100

Rodzicom

Acknowledgements

W pierwszej kolejności chciałbym podziękować moim promotorom: dr hab. Annie Gambin oraz prof. dr. hab. Dariuszowi Grzebelusowi za ich pomoc (naukową i nie tylko) i wsparcie.

Podziękowania kieruję również do współpracowników i współautorów, w szczególności do: Arnaud'a Le Rouzic PhD, dr. Tomasza Gambin, dr. hab. Pawła Stankiewicza, mgr. Krzysztofa Gogolewskiego orac lic. Mateusza Kitlasa.

Na koniec chciałbym podziękować swoim kolegom ze studiów, a szczególnie „mieszkańcom" pokoju 5810. Bez was te studia byłyby dużo mniej interesujące.

Dziękuję!

First, I would like to thank my supervisors: dr hab. Anna Gambin and prof. dr hab. Dariusz Grzebelus for their support (scientific and otherwise) and encouragement.

I would like to thank my collaborators and co-authors, especially: Arnaud Le Rouzic, PhD, dr Tomasz Gambin, dr hab. Paweł Stankiewicz, mgr Krzystof Gogolewski and lic. Mateusz Kitlas, for all the scientific input and discussions.

Finally, I would like to thank my colleagues from the studies, especially the denizens of the 5810 room. Without you the studies would have been much less interesting.

Thank you!

1 Introduction

1.1 Mobile genetic elements

Mobile genetic elements (variously called: transposable elements, TEs) are DNA sequences capable of relocating within the genome, in a process called *transposition*. Since their discovery (McClintock, 1950) transposable elements have been the focus of much scientific study. Their function (or indeed, whether they have any) is still the subject of dispute, as is their status: variously cited as junk DNA (Doolittle and Sapienza, 1980), parasitic DNA (Kidwell and Lisch, 2001), or as being beneficial to the host organism (Kofler et al., 2012). The justification for these classification varies, and it is becoming more and more evident that TEs can fulfill all of these roles, depending on various factors, among them, conditions in which the host organism finds itself.

This work aims to study the dynamics of interaction of TEs with the genome, and to contribute to solving the long-standing question: what functions TEs may have, and how do they affect the host organism?

1.2 Characteristics of transposable elements, classification

1.2.1 Autonomous versus nonautonomous

There are some common structural elements that most TEs possess: most have characteristic sequences at both ends, demarcating the ends of the TEs (be it a Terminal Inverted Repeat (TIR) or a Long Terminal Repeat). Between those sequences, a number of Open Read Frames (ORFs) may sometimes be found, encoding genes: this is a somewhat surprising fact, but the proteins necessary for transposition are often encoded within the TEs themselves. Depending on the mechanism of transposition, these may be either be a transposase, or various integrases, reverse transcriptases and so on. In addition to that, as some of the TEs are remnants of viruses (particularly, retroviruses) which have lost their ability to leave the host cell due to a mutation, some TEs still carry genes encoding virus-associated protein.

This is the first criterion for classification of TEs: the *autonomous* TEs are those that carry within themselves a full set of genes encoding all the proteins necessary for the process of transposition, and which are fully capable of transposing inside the genome on their own, and the *nonautonomous* TEs, which do not have a full set of proteins necessary for transposition. The nonautonomous TEs usually arise as a result of a mutation or deletion occurring within an autonomous TE which renders one or more of the genes non-functional (or completely absent). The non-autonomous TEs may still proliferate, but they may only do so in the presence of autonomous TEs which enable the production of the necessary proteins. These proteins are in turn hijacked by the nonautonomous TEs, and used for their own transposition, in an ironic case of parasitic DNA (the autonomous TEs) being parasitized upon by other parasitic DNA (the nonautonomous TEs).

1.2.2 Active versus inactive

Regarding mutations, of course it is not only the genes within the TEs that may suffer from mutation, but the delimiters too. For example:

damage through single-point mutations to a TE's TIRs may cause them to stop being recognizable by transposase, and thus, inhibit the TE from ever relocating. Such TEs, which have lost their ability to transpose are variously called non-active TEs or TE relics. In fact, going strictly by the definition from the first sentence of this work, these are not TEs at all. However, most TEs are being detected in a bioinformatic fashion, the similarity to the known TEs being a criterion for annotation as a TE. It is notoriously difficult to determine whether a given sequence is in fact capable of transposition within a live cell – as such, the capability of relocation for most TE-like sequences in not actually known, and they are still called TEs, a convention which we will follow here.

1.2.3 Class-I versus class-II

Finally, TEs may be divided based on the mechanism of their transposition. The class-I TEs, or retrotransposons (named as such because of their similarity to retroviruses) proliferate using an RNA intermediate: the TE in the DNA is transcribed onto a RNA strand (similarly to coding genes), next, the RNA enters the cytoplasm where the necessary proteins are synthesized based on the RNA template (in a process of translation) then, the proteins perform reverse transcription of the RNA strand, synthesizing a DNA strand, which in turn, is integrated into the host cell genome at a new position. This is known as the *copy-and-paste mechanism*, as the original copy of the TE remains undisturbed, and a new copy appears at a different site. This is similar to the lifecycle of retroviruses, except that in case of the viruses, after the synthesis of necessary proteins, the RNA (along with reverse transcrpitase, and integrase) is enveloped in a capsid, and expelled from the cell. After it infects another cell the cycle resumes. In fact, some retrotransposons are former retroviruses, which have lost the proteins necessary for construction of the capsid, and thus, the ability to escape the host cells and insert others, but which still proliferate inside the cell like a retrovirus does. One example is the Human Endogenous Retrovirus (HERV), which makes up 8% (Belshaw et al., 2004) of the human genome. For us, the most important feature of this mode of transposition is the fact, that with each transposition results in a new TE appearing.

In contrast to the retrotransposons, the class-II TE, or the DNA transposons, operate directly on the DNA strand. An enzyme called transposase cuts the transposon out of the genome, and reinserts it into a new locus. Then the gap resulting from cutting away the transposon is repaired by standard cellular mechanisms involved in repairing double-strand breaks. Sometimes the repair mechanisms recreate the transposon at the site of excision, resulting in an increase of copy number of the transposon, sometimes they recreate the transposon only partially (most often recreating the ends, but skipping a large portion of the middle of the transposon, which usually creates a nonautonomous transposon, a so-called Miniature Inverted-repeat Transposable Element (MITE)), most frequently they just mend the loose ends, resulting in a complete excision of the transposon. The effect is such that class-II transposons most frequently transpose in a *cut-and-paste* fashion, the transposition only rarely resulting in an increase of copy number of the transposon.

1.2.4 Abundance, effects on the host genome

The transposable elements make up a significant portions of the genomes of various organisms, being the major part of what once was known as "junk DNA". In humans, up to 50% of the genome is annotated as repetitive by RepeatMasker (Smit et al., 2004)

with families such as LINEs, SINEs, Alu, and HERVs being the most prominent.

Transposons have mutagenic effect on the host genome: one, widely known, is due to their ability to disrupt or misregulate genes by their insertions. A gene into which a TE inserts almost always looses its functionality. Furthermore, TEs often carry promoter sequences, which, when inserted into the vicinity of a gene, may cause its up- or downregulation.

1.3 Motivation of this work

The main motivation for the research presented here is our observation that TEs tend to proliferate in discrete "bursts" of activity, interspersed with longer periods of relative stability, such behaviour for example might

be observed in MITEs of *Medicago truncatula* (Grzebelus et al., 2009). We suspected such behaviour to be tied to sudden environmental change: the need to adapt to the new environment reduces the selective pressure against mutability, which, in turn, selects for organisms with active TE families. They use their TEs to adapt to the new environment, resulting in an activity burst. After the adaptation is complete, the selection against mutability resumes, which, again, selects organisms which manage to silence their TEs, resulting in the lack of activity observed between bursts. This has inspired this work: an in-depth study of the impact of transposable elements on the evolution of species, and the interplay between environmental stress a given population is subjected to, and the activity of its TEs, as well as the role of TEs in adaptation.

1.3.1 The computational model

In order to assist the preliminary study of the soundness of the idea, we have developed a computational, stochastic model of TE proliferation. The model is based upon the standard Fisher's Geometric Model (FGM) (Fisher, 1930). The organisms are modelled as having a phenotype described as a real-valued vector. The fitness function of an organism depends upon the distance between the organism's phenotype $\pi(o)$ and an environment-wide optimal phenotype $\hat{\pi}$:

$$F(o) = \exp(-\text{dist}(\boldsymbol{\pi}(o), \hat{\boldsymbol{\pi}})^2) = \exp(-\sum_{i=1}^{n}(\pi_i(o) - \hat{\pi}_i)^2)$$

The environmental change is modelled by moving the phenotypic optimum.

The TEs are included in this model: each organism has a counter of its autonomous and nonautonomous TEs. The model includes only class-I TEs, as these are the ones more abundant in nature. The mutagenic effect of TEs is modelled by having each transposition perturb the organism's phenotype (in a manner similar to random, non-TE-related mutations). The organisms are asexual, and reproduce clonally.

The results obtained suggest that TEs may react to environmental stress even in absence of epigenetic effects, purely through evolutionary pressure

variously favouring organisms with high or low TE activity. The high environmental pressure causes TEs to activate, and to assist the organisms in adapting to the new environment. In turn, stable environment results in TEs loosing their advantageous properties to the host organisms, and being able to persist only in a "parasitic DNA" dynamics, within certain parameter ranges.

1.3.2 Mathematical modelling

The stochasticity and random behaviour of the first model (in particular its propensity for producing qualitatively different results from the same initial conditions in subsequent runs for certain parameter ranges) have inspired us to attempt to construct a mathematical model. After first attempts with differential equation-based approaches, we have produced a model based on probability theory. The model is based on a so-called *generation operator* Φ – which is an operator transforming a probability distribution $\mathbb{P}$ describing a population into a probability distribution describing the population after one generation has passed $\Phi(\mathbb{P})$. The operator for class-I TE has so far proved mathematically intractable, however, we have managed to formulate the operator for class-II transposons, and found its equilibrium state, as well as proved the convergence for a large class of populations. The model for class-II transposons is an extension of a (simpler) mutator model (the theory of which makes up much of Chapter 3). The operator for mutator model has the following form:

$$\Phi(\mathbb{P})(A) = \frac{\int_{\mathbb{R}} \int_{\mathbb{R}} \nu(0,\sigma)(y+z) \cdot \mathbb{1}_{A+\eta}(y+z)\ \mathrm{d}\mathcal{N}(0,\rho)(y)\mathrm{d}\mathbb{P}(z)}{\int_{\mathbb{R}} \int_{\mathbb{R}} \nu(0,\sigma)(y+z)\ \mathrm{d}\mathcal{N}(0,\rho)(y)\mathrm{d}\mathbb{P}(z)}$$

where $\mathcal{N}$ is the Gaussian probability distribution, and ν is a Gaussian PDF (probability density function). We have proved that regardless of the starting probability distribution, the distribution of the population undergoing evolution under this operator strongly converges (in measure) to an equilibrium distribution which is Gaussian:

$$\lim_{n\to\infty} \Phi^n(\mathbb{P})(A) = \mathcal{N}\left(\frac{\eta\sqrt{4\,\sigma^2+\rho^2}-\eta\,\rho}{2\,\rho}, \frac{\sqrt{2\rho(\sqrt{4\sigma^2+\rho^2}-\rho)}}{2}\right)(A)$$

The proof proceeds by first showing that with two applications of Φ all initial populations are brought into a specific class of probability distributions which we named $\mathcal{CND}$:

$$\mathcal{CND} = \{\mathbb{P} : \mathrm{Leb}(\mathbb{R}) \to [0,1] \mid \exists \sigma \in \mathbb{R}\ \exists \mathbb{S} : \mathrm{Leb}(\mathbb{R}) \to [0,1]\ \ \mathrm{pdf}_{\mathbb{P}}(x) =$$

$$= \int_{\mathbb{R}} \nu(y,\sigma)(x)\mathrm{d}\mathbb{S}(y) \text{ and } \mathbb{S} \text{ has a probability density function}\}$$

Then we prove that the distribution $\mathbb{S}$ remains fixed by the action of Φ, only the parameters of the Gaussian part change. We prove convergence in variance using Banach's contraction mapping theorem, then convergence of mean. Finally, the $\lim_{n\to\infty} \Phi^n(A)$ is computed, and strong convergence in measure for any initial population is proven using theorems from the theory of Lebesgue integration.

Next, we present a model for class-II transposons as an extension of the mutator model. Basing on the results from the mutator model we derive a an equilibrium state, and prove weak-* convergence for populations whose distribution has a nonnegative PDF.

1.3.3 Cell suspension culture of *Medicago truncatula*

These results have inspired an attempt to recreate them in a biological setting. Our collaborators from University of Agriculture in Krakow have started a cell suspension culture of *Medicago truncatula* (which is still ongoing) with the aim of determining whether the stress associated with the cell culture has resulted in an increase of TE numbers. The cell culture will be harvested, the DNA extracted and sequenced using Next Generation Sequencing (NGS) techniques. Although the experiment itself is not a part of this thesis (both because it is still ongoing, and because it is purely biological in nature) it is worth mentioning here because it is inspired by the computational model presented here, and its results, in turn, will be used to fine-tune the models.

1.3.4 Transposon detection tools: TRANScendence and TIRfinder

In order to assist the studies of transposable elements we have prepared two bioinformatics tools: the TRANScendence and TIRfinder. TRANScendence is a de-novo TE mining and annotation tool, useful for initial analyses of the TE landscape in newly sequenced genomes, and we plan to use the tool to study the genomes after sequencing. We plan to use TRANScendence as a standardised tool allowing us to compare the TE landscapes of different organisms, using one standardised tool (instead of the public databases with varied levels of curation, which might not reflect the real differences between TE load in organisms).

The validity of this tool has been presented here on a test-case of the reference genome of *Medicago truncatula*, proving the tool's viability for the analysis of this genome. In addition to that, it has already been used in other scientific projects by our collaborators from University of Agriculture in Krakow, in the discovery of a novel transposon family targeting $(TA)_n$ microsatellites (manuscript in preparation).

We have also developed another tool for a deep analysis of TIR-carrying class II transposons, the TIRfinder. A complementary tool to the wide-and-shallow approach of TRANScendence, this tool will enable the in-depth analysis of the behaviour of specific transposon families. This tool has already been successfully used in scientific projects, like the study of PIF/Harbinger-like elements in *Medicago* Grzebelus et al. (2007) or analysis of MITE landscape also in *Medicago* Grzebelus et al. (2009). Here, we present the tool along with two case studies on $ATHPOGO$ and $PIF/HARBINGER$ elements in *Medicago truncatula*.

The results obtained from analysis of the TE activity under conditions of environmental stress using these tools will be used to verify the assumptions and calibrate the parameters of both our mathematical and computational models, especially the future extensions of these models.

1.3.5 A study of mutations caused by inactive LINE elements in human genome

Another venue of study of the impact of transposable elements on the evolution of species is our study in collaboration with Paweł Stankiewicz's team at Baylor College of Medicine (BCM) regarding Nonallelic Homologous Recombination (NAHR) between Long Interspersed Nuclear Elements (LINEs). LINEs are a family of TEs which are abundant in humans, and NAHR is by which they may cause Copy Number Variant (CNV)-type mutations in the genome. So far, it has been considered that NAHR is caused by self-similarity of the genome, however the usual homology length considered necessary for NAHR to occur was thought to be much bigger than 10 kbp, rendering TEs unable to mediate NAHR, despite them being the major source of self-similarity in the genome. Some TE-TE NAHRs have been previously found, however these were single cases, and considered more of a curiosity, and an exception, rather than a common occurrence. Our studies prove that TE-TE NAHR is common enough to be of clinical significance, and therefore, able to affect the course of evolution. This proves that even inactive (fixed) TEs have a mutagenic effect on the genome, something that must be considered in future modelling attempts. Over the course of this research several algorithms had to be developed, these are presented in Chapter 4.

This discovery has affected the mathematical model: it means that the model we have derived for class-II transposons is also applicable to inactive (senescent) families of TEs. This will also have an effect on any future modelling attempts.

1.4 Collaboration

As mentioned previously, this project has grown out of collaboration between Anna Gambin and Dariusz Grzebelus from University of Agriculture in Krakow (with their respective teams) studying the transposon landscape of *Medicago truncatula*, and the "bursty" behaviour of TEs which was discovered. We have collaborated with the team from Kraków on formulating the initial assumptions of the computational model described

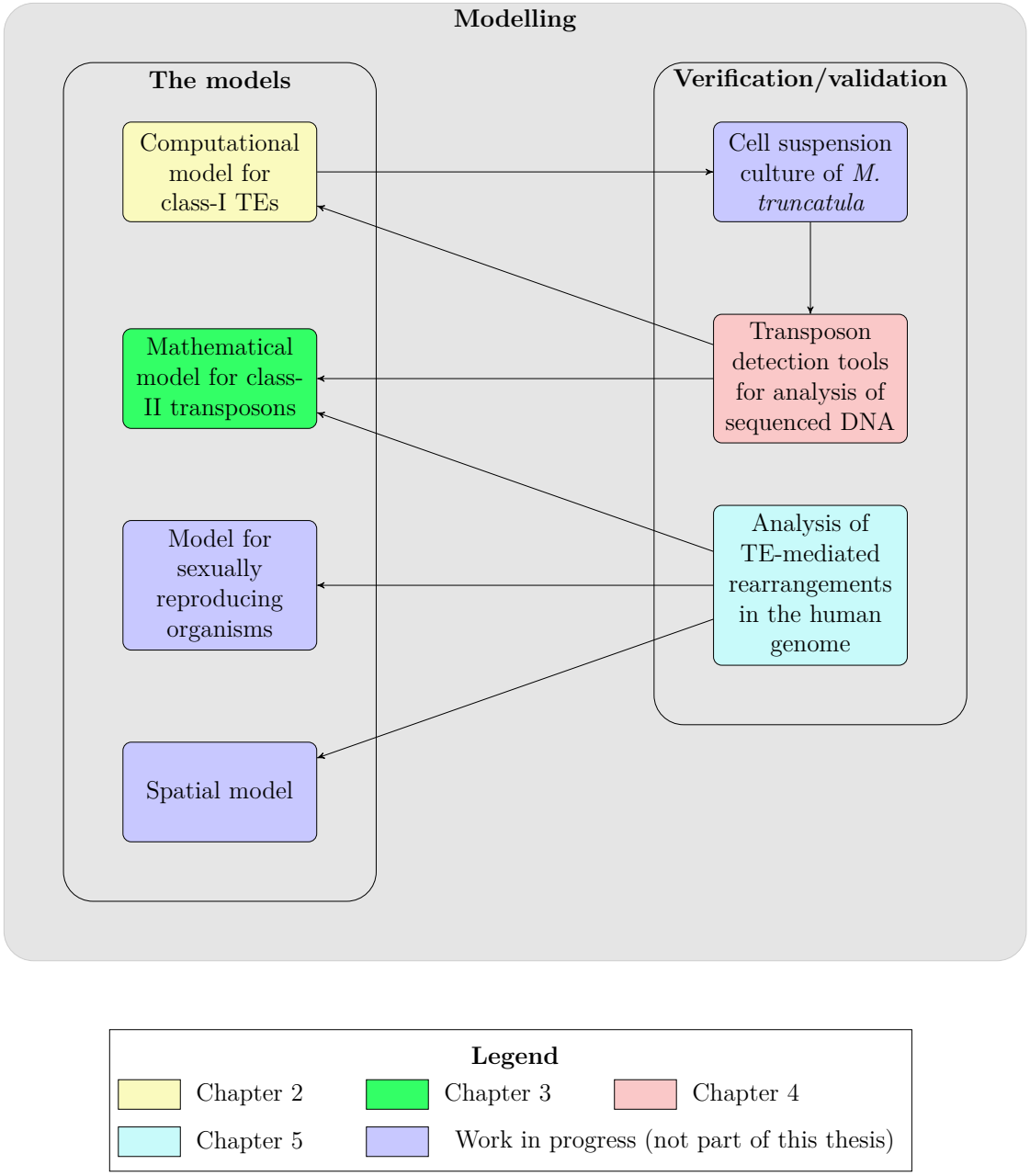


Figure 1.3.1: Overview of the plan of the project described in this thesis. Arrows represent information flow.

in Chapter 2. In addition to that, we have collaborated on the tools presented in Chapter 5, the team from Kraków providing the inspiration for writing them, as well as input on needed functionality, and testing them throughout the development process.

The modelling efforts (both Chapters 2 and 3) were also assisted by Arnaud le Rouzic from Laboratoire Evolution, Génomes et Spéciation (LEGS) at Centre National de la Recherche Scientifique (CNRS), Gif-sur-Yvette, France, in particular, verification of assumptions of the model, as well as fine-tuning the parameters, and providing interesting scenarios to be studied.

The work presented in Chapter 4 was carried out in collaboration with Paweł Stankiewicz's team at Baylor College of Medicine in Houston, USA, with them providing the clinical data, and performing all the laboratory experiments outlined there, while we have handled the bioinformatics part.

1.5 Papers, manuscripts and software tools

The work presented in Chapter 2 has been published in *Theoretical Population Biology*, as an article entitled "Genomic parasites or symbionts? Modeling the effects of environmental pressure on transposition activity in asexual populations".

Manuscript detailing the research from Chapter 3 is still in preparation.

A manuscript detailing the research from Chapter 4 has been sent to Nucleic Acids Research (entitled: "Genome-wide analysis of LINE-LINE-mediated nonallelic homologous recombination") is currently being reviewed for publication.

The TIRfinder tool presented in Chapter 5 has been described in a paper entitled "TIRfinder: A Web Tool for Mining Class II Transposons Carrying Terminal Inverted Repeats" published in Evolutionary Bioinformatics. The tool itself is available online at `http://bioputer.mimuw.edu.pl/tirfindertool/`, while its source code may be downloaded from `http://sourceforge.net/projects/tirfinder/`.

A paper detailing one of the use-cases of TRANScendence, a study of $(TA)_n$ microsatellite-targeting transposon family, entitled "MuTAnT: a family of Mutator-like transposable elements driving evolution of TA

microsatellites in Medicago truncatula" is in preparation. A manuscript describing the tool itself ("TRANScendence: a web tool for de-novo TE identification") is in preparation. The tool is publically available for use at `http://bioputer.mimuw.edu.pl/transcendence`.

2

Transposons in asexual populations – a computational model

In this chapter of the thesis we shall present a computational model of proliferation of TEs in a population of asexual, clonally reproducing organisms. We present here the first, preliminary work we have done to study the response of TEs to the environmental stress – the encouraging results which we obtained here were the motivation for much of the work presented in subsequent chapters. We shall begin with a short introduction, presenting to the reader the state of the art in the field, then we shall progress onto the results.

The evolution of species depends on both the strength of selection and the species' capacity to evolve. Small environmental changes tend to generate moderate stress on populations, which are likely to reach the new phenotypic optimum from standing genetic variation. On the contrary, large and fast shifts in the environment may generate substantial selection pressure, endangering the survival of the species, and adaptation may require the accumulation of several mutational changes (Barrett

and Schluter, 2008; Durand et al., 2010). In any case, the ability for the population to generate new variants through mutation remains a crucial feature that conditions its capacity to cope with environmental challenge. The mechanisms underlying the evolution of the capacity to evolve, or evolvability, are still not fully understood (Hansen, 2006; Partridge and Barton, 2000; Pigliucci, 2008). Both theory and empirical observations suggest that, in some conditions, adaptive evolution of mutation enhancers is realistic (Taddei et al., 1997). In this context, mobile and mutagenic sequences such as Transposable Elements (TEs) appear as natural candidates for evolvability helpers (Blot, 1994; Chao et al., 1983; Schneider and Lenski, 2004).

Transposable elements are self-duplicating DNA sequences that are present in virtually all living species (Biémont, 2010). Yet, understanding their presence, distribution, copy number, insertion patterns, and their propensity to be maintained in constant or changing environments is still under theoretical investigation (Charlesworth et al., 1994; Le Rouzic and Deceliere, 2005). Generally considered as genomic parasites in sexual organisms (Charlesworth and Charlesworth, 1983; Doolittle and Sapienza, 1980; Hickey, 1982; Orgel and Crick, 1980), their mobility promotes both deleterious mutations and genetic innovation. However, the spread of such selfish DNA requires sexual reproduction, and this mechanism cannot explain the persistence of TEs in selfing, parthenogenetic, and clonal organisms (Wright and Finnegan, 2001). Indeed, theoretical developments generally predict that active deleterious TEs should either be eliminated from asexual lineages, or drive them to extinction (Charlesworth and Charlesworth, 1983; Wright and Schoen, 1999; Dolgin and Charlesworth, 2006; Boutin et al., 2012), which has often been supported empirically (Zeyl et al., 1996; Arkhipova and Meselson, 2005). The presence of TE sequences in asexuals is thus generally attributed to rare but recurrent intra- or inter-specific horizontal transfers, compensating the extinction of TE-carrying lineages (Moody, 1988; Basten and Moody, 1991; Bichsel et al., 2010).

Understanding the impact of TEs on evolution and their role in the response to environmental pressure remains particularly challenging, as these sequences can be both beneficial and detrimental for their host (Capy

et al., 2000). Indeed, being mutagenic by nature, they are, on average, deleterious. Most insertions that are not neutral tend to disrupt useful genes, and only a small fraction of TE-driven mutations has the potential to be favored by natural selection, a process often referred to as 'molecular domestication' (Miller et al., 1992, 1997). TE-promoted evolutionary innovations include insertions, deletions, and recombinations, but may also involve TE sequences themselves as new genes or part of chimeric transcripts (Sinzelle et al., 2009). Consequently, TEs are generally considered as major contributors to genomic plasticity (Capy, 1998).

In clonal organisms, the rare occurrence of advantageous mutations may balance the fitness cost of carrying TEs, allowing the persistence of active copies in genomes. Interestingly, the dynamical properties of TEs in asexuals has led to little theoretical investigation compared to sexual populations. The possibility that prokaryotic TEs might act as evolvability enhancers was confirmed theoretically (Sawyer and Hartl, 1986; Martiel and Blot, 2002), but simulations were stopped after a single adaptive walk, leaving unexplored the dynamics of TEs once the fitness peak was reached. In the model proposed in McFadden and Knowles (1997), they are maintained for a long time because TE-promoted mutations allow TE-carrying lineages to cross adaptive valleys, and thus explore more efficiently the adaptive landscape. Although exciting, this model strongly relies on the hypothesis that TE-mediated mutations have significantly larger phenotypic effects than 'regular' background mutations, which does not appear to be supported empirically (Stoebel and Dorman, 2010). The idea that TEs could be maintained on a long-term due to recurrent environmental changes was developed more recently in Edwards and Brookfield (2003) and McGraw and Brookfield (2006), where the authors identified the timing of environmental shifts as the major factor conditioning the survival of TEs in clonal organisms. However, such models were explored only in simple cases (e.g. shifts between only two environments, unconditionally neutral insertions, no or limited evolution of TE sequences).

In particular, intra-genomic competition between TE copies may prevent TE-host systems from reaching an equilibrium. It is well-known that super-parasitic, non-autonomous elements are often successful, and can seriously impact the evolutionary dynamics of autonomous copies (Brook-

field, 1996; Hartl et al., 1992; Le Rouzic and Capy, 2006). Such intra-genomic competition between TE copies may lead to complex evolutionary dynamics, including TE loss or successive bursts of re-invasion, closely matching empirical observation (Le Rouzic et al., 2007).

In this chapter, we develop a general model of TE evolution in clonal organisms accounting for TE polymorphism (including autonomous and non-autonomous copies). Several environmental scenarios were considered (two being shown here), determining the size and the frequency at which TE-related mutations can be favored by natural selection, and the long-term dynamics of the TE-host system were explored for thousands of generations.

2.1 Method and results

Here we present a stochastic computational model of TE proliferation that enables exploration of the interplay between environmental changes and TE activity. We considered populations of $10,000$ clonally propagating individuals carrying both autonomous and non-autonomous TEs. Each organism is defined by its phenotype together with its TE genomic content. Simulations are initialized by introducing a single autonomous element in every individual of a population well-adapted to the current environment (all individuals are at the phenotypic optimum). See Figure 2.1.1 for the general outline of the model.

2.1.1 Phenotype and natural selection

The phenotype-fitness map is adapted from Fisher's geometric model (Fisher, 1930; Martin and Lenormand, 2006) with a moving optimum (Kopp and Hermisson, 2009; Orr, 2005). The phenotype of an individual is represented as a vector of n real numbers, each coordinate representing an independent trait involved in the adaptation of the organism to the environment.

The carrying capacity of the environment is m, i.e. the actual number of organisms fluctuates slightly around m. Associated with the environment

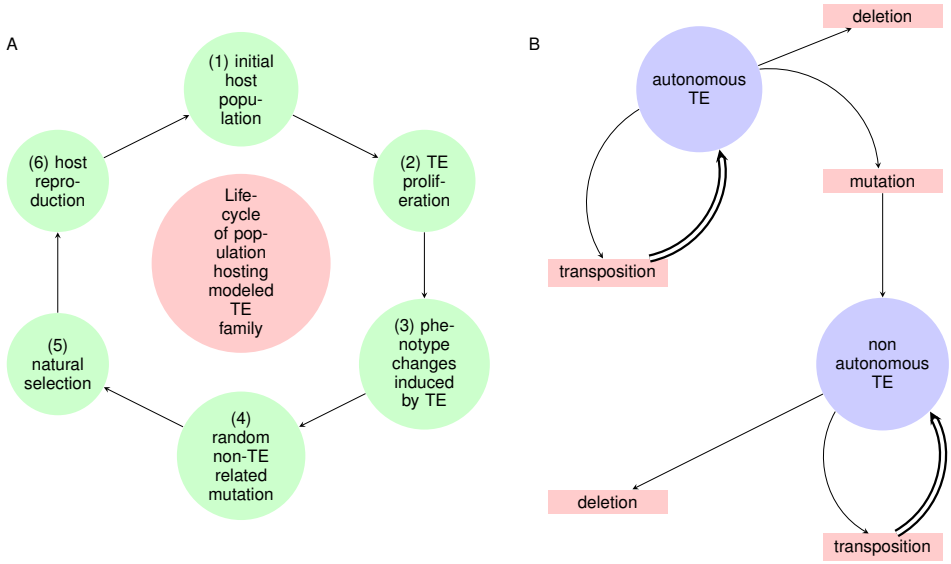


Figure 2.1.1: General outline of the model. (A): Life cycle of the population: (1) simulation starts with a population of 10,000 individuals; (2) both autonomous and non-autonomous transposable elements are mobilized by the transposition machinery produced by autonomous TEs; (3) each TE inserted causes a mutation of the host phenotype; (4) non transposition-related mutations also modify the phenotype (5) better adapted individuals (i.e. closer to the *optimal phenotype*) have greater probability of survival; (6) surviving individuals reproduce to fill the environment back to its capacity. (B): Evolution of transposable elements: an autonomous element can duplicate during the process of transposition or become non-autonomous; a non-autonomous element can still proliferate using the transposition machinery produced by other, autonomous elements.

is an 'optimal phenotype', i.e. a combination of phenotypes for which fitness is maximal.

Organisms whose phenotypes are close to the optimum are considered more 'fit' than organisms with phenotypes distant from the optimal phenotype. The fitness function is calculated from the standard n-dimensional Euclidean distance between the phenotype of an individual o (denoted by $\boldsymbol{\pi}(o) = [\pi_i(o)]_{i=1\ldots n}$) and the optimal phenotype $\hat{\boldsymbol{\pi}} = [\hat{\pi}_i]_{i=1\ldots n}$, as follows:

$$F(o) = \exp(-\text{dist}(\boldsymbol{\pi}(o), \hat{\boldsymbol{\pi}})^2) = \exp(-\sum_{i=1}^{n}(\pi_i(o) - \hat{\pi}_i)^2)$$

The fitness function does not depend on the TE count of an organism, and as such, does not enforce an artificial transposition-selection equilibrium.

Environmental change is modeled by shifting the optimal phenotype. We assume that among the n traits, $n/2$ have invariant optima and the other $n/2$ traits change every T generations by a deterministic factor s, so that the change is directional. The fixed traits are introduced in order to model more realistically a natural environment (which might be changing in some aspects, while remaining stationary in other). Additional simulations (not shown) confirm that the model behaves in a similar fashion for a wide range of 'fixed' traits (between 0 and about $0.8n$). In the scenario called 'Global Warming', the optimal phenotype changes by a small amount ($s_{GW} = 0.0002$) every generation ($T = 1$). In the 'Meteor Impact' scenario, the change is larger ($s_{MI} = 0.075$) and occurs every $T = 500$ generations.

Generations are non-overlapping. The number of offspring produced by an organism is drawn from Poisson distribution with the mean proportional to the organism's fitness. The relative fitness is multiplied by a scaling factor, chosen in each generation in such a way that the expected number of offspring equals the carrying capacity of the environment.

2.1.2 Transposition

Our model considers two kinds of transposable elements: autonomous and non-autonomous copies. Autonomous copies transpose with a constant rate τ per copy and per generation. Non-autonomous copies, which can

"parasitize" the transposition enzymes produced by autonomous copies, transpose at a rate of $\tau \cdot [A]$ per copy and per generation, where $[A]$, the concentration of transposition enzymes, is proportional to the number of autonomous copies in the cell. Therefore, non-autonomous copies cannot transpose in absence of autonomous copies, and their transposition rate increases with the autonomous copy number. In our stochastic simulations, the actual number of transpositions for each copy was sampled in a Poisson distribution. In addition to proliferating, autonomous TEs can spontaneously turn into non-autonomous copies with probability Δ_α, and both autonomous and non-autonomous TEs can disappear (by deletion or by being mutated beyond recognition) with probability Δ_β.

2.1.3 Mutations

Insertion events create *de novo* genetic variation, which in vivo may result in a range of functional alterations, ranging from gene knockouts to subtle regulatory shifts. In addition to transposition-related mutations, transposition-unrelated mutations (e.g. nucleotide substitutions) occur with a constant rate of $\rho = 0.003$. Both types of mutations have the same effect on the phenotype, shifting a single random phenotypic trait by a random number drawn from normal distribution, $Norm(0, \mu)$ where $\mu = 0.1$, the mutational standard deviation, is a parameter of our model. Note that the phenotypic change inflicted by transposition stays with the phenotype regardless of further fates (such as deletion) of the TE which caused it. Mutations are not pleiotropic, i.e. they do not affect several traits at once (in other words, the set of traits can be understood as independent phenotypic directions). Unlike the situation in most models, a positive number being randomly drawn does not necessarily result in a helpful mutation (just like a negative number need not result in a detrimental mutation). The effect of a mutation depends on the relative position of the host organism's phenotype and the optimal phenotype: for well-adapted organisms, most mutations are detrimental, as they push them away from the optimum. With mutational effects being drawn from a normal distribution, some mutations will be almost silent (those coming from the surroundings peak of the bell curve), while others will have

Parameter	Symbol	GW without non-autoTEs (Fig. 3B)	GW with non-autoTEs (Fig. 3C)	GW without TEs (Fig. 3A)	MI without TEs (Fig. 4A)	MI without non-auto TEs (Fig. 4B)	MI with non-auto TEs (Fig. 4C)
Dimension of phenotypic space	n	10	10	10	10	10	10
Mutation stdev.	μ	0.1	0.1	0.1	0.1	0.1	0.1
Non-TE-related mutation rate	ρ	0.003	0.003	0.003	0.003	0.003	0.003
Niche size	m	10000	10000	10000	10000	10000	10000
Autonomy loss probability	Δ_α	0.0	0.0003	—	—	0.0	0.0003
Deletion probability	Δ_β	0.003	0.003	—	—	0.003	0.003
Transposition rate	τ	0.003	0.003	—	—	0.003	0.003
Environmental change*	—	0.0002	0.0002	0.0002	0.075	0.075	0.075

Table 2.1.1: Parameters values used in simulations presented in: Fig. 3B (Global-warming without non-autonomous TEs), Fig. 3C (Global-warming with non-autonomous TEs), Fig. 3A (Global warming without TEs), Fig. 4A (Meteor impact without TEs, Fig. 4B (Meteor impact without non-autonomous TEs), and Fig. 4C (Meteor impact with non-autonomous TEs).
* Measured in phenotypic units per generation for GW and phenotypic units per impact (every 500 generations) for MI model.

noticeable impact on the phenotype, and the mutations coming from the tails of normal distribution are likely to have an immediately lethal effect. Table 2.1.1 presents parameter settings fixed in simulations.

2.1.4 Constant environmental pressure ('Global Warming' (GW) scenario)

The pressure exerted on the host population by slow, gradual environmental changes was modeled by a cumulative, directed shift of the 'optimal phenotype' in each consecutive generation. Both transposition activity and TE copy number increase with the intensity of environmental change (Figure 2.1.2). If the level of environmental change is very low to nonexistent, TEs are only deleterious, and disappear from the population.

A transposition-selection-drift equilibrium can be frequently observed under a moderate environmental change (Fig. 2.1.3(B)). Active transposition maintains a stable number of TE copies, as well as a moder-

(A) (B)

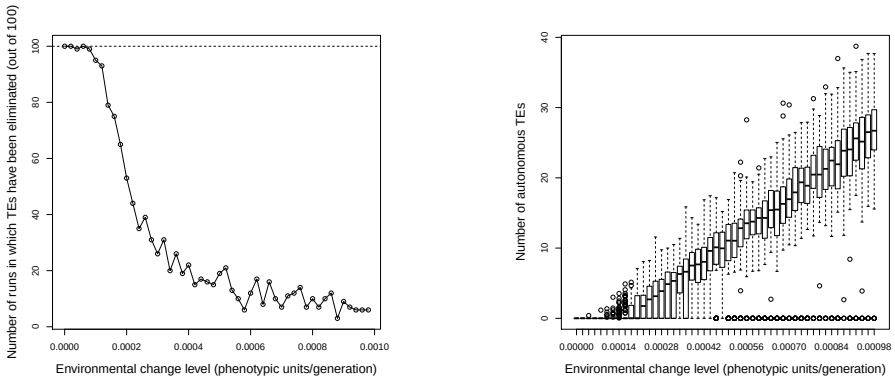


Figure 2.1.2: (A) Number of runs (out of 100) in which autonomous TEs are eliminated by generation 5000 under the "Global Warming" scenario (no non-autonomous copies). In low stress levels all TEs are lost, and the phenotypic optimum is tracked purely through TE-unrelated mutations. With higher levels of environmental change, TEs are maintained more often, and assist in tracking the phenotypic optimum by providing extra mutations. (B) Distribution of the average number of autonomous TE copies at generation 5000, from 100 simulation runs at each change level. The TE copy number at generation 5000 increases linearly with the rate of environmental change above the minimal threshold allowing for TE persistence.

ately high mutation rate (accounting for both transposition-related and transposition-unrelated mutations). Although most mutations are deleterious, some are beneficial and become fixed in the host population. If the environment is constant (not shown), transposition activity is only deleterious, and clones carrying TE copies are lost.

When autonomous TEs can mutate into non-autonomous TEs with frequency $\Delta_\alpha = 0.0003$ per generation, after an initial stage similar to the previous case (autonomous elements are active and stimulate the mutation rate), non-autonomous copies amplify, and the number of autonomous copies decreases (Fig. 2.1.3(C)). The transposition rate (and the induced mutation rate) are maintained, since only a few autonomous copies are enough to stimulate the transposition of many non-autonomous copies. However, this stage is followed by the loss of all autonomous copies, which eventually leads to the loss of transposition activity. At the end of the simulations, all TEs disappear, and the evolvability of populations (its capacity to track environmental changes) is reduced. Fig. 2.1.3(A) presents simulation without TEs for reference. In this case transposition-unrelated mutations manage to track the optimal phenotype, but the average fitness is lower, than with the presence of TEs.

2.1.5 Rapid environmental change ('Meteor Impact' (MI) scenario)

When the environmental change is large and instantaneous, populations do not have the possibility to track the optimum. The optimum shift is thus followed by a stage of directional selection, during which the average fitness in the population remains below the original fitness. If the population is evolvable enough, it can reach the new optimum between two "impacts", otherwise, stages of directional selection follow each other.

Figure 2.1.4 shows a situation in which the changes are too large and too frequent to be tracked efficiently by transposition-unrelated mutations only: if there are no TEs, the population never reaches the phenotypic optimum (absence of individuals having the optimal fitness), see Fig. 2.1.4(A). When autonomous TEs are present, the mutation rate increases, and the population can reach the optimum. Once at the optimum,

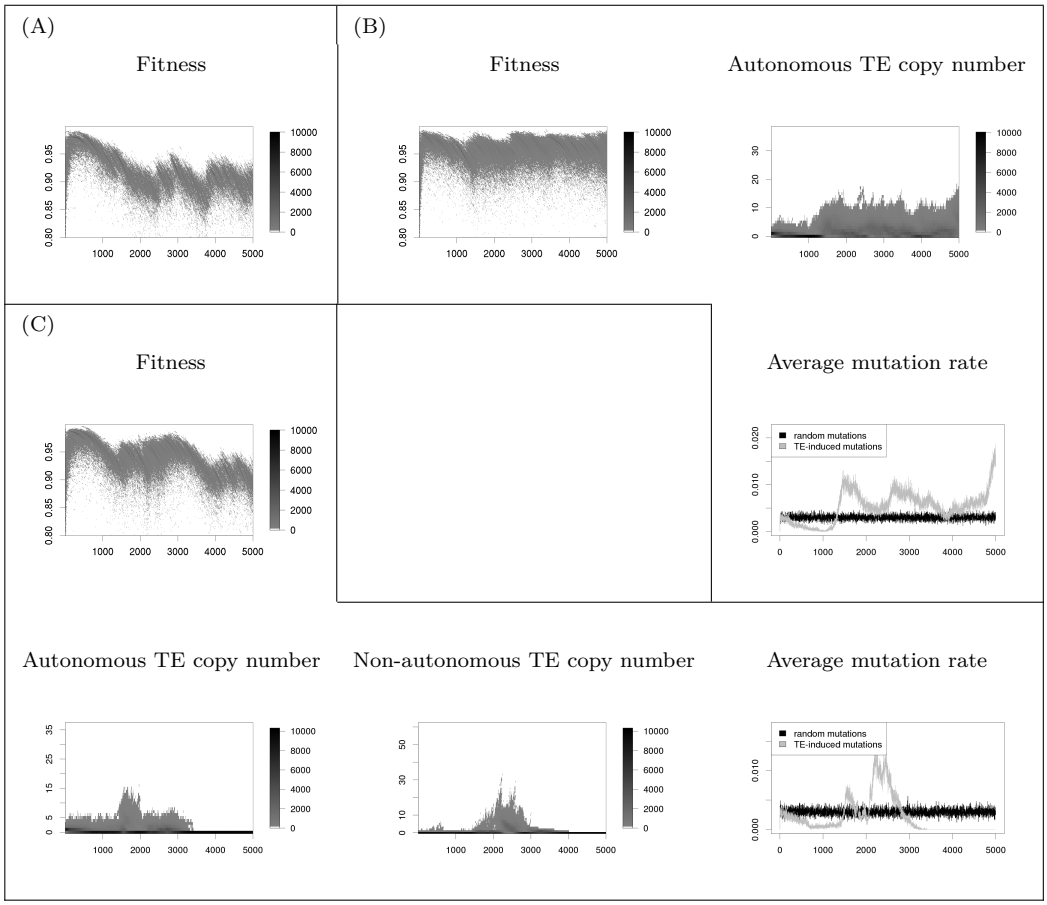


Figure 2.1.3: Effect of smooth environmental change ('Global Warming' scenario). (A) Reference simulation without TEs: transposition-unrelated mutations are enough to track the optimum in the long term. The distance of the population from the optimum is larger than with TEs (the average fitness is lower). The color scale is proportional to the density distribution of fitnesses in the population. (B) Non-autonomous copies disabled. TEs assist in tracking of the phenotypic optimum. (C) Non-autonomous copies enabled.

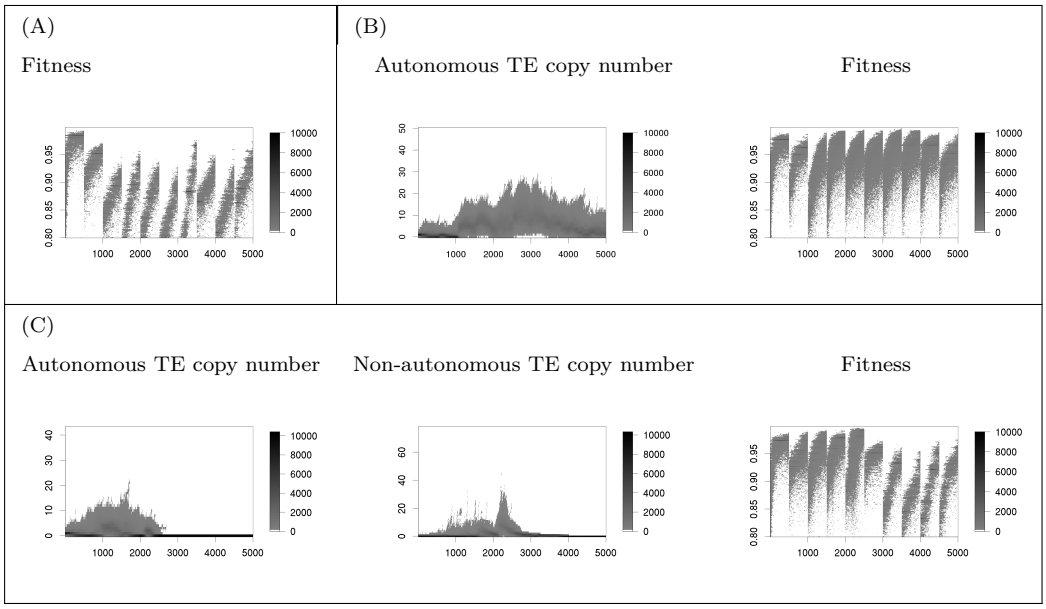


Figure 2.1.4: Effect of periodic dramatic environmental shifts ('Meteor Impact' scenario) (A) Without TEs the population diverges from the phenotypic optimum (B) With autonomous copies the population is able to track more closely the phenotypic optimum (C) With both autonomous and non-autonomous copies, the population follows the phenotypic optimum, until TEs are eliminated.

TE activity is only deleterious, and the copy number tends to drop. If a new environmental change happens before the loss of all copies, active TEs proliferate again, which can lead to the long-term maintenance of active TE copies (Fig. 2.1.4(B)).

Introducing non-autonomous copies has a similar effect as in the GW scenario. Non-autonomous mutants take over the autonomous copies, up to the point where all active copies are eliminated. The major consequence of the loss of transposition activity is a decrease in evolvability, leading to a fitness drop, and the inability to cope efficiently with environmental challenges (Fig. 2.1.4(C)).

2.1.6 General properties of the model

We tested our model using a range of parameter settings, including different population sizes, transposition and mutation rates, etc. In addition to the simulation engine, we have developed a GUI interface for browsing the results accessible from `http://bioputer.mimuw.edu.pl/transp`.

The model generally behaves in a stable fashion. As a rule, increase of the environmental change level results in a raise in TE activity. TEs proliferate most intensely under long-lasting directional changes. In contrast, realistic levels of non-cumulative random environmental fluctuations did not result in any significant TE activation: even if the level of stress imposed by random fluctuations is close to being lethal to the host population, the number of TEs rises, but remains an order of magnitude lower than in the GW scenario. The figures (especially Fig. 2.1.3) show that the main mechanism of evolution relies on selective sweeps by clones carrying beneficial mutations (almost instantaneous fitness increase), followed by stages of slow fitness decay, during which the different lineages tend to accumulate deleterious mutations independently. This mode of evolution is reminiscent of the patterns observed in "mutator" models (see Discussion).

2.2 Discussion

2.2.1 Model

The model described in this chapter aims at understanding and predicting the long-term evolution of transposable elements in asexual species living in changing environments. Compared to the relevant literature, this model brings several significant improvements: (i) the diversity of TEs is represented through autonomous and non-autonomous copies, (ii) the mutational effect of TE mobility on phenotype is modeled explicitly, (iii) fitness is determined according to the distance to a phenotypic optimum, and (iv) this setting makes it possible to model complex and various environmental change scenarios. Our modeling results indicate that TEs can be maintained in asexuals for a wide range of scenarios involving environmental change, as suggested in Edwards and Brookfield (2003); McGraw and Brookfield (2006). However, we conclude that intragenomic competition tends to affect the stability of the host-TE system, and our simulations repeatedly report the loss of the transposition activity when non-autonomous copies are present. If confirmed, this observation precludes the persistence of long-term TE-host symbiosis, restraining TE-

genome cooperation to short periods.

As for any model, our framework remains a simplification of reality, and many details were not accounted for. For instance, even asexual organisms are known to share episodically DNA sequences, through e.g. bacterial conjugation or horizontal transfers (HT). However, a preliminary study we have performed suggests that these do not alter significantly the dynamics of TE families, as long as their frequency remains reasonable when compared to mutation rates, transposition rates, and selection coefficients, which is likely. Very high (unreasonably high) incidence of HTs has, however, the potential to gradually shut down transposition. In any case, sporadic horizontal transfers remain essential for TE invasion, since they constitute the likeliest explanation for the occurrence of the initial copy of the TE family, and for the persistence of TEs in spite of their unstable dynamics when intra-genomic selection is introduced in the models.

For computational reasons, population size remained limited to $10,000$ in most runs, and we could not simulate the evolution of realistic prokaryotic populations, which size often reaches 10^9 or beyond. We have performed one study in which we let the population size vary between 10^2 and 10^7, and found that the impact of genetic drift is small with population sizes $> 1,000$, and is not likely to alter significantly the conclusions of this study, at least with the range of parameters considered. According to previous studies, genetic drift in small populations tends to (i) increase the number of copies (by limiting the efficiency of natural selection against deleterious insertions), and (ii) increase the risk of TE loss or population extinction (Edwards and Brookfield, 2003; Le Rouzic et al., 2007). In contrast, in extremely large populations, TE dynamics are expected to be smoother and more deterministic.

One of the most challenging aspect of TE modeling is the way to introduce the impact of a changing environment in the model. Here, we considered that environmental stochasticity corresponds to a change in the fitness function: when the environment changes, the population is no longer close to the phenotypic optimum, and thus it has to accumulate genetic changes to improve its fitness. In this regard, our model is comparable to the mutator system (Giraud et al., 2001; Taddei et al., 1997).

Mutators are clones characterized by high mutation rates, several orders of magnitude above the base mutation rate of the species. Theoretical models suggest that a long-term coexistence of mutators and non-mutators is possible when the environment changes regularly (Gillespie, 1981; Tanaka et al., 2003; Travis and Travis, 2002), and their dynamical properties was confirmed experimentally (Giraud et al., 2001). Our model thus confirms that TEs could play the role of mutator-like factors (as observed empirically by Fehér et al. (2012)), by increasing the mutation rate in a flexible way when the environment changes. Yet, their capacity to amplify exponentially can also lead to lineage extinction (Rankin et al., 2010; Vinogradov, 2003), making TEs efficient, but dangerous, evolutionary helpers.

Here, we did not consider any direct effect of stress on transposable elements or on mutation rates: in our model, the transposition rate increases solely as a consequence of the accumulation of active copies. Some empirical results suggest that TE mobility might also be directly induced by stress (Capy et al., 2000; Grandbastien et al., 2005; Ogasawara et al., 2009), opening the way towards models considering epigenetic regulation of transposable elements. Therefore, it cannot be excluded that stress-induced transposition might be adaptive if environmental change generates physiological stress (e.g. by threatening the survival of the population). This setting could be similar to the 'SOS' system in bacteria (Janion, 2008; Radman, 1974), involving a stress-induced epigenetic increase in mutation rate.

2.2.2 Impact of TEs on genome evolution

Transposable elements are generally considered as universal, and they may represent most of the genomic DNA, especially in multicellular eukaryotes: 45% in human (Lander et al., 2001), and up to 85% in maize (Schnable et al., 2009). In other eukaryotic phyla, TEs may be less overwhelming, as they constitute around 2% of the genome of *Caenorhabditis elegans*, and 3% in the yeast *Saccharomyces cerevisiae* (Kidwell and Lisch, 2000). Even prokaryotes, with their tiny optimized genomes, are not devoid of TEs, called 'insertion sequences' (IS) (Chandler and Mahillon, 2002).

Although population genetic models generally focus on sexual, random-

mating species, most lineages of living organisms harbor asexual reproduction regimes, with only rare and sporadic gene transfers. Members of two out of three kingdoms of life, Eubacteria and Archaea, reproduce clonally. Eukarya are featured by a higher diversity of reproduction regimes, including perfect asexuality, parthenogenesis, self-fertilization, and sexual mating. Asexuality can be found in multicellular eukaryotes, including fungi, plants, and even some animals.

Our model considers strictly asexual organisms, and could correspond to any clonal prokaryotic or eukaryotic species. Even in asexuals, genetic transfer events might occur, but theoretical models predict that exchange rates need to be very large for the population to behave as a sexual species in terms of TE content (Condit et al., 1988), which excludes the vast majority of asexuals. Mobile DNA content in genomes differs greatly between asexual clades. Eubacteria are generally thought to have a very small number of TEs, most of them being active and recent (Wagner, 2006). Nevertheless, the situation is not homogeneous, and the genome of some strains harbors up to 20% of TE-derived sequences (Newton and Bordenstein, 2011). Archaea do not appear as fundamentally different, although they might contain more copies in average, including non-autonomous insertions (Filée et al., 2007). In contrast, the genome of eukaryotes is much larger, and contain many more TEs. Some asexual animals (such as bdelloid rotifers) tend to have fewer TE copies than sexual relatives (Arkhipova and Meselson, 2000), but the pattern is less clear for plants and fungi (Dufresne et al., 2011; Lockton and Gaut, 2010).

Even if plant, animal, and prokaryotic TEs are not exactly identical, large differences in TE content across organisms do not necessarily reflect different TE properties. Indeed, ecological or environmental factors can also interact with TE dynamics, and condition their evolution. It is suspected that the population size could explain some of the differences in genome size and TE content: the efficiency of natural selection at eliminating slightly deleterious insertions being higher at large population sizes, the accumulation of TEs is much faster in low-population size species (such as multicellular eukaryotes) than in very large population-size prokaryotes ((Lynch, 2007; Lynch and Conery, 2003), but see (Daubin

and Moran, 2004; Charlesworth and Barton, 2004; Whitney and Garland, 2010)). Our results suggest that TE accumulation is also more likely in asexual populations subject to frequent environmental change than in populations living in constant environment, with little evolutionary challenge. This hypothesis is supported by the observation that the TE genomic content in bacteria might be influenced by environmental factors (Newton and Bordenstein, 2011).

2.2.3 Conclusions

In this chapter, we developed a model of TE evolution in asexual organisms that's sufficiently realistic for analysis of real-world phenomena. This model allows the evolution of TE copies, and implements an explicit effect of TE mobility on phenotypic traits which, in conjunction to the environment, determines individual fitness. These simulations evidence that, contrary to what is generally assumed, TE dynamics in asexuals can be extremely rich and complex, featuring losses, re-invasions, bursts of non-autonomous copies, and lineage extinctions. These results show that environment remains a major factor conditioning the genomic content of mobile DNA, through the carrying capacity of the habitat, the frequency at which new evolutionary challenges occur, and the size of the corresponding evolutionary steps. The interplay between intra-genomic competition between TE copies and natural selection at the individual level illustrate the rich and complex coevolutionary nature of the TE-host relationship.

Most importantly, we have confirmed the suspicion that TEs may be stimulated by environmental stress, proving that our suspicions are sound in principle, and laying ground for further investigations. The first direction is the extension of the model presented above, and here two directions present themselves: spatial modelling, and addition of sexual reproduction to the model.

The addition of spatial modelling is the most intuitive way of expanding the model beyond well-mixed populations: an addition of a spatial plane with close-range interactions between plants, such as resource-competition and constrained living space. This will allow the study of the effect of

spatial barriers on transposable elements, more detailed ways of observing colonization of new environments (it is hypothesis ed that colonization front-wave will be marked by an increased TE activity), and a study of mixing between subpopulations. This work is being carried out by Matuesz Kitlas as his Master's Thesis, and the first results are promising.

Another direction of extension of the model is the swapping of the reproduction model from asexual to sexual. This entails the addition of explicit modelling of (a diploid) genome, to account for effects such as insertion site polymorphism, hetero- or homozygosity, and sexual inheritance. This extension is being implemented by Krzysztof Gogolewski as a part of his Master's Thesis. Preliminary results suggest that, surprisingly, even in sexual setting the TEs may act as evolutionary helpers, which runs contrary to the consensus.

Another possible direction is substituting class-I TEs for class-II transposons. However, as we have proposed and solved a mathematical model of class-II transposons (so we have a computational model for class-I TEs and mathematical model for class-II transposons), which will be presented in the next chapter, it seems unlikely that this direction will be pursued in the near future – except perhaps, as a combined model for class-I and class-II TEs, enabling us to study if they possess different roles, and to study their interactions.

more controllable, or at least, possible to quantify. The results are presented in the next chapter.

3

Mathematical modelling of TE-carrying populations in the presence of environmental stress

In this chapter we will present the results of mathematical modelling of the behaviour of transposable elements, their proliferation, and interaction with environmental stress. Mathematical models have several advantages with respect to the computational models, first and foremost being the capability to rigorously prove or disprove a thesis, and to compute the exact values of certain parameters and functions of population, as they evolve with time. Keeping this in mind it is no surprise that we have attempted to apply this approach to the modelling of activity of transposable elements. In this chapter, a full model of transposable elements shall be presented to give the reader an idea of what is the goal. We shall derive models for class I and class II TEs. The model for class I TEs will only be formulated, as so far it seems to be to be mathematically intractable. Next, a model for mutators shall be introduced, and fully analyzed. An

equilibrium state shall be derived, convergence of arbitrary populations to the equilibrium state shall be proven, and various parameters of the equilibrium population shall be derived. Note that while, by volume, the analysis of the mutator model will be the most prominent feature of this chapter, it is only a stepping stone for a model for class II transposons (which turns out to be a relatively small expansion of the mutator model), for which too a solution shall be provided.

3.1 Previous work

This chapter presents work presenting analytical solutions of scenarios within Fisher's Geometric Model (Fisher, 1930), with moving optimum. So far, no such solutions have been obtained (cf. Bürger (2000) p. 324). Scenarios with moving phenotypic optimum have been studied extensively, and there is definitely much scientific interest in them, however all the previous works so far have been based on computer simulations which do not provide a formula for solution (Collins et al., 2007), or sweeping simplifying assumptions and approximations Burger and Lynch (1995); Waxman and Peck (1999) which change the end result – or on a mixture of both approaches Matuszewski et al. (2014). As far as we can reasonably tell, we provide here the first analytical solution of this case which does not use approximations, allows general (in particular: non-Gaussian) initial populations, and is not based on numerical simulations.

3.2 Notation and basic definitions

Before we begin construction of the model, first, several notations and definitions which we are going to use must be established.

- $\nu(\mu, \sigma)$ is the probability density function (PDF) of a Gaussian distribution, with mean μ and standard deviation σ. It is a function: $\nu(\mu, \sigma) : \mathbb{R} \to [0, +\infty)$, evaluating it at point x is written as: $\nu(\mu, \sigma)(x)$. Formally:

$$\nu(\mu, \sigma)(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- From now on, we shall be using the formalism of the Lebesgue integral, with λ denoting the standard Lebesgue measure on $\mathbb{R}$, and $\int_A f(x)\mathrm{d}\mu(x)$ meaning the Lebesgue integral of function f, over a measurable set A, with respect to measure μ. Leb($\mathbb{R}$) shall denote the σ-algebra of Lebesgue-measurable subsets of $\mathbb{R}$. The implicit σ-algebra of measurable sets, unless explicitly said otherwise, will be the Leb($\mathbb{R}$).

- In addition to that, over the course of some proofs, in several places in this section, we shall be using the Fubini-Tonelli's theorem (Rudin (1987) p. 164(. As there is a slight terminological confusion with regard to the theorem, the theorem having several, slightly different versions (and variously being called Fubini's theorem, Fubini-Tonelli's theorem, or just Tonelli's theorem) it appears worthy to recall the form which will be used. From now on, whenever Fubini-Tonelli's theorem is mentioned, it is understood to be the following one:

  Theorem (Fubini-Tonelli's Theorem)**.** Let $(X, \mathfrak{M}, \mu)$ and $(Y, \mathfrak{N}, \nu)$ be two σ-finite measure spaces, and let $f : X \times Y \to [0, +\infty)$ be a measurable (and nonnegative) function. Then:

$$\int_X \left(\int_Y f(x,y)\mathrm{d}\nu(y) \right) \mathrm{d}\mu(x) = \int_Y \left(\int_X f(x,y)\mathrm{d}\mu(x) \right) \mathrm{d}\nu(y) =$$

$$= \int_{X\times Y} f(x,y) d(\mu \times \nu)(x,y)$$

  We are going to be using only the standard Lebesgue measure (which is σ-finite), along with probabilistic measures (which, again, are finite, and therefore, also σ-finite). All the functions that we are going to be using are measurable, and nonnegative – therefore, we shall be applying the above theorem without any further verification that its assumptions hold.

- By $\mathcal{N}(\mu, \sigma)$ we shall understand the probability measure of a Gaussian distribution with mean μ and standard deviation σ. That is:

$\mathcal{N}(\mu, \sigma) : \text{Leb}(\mathbb{R}) \to [0, 1]$. For a Lebesgue-measurable subset A of R, the formula is as follows:

$$\mathcal{N}(\mu, \sigma)(A) = \int_A \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} d\lambda(x)$$

- π_x is orthogonal projection. That is, for example, let $F : \mathbb{R}^2 \to \mathbb{R}^2$ such that $F(a, b) = (c, d)$.
  Then, $\pi_x(F(a, b)) = c$ and $\pi_y(F(a, b)) = d$.

- We shall make use of the fact that the product of two Gaussian PDFs is itself proportional to a Gaussian PDF (Aldershof et al., 1995):

$$\nu(\mu_1, \sigma_1)(x) \cdot \nu(\mu_2, \sigma_2)(x) = c \cdot \nu\left(\frac{\mu_1 {\sigma_2}^2 + \mu_2 {\sigma_1}^2}{{\sigma_2}^2 + {\sigma_1}^2}, \frac{\sigma_1\sigma_2}{\sqrt{\sigma_1^2 + \sigma_2^2}}\right)(x) \tag{3.1}$$

  where:

$$c = \frac{2\sqrt{\pi}\sqrt{{\sigma_2}^2 + {\sigma_1}^2}\, e^{\frac{{\mu_2}^2 - 2\mu_1\mu_2 + {\mu_1}^2}{2{\sigma_2}^2 + 2{\sigma_1}^2}}}{\sqrt{2}}$$

- In addition to that, several parameters will be used in formulas, their meaning, unless specified otherwise is the following:

  - ρ is the random mutation rate, that is, the rate of background mutations which are not associated with transposition.

  - σ is the radius of selection (the lower it is, the more stringent selection is).

  - η is the speed of environmental change, per generation – the optimal phenotype shall be moved by this amount in each generation.

3.3 Transposition model for class I TEs

3.3.1 Introduction – general setting

In the model presented here we remain within the context of asexual, haploidal, clonal organisms. The model is, again, going to be an extension of classical Fisher's geometric model (Fisher, 1930; Martin and Lenormand, 2006) with a moving optimum (Kopp and Hermisson, 2009; Orr, 2005) – that is, it is assumed that every organism in the population shall carry a so-called *phenotype* – a real-valued vector describing some of the properties of the organism which are relevant to its ability to survive in the environment. In addition to that, there is the *optimal phenotype* – a single real-valued vector (of the same dimension), which describes the environment. Organisms whose phenotypes are closest to the optimal phenotype shall thrive, while organisms whose phenotypes are distant – shall die off. In addition to that, each organism carries some TEs – their amount is stored in a single nonnegative integer for each organism. Here, for the sake of simplicity we shall assume that the phenotypic space is one-dimensional, however, the results presented easily carry over to higher dimensions as well.

As such, each organism is completely described by two numbers: a real-valued number, its phenotype, marking its location within the phenotypic space, and a nonnegative integer – the number of TEs its carrying. In this model, we only track active TEs – and we assume that sufficient transposition machinery is always present within a given organism, so that any TEs present may duplicate. Thus, the question of autonomous versus non-autonomous TEs becomes meaningless, and differentiation between them – unnecessary. Again, at least for now, we only focus on class I TEs, the ones which proliferate using a copy-paste mechanism.

In contrast to the computational model we assume a continuous population – as, in mathematical settings these are easier to deal with than the discrete populations favoured by computational models. The population is modelled by a probability density function $p : \mathbb{N} \times \mathbb{R} \to \mathbb{R}$ (with $\sum_{\mathbb{N}} \int_{\mathbb{R}} p = 1$). One might immediately notice that this particular setting enforces "constant" population size, without concerning us about the

actual size of population. The PDF models only relative proportions of organisms with different parameters within the population with respect to each other, it does not provide actual organism counts. Because of that it is impossible for the population to grow, shrink, or most importantly, die off. This might be seen as unrealistic, however, we do not wish to explore the mechanics of species extinction within this work, and our experience with the computational model shows that such effects complicate the model behaviour without adding any new interesting data. And, as mathematical tractability remains an issue which we have to face, we decided against including such effects in our model.

3.3.2 Mathematical formulation

The evolution of the system proceeds in discrete generations, with each new generation completely displacing the previous one. Over one generation the mutations are applied, selection (and renormalization) shifts the proportions of organisms, and, last but not least, the optimal phenotype is shifted by a fixed amount, to represent environmental stress. In mathematical terms, the consecutive generations are obtained with a "generation operator" $\Phi : (\mathbb{N} \times \mathbb{R} \to \mathbb{R}) \to (\mathbb{N} \times \mathbb{R} \to \mathbb{R})$, which transforms population distributed according to p into the next generation of this population $\Phi(p)$.

Next, we shall construct the operator Φ in such a way that is reflects the behaviour of TEs found in nature. The construction, for purposes of facilitating understanding, shall be done in steps, with each step providing the form of Φ for a simplified scenario, with the scenarios getting progressively more complicated, until we arrive at the final form.

Transposition

First, let us tackle the situation in which the organisms only have TEs, and phenotype is disregarded (that is: each organism is described only by its number of TEs, and the population is a function $p : \mathbb{N} \to \mathbb{R}$. The TEs may proliferate and vanish as normal.

- We make the assumption that transposition and TE decay occur

concurrently in a single generation, and do not affect each other. Thus the probabilities of transposition and deletion are independent.

- Probability of d TE deletions occurring in an organism with m initial TEs is modelled by binomial distribution: $\mathbb{P}(\#del_m = d) = \binom{m}{d}\Delta^d(1-\Delta)^{(n-d)}$, where Δ is TE deletion rate. This is consistent with the biological mechanism – each TE may independently of others get excised, or mutate so much that it looses its function.

- Probability of t transpositions occurring in an organism with m initial TEs is modelled by Poisson distribution: $\mathbb{P}(\#tr_m = t) = \frac{(\tau m)^t e^{(-\tau m)}}{t!}$. This is because we assume that there is sufficient transposition machinery for each TE to proliferate, again, independently from others (we do not model the effects of TEs competing for transposition machinery).

- Finally, the evolution operator for TE-carrying organisms has the following form:

$$\Phi(p)(n) = \sum_{d\in\mathbb{N}}\sum_{t\in\mathbb{N}} \mathbb{P}(\#del_{n+d-t} = d)\mathbb{P}(\#tr_{n+d-t} = t)p(n+d-t)$$

Phenotypic landscape

Next, let us consider the reverse situation: we disregard TEs, and only care about the phenotype. The organisms only mutate, with a Gaussian mutation function. We do not model selection yet:

- We assume that mutation shifts the phenotype by a number drawn from a normal distribution $N(0,\rho)$, where ρ is the mutation rate. These are the mutations which are independent of TEs.

- Disregarding the transposition, and assuming a random mutation rate ρ gives us (where $\nu(0,\rho)$ is the PDF of $N(0,\rho)$, star denotes function convolution) the following:

$$\Phi(p)(x) = (p \star \nu(0,\rho))(x)$$

Transposition and phenotypes combined

Combining transpositions and mutations of the phenotype (we still do not model selection, or environmental stress yet) yields the following (we assume that the mutation rate of an organism is the random mutation rate plus the number of transpositions it goes through in this generation: $\rho+t$):

$$\Phi(p)(n,x) =$$

$$\Big((\sum_{d\in\mathbb{N}}\sum_{t\in\mathbb{N}}\mathbb{P}(\#del_{n+d-t}=d)\mathbb{P}(\#tr_{n+d-t}=t)p(n+d-t,\_))\star\nu(0,\rho+t)\Big)(x)$$

Environmental stress

- We will later assume that 0 is the phenotypic optimum. Environmental change may be represented by moving that optimum, however moving optimum in one direction is equivalent to moving the whole population in the phenotypic space in the other direction (mathematically equivalent by a simple parameter substitution), and this is easier to do with the formulas. Therefore, a constant environmental change with magnitude η may be modelled as follows:

  $$\Phi(p)(n,x) =$$

  $$\Big(\sum_{d\in\mathbb{N}}\sum_{t\in\mathbb{N}}\mathbb{P}(\#del_{n+d-t}=d)\mathbb{P}(\#tr_{n+d-t}=t)p(n+d-t,\_)\star\nu(0,\rho+t)\Big)(x-\eta)$$

Selection and final formalism

Last but not least, selection – we will assume that organisms are selected according to normal distribution, centered around the optimum (zero). The selection strength (that is, the standard deviation of this distribution) will be σ. This is done simply by multiplying the PDF describing the population by a Gaussian PDF centered around the phenotypic optimum. Note that this takes us outside of the firm ground of probability theory and into unknown: such an act has no probabilistic interpretation, and indeed, yields a function which, in general, need not be a PDF. However, this can be alleviated by dividing the function by a normalization factor: this can be interpreted as selection killing off some of the organisms (as

the integral of the function after selection, in all but the most pathologic of cases will be less than 1), and renormalization – the remaining organisms producing clones of themselves to refill the environmental niche. Thus, if we let $m = n + d - t$ the "generation operator" with selection looks as following:

$$\Phi(p)(n,x) = \frac{\left(\left(\sum_{d\in\mathbb{N}} \sum_{t\in\mathbb{N}} \mathbb{P}(\#del_m = d)\mathbb{P}(\#tr_m = t)p(m,\_) \star \nu(0,\rho+t) \right)(x-\eta) \right) \cdot \nu(0,\sigma)}{\int_{\mathbb{R}\times\mathbb{N}} \left(\left(\sum_{d\in\mathbb{N}} \sum_{t\in\mathbb{N}} \mathbb{P}(\#del_m = d)\mathbb{P}(\#tr_m = t)p(m,\_) \star \nu(0,\rho+t) \right)(x-\eta) \right) \cdot \nu(0,\sigma) \mathrm{d}\lambda(\mathbb{R}\times\mathbb{N})}$$

The model, as formulated, has so far successfully resisted all attempts at solving – that is, we cannot determine the asymptotic behaviour of iterating this operator.

3.4 Mutator model

Next, we shall formally introduce the so-called mutator model, which is a simplified version of the model for class-I TEs. The model, in addition to its own merits, will be, in next section, used to construct a mathematical formalism for populations carrying class-II transposons.

3.4.1 Introduction

The setting remains the same, that is, clonally-reproducing organisms, subjected to environmental stress, within Fisher's geometric framework. The organisms in a population possess a fixed (parametrised) mutation rate (whether the mechanism of these mutations is related to transposons or not is of no consequence for now, as long as the mutation rate remains constant, and the same for each organism in the population). The goal is to study the effects of various rates of mutation and selection of the organisms on the ability and speed of adaptation of the whole population to the new environment. The results will then be sued to derive a mathematical model for class-II transposons.

3.4.2 Simplification of model for class-I TEs

Let us start with the Φ operator as defined at the end of the previous section. If there are no TEs in the population (that is, when $n = 0$, or, more precisely, $p(n, x) = 0$ if $n > 0$), then new TEs cannot appear (as the only source of new TEs in this setting is transposition, which requires preexisting TEs). This leaves only random mutations as the cause of phenotypic variance, and, as such, we may eliminate the parameter responsible for TEs altogether. In this case, the operator may be simplified into the following form:

$$\Phi(p)(x) = \frac{\hat{\Phi}(p)(x)}{\int_{\mathbb{R}} \hat{\Phi}(p)(t)\mathrm{d}t} \tag{3.2}$$

where:

$$\hat{\Phi}(p)(x) = (p \star \nu(0, \rho))(x - \eta) \cdot \nu(0, \sigma)(x)$$

3.4.3 Stationary state

Description

A fixpoint of Φ is a distribution p which satisfies $\Phi(p) = p$ – that is, a population which remains the same in consecutive generations. Note that, although the mathematical formalism suggests a stationary population, this is in fact far from the truth: the population is stationary with respect to the reference point, namely the optimal phenotype – which is constantly moving to represent environmental change. It is only due to the parameter substitution mentioned earlier that the optimal phenotype is, in each generation, fixed at the origin of the coordinate system, creating illusion of stability. In fact, the fixpoint represents a population which is moving through the phenotypic space, however it is following the optimal phenotype with constant speed at a fixed distance, remaining stationary with respect to it.

Solution in the class of Gaussian distributions

It can be easily verified that in the class of Gaussian functions, Φ has a single, unique one fixpoint:

$$p = \nu \left(\frac{\eta \sqrt{4\,\sigma^2 + \rho^2} - \eta\,\rho}{2\,\rho}, \frac{\sqrt{2\rho(\sqrt{4\sigma^2 + \rho^2} - \rho)}}{2} \right) \tag{3.3}$$

Note that, surprisingly, the phenotypic variance of the fixpoint population does not depend on the strength of the global warming, only the phenotypic mean does.

However, two important questions remain:

- Is this also the single, unique fixpoint in the class of other, non-Gaussian distributed populations?
- Do (all? some?) non-fixpoint populations converge to the fixpoint population with time?

It turns out that the answer to both of the above questions is “yes”, and that can be derived analytically.

3.4.4 Proof of convergence

Φ in terms of probability measures

In fact, it is possible to prove the convergence to the fixpoint for any initial population, not only Gaussian ones, even for populations where the probability distribution describing them does not have a probability density function. Such populations are observed in nature – in fact, every real population is described by a discrete distribution, since the number of organisms in the population is finite. However, for most populations a continuous distribution is a good, and frequently used, approximation. Regardless of that, there are scenarios with populations which cannot reasonably be modelled by a continuous distribution: consider, for example, a single bacterium being dropped onto a surface, and growing into a colony. The initial population consists of only 1 specimen, and, as such, should be modelled by a Dirac delta – and there is no reasonable continuous approximation of that.

However, it is possible to prove convergence for any distribution, be it discrete, continuous, or even singular (or any combination thereof), however, in order to do that, first, the operator Φ must be rephrased in terms of probability measures.

Recall that the convolution of two probability distributions $\mathbb{P}_1$ and $\mathbb{P}_2$ defined on $\mathbb{R}$ (with probability density functions p_1 and p_2, respectively), when expressed on their PDFs has the following form:

$$(p_1 \star p_2)(x) = \int_{\mathbb{R}} p_1(t)p_2(x-t)\mathrm{d}t$$

while, when expressed in terms of the probability measures, becomes:

$$(\mathbb{P}_1 \star \mathbb{P}_2)(A) = \int_{\mathbb{R}}\int_{\mathbb{R}} \mathbb{1}_A(x+y)\mathrm{d}\mathbb{P}_1(x)\mathrm{d}\mathbb{P}_2(y)$$

Keeping that in mind, it is simple to observe, that Φ, when expressed in terms of probability measures, has the following form:

$$\Phi(\mathbb{P})(A) = \frac{\hat{\Phi}(\mathbb{P})(A)}{\hat{\Phi}(\mathbb{P})(\mathbb{R})}$$

where:

$$\hat{\Phi}(\mathbb{P})(A) = \int_{\mathbb{R}} \int_{\mathbb{R}} \nu(0,\sigma)(y+z) \cdot \mathbb{1}_{A+\eta}(y+z) \ \mathrm{d}\mathcal{N}(0,\rho)(y)\mathrm{d}\mathbb{P}(z) =$$

$$= \int_{\mathbb{R}} \int_{\mathbb{R}} \nu(0,\sigma)(y+z) \cdot \mathbb{1}_{A+\eta}(y+z) \cdot \nu(0,\rho)(y) \ \mathrm{d}\lambda(y)\mathrm{d}\mathbb{P}(z)$$

The class $\mathcal{CN}$

We shall begin by defining a certain well-behaved class of probability distributions. Next, we shall prove that the image of Φ lies within the class, thus enabling us to restrict further studies of the behaviour of Φ to within this class.

Definition 1. Let $\mathcal{CN} = \{\mathbb{P} : \mathrm{Leb}(\mathbb{R}) \to [0,1] \mid \exists \sigma \in \mathbb{R} \ \exists \mathbb{S} : \mathrm{Leb}(\mathbb{R}) \to [0,1] \ \ \mathrm{pdf}_{\mathbb{P}}(x) = \int_{\mathbb{R}} \nu(y,\sigma)(x)\mathrm{d}\mathbb{S}(y)\}$ - that is, the set of all distributions which are a compound of Gaussian distribution all with the same standard deviation, and with mean distributed according to some distribution $\mathbb{S}$.

Theorem 1. Let Φ be the operator defined in Equation 3.2. Then:

$$\Phi(P(\mathbb{R})) \subseteq \mathcal{CN}$$

that is, the image of the set of all probability distributions on $\mathbb{R}$ is in the class $\mathcal{CN}$ (or: every possible starting population, after one generation of evolution can be described by a $\mathcal{CN}$-class probability distribution).

Proof. Let $\mathbb{P} : \mathrm{Leb}(\mathbb{R}) \to [0,1]$. For all $A \in \mathrm{Leb}(\mathbb{R})$ we have:

$$\hat{\Phi}(\mathbb{P})(A) = \int_{\mathbb{R}} \int_{\mathbb{R}} \nu(0,\sigma)(y+z) \cdot \mathbb{1}_{A+\eta}(y+z) \cdot \nu(0,\rho)(y) \ \mathrm{d}\lambda(y)\mathrm{d}\mathbb{P}(z) =$$

$$= \int_{\mathbb{R}} \int_{\mathbb{R}} \nu(0,\sigma)(y) \cdot \mathbb{1}_{A+\eta}(y) \cdot \nu(0,\rho)(y-z)\ \mathrm{d}\lambda(y-z)\mathrm{d}\mathbb{P}(z) =$$

because the Lebesgue measure is translation-invariant:

$$= \int_{\mathbb{R}} \int_{\mathbb{R}} \nu(0,\sigma)(y) \cdot \mathbb{1}_{A+\eta}(y) \cdot \nu(0,\rho)(y-z)\ \mathrm{d}\lambda(y)\mathrm{d}\mathbb{P}(z) =$$

using basic properties of Gaussian function:

$$= \int_{\mathbb{R}} \int_{\mathbb{R}} \nu(0,\sigma)(y) \cdot \nu(z,\rho)(y) \cdot \mathbb{1}_{A+\eta}(y)\ \mathrm{d}\lambda(y)\mathrm{d}\mathbb{P}(z) =$$

using Equation (3.1):

$$= c \cdot \int_{\mathbb{R}} \int_{\mathbb{R}} \nu\left(\frac{z\sigma^2}{\sigma^2+\rho^2}, \frac{\sigma\rho}{\sqrt{\sigma^2+\rho^2}}\right)(y) \cdot \mathbb{1}_{A+\eta}(y)\ \mathrm{d}\lambda(y)\mathrm{d}\mathbb{P}(z) =$$

$$= c \cdot \int_{\mathbb{R}} \int_{A+\eta} \nu\left(\frac{z\sigma^2}{\sigma^2+\rho^2}, \frac{\sigma\rho}{\sqrt{\sigma^2+\rho^2}}\right)(y)\mathrm{d}\lambda(y)\mathrm{d}\mathbb{P}(z) =$$

Integrating by substitution: $\overline{z} = \frac{z\sigma^2}{\sigma^2+\rho^2}$

$$= c \cdot \frac{\sigma^2+\rho^2}{\sigma^2} \int_{\mathbb{R}} \int_{A+\eta} \nu\left(\overline{z}, \frac{\sigma\rho}{\sqrt{\sigma^2+\rho^2}}\right)(y)\mathrm{d}\lambda(y)\mathrm{d}\mathbb{P}(\overline{z})$$

Therefore,

$$\Phi(\mathbb{P})(A) = \frac{\hat{\Phi}(\mathbb{P})(A)}{\hat{\Phi}(\mathbb{P})(\mathbb{R})} = \int_{\mathbb{R}} \int_{A+\eta} \nu\left(\overline{z}, \frac{\sigma\rho}{\sqrt{\sigma^2+\rho^2}}\right)(y)\mathrm{d}\lambda(y)\mathrm{d}\mathbb{P}(\overline{z}) =$$

$$= \int_{\mathbb{R}} \int_{A} \nu\left(\overline{z}, \frac{\sigma\rho}{\sqrt{\sigma^2+\rho^2}}\right)(y-\eta)\mathrm{d}\lambda(y)\mathrm{d}\mathbb{P}(\overline{z})$$

Using Fubini's Theorem:

$$= \int_{A} \int_{\mathbb{R}} \nu\left(\overline{z}, \frac{\sigma\rho}{\sqrt{\sigma^2+\rho^2}}\right)(y-\eta)\ \mathrm{d}\mathbb{P}(\overline{z})\mathrm{d}\lambda(y)$$

Therefore, the PDF of $\Phi(\mathbb{P})$ is:

$$\mathrm{pdf}_{\Phi(\mathbb{P})}(x) = \int_{\mathbb{R}} \nu\left(z, \frac{\sigma\rho}{\sqrt{\sigma^2+\rho^2}}\right)(x-\eta)\ \mathrm{d}\mathbb{P}(z) =$$

$$= \int_{\mathbb{R}} \nu\left(z+\eta, \frac{\sigma\rho}{\sqrt{\sigma^2+\rho^2}}\right)(x)\ \mathrm{d}\mathbb{P}(z) =$$

$$= \int_{\mathbb{R}} \nu\left(z, \frac{\sigma\rho}{\sqrt{\sigma^2+\rho^2}}\right)(x)\ \mathrm{d}\mathbb{P}(z-\eta)$$

and is of proper form for $\Phi(\mathbb{P})$ to belong to the class $\mathcal{CN}$. □

The class $\mathcal{CND}$

Next, we shall perform an analogous trick, constraining the space of probability distributions which we need to consider even further, by observing that the image of the class $\mathcal{CN}$ under Φ is even smaller.

Definition 2. Let $\mathcal{CND} = \{\mathbb{P} : \mathrm{Leb}(\mathbb{R}) \to [0,1] \mid \exists\sigma \in \mathbb{R}\ \exists\mathbb{S} : \mathrm{Leb}(\mathbb{R}) \to [0,1]\ \ \mathrm{pdf}_{\mathbb{P}}(x) = \int_{\mathbb{R}} \nu(y,\sigma)(x)\mathrm{d}\mathbb{S}(y)$ and $\mathbb{S}$ has a probability density function$\}$ – that is, the subset of $\mathcal{CN}$ where the distribution of mean of the compound Gaussian distribution has a PDF. Alternatively, one can write:
$\mathcal{CND} = \{\mathbb{P} : \mathrm{Leb}(\mathbb{R}) \to [0,1] \mid \exists\sigma \in \mathbb{R}\ \exists s : \mathbb{R} \to \mathbb{R}_{+,0}\ \ \mathrm{pdf}_{\mathbb{P}}(x) = \int_{\mathbb{R}} \nu(y,\sigma)(x)\cdot s(y)\ \mathrm{d}\lambda(y)\}$

Theorem 2. Let Φ be the operator defined by Equation (3.2). Then:

$$\Phi(\mathcal{CN}) \subseteq \mathcal{CND}$$

Proof. Let $\mathbb{P} \in \mathcal{CN}$. Retrace the steps of proof of Theorem 1 for $\mathbb{P}$, obtaining:

$$\mathrm{pdf}_{\Phi(\mathbb{P})}(x) = \int_{\mathbb{R}} \nu\left(z, \frac{\sigma\rho}{\sqrt{\sigma^2+\rho^2}}\right)(x)\ \mathrm{d}\mathbb{P}(z-\eta)$$

By definition $\mathbb{P}$ has a probability distribution function, and therefore, so has $\mathbb{P}(z-\eta)$, therefore, $\Phi(\mathbb{P}) \in \mathcal{CND}$. □

Action of Φ within $\mathcal{CND}$

Let p be the PDF of a probability measure in $\mathcal{CND}$. Therefore (by definition of $\mathcal{CND}$) $p(x) = \int_{\mathbb{R}} \nu(y,\varsigma)(x) \cdot s(y)\ \mathrm{d}y$ for some s.

Let us find $\Phi(p)$. Note that now that we're within the class $\mathcal{CND}$, we are back on the firm ground of probability distributions which have a density function, therefore we may go back to the form of Φ which operates on PDFs.

$$\hat{\Phi}(p)(x) = (p \star \nu(0,\rho))(x-\eta) \cdot \nu(0,\sigma)(x) = \int_{\mathbb{R}} p(t) \cdot \nu(0,\rho)(x-\eta-t)\mathrm{d}t \cdot \nu(0,\sigma)(x) =$$

$$= \int_{\mathbb{R}}\int_{\mathbb{R}} \nu(y,\varsigma)(t) \cdot s(y)\ \mathrm{d}y \cdot \nu(0,\rho)(x-\eta-t)\mathrm{d}t \cdot \nu(0,\sigma)(x) =$$

$$= \int_{\mathbb{R}}\int_{\mathbb{R}} \nu(y,\varsigma)(t) \cdot s(y) \cdot \nu(0,\rho)(x-\eta-t)\ \mathrm{d}y\mathrm{d}t \cdot \nu(0,\sigma)(x) =$$

using Fubini-Tonelli's Theorem:

$$= \int_{\mathbb{R}}\int_{\mathbb{R}} \nu(y,\varsigma)(t) \cdot s(y) \cdot \nu(0,\rho)(x-\eta-t)\ \mathrm{d}t\mathrm{d}y \cdot \nu(0,\sigma)(x) =$$

$$= \int_{\mathbb{R}}\int_{\mathbb{R}} \nu(y,\varsigma)(t) \cdot \nu(0,\rho)(x-\eta-t)\ \mathrm{d}t\ s(y)\ \mathrm{d}y \cdot \nu(0,\sigma)(x) =$$

$$= \int_{\mathbb{R}}\int_{\mathbb{R}} \nu(y,\varsigma)(t) \cdot \nu(\eta,\rho)(x-t)\ \mathrm{d}t\ s(y)\ \mathrm{d}y \cdot \nu(0,\sigma)(x) =$$

$$= \int_{\mathbb{R}} \Big(\nu(y,\varsigma) \star \nu(\eta,\rho)\Big)(x) \cdot s(y)\ \mathrm{d}y \cdot \nu(0,\sigma)(x) =$$

$$= \int_{\mathbb{R}} \nu\left(y+\eta, \sqrt{\varsigma^2+\rho^2}\right)(x) \cdot s(y)\ \mathrm{d}y \cdot \nu(0,\sigma)(x) =$$

$$= \int_{\mathbb{R}} \nu\left(y+\eta, \sqrt{\varsigma^2+\rho^2}\right)(x) \cdot s(y) \cdot \nu(0,\sigma)(x)\ \mathrm{d}y =$$

using Equation (3.1):

$$= c \cdot \int_{\mathbb{R}} \nu\left(\frac{(y+\eta)\sigma^2}{\varsigma^2+\sigma^2+\rho^2}, \frac{\sigma\sqrt{\varsigma^2+\rho^2}}{\sqrt{\varsigma^2+\rho^2+\sigma^2}}\right)(x) \cdot s(y)\ \mathrm{d}y$$

therefore:

$$\Phi(p)(x) = \int_{\mathbb{R}} \nu\left(\frac{(y+\eta)\sigma^2}{\varsigma^2+\sigma^2+\rho^2}, \frac{\sigma\sqrt{\varsigma^2+\rho^2}}{\sqrt{\varsigma^2+\rho^2+\sigma^2}}\right)(x) \cdot s(y)\ \mathrm{d}y$$

Since Φ is seen to be only operating on the parameters of the normal distribution in the above formula (while leaving s unchanged), let us denote by F the transformation that Φ performs on the parameters:

$$F(x,y) = \left(\frac{(x+\eta)\sigma^2}{y^2+\sigma^2+\rho^2}, \frac{\sigma\sqrt{y^2+\rho^2}}{\sqrt{y^2+\rho^2+\sigma^2}}\right) \tag{3.4}$$

Note: at this point it is possible to double-check the correctness of the proof so far by verifying that:

$$F\left(\frac{\eta\sqrt{4\,\sigma^2+\rho^2}-\eta\,\rho}{2\,\rho}, \frac{\sqrt{2\rho(\sqrt{4\sigma^2+\rho^2}-\rho)}}{2}\right) =$$

$$= \left(\frac{\eta\sqrt{4\,\sigma^2+\rho^2}-\eta\,\rho}{2\,\rho}, \frac{\sqrt{2\rho(\sqrt{4\sigma^2+\rho^2}-\rho)}}{2}\right)$$

It is indeed the case (calculations not shown).

Limit of the standard deviation

At this point, we may abandon Φ altogether, for a while, and study the much simpler object, that is the function F. Proving that iterating F leads to convergence is much easier than proving the same fact for Φ, and yet, it will allow us to derive the same conclusion for Φ later. We will prove convergence separately for both coordinates of F. As the second coordinate of F does not depend in any way on the first (while the first one does depend on second) it will be easier to start with the second coordinate:

Theorem 3. Recall that π_y denotes the orthogonal projection onto the second dimension. Let F be as defined in Equation (3.4). Then:

$$\forall x, y \in \mathbb{R} \quad \lim_{n\to\infty} \pi_y(F^n(x,y)) = \frac{\sqrt{2\rho(\sqrt{4\sigma^2+\rho^2}-\rho)}}{2}$$

Proof. Since $\pi_y(F)$ doesn't depend on x variable, we might as well consider $f(y) = \frac{\sigma\sqrt{y^2+\rho^2}}{\sqrt{y^2+\rho^2+\sigma^2}}$ (which is equal to $\pi_y(F(x,y))$ for all x).

$$\frac{\mathrm{d}f}{\mathrm{d}y} = \frac{\sigma^3 y}{\sqrt{\rho^2+y^2}(\rho^2+\sigma^2+y^2)^{3/2}} \leq \frac{\sigma^3 y}{\sqrt{y^2}(\rho^2+\sigma^2+y^2)^{3/2}} =$$

$$= \frac{\sigma^3}{(\rho^2+\sigma^2+y^2)^{3/2}} \leq \frac{\sigma^3}{(\rho^2+\sigma^2)^{3/2}}$$

therefore f is Lipschitz with constant $\frac{\sigma^3}{(\rho^2+\sigma^2)^{3/2}}$. And since $\rho > 0$ as it is the standard deviation of a distribution, then:

$$\frac{\sigma^3}{(\rho^2+\sigma^2)^{3/2}} < \frac{\sigma^3}{(\sigma^2)^{3/2}} = 1$$

therefore f is a contraction mapping (with this constant, which is strictly less than 1). By applying Banach's contraction mapping theorem (Smart (1980) p. 2) we arrive at the conclusion that f has a unique fixpoint, and that iterating f converges to it. Knowing this, all that remains is to verify the (previously established) actual value of the fixpoint, that is

check that:

$$f\left(\frac{\sqrt{2\rho(\sqrt{4\sigma^2+\rho^2}-\rho)}}{2}\right)=\frac{\sqrt{2\rho(\sqrt{4\sigma^2+\rho^2}-\rho)}}{2}$$

(trivial yet tedious calculations skipped) □

Limit of the mean

Now that we have proved that the second coordinate of F converges, all that's left is to prove the convergence of the first coordinate.

Theorem 4. Let F be as defined in Equation 3.4. Denote: $(x_n, y_n) = F^n(x, y)$. Then:

$$\lim_{n\to\infty} x_n = \frac{\eta\sqrt{4\,\sigma^2+\rho^2}-\eta\,\rho}{2\,\rho}$$

Proof. Observe that x_n is defined by a recurrence relation:

$$x_n = \frac{(x_{n-1}+\eta)\sigma^2}{y_{n-1}^2+\sigma^2+\rho^2}$$

Taking that equation, and converging[1] on both sides with $n \to \infty$:

$$\lim_{n\to\infty} x_n = \lim_{n\to\infty} \frac{(x_{n-1}+\eta)\sigma^2}{y_{n-1}^2+\sigma^2+\rho^2}$$

Denoting: $x = \lim_{n\to\infty} x_n = \lim_{n\to\infty} x_{n-1}$ and $y = \lim_{n\to\infty} y_n = \frac{\sqrt{2\rho(\sqrt{4\sigma^2+\rho^2}-\rho)}}{2}$ we obtain:

$$x = \frac{(x+\eta)\sigma^2}{y^2+\sigma^2+\rho^2}$$

[1]Note: in order to do that we should first prove that the sequence actually converges. Proof omitted, one can do that by noting that for each arbitrarily small $\epsilon > 0$ there exists a M such that $y - \epsilon \leq y_n \leq y + \epsilon$ for each $n > M$. The convergence of sequences with y_n replaced by $y \pm \epsilon$ is easy to prove, they are bounded and monotonous. The original sequence lies between them, and as we converge with $\epsilon \to 0$ their limits approach, so we may perform reasoning similar to the squeeze theorem.

Solving this equation for x yields:

$$x = \frac{\eta\sigma^2}{y^2 + \rho^2}$$

It is easy to verify that substituting for y its value in the above yields expression equal to the limit for x_n from theorem. One way of checking this, is for example using a Computer Algebra System, subtracting the expressions and simplifying the result – this yields 0, proving that they are in fact equal. This was verified using Wolfram Mathematica 7.0.1.0. □

Lemmas for convergence theorems

Before we finally prove the convergence of Φ^n, we still need to prove two more lemmas which we will later use to swap the order of limits and integrals in the final proof.

Lemma 5. The class of functions: $\{\nu\left(\frac{(x+\eta)\sigma^2}{y^2+\sigma^2+\rho^2}, \frac{\sigma\sqrt{y^2+\rho^2}}{\sqrt{y^2+\rho^2+\sigma^2}}\right) | x \in \mathbb{R}, y > 0\}$ is uniformly bounded, and therefore, uniformly integrable with respect to any probability measure $\mathbb{P} : \text{Leb}(\mathbb{R}) \to [0, 1]$

Proof. The PDF of a normal distribution has maximum in the mean. Therefore, its maximum is equal to:

$$\max_{z\in\mathbb{R}} \nu\left(\frac{(x+\eta)\sigma^2}{y^2+\sigma^2+\rho^2}, \frac{\sigma\sqrt{y^2+\rho^2}}{\sqrt{y^2+\rho^2+\sigma^2}}\right)(z) =$$

$$= \frac{1}{\sqrt{2\pi}\left(\frac{\sigma\sqrt{y^2+\rho^2}}{\sqrt{y^2+\rho^2+\sigma^2}}\right)} e^{0/\left(2\left(\frac{\sigma\sqrt{y^2+\rho^2}}{\sqrt{y^2+\rho^2+\sigma^2}}\right)^2\right)} =$$

$$= \frac{\sqrt{y^2+\rho^2+\sigma^2}}{\sqrt{2\pi}\sigma\sqrt{y^2+\rho^2}} = \frac{1}{\sigma\sqrt{2\pi}}\sqrt{\frac{y^2+\rho^2+\sigma^2}{y^2+\rho^2}} =$$

$$= \frac{1}{\sigma\sqrt{2\pi}}\sqrt{1+\frac{\sigma^2}{y^2+\rho^2}} \leq \frac{1}{\sigma\sqrt{2\pi}}\sqrt{1+\frac{\sigma^2}{\rho^2}}$$

which is a common bound for all functions in that class. □

Lemma 6. For all $(x, y) \in \mathbb{R}\times(0, \infty)$ the class of functions: $\{\nu(F^n(x, y))\}_{n\in\mathbb{N}^+}$ is dominated by a single nonnegative function f, such that $\int_{\mathbb{R}} f(z)\mathrm{d}z < \infty$

The proof of this lemma is long and tedious, and does not easily reveal the underlying idea behind it, so before we proceed with a full, formal proof, we shall provide a sketch of proof.

The main idea behind the proof is that the set $\{F^n(x, y)\}_{n\in\mathbb{N}^+}$ forms a convergent sequence, and therefore it is bounded. If it is bounded, then there is a Gaussian distribution with a biggest variance (or, obviously, a sequence of distributions converging to the largest variance, but for now, for the purposes of this sketch, let's disregard the case of sequences, it will be formally handled in the proof), a distribution with smallest variance, a distribution with largest mean and lowest mean. The idea is to construct a dominating function f from blocks: the area between the largest and smallest mean will be covered by a continuous function, at supremum of all the values attained by all the PDFs in the class, while sides are covered by an arm of Gaussian distribution, descending as slow as the slowest function from the class, and rescaled to start at the maximum (c.f. Fig. 3.4.1). The dominating function is obviously integrable as it is a scaled PDF of a Gaussian distribution, with a rectangular block inserted into the middle.

And now, the tedious formal proof:

Proof. Denote, as previously: $(x_n, y_n) = F^n(x, y)$. From Theorems 3 and 4 it follows that both x_n and y_n are convergent, and therefore, bounded. Moreover, clearly $\lim_{n\to\infty} y_n > 0$, and therefore also $\inf_{n\in\mathbb{N}^+}\{y_n\} > 0$ Keeping that in mind, let us calculate:

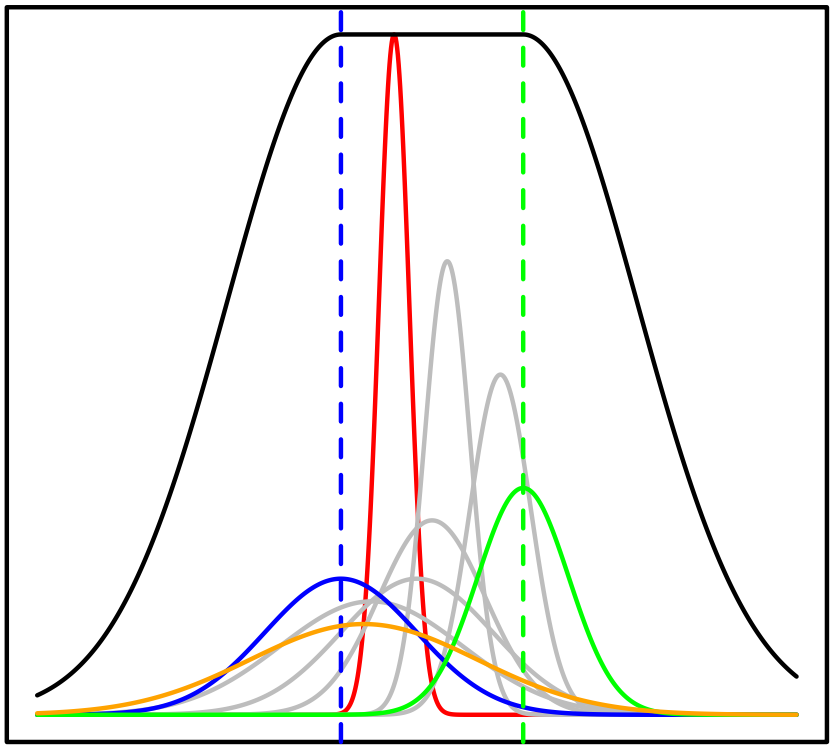


Figure 3.4.1: Construction of the dominating function f, as in proof of Lemma 6. The dominating function is in black, the Gaussian PDF with lowest variance is red, the one with highest variance is orange, the one with lowest mean is blue, and the one with highest is black; (some) other PDFs from the class have also been plotted, in gray.

$$\nu(F^n(x,y))(z) = \frac{1}{y_n\sqrt{2\pi}} \exp\left(-\frac{(z-x_n)^2}{2y_n^2}\right) \leq$$
$$\leq \frac{1}{\inf_{n\in\mathbb{N}^+}\{y_n\}\sqrt{2\pi}} \exp\left(-\frac{(z-x_n)^2}{2\sup_{n\in\mathbb{N}^+}\{y_n\}^2}\right)$$

The exp(...) term is always less than or equal to 1, furthermore, for $z > \sup_{n\in\mathbb{N}^+}\{x_n\}$ the above expression is less than:

$$\frac{1}{\inf_{n\in\mathbb{N}^+}\{y_n\}\sqrt{2\pi}} \exp\left(-\frac{(z-\sup_{n\in\mathbb{N}^+}\{x_n\})^2}{2\sup_{n\in\mathbb{N}^+}\{y_n\}^2}\right)$$

and similarly for $z < \inf_{n\in\mathbb{N}^+}\{x_n\}$.

Therefore, let us define:

$$f(z) = \begin{cases} \frac{1}{\inf_{n\in\mathbb{N}^+}\{y_n\}\sqrt{2\pi}} \exp\left(-\frac{(z-\inf_{n\in\mathbb{N}^+}\{x_n\})^2}{2\sup_{n\in\mathbb{N}^+}\{y_n\}^2}\right) & \text{for } z < \inf_{n\in\mathbb{N}^+}\{x_n\} \\ \frac{1}{\inf_{n\in\mathbb{N}^+}\{y_n\}\sqrt{2\pi}} & \text{for } z \in [\inf_{n\in\mathbb{N}^+}\{x_n\}, \sup_{n\in\mathbb{N}^+}\{x_n\}] \\ \frac{1}{\inf_{n\in\mathbb{N}^+}\{y_n\}\sqrt{2\pi}} \exp\left(-\frac{(z-\sup_{n\in\mathbb{N}^+}\{x_n\})^2}{2\sup_{n\in\mathbb{N}^+}\{y_n\}^2}\right) & \text{for } z > \sup_{n\in\mathbb{N}^+}\{x_n\} \end{cases}$$

From the above it follows that f dominates the class of functions from the theorem, so the only thing left to prove is that it is integrable:

$$\int_{\mathbb{R}} f(z)\,\mathrm{d}z = \int_{(-\infty,\inf_{n\in\mathbb{N}^+}\{x_n\})} \frac{1}{\inf_{n\in\mathbb{N}^+}\{y_n\}\sqrt{2\pi}} \exp\left(-\frac{(z-\inf_{n\in\mathbb{N}^+}\{x_n\})^2}{2\sup_{n\in\mathbb{N}^+}\{y_n\}^2}\right)\,\mathrm{d}z +$$
$$+ \int_{[\inf_{n\in\mathbb{N}^+}\{x_n\},\sup_{n\in\mathbb{N}^+}\{x_n\}]} \frac{1}{\inf_{n\in\mathbb{N}^+}\{y_n\}\sqrt{2\pi}}\,\mathrm{d}z +$$
$$+ \int_{(\sup_{n\in\mathbb{N}^+}\{x_n\},+\infty)} \frac{1}{\inf_{n\in\mathbb{N}^+}\{y_n\}\sqrt{2\pi}} \exp\left(-\frac{(z-\sup_{n\in\mathbb{N}^+}\{x_n\})^2}{2\sup_{n\in\mathbb{N}^+}\{y_n\}^2}\right)\,\mathrm{d}z =$$
$$= \frac{\sup_{n\in\mathbb{N}^+}\{x_n\} - \inf_{n\in\mathbb{N}^+}\{x_n\}}{\inf_{n\in\mathbb{N}^+}\{y_n\}\sqrt{2\pi}} +$$
$$+2\int_{(\sup_{n\in\mathbb{N}^+}\{x_n\},+\infty)} \frac{1}{\inf_{n\in\mathbb{N}^+}\{y_n\}\sqrt{2\pi}} \exp\left(-\frac{(z-\sup_{n\in\mathbb{N}^+}\{x_n\})^2}{2\sup_{n\in\mathbb{N}^+}\{y_n\}^2}\right)\,\mathrm{d}z =$$
$$= \frac{\sup_{n\in\mathbb{N}^+}\{x_n\} - \inf_{n\in\mathbb{N}^+}\{x_n\}}{\inf_{n\in\mathbb{N}^+}\{y_n\}\sqrt{2\pi}} + 2\frac{1}{\inf_{n\in\mathbb{N}^+}\{y_n\}\sqrt{2\pi}} \sup_{n\in\mathbb{N}^+}\{y_n\}\sqrt{\pi/2} < \infty$$

□

Convergence of Φ^n

And the final theorem, establishing a unique stationary population:

Theorem 7. For all $\mathbb{P} : \mathrm{Leb}(\mathbb{R}) \to [0,1]$ and all $A \in \mathrm{Leb}(\mathbb{R})$ the following holds:

$$\lim_{n\to\infty} \Phi^n(\mathbb{P})(A) = \mathcal{N}\left(\frac{\eta\sqrt{4\,\sigma^2+\rho^2}-\eta\,\rho}{2\,\rho}, \frac{\sqrt{2\rho(\sqrt{4\sigma^2+\rho^2}-\rho)}}{2}\right)(A)$$

that is, iterating Φ on any probability measure generates a sequence strongly converging to a normal distribution with the given mean and standard deviation.

Proof. With two applications of Φ we end up in $\mathcal{CND}$ class (by Theorem 1 and 2). Therefore it is enough to prove convergence for $\mathbb{P}$ of the following form:

$$\mathbb{P}(A) = \int_A \int_{\mathbb{R}} \nu\left(F^n(x,y)\right)(z) \cdot s(x)\ \mathrm{d}\lambda(x)\mathrm{d}\lambda(z)$$

Let us start:

$$\lim_{n\to\infty} \Phi^n(\mathbb{P})(A) = \lim_{n\to\infty} \int_A \int_{\mathbb{R}} \nu\left(F^n(x,y)\right)(z) \cdot s(x)\ \mathrm{d}\lambda(x)\mathrm{d}\lambda(z) =$$

Using Fubini-Tonelli's Theorem:

$$= \lim_{n\to\infty} \int_{\mathbb{R}} \int_A \nu\left(F^n(x,y)\right)(z) \cdot s(x)\ \mathrm{d}\lambda(z)\mathrm{d}\lambda(x) =$$

$$\lim_{n\to\infty}\int_{\mathbb{R}}\int_{A}\nu\left(F^{n}(x,y)\right)(z)\ \mathrm{d}\lambda(z)\cdot s(x)\ \mathrm{d}\lambda(x)=$$

$$\lim_{n\to\infty}\int_{\mathbb{R}}\int_{A}\nu\left(F^{n}(x,y)\right)(z)\ \mathrm{d}\lambda(z)\ \mathrm{d}\mathbb{S}(x)=$$

$\nu\left(F^{n}(x,y)\right)(z)$ is a probability density function, therefore $\int_{A}\nu\left(F^{n}(x,y)\right)(z)\ \mathrm{d}\lambda(z)\leq 1$, therefore the class $\{\nu\left(F^{n}(x,y)\right)(z)\}_{n\in\mathbb{N}^{+}}$ is uniformly bounded (by 1). $\mathbb{S}$ is a probability measure, therefore we may use Bounded Convergence Theorem (Rudin (1976) p. 322) to obtain:

$$=\int_{\mathbb{R}}\lim_{n\to\infty}\int_{A}\nu\left(F^{n}(x,y)\right)(z)\ \mathrm{d}\lambda(z)\ \mathrm{d}\mathbb{S}(x)=$$

by Lemma 6 we may use Dominated Convergence Theorem (Rudin (1976) p. 321) (see proof of the lemma for dominating function):

$$=\int_{\mathbb{R}}\int_{A}\lim_{n\to\infty}\nu\left(F^{n}(x,y)\right)(z)\ \mathrm{d}\lambda(z)\ \mathrm{d}\mathbb{S}(x)=$$

Using continuity of ν with respect to its parameters:

$$=\int_{\mathbb{R}}\int_{A}\nu\left(\lim_{n\to\infty}F^{n}(x,y)\right)(z)\ \mathrm{d}\lambda(z)\ \mathrm{d}\mathbb{S}(x)=$$

Using Theorem 3 and 4: $F^{n}(x,y)\xrightarrow[n\to\infty]{}\left(\frac{\eta\sqrt{4\,\sigma^{2}+\rho^{2}}-\eta\,\rho}{2\,\rho},\frac{\sqrt{2\rho(\sqrt{4\sigma^{2}+\rho^{2}}-\rho)}}{2}\right)$, therefore:

$$=\int_{\mathbb{R}}\int_{A}\nu\left(\frac{\eta\sqrt{4\,\sigma^{2}+\rho^{2}}-\eta\,\rho}{2\,\rho},\frac{\sqrt{2\rho(\sqrt{4\sigma^{2}+\rho^{2}}-\rho)}}{2}\right)(z)\,\mathrm{d}\lambda(z)\,\mathrm{d}\mathbb{S}(x)=$$

$$\int_{A}\nu\left(\frac{\eta\sqrt{4\,\sigma^{2}+\rho^{2}}-\eta\,\rho}{2\,\rho},\frac{\sqrt{2\rho(\sqrt{4\sigma^{2}+\rho^{2}}-\rho)}}{2}\right)(z)\ \mathrm{d}\lambda(z)\cdot\int_{\mathbb{R}}\mathrm{d}\mathbb{S}(x)=$$

$\mathbb{S}$ is a probability measure, therefore $\int_{\mathbb{R}}\mathrm{d}\mathbb{S}(x)=1$:

$$=\int_{A}\nu\left(\frac{\eta\sqrt{4\,\sigma^{2}+\rho^{2}}-\eta\,\rho}{2\,\rho},\frac{\sqrt{2\rho(\sqrt{4\sigma^{2}+\rho^{2}}-\rho)}}{2}\right)(z)\ d\lambda(z)=$$

$$= \mathcal{N}\left(\frac{\eta\sqrt{4\,\sigma^2+\rho^2}-\eta\,\rho}{2\,\rho}, \frac{\sqrt{2\rho(\sqrt{4\sigma^2+\rho^2}-\rho)}}{2}\right)(A)$$

□

3.4.5 Analysis of the equilibrium state

Independence of of populational variability from the speed of environmental change

The first, and perhaps most surprising observation is the one we have already made over the course of the proof, namely that the variance of phenotypes in the population is not affected by the amount of environmental stress the population is subjected to. This fact has severe consequences for population genetics and ecological studies – as it points to the fact that it is not possible to estimate the environmental stress that a given population is subjected to by measuring its phenotypic variance.

Survivability

An interesting question is the estimation of survivability in the equilibrium state, that is, the proportion of organisms which survive the selection step. This can be performed by applying the $\hat{\Phi}$ operator (that is, the Φ operator without the scaling factor) to the equilibrium population, and evaluating its integral. Denote by μ^* the stationary mean: $\mu^* = \frac{\eta\sqrt{4\,\sigma^2+\rho^2}-\eta\,\rho}{2\,\rho}$, and by σ^* the stationary standard deviation: $\sigma^* = \frac{\sqrt{2\rho(\sqrt{4\sigma^2+\rho^2}-\rho)}}{2}$. We shall use the PDF form:

$$\int_{\mathbb{R}} \hat{\Phi}\left(\nu\left(\mu^*,\sigma^*\right)\right)(x)\mathrm{d}\lambda(x) =$$

$$= \int_{\mathbb{R}} \left(\nu\left(\mu^*,\sigma^*\right) \star \nu(0,\rho)\right)(x-\eta)\cdot\nu(0,\sigma)(x)\mathrm{d}\lambda(x) =$$

$$= \int_{\mathbb{R}} \left(\nu\left(\mu^*, \sqrt{\left(\sigma^*\right)^2+\rho^2}\right)\right)(x-\eta)\cdot\nu(0,\sigma)(x)\mathrm{d}\lambda(x) =$$

$$= \int_{\mathbb{R}} \left(\nu\left(\mu^*+\eta, \sqrt{\left(\sigma^*\right)^2+\rho^2}\right)\right)(x)\cdot\nu(0,\sigma)(x)\mathrm{d}\lambda(x) =$$

At this point we may apply the formula for product of two Gaussian PDFs (Equation (3.1). Note that only the scaling factor is important, and we may disregard the mean and standard deviation of the resulting Gaussian PDF. The scaling factor was computed using a Computer Algebra System (Wolfram Mathematica 7.0.1.0):

$$= \int_{\mathbb{R}} \frac{e^{-\frac{\eta^2}{2\rho^2}}\sqrt{\frac{2}{\pi}}}{\sigma\sqrt{\rho\left(\rho+\sqrt{\rho^2+4\sigma^2}\right)}\sqrt{\frac{\rho+\sqrt{\rho^2+4\sigma^2}}{\rho\sigma^2}}}\nu(...,...)(x)\mathrm{d}\lambda(x) =$$

$$= \frac{e^{-\frac{\eta^2}{2\rho^2}}\sqrt{\frac{2}{\pi}}}{\sigma\sqrt{\rho\left(\rho+\sqrt{\rho^2+4\sigma^2}\right)}\sqrt{\frac{\rho+\sqrt{\rho^2+4\sigma^2}}{\rho\sigma^2}}}\int_{\mathbb{R}}\nu(...,...)(x)\mathrm{d}\lambda(x) =$$

$$= \frac{e^{-\frac{\eta^2}{2\rho^2}}\sqrt{\frac{2}{\pi}}}{\sigma\sqrt{\rho\left(\rho+\sqrt{\rho^2+4\sigma^2}\right)}\sqrt{\frac{\rho+\sqrt{\rho^2+4\sigma^2}}{\rho\sigma^2}}}$$

and this is the formula for survivability at equilibrium as a function of model parameters. Let us denote this function by $surv(\rho, \sigma, \eta)$. Its behaviour might be observed on Figures 3.4.2 and 3.4.3. Surprisingly, maximum survivability does not in general occur when the mutation rate is equal to the environmental change. Another surprising fact is that survivability *increases* with stringent selection.

Optimal mutation rate

Having computed equilibrium survivability as a function of the model parameters an important question one may ask is: what is the optimal mutation rate for a given speed of environmental change? One would expect that it should be equal to the speed of environmental change, however, as we can already see from Figure 3.4.2 this is not the case. In order to compute the optimal mutation rate for a given set of parameters it is necessary to compute the argument of the maximum of survivability function. In order to do that we must calculate $\frac{\partial}{\partial\rho}surv(\rho, \sigma, \eta)$. The result

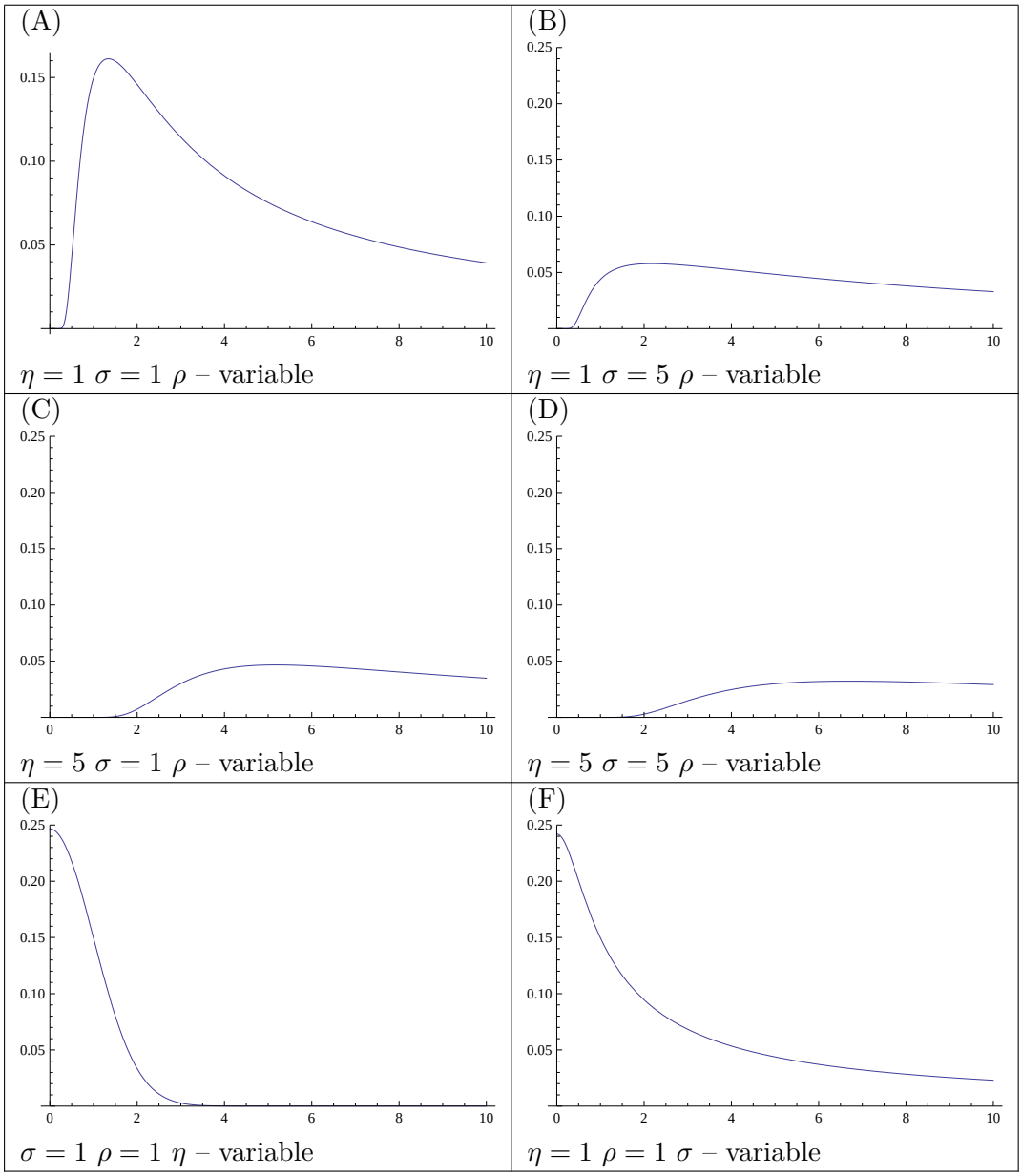


Figure 3.4.2: Plots of survivability function with various parameters fixed. **(A)** Survivability as a function of random mutation rate. A clear optimal random mutation rate is seen. **(B)** Reducing the strength of selection (by increasing the selection radius σ) causes a decrease in survivability, however, the optimal mutation rate does not appreciably change **(C)** Increasing the speed of environmental change reduces survivability, and increases the optimal mutation rate. **(D)** The effects of both lower selection and faster environmental change **(E)** Survivability as a function of environmental change. Highest survivability occurs in a stable environment. **(F)** Survivability as a function of selection radius. Highest survivability occurs with most stringent selection.

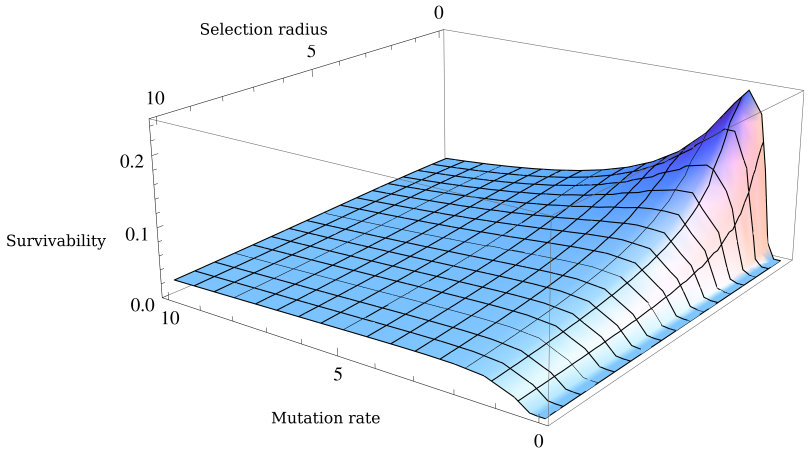


Figure 3.4.3: Plot of survivability function with $\eta = 1$ and varying mutation rates and selection radii.

is (computations, again, performed in Mathematica):

$$\frac{e^{-\frac{\eta^2}{2\rho^2}}\sqrt{\frac{2}{\pi}}\left(-\rho^3+\eta^2\sqrt{\rho^2+4\sigma^2}\right)}{\rho^3\sigma\sqrt{\rho^2+4\sigma^2}\sqrt{\rho\left(\rho+\sqrt{\rho^2+4\sigma^2}\right)}\sqrt{\frac{\rho+\sqrt{\rho^2+4\sigma^2}}{\rho\sigma^2}}}$$

In order to find roots of the above expressions it is enough to solve:

$$-\rho^3+\eta^2\sqrt{\rho^2+4\sigma^2}=0$$

The only solution in $\mathbb{R}^+$ turns out to be:

$$\frac{\sqrt{\left(18\eta^4\sigma^2+\sqrt{3}\sqrt{-\eta^8\left(\eta^4-108\sigma^4\right)}\right)^{1/3}+\frac{\eta^4}{\left(6\eta^4\sigma^2+\sqrt{-\frac{\eta^{12}}{3}+36\eta^8\sigma^4}\right)^{1/3}}}}{3^{1/3}}$$

Average fitness

The average fitness of the equilibrium population can be calculated as follows:

$$\int_{\mathbb{R}} \nu \left(\frac{\eta \sqrt{4\,\sigma^2 + \rho^2} - \eta\,\rho}{2\,\rho}, \frac{\sqrt{2\rho(\sqrt{4\sigma^2+\rho^2} - \rho)}}{2} \right) (x)\nu(0,\sigma)(x)\mathrm{d}\lambda(x) =$$

Now, let us use the formula for the product of two Gaussian PDFs. Note that, like previously, the resulting Gaussian PDF integrates out to 1, and the result is just the scaling constant:

$$= \frac{e^{\frac{\eta^2\left(\rho-\sqrt{\rho^2+4\sigma^2}\right)^2}{4\rho^2\left(\rho^2-2\sigma^2-\rho\sqrt{\rho^2+4\sigma^2}\right)}}}{\sqrt{\pi}\sqrt{-\rho^2+2\sigma^2+\rho\sqrt{\rho^2+4\sigma^2}}}$$

It is important to note that the optimal mutation rate for highest average fitness of the population differs from the optimal mutation rate for highest survivability. It is possible to symbolically compute the optimal mutation rate, again, by differentiating with respect to ρ and searching for roots, however, this time, the result (obtained with a CAS) has over 100 terms and, as such, is useless in terms of understanding the behaviour of the function, so the calculations will be skipped. Instead, it is (again) possible to visualize the behaviour of the function using plots (Figure 3.4.4 and 3.4.5). At a first glance it may come as a surprise that increasing the strength of selection (by decreasing its radius) actually arbitrarily increases the average fitness of the population,however, it is precisely this effect that has been used in plant and animal breeding.

3.4.6 A model for class II transposons

Introduction

Having studied in depth the mutator model, it turns out that actually a model for class II TEs is just a tiny step away. Recall that class II TEs (or DNA transposons) relocate through the genome using a cut-and-paste

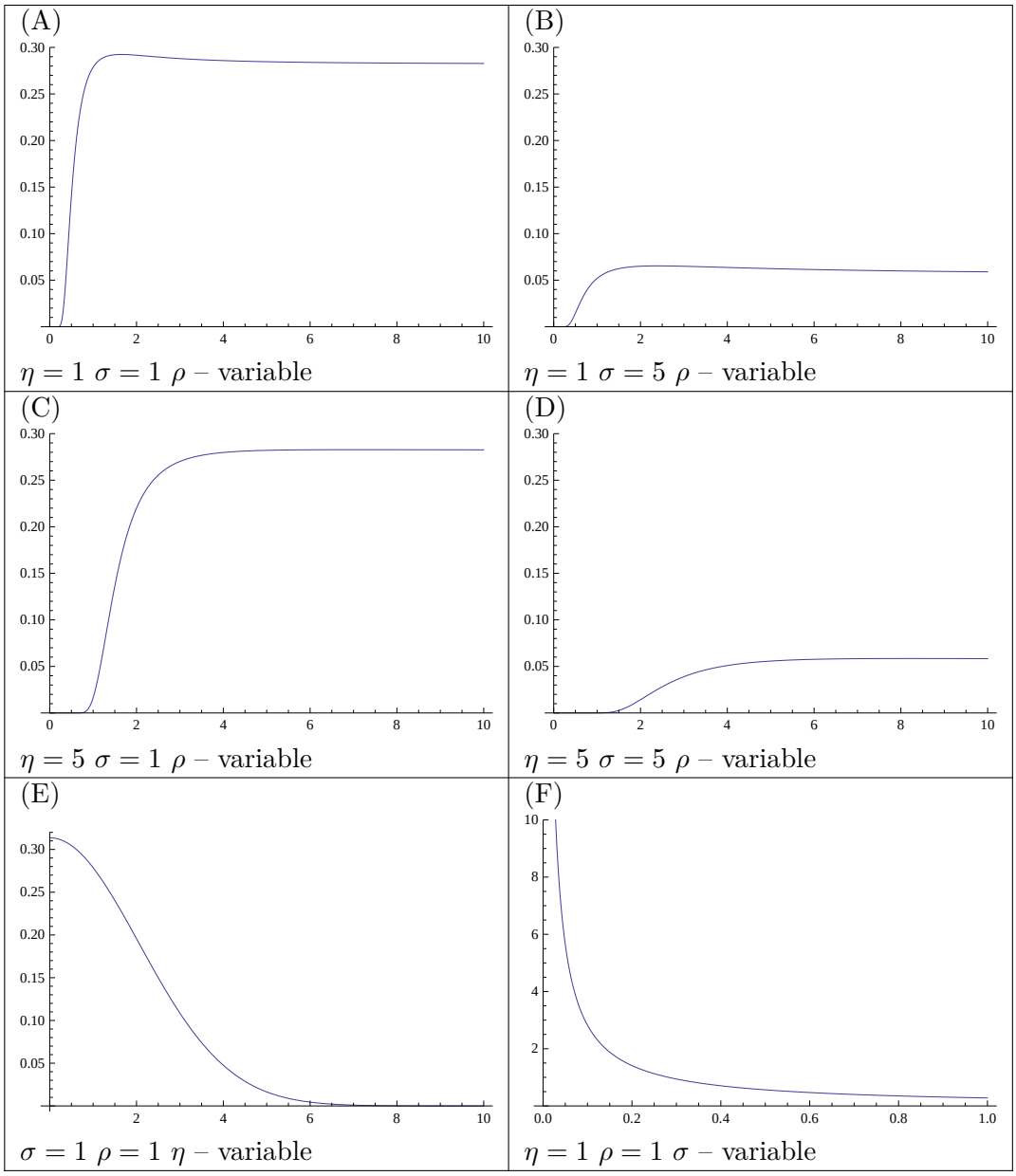


Figure 3.4.4: Plots of average fitness function with various parameters fixed. **(A)** Plot of average fitness as a function of random mutation rate. There is an evident optimal mutation rate, however, increasing mutation rate beyond it causes only a slight drop in average fitness. **(B)** Same plot, with less stringent selection (higher selection radius σ). Decreasing the strength of selection doesn't appreciably relocate the optimal mutation rate, however, it makes adaptation more difficult. **(C)** Increasing the speed of environment change increases the mutation rate needed to keep up with the environment, however, the average fitness does not significantly decrease. **(D)** Combined effects of high environment change and low selection. **(E)** Average fitness as a function of environment change, with fixed mutation rate: if we allow the speed of environmental change to increase without allowing the mutation rate to increase as well, the average fitness of population tends to zero. **(F)** Average fitness as a function of selection radius. With more stringent selection we can arbitrarily increase the population's fitness.

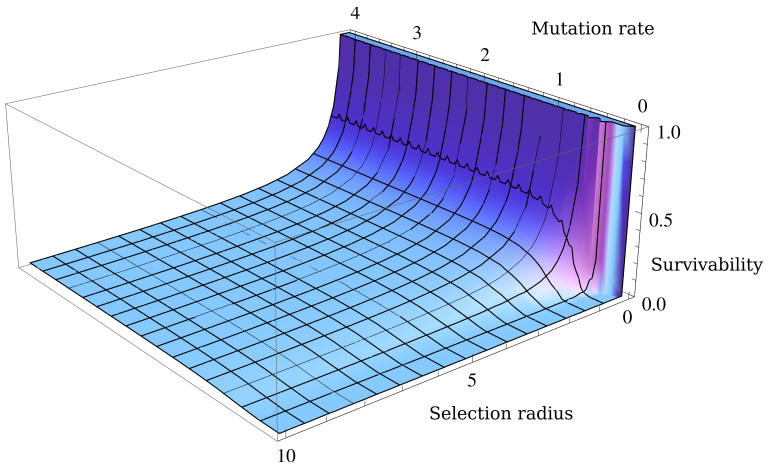


Figure 3.4.5: Plot of average fitness of the population as a function of mutation rate and selection radius, with $\eta = 1$.

mechanism. This has the following effects on the genome:

- Each transposon, by relocating, causes mutations to the genome of the host, for example through their potential to disrupt genes. Note that the damage is permanent: even after the transposon relocates to a new location, the gene containing it remains damaged: the transposon leaves behind a short *Target Site Duplication* (TSD) sequence – which remains in the gene, interrupting it, even after the transposon is gone. The additive mutations caused by transposons are dependant on the number of transposons in a linear fashion. Therefore, we shall assume that each organism in the population mutates with $\rho + t$ rate, where ρ is, as previously, the background mutation rate (rate of non-transposon-related mutations), while t is the number of transposons an organism is carrying. Note that we allow t to be in $\mathbb{R}_0^+$ – while transposons are discrete, their level of activity is certainly not. Therefore, transposons of lower or higher activity may be represented by numbers other than 1.

- The transposition does not affect the number of transposons in general case. While it is true that in certain circumstances the reconstruction of a break in DNA structure caused by a transposon being excised may sometimes recreate a copy of the transposon, even though the original one has already moved into a new locus, we do

not model this effect.

The model

Keeping this in mind, the change needed to turn a mutator model into a model for class II transposons is apparent: we should add another dimension to the model, denoting the number of transposons each organism carries. The population then, may be thought of as a sum of (a continuum of) subpopulations with varying transposon numbers, with each of the subpopulations behaving in accordance to the rules of mutator model, with mutation rate being $\rho + t$ instead of ρ. The organisms may not travel between the subpopulations, however, some subpopulations may proliferate at the expense of other, less adapted ones.

Let us denote by $f : [0, +\infty) \to (0, +\infty)$ the PDF of the distribution of transposons among the population. This time, for simplicity we shall assume that the distribution has a PDF, moreover, that the PDF is positive (that is, for every number, there is a nonzero proportion of organisms with this many transposons) and bounded.

Each of the same-number of transposons subpopulations behaves independently of others, in accordance to the mutator model, therefore, each one of them will converge to the equilibrium. As such, let us study only populations where each of the subpopulations is in its equilibrium state.

Therefore, the initial state of the population may be described by a probability distribution $\mathbb{P} : \mathrm{Leb}(\mathbb{R} \times [0, +\infty)) \to [0, +\infty]$ with the formula:

$$\mathbb{P}(A) =$$

$$= \int_A \nu \left(\frac{\eta \sqrt{4\,\sigma^2 + (\rho + t)^2} - \eta\,(\rho + t)}{2\,(\rho + t)}, \frac{\sqrt{2(\rho + t)(\sqrt{4\sigma^2 + (\rho + t)^2} - (\rho + t))}}{2} \right) (x) f(t) \mathrm{d}\lambda(x, t)$$

Convergence

It is easily observed that with subsequent generations only the $f(t)$ part will evolve (namely, proportions of organisms with different transposon numbers), while the Gaussian part, responsible for the mutator equilibrium shall remain constant. Because of that, let us drop it for now, and

focus purely on f. Selection in this model works as follows: in each subpopulation the selection works like in the mutator model, meanwhile, the scaling is done globally. This is equivalent to multiplying the function $f(t)$ by the survivability of population with mutation rate $\rho + t$ – which we have computed earlier, and then multiplying it by a scaling factor it so that $\int_{[0,+\infty)} f(t)\mathrm{d}\lambda(t) = 1$. Let us denote this transformation by ϕ:

$$\phi(f)(t) = \frac{s(t)f(t)}{\int_{[0,+\infty)} s(x)f(x)\mathrm{d}\lambda(x)}$$

where s is the survivability as a function of mutation rate, the same function as mentioned in subsection 3.4.5:

$$s(\rho) = \frac{e^{-\frac{\eta^2}{2\rho^2}}\sqrt{\frac{2}{\pi}}}{\sigma\sqrt{\rho\left(\rho+\sqrt{\rho^2+4\sigma^2}\right)}\sqrt{\frac{\rho+\sqrt{\rho^2+4\sigma^2}}{\rho\sigma^2}}}$$

With this, we may establish the equilibrium:

Theorem 8. For every bounded, positive PDF $f : [0, +\infty) \to (0, +\infty)$ the sequence of measures μ_n associated with PDFs $\phi^n(f)$ converges in weak-* fashion to the Dirac measure δ whose mass in concentrated at the point:

$$\frac{\sqrt{\left(18\eta^4\sigma^2+\sqrt{3}\sqrt{-\eta^8\left(\eta^4-108\sigma^4\right)}\right)^{1/3}+\frac{\eta^4}{\left(6\eta^4\sigma^2+\sqrt{-\frac{\eta^{12}}{3}+36\eta^8\sigma^4}\right)^{1/3}}}}{3^{1/3}}-\rho$$

if this is positive, or to δ_0 otherwise.

Proof. Note that:

$$\phi^n(f)(t) = \frac{(s(t))^n f(t)}{\int_{[0,+\infty)} (s(x))^n f(x)\mathrm{d}\lambda(x)}$$

Recall that fraction part of the point of mass of δ is the unique point in positive reals where the derivative (with respect to mutation rate) of survivability is zero, and also, the point at which the survivability function

attains its maximum. The mutation rate is now divided between random mutations and transposon-related mutations. If random mutations are already larger than this value, then the survivability function attains its maximum with 0 transposons, otherwise the reminder of the needed mutations must be provided by transposons. Let us denote the point of maximum (that is, the optimal number of transposons) by d. We shall tackle the case where $d > 0$, the case where $d = 0$ is similar (and easier, as it is only necessary to consider one side of the point d).

We shall prove weak-* convergence using the following condition:

Definition 3. Let X be a metric space with Borel (or in this case: Lebesgue, as non-Borel null sets don't play a role here) σ-algebra $\mathfrak{M}$. A sequence of measures μ_n converges in a weak-* manner to a measure μ if $\lim_{n\to\infty} \mu_n(A) = \mu(A)$ for all continuity sets A of the measure μ.

Let A be any continuity set of δ. In this case this means that $d \notin \partial A$. As such, there must exist an open neighbourhood $\mathcal{U}$ of d, which either lies entirely in A or is disjoint with A, and which contains an open ball centered on d, with radius $\epsilon > 0$.

Let $\hat{A}$ be equal to $[0, +\infty) \setminus (d - \epsilon, d + \epsilon)$.

Before we proceed, let us make some remarks about the function s.

Remark 1. As the derivative of the survivability function (with respect to the mutation rate) changes sign at d and only at d (and goes from positive to negative), it follows that either $s(d - \epsilon)$ or $s(d + \epsilon)$ is the function's minimum on $[d - \epsilon, d + \epsilon]$. Without loss of generality, let us assume that it attains minimum at $d - \epsilon$. Let $M = s(d + \epsilon)$ (that is the value on the end opposite of minimum). Because of the behaviour of the derivative it is evident that $M \geq s(a)$ for all $a \in \hat{A}$, and $\exists c_1, c_2$ such that $s([d+\epsilon/2]) \geq c_1 > c_2 \geq s(\hat{A})$[2]. The situation is sketched on Figure 3.4.6

Case 1. $\mathcal{U}$ is disjoint with A.

Note that in this case $A \subseteq \hat{A}$, and $d \notin A$, and as such, $\delta(A) = 0$. Now,

[2]The inequality $X > c$ should be interpreted that all elements of the set X are greater than c

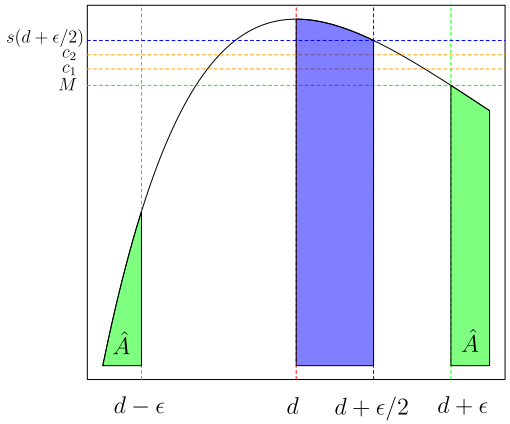


Figure 3.4.6: Plot of an example survivability function with notations introduced in Remark 1

in order to prove $\mu_n \to^* \delta$ we need to compute:

$$\mu_n(A) = \int_A \phi^n(f)(t)\mathrm{d}\lambda(t) \leq$$

$$\leq \int_{\hat{A}} \phi^n(f)(t)\mathrm{d}\lambda(t) = \int_{\hat{A}} \frac{(s(t))^n f(t)}{\int_{[0,+\infty)} (s(x))^n f(x)\mathrm{d}\lambda(x)}\mathrm{d}\lambda(t) \leq$$

$$\leq \int_{\hat{A}} \frac{c_1^n \cdot f(t)}{\int_{[0,+\infty)} (s(x))^n f(x)\mathrm{d}\lambda(x)}\mathrm{d}\lambda(t) = \frac{c_1^n \int_{\hat{A}} f(t)\mathrm{d}\lambda(t)}{\int_{[0,+\infty)} (s(x))^n f(x)\mathrm{d}\lambda(x)} \leq$$

since f is a PDF of a probability distribution, and as such, its integral over any set is ≤ 1:

$$\leq \frac{c_1^n}{\int_{[0,+\infty)} (s(x))^n f(x)\mathrm{d}\lambda(x)} \leq \frac{c_1^n}{\int_{[d,d+\epsilon/2]} (s(x))^n f(x)\mathrm{d}\lambda(x)} \leq$$

$$\leq \frac{c_1^n}{\int_{[d,d+\epsilon/2]} c_2^n f(x)\mathrm{d}\lambda(x)} =$$

Since f is measurable and positive its integral over any nontrivial segment is also positive. Let us denote this value as k.

$$= \frac{c_1^n}{c_2^n \cdot k} \xrightarrow[n\to\infty]{} 0$$

as $c_2 > c_1$. As the limit is obviously bounded by 0 also from below, then, by squeeze theorem (Bartle and Sherbert (2011) p. 64) we arrive at the

conclusion that $\lim_{n\to\infty}\mu_n(A)=0=\delta(A)$.

Case 2. $\mathcal{U}\subseteq A$

In this case, $d\in A$, thus $\delta(A)=1$, and $[d-\epsilon,d+\epsilon]\subseteq A$.

Again, let us compute:

$$\mu_n(A)\geq\mu_n([d-\epsilon,d+\epsilon])=\mu_n(\mathbb{R}^{+,0})-\mu_n([0,d-\epsilon)\cup(d+\epsilon,+\infty))=$$

$$=1-\int_{[0,d-\epsilon)\cup(d+\epsilon,+\infty)}\frac{(s(t))^n f(t)}{\int_{[0,+\infty)}(s(x))^n f(x)\mathrm{d}\lambda(x)}\mathrm{d}\lambda(t)\geq$$

$$\geq 1-\frac{\int_{[0,d-\epsilon)\cup(d+\epsilon,+\infty)}c_1^n f(t)\mathrm{d}\lambda(t)}{\int_{[0,+\infty)}c_2^n f(x)\mathrm{d}\lambda(x)}=1-\frac{c_1^n}{c_2^n}\cdot\frac{\int_{[0,d-\epsilon)\cup(d+\epsilon,+\infty)}f(t)\mathrm{d}\lambda(t)}{\int_{[0,+\infty)}f(x)\mathrm{d}\lambda(x)}=$$

$$=1-\frac{c_1^n}{c_2^n}\cdot k\xrightarrow[n\to\infty]{}1$$

As each of μ_n is a probability measure, therefore $\mu_n(A)\leq 1$, so, by squeeze theorem (Bartle and Sherbert (2011) p.64) we conclude that $\lim_{n\to\infty}\mu_n(A)=1=\delta(A)$.

□

This establishes the following equilibrium population:

$$\mathbb{P}(A)=\int_{\hat{A}}\nu\left(\frac{\eta\sqrt{4\,\sigma^2+(\rho+t)^2}-\eta\,(\rho+t)}{2\,(\rho+t)},\frac{\sqrt{2(\rho+t)(\sqrt{4\sigma^2+(\rho+t)^2}-(\rho+t))}}{2}\right)(x)\mathrm{d}\lambda x$$

where:

$$\hat{A}=\{x\in\mathbb{R}\mid\left(x,\frac{\sqrt{\left(18\eta^4\sigma^2+\sqrt{3}\sqrt{-\eta^8\,(\eta^4-108\sigma^4)}\right)^{1/3}+\frac{\eta^4}{\left(6\eta^4\sigma^2+\sqrt{-\frac{\eta^{12}}{3}+36\eta^8\sigma^4}\right)^{1/3}}}}{3^{1/3}}-\rho\right)\in A\}$$

if:

$$\frac{\sqrt{\left(18\eta^4\sigma^2+\sqrt{3}\sqrt{-\eta^8\,(\eta^4-108\sigma^4)}\right)^{1/3}+\frac{\eta^4}{\left(6\eta^4\sigma^2+\sqrt{-\frac{\eta^{12}}{3}+36\eta^8\sigma^4}\right)^{1/3}}}}{3^{1/3}}-\rho>0$$

and:

$$\hat{A} = \{x \in \mathbb{R} \mid (x, 0) \in A\}$$

otherwise.

3.4.7 Concluding remarks and interpretation

The main setting of the presented model is, as mentioned, the modelling of class-II TEs. The speed of mutations is directly correlated to the number of class-II TEs a given organism carries, because that is proportional to the number of transpositions. The transpositions (in this simplified approach) operate only in a cut-and-paste mode, and so, do not cause the number of transposons to change. However, it is easy to notice that the model, as presented here, actually is not limited to describing transposon-carrying populations: indeed, it may be used to study the behaviour of any population of organisms with variable mutation rates, even if the reason for that has nothing to do with transposons. In such a case the transposon activity parameter is replaced with speed of mutations of a given organism, and all the equations and the theory carry over.

In particular, one application brings us back to class-I TEs: our work, described in the next chapter, proves that the presence of even inactive TEs in the genome has mutagenic effects on the host, by causing the self-similarity of the genome, and as such, enabling the Nonallelic Homologous Recombination events. As such, the inactive TEs (both class-I and class-II) have a fixed mutational effect on the host genome, which is proportional the the square of their content.

4

Analysis of TEs in human genome as contributors to genomic disease through nonallelic homologous recombination

The computational model of TE interaction from Chapter 1 assumed that TEs may only cause mutations through their active transposition in the genome, by disrupting or misregulating genes. However, recently, it has come to the attention of biologists that this may not be the only known mechanism by which TEs might disrupt the genome. In this chapter we will attempt to gauge and estimate the significance of one such method, the nonallelic homologous recombination (NAHR).

This work was done in conjunction with dr Paweł Stankiewicz's team at Baylor College of Medicine (Houston, USA), who suggested this direction of research. In addition to the impact on modelling, the research we have performed here has serious clinical implications, (they were the motivation of the

BCM team), which, too, shall be presented.

4.1 Motivations and related research

Copy-number variation (CNV) contributes significantly both to human genetic variation as well as disease (Sebat et al., 2004; Iafrate et al., 2004; Stankiewicz and Lupski, 2010). NAHR, occurring during meiosis, is the most common mechanism underlying the formation of recurrent CNVs in humans (Gu et al., 2008; Chen, 2012). The product of NAHR can be deletion or reciprocal duplication, as well as inversions and inter- or intrachromosomal translocation (Stankiewicz and Lupski, 2002; Dittwald et al., 2013). In the vast majority of rearrangements characterized thus far, NAHR occurs between segments of the human genome that are present in more than one copy known as low-copy repeats (LCRs or segmental duplications). These LCRs are over 10 kbp in size and share more than 97% DNA sequence identity (Ou et al., 2011; Dittwald et al., 2013; Gu et al., 2008; Stankiewicz and Lupski, 2002). However, in addition to these classically defined LCRs, other sequences have been observed to mediate apparent NAHR events. Specifically, rearrangements mediated by Human Endogenous Retroviruses (HERVs), a small subfamily of long retrotransposons comprising about 5% of the human genome (Shuvarikov et al., 2013), suggest that the lower boundary on the length of the homologous region which is capable of mediating NAHRs might be as low as few kbp. This means that mobile DNA elements (McClintock, 1950) such as TEs may be potential substrates for NAHR. If true, this would suggest that a significantly higher fraction of the human genome is susceptible to NAHR mediated rearrangements, as TEs make up as much as 44% of the reference human genome (Mills et al., 2007).

Since their initial discovery, studies have indicated that the presence of TEs, and particularly active TEs, has mutagenic effects on the genome of their host. The most frequently cited effects of TEs is their capability of disrupting a gene by their insertion and their ability to upregulate genes by inserting nearby owing to TE-borne enhancers (Walisko et al., 2008). This mutational activity of TEs is significant enough that it is being selected against over the course of evolution (Petrov et al., 2011). Computational modeling approaches presented in the second chapter of this work, as well as the mathematical model from Chapter 3, suggest that, in certain circumstances, the activity of TEs may be beneficial to the population by assisting the adaptation of the population to a

new environment.

Another distinct class of the mutagenic effects of TEs, perhaps comparable in scale to that of *de novo* insertions (Kidd et al., 2010) is that caused by their high self-similarity coupled with abundance in the genome. These features create a large number of non-allelic, homologous sites in the genome that can mediate recombination events (Robberecht et al., 2013; Beck et al., 2011). Cases of LINE-LINE/NAHR have been reported previously (Burwinkel and Kilimann, 1998), and indeed some of them linked to disease (Temtamy et al., 2008; Szafranski et al., 2013; Belancio et al., 2009; Higashimoto et al., 2013). Moreover, previous studies of genomic architectural features that stimulate and potentially catalyze pathogenic microdeletions and tandem duplications have found that repetitive elements are enriched at deletion breakpoints (Vissers et al., 2009). Interestingly, comparative analysis of human and chimpanzee genomes (Han et al., 2008) identified (and verified by wet-lab analyses) 73 human specific LINE recombination-associated deletion (55 of them have been classified as NAHR events).

Here, we analyzed the contribution of LINE retrotransposons to NAHR in humans, a much more abundant family of TEs than HERVs. We also performed a comprehensive analysis of the human genome susceptibility to LINE-mediated NAHR deletions/duplications, inversions and translocations using the Baylor College of Medicine clinical database of CNVs. We found that LINE-LINE-mediated NAHR occurs more frequently than previously thought and indeed on a genome-wide scale. We estimate that each healthy individual carries on average three different LINE-mediated NAHR CNVs. Finally, we provide several novel bioinformatic procedures and algorithms for the study of NAHR. This means that the effect needs to be accounted for in models – and, in particular, that the mathematical model presented for class-II transposons may be also used to study the inactive class-I TEs, as indicated in the concluding remarks of the previous chapter.

4.2 Materials and methods

4.2.1 Identification of LINE pairs able to mediate NAHR

We downloaded the reference sequence from the hg19 assembly of the human genome along with coordinates of all LINE elements as denoted by the UCSC *RepeatMasker* track (Tarailo-Graovac and Chen, 2009). Locations of cen-

tromeres were also obtained from the *Gap* UCSC (International Human Genome Sequencing Consortium, 2001; Kent et al., 2002) genome browser track. The sequences of 124,150 LINE elements longer than 1 kbp were extracted, along with 3 kbp flanking sequence. These sequences were then pairwise-aligned using the BLAST algorithm (Altschul et al., 1990) with the low-complexity sequence masking disabled. We have obtained 3,642,718,496 statistically significant High Scoring Segment Pairs (HSPs) with the E-value computed by BLAST being less than 10^{-50}, these were further filtered to eliminate self-alignments. Moreover, we removed cases in which the alignment extended outside the LINE into the flanking regions, suggesting that the duplicated sequence may not be the result of a LINE transposition, but rather that the LINE element was a part of a larger LCR. We also excluded alignments shorter than 1000 bp, those with the identity less than 92%, and pairs that would result in intrachromosomal CNVs greater than 10 Mbp (as such CNVs are unlikely to be observed in a living organism). Alignments were classified into types (deletion, duplication, inversion, or translocation) based on whether the matching LINE pairs map to the same chromosome, on their respective orientation, and on whether the potential NAHR event spans a centromere. Pairs of LINEs on different chromosomes, or on the same chromosome but on the different sides of the centromere were marked as potential translocation substrates. The reminder (pairs mapping on the same chromosome and on the same side of the centromere) were marked as either deletion/duplication substrates (if directly oriented), or inversion substrates (otherwise).

We subsequently intersected the directly-oriented LINE-LINE pairs with our clinical database of CNVs. This database consists of 398,468 CNVs that were identified in 36,285 patients undergoing oligonucleotide chromosomal microarray analysis (CMA) at Medical Genetics Laboratories (MGL) at Baylor College of Medicine (BCM) and were determined to be pathogenic or potentially pathogenic by they clinical cytogeneticist reviewing the case. All DNA samples were anonymised for further study, and no clinical information is unavailable. The precise breakpoint of each CNV was unknown; we narrowed their putative locations to the regions between two adjacent oligo probes showing definitive difference in $\log_2$ ratio. We refer to the intervals as uncertainty regions (cf. Fig. 4.2.1).

The set of CNVs was then filtered to exclude the cases where the uncertain regions at both ends of the CNV were located in or contained directly-oriented

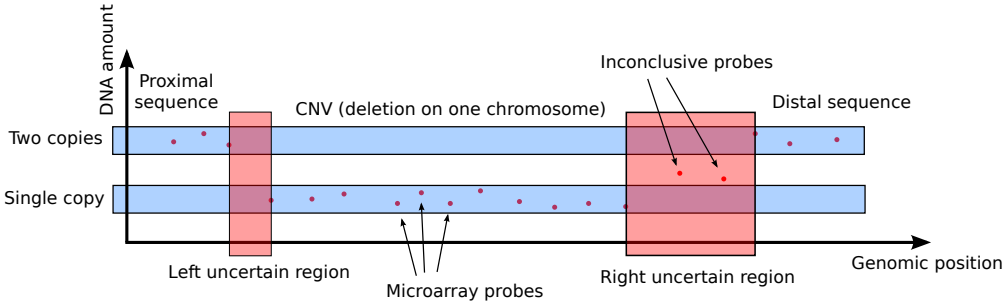


Figure 4.2.1: The schematic representation used to define the *uncertain regions* for breakpoint analyses. The proximal NAHR breakpoint maps within the left red area, the distal one in the right red area. A similar approach was used for duplications.

paralogous LCRs (DP-LCRs). We assumed that in those cases the NAHR might have been mediated by these DP-LCRs, rather than LINEs specifically. After DP-LCR filtering, 358,160 CNVs remained and were compared to the database of possible LINE-LINE/NAHR pairs computed previously. The analysis yielded 112,520 potential CNVs. The parameters were further constrained to include only full- and nearly full-length LINEs (longer than 4 kbp, and aligning over more than 4 kbp of their length) with more than 96% sequence identity.

4.2.2 Clinical CMA

DNA was prepared from peripheral blood using the Puregene DNA isolation kit (Gentra Systems, Minneapolis, MN, USA) according to the manufacturers instructions. CMA was performed with gender-matched controls; labeling, hybridization and scanning procedures as well as computational analysis have been described previously described (Boone et al., 2013). Briefly, BCM MGL oligonucleotide arrays contain both genome-wide backbone probe coverage and enhanced probe resolution within the exons and introns of manually curated known and putative disease genes.

4.2.3 Subjects

Deidentified DNA samples from 44 individuals harboring potentially LINE-LINE/NAHR CNVs, 21 deletions and 23 duplications, from five different genomic regions (Fig. 4.2.2) were obtained from unrelated subjects identified by CMA (CMA oligonucleotide versions V7.1, V7.2, V7.4, V7.6, V8.1, V8.3, V9.1) (Boone et al., 2010; Wiszniewska et al., 2014). Additionally, DNA sam-

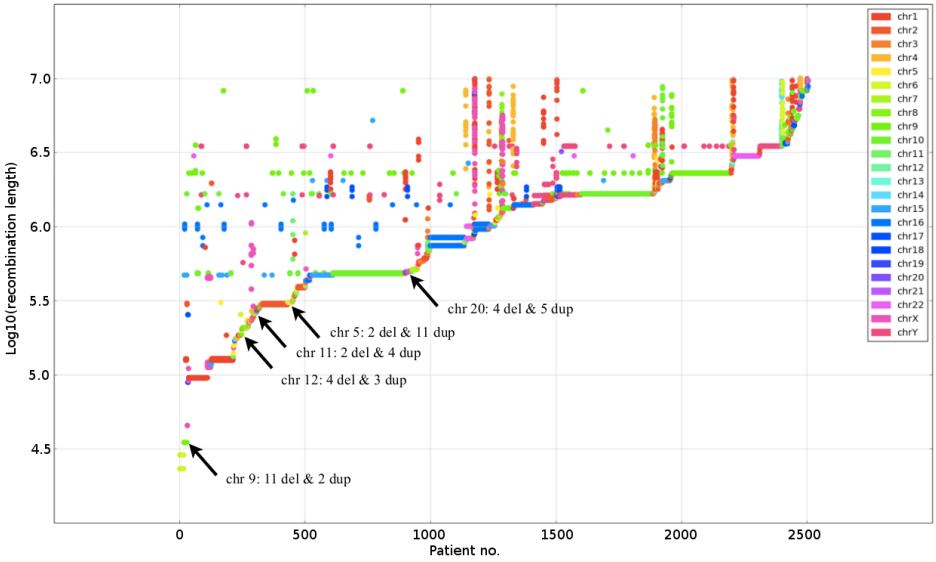


Figure 4.2.2: Scatterplot of patients with CNVs, where a pair of matching LINEs lies in *uncertain regions* for the ends of the CNV. Only cases where the alignment between LINEs is longer than 4 bp with over 96% identity are shown. Patients are sorted by the length of their shortest CNV; one patient may possess multiple CNVs. Cases selected for PCR confirmation are highlighted with arrows.

ples were obtained from six healthy individuals. These samples were obtained following informed consent (BCM IRB protocol H33409).

4.2.4 Long Range PCR and DNA sequencing

Long-range PCR (LR-PCR) primers flanking LINE elements were automatically designed using custom software including code from Primer3 (`http://primer3.sourceforge.net/`) (Untergasser et al., 2012). The program automatically generates a hybrid LINE sequence (assuming that the breakpoint maps within LINE-LINE homology region) along with unique flanking sequence for all possible NAHR rearrangements (deletion, duplication, inversion, or translocation) (Fig. 4.2.3). These hybrid sequences were then analyzed by a custom Primer3 script to obtain LR-PCR primers.

LR-PCR amplification of 7-15 kbp fragments was performed using LA *Taq* Polymerase (TaKaRa Bio USA, Madison, WI, USA) following the manufacturer's protocol. Briefly, we used $25\mu l$ reaction mixtures containing $100ng$ genomic DNA, $0.4mM$ dNTPs, $0.2\mu M$ of each primer, and $1.25U$ of LA Taq polymerase mix. PCR conditions were: 94°C for 1 minute, followed by 30 cycles

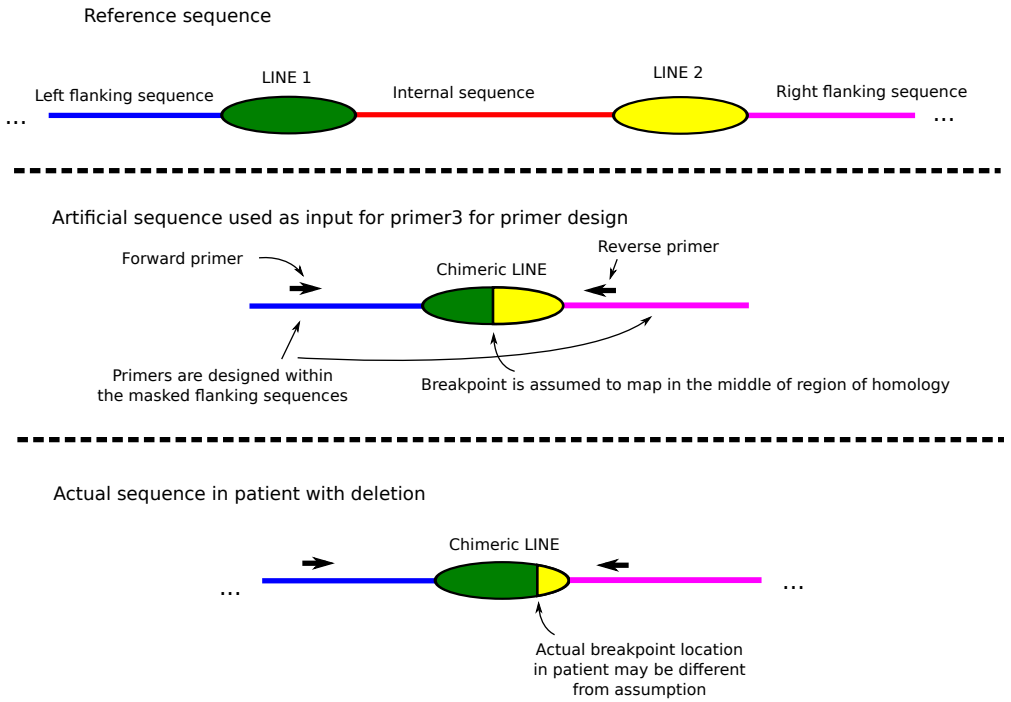


Figure 4.2.3: Artificial sequences computed for primer design for detection of chimeric LINE sequences. Shown on figure is the process for deletion, with duplications and inversions being handled in similar fashion.

at 94°C for 30 seconds and 68°C for 12 minutes, and 72°C for 10 min. PCR products were treated with ExoSAP-IT (USB, Cleveland, OH, USA) to remove unconsumed dNTPs and inactivate primers. The treated amplicons were then sequenced by the Sanger method (Lone Star Labs, Houston, TX, USA) using the initial primers and primers specific for both unmasked proximal and distal copies of the LINEs.

4.2.5 LR-PCR analysis of healthy subjects

In addition, from the set of computationally predicted LINE-LINE flanked loci, we selected 95 directly-oriented pairs with high homology parameters for studies of deletions, 95 for duplications, and 95 inverted pairs for study of inversions. We have designed PCR primers as described above (one set of 95 primer pairs for deletions, one set for duplications, and one for inversions). The LR-PCR reaction (with the same parameters as described earlier) was ran on a mixture of DNA (8x50mg) from 8 donors not known to suffer from genetic disease. The existence of CNV was confirmed by visualizing a band during gel electrophoresis, and comparing the expected amplicon length (computed during design of

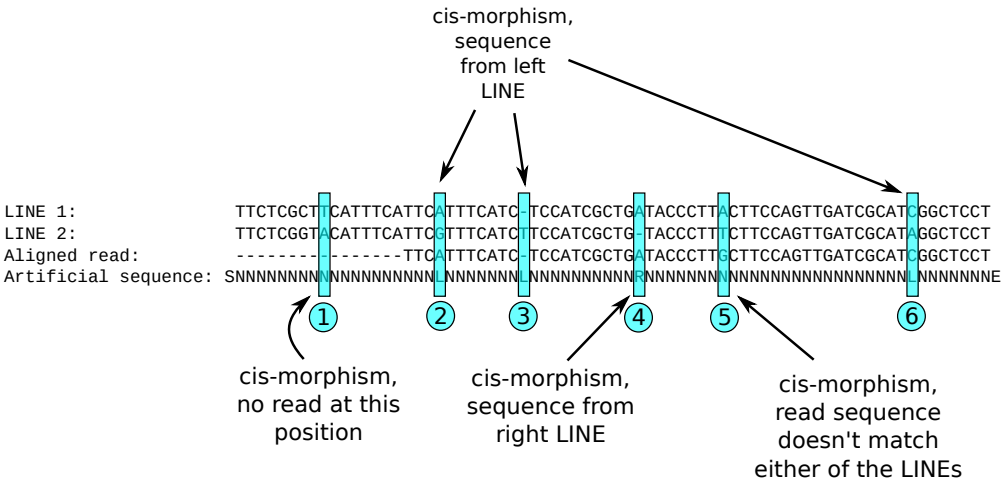


Figure 4.2.4: Construction of input sequence for estimation of NAHR breakpoint location. In artificial sequence, the S and E are special markers, for beginning and end of the sequence, L means that the observed sequence seems to come from the left (first) LINE, R means it comes from the right (second) one, N means that the source LINE cannot be determined from this location.

primers) to the actual length of the DNA band.

4.2.6 Array CGH analysis of healthy subjects

Genomic CNVs were analyzed using a custom-designed genome-wide LINE-LINE-targeted CGH 4x180 K microarrays (Agilent Technologies, Santa Clara, CA). The arrays were designed using a set of custom scripts, written in Python programming language. All probes were selected from the database of 26 million Agilent high density oligonucleotide probes. In addition to backbone probes used for calibration, each LINE from the set of directly oriented LINE-LINE/NAHR pairs was flanked with five oligonucleotide probes on each side to detect CNVs with both breakpoints mapping within LINE elements. For each array, one healthy individual was labeled with Cy3 and different, sex-matched healthy individual was labeled with Cy5. The labeling and hybridization procedures were performed according to manufacturer's protocols (Agilent Technologies, Santa Clara, CA).

4.2.7 DNA sequence analysis

Genomic sequences defined by coordinates identified in the array CGH experiments, were downloaded from the UCSC genome browser (NCBI build

37, May 2009, `http://www.genome.ucsc.edu`) and assembled using the Sequencher v4.8 software (Gene Codes, Ann Arbor, MI, USA). Interspersed repeat sequences were identified using RepeatMasker UCSC track (`http://www.repeatmasker.org`).

4.2.8 Preparation of sequences for analysis

For each pair of LINEs, a consensus sequence was computed (replacing mismatches with N), and a custom version of the Needleman-Wunsch algorithm (Needleman and Wunsch, 1970) modified to compute a semi-global alignment (See Algorithm 1) was used to align the Sanger reads to the consensus. An artificial sequence containing the information about sequence *cis*-morphisms was computed for each case (Fig. 4.2.4).

4.2.9 Hidden Markov model for breakpoints identification

Next, we define a Hidden Markov Model which will be used for the analysis of the sequences.

It is possible to think of a Hidden Markov Model as if it was a Markov chain in which the current state is not directly observable, instead the chain emits output symbols, with different probabilities in each state, and only these may be observed. The only way to infer the state of the Markov Chain is by observing the emitted sequence of symbols. Formally, let $\mathcal{S}$ be a finite set called the set of hidden states (we shall only deal with finite HMMs with discrete time steps). A finite Markov Chain with discrete time is a sequence of random variables $X_1, X_2, ...$ with a so-called *Markov property*, that is: $\mathbb{P}(X_{n+1} = x|X_1 = x_1, X_2 = x_2, ..., X_n = x_n) = \mathbb{P}(X_{n+1} = x|X_n = x_n)$. Additionally, we can say that a Markov chain is *time-homogeneous* if $\mathbb{P}(X_{n+1} = x|X_n = y) = \mathbb{P}(X_n = x|X_{n-1} = y)$ for all n.

A finite time-homogeneous Markov Chain with discrete time can be represented as a graph with edge weights: let $\mathcal{T} = \{(s_1, s_2) \in \mathcal{S} \times \mathcal{S} \mid \mathbb{P}(X_{n+1} = s_2|X_n = s_1) > 0\}$ be the set of transitions, and let $t : \mathcal{T} \to (0, 1]$ be a probability transition function defined as: $t((s1, s2)) = \mathbb{P}(X_{n+1} = s_2|X_n = s_1)$. Note that a sufficient condition for $(\mathcal{S}, \mathcal{T}, t)$ to be a representation of a Markov Chain is:

$$\forall s \in \mathcal{S} \mid \sum_{\substack{s' \in \mathcal{S} \\ (s,s') \in \mathcal{T}}} t(s, s') = 1$$

Algorithm 1: A modified version of Needleman-Wunsch algorithm for semiglobal alignment of reads to the consensus LINE sequence

Data: $LINE\_consensus$, $read\_sequence$,
$scores$ - a table of PHRED quality scores for read sequence
$similarity\_score$ - a scoring function based on substitution matrix
Result: An optimal semiglobal alignment of $read\_sequence$ to $LINE\_consensus$

$gap\_penalty \leftarrow -3.0$;
$edge\_gap\_penalty \leftarrow -0.1$;
for $i \in \{0..\text{len}(LINE\_consensus)\}$ **do**
  $T[i][0] \leftarrow (i * edge\_gap\_penalty, 'RightGap')$
end
for $i \in \{0..\text{len}(read\_sequence)\}$ **do**
  $T[0][i] \leftarrow (i * edge\_gap\_penalty, 'LeftGap')$
end
for $i \in \{0..\text{len}(LINE\_consensus)\}$ **do**
  for $i \in \{0..\text{len}(read\_sequence)\}$ **do**
    if $i = \text{len}(read\_sequence)$ **then**
      $LeftGapScore \leftarrow$
      $edge\_gap\_penalty * scores[j-1] + T[i][j-1][0]$
    else
      $LeftGapScore \leftarrow gap\_penalty * scores[j-1] + T[i][j-1][0]$
    end
    if $j = \text{len}(LINE\_consensus)$ **then**
      $RightGapScore \leftarrow$
      $edge\_gap\_penalty * scores[j-1] + T[i-1][j][0]$
    else
      $RightGapScore \leftarrow gap\_penalty * scores[j-1] + T[i-1][j][0]$
    end
    $MatchScore \leftarrow similarity\_score(seq1[i-1], seq2[j-1]) * scores[j-1] + T[i-1][j-1][0]$ **if** $MatchScore = max(RightGapScore, LeftGapScore, MatchScore$ **then**
      $T[i][j] = (MatchScore, 'MATCH')$
    else
      if $RightGapScore = max(RightGapScore, LeftGapScore, MatchScore$ **then**
        $T[i][j] = (RightGapScore, 'RIGHT')$
      else
        $T[i][j] = (LeftGapScore, 'LEFT')$
      end
    end
  end
end
Recover alignment from T as in standard Needleman-Wunsch algorithm;

From now on we shall identify a Markov chain with its graph representation $(\mathcal{S}, \mathcal{T}, t)$.

Next, let Σ be a finite set called the set of *emissions* (also called the *input alphabet*. Let $\mathcal{E} \subseteq \mathcal{S} \times \Sigma$, and let $e : \mathcal{E} \to (0, 1]$ be a function such that:

$$\forall s \in \mathcal{S} \quad \sum_{\substack{s' \in \Sigma \\ (s,s') \in \mathcal{E}}} e(s, s') = 1$$

Under all these assumptions, $(\mathcal{S}, \mathcal{T}, t, \Sigma, \mathcal{E}, e)$ is a Hidden Markov Model.

The formalism of Hidden Markov Model reflects a system with unknown internal state in which the internal state is governed by the rules of a Markov Chain. The system produces observable output dependent on the unknown internal state. A commonly cited example is the unfair casino, with a game of coin tossing. The casino has two coins, one fair and one biased in favour of heads. The casino, from time to time (with low probability) swaps the coin being used for the unfair one, and again from time to time (again, with low probability) swaps it back. A customer coming to the casino does not know which coin is currently being used by the casino, nor the points of time when it is swapped, all he can observe is the sequence of coin toss results.

This is modelled by a Hidden Markov Model with 2 hidden states, fair and biased coin: $(\mathcal{S} = \{F, B\})$ and an emission set of heads and tails: $\Sigma = \{H, T\}$. All transitions are possible $(\mathcal{T} = \mathcal{S} \times \mathcal{S})$, the casino keeps the coin currently being used with probability $0.9 = t(F, F) = t(B, B)$, and with probability 0.1 swaps the coin for the other: $0.1 = t(B, F) = t(F, B)$.

When the fair coin is in use the probability of getting heads is equal to the probability of getting tails: $e(F, H) = e(F, T) = 0.5$, unfair coin favours heads: $e(B, H) = 0.6$ and $e(B, T) = 0.4$.

A nice if quite expected property of Hidden Markov Models is that for a given sequence of observations from Σ^* it is possible to generate a sequence of hidden states, called a *Viterbi path* which maximizes the probability of generating the observed sequence – in our example this translates to the customer of the casino being able to deduce when are the most likely points when the coin was swapped, just by observing the sequence of coin toss results, assuming one can guess the bias of the unfair coin and the likelihood of coins being swapped. This is done using the Viterbi algorithm (Viterbi, 1967).

However, there is an even nicer and very unexpected fact: given only $\mathcal{S}, \mathcal{T}, \Sigma, \mathcal{E}$ and a set of sequences of observations (or: *even without $\mathcal{T}$ and $\mathcal{E}$!* - then it is

assumed that $\mathcal{T} = \mathcal{S} \times \mathcal{S}$ and $\mathcal{E} = \mathcal{S} \times \Sigma$) it is possible to estimate the functions e and t. The customer of the casino can even estimate the bias of the coins being used, and the likelihood of them being swapped, just by looking at the sequence of the result (if the sequence is long enough). This is done using the Baum-Welch algorithm (Welch, 2003).

This is a potent tool for analysis of biological sequences: in our case the unknown variable represented by the hidden states, which we would like to deduce, is whether at a given point in the sequence, we are before the breakpoint between LINEs, or after it.

In our case the HMM has 4 hidden states: $\mathcal{S} = \{S_0, S_1, ..., S_3\}$, the input alphabet is $\Sigma = \{S, N, L, R, E\}$, and the structure of the HMM is shown on Figure 4.2.5. The states S_0 and S_3 are added for technical reasons (the starting and ending state), the state S_1 represents being before the breakpoint, and S_2 represents being after the breakpoint. The input sequence consists of the letters S, N, R, L and E and is constructed as shown (see Fig. 4.2.4)

Then, the sequences were analyzed with a Hidden Markov Model (Eddy, 2004) trained using a custom version of the Baum-Welch algorithm (Welch, 2003).

The modified algorithm (Algorithm 2) differs from the standard version in that it enforced the following constraints during training:

- $\mathbb{P}(S_1 \to S_2) = \mathbb{P}(S_2 \to S_3)$: ensures the model does not favour placement of breakpoints near the beginning or end of alignments because the training data happens to be skewed as such

- $\mathbb{P}(S_1 \text{ emits } N) = \mathbb{P}(S_2 \text{ emits } N)$, $\mathbb{P}(S_1 \text{ emits } L) = \mathbb{P}(S_2 \text{ emits } R)$, $\mathbb{P}(S_1 \text{ emits } R) = \mathbb{P}(S_2 \text{ emits } L)$: assumes that SNVs with respect to the reference sequence, which would make the source LINE ambiguous (such as Fig. 4.2.4, location 5), or even suggest the wrong LINE (location 6) are equally likely to occur on either side of the breakpoint.

The model with parameters obtained from the Baum-Welch algorithm were then used to compute the posterior probabilities of transition from the S_1 state to S_2 at all locations, which correspond to the probability that the NAHR cross-over event having occurred at each location. These were computed using a custom version of the forward-backward algorithm (Lawrence R. Rabiner,

Algorithm 2: A modified version of Baum-Welch algorithm with constraints

Data: $\mathbb{P}$ - initial HMM parameters, as given in Table 4.2.1
Result: HMM parameters maximizing the probability of observations while adhering to the constraints

while not converged **do**
  $\mathbb{P}^* \leftarrow \text{Single\_iteration\_of\_Baum} - \text{Welch}(\mathbb{P})$;
  $\mathbb{P}^*(S_1 \to S_2), \mathbb{P}^*(S_2 \to S_3) \leftarrow \frac{\mathbb{P}^*(S_1 \to S_2) + \mathbb{P}^*(S_2 \to S_3)}{2}$;
  $\mathbb{P}^*(S_1 \text{ emits } N), \mathbb{P}^*(S_2 \text{ emits } N) \leftarrow \frac{\mathbb{P}^*(S_1 \text{ emits } N) + \mathbb{P}^*(S_2 \text{ emits } N)}{2}$;
  $\mathbb{P}^*(S_1 \text{ emits } L), \mathbb{P}^*(S_2 \text{ emits } R) \leftarrow \frac{\mathbb{P}^*(S_1 \text{ emits } L) + \mathbb{P}^*(S_2 \text{ emits } R)}{2}$;
  $\mathbb{P}^*(S_1 \text{ emits } R), \mathbb{P}^*(S_2 \text{ emits } L) \leftarrow \frac{\mathbb{P}^*(S_1 \text{ emits } R) + \mathbb{P}^*(S_2 \text{ emits } L)}{2}$;
  $\mathbb{P} \leftarrow \mathbb{P}^*$;
end
Output $\mathbb{P}$;

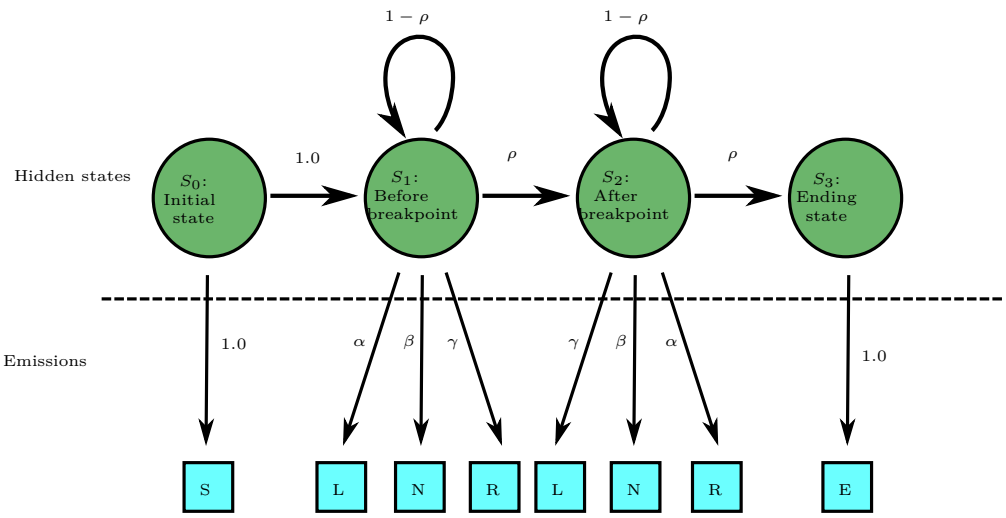


Figure 4.2.5: Hidden Markov model used for estimation of breakpoint location. The NAHR site maps at the point of $S_1 \to S_2$ transition. The prior and posterior values of $\alpha, \beta, \gamma, \rho$ can be found in Table 4.2.1.

Parameter name	Prior value	Posterior value
α	0.1	0.0092489671329
β	0.89	0.9899202188834
γ	0.01	0.0008308139835
ρ	0.05	0.0003545644262

Table 4.2.1: Table of HMM parameters used. Parameter names from Figure 4.2.5

1986), in which the observation matrices corresponding to the L and R emissions were replaced with an affine combination of matrices for L and R with weights based on the PHRED quality score (Ewing et al., 1998; Ewing and Green, 1998) of the sequence from which the L or R signals originated. The posterior probabilities were calculated, and in most cases a single location of the breakpoint was obtained. The computed locations were later confirmed by visual inspection using the Sequencher software `http://www.genecodes.com/`.

4.2.10 Enrichment of CMA instability regions with LINEs

Statistical significance of correlation between the paired LINE insertions, and the regions containing CNVs of CMA patients was estimated. The CMA database of patients with CNVs identified by microarrays was filtered in order to remove the duplicate entries (different patients, who had CNVs in the same region), and cases where the uncertain region for breakpoints (on both ends of the CNV) contained matching LCRs. For these data we computed the expected number ($\epsilon = 0.0583$) of CNVs that a single randomly inserting LINE pair will match.

Then, we calculated the enrichment ($\mathcal{E}(l, id)$) of instability regions in LINEs (as a function of LINE pairs homology length l and significance id, measured as sequence identity percent):

$$\mathcal{E}(l, id) = \frac{\#\text{matched\_regions}(l, id)}{\epsilon \cdot \#\text{LINE\_pairs}(l, id)}$$

where $\#\text{matched\_regions}(l, id)$ is the number of instability regions matched by LINE pairs with the homology length of l or more, and the sequence identity percentage of id or more, while $\#\text{LINE\_pairs}(l, id)$ is the total number of LINE pairs with homology of l or more base pairs, and identity percentage greater than id. The above formula is a simplified version of the actually used algorithm, which, in addition, took into account also border effects CNVs lying near the edges of chromosome and centromeres. The resulting plot of the function $\mathcal{E}$ is shown as Figure 4.2.6. In essence, we took the known positions of CNVs and calculated how many of them would be triggered if the LINEs (in the number found in genome) were inserting in random locations, and compared them to the number actually triggered by the known LINEs.

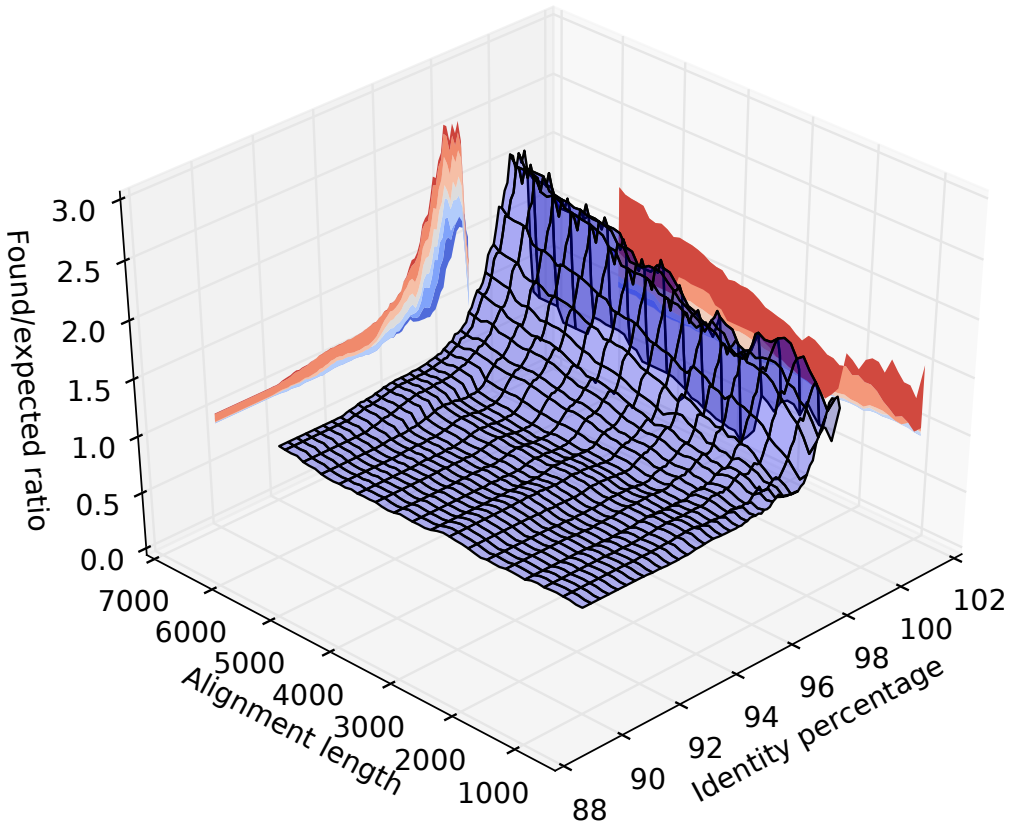


Figure 4.2.6: A plot of observed/expected ratio of matching LINE pairs lying within CNV breakpoint regions. For each point on X and Y axes the observed/expected ratio of LINE pairs with parameters equal or better is shown. It is evident that identity percentage of 96 or better is needed to mediate NAHR, while there is no sharp restriction on minimal length of the homology.

4.3 Results

4.3.1 Regions potentially prone to LINE-LINE-mediated NAHR

Because of the relative abundance of TEs in the human genome compared to LCRs, they have the potential to mediate NAHRs between a wider array of loci, thus potentially posing a significant contribution to genetic instability. From our bioinformatic analysis of the genome, we found 416,180 potential deletion/duplication, 415,581 inversion, and 59,678,570 translocation sites. Figure 4.3.1 indicates the genomic regions potentially susceptible to deletions or duplications due to LINE-LINE mediated NAHR events. Our analysis suggests that 82.8% of the human genome is potentially susceptible to such events, with most of the genome being overlapped by multiple combinations.

Of note, the number of different potential LINE-LINE/NAHR encompassing a given genomic locus (and thus, the probability of its variation in copy number) is dependent on the *square* of the density of homologous LINEs in its vicinity. Therefore, stochastically occurring clusters of LINE elements greatly increase the predicted instability of the genome in a given area.

4.3.2 Analyses of NAHR breakpoints observed in individuals

To gain further insight into LINE-LINE/NAHR, we sought to amplify and characterize the breakpoints of LINE-LINE CNVs identified among patients tested at our clinical laboratory. We cross-referenced our predicted LINE-LINE/NAHR susceptibility regions with the database of CNVs (Fig. 4.2.2). We selected CNVs where the uncertainty regions identified by the clinical aCGH overlapped pairs of LINE elements predicted to mediate LINE-LINE/NAHR. We focused additional attention on loci where two or more distinction patients had CNVs at the same location. We subsequently designed primers complementary to unique genomic sequences that flanked the predicted LINE elements. In 44 cases, we successfully amplified the junction fragment of the putative LINE-LINE/NAHR CNV. Using Sanger sequencing, the location of each NAHR site was narrowed to a homology region between single base mismatches, one belonging to the proximal LINE and the other to the homologous distal LINE. In most cases, it was possible to pinpoint a single breakpoint location for a given patient both manually and using the probability plots obtained from the Hidden Markov Model. The NAHR breakpoints mapped to 80 ± 53 base pairs

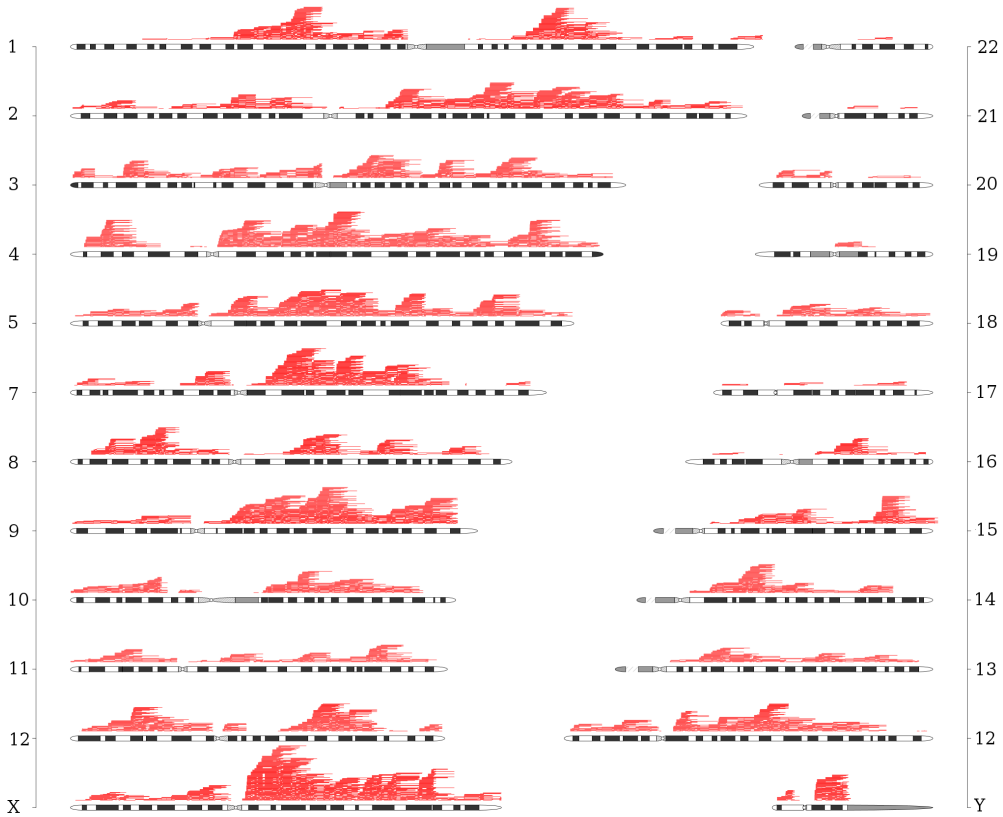


Figure 4.3.1: Ideogram showing the susceptibility of human genome to LINE-LINE-mediated NAHR. Each horizontal red line corresponds to one potentially NAHR-mediating LINE pair: the LINE elements map at the ends of the line, whereas the segment covers the potentially deleted or duplicated regions. For clarity of the figure, inversions and translocations are not shown.

(excluding outliers) indicative of the highly identical nature of the LINE-LINE pairs. In 15 cases, the microhomology was much longer (2.4 ± 0.7kbp). In two cases, the breakpoint could be determined to the basepair.

4.3.3 Enrichment of predicted LINE-LINE/NAHR pairs in CNV uncertainty regions

Because we observed a number of LINE-LINE/NAHR mediated CNVs among patients tested at the diagnostic laboratory, we hypothesized that a considerable fraction of all CNVs could be mediated by LINEs and that such elements would be found within the uncertainty region of CNVs more often than expected by random chance. To test this, we computationally randomized LINE elements throughout the genome to calculate the expected occurrence of LINE-LINE/NAHR pairs in uncertainty regions. We did not identify a significant enrichment of LINE/LINE pairs of a specific size (Fig. 4.2.6). Interestingly, LINE-LINE pairs with high sequence identity occur in uncertainty regions more often than expected, potentially indicating that sequence identity is, comparatively, a more important feature of NAHR promoting sequence.

4.3.4 NAHR hotspots in flanking LINE elements

In some cases, the precise NAHR cross-over site in a given LINE-LINE pair varied between patients, *e.g.* in the case of duplication on chromosome 20 (see Fig. 4.3.2), suggesting independent *de novo* events. In all cases, the NAHR sites either mapped between the same two *cis*-morphisms or were clustered together, typically within 500 bp of each other. This observation could suggest that inside the LINE elements there could exist NAHR-facilitating motifs which make some regions of the LINE more prone to recombination than others. We performed computational analysis of potential hotspot motifs, including the canonical motif associated with PRDM9 binding (Segurel, 2013); however, we were unable to identify significant enrichments near the identified breakpoints. It should, however be emphasized that such studies are very difficult if not impossible due to the nature of LINE elements, which are all very similar to each other, and it is impossible to distinguish motifs associated with LINEs from motifs associated with NAHR using the data we have.

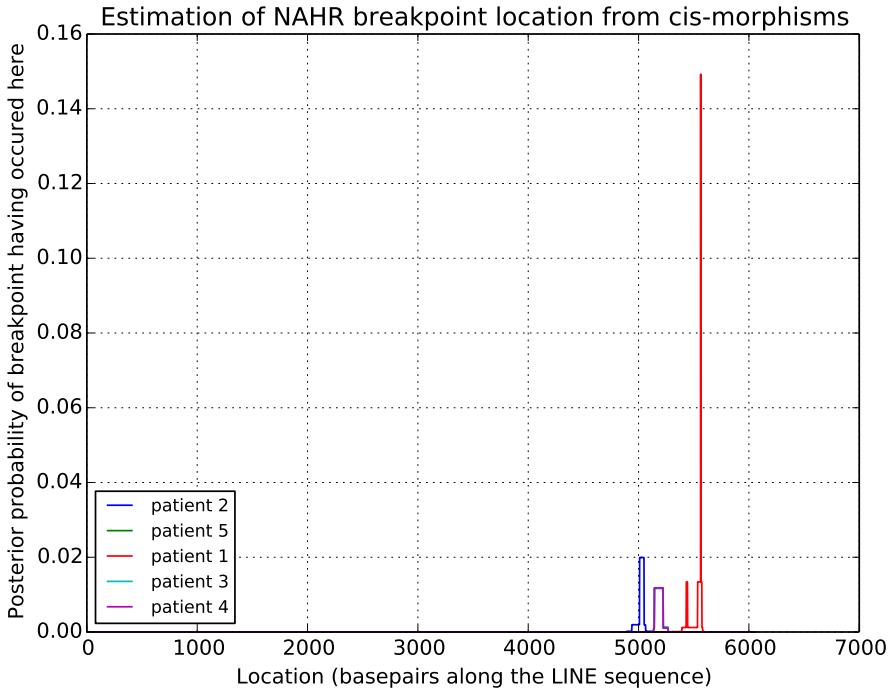


Figure 4.3.2: Estimated NAHR breakpoint location probabilities from the Hidden Markov Model for duplications between LINEs on chromosome 20. Three distinct NAHR loci were identified among the tested patients.

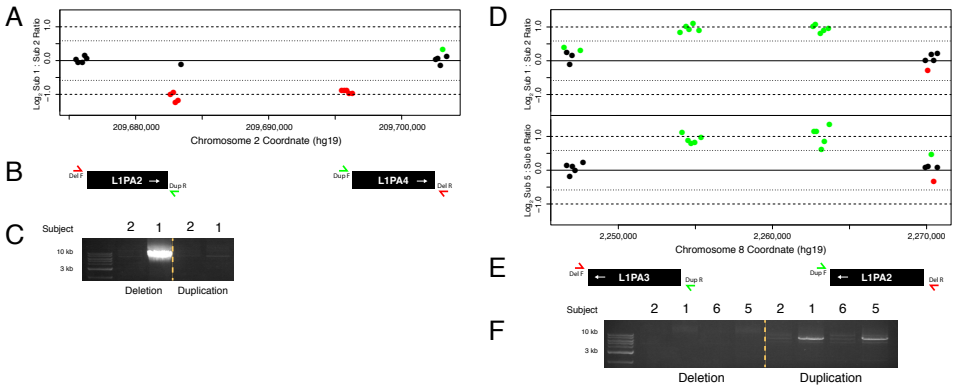


Figure 4.3.3: Molecular validation of predicted LINE-LINE CNVs identified among healthy individuals by aCGH. A) Array CGH data indicates a CNV at 2q34 in subject 1 or 2. B) Schematic representation of the L1PA elements that mediate the CNV and LR-PCR primers testing for the CNV. C) LR-PCR identifies the presence of a deletion in subject 1. D) Array CGH data indicates a CNV at 8p23.3 in subject 1 or 2 and subject 5 or 6. E) Schematic representation of the L1PA elements that mediate the CNVs and LR-PCR primers testing for the CNVs. F) LR-PCR identifies the presence of homozygous duplications in subjects 1 and 5.

4.3.5 aCGH Analysis of Healthy Individuals

Given our identification of LINE-mediated CNVs among individuals tested for suspected genetic disease, we hypothesized that LINEs may also contribute to genome variation in the normal population. To test this, we performed high-resolution aCGH with probes flanking the LINE elements, that we computationally predicted to contribute to genome instability, on peripheral blood DNA from six healthy individuals. We identified 13 potential CNVs mediated by the LINE pairs predicted in our computational analysis (Table 1). The CNVs identified in control individuals were small, each less than 25 kbp, including the deleted or duplicated segment of LINE. At two loci, the CNVs involved intronic sequences of RefSeq genes. Each CNV identified in the healthy subjects overlapped similarly sized deletions or duplications in the Database of Genomic Variants (MacDonald et al., 2014), suggesting their widespread occurrence.

Because genomic DNA from the healthy individuals was hybridized together, the aCGH data alone are insufficient to differentiate a deletion in one individual from a homozygous duplication in the other – the log2 ratio of either event is ± 1.0. For two loci (2q34 and 8p23.2) where the genomic architecture surrounding the CNV was conducive to unique primer design and long range PCR,

we validated the CNVs molecularly (Fig. 4.3.3). We used two pairs of primers at each locus to test for both deletions and duplications separately. The PCR revealed a heterozygous deletion in subject 1 at 2q24 and homozygous duplications in subjects 1 and 5 at 8p23.2. These data suggest that small LINE-LINE mediated CNVs are present in the normal population and are common enough to be in the homozygous state. Thus, such CNVs likely contribute to normal genetic variation.

4.4 Discussion

Here we have performed a comprehensive, genome-wide study to determine the extent and frequency of LINE-LINE/NAHR events and assess the impact of such events on the genome and potential to cause genetic disease. Our results indicate that LINE-mediated NAHR does occur frequently on a genome scale, and can be responsible for rearrangements resulting in genomic disorders. We propose that the traditional lower bound of minimal length of homology required for NAHR should be reevaluated. We have found NAHR events occurring between elements with as little as 4 kbp of homology. Furthermore, our statistical analysis (Fig. 4.2.6) suggests that, at least for LINE-LINE/NAHR, high homology may be a more important property than sequence length.

The analyses of breakpoints within sequenced individuals showed slight discrepancies in the breakpoint location, which points to the fact that at least some of the CNVs arose as independent events, rather than being inherited from a common ancestor. This points to the destabilizing influence which the LINE elements still exert upon the human genome, and confirms NAHR as the mechanism of this influence, rather than causing replication errors. Our data shows that LINE elements contribute to human genetic variability by mediation of NAHR in addition to active retrotransposition (Kidd et al., 2010; Lupski, 2010). The scale of such influence is proportionally higher than the influence of HERV elements (Shuvarikov et al., 2013), and is significant enough that LINE elements should be considered as one of the major genomic features responsible for promotion of NAHR events, in addition to the ones already known (Dittwald et al., 2013).

The clustering of NAHR breakpoints in particular hotspots within LINE elements suggests that at least some LINEs carry a recombination-promoting motif or sequence. Our current analysis failed to identify a statistically significant enrichment of any motifs within the LINE elements tested in this study. However,

clustering seems to be corroborated by previous studies of NAHR (McVean, 2010). Our findings may have dramatic consequences for population genetics studies concerning the role of TEs in evolution, particularly in differences between TE behavior in sexual versus asexual species. One particularly interesting possibility is that TEs could have been co-opted during evolution of sexuality to spread recombination sites through the genome.

However, in addition to the clinical relevance summarized above, the research presented in this chapter suggests that TE-mediated NAHR is, as suspected, a real phenomenon, and, more importantly, it does occur frequently enough to affect the evolution of the genome. This means that even static, inactive TEs may exert a mutagenic pressure on the host organism, which fact is relevant to the modelling. In particular, it confirms the ability of mathematical model from Chapter 2 to be used for modelling of organisms carrying inactive TEs.

5

TIRfinder and TRANScendence: transposable element detection tools

In this chapter we shall present two computational tools for the detection of TEs in the sequenced genomes of various organisms. The goal of these tools is to enable the large-scale analysis of genomes in a standardised manner, so that the transposon landscapes of different organisms may be compared, and the results of that shall be used for further fine-tuning of the parameters of the models presented. The first of the tools, the TIRfinder, suited for an in-depth study of a single, already known family of class-II, TIR-carrying transposons, while the second TRANScendence is a general, genome-wide *de-novo* TE detection and annotation tool.

5.1 TIRfinder

First tool developed for the study of transposable elements, during the initial phases of collaboration with the team at University of Agriculture in Krakow is the TIRfinder. This tool is specifically made to facilitate the search for class-II transposons carrying Terminal Inverted Repeats (TIRs). In addition

to carrying TIRs, the transposons, upon their insertion, cause a Target Site Duplication (TSD) – and both of these features are used in the recognition of transposons.

The importance of accurate TE annotation and masking for the structural characterization of newly sequenced genomes drives the interest in developing new methods for TE detection and analysis (Bergman and Quesneville, 2007). The exhaustive list of tools and resources for TE analysis compiled by Bergman Lab (http://bergmanlab.smith.man.ac.uk/) contains about 120 items. A large number of them are designed for particular TE families (eg, Helitrons (Du et al., 2008) LTR retrotransposons (Ellinghaus et al., 2008)) or for analysis of particular species (like Drosophila melanogaster), several of these are for general use, i.e., for all kinds of repeats (RepeatMasker (Tarailo-Graovac and Chen, 2009), REPuter (Kurtz and Schleiermacher, 1999), PILER (Edgar and Myers, 2005). RepeatScout (Price et al., 2005) RECON (Bao and Eddy, 2002)), and finally six of them are suitable for the structural analysis of class II elements (Inverted Repeat Finder (Warburton et al., 2004), MAK (Yang and Hall, 2003), MITE-hunter (Han and Wessler, 2010), MUST (Chen et al., 2009), STAN (Nicolas et al., 2005), TRANSPO (Santiago et al., 2002)),with the latter three providing a web interface.

MUST (Chen et al., 2009) allows the user to search for all Miniature Inverted-repeat TEs (MITEs) that satisfy given criteria corresponding to minimum and maximum length of TIR, TSD and size of MITEs (up to 1000 bp). TRANSPO (Santiago et al., 2002) in addition to the functionality of MUST, enables the user to specify the sequence of TIRs and maximum number of errors allowed in TIRs. STAN (Nicolas et al., 2005) is the most flexible tool. It finds all sequences containing a given pattern specified in the SVG grammar. However, STAN uses fixed (non-parametrised) definition of inverted repeats which makes it less suitable for searching class II transposons.

All of the tools mentioned above have been developed to facilitate identification of MITEs, i.e., elements that have TIRs but lack any coding capacity. Our TE discovery tool, TIRfinder, (which also captures specific structural features) offers functionality that goes beyond the proposed methods. It combines an efficient approach based on suffix trees that allows for de novo TE detection, with the possibility of a deep analysis of a specific TE family, based on its structural characteristics. In particular, while searching for all putative TEs, TIRfinder allows the user to specify TIR and TSD patterns as a sequence of A, C, T, G or symbols from extended IUPAC nomenclature (Cornish-Bowden, 1985). This

provides the ability to define TIR/TSD patterns as consensus sequences which combine conserved and non-conserved positions.

Recently, deep sequencing projects have increased dramatically, raising the possibility of the repetitive DNA characterization of not completely sequenced organisms. Notice that our approach based on a suffix tree requires a large contiguous genomic sequence. Therefore for analysis of non-assembled raw reads produced by next generation sequencing, recently developed tools which utilize a clustering approach (Novak et al., 2010) should be applied.

5.1.1 Methods of TE detection and analysis

TIRfinder allows efficient searches of the DNA for structured set of motifs that define TEs to be conducted. Then it performs a BLAST search to find transposase or any other TE-related open reading frames (ORFs), aiming at the detection of autonomous copies, as well as elements directly derived from them. The method works in three stages: structural analysis, functional analysis and MITE analysis (see Fig. 5.1.1).

In the first step, TIRfinder scans the input sequence for all putative TEs, based on the given structural characteristic of a particular TE family, ie, patterns describing TIR, TSD, the number of allowed mismatches and size limits of desired TEs. The algorithmic details of the structural analysis are described further in this section.

Next, all elements identified in the structural analysis stage are analyzed to check whether they contain the ORF coding for a protein specified by the user, most often it would be a transposase. For this purpose, they are aligned with the protein sequence (using an appropriate version of the BLAST algorithm (Altschul et al., 1990)) and elements with statistically significant similarity are marked as putative autonomous. Each autonomous element is then used to search for so called *derivatives*. These are non-autonomous elements that have emerged from an autonomous elements during the course of evolution, often by internal deletions disrupting the ORFs. To classify the TE as a derivative of a given autonomous element, we require that both pairs of subterminal regions share a significant level of similarity which is defined by the user (e.g., BLAST E-value $<$ E-10).

Finally, the remaining set of putative TEs (ie, those not classified as autonomous TEs and their derivatives) is analyzed to find short elements which cluster together based on their sequence similarity. The level of similarity is de-

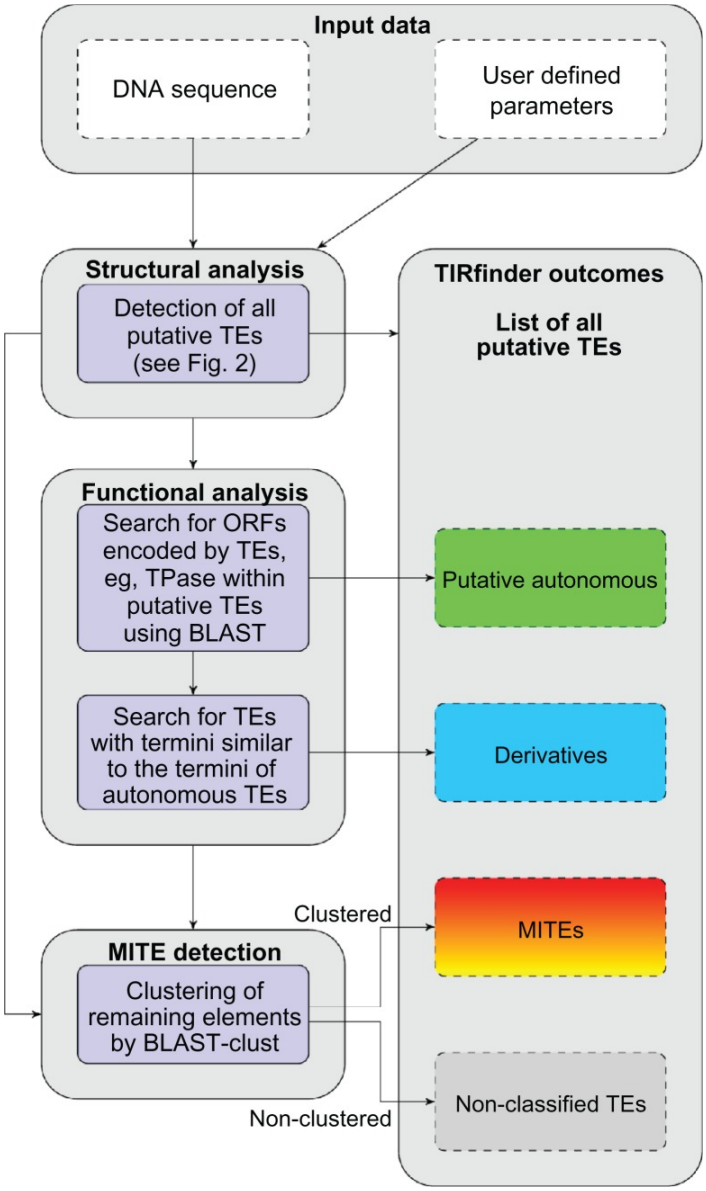


Figure 5.1.1: The control flow through different phases of the TIRfinder TEs detection method.

fined by the user. Clustering is performed using BLAST-clust algorithm (Altschul et al., 1990). The TEs assigned to clusters are labeled as MITEs.

Structural analysis

Our application follows a structure-based approach, i.e., it relies on the detection of specific models of TE architecture consisting of a pair of TIRs (inverted repeats) that are flanked by TSDs (direct repeats).

For detecting TEs we use suffix trees which are very efficient data structures, commonly used in computer science over last four decades. They consists of a root, nodes and labeled edges representing one or more characters from input sequence. All suffixes of the input sequence can be obtained by traversing the suffix tree from the root to leaves. The core idea is that all suffixes that share the common prefix are hanged off on the common node, which reduces the total number of nodes and memory usage.

The algorithm implemented in TIRfinder takes as an input a DNA sequence, a mask corresponding to combined TSD and TIR patterns, and other parameters (e.g., the number of mismatches), see Figure 5.1.2 for the pseudocode of the algorithm and Figure 5.1.3 for graphical explanation of the mask notion.

First, the DNA sequence is divided into a set of smaller fragments of the same length, corresponding to the maximal TE size. A suffix tree is built for every two consecutive fragments, with overlaps of one fragment length (see Fig. 5.1.2B). This is to ensure that we do not miss any match. The algorithm for each fragment independently searches all matches for the mask. In this step, all potential 3'-ends of a TIR should be found. Then we determine if the complementary (5'-end) part of the repeat exists, which is the most time-consuming part of the algorithm. In order to do this efficiently the suffix tree is used (it is built in linear time with respect to a length of the fragment). For each match (TIR + TSD) found in the previous step, the reverse complement of TIR followed by TSD is searched in the suffix tree (see Fig. 5.1.2).

The principal advantage of our approach is memory efficiency: the genomic DNA sequence is split into smaller fragments and the suffix tree data structure needs only a linear amount of space. Thus, we do not impose any limits to the total length of the sequence and the stand-alone version of TIRfinder can be run even on standard PC computers.

Data: DNA sequence G, TIR pattern, TSD pattern, max_size, mismatch_thresholds

Result: all regions flanked by TIR's and TSD up to a predefined mismatch threshold

```
1  Split the sequence G into fragments g_i = 1..n of size = max_size;
2  MASK = TSD · TIR;
3  foreach sequence g_i · g_{i+1} do
4  |  build the suffix tree ST_i;
5  |  find all matches to MASK in g_{i+1}, for example
   |  m_1, m_2, ..., m_j, ..., m_k;
6  |  foreach m_j do
7  |  |  find in the suffix tree ST_i all positions that match to
   |  |  revcomp(m_j);
   |  |  /* revcomp = reverse complement                          */
8  |  |  check the number of mismatches between MASK and
   |  |  revcomp(m_j);
9  |  end
10 end
```

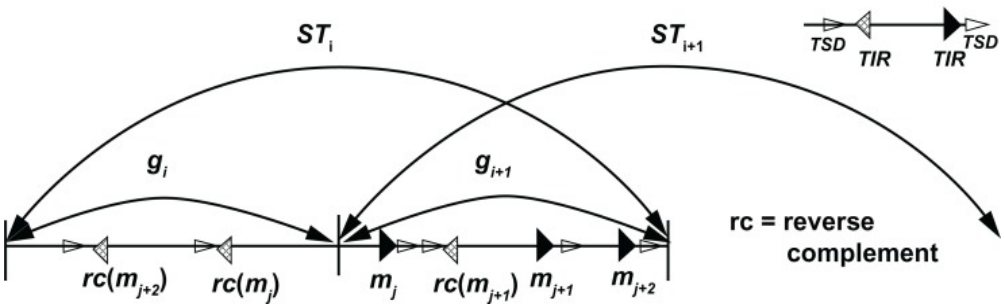


Figure 5.1.2: TIRfinder algorithm. The suffix trees are overlapping to ensure that sequences on the boundaries between two suffix trees are detected correctly.

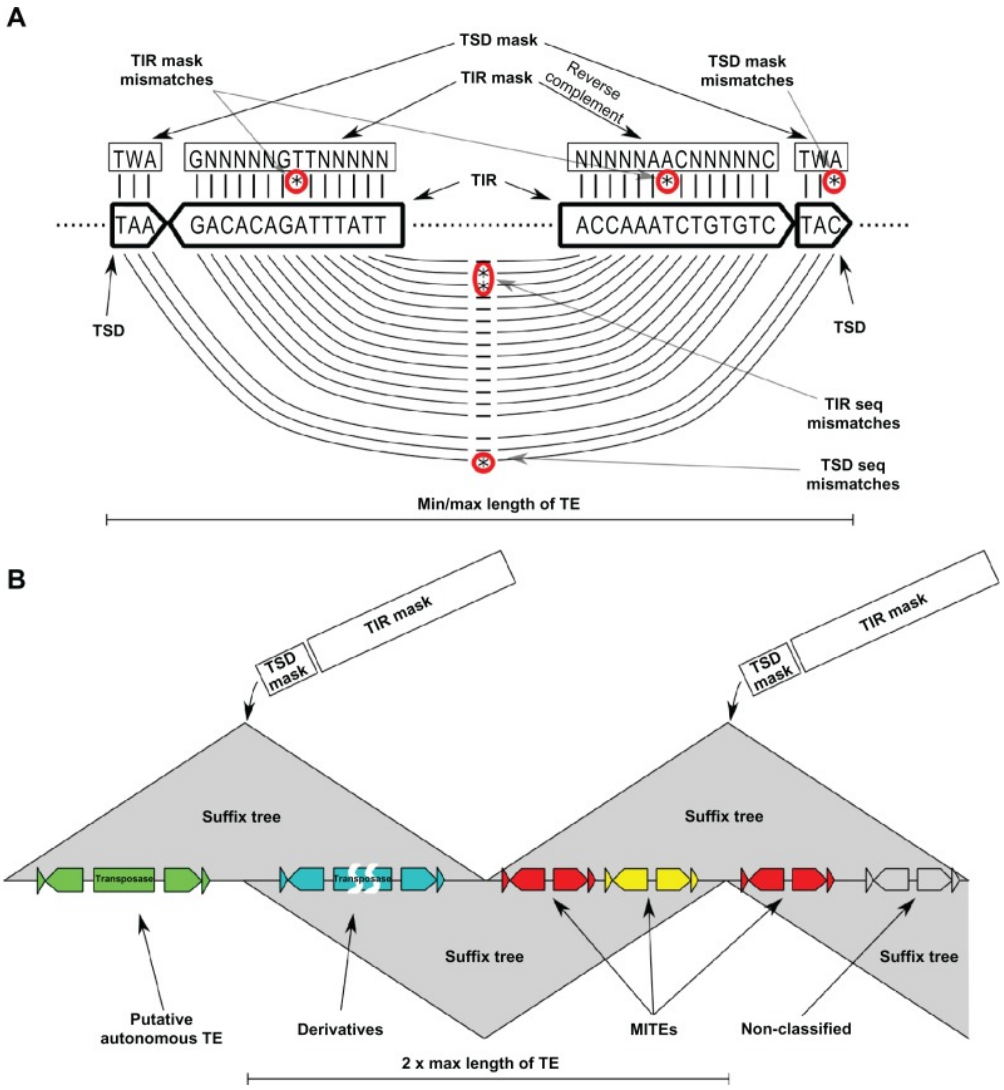


Figure 5.1.3: TIRfinder – structural analysis. **(A)** Explanation of TIR and TSD mask concept. **(B)** Overview of TEs detection phase

5.1.2 Usage

The user must first select the genomic sequence to be searched for TEs. Several plant genomes are currently available in the TIRfinder website (*Arabidopsis thaliana*, *Arabidopsis lyrata*, *Oryza sativa japonica*, *Oryza sativa indica* and *Medicago truncatula*); other genomes can be provided upon request. Alternatively, the user can provide any sequence by uploading a FASTA file not exceeding 100 MB.

Next, the user has to define several parameters such as a minimal and maximal distance between TIRs and patterns of inverted and direct repeats. Patterns of the particular TIR or TSD can be determined by means of a finite string composed of extended IUPAC nucleotide alphabet (Cornish-Bowden, 1985).

Note that the user has to define only one side of the repeat. The second part will be computed as a reverse complement in the case of TIR or simply duplicated in the case of TSD. Furthermore, there is a possibility to specify a maximal number of mismatches between TSDs and TIRs flanking each copy of identified TEs. TIRfinder reads a DNA sequence given in a FASTA file to search for results and saves them in a simple text format. Then the file can be further processed in order to obtain more detailed and accurate data.

Carefully prepared input data, such as TIR and TSD masks, are extremely important in order to get satisfactory results. One possible way is to take TIR sequences from individual copies of TEs representing the desired family and their corresponding flanking TSDs, and create consensus for TSD and TIR sequences. Afterwards, the user can set up other parameters of TIRfinder, which allow for the control of similarity between mined elements and the mask (MaskMismatches) or between corresponding 5'- and 3'- TIRs and TSDs of found elements (SeqMismatches) see Figure 5.1.3A. The ability to manipulate these parameters makes it easier to search for known, well conserved elements as well as to mine new (sub)families of TEs de novo.

Subsequently, the user may specify parameters for the identification of TE copies carrying ORFs, which for simplicity was dubbed functional analysis. First, the protein sequence (in FASTA format) is provided to detect ORFs (a significance threshold is required). Then, long non-autonomous copies – direct derivatives are identified as elements sharing similarity of subterminal regions (length parameter in bp and alignment significance threshold are provided by the user) with one of the previously found autonomous TEs. If the ORF sequence is unknown to the user, the functional analysis step can be omitted in

the course of analysis.

Finally, the user defines constraints for MITE analysis, ie,: minimal and maximal length of MITEs, maximal number of MITE clusters and the level of in-cluster similarity.

5.1.3 TIRfinder implementation

Current release of TIRfinder is an open source web application built with Java, Perl, Apache and other tools (ie, BLAST, BLAST-clust) hosted at `http://bioputer.mimuw.edu.pl/tirfindertool/`. Previous, stand-alone release of TIRfinder, used for case studies reported in Grzebelus et al. (2007, 2009) is available at `http://sourceforge.net/projects/tirfinder/`.

5.1.4 Case Studies

As an example of TIRfinder application, we searched the genome of *Arabidopsis thaliana* and *Medicago truncatula* for the *ATHPOGON3* class II TE family (Le et al., 2000). We fixed TIR and TSD strings as consensus subsequences from given sequences of *ATHPOGO, ATHPOGON1, ATHPOGON2* and *ATHPOGON3* from Repbase (Jurka et al., 2005). Finally, we obtained the following parameters: TIR pattern = CAGTARAAMCTC-TATAAATTAATA, TSD pattern = TA. We decided to set min distance = 300, max distance = 5000, max TSD mask and TSD seq mismatches = 0, max TIR mask and TIR seq mismatches = $i \in (0, 1, 2, 3, 4, 5)$. The experiment allowed for efficient mining of pogo-like transposons in *A. thaliana* and *M. truncatula*. The number of TEs found by TIRfinder and annotated in Repbase for each chromosome of *A. thaliana* is shown in the Table 5.1.1. The breakdown analysis of the found TE sizes (shown in Fig. 5.1.4) reveals that lengths of the majority of putative TEs correspond to the sizes of known pogo-like elements. Most of these elements were MITEs, while in *M. truncatula* we found only 28 pogo-like elements and no MITEs. It fully confirmed previous reports on these elements, reflecting species-specific behavior of related TEs in the two species (Guermonprez et al., 2008) and confirming TIRfinder efficiency.

Moreover, we performed similar search of *M. truncatula* genome for *PIF/Harbinger* TEs family (using TIR pattern = GNNNNNGTTNNNNN and TSD pattern = TWA). The exemplary results of this analysis, presented by TIRfinder (see

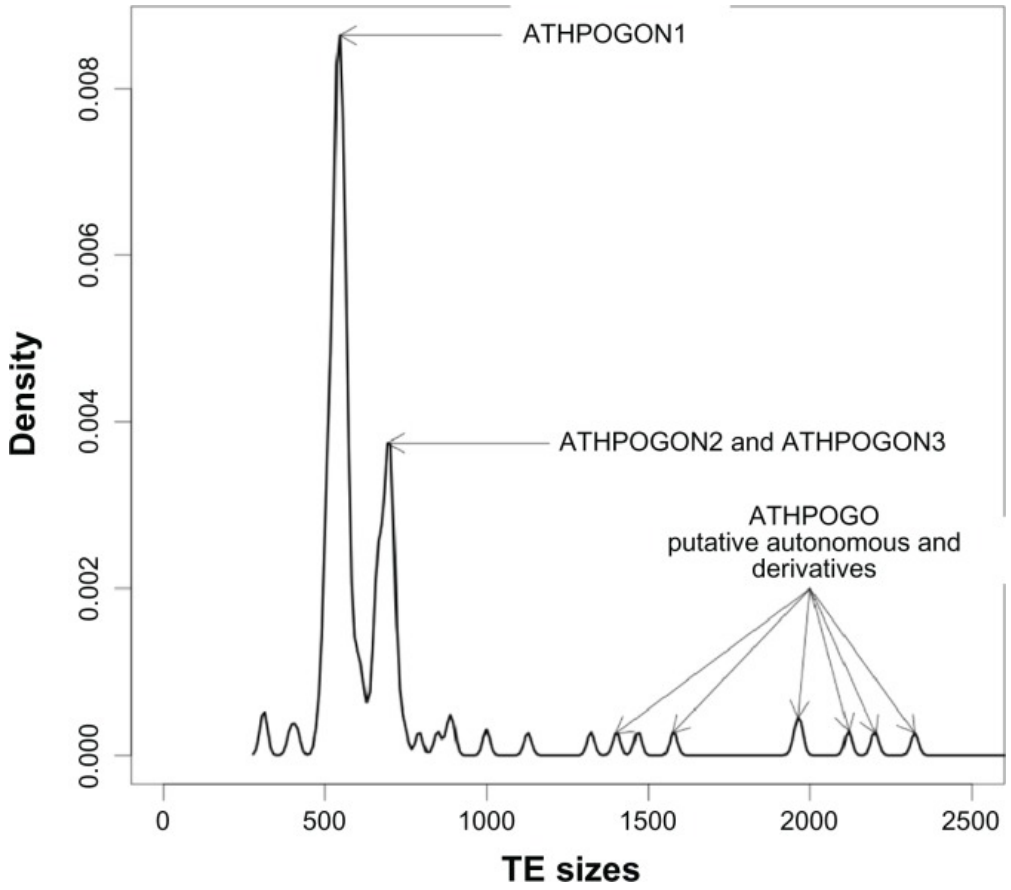
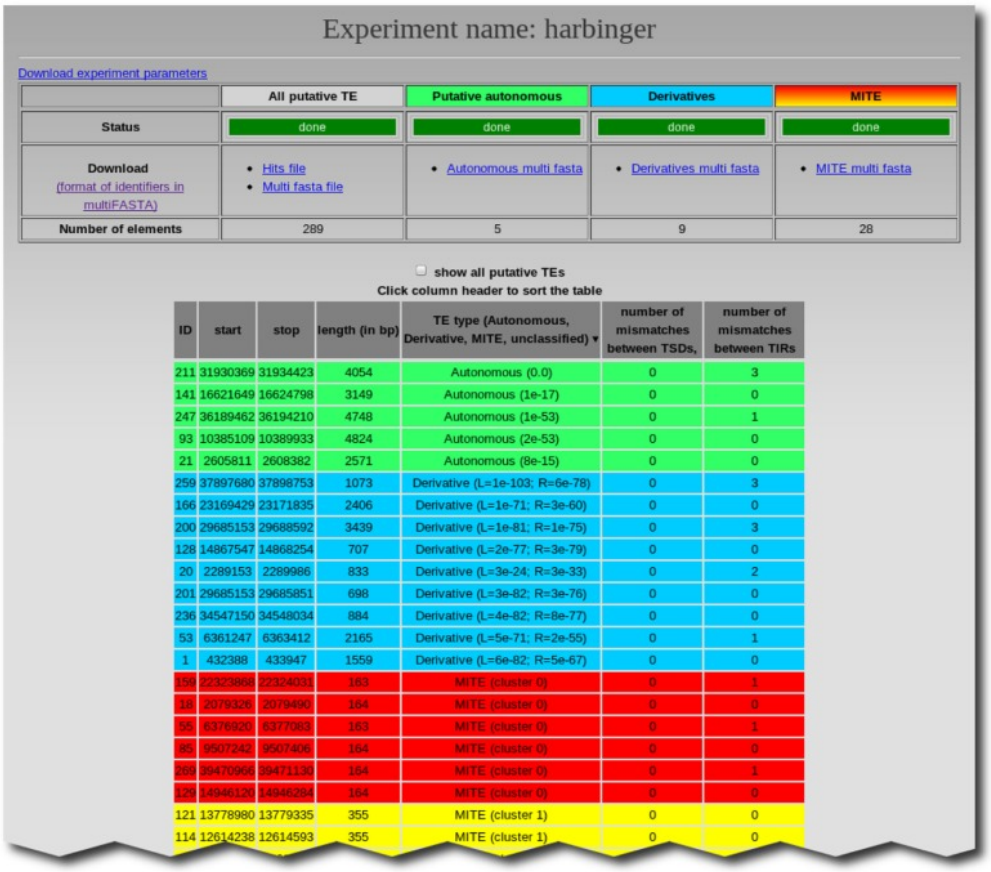


Figure 5.1.4: Pogo-like TEs landscape detected by TIRfinder in *A. thaliana.*

i[a]	chr1	chr2	chr3	chr4	chr5	Sum	Positive predictive value[b]
0	1	2	0	4	2	9	100
1	11	6	7	10	9	43	100
2	20	10	16	17	13	76	95
3	25	20	18	20	14	97	94
4	29	22	22	22	17	112	88
5	33	22	25	24	20	124	84
Rep[c]	25	22	20	22	8	97	

Notes:
[a]Number of allowed TIR mask and TIR seq mismatches
[b]% of TIRfinder output masked by Repbase data
[c]numbers of ATHPOGO elements (>300 bp) annotated in Repbase.

Table 5.1.1: Pogo-like TEs found by TIRfinder vs. annotated in Repbase.



Experiment name: harbinger

Download experiment parameters

	All putative TE	Putative autonomous	Derivatives	MITE
Status	done	done	done	done
Download (format of identifiers in multiFASTA)	• Hits file • Multi fasta file	• Autonomous multi fasta	• Derivatives multi fasta	• MITE multi fasta
Number of elements	289	5	9	28

show all putative TEs

Click column header to sort the table

ID	start	stop	length (in bp)	TE type (Autonomous, Derivative, MITE, unclassified)	number of mismatches between TSDs,	number of mismatches between TIRs
211	31930369	31934423	4054	Autonomous (0.0)	0	3
141	16621649	16624798	3149	Autonomous (1e-17)	0	0
247	36189462	36194210	4748	Autonomous (1e-53)	0	1
93	10385109	10389933	4824	Autonomous (2e-53)	0	0
21	2605811	2608382	2571	Autonomous (8e-15)	0	0
259	37897680	37898753	1073	Derivative (L=1e-103; R=6e-78)	0	3
166	23169429	23171835	2406	Derivative (L=1e-71; R=3e-60)	0	0
200	29685153	29688592	3439	Derivative (L=1e-81; R=1e-75)	0	3
128	14867547	14868254	707	Derivative (L=2e-77; R=3e-79)	0	0
20	2289153	2289986	833	Derivative (L=3e-24; R=3e-33)	0	2
201	29685153	29685851	698	Derivative (L=3e-82; R=3e-76)	0	0
236	34547150	34548034	884	Derivative (L=4e-82; R=8e-77)	0	0
53	6361247	6363412	2165	Derivative (L=5e-71; R=2e-55)	0	1
1	432388	433947	1559	Derivative (L=6e-82; R=5e-67)	0	0
159	22323868	22324031	163	MITE (cluster 0)	0	1
18	2079329	2079490	164	MITE (cluster 0)	0	0
55	6376920	6377083	163	MITE (cluster 0)	0	1
85	9507242	9507406	164	MITE (cluster 0)	0	0
269	39470966	39471130	164	MITE (cluster 0)	0	1
129	14946120	14946284	164	MITE (cluster 0)	0	0
121	13778980	13779335	355	MITE (cluster 1)	0	0
114	12614238	12614593	355	MITE (cluster 1)	0	0

Figure 5.1.5: The example of TIRfinder outcomes: search for *PIF/Harbinger* TEs family in chromosome 5 of *M. truncatula* genome.

Fig. 5.1.5), reveals the occurrence of all functional classes of *PIF/Harbinger* TEs, i.e., putative autonomous, derivatives and MITEs.

5.2 TRANSCENDECE

It has recently come to general attention that TEs may be a major (if not the main) driving force behind speciation and evolution of species (Giordano et al., 2007; Ginzburg et al., 1984) Thus understanding of TE behavior and evolution seems crucial to deepening our knowledge on evolution of species (Kazazian, 2004; Britten, 2010). However, the lack of general, easily-usable and freely available tools for TE detection and annotation, hinders scientific progress in this area. This is especially apparent with the advent of next-generation sequencing techniques, and the resultant abundance of genomic data, most of

which has not been scanned for transposable elements yet.

TRANScendence aims to fill that niche. In contrast with previous methods, which either require manual assistance (and thus, are unsuitable for high-throughput analyses), or require deep programmer's experience to set-up and use, proposed tool is fully automatic (though it is possible to manually curate the results if desired).

In addition to that, however, our motivation is enabling the calibration and parameter estimation of models presented in Chapters 1 and 2 of this thesis, as well as to gather supporting data. Our goal is to produce a reliable, high-throughput tool for analysis of various genomes in a standardized manner. In many organisms the available databases of TEs have varied states of completeness (for example, the TE landscape of humans is much better annotated than that of some newly-sequenced organisms, making direct comparison between the impossible). Our goal is to ascertain whether there is correlation between observed TE counts and the amount of environmental stress a given organism has been subjected to over the course of its recent evolution. The creation of a *de novo* TE detection tool will allow us to avoid bias associated with different levels of annotation for different organisms, and the bias associated with the incompleteness of any preexisting database such as that of Repbase (Kapitonov and Jurka, 2008).

Another aspect, associated with modelling of TE evolution is the tool's capability to defragment nested TE insertions and create a nesting graph. This will (in future) be used to estimate the ages of various TE families, and to study their past activity, which will enable us to see if there is any correlation between a species' history, and the activity pattern of its TEs, as revealed by the present TE landscape. Such correlation, if found, would be useful to strengthen the arguments behind the TE models we have created, and to fine-tune their parameters.

However we have decided to make the tool publicly available to other groups, and we have included some extra functionality that might be useful to groups with which we are not directly collaborating. In addition to tagging of TEs in genomes, our tool is capable of performing different qualitative and quantitative analyses. It classifies TEs into families, superfamilies and orders, allowing us to estimate relative abundances of TEs in selected genomes, and to perform comparative genomic studies. It also performs searches for TE clusters, i.e. regions containing high concentration of TEs nested in one another.

The main objective of the proposed solution is the support of TE evolutionary

studies. We put special emphasis on the design of the web-interface to assure simplicity and flexibility of data manipulation process.

The presented tool is not built from scratch, but consists of several components, which are extended for our purposes. One of the main tool components is REPET package (Flutre et al., 2011), which enables us to implement *de-novo* TE mining and annotation pipeline. The REPET itself combines several different programs for the clustering of interspersed repeats, like GROUPER (Quesneville et al., 2003), RECON (Bao and Eddy, 2002) and PILER (Edgar and Myers, 2005). Also the annotation phase requires the use of multiple mechanisms, mainly based on comparisons to TE elements stored in the Repbase (Kapitonov and Jurka, 2008).

We will present a general description of the tool, and demonstrate its capabilities on an example case: the study of *Medicago Truncatula* genome.

The tool is freely available for use by the general public at: `http://bioputer.mimuw.edu.pl/transcendence`

5.2.1 Functionality of TRANScendence

The standard use-case consists of three steps: TE detection phase, TE annotation phase and TE nesting analysis, see Figure 5.2.1. The workflow of our utility usually begins with a user uploading (through a web-based interface) a (possibly zipped) set of FASTA files, containing the genome of organism to be searched. The user then creates an 'experiment' on the genome. Within the experiment, the genome gets automatically searched for TEs with the help of REPET pipeline (Flutre et al., 2011).

The results of the detection phase are all putative transposable elements stored in the TE repository. Further in the annotation phase all elements found within the experiment, are automatically annotated (against Repbase), and are made available for the user to either view or download. Furthermore, the TEs get clustered into families, with consensus sequences for each family being computed, and annotated. The consensus sequences for detected TE families may be downloaded as well. For the convenience of users the additional output is an annotated version of the genome, in formats suitable for widely-used genomic browsers such as GBrowse (Donlin, 2002) or Apollo (Lewis et al., 2002).

In next step the tool analyzes the nesting structure of detected transposable elements. A graphical visualization of TE interruptions is generated, allowing

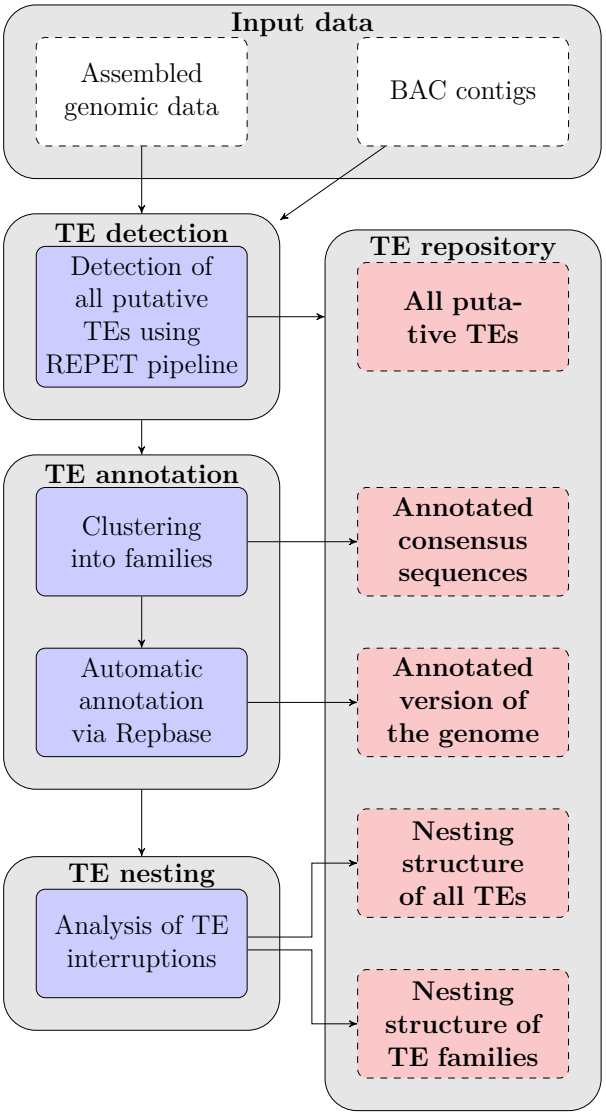


Figure 5.2.1: Overview of the TRANScendence tool.

the study of genomic sites with deeply nested TEs, as well as the display of more detailed information about such sites. In addition to that, the tool presents the user with a graph of interruptions on TE family level, allowing for chronological study of TE families. To ensure the high quality of data stored in the repository the user is allowed to perform manual curation of all annotations. The useful option available to the user is to present his own uploaded genomes as 'public', i.e. and viewable by other users of the system. Experiments performed by users may be shared with their consent as well. Using this mechanism, the database of *Medicago Truncatula* TEs has been made available for public viewing.

5.2.2 The TE Landscape of the *Medicago truncatula*

Medicago truncatula, also called barrel medic is the primary model, or reference legume species for genomic and functional genomic research. The sequenced part of its genome (ca. 313 Mbp) was scanned for TEs and 121 509 elements have been found and grouped into 2456 families. Approximately 80 Mbp have been found to be contained in a TE.

1356 families group 51740 class I TEs, while 59855 class II elements were classified into 803 families. Among DNA TEs 4907 elements are MITEs and the set of other TEs carrying terminal inverted repeats encompasses: 4475 Harbinger, 3181 hAT, 1015 Mariner/Tc1, 210 Enhancer/Suppressor mutator (En/Spm)-like TEs, 22657 MuDR elements and only 3 Polinton elements. It should be noted that in the latter case the homology to known Polintons from Repbase is rather weak, so these should probably be considered artifacts. Finally 2557 elements were classified as Helitrons – this is, surprisingly, much more than in closely related species like *Lotus japonicus* (Holligan et al., 2006).

Detected retrotransposons were divided into most abundant Gypsy superfamily (16440 TEs), Copia (15930 TEs), 26 ERV1-type repeats, 8195 L1 elements, and 134 RTE superfamily elements.

TE detection phase is summarized with several statistics presented at Experiment's results page, see Figure 5.2.2.

In addition to the quantitative results which allow to compare genomes composition in various TE families, all the data stored in our database can be easily manipulated and visualized. The example of usage may be downloading the TE flanking sequences for designing the PCR starters. As stated above all putative TE elements are automatically classified into orders and superfamilies. Moreover we annotate each detected TE family by aligning the family consensus

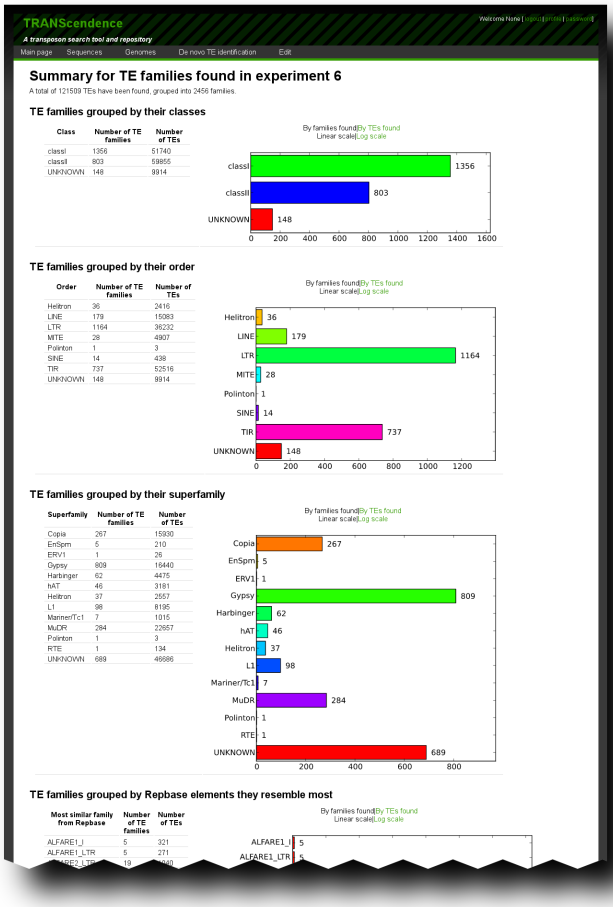


Figure 5.2.2: All found putative TE elements are classified into classes, orders and superfamilies.

sequence against elements stored in Repbase Update (Kapitonov and Jurka, 2008), see Figure 5.2.3.

5.2.3 TEs interruption graph

High density of TEs, especially in plant genomes, is the result of the constant bombardment of the genome by different TEs over millions of years. The insertion activity may cause the splitting of TEs already existing in the genome into non-contiguous fragments separated by the sequence of newly inserted elements. The identification of nested transposable elements is important for

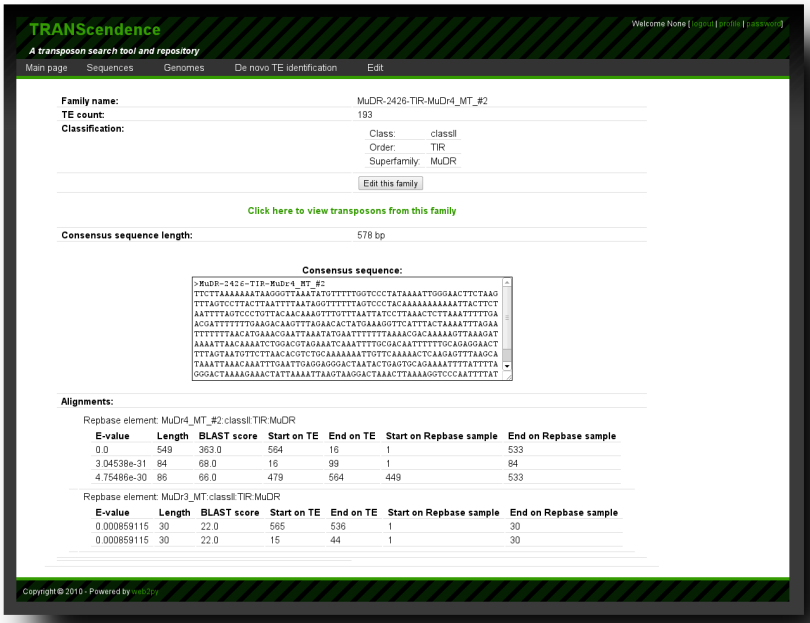


Figure 5.2.3: Each TE family is annotated by BLASTing the consensus sequence against Repbase content.

evolutionary comparisons among various regions of the genome. For human genome the chronology of TE families based on the TEs nesting was studied in Giordano et al. (2007). For highly repetitive plants genomes the TEnest tool have been proposed (Kronmiller and Wise, 2008) for visual representation of TE integration history.

The latter tool focuses on LTR carrying TEs and more importantly is no longer unavailable to use at PlantGDB site. Thus in our service we implemented the nesting repeats identification scheme analogous to the approach previously applied to human genome. All found TE nestings are displayed in a graphical format, and may be reviewed online. In addition to that, a graph of nesting dependencies between TE families is constructed, along with interruption matrix. These are made available for download by the user for further analyses.

The Figure 5.2.4 represents the interruption graph of TE families. The nodes are colored in such a way that families belonging to the same superfamily have the same color, e.g. on this picture the MuDR superfamily is green, L1 superfamily is assigned a different shade of green, Gypsy is purple, Harbinger – blue, and so on. Unknown (unclassified) elements are colored in orange. This serves as a visual aid, to assist in recognizing clusters of interesting TEs from the same superfamily.

The graph is useful for assessing relative ages of TE families – if an edge exists from family A to family B, then it means that at least one TE from family B has been found to be interrupting a TE from family A. This means that the family B must have been active when at least some of the TEs from family A have settled, and so, cannot be older than family A.

For example, from the zoomed part of the picture, we can conclude that the family Gypsy-480-LTR-Ram9B_I_#2 is probably older than L1-1403-LINE-SHALINE2_MT_#2. On the other hand, consider families MuDR-1264-TIR-MuSHAN_MT, MuDR-1321-TIR-MuSHAN_MT, and MuDR-1209-TIR-MuSHAN_MT. They are joined by a cyclic path, and therefore, probably have been active concurrently, and as such, have similar age.

5.2.4 Automatic TE mining

The task of searching for TEs is not an easy one. It has been split into many smaller sub-tasks, such as searching for repeat genomic sequences, annotation of structural elements of TEs, clustering TEs into families, annotating TEs

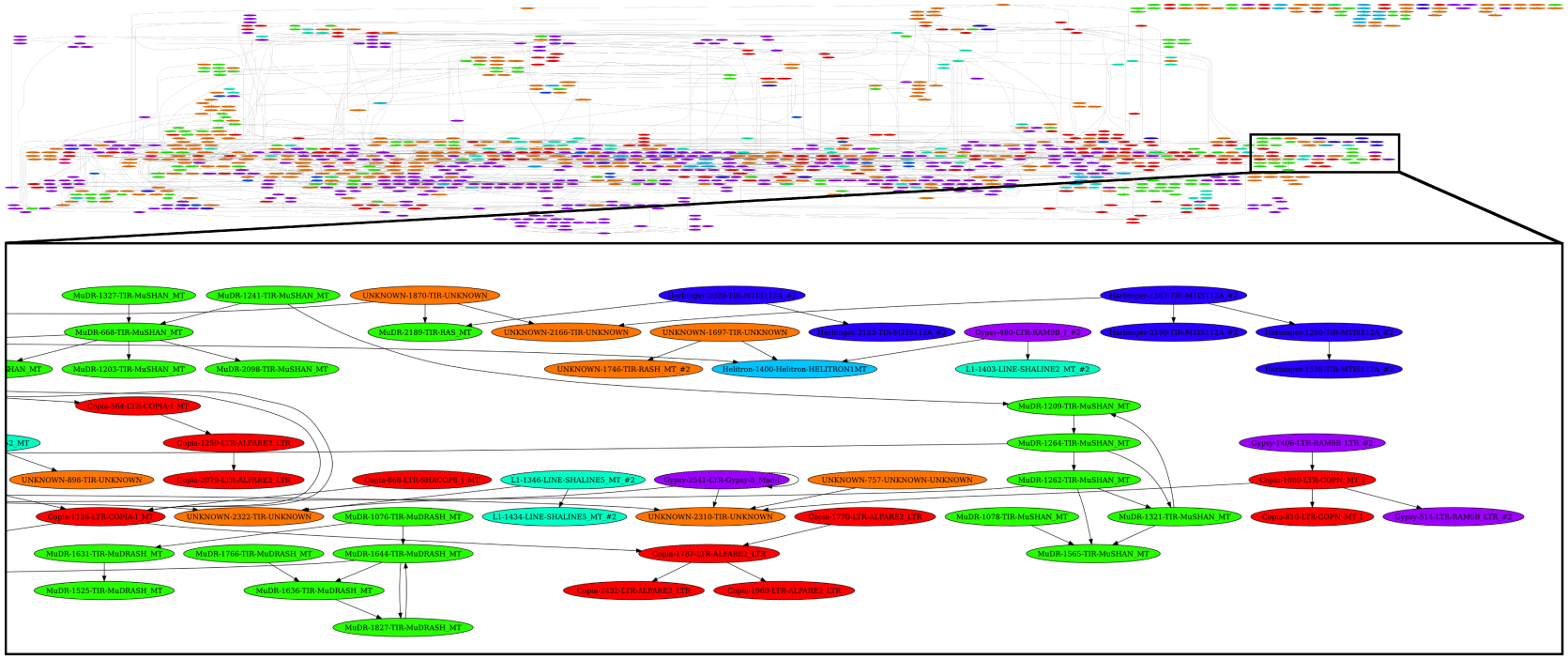


Figure 5.2.4: Nesting structure for TE families detected in *Medicago truncatula*.

in genomes, de-novo searching for repeatable elements, classifying repeatable elements, and so on.

A multitude of tools that can perform each of these steps have been written, often with one task being covered by several tools that perform it in slightly different way (and with slightly different results).

Therefore a need has arisen to merge the appropriate tools into one generic pipeline, which would be capable of performing a complete, *de-novo* annotation of TEs in a whole genome, from the ground up, starting only from the sequence of an organism's genome.

One such pipeline has been created, called REPET (Flutre et al., 2011). It joins over ten distinct tools from the field of computational biology, in an effort to provide a comprehensive utility for *de-novo* TE annotation.

However, in order to use the pipeline, each one of its components has to be separately installed and configured. The REPET utility runs in many stages, each one of them controlled by a command-line interface, and this makes the tool hard to use for users without profound programming skills. Last but not least, the REPET tool requires a complex setup of a grid environment and database, even if it is to be run on a single computer. Therefore, the setup process is rather difficult and time-consuming causing that the tool remains practically unused, despite its enormous scientific potential.

Here we have decided to integrate REPET pipeline with a flexible, relational TE repository, along with a web interface to benefit from a wide range of efficient services, but at the same time to eliminate the inconvenience of the lack of user-friendly interface.

For this reason the TRANScendence software is not available as a stand-alone software package, only as a web server – our goal was simplicity and ease-of-use, meanwhile setting up TRANScendence on any local server would require the configuration of the many building blocks we used - MySQL database, Apache web server, Sun Grid Engine, web2py environment, and, most importantly, REPET along with all of its modules.

5.2.5 TE repository

TRANScendence tool stores all detected TE elements and associated information in a relational database. The database was created using MySQL and consists of the several tables containing the data about: *(i)* analyzed genomes (BAC contigs, TE mining experiments performed, etc.), *(ii)* detected TEs (also

TE fragments) and *(iii)* TE families (consensus sequences, alignments with Repbase elements, etc). The simple relationships between the tables facilitate the possible expansion of the database to other genomic features potentially valuable in evolutionary studies (e.g. transcription factor binding sites).

5.2.6 Technologies used

The user interface of our utility has been implemented mainly in Python, within the web2py framework, with matplotlib and graphviz being used for data visualization. The website make heavy use of SVG vector graphics, therefore a browser supporting SVG (such as Mozilla Firefox, Google Chrome or newer versions of Internet Explorer) is recommended for viewing it. At the core of the application is the REPET pipeline, which is used for TE searches. REPET, in turn is made of multiple bioinformatics tools, such RECON(Bao and Eddy, 2002), CENSOR (Kohany et al., 2006), PILER (Edgar and Myers, 2005), RepeatMasker (Smit et al., 2004), TRF (Benson, 1999), Mreps (Kolpakov et al., 2003), and others. We use NCBI BLAST (Altschul et al., 1990) for homology searches.

The service uses MySQL database as its backend, as well as the backend for REPET. The genome files are stored in the filesystem as FASTA files. Most of the computational load is shared between nodes with the use of Sun Grid Engine.

5.2.7 Discussion

We would like to emphasize that, our service, however, is more than just 'online, user-friendly REPET'. In fact we have integrated REPET with appropriately designed TE database, which allows for storing the results of computations, and their visualization. The utility also provides an interface that assists experts in performing a manual curation of obtained results, making it especially useful for de-novo analysis of newly sequenced species. In such cases the Repbase (Kapitonov and Jurka, 2008) is usually insufficient for annotation of found TEs in a satisfactory manner.

Predictably, the lack of easily usable TE annotation pipelines hinders the comparative genomics of TE families. Especially the problem of establishing chronologies of activity of various TE families on relation to one another remains elusive. The standard approach of reconstructing the phylogenetic trees of TEs is somewhat hindered by the junk-DNA status of TEs. Because they

are not conserved, they are vulnerable to large-scale genomic deletions and other rearrangements. This causes TE copies to be mangled and fragmented, which presents a difficulty for standard phylogenetic tools, as they have mostly been developed with (much better-conserved) genes in mind. As such, the chronologies obtained from sequence-similarity-based tools are not fully trusted, and could stand to be verified with other methods.

One such attempt at verification has already been performed for human genome (Giordano et al., 2007) establishing chronology based on the insertions of TEs within each other. The approach however, while being a significant improvements over the previous methods, depends on a preexisting TE database, and employs an ad-hoc TE defragmentation method. Obviously, the state-of-the-art utilities for de-novo TE detection and analysis could widen the scope of application of this method, as well as produce better results.

To this end we focused our tool on obtaining interruption matrices of TEs (that is, data representing how TEs nest in each other), and provides an option of visualizing the interruptions graph – which constitute the useful tool for assessing the periods of activity of TE families, as well as for their dating. The interruption matrix may be downloaded in plain-text format for use in the user's own analyses, however, we are currently working on the implementation of an automated TE dating module.

5.3 Conclusions

We have developed two tools designed for the study of TEs in the genome. In addition to being useful to a wide audience, the tools will be used to calibrate the parameters of the presented models of transposition. The ability to perform comparative studies of TE landscape in a wide array of species, as well as to reconstruct the TE activity from the interruption matrix should greatly enhance the value of the presented models. This, in turn, will result in a deeper understanding of the role of TEs as evolution helpers.

6

Conclusions and further research

We have presented a body of research regarding the interdependence of evolution and TEs, particularly in conditions of environmental stress. Our research suggests that in addition to being genomic parasites, TEs may act as evolutionary helpers, assisting adaptation to new environments, and enabling faster and more efficient colonization of new environmental niches. However, the study of the role of TEs is far from finished, and here we shall present some ongoing and future directions of research.

6.1 Spatial extension to computational model

One of the directions for further research is a spatial extensions of the computational, and perhaps the mathematical model. Such a model will allow us to perform an in-depth study of the dynamics of TEs in different subpopulations, particularly competition between TEs-carrying organisms, and organisms dispossessed of them. One particularly interesting scenario to be studied is the co-evolution of both TE-carrying organisms and ones without them while climbing a phenotypic gradient. It is expected that the forefront of colonization will consist of organisms with high copy numbers of transposable elements, and

these will be followed by organisms with low copy numbers. However, the interesting question is whether this will be the result of the TE-carrying population silencing and loosing their TEs after the initial colonization frontwave passes, or whether the high-TE organisms will be just pushed out by a slower-adapting subpopulation without TEs. Some initial work in this direction has been performed by M. Kitlas, we are currently in the process of fine-tuning the parameters of the model.

6.2 Model for sexually-reproducing organisms

Both the computational and mathematical models presented here assume a population of asexual, clonally-reproducing organisms. An interesting extension of both models would be changing of the reproduction model from asexual to diploid-based, sexual one. In case of the computational model this would involve the addition of a (diploid) genetic model, as the tracking of insertion sites for each TE becomes important, to account for effects such as homo/heterozygosity, and sexual mode of trait inheritance. A wide consensus seems to be that TEs may persist in sexual organisms due to their properties as selfish, parasitic DNA (in asexuals such a mechanism is impossible). Our preliminary results suggest that in addition to the parasitic DNA dynamics, the mode of evolutionary helpers can also be achieved in certain conditions, with TEs being able to proliferate as evolutionary helpers under conditions of environmental stress, even if their basal level of transposition is insufficient for them to persist as parasitic DNA in conditions of stable environment. The work on this extension is being performed by K. Gogolewski, and right now we're fine-tuning model parameters and performing an initial exploration of interesting scenarios.

6.3 Mathematical modelling of class-I TE and sexually-reproducing organisms

In addition to the model already presented, we plan to continue the attempts to derive an equilibrium state for class-I TEs. One approach we plan to use is to discard the notion of discrete TEs (in similar fashion to the model used for class-I TEs), and to allow TE counts to evolve following the Gaussian distribution. This should simplify the equations while still remaining realistic enough to

represent real-world phenomena.

In addition to that, we plan to extend the mutator model to sexual organisms. Sexual reproduction should not drastically change the mathematical apparatus involved in the solution, and so, should be tractable – while allowing us to study the response of sexual versus asexual organisms to environmental change. This might allow us to answer important questions about evolution of sexuality.

6.4 Further development of TE-detection tools

As the cell culture of *Medicago truncatula* experiment mentioned in Chapter 1 nears its completion, the TE detection tools will probably need to be expanded in order to study any interesting phenomena we will uncover. While currently it is difficult to predict the direction in which the data analysis might take us, there is one extension we can already anticipate: it is the addition of a module to the TRANScendence tool which will enable the study of TE nesting matrix. This will enable the recovery of the chronology of activity of various TE families, and will shed some further light on the problem of the dynamics of activity of transposable elements, this time from an experimental viewpoint. The challenging part is that even in its simplest form the problem is NP-complete (easily reduces to the minimal feedback arc set problem), which precludes a direct solution, and forces the use of a heuristic algorithm. Some initial heuristic algorithm ideas have been already sketched, however very little work has already been done in this direction yet.

Glossary

Abbreviation	Meaning
aCGH	Array-comparative genomic hybridization – a technique for detecting copy number changes in a genome.
CMA	Chromosomal microarray.
CNV	Copy number variation – an alteration of DNA structure resulting in abnormal number of copies of a given sequence (e.g. duplication or deletion).
DNA	Deoxyribonucleic acid.
FGM	Fisher's geometric model – a type of model used in population genetics.
HERV	Human endogenous retrovirus – a family of virus-derived class I TEs.
HT	Horizontal transfer – a transfer of genetic material between organisms outside of normal parent-child inheritance mechanisms.
LINE	Long INterspersed Element, or alternatively: Long Interspersed Nuclear Element – a type of class-I TEs abundant in humans.
LR-PCR	Long-range PCR.
LTR	Long Terminal Repeat – a sequence carried by certain class-I transposons, demarcating their ends.
MITE	Miniature inverted-repeat transposable element – a family of class-II transposons.
NAHR	Nonallelic homologous recombination – one of the mechanisms of mutation, resulting in appearance of CNVs.
ORF	Open Read Frame – a framgnet of DNA strand which is transcribed to RNA, often containing a gene.
PCR	Polymerase chain reaction – a reaction allowing to check for presence, and to aplify a selected fragment of DNA.

Abbreviation	Meaning
PDF	Probability density function.
RNA	Ribonucleic acid.
SINE	Short INterspersed Element, or alternatively: Short Interspersed Nuclear Element – a type of class-I TEs nonautonomous TEs, which uses enzymes carried by LINE TEs for transposition.
SNV	Single nucleotide variant.
TE	Transposable element.
TIR	Terminal Inverted Repeat – a sequence carried by certain class-II transposons, demarcating their ends.
TSD	Target Site Duplication – a short sequence at both ends of certain transposons, resulting from a duplication of a short DNA fragment during the process of their insertion.

Bibliography

Aldershof, B., Marron, J., Park, B., Wand, M., 1995. Facts about the gaussian probability density function, Applicable Analysis, 59(1-4), p. 289.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., Lipman, D. J., 1990. Basic local alignment search tool, Journal of Molecular Biology, 215(3), p. 403.

Arkhipova, I., Meselson, M., 2000. Transposable elements in sexual and anciant asexual taxa., Proc. Natl. Acad. Sci. USA, 97, p. 14473.

Arkhipova, I., Meselson, M., 2005. Deleterious transposable elements and the extinction of asexuals., Bioessays, 27(1), p. 76.

Bao, Z., Eddy, S. R., 2002. Automated de novo identification of repeat sequence families in sequenced genomes, Genome Res., 12(8), p. 1269.

Barrett, R. D., Schluter, D., 2008. Adaptation from standing genetic variation., Trends Ecol. Evol., 23(1), p. 38.

Bartle, R., Sherbert, D., 2011. Introduction to Real Analysis, Wiley.

Basten, C. J., Moody, M. E., 1991. A branching-process model for the evolution of transposable elements incorporating selection., J. Math. Biol., 29(8), p. 743.

Beck, C. R., Garcia-Perez, J. L., Badge, R. M., Moran, J. V., 2011. LINE-1 elements in structural variation and disease, Annu. Rev. Genomics Hum. Genet., 12, p. 187.

Belancio, V. P., Deininger, P. L., Roy-Engel, A. M., 2009. LINE dancing in the human genome: transposable elements and disease, Genome Med, 1(10), p. 97.

Belshaw, R., Pereira, V., Katzourakis, A., Talbot, G., Paces, J., Burt, A., Tristem, M., 2004. Long-term reinfection of the human genome by endogenous retroviruses, Proc. Natl. Acad. Sci. U.S.A., 101(14), p. 4894.

Benson, G., 1999. Tandem repeats finder: a program to analyze dna sequences., Nucleic Acids Research, 27(2), p. 573.

Bergman, C. M., Quesneville, H., 2007. Discovering and detecting transposable elements in genome sequences, Brief. Bioinformatics, 8(6), p. 382.

Bichsel, M., Barbour, A. D., Wagner, A., 2010. The early phase of a bacterial insertion sequence infection, Theor. Pop. Biol., 78(4), p. 278.

Biémont, C., 2010. From genotype to phenotype. what do epigenetics and epigenomics tell us?, Heredity, 105(1), p. 1.

Blot, M., 1994. Transposable elements and adaptation of host bacteria., Genetica, 93(1-3), p. 5.

Boone, P. M., Bacino, C. A., Shaw, C. A., Eng, P. A., Hixson, P. M., Pursley, A. N., Kang, S. H., Yang, Y., Wiszniewska, J., Nowakowska, B. A., del Gaudio, D., Xia, Z., Simpson-Patel, G., Immken, L. L., Gibson, J. B., Tsai, A. C., Bowers, J. A., Reimschisel, T. E., Schaaf, C. P., Potocki, L., Scaglia, F., Gambin, T., Sykulski, M., Bartnik, M., Derwinska, K., Wisniowiecka-Kowalnik, B., Lalani, S. R., Probst, F. J., Bi, W., Beaudet, A. L., Patel, A., Lupski, J. R., Cheung, S. W., Stankiewicz, P., 2010. Detection of clinically relevant exonic copy-number changes by array CGH, Hum. Mutat., 31(12), p. 1326.

Boone, P. M., Soens, Z. T., Campbell, I. M., Stankiewicz, P., Cheung, S. W., Patel, A., Beaudet, A. L., Plon, S. E., Shaw, C. A., McGuire, A. L., Lupski, J. R., 2013. Incidental copy-number variants identified by routine genome testing in a clinical population, Genet. Med., 15(1), p. 45.

Boutin, T. S., Le Rouzic, A., Capy, P., 2012. How does selfing affect the dynamics of selfish transposable elements?, Mob. DNA, 3(1), p. 5.

Britten, R. J., 2010. Transposable element insertions have strongly affected human evolution, Proceedings of the National Academy of Sciences, 107, p. 19945.

Brookfield, J., 1996. Models of the spread of non-autonomous selfish transposable elements when transposition and fitness are coupled, Genet. Res., 67, p. 199.

Bürger, R., 2000. The Mathematical Theory of Selection, Recombination, and Mutation, Wiley Series in Mathematical & Computational Biology, Wiley.

Burger, R., Lynch, M., 1995. Evolution and extinction in a changing environment: a quantitative-genetic analysis, Evolution, p. 151.

Burwinkel, B., Kilimann, M. W., 1998. Unequal homologous recombination between LINE-1 elements as a mutational mechanism in human genetic disease, J. Mol. Biol., 277(3), p. 513.

Capy, P., 1998. A plastic genome., Nature, 396, p. 522.

Capy, P., Gasperi, G., Biémont, C., Bazin, C., 2000. Stress and transposable elements: co-evolution or useful parasites?, Heredity, 85(2), p. 101.

Chandler, M., Mahillon, J., 2002. Mobile DNA II, chap. Insertion sequences revisited., American society for microbiology press, p. 305.

Chao, L., Vargas, C., Spear, B. B., Cox, E. C., 1983. Transposable elements as mutator genes in evolution., Nature, 303(5918), p. 633.

Charlesworth, B., Barton, N., 2004. Genome size: does bigger mean worse?, Curr. Biol., 14(6), p. R233.

Charlesworth, B., Charlesworth, D., 1983. The population dynamics of transposable elements., Genet. Res., 42, p. 1.

Charlesworth, B., Sniegowski, P., Stephan, W., 1994. The evolutionary dynamics of repetitive DNA in eukaryotes., Nature, 371, p. 215.

Chen, J.-M., 2012. The first human mitotic nonallelic homologous recombination hotspot associated with genetic disease, Human Mutation, 33(11), p. v.

Chen, Y., Zhou, F., Li, G., Xu, Y., 2009. MUST: a system for identification of miniature inverted-repeat transposable elements and applications to Anabaena variabilis and Haloquadratum walsbyi, Gene, 436(1-2), p. 1.

Collins, S., de Meaux, J., Acquisti, C., 2007. Adaptive walks toward a moving optimum, Genetics, 176(2), p. 1089.

Condit, R., Stewart, F. M., Levin, B. R., 1988. The population biology of bacterial transposons: a priori conditions for maintenance as parasitic DNA, Am. Nat., 132, p. 129.

Cornish-Bowden, A., 1985. Nomenclature for incompletely specified bases in nucleic acid sequences: recommendations 1984, Nucleic Acids Res., 13(9), p. 3021.

Daubin, V., Moran, N. A., 2004. Comment on "the origins of genome complexity"., Science, 306(5698), p. 978.

Dittwald, P., Gambin, T., Szafranski, P., Li, J., Amato, S., Divon, M. Y., Rodriguez Rojas, L. X., Elton, L. E., Scott, D. A., et al., 2013. NAHR-mediated copy-number variants in a clinical population: mechanistic insights into both genomic disorders and Mendelizing traits, Genome Res., 23(9), p. 1395.

Dolgin, E. S., Charlesworth, B., 2006. The fate of transposable elements in asexual populations, Genetics, 174, p. 817.

Donlin, M. J., 2002. Using the Generic Genome Browser (GBrowse), John Wiley & Sons, Inc.

Doolittle, W., Sapienza, C., 1980. Selfish genes, the phenotype paradigm and genome evolution., Nature, 284(5757), p. 601.

Du, C., Caronna, J., He, L., Dooner, H. K., 2008. Computational prediction and molecular confirmation of Helitron transposons in the maize genome, BMC Genomics, 9, p. 51.

Dufresne, M., Lespinet, O., Daboussi, M.-J., Hua-Van, A., 2011. Genome-wide comparative analysis of pogo-like transposable elements in different Fusarium species, J. Mol. Evol., 73, p. 230.

Durand, E., Tenaillon, M. I., Ridel, C., et al., 2010. Standing variation and new mutations both contribute to a fast response to selection for flowering time in maize inbreds., BMC Evol. Biol., 10, p. 2.

Eddy, S. R., 2004. What is a hidden markov model?, Nat. Biotech., 22(10), p. 1315.

Edgar, R. C., Myers, E. W., 2005. PILER: identification and classification of genomic repeats., Bioinformatics (Oxford, England), 21 Suppl 1(suppl_1), p. i152.

Edwards, R., Brookfield, J., 2003. Transiently beneficial insertions could maintain mobile DNA sequences in variable environments., Mol. Biol. Evol., 20(1), p. 30.

Ellinghaus, D., Kurtz, S., Willhoeft, U., 2008. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons, BMC Bioinformatics, 9, p. 18.

Ewing, B., Green, P., 1998. Base-calling of automated sequencer traces using phred. II. Error probabilities, Genome Res., 8(3), p. 186.

Ewing, B., Hillier, L., Wendl, M. C., Green, P., 1998. Base-calling of automated sequencer traces using phred. I. Accuracy assessment, Genome Res., 8(3), p. 175.

Fehér, T., Bogos, B., Méhi, O., et al., 2012. Competition between transposable elements and mutator genes in bacteria., Mol. Biol. Evol., 29(10), p. 3153.

Filée, J., Siguier, P., Chandler, M., 2007. Insertion sequence diversity in archaea., Microbiol. Mol. Biol. Rev., 71(1), p. 121.

Fisher, R. A., 1930. The genetical theory of natural selection, Oxford University Press.

Flutre, T., Duprat, E., Feuillet, C., Quesneville, H., 2011. Considering transposable element diversification in de novo annotation approaches, PLoS ONE, 6(1), p. 15.

Gillespie, J. H., 1981. Evolution of the mutation rate at a heterotic locus., Proc. Natl. Acad. Sci. USA, 78(4), p. 2452.

Ginzburg, L. R., Bingham, P. M., Yoo, S., 1984. On the theory of speciation induced by transposable elements., Genetics, 107(2), p. 331.

Giordano, J., Ge, Y., Gelfand, Y., Abrusán, G., Benson, G., Warburton, P. E., 2007. Evolutionary history of mammalian transposons determined by genome-wide defragmentation, PLoS Computational Biology, 3(7), p. 14.

Giraud, A., Matic, I., Tenaillon, O., et al., 2001. Costs and benefits of high mutation rates: adaptive evolution of bacteria in the mouse gut., Science, 291(5513), p. 2606.

Grandbastien, M.-A., Audeon, C., Bonnivard, E., et al., 2005. Stress activation and genomic impact of tnt1 retrotransposons in solanaceae, Cytogenet. Genome Res., 110(1-4), p. 229.

Grzebelus, D., Gładysz, M., Maćko-Podgórni, A., Gambin, T., Golis, B., Rakoczy, R., Gambin, A., 2009. Population dynamics of miniature inverted-repeat transposable elements (MITEs) in medicago truncatula, Gene, 448(2), p. 214.

Grzebelus, D., Lasota, S., Gambin, T., Kucherov, G., Gambin, A., 2007. Diversity and structure of PIF/Harbinger-like elements in the genome of Medicago truncatula, BMC Genomics, 8, p. 409.

Gu, W., Zhang, F., Lupski, J. R., 2008. Mechanisms for human genomic rearrangements, Pathogenetics, 1(1), p. 4.

Guermonprez, H., Loot, C., Casacuberta, J. M., 2008. Different strategies to persist: the pogo-like Lemi1 transposon produces miniature inverted-repeat transposable elements or typical defective elements in different plant genomes, Genetics, 180(1), p. 83.

Han, K., Lee, J., Meyer, T. J., Remedios, P., Goodwin, L., Batzer, M. A., 2008. L1 recombination-associated deletions generate human genomic variation, Proc. Natl. Acad. Sci. U.S.A., 105(49), p. 19366.

Han, Y., Wessler, S. R., 2010. MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences, Nucleic Acids Res., 38(22), p. e199.

Hansen, T. F., 2006. The evolution of genetic architecture, Annu. Rev. Ecol. Evol. Syst., 37, p. 123.

Hartl, D. L., Lozovskaya, E. R., Lawrence, J. G., 1992. Nonautonomous transposable elements in prokaryotes and eukaryotes, Genetica, 86, p. 47.

Hickey, D. A., 1982. Selfish DNA : a sexually-transmitted nuclear parasite., Genetics, 101, p. 519.

Higashimoto, K., Maeda, T., Okada, J., Ohtsuka, Y., Sasaki, K., Hirose, A., Nomiyama, M., Takayanagi, T., Fukuzawa, R., Yatsuki, H., Koide, K., Nishioka, K., Joh, K., Watanabe, Y., Yoshiura, K., Soejima, H., 2013. Homozygous deletion of DIS3L2 exon 9 due to non-allelic homologous recombination between LINE-1s in a Japanese patient with Perlman syndrome, Eur. J. Hum. Genet., 21(11), p. 1316.

Holligan, D., Zhang, X., Jiang, N., Pritham, E. J., Wessler, S. R., 2006. The transposable element landscape of the model legume lotus japonicus., Genetics, 174(4), p. 2215.

Iafrate, A. J., Feuk, L., Rivera, M. N., Listewnik, M. L., Donahoe, P. K., Qi, Y., Scherer, S. W., Lee, C., 2004. Detection of large-scale variation in the human genome, Nat. Genet., 36(9), p. 949.

International Human Genome Sequencing Consortium, 2001. Initial sequencing and analysis of the human genome, Nature, 409(6822), p. 860.

Janion, C., 2008. Inducible SOS response system of DNA repair and mutagenesis in *Escherichia coli.*, Int. J. Biol. Sci., 4(6), p. 338.

Jurka, J., Kapitonov, V. V., Pavlicek, A., Klonowski, P., Kohany, O., Walichiewicz, J., 2005. Repbase Update, a database of eukaryotic repetitive elements, Cytogenet. Genome Res., 110(1-4), p. 462.

Kapitonov, V. V., Jurka, J., 2008. A universal classification of eukaryotic transposable elements implemented in repbase, Nat Rev Genet, 9(5), p. 411.

Kazazian, H. H., 2004. Mobile elements: Drivers of genome evolution, Science, 303(5664), p. 1626 .

Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., Haussler, D., 2002. The human genome browser at UCSC, Genome Res., 12(6), p. 996.

Kidd, J. M., Graves, T., Newman, T. L., Fulton, R., Hayden, H. S., Malig, M., Kallicki, J., Kaul, R., Wilson, R. K., Eichler, E. E., 2010. A human genome structural variation sequencing resource reveals insights into mutational mechanisms, Cell, 143(5), p. 837.

Kidwell, M. G., Lisch, D. R., 2000. Transposable elements and host genome evolution., Trends Ecol. Evol., 15(3), p. 95.

Kidwell, M. G., Lisch, D. R., 2001. Perspective: transposable elements, parasitic DNA, and genome evolution, Evolution, 55(1), p. 1.

Kofler, R., Betancourt, A. J., Schlotterer, C., 2012. Sequencing of pooled DNA samples (Pool-Seq) uncovers complex dynamics of transposable element insertions in Drosophila melanogaster, PLoS Genet., 8(1), p. e1002487.

Kohany, O., Gentles, A. J., Hankus, L., Jurka, J., 2006. Annotation, submission and screening of repetitive elements in repbase: Repbasesubmitter and censor, BMC Bioinformatics, 7(1), p. 474.

Kolpakov, R., Bana, G., Kucherov, G., 2003. mreps: efficient and flexible detection of tandem repeats in dna, Nucleic Acids Research, 31(13), p. 3672.

Kopp, M., Hermisson, J., 2009. The genetic basis of phenotypic adaptation I: fixation of beneficial mutations in the moving optimum model., Genetics, 182(1), p. 233.

Kronmiller, B. A., Wise, R. P., 2008. Tenest: Automated chronological annotation and visualization of nested plant transposable elements, Plant Physiology, 146(1), p. 45.

Kurtz, S., Schleiermacher, C., 1999. REPuter: fast computation of maximal repeats in complete genomes, Bioinformatics, 15(5), p. 426.

Lander, E., Linton, L., Birren, B., et al., 2001. Initial sequencing and analysis of the human genome, Nature, 409(6822), p. 860.

Lawrence R. Rabiner, B. H. J., 1986. An introduction to hidden markov models, IEEE ASSP Magazine, p. 1.

Le, Q. H., Wright, S., Yu, Z., Bureau, T., 2000. Transposon diversity in Arabidopsis thaliana, Proc. Natl. Acad. Sci. U.S.A., 97(13), p. 7376.

Le Rouzic, A., Boutin, T. S., Capy, P., 2007. Long-term evolution of transposable elements., Proc. Natl. Acad. Sci. USA, 104(49), p. 19375.

Le Rouzic, A., Capy, P., 2006. Population genetics models of competition between transposable element subfamilies, Genetics, 174(2), p. 785.

Le Rouzic, A., Deceliere, G., 2005. Models of the population genetics of transposable elements, Genet. Res., 85, p. 171.

Lewis, S. E., Searle, S. M., Harris, N., Gibson, M., Lyer, V., Richter, J., Wiel, C., Bayraktaroglir, L., Birney, E., Crosby, M. A., Kaminker, J. S., Matthews, B. B., Prochnik, S. E., Smithy, C. D., Tupy, J. L., Rubin, G. M., Misra, S., Mungall, C. J., Clamp, M. E., 2002. Apollo: a sequence annotation editor, Genome Biol., 3(12), p. 82.

Lockton, S., Gaut, B., 2010. The evolution of transposable elements in natural populations of self-fertilizing *Arabidopsis thaliana* and its outcrossing relative *Arabidopsis lyrata*, BMC Evol. Biol., 10(1), p. 10.

Lupski, J. R., 2010. Retrotransposition and structural variation in the human genome, Cell, 141(7), p. 1110.

Lynch, M., 2007. The origins of genome architecture, Sinauer Associates Inc, Sunderland, MA, USA.

Lynch, M., Conery, J., 2003. The origins of genome complexity., Science, 302(5649), p. 1401.

MacDonald, J. R., Ziman, R., Yuen, R. K., Feuk, L., Scherer, S. W., 2014. The Database of Genomic Variants: a curated collection of structural variation in the human genome, Nucleic Acids Res., 42(Database issue), p. D986.

Martiel, J., Blot, M., 2002. Transposable elements and fitness of bacteria., Theor. Popul. Biol., 61(4), p. 509.

Martin, G., Lenormand, T., 2006. A general multivariate extension of Fisher's geometrical model and the distribution of mutation fitness effects across species., Evolution, 60(5), p. 893.

Matuszewski, S., Hermisson, J., Kopp, M., 2014. Fisher's geometric model with a moving optimum, Evolution, 68(9), p. 2571.

McClintock, B., 1950. The origin and behavior of mutable loci in maize, Proc. Natl. Acad. Sci. U.S.A., 36(6), p. 344.

McFadden, J., Knowles, G., 1997. Escape from evolutionary stasis by transposon-mediated deleterious mutations, J. Theor. Biol., 186, p. 441.

McGraw, J. E., Brookfield, J. F. Y., 2006. The interaction between mobile DNAs and their hosts in a fluctuating environment., J. Theor. Biol., 243(1), p. 13.

McVean, G., 2010. What drives recombination hotspots to repeat DNA in humans?, Philos. Trans. R. Soc. Lond., B, Biol. Sci., 365(1544), p. 1213.

Miller, W. J., Hagemann, S., Reiter, E., Pinsker, W., 1992. P-element homologous sequences are tandemly repeated in the genome of *Drosophila guanche.*, Proc. Natl. Acad. Sci. USA, 89(9), p. 4018.

Miller, W. J., Mcdonald, J. F., Pinsker, W., 1997. Molecular domestication of mobile elements, Genetica, p. 261.

Mills, R. E., Bennett, E. A., Iskow, R. C., Devine, S. E., 2007. Which transposable elements are active in the human genome?, Trends Genet., 23(4), p. 183.

Moody, M. E., 1988. A branching process model for the evolution of transposable elements., J. Math. Biol., 26(3), p. 347.

Needleman, S. B., Wunsch, C. D., 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins, J. Mol. Biol., 48(3), p. 443.

Newton, I. L. G., Bordenstein, S. R., 2011. Correlations between bacterial ecology and mobile DNA., Curr. Microbiol., 62(1), p. 198.

Nicolas, J., Durand, P., Ranchy, G., Tempel, S., Valin, A. S., 2005. Suffix-tree analyser (STAN): looking for nucleotidic and peptidic patterns in chromosomes, Bioinformatics, 21(24), p. 4408.

Novak, P., Neumann, P., Macas, J., 2010. Graph-based clustering and characterization of repetitive sequences in next-generation sequencing data, BMC Bioinformatics, 11, p. 378.

Ogasawara, H., Obata, H., Hata, Y., Takahashi, S., Gomi, K., 2009. Crawler, a novel Tc1/mariner-type transposable element in *Aspergillus oryzae* transposes under stress conditions., Fung. Genet. Biol., 46(6-7), p. 441.

Orgel, L. E., Crick, F. H. C., 1980. Selfish DNA: the ultimate parasite., Nature, 284, p. 604.

Orr, H. A., 2005. Theories of adaptation: what they do and don't say, Genetica, 123, p. 3.

Ou, Z., Stankiewicz, P., Xia, Z., Breman, A. M., Dawson, B., Wiszniewska, J., Szafranski, P., Cooper, M. L., Rao, M., Shao, L., et al., 2011. Observation and prediction of recurrent human translocations mediated by NAHR between nonhomologous chromosomes, Genome Res., 21(1), p. 33.

Partridge, L., Barton, N., 2000. Evolving evolvability, Nature, 407, p. 457.

Petrov, D. A., Fiston-Lavier, A. S., Lipatov, M., Lenkov, K., Gonzalez, J., 2011. Population genomics of transposable elements in Drosophila melanogaster, Mol. Biol. Evol., 28(5), p. 1633.

Pigliucci, M., 2008. Is evolvability evolvable?, Nat. Rev. Genet., 9(1), p. 75.

Price, A. L., Jones, N. C., Pevzner, P. A., 2005. De novo identification of repeat families in large genomes, Bioinformatics, 21 Suppl 1, p. i351.

Quesneville, H., Nouaud, D., Anxolabéhère, D., 2003. Detection of new transposable element families in drosophila melanogaster and anopheles gambiae genomes., J Mol Evol, 57 Suppl 1.

Radman, M., 1974. Molecular and Environmental aspects of mutagenesis, chap. Phenomenology of an inducible mutagenic DNA repair pathway in *Escherichia coli*: SOS repair hypothesis, Springfield IL: Charles C Thomas publisher, p. 128.

Rankin, D. J., Bichsel, M., Wagner, A., 2010. Mobile DNA can drive lineage extinction in prokaryotic populations., J. Evol. Biol., 23(11), p. 2422.

Robberecht, C., Voet, T., Zamani Esteki, M., Nowakowska, B. A., Vermeesch, J. R., 2013. Nonallelic homologous recombination between retrotransposable elements is a driver of de novo unbalanced translocations, Genome Res., 23(3), p. 411.

Rudin, W., 1976. Principles of Mathematical Analysis, International series in pure and applied mathematics, McGraw-Hill.

Rudin, W., 1987. Real and complex analysis, Mathematics series, McGraw-Hill.

Santiago, N., Herraiz, C., Goni, J. R., Messeguer, X., Casacuberta, J. M., 2002. Genome-wide analysis of the Emigrant family of MITEs of Arabidopsis thaliana, Mol. Biol. Evol., 19(12), p. 2285.

Sawyer, S., Hartl, D., 1986. Distribution of transposable elements in prokaryotes, Theor. Pop. Biol., 30(1), p. 1.

Schnable, P. S., Ware, D., Fulton, R. S., et al., 2009. The b73 maize genome: Complexity, diversity, and dynamics, Science, 326(5956), p. 1112 .

Schneider, D., Lenski, R. E., 2004. Dynamics of insertion sequence elements during experimental evolution of bacteria., Res. Microbiol., 155(5), p. 319.

Sebat, J., Lakshmi, B., Troge, J., Alexander, J., Young, J., Lundin, P., Maner, S., Massa, H., Walker, M., Chi, M., et al., 2004. Large-scale copy number polymorphism in the human genome, Science, 305(5683), p. 525.

Segurel, L., 2013. The complex binding of PRDM9, Genome Biology, 14(4), p. 112.

Shuvarikov, A., Campbell, I. M., Dittwald, P., Neill, N. J., Bialer, M. G., Moore, C., Wheeler, P. G., Wallace, S. E., Hannibal, M. C., Murray, M. F., et al., 2013. Recurrent HERV-H-mediated 3q13.2-q13.31 deletions cause a syndrome of hypotonia and motor, language, and cognitive delays, Hum. Mutat., 34(10), p. 1415.

Sinzelle, L., Izsvák, Z., Ivics, Z., 2009. Molecular domestication of transposable elements: from detrimental parasites to useful host genes., Cell. Mol. Life. Sci., 66(6), p. 1073.

Smart, D., 1980. Fixed Point Theorems, Cambridge tracts in mathematics, Cambridge University Press.

Smit, A. F. A., Hubley, R., Green, P., 2004. Repeat masker open-3.0, http://www.repeatmasker.org.

Stankiewicz, P., Lupski, J. R., 2002. Genome architecture, rearrangements and genomic disorders, Trends Genet., 18(2), p. 74.

Stankiewicz, P., Lupski, J. R., 2010. Structural variation in the human genome and its role in disease, Annu. Rev. Med., 61, p. 437.

Stoebel, D. M., Dorman, C. J., 2010. The effect of mobile element IS10 on experimental regulatory evolution in *Escherichia coli.*, Mol. Biol. Evol., 27(9), p. 2105.

Szafranski, P., Dharmadhikari, A. V., Brosens, E., Gurha, P., Kolodziejska, K. E., Zhishuo, O., Dittwald, P., Majewski, T., Mohan, K. N., Chen, B., Person, R. E., Tibboel, D., de Klein, A., Pinner, J., Chopra, M., Malcolm, G., Peters, G., Arbuckle, S., Guiang, S. F., Hustead, V. A., Jessurun, J., Hirsch, R., Witte, D. P., Maystadt, I., Sebire, N., Fisher, R., Langston, C., Sen, P., Stankiewicz, P., 2013. Small noncoding differentially methylated copy-number variants, including lncRNA genes, cause a lethal lung developmental disorder, Genome Res., 23(1), p. 23.

Taddei, F., Radman, M., Maynard-Smith, J., et al., 1997. Role of mutator alleles in adaptive evolution., Nature, 387(6634), p. 700.

Tanaka, M. M., Bergstrom, C. T., Levin, B. R., 2003. The evolution of mutator genes in bacterial populations: the roles of environmental change and timing., Genetics, 164(3), p. 843.

Tarailo-Graovac, M., Chen, N., 2009. Using RepeatMasker to identify repetitive elements in genomic sequences, Curr. Protoc. Bioinformatics, Chapter 4, p. Unit 4.10.

Temtamy, S. A., Aglan, M. S., Valencia, M., Cocchi, G., Pacheco, M., Ashour, A. M., Amr, K. S., Helmy, S. M., El-Gammal, M. A., Wright, M., Lapunzina, P., Goodship, J. A., Ruiz-Perez, V. L., 2008. Long interspersed nuclear element-1 (LINE1)-mediated deletion of EVC, EVC2, C4orf6, and STK32B in Ellis-van Creveld syndrome with borderline intelligence, Hum. Mutat., 29(7), p. 931.

Travis, J. M. J., Travis, E. R., 2002. Mutator dynamics in fluctuating environments, Proc. Roy. Soc. Lond. B, 269(1491), p. 591.

Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B. C., Remm, M., Rozen, S. G., 2012. Primer3–new capabilities and interfaces, Nucleic Acids Res., 40(15), p. e115.

Vinogradov, A., 2003. Selfish DNA is maladaptive: evidence from the plant Red List., Trends Genet., 19(11), p. 609.

Vissers, L. E., Bhatt, S. S., Janssen, I. M., Xia, Z., Lalani, S. R., Pfundt, R., Derwinska, K., de Vries, B. B., Gilissen, C., Hoischen, A., Nesteruk, M., Wisniowiecka-Kowalnik, B., Smyk, M., Brunner, H. G., Cheung, S. W., van Kessel, A. G., Veltman, J. A., Stankiewicz, P., 2009. Rare pathogenic microdeletions and tandem duplications are microhomology-mediated and stimulated by local genomic architecture, Hum. Mol. Genet., 18(19), p. 3579.

Viterbi, A., 1967. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm, Information Theory, IEEE Transactions on, 13(2), p. 260.

Wagner, A., 2006. Cooperation is fleeting in the world of transposable elements, PLoS Comp. Biol., 2, p. 1522.

Walisko, O., Schorn, A., Rolfs, F., Devaraj, A., Miskey, C., Izsvak, Z., Ivics, Z., 2008. Transcriptional activities of the Sleeping Beauty transposon and shielding its genetic cargo with insulators, Mol. Ther., 16(2), p. 359.

Warburton, P. E., Giordano, J., Cheung, F., Gelfand, Y., Benson, G., 2004. Inverted repeat structure of the human genome: the X-chromosome contains a preponderance of large, highly homologous inverted repeats that contain testes genes, Genome Res., 14(10A), p. 1861.

Waxman, D., Peck, J. R., 1999. Sex and adaptation in a changing environment, Genetics, 153(2), p. 1041.

Welch, L. R., 2003. Hidden markov models and the baum-welch algorithm, IEEE Information Theory Society Newsletter, 53(4), p. 1.

Whitney, K. D., Garland, T., 2010. Did genetic drift drive increases in genome complexity?, PLoS Genet., 6(8).

Wiszniewska, J., Bi, W., Shaw, C., Stankiewicz, P., Kang, S. H., Pursley, A. N., Lalani, S., Hixson, P., Gambin, T., Tsai, C. H., Bock, H. G., Descartes, M., Probst, F. J., Scaglia, F., Beaudet, A. L., Lupski, J. R., Eng, C., Cheung, S. W., Bacino, C., Patel, A., 2014. Combined array CGH plus SNP genome analyses in a single assay for optimized clinical testing, Eur. J. Hum. Genet., 22(1), p. 79.

Wright, S., Finnegan, D., 2001. Sex and transposable elements., Cur. Biol., 11, p. R296.

Wright, S. I., Schoen, D. J., 1999. Transposon dynamics and the breeding system, Genetica, 107, p. 139.

Yang, G., Hall, T. C., 2003. MAK, a computational tool kit for automated MITE analysis, Nucleic Acids Res., 31(13), p. 3659.

Zeyl, C., Bell, G., Green, D. M., 1996. Sex and spread of retrotransposon *ty3* in experimental populations of *Saccharomyces cerevisiae*., Genetics, 143, p. 1567.