Statistics and Evolution of Functional Genomic Sequence

Inauguraldissertation
zur Erlangung des Doktorgrades der
Mathematisch–Naturwissenschaftlichen Fakultät
der Universität zu Köln

vorgelegt von

Dominic Grün

aus Bergisch Gladbach

Köln 2006

Berichterstatter: Prof. Dr. Michael Lässig
Prof. Dr. Nikolaus Rajewsky

Tag der mündlichen Prüfung: 13. Dezember 2006

Contents

Chapter 1

General Introduction

"I have called this principle, by which each slight variation, if useful, is preserved, by the term Natural Selection."

Charles Darwin from "The Origin of Species".

It was Charles Darwin who established evolution as a *common descent* in order to explain the emergence of new species as a process of accumulation of slight random variations, which were preserved if they proved beneficial for the organism.
Darwin's theory of natural selection, although some 150 years old, is still the basis for the modern understanding of evolution. However, until the late twentieth century it remained much less clear how complexity of phenotypical traits is encoded by the genome. When the human genome was sequenced in the last decade, the small number of protein coding genes, which is in fact not much higher than in organisms as simple as the fruit fly, came as a surprise to many researchers who assumed this number to correlate with the complexity of the species.
This discrepancy was resolved once it was recognized that it is not the protein itself but rather its regulation, which is mainly affected by evolution. In eukaryotes, the *cis*-regulatory region of an average protein coding gene is targeted by a variety of *trans*-acting factors, stimulating or repressing transcription of the gene and thus subsequent protein production. The gene itself could code for another *trans*-acting factor targeting another set of genes. One can easily imagine that evolution acts much more efficiently by modulating and rewiring this intricate gene network than by introducing novel genes. Importantly, not only the sequence turnover in gene regulatory regions, but also the evolution of transcription factors has significant impact on the evolution of the gene regulatory networks. Understanding and deciphering gene regulatory relationships and their evolution can be considered one of the main goals of genomics.

With the recent availability of whole genome sequence data across many different clades of life as well as high-throughput methods to measure genomic data such as gene expression levels on a large scale, sophisticated methods for computational analysis and statistical evaluation of this wealth of data became indispensable for the field of genomics.
The classical tasks of *bioinformatics*, i. e. organization and statistical analysis of genomic data, increasingly merge with the field of *systems biology* that aims at integrating different levels of information to understand how biological systems function. The body of

statistical methods such as those for biological sequence analysis, summarized in Durbin et al. (1998) [39], needs to keep up with the growing complexity of biological data. However, to understand how information is encoded by the genome, the first step is to devise models of genomic interactions, which can be tested statistically as well as experimentally. Exploring whole genome data, for instance when inferring position weight matrices of transcription factor binding sites, one typically deals with a large number of instances of the entity under consideration. It is thus most natural to apply concepts of statistical physics, which are generally suited to analyze large ensembles of interacting particles. In the context of genomic data, functional constraints that induce correlations correspond to the interaction between particles and can be modeled accordingly. For instance, transcription factor binding sites of particular motif compositions are expected to be overrepresented in promoter regions of genes co-expressed with the transcription factor. The validity of specific models can be statistically assessed by a comparison to uncorrelated background models, which can be interpreted as an analog to ensembles of non-interacting particles.

The derivation of refined statistical models allows for the development of improved bioinformatic tools to examine genomic sequences at a higher sensitivity and specificity with regard to the properties of interest. Thus, with the availability of increasing amounts of data, progress in analyzing genomic sequence crucially depends on the existence of sufficiently complex statistical models.

In this thesis, three separate but yet related biological problems are addressed with the mindset of a physicist.

After the recent discovery of *microRNAs* in 1993 [81], the ubiquity and physiological importance of these tiny RNA molecules that regulate protein production by binding to the mRNA of their target genes, was soon revealed for all multi-cellular eukaryotic organisms examined so far. However, detection of genes targeted by specific microRNAs proved rather difficult and only the massive use of cross-species comparison substantially enhanced the reliability of target predictions. In the first chapter of this work, one of the first algorithms for the genome-wide detection of microRNA target genes is introduced. Multiple-sequence-alignments of various species are screened for putative target genes and a probabilistic method assesses the probability of these loci to be true target genes. The identification of target sites relies on a thermodynamical model of binding of the microRNA to the mRNA, which is based on the statistical evaluation of validated instances of microRNA-regulated genes [110].

A detailed analysis of target predictions is presented for three different clades of life, vertebrates, insects, and nematodes, yielding a first glimpse at the evolution of microRNAs and their regulatory relationships.

In the second chapter, evolutionary dynamics of *microsatellites*, an abundant class of repetitive sequence in eukaryotic genomes is addressed. Inspired by the putative functionality of some of these elements and the difficulty of constructing correct sequence alignments that reflect the evolutionary relationships between microsatellites, we propose and validate a neutral model for microsatellite evolution in *Drosophila melanogaster*. By implementing specific mutational processes, the model aims at discriminating the sequence composition of random background and microsatellite sequence. A comparison to evolutionary rates, independently measured from multiple-sequence-alignments of microsatellite loci, serves to validate the model. It adds on previously existing studies by providing an improved

framework for the description of microsatellite evolution and genome-wide application broken down to sequence classes of different functional annotations in *D. melanogaster*.
In the third chapter, we propose a general population genetic model in order to compute the population distribution of functional versus non-functional genotypes at a single transcription factor binding site across the entire range of a limited set of evolutionary parameters. The increased conservation of the binding locus versus neutral background sequence is explained as an equilibrium state of evolutionary forces. More precisely, the selection pressure that acts on a functional sites prevents the sequence from freely diffusing in genotype space due to point mutations and genetic drift.
The model is used to explain the loss of binding site stability as a function of effective population size, such as observed at the evolutionary transition from prokaryotes to eukaryotes, and the compensatory benefit of multiple sites is assessed by an extension of the single-site model.

Chapter 2

Computational identification of MicroRNA target genes

2.1 Biology of microRNAs

2.1.1 Introduction

MicroRNAs are a novel class of small endogenous non-coding RNAs, typically of length 21-23 nucleotides (nt) in their mature form. They suppress protein production by binding to the mRNA of their target genes. Genetic screens of mutant phenotypes in the nematode *C. elegans* led to the discovery of the first microRNA, *lin-4*, in 1993 [81]. *lin-4* was observed to repress the nuclear protein encoded by *lin-14*, whose downregulation is required for the developmental transition between the first two larval stages of the worm embryo [116]. This kind of post-transcriptional regulation was first considered to be exceptional, since *lin-4* was only found in nematodes. After the discovery of the second microRNA, *let-7*, again in *C.elegans* [113], the view of post-transcriptional control imposed by these tiny regulators drastically changed, since *let-7* was observed to be conserved in all metazoans [99], suggesting a fundamental role for this class of non-coding RNAs. In nematodes, expression of *let-7* proved responsible for post-transcriptional down-regulation of the proteins encoded by *lin-41* and *hbl-1* [113, 129, 87]. Similar to the *lin-4*/*lin-14* interaction at the end of the first larval stage, the repression of *lin-41* by *let-7* induces the developmental transition from the third to the fourth larval stage. The apparent fundamental role of microRNAs in nematode development together with the outstanding degree of conservation of *let-7* triggered large scale searches for microRNAs in various organisms and of the order of 100-400 microRNAs per species have already been identified experimentally in *C. elegans*, *D. melanogaster*, *H. sapiens*, *M. musculus*, *D. rerio* and also in plants, many of them being conserved over large evolutionary distances. Recent computational searches for microRNA genes in human indicate that the true number of different microRNAs per species in mammals could exceed 500, including a substantial fraction of species or lineage specific microRNAs [15]. Already these numbers suggest that microRNAs serve as an important layer of post-transcriptional control.
Mature microRNAs are specifically expressed at high numbers in various tissues [8, 11], suggesting substantial regulation in various physiological processes. MicroRNA genes are typically transcribed by RNA polymerase II as large primary transcripts (pri-microRNA)

with CAP-structure and poly(A)-tail and are subject to subsequent processing [9]. They often occur in clusters and clustered microRNAs generally show similar expression profiles [11], suggesting their transcription as polycistrons. Approximately half of all microRNAs reside in introns of protein-coding genes and are presumably co-transcribed with their host genes, although they can also have independent promoters.
The pri-microRNA is processed into a $\sim$70 nt long hairpin-precursor by the nuclear ribonuclease III (RNase III) endonuclease *Drosha* and transported into the cytoplasm by *Exportin5/RanGTP* where another RNase III, *Dicer*, cleaves out a $\sim$21 nt RNA duplex. In most of the cases, only the one strand of this duplex with the smaller 5' end binding energy becomes a mature microRNA and is loaded into the RNA induced silencing complex (RISC), while the remaining strand is subject to rapid degradation. RISCs are ribonucleoprotein complexes of various subtypes reflecting probably different modes of function. As a core component, RISCs always contain one family member of the *Argonaute* proteins. Although the composition and structure of these complexes has been studied in detail, little is known about their biochemical function. However, a recent study revealed how RISCs guide cleavage of mRNA targeted by small interfering RNAs (siRNAs) in the RNAi pathway [88]. Here, the siRNA, incorporated into RISC, binds to a perfectly complementary site in the mRNA of its target gene and induces cleavage and subsequent degradation of the transcript. In this RISC subtype in animals, *Argonaute2* has been identified as the catalytic center inducing the mRNA cleavage. However, the biochemical function of RISC loaded with microRNA differs from this scenario. MicroRNAs guide RISC to target sites of imperfect complementarity and repress translation without cleaving their targets at a particular catalytic sites. Experimental examination of post-transcriptional regulation by *let-7* revealed that microRNAs could repress translation initiation by interfering with the recognition of the m^7G-cap by the protein eIF4E [104]. In this mode, RISC presumably primarily blocks translation without inducing significant degradation of the target gene. The degree of complementarity between the microRNA and its target thus determines whether RISC induces instability of the target gene (e. g. by inducing deadenylation and exonucleolytic degradation) or just blocks translation. While the latter mode is common in animals, target cleavage prevails in plants. A further difference between these two modes is the typical location of target sites. Cleavage sites in plants are located mainly inside of coding regions and target sites of imperfect complementarity typically reside in 3' untranslated regions (UTRs) of mRNAs.
Since the predominant targeting mode in animals appears to be translational repression mediated by imperfect base pairing between the microRNA and the 3'UTR of their target mRNA [98, 138], it was originally thought that the expression of transcripts targeted in this mode remains largely unaffected. However, recent studies indicated regulation of mRNA expression levels also in animals. In *C. elegans*, degradation of transcripts targeted by *lin-4* and *let-7* following the blocking of the translational machinery has been supported by experimental evidence [6]. It was also shown in mammals, that a single microRNA can strongly influence the overall mRNA expression profile of a single cell [86]. In this study, human cells were transfected with *miR-1* and *miR-124*, two microRNAs specifically expressed in muscles and brain, respectively. Transfection with either microRNA shifted the overall transcription profile of the cell towards an expression profile characteristic for the respective tissue. A recent work [71] provided statistical and experimental evidence that

the mammalian liver expressed microRNA *miR-122* could specifically drive the expression level of roughly 500 genes. The 3'UTRs of genes upregulated upon downregulation of *miR-122* are found to be strongly enriched with *miR-122* recognition motifs, whereas downregulated genes are depleted of these motifs.
In order to assess the physiological importance of microRNA regulation, it is crucial to identify their target genes. The small number of experimentally validated targets anticipates microRNA regulation of various processes. As already discussed, the founding members *lin-4* and *let-7* control developmental timing in nematodes [81, 113]. The *let-7* family comprises three additional microRNAs very similar in sequence to *let-7*, but with differing expression patterns. This microRNA family is further implicated in regulation of the *C. elegans* gene *let-60* and its human homolog, the *ras*-oncogene [60]. *let-60* controls vulva development in nematodes and activating mutations of *let-60* lead to a multivulva phenotype that can be suppressed by overexpressing the *let-7* family member *miR-84*.
The *C. elegans* microRNAs *lsy-6* and *miR-273* control left-right asymmetry of chemosensory neurons by repressing *cog-1* and *die-1*, respectively [61, 28].
In *Drosophila*, the microRNA *bantam* was discovered to control cell proliferation. In particular, the pro-apoptotic gene *hid* was identified as a likely direct target of *bantam* [19]. Furthermore, there is experimental *in vivo* evidence for *miR-14* [147] and the *miR-2* family [132] to suppress programmed cell death.
A recent study revealed the expression pattern of *miR-1* in the *Drosophila* embryo and demonstrated, that knockout of *miR-1* leads to lethality of second instar larvae due to severely deformed musculature [130]. Overexpression assays provided experimental evidence for regulation of the *Notch* signaling pathway by microRNAs of the *miR-4* and the *miR-2* family [132] and *miR-7*. These observations together suggest an important involvement of microRNAs in *Drosophila* development. In zebrafish, microRNA regulation appears to be dispensable for early embryonic development, whereas morphogenesis at later stages was severely affected [46, 143]. Injecting synthetic microRNAs into *Dicer* null mutants, depleted of microRNAs, demonstrated a crucial role of *miR-430* in early brain morphogenesis and discarding of maternal transcripts.
In mammals, *miR-375* was shown to control insulin secretion by direct interaction with Myotrophin [105].
Overexpression experiments in mice indicated the regulation of heart organogenesis by *miR-1* [149]. Expression of *miR-1* in the developing heart was found to be driven by the *serum response factor (Srf)*, *myocardin*, *Mef2* and *MyoD* and the transcription factor *Hand2* was identified as a likely target of *miR-1*.
Ample evidence for a connection between microRNAs and cancer has been reported independently by different groups [24, 25, 26, 136, 60, 55, 89, 95]. A common insight of these works is that misexpression of particular microRNAs is responsible for certain types of cancer. Lu et al. (2005) [89] reported that different kinds of tumor can be recognized by the microRNA expression signature of the cancer tissue.
Even viruses seem to produce their own microRNAs [102, 103], opening new ways to escape the anti-viral defense of the host organism or to regulate their own replication [135]. On the other hand, involvement of cellular microRNAs in repression of viral replication has been reported [80].
Although the first targets of microRNAs were discovered by genetic screens, experimental techniques are still too expensive and time-consuming to validate targets of microRNAs

on a larger scale. Hence, computational prediction based on statistical properties of a microRNA target site serves as an important first step to filter out likely candidates of microRNA target genes. A significant fraction of the reported confirmed target genes has been predicted computationally prior to experimental validation. However, current computational target prediction methods still rely on a rather limited amount of information that can be deduced from common features of only a handful of experimentally validated target sites. Even examination of these few examples anticipates different modes of target recognition. Existing target prediction algorithms are thus restricted to particular classes of target sites, or operate with rather high false positive rates.

2.1.2 Characteristics of a typical microRNA target sites

The strategy of choice to computationally predict microRNA genes is based on machine learning. The idea is to train an algorithm with a "training set" of known microRNAs such that it can assign a score to a new candidate to assess the likelihood of being a true microRNA. This technique obviously requires enough instances of validated microRNA genes with somewhat overlapping properties to provide sufficient training data. Due to the large number of known microRNA genes, machine learning based algorithms became established as a prediction method of microRNA genes.
As opposed to microRNAs themselves, prediction of their target genes is complicated by a small training set. Most of the current target prediction algorithms employ an empirical model of target site recognition based on experimental evidence and statistical properties deduced from the training set.

The most prominent feature common to nearly all validated sites is the "binding nucleus" [110] (also addressed as "seed" [82]), a stretch of consecutive base pairings between the microRNA and the 3'UTR of the targeted transcript, in most cases 6-8 nt long and located at the 5' end of the microRNA. The idea of such nuclei has already been suspected after the discovery that *lin-4* partially complementary sites in the 3'UTR of *lin-14* contain a core region of perfect complementarity [144].
After the discovery of numerous vertebrate and invertebrate microRNAs, it was observed that 3'UTRs in these organisms are often enriched with stretches of nucleotides perfectly complementary to 5' segments of microRNAs [76, 145]. A thermodynamical explanation of the binding nucleus was deduced by introducing a scoring model to obtain the maximum discrimination between microRNAs binding to known targets versus random background sequence [110]. This model assigns a score to a stretch of consecutive base pairings between the microRNA and its target site, defined by the sum of single scores for each base pair belonging to the nucleus.
The best nucleus score discrimination between microRNAs binding to known sites versus background sequence yielded single base pair scores corresponding to the relative free energies of AU- and CG-base pairs. Interestingly, G:U wobble base pairs turned out to be strongly disfavored. The nucleus could thus be interpreted as a locus of favorable free energy where a rapid zip-up between the microRNA and the mRNA is guaranteed by perfect complementarity. This quick zip-up helps to overcome thermal diffusion and allows for slower binding of the remaining 3' portion of the microRNA to its target site.
Experimental work by Doench and Sharp [37] yielded deeper insights into the dependence of the repression efficiency of the target gene imposed by the microRNA on the presence

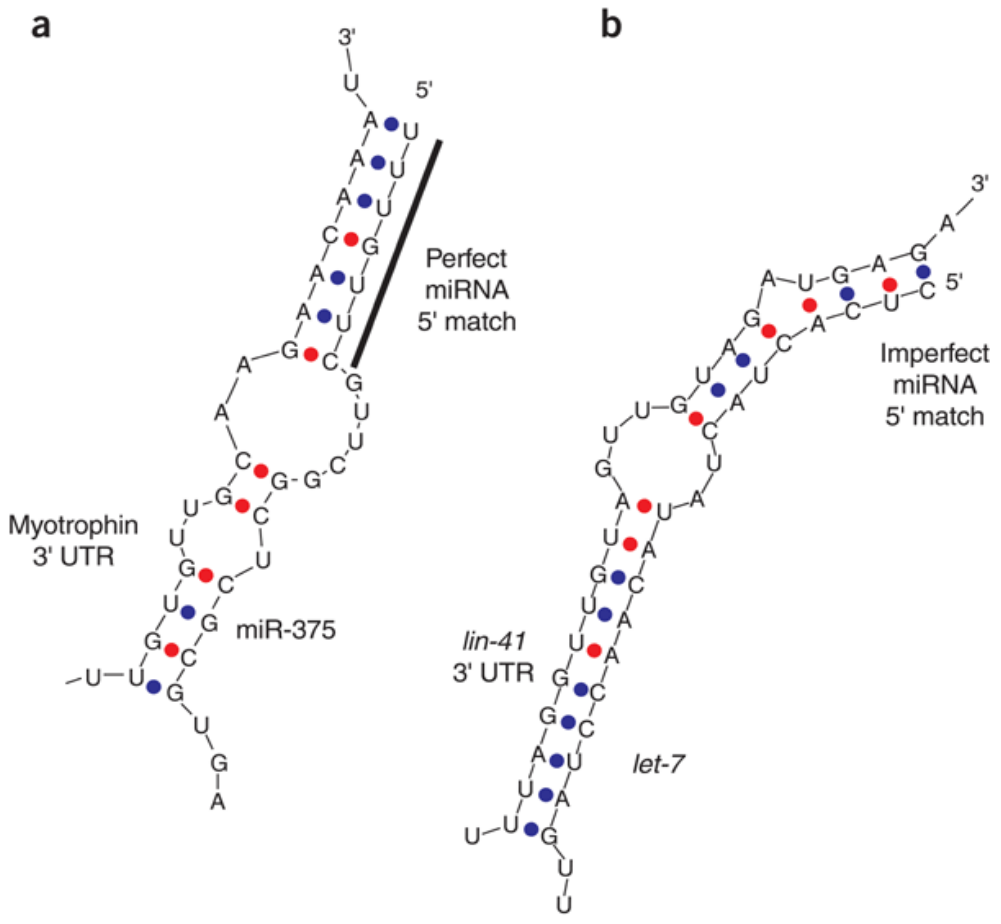


Figure 2.1: **Secondary structure of microRNA:mRNA duplexes**
Examples for secondary structures of a microRNA:mRNA duplex of an experimentally validated target site with a perfect nucleus (a) and an imperfect nucleus (b). In the latter case, extensive compensatory binding of the microRNA 3' end can be recognized, while in the former case larger loops and bulges at the microRNA 3' end are permitted. (Courtesy of N. Rajewsky)

of a minimal nucleus. To test a synergistic effect of multiple sites [36], the authors implanted two target sites subject to single nucleotide mutations flanked by two original sites of a particular microRNA into a reporter mRNA and measured the repression efficacy for each mutant in HeLa cells transfected with the reporter construct. First, only the nucleus region was mutated in order to examine the mutability of the nucleus without significant loss of repression. The results suggest a dependence of the repression level on the free energy of the nucleus section of the microRNA:mRNA duplex, such that single point mutations of the nucleus region are only allowed if the free energy of the nucleus does not increase substantially. It was also shown that G:U wobble base pairs in the nucleus region hinder repression. Mutating the region of the target site predicted to bind the 3' end of the microRNA revealed that for a binding site with 7mer nuclei of perfect complementarity (*perfect nuclei*) single point mutations in this region only mildly

affect repression efficiency. Furthermore, the authors examined in how far target sites are allowed to overlap and observed that for sites with perfect nuclei, repression remains unaffected as long as the nuclei regions of both sites do not overlap.
Along these lines, a recent study distinguished different classes of microRNA target sites based on an *in vivo* assay in the *Drosophila* wing imaginal disc [20]. A single target site was implanted in the 3'UTR of a fluorescent reporter gene expressed in the imaginal disc and the degree of repression was compared for microRNA-expressing and adjacent microRNA-non-expressing cells. The authors tested various types of mutations and came up with the definition of two different classes of target site: 5' dominant sites have sufficient complementarity in the nucleus region and need nearly no support of the microRNA 3' end binding to the mRNA (fig. 2.1a). Target sites with weak nuclei, i. e. nuclei shorter than 7 base pairs or nuclei containing mismatches require compensatory binding of the microRNA 3' end to its target site above the background level. As opposed to 5' dominant sites, 3' compensatory sites form microRNA:mRNA duplexes of free energies lower than expected by matching the microRNA 3' end to randomly chosen background sequence (fig. 2.1b). The contribution of this class of sites leads to a higher specificity, if minimal free energy requirements are incorporated in target prediction algorithms as it is the case for nearly all existing ones.
As a prominent example of 3' compensatory binding, *lin-41* was shown to be bound by *let-7* in nematodes *in vivo* (fig. 2.1b) via two neighboring sites with imperfect nuclei residing in the 3'UTR of this gene [139]. In accordance with former experiments [36], which demonstrated that increasing the number of microRNA or siRNA target sites increases the degree of repression of the target gene, the authors observed that only a single *let-7* binding site is insufficient to mediate repression of the target gene. The synergistic effect of both sites proved necessary to repress translation of *lin-41*. In case of weak 3' compensatory binding sites, the synergistic effect is presumably even more crucial than for multiple sites comprising binding sites with perfect nuclei. However, merely multimerizing single microRNA binding sites does not necessarily suffice to acquire repression. For the two *lin-41* target sites, the authors showed that also the spacer sequence in between these sites plays an important role. Repression is significantly diminished if the spacer segment is mutated or if it is shortened too much. This observation suggests a more complicated architecture of multiple sites and thus anticipates the involvement of interactions between different sites mediated by additional proteins similar to the case of multiple transcription factor binding sites in gene regulatory modules. These details appear to be specific for each instance of multiple sites residing in a single transcript and are thus difficult to include in a general model of target recognition.
Beyond this observation, it is not unlikely that microRNA target sites divide in even more subclasses with target duplexes of each subclass characterized by specific features, e. g. loops of particular sizes. It has been attempted to determine more elaborate characteristics of micorRNA:mRNA duplexes by mutation experiments of a single *let-7* site in human and mouse cell lines [67]. From these experiments, a set of rules was deduced, that was subsequently applied to predict targets of arbitrary microRNAs. However, it remains unclear to what extent specific rules derived for a single particular microRNA hold for other microRNAs or different classes of targets. Moreover, one is at risk to loose many true target sites, which do not obey these rules.
Based on the finding that multiple sites could act synergistically [36], the specificity of

target prediction algorithms can be increased by searching for multiple target sites of a particular microRNA or the co-occurrence of binding sites for various microRNAs in the same transcript [43, 59, 68, 50]. In particular, searching for the co-occurrence of binding sites for co-expressed microRNAs led to the prediction and subsequent validation of the common regulation of the insulin-secretion regulating gene Mtpn by *miR-375*, *miR-124* and *let-7b* in mouse, providing evidence for coordinate microRNA control in mammals [68].

It is still an open question, in how far the folding structure of the mRNA influences the accessability of microRNA target sites. It has been attempted in two studies to incorporate the secondary structure of the potential target into target prediction algorithms [114, 149]. In both studies, it was assumed that the nucleus part of a microRNA binding site residing in the 3'UTR of the target mRNA requires a minimal overlap with single stranded sub-stretches of the folded transcript, e. g. hairpin-loops or free terminal ends, since the pairing probability between the microRNA and double-stranded elements is strongly reduced compared to the pairing probability between the microRNA and such single stranded elements.

In order to incorporate structural features, either the entire 3'UTR of the putative target mRNA [114] or a small segment of sequence surrounding the preselected microRNA binding site [149] was folded. Since mRNA folding software becomes unreliable at a sequence length comparable to the average 3'UTR length of the organisms under consideration it is questionable whether predictions should be based on foldings of whole 3'UTRs. On the other hand, if the mRNA is folded only locally, it is unclear whether this reflects the true local folding structure, since inclusion of surrounding sequence could entirely change the folding of a local element. In both studies, the authors provide validations of some of their target predictions providing supporting evidence for their methods. However, both methods impose additional filters to extract candidates for target genes. In particular, both algorithms require the presence of a 7mer binding nucleus which is known to be the consensus feature for a big class of target sites.

The current amount of information inferred from examination of validated sites is still too small to search the genome of a single species for target sites of microRNAs. For many of the current prediction algorithms the first step consists of searching the 3'UTR of all genes for binding nuclei of microRNAs, i. e. sub-stretches of seven specific nucleotides. Since a given 7mer occurs roughly every 16,000 bases by chance, this counting presumably contains many false positives. By using cross-species comparison, the number of false positives can be significantly reduced. Cross-species comparison assumes enhanced conservation of true target sites due to selective pressure, while random 3'UTR background sequence is assumed to mutate at a higher rate. Randomly chosen 7mers are thus expected to have on average significantly different conservation levels than 7mers complementary to the nucleus region of microRNAs. This expectation was independently confirmed by several studies on a genome-wide scale in vertebrates, fruit flies and nematodes [145, 68, 50, 83, 79].

However, cross-species comparison only allows for target predictions of microRNAs conserved between the species under consideration. Furthermore, species- or lineage-specific target genes even of conserved microRNAs are missed by this technique. Nonetheless, as will be discussed below, this approach yielded deeper insight into microRNA biology and led to an increasing number of target validations.

2.2 PicTar – A probabilistic approach to detect genes coordinately regulated by microRNAs

In this chapter, an algorithm is introduced, which screens the input sequence, usually the 3'UTR of an mRNA, for binding sites of a given search set of microRNAs and assigns a log likelihood score for this sequence to be bound by microRNAs drawn from this search set. In the simplest model, the mere presence of a binding nucleus for a particular microRNA could be sufficient to assume a regulatory relationship. However, it is crucial to take into account the possibility of such binding nuclei to emerge by chance without any functional bearing. In the presence of binding nuclei, the sequence composition of the input sequence can thus be explained by two competing models, a neutral background model versus a functional model. We took advantage of evolutionary conservation in order to assign the probability of a binding nucleus to be functional. Background sequence, on the other hand, is assumed to be emitted with uncorrelated single nuncleotide probabilities. A maximum likelihood procedure is then used to determine the probability of the functional model to explain the occurrence of the binding nucleus in the input sequence. If the resulting score is low, the binding nucleus is likely to have emerged by chance and the probability of the gene to be regulated by the corresponding microRNA is low.
To assess the significance of a putatively functional site, it is thus crucial to allow for a competing neutral model. The competition of overlapping binding nuclei for different microRNAs is treated analogously.
Hence, a bioinformatic tool for the prediction of microRNA target genes, that goes beyond a simple pattern search, should ideally reflect the competition of all conceivable options to explain the observed sequence.
To infer the parameters of the model, the statistical evaluation of a large number of putative instances of functional and non-functional sequence is needed. The latter seem to be reasonably well characterized by single letter probabilities. The specificity of functional sequence, i. e. binding nuclei, was first derived from validated instances [110] and model parameters were finally inferred using the average level of evolutionary conservation of binding nuclei in the ensemble of all 3'UTR sequences.

2.2.1 General logic of PicTar

Expression profiling of microRNAs by cloning and sequencing, Northern blotting, or microarrays has shown that oftentimes a small number of microRNAs (typically 1-10) is expressed in specific tissues and developmental stages [8]. Many known target genes of microRNAs contain several microRNA binding sites, and experiments have shown that the degree of translational repression can be exponential in the number of microRNA binding sites in the 3'UTR [36]. Thus, similar to transcriptional regulation, the concentrations of the trans-acting microRNAs in a cell may be read out by *cis*-regulatory sites and used to fine-tune gene expression [57]. Hence, to understand biological microRNA function, it may be important to search for combinations of microRNA binding sites for sets of co-expressed microRNAs. Previously developed computational algorithms can identify targets for single microRNAs, but it is unclear how to use them to score common targets of several microRNAs. Furthermore, they typically have significantly enhanced false positive rates when the number of binding sites for a given microRNA in a 3'UTR is small

[2]. **PicTar** (**P**robabilistic **I**dentification of **C**ombinations of **T**arget sites) overcomes these problems by generalizing previous methods and allows the identification of targets for both single microRNAs and combinations of microRNAs.

Input to PicTar (fig. 2.2a) is a fixed search set of microRNAs and multiple alignments of orthologous nucleotide sequences (3'UTRs). Output are scores that rank genes by their likelihood to be a common target of members (subsets) of the search set, and probabilities for the predicted binding sites in each UTR. The algorithm follows the general logic of Ahab, a validated probabilistic algorithm for the identification of combinations of transcription factor binding sites [109, 122]. PicTar tallies all segmentations of a sequence into binding sites and background, and computes the maximum likelihood score that the sequence is bound by combinations of microRNAs (fig. 2.2b). In this probabilistic model, microRNAs compete with each other and background for binding. The model accounts for synergistic effects of multiple binding sites of one microRNA, or several microRNAs acting together, as well as for the appropriate scoring of overlapping sites. Cross-species comparisons are crucial for filtering out false positives: candidate target genes are defined as UTRs with a minimal (user-defined) number of evolutionarily conserved putative binding sites. PicTar then scores the candidate sequences for each species separately. The resulting scores are combined to obtain the final PicTar score for a gene.

2.2.2 Detailed description of the PicTar algorithm

Definition of nuclei

It has been shown that a key role in both target site recognition and repression of the target transcript is played by the nucleus, typically a perfectly Watson-Crick base paired stretch of $\sim$7 nt in the microRNA:mRNA duplex. The nucleus is usually located in the 5' end of the microRNA starting at the first or second position of the microRNA 5' end [2]. According to the already discussed insights gained from the experimental examination of validated target sites as well as on statistical evidence we define a "perfect nucleus" as a perfectly W-C base paired stretch of 7 nucleotides either starting at the first, or the second base of the microRNA (counted from 5'). A single insertion or mutation, unless it leads to a G:U wobble base pair, in the mRNA sequence of a perfect nucleus is allowed as long as compensatory binding of the microRNA 3' end is observed, reducing the overall free energy. These mutated nuclei are called "imperfect nuclei". Hence, for imperfect nuclei we require that the free energy of the entire microRNA:mRNA duplex, calculated with the standard folding software RNAhybrid [112], is below a cutoff which is set to a restrictive value of 60% of the optimal free energy of the entire mature microRNA binding to a perfectly complementary target site.

Definition of anchor sites and identification of target candidates

We first precompute the positions of all possible microRNA nuclei in all UTR sequences. We check if nuclei for the same microRNA fall into overlapping alignment positions for all species under consideration. To enhance the sensitivity, the conservation requirement is alleviated: Whenever a group of species with similar evolutionary distance to the reference

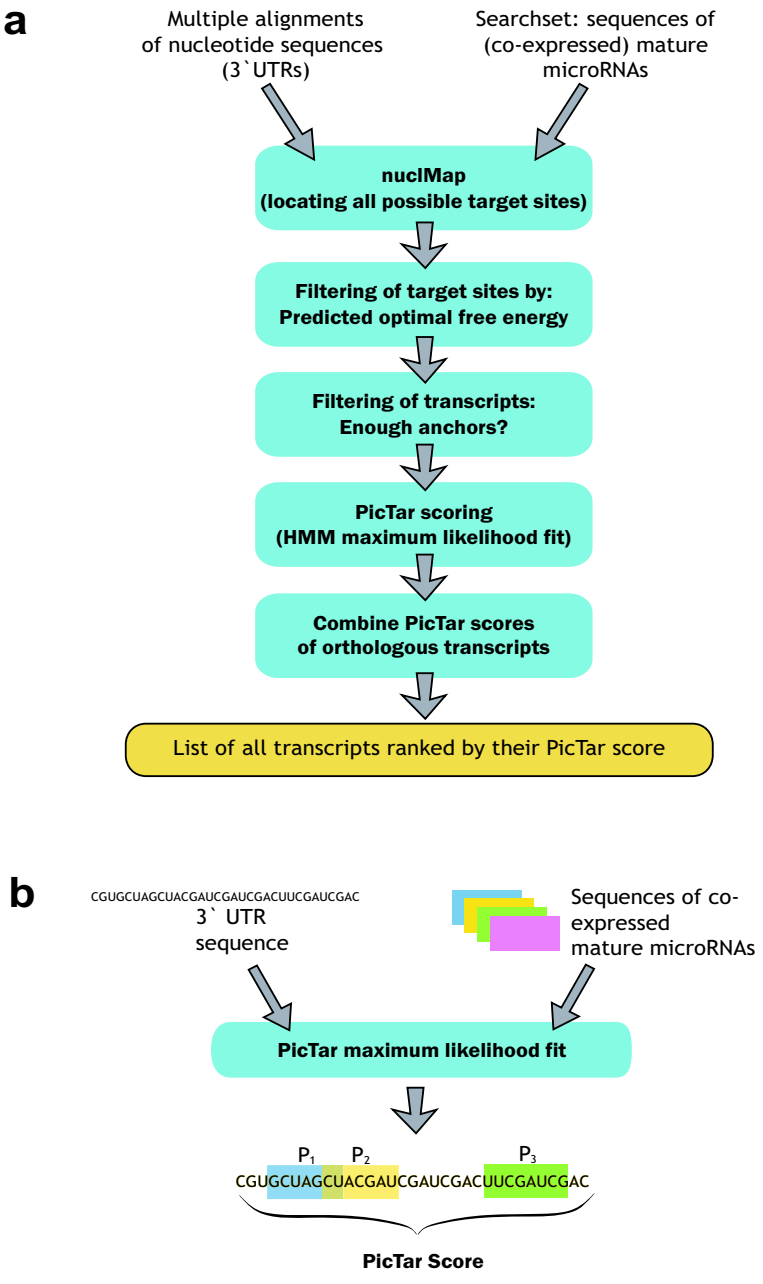


Figure 2.2: **Flowchart of the PicTar algorithm.**
See text for details.

species is present in the set of input sequences, but the whole 3'UTR sequence of a particular species is missing, each target site found in the sequences of the remaining species is considered to be conserved in the whole group. In case of vertebrates, for

instance primates (human, chimp) and rodents (mouse, rat) are grouped independently. This strategy reduces the influence of incomplete alignments, e.g. due to sequencing gaps. If nuclei are conserved by these criteria, we check if the optimal free energy of their predicted microRNA:mRNA duplexes passes our filtering criteria (see above). Nuclei that survive these steps are termed anchors. The number of anchors in a UTR is a parameter which determines if a transcript will be scored by PicTar. If so, the optimal free energy of all sites with imperfect nuclei in each UTR sequence is used to filter out improbable target sites. The remaining sites for each UTR are input to the core hidden Markov model (HMM) of PicTar to compute a score for each UTR in the multiple alignment.

Scoring target candidates by an HMM

The core algorithm of our target identification procedure scores the statistical likelihood for each 3'UTR sequence to be targeted for repression by a given set of microRNAs $\mu_i, i = 1, ..., M$.
Hidden Markov models can be thought of as sequence-generating models (see Appendix A). A random sequence of states is generated by a set of independent "drawings", governed by the transition probabilities between the states. At each state, a certain nucleotide is chosen ("emitted") according to emission probabilities of this particular state. This way, any sequence can be generated with a certain probability. Given the particular sequence x, one can calculate the probability $P(x)$ that x was generated by the specific model (defined by the transition and emission probabilities). Another important task can be to determine the parameters of the model that will maximize the probability that x was generated (this procedure is usually referred to as Expectation Maximization, see Appendix A).
We modeled the 3'UTR of a gene as generated by an HMM whose states are binding sites of each of the microRNAs from the set and the background. The microRNA binding sites are represented by the 7-8 nucleotides long nuclei. Different microRNAs, or, more precisely, their nuclei compete with each other and background for binding (it was experimentally shown [37] that two overlapping nuclei can not act cooperatively). The final score of a sequence is the log ratio of the probability of the sequence being generated by this model, versus the probability that it was generated by background process alone, after the optimal parameters were found using the Expectation Maximization procedure. The sequence is generated in the following way: at each step one of the states of the system is chosen with prior probability (the equivalent of transition probabilities) ρ_i ($i = 0, ..., M$, where ρ_0 is the prior probability for the background, and M is the total number of different microRNAs whose combinatorial effect we are assessing). Depending on the nature of the state, a certain sequence will be emitted: in the background state, one nucleotide will be emitted, and in microRNA target site state the 7-mer or 8-mer will be emitted. The emitted sequence is appended to the sequence previously generated, and the process proceeds to the next step until the required 3'UTR length is reached. The sequence of states chosen from positions 1 to L is called the "parse". Obviously, the same sequence of nucleotides can be created by different parses. Background is modeled with the Markov model of order 0. A short 3'UTR ($<$ 300 nt) cannot be used to reliably estimate its own background nucleotide frequencies. In these cases we take the linear combination of the background nucleotide frequencies estimated from the UTR

and background frequencies estimated from all UTRs for the same species in our data set.
Given the set θ of microRNAs μ_i and their prior probabilities of binding ρ_i, $i = 1, ..., M$, the probability for the particular sequence to be generated is the sum of probabilities of all possible "parses" of this sequence:

$$P(x \mid \theta) = \sum_{\pi} P(x \mid \theta, \pi) \tag{2.1}$$

Here the probability of a particular parse π, $P(x \mid \theta, \pi)$, is simply a product of transition probabilities for each of the states in this particular parse and the emission probabilities of subsequences present in those states ($N =$ total number of states in the parse):

$$P(x \mid \theta, \pi) = \prod_{i=1}^{N(T)} \rho_i \cdot m_i(s) \tag{2.2}$$

The emission probabilities $m_i(s)$ for the microRNA binding sites (i.e. the probabilities that a certain mRNA subsequence will be the microRNA binding site) were modeled to best discriminate between true and spurious, i. e. non-functional nuclei. The idea is based on the increased evolutionary conservation of functional sites. For each possible nucleus, we recorded the number of occurrences in anchor sites conserved in all species of the search set for each species separately. The ratio of these numbers for the species most distantly related to the reference species and the total number of occurrences (conserved and non-conserved instances) of the same nucleus in the reference species itself defines a "conservation score" that reflects the degree of conservation for each particular nucleus. We defined the emission probability of a nucleus exactly by this conservation score.
The identification of the emission probability with the conservation score was introduced only in a second advanced version of the algorithm. The original definition was chosen according to experimental insights. In this version, imperfect nuclei were only accepted, if their free energy was lower or equal than the free energy of the corresponding perfect nucleus duplex. Moreover, the free energy filtering of the whole microRNA:mRNA duplex was somewhat different: duplexes with perfect nuclei were required to have at least 33% of the free energy of the entire microRNA hybridized to a perfectly complementary sequence while this threshold was set to 66% for imperfect nuclei. For perfect nulcei this was a very mild filtering with substantially higher impact in case of imperfect nuclei. A perfect nucleus that survived the filtering was assigned a probability p to be a binding site for the microRNA. The probability for imperfect nuclei was (1-p) divided by the total number of imperfect nuclei (typically in the range of 2-20). We worked with a high p (p $\sim$0.8) since most of the known target sites do not have imperfect nuclei. Another important difference of the original version is, that imperfect nuclei were not accepted in anchor sites.
In the subsequent discussion of target predictions for vertebrates and flies we refer in part to results of the original version [68, 50], since not the entire analysis was redone with the new version [79].
The set of optimal prior probabilities is determined by maximizing the probability $P(x)$ to observe the sequence we are scoring. Since the prior probabilities in principle represent the probability that a given microRNA is present in the cell (and sequence emission

probabilities in the target site state represent the probability that microRNA will actually bind to this particular subsequence), we can say that by finding the priors that maximize the likelihood to observe our sequence, we are actually finding the relative effective concentrations of different microRNAs in the cell that will maximize the probability that this gene is translationally repressed. The optimization is performed using the Baum-Welch algorithm for Expectation Maximization, as described in [127] and [39] (see Appendix A). This procedure is performed because we in principle do not have sufficient information about concentrations of all messenger RNA and microRNA molecules present in the cell. One can imagine, however, that this will change in the future (e.g. availability of more microarray data), and it can be in a natural way incorporated into the model.
Since the sequence is modeled by the HMM, consecutive steps while generating it are independent, which means that one can use the forward equation to calculate $P(x \mid \theta)$ in time proportional to the length of the sequence. Posterior decoding (see Appendix A) is used to calculate the probability for each of the positions in the sequence to be bound by a microRNA.
The advantage of the HMM treatment of the microRNA target prediction problem is that it naturally captures many characteristics of microRNA:mRNA interaction: coordinate action of several microRNAs, amplification of repression efficacy in case of multiple binding sites, competition of microRNAs in case of overlapping binding sites etc. In case of multiple binding sites (either from one or from several microRNAs) the PicTar score will reflect synergistic action: it will be higher than a simple sum of scores for single sites. The length of the sequence influences the score as well: the longer UTR containing n binding sites is statistically less significant than the shorter one with the same number of sites. Beside the features of translational repression that we described in connection to our model, different factors in the structure of microRNA binding sites have begun to emerge that can affect the repression efficacy [139, 67]. In addition, there is a possibility that the process can be affected by correlations in the positions of the target sites relative to each other [37, 139]. It was already shown in case of the transcriptional regulation model described in [127] that these correlations can be incorporated in the HMM.
To obtain a final cross-species score, which reflects the probability that the UTR is regulated by the given set of microRNAs, we averaged the scores for all species that were used to define anchor sites. This average should reflect the different evolutionary distances between species. In vertebrates, for instance, we averaged the human and chimpanzee score, and, independently, the mouse and rat score, to obtain a primate score and a rodent score. These scores were then averaged with the dog score to obtain a score reflecting conservation in all mammals. Similarly, we averaged this mammalian score, the chicken score, and the averaged score of the zebra- and the pufferfish.

2.3 PicTar target predictions for different clades of life

We applied the PicTar method to compute genomewide target predictions in three different clades of life: vertebrates, insects and nematodes. All target predictions are made available online (*http://pictar.bio.nyu.edu*) and are linked to various databases, providing

information on the microRNA itself and the target gene. The output can be viewed as a ranked list of target genes for each particular microRNA linked with a detailed representation of the alignment and information on each particular target site. Moreover, our most recent target predictions [79] can be accessed at and downloaded from the widely used UCSC genome browser (*http://genome.ucsc.edu/*).

2.3.1 Vertebrates

Vertebrate 3'UTR sequences and alignments

We extracted genome-wide multiple alignments of eight vertebrates from the UCSC Genome Browser (*http://www.genome.ucsc.edu/*). These alignments were built from the following genome assemblies: human May 2004 (hg17), chimp Nov. 2003 (panTro1), mouse May 2004 (mm5), rat Jun. 2003 (rn3), dog Jul. 2004 (canFam1), chicken Feb. 2004 (galGal2), fugu Aug. 2002 (fr1), and zebrafish Nov. 2003 (danRer1). We used the UCSC mappings of the human RefSeq mRNA data [106] (Release 6, July 5, 2004) to the human genome to define multiple alignments of 3'UTRs. These alignments cover 19,971 sequences for human/chimp, 19,289 for human/chimp/mouse, 18,717 for human/chimp/mouse/rat, 18,567 for human up to dog, 11,190 up to chicken, 6,136 up to fugu and 4,355 up to zebra fish. The multiple alignments cover human/chimp/mouse/rat/dog for 90% of all human 3'UTR sequence nucleotides. Sequences of all eight species are aligned for 21% of all human 3'UTRs. The coverage for human/chimp/mouse/rat/dog/chicken (55%) is consistent with the estimated number of orthologous human/chicken genes. For generating our statistics, we produced a final dataset of 3'UTRs by restricting human 3'UTR sequences to unique sequences.
Comparing human/mouse sequence pairings of the alignments to pairings independently defined via a gene orthology table yielded a low error rate of $\sim$3%.

Data sets of known and randomized mature microRNA sequences

We downloaded mature microRNA sequences from Rfam [48] (Release 5.0) and added nine microRNAs [105]. We extracted a subset of microRNAs conserved between human/chimp/mouse/rat/dog/chicken using Rfam annotations of mature vertebrate microRNAs homologous to a human microRNA. Whenever no annotation was available, we used stringent criteria to check conservation of the precursor and for the mature microRNA. We constructed a set of unique microRNAs by lumping together microRNAs with identical bases at positions 1 through 7 or 2 through 8 (starting at the 5' end). We obtained 58 unique microRNAs that are conserved up to chicken. Similar to Lewis et al. 2003 [82], we recruited cohorts of unique randomized microRNAs by extracting 8mers with approximately the same abundance ($\pm$15%) of the 7mer starting at position 1 and 2 and the corresponding 7mer of the respective real microRNA in all human 3'UTRs. Experimenting with numerous other randomization schemes led to comparable signal-to-noise ratios. We attached the 3' end of each microRNA to the corresponding random 8mer.

Single microRNA target predictions

The set of 3'UTR alignments served as input sequence. We compute targets at two different levels of conservation. The mammalian predictions require conservation of anchor sites between human, chimp, mouse, rat and dog. The vertebrate predictions require a nucleus to be additionally conserved in chicken. Both human and chimp as well as mouse and rat are considered to reside on a similar evolutionary level when compared to all species included in the analysis. Hence, for both pairs of species a nucleus in just one species was considered sufficient if the sequence of the entire 3'UTR is missing in the other species. Target predictions were computed for 164 microRNAs conserved in mammals and 125 microRNAs additionally conserved in chicken.
To compute the signal-to-noise ratio we restrict the predictions only to the 58 microRNAs with unique nuclei conserved in mammals and in chicken and generated 3 cohorts for a subset of 23 out of these 58 unique microRNAs. To reduce computation time we used only half of all 3'UTRs for target predictions of these randomized microRNAs. For

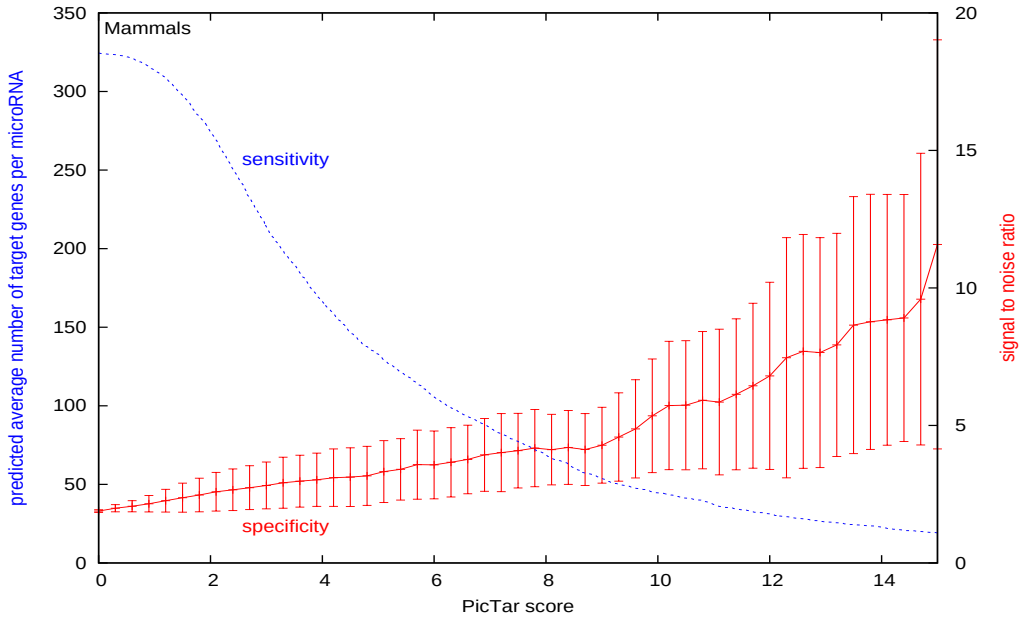


Figure 2.3: **Specificity and sensitivity of target predictions in mammals** signal-to-noise ratio (specificity, solid line) and average number of predicted target genes (sensitivity, dashed line) as a function of a PicTar score cutoff for the modified PicTar algorithm in mammals, requiring conservation in human, chimp, mouse, rat and dog. The signal-to-noise ratio was averaged over three cohorts of randomized microRNAs in mammals. The standard deviation is indicated by error-bars.

mammals, we predict on average 320 target genes per microRNA at a signal-to-noise ratio of 1.9. Dependence of the specificity and the sensitivity on the PicTar score of the

target gene is shown in fig. 2.3. For high score cutoffs, the big error bars are due to the small numbers of predicted targets for randomized microRNAs. At a PicTar score cut off of 10 we still predict on average 50 target genes per microRNA at a signal-to-noise ratio of 5. Requiring conserved anchor sites in mammals and chicken (vertebrates) yields on average 130 target genes per microRNA at a signal-to-noise ratio of 3. At a score cutoff of 2 we still predict on average 60 targets per microRNA at a signal-to-noise ratio of 8. With a score cutoff ranging from 0 to 3, the signal-to-noise ratio increases linearly from 3 to 10, while the sensitivity decreases from 130 to 30 predicted target genes per microRNA.
The contribution of imperfect nuclei to conserved anchor sites is found to be very minor. Among the target predictions for mammals, imperfect nuclei make up for 1.5% of all nuclei in anchor sites while this fraction raises to 2% when requiring conservation of anchor sites also in chicken.
We predict that 44% (16%) of all 17,450 unique genes have at least one anchor site conserved in mammals (vertebrates).
A comparison of the estimated signal-to-noise ratios for mammals and vertebrates demonstrates the power of cross species comparison to filter out false positives. The more distantly relates species are included into the analysis the higher becomes the signal-to-noise ratio already without applying any score cutoff.
The strongly enhanced specificity at the expense of lower sensitivity with increasing score cutoff supports the hypothesis that the conservation score is in fact biologically meaningful and provides a good means of discriminating between true and spurious target sites.

Vertebrate microRNAs are unlikely to target promoter sequence

It has been proposed that microRNAs can target genomic DNA in animals and induce transcriptional silencing of genes via chromatin modification [9]. To test this hypothesis, we ran PicTar on genome-wide sets of non-transcribed upstream sequences.
Analogously to the construction of vertebrate 3'UTR sequences, we used UCSC mappings of human RefSeq mRNA sequences to define 500 base pairs upstream of transcription start sites (termed promoters). To exclude as best possible overlaps with 3'UTRs, we did not include sequences, which were overlapping with any transcript, arriving at a total of 17,883 human sequences. However, in a number of cases our promoters will overlap with 5'UTRs since transcription start sites are often not known. Multiple alignments for promoters across vertebrates were constructed as described above.
In striking contrast to our results for 3'UTRs, we found neither correlation of binding site positions with evolutionary conservation nor significant differences in the conservation of putative target sites for real or randomized microRNAs. Our data suggest that most animal microRNAs either recognize targets in genomic DNA by mechanisms not captured by our algorithm, target sequences other than proximal upstream sequences, or do not target genomic DNA to a significant extent.

Multiplicity of sites enhances specificity

We recorded signal-to-noise ratios for the number of transcripts with at least N conserved anchors for each microRNA separately for random and real microRNAs (fig. 2.4a). Clearly, the multiplicity of sites in a UTR, which is scored by PicTar, also leads to a significant increase of signal-to-noise [82]. To provide a crude estimate of the number

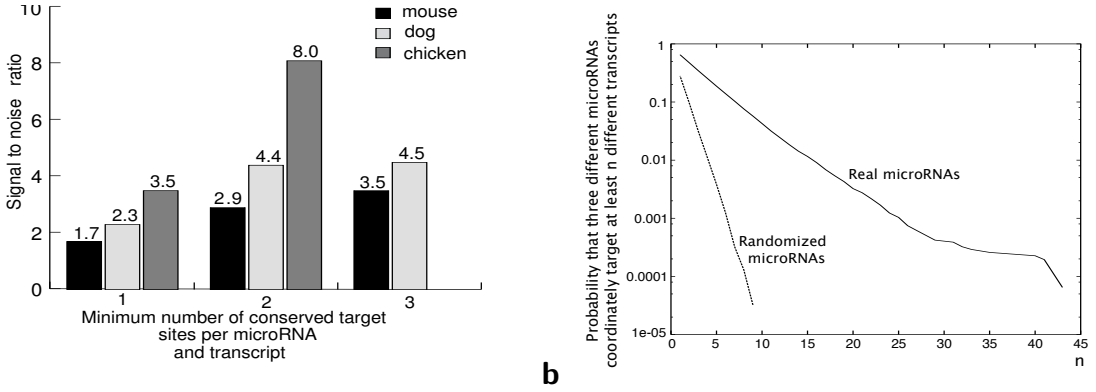


Figure 2.4: **Statistics of predicted targets of multiple microRNAs**
(a) Multiplicity of target sites boosts the signal-to-noise ratio. The ratio of the number of transcripts with at least N anchor sites per microRNA for real versus random microRNAs provides an estimate of the signal-to-noise ratio for sites conserved in human sequences up to mouse (black bars), up to dog (light grey bars), and up to chicken (dark grey bars). The multiplicity of target sites (which is scored by PicTar) clearly helps to raise the signal-to-noise ratio. The plotted results were obtained with the original version of PicTar [68], but differ only slightly from the same numbers computed with the current version of PicTar.
(b) Estimate of the number of coordinately regulated targets for sets of three microRNAs. The probability P(n) that a set of three microRNAs hits at least n transcripts is plotted on a log scale for real (upper curve) and random (lower curve) microRNAs as a function of n. P(n) drops off exponentially and much more steeply for random microRNAs, demonstrating that PicTar runs with random microRNAs will typically yield a vastly reduced number of predictions. P(25) is $\sim$0.001 for real microRNAs, thereby indicating that only $\sim$30 out of $\sim$30,000 possible sets of microRNA triples (sampled from our set of 58 microRNAs) are likely candidates to coordinately regulate at least 25 different transcripts each.

of microRNAs that may coordinately regulate target genes, we counted how many sets of three microRNAs have anchor sites in common transcripts. We plot the probability that a fixed triplet of microRNAs has an anchor site for each microRNA in at least n different transcripts (fig. 2.4b). The criterion for a "hit" was the presence of at least one anchor site, conserved between human/chimp/mouse/rat/dog, for each microRNA in the triplet. The probability of obtaining not a single hit for a triple of real microRNAs is 1-P(1) = 0.35. This probability decays exponentially with n. Roughly two thirds of all possible microRNA triplets could coordinately regulate transcripts. The probability that a fixed set of three microRNAs hits more than 10 (25) targets is 0.04 (0.001), respectively. Roughly 600 triplets out of all triplets drawn from the 58 microRNAs could coordinately

regulate more than 10 targets each. We also computed the same statistics for randomized microRNA sequences (fig. 2.4b), which demonstrated that running PicTar with sets of randomized microRNAs will result in a vastly reduced number of predicted targets.
Along with the genomewide single microRNA target predictions, we also computed predictions for genes coordinately targeted by microRNAs that are co-expressed in certain tissues. We assembled these sets of unique co-expressed microRNAs, conserved up to dog, using published microRNA expression data [8] for brain, thymus, testes and placenta. For each tissue, we required at least two anchor sites of different microRNAs to be conserved between mammals.

Experimental validation of PicTar prediction

We hypothesized that the three most highly expressed microRNAs in the murine pancreatic cell line MIN621, *miR-124* (11.5% of the total microRNA profile), *miR-375* (6.5%), and *let-7b* (6.5%), may act together on a target gene. We used them as input for PicTar, requiring at least one anchor site to be conserved between human/chimp/mouse/rat/dog for each microRNA. We examined the results for *Mtpn*, a known target of *miR-375* [105]. When searching ∼18,500 3'UTR alignments with *miR-375* only, Mtpn was ranked 102, with one predicted binding site with a nucleus 3135 nt downstream of the stop codon and a probability of 0.74. Functionality of this site was validated previously [105]. Searching for *miR-124* targets only puts *Mtpn* at rank 727. However, using both *miR-124* and *miR-375* as the search set boosted the rank to 14. Finally, searching for targets of *miR-124*, *miR-375*, and *let-7b* resulted in rank 4 for *Mtpn*. These ranks refer to the original version of PicTar [68] and have undergone small changes in the current version.
To experimentally validate these predictions, we tested Mtpn regulation by *miR-124* and *let-7b* using two methods. First, our collaborators from the Stoffel lab at Rockefeller University transfected neuroblastoma N2A cells with siRNA duplexes that are homologous in sequence to *miR-124* (si-124) and to *let-7b* (si-let-7b), respectively. We observed in both cases a decrease in endogenous Mtpn expression by Western blotting (fig. 2.5). Second, to test if Mtpn is a target of *miR-124* or *let-7b*, we subcloned the Mtpn 3'UTR downstream of a luciferase reporter gene. Co-transfection of this reporter construct with si-124 or si-let-7b significantly decreased luciferase activity compared to a control siRNA targeting eGFP (si-GFP) (fig. 2.5). The down-regulation by si-124 and si-let-7b was similar to that of si-375. Furthermore, co-transfection of the luciferase reporter with a pool of si-124, si-let-7b, and si-375 resulted in normalized luciferase activity that was significantly less than the activity in any of the other co-transfections, suggesting that Mtpn is regulated by the coordinate action of all three microRNAs. Together, our results provide evidence for a direct and microRNA concentration-dependent regulation of Mtpn by *miR-375*, *miR-124* and *let-7b* and thus establish that Mtpn expression is coordinately regulated by three highly expressed pancreatic microRNAs.
In addition, we validated 7 out of 13 predicted targets of *miR-124* and *miR-375* by either western blotting or luciferase reporter assays, consistent with our false positive estimates (Table 2.1).

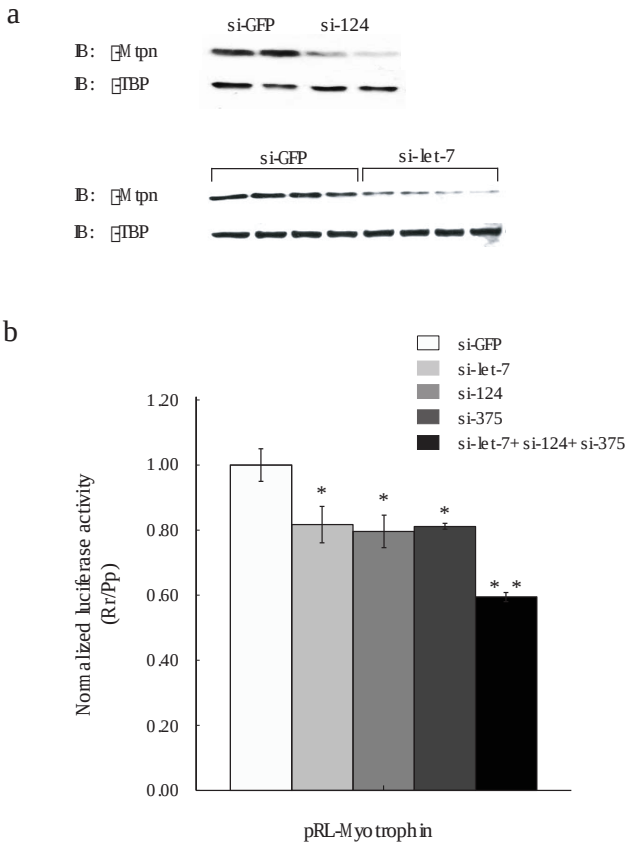


Figure 2.5: **Regulation of Myotrophin by *miR-375*, *miR-124* and *let-7b* (Experiments were done by the Stoffel lab at Rockefeller University)**
a) Immunoblotting. N2A cells were transiently transfected with siRNAs designed against eGFP (si-GFP), *miR-124* (si-124), or *let-7b* (si-let-7b) and lysed after 48h. *Mtpn* expression was assessed following SDS-PAGE, and immunoblotting with anti-*Mtpn* antibodies. Tbp (TATA-binding protein) expression (α-Tbp) was analyzed for a loading control.
b) Dual luciferase assay of transfected N2A cells. A *Renilla reniformis luciferase* (*Rr*-luc) construct containing the full length 3'UTR of *Mtpn* was transiently transfected with either si-GFP, si-124, si-let-7b, or si-375, or all three si-RNAs for 48h and lysed. A *Photinus pyralis luciferase* (*Pp*-luc) served as an internal transfection control. The ratios of *Rr*-luc to *Pp*-luc expression were normalized to the si-GFP transfections. Error bars represent the standard error (S.E.) from three independent experiments. Asterisks (*) and (**) indicate significance levels at a p-value of $\leq$0.05 and $\leq$0.01, respectively.

Conclusions

In summary, we have developed a computational approach that not only can successfully identify microRNA target genes for single microRNAs but also targets, which are likely to be regulated by microRNAs that are co-expressed or act in a common pathway. We have shown that massive sequence comparisons utilizing previously unavailable genome-wide

microRNA	RefSeq gene ID	Gene functional annotation	blot	luc
miR-375	NM_013464.2	aryl-hydrocarbon receptor (Ahr)	-	n.d.
miR-375	NM_010847.1	Max interacting protein 1 (Mxi1)	-	n.d.
miR-375	NM_016889.1	insulinoma-associated 1 (Insm1)	-	n.d.
miR-375	**NM_008413.1**	Janus kinase 2 (Jak2)	n.d.	+
miR-375	**NM_007573.1**	compl. comp. 1, q subcomp. binding protein (C1qbp)	n.d.	+
miR-375	**NM_146144.1**	ubiquitin specific protease 1 (Usp1)	n.d.	+
miR-375	**NM_008098.2**	myotrophin (Mtpn)	+	+
miR-375	**NM_197985**	adiponectin receptor 2 (Adipor2)	+	+
miR-124	NM_009498.3	vesicle-associated membrane protein 3 (Vamp3)	-	n.d
miR-124	NM_013464.2	aryl-hydrocarbon receptor (Ahr)	-	n.d.
miR-124	**NM_011951.1**	mitogen activated protein kinase 14 (Mapk14)	+	n.d.
miR-124	**NM_008098.2**	myotrophin (Mtpn)	+	+
miR-124	NM_197985	adiponectin receptor 2 (Adipor2)	-	-

Table 2.1: **Experimentally validated PicTar predictions**
Target validation for *miR-124* and *miR-375* by immunoblotting and/or luciferase reporter assays. The genes tested were selected from the single microRNA target prediction lists requiring different levels of conservation (human/chimp/mouse or human/chimp/mouse/rat/dog). Genes indicated in bold were validated by either immunoblotting (blot) and/or luciferase (luc) reporter assay ("+"). "-" indicates no detectable decrease of endogenous target protein levels or luciferase reporter activity following microRNA overexpression. "n.d.": not determined. Predictions to be tested were not selected by their PicTar score. The average score of the predicted *miR-375* (*miR-124*) targets is rather low. Therefore our number of false positives is comparable to the predicted signal-to-noise ratio. RefSeq identifiers refer to Mus musculus.

alignments across eight vertebrate species strongly decreased the false positive rates of microRNA target predictions, allowing PicTar to predict (above noise) on average roughly 200 targeted transcripts per microRNA. PicTar's combinatorial microRNA target predictions led to the experimental validation of *Mtpn* as the first mammalian gene shown to be regulated coordinately by three microRNAs.

2.3.2 Drosophila

We used PicTar and cross-species comparisons of seven recently sequenced *Drosophila* species to predict and analyze microRNA targets in flies. We also computed predictions for common targets of clustered microRNAs, since recent experiments [125, 11] have suggested that microRNA genes that reside in clusters spanning roughly 50 kilobases of genomic DNA tend to be co-expressed. To shed light on the specific function of microRNAs, we analyzed the functional annotation for predicted target sets using Gene Ontology (GO) terms [47]. However, to arrive at a more global understanding of microRNA function we then asked if the extent of microRNA targeting in flies is comparable to targeting in vertebrates, if certain microRNA:mRNA regulatory relationships are conserved between both clades, and if individual microRNAs could potentially play a role in clade-specific gene regulation. We present the results obtained with the original PicTar version along with a comparison to the improved version, this way demonstrating the improvements of the new scoring scheme.

Genome-wide cross-species comparison of seven fly species allow high-specificity and high-sensitivity microRNA target predictions

It has been widely demonstrated that the success of the computational identification of microRNA target sites can be significantly boosted by searching for evolutionarily conserved target sites, which are therefore likely to be functional. Thus we set out to make use of the very recent whole-genome sequencing of a number of fly species (fig. 2.6). The genomic sequence for eight of these species, including members of the *melanogaster*, *obscura*, *repleta*, and *virilis* groups, have been already assembled (*D. melanogaster*, *D. simulans*, *D. yakuba*, *D. erecta*, *D. ananassae*, *D. pseudoobscura*, *D. virilis*, and *D. mojavensis*). We discarded the *D. simulans* assembly since at this point it still contained large gaps. The estimated divergence time for these species ranges from a few million years to roughly 40 million years (fig. 2.6).

Two sets of 3'UTR alignments

To identify evolutionarily conserved microRNA target sites in 3'UTR sequences, it was critical to identify orthologous mRNAs. We experimented with two independently produced sets of genome-wide alignments of the eight species. The first set of alignments (termed set 1), which does not contain sequence for *D. erecta*, was produced by the UCSC Genome database (*http://genome.ucsc.edu/*) and is based on pairwise alignments, which were subsequently multiply aligned. The following assemblies were used to construct the multiz alignments [17]: *D. melanogaster* Apr. 2004 (dm2), *D. yakuba* Apr. 2004 (droYak1), *D. ananassae* Jul. 2004 (droAna1), *D. pseudoobscura* Aug. 2003 (dp2), *D. virilis* Jul. 2004 (droVir1), *D. mojavensis* Aug. 2004 (droMoj1), *A. gambiae* Feb. 2003 (anoGam1), and *A. mellifera* Jul. 2004 (apiMel1). The detailed amount of nucleotides and aligned sequence for all flies is shown in Table 2.2. The second set (termed set 2)

Set	Category	*D. melanogaster*	*D. yakuba*	*D. erecta*	*D. ananassae*	*D. pseudoobscura*	*D. virilis*	*D. mojavensis*
1	All genes	18,892	18,718	—	17,380	16,032	14,351	13,465
	Unique genes	9,958	9,923	—	9,411	8,744	7,878	7,425
2	All genes	18,381	17,696	17,061	15,765	14,601	13,366	13,030
	Unique genes	9,771	9,521	9,283	8,826	8,354	7,795	7,614

Total number of UTR alignments with sequence for all species up to the indicated one, referring to the order *D. melanogaster, D. yakuba, D. erecta, D. ananassae, D. pseudoobscura, D. virilis, D. mojavensis.*
DOI: 10.1371/journal.pcbi.0010013.t001

Table 2.2: **Statistics of 3'UTR alignments**
Total number of nucleotides per species in the multiple alignments for set 1 and set 2 (for all genes and for unique genes with both masked and unmasked repeats).

came from true genome-wide multiple alignments generated by the Pachter group at UC Berkeley [18] using the following assemblies: *D. melanogaster* Apr. 2004 (dm2), *D. ananassae* Jul. 2004 (droAna1), *D. yakuba* Apr. 2004 (droYak1), *D. erecta* Oct. 2004, *D. pseudoobscura* (dp1), *D. virilis* Jul. 2004 (droVir1), *D. mojavensis* Dec. 2004. For both datasets we used FlyBase Release 4.1 to extract 3'UTRs in *D. melanogaster*. For both sets, we extracted multiple alignments of *D. melanogaster* 3'UTRs using the *D. melanogaster* FlyBase annotation for 18,892 gene transcripts and obtained 3'UTR

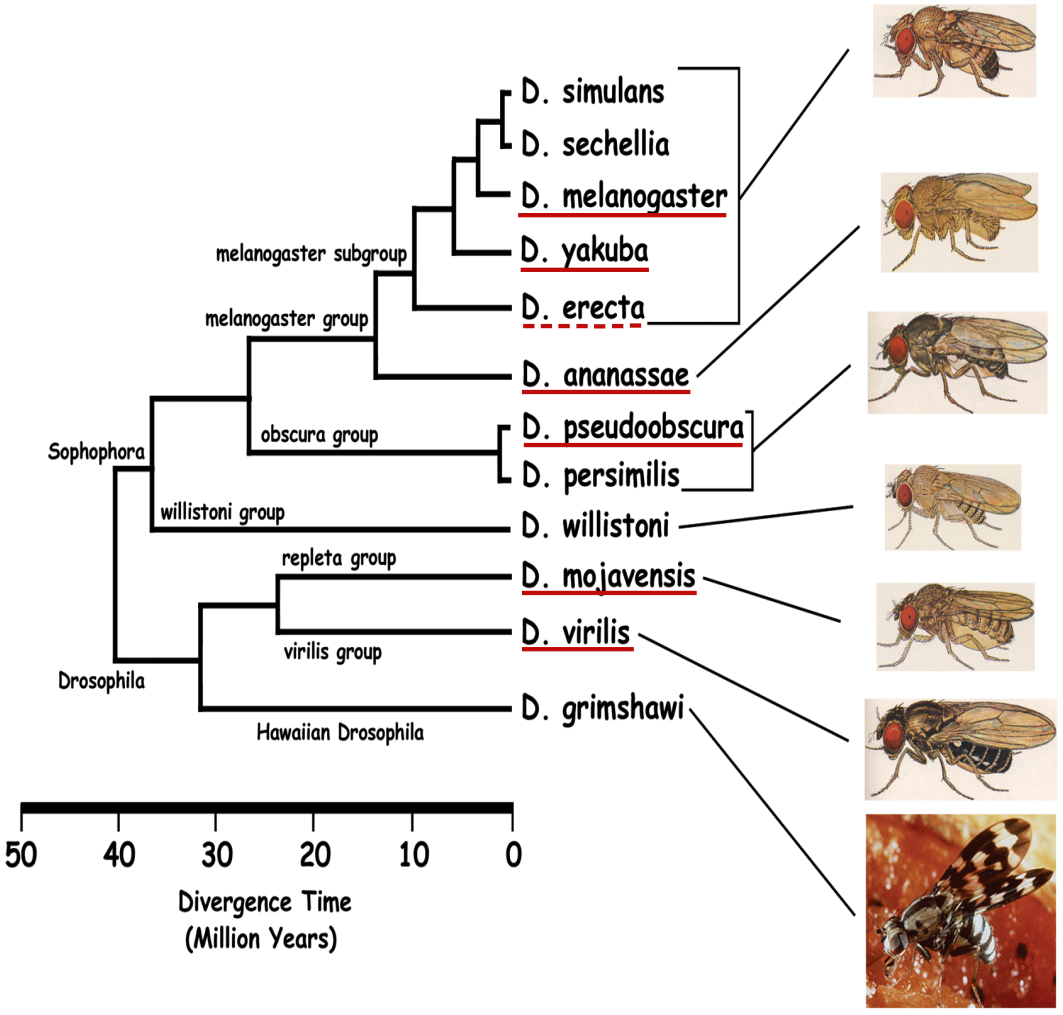


Figure 2.6: **Phylogenetic tree of 12 *Drosophila* species**
Phylogenetic tree of 12 *Drosophila* species (retrieved from *http://species.flybase.net/*). Our datasets include 3'UTRs for seven of these species: *D. melanogaster*, *D. yakuba*, *D. erecta*, *D. ananassae*, *D. pseudoobscura*, *D. virilis*, and *D. mojavensis*. Species underlined in solid are present in set 1 and set 2. *D. erecta* (broken line) is present only in set 2.

alignments across all eight species for 13,465 transcripts (set 1) and 13,030 transcripts (set 2). We also defined sets of alignments by keeping only the longest 3'UTR from all transcript variants for the same gene, resulting in ∼9,800 alignments for each set (termed unique alignments). The coverage of genes is thus roughly comparable between both sets. Additionally, we masked repeats in the unique alignments using the UCSC repeat masks for set 1 and using the Tandem Repeat Remover [14] following Rajewsky et al. [109] for set 2. The nucleotide space of the various alignment sets is listed in Table 2.3 and comprises for each set a total of 2.2-4.1 Mb per species for the repeat-masked

unique alignments. Masking repeats thus removed substantial amounts of sequence (up to 22% per species).

Set	Category	*D. melanogaster*	*D. yakuba*	*D. erecta*	*D. ananassae*	*D. pseudoobscura*	*D. virilis*	*D. mojavensis*
1	All genes	6,833,600	6,837,151	—	6,248,338	6,013,857	4,811,921	4,510,597
	Unique genes	3,906,057	3,910,995	—	3,494,974	3,292,411	2,600,794	2,453,927
	Unique genes, masked repeats	3,761,764	3,766,941	—	3,301,278	3,092,154	2,326,277	2,159,408
	Percent of repeats	3.69%	3.68%	—	5.54%	6.08%	10.56%	12.00%
2	All genes	6,389,344	6,559,084	6,084,950	9,936,560	7,194,840	8,773,428	8,838,383
	Unique genes	3,681,969	3,813,324	3,546,121	4,700,857	4,204,005	5,123,907	5,062,488
	Unique genes, masked repeats	3,190,257	3,299,538	3,082,864	3,938,559	3,437,270	4,078,727	3,949,839
	Percent of repeats	13.35%	13.47%	13.06%	16.22%	18.24%	20.40%	21.98%

Total number of nucleotides per species in the multiple alignments for set 1 and set 2 (for all genes and for unique genes with both masked and unmasked repeats).
DOI: 10.1371/journal.pcbi.0010013.t002

Table 2.3: **Statistics of 3'UTR alignments**
Total number of nucleotides per species in the multiple alignments for set 1 and set 2 (for all genes and for unique genes with both masked and unmasked repeats).

microRNA sequences

We downloaded all *D. melanogaster* microRNA precursors and mature microRNAs from the microRNA registry at Rfam [48] (Release 5.0). For each microRNA, we checked for conservation of the precursor sequence in all fly species, using multiple alignments retrieved from the UCSC Genome Database. We required the first 8mer of the mature microRNA to be perfectly conserved, but applied a less stringent conservation constraint, a percentage identity of 75%, to the precursor sequence. From the 79 mature *D. melanogaster* microRNAs we found 69 to be conserved in all flies and 73 to be conserved in the *melanogaster* and *obscura* groups. Statistics were generated with a subset of 46 microRNAs with unique nuclei, i.e. each nucleus is specific for only one microRNA in this list. To compute signal-to-noise ratio, we produced five cohorts of unique randomized microRNAs for set 1 and set 2, respectively, in either case both with masked and unmasked repeats.

Sensitivity-Specificity estimation for different parameter settings

To estimate the extent of microRNA targeting in *Drosophila*, we used PicTar to count conserved putative target sites with perfect nuclei (anchors). The microRNAs used for these searches consisted of all currently known microRNAs that seemed to be conserved in all species under consideration. To avoid counting of target sites more than once, we represented all microRNA "families" that share identical nuclei by just one member of each family. The final set contained 46 microRNAs with unique nuclei conserved in all flies. As in our previous study [68], we recruited cohorts of randomized microRNA sequences to estimate the number of false positives. Specifically, we computed all anchor sites (single conserved nuclei) for set 1 and set 2 with masked and unmasked repeats for real microRNAs, as well as for five sets of randomized cohorts in each case (fig. 2.7). A measure for the specificity is the signal-to-noise ratio, which is defined as the ratio of the

number of anchor sites for real versus randomized microRNAs. In each case, we averaged the result over five cohorts and computed the mean and the standard deviation of the signal-to-noise ratio. We computed specificity and sensitivity, requiring different degrees of evolutionary conservation of anchor sites both with and without free energy filtering of perfect nuclei (fig. 2.7). Overall, we observed that using the free energy filter or masking repeats tends to enhance specificity with modest losses in sensitivity. We obtained higher signal-to-noise ratios with set 2, but a higher sensitivity with set 1. We also found that requiring different degrees of evolutionary conservation of anchor sites strongly affects sensitivity and specificity. More precisely, searching for anchor sites conserved between all flies (at various parameter settings) yielded a signal-to-noise ratio of 2.8-3.6 (set 1) and 3.3-3.9 (set 2). The sensitivity was, on average, 25-33 (set 1) and 15-29 (set 2) anchor sites per microRNA above noise. Anchor sites conserved in the *melanogaster* and *obscura groups* yielded signal-to-noise ratios of 2.1-2.4 (set 1) and 2.3-2.7 (set 2) with a sensitivity of 47-57 (set 1) and 29-40 (set 2) anchor sites per microRNA above noise (2.7).

A major determinant of sensitivity and specificity is the required level of conservation of anchor sites. Based on these results we defined three settings, termed S1, S2, and S3, that allowed us to adjust the trade-off between sensitivity and specificity, and to generate predictions of high sensitivity, high specificity, and medium specificity/sensitivity, respectively. Masking repeats and applying free energy filtering of anchor sites serves to fine-tune the trade-off between sensitivity and specificity for each setting. The high sensitivity setting (S1) has repeat masked UCSC alignments (set 1) as input sequences, requires conservation of anchor sites only between species of the *melanogaster* and *obscura* groups and applies no free energy filtering of perfect nuclei. Setting S2, providing high specificity predictions, uses alignments of set 2 with unmasked repeats as input sequences and requires conservation of anchors in all flies and free energy filtering of perfect nuclei. The medium sensitivity/medium specificity setting S3 is equal to setting S1, but uses conservation of anchors in all flies.

Phylogenetic PicTar score

Given an alignment of a 3'UTRs for all flies, PicTar computes a likelihood score for the UTR of each species separately. The final score of the whole alignment is a weighted average of the single species scores, with weights reflecting the phylogenetic grouping of the species. More precisely, the score of all flies in the *melanogaster* subgroup is averaged and the resulting score is further averaged with the score for *D. ananassae* and *D. pseudoobscura* yielding a score for the *melanogaster* and *obscura* groups. The score for *D. mojavensis* and *D. virilis*, which have similar evolutionary distances to the *melanogaster* group, are averaged. This outgroup score and the score of the *melanogaster* and *obscura* groups are averaged to obtain the final PicTar score for all flies.

Score dependence of specificity/sensitivity

For each of the settings S1-3 we recorded the specificity and number of targeted transcripts as a function of the PicTar score cutoff, i. e. discarding all predictions with a score lower than a given threshold (fig. 2.8). We found that high scoring transcripts tend to have a significantly improved specificity. For example, when using setting S2, the signal-to-noise

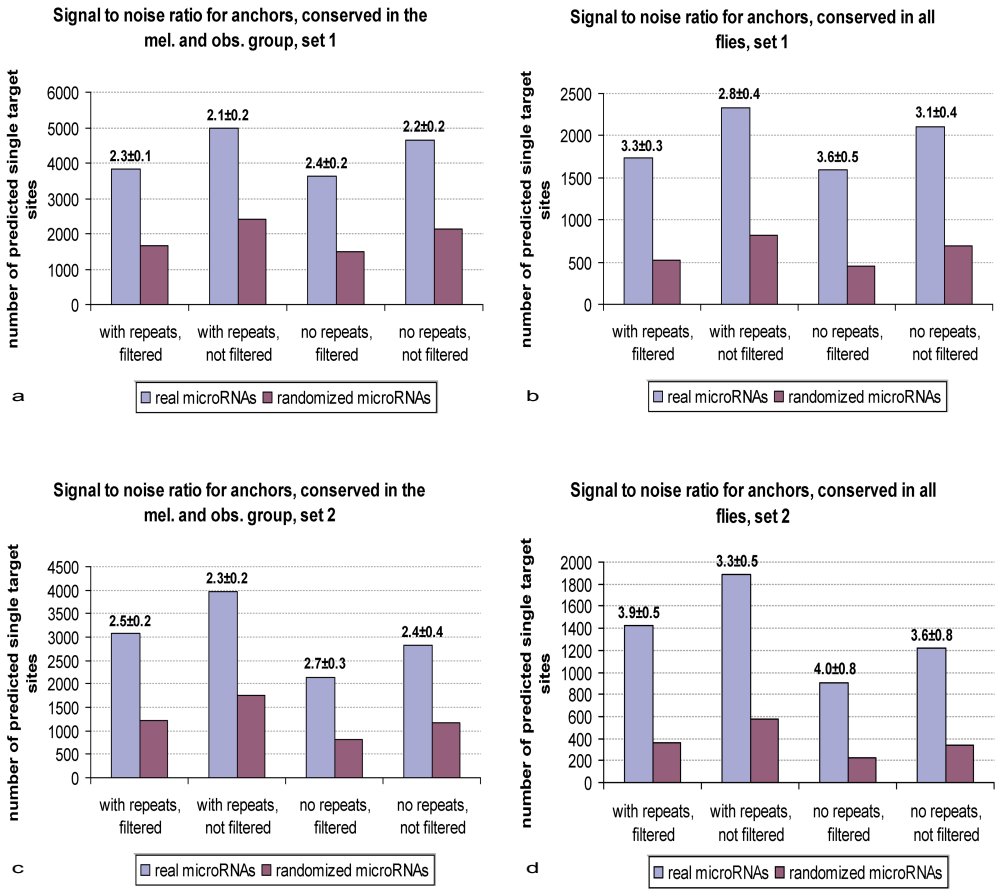


Figure 2.7: **signal-to-noise ratios of the PicTar single target site predictions** For both set 1 and set 2 the predicted number of anchor sites for 46 unique microRNAs, conserved in all flies, and corresponding randomized microRNAs (averaged over five cohorts) and the respective signal-to-noise ratio (indicated above the bars), are shown with and without using free energy filtering of anchor sites for UTRs with masked and unmasked repeats, respectively. (a) Predictions for set 1 with anchor sites conserved in the *melanogaster* and *obscura* groups. (b) Predictions for set 1 with anchor sites conserved in all flies. (c) Predictions for set 2 with anchor sites conserved in the *melanogaster* and *obscura* groups. (d) Predictions for set 2 with anchor sites conserved in all flies.

ratio can be improved by a factor of 1.7 while retaining a sizable number of predicted transcripts per microRNA. The positive correlation between specificity and PicTar score is consistent with our observation that some non-anchor sites make a contribution to the score. These sites appear to be "scattered", i.e. are present only in some species or are not found in all species at the same position in the alignment. We experimented with relaxing our anchor site definition to include cases where a perfect nucleus is found in all species under consideration but not necessarily at overlapping positions in the alignments. The signal-to-noise ratio decreased in all settings S1-3 (for example for S3 from 3.3 to 2.6), at no significant gain in sensitivity. We thus concluded that many scattered sites

could be functional but should be scored only when they occur in conjunction with anchor sites, as implemented in the PicTar algorithm. These results, obtained with the original

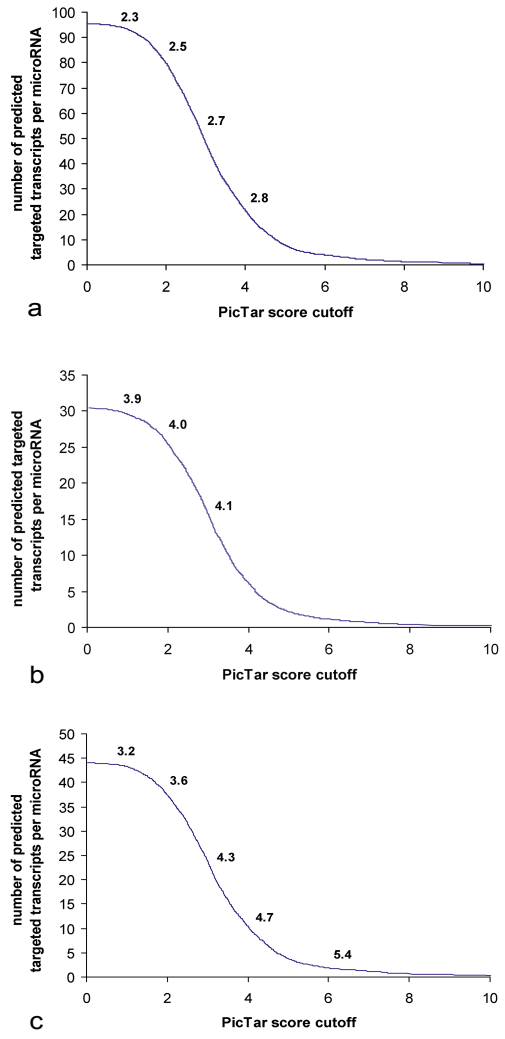
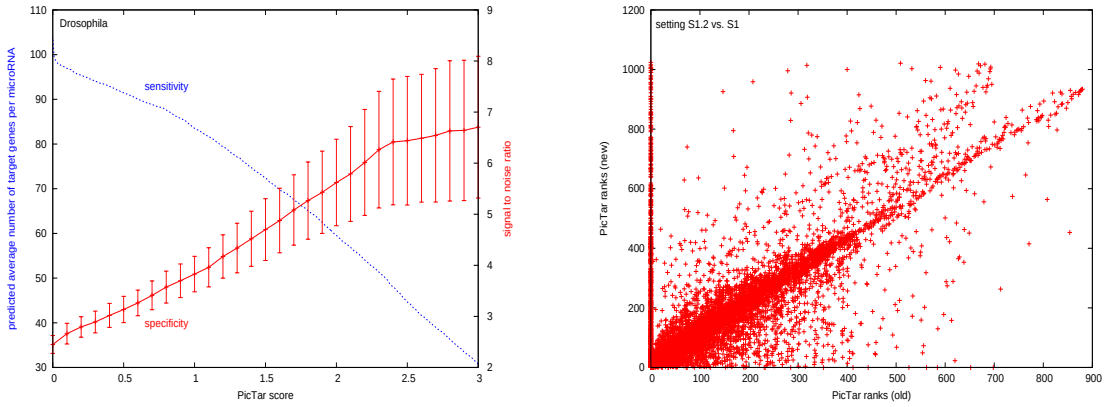


Figure 2.8: **PicTar score dependent sensitivity and specificity.**
Shown is the average number of predicted targeted genes as a function of a PicTar score cutoff (discarding all target genes with a score below this cutoff) for three different PicTar settings (S1-3, see text): (a) high sensitivity settings (S1), (b) high specificity settings (S2), (c) medium sensitivity/medium specificity settings (S3). The signal-to-noise ratio also depends on the score cutoff and is indicated above the curve for certain cutoff values. All predictions for all settings can be accessed on the PicTar web server (not filtered by score cutoffs).

algorithm, we can compare to predictions produced with the improved algorithm [79]. We computed new target predictions based on our UCSC 3'UTR alignments at two different levels of conservation, corresponding to modified versions of the settings S1 and S3. The input data, i.e. the datasets of conserved microRNAs and the UTR alignments as well

as the cohorts of randomized microRNAs remained unchanged compared to the original version. The modified settings, used for the new predictions are termed S1.2 and S3.2, respectively. Setting S1.2 requires conservation of target sites between *D. melanogaster*, *D. yakuba*, *D. ananassae* and *D. pseudoobscura*. However, since the latter two species are of similar evolutionary distance to the reference species *D. melanogaster* we also consider an anchor site conserved, if a nucleus is found in only one of those species and the sequence of the entire UTR is missing in the third species. Setting S3.2 requires an anchor site to be additionally conserved in *D. virilis* and *D. mojavensis*, and these two species are again considered to be of equal evolutionary distance to *D. melanogaster* and treated accordingly. As opposed to the original version of PicTar, we do not mask repeats in both cases anymore in order to maintain the highest possible sensitivity. Hence, the only remaining difference between settings S1.2 and S3.2 is the different level of conservation. All additional modifications with respect to free energy filtering and the treatment of imperfect nuclei were applied as already explained. The conservation score was determined by the number of occurrences of a particular nucleus in anchor sites conserved in *D. pseudoobscura* (for S1.2) and *D. mojavensis* (for S3.2), respectively, divided by the number of occurrences in *D. melanogaster*. For the setting S1.2 we predict on average 105 target genes per microRNA compared to 95 for the setting S1. We computed the specificity by averaging the ratio of the number of predicted targets for real and randomized microRNAs over four cohorts and obtained a signal-to-noise ratio of 2.5 (compared to 2.3 for S1). For the Setting S3.2 PicTar computes on average 49 target genes per microRNA at a signal-to-noise ratio of 3.6. Compared to the first version of PicTar this means an increase of $\sim$5 predicted target genes per microRNA at an enhanced signal-to-noise ratio (3.2 for setting S3). Compared to setting S1 and S3, respectively, the specificity of the new PicTar predictions depends much stronger on a PicTar score cutoff (shown for S1.2 in fig. 2.9a). As can be seen from this Figure, applying a score cutoff allows us to extract highly specific target predictions without losing too many predicted targets. For setting S1.2 (S3.2) we obtained a linearly increasing signal-to-noise ratio, from 2.5 (3.6) with no score cutoff to 6.5 (15) at a score cutoff of 2.5. The sensitivity drops down linearly from 105 (49) predicted target genes per microRNA with no score cut off to 45 (15) genes per microRNA at a score cutoff of 2.5. Although the PicTar modifications result in a less restrictive treatment of imperfect nuclei, the strong requirement of compensatory binding of the microRNA 3' end keeps their contribution very low. Among all nuclei in anchor sites, imperfect nuclei are only observed in 1% of all cases when using setting S1.2 and 0.5% of all cases when using setting S3.2. The fraction of genes with at least one predicted microRNA target site also increases by approximately 10% in the new version. We predict that 29% (16%) of all 9958 unique genes are predicted targets of microRNAs for setting S1.2 (S3.2) as opposed to 27%(15%) for setting S1(S3). To assess the influence of the PicTar modifications on the ranks of microRNA/target gene pairs, we plotted all PicTar ranks of the previous version against all ranks computed using the new version (fig. 2.9b). Whenever a gene is a predicted target of a particular microRNA for one PicTar version, but not contained in the target list of the other it is assigned rank zero for the version missing this prediction. Thus, all newly predicted targets of the new version reside on the y-axis, while all predictions present in the old but not in the new version are located on the x-axis. In the new version, we loose none of our previously predicted targets, while the number of predicted



a **b**

Figure 2.9: **PicTar score dependent sensitivity and specificity of the new version and comparison to the old ranks.**
(a) Signal-to-noise ratio (specificity, solid line) and average number of predicted target genes (sensitivity, dashed line) as a function of a PicTar score cutoff for the modified PicTar algorithm in flies for setting S1.2. The signal-to-noise ratio was averaged over four cohorts of randomized microRNAs in flies. The standard deviation is indicated by error-bars. (b) Scatterplot of target gene ranks of the previous list of PicTar predictions against the PicTar ranks extracted from the new prediction list for setting S1.2 versus S1, respectively. Each dot represents a particular microRNA/target gene pair. If a microRNA/target gene pair is absent in the predictions of one version, it is assigned rank zero for the respective version.

targets grows by ∼10%. For both settings, the plot shows a systematic deviation from the diagonal. This can be explained by a shift to higher ranks due to the insertion of new predictions into the list of target genes. Because the new scoring scheme assigns a distinct probability to each possible nucleus of a particular microRNA, many ranks are observed to change significantly. To assess the degree of similarity between both versions, we asked for each microRNA for how many of the predicted targets the difference between the old and the new rank is less than 25% of the number of targets predicted for the microRNA under consideration. We found that this is the case for roughly 70% of all predictions with setting S1.2 and for 74% when using setting S3.2. We conclude, that the modifications of PicTar strongly enhance the specificity as well as the sensitivity of the algorithm virtually without discarding any of the previous predictions.

Identifying putative targets of multiple microRNAs

Previous analyses of microRNA targeting in vertebrates [59, 83, 68, 145] and flies [43, 20] suggested that a substantial fraction (10%-30%) of all protein-coding genes in both clades is regulated by microRNAs. Using settings S3 (S2), we found that 15% (13%) of all annotated ∼10,000 unique *melanogaster* 3'UTR transcripts (corresponding to ∼10,000 genes) have at least one anchor site that is conserved in all seven fly species at a

signal-to-noise ratio of ∼3 (4). Thus, with settings S3 or S2 roughly 10% of all transcripts are predicted to be targeted by microRNAs above noise in all flies. To estimate how many genes could be regulated by more than one microRNA, we counted all transcripts with at least two anchor sites. Applying the high specificity setting S2, we found that searching for multiply targeted transcripts further enhances the specificity to a significant degree (fig. 2.10). For example, we found seven times as many targeted transcripts with at least

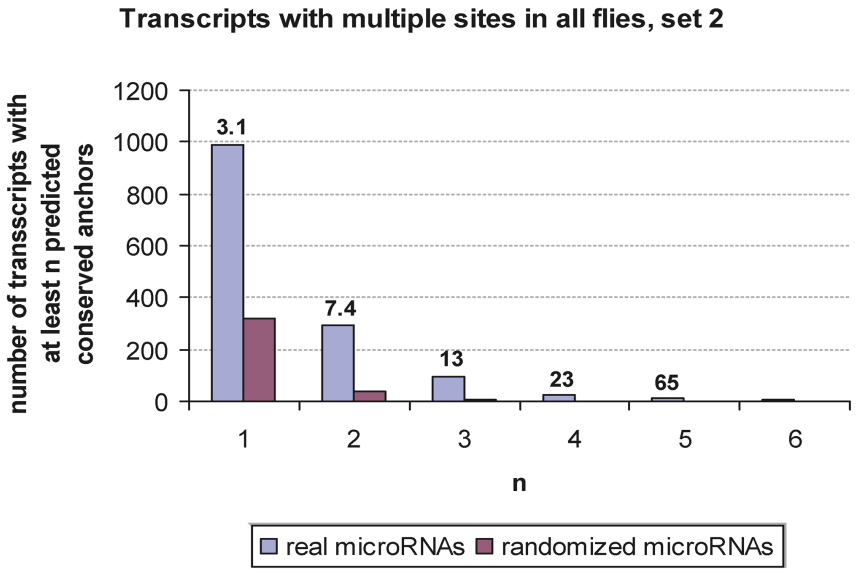


Figure 2.10: **Specificity of PicTar predictions of genes with multiple putative target sites.**
Number of unique genes as a function of the minimal number of anchor sites for 46 unique, conserved microRNAs and for randomized microRNAs (averaged over 5 cohorts). The ratio of these numbers, reflecting the specificity, is indicated above each bar.

two anchor sites for real microRNAs compared to randomized microRNAs. With settings S2 and S3, we predicted that 30% of all targeted transcripts have more than one anchor site. Finally, for our high sensitivity setting S1 we found that 27% of all transcripts have at least one anchor site at a single site signal-to-noise ratio of ∼2.2. Of these, 40% are found to have at least two anchor sites.
In summary, based on our high sensitivity settings, we predicted that at least 15% of all *D. melanogaster* genes with currently annotated 3'UTR sequences are regulated by at least one known microRNA, and that at least one fifth of these *Drosophila* microRNA targets could be subject to coordinate control of two or more microRNAs from different microRNA families (above noise). We provide ranked PicTar target predictions for all conserved microRNAs, all FlyBase transcripts, and settings S1-3 at our searchable website *http://pictar.bio.nyu.edu*. The results, linked to various other public databases, can be queried for genes of interest or microRNAs of interest.

Recovery of experimentally validated microRNA targets in *Drosophila*

To analyze the recovery of experimentally validated targets in *Drosophila*, we collected 19 microRNA:target regulatory relationships from the literature [132, 114, 20]. The overlap with PicTar predictions across settings S1-3 is summarized in Table 2.5. The apoptosis gene *hid/wrinkled* (W) is targeted by the microRNA *bantam* [19]. For all settings S1-3, *hid* is the top scoring *bantam* target (PicTar score 17.3) and has five anchor sites conserved in all flies. Notably, *hid* targeted by *bantam* has the second highest PicTar score within all our target predictions. The only gene with a higher score (40.5) is *nerfin-1*, which contains two anchor sites for *miR-286* (or equivalently *miR-279*) conserved in all flies and many additional sites for the same microRNA (see below). The *Notch* signaling gene *hairy* was recently predicted [132, 110] and validated as a target of *miR-7* with a single binding site [132]. PicTar found a *miR-7* anchor site conserved in all flies of the *melanogaster* and *obscura* groups, whereas the site in *D. virilis* appears to be slightly shifted upstream. Hence, this target is recovered with setting S1 but not with settings S2-3. There is experimental evidence that *miR-7* also targets *HLHm3* and *E(spl)m4*, two genes located in the *E(spl)*-complex [132]. For *HLHm3*, PicTar predicts one *miR-7* target site conserved in all flies (at all settings). The gene *E(spl)m4* did not have an annotated 3'UTR but was recovered after adding the likely 3'UTR sequence to our dataset [132]. Another gene of the *E(spl)*-complex, *HLHm5*, is the highest ranking target gene of *miR-7* when searching for targets conserved in all flies (with setting S2; rank 2 with setting S3). Target predictions at a reduced level of conservation (setting S1) also yield *HLHm5* as the top-ranking *miR-7* target. The *Notch* gene *Bearded* is recovered as a target of *miR-4* (or *miR-79*, equivalently). With setting S1 we found three conserved sites in its 3'UTR. These so called *Brd* boxes have been shown to mediate repression of a reporter gene with a *Brd* 3'UTR *in vivo* [75]. This gene is again very high scoring (15.6) and is found on rank 2 in the list of *miR-4* target predictions (setting S1). This target is not recovered with the other settings, because the alignments of this gene do not contain sequence for *D. mojavensis* and *D. virilis*. The same microRNA is thought to repress *bagpipe* [20], which is found on rank 2 in the list of *miR-4* target predictions (S3).
The proapoptotic genes *reaper*, *grim* and *sickle* are validated targets of the *miR-2* family [132]. For *sickle* we found one conserved site in all flies for *miR-2*, *miR-13* and *miR-6*, which share the same nucleus. For *reaper*, we recovered one site for the same microRNAs in the *melanogaster* and *obscura* group with setting S1, while the other settings failed to identify this target because of missing sequence for this gene in *D. mojavensis*. *grim* is the only target of this group not recovered by PicTar, because it has only a 6mer nucleus for *miR-2*.
A recent algorithm for the prediction of microRNA targets did not rely on evolutionary information, but incorporated the 3'UTR secondary structure to compute putative microRNA targets [114]. Some of the high scoring predictions could then be supported by luciferase reporter constructs in cell lines. We recovered four targets from this list (*miR-7/HLHm5*, *miR-279/SP555*, *miR-124/Gli*, *miR-310/imd*) but failed to locate conserved nuclei for the other six targets (see comments in Table 2.5). Strikingly, out of nine computationally predicted targets that were experimentally assayed but did not show any repression activity (likely false positives) [114], we only predicted one microRNA:target regulatory relationship (*miR-286/boss*). In summary, PicTar recovered 8/9 (89%) of all

Category	microRNA–Target	S1	S2	S3	Comments
microRNA targets with experimental support [4,14,24]	*bantam–hid*	+	+	+	
	miR-7–hairy	+	–	–	Not strictly conserved in all flies but scattered sites present
	miR-7–HLHm3	+	+	+	
	miR-7–m4	+	+	+	3′ UTR absent in FlyBase 4.1 annotation
	miR-4–Bearded	+	–	–	Not conserved in all flies
	miR-4–bagpipe	+	+	+	
	miR-2–sickle	+	+	+	
	miR-2–reaper	+	–	–	Not conserved in all flies
	miR-2–grim	–	–	–	Nucleus consists of six Watson–Crick basepairings and one G/U
microRNA targets with experimental support [12] (Luciferase reporter assays in cell lines)	*bantam–MAD*	–	–	–	
	miR-287–CRMP	–	–	–	
	miR-7–HLHm5	+	+	+	
	miR-279–SP555	+	+	+	
	miR-310–imd	+	+	+	Recovered if *miR-310* presumed to be conserved in all flies
	miR-1–tutl	–	–	–	
	miR-34–su(z) 12	–	–	–	Not recovered because nucleus overlaps with repeat
	miR-12–rt	–	–	–	
	miR-124–gli	+	+	+	
	miR-7–fng	–	–	–	
False positives according to experiments [12]	*miR-287–dip1*	–	–	–	
	miR-303–CG14991	–	–	–	
	miR-278–tup	–	–	–	
	miR-317–yellow-c	–	–	–	
	miR-318–CG13380	–	–	–	
	miR-286–boss	+	+	+	
	miR-288–CG32057	–	–	–	
	miR-276b–ke1	–	–	–	
	miR316–ia2	–	–	–	

Experimentally assayed microRNA target sites are listed in the second column, comprising 19 microRNA–gene regulatory relationships with various degrees of experimental support and nine sites that did not show regulatory activity. Columns labeled by S1–S3 refer to the recovery of sites at the corresponding PicTar setting.

DOI: 10.1371/journal.pcbi.0010013.t003

Table 2.4: **Recovery of published Drosophila microRNA targets with experimental support.**
Experimentally assayed microRNA target sites are listed in the first column, totaling 19 microRNA:gene regulatory relationships with various degrees of experimental support and 9 sites which did not show regulatory activity. Columns labeled by S1-S3 refer to the recovery of sites at the corresponding PicTar setting.

known targets with experimental *in vivo* evidence and 4/10 or 40% of targets with other experimental support with setting S1, i.e. requiring conservation of anchor sites only in flies of the *melanogaster* and *obscura* groups. Only three of all targets with experimental support were lost when requiring conservation between all fly species and thus were not recovered with settings S2 and S3.

Clustered microRNAs are likely to coordinately regulate gene expression

Expression assays have shown that human microRNA genes that are located in the same genomic region within 50 kilobases of each other are often co-expressed in specific tissues [125, 11], suggesting the possibility that they may coordinately regulate common target genes. In *D. melanogaster*, we identified 7 clusters within 50 kilobase regions that contain precursors of at least two conserved microRNAs from different families. To identify common targets of clustered microRNAs in flies, we used PicTar to predict coordinate targets for each of these microRNA clusters (available on the PicTar server). Table 2.5 gives an overview of all clusters, their location in the *Drosophila* genome, the abundance of targeted transcripts and, whenever all microRNA genes of a given cluster are located in an intron of another gene, the identifier of this gene. To evaluate whether clustered microRNAs target the same gene more often than expected by chance, we considered all 1,128 pairwise combinations of all 48 unique conserved microRNAs. While pairs of microRNAs from the same cluster comprise only 2.1% of these pairs, 132 genes contained at least one anchor site for each microRNA of these clustered pairs (using setting S1), or 12% of the 1,104 genes that contain at least two different anchor sites for any combination of these 48 microRNAs. Thus, pairs of microRNAs from clusters are likely to coordinately regulate a significantly higher proportion of genes (12%) than expected (2.1%). To assess the significance of targeting by 24 pairs of microRNAs extracted from clustered microRNA genes, we used 1,000 sets of 24 pairs of microRNAs drawn randomly from the set of all possible 1,128 distinct pairs (using all 48 unique microRNAs conserved in the melanogaster and obscura groups). For conservation of anchor sites in the *melanogaster* and *obscura* groups, on average 18($\pm$2) out of 24 random pairs have at least one target gene, compared to 22 (Z=2) of the co-expressed pairs (Z=2). We obtained on average 70($\pm$21) unique target genes per random set, compared to 132 unique targets of the clustered pairs with a high Z-value ($Z = 3$). When requiring conservation between all flies, the results are more significant: 19 out of 24 clustered pairs target 50 unique genes, while on average 11 (2) out of 24 randomly drawn doublets (Z=4) are predicted to target approximately 23(8) unique genes (Z=3.5). These findings support the hypothesis of coordinate control executed by clustered microRNAs.

Biological and molecular classification of predicted microRNA targets

To gain insights into the function of *Drosophila* microRNAs, we used GeneMerge [27] to analyze the over-representation of specific Gene Ontology (GO) terms [47] in the functional annotation of genes predicted to be targeted by a particular microRNA versus a background gene set. GeneMerge computes the significance of occurrences of particular GO terms for a set of genes compared to a background gene set. Briefly, GeneMerge takes as input the set of genes of interest (predicted targets for a specific microRNA), a background set representing the universe of genes (all genes that are potential PicTar

Cluster Number	Chromosome	Start Position	Stop Position	Strand	microRNA Precursor	Number of Unique Targeted Transcripts Conserved in the *melanogaster* and *obscura* Groups (Set 1)	Number of Unique Targeted Transcripts Conserved in All Flies (Set 1)	Number of Unique Targeted Transcripts Conserved in the *melanogaster* and *obscura* Groups (Set 2)	Number of Unique Targeted Transcripts Conserved in All Flies (Set 2)	FlyBase HostGene
1	chr2L	7425795	7425892	+	*dme-mir-275*	2	0	3	2	
	chr2L	7425972	7426044	+	*dme-mir-305*					
2	chr2L	16693853	16693944	+	*dme-mir-9c*	17	7	11	4	CG17161
	chr2L	16694333	16694417	+	*dme-mir-306*					
	chr2L	16694483	16694579	+	*dme-mir-79*					
3	chr2L	17562299	17562398	+	*dme-mir-124*	0	0	0	0	
	chr2L	17570539	17570631	+	*dme-mir-287*					
4	chr2L	18467363	18467462	+	*dme-mir-100*	1	1	2	1	
	chr2L	18467963	18468040	+	*dme-let-7*					
	chr2L	18468244	18468353	+	*dme-mir-125*					
5	chr2R	15175579	15175658	–	*dme-mir-6-3*	55	33	38	26	
	chr2R	15176021	15176089	–	*dme-mir-5*					
	chr2R	15176152	15176232	–	*dme-mir-4*					
	chr2R	15176285	15176384	–	*dme-mir-286*					
	chr2R	15176458	15176526	–	*dme-mir-3*					
6	chr3R	5916848	5916939	+	*dme-mir-317*	31	9	19	3	
	chr3R	5925744	5925843	+	*dme-mir-277*					
	chr3R	5926658	5926756	+	*dme-mir-34*					
7	chrX	15341893	15341992	+	*dme-mir-283*	34	5	16	3	CG33206
	chrX	15342896	15342983	+	*dme-mir-304*					
	chrX	15343410	15343483	+	*dme-mir-12*					
Overall number of unique targeted genes						132	50	85	38	

Clusters of unique microRNAs, conserved in all flies, with precursor sequences, originating from a genomic region of less than 50 kB. The number of unique genes with at least two anchor sites for different microRNAs of a given cluster is indicated. Predictions are computed for both set 1 and set 2, and for anchors conserved in the *melanogaster* and *obscura* groups, and in all seven fly species. If clustered microRNA precursors reside in an intron of an annotated FlyBase gene, the identifier is also indicated.

DOI: 10.1371/journal.pcbi.0010013.t004

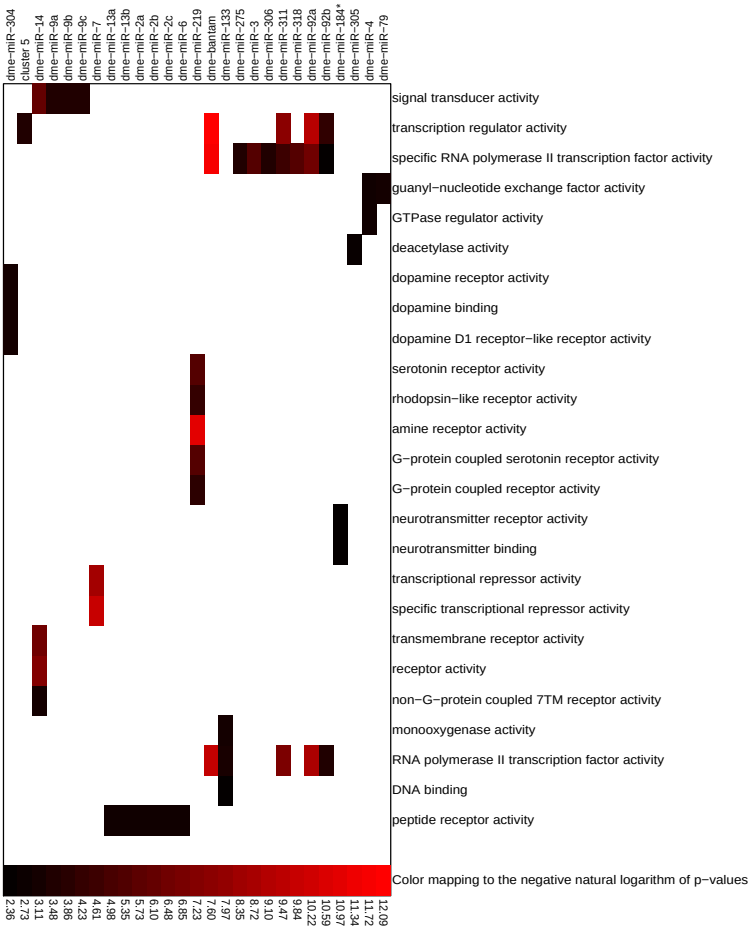
Table 2.5: **Clusters of microRNAs and their number of predicted target genes.** Clusters of unique microRNAs, conserved in all flies, with precursor sequences, originating from a genomic region of less than 50 kilobases. The number of unique genes with at least two anchor sites for different microRNAs of a given cluster is indicated. Predictions are computed for both set 1 and set 2, anchors conserved in the *melanogaster* and *obscura* groups and in all seven fly species, respectively. If clustered microRNA precursors reside in an intron of an annotated FlyBase gene, the identifier is also indicated.

targets), and an association file containing GO term annotations for each gene. It then outputs a list of GO terms which are overrepresented in the study set vs. the background set, using a hypergeometric distribution to compute a p-value, followed by a Bonferroni correction to account for multiple testing.
To use an extensive background gene set, which captures features of genes targeted by microRNAs as best possible, we lumped together all genes predicted to be targeted by all microRNAs (setting S1) or genes, which were hit by the five cohorts of randomized microRNAs. Finally, we recorded all p-values below a cutoff of 0.1. These p-values are conservatively corrected for multiple testings, as provided by GeneMerge. We performed the analysis separately for all GO terms in the"Biological Processes" ontology, and the most specific "Biological Processes" GO term for each gene, as well as for all GO terms in the "Molecular Function" ontology. These three classes of GO terms are provided by GeneMerge. Results from the first two analyses were merged into one output file, keeping the lower p-value for GO terms which were present twice. To visualize the results, we used two-way hierarchical clustering based on the linear correlation coefficient of the negative logarithm of the p-value [56].
From the "Biological Process" ontology, a total of 112 significantly over-represented GO terms were identified; 70% of the gene sets targeted individually by conserved microRNAs and two sets of combinatorial target predictions for microRNA clusters contained at least one over-represented GO term (fig. 2.11a). For the "Molecular Function" ontology, a total of 25 significantly over-represented GO categories were obtained among 36% of all individual microRNA target gene sets and one set of microRNA cluster targets (fig. 2.11b). Consistent with previous estimates [2, 10], our data indicate that microRNAs regulate a large variety of genes in many different biological processes. Globally prominent GO terms were morphogenesis, organogenesis, development (including embryonic development, anterior/posterior and dorsal/ventral axis specification), neurogenesis, signal transduction (including *Notch*, *Torso*, *Sevenless*, and Frizzled signaling), and transcriptional regulation. Our overall overlap with another GO analysis for fly microRNA targets in a recent study was marginal, very likely due not only to the differences in approaches for identifying over-represented GO terms, but also to the different nature of target site predictions made by PicTar and the published miRanda algorithm [43].

Our data were consistent with and extended results from a recent study that used GO functional analysis to predict microRNA target genes [132], in which *miR-7* was predicted to be active in *Notch* signaling and *miR-277* in *valine*, *leucine*, and *isoleucine* degradation. For *miR-277*, we recovered all nine predicted targets and found five additional genes (CG3267, CG4389, CG4600, CG6638, CG8778) at a p-value $< 10^{-7}$. Targets of *miR-7* predicted by PicTar included many *Notch* pathway genes as well as targets of *Notch* signaling, including *E(spl)m5*, *Tom*, *Bob*, *E(spl)m*γ, *Bearded*, *E(spl)m3*, and *E(spl)m4*, most of which were very high scoring (using setting S1). Furthermore, many targets of *Notch* signaling were also predicted as targets of the *Bearded*-box microRNAs *miR-4* and *miR-79* (*E(spl)m5*, *Bearded*, *E(spl)m*γ, *Tom*) and of the K-box microRNAs *miR-2* and *miR-11* (*E(spl)m5*, *E(spl)m2*, *E(spl)m*δ, *E(spl)m3*), consistent with previous observations [77]. Other known *Notch* targets would have been included in PicTar's target lists if their 3'UTRs were annotated in the current FlyBase release (data not shown). We note that the majority of *Notch* targets predicted by PicTar would not have been predicted if stringent free energy filtering was applied for predicted microRNA:target duplexes with


dme−bantam
dme−miR−3
dme−miR−318
dme−miR−iab−4−5p
cluster 5
dme−miR−263b
dme−miR−279
dme−miR−286
dme−miR−283
dme−miR−289
dme−miR−316
dme−miR−14
dme−miR−309
dme−miR−124
dme−miR−79
dme−miR−285
dme−miR−iab−4−3p
dme−miR−287
dme−miR−311
dme−miR−92a
dme−miR−92b
dme−miR−304
dme−miR−184
dme−miR−281
cluster 7
dme−miR−315
dme−miR−31a
dme−miR−31b
dme−miR−9a
dme−miR−9b
dme−miR−9c
dme−miR−263a
dme−miR−34
dme−miR−133
dme−miR−7
dme−miR−13a
dme−miR−13b
dme−miR−2a
dme−miR−2b
dme−miR−2c
dme−miR−6
dme−miR−1
dme−miR−317
dme−miR−281−1*
dme−miR−281−2*
dme−miR−12
dme−miR−219
dme−miR−314
dme−miR−210
dme−miR−277
transcription initiation from RNA polymerase II promoter
regulation of photoreceptor differentiation (sensu Endopterygota)
regulation of photoreceptor cell differentiation
locomotion
chromosome organization and biogenesis
cell−cell signaling
adult locomotory behavior
organic acid metabolism
locomotory behavior
jump response
flight behavior
fatty acid oxidation
fatty acid metabolism
carboxylic acid metabolism
branched chain family amino acid metabolism
behavior
adult behavior
defense response
cell−matrix adhesion
cell−cell adhesion
negative regulation of biological process
Notch signaling pathway
transcription from RNA polymerase II promoter
transcription, DNA−dependent
transcription
regulation of transcription, DNA−dependent
regulation of transcription
regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolism
regulation of metabolism
wing morphogenesis
tracheal system development (sensu Insecta)
segment specification
retinal programmed cell death (sensu Endopterygota)
regulation of transcription from RNA polymerase II promoter
regulation of physiological process
nucleobase, nucleoside, nucleotide and nucleic acid metabolism
metamorphosis (sensu Insecta)
metamorphosis
limb morphogenesis
leg morphogenesis
leg disc development
larval or pupal development (sensu Insecta)
imaginal disc metamorphosis
imaginal disc development
genital disc metamorphosis
genital disc development
eye morphogenesis (sensu Endopterygota)
eye morphogenesis
eye−antennal disc metamorphosis
eye−antennal disc development
appendage morphogenesis (sensu Endopterygota)
appendage morphogenesis
apoptotic program
antennal morphogenesis
regulation of tracheal tube size
heart development
establishment of blood−nerve barrier
dopamine receptor signaling pathway
vesicle−mediated transport
synaptic vesicle transport
synaptic vesicle targeting
protein amino acid glycosylation
intracellular protein transport
endocytosis
female gamete generation
torso signaling pathway
sevenless signaling pathway
dorsal/ventral axis determination, follicular epithelium (sensu Insecta)
negative regulation of transcription
lipid metabolism
neurogenesis
ectoderm development
response to stimulus
regulation of cell size
regulation of biological process
positive regulation of cell growth
muscle development
embryonic morphogenesis
dorsal closure
MAPKKK cascade
peripheral nervous system development
periodic partitioning by pair rule gene
organogenesis
organ development
neurite morphogenesis
frizzled signaling pathway
central nervous system development
histogenesis
cell communication
cell surface receptor linked signal transduction
signal transduction
calcium−mediated signaling
specification of segmental identity, abdomen
embryonic heart tube development
embryonic heart tube anterior/posterior pattern formation
cardioblast differentiation
cardiac cell differentiation
anterior/posterior axis specification
morphogenesis
specification of organ identity
midgut development
mesoderm development
larval visceral muscle development
development
axonogenesis
transmission of nerve impulse
serotonin receptor signaling pathway
neurophysiological process
cell motility
axon guidance
germ−band shortening
amnioserosa maintenance
Color mapping to the negative natural logarithm of p−values
2.32
2.60
2.87
3.15
3.42
3.70
3.98
4.25
4.53
4.80
5.08
5.36
5.63
5.91
6.18
6.46
6.73
7.01
7.29
7.56
7.84
8.11
8.39
8.67
8.94
9.22
9.49
9.77
10.05
10.32
10.40
10.87
11.15
11.43
11.70
11.98
12.25
12.53
12.80
13.08
13.36
13.63
13.91
14.18
14.46
14.74
15.01
15.29
15.56
15.84

a



b

Figure 2.11: **Significant GO-terms among the predicted target genes of all single microRNAs and clusters of co-expressed microRNAs.**
Significantly enriched GO terms for (a) "Biological Processes" and (b) "Molecular Function" ontologies. Shown are GO terms with p-values smaller than 0.1, corrected for multiple testing. Hierarchical clustering was performed separately for GO terms and microRNAs.

perfect nuclei.

Comparison of microRNA targets between flies and vertebrates

Previously, we applied PicTar to exhaustively search 3'UTR alignments of eight vertebrates (human, chimpanzee, mouse, rat, dog, chicken, pufferfish and zebrafish) for microRNA target sites [68]. To compare the extent of microRNA targeting in flies and vertebrates, we first compared length, repeat content, and conservation of 3'UTRs between both clades, using our datasets derived from the UCSC database for consistency. We focused

on the comparison of 3'UTRs between *D. melanogaster* and human since 3'UTRs from these species were extracted based on annotated transcripts. We found that the length distribution of 3'UTRs and the distribution of repeats within them are very similar between all mammals and between all flies, respectively, so comparisons between human and *D. melanogaster* UTRs should reveal essential differences between the two clades. We found a much broader distribution of 3'UTR lengths in mammals compared to flies, yielding on average ∼900 nucleotides per 3'UTR for human and ∼400 nucleotides per 3'UTR in *D. melanogaster* (fig. 2.12), consistent with previous results [92]. Examining the contribution

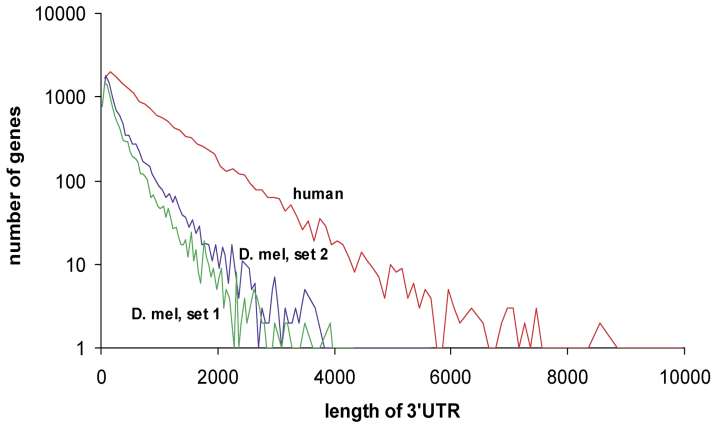


Figure 2.12: **Lengths distribution of 3'UTRs in human and** *D. melanogaster* **(for set 1 and set 2, logarithmic scale).**
The distribution decays exponentially with increasing length, in human much faster than in *D. melanogaster*. The average 3'UTR length in human (*D. melanogaster*) is ∼900 (∼400) nucleotides.

of repeat elements, we found that repeats constitute 11% of all human 3'UTR sequences compared to 4% in *D. melanogaster* (Table 2.6a). Interestingly, for short repeats (up to ∼50 nucleotides), the length distribution in *D. melanogaster* and human is similar (fig. 2.13). For longer elements the distribution in flies continues to decay exponentially with the same slope, whereas the human distribution displays a broad tail with another significant peak centered around ∼300 nucleotides. To analyze 3'UTR conservation, we counted all 7mers that appear to be perfectly conserved in each 3'UTR multiple alignment and divided these counts by the length of the 3'UTR sequence. We found that the probability of a nucleotide to reside in a conserved 7mer is comparable between vertebrate alignments (including human, chimp, mouse, rat, dog, and chicken) and alignments covering all fly species in our dataset (0.02 and 0.03, respectively). Similarly, 3'UTR conservation is comparable between mammals and flies in the *melanogaster* and *obscura* groups (0.06 and 0.08, respectively). The contribution of repeat elements to conserved 7mers is substantially different in vertebrates and flies (Table 2.6b). Masking repeats reduced the number of bases in conserved 7mers by ∼1% in vertebrates and ∼10% in flies, respectively. Thus, repeats in 3'UTRs appear to be much better conserved in flies than in vertebrates and thus may be of functional importance in flies.

The extent of microRNA regulation seems roughly comparable between mammals and flies overall, with several interesting clade-specific differences. In vertebrates, we and

a

Dataset	Genome-Wide Number of Nucleotides		
	Human	*D. melanogaster*, Set 1	*D. melanogaster*, Set 2
Unmasked repeats	16,311,781	3,906,057	3,681,969
Masked repeats	14,575,934	3,761,764	3,190,257
Percent difference	11%	4%	13%

Fraction of repeats in the 3′ UTRs of human and *D. melanogaster*.
DOI: 10.1371/journal.pcbi.0010013.t005

b

Level of Conservation	Number of Nucleotides in Conserved 7mers for a Given Level of Conservation					
	Mammals + Chicken	*D. mojavensis*, Set 1	*D. mojavensis*, Set 2	Mammals	*D. pseudoobscura*, Set 1	*D. pseudoobscura*, Set 2
with repeats	265,828	100,140	75,908	1,014,989	306,700	234,165
without repeats	263,990	85,956	48,559	1,004,870	277,586	162,227
%-difference	1%	14%	36%	1%	10%	31%

Fraction of nucleotides residing in 7mers conserved in all flies up to the indicated one (referring to the order *D. melanogaster, D. yakuba, D. erecta, D. ananassae, D. pseudoobscura, D. virilis, D. mojavensis*) and in vertebrates, with and without the inclusion of repeat elements. Comparison of Table 4 and 5 demonstrates that in vertebrates (flies), repeat elements share less (more) nucleotides than expected with conserved 7mers.
DOI: 10.1371/journal.pcbi.0010013.t006

Table 2.6: **Repeat elements in 3'UTRs of human and** *D.melanogaster* **and their conservation.**
(a) Fraction of repeats in the 3'UTRs of human and *D. melanogaster*. (b) Fraction of nucleotides residing in 7mers conserved in all flies up to the indicated one (referring to the order *D. melanogaster*, *D. yakuba*, *D. erecta*, *D. ananassae*, *D. pseudoobscura*, *D. virilis*, *D. mojavensis*) with and without the inclusion of repeat elements. Comparison of Table (a) and (b) demonstrates that in vertebrates (flies) repeat elements share less (more) nucleotides than expected with conserved 7mers.

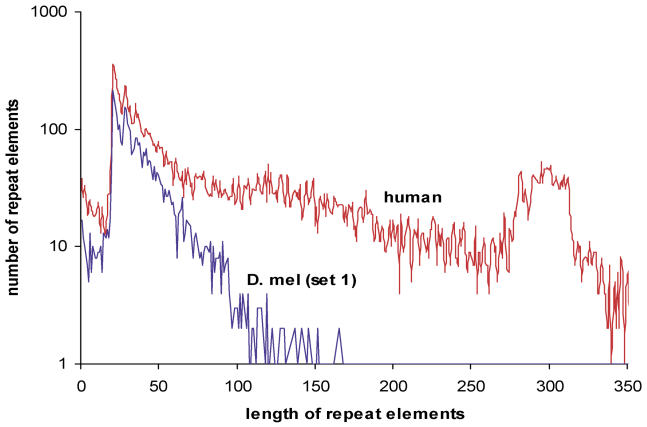


Figure 2.13: **Length distribution of repeat elements in 3'UTRs of human and** *D. melanogaster* **(set 1, logarithmic scale).**
At a length of 11 nucleotides, the distribution is strongly peaked for both species and decays exponentially for longer repeat elements in *D. melanogaster*. Up to a length of roughly 50 nucleotides, both distributions are very similar, while for longer elements the distribution for human no longer decays exponentially, but has a broad tail with another significant peak at a length of approximately 300 nucleotides.

others [59, 83] found that roughly 30% of all genes may be regulated by microRNAs. This is twice the number we found in flies (15%), but this could be explained by the smaller number of known microRNAs in flies (Discussion). More interestingly, we checked if individual microRNAs appear to target similar or significantly different numbers of genes in mammals compared to flies, since such differences could be indicative of clade-specific changes in microRNA function. To retain a reasonable sensitivity in target predictions for this analysis, we used human, chimp, mouse, rat, and dog for target predictions in mammals and the *melanogaster* and *obscura* groups for predictions in flies. We first defined sets of homologous microRNAs as well as homologous coding genes in mammals and flies. For microRNAs we applied a relaxed definition of homology according to the experimental insight that the nucleus is the crucial instance of regulation. Whenever the first or second 7mer of a microRNA in *Drosophila* is also present as one of the nuclei in a human microRNA, these two microRNAs are assumed to be homologs. Comparing all microRNAs conserved in the *melanogaster* and *obscura* groups with all microRNAs conserved in mammals, we obtained 47 pairs of homologous microRNAs between mammals and flies. Homologous genes between *D. melanogaster* and human were extracted from HomoloGene [142] (*ftp://ftp.ncbi.nih.gov/pub/HomoloGene/current/*) with annotations of 3/14/2005. This list contained 19,685 human genes and 7,983 fly genes. Keeping only pairs of homologous genes for which we were able to assign a FlyBase CG-number and a RefSeq gene identifier [106], respectively, our final list contained 4,623 pairs of homologous genes. We then set out to compute the average number of microRNA targets in both clades and finally calculated the ratio of predicted targets per microRNA to the average separately for each clade (Table 2.7).

A scatter plot of these ratios (fig. 2.14) demonstrates a correlation between the numbers of targeted genes for homologous microRNAs in mammals and flies. However, certain

D. melanogaster microRNA	Number of Putative Target Genes	Relative Number of Putative Targets[a]	Human microRNA	Number of Putative Target Genes	Relative Number of Putative Targets[b]	Ratio of Relative Numbers of Targets in Mammals and Flies
dme-miR-9c	309	1.90	*hsa-miR-9*	829	2.25	1.18
dme-miR-9b	313	1.92	*hsa-miR-9*	829	2.25	1.17
dme-miR-9a	310	1.91	*hsa-miR-9*	829	2.25	1.18
dme-miR-124	221	1.36	*hsa-miR-124a*	787	2.14	1.57
dme-miR-263b	175	1.08	*hsa-miR-96*	735	2.00	1.85
dme-miR-285	54	0.33	*hsa-miR-29c*	684	1.86	5.64
dme-miR-285	54	0.33	*hsa-miR-29b*	684	1.86	5.64
dme-let-7	79	0.49	*hsa-miR-98*	602	1.64	3.35
dme-let-7	79	0.49	*hsa-let-7i*	602	1.64	3.35
dme-let-7	79	0.49	*hsa-let-7g*	602	1.64	3.35
dme-let-7	79	0.49	*hsa-let-7f*	602	1.64	3.35
dme-let-7	79	0.49	*hsa-let-7e*	602	1.64	3.35
dme-let-7	79	0.49	*hsa-let-7c*	602	1.64	3.35
dme-let-7	79	0.49	*hsa-let-7b*	602	1.64	3.35
dme-let-7	79	0.49	*hsa-let-7a*	602	1.64	3.35
dme-miR-92b	209	1.29	*hsa-miR-32*	584	1.59	1.23
dme-miR-92a	223	1.37	*hsa-miR-32*	584	1.59	1.16
dme-miR-1	274	1.68	*hsa-miR-1*	535	1.45	0.86
dme-miR-1	274	1.68	*hsa-miR-206*	531	1.44	0.86
dme-miR-125	27	0.17	*hsa-miR-125b*	531	1.44	8.47
dme-miR-125	27	0.17	*hsa-miR-125a*	531	1.44	8.47
dme-miR-79	297	1.83	*hsa-miR9**	508	1.38	0.75
dme-miR-4	336	2.07	*hsa-miR-9**	508	1.38	0.67
dme-let-7	79	0.49	*hsa-let-7d*	468	1.27	2.59
dme-miR-92b	209	1.29	*hsa-miR-367*	440	1.20	0.93
dme-miR-92a	223	1.37	*hsa-miR-367*	440	1.20	0.88
dme-miR-34	142	0.87	*hsa-miR-34c*	439	1.19	1.37
dme-miR-34	142	0.87	*hsa-miR-34b*	439	1.19	1.37
dme-miR-34	142	0.87	*hsa-miR-34a*	422	1.15	1.32
dme-miR-133	32	0.20	*hsa-miR-133b*	400	1.09	5.45
dme-miR-92b	209	1.29	*hsa-miR-92*	389	1.06	0.82
dme-miR-92a	223	1.37	*hsa-miR-92*	389	1.06	0.77
dme-miR-92b	209	1.29	*hsa-miR-25*	380	1.03	0.80
dme-miR-92a	223	1.37	*hsa-miR-25*	380	1.03	0.75
dme-miR-133	32	0.20	*hsa-miR-133a*	365	0.99	4.95
dme-miR-7	116	0.71	*hsa-miR-7*	330	0.90	1.27
dme-miR-285	54	0.33	*hsa-miR-29a*	326	0.89	2.70
dme-miR-219	103	0.63	*hsa-miR-219*	226	0.61	0.97
dme-miR-31b	95	0.58	*hsa-miR-31*	198	0.54	0.93
dme-miR-31a	95	0.58	*hsa-miR-31*	198	0.54	0.93
dme-miR-10	17	0.10	*hsa-miR-10b*	181	0.49	4.90
dme-miR-10	17	0.10	*hsa-miR-10a*	181	0.49	4.90
dme-miR-304	166	1.02	*hsa-miR-216*	119	0.32	0.31
dme-miR-100	16	0.10	*hsa-miR-99b*	40	0.11	1.10
dme-miR-100	16	0.10	*hsa-miR-99a*	41	0.11	1.10
dme-miR-100	16	0.10	*hsa-miR-100*	41	0.11	1.10
dme-miR-184	60	0.37	*hsa-miR-184*	17	0.05	0.14
dme-miR-210	134	0.82	*hsa-miR-210*	15	0.04	0.05

The ratio of the number of target genes for a particular microRNA to the number of target genes averaged over all microRNAs is indicated for flies and for vertebrates (termed relative abundances). The ratio of the relative abundances between flies and mammals is plotted in Figure 8.

[a]In *melanogaster* and *obscura*, in units of the average number of targeted genes per microRNA.

[b]In mammals, in units of the average number of targeted genes per microRNA.

DOI: 10.1371/journal.pcbi.0010013.t007

Table 2.7: **Homologous microRNAs between mammals and flies of the *melanogaster* and *obscura* groups and their respective number of target genes.** The ratio of the number of target genes for a particular microRNA to the number of target genes averaged over all microRNAs is indicated for flies and for vertebrates (termed relative abundances). The ratio of the relative abundances between flies and mammals is plotted in fig. 2.14.

microRNAs appear to have a significantly higher number of target genes in humans (*miR-10*, *miR-133*, *miR-125*, *let-7*, *miR-285*) or in flies (*miR-184*, *miR-210*). For example, for *let-7* we found 1.64 as many target genes as expected on average in mammals, but only around 50% of the average expected number in flies. It is impossible to determine from this analysis if microRNAs have acquired more targets in one clade or lost targets in the other, but it is striking that both human homologs of the fly microRNAs *miR-184* and *miR-210* are expressed at low abundance across many human tissues, while the homologs of *miR-10*, *miR-133*, *miR-125*, *let-7*, and *miR-285* are expressed overall at much higher levels [11]. We stress that the human homologs of *miR-10* and *miR-133* have average or below average numbers of predicted targets in human. Our data indicate that the above seven microRNAs may function in clade-specific modes of gene regulation.

Finally, we computed which regulatory microRNA:mRNA relationships seemed to be con-

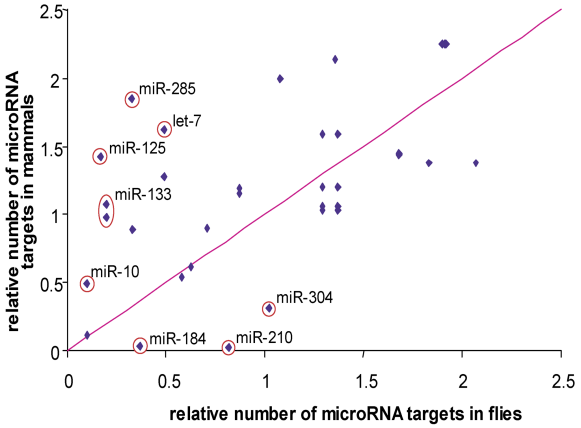


Figure 2.14: **Number of predicted target genes for homologous microRNAs between mammals and flies.**
Scatter plot for relative numbers of targeted genes predicted for homologous microRNAs in mammals and flies. The ratio of the number of predicted target genes of a microRNA and the average number of putative targeted genes per microRNA are plotted in mammals (y-axis) versus flies (x-axis). Conservation in flies included the *melanogaster* and *obscura* groups. Outliers (with a ratio of relative numbers of predicted target genes larger than 3 or smaller then 0.33) are circled. The microRNA identifiers refer to microRNAs annotated in *D. melanogaster*.

served between flies and mammals. From all 4,623 homologous human-*D. melanogaster* gene pairs in our dataset, 34 gene pairs were predicted to be targeted by homologous microRNAs. To asses the significance of the number of conserved microRNA-target relations of homologous target genes and microRNAs between vertebrates and flies, we shuffled homology relations in vertebrates and flies in the following way: All non homologous genes and microRNAs were discarded from our table with microRNA/target gene assignments. All microRNAs of a given family with equal 7mers at the 5' end were represented by one specific member of this family. Similarly, we discarded multiple transcript variants, keeping only the longest variant for each gene. We constructed a list with assignments of each microRNA to all its target genes. Shuffling was performed by permuting the microRNA entries of this list, thereby assigning a new set of target

genes to each microRNA. We counted the number of homology relationships for these permuted microRNA-target assignments and averaged the results over 100 runs. We obtained on average 24($\pm$6) homology relationships for the shuffled lists, while we counted 37 real homology relationships, when only using unique lists of genes and microRNAs. The described shuffling strategy models a situation of non conserved microRNA-target relations, but keeps the number of microRNAs targeting a particular gene constant. Thus, although the numbers are small, the conservation of microRNA/target gene relationships appears to be significant (Z=2).

Discussion

The extent of post-transcriptional gene regulation in *Drosophila* mediated by microRNAs. The sequencing of the genomes of several *Drosophila* species proved to be an invaluable resource for the analysis of microRNA targets in flies. Cross-species comparisons allowed us to arrive at significantly enhanced sensitivity and specificity for microRNA target predictions in comparison with recent approaches. For example, previous studies have predicted on average 8 target genes per microRNA (see [20] and references therein), while our data allow us (with high sensitivity settings S1) to predict 54 target genes per microRNA above noise in *D. melanogaster*. Requiring conservation in all flies (settings S2 and S3), we still predict on average more than 23 and 30 target genes per microRNA, respectively, at a strongly enhanced specificity.

Based on our target predictions, we found that currently known microRNAs are expected to regulate a large fraction of all *D. melanogaster* genes (15%). This number is almost certainly an underestimate, since (a) the annotation of 3'UTRs is incomplete, and (b) it is expected that many more microRNAs in fly remain to be discovered. Indeed, using an approach analogous to that of a recent comparative study of mammals [145], we analyzed fly 3'UTRs across all seven species and found strong evidence for the existence of a substantial number of yet undiscovered fly microRNA genes (N. Rajewsky, unpublished results).

The number of targets per microRNA we predicted is consistent with recent estimates of the true number of microRNA targets by Brennecke et al. [20]. In that study, the authors analyzed the statistical significance of conserved 8mer nuclei (starting at position one at the 5' end in the mature microRNAs) and conserved 7mer nuclei (starting at position two) and concluded that the vast majority of computationally detectable target sites possesses at least one conserved 7mer nucleus according to their definition. Our method is similar to this approach, but differs in the definition of the nucleus and the number of species included in the conservation analysis. Regardless of these differences, similar numbers of target genes per microRNA were predicted by both methods.

The highest scoring gene from all single microRNA target site predictions was *nerfin-1*, with two anchor sites for *miR-286* conserved in all flies and many additional, non-aligned sites present in all flies. Errors or ambiguities in the alignment can oftentimes explain the presence of these "scattered" sites. Additionally, compensatory mutations could lead to non-aligned and yet functionally conserved target sites in a 3'UTR. At present, PicTar scores these scattered sites in the same way as it scores conserved sites, as long as both of them occur in the same UTR. Future refinements of the algorithm should explore (a) explicit evolutionary models for the evolution of 3'UTR sequences and microRNA

target sites, (b) improved probabilistic scoring for sites with imperfect nuclei [20], (c) the incorporation of secondary structure information [114], (d) incorporation of mRNA expression levels (for example from microarray experiments), and (e) expression levels of microRNAs.
Our data indicated that clustered microRNAs are likely to coordinately regulate target genes. In addition, it has been shown that clustered microRNAs are likely to be co-expressed. Using multiple co-expressed microRNAs to coordinately regulate target genes could be an efficient way to increase the specificity of target gene regulation, and may also enhance the robustness of target gene expression levels against fluctuations in individual microRNA concentrations. We note that our data only suggest that clustered microRNAs are more likely to coordinately regulate target genes by coordinate binding to their 3'UTRs than non-clustered microRNAs. Many microRNAs that reside in clusters also seem to target genes without additional binding sites for microRNAs in the same cluster. Conversely, there appear to be many possibilities for microRNAs from different clusters to coordinately bind the same target genes.

The evolution of microRNA function across large evolutionary distance. microRNAs offer the exciting possibility to study the evolution of *trans*-acting regulatory genes together with the evolution of their *cis*-regulatory target sites using computational methods. In this study, we have only touched upon this problem by comparing the estimated number of targeted genes per microRNA in one clade to the predicted number of targets for the homologous microRNA which, by our definition of homology, is likely to bind to the same *cis*-regulatory sites. We caution that our definition of homology would also refer to microRNAs which may have evolved independently in one or both clades. However, our comparison yielded a non-trivial correlation between the numbers of targeted genes per microRNA in flies and vertebrates, indicating that the relative number of microRNA targets per microRNA tends to be conserved over very large evolutionary distances. In contrast, only a relatively modest number of specific microRNA:mRNA regulatory relationships seemed to be conserved between both clades. This scenario could be explained either by a high degree of combinatorial control mediated by microRNAs since binding sites could then be subject to compensatory mutations, or by frequently redundant functional roles of microRNAs.
It was striking that some microRNAs (including *let-7*) that are likely to have a large number of target genes in vertebrates seem to have a strongly reduced relative number of targets in flies, and vice versa. We singled out three microRNAs (*miR-184*, *miR-304*, and *miR-210*) with a drastically enhanced relative number of targets in flies compared to vertebrates. Our GO term analysis for microRNA targets revealed that one them (*miR-210*) had over 70 predicted target genes, which as a group were significantly enriched ($p < 0.03$ after correcting for multiple testing) for 11 genes with the GO annotation "female gamete generation" (fig. 2.11a). These 11 predicted *miR-210* targets are *cut*, *egghead*, *germ cell-less*, *gurken*, *lozenge*, *par-1*, *Ras oncogene at 85D*, *rhomboid-4*, *RNA-binding protein 9*, *singed*, and *slalom*. Most of these genes are evolutionarily conserved and have a known role in *Drosophila* oogenesis, either in development and patterning of the oocyte or in differentiation of the somatic follicle cells that surround the developing egg chamber, and seven of the 11 are implicated in developmentally critical signaling pathways involving *receptor tyrosine kinases*, *Notch*, *wingless*, or *hedgehog*. Development

of a mature *Drosophila* oocyte involves an elaborate sequence of events that must be precisely orchestrated in time. A surprising number of the genes in the above list play roles in important events that must take place within a specific window of time during oogenesis, many of which involve signaling between the germline and soma. Thus an important emergent theme of microRNA regulation may revolve around the widespread need for precise control of spatio-temporally restricted events during development. In addition, oogenesis in *Drosophila* occurs through a very different developmental program than in vertebrates. It is thus intriguing that a single microRNA has potentially evolved to include a wide array of target genes that are important for this developmentally divergent process. However, many of these potential targets are not restricted to oogenesis but also function at other times and places, including the eye, nervous system, and epithelia, and a number of other predicted *miR-210* targets also function in these tissues (e.g. *arrowhead*, *cacophony*, *trio*, *Sema-1b*, *makorin*, *Van Gogh*, *Syntaxin 17*, *G-oα47A*, *RhoGAP92B*, *cul-2*, *Apc*, and *Scm*). Thus this microRNA may play more complex pleiotropic roles in developmental networks. We conclude that some microRNAs could be candidates for genes that mediate clade-specific differences in gene expression, and could play an important role in shaping the diversity of life.

2.3.3 Nematodes

We used the improved version of PicTar to predict a global map of conserved *C. elegans* microRNA/target interactions. Genomic sequences from three *Caenorhabditis* species (*C. elegans*, *C. briggsae*, and *C. remanei*) were used for this analysis.

Genome-wide multiple alignments of nematode sequences

Genome-wide alignments between *C. elegans* and *C. briggsae* are already available at *www.wormbase.org* and the UCSC database [38, 134, 30]. However, a third *Caenorhabditis* species, *C. remanei*, has been sequenced recently, and since the three species have roughly the same pairwise evolutionary distances to each other [66], we reasoned that including *C. remanei* in the alignments would substantially boost our power to reliably detect evolutionary conserved 3'UTR sequence elements.
Whole genome alignments of *C. elegans* (WormBase Release WS120, March 2004), *C. briggsae* (cb25.agp8 assembly, July 2002), and *C. remanei* (Washington University, St. Louis, December 2004 assembly) were produced by the group of Lior Pachter from the Universiy of California at Berkeley in a two-step approach that combines orthology mapping with sequence alignment [79]. In essence, we used two existing programs (MAVID and MERCATOR) to construct genome-wide multiple alignments between all three *Caenorhabditis* species. These alignments covered 74% of the *C. elegans* genome and almost 80% of all known and predicted exons in *C. elegans*. Roughly 90% of these covered exons are shared between all three species. 14,874 of our *C. briggsae* gene annotations had orthologs in *C. elegans*, very similar to a comparison of the *C. elegans* and *C. briggsae* genomes [134]. The resulting alignments allow us to compare the evolution of 14,874 *C. elegans* 3'UTRs across all three nematode species. *C. elegans* 3'UTR sequences were extracted as described in [68] and mapped to the multiple alignments. Of all 21,623 *C. elegans* 3'UTRs, 16,965 were aligned between all three species.

MicroRNA sequences

Target genes were predicted for all 117 *C. elegans* microRNAs retrieved from Rfam (Release 6.0) [48]. To compute statistics (such as signal-to-noise ratio) with PicTar, a reduced set of all 73 microRNAs with unique nuclei was produced, containing the 55 unique microRNAs and one representative for each of 18 families with shared sequence identity at the 5 end. For these 73 microRNAs, we produced a randomized microRNA as explained previously.

Genomewide single microRNA target site predictions

Emission probabilities of nuclei were estimated by dividing the number of occurrences of each nucleus in conserved anchor sites by the total number of occurrences in the 3'UTR of *C.elegans* (conserved and non-conserved occurrences). This ratio should reflect the probability that the site is functional. Notably, only four sites with *let-7* imperfect nuclei had a non-zero probability of being binding sites by this criterion, and three of these match the validated target sites in *lin-41* supporting our hypothesis of a strong correlation between conservation and functionality. The power of this conservation based approach is evident as *lin-41* ranks 12/57 unique *let-7* targets, a rank that would be unattainable in algorithms that assess perfect complementarity with the 5' end of the microRNA.
We ran PicTar using all *C. elegans* microRNAs annotated in Rfam Release 6.0 along with the orthologous 3'UTR alignments described above. To estimate the signal-to-noise ratio for target predictions, we selected all 73 microRNAs with a unique binding nucleus sequence, and recruited four cohorts of randomized microRNAs. The signal-to-noise ratio was then calculated as the number of all target genes of the 73 unique real microRNAs with a PicTar score higher than a given score threshold, divided by the number of predicted target genes averaged over four cohorts of randomized microRNAs (fig. 2.15). A moderate score cutoff of 1.5 resulted in roughly 35 predicted target genes per microRNA at a drastically improved signal-to-noise ratio of 7.5. Only 42 of 117 microRNAs are found to have at least one imperfect target site residing in an anchor site. Altogether, we estimated that at least 10% of all *C. elegans* genes may be regulated by at least one known microRNA.

PicTar *let-7* target predictions

We initially chose to test the PicTar predicted targets of the most-studied microRNA, *let-7*, whose spatiotemporal expression pattern is known [113, 61]. Since the *let-7* mutant phenotype is suppressed by decreased activity of some of its targets, including *lin-41*, this provides an additional assay for testing its predicted targets. Although not specifically trained to identify *let-7* binding sequences, PicTar predicts many of the previously defined *let-7* targets [79]. GO term analysis shows that *let-7* targets predicted by PicTar are enriched for transcription factors, heterochronic genes, and genes involved in ectoderm development [79], consistent with previous data [49].
The *C. elegans* genome contains three additional microRNAs that are homologous to *let-7*: *miR-48*, *miR-84*, and *miR-241*. As expected given 5' sequence identity of these microRNAs, 98% of predicted *let-7* targets are also shared by *let-7* homologs. However, PicTar predicts that *lin-41* is the target of *let-7* alone, consistent with the idea that

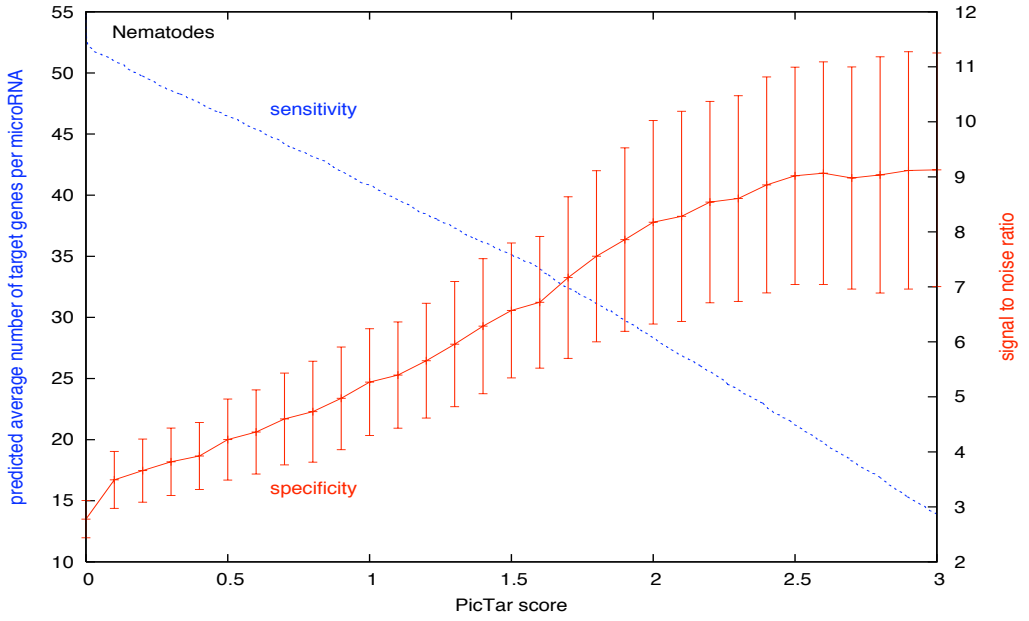


Figure 2.15: **Dependence of sensitivity and specificity on the PicTar score.** Shown is the average number of predicted target genes per microRNA (sensitivity, dashed line) and the signal-to-noise ratio (specificity, solid line) as a function of the PicTar score cutoff (i.e. all predicted targets with a score lower than this threshold are discarded). The signal-to-noise ratio was averaged over four cohorts of randomized microRNAs, and the standard deviation is indicated by error bars. The strong score dependence of the signal-to-noise ratio reflects the ability of PicTar to distinguish real sites from random sites. Top scoring predictions will therefore have a much higher signal-to-noise ratio than the average.

let-7 homologs do not act redundantly to repress *lin-41* translation, and suggesting that the improved PicTar algorithm can distinguish distinct targets of microRNAs that have highly homologous 5' binding nucleus sequences. Although cases of distinct targets for homologous microRNAs represent less than 5% of the predictions, algorithms that focus only on high complementarity with the most 5' sequence of the microRNA may not identify these. Recently, it has been shown that *let-7* homologs may have roles in larval development mediated by *hbl-1*, but not by *lin-41* [1], providing some indication that the PicTar prediction that *lin-41* is not a target of these microRNAs may be biologically relevant.

We tested all of the predicted *let-7* targets from various versions of PicTar for their ability to suppress the *let-7(n2853)* vulval bursting phenotype. In all, 74 predicted targets were tested, 57 of which are conserved across the three nematode genomes and are thus found by the latest PicTar version. Of the predicted targets tested, RNAi aimed at seven predicted *let-7* targets suppresses vulval bursting significantly ($p<0.05$ in a one tailed

t-test). As a negative control, RNAi tests of a control group of 44 non-lethal genes that are not predicted to be *let-7* targets found only one weak suppressor [79]. Novel suppressors of the *let-7* phenotype include F29G9.4, C27D6.4, C48A7.2, uba-1, K08F8.1, and K07A6.2, though suppression in all these cases is considerably weaker than that elicited by *lin-41*. C27D6.4 and F29G9.4 are predicted transcription factors, a result consistent with previously observed trends [49].

Global analysis of predicted microRNA target functions and evolution of microRNA targets

We used the predicted target dataset from PicTar to predict functions for each microRNA using GeneMerge v.1.2 [27]. Despite the diversity of regulation found upon analysis of particular microRNA targets, our global analysis of the predicted targets of the *C. elegans* microRNAs indicates that nearly half of *C. elegans* microRNAs regulate a set of genes enriched for particular functions. These functions are diverse, ranging from adhesion (*miR-233, 356, 251*) to ATP metabolism (*miR-1* and *256*). In general, enriched categories differ from the predicted functions of the *Drosophila* and human microRNAs [50]. For example, the *C. elegans miR-1* target set is enriched for functions involving ion transport and ATP metabolism, while *D. melanogaster miR-1* targets are enriched for a number of vesicle transport functional categories. In a systematic search for shared functions between homologous microRNAs in humans, *D. melanogaster* and *C. elegans*, we found no statistical enrichment for shared functional annotations above random noise (unpublished data). This suggests that there have been considerable changes in the functional categories targeted by related microRNAs, even though some microRNA sequences are highly conserved across phyla. Thus there seems to be selective pressure to maintain the sequence of some microRNAs across species, but changes in the nature of their target sets. This may be due to the potentially massive effect of altering a nucleotide in the 5' end of the microRNA; indeed a single substitution may subject a random set of targets to microRNA regulation. Alternatively, the sequence at the 5' end of the microRNA may be key to activating effector proteins involved in repression, reducing tolerance for alterations in this sequence.

A simple clustering algorithm that sorts microRNAs based on the enriched GO terms of their targets also tends to group the microRNAs by family (with some exceptions), as well as by expression pattern on developmental Northern blots. These rough correlations are consistent with the co-expression of certain microRNAs, but also suggest that families of *C. elegans* microRNAs tend to regulate similar suites of genes. This is evident from a simple perusal of microRNA target lists: 98% of *let-7* targets are also predicted targets of *let-7* homologs.

Although some individual microRNAs are predicted to regulate metabolic and developmental genes, the pooled targets of all microRNAs show such GO terms as under-represented categories. One intriguing explanation for this under-representation is that many microRNAs may regulate 'repressors' of metabolic and developmental gene function, such as transcription factors. In this case it would be counter-productive for microRNAs to also target metabolic genes directly. A major next step will be to analyze the functional categories suggested in our analysis experimentally, and such an analysis in multiple clades should shed light on the functions of microRNAs, as well as the evolution of their targets.

Integration of PicTar predictions with functional genomic data allows exploration of microRNA mediated regulatory networks

PicTar predictions are available from the NYU PicTar website (*http://pictar.bio.nyu.edu/*). The server currently provides access to predicted target gene sets for single microRNAs, listed in rank order based on their PicTar score, as well as links to the tri-nematode alignment and public databases containing information on the microRNA and gene (*http://microrna.sanger.ac.uk*, and *http://www.wormbase.org*, [30]). In addition a hyperlink is included pointing to a network view of all predicted microRNA/target gene links for the microRNA or target in question, generated using the Generic Network Browser at NYU (N-Browse, *http://gnetbrowse.org*; H-L Kao, F. Piano and K.C. Gunsalus, unpublished). These data are available for nematodes as well as vertebrates and flies. N-Browse is a dynamic, web-based graphical network browser that uses a client-server system to visualize network data residing in a remote database. The current version provides access to microRNA target predictions from PicTar and integrates them with combined functional genomics data from *C. elegans*, including protein-protein interactions, gene expression similarity, and phenotypic similarity data (as described [52]). N-Browse also provides information on gene attributes such as GO terms, RNAi phenotypes, and brief gene descriptions, and it offers links to external database resources such as WormBase [30], RNAiDB [51], and WormGenes (*http://www.wormgenes.org*; D. Thierry-Mieg, J. Thierry-Mieg and Y. Thierry-Mieg, M. Potdevin, M. Sienkiewicz, V. Simonyan, unpublished). Local gene neighborhoods in the network can be explored interactively by clicking on individual genes in the graph, which dynamically expands the current network around that gene to include all of its nearest neighbors and all their links to other genes in the current graph. An example of a network generated by querying for a subset of predicted *miR-256* targets is shown in fig. 2.16.
The network illustrates the interactions between genes involved in ATP metabolism, including multiple *C. elegans* homologs of the vacuolar ATPase subunits (*vha* genes). N-Browse illustrates that *miR-256* is predicted to regulate multiple components of this machinery, indicating that *miR-256* targets are enriched for genes involved in ATP metabolism. In addition both *miR-256* and *dpy-23* interact with a shared set of genes, a network whose visualization is facilitated by N-Browse. N-Browse therefore enables the analysis of integrated gene interaction networks that can be used to generate informed hypotheses regarding microRNA function.

Conclusions

Our analysis predicts that at least 10% of the transcripts in *C. elegans* contain conserved microRNA binding sites. This is likely an underestimate of microRNA regulation as it does not account for non-conserved microRNA target sites. Furthermore, many 3'UTR sequences may be missing from our set of 3'UTR sequences due to the lack of available full-length mRNA sequences and the imposed cutoff in 3'UTR length when predicting unannotated 3'UTR sequences. Using GO functional annotations, we find that roughly half of the known *C. elegans* microRNAs may regulate sets of genes enriched for particular functional categories. We have integrated our microRNA target predictions into an accessable network-browsing tool (N-Browse). Together our data therefore elucidate the functions of microRNAs in terms of the networks of interacting genes that they are

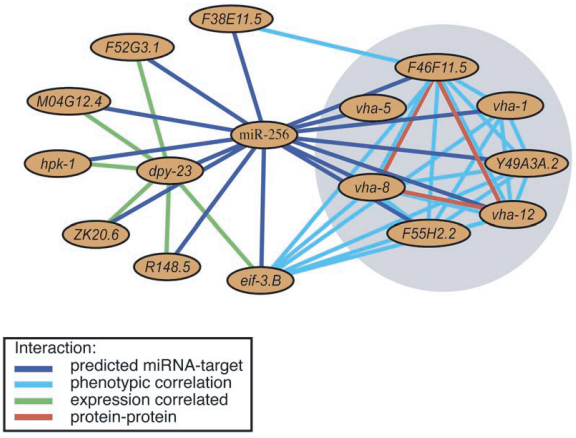


Figure 2.16: **N-Browse allows integration of predicted microRNA-target interactions with genomic datasets into networks.**
Shown is an example in which *miR-256* is seen to interact with multiple vacuolar ATPase subunits and other ATP metabolism genes (highlighted in grey circle). Multiple functional links between these targets, based on phenotypic correlation and protein-protein interactions, are also shown. *miR-256* and *dpy-23* both interact with a common set of genes in the graph, defining an integrated network that emerges facilitated by N-Browse. Thus N-Browse queries allow probing of interaction networks that can be used to generate informed hypotheses regarding microRNA biology.

predicted to regulate.

2.4 Outlook

Target prediction algorithms such as PicTar have largely extended the knowledge about post-transcriptional gene regulation imposed by microRNAs. Insights into the extent of microRNA dependent regulation across different clades of life as well as the observation of a statistically significant enrichment of a variety of biological processes with predicted microRNA target genes are among the main results of computational studies. However, the successful prediction of a large class of target genes should not draw off the attention from several caveats.
It is yet unknown, if modes of microRNA targeting other than the mechanisms discussed in this chapter exist. Although statistical overrepresentation of microRNA nuclei in animals has not yet been found in genomic sequence other than 3'UTRs, it is not clear if sequence categories of different functional annotations (such as genomic DNA, or 5'UTRs, or coding sequences) are targeted by yet unknown mechanisms. Furthermore, current target prediction algorithms still heavily rely on cross-species comparison, a criterion which is obviously unsuited for the identification of species-specific target genes. Although several attempts have been made to infer higher specificity from single target site data, e. g. by considering the secondary structure of the entire microRNA:mRNA duplex, these approaches have not improved the quality of single species target predictions. Another shortcoming is the insufficient understanding of the protein complex, which incorporates

the microRNA. Future experiments should elucidate the structural constraints on the microRNA:mRNA duplex imposed by this complex as well as its possible interaction with the target mRNA. For instance, it is still an open question how it is decided if the mRNA of a microRNA target gene is degraded and/or only translationally repressed.
Apart from specific features of the target site itself, combined analysis of expression data of the microRNA and its putative target genes could enhance the specificity of target predictions. Since large scale expression data become increasingly available, the next generation of target prediction algorithms should incorporate these data. Correlated expression data, such as upregulation of sets of genes upon knockdown of specific microRNAs are particularly helpful in this regard [71, 131].
In summary, although the last few years elucidated many aspects of microRNA biology, many questions have remained unanswered and considerable scientific effort is still required in order to fully understand microRNA dependent gene regulation.

Chapter 3

Evolutionary dynamics of microsatellites in *Drosophila*

3.1 Introduction

Microsatellites are tandem repeat sequences of motif length 1-6 bases that contribute a sizable fraction to eukaryotic genomes (reviewed in [85]). As opposed to non-repetitive sequence, insertion- and deletion-processes strongly dominate simple base substitutions in microsatellites producing substantial length variability at homologous loci (e.g.[123, 40, 141]). The main cause for this observation is thought to be strand slippage during DNA replication: the nascent strand dissociates and its terminal repeat unit erroneously realigns to the wrong unit of the template strand, leaving a loop confirmation. If the latter is missed by the DNA repair mfoachinery the newly synthesized strand is left with an altered number of repeat units [137, 118, 140].
A strong interest in understanding evolutionary dynamics of repeats emerges from their widespread use in phylogenetic reconstruction [53], as genetic markers to identify specific regions under natural selection [119], in population genetics [58], and because of the implication of certain classes of microsatellites in human cancers and genetic disorders (reviewed in [85]). Moreover, microsatellites seem to be involved in chromatin organization, DNA metabolism and regulation of gene activity (reviewed in [85]). Although microsatellites were claimed to evolve neutrally [120, 121], there are also indications for non-randomness. For instance, Bachtrog *et al.* (1999) [4] found that 39% of the analysed repetitive sequence in *Drosophila melanogaster* deviates from a random distribution. Furthermore, the identified function of several instances of simple tandem repeats suggests selective constraint at least on specific repeat sequences.
Compared to the prevailing evolutionary process that shapes non-repetitive sequence, which is single point mutations, microsatellites obey more complicated evolutionary dynamics. The genome-wide distribution of microsatellites in a species, if it can be assumed to be stationary, is the outcome of three different mutational mechanisms: replication slippage induced growth and decay and single point mutations, disrupting the ability of a tandem repeat unit to undergo slippage and thus also inducing repeat decay. Since microsatellites are of relatively short length, and their length distribution is observed to be decaying, degrading events that comprise backward slippage and point mutations outweigh slippage induced growth. Therefore, each microsatellite has only a

limited lifetime and will decay to a state of exclusively mutated repeat units that do not undergo further slippage. If stationarity holds, this extinction of microsatellites must be balanced by the emergence of new repeat elements. A single motif duplication, which can be considered a seed for a new microsatellite can emerge, for instance, by a misalignment after double stranded nicks or in case of short motifs also by single point mutations. The ensemble of microsatellites is thus a stationary non-equilibrium system that does not obey detailed balance.

Formal models of repeat evolution have been introduced in numerous studies. They all rest on the assumption of stationarity, i. e. the distributions of the current generation are supposed to be the same as those of the next generation. Another common feature is the modeling of repeat evolution as a single-stepwise process, which was first proposed by Jurka and Bell (1997) [13]. This model rests upon the idea that at each generation a tandem repeat element either grows or decays by one unit due to a single slippage event with specific backward- and forward-slippage rates. In addition to slippage, single base substitutions abrogate the ability of the affected repeat unit to undergo slippage, thus cutting the repeat element into shorter pieces. The most highly debated issue of the single-stepwise model is the magnitude of the slippage rates and their dependence on the length of the perfect repeat, i. e. number of adjacent repeat units. Moreover, a minimum number of repeat units required for slippage mutations to occur has been proposed [91] and a minimum length of 8 bp irrespective of the motif size has been suggested by a simple mathematical model [115]. However, the existence of such a threshold size is still controversial [108].

In several studies the single-stepwise time evolution of microsatellites was modeled by a Markov chain, following the evolution of a single repeat element consisting only of perfect adjacent repeat units [69, 70, 126]. Only the longer piece is retained if the perfect repeat sequence is cut into two pieces by a single point mutation. As an improvement, Lai et al. (2003) [78] modeled the decay process as a Markov chain multitype branching process, keeping track of the further evolution of both pieces after a point mutation event. In all studies, it was either assumed that slippage rates linearly increase with the repeat length [69, 70] or subsequently demonstrated that slippage rates grow with the number of repeat units [78] by fitting to data. Sibly et al. (2001) [126] developed a maximum-likelihood approach for assessing several competing hypotheses and also conclude that slippage rates are proportional to the repeat locus length. A simple and intuitive explanation for this behavior could be the possibility of slippage for each pair of adjacent repeat units. However, a minimum motif length is likely to be required because of the limited flexibility of the polymer. The linear dependence further outrules the possibility of multiple-unit-slippage, which would follow from realigning to a more distant locus and result in an exponential growth of the slippage rate with the number of repeat units. We conjecture that these higher order processes interfere with the single-stepwise slippage preferentially at short motif length, i. e. for mono- and dinculeotide repeats, which have been shown to obey more complicated dynamics [23].

Another degree of freedom is the ratio of forward- and backward-slippage rates. For the symmetric single-stepwise model that assumes the expansion rate to be the same as the contraction rate [69, 70, 126] it was shown that it cannot explain human sequence data. Hence, a mutual dependence of these rates should not be postulated a priori [78].

Although some more subtle length dependencies for particular repeat classes were sug-

gested by experiments [146, 5, 54], the overall slippage rates predicted by these models compared reasonably well to experimentally inferred rates suggesting that single-stepwise mutational events are the predominant driving force microsatellite evolution. However, the amount of experimental data is still limited and probably insufficient to test more subtle differences between competing evolutionary models. In this study, we try to overcome this problem by comparing the rates predicted from our model with rates independently measured from multiple alignments of closely related *Drosophila* species. We believe that such *in silico* experiments contribute a valuable source of information in the era of whole genome sequencing.

Using similar comparisons a careful examination of the existing models should be required, since all these models suffer the shortcoming to neglect the microsatellite structure. A point mutation inside of a microsatellite cuts the repeat element into two smaller ones. The spatial vicinity of these neighboring repeat elements, however, is not regarded by previous models. The length distribution of microsatellites, captured by these models, thus neglects the spatial correlation of neighboring repeat elements. Hence, it is arguable whether these models correctly predict slippage rates and repeat length distributions.

Here we introduce a new method for the description of microsatellite evolution that is based on a novel representation of tandem repeat elements, which are looked upon as entities comprising perfect as well as mutated instances of the repeat unit. Time-evolution of the whole object is encoded by a general master equation. Assuming stationarity of the repeat distribution, we derive two observables, which are easily measurable from genomic data and which allow for the inference of backward- and forward-slippage rates. This modeling scheme of repeat evolution seems to be superior to previous methods, since spatial correlations of "mutationally separated" neighboring perfect repeat elements of a common origin are not neglected any more. More precisely, a perfect repeat element is observed with increased likelihood in the neighborhood of another perfect repeat element. We compute the stationary distribution of tri-, tetra- and pentha-nucleotide microsatellites in *Drosophila melanogaster* across sequence classes of diverse functional annotations and provide strong evidence for the validity of our model in case of putatively neutral sequence by comparing to optimized three-way alignments of *D. melanogaster*, *D. simulans* and *D. yakuba* as well as to experimentally measured rates. We further demonstrate that deviations from the predicted distribution increase with the content of functional sequence. Hence, we present a new model that, opposed to all existing models, traces the decay of tandem repeats into smaller parts by base substitutions maintaining the mutated units and thus its representation as a single object. This obviously increases the amount of information drawn from the data. Moreover, we present the first whole genome analysis of microsatellite distribution in *D. melanogaster* resolved into sequence classes of different functional annotation and conduct a large scale comparison of our rate predictions to insertion and deletion rates measured from multiple alignments.

3.2 Sequence Data

3.2.1 *Drosophila* alignments

Sequence data were extracted from true genome-wide multiple alignments generated by the Pachter group at UC Berkeley [18] using the following assemblies: *D. melanogaster* Apr. 2004 (dm2), *D. yakuba* Apr. 2004 (droYak1), *D. simulans* Sept. 2004. We used FlyBase Release 4.1 to extract gene annotations in *D. melanogaster*.
A set of intergenic background sequences was constructed by extracting pairs of orthologous RefSeq genes that are neighboring and syntenic in all three fly species under consideration. To determine orthologous RefSeq genes in *D. simulans* and *D. yakuba* we used the "GeneMapper" program [29]. From these syntenic contigs, that contained the annotated genes plus the intergenic regions we removed all overlaps with RefSeq and FlyBase gene annotations together with a region of 20,000 (10,000) bases upstream (downstream) of each transcript. The remaining intergenic sequence covered $\sim$7 megabases (Mb).
We also used the syntenic contigs to extract 1 kilobase (kb) of sequence upstream of the transcription start site for 2,077 unique FlyBase genes, totaling 2 Mb.
Coding sequence, 3'UTRs and 5'UTRs were extracted for 13,423 unique FlyBase genes (keeping always only the longest transcript variant), summing up to 21 Mb, 4 Mb and 2 Mb, respectively.
Intronic sequence was extracted for 4,982 unique introns from coding regions of 2,905 unique FlyBase genes, covering altogether 12 Mb. For each intron, the first and the last 100 bases were cut off to discard conserved intron-exon boundaries containing splice sites.

3.2.2 Alignment Optimization

In order to detect insertions and deletions in a set of evolutionary related sequences at a high level of specificity and sensitivity, the alignment parameters need to be adjusted to their actual values matching the evolutionary process that generated the sequences under consideration. In practice, since the true evolutionary process that generated a given pair of sequences from a common ancestor is unknown, a simple model is postulated that is based on experimental evidence as well as statistical evaluation of the related sequences. We started out with the BLASTZ scoring matrix, a scoring scheme derived and evaluated by Chiaromonte et al. (2002) [31], which was shown to work well for human-mouse alignments [124]:

	A	C	G	T
A	91	-114	-31	-123
C	-114	100	-125	-31
G	-31	-125	-100	-114
T	-123	-31	-114	91

As initial values for the gap-opening cost s_o and the gap-extension cost s_e we chose $s_o = -750$ and $s_e = -25$, respectively. This choice was motivated by the scoring scheme of the multiple alignment tool Multi-LAGAN [21], which we used for the computation of

optimal alignments.
By variation of the gap-opening and gap-extension costs as well as the ratio of match- and mismatch-probability after initialization with the BLASTZ scoring matrix, we achieve the optimization of the full probabilistic alignment in order to obtain the most likely parameters to explain the data with the assumed model.
For the pairwise alignment of two sequences S_1 and S_2, the full partition sum as a function of the gap-opening cost s_o, the gap-extension cost s_e and the single-letter (mis-)match-scores $s(a_i, a_j)$ ($a_i \in \{A, C, T, G\}$) reads

$$Z = \sum_{paths} \prod_{pairs\, a_i \in S_1, a_j \in S_2} e^{s(a_i,a_j)/\tau} \prod_{gaps\, i\, of\, length\, l_i} e^{(s_o + s_e(l_i-1))/\tau} \prod_{a_i \in S_1} q(a_i) \prod_{a_j \in S_2} q(a_j) \tag{3.1}$$

Once an alignment "temperature" τ has been fixed, each score determines a probability of observing a specific letter pairing or gap, respectively. As shown in Yu and Hwa (2001) [148], these probabilities are constrained by a normalization condition. Given the scores $s(a_i, a_j)$ ($a_i \in \{A, C, T, G\}$) of all letter pairings and the probability $q(a_i)$ to observe letter a_i in a single sequence, this condition reads

$$\sum_{\{i=1,..,4;j,1,..,4\}} q(a_i) q(a_j) e^{s(a_i,a_j)/\tau} \left(1 + \frac{2e^{s_o/\tau}}{1 - e^{s_e/\tau}}\right) = 1. \tag{3.2}$$

It ensures that the probabilities of all possible configurations at any alignment grid point sum up to one. Using (3.2) and our initial choice of parameters, the alignment temperature τ can be computed numerically. We proceed to compute the alignment partition sum, divided by the partition sum for an uncorrelated gapless background model. By this procedure, all single letter emission probabilities cancel out and one obtains

$$\frac{Z}{Z_b} = \sum_{paths} \prod_{pairs\, a_i \in S_1, a_j \in S_2} e^{s(a_i,a_j)/\tau} \prod_{gaps\, i\, of\, length\, l_i} e^{(s_o + s_e(l_i-1))/\tau}. \tag{3.3}$$

To compute (3.3) recursively, we follow Yu and Hwa (2001) [148]. For a given pair of evolutionary related sequences, the logarithm of this quantity, i. e. the score gain relative to the uncorrelated model (see Appendix B) was numerically maximized as a function of the two gap cost parameters and the relative weight of matches and mismatches under the constraint (3.2). Introducing the variables

$$\begin{aligned}
\alpha &= \sum_i q(a_i)^2 e^{s(a_i,a_i)/\tau} \text{ (match-probability)} \\
\beta &= \sum_{i \neq j} q(a_i) q(a_j) e^{s(a_i,a_j)/\tau} \text{ (mismatch-probability)} \\
\gamma &= e^{s_o/\tau} \text{ (gap-opening probability)} \\
\delta &= e^{s_e/\tau} \text{ (gap-extension probability)}
\end{aligned} \tag{3.4}$$

equation (3.2) reads

$$(\alpha + \beta)\left(1 + \frac{2\gamma}{1 - \delta}\right) = 1. \tag{3.5}$$

Varying a score parameter corresponds to multiplicative variation of one of the indicated quantities. The optimization procedure is implemented as a stepwise process. First, the quantities α and β are varied keeping their sum and thus equation (3.5) constant. Second, the quantity γ is varied and simultaneous variation of $(\alpha + \beta)$ (i. e. all letter pairing scores simultaneously change by the same amount) ensures the condition (3.5). Similarly, the parameter δ is tuned. These three subsequent steps are repeated until the parameter values converge.
The optimization procedure was used to compute specific scoring matrices for Multi-LAGAN with the final goal to produce improved optimal alignments. To obtain good parameter estimates for the multiple alignment of *D. melanogaster*, *D. simulans* and *D. yakuba*, we optimized the sum of (3.3) over all possible pairings. Although the ingroup species *D. melanogaster* and *D. simulans* are more closely related to each other than to the outgroup species *D. yakuba*, this procedure seems to be a good working compromise, since the evolutionary distance of *D. yakuba* is still smaller than twice the evolutionary distance between *D. melanogaster* and *D. simulans*. Multi-LAGAN requires only a single scoring table and it is thus impossible to assign specific scoring matrices to different pairings from the set of species to be aligned.
To test the improvement of the optimized scoring versus the default (BLASTZ) parameters, we simulated sequences of relative evolutionary distances similar to those of *D. melanogaster*, *D. simulans* and *D. yakuba*. More precisely, the phylogenetic distance between the outgroup species and an ingroup species was chosen to be 1.5 times the distance between the ingroup species. We simulated a uniform mismatch-probability of 10% and uniform insertion/deletion rates of the same order of magnitude. Our parameter optimization procedure not only recovered these rates with low errors ($< 10\%$), but also recovered gaps with higher sensitivity when applied to compute the optimal path using Multi-Lagan: While only 80% of all gapped nucleotides are recovered when BLASTZ default parameters are used, we recover 94% after the parameter optimization.
In order to optimize alignment scores for our sets of genomic sequences, we randomly selected 20 alignments of 1,000 nucleotides from each set and ran the optimization procedure. We obtained the following modifications of the BLASTZ matrix with δMM (δM) denoting the additive difference to the BLASTZ match-(mismatch-)scores:

	s_o	s_e	δMM	δM
synt. back.	-540	-17	-160	27
upstream	-578	-11	-106	25
reg. modules	-578	-9	-159	27
3'UTRs	-606	-12	-194	31
5'UTRs	-558	-15	-158	31
CDS	-846	-11	-153	34
intronic	-554	-7	-124	26

Note that the alignment temperature in each case is $\tau \sim 100$ and thus a score difference of 65 changes the probability by a factor of 2.

3.2.3 Repeat Data

To identify repeats in the genomic sequence of *D. melanogaster*, *D. simulans* and *D. yakuba*, we use the Tandem Repeats Finder (TRF) [14], which models tandem repeats by percentage identity and indel frequency between adjacent pattern copies and uses parameter dependent statistical recognition criteria. The parameter settings for the TRF were chosen based on base substitution rates and indel rates of the alignment of the detected repeat sequence to the perfect run of repeat motifs, which is part of the TRF output.
With settings *match score 2, mismatch cost 5, indel cost 5, match probability 80, mismatch probability 10, minimal alignment score 15, maximal repeat length 500*, one obtains a mismatch probability of 4%, an insertion rate of 1.6% and a deletion rate of 1.5%. These values are higher than those obtained from the alignment of intergenic sequence of *D. simulans* and *D. melanogaster* but lower than those between the ingroup species and *D. yakuba*. Applying these settings we thus focus on repeat elements that on average emerged on the ingroup branch after the split of *D. yakuba* and the common ancestor of *D. melanogaster* and *D. simulans*. Although the minimal required alignment score of 15 is rather low, the TRF will miss the shortest instances of repeat elements with only two perfect adjacent motifs at short motif length $l = 1, 2, 3$. However, since we are interested in repeats with motif length $l \geq 3$ and more than just a single duplictaion this cutoff is appropriate for our purposes.

3.3 Analysis of Microsatellites

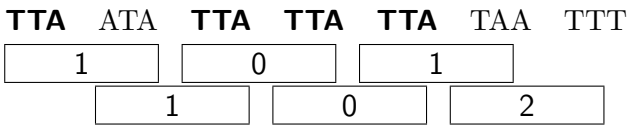
3.3.1 Statistical model of microsatellite evolution

Theory of repeat evolution

Elementary processes and rates. The "minimal" sequence evolution model used in this work contains three elementary molecular processes: (i) *Base substitutions* act on the sequence as a uniform Markov process with a substitution rate μ. (ii) *Repeat initiation* processes generate a new repeat from non-repetitive (background) sequence. The new repeat is a *perfect doublet*, i.e., two adjacent identical units of a sequence motif of length ℓ. Repeat initiation can take place by base substitutions but may also involve other processes, such as imperfect lesion repair after double-stranded nicks. Therefore, it is modeled as a Markov process with an independent rate $\gamma_i(\ell)$ per unit sequence length. (iii) *Slippage* processes act independently on each perfect doublet, producing a further identical unit with rate $\gamma_+(\ell)$ or reducing the original pair to a single unit with rate $\gamma_-(\ell)$. Thus, slippage is assumed to be suppressed whenever two subsequent repeat units are separated by other sequence elements or if they are an *imperfect doublet*, i.e., they are not identical as a result of preceding base substitutions in one of them. These constraints approximate the observed strong decrease of the slippage rate with the distance between the units and with the number of base substitutions between them [140]. Repeats containing at least one perfect doublet are called *active*, repeats with only imperfect doublets are *silent*, since they no longer undergo slippage.
This model contains the main distinguishing feature of repeat evolution, namely the

coupling between slippage and base substitutions and the resulting breakdown of time reversal symmetry. There are only four independent parameters, μ, γ_i, γ_+, and γ_-, which can be inferred efficiently from sequence analysis. Of course, the elementary processes of the minimal model are strong simplifications of the actual molecular evolution of repeats, which may also involve additional modes such as simultaneous slippage and base substitutions. Nevertheless, as will be demonstrated below, the minimal model provides a good statistical description of the observed repeat statistics at least in intergenic regions of the *Drosophila* genome, implying that it would be difficult to find statistical evidence for a more complicated model. Hence, the processes and the inferred parameters of the minimal model are to be regarded as efficient summaries of the full molecular machinery.

Mean-field approximation. A typical repeat is a sequence of perfect and mutated units of its sequence motif, as illustrated by the following example:



These units generate a sequence of perfect and imperfect doublets (marked by the number within the boxes denoting the Hamming distance to the perfect doublet). The total slippage rate of a repeat depends only on the number n_0 of perfect doublets. Most base substitutions change perfect into mutated units and, hence, perfect into imperfect doublets; however, the number of perfect doublets lost depends also on the positioning of the mutated units within the repeat. To capture the effect of base substitutions in an approximate way, we classify all doublets by their Hamming distance $d = 0, 1, 2, \ldots$ from a perfect doublet, i.e., by the total number of base substitutions away from the sequence motif in both its units (marked within the boxes). This is reminiscent of the notion of error classes for a molecular quasispecies [42], which are defined with the perfect doublet of the repeat motif as Master sequence. However, here the occupation numbers n_d count *fixed* sequence states of doublets within one repeat, while a quasispecies is a *polymorphic* population and n_d are the population numbers of coexisting sequence states (alleles) at the same genomic locus. In the following, we keep track of the doublet content $(n_0, n_1, \ldots)$ of a repeat, but the position of mutated units is taken into account only in an average way, decoupling their effects on individual doublets. This type of treatment is known in statistical physics as mean-field approximation. Within the framework of the model, single doublets are considered as uncorrelated entities evolving independently. In this approximation we neglect the effect of mutations occurring in single units shared by two overlapping doublets, e. g. instead of simultaneously mutating two perfect doublets within the overlapping unit at rate $\mu\ell$ in our model the overlapping unit is subject to two independent mutations at rate $\mu\ell$ each reducing the number of perfect doublets by one. However, a specific correlation has to be incorporated since simulation of the system has shown that it crucially affects the evolutionary dynamics: If a unit is shared by a perfect doublet and a doublet carrying a single base substitution, a base substitution within the shared unit changes the occupation number of error class zero while the occupation number of the first error class remains unchanged: the loss of a doublet in the first error class is compensated by the simultaneous creation of a novel doublet. With respect to the

dynamics in the (n_0, n_1)-plane this process can be summarized as $n_0 \to n_0 - 1 \wedge n_1 \to n_1$ and gives rise to an additional term in the Masterequation (3.6) and two constants ρ and ρ' reflecting the averaged probabilities to find configurations affected by this process. The parameter ρ determines the fraction of processes $n_0 \to n_0 - 1$ with $n_1 \to n_1$. With the definitions

$\rho \equiv \frac{P(0|00|1)}{P(00)} = \frac{P(1|00|0)}{P(00)}$: probability for a unit of a perfect doublet to be flanked by a unit carrying a single base substitution
p_1: probability for the process $n_0 \to n_0 - 1 \wedge n_1 \to n_1 + 1$
p_2: probability for the process $n_0 \to n_0 - 1 \wedge n_1 \to n_1$

we obtain the possible configurations with their respective probabilities:

i_0	i_1	i_2	i_3	p_1	p_2
0/>1	0	0	0/>1	$(1-\rho)^2 2\mu\ell$	0
1	0	0	1	0	$\rho^2 2\mu\ell$
1	0	0	0	$\rho(1-\rho)2\mu\ell$	$\rho(1-\rho)2\mu\ell$
			$\sum$	$2(1-\rho)\mu\ell$	$2\rho\mu\ell$

Similarly, with the definitions

$\rho' \equiv \frac{P(0|01)}{P(01)} = \frac{P(10|0)}{P(10)}$: probability for the perfect unit of a doublet with a single base substitution to be flanked by another perfect unit
p_1': probability for the process $n_1 \to n_1 - 1$

we can distinguish all possible configuration and obtain the likelihood that a mutation of a doublet drawn from the first error class results in a decrease of the occupation number of this error class:

i_0	i_1	i_2	p_1'
0	0	1	$\rho'\mu(\ell-1)$
>0	0	1	$(1-\rho')(2\ell-1)\mu$
		$\sum$	$((2\ell-1)-\rho'\ell)\mu$

The actual values of ρ and ρ' are determined directly from the repeat data. However, in the limit $\beta \to 0$ it is easy to verify that $\rho = \beta/2 + O(\beta^2)$ and $\rho' = 1 - O(\beta)$ which is supported by the numerical parameter values.

Master equation. We thus arrive at a closed evolution equation for the probabilities $W(n_0, n_1, \dots)$ per unit sequence length to find a repeat with doublet composition $(n_0, n_1, \dots)$ for a given motif length ℓ. Restricting the analysis to perfect and single-

mutation doublets, this takes the form

$$\begin{aligned}\partial_t W(n_0,n_1,t) = \\ & J_0(n_0-1,n_1,t) - J_0(n_0,n_1,t) \\ & +J_{01}(n_0+1,n_1-1,t) - J_{01}(n_0,n_1,t) \\ & +J_1(n_0,n_1+1,t) - J_1(n_0,n_1,t), \end{aligned} \tag{3.6}$$

where

$$\begin{aligned} J_0(n_0,n_1,t) &\equiv [\gamma_+ n_0 + \gamma_i \delta_{n_0,0}] W(n_0,n_1,t) \\ &\quad -\gamma_-(n_0+1)W(n_0+1,n_1,t) \\ &\quad -2\rho\mu\ell(n_0+1)W(n_0+1,n_1,t), \end{aligned} \tag{3.7}$$

$$J_{01}(n_0,n_1,t) \equiv 2(1-\rho)\mu\ell n_0 W(n_0,n_1,t), \tag{3.8}$$

$$J_1(n_0,n_1,t) \equiv ((2\ell-1)-\rho'\ell)\mu n_1 W(n_0,n_1,t) \tag{3.9}$$

are the probability currents due to repeat initiation and slippage, due to the first mutation of a perfect ($d=0$) doublet, and to the second mutation of a $(d=1)$-doublet, respectively, and we have suppressed the dependence of the rate constants on ℓ in the notation. Furthermore, we have neglected "backward" mutations, which reduce the Hamming distance d to a perfect doublet. The approximations underlying eq. (3.6) are justified for the intermediate motif lengths $\ell = 3,4,5$ studied in this work, for which we have verified the validity of the mean-field approximation by comparison with direct simulations of sequences evolving by the molecular processes described above (see below). For smaller motif lengths, backward mutations play a role as well as simultaneous slippage of more than one unit, which couples the evolution for a given ℓ with its integer multiples. For larger values of ℓ, slippage events of $(d=1)$-repeats start to become significant.
The majority of novel repeats originates from background at $n_1 = 0$; resurrection of inactive repeats also contributes only weakly to the ensemble of nascent repeats. However, a sizable fraction of novel repeats emerges from background with $n_1 > 0$: assuming a random sequence composition, the probability that a nascent doublet of motif length l adjoins a single doublet of the first error class reads $2\frac{l!}{k!(l-k)!}\left(\frac{1}{4}\right)^{l-k}\frac{1}{4}^k$. For instance, the likelihood of emerging at $n_1 = 1$ for $\ell = 3,4,5$ equals $0.28, 0.09, 0.03$.

For the genome analysis below, we write the doublet composition distribution $W(n_0,n_1,t)$ in the form

$$\begin{aligned} W(n_0,n_1,t) &= \lambda(t)\, P_a(n_0,n_1,t) \\ &\quad +[1-\lambda(t)]\, P_0(n_1,t)\, \delta_{n_0,0}, \end{aligned} \tag{3.10}$$

where $\lambda(t) \equiv \sum_{n_0=1}^{\infty}\sum_{n_1=0}^{\infty} W(n_0,n_1,t)$ measures the density of active repeats per unit sequence length, $P_a(n_0,n_1,t) \equiv \lambda^{-1}(t)W(n_0,n_1,t)(1-\delta_{n_0,0})$ is the normalized composition distribution of active repeats, and $P_0(n_1,t)$ is the "background" distribution of silent repeats and non-repetitive sequence. From a computational point of view, this is a Hidden Markov model containing two ensembles with prior probabilities λ and $1-\lambda$, respectively. Of course, the decomposition (3.10) is not unique, and depending on the sensitivity of genomic repeat finding, one may choose to include silent

repeats or repeats containing units with multiple mutations ($d > 1$) in the repeat coverage.

Stationary state. Eq. (3.6) determines a unique time-independent doublet composition distribution $W(n_0, n_1)$, which is readily obtained by numerical iteration. It depends only on the scaled slippage rates γ_i/μ, γ_+/μ and γ_-/μ. Furthermore, the normalized distributions P_a and P_0 in the decomposition (3.10) are independent of γ_i. Fig. 3.1(a) shows as an example the doublet composition distribution $P_a(n_0, n_1)$ (with $n_0 > 0$) for active trinucleotide ($\ell = 3$) repeats. The stationary marginal distribution of perfect doublets, $p_a(n_0) \equiv \sum_{n_1=0}^{\infty} P_a(n_0, n_1)$, can be computed analytically as

$$p_a(n_0) = \frac{C}{n_0}\,\mathrm{e}^{-\alpha n_0} \quad \text{with } \alpha = -\ln\left(\frac{\gamma_+}{\gamma_- + 2\ell\mu}\right) \tag{3.11}$$

and the normalization $C = \sum_{n_0=1}^{\infty} n_0^{-1}\mathrm{e}^{-\alpha n_0}$, see fig. 3.1(b). A further characteristic of the full distribution $P_a(n_0, n_1)$, the expected ratio of single-mutation doublets and perfect doublets, is numerically observed to be approximately independent of n_0, which defines an independent function of the scaled slippage rates,

$$\beta(\gamma_+/\mu, \gamma_-/\mu) = \frac{\langle n_1\rangle(n_0)}{n_0} \equiv \frac{1}{n_0}\sum_{n_1=0}^{\infty} n_1 P_a(n_0, n_1), \tag{3.12}$$

see also fig. 3.1(b). Eqs. (3.11) and (3.12) determine the observed exponential decay of the marginal distribution of the total doublet number $n \equiv n_0 + n_1$,

$$\tilde{p}_a(n) \sim \frac{1}{n}\,\mathrm{e}^{-\tilde{\alpha} n} \quad \text{with } \tilde{\alpha} = \frac{\alpha}{1+\beta}, \tag{3.13}$$

which is also shown in fig. 3.1(b). Using (3.11), the stationary density λ of active repeats is then given by flux balance between active repeats and background sequence,

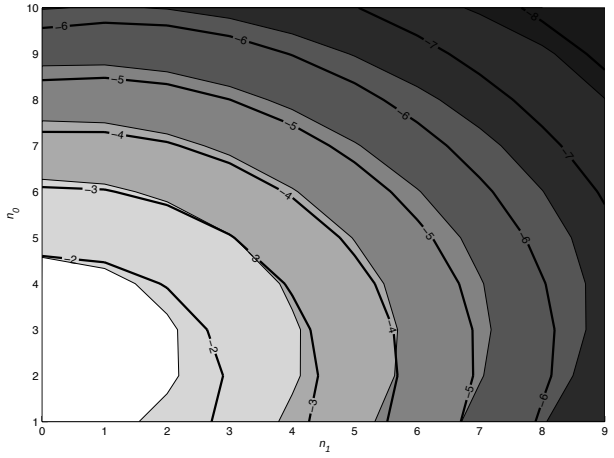
$$\lambda = \frac{1}{C\mathrm{e}^{-\alpha}}\,\frac{\gamma_i}{\gamma_- + 2\ell\mu} + O(\gamma_i^2/\gamma_-^2). \tag{3.14}$$

Time-dependent distributions. It is straightforward to analyze the statistics of repeat life histories by numerical iteration of the time-dependent solution $W(n_0, n_1, t)$ of eq. (3.6) with $\gamma_i = 0$ and with the initial condition $W(n_0, n_1, 0) = \delta_{n_0,1}\delta_{n_1,0}$ describing an initiation event at $t = 0$. This solution can be decomposed in the form (3.10) with a time-dependent density $\lambda_+(t)$ and composition distribution $P(n_0, n_1, t)$ of active repeats, which determines their age-dependent expectation values

$$\langle n_d\rangle(t) \equiv \sum_{n_0=1}^{\infty}\sum_{n_1=0}^{\infty} n_d P(n_0, n_1, t) \quad (d = 0, 1) \tag{3.15}$$

shown in fig. 2(b) and the cumulative distribution of their lifetimes τ, $Prob(\tau < t) = \lambda_+(t)$ (see fig. 2(c)).

a



b

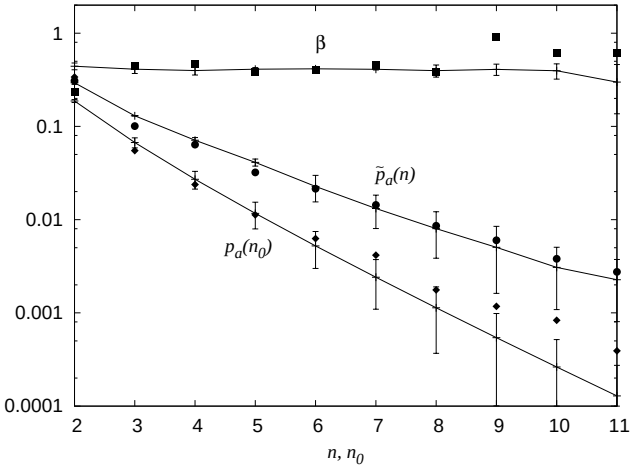


Figure 3.1: **Repeat composition at stationarity.**
(a) **Doublet composition distribution** $P_a(n_0, n_1)$ for active trinucleotide repeats ($\ell = 3$). Theoretical distribution obtained from the mutation-slippage model with scaled rates $\gamma_+/\mu = 4.6$, $\gamma_-/\mu = 2.7$, genomic distribution for intergenic regions of *Drosophila melanogaster* (contour levels). (b) **Marginal doublet number distributions** $p_a(n_0)$ for perfect doublets and $\tilde{p}_a(n = n_0 + n_1)$ for all doublets; **conditional expectation value** $\langle n_1 \rangle(n_0)/n_0 \approx \beta$. Theoretical curves given by eqs. (3.11), (3.13), and (3.12); genomic data for the same sequences as in (a) (diamonds, squares and circles).

3.3.2 Numerical simulation of simple tandem repeats

To obtain numerical results for the configuration distribution of microsatellites as a function of the occupation number of the zeroth and first error class, we performed two independent computations. Assuming constant emergence of new repeat seeds, thus enforcing stationarity of the repeat distribution, the stationary state of the general master

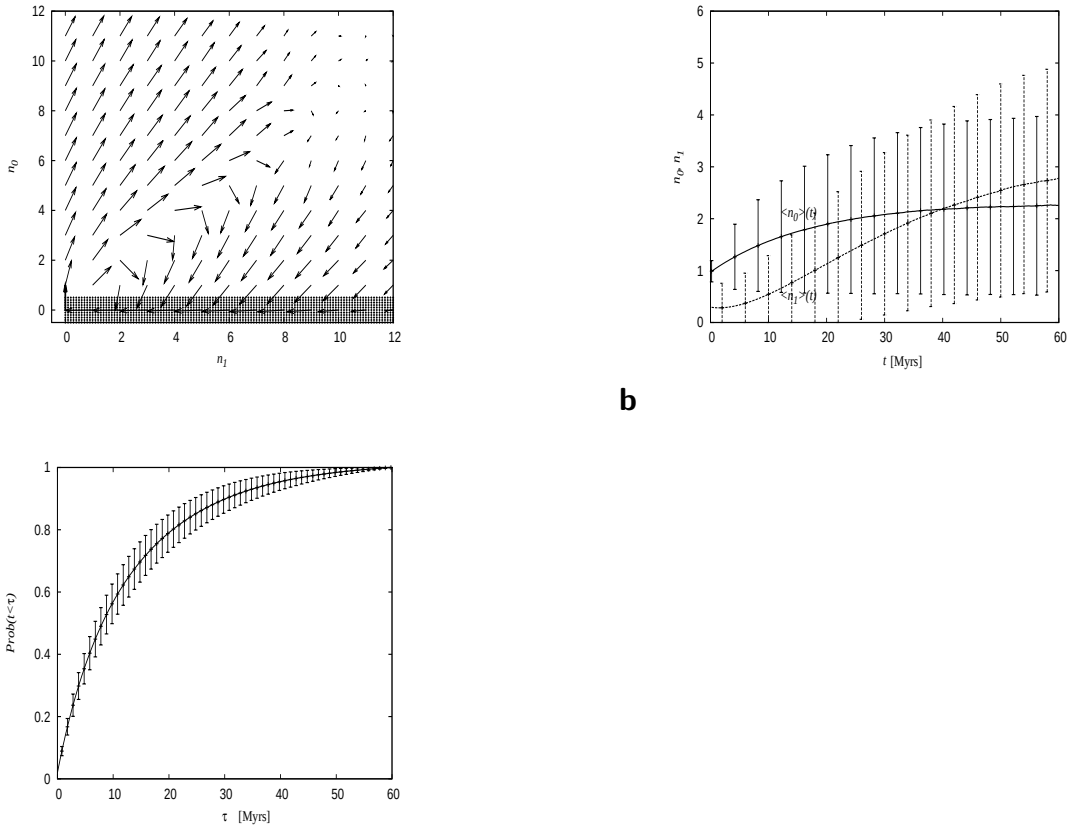


Figure 3.2: **Age-dependent repeat characteristics.**
(a) **Probability flux J**(n_0, n_1) at stationarity, quantifying composition changes during repeat life histories and expressing the breakdown of time reversal symmetry (arrows not to scale for better readability). The region of active repeats ($n_0 > 0$) is shown unshaded, the background region ($n_0 = 0$) shaded, and the repeat initiation current $J_0(0, 0)$ is highlighted (thick arrow). (b) **Doublet content**. Model predictions of the expectation values $\langle n_0 \rangle(t)$ (dashed line) and $\langle n_1 \rangle(t)$ (solid line) with standard deviations (bars) as a function of the repeat age t for model parameters as in fig. 1. (c) **Activity lifetimes** τ. Predicted cumulative age distribution $Prob(\tau < t)$.

equation (3.6) was obtained by simulating all processes contributing to (3.6) in the (n_0, n_1)-plane in occupation number representation. Transitions were computed, using a binomial distribution with appropriate rates. The (n_0, n_1)-plane was limited in size, reflecting a length limitation of repeat elements. Stationarity was ensured by feeding the probability flux leaving the size-limited plane into new seeds emerging at $(n_0 = 0, n_1)$. The upper length boundary was chosen large enough, such that it was practically never touched during the simulation.
To infer time dependent repeat properties we further simulated a single repeat element, starting with a single doublet seed and allowing slippage in either direction and mutation to occur at the respective rates. Whenever the last perfect doublet became extinct by mutation or backward slippage, a new seed was created. At each point in time, n_0

and n_1 were recorded, yielding a time series for (n_0, n_1) and a corresponding probability distribution $P(n_0, n_1)$. At all tested parameter values both approaches perfectly agree.

3.3.3 Inference of slippage rates from single species sequence data

We identified tandem repeat content using the Tandem Repeats Finder [14] for all sets of functionally annotated sequence and intergenic background sequence independently. From these repeat data, the exponent α and its error was determined by a least square fit of the logarithm of the repeat distribution 3.11 to the data for $l = 3, 4, 5$ and all sequence categories independently (see fig. 3.3). For each category and motif length, β was measured from the repeat data. Variation of $beta$ at different values of n_0 indicated a statistical error of the order of 10%. Knowing α and β allows for the inference of γ_-/μ and γ_+/μ for each sequence category and motif length independently. The error of the slippage rates was inferred by varying α and β within their respective error intervals and repeating the data fit. The maximal observed deviation of the rates was then chosen as the error interval. The results are listed in table 3.1.

$l = 3$	3'UTR	5'UTR	CDS	upstream	background	intronic
α	0.54±0.06	0.51±0.07	0.53±0.09	0.61±0.06	0.62±0.07	0.59±0.08
β	0.50±0.05	0.49±0.05	0.50±0.05	0.44±0.04	0.44±0.04	0.37±0.04
γ_+/μ	5.2±1.2	6.0±1.5	5.1±1.5	5.2±0.9	4.6±0.9	6.4±1.2
γ_-/μ	3.0±1.5	3.9±1.8	3.0±1.8	3.0±1.2	2.7±1.2	5.7±1.5
$l = 4$	3'UTR	5'UTR	CDS	upstream	background	intronic
α	0.67±0.04	0.64±0.05	1.00±0.10	0.64±0.06	0.75±0.07	0.64±0.03
β	0.33±0.03	0.30±0.03	0.17±0.02	0.35±0.04	0.34±0.03	0.31±0.03
γ_+/μ	6.8±0.8	8.8±1.2	4.8±1.2	6.8±1.2	4.8±0.8	7.6±0.8
γ_-/μ	6.0±1.2	9.2±2.0	6.4±1.6	4.8±1.2	2.4±1.2	8.0±1.6
$l = 5$	3'UTR	5'UTR	CDS	upstream	background	intronic
α	0.80±0.05	0.86±0.05	1.59±0.09	0.78±0.06	0.88±0.04	0.79±0.06
β	0.23±0.02	0.25±0.03	0.09±0.01	0.26±0.03	0.23±0.02	0.26±0.03
γ_+/μ	8.5±1.5	6.0±1.0	1.0±0.5	8.0±1.5	6.5±1.0	8.0±1.5
γ_-/μ	10.0±2.5	5.0±2.0	0.5±0.5	8.5±2.0	6.5±1.0	8.5±2.0

Table 3.1: **Inferred slippage rates for all sequence categories and motif length** $l = 3, 4, 5$

Examining fig. 3.3 shows that for $l = 4$ and $l = 5$ the observed length distribution is in good agreement with (3.11), except for the tail of the distribution at very small numbers of repeat elements.
For $l = 3$ one observes a significant underrepresentation at $n_0 = 1$, which originates from the alignment score cutoff of the TRF, selected to be 15. If a matched base pair is assigned a score of 2, a single perfect repeat of a motif of length $l = 3$ yields a score of 12 and does not pass the score filter. At higher values of n_0, a zik-zak pattern of the distribution with maxima at even and minima at odd values of n_0 can be identified. A possible reason for this behavior is simultaneous slippage of two motifs in cases where

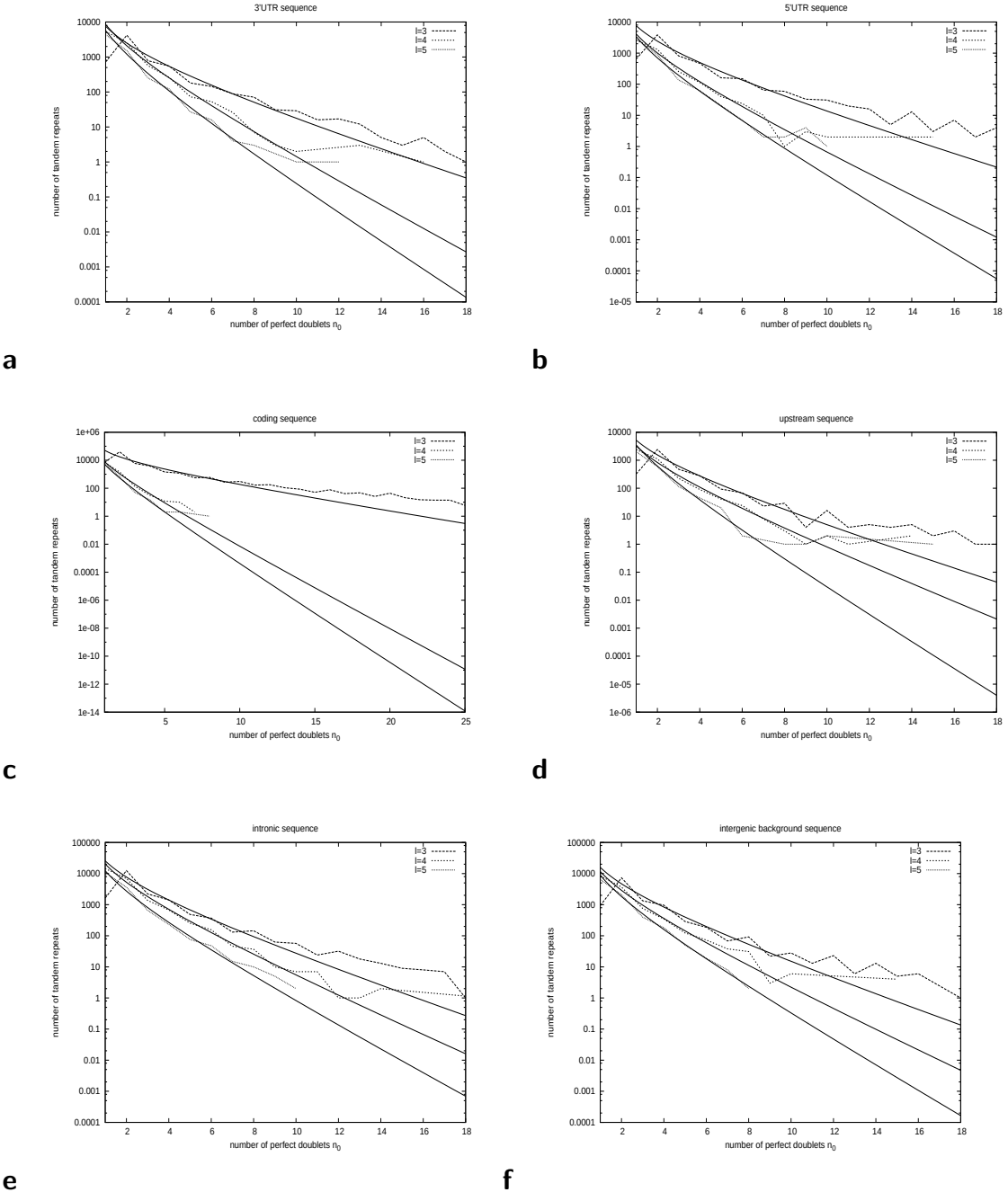


Figure 3.3: **Stationary distribution of perfect duplications in microsatellites** Number of tandem repeat elements as a function of the number of perfectly duplicated motifs n_0 contained in the tandem repeat element, for all sequence categories. α was determined by fitting distribution (3.11) to the data (also shown in the plot).

an uninterrupted run of at least four adjacent $l = 3$-motifs resides in the sequence, a phenomenon that was expected to complicate repeat dynamics for short motifs (see above). Since longer uninterrupted motif iterations are expected to be more abundant in longer repeats, this phenomenon may also account for the more substantial tail at $l = 3$.

$l = 3$	3'UTR	5'UTR	CDS	upstream	background	intronic
γ_+/μ	**3.3**	2.1	1.5	2.4	**4.2**	**3.9**
γ_-/μ	**2.7**	**3.0**	**1.5**	**1.8**	**3.3**	2.7
$l = 4$	3'UTR	5'UTR	CDS	upstream	background	intronic
γ_+/μ	**6.4**	3.6	0.4	2.4	**6.4**	**6.0**
γ_-/μ	**3.2**	2.4	0.8	**3.2**	**4.0**	3.2
$l = 5$	3'UTR	5'UTR	CDS	upstream	background	intronic
γ_+/μ	**6.0**	3.0	**1.0**	**5.0**	**7.0**	**6.0**
γ_-/μ	4.5	**2.0**	**1.0**	2.0	**4.0**	4.0

Table 3.2: **Rate estimates from multiple alignments.**
Estimated slippage rates inferred from parameter optimized alignments of *D. melanogaster*, *D. simulans* and *D. yakuba*. The estimates concordant to the rates inferred from the repeat data within their error intervals are printed in boldface.

Note that the quantities α and β only weakly depend on the TRF parameter settings. Choosing significantly less restrictive parameters (*match score 2, mismatch cost 3, indel cost 3, match probability 80, mismatch probability 10, minimal alignment score 10, maximal repeat length 500*) only marginally affects the rate measurements.

3.3.4 Comparison of repeat-inferred and alignment-inferred slippage rates

To validate our model of repeat dynamics, we measured insertion, deletion and base substitution rates in three-way alignments of *D. melanogaster*, *D. simulans* and *D. yakuba*. The score matrix we used was inferred by parameter optimization for each set of functional sequence separately. From the alignments, we extracted the regions identified as repeats and counted base substitutions, insertions and deletions based on maximum parsimony. Phylogenetic criteria to distinguish insertions from deletions were applied following Sinha and Siggia (2005) [128]. To safeguard against rate biases of very long tandem repeat elements as well as very short ones, which probably do not undergo slippage at all [108], we only recorded rates for repeat elements longer than four times and shorter than twenty times the motif length. Using the entire sets of functional sequence, we measured rates for repeat elements of different motif length separately (see table 3.2).

Slippage rates measured from the alignment that match the repeat-inferred rates within their error intervals are printed in boldface. To estimate the error of the rate measurements, we performed the whole alignment procedure again, this time using the BLASTZ default parameters, which are found to be significantly more restrictive than the globally optimized parameters. Using the BLASTZ scoring matrix, the number of gaps decreases by approximately 50%. We take this number as an estimate for the relative error of the measured rates, although the true error induced by a particular choice of the scoring matrix remains very uncertain. However, since globally optimized parameters were used and repeat elements are more permissive with regard to indels compared to non-repetitive-sequence ([128], our own unpublished observations), the inferred rates are

very likely to be too low. However, single repeat elements in *D. melanogaster* are too short to confidently optimize parameters specifically for repeats. Optimized parameters are found to vary strongly among different instances of tandem repeat elements.
Another reason for the observation of fewer gaps than expected is the fact that gaps in particular configurations will never be observed by any alignment algorithm. For instance, a configuration of a neighboring motif-deletion and -insertion intermitted by any number of perfect (or even weakly mutated) motifs will always disappear for reasonable alignment parameters. In general, insertions and deletions tend to get lost at the expense of point mutations, a phenomenon that becomes particularly pronounced at high ratios of gap rate over point mutation rate. Importantly, this spurious result is amplified in repeat elements, since all gaps tend to be of the same size. These limitations of common alignment algorithms based on standard gap models demonstrate the need of a sophisticated algorithm for aligning repeat enriched sequence.
In fig. 3.4 concordance of the two independent measurements can be read from scatter plots with the repeat-inferred rates on the x-axis and the alignment-inferred rates on the y-axis for all functional annotations. In order to better visualize the expected linear relationship, we plotted the slippage rate divided by the single nucleotide mutation rate. If the two measurements were in perfect agreement, all points would reside on the diagonal. Within the assigned error intervals, rate estimates from the alignments coincide with repeat-inferred rates at all motif-lengths only for intergenic background sequence. For most data points of 3'UTR and intronic sequence, the repeat model is also found to explain the alignment based rate estimates. For coding sequence, the predictions of both independent approaches most poorly agree. Also for the other classes of functional sequence (upstream sequence, 5'UTRs) the two measurements are in bad agreement. This could indicate functional constraints of a more complicated nature, which cannot be captured by models as simple as the existing ones and the one presented here. A possible explanation of the lower correspondence for intronic sequence compared to background sequence is alternative splicing, i. e. the overlap of introns with exons of different transcript variants and stronger enrichment with functional sequence elements. We tried to more precisely assess the nature of the observed deviations across the sequence categories.
First, we computed Pearson's correlation coefficient for the set of all forward- and all backward-slippage rates divided by the single point mutation rate. For the forward- (backward-) slippage rates we obtained a correlation coefficient of 0.46 (0.49), indicating a medium positive correlation in either case. The correlation coefficient for the whole set of all forward- and backward-slippage rates was found to be 0.44. The correlation coefficients r for the sets of all slippage rates in each category are shown in the table below. Note, that a medium positive correlation is found for each category except for coding sequence.
We next tried to classify the systematic deviation by computing the best fit of a straight line through the origin (also shown in fig. 3.4) for each set separately. The table lists the slope a of the best fit as well as the shortest correlation coefficient r:

set	a	r
background	0.99	0.57
3'UTR	0.63	0.63
intronic	0.59	0.46
upstream	0.43	0.50
5'UTR	0.39	0.26
CDS	0.24	0.03

These data show a hierarchy with the minimum deviation for intergenic background, followed by 3'UTR, intronic, upstream, 5'UTR, and finally coding sequence. This order correlates with a putative enrichment with functional elements. Intergenic background presumably contains the lowest amount of functional sequence. Introns do sometimes overlap with exonic sequence of alternative splicing variants and oftentimes contain regulatory elements. To find 5'UTR and upstream sequence at the bottom of the list, with only coding sequence showing larger deviations, corresponds to the expectation that these regulatory regions are under strong functional constraint. Deviations for 3'UTR sequence are of the same order as for intronic sequence. It is further remarkable that the deviations for 3'UTR and intronic sequence as well as for 5'UTR and upstream sequence, respectively, have a similar magnitude, suggesting related evolutionary constraints for microsatellites in the these two pairs of sequence classes. However, further conclusions require a deeper analysis.
In summary, these observations suggest that our simple repeat model can account for the repeat distribution only if the sequence under consideration is not strongly enriched with functional elements, i. e. the overall selective constraint is low. Following our observations, one could thus rank the sequence by the degree of neutrality with respect to tandem repeat evolution. In decreasing order, this yields background, 3'UTR, intronic, upstream, 5'UTR, and coding sequence.

3.3.5 Discussion

We introduced a probabilistic model to compute the stationary distribution of microsatellites as a result of forward- and backward-slippage with independent, motif-length dependent rates as well as single point mutations inhibiting further slippage of the mutated repeat unit. Although modeling tandem repeat evolutionary dynamics as a single-stepwise process with repeat-motif length dependent slippage rate is not new (see Introduction), we exceed previous studies in several regards.
First, we do not restrict the definition of a tandem repeat element only to perfect doublets of the repeat motif, but also include mutated units in order to increase the evolutionary information encoded at each microsatellite locus. We therefore increase the amount of information drawn from the sequence data, enhancing the statistical power to infer slippage rates (see fig. 3.5). Moreover, as opposed to previous microsatellite evolutionary models, we do not neglect spatial correlations induced by point mutations that split larger repeat elements into smaller ones. These correlations can be interpreted as an effective attractive interaction that couples neighboring perfect tandem repeats. Since we only rely on the joint distribution of perfectly repeated units and motif duplications mutated only at a single position, we suppress the inclusion of non-repeat originated background

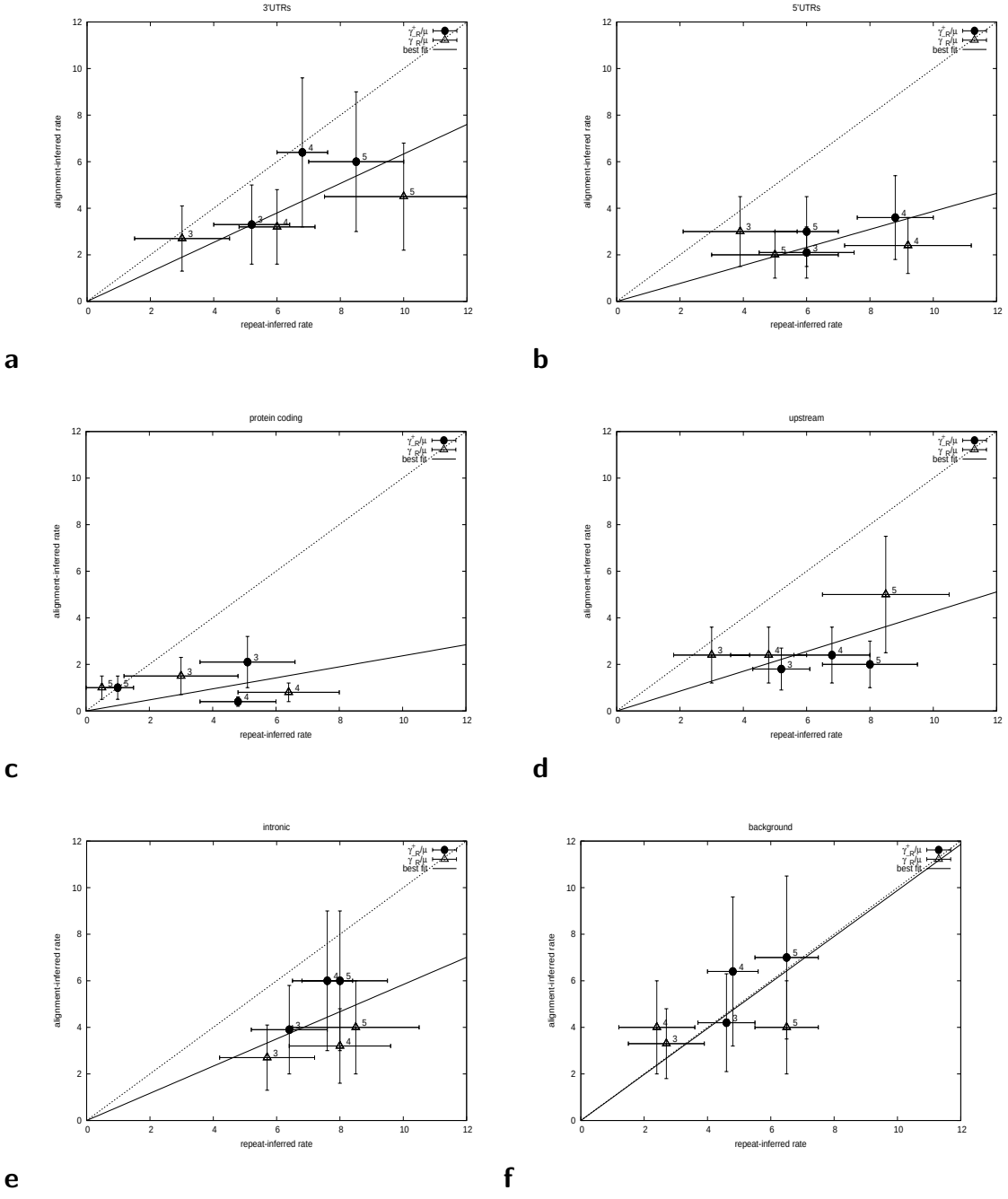


Figure 3.4: **Comparison of rates, predicted by the repeat model and rate measurements from the alignments**
Scatter plots of the rates inferred from the repeat data and inferred from the multiple alignments. The dashed line indicates the diagonal. The numbers next to the data points denote the length of the repeat motif.

sequence, i. e. false positive repeat sequence. For this reason and because of our general concerns regarding the validity of the single-stepwise model for mono- and dinucleotide repeats, we only measure slippage rates for motif length $l \geq 3$.

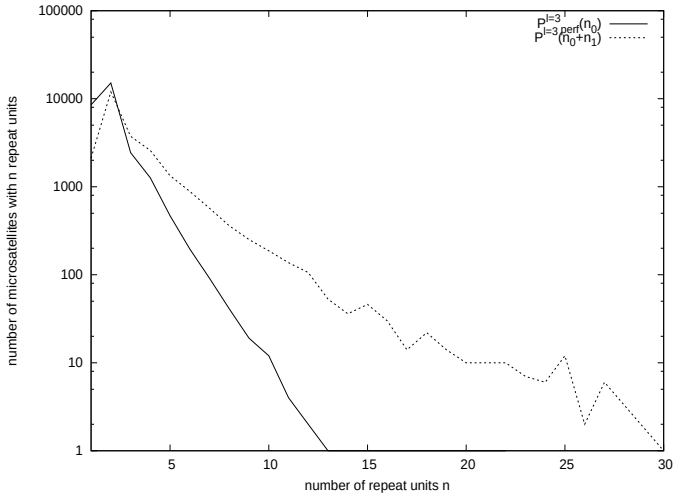


Figure 3.5: **Gain of statistical power due to the inclusion of mutated repeat units**
The figure demonstrates the increased amount of data, if the distribution $P^l(n_0, n_1)$ (here, for convenience, in a one-dimensional representation $P^l(n_0 + n_1)$) is analyzed instead of the length distribution only of perfect repeat elements, $P^l_{perf}(n_0)$. The latter has not only a reduced statistical power due to a decreased amount of data, but also neglects spatial correlations of perfect tandem repeats.

It is not really surprising to observe an exponential distribution of microstaellite length. However, we derive two observables that exhibit a concerted unique dependence on the effective slippage rates, $\gamma_R^+(l)/\mu$ and $\gamma_R^-(l)/\mu$. In order to validate our predictions, we compare to rate measurements from large sets of multiple sequence alignments. We use the model to predict slippage rates for genomic sequence of various functional annotations in *D. melanogaster* and by comparison to multiple alignments of repeat sequence across three closely related *Drosophilids* we demonstrate the validity of our model for intergenic background sequence.
We find that our model appears to be insufficient to explain the repeat distribution for functional sequence. In this case, varying selective constraint on certain repeat classes at different loci as well as the possible dependence of slippage rates on the motif's sequence composition complicates the description of all microsatellites with fixed motif length by a simple general model such as the one presented in this work. The degree of functionality of microsatellites is still highly debated. To this discussion we contribute our insight that the validity of the single-stepwise model negatively correlates with the degree of functionality. Our observations supports the assumption of selective constraint on microsatellites in upstream sequence, UTRs and coding sequence. For intronic sequence, correspondence of repeat- and alignment-inferred rates is also found to be rather low, a finding that can be probably accounted for by alternative splicing. The nature of the systematic deviation of functionally enriched sequence is observed to be a general

suppression of slippage events compared to the rates predicted by the model. The order of the functional sequence categories ranked by the magnitude of the systematic deviation reflects the relative putative fraction of repeat sequence under functional constraint. However, we point out that current alignment algorithms rely on a simplistic modeling scheme of gaps that does not reflect the true processes leading to gaps within alignments of tandem repeats. Since microsatellites contribute an abundant class of sequence to eukaryotic genomes, a more reliable alignment tool for these sequences is most desirable.

In summary, we found that backward- and forward-slippage are rather balanced with the additional action of base substitutions causing a deterministic decay of microsatellites. The repeat analysis indicates that the slippage probability per base pair is only slightly bigger than the probability of a point mutation in tandem repeat elements, which is different from the slippage rates in mammals observed to be about one magnitude higher [74, 101, 35, 33].

Because of the scarcity of experimental data on microsatellite slippage rates, we can only present a very limited comparison. Schug et al. (1998) [123] directly measured the slippage rate of dinucleotide repeats in *D. melanogaster* and found a forward slippage rate of 9.3×10^{-6} per locus which corresponds to 7.8×10^{-7} per repeat unit for an average locus length of 12 repeat units (as it was the case for their data set). Slippage rates for tri- and tetranucleotide repeats were then theoretically inferred by comparing the variance of repeat numbers within the population at different motif length. The ratio of dinucleotide and tri-(tetra-)nucleotide repeats was determined to be 6.4 (8.4), which corresponds to absolute rates of $1.2 \times 10^{-7}(9.3 \times 10^{-8})$ for tri- (tetra-) nucleotide repeats. Since Schug et al. [123] only measured three events of forward slippage, we can only compare to our predicted forward slippage rate. With an assumed point mutation rate of $\mu = 10^{-8}$ we predict the forward slippage rate to be $\sim 5 \times 10^{-8}$ for tri- and tetranucleotide repeats. The rates predicted in this study are only about half of the rates measured experimentally. However, in accordance with the experimental result, slippage rates for tri- and tetra-nucleotides are of the same order of magnitude. Our observations suggest that this is also the case for penthanucleotide repeats. Since the rates in Schug et al. (1998) [123] are inferred based on a very small number of events, we consider the comparison of these rates to our results insufficient to test our model.

3.4 Conclusions

We developed an advanced probabilistic model for the computation of the stationary microsatellite distribution and succeeded to derive slippage rates for microsatellites in intergenic background sequence of *D. melanogaster*. We assessed the validity of our model by comparing to insertion and deletion rates measured from multiple alignments of *D. melanogaster*, *D. simulans* and *D. erecta* and observed that the single-stepwise model is insufficient to explain the microsatellite distribution of functionally enriched sequence. Since these results strengthen the hypothesis that certain classes of simple tandem repeats, or, more precisely, tandem repeats residing in particular sequence classes are under selective constraint, a deeper understanding of their evolution is essential. To this end,

current repeat evolutionary models need to be integrated into alignment algorithms, and more sophisticated models need to be developed for specific classes of tandem repeats, putatively under functional constraint. Although the final validation of proposed evolutionary models requires experiments, *in silico* analyses such as the comparison to independent alignment data discussed in this work provide a valuable contribution to model assessment.

Chapter 4

Evolutionary origins of promoter complexity – the influence of finite population size

4.1 Introduction

Regulatory gene interactions are believed to be vital for the functional diversity in eukaryotes. They govern the differentiation between cells and, to some extent, also between species. Accordingly, the regulatory machinery in eukaryotes is much more sophisticated than in prokaryotes. A typical eukaryotic promoter is longer and contains more transcription factor binding sites. The complicated spatial arrangement of these sites allows for binding cooperativity between factors and for the recruitment of cofactors [107].

The evolutionary modes of eukaryotic promoters remain largely to be uncovered. In particular, it is an open question whether growth in promoter complexity is a response to selection pressures towards functional differentiation. Here, we analyze the evolution of promoter architecture from a population genetic point of view. We argue that the growth in complexity from prokaryotes to eukaryotes is a primary process driven by broad population characteristics, notably the decrease in effective population size. This opens the stage for functional diversification as a secondary process.

In this picture, changes in promoter architecture appear in the context of broader growth in overall genomic complexity. Recently, Lynch and Conery have argued on the basis of a broad population genetic survey that the increase in genome size from prokaryotes to eukaryotes is caused by the accumulation of slightly deleterious repeat elements, which is possible due to the smaller eukaryotic population sizes [90].

It is such repeat insertion processes by which eukaryotes may have acquired complex regulatory modules with groups of adjacent binding sites for the same transcription factor [128].

Here we analyze how background changes between lower and higher taxa, specifically the decrease in effective population sizes and the growth of intergenic regions, affect the architecture and function of transcriptional regulation. We show that generic promoters with a single binding site are stable in prokaryotes but become unstable to decay by genetic drift already in unicellular eukaryotes. This transition is due to the decrease in

effective purifying selection pressure as a result of the smaller population size in eukaryotes. On the other hand, a more complex promoter architecture with multiple loci evolving as (potential) binding sites for the same factor is seen to stabilize the functionality. Since multiple active sites are then likely to be present simultaneously, subsequent increase in gene regulatory complexity is assumed to originate as a secondary consequence, e. g. via enhancer subfunctionalization [45].
Our results are obtained from a generic evolutionary theory for binding site sequences containing mutations, genetic drift, and selection. We present a unified theoretical approach applicable to viruses, prokaryotes, and eukaryotes, i.e., to a range of effective population sizes and of mutation rates varying over several orders of magnitude. The product of these two parameters, μN, governs the balance between mutations and genetic drift. Our general theory contains the known Kimura-Ohta dynamics [65, 96, 97] and the Eigen-Schuster quasispecies model [41, 42] as its limit cases for μN taking very small and very large values, respectively.
In a previous study by Rouzine et al. (2005) [117], a mathematical model, valid for all values of μN, but specific to *single-nucleotide loci* was already introduced and applied to viral systems. Although we started out with a different mathematical framework, we arrive at identical results for the specific case of single-nucleotide loci. However, the extensions of the model presented here allow for a first application of this population genetic theory to the evolution of transcription factor binding sites and even to regulatory modules, containing multiple sites.
To arrive at our conclusions on the compensatory benefit of multiple binding sites in regulatory modules of higher eukaryotes, we argue along similar lines as Force et al. (2005) [45]. In this study, it was also proposed that the decaying effective population size in higher organisms leads to fixation of slightly deleterious sequence and opens the stage for emergence of complex gene regulatory modules by enhancer duplication with subsequent subfunctionalization.
However, our own analysis exceeds these conclusions, since we do not restrict our model to neutral duplication events. More precisely, we demonstrate that a single transcription factor binding site under constant selection pressure could become unstable at a decay of effective population size as it can be observed, for instance, at the evolutionary transition from prokaryotes to eukaryotes. We propose the emergence of additional sites in the same regulatory region as a necessary mechanism to compensate for the decreased stability of a single site in order to maintain the regulatory relationship between the gene and the targeting transcription factor.

4.2 Population genetic model and observables

4.2.1 Definition of the model and derivation of the genotype distribution

In this methodological section we introduce the general population genetics framework and define observables. A master equation that describes the evolution of genotypes driven by mutations and selection serves as a starting point of our analysis. The genotypes under consideration could be whole genomes of a particular species as well as single genes

or even smaller genomic loci, e. g. transcription factor binding sites in gene regulatory models.
Each genotype can be represented by a sequence $\mathbf{s}_i = (s_i^1, \ldots, s_i^L)$ of length L with letters s_i^j drawn from the 4-letter-alphabet $\{A, C, T, G\}$. The genotypic space can thus be thought of as a 4^L-dimensional hypercubus, with the distance of two points, i. e. two sequences $\mathbf{s}_i$ and $\mathbf{s}_j$, measured by the number of differing entries $s_i^k \neq s_j^k$ for $k = 0, .., L$. This metric is termed Hamming distance $d(\mathbf{s}_i, \mathbf{s}_j)$. Each genotype $\mathbf{s}_i$ is assigned a replication rate $F(i)$ to reflect the selective constraint or *fitness* of an individual carrying this genotype by its reproductive success. During the process of genotype replication mutations are assumed to occur at a constant rate μ per nucleotide. We only allow for single point mutations per sequence per generation, which is justified for $\mu L \ll 1$. This generally holds true for all organisms and sequences under consideration. For prokaryotes and eukaryotes, which are subject to our analysis, we assume constant effective population sizes N determined by a limited availability of resources, such as nutrients or host organisms. Technically, the population size is kept constant by requiring the extinction of one individual whenever another one replicates. The probability distribution $\mathcal{P}(\mathbf{n})$ of population states $\mathbf{n}$ is then the solution of the genotypic master equation

$$\partial_t \mathcal{P}(\mathbf{n}) = \sum_{\mathbf{n}'} A(\mathbf{n}, \mathbf{n}') \mathcal{P}(\mathbf{n}') \tag{4.1}$$

with

$$A(\mathbf{n}, \mathbf{n}') = \sum_{i,j=0}^{4^L} (1 - \delta_{ij}) \prod_{\substack{k=0 \\ k \neq i,j}}^{4^L} \delta_{n'_k n_k} \quad \Big(n'_i \mu \delta_{1, d(s_i, s_j)} (\tfrac{1}{3} \delta_{n'_i (n_i+1)} \delta_{n'_j (n_j-1)} - \delta_{n'_i n_i} \delta_{n'_j n_j}) + n'_i n'_j F(j)/N (\delta_{n'_i (n_i+1)} \delta_{n'_j (n_j-1)} - \delta_{n'_i n_i} \delta_{n'_j n_j}) \Big) \ . \tag{4.2}$$

Since we aim at deducing the population states of transcription factor binding sites in the promoter region of a gene, we assign appropriate replication rates $F(i)$ in accordance with known theory [16]. The functionality of a given locus is determined by its ability to be bound by a particular transcription factor. For each DNA binding protein a particular DNA motif can be identified that minimizes the thermodynamic protein-DNA binding energy. Although binding becomes weaker if single nucleotide substitutions occur in the optimal motif, a certain number of point mutations is tolerated without loosing the ability of the transcription factor to specifically recognize the motif. Sampling all known binding sites of a particular transcription factor in individuals belonging to the same population, yields a consensus motif of length L containing the most likely nucleotide at each position. Functionality is assumed to get lost if the number of mutations, i. e. the Hamming distance to the consensus motif, exceeds a certain threshold r. Hence, we projected the whole 4^L-dimensional genotype space onto a one-dimensional subspace centered around the consensus motif and parametrized by the binding energy or, equivalently, by the Hamming distance (single point mutations are assumed to contribute additively to a change in binding energy). For a coarse-grained examination of the underlying evolutionary dynamics we project the one-dimensional subspace once more onto a two-state system: All genotypes of a Hamming distance smaller or equal than r to the consensus motif are considered equally functional and termed master genotypes. Genotyes at larger Hamming distances are considered non-functional (non-master genotypes). This classification corresponds to an approximation of the established sigmoid dependence of

the binding probability on the Hamming distance by a simple step function. The increased fitness of master genotypes is expressed by their enhanced fitness $f_0 = 1$ compared to non-master genotypes, replicating at rate $f_1 = 1 - \sigma$ $(0 < \sigma < 1)$. Introduction of effective mutation rates ($\alpha\mu$ and $\beta\mu$, respectively) allows for the computation of a marginal master equation for the probability distribution $P(n)$ of the population n of functional-sites:

$$\begin{aligned}
&\partial_t P(n) = \\
&\mu L\,[\alpha((n+1)P(n+1) - nP(n)) \\
&+ \beta\tfrac{1}{3}((N-(n-1))P(n-1) - (N-n)P(n))\big] \\
&+ (f_1/N((n+1)(N-(n+1))P(n+1) - n(N-n)P(n)) \\
&+ f_0/N((n-1)(N-(n-1))P(n-1) - n(N-n)P(n)))\,.
\end{aligned} \tag{4.3}$$

If the maximum Hamming distance to the consensus motif with maintained functionality is r, $\mu L\alpha n$ equals the number of individuals with sites at Hamming distance r multiplied by the probability to increase the Hamming distance by a further mutation. Accordingly, $\mu L(1/3)\beta(N-n)$ is the number of individuals with sites at Hamming distance $r+1$ multiplied by the probability to decrease the Hamming distance by a further mutation. In the first case, mutations lead to a loss of functionality, while in the second case, functionality is restored by backward mutations. The continuous Fokker-Planck-equation (see Appendix C) for $x = n/N$ is obtained as an $\mathcal{O}(1/N)$-expansion of (4.3):

$$\begin{aligned}
&\partial_t \bar{P}(x) = \\
&-\partial_x\left((\sigma x(1-x) - \mu L(\alpha x - \beta(1-x)))\bar{P}(x)\right) \\
&+\tfrac{1}{2N}\partial_x^2\left(((2-\sigma)x(1-x) + \mu L(\alpha x + \beta(1-x)))\bar{P}(x)\right).
\end{aligned} \tag{4.4}$$

Assuming stationarity, the Fokker-Planck-equation (4.4) is analytically solvable. Neglecting contributions of $\mathcal{O}(\sigma/N, \mu L/N)$ yields,

$$\bar{P}(x) = Cx^{-1+\mu L\beta/3N}(1-x)^{-1+\mu L\alpha N}e^{\sigma N x}, \tag{4.5}$$

where C is a normalization constant that can be expressed in terms of Bessel- and Gamma-functions.
The coefficients α and β determine the likelihood that a mutation of a master genotype leads to a loss of function and that the mutation of a non-master genotype restores functionality, respectively. Assuming approximately equally distributed genotypes in both

the master- and the non-master genotype class, respectively, one obtains

$$\alpha = \frac{L-r}{L} \frac{\left(\begin{array}{c} L \\ r \end{array} \right) 3^r}{\sum_{k=0}^{r} \left(\begin{array}{c} L \\ k \end{array} \right) 3^k} \tag{4.6}$$

$$\beta = \frac{r+1}{L} \frac{\left(\begin{array}{c} L \\ (r+1) \end{array} \right) 3^{r+1}}{\sum_{k=r+1}^{L} \left(\begin{array}{c} L \\ k \end{array} \right) 3^k}. \tag{4.7}$$

This approximation appears to be reasonable, because for small r we have $\alpha \sim \mathcal{O}(1)$ and $\beta \ll 1$ and the exact value of β only becomes important if $x \approx 0$. In this case, the localizing effect of the master genotype population is also lost. Because mutations increase the Hamming distance with strongly enhanced probability for $L \gg 1$, the population will tend to loose its localization before new master genotype individuals emerge by backward mutations. Hence, the indicated estimate gives a lower bound for β. An upper boundary for β results from the assumption that all non-master genotype individuals are found at Hamming distance $r+1$, i. e. $\beta = (r+1)/L$. The true value of β is likely to be closer to the lower boundary. In the Kimura-Ohta regime at low values of μN the population stays in a monomorphic state [65] and this approximation becomes even better. Once the monomorphic population is found at loci other than the master genotypes, it freely diffuses in genotype space.

To test the quality of our approximations we performed a Monte Carlo simulation of the evolution of a finite population subject to mutations and selection. More precisely, the population was parametrized by the Hamming distance to the consensus motif and all individuals at a given Hamming distance were allowed to change their Hamming distance by one with the probability μ per time step. Sub-populations at each Hamming distance were subject to replication probabilites f_0 or f_1, respectively, per time step and individual. To keep the population size constant, a death rate was introduced, which equals the average replication rate of the whole population at every time step. For three qualitatively different regimes, the comparison of the simulation and the analytical prediction is shown in fig. 4.1. At high mutation rates ($\mu LN \gg 1$) the assumption of a flat distribution of master- and non-master genotypes in their respective genotype sub-space leads to an obvious deviation between the prediction and the actual behavior. In fact, our results indicate an accumulation of master genotypes at Hamming distance zero and a linear decay of the population density up to Hamming distance r (data not shown). At decreasing mutation rate this accumulation is less pronounced and assuming a flat distribution of master- and non-master genotypes in their respective genotype sub-spaces yields a good agreement of the distribution (4.5) and the simulated behavior. Importantly, for any choice of the parameters the theoretical prediction is conservative, i. e. it slightly underestimates the true number of master genotypes by overestimating the mutational flux from master genotypes to non-master genotypes.

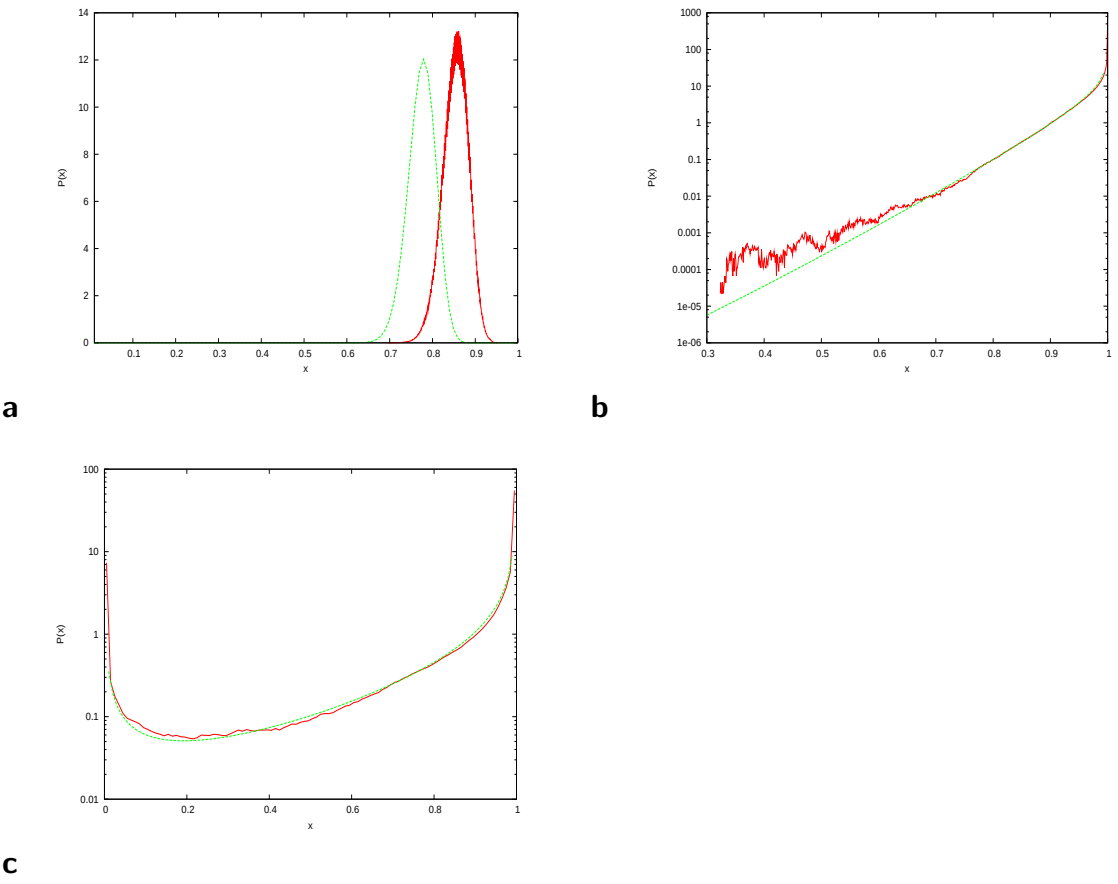


Figure 4.1: **The probability distribution of master genotype-population.** The probability distribution $P(x)$ of the master genotype population fraction x, analytically computed (solid line) and simulated (dashed line). In the quasispecies regime (at $\sigma/(\mu L) = 2$, $1/(\mu LN) = 0.01$ (a)), in the intermediate regime (at $\sigma/(\mu L) = 20$, $1/(\mu LN) = 1$ (b)) and in the Kimura-Ohta regime (at $\sigma/(\mu L) = 10$, $1/(\mu LN) = 0.4$ (c)). $L = 10$ and $r = 3$ in all cases.

4.2.2 Definition of observables to characterize the population state

In order to characterize evolutionary dynamics at arbitrary parameter values, we define three biologically meaningful observables. Considering a transcription factor binding site at a particular genomic locus, equation (4.5) gives the long-time distribution of the fraction of master genotypes in a given population, or equivalently, the distribution over an ensemble of independent populations (e.g., different species sufficiently distant from each other in a phylogeny). The probability distribution (4.5) allows for the computation of the average fraction of working sites:

$$\bar{x} = \int_0^1 x\bar{P}(x)dx. \tag{4.8}$$

In the mean-field limit of infinite population size, this quantity indicates the constant fraction of working sites in a stationary population. However, in finite populations,

stochastic fluctuations become increasingly important. In this case, the mean value of working sites is no longer sufficiently meaningful. Genetic drift becomes a prevalent evolutionary driving force. Fluctuations of the master genotype population between different points in time or, equivalently, between independent populations subject to the same parameters at the same time are quantified by

$$\nu_1^2 = \int_0^1 (x - \bar{x})^2 \bar{P}(x) dx. \tag{4.9}$$

These *inter-population fluctuations* reflect the influence of genetic drift on the population state.
In order to compute the degree of polymorphism in a given population, we define a third observable, termed *intra-population fluctuations*:

$$\nu_2^2 = \int_0^1 x(1 - x) \bar{P}(x) dx. \tag{4.10}$$

In a monomorphic population, consisting of either exclusively master genotypes or non-master genotypes, ν_2 becomes zero. The maximal degree of polymorphism coincides with the maximum of ν_2 when all genotypes are equally distributed among functional and non-functional sites. As it will be shown in the following, these observables provide an appropriate framework to characterize the stationary population state for any choice of the parameters.

4.2.3 The known population genetics theories are recovered as limiting cases of our model.

The probability distribution $\bar{P}(x)$ is seen to be invariant under the transformation $\sigma \to c\sigma$, $\mu \to c\mu$ and $N \to N/c$, and thus has the form $\bar{P}(\sigma N, \mu L N)$. The limiting cases of our model are found to correspond to well known theories:

(i) Selection-mutation equilibrium for $\mu L N \gg 1$: $\bar{P}(\sigma N, \mu L N) \to \bar{P}(\sigma/\mu L)$. Evolutionary dynamics in this regime are governed by the interplay of mutation and selection. Genetic drift becomes negligible. Hence, the fraction of working sites follows the deterministic behavior, predicted by the mean-field limit of our model, which coincides with the well known traditional quasispecies model [41, 42]. No fixation of single genotypes is observed and the population is generically polymorphic with a stationary fraction $\bar{x}(\sigma/\mu L)$ of functional sites (see fig. 4.1a). For sufficiently large sequence length and $r \ll L$ this model predicts a delocalization transition at $\alpha \mu L = \sigma$, termed error threshold. For higher μ selection does not localize the population in proximity to the consensus motif in genotype space anymore.

(ii) Equilibrium between selection and genetic drift for $\mu L N \ll 1$: $\bar{P}(\sigma N, \mu L N) \to \bar{P}(\sigma N)$. In this regime, much less than one mutation per generation is observed on average. The population is generically monomorphic, i.e. most of the time either all individuals are master genotypes or non-master genotypes, respectively. If a competing

allele with macroscopic population emerges from a rare mutation, one allele is driven to fixation after $\mathcal{O}(N)$ generations, while the other one becomes extinct. The fixation rates are predicted by the Kimura-Ohta theory [65], here for time-independent purifying selection. The long-term average $\bar{x}(\sigma N)$ between functionality and non-functionality is determined by the probabilities of the population states $x = 0$ and $x = 1$. The ratio of these probabilities, computed with (4.5), roughly equals $e^{\sigma N - \ln(3\alpha/\beta)}$ in accordance with the Kimura-Ohta theory (see fig. 4.1b).

(iii) The intermediary regime ($\mu LN \sim 1$) describes a crossover between quasispecies behavior and Kimura-Ohta fixation dynamics. Population characteristics are more complex in this case, depending on both variables, $\sigma/\mu L$ and μLN (see fig. 4.1c).

The three observables $\bar{x}$, ν_1 and ν_2 are convenient quantities to delineate these three different parameter regimes. In fig. 4.2, $\bar{x}$, ν_1 and ν_2 are plotted as a function of μLN and σN for a sequence of length $L = 10$ and $r = 3$.

The long term average of the fraction of master genotypes in fig. 4.2a shows a clear separation of a parameter regime with a macroscopic population and a delocalized regime with vanishing population of master genotypes. For small values of μLN the population state is governed by Kimura-Ohta dynamics. The transition between a monomorphic population at the master genotype locus and a population freely diffusing in genotype space is found to obey the Kimura-Ohta predictions: The probabilities of the population states $x = 0$ and $x = 1$ become approximately equal at

$$\sigma N \sim \ln(3\alpha/\beta). \tag{4.11}$$

In fig. 4.2a, this horizontal transition line at constant values of σN is recovered. Hence, it can already be concluded that a binding site of length $L = 10$ at $r = 3$ requires a selection coefficient of $\sigma N \sim 10$ in order to maintain in a population residing in this regime. For $\mu LN \gg 0$ we approach the quasispecies regime and recover the delocalization transition at the error-threshold $\sigma = \alpha\mu L$. This equation describes a straight line with slope α through the origin in fig. 4.2a, delimiting the localized regime.

For a given value of μLN, we define the error threshold as the value $\sigma/(\mu L)$, where the inter-population fluctuations of the master genotype reach their maximum. A further increase of μ at a constant selection coefficient σ would then lead to decreasing fluctuations because the average population of the master genotype approaches rapidly $x = 0$. The population distribution (4.5) is peaked in proximity to the mean-field value, predicted by the quasispecies model, but diverges at $x = 0$. This divergence (with finite probability) is due to the extinction of the master genotype population caused by fluctuations. In the absence of backward mutations (for an infinitely long sequence), the population would remain extinct. The sequence length determines the average waiting time for a backward mutation to repopulate the master genotype. Given a fixed value of μLN, the inter-population fluctuations grow with decreasing $\sigma/\mu L$, and become maximal when the peak of the distribution (4.5) vanishes and the distribution decays monotonically. The finite population correction of the error-threshold in the quasispecies regime, $\sigma/(\mu L) = \alpha$, with respect to μLN (at $\beta \sim 0$, i. e. neglecting backward

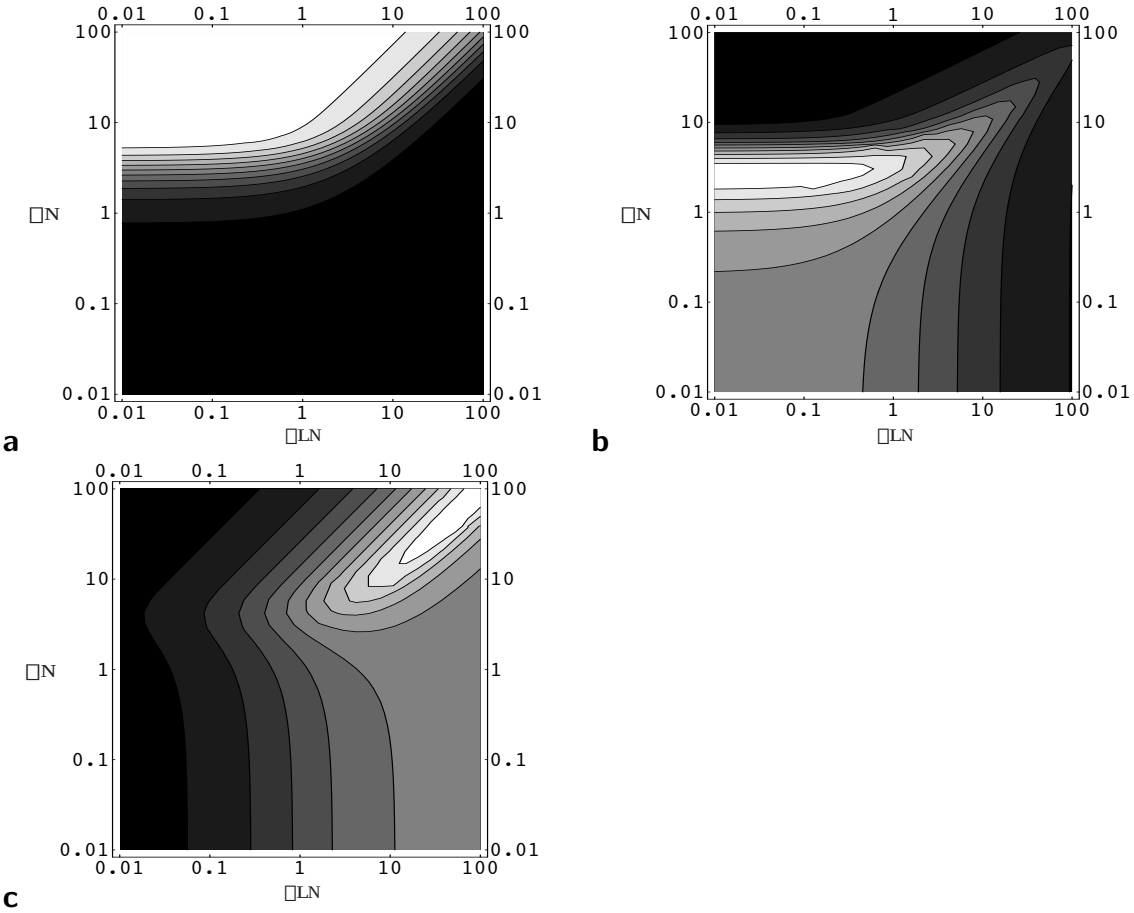


Figure 4.2: **Parameter dependence of the observables**
Shown is (a) the average fraction of master genotypes (black (white) shading corresponds to a minimal (maximal) value of < 0.1 (1)), (b) the inter-population fluctuations ν_1 and (c) the intra-population fluctuations ν_2 as a function of μLN and σN on a logarithmic scale. In (b) and (c) black (white) shading corresponds to a minimal (maximal) value of < 0.1 (0.5). As described in the main text a clear separation of localized and non-localized population states, depending on the parameter values is recovered. The transition line to the delocalized regimes coincided with maximal inter- and intra-population fluctuations in the Kimura-Ohta and the quasispecies regime, respectively.

mutations at $r \ll L$) is thus the solution of

$$\left(\frac{\sigma}{\alpha\mu L} - 1\right)^2 - \frac{4}{\alpha\mu LN}\left(1 - \frac{1}{\alpha\mu LN}\right) = 0, \tag{4.12}$$

describing the parameter combination, where maximum and minimum of (4.5) coincide. At first order in $\frac{1}{\mu LN}$ this yields

$$\frac{\sigma}{\mu L} = \alpha + \frac{2\sqrt{\alpha}}{\sqrt{\mu LN}}, \tag{4.13}$$

in accordance with an earlier study [94]. The inter-population fluctuations (fig. 4.2b) are maximal in the Kimura-Ohta regime and decay along the transition to the quasispecies

regime. At constant values of μLN in the Kimura-Ohta regime, these fluctuations are found to be maximal, if the population states $x = 0$ and $x = 1$ have equal probabilities, i. e. at the delocalization transition. At constant values of σN, the intra-population fluctuations (fig. 4.2c) vanish in the Kimura-Ohta regime and grow with increasing μLN until they become maximal in the quasispecies regime when approaching the error-threshold. At this point the population is equally distributed among master- and non-master genotypes. Larger values of μLN lead to a reduction of master genotypes and therefore to decreasing intra-population fluctuations.
Hence, the parameter regime with a macroscopic fraction of functional sites (fig. 4.2a) is delineated by the maximum of the intra-population fluctuations in the quasispecies-regime and by the maximum of the inter-population fluctuations in the Kimura-Ohta regime. In the intermediate regime $\mu LN \sim 1$ both kinds of fluctuations govern the population state. Therefore neither the quasispecies nor the Kimura-Ohta theory provides an appropriate description in such cases.

4.2.4 Modeling population dynamics of promoters with multiple sites

The condensed gene regulatory regions of prokaryotic organisms exhibit only very few transcription factor binding sites. With growing complexity of the evolving species, in particular at the transition from prokaryotes to eukaryotes, the complexity of gene regulation increases, accomplished by long promoter regions spanning up to thousands of nucleotides covered with groups of various transcription factor binding sites in higher eukaryotes. New biological functions are thus assumed to evolve by modulating cell- or tissue-specific gene expression rather than increasing the number of genes. However, the growth of promoter regions with organismic complexity and their intricate architecture could be interpreted as a consequence of decreasing effective population size. Assuming a transcription factor binding site under constant selection pressure σ, the substitution probability by a mutated non-functional allele grows with a drop off in μN. If at the same time the promoter size increases and the positioning of the binding site with respect to the transcription start site is flexible within certain boundaries, the enhanced risk to become extinct is compensated by an augmented number of putative loci for this site. As a consequence, multiple copies of sites for the same and/or for a different transcription factor may be present simultaneously. It could be imagined that complex tissue- or cell-type specific expression patterns emerge from these configurations by an accumulation of neutral evolutionary events. To assess the benefit of many possible loci over a single locus for a given transcription factor binding site, the single-site model (4.3) is modified as follows: Considering a gene under the control of a particular transcription factor, we require at least a single binding site for this factor in the gene regulatory region in order to maintain a functional allele. However, the motif is allowed to reside at a number of different loci in the promoter region. We call these loci *candidate sites*. Given a typical eukaryotic gene regulatory module of ~ 500 nucleotides, a number of at least ten independent candidate sites seems to be a conservative estimate. The whole population is again divided into two classes, functional genotypes with at least one working site and non-functional genotypes with a selective disadvantage σ. In order to allow for multiple candidate sites in (4.3), the effective mutation rates α and β have to

be modified. Assuming k sites, the quantity β acquires an additional factor k, because each individual with no working site possesses k candidate sites to restore functionality. If a certain individual possesses more than one working site, a single mutation cannot lead to loss of functionality. However, in order to keep the model conservative α is left unchanged, i. e. every master genotype is assumed to harbour only a single working site. In the next section, we demonstrate how a higher multiplicity of sites enhances the fraction of functional alleles in a given population and compensates for a disadvantageous shift of the delocalization transition due to a reduction in effective population size. It can be read from (4.11) that in the Kimura-Ohta regime a multiplicity of k shifts the error threshold to lower values of σN by a factor of $\ln k$. Equivalently, the probability ratio between the population states at $x = 1$ and $x = 0$ in the Kimura-Ohta regime multiplies by a factor of k.

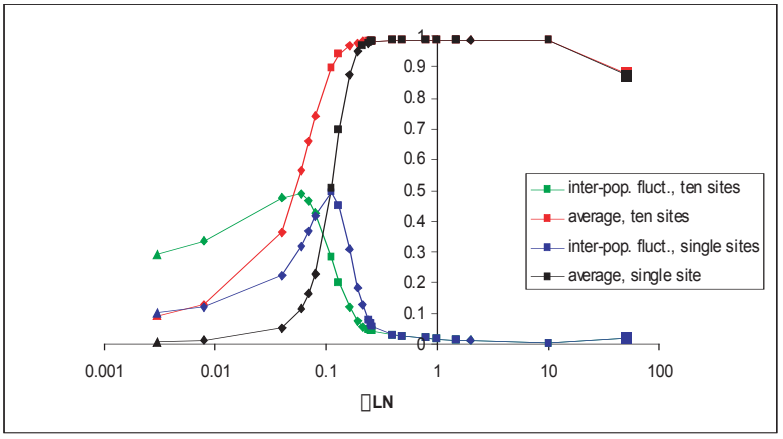


Figure 4.3: **Fraction of master genotypes at fixed selection coefficient** σ
Fraction of master genotypes and inter-population fluctuations across different organisms (triangle: higher vertebrates, diamonds: unicellular eukaryotes, small squares: prokaryotes, large square: viruses (see Appendix D for details)), depending on μLN for a single site and a regulatory module with ten candidate loci ($k = 10$) subject to a constant selection pressure. For the selection coefficient we assumed $\sigma \sim 10^{-7}$. Each site is assumed to have the same length $L = 10$ with $r = 3$. Among prokaryotes and unicellular eukaryotes, the single nucleotide mutation rate μ is roughly $2 \cdot 10^{-10}$. For higher vertebrates the mutation rate is increased by a factor of 100 reflecting the number of germ line divisions per generation. To represent viral system, a data point with $\mu \sim 5 \cdot 10^{-9}$ and $N \sim 10^9$ was added. While promoter sites in viral system tend to reside in the quasispecies regime and have a certain degree of polymorphism, they shift to the intermediate and the Kimura-Ohta regime for prokaryotic organisms. The transition from prokaryotes to unicellular eukaryotes is found to coincide with a steep decay of the average fraction of working sites in a given population. As the plot demonstrates, this trend is significantly alleviated by introducing multiple sites.

4.3 Application to promoter regions across distant domains of life

In this section we apply the population genetic model for binding site evolution to conduct a comparative analysis of several domains of life, from viral over prokaryotic to eukaryotic systems.
The transition from prokaryotes to eukaryotes correlates with a decrease in effective population size. The mutation rate per nucleotide and germline cell division is approximately equal in eukaryotes and prokaryotes with an average value of $\sim 2.3 \cdot 10^{-10}$. To obtain the mutation rate per generation, this rate needs to be multiplied by the number of germline cell divisions per generation.
The effective population size for prokaryotic organisms is larger than 10^8. Effective population sizes range from 10^7 to 10^8 for unicellular eukaryotes, from 10^5 to 10^6 for invertebrates and form 10^4 to 10^5 for higher vertebrates.
Mutation rates of DNA-viruses are between 10^{-7} and 10^{-10} and an effective population

size of the order of $\sim 10^9$ is observed.
Using these data, groups of organisms by can be classified by distinct values of the product of mutation rate and effective population size μN. This quantity is largest in viruses, it decreases in prokaryotes and is further reduced with increasing complexity of the organism from unicellular eukaryotes to higher vertebrates.
We applied our model to estimate the average fraction of functional sites for a typical promoter binding site locus under constant selection pressure across populations of different representative organisms (see Appendix D).
A recent study in *E. coli* yielded an estimate for σN of the order of 10, i. e. $\sigma \sim 10^{-7}$ [93]. Although we assumed a similar selection pressure on a single site for all organisms under consideration, values of σ are expected to vary drastically between essential regulatory functions (e.g. embryonic development) and typical adult regulatory functions. However, we suppose this estimate to be reasonable if the regulated gene is dispensable to a certain degree and therefore a putative target of evolution.
We examine a transcription factor binding site of length $L = 10$, which tolerates a maximum number of three point mutations without loosing its ability to attract a sufficient amount of transcription factors, i. e. all sequences up to a Hamming distance $r = 3$ from the consensus motif are assumed to be functional. We computed the fraction of functional sites for a variety of different organisms and plotted the result in fig. 4.3 as a function of μLN.

Between viral and prokaryotic systems, there is a crossover from generically polymorphic to monomorphic populations. Compared to prokaryotic and eukaryotic systems, the population distribution of an average transcription factor binding site in viral systems resides closest to the quasispecies regime.
For DNA-viruses with effective population sizes of the order of 10^{-9} the quantity μLN ranges from 1 to 1000 for a binding site of length 10. As it can be seen from the intra-population fluctuations, populations residing in this regime are polymorphic. The data point in fig. 4.3, representing populations of typical DNA-viruses, indicates that roughly 10% of all sites in individuals of such populations are distributed among non-functional genotypes at $\sigma \sim 10^{-7}$. This fraction decreases significantly in bacteria. Binding sites of the same length and subject to the same selection pressure show a much higher degree of monomorphism in prokaryotes. A transition from DNA-viruses to prokaryotes with values of μLN between 0.1 and 10 can thus be interpreted as a crossover from the localized quasispecies to the localized intermediate and Kimura-Ohta regime, i. e. from generically polymorphic to generically monomorphic populations. This transition, driven by a decreasing effective population size in prokaryotes compared to viral systems follows a trajectory inside of the localized regime in fig. 4.2 from high to low values of μLN. Evolutionary dynamics of viral systems have been described by the quasispecies theory in previous works [62, 63]. More precisely, both viruses and the immune system of the host organism can be represented by quasispecies with coupled selection pressures, co-evolving in permanent competition. The ability of the immune system to adapt to a particular virus requires the polymorphic population structure of viral systems to escape the immune system. Prokaryotes, on the other hand, have already more complex and therefore more conserved genomes with a significantly reduced degree of polymorphism. These organisms reside in the intermediate regime ($\mu LN \sim 1$) with

inter-population fluctuations ranging from very small to moderate values (at $\mu LN \sim 0.1$).

Promoters with a single binding sites are generically stable in prokaryotes, but become unstable to genetic drift in eukaryotes. According to fig. 4.3, a typical transcription factor binding site with $\sigma \sim 10^{-7}$ is found to be highly monomorphic and therefore stable among the majority of prokaryotic organisms. It can be read from this Figure, that the transition from prokaryotes to unicellular eukaryotes at values of $\mu LN \sim 0.1$ coincides with a decay of the effective population size, which destabilizes the fraction of functional sites. This evolutionary trajectory drives the population to the Kimura-Ohta regime, but crosses over to the delocalized regime at the same time. Within the parameter range of μLN for unicellular eukaryotes from 0.01 to 1 the inter-population fluctuations reach their maximum and the average fraction of working sites strongly decays (fig. 4.3). With further increasing genomic complexity in multicellular eukaryotes, the effective population size is reduced by another order of magnitude and the fraction of functional sites decays to its neutral value. Hence, a single promoter site under constant selection pressure could become extinct due to a reduction in effective population size along the transition from prokaryotes to eukaryotes.

In eukaryotes, multiple binding sites for the same factor increase the stability of the regulatory function. Comparison of unicellular eukaryotes and prokaryotes reveals fundamental differences in promoter architecture. Promoter regions in prokaryotes are short and single transcription factor binding sites are predominant, whereas in promoter regions of unicellular eukaryotes modules of multiple transcription factor binding sites are ubiquitous. We devised a simple model to examine the influence of the introduction of multiple sites on the fraction of functional alleles (promoter regions with at least one functional site) in a population. Our model is based on the assumption that in eukaryotic promoter regions *cis*-regulatory sites are allowed to emerge at various positions, which is supported by experiments for certain classes of eukaryotic enhancers [72, 73]. For typical eukaryotic promoters a reasonable estimate of the number of these candidate sites ranges from 10 to more than 100. If there is no further constraint, the number of candidate sites can reach the order of the promoter module length. We proceeded to assign the full fitness advantage to alleles with at least a single functional site. Fig. 4.3 shows the enhancement of functional alleles if ten candidate sites are present, demonstrating how an increased multiplicity of sites shifts the delocalization transition to smaller values of μLN and reduces the slope of the decaying fraction of functional sites with decreasing μLN. We conclude that the introduction of multiple sites due to neutral inflation of promoter regions with decreasing effective population size, e. g. by fixation of slightly deleterious enhancer duplications [90, 45], provide a compensatory mechanism to maintain a regulatory relationship between a particular gene-transcription factor pair if the selection coefficient is left unaltered.

The emergence of complex regulatory networks could thus be interpreted as a secondary consequence of increased multiplicity of promoter sites that originally served to maintain existing regulatory relationships. This may explain the change in promoter architecture without changing selection pressure towards functional differentiation as a secondary process.

It must be stressed again that our model reflects only the lower boundary of the benefit

of multiple sites. Furthermore, the number of ten candidate sites is likely to be an underestimate in higher eukaryotes, where promoter regions are very long.
Our hypotheses agree with several recent studies on the interplay of genome expansion and functionalization with increasing complexity of the organism. Lynch and Conery (2003) [90] found that the reduced effective population size in higher organisms allows for the fixation of neutral or even slightly deleterious insertions, arisen by gene duplication or other transposable elements. The authors observed a significant increase in the number of transposable elements at the transition from prokaryotes to eukaryotes and argue that fixation of neutral or slightly deleterious insertions by genetic drift allows for rarely occurring beneficial mutations.
According to a recent cross-species comparison in three *Drosophila* species by Sinha and Siggia [128] and our own analysis, a large fraction of regulatory non-coding sequence can be explained by insertions of repeat elements which outweigh deletions. However, Petrov and Hartl [100] predicted a 1:8 ratio between insertions and deletions for neutral intergenic sequence. This suggests that insertions in regulatory sequence are protected from deletion to provide a resource for the emergence of functional sites by beneficial mutations. We conclude that this mechanism could have been crucial at the evolutionary transition from prokaryotes to eukaryotes in order to maintain functionality of regulatory relationships and allowed at the same time for the evolution of increasingly complex regulatory interactions. It is conceivable that in higher eukaryotes integration could even exist on larger scales, e. g. cooperativity of whole regulatory modules or even compensatory networks of different genes, opening yet undiscovered paths for the evoltion of organismal complexity.

4.4 Conclusions

We introduced a general population genetic theory for the computation of the stationary state population distribution of transcription factor binding sites. Supported by Monte Carlo simulations we argue that our approximations capture the main evolutionary characteristics of a single transcription factor binding site and introduce an extension to this model that allows for lower bound estimates of the fraction of functional regulatory modules with multiple putative loci for a transcription factor binding site.
The analytical solution of our model allows us to derive a "phase-diagram" of binding site populations governed by the parameters μN and σN that delineates qualitatively distinct regimes and knows the well established quasispecies [41, 42] and the theory of neutral evolution by Kimura and Ohta [65, 96, 97] as two limiting cases of a general distribution. Parametrizing a variety of taxa by the product of mutation rate and effective population size, we show that a single binding site under fixed selection pressure could become unstable at the evolutionary transition from prokaryotes to eukaryotes as a consequence of decreasing population size. We propose the introduction of multiple putative loci for binding sites in the same promoter, as a neutral result of a reduction in population size, as a compensatory mechanism to maintain existing regulatory relationships. Emergence of a complex promoter architecture can then be easily thought of as a consequence of this compensatory mechanism, since a higher multiplicity of sites opens the stage for the evolution of tissue-specific regulation by compensatory mutations [45].

Bibliography

[1] Abbott, A.L., Alvarez-Saavedra, E., Miska, E.A., Lau, N.C., Bartel, D.P., Horvitz, H.R. & Ambros, V. The let-7 MicroRNA family members mir-48, mir-84, and mir-241 function together to regulate developmental timing in Caenorhabditis elegans. *Developmental Cell* **9**, 403–414 (2005).

[2] Ambros, V. The functions of animal microRNAs. *Nature* **431**, 350–355 (2004).

[3] Andolfatto, P. Adaptive evolution of non-coding DNA in Drosophila. *Nature* **437**, 1149–52 (2005).

[4] Bachtrog, D., Weiss, S., Zangerl, B., Brem, G. & Schlötterer, C. Distribution of dinucleotide microsatellites in the Drosophila melanogaster genome. *Molecular Biology and Evolution* **16**, 602–610 (1999).

[5] Bacon, A.L., Dunlop, M.G. & Farrington, S.M. Hypermutability at a poly(A/T) tract in the human germline. *Nucleic Acids Research* **29**, 4405–4413 (2001).

[6] Bagga, S., Bracht, J., Hunter, S., Massirer, K., Holtz, J., Eachus, R. & Pasquinelli A.E. Regulation by let-7 and lin-4 miRNAs results in target mRNA degradation. *Cell* **122**, 553–563 (2005).

[7] Banerjee, D. & Slack, F. Control of developmental timing by small temporal RNAs: a paradigm for RNA-mediated regulation of gene expression. *Bioessays* **24**, 119–129 (2002).

[8] Barad, O. et al. MicroRNA expression detected by oligonucleotide microarrays: system establishment and expression profiling in human tissues. *Genome Research* **14**, 2486–2494 (2004).

[9] Bartel, D.P. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* **116**, 281–297 (2004).

[10] Bartel, D.P. & Chen, C.Z. Micromanagers of gene expression:the potentially widespread influence of metazoan microRNAs. *Nature Review Genetics* **5**, 396–400 (2004).

[11] Baskerville, S. & Bartel, D.P. Microarray profiling of microRNAs reveals frequent coexpression with neighboring miRNAs and host genes. *RNA* **11**, 241–247 (2005).

[12] Baum, L.E An equality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes. *Inequalities* **3**, 1–8 (1972).

[13] Bell, G.I. & Jurka, J. The length distribution of perfect dimer repetitive DNA is consistent with its evolution by an unbiased single-step mutation process. *Journal of Molecular Evolution* **44**, 414–421 (1997).

[14] Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Research* **27**, 573-580 (1999).

[15] Berezikov, E., Guryev, V., van de Belt, J., Wienholds, E., Plasterk, R.H. and Cuppen, E. Phylogenetic shadowing and computational identification of human microRNA genes. *Cell*, **120**, 21–24 (2005).

[16] Berg, O.G., von Hippel, P.H. Selection of DNA Binding Sites by Regulatory Proteins *Journal of Molecular Biology* **193**, 723–750 (1987).

[17] Blanchette, M., Kent, W.J., Riemer, C., Elnitski, L., Smit, A.F., Roskin, K.M., Baertsch, R., Rosenbloom, K., Clawson, H., Green, E.D., Haussler, D. & Miller, W. Aligning Multiple Genomic Sequences with the Threaded Blockset Aligner. *Genome Research* **14**, 708–715 (2004).

[18] Bray, N. & Pachter, L. MAVID: Constrained ancestral alignment of multiple sequences. *Genome Research* **14**, 693–699 (2004).

[19] Brennecke, J., Hipfner, D.R., Stark, A., Russell, R.B. & Cohen, S.M. Bantam encodes a developmentally regulated microRNA that controls cell proliferation and regulates the proapoptotic gene hid in Drosophila. *Cell* **113**, 25–36 (2003).

[20] Brennecke, J., Stark, A., Russell, R.B.& Cohen, S.M. Principles of MicroRNA-Target Recognition. *PLoS Biology* **3**, e85 (2005).

[21] Brudno, M., Do, C.B., Cooper, G.M., Kim, M.F., Davydov, E.; NISC Comparative Sequencing Program; Green, E.D., Sidow, A. & Batzoglou, S. LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. *Genome Research* **13**, 721–731 (2003).

[22] Burgler, C. & Macdonald, P.M. Prediction and verification of microRNA targets by MovingTargets, a highly adaptable prediction method. *BMC Genomincs* **6**, 88 (2005).

[23] Calabrese, P. & Durrett, R. Dinucleotide repeats in the Drosophila and human genomes have complex, length-dependent mutation processes. *Molecular Biology and Evolution* **20**, 715–725 (2003).

[24] Calin, G.A., Dumitru, C.D., Shimizu, M., Bichi, R., Zupo, S., Noch, E., Adler, H., Rattan, S., Keating, M., Rai, K., Rassenti, L., Kipps, T., Negrini, M., Bullrich, F. & Croce, C.M. Frequent deletions and down-regulation of microRNA genes miR15 and miR16 at13q14 in chronic lymphocytic leukemia. *Proceedings of the National Academy of Sciences USA* **99**, 15524–15529 (2002).

[25] Calin, G.A., Sevignani, C., Dumitru, C.D., Hyslop,T., Noch, E., Yendamuri, S., Shimizu, M., Rattan, S., Bullrich, F., Negrini, M. & Croce, C.M. Human microRNA genes are frequently located at fragile sites and genomic regions involved in cancers. *Proceedings of the National Academy of Sciences USA* **101**, 2999-3004 (2004).

[26] Calin, G.A., Liu, C.G., Sevignani, C., Ferracin, M., Felli, N., Dumitru, C.D., Shimizu, M., Cimmino, A., Zupo, S., Dono, M., Dell'Aquila, M.L., Alder, H.J., Rassenti, L., Klipps, T. J., Bullrich, F., Negrini, M. & Croce, C.M. MicroRNA profiling reveals distinct signatures in B cell chronic lymphocytic leukemias. *Proceedings of the National Academy of Sciences USA* **101**, 11755–11760 (2004b).

[27] Castilllo-Davis, C.I. & Hartl, D.L. GeneMerge–post-genomic analysis, data mining, and hypothesis testing. *Bioinformatics* **19**,891–892 (2003).

[28] Chang, S., Johnston, R.J. Jr., Frokjaer-Jensen, C., Lockery, S. & Hobert, O. MicroRNAs act sequentially and asymmetrically to control chemosensory laterality in the nematode. *Nature* **430**, 785–789 (2004).

[29] Chatterji, S. & Pachter, L. Reference based annotation with GeneMapper. *Genome Biology* **7**, R29 (2006).

[30] Chen, N., Harris, T.W., Antoshechkin, I., Bastiani, C., Bieri, T., Blasiar, D., Bradnam, K., Canaran, P., Chan, J., Chen, C.K., Chen, W.J., Cunningham, F., Davis, P., Kenny, E., Kishore, R., Lawson, D., Lee, R., Muller, H.M., Nakamura, C., Pai, S., Ozersky, P., Petcherski, A., Rogers, A., Sabo, A., Schwarz, E.M., Van Auken, K., Wang, Q., Durbin, R., Spieth, J., Sternberg, P.W. & Stein, L.D. WormBase: a comprehensive data resource for Caenorhabditis biology and genomics. *Nucleic Acids Research* **33**, D383–389 (2005).

[31] Chiaromonte, F., Yap, V.B. &, Miller, W. Scoring pairwise genomic sequence alignments. *Pacific Symposium on Biocomputing* **7**, 115–126 (2002).

[32] Cullen, B.R. Transcription and Processing of Human microRNA Precursors. *Molecular Cell* **16**, 861–865 (2004).

[33] Dallas, J.F. Estimation of microsatellite mutation rates in recombinant inbred strains of mouse. *Mammalian Genome* **3**, 452–456 (1992).

[34] Dieringer, D. & Schlötterer, C. Two distinct modes of microsatellite mutation processes: evidence from the complete genomic sequences of nine species. *Genome Research* **13**, 2242–2251 (2003).

[35] Dietrich, W., Katz, H., Lincoln, S.E., Shin, H.S., Friedman, J., Dracopoli, N.C. & Lander, E.S. A genetic map of the mouse suitable for typing intraspecific crosses. *Genetics* **131**, 423–47 (1992).

[36] Doench, J.G., Sharp, P.A. SiRNAs can function as miRNAs. *Genes & Development* **17**, 438–442 (2003).

[37] Doench, J.G. & Sharp, P.A. Specificity of microRNA target selection in translational repression. *Genes & Development* **18**, 504–511 (2004).

[38] Dupuy, D., Li, Q.R., Deplancke, B., Boxem, M., Hao, T., Lamesch, P., Sequerra, R., Bosak, S., Doucette-Stamm, L., Hope, I.A., Hill, D.E., Walhout, A.J. & Vidal, M. A first version of the Caenorhabditis elegans Promoterome. *Genome Research* **14**, 2169–2175 (2004).

[39] Durbin, R., Eddy S., Krogh A. and Mitchinson, G., Markov chains and hidden Markov models. in *Biological Sequence Analysis*. 46 - 79 (Cambridge University Press, Cambridge, 1998).

[40] Edwards, A., Hammond, H.A., Jin, L., Caskey, C.T. & Chakraborty, R. Genetic variation at five trimeric and tetrameric tandem repeat loci in four human population groups. *Genomics* **12**, 241–253 (1992).

[41] Eigen, M. Selforganization of matter and the evolution of biological macromolecules. *Die Naturwissenschaften* **58**, 465-523 (1971).

[42] Eigen, M. McCaskill, J. & Schuster, P. The molecular quasispecies. *Advances in Chemical Physics* **75**, 149 (1989).

[43] Enright, A.J., John, B., Gaul, U., Tuschl, T., Sander, C. & Marks, D.S. MicroRNA targets in Drosophila. *Genome Biology* **5**, R1 (2003).

[44] Farh, K.K., Grimson, A., Jan, C., Lewis, B.P., Johnston, W.K., Lim, L.P., Burge C.B. & Bartel, D.P. (2005). The Widespread Impact of Mammalian MicroRNAs on mRNA Repression and Evolution. *Science* **310**, 1817–1821 (2005).

[45] Force, A., Cresko, W. A. ,Pickett, F. B., Proulx, S. R., Amemiya, C. & Lynch, M. The Origin of Subfunctions and Modular Gene Regulation. *Genetics* **2005**, 433–446 (2005).

[46] Giraldez, A.J., Cinalli, R.M., Glasner, M.E., Enright, A.J., Thomson, J.M., Baskerville, S., Hammond, S.M., Bartel, D.P. & Schier, A.F. MicroRNAs regulate brain morphogenisis in zebrafish. *Science*, **308**, 833–838 (2005).

[47] The Gene Ontology Consortium. Gene Ontology: tool for the unification of biology. *Nature Genetics* **25**: 25–29 (2000).

[48] Griffiths-Jones, S. The microRNA Registry. *Nucleic Acids Research* **32**, D109–111 (2004).

[49] Grosshans, H., Johnson, T., Reinert, K.L., Gerstein, M., & Slack, F.J. The temporal patterning microRNA let-7 regulates several transcription factors at the larval to adult transition in C. elegans. *Developmental Cell* **8**, 321–330 (2005).

[50] Grün, D., Wang, Y.-L., Langenberger, D., Gunsalus, K.C. & Rajewsky, N. microRNA Target Predictions across Seven Drosophila Species and Comparison to mammalian Targets. *PLoS Computational Biology* **1**, e13 (2005).

[51] Gunsalus, K.C., Yueh, W.C., MacMenamin, P. & Piano, F. RNAiDB and PhenoBlast: web tools for genome-wide phenotypic mapping projects. *Nucleic Acids Research* **32**, D406–410 (2004).

[52] Gunsalus, K.C., Ge, H., Schetter, A.J., Goldberg, D.S., Han, J.D., Hao, T., Berriz, G.F., Bertin, N., Huang, J., Chuang, L.S., Li, N., Mani, R., Hyman, A.A., Sonnichsen, B., Echeverri, C.J., Roth, F.P., Vidal, M. & Piano, F. Predictive models of molecular

machines involved in Caenorhabditis elegans early embryogenesis. *Nature* **436**, 861–865 (2005).

[53] Harr, B., Weiss, S., David, J.R., Brem, G. & Schlötterer, C. A microsatellite-based multilocus phylogeny of the Drosophila melanogaster species complex. *Current Biology* **8**, 1183–1186 (1998).

[54] Harr, B. & Schlötterer, C. Long microsatellite alleles in Drosophila melanogaster have a downward mutation bias and short persistence times, which cause their genome-wide underrepresentation. *Genetics* **155**, 1213–1220 (2000).

[55] He, L., Thomson, J.M., Hemann, M.T., Hernando-Monge, E., Mu, D., Goodson, S., Powers, S., Cordon-Cardo, C., Lowe, S.W., Hannon, G.J. & Hammond, S.M. A microRNA polycystron as a potential human oncogene. *Nature*, **435**, 828–833 (2005).

[56] Herrero, J., Al-Shahrour, F., Daz-Uriarte, R., Mateos, ., Vaquerizas, J.M., Santoyo, J. & Dopazo, J. GEPAS, a web-based resource for microarray gene expression data analysis. *Nucleic Acids Research* **31**, 3461–3467 (2003).

[57] Hobert, O. Common logic of transcription factor and microRNA action. *Trends in Biochemical Sciences* **29**, 426–428 (2004).

[58] Jarne, P., & Lagoda, P.J.L. Microsatellites, from molecules to populations and back. *Trends in Ecology & Evolution* **11**, 424–429 (1996).

[59] John, B., Enright, A.J., Aravin, A., Tuschl, T., Sander, C. & Marks, D.S. Human microRNA targets. *PLoS Biology* **2**, e363 (2004).

[60] Johnson, S.M., Grosshans, H., Shingara, J., Byrom, M., Jarvis, R., Cheng, A., Labourier, E., Reinert, K.L., Brown, D. & Slack F.J. RAS is regulated by the let-7 family. *Cell* **120**, 635–647 (2005).

[61] Johnston, R.J. & Hobert, O. A microRNA controlling left/right neuronal asymmetry in Caenorhabditis elegans. *Nature* **426**, 845–849 (2003).

[62] Kamp, C., Wilke, C. O., Adami, C. & Bornholdt, S. Viral evolution under the pressure of an adaptive immune system - optimal mutation rates for viral escape. *Complexity* **8**, 28–33 (2002).

[63] Kamp, C. & Bornholdt, S. (2002) Co-Evolution of quasispecies: B-cell mutation rates maximize viral error catastrophes. *Physical Review Letters* **88**, 068104 (2002).

[64] Karolchik, D., Hinrichs, A.S., Furey, T.S., Roskin, K.M., Sugnet, C.W., Haussler, D. & Kent, W.J. The UCSC Genome Browser Database. *Nucleic Acids Research* **31**, 51–54 (2003).

[65] Kimura, M. (1983) *The neutral theory of molecular evolution*, (Cambridge: Cambridge U. P.).

[66] Kiontke, K., Gavin, N.P., Raynes, Y., Roehrig, C., Piano, F., & Fitch, D.H. Caenorhabditis phylogeny predicts convergence of hermaphroditism and extensive intron loss. *Proceedings of the National Academy of Science USA* **101**, 9003–9008 (2004).

[67] Kiriakidou, M., Nelson, P.T., Kouranov, A., Fitziev, P., Bouyioukos, C., Mourelatos, Z. & Hatzigeorgiou, A. A combined computational-experimental approach predicts human microRNA targets. *Genes & Development* **18**, 1165–1178 (2004).

[68] Krek, A., Grün, D., Poy, M.N., Wolf, R., Rosenberg, L., Epstein E.J., MacMenamin, P., da Piedade, I., Gunsalus, K.C., Stoffel, M. & Rajewsky, N. Combinatorial microRNA target predictions. *Nature Genetics* **37**, 495–500 (2005).

[69] Kruglyak, S., Durrett, R.T., Schug, M.D. & Aquadro, C.F. Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations. *Proceedings of the National Academy of Sciences USA* **95**, 10774–10778 (1998).

[70] Kruglyak, S., Durrett, R.T., Schug, M.D. & Aquadro, C.F. Distribution and abundance of microsatellites in the yeast genome can Be explained by a balance between slippage events and point mutations. *Molecular Biology and Evolution* **17**, 1210–1219 (2000).

[71] Krützfeldt, J., Rajewsky, N., Braich, R., Rajeev, K.G., Tuschl, T., Manoharan, M. & Stoffel, M. *Nature* **438**, 685-689 (2005).

[72] Kulkarni, M.M. & Arnosti, D.N. Information display by transcriptional enhancers. *Development* **131**, 2419–2429 (2003).

[73] Arnosti, D.N. & Kulkarni, M.M. Transcriptional enhancers: Intelligent enhanceosomes or flexible billboards. *Journal of Cellular Biochemitry* **94**, 890–898 (2005).

[74] Kwiatkowski, D.J., Henske, E.P., Weimer, K., Ozelius, L., Gusella, J.F. & Haines, J. Construction of a GT polymorphism map of human 9q. *Genomics* **12**, 229–40 (1992).

[75] Lai, E.C. & Posakony, J.W. The Bearded box, a novel 3'UTR sequence motif, mediates negative posttranscriptional regulation of Bearded and Enhancer of split complex gene expression. *Development* **124**, 4847–4856 (1997).

[76] Lai, E.C. Micro RNAs are complementary to 3'UTR sequence motifs that mediate negative post-transcriptional regulation. *Nature Genetics* **30**, 363–364 (2002).

[77] Lai, E.C., Tomancak, P., Williams, R.W. & Rubin, G.M. Computational identification of Drosophila microRNA genes. *Genome Biology* **4**, R42 (2003).

[78] Lai, Y. & Sun, F. The relationship between microsatellite slippage mutation rate and the number of repeat units. *Molecular Biology and Evlution* **20**, 2123–2131 (2003).

[79] Lall, S., Grün, D., Krek, A., Chen, K., Wang, Y.L., Dewey, C.N., Sood, P., Colombo, T., Bray, N., MacMenamin, P., Kao, H.L., Gunsalus, K.C., Pachter, L., Piano, F. & Rajewsky, N. A genome wide map of conserved microRNA targets in C. elegans. *Current Biology* **16**, 460–471.

[80] Lecellier, C.H., Dunoyer, P., Arar, K., Lehmann-Che, J., Eyquem, S., Himber, C., Saib, A. & Voinnet, O. A cellular microRNA mediates antiviral defense in human cells. *Science* **308**, 557–560.

[81] Lee, R.C., Feigenbaum, R.L. & Ambros, V. The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell* **75**, 843–854 (1993).

[82] Lewis, B.P., Shih, I.H., Jones-Rhoades, M.W., Bartel, D.P. & Burge, C.B. Prediction of mammalian microRNA targets. *Cell* **26**, 787–798 (2003).

[83] Lewis, B.P., Burge, C.B., Bartel, D.P. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* **120**,15–20 (2005).

[84] Li, W. H. Molecular evolution. *Sinauer Associates*, Sunderland, Mass. (1997).

[85] Li, Y.C., Korol A.B., Fahima, T., Beiles, A. & Nevo, E. Microsatellites: genomic distribution, putative functions and mutational mechanisms: a review. *Molecular Ecology* **11**, 2453–2465 (2002).

[86] Lim, L.P., Lau, N.C., Garrett-Engele, P., Grimson, A. & Schelter, J.M. Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. *Nature* **433**, 769–773 (2005).

[87] Lin, S.Y., Johnson, S.M., Abraham, M., Vella, M.C., Pasquinelli, A., Gamberi, C., Gottlieb, E. & Slack, F.J. The C. elegans hunchback homolog, hbl-1, controls temporal patterning and is a probable microRNA target. *Developmental Cell* **4**, 639–650 (2003).

[88] Liu, J., Carmell, M.A., Rivas, F.V., Marsden, C.G., Thomson, J.M., Song, J.J., Hammond, S.M., Joshua-Tor, L. & Hannon, G.J. Argonaute2 is the catalytic engine of mammalian RNAi. *Science* **305**, 1437–1441 (2004).

[89] Lu, J., Getz, G., Miska, E.A., Alvarez-Saavedra, E., Lamb, J., Peck, D., Sweet-Cordero, A., Ebert, B.L., Mak, R.H., Ferrando, A.A., Downing, J.R., Jacks, T., Horvitz, H.R. & Golub T.R. MicroRNA expression profiles classify human cancers. *Nature*, **435**, 834–838 (2005).

[90] Lynch, M. & Conery, J.S. The origins of genome complexity. *Science* **302**, 1401–1404 (2003).

[91] Messier, W., Li, S.H. & Stewart, C.B. The birth of microsatellites. *Nature* **381**, 483 (1996).

[92] Mignone, F., Grillo, G., Licciulli, F., Iacono, M., Liuni, S., Kersey, P.J., Duarte, J., Saccone, C. & Pesole, G. UTRdb and UTRsite: a collection of sequences and regulatory motifs of the untranslated regions of eukaryotic mRNAs. *Nucleic Acids Research* **33**,141–146 (2005).

[93] Mustonen, V. & Lässig, M. Evolutionary population genetics of promoters: predicting binding sites and functional phylogenies. *Proceedings of the National Academy of Sciences USA* **102**, 15936–15941 (2005).

[94] Nowak, M. & Schuster, P. Error thresholds of replication in finite populations, mutation frequencies and the onset of Muller's ratchet. *Journal of theoretical Biology* **137**, 375-395 (1989).

[95] O'Donnell, K.A., Wentzel, E.A., Zeller, K.I., Dang, C.V. & Mendell J.T. c-Myc-regulated microRNAs modulate E2F1 expression. *Nature* **435**, 839–843.

[96] Ohta, T. Slightly deleterious mutant substitutions in evolution. *Nature* **246**, 96–98 (1973).

[97] Ohta, T. Role of very slightly deleterious mutations in molecular evolution and polymorphism. *Theor. Pop. Biol.* **10**, 254–275 (1976).

[98] Olson, P.H., Ambros, V. The lin-4 regulatory RNA controls developmental timing in Caenorhabditis elegans by blocking LIN-14 protein synthesis after the initiation of translation. *Developmental Biology* **216**, 671–680 (1999).

[99] Pasquinelli, A.E., Reinhart, B.J., Slack, F., Martindale, M.Q., Kuroda, M.I., Maller, B., Hayward, D.C., Ball, E.E., Degnan, B., Muller, P., Spring, J., Srinivasan, A., Fishman, M., Finnerty, J., Corbo, J., Levine, M., Leahy, P., Davidson, E. & Ruvkun, G. Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA. *Nature* **408**, 86–89 (2000).

[100] Petrov, D.A. & Hartl, D.L. High rate of DNA loss in the Drosophila melanogaster and Drosophila virilis species groups. *Molecular Biology and Evolution* **205**, 279–298 (1998).

[101] Petrukhin, K.E., Speer, M.C., Cayanis, E., Bonaldo, M.F., Tantravahi, U., Soares, M.B., Fischer, S.G., Warburton, D., Gilliam, T.C. & Ott, J. A microsatellite genetic linkage map of human chromosome 13. *Genomics* **15**, 76–85 (1993).

[102] Pfeffer, S., Zavolan, M., Grasser, F.A., Chien, M., Russo, J.J., Ju, J., John, B., Enright, A.J., Marks, D., Sander, C. & Tuschl T. Identification of virus-encoded microRNAs. *Science* **304**, 734–736 (2004).

[103] Pfeffer, S., Sewer, A., Lagos-Quintana, M., Sheridan, R., Sander, C., Grasser, F.A., van Dyk, L.F., Ho, C.K., Shuman, S., Chien, M., Russo, J.J., Ju, J., Randall, G., Lindenbach, B.D., Rice, C.M., Simon, V., Ho, D.D., Zavolan, M. & Tuschl, T. Identification of microRNAs of the herpesvirus family. *Nature Methods* **2**, 260–276 (2005).

[104] Pillai, R.S., Bhattacharyya, S.N., Artus, C.G., Zoller T., Courgot, N., Basyuk, E., Bertrand, E. & Filipowicz, W. Inhibition of translational initiation by Let-7 MicroRNA in human cells. *Science*, **309**, 1573–1576 (2005).

[105] Poy, M.N., Eliasson, L., Krutzfeldt, J., Kuwajima, S., Ma, X., Macdonald, P.E., Pfeffer, S., Tuschl, T., Rajewsky, N., Rorsman, P. & Stoffel, M. A pancreatic islet-specific microRNA regulates insulin secretion. *Nature* **432**, 226–230 (2004).

[106] Pruitt, K.D., Tatusova, T. & Maglott, D. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts, and proteins. *Nucleic Acids Research* **33**, D501–D504 (2005).

[107] Ptashne M. Regulated recruitment and cooperativity in the design of biological regulatory systems. *Philos Transact A Math Phys Eng Sci.* **361**, 1223–1234 (2003).

[108] Pupko, T. & Graur, D. Evolution of microsatellites in the yeast Saccharomyces cerevisiae: role of length and number of repeated units. *Journal of Molecular Evolution* **48**, 313–316 (1999).

[109] Rajewsky, N., Vergassola, M., Gaul U. & Siggia, E.D. Computational detection of genomic cis-regulatory modules applied to body patterning in the early Drosophila embryo. *BMC Bioinformatics* **3**, 30 (2002).

[110] Rajewsky, N.& Socci, N.D. Computational identification of microRNA targets. *Developmenal Biology* 267, 529–535 (2004).

[111] Rajewsky, N. Computational microRNA target predictions in animals. *Nature Genetics* **38**, 8–13 (2006).

[112] Rehmsmeier, M., Steffen, P., Hochsmann, M. & Giegerich, R. Fast and effective prediction of microRNA/target duplexes. *RNA* **10**, 1507–1517 (2004).

[113] Reinhart, B.J., Slack, F.J., Basson, M., Pasquinelli, A.E., Bettinger, J.C., Rougvie, A.E., Horvitz, H.R. & Ruvkun, G. The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans. *Nature* **24**, 901–906 (2000).

[114] Robins, H., Li, Y. & Padgett, R.W. Incorporating structure to predict microRNA targets. *Proceedings of the Natational Academy of Science USA* **102**, 4006–4009 (2005).

[115] Rose, O. & Falush, D. A threshold size for microsatellite expansion. *Molecular Biology and Evolution* **15**, 613–615 (1998).

[116] Rougvie, A.D. Keeping time with microRNAs. *Development*, **132**, 3787–3798 (2005).

[117] Rouzine, I.M., Rodrigo, A. & Coffin, J.M. Search for the mechanism of genetic variation in the pro gene of human immunodeficiency virus. *Microbiology and Molecular Biology Reviews* **65**, 151–185 (2001).

[118] Schlötterer, C. & Tautz, D. Slippage synthesis of simple sequence DNA. *Nucleic Acids Research* **20**, 211–215 (1992).

[119] Schlötterer, C., Vogl, C., Tautz, D. Polymorphism and locus-specific effects on polymorphism at microsatellite loci in natural Drosophila melanogaster populations. *Genetics* **146**, 309–320.

[120] Schlötterer, C. & Wiehe, T. Microsatellites, a neutral marker to infer selective sweeps. In: *Microsatellites: Evolution and Applications* (eds Goldstein D.B., Schlötterer, C.), 238–247. Oxford University Press, Oxford (1999).

[121] Schlötterer, C. Evolutionary dynamics of microsatellite DNA. *Chromosoma* **109**, 365–371 (2000).

[122] Schroeder, M.D., Pearce, M., Fak, J., Fan, H., Unnerstall, U., Emberly, E., Rajewsky, N., Siggia, E.D. & Gaul, U. Transcriptional control in the segmentation gene network of Drosophila. *PloS Biology* **2**, e271 (2004).

[123] Schug, M.D., Hutter, C.M., Wetterstrand, K.A., Gaudette, M.S., Mackay, T.F. & Aquadro, C.F. The mutation rates of di-, tri- and tetranucleotide repeats in Drosophila melanogaster. *Molecular Biology and Evolution* **15**, 1751–1760 (1998).

[124] Schwartz, S., Kent, W.J., Smit, A., Zhang, Z., Baertsch, R., Hardison, R.C., Haussler, D. & Miller, W. Human-mouse alignments with BLASTZ. *Genome Research* **13**, 103–107 (2003).

[125] Sempere, L.F., Sokol, N.S., Dubrovsky, E.B., Berger, E.M. & Ambros, V. Temporal regulation of microRNA expression in Drosophila. *Developmental Biology* **259**, 9–18 (2003).

[126] Sibly, R.M., Whittaker, J.C. & Talbot, M. A maximum-likelihood approach to fitting equilibrium models of microsatellite evolution. *Molecular Biology and Evolution* **18**, 413–417 (2001).

[127] Sinha, S., van Nimwegen, E. & Siggia, E.D. A probabilistic method to detect regulatory modules. *Bioinformatics* **19**, i292–301 (2003).

[128] Sinha, S. & Siggia, E.D. Sequence turnover and tandem repeats in cis-regulatory modules in drosophila. *Molecular Biology and Evolution* **22**, 874–858 (2005).

[129] Slack, F.J., Basson, M., Liu, Z., Ambros, V., Horvitz, H.R. & Ruvkun, G. The lin-41 RBCC gene acts in the C. elegans heterochronic pathway between the let-7 regulatory RNA and the LIN-29 transcription factor. *Molecular Cell* **5**, 659–669 (2000).

[130] Sokol, N.S. & Ambros, V. Mesodermally expressed Drosophila microRNA-1 is regulated by Twist and is required in muscles during larval growth. *Genes & Development* **19**, 2343–2354 (2005).

[131] Sood, P., Krek, A., Zavolan, M., Macino, G. & Rajewsky, N. Cell-type specific signatures of microRNAs on target mRNA expression. *Proceedings of the Natational Academy of Science USA* **103**, 2746–2751 (2005).

[132] Stark, A., Brennecke, J., Russell, R.B. & Cohen, S.M. Identification of Drosophila MicroRNA Targets. *PLoS Biology* **1**, e60 (2003).

[133] Stark, A., Brennecke, J., Bushati, N., Russell, R.B. & Cohen, S.M. Animal MicroRNAs Confer Robustness to Gene Expression and Have a Significant Impact on 3'UTR Evolution. *Cell* **123**, 1133–1146 (2005).

[134] Stein, L.D., Bao, Z., Blasiar, D., Blumenthal, T., Brent, M.R., Chen, N., Chinwalla, A., Clarke, L., Clee, C., Coghlan, A., Coulson, A., D'Eustachio, P., Fitch, D.H., Fulton, L.A., Fulton, R.E., Griffiths-Jones, S., Harris, T.W., Hillier, L.W., Kamath, R., Kuwabara, P.E., Mardis, E.R., Marra, M.A., Miner, T.L., Minx, P., Mullikin, J.C., Plumb, R.W., Rogers, J., Schein, J.E., Sohrmann, M., Spieth, J., Stajich, J.E., Wei, C., Willey, D., Wilson, R.K., Durbin, R. & Waterston, R.H. The genome sequence of Caenorhabditis briggsae: a platform for comparative genomics. *PLoS Biology* **1**, e45 (2003).

[135] Sullivan, C.S., Grundhoff, A.T., Tevethia, S., Pipas, J.M. & Ganem, D. SV40-encoded microRNAs regulate viral gene expression and reduce susceptibility to cytotoxic T cells. *Nature* **435**, 682–686 (2005).

[136] Takamizawa, J., Konishi, H., Yanagisawa , K., Tomida, S., Osada, H., Endoh, H., Harano, T., Yatabe, Y., Nagino, M., Nimura, Y., Mitsudomi ,T. & Takahashi T. Reduced expression of the let-7 microRNAs in human lung cancers in association with shortened postoperative survival. *Cancer Research* **64**, 3752–3756.

[137] Tautz, D. & Schlötterer, C. Simple Sequences. *Current Opinion in Genetics and Development* **4**, 832–837 (1994).

[138] Vella, M.C., Choi, E.Y., Lin, S.Y., Reinert, K. & Slack F.J.The C.elegans microRNA let-7 binds to imperfect complementary sites from the lin-41 3'UTR. *Genes & Development* **18**, 132–137 (2004).

[139] Vella, M.C., Reinert, K. & Slack F.J. Architecture of a validated microRNA::target Interaction. *Chemistry & Biology* **11**, 1619–1623 (2004).

[140] Viguera, E., Canceill, D. & Ehrlich, S.D. Replication slippage involves DNA polymerase pausing and dissociation. *The EMBO Journal* **20**, 2587–2595 (2001).

[141] Weber, J.L. & Wong, C. Mutation of human short tandem repeats. *Human Molecular Genetics* **2**, 1123–1128 (1993).

[142] Wheeler, D.L., Barrett, T., Benson, D.A., Bryant, S.H., Canese, K., Chetvernin V, Church, D.M., DiCuccio, M., Edgar, R., Federhen, S., Geer, L.Y., Helmberg, W., Kapustin, Y., Kenton, D.L., Khovayko, O., Lipman, D.J., Madden, T.L., Maglott, D.R., Ostell, J., Pruitt, K.D., Schuler, G.D., Schriml, L.M., Sequeira, E., Sherry, S.T., Sirotkin, K., Souvorov, A., Starchenko, G., Suzek, T.O., Tatusov, R., Tatusova, T.A., Wagner, L. & Yaschenko, E. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research* **33**, D39–45 (2005).

[143] Wienholds, E., Kloostermann, W.P., Miska, E., Alvarez-Saavedra, E., Berezikov, E., de Bruijn, E., Horvitz, H.R., Kauppinen, S. & Plasterk, R.H. MicroRNA expression profiling in zebrafish embryonic development. *Science* **309**, 310–311 (2005).

[144] Wightman, B., Ha, I. & Ruvkun, G. Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in C. elegans. *Cell* **75**, 855–862 (1993).

[145] Xie, X., Kulbokas , E.J., Golub T.R., Mootha , V., Lindblad-Toh, K., Lander, E.S. & Kellis, M. Systematic discovery of regulatory motifs in human promoters and 3 UTRs by comparison of several mammals. *Nature* **434**, 338–345 (2005).

[146] Xu, X., Peng, M. & Fang, Z. The direction of microsatellite mutations is dependent upon allele length. *Nature Genetics* **24**, 396–399.

[147] Xu, P., Vernooy, S.Y., Guo, M. & Hay, B.A. The Drosophila microRNA Mir-14 suppresses cell death and is required for normal fat metabolism. *Current Biology*, **13**, 790–795 (2003).

[148] Yu, Y.K. & Hwa, T. Statistical significance of probabilistic sequence alignment and related local hidden Markov models. *Journal of Computational Biology* **8**, 249–282 (2001).

[149] Zhao, Y., Samal, E. & Srivastava, D. Serum response factor regulates a muscle-specific microRNA that targets Hand2 during cardiogenesis. *Nature* **436**, 214–220 (2005).

Appendix A

Hidden Markov Models

Definition. A hidden Markov model assigns a sequence of states to a sequence $x = (x_1, ..., x_L)$ of symbols $x_i \in \mathcal{B}$. In each state $k \in \{1, ..., N\}$ a symbol b drawn from set $\mathcal{B}$ can be emitted with state specific probabilities $e_k(b)$. Emitting a sequence of symbols, the model can switch from state k to state l with transition probability a_{kl}. Since the same symbols can be emitted in different states, a given sequence x can be explained by a number of different compositions of states, so called *paths* $\pi = (\pi_1, ..., \pi_L)$. Given all transition- and emission-probabilities, the probability of a given path π realizing sequence x, $P(x, \pi)$, can be computed,

$$P(x, \pi) = a_{0\pi_1} \prod_{i=1}^{L} e_{\pi_i} a_{\pi_i \pi_{i+1}}, \tag{A.1}$$

where a_{0k} can be interpreted as the probability of starting in state k. The total probability of the model to explain the sequence x is then given by the sum over the probabilities of all possible paths,

$$P(x) = \sum_{\pi} P(x, \pi). \tag{A.2}$$

The forward algorithm. However, the number of possible paths increases exponentially and simple enumeration of all paths in order to compute (A.2) becomes impossible already for short sequences. A solution to this problem is given by the *forward algorithm*, a dynamic programming strategy that introduces probabilities $f_k(i)$ for the sequence component x_i to be emitted in state k,

$$f_k(i) = P(x_1 ... x_i, \pi_i = k), \tag{A.3}$$

and follows the sequence x from $i = 1$ to $i = L$ in order to recursively compute all the probabilities $f_k(i)$, which finally allow for the computation of $P(x)$ itself:

Initialization ($i = 0$): $f_0(0) = 1$, $f_k(0) = 0$ for $k > 0$.
Recursion ($i = 1, ..., L$): $f_l(i) = e_l(x_i) \sum_k f_k(i-1) a_{kl}$.
Termination: $P(x) = \sum_k f_k(L) a_{k0}$
where a_{k0} is the probability to end in state k.

The backward algorithm. What we really want to compute is not only the overall probability $P(x)$ but the probability that a given sequence position x_i was emitted in a particular state k given the sequence x, $P(\pi_i = k|x)$. This quantity is called the *posterior probability* of state k at sequence position i.
To compute this probability we first need to know the quantity

$$b_k(i) = P(x_{i+1}...x_L|\pi_i = k) \tag{A.4}$$

which denotes the probability of the subsequent path $j = i+1, .., L$ given that position i was emitted in state k. $b_k(i)$ is computed by the *backward algorithm*:

Initialization ($i = L$): $b_k(L) = a_{k0}$ for all k.
Recursion ($i = L-1, ..., 1$): $b_k(i) = \sum_l a_{kl} e_l(x_{i+1}) b_l(i+1)$.
Termination: $P(x) = \sum_l a_{0l} e_l(x_1) b_l(1)$.

The joint probability of observing sequence x with the $i-$th symbol emitted in state k can then simply be written as

$$P(x, \pi_i = k) = f_k(i) b_k(i) \tag{A.5}$$

and the posterior probability of observing state k at position i given sequence x follows as

$$P(\pi_i = k|x) = \frac{f_k(i) b_k(i)}{P(x)}. \tag{A.6}$$

Baum-Welch algorithm. If the parameters a_{kl} and/or $e_k(b)$ are not known in the beginning, they also have to be inferred from the data. One particular standard method to find optimal parameters, i. e. parameters that maximize the log likelihood $\log P(x, \Theta)$ of the model to generate the sequence x is the *Baum-Welch* algorithm [12]. It starts with a given set Θ of parameters a_{kl} and $e_k(b)$. It then counts the number of all particular transition events, A_{kl} and emissions, $E_k(b)$ from the probable paths identified with parameter set Θ and estimates a new set of parameters,

$$\begin{aligned} a_{kl} &= \frac{A_{kl}}{\sum_{l'} A_{kl'}} \\ e_k(b) &= \frac{E_k(b)}{\sum_{b'} E_k(b)}. \end{aligned} \tag{A.7}$$

This iteration proceeds until some stopping criterion is reached. It can be shown that the log likelihood increases by this iteration and hence it converges to a local maximum. Since there are usually many local maxima, the parameters Θ must be initialized carefully to end up in the correct local maximum.
More precisely, the algorithm does not enumerate all transition and emission events but determines their expectation values using the *posterior decoding* after applying the forward- and the backward-algorithm to sequence x,

$$P(\pi_i = k, \pi_{i+1} = l|x, \Theta) = \frac{f_k(i) a_{kl} e_l(x_{i+1}) b_l(i+1)}{P(x, \Theta)}. \tag{A.8}$$

The expected number of state transitions from k to l estimated from sequence x can then be written as

$$A_{kl} = \frac{\sum_i f_k(i) a_{kl} e_l(x_{i+1}) b_l(i+1)}{P(x, \Theta)}. \tag{A.9}$$

Similarly, the expectation value of emitting symbol b in state k is given by

$$E_k(b) = \frac{\sum_{i|x_i=b} f_k(i) b_k(i)}{P(x, \Theta)} \tag{A.10}$$

where the sum is only taken over those positions i where symbol b is emitted. Using these expectation values, the new set of parameters is computed according to (A.7).
Since th algorithm converges in a continuous-valued space, the maximum will never in fact be reached. A possible stopping criterion could be a sufficiently small change of the log likelihood.
Running the Baum-Welch algorithm on a set of sequences $j = 1, ..., M$ it can be summarized as follows:

Initialization:

Selection of arbitrary model parameters.

Recurrence:

Set all the A and E variables to zero or some pseudocount value r.
For each sequence $j = 1, .., M$:

Calculate $f_k(i)$ for sequence j.
Calculate $b_k(i)$ for sequence j
Add the contribution of sequence j to A (A.9) and E (A.10).

Calculate the new model parameters (A.7).
Calculate the new log likelihood of the model.

Termination:

Stop if the change in log likelihood is smaller then some threshold or the maximum number of iterations is reached.

Appendix B

Bayesian approach to model comparison

Given a pair of sequences x and y without any prior knowledge on their relatedness, the Bayesian approach provides a means of quantitatively comparing different models to explain the relation between the two sequences. For convenience, we demonstrate this approach by comparing two models, an uncorrelated background model B with prior probability $P(B)$, postulating that the sequences are unrelated, and a correlated model R with prior probability $P(R)$ that assumes the sequences to be related. If we allow only for the two competing models, their prior probabilities sum up to one, $P(B) + P(R) = 1$. Once we have seen the data, the posterior probability that the correlated model R is correct is

$$\begin{aligned} P(R|x,y) &= \frac{P(x,y|R)P(R)}{P(x,y)} \\ &= \frac{P(x,y|R)P(R)}{P(x,y|R)P(R) + P(x,y|B)P(B)} \\ &= \frac{P(x,y|R)P(R)/P(x,y|B)P(B)}{1 + P(x,y|R)P(R)/P(x,y|B)P(B)} \end{aligned} \tag{B.1}$$

Introducing the quantity

$$S' = S + \log\left(\frac{P(R)}{P(B)}\right), \tag{B.2}$$

where

$$S = \log\left(\frac{P(x,y|R)}{P(x,y|B)}\right) \tag{B.3}$$

is the log-odds score of the alignment, the posterior probability can be written as

$$P(R|x,y) = \sigma(S') \tag{B.4}$$

with the sigmoid *logistic* function

$$\sigma(x) = \frac{e^x}{1+e^x}, \tag{B.5}$$

increasing monotonically from zero to one (fig. B.1). From (B.5) we can see that

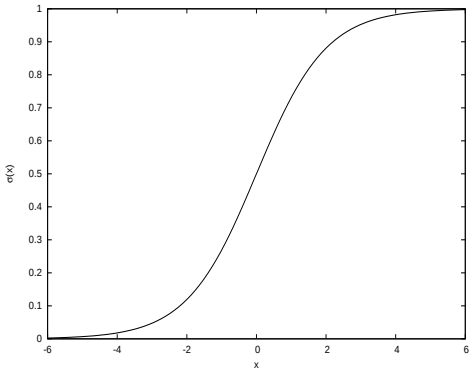


Figure B.1: **The logistic function.**

maximizing S' by testing different correlated models R maximizes the posterior probability of the model R to explain the data (B.4). If the prior probabilities of the models are unknown, one could assign equal probabilities to all tested correlated models and the background model. If a particular model R is then tested against the background model B, one has $P(R) = P(B) = 1/2$ and (B.4) simplifies to

$$P(R|x, y) = \sigma(S) \tag{B.6}$$

Under this assumption the optimal model maximizes the log-odds score of the alignment (B.3).

Appendix C

Derivation of the Fokker-Planck equation

The Fokker-Planck equation can be derived as an approximation of a general master equation. The transition probabilities T are first expressed as a function of the size δx of the jump and of the staring point x':

$$T(x|x') = T(x'; \delta x),\ \delta x = x - x' \tag{C.1}$$

The general master equation

$$\partial_t P(x,t) = \int \{T(x|x')P(x',t) - T(x'|x)P(x,t)\}dx' \tag{C.2}$$

can then be written as

$$\partial_t P(x,t) = \int T(x-\delta x;\delta x)P(x-\delta x,t)d(\delta x) - \int P(x,t)\int T(x;-\delta x)d(\delta x) \tag{C.3}$$

If $T(x';\delta x)$ is now assumed to be a sharply peaked function of δx and at the same time slowly varying with x' and if furthermore $P(x,t)$ is assumed to be a slowly varying function, one can approximate (C.3) by a Taylor expansion of the first integral up to second order:

$$\begin{aligned}\partial_t P(x,t) &= \int T(x;\delta x)P(x,t)d(\delta x) - \int \delta x \partial_x\{T(x;\delta x)P(x,t)\}d(\delta x) \\ &\quad + \frac{1}{2}\int (\delta x)^2 \partial_x^2\{T(x;\delta x)P(x,t)\}d(\delta x) - P(x,t)\int T(x;\delta x)d(\delta x)\end{aligned} \tag{C.4}$$

The first and fourth term cancel out and the remaining two terms can be written with the help of jump moments

$$a_\nu(x) = \int_{-\infty}^{\infty} (\delta x)^\nu T(x;\delta x)d(\delta x) \tag{C.5}$$

yielding

$$\partial_t P(x,t) = -\partial_x\{a_1(x)P(x,t)\} + \frac{1}{2}\partial_x^2\{a_2(x)P(x,t)\}\ . \tag{C.6}$$

This equation is called the Fokker-Planck equation and it keeps only the terms up to $\mathcal{O}(\delta x)^2$ of the full Taylor expansion of the master equation,

$$\partial_t P(x,t) = \sum_{\nu=1}^{\infty} \frac{(-1)^\nu}{\nu!} \partial_x^\nu \{a_\nu(x) P(x,t)\}, \tag{C.7}$$

which is called the *Kramers-Moyal expansion*.

Appendix D

List of μLN for various organisms

organism	μLN
prokaryotes	
Prochlorococcus	10.00
Salmonella enterica	1.5
Legionella pneumophila	1.5
Heliobacter pylori	1.00
Neisseria meningitidis	0.8
Escherichia coli	0.48
Vibrio cholerae	0.4
Enterococcus faecium	0.26
Campylobacter jejuni	0.25
Streptococcus pyogenes	0.13
Pseudomonas aeruginosa	0.11
unicellular eukaryotes	
Tetrahymena thermophila	2.00
Cryptococcus neoformans	0.26
Cryptosporidium parvum	0.24
Saccharomyces cerevisia	0.24
Chlamydomonas reinharditii	0.21
Dictyostelium discoideum	0.19
Neurospora crassa	0.16
Giardia lamblia	0.08
Toxoplasma gondii	0.07
Trypanosoma cruzi	0.06
Encephalitozoon cuniculi	0.04
higher vertebrates	
Pan troglodytes	0.003
Homo sapiens	0.003
Mus musculus	0.003

Acknowledgment

It is my great pleasure to acknowledge the excellent supervision of this doctoral thesis by Prof. Michael Lässig who stimulated my interest in the field of quantitative biology already during my diploma thesis and gave me the freedom to also pursue my own scientific ideas. I am indebted to Prof. Nikolaus Rajewsky who gave me the opportunity to work for one year under his inspiring guidance at the New York University. He introduced me to the exciting field of microRNA research and stimulated my persisting interest in computational biology.
On the scientific side, I further want to thank all members of the Lässig group at the University of Cologne and the Rajewsky lab at NYU. In particular, I want to mention Azra Krek who had already developed the "scaffold" of our microRNA target prediction algorithm at the time I entered this project and who essentially influenced the success of my work on microRNAs.
Beyond the scientific side, I am deeply grateful for the unceasing support of my enchanting wife Natalia who always cheered me up and gave me confidence. Not less important was the help and encouragement of my family. Finally, I want to thank my friends, Brian Cooper, Jan Müller, Tobias Micklitz, Peter Jung, Philipp Messer, Azra Krek and Teresa Colombo for making the past few years a marvelous period of my life.

Zusammenfassung

In dieser Arbeit wird die Behandlung dreier separater biologischer Fragestellungen mit Hilfe aus der statistischen Physik entnommener Methoden vorgestellt.
Ziel des im ersten Kapitel vorgestellten Projektes ist es, ein besseres Verständnis der Posttranskriptionskontrolle zu erlangen, wie sie von sog. *microRNAs*, einer Klasse nicht-kodierender, winziger RNAs, ausgeübt wird. Es wird ein probabilistischer Algorithmus zur Identifizierung von microRNA-regulierten Genen vorgestellt, entwickelt auf Grundlage experimentell erlangter Erkentnisse sowie statistischer Analyse genomischer Daten. Die Anwendung dieses Algorithmus auf Alignments evolutionär verwandter Spezies ermöglicht die spezifische und sensitive Identifizierung durch microRNA regulierter Gene. Analyse und Vergleich dieser Daten für verschiedene Tierstämme und daraus resultierende Einsichten in die microRNA-Biologie sowie erste Erkentnisse über die Evolution von microRNA-Regulation werden ausführlich diskutiert.
Gegenstand des zweiten Kapitels ist die Modellierung der Evolution von *Microsatellites*, einer Klasse repetitiver Sequenz, die häufig im Genom von Eukaryonten anzutreffen ist. Inspiriert durch die Vermutung, dass derartige Sequenzen in bestimmten Fällen funktional sein können und der nach wie vor grossen Schwierigkeit, korrekte Alignments von Microsatellites zu bestimmen, haben wir ein neutrales Modell für Microsatellite-Evolution entwickelt, das in dieser Arbeit diskutiert wird. Um die durch das Modell gemachten Vorhersagen der evolutionären Raten im Genom der Fruchtfliege *Drosophila melanogaster* zu überprüfen, werden diese mit unabängig gemessenen Raten aus Alignments von drei nahe verwandten *Drosophila*-Spezies verglichen. Das Evolutionsmodell wird dabei separat für Sequenzkategorien unterschiedlicher funktionaler Annotierung analysiert, um den möglichen Einfluss von Selektionsdruck auf die Microsatellite-Evolution zu untersuchen.
Im letzten Kapitel wird ein populationsgenetisches Modell entworfen, um die Stabilität von Transkriptionsfaktor-Bindungsstellen als Funktion von Selektionsdruck, Mutationsrate und effektiver Populationsgrösse für beliebige Werte dieser Parameter vorherzusagen. Die analystische Lösung dieses Modells beschreibt die Wahrscheinlichkeit, dass eine gegebene Bindungsstelle funktional ist. Mit Hilfe des Modells wird der Anteil funktionaler Bindungsstellen unter konstantem Selektionsdruck in Populationen verschiedener Taxone bestimmt. Die Interpretation der Ergebnisse zeigt, dass eine Abnahme der effektiven Populationsgrösse, wie beim evolutionären Übergang von Prokaryonten zu Eukaryonten zu beobachten, den Verlust der Stabilität einer Bindungsstelle bedeuten kann. Eine Erweiterung des Modells ermöglicht es, den kompensatorischen Effekt des gleichzeitigen Auftretens einer grösseren Anzahl von Bindungsstellen zu bewerten, der den Erhalt der Stabilität einer gegebenen regulatorischen Beziehung erleichtern kann.

Abstract

In this thesis, three separate problems of genomics are addressed, utilizing methods related to the field of statistical mechanics.
The goal of the project discussed in the first chapter is the elucidation of post-transcriptional gene regulation imposed by *microRNAs*, a recently discovered class of tiny non-coding RNAs. A probabilistic algorithm for the computational identification of genes regulated by microRNAs is introduced, which was developed based on experimental data and statistical analysis of whole genome data. In particular, the application of this algorithm to multiple-alignments of groups of related species allows for the specific and sensitive detection of genes targeted by microRNAs on a genome-wide level. Examination of clade-specific predictions and cross-clade comparison yields deeper insights into microRNA biology and first clues about long-term evolution of microRNA regulation, which are discussed in detail.
Modeling evolutionary dynamics of *microsatellites*, an abundant class of repetitive sequence in eukaryotic genomes, was the objective of the second project and is discussed in chapter two. Inspired by the putative functionality of some of these elements and the difficulty of constructing correct sequence alignments that reflect the evolutionary relationships between microsatellites, a neutral model for microsatellite evolution is developed and tested in the fruit fly *Drosophila melanogaster* by comparing evolutionary rates predicted by the model to independent measurements of these rates from multiple alignments of three closely relates *Drosophila* species. The model is applied separately to genomic sequence categories of different functional annotations in order to assess the varying influence of selective constraint among these categories.
In the last chapter, a general population genetic model is introduced that allows for the determination of transcription factor binding site stability as a function of selection strength, mutation rate and effective population size at arbitrary values of these parameters. The analytical solution of this model indicates the probability of a binding site to be functional. The model is used to compute the population fraction of functional binding sites at fixed selection pressure across a variety of different taxa. The results lead to the conclusion that a decreasing effective population size, such as observed at the evolutionary transition from prokaryotes to eukaryotes, could result in loss of binding site stability. An extension to our model serves us to assess the compensatory effect of the emergence of multiple binding sites for the same transcription factor in order to maintain the existing regulatory relationship.

Erklärung

Ich versichere, dass ich die von mir vorgelegte Dissertation selbständig angefertigt, die benutzten Quellen und Hilfsmittel vollständig angegeben und die Stellen der Arbeit — einschliesslich Tabellen, Karten und Abbildungen, die anderen Werken im Wortlaut oder dem Sinn nach entnommen sind — in jedem Einzelfall als Entlehnung kenntlich gemacht habe; dass diese Dissertation noch keiner anderen Fakultät oder Universität zur Prüfung vorgelegen hat, dass sie — abgesehen von unten angegebenen Teilpublikationen — noch nicht veröffentlicht worden ist sowie, dass ich eine solche Veröffentlichung vor Abschluss des Promotionsverfahrens nicht vornehmen werde. Die Bestimmungen der Promotionsordnung sind mir bekannt. Die von mir vorgelegte Dissertation ist von Herrn Prof. Dr. Michael Lässig betreut worden.

Köln, den 25. September 2006

Teilpublikationen

- Krek, A.*, Grün, D.*, Poy, M.N.*, Wolf, R., Rosenberg, L., Epstein E.J., MacMenamin, P., da Piedade, I., Gunsalus, K.C., Stoffel, M. & Rajewsky, N. Combinatorial microRNA target predictions. *Nature Genetics* **37**, 495–500 (2005).
- Grün, D., Wang, Y.-L., Langenberger, D., Gunsalus, K.C. & Rajewsky, N. microRNA Target Predictions across Seven Drosophila Species and Comparison to mammalian Targets. *PLoS Computational Biology* **1**, e13 (2005).
- Lall, S.*, Grün, D.*, Krek, A., Chen, K., Wang, Y.L., Dewey, C.N., Sood, P., Colombo, T., Bray, N., MacMenamin, P., Kao, H.L., Gunsalus, K.C., Pachter, L., Piano, F. & Rajewsky, N. A genome wide map of conserved microRNA targets in C. elegans. *Current Biology* **16**, 460-471 (2006).
- Grün, D. & Rajewsky, N. Computational prediction of microRNA targets in vertebrates, fruitflies and nematodes. in *microRNAs: From Basic Science to Disease Biology* edited by Dr. Krishnarao Appasani, (Cambridge University Press, Cambridge) to be published.

*equal contributions

Lebenslauf

Persönliche Daten

Name	Dipl.-Phys. Dominic Grün
Anschrift Geburtsdatum	30. September 1977
Geburtsort	Bergisch Gladbach
Familienstand	verheiratet
Staatsangehörigkeit	deutsch

Schulbildung

1984-1988	Grundschule Herkenrath
1988-1997	Gymnasium Herkenrath, Bergisch Gladbach
Juni 1997	Abitur

Hochschulstudium

Oktober 1997	Immatrikulation an der Universität zu Köln Studiengang: Betriebswirtschaftslehre
Oktober 1998	Fachwechsel: Physik (Diplom)
August 2000	Vordiplom
Oktober 2003	Diplom
seit Oktober 2003	Anfertigung der vorliegenden Dissertation

Arbeitsverhältnisse

April 2001–Oktober 2003	Studentische Hilfskraft am Institut für Theoretische Physik
seit Oktober 2003	Wissenschaftlicher Mitarbeiter (ibidem)