



FUNDAÇÃO OSWALDO CRUZ

INSTITUTO OSWALDO CRUZ

PROGRAMA DE PÓS-GRADUAÇÃO EM BIOLOGIA COMPUTACIONAL E SISTEMAS

**Filodinâmica de Elementos
Transponíveis e seu uso no controle
genético de vetores de doenças
infecciosas**

Doutorando:

Felipe Soares FIGUEIREDO

Orientador:

Dr. Claudio STRUCHINER

Rio de Janeiro, Julho de 2012

Tese elaborada como parte dos requisitos para a obtenção do título de Doutor em Ciências no Programa de Biologia Computacional e Sistemas.

Folha destinada às assinaturas da banca

Figueiredo, Felipe .

Filodinâmica de Elementos Transponíveis e seu uso no controle genético de vetores de doenças infecciosas / Felipe Figueiredo. - Rio de Janeiro, 2012. xvi, 174 f.; il.

Tese (Doutorado) - Instituto Oswaldo Cruz, Pós-Graduação em Biologia Computacional e Sistemas, 2012.

Orientador: Claudio Struchiner.

Bibliografia: f. 127-139

1. modelo baseado em indivíduos. 2. simulação. 3. gene drive systems. 4. elementos transponíveis. 5. evolução molecular. I. Título.

Agradecimentos

Ao meu orientador Claudio J. Struchiner, que me proporcionou os desafios que precisei para me manter motivado, e ainda alguns tantos outros. Sua contribuição, na forma de conversas e correspondências não pode ser medida, mas seus insights e ideias peculiares certamente moldaram uma parte da forma como encaro a pesquisa acadêmica hoje em dia.

R. Daniela Medina, fonte constante de inspiração e motivação, excelentes discussões sobre o tema relativo a TEs ao longo do percurso. Um exemplo vivo de determinação e perseverança, e sobretudo amor ao estudo. Sempre que eu pensava que minha vida era difícil por causa de um projeto complicado e multifacetado, eu sempre imaginava a dificuldade de se ter dois filhos durante o doutorado. Perspectiva é fundamental nas crises.

Oswaldo Cruz, por suas longas e sempre interessantes conversas e dicas variadas durante as pausas para o café. Tecnófilos se reconhecem.

Aos alunos mais próximos do programa, primeiro colegas e posteriormente amigos, Bruno Manoel Silva, Rafael Cuadrat, Diogo Tschoeke, Marcelo Pontes e Anna Beatriz “Xuxu” Ferreira, que mesmo sem entender exatamente o que eu fazia, sempre tentaram me ajudar de uma forma ou outra.

Às amigas Alessandra Brandão, Amanda Sutter, Fábria Andréz, Joana Oliveira, Vanessa Koehler, que tanto me ajudaram nos momentos mais difíceis. Apoio emocional fundamental.

Bel, que pela maior parte do percurso acadêmico foi meu maior suporte emocional, e fonte de motivação frente às adversidades. Obrigado por ter me aturado, com todas as minhas idiossincrasias e complexidades por tantos anos. Sempre serei grato ao carinho.

Adriana, que nos momentos mais difíceis da reta final me proporcionou o equilíbrio mental e emocional para finalizar a jornada.

Aos colegas e amigos do PROCC/Fiocruz Rodrigo, Daniela, Luciane, Ronaldo, Mônica, Ernesto, Oswaldo, Elaine, Marília, Amanda, e Thaís e que acompanharam de perto, em maior ou menor grau essa longa e árdua jornada.

Aos amigos de longa data, Guilherme Laña, Gustavo Rios, Léo Mahfuz, André Luiz

“Cachorro”, e tantos outros que me ajudaram a espairecer nos momentos de estresse, e sempre estiveram presentes desde a adolescência. Amizades não se mantêm por mais de 15 anos sem razão.

Ao amigo madrugueiro, que me mantinha intelectualmente ativo com conversas interessantes, Fabrício Moura. Tantas discussões técnicas, filosóficas e intelectuais sobre política, guerras, religião, drogas, condicionamento físico, bodybuilding e video games, certamente me tornaram uma pessoa melhor, e mais informada, e com a constantemente desafiada.

Ao amigo Fábio Ramos, que sempre me manteve em perspectiva sobre a real dificuldade do projeto, e serviu de modelo de dedicação e exemplo sobre a possibilidade de se ser um profissional multitarefas, mesmo sob pressão e estresse extremos. A experiência com nossos “boosts” no mestrado foram fundamentais para a conclusão dessa etapa!

A Márcio Pavan e Fernando Monteiro, por me providenciarem a oportunidade de lecionar conceitos básicos de filogenética em cursos da PG em Biologia Parasitária, que tanto contribuiu para meus estudos e entendimento do assunto.

Ao professor Orlando B. Martins, da Bioquímica Médica da UFRJ, por ter me apresentado à bioinformática, e me proporcionado livre trânsito e farto aprendizado em seu laboratório de biologia molecular desde meu primeiro ano de graduação.

Ao meu professor de matemática no segundo grau, Sérgio Antinarelli, que sem saber me motivou a entrar na carreira acadêmica, por meio da Matemática. Sua calma diante das dificuldades e seu amor pelos estudos foram o exemplo e impulso que precisava no momento do vestibular.

À Márcia Verônica e Alessandra Portugal, secretárias da PG-BCS, que tornaram a burocracia suportável. Muito obrigado por tudo!

Sobretudo à minha família, a base de minha formação moral, e sempre me motivou a expandir meus horizontes intelectuais. Ter sido criado por uma família secular é algo que não tem preço nesta carreira que escolhi. Minha mãe Marilene, sempre presente, e disposta a amar. Minha irmã Talita, seu marido Tasso Marcelo e seu bebê Tiago, que sempre ajudam no que podem. Meu pai, principal coadjuvante da minha vida, sempre me proporcionou e motivou a ler e aprender sobre tudo, e disponibilizou todos os livros que podia ler e tantos outros que ainda prometo ler. . . que descanse em paz.

*And now the end is here
And so I face the final curtain
My friend I'll say it clear
I'll state my case of which I'm certain I've lived a life that's full
I traveled each and every highway
And more, much more than this
I did it my way*

*Regrets I've had a few
But then again too few to mention
I did what I had to do
And saw it through without exemption I planned each charted course
Each careful step along the byway
And more, much more than this
I did it my way*

*Yes, there were times, I'm sure you knew
When I bit off more than I could chew
But through it all, when there was doubt
I ate it up and spit it out
I faced it all and I stood tall and did it my way*

*I've loved, I've laughed and cried
I've had my fill, my share of losing
And now, as tears subside, I find it all so amusing
To think I did all that
And may I say, not in a shy way,
"Oh, no, oh, no, not me, I did it my way"*

*For what is a man, what has he got?
If not himself, then he has naught
To say the things he truly feels and not the words of one who kneels
The record shows I took the blows and did it my way!*

Yes, it was my way

Sumário

Lista de Figuras	xi
Lista de Tabelas	xiii
Lista de Abreviações	xiv
Resumo	xv
Abstract	xvi
1 Introdução	1
1.1 Controle genético de pragas	2
1.1.1 Técnicas clássicas	3
1.1.2 Novas propostas	4
1.2 Gene Drive Systems	4
1.3 Elementos Transponíveis - TEs	6
1.3.1 Ciclo e dinâmica de invasão dos TEs	8
1.3.2 Impacto evolutivo de TEs	10
1.4 Filogenética	13
1.4.1 Características gerais	13
1.4.2 Métodos de reconstrução de árvores	14
1.5 Coalescência	15
1.6 Filodinâmica	16
1.7 Simulações	19
1.7.1 Simulador	19
2 Justificativa	20
3 Objetivos	23
3.1 Objetivos gerais	24
3.2 Objetivos específicos	24

4	Metodologia	25
4.1	O simulador	26
4.1.1	A população de vetores	27
4.1.2	Modelos ecológicos	29
4.1.3	Estrutura do genoma	32
4.1.4	Representação de TEs	34
4.1.5	TEs de Classe I e II	35
4.1.6	Eventos de transposição	36
4.1.7	Dinâmica de transposição	37
4.1.8	Atividade e inativação de TEs	40
4.1.9	Eventos evolutivos	41
4.1.10	Rastreamento da origem das sequências	43
4.1.11	Algoritmo das gerações	44
4.2	Desenho experimental	45
4.2.1	Cenários de simulação	46
4.3	Simulações	48
4.3.1	Protocolo comum aos experimentos	49
4.3.2	Experimento MG	50
4.3.3	Experimento A	51
4.3.4	Experimento B	54
4.4	Análises dos dados simulados	57
4.5	Análises gráficas	58
4.5.1	Experimento A	58
4.6	Filodinâmica	60
4.6.1	Protocolo comum às topologias	61
4.6.2	Filogenética - componente temporal	62
4.6.3	Filogenética - componente demográfico	65
4.6.4	Tratamento das árvores obtidas	66
5	Resultados	67
5.1	Experimento MG	68
5.1.1	Topologias esperadas para um processo Master Gene	68
5.1.2	Neighbor-Joining	73

5.1.3	Considerações gerais	77
5.2	Experimento A	78
5.2.1	Sem custo de <i>fitness</i> (controle)	79
5.2.2	Com custo de <i>fitness</i>	90
5.3	Experimento B	104
5.3.1	Sem custo de <i>fitness</i>	104
5.3.2	Com custo de <i>fitness</i>	115
6	Discussão	116
6.1	Do simulador	117
6.2	Dos experimentos	117
6.2.1	Experimento MG	117
6.2.2	Experimento A	118
6.2.3	Experimento B	120
6.2.4	Modelagem matemática	121
6.3	Perspectivas futuras	123
7	Conclusões	125
8	Referências Bibliográficas	127
A	Manuscritos de artigos	140
A.1	An individual-based model forward simulator for transposable elements dynamics and evolution	140
A.1.1	Manuscrito principal	140
A.1.2	Material suplementar (SOM)	143
A.2	Phylogenetics of Transposable Elements	160
A.2.1	Manuscrito principal	160
A.2.2	Material suplementar (SOM)	161
	Índice Remissivo	173

Lista de Figuras

1.1	Diferenças fenotípicas em milho devidas a atividade de TEs	7
1.2	Mecanismos de transposição de TEs Classes I e II	8
1.3	Mecanismo de expansão dos TEs em populações sexuadas	9
1.4	Ciclo de vida e fixação de uma família de TEs	10
1.5	Impacto da invasão de TEs em uma população hospedeira	11
1.6	Diagrama do modelo de coalescência	15
1.7	Filogenias de vários patógenos	17
1.8	Filogenias esperadas para vários cenários epidemiológicos	18
4.1	Protocolo de simulações do experimento MG	51
4.2	Protocolo de simulações do experimento A com <i>template</i> contendo 1 TE e sem impacto no <i>fitness</i>	52
4.3	Protocolo de simulações do experimento A com <i>template</i> contendo 20 TEs e sem impacto no <i>fitness</i>	53
4.4	Protocolo de simulações do experimento A com <i>template</i> contendo 1 TE e impactos fraco e moderado no <i>fitness</i>	54
4.5	Protocolo de simulações do experimento A com <i>template</i> contendo 20 TEs e impacto moderado no <i>fitness</i>	55
4.6	Protocolo de simulações do experimento B com população de mosquitos constante	56
4.7	Protocolo de simulações do experimento B com população de mosquitos exponencial decrescente	57
4.8	Protocolo de simulações do experimento B com população de mosquitos exponencial crescente	58
4.9	Protocolo de simulações do experimento B com população de mosquitos logística	59
4.10	Protocolo de simulações do experimento B com população de mosquitos constante, com custo de <i>fitness</i>	60
5.1	Experimento MG: $u = 1$, 1 MG	69
5.2	Experimento MG: $u = 2$, 1 MG	70
5.3	Experimento MG: $u = 1$, 2 MG	71

5.4	Experimento MG: $u = 2$, 2 MG	72
5.5	Experimento MG: $u = 1$, 1 MG, NJ	73
5.6	Experimento MG: $u = 2$, 1 MG, NJ	74
5.7	Experimento MG: $u = 1$, 2 MG, NJ	75
5.8	Experimento MG: $u = 2$, 2 MG, NJ	76
5.9	Experimento A, $template = 1$, $s = 0$, demografia	80
5.10	Experimento A, $template = 1$, $s = 0$, TEs	81
5.11	Experimento A, $template = 1$, $s = 0$, demografia de populações pequenas	84
5.12	Experimento A, $template = 20$, $s = 0$, demografia	87
5.13	Experimento A, $template = 20$, $s = 0$, TEs	88
5.14	Experimento A, $template = 1$, $s = 0,01$, demografia	91
5.15	Experimento A, $template = 1$, $s = 0,01$, TEs	92
5.16	Experimento A, $template = 1$, $s = 0,05$, demografia	95
5.17	Experimento A, $template = 1$, $s = 0,05$, TEs	96
5.18	Experimento A, $template = 1$, $s = 0,05$, demografia de populações pequenas	98
5.19	Experimento A, $template = 20$, $s = 0,05$, demografia	100
5.20	Experimento A, $template = 20$, $s = 0,05$, TEs	101
5.21	Experimento B: Árvore de população constante	106
5.22	Experimento B: Árvore de população exponencial (crescente)	107
5.23	Experimento B: Árvore de população logística	108
5.24	Experimento B: Árvore de população exponencial (decrecente)	109
5.25	Experimento B: Skyline plot linear - população constante.	110
5.26	Experimento B: Skyline plot linear - população exponencial (crescente)	111
5.27	Experimento B: Skyline plot linear - população exponencial (decrecente)	111
5.28	Experimento B: Skyline plot linear - população logística	112
5.29	Experimento B: Skyline plot linear - 4 modelos	113

Lista de Tabelas

4.1	Protocolo das topologias do experimento MG	64
4.2	Protocolo das topologias do experimento B	65
4.3	Protocolo das demografias do experimento B	66
5.1	Exp A: Quantidades de TEs ($template = 1, s = 0$)	82
5.2	Exp A: Quantidades de TEs ($template = 20, s = 0$)	86
5.3	Exp A: Quantidades de TEs ($template = 1, s = 0,01$)	90
5.4	Exp A: Quantidades de TEs ($template = 1, s = 0,05$)	97
5.5	Exp A: Quantidades de TEs ($template = 20, s = 0,05$)	102
5.6	Experimento B: Resultados para cenários com $s = 0,01$, e diversos modelos populacionais	115

Lista de Abreviações

BI	Bayesian Inference (Inferência bayesiana)
CPU	Central Processing Unit
DDT	Dicloro-Difenil-Tricloroetano
DOA	Dead on arrival
ESS	Effective Sample Size
FK	Female Killing
GM	Genetically Modified (subpopulação transgênica simulada)
GMI	Genetically Modified Insect
HERV	Human endogenous retrovirus
IBM	Individual Based Model
IRS	Indoor residual spraying
MCC	Maximum Clade Credibility
MCMC	Markov Chain Monte-Carlo
MEDEA	Maternal Effect Dominant Embryonic Arrest
MG	Master Gene
MRCA	Most recent common ancestor
NJ	Neighbor-Joining
SIT	Sterile Insect Technique
TE	Transposable Element

Resumo

As doenças transmitidas por mosquitos têm grande custo de vidas e socio-econômico, especialmente em países tropicais em desenvolvimento, e por isso é uma prioridade para a Organização Mundial da Saúde. Novas propostas para o controle destas doenças incluem a modificação genética dos vetores e para isso, além da identificação e inserção de genes de resistência ao patógeno no mosquito, é necessária a obtenção de métodos eficientes para difundir e fixar tais transgenes nas populações naturais. O uso de elementos transponíveis (TEs) tem sido proposto como mecanismo de propagação devido a suas características egoístas e capacidade de invasão em populações inteiras. Nesta tese examinamos modelos matemáticos sob a ótica de simulações dos fenômenos ecológicos e evolutivos envolvidos nos processos de invasão de uma população selvagem de mosquitos por uma família de TEs que carregue um gene de resistência que confira refratoriedade contra o patógeno de uma doença transmissível. Elaboramos as premissas de um estudo recente que adaptou técnicas de filodinâmica usadas com sequências de vírus para sequências de TEs, e como essa analogia pode ser usada para estimar o tempo de invasão de TEs em uma população de mosquitos. Foi desenvolvido um simulador *de novo* baseado em indivíduos capaz de representar três níveis de organização biológica: a população dos mosquitos, as quantidades e *loci* dos elementos e as sequências individuais que sofrem mutações ao longo das gerações. Exploramos tanto a influência do custo de *fitness* dos TEs como a influência de diferentes dinâmicas populacionais nas quantidades totais de elementos por indivíduo e na população, tomando como base uma família de TEs que se expande de acordo com um modelo *master gene*. Observamos que topologias reconstruídas de uma família com essa característica exibem as estruturas pectinadas previstas na literatura teórica, e em casos simples, o tempo entre eventos de transposição pode ser observado graficamente na árvore. Mostramos também o conflito entre a taxa de transposição, a perda de TEs ativos e o impacto individual no fitness do hospedeiro, e como essas grandezas devem ser consideradas em conjunto para futuros estudos. Mostramos que ao fazer a filodinâmica com sequências de TEs, é possível observar a influência da demografia dos hospedeiros na estimativa da população dos TEs.

Abstract

Mosquito-borne diseases present a severe socio-economical burden for tropical developing countries, and thus is a high priority for the World Health Organization. New proposals for the control of these diseases include the genetic modification of the vectors that transmit the pathogen, and to achieve this goal, not only it is required to identify and insert genes that inhibit the pathogen cycle in the mosquito, but also to drive these genes into whole wild populations. The use of transposable elements (TEs) has been suggested as a possible gene drive system due to its selfish behaviour and ability to invade whole populations. In this thesis we revisit mathematical models by the use of computer simulation of ecological and evolutionary processes involved in the invasion of a mosquito population by a family of TEs that carries a resistance gene against the pathogen of a transmitted disease. We elaborated on premises from a recent study that adapted phylogenetics techniques usually employed with virus sequences to TE data, and how this analogy can be used to estimate the time since invasion in a mosquito population. A *de novo* individual-based computer simulation model was developed, capable of representing three levels of biological organization: the mosquito population, the quantities and *loci* of TEs and the individual DNA sequences that mutate over generations. We explored both the influence of the fitness cost from TEs and the influence of different population dynamics in the total quantities of TEs in individuals and in the whole population, considering a master gene model. We observed that phylogenies reconstructed from such a TE family meet theoretical properties predicted in the literature. Our simulations also show the compromise between the transposition rate, the loss of active TEs and the fitness impact in the host and how these variables must be jointly considered in future studies. We show that when phylogenetics is applied to TE sequences, the influence of the host demography can be observed in the estimated TE population.

Palavras chave: Modelo baseado em indivíduos, simulação, gene drive systems, elementos transponíveis, evolução molecular, filodinâmica, modelos de populações.

Capítulo 1

Introdução

“(...) it is necessary that the reasoner should be able to utilise all the facts which have come to his knowledge; and this in itself implies, as you will readily see, a possession of all knowledge, which, even in these days of free education and encyclopaedias, is a somewhat rare accomplishment. It is not so impossible, however, that a man should possess all knowledge which is likely to be useful to him in his work, and this I have endeavoured in my case to do. If I remember rightly, you on one occasion, in the early days of our friendship, defined my limits in a very precise fashion.”

Sherlock Holmes in “The Five Orange Pips” (1892)

1.1 Controle genético de pragas

Doenças transmitidas por vetores têm fortes efeitos negativos na saúde humana, agricultura e pecuária. Com a melhora das condições sanitárias nos últimos séculos, doenças como a peste bubônica, foram mitigadas ou erradicadas em diversas áreas do mundo. Adicionalmente, avanços no estudo e produção de vacinas também erradicaram doenças como varíola em muitos países.

No entanto não existem vacinas disponíveis para todas as doenças conhecidas. Na ausência de uma vacina, o controle das doenças transmitidas por vetores é feito principalmente com o controle da população desse vetor. O ciclo da malária, por exemplo, depende tanto da presença do vetor, um mosquito do gênero *Anopheles*, quanto do parasito, um protozoário do gênero *Plasmodium*. Para isso, medidas são tomadas para reduzir o tamanho da população do vetor, o que implica numa redução na taxa de encontros entre o vetor (mosquito) e o hospedeiro (humano). Campanhas de redução e controle de mosquitos foram bem sucedidos na erradicação da malária na maior parte dos países desenvolvidos, e fora das zonas tropicais. Infelizmente esta doença ainda persiste com epidemias frequentes e em alguns casos de forma endêmica na maior parte da África e alguns países da Ásia.

De acordo com a Organização Mundial da Saúde (OMS), houve algo em torno de 216 milhões de casos de malária, implicando em cerca de 665.000 mortes em 2010 [1]. As taxas de mortalidade por malária ao redor do mundo caíram mais de 25% desde o ano de 2000, e mais de 33% no continente africano. A maior parte das mortes ocorre entre crianças na África, onde uma criança morre a cada minuto de malária e a doença provoca cerca de 24% das causas de mortalidade infantil.

A OMS recomenda que larvas e mosquitos adultos sejam eliminados e repelidos através do uso de inseticidas no interior de moradias. Existem 12 tipos de inseticidas sugeridos para pulverização residual em domicílios (*IRS = indoor residual spraying*, incluindo-se o DDT (*Dicloro-Difenil-Tricloroetano*) apesar de sua proibição na Convenção de Estocolmo¹. Além da pulverização nas paredes e telhados, o número de picadas por noite pode ser reduzido com a utilização de telas e mosquiteiros tratados com inseticidas. O uso desses métodos, apesar de eficazes a curto prazo, perdem sua eficiência a longo prazo devido aos mecanismos adaptativos que originam resistências aos inseticidas por parte

¹http://www.who.int/malaria/publications/atoz/who.htm_gmp_2011/en/index.html

dos vetores [2]. Nos casos de doenças para as quais há vacinas disponíveis, campanhas de vacinação maciça também tem um bom desempenho a médio prazo. Contudo, analogamente ao caso dos inseticidas, estas também criam uma pressão seletiva no parasito, e se a campanha não obtiver abrangência de 100% da população alvo, os parasitos remanescentes podem desenvolver mecanismos de evasão da pressão causada pelo sistema imunológico do hospedeiro [3]. Sendo assim, faz-se necessário que esses métodos sejam constantemente revistos e atualizados, como também ocorre com inseticidas.

Os mosquitos do gênero *Anopheles*, possuem uma resistência natural ao plasmódio causador da malária e por isso quando considerados individualmente, são em geral vetores ineficientes do parasito [4]. O estudo de genes específicos que garantam resistência do vetor ao patógeno objetiva a criação de mosquitos geneticamente modificados (**GMI**s = **Genetically Modified Insects**) [5, 6]. Especula-se que seja possível explorar esse mecanismo natural de resistência, e difundi-lo de maneira programada para toda a população do mosquito em um dado ambiente, de modo a interromper o ciclo de transmissão da doença quando o parasito entra em contato com o vetor [7]. A criação de um mosquito transgênico que seja incapaz de atuar como vetor também é foco de muita discussão na comunidade científica [8, 9], e tem-se observado sucesso na obtenção de cepas em laboratório refratárias ao patógeno [10, 11].

1.1.1 Técnicas clássicas

A ideia de se modificar geneticamente insetos para o controle de pragas não é nova. O pesquisador russo Alexander Serebrovskii propôs uma abordagem teórica baseada em translocações cromossômicas nos anos 1940, mas foi impedido de fazer seus experimentos pela Segunda Guerra Mundial [8, 12]. Entre os anos 1940 e 1950 os pesquisadores Knipling e Bushland, desenvolveram a técnica SIT (*Sterile Insect Technique*) [13], que consiste em criar mutações genéticas nos gametas dos machos de modo a produzir progênia estéril [8]. Para isso, os insetos são expostos a radiação em doses controladas e posteriormente liberados no ambiente para se reproduzir livremente com as fêmeas silvestres.

Embora campanhas de redução de populações de insetos que causam danos na agricultura e pecuária tenham tido sucesso no passado em ilhas e áreas geograficamente isoladas, o controle genético de pragas não está restrito à tentativa de extinção da população. De fato, em muitos casos não é interessante, viável ou mesmo ético extinguir toda uma po-

pulação ou espécie de vetores. A erradicação de uma população de insetos pode ter impactos secundários difíceis de serem previstos ou mesmo mensurados [5].

1.1.2 Novas propostas

Uma proposta alternativa à erradicação de vetores de doenças, que vem ganhando tração na literatura na última década é a de **substituição de população** (*population replacement*) [6, 14]. Nesse contexto, a população silvestre deve ser substituída por uma variante menos danosa, ou incapaz de se contaminar com o parasito, evitando assim, a transmissão da doença [8].

Existem diversas propostas de estratégias de substituição de população, como veremos na próxima seção. Dado que um mosquito refratário ao patógeno seja produzido em laboratório por intermédio de um transgene que confira resistência ou imunidade ao parasito[15], um dos próximos passos a ser investigado é a propagação desse transgene numa população silvestre. Sabe-se no entanto que mosquitos de laboratório têm *fitness* reduzido quando comparado aos mosquitos silvestres [16, 17]. Mesmo que os mosquitos de cativeiro competissem igualmente com os mosquitos silvestres, a quantidade de indivíduos de cativeiro seria tão diminuta em relação à população silvestre que o gene introduzido estaria sujeito a fortes efeitos de deriva genética. Dessa forma, seria pequena a probabilidade de que um gene exógeno inserido em mosquitos de cativeiro vá se fixar na população silvestre, considerando apenas o mecanismo mendeliano. A fim de que haja a substituição da população, é necessário portanto um mecanismo eficiente em propagar e fixar o transgene na população a despeito da pouca competitibilidade dos mosquitos transgênicos em relação aos mosquitos silvestres [8, 18, 19].

1.2 Gene Drive Systems

O objetivo de propagar um gene que confira refratoriedade a um patógeno, numa população natural de insetos necessita de um mecanismo mais eficiente do que o acaso que segrega um dado gene na taxa de 50% dos gametas produzidos pela população original. Como a quantidade de mosquitos transgênicos que pode ser produzida em laboratório é muito pequena em comparação ao tamanho esperado da população silvestre, o transgene sujeito às condições mendelianas seria perdido em poucas gerações por deriva genética

[19, 20].

Mecanismos que satisfazem esse critério são chamados *Genetic Drive Mechanisms*, [5]. Um **Gene Drive System** é um sistema que utiliza um mecanismo de *drive* genético para replicar o gene de interesse a uma taxa maior do que este é perdido [20, 21]. James (2005, [5]) define o *drive system* como o produto sintético que corresponde ao *drive* genético, enquanto o *drive mechanism* é o sistema biológico subjacente incorporado no sistema.

Um gene ou elemento genético é chamado “egoísta” quando ele é capaz de se propagar e se fixar na população contrariando a premissa da seleção natural, isto é, ele se expande sem conferir vantagem evolutiva ao hospedeiro, de forma análoga ao comportamento de um parasito [22, 23]. Vários tipos de mecanismos egoístas que se qualificam nessa categoria por serem eficientes na fixação de determinados genes ou alelos na população em poucas gerações, geralmente possuem um custo em termos de *fitness* da espécie. Alguns mecanismos promovem o suicídio programado de filhotes que não possuam o gene de interesse, enquanto outros modificam a razão sexual da população, minando a capacidade reprodutiva dos indivíduos. Uma breve lista de exemplos de mecanismos candidatos a *gene drive* inclui:

- *Female Killing* (FK) - morte de fêmeas [12],
- *Maternal Effect Dominant Embryonic Arrest* (MEDEA) [24],
- *Wolbachia* – uma bactéria endossimbionte que é transmitida pela linhagem materna, [14, 25],
- *Meiotic Drive* - alteração progressiva da proporção entre machos e fêmeas da população [24, 26, 27].
- Elementos Transponíveis (TEs) [28–31] - sequências de DNA com capacidade auto-replicativa e mecanismo de propagação egoísta.

Este último item corresponde ao mecanismo de interesse desta tese, que será descrito na próxima seção.

É importante salientar que a introdução de novos genes em um organismo, mesmo por um mecanismo natural, geralmente influencia negativamente seu sucesso reprodutivo, seja fenotipicamente na competição pela atenção do sexo, ou em sua fecundidade [15, 32]. Esses mecanismos egoístas, no entanto, são capazes de superar essa pressão seletiva e invadir

populações inteiras, migrar para populações em áreas vizinhas espalhando-se por diversas regiões geográficas e até colonizar toda uma espécie, desde que esse impacto negativo se mantenha menor que a capacidade reprodutiva do organismo. Um exemplo dessa capacidade de invasão pandêmica é a invasão por parte do elemento transponível conhecido como elemento P em populações de *Drosophila melanogaster* ao redor do mundo após o início do monitoramento de sua atividade [33, 34]. Isso é particularmente importante pois GMIs tipicamente têm indicadores de *fitness* desvantajosos, sendo mais frágeis, com tamanho corporal, longevidade ou reservas lipídicas comparativamente menores em relação aos espécimes selvagens [15, 16, 19, 32, 35].

1.3 Elementos Transponíveis - TEs

Hartl e Clark (1998, [36]) definem elementos transponíveis (**TEs = Transposable Elements**) como sequências de DNA capazes de mudar sua posição num genoma, inserindo-se em novas posições nos cromossomos e assim aumentando seu número ao longo de gerações. Desde a sua identificação nos anos 1950 por Barbara McClintock no genoma do milho [37] (cf. figura 1.1, [38]), TEs tem sido alvo de vários estudos para identificar suas funções e os resultados de suas inserções [39, 40]. Estas sequências podem se inserir em regiões codificantes, inativando um gene, ou se inserir em regiões promotoras e regulatórias, alterando o padrão de expressão gênica do indivíduo. Além disso, pares de elementos podem agir de maneira a criar recombinações e criar rearranjos cromossômicos, sendo portanto agentes de variação genética considerável nas populações hospedeiras.

Em 2007, Wicker *et al.* [41], propuseram uma nova classificação universal para os diversos tipos de elementos móveis identificados quanto a suas estruturas genéticas e mecanismos de replicação. Os mecanismos que determinam a mobilidade dos TEs no genoma geralmente criam novas cópias desses elementos mantendo a sequência original em sua posição, aumentando assim o número de cópias do elemento. Existem vários mecanismos descritos para essa mobilização, (cf. figura 1.2 adaptada de [42]) e os TEs são organizados em duas classes, Classe I e Classe II quanto à sua estrutura e mecanismo intermediário de transposição. Os TEs da Classe I, ou retro-transposons, usam uma transcriptase reversa como mecanismo intermediário de transposição similarmente ao usado por retrovírus para inserir seu material genético no genoma hospedeiro. TEs Classe II se transpõem usando uma DNA transposase, que excisa o elemento de uma das cadeias de DNA e o insere em

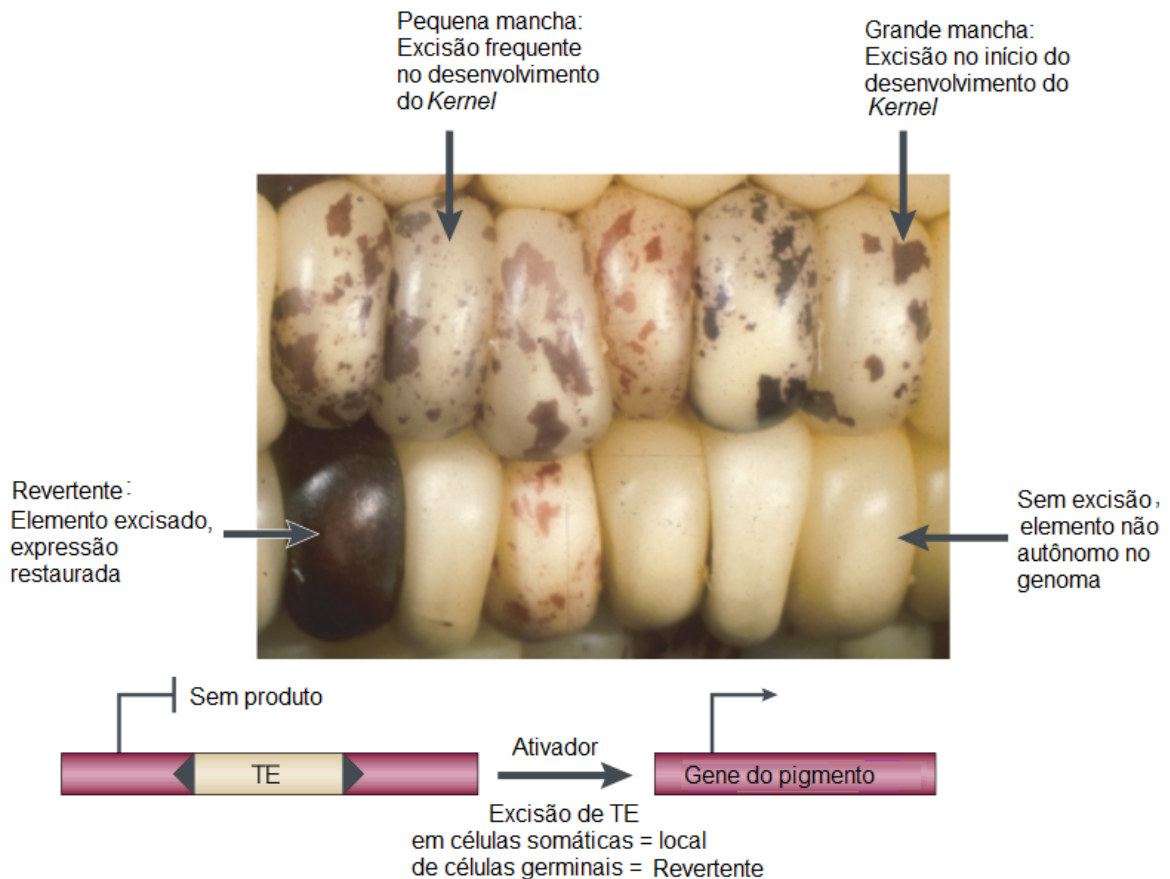


Figura 1.1: Diferenças fenotípicas em milho devidas a atividade de TEs, ilustrando sua descoberta por McClintock em 1950. Fonte: Feschotte, *et al.*, (2002)

outra posição do cromossomo. Estes se subdividem em duas subclasses, I e II. Nesse processo de transposição os TEs da subclasse I são inteiramente removidos do cromossomo, e transportados para outra região. O mecanismo usado pelos TEs da subclasse II só remove a cópia presente em uma das cadeias de DNA, mantendo a informação original na cadeia complementar. Posteriormente o mecanismo de reparo nuclear reconstrói a informação removida, restaurando o elemento *template*. Assim os dois mecanismos descritos para Classe I e II são comumente referidos na literatura por *copy and paste* e *cut and paste* respectivamente.

Por conta desse processo replicativo, TEs contribuem com grandes quantidades de material genético nos genomas hospedeiros, totalizando porções consideráveis na maioria

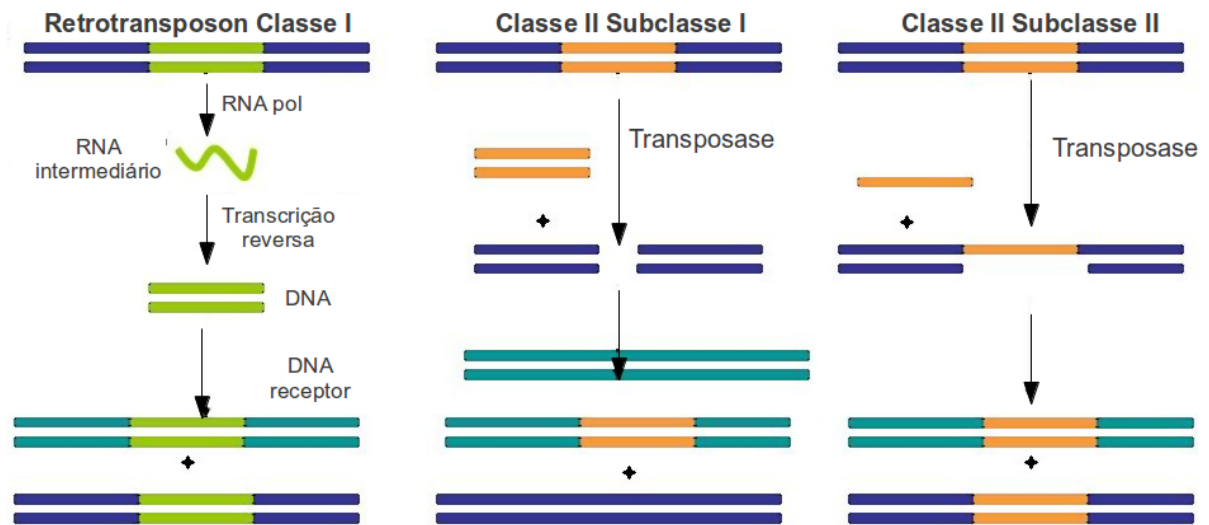


Figura 1.2: Mecanismos de transposição de TEs Classes I e II (adaptado de Fernandez-Medina, 2009)

dos genomas colonizados [30, 43, 44]. Os TEs que se acumulam nos genomas da espécie hospedeira lá ficam até que mutações deletérias descaracterizem seus sítios específicos para o reconhecimento da transcriptase reversa (no caso dos TEs Classe I) ou da transposase (TEs Classe II) inativando o elemento. A partir desse momento, a sequência remanescente passa a evoluir neutralmente, sem nenhum efeito de seleção. Como esses processos de inativação e degeneração ocorrem independentemente para cada elemento, o conjunto dessas sequências representa um retrato da história da invasão desse elemento na espécie hospedeira [31].

1.3.1 Ciclo e dinâmica de invasão dos TEs

Após a inserção de um ou mais elementos em uma nova população hospedeira por transferência horizontal, tem início um novo processo de invasão com algumas etapas características [45, 46]. Os TEs tem a capacidade de se replicar e se inserir em novas regiões nos cromossomos (figura 1.3 [28]) aumentando o número de cópias representadas nos gametas, em relação aos cromossomos parentais.

Se a taxa de transposição não for suficientemente elevada, a família de TEs pode ser eliminada pela maquinaria nuclear, ou perda da população por deriva. Por outro lado, a presença dos TEs cria um gasto energético para sua replicação e manutenção no

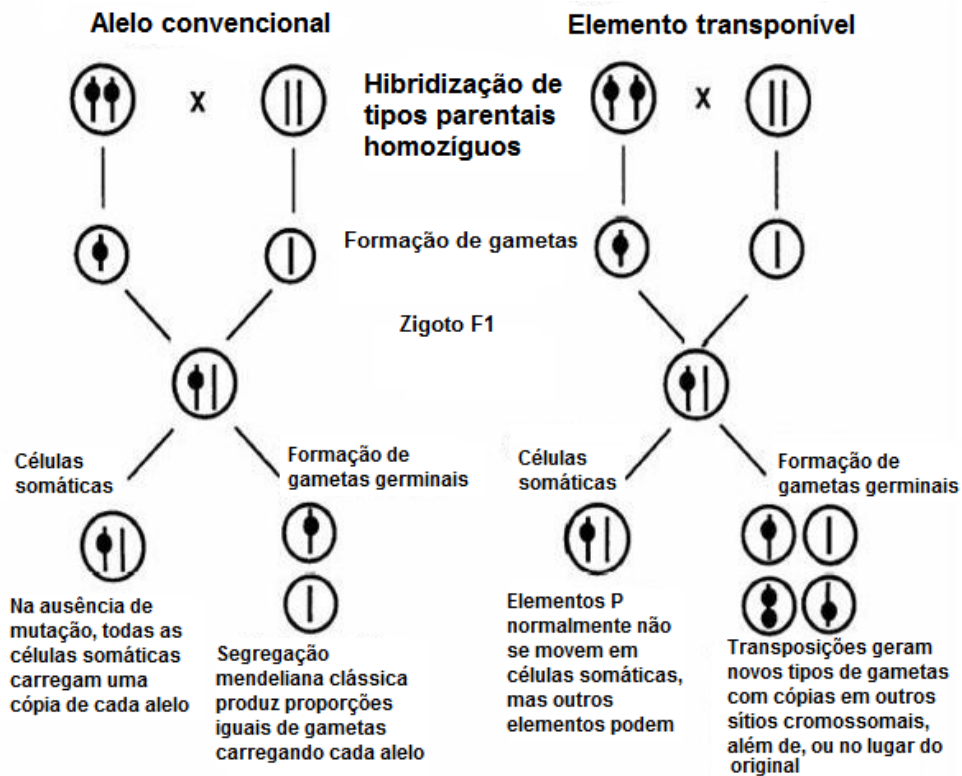


Figura 1.3: Mecanismo de expansão dos TEs em populações sexuadas (adaptado de Kidwell, 1992)

genoma, e o crescimento não-regulado desses elementos egoístas pode levar a espécie hospedeira à extinção. Existem diversos mecanismos propostos para o estudo da regulação do número de cópias de TEs ao invadir uma nova população hospedeira, incluindo mecanismos intrínsecos aos TEs, e extrínsecos, relativos à população de hospedeiros [47, 48]. Algumas famílias de TEs parecem ser capazes de autorregulação, proporcionando taxas de transposição cada vez menores conforme o número de TEs presente se aproxima de um equilíbrio [49, 50]. Por outro lado, a inserção de novos elementos genéticos oferece risco de mutações deletérias ao hospedeiro que são eliminadas por seleção negativa.

A figura 1.4 ilustra as diversas etapas do ciclo de um TE típico. Após a transferência horizontal para uma nova espécie, é necessária uma fase de amplificação, onde o número de cópias do elemento colonizador aumenta de modo a popular os cromossomos de todos os gametas, aumentando suas chances de transmissão vertical na linhagem do hospedeiro. Conforme esse processo se itera ao longo de algumas gerações, há a expansão do número de

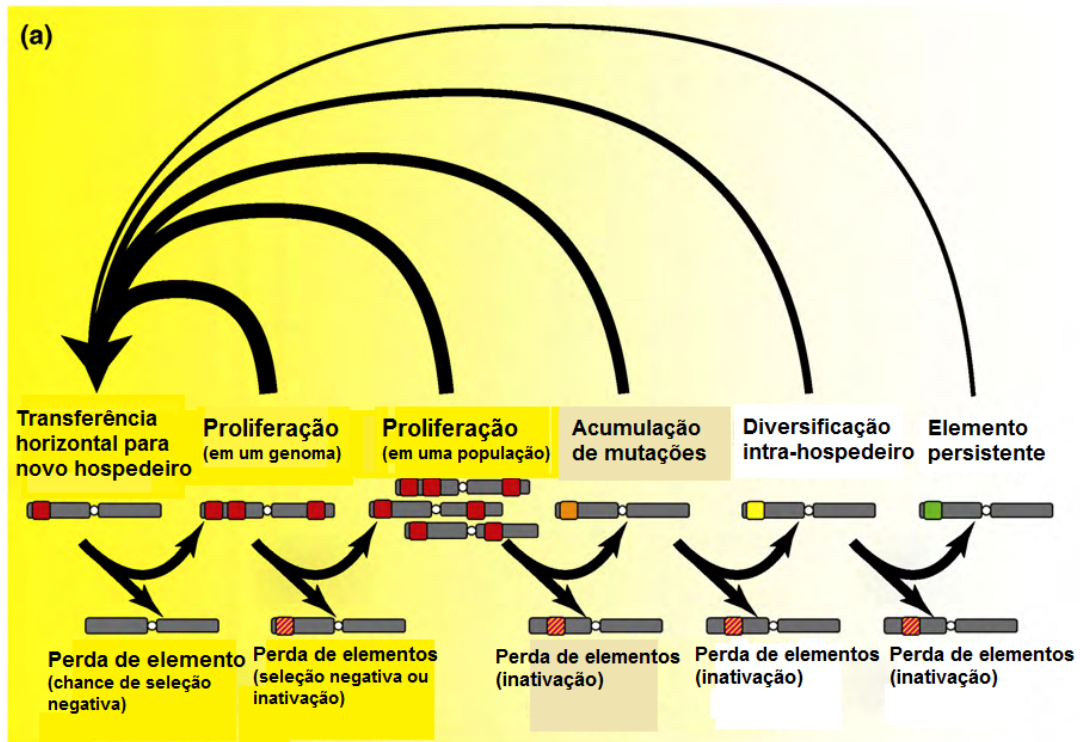


Figura 1.4: Ciclo de vida e fixação de uma família de TEs (adaptado de Schaack, 2010)

indivíduos infectados na população, até o ponto da invasão de toda a população. Em todas essas etapas, novas cópias são criadas ou perdidas aleatoriamente. Conforme o tempo passa, mutações se acumulam nos elementos que se fixaram no genoma da população e com isso alguns elementos se tornam inativos para novas transposições.

1.3.2 Impacto evolutivo de TEs

O acúmulo e perda de TEs em genomas hospedeiros proporciona os mais variados impactos mutacionais. Esses elementos que originalmente eram considerados “DNA lixo”, e “parasitos genômicos” devido a seu mecanismo egoísta de replicação e ausência de função aparente, passaram a gerar mais interesse nas últimas décadas com a disponibilização de genomas completos de insetos como *D. melanogaster* [51], *Anopheles gambiae* [52] e *Aedes aegypti* [53].

Além de ser encontrados em diversas espécies de organismos de todos os reinos e níveis de complexidade, observou-se que algumas famílias de TEs se apresentam simul-

taneamente em espécies próximas, mas diversos estudos apontam para invasões recentes, cujos tempos estimados são várias ordens de grandeza menores que os tempos envolvidos nos processos de especiação [54]. Sabe-se que os TEs são capazes de invadir novas espécies principalmente por transferência horizontal [55]. Embora a transferência horizontal não explique todas as disparidades encontradas entre as filogenias das espécies e dos TEs [56], esta permanece como uma das principais hipóteses para a presença de famílias de TEs próximas ou semelhantes estarem fixadas em espécies distantes. Por exemplo, um estudo recente indica que apesar de os mosquitos dos gêneros *Aedes* e *Anopheles* terem divergido há cerca de 145-200 milhões de anos, ambos compartilham uma família de TEs (da superfamília *mariner*) com 99% de identidade, o que suporta a hipótese de uma transferência horizontal recente [57].

Diversos estudos recentes mostram que aparecimento de novos genes, reorganização cromossomal e mutações fixadas são atribuídos à atividade de TEs [43, 58–62]. Quando dois TEs se encontram próximos, ambos são movidos simultaneamente, e junto com eles toda a região que os separa. Assim genes inteiros podem ser transportados por pares de TEs. De acordo com Schaack (2010 [63]), esse mecanismo é um veículo comum para transferência de genes entre espécies de procaríotos, e embora nunca tenha sido demonstrado que TEs transfiram genes entre espécies de eucariotos, eles são capazes de capturar e transportar sequências com grande frequência dentro de uma mesma espécie.

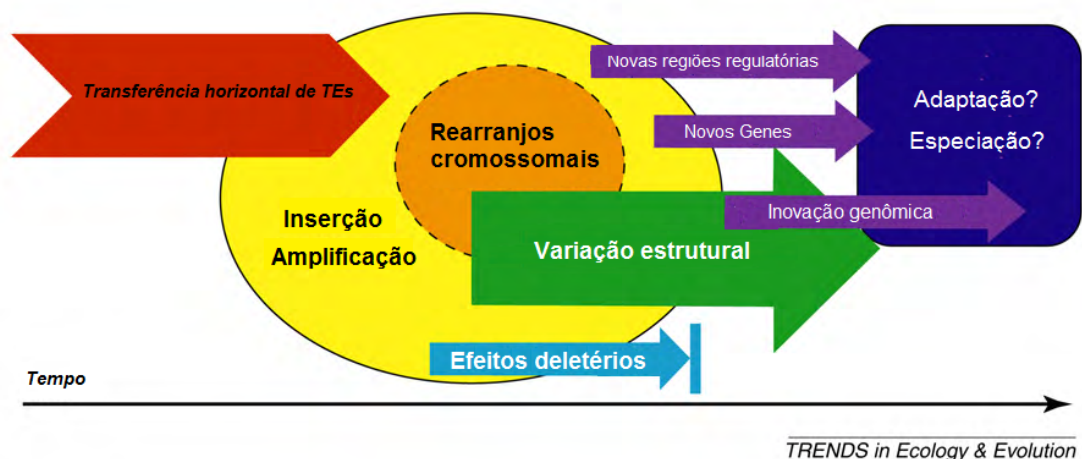


Figura 1.5: Impacto da invasão de TEs em uma população hospedeira. Fonte: Schaack (2010)

Além disso, especula-se que os rearranjos genômicos causados por TEs posteriormente

a uma invasão e colonização em eucariotos tenham um papel relevante na origem de novas espécies de modo que estes têm sido considerados agentes chaves da evolução [44, 64–69]. A figura 1.5, retirada de [63], mostra um diagrama de alguns efeitos evolutivos hoje atribuídos à atividade de TEs. Schaack *et al.* (2010 [63]) apontam que, além de os mecanismos de transferência horizontal serem importantes para a persistência de TEs e outros elementos egoístas a despeito de não oferecerem vantagem óbvia ao hospedeiro, é notável que uma invasão tenha coincidido com um dos eventos de especiação documentado em mamíferos, o morcego *Myotis lucifugis* [70, 71]. Outro exemplo interessante do impacto da atividade de TEs é a identificação de um gene quimérico em primatas como consequência da atividade de TEs, aparentemente originado da fusão entre um gene SET preexistente na linhagem dos primatas e o gene da transposase do elemento *Hsmar1*, um TE com características semelhantes ao *mariner* [58].

Ecologia do genoma

O termo **ecologia do genoma** foi proposto em 1997 por Kidwell e Lisch [40] para representar a complexa interação entre esses elementos e o genoma hospedeiro. Diversos estudos indicam interações positivas e negativas que podem ser interpretadas de maneira análoga às interações típicas de populações. Exemplos incluem competição entre diferentes famílias de TEs e o parasitismo entre TEs da mesma família quando há presença de elementos autônomos (capazes de expressar a transposase necessária para sua replicação) e elementos não-autônomos (elementos que perderam total ou parcialmente esse gene, mas ainda são capazes de ser transpostos pela transposase de um TE autônomo [46, 59, 72]. Esses mecanismos formam redes complexas de interações, e embora avanços tenham sido feitos nos últimos anos para entendê-las, nem todos os mecanismos são conhecidos [73].

Estudos de história evolutiva de TEs

Estudos de comparação de genomas inteiros têm se mostrado muito eficientes para análises filogenéticas de várias espécies e têm tido seu poder explicativo aumentado em vista dos avanços recentes nas tecnologias de sequenciamento e tamanhos sempre crescentes das bases de dados genômicas [74]. Alguns desses estudos têm focado direta ou parcialmente a busca e identificação de TEs ou análise da sua história evolutiva [56, 75].

Embora a origem dos TEs como um todo seja desconhecida, em alguns casos é possível

associar sua origem com vírus e retrovírus que perderam a capacidade de se mobilizar para fora da célula, como por exemplo os HERVs (*Human endogenous retrovirus*) [76, 77], que compõem cerca de 5-8% do genoma humano [78].

1.4 Filogenética

1.4.1 Características gerais

Reconstruções de árvores filogenéticas tipicamente são constituídos por três pilares principais: um modelo demográfico, e um modelo de relógio molecular, e um modelo de evolução molecular [79].

Componente molecular - modelo evolutivo

Um modelo de evolução determina que tipos de mutações ocorrem e as proporções relativas com que ocorrem cada tipo; exemplos comuns incluem **Jukes-Cantor** [80] (todas as mutações ocorrem com a mesma frequência), **Kimura 2 parâmetros** [81] (mutações do tipo transição ocorrem com frequência diferente de mutações do tipo transversão), e modelos com mais parâmetros, como **HKY** [82] e **GTR** [83] (cada mutação possível tem um parâmetro independente dos outros).

Componente demográfico

Um modelo demográfico determina o tamanho da população e como este varia com o tempo. Muitos métodos de reconstrução assumem uma população infinita, o que simplifica a modelagem matemática e torna os cálculos tratáveis analiticamente, enquanto alguns permitem a escolha de um modelo populacional paramétrico. Exemplos comuns de modelos paramétricos de populações finitas são modelos de **população constante**, modelo de **expansão exponencial** e **modelo logístico**.

Componente temporal

A taxa de mutação relativa entre um grupo de espécies relacionadas pode variar ao longo do tempo evolutivo, devido a fatores ambientais ou inerentes às espécies.

Um modelo de relógio molecular determina como a taxa de mutação varia com o tempo; exemplos incluem o **relógio estrito** (todas as sequências envolvidas divergem umas das outras com mesma taxa) e **relógios relaxados** [84] (taxas podem variar ao longo da árvore de relações filogenéticas).

1.4.2 Métodos de reconstrução de árvores

Existem diversos métodos para reconstrução de árvores filogenéticas, que assumem premissas aplicáveis a conjuntos de dados e cenários específicos, sendo comumente divididos entre métodos de distância, e métodos de busca de árvores (*tree searching*) [85, 86].

Métodos de distância estimam a distância evolutiva entre os *taxa* par a par, de acordo com uma métrica predeterminada. As distâncias dos pares de *taxa* são organizados em uma **matriz de distâncias**, que é então utilizada para o agrupamento hierárquico dos *taxa*. Exemplos de métodos de distância incluem UPGMA e *Neighbor-Joining* [87].

Métodos de buscas de árvores percorrem o espaço de árvores em busca de uma árvore que satisfaça algum critério de otimização. Exemplos incluem o método de Máxima Parcimônia, que agrupa os *taxa* de modo a minimizar a quantidade de mutações necessária para representar a árvore, e os métodos baseados em estimativas estatísticas, como o método da Máxima Verossimilhança e a Inferência Bayesiana, que têm a vantagem de considerar na análise as incertezas inerentes à captação dos dados, processos evolutivos complexos que podem ter agido para gerar a diversidade observada, e incertezas e vieses dos próprios estimadores utilizados na análise. Assim, os métodos estatísticos de reconstrução filogenética resultam em análises com estimativas de erro de teste de hipóteses (no caso da Máxima verossimilhança) ou a probabilidade de o estimador estar correto (no caso da inferência bayesiana). Embora tanto os os métodos de distância como os de busca de árvores tenham vantagens e desvantagens, nenhuma destas técnicas é reconhecida como globalmente eficaz para qualquer conjunto de dados. Deste modo a escolha de um método deve levar em conta criteriosamente o tamanho e características particulares do conjunto de dados a ser analisado.

1.5 Coalescência

Os modelos de coalescência estimam o tempo de divergência entre uma amostra de alelos de um determinado gene, e o ancestral comum mais recente (*MRCA = Most recent common ancestor*) a todos os membros da amostra.

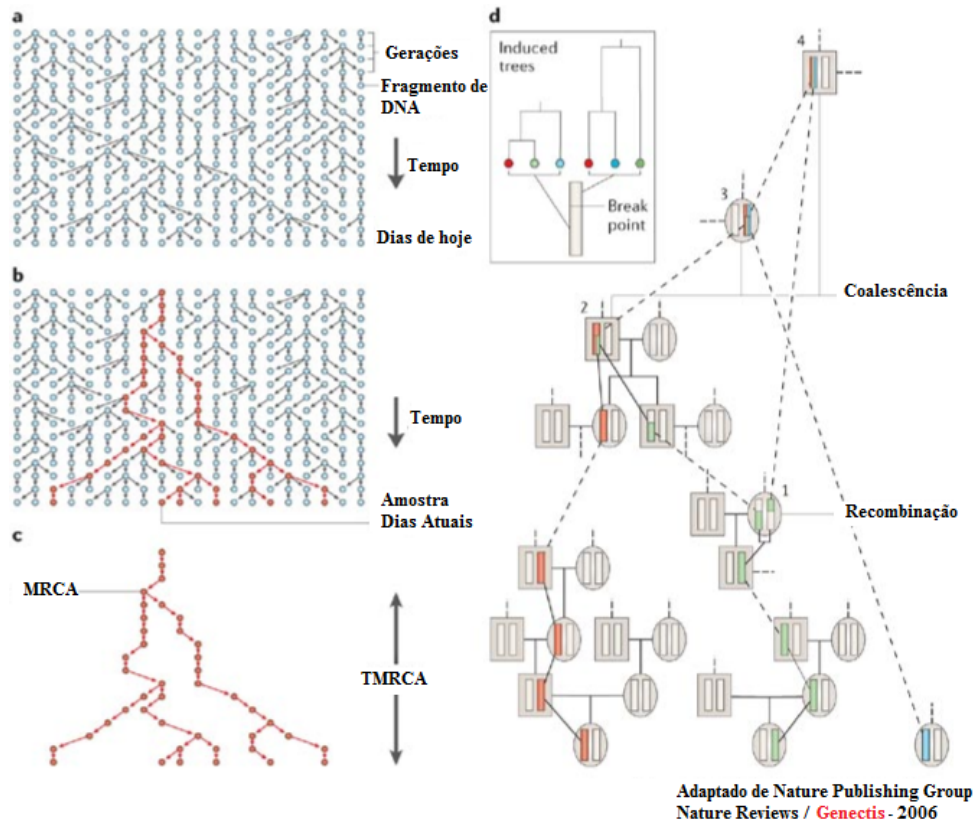


Figura 1.6: Diagrama do modelo de coalescência. Fonte: Marjoram e Tavaré, (2006)

Sob condições simplificadoras, como população constante, e ausência de mutação, os modelos são extremamente simples e a quantidade de gerações necessária para uma amostra obtida no tempo presente **coalescerem** pode ser obtido explicitamente [36] (figura 1.6 [88]).

Trabalhos da última década que expandiram esses conceitos para hipóteses mais realistas e aplicáveis a dados de sequências em filogenias em geral [89, 90] deram um novo interesse ao estudo desses modelos, que motivaram aplicações nas áreas de filo-geografia [91–93] e demografias de parasitos [94, 95].

1.6 Filodinâmica

Com o advento da melhora tecnológica dos sequenciadores na última década, e a consequente disponibilização de enormes quantidades de dados de organismos de várias espécies, é natural que várias áreas aplicadas se beneficiem dessa abundância [96]. Modelos epidemiológicos e estudos quantitativos e qualitativos passaram a incorporar informações evolutivas para a melhor compreensão dos mecanismos parasitários que geram as doenças [97], no estudo da virulência [98] e como esse conhecimento pode ser incorporado em campanhas de redução do número de casos e controle epidemiológico [99, 100], bem como na busca de novos alvos de fármacos mais específicos para diminuir o impacto das doenças [101].

Notavelmente as informações sobre evolução de vírus são mais abundantes, pois devido às altas taxas de mutação envolvidas, os processos evolutivos podem ser observados em um curto período de tempo (ao contrário das escalas de tempo da ordem de grandeza de milhões de anos, normalmente associadas a processos evolutivos de eucariotos).

Em 2004, foi proposta uma nova maneira de aliar os campos da epidemiologia e da dinâmica evolutiva dos parasitos [102]. Essa proposta, que recebeu o nome de **Filodinâmica** tem como um dos principais objetivos a obtenção de informação epidemiológica a partir de dados e metodologias de filogenética molecular. Observou-se que diferentes processos epidemiológicos imprimem padrões topológicos distintos nas filogenias reconstruídas a partir dos parasitos (exemplificados na figura 1.7), fato que foi observado e confirmado em diversos estudos desde então. Esses padrões de ramificação representam as pressões seletivas a que os patógenos estão sujeitos em sua expansão na população hospedeira, e os mecanismos possivelmente usados para evadí-las. A figura 1.8 resume os principais cenários em que diferentes hipóteses sobre a pressão causada pelo sistema imune do hospedeiro determinam estruturas observáveis na filogenia de patógenos cujos processos epidemiológicos são conhecidos. Por exemplo, uma pressão seletiva constante gerada pelo sistema imune, que é continuamente evadida por um patógeno com grande capacidade mutacional como o vírus da gripe (*Influenza*), tem uma filogenia caracterizada por muitas ramificações, com ramos curtos. Por outro lado, um patógeno que está sujeito a uma fraca pressão seletiva como o vírus da Hepatite C (HCV) possui ramos mais longos, o que indica um crescimento da população patogênica, sem a necessidade de uma grande diversificação intra-hospedeiro. Árvores filogenéticas têm sido reconstruídas usando-se

sequências de patógenos e os padrões de ramificação nas árvores são caracterizados a partir de padrões correspondentes de infecção, transmissão, virulência e imunização dos respectivos tipos de parasitos [103–108].

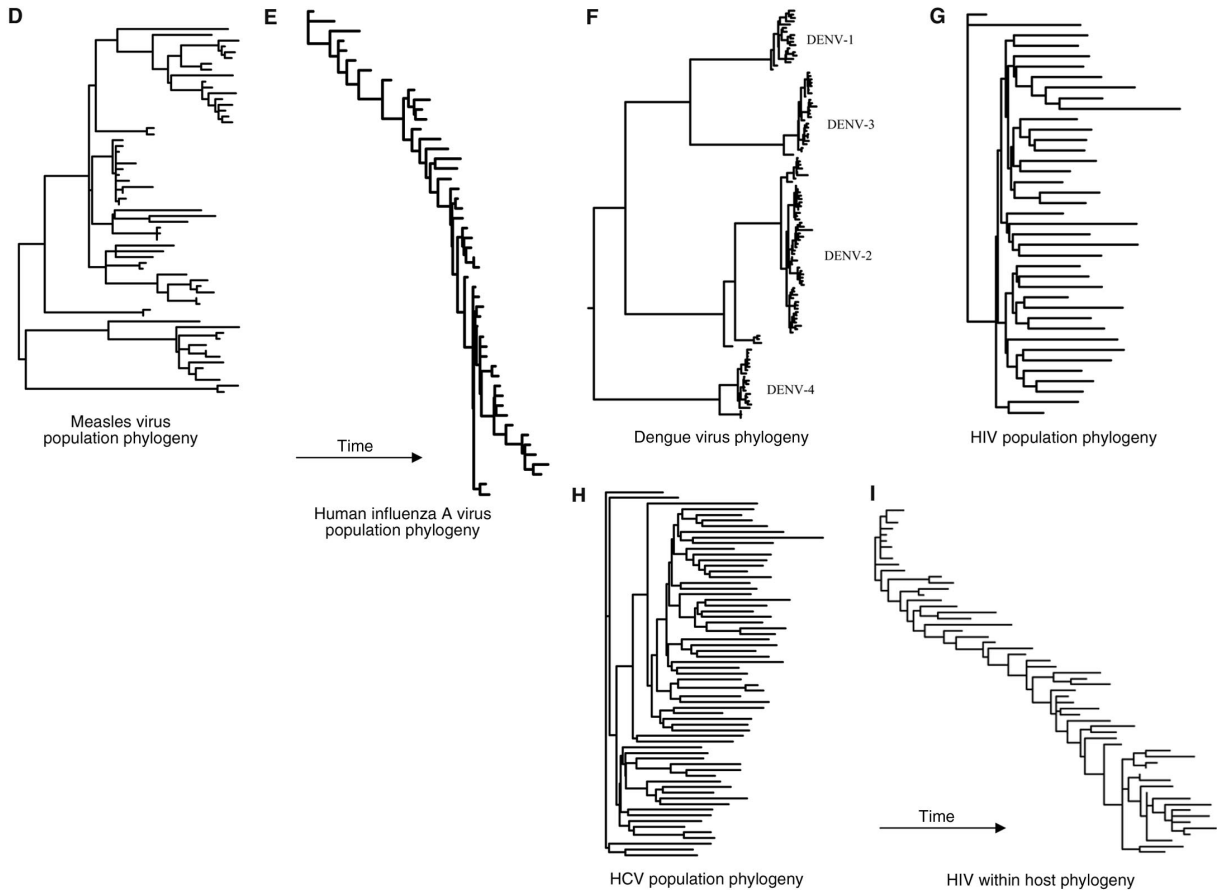


Figura 1.7: Filogenias de vários patógenos. Fonte: Grenfell, *et al.*, (2004)

Revisões recentes dos avanços dessa área foram elaborados [105, 106], observando os sucessos e desafios da área para os próximos anos. Recentemente foi organizada uma conferência para discutir os últimos avanços, e determinar quais os principais desafios observados pelos pesquisadores [109].

A filodinâmica compreende um conjunto de métodos tipicamente utilizados em análises filogenéticas temporais em dados de patógenos. A análise das pressões evolutivas a que os patógenos estão expostos indica as respostas impostas pelo sistema imune do hospedeiro.

Dentre as principais técnicas que caracterizam a filodinâmica, temos:

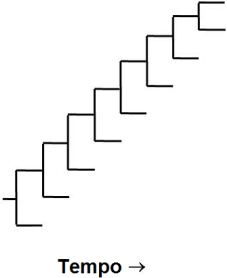
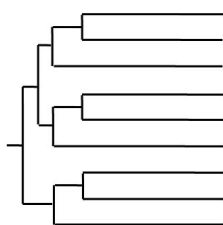
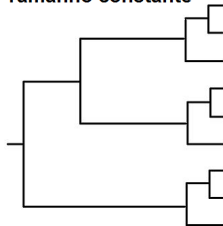
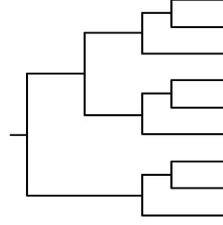
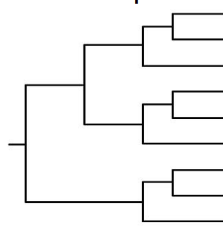
	Continual Immune Selection	Weak or Absent Immune Selection	
		Tree shape controlled by non-selective Processo de dinâmica de população	
Formas de Filogenia idealizadas		Dinâmica de tamanho da população	Dinâmica espacial
		Crescimento exponencial  Tamanho constante 	Estrutura espacial forte  Estrutura espacial fraca 
Exemplos	Human influenza A virus intra-host HIV	inter-host HIV inter-host HCV	Measles, rabies inter-host HIV
Três Inferências	Detection of antigenic escape mutations	Estimativas das taxas de crescimento populacional	Estimativas das taxas de migração da população

Figura 1.8: Filogenias esperadas para vários cenários epidemiológicos. Fonte: Grenfell, *et al.*, (2004)

- Simulações para tempo passado (*backward in time*) e para tempo futuro (*forward in time*) são feitas para se estudar a dinâmica do sistema epidemiológico;
- Análises de coalescência de sequências moleculares para inferir as demografias dos parasitos e com isso estabelecer aproximações para parâmetros de interesse epidemiológico.

Recentemente foi proposto que as técnicas empregadas para o estudo da filodinâmica de vírus e retrovírus sejam adaptados para dados de sequências de TEs [110]. A proposta sugere que, dado que existem semelhanças marcantes tanto na estrutura de alguns vírus e retrovírus e respectivamente transposons e retro-transposons, o sucesso no emprego dos modelos baseados em coalescência para inferência de parâmetros de modelos epidemiológicos de doenças infecciosas indica que seria possível estimar as quantidades e dinâmica infestatória desses elementos genômicos. Essas descobertas possibilitariam uma

nova fonte de informação para detecção e avaliação de possíveis candidatos a *gene drive systems*.

1.7 Simulações

Assim como ocorre há algum tempo na ecologia [111, 112], e em outras áreas [113, 114] a Biologia Computacional Molecular tem observado um crescente número de publicações que descrevem simuladores que se destinam a avaliar questões específicas, particularmente na área de Genética de Populações [115].

Nosso interesse no estudo da evolução e dinâmica de TEs nos motivou a desenvolver uma ferramenta própria, capaz de investigar questões que estão fora do alcance de outros simuladores previamente descritos. Foi elaborado um manuscrito que descreve o simulador aqui referido seguindo os moldes e tradições da área, e o incluímos como anexo [116].

1.7.1 Simulador

A abordagem do simulador proposto (denominado **TRepid**) é válida para modelos paramétricos e determinísticos tanto contínuos como discretos de demografia, dinâmica de transposição e evolução molecular. Além disso, pode ser também generalizada na utilização de modelos estocásticos, conforme indicado no manuscrito em anexo.

Apesar de o simulador manipular modelos determinísticos, os parâmetros demográficos a ser inferidos precisam incorporar a incerteza gerada pela autonomia dos indivíduos (modelo baseado em indivíduos), recombinação dos gametas, formação de casais aleatórios (*random mating*, pan-mixia), e efeitos de deriva genética (*genetic drift*) em populações finitas, estas descritas como modelos de *Wright-Fisher* na literatura de Genética de Populações [36].

Capítulo 2

Justificativa

“My mind,” he said, “rebels at stagnation. Give me problems, give me work, give me the most abstruse cryptogram or the most intricate analysis, and I am in my own proper atmosphere. I can dispense then with artificial stimulants. But I abhor the dull routine of existence. I crave for mental exaltation. That is why I have chosen my own particular profession, or rather created it, for I am the only one in the world”

Sherlock Holmes in “The Sign of the Four” (1890)

As diversas áreas da biologia têm recebido enorme contribuição e influência de diversos ramos da Matemática, pura ou aplicada, e isso tem moldado o processo de formulação de hipóteses para incluir modelos matemáticos, estatísticos e computacionais ao longo das últimas décadas [117]. Grandes quantidades de dados genéticos e moleculares têm sido continuamente disponibilizados após a evolução tecnológica dos sequenciadores, cada vez mais automáticos e eficientes.

Modelos baseados em indivíduos (IBMs) têm tido grande influência nesse processo de metamorfose, com aplicações em diversos problemas de ecologia [118] e epidemiologia [114].

A filodinâmica é uma área relativamente nova (o termo foi proposto em 2004 [102]), que propõe utilizar-se das técnicas rotineiramente utilizadas em duas áreas: evolução e epidemiologia.

O objetivo na área de pesquisa em filodinâmica é unir os dois extremos do espectro, isto é obter métodos que considerem tanto modelos mecanicistas de tempo positivo (a subárea de “-dinâmica”) como os já convencionais modelos de coalescência de tempo negativo (a subárea de “filo-”), em uma abordagem integrada.

Estudos de **dinâmica** de populações são praxe tanto na epidemiologia como na ecologia de vetores. Incontáveis modelos, tanto usando teoria clássica ecológica, como mais recentemente nas duas últimas décadas os modelos baseados em indivíduo (IBMs) [111], foram criados para resolver diversos problemas específicos, disponibilizando uma miscelânea de ferramentas.

Modelos que fazem inferência para tempo negativo, baseados no modelo de **coalescência** têm sido usados com sucesso para estimar modelos demográficos e tempo de divergência em sequências de organismos superiores e vírus [89]. Por não existir uma validação teórico-empírica que justifique sua aplicação em TEs, precisamos obtê-la através de simulações detalhadas e do emprego de análises estatísticas [110].

Esse projeto tem características de ambas as áreas. Embora não ofereça uma abordagem totalmente integrativa, como seria o padrão de ouro desejado pelos pesquisadores da área, acreditamos que a contribuição teórica proposta aqui é um pequeno passo na direção desse objetivo.

Mostraremos simulações refletindo **dinâmica** de populações e evolução molecular, e posteriormente as sequências amostradas serão analisadas com as técnicas convencionais da área para a reconstrução demográfica das sequências (**coalescência**).

A comparação desses dois tipos intrinsecamente diferentes de resultados é apenas um pequeno incremento no sentido de tentar unificar as abordagens. Nossa maior contribuição é a de expandir o contexto de discurso típico da filodinâmica para o universo dos TEs. Para uma tentativa mais abrangente no sentido de integrar os dois extremos do espectro, trabalhos futuros deverão focar mais na parte matemática dos modelos.

Capítulo 3

Objetivos

“Before turning to those moral and mental aspects of the matter which present the greatest difficulties, let the inquirer begin by mastering more elementary problems.”

Sherlock Holmes in “A Study in Scarlet” (1887)

3.1 Objetivos gerais

Estudar o comportamento dinâmico de elementos transponíveis de mosquitos em diversos cenários populacionais, e a influência que a demografia de seus hospedeiros tem em suas distribuições quantitativa e qualitativa, considerando o interesse nesses elementos como possível mecanismo de *drive* genético para a transformação de mosquitos causadores de doenças infecciosas.

3.2 Objetivos específicos

- Simular cenários de expansão de elementos transponíveis (TEs) e observar os padrões filogenéticos decorrentes desses processos;
- Simular cenários de invasão de TEs em populações de mosquitos, e inferir parâmetros dos modelos ecológicos e de transposição que viabilizam invasões;
- Simular sequências resultantes de vários cenários de invasão de TEs para análise filogenética;
- Estimar o componente demográfico dos TEs a partir das sequências de um indivíduo amostrado;
- Observar a influência da dinâmica populacional do hospedeiro na demografia dos TEs;

Capítulo 4

Metodologia

“It is a capital mistake to theorize before one has data. Insensibly one begins to twist facts to suit theories, instead of theories to suit facts.”

Sherlock Holmes in “A Scandal in Bohemia” (1892)

4.1 O simulador

Como a maior parte dos modelos matemáticos de interesse aplicado são não-lineares e acoplados, uma resolução analítica completa está fora de questão. Além disso, embora os resultados qualitativos observáveis para esses modelos sejam importantes e valiosos, não auxiliam a formulação de perguntas quantitativas necessárias para a tomada de decisões no campo de saúde pública. Embora seja possível formular modelos simples e tratáveis para determinados problemas da Biologia, é frequente a formulação de problemas que não podem ser resolvidos com os métodos clássicos de resoluções de equações diferenciais ou a diferenças finitas (equações de recorrência). Esse fenômeno acontece nos modelos que descrevem as interações em diversos níveis de organização biológica relevantes ao estudo da dinâmica de invasão e fixação de TEs em uma nova espécie.

Na falta de soluções analíticas gerais ou específicas, é prevalente nas pesquisas afins o uso de simulações para testar e validar os modelos matemáticos empiricamente. Como não havia um modelo computacional capaz de simular todos os níveis de organização biológica de interesse desse projeto, nem um que fosse próximo o suficiente para ser modificado, um simulador *ad hoc* foi desenvolvido [116] para esse projeto. Esse sistema (chamado TRepid) tem características em comum com diversos outros programas disponíveis da área de Genética de Populações mas, conforme mencionado, nenhum outro incorpora todos os três componentes de modelagem destacados nesse projeto.

O simulador é do tipo *forward-time* e *individual-based* (**IBM = individual-based model**, também referido por *Agent-based model*), isto é, simula gerações sucessivas de uma população para tempo positivo (*forward in time*) e representa cada indivíduo da população de maneira independente (*individual-based*). Cada um destes indivíduos ou agentes tem portanto algum grau de autonomia, e no caso do nosso modelo, cada indivíduo tem um genoma com dois cromossomos, gênero definido e durante sua vida adulta é capaz de contribuir para o *pool* genético da próxima geração (espécie sexuada e diploide)¹.

As sequências de TEs são também representadas como agentes independentes na simulação e podem evoluir independentemente umas das outras, gerar novas cópias ou ser removidas. As famílias de TEs representadas pelos elementos ancestrais e derivados podem se fixar na população ou ser extintas.

¹O conteúdo dos dois cromossomos pode opcionalmente ser fundido em um único cromossomo, a fim de simular populações sexuadas haploides

A presente seção descreve integralmente os modelos implementados no simulador. Os modelos e parâmetros utilizados especificamente para as simulações feitas nesta tese serão descritos na seção 4.2.

4.1.1 A população de vetores

As populações representadas no simulador têm diversas estruturas características de IBMs, que descrevemos a seguir.

***Fitness* dos indivíduos**

Em diversos programas disponíveis para simulação *forward in time* de Genética de Populações o conceito de *fitness* é crucial quando se tenta entender o papel da pressão seletiva (positiva ou negativa) de determinado marcador genético. Existem no entanto duas interpretações possíveis para esse conceito, no tocante o uso em populações simuladas. Ambas são equivalentes ao se estudar populações biológicas reais.

Nesses simuladores tipicamente o *fitness* (representado quase ubiquamente pela letra **W**) é um número que afeta o quanto o marcador genético afeta o desempenho reprodutivo do indivíduo. Esse conceito determina a chance de sucesso de um indivíduo em contribuir para o *pool* genético da próxima geração.

Uma possível implementação considera o sucesso do indivíduo em conseguir realizar a cópula, i.e., quanto maior o fitness, maior a probabilidade de se gerar descendentes. Alguns simuladores que usam essa interpretação incluem `simuPOP` [119], `SFS_code` [120], e `GenomePOP` [121]. Le Rouzic *et al* ([122]) também usam essa definição em seu trabalho de simulação de TEs, embora o ambiente de simulação desenvolvido não tenha sido publicado a parte.

Outra representação possível consiste em tomar o *fitness* diretamente como uma modificação na reprodutividade ou fecundidade, isto é, o número total de descendentes que o indivíduo pode gerar, ou gerou. O programa FPG [123] é um exemplo de aplicação dessa definição. Consideramos que essa maneira foi mais conveniente.

Hipótese de modelagem: O *fitness* de um indivíduo é definido como a quantidade de prole (*offspring*) que este gerou em sua vida adulta.

Estrutura de gênero

A população tem estrutura de gênero, com cada indivíduo tendo sexo definido entre macho e fêmea. Dessa forma, são formados casais aleatórios na população (*random mating*).

Hipóteses de modelagem: A população representada se reproduz por formação de casais aleatórios, sem viés espacial ou preferência fenotípica. A população não é monogâmica. Cada fêmea pode se reproduzir uma única vez por geração. Cada macho pode se reproduzir com várias fêmeas por geração.

No início de cada geração, são formados casais para criar a próxima geração da população. As fêmeas são sorteadas aleatoriamente, sem reposição, em quantidade suficiente para gerar o número de indivíduos especificado pelo **modelo ecológico**, levando em conta a quantidade esperada de filhotes escolhida para a simulação. Os machos são sorteados aleatoriamente com reposição, portanto um mesmo macho pode copular com várias fêmeas, ao acaso.

Estrutura etária

A população tem estrutura etária discreta, sendo a unidade de tempo (também discreta) medida em passos de simulação. A idade em que cada indivíduo se torna apto a se reproduzir (adulto) e a idade máxima (longevidade) são ambas definidas como parâmetros numéricos pelo usuário previamente à simulação.

Como nosso interesse imediato é simular populações de mosquitos, os parâmetros foram escolhidos de modo a eliminar a sobreposição de gerações, embora o sistema possa representar populações mais gerais. Esse comportamento pode ser obtido definindo-se a idade máxima igual a uma unidade maior que a idade reprodutiva. Se além disso, a idade reprodutiva for a primeira unidade disponível para os indivíduos nascidos, a estrutura de idade é efetivamente desabilitada. Isso representa uma hipótese comum em vários simuladores de Genética de Populações, em que a cada geração a população inteira da próxima geração é criada, e toda a população existente é descartada [115, 121, 123].

Estrutura genética (perfil cromossomal)

Hipótese de modelagem: Todo TE presente no genoma, ativo ou inativo para transposições, é expresso e tem o potencial de reduzir a fecundidade de seu hospedeiro.

Esse impacto depende apenas de sua localização no cromossomo.

Subpopulação silvestre: Indivíduos virgens são aqueles indivíduos que não possuem nenhum TE, e portanto nenhum custo de fitness associado na análise. Estes desfrutam portanto do *fitness* máximo possível para a espécie.

Subpopulação GM: Indivíduos **GM** são aqueles que possuem pelo menos um TE em seu genoma. Essa é a subpopulação de interesse das análises, e que está sujeita à diminuição condicional de seu *fitness*, no caso em que um ou mais TEs ocupem determinadas posições no cromossomo (cf. seção 4.1.3).

Ploidia: Populações sexuais são geralmente diploides, e esse é o comportamento padrão do simulador. No entanto, para examinar a alternativa de populações haploides foi criada uma alteração na formação de novos indivíduos, semelhante ao que acontece na formação de gametas. Esse mecanismo dá origem a genomas haploides, após a reprodução sexuada.

4.1.2 Modelos ecológicos

Hipóteses de modelagem: A população de hospedeiros é finita e isolada de outras espécies ou populações, se distribui homogeneamente em um ambiente espacialmente não-limitado, e se reproduz sem qualquer viés espacial ou sexual (*random mating*).

A cada geração uma porção da população presente em idade reprodutiva é amostrada (uniformemente) para formar os casais necessários para gerar a próxima geração (*random mating*). A quantidade de novos indivíduos a ser criada nessa etapa depende de vários fatores, mas principalmente de um modelo ecológico predeterminado que é consultado para determinar uma dinâmica paramétrica à população de hospedeiros.

A literatura de ecologia matemática de populações abunda em exemplos de modelos que descrevem dinâmicas arbitrárias ou de acordo com fenômenos ou ambientes específicos para diversas espécies.

Tipicamente é desejável que algumas características sejam observadas num modelo ecológico implementado no sistema TRepid:

- o tamanho da população deve ser limitado superiormente, de modo que a população não cresça indefinidamente, causando um custo computacional em termos de processamento e memória excessivo.

- o modelo deve ser representado matematicamente por uma fórmula explícita, ou uma equação de variação, como é típico na literatura de modelagem.

A primeira dessas premissas é declarada apenas para a conveniência de simulações longas (grande número de gerações), não sendo obrigatória na modelagem computacional, caso haja recursos de CPU e memória suficientes. Diversos modelos satisfazem essas simples premissas, incluindo em nossa implementação os modelos com saturação populacional e o modelo de população constante, para o primeiro caso. Todos os modelos implementados, descritos a seguir, satisfazem o segundo requisito.

Os modelos populacionais do hospedeiro implementados incluem:

1. População constante
2. População linear (crescimento e decrescimento)
3. População exponencial (crescimento e decrescimento)
4. Crescimento com saturação (denso-dependência - equação logística)
5. Equação de Hassel (competição intraespecífica e denso-dependência).

Os modelos implementados na versão atual do simulador podem ser genericamente representados por uma equação dinâmica com tempo discreto, onde p_n representa a população na geração atual (tempo n), de modo que a quantidade de indivíduos na próxima geração é p_{n+1} , no caso particular em que seja desconsiderada a estrutura de idade dos indivíduos².

Nas seções a seguir, p_n representa o tamanho da população p na geração n , e p_{n+1} o tamanho da população na geração $n + 1$.

População constante

Uma **população constante** no simulador TRapid consiste em uma população que a cada geração cria sempre a mesma quantidade de novos indivíduos.

²Caso contrário, a formulação explícita deverá levar em conta as quantidades esperadas de cada classe etária que sobreviverão até a próxima geração. Essa formulação não está presente no texto pois está fora do escopo do projeto que considera apenas populações de insetos com curta longevidade e sem sobreposição de gerações.

Dessa forma, o parâmetro r que determina a taxa de reprodutividade é interpretado como (ou truncado para) um número inteiro, que corresponde a quantidade exata de novos indivíduos da próxima geração, na ausência de fatores que influenciem o *fitness* dos indivíduos reprodutores.

Dos modelos ecológicos implementados, mencionados nesse trabalho, é o único que não usa como parâmetro de entrada o tamanho atual da população.

$$p_n = r \quad (4.1)$$

População linear

Uma **população linear** é uma população que cresce ou decresce a cada geração em incrementos proporcionais.

Assim como na **população constante**, o parâmetro r que determina a taxa de natalidade é interpretado como um número inteiro, mas no modelo linear este é acrescentado ao (ou subtraído do) tamanho atual da população para determinar o tamanho da próxima geração a ser criada.

$$p_{n+1} = p_n + r \quad (4.2)$$

População exponencial

Uma **população exponencial** cresce ou decresce a uma taxa de natalidade constante que é proporcional ao tamanho atual da população.

Nesse modelo, o parâmetro r que determina a taxa de crescimento da população é interpretado como um número real e usado como proporção do tamanho atual da a ser criado.

Essa implementação permite que o tempo não seja interpretado explícita ou implicitamente na formulação, de modo que o crescimento é interpretado como uma expansão exponencial **instantânea** da população a cada geração. Isto é, na ausência de fatores externos (por exemplo, no caso de uma população virgem) isso corresponde a um crescimento exponencial típico de uma equação diferencial. Na presença de fatores que impactam na fertilidade da população, como por exemplo o acúmulo de TEs que diminuem o *fitness* de seus hospedeiros, podem ocorrer oscilações no crescimento.

$$p_{n+1} = p_n + rp_n \quad (4.3)$$

População logística

Uma **população logística** é uma população que tem um crescimento sujeito a um ambiente com recursos finitos ao qual está sujeita. Esse tipo de crescimento é chamado **denso-dependente**.

Se a população é muito pequena em relação à **capacidade suporte** K do ambiente, ela cresce rapidamente, de forma semelhante a uma população exponencial com parâmetro reprodutivo r . Quando o tamanho passa a ser comparável à capacidade suporte, esse crescimento é desacelerado, e apesar da população continuar crescendo, isso ocorre em incrementos cada vez menores. No caso extremo, se a população atingir um patamar quase idêntico à capacidade suporte, sua taxa de crescimento se torna negligível, e na ausência de perturbações externas ela se comporta como uma população constante.

Esse modelo possui portanto dois parâmetros de interesse: a taxa de crescimento r da população, e a capacidade suporte K do ambiente que determina o tamanho da população em equilíbrio.

$$p_{n+1} = p_n + rp_n \left(1 - \frac{p_n}{K}\right) \quad (4.4)$$

População de Hassel

O modelo de **população de Hassel** [124] é muito semelhante ao modelo de população logística descrito acima. Este também considera um crescimento denso-dependente, representado pelos mesmos parâmetros r de crescimento e K de capacidade suporte. Além desses, é também considerada uma competição intraespecífica por recursos, representada pelo parâmetro β .

$$p_{n+1} = p_n + rp_n \left(1 - \frac{p_n}{K}\right)^\beta \quad (4.5)$$

4.1.3 Estrutura do genoma

Hipóteses de modelagem: O conteúdo específico do genoma do hospedeiro não é relevante para a dinâmica dos TEs invasores. Os fatores determinantes a ser considerados

são a quantidade e localização dos TEs, enquanto os fatores determinantes para a evolução dos mesmos são suas sequências. As sequências dos TEs assumem o modelo de sítios finitos [125], possibilitando que mutações sucessivas se acumulem em um mesmo sítio. Essa possibilidade aumenta o realismo da simulação, e riqueza do processo, em comparação com o modelo de infinitos sítios, no qual cada mutação ocorre sempre em um novo sítio, criando efetivamente um novo alelo [126].

O genoma dos indivíduos é representado por uma lista discreta de **loci** estruturados que podem ou não conter TEs. Esses *loci* estão distribuídos ao longo dos cromossomos e são classificados em três categorias, de acordo com o impacto potencial que causam no hospedeiro:

1. *Loci* neutros;
2. *Loci* severos;
3. *Loci* fatais.

Loci neutros são aqueles *loci* que não causam alteração fenotípica perceptível, ou não causam mutação que influencie de forma notável a capacidade reprodutiva do hospedeiro. Esses *loci* podem representar regiões não codificantes dos cromossomos, regiões intergênicas, *introns*, ou mesmo novos blocos que passaram a existir devido a presença de TEs novos.

Loci severos correspondem a *loci* que causam mutações perceptíveis que diminuem a capacidade reprodutiva do hospedeiro, seja interferindo no processo de corte, cópula, ou fecundidade. O impacto de fato que cada TE inflige no hospedeiro ao ocupar uma dessas posições corresponde a um valor numérico cumulativo. O valor individual de impacto para cada é determinado por um parâmetro numérico predeterminado, escolhido na configuração da simulação pelo parâmetro `severe_impact` e denotado neste manuscrito por s . O impacto total acumulado por todos os TEs em *loci* severos é arredondado e esse inteiro é subtraído da quantidade esperada de filhotes que o indivíduo gerará, caso seja selecionado para procriação (cf. seção 4.1.4).

Loci fatais representam posições intragênicas essenciais, e qualquer alteração nessas regiões implica em mutação deletéria, que é negativamente selecionada. O indivíduo pode ter uma má formação embrionária, ou logo após o nascimento, e qualquer inserção por TE em um *locus* fatal é suficiente para inviabilizar o hospedeiro, tendo-o removido da população no momento de seu nascimento.

4.1.4 Representação de TEs

No modelo computacional **TRepid**, diversas hipóteses simplificadoras são feitas em relação à representação de TEs. Nem todas as características de TEs reais [41] são consideradas, a fim de tornar o sistema computacionalmente tratável.

O simulador foi desenvolvido na linguagem de programação Perl5³, que é uma linguagem orientada a objetos. Cada TE corresponde a um objeto, representado por uma lista associativa (*associative array* ou *hash*). Nesta lista estão representados diversos atributos que determinam todas as características individuais do TE. Assim sendo, os principais atributos de um TE, correspondem a variáveis definidas em escopo local (*object data* para cada TE), de modo que cada TE de cada indivíduo da simulação é de fato representado de forma independente de todos os outros.

A seguir, segue a descrição superficial da estrutura de dados desses objetos. Mais detalhes sobre essa e outras estruturas de dados serão fornecidos no apêndice.

Atributos de TEs

Posição

Uma das informações necessárias que definem o TE é sua localização no genoma. Para tanto, usamos dois números para identificar sua posição no cromossomo (*locus*) e em qual cromossomo ele está.

Status de atividade

Hipótese de modelagem: Os TEs têm toda a estrutura necessária para sua própria transposição em estado completamente funcional, ou não tem as estruturas necessárias para ser transpostos. Isto é, sendo ativos ou inativos, todos são autônomos (cf. seção 1.3.2).

Na versão atual do simulador, a única classificação implícita entre TEs quanto às suas características estruturais é entre **ativos** e **inativos**⁴. Representamos essa ideia no sistema com um número real entre zero e um. Qualquer elemento com status maior que zero é considerado ativo, e o mecanismo de autorregulação da família de TEs é considerado de forma implícita nesse valor de status (cf. seção 4.1.8).

Impacto de *fitness*

³<http://www.perl.org>

⁴não foi implementada uma classificação de TEs entre autônomos e não-autônomos, embora isso esteja previsto para uma versão futura (cf. seção 6.3).

Existem várias possíveis maneiras de se modelar o impacto que cada TE específico inflige em seu hospedeiro (cf. seção 4.1.1). Uma proposta simples é considerar o custo s_i que cada elemento i causa, e contabilizar o custo total dos TEs de maneira aditiva, ou multiplicativa (e.g. [123]).

O sistema **TRepid** usa um custo aditivo e constante (denotado s). Dessa forma, para calcular o impacto total S que os TEs causam ao indivíduo, basta contabilizar todos os elementos que estão em posição de causar algum custo

$$S = \sum_i s_i = \sum_i s = i \times s \quad (4.6)$$

O impacto total S no indivíduo é então arredondado e subtraído do número total de filhotes que ele é capaz de gerar.

4.1.5 TEs de Classe I e II

Fizemos algumas hipóteses simplificadoras ao representar retro-transposons e transposons no simulador.

Hipóteses de modelagem: Não há diferença intrínseca na eficiência dos mecanismos replicativos dos TEs que se replicam via DNA ou RNA. Ambos se replicam de forma quantitativa idêntica e esta quantidade é definida previamente à simulação. A diferença entre esses mecanismos é qualitativa, na forma como as sequências evoluem ao longo do tempo.

Classe I

Hipóteses de modelagem: Um TE classe I é copiado por um mecanismo baseado em retro-transposase, de modo que uma cópia é feita a partir da porção de DNA presente no cromossomo, deixando-a intacta. A nova cópia é inserida em um novo *locus* do genoma por um mecanismo imperfeito.

Os TEs de Classe I, ou retro-transposons se replicam usando uma transcriptase reversa como etapa intermediária, que transcreve a informação contida no DNA para RNA, mantendo a cópia do DNA em sua localização original. Por essa razão, esse mecanismo é conhecido como “*copy and paste*”. Como a cópia original é preservada, ela permanece inalterada até o próximo **evento evolutivo** (cf. seção 4.1.9).

Na implementação, o elemento ativo original é sempre preservado em sua sequência intacta no início do **evento de transposição** (cf. seção 4.1.6). No entanto, a nova cópia está sujeita a erros na replicação, portanto ela passa por um evento evolutivo durante o evento de transposição.

Classe II

Hipóteses de modelagem: Um TE de Classe II é excisado pela transposase do cromossomo, e copiado para um novo *locus* do genoma por um mecanismo imperfeito. O *locus* original é então reconstruído por um mecanismo também imperfeito.

Um TE de classe II utiliza uma transposase, enzima que reconhece as regiões terminais do elemento e o excisa inteiramente do cromossomo, transportando-o para outro *locus* de inserção. Por essa razão, esse mecanismo é conhecido na literatura como “*cut and paste*”. A maquinaria de reparo nuclear é então responsável por usar a cópia reverso complementar para reconstruir a informação original, e assim o elemento é replicado de um *locus* para outro. É levado em conta uma chance de erro desse mecanismo de reparo nuclear, e portanto no momento da transposição é contabilizado um **evento evolutivo** para o TE ativo que original a transposição.

Neste modelo computacional o elemento ativo original passa por um **evento evolutivo** (cf. seção 4.1.9), assim como as cópias que ele criou nesse **evento de transposição** (cf. seção 4.1.6).

4.1.6 Eventos de transposição

Chamamos **evento de transposição** o momento em que o número de TEs muda, na gametogênese. A variação do número de TEs no cromossomo (e conseqüentemente o número de elementos total na população naquela geração) ocorre principalmente por dois fatores. O **modelo de transposição** é consultado em duas etapas distintas com as seguintes finalidades:

- Determinar o número de novos TEs;
- Gerar novos TEs em quantidade suficiente para satisfazer o modelo de transposição e incluí-los no gameta;
- Determinar o número de TEs a ser excisados (removidos).

- Remover TEs do gameta em quantidade suficiente para satisfazer o modelo de excisão;

4.1.7 Dinâmica de transposição

Criação de elementos

Os modelos de dinâmica de transposição que determinam a quantidade de TEs do cromossomo a ser criado (gameta) em relação aos cromossomos parentais advém de várias formulações possíveis [122]. Estes são tipicamente representados por equações ou sistemas de equações de variação de quantidade (equações diferenciais ou equações a diferenças finitas) e podem ser classificados em duas categorias: modelos neutros, que não levam em conta pressão seletiva da existência ou abundância, e modelos que consideram a seleção natural no hospedeiro como fator regulador da quantidade de elementos no genoma.

Os modelos atualmente implementados no sistema incluem:

- Crescimento linear
- Crescimento exponencial
- Modelo SKR [127]

Nas seções a seguir, c denota o número de cópias presentes no genoma parental, e dc o número de novas cópias a ser geradas para o gameta.

Crescimento linear

O modelo linear corresponde a um crescimento com uma taxa fixa, esta representada por um parâmetro numérico (inteiro) que determina quantos novos elementos serão criados em cada **evento de transposição**.

$$dc = u \tag{4.7}$$

onde u é a taxa de transposição, representada pelo número (inteiro) de novas cópias a ser criadas para o gameta.

Crescimento exponencial

O modelo exponencial corresponde a um crescimento proporcional a quantidade de elementos existentes no genoma parental. O parâmetro numérico (número real) corresponde a proporção da quantidade de TEs que deve ser criada, em relação à quantidade existente no genoma parental em cada **evento de transposição**.

$$dc = uc \quad (4.8)$$

onde u é a taxa de transposição, representada pela proporção de novas cópias a ser criadas para o gameta referente ao número existente de cópias c .

Modelo SKR

O modelo SKR [127] é um modelo que originalmente acopla tanto equações que regem as quantidades de TEs, como o tamanho da população. Como o modelo foi proposto num *framework* matemático, e o simulador proposto nesse projeto é um IBM, esse modelo foi adaptado para incorporar apenas as equações e fórmulas relacionadas a TEs. A equação populacional do artigo original de 2005 é a equação de Hassel [124] e este é um dos modelos ecológicos implementados no simulador **TRepid**. Além de ser possível usar o modelo de transposição SKR com a equação de Hassel assim reproduzindo a metodologia do artigo original, pode-se usá-lo com qualquer outro modelo ecológico aumentando-se assim a gama de possíveis **cenários de simulação**.

$$dc(t+1) = \text{ceil}(c(t) \times T_0 \times U(c(t))) \quad (4.9)$$

Nesse modelo, $U(c)$ é uma função crescente limitada que define a forma da dinâmica, que pode ser arbitrada satisfazendo certos critérios. No artigo, são propostas três formas para essa função que foram implementadas no simulador segundoss suas formulações originais:

$$U_1(c) = 2\left(\frac{-c}{C_{0.5}}\right) \quad (4.10)$$

$$U_2(c) = \frac{1 - c^5}{(C_{0.5}^5 + c^5)} \quad (4.11)$$

$$U_3(c) = 1 + (c - \frac{.5}{C_{0.5}}) \quad (4.12)$$

onde T_0 e $C_{0.5}$ são constantes passadas como parâmetros pelo usuário.

Excisão de elementos

Mecanismos de perda ou remoção de TEs existem e são responsáveis por controlar a quantidade dos elementos no genoma da espécie hospedeira. Esse tipo de mecanismo já amplamente reportado em estudos anteriores, podem ser atribuídos em alguns casos a uma resistência natural da espécie em acumular DNA (cf. seção 1.3.2), ou como resultado de seleção negativa de indivíduos com *fitness* reduzido pela atividade dos elementos, e ainda em outros casos a mecanismos de autorregulação das próprias famílias de TEs em atividade de expansão [47, 50, 128]. Existem diversas tentativas de incorporá-los direta ou indiretamente em modelos a fim de estudar o fenômeno mais amplamente do que já foi testado *in vivo* [122, 129–131]. Decelière *et al* [132] usam uma formulação comum na literatura, a de uma deleção a taxa constante por elemento por geração, resultando num modelo exponencial.

Hipótese de modelagem: A perda aleatória de TEs é causada principalmente pelo mecanismo de reparo nuclear, e é representada indiretamente no modelo de excisão de TEs. Esta regulação é intra-hospedeiro, mas não depende diretamente das características da família de TEs em questão.

Assim, além de determinar o número de novos elementos conforme descrito na seção anterior, o sistema também precisa determinar o número de cópias existentes que são removidas aleatoriamente em cada **evento de transposição**.

O simulador TRepid incorpora modelos de excisão específicos para cada modelo de transposição, e nestes descritos acima a implementação da excisão é semelhante à de recrutamento de novas cópias.

Para os modelos linear e exponencial são utilizados fórmulas idênticas às de replicação, porém usando um parâmetro separado para a taxa. Para o modelo SKR, é utilizado o modelo exponencial.

4.1.8 Atividade e inativação de TEs

Elementos transponíveis podem ser ativos ou inativos, de acordo com a capacidade de se transpor (TEs autônomos) ou ser transposto caso a enzima necessária seja viabilizada por outro TE (TEs não autônomos). Na natureza isso geralmente se reflete na presença e conservação de estruturas terminais que podem ser **regiões terminais invertidas** (TIR), **regiões terminais diretas** (LTR), e **sequências palindrômicas**.

No sistema TRepid a informação relevante dessas estruturas está incorporada em um único valor numérico, o **status de atividade** que assume valores entre 0 e 1. Qualquer valor não-nulo determina um TE ativo. Dessa forma, não precisamos considerar ou manipular subsequências correspondendo a estruturas conservadas, simplificando os algoritmos e diminuindo o armazenamento de dados em memória e disco, e o número de operações envolvidas.

TEs podem se multiplicar desde que haja pelo menos uma cópia ativa disponível, assumindo que o mecanismo de transposição esteja disponível (transposase ou retrotransposase). Dessa forma, todas as cópias ativas são atualmente consideradas autônomas, também para simplificar os algoritmos⁵.

Na versão atual, são implementados os seguintes mecanismos de inativação:

1. Master gene;
2. Progressivo;
3. Aleatório.

Seguem descrições sucintas sobre suas premissas e implementações. Todos os mecanismos descritos abaixo têm um limiar mínimo de 0,01. Qualquer **status** menor que 0,01 é arredondado (truncado) para zero e portanto considerado inativo.

Modelo Master Gene

O modelo *Master Gene* é favorecido na literatura como o mecanismo de replicação de TEs mais simples de ser modelado e interpretado [78, 133–135], além de explicar a

⁵Uma análise mais abrangente considerando TEs ativos não-autônomos é possível e está prevista para trabalhos futuros (cf. seção 6.3).

aparente pequena quantidade de elementos que seriam observados em um crescimento exponencial, em algumas famílias de TEs em genomas observados [136].

De acordo com a hipótese *master gene*, as cópias criadas são ligeiramente degeneradas ou incompletas, o que as impedem de ser replicadas, mesmo que as enzimas necessárias estejam presentes devido a outros elementos autônomos. Essas novas cópias são consideradas inativas., ou *Dead on Arrival (DOA)*.

Hipótese de modelagem: Todos os elementos criados a partir de um TE ativo são inativos, e não há possibilidade de reativação de elementos previamente inativos. Todas as cópias ativas são autônomas.

Na implementação do modelo *Master Gene* no simulador, eventos de transposição em um dado indivíduo só são possíveis se existir pelo menos um TE ativo (portanto com **status** positivo). Todas as cópias em um cenário *Master Gene* são criadas automaticamente com **status** zero.

Modelo progressivo

Hipótese de modelagem: Todos os TEs criados a partir de um elemento ativo têm o **status** menor que o TE original. Cada novo status é igual à metade do status do elemento original. Todas as cópias ativas são autônomas.

Modelo aleatório

Hipótese de modelagem: Todos os TEs criados a partir de um elemento ativo têm o **status** menor ou igual ao TE original. É escolhido um valor aleatoriamente entre zero e o status original. Todas as cópias ativas são autônomas.

4.1.9 Eventos evolutivos

Chamamos **evento evolutivo** cada evento em que é inserida uma ou mais mutações nas sequências de TEs em questão. No modelo computacional do simulador são consideramos os dois seguintes eventos evolutivos:

1. Uma nova transposição;
2. O final da gametogênese.

Em todos os casos em que um evento evolutivo pode ser desencadeado, a ocorrência de fato de um evento é aleatória, com probabilidade definida por parâmetro pelo usuário.

Nova transposição

A cada nova transposição, isto é, a cada novo TE criado, uma ou mais mutações são inseridas na nova cópia, o que a diferencia do elemento *template* (o elemento que originou a nova cópia). Se a família de TEs em questão for de Classe I (retro-transposase), o elemento *template* é preservado em sua forma original. Se por outro lado a família de TEs é de Classe II (DNA transposase), a cópia original pode receber mutações representando erros no mecanismo de reparo da célula.

Fim da gametogênese

Ao final da gametogênese, quando todas as novas cópias já foram criadas, e os elementos a ser excisados também já foram removidos, todos os TEs presentes no gameta passam por um evento evolutivo, recebendo uma ou mais mutações adicionais. Esse mecanismo foi inserido para garantir que haja um envelhecimento padronizado do cromossomo, caracterizando uma escala de tempo evolutivo para as sequências. Ele também tem como efeito o aumento da velocidade do processo evolutivo, já que é possível arbitrar o número de substituições que ocorrem nesse evento.

Quantidade de substituições por evento

Todas as mutações inseridas em cada **evento evolutivo** são substituições de bases nucleotídicas sorteadas com reposição tanto no alfabeto de nucleotídeos quanto na extensão da sequência alvo. Isto possibilita a ocorrência de substituições silenciosas com probabilidade inversamente proporcional ao tamanho da sequência.

Ocorrência condicional de eventos evolutivos

Todo **evento evolutivo** é condicional, i.e. é acionado ou não de acordo com uma probabilidade preestabelecida, definida no arquivo de configurações da simulação por meio do parâmetro `mutation_prob`. É possível fixar a probabilidade desse evento ocorrer como 1 (respec. prob. = 0), para garantir que ela sempre ocorra (respec. nunca ocorra).

<p>Redução do cabeçalho de TEs simulados</p> <p>Nome geração elemento original</p> <p>Exemplo:</p> <p>15_1_0 15_1_0 15_1_0 10_2_1 10_2_1 10_2_1 \implies 15_1_1 3 10_2_1</p>
--

4.1.11 Algoritmo das gerações

O algoritmo que ocorre a cada geração modela o ciclo de vida básico de uma espécie sexuada, sujeita a eventos de transposição e recombinação durante a gametogênese.

1. Casais de hospedeiros são sorteados aleatoriamente dos hospedeiros sexualmente maduros. Machos e fêmeas são escolhidos com e sem reposição, respectivamente;
2. Cada adulto gera um novo gameta, de acordo com as definições para transposição e recombinação;
3. A transposição ocorre de acordo com a quantidade de novos TEs a ser gerados, e a quantidade de TEs existentes a ser excisados do cromossomo. Isso efetivamente modifica o número de TEs no gameta, em relação à quantidade presente nos cromossomos parentais;
4. Mutações são sorteadas do modelo evolutivo para elementos criados na gametogênese. Se a transposição for *cut and paste*, as cópias originais também recebem mutações; se for *copy and paste*, as cópias originais permanecem inalteradas.
5. A recombinação dos cromossomos proporciona um embaralhamento adicional do conteúdo do gameta;
6. Mutações são sorteadas do modelo evolutivo e substituídas em todas as cópias de TEs do gameta;
7. O custo de *fitness* imposto pelos TEs em futuros hospedeiros é calculado a partir dos gametas e qualquer filhote que exceda um limiar preestabelecido é considerado morto antes de nascer, e removido da população.
8. Cada casal formado gera uma quantidade de filhotes definido pelo usuário como parâmetro, excluindo-se filhotes que excedam o limiar de custo de *fitness*;

9. A idade de todos os indivíduos sobreviventes da população é incrementada, e indivíduos que excedam a idade máxima são removidos;

4.2 Desenho experimental

As simulações desenhadas para este projeto foram divididas em três experimentos com objetivos distintos.

- O **experimento MG** teve a finalidade de observar a topologia básica de uma família de TEs que se expande segundo um modelo *Master Gene*, e observar a relação entre a estrutura temporal da topologia com as gerações da população simuladas;
- O **experimento A** proporcionou uma análise detalhada das condições necessárias para uma invasão bem sucedida por uma família de TEs em uma população hospedeira, após a liberação de uma pequena quantidade de indivíduos modificados geneticamente;
- O **experimento B** proporciona a visualização de um panorama mais realista de cenários de invasão, seguindo as condições exibidas nos dois experimentos anteriores, e a análise demográfica das sequências para obter tanto estimativas do tempo de invasão, como o tamanho efetivo da família de TEs.

Embora tenham objetivos bastante diversos, algumas características comuns podem ser observadas nos conjuntos de parâmetros dos três experimentos. Descreveremos as metodologias de cada um em seção separada, mas para simplificar a leitura, isolamos as escolhas de parâmetros comuns aos três.

Antes de detalhar os parâmetros utilizados nas simulações e análises, alguns termos serão definidos nas próximas seções, para simplificar a linguagem ao fazer menção a grandes quantidades de informações ou parâmetros que são costumeiramente interpretados sempre em conjunto e no mesmo contexto.

Esses termos incluem:

1. Cenários de simulação
2. Eventos de transposição

3. Eventos evolutivos

Essas expressões se referem a detalhes específicos da forma como os dados foram simulados, e não tem necessariamente relação com conceitos biológicos amplos, ou descritos na literatura.

4.2.1 Cenários de simulação

O *core* das simulações segue um protocolo comum. Os resultados estão sendo compilados em artigos ainda em fase de escrita, cujos rascunhos estão disponíveis em anexo com essa tese. [116, 137]. Uma breve apresentação das semelhanças e diferenças entre esta tese e os manuscritos em preparação se encontra no apêndice.

A cada **cenário de simulação** simulado corresponde um conjunto de parâmetros passados ao simulador **TRepid**, armazenado em um arquivo de configurações (que usa o nome padrão `trepid.conf`). Cada cenário é armazenado em uma estrutura de diretórios independente, o que facilita a organização e manipulação dos diversos arquivos criados⁶.

Os parâmetros comuns a todos os cenários incluem: modelo de inativação *Master Gene*, modelo de transposição constante. Os TEs simulados são da classe I, portanto se multiplicam usando o método *copy and paste*, que preserva a cópia original (cópia *template*) durante o evento de transposição.

O número de mutações em cada evento evolutivo é 2 substituições, sorteadas com reposição, isto é, substituições silenciosas são permitidas em todos os cenários.

Ao final de cada gametogênese, todos os TEs presentes - novos ou preexistentes - passam por um novo evento mutacional, efetivamente marcando um processo de “envelhecimento” dos mesmos.

A estrutura dos cromossomos foi parcialmente desativada, e a maior parte dos *loci* são considerados **neutros**. A quantidade de *loci* **severos** é especificada para cada um dos experimentos, e o coeficiente de seleção s (custo de *fitness*) utilizado é especificado para cada simulação. Nenhum dos experimentos presentes nesta tese consideram *loci* **fatais** ou essenciais, que causariam inviabilidade do hospedeiro (vide seção 4.1.3 para mais informações sobre a estrutura dos cromossomos nestas classes de *loci*).

⁶são criados 6 arquivos de saída para cada réplica de cada cenário

Réplicas

A cada cenário descrito correspondem 10 simulações com exatamente os mesmos parâmetros, e ao final de cada simulação individual é amostrado um indivíduo GM da última geração da população e seu genoma é armazenado em disco nos formatos FASTA e NEXUS para análise posterior.

cenário \iff conjunto de parâmetros \iff 10 réplicas (simulações)

Amostragem

Ao fim de cada simulação, um indivíduo é amostrado aleatoriamente da subpopulação GM, sem nenhuma preferência ou viés introduzido artificialmente. Isso pode ter o inconveniente de que o indivíduo amostrado tenha um perfil cromossômico com muito poucas sequências, devido a grande variabilidade intrínseca do processo. Nesse processo de amostragem, no entanto, são eliminados do universo amostral os indivíduos virgens, isto é, indivíduos que não contém nenhum TE.

Para análise de cada cenário, é escolhido dentre as amostras o indivíduo amostrado com maior número de sequências, desde que seja observado um número mínimo de sequências igual a 50.

O modelo **coalescente** prevê que uma amostra aleatória de tamanho n tem uma probabilidade de $\frac{n-1}{n+1}$ de conter o ancestral comum mais recente (MRCA) [110]. Uma amostra de tamanho 39 teria portanto uma chance de 95% de que exista um ancestral comum mais recente.

Nos experimentos pilotos realizados para afinar o desenho experimental, observamos que o número médio de TEs por indivíduo era muito maior que 40 após algumas gerações, chegando em alguns casos a mais de 100. Uma amostra de tamanho 50 é portanto uma realização aceitável no contexto de nossas simulações.

Se em algum cenário não for obtido nenhuma amostra com pelo menos 50 sequências, mas o número médio de sequências por indivíduo for superior a 50, é feita uma nova simulação, mas ao final desta réplica, a amostragem é feita manualmente para selecionar um indivíduo com pelo menos 50 sequências.

Se o número médio de TEs por indivíduo for menor que 50 (na média) para todas as réplicas, considerar-se-á que esse cenário está gerando inerentemente poucas sequências, e portanto sem condições de representar invasões.

Critérios de parada para as simulações

Existem três critérios de parada para cada simulação. A simulação pode parar (i) caso haja uma invasão bem sucedida, definida como uma proporção de indivíduos contendo pelo menos um TE ser igual ou maior a 90% da população total; (ii) caso o TE seja perdido, definido como a proporção da população que contem pelo menos um TE cair abaixo de 5%; (iii) um número máximo pré-determinado de gerações sejam rodadas, caso não haja convergência da invasão, ou a convergência for muito lenta.

Modelos ecológicos

Uma população de referência, com dinâmica ecológica constante foi usada nos cenários dos três experimentos MG, A e B, para simplificar a comparação dos resultados. Além deste, no experimento B foram observados outros três modelos ecológicos. Os tamanhos das populações variam para cada experimento e todos os parâmetros serão detalhados em cada respectiva seção.

Todas as simulações iniciam com uma proporção de 10% da população contendo TEs (subpopulação GM) e os outros 90% compostos de mosquitos virgens (subpopulação silvestre, ou *wild*). Todos os mosquitos GM da geração F_0 são idênticos e gerados a partir de um *template* contendo um ou mais elementos, conforme mencionado nas seções que descrevem cada simulação específica.

4.3 Simulações

Diversos cenários foram simulados, divididas em três diferentes experimentos, de acordo com objetivos específicos distintos. Começaremos a descrição pelos componentes e opções que todos têm em comum.

Hipóteses de modelagem: Todas as simulações assumem um conjunto de premissas simplificadoras.

- A população silvestre é originalmente virgem, i.e., não contém o transgene de interesse, a ser inserido via TEs;
- A população de cativeiro é formada com perfil cromossomal idêntico;
- A população de cativeiro é liberada uma única vez no ambiente;

- Ambos os indivíduos de cativeiro e os silvestres são sexualmente compatíveis, e produzem prole fértil.
- Mutações ocorrem com frequência grande (a cada evento evolutivo)
- Mutações ocorrem em quantidades acumuladas (2 por evento evolutivo)
- Uma geração simulada não representa uma única geração de uma população realista

A população F_0 de cativeiro é liberada no ambiente e entra em contato com a população silvestre pela primeira vez, e tem início um processo de fluxo gênico. A partir desse momento, que define o início da simulação, chamaremos tanto a população silvestre e a de cativeiro de subpopulações e nos referindo como população apenas à união destas subpopulações.

4.3.1 Protocolo comum aos experimentos

Subpopulação F_0

Todos os indivíduos colonizadores são gerados a partir de um único *template* para cada cenário. Ao total, foram utilizados três *templates* diferentes, todos com pelo menos um TE ativo.

Todos os indivíduos GM da população F_0 são idênticos a um único *template*. Existem três *templates* em uso nos experimentos, um com uma única sequência de TE ativo, o segundo com dois TEs ativos, e o terceiro com 20 TEs, dos quais um ativo e os outros 19 inativos. Todos os TEs estão localizados nas primeiras *loci* do cromossomo, de modo que nos experimentos em que é contabilizado custo de *fitness*, todos os elementos iniciais causam impacto.

No caso dos *templates* com 1 e 20 elementos, a sequência nucleotídica de cada TE tem 400 pares de bases, e todos são idênticos (sequência = TTTTTT...). No caso do *template* com 2 TEs, a sequência é semelhante à do caso com 1 TE, exceto que 5 nucleotídeos são substituídos por bases C.

Transposição

Todos os cenários usam o modelo de transposição linear, que exhibe crescimento do número de TEs por gameta com taxa constante, em relação à quantidade preexistente no

genoma parental (cf. seção 4.1.7).

Em todos os cenários, foi desativado o modelo de excisão, que é responsável por limitar o crescimento do número de TEs, removendo uma certa quantidade de TEs a cada **evento de transposição**. Dessa forma, as únicas maneiras como TEs podem ser perdidos são a perda aleatória durante a segregação ou pela recombinação. Assim a quantidade de TEs tem uma menor interferência, oferecendo uma menor complexidade a ser analisada.

Evolução

Foram escolhidos parâmetros para simplificar o processo evolutivo, de modo que a cada **evento evolutivo** ocorra sempre o conjunto de mutações estabelecido.

Além disso, o processo evolutivo é enormemente acelerado, de modo a observar os fenômenos em escala de tempo reduzida. Assim, a cada **evento evolutivo**, são incluídas 2 mutações aleatórias, com probabilidade 1 (cf. seção 4.1.9).

4.3.2 Experimento MG

A fim de ilustrar as capacidades básicas do simulador, estabelecemos um experimento que representa os elementos mais fundamentais dos modelos de dinâmica de expansão de TEs. Esse experimento não corresponde à observação de uma invasão, mas tem a finalidade de observar características marcantes de um processo de expansão de uma família de TEs segundo o modelo *Master Gene*.

Quatro cenários foram simulados, para comparar todas as combinações entre duas taxas de transposição e número de elementos ativos (MGs) em um curto período de tempo, em uma pequena população haploide (cf. figura 4.1).

Foi escolhida para esse experimento uma população constante de 1000 indivíduos, todos GM, que se reproduz livremente por 5 gerações.

Os dois *templates* utilizados são:

- 1 único TE ativo, com sequência de 400 bases, todas iguais a T.
- 2 TEs ativos, ambos com sequências de 400 bases. O primeiro TE tem a sequência idêntica ao *template* anterior. O segundo tem 5 posições substituídas por C.

Cada um desses *templates* foram expostos a duas diferentes taxas u de transposição. Foram feitos para cada caso um cenário com $u = 1$ e $u = 2$.

```

5 generations
Constant GM population (F0 = 1000)
constant transposition (1 and 2 copies)
100% initial GM
template with 1 and 2 active elements
no recombination
no fitness cost
2 mutations per event
loci (F/S/N = 0/0/500)
master gene
copy and paste
haploid

10 replicates

```

Figura 4.1: Parâmetros do protocolo de simulações do experimento MG. GM é a quantidade de indivíduos na subpopulação geneticamente modificada. F_0 é a população inicial total gerada a partir do *template*. Os *loci* são divididos em fatais (F), severos (S) e neutros (N), com suas quantidades indicadas respectivamente.

Árvores filogenéticas foram geradas usando dois métodos diferentes - inferência bayesiana (**BI = Bayesian Inference**) e **Neighbor-Joining** (NJ) - para comparar os padrões de ramificação gerados por esses diferentes processos de invasão. Esse processo será detalhado em uma seção posterior (seção 4.6.2), dedicada ao protocolo de reconstrução filogenética dos experimentos que analisam árvores das sequências simuladas.

4.3.3 Experimento A

Para ilustrar a relevância de uma elevada taxa de transposição na capacidade de uma família de TEs suplantarem o custo de *fitness* que ele impõe à população hospedeira ao invadi-la, foi elaborado um experimento com população constante com taxas de transposição simples.

Na primeira etapa, definimos o controle do experimento, isto é um processo de expansão dos TEs na população hospedeira sem nenhum custo de *fitness*. Diversos cenários distintos foram simulados, todos usando o modelo de transposição linear.

Dois *templates* com quantidades iniciais de TEs foram usados, o primeiro com um único TE, e o segundo com 20 TEs. Ambos *templates* foram submetidos a cenários com

três diferentes taxas de transposição u : a primeira taxa de $u = 1$ nova cópia por evento de transposição, a segunda de $u = 5$ novos elementos e a terceira de $u = 10$ novas cópias por evento de transposição. Essas taxas foram simuladas ao longo de 30 gerações numa população constante de 10.000 indivíduos, com uma quantidade inicial de 1000 indivíduos GM, representando portanto 10% de proporção de invasores. Os parâmetros que definem estes cenários estão resumidos nos protocolos 4.2 e 4.3.

```

30 generations
constant population (F0 = 10k)
constant transposition (1, 5 and 10 copies)
10% initial GM
template with 1 elements
no recombination
no fitness cost
2 mutations per event
loci (F/S/N = 0/0/250)
master gene
copy and paste
diploid

10 replicates

```

Figura 4.2: Parâmetros do protocolo de simulações do experimento A com *template* contendo 1 TE e sem impacto no *fitness* ($s = 0$). GM é a quantidade de indivíduos na subpopulação geneticamente modificada. F_0 é a população inicial total gerada a partir do *template*. Os *loci* são divididos em fatais (F), severos (S) e neutros (N), com suas quantidades indicadas respectivamente.

Os experimentos de interesse no entanto, precisam levar em conta uma diminuição na fecundidade dos hospedeiros induzidos pela inserção e expansão de TEs em seus cromossomos. Para tanto, impusemos dois valores relativos a um custo pequeno ($s = 0,01$) e um custo moderado ($s = 0,05$) por TE.

Esses dois valores de impacto seletivo foram testados para o *template* com 1 TE. Para esse *template*, simulamos cenários para as três taxas de transposição $u = 1$, $u = 5$ e $u = 10$ novas cópias por gameta por geração, ao longo de 30 gerações em população constante de 10.000 indivíduos, com uma proporção inicial de 1000 indivíduos GM. O TE inicial do *template* usado para compor a população F_0 está localizado em um *locus* **severo** (c. seção 4.1.3) e portanto é sempre contabilizado para a redução de fecundidade dos hospedeiros

```
30 generations
constant population (F0 = 10k)
constant transposition (1, 5 and 10 copies)
10% initial GM
template with 20 elements
no recombination
no fitness cost
2 mutations per event
loci (F/S/N = 0/0/250)
master gene
copy and paste
diploid

10 replicates
```

Figura 4.3: Parâmetros do protocolo de simulações do experimento A com *template* contendo 20 TEs e sem impacto no *fitness* ($s = 0$). GM é a quantidade de indivíduos na subpopulação geneticamente modificada. F_0 é a população inicial total gerada a partir do *template*. Os *loci* são divididos em fatais (F), severos (S) e neutros (N), com suas quantidades indicadas respectivamente.

GM. Estes parâmetros estão sumarizados no protocolo 4.4.

Por fim, para testar um impacto excessivo na fecundidade da população hospedeira, utilizamos o *template* com 20 TEs iniciais, submetido ao impacto seletivo $s = 0,05$. Assim como nos cenários do *template* de 1 único transposon seletivos, simulamos cenários para as mesmas três taxas de transposição $u = 1$, $u = 5$ e $u = 10$ novas cópias por gameta por geração, ao longo de 30 gerações na população constante de 10.000 indivíduos, com uma proporção inicial de 1000 indivíduos GM. Todos os 20 TEs do *template* usados na população F_0 estão localizados em *loci* **severos**, portanto iniciam causando impacto na fecundidade dos hospedeiros que os carregam (cf. seção 4.1.3). É necessário observar no entanto que todos os 20 TEs iniciais estão localizados em um único cromossomo, e portanto na ausência de transposição e recombinação *crossover*, haveria apenas 50% de chance de que um filho de um indivíduo GM da população F_0 com um indivíduo virgem receba TEs. Estes parâmetros estão sumarizados no protocolo 4.5.

Todos os cenários descritos acima usam o modelo *Master Gene* de inativação, de modo que não há criação de novas cópias ativas nos eventos de transposição. É esperado dessa forma um declínio gradual no número total de elementos ativos, conforme estes são

```

30 generations
constant population (F0 = 10k)
constant transposition (1,5 and 10 copies)
10% initial GM
template with 1 elements
no recombination
fitness cost 0.01 and 0.05
2 mutations per event
loci (F/S/N = 0/50/450)
master gene
copy and paste
diploid

10 replicates

```

Figura 4.4: Parâmetros do protocolo de simulações do experimento A com *template* contendo 1 TE e impactos no *fitness* fraco ($s = 0,01$) e moderado ($s = 0,05$). GM é a quantidade de indivíduos na subpopulação geneticamente modificada. F_0 é a população inicial total gerada a partir do *template*. Os *loci* são divididos em fatais (F), severos (S) e neutros (N), com suas quantidades indicadas respectivamente.

perdidos ao acaso por deriva genética.

A quantidade total de TEs na população, e o número de TEs ativos são registrados a cada geração pelo simulador, e esses valores foram recuperados e sumarizados em gráficos e tabelas para ilustrar o comportamento de cada cenário.

4.3.4 Experimento B

Para testar o poder preditivo das técnicas de filodinâmica em dados de sequências de transposons, é necessário observar a dinâmica de expansão de uma família de TEs em diversos cenários de dinâmica populacional. Para isso, contrapomos a expansão populacional, sem custo s de fitness, em vários cenários ecológicos.

Para o experimento B, assim como nos experimentos anteriores, também simulamos usando o modelo *Master Gene*, de forma similar ao experimento A, mas dessa vez a comparação se deu principalmente quanto aos diferentes possíveis modelos ecológicos. O objetivo é observar o efeito indireto do tamanho populacional do vetor, na dinâmica de invasão dos TEs.

```
30 generations
constant population (F0 = 10k)
constant transposition (1, 5 and 10 copies)
10% initial GM
template with 20 elements
no recombination
fitness cost 0.05
2 mutations per event
loci (F/S/N = 0/250/250)
master gene
copy and paste
diploid

10 replicates
```

Figura 4.5: Parâmetros do protocolo de simulações do experimento A com *template* contendo 20 TEs e impacto moderado no *fitness* ($s = 0,05$). GM é a quantidade de indivíduos na subpopulação geneticamente modificada. F_0 é a população inicial total gerada a partir do *template*. Os *loci* são divididos em fatais (F), severos (S) e neutros (N), com suas quantidades indicadas respectivamente.

Para cada modelo ecológico, foi usado um número máximo de 100 gerações em cada cenário, com população sexual diploide, modelo de transposição linear, com taxa de crescimento constante do número de TEs por geração igual a 5 novas cópias a cada evento reprodutivo, independente do número existente de elementos nos cromossomos parentais.

- População constante
- População com crescimento exponencial
- População com decrescimento exponencial
- População logística

O cenário de população constante inicia com população F_0 de 10.000 indivíduos (protocolo 4.6). Tanto o cenário de população com crescimento exponencial como o de crescimento logístico iniciam com $F_0 = 1.000$ indivíduos. O cenário de população com decrescimento exponencial inicia com $F_0 = 20.000$ indivíduos.

Os parâmetros populacionais do modelo logístico são: taxa de crescimento $r = 1$ e capacidade suporte $K = 10.000$ (protocolo 4.9). Os parâmetros de crescimento dos modelos exponenciais decrescente e crescente são, respectivamente, $r = -0,1$ e $r = 0,1$ (protocolos 4.7 e 4.8, respectivamente).

Todos os cenários iniciam com uma subpopulação GM correspondente a 10% da população F_0 , assim como no experimento A.

```

100 generations
constant population (F0 = 10k)
constant transposition (5 copies)
10% initial GM
template with 1 active element
no recombination
no fitness cost
2 mutations per event
loci (F/S/N = 0/50/450)
master gene
copy and paste
diploid

10 replicates

```

Figura 4.6: Parâmetros do protocolo de simulações do experimento B com população de mosquitos constante ($N = 10.000$). GM é a quantidade de indivíduos na subpopulação geneticamente modificada. F_0 é a população inicial total gerada a partir do *template*. Os *loci* são divididos em fatais (F), severos (S) e neutros (N), com suas quantidades indicadas respectivamente.

Além dos vários cenários populacionais, foi simulado um cenário com custo de fitness $s = 0,05$ para população constante. Foi usada uma taxa de transposição de $u = 5$ novas cópias por gameta, conforme descrito no Protocolo 4.10.

Assim como nos experimentos MG e A, foram feitas 10 réplicas de cada um desses cenários, e em cada réplica é amostrado um indivíduo da última geração e seu perfil cromossômico é separado para análise das sequências.

O número de TEs de cada amostra foi observada, e foi escolhido para análise a amostra com maior número de sequências, observando-se o número mínimo de 50 sequências. Nos casos em que não tenha sido amostrado pelo menos um indivíduo com 50 sequências, foi


```
100 generations
exponential population (F0 = 20k)
r = -0.01
constant transposition (5 copies)
10% initial GM
template with 1 active element
no recombination
no fitness cost
2 mutations per event
loci (F/S/N = 0/0/500)
master gene
copy and paste
diploid

10 replicates
```

Figura 4.7: Parâmetros do protocolo de simulações do experimento B com população de mosquitos exponencial decrescente ($r = -0,1$). GM é a quantidade de indivíduos na subpopulação geneticamente modificada. F_0 é a população inicial total gerada a partir do *template*. Os *loci* são divididos em fatais (F), severos (S) e neutros (N), com suas quantidades indicadas respectivamente.

feita uma nova replicada e ao final dela, foi amostrado manualmente um indivíduo que tivesse pelo menos 50 TEs em seu genoma.

4.4 Análises dos dados simulados

Cada um dos experimentos tem requerimentos de análises distintas.

- No experimento MG foi feita a reconstrução de árvores filogenéticas para a observação dos padrões de ramificações, usando-se dois métodos diferentes (BI e NJ).
- No experimento A foram analisadas as estatísticas das populações e subpopulações de cada cenário.
- No experimento B foi feita a reconstrução filogenética bayesiana dos genomas amostrados, e a estimação do modelo demográfico da população de TEs.

Nas seções a seguir serão detalhados os procedimentos de cada uma dessas etapas.

```
100 generations
exponential population (F0 = 1k)
constant transposition (5 copies)
r=0.1
10% initial GM
template with 1 active element
no recombination
no fitness cost
2 mutations per event
loci (F/S/N = 0/0/500)
master gene
copy and paste
diploid

10 replicates
```

Figura 4.8: Parâmetros do protocolo de simulações do experimento B com população de mosquitos exponencial crescente ($r = 0,1$). GM é a quantidade de indivíduos na subpopulação geneticamente modificada. F_0 é a população inicial total gerada a partir do *template*. Os *loci* são divididos em fatais (F), severos (S) e neutros (N), com suas quantidades indicadas respectivamente.

4.5 Análises gráficas

4.5.1 Experimento A

Os dados gerados pelo simulador a cada geração (e para cada réplica) incluem

- o tamanho total da população;
- o tamanho da subpopulação GM;
- a quantidade total de TEs presente na população;
- a quantidade total de TEs ativos na população;
- a quantidade média de TEs por indivíduo.

Para o experimento A, todos esses dados foram exibidos em gráficos de dinâmica temporal, para cada taxa de transposição examinada. A métrica de interesse para detectar

```
100 generations
logistic population (F0 = 1k)
r= 1, K=10k
constant transposition (5 copies)
10% initial GM
template with 1 active element
no recombination
no fitness cost
2 mutations per event
loci (F/S/N = 0/0/500)
master gene
copy and paste
diploid

10 replicates
```

Figura 4.9: Parâmetros do protocolo de simulações do experimento B com população de mosquitos logística ($r = 1$ e $K = 10.000$). GM é a quantidade de indivíduos na subpopulação geneticamente modificada. F_0 é a população inicial total gerada a partir do *template*. Os *loci* são divididos em fatais (F), severos (S) e neutros (N), com suas quantidades indicadas respectivamente.

e analisar o comportamento invasivo é a proporção de indivíduos GM na população, mas como todos os cenários assumem população constante, essa proporção corresponde isometricamente à quantidade total da subpopulação GM. Por essa razão, não foi necessário fazer a divisão dos dados pelo tamanho total da população em cada geração.

Os dados da dinâmica temporal da subpopulação GM foram sobrepostos graficamente para todas as réplicas de cada cenário. Para análise global de cada cenário, a média de todas as réplicas foi tomada para cada geração.

Os dados da dinâmica temporal dos TEs foram considerados apenas por suas médias, de maneira análoga aos da subpopulação GM descritos acima. Além desses, fizemos também os gráficos para cada cenário da evolução do número de TEs ativos de cada cenário, representados também pelas médias das replicadas, tomadas a cada geração.

Elaboramos também tabelas para sumarizar os dados de TEs para cada cenário. Para muitos cenários, nem todas as réplicas chegaram ao número máximo de gerações programado, tendo sido interrompidas por invasão ou perda de TEs. Dessa forma, o valor representado na tabela como a média do número de TEs ou média de TEs ativos foram

```

100 generations
constant population (F0 = 10k)
constant transposition (5 copies)
10% initial GM
template with 1 active element
no recombination
fitness cost 0.05
2 mutations per event
loci (F/S/N = 0/50/450)
master gene
copy and paste
diploid

10 replicates

```

Figura 4.10: Parâmetros do protocolo de simulações do experimento B com população de mosquitos constante ($N = 10.000$), e custo moderado de fitness ($s = 0,05$). GM é a quantidade de indivíduos na subpopulação geneticamente modificada. F_0 é a população inicial total gerada a partir do *template*. Os *loci* são divididos em fatais (F), severos (S) e neutros (N), com suas quantidades indicadas respectivamente.

tomados considerando a última geração em que todas as réplicas estavam ativas. Como esse valor não representa necessariamente o potencial daquele cenário, os valores máximos de TEs e TEs ativos de cada cenário também foram incluídos nas tabelas.

4.6 Filodinâmica

Para a análise filodinâmica dos dados simulados são necessárias duas etapas:

1. a reconstrução filogenética com um componente temporal;
2. a inferência de parâmetros demográficos a partir das genealogias inferidas;

A reconstrução filogenética é necessária para a observação dos padrões de ramificação, que insinuam informações sobre os padrões e pressões seletivas que sublinham o processo evolutivo em questão.

A inferência de um modelo demográfico objetiva determinar o tamanho da “população” de patógenos, no caso da filodinâmica clássica, ou de TEs no nosso caso, que é influenciada

pelo fenômeno epidemiológico ou de invasão, respectivamente.

4.6.1 Protocolo comum às topologias

O protocolo resumido para geração de árvores filogenéticas e modelos bayesianos segue as seguintes hipóteses simplificadoras:

- As sequências geradas são homólogas próximas;
- Todos os *loci* alinhados são homólogos, devido a ausência de *indels*;
- Cada sítio nucleotídico é modificado de acordo com um modelo evolutivo uniforme (Jukes-Cantor) [80];

Essas premissas são garantidas pelo desenho experimental das simulações, que por sua vez corresponde às premissas feitas na escolha de parâmetros dos **cenários de simulação** em questão.

Amostragem de sequências para análise

Conforme mencionado anteriormente, para cada cenário foram feitas 10 simulações com parâmetros idênticos, correspondendo a 10 réplicas para cada parte do experimento. Ao final de cada réplica é amostrado um indivíduo da geração mais recente da população. Quando as amostras de todas as réplicas são comparadas, é escolhido para análise o indivíduo com maior número de TEs para a reconstrução filogenética (cf. seção 4.2.1). As réplicas amostradas para o experimento B buscam um número mínimo de 50 sequências.

Outgroups

Em todos os experimentos em que são feitas árvores, a raiz da mesma é determinada utilizando-se um *outgroup* definido por duas sequências de 400 nucleotídios, dos quais os primeiros 25 e 30 (respectivamente) correspondem à letra C, e todos os outros iguais à letra T. O mesmo grupo de outgroups foi utilizado em todas as análises.

4.6.2 Filogenética - componente temporal

As topologias de destaque foram reconstruídas de acordo com um processo de inferência bayesiana (BI) para os experimentos MG e B.

Adicionalmente, foi utilizado um método de distâncias (**NJ = Neighbor-Joining**) para o experimento MG.

O programa de filogenias bayesiano **BEAST 1.7.0** [138, 139] usa como *input* um arquivo XML com todos os parâmetros necessários para suas análises, bem como as sequências e nomes dos arquivos a ser gerados. Esse arquivo XML é gerado por um programa de interface chamado **BEAUTI 1.7.0**, presente no mesmo pacote do **BEAST**, a partir de um alinhamento previamente estabelecido. As sequências obtidas em cada amostra foram inseridas no programa **BEAUTI** para preparar o arquivo XML que foi inserido no programa **BEAST**, de acordo com as premissas acima, e seguindo as escolhas de parâmetros a seguir.

Neste programa **BEAUTI**, foram escolhidos os seguintes parâmetros a ser estimados:

1. Os parâmetros do modelo evolutivo;
2. O componente temporal, representado pela taxa do relógio molecular;
3. O componente demográfico determinado pela premissa filogenética;
4. A topologia da árvore, e seus comprimentos de ramos;

As inferências dos programas do pacote **BEAST** são analisadas usando o programa **TRACER**, desenvolvido pelo mesmo grupo de pesquisa que desenvolve o pacote **BEAST**. Nele podem ser observados os valores estimados para o tamanho efetivo das amostras (ESS - Effective Sample Size), e o traço da análise. O ESS indicado pelos autores como valor de corte para os parâmetros de interesse é 100. O traço é uma maneira gráfica de observar a evolução das cadeias MCMC (Markov Chain Monte-Carlo), e sua inspeção visual pode indicar o estado de convergência das cadeias. Essa inspeção visual tem o objetivo de determinar se este se encontra em um estado estacionário, ou se há alguma tendência clara de crescimento, decrescimento ou oscilação que indicaria que as cadeias precisam rodar por mais tempo, ou alguma alteração nos *prioris* ou dados é necessária.

Protocolo comum aos experimentos MG e B

Sendo uma abordagem bayesiana, é necessário introduzir os *priori* de todos esses parâmetros, de modo que para cada estimador acima os seguintes *prioris* foram respectivamente arbitrados (detalhados a seguir):

1. Utilizamos o modelo Jukes-Cantor na análise de evolução molecular;
2. O modelo de relógio molecular selecionado foi o relógio estrito (*strict clock*), com taxa igual a 1;
3. O modelo demográfico utilizado para as sequências foi o *bayesian skyline plot*);
4. A topologia inicial da árvore foi obtida com o método de distâncias UPGMA.

O número de iterações para as cadeias MCMC, e números de amostras foi diferente entre os experimentos MG, e o experimento B.

Nos experimentos MG e B estamos simulando um TE Classe I, que usa como estágio intermediário da transposição uma retro-transposase, e portanto se replica pelo modo *copy and paste* para gerar novas cópias (cf. seção 4.1.4).

Foram escolhidos para as simulações parâmetros que simplificam a observação do processo evolutivo, de modo que a cada **evento evolutivo** ocorra sempre o mesmo número de mutações. Além disso, o processo evolutivo foi consideravelmente acelerado, com cada **evento evolutivo** observando a inclusão de 2 mutações aleatórias (cf. seção 4.1.9).

Como a posição de cada mutação inserida é uniformemente sorteada ao longo dos sítios, e não há partição de códons ou nenhuma estrutura intrínseca para as sequências, um método de reconstrução filogenética de distâncias deve obter um *priori* adequado para a topologia da árvore. Nos casos em que há acumulação de substituições em um mesmo sítio, algum sinal filogenético pode ser perdido, mas a probabilidade desses eventos diminui conforme o tamanho das sequências aumenta. Um corolário dessa construção é que as *taxa* satisfazem a hipótese do relógio molecular.

Para representar o modelo Jukes-Cantor no BEAST, usamos o protocolo sugerido pelos autores de usar o modelo GTR, com todas as taxas iguais, e desabilitar a inferência dos parâmetros relativos a taxas de substituição⁷.

⁷http://beast.bio.ed.ac.uk/FAQ#How_to_make_use_of_other_substitution_models_not_given_in_Beauti.3F_Specifically.2C_Jukes-Cantor_model.3F

Topologias do experimento MG

O objetivo desse experimento é observar as topologias pectinadas quando há um único elemento ativo (MG). Para comparar com múltiplos elementos ativos, foram simulados também cenários com dois elementos ativos. O protocolo para análise filogenética do experimento está sumarizado na tabela 4.1.

Para esse experimento, quatro cenários de simulação foram executados, para uma população haploide (cf. protocolo 4.1). Nos dois primeiros, existe um único elemento ativo (denominado MG no que se segue). Nos dois cenários seguintes, existem dois elementos MG.

As cadeias MCMC foram rodadas por um total de 10 milhões de iterações, com amostragem a cada 100 iterações tanto para os parâmetros quanto para as topologias.

Tabela 4.1: Protocolo para obtenção das topologias do experimento MG conforme inserido no programa BEAUTI.

Parâmetro	opção principal	opção(ões) auxiliar(es)
Modelo de sítios	GTR	all equal
Modelo de relógio	relógio estrito	estimar parâmetro, taxa=1
Demografia	skyline plot	5 grupos, variante linear
Prior de topologia	árvore de distâncias (UPGMA)	NA
Prior do relógio	uniforme [0,1]	NA
Operadores	desabilitar parâmetros GTR	ac, ag, at, cg, gt
MCMC	10 milhões de iterações	amostrar a cada 100 cadeias

Como o modelo de evolução molecular envolvido nas simulações é extremamente simples, e a distância evolutiva entre as sequências é pequena (dado que o tempo de divergência é extremamente pequeno), especulamos inicialmente que um método de reconstrução filogenética de distâncias seria razoavelmente apropriado para obter as topologias das árvores.

Para testar essa hipótese, reconstruímos uma árvore adicional com o método **NJ** (**Neighbor-Joining**) utilizando o programa **ClustalW 2.1** [140] e comparamos essa à árvore **BI** criada com o **BEAST 1.7**.

Topologias experimento B

No experimento B, não só as topologias das árvores devem ter bom padrão de convergência, mas também diversos outros parâmetros e estatísticas serão analisados, por-

tanto as cadeias MCMC devem rodar por mais tempo a fim de observar um estado estacionário das distribuições amostradas. A tabela 4.2 descreve os parâmetros utilizados na análise.

Tabela 4.2: Protocolo para obtenção das topologias do experimento B, conforme inserido no programa BEAUTI.

Parâmetro	opção principal	opção(ões) auxiliar(es)
Modelo de sítios	GTR	all equal
Modelo de relógio	relógio estrito	estimar parâmetro, taxa=1
Demografia	skyline plot	5 grupos, variante linear
Prior de topologia	árvore de distâncias (UPGMA)	NA
Prior do relógio	uniforme [0,1]	NA
Operadores	desabilitar parâmetros GTR	ac, ag, at, cg, gt
MCMC	100 milhões de iterações	amostrar a cada 1000 cadeias

Para as análises desse experimento, foi utilizado um número de 100 milhões de iterações para as cadeias MCMC e topologias, com amostras feitas a cada 1000 cadeias. Inspeção dos valores de ESS e dos traços obtidos pelo programa TRACER indicaram que a convergência foi atingida para as cadeias de todos os cenários do experimento.

4.6.3 Filogenética - componente demográfico

Experimento B

Para inferir um modelo demográfico a partir das topologias do experimento B, alimentamos o programa BEAST com opções de parâmetros compatíveis com as simulações.

Foram gerados modelos demográficos não paramétricos diretamente a partir das topologias usando o método *Bayesian skyline plot* [89] implementado no pacote BEAST 1.7 [138, 139]. Os parâmetros inseridos na seção 4.6.2, subseção **Topologias experimento B**, detalham as opções feitas para a inferência desse componente demográfico. Aqui detalhamos a análise demográfica feita com aqueles resultados.

A análise demográfica foi feita também no programa TRACER, usando como parâmetros o modelo bayesian skyline plot, com sua variante linear. Os traços analisados correspondem à mediana das amostras.

Tabela 4.3: Protocolo para obtenção dos modelos demográficos não-paramétricos do experimento B via bayesian skyline plot, conforme inserido no TRACER

Parâmetro	opção principal	opção(ões) auxiliar(es)
Demografia	skyline plot	5 grupos, variante linear
altura da árvore	Mediana	

4.6.4 Tratamento das árvores obtidas

As topologias de árvores amostradas no programa BEAST foram sumarizadas utilizando o programa TREEANNOTATOR 1.7.0, fornecido conjuntamente com o BEAST, para obtenção da árvore de maior credibilidade (**MCC = Maximum Clade Credibility** - Clado de máxima credibilidade). Foram usados todos os valores *default*, via linha de comando. Foram descartadas as primeiras 10% amostras (*burn-in*).

Foram analisadas, no caso da topologia BI, a topologia de maior credibilidade (árvore MCC) dentre as amostradas, e no caso da topologia NJ a única árvore obtida. As árvores a ser exibidas foram formatadas utilizando-se o programa FIGTREE 1.3.1.

Os ramos e cabeçalhos correspondentes às linhagens MG foram coloridas para evidenciar os TEs ativos, e as ramificações foram ordenadas de forma a facilitar a visualização da ordem prevista pela idade de cada sequência, de acordo com a informação em seus cabeçalhos individuais que determina em que geração cada uma foi criada. Nos casos em que duas ou mais sequências criadas na mesma geração ficaram agrupadas em sub-clados monofiléticos, estes foram agrupados usando a opção *cartoon*, que representa o clado como um triângulo.

Nas árvores em que foram utilizados um *outgroup* e o método de reconstrução filogenética não foi capaz de posicionar a raiz do clado de interesse no ramo ligado ao *outgroup*, foi feito um reenraizamento para classificar corretamente o clado.

Capítulo 5

Resultados

“There is nothing more deceptive than an obvious fact.”

Sherlock Holmes in “The Boscombe Valley Mystery” (1891)

5.1 Experimento MG

5.1.1 Topologias esperadas para um processo Master Gene

Nosso modelo computacional indicou que uma família de TEs que se expande a partir de um único elemento ativo cria topologias **pectinadas** (em forma de pente), com todas as ramificações partindo da linhagem do elemento MG (o único TE ativo). Esse é o comportamento esperado para esse fenômeno [135, 141].

Todos os quatro cenários simulados para testar as hipóteses master gene fazem uso de uma premissa simplificadora: o organismo representado é haploide. Isso concentra uma maior quantidade de TEs no genoma, proporcionalmente ao que seria esperado caso houvessem dois cromossomos.

Lembramos que os cabeçalhos das sequências contém uma codificação da genealogia das mesmas, determinando a geração em que cada sequência passou a existir, e o MG de origem (cf. seção 4.1.10).

As árvores feitas com o método bayesiano têm uma estrutura temporal implicada pela adição da hipótese do relógio molecular. Isso torna possível estimar o tempo de divergência entre as linhagens envolvidas, e se a análise estiver correta, pode ser usado para inferir o número aproximado de gerações que ocorreu na simulação.

Observação: Se a inferência temporal estiver correta, os comprimentos dos ramos devem codificar a cronologia em que as sequências foram criadas, de modo que as ramificações devem ficar aproximadamente alinhadas de acordo com sua geração de origem.

Veremos que nos cenários em que houve boa aproximação pelo componente temporal, as linhagens criadas em uma mesma geração ficam agrupadas, e os clados correspondentes a cada geração ficam ordenados hierarquicamente, representando a ordem das mesmas.

No cenário mais simples, com um único MG que cria um único novo TE por geração, pode-se observar na figura 5.1 a estrutura pectinada da topologia da árvore, com cada nova linhagem de TE se ramificando a partir da linhagem do MG.

No cenário com maior quantidade de ramificações, com um único MG mas que gera duas novas linhagens inativas a cada intervalo de tempo, observa-se na figura 5.2 que todos os TEs criados simultaneamente são corretamente agrupados em clados duplos,

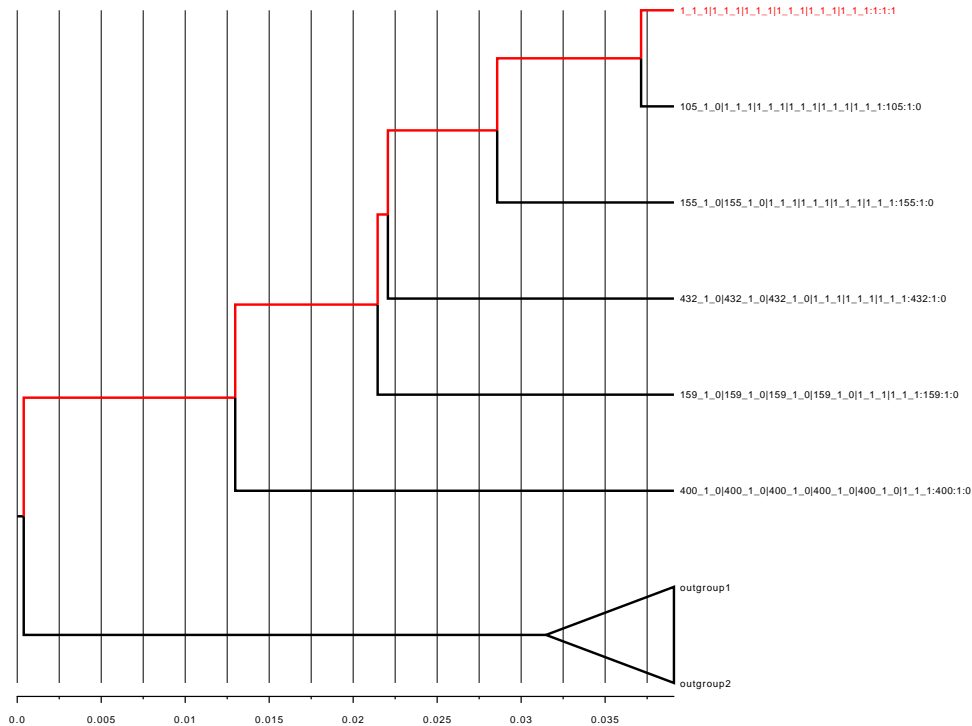


Figura 5.1: Experimento MG: topologia reconstruída com o método bayesiano a partir de uma invasão de TEs numa população de mosquitos virgens com $u = 1$ transposição por geração, e um único elemento ativo (MG), marcado em vermelho. A grade em escala temporal foi introduzida para evidenciar a estrutura determinada pelas sucessivas gerações.

com exceção dos TEs criados na última geração, que ficam agrupados num clado com três sequências, incluindo-se o elemento MG. Esse comportamento foi observado também em todos os outros cenários com taxa de transposição maior que 1.

As três sequências desse clado tem a mesma idade, expressa implicitamente pelo número esperado de mutações que acumularam ao longo do processo evolutivo, e deveriam portanto ser agrupadas em um clado triplo. Atribuímos essa falta de resolução no clado correspondente à geração mais recente não à perda de sinal ou presença de ruído filogenético, mas ao processo de criação de árvores que assume que as linhagens ocorrem sempre em bifurcações.

Na figura 5.3, vemos o cenário que corresponde pela primeira vez à introdução de mais de uma linhagem ativa na população. Com dois MGs, a topologia fica consideravelmente mais rica e portanto perde a forma pectinada ao longo da árvore. Porém, a estrutura

5.1.2 Neighbor-Joining

O método do Neighbor-Joining utilizado encontrou uma topologia com algumas características pontuais corretas nos cenários mais simples examinados no experimento MG, mas falhou em agrupar os TEs de acordo com suas origens, isto é, com respeito a ordem de criação definida pela genealogia. Além disso, o posicionamento da raiz do *ingroup* implicado pelo *outgroup* ficou incorreto em todos os cenários. Assim, nenhuma das árvores reconstruídas usando o método NJ ficaram com estruturas pectinadas. Por essas razões, esse método foi descartado para análises de cenários que envolvam premissas mais realistas do que esse exemplo, como o experimento B.

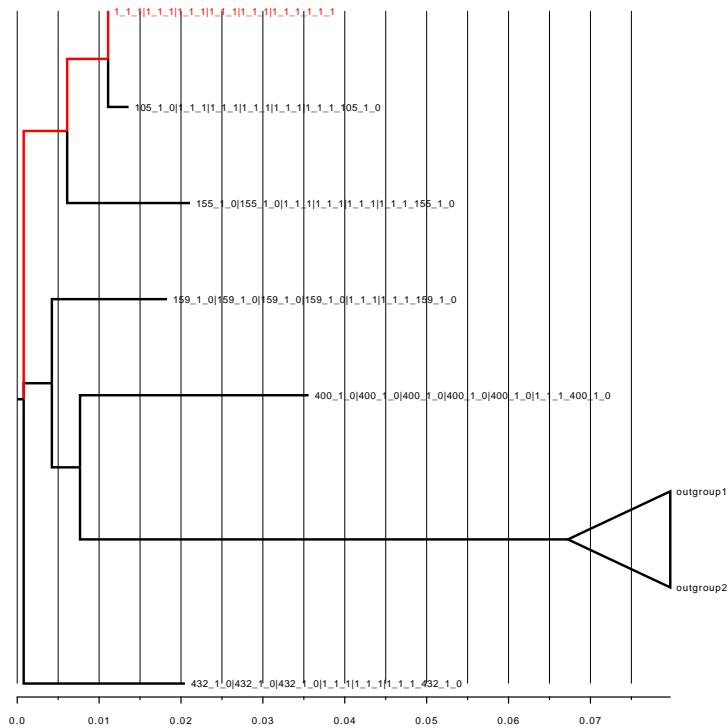


Figura 5.5: Experimento MG: topologia reconstruída com o método NJ a partir de uma invasão de TEs numa população de mosquitos virgens com $u = 1$ transposição por geração, e um único elemento ativo (MG), marcado em vermelho. A grade em escala temporal foi introduzida para evidenciar a estrutura determinada pelas sucessivas gerações.

Na figura 5.5 pode-se observar que a estrutura pectinada não foi encontrada ao longo de toda a árvore, com três TEs (159_1_0, 400_1_0 e 432_1_0) se agrupando distantes do clado que contém o MG. Além disso, esses três TEs não formam em si um clado pectinado,

o que seria desejável pelo desenho experimental. O erro na inferência da topologia parece se basear na ausência do TE 432_1_0 no clado, que ficou numa linhagem isolada e pertence à terceira geração, que é justamente a que faltou na estrutura pectinada.

A ordem relativa dos TEs 159_1_0 e 400_1_0, representada por seus comprimentos de ramos, está de acordo com as suas respectivas gerações de origem (primeira e segunda, respectivamente). Apesar da topologia estar equivocada, os comprimentos dos ramos são compatíveis com as idades esperadas para cada TE.

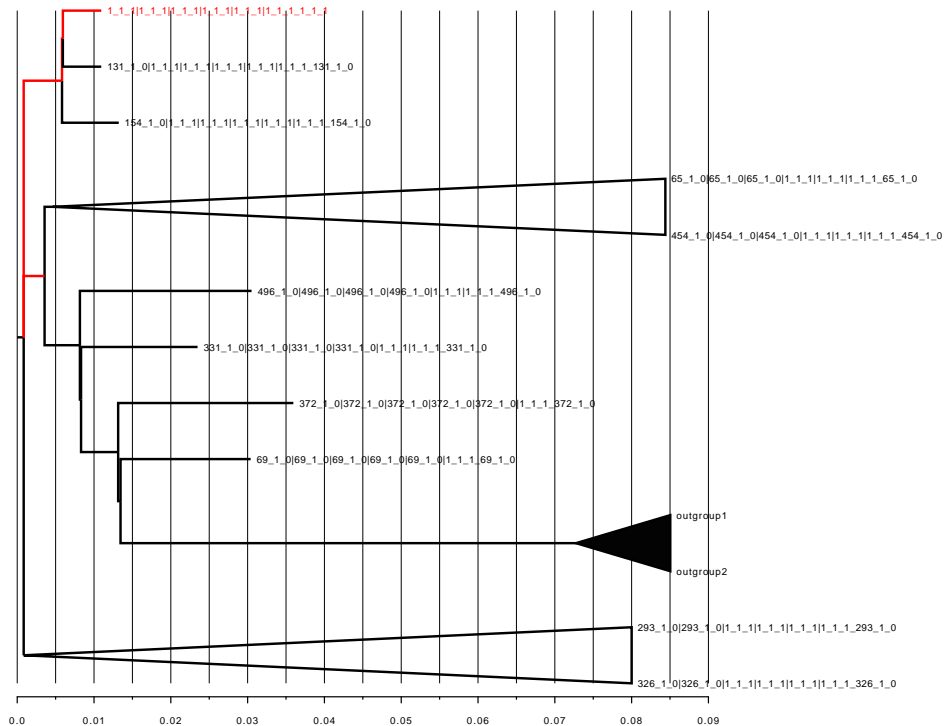


Figura 5.6: Experimento MG: topologia reconstruída com o método NJ a partir de uma invasão de TEs numa população de mosquitos virgens com $u = 2$ transposições por geração, e um único elemento ativo (MG), marcado em vermelho. Múltiplas transposições em uma mesma geração foram agrupadas quando possível. A grade em escala temporal foi introduzida para evidenciar a estrutura determinada pelas sucessivas gerações.

Esse problema pode ter sido causado pela pequena quantidade de dados, no entanto. Um indício disso pode ser observado no cenário onde duas novas cópias são geradas a cada geração, ao invés de uma única (figura 5.6). Neste cenário, todos os pares de TEs criados simultaneamente ficam corretamente agrupados próximos. Apesar disso, não foram formados clados duplos em todos os casos, e esses agrupamentos parcialmente corretos não

ficam organizados na hierarquia temporal real, representando uma ordem incorreta das gerações em que foram criados.

Além disso, parece haver uma falta de sinal filogenético suficiente para resolver as ramificações próximas da raiz, fazendo com que o agrupamento entre os TEs 69_1_0 e 372_1_0 fica incorretamente agrupado com os TEs 331_1_0 e 496_1_0 e o outgroup.

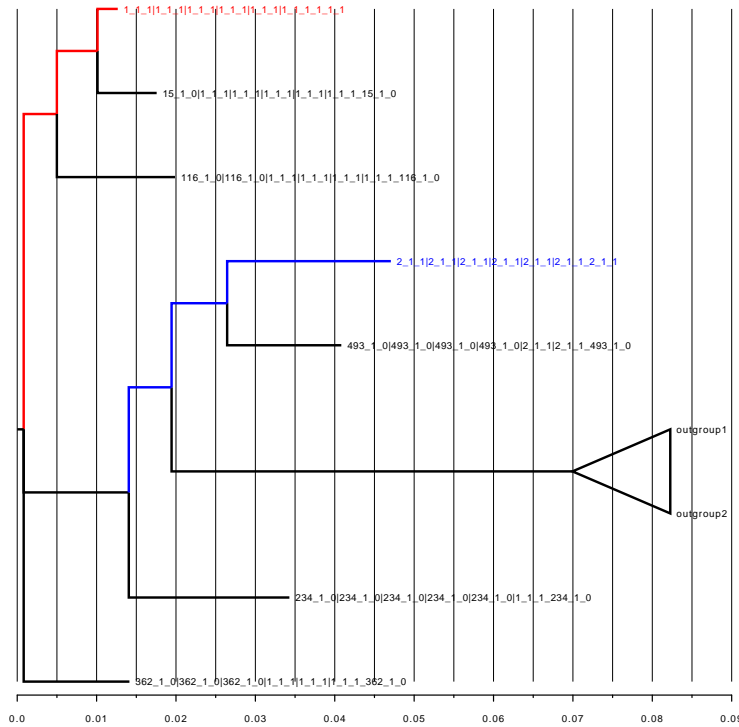


Figura 5.7: Experimento MG: topologia reconstruída com o método NJ a partir de uma invasão de TEs numa população de mosquitos virgens com $u = 1$ transposição por geração, e 2 elementos ativos (MG), marcados em vermelho e azul. A grade em escala temporal foi introduzida para evidenciar a estrutura determinada pelas sucessivas gerações.

Alguns problemas também ocorrem ao se considerar duas linhagens independentes, fontes de novos TEs. Considerando o caso do cenário com dois MGs e uma única transposição por geração da figura 5.7, observa-se que a inferência foi próxima da correta, com as ramificações pectinadas encontradas corretamente, e representando a ordem de criação dos elementos. Porém, como a raiz ficou incorretamente posicionada, o TE 362_1_0, que deveria ter sido agrupado na linhagem do MG 1_1_1, ficou isolado do *ingroup* e o TE 234_1_0 ficou agrupado mais próximo da linhagem do MG 2_1_1.

Nesse caso a falta de sinal filogenético próximo da raiz da árvore pode ser um indício

5.1.3 Considerações gerais

Para a execução desse experimento, o modelo foi adaptado para representar uma população hipotética sexuada com genoma haploide. Isso foi feito para aproximá-lo das premissas dos modelos representados nos trabalhos de Brookfield (2005 [135]), e Brookfield, Johnson (2006 [141, 142]). Além disso, a representação dos dados fica consideravelmente simplificada, pois: (1) um genoma haploide atenua sensivelmente a amplificação dos TEs, fazendo com que o tamanho final da família de TEs seja menor que em um genoma diploide; (2) eliminamos assim o risco de haver dois TEs ativos no indivíduo amostrado, ocupando o *locus* previsto em cada um dos dois cromossomos.

A primeira observação a ser feita é de que de fato os resultados do modelo computacional são coerentes com os resultados analíticos e simulações das equações propostas em [141], pois as árvores encontradas têm estrutura pectinada, seguindo a linhagem de cada elemento MG. Como a população é relativamente grande, o fenômeno de *genetic drift* não têm efeito perceptível na expansão da família de TEs, mesmo considerando que foi simulado apenas um pequeno número de gerações para aproximação de um equilíbrio do número de elementos.

5.2 Experimento A

No experimento A, cujo objetivo era observar as condições em que TEs podem invadir ou ser perdidos numa nova população hospedeira, foram simuladas taxas de transposição de $u = 1$, $u = 5$ e $u = 10$ novos elementos por gameta por geração. Esses cenários foram divididos entre os casos em que os TEs nunca causam redução de *fitness* em seus hospedeiros, e dois casos em que há impacto pequeno e moderado em sua capacidade reprodutiva, representados por custos de *fitness* iguais a $s = 0,01$ e $s = 0,05$, respectivamente.

5.2.1 Sem custo de *fitness* (controle)

Template com 1 TE, sem custo de *fitness*

A primeira configuração consistiu em simular populações sendo invadidas por uma subpopulação F_0 contendo um único TE ativo, e nesses cenários não há impacto de *fitness* perceptível na população hospedeira. Essa etapa portanto é o controle do experimento que visa analisar a característica “egoísta” do mecanismo de invasão dos TEs, a ser analisado a seguir quando é introduzido um custo de *fitness* por TE para os hospedeiros portadores.

Foram simulados diversos cenários com três diferentes taxas constantes de transposição: $u = 1$, $u = 5$ e $u = 10$ novas cópias por indivíduo por geração, representando um crescimento esperado médio linear no número de cópias por indivíduo.

Comportamento esperado: Se a presença de TEs não causa impacto no *fitness*, não há atenuação na curva de invasão.

Os gráficos apresentados a seguir da dinâmica da subpopulação GM na figura 5.9 correspondem às densidades dessa subpopulação ao longo do tempo, para todas as réplicas. Os gráficos das médias correspondem à média tomada por geração em todas as réplicas. Os gráficos referentes aos dados de TEs na figura 5.10 correspondem às médias tomadas por geração para o número total de TEs na população, e o número de TEs ativos na população (cf. seção 4.5.1).

Em todos os casos em que não há custo de *fitness*, a família de TEs se expande livremente na população diploide, seguindo o potencial máximo de replicação determinado pelo modelo de transposição. Esse é o comportamento esperado para um mecanismo autoexplicativo não-limitado.

A taxa de transposição da família de TEs nesses cenários extremamente simplificados parece não afetar a velocidade de invasão, o que pode ser observado pela semelhança número de médio indivíduos GM ao final de 30 gerações em cada cenário da figura 5.9. Todas as três taxas de transposição $u = 1$, $u = 5$ e $u = 10$ atingiram patamares de penetração na população semelhantes, da ordem de 8.000 indivíduos (representando 80% da população).

Experimento A: histórias demográficas da subpopulação GM nas simulações com *template* com 1 TE, no caso controle sem custo de *fitness* ($s = 0$). Os gráficos da coluna da esquerda representam o tamanho da subpopulação GM nas 10 réplicas simuladas para cada cenário de simulação. Os gráficos da coluna da direita correspondem à média aritmética

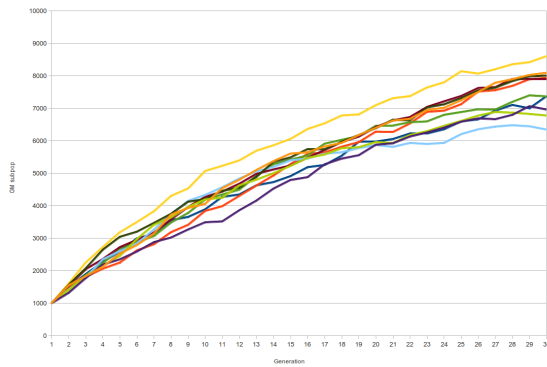
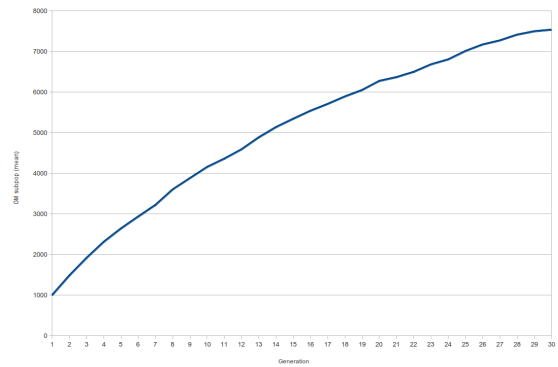
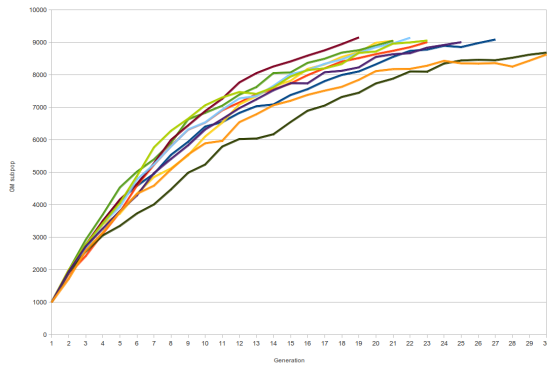
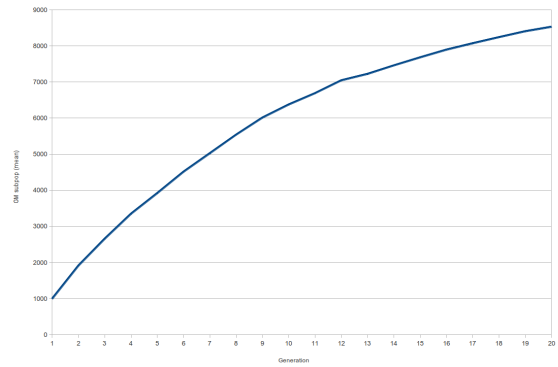
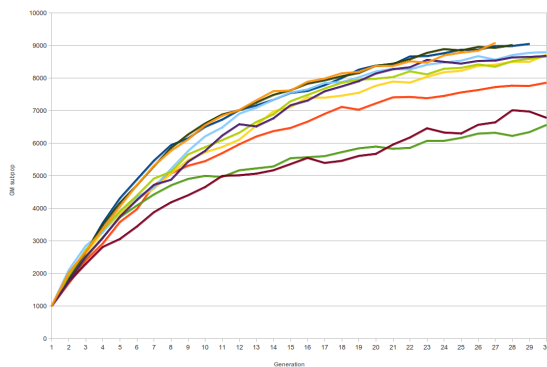
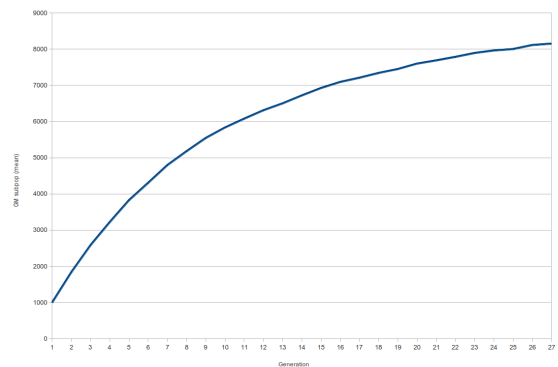
(a) $u = 1$ (b) $u = 1$ (média)(c) $u = 5$ (d) $u = 5$ (média)(e) $u = 10$ (f) $u = 10$ (média)

Figura 5.9: Experimento A: histórias demográficas da subpopulação GM nas simulações com *template* com 1 TE, no caso controle sem custo de *fitness* ($s = 0$). Os gráficos da coluna da esquerda representam o tamanho da subpopulação GM nas 10 réplicas simuladas para cada cenário de simulação. Os gráficos da coluna da direita correspondem à média aritmética dos dados mostrados explicitamente na coluna da direita.

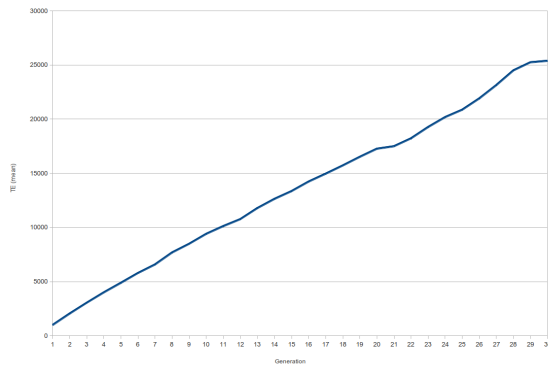
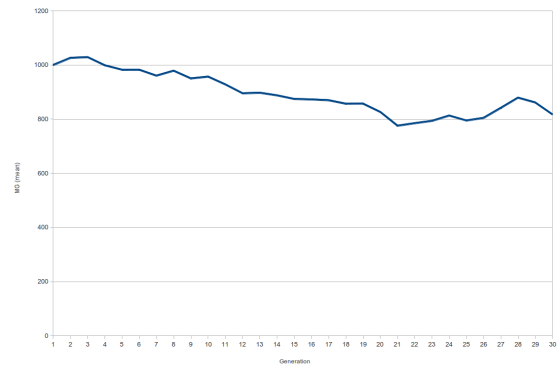
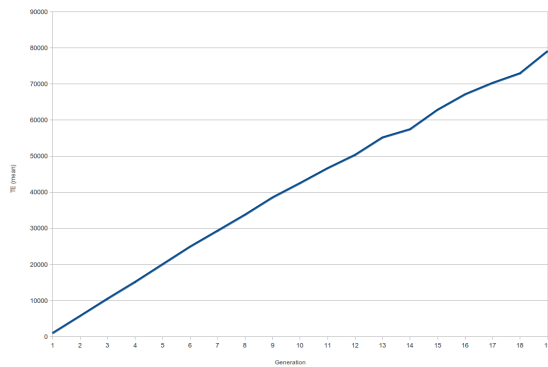
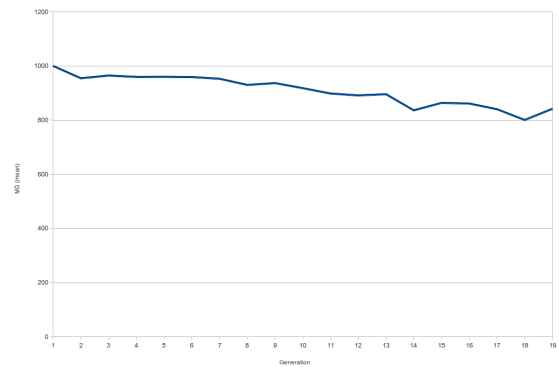
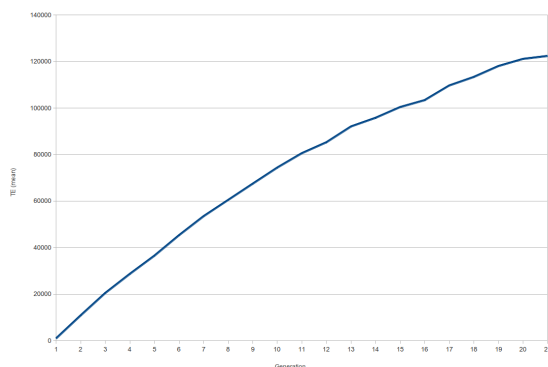
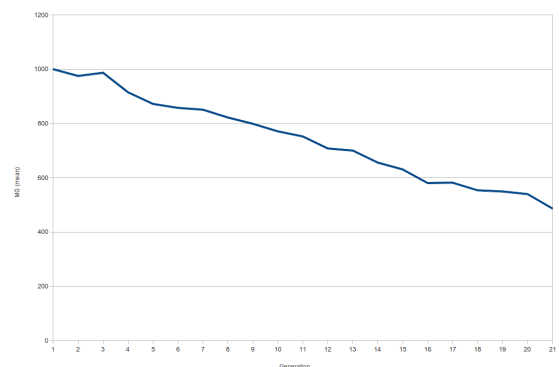
(a) $u = 1$ TEs totais (média)(b) $u = 1$ TEs ativos (média)(c) $u = 5$ TEs totais (média)(d) $u = 5$ TEs ativos (média)(e) $u = 10$ TEs totais (média)(f) $u = 10$ TEs ativos (média)

Figura 5.10: Experimento A, *template* com 1 TE, evolução dos TEs na população, Os gráficos das médias correspondem à média aritmética do total de TEs, e o número de TEs ativos, tomada por geração para todas as réplicas.

dos dados mostrados explicitamente na coluna da direita.

A diferença no comportamento quantitativo no entanto fica evidente nas quantidades média e total de TEs acumulados em cada cenário, sumarizados na tabela 5.1 e ilustrados na figura 5.10. O número total de TEs acumulados em cada cenário cresce consideravelmente com a taxa de transposição.

Tabela 5.1: Quantidades de TEs ($s = 0$, $template=1$). Médias correspondem à última geração comum a todas as réplicas. Valores máximos foram obtidos a partir da última geração de cada réplica.

u	TEs total	TEs ativos	máximo TEs	máximo ativos
$u = 1$	25386,7	817,1	35174	1280
$u = 5$	80867,3	806,5	107241	1257
$u = 10$	122359,0	486,5	212636	909

Uma vez que não há nenhuma limitação ou regulação em efeito, a expansão do número total de TEs ocorre segundo um crescimento linear por partes. Como se pode observar na figura 5.10, o número total de TEs na população cresce linearmente, atendendo a expectativa para um modelo de transposição linear.

Observe-se no entanto, que para esse crescimento ser linear, basta que existam TEs ativos em quantidade suficiente para gerar novas cópias, e que quanto maior a taxa de transposição, menor é a quantidade necessária de TEs ativos para que esse crescimento seja sustentado. De fato, o número de TEs ativos na população parece decrescer com o tempo, conforme pode ser observado nas porções respectivas da figura 5.10. Caso o número de TEs ativos crescesse linearmente, se expandindo por mais indivíduos na população, o número total de TEs seria retroalimentado por esse excedente, exibindo uma velocidade de crescimento polinomial de ordem 2 (crescimento quadrático).

Existem várias razões que podem justificar o decréscimo do número de TEs ativos (MG), tais como:

1. Por construção, seu número não cresce como consequência do modelo *Master Gene* só originar cópias inativas;
2. Perda do elemento ativo por sobreposições espúrias, ou na segregação de cromossomos.

Como novas transposições não geram TEs ativos, esses elementos nunca são mobilizados para outros *loci* do cromossomo. Dessa forma, TEs ativos podem ser mantidos na mesma posição, ou perdidos ao acaso na segregação de cromossomos.

Uma vez que cada novo elemento inativo é inserido em um *locus* aleatório do cromossomo, em cada gametogênese existe uma probabilidade de $\frac{1}{\#\text{loci}}$ de o novo elemento inativo sobrescrever o TE ativo, para cada novo TE criado.

Esse resultado está de acordo com o modelo apresentado na seção 6.2.4, que determina que o número de elementos ativos no equilíbrio nessas condições é zero, mesmo na ausência de custo de *fitness*, ou um mecanismo explícito de perda de TEs.

Por outro lado, o único mecanismo que pode gerar aumento do número de TEs ativos é a segregação na gametogênese, isto é, a contribuição que um indivíduo GM dá à sua prole. Porém para que seu número total na população aumentasse, essa contribuição teria necessariamente que ser maior que 50% da prole, e portanto mais eficiente que as regras básicas da genética mendeliana. Tratando apenas dos TEs ativos, em um modelo do tipo *Master Gene*, em cenários onde não há impacto perceptível no *fitness* dos indivíduos, isso não é viável.

Nesse contexto, cada TE ativo tem correspondência com um gene comum restrito a um único *loci* do cromossomo, e sua frequência na população pode ser estudada de acordo com os modelos clássicos de genética de populações. Como sua presença no genoma não confere vantagem ou desvantagem seletiva ao indivíduo que o carrega, sua expansão, fixação ou perda na população depende principalmente do tamanho efetivo da mesma. Se a população for muito grande, sua fixação se dará de acordo com as condições de Hardy-Weinberg. Se por outro lado a população for pequena, o único fator que pode causar um aumento do número de TEs ativos na população é a fixação aleatória por deriva genética. Assim, se um indivíduo possui apenas um MG, e nenhum novo elemento ativo é criado na gametogênese, cada filhote gerado possui apenas 50% de probabilidade de receber o elemento MG.

Cenários com populações pequenas tornam mais evidente esse comportamento de perda de TEs, tanto ativos como inativos (cf. figura 5.11). Adicionalmente aos cenários de interesse descritos acima, simulamos também populações de tamanhos 1000 e 100 para as três taxas de transposição ($u = 1$, $u = 5$ e $u = 10$) e em todos eles o pequeno tamanho implica em uma maior variância ao comportamento dinâmico observado, quando comparado com os cenários do experimento (cf. figura 5.9).

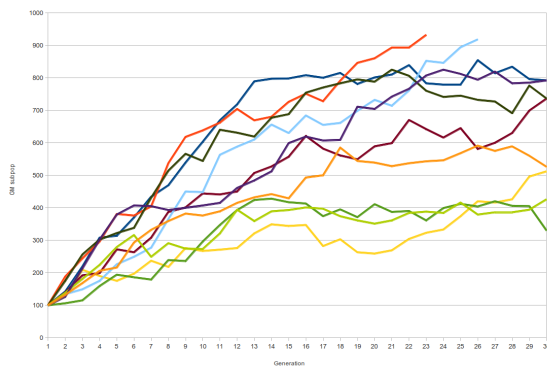
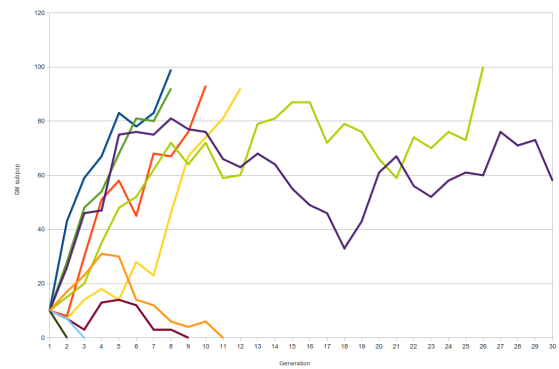
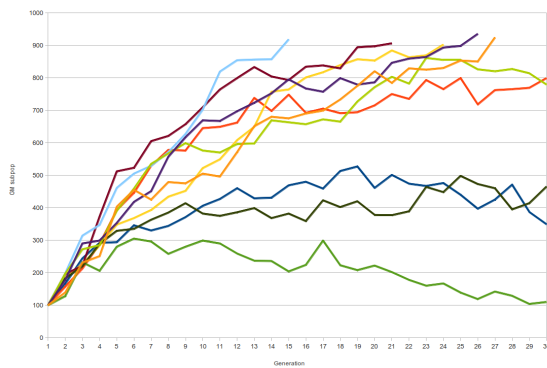
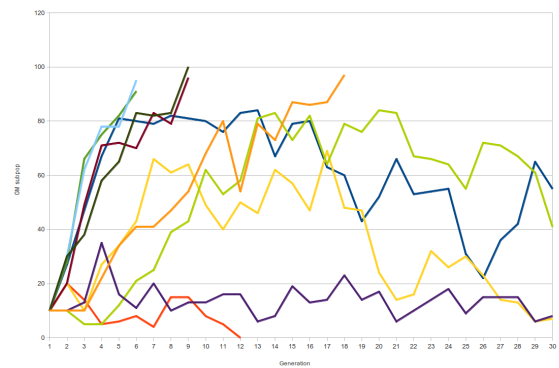
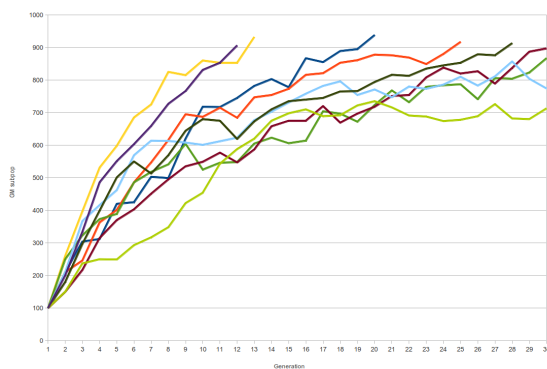
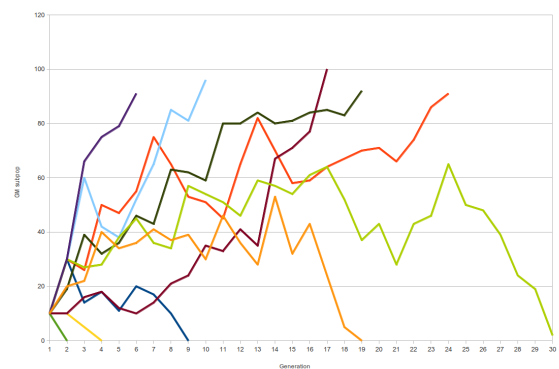
(a) $u = 1$, pop=1000(b) $u = 1$, pop=100(c) $u = 5$, pop=1000(d) $u = 5$, pop=100(e) $u = 10$, pop=1000(f) $u = 10$, pop=100

Figura 5.11: Experimento A, *template* com 1 TE, $s = 0$, histórias demográficas da subpopulação GM para populações pequenas (1000 e 100 indivíduos). Os gráficos das médias correspondem à média aritmética da subpopulação GM tomada por geração para todas as réplicas.

Nos cenários com população 1000 a possibilidade de perda de TEs parece não operar com intensidade suficiente para a perda dos TEs na população como um todo. O comportamento de todas as réplicas é coerente com os cenários estudados no experimento, descritos anteriormente, embora a variância dessas medidas seja maior, o que é evidente pelo maior espalhamento das réplicas em relação ao comportamento médio (curva média não mostrada nesta tese).

No entanto ao observarmos os cenários com população de tamanho 100, vê-se que a perda de TEs na população ocorre em pelo menos algumas réplicas de cada caso. O cenário com $u = 1$ parece não oferecer crescimento suficiente para que populações que comecem perdendo parte de sua proporção inicial nas primeiras gerações recuperem volume nas gerações subsequentes. Esse efeito de perda inicial irreversível parece ser mitigado ao aumentar progressivamente a taxa de transposição (cf. cenários para $u = 5$ e $u = 10$).

Um exemplo específico disso, que destoa dos cenários com população 1000, é a réplica representada pela curva verde, no cenário com população 100 e $u = 5$ na figura 5.11. Nas primeiras gerações a proporção GM da população se mantém estável em 10%, depois decresce para algo em torno de 5%, e a partir da quarta geração começa a crescer rapidamente por 10 gerações, atingindo seu ponto máximo em torno de 80% da população. Esse exemplo é particularmente ilustrativo também pelo fato que, ao contrário do que ocorreu em todos os cenários com populações grandes, o comportamento de cada replicada é difícil de ser previsto. Nessa réplica específica, o fato de a subpopulação GM atingir um patamar superior a 80% da população não lhe garante a subsistência ao longo do tempo, o que pode ser visto por sua posterior queda até a proporção de cerca 40% da população em torno da geração 30.

Esses experimentos portanto indicam que as primeiras gerações são cruciais na capacidade de invasão de TEs em populações pequenas, o que diminui a eficácia de TEs como *gene drive systems*.

Template com 20 TEs, sem custo de *fitness*

Como o fator determinante para a invasão é a capacidade da família de TEs se expandir mais rapidamente do que é removida do genoma [21, 45, 127], tentamos observar o comportamento de uma família que expande com várias taxas, mas a partir de um número inicial moderadamente grande. Simulamos três cenários com as mesmas três taxas de transposição anteriores ($u = 1$, $u = 5$ e $u = 10$ novas cópias por indivíduo por geração), mas desta vez partindo de um *template* inicial com 20 cópias (cf. 4.3.3).

Como estes cenários também se utilizam do modelo *Master Gene* de inativação de cópias, apenas um destes 20 elementos iniciais é ativo.

Comportamento esperado: Mesmo com uma carga inicial grande de TEs, se estes não causam impacto perceptível no *fitness*, não há atenuação na curva de invasão. O comportamento geral deve ser semelhante ao do cenário respectivo com *template* de 1 TE, a partir do momento que os indivíduos tenham acumulado em torno de 20 elementos.

Como se pode ver na figura 5.13, a quantidade de TEs acumulada ao longos das gerações foi bastante incrementada em relação as cenários do *template* anterior (figura 5.10).

Ainda assim o comportamento qualitativo da invasão seja essencialmente o mesmo, com ambos *templates* atingindo penetração em torno de 80% da população (cerca de 8000 indivíduos), para todas as taxas de transposição estudadas (figuras 5.9 e 5.12). Isto é, apesar da maior quantidade total de TEs acumulada na população, isso não se refletiu numa difusão mais eficiente, tanto em termos de velocidade da invasão, como da penetração em uma proporção maior da população.

Tabela 5.2: Quantidades de TEs ($s = 0$, *template*=20). Médias correspondem à última geração comum a todas as réplicas. Valores máximos foram obtidos a partir da última geração de cada réplica.

u	TEs total	TEs ativos	máximo TEs	máximo ativos
$u = 1$	47355,9	977,5	58837	1308
$u = 5$	96893,5	814,7	137876	1153
$u = 10$	144209,1	698,9	220639	1137

É importante também observar dois fenômenos que parecem emergir das simulações. O primeiro deles é que parece haver uma correlação entre o aumento da taxa de transposição, e a perda de elementos ativos (especialmente nos cenários $u = 10$ nas figuras 5.10 e 5.13).

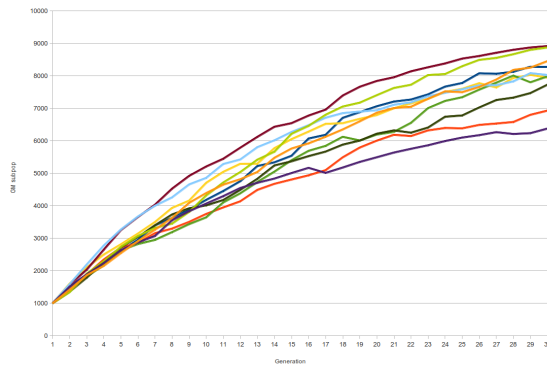
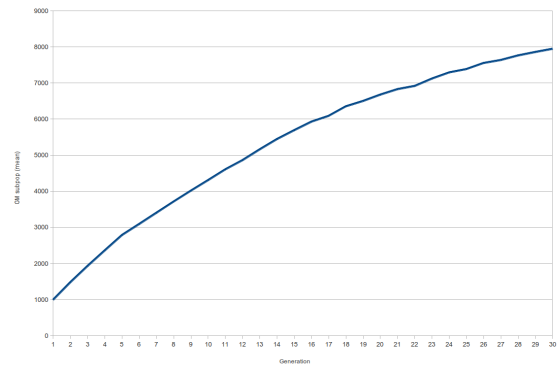
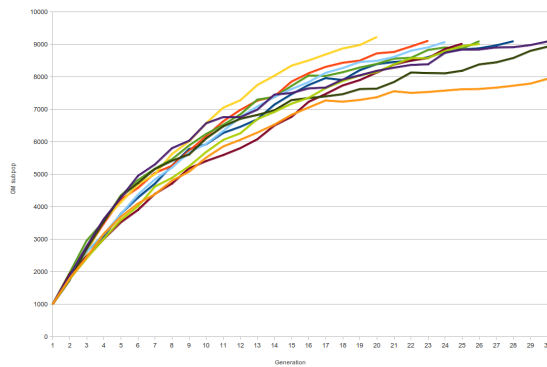
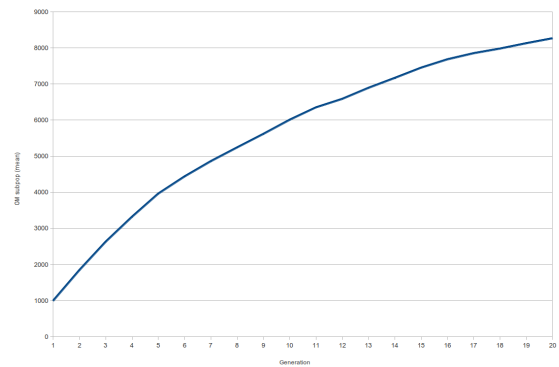
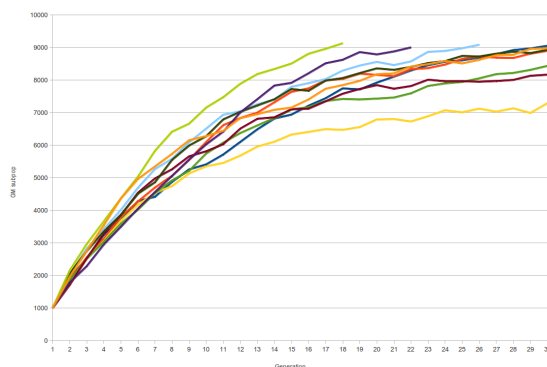
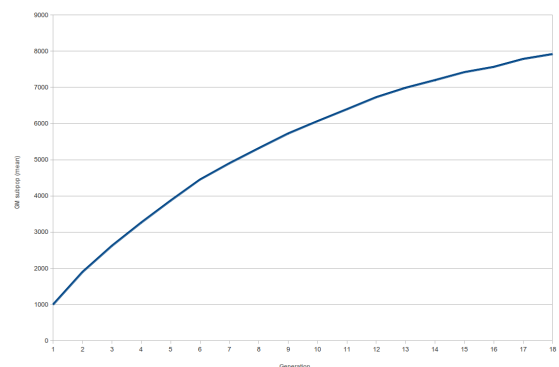
(a) $u = 1$ (b) $u = 1$ (média)(c) $u = 5$ (d) $u = 5$ (média)(e) $u = 10$ (f) $u = 10$ (média)

Figura 5.12: Experimento A, *template* com 20 TEs, $s = 0$, histórias demográficas da subpopulação GM. Os gráficos das médias correspondem à média aritmética da subpopulação GM tomada por geração para todas as réplicas.

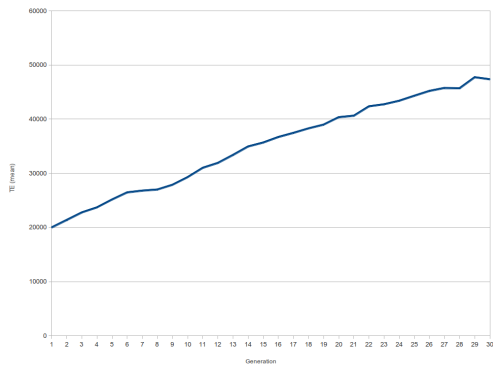
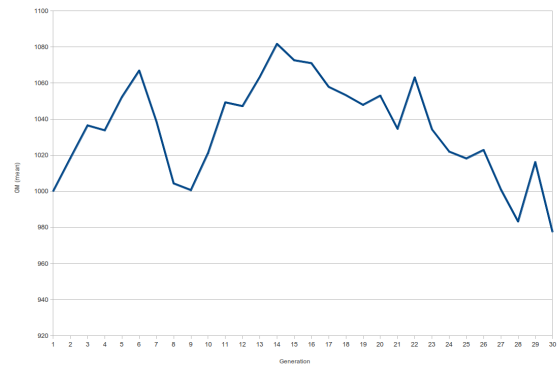
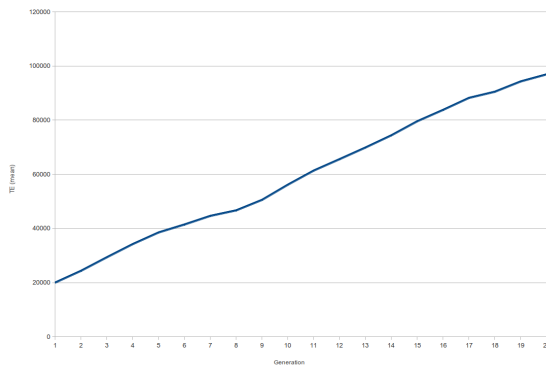
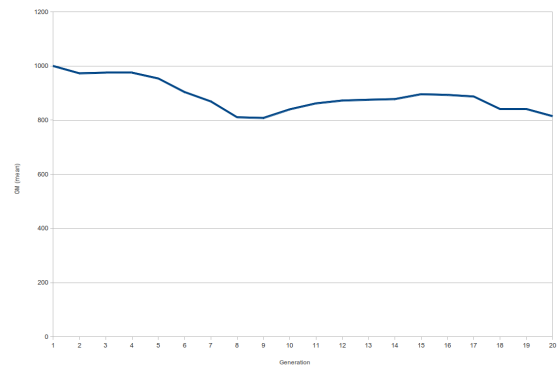
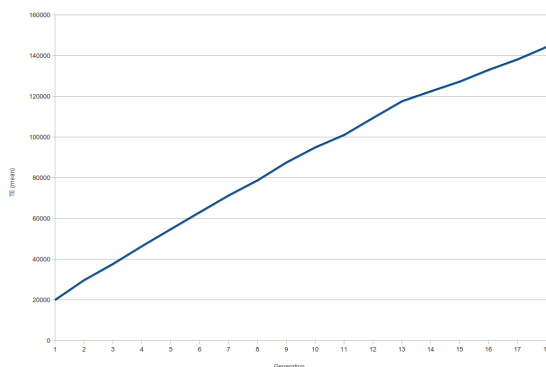
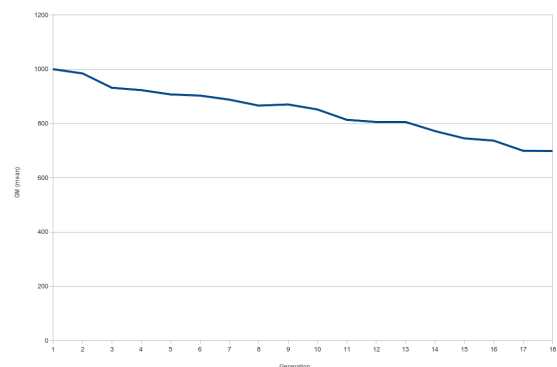
(a) $u = 1$ TEs totais (média)(b) $u = 1$ TEs ativos (média)(c) $u = 5$ TEs totais (média)(d) $u = 5$ TEs ativos (média)(e) $u = 10$ TEs totais (média)(f) $u = 10$ TEs ativos (média)

Figura 5.13: Experimento A, *template* com 20 TE, evolução dos TEs na população, Os gráficos das médias correspondem à média aritmética do total de TEs, e o número de TEs ativos, tomada por geração para todas as réplicas.

Nas próximas seções e cenários, quando considerarmos o impacto na fecundidade, esse fenômeno poderá ser observado de forma mais expressiva. O segundo fenômeno é que a dinâmica dos cenários com taxa $u = 5$ parecem ter um comportamento mais suave que os de $u = 1$.

O fato de a dinâmica deste *template* ser semelhante aos cenários com o *template* menor não é de forma alguma surpreendente, porém é um indicativo de robustez na dinâmica dessa classe de cenários.

A seguir, estudaremos cenários mais realistas, com diferentes cargas de impacto na fecundidade do hospedeiro.

5.2.2 Com custo de *fitness*

Template com 1 TE, custo pequeno de *fitness*

Os cenários a seguir finalmente enquadram as hipóteses de maior interesse biológico, pois mais se aproximam do fenômeno natural a que se pretende modelar. Conforme definido na seção 4.1.4, o impacto total de *fitness* aplicado a cada indivíduo é modelado como um custo em sua fecundidade. O impacto individual de cada TE deve ser somado e arredondado para que esse custo total seja contabilizado. Esses parâmetros devem ser escolhidos para que o número de equilíbrio de TEs tenha um valor total de impacto menor que a capacidade reprodutiva da espécie, caso contrário a invasão não se converterá em fixação do TE na população, seja por perda da subpopulação GM, ou por extinção da população como um todo.

Comportamento esperado: Com um impacto de *fitness* pequeno em comparação à taxa de transposição, a invasão deve ocorrer a despeito do impacto negativo no hospedeiro.

Nos cenários ilustrados na figura 5.14, pode-se observar o crescimento da subpopulação GM, substituindo a população silvestre, de modo semelhante ao caso controle em que não há custo de *fitness*, mas o crescimento é ligeiramente atenuado pela diminuição de fecundidade dos indivíduos GM, especialmente nas últimas gerações. Quanto maior a quantidade total de TEs na população (figura 5.15), maior o impacto na fecundidade média da população.

Como podemos observar na figura 5.15, a quantidade total de TEs parece crescer conforme o controle, atingindo patamares máximos próximos dos obtidos no controle. A única exceção ocorreu no cenário com taxa de transposição $u = 10$. No entanto, nesse cenário na perda de TEs ativos foi menor que no controle, com número final em torno de 700 TEs, em comparação aos 500 do controle.

Tabela 5.3: Quantidades de TEs ($s = 0,01$, *template*=1). Médias correspondem à última geração comum a todas as réplicas. Valores máximos foram obtidos a partir da última geração de cada réplica.

u	TEs total	TEs ativos	máximo TEs	máximo ativos
$u = 1$	29018,4	971,3	37628	1583
$u = 5$	72918,8	906,3	131445	1245
$u = 10$	158418,5	668,9	200040	989

O comportamento oscilatório das curvas de TEs ativos na figura 5.15 para os cenários

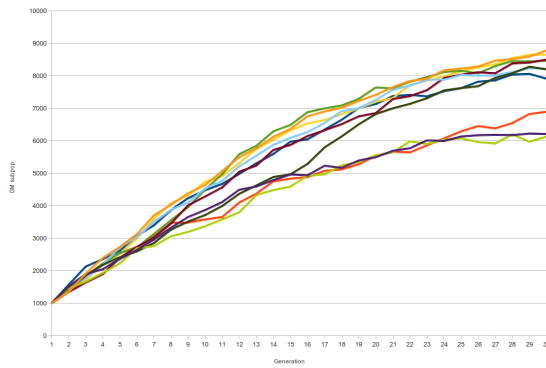
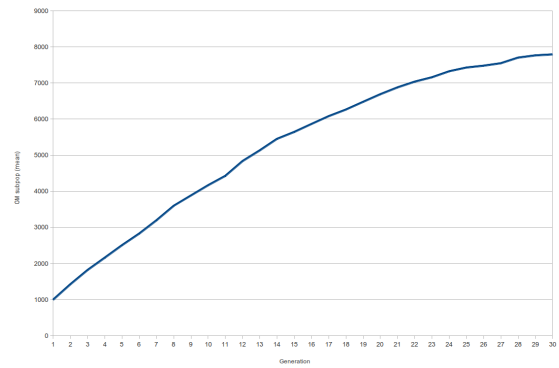
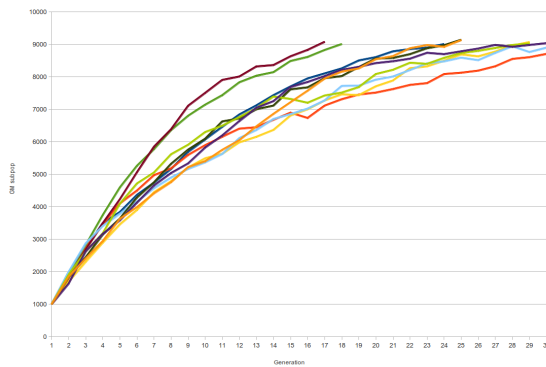
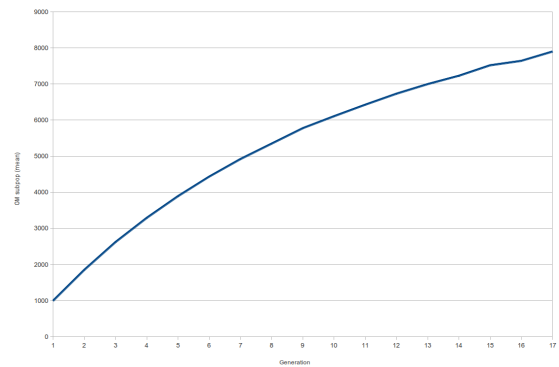
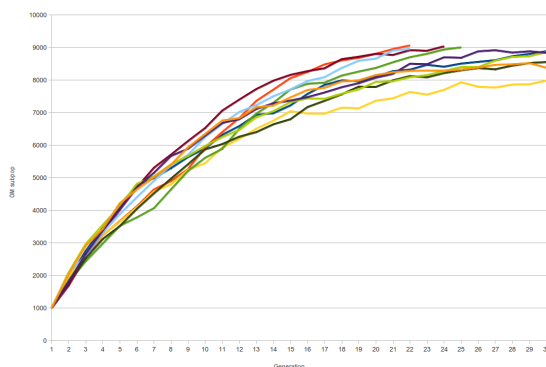
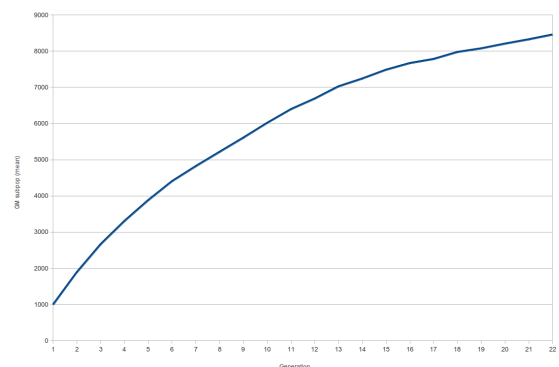
(a) $u = 1$ (b) $u = 1$ (média)(c) $u = 5$ (d) $u = 5$ (média)(e) $u = 10$ (f) $u = 10$ (média)

Figura 5.14: Experimento A, *template* com 1 TE, $s = 0,01$, histórias demográficas da subpopulação GM. Os gráficos das médias correspondem à média aritmética da subpopulação GM tomada por geração para todas as réplicas.

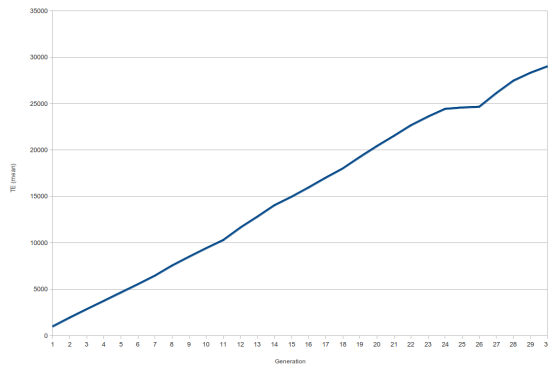
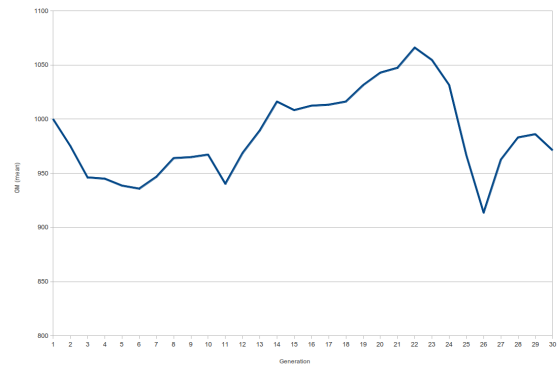
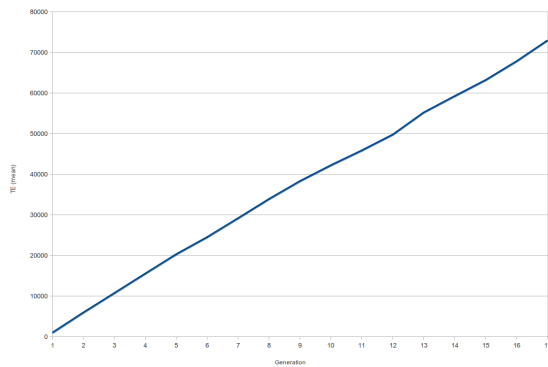
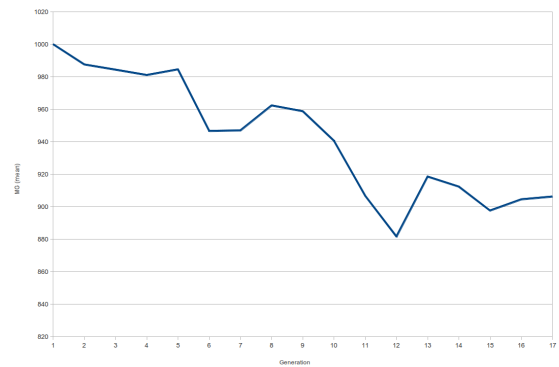
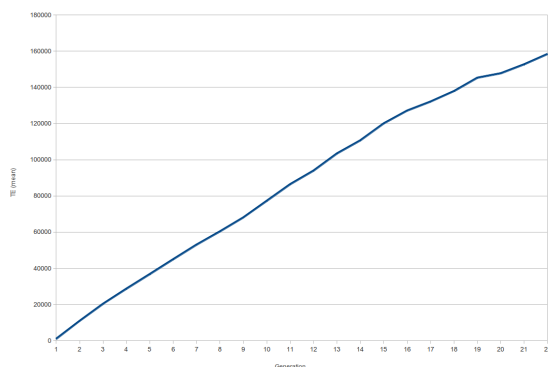
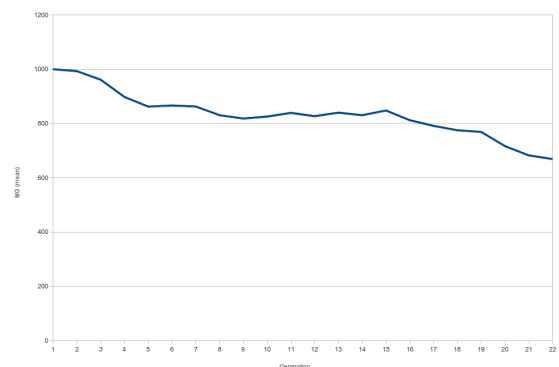
(a) $u = 1$ TEs totais (média)(b) $u = 1$ TEs ativos (média)(c) $u = 5$ TEs totais (média)(d) $u = 5$ TEs ativos (média)(e) $u = 10$ TEs totais (média)(f) $u = 10$ TEs ativos (média)

Figura 5.15: Experimento A, *template* com 1 TE, $s = 0,01$, evolução dos TEs na população, Os gráficos das médias correspondem à média aritmética do total de TEs, e o número de TEs ativos, tomada por geração para todas as réplicas.

$u = 1$ e $u = 5$ pode ser explicada pela grande variância dessas distribuições. Esses cenários foram bastante diversos, e uma composição das curvas de todas as réplicas mostra um comportamento de espalhamento em torno da média, sem tendência óbvia (dados não mostrados nesta tese). Ainda assim, como se pode ver nos respectivos gráficos para o total de TEs nesses cenários, observa-se que mesmo nas réplicas em que houve queda expressiva do número de TEs ativos, a contribuição dos ativos remanescentes foi suficiente para alimentar o crescimento linear do total.

Template com 1 TE, custo moderado de *fitness*

Além da velocidade da invasão, em termos do número de gerações, é preciso considerar que diferentes famílias de TEs interagem de forma diferente com o genoma do hospedeiro, e dessa forma impõe impactos diferentes em sua fecundidade.

Para avaliar as diferenças qualitativas que emergem dessa variante, dois casos diferentes em termos de custo de *fitness* foram simulados. Na figura 5.14 foi representada uma família de TEs que causa um impacto pequeno (custo individual de *fitness* por TE por indivíduo igual a 0,01 filhotes). A figura 5.16 ilustra o caso em que esse impacto é aumentado para 0,05 filhotes por TE por indivíduo.

Comportamento esperado: Com um impacto de *fitness* moderado em comparação à taxa de transposição, a invasão ocorre de maneira mais lenta.

Os cenários desse *template* exibem uma diferença qualitativa dramática em relação aos do caso anterior (figura 5.16). Observa-se que a subpopulação GM tem seu crescimento desacelerado após um certo número de gerações em todas as taxas. Nos cenários com $u = 5$ a subpopulação cresce até um certo patamar e depois prossegue relativamente estável. No cenário com $u = 10$, ocorre uma estabilização, mas com leve declínio.

Esse comportamento de estabilidade da subpopulação GM fica evidente ao observar as curvas relativas ao total de TEs, na figura 5.17. Mesmo no cenário com taxa $u = 1$ há um leve declínio no total de TEs na população nas últimas gerações.

Esse amortecimento e posterior estabilização é explicado pela queda acentuada do número de TEs ativos no cenário $u = 1$, e na perda de todos os ativos nos cenários $u = 5$ e $u = 10$ (5.17).

Ambos os cenários em que houve perda dos TEs ativos parecem se aproximar de um valor próximo de 20.000 TEs da população, após a inativação completa da família de TEs. No cenário com $u = 10$, como o ganho de TEs foi muito abrupto nas primeiras gerações, o decréscimo parece mais acentuado que no caso $u = 5$, embora ambos pareçam se aproximar do mesmo valor. Essa quantidade parece ser o valor próximo de equilíbrio da população na presença desse custo de *fitness*.

De acordo com os modelos matemáticos mais simples, que assumem apenas ganho e perda lineares, em uma população constante (e.g. o modelo 6.3), esse é o comportamento esperado a longo prazo. Após uma fase inicial da invasão caracterizada por uma alta taxa de transposição, o número de TEs entra em equilíbrio. Isso é explicitamente justificado na seção 6.2.4.

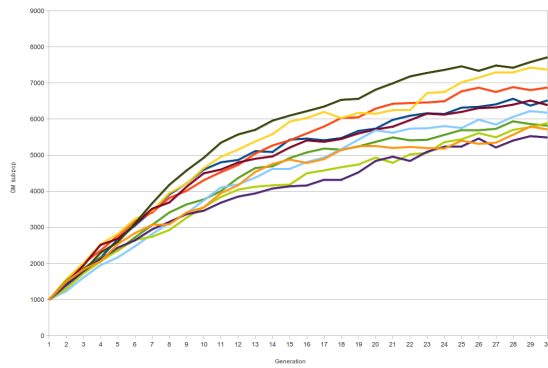
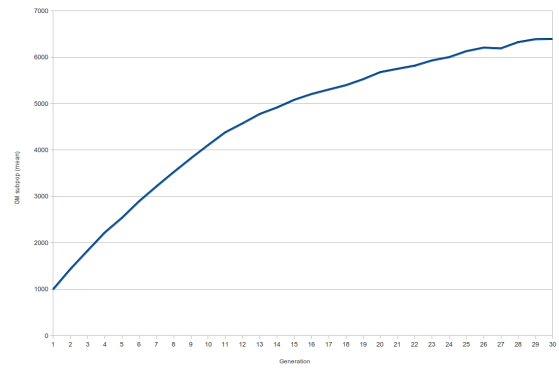
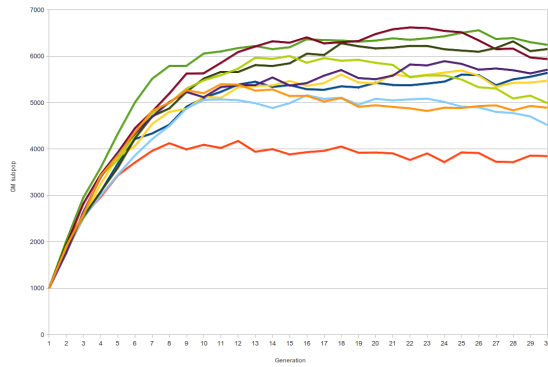
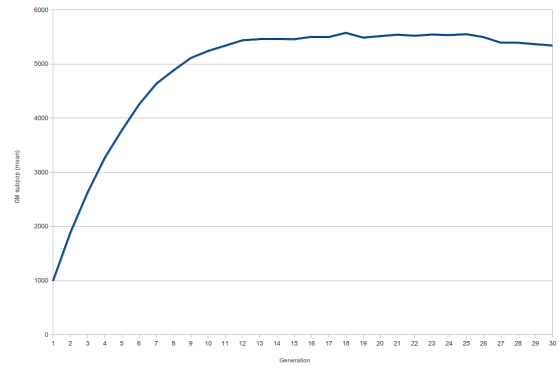
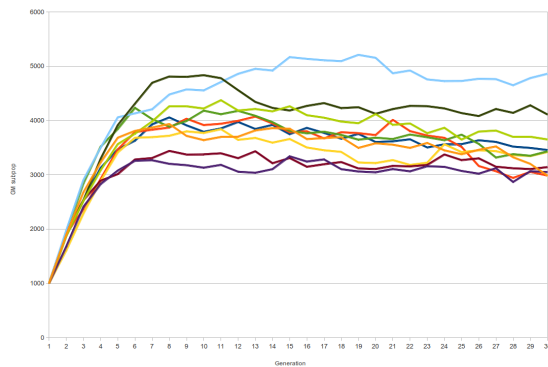
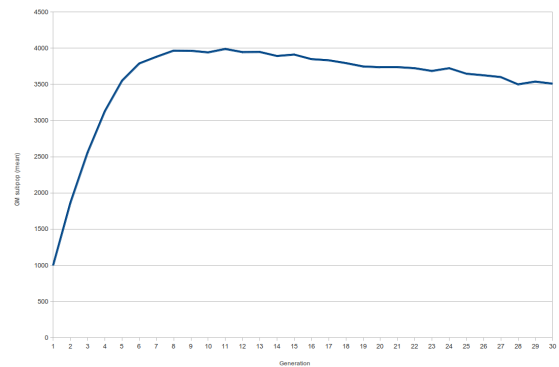
(a) $u = 1$ (b) $u = 1$ (média)(c) $u = 5$ (d) $u = 5$ (média)(e) $u = 10$ (f) $u = 10$ (média)

Figura 5.16: Experimento A, *template* com 1 TE, $s = 0,05$, histórias demográficas da subpopulação GM. Os gráficos das médias correspondem à média aritmética da subpopulação GM tomada por geração para todas as réplicas.

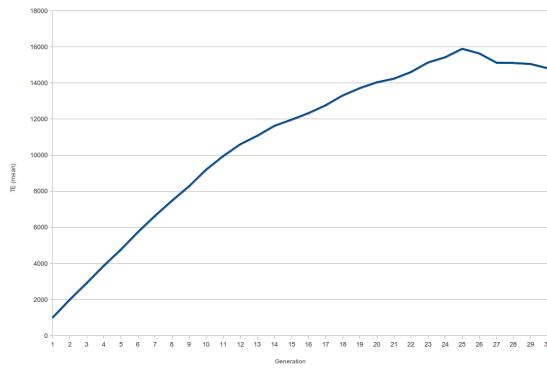
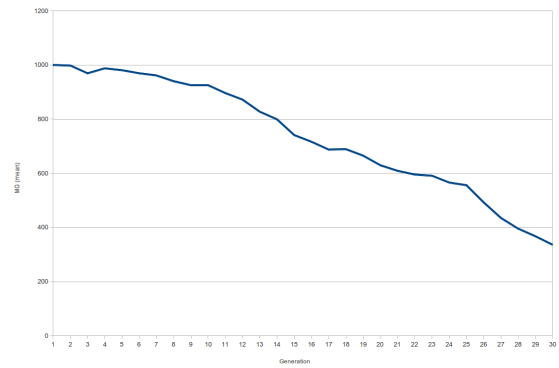
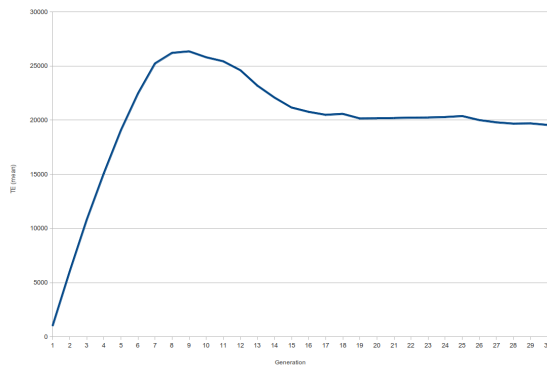
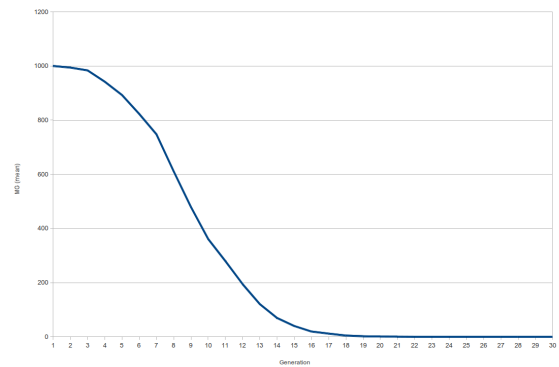
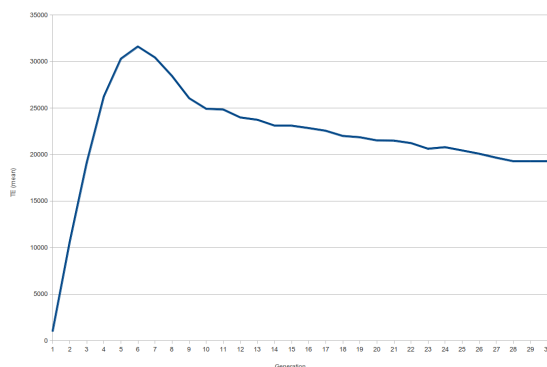
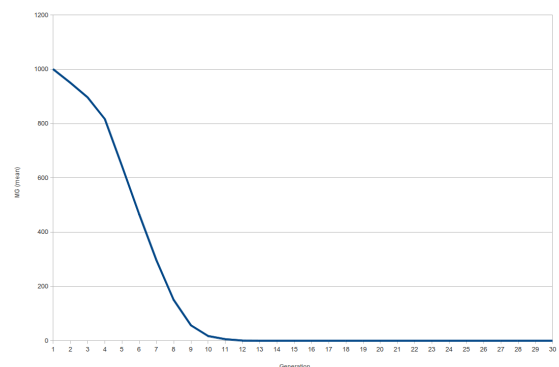
(a) $u = 1$ TEs totais (média)(b) $u = 1$ TEs ativos (média)(c) $u = 5$ TEs totais (média)(d) $u = 5$ TEs ativos (média)(e) $u = 10$ TEs totais (média)(f) $u = 10$ TEs ativos (média)

Figura 5.17: Experimento A, *template* com 1 TE, $s = 0,05$, evolução dos TEs na população, Os gráficos das médias correspondem à média aritmética do total de TEs, e o número de TEs ativos, tomada por geração para todas as réplicas.

Tabela 5.4: Quantidades de TEs ($s = 0,05$, $template=1$). Médias correspondem à última geração comum a todas as réplicas. Valores máximos foram obtidos a partir da última geração de cada réplica.

u	TEs total	TEs ativos	máximo TEs	máximo ativos
$u = 1$	14819,6	336,2	20699	569
$u = 5$	19556,9	0	26194	0
$u = 10$	19268,7	0	26920	0

Se comportamento esperado para esses cenários era de uma invasão mais lenta porém com maior abrangência na população, em última instância, não foi o que ocorreu. Embora a quantidade de indivíduos GM na população tenha subido até uma proporção expressiva da população (65% para $u = 1$, 55% para $u = 5$ e 40% para $u = 10$), e a família de TEs pareça se fixar, o objetivo de substituição de população claramente não é satisfeito.

Esse resultado indica a importância de se balancear a taxa de transposição com o custo de *fitness* implicado na população. Caso a fecundidade seja severamente reduzida, a população pode ser extinta, em casos extremos.

Simulamos também cenários com populações pequenas sujeitas a custo de *fitness* $s = 0,05$ (cf. figura 5.18). Assim como nos cenários com populações pequenas no controle ($s = 0$), nesses cenários a deriva genética aumenta sensivelmente a variância do comportamento das subpopulações GM, e quanto menor a população (cf. nos cenários com $N = 100$), mais errático é o comportamento do sistema. Invasões ou perdas da família de TEs parecem ocorrer sem uma tendência previsível. Assim como nos cenários com $N = 10.000$, o aumento na taxa de transposição implica numa aceleração do processo de perda do elemento MG, de modo que a família de TEs perde a capacidade de manter seu crescimento.

A seguir, veremos um caso um pouco mais exacerbado, mas ainda não suficiente para causar a extinção da população.

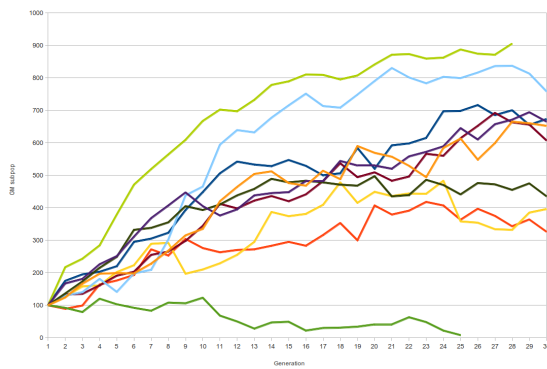
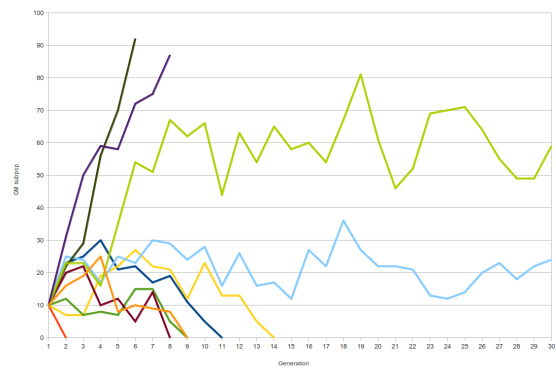
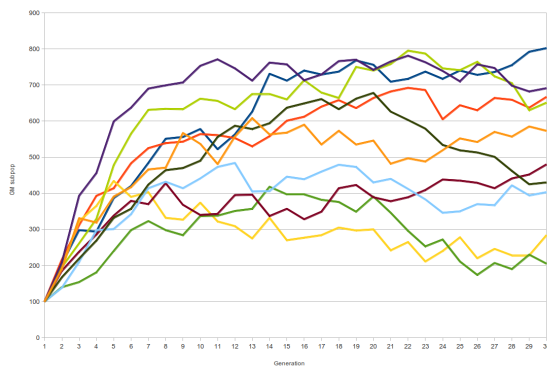
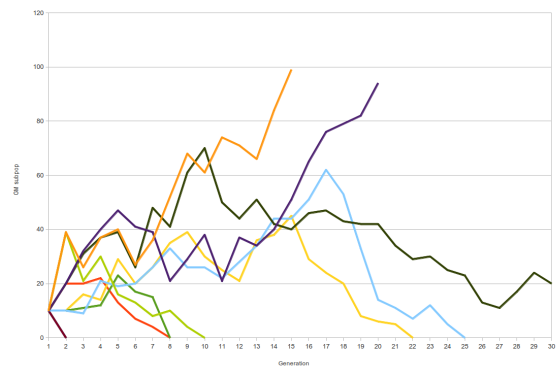
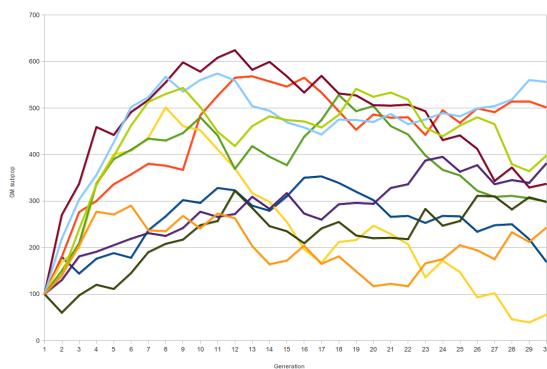
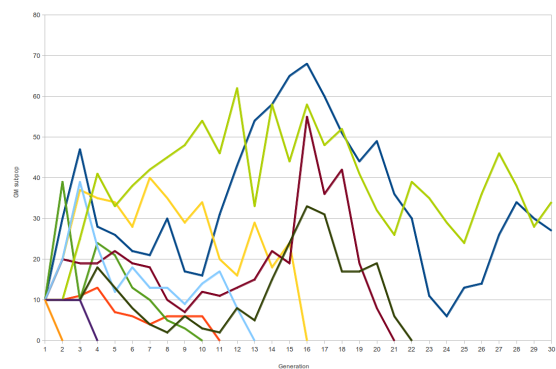
(a) $u = 1$, pop=1000(b) $u = 1$, pop=100(c) $u = 5$, pop=1000(d) $u = 5$, pop=100(e) $u = 10$, pop=1000(f) $u = 10$, pop=100

Figura 5.18: Experimento A, *template* com 1 TE, $s = 0,05$, histórias demográficas da subpopulação GM para populações pequenas (1000 e 100 indivíduos). Os gráficos das médias correspondem à média aritmética da subpopulação GM tomada por geração para todas as réplicas.

Template com 20 TEs, custo moderado de *fitness*

Além dos cenários anteriores, consideramos agora o caso em que um mosquito GM já seria introduzido na natureza com uma carga grande TEs, a fim de acelerar a chance inicial de dispersão dos transgenes, e possibilitar a invasão. É claro que na prática essa situação limítrofe gera a expectativa de que a primeira geração de descendentes seja extremamente ineficiente ou mesmo infértil numa estratégia de substituição de população. Nossas simulações mostram que não só a geração F_1 (a primeira geração descendente da população inicial F_0) é fortemente impactada, como toda a população silvestre sofre o impacto do custo inerente à introdução de muitos elementos de uma vez.

Comportamento esperado: Com um custo inicial elevado, a subpopulação GM invasora tem uma desvantagem considerável na competição com a população silvestre, oferecendo um cenário improvável para a substituição da população, e portanto o uso como *drive* genético.

Primeiramente consideremos uma observação importante sobre o impacto total dos TEs nos indivíduos GM da população F_0 . Com 20 TEs impactantes, a um custo de $s = 0,05$ cada, o custo total na fecundidade é de $S = 1$ filhote por indivíduo. A população de mosquitos que estamos simulando tem fecundidade máxima de 10 filhotes por indivíduo, impondo então uma perda inicial de 10% de fecundidade. Ao longo das gerações esse impacto por indivíduo tende a aumentar, especialmente com altas taxas de transposição.

Dado que a população se reproduz sem viés por *random mating*, o número esperado de indivíduos GM da população F_0 que contribuirá para a próxima geração é de 10%, cada um gerando 9 descendentes ao invés de 10. Como não estamos considerando recombinação *crossover*, cada um desses descendentes tem 50% de receber a carga inicial de TEs definida no *template*, desconsiderando os novos TEs a ser gerados¹. Como a proporção de indivíduos virgens é muito maior que a de GMs, a probabilidade de formação de casais com ambos os indivíduos GM a partir da população F_0 é relativamente pequena.

Uma segunda observação importante, é a diferença nos parâmetros dos cenários desse caso. Para essas simulações, foi usada uma estrutura diferente nos cromossomos. Os 500 *loci* possíveis para inserção foram reparticionados de modo que 50% destes sejam severos, comparado aos 10% dos outros casos. Isto é, a probabilidade de que cada novo TE cause impacto é grandemente aumentada aqui, de modo que os efeitos mais severos

¹Observe que todos os 20 TEs do *template* estão localizados em um único cromossomo

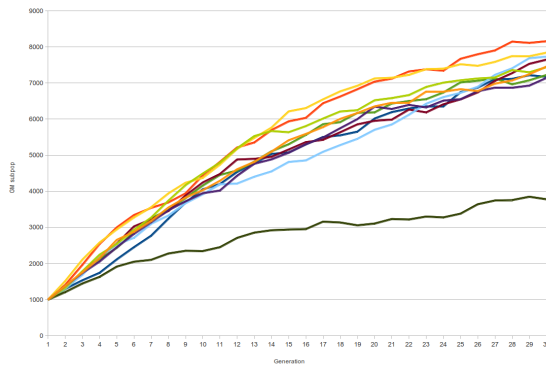
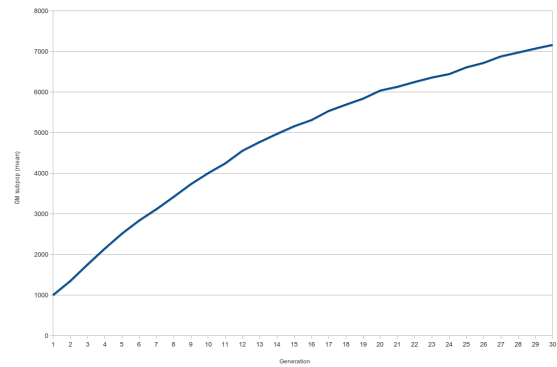
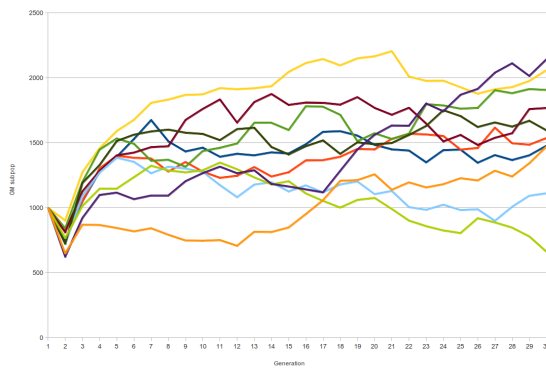
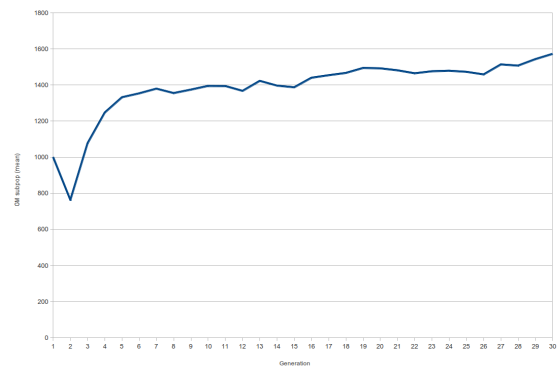
(a) $u = 1$ (b) $u = 1$ (média)(c) $u = 5$ (d) $u = 5$ (média)

Figura 5.19: Experimento A, *template* com 20 TEs, $s = 0,05$, histórias demográficas da subpopulação GM. Os gráficos das médias correspondem à média aritmética da subpopulação GM tomada por geração para todas as réplicas.

na fecundidade podem ser observados com menos gerações.

A figura 5.19 evidencia considerável impacto na primeira geração de descendentes da subpopulação GM introduzida no ambiente silvestre. De fato, o impacto repercute na população inteira, pois nessa primeira geração e nas seguintes não são gerados indivíduos suficientes para manter o patamar constante de 10.000 indivíduos (dados não mostrados nesta tese) previsto pelo modelo ecológico. O equilíbrio da população em patamar constante de 10.000 indivíduos só volta a ocorrer após algumas poucas gerações, quando a família de TEs começa a se homogenizar por um número maior de indivíduos (em torno da terceira geração, dados não mostrados).

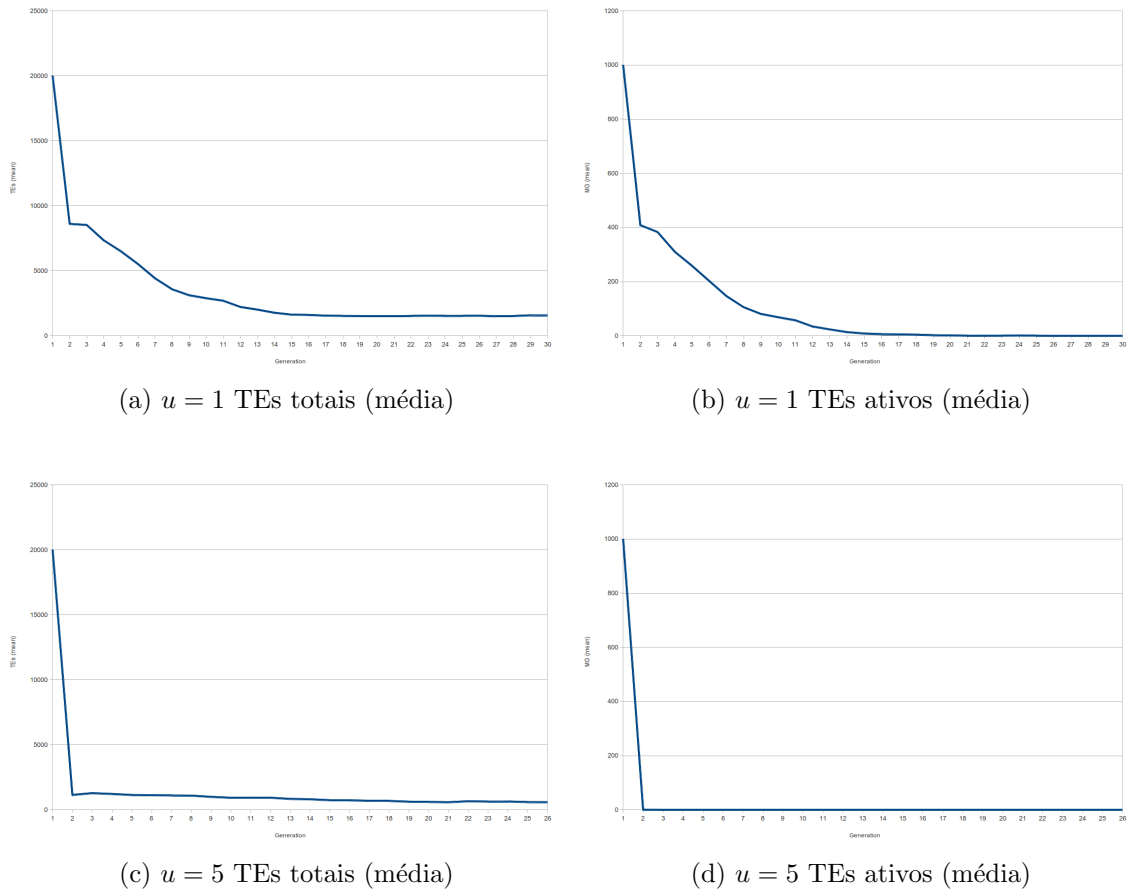


Figura 5.20: Experimento A, *template* com 20 TEs, $s = 0,05$, evolução dos TEs na população, Os gráficos das médias correspondem à média aritmética do total de TEs, e o número de TEs ativos, tomada por geração para todas as réplicas.

Observamos na figura 5.20 que a perda de TEs ativos é abrupta no cenário com $u = 1$ e quase imediata no cenário $u = 5$ (fato também evidenciado na tabela 5.5). A perda prematura de elementos ativos faz com que o número total de TEs também despenque logo nas primeiras gerações, e permaneça baixo até o final da simulação (comparar com tabelas 5.3 e 5.4).

Apesar dessa perda acentuada de TEs na população já nas primeiras gerações, um fenômeno interessante emerge nesse cenário. Mesmo após o desaparecimento quase imediato dos TEs ativos, e conseqüente falta de suporte para criação e sustentação de um elevado número de TEs como um todo na população, observa-se na figura 5.19 que a

Tabela 5.5: Quantidades de TEs ($s = 0,05$, $template=20$). Médias correspondem à última geração comum a todas as réplicas. Valores máximos foram obtidos a partir da última geração de cada réplica.

u	TEs total	TEs ativos	máximo TEs	máximo ativos
$u = 1$	1570,1	0	2288	0
$u = 5$	587,1	0	1094	0

população GM cresce, mesmo que lentamente. Observa-se um crescimento lento e menor que linear no cenário $u = 1$, e um crescimento mais rápido em direção à estabilidade no cenário $u = 5$. O número total de TEs relativamente baixo em relação aos outros casos é contraposto a uma maior dispersão desses TEs na população, a despeito de custo de *fitness*.

Uma possível explicação para esse crescimento é que os TEs remanescentes sejam em sua maioria neutros. Como as populações de todas as réplicas foram descartadas ao final das simulações, não é possível examinar a distribuição dos TEs existentes nas últimas gerações quanto ao custo de *fitness* deles. No entanto, para testar essa hipótese examinamos as 10 amostras obtidas em cada réplica e a proporção de TEs neutros encontrada foi:

$$u = 1 \implies \frac{5}{10} = 50\% \text{ TEs neutros}$$

$$u = 5 \implies \frac{12}{20} = 60\% \text{ TEs neutros}$$

Como uma proporção expressiva de TEs impactantes se faz presente nas amostras, o crescimento não ocorre devido à perda de TEs classificados como severos e concentração de TEs classificados como neutros. Por outro lado, essa classificação só tem efeito prático quando há quantidade suficiente para reduzir a fecundidade em uma unidade (no caso, no mínimo 20 TEs, como no *template*), pois o custo de *fitness* é arredondado para baixo. A quantidade de elementos na população é muito pequena quando comparada aos outros cenários, e isso se reflete nas amostras. Muitos indivíduos amostrados tem um único TE, e o número máximo de TEs em uma amostra foi de 3 elementos.

Nessa situação, o número de TEs por indivíduo é tão baixo que nenhum dos indivíduos sofreu impacto em sua fecundidade. Esses cenários portanto não caracterizam uma fixação devido a um mecanismo egoísta, embora para a expansão inicial o mecanismo egoísta

tenha sido fundamental. Ainda assim, parece haver a fixação de uma pequena quantidade de TEs por indivíduo em uma certa proporção considerável da população, satisfazendo o critério necessário para *gene drive system*.

5.3 Experimento B

5.3.1 Sem custo de *fitness*

Árvores geradas para vários modelos ecológicos

Os resultados do experimento B indicam que a dinâmica da população de hospedeiros tem um impacto mensurável na dinâmica de TEs. Isso se reflete em uma assinatura observável em topologias reconstruídas a partir de sequências de TEs.

Todas as árvores e estimativas demográficas a seguir foram feitas de acordo com o protocolo 4.6.2, a partir das simulações definidas nos protocolos 4.6, 4.7, 4.8 e 4.9 para expansões de famílias de TEs isentas de impacto de *fitness*.

Como parâmetro de comparação para modelos ecológicos variáveis, consideramos a árvore reconstruída a partir de uma população constante (figura 5.21).

Analizamos também dois casos em que a população começa pequena e depois se expande. A árvore na figura 5.22 foi reconstruída de uma população com crescimento exponencial. A população F_0 era consideravelmente pequena, com 1.000 indivíduos. A taxa de natalidade $r = 0,1$ implica num crescimento relativamente lento, mas ao longo de várias gerações, a população ficou consideravelmente grande. A árvore da figura 5.23 foi feita a partir de uma população logística, com taxa de crescimento $r = 1$, e estabilizando em sua capacidade suporte $K = 10.000$.

Uma observação que pode ser feita sobre o comportamento qualitativo desses cenários é que para populações iniciais pequenas, parece haver maior proporção de TEs originados de um mesmo indivíduo, e isso se reflete na topologia da árvore como maior quantidade de clados múltiplos, relativos aos vários descendentes de um mesmo MG em um indivíduo, ou indivíduos com perfil cromossômico muito parecido. Isso pode ser visto nas primeiras gerações dos cenários de populações crescentes, exponencial (figura 5.22) e logística (figura 5.23).

Uma segunda observação é a de que, como o *template* usado nos cenários desse experimento contem um único TE ativo, todas as árvores tem forma moderadamente pectinada, exceto quando a história da população do hospedeiro influencia fortemente na capacidade de expansão da demografia dos TEs. Esse tipo de complicação, no entanto, não ocorreu de forma generalizada, ficando restrita a apenas uma pequena porção de algumas árvores.

Populações muito grandes, parecem não originar a estrutura pectinada prevista. Isso

pode ser observado nas ramificações primeiras gerações da exponencial decrescente, que iniciou com uma geração F_0 de 20.000 indivíduos (ramos mais longos da árvore, mais próximos da raiz, na figura 5.24).

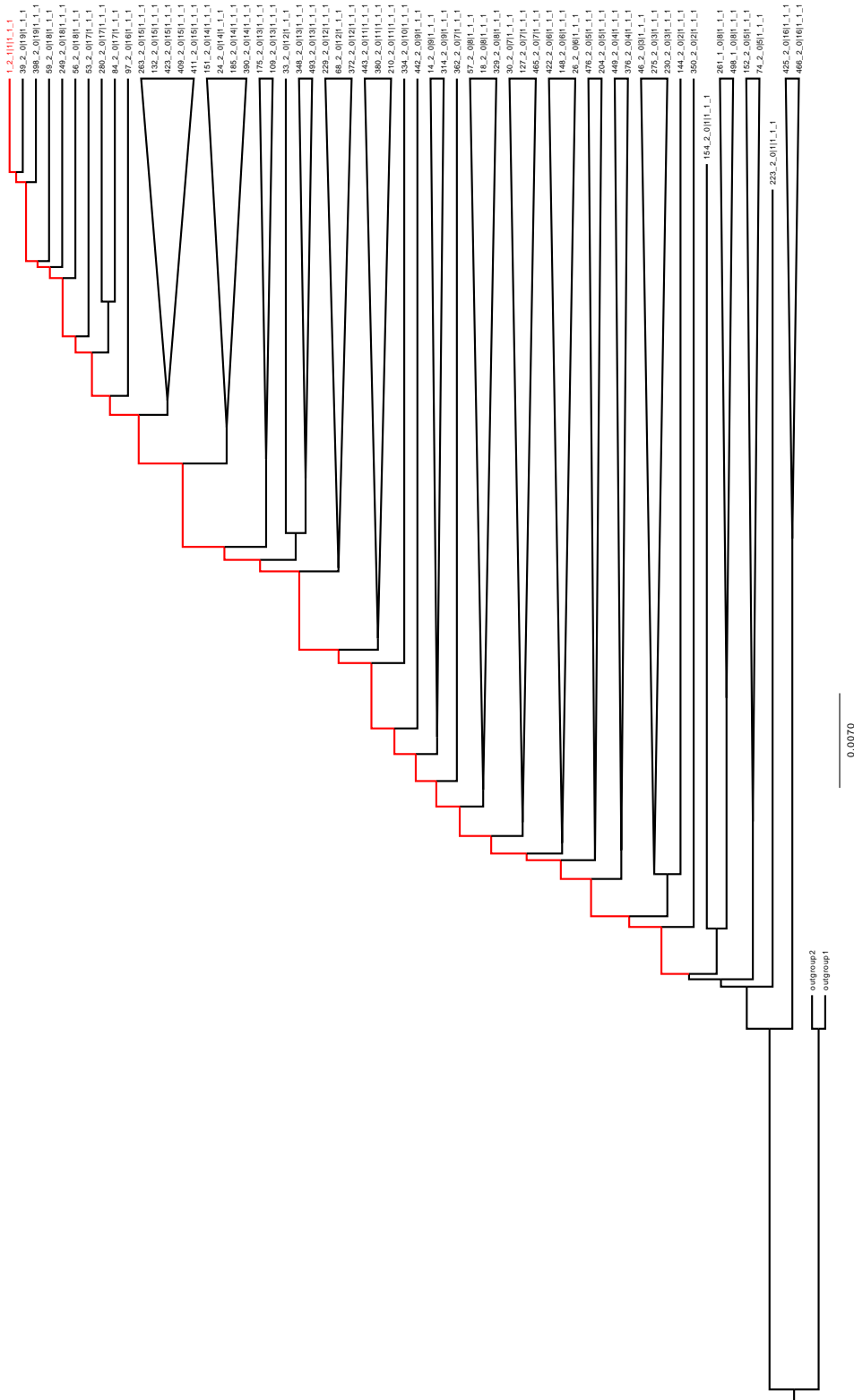


Figura 5.21: Experimento B: Árvore de população constante

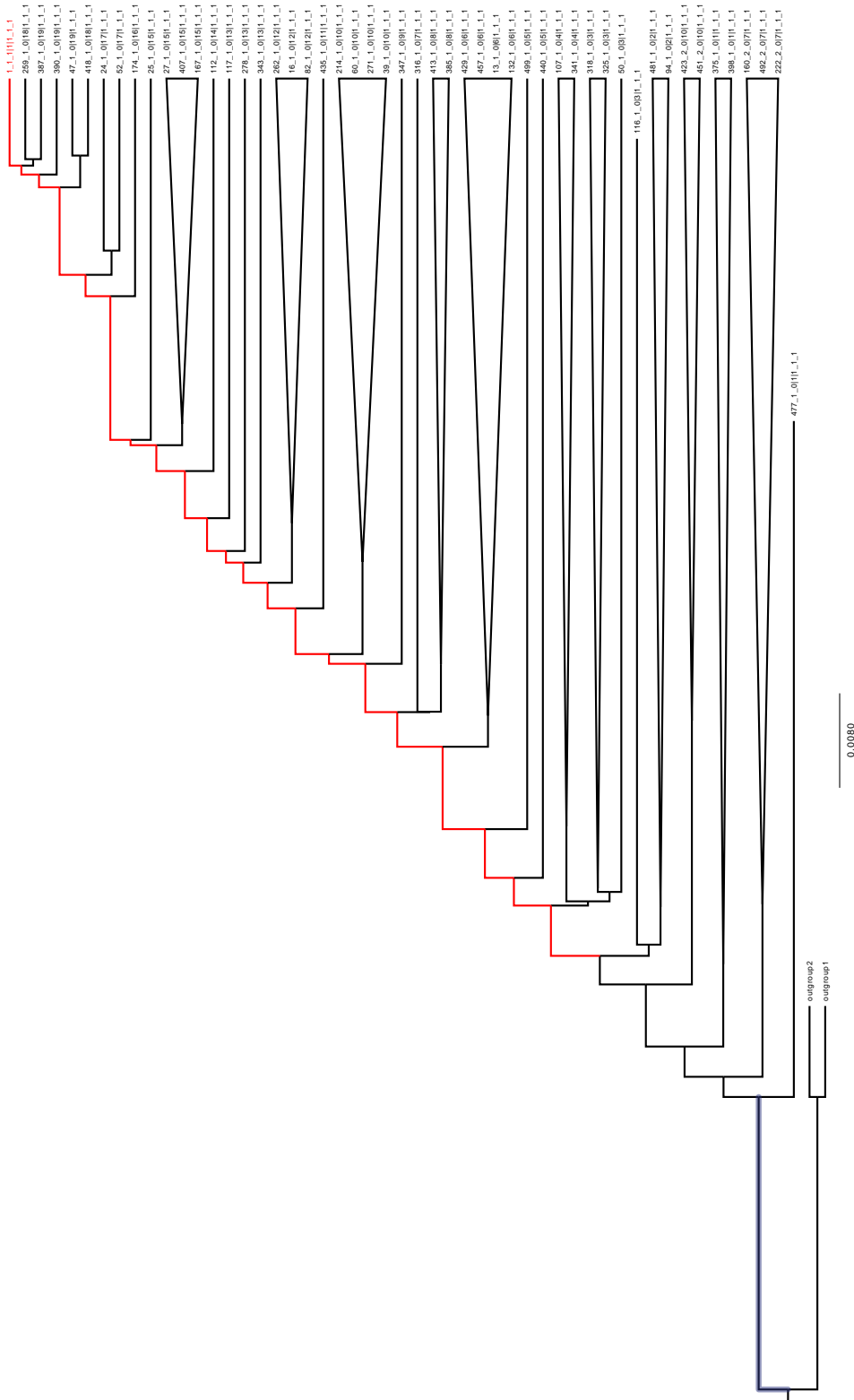


Figura 5.23: Experimento B: Árvore de população logística

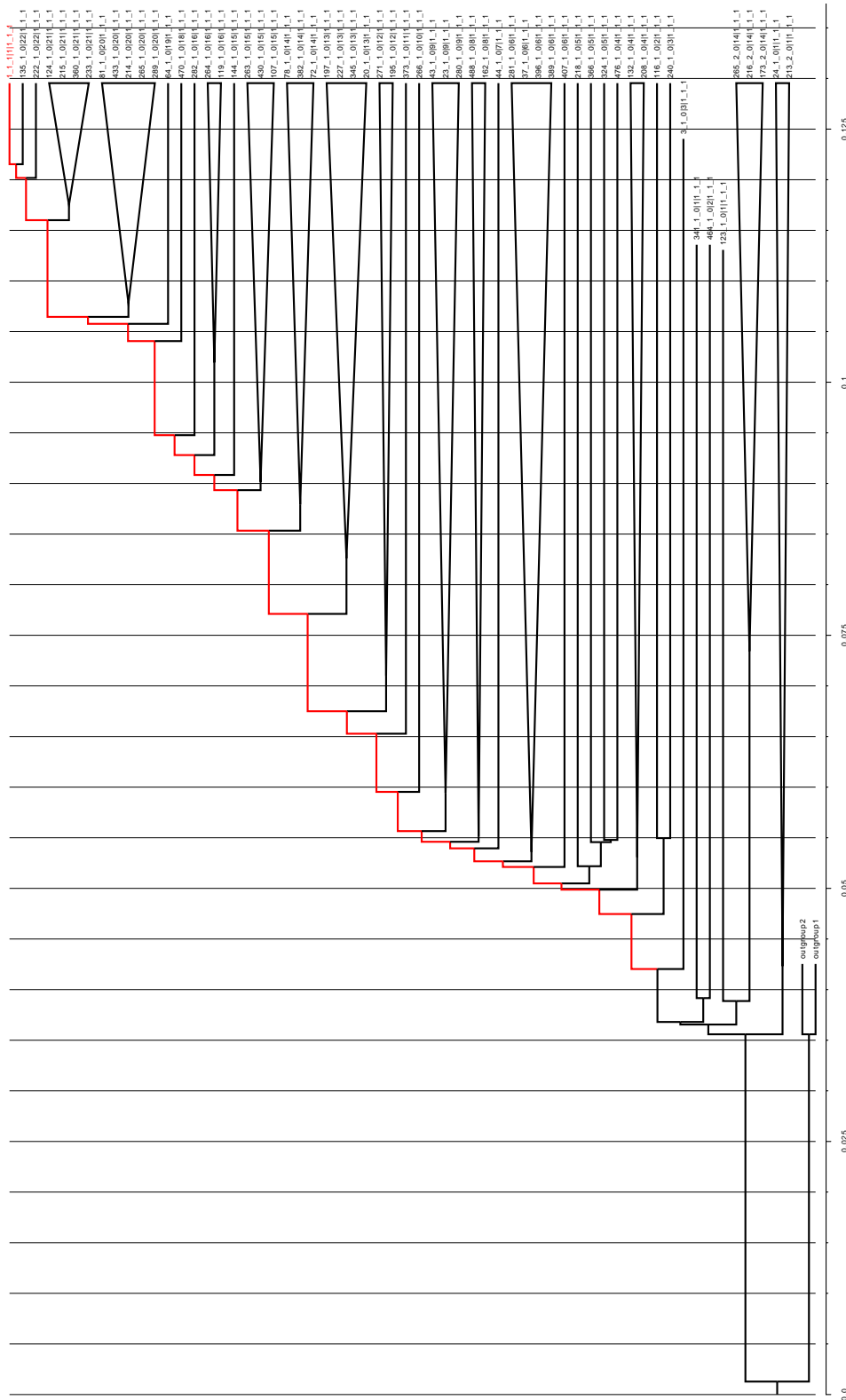


Figura 5.24: Experimento B: Árvore de população exponencial (decrecente)

Modelo demográfico não-paramétrico

Para este experimento, as condições de crescimento do número total de TEs na população são semelhantes ao cenário do experimento A, com $u = 5$ e *template* de 1 TE para a população F_0 . No entanto, a variável em estudo aqui é o modelo ecológico.

Cabe aqui uma observação sobre notação. De acordo com a terminologia típica da literatura de filodinâmica, o que estamos estimando é o tamanho efetivo da “população” de TEs, considerando-os como parasitos genômicos, diretamente ligados à sua população de hospedeiros. Porém, para evitar a confusão notacional entre a população do parasito e a população do hospedeiro, nos referiremos aos TEs apenas pela palavra “demografia” (e.g. “demografia dos TEs”). Continuaremos com a notação já estabelecida neste texto de chamar de população os modelos representados pelos hospedeiros que os carregam.

Nas figuras 5.25, 5.26, 5.27 e 5.28, podemos observar as inferências para os tamanhos efetivos da demografia de TEs nesses respectivos cenários. Nesse gráficos, o eixo Y foi transformado para uma escala logarítmica, típica desse tipo de análise. Uma composição das demografias de TEs dos quatro cenários em escala natural pode ser vista na figura 5.29, de forma a facilitar a comparação entre os cenários.

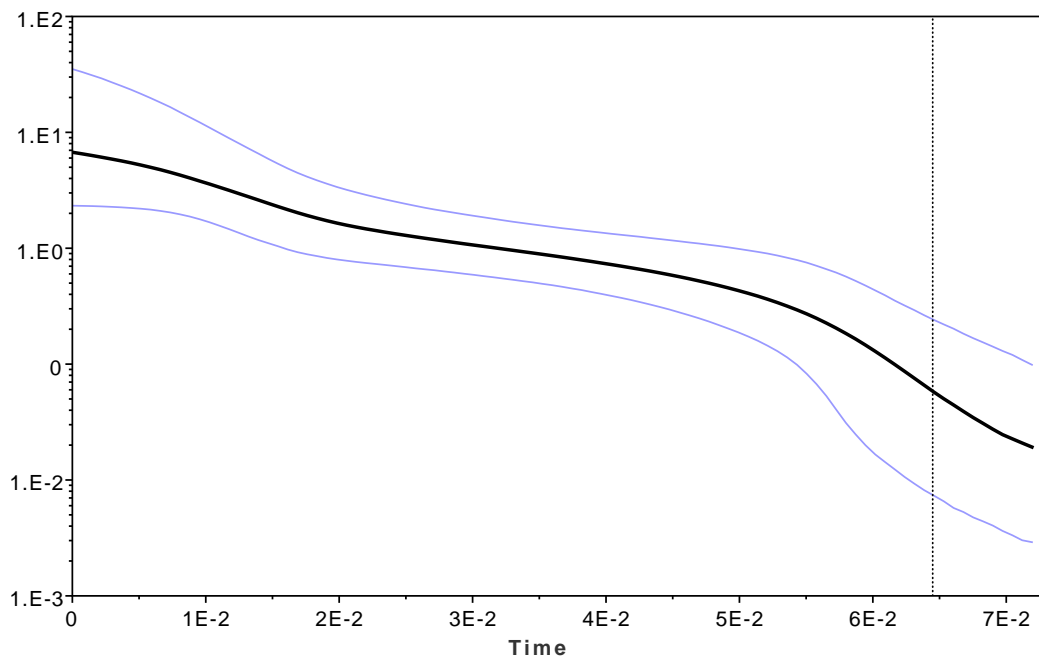


Figura 5.25: Experimento B: Skyline plot linear - população constante.

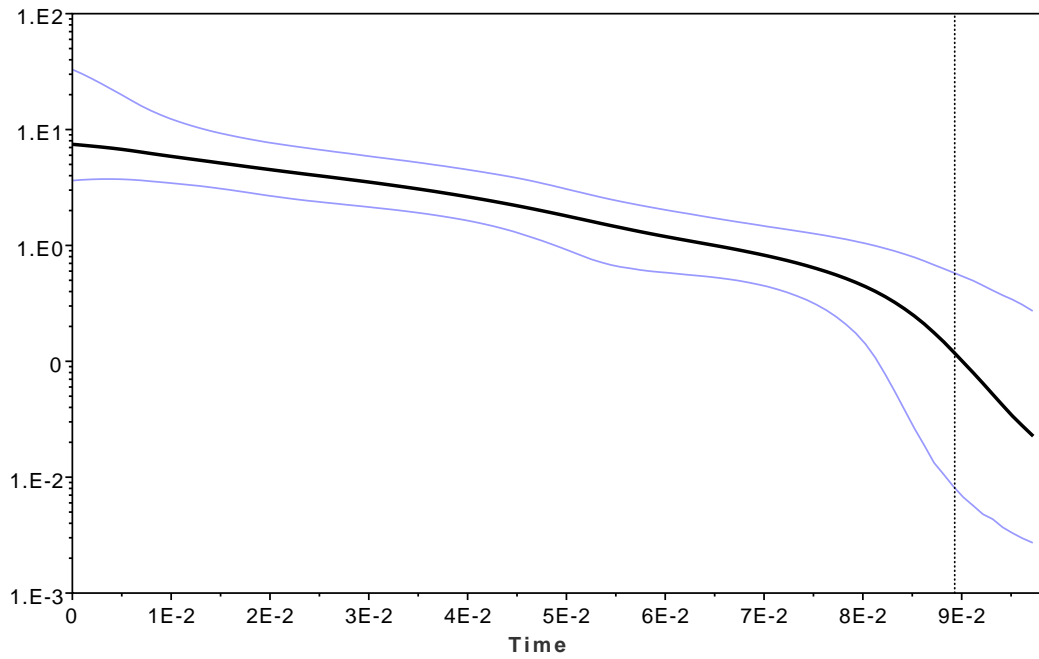


Figura 5.26: Experimento B: Skyline plot linear - população exponencial (crescente)

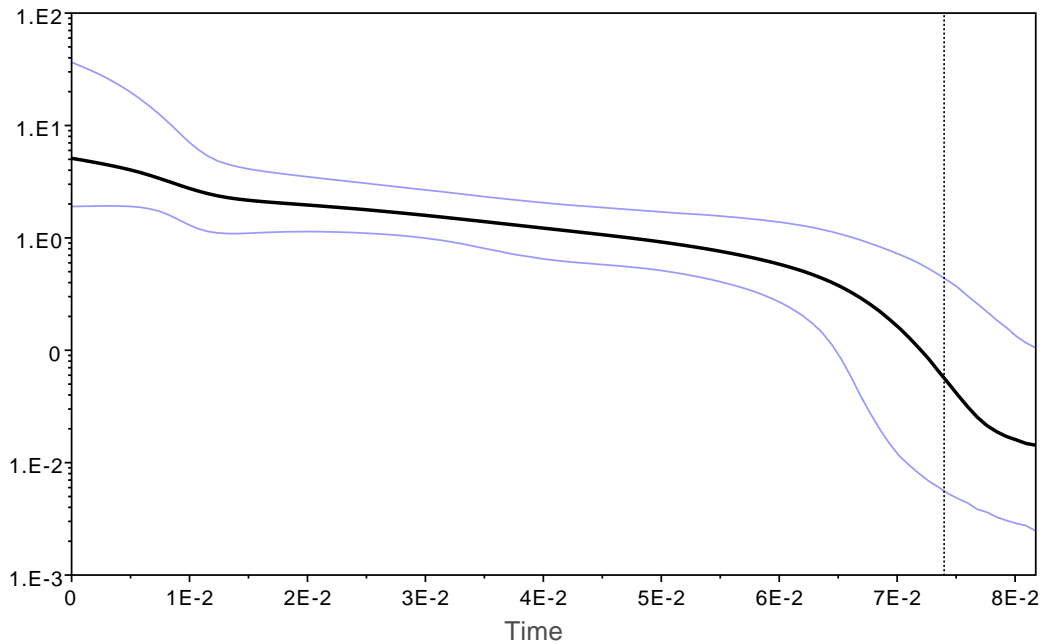


Figura 5.27: Experimento B: Skyline plot linear - população exponencial (decrecente)

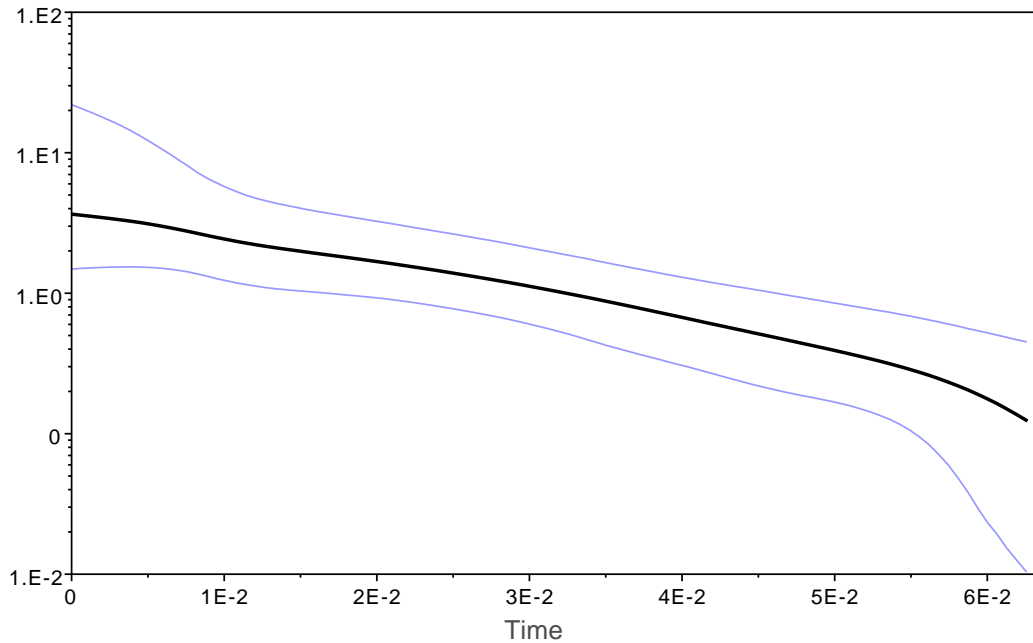


Figura 5.28: Experimento B: Skyline plot linear - população logística

Para o modelo ecológico de população constante (figura 5.25 e curva em verde na figura 5.29), observamos que a estimaco apresenta um crescimento com acentuaco em duas etapas, isto é, duas velocidades de crescimento correspondendo a duas fases, antes e após a perda do elemento ativo.

Conforme observamos nos resultados do experimento A, a existncia de dois momentos característicos na dinmica dos TEs se deve ao fato de que o número de TEs ativos diminui ao longo das geraçes. É possvel que essa alteraco qualitativa se deva à diminuico ou mesmo perda de elementos ativos na populaco, o que desaceleraria o crescimento da demografia dos TEs. Além disso, como no experimento B as simulaces foram executadas por um número de geraçes consideravelmente maior que no experimento A (até 100 geraçes no experimento B, contra até 30 geraçes no experimento A), a diferenca entre essas duas etapas se torna mais evidente.

No cenrio de populaco logística (figura 5.28 e curva em azul na figura 5.29), o tamanho efetivo da demografia parece ser aproximadamente linear ao longo das geraçes.

Observando os tamanhos finais das demografias inferidas (ponto em que tocam o eixo Y na figura 5.29), o cenrio de populaco exponencial crescente gerou um tamanho final para a demografia maior que o de populaco constante, enquanto a populaco exponencial

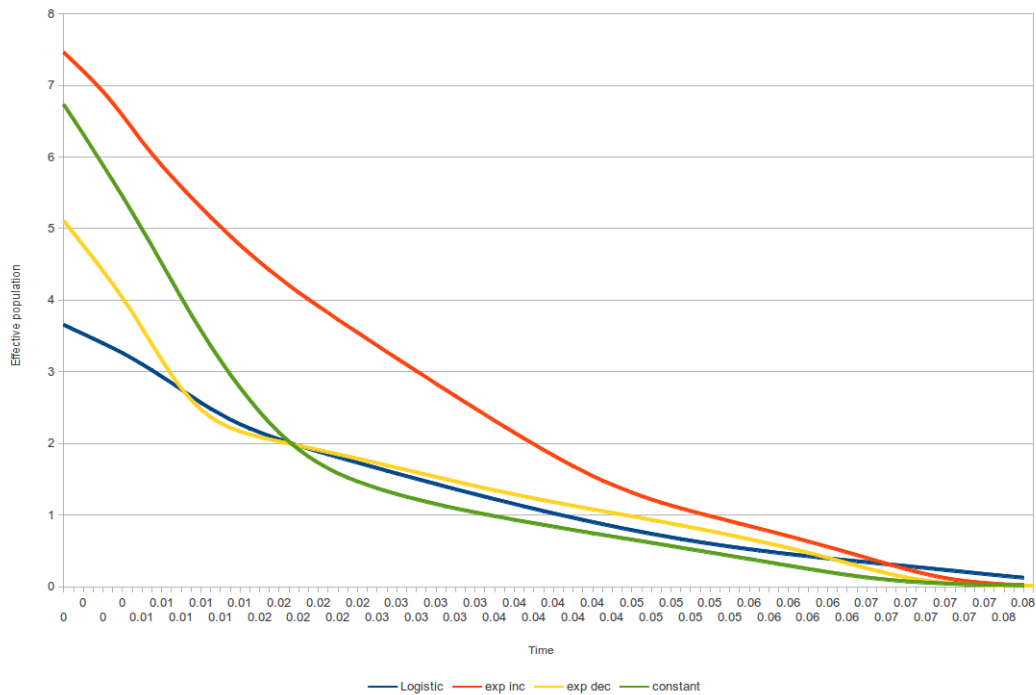


Figura 5.29: Experimento B: Skyline plot linear - 4 modelos

decrecente oferece um tamanho final menor. A população logística gerou um tamanho final menor que esses três.

Uma possível explicação para isso é que o crescimento da população expande o espaço no qual a demografia de TEs pode variar, aumentando a gama de possíveis hospedeiros, enquanto a contração populacional do modelo decrescente a limita, gerando um gargalo nas últimas gerações.

Embora a população logística tenha o potencial de atingir o mesmo patamar da população constante (10.000 indivíduos), isso não se daria num número pequeno de gerações. De fato, em todas as réplicas desse cenário a subpopulação GM atingiu a proporção de 90% da população, encerrando prematuramente a simulação. Para a população crescer até cerca de 9.000 indivíduos, as poucas réplicas que atingiram esse patamar levaram pelo menos 40 gerações. A maior parte das réplicas no entanto se encerraram em torno de 20 gerações.

Uma possível explicação para o baixo desempenho demográfico dos TEs, conforme aferido pela inferência da figura 5.28 pode ser feita por analogia aos resultados do ex-

perimento A. Observamos naqueles cenários que dificilmente o número de TEs cresce, e quando o faz parece ser apenas ao acaso por deriva genética. De fato, com a análise conjunta de várias réplicas, na média o número de TEs ativos tende a cair. Como esse cenário logístico inicia com uma população pequena, o número de hospedeiros com TEs ativos disponíveis para procriação parece ser muito pequeno, proporcionalmente aos outros cenários.

5.3.2 Com custo de *fitness*

Foram simulados cenários com custo de *fitness* $s = 0,01$, que contemplariam situações mais realistas. O experimento B foi executado com uma família de TEs *master gene*, e observamos que ao acrescentar uma pressão seletiva considerável, o tamanho da família despenca para números muito pequenos, mesmo após a perda do elemento ativo (MG). Johnson e Brookfield (2006, [142]) sugerem que um único Master Gene estrito dificilmente seria capaz de amplificar a família em quantidade suficiente para invadir uma população, o que sugere que nos estágios iniciais uma família desse tipo utilize uma estratégia mais eficiente, e passe a usar a estratégia *master gene* apenas após o equilíbrio do número de elementos.

Tabela 5.6: Experimento B: Resultados para cenários com $s = 0,01$, e diversos modelos populacionais. As médias de TEs por indivíduo foram tomadas pela última geração de cada réplica. As médias de TEs nas amostras foram calculadas observando o número de sequências em cada réplica. O cenário de população exponencial crescente não convergiu, e portanto não gerou dados analisáveis.

População	Média de TEs por indivíduo	Média de TEs nas amostras
constante	3,194	2,6
exponencial crescente	*	*
exponencial decrescente	4,038	5,4
logística	3,414	4

De fato, nos cenários em que foi utilizado um custo de *fitness* como pressão seletiva o número de sequências amostradas ao final de cada réplica é muito pequeno para obter convergência nas cadeias de Markov (cf. tabela 5.6), e embora isso inviabilize a nossa proposta de análise dos parâmetros evolutivos e demográficos da família com as técnicas de filodinâmica, o resultado é suficiente para satisfazer o objetivo de um *gene drive system*. Para este fim, basta que um único TE persista ao final da simulação, fixando-se na população.

O que de fato ocorre na natureza ainda é incerto. Embora existam evidências que suportem a aplicação de modelos com premissas do tipo *master gene*, observamos que nas fases iniciais de invasão é necessário um crescimento maior do que esse tipo de modelo de expansão é capaz de providenciar.

Capítulo 6

Discussão

“You will not apply my precept”, he said, shaking his head. “How often have I said to you that when you have eliminated the impossible, whatever remains, however improbable, must be the truth?”

Sherlock Holmes in “The Sign of the Four” (1890)

6.1 Do simulador

O simulador TRepid é grande em funcionalidade (> 50 parâmetros, muitos com várias opções), e essa complexidade permite realizar uma enorme gama de cenários de simulação, com diferentes níveis de realismo, de acordo com diversas hipóteses e premissas.

É uma plataforma ampla de simulação, flexível, expansível, com muitas opções de parâmetros oferecidas ao usuário. Sua interface mais simples é básica, bastando ao usuário criar um pequeno arquivo de texto segundo uma sintaxe intuitiva.

Os resultados dos experimentos MG e A indicam que o simulador é capaz de gerar cenários satisfatórios de expansão e invasão, estando de acordo com os modelos propostos na literatura [21, 122, 141]. Nossos resultados têm paralelo com resultados de outros estudos que analisam cenários semelhantes [45], com a vantagem adicional de considerarem pelo menos um nível de organização biológica a mais que qualquer outro trabalho de simulação de TEs.

Meus planos para o futuro próximo para o simulador pretendem expandir ainda mais as funcionalidades para cobrir todos os cenários típicos discutidos na literatura. Espero que isso torne o programa utilizável em uma gama ainda maior de estudos, gerando visibilidade ao grupo de pesquisa, e expandindo minhas potenciais linhas de pesquisa.

6.2 Dos experimentos

6.2.1 Experimento MG

Através de uma observação minuciosa do processo de geração de ramificações em torno de uma única linhagem, elaborada no experimento MG, podemos inferir a proporção de elementos ativos e inativos em árvores geradas a partir de sequências de TEs, como as do experimento B. Quando um único elemento é ativo para transposições, todas as folhas da árvore se ramificam a partir de uma única linhagem, gerando uma estrutura pectinada.

Simulações de dinâmica de expansão de TEs do tipo **master gene** são consideravelmente mais simples que modelos do tipo **transposon**, pois as relações entre membros da família de TEs podem ser representadas por topologias relativamente simples, com árvores pectinadas [141]. Famílias que seguem expansão do tipo **transposon** podem ter muitas linhagens gerando ramificações, e em alguns casos só podem ser convenientemente

representadas por redes filogenéticas [42]. Além disso, ao se analisar uma família nova, essas duas hipóteses podem ser confrontadas, e é possível testar a existência de um (ou poucos) elementos ativos que geram novas cópias numa família de TEs, conforme proposto por Johnson e Brookfield [142].

Observamos que mesmo com relativamente poucas iterações das cadeias MCMC da inferência filogenética bayesiana, os comprimentos dos ramos e distâncias entre clados parecem estar de acordo, na média, com o tempo marcado pelas gerações da população hospedeira.

Apesar das topologias NJ serem de fato incorretas, os comprimentos parecem indicar na maioria das ramificações a origem temporal das sequências, o que indica que nesse caso a hipótese do relógio molecular parece estar sendo satisfeita. Esse comportamento foi aproximado nas árvores bayesianas, mas não foi observado em nenhum dos quatro cenários examinados com o método NJ. É importante observar que isso provavelmente não se deve a um erro de inferência inerente ao método NJ, e sim à saturação evolutiva determinada pelas taxas de mutação atribuídas nas simulações, o que foi necessário para reduzir o número de gerações necessários para se observar um número razoável de novos TEs.

6.2.2 Experimento A

Examinamos no experimento A vários cenários de simulação em que diversas taxas de transposição u são contrapostas a vários níveis de pressão seletiva, representados pelo custo na fecundidade causado por TEs específicos que se acumulam. A comparação desses cenários evidencia a contribuição de cada parâmetro no fenômeno.

Ao longo da discussão dos resultados do experimento A, observamos que mesmo um modelo *master gene*, que não cria novos elementos ativos, é capaz de gerar grandes números totais e médios de TEs na população. Isso suporta a hipótese que, mesmo o cenário com menor capacidade replicativa tem ainda a capacidade de expandir de forma considerável, e potencialmente fixar pelo menos alguns elementos no genoma da população. No entanto, em condições realistas, a baixa capacidade replicativa em um sistema com propriedades *master gene* não é suficiente para gerar grandes quantidades de TEs em poucas gerações [142]. Os cenários aqui estudados indicam que, apesar desse tipo de sistema ser capaz de invadir espécies hospedeiras, a velocidade da invasão é pequena.

Por isso, uma família de TEs *master gene* não seria um candidato viável como sistema de *gene drive* para uso de substituição de populações silvestres de vetores de doenças transmissíveis, conforme as propostas de James [5].

Observamos que os cenários que não impõem custo de *fitness* mesmo não sendo exatamente razoáveis do ponto de vista biológico, servem de referência de comparação para os experimentos semelhantes em que há tal impacto. Sabe-se que invasões de TEs tipicamente causam redução da fecundidade nos hospedeiros [143]. De fato, uma das características que determina o sucesso na invasão e posterior domesticação dos elementos é a regulação entre os três principais fatores relacionados com a presença de uma nova família de TEs numa população de hospedeiros [45]: (i) uma fase inicial com alta taxa de transposição que possibilita a fixação dos TEs na população hospedeira, (ii) um custo de *fitness* que não sobrepuja a capacidade do hospedeiro em tolerar a presença e acumulação de TEs em seu genoma, (iii) uma fase de equilíbrio caracterizada pela redução da taxa de transposição possibilitando o equilíbrio do número de TEs.

O exemplo observado no caso com custo inicial elevado, e custo por TE moderado, indica que mesmo uma invasão definida por um impacto elevado na fecundidade dos hospedeiros, se a taxa de transposição não for suficientemente alta de modo a superar a capacidade da população em se recuperar, uma quantidade limitada de elementos pode ser fixada na população e passar a evoluir como genes ou pseudogenes. Apesar desses cenários não indicarem um bom candidato a *gene drive system* para a substituição total da população, ainda assim, o mecanismo dos TEs foi suficientemente eficiente para fixar uma subpopulação no ambiente.

Cenários sem capacidade de gerar pelo menos 50 sequências em indivíduos amostrados não foram encontrados em nossas simulações para este experimento. Mesmo nos cenários em que não houve invasão, o número médio de TEs por indivíduo na população cresceu consideravelmente, representando uma concentração do número de TEs em poucos indivíduos.

Uma diferença importante entre as premissas de nossos experimentos e as encontradas em resultados analíticos ou de simulações anteriores (como [122, 141]), é a de que não consideramos a família de TEs em equilíbrio numérico. Pelo contrário, estamos interessados nos processos e forças evolutivas envolvidos nos estágios iniciais da invasão em uma nova população de hospedeiros, fenômeno que também foi amplamente estudado em trabalhos como [21, 45, 50, 127, 144, 145].

Um processo lento de invasão é obviamente mais parcimonioso com a população de hospedeiros, porém em populações pequenas a família de TEs pode ser rapidamente perdida por deriva gênica caso a fixação não ocorra rapidamente (em escala de tempo genealógico, ao invés de escala evolutiva). Para se fixar em uma população, uma família de TEs precisa se expandir em número suficiente logo nos primeiros estágios da invasão, antes que esse crescimento seja atenuado pelos mecanismos de autorregulação, regulação por seleção contra inserções deletérias e rearranjos cromossômicos no hospedeiro e ainda perda aleatória de TEs por deriva genética.

Como observação final, destacamos que no caso hipotético em que a família de TEs não causa impacto perceptível, mesmo uma taxa de transposição baixa e um modelo linear já são suficientes para garantir a expansão do número de TEs na população hospedeira. Isso torna evidente o fato de que o mecanismo usado pelos TEs é intrinsecamente mais eficiente que a genética mendeliana comum para fixação dos elementos. Este é um requisito importante para o uso desses elementos como candidatos a *gene drive systems*.

6.2.3 Experimento B

No experimento B sem pressão seletiva, as sequências resultantes de vários cenários de invasão bem sucedida (mensurado pela capacidade da família de TEs em se expandir na população hospedeira até um patamar mínimo de 90% desta) foram analisadas quanto ao tipo de topologia que caracteriza cada cenário populacional. Essas sequências foram ainda examinadas sob a ótica dos modelos baseados em coalescência, de modo a comparar a inferência demográfica com a demografia observada na simulação. As inferências demográficas foram informativas quanto aos respectivos cenários, e mesmo na ausência de pressão seletiva, geram topologias com características observáveis.

Vimos na figura 5.29 que a mudança de modelo evolutivo para a população de hospedeiros tem impacto perceptível na demografia de TEs, quando esta é a única variável nos cenários de simulação. Mesmo na ausência de pressão seletiva, isso pode ser detectado pela análise das sequências dos TEs usando as técnicas de coalescência descritas por Drummond *et al* [89] em um contexto bayesiano (*skyline plot*).

Especulamos que diferentes cenários de pressão seletiva sobre os TEs, gerarão assinaturas mais marcantes nas topologias, analogamente a resultados amplamente observados em estudos de filogenias de vírus sob pressão seletiva de sistema imune ou vacinação em

massa. Especificamente, nosso modelo de pressão seletiva é ativado em etapas discretas, conforme os TEs impactantes se acumulam no genoma de cada indivíduo, pois o custo é arredondado para o maior inteiro menor que o custo total calculado. Assim, como o custo por TE impactante foi fixado em $s = 0,01$, o impacto final no indivíduo só é diferenciado a cada 100 TEs - um indivíduo com 15 TEs impactantes sofre a mesma penalidade que outro com 20 elementos.

A metodologia de filodinâmica foi adaptada para uso com dados de TEs por Struchiner *et al* [110], mas as simulações deste experimento que consideraram impacto no *fitness* não geraram sequências suficientes para a reconstrução filogenética (cf. tabela 5.6). Como as famílias têm em geral o elemento MG perdido após algumas gerações (dados não mostrados nesta tese, caso análogo ao experimento A), o crescimento da família perde a força aceleradora causada pelo mecanismo replicativo dos TEs. Mais tentativas com novos conjuntos de parâmetros devem ser feitas para a otimização dos cenários e posterior análise desses dados com inferência de demografias não-paramétricas (*skyline plot*).

6.2.4 Modelagem matemática

Diversos modelos paramétricos foram propostos na literatura para explicar o comportamento de invasão e fixação de TEs em populações hospedeiras. Le Rouzic [122] fez uma revisão de muitas das técnicas e modelos presentes na literatura.

Modelos da dinâmica de invasão

Katzourakis *et al* [77] propõem um modelo simples que leva em conta as variações entre elementos ativos A e inativos I , com relação às suas taxas de criação e deleção.

$$\begin{aligned}\dot{A} &= bA(1-p) - iA - dA \\ \dot{I} &= bAp + iA - dI\end{aligned}\tag{6.1}$$

Aqui, b é a taxa de criação de novos elementos fixados no genoma, i a taxa de inativação de elementos ativos por mutações, p é a proporção de novos elementos que são inativados devido a mutações durante a transposição e d é a taxa de deleção de elementos, tanto ativos como inativos.

Struchiner *et al* [110] propõem um modelo semelhante, com as seguintes diferenças:

$$\begin{aligned}\dot{A} &= bA \left(\frac{1}{(k+I)} \right) - iA - d_A A \\ \dot{I} &= bAp + iA - d_I I\end{aligned}\tag{6.2}$$

Esse modelo incorpora taxas de deleção diferentes para TEs ativos e inativos, d_A e d_I respectivamente, e implementa um mecanismo de regulação da transposição que gera elementos ativos, considerando a taxa de criação inversamente proporcional ao número de TEs inativos. Outra hipótese deste modelo é que todos os TEs criados são ativos ($p = 0$), o que o torna um modelo do paradigma **transposon**, e portanto incompatível com o desenho experimental dessa tese.

Como nossos dados são simulados em um contexto mais simples que o observado na natureza, propomos um modelo reduzido:

$$\begin{aligned}\dot{A} &= -dA \\ \dot{I} &= uA - dI\end{aligned}\tag{6.3}$$

Esse modelo difere dos anteriores para representar as características do modelo *Master Gene*, isto é, todos os TEs criados a partir de um TE ativo são inativos ($p = 1$), e respeitando nossa hipótese que elementos ativos não são inativados ao longo de sua história ($i = 0$). Todos os TEs presentes na população apenas podem ser perdidos ao acaso na segregação, ou quando um novo elemento é sobreposto a um TE existente. A taxa de transposição (criação de novos elementos) b foi renomeada para u , para manter a coerência de nossa notação nesta tese.

Modelo da situação de equilíbrio

Em um trabalho recente, Brookfield [141] indica que resultados anteriores mostraram que a hipótese de que todos os novos TEs criados forem ativos e tiverem a mesma taxa de transposição, gera conclusões de valores excessivamente grandes para o tempo de ancestralidade comumente observado para sequências repetitivas de humanos, da ordem de bilhões de gerações. Isso indica que nem todas as famílias seguem o modelo **transposon** de replicação, e portanto uma parte ou a maioria dos novos TEs são criados inativos.

Considere uma quantidade constante ν de transposições por genoma. Desse total, ν_a é a taxa de transposição que cria elementos ativos, e ν_i a que cria elementos inativos.

Considere ainda uma taxa de deleção d , e um efeito pseudo-gene κ , que representa

a inativação de TEs previamente ativos por mutações nas estruturas terminais que são responsáveis pelo reconhecimento da transposase, ou retro-transposase.

A taxa de crescimento do número de elementos ativos é ν_a e a taxa de perda é $n_1(\kappa+d)$, onde n_1 é o número de elementos ativos no genoma.

Assim, no equilíbrio, o número de elementos ativos é

$$n_1 = \frac{\nu_a}{\kappa + d}.$$

O equilíbrio do número de elementos inativos n_2 é

$$n_2 = \frac{\nu_i + n_1\kappa}{d} = \frac{(\nu_i + \nu_a)\kappa + \nu_id}{d(\kappa + d)}$$

de modo que

$$n = n_1 + n_2 = \frac{\nu_a + \nu_i}{d}$$

É fácil ver que nas expressões acima, um modelo do tipo *Master Gene* tem como equilíbrio n_1 uma quantidade nula de elementos ativos.

6.3 Perspectivas futuras

Novos cenários de simulações:

- Modelos de transposição diferentes de *Master Gene*;
- População com estrutura de idade;
- Sequências estruturadas (Inativação considerando mutações deletérias nas regiões terminais que promovem a transposição/retro-transposição);
- Comparação entre TEs de Classe I e Classe II;
- Populações estruturadas mais gerais, que não correspondam necessariamente a insetos (e.g. relaxar a condição de não-superposição de gerações);
- Meta-populações;
- Regulação infrafamiliar de TEs, com TEs ativos não-autônomos regulando TEs autônomos;

- Competição entre duas famílias de TEs;
- Formulações analíticas e análise matemática qualitativa do modelo composto, envolvendo todos os níveis de organização biológica;
- Populações assexuadas.

Capítulo 7

Conclusões

“Education never ends, Watson. It is a series of lessons, with the greatest for the last.”

Sherlock Holmes in “The Adventure of the Red Circle” (1911)

- Modelos simples como *master gene* geram topologias previsíveis, com estruturas pectinadas ramificando-se a partir da linhagem do elemento ativo;
- Nessas condições, a ordem dos eventos de transposição fica discriminada na topologia, e em alguns casos o tempo entre esses eventos pode ser observado graficamente na árvore;
- O conflito entre a taxa de transposição, a perda de elementos ativos e o impacto no fitness do hospedeiro emerge em simulações de base individual;
- Ao estimar um modelo demográfico não-paramétrico a partir de sequências de TEs, pode-se observar a influência da demografia de seus hospedeiros.

Capítulo 8

Referências Bibliográficas

- [1] WHO. World Malaria Report 2011. World Health Organization; 2011. Acessado em Março de 2012.
- [2] Ffrench-Constant RH. Which came first: insecticides or resistance? *Trends Genet.* 2007;23(1):1–4.
- [3] Day T, Galvani A, Struchiner C, Gumel A. The evolutionary consequences of vaccination. *Vaccine.* 2008;26 Suppl 3:C1–3.
- [4] Osta MA, Christophides GK, Kafatos FC. Effects of mosquito genes on Plasmodium development. *Science.* 2004;303(5666):2030–2.
- [5] James AA. Gene drive systems in mosquitoes: rules of the road. *Trends Parasitol.* 2005;21(2):64–7.
- [6] Christophides GK. Transgenic mosquitoes and malaria transmission. *Cell Microbiol.* 2005;7(3):325–33.
- [7] Boete C, Koella JC. A theoretical approach to predicting the success of genetic manipulation of malaria mosquitoes in malaria control. *Malar J.* 2002;1:3.
- [8] Gould F, Schliekelman P. Population genetics of autocidal control and strain replacement. *Annu Rev Entomol.* 2004;49:193–217.
- [9] Hemingway J, Craig A. Parasitology. New ways to control malaria. *Science.* 2004;303(5666):1984–5.
- [10] Riehle MA, Srinivasan P, Moreira CK, Jacobs-Lorena M. Towards genetic manipulation of wild mosquito populations to combat malaria: advances and challenges. *J Exp Biol.* 2003;206(Pt 21):3809–16.
- [11] Ito J, Ghosh A, Moreira LA, Wimmer EA, Jacobs-Lorena M. Transgenic anopheline mosquitoes impaired in transmission of a malaria parasite. *Nature.* 2002;417(6887):452–5.

- [12] Schliekelman P, Gould F. Pest control by the release of insects carrying a female-killing allele on multiple loci. *J Econ Entomol.* 2000;93(6):1566–79.
- [13] Knippling EF, Laven H, Craig GB, Pal R, Kitzmiller JB, Smith CN, et al. Genetic control of insects of public health importance. *Bull World Health Organ.* 1968;38(3):421–38.
- [14] Rasgon J. Population replacement strategies for controlling vector populations and the use of *Wolbachia pipientis* for genetic drive. *J Vis Exp.* 2007;(5):225.
- [15] Moreira LA, Wang J, Collins FH, Jacobs-Lorena M. Fitness of anopheline mosquitoes expressing transgenes that inhibit *Plasmodium* development. *Genetics.* 2004;166(3):1337–41.
- [16] Huho BJ, Ng’habi KR, Killeen GF, Nkwengulila G, Knols BG, Ferguson HM. Nature beats nurture: a case study of the physiological fitness of free-living and laboratory-reared male *Anopheles gambiae* s.l. *J Exp Biol.* 2007;210(Pt 16):2939–47.
- [17] Irvin N, Hoddle MS, O’Brochta DA, Carey B, Atkinson PW. Assessing fitness costs for transgenic *Aedes aegypti* expressing the GFP marker and transposase genes. *Proc Natl Acad Sci U S A.* 2004;101(3):891–6.
- [18] O’Brochta DA, Handler AM. Perspectives on the state of insect transgenics. *Adv Exp Med Biol.* 2008;627:1–18.
- [19] Rasgon JL. Multi-locus assortment (MLA) for transgene dispersal and elimination in mosquito populations. *PLoS One.* 2009;4(6):e5833.
- [20] Sinkins SP, Gould F. Gene drive systems for insect disease vectors. *Nat Rev Genet.* 2006;7(6):427–35.
- [21] Le Rouzic A, Capy P. Reversible introduction of transgenes in natural populations of insects. *Insect Mol Biol.* 2006;15(2):227–34.
- [22] Dawkins R. *The Selfish Gene.* Oxford University Press; 1976.
- [23] Doolittle WF, Sapienza C. Selfish genes, the phenotype paradigm and genome evolution. *Nature.* 1980;284(5757):601–3.

- [24] Chen CH, Huang H, Ward CM, Su JT, Schaeffer LV, Guo M, et al. A synthetic maternal-effect selfish genetic element drives population replacement in *Drosophila*. *Science*. 2007;316(5824):597–600.
- [25] Ruang-Areerate T, Kittayapong P. Wolbachia transinfection in *Aedes aegypti*: a potential gene driver of dengue vectors. *Proc Natl Acad Sci U S A*. 2006;103(33):12534–9.
- [26] Schliekelman P, Ellner S, Gould F. Pest control by genetic manipulation of sex ratio. *J Econ Entomol*. 2005;98(1):18–34.
- [27] Kondo N, Nikoh N, Ijichi N, Shimada M, Fukatsu T. Genome fragment of Wolbachia endosymbiont transferred to X chromosome of host insect. *Proc Natl Acad Sci U S A*. 2002;99(22):14280–5.
- [28] Kidwell MG, Ribeiro JM. Can transposable elements be used to drive disease refractoriness genes into vector populations? *Parasitol Today*. 1992;8(10):325–9.
- [29] Catteruccia F, Nolan T, Loukeris TG, Blass C, Savakis C, Kafatos FC, et al. Stable germline transformation of the malaria mosquito *Anopheles stephensi*. *Nature*. 2000;405(6789):959–62.
- [30] Ribeiro JM, Kidwell MG. Transposable elements as population drive mechanisms: specification of critical parameter values. *J Med Entomol*. 1994;31(1):10–6.
- [31] Rasgon JL, Gould F. Transposable element insertion location bias and the dynamics of gene drive in mosquito populations. *Insect Mol Biol*. 2005;14(5):493–500.
- [32] Catteruccia F, Godfray HC, Crisanti A. Impact of genetic manipulation on the fitness of *Anopheles stephensi* mosquitoes. *Science*. 2003;299(5610):1225–7.
- [33] Anxolabehere D, Kidwell MG, Periquet G. Molecular characteristics of diverse populations are consistent with the hypothesis of a recent invasion of *Drosophila melanogaster* by mobile P elements. *Mol Biol Evol*. 1988;5(3):252–69.
- [34] Kidwell MG, Lisch DR. Perspective: transposable elements, parasitic DNA, and genome evolution. *Evolution Int J Org Evolution*. 2001;55(1):1–24.

- [35] Magori K, Gould F. Genetically engineered underdominance for manipulation of pest populations: a deterministic model. *Genetics*. 2006;172(4):2613–20.
- [36] Hartl DL, Clark AG. *Principles of population genetics*. 3rd ed. Sinauer Associates; 1998.
- [37] McCLINTOCK B. The origin and behavior of mutable loci in maize. *Proc Natl Acad Sci U S A*. 1950;36(6):344–55.
- [38] Feschotte C, Jiang N, Wessler SR. Plant transposable elements: where genetics meets genomics. *Nat Rev Genet*. 2002;3(5):329–41.
- [39] Capy P. Evolutionary biology. A plastic genome. *Nature*. 1998;396(6711):522–3.
- [40] Kidwell MG, Lisch D. Transposable elements as sources of variation in animals and plants. *Proc Natl Acad Sci U S A*. 1997;94(15):7704–11.
- [41] Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, et al. A unified classification system for eukaryotic transposable elements. *Nat Rev Genet*. 2007;8(12):973–82.
- [42] Fernandez-Medina RD. ELEMENTOS DE TRANSPOSICIÓN EN EL GENOMA DEL MOSQUITO ANOPHELES GAMBIAE [Tese de Doutorado]. Escola Nacional de Saúde Pública Sérgio Arouca, Fiocruz. Rio de Janeiro, Brasil; 2009.
- [43] Kidwell MG. Transposable elements and the evolution of genome size in eukaryotes. *Genetica*. 2002;115(1):49–63.
- [44] Kazazian HH Jr. Mobile elements: drivers of genome evolution. *Science*. 2004;303(5664):1626–32.
- [45] Le Rouzic A, Capy P. The first steps of transposable elements invasion: parasitic strategy vs. genetic drift. *Genetics*. 2005;169(2):1033–43.
- [46] Brookfield JF. The ecology of the genome - mobile DNA elements and their hosts. *Nat Rev Genet*. 2005;6(2):128–36.

- [47] Coen D, Lemaitre B, Delattre M, Quesneville H, Ronsseray S, Simonelig M, et al. *Drosophila* P element: transposition, regulation and evolution. *Genetica*. 1994;93(1-3):61–78.
- [48] Pasyukova EG, Nuzhdin SV, Filatov DA. The relationship between the rate of transposition and transposable element copy number for copia and Doc retrotransposons of *Drosophila melanogaster*. *Genet Res*. 1998;72(1):1–11.
- [49] Vieira C, Biemont C. Transposition rate of the 412 retrotransposable element is independent of copy number in natural populations of *Drosophila simulans*. *Mol Biol Evol*. 1997;14(2):185–8.
- [50] Quesneville H, Anxolabehere D. Dynamics of transposable elements in metapopulations: a model of P element invasion in *Drosophila*. *Theor Popul Biol*. 1998;54(2):175–93.
- [51] Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, et al. The genome sequence of *Drosophila melanogaster*. *Science*. 2000;287(5461):2185–95.
- [52] Holt RA, Subramanian GM, Halpern A, Sutton GG, Charlab R, Nusskern DR, et al. The genome sequence of the malaria mosquito *Anopheles gambiae*. *Science*. 2002 Oct;298(5591):129–149.
- [53] Nene V, Wortman JR, Lawson D, Haas B, Kodira C, Tu ZJ, et al. Genome sequence of *Aedes aegypti*, a major arbovirus vector. *Science*. 2007;316(5832):1718–23.
- [54] Silva JC, Loreto EL, Clark JB. Factors that affect the horizontal transfer of transposable elements. *Curr Issues Mol Biol*. 2004;6(1):57–71.
- [55] Biemont C, Cizeron G. Distribution of transposable elements in *Drosophila* species. *Genetica*. 1999;105(1):43–62.
- [56] Capy P, Anxolabehere D, Langin T. The strange phylogenies of transposable elements: are horizontal transfers the only explanation? *Trends Genet*. 1994;10(1):7–12.
- [57] Diao Y, Qi Y, Ma Y, Xia A, Sharakhov I, Chen X, et al. Next-generation sequencing reveals recent horizontal transfer of a DNA transposon between divergent mosquitoes. *PLoS One*. 2011;6(2):e16743.

- [58] Cordaux R, Udit S, Batzer MA, Feschotte C. Birth of a chimeric primate gene by capture of the transposase gene from a mobile element. *Proc Natl Acad Sci U S A*. 2006;103(21):8101–6.
- [59] Kidwell MG, Lisch DR. Perspective: transposable elements, parasitic DNA, and genome evolution. *Evolution*. 2001;55(1):1–24.
- [60] Feschotte C, Pritham EJ. Mobile DNA: genomes under the influence. *Genome Biol*. 2006;7(6):320.
- [61] Feschotte C, Pritham EJ. DNA transposons and the evolution of eukaryotic genomes. *Annu Rev Genet*. 2007;41:331–68.
- [62] Jurka J. Conserved eukaryotic transposable elements and the evolution of gene regulation. *Cell Mol Life Sci*. 2008;65(2):201–4.
- [63] Schaack S, Gilbert C, Feschotte C. Promiscuous DNA: horizontal transfer of transposable elements and why it matters for eukaryotic evolution. *Trends Ecol Evol*. 2010;25(9):537–46.
- [64] Gotea V, Makalowski W. Do transposable elements really contribute to proteomes? *Trends Genet*. 2006;22(5):260–7.
- [65] Shapiro JA, von Sternberg R. Why repetitive DNA is essential to genome function. *Biol Rev Camb Philos Soc*. 2005;80(2):227–50.
- [66] Shapiro JA. A 21st century view of evolution: genome system architecture, repetitive DNA, and natural genetic engineering. *Gene*. 2005;345(1):91–100.
- [67] Shapiro JA. Transposable elements as the key to a 21st century view of evolution. *Genetica*. 1999;107(1-3):171–9.
- [68] Shapiro JA. Revisiting the central dogma in the 21st century. *Ann N Y Acad Sci*. 2009;1178:6–28.
- [69] Shapiro JA. Mobile DNA and evolution in the 21st century. *Mob DNA*. 2010;1(1):4.

- [70] Pritham EJ, Feschotte C. Massive amplification of rolling-circle transposons in the lineage of the bat *Myotis lucifugus*. *Proc Natl Acad Sci U S A*. 2007;104(6):1895–900.
- [71] Ray DA, Feschotte C, Pagan HJ, Smith JD, Pritham EJ, Arensburger P, et al. Multiple waves of recent DNA transposon activity in the bat, *Myotis lucifugus*. *Genome Res*. 2008;18(5):717–28.
- [72] Le Rouzic A, Dupas S, Capy P. Genome ecosystem and transposable elements species. *Gene*. 2007;390(1-2):214–20.
- [73] Feschotte C. Transposable elements and the evolution of regulatory networks. *Nat Rev Genet*. 2008;9(5):397–405.
- [74] Koonin EV. Darwinian evolution in the light of genomics. *Nucleic Acids Res*. 2009;37(4):1011–34.
- [75] Clark AG, Eisen MB, Smith DR, Bergman CM, Oliver B, Markow TA, et al. Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature*. 2007;450(7167):203–18.
- [76] Belshaw R, Pereira V, Katzourakis A, Talbot G, Paces J, Burt A, et al. Long-term reinfection of the human genome by endogenous retroviruses. *Proc Natl Acad Sci U S A*. 2004;101(14):4894–9.
- [77] Katzourakis A, Rambaut A, Pybus OG. The evolutionary dynamics of endogenous retroviruses. *Trends Microbiol*. 2005;13(10):463–8.
- [78] Hedges DJ, Cordaux R, Xing J, Witherspoon DJ, Rogers AR, Jorde LB, et al. Modeling the amplification dynamics of human Alu retrotransposons. *PLoS Comput Biol*. 2005;1(4):e44.
- [79] Higgs PG, Attwood TK. *Bioinformatics And Molecular Evolution*. Blackwell; 2005.
- [80] Jukes TH, Cantor CR. Evolution of protein molecules. In: Munro HN, editor. *Mammalian protein metabolism*. New York: Academic Press; 1969. p. 21–123.
- [81] Kimura M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J Mol Evol*. 1980;16(2):111–20.

- [82] Hasegawa M, Kishino H, Yano T. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J Mol Evol.* 1985;22(2):160–74.
- [83] Rodriguez F, Oliver JL, Marin A, Medina JR. The general stochastic model of nucleotide substitution. *J Theor Biol.* 1990;142(4):485–501.
- [84] Drummond AJ, Ho SY, Phillips MJ, Rambaut A. Relaxed phylogenetics and dating with confidence. *PLoS Biol.* 2006;4(5):e88.
- [85] Nei M, Kumar S. *Molecular Evolution and Phylogenetics.* Oxford University Press; 2000.
- [86] Yang Z. *Computational Molecular Evolution.* Oxford Series in Ecology and Evolution. Oxford University Press; 2006.
- [87] Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol.* 1987;4(4):406–25.
- [88] Marjoram P, Tavaré S. Modern computational approaches for analysing molecular genetic variation data. *Nat Rev Genet.* 2006;7(10):759–70.
- [89] Drummond AJ, Rambaut A, Shapiro B, Pybus OG. Bayesian coalescent inference of past population dynamics from molecular sequences. *Mol Biol Evol.* 2005;22(5):1185–92.
- [90] Opgen-Rhein R, Fahrmeir L, Strimmer K. Inference of demographic history from genealogical trees using reversible jump Markov chain Monte Carlo. *BMC Evol Biol.* 2005;5:6.
- [91] Lemey P, Rambaut A, Drummond AJ, Suchard MA. Bayesian phylogeography finds its roots. *PLoS Comput Biol.* 2009;5(9):e1000520.
- [92] Bloomquist EW, Lemey P, Suchard MA. Three roads diverged? Routes to phylogeographic inference. *Trends Ecol Evol.* 2010;25(11):626–32.
- [93] Allicock OM, Lemey P, Tatem AJ, Pybus OG, Bennett SN, Mueller BA, et al. Phylogeography and Population Dynamics of Dengue Viruses in the Americas. *Mol Biol Evol.* 2012;.

- [94] Nieberding CM, Olivieri I. Parasites: proxies for host genealogy and ecology? *Trends Ecol Evol.* 2007;22(3):156–65.
- [95] Moratorio G, Costa-Mattioli M, Piovani R, Romero H, Musto H, Cristina J. Bayesian coalescent inference of hepatitis A virus populations: evolutionary rates and patterns. *J Gen Virol.* 2007;88(Pt 11):3039–42.
- [96] Brand A, Brand H, Schulte in den Baumen T. The impact of genetics and genomics on public health. *Eur J Hum Genet.* 2008;16(1):5–13.
- [97] Loscalzo J, Kohane I, Barabasi AL. Human disease classification in the postgenomic era: a complex systems approach to human pathobiology. *Mol Syst Biol.* 2007;3:124.
- [98] Ebert D, Bull JJ. Challenging the trade-off model for the evolution of virulence: is virulence management feasible? *Trends Microbiol.* 2003;11(1):15–20.
- [99] Galvani AP. Epidemiology meets evolutionary ecology. *TRENDS in Ecology and Evolution.* 2003;18(3):132–139.
- [100] Medlock J, Luz PM, Struchiner CJ, Galvani AP. The impact of transgenic mosquitoes on dengue virulence to humans and mosquitoes. *Am Nat.* 2009;174(4):565–77.
- [101] Frearson JA, Wyatt PG, Gilbert IH, Fairlamb AH. Target assessment for antiparasitic drug discovery. *Trends Parasitol.* 2007;23(12):589–95.
- [102] Grenfell BT, Pybus OG, Gog JR, Wood JL, Daly JM, Mumford JA, et al. Unifying the epidemiological and evolutionary dynamics of pathogens. *Science.* 2004;303(5656):327–32.
- [103] Koelle K, Cobey S, Grenfell B, Pascual M. Epochal evolution shapes the phylodynamics of interpandemic influenza A (H3N2) in humans. *Science.* 2006;314(5807):1898–903.
- [104] Volz EM, Kosakovsky Pond SL, Ward MJ, Leigh Brown AJ, Frost SD. Phylodynamics of infectious disease epidemics. *Genetics.* 2009;183(4):1421–30.
- [105] Holmes EC, Grenfell BT. Discovering the phylodynamics of RNA viruses. *PLoS Comput Biol.* 2009;5(10):e1000505.

- [106] Frost SD, Volz EM. Viral phylodynamics and the search for an 'effective number of infections'. *Philos Trans R Soc Lond B Biol Sci.* 2010;365(1548):1879–90.
- [107] Bennett SN, Drummond AJ, Kapan DD, Suchard MA, Munoz-Jordan JL, Pybus OG, et al. Epidemic dynamics revealed in dengue evolution. *Mol Biol Evol.* 2010;27(4):811–8.
- [108] Rasmussen DA, Ratmann O, Koelle K. Inference for nonlinear epidemiological models using genealogies and time series. *PLoS Comput Biol.* 2011;7(8):e1002136.
- [109] Katia Koelle. Meeting Report from the RAPIDD Workshop on Phylodynamics. Durham, NC: NESCent; 2011.
- [110] Struchiner CJ, Massad E, Tu Z, Ribeiro JM. The tempo and mode of evolution of transposable elements as revealed by molecular phylogenies reconstructed from mosquito genomes. *Evolution.* 2009;63(12):3136–46.
- [111] Grimm V. Ten years of individual-based modelling in ecology: what have we learned and what could we learn in the future? *Ecological Modelling.* 1999;115(2-3):129–148.
- [112] Grimm V, Berger U, Bastiansen F, Eliassen S, Ginot V, Giske J, et al. A standard protocol for describing individual-based and agent-based models. *Ecological Modelling.* 2006 Sep;198(1-2):115–126.
- [113] Bonabeau E. Agent-based modeling: Methods and techniques for simulating human systems. *Proceedings of the National Academy of Sciences of the United States of America.* 2002;99(Suppl 3):7280–7287.
- [114] Gu W, Killeen GF, Mbogo CM, Regens JL, Githure JI, Beier JC. An individual-based model of *Plasmodium falciparum* malaria transmission on the coast of Kenya. *Trans R Soc Trop Med Hyg.* 2003;97(1):43–50.
- [115] Hoban S, Bertorelle G, Gaggiotti OE. Computer simulations: tools for population and evolutionary genetics. *Nat Rev Genet.* 2011;13(2):110–22.
- [116] Figueiredo F, Struchiner C. An individual-based model forward simulator for transposable elements dynamics and evolution; 2012. Draft to be submitted to the *Bioinformatics Journal*.

- [117] Cohen JE. Mathematics is biology's next microscope, only better; biology is mathematics' next physics, only better. *PLoS Biol.* 2004;2(12):e439.
- [118] Judson OP. The rise of the individual-based model in ecology. *Trends Ecol Evol.* 1994;9(1):9–14.
- [119] Peng B, Kimmel M. simuPOP: a forward-time population genetics simulation environment. *Bioinformatics.* 2005;21(18):3686–7.
- [120] Hernandez RD. A flexible forward simulator for populations subject to selection and demography. *Bioinformatics.* 2008;24(23):2786–7.
- [121] Carvajal-Rodriguez A. GENOMEPOP: a program to simulate genomes in populations. *BMC Bioinformatics.* 2008;9:223.
- [122] Le Rouzic A, Deceliere G. Models of the population genetics of transposable elements. *Genet Res.* 2005;85(3):171–81.
- [123] Hey J. A computer program for forward population genetic simulation.; 2004. Software not submitted to peer-review.
- [124] Hassell MP. Density-dependence in single-species populations. *Journal of Animal Ecology.* 1975;44(1):pp. 283–295.
- [125] Yang Z. Statistical properties of a DNA sample under the finite-sites model. *Genetics.* 1996;144(4):1941–50.
- [126] Tajima F. Infinite-allele model and infinite-site model in population genetics. *Journal of Genetics.* 1996;75(1):27–31.
- [127] Struchiner CJ, Kidwell MG, Ribeiro JMC. Population Dynamics of Transposable Elements: Copy Number Regulation and Species Invasion Requirements. *Journal of Biological Systems.* 2005 Dec;13(4):455–475.
- [128] Blumenstiel JP, Hartl DL, Lozovsky ER. Patterns of insertion and deletion in contrasting chromatin domains. *Mol Biol Evol.* 2002;19(12):2211–25.
- [129] Petrov DA, Hartl DL. Trash DNA is what gets thrown away: high rate of DNA loss in *Drosophila*. *Gene.* 1997;205(1-2):279–89.

- [130] Petrov DA, Hartl DL. High rate of DNA loss in the *Drosophila melanogaster* and *Drosophila virilis* species groups. *Mol Biol Evol.* 1998;15(3):293–302.
- [131] Petrov DA, Sangster TA, Johnston JS, Hartl DL, Shaw KL. Evidence for DNA loss as a determinant of genome size. *Science.* 2000;287(5455):1060–2.
- [132] Deceliere G, Charles S, Biemont C. The dynamics of transposable elements in structured populations. *Genetics.* 2005;169(1):467–74.
- [133] Lozovskaya ER, Nurminsky DI, Petrov DA, Hartl DL. Genome size as a mutation-selection-drift process. *Genes Genet Syst.* 1999;74(5):201–7.
- [134] Brookfield JF, Badge RM. Population genetics models of transposable elements. *Genetica.* 1997;100(1-3):281–94.
- [135] Brookfield JF. Evolutionary forces generating sequence homogeneity and heterogeneity within retrotransposon families. *Cytogenet Genome Res.* 2005;110(1-4):383–91.
- [136] Batzer MA, Deininger PL. Alu repeats and human genomic diversity. *Nat Rev Genet.* 2002;3(5):370–9.
- [137] Figueiredo F, Struchiner C. *Phylodynamics of Transposable Elements*; 2012. Early draft.
- [138] Drummond AJ, Rambaut A. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol Biol.* 2007;7:214.
- [139] Drummond AJ, Suchard MA, Xie D, Rambaut A. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol Biol Evol.* 2012;.
- [140] Thompson JD, Gibson TJ, Higgins DG. Multiple sequence alignment using ClustalW and ClustalX. *Curr Protoc Bioinformatics.* 2002;Chapter 2:Unit 2.3.
- [141] Brookfield JF, Johnson LJ. The evolution of mobile DNAs: when will transposons create phylogenies that look as if there is a master gene? *Genetics.* 2006;173(2):1115–23.

- [142] Johnson LJ, Brookfield JF. A test of the master gene hypothesis for interspersed repetitive DNA sequences. *Mol Biol Evol.* 2006;23(2):235–9.
- [143] Medstrand P, van de Lagemaat LN, Dunn CA, Landry JR, Svenback D, Mager DL. Impact of transposable elements on the evolution of mammalian gene regulation. *Cytogenet Genome Res.* 2005;110(1-4):342–52.
- [144] Quesneville H, Anxolabehere D. A simulation of P element horizontal transfer in *Drosophila*. *Genetica.* 1997;100(1-3):295–307.
- [145] Le Rouzic A, Capy P. Population genetics models of competition between transposable element subfamilies. *Genetics.* 2006;174(2):785–93.
- [146] Posada D, Crandall KA. MODELTEST: testing the model of DNA substitution. *Bioinformatics.* 1998;14(9):817–8.
- [147] Frech K, Danescu-Mayer J, Werner T. A novel method to develop highly specific models for regulatory units detects a new LTR in GenBank which contains a functional promoter. *J Mol Biol.* 1997;270(5):674–87.

Apêndice A

Manuscritos de artigos

Com relação aos manuscritos dos artigos incluídos em anexo ao texto principal desta tese, algumas observações são pertinentes.

A.1 An individual-based model forward simulator for transposable elements dynamics and evolution

A.1.1 Manuscrito principal

O manuscrito do artigo [116] encontra-se em fase final de preparação. No entanto, para ser submetido à revista **Bioinformatics** ele precisa ser reduzido para caber em 2 páginas.

An individual-based model forward simulator for transposable elements dynamics and evolution

Felipe Figueiredo^{1,2*}, Claudio Struchiner²

¹Programa de Biologia Computacional e Sistemas, Instituto Oswaldo Cruz (BCS/IOC/Fiocruz)

²Programa de Computação Científica, PROCC/Fiocruz

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Motivation: Transposable Elements (TEs) are small genomic elements present in almost all genomes sequenced so far, that appear repeatedly in the individuals genome. Much debate is ongoing about their origin, but some mechanisms for the invasion of a host individual and the later spread within the population are known from both *in silico* and wet laboratory experiments.

In order to assess both the evolutionary forces that promote variation in TEs and their impact on the host fitness, three independent levels of biological organization must be observed simultaneously: (a) the population demographics should be represented by a proper ecological model, (b) the TEs' population genetics should be given by a transposition model and (c) the TE sequences must evolve according to a known evolutionary model.

Results:

We present TRepid, an individual-based model that can be used to simulate TE invasion and fixation on an age-structured host population.

Host population dynamics features include random mating, selectable ecological models and recombination by crossing over of gametes. The TEs spread dynamics features include: transposition by either copy and paste or cut and paste, models that determine transposition rate, nucleotide substitution according to a selectable molecular evolutionary model, active and inactive TEs, inactivation of TE over generations, selective pressure on host individuals by accumulation of TEs and deleterious transposition events, among others.

Availability: The software is available with an open-source license from <https://launchpad.net/trepid>

Contact: ffigueiredo@ioc.fiocruz.br

1 INTRODUCTION

Several approaches have been proposed to understand and predict how the amount of TEs varies in hosts genomes (Le Rouzic and Deceliere, 2005; Hedges *et al.*, 2005; Struchiner *et al.*, 2005). Such mathematical models try to assess the invasion capabilities, accumulation and fixation of new copies and are usually based on mathematical and computational techniques.

Previous attempts in the modelling community focused in reproducing the dynamics predicted by differential or difference equations

that express how the total amount of TE copies vary in terms of acquisition of new copies and excision of old ones (Quesneville and Anxolabehere, 1997, 1998; Deceliere *et al.*, 2005, 2006). These types of models can typically be characterized in a continuum between two paradigm extremes: the *master gene* model, in which only one TE is active and generates inactive copies, and the *transposon* model in which every TE actively produces active copies, leading to an exponential growth of TE copy number in the absence of some regulation mechanism (Katzourakis *et al.*, 2005). Phylogenetic analysis can then provide the means to assess where in this spectrum a given TE family resides (Brookfield and Johnson, 2006; Johnson and Brookfield, 2006).

Most of the literature on models describing TE dynamics suffer from the lack of empirical data against which these models could be tested. It means that model diagnosis, an important step in model development, is missing. By looking at the sequences of families of transposable elements from the same genome, one can make inferences about their phylogenetic tree. This tree is influenced by, among other things, the changes in the number of elements of a given family over evolutionary time. Thus, in principle, one can make inferences about changes in the number of mobile elements in the genome from the phylogeny of the elements. Therefore, it becomes clear that examining the population dynamics of TEs through coalescent approaches is going to be highly informative (Brookfield, 2005; Brookfield and Johnson, 2006; Struchiner *et al.*, 2009). The main contribution of our work is to devise a simulation framework that mimics the empirical sequence data on TE dynamics where this dynamics is known. By exploring this simulation framework, we hope to validate the use of coalescent models to estimate potential parameters, and associated uncertainty in the estimation process, that describe TE invasion. In doing so, we hope to contribute an important tool to the debate about the introduction of transposable elements as part of genetic drive systems moving genes through disease vector populations.

2 METHODS

The simulator takes into consideration distinct and independent modeling techniques for each level of biological organization: a population model, a transposition model and an evolutionary model of the DNA sequences. Each of these sections are based on either differential or difference equations, or probabilistic models.

*to whom correspondence should be addressed

At the ecological level the population model produces a trend for the population dynamics. Currently there are models for constant and exponential growth, and growth with saturation (logistic and Hassel equations, see SOM for details). At the start of each generation, the necessary amount of sexually mature individuals is sampled and coupled to generate the required amount of offspring to follow the ecological model as closely as possible, notwithstanding fitness effects from TEs. The modular framework provide the means to implement any other ecological model. An age structure is also optionally available in discrete age classes.

At the population genetics level, the transposition model does the same thing for the amount of TEs in each new individual, based on the amount of TEs in the parental gametes. Transposition models are available with selection impact (Struchiner *et al.*, 2005) and without (Le Rouzic and Deceliere, 2005).

At the sequence level the evolutionary model determines how the TE sequences change over time, after successive transposition events, and an aging structure for TEs that escape excision from the host chromosomes.

2.1 The generation algorithm

The algorithm that happens at each generation models the basic life cycle of a diploid sexual species subjected to transposition events during gametogenesis.

1. Host couples are chosen randomly from available mature hosts at the beginning of each reproductive season. Males and females are chosen with and without replacement, respectively.
2. Each adult bears new gametes after transposition and recombination.
3. Transposition draws a recruitment amount of new TE copies and the deletion amount of excised copies from the transposition model. This changes the content of the gametes in terms of availability of TEs.
4. Mutations are sampled from the evolutionary model for newly created TE copies. If “cut and paste” transposition is being used, sample mutations for the original copy also. The same does not happen for “copy and paste”.
5. Recombination provides additional shuffling of gamete contents.
6. Mutations are sampled from the evolutionary model for all existing TE copies.
7. Each couple gives birth to a number of offspring defined by the user as a parameter.
8. The fitness cost from TEs in newborn individuals is calculated and any that exceeds a given threshold is killed before birth and removed from the population.
9. The age of every surviving individual is incremented at the end of the generation.

3 CONCLUSION

In this article we describe a computational model composed of a forward-time individual-based model for the population genetics and molecular evolution of transposable elements.

Population genetics software exist for both forward simulations (Carvajal-Rodriguez, 2008; Guillaume and Rougemont, 2006; Peng and Kimmel, 2005; Padhukasahasram *et al.*, 2008; Hernandez, 2008) and backward-time simulations (Hudson, 2002; Teshima and Innan, 2009), although most of them simply count the distribution and availability of a set of alleles that populate a given *locus* or *loci*. Similarly, there are simulators for transposition phenomena in host populations (Deceliere *et al.*, 2006) but they assume that TEs don't

change over time. As far as we are aware there is no simulator dealing with all three levels of biological organization concomitantly as well as considering how one level affect each other.

Additionally, at the end of each simulation an individual is optionally sampled from the population so an additional level of ecological modeling is being implicitly considered. This provides the means to take into account a sampling distribution in a statistical ecology framework.

ACKNOWLEDGEMENT

Funding: This work was partially supported by a Bill & Melinda Gates Foundation grant (FIXME), and a PhD scholarship by CAPES (FIXME).

REFERENCES

- Brookfield, J. F. (2005). Evolutionary forces generating sequence homogeneity and heterogeneity within retrotransposon families. *Cytogenet Genome Res*, **110**(1-4), 383–91.
- Brookfield, J. F. and Johnson, L. J. (2006). The evolution of mobile DNAs: when will transposons create phylogenies that look as if there is a master gene? *Genetics*, **173**(2), 1115–23.
- Carvajal-Rodriguez, A. (2008). GENOMEPOP: a program to simulate genomes in populations. *BMC Bioinformatics*, **9**, 223.
- Deceliere, G., Charles, S., and Biemont, C. (2005). The dynamics of transposable elements in structured populations. *Genetics*, **169**(1), 467–74.
- Deceliere, G., Letrillard, Y., Charles, S., and Biemont, C. (2006). TESD: a transposable element dynamics simulation environment. *Bioinformatics*, **22**(21), 2702–3.
- Guillaume, F. and Rougemont, J. (2006). Nemo: an evolutionary and population genetics programming framework. *Bioinformatics*, **22**(20), 2556–7.
- Hedges, D. J., Cordaux, R., Xing, J., Witherspoon, D. J., Rogers, A. R., Jorde, L. B., and Batzer, M. A. (2005). Modeling the amplification dynamics of human Alu retrotransposons. *PLoS Comput Biol*, **1**(4), e44.
- Hernandez, R. D. (2008). A flexible forward simulator for populations subject to selection and demography. *Bioinformatics*, **24**(23), 2786–7.
- Hudson, R. R. (2002). Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics*, **18**(2), 337–8.
- Johnson, L. J. and Brookfield, J. F. (2006). A test of the master gene hypothesis for interspersed repetitive DNA sequences. *Mol Biol Evol*, **23**(2), 235–9.
- Katzourakis, A., Rambaut, A., and Pybus, O. G. (2005). The evolutionary dynamics of endogenous retroviruses. *Trends Microbiol*, **13**(10), 463–8.
- Le Rouzic, A. and Deceliere, G. (2005). Models of the population genetics of transposable elements. *Genet Res*, **85**(3), 171–81.
- Padhukasahasram, B., Marjoram, P., Wall, J. D., Bustamante, C. D., and Nordborg, M. (2008). Exploring population genetic models with recombination using efficient forward-time simulations. *Genetics*, **178**(4), 2417–27.
- Peng, B. and Kimmel, M. (2005). simuPOP: a forward-time population genetics simulation environment. *Bioinformatics*, **21**(18), 3686–7.
- Quesneville, H. and Anxolabehere, D. (1997). A simulation of P element horizontal transfer in *Drosophila*. *Genetica*, **100**(1-3), 295–307.
- Quesneville, H. and Anxolabehere, D. (1998). Dynamics of transposable elements in metapopulations: a model of P element invasion in *Drosophila*. *Theor Popul Biol*, **54**(2), 175–93.
- Struchiner, C. J., Kidwell, M. G., and Ribeiro, J. M. C. (2005). Population Dynamics of Transposable Elements: Copy Number Regulation and Species Invasion Requirements. *Journal of Biological Systems*, **13**(4), 455–475.
- Struchiner, C. J., Massad, E., Tu, Z., and Ribeiro, J. M. (2009). The tempo and mode of evolution of transposable elements as revealed by molecular phylogenies reconstructed from mosquito genomes. *Evolution*, **63**(12), 3136–46.
- Teshima, K. M. and Innan, H. (2009). mbs: modifying Hudson's ms software to generate samples of DNA sequences with a biallelic site under selection. *BMC Bioinformatics*, **10**, 166.

A.1.2 Material suplementar (SOM)

Parte desse texto foi escrita de acordo com uma versão anterior simulador, portanto existem trechos desatualizados. Além disso, alguns detalhes dos algoritmos serão revisados, e outros incluídos, considerando a versão final do software.

O material suplementar online (SOM) incluirá também um exemplo de uso do programa, provavelmente baseado no experimento MG. Nessa seção deverão constar gráficos descrevendo diferentes cenários de população, e de modelos de transposição, bem como filogenias descrevendo os dois espectros de modelos de inativação, *Master Gene* e *transposon*, com seus diferentes padrões de ramificação de linhagens.

Supplementary online material for “An individual-based model simulator for transposable elements dynamics and evolution”

Felipe Figueiredo*

Claudio Struchiner

November 9, 2013

Contents

1 Overview of the simulator	2
1.1 The main players	2
1.2 Basic usage: Input	2
1.3 The default config file	3
1.4 General	3
1.5 Initial	3
1.6 Reproduction	3
1.7 Transposition	4
1.8 Evolution	6
1.9 Output	6
1.9.1 Log file	6
1.9.2 CSV tables	6
1.9.3 Sampled individual	6
1.10 Advanced usage: API for Perl programmers	6
2 Algorithms and models	7
2.1 Population	7
2.1.1 Ecological Model	8
2.1.2 Age structure	8
2.2 Molecular evolution	9
2.2.1 The genome	9
2.2.2 Recombination and ploidy	9
2.2.3 The evolutionary model	9
2.3 Transposition model	10
2.3.1 Activity cycle	10
2.3.2 Forms of impact on hosts	10
2.3.3 TE activity dynamics	11
2.4 The population genealogy	11
2.5 Sources of stochasticity	11
3 Availability	12

*ffigueiredo@ioc.fiocruz.br

4	Implementation details	12
4.1	Population models	12
4.1.1	Implementation of the logistic model	12
4.1.2	Constant model	12
4.2	Transposition models	13
4.2.1	Exponential model	13
4.2.2	Constant model	13
4.2.3	SKR model	13
4.3	Caching of deterministic models	14
5	Data structures	14
5.1	Host data structure	14
5.2	Population data structure	15
5.3	TE data structure	15
5.4	Meta-data format for Fasta sequences	15

1 Overview of the simulator

The **TRepid** simulator is an Object-Oriented system, with individual hosts and TEs each represented by objects. It's main interface with the user is provided by a plain-text configuration file that should be present in directory of the simulation. The simulation sub-directory tree will be created, and data and log files written by the simulator as required.

1.1 The main players

The main class defines simulation objects that contain all the information and data related to a given simulation. Associated with each simulation object is a configuration object to access or modify any parameter of the simulation, including equations parameters, file names, etc. It collects all configuration parameters and variables from configuration files, or use defaults if values are not provided by the user.

1.2 Basic usage: Input

The principal way of setting up a new simulation is through a configuration file called **trepid.conf**, inside which the user defines parameters for all the models.

The default configuration file that comes with the simulator has all parameters filled with default values, and comments explain what they mean. This file is divided in sections to organize semantics for the user. These categories are not interpreted though, since only variable names are used and users are free to remove the section names and define the config file in the order they prefer.

After preparing the configuration file, the user can run the **mmrrsim** program inside the directory where the config file is located. Upon starting, the simulator looks for the **trepid.conf** file, and if found, create a directory structure for the population (even if the population is not saved at the end of the simulation), a directory for logs and results, and finally a directory for the Templates for the hosts with TEs.

The Templates dir is where one or more genome templates are located, and used to create the GM (Genetically Modified, i.e. with TEs) subpopulation. If the user wants to use a GM population, at least one GM template must exist. If the templates directory is previously non-existent, a GM population won't be created, and in this case the simulator aborts.

If the configuration file is not found or is empty, default values will be used for all parameters. The same happens to any parameter not defined in the config file, if it is incomplete. The user can check what parameters were set for each simulation in the log header.

1.3 The default config file

In this section we describe all of the variables definable in the config file, showing the default values as illustration. These are the parameters used for a very simple simulation of a Master Gene model in a small host population that does not have generation overlap.

1.4 General

Simulation time and initial subpopulations

```
#####  
[General]  
# Number of generations (default: 10)  
it_max = 10
```

FIXME: include other general variables, save_*, sample_gm, interactive, often_max.

1.5 Initial

FIXME: include other initial variable: init_gender .

```
#####  
[Initial]  
# Initial Susceptibles (default: 900)  
S_0 = 0  
# Initial Infectious (default: 100)  
I_0 = 10000  
# Initial Removed (default: 0)  
R_0 = 0
```

The parameter `it_max` determines how many iterations of the model, i.e., how many generations will be simulated.

The F_0 population is determined by three initial subpopulations. In these parameters, the notations inspired by the SIR epidemic model of the early prototypes were maintained. `S_0` symbolizes the initial wild subpopulation, `I_0` the initial GM subpopulation (i.e. individuals that carry TEs) and `R_0` the Deceased or past individuals.

If a population was previously saved to disk, the files will be automatically read and the F_0 population will be constructed using those individuals. If the initial parameters for either of `S_0`, `I_0` and `R_0` are greater than the pre-existing individuals, additional ones will be created to match the simulation criteria.

1.6 Reproduction

Age structure

```
#####  
[Reproduction]  
# Life expectancy (default: 4)  
age_max = 4  
# Reproductive age (default: 3)  
age_reprod = 3  
# Offspring count (default: 100)  
offspring_count = 100
```

The two parameters `age_max` and `age_reprod` determine the age structure in the population. They are measured in generations. Setting the `age_reprod` to one unit lower than `age_max` configures a population without generation overlap. Setting `age_max` to 2 and `age_reprod` to 1 disables the age structure in the sense that all individuals last only one generation after being created.

```
# Recombination model (default: 1step)
recombination_model = "none"
```

The above parameters set the total offspring per couple, and turn off the recombination for the simulation.

```
# Reproduction model (default: Logistic)
reproduction_model="Logistic"
```

```
# Logistic parameters
# reproductive (default: 0.1)
r = 20
```

```
# carrying capacity (default: 5000)
K = 20000
```

The above options select the ecological model (in this case a logistic population) and define the model parameters (in this case r and K).

```
# ploidy parameters (default = 2)
ploidy = 1
```

The ploidy can be artificially set to diploid or haploid. Haploid populations are created in the following manner: upon creation of a new individual, the first chromosome comes from the father and the second from the mother, but since only one chromosome will be considered throughout the simulation. The contents of both chromosomes are merged, possibly overwriting

FIXME: portuguese.

Obs: A haploidia feita da seguinte forma: na criação do novo indivíduo, o primeiro cromossomo vem do pai, e o segundo vem da mãe, mas apenas um cromossomo final será considerado. Dessa forma, feita uma fusão dos dois cromossomos originais, possivelmente sobrescrevendo um TE, caso haja um TE no mesmo sítio em ambos os cromossomos originais. Isso foi feito para garantir que haja sempre um nico TE ativo, ao contrário do que aconteceria no caso diploide, em que um ativo herdado do pai, e outro da mãe. Na geração subsequente, dois seriam herdados do pai e dois da mãe e assim por diante, em progresso geométrica.

1.7 Transposition

```
#####
[Transposition]
# Species' categories sites (default: k=0, S=30, N=70)
killer_sites = 0
severe_sites = 0
neutral_sites = 200
```

FIXME: portuguese.

Obs: Tamanhos dos setores do cromossomo. Killer sites são os sites que matam o indivíduo imediatamente ao nascer, se algum estiver ocupado. Severe contabilizam custo de fitness e neutros não.

```
# Fitness cumulative impact of transposition (default .05)
severe_impact = .05
```

FIXME: portuguese.

Obs: Esse valor somado para cada TE que ocupar uma posio num severe site. O total arredondado para cima, e esse nmero inteiro ser a quantidade de prole que aquele indivduo no ser capaz de gerar devido aos TEs. Obs2: esse valor calculado para cada indivduo, e portanto deve ser contabilizado tanto pra o pai como para a me.

```
# Transposition model: constant, exponential or SKR (default: SKR)
transposition_model = "constant"
```

FIXME: portuguese.

Obs: O modelo constant acrescenta sempre a mesma quantidade de novas cpias em cada gametognese, dado que exista pelo menos uma cpia ativa.

```
# Transposition rate for constant and exponential model (default: .5)
transposition_rate=1
```

FIXME: portuguese.

Obs: A taxa de transposio tem significados diferentes para cada modelo. No caso do modelo constante, significa a quantidade de cpias novas a ser criadas. No caso do exponencial e do SKR, a probabilidade de que novas cpias sero criadas, isto , a prob. de ocorrer um evento de transposio naquela gametognese.

```
# Excision rate for constant and exponential models
excision_rate = 0
```

FIXME: portuguese.

Obs: anlogo a taxa de transposio.

```
# Deleterious threshold (default: severe + neutral)
# could be set to any arbitrary value
#delete = 10
```

FIXME: explain.

```
#Inactivation model: mastergene, random and progressive (default: mastergene)
inact_model = "mastergene"
```

FIXME: portuguese.

Obs: A atividade (STATUS na documentao) representada por um nmero entre 0 e 1. Qualquer nmero positivo representa um elemento ativo. Modelos de inativao: Master gene significa que TODA nova cpia inativa. Random sorteia um nmero entre 0 e o status do TE que est gerando a cpia. Progressive diminui o status em razo constante. usado um threshold inferior para que nmeros muito pequenos sejam considerados zero, e portanto inativar cpias.

```
# Invasion threshold for GM sub population (default: 0.9). Set to 1 to disable.
invasion_threshold = 0.9
```

```
# Loss threshold for GM sub population (default: 0.01). Set to 0 to disable.
loss_threshold = 0
```

FIXME: explain.

1.8 Evolution

```
#####
```

```
[Evolution]
```

```
# Probability that a mutation (substitution) will occur (default: .1)
```

```
mutate_prob = 1
```

FIXME: portuguese.

Obs: A probabilidade de que uma mutação vai ocorrer numa transposição.

```
# Number of point mutations to be introduced in each evolutionary event (default: 2)
```

```
mutation_count = 2
```

FIXME: explain.

1.9 Output

Several files are produced after a simulation ends.

1.9.1 Log file

Ao longo de uma simulação, são gerados um arquivo de log, dois arquivos CSV (um para a demografia e um com dados de TEs) e se um indivíduo puder ser amostrado para que suas sequências sejam analisadas, também ser gerado um arquivo com as sequências no formato FASTA e um no formato NEXUS (representando um alinhamento pronto para ser usado no BEAST). Todos os arquivos têm como prefixo a data seguida de um número serial, de modo que várias simulações consecutivas fiquem organizadas em ordem cronológica. Um exemplo com os arquivos output está em anexo.

FIXME: portuguese.

1.9.2 CSV tables

FIXME: explain.

1.9.3 Sampled individual

FIXME: explain.

1.10 Advanced usage: API for Perl programmers

Although the configuration file can provide a practical input for several simulation scenarios, some users might have more complex needs. Besides the configuration file there is an API which can be directly accessed by softwares, scripts or an interactive shell in Perl language which facilitates step-by-step procedures, analysis, reanalysis and more detailed data acquisition.

All data generated in each component can be accessed from the API, and accessor methods are provided for each data type for all object classes. Additionally, the whole generation algorithm can be bypassed and simulations can be manually run, step by step, and scrutinized in any way desired. One can change configuration parameters after a few generations for any of the models, or even the models themselves, insert or remove individuals in the population, et cetera.

The API documentation along with the complete list of functions and class and object methods are provided inline with the code and can be accessed by the usual `perldoc` command. PDF and HTML documents are also provided in the software package.

Each simulation run is identified by a unique ID string, and all objects related to that simulation are tagged with this ID, so several simulations can be run by the same instance or script without risk of

collision of variable values, either in batch or in parallel. We provide an example script to run a batch of simulations with different parameters.

2 Algorithms and models

The simulator is flexible in construction and able to produce a realistic scenario in which transposition events occur in finite populations that agree with the Wright-Fisher conditions [Hartl and Clark(1998)Hartl and Clark]. Transposable elements (TEs) replicate according to an arbitrary transposition model and TE sequences' evolve according to Kimura's infinite-sites model [Tajima(1996)Tajima], and an arbitrary model of molecular evolution [Yang(2006)Yang].

2.1 Population

As described above, the population has two independent structures: gender and age. We have drawn inspiration from the SIR epidemiological model to structure the population in discrete compartments. The age structure is also discrete, as described below.

The population is divided into **Wild** and **GM** (genetically modified) compartments of hosts similarly to the SIR compartmental model classes *Susceptible* and *Infected*. There's also a **Deceased** compartment, in analogy to the *Recovered* compartment, where individuals from earlier generations are kept for future reference or analysis. A given individual cannot change from the **Wild** to the **GM** compartment, though, and vice versa. Instead, these compartments' dynamics evolve through generations of sexual reproduction of the host population.

Time is measured in simulation steps representing discrete generations. The age structure is defined by two classes, regarding wither or not a host is mature enough to reproduce. Each new individual remains non-reproductive for a maturation period, after which it becomes mature and enter the reproductive pool in the population.

Age of sexual maturity and maximum lifespan are passed to the simulator as configuration parameters by the user. Default parameters for maturation, lifespan and offspring count reflect behaviour expected for strongly r -selected species with no generation overlap, suitable for modeling some insect populations such as mosquitoes or fruit flies.

Default age parameters	
Initial age	1
Reproductive age	3
Maximum age	4

The growth rate of the population is determined at the beginning of each generation from an arbitrary ecological model (see below), and from this rate the net income of new individuals for the next generation is obtained. The necessary amount of reproductive encounters to reach the appropriate number of new individuals for the next generation is then determined by dividing the total income by the expected number of offspring per couple as indicated in equation (1). Note that assuming a non-monogamous species, this is equal to the number of females expected to mate in this generation, as there should always be enough males to impregnate any available fertile females.

$$\text{reproductive encounters} = \frac{\text{total income of new hosts}}{\text{expected number of offspring}} \quad (1)$$

Couples are then pooled from the available mature population to satisfy the necessary number of reproductive encounters. We are modelling host species that are not monogamous, so males are chosen

from the population with replacement while females are chosen without replacement. This approximates the random mating behaviour (i.e., no spatial structure), while retaining the realistic behaviour that a given female can only copulate at most once per generation.

The host fitness is defined as the total offspring count that host has generated. The mean fitness over the population for a given generation is the arithmetic mean of the fitnesses of all live individuals at the end of the generation.

The offspring amount for each reproductive encounter could be drawn from a distribution whose mean is determined for the population. For simplicity, however, we consider in this version of the simulator a fixed initial amount of offspring for each couple. From this number it will be subtracted a quantity proportional to the impact TEs might have on the host fecundity. We postpone the definition of fitness impact and how it is calculated to the transposition model section.

Migration between meta-populations can be introduced in the simulator by the use of several geographic patches. Each of these patches is a micro-environment of its own, with the same compartment structure. For simplicity, in this version of the simulator we consider only one geographic patch, so migration events are ignored. This feature is planned for a future release.

2.1.1 Ecological Model

At each generation, the total number of new offspring must be determined before mating begins. To this end, an ecological model is consulted to determine the dynamics of the population in the absence of external influence (in our case, impact from TEs, which is described below).

The simulator is modular in the sense that the ecological and transposition models are implemented and considered independently from the rest of the system framework, so they function as exchangeable parts in the mechanism. This way, many other models can be included in the simulator in order to suit particular demands for each of these sections.

Indeed, any population model can be used in this system provided it can be represented in the form:

$$p_{n+1} = F(p_n) \tag{2}$$

where F is a function that depends on the current population size p_n . It is not required that F be deterministic, i.e., depends only on p_n . It can optionally depend on a random variable ξ_n to introduce intrinsic stochasticity in the growth rate:

$$p_{n+1} = F(p_n, \xi_n) \tag{3}$$

As a result, it is also not required that the model should be written explicitly as a differential or difference equation. The only requisite is that the input is the current population size, and the output is the expected quantity for the next generation.

We currently implement five population models, two models of unrestricted growth and two models for saturated growth: constant population, constant growth (linear), exponential growth (or decay), the Logistic and the Hassel [[Hassell\(1975\)Hassell](#)] equations. All of these are discretized from Ordinary Differential Equations, as described in section [4.1](#).

2.1.2 Age structure

The population has an age structure in which it is divided in discrete classes. The behavior of such models has been thoroughly studied in the ecological literature [[Nisbet and Gurney\(1982\)Nisbet and Gurney](#)]. One can configure the parameters to set any arbitrary number of age classes for a simulation, but there are two age classes of obvious interest: before and after reaching the reproductive age. The durations for the maturation stage and the adult stage last are selected as configuration parameters to the simulator, and

can be chosen to reproduce several scenarios, according to the organism being modeled. The default values for maturation and longevity are appropriate to species that don't have generation overlap, such as most insects.

(FIXME: mover para outra seo) The default value total offspring for each reproductive encounter is also selected as to represent r -selected species with high fecundity and high mortality, such as insects.

2.2 Molecular evolution

2.2.1 The genome

The genome in each individual is composed by two chromosomes, and each chromosome is represented discretely as a list of insertion sites that act as *loci* for TEs. Each insertion site in each chromosome can therefore be either empty or occupied by a TE. To reflect the fact that some insertions can cause disadvantageous, deleterious or even fatal mutations on the host, we categorize the insertion sites into three classes: fatal (*killer sites*), meaning they disrupt essential genes in a way they render the cell useless; severe (*severe sites*), which may disrupt non-essential genes, or metabolic pathways, and neutral (*neutral sites*). Under the hypothesis that within a species genomes from different hosts should bear more similarities than differences, the number of insertion sites in each of the above classes should not vary much within a species, and could be considered species specific if such numbers can be estimated based on genome size.

The genome size in this simulator is thus defined as a result of config parameters choices. Each chromosome has $\eta = k + s + n$ *loci*, each of which can contain one or zero TEs. These sites can be reordered, without loss of generality, in such a way that the first k sites are *killer sites*, the next s are *severe sites*, and the remaining n are of *neutral sites*.

2.2.2 Recombination and ploidy

A model of crossing-over recombination is optionally used to promote additional variability of chromosome content. It is implemented in three variants called *1-step*, *2-step* and *all-step*. The resulting gamete retains the original size in insertion sites, and each site is filled consulting one of the parent's corresponding site from one of the parental gamete.

In the *1-step* model a site is chosen at random and this site is the cutting point between the parts coming from each chromosome. The resulting gamete will be filled up to this site identical to the first chromosome and then the remaining sites are merged from the second chromosome.

The *2-step* model is similar to the above, but there are two cutting points instead of one. This means the first and third sections of the gamete are merged from the first chromosome and the second is merged from the other chromosome.

There's also an *all-step* model that abstracts from the two models above in the sense that every site is a cutting point. This way every site is taken from a random chromosome preserving only the order from which it came. This will provide maximum variability considering recombination events.

If none of the above recombination models is selected, the resulting gamete will be equal to one of the parental chromosomes, chosen at random.

Additionally, although we are mainly interested in simulating sexual populations, a workaround is provided to represent haploid simulations, for the benefit of simplicity. If this option is chosen, the second chromosome of the offspring being created is always ignored, and its contents copied to the first one respecting the original chromosome site, possibly overwriting an existing TE that previously existed there.

This "haploid sexual population" is a practical way of creating simple simulations and hypotheses.

2.2.3 The evolutionary model

Over time mutations occur and are accumulated in the genome of simulated hosts. Whenever a mutation occurs a substitution drawn from an evolutionary model.

We implemented a simple model that follow the assumptions of the Jukes-Cantor substitution model [Jukes and Cantor(1969)Jukes and Cantor]. All mutations are substitutions and equally likely to occur. As a result no special assumption is made over sites in the sequence which are more likely to change, nor transition/transversion bias. *All sites are considered equally likely to change as well as all nucleotides changes are considered equally likely to occur.*

As happens with the ecological section, almost any evolutionary model can be implemented in the simulator. The requirement here is that the input is the original sequence, and the output is the mutated sequence. Which changes are allowed or disallowed are completely up to the model in question. The evolutionary model is any model or function that, given an input nucleotide sequence, outputs a similar homologous sequence.

2.3 Transposition model

The existence of TEs in a host genome can cause several kinds of impacts in the host, ranging from fitness reduction, longevity reduction and even host inviability. There are several ways to model both how the TEs replicate in the genome, and the impact that such replication causes in the host that carry them. The three key elements to be considered in the modeling of these phenomena are the way the total amount of TE copies vary, the ability of TEs to be transposed and the impact that each TE might cause in the host.

2.3.1 Activity cycle

The initial phase of the TE invasion, should be regarded as the period during which the TE is most active. Otherwise, genetic drift may lead the TE to extinction [Le Rouzic and Capy(2005)Le Rouzic and Capy]. After this initial phase, if the TE succeeds in fixating in the host population, it must decrease its activity so as not to disrupt too much the host's fertility.

Transposable elements usually undergo some degeneration process, that inactivates TE copies. We take this phenomenon into account in the model in the form of an *status score*, that is a fixed number between zero and one for each TE. When a status of zero is selected, it renders the TE inactive. This score is also used to determine whether or not the TE is actively transposing.

Every time a new TE is created from an existing TE, the new copy's *status score* is chosen to be less than the original score. How fast this score drops to zero in successive transpositions can be implemented in several ways, and could be used to reproduce several interesting scenarios:

- if the status is constant and zero, then every new copy is inactive, and the only active copy is the original one. This is known as the *master gene model*;
- if the status is constant and non-zero, then every new copy is active, which represents the *transposon model*;
- the status can decrease in a constant rate, which will render new copies inactive after a few
- the new status can be chosen randomly, between zero and the original value

2.3.2 Forms of impact on hosts

New TEs can cause an impact on the host, depending on where in the genome it lands when created. This is implemented in the three categories of insertion sites where TEs might appear. If a TE is created in a *killer site*, it will be considered a deleterious transposition. Hosts can be born dead due to deleterious transposition, and otherwise the impact will be considered a fitness toll, thus the host will have a lower offspring count when compared.

All transposition models can be used with or without individual fitness impact on hosts, as an additive fitness impact can be optionally defined as a config parameter, which is a real number. All elements that have such an effect on the host are accounted for and their relative impact is summed and rounded up.

This total impact is the total amount of offspring this particular host will be unable to produce due to disadvantageous transpositions in its genome.

Lifespan can also be decreased if the impact is too great, albeit not fatal (as proposed in [Le Rouzic and Deceliere(2005)Le Rouzic and Deceliere]). We include this characteristic in the model as follows. First we take the ratio between the total fitness impact as calculated above and the maximum age defined in the config for the simulation, then this ratio is rounded down. This integer is the total age classes the host will be unable to achieve due to deleterious transpositions.

2.3.3 TE activity dynamics

Several transposition models can be implemented in the system. We implemented both neutral models in respect to natural selection (constant and exponential growth), and a transposition model that takes into account intrinsic deleterious effects by insertion of new elements [Struchiner *et al.*(2005)Struchiner, Kidwell, and Ribeiro].

Transposition models implemented in our framework have two main components: one for the the acquisition of new copies and one for the excision of existing copies.

The acquisition model (usually called by a `transpose()` function) determines the total amount of new copies that should be created for the gamete. The excision model (usually called by a `excise()` function) is the exact opposite. Parameter choices in the simulation configuration should consider cases where the creation of new copies never falls behind the deletion of old copies, otherwise the

2.4 The population genealogy

We are interested in analyzing sequence data from an individual to reconstruct the phylogenetic relations between its TEs. Therefore it becomes necessary to compare the reconstructed history with the real parental history we simulated.

Each host is identified by a name defined as a serial number, and carries the names of both parents. A simple recursive breadth-first algorithm is used to recover the names of parents, grandparents and so on. As such the genealogy can be fully reconstructed up to any valid time interval. The population stored in the **Deceased** compartment can be consulted to create genealogies of different depths after the execution of the simulation, or inspect each individual for it's genome, in order to compare the elements present in the sampled individual, and its ancestors.

2.5 Sources of stochasticity

Although most of models described so far involve deterministic models the overall behavior is stochastic. This happens due to both sampling effects and the fact that the system is an individual based model, or agent based model [Bonabeau(2002)Bonabeau]. The following are sources of such uncertainty:

1. sampling of individuals in finite population
2. recombination
3. replication and excision of TEs
4. mutations (substitutions) within TE sequences

The item 1 refers to Wright-Fisher models of populations [Hartl and Clark(1998)Hartl and Clark]. Since we're simulating small populations (typically $< 10^5$, but there's no restriction to population size) there is a sampling effect to be considered when an individual is chosen for reproduction. Should we simulate very large simulations the importance of sampling effects could be diminished, and we could approximate the behaviour of an infinite population (as in a Hardy-Weinberg population).

Since we're simulating sexual populations, we incorporate recombination by crossing over of games. This can be implemented in several ways, and the item 2 is related to the recombination model used to promote additional variability to the chromosome pool.

FIXME: falta terminar essa seo.

The item 3 is related to the availability of new TEs in the chromosomes (as defined by different *loci* and sequence).

The item 4 is related to the evolutionary model used to promote variability across generations.

3 Availability

The simulator has been developed and tested in GNU/Linux systems running Ubuntu Linux. As it is a portable language, it should run on any system for which its dependencies are available. Those include Mac OS X and Windows, besides UNIX/Linux in general. Packaging has been prepared following guidelines typical of Perl modules and applications, which should ease the installation in any platform Perl is available with the help of the CPAN framework¹, as well as Debian/Ubuntu DEB packages. The software is released in the open-source license GPL and is available from <https://launchpad.net/trepid>

4 Implementation details

4.1 Population models

4.1.1 Implementation of the logistic model

We will use the logistic equation as an example to describe the implementation because of its simplicity; the implementation of the Hassel model is analogous.

The differential equation is discretized as a single iteration of Euler's method, with a step size $h = 1$. This results in formula (4):

$$p_{n+1} - p_n = \Delta p = \text{Ceiling} \left(\frac{r \times p \times \left(1 - \frac{p}{K}\right)}{\text{offspring count}} \right) \quad (4)$$

where $\text{Ceiling}(x)$ ($\text{Floor}(x)$) is the smallest (greatest) integer greater (less) than or equal to x . Note that this implementation already takes into account the normalization by the total offspring count as per equation (1), so the result is the number of reproducing couples for that generation, instead of the total income of new individuals.

This model is deterministic so the caching system described in section 4.3 is used for it.

4.1.2 Constant model

The simplest population model available is arguably the constant population, which assumes equilibrium and zero net migration.

The parameter considered for the constant growth is the r config parameter (reproductive rate), which is used in this case as the total amount of new individuals to be created for each generation, rounded up if it's not already an integer.

$$\Delta p_n = r \quad (5)$$

¹<http://www.cpan.org>

4.2 Transposition models

4.2.1 Exponential model

The exponential model is a generalization of the constant model, but instead of only depending on a fixed quantity, it also depends on the relative number of copies existing at each time. This model also belongs to a wider class of neutral models that don't take into account any intrinsic regulation of copy number by fitness disadvantage to the host [Le Rouzic and Deceliere(2005)Le Rouzic and Deceliere], where the transposition rate u_n is a bounded function of c_n and equilibrium is attained when it saturates at the same value of the excision rate e .

$$\Delta c_n = (u_n - e)c_n \quad (6)$$

Considering the transposition rate u_n constant (therefore u), it can be rewritten as:

$$c_{n+1} - c_n = \Delta c = (u - e)c \quad (7)$$

that only depends (deterministically) on the current quantity of copies c in the reproducing host. As such, the same discretization scheme used in the population models can be used to provide a cache-able model that continually increases or decreases copy number over generations, depending on the value of $u - e$. If $u - e > 0$, the total copy number will always increase, whereas if $u - e < 0$ the simulator will tend to remove more copies that are added.

The model is consulted each time a host produces a gamete both for the amount of new TE copies to be created, and the amount of existing copies to be excised. If there must be new TE copies created, for each new copy a random active TE is sampled (with replacement) from the host genome, and replicated. Afterwards, if there must be excision of existing copies, for each excised copy the genome is sampled for a random TE (without replacement) and this copy is removed from the genome. Note that only active TEs get transposed, but any TE can be excised.

4.2.2 Constant model

The constant model of transposition is analogous to the respective population model counterpart. At each time the model is consulted for the amount of new TE copies to be created, the transposition rate config parameter *transposition_rate* is used as the total amount of new copies to be generated (rounded up). There is no significant dynamics in regard to the copy number growth, and it's not bounded per individual, except for the total number of insertion sites in the genome, defined in the config file (*killer_sites*, *severe_sites* and *neutral_sites*).

The excision of copies is also constant, defined by the config parameter e as the total amount of previously existing TE copies to be removed from the gamete.

$$\Delta c_n = u - e \quad (8)$$

4.2.3 SKR model

This model assumes an intrinsic self-regulation of TE copy number, and can assume various shapes and growth forms. The original article proposes that the form of the $U(c_n)$ function can be chosen as any functions that share the qualitative behaviour described, and suggests (and tests) three forms indexed as U_1 , U_2 and U_3 that we reproduced in this framework.

$$c_{n+1} = \text{Ceiling}(c_n + c_n \times T_0 \times U(c_n)) \quad (9)$$

$$U_1(c) = 2^{(-c/C_{0.5})} \quad (10)$$

$$U_2(c) = 1 - \frac{c^5}{(C_{0.5}^5 + c^5)} \quad (11)$$

$$U_3(c) = 1 + (c - \frac{0.5}{C_{0.5}}) \quad (12)$$

Transposition with this model is provided by the preceding equations, while excision is provided by the exact same function of the exponential model.

4.3 Caching of deterministic models

In order to save unnecessary calculations, caching of model data is implemented for both population and transposition models. All values calculated for deterministic models are cached to avoid the performance toll of unnecessary repeated calculations. This cache is implemented as a hash², and is consulted before each calculation to see if it's necessary.

The cached data are stored independently per simulation and per data type (population or transposition), so in case a user wants to instantiate several different simulations in the same session³, data from different simulations are guaranteed not to mix.

5 Data structures

The model is implemented in Perl5, using its object-oriented paradigm. The main classes `TRepid::Population`, `TRepid::Host` and `TRepid::TE` define, respectively, objects for the population, the hosts that constitute the population and the TEs that populate each host genome.

Two classes inherit methods from Bioperl classes [Stajich *et al.*(2002)Stajich, Block, Boulez, Brenner, Chervitz, Dagdigian, Fuellen, Gilbert, Korf, Lapp, Lehvaslaiho, Matsalla, Mungall, Osborne, Pocock, Schattner, Senger, Stein, Stupka, Wilkinson, and Birney]. The TE class inherits the `Bio::Seq` class, and the Host class inherits the `Bio::SeqIO` class. This guarantees the simulator can import and export sequences using a great range of sequence formats. It also uses the abstract model of representing DNA sequences provided by the Bioperl project in a way that unifies the usage of any sequence format (Fasta, genbank, embl, swissprot, etc) so that any of these formats can be used at the discretion of the user. We provide a wrapper for including header information in the Fasta format, which we describe in detail in a later section.

The documentation of the API is bundled within the code with the POD (Plain Old Documentation) format, which can be accessed by the ordinary means of any Perl distribution. Besides being available inline in the code, all this material is also converted and provided in more convenient formats like HTML and PDF.

5.1 Host data structure

The structure inherited from `Bio::SeqIO` provide methods to collect and save sequences from a file, which inspired an implementation for a file-based storage. Besides these general-purposed methods we defined

²Sometimes referred to as an associative array.

³This functionally requires usage of the library via the API. A sample script is provided as an example for programmers.

some object attributes to store meta-data related to each individual like its name, age, gender, fitness, the two chromosomes the file name to which the `Host` object is associated. Each of the two chromosomes is an array, and each of its positions stores a reference for a TE object if one is present. Methods for consulting the number or position of TEs and de-reference them are provided in the class API.

5.2 Population data structure

The object attributes for the `Population` class store information related to where the files are stored in disk, and some statistics that are frequently consulted.

5.3 TE data structure

Besides the structure inherited from `Bio:Seq`, which includes methods for setting and retrieving several meta-data as the sequence string, header information, etc.

5.4 Meta-data format for Fasta sequences

We default to using sequences in Fasta format. To this end, we propose a protocol for inserting meta-data into the Fasta header and a set of corresponding parser (writer) methods in order to retrieve (save) the corresponding information.

```
>string:int:int:real
```

The first field (string) denotes the **TE description**. It can vary in size and can contain spaces. Usually it comprises the whole Fasta header if a real sequence is used, as acquired from any genomic database, should one such sequence be used.

The next two fields that contain integer numbers. The first is the **site** position in the chromosome, and the second defines in which **chromosome** this particular TE is located. The total number of sites available can be chosen by the user, and the default is 100 positions (from 0 to 99). There are always two chromosomes (denoted by the numbers 0 and 1) since we are modelling sexual diploid populations. The last field is a real number, in the interval $[0, 1]$. It describes the **status score** of activity for that TE, where any non-zero value means the TE is active. Each new copy generated from this copy should have a lower score, effectively reproducing a deactivation function for the TE family.

References

- [Bonabeau(2002)Bonabeau] Bonabeau, E. (2002). Agent-based modeling: Methods and techniques for simulating human systems. *Proceedings of the National Academy of Sciences of the United States of America*, **99**(Suppl 3), 7280–7287.
- [Hartl and Clark(1998)Hartl and Clark] Hartl, D. L. and Clark, A. G. (1998). *Principles of population genetics*. Sinauer Associates, 3 edition.
- [Hassell(1975)Hassell] Hassell, M. P. (1975). Density-dependence in single-species populations. *Journal of Animal Ecology*, **44**(1), pp. 283–295.
- [Jukes and Cantor(1969)Jukes and Cantor] Jukes, T. H. and Cantor, C. R. (1969). Evolution of protein molecules. In H. N. Munro, editor, *Mammalian protein metabolism*, pages 21–123. Academic Press, New York.

- [Le Rouzic and Capy(2005)Le Rouzic and Capy] Le Rouzic, A. and Capy, P. (2005). The first steps of transposable elements invasion: parasitic strategy vs. genetic drift. *Genetics*, **169**(2), 1033–43.
- [Le Rouzic and Deceliere(2005)Le Rouzic and Deceliere] Le Rouzic, A. and Deceliere, G. (2005). Models of the population genetics of transposable elements. *Genet Res*, **85**(3), 171–81.
- [Nisbet and Gurney(1982)Nisbet and Gurney] Nisbet, R. M. and Gurney, W. S. C. (1982). *Modelling fluctuating populations*. John Wiley, Chichester. US.
- [Stajich *et al.*(2002)Stajich, Block, Boulez, Brenner, Chervitz, Dagdigian, Fuellen, Gilbert, Korf, Lapp, Lehvaslaiho, Matsall
Stajich, J. E., Block, D., Boulez, K., Brenner, S. E., Chervitz, S. A., Dagdigian, C., Fuellen, G., Gilbert, J. G., Korf, I., Lapp, H., Lehvaslaiho, H., Matsalla, C., Mungall, C. J., Osborne, B. I., Pocock, M. R., Schattner, P., Senger, M., Stein, L. D., Stupka, E., Wilkinson, M. D., and Birney, E. (2002). The Bioperl toolkit: Perl modules for the life sciences. *Genome Res*, **12**(10), 1611–8.
- [Struchiner *et al.*(2005)Struchiner, Kidwell, and Ribeiro] Struchiner, C. J., Kidwell, M. G., and Ribeiro, J. M. C. (2005). Population Dynamics of Transposable Elements: Copy Number Regulation and Species Invasion Requirements. *Journal of Biological Systems*, **13**(4), 455–475.
- [Tajima(1996)Tajima] Tajima, F. (1996). Infinite-allele model and infinite-site model in population genetics. *Journal of Genetics*, **75**(1), 27–31.
- [Yang(2006)Yang] Yang, Z. (2006). *Computational Molecular Evolution*. Oxford Series in Ecology and Evolution. Oxford University Press.

A.2 Phylodynamics of Transposable Elements

A.2.1 Manuscrito principal

O manuscrito [137] encontra-se em fase inicial de escrita, pois sua elaboração foi interrompida para reformulação do desenho experimental. A metodologia deste será reformulada para detalhar os experimentos seguindo o desenho experimental descrito nos experimentos MG, A e B apresentados nesta tese.

A fim de ampliar o panorama de investigação dos fenômenos envolvidos, serão feitas as seguintes adições aos experimentos:

- Expansão do número de cenários, para incluir outros (todos?) modelos ecológicos e de transposição;
- Aumento do número de réplicas de cada cenário;
- Aumento do número de árvores analisadas para cada cenário;
- Comparação das topologias usando uma métrica de topologias, para estabelecer estatísticas explanatórias sobre os fenômenos e estruturas;
- Adaptação da metodologia de filodinâmica para sequências reais de TEs de Classe II presentes em mosquitos dos gêneros *Anopheles* e *Aedes*.

A adaptação da metodologia para uso com sequências de bases de dados biológicos públicos consistirá essencialmente em relaxar as hipóteses garantidas nos cenários de simulação. Essas hipóteses simplificadoras incluem: existência de relógio molecular estrito, modelo evolutivo Jukes-Cantor, *prior* da árvore aleatório ao invés de UPGMA. Ao invés desses parâmetros, serão utilizados o relógio molecular relaxado log-normal, modelo evolutivo a ser determinado empiricamente utilizando um programa como `ModelTest` [146] ou `ModelGenerator` [147], e *prior* da árvore aleatória (caso sejam rodadas várias cadeias independentes), ou NJ, caso as distâncias evolutivas envolvidas sejam pequenas ou moderadas.

Com uma introdução mais detalhada, resultados sólidos de simulação e filogenias convincentes de TEs de bases de dados, consideramos que o trabalho poderá ser submetido para uma revista de impacto Qualis A. O periódico alvo nesse momento é o **PLoS Computational Biology**.

A.2.2 Material suplementar (SOM)

Para evitar excesso de descrição da metodologia no texto principal, os detalhes e algumas escolhas de parâmetros serão incluídas em material suplementar, seguindo o modelo do artigo de Struchiner *et al* (2009, [110]).

Detalhes e informações que satisfazem esse critério incluem:

- parâmetros de importância secundária que descrevem os cenários de simulação;
- parâmetros das análises filogenéticas;
- parâmetros da reconstrução demográfica.

Phylodynamics of transposable elements

Felipe Figueiredo¹BCS, Claudio Struchiner²PROCC

June 1, 2012

Abstract

We propose a novel methodology for combining the inference of the dynamics for both population genetics and molecular evolution of Transposable Elements (TEs) based on sequence data. We simulate forward in time an invasion process taking into consideration the inherent evolution that occurs within several generations, and use phylodynamics techniques to reconstruct the backwards process. Although we use techniques developed for virus sequence data against TE sequence data, we argue that to do so some analogies must be made, but reasonable ones. We justify each such analogy made.

Individual Based model — Agent Based Model — simulation — transposition dynamics — population genetics — wright-fished model — finite population — infinite alleles model — infinite-sites model

1 Background

According to the World Health Organization (WHO), malaria is the vector-borne infectious disease with greatest death toll, estimated to kill over 1 million people worldwide each year[1]. In the absence of a vaccine, one of the propositions to eradicate malaria spread is by substitution of the vector population of the mosquito of the genus *Anopheles* by a genetically modified (GM) variant that has a reduced competence in transmitting the pathogen. This can be done by inserting a gene that codes an enzyme that breaks the plasmodium cycle within the mosquito[2–4]. This in turn poses the question of how to push this extraneous gene through the whole wild population, i.e., to effectively substitute the wild population for a variant refractory to the pathogen and thus incompetent as a disease vector.

The creation of genetically modified insects (GMI) has become viable with current technology [5–7]; the introduction of these subjects in the field, however, is a whole other challenge. With major fitness disparity between laboratory strain and their wild counterparts [7–12], it is unlikely that they would succeed in colonizing a given region if it would depend only on the canonical mendelian rules of genetics.

To deal with this difficulty it was proposed the use of Gene Drive Systems that can optimize this fixation process[4] amplifying the number of copies of the

subject gene at each generation. One of the proposed mechanisms is the use of active Transposable Elements (TEs) that could carry nested transgenes, which we model in this work [3, 13–17].

Additional topics related to the substitution of a wild population include the impact virulence [18] and evolution of resistance [19, 20].

1.1 Transposable Elements

Transposable elements are DNA segments that appear in several genomic sites in a host. They occur in several sizes and structures, ranging from small segments flanked by inverted repeats that get copied whenever there's an appropriate enzyme, to autonomous virus-like elements that contain the necessary genes to produce its own transposase or retrotransposase.

They are so common, in fact, that they have been observed in such quantities corresponding to 40-50% of mammals genomes, and almost 90% in some plants [21], and seem to be present inside every genome they have been looked for so far [22]. This is why these elements previously thought to be *junk DNA* have now been promoted to agents of evolutionary force [23–25]. A throughout classification review of TEs is available in [26].

1.2 TEs as Gene Drive systems

Previous simulations indicate autonomous transposable elements should have a too high transposition rate, lowering the overall fitness of the mosquito [11], whereas the use of non-autonomous transposable elements should allow both lower equilibria quantity and rate of transposition.

GMI should be as similar as possible to the indigenous species it is supposed to replace, preferably with the only perceptible difference being the refractoriness to the disease pathogen. But genetic modification is known to lower the overall fitness of mosquitos, as indicated in laboratory studies. In fact, most TE invasions are believed to cause negative reproductive impact and/or induce mutation in every species observed [7, 8, 10, 12, 27]. Even so, these elements are now highly accepted as key agents in species' evolutionary capacity, and genome organization [21, 25, 28, 29].

1.3 Modeling by simulation

Since an analytical-driven framework consisting of only mathematical models, considering all the nested agents and processes in play is impractical, the next best thing is to model empirically by simulating all the players in the three levels of biological organization.

To this end we employ an Individual-based Model (TRapid, to be published), that represents independently each individual in the population, each TE per chromosome and each TE's sequence. This approach frees the researchers of dealing with many models simultaneously, while still providing a rich environment for the representation of most simplistic models. Complexity can be added

gradually so simulation results can be compared hierachically with standard statistics techniques.

Although the comparison of complex models with simple scenarios is not always straightforward, we were still able to achieve the same level of qualitative conclusions for TEs that typical phylodynamics analysis do for virus sequence data.

1.4 Analogies to virus phylodynamics

One of the major goals of a phylodinamics analysis is to estimate the effective number of the population (N_e), which is an ubiquitous value in the Population Genetics area and can be used to estimate the basic reproductive ratio (R_0), which is similarly important in epidemiology.

There is a clear biological analogy between TEs and viruses (FIXME: insert citation). The step that needs to be addressed is the usage of coalescence-based techniques for TE sequence data. Which populations are being considered for the dynamical system and ... (FIXME)?

Sequence data widely available in public sequence databases are being largely overlooked for their potential and hidden information. We hope the evidence we provide here will be compelling enough for future works to assess TE invasions in either natural or laboratory cage populations will include phylodynamics techniques in order to extract more information and knowledge from data (FIXME: clarify).

...
FIXME: cite [30-32]
[22]

2 Conclusions

We conducted several experiments to observe the simulated molecular evolution of an invading TE family in a host population.

We observe that Master Gene simulations produce comb-like phylogenies, as proposed by [33].

Further analysis of experiment A indicates the importance of a high transposition rate at the beginning of the invasion. Low transposition rates are not sufficient to maintain TE presence in the population, even in the absence of a fitness toll per element. Similar results were achieved by [11]. Observing 30 generations with a constant transposition rate of 1 new copy per reproductive event rarely produces an invasion scenario as seen in Fig 1. This parameter set gives a very low success rate in invasion, and is indistinguishable from random fixation due to genetic drift.

On the other hand, while a single new TE copie per generation is not enough for a transposition rate, experiment A2 shows that 5 new copies per gametogenesis accumulates enough copies in few generations to achive a successful invasion (Fig 2). This is a simplistic transposition model, with no self-regulation and can

produce unrealistic quantities of TE copies per individual if enough generations are permitted to pass.

...

2.1 Experiment A

2.2 Experiment B

Coalescent-based demographic inference indicate that the host population growth or decrease may have a subtle but perceptible impact on the

We used TRepid (version XXX), to conduct several experiments. Each experiment consisted of several parameter sets for simulations, and each parameter set was run 5-10 times, in order to observe the frequency of the phenomena for that set.

For the first experiments we used the simplest models available in the simulator: a small constant population for the ecological model and constant transposition. For the molecular evolution component we used a model that doesn't take any particular premises about site-specific variability, so it's regarded as a jukes-Cantor model[34]. Each parameter set was run several times, and the proportion of success rate of invasion is calculated

Each simulation experiment can stop at one of the following conditions: (a) the proportion of the population that has at least one TE reaches 90% or more, (b) the quantity of individuals that have TEs drops to zero, (c) the maximum number of generations have passed.

3 Simulations

3.1 Experiment A

For experiment A we simulated a Master Gene (MG) amplification for a total of 5 generations on a saturated logistic population. The F_0 population was entirely identical to a template, containing a single active TE. At each generation a single additional inactive TE was included per host, and no excision was allowed. Template sequences for the active element are formed by aprox. 200 identical characters (sequence = TTTTTT...), and since we're simulating a Class I TE (retrotransposase-based TE [26]), the copy and paste transposition method is used as follows. The original active copy is preserved during transposition (as oposed to the cut and paste method, assumed for Class II DNA transposase-based TEs). Mutations are inserted at random in two events: in the beginning of gametogenesis one mutation is introduced at transposition for each new copy (and not for the original copy), and at the end of gametogenesis, all existing copies receive an extra mutation to induce an aging pattern to be observable afterwards in the phylogenetic tree. This means that at the end of the gametogenesis, the active TE will have one random mutation, while the new inactive TE will have two, and so on. New TE sequences are created with informative headers that represent their creation history and age, and all headers

are updated to reflect current age (time is measured in terms of generations, or simulation steps).

Since every mutation inserted is uniformly sampled along sites, and there is no codon partition or any other intrinsic structure for the sequences, a distance based phylogenetic approach by definition should provide an adequate prior for the tree topology (save for random substitutions on sites where there has been a previous mutation, in which case the phylogenetic signal is diminished/lost). A corollary of this construction is that the *taxa* represented by these sequences satisfy the molecular clock hypothesis. To test this hypothesis, we reconstructed a Neighbor-Joining tree using ClustalW[35] and compared this to the tree created with a Bayesian package (BEAST)[36], assuming the strict clock model and an UPGMA tree as a tree prior, while co-estimating a non-parametric demographic model using bayesian skyline.

3.2 Experiment B

For experiment B, we also simulate a Master Gene transposition model very similar to experiment A, but this time we test it against several different host population scenarios. For each ecological model we conducted a maximum of 100 generations of sexual diploid population, and with a linear transposition model of constant growth, of 5 new copies per reproductive event. The ecological models: constant population, exponential growth and exponential decay.

We will compare the inference for the demographic component using two different inference approaches. The

“BEAST outputs a smooth estimate of the demographic function using the Bayesian skyline method [37], which does take the phylogenetic error into account but requires that the amount of smoothing be fixed a priori. A related method developed by [38] relaxes this requirement; however, this approach takes the phylogeny tree as fixed and, unlike the Bayesian skyline plot, does not take phylogenetic error into account.”

– text of acknowledgments here, including grant info –
Grant from CAPES number XXX

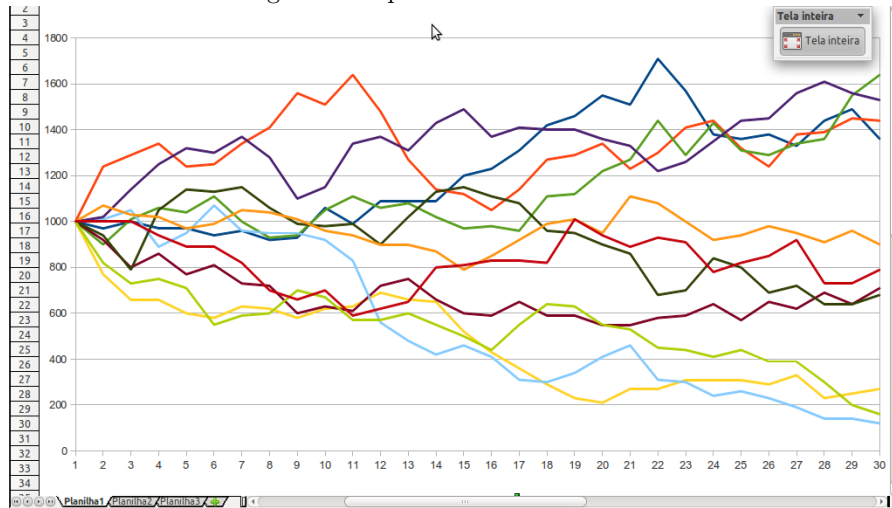
References

- [1] World Health Organization (2005) WHO roll back malaria report., Technical report.
- [2] Hemingway J, Craig A (2004) Parasitology. New ways to control malaria. *Science* 303:1984–5.
- [3] Christophides GK (2005) Transgenic mosquitoes and malaria transmission. *Cell Microbiol* 7:325–33.
- [4] James AA (2005) Gene drive systems in mosquitoes: rules of the road. *Trends Parasitol* 21:64–7.

- [5] Catteruccia F, et al. (2000) Stable germline transformation of the malaria mosquito *Anopheles stephensi*. *Nature* 405:959–62.
- [6] Ito J, Ghosh A, Moreira LA, Wimmer EA, Jacobs-Lorena M (2002) Transgenic anopheline mosquitoes impaired in transmission of a malaria parasite. *Nature* 417:452–5.
- [7] Moreira LA, Wang J, Collins FH, Jacobs-Lorena M (2004) Fitness of anopheline mosquitoes expressing transgenes that inhibit *Plasmodium* development. *Genetics* 166:1337–41.
- [8] Catteruccia F, Godfray HC, Crisanti A (2003) Impact of genetic manipulation on the fitness of *Anopheles stephensi* mosquitoes. *Science* 299:1225–7.
- [9] Riehle MA, Srinivasan P, Moreira CK, Jacobs-Lorena M (2003) Towards genetic manipulation of wild mosquito populations to combat malaria: advances and challenges. *J Exp Biol* 206:3809–16.
- [10] Irvin N, Hoddle MS, O’Brochta DA, Carey B, Atkinson PW (2004) Assessing fitness costs for transgenic *Aedes aegypti* expressing the GFP marker and transposase genes. *Proc Natl Acad Sci U S A* 101:891–6.
- [11] Le Rouzic A, Capy P (2006) Reversible introduction of transgenes in natural populations of insects. *Insect Mol Biol* 15:227–34.
- [12] Huho BJ, et al. (2007) Nature beats nurture: a case study of the physiological fitness of free-living and laboratory-reared male *Anopheles gambiae* s.l. *J Exp Biol* 210:2939–47.
- [13] Kidwell MG, Ribeiro JM (1992) Can transposable elements be used to drive disease refractoriness genes into vector populations? *Parasitol Today* 8:325–9.
- [14] Carareto CM, et al. (1997) Testing transposable elements as genetic drive mechanisms using *Drosophila P* element constructs as a model system. *Genetica* 101:13–33.
- [15] Ribeiro JM, Kidwell MG (1994) Transposable elements as population drive mechanisms: specification of critical parameter values. *J Med Entomol* 31:10–6.
- [16] Rasgon JL, Gould F (2005) Transposable element insertion location bias and the dynamics of gene drive in mosquito populations. *Insect Mol Biol* 14:493–500.
- [17] Medstrand P, et al. (2005) Impact of transposable elements on the evolution of mammalian gene regulation. *Cytogenet Genome Res* 110:342–52.
- [18] Medlock J, Luz PM, Struchiner CJ, Galvani AP (2009) The impact of transgenic mosquitoes on dengue virulence to humans and mosquitoes. *Am Nat* 174:565–77.

- [19] Boete C, Koella JC (2002) A theoretical approach to predicting the success of genetic manipulation of malaria mosquitoes in malaria control. *Malar J* 1:3.
- [20] Boete C, Koella JC (2003) Evolutionary ideas about genetically manipulated mosquitoes and malaria control. *Trends Parasitol* 19:32–8.
- [21] Kazazian, Jr HH (2004) Mobile elements: drivers of genome evolution. *Science* 303:1626–32.
- [22] Batzer MA, Deininger PL (2002) Alu repeats and human genomic diversity. *Nat Rev Genet* 3:370–9.
- [23] Biemont C, et al. (2003) Worldwide distribution of transposable element copy number in natural populations of *Drosophila simulans*. *Evolution* 57:159–67.
- [24] Le Rouzic A, Capy P (2005) The first steps of transposable elements invasion: parasitic strategy vs. genetic drift. *Genetics* 169:1033–43.
- [25] Gotea V, Makalowski W (2006) Do transposable elements really contribute to proteomes? *Trends Genet* 22:260–7.
- [26] Wicker T (2007) A unified classification system for eukaryotic transposable elements. *Nature* 8.
- [27] Cha SJ, Mori A, Chadee DD, Severson DW (2006) Cage trials using an endogenous meiotic drive gene in the mosquito *Aedes aegypti* to promote population replacement. *Am J Trop Med Hyg* 74:62–8.
- [28] Kidwell MG, Lisch DR (2001) Perspective: transposable elements, parasitic DNA, and genome evolution. *Evolution Int J Org Evolution* 55:1–24.
- [29] Kidwell MG (2002) Transposable elements and the evolution of genome size in eukaryotes. *Genetica* 115:49–63.
- [30] Volz EM, Kosakovsky Pond SL, Ward MJ, Leigh Brown AJ, Frost SD (2009) Phylodynamics of infectious disease epidemics. *Genetics* 183:1421–30.
- [31] Bennett SN, et al. (2010) Epidemic dynamics revealed in dengue evolution. *Mol Biol Evol* 27:811–8.
- [32] Allicock OM, et al. (2012) Phylogeography and Population Dynamics of Dengue Viruses in the Americas. *Mol Biol Evol*.
- [33] Brookfield JF, Johnson LJ (2006) The evolution of mobile DNAs: when will transposons create phylogenies that look as if there is a master gene? *Genetics* 173:1115–23.

Figure 1: Experiment A - no invasion



- [34] Jukes TH, Cantor CR (1969) in *Mammalian protein metabolism*, ed Munro HN (Academic Press, New York), pp 21–123.
 - [35] Thompson JD, Gibson TJ, Higgins DG (2002) Multiple sequence alignment using ClustalW and ClustalX. *Curr Protoc Bioinformatics* Chapter 2:Unit 2.3.
 - [36] Drummond AJ, Rambaut A (2007) BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol Biol* 7:214.
 - [37] Drummond AJ, Rambaut A, Shapiro B, Pybus OG (2005) Bayesian coalescent inference of past population dynamics from molecular sequences. *Mol Biol Evol* 22:1185–92.
 - [38] Opgen-Rhein R, Fahrmeir L, Strimmer K (2005) Inference of demographic history from genealogical trees using reversible jump Markov chain Monte Carlo. *BMC Evol Biol* 5:6.
 - [39] Paradis E, Claude J, Strimmer K (2004) APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* 20:289–90.
- FIXME: redo figures (seriously)

Figure 4: Experiment A - Demographic component

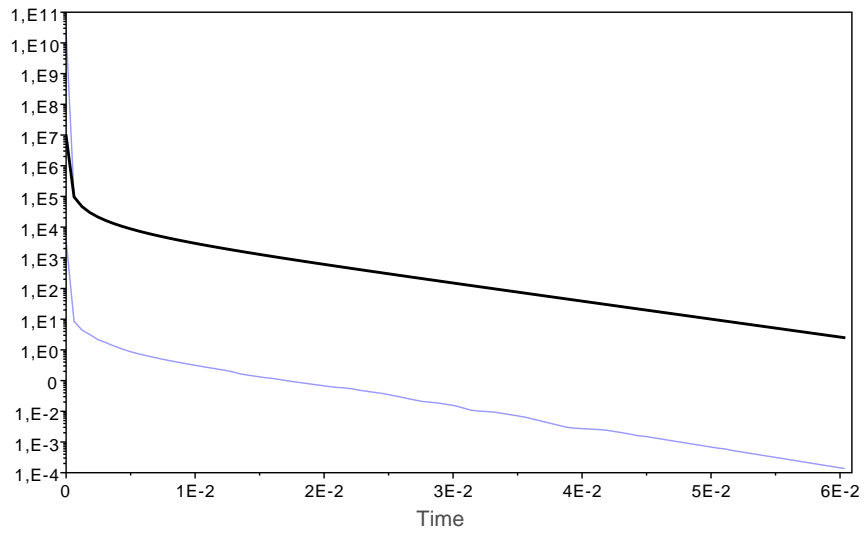


Figure 5: Experiment B - High success in invasion rate

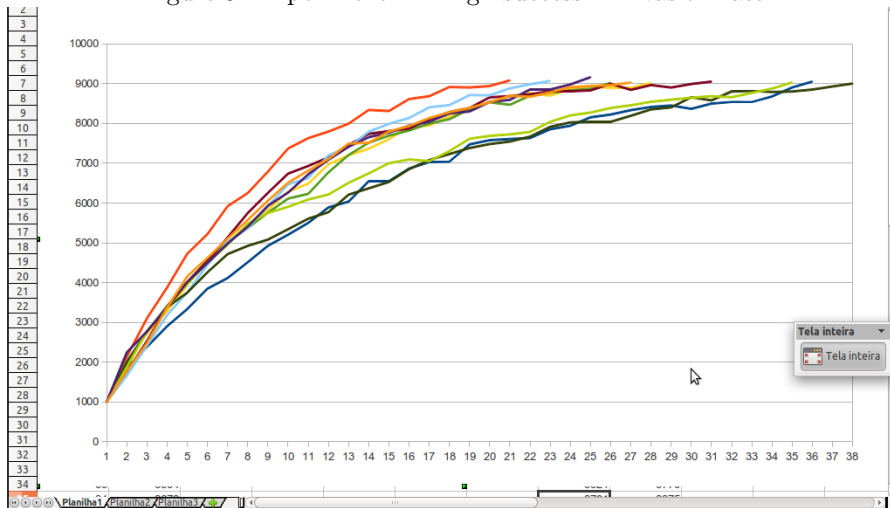
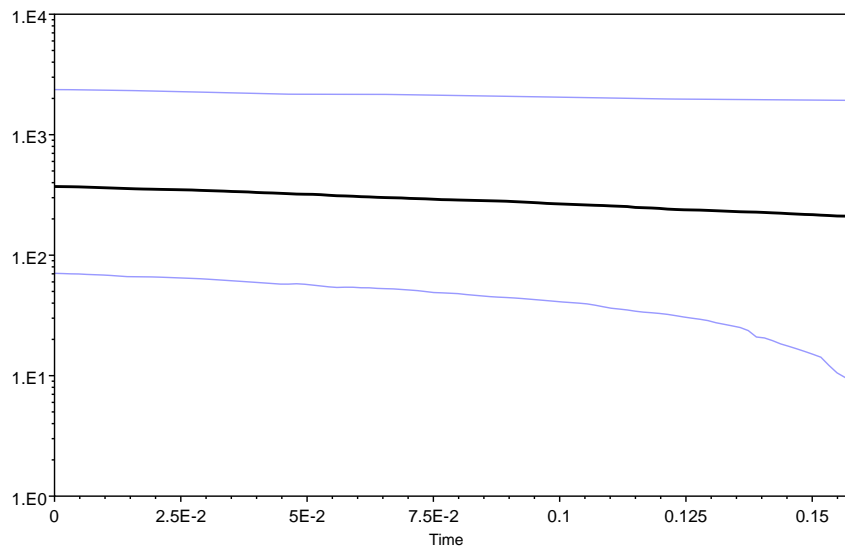


Figure 6: Experiment B - Demographic component



Índice Remissivo

- árvore
 - bipectinada, 70, 71
 - pectinada, 50, 64, 68, 70, 117
- ABMs, 26
- amostragem, 47
- backward simulation, 17
- BEAST, 62, 64
- BI, 51, 62, 64
- Cenário de simulação, 38, 46–48, 52, 55, 61, 63
- Coalescência, 15, 18, 21
- DDT, 2
- ESS, 62, 65
- Evento de transposição, 36–39, 43, 50, 52
- Evento evolutivo, 35, 36, 41, 42, 46, 50, 63
- Excisão de TEs, 39
- Experimentos
 - A, 48, 51, 58, 78, 118
 - B, 48, 54, 64, 65, 104, 120
 - MG, 48, 50, 64, 68, 117
- Filodinâmica, 16, 21, 54, 62, 65, 120
- Filogenética, 13, 62, 64
- Fitness, 4, 5, 31, 34, 44, 46, 79
- FK, 5
- forward simulation, 17, 26
- Gene Drive Systems, 5
- genetic drift, 19
- GM, 58
- GMI, 3
- HERV, 12
- IBMs, 19, 21, 26, 27, 38
- Invasão, 118
- IRS, 2
- Malária, 2
- Master Gene, 46, 50, 54, 115, 122, 123
- MCC, 66
- MCMC, 62–65, 115
- MEDEA, 5
- MG, 8
- Modelo Baseado em Indivíduos, *veja* IBMs
- MRCA, 15, 47
- Neighbor-Joining, 14, 51, 62, 64, 73
- NJ, *veja* Neighbor-Joining
- random mating, 19, 28, 29, 99
- Relógio molecular
 - Estrito, 13, 63
 - Relaxado, 13
- SIT, 3
- Skyline plot, 15, 21, 63, 65, 110, 112, 120, 121
- TEs, 5, 6, 21, 34, 42, 44, 59, 122
 - Classe I, 6, 35, 123
 - Classe II, 6, 36, 123
- Transferência horizontal, 8, 9
- Transposon (modelo), 50, 122
- TRepid, 19, 26, 29, 30, 34, 35, 38, 39, 46

174

ÍNDICE REMISSIVO

UPGMA, 14