EXPLOITING NATURAL SELECTION TO STUDY ADAPTIVE BEHAVIOR

Sergio Pulido Tamayo

Promotor:
Kathleen Marchal, UGent
Co-promotor:
Jan Fostier, UGent

Promotor:
Prof. Jos Vanderleyden, KULeuven

Examination Committee:
Prof. Luc De Raedt, KULeuven
Prof. Yves Van de Peer
Prof. Wim Vanden Berghe
Prof. Tom Dhaene
Prof. Peter Dawyndt

Dissertation presented in partial fulfilment of the requirements for the degree of PhD in Bioscience engineering

SUMMARY

The research presented in this dissertation explores different computational and modeling techniques that combined with predictions from evolution by natural selection leads to the analysis of the adaptive behavior of populations under selective pressure.

For this thesis three computational methods were developed: EXPLoRA, EVORhA and SSA-ME. EXPLoRA finds genomic regions associated with a trait of interests (QTL) by explicitly modeling the expected linkage disequilibrium of a population of sergeants under selection. Data from BSA experiments was analyzed to find genomic loci associated with ethanol tolerance. EVORhA explores the interplay between driving and hitchhiking mutations during evolution to reconstruct the subpopulation structure of clonal bacterial populations based on deep sequencing data. Data from mixed infections and evolution experiments of E. Coli was used and their population structure reconstructed. SSA-ME uses mutual exclusivity in cancer prioritize cancer driver genes. TCGA data of breast cancer tumor samples were analyzed.

SAMENVATTING

Dit onderzoek handelt over verschillende computationele technieken die, samen met voorspellingen gebaseerd op evolutie door natuurlijke selectie, leiden tot de analyse van adaptief gedrag van populaties onder selectiedruk.

In het kader van deze thesis werden drie computationele methodes ontwikkeld: EXPLoRA, EVORhA en SSA-ME. EXPLoRA vindt genomische regio's die geassocieerd zijn met bepaalde interessante eigenschappen (QTL) door expliciet de verwachtte "linkage disequilibrium" van een populatie segreganten onder selectie te modeleren. Data van BSA experimenten werd geanalyseerd om de genomische loci die geassocieerd zijn met ethanol tolerantie te vinden. EVORhA onderzoekt hoe causale ("driver") mutaties en niet-causale ("hitchhiking") mutaties in een bepaald fenotype zich tot elkaar verhouden tijdens evolutie om zo de substructuur van een klonale bacteriële populatie, gebaseerd op deep sequencing data, te reconstrueren. Data van gemengde infectie experimenten en van evolutie experimenten voor *E. coli* werd hier gebruikt. SSA-ME gebruikt het concept van mutuele exclusiviteit in kanker om causale ("driver") genen te prioriteren. Hier werd borstkanker data uit TCGA (The Cancer Genome Atlas) geanalyseerd.

ABBREVIATIONS

QTL	Quantitative Trait Loci
BSA	Bulk Segregant Analysis
SSA	Small Subnetwork Analysis
CADD	Combined Annotation Dependent Depletion
ME	Mutual Exclusivity
HMM	Hidden Markov Model
LD	Linkage Disequilibrium
SNP	Single Point Nucleotide
SNV	Single Nucleotide Variant
CNV	Copy Number Variation
FDR	False Discovery Rate
PCR	Polymerase Chain Reaction
YDP	yeast extract peptone dextrose
BED	Browser Extensible Data
VCF	Variant Call Format
NGS	Next Generation Sequencing
ORM	Object-Relational Mapping
BLOSUM	Blocks Substitution Matrix
MAE	Mean Absolute Error
RMSE	Mean Squared Error
BWA	Burrows-Wheeler Aligner
TCGA	The Cancer Genome Atlas
ICGC	International Cancer Genome Consortium
MES	Mutual Exclusivity Score
ROC	Receiver Operating Characteristic
AUC	Area Under the Curve
GO	Gene Ontology
PPV	Positive Predictive Value
CGC	Cancer Gene Census
NCG	Network of Cancer Genes
BRCA	Breast Cancer invasive carcinoma
eQTL	Expression QTL
pQTL	Protein level QTL
DNA	Deoxyribonucleic acid
RNA	Ribonucleic acid
NAR	Nucleic Acids Research

TABLE OF CONTENTS

Summary i

Samenvatting ii

Abbreviations iii

Table of Contents v

Table of Figures ix

List of Tables xi

Chaper 1. Introduction 1

Dynamics of natural selection 2
Natural selection in sexual populations 2
Natural selection in clonal populations 4

Outline of the thesis 6

Chaper 2. Linkage analysis of Quantitative Trait Loci using bulk segregants 9

Abstract 9

Introduction 10

Materials and Methods 14
Simulated data 14
Performance analysis 15
Comparison with state-of-the-art 15
Real dataset 16
Experimental validation 17

Results *18*
Development of EXPLoRA, a HMM for the analysis of BSA data 18
Parameter sensitivity of EXPLoRA 22
Comparison with state of the art 27
Application of EXPLoRA to real datasets 30

Web Server *36*
EXPLoRA web 36

Duscussion *41*

Conclusion **42**

Chaper 3. Frequency-based haplotype reconstruction from deep sequencing data of bacterial populations **43**

Abstract *43*

Introduction *44*

Material and Methods *46*
EVORhA 46
Simulation experiments 58
Haplotype reconstruction to infer evolutionary trajectories 60
Haplotype reconstruction to identify mixed infections 61

Results *62*
Performance of EVORhA on simulated data 63
Comparison of EVORhA with state-of-the-art haplotype reconstruction 67
Haplotype reconstruction to reconstruct evolutionary trajectories 70
Haplotype reconstruction to identify mixed infections 73

Discussion *75*

Chaper 4. Detection of cancer driving genes using mutual exclusivity **79**

Abstract 79

Introduction 79

Materials and Methods 82
SSA-ME 82
Simulated data 86
Breast Cancer TCGA Data 88

Results 89
SSA-ME Implementation 89
Performance on simulated data 90
Analysis of the TCGA breast cancer data 94
Comparison with TCGA Analysis 98

Discussion 102

Chaper 5. Conclusions and perspectives 105

Conclusions 105

Perspectives 107

Future work 108
Integration of EXPLoRA-web with network based analyses 108
Evolutionary trajectories using time resolution 109
Using SSA for more complex tasks 109

Curriculum 111

Appendix A. Supplementary data Chapter 2 113

Appendix B. Supplementary data Chapter 3 115

Appendix C. Supplementary data Chapter 4 121

Literature based evidence for newly predicted cancer drivers 121

Independent validation of the predicted drivers 122

References **129**

TABLE OF FIGURES

Figure 2.1 Bulk segregant analysis for mapping genomic regions linked to a phenotype of interest in yeast. 11
Figure 2.2 Hidden Markov Model used to predict genomic regions linked to the phenotype of interest. 20
Figure 2.3 Effect of the recombination rate (r) on the performance of EXPLoRA. ... 23
Figure 2.4 Effect of α_P/β_P on the performance of EXPLoRA. 25
Figure 2.5 Comparison with the state-of-the-art. 28
Figure 2.6 Linkage scores obtained by EXPLoRA for the five QTLs identified in the 16% pool (left) and in the 17% pool (right). 33
Figure 2.7 Experimental validation of QTL2 on chromosome X. 35
Figure 2.8 Overview of the web service. 38
Figure 3.1 Method Overview. 47
Figure 3.2 Reliability of haplotype reconstruction by EVORhA on simulated data. . 66
Figure 3.3 Performance comparison of EVORhA, ShoRAH and QuasiRecomb on simulated data. 69
Figure 3.4 Haplotype reconstruction to infer evolutionary trajectories. 71
Figure 3.5 Haplotype reconstruction to identify mixed infections. 74
Figure 4.1 Pseudocode of SSA.ME algorithm 83
Figure 4.2 Calculation of *MES* and corresponding *rMES* scores for three different small subnetworks. 85
Figure 4.3 Overview of SSA-ME. 91
Figure 4.4 Performance on Simulated Data. 93
Figure 4.5 Application of SSA.ME on TGCA Breast Cancer dataset. 95
Figure 4.6 CADD scores analysis of selected genes. 97
Figure 4.7 Comparison between SSA.ME and MEMo. 99

Supp Figure A. 1 Average scaled linkage score at the causal site reported by the method of Magwene et al. as a function of the coverage and noise levels... 113
Supp Figure A. 2 Comparison with the state-of-the-art. 114

Supp Figure B. 1 Inferring template haplotypes and performing error correction at the local scale. 115
Supp Figure B. 2 Window extension procedure 117
Supp Figure B. 3 Figure 3S. Frequency analysis 119

Supp Figure C. 1 Complete ROC Curve 127
Supp Figure C. 2 New Predictions mutations 128

LIST OF TABLES

Table 1 Overview of the 49 prioritized drivers and their previous associations with cancer: Columns indicate true positive sets (CGC, Malacards or NCG). 124

Table 2 CADD scores of mutations in each of the 9 predicted drivers. CRK and TK1 are excluded as they did not carry any mutations and MCL1 is excluded as it only contains CNVs. 126

Chaper 1. INTRODUCTION

The observation of natural selection goes a long way back, from the time of the Origin of the Species to observe the effects of natural selection on a population have been an important step towards better scientific understanding. One of the best known cases of observing natural selection is that of the industrial melanisms in the peppered moth (*Betularia F. Typical*) (Majerus 2008). The peppered moth sleeps by day and its coloring is an important phenotype to survive predators that want to eat them. It wasn't until 1950's that several experiments were conducted to observe how selection affected the population of moths. In the non-polluted, lichen filled trees, the common peppered moth thrived thanks to the hiding provided by its whitish coloring. On the contrary, in industrial areas where trees were blackened by pollution, the traditionally colored moth was an easy target for predators and the individuals with dark coloring survived. Following in the same line of experiments, fifty years ago the research to observe natural selection was focused on phenotypic traits easy to evaluate. Research at the time was mainly focused on animal color patterns, mimicry and distastefulness (Endler 1983). Another famous example of this was the characterization of the guppy fish (*Poecilia Reticulata*) colorful skin, and the experiments to explain how two conflicting selective pressures, that of being more attractive to the guppy females (i.e. more colorful) and that of being a harder target for predators (i.e. less colorful) played a role on determining the color phenotype of the populations of guppies.

Today, thanks to the recent advancements and wide availability of sequencing data it is possible to observe the effects of natural selection at the genomic level with unpresented resolution.

DYNAMICS OF NATURAL SELECTION

To be able to observe the marks left by natural selection it is necessary to first understand the possible effects it has on the genome and how it is affected. The tenets of natural selection are simple: there is variation in traits, there is differential reproduction and, last but not least, these traits are inherited. These three conditions can occur in countless ways and the effects depend on the nature of the traits and how reproduction followed.

These variables create different dynamics for different organisms. Are the traits being selected advantageous for the individual? Or on the other hand, are they problematic for the current environment and getting rid of them will be an advantage to the individual? Is sexual reproduction, and therefore recombination, available to remove useless or deleterious mutations? Or are offspring stuck with all mutations of its ancestors independent of their selective advantage? Was there genotypic variation already available on the population? Or, on the contrary, the original population consisted of clones and variation needed to wait for new mutations to arise? We will discuss the implications of these variables and what marks are left in the genome in the rest of this section.

NATURAL SELECTION IN SEXUAL POPULATIONS

Sexual reproduction involve two parents contributing one gamete each. Gametes are produced during meiosis: chromosomes are duplicated and then the cells are divided twice, ending in sexual cells with half the number of chromosomes of a normal cell. One key feature in this process is recombination or crossover.

Recombination is a fundamental source of genomic variation. It has a huge impact on the lifecycle of mutations. Recombination cause the offspring to inherit a completely new combination of parental DNA. From the point of view of the dynamics of evolution this allows the beneficial mutations to be selected

independently of the rest of the ancestor's mutations to some degree. The independence of genes is not complete because the frequency of crossover events between two mutations is proportional to their physical distance in the chromosomes. Mutations close to each other are less likely to suffer a crossover event that separate them.

The degree to which mutations are separated is related to the recombination rate (i.e. the average number of recombination events that happen in a chromosome). The more recombination events in a chromosome, the easier it becomes for a beneficial mutation to be selected alone, just for the fitness advantage it confers. But this also implies that mutations in close vicinity tend to be inherited together, and therefore "linked". The link between two mutations is known as linkage disequilibrium.

When observing the genomes of descendants that are selected for the traits of their parents, being the degree of melanism in the moths, the colorful spots of the guppy fish or any other trait, the frequencies of the mutations responsible for the phenotype under selection start to become more frequent in the population (Barrick & Lenski 2013). This can happen for many reasons, e.g. those with the mutation are more likely to survive (like the dark moths were less likely to be eaten by predators) or reproduce more (like the guppy with flashier spots can court more females), but independent of the reason behind the selection the effect is that the causal mutation become more common in the population. Interestingly, this does not only happen to the responsible mutation, it also happen to those mutation in close proximity to the causal mutation. The closer a neutral mutation is to the causal mutation the less likely is that a recombination event happen in the area. If a neutral mutation is in very close proximity to the causal mutation, the chance that a recombination event happens in the region between them becomes very low.

Natural selection in clonal populations

Clonal populations do not have the advantages given by sex. Recombination is not readily available and once a mutation occurs in an individual, all of its descendants will possess that same mutation. Given this constrain, the fate of a mutation will depend on the survival of the individuals and not only on the fitness advantages given by the mutation *per se*. A neutral mutation could be fixated in the population if it occurs in the same individual as a highly advantageous mutation. Most mutations are neutral and do not confer any advantage or disadvantage to the individual. They occur and stay on the individuals and its descendants. The mutations that increase in frequency in the population without conferring a fitness advantage are known as hitchhikers or passenger mutations. While the ones conferring the advantage are called driver mutations.

Imagine a simple scenario of clonally reproducing cells where beneficial mutations rarely occur. Imagine now that one of those rare beneficial mutation just occurred in an individual and that it, for example, allows the cell to obtain more nutrients from the environment. The mutation (assuming it survive being removed by genetic drift) will provide an advantage to the individual and its descendants. They will get more nutrients and outcompete the rest of the population in a short time. As beneficial mutations rarely occur, the competitors will not be able to compete and will disappear while the beneficial mutation gets to fixation. After this selection sweep, all the population will be descendants of the original individual and all will have the same fitness. Until sometime afterwards a new beneficial mutation occurs and the process repeats itself. This scenario is called *periodic selection* and constitutes a set of selective sweeps one after another (ATWOOD et al. 1951).

However, real life rarely is as simple as the scenario described before. In a more complex scenario, additional beneficial mutations can occur in other individuals before the mutation is fixated in the population. Individuals

possessing two different beneficial mutations will compete between them, taking a much longer time for any of the mutations to reach fixation. When this competition between subpopulations containing different beneficial mutations happens, it is known as clonal interference.

CANCER

Cancer is a very complex disease. It evolves by a repeated process of clonal expansion in which the genotype and population structure change during the process. The disease is very diverse, it can evolve in less than a year or over several, it can present few mutations or over a thousand. Each cancer is different and have followed a unique evolutionary trajectory.

As a clonal evolutionary process similar to those described above, with most mutations being neutral, many of the mutations encountered in the population of cancerous cells are expected to be passenger mutations. This increases the difficulty of finding the few driver mutations in a sea of hitchhikers. But one special feature of (some) cancers is that for them to be successful they need to disrupt cellular processes or pathways controlling mechanisms that could empower invasion, self-renewal and growth (Greaves & Maley 2012). The disruption of these cellular processes requires no more than one well-placed mutation in one of the components that make up the complex cell machinery necessary to fulfill them.

The interplay of this cancer feature and the unique evolutionary paths of each tumor create a very interesting scenario: two different cancers can disrupt the same pathway at different genes to get the same advantageous phenotype. And, any additional mutation in the same pathway will be redundant, and therefore will not confer any further evolutionary advantage. When several tumors are observed simultaneously, a mutual exclusivity pattern form in these pathways (Yuan & Cantley 2008; Vandin et al. 2012; Babur et al. 2014; Ciriello et al. 2012; Pulido-Tamayo et al. 2015).

OUTLINE OF THE THESIS

This thesis outline three methods developed to interpret DNA sequencing data and observe natural selection in action at the genomic level.

The first part of the thesis deals with finding Quantitative Trait Loci (QTL) using sequencing data obtained by Bulk Segregant Analysis (BSA). In this research we used the expected outcome of selecting a population of descendants (segregants) crossed from a superior and inferior parent in ethanol tolerance to determine the mutations that give the superior parent its phenotype. The developed tool uses the information intrinsic to linkage disequilibrium to accomplish the task at hand. This tool was published as Jorge Duitama, Aminael Sánchez-Rodríguez, Annelies Goovaerts, Sergio Pulido-Tamayo, Georg Hubmann, María R Foulquié-Moreno, Johan M Thevelein author, Kevin J Verstrepen author and Kathleen Marchal (2014) "Improved linkage analysis of Quantitative Trait Loci using bulk segregants unveils a novel determinant of high ethanol tolerance in yeast". *BMC Genomics*. Additionally, the tool was improved to be provided as a web server available at http://bioinformatics.intec.ugent.be/explora-web/ and accepted for the NAR Web Server Issue 2016 as Sergio Pulido-Tamayo, Jorge Duitama and Kathleen Marchal "EXPLoRA-web: linkage analysis of Quantitative Trait Loci using bulk segregant analysis".

The second part of the thesis explores the DNA sequencing data of clonal evolution experiments in bacteria for ethanol tolerance. In it we study and track the evolution of a population of E. coli from 5% ethanol tolerance to 8.5%. We developed EVORhA a tool to reconstruct haplotypes, or the genome of subpopulations, from deep sequencing data at specific time points. We reconstructed the haplotypes at three different time points and used the reconstruction to build the evolutionary trajectory of the population. This work was published as Sergio Pulido-Tamayo, Aminael Sánchez-Rodríguez, Toon Swings, Bram Van den Bergh, Akanksha Dubey, Hans Steenackers, Jan Michiels,

Jan Fostier and Kathleen Marchal (2015) Frequency-based haplotype reconstruction from deep sequencing data of bacterial populations. *Nucleic Acids Research*.

The third part of the thesis make use the mutual exclusivity expected by the parallel evolution of different patient tumors to search for cancer driver genes. In this research we proposed a new framework to tackle large search space computational problems using biological networks called Small Subnetwork Analysis and we made an application by using mutual exclusivity. We searched for driver genes in breast cancer data, and we were able to find the usual suspects driving cancer genes plus some new genes with few genomic alterations previously not associated with cancer. Genes with few alterations are normally not found in driver analyses because the lack of enough mutations to make any decision on mutation clustering. The framework and mutual exclusivity application was submitted to Nature Scientific Reports as Sergio Pulido-Tamayo, Bram Weytjens, Dries De Maeyer, Kathleen Marchal (2016) SSA-ME Detection of cancer mutual exclusivity patterns by small subnetwork analysis.

Chaper 2. LINKAGE ANALYSIS OF QUANTITATIVE TRAIT LOCI USING BULK SEGREGANTS

This chapter bundle two manuscripts. The first manuscript is Duitama, J. et al., 2014. Improved linkage analysis of Quantitative Trait Loci using bulk segregants unveils a novel determinant of high ethanol tolerance in yeast. *BMC genomics* and Pulido-Tamayo, S. et al. 2016. Improved linkage analysis of Quantitative Trait Loci using bulk segregants unveils a novel determinant of high ethanol tolerance in yeast. *Nucleic acids research*. The first manuscript is a methodological paper that describe a HMM to analyse BSA experiments. My contribution to it revolved around testing the method performance and paremeter sensitivity in simulated data. And benchmarking it versus state-of-the-art competitors. The sections not related to real data are of my authorship, including Figures 2.3, 2.4 and 2.5. The second manuscript is Pulido-Tamayo, S., et al. 2016. Improved linkage analysis of Quantitative Trait Loci using bulk segregants unveils a novel determinant of high ethanol tolerance in yeast (submitted). In it I developed a web service to improve the accessibility and usability of the original method.

ABSTRACT

Bulk segregant analysis (BSA) coupled to high throughput sequencing is a powerful method to map genomic regions related with phenotypes of interest. It relies on crossing two parents, one inferior and one superior for a trait of interest. Segregants displaying the trait of the superior parent are pooled, the DNA extracted and sequenced. Genomic regions linked to the trait of interest are identified by searching the pool for overrepresented alleles that normally originate from the superior parent. BSA data analysis is non-trivial due to sequencing, alignment and screening errors. To increase the power of this technology and obtain a better distinction between spuriously and truly linked

regions, we developed EXPLoRA (EXtraction of over-rePresented aLleles in BSA), an algorithm for BSA data analysis that explicitly models the dependency between neighboring marker sites by exploiting the properties of linkage disequilibrium through a Hidden Markov Model (HMM). Comparing the performance of EXPLoRA with an existing method showed that EXPLoRA has a performance at least as good as the state-of-the-art and that it is robust even at low signal to noise ratio's i.e. when the true linkage signal is diluted by sampling, screening errors or when few segregants are available. Reanalyzing a BSA dataset for high ethanol tolerance in yeast with EXPLoRA allowed reliably identifying QTLs linked to ethanol tolerance that could not be identified with statistical significance in the original study. Experimental validation of one of the least pronounced linked regions, by identifying its causative gene *VPS70*, confirmed the potential of our method.

INTRODUCTION

Bulk segregant analysis (BSA) is an elegant method that allows simultaneous identification of genetic loci that contribute to a specific trait or phenotype (for a review see Liti and Schacherer (Liti & Schacherer 2011) and references therein). Recently, BSA has been coupled to high throughput sequencing methods (for a review see (Swinnen, Thevelein, et al. 2012)) and references therein). In such a BSA set up, an individual displaying a phenotype of interest (superior parent) is crossed with a reference (inferior) parent lacking this phenotype to generate a population of segregants. Subsequently, the segregants are screened to identify a subset displaying the phenotype of interest. These selected individuals are pooled together (here referred to as the "selected pool"), and the genomic DNA of the pool isolated. High-coverage sequencing of this pooled genomic DNA allows identifying for each polymorphic genomic site (referred to as genetic marker sites) the relative frequency of the two (superior and inferior) parental variants in the pool. Variant frequencies of these SNPs should theoretically be 50% for either parent

variant, except for those regions that are genetically linked to the phenotype of interest. At those regions, often referred to as Quantitative Trait Loci (QTLs), the causative allele from the superior parent will be over-represented. The corresponding allele of the inferior parent will be under-represented. Figure 2.1 shows a schematic representation of this approach, which has been successfully applied amongst others in *Saccharomyces cerevisiae* for high ethanol tolerance (Swinnen, Schaerlaekens, et al. 2012), impaired vacuole inheritance (Birkeland et al. 2010), xylose utilization (Wenger et al. 2010), heat tolerance (Parts et al. 2011), variation in colony morphology (Magwene et al. 2011), tolerance to 23 different ecologically relevant environments (Cubillos et al. 2011) and 17 chemical resistance traits (Ehrenreich et al. 2010); in *Zea mays* for drought resistance (Quarrie et al. 1999); in *Arabidopsis thaliana* for growth defects (Schneeberger et al. 2009) and cell wall composition (Austin et al. 2011); in *Oryza sativa* to find agronomically important loci (Austin et al. 2011) and in *Danio rerio* to study developmental mutants (Leshchiner et al. 2012).

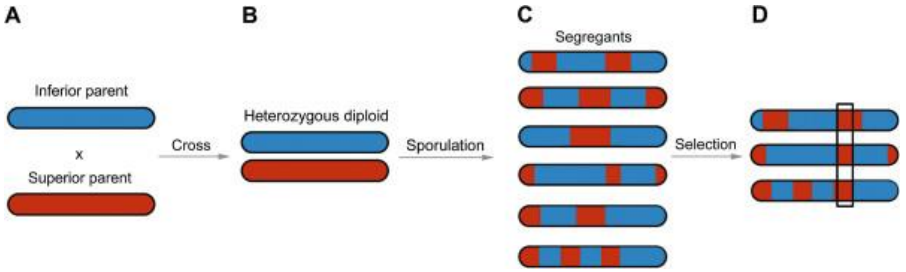


Figure 2.1 Bulk segregant analysis for mapping genomic regions linked to a phenotype of interest in yeast.

A: A parent displaying the phenotypic trait of interest (superior parent) is crossed with a reference strain lacking the trait (inferior parent). **B**: The resulting heterozygous diploid strain is then sporulated to generate haploid segregants. **C**: Segregating offspring carry a mosaic of genetic material derived from both parents (red and blue segments) due to the recombination events in meiosis. After phenotyping, the subset of segregants displaying the trait of the superior parent is selected. **D**: Genomic DNA extracted from the pooled selected segregants is submitted to whole-genome sequence analysis. Polymorphic genomic regions (marker sites) are identified that allow distinguishing between the parental variants. Counting for each marker site how many variants originate from the superior versus the

inferior parent allows determining the variant frequency in the pool for each marker site. Regions linked to the phenotype of interest are expected to originate predominantly from the superior parent (black boxed region). The principle of BSA with diploid organisms is similar, but usually inbred (homozygous) lines are used as parents and two generations are needed to observe segregation of the phenotype.

Theoretically, for any marker site not linked to the phenotype of interest, the alleles in the pool of segregants should be inherited in nearly equal proportions (50%) from either parent. A statistical test e.g., (Birkeland et al. 2010; Swinnen, Schaerlaekens, et al. 2012) can be applied for each genetic marker separately to assess the extent to which the variant frequency at the marker site deviates from the expected inheritance probability of 50%. Hence, the power of QTL mapping by BSA depends on the size of the initial population of segregants, the size of the selected pool and the strength on the phenotype (QTL effect). However, the sequencing procedure can compromise the QTL-mapping power: the sequencing coverage should at least be equal to the number of segregants to ensure information retrieval from all segregants (Magwene et al. 2011). When the coverage is too low, variant frequencies at marker sites will deviate significantly from the theoretical 50% in phenotype-neutral regions due to sampling error. In addition, errors introduced during library preparation, sequencing, read alignment and SNP calling can also cause bias in variant frequency and result in falsely linked regions (regions not truly related to the phenotype). As a result, in reality, spurious deviations of the observed variant frequencies from the theoretical 50% at marker sites will occur due to different sources of experimental error.

To increase the power of QTL mapping by BSA the properties of linkage disequilibrium can be exploited. Linkage disequilibrium (LD) arises because proximal marker sites are co-inherited (Hill & Robertson 1968): in a BSA set up, a causative mutation will thus always be embedded in a larger region of marker sites that all display a deviation from the theoretical 50% inheritance of either parental variant. The extent of the deviation decreases with the distance to

the causative mutation and depends on the resolution of the BSA. Linkage disequilibrium produces deviations of variant counts towards the superior variant, not only at the genetic marker site(s) causative to the phenotype of interest, but also in genetic marker sites closely located to these causative marker sites.

State-of-the-art BSA methods exploit LD to increase the power of BSA analysis but they differ in the way LD is modeled. A first set of methods model LD in a mere data driven way: relative variant frequencies are fitted robustly fit using a sliding window based strategy followed by different smoothing functions (Swinnen, Schaerlaekens, et al. 2012; Ehrenreich et al. 2010; Magwene et al. 2011). More recently, Edwards and Gifford (Edwards & Gifford 2012) developed a Bayesian network called MULTIPOOL to estimate the probability of linkage for each site and (Leshchiner et al. 2012) developed an HMM tailored to perform fine mapping of causative sites in mutagenesis experiments.

We developed a Hidden Markov Model (HMM) called EXPLoRA that explicitly models the effects of linkage disequilibrium to explain the dependencies between neighboring variant frequencies in the observed data. In contrast with other methods, EXPLoRA models the relationship between a genomic variant and the phenotype of interest as a hidden state and use beta-binomial distributions to calculate emission probabilities of the observed data. Tests on simulated data show that EXPLoRA outperforms currently available state-of-the-art algorithms especially in cases where only a limited number of selected segregants can be produced. To further assess the performance of EXPLoRA we analyzed a recently published dataset, described in Swinnen et al., in which three different pools of yeast segregants were used, two of which were selected for tolerance to a different high level of ethanol and one which was used as unselected control pool. Upon re-analysis of the data of Swinnen et al. with our HMM model, we were able to identify reliably QTLs linked to ethanol tolerance that could not be identified with statistical significance in the original

study. An open source java implementation of EXPLoRA, useful for external use and independent validation is available at:

http://homes.esat.kuleuven.be/~kmarchal/Supplementary_Information_Duitama_2013/.

MATERIALS AND METHODS

SIMULATED DATA

To assess the robustness of EXPLoRA we conducted simulations as follows: an artificial chromosome of length 750 kbp with random polymorphic sites was simulated. A single site was randomly chosen to be causative. For each simulation we defined in advance a proportion of segregants in the selected pool with the causative site (referred to as the PSC)). This proportion is used to construct a selected pool as follows: each segregant originates by randomly combining both parental alleles. So each segregant has a probability of 50% to contain the causal variant. Each segregant with the causal variant has a probability equal to the PSC to be present in the final pool whereas a sergeant without the causal variant has a probability of 1-PSC. Segregants are added to the pool until the final number of selected segregants is reached (*n*). By defining in the simulations the 'noise level' as the PSC we avoid to make any assumptions on the cause of the 'noise level' (which can both be attributed to an incomplete QTL effect or to a difficult selection procedure of the selected segregrants) and the subsequent choice of an explicit model to describe the 'QTL effect' of the segregrants. It is important to note that in this simulation set up, a higher number of segegrants (*n*) does not increase the noise level (as is the case for simulations that rely on an explicit phenotypic model (Magwene et al. 2011)). The effect of *n* only affects the results through its effect on the statistical power (if applicable) or because at low values of *n* the relative impact of a sampling error will be higher. Pools of selected segregants of size *n* were created by recombining the parental strains at a constant recombination rate of 0.37 centimorgans (cM) per kilobase, which is the average value for a yeast

chromosome (Ruderfer et al. 2006). Sequences of the selected pools were simulated at variable coverage (*c*) with a constant sequencing error rate of 0.01 (corresponding to the reported Illumina sequencing error (Cherry et al. 1997)). A total of 100 datasets were created for each tested combination of simulation parameters.

Performance analysis

To test the effect of the parameters on the performance of EXPLoRA we used the fixed simulation parameters mentioned above and the following variable ones: n = 30, c = 200. The PSC was varied from 0.6 to 0.95 and the number of polymorphic sites (marker sites) was changed from 10 to 10000. The α_P/β_P ratio was varied from 5 to 40 and the assumed recombination rate (r) was changed from 3.5×10^{-8} to 3.5×10^{-3}. For each setting we report the recovery rate (i.e. the capacity to retrieve the region in which the causal site is embedded), the size of the linked region containing the position of the true causal site and the number of false positive linked regions.

Comparison with state-of-the-art

To perform a comparison with Magwene et al. (Magwene et al. 2011) we used the fixed simulation parameters mentioned above and the following variable ones: 2 500 random polymorphic sites of which a single site was randomly chosen to be causative. Two noise scenarios are presented: Low Noise with a PSC of 0.95 indicating that around 95% of the selected segregants contained the causative allele of the superior parent, and High Noise scenario with a PSC of 0.85. Pools with an increasing number of segregants (n = 5, 10, 20, 30, 200, 500, 1000 and 2000) were simulated. Sequencing of the selected pools was simulated at variable coverage (c = 30, 50, 100, 200, 500 and 1000).

A standalone version of the method described by Magwene et al. was obtained from the authors. For the purpose of comparison both tools were run on the simulated data (see above). To assess sensitivity we measured the power of

each method as the number of times that the region in which the causative site was embedded was found to be significantly linked divided by 100 (the number of repeats for each experimental setup). For EXPLoRA a marker is significantly identified if the posterior probability assigned to the marker is larger than 0.95. For the method of Magwene et al. we calculated for each experiment the null distribution of the G' score using the non-parametric method described by the authors. Based on this null distribution, we calculated a p-value for each marker, also following the method described in Magwene et al. A marker is significantly linked with the phenotype if its p-value passes FDR correction at a 0.05 significance level (Magwene et al. 2011; Glenn 2011). Specificity is measured using two metrics: the size of the linked region at the causal position and the number of false positive linked regions found. We ran the method of Magwene et al. with a default genetic window size of 30 cM, as recommended by the authors. For EXPLoRA we fixed the α_P/β_P ratio at 15 which gives the best tradeoff between the recovery rate and the size of the predicted regions.

Real dataset

To test our method, we used the dataset reported by Swinnen et al. In their work, a segregant, VR1-5B (superior parent) from a Brazilian bioethanol production strain VR1 was crossed with the BY4741 lab strain (inferior parent). A total of 136 segregants tolerant to 16% ethanol and out of these, 31 segregants also tolerant to 17% ethanol, were pooled. DNA of the pools and also of the VR1-5B parental strain was extracted and sequenced using Illumina technology (100 bp reads). A total of 131 unselected segregants from the same cross were also pooled and sequenced as control experiment (unselected pool).

Marker sites were identified as follows: the yeast S288c reference genome (3 Feb. 2011 release) available in the Saccharomyces Genome Database (http://www.yeastgenome.org) was used as reference. All reads from the

parental strain VR1-5B were mapped to the reference sequence using bowtie2 (Benjamini & Yekutieli 2005). We used the -a option to retain as many good alignments as possible for each read. Over 93% of the reads from VR1-5B, 84% and 86% of the reads from the pools of segregants under selection, and 98% of the reads from the pool of unselected segregants could be mapped to the latest reference genome. We ignored the last 25 bp of each read from the VR1-5B strain and the two pools of selected segregants based on the base calling error rate estimated from unique alignments.

SNPs and small indels between the two parents VR1-5B and S288c (the reference sequence) were identified with the SNVQ algorithm (Langmead & Salzberg 2012). We filtered out predicted variants with genotype quality scores lower than 40, falling into annotated repetitive regions (i.e., transposons, telomeres, centromeres), or falling into duplicated regions predicted either by reads with multiple alignments or by the CNVnator algorithm (Duitama et al. 2012). Finally, we filtered out predicted variants located less than 30bp from each other to avoid undesired local errors due to misaligned reads. We obtained 25,972 SNPs and 1,429 indels which were used for analysis of segregant pools.

To identify the relative variant frequencies in the pools of segregants at marker sites, we implemented a custom script to count at each marker site the number of read alignments that support the variant originating from the superior parent (VR1-5B) and the total number of alignments. Within each pool variants with read coverage less than 20 or over 100 were ignored. We retained 26,913 variants for the 16% pool, 26,865 variants for the 17% pool, and 24553 variants for the pool of unselected segregants.

Experimental validation

Experimental verification of QTL2 on chromosome X was based on determining for a selected set of marker sites in this region, the number of times individual

segregants selected for high ethanol tolerance displayed the variant originating from the superior parent (relative variant frequency in individual segregants) (Swinnen, Schaerlaekens, et al. 2012). Relative variant frequencies in individual segregants were used to calculate the posterior probability of each marker site to be linked to the phenotype of interest using an exact binomial test with a confidence level of 95% and correction for multiple testing by a false discovery rate (FDR) control according to (Benjamini & Yekutieli 2005). Ethanol tolerance assays and reciprocal hemizygosity analysis were carried out as described previously (Swinnen, Schaerlaekens, et al. 2012).

RESULTS

DEVELOPMENT OF EXPLORA, A HMM FOR THE ANALYSIS OF BSA DATA

As indicated above, BSA is the first step towards finding sequence variations (also referred to as "alleles", "variants”") that cause a given phenotype. Causative sequence variations originating from the superior parent are expected to be over-represented in the selected segregant pool. Due to linkage disequilibrium (LD), other variants at marker sites that surround the causative site will also be over-represented in the selected pool. LD thus limits the resolution of the BSA analysis towards identifying the region in which the true causal site is embedded rather than the true causal site. However, this dependency between neighboring sites (LD) can be exploited to increase the power of the statistical linkage of the identified loci to the phenotype of interest by filter out spuriously linked regions. To exploit the information contained in the dependency between neighboring marker sites, we developed a Hidden Markov Model (HMM) called EXPLoRA (Figure 2.2).

EXPLoRA explicitly models the effect of linkage disequilibrium to explain the dependencies between neighboring sites in the data. EXPLoRA models for each marker site, two possible states: one state (P-state) expresses that the variants in the pool at that marker site originate predominantly (but not always in all segregants) from the superior parent and are thus linked to the phenotype of

interest. A second state (N-state) models that the variants in the pool at a given marker site originate to an equal extent from either parent, in which case the marker site is assumed to be located in a neutral region not linked to the phenotype of interest. The effect of linkage disequilibrium is modeled by the transition probabilities τ between two neighboring marker sites. The transition probability τ models the chance that a neighboring site remains in the same state as its preceding site state. Its distribution is described by a negative exponential model as a function of the recombination rate and the physical distance between neighboring marker sites (Scheet & Stephens 2006) (Figure 2.2C). The probability to change states upon transition from one marker site to its direct neighboring marker site (from a neutral N-state to a phenotype-linked P-state or vice versa) is then described by 1-τ and takes into account the true distance between them (i.e. no distance binning is involved). The model captures the fact that marker sites located in each other's physical neighborhood are likely to be in linkage disequilibrium and less likely to change their state (from P to N or from N to P). Given a random state N_i or P_i at a marker site 'i', the transition probabilities to the states N_i+1 or P_i+1 for the neighboring marker site 'i + 1' are given by:

$$\tau_{N_i \to N_{i+1}} = 1 - e^{-rl_i}$$

or

$$\tau_{P_i \to P_{i+1}} = 1 - e^{-rl_i}$$

where l_i is the physical distance between the marker sites *i* and *i*+1 and *r* is a recombination rate, which is determined by the average number of crossing-overs occurring during meiosis over a given distance in a chromosome. The default level of *r* was fixed at 3.5×10^{-6}, based on the estimations derived by (Ruderfer et al. 2006).

Each state in the model emits a random variable n_A, corresponding to the number of variant counts at a given marker site originating from the superior parent. n_A ranges from 0 to n, with n being equal to the (known) total variant count for the marker site. n_A is described by a beta binomial distribution, which allows capturing different emission probabilities in phenotype-linked versus neutral states by choosing different α and β parameters for their corresponding distributions (Figure 2.2B). We modeled all neutral states with the same parameters α_N and β_N, and all phenotype-linked states with the same parameters α_P and β_P. While for the neutral states α_N should almost equal β_N to make values of n_A closer to $n/2$ more likely to be sampled, for the phenotype-linked states α_P should be much larger than β_P to make values of n_A close to n more likely to be sampled.

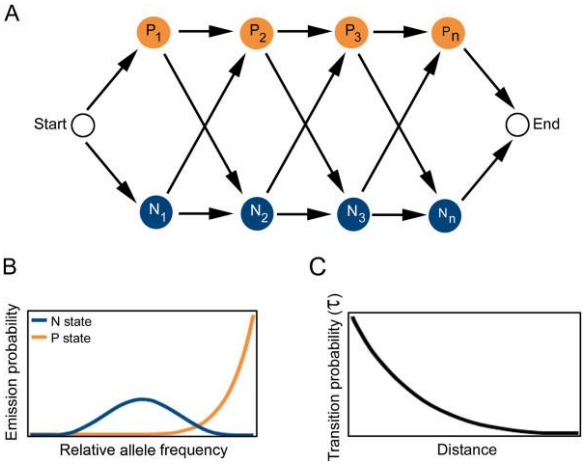


Figure 2.2 Hidden Markov Model used to predict genomic regions linked to the phenotype of interest.
A: each marker site is modeled to be in a neutral state (N-state, blue circles) or in a state of being linked to the phenotype of interest (P-state, orange circles) based on its observed relative variant frequency in the pool of segregants. **B**: emission probabilities for respectively the neutral (blue curve) and the phenotype-linked states (orange line) as a function of the relative variant frequencies, modeled by a beta-binomial distribution with respective parameters α and β. **C**: transition probability as a function of the physical distance between neighboring marker sites.

Given the observed total variant count and the variant counts that originate from the superior parent at each marker site (*D*) and fixed values for the parameters α_N, β_N, α_P, β_P, and τ, we can calculate the posterior probability of each state in the HMM with a standard forward-backward algorithm (Scheet & Stephens 2006). For each marker site, we then estimate its probability to be linked to the phenotype of interest as the normalized probability $P(P_i \mid D) / (P(P_i \mid D) + P(N_i \mid D))$.

Since most of the genomic regions are supposed to be neutral with respect to the phenotype of interest, the parameters α_N and β_N of the emission probabilities in the neutral state can be estimated directly from the observed variant frequencies. To this end, we implemented a two-step process in which we first assume that most of the genomic regions are phenotype-neutral. We estimate with the method of moments the most likely values of α_N and β_N given the variant frequencies at each marker site. Then in a second step we identify the marker sites linked to the phenotype of interest using the model, and we estimate again α_N and β_N leaving out the marker sites identified to be linked to the phenotype. The ratio between α_P and β_P thus defines the degree to which the relative variant frequency at a marker site needs to differ from the one obtained through random inheritance for it to be called linked to the phenotype (stringency of the method). Changing the ratio affects the probability with which an observed relative variant frequency is interpreted by the model as a phenotype linked region (see also below). In our experiments, we altered the ratio between α_P and β_P by fixing β_P equal to 1 and testing different values of α_P. A cut-off on the obtained posterior probability of each marker site to be linked to the phenotype was used to prioritize the most likely causative marker sites for the phenotype of interest.

PARAMETER SENSITIVITY OF EXPLORA

We tested to what extent changing the model parameters (i.e. the α_P/β_P ratio and the recombination rate) affect the results in terms of the recovery rate, the number of falsely predicted linked regions and the average size of the predicted regions. Tests were performed under two different settings that assess respectively the effect of diluting the signal to noise ratio and the resolution of the BSA. Changing signal to noise ratio's is simulated as explained in Materials and methods (PSC) and mimics the effect of e.g. having an incomplete QTL effect of the causal genes, because for instance several minor alleles might be involved or an imperfect selection procedure of the segregants. The BSA resolution was altered by varying the number of marker sites in the artificial set up (see Material and methods).

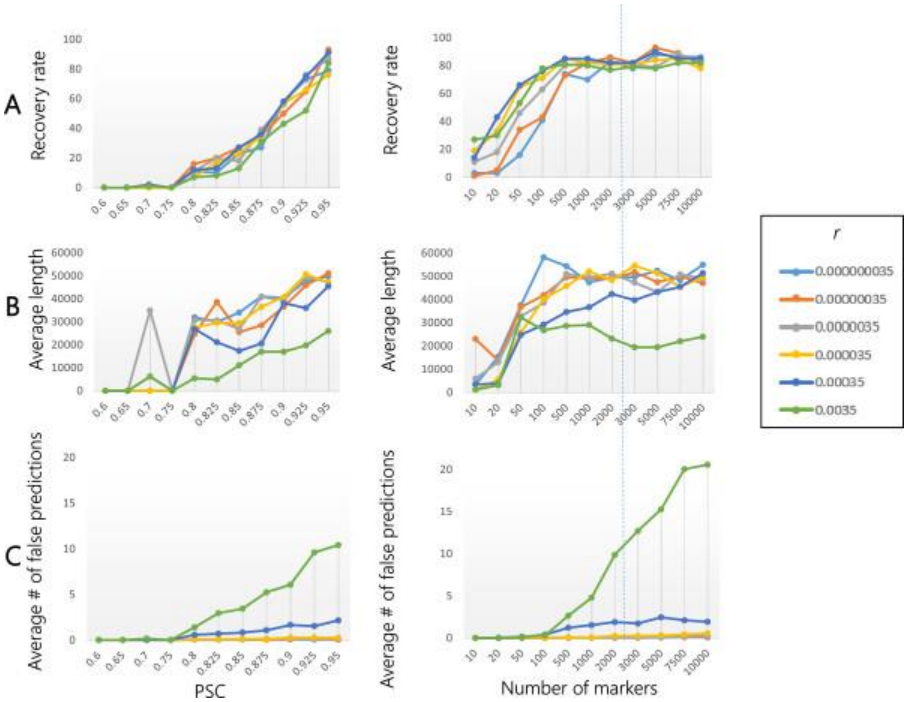


Figure 2.3 Effect of the recombination rate (r) on the performance of EXPLoRA.
The recovery rate (panel A), average size of the linked region (panel B) and number of falsely predicted regions (Panel C) as a function of the noise level (left sided plots) and the number of marker sites (right sided plots). The noise level is represented by the ratio of the segregants in the pool that have the causal allele versus those that have not (PSC). Results obtained with a number of markers that occur in real experimental settings are indicated with a dotted line.

Both Figure 2.3 and Figure 2.4, show that irrespective of the choice of the parameters, the recovery rate will drop with the noise in the dataset (noise equals lower QTL effect). The number of falsely predicted linked regions is in general quite noise independent (except for extreme overestimations of r, see also below). Counterintuitively, we observe that the average region size becomes more refined with increasing noise levels (an observation we also made in the real data). The latter can be explained by the fact that when the signal/noise level decreases, a longer region with truly deviating relative allele

frequencies will have more chance to become interrupted (as the distinction between signal and noise is not that clear). Figures 3 and 4 also show that the recovery rate, the region sizes and the number of falsely predicted linked regions (except for extreme overestimations of *r*, see also below) are almost independent of the BSA resolution (the number of marker sites), provided a minimal number of markers is available. In the following, we will focus on the effect of the parameter choices on the results of EXPLoRA.

The parameter 'recombination rate (*r*)' determines the shape of the transition probability function which models the change from the N-state to the P-state and vice versa. EXPLoRA predicts causal sites by transitioning between these states. Gradually overestimating/underestimating the recombination rate, decreases the impact of linkage disequilibrium in modeling the effect between neighboring sites. How this affects EXPLoRA is shown in Figure 2.3 (both for different noise levels value and number of markers). In general, as *r* is gradually more overestimated, markers sites will be treated increasingly independent and each region with a sufficiently deviating relative allele frequency will be predicted as being linked to the phenotype, even spurious signals. This is clear in Figure 2.3 that shows that independent of the noise level or the number of markers (provided you have a minimal number of 1000 markers), seriously overestimating *r* results in smaller linked region sizes of the true peaks. This, however, comes at the expense of selecting a much higher number of false positive regions. Expectedly, this behavior is most pronounced under conditions with a high number of markers as under those conditions the chance of introducing spurious signals is higher. The behavior is also more present at low noise levels which is counterintuitive but can simply be explained by the fact that at high noise levels EXPLoRA does not identify any linked regions, not even spurious ones. However, at low noise levels when regions are identified, overestimating *r* results in splitting up a truly linked region into smaller regions because the method becomes more sensitive to the small noisy variations in allele frequencies. So rather than identifying truly

falsely linked regions, a high value of *r* only results in splitting up a truly linked region.

In contrast to the number of false linked regions and the region size, the recovery rate is unaffected by the choice of the parameter *r*. Contrarily to overestimating *r*, underestimating r almost does not affect the results.

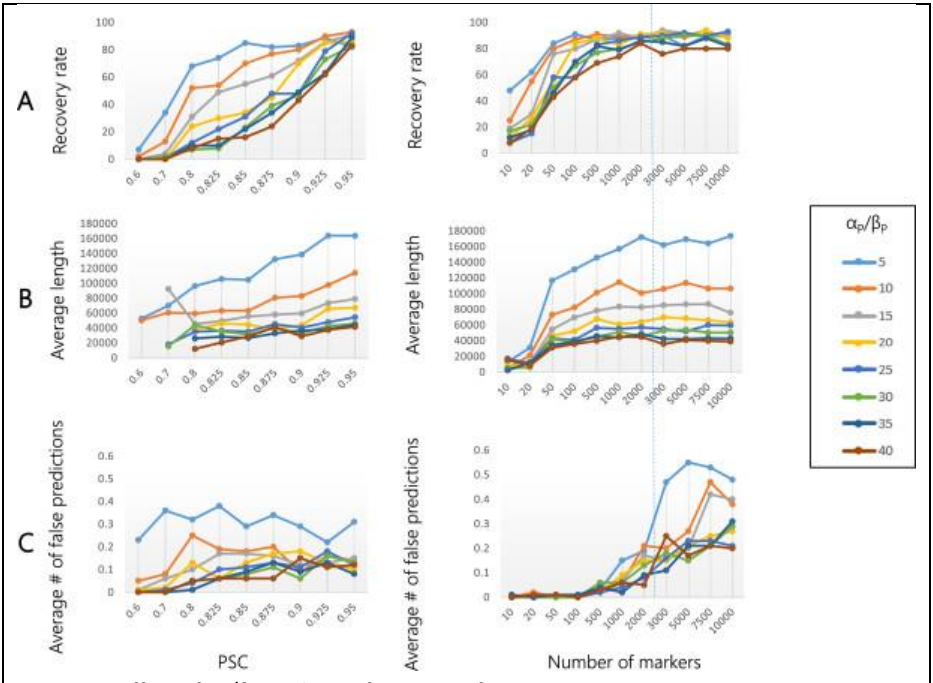


Figure 2.4 Effect of α_P/β_P on the performance of EXPLoRA.
The recovery rate (panel A), average size of the linked region (panel B) and number of falsely predicted regions (Panel C) as a function of the noise level represented by the ratio of the segregants in the pool that have the causal allele versus those that have not (PSC) (left sided plots) and the number of marker sites (right sided plots). Results obtained with a number of markers that occur in real experimental settings are indicated with a dotted line.

Changing the α_P/β_P ratio affects the emission probability or the probability with which an observed relative variant frequency is interpreted by the model as a

phenotype linked region. Increasing the α_P/β_P ratio makes the prediction more stringent, meaning that a higher deviation of the relative allele frequency is needed before the region is considered linked.

The results in Figure 2.4 are consistent with this explanation: expectedly a lower α_P/β_P (less obvious relative allele frequency deviations needed) increase the recovery rate. Interestingly, the choice of α_P/β_P does not affect the number of falsely linked regions (except maybe for $\alpha_P/\beta_P = 5$, but also here the number of falsely linked regions is still lower than one per dataset), but it rather affects the average size of the linked regions. This means that provided the parameter *r* is not overestimated and linkage disequilibrium is taken into account, consistency between neighboring marker sites will compensate for the spurious deviations in relative allele frequencies. Making the ratio α_P/β_P less stringent will thus only extend the size of the truly linked region, but does not affect the number of false positive predictions.

Also the recovery rate, region size and the number of false positive linked regions (note the scale of the plot in this case) as a function of the number of marker sites is relatively independent of the choice of α_P/β_P. For a high number of markers, it seems that a less stringent α_P/β_P ratio results in a relatively higher number of false positives (although again the absolute numbers are still lower than 1 false positive peak per dataset). To some extent introducing more markers will result in a higher chance of also detecting spuriously deviating relative allele frequencies.

Conclusively, at a number of available marker sites comparable to those found in real life situations (e.g. ~2 500 marker sites in 750 Kb is comparable to the yeast real data analyzed in this paper), and choosing a value for *r* that approximates the real recombination rate (which can be estimated from real data), EXPLoRA will be able to predict truly linked regions with very little false positive regions, even for experimental settings with low QTL effect (meaning

that the expected relative allele frequency at the causal site is low). The choice of α_P/β_P allows tuning the tradeoff between the recovery rate and the size of the linked region but does not interfere too much with the number of false positive regions.

COMPARISON WITH STATE OF THE ART

To illustrate the added value of explicitly modeling linkage disequilibrium (LD) in EXPLoRA, we ran our tool on simulated datasets and compared its performance to that obtained with the method of (Magwene et al. 2011). This model is a state-of-the-art method for the analysis of BSA results, belonging to the class of statistical methods that apply a windows-based strategy to capture the block-like behavior of the relative allele frequencies plotted along the genome.

Simulations mimicked different BSA experiments, differing from each other in their noise level (high and low noise level), the number of selected segregants (*n*) and the coverage (*c*) at which this pool was sequenced. Note that in our simulation set up, the noise level is mimicked by fixing the ratio of the segregants in the pool that have the causal allele versus those that have not. As a result, except for the higher impact of sampling errors at low *n*, the noise level in our simulation set up is independent of the number of selected segregants *n*.

For each experimental set up 100 different datasets were simulated and performances were assessed by the recovery rate, the false positive detection rate, and the average region size as described in Materials and methods.

Figure 2.5 shows that expectedly for both methods the recovery rate decreases with the noise in the dataset. The number of false positives is quite noise independent for both methods. For the method of Magwene et al., and for a given *n, c* combination, the size of the linked region is relatively independent

of the noise level, whereas for EXPLoRA we again observed a decrease in region size with the increase in noise level (as was already noted above).

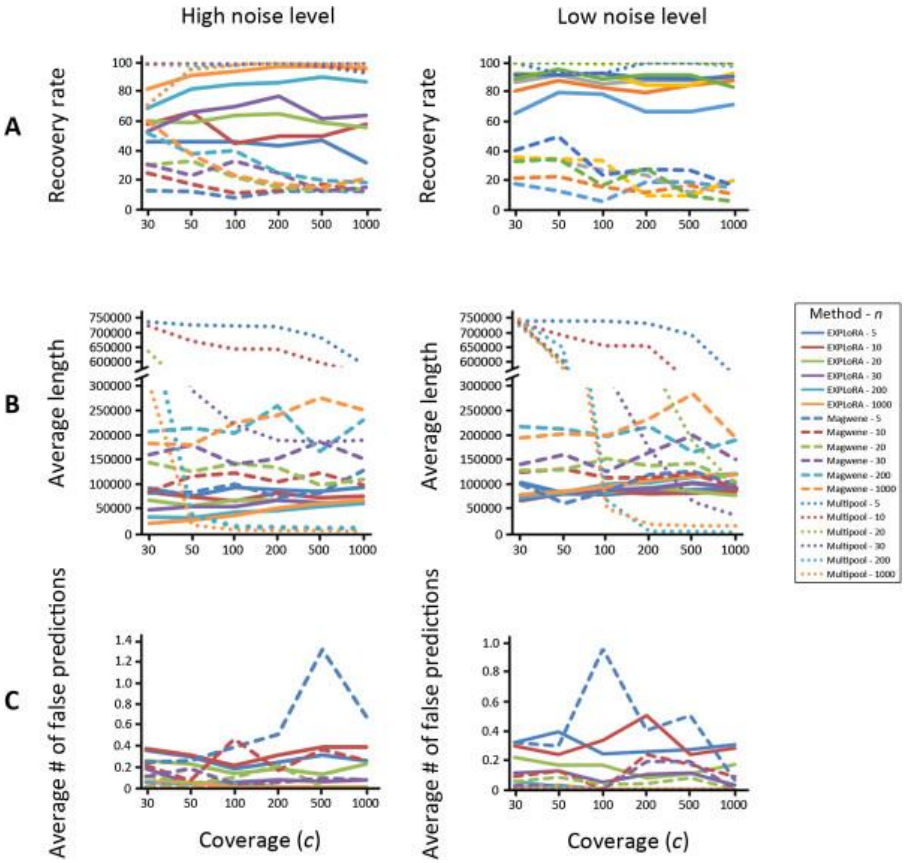


Figure 2.5 Comparison with the state-of-the-art.
The recovery rate (panel A), average size of the linked region (panel B) and number of falsely predicted regions (Panel C) under high (left sided plots) and low (right sided plots) noise levels were assessed for EXPLoRA the method of Magwene et al. and MULTIPOOL. In the plots of panel **B** (average size of the linked region) the y-axis was split into two scales to facilitate showing the results of MULTIPOOL without compressing the curves obtained by EXPLoRA and the method of Magwene et al.

For both methods the performance (recovery rate, number of falsely linked regions) decreases with a lower number of segregants. This is due to the fact that at low *n* values, sampling errors increase i.e. the relative impact of by chance including a segregant that does not carry the causal allele is higher. For the method of Magwene et al. a low *n* also interferes with the used statistics, further exacerbating the drop in performance at low *n*. This is also the reason why Magwene et al. specifically recommend against applying their method on data obtained from small segregant pools.

Given the used parameter and FDR correction settings, EXPLoRA obtains a higher recovery rate with smaller regions sizes for both noise levels than the method of Magwene et al. This low recovery rate of Magwene et al. is mainly due to the stringency in the selection imposed by the robust FDR as the raw linkage scores prior to the FDR selection were observed to be genuinely high at truly linked regions. The FDR also results in the counterintuitive decrease of recovery rate of Magwene et al. with increasing coverage, a behavior that was not expected based on the visual interpretation of the raw linkage score (G') (see Supp Figure A. 1). Using the simpler version of the FDR (which does not take into account dependency between tests) compensates for this loss of recovery rate, but comes at the expense of a much larger linked regions (see Supp Figure A. 2).

Conclusively, EXPLoRA shows state-of-the-art performance. More importantly its performance remains extremely robust even when lowering the number of selected segregants or when the signal/noise level is low. These properties make the method particularly useful under BSA conditions for which segregant selection is non-trivial or the QTL effect is minor (e.g. when several minor alleles are contributing to the phenotype).

APPLICATION OF EXPLORA TO REAL DATASETS

To evaluate the performance of our analysis method with a real BSA experiment, we applied EXPLoRA to the data described in Swinnen et al. In their analysis they used a statistical smoother to facilitate detecting from the raw data regions with deviations in relative allele frequencies. Based on visual inspection and comparing the results from the 16 and 17 % pool allowed them to predict 6 loci as being significantly linked to the phenotype, all of which were also explicitly mentioned in the paper. Of those the QTLs 1, 2 and 3 were further proven to be statistically linked by individual genotyping of SNP markers surrounding each QTL.

To test to what extent we could recapitulate their results, we ran EXPLoRA with both $\alpha_P/\beta_P = 30$ and $\alpha_P/\beta_P = 10$ ratios and a cut off on the posterior probability score of 0.95 on the pools selected for 16 and 17% ethanol separately. In Figure 2.6 the most confident results are shown i.e. those results that either could be confirmed with both parameter settings (the most and the least stringent that is $\alpha_P/\beta_P = 30$ and $\alpha_P/\beta_P = 10$) or that could be confirmed in both pools (16 and 17 % ethanol) with at least one parameter setting. With $\alpha_P/\beta_P = 10$ and setting a minimum posterior probability of linkage of 0.95 we predicted in the 16% pool 923 marker sites clustered in four QTL regions. In agreement with the initial study of Swinnen et al. we identified the experimentally verified QTL1 located on chromosome V between coordinates 116,000 and 117,000, containing the causative gene *URA3*. QTL2 located on chromosome X between coordinates 646,155 and 662,146 (for which no causative gene was reported in the original work of Swinnen et al.) and QTL3 encompassing a gene cluster on chromosome XIV between coordinates 466,000 and 486,000, containing the causative genes *MKT1* and *APJ1*. In addition, we detected one QTL that was mentioned but not further validated in the initial publication: a small, but still significant region on chromosome II (referred to as QTL4 encompassing 18 of the marker sites (Figure 2.6)). The length of the linked regions identified with

α_P/β_P= 10 varies from as small as 4.3 kbp for QTL4 to as large as 226 kbp in QTL3.

These four QTLs (QTL1, 2, 3, and 4) identified in the 16% pool were also detected in the analysis of the 17% ethanol pool using EXPLoRA with the same parameter settings (α_P/β_P =10), further increasing the confidence that these QTLs were truly linked to ethanol tolerance (these regions encompassed a 757 (37.2%) of the total number of linked marker sites (2,034) in the 17% pool). In addition the more stringently phenotypic selection of the 17% pool allowed drastically decreasing the length of QTL1 and QTL2 (reducing them from 123 kbp and 16 kbp to 58 kbp and 5.3 kbp respectively) as detected by EXPLoRA with α_P/β_P =10.

The remaining 607 linked markers in the 17% pool mapped to a QTL encompassing a region of 105 kbp in chromosome XV (referred to as QTL5) and to three small regions on chromosomes I, VI, and XII. Neither of those QTLs was detected in the 16% pool, indicating that they are specifically enriched at more extreme ethanol levels (17%). The fact that the region at chromosome XV (QTL5) could also be confirmed with the more stringent value of α_P/β_P = 30 (see also below) indicates that from these additional QTLs, this region is the best candidate to be an additional truly linked region. Using the same settings (α_P/β_P= 10 and α_P/β_P= 30 and a cut off on the posterior probability score of 0.95), EXPLoRA did not report significant relationship with ethanol tolerance for any polymorphic site in the control pool of unselected segregants.

Figure 2.6 further illustrates the effect of changing the α_P/β_P ratio on the recovery rate and the size of the linked region for the identified QTLs on respectively the 16 and 17% pool. As predicted by the simulation experiments, changing the ratio α_P/β_P from less (10, solid line) to more stringent values (30; dashed lines) reduces the length of the linked region size, but comes at the expense of missing the least pronounced QTLs. For instance, for the 16%

ethanol pool increasing the α_P/β_P ratio, reduces the length of QTLs from 123 kbp to 66 kbp and from 226 kbp to 93 kbp in QTL1 and QTL3 respectively. However, this more stringent setting results in missing QTL2 and QTL4 (dashed lines in Figure 2.6) in the 16% pool, indicating that for this pool the signals of these QTLs are not very pronounced (minor QTLs in 16% ethanol). Equally, in the 17% pool increasing the stringency of EXPLoRA, reduces the length of the linked regions in QTL3, 4 and 5, but results in missing QTL1 and QTL2 and the additional smaller linked regions in chromosomes I, VI, and XII.

These results indicate that the signal of QTL3 is prominent in both pools and thus very relevant for ethanol tolerance under both ethanol conditions. The signal of QTL1 is clearly more pronounced in the 16% pool than in the 17% pool, whereas for the signals of QTL4 and QTL5 the opposite is true implying that under both ethanol conditions other protection mechanisms tend to play a role. The region in QTL2, despite being a minor locus (not such pronounced signal) might play an equally important role under both ethanol conditions as it is recovered in both pools.

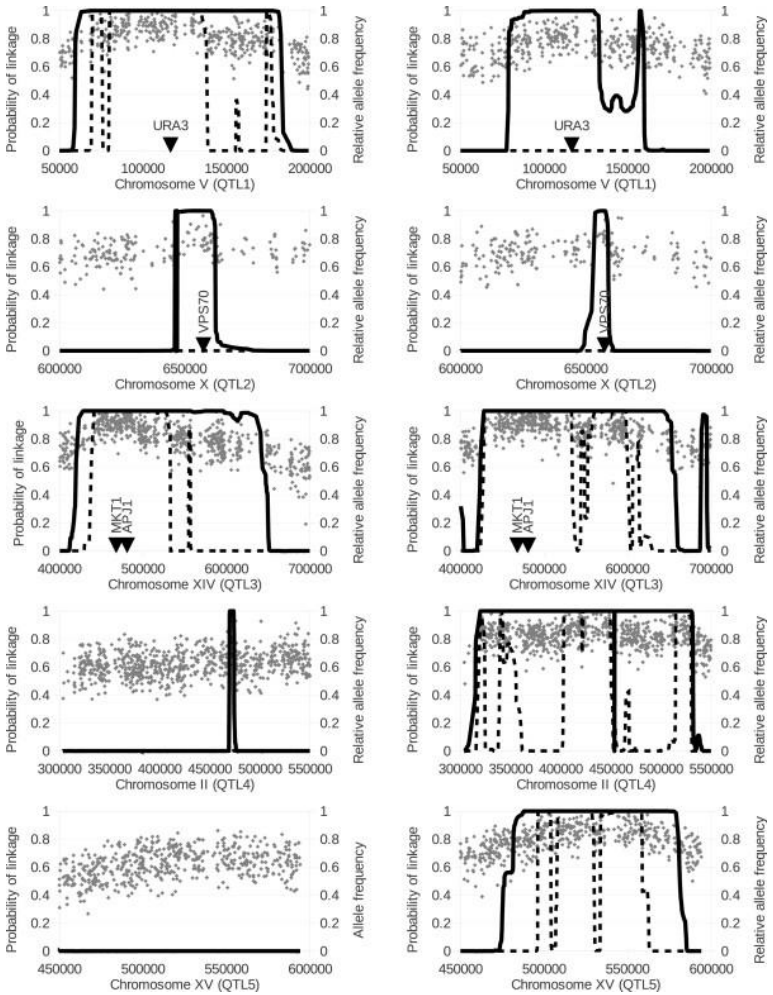


Figure 2.6 Linkage scores obtained by EXPLoRA for the five QTLs identified in the 16% pool (left) and in the 17% pool (right).

The original relative variant frequencies as determined by genome sequencing are displayed for each plot (light gray dots). Solid lines show the posterior probabilities for $\alpha_P/\beta_P = 10$ whereas dashed lines show the posterior probabilities for $\alpha_P/\beta_P = 30$.

EXPERIMENTAL VALIDATION OF THE NEWLY PREDICTED QTL2 ON CHROMOSOME X

To assess the validity of our predictions, we selected QTL2 (on chromosome X) for experimental validation as this QTL, despite being important in both the 16% and 17% pool seemed to be one of the more difficult QTLs to detect (only confirmed by the least stringent selection criteria). Fine-mapping of the region by PCR-based scoring of the markers in the individual segregants (Materials and methods), allowed us to confirm the area with the strongest link. Mutations in this confirmed region were verified by Sanger sequencing. All genes carrying non-synonymous mutations in their coding region were first selected as candidate causative genes (Figure 2.7A). True causative genes in QTL2 were identified using reciprocal hemizygosity analysis (Steinmetz et al. 2002). For each candidate causative gene a set of two diploid strains was constructed by crossing the parental strains, either containing or lacking the candidate gene. As a result each diploid has a different allele of the candidate gene while the other copy of the gene is deleted (Figure 2.7B). Phenotypic analysis on YPD plates with 16% ethanol showed a clear difference in ethanol tolerance between the two diploid strains carrying a different allele of *VPS70:* the strain with the allele derived from the VR1-5B superior parent grew very well in the presence of 16% ethanol, whereas the strain with the allele from the BY4741 inferior parent did not grow at all (Figure 2.7B), indicating that *VPS70* carries a causative mutation responsible for the link of QTL2 with high ethanol tolerance. Except for a putative role in sorting of vacuolar carboxypeptidase Y to the vacuole (Bonangelino et al. 2002), no link to ethanol tolerance for *VPS70* has been reported previously. This may be due to the fact that all previous analyses of yeast ethanol tolerance were performed with laboratory strains and with much lower ethanol concentrations e.g., (van Voorst et al. 2006).

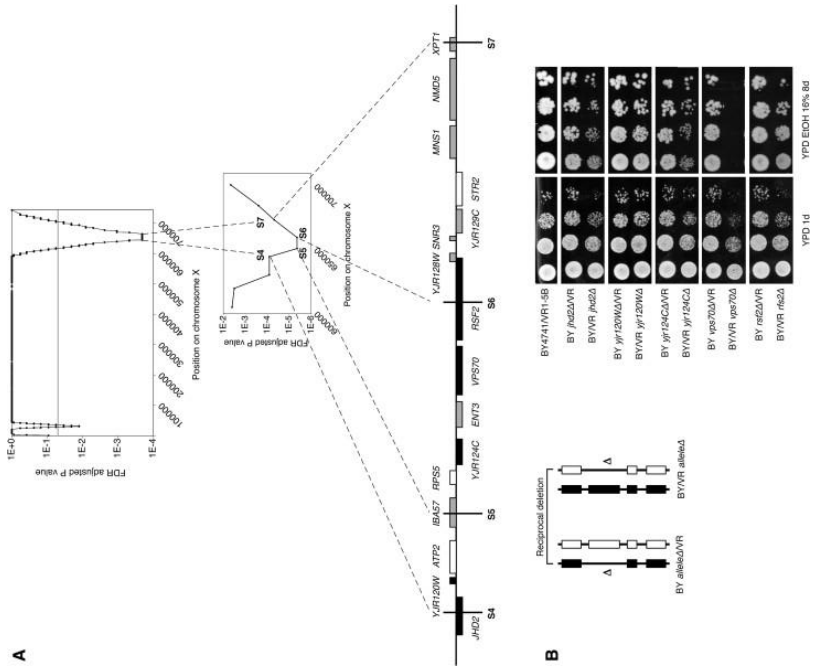


Figure 2.7 Experimental validation of QTL2 on chromosome X.

A**:** upper plot shows the region corresponding to QTL2 of which linkage to the phenotype of interest was confirmed by scoring selected marker sites in individual segregants. Scored marker sites are indicated (S4-S7). For each marker site, the p-value indicates the probability to be linked to the phenotype by chance according to a binomial distribution (see materials and methods). Lower plot: zoom in on the genes in the experimentally confirmed region corresponding to QTL2 (29 kb). Black bars: genes with non-synonymous mutations in the coding region; grey bars: genes with mutations in the promotor or terminator; white bars: genes without mutations. B**:** Reciprocal hemizygosity analysis for the genes with non-synonymous mutations in the coding regions located in the fine-mapped region. To that end, two different diploid strains were constructed by crossing the original superior parent VR1-5B with the inferior parent BY4741, carrying a deletion in its allele of the candidate causative gene or the other way around. Hence, this resulted in two different diploid strains, each with only one functional allele of the candidate causative gene, originating from either the 'superior' or the 'inferior' parent. The ethanol tolerance of the two diploid strains was compared with dilution spot growth assays on a YPD plate with 16% ethanol and a YPD plate without ethanol as control.

WEB SERVER

We also developed EXPLoRA-web, an intuitive webserver that facilitates users performing data analysis of BSA experiments. EXPLoRA-web is wrapped around our BSA data analysis method that was shown to maximize power and accuracy in detecting QTLs by exploiting the properties of LD. The web service is available at http://bioinformatics.intec.ugent.be/explora-web/.

EXPLORA WEB

EXPLoRA web is accessible using any internet browser (Google Chrome, Microsoft Edge and Firefox, among others). The webserver's help pages http://bioinformatics.intec.ugent.be/explora-web/help provide detailed guidelines on how to perform the analysis, tune the parameters and interpret the results. The EXPLoRA web server is freely accessible and does not require login, although an optional account can be created to have easy access to the results of previously analyzed experiments.

EXPLoRA navigates the variable sites of the genome using a Hidden Markov Model (HMM) that calculates the probability of allele frequencies at each marker site to be emitted by two possible states: a phenotype linked state and a neutral state. While markers linked to the phenotype are expected to show predominantly the allele of the superior parent, neutral markers are expected to show the alleles of the two parents at a ratio that reflects random segregation.

The effect of linkage disequilibrium is modeled by the transition probabilities between two neighboring marker sites. The transition probability models the chance of a change of state between two neighbor sites. Its distribution is described by a negative exponential function of the recombination rate and the physical distance between neighboring marker sites following Haldane's recombination model. The model captures the fact that neighboring marker

sites are likely to be in linkage disequilibrium and hence the probability of a state change between them is small.

Emission probabilities of marker site states are modeled by two beta binomial distributions, one for the emission probabilities in phenotype linked states and another for emission probabilities in neutral states. The beta-distribution for the neutral states is automatically estimated from the read count data. The distribution describing the expected frequencies of the phenotype-linked variants is defined by the user selecting the α and β parameters. The ratio between α and β defines the degree to which the relative variant frequency at a marker site needs to differ from the one expected based on random segregation in order to be called 'linked to the phenotype'.

INPUT

The data necessary to run EXPLoRA consists of the information on the experimental setup, the allele counts from the pooled sequencing and the selection of the method's parameters.

Information related to the experimental setup consists of the number of segregrants that were pooled prior to sequencing and an approximation of the true recombination rate of the organism under study. The latter is required to take into account the effect of LD between neighboring markers. Additional fields to name and describe the experiment are also provided (Figure 2.8A).

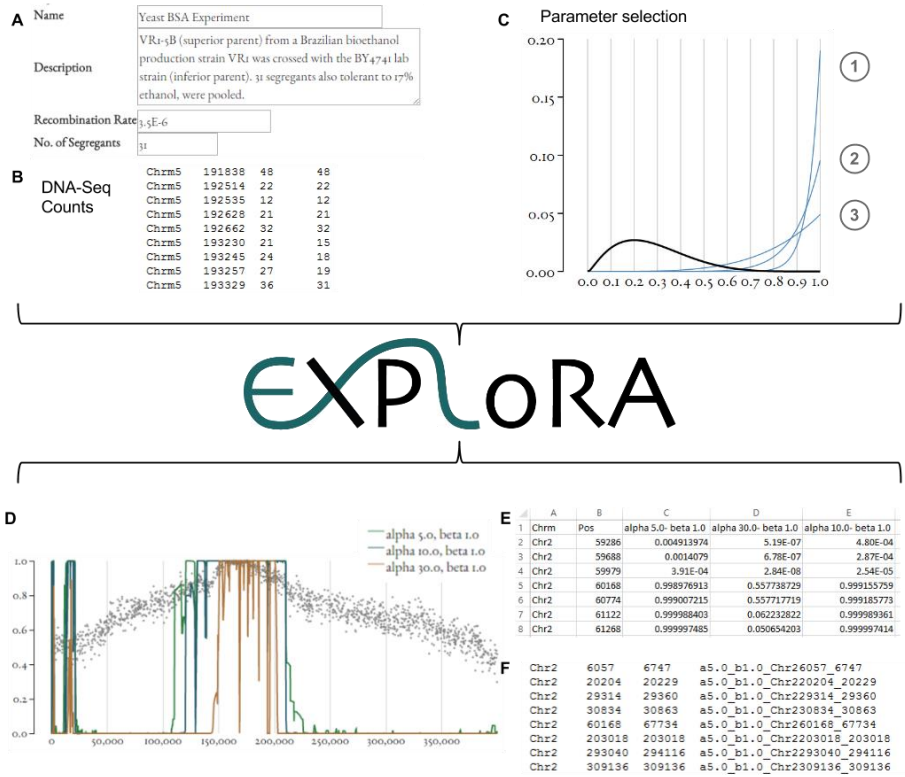


Figure 2.8 Overview of the web service.
A) Input experimental information. B) Upload count data. C) Parameter selection. The black line corresponds to the cumulative distribution of allele frequencies (alternative read count/total read count) derived from the uploaded data and is used to estimate the probability distribution that models the emission probability of the neutral state allele distribution. The blue lines represent the beta-distributions model for each of the 3 different ratios of alfa/beta where *line 1* corresponds to a setting reflecting high specificity and low sensitivity, *line 2* medium specificity and sensitivity and *line 3* low specificity and high sensitivity. D) Visual Output. The X-axis corresponds to the chromosomal positions and the Y-axis to the posterior probabilities obtained for each marker site. E) Posterior distributions of the marker sites for each parameter setting. F) BED file indicating the regions linked to the phenotype.

The allele count at the marker sites derived from the pooled DNA sequencing should be uploaded as a variant call format (VCF) file or as a simple tab delimited file in which rows correspond to the different marker sites (Figure 2.8B). For the simple text tab delimited file the columns correspond to respectively the chromosome at which the marker sites are located (marker chromosome), the genomic position of the marker sites (marker site position), and two columns describing the read counts containing respectively the total read count at a marker site and the alternative allele read count (which usually corresponds to the allele of the superior parent). VCF files containing these allele counts are produced from raw reads by analysis pipelines for high throughput sequencing data such as NGSEP (Duitama et al. 2014) and GATK (Van der Auwera et al. 2013), so the outputs of these pipelines can be directly uploaded to EXPLoRA-web.

Besides the data and experimental information, EXPLoRA also requires specifying the parameters that control the model. The main parameter to be selected is the α/β ratio that determines the shape of the beta distribution that models the emission probability for the phenotype-linked states (Figure 2.8C). Changing the α/β ratio affects the probability with which an observed relative variant frequency is interpreted by the model as representative for a region linked to the trait of interest. Increasing the α/β ratio makes the identification of phenotype-linked regions more stringent, meaning that a higher deviation of the relative allele frequency from the one expected under random segregation is needed before the region is considered linked. The higher the α/β ratio, the less phenotype-linked markers are called and the smaller the size of the called regions: identifying only the most pronouncedly linked regions results in more precisely pinpointing the linked region. However, this comes at the expense of potentially missing some truly linked markers/regions (lower sensitivity). By default, the web server proposes 3 different α/β ratio's corresponding to different trade-offs between sensitivity and specificity. A region that is identified with a certain threshold and that thus corresponds to

the most reliable signal in the data will by definition also be identified by the less stringent α/β ratio. Providing the results obtained with different thresholds thus allows the user to assess the reliability of the predictions based on the stringency of the threshold at which the linked region was first detected. The effect of changes in the α/β ratio can be observed visually in the graphical interface.

The emission probability of markers to be in the neutral state (i.e. not linked to the trait of interest) is directly estimated from the count data uploaded by the user, hereby assuming that most of the markers are not linked to the phenotype and their count data thus display the allele frequency distribution expected under random segregation. The web server also displays the distribution inferred from the real counts, allowing the user to select a set of running parameters in agreement with the data to be analyzed.

OUTPUT

The EXPLoRA web service determines per α/β ratio the posterior probability of each marker site to be linked to the phenotype of interest. Because of LD, neighboring markers on the chromosome will together display either high or low posterior probabilities. Consecutive sets of neighboring markers displaying high posterior probabilities are thus used to identify the QTL. The server displays the marker-specific posterior probability scores graphically (Figure 2.8D) along the chromosome for each α/β combination together with the observed allele frequencies at the variant sites and accordingly derives the genomic regions covered by the identified phenotype-linked markers. Results can be downloaded in two formats: 1) as a comma separated file, listing the posterior distributions per marker and parameter setting, and 2) as a BED file containing the genomic information to identify the regions found to be linked to the phenotype. The BED file can be imported into other tools such as the Integrative Genomic Viewer (Robinson et al. 2011).

IMPLEMENTATION

The web service was implemented in Java and connects to a MySQL database for information storage. The software was built using the model view controller pattern implemented via Tapestry 5.2. The access to the database was achieved using Hibernate for object-relational mapping (ORM). The graphical interface was made using bootstrap, jQuery and D3js. The server runs on a 16-core, 64bit CentOS 6.2 system with 128GB of memory.

DUSCUSSION

In contrast to previously applied single locus models (Swinnen, Thevelein, et al. 2012; Birkeland et al. 2010), most state-of-the-art methods to analyse the results of BSA exploit the dependencies between neighbouring sites to better distinguish truly from spuriously linked regions. Whereas classical data-driven statistical approaches fit a complex smoothing function to the data to facilitate the identification of patterns in the relative variant frequency plots, EXPLoRA explicitly models linkage disequilibrium to explain the observed patterns in the data, which allows to compensate for noise caused by sampling and sequencing errors, and for the low statistical power in case of small pools or incomplete QTL effects. This was clearly illustrated in the simulation experiments where under conditions that become restrictive for a state-of-the-art statistical method such as the one of Magwene et al. EXPLoRA was still able achieve a high recovery rate while keeping a permissible low number of falsely linked regions.

A similar philosophy as the one adopted by EXPLoRA is also used in the recently published methods MULTIPOOL (Edwards & Gifford 2012) and the model of (Leshchiner et al. 2012). However, results obtained with EXPLoRA on simulated data show that the specific way in which EXPLoRA models the effect of LD results in efficiently identifying phenotype-linked regions, even at low signal/noise ratio's. These results were confirmed by reanalyzing a real dataset in which EXPLoRA was indeed able to detect additional QTLs in the 17% pool

that were confirmed by the 16% pool despite the much lower number of segregants in this 17% pool. It was also able to recover for both pools a minor allele (in QTL2) for which the true contribution to ethanol tolerance was confirmed by experimentally identifying its causal gene.

To improve the usability of the method, we developed EXPLoRA web, a novel web service. The use of standard formats such as the variant calls format (VCF) allows users to upload directly the outputs of software pipelines to obtain allele counts and genotype calls from whole genome sequencing data such as NGSEP (Duitama et al. 2014) or GATK (Van der Auwera et al. 2013) as inputs for QTL analysis. With EXPLoRA we anticipate on the increasing use of BSA in combination with pooled sequencing both for fundamental and applied purposes and on the concomitant need for user friendly applications to facilitate its complex data analysis activities.

CONCLUSION

By using linkage disequilibrium to model the dependency between neighboring marker sites, EXPLoRA allows to reliably detect QTLs using bulk-segregant whole genome sequencing data. Results obtained with both simulated and experimental data show that EXPLoRA displays superior performance under conditions with a low signal to noise level (e.g. small selected pool size, sampling errors, incomplete QTL effects e.g. by the contribution of multiple minor alleles).

Chaper 3. FREQUENCY-BASED HAPLOTYPE RECONSTRUCTION FROM DEEP SEQUENCING DATA OF BACTERIAL POPULATIONS

This chapter contain the manuscript Pulido-Tamayo, S., et al., 2015. Frequency-based haplotype reconstruction from deep sequencing data of bacterial populations. *Nucleic acids research*.

ABSTRACT

Clonal populations accumulate mutations over time, resulting in different haplotypes. Deep sequencing of such a population in principle provides information to reconstruct these haplotypes and the frequency at which the haplotypes occur. However, this reconstruction is technically not trivial, especially not in clonal systems with a relatively low mutation frequency. The low number of segregating sites in those systems adds ambiguity to the haplotype phasing and thus obviates the reconstruction of genome-wide haplotypes based on sequence overlap information.

Therefore, we present EVORhA, a haplotype reconstruction method that complements phasing information in the non-empty read overlap with the frequency estimations of inferred local haplotypes. As was shown with simulated data, as soon as read lengths and/or mutation rates become restrictive for state-of-the-art methods, the use of this additional frequency information allows EVORhA to still reliably reconstruct genome-wide haplotypes. On real data, we show the applicability of the method in reconstructing the population composition of evolved bacterial populations and in decomposing mixed bacterial infections from clinical samples.

INTRODUCTION

The genetic heterogeneity of clonal populations is key to their adaptive behavior. Environment-specific genes, subject to relaxed selection in a non-inducing environment, build up cryptic variation, that enhances the adaptive potential (Pfennig et al. 2010; Hayden et al. 2011). Even when starting evolution from a single clone (haplotype) under severe selection pressure, the combination of mutation rate and population size appears to be sufficiently high to build up genetic variation in the population (Barrick & Lenski 2009; Lang et al. 2011), resulting in a mixture of closely related haplotypes (or quasispecies). As a result, a population is not genetically uniform most of the time (Kao & Sherlock 2008). Although single cell sequencing would be ideal to determine the composition of such a heterogeneous population, it is still technically very difficult and cost-inefficient (Heywood et al. 2011; Stepanauskas 2012; Lasken & McLean 2014). However, deep sequencing a clonal population in its entirety, referred to as pooled or metagenomic sequencing (Barrick & Lenski 2013) inherently contains information to determine the haplotypic variation of the population, i.e., the identity of the occurring haplotypes and their frequencies.

However, resolving haplotypes from deep sequencing data of clonal populations, also referred to as haplotype reconstruction or quasi-species assembly is technically not trivial and methods to do so are still lacking for most clonal species, other than viruses.

At first because the problem of error correction is confounded with the haplotype reconstruction itself (Beerenwinkel et al. 2012) and therefore error correction and haplotype reconstruction should ideally be performed simultaneously. The reconstruction problem itself is non-trivial either. For this reconstruction step all current haplotype reconstruction methods rely on the presence of a sufficient number of segregating sites and relatively long reads

to allow phasing the segregating polymorphic sites into unique haplotypes using either single end (Zagordi et al. 2011; Astrovskaya et al. 2011; Prosperi & Salemi 2012; Huang et al. 2011; Prabhakaran et al. 2013; Töpfer et al. 2013) or paired end read information (Töpfer et al. 2014). This strategy implies that most reads should contain segregating sites and that a sufficient amount of overlap between reads is available to resolve the reconstruction problem. As a result, current methods are restricted to haplotype reconstruction from relatively long-read population sequencing (mainly Roche 454) of clonal organisms with a high mutation frequency (Giallonardo et al. 2014), such as viruses: a high mutation frequency guarantees a large number of segregating sites and long-read based sequencing allows for a large degree of overlap between the reads.

However, for most clonal populations the number of observed segregating sites is much lower than what is observed in viral populations. In a bacterial setting, for instance, haplotypes consist of millions of base pairs corresponding to the size of a bacterial genome, but populations typically contain less than a few hundreds of mutations even in the presence of a mutator phenotype (e.g., bacterial populations originating from a mutator phenotype accumulate after 300 generations approximately 1000 mutations). This relatively low mutation frequency implies an average distance between segregating sites that is in the order of kilobases, which is a lot larger than the maximal read length for Illumina and Roche 454 technologies. Due to the lack of segregating sites, phasing becomes extremely difficult. As a result, state-of-the-art viral haplotype reconstruction methods cannot infer haplotypes from bacterial population samples.

With EVORhA (EVOlutionary Reconstruction of hAplotypes) we propose to the best of our knowledge the first bacterial haplotype reconstruction method. EVORhA combines local haplotype inference with error correction and uses a

probabilistic approach for the genome-wide reconstruction. Key to the method is the use of the inferred frequency ratios of the contributing haplotypes to improve the extension of local haplotypes into genome-wide ones, particularly in those cases where the non-empty overlap between reads does not allow for a non-ambiguous phasing or where partially reconstructed regions have no sequence overlap at all. Because of this key step EVORhA is applicable to the analysis of pooled sequence data obtained from populations of clonal organisms with a low mutation frequency and/or to data obtained with a short-read based technology. We demonstrated the performance of EVORhA under different settings on simulated data. In addition, we showed its ability to reconstruct genome-wide haplotypes in a real setting by analyzing data obtained from a mixed bacterial infection and from pooled sequence samples of an evolving bacterial population. The implementation can be downloaded from http://bioinformatics.intec.ugent.be/kmarchal/EVORhA/. The source code is available upon request.

Material and Methods

EVORhA

Our method uses a two-step procedure: the first step reconstructs haplotypes at the local level and joins locally reconstructed haplotypes into so-called extended haplotypes based on information contained within the read overlap. The second step assigns these extended haplotypes to haplotype sets by using a mixture model of Gaussian distributions that describes for each set the frequency at which the haplotypes assigned to the set are observed. These haplotype sets are eventually joined into genome-wide haplotypes, following a procedure that explicitly assumes that the different haplotypes in the population have evolved from a common ancestor by clonal reproduction. The procedure is outlined in Figure 3.1.

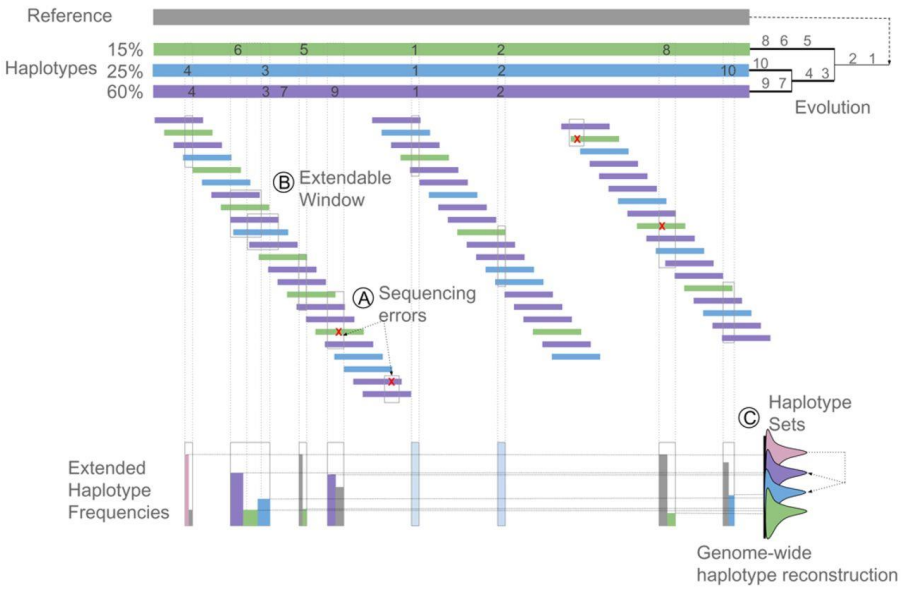


Figure 3.1 Method Overview.
A clonal community where three descendant haplotypes evolved from a reference genome is depicted. The haplotypes are present at 60%, 25% and 15% in the population. A total of 10 different mutations accumulated in the evolving population. Mutations 1 and 2 were acquired by the last common ancestor of extant haplotypes and are therefore shared by all haplotypes in the population. Mutations 3 and 4 were acquired before the origin of the blue and purple haplotypes, whereas the remaining mutations are unique to one of the haplotypes. Small colored horizontal bars represent the reads obtained by deep sequencing the aforementioned population. Step 1: Local haplotype reconstruction and window extension. (A) Haplotype templates and their frequencies are first inferred per window. Windows are represented by gray rectangles. Per window, accepted template windows will be defined by performing a local haplotype reconstruction simultaneously with the error correction (see Supp Figure B. 1). (B) Template haplotypes are extended over flanking windows with overlapping polymorphic sites based on the consistency in the polymorphisms present in the non-empty read overlap (see Supp Figure B. 2). (C) Step 2: Genome-wide haplotype reconstruction. Extended haplotypes from the different concatenated windows will be merged into genome-wide haplotypes based on the frequency information (see also Supp Figure B. 3).

STEP 1: LOCAL HAPLOTYPE RECONSTRUCTION AND WINDOW EXTENSION

Window definition

A window is defined as a genomic region that is covered by a sufficient number of reads and that contains a set of one or more consecutive tentative polymorphic sites. As at this point, variation observed at these sites can refer to both sequencing errors and true polymorphisms, we refer to them as 'tentative'. All windows that contain a unique combination of consecutive (tentative) polymorphic sites are enumerated, with the restriction that windows should be shorter than a pre-specified maximal window length (60% of the read length by default) and that windows should be entirely covered by a certain minimum number of reads (30 reads by default). Regions that are not covered by the minimal number of reads will be ignored. Both window-defining parameters could be adjusted if necessary.

Inferring template haplotypes and their frequencies per window

A list of possible template haplotypes is generated per window (see Figure 1S). A template haplotype is defined as a unique combination of one or more consecutive (tentative) polymorphisms observed in at least one of the reads that fully covers the window. For each template haplotype $\boldsymbol{h} = \{h_1, h_2, \dots\}$ found in window W a corresponding 'support' $\boldsymbol{\tau} = \{\tau_1, \tau_2, \dots\}$ is based on all reads that are consistent with that template haplotype. First, we consider only the reads that fully overlap with the window and calculate a *base support* τ_i^0 for template haplotype h_i as follows:

$$\tau_i^0 = \sum_{r \in F_i} w(r)$$

where F_i denotes the set of reads that fully overlap with the window and that are consistent with template haplotype h_i and where $w(r)$ is given by

$$w(r) = \min_{j=1,l} P(r[j])$$

with $P(r[j])$ the base call accuracy at tentative polymorphic site j of read r containing l tentative polymorphisms. It is related to the Phred quality score $Q(r[j])$ as follows:

$$P(r[j]) = 1 - 10^{\frac{-Q(r[j])}{10}}$$

We choose w(r) to depend on min $P(r[j])$, assuming that the contribution of the read to the support depends on its lowest quality polymorphism. This allows to have a scoring independent of the window length.

Additional support τ_i^1 for template haplotype h_i is derived from reads that only partially overlap with window W.

$$\tau_i^1 = \sum_{r \in P_i} \frac{\tau_i^0}{\sum_{j|r \in P_j} \tau_j^0} w(r)$$

Where P_i denotes the set of reads that partially overlap with the window and that are consistent with template haplotype h_i. Note that partially overlapping reads can be consistent with multiple template haplotypes. The fraction within the summation therefore denotes that a given read gives a support to the haplotype it is compatible with, weighted proportionally to the base support of that haplotype. The total support of a template haplotype τ_i is then given by

$$\tau_i = \tau_i^0 + \tau_i^1$$

The template haplotypes $\boldsymbol{h}$ might contain errors, i.e., tentative polymorphisms that arose due to sequencing errors. To prune templates in the windows, we retain per window only the template haplotypes with support τ_i greater than

the template haplotype threshold. The threshold for the support is different for different templates and is determined by the following equation:

$$\tau_{\text{threshold}} = -11.86 + 4.24\ln(f) + E(c)$$

$$E(c) = \begin{cases} 0 \; if \; InterGenicRegion \\ 0 \; if \; BLOSUM \geq 0 \\ -1 \times BLOSUM \; if \; BLOSUM \; < 0 \end{cases}$$

Where f represents the average fold coverage in the considered window to which a polymorphism belongs and c the codon in the same window with the lowest BLOSUM score. The BLOSUM matrix used is obtained by comparing already calculated matrices against the analysed data. In most cases, the most similar matrix is the BLOSUM100. The equation describes that the minimal threshold on the support for accepting template haplotypes will increase with the coverage of the window and with the severity of the amino acid changes induced by the polymorphisms, i.e., template haplotypes with a higher coverage and more unlikely amino acid changes need more support to be retained. The parameters in this formula, being the contribution of the window coverage and the codon changes, were determined by maximizing the accuracy of reconstructing true haplotypes in a simulated setting (see simulated data).

A haplotype with a support below the threshold is assumed to be a 'false positive' and will no longer be considered as a possible template. Reads that contributed to the support of a rejected template will be reassigned to the accepted template haplotype that is evolutionary most related to it (using the BLOSUM matrix mentioned above). Note that sequence error correction is performed simultaneously with haplotype reconstruction: rather than filtering upfront tentative polymorphisms that occur infrequently (i.e., standard error correction), polymorphisms are filtered when they belong to template

haplotypes with insufficient support. This prevents the deletion of infrequently observed polymorphisms when they belong to a template haplotype with sufficient support.

Window extension: concatenating windows that share polymorphisms

Here, we start with a set of windows and their respective accepted template haplotypes. Some windows will share polymorphic sites and will be extended (see Figure 2S). The extension is performed window by window, starting from a so called seed window which corresponds to the window with the largest number of 'accepted' template haplotypes, the largest number of polymorphic sites and the highest coverage. If no window meets all criteria simultaneously, we select the seed window by prioritizing first on the number of template haplotypes, then on the number of polymorphic sites and, lastly, on coverage. The goal of the extension is to find the best combination of haplotypes from the first and second window that can be concatenated to generate an extended haplotype, where 'best' is defined in terms of matching frequencies and shared polymorphisms.

The extension, which is conceptually very similar to what is referred to as the 'global reconstruction' in graph-based haplotype reconstruction approaches (Prosperi & Salemi 2012; Töpfer et al. 2013; Zagordi et al. 2011; Astrovskaya et al. 2011), is performed as follows: for each pair of overlapping windows $W = \{W_a, W_b\}$, a set of *groups* G is declared where one group g_i is defined per unique combination of consecutive polymorphisms in the overlap of both windows. The template haplotypes in the windows are then assigned to those groups. Note that a specific template haplotype can only be assigned to a single group; however, a certain group can contain multiple template haplotypes. We can now distinguish three cases:

(1) If a group contains a single template haplotype in one window and at least one template haplotype in the other window, the extension is straightforward

and the extended haplotypes consist of the concatenation of the sequences of the template haplotypes in both windows.

(2) If a group contains multiple template haplotypes in both windows, the extension is ambiguous. In that case, first, it is assumed that the number of extended haplotypes equals the maximum number of template haplotypes present in either window. The assignment of template haplotypes to the extended haplotypes and the frequencies of the extended haplotypes $\theta = \{\theta_1, \theta_2, \ldots\}$ are determined using an expectation maximization algorithm. First, the frequencies θ are initialized randomly. During the expectation step, template haplotypes in both windows are assigned to the extended haplotypes such that the observed frequencies f of the template haplotypes best match the frequencies θ. In case a template haplotype can be assigned to multiple extended haplotypes, the frequency of the template haplotype is split according to the frequencies θ of the extended haplotypes to which it was assigned. All possible combinations of assignments are exhaustively enumerated and the one that maximizes the log-likelihood according to the Poisson distribution is selected:

$$L = \sum_{x \in X} \left[\log \frac{{\lambda_{x,a}}^{k_{x,a}}}{k_{x,a}!} e^{-\lambda_{x,a}} + \log \frac{{\lambda_{x,b}}^{k_{x,b}}}{k_{x,b}!} e^{-\lambda_{x,b}} \right]$$

where $\lambda_{x,a} = c_a \theta_x$ and $\lambda_{x,b} = c_b \theta_x$ denote the expected number of reads matching extended haplotype x in windows a and b, respectively. Here, c_a and c_b are the coverages in windows a and b, respectively and θ_x is the frequency of extended haplotype x. Similarly, $k_{x,a} = c_a f_x$ and $k_{x,b} = c_b f_x$ denote the observed number of reads matching extended haplotype x in windows a and b, respectively, with f_x being the observed frequency of the template haplotypes.

The maximization step then computes the new frequencies θ for extended haplotypes by computing the average frequencies of the contributing template haplotypes.

This process is repeated until the likelihood difference between consecutive iterations becomes sufficiently small or until a maximum number of iterations has been reached. We perform multiple starts with random initial frequencies to avoid local maxima.

(3) If a certain group only contains template haplotypes from one window (not both), the template haplotypes in the group are moved to a different group that contains the haplotypes for which the evolutionary distance (BLOSUM) to the haplotypes under consideration is the smallest. This situation can occur with low frequency haplotypes where reads derived from these haplotype might not be available for all windows. After reassigning the haplotypes to another the procedure is as described in (1) and (2).

The concatenated window containing the extended haplotypes is subsequently used as a seed to concatenate the next set of flanking windows. If no more flanking windows exist for the current concatenated window, a new initial window is defined. The procedure continues until all windows that display overlapping polymorphisms have been concatenated.

The window extension thus produces a set of extended windows and their respective extended haplotypes. An extended window by definition does not share any polymorphic sites with other windows and cannot be further extended (phased) by analysing read overlap.

STEP 2: GENOME-WIDE HAPLOTYPE RECONSTRUCTION

Extended haplotypes are those that can no further be concatenated, because they do not longer contain polymorphisms in their read overlap. This occurs

when either the read overlap between two extended haplotypes is empty or non-informative. The latter situation arises in case of low mutation frequency when the genomic distances between segregating sites are usually larger than the median read length. To compensate for this lack of information, we will use the frequency information of each of these extended haplotypes, to infer a possible genome-wide haplotype. To combine extended haplotypes into genome-wide haplotypes, we first perform a frequency analysis by grouping together extended haplotypes that occur at similar frequency (referred to as haplotype sets) and subsequently use a power set approach to search for a final genome-wide haplotype that can explain the frequencies of the observed sets of 'extended haplotypes'.

Frequency Analysis:

Let $\boldsymbol{R} = \{R_1, R_2, \dots, R_n\}$ denote the number of reads that correspond to the extended haplotypes $\boldsymbol{h} = \{h_1, h_2, \dots, h_n\}$, observed in a concatenated window. $\boldsymbol{R}$ then follows a multinomial distribution:

$$\boldsymbol{R} \sim \text{Multi}(C, \boldsymbol{P})$$

where C denotes the number of reads observed in the concatenated window and $\boldsymbol{P} = \{P_1, P_2, \dots, P_n\}$ denotes the true frequencies of the haplotypes. If the window coverage is sufficiently high, the multinomial distribution $\boldsymbol{R}$ can be approximated by n normal distributions:

$$R_i = CX_i \sim \text{N}(CP_i, CP_i(1 - P_i))$$

where $\boldsymbol{X} = \{X_1, X_2, \dots, X_n\}$ are the observed frequencies of the extended haplotypes in the concatenated window. It then follows that:

$$X_i \sim \text{N}\left(P_i, \frac{P_i(1 - P_i)}{C}\right)$$

Given that the extended haplotypes and the frequencies $\boldsymbol{X}$ at which they are observed, should be consistent over at least several windows, the frequencies at which extended haplotypes are observed in each of the windows can be assumed to be generated by a mixture model of Gaussian distributions (see Figure 3S). If polymorphisms occurring at the same site are shared by different haplotypes, they will occur at a frequency (or Gaussian) different than polymorphisms that are unique to a single haplotype.

This mixture model is inferred as follows: the method starts by assigning one Gaussian distribution to each of the extended haplotypes observed in an initially selected concatenated window (using the same initialization criteria defined above i.e. the concatenated window with the largest number of 'accepted' template haplotypes, the largest number of polymorphisms and the highest coverage is selected as a seed. If no window meets all criteria simultaneously, we select the seed prioritizing first on the number of template haplotypes, then on the number of polymorphisms and, lastly, on coverage). Subsequently we assign per concatenated window each extended haplotype to the Gaussian that currently best explains its observed frequency provided the mean of this Gaussian is located less than one standard deviation from the observed haplotype frequency. If the mean of the best explaining distribution is more than one standard deviation away from the observed frequency of the given haplotype, we create a new Gaussian in the mixture model and we continue until all haplotypes have been assigned to a Gaussian distribution that is less than one standard deviation away from the observed frequency of the given haplotype. The resulting model is referred to as the full mixture model.

As this is a very relaxed way of extending the mixture, we include a final Bregman hierarchical clustering step (Banerjee et al. 2005) to reduce the model complexity and find an optimal mixture model for which the difference

with the full mixture model in fitting the observed frequencies of the haplotypes is less than 1%. By considering the frequencies of all extended haplotypes in all extended windows, the frequency analysis results in a mixture model of Gaussian distributions that groups all extended haplotypes, occurring at a similar frequency in haplotype sets.

Inferring the final genome-wide haplotype

Each inferred distribution in the mixture model corresponds to a set of extended haplotypes that are likely to co-occur in one or more genome-wide haplotypes (referred to as a haplotype set). At this moment each true haplotype can still be characterized by several Gaussians from the mixture (see Figure 3S). This is because haplotype sets containing polymorphisms unique for the haplotype will occur at a lower frequency than haplotype sets that contain polymorphisms shared by several genome-wide haplotypes.

To join haplotype sets that can safely be assumed to belong to the same haplotype we use the following approach: a distance matrix D is calculated between all pairs i, j of haplotype sets, where $D_{i,j} = \left|P_i \cup P_j\right| - \left|P_i \cap P_j\right|$ and P_i, P_j are the polymorphisms composing each haplotype set, respectively. Obviously, different polymorphisms at the same polymorphic site are considered as different objects in the haplotype sets.

Subsequently, for each haplotype set h from the full list of haplotype sets $\mathcal{H}$ (always starting with the haplotype set with the highest observed frequency) we construct a subset $\mathcal{H}_h = \{g \in \mathcal{H} | \mu_g < \mu_h \wedge D_{h,g} = 0\}$ where μ_g and μ_h are the means of the Gaussian distributions representing haplotypes sets g and h, respectively. Then, for the power set $P(\mathcal{H}_h)$ we construct $P'(\mathcal{H}_h) = \{\omega \in P(\mathcal{H}_h) | f(\omega) < 2 \times \sigma_h\}$ where σ_h is the standard deviation of the Gaussian distribution representing haplotype set h (i.e. those subsets where

the sum of the frequencies of the haplotype sets is in the 95% confidence interval of the Gaussian distribution of h):

$$\varpi = \underset{\omega \in P'(\mathcal{H}_h)}{\arg\min}\, f(\omega)$$

$$f(\omega) = \left[\mu_g - \sum_{i \in \omega} \mu_i\right]$$

If ϖ exists, we conclude that haplotype set h contains a set of polymorphisms shared by all haplotypes in ϖ and therefore will no longer be considered as an individual haplotype. Therefore, we remove h from the list of haplotypes and add the polymorphisms in h to all haplotype sets in ϖ. This final step in the analysis results in the reconstructed genome-wide haplotypes, their inferred frequencies and polymorphisms.

The following running parameters are used for EVORhA by default:

The parameters of the method are those that define valid windows, being the 'maximal window length' and the 'minimal read coverage'. For the minimal read coverage a default value of 30 reads was chosen. This, to guarantee a sufficient number of reads in each window so that the distribution of the reads corresponding to template haplotypes in the window can be approximated by a multinomial, and therefore can be modelled as a mixture model of Gaussian distributions, such as described in step 2. The default of the maximum window length was chosen at 60% of the read length so that the template haplotypes and their *base support*, both derived from reads that fully overlap with the window, are representative for the true haplotypes present in the window.

SIMULATION EXPERIMENTS

PERFORMANCE ASSESSMENT OF EVORHA

To test EVORhA, a first set of population sequence data was generated using the sequence of *Salmonella Typhimurium* 14028S (Accession number CP001362) as the ancestral reference strain. In the simulated populations, the number of polymorphisms varied between 100, 1000 and 2500, the number of haplotypes varied from 2 to 7 and their frequencies were set randomly. For each simulated population a random phylogeny was constructed prior to assigning polymorphisms to haplotypes. Polymorphisms were added at a random branch of the phylogeny and propagated to all haplotypes descending from that branch, ensuring the simulation of evolutionary related haplotypes (assuming that the different haplotypes in the population have developed by clonal reproduction from a common ancestor). For each of these populations we simulated reads at the polymorphic sites for different sequence coverages (ranging between 50, 200 and 500) and using a sequencing error probability of 1%. Per parameter combination (number of haplotypes, number of mutations and coverage), 100 datasets were generated.

The degree to which an inferred haplotype could correctly be reconstructed was assessed by a 'reliability score', which is the proportion of shared polymorphisms between the reconstructed haplotype and its most similar true counterpart i.e. the simulated haplotype.

$$Reliability = \frac{|P_h \cap P_s|}{|P_h| + |P_s|}$$

where P_h and P_s are the polymorphisms present in the reconstructed haplotype and its most similar true counterpart , respectively.

To test the extent to which the true frequencies of the reconstructed haplotypes could be inferred, we used the mean absolute error (MAE)

between the true frequencies at which a haplotypes occurred (y_i) in the simulated population and the estimated frequencies of the matching reconstructed haplotypes ($\hat{y}_i$) divided by the number of true haplotypes (N).

$$MAE = \frac{1}{N}\sum(|y_i - \hat{y}_i|)$$

COMPARISON WITH STATE-OF-THE-ART HAPLOTYPE RECONSTRUCTION METHODS

To perform a comparison with ShoRAH (Zagordi et al. 2011), QuasiRecomb (Töpfer et al. 2013) and Predicthaplo (Prabhakaran et al. 2013) we simulated a second set, this time consisting of raw population sequence data (as we need an alignment file as input for each of the respective algorithms we compared with). Raw datasets were generated using GemSIM v1.6 (McElroy et al. 2012), derived from a single gene of 2562 bp long, flanked on both sides by 700 bp regions. The number of haplotypes ranged between 2 and 4. The number of polymorphisms in the simulation was varied between 7, 10, 20 and 50. The same phylogenetic approach mentioned above was used to generate evolutionary related haplotypes. Raw data mimicked 100 and 700 bp reads, produced under respectively an Illumina and a Roche 454 error model provided by the simulator. The coverage varied between 50, 100, 200 and 500. One hundred (100) datasets were generated for each combination of parameters. Note that we focused on simulating one gene rather than a full bacterial genome in order to design a set up optimized for the state-of-the-art methods we intended to compare with (as these cannot handle a genome-wide haplotype reconstruction).

The latest version of ShoRAH was downloaded from: http://www.bsse.ethz.ch/cbg/software/shorah. ShoRAH was run according to the authors' recommendations with a window size that is about one third of the read length, i.e. with a window of 30 bp for simulations with a read length of 100 bp reads and of 252 bp for simulations with read lengths of 700 bp. The

latest version of PredictHaplo was obtained from: http://bmda.cs.unibas.ch/HivHaploTyper/. In our hands, most of the simulated datasets could not be processed with PredictHaplo, preventing us from comparing its performance with that of the other methods. The latest version of QuasiRecomb was obtained from: http://www.silva.bsse.ethz.ch/cbg/software/quasirecomb. The method was run with flags "noRecomb", "conservative" and "unpaired" as recommended by the authors for a comparable setting . All tools were run on the same simulated datasets.

HAPLOTYPE RECONSTRUCTION TO INFER EVOLUTIONARY TRAJECTORIES

The data used for haplotype reconstruction during bacterial evolution was generated as follows: *Escherichia coli* SX4 was grown under selective pressure (high concentration of ethanol) in a serial transfer experiment in which the concentration of ethanol was gradually increased over time (0.5 % each time) as soon as the population resumed growth under a current selection pressure. At three consecutive time points, population samples and individual clones, selected from these sampled populations were subjected to Illumina sequencing HiSeq2000 (using 100 bp paired end read mode, with a coverage of approximately 200 fold for the population samples). Sampling and DNA isolation were done according to standard procedures (Qiagen® Blood & Tissue kit). Sequences are stored under BioProject PRJNA262000. Read mapping of both the sequenced pooled samples and individual clones was performed with Burrows–Wheeler Aligner BWA-MEM using the sequence of the original unevolved ancestral clone as a reference. For the sampled clones variants were called after alignment using SAMtools (Li et al. 2009) with default parameters. As a control we confirmed that the called variants were also obtained using the CLC Bio pipeline (http://www.clcbio.com) with default parameters.

The true haplotypes of the individual clones sampled from the pool were compared with the haplotypes reconstructed from the pooled data. To this end we calculated the ratio of the number of polymorphisms shared between an individual clone and its best matching reconstructed haplotype versus their total number of polymorphisms. Polymorphisms that reached fixation in the population were not taken into account as they are present in all haplotypes and therefore not informative.

The phylogenetic relations between the haplotypes reconstructed at different time points was inferred using a Levenshtein distance measure between a haplotype observed at a current time point and the ones observed at the closest preceding time point, hereby assuming that the closest haplotype in the preceding time point is the evolutionary ancestor of the haplotype observed at the current time point.

Haplotype reconstruction to identify mixed infections

Publicly available read mapping data of in-vitro mixed infections was obtained from Eyre et al. (Eyre et al. 2013). They generated 36 mixed infections by pairwisely combining the DNA obtained from different clones in 3 different proportions – 50%/50%, 70%/30% and 90%/10%. For each proportion 12 mixed infections were generated, each consisting of different pairwise combinations of clones. The pools of the in vitro generated mixed infections were subjected to Illumina sequencing at 150 fold coverage. We performed a genome-wide reconstruction using all variant loci detected in the population sequencing data of the mixed infection and, as an alternative, we also performed a reconstruction by using a preselected set of 151 polymorphic sites present in 3 different genes as outlined in Eyre et al. (Eyre et al. 2013). For both reconstructions the Mean Squared Error (RMSE) was calculated to assess the correctness of predicting the correct haplotype frequencies in the mixed infection. We use the RMSE rather than the above mentioned MAE to be

consistent with the original paper of Eyre et al. (Eyre et al. 2013). The reliability of the reconstruction was assessed as described in section 'performance assessment' by comparing the polymorphisms present in the reconstructed haplotypes with the polymorphisms observed in the sequences of the single clones.

RESULTS

Our method consist of two steps: a first step comprising a local haplotype reconstruction followed by a window extension, in which haplotypes are defined at the local level, sequencing errors are removed and overlapping regions sharing polymorphisms are extended into longer haplotypes; a second genome-wide reconstruction, during which the final haplotypes and their relative frequencies are inferred by using the frequency observations of the extended haplotypes.

Based on concepts developed in the context of viral haplotype reconstruction (Zagordi et al. 2011; Astrovskaya et al. 2011; Töpfer et al. 2013; Prabhakaran et al. 2013), the first step of our method performs the error correction simultaneously with the haplotype inference on a local scale (Fig 1 A and B). The local scale is defined by a set of consecutive polymorphic sites that map on a single genomic region of the reference genome and that are 'covered' by a sufficient number of reads (referred to as a window). The simultaneous estimation of the sequence errors with the local haplotype inference is based on a method that iterates between assigning reads to haplotypes and using these read-to-haplotype assignments to infer the probabilities that either the observed reads were the result of random sequence error or originated through mutations in the ancestral genome. We used information contained in BLOSUM substitution matrices to lower the allowance of mutations in coding regions, occurring rarely in nature. This local reconstruction step results per genomic window in haplotypes and their observed frequencies in the

pooled sample (referred to as local haplotypes, consistent with the literature (Beerenwinkel et al. 2012)). Subsequently, these local haplotypes are extended using a heuristic approach that takes into account both phasing information in the non-empty read overlap between flanking windows, and also the inferred local haplotype frequencies. Because of the sparse number of expected segregating sites between the individuals in the population, the haplotype extension step results in slightly larger contigs, but rarely covers more than a few hundred base pairs.

Therefore, in the second step, referred to as the genome-wide haplotype reconstruction, extended haplotypes are joined into genome-wide haplotypes that ideally cover a full haplotype in the population (Fig 1 C). This step is entirely dependent on the frequency information of the extended haplotypes: sets of extended haplotypes that occur at a similar frequency in the population and that do not show any inconsistencies in their polymorphisms (i.e. do not have a different mutation at exactly the same genomic position) are assumed to belong to the same genome-wide haplotypes. This genome-wide haplotype reconstruction step is solved by first estimating sets of locally extended haplotypes that occur at a similar frequency and subsequently searching for the set of genome-wide haplotypes and their frequencies that best explain the observed frequencies of the extended haplotypes. This latter step assumes that the haplotypes in the population have developed by clonal reproduction from a common ancestor and therefore haplotype sets that are shared by at least two genome-wide haplotypes should occur at a frequency in the population that approximates the sum of the frequencies of each of the individual genome-wide haplotypes containing the shared haplotype set.

PERFORMANCE OF EVORHA ON SIMULATED DATA

To test the performance of EVORhA in reconstructing haplotypes, whole genome sequencing datasets were simulated for clonal populations, differing

from each other in the number of haplotypes (ranging from 2-7), the frequencies at which the haplotypes occur in each of these populations, the number of polymorphisms (ranging between 100, 1000 and 2500) and the sequencing coverage (ranging between 50, 200 and 500 fold). For each simulation setting, 100 simulations were performed.

To assess the reliability of the reconstruction, we compared the reconstructed haplotypes with the simulated ones. At the same time we assessed the ability of the reconstruction to correctly estimate the true frequencies at which the haplotypes occurred in the population. As expected, the reliability of the reconstruction increases with the coverage and this effect was most pronounced for haplotypes occurring at low frequencies (Figure 3.2 panel A), as especially for those haplotypes an increase in coverage has a relatively larger effect on the ability to distinguish a true polymorphism from a sequencing error. Figure 3.2 panel A also shows that even at low coverage (50 fold), haplotypes were reconstructed with an average reliability of 70%. For the same coverage the performance improves with a decrease in the complexity of the pool (less polymorphisms and less haplotypes), with a maximum of 92% in average reliability observed for the least complex problem (100 polymorphisms, 2 haplotypes, at the highest coverage of 500 fold) (Figure 4S).

Figure 4S shows for the different simulated set ups, the degree to which EVORhA could correctly estimate the true haplotype frequencies in the population (expressed by the Mean Absolute Error - MAE). The ability to estimate true frequencies seems largely independent of the number of haplotypes in the population or the number of polymorphisms. The latter is to be expected as the total number of polymorphisms in the simulation is sparse anyhow and does confer little information to the final frequency estimation. Figure 3.2 panel B shows how the true frequency estimation is largely affected

by the coverage: when the coverage is low (50 fold), the average error rate on estimating the true frequency of the haplotypes is around 10%. This is understandable given that at low coverage the sampling that produces the reads is more prone to random effects.

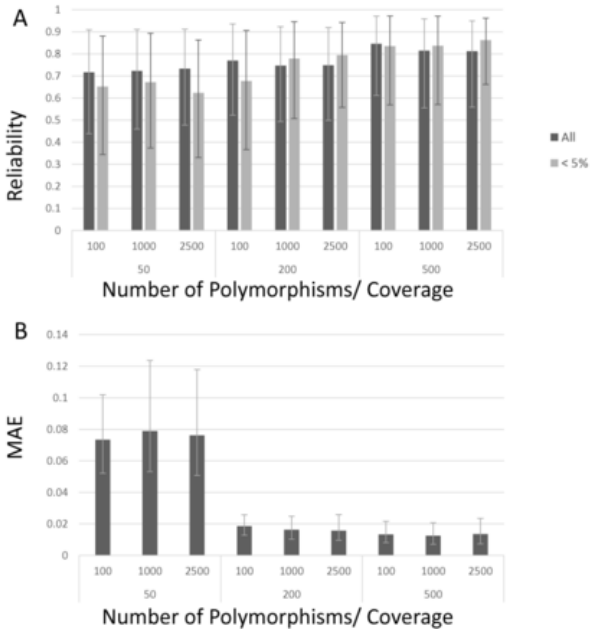


Figure 3.2 Reliability of haplotype reconstruction by EVORhA on simulated data.
The X-axis displays the different combinations of coverage (respectively 50, 200, 500) and number of polymorphisms in the population (respectively 100, 1000, 2500) used for each simulated set up, hereby collapsing the results obtained for simulations with a different number of haplotypes. The degree to which the simulated haplotypes was correctly reconstructed was assessed by the reliability. The correctness of the frequency estimation of the reconstructed haplotypes was assessed by the MAE. (A) Average reliability of the haplotype reconstruction, derived by either considering the results of all reconstructed haplotypes (dark bars) or only the results obtained for haplotypes that occur in the population at a frequency below 5% (light bars). Y-axis: average reliability; values are obtained by averaging the reliability scores of the considered haplotypes resulting from simulations obtained with the same coverage and same number of polymorphisms, irrespective of the number of haplotypes (so showing the average reliability of haplotypes obtained from 500 simulations). Error bars indicate the 90% confidence interval of the reconstruction. (B) Y-axis: MAE of the frequency estimation of all haplotypes resulting from simulations obtained with the same coverage and same number of polymorphisms, irrespective of the number of the haplotypes (see panel (A)). Error bars indicate the MAE 90% confidence interval.

COMPARISON OF EVORHA WITH STATE-OF-THE-ART HAPLOTYPE RECONSTRUCTION

Because our method builds for its initial step on concepts that were first described in the context of viral haplotype reconstruction, we compared our tool with state-of-the-art viral haplotype reconstruction tools. As representatives of read-graph based tools we used ShoRAH (Zagordi et al. 2011) and QuasiRecomb (Töpfer et al. 2013), both widely used for viral haplotype reconstruction. In addition, we used PredictHaplo (Prabhakaran et al. 2013) as a representative of probabilistic haplotype reconstruction methods.

As none of the above mentioned viral haplotype reconstruction tools (ShoRAH, QuasiRecomb and PredictHaplo) was able to run in the bacterial setting, we compared our method in a setting more appropriate for these state-of-the-art tools (reconstruction of viral sized haplotypes in the presence of a large number of polymorphisms). Hereto, we designed a simulation experiment, mimicking the data resulting from the population sequencing of a small region obtained with either a relatively short or long read sequencing technology (respectively mimicking Illumina and Roche 454 reads). Simulated populations differed from each other in the number of haplotypes (ranging from 2 to 4), the used sequence coverage (50, 100, 200 and 500 fold) and the number of polymorphisms in the population (7, 10, 20 and 50 sites). For each experimental setup 100 different datasets were simulated and performances of respectively EVORhA, ShoRAH and QuasiRecomb were assessed as outlined above and in the material and methods.

Figure 3.3 shows that in general, and irrespective of the read length used for sequencing, using an increased sequencing coverage and having intrinsically more polymorphic sites in the population positively influences the performance of all methods, mainly in terms of reliability i.e. correctly

reconstructing the true haplotypes in the population. Only for ShoRAH the reconstruction reliability seems to decrease with the coverage in case haplotypes were obtained from populations with few polymorphisms (<10).

For relatively long reads (700 bp), EVORhA reaches performances similar to those of QuasiRecomb and ShoRAH: in the tested setting the reconstruction reliability obtained with QuasiRecomb was slightly higher than the one obtained with EVORhA, but this came at the expense of QuasiRecomb having a relatively lower performance in terms of the frequency estimation (relatively higher MAE) than EVORhA.

For shorter reads (100 bp) EVORhA consistently outperforms QuasiRecomb and ShoRAH, both in terms of having a higher reconstruction reliability and having a better frequency estimation.

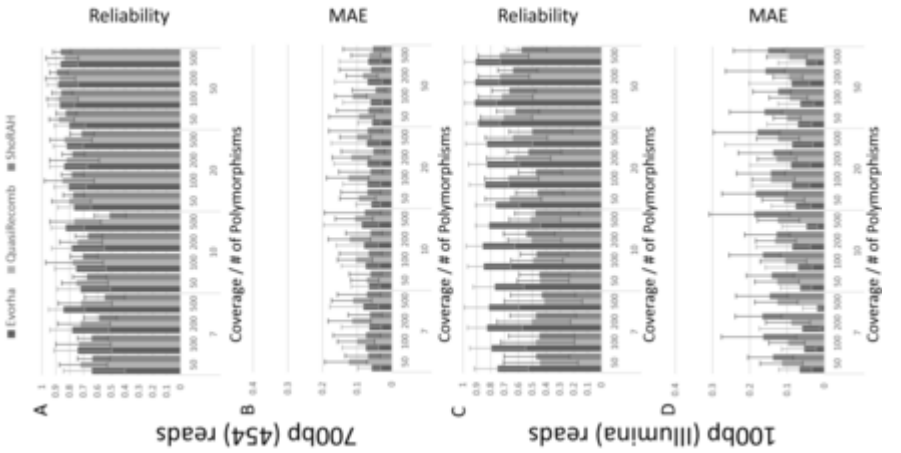


Figure 3.3 Performance comparison of EVORhA, ShoRAH and QuasiRecomb on simulated data.

(A) Comparison of the reconstruction reliability using long read sequencing (700 bp), hereby collapsing the results obtained for simulations with a different number of haplotypes. Data sets were obtained by simulating long read sequencing. The X-axis displays the different combinations of coverage (respectively 50, 100, 200, 500) and number of polymorphisms in the population (respectively 7, 10, 20, 50) used for each experimental set up. The Y-axis shows the average reliability of the reconstruction. Bars indicate the performance per method. Reliability values are obtained by averaging the reliability scores of all haplotypes resulting from simulations obtained with the same coverage and same number of polymorphic sites, irrespective of the number of haplotypes. Error bars indicate the 90% confidence interval of the reconstruction. (B) Comparison of the frequency estimation of the haplotypes using long read sequencing. Experimental set up and legend as in panel (A) except for the Y-axis which displays the MAE of the frequency estimation for all haplotypes resulting from simulations obtained with the same coverage and same number of polymorphic sites, irrespective of the number of haplotypes. Error bars indicate the MAE 90% confidence interval. (C) Comparison of the reconstruction reliability using short read sequencing (100 bp). Legend and experimental set up as in panel (A), but displaying results obtained on data simulating short reads. (D) Comparison of the reconstruction reliability using short read sequencing. Legend and experimental set up as in panel (B), but displaying results obtained on data simulating short reads.

HAPLOTYPE RECONSTRUCTION TO RECONSTRUCT EVOLUTIONARY TRAJECTORIES

To test EVORhA in a real setting, we reconstructed haplotypes from pooled sequencing data of population samples, taken during an evolution experiment. In this experiment a lab *E. coli* strain was subjected to high ethanol concentrations and growth in the presence of ethanol, referred to as ethanol tolerance was estimated as a focal phenotype. The trajectory of the population phenotype clearly showed that after about 40 days, the cell's ethanol tolerance steadily increased from 6 to 7 % after which a plateau was reached (Figure 3.4A). To evaluate the evolutionary trajectories of the haplotypes during this switch in the population phenotype, samples were taken at three consecutive time points: at T0, the beginning of a 40 days stationary phase when no increased tolerance against ethanol was observed yet, at T1 right before the phenotypic switch and at T2 the focal end point after which no further increase in ethanol tolerance was observed (Figure 3.4A). Population samples were subjected to Illumina pooled sequencing and applying EVORhA to the data obtained for each of the pooled samples allowed reconstructing per population its composition i.e. the different haplotypes that were present and the frequencies at which they were present in the respective populations. As haplotypes present in consecutive time points are related to each other through their common ancestry, the phylogenetic relations between the reconstructed haplotypes was inferred, the evolutionary history is represented by means of a concept map using CmapTools (Novak & Cañas 2008) (Figure 3.4B).

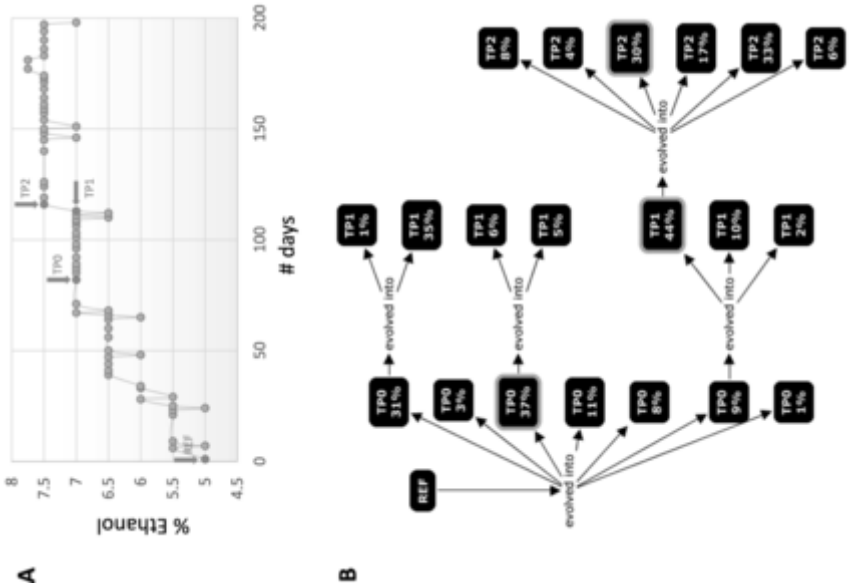


Figure 3.4 Haplotype reconstruction to infer evolutionary trajectories.

(**A**) shows the phenotypic trajectory of a population during an evolution experiment in which *E. coli* strains were subjected to increasing ethanol concentrations. The measured focal phenotype is the ethanol tolerance of the population (i.e. the % of ethanol at which growth still occurs). Arrows indicate the time points at which population samples were taken that were subjected to sequencing and haplotype reconstruction. Y-axis indicates the % of Ethanol to which the population was subjected. (**B**) concept map illustrating the evolutionary relations between the haplotypes reconstructed from each of sampled time points described in panel (A). Ref indicates the unevolved parental strain of which the genomic sequence was used as a reference. TP0, TP1 and TP2 represent the 3 time points at which population samples were taken (see panel (A)), i.e. TP0 is Time Point Zero (0). Each square corresponds to a different reconstructed haplotype and '%' indicates the frequency at which this haplotype was estimated to occur in the population. Arrows indicate the phylogenetic relatedness between the reconstructed haplotypes (or ancestry). Indicated with a lighter gray border are the reconstructed haplotypes that best match the individual clones, sampled at each time point.

To verify the correctness of the haplotype reconstruction, we sampled and sequenced one clone per time point and determined their polymorphisms. Per time point each sampled clone was assigned to its 'best matching reconstructed haplotype' based on a minimal number of inconsistencies between polymorphisms of the sampled clone and polymorphisms present in the inferred haplotypes. At TP0, the genome of the sampled clone contained 296 polymorphisms that were still rising to fixation in the population. 295 of these polymorphisms could uniquely be assigned to one of the reconstructed haplotypes (i.e. with a reliability of 99.7%). At TP1 the sampled clone contained 152 polymorphisms rising to fixation in the population, of which 151 could uniquely be assigned to a reconstructed haplotype (with 99.3% reliability). At TP2, the sequenced clone contained 122 polymorphisms still rising to fixation in the population of which 111 could uniquely be assigned (with 91% reliability). The haplotypes best matching the genomic sequences of the sampled clones are indicated in Figure 3.4B.

Although the sampled clones did most often correspond to the haplotypes that were high in frequency at the time of sampling, reconstructing the evolutionary trajectory of the focal end point clone would not have been possible using the sequence information from the sampled clones only. This because, although phenotypically the population undergoes a clear increase in ethanol tolerance (Figure 3.4A), it remains at each single time point largely heterogeneous (Figure 3.4B). The two haplotypes dominating at TP0, before the sweep with a frequency of respectively 31% and 37 %, of which one was sampled as an individual clone, seem to have been dead ends and therefore were not ancestral. In contrast, a minor haplotype rapidly increasing in frequency between TP0 and TP1 (from 9 to 44 %) took over the population and eventually gave rise to all haplotypes at TP2, the focal end point.

Note that in Figure 3.4B, frequencies inferred at a single time point do not sum to one. This because, we intentionally choose to only perform a partial haplotype reconstruction in case the genome-wide reconstruction was still ambiguous (if a haplotype set occurs at a low frequency and has no conflicts with more than one haplotypes occurring at higher frequency, we have insufficient information to decide to which genome-wide haplotype the haplotype set at low frequency belongs).

HAPLOTYPE RECONSTRUCTION TO IDENTIFY MIXED INFECTIONS

Currently, tracing the origin of an infectious disease during an outbreak is based on determining the genetic similarity between individual strains sampled from different infected entities (individuals), hereby assuming that the contaminating population isolated from each entity is largely homogenous. However, in case of mixed infections such approach might fail (Eyre et al. 2013) unless the different contaminators within one individual can be disentangled.

To assess the applicability of EVORhA in identifying mixed infections, we used the benchmark data generated by Eyre et al. (Eyre et al. 2013). These authors mimicked in vitro 36 mixed *Clostridium difficile* infections by pairwisely combining in different proportions (50%/50%, 70%/30% and 90%/10%) DNA extracted from single clones. We reconstructed genome-wide haplotypes from the Illumina based sequence data of these mixed samples. To this end, we used EVORhA either in combination with 'all' polymorphisms detected in the population or as an alternative, and consistent with the original approach described in Eyre et al., in combination with a preselected set of polymorphisms a priori known to be discriminative for the haplotypes in the mixture (Eyre et al. 2013). Also here, the reliability of the haplotype reconstruction was assessed by comparing the reconstructed genomes with the ones known to be present in the mixtures. The correctness of the inferred

haplotype frequencies was assessed by the RMSE (Mean Squared Error). Results for the reconstruction are displayed in Figure 3.5.

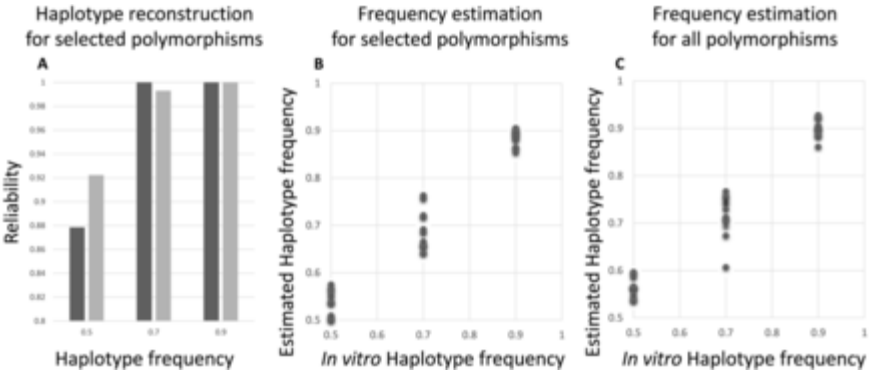


Figure 3.5 Haplotype reconstruction to identify mixed infections.
(**A**) Reliability of the haplotype reconstruction of the mixed infection set up when using a selection of polymorphisms. The X-axis contains the three different *in vitro* proportions at which the mixed infection set ups were generated (50–50%, 30–70% and 10–90%). For each of the three proportions, 12 mixtures were obtained by mixing two different clonal DNA samples according to the indicated frequencies. The Y-axis indicates the average reliability of the two reconstructed haplotypes compared to the known haplotypes at the polymorphic sites. (**B** and **C**) Correctness of the frequency estimation in the mixed infection set up. For panels (B) and (C) the X-axis represents the known frequency of this haplotype in the mixture, whereas the Y-axis represents the estimated frequencies of the most frequent haplotype in each combination (the least frequent haplotype is not displayed as it would have a frequency of 1—the shown frequency). (B) Correctness of the frequency estimation based on a haplotype reconstructing using the 151 selected polymorphisms only (RMSE = 0.037, MAE = 0.030). (C) Correctness of the frequency estimation based on a haplotype reconstructing using all polymorphisms (RMSE = 0.047, MAE = 0.038).

For both approaches, it is clear that reconstructing haplotypes from a mixed infection improves if the haplotypes in the mixture occur at different frequencies (such as observed at 90%/10% or 70%/30%). This is reflected by the high reliabilities (Figure 3.5A) and good frequency estimates (Figure 3.5 panel B and C) obtained at these frequency ratios in the mixture. This is not unexpected as in this bacterial setting (low mutation frequency) the haplotype

reconstruction largely relies on the frequency information: the more the frequencies differ between the haplotypes in the mixture, the more discriminative this feature is in assigning polymorphisms to correct haplotypes. Although in 50%/50% mixtures many ambiguous assignments are expected to occur, resulting in the lowering of the reconstruction reliability, it is remarkable that EVORhA is still able to reconstruct the haplotypes relatively well in the 50%/50% mixture (a reliability of at least 85% which is significantly higher than what would be expected from randomly combining contigs occurring at 50% into two haplotypes). At the given sequencing coverage, the mixture model underlying EVORhA seems to have a rather high resolution (allowing to separate a haplotype occurring at 53% from a haplotype occurring at 47%). Although this deviation from 50% is now penalized in the RMSE, which assumes that in the in vitro mixed sample the haplotypes truly occur at 50%/50%, experimental and sampling biases might have resulted in the inferred small deviations of this intended 50%-50%. Despite being very small, these frequency deviations can still be captured by the haplotype reconstruction, resulting in a correct reconstruction.

DISCUSSION

In this work we present EVORhA, a method for reconstructing haplotypes from deep sequencing data of clonal populations that have a relatively low mutation rate, such as bacteria.

Haplotype reconstruction in general is complicated because polymorphisms of infrequent haplotypes are difficult to distinguish from sequencing errors. The solution to this problem, referred to as local haplotype reconstruction has been proposed in the context of viral haplotype reconstruction and relies on simultaneously, rather than sequentially identifying sequencing errors and reconstructing haplotypes (Zagordi et al. 2011; Astrovskaya et al. 2011; Prosperi & Salemi 2012; Huang et al. 2011; Prabhakaran et al. 2013; Töpfer et

al. 2013). EVORhA uses a local haplotype reconstruction based on similar principles, but in addition exploits the information contained in a BLOSUM matrix to better distinguish true polymorphisms from likely sequencing errors.

In contrast to viral reconstruction, however, in our bacterial setting read lengths are short compared to the average distance between polymorphic sites. This prevents us from using viral haplotype reconstruction approaches to infer bacterial haplotypes, because in order to extend locally inferred haplotypes into more global ones, all viral methods rely on the presence of a sufficient number of segregating sites to reconstruct haplotypes from phasing information (Beerenwinkel et al. 2012).

Key to our method, therefore, is the use of the frequency ratios of the inferred haplotypes, not only to improve the extension of haplotypes for which the non-empty overlap between flanking windows results in an ambiguous phasing (i.e. the window extension) (Töpfer et al. 2014), but also to further link distant phased regions that have no sequence overlap at all (i.e. during the genome-wide reconstruction step).

As was shown on the simulated data, as soon as read lengths and/or mutation rates become restrictive for state-of-the-art methods, the additional frequency information, mainly through the genome-wide reconstruction step allows EVORhA to still reliably reconstruct haplotypes.

This frequency-based genome-wide reconstruction is also the key enabling step to resolve the bacterial haplotypes in the real data applications. This step, based on using a mixture model assumes that locally extended haplotypes observed at similar frequencies are likely to belong to the same global haplotype. This assumption gets violated however, if two haplotypes occur at similar frequencies, in which case the haplotype reconstruction might result in hybrids. However, our results on the mixed infection dataset showed that even

marginal frequency deviations between haplotypes allow the mixture model to resolve these haplotypes with high accuracy, provided the sequencing coverage of the sample is sufficiently high.

As was shown in the results, sequencing coverage highly impacts the reconstruction performance of EVORhA: at first indirectly because it affects the correctness of the reference-based assembly which is used as input. More directly because a too low coverage complicates distinguishing sequencing errors from true polymorphisms. In addition, the coverage determines the maximum number of haplotypes that can be detected. This is mainly because the standard deviation used when inferring the mixture model is dependent on the coverage, i.e. a lower coverage implies larger standard deviations of the Gaussian distributions of the mixture model which may cause haplotypes occurring at similar frequencies to become confounded.

Conclusively, EVORhA, by enabling bacterial haplotype reconstruction opens a whole new area of applications for bacterial population sequencing (or metagenome sequencing). As was illustrated by the real data examples, bacterial haplotype reconstruction can aid in resolving mixed infections or in reconstructing the dynamics of evolving clonal populations. In addition, it can potentially be useful to further resolve genomes from reads with similar sequence composition in bacterial metagenomics datasets of which the complexity has been reduced with binning approaches (Alneberg et al. 2014; Albertsen et al. 2013; Nielsen et al. 2014).

Haplotype reconstruction thus provides a quick view on the composition of a mixed sample and allows pinpointing haplotypes with interesting characteristics that can be further focused on by downstream molecular analysis.

Chaper 4. DETECTION OF CANCER DRIVING GENES USING MUTUAL EXCLUSIVITY

ABSTRACT

Because of its clonal evolution a tumor rarely contains multiple genomic alterations in the same pathway, as disrupting the pathway by one gene often is sufficient to confer the complete fitness advantage. As a result mutated genes display patterns of mutual exclusivity across tumors. The identification of such patterns have been exploited to detect cancer drivers. The complex problem of searching for mutual exclusivity across individuals has previously been solved by filtering the input data upfront, analyzing only genes mutated in numerous samples. These stringent filtering criteria come at the expense of missing rarely mutated driver genes. To overcome this problem, we present SSA-ME, a network-based method to detect mutually exclusive genes across tumors that does not depend on stringent filtering. Analyzing the TCGA breast cancer dataset illustrates the added value of SSA-ME: despite not using mutational frequency based-prefiltering, well-known recurrently mutated drivers could still be highly prioritized. In addition, we prioritized several genes that displayed mutual exclusivity and pathway connectivity with well-known drivers, but that were rarely mutated. We expect the proposed framework to be applicable to other complex biological problems because of its capability to process large datasets in polynomial time and its intuitive implementation.

INTRODUCTION

Because of internationally coordinated efforts such as TCGA (Cancer Genome Atlas Research Network et al. 2013; Cancer Genome Atlas Network 2012) and ICGC (International Cancer Genome Consortium et al. 2010), a vast number of cancer datasets are publicly available. Using these datasets to identify mutations and pathways driving cancer phenotypes has become an active field of research(Gonzalez-Perez & Lopez-Bigas 2012; Ciriello et al. 2012; Vandin et

al. 2012; Ng et al. 2012). Tumorigenesis and tumor progression follow a clonal evolutionary model (Yeang et al. 2008; Alexandrov et al. 2013; Vogelstein et al. 2013; Hahn & Weinberg 2002). In this view, the disruption of a single gene in a molecular pathway often yields the complete fitness advantage associated with disruption of that pathway, making additional mutations in the same pathway redundant (Yeang et al. 2008). This evolutionary property can be exploited to understand cancer mechanisms by searching for patterns of genes that display mutual exclusivity (*i.e.* groups of genes which mostly have maximum one mutation per tumor). The identification of groups of genes showing patterns of mutual exclusivity across patients in large datasets has already been proven useful for the detection of driver mutations/pathways in single cancer types such as triple-negative breast cancer(Shah et al. 2012), Lung Adenocarcinoma (Leiserson et al. 2013) and in a pan-cancer setting(M. D. M. Leiserson, Vandin, et al. 2015; Kandoth et al. 2013).

In practice the mutual exclusivity patterns are not always strict (hard patterns), *i.e.* most patterns occasionally show the presence of multiple mutations in a single tumor. This is possible because for example tumorigenesis can start in an initially less potent driver, but more potent drivers in the same pathway can accumulate at later times, providing an additional marginal beneficial effect (diminishing returns)(Barrick & Lenski 2013). Therefore, exploiting clonal behavior for identifying driver pathways requires searching for “soft” mutual exclusivity where two otherwise independent mutational events co-occur less than expected by chance(Bradley & Farnsworth 2009).

In order to discover genes that exhibit a mutual exclusive pattern in cancer, all possible sets of genes have to be examined. Due to the factorial computational complexity of this problem i.e. adding an extra gene to the pattern implies that the algorithm’s processing time increases factorially (Bassil 2012), this problem cannot be solved for large data sets. Current methods mainly cope with this by prioritizing potential important genes upfront, filtering out genes which seem

to be less important mainly based on the frequency with which they are mutated across tumor samples. Methods such as Dendrix (Vandin et al. 2012), CoMEt (M. D. Leiserson et al. 2015) and Multidendrix (Leiserson et al. 2013) explicitly try to find the largest set of genes that exhibit a mutual exclusivity pattern after a filtering step, using an integer linear program or a Markov chain Monte Carlo approach while methods such as MEMo (Ciriello et al. 2012) and Mutex (Babur et al. 2014) rely on the use of the human interaction network to further constrain the search space by using the knowledge that mutually exclusive genes are likely to be located in the same molecular pathways.

MEMo relies on a human protein-protein interaction network to search for the largest set of genes that are closely related in the network and that exhibit mutual exclusivity, whereas Mutex uses a directed signaling network. Although using a network restricts the search space, searching for patterns of mutual exclusivity is still a difficult task. For these reasons, both MEMo and Mutex require a stringent filtering of the input (the input of Mutex is required to consist of less than 500 genes, MEMo is capable to analyze about 250 genes). As a result, potential drivers that are rarely mutated are likely to be missed.

Therefore we developed SSA-ME (Small Subnetwork Analysis with reinforced learning for detecting Mutual Exclusivity patterns), a computational tool that searches for genes that belong to common patterns of mutual exclusivity and that are closely connected on an interaction network to prioritize drivers. It uses a novel methodology named Small Subnetwork Analysis with reinforced learning (SSA) that divides a complex problem, *i.e.* finding the largest set of genes that exhibit a mutual exclusivity pattern, into many simpler ones by calculating measures for mutual exclusivity in many small subnetworks. By solving these simpler problems iteratively, each time biasing the search space based on results of previous iterations, SSA-ME can prioritize potential driver genes with linear algorithmic complexity. This, in principle, allows it to process

large input datasets in short computational times and therefore, in contrast to previous approaches, requires little prior filtering.

To assess the performance of SSA-ME we re-analyzed the breast cancer dataset from the 2012 cancer genome atlas (TCGA)(Cancer Genome Atlas Network 2012) without filtering the genomic variants up front. Despite adding many more mutations in the input, we could prioritize well-known drivers that are found to be recurrently mutated in different tumors. However, in addition to prior findings we could prioritize several genes that displayed mutual exclusivity and pathway connectivity with well-known drivers, but that were rarely mutated in the different tumors and therefore were missed by other methods that search for mutual exclusivity.

MATERIALS AND METHODS

SSA-ME

Small Subnetwork Analysis with reinforced learning for detecting Mutual Exclusivity patterns (SSA-ME) is an algorithm that uses a reference network to detect mutual exclusive gene patterns in cancer. To accomplish this, SSA-ME performs two independent functions in an iterative manner: small subnetwork selection/scoring and reinforced learning. Each gene (node) in the reference network is initialized with an initial uniform gene score. Then, iteratively: starting from a set of seed genes, small subnetworks are selected favoring genes with high gene scores. Each selected small subnetwork is then scored based on how well the genes composing the small subnetwork belong to a mutual exclusivity pattern. Genes that consistently belong to small subnetworks with high scores thus exhibit mutual exclusivity with other genes in their neighborhood very well and are more likely to be selected in subsequent iterations. This will lead to high gene scores for genes which are part of a mutual exclusivity pattern. The pseudocode describing the algorithm can be found in Figure 4.1.

```
network := initialize
for n in seeds:
createSubnetworkSelector(n)
for 1 to number_of_iterations or converged:
# Subnetwork selection and scoring
      for<parallel> ss in subnetworkSelectors:
      subnetwork := ss.selectSubnetwork
      store&ScoreSubnetwork(subnetwork)
# Reinforced learning
for n in nodes:
reinforceLearning(n)
```

Figure 4.1 Pseudocode of SSA.ME algorithm

INITIALIZATION

The algorithm is initialized by giving each gene (node) a uniform initial gene score of 0.5. A static list of seed genes is defined that contains genes that possibly belong to a mutually exclusive pattern. Any type of biologically relevant filtering can be used to generate such gene list. In the context of this paper, seed genes are defined as all genes that were found to be mutated in at least one sample (tumor).

SMALL SUBNETWORK SELECTION AND SCORING

In each iteration small subnetworks of equal size are selected. Starting from every seed gene, subnetworks are selected by subsequently adding a gene adjacent to the current subnetwork. In order to be able to detect mutual exclusivity patterns of different sizes, the size of the small subnetworks varies from 3 to 6 genes between iterations. The probability of adding an adjacent gene to a small subnetwork is proportional to the gene scores of adjacent genes, expressing the assumption that mutually exclusive genes are likely to be located in the same adaptive pathway. Once constructed, each small

subnetwork receives a mutual exclusivity score (MES). Each sample contributes to this score with a weight that is inversely related to the number of genes from the small subnetwork that were found mutated in that sample. This is calculated using the following equation:

$$MES(sn) = \sum_V \sqrt{\sum_{s \in S} \frac{1}{m(s,V)}} \quad \text{Formula 1}$$

Where V are the genes present in small subnetwork sn ordered according to the number of samples in which these genes were found to be mutated. S is the set of samples pending to contribute to the mutual exclusivity score. Initially S includes every sample with a mutation in one of the genes in the small subnetwork, but every time a sample is used to calculate a mutual exclusivity score it is removed from S. In this way a sample can only contribute once to the MES. $m(s,V)$ is the number of genes in V which are mutated in sample s. This value would be equal to 1 if the genes in gene set V are all members of a perfect mutual exclusive pattern and $|V|$ if all genes in V are mutated in all samples. The square root allows giving relatively higher mutual exclusivity scores to small subnetworks for which each gene is mutated in approximately the same number of samples.

Next, the MES are ranked from highest to lowest and their ranks are divided by the maximum rank (Figure 7). We end up with a ranked MES ($rMES$) between zero and one where zero refers to the small subnetwork having the least evidence for mutual exclusivity and one refers to the small subnetwork having the most evidence for mutual exclusivity.

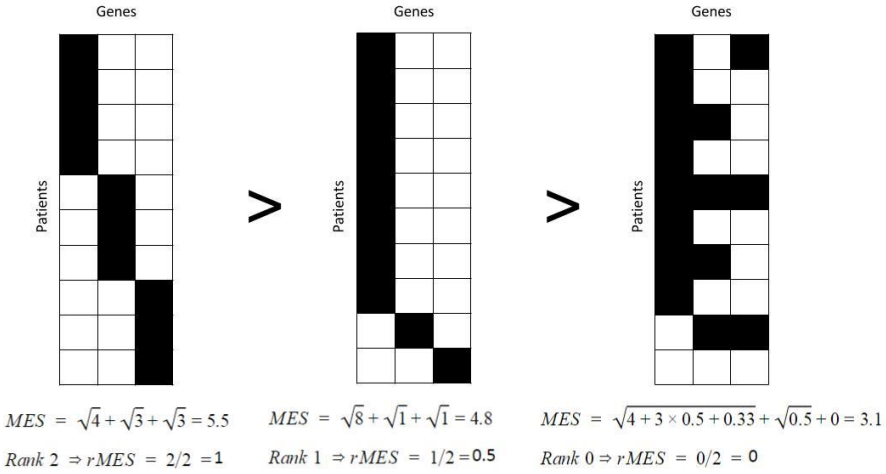


Figure 4.2 Calculation of *MES* and corresponding *rMES* scores for three different small subnetworks.
Genes which make up the small subnetwork are represented as columns of matrices, patients are represented as rows. Genes with alterations in a specific patient are depicted as black tiles. Small subnetworks exhibiting perfect mutually exclusivity patterns (two most left small subnetworks) have higher *rMES* scores than small subnetworks with non-perfect mutual exclusivity patterns (most right small subnetwork). Also, small subnetworks having a more uniform distribution of gene alterations across patients have higher *rMES* scores as shown by the two most left small subnetworks.

REINFORCED LEARNING

Using the $rMES$ for each small subnetwork, the reinforced learning step updates gene scores based on two parameters: *reinforcement* and *forgetfulness*. The *reinforcement* is a parameter that determines the maximal value by which a gene score can be increased in the next iteration. The reinforcement is multiplied by the highest $rMES$ score of all small subnetworks to which the gene belongs, so the gene score of genes which are consistently in small subnetworks with high $rMES$ scores will further increase with iterations.

The *forgetfulness* determines the fraction of the gene score that is retained in every subsequent iteration. This means that part of the gene score is effectively lost every iteration step and thus the gene scores of genes having persistently low $rMES$ scores will go to zero. To calculate gene scores the following formula is used:

$$g_{i+1} = g_i . f . \left[1 + r . \max_{sn \in SN_g} rMES(sn)\right] \text{ Formula 2}$$

Where g_i is the gene score at iteration i, f is the *forgetfulness*, r the *reinforcement*, SN_g the set of small subnetworks containing the gene. If the gene score resulting from the formula is larger than 1, it is topped off at 1 as the maximal gene score can never be larger than 1. The default parameters of the method are forgetfulness $f = 0.995$, reinforcement $r = 0.005$ and 5000 iterations. In general, the sum of forgetfulness and reinforcement should be close to 1 and the reinforcement should be small (smaller than 0.01). This because small values for forgetfulness or large values for reinforcement would make the algorithm prone to stochastic effects. Note that genes which are not part of any small subnetwork are assigned a value of zero for $\max_{sn \in SN_g} rMES(sn)$.

In a final step we assign a rank to each gene that reflects to what extent a gene belongs to a mutual exclusivity pattern. Hereto we exploit the fact that genes belonging to a mutual exclusivity pattern tend to have a consistent increase in their gene score between iterations over time. Genes are ranked according to the maximal gene score they reach and in case of ties are based on how fast their score converges.

Simulated data

To assess the performance of SSA-ME we used simulated data. The set of true positive driver genes was defined first by creating a target mutually exclusive pattern which in biological terms corresponds to a driver pathway. The target

mutual exclusivity pattern was generated using a random walker with restart (5% restart chance) to select genes from the local network neighborhood of a randomly selected gene until 20 interactions have been visited in a high quality human reference network. This high quality human reference network was composed of HINT(Das & Yu 2012), Interactome (HI-II-14) (Rolland et al. 2014) and Reactome (Croft et al. 2014) interaction data.

To mimic real tumor data, we counted the number of mutated genes present in each tumor sample in the TCGA 2012 study and assigned an equal number of alterations to random genes, thus conserving the distribution of mutated genes per sample. We added mutually exclusive mutations to genes present in the target pattern in 30 % of the samples. Each sample had 5% chance to also be mutated in any of the other genes belonging to the same mutual exclusivity pattern as we allowed for "soft" mutual exclusivity patterns which are non-perfect across samples.

To evaluate the robustness of the method with respect to the used reference network, multiple simulated datasets were analyzed for different degrees of connectedness in the high quality human reference network: highly underconnected (50% of the edges were deleted from the reference network), mildly underconnected (25% of the edges deleted), lowly underconnected (10% edges deleted), true network (i.e. the high quality human reference network), lowly overconnected (10% additional random edges added to the reference network), mildly overconnected (25% additional edges) and highly overconnected (50% additional edges). We generated 100 different simulated datasets per network and ran SSA-ME. Performance was measured by receiver operating characteristic (ROC) curves.

To assess parameter sensitivity we tested the effect of using different parameter combinations on the performance. This included 400 simulations for all combinations of reinforcement r (from 0.0005 to 0.0100 in steps of

0.0005) and forgetfulness f (from 0.99 to 0.9995 in steps of 0.0005). Performance for each parameter combination was measured using the area under the curve (AUC).

Breast Cancer TCGA Data

The TCGA Breast Cancer (BRCA) data published in 2012(Cancer Genome Atlas Network 2012) was downloaded from https://tcga-data.nci.nih.gov/docs/publications/brca_2012/. Level 2 files were used containing somatic mutations, RNA expression and copy number variations. Copy number alterations obtained from the original TCGA Breast cancer data were inferred with GISTIC(Beroukhim et al. 2007). In our analysis only genes in samples with high-level thresholds for amplifications/deletions and for which copy number alteration showed positive correlation with expression level were used. Priorization results were obtained by running SSA-ME on a non-stringently filtered input set, consisting of all genes having at least one genetic alteration (mutation or amplification/deletion correlated with expression) in the dataset. As a high quality human reference network we compiled information data from HINT (Das & Yu 2012), Interactome (HI-II-14) (Rolland et al. 2014) and Reactome (Croft et al. 2014).

Using the TCGA breast cancer data also allowed us to compare our results with the ones originally published by MEMo, a representative state-of-the-art method that searches for mutual exclusivity patterns using a reference network. To maximize the comparison, we ran our approach with the same reference network and with the same data as originally used by MEMo. This reference network is a non-curated reference network consisting of Reactome (Croft et al. 2014), Panther (Mi et al. 2010), KEGG (Kanehisa et al. 2008), INOH (Yamamoto et al. 2011) and interactions from non-curated sources (like high-throughput derived protein–protein interactions, gene co-expression, protein domain interaction, GO annotations, and text-mined protein interactions) (Wu et al. 2010). Data were reproduced according to the description in the original

paper, i.e. only retaining genes that were altered in at least ten samples. To illustrate how prioritizations were not largely affected by omitting this stringent prefiltering we redid the analysis in the same setting but using the less filtered input described above.

RESULTS

SSA-ME IMPLEMENTATION

To identify cancer drivers we develop SSA-ME (Small Subnetwork Analysis for mutual exclusivity), a method that searches for small subnetworks of the interaction network containing mutated genes that show a pattern of mutual exclusivity. SSA-ME reformulates the complex problem of finding the largest set of mutually exclusive genes into many independent and less complex sub-problems. SSA-ME scores many small subnetworks for their potential to contain genes that belong to a mutual exclusivity pattern, instead of explicitly searching for the largest set of mutual exclusive genes. Using these small subnetwork scores in a reinforced learning framework allows prioritizing individual genes that are likely to belong to a mutual exclusivity pattern, without ever having to explicitly evaluate the largest set of mutually exclusive genes.

The method is outlined in Figure 4.3. SSA-ME searches the local neighborhood around a set of predefined seed genes. In this case, the seed genes correspond to all genes mutated in at least one sample. In each iteration step of the algorithm, genes in the neighborhood of a seed gene are selected into a small subnetwork with a chance proportional to their gene scores (which are chosen to be uniformly distributed in the first iteration). These small subnetworks are subsequently scored based on the mutual exclusivity signal of the genes in each small subnetwork. Individual gene scores are updated proportional to the mutual exclusivity scores of the selected small subnetworks to which they belonged. Updating of the gene scores modifies the likelihood with which each gene will be selected in subsequent iteration steps. The iterative process

continues until the method converges to a solution or a maximum number of iterations is reached. The output of SSA-ME consists of an interactive network together with supporting files compatible with Cytoscape (Shannon et al. 2003).

Performance on simulated data

To evaluate the robustness of the method with respect to the used reference network, we applied SSA-ME on a simulated dataset in combination with a high quality human reference network (see Materials and Methods) and underconnected/overconnected versions of this reference network (with respectively 10%, 25% and 50% of the network edges being deleted or added). Per network, 100 simulations were performed. Each simulated dataset contained a target mutual exclusivity pattern consisting of maximally 20 genes interacting on the reference network that were mutated in 30% of the samples (see Materials and Methods).

Applying SSA-ME on each simulated dataset resulted in a ranked gene list. The top x% of the gene list were considered as genes belonging to a mutual exclusivity pattern, whereas the remainder of the genes were considered not to exhibit mutual exclusivity. Performance was evaluated by plotting the sensitivity versus the specificity where the sensitivity is defined as the percentage of genes belonging to the target mutual exclusivity pattern that were retrieved amongst the x% highest ranked genes and the specificity, defined as the proportion of genes not present in the target pattern that were correctly classified as such. The results are shown in Figure 4.4A for the highest ranked genes as this is the range that is of biological relevance (correctly identifying positives). The full ROC plot and the sensitivity/PPV plots can be found in the Supp Figure C. 1.

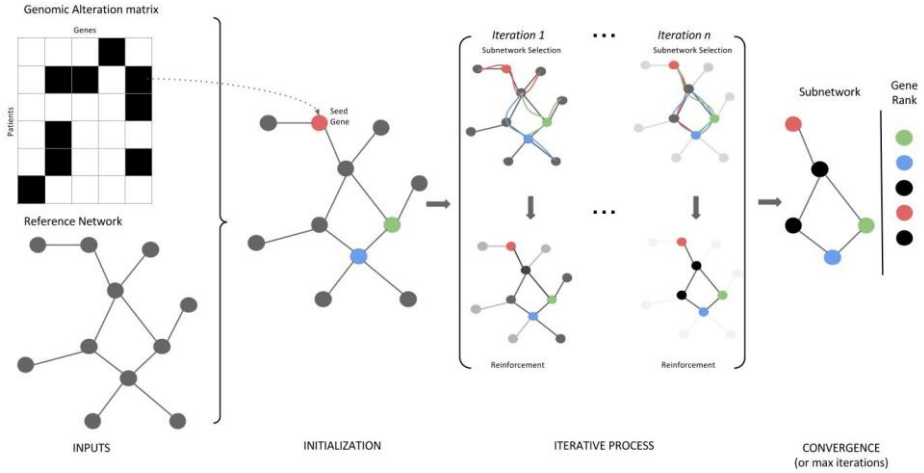


Figure 4.3 Overview of SSA-ME.
The input consists of a matrix containing genomic alterations (i.e. mutations or copy number alterations, among others) across patients (depicted as black tiles) and a human reference network. In a first initialization step, every gene which has at least one genomic alteration across all patients is selected as a seed gene (colored genes in the network). The gene scores (represented as the opacity of the genes in the networks) are uniformly set to a value of 0.5. In every subsequent iteration step, small subnetworks will be generated, starting at every seed gene. Every gene adjacent to the small subnetwork has a chance proportional to its score to be incorporated in the small subnetwork. When a certain size has been reached the small subnetwork generation will stop and a score for each selected small subnetwork will be calculated based on the mutually exclusivity pattern found within this small subnetwork. At the end of every iteration step these small subnetwork scores will be used to update gene scores, altering the chance of genes to be incorporated into the small subnetwork in subsequent iteration steps. Upon convergence it can be seen that a few genes have high scores while others have scores close to 0. Genes are ranked based on their gene scores which reflects their potential to belong to a mutual exclusivity pattern.

Figure 4.4A indicates that the best performance is obtained using the reference network without added or deleted edges, as for the same relative increase in sensitivity less false positives are predicted (lower relative increase in 1-sensitivity). The method shows in general a high resilience of the results to using an overconnected network. In this case the method is capable of successfully prioritizing most of the mutually exclusive genes with a low

number of false positives (which is the range we envisage when only showing the values of the 1-specificty between 0 and 0.01). With an underconnected network the maximal sensitivity that can be reached will get restricted as some of the genes that show mutual exclusivity can no longer be connected in the network.

To assess the sensitivity of the method versus its parameter settings we ran SSA-ME on the same simulated data each time using a different combination of the reinforcement and forgetfulness parameters. Hereby reinforcement values were varied from 0.0005 to 0.0100 in steps of 0.0005. Forgetfulness values varied from 0.99 to 0.9995 in steps of 0.0005. Note that values of the forgetfulness closer to 1 imply that less is 'forgotten' and values of reinforcement are consistently lower than the ones of the forgetfulness to ensure that only few true positives will be reinforced. For each parameter combination 10 simulated datasets were analyzed. The performance per parameter combination was assessed using the area under the ROC (Figure 4.4B). In general a low performance is obtained if the forgetfulness is relatively low compared to the reinforcement. In those settings false positives might become reinforced relatively more than some weak or isolated true positives. However, in ranges where the forgetfulness is close to 1, the performance is more robust to the choice of the reinforcement value. Best performances were obtained on the diagonal where irrespective of their absolute values the sum of the values of r and f are close to each other $r + f \approx 1$. In most cases, a combination where the sum of the reinforcement and the forgetfulness is higher than one results in lower performances because then again the reinforcement becomes relatively high compared to the forgetfulness, resulting in relatively more false positives.

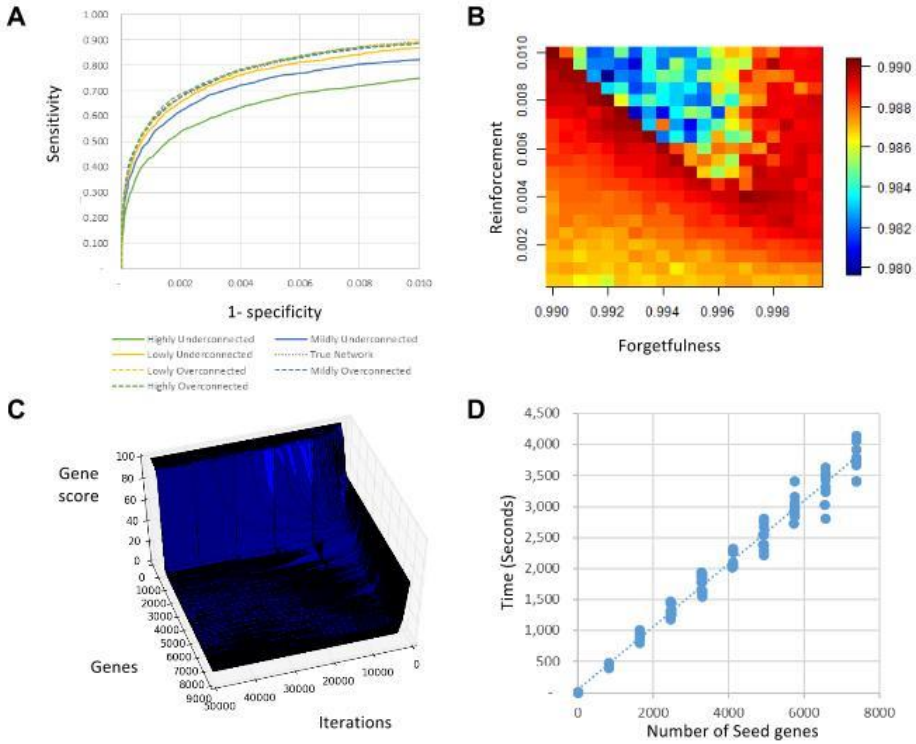


Figure 4.4 Performance on Simulated Data.
Robustness of the predictions with respect to the used reference network. The X-axis represents 1-specificity and the Y-axis represents sensitivity. Underconnected networks lead to lower performance while overconnected networks result in similar, although lower, performance to when using the the true network. Note that, for clarity reasons, the range of the x-axis is restricted to [0, 0.01]. **B)** Heat map depicting parameter sensitivity. AUC values for every analyzed parameter pair are depicted. It can be seen that the best performance is achieved on the diagonal for combinations of reinforcement and forgetfulness of 1. **C)** Plot visualizing convergence and stability of convergence of gene scores. The X-axis represents the number of performed iterations, the Y-axis displays all genes in the reference network (black lines in the plot) and the Z-axis represents the gene scores. All genes start on the right side with a gene score of 0.5. Most of them converge fastly to 0 or 1. **D)** Plot showing linear time complexity of the algorithm with respect to the number of seed genes. Each dot on the plot represents the time to convergence of a separate run. Per tested number of seed genes, 10 simulations were performed. Results were obtained by running the algorithm on one single processor Intel(R) Xeon(R) CPU E5-2670 0 @ 2.60GHz.

To show that the method converges to a stable solution, we ran it on one simulated dataset for 50.000 iterations. Figure 4.4C shows that the method exhibits a consistent behavior, i.e. after a gene obtains a high gene score, it will remain consistently high or vice versa. Furthermore this figure shows that the algorithm converges, provided a sufficient number of iterations have been performed.

To analyze its complexity with respect to the number of seed genes, we ran SSA-ME on 10 different simulated dataset, each time using an increasing number of seed genes (ranging from 1 to 8000 genes). Datasets contained incrementally more added seed genes. Seed genes were added gradually according to the frequency with which they were found mutated in the different tumor samples, hereby assuming that the most frequently mutated genes are the ones that in a real setting would also be prioritized as the most promising seeds. These runs were repeated on 10 different simulated datasets. Results are visualized in Figure 4.4D and clearly show the linear complexity of the algorithm with respect to the number of seed genes.

Analysis of the TCGA breast cancer data

To test the biological relevance of the predictions we applied SSA-ME on the well-studied TCGA Breast Cancer 2012 dataset (Cancer Genome Atlas Network 2012) using a high quality human reference network (see Materials and Methods). As seed genes we used all genes carrying somatic mutations or copy number alterations, provided the latter alterations also resulted in positively correlated expression values of those copy number altered genes. After running SSA-ME, genes were ranked according to their gene score and the highest ranked genes were prioritized as putative drivers. The cut-off on the ranked list was chosen so that, given a set of known cancer genes, a good trade-off between sensitivity and precision was obtained, i.e. we use the cut-off so that a maximal sensitivity was obtained with a PPV higher than 80%

(Figure 4.5A). Note that the PPV represents a lower boundary on the actual number of true positive predictions as all genes not previously associated with cancer according to CGC (Futreal et al. 2004), Malacard (Rappaport et al. 2013) or NCG (An et al. 2015) are regarded as false positives. Applying SSA-ME on the TCGA BRCA 2012 dataset resulted in 49 genes being prioritized as potential breast cancer associated genes. Because of the nature of the method this prioritized gene list contains both putative drivers, but also 'linker genes' that connect genes that are part of a mutual exclusivity pattern but that are not mutated themselves. These 'linker genes' are therefore no true drivers, but have driver potential as they were found in the network neighborhood of true drivers.

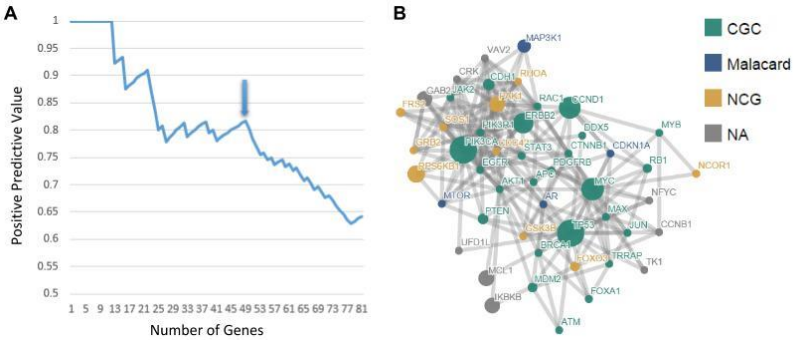


Figure 4.5 Application of SSA.ME on TGCA Breast Cancer dataset.
Determination of the number of genes to be prioritized as cancer drivers. Genes were ranked according to their gene score obtained with SSA.ME. The X-axis represents the number of genes in the list of prioritized genes obtained by setting a cut-off on the rank. The Y-axis represents the positive predictive value (PPV) for the genes present in each list that corresponds to a given rank threshold. At the chosen threshold (arrow) 49 potential cancer drivers were prioritized. **B)** Subnetwork obtained after using SSA.ME on the TGCA breast cancer dataset. Seed genes and network were as defined in the main text. Genes are represented by nodes. If the gene had been associated with cancer, this is indicated by the color of the database in which the association was described. Gray genes correspond to genes not present in the Census of Cancer Genes, Malacard or the Network of Cancer Genes database. The size of the node reflects the number of samples in which a gene was found mutated.

The subnetwork in Figure 4.5 displays the 49 prioritized genes and all edges in the reference network connecting them. Most of these genes (40 out of 49) have previously been associated with either breast cancer or cancer in general (Appendix C Table 1). 5 genes of those 40 were selected as 'linker genes' (i.e. they did not display alterations in the breast cancer dataset), but have been associated with other cancer types (i.e., *CDC42*, *CDKN1A*, *RAC1*, *GSK3B* and *CTNNB*). This indicates linker genes are potential cancer drivers.

Of all 49 predictions, 9 genes were not previously associated with breast cancer in Malacards (Rappaport et al. 2013), or associated with cancer in general by the Cancer Gene Census (Futreal et al. 2004) or Network of Cancer Genes(An et al. 2015) (*CCNB1*, *CRK*, *GAB2*, *IKBKB*, *MCL1*, *NFYC*, *TK1*, *VAV2* and *UFD1L*). Of those genes, two were selected as 'linker genes' (CRK and TK1). Of the remaining 7 genes, 4 were missed by previous analyses on the same TCGA breast cancer dataset because they were mutated in less than 2% of the samples and therefore did not pass the filtering strategies commonly applied prior to the driver analysis (i.e. *VAV2, UFD1L, NFYC*, and CCNB1). The genes IKBKB, MCL1 and GAB2 pass the filtering criteria used by previous methods. Of those IKBKB was also detected in the original TCGA analysis, reported to belong to a pattern of mutual exclusive genes based on the MEMo analysis, whereas MCL1 and GAB2 were not.

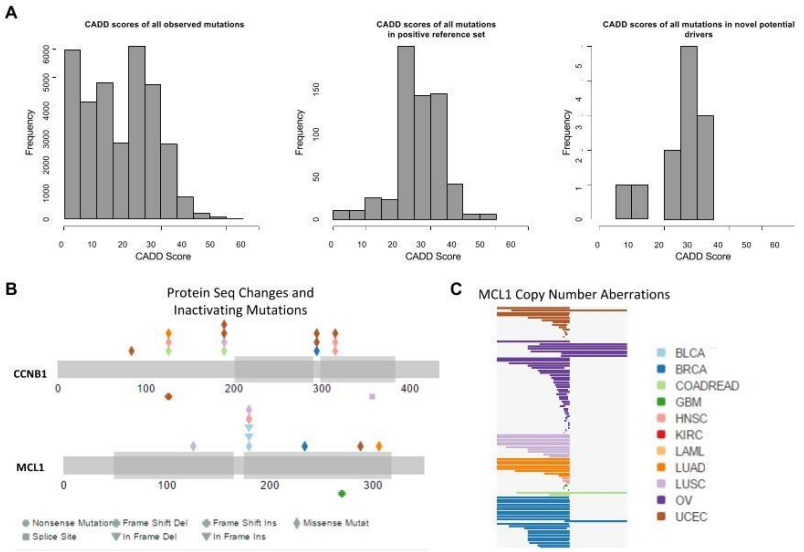


Figure 4.6 CADD scores analysis of selected genes.
A) CADD score distribution of mutations of the unselected set (left histogram), the positive set (middle histogram) and the set containing the mutations in the novel predicted driver genes (right histogram). The X-axis depicts the CADD score and the Y-axis depicts the frequency of mutations having a CADD score within a certain range. **B)** Positioning of the mutations found in the TCGA Pan-Cancer dataset along the CCNB1 and MCL1 gene loci. Only a subset of copy number aberrations are included in this representation for MCL1. Figure obtained with MAGi. **C)** Copy number aberrations observed in the MCL1 gene in the TCGA Pan-Cancer dataset. Figure obtained with MAGi (M. D. M. Leiserson, Gramazio, et al. 2015).

We could show that the mutations carried by the 49 prioritized genes and by the set of 9 cancer related genes not previously reported in cancer reference databases followed a CADD (Kircher et al. 2014) score distribution significantly different from the CADD score distribution of mutations in non-cancer related genes (Figure 4.6A), pointing towards the functional relevance of at least some of the mutations carried by the predicted drivers. In addition, we could find clear associations for the novel predictions with cancer in literature (see Appendix C). Although at least visually for some of these driver candidates, the mutations they carry seem to cluster at the same genomic positions (Figure 4.6B and Figure 2S), none of them scores highly significant for clustering of their mutations according to the results provided in the pan-cancer analysis(Cancer Genome Atlas Research Network et al. 2013) or the results we obtained by running SomInaClust (Van den Eynden et al. 2015) (see Appendix C). This indicates that indeed without using network-based information, it would be difficult to prioritize these rarely mutated genes.

Comparison with TCGA Analysis

We compared the previously obtained predictions(Cancer Genome Atlas Network 2012) of MEMo, with our predictions. To maximize comparability between our results and those of MEMo on the same TCGA dataset, we reproduced the same filtering approach and network of the original breast cancer study and ran SSA-ME (see Materials and Methods).

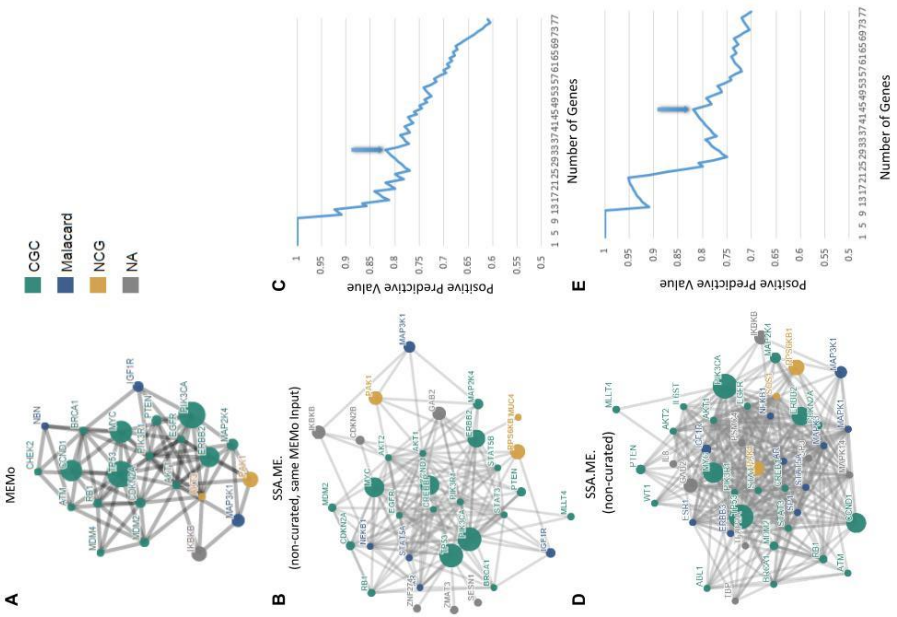


Figure 4.7 Comparison between SSA.ME and MEMo.
Prioritized driver networks obtained by MEMo as retrieved from the original mutually exclusive modules outlined in the breast cancer TCGA paper (Panel **A**), obtained by SSA.ME using the filtered data (Panel **B**), and using the non-filtered data as input (Panel **D**). Genes are represented as nodes. Colors refer to the databases in which associations of the indicated genes with breast cancer or cancer have been described. Gray genes were not found to be associated with breast cancer/cancer according to the used reference databases. Panel **C** and **E** represent the PPV analysis of results obtained by applying SSA.ME on respectively the filtered and non-filtered datasets. Y-axis represents the PPV according to the reference databases. X-axis represents the number of genes in lists of prioritized genes of increasing order. Sizes of gene lists were determined by ranking genes according to their gene scores and counting the number of genes with a rank lower than a given threshold. Arrows indicate the thresholds that were chosen to select the genes in the represented networks. We choose the threshold on the ranked list so that a good trade-off between sensitivity and precision was obtained given a set of known cancer genes, i.e. we use the cut-off so that a maximal sensitivity was obtained with a PPV higher than 80%.

Because of the high similarity of the mutual exclusivity patterns detected by MEMo in the original paper (patterns consisting of maximally 8 genes that varied in most cases in no more than one gene), we collapsed the 23 genes of

all patterns and depicted them as a network (Figure 4.7A). The subnetwork obtained by SSA-ME using the same filtered dataset consists of 33 genes (applying the same cut-off criteria as mentioned above) of which 18 were also found in the MEMo network (Figure 4.7B). 5 genes retrieved by MEMo were not detected by SSA-ME (*NBN*, *CHECK2* and *MDM4)* as they were no longer present in the filtered list we used as input, whereas they must have been present in the original input of MEMo: in contrast to what has been described in the original TGCA paper we found these genes to be mutated in less than 2 samples and therefore removed them from our analysis. The score of *ATM* just fell below the chosen threshold (it ranked 36 with SSA-ME and the cut-off we used was set at 33) and was therefore also missing from our prioritization. *ATK3* was truly missed in our analysis as the small subnetworks to which it belonged never received consistently high scores during subsequent iteration steps.

On the other hand we found 10 additional genes that were not retrieved by MEMo (of these, 4 were also found using the high quality network in the analysis described above: *AR*, *STAT3*, *RPS6KB1* and *GAB2*). Some of the additional genes had previously been associated to breast cancer (*AR* and *ESR1)* or to cancer in general (*MUC4* and *CCDN1)*. The reason why we detect more genes than MEMo is partially due to the choice on the cut-off, but also because of the inherent differences in selection criteria between the methods: first, our method only requires that the selected genes are members of the local neighborhood of genes that exhibit a mutual exclusivity pattern across patients. Second, our method does not require stringent filtering which leaves the possibility of selecting rarely mutated genes.

These results thus show that SSA-ME is able to reproduce largely the same results as MEMo, provided the same input data are used or said otherwise genes that are highly ranked by MEMo are also highly ranked by our method. The discrepancies between the number of driver genes detected in this

comparative analysis and the one above are due to the differences in the used networks. Above we choose to use a higher quality human network to reduce the possibility of including false positive interactions, whereas here we used the more connected interaction network used in the original TCGA dataset.

As shown above, the advantage of SSA-ME is that because of its reduced computational complexity it does not require stringent prior filtering of the data and therefore can also predict cancer drivers that are, for instance, infrequently mutated across the different samples. One could argue that not filtering the data can deteriorate the results as having more potentially false positives in the input list could dilute the true signals in the data and prevent the method from finding these true positives. To prove this is not happening we also applied SSA-ME on the less filtered data using the same reference network as in the original TCGA paper (Cancer Genome Atlas Network 2012). Less filtered data here correspond to all genes having at least one genomic alteration (5641 genes). Applying SSA-ME to these data and applying the same criteria to set the cut-off as mentioned above resulted in a driver network of 44 genes being selected from a total 5641 of genes (Figure 4.7D). For comparison, with the filtered input set, 33 genes were prioritized out of the 237 using the same heuristics to set the cut-off on the size of the prioritized gene lists. Assuming that filtering already prioritizes the most frequently mutated genes and thus the most promising candidates, we can argue that the list obtained with the filtered input is the most reliable. Remarkably, of the 33 genes prioritized with the filtered input, 27 also occurred amongst the 44 genes prioritized with the unfiltered set. This indicates that despite the much larger number of input genes, the presumably true signals in the data are still best recovered and, compared to the much larger input that was used only relatively few additional candidates are prioritized. Not relying on pre-filtering in contrast offers the additional advantage of also recovering candidates that would not have passed the standard used stringent filtering criteria.

DISCUSSION

We introduce SSA, a small subnetwork analysis technique with reinforced learning which solves a complex combinatorial search over an interaction network by calculating measures for mutual exclusivity in many small subnetworks of the interaction network. The method can be generically applied to any problem in which local neighborhoods in a network hold useful information.

Here we applied SSA to prioritize cancer driver genes that are located in each other's neighborhood on the interaction network and of which the genetic alterations display patterns of mutual exclusivity across different tumor samples (referred to as SSA-ME). To overcome the inherent high algorithmic complexity posed by its combinatorial nature, the problem of identifying drivers is iteratively solved and in each iteration multiple small subnetworks are independently analyzed. All results of these small subnetwork analyses are used in subsequent steps to bias the search space. The advantage of splitting the complex problem into multiple less complex problems, is that SSA-ME is not restricted by the number of mutated genes in the input data. As such by circumventing the stringent filtering strategy that is required by most other methods to enable the search for mutual exclusivity, SSA-ME can identify drivers carrying rare mutations and is able to identify genes based on relatively small-sized tumor cohorts of which the genetic variants cannot be pre-filtered based on the mutation frequency across samples.

Because we never explicitly evaluate the largest set of mutual exclusive genes in the interaction network, the prioritized mutated genes are not guaranteed to be all mutually exclusive or to all belong the same network neighborhood. SSA-ME rather prioritizes genes that belong to local mutual exclusivity patterns. If one would be interested in finding the largest set of mutual exclusive genes or independent modules, the prioritized gene list would suffice as input for *de novo* discovery methods for mutual exclusivity, such as Dendrix,

Multidendrix or CoMEt. However, given the incompleteness of the interaction network and the biology of clonal systems, imposing too strong global constrains, e.g. requiring that all genes belonging to a mutual exclusivity pattern should also all be closely connected in the network, might reduce the number of selected potential driver genes. This because patterns of mutual exclusivity can be broken because genes belonging to a specific pattern can be unconnected in the interaction network due to missing interactomics data. In addition, if mutations trigger different adaptive pathways that are, when occurring in the same tumor, synthetically lethal, the genes carrying the mutually exclusive mutations would belong to different local regions in the interaction network (incompatible pathways) that cannot co-occur in the same tumor.

We showed that the results obtained by SSA-ME were largely consistent with those obtained by MEMo on the same TCGA 2012 dataset(Cancer Genome Atlas Network 2012) and that the use of a stringently versus an non-stringently filtered input set did not deteriorate the quality of those findings. By applying on the same breast cancer dataset SSA-ME with mutational data that were not a priori filtered based on recurrence across samples, we could show the potential of the method in discovering rarely mutated driver genes. In addition to drivers reported in well-known databases, our method prioritized an additional 9 drivers, not yet covered by conventional cancer databases. Several of these additional drivers were found to be infrequently mutated in the breast cancer and pan-cancer datasets and therefore have been missed by statistical prioritization methods that rely on recurrence of mutations across different tumors.

Conclusively, SSA-ME allows exploiting network connectivity and mutual exclusivity to identify drivers. Because of its computational efficiency, it can be used without relying on filtering non-recurrent alterations and as such allows for the detection of infrequently mutated drivers.

Chaper 5. CONCLUSIONS AND PERSPECTIVES

CONCLUSIONS

This is thesis is a compilation of bioinformatics methods developed in the last four years capable of using different scenarios and predictions made by selection to search for marks left in the DNA to accomplish their specific goals.

Detection of QTLs is widely used in industry and research to pinpoint genomic loci that control traits of interest. These methods are used in biotechnology for food production (Takagi et al. 2015; Abe et al. 2012) and the biofuel industry (Meijnen et al. 2016; Wenger et al. 2010; Pais et al. 2013), among others. With EXPLoRA, we were able to use the linkage disequilibrium expected of sexually reproducing populations under selection to obtain a smaller and more defined QTLs. The method was able to work on difficult settings like having a small number of successful individuals because of very stringent selection and to also detect not very strong effects. As QTL detection is a mandatory step in several areas of research, we also provided an easy to use, user friendly and accessible web server able to perform the analysis. Something that is not yet accomplished is a more seamless integration with different pipelines of data generation and those downstream analysis made possible by the identification of QTL for gene prioritization.

Trying to decipher the evolutionary trajectory of a population have always been an important aspect of biology. Some methods look at this problem at the level of species or distant relatives and propose phylogenetic trees where the evolution can be seen. Several tools exists to analyze quasispecies of virus after infection (Giallonardo et al. 2014; Töpfer et al. 2014; Zagordi et al. 2011; Huang et al. 2011; Prabhakaran et al. 2013), in which several subpopulations become common and start competing between them in clonal interference.

But there were no methods capable to solve this problem for bacteria, organisms with clonal reproduction and a relatively low mutation rate. EVORhA was built under the expectation that once a fitness increasing mutation occurs, it will happen over a genomic background possessing several hitchhiking mutations that will become prominent due to the lack of recombination. And those mutations, as group can be detected. Additionally, the method possess some features that can be adapted to improve or build new tools used in similar fields: the use of codon information, the unbound length of windows. The type of analysis performed by EVORhA open several possibilities to further study and understand mixed infections, metagenomics samples and evolutionary dynamics.

The analysis of biological networks is a great tool to interpret high-throughput data. With the advent of larger quantities of data it becomes imperative to propose algorithms that scale with the amount of data. Many problems in bioinformatics suffer from a high computational complexity, which make its use cumbersome and resource intensive even for small problems. In most network analysis tools is difficult to couple the computational calculations in a biological network (be it the shortest paths, diffusion, similarity matrices, or any other method) to the biology underneath. For these reasons, we introduced SSA to solve some of the problems that could benefit with the use of a biological network. SSA is an algorithm of linear complexity that can be used to evaluate different biological questions when the answer can be found partially in close proximity to the causal genes. To test the SSA framework we searched for driver genes in cancer using mutual exclusivity as its driving force. We expect the framework to be relevant for other complex problems as it is capable to process large datasets in polynomial time and its intuitive implementation.

PERSPECTIVES

QTL analysis is a widely used technique that have changed with the appearance of new technologies and new research fields. QTL finding remains a mayor challenge in genetics. Understanding the genetic basis of variation of quantitative traits not only help us to make use of those genetic differences in biotechnological applications but also to gain insights about the underlying genetic architecture of the traits (Mackay et al. 2009). Additionally, QTL analysis have been adapted to new technologies. For example the use of expression as a phenotype (eQTL) or protein levels as a phenotype (pQTL), among other phenotypes of interest have shown the potential of QTL as a basic concept of genetics (Albert & Kruglyak 2015). The future of QTL mapping hold great challenges like the detection of epistatic interactions of small effect loci, or detecting QTLs whose effect dependent on the environment. With the work of this thesis we provided tools capable of higher precision for detecting QTLs and, no only better in performance, but also tools that are easier to use and more reachable. DNA sequencing data is more pervasive every day, there is a need for biologists to use more advanced bioinformatics tools as more data is available but also a need for bioinformatics to make more easily accessible and usable methods. We need to meet half-way.

Analyses of populations and evolutionary dynamics depend on understanding and knowing the composition of the population under study. Fast adapting clonal populations are genetically diverse and far from homogeneous. This is a feature observed in viruses, bacteria and cancer. It is common to observe infections or tumors that become resistant to drugs by natural selection (Beerenwinkel et al. 2012; Yates & Campbell 2012; Eyre et al. 2013). Before the drug treatment populations are genetically diverse with few drug resistant individuals that struggle to compete with the rest of the population. Once the drug is administered, those individuals resistant to the drug have a selective advantage in the new drug filled environment and take over it. In this thesis

we made advancements related to the study of bacterial populations. Further research will be necessary to understand and analyze the interplay between the different sub-populations and how it affects the population dynamics in time.

With Every passing day more complex biological questions are being asked. Biological network methods have been used in many ways to answer these questions. In this endeavor the bioinformatics community have used diffusion methods (Verbeke et al. 2015; M. D. M. Leiserson, Vandin, et al. 2015), shortest paths methods (De Maeyer et al. 2013) and identification of network motifs (Yeger-Lotem et al. 2004; Alon 2007), among others. A common difficulty in network analysis is how to interpret a graph algorithm result as a biological explanation, e.g. how to biologically interpret a dijkstra minimum spanning tree. Additionally, network methods that want to solve problems that are difficult to interpret using any graph algorithm (e.g. mutual exclusivity) possess a high algorithmic complexity, therefore are not scalable to large datasets. With the work presented in this thesis we have provided a framework that use biological networks and is capable of scalable solutions for problems that cannot (or at least are very difficult to) be translated into known algorithms or graph properties. With the proposed framework, answering a complex biological question is transformed into finding an evaluation function that score small subnetworks for how well they fit what the problem is looking for.

Future work

Integration of EXPLoRA-web with network based analyses

After detecting QTLs it is normally necessary to use some kind of gene prioritization technique. QTLs normally encompass genomic regions spanning several genes and is not experimentally feasible to test all these genes. EXPLoRA could be seamlessly integrated with gene prioritization tools. These tools could make use of additional omics data produced in the experiment. The combination of EXPLoRA with other tools developed in our group like PheNetic

(De Maeyer et al. 2013) or SSA (Pulido-Tamayo et al. 2015) in a user friendly and straight manner will increase the usability of the complete pipeline and therefore increase the usage of the tools than what would be accomplished if the tools were to be presented independently.

Evolutionary trajectories using time resolution

EVORhA is a tool capable to reconstruct the population structure using deep sequencing data of bacterial pooled sequenced samples. An interesting next research step is to not only analyze the information per time point as independent but to use the samples from different time points to take better decisions and reconstruct with greater quality the evolutionary dynamics of the population.

Using SSA for more complex tasks

SSA can be used for different problems that could be solved by evaluating the local neighborhood in a protein interaction network.

Finding marks of positive selection

Finding marks of positive selection in genes to determine cancer genes is a very active field of research (Gonzalez-Perez & Lopez-Bigas 2012; Van den Eynden et al. 2015; Lawrence et al. 2013). All current methods analyses mutation on a gene level. To our knowledge only one have extrapolated their analysis to the level of known cancer pathways (Gonzalez-Perez & Lopez-Bigas 2012) and the results of this kind of analysis are not highly valuable as those pathways are already known.

We propose to use SSA by scoring subnetworks by using deleteriousness scores like CADD (Kircher et al. 2014) to search for driver cancer genes that show certain mutation signatures concordant to the known cancer genes.

FINDING EPISTASIS IN QTLS

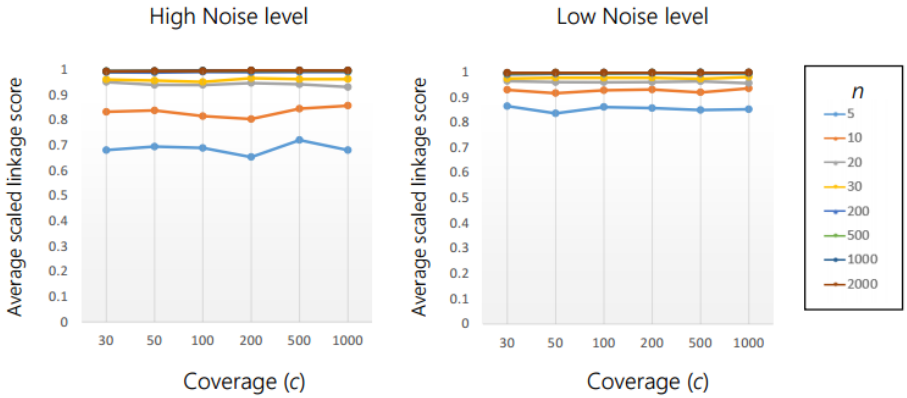
QTL analysis are used on a daily basis for industry and medical applications to find genes responsible of almost any phenotype of interest, e.g. ethanol resistance for beer and biofuel industries, biomass production for bio-energy companies, meat production for the cattle industry, hypertension susceptibility experiments, etc.

We can use SSA to evaluate in conjunction to the information from linkage disequilibrium, the information available in protein interaction networks. We will need to define a subnetwork evaluation score to find alleles that show an insignificant effect on the phenotype if they are analyzed independently, but show an important effect if they are analyzed in conjunction. The algorithm could be used to detect QTLs related to bio-ethanol production in saccharomyces cerevisiae and biomass production in Arabidopsis Thaliana.

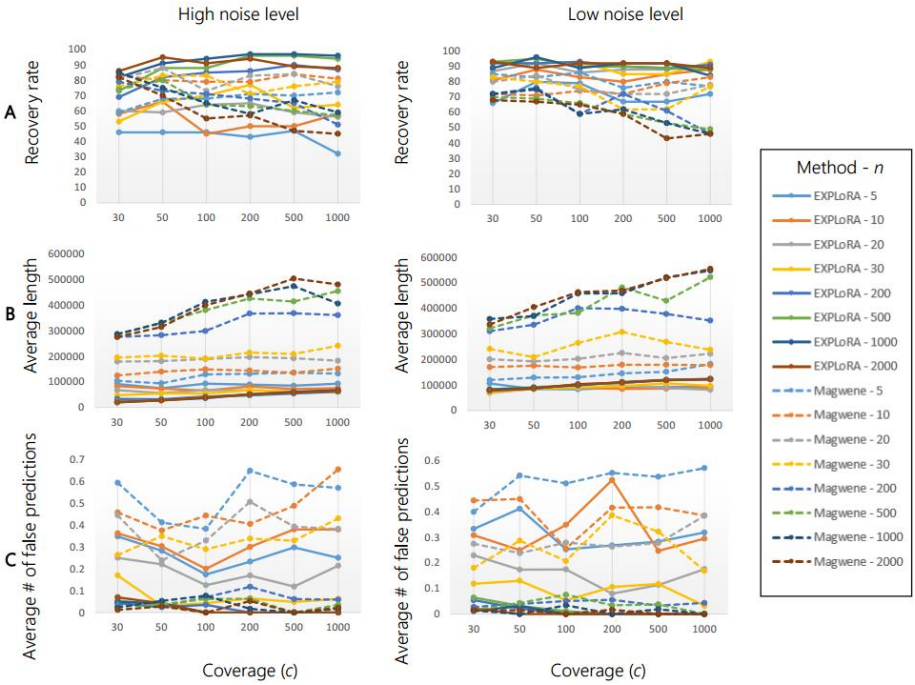
CURRICULUM

Sergio Pulido Tamayo was born in Medellín (Colombia) on March 18th, 1982. In 1999 he started his education in EAFIT University and received his bachelor degree in Computer Science in 2005. As bachelor student he did his first activities as a researcher while doing a one year internship in the Institute for Human and Machine Cognition (IHMC), USA in 2003. After his bachelor degree he worked for a couple of software companies developing products and services for insurance, banking and finance companies. His interest in biology and evolution led him to start a master in bioinformatics at KU Leuven in 2010. He is currently doing a joint PhD in UGent and KU Leuven, developing computational methods for systems biology by combining advanced statistical and modeling techniques with evolutionary assumptions.

Appendix A. SUPPLEMENTARY DATA CHAPTER 2

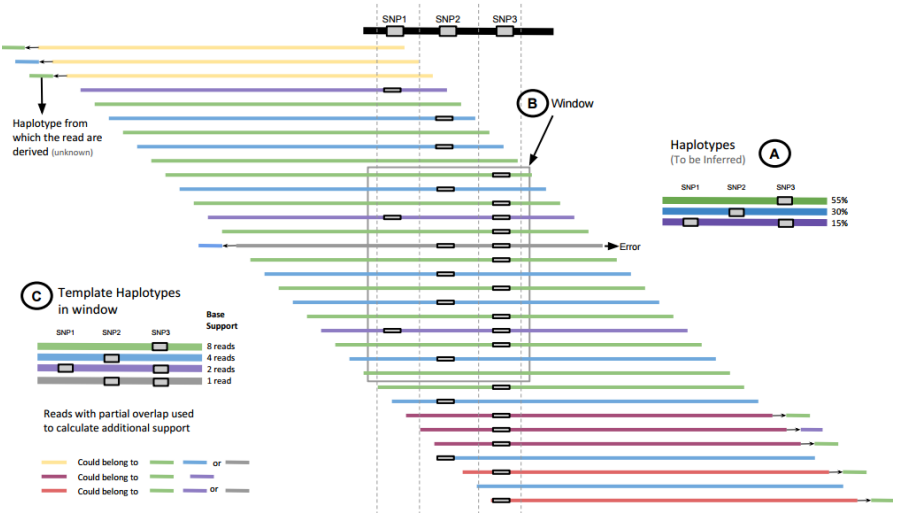


Supp Figure A. 1 Average scaled linkage score at the causal site reported by the method of Magwene et al. as a function of the coverage and noise levels **(panel A)** and low **(panel B)** noise levels. Raw values of the G' statistics at the causal site (G'_{causal}) were scaled taking into the maximum (G'_{max}) and minimum (G'_{min}) G's values from the entire artificial chromosome according to the following formula: $G'_{scaled} = (G'_{causal} - G'_{min})/(G'_{max} - G'_{min})$. Reported values correspond to the average of 100 repetitions.



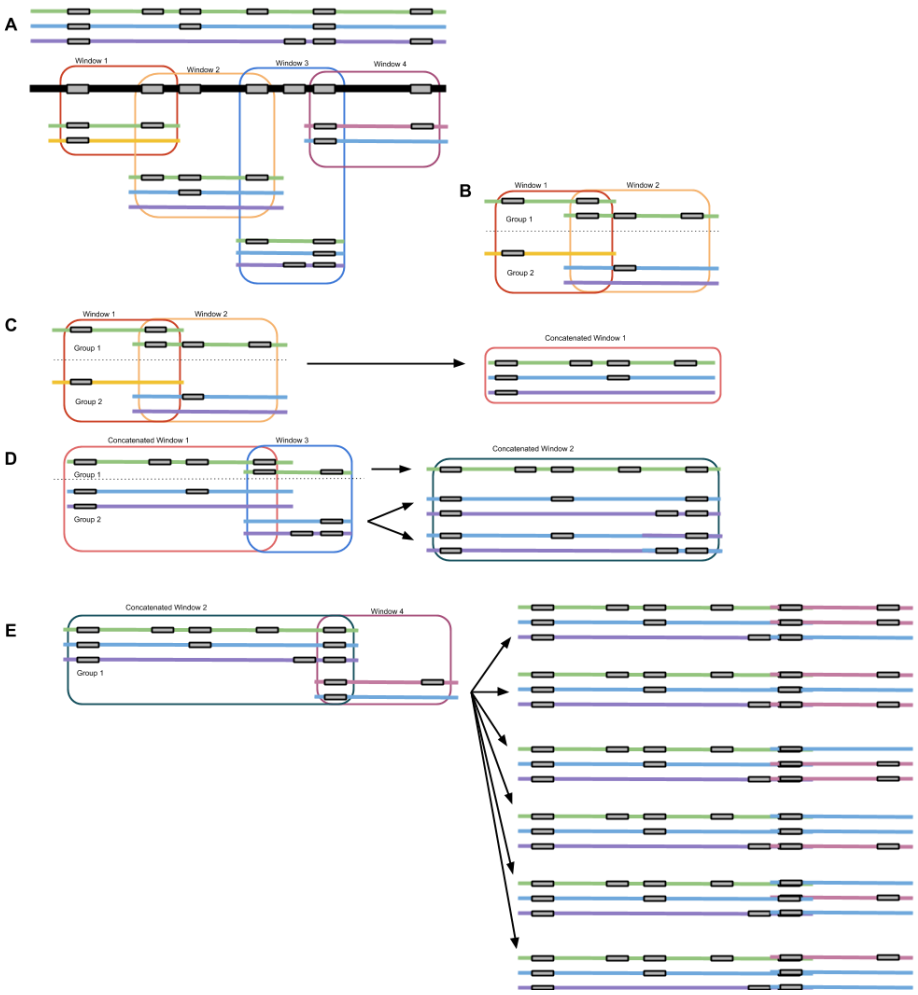
Supp Figure A. 2 Comparison with the state-of-the-art.
The recovery rate **(panel A)**, average size of the linked region **(panel B)** and number of falsely predicted regions **(Panel C)** under high (left sided plots) and low (right sided plots) noise levels were assessed for EXPLoRA and the method of Magwene et al. For the method of Magwene et al. [7] the less stringent correction for multiple testing, which does not take into account dependency between tests, was used.

Appendix B. SUPPLEMENTARY DATA CHAPTER 3



Supp Figure B. 1 Inferring template haplotypes and performing error correction at the local scale. The example population consists of three haplotypes (**A**: *Green* at 55%, *Blue* at 30% and *Purple* at 15%). Reads obtained from sequencing this population were mapped to a reference genomic region (Black), allowing the detection of polymorphic sites. To differentiate true variants from sequencing errors in a window (**B**), a set of template haplotypes is defined from those reads that fully cover the window (indicated in respectively lighter *green*, *blue*, *purple* and *grey)*. For each template haplotype a *support* (**C**) is calculated using 1) the reads that fully overlap with the window (i.e. *base support*, hereby assuming a perfect base call accuracy for simplification purposes, i.e. $w(r) = 1$) , but also using 2) all remaining reads that partially overlap with the window and that are consistent with the template haplotypes (i.e. *additional support*). These partially overlapping reads give an additional support to each template haplotype with which they are consistent, proportional to the *base support* of each template haplotype. In the example represented here, the three yellow reads match to the *green*, *blue* and *grey* haplotypes (as they do not contain a mutation in SNP1). Each yellow read gives additional support to all consistent template haplotypes, taking into account the *base support* of the matching haplotypes e.g. 8 is the base support of the green haplotype, 4 is the base support of the blue and 1 of the grey, see figure insets **B** and **C**. The sum of the haplotypes base supports matching the yellow reads is 13. Therefore, the contribution of each yellow read to the additional support

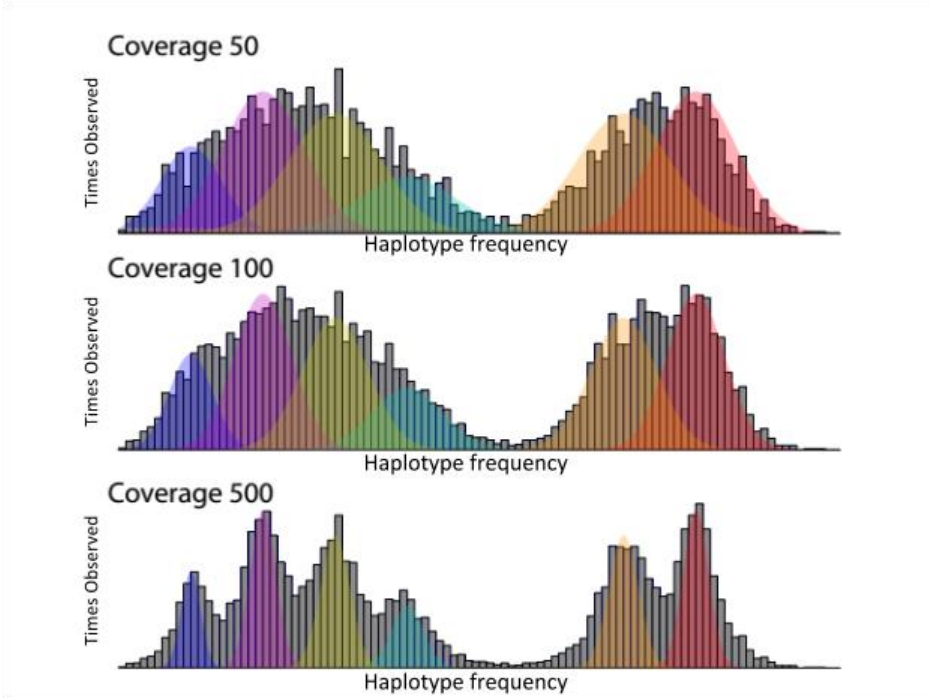
of the matching haplotypes is calculated as follows: 8/13 for green, 4/13 for blue and 1/13 for grey. Only template haplotypes with a sufficient total support will be maintained for further analysis, hereby assuming that erroneous template haplotypes that are the result of a sequencing error will not have sufficient total support. In the example, the grey template ends up with a total support of 1.42 (1 base support + 3 yellow reads that each gives 1/13 additional support + 2 pink reads that each gives 1/11 additional support) while the other haplotypes obtain a much larger total support: 3.96 for purple (2 base support + 1 partially overlapping read only consistent to the purple template + 3 magenta reads with 2/11 additional support each + 2 pink reads with 2/11 additional support), 9.93 for blue (4 base support + 5 additional support from reads with partial overlap consistent only to the blue template + 3 yellow reads with 4/13 additional support each) and 17.70 for green (8 base support + 4 partially overlapping reads consistent only to this template + 3 yellow reads with 8/13 additional support each + 3 magenta reads with 8/10 additional support each + 2 pink reads with 8/11 additional support each).



Supp Figure B. 2 Window extension procedure

In the example four windows contain reads from the sequencing of a population of three haplotypes (green, blue and purple). These windows can be extended because they have polymorphisms in their read overlap. EVORhA choses a seed window to start the window extension (say, Window 2). Window 2 is extended with its flanking windows (window 1 and window 3) starting by the one on the left. Groups are declared by combining template haplotypes from both windows that share the same unique combination of polymorphisms in the overlap region of the windows. In this case (B) there are two groups: template

haplotypes belonging to group 1 share the SNP in their overlap region, whereas those of group 2 share “not having the SNP”. The goal of the extension is to find within a group the best concatenation of template haplotypes from respectively the first and second window to generate an extended haplotype, where ‘best’ is defined in terms of matching frequencies and shared polymorphisms. (A) A toy example of an extension is given below: we start by joining window 1 and 2 (B), in which two groups can be distinguished. The extension is performed for each group separately. This is an example of two straightforward extensions as group 1 only contains exactly one template haplotype in either window, whereas group 2 only contains one haplotype in window 1. For both groups the extension consists of concatenating the template haplotypes of the flanking windows. The extended haplotypes belong to an artificial window referred to as Concatenated Window 1 (C), which will be further extended with window 3 (D). Again, the template haplotypes are separated in two groups, a group containing the template haplotypes that share the SNP (group 1) and those without the SNP (group 2) in their overlap region. The extension of group 1 is again straightforward and a concatenation is sufficient. For group 2 the extension is ambiguous (E) as the concatenation can result either in the true haplotypes or a chimera. To solve this, an Expectation Maximization algorithm is used to find which concatenation is the most likely based on the frequencies of the template haplotypes within a group. So, we assume that a template haplotype in a window will have a more similar frequency with its true counterpart in the flanking window than with any other template haplotype in the same flanking window. In the next extension, when extending using window 4, we can only divide the haplotypes into a high complexity extendable group where many more combinations are possible. The window overlap information here becomes null as all haplotypes share the same mutations. The responsibility of the EM is to find the most likely way to concatenate the haplotypes based on the haplotype frequencies in the group.



Supp Figure B. 3 Figure 3S. Frequency analysis
Four genome-wide haplotypes were simulated, occurring at a frequency of respectively 10%, 20%, 30% and 40% in a given population. Pooled sequence data were simulated and the frequency with which 'extended haplotypes' were detected in each concatenated window was recorded. The histogram displays the number of extended haplotypes (Y-axis) that occur at a given frequency (X-axis). This histogram shows how several extended haplotypes, here referred to as haplotype sets occur at approximately the same frequency. The simulation shows that the distribution of the frequency at which extended haplotypes sets occur can be modelled by six independent Gaussian distributions. Four of these Gaussians correspond to haplotype sets containing polymorphisms unique to each of the respective simulated genome-wide haplotypes (*blue*, *purple*, *yellow* and *cyan*), whereas the two Gaussians observed at higher frequencies correspond to haplotype sets containing polymorphisms shared by several genome-wide haplotypes (*orange* and *red*). The goal of the frequency analysis is to determine a mixture model of Gaussian distributions that optimally describes the observed frequencies at which the haplotype sets occur. Extended haplotypes belonging to the same set likely belong to the same 'genome-wide haplotype' provided they do not contain any conflicting polymorphisms at the same polymorphic site. The haplotype sets are subsequently used to infer genome-wide haplotypes, hereby assuming that for each

considered haplotype set a combination of haplotype sets might exist, occurring at a frequency lower than the frequency of the considered haplotype set and for which the sum of these respective frequencies equals the frequency of the haplotype set under consideration e.g. the *orange* haplotype set frequency is the sum of the *cyan* and *purple* frequencies.

Appendix C. SUPPLEMENTARY DATA CHAPTER 4

LITERATURE BASED EVIDENCE FOR NEWLY PREDICTED CANCER DRIVERS

Constitutive expression of **IKBKB** (IKKβ) has been identified in tumors of epithelial origin, including breast cancer. In addition IKKβ signaling has in ovarian cancer been shown to regulate the transcription of genes involved in a wide range of cellular effects known to increase the aggressive nature of the cells(Hernandez et al. 2010).

Of the genes we prioritized both **GAB2** and *PAK1* belong to the same amplicon as the well-known breast cancer driver *CCND1*(Hennessy et al. 2006). However, because it cannot be excluded that possibly more genes in the same amplicon are causal to cancer and because *CCND1*, *GAB2*, and *PAK1* each show a strong mutual exclusivity with a subset of selected genes closely located in the network, but not between themselves, each of them might act independently from one another as a true driver. Whereas both CCND1, a regulatory protein involved in mitosis, and PAK1, a protein belonging to the family of serine/threonine p21-activating kinases that are involved in cytoskeleton reorganization and nuclear signaling, have been reported in at least one of the cancer related databases, GAB2 has not. GAB2 was in our analysis prioritized because of its mutual exclusivity and close network connectivity with amongst others PIK3CA. In addition, the mutations within GAB2 have high CADD(Kircher et al. 2014) scores in breast cancer. GAB2 is a scaffolding adapter protein that transduces cellular signals between receptors (tyrosine kinase receptors) and intracellular downstream effectors (PI3K, SpH2) and is required for efficient ErbB2-driven mammary tumorigenesis and metastatic spread by acting downstream of ErbB2(Ding et al. 2015; Bocanegra et al. 2010). Interestingly, it was also shown that a focal amplification of GAB2 independent of CCND1 in breast tumors, contributes to diverse oncogenic phenotypes in breast cancer by activating amongst others the PI3K pathway,

further confirming its role as primary driver in breast cancer(Bocanegra et al. 2010).

MCL1 is involved in apoptosis modulation and signaling(Rückert et al. 2010). It is mainly altered by copy number alterations in both the breast cancer and pan-cancer datasets. MCL1 carries CNVs in 133 samples belonging to different cancer types, amongst which also breast cancer. It has been associated with a number of cancers because of its involvement in the regulation of apoptosis versus cell survival(Ertel et al. 2013).

VAV2, is a protein, known to be abundantly expressed in human breast cancer cells. Vav2 is involved in altering cell shape and migration by triggering cytoskeleton changes through affecting the activity of Rho family GTPases (Rho, RhoG, Rac, Cdc42)(Barrio-Real & Kazanietz 2012; Jones et al. 2013). As such it has been associated with metastasis in breast cancer.

INDEPENDENT VALIDATION OF THE PREDICTED DRIVERS

As an independent evaluation of the predicted cancer related genes, we tested to what extent these genes were enriched in carrying mutations with a likely phenotypic impact. We hereby assumed that mutations in true drivers are more likely to carry mutations with a phenotypic impact. To this end, we used previously described CADD scores, which measure the deleteriousness of mutations(Kircher et al. 2014). CADD scores were downloaded from http://cadd.gs.washington.edu/ version 1.3. Figure 4.6A shows the empirical distribution of the CADD scores of the mutations (SNVs) present in 6 out of the 9 novel prioritized genes (TK1 and CRK were excluded because they do not contain any mutations, whereas MCL1 was excluded because it contains CNVs only). As a positive and negative reference, we also show the CADD score distribution of respectively the mutations in the 40 prioritized genes previously associated with cancer and of a set containing all mutations. Mutations in the positive reference set have on average higher CADD scores than the ones in

the negative reference set ($p = 2{,}2*10^{-16}$ using a two-sample Wilcoxon rank sum test), indicating that for true positive drivers indeed we can expect higher CADD scores. The mutations of the set of novel predictions had on average significantly higher CADD scores than the mutations in the genes of the negative reference set ($p = 0.004735$ using a two-sample Wilcoxon rank sum test), indicating the putative functionality of these mutations. More specifically, out of the 6 potential novel driver genes which have at least one mutation, 5 genes have at least one mutation with a CADD score exceeding the 80% highest score of the negative reference (set containing all observed mutations). Appendix C Table 2 provides a list of all mutations present in these novel predicted driver genes together with their CADD scores.

Because the number of mutations in the genes we prioritized were rare, we also, to further back up their putative role in driving cancer, assessed to what extent each of these prioritized drivers were found mutated in the larger TCGA Pan-Cancer dataset(Cancer Genome Atlas Research Network et al. 2013) and how, if available, the retrieved mutations were distributed along the genomic positions of these genes. Pan-Cancer analysis was performed using MAGI(M. D. M. Leiserson, Gramazio, et al. 2015) for the visual examination, and SomInaClust(Van den Eynden et al. 2015) for detecting clustering of mutations as a signal of positive selection. Pan-cancer data containing information for 12 tumor types and 3281 tumors samples was download from Synapse (syn1710680).

Results of the genomic clustering are visualized on Supp Figure C. 2 and in Figure 4.6B for two representative candidates. Although at least visually for some of these driver candidates, the mutations they carry seem to cluster at the same genomic positions, none of them scores highly ‘significant’ for clustering of their mutations according to the results provided in the pan-cancer analysis(Cancer Genome Atlas Research Network et al. 2013) or the results we obtained by running SomInaClust(Van den Eynden et al. 2015). This

is to be expected as the p-values to prioritize drivers based on the clustering of their mutations requires a rather high power which is difficult to obtain for these rarely mutated genes. Note, however, that we prioritized these genes not because of their genomic clustering, but because they consistently displayed a mutual exclusivity pattern with other mutated genes located in their network neighborhood.

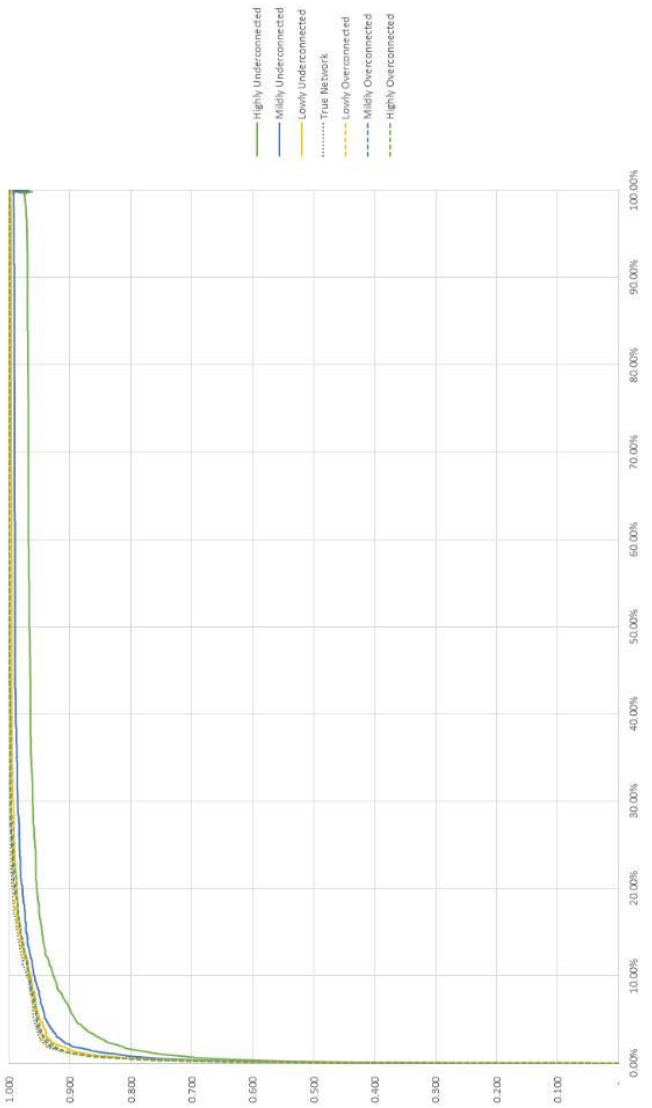
Table 1 Overview of the 49 prioritized drivers and their previous associations with cancer: Columns indicate true positive sets (CGC, Malacards or NCG).

Rank	GeneSymbol	Malacard	CGC	NCG
1	PIK3CA	TRUE	TRUE	TRUE
2	TP53	TRUE	TRUE	TRUE
3	MYC	TRUE	TRUE	TRUE
4	CCND1	TRUE	TRUE	TRUE
5	ERBB2	TRUE	TRUE	TRUE
6	AKT1	TRUE	TRUE	TRUE
7	PTEN	TRUE	TRUE	TRUE
8	CDH1	TRUE	TRUE	TRUE
9	STAT3	TRUE	TRUE	TRUE
10	MYB	TRUE	TRUE	TRUE
11	EGFR	TRUE	TRUE	TRUE
12	MDM2	TRUE	TRUE	TRUE
13	BRCA1	TRUE	TRUE	TRUE
14	ATM	TRUE	TRUE	TRUE
15	JUN	TRUE	TRUE	TRUE
16	CTNNB1	TRUE	TRUE	TRUE
17	MAP3K1	TRUE	FALSE	TRUE
18	AR	TRUE	FALSE	TRUE
19	MTOR	TRUE	FALSE	FALSE
20	CDKN1A	TRUE	FALSE	FALSE
21	PIK3R1	FALSE	TRUE	TRUE

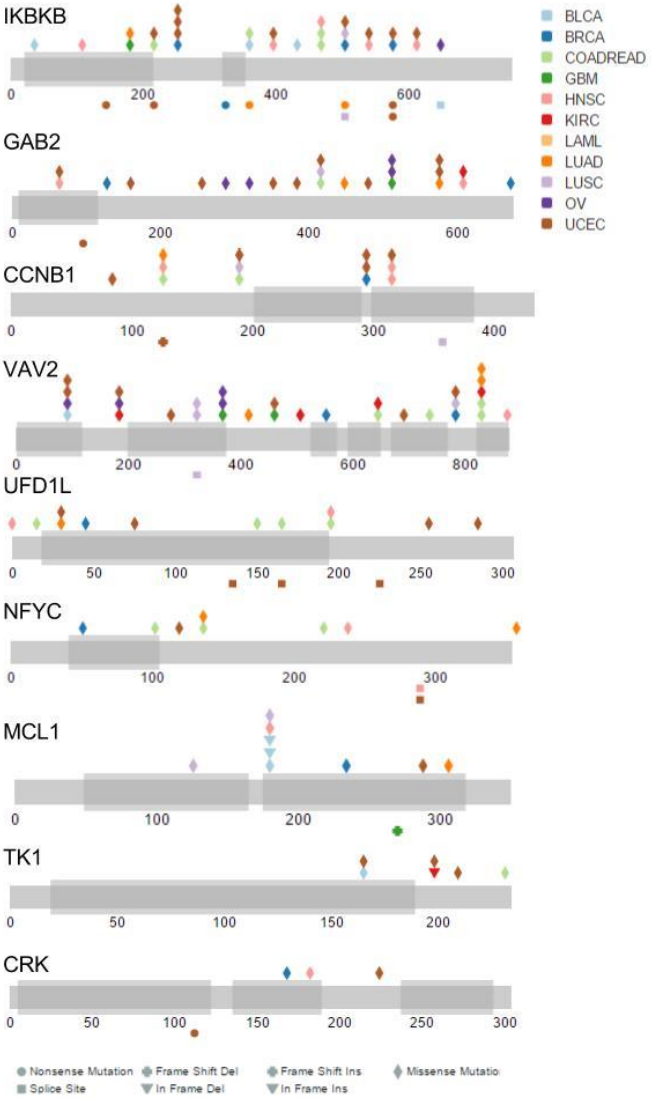
22	RB1	FALSE	TRUE	TRUE
23	DDX5	FALSE	TRUE	TRUE
24	FOXA1	FALSE	TRUE	TRUE
25	APC	FALSE	TRUE	TRUE
26	JAK2	FALSE	TRUE	TRUE
27	TRRAP	FALSE	TRUE	TRUE
28	PDGFRB	FALSE	TRUE	TRUE
29	PAK1	FALSE	FALSE	TRUE
30	RPS6KB1	FALSE	FALSE	TRUE
31	CDC42	FALSE	FALSE	TRUE
32	FOXO3	FALSE	FALSE	TRUE
33	SOS1	FALSE	FALSE	TRUE
34	RHOA	FALSE	FALSE	TRUE
35	GSK3B	FALSE	FALSE	TRUE
36	FRS2	FALSE	FALSE	TRUE
37	NCOR1	FALSE	FALSE	TRUE
38	GRB2	FALSE	FALSE	TRUE
39	MAX	FALSE	TRUE	FALSE
40	RAC1	FALSE	TRUE	FALSE
41	GAB2	FALSE	FALSE	FALSE
42	MCL1	FALSE	FALSE	FALSE
43	UFD1L	FALSE	FALSE	FALSE
44	CRK	FALSE	FALSE	FALSE
45	VAV2	FALSE	FALSE	FALSE
46	CCNB1	FALSE	FALSE	FALSE
47	TK1	FALSE	FALSE	FALSE
48	IKBKB	FALSE	FALSE	FALSE
49	NFYC	FALSE	FALSE	FALSE

Table 2 CADD scores of mutations in each of the 9 predicted drivers. CRK and TK1 are excluded as they did not carry any mutations and MCL1 is excluded as it only contains CNVs.

chromosome	Position on chromosome	Gene name	CADD score
1	41213269	NFYC	34
5	68467279	CCNB1	27.7
5	68470891*	CCNB1	26.3
5	68470891*	CCNB1	26.3
8	42171889	IKBKB	31
8	42177138	IKBKB	24.7
8	42179428	IKBKB	20.7
9	136634607	VAV2	11.24
9	136643908	VAV2	9.349
11	77930333	GAB2	28.8
11	77991673	GAB2	29.2
22	19462608	UFD1L	34



Supp Figure C. 1 Complete ROC Curve



Supp Figure C. 2 New Predictions mutations

REFERENCES

Abe, A. et al., 2012. Genome sequencing reveals agronomically important loci in rice using MutMap. *Nature biotechnology*, 30(2), pp.174–8. Available at: http://dx.doi.org/10.1038/nbt.2095 [Accessed December 29, 2015].

Albert, F.W. & Kruglyak, L., 2015. The role of regulatory variation in complex traits and disease. *Nature Reviews Genetics*, 16(4), pp.197–212. Available at: http://dx.doi.org/10.1038/nrg3891 [Accessed February 24, 2015].

Albertsen, M. et al., 2013. Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nature biotechnology*, 31(6), pp.533–8. Available at: http://www.ncbi.nlm.nih.gov/pubmed/23707974 [Accessed July 9, 2014].

Alexandrov, L.B. et al., 2013. Signatures of mutational processes in human cancer. *Nature*, 500(7463), pp.415–421. Available at: http://www.nature.com/doifinder/10.1038/nature12477 [Accessed August 14, 2013].

Alneberg, J. et al., 2014. Binning metagenomic contigs by coverage and composition. *Nature Methods*, (August). Available at: http://www.nature.com/doifinder/10.1038/nmeth.3103 [Accessed September 14, 2014].

Alon, U., 2007. Network motifs: theory and experimental approaches. *Nature reviews. Genetics*, 8(6), pp.450–61. Available at: http://www.ncbi.nlm.nih.gov/pubmed/17510665 [Accessed February 22, 2016].

An, O. et al., 2015. NCG 5.0: updates of a manually curated repository of cancer genes and associated properties from cancer mutational screenings. *Nucleic acids research*, p.gkv1123–. Available at: http://nar.oxfordjournals.org/content/early/2015/10/29/nar.gkv1123.full [Accessed November 5, 2015].

Astrovskaya, I. et al., 2011. Inferring viral quasispecies spectra from 454 pyrosequencing reads. *BMC bioinformatics*, 12, pp.90–101. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3194189&tool=pmcentrez&rendertype=abstract [Accessed November 9, 2012].

ATWOOD, K.C., SCHNEIDER, L.K. & RYAN, F.J., 1951. Periodic selection in Escherichia

coli. *Proceedings of the National Academy of Sciences of the United States of America*, 37(3), pp.146–55. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1063322&tool=pmcentrez&rendertype=abstract [Accessed March 17, 2016].

Austin, R.S. et al., 2011. Next-generation mapping of Arabidopsis genes. *The Plant journal : for cell and molecular biology*, 67(4), pp.715–25. Available at: http://www.ncbi.nlm.nih.gov/pubmed/21518053 [Accessed January 22, 2016].

Van der Auwera, G.A. et al., 2013. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. A. Bateman et al., eds. *Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.]*, 11(1110), pp.11.10.1–11.10.33. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4243306&tool=pmcentrez&rendertype=abstract [Accessed July 16, 2014].

Babur, O. et al., 2014. *Systematic identification of cancer driving signaling pathways based on mutual exclusivity of genomic alterations*, Cold Spring Harbor Labs Journals. Available at: http://genomebiology.com/2015/16/1/45 [Accessed February 22, 2015].

Banerjee, A. et al., 2005. Clustering with Bregman Divergences. *The Journal of Machine Learning Research*, 6, pp.1705–1749. Available at: http://dl.acm.org/citation.cfm?id=1046920.1194902 [Accessed August 19, 2014].

Barrick, J. & Lenski, R., 2009. Genome-wide mutational diversity in an evolving population of Escherichia coli. *Cold Spring Harb Symp Quant Biol*, 74(517), pp.119–129. Available at: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2890043/ [Accessed January 30, 2013].

Barrick, J.E. & Lenski, R.E., 2013. Genome dynamics during experimental evolution. *Nature reviews. Genetics*, 14(12), pp.827–39. Available at: http://dx.doi.org/10.1038/nrg3564 [Accessed July 10, 2014].

Barrio-Real, L. & Kazanietz, M.G., 2012. Rho GEFs and cancer: linking gene expression and metastatic dissemination. *Science signaling*, 5(244), p.pe43. Available at: http://www.ncbi.nlm.nih.gov/pubmed/23033535 [Accessed December 4, 2015].

Bassil, Y., 2012. A Comparative Study on the Performance of Permutation Algorithms. *arXiv preprint arXiv:1205.2889*, pp.7–19. Available at: http://arxiv.org/abs/1205.2888 [Accessed March 4, 2015].

Beerenwinkel, N. et al., 2012. Challenges and opportunities in estimating viral genetic diversity from next-generation sequencing data. *Frontiers in microbiology*, 3(September), p.329. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3438994&tool=pmcentrez&rendertype=abstract [Accessed January 31, 2014].

Benjamini, Y. & Yekutieli, D., 2005. Quantitative trait Loci analysis using the false discovery rate. *Genetics*, 171(2), pp.783–90. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1456787&tool=pmcentrez&rendertype=abstract [Accessed February 17, 2016].

Beroukhim, R. et al., 2007. Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma. *Proceedings of the National Academy of Sciences of the United States of America*, 104(50), pp.20007–12. Available at: http://www.pnas.org/content/104/50/20007.long [Accessed September 11, 2015].

Birkeland, S.R. et al., 2010. Discovery of mutations in Saccharomyces cerevisiae by pooled linkage analysis and whole-genome sequencing. *Genetics*, 186(4), pp.1127–37. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2998298&tool=pmcentrez&rendertype=abstract [Accessed March 15, 2016].

Bocanegra, M. et al., 2010. Focal amplification and oncogene dependency of GAB2 in breast cancer. *Oncogene*, 29(5), pp.774–9. Available at: http://dx.doi.org/10.1038/onc.2009.364 [Accessed November 20, 2015].

Bonangelino, C.J., Chavez, E.M. & Bonifacino, J.S., 2002. Genomic screen for vacuolar protein sorting genes in Saccharomyces cerevisiae. *Molecular biology of the cell*, 13(7), pp.2486–501. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=117329&tool=pmcentrez&rendertype=abstract [Accessed March 15, 2016].

Bradley, J.R. & Farnsworth, D.L., 2009. Testing for Mutual Exclusivity. *Journal of Applied Statistics*, 36(11), pp.1307–1314. Available at: http://www.tandfonline.com/doi/abs/10.1080/02664760802582306?journalC

ode=cjas20 [Accessed November 5, 2015].

Cancer Genome Atlas Network, 2012. Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418), pp.61–70. Available at: http://dx.doi.org/10.1038/nature11412 [Accessed July 9, 2014].

Cancer Genome Atlas Research Network et al., 2013. The Cancer Genome Atlas Pan-Cancer analysis project. *Nature genetics*, 45(10), pp.1113–20. Available at: http://www.nature.com/doifinder/10.1038/ng.2764 [Accessed July 11, 2014].

Cherry, J.M. et al., 1997. Genetic and physical maps of Saccharomyces cerevisiae. *Nature*, 387(6632 Suppl), pp.67–73. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3057085&tool=pmcentrez&rendertype=abstract [Accessed February 26, 2016].

Ciriello, G. et al., 2012. Mutual exclusivity analysis identifies oncogenic network modules. *Genome research*, 22(2), pp.398–406. Available at: http://genome.cshlp.org/cgi/doi/10.1101/gr.125567.111 [Accessed July 11, 2014].

Croft, D. et al., 2014. The Reactome pathway knowledgebase. *Nucleic acids research*, 42(Database issue), pp.D472–7. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3965010&tool=pmcentrez&rendertype=abstract [Accessed October 17, 2014].

Cubillos, F.A. et al., 2011. Assessing the complex architecture of polygenic traits in diverged yeast populations. *Molecular ecology*, 20(7), pp.1401–13. Available at: http://www.ncbi.nlm.nih.gov/pubmed/21261765 [Accessed March 15, 2016].

Das, J. & Yu, H., 2012. HINT: High-quality protein interactomes and their applications in understanding human disease. *BMC systems biology*, 6(1), p.92. Available at: http://www.biomedcentral.com/1752-0509/6/92 [Accessed November 5, 2015].

Ding, C.-B. et al., 2015. Structure and function of Gab2 and its role in cancer (Review). *Molecular medicine reports*, 12(3), pp.4007–14. Available at: http://www.ncbi.nlm.nih.gov/pubmed/26095858 [Accessed November 10, 2015].

Duitama, J. et al., 2014. An integrated framework for discovery and genotyping of genomic variants from high-throughput sequencing experiments. *Nucleic acids*

research, 42(6), p.e44. Available at: http://nar.oxfordjournals.org/content/42/6/e44.long [Accessed January 3, 2016].

Duitama, J., Srivastava, P.K. & Măndoiu, I.I., 2012. Towards accurate detection and genotyping of expressed variants from whole transcriptome sequencing data. *BMC genomics*, 13 Suppl 2, p.S6. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3394419&tool=pmcentrez&rendertype=abstract [Accessed March 15, 2016].

Edwards, M.D. & Gifford, D.K., 2012. High-resolution genetic mapping with pooled sequencing. *BMC bioinformatics*, 13 Suppl 6(Suppl 6), p.S8. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3358661&tool=pmcentrez&rendertype=abstract [Accessed July 18, 2012].

Ehrenreich, I.M. et al., 2010. Dissection of genetically complex traits with extremely large pools of yeast segregants. *Nature*, 464(7291), pp.1039–42. Available at: http://dx.doi.org/10.1038/nature08923 [Accessed December 11, 2015].

Endler, J.A., 1983. Natural and sexual selection on color patterns in poeciliid fishes. *Environmental Biology of Fishes*, 9(2), pp.173–190. Available at: http://link.springer.com/10.1007/BF00690861 [Accessed March 17, 2016].

Ertel, F. et al., 2013. Programming cancer cells for high expression levels of Mcl1. *EMBO reports*, 14(4), pp.328–36. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3617356&tool=pmcentrez&rendertype=abstract [Accessed September 20, 2015].

Van den Eynden, J. et al., 2015. SomInaClust: detection of cancer genes based on somatic mutation patterns of inactivation and clustering. *BMC bioinformatics*, 16(1), p.125. Available at: http://www.biomedcentral.com/1471-2105/16/125 [Accessed November 20, 2015].

Eyre, D.W. et al., 2013. Detection of mixed infection from bacterial whole genome sequence data allows assessment of its role in Clostridium difficile transmission. *PLoS computational biology*, 9(5), p.e1003059. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3642043&tool=pmcentrez&rendertype=abstract [Accessed February 25, 2014].

Futreal, P.A. et al., 2004. A census of human cancer genes. *Nature reviews. Cancer*, 4(3), pp.177–83. Available at:

http://www.nature.com/doifinder/10.1038/nrc1299 [Accessed July 10, 2014].

Giallonardo, F. Di et al., 2014. Full-length haplotype reconstruction to infer the structure of heterogeneous virus populations. *Nucleic acids research*, p.gku537–. Available at: http://nar.oxfordjournals.org/content/early/2014/06/27/nar.gku537.abstract [Accessed July 23, 2014].

Glenn, T.C., 2011. Field guide to next-generation DNA sequencers. *Molecular ecology resources*, 11(5), pp.759–69. Available at: http://www.ncbi.nlm.nih.gov/pubmed/21592312 [Accessed July 6, 2015].

Gonzalez-Perez, A. & Lopez-Bigas, N., 2012. Functional impact bias reveals cancer drivers. *Nucleic acids research*, 40(21), p.e169. Available at: http://nar.oxfordjournals.org/content/early/2012/08/14/nar.gks743.long [Accessed March 10, 2015].

Greaves, M. & Maley, C.C., 2012. Clonal evolution in cancer. *Nature*, 481(7381), pp.306–13. Available at: http://dx.doi.org/10.1038/nature10762 [Accessed July 9, 2014].

Hahn, W.C. & Weinberg, R.A., 2002. Modelling the molecular circuitry of cancer. *Nature reviews. Cancer*, 2(5), pp.331–41. Available at: http://www.nature.com/doifinder/10.1038/nrc795 [Accessed August 15, 2015].

Hayden, E.J., Ferrada, E. & Wagner, A., 2011. Cryptic genetic variation promotes rapid evolutionary adaptation in an RNA enzyme. *Nature*, 474(7349), pp.92–5. Available at: http://www.ncbi.nlm.nih.gov/pubmed/21637259 [Accessed July 16, 2014].

Hennessy, B.T. et al., 2006. Identification of remodeling and spacing factor 1 (rsf-1, HBXAP) at chromosome 11q13 as a putative oncogene in ovarian cancer. *European journal of human genetics : EJHG*, 14(4), pp.381–3. Available at: http://dx.doi.org/10.1038/sj.ejhg.5201570 [Accessed November 5, 2015].

Hernandez, L. et al., 2010. Activation of NF-kappaB signaling by inhibitor of NF-kappaB kinase beta increases aggressiveness of ovarian cancer. *Cancer research*, 70(10), pp.4005–14. Available at: http://cancerres.aacrjournals.org/content/70/10/4005.long [Accessed November 20, 2015].

Heywood, J.L. et al., 2011. Capturing diversity of marine heterotrophic protists: one cell at a time. *The ISME journal*, 5(4), pp.674–84. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3105736&tool=pmcentrez&rendertype=abstract [Accessed August 6, 2013].

Hill, W.G. & Robertson, A., 1968. Linkage disequilibrium in finite populations. *TAG. Theoretical and applied genetics. Theoretische und angewandte Genetik*, 38(6), pp.226–31. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24442307 [Accessed March 15, 2016].

Huang, A. et al., 2011. QColors : An Algorithm for Conservative Viral Quasispecies Reconstruction from Short and Non-Contiguous Next Generation Sequencing Reads. *IEEE International Conference on Bioinformatics and Biomedicine Workshops*, 11(5), pp.193–201.

International Cancer Genome Consortium et al., 2010. International network of cancer genome projects. *Nature*, 464(7291), pp.993–8. Available at: http://www.nature.com/doifinder/10.1038/nature08987 [Accessed July 9, 2014].

Jones, M.C. et al., 2013. Paxillin kinase linker (PKL) regulates Vav2 signaling during cell spreading and migration. *Molecular biology of the cell*, 24(12), pp.1882–94. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3681694&tool=pmcentrez&rendertype=abstract [Accessed December 4, 2015].

Kandoth, C. et al., 2013. Mutational landscape and significance across 12 major cancer types. *Nature*, 502(7471), pp.333–9. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3927368&tool=pmcentrez&rendertype=abstract.

Kanehisa, M. et al., 2008. KEGG for linking genomes to life and the environment. *Nucleic acids research*, 36(Database issue), pp.D480–4. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2238879&tool=pmcentrez&rendertype=abstract [Accessed April 17, 2015].

Kao, K.C. & Sherlock, G., 2008. Molecular characterization of clonal interference during adaptive evolution in asexual populations of Saccharomyces cerevisiae. *Nature genetics*, 40(12), pp.1499–504. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2596280&tool=p

mcentrez&rendertype=abstract [Accessed August 25, 2014].

Kircher, M. et al., 2014. A general framework for estimating the relative pathogenicity of human genetic variants. *Nature genetics*, 46(3), pp.310–5. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24487276 [Accessed July 9, 2014].

Lang, G.I., Botstein, D. & Desai, M.M., 2011. Genetic variation and the fate of beneficial mutations in asexual populations. *Genetics*, 188(3), pp.647–61. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3176544&tool=pmcentrez&rendertype=abstract [Accessed August 14, 2014].

Langmead, B. & Salzberg, S.L., 2012. Fast gapped-read alignment with Bowtie 2. *Nature methods*, 9(4), pp.357–9. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3322381&tool=pmcentrez&rendertype=abstract [Accessed July 10, 2014].

Lasken, R.S. & McLean, J.S., 2014. Recent advances in genomic DNA sequencing of microbial species from single cells. *Nature Reviews Genetics*, 15(9), pp.577–584. Available at: http://www.nature.com/doifinder/10.1038/nrg3785 [Accessed August 5, 2014].

Lawrence, M.S. et al., 2013. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*, 499(7457), pp.214–8. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3919509&tool=pmcentrez&rendertype=abstract [Accessed July 11, 2014].

Leiserson, M.D. et al., 2015. CoMEt: a statistical approach to identify combinations of mutually exclusive alterations in cancer. *Genome Biology*, 16(1), p.160. Available at: http://genomebiology.com/2015/16/1/160 [Accessed August 9, 2015].

Leiserson, M.D.M., Gramazio, C.C., et al., 2015. MAGI: visualization and collaborative annotation of genomic aberrations. *Nature methods*, 12(6), pp.483–4. Available at: http://www.nature.com/doifinder/10.1038/nmeth.3412 [Accessed May 29, 2015].

Leiserson, M.D.M., Vandin, F., et al., 2015. Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nature genetics*, 47(2), pp.106–14. Available at: http://dx.doi.org/10.1038/ng.3168 [Accessed December 15, 2014].

Leiserson, M.D.M. et al., 2013. Simultaneous identification of multiple driver pathways

in cancer. *PLoS computational biology*, 9(5), p.e1003054. Available at: http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003054#pcbi-1003054-g003 [Accessed March 4, 2015].

Leshchiner, I. et al., 2012. Mutation mapping and identification by whole-genome sequencing. *Genome research*, 22(8), pp.1541–8. Available at: http://genome.cshlp.org/content/22/8/1541.full.html [Accessed December 14, 2015].

Li, H. et al., 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)*, 25(16), pp.2078–9. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2723002&tool=pmcentrez&rendertype=abstract [Accessed July 9, 2014].

Liti, G. & Schacherer, J., 2011. The rise of yeast population genomics. *Comptes rendus biologies*, 334(8-9), pp.612–9. Available at: http://www.sciencedirect.com/science/article/pii/S1631069111001429 [Accessed February 27, 2016].

Mackay, T.F.C., Stone, E. a & Ayroles, J.F., 2009. The genetics of quantitative traits: challenges and prospects. *Nature reviews. Genetics*, 10(8), pp.565–77. Available at: http://www.ncbi.nlm.nih.gov/pubmed/19584810.

De Maeyer, D. et al., 2013. PheNetic: network-based interpretation of unstructured gene lists in E. coli. *Molecular bioSystems*, 9(7), pp.1594–603. Available at: http://www.ncbi.nlm.nih.gov/pubmed/23591551 [Accessed February 21, 2016].

Magwene, P.M., Willis, J.H. & Kelly, J.K., 2011. The statistics of bulk segregant analysis using next generation sequencing. *PLoS computational biology*, 7(11), p.e1002255. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3207950&tool=pmcentrez&rendertype=abstract [Accessed July 13, 2012].

Majerus, M.E.N., 2008. Industrial Melanism in the Peppered Moth, Biston betularia: An Excellent Teaching Example of Darwinian Evolution in Action. *Evolution: Education and Outreach*, 2(1), pp.63–74. Available at: http://www.evolution-outreach.com/content/2/1/ [Accessed March 17, 2016].

McElroy, K.E., Luciani, F. & Thomas, T., 2012. GemSIM: general, error-model based simulator of next-generation sequencing data. *BMC genomics*, 13(1), p.74. Available at: http://www.biomedcentral.com/1471-2164/13/74 [Accessed May

23, 2013].

Meijnen, J.-P. et al., 2016. Polygenic analysis and targeted improvement of the complex trait of high acetic acid tolerance in the yeast Saccharomyces cerevisiae. *Biotechnology for Biofuels*, 9(1), p.5. Available at: http://biotechnologyforbiofuels.biomedcentral.com/articles/10.1186/s13068-015-0421-x [Accessed January 8, 2016].

Mi, H. et al., 2010. PANTHER version 7: improved phylogenetic trees, orthologs and collaboration with the Gene Ontology Consortium. *Nucleic acids research*, 38(Database issue), pp.D204–10. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2808919&tool=pmcentrez&rendertype=abstract [Accessed September 18, 2015].

Ng, S. et al., 2012. PARADIGM-SHIFT predicts the function of mutations in multiple cancers using pathway impact analysis. *Bioinformatics (Oxford, England)*, 28(18), pp.i640–i646. Available at: http://bioinformatics.oxfordjournals.org/cgi/doi/10.1093/bioinformatics/bts402 [Accessed November 5, 2015].

Nielsen, H.B. et al., 2014. Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. *Nature biotechnology*, 32(8). Available at: http://www.ncbi.nlm.nih.gov/pubmed/24997787 [Accessed July 9, 2014].

Novak, J.D. & Cañas, A.J., 2008. The Theory Underlying Concept Maps and How to Construct and Use Them. *Technical Report IHMC CmapTools 2006-01 Rev 01-2008*, pp.1–36.

Pais, T.M. et al., 2013. Comparative polygenic analysis of maximal ethanol accumulation capacity and tolerance to high ethanol levels of cell proliferation in yeast. *PLoS genetics*, 9(6), p.e1003548. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3675000&tool=pmcentrez&rendertype=abstract [Accessed December 22, 2015].

Parts, L. et al., 2011. Revealing the genetic structure of a trait by sequencing a population under selection. *Genome research*, 21(7), pp.1131–8. Available at: http://genome.cshlp.org/content/21/7/1131.long [Accessed December 22, 2015].

Pfennig, D.W. et al., 2010. Phenotypic plasticity's impacts on diversification and

speciation. *Trends in ecology & evolution*, 25(8), pp.459–67. Available at: http://www.ncbi.nlm.nih.gov/pubmed/20557976 [Accessed July 11, 2014].

Prabhakaran, S. et al., 2013. HIV Haplotype Inference Using a Propagating Dirichlet Process Mixture Model. *IEEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM*, 6(1). Available at: http://www.ncbi.nlm.nih.gov/pubmed/24216116.

Prosperi, M. & Salemi, M., 2012. QuRe: software for viral quasispecies reconstruction from next-generation sequencing data. *Bioinformatics*, 21(1), pp.132–133. Available at: http://bioinformatics.oxfordjournals.org/content/28/1/132.short [Accessed October 2, 2014].

Pulido-Tamayo, S. et al., 2015. *SSA.ME Detection of cancer mutual exclusivity patterns by small subnetwork analysis*, Cold Spring Harbor Labs Journals. Available at: http://biorxiv.org/content/early/2015/12/10/034124.abstract [Accessed January 12, 2016].

Quarrie, S.A. et al., 1999. Bulk segregant analysis with molecular markers and its use for improving drought resistance in maize. *Journal of Experimental Botany*, 50(337), pp.1299–1306. Available at: http://jxb.oxfordjournals.org/content/50/337/1299.abstract [Accessed March 15, 2016].

Rappaport, N. et al., 2013. MalaCards: an integrated compendium for diseases and their annotation. *Database : the journal of biological databases and curation*, 2013, p.bat018. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3625956&tool=pmcentrez&rendertype=abstract [Accessed November 5, 2015].

Robinson, J.T. et al., 2011. Integrative genomics viewer. *Nature Biotechnology*, 29(1), pp.24–26. Available at: http://dx.doi.org/10.1038/nbt.1754 [Accessed November 19, 2014].

Rolland, T. et al., 2014. A proteome-scale map of the human interactome network. *Cell*, 159(5), pp.1212–26. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0092867414014226\nhttp://www.ncbi.nlm.nih.gov/pubmed/25416956 [Accessed November 20, 2014].

Rückert, F. et al., 2010. Examination of apoptosis signaling in pancreatic cancer by computational signal transduction analysis. *PloS one*, 5(8), p.e12243. Available

at:
http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2924379&tool=pmcentrez&rendertype=abstract [Accessed October 13, 2015].

Ruderfer, D.M. et al., 2006. Population genomic analysis of outcrossing and recombination in yeast. *Nature genetics*, 38(9), pp.1077–81. Available at: http://www.ncbi.nlm.nih.gov/pubmed/16892060 [Accessed March 15, 2016].

Scheet, P. & Stephens, M., 2006. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *American journal of human genetics*, 78(4), pp.629–44. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1424677&tool=pmcentrez&rendertype=abstract [Accessed December 19, 2015].

Schneeberger, K. et al., 2009. SHOREmap: simultaneous mapping and mutation identification by deep sequencing. *Nature methods*, 6(8), pp.550–1. Available at: http://www.ncbi.nlm.nih.gov/pubmed/19644454 [Accessed March 15, 2016].

Shah, S.P. et al., 2012. The clonal and mutational evolution spectrum of primary triple-negative breast cancers. *Nature*, 486(7403), pp.395–9. Available at: http://dx.doi.org/10.1038/nature10933 [Accessed July 13, 2012].

Shannon, P. et al., 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*, 13(11), pp.2498–504. Available at: http://genome.cshlp.org/content/13/11/2498.full [Accessed July 9, 2014].

Steinmetz, L.M. et al., 2002. Dissecting the architecture of a quantitative trait locus in yeast. *Nature*, 416(6878), pp.326–30. Available at: http://www.ncbi.nlm.nih.gov/pubmed/11907579 [Accessed November 14, 2015].

Stepanauskas, R., 2012. Single cell genomics: an individual look at microbes. *Current opinion in microbiology*, 15(5), pp.613–20. Available at: http://www.ncbi.nlm.nih.gov/pubmed/23026140 [Accessed August 6, 2013].

Swinnen, S., Schaerlaekens, K., et al., 2012. Identification of novel causative genes determining the complex trait of high ethanol tolerance in yeast using pooled-segregant whole-genome sequence analysis. *Genome research*, 22(5), pp.975–84. Available at:

http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3337442&tool=pmcentrez&rendertype=abstract [Accessed October 25, 2015].

Swinnen, S., Thevelein, J.M. & Nevoigt, E., 2012. Genetic mapping of quantitative phenotypic traits in Saccharomyces cerevisiae. *FEMS yeast research*, 12(2), pp.215–27. Available at: http://www.ncbi.nlm.nih.gov/pubmed/22150948 [Accessed March 15, 2016].

Takagi, H. et al., 2015. MutMap accelerates breeding of a salt-tolerant rice cultivar. *Nature biotechnology*, 33(5), pp.445–9. Available at: http://dx.doi.org/10.1038/nbt.3188 [Accessed January 4, 2016].

Töpfer, A. et al., 2013. Probabilistic inference of viral quasispecies subject to recombination. *Journal of computational biology : a journal of computational molecular cell biology*, 20(2), pp.113–23. Available at: http://online.liebertpub.com/doi/abs/10.1089/cmb.2012.0232 [Accessed August 6, 2014].

Töpfer, A. et al., 2014. Viral quasispecies assembly via maximal clique enumeration. *PLoS computational biology*, 10(3), p.e1003515. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3967922&tool=pmcentrez&rendertype=abstract [Accessed October 2, 2014].

Vandin, F., Upfal, E. & Raphael, B.J., 2012. De novo discovery of mutated driver pathways in cancer. *Genome research*, 22(2), pp.375–85. Available at: http://genome.cshlp.org/cgi/doi/10.1101/gr.120477.111 [Accessed July 22, 2014].

Verbeke, L.P.C. et al., 2015. Pathway Relevance Ranking for Tumor Samples through Network-Based Data Integration. *PloS one*, 10(7), p.e0133503. Available at: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0133503 [Accessed March 17, 2016].

Vogelstein, B. et al., 2013. Cancer genome landscapes. *Science (New York, N.Y.)*, 339(6127), pp.1546–58. Available at: http://www.sciencemag.org/cgi/doi/10.1126/science.1235122 [Accessed July 9, 2014].

van Voorst, F. et al., 2006. Genome-wide identification of genes required for growth of Saccharomyces cerevisiae under ethanol stress. *Yeast (Chichester, England)*, 23(5), pp.351–9. Available at: http://www.ncbi.nlm.nih.gov/pubmed/16598687

[Accessed March 15, 2016].

Wenger, J.W., Schwartz, K. & Sherlock, G., 2010. Bulk segregant analysis by high-throughput sequencing reveals a novel xylose utilization gene from Saccharomyces cerevisiae. *PLoS genetics*, 6(5), p.e1000942. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2869308&tool=pmcentrez&rendertype=abstract [Accessed February 29, 2016].

Wu, G., Feng, X. & Stein, L., 2010. A human functional protein interaction network and its application to cancer data analysis. *Genome biology*, 11(5), p.R53. Available at: http://genomebiology.com/2010/11/5/R53 [Accessed November 9, 2015].

Yamamoto, S. et al., 2011. INOH: ontology-based highly structured database of signal transduction pathways. *Database : the journal of biological databases and curation*, 2011, p.bar052. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3225078&tool=pmcentrez&rendertype=abstract [Accessed November 9, 2015].

Yates, L.R. & Campbell, P.J., 2012. Evolution of the cancer genome. *Nature reviews. Genetics*, 13(11), pp.795–806. Available at: http://www.nature.com/doifinder/10.1038/nrg3317 [Accessed January 29, 2013].

Yeang, C.-H., McCormick, F. & Levine, A., 2008. Combinatorial patterns of somatic gene mutations in cancer. *FASEB journal : official publication of the Federation of American Societies for Experimental Biology*, 22(8), pp.2605–22. Available at: http://www.fasebj.org/cgi/doi/10.1096/fj.08-108985.

Yeger-Lotem, E. et al., 2004. Network motifs in integrated cellular networks of transcription-regulation and protein-protein interaction. *Proceedings of the National Academy of Sciences of the United States of America*, 101(16), pp.5934–9. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=395901&tool=pmcentrez&rendertype=abstract [Accessed March 17, 2016].

Yuan, T.L. & Cantley, L.C., 2008. PI3K pathway alterations in cancer: variations on a theme. *Oncogene*, 27(41), pp.5497–510. Available at: http://www.ncbi.nlm.nih.gov/pubmed/18794884.

Zagordi, O. et al., 2011. ShoRAH: estimating the genetic diversity of a mixed sample from next-generation sequencing data. *BMC bioinformatics*, 12(1), p.119.

Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3113935&tool=pmcentrez&rendertype=abstract [Accessed July 12, 2012].