Zur Evolution von Short Tandem Repeats

Dissertation
zur Erlangung des Doktorgrades
der Mathematisch-Naturwissenschaftlichen Fakultät
der Christian-Albrechts-Universität zu Kiel

vorgelegt von
Arne Jochens

Kiel 2014

Erster Gutachter:	Prof. Dr. Michael Krawczak
Zweite Gutachterin:	Prof. Dr. Manuela Dittmar
Tag der mündlichen Prüfung:	15.4.2014
Zum Druck genehmigt:	15.4.2014
gez.	Prof. Dr. Wolfgang J. Duschl, Dekan

Summary

Short tandem repeats (STRs) are stretches of DNA that consist of a tandemly repeated short nucleotide sequence. About 3% of the human genome are STRs, and mutations at some of these loci are responsible for certain diseases. Due to their exceptionally high mutation rates, STRs, in particular those on the Y chromosome (Y-STRs), are often used as genetic markers, e.g. in forensics and for paternity tests.

Many of these applications require models of STR evolution. Such models are obtained by combining a genealogical model with a mutation model. Concerning STR mutation, *slipped-strand mispairing* during DNA replication is the most important mechanism. This results in STR mutations usually comprising the gain or loss of one repeat unit. The *Stepwise Mutation Model* (SMM), which augments the genealogical Wright-Fisher model accordingly, is thus the simplest STR mutation model.

Applying Markov chain theory, we characterize the globally wandering but locally coherent behaviour of the allele frequency distribution in a population that is expected under the SMM. To this end, the appropriate allele process X is considered and shown to be a null recurrent Markov chain (global behaviour). On the other hand, a modified version V of this process, obtained by subtracting the allele of a specific individual (or the mean of all alleles) in the current generation from each allele, is positive recurrent (local behaviour). The Markov chain V converges exponentially fast towards a unique unimodal stationary distribution η. It is shown how η can be approximated numerically.

Beyond the SMM, a multitude of STR mutation models exist. Many of these take into account the fact that the STR mutation rate depends on the given allele length. But it is also known that locus-specific factors play a role, too. To accomodate this, we accumulated data on six Y-STR loci in 15,285 father-son pairs from the forensics literature. These data allow the virtually direct observation of 162 mutations, and they are used for a locus-specific comparison of some STR mutation models via a maximum-likelihood approach. Considered are the SMM, the Linear Model (in which the mutation rate increases linearly with the father's allele length) and a novel Logistic Model in various versions. For five of the six Y-STRs, a certain version of our Logistic Model fits best. It is discussed in how far these results are generalizable to autosomal loci.

One example for the practical use of STR evolution models is the estimation of the *random match probability* (RMP) to quantify the evidence of a profile match in forensic DNA typing. This estimation is especially problematic if the profile in question is a Y-STR haplotype that has not been observed previously (a *singleton* haplotype). Regarding this case, we describe several RMP estimators and conduct simulations to compare them. In particular, we consider a coalescent-based method that requires an STR mutation model. Although this estimator is computationally demanding, it performs best otherwise.

Finally, the question whether STRs possess an evolutionary function is discussed. This question hinges on the proportion of functional sequences in the human genome. The importance of these considerations is underlined by the fact that in many countries, e.g. Germany, the law allows only neutral markers to be used for identification purposes.

Zusammenfassung

Short Tandem Repeats (STRs) sind DNA-Abschnitte, in denen sich eine kurze Nukleotid-Sequenz aufeinanderfolgend wiederholt. Das Genom des Menschen besteht zu etwa 3% aus solchen Sequenzen, und Mutationen an einigen dieser Loci sind für bestimmte Krankheiten verantwortlich. Wegen ihrer sehr hohen Mutationsraten werden STRs, insbesondere solche auf dem Y-Chromosom (Y-STRs), oft als genetische Marker verwendet, z.B. in der Forensik und für Vaterschaftstests.

Viele dieser Anwendungen setzen Modelle für die STR-Evolution voraus. Solche Modelle bestehen aus der Kombination eines genealogischen Modells mit einem Mutationsmodell. Der für STRs wichtigste Mutationsmechanismus ist das *Slipped-Strand Mispairing* während der DNA-Replikation. Dies führt dazu, dass eine STR-Mutation in der Regel im Gewinn oder Verlust einer Repeat-Einheit besteht. Daraus ergibt sich als einfachstes STR-Mutationsmodell das *Stepwise Mutation Model* (SMM), welches das genealogische Wright-Fisher-Modell um einen entsprechenden Mutationsprozess erweitert.

Mittels Markov-Ketten-Theorie charakterisieren wir das global wandernde, aber lokal kohärente Verhalten der Allelfrequenzen-Verteilung in einer Population, das unter dem SMM erwartet wird. Dazu wird der entsprechende Allelprozess X betrachtet, der sich als nullrekurrente Markov-Kette erweist (globales Verhalten). Eine modifizierte Version V dieses Prozesses, bei der von jedem Allel das Allel eines bestimmten Individuums (oder der Mittelwert aller Allele) der jeweiligen Generation subtrahiert wird, ist dagegen positiv rekurrent (lokales Verhalten). Die Markov-Kette V konvergiert exponentiell schnell gegen eine eindeutige unimodale stationäre Verteilung η. Es wird gezeigt, wie man η numerisch approximieren kann.

Über das SMM hinaus existiert eine Vielzahl an STR-Mutationsmodellen. Viele davon berücksichtigen die Tatsache, dass die STR-Mutationsrate von der vorhandenen Allel-Länge abhängt. Bekannt ist aber auch, dass locus-spezifische Faktoren ebenfalls eine Rolle spielen. Um dies zu berücksichtigen, haben wir zu sechs Y-STR-Loci die Daten von 15.285 Vater-Sohn-Paaren aus der forensischen Literatur zusammengetragen. Dies ermöglicht die quasi direkte Beobachtung von 162 Mutationen. Mit einem Maximum-Likelihood-Ansatz werden anhand dieser Daten das SMM, das Lineare Modell (in dem die Mutationswahrscheinlichkeit li-

near mit der Allel-Länge des Vaters zunimmt) und ein neues Logistisches Modell in verschiedenen Versionen locus-spezifisch miteinander verglichen. Für fünf der sechs betrachteten Y-STRs zeigt sich, dass eine bestimmte Version unseres Logistischen Modells am besten passt. Es wird diskutiert, inwieweit diese Ergebnisse auf autosomale Loci übertragbar sind.

Ein Beispiel für die praktische Anwendung von STR-Evolutionsmodellen ist das Schätzen der *Random Match Probability* (RMP) für die Quantifizierung der Evidenz einer Profilübereinstimmung im Rahmen forensischer DNA-Analysen. Dieses ist bei zuvor unbekannten Y-STR-Profilen (*Singleton*-Haplotypen) besonders problematisch. Für diesen Fall werden verschiedene RMP-Schätzer beschrieben und simulationsbasiert miteinander verglichen. Insbesondere betrachten wir eine koaleszenzbasierte Methode, die ein STR-Mutationsmodell voraussetzt. Dieser Schätzer ist zwar sehr rechenaufwändig, erweist sich den anderen Methoden ansonsten aber als überlegen.

Abschließend wird diskutiert, ob STRs eine evolutionäre Funktion besitzen, bzw. wie groß der Anteil funktioneller Sequenzen am menschlichen Genom ist. Dies ist deshalb bedeutsam, weil z.B. in Deutschland nur neutrale Marker für Identifikationszwecke verwendet werden dürfen.

Inhaltsverzeichnis

Tabellenverzeichnis

Abbildungsverzeichnis

Vorwort

> „Ever since their discovery in the early 1980s, the ubiquitous occurrence of microsatellites — also referred to as short tandem repeats (STRs) or simple sequence repeats (SSRs) — has puzzled geneticists. Why are they so common? Do they fulfill some function or are they simply junk DNA sequences that should perhaps be viewed as 'selfish DNA' [Ref. auf Doolittle & Sapienza 1980 [52] und Orgel & Crick 1980 [186]]? Addressing these questions is important if we wish to understand how genomes are organized and why most genomes are filled with sequences other than genes."
>
> Hans Ellegren [61, S. 435]

In der vorliegenden Arbeit nähern wir uns der Evolution von Short Tandem Repeats (STRs) über deren Mutationsverhalten. Dazu werden in Kapitel 1 die wichtigsten Konzepte und Probleme eingeführt. Genauer gesagt werden dort zunächst die Definitionen von STRs und verwandten Begriffen behandelt (Abschnitt 1.1). Dann werden zur Motivation der folgenden Arbeiten einige praktische Aspekte von STR-Mutationen erläutert (Abschnitt 1.2), insbesondere ein Beispiel einer populationshistorischen Studie und Morbus Huntington als Beispiel einer STR-assoziierten Krankheit. Da wir nicht nur Mutationen, sondern ganze Populationen betrachten wollen, enthält Abschnitt 1.3 eine Beschreibung zweier grundlegender genealogischer Modelle, nämlich des Wright-Fisher-Modells und der Koaleszenztheorie. Darauf folgt ein Überblick, was über das Mutationsverhalten von STRs bekannt ist, sowie eine Zusammenfassung interessanter STR-Mutationsmodelle (Abschnitt 1.4). Die populationsgenetischen Implikationen des einfachsten dieser Modelle sind bereits eingehend mathematisch analysiert worden, wie Abschnitt 1.5 zeigt.

Der Verwendung von STRs als forensische Marker ist Abschnitt 1.6 gewidmet. Wir betrachten die Geschichte des genetischen Fingerabdrucks und das Konzept der *Random Match Probability* (RMP), zunächst in allgemeiner Form. Dann wenden wir uns dem Y-Chromosom zu, dessen Eigenschaften zusammengefasst werden. Auch dieses Chromosom enthält STR-Loci (Y-STRs), von denen die praktisch bedeutsamsten vorgestellt werden. Die Rolle von Vaterschaftstests in Bezug

auf Y-STR-Mutationsdaten wird erläutert. Schließlich wird die RMP noch einmal aufgegriffen und in den Y-STR-Kontext eingebunden.

Den Kern dieser Arbeit bilden drei Veröffentlichungen, die in Kapitel 2 (ab S. 33) vollständig wiedergegeben werden. Dort ist diesen Publikationen jeweils eine deutsche Zusammenfassung vorangestellt. Hier werden zunächst nur ihre bibliographischen Daten und die Original-Abstracts zitiert. Im vorliegenden Text wird mit „Publikation (i)“ usw. auf diese Veröffentlichungen Bezug genommen.

(i) Caliebe[1], Jochens[1], Krawczak und Rösler (2010)
A Markov Chain Description of the Stepwise Mutation Model: Local and Global Behaviour of the Allele Process.
Journal of Theoretical Biology 266:2, 336–342.

Abstract:
The stepwise mutation model (SMM) is a simple, widely used model to describe the evolutionary behaviour of microsatellites. We apply a Markov chain description of the SMM and derive the marginal and joint properties of this process. In addition to the standard SMM, we also consider the normalised allele process. In contrast to the standard process, the normalised process converges to a stationary distribution. We show that the marginal stationary distribution is unimodal. The standard and normalised processes capture the global and the local behaviour of the SMM, respectively.

(ii) Jochens, Caliebe, Rösler und Krawczak (2011)
Empirical Evaluation Reveals Best Fit of a Logistic Mutation Model for Human Y-chromosomal Microsatellites.
Genetics 189:4, 1403–1411.

Abstract:
The rate of microsatellite mutation is dependent upon both the allele length and the repeat motif, but the exact nature of this relationship is still unknown. We analysed data on the inheritance of human Y-chromosomal microsatellites in father-son duos, taken from 24 published reports and comprising 15,285 directly observable meioses. At the six microsatellites analysed (DYS19, DYS389 I, DYS390, DYS391, DYS392 and DYS393), a total of 162 mutations were observed. For each locus, we employed a maximum likelihood approach to evaluate one of several single-step mutation models on the basis of the data. For five of the six loci considered, a novel logistic mutation model was found to provide the best fit according to Akaike's Information Criterion. This implies that the mutation probability at the loci increases (non-linearly) with allele length at a rate that differs between upward and downward mutations. For DYS392, the best fit was provided by a

[1] geteilte Erstautorenschaft

linear model in which upward and downward mutation probabilities increase equally with allele length. This is the first study to empirically compare different microsatellite mutation models in a locus-specific fashion.

(iii) Andersen, Caliebe, Jochens, Willuweit und Krawczak (2013)
Estimating Trace-Suspect Match Probabilities for Singleton Y-STR Haplotypes Using Coalescent Theory.
Forensic Science International: Genetics 7:2, 264–271.

Abstract:
Estimation of match probabilities for singleton haplotypes of lineage markers, i.e. for haplotypes observed only once in a reference database augmented by a suspect profile, is an important problem in forensic genetics. We compared the performance of four estimators of singleton match probabilities for Y-STRs, namely the count estimate, both with and without Brenner's so-called 'kappa correction', the surveying estimate, and a previously proposed, but rarely used, coalescent-based approach implemented in the BATWING software. Extensive simulation with BATWING of the underlying population history, haplotype evolution and subsequent database sampling revealed that the coalescent-based approach is characterized by lower bias and lower mean squared error than the uncorrected count estimator and the surveying estimator. Moreover, in contrast to the two count estimators, both the surveying and the coalescent-based approach exhibited a good correlation between the estimated and true match probabilities. However, although its overall performance is thus better than that of any other recognized method, the coalescent-based estimator is still computation-intense on the verge of general impracticability. Its application in forensic practice therefore will have to be limited to small reference databases, or to isolated cases of particular interest, until more powerful algorithms for coalescent simulation have become available.

Kapitel 3 enthält Resultate, die zuvor nicht veröffentlicht wurden. Diese Resultate lassen sich thematisch entweder Publikation (i) oder (ii) zuordnen und sind dementsprechend auf die Abschnitte 3.1 und 3.2 verteilt.

In Kapitel 4 werden die vorangegangenen Resultate diskutiert. Insbesondere wird auf einige Y-STR-spezifische Aspekte eingegangen (Abschnitt 4.1) und ein Ausblick gegeben, welche Arbeiten sich in Zukunft anschließen können (Abschnitt 4.2). Abschließend diskutieren wir die im obigen Ellegren-Zitat angesprochenen Fragen (Abschnitt 4.3). Nach dem Literaturverzeichnis folgen dann noch ein Verzeichnis der im Text verwendeten Abkürzungen und die Danksagung.

Kapitel 1

Einleitung

1.1 Short Tandem Repeats

Als *Short Tandem Repeats* (STRs) werden solche Abschnitte eines Genoms bezeichnet, die aus der aufeinanderfolgenden Wiederholung einer kurzen Nukleotidsequenz bestehen [65]. Diese einfache Sequenz heißt das *Repeat-Motiv* und kann nach der hier verwendeten Definition aus zwei bis sieben Basenpaaren bestehen. Damit folgen wir der Definition des US-amerikanischen *National Institute of Standards and Technology* (NIST) [25, S. 148]. Die Anzahl der Wiederholungen (plus eins) wird als *Repeat-Länge*, *Allel-Länge* oder kurz als Allel bezeichnet und kann bis zu etwa 100 betragen [219]. In der älteren Literatur werden die Allele manchmal auch anhand der Gesamtlänge des bei der Typisierung herausgeschnittenen bzw. amplifizierten Fragments (siehe Abschnitt 1.6.1) benannt, aber in dieser Arbeit bezieht sich *Allel* stets auf die Repeat-Länge. Die International Society of Forensic Genetics (ISFG) gibt Richtlinien heraus [88], in denen unter anderem die Nomenklatur für forensisch genutzte STRs (insbesondere Loci- und Allelnamen) geregelt wird (siehe Abschnitt 1.6.4).

In der älteren Literatur findet sich auch die Abkürzung *STRP* (wobei das P für Polymorphismus steht) als Bezeichnung für einen gegebenen STR-Locus im Genom, insbesondere wenn dieser eine besonders hohe interindividuelle Variabilität aufweist. Manche Autoren benutzen auch STR (ohne P) in diesem engeren Sinne. Desweiteren werden die Begriffe *Simple* oder *Short Sequence Repeat* (SSR) in der Literatur in der Regel synonym mit STR verwendet. Wenn eine Teilmenge von STRs anhand gleich langer Repeat-Motive identifiziert werden soll, wird oft die entsprechende griechische Vorsilbe verwendet. So bezeichnet man als *Tetranukleotid-Repeats* alle STRs, deren Repeat-Motiv aus genau vier Basenpaaren besteht.

STRs sind Teil einer breiteren Klasse von repetitiven Sequenzen, die allgemein als *Variable Number of Tandem Repeats* (VNTR) bezeichnet werden. Die Gemeinsamkeit aller VNTRs ist die aufeinanderfolgende Wiederholung eines Repeat-

Motivs. Anhand der Länge des Repeat-Motivs werden Subklassen unterschieden. Dazu gehören neben den STRs auch Minisatelliten (Repeat-Motive von acht bis hundert Basenpaaren) und Satelliten (noch längere Repeat-Motive, vor allem in Zentromeren und Heterochromatin zu finden). Wegen dieser Verwandtschaft werden STRs auch als *Mikrosatelliten* bezeichnet. Diese beiden Begriffe werden in der vorliegenden Arbeit synonym verwendet. In der Literatur finden sich auch Abweichungen von den hier verwendeten Definitionen, z.B. Repeat-Motive von ein bis sechs Basenpaaren für STRs [65], was verdeutlicht, dass die Abgrenzungen zwischen den VNTR-Subklassen unscharf sind [219]. Zudem wird manchmal die Bezeichnung VNTR als Synonym für Minisatelliten verwendet.

Der Begriff Satelliten-DNA geht auf den russischen Botaniker Sergei Nawaschin zurück, der 1912, bevor das Wort in der Raumfahrttechnik Verwendung fand, die morphologisch vom Rest des Chromosoms durch eine zweite Einschnürungsstelle abgegrenzten Enden mancher Chromosomen als *Sputnik* bezeichnete (siehe Gregory [85], der dazu Battaglia [11] zitiert). Als in den 1960-er Jahren bei der Anwendung der Caesiumchlorid-Zentrifugation auf Zell-DNA-Proben schmale Bänder bemerkt wurden, deren Dichte (wegen des hohen AT-Gehalts der meisten VNTRs) in der Regel geringer als die der übrigen DNA war, wurde auch dafür der Begriff Satelliten-DNA geprägt [85].

1.2 Motivation

Nahezu alle bislang sequenzierten Eukaryoten-Genome sind reich an STRs, die in der Regel über das gesamte Genom verteilt sind. Auch das menschliche Genom bildet hier mit seinen circa 380.000 STRs in der haploiden Referenzsequenz (build 36) keine Ausnahme [165, Table 3]. Diese Zahl schwankt je nach Definition und Detektions-Algorithmus, aber die Varianz der Schätzungen ist bemerkenswert niedrig. So ergibt z.B. eine Extrapolation der Daten aus dem ersten Entwurf der Humangenomsequenz [150, Table 14] auf eine Genomlänge von 2,9 Giga-Basenpaaren (Gb) circa 350.000 STRs. Man kann davon ausgehen, dass das Humangenom zu etwa 3% aus STR-Sequenzen besteht [5, Table 1.1].

Die sehr hohen Mutationsraten (siehe Abschnitt 1.4) und die damit verbundene große interindividuelle Variabilität machen STRs zu sehr nützlichen genetischen Markern. Die Anwendungen reichen dabei von der Genkartierung [234, 47] über populationshistorische Studien (siehe Abschnitt 1.2.1), „genetische Fingerabdrücke“ (siehe Abschnitt 1.6.1), Verwandtschaftsanalysen (z.B. Vaterschaftstests, siehe Abschnitt 1.6.5) und genetische Ahnenforschung [207] bis zur genetischen Epidemiologie [111, 110].

Wie John Butler bemerkt [25, S. 6 und 16], ist es wahrscheinlich, dass STRs noch lange die wichtigsten Marker für forensische Anwendungen bleiben, weil bereits sehr umfangreiche Datenbanken mit entsprechenden Profilen existieren (siehe Abschnitt 1.6.2). Über die Anwendung als Werkzeug hinaus sind STRs

auch an sich von wissenschaftlichem Interesse, weil sie z.B. eine wichtige Rolle in der Genomevolution spielen könnten [154, 116]. Außerdem sind einige STR-Mutationen beim Menschen für bestimmte Krankheiten verantwortlich (siehe Abschnitt 1.2.2). Aber auch wenn in einer bestimmten Bevölkerung ein spezieller STR-Locus eventuell evolutionär neutral ist (d.h. keiner natürlichen Selektion unterliegt), bzw. wenn man dies mittels eines statistischen Tests prüfen möchte, sind Modelle für die STR-Evolution unabdingbar, um die erwartete Verteilung der Allele unter der Nullhypothese (keine Selektion) zu bestimmen.

Auf zwei der genannten Punkte soll in den folgenden Unterabschnitten beispielhaft etwas näher eingegangen werden.

1.2.1 Populationshistorische Studien

STRs werden oft als Marker in populationshistorischen Studien verwendet, d.h. für die Rekonstruktion von demographischer Entwicklung und Migrationsbewegungen, insbesondere solche menschlicher Populationen [202, 81, 111, 110]. Eine besondere Rolle spielt dabei häufig das Y-Chromosom (siehe dazu Abschnitt 1.6.3), aber in diesem Abschnitt betrachten wir autosomale Marker.

Ein klassisches Beispiel für solche Studien ist die Arbeit von Marek Kimmel und Kollegen [124], siehe auch [123]. In einem Teil dieser Arbeit gingen die Autoren von STR-Daten extanter menschlicher Populationen [112, 113] aus und verwendeten Simulationen unter einem Modell mit Bottleneck und anschließend (nach verschiedenen Szenarien) wachsender Bevölkerung, Migration und schrittweisen Mutationen (siehe Abschnitt 1.4.2) der STRs, um die *Out-of-Africa*-Hypothese im Kontrast zum multiregionalen Modell zu evaluieren. (Für eine Diskussion der verschiedenen Modelle zum Ursprung moderner Menschen siehe [110, Chapter 8.4].)

Als Statistiken wählten sie die genetische Varianz, definiert als Erwartungswert der quadrierten Differenz zwischen zwei zufällig aus der Population gezogenen Allelen (wobei mit Allel die Anzahl der Wiederholungen des Repeat-Motivs gemeint ist, siehe Abschnitt 1.1), sowie die Homozygotie, definiert als die Wahrscheinlichkeit, dass zwei zufällig aus der Population gezogene Allele übereinstimmen. Diese beiden Statistiken können einerseits aus den vorliegenden Stichproben-Daten der extanten Populationen geschätzt werden. Andererseits kann unter der Annahme, dass die Population sich in Bezug auf die betrachteten Allelfrequenzen im Gleichgewicht befindet, die genetische Varianz auch als Produkt aus Populationsgröße und Mutationsrate, $N\mu = \theta$, berechnet werden, und die Homozygotie als $1/\sqrt{1+2\theta}$, wie Kimmel [123] mit Hilfe der Koaleszenztheorie (siehe Abschnitt 1.3.2) zeigt. Man erhält somit zwei Schätzwerte für θ, $\hat{\theta}_1$ (berechnet über die genetische Varianz) und $\hat{\theta}_2$ (berechnet über die Homozygotie). Die dafür erforderlichen Simulationen erfolgen unter der Nullhypothese, die hier z.B. besagt, dass N dem durch das Modell vorgegebenen Wachstum folgt. Ein eventueller Unterschied zwischen $\hat{\theta}_1$ und $\hat{\theta}_2$ liefert dann also Evidenz gegen

die Nullhypothese. Die Ergebnisse sind in Tabelle 1.1 zusammengefasst.

Population	$\ln(\hat{\theta}_1/\hat{\theta}_2)$	P-Wert
Asiaten	$0,60$	$< 0,01$
Europäer	$0,29$	$< 0,05$
Afrikaner	$0,11$	$0,30$

Tabelle 1.1: Zusammenfassung einiger Ergebnisse von Marek Kimmel und Kollegen [124, 123]. Die Schätzwerte der Allelhäufigkeiten sind über 60 STR-Loci, allesamt Tetranukleotid-Repeats, gemittelt (weltweite Daten von Jorde et al. [112, 113]). Die Verteilung von $\hat{\theta}_1/\hat{\theta}_2$ unter der Nullhypothese zur Bestimmung der P-Werte wurde mittels Simulationen berechnet.

Wie Kimmel [123] ausführt, sind diese Werte mit der *Out-of-Africa*-Hypothese konsistent, da sie sich durch ein Populations-Bottleneck erklären lassen, das für Afrikaner am weitesten zurückliegt und für Asiaten am rezentesten ist, während Europäer einen Wert dazwischen aufweisen.

Im Hinblick auf die vorliegende Arbeit ist festzuhalten, dass für die beschriebene Methodik ein STR-Mutationsmodell unabdingbar ist, da die Evolution unter der Nullhypothese simuliert werden muss, um die P-Werte zu berechnen. Dies ist nur ein Beispiel von vielen, in denen die Mutation von STRs von Bedeutung ist. Abschnitt 1.4.2 gibt einen Überblick, welche STR-Mutationsmodelle existieren, und in Publikation (ii) vergleichen wir einige dieser Modelle.

1.2.2 Pathogene STRs

Beim Menschen und in einigen Tiermodellen kann die Expansion (d.h. das mutationsbedingte Auftreten sehr langer Allele) bestimmter STRs, typischerweise solcher mit Trinukleotid-Repeat-Motiv, bestimmte oft neurodegenerative Erkrankungen auslösen [4, 200, 43, 74, 187, 30, 145, 176]. Diese STRs liegen in der Regel in proteincodierenden Sequenzen, oder zumindest in der Nähe von Genen, und sie sind auch in gesunden Menschen polymorph. Wenn aber die Repeat-Länge durch Mutation eine bestimmte Schwelle überschreitet, manifestiert sich die Erkrankung. Die klinischen Symptome sind dabei oft nicht von Geburt an sichtbar, sondern erst nach einigen Jahren. Das Eintrittsalter ist typischerweise mit der Allel-Länge negativ korreliert. Außerdem unterliegen die Mutationen an solchen Loci in der Regel einem Expansions-Bias [188], d.h. eine Mutation führt in mehr als der Hälfte aller Fälle zu einem längeren Allel. Dies erklärt das schon vor der Aufklärung des molekularen Hintergrundes bekannte Phänomen der *Antizipation*: In betroffenen Familien sinkt oft das Eintrittsalter von Generation zu Generation, und die Symptome werden ausgeprägter. Einige Beispiele für solche Erkrankungen werden in Tabelle 1.2 aufgelistet.

			Repeat-Anzahl		
Chr.	Protein	Motiv	normal	pathogen	Krankheit
4	Huntingtin	CAG	10–30	>35	Chorea Huntington
6	Ataxin 1	CAG	6–39	41–81	Spinozerebelläre Ataxie Typ 1
9	Frataxin	GAA	7–34	>100	Friedreich-Ataxie
12	Atrophin	CAG	7–25	49–75	Dentatorubro-Pallidoluysische Atrophie
14	Ataxin 3	CAG	12–37	61–84	Spinozerebelläre Ataxie Typ 3
19	Myotonin-Proteinkinase	CTG	5–37	50–3000	Myotone Dystrophie Typ 1
22	Ataxin 10	ATTCT	≈14	>4000	Spinozerebelläre Ataxie Typ 10
X	FMR-1	CGG	6–52	6–1000	Fragiles-X-Syndrom
X	Androgen-rezeptor	CAG	11–33	38–66	Kennedy-Krankheit

Tabelle 1.2: Beispiele für pathogene STRs. Modifiziert nach Avise [5, Table 4.2]. Alle aufgeführten STRs liegen entweder direkt in einem proteincodierenden Gen, oder in einer einem solchen Gen zugeordneten regulatorischen Sequenz. Angegeben ist jeweils das Chromosom, das von dem betroffenen Gen codierte Protein, das Repeat-Motiv des jeweiligen STRs, die minimalen und maximalen beobachteten Allele (aufgeteilt in den phänotypisch unauffälligen und den pathogenen Bereich), sowie der Name der Erkrankung.

Wie bereits diese kleine Auswahl zeigt, sind viele dieser Erkrankungen insbesondere mit CAG-Repeat-Expansionen assoziiert. Diejenigen der in Tabelle 1.2 angegebenen STRs, die ein CAG-Repeat-Motiv haben, liegen darüber hinaus alle in proteincodierenden Sequenzen [5, Table 4.2], siehe auch [200, Table 7.1] und [145, Table 1]. CAG codiert für Glutamin, die entsprechenden Erkrankungen werden also durch besonders lange Polyglutamin-Ketten innerhalb des jeweiligen Proteins verursacht [240]. Allerdings ist noch nicht geklärt, ob das Polyglutamin an sich toxisch wirkt, diesen Erkrankungen also neben dem molekulargenetischen auch ein gemeinsamer pathologischer Mechanismus zugrundeliegt, oder ob die Wirkung langer Polyglutamin-Ketten für jedes betroffene Protein spezifisch ist [176, S. 829].

Eine der bekanntesten Polyglutamin-Erkrankungen ist der Morbus Huntington (*Huntington disease*, HD oder auch Chorea Huntington; für Reviews siehe [10], [198], [237] und [178]). HD wird autosomal-dominant vererbt und manifestiert sich in der Regel erst im mittleren Alter. Die klinischen Symptome umfassen die charakteristische Chorea (unwillkürliche krampfhafte Bewegungen der Glieder und Gesichtsmuskeln), Dystonie, Sprachschwierigkeiten, Appetitlosigkeit, Erinnerungs- und andere kognitive Defizite sowie dementsprechende Persönlichkeitsveränderungen, Depression, bulbare Dysfunktion, Gehirnatrophie und schließlich unweigerlich den Tod, in der Regel 15 bis 20 Jahre nach Eintreten der ersten Symptome [93, 187, 178]. Die Prävalenz beträgt weltweit etwa 1/20.000 [93], variiert aber geographisch sehr stark [178].

Das HD-Gen konnte bereits 1983 durch Linkage-Analysen grob lokalisiert werden [87]; es liegt am Ende des kurzen Arms von Chromosom 4 (4p16.3). 1993 gelang es dann, das Gen und den darin enthaltenen STR (dessen besonders lange Allele HD verursachen) zu isolieren [101]. Es handelt sich um ein proteincodierendes Gen, dessen Produkt Huntingtin getauft wurde. Dieses Protein ist relativ groß (348 ku) und wird lebenslang exprimiert, besonders im zentralen Nervensystem [187]. Seine Struktur ist die eines multifunktionellen Gerüstproteins [2, 218, 159], und in der Tat interagiert es mit sehr vielen anderen Proteinen [79]. Die HD-Pathogenese ist, soweit bekannt, dementsprechend komplex [187, S. 592-596], [198]. Vermutlich führen die besonders langen Polyglutamin-Ketten, die durch Expansions-Mutationen des CAG-STRs entstehen, zu einem Funktionsgewinn des Huntingtins [200, S. 85].

Die schon im allgemeinen Kontext erwähnte negative Korrelation zwischen der Repeat-Anzahl des längeren der zwei Allele eines Patienten und dem Eintrittsalter ist für HD besonders gut dokumentiert [54, 210, 3, 201, 236, 152]. Eine unter genetischen Erkrankungen äußerst seltene Eigenschaft der HD ist die vollständige Dominanz: Patienten, die homozygot für Allele im pathogenen Bereich sind, lassen sich phänotypisch nicht von heterozygoten Patienten unterscheiden [238]. Darüber hinaus hat bei Heterozygoten die Länge des kürzeren Allels keinen Einfluss auf das Eintrittsalter [152]. Eine weitere Besonderheit ist, dass besonders große Expansions-Mutationen fast ausschließlich bei den Nachkommen betroffe-

ner Väter auftreten [141]. Auf all diesen molekularen Erkenntnissen aufbauend werden zur Zeit verschiedene Therapieansätze erforscht [187, 198, 237]. Bislang verlief aber noch keine klinische Phase-III-Studie einer neuroprotektiven Therapie erfolgreich [178].

Generell lässt sich über das Mutationsverhalten pathogener STRs sagen, dass es für normale Allel-Längen im Wesentlichen dem von gewöhnlichen STRs entspricht (siehe dazu Abschnitt 1.4). Die besonders langen pathogenen Allele haben indes extrem erhöhte Mutations-Wahrscheinlichkeiten und -Beträge [200, Section 7.3]. Hier wirken also besondere Mechanismen (vgl. Abschnitt 1.4.1); insbesondere spielen die DNA-Reparaturprozesse eine wichtige Rolle [189]. Obwohl diese Mutationsmechanismen sehr komplex sind, gibt es Ansätze für Mutationsmodelle, die speziell auf pathogene STRs zugeschnitten sind, siehe dazu [102] und [226].

1.3 Genealogische Modelle

Genealogische Modelle sind stochastische Prozesse, die eine Beschreibung der Verwandschaftsbeziehungen in einer Population im zeitlichen Verlauf liefern. In diesem Abschnitt werden zwei der wichtigsten solchen Modelle vorgestellt, nämlich das Wright-Fisher-Modell sowie der koaleszenzbasierte Ansatz.

1.3.1 Das Wright-Fisher-Modell

Das auf Sewall Wright [246] und Ronald Fisher [69, 1. Auflage 1930] zurückgehende Modell bildet die Grundlage für viele populationsgenetische Modelle. Auch wir gehen in den Publikationen (i) und (iii) von diesem Modell aus. Es beschreibt in seiner einfachsten Form zunächst einen Locus in einer Population von haploiden Individuen, für den folgende Annahmen gelten:

(i) konstante, endliche Populationsgröße N;

(ii) diskrete, nicht überlappende Generationen;

(iii) keine Mutationen, keine Selektion, keine Migration und keine Populations-Substruktur.

Die Idee des Modells ist, dass jede Generation aus der vorherigen durch N-maliges (unabhängiges) „Ziehen mit Zurücklegen“ hervorgeht. Dies ist äquivalent zu der zeitlich rückwärts gerichteten Betrachtungsweise, dass sich die Individuen einer Generation ihre Vorfahren mit Zurücklegen ziehen. Den Individuen einer jeden Generation werden die Zahlen Eins bis N zugeordnet. Die Reihenfolge spielt dabei keine Rolle, d.h. die Individuen einer Generation sind im mathematischen Sinne *austauschbar*. In Publikation (i) bezeichnen wir mit $Y_n(i)$ den direkten Vorfahren (bzw. dessen Nummer) des i-ten Individuums der n-ten Generation, d.h. die $Y_n(i)$ sind unabhängig identisch verteilte Zufallsgrößen, die jeweils mit gleicher

Wahrscheinlichkeit eine der natürlichen Zahlen von 1 bis N als Wert annehmen. $(Y_n)_{n\in\mathbb{N}} = (Y_n(1), \ldots, Y_n(N))_{n\in\mathbb{N}}$ nennen wir den genealogischen Prozess.

Die Anzahl $D_n(i)$ der Nachkommen des i-ten Individuums der n-ten Generation ist also binomialverteilt. Die Anzahlen der Nachkommen aller Individuen einer Generation folgen somit einer symmetrischen Multinomialverteilung, d.h. für alle $n \in \mathbb{N}$, $d_1, \ldots, d_N \in \mathbb{N}_0$ mit $\sum_{i=1}^{N} d_i = N$ gilt:

$$P(D_n = (d_1, \ldots, d_N)) = \frac{1}{N^N} \cdot \frac{N!}{d_1! \cdots d_N!}. \tag{1.1}$$

Wie z.B. Warren Ewens [66, S. 20] anmerkt, kann man den Allelprozess im Wright-Fisher-Modell als Markov-Kette auffassen. Das liegt daran, dass die Allelverteilung in jeder neuen Generation nur von der Allelverteilung in der vorherigen Generation abhängt. Falls z.B. an dem betrachteten Locus genau zwei Allele (A_1 und A_2) vorkommen, so ist die bedingte Wahrscheinlichkeit, dass in Generation $n+1$ genau j-mal das Allel A_1 vorkommt, gegeben dass es in der n-ten Generation i-mal vorkommt:

$$p_{ij} = \binom{N}{j} \left[\frac{i}{N}\right]^j \left[\frac{N-i}{N}\right]^{N-j} \tag{1.2}$$

D.h. die Übergangswahrscheinlichkeiten p_{ij} der Markov-Kette der Allelhäufigkeiten sind binomialverteilt.

Dies ist nur eines von vielen Beispielen für die Nützlichkeit der Markov-Ketten-Theorie in der Populationsgenetik. Viele klassische Resultate wurden so nachträglich noch eleganter bewiesen und miteinander verknüpft [220]. Zur Anwendung von Markov-Prozessen auf STR-Mutationsmodelle siehe insbesondere Watkins [229]. Auch wir verwenden in Publikation (i) Markov-Ketten-Theorie, um das Verhalten des Allelprozesses unter dem einfachsten STR-Mutationsmodell zu analysieren.

1.3.2 Koaleszenztheorie

Die Koaleszenztheorie ist ein wichtiger Pfeiler der modernen Populationsgenetik [50, 71, 180, 197, 96]. Sie beruht auf der einfachen aber mächtigen Idee, die Genealogie einer Population nicht zeitlich vorwärts zu modellieren (wobei man jedes Individuum erfasst), sondern stattdessen von einer gegenwärtigen Stichprobe von Individuen (oder Genen) ausgehend zeitlich rückwärts deren Genealogie zu betrachten. Wenn bei dieser Rückwärtsbetrachtung zu einem Zeitpunkt zwei Individuen einen gemeinsamen Vorfahren haben, fasst man sie entsprechend zusammen und spricht von einem Koaleszenzereignis (nach dem englischen *to coalesce* — verschmelzen). Auf diese Weise kann man die gesamte Genealogie für die Stichprobe simulieren. Die Koaleszenzereignisse finden dabei zufällig statt, mit einer Wahrscheinlichkeit (pro Generation, im diskreten Modell) bzw. Rate (im stetigen Modell), die im einfachsten Fall nur von der Anzahl der verbleibenden

Linien und der Populationsgröße abhängt. Ein so definierter stochastischer Prozess heißt *Coalescent* [66, Chapter 10].

Diese Ideen wurden bereits in den 1970-er Jahren angedacht [136], unter anderem im Zusammenhang mit P.A.P. Morans Arbeiten über wandernde Verteilungen (siehe Abschnitt 1.5 und Publikation (i)). Aber erst 1982 wurden sie dann von J.F.C. Kingman ausformuliert [135, 132, 133].

Der Koaleszenzansatz hat zwei offensichtliche Vorteile: Erstens muss nicht die gesamte Population simuliert und gespeichert werden, sondern nur die Stichprobe bzw. die verbleibenden Linien. Zweitens braucht man keine willkürlich gewählte lange Zeit zu simulieren, um einen Gleichgewichtszustand zu erreichen, sondern man simuliert genau so weit, bis alle Linien zusammengefallen sind. Das damit erreichte Individuum heißt letzter gemeinsamer Vorfahre (*most recent common ancestor*, MRCA). In der Regel kann der Mutationsprozess als unabhängig von der Genealogie angenommen werden. In diesem Fall können nach der Koaleszenzsimulation die Mutationen simuliert werden, indem dem MRCA ein Allel zugeordnet wird und Mutationsereignisse auf den Zweigen des Koaleszenzbaumes mit konstanten Raten, also mit Wahrscheinlichkeiten proportional zu den Zweiglängen, stattfinden.

In unserer Publikation (i) über wandernde Verteilungen benötigen wir die Koaleszenztheorie nicht, aber andere Autoren haben wandernde Verteilungen (siehe Abschnitt 1.5) auch mit Hilfe von Koaleszenzsimulationen erforscht (siehe z.B [192]). Für unsere Publikation (iii) ist die Koaleszenztheorie unabdingbar, wie in Abschnitt 1.6.6 noch näher erläutert wird.

1.4 Mutation von STRs

STRs weisen mit etwa 10^{-6} bis 10^{-2} Mutationen pro Locus und Generation sehr hohe Mutationsraten auf [157]. Für einzelne Loci wurden sogar schon Mutationsraten von bis zu 0,07 geschätzt [8]. Eine STR-Mutation besteht in der Regel im Gewinn oder Verlust einer Repeat-Einheit, aber es sind auch schon größere Änderungen der Allel-Länge von einer Generation zur nächsten beobachtet worden (siehe z.B. Publikation (ii), Table 3).

Eine Mutation, die zu einem längeren Allel führt, bezeichnen wir als *Aufwärtsmutation*. Ist das resultierende Allel kürzer, sprechen wir von einer *Abwärtsmutation*. Unter Umständen können Aufwärts- und Abwärtsmutationen im Mittel als nahezu gleich wahrscheinlich erscheinen [8], aber siehe dazu unsere Publikation (ii).

Es gibt viele Hinweise auf eine Beschränkung der durchschnittlichen Allel-Länge an bestimmten STR-Loci (z.B. [73]). Desweiteren ist es gut dokumentiert, dass die STR-Mutationsrate

(i) von Locus zu Locus stark variiert, was zum Teil an der Länge und Zusammensetzung des Repeat-Motivs liegt [119, 8];

(ii) von der jeweiligen Repeat-Anzahl abhängt [253, 146, 119, 8, 216];

(iii) mit dem Alter des Vaters zunimmt [8, 216].

In Publikation (ii) haben wir Daten zur Mutation einiger Y-STRs beim Menschen aus der forensischen Literatur zusammengetragen und dann einige Mutationsmodelle bezüglich dieser Daten miteinander verglichen. Hier werden zunächst in Abschnitt 1.4.1 zwei Mutationsmechanismen vorgestellt, die für repetitive Sequenzen besondere Bedeutung haben können. In Abschnitt 1.4.2 wird dann ein Überblick gegeben, welche Modelle für STR-Mutationen existieren.

1.4.1 Mutationsmechanismen

Slipped-Strand Mispairing

Während der DNA-Replikation kann es vorkommen, dass (i) die bereits in der Replikationsgabel getrennten nativen Stränge wieder aneinander binden oder (ii) ein naszenter Strang sich kurzzeitig vom nativen Strang löst und wieder anbindet [140, Section 3.2.2]. In beiden Fällen kann es dabei besonders in repetitiven Bereichen zu einer Schleifenbildung in einem der Stränge kommen (siehe Abbildung 1.1). Wenn nun die zellulären Reparaturenzyme eine solche Schleife herausschneiden oder der andere Strang entsprechend erweitert wird, um die Schleife auszugleichen, resultiert das in einer Mutation, sofern kein weiterer Fehlerkorrekturmechanismus eingreift [140, Section 3.2.3]. Dieses Phänomen heißt *Slipped-Strand Mispairing* (SSM) oder *Replication Slippage* [154].

Vieles deutet darauf hin, dass SSM für STRs in der Regel der dominierende Mutationsmechanismus ist [153, 206]. Das ist auch mit der Tatsache konsistent, dass sich STR-Mutationsraten beim Menschen nicht systematisch zwischen Autosomen und Geschlechtschromosomen unterscheiden [61]. Der in Abschnitt 1.2.2 behandelte Expansions-Bias pathogener STRs resultiert wahrscheinlich daraus, dass SSM-Ereignisse in CAG-Repeats unterschiedliche Reparaturaussichten haben, je nachdem in welchem Strang sich die Schleife gebildet hat [188].

Unequal Recombination

Wenn sich während der Meiose zwei Chromatide überkreuzen, kann es neben dem gewöhnlichen Crossing-Over auch vorkommen, dass der Kreuzungspunkt in den beiden Sequenzen an verschiedenen Stellen liegt. Das Resultat ist dann eine sogenannte ungleiche Rekombination, bei der die eine Sequenz länger und die andere kürzer wird.

Zwar liegt die Vermutung nahe, dass dieser Mechanismus für alle repetitiven Sequenzen stark zur Mutationsrate beiträgt, aber während die ungleiche Rekombination bei Minisatelliten-Mutationen tatsächlich eine wesentliche Rolle spielt [21], gibt es dafür bei STRs kaum Evidenz [61].

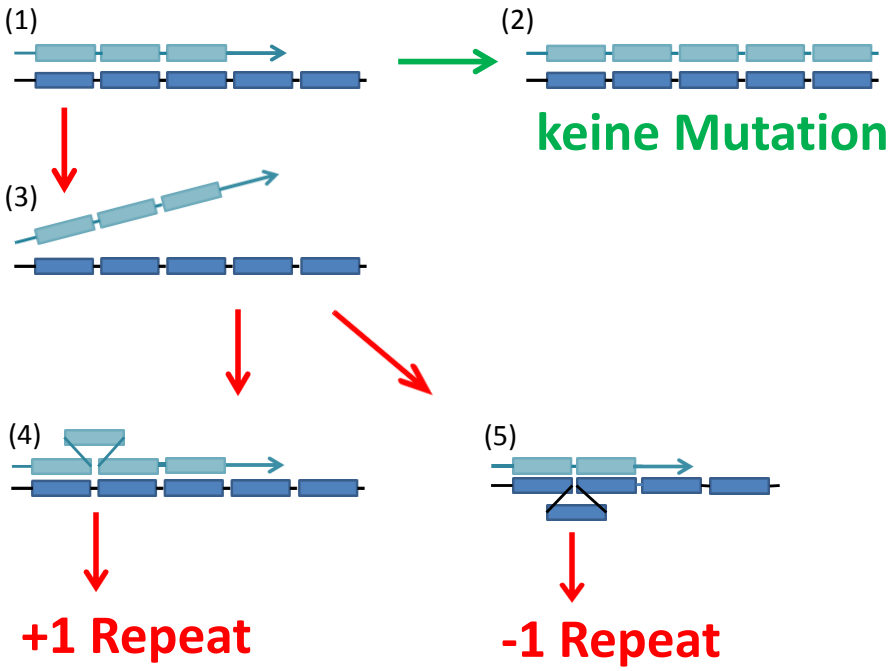

Abbildung 1.1: Schematische Darstellung des Slipped-Strand Mispairing (SSM). Links oben ist ein STR während der DNA-Replikation skizziert (1). Jedes Kästchen symbolisiert das Repeat-Motiv. Im Normalfall wird die Sequenz originalgetreu repliziert, und es gibt keine Mutation (2). Wenn sich jedoch der naszente Strang während der Replikation vom nativen Strang löst (3), dann kann es passieren, dass sich bei der Wiederanlagerung eine Schleife bildet. Dies wird oft durch Reparaturmechanismen noch korrigiert (nicht gezeigt). Wenn letzteres aber ausbleibt, was für Schleifen, die aus nur einem Repeat-Motiv bestehen besonders wahrscheinlich ist, dann kommt es zu einer Mutation, deren Betrag der Länge der Schleife entspricht. Dabei handelt es sich um eine Aufwärts-Mutation, wenn die Schleife sich im naszenten Strang befindet (4) und um eine Abwärts-Mutation, wenn die Schleife sich im nativen Strang befindet (5).
Modifiziert nach Jonathan Eisen [60, Fig. 4.1].

1.4.2 Mutationsmodelle

Dieser Abschnitt soll einen Überblick geben, welche stochastischen Modelle in der Literatur für STRs vorgeschlagen, bzw. für die Modellierung von STR-Mutationen verwendet wurden. Weitere Details zu vielen dieser Modelle werden von Calabrese und Sainudiin [27] sowie Watkins [229] angegeben. In Anbetracht der biologischen Komplexität ist es nicht zu erwarten, dass irgendeines dieser Modelle den STR-Mutationsprozess perfekt beschreibt. Stattdessen sollte jedes Modell als eine Approximation der Realität betrachtet werden, wobei natürlich in einem gegebenen Szenario manche Modelle besser geeignet sein werden als andere. In diesem Sinne vergleichen wir in Publikation (ii) einige der folgenden Modelle anhand von Y-STR-Mutationsdaten (siehe auch Abschnitt 1.6.5).

Zunächst betrachten wir Modelle mit diskreter Zeit, ausgehend vom Wright-Fisher-Modell (Abschnitt 1.3.1).

Einfaches Stepwise-Mutation-Modell

Das einfachste und beliebteste aller für STRs geeigneten Mutationsmodelle ist das (einfache) schrittweise Mutationsmodell (*Stepwise Mutation Model*, SMM). Wir können es als Erweiterung des Wright-Fisher-Modells (siehe Abschnitt 1.3.1) um den folgenden Mutationsmechanismus definieren: Jedem Individuum ist ein Allel $z \in \mathbb{Z}$ zugeordnet, und beim Übergang von einer Generation zur nächsten kommt es mit einer Wahrscheinlichkeit von je $\mu/2$ zu einer Mutation $+1$ bzw. -1. Der Parameter μ wird oft als Mutationsrate bezeichnet, obwohl er eigentlich eine Wahrscheinlichkeit und keine Rate ist, da es sich um ein zeitlich diskretes Modell handelt.

Das SMM stammt von Ohta und Kimura [183], die es für durch Elektrophorese unterscheidbare Allele entwickelten. Daraufhin untersuchte P.A.P. Moran [168, 169] das Modell mathematisch. Er entwickelte dabei die Theorie der wandernden Verteilungen (siehe Abschnitt 1.5). In Publikation (i) bauen wir darauf auf.

Später wurde das SMM auch und gerade für STRs benutzt. Es ist in diesem Zusammenhang immer noch das am häufigsten verwendete Modell. Die Werte $z \in \mathbb{Z}$ sind dann die Allel-Längen (siehe Abschnitt 1.1), bzw. deren normierte Werte, z.B. jeweils abzüglich der am häufigsten beobachteten Allel-Länge (siehe Publikation (i)).

In den folgenden Abschnitten werden einige Erweiterungen bzw. Modifikationen des einfachen SMM betrachtet.

SMM mit zirkulärem Zustandsraum

Neben dem einfachen SMM in der oben angegebenen Form untersuchten Ohta und Kimura mittels Computersimulationen auch die folgende Variante des einfachen SMM [183, 184]: Die Menge der möglichen Allele, also der Zustandsraum der Markov-Kette, ist endlich: $\{1, \ldots, n\}$. Zusätzlich zu den einfachen Mutationen zu

benachbarten Allelen treten auch Mutationen von n nach 1 und umgekehrt jeweils mit Wahrscheinlichkeit $\mu/2$ auf. Das einfache SMM ist dann der Grenzfall für $n \to \infty$. Moran [168, 169] unterzog dieses Modell einer kurzen theoretischen Betrachtung.

SMM mit Beschränkung der Allel-Länge

Wie bereits erwähnt wurde, kann die Allel-Länge von STRs nach oben beschränkt sein. Ursache kann ein Mutations-Bias sein (siehe oben), oder auch Selektion, die gegen Allele jenseits einer bestimmten Länge wirkt. Eine untere Schranke für die Allel-Länge ist offensichtlich realistisch, da STRs mit negativer Länge nicht existieren. Aber auch für eine untere Schranke, die größer ist als 1 gibt es biologische Argumente [196].

Nauta et al. [174] schlugen ein einfaches SMM mit fester unterer und oberer Schranke für die Allel-Längen vor. Feldman et al. [68] untersuchten dasselbe Modell genauer.

Asymmetrisches SMM

Als asymmetrisches SMM bezeichnen wir die Version des einfachen SMM, in der die Aufwärts- (μ_u) ungleich der Abwärts-Mutationswahrscheinlichkeit (μ_d) sein kann. Cooper et al. [40] verwendeten dieses Modell in Verbindung mit Koaleszenzsimulationen, ausgehend von realen Daten.

Das asymmetrische SMM ist eines von den Modellen, die wir in Publikation (ii) betrachten.

Multistep-Modelle

Als *Multistep-Modelle* bezeichnen wir Varianten des SMM, in denen Mutationen das Allel auch um mehr als eine Einheit ändern können. Den Betrag einer Mutation, also die Änderung der Allel-Länge in Repeat-Einheiten, nennen wir dabei die *Schrittlänge*. Ein solches Modell wurde bereits 1975 erstmals analysiert [19]. Auch Moran betrachtete kurz zumindest den Spezialfall mit Mutationen ± 1 oder ± 2 [168]. Im *Zwei-Phasen-Modell* [46] wird das einfache SMM mit geometrisch verteilten größeren Mutationen kombiniert. Dieses Modell wurde dann noch wie folgt verallgemeinert: Gegeben, dass eine Mutation stattfindet, unterliegt die Schrittlänge einer beliebig festgelegten Verteilung mit Erwartungswert 0 und fester Varianz [73].

Für krankheitsverursachende lange Trinukleotid-Wiederholungen spielen, wie in Abschnitt 1.2.2 erläutert, Multistep-Mutationen eine wesentliche Rolle, und dafür gibt es eigene Modelle, siehe z.B. [102] und [226].

Modelle mit Längen-Bias

Motiviert durch die zu Beginn von Abschnitt 1.4 zitierten Beobachtungen wurden mehrere Mutationsmodelle entwickelt, die sich vom einfachen SMM dadurch unterscheiden, dass die Schrittlänge oder die Mutationswahrscheinlichkeit von der gegenwärtigen Allel-Länge abhängt.

Eine Idee ist, dass es eine bestimmte „Zielgröße“ für die Allel-Längen gibt, zu der Mutationen tendieren. Dies modellieren Garza et al. [73] folgendermaßen: Die Allel-Längen seien so normiert, dass die Zielgröße 0 ist. Wenn ein Allel der Länge a mutiert, unterliegt die Änderung seiner Länge einer Verteilung mit Erwartungswert $-\beta a$ und fester Varianz. Es liegt also ein Ornstein-Uhlenbeck-Prozess vor. Zhivotovsky et al. [254] unterziehen dasselbe Modell einer genaueren mathematischen Untersuchung. Kimmel et al. [125] lassen noch allgemeiner alle Variationen des SMM zu, bei denen die Änderung der Allel-Länge einer beliebigen Verteilung unterliegt.

Später wurden dann verschiedene Modelle vorgeschlagen, in denen auch die Mutationswahrscheinlichkeit selbst nicht mehr konstant ist, sondern von der Allel-Länge abhängt. Ein frühes Beispiel dafür ist die Arbeit von Falush und Iwasa [67]. Auch die meisten Modelle (mit Ausnahme des einfachen SMM), die wir in Publikation (ii) untersuchen, gehören zu dieser Gruppe. Da diese Modelle die gemeinsame Eigenschaft haben, dass für jede Generation die Wahrscheinlichkeitsverteilung der Allel-Änderung (hier -1, +1 oder 0) nur von der vorherigen Allel-Länge abhängt, lassen sie sich als (zeitlich homogene) Markov-Ketten auf der Menge der möglichen Allele (hier z.B. die Menge der ganzen Zahlen $\mathbb{Z}$) auffassen. Das heißt insbesondere, dass wir diese Modelle vollständig spezifizieren können, indem wir für alle Allele $i, j \in \mathbb{Z}$ die Übergangswahrscheinlichkeiten p_{ij} angeben. Für jedes Modell ist p_{ij} die bedingte Wahrscheinlichkeit, dass ein gegebenes Allel i in einer Generation zu Allel j mutiert (bzw. nicht mutiert falls $i = j$). Im folgenden betrachten wir keine Multistep-Modelle, d.h. für alle $i, j \in \mathbb{Z}$ mit $|j - i| > 1$ gilt stets $p_{ij} = 0$.

Als *lineares Modell* bezeichnen wir in Publikation (ii) den Fall, dass die Aufwärts- und Abwärtsmutationswahrscheinlichkeit jeweils eine lineare Funktion der Allel-Länge ist. Dieses Modell wurde in anderer Formulierung schon von Kruglyak [143] benutzt. Um es sauber definieren zu können, wählen wir ein festes $k \in \mathbb{N}$ und beschränken den Zustandsraum auf $S = \{1, 2, \dots k\}$. Für gegebene Parameter $\nu_u, \nu_d \in (0, \frac{1}{2k})$ sind dann die Übergangswahrscheinlichkeiten

$$p_{ij} = \begin{cases} i \cdot \nu_u & \text{falls} \quad j = i + 1, \\ i \cdot \nu_d & \text{falls} \quad j = i - 1, \\ 1 - i \cdot (\nu_u + \nu_d) & \text{falls} \quad j = i, \end{cases}$$

für alle $i \in S \backslash \{1, k\}, j \in S$. Und außerdem gilt für die Grenzen: $p_{12} = \nu_u$, $p_{11} = 1 - \nu_u$, $p_{k,k-1} = k \cdot \nu_d$ und $p_{kk} = 1 - k \cdot \nu_d$. Als *symmetrisches lineares Modell* bezeichnen wir den Spezialfall $\nu_u = \nu_d = \nu$.

Als *exponentielles Modell* bezeichnen wir das folgende, von Whittaker et al. 2003 [239] vorgeschlagene Modell. Es hat die Parameter $\alpha_u, \alpha_d, \gamma_u, \gamma_d, \lambda_u$ und λ_d. Die Übergangswahrscheinlichkeiten sind

$$p_{ij} = \begin{cases} \gamma_u \cdot \exp(\alpha_u \cdot i - \lambda_u \cdot (j-i)) & \text{falls} \quad j > i, \\ \gamma_d \cdot \exp(\alpha_d \cdot i - \lambda_d \cdot (i-j)) & \text{falls} \quad j < i, \\ 1 - \sum_{k \neq i} p_{ik} & \text{falls} \quad j = i. \end{cases}$$

In dieser Form ist das Modell allerdings nicht wohldefiniert, da es für gewisse Parameter-Werte zu Wahrscheinlichkeiten größer als Eins führen kann, wie wir in Publikation (ii) zeigen. Dies war ein Grund für die Entwicklung des folgenden Modells.

In Publikation (ii) führen wir das folgende *logistische Modell* ein, das zuvor in der Literatur noch nicht betrachtet wurde. Es hat in seiner allgemeinen Form den Zustandsraum $S = \mathbb{Z}$ und die Parameter $\alpha_u, \alpha_d \in \mathbb{R}$, $\beta_u, \beta_d \in \mathbb{R}^+$ und $\gamma_u, \gamma_d \in (0, 0.5)$. Die Übergangswahrscheinlichkeiten sind

$$p_{ij} = \begin{cases} \gamma_u/(1 + \exp(\alpha_u \cdot (\beta_u - i))) & \text{falls} \quad j = i+1, \\ \gamma_d/(1 + \exp(\alpha_d \cdot (\beta_d - i))) & \text{falls} \quad j = i-1, \\ 1 - \gamma_u/(1 + \exp(\alpha_u \cdot (\beta_u - i))) - \gamma_d/(1 + \exp(\alpha_d \cdot (\beta_d - i))) & \text{falls} \quad j = i. \end{cases}$$

Die Rollen der verschiedenen Parameter sind in Abbildung 1.2 illustriert. In Publikation (ii) passen wir dieses Modell an Y-STR-Mutationsdaten an.

Punktmutationen

Kruglyak, Durrett et al. [143] erweiterten das SMM um die zusätzliche Berücksichtigung von selten auftretenden Punktmutationen innerhalb der repetitiven Sequenz. In ihrem Modell wird nach einem solchen Ereignis nur einer der beiden durch die neue Punktmutation getrennten Repeat-Teile weiter verfolgt, so dass sich dann die Allel-Länge entsprechend reduziert. Auch dieses Modell lässt sich als Markov-Kette formulieren, aber im Gegensatz zu den oben behandelten Modellen ist die Zeit hier stetig. Der Zustandsraum ist $S = \mathbb{N}$, und die Parameter sind die Punktmutationsrate α, sowie β, das die STR-typische Mutationsrate determiniert, und schließlich die Grenzrate γ. Es gibt drei möglichen Typen von Übergängen (aus einem gegebenen Zustand $k \in \mathbb{N}$):

(i) ± 1 mit Rate $(k-1) \cdot \beta$,

(ii) Übergang zu $1, 2, \ldots k-1$ jeweils mit Rate α,

(iii) falls $k = 1$: Übergang zu 2 mit Rate γ.

Durrett und Kruglyak [53] zeigten, dass in ihrem Modell der Allelprozess (siehe Abschnitt 1.5) ohne Normierung konvergiert. Calabrese et al. [28] betrachteten

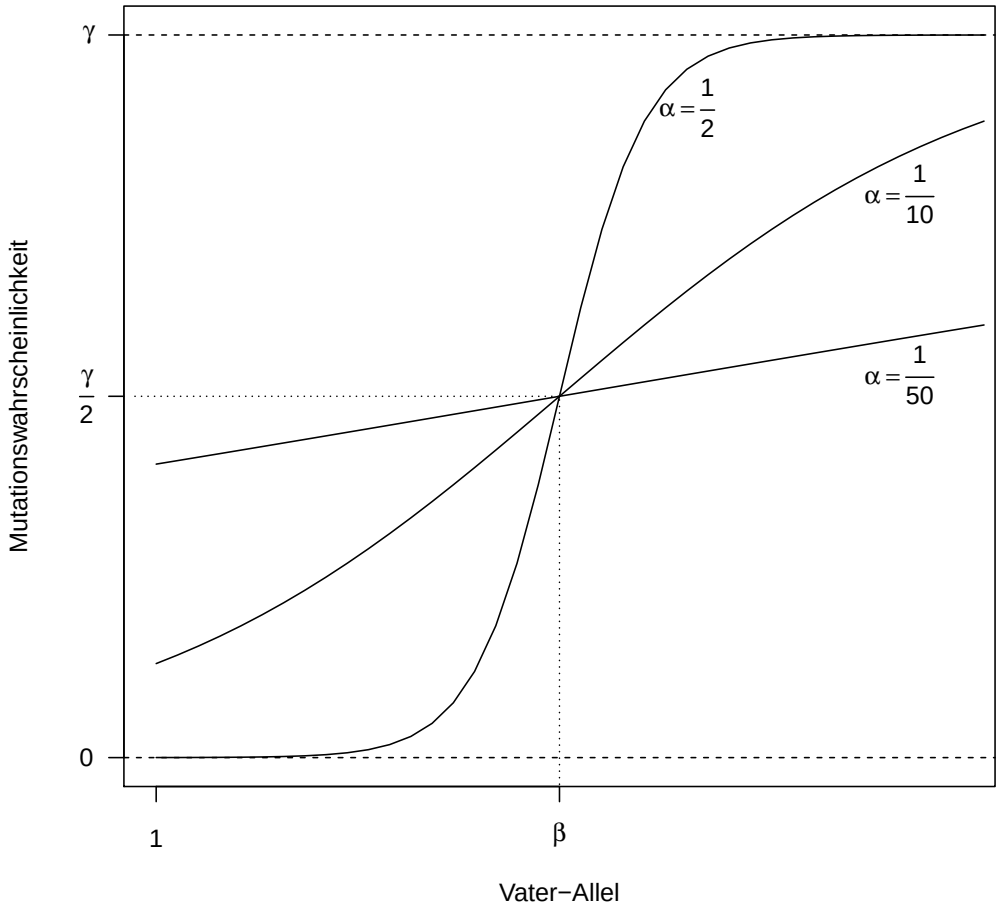

Abbildung 1.2: Das Logistische Modell und seine Parameter. Dargestellt ist die Aufwärts-Mutationswahrscheinlichkeit als Funktion der Allel-Länge des Vaters. Die hier fest zugrundegelegten Parameter-Werte sind $\gamma = 0,1$ und $\beta = 20$. Dagegen wurde α zwischen drei verschiedenen Werten variiert, um den Einfluss dieses Parameters auf die Form der Kurve zu illustrieren.

eine etwas verfeinerte Version dieses Modells, das sogenannte *PS/PM-Modell*, und untersuchten auch den Spezialfall mit Punktmutationsrate 0 (*PS/0M-Modell*). Außerdem führten sie das *PCR-Modell* ein, bei dem nach einer Punktmutation beide Teil-STRs weiterverfolgt werden. Dieringer und Schlötterer [48] empfehlen, das PS/PM-Modell um einen zusätzlichen Mutationsprozess mit konstanter Rate für kurze Allele zu erweitern, was sie als *Indel Slippage* bezeichnen.

Verallgemeinerungen

J.F.C. Kingman [134] betrachtete die folgende Verallgemeinerung des SMM: Die Allele können beliebige reelle Zahlen sein ($S = \mathbb{R}$), und die Mutationen werden jeweils durch eine beliebige Wahrscheinlichkeitsverteilung auf $\mathbb{R}$ gegeben. Kingman zeigte [134], dass dann der Allelprozess (siehe Abschnitt 1.5) immer noch eine Markov-Kette ist, aber keine stationäre Verteilung hat. (Denn eine solche müsste translationsinvariant auf $\mathbb{R}$ sein, könnte also kein endliches Maß haben.)

Später betrachtete Kingman [131] auch den Fall mehrerer Loci (d Stück), d.h. Haplotypen statt einzelner Allele, auf dem Zustandsraum $S = \mathbb{R}^d$. Er bemerkte, dass die von ihm betrachteten Größen robust gegen diese Modellerweiterung sind und berechnete beispielhaft die effektive Zahl der Allele.

Auch für das SMM mit Längen-Bias gibt es Ansätze zur weiteren Verallgemeinerung, z.B. das *Snakes-and-Ladders-Modell* [32], ein Multistep-Modell, bei dem die Mutationsverteilung auch von Umwelteinflüssen abhängen kann.

1.5 Theorie der wandernden Verteilungen

Die Theorie der wandernden Verteilungen wurde von P.A.P. Moran [168, 169] initiiert und befasst sich mit der unter einem SMM erwarteten Verteilung der Allelfrequenzen in einer Population. In diesem Abschnitt werden einige klassische Arbeiten zusammengefasst, die ein wanderndes, aber kohärentes Verhalten dieser Verteilung herleiten und illustrieren. In Publikation (i) komplementieren wir diese klassischen Resultate, indem wir das Phänomen der wandernden Verteilungen mittels Markov-Ketten-Theorie charakterisieren.

1.5.1 Morans Arbeiten

Moran [168] ging vom einfachen SMM aus (siehe Abschnitt 1.4.2) und betrachtete die Anzahl $A_n(z)$ der Individuen mit Allel z in Generation n. Da wir ein Modell mit konstanter Populationsgröße N haben, gilt für alle $n \in \mathbb{N}$ stets $\sum_{z\in\mathbb{Z}} A_n(z) = N$. In jeder Generation n bildet $(A_n(z))_{z\in\mathbb{Z}}$ eine Folge von Zufallsgrößen, deren gemeinsame Verteilung von Interesse ist. Egal wie groß N ist, wenn die Genealogie der Population sich nach einem Wright-Fisher-Modell entwickelt (siehe Abschnitt 1.3.1), dann konvergiert diese gemeinsame Verteilung mit $n \to \infty$ nicht gegen eine

feste Verteilung, sondern wandert unbeschränkt auf $\mathbb{Z}$ umher, und zwar so, dass die Menge der $z \in \mathbb{Z}$ mit $A_n(z) > 0$ eine mehr oder weniger kompakte Gruppe bleibt. Dies nannte Moran [168] eine *wandernde Verteilung*. In Publikation (i) charakterisieren wir dieses Verhalten, indem wir es in einen globalen und einen lokalen Aspekt zerlegen. Lokal betrachtet nimmt die Allel-Verteilung eine feste Form an, wenngleich sie global betrachtet nicht konvergiert.

Moran [168] beschritt einen anderen Weg, um diese Eigenschaften der Allelverteilung herzuleiten. Er zeigte zunächst, dass $((A_n(z))_{z\in\mathbb{Z}})_{n\in\mathbb{N}_0}$ mit $n \to \infty$ nicht konvergiert, d.h. auch nach Ablauf vieler Generationen gibt es keinen Gleichgewichtszustand für die gemeinsame Verteilung der Allele. Deshalb betrachtete Moran dann stattdessen die folgenden Größen. Für alle $n \in \mathbb{N}$, $k \in \mathbb{Z}$ definiere

$$C_k(n) := \frac{1}{N^2} \sum_{z\in\mathbb{Z}} A_n(z) \cdot A_n(z+k). \tag{1.3}$$

Dann gilt für alle $k \in \mathbb{Z}$, dass der Limes $\lim_{n\to\infty} C_k(n) =: C_k$ existiert [168]. Wie Moran [168] bemerkte, sind die Größen $C_k(n)$ eine Verallgemeinerung dessen, was Ohta und Kimura [183] als n_e, die *effektive Anzahl neutraler Allele* (in der Population) bezeichnen, denn für die gegenwärtige Generation n gilt:

$$C_0(n) = \frac{1}{N^2} \sum_{z\in\mathbb{Z}} A_n^2(z) = \frac{1}{n_e}. \tag{1.4}$$

Insbesondere gilt also, dass die effektive Anzahl neutraler Allele mit $n \to \infty$ konvergiert. Dieses Resultat ist ein Verbindungspunkt zu unserer Publikation (i), denn dort ergibt es sich als Korollar aus Satz 11.

In der Diskussion seiner Ergebnisse wies Moran [168] darauf hin, dass seine Methode auch auf allgemeinere Mutationsmodelle (z.B. Multistep) anwendbar wäre. Er bemerkte ferner, dass auf einem zirkulären Zustandsraum dasselbe Phänomen der wandernden Verteilung auftreten kann, falls θ klein genug ist. Der Extremfall mit dem Zustandsraum $\{1, 2\}$ entspricht dann dem klassischen (symmetrischen) Zwei-Allele-Modell. Die Frage, was der Erwartungswert für die Anzahl verschiedener Allele in der Population ist, ließ Moran offen.

1.5.2 Kingmans Ansatz

Um sein verallgemeinertes SMM (siehe Abschnitt 1.4.2) zu untersuchen, bediente sich Kingman [134] des wahrscheinlichkeitstheoretischen Konzepts der charakteristischen Funktionen (siehe z.B. [12, § 22]). In Publikation (i) bezeichnen wir mit $X_n(i)$ das Allel des i-ten Individuums der n-ten Generation und nennen $X = (X_n)_{n\in\mathbb{N}_0}$ den Allelprozess. In dieser Notation lautet Kingmans Definition wie folgt. Für jede Generation $n \in \mathbb{N}_0$ ist die gemeinsame charakteristische

Funktion Ψ_n der Allele $X_n(j)$ (mit $j \in \{1, \dots N\}$) gegeben durch

$$\Psi_n(u) = \Psi_n(u_1, \dots, u_N) := E \exp\left(i \sum_{j=1}^{N} u_j \cdot X_n(j) \right) \tag{1.5}$$

für alle $u = (u_1, \dots, u_N) \in \mathbb{R}^N$. Damit ist Ψ_n für jedes $n \in \mathbb{N}_0$ wegen der Austauschbarkeit der $X_n(j)$ (siehe Publikation (i)) eine symmetrische Funktion.

Kingman [134] definierte außerdem für alle $n \in \mathbb{N}_0$, $j \in \{1, \dots, N-1\}$

$$V_n(j) := X_n(j) - X_n(N). \tag{1.6}$$

Der Prozess $V := (V_n)_{n \in \mathbb{N}_0}$ heißt *normierter Allelprozess*. In Publikation (i) analysieren wir diesen normierten Prozess für das einfache SMM. Auch für den normierten Allelprozess definierte Kingman [134] analog die charakteristischen Funktionen Ψ_n^0.

Auf diesem Wege konnte Kingman [134] die folgenden Konvergenzresultate für den normierten Allelprozess beweisen:

(i) Es ex. $\Psi^0 : \mathbb{R}^{N-1} \to \mathbb{C}$, stetig in 0_{N-1}, mit

$$\Psi_n^0 \overset{n \to \infty}{\longrightarrow} \Psi^0. \tag{1.7}$$

(ii) Für alle $n \in \mathbb{N}_0$ gilt:

$$|\Psi_n^0 - \Psi^0| \leq 6 \left(\frac{N-1}{N} \right)^n. \tag{1.8}$$

Qualitativ formuliert gilt also, dass V in Verteilung konvergiert, wobei die Konvergenzgeschwindigkeit von N abhängt. X hat somit auch in diesem Fall eine wandernde Verteilung, oder mit Kingmans Worten: X ist ein *coherent random walk*. Kingman gab dafür laut Ewens [66] auch noch die folgende intuitive Erklärung, in der bereits einige seiner Koaleszenzideen sichtbar werden (vgl. Abschnitt 1.3.2):

- Die Wahrscheinlichkeit, dass zwei Individuen aus Generation n einen gemeinsamen Vorfahren in Generation m haben, ist $1 - (\frac{N-1}{N})^{n-m}$.
- Diese Wahrscheinlichkeit ist nahe 1, wenn $n - m$ groß ist verglichen mit N.
- Also stammt die gesamte n-te Generation von einem Vorfahren aus Generation $n - \Delta$ ab, wobei $\Delta = \Delta(n)$ eine Zufallsgröße ist, die stochastisch beschränkt bleibt für $n \to \infty$.
- V ist also das Resultat von Mutationen in nur Δ Generationen.

1.5.3 Kestens Verallgemeinerung

Harry Kesten [121, 120] griff die von Moran aufgeworfene Frage nach der Anzahl verschiedener Allele auf, indem er einen verallgemeinerten Mutationsprozess betrachtete. In Publikation (i) bezeichnen wir mit $Z_n(i)$ die Änderung der Allel-Länge (durch Mutation) des i-ten Allels der n-ten Generation im Vergleich zu seinem Vorfahren (siehe dort Abschnitt 2). Kesten ging von Kingmans verallgemeinertem SMM aus (siehe Abschnitt 1.4.2), definierte den Mutationsprozess Z aber für beliebige Wahrscheinlichkeitsverteilungen $\varphi : \mathbb{Z} \to [0, 1]$ mit $\varphi(0) = 0$ — und für Mutationsraten, die von der Populationsgröße abhängen können ($\mu_N \in (0, 1]$ für alle $N \in \mathbb{N}$) — durch

$$P(Z_n(i) = k) = \begin{cases} \mu_N \cdot \varphi(k) & \text{falls } k \neq 0, \\ 1 - \mu_N & \text{falls } k = 0; \end{cases} \tag{1.9}$$

für alle $n \in \mathbb{N}$, $i \in \{1, \dots N\}$, $k \in \mathbb{Z}$. Das SMM ist als Spezialfall (mit $\varphi(1) = \varphi(-1) = \frac{1}{2}$) in diesem Modell enthalten.

Kesten definierte $\Lambda(s, N, n)$ (für alle $n \in \mathbb{N}_0$, $s \in \{1, \dots N\}$) als die Anzahl verschiedener Allele in einer Stichprobe der Größe s aus der Generation n. Insbesondere ist also $\Lambda(N, N, n)$ die Anzahl verschiedener Allele in Generation n. Λ kann direkt aus den Werten des Allelprozesses X, oder indirekt aus den Werten des normierten Allelprozesses V berechnet werden. Somit folgt aus der Konvergenz von V, dass $\Lambda(s, N, n)$ mit $n \to \infty$ in Verteilung konvergiert. $\Lambda(s, N)$ sei der Limes von $\Lambda(s, N, n)$ für $n \to \infty$. Man kann $\Lambda(s, N)$ als die Anzahl verschiedener Allele in einer Stichprobe der Größe s aus einer Population im Gleichgewicht interpretieren. Allerdings stellte Kesten [121, 120] fest, dass $\Lambda(s, N)$ mit $s \to \infty$ divergiert. Unter einem so allgemeinen Modell liegt also nicht notwendigerweise eine wandernde Verteilung vor. Deshalb konzentrieren wir uns in Publikation (i) auf das einfache SMM.

1.6 Forensische Anwendungen

Die Beziehung zwischen forensischer Genetik und populationsgenetischer Forschung kann in beiden Richtungen fruchtbar sein. Einerseits entwickelt und verbessert die Forensik ihre Methoden auf der Grundlage wissenschaftlicher Erkenntnisse, und andererseits kann die forensische Arbeit durch ihre Datensammlungen selbst wieder als Forschungsgrundlage dienen. Der letztgenannte Punkt wird durch unsere Publikation (ii) illustriert.

Für jene Arbeit haben wir Y-STR-Daten von Vater-Sohn-Paaren aus zahlreichen forensischen Publikationen aggregiert. So wurde es möglich, verschiedene STR-Mutationsmodelle für jeden Locus separat anzupassen und miteinander zu vergleichen. Die Locusabhängigkeit des Mutationsprozesses wird somit berücksichtigt, ohne in den einzelnen Modellen explizit formuliert zu sein.

In Publikation (iii) beschreiten wir den umgekehrten Weg, indem wir mittels populationsgenetischer Methoden einen Beitrag zur forensischen Praxis der Schätzung von Random-Match-Wahrscheinlichkeiten (siehe Abschnitt 1.6.2) leisten.

1.6.1 Der genetische Fingerabdruck

Als Alec Jeffreys und Kollegen 1985 [106, 107] mit Hilfe von Restriktionsenzymen, Elektrophorese und eigens entwickelten Hybridisierungs-Sonden (Southern Blotting) die Multi-Locus-Technik zur gleichzeitigen Darstellung mehrerer hochvariabler Minisatelliten im menschlichen Genom entwickelten, erkannten sie sofort das Potenzial dieser Technik zu Identifikationszwecken und anderen forensischen Anwendungen [77]. Die erste solche Anwendung in Form der Verwandtschaftsanalyse für einen Einwanderungsfall [104] folgte wenig später.

Jeffreys entschied sich nach eigenen Angaben spontan dafür, seine Methode den *genetischen Fingerabdruck* zu nennen, als sein Freund Nick Proudfoot ihn nach einem Seminar-Vortrag auf die Analogie aufmerksam machte [103]. Gelegentlich wird ihm unterstellt, dass er diese Metapher bewusst wählte, um damit die Präzision zu betonen und so die breite Akzeptanz der Technik zu beschleunigen (siehe z.B. [167, S. 40 und S. 57]). Sollte das tatsächlich damals schon die Absicht gewesen sein, so war diese Strategie durchaus erfolgreich, und später wurde der Begriff auch für andere forensische Genotypisierungstechniken (siehe unten) übernommen [140]. Das US-amerikanische National Institute of Standards and Technology (NIST) vermeidet allerdings die Metapher des genetischen Fingerabdrucks und verwendet stattdessen den neutraleren Begriff *DNA Typing* [25]. Oft wird auch von DNA-Profilen gesprochen, insbesondere in Bezug auf mehrere einzeln typisierte Loci.

Trotz aller offensichtlichen Parallelen sollte man nämlich nicht vergessen, dass es auch wesentliche Unterschiede zwischen genetischen und echten Fingerabdrücken gibt [214], [140, S. 95], [167, insbes. S. 40, Note 79], [142, Box 14.2]:

(i) Die Molekulargenetik ist ein weites und extrem aktives Forschungsfeld, während über Fingerabdrücke seit Francis Galton [72] vergleichsweise wenige neue Erkenntnisse gewonnen wurden.

(ii) Die Technik des genetischen Fingerabdrucks ist anspruchsvoll und fehleranfällig, die strikte Einhaltung von Protokollen und Qualitätsstandards ist unabdingbar [25, Chapter 13].

(iii) Aus klassischen Fingerabdrücken lassen sich in der Regel keine Rückschlüsse über Geschlecht, Hautfarbe oder Erkrankungen einer Person ziehen.

(iv) Klassische Fingerabdrücke zeigen zwar eine gewisse Heritabilität, diese ist aber zu gering, um praktischen Nutzen für forensische Verwandschaftsanaly-

sen zu haben. Sogar monozygote Zwillinge haben voneinander verschiedene Fingerabdrücke.

Die Einführung der neuen Technik als Werkzeug zur Verbrechensaufklärung war besonders in den USA von zum Teil vehement geführten Debatten zwischen Staatsanwälten, Strafverteidigern, Bürgerrechtlern, FBI-Mitarbeitern, Humangenetikern, Populationsgenetikern und Mathematikern begleitet, die sowohl in Gerichtssälen (bzw. in den einigen Stafverfahren vorgeschalteten Anhörungen zur Zulässigkeit von Beweismitteln) als auch in der wissenschaftlichen Fachliteratur ausgetragen wurden (siehe z.B. [147], [148] und die darauf folgenden Leserbriefe, [156], [31] und [14]). Im Zuge dieser Kontroversen wurde durch den *National Research Council* (NRC), also innerhalb der *National Academy of Sciences*, zwecks Konsensfindung eine Arbeitsgruppe unter Leitung von Victor McKusick gegründet, die 1992 ihren Bericht mit Empfehlungen zum genetischen Fingerabdruck vorlegte [38]. Danach flammte die Debatte zwar noch einmal auf [231, 45], gerade in Bezug auf die Quantifizierung der Evidenz einer Profilübereinstimmung (siehe Abschnitt 1.6.2), aber allmählich glätteten sich die Wogen [149]. Dazu trug auch eine weitere Arbeitsgruppe des NRC bei, diesmal unter Leitung von James Crow, die 1996 ihren Abschlussbericht veröffentlichte [37].

Auch dank dieser Auseinandersetzungen ist die DNA-Typisierung heute eine ausgereifte und vermutlich die wissenschaftlich fundierteste und am besten regulierte Technik der gesamten Forensik. Es ist eine Ironie der Geschichte, dass dadurch inzwischen andere forensische Techniken, nicht zuletzt der klassische Fingerabdruck [177], verstärkter Kritik ausgesetzt sind, da sie keinen vergleichbaren Standards genügen [39]. Dies wird besonders dramatisch durch die Fälle illustriert, in denen zu Unrecht Verurteilte dank nachträglicher DNA-Analysen rehabilitiert wurden [91]. Sogar Richard Lewontin, der in den 1990-er Jahren zu den schärfsten Kritikern der Praxis forensischer DNA-Typisierung zählte, ist dank zahlreicher Verbesserungen inzwischen von der Technik überzeugt [155].

Neben den theoretischen Entwicklungen gab es auch auf der technischen Seite wesentliche Fortschritte. Die ersten für den genetischen Fingerabdruck genutzten Marker waren wie bereits dargestellt Minisatelliten. Wegen der erwähnten Verwendung von Restriktionsenzymen wurden diese (und andere mit dieser Technik typisierte) Loci oft auch als Restriktionsfragment-Längen-Polymorphismen (RFLPs) bezeichnet. Die Erfindung der Polymerase-Kettenreaktion (PCR), ebenfalls 1985 publiziert [204, 173], in Verbindung mit der aus *Thermus aquaticus* isolierten hitzebeständigen DNA-Polymerase [203] sollte dann auch die forensische DNA-Typisierung revolutionieren. Die Vorteile der PCR-Technik gerade für forensische Anwendungen sind unverkennbar, denn Tatortspuren enthalten oft nur geringe Mengen verwertbarer DNA, d.h. PCR-Amplifikation ist die einzige Möglichkeit, diese Spuren überhaupt auszuwerten. Auch die für die ersten genetischen Fingerabdrücke verwendeten Minisatelliten lassen sich mittels PCR amplifizieren [105]. Aber es stellte sich heraus, dass Mikrosatelliten (STRs) noch besser

für forensische Anwendungen, gerade in Verbindung mit der PCR, geeignet sind [58, 59]. Für die Verwendung von STRs als forensische Standard-Marker sprechen insgesamt die folgenden Gründe [109], [24], [25, Table 3.6], [26, S. 99–103].

(i) Es existiert, über das gesamte Genom verteilt, eine reichhaltige Auswahl an STRs mit hoher interindividueller Variabilität (siehe Abschnitt 1.2).

(ii) Ein historischer Grund für die Verwendung von STRs in der Forensik war die Tatsache, dass viele solche Loci im Zuge humangenetischer Forschung, nämlich zur Kartierung von genetischen Erkrankungen, charakterisiert worden waren.

(iii) Die STRs selbst sind aber meist evolutionär neutral (siehe Abschnitt 4.3). Die in Abschnitt 1.2.2 behandelten krankheitsassoziierten STRs machen nur einen kleinen Teil aller STR-Loci aus und werden in der Forensik nicht verwendet.

(iv) STRs lassen sich auch bei heterozygoten Personen gut mittels PCR amplifizieren, da der Längenunterschied zwischen den beiden Allelen in der Regel nicht sehr groß ist. Bei Minisatelliten besteht dagegen oft das Problem, dass die PCR überwiegend das kürzere Allel repliziert und das längere Allel dann nicht mehr sichtbar ist (*allelic dropout*).

(v) STRs haben relativ kurze Allele, und auch die entstehenden PCR-Fragmente lassen sich durch entsprechendes Primer-Design kurz halten (man spricht dann von *miniSTRs*). Das ist wichtig, weil forensisch relevante DNA-Proben oft stark fragmentiert sind.

(vi) Die geringe Länge der PCR-Fragmente ermöglicht sogar die nukleotidgenaue Bestimmung der Allel-Länge mittels Kapillar-Elektrophorese.

(vii) Slipped-Strand-Mispairings (siehe Abschnitt 1.4.1) sind bei Tetranukleotid-Repeats während der PCR relativ selten, d.h. der Anteil der sogenannten *Stotter-Fragmente* (PCR-Fragmente, die meist um eine Repeat-Länge von der Mehrheit abweichen) am Gesamtvolumen ist gering, in der Regel unter 15%.

(viii) Die geringe Spannweite der Allel-Längen eines STRs ist vorteilhaft für die gleichzeitige Amplifizierung mehrerer Loci (siehe unten).

(ix) Das Typisierungsverfahren ist heute weitgehend automatisiert und dauert nur wenige Stunden.

(x) Die breite Verwendung kommerzieller Kits für einige Standard-Loci (siehe unten) schafft Vergleichbarkeit.

Diese Vorteile der Verwendung von STRs in Verbindung mit der PCR für forensische Zwecke sind so überzeugend, dass die folgenden Nachteile [25, Table 3.6] in Kauf genommen werden.

(i) Die Variabilität pro Locus ist nicht so groß wie bei Minisatelliten.

(ii) Die PCR-Technik ist besonders anfällig für Kontamination.

(iii) Die benötigte Ausrüstung ist teuer.

(iv) Stotter-Fragmente, unterschiedliche Mengen der verschiedenen Allelen entsprechenden PCR-Fragmente und andere Artefakte können die Interpretation erschweren.

Heute sind STRs die in der Forensik am weitaus häufigsten verwendeten Marker. Dieser Status wurde insbesondere durch die Einrichtung von entsprechenden DNA-Datenbanken (siehe Abschnitt 1.6.2) zementiert. Für die Standardisierung spielt dabei neben den schon in Abschnitt 1.1 erwähnten ISFG-Richtlinien [88] die Existenz kommerzieller Typisierungs-Kits [163] eine wesentliche Rolle. Die Auswahl der Loci für Kits und Datenbanken sind eng miteinander verflochten, so dass heute alle Standard-Loci einer Datenbank mittels eines kommerziellen Kits in einer einzigen Reaktion typisiert werden können (*Multiplexing*).

1.6.2 Datenbanken und Random Match Probability

Bei den meisten forensischen Anwendungen von DNA-Profilen zur Verbrechensaufklärung in den 1980-er Jahren wurde jeweils ein Tatortprofil mit dem Profil eines Tatverdächtigen verglichen. Die DNA-Probe des Verdächtigen wurde also genommen, nachdem dieser aus von seinem DNA-Profil unabhängigen Gründen in den Ermittlungsfokus geraten war. Die Ausnahme waren DNA-Reihenuntersuchungen (*Dragnets*), bei denen alle potentiellen Täter aus einem bestimmten Gebiet von der Polizei zur Abgabe einer Probe aufgefordert wurden, wie z.B. im Fall des Doppelmordes in zwei englischen Dörfern [228], dem ersten Kriminalfall, bei dem der genetische Fingerabdruck eine wichtige Rolle spielte. In den 1990-er Jahren begannen die Strafverfolgungsbehörden im UK und in den USA damit, die DNA-Profile von verurteilten Straftätern in Datenbanken zu speichern [142, S. xvi]. Somit kann ein Tatortprofil routinemäßig mit allen Profilen in der Datenbank verglichen werden, was im Falle einer Übereinstimmung dann erst den Verdacht generiert. In den USA hat das FBI dafür ein System (CODIS) entwickelt, in dem 13 autosomale STRs sowie der Amelogenin-Locus (zur Geschlechtsbestimmung) enthalten sind. Auch viele andere Staaten nutzen inzwischen dieses System, aber im UK und in Deutschland wurden eigene Systeme entwickelt.

In einigen Bundesstaaten der USA wird heute sogar von jedem Verhafteten ein DNA-Profil gespeichert, unabhängig von einer Verurteilung [142, S. xvi]. Einige Richter und Politiker in den USA und im UK, so z.B. Rudolph Giuliani und Tony

Blair, haben bereits die Erfassung der DNA-Profile sämtlicher Bürger gefordert, konnten sich damit aber bisher nicht durchsetzen [142, S. 144], [170].

In Deutschland wurde im Vergleich zu den USA und dem UK erst spät, und zwar 1997, mit dem DNA-Identitätsfeststellungsgesetz die rechtliche Grundlage für die Einrichtung einer forensischen DNA-Datenbank geschaffen [142, S. 207]. Ein weiterer Unterschied besteht darin, dass in Deutschland die biologische Probe unmittelbar nach der Bestimmung des DNA-Profils vernichtet werden muss, d.h. es wird nur das Profil gespeichert (Strafprozessordnung, § 81g, Absatz 2). Im Dezember 2008 urteilte der europäische Gerichtshof für Menschenrechte (ECHR) in einem englischen Fall, dass die jahrelange Speicherung von DNA-Profilen ehemaliger Verdächtiger, gegen die nicht mehr ermittelt werde, deren Recht auf Privatsphäre verletze.

Grundsätzlich kann man feststellen, dass im Falle eines Unterschiedes zwischen zwei DNA-Profilen die beiden Proben höchstwahrscheinlich von verschiedenen Personen stammen (*Exclusion*). Wenn aber zwei Profile gleich sind (*Inclusion*), so könnten sie möglicherweise dennoch von verschiedenen Personen stammen. Das ist natürlich umso unwahrscheinlicher, je seltener die beobachteten Allele sind. Man muss also die Allelfrequenzen in einer Population kennen, um die Aussagekraft einer Profilübereinstimmung beurteilen zu können. Das ist auch ein Grund für die Einrichtung von DNA-Datenbanken.

Eine solche Datenbank kann folglich genutzt werden, um die Aussagekraft einer Profilübereinstimmung zu quantifizieren. Hierbei werden die Allelfrequenzen in verschiedenen Populationen und Subpopulationen geschätzt. Oft werden zu diesem Zweck sogar eigene Datenbanken angelegt, die sich aus entsprechenden Studien speisen. Auch genealogisch orientierte Datenbanken kommerzieller Anbieter können in diesem Zusammenhang nützlich sein.

Die Quantifizierung der Aussagekraft einer Profilübereinstimmung erfolgt dann in der Regel über die sogenannte *Random Match Probability* (RMP), das ist die Wahrscheinlichkeit, mit der eine zufällig aus der Referenzpopulation gezogene Person das beobachtete Profil aufweist. Wenn dabei jede Person mit gleicher Wahrscheinlichkeit gezogen wird, dann ist die RMP also die relative Häufigkeit des beobachteten Profils, aber für die Berechnung bzw. Schätzung der RMP werden oft nahe Verwandte ausgeschlossen. Eine Alternative zur RMP, auf die hier nicht weiter eingegangen wird, ist die Verwendung von Likelihood-Ratios, siehe dazu z.B. Collins und Morton [35] oder Weir [232, Chapter 6]. Jedoch wird unter dem Likelihood-Ansatz die RMP oft etwas anders definiert, nämlich als die *bedingte* Wahrscheinlichkeit für das Profil, gegeben dass dasselbe Profil bereits einmal beobachtet wurde [233, S. 723].

1.6.3 Das Y-Chromosom

In den Publikationen (ii) und (iii) betrachten wir STR-Loci auf dem Y-Chromosom, kurz *Y-STRs*. Hier soll ein kurzer Überblick über die Evolution des Y-Chromosoms

und seine daraus resultierenden speziellen Eigenschaften gegeben werden. Wegen der folgenden Anwendungen liegt der Focus auf dem Y-Chromosom des Menschen, aber es werden gelegentlich andere Arten erwähnt, um die evolutionären Mechanismen besser beurteilen zu können. In Abschnitt 1.6.4 werden dann die wichtigsten Y-STRs vorgestellt.

Wie bei anderen Säugetieren werden beim Menschen die Geschlechtschromosomen (Gonosomen) mit X und Y bezeichnet, und das homogametische Geschlecht (XX) ist das weibliche. Gonosomen-Paare sind in der Evolution mehrmals entstanden, und man geht davon aus, dass sie sich in der Regel aus Autosomen entwickelt haben. Es ist aber noch nicht geklärt, ob dabei die Entstehung geschlechtsspezifischer Loci (z.B. solche Gene, die nur in den Testikeln exprimiert werden) der Unterdrückung der Rekombination vorausgeht, oder umgekehrt [158, S. 349-350].

Vieles spricht dafür, dass das menschliche Y-Chromosom einen weit verbreiteten Evolutionsmechanismus durchlaufen hat (für aktuelle Reviews siehe [62], [98] und [6]). Im Rahmen des Humangenomprojekts war ein erster Entwurf der Y-Chromosom-Sequenz entstanden [150, Figure 9], siehe auch [225]. Dieser war jedoch sehr ungenau, weil das Y-Chromosom wegen der vielen repetitiven Sequenzen besonders schwierig zu sequenzieren ist (siehe z.B. [225, Table 10]). 2003 wurde dann von Helen Skaletsky und Kollegen unter Leitung von David Page ein weiterer Sequenzentwurf des euchromatischen Teils veröffentlicht [208]. Genauer gesagt wurde in dieser Arbeit erstmals die *male-specific region of the Y chromosome* (MSY) einigermaßen vollständig und zuverlässig sequenziert. Die MSY macht mit 22,5 Mb etwa 90% des euchromatischen Teils aus und komplementiert die pseudoautosomalen Regionen an den beiden Enden des Y-Chromosoms, welche nach wie vor während der männlichen Meiose mit den entsprechenden Abschnitten auf dem X-Chromosom rekombinieren. Es gibt in der MSY nur noch 78 proteincodierende Gene, die zusammen aber aufgrund von Gen-Duplikationen sogar nur 27 funktionell verschiedene Proteine codieren.

Skaletsky et al. [208] unterscheiden drei Klassen von Sequenzen innerhalb des euchromatischen Teils der MSY anhand ihrer Zusammensetzung und damit der evolutionären Geschichte: X-transponiert, X-degeneriert und amplikonisch (*ampliconic*). Die X-transponierten Sequenzen haben zusammen eine Länge von 3,4 Mb und stimmen zu 99% mit Sequenzen in Xq21 überein, d.h. sie lassen sich auf eine rezente (3-4 mya) Transposition vom X-Chromosom zurückführen. Auch zu den X-degenerierten Sequenzen (insgesamt 8,6 Mb) lassen sich noch Homologe auf dem X-Chromosom finden, aber mit deutlich weniger Übereinstimmung. Die amplikonischen Abschnitte (insgesamt 10,2 Mb) zeigen dagegen eine starke intrachromosomale Sequenzübereinstimmung. Insbesondere gehören zu dieser Klasse acht große Palindrome.

Während die X-transponierten Sequenzen nur zwei Gene enthalten, liegen im X-degenerierten Teil 16 verschiedene proteincodierende Gene, die oft im ganzen Körper exprimiert werden. Insbesondere ist hier die *sex-determining region Y*

(SRY) zu nennen, ein Gen, das einen Transkriptionsfaktor codiert, der die Ausbildung der männlichen Geschlechtsmerkmale auslöst. Auch *Amelogenin Y*, ein Gen das mit dem forensischen Standard-Marker zur Geschlechtsbestimmung assoziiert ist, gehört zu dieser Gruppe. Alle bislang genannten Gene kommen bei den meisten Menschen nur einfach vor. Die Gene in den amplikonischen Abschnitten liegen dagegen in der Regel in Form mehrerer Kopien vor und codieren neun Proteine, die geschlechtsspezifische Funktionen erfüllen.

Vor kurzem sind auch Y-Chromosom-Sequenzen eines Schimpansen (*Pan troglodytes*, [100]) und eines Rhesus-Makaken (*Macaca mulatta*, [99]) publiziert worden. Auch in diesen beiden Sequenzen wurden X-degenerierte und amplikonische (in *M. mulatta* aber nur 0,5 Mb) Bereiche gefunden. Die X-transponierten Abschnitte finden sich nur beim Menschen, was zu erwarten war, denn das erwähnte Transpositionsereignis wurde auf einen Zeitpunkt nach unserem letzten gemeinsamen Vorfahren mit den Schimpansen datiert. Bei *P. troglodytes* und *M. mulatta* sind die X-degenerierten Abschnitte wie beim Menschen tendenziell stark konserviert, während die amplikonischen Sequenzen relativ vielen und großen Veränderungen (Deletionen, Duplikationen und Inversionen) unterliegen. Der Mechanismus, der diesen Veränderungen zugrundeliegt, wird als Genkonversion bezeichnet. Dabei lagern sich während der Meiose verschiedene Abschnitte der MSY aneinander an, und es kommt zu einer nichtreziproken Rekombination, d.h. der eine beteiligte Abschnitt verschwindet und wird durch eine Kopie des anderen ersetzt. Analog kann auch reziproke Rekombination innerhalb der MSY vorkommen. Diese Rekombinationsereignisse sind alle intrachromosomal, d.h. das X-Chromosom ist daran nicht beteiligt.

Zur Populationsgenetik des Y-Chromosoms [193] ist anzumerken, dass bei Säugetieren die effektive Populationsgröße für Y-chromosomale Loci, $N_{g,Y}$, in der Regel deutlich geringer ist als die effektive Populationsgröße für autosomale Loci, $N_{g,A}$. Dafür gibt es mehrere Gründe. So kommt z.B. in Populationen mit einem Geschlechterverhältnis von 1:1 jedes Autosom in der Population viermal so oft vor wie das Y-Chromosom, da letzteres in Männern nur haploid vorliegt und in Frauen gar nicht. Michael Lynch [158, Table 12.3] schätzt das Verhältnis $N_{g,Y}/N_{g,A}$ beim Menschen auf 8,0 %.

Aufgrund ihrer haploiden Natur ist die MSY als Marker für populationshistorische Studien (siehe Abschnitt 1.2.1) besonders beliebt [111], [110, Section 8.6.2], [223]. Der mittlerweile etablierte Weg, eine Phylogenie aller extanten menschlichen Y-Chromosomen zu rekonstruieren, führt über Einzelnukleotid-Polymorphismen (*Single Nucleotide Polymorphisms*, SNPs) [115]. Inzwischen sind die nichtrepetitiven, euchromatischen Abschnitte des Y-Chromosoms in Dutzenden Männern sequenziert worden. Anhand der dabei gefundenen Einzelnukleotid-Varianten (SNVs) konnte eine detaillierte Phylogenie rekonstruiert werden, deren Zeit bis zum MRCA auf etwa 120.000 bis 200.000 Jahre geschätzt wird [191, 70].

1.6.4 Y-STR-Loci

Für forensische Anwendungen hat die MSY den großen Vorteil, dass sich darüber leicht der männliche Beitrag zu Mischspuren, z.B. in Vergewaltigungsfällen, herausfiltern lässt. Reviews zu solchen forensischen Anwendungen finden sich in [194], [25, Chapter 16] und [26, Chapter 13]. Wie allgemein in der forensischen DNA-Typisierung (siehe Abschnitt 1.6.1) hat sich auch für das Y-Chromosom die Nutzung von STR-Markern durchgesetzt. Auch hier gelten die ISFG-Richtlinien [88] zur Nomenklatur von Loci und Allelen.

Das humane Y-Chromosom enthält über 400 bekannte STRs, von denen aber nur ein kleiner Teil routinemäßig in der Forensik eingesetzt wird [25, S. 369]. 1997 wurde im Rahmen einer großen europaweiten Studie die Empfehlung ausgesprochen, bei der forensischen Y-Typisierung stets sieben bestimmte Kern-Loci zu erfassen [118]. Dieser sogenannte *minimale Haplotyp* umfasst die folgenden Loci: DYS19, DYS389 I, DYS389 II, DYS390, DYS391, DYS392 und DYS393. Dies sind infolgedessen auch die Loci, zu denen mit großem Abstand die meisten Daten vorliegen, und auf die wir uns in Publikation (ii) konzentrieren (mit Ausnahme von DYS389 II). Es existieren verschiedene Y-STR-Kits kommerzieller Anbieter, die neben dem minimalen Haplotyp noch spezifische weitere Loci abdecken. Das bisher am häufigsten verwendete Kit ist der sogenannte *Yfiler* [171] von *Applied Biosystems* mit insgesamt 17 Loci. Diese Loci und einige ihrer Eigenschaften sind in Tabelle 1.3 zusammengefasst. Die entsprechenden Daten stammen aus der *Y Haplotype Reference Database* (YHRD) von Lutz Roewer und Sascha Willuweit [243], der weltweit größten Sammlung von Y-STR-Daten. Zur Zeit plant Lutz Roewer ein internationales Projekt, in dem das *Promega*-Kit *PowerPlex Y23* mit 23 Loci breite Anwendung finden soll.

Kaye Ballantyne und Kollegen haben 13 Y-STRs mit geschätzten Mutationsraten über 1% entdeckt [8], und sie empfehlen, diese RM- (*rapidly mutating*) Y-STRs zur Differenzierung paternaler Linien in Populationen mit geringer Y-Diversität einzusetzen [9]. Anhand von Stichproben aus aller Welt wurde zunächst geschätzt, dass sogar die Unterscheidung zwischen Vater und Sohn mit diesen 13 RM-Y-STRs für fast die Hälfte aller Vater-Sohn-Paare möglich ist [9], aber eine größere Studie ergab, dass lediglich 26,9% aller Vater-Sohn-Paare mit diesem Marker-Satz differenziert werden konnten [7].

1.6.5 Vater-Sohn-Daten

Die Standardmethode bei Vaterschaftstests ist es, autosomale STRs in Mutter, Kind und einem oder mehreren hypothetischen Vätern zu genotypisieren [25, S. 397–404]. Aber auch Y-STRs werden oft für Vaterschaftstests hinzugezogen, wenn nur ein Mann aus einer männlichen Linie als Vater in Betracht kommt. Siehe dazu die entsprechenden Richtlinien der ISFG [78, insbes. R2.2].

Da das Aufkommen an Vaterschaftstests in forensischen und anderen La-

Nr.	Name	$\mu\,[10^{-3}]$	Allele	Motiv
1.	**DYS19**	2,3	14 (9–19)	TAGA
2.	**DYS389 I**	2,5	13 (9–17)	TCTG / TCTA
3.	DYS389 II	3,6	30 (24–35)	TCTG / TCTA
4.	**DYS390**	2,1	24 (17–29)	TCTG / TCTA
5.	**DYS391**	2,6	10 (5–16)	TCTA
6.	**DYS392**	0,4	11 (4–20)	TAT
7.	**DYS393**	1,0	13 (7–18)	AGAT
8.	DYS385 a/b	2,1	14,15 (6,16–24,24)	GAAA
9.	DYS438	0,3	11 (7–18)	TTTTC
10.	DYS439	5,2	12 (5–19)	GATA
11.	DYS437	1,2	14 (10–18)	TCTA / TCTG
12.	DYS448	1,6	20 (14–24)	AGAGAT
13.	DYS456	4,2	15 (10–23)	AGAT
14.	DYS458	6,4	17 (11–24)	GAAA
15.	DYS635	3,5	21 (16–29)	TCTA / TGTA
16.	Y-GATA-H4	2,4	12 (8–15)	TAGA

Tabelle 1.3: Charakteristika der am häufigsten verwendeten YSTR-Loci (*Yfiler*-Kit). Für jeden Locus ist μ die relative Häufigkeit beobachteter Mutationen pro Generation, ohne Berücksichtigung der Allel-Länge des Vaters. Unter „Allele“ ist die am häufigsten beobachtete Repeat-Anzahl angegeben, mit dem Wertebereich (Minimum–Maximum) in Klammern. Die sechs Kern-Loci, auf die wir uns in Publikation (ii) konzentrieren, sind durch Fettdruck ihrer Namen hervorgehoben. DYS385 a/b besteht aus zwei Kopien eines STRs, die auf den beiden Armen eines Palindroms liegen. Die beiden Allele an diesem Locus werden laut Konvention durch ein Komma getrennt geschrieben, wobei stets das kürzere Allel an erster Stelle steht. Daten nach YHRD [243], Release 33.

boren weltweit sehr groß ist (allein in den USA mehr als 300.000 Fälle pro Jahr [25, S. 397]), liegt hier ein enormes Potential für die wissenschaftliche Nutzung der dabei anfallenden Daten. Einige Labore speisen die von ihnen erhobenen Y-STR-Haplotypen unverwandter Männer (die oft die potentiellen Väter aus Vaterschaftstest-Fällen sind) routinemäßig in anonyme Datenbanken wie die YHRD ein, oft begleitet von wissenschaftlichen Publikationen. Der praktische Hauptzweck dieser Vorgehensweise liegt darin, eine breite Datenbasis für die Schätzung von Haplotypfrequenzen, in der Regel beschränkt auf eine bestimmte Subpopulation, zu gewinnen. Gleichzeitig werden mitunter die Y-STR-Haplotypen kompletter Vater-Sohn-Paare veröffentlicht. In Publikation (ii) listen wir alle Publikationen mit solchen Daten auf, die wir bis dato finden konnten (Table 2) und verwenden diese Daten zur Schätzung der Parameter für verschiedene Mutationsmodelle.

1.6.6 Evidenz-Quantifizierung für Y-STR-Haplotypen

Wie in Abschnitt 1.6.3 erläutert, unterscheidet sich das Y-Chromosom molekular- und populationsgenetisch wesentlich von den Autosomen. Diese Unterschiede müssen in der forensischen Anwendung berücksichtigt werden, insbesondere für die Quantifizierung der Evidenz, die bei der Übereinstimmung von zwei Y-STR-Profilen vorliegt [22]. So wird allgemein akzeptiert, dass es aufgrund der Abwesenheit von Rekombination unzulässig wäre, die Allel-Häufigkeiten mehrerer Y-STRs zu multiplizieren, um die Haplotyp-Häufigkeit zu schätzen. Dies steht im Kontrast zum Verfahren bei autosomalen Loci. Darüber hinaus liegt die Vermutung nahe, dass bei der forensischen Verwendung von Y-STRs deren Mutationseigenschaften, sowie die konkreten Allel-Längen, nicht mehr ignoriert werden können. Wir argumentieren in Publikation (iii), dass diese Vermutung zutrifft und zeigen einen Weg, dem Rechnung zu tragen. Hier geben wir zunächst einen kurzen Überblick der wichtigsten Verfahren.

Wie bei autosomalen Loci führt der gebräuchlichste Weg zur Quantifizierung der Evidenz einer Übereinstimmung von zwei Y-STR-Profilen über die RMP (vgl. Abschnitt 1.6.2). Aber da die MSY, in der alle forensisch genutzten Y-STRs liegen [194], keiner Rekombination unterliegt, sind die verschiedenen Y-STRs natürlich nicht statistisch unabhängig voneinander. Die einfachste Methode dies zu berücksichtigen besteht darin, die Häufigkeit des gesamten beobachteten Haplotypen (in der relevanten Population) durch seinen Anteil in einer Referenzdatenbank wie der YHRD zu schätzen. Eines der beiden beobachteten Profile wird dabei in der Regel der Referenzdatenbank zugerechnet. Dies ist die traditionelle *Counting*-Methode [76].

Ein Problem dabei, das sich durch die neueren Kits mit mehr Loci verschärft hat, ist die Tatsache, dass oft die beiden beobachteten Haplotypen in der Referenzdatenbank noch gar nicht vorkommen. Ein solcher Haplotyp wird als *Singleton* bezeichnet, weil er nach der Augmentierung der Referenzdatenbank um den

einen beobachteten Haplotypen eben genau einmal darin vorkommt. In diesem Fall ist der *Counting*-Schätzer $1/(n+1)$ (für eine Subpopulation der Größe n in der Referenzdatenbank) besonders unbefriedigend, weil er offensichtlich durch n beschränkt ist. Charles Brenner schlug für Singletons eine Modifikation des *Counting*-Schätzers, die sogenannte κ-Korrektur, vor [18].

Eine weitere Alternative ist die *Haplotype Frequency Surveying Method* [195, 139, 242], die auf einem Bayes-Ansatz basiert. Und schließlich gibt es eine weniger bekannte, aber vielversprechende Methode, die von Ian Wilson und David Balding entwickelt wurde [244, 245]. Sie bedient sich der Koaleszenztheorie (siehe Abschnitt 1.3.2) und bildet die Grundlage für unsere Publikation (iii). Für eine Zusammenfassung dieser Methode siehe dort Section 2.

Kapitel 2

Veröffentlichte Resultate

2.1 A Markov Chain Description of the Stepwise Mutation Model: Local and Global Behaviour of the Allele Process

Caliebe[1], Jochens[1], Krawczak und Rösler (2010)
Journal of Theoretical Biology 266:2, 336–342.

Zusammenfassung

STRs weisen unter bestimmten Bedingungen ein Verhalten auf, das als *wandernde Verteilung* [168, 169] bekannt ist. In dieser Publikation benutzen wir die mathematische Theorie der Markov-Ketten, um dieses Phänomen zu charakterisieren.

Wir gehen vom Wright-Fisher-Modell in Verbindung mit dem schrittweisen Mutationsmodell (SMM) aus. Insbesondere betrachten wir somit eine Population konstanter Größe N mit diskreten Generationen, und die verschiedenen Allele werden durch ganze Zahlen repräsentiert. Wir bezeichnen mit $X_n(i)$ das Allel des i-ten Individuums der n-ten Generation und nennen $X = (X_n)_{n\in\mathbb{N}_0}$ den Allelprozess. Außerdem definieren wir den normalisierten Allelprozess V, der sich von X dadurch unterscheidet, dass von jedem Allel das Allel eines bestimmten Individuums der jeweiligen Generation subtrahiert wird.

Sowohl X als auch V lassen sich als Markov-Ketten auffassen, d.h. die Verteilung der Allele (bzw. der Alleldifferenzen im Fall von V) in einer Generation hängt nur von den Allelen (bzw. Alleldifferenzen) der vorherigen Generation ab. Wir zeigen, dass der Allelprozess X nullrekurrent ist, d.h. für jeden Zustand $x \in \mathbb{Z}^N$ ($x = (x(1), \ldots x(N))$ steht für die Allele der N Individuen) gilt: Bei Start in x kehrt der Prozess mit Wahrscheinlichkeit 1 irgendwann zu x zurück, aber die erwartete Rückkehrzeit ist unendlich. Im Gegensatz dazu ist der normalisierte Allelprozess V positiv rekurrent, d.h. die erwarteten Rückkehrzeiten sind endlich. Daraus folgt, dass V (aber nicht X) gegen eine eindeutige stationäre Verteilung η konvergiert. Wir zeigen, dass diese Konvergenz exponentiell schnell erfolgt und dass η unimodal ist.

Wir leiten sowohl eine Rekursionsformel als auch einen geschlossenen Ausdruck für die Berechnung der Verteilung η_n von V_n her. Wegen der exponentiell schnellen Konvergenz gegen die stationäre Verteilung η kann dieses Resultat praktisch genutzt werden, um η zu approximieren. Dafür zeigen wir einige Beispiele.

Den Allelprozess X interpretieren wir als das globale Verhalten einer Population unter dem SMM. Die Menge der Allele wandert auf ganz $\mathbb{Z}^N$ umher. Dagegen erfasst der normalisierte Prozess V das lokale Verhalten. Die Allele bleiben auf ihrer Wanderung dicht beieinander, wobei die Form ihrer Gruppierung durch die Verteilungen η_n charakterisiert wird.

[1] geteilte Erstautorenschaft

Journal of Theoretical Biology 266 (2010) 336–342

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

A Markov chain description of the stepwise mutation model: Local and global behaviour of the allele process

Amke Caliebe [a,*,1], Arne Jochens [a,b,1], Michael Krawczak [a], Uwe Rösler [b]

[a] Institut für Medizinische Informatik und Statistik, Haus 31, Christian-Albrechts-Universität Kiel, Arnold-Heller-Str. 3, 24105 Kiel, Germany
[b] Mathematisches Seminar, Ludewig-Meyn-Str. 4, Christian-Albrechts-Universität Kiel, 24098 Kiel, Germany

ARTICLE INFO

Article history:
Received 2 February 2010
Received in revised form
24 June 2010
Accepted 24 June 2010
Available online 30 June 2010

Keywords:
Microsatellite
Evolution
Population genetics
Asymptotic behaviour

ABSTRACT

The stepwise mutation model (SMM) is a simple, widely used model to describe the evolutionary behaviour of microsatellites. We apply a Markov chain description of the SMM and derive the marginal and joint properties of this process. In addition to the standard SMM, we also consider the normalised allele process. In contrast to the standard process, the normalised process converges to a stationary distribution. We show that the marginal stationary distribution is unimodal. The standard and normalised processes capture the global and the local behaviour of the SMM, respectively.

1. Introduction

Microsatellites are successive iterations of a given short DNA sequence motif (usually 2–6 nucleotides long) that is repeated 5–100 times (Tautz, 1993; Chambers and MacAvoy, 2000). The number of iterations (the "repeat number") serves to identify a given microsatellite allele. Microsatellites are abundant in many species and have very high mutation rates (up to 10^{-2} per generation, Li et al., 2002). Owing to their high degree of variability, microsatellites are frequently used as markers in population genetics (Goldstein et al., 1999; Kashi and King, 2006), DNA fingerprinting (Cassidy and Gonzales, 2005; Bindu et al., 2007), whole genome mapping (Weissenbach et al., 1992) and genetic epidemiology (Thibodeau et al., 1993; Ashley and Warren, 1995).

The stepwise mutation model (SMM) was first introduced by Ohta and Kimura (1973) to describe the behaviour of electrophoretically detectable alleles in a population. Since then, the SMM has been widely used for modelling microsatellite mutation and evolution (Tishkoff et al., 1996; Zhivotovsky et al., 2003; De Iorio et al., 2005; Vardo and Schall, 2007). The SMM assumes that, in one generation, the repeat number can only increase or decrease by at most one, usually with equal probability. More refined models have been proposed that include mutations of greater length, mutation rates that depend upon repeat number, or the additional introduction of point mutations (Di Rienzo et al., 1994; Garza et al., 1995; Feldman et al., 1997; Zhivotovsky et al., 1997; Kruglyak et al., 1998; Durrett and Kruglyak, 1999; Falush and Iwasa, 1999; Calabrese et al., 2001); for an overview, see Watkins (2007) or Calabrese and Sainudiin (2005). As yet, however, it has remained controversial to what extent these models approximate the reality (Chambers and MacAvoy, 2000; Whittaker et al., 2003; Sainudiin et al., 2004; Cornuet et al., 2006).

In the following, we will consider the classical SMM. In 1975, Moran discovered that the distribution of the absolute frequencies $n_i(t)$ of alleles (as identified by their repeat number i) at time t does not converge, but has bounded variance. He subsequently conjectured that the distribution "remains in a bunch" and characterised its behaviour as "wandering", without being more specific as to the existence of a limiting distribution (Moran, 1975). To investigate convergence, Moran considered quantities $C_k(t) := N^{-2}\sum_i n_i(t)n_{i+k}(t)$, where N is the population size. For $k=0$, this is the "effective number of neutral alleles in the population" of Ohta and Kimura (1973). Moran was able to show that "unlike most problems in population genetics that have been discussed in the past, we do not obtain a limiting distribution or convergence in probability [of $C_k(t)$]." (Moran, 1975). Shortly after Moran's publication, Kingman investigated the normalised Markov chain of the SMM, given by the repeat number difference to the allele of the Nth (or any other) individual in each generation (Kingman, 1976). Using characteristic functions, he could prove exponentially fast convergence in distribution for a generalised model.

* Corresponding author. Tel.: +49 431 597 3199; fax: +49 431 597 3193.
E-mail addresses: caliebe@medinfo.uni-kiel.de (A. Caliebe), jochens@medinfo.uni-kiel.de (A. Jochens), krawczak@medinfo.uni-kiel.de (M. Krawczak), roesler@math.uni-kiel.de (U. Rösler).
[1] Equal contribution of authors.

0022-5193/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jtbi.2010.06.033

He also obtained results about the limiting distribution of samples from a population when the population size tends to infinity conditioned that a certain relationship between time and population size holds.

Here, we will give a detailed analysis of the behaviour of the allele process under the SMM, where our focus will be upon the resulting Markov chain. Markov processes have been applied before to the characterisation of microsatellite mutation models by Watkins (2007). In contrast to Kingman (1976), who used the analytic tool of characteristic functions, we will apply the stochastic method of recurrence of Markov chains. In Section 2 we will introduce the stepwise mutation model which is the basis for all subsequent results. In Section 3 the allele process X is investigated. We will make use of the fact that every population which does not die out, such as under a Wright–Fisher model, contains a genealogical lifeline that does not die out. Adding independent mutations to the genealogy generates an inherent random walk, and thereby results for the marginal distribution of X. In the second subsection, we will show that X is an irreducible, aperiodic and null recurrent Markov chain. The behaviour of X represents the global aspect of the SMM. The normalised allele process V is analysed in Section 4, characterising the local view of the SMM. Again, marginal results such as moments and exponential moments will be given in the first subsection. Then, it will be proven that V is a positive recurrent Markov chain with exponentially fast convergence to the invariant distribution. A central result is provided in the third subsection where it will be shown that the marginal invariant distribution is unimodal. Finally, some simulation results for this distribution are given.

2. Wright–Fisher model with stepwise mutations

The microsatellite allele process under neutral evolution will be studied using a Wright–Fisher model with stepwise mutation. Let $(\Omega,\mathcal{A},\mathbb{P})$ be the underlying probability space, and let $N\in\mathbb{N}:=\{1,2,3,\ldots\}$ be the constant population size. A microsatellite allele will be represented by the number of iterations of the sequence motif, the repeat number. Alleles are normalised such that allele 0 corresponds to a particular basic state $m\in\mathbb{N}$, e.g. the most commonly observed repeat number. For simplicity in the classical SMM, which we apply here, there are no length restrictions on the allele size and even negative repeat numbers are theoretically possible. Thus, the set of possible alleles equals $\mathbb{Z}$, and an allele $z\in\mathbb{Z}$ then has repeat number $m+z$.

a) *Genealogy*: The genealogy is assumed to be given by a Wright–Fisher model. For ease of notation and terminology, we will consider only haploid individuals. However, our results can easily be transferred to diploid individuals by regarding each of their two alleles separately.
Let $Y_n(i)$ be the direct ancestor of the ith individual in the nth generation, $i\in\{1,\ldots,N\}$, $n\in\mathbb{N}$. Clearly, $Y_n(i)$ is an individual of the $(n-1)$th generation. According to the Wright–Fisher model, $Y_n:\Omega\to\{1,\ldots,N\}^N$ satisfies $\mathbb{P}(Y_n(i)=j)=1/N$ for all $i,j\in\{1,\ldots,N\}$ and the $(Y_n(i))_{i\in\{1,\ldots,N\},n\in\mathbb{N}}$ are independent.
b) *Mutation process*: Let $Z_n(i)$ be the mutational event preceding inheritance, from $Y_n(i)$, of the allele of the ith individual in the nth generation. We only consider mutation events that either increase or decrease the repeat number by 1, or leave the repeat number unchanged, i.e. $Z_n(i)\in\{-1,0,1\}$. Let $\mu\in(0,1)$ be the mutation rate, i.e. the probability of a change in repeat number per generation and per individual. Then, $Z_n:\Omega\to\{0,1,-1\}^N$ is assumed to satisfy $\mathbb{P}(Z_n(i)=0)=1-\mu$, $\mathbb{P}(Z_n(i)=1)=\mathbb{P}(Z_n(i)=-1)=\mu/2$.

As usual, we assume an *Independence Property* for the genealogical and mutational processes, namely that

the whole family $Z_n(i),Y_n(i);n\in\mathbb{N},i\in\{1,\ldots,N\}$ is independent. (1)

c) *Allele process*: Let $X_n(i)$ denote the allele of the ith individual in the nth generation. For all $n\in\mathbb{N}_0$, $X_n:\Omega\to\mathbb{Z}^N$ can be written as

$$X_n(\omega)(i):=X_{n-1}(\omega)(Y_n(\omega)(i))+Z_n(\omega)(i)\quad\text{with } X_0\equiv 0. \tag{2}$$

$X:=(X_n)_{n\in\mathbb{N}_0}$ is called the *allele process* of the Wright–Fisher SMM. The distribution of the initial states X_0 is arbitrary and does not influence the asymptotic behaviour. For the sake of simplicity, we assume $X_0\equiv 0$ which means that all alleles have the same repeat number m at time 0.
d) Fundamental properties of the allele process X: Let $\mathcal{A}_n:=\sigma(Y_1,\ldots,Y_n,Z_1,\ldots,Z_n)$ be the σ algebra generated by $Y_1,\ldots,Y_n,Z_1,\ldots,Z_n$. Then the following property follows directly from the definition of X.

Proposition 1.

(i) *X_n is $\mathcal{A}_n$-measurable for all $n\in\mathbb{N}$.*
(ii) *For all $n\in\mathbb{N}$ the family $(X_n(i))_{i\in\{1,\ldots,N\}}$ is exchangeable.* (*Exchangeability Property*)

3. Global behaviour: The allele process X

3.1. Marginal properties of X

To investigate the marginal distribution of the allele process X, we will use an immanent random walk. This is generated by the "lifeline" of the genealogy, i.e. the line of descent that never dies out. J_n is the index, in generation n, of the (unique) member of the lifeline.

Proposition 2.

(i) *There exists an almost surely unique $J:\Omega\to\{1,\ldots,N\}^{\mathbb{N}}$ such that $Y_n(J_n)=J_{n-1}$ for all $n\in\mathbb{N}$, and J_n is $\sigma(Y_k,k\in\mathbb{N},k>n)$ measurable. Furthermore, for $n\in\mathbb{N}$, $X_n(J_n)$ has the same distribution as $X_n(1)$.*
(ii) *$(X_n(J_n))_{n\in\mathbb{N}_0}$ is a random walk. For $k\in\mathbb{Z}$, the transition probabilities are $\mathbb{P}(X_n(J_n)=k|X_{n-1}(J_{n-1})=k)=1-\mu$ and $\mathbb{P}(X_n(J_n)=k-1|X_{n-1}(J_{n-1})=k)=\mathbb{P}(X_n(J_n)=k+1|X_{n-1}(J_{n-1})=k)=\mu/2$.*

Proof. (ii) follows from the definition of X once the existence of J has been established. Let τ_n be the first generation (after n) in which all individuals have a common ancestor in generation n, i.e.

$$\tau_n:=\inf\{k>n:\forall i,j\in\{1,\ldots,N\}\ Y_{n+1}\circ\cdots\circ Y_k(i)=Y_{n+1}\circ\cdots\circ Y_k(j)\}.$$

τ_n is $\sigma(Y_k,k\in\mathbb{N},k>n)$ measurable and almost surely finite.
Then, for $n\in\mathbb{N}_0$, define on $\tau_n<\infty$

$$J_n:=Y_{n+1}\circ Y_{n+2}\circ\cdots\circ Y_{\tau_{n-1}}\circ Y_{\tau_n}(1).$$

J_n is almost surely well defined. For $\tau_n=\tau_{n-1}$ the equality $Y_n(J_n)=J_{n-1}$ is clear. For $\tau_n>\tau_{n-1}$ define $Z:=Y_{\tau_{n-1}+1}\circ\cdots\circ Y_{\tau_n}(1)$. Then

$$Y_n(J_n)=Y_n\circ Y_{n+1}\circ\cdots\circ Y_{\tau_{n-1}}(Z)=J_{n-1}.$$

Hence J_n satisfies the required properties. □

The first and second moment of the marginal distribution of X and a recurrence equation follow immediately from this

proposition and from Proposition 1(i). A limit result for the first absolute moment can be derived by applying the central limit theorem to the random walk $(X_n(J_n))_{n\in\mathbb{N}_0}$. For all $n\in\mathbb{N}_0, z\in\mathbb{Z}$ define

$\rho_n(z) := \mathbb{P}(X_n(i)=z)$.

Note that, owing to the exchangeability property of Proposition 1(ii), $\rho_n(z)$ is independent of the choice of $i\in\{1,\dots,N\}$.

Lemma 3. *For any* $i\in\{1,\dots,N\}$, $n\in\mathbb{N}$

(i) $\mathbb{E}(X_n(i))=0$
(ii) $\mathrm{Var}(X_n(i))=\mu n$
(iii) $\rho_n(z)=(1-\mu)\rho_{n-1}(z)+\mu/2(\rho_{n-1}(z-1)+\rho_{n-1}(z+1))$
(iv) $\lim_{m\to\infty}\mathbb{E}|X_m(i)|=\infty$

Note that $\lim_{n\to\infty}\mathrm{Var}(X_n(i))=\infty$.

Lemma 4. *For any* $i,j\in\{1,\dots,N\}$, $i\neq j$, $n\in\mathbb{N}$

$$\mathrm{Cov}(X_n(i),X_n(j))=\mu\left(n+\frac{(N-1)^n-N^n}{N^{n-1}}\right).$$

A proof of Lemma 4 is given in the appendix.

3.2. Characterisation of X as a Markov chain

The following theorem shows that, in our new representation as a Markov chain, the allele process X is null recurrent (see Breiman, 1992, p. 140, for the definition of null recurrent). Therefore, no asymptotic distribution exists. In the following, we will write 0_N for $(0,\dots,0)\in\mathbb{Z}^N$. For the definition of $\mathcal{A}_n$, see Proposition 1.

Theorem 5.

(i) *The allele process X is an irreducible, aperiodic Markov chain on* $\mathbb{Z}^N$ *with respect to* $(\mathcal{A}_n)_{n\in\mathbb{N}_0}$.
(ii) *The allele process X is null recurrent.*

Proof. (i) follows directly from the definition of X. For the proof of the recurrence, it suffices to verify recurrence for state $0_N\in\mathbb{Z}^N$ because of irreducibility. Remember that $X_0\equiv 0_N$. We will prove the criterion $\sum_{n=1}^{\infty}P(X_n=0_N)=\infty$ (Chung, 1967, p. 23, Theorem 4). One possibility for process X to get from state 0_N at time 0 to state 0_N at time $2n+1$, is that $X_{2n}(1)=0$, $Y_{2n+1}(i)=1$ and $Z_{2n+1}(i)=0$ for all $i\in\{1,\dots,N\}$. Therefore,

$$P(X_{2n+1}=0_N)\geq P(X_{2n}(1)=0)\left(\frac{1}{N}\right)^N(1-\mu)^N.$$

Now choose J according to Proposition 2(i). Then

$$\sum_{n=1}^{\infty}P(X_n=0_N)\geq\sum_{n=1}^{\infty}P(X_{2n}(J_{2n})=0)\left(\frac{1}{N}\right)^N(1-\mu)^N=\infty$$

since the random walk of Proposition 2(ii) is known to be recurrent.

Let $\tau := \inf\{n\in\mathbb{N}: X_n=0_N\}$. For null recurrence, it remains to be shown that $\mathbb{E}(\tau)=\infty$. This follows from $\tau\geq\inf\{n\in\mathbb{N}: X_n(J_n)=0\}$ and from the fact that the random walk of Proposition 2(ii) is null recurrent. □

4. Local behaviour: The normalised allele process *V*

Since no asymptotic distribution exists for the allele process X, we will now consider the normalised allele process V, corresponding to the differences between the repeat numbers of each allele and the allele of the Nth individual in each generation. Note that because of the exchangeability, any other individual may take the place of the Nth individual.

Definition 6. The process $V := (V_n)_{n\in\mathbb{N}_0}$, defined by

$V_n:\Omega\to\mathbb{Z}^{N-1}$ with $V_n(i) := X_n(i)-X_n(N)$,

is called the normalised allele process.

4.1. Marginal properties of V

In this subsection several marginal properties of V are derived. The proofs are given in the appendix.

The first and second moments of the marginal distribution of V can be calculated directly from the corresponding moments of X (see appendix). Because of the exchangeability property, the distribution of $V_n(i)$ is symmetric around zero.

Lemma 7. *For any* $i,j\in\{1,\dots,N-1\}, i\neq j, n\in\mathbb{N}$

(i) $\mathbb{E}(V_n(i))=0$,
(ii) $\mathrm{Var}(V_n(i))=2\mu N(1-(1-1/N)^n)$,
(iii) $\mathrm{Cov}(V_n(i),V_n(j))=\frac{1}{2}\mathrm{Var}(V_n(i))$.

Note that, in contrast to the behaviour of process X (see Lemma 3), $\lim_{n\to\infty}\mathrm{Var}(V_n(i))=2\mu N$ is finite.

We now derive a recursion for the marginal distribution of V. Note that, because of the exchangeability property of Proposition 1(ii), the distribution of $V_n(i)$ is independent of i for $i\leq N-1$. Thus, define

$$\eta_n(z) := \mathbb{P}(V_n(i)=z)\quad\text{for all } n\in\mathbb{N}_0,\ z\in\mathbb{Z}. \tag{3}$$

$$\text{For } k\in\mathbb{Z} \text{ let } r(k) := \mathbb{P}(Z_n(1)-Z_n(2)=k). \tag{4}$$

Obviously r does not depend on n, $r(0)=1-2\mu+\frac{3}{2}\mu^2$, $r(1)=r(-1)=\mu-\mu^2$, $r(2)=r(-2)=\frac{1}{4}\mu^2$ and $r(k)=0$ for any other k.

Lemma 8. *For any* $n\in\mathbb{N}$, $z\in\mathbb{Z}$

$$\eta_n(z)=\frac{N-1}{N}\sum_{k=-2}^{2}r(k)\eta_{n-1}(z-k)+\frac{1}{N}r(z).$$

The next lemma provides a recursion for the higher moments and allows determination of the exponential moments of $V_n(i)$. For $\lambda>0$ and $i\leq N-1$, define $c(\lambda) := \mathbb{E}(\exp(\lambda(Z_n(i)-Z_n(N))))$. Then $c(\lambda)=r(0)+(\exp(\lambda)+\exp(-\lambda))r(1)+(\exp(2\lambda)+\exp(-2\lambda))r(2)$.

Lemma 9. *Let* $i\in\{1,\dots,N-1\}$, $n,m\in\mathbb{N}$, $\lambda>0$.

(i) *All moments of* $V_n(i)$ *are finite and emerge from the following recursion*:
$\mathbb{E}(V_n(i))^m=0$ *for odd* m.
$$\mathbb{E}(V_n(i))^m=\frac{1}{N}(2\mu+\mu^2(2^{m-1}-2))+\left(1-\frac{1}{N}\right)\sum_{\substack{k=0\\ k\ \text{even}}}^{m}\binom{m}{k}(2\mu+\mu^2(2^{m-k-1}-2))\mathbb{E}(V_{n-1}(i))^k$$
for even m.

(ii) *All exponential moments of* $V_n(i)$ *are finite and are given by*
$$\mathbb{E}\exp(\lambda V_n(i))=\frac{1}{N}\sum_{k=0}^{n-1}\left(1-\frac{1}{N}\right)^k c(\lambda)^{k+1}+\left(1-\frac{1}{N}\right)^n c(\lambda)^n$$
$$=\frac{c(\lambda)}{N}\frac{1-\left(1-\frac{1}{N}\right)^n c(\lambda)^n}{1-\left(1-\frac{1}{N}\right)c(\lambda)}+\left(1-\frac{1}{N}\right)^n c(\lambda)^n.$$

The following corollary is straightforward and reveals the behaviour of the moments of $V_n(i)$ for $n\to\infty$.

Corollary 10. *Let* $i\in\{1,\dots,N-1\}$, $m\in\mathbb{N}$, $\lambda>0$.

(i) $\lim_{n\to\infty}\mathbb{E}(V_n(i))^m<\infty$.

(ii)
$$\lim_{n\to\infty}\mathbb{E}\exp(\lambda V_n(i))=\begin{cases}<\infty & \text{if } \left(1-\frac{1}{N}\right)c(\lambda)<1\\ =\infty & \text{if } \left(1-\frac{1}{N}\right)c(\lambda)\geq 1.\end{cases}$$

4.2. Characterisation of V as a Markov chain

Like the original allele process X, the normalised process V is a Markov chain. Contrary to X, however, V can be shown to be positive recurrent (see below; for the definition of positive recurrent see Breiman, 1992, p. 140). Therefore, there is an invariant distribution that characterises the asymptotic behaviour of V, and V can even be shown to converge to this distribution exponentially fast. It should be pointed out that, whereas our Markov chain characterisation of the normalised allele process V is new, the convergence result was already obtained by Kingman, using characteristic functions (Kingman, 1976).

Theorem 11.

(i) *V is an irreducible, aperiodic Markov chain on $\mathbb{Z}^{N-1}$, with respect to $(\mathcal{A}_n)_{n\in\mathbb{N}_0}$.*
(ii) *V is positive recurrent.*
(iii) *V converges exponentially fast to the unique invariant distribution.*

Proof. Using Eq. (2), section (i) follows from the fact that

$$\begin{aligned}V_n(i)&=X_n(i)-X_n(N)=X_{n-1}(Y_n(i))+Z_n(i)-X_{n-1}(Y_n(N))-Z_n(N)\\&=V_{n-1}(Y_n(i))-V_{n-1}(Y_n(N))+Z_n(i)-Z_n(N).\end{aligned}$$

For the proof of (ii) and (iii), write 0_{N-1} for $(0,\dots,0)\in\mathbb{Z}^{N-1}$ and note that, for every $z\in\mathbb{Z}^{N-1}$,

$$\mathbb{P}(V_n=0_{N-1}|V_{n-1}=z)\geq \mathbb{P}(\forall i,j\in\{1,\dots,N\}: Y_n(i)=Y_n(j),Z_n(i)=Z_n(j))>0.$$

Thus, process V fulfills the Doeblin condition and sections (ii) and (iii) follow (see, e.g. Doob, 1953, pp. 192 ff., case b). □

4.3. Unimodality of the asymptotic marginal distribution of V

Theorem 11 implies that the distribution η_n of $V_n(i)$ (see Eq. (3)) converges in distribution as $n\to\infty$. Let $\eta=\lim_{n\to\infty}\eta_n$. We will now show that η is a unimodal discrete distribution, which is one of our main novel results.

Following Keilson and Gerber (1971), we call a distribution p on $\mathbb{Z}$ *unimodal*, if at least one $M\in\mathbb{Z}$ exists such that

$$p(z)\geq p(z-1)\quad\text{for all } z\leq M$$

$$p(z+1)\leq p(z)\quad\text{for all } z\geq M.$$

For proving the unimodality of η we need the following preparatory lemma, the proof of which can be found in the appendix.

Lemma 12. *Let $\mathbb{R}^+$ denote the set of strictly positive real numbers and $\mathbb{R}_0^+$ the set of positive real numbers including zero. If*

$$M:=\{v:\mathbb{Z}\to\mathbb{R}_0^+\,|\,\exists n\in\mathbb{N};a_1,\dots,a_n\in\mathbb{R}^+;b_1,\dots,b_n\in\mathbb{N}_0:\ v=\sum_{i=1}^n a_i\cdot\mathbf{1}_{\{-b_i,-b_i+1,\dots,b_i\}}\},$$

then M is closed under convolution.

With this, we can show that η is unimodal. The critical assumption of the following theorem, namely that $\mu\leq 0.8$, can safely be assumed for microsatellites.

Theorem 13.

(i) *If $r:\mathbb{Z}\to\mathbb{R}$ is defined as in Eq. (4), then $\eta_1=r$ and for all $n\in\mathbb{N}\backslash\{1\}$*
$$\eta_n=\frac{1}{N}\,r*\left(\sum_{i=0}^{n-2}\left(\frac{N-1}{N}\right)^i r^i\right)+\left(\frac{N-1}{N}\right)^{n-1}r^n,\tag{5}$$
where $*$ denotes the convolution of two functions and r^i the ith convolution of r.
(ii) *If $\mu\leq 0.8$, then η is unimodal and symmetric around zero.*

Proof. Recalling that $X_0\equiv 0_N$, it follows that, for all $z\in\mathbb{Z}$,

$$\begin{aligned}\eta_1(z)&=\mathbb{P}(V_1(1)=z)=\mathbb{P}(X_1(1)-X_1(N)=z)\\&=\mathbb{P}(Z_1(1)-Z_1(N)=z)=r(z),\end{aligned}$$

according to the definition of r, see Eq. (4). Since by definition $\sum_{k\in\mathbb{Z}}r(k)\eta_{n-1}(z-k)=(r*\eta_{n-1})(z)$, we can reformulate the recursive equation in Lemma 8 as follows:

$$\eta_n=\frac{N-1}{N}r*\eta_{n-1}+\frac{1}{N}r.\tag{6}$$

We will now prove Eq. (5) by induction. For $n=2$, Eq. (5) follows from Eq. (6). Now, let $n\in\mathbb{N}\backslash\{1,2\}$. Assuming that Eq. (5) holds for $n-1$, we have

$$\begin{aligned}\eta_n&=\frac{N-1}{N}r*\eta_{n-1}+\frac{1}{N}r\\&=\frac{1}{N}r+\frac{N-1}{N}r*\left(\frac{1}{N}r*\sum_{i=0}^{n-3}\left(\frac{N-1}{N}\right)^i r^i+\left(\frac{N-1}{N}\right)^{n-2}r^{n-1}\right)\\&=\frac{1}{N}r*\left(1+\sum_{i=0}^{n-3}\left(\frac{N-1}{N}\right)^{i+1}r^{i+1}\right)+\left(\frac{N-1}{N}\right)^{n-1}r^n\\&=\frac{1}{N}r*\sum_{i=0}^{n-2}\left(\frac{N-1}{N}\right)^i r^i+\left(\frac{N-1}{N}\right)^{n-1}r^n.\end{aligned}$$

To prove section (ii), we will first show that η_n is unimodal for all $n\in\mathbb{N}$. Since $0<\mu\leq 0.8$, the following inequalities hold:

$$1-2\mu+\tfrac{3}{2}\mu^2>\mu-\mu^2\geq\tfrac{1}{4}\mu^2.$$

Thus, in the notation of Lemma 12, $r\in M$ with $n=3$, $b_1=0$, $b_2=1$, $b_3=2$, $a_1=1-2\mu+\frac{3}{2}\mu^2-(\mu-\mu^2)$, $a_2=-\frac{1}{4}\mu^2+(\mu-\mu^2)$, $a_3=\frac{1}{4}\mu^2$. Now, considering Eq. (5), Lemma 12 implies that $\eta_n\in M$ for all $n\in\mathbb{N}$, and all elements of M are clearly unimodal. Unimodality of η follows from the fact that the limit of a convergent sequence of unimodal discrete distributions is itself unimodal (see "Statement 4" in Keilson and Gerber, 1971). □

4.4. Simulation of the asymptotic marginal distribution η

Lemma 8 can be used to simulate the marginal distribution of $V_n(i)$. From Theorem 11, we know that $V_n(i)$ converges in distribution as $n\to\infty$. Figs. 1 and 2 show the behaviour in time of the distribution of $V_n(i)$, assuming $\mu=0.01$ and either $N=100$

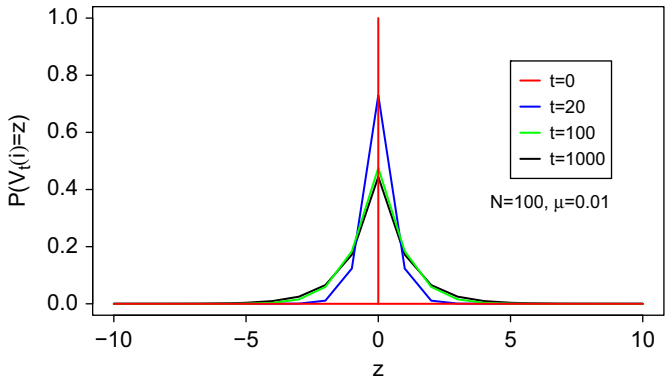

Fig. 1. Convergence of the marginal distribution of the normalised allele process V. For illustration, the discrete probabilities $\mathbb{P}(V_t(i)=z)$ obtained for integer z are connected by lines. $N=100$, $\mu=0.01$, t: number of generations.

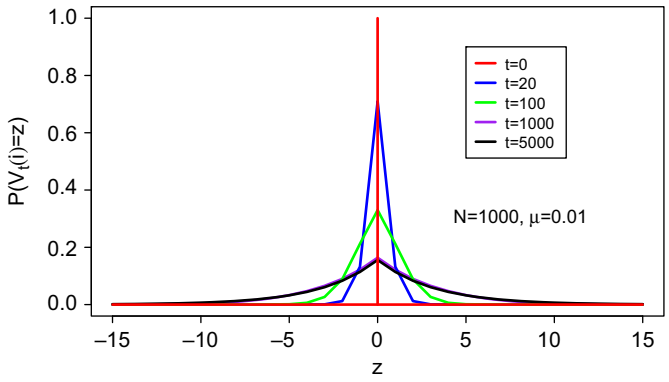

Fig. 2. Convergence of the marginal distribution of the normalised allele process V. For illustration, the discrete probabilities $\mathbb{P}(V_t(i)=z)$ obtained for integer z are connected by lines. $N=1000$, $\mu=0.01$, t: number of generations.

or 1000, respectively. For $N=100$, the distribution of $V_n(i)$ is close to the stationary distribution at $n=100$, and the domain is mainly concentrated in the interval $[-7,7]$. For $N=1000$, convergence is slower and the domain is larger. The distribution of $V_n(i)$ is close to the stationary distribution at $n=1000$, and the domain is mainly concentrated in $[-13,13]$.

5. Discussion

We have shown that the allele process of the stepwise mutation model is characterised by two different types of behaviour. The expectation of the absolute value of the repeat number of a given individual converges to infinity. This signifies the global behaviour, where no convergence occurs. However, when the allelic state of an individual is chosen as a reference point for the other individuals of the population, then a limiting invariant distribution of the resulting allele difference process emerges. This is the local behaviour of the allele process, which implies that the alleles stay "clumped together" during convergence to infinity. These results confirm Moran's notion of the term "wandering distributions" (Moran, 1975).

The convergence of the allelic differences is exponentially fast, as was already noted by Kingman (1976). This is reassuring because it means that estimates or test statistics obtained from the allele differences not only approach a limiting distribution, but do so very quickly. As we showed, the resulting limiting marginal distribution is unimodal.

It should be noted that the SMM is a very simple model of microsatellite mutation. In some cases, it would be reasonable to assume not only mutations that change the repeat number by one unit but to allow a wider range of mutations (Huang et al., 2002). Kingman also considered generalised forms of mutations (Kingman, 1976). As long as the individual mutation events $Z_n(i)$ remain independent, which is biologically plausible, central Theorem 11 of this paper will hold true. If the random walk corresponding to the mutation process Z is null recurrent, Theorem 5 will apply. Another limitation of the SMM is the unboundedness of the state space whereas, in reality, negative repeat numbers cannot occur. Also, very large repeat numbers can result in physically unstable microsatellites and stop the evolutionary process at certain thresholds. One way to account for these limitations would be to restrict the state space of the allele process X by reflecting boundaries. The result would be a Markov chain with finite state space, and convergence to an invariant distribution would follow even for the non-normalised process X. However, differences between the normalised and non-normalised behaviour of the allele process remain possible, for instance, in the form of different convergence rates or different shapes of the invariant distribution. Because of the Markov structure and the assumed one-unit-up-or-down mutations, the process would only "realise" the existence of boundaries when it would be very close to them. Most of the time, the process would stay away from the boundaries and behave according to the stationary distribution of the normalised process V, as if no boundaries would exist.

Regarding the simplicity of the SMM, our results are only a first step towards a better understanding of the real-life situation, and investigations of how the allele process behaves under more realistic models incorporating, for example, variable mutation rates or migration, are warranted.

Acknowledgements

This work was partly funded by the German Ministry of Science and Education (BMBF) through an NGFN SMP-GEM Grant to Michael Krawczak (01GS0426).

Appendix

In order to calculate the respective covariances of Lemma 4 from Eq. (2), we need the following little lemma that can be proven by induction.

Lemma 14. *Let $a,b\in\mathbb{R}$, $b\neq 1$. Then, for the real valued sequence $(x_n)_{n\in\mathbb{N}_0}$ defined by $x_n=(n-1)a+bx_{n-1}$ and $x_0=0$,*

$$x_n = a\frac{b^n+n(1-b)-1}{(1-b)^2}.$$

Proof of Lemma 4. Note that the exchangeability property implies that $\mathbb{P}(X_n(i)=y,X_n(j)=z)$ is independent of $i,j\in\{1,\ldots,N\}$ as long as $i\neq j$. We can calculate a recursion for the covariances using Eq. (2), Proposition 1(i), Lemma 3(i), (ii) and the independence property (1):

$$\begin{aligned}\mathrm{Cov}(X_n(i),X_n(j)) &= \mathrm{Cov}(X_{n-1}(Y_n(i)),X_{n-1}(Y_n(j)))\\ &= \sum_{k,l=1}^{N}\mathrm{Cov}(X_{n-1}(k),X_{n-1}(l))\mathbb{P}(Y_n(i)=k,Y_n(j)=l)\\ &= \frac{1}{N^2}\sum_{k,l=1}^{N}\sum_{y,z\in\mathbb{Z}} yz\,\mathbb{P}(X_{n-1}(k)=y,X_{n-1}(l)=z)\end{aligned}$$

$$= \frac{1}{N^2}\sum_{k=1}^{N}\sum_{y\in\mathbb{Z}} y^2\,\mathbb{P}(X_{n-1}(k)=y) + \frac{1}{N^2}\sum_{\substack{k,l=1\\k\neq l}}^{N}\sum_{y,z\in\mathbb{Z}} yz\,\mathbb{P}(X_{n-1}(k)=y,X_{n-1}(l)=z)$$
$$= \frac{1}{N}\mathrm{Var}(X_{n-1}(1)) + \frac{N-1}{N}\mathrm{Cov}(X_{n-1}(i),X_{n-1}(j))$$
$$= \frac{1}{N}\mu(n-1) + \frac{N-1}{N}\mathrm{Cov}(X_{n-1}(i),X_{n-1}(j)).$$

Therefore, from Lemma 14,

$$\mathrm{Cov}(X_n(i),X_n(j)) = \frac{\mu}{N}\frac{\left(\frac{N-1}{N}\right)^n + n\left(1-\frac{N-1}{N}\right)-1}{\left(1-\frac{N-1}{N}\right)^2} = \mu\left(n+\frac{(N-1)^n-N^n}{N^{n-1}}\right). \quad \square$$

Proof of Lemma 7. (i) follows directly from the definition of V. Variance and covariance can be derived using Lemmas 3 and 4:

$$\mathrm{Var}(V_n(i)) = \mathrm{Var}(X_n(i)) + \mathrm{Var}(X_n(N)) - 2\mathrm{Cov}(X_n(i),X_n(N)) = 2\mu n - 2\mu\left(n+\frac{(N-1)^n-N^n}{N^{n-1}}\right)$$
$$\mathrm{Cov}(V_n(i),V_n(j)) = \mathrm{Var}(X_n(N)) - \mathrm{Cov}(X_n(N),X_n(j)) - \mathrm{Cov}(X_n(i),X_n(N)) + \mathrm{Cov}(X_n(i),X_n(j)) = \mathrm{Var}(X_n(N)) - \mathrm{Cov}(X_n(N),X_n(j)) \quad \square$$

Proof of Lemma 8. Using recursion (2), it follows that

$$\begin{aligned}\eta_n(z) &= \mathbb{P}(X_n(1)-X_n(N)=z) = \mathbb{P}(X_n(1)-X_n(2)=z)\\ &= \mathbb{P}(X_{n-1}(Y_n(1))+Z_n(1)-(X_{n-1}(Y_n(2))+Z_n(2))=z)\\ &= \mathbb{P}(Y_n(1)\neq Y_n(2))\cdot\mathbb{P}(X_{n-1}(Y_n(1))+Z_n(1)-X_{n-1}(Y_n(2)) - Z_n(2)=z|Y_n(1)\neq Y_n(2)) + \mathbb{P}(Y_n(1)=Y_n(2))\cdot\mathbb{P}(Z_n(1) - Z_n(2)=z|Y_n(1)=Y_n(2))\\ &= \frac{N-1}{N}\sum_{k\in\mathbb{Z}}\mathbb{P}(Z_n(1)-Z_n(2)=k)\cdot\mathbb{P}(X_{n-1}(Y_n(1)) + Z_n(1)-X_{n-1}(Y_n(2))-Z_n(2) = z|Y_n(1)\neq Y_n(2),Z_n(1)-Z_n(2)=k) + \frac{1}{N}\mathbb{P}(Z_n(1)-Z_n(2)=z)\\ &= \frac{N-1}{N}\sum_{k\in\mathbb{Z}} r(k)\mathbb{P}(X_{n-1}(1)-X_{n-1}(2)=z-k) + \frac{1}{N}\,r(z)\\ &= \frac{N-1}{N}\sum_{k\in\mathbb{Z}} r(k)\eta_{n-1}(z-k) + \frac{1}{N}\,r(z). \quad \square\end{aligned}$$

Proof of Lemma 9. (i): From Eq. (2), we obtain

$$\begin{aligned}\mathbb{E}(V_n(i))^m &= \mathbb{P}(Y_n(i)=Y_n(N))\mathbb{E}(Z_n(i)-Z_n(N))^m + \sum_{\substack{k,l=1\\k\neq l}}^{N}\mathbb{P}(Y_n(i)=k,Y_n(N)=l)\\ &\cdot\mathbb{E}(X_{n-1}(k)+Z_n(i)-X_{n-1}(l)-Z_n(N))^m\\ &= \frac{1}{N}\mathbb{E}((Z_n(i)-Z_n(N))^m) + \left(1-\frac{1}{N}\right)\mathbb{E}((V_{n-1}(i)+Z_n(i)-Z_n(N))^m)\\ &= \frac{1}{N}\mathbb{E}((Z_n(i)-Z_n(N))^m) + \left(1-\frac{1}{N}\right)\sum_{k=0}^{m}\binom{m}{k}\mathbb{E}((Z_n(i)-Z_n(N))^{m-k})\mathbb{E}(V_{n-1}(i))^k.\end{aligned}$$

Using $\mathbb{E}((Z_n(i)-Z_n(N))^m)=0$ for m odd and $\mathbb{E}((Z_n(i)-Z_n(N))^m) = 2\mu+\mu^2(2^{m-1}-2)$ for m even, section (i) follows by induction and Lemma 7.

(ii): Treating the exponential moments in the same way yields

$$\begin{aligned}d_n &:= \mathbb{E}(\exp(\lambda V_n(i)))\\ &= \frac{1}{N}\mathbb{E}(\exp(\lambda(Z_n(i)-Z_n(N)))) + \left(1-\frac{1}{N}\right)\mathbb{E}(\exp(\lambda V_{n-1}))\mathbb{E}(\exp(\lambda(Z_n(i)-Z_n(N))))\\ &= \frac{1}{N}c(\lambda) + \left(1-\frac{1}{N}\right)c(\lambda)d_{n-1}. \quad \square\end{aligned}$$

Proof of Lemma 12. Let $*$ denote the convolution of two functions and define $f_b := \mathbf{1}_{\{-b,-b+1,\ldots,b\}}$ for $b\in\mathbb{N}_0$. First note that, for all $b\le b'\in\mathbb{N}_0$ and for all $z\in\mathbb{Z}$, the following equation holds

$$(f_b * f_{b'})(z) = \sum_{i\in\mathbb{Z}} f_b(i)\cdot f_{b'}(z-i) = \sum_{i=-b}^{b} f_{b'}(z-i) = \sum_{i=-b}^{b}\mathbf{1}_{\{-b'+i,\ldots,b'+i\}}(z).$$

Dropping argument z and decomposing the sum on the right-hand side,

$$\begin{aligned}f_b * f_{b'} &= f_{b'} + \sum_{i=1}^{b}\mathbf{1}_{\{-b'+i,\ldots,b'+i\}} + \sum_{i=-b}^{-1}\mathbf{1}_{\{-b'+i,\ldots,b'+i\}}\\ &= f_{b'} + \sum_{i=1}^{b}(\mathbf{1}_{\{-b'+i,\ldots,b'+i\}} + \mathbf{1}_{\{-b'-i,\ldots,b'-i\}})\\ &= f_{b'} + \sum_{i=1}^{b}(f_{b'+i} + f_{b'-i}).\end{aligned}$$

Now let $v,v'\in M$, which can be written as $v=\sum_{i=1}^{n} a_i\cdot f_{b_i}$ and $v'=\sum_{i=1}^{n'} a'_i\cdot f_{b'_i}$. Let $m(ij) := \max\{b_i,b'_j\}$. Taking into account the distributivity and linearity of the convolution, the above implies that

$$\begin{aligned}v * v' &= \left(\sum_{i=1}^{n} a_i\cdot f_{b_i}\right) * \left(\sum_{i=1}^{n'} a'_i\cdot f_{b'_i}\right) = \sum_{i=1}^{n}\sum_{j=1}^{n'} a_i a'_j\cdot(f_{b_i} * f_{b'_j})\\ &= \sum_{i=1}^{n}\sum_{j=1}^{n'} a_i a'_j\cdot\left(f_{m(ij)} + \sum_{k=1}^{\min\{b'_j,b_i\}}(f_{m(ij)+k} + f_{m(ij)-k})\right)\\ &= \sum_{i=1}^{n}\sum_{j=1}^{n'} a_i a'_j\cdot f_{m(ij)} + \sum_{i=1}^{n}\sum_{j=1}^{n'}\sum_{k=1}^{\min\{b'_j,b_i\}} a_i a'_j\cdot f_{m(ij)+k} + \sum_{i=1}^{n}\sum_{j=1}^{n'}\sum_{k=1}^{\min\{b'_j,b_i\}} a_i a'_j\cdot f_{m(ij)-k}.\end{aligned}$$

Since all the characteristic functions in the last expression are symmetrical around zero, it follows that $v * v'\in M$. $\square$

References

Ashley, C.T., Warren, S.T., 1995. Trinucleotide repeat expansion and human disease. Ann. Rev. Genet. 29, 703–728.

Bindu, G.H., Trivedi, R., Kashyap, V.K., 2007. Allele frequency distribution based on 17 STR markers in three major Dravidian linguistic populations of Andhra Pradesh, India. Forensic Sci. Int. 170, 76–85.

Breiman, L., 1992. Probability. SIAM, Philadelphia.

Calabrese, P.P., Durrett, R.T., Aquadro, C.F., 2001. Dynamics of microsatellite divergence and proportional slippage/point mutation models. Genetics 159, 839–852.

Calabrese, P.P., Sainudiin, R., 2005. Models of microsatellite evolution. In: Nielsen, R. (Ed.), Statistical Methods in Molecular Evolution. Springer, London, pp. 289–306.

Cassidy, B.G., Gonzales, R.A., 2005. DNA testing in animal forensics. J. Wildl. Manage. 69, 1454–1462.

Chambers, G.K., MacAvoy, E.S., 2000. Microsatellites: consensus and controversy. Comp. Biochem. Physiol. B 126, 455–476.

Chung, K.L., 1967. Markov Chains with Stationary Transition Probabilities, second ed. Springer, Berlin.
Cornuet, J.M., Beaumont, M.A., Estoup, A., Solignac, M., 2006. Inference on microsatellite mutation processes in the invasive mite, Varroa destructor, using reversible jump Markov chain Monte Carlo. Theor. Popul. Biol. 69, 129–144.
De Iorio, M., Griffiths, R.C., Leblois, R., Rousset, F., 2005. Stepwise mutation likelihood computation by sequential importance sampling in subdivided population models. Theor. Popul. Biol. 68, 41–53.
Di Rienzo, A., Peterson, A.C., Garza, J.C., Valdes, A.M., Slatkin, M., Freimer, N.B., 1994. Mutational processes of simple-sequence repeat loci in human populations. Proc. Natl. Acad. Sci. USA 91, 3166–3170.
Doob, J.L., 1953. Stochastic Processes. John Wiley, New York reprinted Wiley Classics Library Edition 1990.
Durrett, R., Kruglyak, S., 1999. A new stochastic model of microsatellite evolution. J. Appl. Probab. 36, 621–631.
Falush, D., Iwasa, Y., 1999. Size-dependent mutability and microsatellite constraints. Mol. Biol. Evol. 16, 960–966.
Feldman, M.W., Bergman, A., Pollock, D.D., Goldstein, D.B., 1997. Microsatellite genetic distances with range constraints: analytic description and problems of estimation. Genetics 145, 207–216.
Garza, J.C., Slatkin, M., Freimer, N.B., 1995. Microsatellite allele frequencies in humans and chimpanzees, with implications for constraints on allele size. Mol. Biol. Evol. 12, 594–603.
Goldstein, D.B., Roemer, G.W., Smith, D.A., Reich, D.E., Bergman, A., Wayne, R.K., 1999. The use of microsatellite variation to infer population structure and demographic history in a natural model system. Genetics 151, 797–801.
Huang, Q.-Y., Xu, F.-H., Shen, H., Deng, H.-Y., Liu, Y.-J., Liu, Y.-Z., Li, J.-L., Recker, R.R., Deng, H.-W., 2002. Mutation patterns at dinucleotide microsatellite loci in humans. Am. J. Hum. Genet. 70, 625–634.
Kashi, Y., King, D.G., 2006. Simple sequence repeats as advantageous mutators in evolution. Trends Genet. 22, 253–259.
Keilson, J., Gerber, H., 1971. Some results for discrete unimodality. J. Am. Stat. Assoc. 66, 386–389.
Kingman, J.F.C., 1976. Coherent random walks arising in some genetical models. Proc. R. Soc. London A 351, 19–31.
Kruglyak, S., Durrett, R.T., Schug, M.D., Aquadro, C.F., 1998. Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations. Proc. Natl. Acad. Sci. USA 95, 10774–10778.
Li, Y.-C., Korol, A.B., Tzion, F., Beiles, A., Nevo, E., 2002. Microsatellites: genomic distribution, putative functions and mutational mechanisms: a review. Mol. Ecol. 11, 2453–2465.
Moran, P.A.P., 1975. Wandering distributions and the electrophoretic profile. Theor. Popul. Biol. 8, 318–330.
Ohta, T., Kimura, M., 1973. A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. Genet. Res. 22, 201–204.
Sainudiin, R., Durrett, R.T., Aquadro, C.F., Nielsen, R., 2004. Microsatellite mutation models: insights from a comparison of humans and chimpanzees. Genetics 168, 383–395.
Tautz, D., 1993. Notes on the definition and nomenclature of tandemly repetive DNA sequences. In: Pena, S.D.J., Chakraborty, R., Epplen, J.T., Jeffreys, A.J. (Eds.), DNA Fingerprinting: State of the Science. Birkhäuser Verlag, Basel, pp. 21–28.
Thibodeau, S.N., Bren, G., Schaid, D., 1993. Microsatellite instability in cancer of the proximal colon. Science 260, 816–819.
Tishkoff, S.A., Dietzsch, E., Speed, W., Pakstis, A.J., Kidd, J.R., Cheung, K., Bonné-Tamir, B., Santachiara-Benerecetti, A.S., Moral, P., Krings, M., Pääbo, S., Watson, E., Risch, N., Jenkins, T., Kidd, K.K., 1996. Global patterns of linkage disequilibrium at the CD4 locus and modern human origins. Science 271, 1380–1387.
Vardo, A.M., Schall, J.J., 2007. Clonal diversity of a lizard malaria parasite, Plasmodium mexicanum, in its vertebrate host, the western fence lizard: role of variation in transmission intensity over time and space. Mol. Ecol. 16, 2712–2720.
Watkins, J.C., 2007. Microsatellite evolution: Markov transition functions for a suite of models. Theor. Popul. Biol. 71, 147–159.
Weissenbach, J., Gyapay, G., Dib, C., Vignal, A., Morissette, J., Millasseau, P., Vaysseix, G., Lathrop, M., 1992. A second-generation linkage map of the human genome. Nature 359, 794–801.
Whittaker, J.C., Harbord, R.M., Boxall, N., Mackay, I., Dawson, G., Sibly, R.M., 2003. Likelihood-based estimation of microsatellite mutation rates. Genetics 164, 781–787.
Zhivotovsky, L.A., Feldman, M.W., Grishechkin, S.A., 1997. Biased mutations and microsatellite variation. Mol. Biol. Evol. 14, 926–933.
Zhivotovsky, L.A., Rosenberg, N.A., Feldman, M.W., 2003. Features of evolution and expansion of modern humans, inferred from genomewide microsatellite markers. Am. J. Hum. Genet. 72, 1171–1186.

2.2 Empirical Evaluation Reveals Best Fit of a Logistic Mutation Model for Human Y-Chromosomal Microsatellites

Jochens, Caliebe, Rösler und Krawczak (2011)
Genetics 189:4, 1403–1411.

Zusammenfassung

Es ist bekannt, dass STR-Mutationswahrscheinlichkeiten sowohl von der Allel-Länge als auch vom Repeat-Motiv abhängen und auch darüber hinaus von Locus zu Locus variieren (siehe Abschnitt 1.4). Insofern besteht ein vielversprechender Ansatz zum Vergleich der zahlreichen vorgeschlagenen Mutationsmodelle (Abschnitt 1.4.2) darin, die Modelle für jeden Locus separat anzupassen. Dies erfordert eine große Menge locusspezifischer Mutationsdaten und ist uns mit dieser Publikation erstmals gelungen.

Zu diesem Zweck haben wir 24 veröffentlichte Studien ausgewertet, die Y-STR-Daten von Vater-Sohn-Paaren enthalten. Insgesamt konnten wir so 15.285 Vater-Sohn-Paare aggregieren, die sich jeweils als eine direkt beobachtete Meiose auffassen lassen. Für die Analyse konzentrieren wir uns auf sechs Kernloci (DYS19, DYS389 I, DYS390, DYS391, DYS392 und DYS393), die in nahezu allen Individuen typisiert wurden. An diesen Loci wurden insgesamt 162 Mutationen beobachtet, wobei drei Mutationen den Betrag zwei aufwiesen und der Rest Einschritt-Mutationen waren.

Wir konzentrieren uns auf solche Mutationsmodelle, die sich als Übergangswahrscheinlichkeiten von Markov-Ketten beschreiben lassen, d.h. die Mutationswahrscheinlichkeit an einem gegebenen Locus hängt nur vom Allel des Vaters ab. Außerdem beschränken wir uns in Anbetracht der Daten auf Einschritt-Mutationsmodelle. Zu der so definierten Modellklasse gehören das einfache schrittweise Mutationsmodell (SMM, siehe Abschnitt 1.4.2), das Lineare Modell (in dem die Mutationswahrscheinlichkeit linear mit der Allel-Länge des Vaters zunimmt) und das Logistische Modell, das wir in dieser Publikation einführen.

Unser Logistisches Modell mit den Parametern α, β und γ beruht auf der Idee, dass der Zusammenhang zwischen Vater-Allel und (Auf- bzw. Abwärts-) Mutationswahrscheinlichkeit durch eine logistische Funktion approximiert wird (siehe Abbildung 1.2). Somit ist die Mutationswahrscheinlichkeit für kurze Allele nahezu Null und steigt mit zunehmender Allel-Länge zunächst langsam an. Bei einer mittleren Allel-Länge, die dem Modellparameter β entspricht, ist dieser Anstieg dann maximal. Wie groß der Anstieg an dieser Stelle genau ist, wird durch den Parameter α kontrolliert. Für längere Allele wird die Steigung wieder geringer, und mit weiter zunehmender Allel-Länge konvergiert die Mutationswahrscheinlichkeit gegen einen festen Wert, der dem Modellparameter $\gamma/2$ entspricht.

Für jeden Locus verwenden wir einen Maximum-Likelihood-Ansatz, um die verschiedenen Modelle an die Daten anzupassen, d.h. ihre Parameter zu schätzen. Mittels Akaikes Informationskriterium (*Akaike Information Criterion*, AIC) lässt sich dann beurteilen, welches Modell die Daten am besten erklärt. Für fünf der sechs betrachteten Kernloci zeigte sich, dass eine bestimmte Version unseres neuen Logistischen Modells am besten passt. Diese Version wird dadurch charakterisiert, dass der Parameter γ für Auf- und Abwärtsmutationen gleich ist, während α und β sich zwischen Auf- und Abwärtsmutationen unterscheiden. Für DYS392 passte ein lineares Modell am besten, in dem Aufwärts- und Abwärts-Mutationswahrscheinlichkeiten gleichermaßen linear mit der Allel-Länge zunehmen.

Empirical Evaluation Reveals Best Fit of a Logistic Mutation Model for Human Y-Chromosomal Microsatellites

Arne Jochens,*,[1] Amke Caliebe,* Uwe Rösler,[†] and Michael Krawczak*
*Institut für Medizinische Informatik und Statistik, Christian-Albrechts-Universität zu Kiel, 24105 Kiel, Germany, and
[†]Mathematisches Seminar, Christian-Albrechts-Universität zu Kiel, 24098 Kiel, Germany

ABSTRACT The rate of microsatellite mutation is dependent upon both the allele length and the repeat motif, but the exact nature of this relationship is still unknown. We analyzed data on the inheritance of human Y-chromosomal microsatellites in father–son duos, taken from 24 published reports and comprising 15,285 directly observable meioses. At the six microsatellites analyzed (DYS19, DYS389I, DYS390, DYS391, DYS392, and DYS393), a total of 162 mutations were observed. For each locus, we employed a maximum-likelihood approach to evaluate one of several single-step mutation models on the basis of the data. For five of the six loci considered, a novel logistic mutation model was found to provide the best fit according to Akaike's information criterion. This implies that the mutation probability at the loci increases (nonlinearly) with allele length at a rate that differs between upward and downward mutations. For DYS392, the best fit was provided by a linear model in which upward and downward mutation probabilities increase equally with allele length. This is the first study to empirically compare different microsatellite mutation models in a locus-specific fashion.

"MICROSATELLITES", also known as "short tandem repeats" (STRs), are stretches of DNA in which a short piece of sequence (1–6 bp, the repeat motif) is tandemly repeated a number of times. Microsatellites are abundant throughout the human genome (Lander *et al.* 2001). They have high mutation rates, which renders them useful for a number of practical applications, including genetic epidemiology (Jobling and Tyler-Smith 2003), gene mapping (Weissenbach 1993), studies of population history (Jobling and Tyler-Smith 2003), genetic fingerprinting (Krawczak and Schmidtke 1998), kinship testing (Weir *et al.* 2006), and genetic ancestry analysis (Shriver and Kittles 2004). Microsatellites also may have played a major role in genome evolution (Kashi and King 2006). Microsatellite mutations typically comprise the gain or loss of 1 repeat unit, but larger changes of the repeat number in a single meiosis also have been observed (see below).

doi: 10.1534/genetics.111.132308
Manuscript received July 6, 2011; accepted for publication September 23, 2011
[1]Corresponding author: Institut für Medizinische Informatik und Statistik, Christian-Albrechts-Universität zu Kiel, Haus 31, Arnold-Heller-Str. 3, 24105 Kiel, Germany. E-mail: jochens@medinfo.uni-kiel.de

Various statistical models have been proposed for the microsatellite mutation process (Calabrese and Sainudiin 2005), but no model has yet been identified as representing the single best, thereby testifying to the complexity of the mutation process. In any case, it is well established that the mutation rate of microsatellites (i) depends upon the respective repeat number (Xu *et al.* 2000; Lai and Sun 2003; Kelkar *et al.* 2008) and (ii) varies greatly across loci, which is partly due to the variable nature of the repeat motif itself (Kelkar *et al.* 2008).

The stepwise mutation model (SMM) was the first one used to describe microsatellite evolution. The SMM implies that each mutation adds or deletes 1 repeat unit at a time with a symmetric mutation rate that is independent of the repeat number (Ohta and Kimura 1973). The SMM was used in simulation-based attempts to explain the observed population patterns of microsatellite allele frequencies (Shriver *et al.* 1993; Valdes *et al.* 1993). Since then, various modifications of the SMM have been proposed. For example, the two-phase model (Di Rienzo *et al.* 1994) combines the SMM with geometrically distributed mutational changes of more than 1 repeat unit. This model was itself generalized (Garza *et al.* 1995), allowing for any prespecified distribution of the repeat number change. Other approaches included lower

and upper bounds for the allele length (Nauta and Weissing 1996; Feldman *et al.* 1997).

Several microsatellite mutation models have been proposed to account for the relationship between the mutation rate and the respective allele length (Falush and Iwasa 1999). More restricted versions of this approach entail the idea that only the difference in repeat number before and after mutation, but not the mutation rate itself, depends upon the allele length (Garza *et al.* 1995; Kimmel *et al.* 1996). Kruglyak, Durrett, and co-workers (Kruglyak *et al.* 1998; Durrett and Kruglyak 1999) combined the SMM with the possibility of (rare) point mutations that reduce the repeat number. This model, later expanded (Calabrese *et al.* 2001; Dieringer and Schlötterer 2003), is motivated by the idea that a point mutation divides a microsatellite into two parts, only one of which is kept track of.

Finally, Whittaker *et al.* (2003) proposed an exponential microsatellite mutation model with parameters α_u, α_d, γ_u, γ_d, λ_u, and λ_d from an unspecified parameter space. Their transition probabilities (*cf. Materials and Methods*) from paternal allele i to offspring allele j were as follows:

$$p_{ij} = \begin{cases} \gamma_u \cdot \exp(\alpha_u \cdot i - \lambda_u \cdot (j-i)) & \text{if} \quad j > i, \\ \gamma_d \cdot \exp(\alpha_d \cdot i - \lambda_d \cdot (i-j)) & \text{if} \quad j < i, \\ 1 - \sum\limits_{k \neq i} p_{ik} & \text{if} \quad j = i. \end{cases}$$

However, in mathematical terms this model is not well defined because, for fixed parameters with $\alpha_u, \gamma_u > 0$, $p_{i,i+1}$ may exceed 1 for large i, and $p_{i,i+1}$ even approaches infinity as $i \to \infty$. Nevertheless, the model can be remedied if written as an instantaneous rate matrix over a finite state space as in Calabrese and Sainudiin (2005).

In the present study, we compared different models of microsatellite mutation, taking into account the locus dependency of the mutation process by evaluating each model seperately for each locus. We chose to confine our study to microsatellites from the male-specific region of the human Y chromosome (Y-STRs) for which mutations can be inferred directly from genotype data on father–son duos. A large amount of such data have become available for opportunistic use in scientific analyses because large numbers of father–son pairs are regularly genotyped in paternity cases. Moreover, these data allow us to directly infer the one-step probability matrix of the discrete-time Markov chain model scaled in generations, as opposed to the continuous-time Markov process usually applied to evolutionary genomic data.

Materials and Methods

Y-STRs

In forensics as well as for other practical applications, a few commercially available, PCR-based kits are widely used for microsatellite genotyping (Mayntz-Press and Ballantyne 2007). The Y chromosome is of particular interest in forensics because it allows easy discrimination between male and female contributions to traces and trace mixtures. This has led to a situation where a few Y-STRs are routinely genotyped in a vast number of individuals (Roewer 2009). Table 1 lists the six Y-STRs that have been included in the present study, along with their most important molecular characteristics. For labeling Y-STR loci and alleles, we employed the nomenclature recommended by the International Society of Forensic Genetics (Gusmão *et al.* 2006). In particular, the allele designations refer to the actual number of repeats rather than the crude PCR fragment length.

Table 1 Y-STR loci

Name	μ [10^{-3}]	Alleles	Repeat motif
DYS19	2.3	14 (9–19)	TAGA
DYS389I	2.5	13 (9–17)	TCTG/TCTA
DYS390	2.1	24 (17–29)	TCTG/TCTA
DYS391	2.6	10 (5–16)	TCTA
DYS392	0.4	11 (4–20)	TAT
DYS393	1.0	13 (7–18)	AGAT

Characteristics of six Y-STR loci considered in the present study are shown. μ is the relative frequency of observed mutations across all paternal alleles for each locus. The "Alleles" column contains the most frequently observed repeat number, and the range is given in parantheses. Data are from the Y Chromosome Haplotype Reference Database (Willuweit and Roewer 2007), Release 33.

Mutation data

Table 2 lists 24 published reports of directly observed Y-STR mutations (or the lack thereof) in father–son pairs, which formed the basis of our present study. Taken together, these reports comprised 15,285 father–son pairs in whom 162 mutations were observed across the six Y-STRs considered. Paternity was confirmed in all cases by independent genotyping of autosomal loci. All studies were original reports so that there should be no overlap between the samples included. Most of the recent reports included the population frequencies of the paternal alleles along with the mutation data, as was recommended by the International Society of Forensic Genetics in 2006 (Gusmão *et al.* 2006). This information is required for fitting mutation models with allele-dependent mutation rates. For studies that did not provide paternal allele frequencies, we tried to approximate these as follows: For four studies (Kayser *et al.* 2000; Budowle *et al.* 2005; Ge *et al.* 2009; Goedbloed *et al.* 2009), estimates were equated to the respective allele frequencies as reported for the relevant subpopulations in the Y Chromosome Haplotype Reference Database (Willuweit and Roewer 2007), Release 31. For the remaining studies, we adopted the allele frequencies reported in the studies themselves, although these estimates included not only fathers, but also unrelated individuals.

Mutation models

The microsatellite mutation models considered here share a number of characteristics. Thus, alleles are consistently represented by an integer corresponding to the number of

Table 2 Reports on father–son duo Y-STR data

Reference	Sample origins	*N*	*n*
Bianchi *et al.* (1998)	United States, France, Venezuela	249	0
Lessig and Edelmann (1998)	Leipzig (Germany)	41	1
Pestoni *et al.* (1999)	Galicia (Spain)	35	0
Kayser *et al.* (2000)	Germany, Poland	996	9
Dupuy *et al.* (2001)	Norway	150	3
Tsai *et al.* (2002)	Taiwan	109	1
Dupuy *et al.* (2004)	Norway	1,766	24
Kurihara *et al.* (2004)	Japan	161	2
Ballard *et al.* (2005)	United Kingdom	248	4
Berger *et al.* (2005)	Tyrol (Austria)	70	1
Budowle *et al.* (2005)	North America	692	5
De Souza Góes *et al.* (2005)	Rio de Janeiro (Brazil)	119	4
Gusmão *et al.* (2005)	South America, Portugal, Spain	2,816	23
Turrina *et al.* (2006)	Northeast Italy	50	1
Domingues *et al.* (2007)	Rio de Janeiro (Brazil)	135	1
Hohoff *et al.* (2007)	Westphalia (Germany)	1,029	12
Lee *et al.* (2007)	Korea	369	5
Pontes *et al.* (2007)	North Portugal	45	0
Decker *et al.* (2008)	Ohio	389	7
Padilla-Gutiérrez *et al.* (2008)	Mexico	218	0
Sánchez-Diz *et al.* (2008)	Argentina, Brazil, Portugal	701	9
Soares-Vieira *et al.* (2008)	São Paulo (Brazil)	222	6
Ge *et al.* (2009)	Texas	2,913	17
Goedbloed *et al.* (2009)	Germany, Poland	1,762	27
Sum		15,285	162

Reports on directly observed Y-STR mutations in father–son duos are shown. *N* denotes the number of observed father–son duos included, and *n* is the number of observed mutations across all six loci considered in this study.

repeats. In a father-to-son transmission, an allele may change according to a probability distribution that depends only upon the paternal allele. Mathematically, this means that each model entails a time-homogeneous Markov chain $(X_n)_{n\in\mathbb{Z}}$ with state space $S = \mathbb{Z}$, the set of all integers. Each mutation model is thus defined by specifying, for all $i, j \in \mathbb{Z}$, transition probabilities p_{ij} from paternal allele i to offspring allele j. For values of j not explicitly addressed in the definitions given below, we assume $p_{ij} = 0$. Since the data considered here contained only three mutations that changed the paternal allele by more than 1 repeat unit (Table 3), we confined our analyses to single-step models, characterized by $p_{ij} = 0$ for all $i, j \in \mathbb{Z}$ with $|j - i| > 1$.

Stepwise mutation model: The most popular model of microsatellite mutation is the SMM, originally proposed by Ohta and Kimura (1973), although in a slightly different context. The parameters of the SMM are $0 < \mu_u, \mu_d < 0.5$, and the transition probabilities are

$$p_{ij} = \begin{cases} \mu_u & \text{if} \quad j = i+1, \\ \mu_d & \text{if} \quad j = i-1, \\ 1-\mu_u-\mu_d & \text{if} \quad j = i. \end{cases}$$

Parameters μ_u and μ_d correspond to the upward and downward mutation rates, respectively. The symmetric SMM is a special case in which $\mu_u = \mu_d = \mu$.

Linear model: To be able to properly define the linear model (Kruglyak *et al.* 1998), we must restrict the state space to $S = \{1, 2, \ldots, k\}$, with $k \in \mathbb{N}$ fixed. The linear model has parameters $\nu_u, \nu_d \in [0, 1/(2k)]$, and the transition probabilities are

$$p_{ij} = \begin{cases} i \cdot \nu_u & \text{if} \quad j = i+1, \\ i \cdot \nu_d & \text{if} \quad j = i-1, \\ 1 - i \cdot (\nu_u + \nu_d) & \text{if} \quad j = i, \end{cases}$$

for all $i \in S\backslash\{1, k\}$, $j \in S$. Furthermore, $p_{12} = \nu_u$, $p_{11} = 1 - \nu_u$, $p_{k,k-1} = k \cdot \nu_d$, and $p_{kk} = 1 - k \cdot \nu_d$. The symmetric linear model is a special case in which $\nu_u = \nu_d = \nu$.

Logistic model: Here, we newly introduce the logistic model, defined on state space $S = \mathbb{Z}$. To our knowledge, this mutation model has not been evaluated yet, neither theoretically nor empirically. In its most general form, the logistic model has parameters α_u, $\alpha_d \in \mathbb{R}$, β_u, $\beta_d \in \mathbb{R}^+$, and γ_u, $\gamma_d \in (0,0.5)$ and is characterized by transition probabilities

$$p_{ij} = \begin{cases} \dfrac{\gamma_u}{1+\exp(\alpha_u \cdot (\beta_u - i))} & \text{if} \quad j = i+1, \\ \dfrac{\gamma_d}{1+\exp(\alpha_d \cdot (\beta_d - i))} & \text{if} \quad j = i-1, \\ 1 - \dfrac{\gamma_u}{1+\exp(\alpha_u \cdot (\beta_u - i))} & \\ \quad - \dfrac{\gamma_d}{1+\exp(\alpha_d \cdot (\beta_d - i))} & \text{if} \quad j = i. \end{cases}$$

Thus, α_u, β_u, and γ_u together with the repeat number determine the probability of a gain of 1 repeat unit. Similarly,

Table 3 Observed Y-STR mutations

	DYS19			DYS389I			DYS390			
Allele	−1	0	+1	−1	0	+1	−1	0	+1	+2
8					5					
9					816	1				
10					1123	1				
11		2		2	502					
12		26		2	2126	3				
13		1438		3	6712	7				
14		7017	10	9	2093	2				
15		3847	3	3	45					
16	5	1887	3							
17	7	954	4							
18	2	11								
19		3								
20								21		
21								1055	1	
22								1784		1
23							6	3691	2	
24							3	5438	10	
25							5	2495	3	
26								203	1	
27								13		

	DYS391					DYS392			DYS393		
Allele	−2	−1	0	+1	+2	−1	0	+1	−1	0	+1
5			1								
6			5								
7			3				1				
8			29				5				
9			526				10			4	
10			8446	11			90			12	
11		14	5307	12	1	2	6813			40	
12	1	3	215				870			1468	
13			20				5356	3	2	9490	5
14			1			1	1128	1	2	1699	3
15							164		1	426	
16							100		1	21	
17							4			4	
18							1				

Numbers of Y-STR mutations observed in 24 studies (*cf.* Table 2) of father–son duos are shown. "Allele": repeat number of the paternal allele. "−1" and "+1" indicate loss or gain, respectively, of 1 repeat unit in the son. "−2" and "+2" indicate losses or gains of 2 repeat units in one generation. "0": no mutation.

α_d, β_d, and γ_d determine the probability of a downward mutation. Specifically, the α-values determine the increase in mutation rate with increasing allele length while the γ-values can be interpreted as maximum mutation probabilities (mutation probabilities for very long alleles). More formally, $\lim_{i\to\infty} p_{i,i+1} = \gamma_u$ if $\alpha_u > 0$, and $\lim_{i\to\infty} p_{i,i-1} = \gamma_d$ if $\alpha_d > 0$. Parameter β can be thought of as the repeat number that corresponds to an intermediate mutation probability $\frac{1}{2}\gamma$.

Statistical approach

Similar to Whittaker *et al.* (2003), we employed a maximum-likelihood approach (Pawitan 2001) to estimate the parameters of a given mutation model. Assuming pairwise independence of the father-to-son transmissions, the likelihood L of a fixed parameter vector θ equals the product of the respective transition probabilities, taken over all observed father–son duos. Thus

$$L(\theta) = \prod_{i=1}^{N} p_i(\theta), \tag{1}$$

where N is the number of observed father–son duos and $p_i(\theta)$ is the respective transition probability (with parameter vector θ) for the ith father–son duo. Note that Equation 1 includes all observed father–son duos, not only those in whom a mutation occurred, because transition probabilities include the probability of the Markov chain staying in the current state. This is why paternal allele frequencies are required for alleles not involved in any observed mutation.

Since the likelihoods in Equation 1 are typically very small, it is convenient to use logarithms; *i.e.*,

$$\mathcal{L}(\theta) = \log(L(\theta)) = \sum_{i=1}^{N} \log(p_i(\theta)). \tag{2}$$

For each of the six Y-STRs and for each model considered, we maximized the log-likelihood function $\mathcal{L}$ using numerical optimization (Nocedal and Wright 2006) as implemented in the *R* environment (Ihaka and Gentleman 1996), particularly the nlminb function from the stats package.

To allow comparison of unnested mutation models that have different numbers of parameters, we used Akaike's information criterion (AIC), computed as

$$\mathrm{AIC} = -2 \cdot \max_{\theta} \{\mathcal{L}(\theta)\} + 2k, \tag{3}$$

where k is the number of parameters in the mutation model of interest (Akaike 1973; Burnham and Anderson 2002). The smaller the AIC value is for a given Y-STR, the better the model fit. In the case of nested models, any additional parameter has to improve the model fit by at least 1 log-likelihood unit to reduce the AIC value. As a general rule of thumb, all models with an AIC value less than 2 units above the smallest AIC value warrant further attention (Burnham and Anderson 2002).

Results

Observed mutations

The observed distribution of mutations (Table 3) reveals a consistent pattern of upward bias for shorter alleles and downward bias for longer alleles. The only apparent exception is DYS392, but since only seven mutations were observed at this locus, the observed pattern still seems inconclusive.

All observed mutations were single step except for a −2 mutation at DYS391 and one +2 mutation each at DYS390 and DYS391. These mutations were counted as −1 and +1, respectively, for the purpose of model fitting.

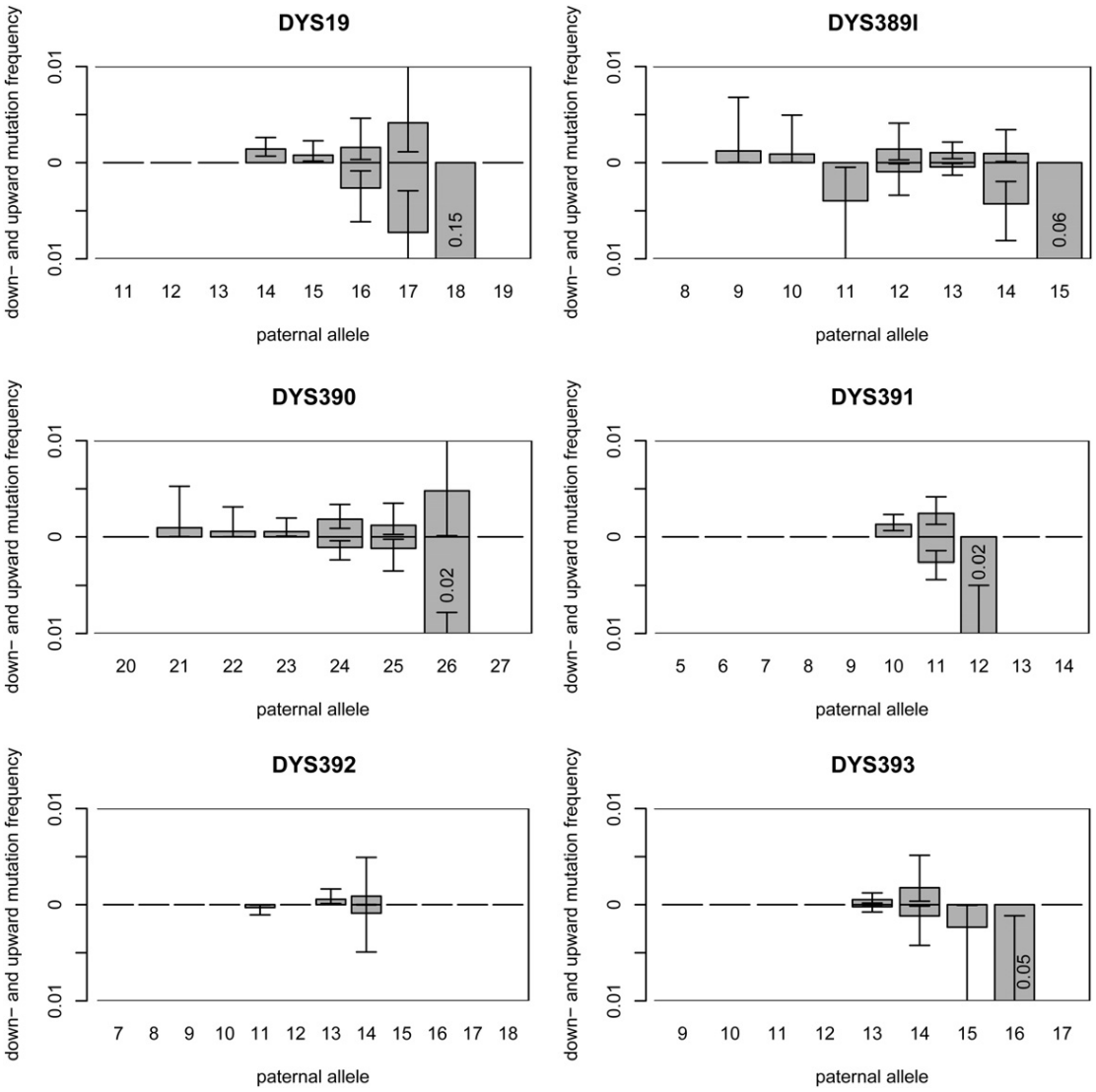

Figure 1 Allele-specific mutation frequencies observed at six Y-STR loci. In each panel, top bars depict upward mutations while bottom bars correspond to downward mutations. Where observed frequencies are larger than zero, 95% confidence intervals are plotted as error bars. Note that the depicted frequency range ends at 0.01, which results in the largest downward mutation bar being clipped for five loci. In these cases, the true height of the bar is given inside the bar.

Model fitting

A straightforward approach to estimate locus- and allele-specific mutation probabilities would be by way of observed relative frequencies of mutations (Figure 1). This simplistic assessment already reveals an asymmetry between upward and downward mutations. In particular, long alleles appear to have high downward mutation probabilities. However, since allele-specific probability estimates are not reliable for alleles with few or no observed mutations, a model-based approach appeared warranted. Moreover, mathematical models potentially yield additional insights into the biological nature of the microsatellite mutation process. We used a maximum-likelihood approach to evaluate several mutation models using the Y-STR data summarized in Table 3. The resulting parameter estimates for some of these models are given in Table 4.

Inspection of the model- and locus-specific maximum log-likelihoods and AIC values (Table 5) reveals a remarkable consistency across the first four loci (DYS19, DYS389I, DYS390, and DYS391). For each of these, the symmetric SMM was found to be only a crude approximation, and the asymmetric SMM was not significantly better. The linear model turned out to be superior to the SMM although allowance for asymmetry again did not seem to improve the model fit substantially. The logistic model provided the best fit to the available data for all four loci, and it was found to be substantially better than both the SMM and the linear model. Within the logistic models, symmetry with respect to

Table 4 Parameter estimates

Model	Parameter	Locus-specific estimates					
		DYS19	DYS389I	DYS390	DYS391	DYS392	DYS393
1	μ	1.14×10^{-3}	1.22×10^{-3}	1.07×10^{-3}	1.43×10^{-3}	2.22×10^{-4}	5.41×10^{-4}
2	μ_u	1.31×10^{-3}	1.04×10^{-3}	1.22×10^{-3}	1.64×10^{-3}	2.75×10^{-4}	6.07×10^{-4}
	μ_d	9.20×10^{-4}	1.41×10^{-3}	9.50×10^{-4}	1.23×10^{-3}	2.06×10^{-4}	4.55×10^{-4}
3	ν	7.65×10^{-5}	9.86×10^{-5}	4.62×10^{-5}	1.39×10^{-5}	1.99×10^{-5}	4.06×10^{-5}
4	ν_u	9.00×10^{-5}	8.37×10^{-5}	5.20×10^{-5}	1.59×10^{-5}	2.27×10^{-5}	4.64×10^{-5}
	ν_d	6.30×10^{-5}	1.14×10^{-4}	4.05×10^{-5}	1.19×10^{-5}	1.70×10^{-5}	3.48×10^{-5}
5	α	−0.886	−0.453	−0.766	−1.65	−0.701	−1.13
	β	22.0	24.7	31.9	11.7	13.4	16.7
	γ	0.500	0.278	0.500	0.0115	7.87×10^{-4}	0.0228
8	α_u	0.374	0.0183	0.323	0.592	58.3	0.611
	α_d	2.08	1.18	1.57	17.7	3.47×10^{-3}	1.57
	β_u	27.4	349	37.0	14.1	13.0	20.8
	β_d	18.3	18.0	27.2	11.1	270	16.7
	γ	0.149	0.500	0.0903	0.0163	7.13×10^{-4}	0.0641

Locus-specific maximum-likelihood estimates of the parameters of selected single-step microsatellite mutation models are shown. Models were fitted to the data of Table 3. For model numbers, see Table 5.

γ seems to be acceptable whereas α and β apparently need to be asymmetric. Thus, the newly introduced logistic model with parameters α_u, α_d, β_u, β_d, and γ turned out to provide the best fit according to the AIC, closely followed by the full logistic model. All other model fits were significantly worse.

For DYS392, the symmetric linear model yielded the smallest AIC, but this should be interpreted with caution because only few mutations were observed at this locus (Table 3). For DYS393, the logistic model with symmetrical α-, β-, and γ-values was the best fit according to AIC, but other versions of the logistic model yielded very similar AIC values.

A plot of three models fitted to the DYS19 data (Figure 2) serves to illustrate further that the novel logistic model is appropriate to explain the respective mutation data. The corresponding plots for the other five loci look similar and are therefore not shown.

Table 5 Maximum-likelihood and AIC values

	Model	Parameters	DYS19 (*N* = 15,219)		DYS389I (*N* = 13,455)		DYS390 (*N* = 14,732)	
			ΔML	ΔAIC	ΔML	ΔAIC	ΔML	ΔAIC
1	SMM	μ	0	0	0	0	0	0
2		μ_u, μ_d	0.5	0.9	0.4	1.2	0.3	1.5
3	Linear	ν	2.5	−5.0	1.4	−2.7	1.1	−2.2
4		ν_u, ν_d	3.0	−4.1	1.7	−1.5	1.4	−0.7
5	Logistic	α, β, γ	18.3	−32.5	3.9	−3.8	9.8	−15.5
6		α_u, α_d, β, γ	26.7	−47.4	4.0	−1.9	12.0	−17.9
7		α_u, β_u, β_d, γ	21.8	−37.6	4.3	−2.5	10.0	−14.0
8		α_u, α_d, β_u, β_d, γ	29.0	−50.1[a]	9.2	−10.3[a]	14.8	−21.5[a]
9		α_u, α_d, β_u, β_d, γ_u, γ_d	29.0	−48.1	9.2	−8.3	14.8	−19.6

	Model	Parameters	DYS391 (*N* = 14595)		DYS392 (*N* = 14549)		DYS393 (*N* = 13178)	
			ΔML	ΔAIC	ΔML	ΔAIC	ΔML	ΔAIC
1	SMM	μ	0	0	0	0	0	0
2		μ_u, μ_d	0.4	1.1	0.1	1.9	0.1	1.7
3.	Linear	ν	1.8	−3.7	0.3	−0.6[a]	0.6	−1.3
4		ν_u, ν_d	2.3	−2.5	0.4	1.2	0.8	0.4
5	Logistic	α, β, γ	13.2	−22.4	1.0	2.1	5.3	−6.6[a]
6		α_u, α_d, β, γ	20.9	−35.8	0.9	4.2	6.1	−6.1
7		α_u, β_u, β_d, γ	19.4	−32.8	1.0	4.1	6.1	−6.1
8		α_u, α_d, β_u, β_d, γ	22.6	−37.3[a]	3.2	1.7	6.7	−5.3
9		α_u, α_d, β_u, β_d, γ_u, γ_d	23.4	−36.9	3.2	3.7	7.7	−5.3

Differences in maximum log-likelihood (ML) and corresponding Akaike information criterion (AIC) values for some single-step microsatellite mutation models are shown. *N*: number of father–son duos with data available for the respective locus. The stepwise mutation model (SMM) with a single parameter was chosen as the reference model so that ΔML is the maximum log-likelihood minus this value for the symmetric SMM. Similarly, ΔAIC is the AIC value minus the value for the symmetric SMM.

[a] For each locus, the smallest AIC value, indicating the best model.

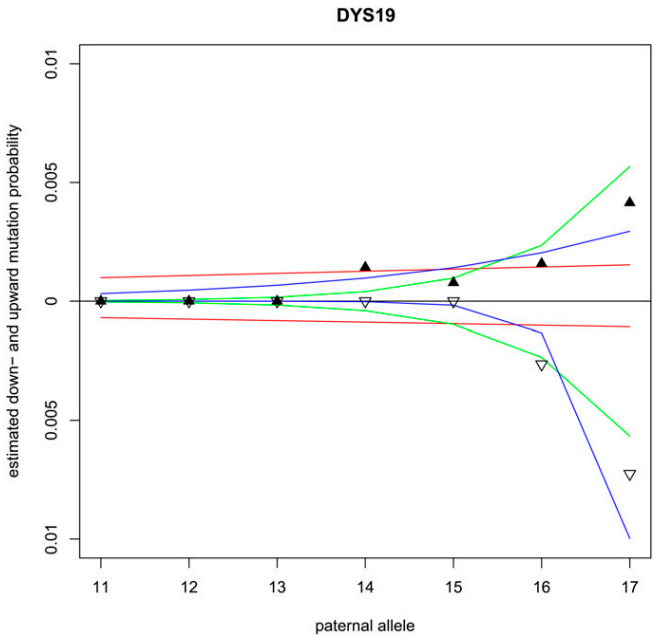

Figure 2 Mutation probabilities of three models fitted to the DYS19 data. The top half of the diagram depicts upward mutation probabilities whereas the bottom half depicts downward mutation probabilities. Triangles mark the observed mutation frequencies (solid, upward; open, downward). Models fitted are the linear model without symmetry (red lines), the fully symmetrical logistic model (green lines), and the logistic model with parameters α_u, α_d, β_u, β_d, and γ (blue lines). Note that the plot is truncated beyond paternal allele 17 because the observed downward mutation frequency for paternal allele 18 is very high, but unreliable, due to the small number of fathers carrying allele 18. Mutation frequencies not shown are $\hat{p}_{18,19} = 0$, $\hat{p}_{19,20} = 0$, $\hat{p}_{18,17} = 0.15$, and $\hat{p}_{19,18} = 0$.

Discussion

It is well known that the pattern of microsatellite mutation varies across loci (Kelkar *et al.* 2008). To our knowledge, however, the present study is the first to systematically compare novel as well as previously described microsatellite mutation models for Y-STRs in a locus-specific fashion. This comparison was made possible by the accumulation of suitable genotype data from 15,285 father–son duos. The major advantage of this type of data is that all father–son relationships had been confirmed by independent genotyping of other markers. In studies using deep pedigree data for mutational analyses, this is typically not the case (Heyer *et al.* 1997), which renders discrimination between false paternity records and genuine mutations notoriously difficult. Furthermore, by basing our analysis on directly observed mutations (*i.e.*, Y-STR mutations in father–son duos), we avoided the need for additional assumptions about the underlying population dynamics, mating behavior, or selective pressure. This is a clear advantage over studies that sought to investigate microsatellite mutation processes by comparing distantly related genomes (Dieringer and Schlötterer 2003).

In our analysis, we also avoided the complicating effects of recombination through choosing loci from the male-specific region of the Y chromosome. Although this restriction may at first glance seem to limit the general applicability of our results, it may be surmised that Y-chromosomal and autosomal microsatellite loci obey similar mutation models because they have similar biochemical properties and because replication slippage is responsible for STR mutations in both instances (Heyer *et al.* 1997; Kayser *et al.* 2000). This contrasts with minisatellite mutations, where recombination plays a significant role (Buard *et al.* 2000).

One caveat of our study is that the loci considered were originally selected for forensic applications because of their high variability. Therefore, we cannot exclude that our parameter estimates are biased toward higher mutation probabilities, but this seems unlikely to affect the general conclusion as to which models are most appropriate for microsatellites.

As was mentioned before, many statistical models have been proposed for the microsatellite mutation process (Calabrese and Sainudiin 2005). However, only a few of these turned out to be applicable to our data. For example, the model proposed by Kruglyak *et al.* (1998), which includes point mutations, was not deemed relevant to our study because, with the genotyping systems used in forensics, point mutations are not altering repeat counts (Gusmão *et al.* 2006). Moreover, we chose to restrict ourselves to one-step models owing to the scarcity of data on multistep mutations (Table 3). The three instances of mutations resulting in a change by 2 repeat units in our data were counted as single-step changes for model-fitting purposes. This concerned two of the six loci considered, for which the models should therefore be interpreted as dichotomizing all possible mutation events into up- and downward mutations, regardless of step size. However, since multistep mutations are very rare, this dichotomization should not affect our conclusions substantially. Nevertheless, should more data on multistep mutations become available in the future, a study of more sophisticated models may become worthwhile.

Our study was in part inspired by Whittaker *et al.* (2003), who were the first to suggest the use of maximum likelihood to fit microsatellite mutation models. However, as explained above, their exponential mutation model was not well defined, resulting in unbounded mutation probabilities. This problem does not occur with a logistic model, which to our knowledge has not been investigated before, because in the logistic model mutation probabilities are always bounded by parameter γ. Notably, for small repeat numbers, Whittaker's model and the logistic model are similar in that both entail an exponential increase in mutation probabilities with increasing repeat number. Qualitatively, the logistic model is also similar to the best models emerging from genome comparisons, *e.g.*, the PL1 model in Sainudiin *et al.* (2004). The main features in both instances are an allele-length–dependent mutation rate and a confinement to single-step mutations.

In principle, it would be possible to combine mutation models to obtain better fits. For example, a model with

linearly increasing upward mutation probabilities but with downward mutation probabilities according to the logistic model fits our data for DYS390 and DYS391 somewhat better than the logistic model alone ($\Delta AIC = -21.9$ and $\Delta AIC = -37.7$, respectively, *cf*. Table 5). However, we decided to focus on pure models here to not exacerbate the multiple-testing problem.

Practical applications of our results are vast because many uses of microsatellite data require estimates of the respective mutation probabilities. The logistic model, which was shown here to provide the best fit to empirical mutation data, is readily applicable to likelihood-based kinship analysis, phylogenetic analysis, and coalescence methods used in population genetics. Our statistical evaluation of mutation models may also contribute to a better understanding of the underlying biological mutation mechanisms. In particular, the fact that the combined models fit the data better than the original ones suggests possible differences between the mechanisms of upward and downward mutation.

With these applications in mind, gathering of further mutation data, for example, in an international STR mutation database, seems to be warranted. With a growing database, it will become possible to further refine parameter estimates as well as the models themselves.

Acknowledgments

We thank all authors whose Y-STR mutation data were included in the present study and Lutz Roewer and Sascha Willuweit for providing access to summary statistics for the Y Chromosome Haplotype Reference Database (Willuweit and Roewer 2007). We thank Raazesh Sainudiin and an anonymous reviewer for helpful suggestions.

Literature Cited

Akaike, H., 1973 Information theory and an extension of the maximum likelihood principle, pp. 267–281 in *Second International Symposium on Information Theory*, edited by B. N. Petrov. Akademiai Kiado, Budapest.

Ballard, D., C. Phillips, G. Wright, C. Thacker, C. Robson *et al.*, 2005 A study of mutation rates and the characterisation of intermediate, null and duplicated alleles for 13 Y chromosome STRs. Forensic Sci. Int. 155: 65–70.

Berger, B., A. Lindinger, H. Niederstätter, P. Grubwieser, and W. Parson, 2005 Y-STR typing of an Austrian population sample using a 17-loci multiplex PCR assay. Int. J. Legal Med. 119: 241–246.

Bianchi, N., C. Catanesi, G. Bailliet, V. Martinez-Marignac, C. Bravi *et al.*, 1998 Characterization of ancestral and derived Y-chromosome haplotypes of New World native populations. Am. J. Hum. Genet. 63: 1862–1871.

Buard, J., A. Shone, and A. Jeffreys, 2000 Meiotic recombination and flanking marker exchange at the highly unstable human minisatellite CEB1 (D2S90). Am. J. Hum. Genet. 67: 333–344.

Budowle, B., M. Adamowicz, X. Aranda, C. Barna, R. Chakraborty *et al.*, 2005 Twelve short tandem repeat loci Y chromosome haplotypes: genetic analysis on populations residing in North America. Forensic Sci. Int. 150: 1–15.

Burnham, K. P., and D. R. Anderson, 2002 *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. Springer-Verlag, New York.

Calabrese, P., and R. Sainudiin, 2005 Models of Microsatellite Evolution. In: *Statistical Methods in Molecular Evolution*, edited by Nielsen, R., pp. 290–305. Springer-Verlag, New York.

Calabrese, P. P., R. T. Durrett, and C. F. Aquadro, 2001 Dynamics of microsatellite divergence under stepwise mutation and proportional slippage/point mutation models. Genetics 159: 839–852.

de Souza Góes, A., E. de Carvalho, I. Gomes, D. da Silva, E. Fonseca Gil *et al.*, 2005 Population and mutation analysis of 17 Y-STR loci from Rio de Janeiro (Brazil). Int. J. Legal Med. 119: 70–76.

Decker, A., M. Kline, J. Redman, T. Reid, and J. Butler, 2008 Analysis of mutations in father-son pairs with 17 Y-STR loci. Forensic Sci. Int. Genet. 2: e31–e35.

Dieringer, D., and C. Schlötterer, 2003 Two distinct modes of microsatellite mutation processes: evidence from the complete genomic sequences of nine species. Genome Res. 13: 2242–2251.

Di Rienzo, A., A. C. Peterson, J. C. Garza, A. M. Valdes, M. Slatkin *et al.*, 1994 Mutational processes of simple-sequence repeat loci in human populations. Proc. Natl. Acad. Sci. USA 91: 3166–3170.

Domingues, P., L. Gusmão, D. da Silva, A. Amorim, R. Pereira *et al.*, 2007 Sub-Saharan Africa descendents in Rio de Janeiro (Brazil): population and mutational data for 12 Y-STR loci. Int. J. Legal Med. 121: 238–241.

Dupuy, B., R. Andreassen, A. Fløthings, K. Tomassen, T. Egeland *et al.*, 2001 Y-chromosome variation in a Norwegian population sample. Forensic Sci. Int. 117: 163–173.

Dupuy, B., M. Stenersen, T. Egeland, and B. Olaisen, 2004 Y-chromosomal microsatellite mutation rates: differences in mutation rate between and within loci. Hum. Mutat. 23: 117–124.

Durrett, R., and S. Kruglyak, 1999 A new stochastic model of microsatellite evolution. J. Appl. Probab. 36: 621–631.

Falush, D., and Y. Iwasa, 1999 Size-dependent mutability and microsatellite constraints. Mol. Biol. Evol. 16: 960–966.

Feldman, M. W., A. Bergman, D. D. Pollock, and D. B. Goldstein, 1997 Microsatellite genetic distances with range constraints: analytic description and problems of estimation. Genetics 145: 207–216.

Garza, J., M. Slatkin, and N. Freimer, 1995 Microsatellite allele frequencies in humans and chimpanzees, with implications for constraints on allele size. Mol. Biol. Evol. 12: 594–603.

Ge, J., B. Budowle, X. Aranda, J. Planz, A. Eisenberg *et al.*, 2009 Mutation rates at Y chromosome short tandem repeats in Texas populations. Forensic Sci. Int. Genet. 3: 179–184.

Goedbloed, M., M. Vermeulen, R. Fang, M. Lembring, A. Wollstein *et al.*, 2009 Comprehensive mutation analysis of 17 Y-chromosomal short tandem repeat polymorphisms included in the AmpFlSTR Yfiler PCR amplification kit. Int. J. Legal Med. 123: 471–482.

Gusmão, L., P. Sánchez-Diz, F. Calafell, P. Martín, C. Alonso *et al.*, 2005 Mutation rates at Y chromosome specific microsatellites. Hum. Mutat. 26: 520–528.

Gusmão, L., J. Butler, A. Carracedo, P. Gill, M. Kayser *et al.*, 2006 DNA Commission of the International Society of Forensic Genetics (ISFG): an update of the recommendations on the use of Y-STRs in forensic analysis. Forensic Sci. Int. 157: 187–197.

Heyer, E., J. Puymirat, P. Dieltjes, E. Bakker, and P. de Knijff, 1997 Estimating Y chromosome specific microsatellite mutation frequencies using deep rooting pedigrees. Hum. Mol. Genet. 6: 799–803.

Hohoff, C., K. Dewa, U. Sibbing, K. Hoppe, P. Forster *et al.*, 2007 Y-chromosomal microsatellite mutation rates in a popu-

lation sample from Northwestern Germany. Int. J. Legal Med. 121: 359–363.
Ihaka, R., and R. Gentleman, 1996 R: a language for data analysis and graphics. J. Comput. Graph. Stat. 5: 299–314.
Jobling, M. A., and C. Tyler-Smith, 2003 The human Y chromosome: an evolutionary marker comes of age. Nat. Rev. Genet. 4: 598–612.
Kashi, Y., and D. G. King, 2006 Simple sequence repeats as advantageous mutators in evolution. Trends Genet. 22: 253–259.
Kayser, M., L. Roewer, M. Hedman, L. Henke, J. Henke *et al.*, 2000 Characteristics and frequency of germline mutations at microsatellite loci from the human Y chromosome, as revealed by direct observation in father/son pairs. Am. J. Hum. Genet. 66: 1580–1588.
Kelkar, Y. D., S. Tyekucheva, F. Chiaromonte, and K. D. Makova, 2008 The genome-wide determinants of human and chimpanzee microsatellite evolution. Genome Res. 18: 30–38.
Kimmel, M., R. Chakraborty, D. N. Stivers, and R. Deka, 1996 Dynamics of repeat polymorphisms under a forward-backward mutation model: within- and between-population variability at microsatellite loci. Genetics 143: 549–555.
Krawczak, M., and J. Schmidtke, 1998 *DNA Fingerprinting*. Springer-Verlag, New York.
Kruglyak, S., R. T. Durrett, M. D. Schug, and C. F. Aquadro, 1998 Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations. Proc. Natl. Acad. Sci. USA 95: 10774–10778.
Kurihara, R., T. Yamamoto, R. Uchihi, S. Li, T. Yoshimoto *et al.*, 2004 Mutations in 14 Y-STR loci among Japanese father-son haplotypes. Int. J. Legal Med. 118: 125–131.
Lai, Y., and F. Sun, 2003 The relationship between microsatellite slippage mutation rate and the number of repeat units. Mol. Biol. Evol. 20: 2123–2131.
Lander, E. S., L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody *et al.*, 2001 Initial sequencing and analysis of the human genome. Nature 409: 860–921.
Lee, H., M. Park, U. Chung, H. Lee, W. Yang *et al.*, 2007 Haplotypes and mutation analysis of 22 Y-chromosomal STRs in Korean father-son pairs. Int. J. Legal Med. 121: 128–135.
Lessig, R., and J. Edelmann, 1998 Y chromosome polymorphisms and haplotypes in West Saxony (Germany). Int. J. Legal Med. 111: 215–218.
Mayntz-Press, K. A., and J. Ballantyne, 2007 Performance characteristics of commercial Y-STR multiplex systems. J. Forensic Sci. 52: 1025–1034.
Nauta, M. J., and F. J. Weissing, 1996 Constraints on allele size at microsatellite loci: implications for genetic differentiation. Genetics 143: 1021–1032.
Nocedal, J., and S. J. Wright, 2006 *Numerical Optimization*. Springer-Verlag, New York.
Ohta, T., and M. Kimura, 1973 A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. Genetics 22: 201–204.
Padilla-Gutiérrez, J. R., Y. Valle, A. Quintero-Ramos, G. Hernández, K. Rodarte *et al.*, 2008 Population data and mutation rate of nine Y-STRs in a mestizo Mexican population from Guadalajara, Jalisco, México. Legal Med. 10: 319–320.
Pawitan, Y., 2001 *In All Likelihood: Statistical Modelling and Inference Using Likelihood*. Oxford University Press, New York.
Pestoni, C., M. Cal, M. Lareu, M. Rodríguez-Calvo, and A. Carracedo, 1999 Y chromosome STR haplotypes: genetic and sequencing data of the Galician population (NW Spain). Int. J. Legal Med. 112: 15–21.
Pontes, M., L. Cainé, D. Abrantes, G. Lima, and M. Pinheiro, 2007 Allele frequencies and population data for 17 Y-STR loci (AmpFlSTR Y-filer) in a Northern Portuguese population sample. Forensic Sci. Int. 170: 62–67.
Roewer, L., 2009 Y chromosome STR typing in crime casework. Forensic Sci. Med. Pathol. 5: 77–84.
Sainudiin, R., R. T. Durrett, C. F. Aquadro, and R. Nielsen, 2004 Microsatellite mutation models: insights from a comparison of humans and chimpanzees. Genetics 168: 383–395.
Shriver, M. D., and R. A. Kittles, 2004 Genetic ancestry and the search for personalized genetic histories. Nat. Rev. Genet. 5: 611–618.
Shriver, M. D., L. Jin, R. Chakraborty, and E. Boerwinkle, 1993 VNTR allele frequency distributions under the stepwise mutation model: a computer simulation approach. Genetics 134: 983–993.
Sánchez-Diz, P., C. Alves, E. Carvalho, M. Carvalho, R. Espinheira *et al.*, 2008 Population and segregation data on 17 Y-STRs: results of a GEP-ISFG collaborative study. Int. J. Legal Med. 122: 529–533.
Soares-Vieira, J. A., A. E. Billerbeck, E. S. Iwamura, B. B. Mendonca, L. Gusmão *et al.*, 2008 Population and mutation analysis of Y-STR loci in a sample from the city of São Paulo (Brazil). Genet. Mol. Biol. 31: 651–656.
Tsai, L., T. Yuen, H. Hsieh, M. Lin, C. Tzeng *et al.*, 2002 Haplotype frequencies of nine Y-chromosome STR loci in the Taiwanese Han population. Int. J. Legal Med. 116: 179–183.
Turrina, S., R. Atzei, and D. De Leo, 2006 Y-chromosomal STR haplotypes in a Northeast Italian population sample using 17plex loci PCR assay. Int. J. Legal Med. 120: 56–59.
Valdes, A. M., M. Slatkin, and N. B. Freimer, 1993 Allele frequencies at microsatellite loci: the stepwise mutation model revisited. Genetics 133: 737–749.
Weir, B. S., A. D. Anderson, and A. B. Hepler, 2006 Genetic relatedness analysis: modern data and new challenges. Nat. Rev. Genet. 7: 771–780.
Weissenbach, J., 1993 A second generation linkage map of the human genome based on highly informative microsatellite loci. Gene 135: 275–278.
Whittaker, J. C., R. M. Harbord, N. Boxall, I. Mackay, G. Dawson *et al.*, 2003 Likelihood-based estimation of microsatellite mutation rates. Genetics 164: 781–787.
Willuweit, S., and L. Roewer, 2007 Y chromosome haplotype reference database (YHRD): update. Forensic Sci. Int. Genet. 1: 83–87.
Xu, X., M. Peng, Z. Fang, and X. Xu, 2000 The direction of microsatellite mutations is dependent upon allele length. Nat. Genet. 24: 396–399.

Communicating editor: M. A. Beaumont

2.3 Estimating Trace-Suspect Match Probabilities for Singleton Y-STR Haplotypes Using Coalescent Theory

Andersen, Caliebe, Jochens, Willuweit und Krawczak (2013)
Forensic Science International: Genetics 7:2, 264–271.

Zusammenfassung

Mit dieser Publikation wenden wir uns einer praktischen Anwendung von STR-Mutationsmodellen zu. Wie in Abschnitt 1.6.2 erläutert wurde, spielt das Schätzen der *Random Match Probability* (RMP) eine zentrale Rolle für die Quantifizierung der Evidenz in forensischen DNA-Analysen. Wenn es sich bei den dabei verwendeten Loci um mehrere *Lineage Marker* handelt, was beim Menschen gleichbedeutend mit Y-chromosomalen oder mitochondriellen Loci ist, dann ist der Konsens, dass die stochastische Abhängigkeit der Loci voneinander, die aus der Abwesenheit von Rekombination resultiert, bei der RMP-Schätzung berücksichtigt werden muss (siehe Abschnitt 1.6.6). Insbesondere heißt das, dass die geschätzten Allel-Häufigkeiten für die verschiedenen Loci nicht einfach miteinander multipliziert werden dürfen. Die einfachste Vorgehensweise besteht darin, die Populations-Häufigkeit des gesamten beobachteten Haplotypen durch seinen Anteil an allen Haplotypen in einer geeigneten Referenzdatenbank zu schätzen, wobei die Datenbank in der Regel zuvor noch um den beobachteten Haplotypen erweitert wird. Dies wird als *Counting*-Methode bezeichnet.

Da die in der Forensik gebräuchlichen Y-STR-Kits oft schon 17 sehr variable Loci umfassen, kommt es inzwischen häufig vor, dass der beobachtete Haplotyp zuvor in der Referenzdatenbank noch nicht enthalten war. Man spricht dann von einem *Singleton*-Haplotypen, dessen RMP nach der Counting-Methode auf $1/(n+1)$ geschätzt wird, wenn n die ursprüngliche Größe der Datenbank ist. Dieser Schätzer ist jedoch unbefriedigend, weil die inverse Datenbankgröße somit die RMP für Singleton-Haplotypen nach unten beschränkt. Deshalb wurde von C. Brenner eine Verfeinerung vorgeschlagen, die als κ-Korrektur bekannt ist. Darüber hinaus existiert ein umstrittener Ansatz, der als *Haplotype Surveying* bezeichnet wird. Diese Methode basiert darauf, die Häufigkeit des fraglichen Haplotypen mittels solcher Haplotypen aus der Referenzdatenbank zu schätzen, die möglicherweise evolutionär eng mit dem beobachteten Haplotypen verwandt sind. Ein weniger bekannter Ansatz bedient sich der Koaleszenztheorie (siehe Abschnitt 1.3.2) und wurde von I. Wilson und Kollegen in der BATWING-Software implementiert [244, 245].

Hier vergleichen wir die Leistung dieser vier RMP-Schätzer für Singleton-Y-STR-Haplotypen. Zu diesem Zweck modifizieren wir das Wright-Fisher-Modell um ein exponentielles Populationswachstum und simulieren mittels BATWING

die zugrundeliegende Populationsgeschichte. Außerdem gehen wir wieder vom einfachen schrittweisen Mutationsmodell (SMM) aus. So simulieren wir die Evolution der gesamten Y-STR-Haplotypen, sowie die Ziehung der Referenz-Datenbank.

Wir zeigen, dass der koaleszenzbasierte Ansatz sich durch geringeren Bias und geringeren mittleren quadratischen Fehler (*Mean Squared Error*, MSE) als die unkorrigierte Counting-Methode und der Surveying-Schätzer auszeichnet. Außerdem weisen sowohl der Surveying- als auch der koaleszenzbasierte Schätzer im Gegensatz zu den beiden Counting-Schätzern (ohne bzw. mit κ-Korrektur) eine gute Korrelation zwischen geschätzten und tatsächlichen RMPs auf. Als Beispiel einer realen Referenz-Datenbank betrachten wir 1757 deutsche 15- bzw. 7-Locus-Haplotypen aus der YHRD (siehe Abschnitt 1.6.4).

Insgesamt zeigt der koaleszenzbasierte Schätzer eine Leistung, die besser als die jeder anderen verfügbaren Methode ist. Allerdings sind die erforderlichen Koaleszenzsimulationen sehr rechenaufwändig, an der Grenze zur nicht allgemein gegebenen Anwendbarkeit. Die Anwendung in der forensischen Praxis wird sich somit zunächst auf kleine Referenzdatenbanken beschränken müssen — oder auf vereinzelte Fälle von besonderem Interesse — bis leistungsstärkere Algorithmen für Koaleszenzsimulationen oder schnellere Hardware zur Verfügung stehen.

Forensic Science International: Genetics 7 (2013) 264–271

Contents lists available at SciVerse ScienceDirect

Forensic Science International: Genetics

journal homepage: www.elsevier.com/locate/fsig

Estimating trace-suspect match probabilities for singleton Y-STR haplotypes using coalescent theory

Mikkel Meyer Andersen [a,*], Amke Caliebe [b,1], Arne Jochens [b,2], Sascha Willuweit [c,3], Michael Krawczak [b,4]

[a] Department of Mathematical Sciences, Aalborg University, Fredrik Bajers Vej 7G, 9220 Aalborg East, Denmark
[b] Institute of Medical Informatics and Statistics, Christian-Albrechts University, UK SH Campus Kiel, Arnold-Heller-Strasse 3, 24105 Kiel, Germany
[c] Institute of Legal Medicine, Charité – Universitätsmedizin Berlin, Augustenburger Platz 1, 13353 Berlin, Germany

ARTICLE INFO

Article history:
Received 20 June 2012
Received in revised form 7 November 2012
Accepted 24 November 2012

Keywords:
Y-STR
Singleton haplotype
Lineage marker
Match probability
Likelihood ratio
Coalescent theory

ABSTRACT

Estimation of match probabilities for singleton haplotypes of lineage markers, i.e. for haplotypes observed only once in a reference database augmented by a suspect profile, is an important problem in forensic genetics. We compared the performance of four estimators of singleton match probabilities for Y-STRs, namely the count estimate, both with and without Brenner's so-called 'kappa correction', the surveying estimate, and a previously proposed, but rarely used, coalescent-based approach implemented in the BATWING software. Extensive simulation with BATWING of the underlying population history, haplotype evolution and subsequent database sampling revealed that the coalescent-based approach is characterized by lower bias and lower mean squared error than the uncorrected count estimator and the surveying estimator. Moreover, in contrast to the two count estimators, both the surveying and the coalescent-based approach exhibited a good correlation between the estimated and true match probabilities. However, although its overall performance is thus better than that of any other recognized method, the coalescent-based estimator is still computation-intense on the verge of general impracticability. Its application in forensic practice therefore will have to be limited to small reference databases, or to isolated cases of particular interest, until more powerful algorithms for coalescent simulation have become available.

1. Introduction

In forensic genetics, it is often necessary to compare the plausibility of two case-relevant hypotheses on the basis of some genetic data, and the most consistent (and therefore generally recommended) way of doing so is by means of the likelihood ratio [1]. Calculating the likelihood ratio in forensic case work is usually tantamount to quantifying the match probability between two genetic profiles under different assumptions about their degree of relatedness. One particularly important match probability in this context is the probability that a certain individual (e.g. the donor of a trace found at a crime scene) has the same DNA profile as another individual (usually a suspect) drawn randomly from the same population. Methods to estimate this so-called 'trace-suspect' match probability are well established for autosomal STRs [2], with most of them assuming statistical independence between the markers included in the profile.

Lineage markers, such as Y-chromosomal short tandem repeats (Y-STRs) or mtDNA polymorphisms, have several advantages over autosomal markers [3,4], for example, when solving cases of sexual assault [5]. However, due to the lack of recombination and, therefore, lack of statistical independence, the calculation of match probabilities is more challenging for lineage than for autosomal markers [6]. In particular, when considering Y-STR haplotypes comprising up to 17 loci [7], the proportion of cases involving singletons, defined as haplotypes observed only once in a reference database augmented by the suspect profile, may become so large that use of traditional count estimates of the corresponding match probabilities becomes unsatisfactory.

To detail the inference problem arising with singleton haplotypes, let us assume that a reference database of size n is given, and that a trace and suspect carry a new haplotype not yet observed in the database. Initially, the count estimator $1/(n+1)$ was used to derive match probabilities in such cases. However, this estimator is rather conservative because it is limited from below by

* Corresponding author. Tel.: +45 99408866.
E-mail addresses: mikl@math.aau.dk (M.M. Andersen), caliebe@medinfo.uni-kiel.de (A. Caliebe), jochens@medinfo.uni-kiel.de (A. Jochens), sascha.willuweit@charite.de (S. Willuweit), krawczak@medinfo.uni-kiel.de (M. Krawczak).
[1] Tel.: +49 431 597 3199.
[2] Tel.: +49 431 597 3192.
[3] Tel.: +49 30 450 525074.
[4] Tel.: +49 431 597 3200.

1872-4973/$ – see front matter © 2012 Elsevier Ireland Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.fsigen.2012.11.004

the inverse of the database size. Therefore, a more advanced method referred to as 'haplotype surveying' was proposed [8,9] that tried to exploit the information about evolutionary relatedness inherent in a given database of Y-STR haplotypes. In view of the criticisms raised against it [10,11], the surveying method was later refined [12] and a new version is now implemented, for example, at the YHRD website [13,7] (see http://www.yhrd.org). Recently, Charles Brenner suggested an alternative, comparatively simple method of estimating the match probability for singletons for any kind of markers [11], the so-called 'κ correction' of the count estimator inspired by Robbins [14]. In short, the κ correction entails estimating a match probability by $(1-\kappa)/(n+1)$, where $\kappa = \alpha/(n+1)$ and α denotes the total number of singletons in the database.

Interestingly, there is yet another estimator of forensic match probabilities that unfortunately never got much attention, most probably due to its computational demands. The approach was first described by Ian Wilson and colleagues in 2003 [15] and involves the refinement of a previously published Markov Chain Monte Carlo method to sample coalescent trees [16–18]. In the present paper, we will briefly recall the original work [15] before comparing it to the other three estimators mentioned above. Using both simulated and real data, we will highlight the power and limits of coalescent-based estimation of match probabilities for singleton Y-STR haplotypes.

2. Coalescent-based estimation of match probabilities

The main idea of the coalescent-based approach is as follows [17,18]: Adopting a sensible population history and an appropriate mutation model, a large number of coalescent trees is simulated linking the haplotypes in the reference database $H = (h_1, h_2, \ldots, h_n)$ to one another and to the suspect haplotype h_S. Then, the unknown trace donor X is linked randomly to each tree assuming the same population history as in the simulation of the tree. After the tree-specific probabilities have been calculated that the trace donor possesses the same DNA profile as the suspect, the average of these probabilities, taken over all simulated trees, serves as an estimate of the sought-after match probability.

In 1998, Ian Wilson and David Balding [16] introduced a Bayesian Markov Chain Monte Carlo model to generate random coalescent trees according to their probability of occurrence. This model was expanded in 2003 to include population growth, among other generalizations [15]. To our knowledge, the 2003 paper was also the first one to put the calculation of forensic match probabilities into a coalescent theory context: "*In addition to the genealogical tree underlying the $n+1$ observed [haplotypes], we introduce a branch connecting the unobserved [haplotype] of [a random individual] X with the tree, writing Z for the new node thus introduced*". In our terminology, individual Z is the most recent common ancestor of trace donor X and the most closely related individual(s) in the database, including the suspect. In the Bayesian approach taken [16,15], the haplotypes are assumed to be known at all internal nodes of the tree, including h_Z. This implies that the match probability for a given tree equals the probability that h_Z mutates to the suspect haplotype h_S during the time span separating Z and X (Fig. 1).

The approach proposed by Wilson and colleagues [15] is implemented in the computer program 'Bayesian Analysis of Trees With Internal Node Generation' (BATWING), which is publicly available at http://www.mas.ncl.ac.uk/~nijw/. However, the BATWING program does not explicitly support the calculation of forensic match probabilities but had to be adapted to this task for the present study. The modified BATWING program with the forensic match probability module included can be downloaded from the 'Software' page at http://people.math.aau.dk/~mikl/?p=software.

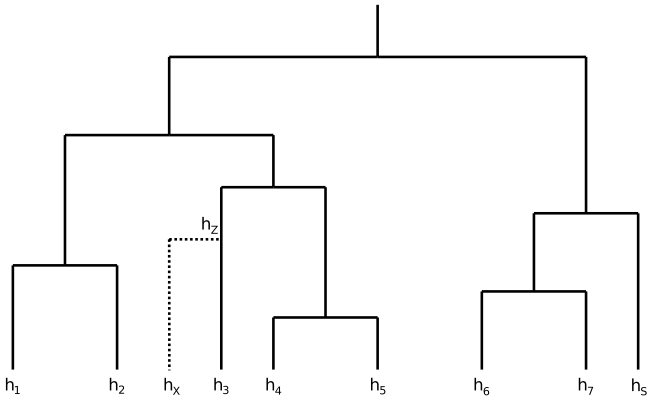

Fig. 1. Calculation of forensic match probabilities using coalescent theory (after [15]). $h_1, h_2, \ldots, h_7$: haplotypes in a reference database H of size $n = 7$; h_S: suspect haplotype; h_X: haplotype of trace donor X; h_Z: haplotype of the most recent common ancestor Z of trace donor X and the most closely related individual(s) in the database, including suspect S. The contribution to the match probability of this particular tree would be the probability that h_Z mutates to h_S during the time span indicated by the dotted line, thereby creating a match between the suspect and trace haplotype.

2.1. Branch-wise contribution to the tree probability

Calculating match probabilities with BATWING is based upon use of the probabilities that a given haplotype mutates to another given haplotype within a specified period of time. In principle, any realistic mutation model can be employed to quantify these probabilities but, in the case of Y-STRs, it appears reasonable to draw upon a single-step mutation model. Under the single-step mutation model used here, the marker-specific numbers of upward and downward mutations (by one repeat unit) in a given number of generations, M_u and M_d, follow independent Poisson distributions with parameters λ_u and λ_d. For the consequent allelic change, only the net effect of the two opposite mutation processes is important, and this difference, $\Delta = M_u - M_d$, follows a Skellam distribution [19] with probability function

$$f(\delta;\lambda_u,\lambda_d) = e^{-(\lambda_u+\lambda_d)}\left(\frac{\lambda_u}{\lambda_d}\right)^{\delta/2} I_{|\delta|}(2\sqrt{\lambda_u\lambda_d}). \quad (1)$$

Here, $I_{|\delta|}$ is the modified $|\delta|$th order Bessel function of the first kind. For the sake of simplicity, we will henceforth assume that upward and downward mutations occur at the same rate. In this case, $\lambda = \lambda_u = \lambda_d$ and the Skellam probability function simplifies to

$$f(\delta;\lambda) = e^{-2\lambda} I_{|\delta|}(2\lambda). \quad (2)$$

Now, let N be the effective population size appropriate for a given forensic context, and let $\theta = 2N\mu$ where μ denotes the total mutation rate per generation per marker. Then, the expected number of (upward plus downward) mutations occurring on a tree branch of length t equals $t\theta/2 = tN\mu$. Assuming equal rates for upward and downward mutation, the mutation process can be thought of as creating two independent random variables, each with a Poisson distribution with parameter $(t\theta/2)/2 = t\theta/4$. In summary, the net allelic change $\Delta_t = M_{u,t} - M_{d,t}$ along a tree branch of length t generations thus follows a Skellam distribution with probability function

$$f(\delta; t, \theta) = e^{-t\theta/2} I_{|\delta|}\left(\frac{t\theta}{2}\right). \quad (3)$$

2.2. Estimation of the match probability

For a given tree, let t denote the time (in generations) between (i) trace donor X and (ii) the most recent common ancestor Z of X and the most closely related individual(s) in the database, including the suspect (Fig. 1). As was noted above, the conditional match probability $P(h_X = h_S \mid H,\ h_S,\ h_Z,\ t)$ equals the probability that h_Z mutates into h_S when passed down from Z to X. Since all trees are simulated (approximately) independently according to their conditional probability of occurrence, given reference database H and suspect haplotype h_S, the sought-after match probability $P(h_X = h_S \mid H,\ h_S)$ can be estimated by

$$\hat{p}_{H,h_S,m} = m^{-1}\sum_{i=1}^{m} P(h_X = h_S | H, h_S, h_Z(i), t(i)), \quad (4)$$

where m equals the number of simulated trees, and where $h_Z(i)$ and $t(i)$ refer to the ith tree.

Under the single-step mutation model used here, the conditional probability $P(h_X = h_S \mid H,\ h_S,\ h_Z,\ t)$ can be quantified using the Skellam probability function given in Eq. (3). Let $\delta(j) = h_S(j) - h_Z(j)$ be the allelic change required at the jth out of r markers. Then

$$P(h_X(j) = h_S(j) | H, h_S, h_Z, t) = f(\delta(j); t, \theta) \quad (5)$$

and, because of independence between mutations,

$$P(h_X = h_S | H, h_S, h_Z, t) = \prod_{j=1}^{r} f(\delta(j); t, \theta). \quad (6)$$

It is worthy of note that coalescent trees are simulated (approximately) independently and according to the same distribution. Therefore, the average of the resulting conditional probabilities $P(h_X = h_S \mid H,\ h_S,\ h_Z(i),\ t(i))$, taken over all m simulations, automatically constitutes a maximum likelihood estimate of the sought-after match probability $P(h_X = h_S \mid H,\ h_S)$ under the employed coalescent and mutation model.

2.3. Convergence issues

The simulation of coalescent trees as described above entails (at least) two different types of convergence of the ensuing match probability estimates:

(i) For a given reference database H and a given suspect haplotype h_S, estimates $\hat{p}_{H,h_S,m}$ from Eq. (4) converge to $P(h_X = h_S \mid H,\ h_S)$ when the number of simulations m increases.
(ii) $P(h_X = h_S \mid H,\ h_S)$ converges to the true match probability $P(h_X = h_S)$ when the reference database H expands towards the whole population.

This means that, in a given case and with a given reference database, increasing the number of simulations ensures that the coalescent-based estimate of the match probability converges to $P(h_X = h_S \mid H,\ h_S)$. The latter is an estimate of $P(h_X = h_S)$ and has sampling variance that can only be reduced by increasing the size of the reference database. However, the larger the database, the more simulations would be required for $\hat{p}_{H,h_S,m}$ to approximate $P(h_X = h_S \mid H,\ h_S)$ sufficiently well, owing to the larger space of coalescent trees to sample from.

3. Methods

The performance of the coalescent-based estimator of singleton match probabilities was compared to that of three other methods, namely (i) the count estimator $1/(n+1)$, where n denotes the database size, (ii) the surveying method in its most recent form [12], and (iii) Brenner's κ correction of the count estimator [11,14].

Each estimator was evaluated on singleton haplotypes from both simulated and real Y-STR data. Simulated data allow a comparison to be made between estimated and true match probabilities by first simulating a big population from which realistically sized databases are then drawn for estimation. As performance measures, we employed the bias and mean squared error (MSE) of each estimator as well as the correlation between the estimated and the truly underlying match probabilities.

Let $\hat{p}_{H_j,h_{S_j}}$ be any estimate of the match probability (coalescent-based, count or surveying) assuming that the jth singleton h_{S_j}, out of ν singletons considered, belongs to the suspect. Thus, H_j is the database with the jth singleton excluded. Let $p_{h_{S_j}}$ be the population frequency of h_{S_j} which, for the sake of simplicity, was taken to coincide with the match probability in our study (i.e. the underlying population was assumed to be panmictic). Then the bias of the estimator was estimated by

$$\frac{1}{\nu}\sum_{j=1}^{\nu}(\hat{p}_{H_j,h_{S_j}} - p_{h_{S_j}}). \quad (7)$$

Similarly, the mean squared error was estimated by

$$\frac{1}{\nu}\sum_{j=1}^{\nu}(\hat{p}_{H_j,h_{S_j}} - p_{h_{S_j}})^2. \quad (8)$$

Finally, we also calculated the Spearman rank correlation coefficient between $\hat{p}_{H_j,h_{S_j}}$ and $p_{h_{S_j}}$. All analyses were carried out with R [20].

3.1. Generation and analysis of simulated data

BATWING [16,15] was not only used for the estimation of match probabilities but also for simulating a large population from which small databases of size $n = 100$ and $n = 200$ were repeatedly sampled for the evaluation of the different estimators. In principle, BATWING supports three different types of population dynamics, namely a constant population size and two exponential growth models (one with constant growth and one with growth after some point in time [15]). Here, we simulated a single source population of 50 million haplotypes that resulted from the constant exponential expansion, over 2000 generations, of an initial population of 20,000 haplotypes. The two-sided (single-step) mutation rate μ was set equal to 0.003 per generation per marker. The number of markers was set equal to 7 as a compromise between computational feasibility and the possibility to obtain realistic data. As can be inferred from Figure A.1 in the Supplementary material, the computation time required for coalescent-based match probability estimation for a fixed number of simulations increased dramatically with both the marker number and the database size.

For each database size (i.e. $n = 100$ or $n = 200$), five databases were drawn randomly from the simulated source population. Next, the forensic match probability was estimated for each singleton haplotype in the database (for the respective proportions of singletons, see Table 1) assuming that the haplotype came from a

Table 1
Number and percentage (in brackets) of singletons observed in ten databases of different size n, sampled from a large simulated source population. Sample numbers are consistent across Tables 1–3.

Sample	$n = 100$	$n = 200$
1	84 (84.0%)	148 (74.0%)
2	85 (85.0%)	135 (67.5%)
3	82 (82.0%)	133 (66.5%)
4	82 (82.0%)	131 (65.5%)
5	92 (92.0%)	152 (76.0%)

suspect and was not included in the reference database itself. Estimation was based upon either 500,000 ($n = 100$) or 200,000 ($n = 200$) simulated coalescent trees per singleton. The larger the database, the larger is the space of coalescent trees to sample from. This means that, in principle, more simulations should be performed for larger databases. Due to computational constraints, however, a substantial increase of the simulation number was not feasible in our study. We therefore conducted a partial in-depth analysis for the five databases of size $n = 200$ by randomly selecting 10 singletons from each database and simulating one million trees for each of these.

In the coalescent-based estimation of the match probabilities with BATWING, we used the same distributions of population size, growth rate and mutation rates as employed in the simulation of the source population. This was done in order to verify whether coalescent-based estimation was feasible at all. In practice, such population and mutation parameters may not be known. However, BATWING [15] allows the specification of locus-specific prior distributions that would enable meaningful application of the coalescent-based approach even in cases of uncertainty about the parameters (see subsection "Real data" below).

BATWING's thinning parameters `Nbetsamp` and `treebetN` were both set equal to 15 after minor initial calibration (see the BATWING documentation for further details).

3.2. Real data

We analyzed the 1774 German 17-loci haplotypes from release 37 of the YHRD (http://www.yhrd.org) [7]. To render the data amenable to both coalescent-based estimation and frequency surveying, some markers and haplotypes had to be excluded. Thus, `DYS385a/b` was ignored because of its inherent genotype ambiguity [8], leaving 15 markers for further analysis. Next, four haplotypes with two alleles reported at `DYS19` and 13 haplotypes with intermediate alleles were excluded, leaving $n = 1757$ haplotypes in the data set. Finally, alleles at `DYS389II` were replaced by `DYS389II` minus `DYS389I` [21]. Of the 1757 haplotypes analyzed, 1469 were singletons (83.6%).

When restricting the genotype information to the 7-loci so-called 'minimal haplotype' comprising `DYS19`, `DYS389I`, `DYS389II`, `DYS390`, `DYS391`, `DYS392`, and `DYS393`, a total of 392 singletons (22.3%) were observed in the German data. Ten singletons were drawn randomly from the database and the match probability estimates obtained with the different estimators were compared.

Coalescent-based match probabilities were estimated from 5 million simulations per singleton, after a 50,000 simulations burn-in of the Monte Carlo Markov Chain. All estimations were carried out assuming exponential growth with a `Gamma`(1, 1) prior on the growth rate [15], no migration, a `Gamma`(3, 0.0001) prior on the effective population size, and fixed mutation rates from http://www.yhrd.org as of September 26th, 2012 (`DYS19`: 0.002299, `DYS389I`: 0.002523, `DYS389II`: 0.003644, `DYS390`: 0.002102, `DYS391`: 0.002599, `DYS392`: 0.004123, `DYS393`: 0.001045).

The same thinning parameters as for the simulated data were used (i.e. both `Nbetsamp` and `treebetN` were set equal to 15).

3.3. Frequency surveying

Let n_i be the number of times the ith haplotype has been observed in the database including the suspect profile, with $n = \sum_i n_i$ equal to the size of this augmented database and let d_{ij} be the minimum number of mutational steps separating the ith from the jth haplotype. In its revised form, haplotype surveying [12] is based upon an exponential regression model

$$\mu_i = \exp(r_1 W_i + r_2), \tag{9}$$

$$\sigma_i = \exp(s_1 W_i + s_2), \tag{10}$$

that links mean μ_i and standard deviation σ_i of the population frequency of the ith haplotype to the weighted inverse molecular distance, $W_i = n^{-1}\sum_{j\neq i}(n_j/d_{ij})$, between this haplotype and all other haplotypes in the database. Once the regression parameters r_1, r_2, s_1 and s_2 have been determined, the model serves to define a prior Beta distribution for the frequency of any haplotype h_0 with inverse distance value W_0. The parameters for this prior distribution are

$$\alpha_0 = \frac{\mu_0^2(1-\mu_0)}{\sigma_0^2} - \mu_0, \tag{11}$$

$$\beta_0 = \alpha_0\left(\frac{1-\mu_0}{\mu_0}\right). \tag{12}$$

Maximum likelihood estimates of the regression parameters were obtained in our study by numerical optimization [12] using the Nelder-Mead simplex algorithm with up to 1500 iterations, as implemented in `R` [20]. Several different starting values of (r_1, r_2, s_1, s_2) were tried, and the vector resulting in the highest likelihood was chosen. For the simulated data, starting values were taken from the Cartesian product $\{15,\ 20\} \times \{-10,\ -15\} \times \{15,\ 20\} \times \{-10,\ -15\}$, resulting in 16 possible vectors to choose from. For the real data, starting values were taken from $\{15,\ 20,\ 30.82\} \times \{-10,\ -15,\ -13.17\} \times \{15,\ 20,\ 28.95\} \times \{-10,\ -15,\ -11.71\}$. The additional elements for the real data are the respective binning estimates for the Western-European population adopted from Table 3 of [12].

For comparison to the other estimators, we used the mean of the posterior Beta($\alpha_i + n_i - 1$, $\beta_i + n - n_i$), given by

$$\frac{\alpha_i + n_i - 1}{\alpha_i - 1 + \beta_i + n}, \tag{13}$$

as the haplotype surveying estimate of the sought-after match probability for h_i. Note that $n_i = 1$ as far as singletons were concerned.

4. Results

4.1. Comprehensive analysis of all singletons

Fig. 2 illustrates a comparison of the different singleton match probability estimators for database size $n = 100$. Obviously, both the uncorrected count estimator $1/(n+1)$ and the surveying estimator are rather conservative in that almost all estimates were larger than the corresponding true match probability. Brenner's and the coalescent-based estimator, on the other hand, yielded consistently lower estimates and were found to have small and comparable bias. However, whereas the coalescent-based and surveying estimates were moderately correlated with the true

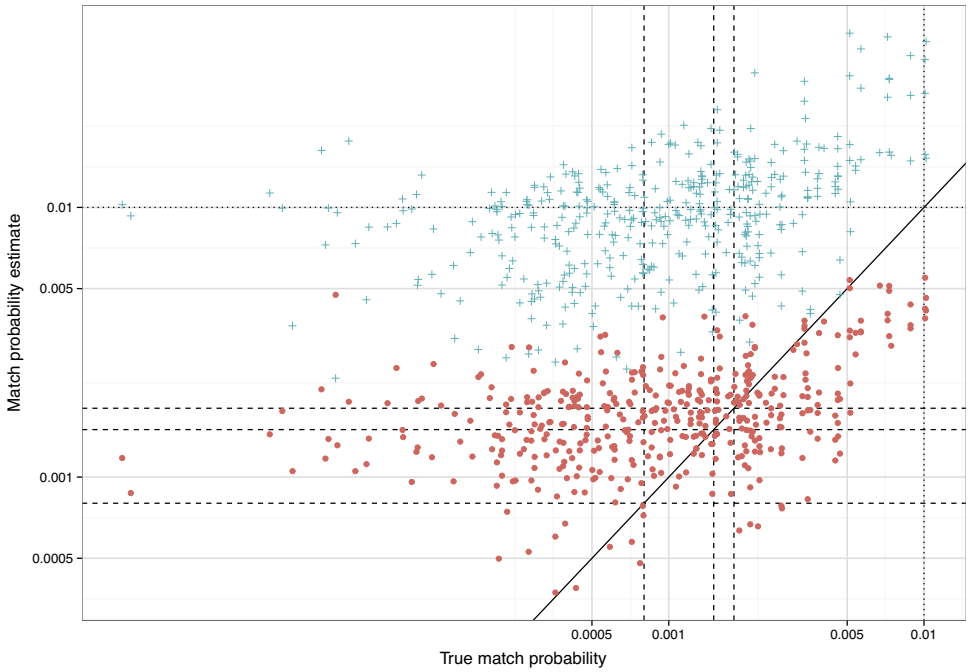

Fig. 2. Singleton match probability estimates for five sample databases of size $n = 100$. The uncorrected count estimate (dotted line) was $1/(99 + 1) = 0.01$ throughout whereas Brenner's estimate varied between 0.0008 and 0.0018, with a mean of 0.0015 (dashed lines). Blue crosses: surveying estimates; red dots: coalescent-based estimates. Each point corresponds to one singleton haplotype. The solid line equates the estimated with the true match probability (i.e. the underlying population frequency). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

match probabilities, by definition, no such relationship exists for the two count estimators (uncorrected and Brenner's).

Inspection of Fig. 2 also reveals that, for singletons with a true match probability smaller than the average of Brenner's estimates, this probability may be difficult to assess by the coalescent-based method in general. On the other hand, singletons with a true match probability above Brenner's average appear to contain sufficient evolutionary information to allow much more precise estimation.

For databases of size $n = 100$ and $n = 200$, Brenner's and the coalescent-based estimator are obviously less biased and have smaller MSE than the surveying estimator (Table 2). In fact, the latter was consistently found to overestimate the true match probability. Also, for $n = 100$, the coalescent-based estimator had lower MSE than Brenner's. This relationship became reverted for $n = 200$ (Table 2), but there is good reason to believe that this observation essentially reflects insufficient convergence of the coalescent-based estimator because MSE is also a function of the variance of an estimator. As was mentioned above, inspection of the Spearman rank correlation coefficients revealed a moderate correlation with the true match probability for both the surveying and the coalescent-based estimates (Table 2). The correlation between the coalescent-based estimates and the true match probabilities was also found to increase with the number of simulations performed (Fig. 3). The same was true for the bias and MSE, both of which converged when the number of simulations increased (Figures A.2 and A.3 in the Supplementary material).

4.2. In-depth analysis of coalescent-based estimates for selected singletons

Table 3 summarizes an in-depth analysis of the coalescent-based match probability estimates obtained for 10 randomly selected singletons per sample database of size $n = 200$, using a much larger number of simulations than before. In general, a substantial increase in simulation number from 200,000 to one million reduced

Table 2
Comparative analysis of singleton match probability estimators. MSE: mean squared error; Spearman: Spearman rank correlation coefficient between estimated and true match probabilities. Sample database numbers are consistent across Tables 1–3.

Size	Sample database	Bias			MSE			Spearman		
		Brenner	Surveying	Coalescent	Brenner	Surveying	Coalescent	Brenner	Surveying	Coalescent
100	1	-9.4×10^{-5}	9.1×10^{-3}	2.9×10^{-4}	4.3×10^{-6}	1.5×10^{-4}	2.7×10^{-6}	0	0.528	0.446
	2	-3.3×10^{-4}	8.9×10^{-3}	2.7×10^{-4}	4.7×10^{-6}	8.3×10^{-5}	2.4×10^{-6}	0	0.566	0.509
	3	4.3×10^{-4}	9.6×10^{-3}	4.2×10^{-4}	2.4×10^{-6}	9.5×10^{-5}	2.0×10^{-6}	0	0.413	0.327
	4	5.4×10^{-5}	8.8×10^{-3}	-4.0×10^{-5}	3.4×10^{-6}	9.8×10^{-5}	2.3×10^{-6}	0	0.401	0.274
	5	-6.4×10^{-4}	8.1×10^{-3}	2.5×10^{-4}	2.5×10^{-6}	9.6×10^{-5}	1.9×10^{-6}	0	0.389	0.266
200	1	7.7×10^{-5}	4.1×10^{-3}	5.6×10^{-5}	1.8×10^{-6}	2.7×10^{-5}	2.3×10^{-6}	0	0.309	0.154
	2	3.5×10^{-4}	4.9×10^{-3}	3.3×10^{-4}	2.6×10^{-6}	2.6×10^{-5}	3.8×10^{-6}	0	0.490	0.267
	3	6.1×10^{-4}	4.2×10^{-3}	1.4×10^{-4}	1.8×10^{-6}	2.5×10^{-5}	1.9×10^{-6}	0	0.283	0.343
	4	4.9×10^{-4}	4.7×10^{-3}	1.2×10^{-4}	1.7×10^{-6}	2.7×10^{-5}	3.3×10^{-6}	0	0.381	0.184
	5	-8.4×10^{-5}	4.3×10^{-3}	9.8×10^{-5}	2.0×10^{-6}	2.2×10^{-5}	2.8×10^{-6}	0	0.389	0.250

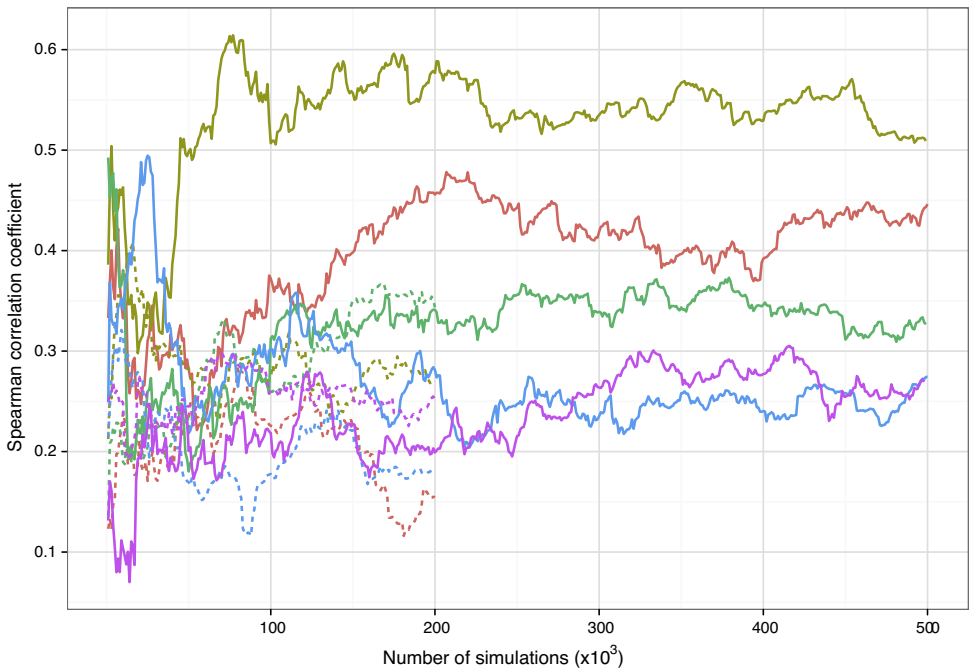

Fig. 3. Trace plots of the Spearman rank correlation coefficient between true match probabilities and coalescent-based estimates, after a given number of simulations. Each line corresponds to one of five databases per database size n, sampled at random from a large simulated source population. Solid lines: $n = 100$; dashed lines: $n = 200$.

both bias and MSE. We also generated two individual trace plots of one million simulations and included these into the Supplementary material. For one singleton (Figure A.4) convergence of the match probability estimate was lacking while, for the other singleton (Figure A.5), the match probability estimate converged quite rapidly.

4.3. Real data

Trace plots for the 7-loci match probabilities of ten singletons randomly chosen from the real German Y-STR data are given in Fig. 4. In some instances, but not all, the coalescent-based estimates seem to have converged to a value near Brenner's and the surveying estimates. A singleton that does not seem to have converged at all is H01675. Inspired by Felsenstein [22], we drew 10 random subsamples of 50 haplotypes each from the original database to see if the subsample estimates for H01675 approximated the other estimates. The trace plots can be found in Figure A.8 of the Supplementary material. Since the true match probabilities were unknown for the real data, a comparison of the different estimators in terms of their accuracy was not possible. However, the mean subsample estimates for H01675 were in the range of 10^{-2} to 10^{-3}, indicating that the original coalescent-based estimate had indeed not converged, despite the large number of simulations performed.

5. Discussion

5.1. General appraisal of coalescent-based match probability estimation

Our simulation study revealed that, overall, the coalescent-based estimator of trace-suspect match probabilities performs better for Y-STR singleton haplotypes than any other previously proposed estimator, at least under the conditions of our simulation study. In terms of both its bias and mean squared error (MSE), the coalescent-based approach was found to be clearly superior to the surveying method [8,9,12]. Moreover, it also outperformed Brenner's κ correction [11] regarding the correlation between estimated and true match probability which, by definition, equals zero for Brenner's estimator. The said correlation also indirectly corroborated the claim, made in connection with the first introduction of the surveying method [8], that the allelic spectrum of a given database contains valuable information about the evolutionary relatedness of its constituent haplotypes, and therefore about match probabilities.

This view is further supported by the observation that, for all the singletons analyzed in our simulation study combined (Table 1), the correlation between the true match probabilities and their coalescent-based estimates increases with the key parameter of the surveying method [12], namely the weighted inverse

Table 3
In-depth analysis for 10 selected singletons per sample database of the coalescent-based estimator of match probabilities, using different numbers of simulations (2×10^5 and 10^6 per singleton). Sample database numbers are consistent across Tables 1–3.

Size	Sample database	Bias			MSE		
		2×10^5	10^6	Brenner	2×10^5	10^6	Brenner
200	1	-4.1×10^{-4}	-1.8×10^{-4}	-4.7×10^{-4}	6.0×10^{-6}	6.0×10^{-6}	8.9×10^{-6}
	2	-6.7×10^{-4}	-4.6×10^{-4}	-2.0×10^{-4}	3.5×10^{-6}	2.3×10^{-6}	4.8×10^{-6}
	3	-6.9×10^{-4}	-5.6×10^{-4}	1.9×10^{-5}	1.3×10^{-6}	1.3×10^{-6}	1.7×10^{-6}
	4	1.4×10^{-3}	4.6×10^{-4}	6.7×10^{-4}	6.4×10^{-6}	1.8×10^{-6}	1.1×10^{-6}
	5	-3.6×10^{-4}	-3.2×10^{-4}	-2.8×10^{-4}	1.2×10^{-6}	1.2×10^{-6}	2.5×10^{-6}

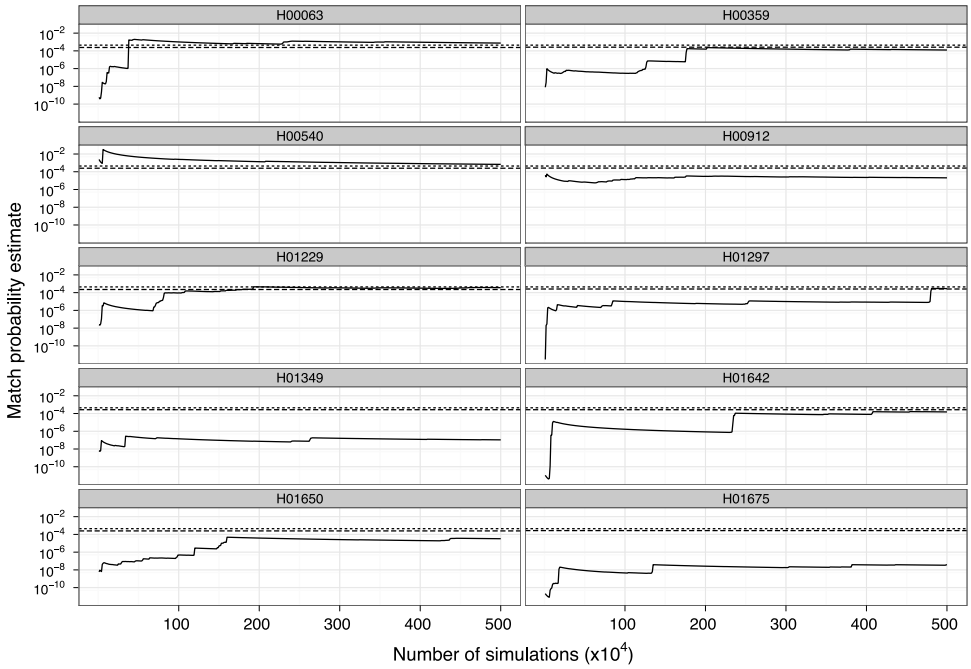

Fig. 4. Trace plots of selected match probability estimates from the real German 7-loci Y-STR data. Match probabilities were estimated for 10 singletons using the coalescent-based approach with 5 million simulations (solid lines), Brenner's κ correction (dotted lines) and the surveying estimator (dash-dotted lines). Note: Brenner's estimate equaled 4.0×10^{-4} while the surveying estimate ranged from 2.3×10^{-4} to 2.7×10^{-4} for the 10 haplotypes. Since the vertical axis has a logarithmic scale, the less than two-fold difference between the two types of estimates implied that they were depicted in close proximity. Trace plots of a subsample study of H01675 can be found in Figure A.8 of the Supplementary material.

molecular distance W between a singleton and the rest of the corresponding reference database (Table 4).

The major downside of the coalescent-based approach consists in its enormous computational demands. These render any widespread practical application of the method difficult, at least until more powerful algorithms to sample coalescent trees have been developed and implemented in suitable software packages. Moreover, because of the large number of singletons assessed in our study, the number of simulations performed for each individual estimate had to be comparatively low. Therefore, the resulting biases and MSEs still have to be interpreted with some caution. This notwithstanding, if applied to derive only one or a few match probability estimates, and with a greater number of simulations thus possible, our in-depth analysis of selected singletons suggests that the accuracy of the coalescent-based method will surpass that of the other approaches tested.

As has been mentioned in Section 3, the Bayesian framework of BATWING [15] allows the specification of prior distributions for the coalescent parameters, including the effective population size, (locus-specific) mutation rates and population growth. This way, any uncertainty about the respective quantities (as would arise in practical casework) can be incorporated into the population model and the posterior distributions derived. Here, we used fixed mutation rates and standard prior distributions for the other parameter values because our main interest was to determine if and how the coalescent-based method would work in principle. Along the same vein, we employed a simplified mutation model in our study for which the upward and downward mutation rates were assumed to be equal. In practice, if the coalescent-based approach was to be used to estimate real match probabilities, this assumption can be abandoned in favor of allele- and direction-specific mutation rates for the Y-STRs of interest, although such modifications may require substantial alteration of the software used.

To assess the robustness of the coalescent-based estimate, we also varied the mutation rate and the prior distribution of the effective population size. The resulting trace-plots can be found in Figures A.6 and A.7 of the Supplementary material. With all the different values and priors tested, the coalescent-based estimator turned out to be quite robust.

Table 4
For each range of W_i values, the Spearman rank correlation coefficient between the coalescent-based estimates and the true match probabilities is given together with the number of singletons in each range.

W_i range	No. singletons	Spearman
(0.05, 0.10]	138	0.077
(0.10, 0.15]	451	0.082
(0.15, 0.20]	331	0.130
(0.20, 0.25]	154	0.155
(0.25, 0.40]	50	0.442

5.2. Match probabilities for non-singletons

In our study, we focused upon singleton haplotypes, i.e. haplotypes for which the estimation of match probabilities appears to be most problematic because the commonly used count estimator $1/(n+1)$ is rather conservative. Moreover, singleton proportions are bound to increase with the number of markers included in a genetic profile, and particularly so when rapidly mutating Y-STRs [23] are involved. However, one important advantage of the coalescent-based (and the surveying) estimator over Brenner's κ correction of the count estimate is that singletons are not treated differently from other, more frequent haplotypes. Therefore, the coalescent-based method can be expected to work as reliably for non-singletons as for singletons, although this supposition still needs to be confirmed systematically.

5.3. Computational recommendations

The coalescent-based method is still on the verge of being too slow for practical application, at least with the software used here.

This is because the computation time required grows exponentially with both the database size and the number of loci involved (Figure A.1). In addition, the more markers are included in a genetic profile, and the larger the database used to quantify the evidential value of a match, the more simulations are required to guarantee proper convergence of the coalescent-based estimate of the match probability. Therefore, the practical application of the coalescent-based approach would currently be limited to rather small databases and to small numbers of markers.

The above notwithstanding, some recommendations can still be made to facilitate efficient and sensible use of the existing simulation software. First, when using Metropolis-Hastings sampling [24,25] as done in BATWING [15], it is important to carefully choose the acceptance rates so as to ensure that the algorithm visits a sufficiently large proportion of the parameter space. There are guidelines regarding the best choice of proposal functions and acceptance rates [26] and these should be adopted if and when meaningful. Second, thinning parameters such as `Nbetsamp` and `treebetN` should be calibrated to individual cases, for example, by consulting autocorrelation plots and statistics, so that the simulations are made approximately independent. Third, the rate and quality of the convergence of individual estimates should be assessed by trace plots similar to those of Fig. 4. Finally, like with other Markov Chain Monte Carlo methods, a burn-in is recommended for the use of BATWING.

Acknowledgements

We wish to thank Ian J. Wilson, Newcastle upon Tyne, for providing non-released parts of the forensic match probability extension of the BATWING program and for helping us with implementing them in BATWING. We also wish to thank Charles Brenner, Oakland, for additional comments on [11] and for a fruitful discussion of the topic in general. Two anonymous reviewers are gratefully acknowledged for helping us to improve our paper.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.fsigen.2012.11.004.

References

[1] I.W. Evett, B.S. Weir, Interpreting DNA Evidence, Sinauer Associates, Sunderland, Massachusetts, USA, 1998.
[2] D.J. Balding, R.A. Nichols, DNA profile match probability calculation: how to allow for population stratification, relatedness, database selection and single bands, Forensic Sci. Int. 64 (1994) 125–140.
[3] P. Gill, A.J. Jeffreys, D.J. Werrett, Forensic application of DNA fingerprints, Nature 318 (1985) 577–579.
[4] L. Roewer, Y chromosome STR typing in crime casework, Forensic Sci. Med. Pathol. 5 (2009) 77–84.
[5] I. Sibille, C. Duverneuil, G.L. de la Grandmaision, K. Guerrouache, M.D.F. Teissière, P. de Mazancourt, Y-STR DNA amplification as biological evidence in sexually assaulted female victims with no cytological detection of spermatozoa, Forensic Sci. Int. 125 (2002) 212–216.
[6] J. Buckleton, M. Krawczak, B. Weir, The interpretation of lineage markers in forensic DNA testing, Forensic Sci. Int. Genet. 5 (2) (2011) 78–83, haploid DNA markers in Forensic Genetics.
[7] S. Willuweit, L. Roewer, Y chromosome haplotype reference database (YHRD): update, Forensic Sci. Int. Genet. 1 (2) (2009) 83–87.
[8] L. Roewer, M. Kayser, P. de Knijff, et al., A new method for the evaluation of matches in non-recombining genomes: application to Y-chromosomal short tandem repeat (STR) haplotypes in European males, Forensic Sci. Int. 114 (1) (2000) 31–43.
[9] M. Krawczak, Forensic evaluation of Y-STR haplotype matches: a comment, Forensic Sci. Int. 118 (2-3) (2001) 114–115.
[10] M.M. Andersen, Y-STR: Haplotype Frequency Estimation and Evidence Calculation, Master's thesis, Aalborg University, Denmark, 2010.
[11] C.H. Brenner, Fundamental problem of forensic mathematics – the evidential value of a rare haplotype, Forensic Sci. Int. Genet. 4 (5) (2010) 281–291.
[12] S. Willuweit, A. Caliebe, M.M. Andersen, L. Roewer, Y-STR frequency surveying method: a critical reappraisal, Forensic Sci. Int. Genet. 5 (2) (2011) 84–90, haploid DNA markers in Forensic Genetics.
[13] L. Roewer, M. Krawczak, S. Willuweit, et al., Online reference database of European Y-chromosomal short tandem repeat (STR) haplotypes, Forensic Sci. Int. 2-3 (2001) 106–113.
[14] H.E. Robbins, Estimating the total probability of the unobserved outcomes of an experiment, Ann. Math. Stat. 39 (1) (1968) 256–257.
[15] I.J. Wilson, M.E. Weale, D.J. Balding, Inferences from DNA data: population histories, evolutionary processes and forensic match probabilities, J. R. Stat. Soc. Ser. A 166 (2003) 155–201.
[16] I.J. Wilson, D.J. Balding, Genealogical inference from microsatellite data, Genetics 150 (1998) 499–510.
[17] J. Kingman, The coalescent, Stoch. Process. Appl. 13 (3) (1982) 235–248.
[18] J. Hein, M.H. Schierup, C. Wiuf, Gene Genealogies, Variation and Evolution: A Primer in Coalescent Theory, Oxford University Press, New York, USA, 2005.
[19] J.G. Skellam, The frequency distribution of the difference between two Poisson variates belonging to different populations, J. R. Stat. Soc. Ser. A 109 (3) (1946) 296.
[20] R Development Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2010.
[21] J.M. Butler, Forensic DNA Typing: Biology, Technology, and Genetics of STR Markers, 2nd ed., Academic Press, Burlington, Massachusetts, USA, 2005.
[22] J. Felsenstein, Accuracy of coalescent likelihood estimates: do we need more sites, more sequences, or more loci? Mol. Biol. Evol. 23 (2006) 691–700.
[23] K.N. Ballantyne, et al., Mutability of Y-chromosomal microsatellites: rates, characteristics, molecular bases, and forensic implications, Am. J. Hum. Genet. 87 (3) (2010) 341–353.
[24] N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller, E. Teller, Equations of state calculations by fast computing machines, Genetics 21 (1953) 1087–1092.
[25] W.K. Hastings, Monte Carlo sampling methods using Markov chains and their applications, Biometrika 57 (1970) 97–109.
[26] A. Gelman, G.O. Robers, W.R. Gilks, Efficient metropolis jumping rules, Bayesian Stat. 5 (1996) 599–607.

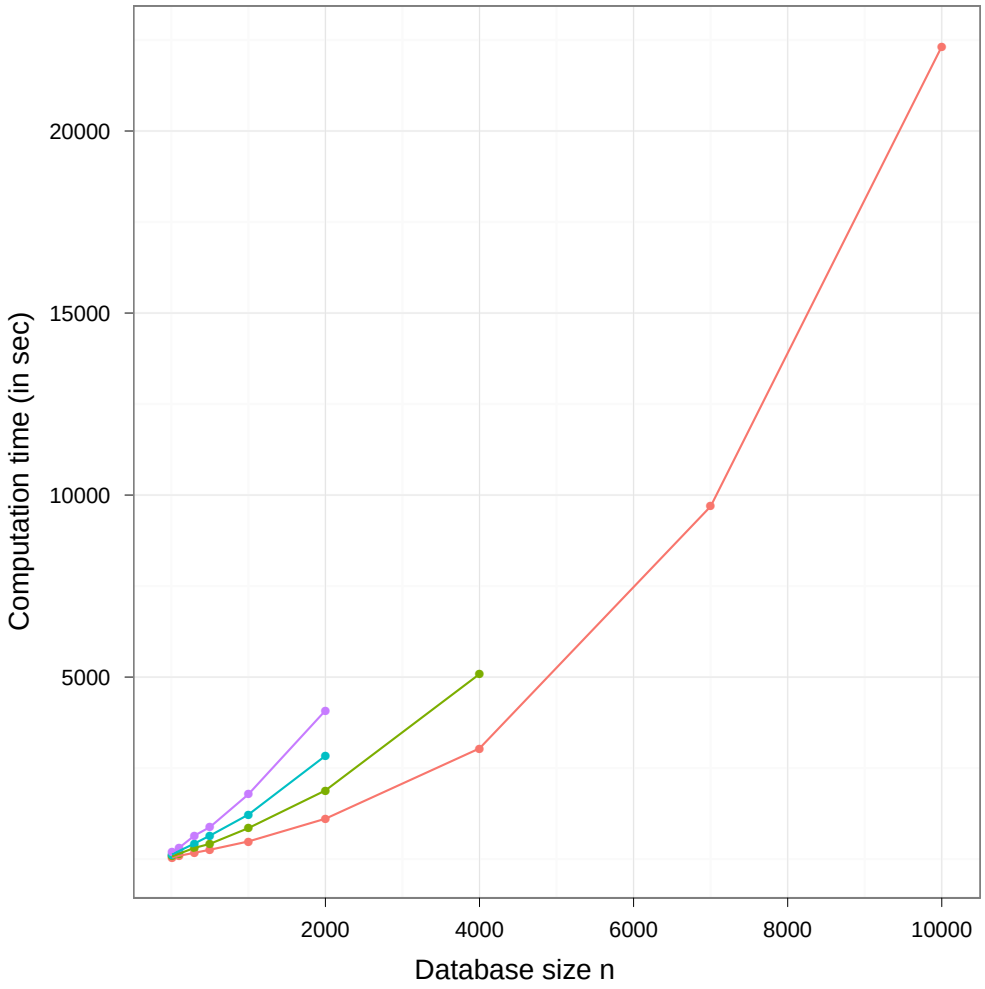

Figure A.1: Computational demand of coalescent-based estimation of singleton match probabilities as a function of database size n and number of loci included in a genetic profile. Calculations were run on an Intel Xeon CPU E5420 at 2.50GHz. Computation times are averages per singleton haplotype. Parameters used were: 10,000 simulations per coalescent tree, a starting population size of 20,000, no population growth, no migration, and a mutation rate of 0.003 per locus per generation. Red dots: 5 loci; green dots: 10 loci; blue dots: 15 loci; black dots: 20 loci.

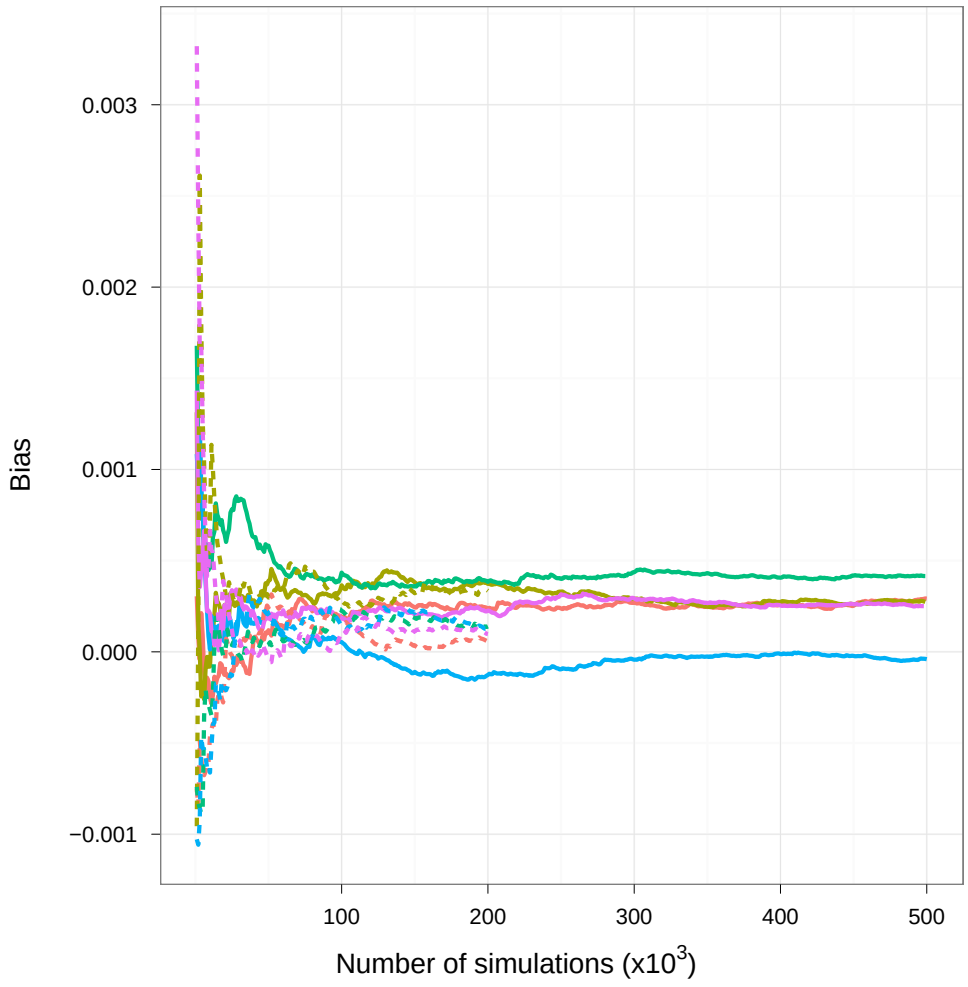

Figure A.2: Bias (see Equation 3 of the main text) of the coalescent-based estimator of singleton match probabilities. Calculation of the bias was based upon all singletons in each of five databases per database size. Solid lines: database size $n = 100$; dashed lines: database size $n = 200$.

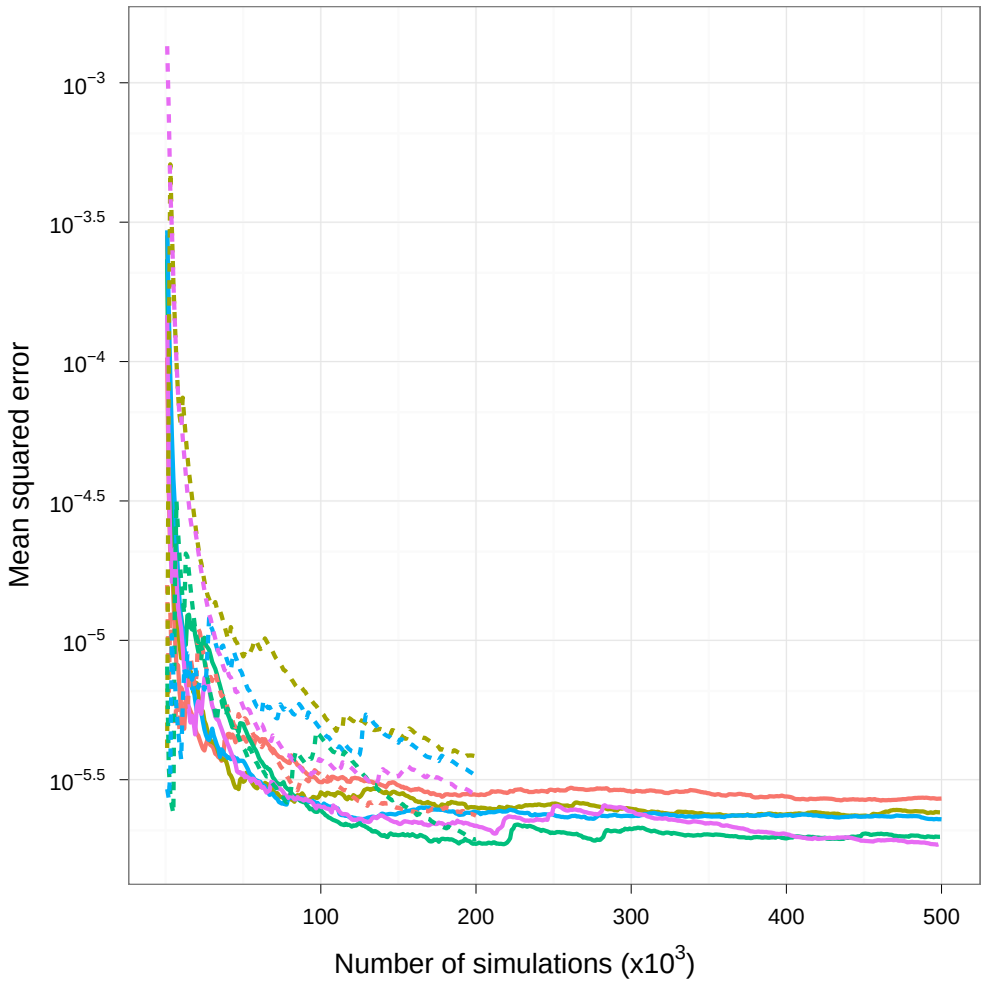

Figure A.3: Mean squared error (see Equation 4 of the main text) of the coalescent-based estimator of singleton match probabilities. Calculation of the MSE was based upon all singletons in each of 5 databases per database size. Solid lines: database size $n = 100$; dashed lines: database size $n = 200$.

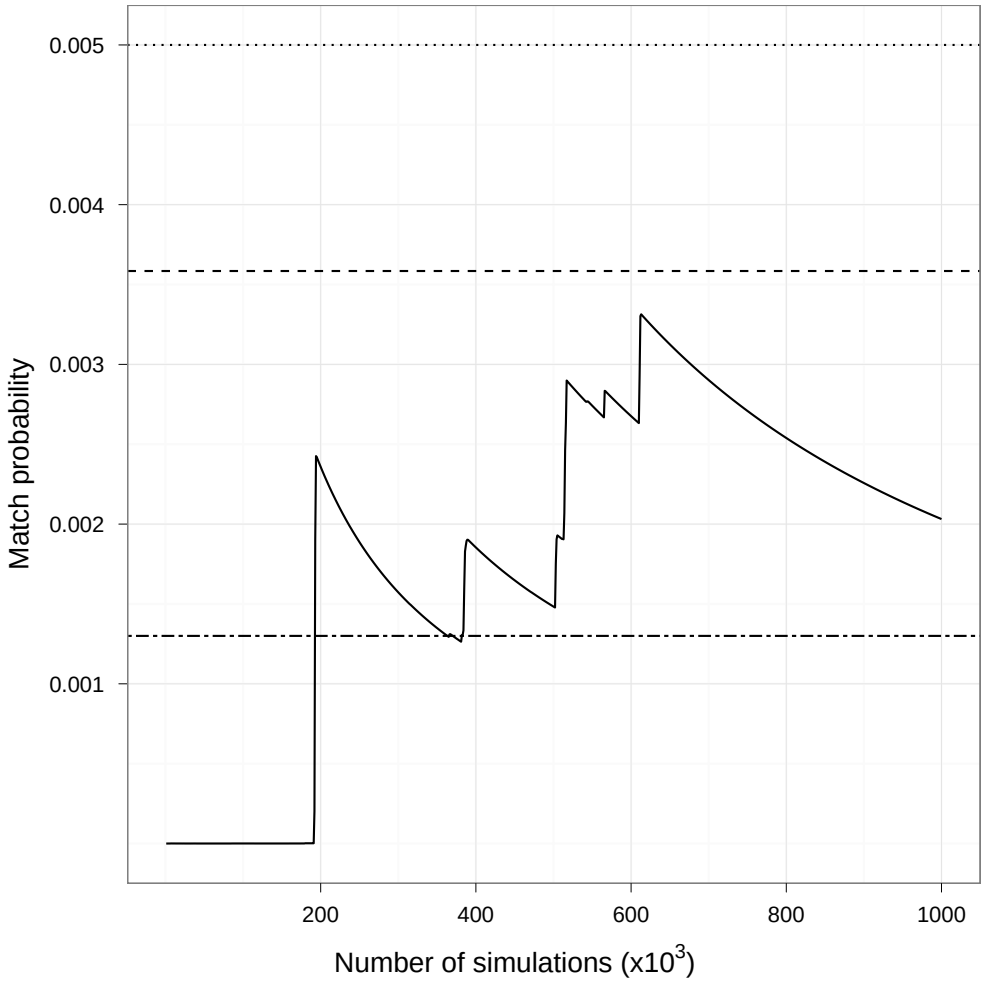

Figure A.4: Trace plot of the coalescent-based match probability estimate (solid line) for a selected singleton from sample database no. 1 of size $n = 200$. Dotted line: uncorrected count estimate; dash-dotted line: Brenner's κ-corrected count estimate, dashed line: true match probability (i.e. the underlying population frequency).

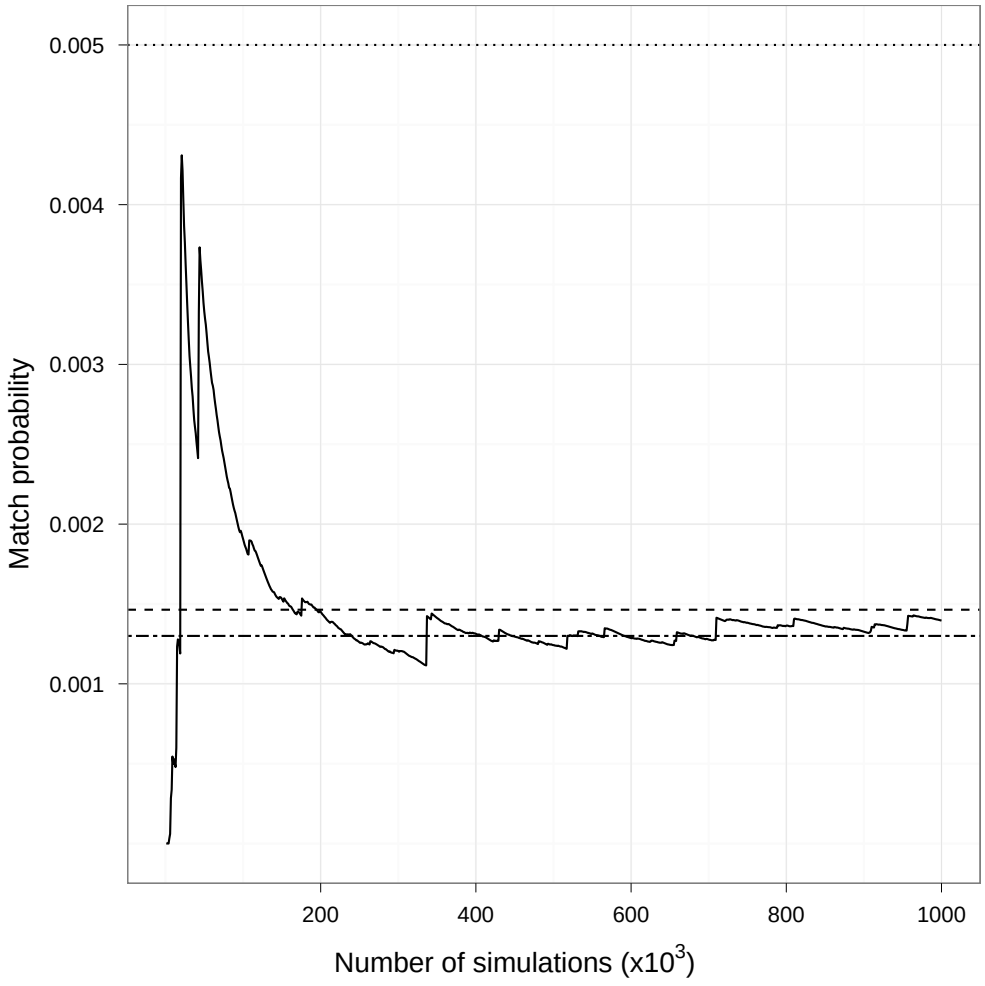

Figure A.5: Trace plot of the coalescent-based match probability estimate (solid line) for a selected singleton from sample database no. 1 of size $n = 200$. Dotted line: uncorrected count estimate; dash-dotted line: Brenner's κ-corrected count estimate; dashed line: true match probability (i.e. the underlying population frequency).

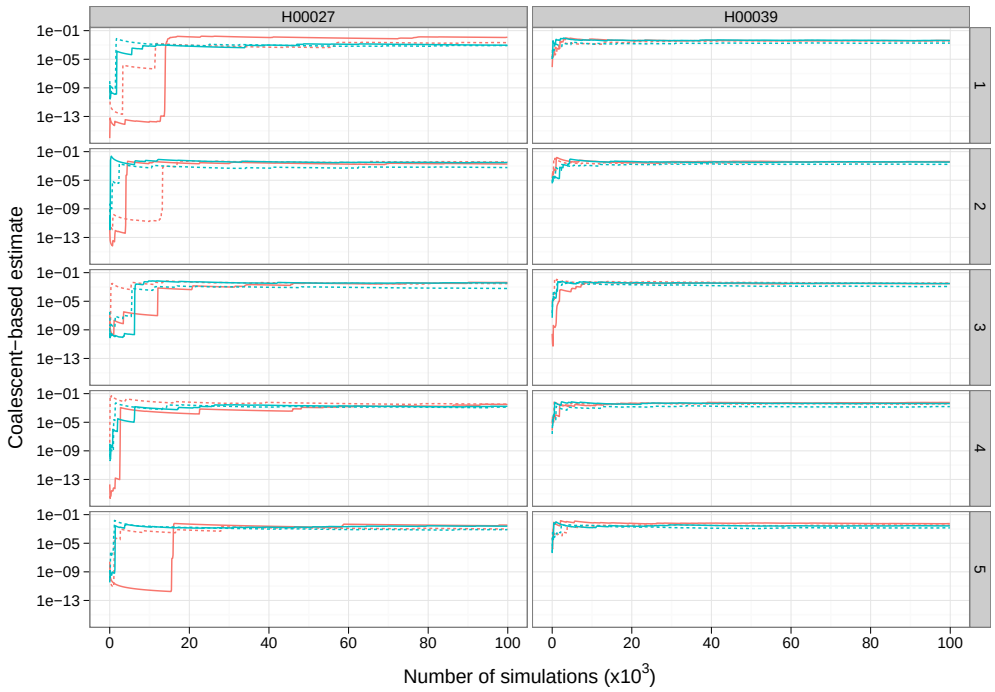

Figure A.6: Trace plots from a robustness study. For each of two randomly selected singletons (columns), five subsamples (rows) of size 20 were drawn from the original database of size 100 (database 1 in Tables 1 and 2 of the main text). Match probabilities were then estimated from each of these subsamples alone, using the coalescent approach. Mutation rates were fixed at either 0.001 (red lines) or 0.003 (blue lines). The effective population size was assumed to be normally distributed with a mean of either 10,000 (solid lines) or 20,000 (dashed lines) and a standard deviation of 3,000. See Figure A.7 for a magnified traceplot where the first 20,000 simulations were discarded as burn-in.

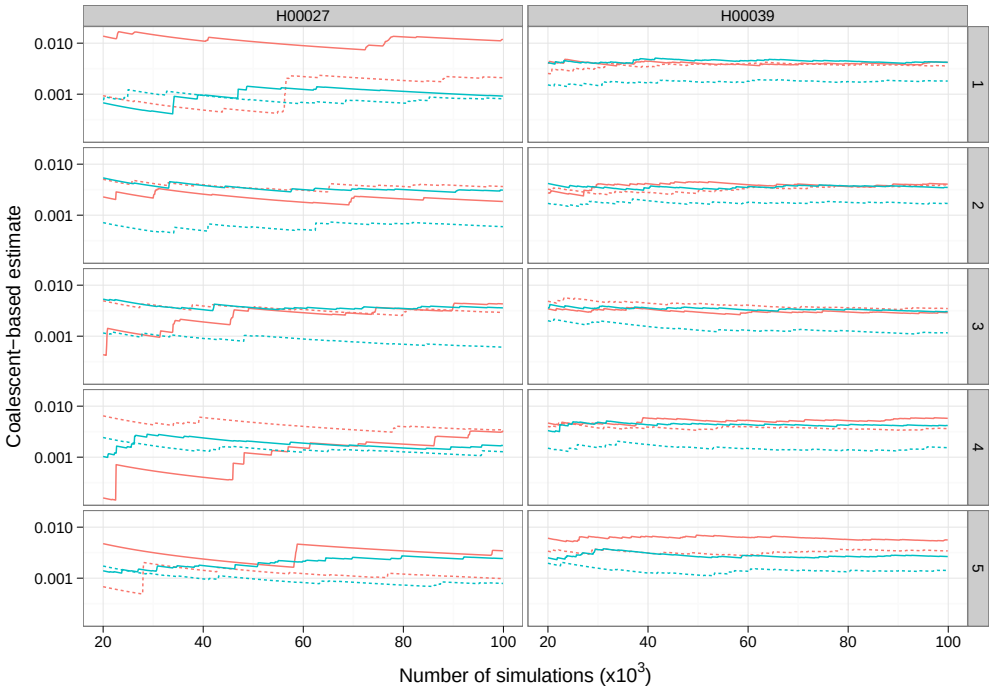

Figure A.7: Same as Figure A.6 but with the first 20,000 simulations discarded as burn-in.

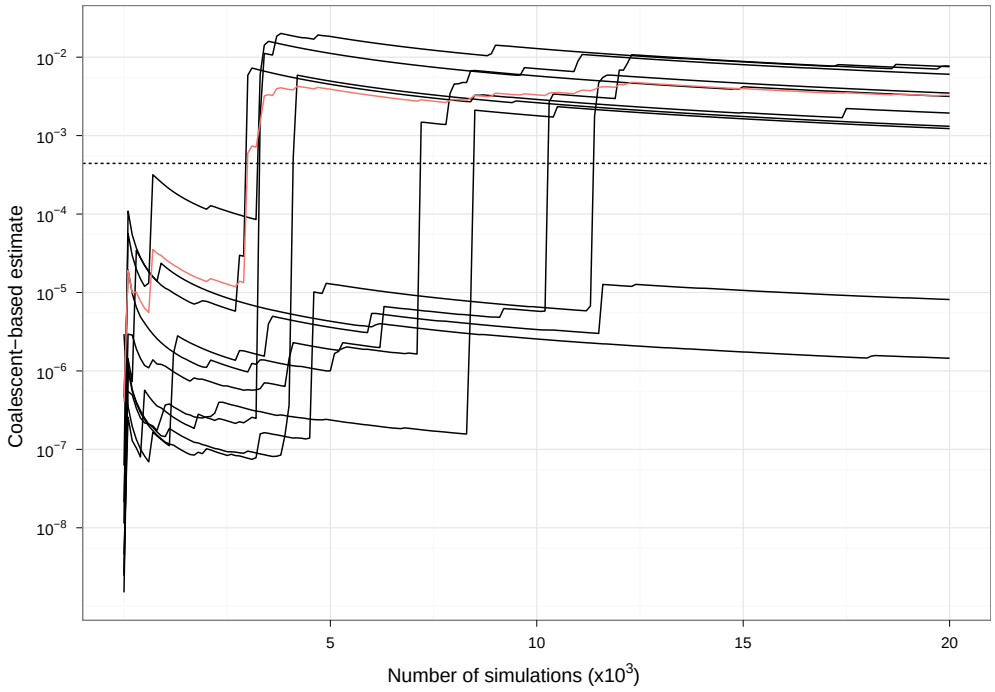

Figure A.8: Subsampling study on singleton haplotype H01675 from the German 7-loci database (1,757 haplotypes). Ten subsamples of 50 haplotypes each were randomly drawn from the database. With each of these subsamples, a coalescent-based estimate was calculated using 20,000 simulations. All estimations were carried out assuming exponential population growth with a `Gamma`$(1, 1)$ prior on the growth rate, no migration, a `Gamma`$(3, 0.0001)$ prior on the effective population size, and the following mutation rates from `http://www.yhrd.org` as of September 26th, 2012: `DYS19`, 0.002299; `DYS389I`, 0.002523; `DYS389II`, 0.003644; `DYS390`, 0.002102; `DYS391`, 0.002599; `DYS392`, 0.004123; `DYS393`, 0.001045. The black solid lines depict the individual coalescent-based estimation processes. The red solid line depicts the mean of individual runs. The dashed black line corresponds to Brenner's estimate.

Kapitel 3

Weitere Resultate

3.1 Wandernde Verteilungen

In Publikation (i) haben wir gezeigt, dass der Allelprozess des einfachen SMM je nach Perspektive zwei verschiedene Verhaltensweisen aufweist. Als globales Verhalten bezeichnen wir dabei die Eigenschaften des gewöhnlichen Allelprozesses, der insbesondere keine Konvergenz zeigt: Wenn man in jeder Generation n ein Individuum zufällig auswählt, geht der Betrag der Allel-Längen mit $n \to \infty$ gegen unendlich. Dem entgegen steht das lokale Verhalten, das man z.B. erfassen kann, indem man nicht die Allele selbst betrachtet, sondern deren Differenzen zu einem in jeder Generation zufällig ausgewählten Referenzallel. Dies war unser Ansatz in Publikation (i), und wir konnten zeigen, dass die zugehörige Markov-Kette im Limes eine stationäre Verteilung hat.

Eine naheliegende Frage ist nun, wie sich die Wahl des Referenzallels auf unsere Ergebnisse auswirkt. Darauf soll im folgenden Abschnitt 3.1.1 näher eingegangen werden, indem eine alternative Normierungsmethode betrachtet wird. Auch dieser Fall zeigt, dass die Allele auf ihrer Drift gegen unendlich gebündelt beieinander bleiben. Dies ist es, was Moran [168] mit seinem Begriff der wandernden Verteilungen meinte. Übrigens wäre der Originalausdruck *wandering distributions* inhaltlich treffender (wenngleich etymologisch entfernter) als *umherschweifende Verteilungen* zu übersetzen, um die Konnotation der Ziellosigkeit nicht zu verlieren [241, S. 1222]. Man könnte auch von Nomadenverteilungen oder ruhelosen Verteilungen sprechen.

In Abschnitt 3.1.2 befassen wir uns noch einmal mit der Existenz der Grenzverteilung η (vgl. S. 339 in Publikation (i)). Und in Abschnitt 3.1.3 diskutieren wir schließlich einige numerische Resultate über η.

3.1.1 Der Mittelwert als Referenzallel

In Publikation (i) haben wir den Allelprozess normalisiert, indem wir das Allel eines Individuums als Referenzallel verwendet haben (Definition 6). Um zu un-

tersuchen, inwieweit sich die Wahl des Referenzallels auf die in Publikation (i) gezeigten grundlegenden Eigenschaften des normalisierten Allelprozeses (Lemma 7) auswirkt, betrachten wir hier die Alternative, den Mittelwert aller Allele in der jeweiligen Generation als Referenzwert zu verwenden. Dazu beweisen wir zunächst zwei Beziehungen, die allgemein für die Varianz bzw. Kovarianz von Zufallsgrößen gelten, wenn diese sich als Differenz zwischen einer Zufallsgröße und dem Mittelwert dieser und weiterer Zufallsgrößen (mit derselben Varianz bzw. Kovarianz) darstellen lassen.

Proposition 3.1.1 *Sei $k \in \mathbb{N}$ und $X_1, \dots X_k$ seien Zufallsgrößen, für die die zweiten Momente existieren. Für alle $i, j, i', j' \in \{1, \dots k\}$ mit $i \neq j$ und $i' \neq j'$ gelte $Var(X_i) = Var(X_j)$ und $Cov(X_i, X_j) = Cov(X_{i'}, X_{j'})$. Dann gilt:*

(i)

$$Var\left(X_1 - \frac{1}{k}\sum_{i=1}^{k} X_i\right) = \frac{k-1}{k}\left(Var(X_1) - Cov(X_1, X_2)\right)$$

(ii)

$$Cov\left(X_1 - \frac{1}{k}\sum_{i=1}^{k} X_i,\ X_2 - \frac{1}{k}\sum_{i=1}^{k} X_i\right) = \frac{1}{k}(Cov(X_1, X_2) - Var(X_1))$$

Beweis: zu (i): Es gilt:

$$\begin{aligned}
&\mathrm{Var}\left(X_1 - \frac{1}{k}\sum_{i=1}^{k} X_i\right) \\
&= \mathrm{Var}\left(X_1 + \sum_{i=1}^{k}\left(-\frac{1}{k}\right) X_i\right) \\
&= \mathrm{Var}(X_1) + \sum_{i=1}^{k}\mathrm{Var}\left(-\frac{1}{k}X_i\right) + 2\sum_{i=1}^{k}\mathrm{Cov}\left(X_1, -\frac{1}{k}X_i\right) + \sum_{i\neq j}\mathrm{Cov}\left(-\frac{1}{k}X_i, -\frac{1}{k}X_j\right) \\
&= \mathrm{Var}(X_1) + \sum_{i=1}^{k}\frac{1}{k^2}\mathrm{Var}(X_i) - \frac{2}{k}\sum_{i=1}^{k}\mathrm{Cov}(X_1, X_i) + \frac{1}{k^2}\sum_{i\neq j}\mathrm{Cov}(X_i, X_j) \\
&= \frac{k+1}{k}\mathrm{Var}(X_1) - \frac{2}{k}(\mathrm{Var}(X_1) + (k-1)\mathrm{Cov}(X_1, X_2)) + \frac{k-1}{k}\mathrm{Cov}(X_1, X_2) \\
&= \frac{k-1}{k}(\mathrm{Var}(X_1) - \mathrm{Cov}(X_1, X_2))
\end{aligned}$$

zu (*ii*): Es gilt:

$$\begin{aligned}
&\mathrm{Cov}\left(X_1 - \frac{1}{k}\sum_{i=1}^{k} X_i,\ X_2 - \frac{1}{k}\sum_{i=1}^{k} X_i\right) \\
&= \mathrm{Cov}(X_1, X_2) - \mathrm{Cov}\left(X_1, \frac{1}{k}\sum_{i=1}^{k} X_i\right) - \mathrm{Cov}\left(\frac{1}{k}\sum_{i=1}^{k} X_i, X_2\right) + \mathrm{Cov}\left(\frac{1}{k}\sum_{i=1}^{k} X_i, \frac{1}{k}\sum_{i=1}^{k} X_i\right) \\
&= \mathrm{Cov}(X_1, X_2) - \frac{2}{k}\mathrm{Cov}\left(X_1, \sum_{i=1}^{k} X_i\right) + \frac{1}{k^2}\mathrm{Cov}\left(\sum_{i=1}^{k} X_i, \sum_{i=1}^{k} X_i\right) \\
&= \mathrm{Cov}(X_1, X_2) - \frac{2}{k}\sum_{i=1}^{k}\mathrm{Cov}(X_1, X_i) + \frac{1}{k^2}\left(\sum_{i=1}^{k}\mathrm{Var}(X_i) + \sum_{i\neq j}\mathrm{Cov}(X_i, X_j)\right) \\
&= \mathrm{Cov}(X_1, X_2) - \frac{2}{k}\mathrm{Var}(X_1) - \frac{2(k-1)}{k}\mathrm{Cov}(X_1, X_2) + \frac{k}{k^2}\mathrm{Var}(X_1) + \frac{k(k-1)}{k^2}\mathrm{Cov}(X_1, X_2) \\
&= \frac{1}{k}(\mathrm{Cov}(X_1, X_2) - \mathrm{Var}(X_1))
\end{aligned}$$

□

Diese Proposition können wir nun auf unseren Allelprozess X anwenden. Dafür verwenden wir die Notation aus Publikation (i) und erhalten folgendes Lemma.

Lemma 3.1.2 *Für alle $n \in \mathbb{N}$, $j \in \{1, \dots N\}$ gilt:*

(i)

$$E\left(X_n(j) - \frac{1}{N}\sum_{i=1}^{N} X_n(i)\right) = 0$$

(ii)

$$Var\left(X_n(j) - \frac{1}{N}\sum_{i=1}^{N} X_n(i)\right) = \left(N - 1 - \frac{(N-1)^{n+1}}{N^n}\right)\cdot\mu$$

(iii)

$$Cov\left(X_n(1) - \frac{1}{N}\sum_{i=1}^{N} X_n(i),\ X_n(2) - \frac{1}{N}\sum_{i=1}^{N} X_n(i)\right) = \left(\left(\frac{N-1}{N}\right)^n - 1\right)\cdot\mu$$

Beweis: Sei $n \in \mathbb{N}$ und $j \in \{1, \dots N\}$.
zu (i): Nach Lemma 3 (i) aus Publikation (i) gilt:

$$\begin{aligned} E\left(X_n(j) - \frac{1}{N}\sum_{i=1}^{N} X_n(i)\right) &= E(X_n(j)) - \frac{1}{N}\sum_{i=1}^{N} E(X_n(i)) \\ &= 0 - \frac{1}{N}\sum_{i=1}^{N} 0 \\ &= 0 \end{aligned}$$

zu (ii): Nach Lemma 3 (ii) und Lemma 4 aus Publikation (i), sowie Proposition 3.1.1 (i) gilt:

$$\begin{aligned} \mathrm{Var}\left(X_n(j) - \frac{1}{N}\sum_{i=1}^{N} X_n(i)\right) &= \frac{N-1}{N}(\mathrm{Var}(X_n(1)) - \mathrm{Cov}(X_n(1), X_n(2))) \\ &= \frac{N-1}{N}\left(n\mu - \left(n + \frac{(N-1)^n - N^n}{N^{n-1}}\right)\mu\right) \\ &= (N-1)\left(1 - \left(\frac{N-1}{N}\right)^n\right)\mu \\ &= \left(N - 1 - \frac{(N-1)^{n+1}}{N^n}\right)\mu \end{aligned}$$

zu (iii): Nach Lemma 3 (ii) und Lemma 4 aus Publikation (i), sowie Proposition 3.1.1 (ii) gilt:

$$\begin{aligned} &\mathrm{Cov}\left(X_n(1) - \frac{1}{N}\sum_{i=1}^{N} X_n(i),\ X_n(2) - \frac{1}{N}\sum_{i=1}^{N} X_n(i)\right) \\ &= \frac{1}{N}(\mathrm{Cov}(X_n(1), X_n(2)) - \mathrm{Var}(X_n(1))) \\ &= \frac{1}{N}\left(\left(n + \frac{(N-1)^n - N^n}{N^{n-1}}\right)\mu - n\mu\right) \\ &= \frac{(N-1)^n - N^n}{N^n}\mu \\ &= \left(\left(\frac{N-1}{N}\right)^n - 1\right)\mu \end{aligned}$$

□

So wie aus Lemma 7 in Publikation (i) direkt folgte, dass $\lim_{n\to\infty} \mathrm{Var}(V_n(i)) = 2\mu N$, also (im Gegensatz zu $\mathrm{Var}(X_n(i))$) insbesondere endlich ist, können wir auch aus Lemma 3.1.2 zwei direkte Folgerungen ziehen:

(i)

$$\mathrm{Var}\left(X_n(1) - \frac{1}{N}\sum_{i=1}^{N} X_n(i)\right) \underset{n\to\infty}{\longrightarrow} (N-1)\mu \tag{3.1}$$

(ii)

$$\mathrm{Cov}\left(X_n(1) - \frac{1}{N}\sum_{i=1}^{N} X_n(i),\ X_n(2) - \frac{1}{N}\sum_{i=1}^{N} X_n(i)\right) \underset{n\to\infty}{\longrightarrow} -\mu \tag{3.2}$$

Es zeigt sich also, dass auch der anhand des Mittelwertes aller Allele normalisierte Allelprozess im Limes endliche Varianzen und Kovarianzen hat.

3.1.2 Existenz der invarianten Verteilung

In Publikation (i) haben wir die Existenz der invarianten Verteilung η bewiesen, indem wir auf Resultate aus der Markov-Ketten-Theorie zurückgegriffen haben (siehe dort den Beweis von Theorem 11). Außerdem gibt es wie erwähnt einen Beweis von Kingman [134] mittels charakteristischer Funktionen. Es existiert aber auch ein elementarer analytischer Beweis, der die erforderlichen Bedingungen noch zusätzlich illustriert und deshalb hier ergänzt werden soll.

Sei M der Raum der Wahrscheinlichkeitsmaße auf $\mathbb{Z}$. Wir benutzen wieder die Notation aus Publikation (i) — insbesondere siehe dort Gleichung (4) für die Definition von $r(\cdot)$ — und definieren zusätzlich den Totalvariationsabstand

$$d : M \times M \to \mathbb{R};\ (\nu, \xi) \mapsto \sum_{z\in\mathbb{Z}} \frac{1}{2}|\nu(z) - \xi(z)| \tag{3.3}$$

und

$$K : M \to M;\ \nu \mapsto \frac{N-1}{N}\sum_{k=-2}^{2} r(k)\cdot\nu(z+k) + \frac{1}{N}r(z). \tag{3.4}$$

Wir bemerken zunächst, dass man leicht mit elementaren Mitteln beweisen kann, dass d eine Metrik, (M, d) ein vollständiger Raum und K wohldefiniert ist.

Proposition 3.1.3 *K ist eine Kontraktion bezüglich d.*

Beweis: Seien $\nu, \xi \in M$. Dann gilt:

$$
\begin{aligned}
&d(K(\nu), K(\xi)) \\
&\quad = \sum_{z\in\mathbb{Z}} \frac{1}{2} \left| \frac{N-1}{N} \sum_{k=-2}^{2} r(k)\nu(z+k) + \frac{1}{N} r(z) - \frac{N-1}{N} \sum_{k=-2}^{2} r(k)\xi(z+k) - \frac{1}{N} r(z) \right| \\
&\quad = \frac{N-1}{2N} \sum_{z\in\mathbb{Z}} \left| \sum_{k=-2}^{2} r(k)(\nu(z+k) - \xi(z+k)) \right| \\
&\quad \leq \frac{N-1}{2N} \sum_{z\in\mathbb{Z}} \sum_{k=-2}^{2} r(k)|\nu(z+k) - \xi(z+k)| \\
&\quad = \frac{N-1}{2N} \sum_{k=-2}^{2} r(k) \sum_{z\in\mathbb{Z}} |\nu(z+k) - \xi(z+k)| \\
&\quad = \frac{N-1}{2N} \sum_{k=-2}^{2} r(k) \sum_{z\in\mathbb{Z}} |\nu(z) - \xi(z)| \\
&\quad = \frac{N-1}{2N} \left(\sum_{z\in\mathbb{Z}} |\nu(z) - \xi(z)| \right) \sum_{k=-2}^{2} r(k) \\
&\quad = \frac{N-1}{2N} \sum_{z\in\mathbb{Z}} |\nu(z) - \xi(z)| \\
&\quad = \frac{N-1}{N} d(\nu, \xi)
\end{aligned}
$$

□

Korollar 3.1.4 *$(\eta_n)_{n\in\mathbb{N}}$ konvergiert in Totalvariationsabstand gegen ein Wahrscheinlichkeitsmaß η.*

Beweis: Für die Verteilungen η_n von $V_n(1)$ haben wir mit Lemma 8 aus Publikation (i) die folgende Rekursionsgleichung bewiesen (für alle $n \in \mathbb{N}$, $z \in \mathbb{Z}$):

$$\eta_n(z) = \frac{N-1}{N} \sum_{k=-2}^{2} r(k)\eta_{n-1}(z-k) + \frac{1}{N} r(z).$$

Nach der Definition von K gilt also für alle $n \in \mathbb{N}$:

$$\eta_{n+1} = K(\eta_n).$$

Da K wie gezeigt eine Kontraktion bezüglich der Metrik d ist, folgt das Korollar somit aus dem Fixpunktsatz von Banach.

□

3.1.3 Approximation der invarianten Verteilung

Mittels der Rekursionsgleichung (Lemma 8) aus Publikation (i) bzw. ihrer Lösung (Theorem 13 (i)) kann für jede Generation n die Verteilung η_n von $V_n(1)$ berechnet werden. Da die Folge $(\eta_n)_{n\in\mathbb{N}}$, wie wir in Publikation (i) gezeigt haben (Theorem 11), mit $n \to \infty$ exponentiell schnell in Verteilung gegen die invariante Verteilung η konvergiert, hat man so auch die Möglichkeit, η numerisch zu approximieren. In Tabelle 3.1 sind einige solche Resultate zusammengefasst.

		$\mu = 0,02$	$\mu = 0,04$	$\mu = 0,06$	$\mu = 0,08$	$\mu = 0,1$
		$\theta = 0,4$	$\theta = 0,8$	$\theta = 1,2$	$\theta = 1,6$	$\theta = 2$
	$\eta(0)$	0,739	0,606	0,522	0,463	0,418
$N = 10$	$\eta(1)$	0,113	0,153	0,171	0,180	0,183
	$\eta(2)$	0,0155	0,0342	0,0487	0,0596	0,0679
		$\theta = 0,8$	$\theta = 1,6$	$\theta = 2,4$	$\theta = 3,2$	$\theta = 4$
	$\eta(0)$	0,613	0,476	0,400	0,350	0,314
$N = 20$	$\eta(1)$	0,149	0,174	0,178	0,177	0,174
	$\eta(2)$	0,0342	0,0587	0,0725	0,0807	0,0859
		$\theta = 1,2$	$\theta = 2,4$	$\theta = 3,6$	$\theta = 4,8$	$\theta = 6$
	$\eta(0)$	0,536	0,405	0,337	0,293	0,263
$N = 30$	$\eta(1)$	0,164	0,176	0,173	0,167	0,161
	$\eta(2)$	0,0481	0,0719	0,0828	0,0882	0,0908
		$\theta = 1,6$	$\theta = 3,2$	$\theta = 4,8$	$\theta = 6,4$	$\theta = 8$
	$\eta(0)$	0,482	0,359	0,297	0,258	0,231
$N = 40$	$\eta(1)$	0,171	0,173	0,166	0,158	0,151
	$\eta(2)$	0,0582	0,0796	0,0877	0,0908	0,0918
		$\theta = 2$	$\theta = 4$	$\theta = 6$	$\theta = 8$	$\theta = 10$
	$\eta(0)$	0,442	0,326	0,268	0,233	0,208
$N = 50$	$\eta(1)$	0,173	0,169	0,160	0,151	0,143
	$\eta(2)$	0,0657	0,0843	0,0900	0,0915	0,0913

Tabelle 3.1: Die invariante Verteilung η für verschiedene Werte der Populationsgröße N und der Mutationswahrscheinlichkeit μ. Angegeben sind jeweils $\theta = 2N\mu$ sowie $\eta(0)$, $\eta(1) = \eta(-1)$ und $\eta(2) = \eta(-2)$. Die Werte basieren jeweils auf 10.000 Iterationen (siehe Publikation (i), Lemma 8).

Es fällt auf, dass die invariante Verteilung η für verschiedene Werte von N und μ, die dasselbe Produkt $\theta = 2N\mu$ ergeben, grob übereinstimmt. θ kontrolliert somit insbesondere die Breite der Verteilung: Für $\theta = 0,4$ sind 99,5% der Wahrscheinlichkeitsmasse bereits durch den aufgelisteten Träger $\{-2,-1,0,1,2\}$ abgedeckt, für $\theta = 2$ sind es 92% und für $\theta = 10$ nur 68%.

Dass η eine unimodale Verteilung ist (falls $\mu < 0,8$), haben wir bereits in Publikation (i) bewiesen (Theorem 13 (ii)). Zwar ist die Voraussetzung $\mu < 0,8$

für STRs immer erfüllt, aber aus theoretischer Sicht stellt sich die Frage, ob diese Voraussetzung notwendig ist. Mit anderen Worten: Ist η für $\mu \geq 0,8$ nicht mehr unimodal? Um sich der Antwort auf diese Frage zu nähern ist es hilfreich, den Extremfall $\mu = 1$ zu betrachten (den wir in Publikation (i) durch die Definition von μ von vorneherein ausgeschlossen hatten). In diesem Fall ist die Markov-Kette X periodisch, denn ausgehend von $X_0 \equiv 0$ können die geraden Zustände nur in einer geraden Anzahl von Generationen erreicht werden (und die ungeraden Zustände nur in einer ungeraden Anzahl von Generationen), da immer alle Allele um ± 1 mutieren. V kann in diesem Fall also in jeder Generation nur gerade Zustände annehmen, da zu jedem Zeitpunkt entweder alle Allele gerade oder alle Allele ungerade sind. Der Beweis für die Existenz von η aus Publikation (i) ist auf diesen Fall demnach nicht anwendbar, denn für die dort verwendete Bedingung ist eine notwendige Voraussetzung, dass V irreduzibel ist. Der alternative Beweis aus Abschnitt 3.1.2 ist jedoch auch in diesem Fall durchführbar, d.h. η existiert, ist aber nicht unimodal. Numerische Betrachtungen analog zu Tabelle 3.1 zeigen, dass die η_n sich mit $\mu \to 1$ stetig diesem Extremfall annähern, d.h. für Werte von μ etwas unterhalb von 1 liegt ein großer Teil der Wahrscheinlichkeitsmasse von η_n auf den geraden Zahlen und nur ein kleiner Teil auf den ungeraden. Die Ergebnisse der Iteration dieser Berechnungen nach Publikation (i), Lemma 8, über viele Generationen deuten stark darauf hin, dass dies nicht nur für die η_n gilt, sondern auch für deren Grenzverteilung η. Für die genaue Schwelle, ab der η nicht mehr unimodal ist, ist nach den oben gemachten Beobachtungen zu erwarten, dass sie nicht nur von μ, sondern auch von N abhängt.

Ein Nachteil der Approximation der invarianten Verteilung mittels der beschriebenen Methode ist, dass sie wegen der erforderlichen Iterationen (bei Berechnung nach Publikation (i), Lemma 8) bzw. Faltungen (bei Berechnung nach Publikation (i), Theorem 13) sehr rechenintensiv ist. Eine Alternative wurde von unserem Kooperationspartner Mikkel Meyer Andersen und Kollegen entwickelt [1]. Sie zeigen, dass η sich sehr gut durch eine diskrete Laplace-Verteilung approximieren lässt. Eine Zufallsvariable D hat eine diskrete Laplace-Verteilung mit Parameter $p \in (0, 1)$ genau dann, wenn

$$P(D = d) = \frac{1 - p}{1 + p} \, p^{|d|} \tag{3.5}$$

für alle $d \in \mathbb{Z}$ [1]. Diese Verteilung lässt sich sehr schnell berechnen und kann daher z.B. für forensische Anwendungen nützlich sein [1]. Warum diese Approximation so präzise ist, bleibt bisher eine offene Frage.

3.2 Modellanpassungen für weitere Y-STRs

In Publikation (ii) haben wir die Anpassung von drei ausgewählten Modellen an die Vater-Sohn-Daten für den Locus DYS19 grafisch dargestellt (Figure 2).

Diese Modelle waren das Lineare Modell ohne Symmetrie, das voll symmetrische Logistische Modell (mit den Parametern α, β und γ), sowie das Logistische Modell mit symmetrischem γ (aber zwischen Auf- und Abwärtsmutationen verschiedenen Werten von α und β). Hier ergänzen wir dazu die entsprechenden Diagramme für die übrigen fünf von uns untersuchten Loci (Abbildungen 3.1–3.5). Zwecks Vergleichbarkeit ist in allen diesen Diagrammen für die Mutationswahrscheinlichkeiten der Bereich von 0 bis 0,01 dargestellt. Dadurch entziehen sich einige besonders große beobachtete Mutationshäufigkeiten der Darstellung, diese werden in den Bildunterschriften gesondert aufgeführt. Die insgesamt drei beobachteten Mutationen mit Betrag 2 wurden für die Modellanpassungen als Einschritt-Mutationen gewertet, wie in Publikation (ii) diskutiert.

Wie man feststellen kann, sind die Ergebnisse qualitativ für alle Loci (ausgenommen DYS392, an dem nur sieben Mutationen beobachtet wurden) sehr ähnlich: Das Lineare Modell ist ungeeignet, weil die Zunahme der Mutationswahrscheinlichkeiten mit steigender Allel-Länge nicht linear ist. Das symmetrische Logistische Modell ist vertretbar, aber das Logistische Modell, in dem nur γ symmetrisch ist, passt in der Regel wesentlich besser. Insbesondere kann das letzgenannte Modell die Tatsache abbilden, dass im Bereich der besonders langen Allele die Abwärtsmutationswahrscheinlichkeit oft noch deutlich größer ist als die Aufwärtsmutationswahrscheinlichkeit.

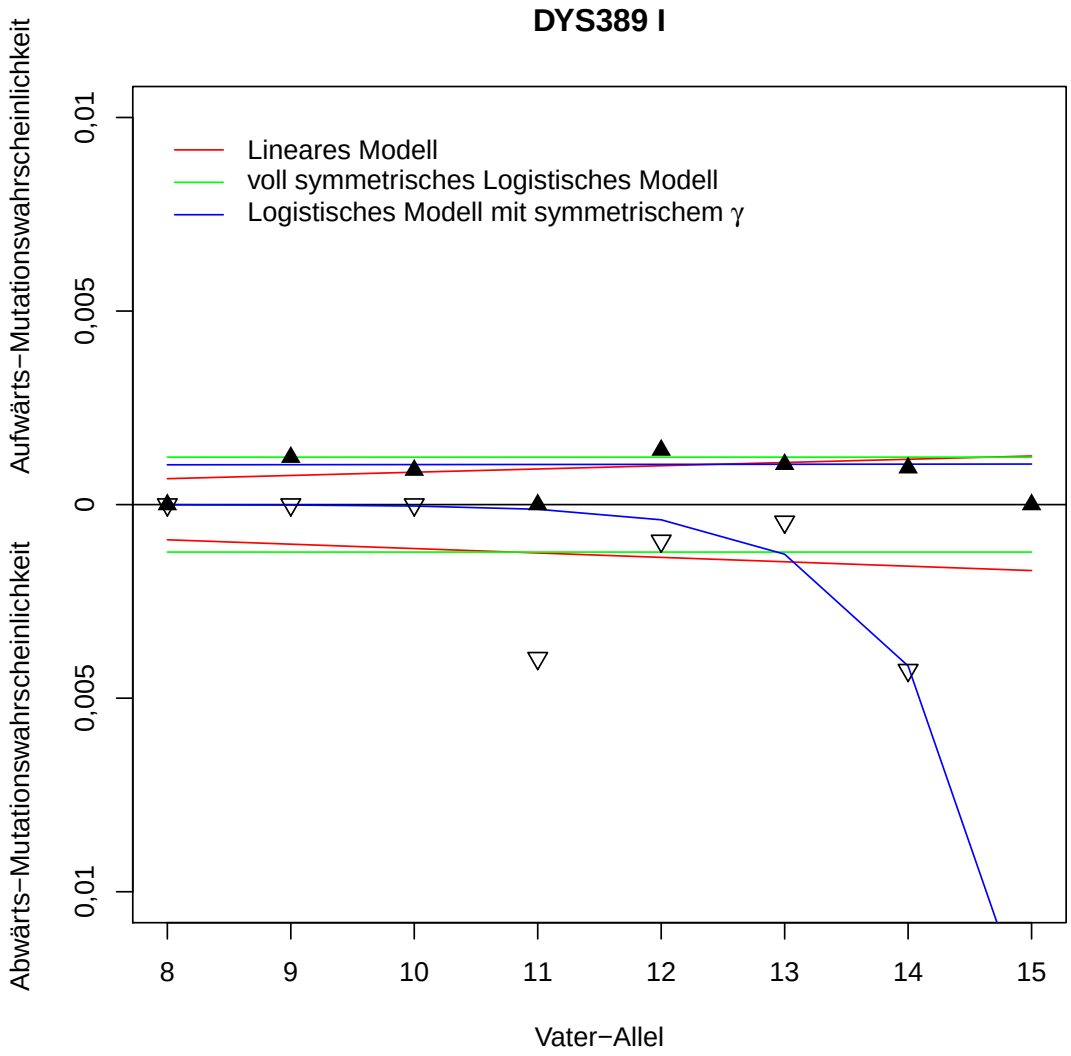

Abbildung 3.1: Mutationswahrscheinlichkeiten unter drei Modellen, angepasst an die Vater-Sohn-Daten (siehe Publikation (ii), Table 3) für DYS389 I. In der oberen Hälfte des Diagramms sind die Aufwärtsmutationswahrscheinlichkeiten dargestellt, während die untere Hälfte die Abwärtsmutationswahrscheinlichkeiten enthält. Dreiecke stehen für die beobachteten relativen Mutationshäufigkeiten (ausgefüllt: aufwärts, leer: abwärts). Die beobachtete Abwärts-Mutationshäufigkeit von Allel 15 liegt mit $3/48 \approx 0,06$ außerhalb des dargestellten Bereichs. Die hier gezeigten Modelle sind das Lineare Modell ohne Symmetrie (rote Linien), das voll symmetrische Logistische Modell (grüne Linien) und das Logistische Modell mit symmetrischem γ, d.h. mit den Parametern α_u, α_d, β_u, β_d und γ (blaue Linien).

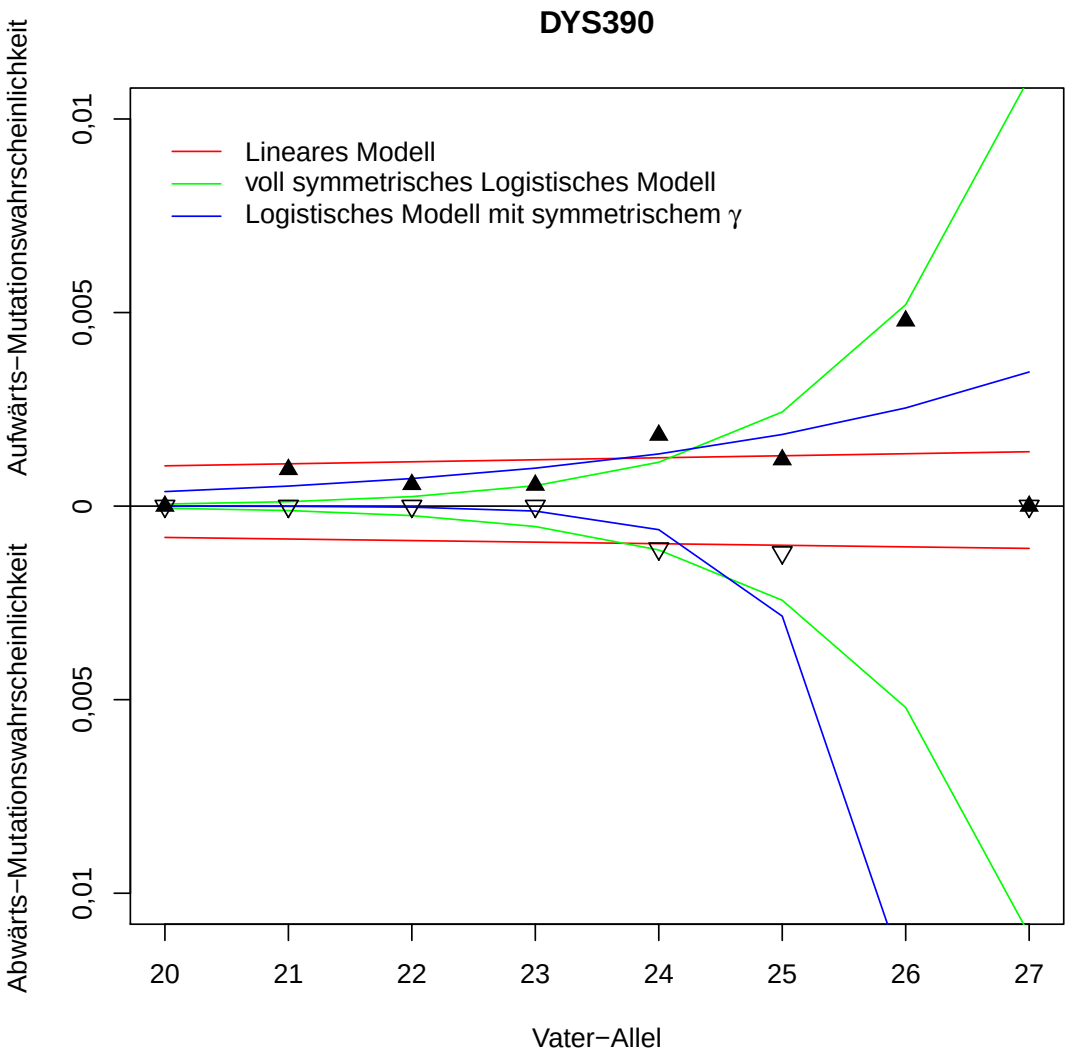

Abbildung 3.2: Mutationswahrscheinlichkeiten unter drei Modellen, angepasst an die Vater-Sohn-Daten (siehe Publikation (ii), Table 3) für DYS390. Symbole und Farben wie in Abbildung 3.1. Die beobachtete Abwärts-Mutationshäufigkeit von Allel 26 liegt mit $5/209 \approx 0,02$ außerhalb des dargestellten Bereichs. Die eine beobachtete +2-Mutation wurde als +1 gewertet.

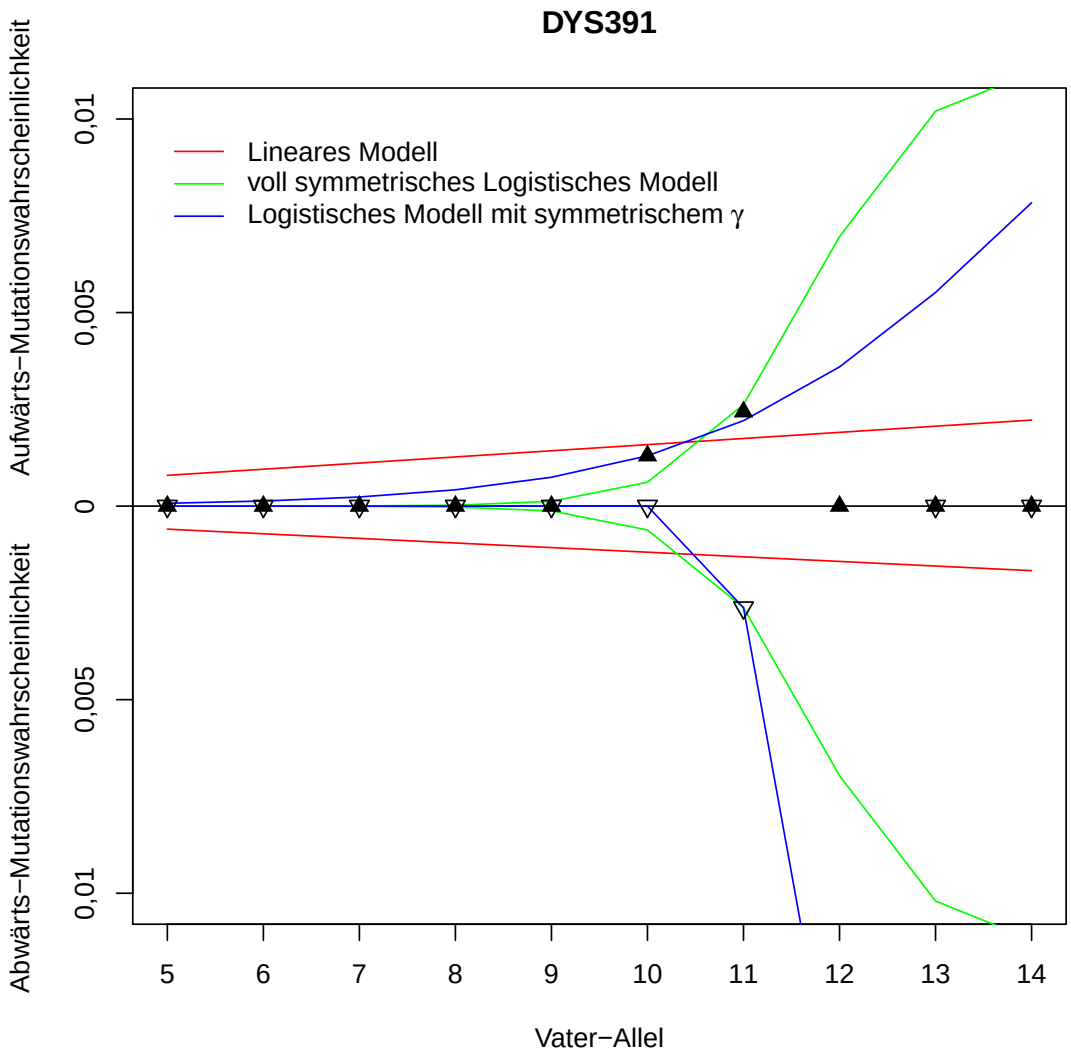

Abbildung 3.3: Mutationswahrscheinlichkeiten unter drei Modellen, angepasst an die Vater-Sohn-Daten (siehe Publikation (ii), Table 3) für DYS391. Symbole und Farben wie in Abbildung 3.1. Die beobachtete Abwärts-Mutationshäufigkeit von Allel 12 liegt mit $4/219 \approx 0,02$ außerhalb des dargestellten Bereichs. Die beiden beobachteten ± 2-Mutationen wurden als ± 1 gewertet.

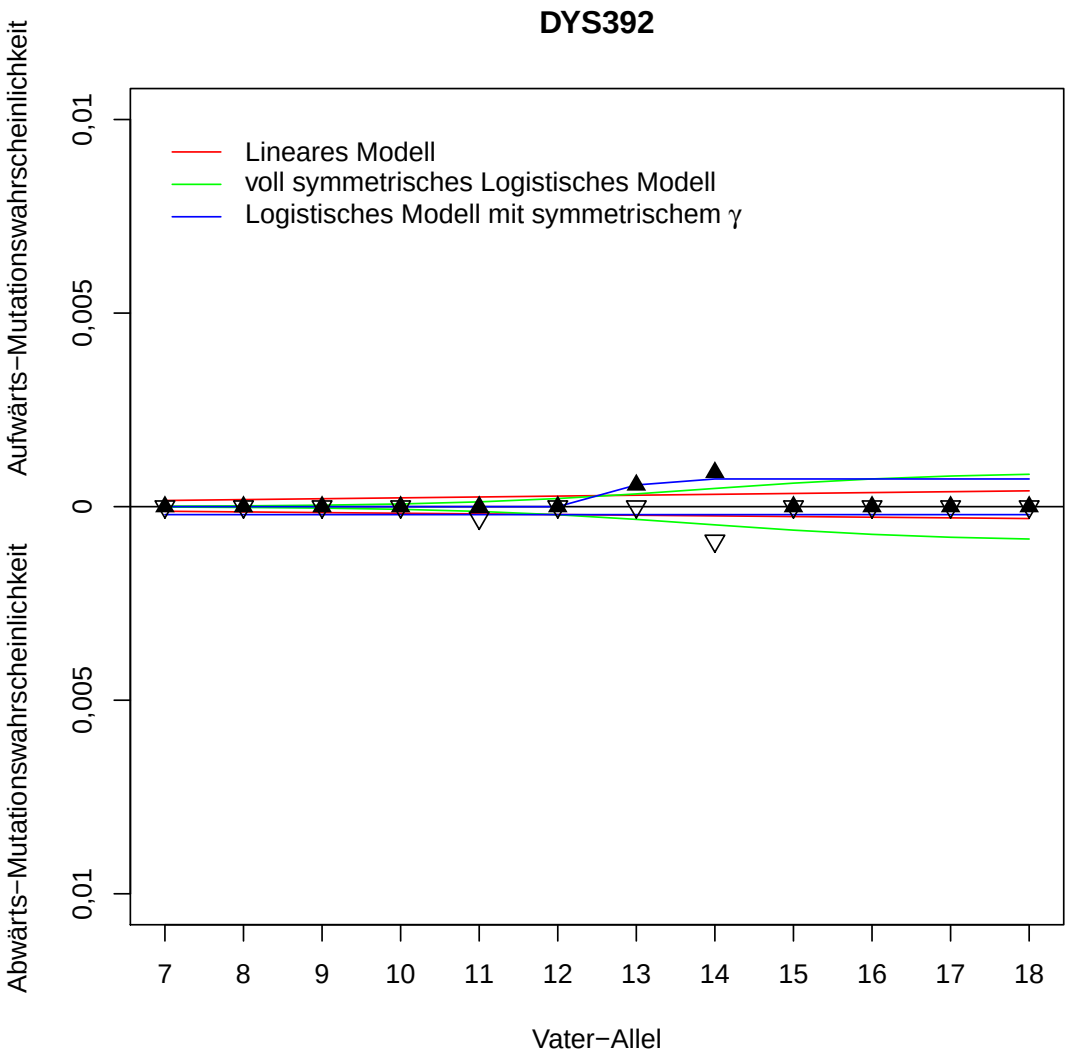

Abbildung 3.4: Mutationswahrscheinlichkeiten unter drei Modellen, angepasst an die Vater-Sohn-Daten (siehe Publikation (ii), Table 3) für DYS392. Symbole und Farben wie in Abbildung 3.1.

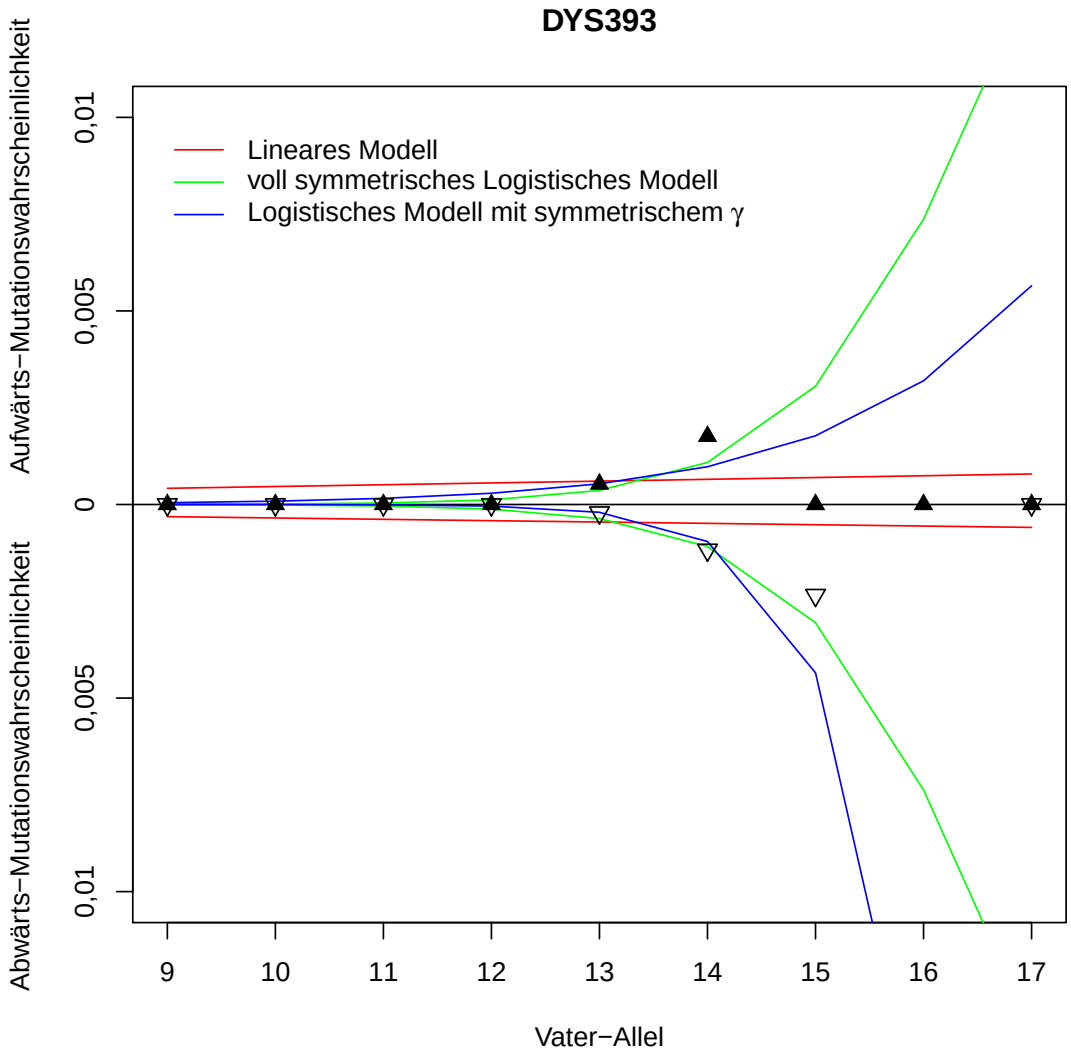

Abbildung 3.5: Mutationswahrscheinlichkeiten unter drei Modellen, angepasst an die Vater-Sohn-Daten (siehe Publikation (ii), Table 3) für DYS393. Symbole und Farben wie in Abbildung 3.1. Die beobachtete Abwärts-Mutationshäufigkeit von Allel 16 liegt mit $1/22 \approx 0,05$ außerhalb des dargestellten Bereichs.

Kapitel 4

Diskussion

4.1 Y-STRs

In den Publikationen (ii) und (iii) haben wir uns mit Y-STRs befasst. Publikation (ii) war die erste Studie überhaupt, in der verschiedene STR-Mutationsmodelle systematisch Locus für Locus miteinander verglichen wurden. Mit Publikation (iii) liefern wir einen Beitrag zur Lösung des in Abschnitt 1.6.6 eingeführten Problems der forensischen Evidenz-Quantifizierung für Y-STR-Haplotypen. Die wichtigsten Aspekte wurden schon innerhalb dieser beiden Publikationen diskutiert. Hier sollen daraus einige Punkte herausgegriffen und ausführlicher behandelt werden.

4.1.1 Vater-Sohn-Daten

In Publikation (ii), Table 2, haben wir alle 24 uns bekannten Veröffentlichungen mit Y-STR-Daten von Vater-Sohn-Paaren aufgelistet, die bis Mitte 2009 erschienen waren. Seitdem sind weitere derartige Studien erschienen, z.B. [8], [235] und [151]. Wenn sich genug Daten akkumuliert haben, planen wir ein Update von Publikation (ii), nach Möglichkeit ausgeweitet auf zusätzliche Loci und Einflussgrößen (siehe unten). Die Vorteile dieses Datensatzes wurden in Publikation (ii) bereits ausführlich diskutiert. Hier wollen wir einige Gedanken zu den Nachteilen bzw. Problemen ergänzen, auch im Hinblick auf zukünftige Arbeiten.

Je mehr Studien akkumuliert werden, desto schwieriger ist es sicherzustellen, dass es zwischen ihren Stichproben keine Überschneidungen gibt. Falls solche Überschneidungen unbemerkt auftreten, so ist die tatsächliche Gesamtfallzahl kleiner als angenommen. Das führt dann zu deflationierten Schätzungen der Standardfehler, etwa bei den Modellparametern. Ein Bias in unseren primären Schätzungen der Modellparameter wäre infolgedessen aber nicht zu befürchten, es sei denn, es gibt einen Zusammenhang zwischen dem Mutationswert und der unbemerkten Verdoppelung von Vater-Sohn-Paaren. Jedenfalls ist es gute wissenschaftliche Praxis, in jeder Publikation derartiger Daten klar zu kennzeichnen, ob

ein Teil der Daten (und wenn ja, welcher) bereits vorher veröffentlicht wurde.

Eine weitere Information, die leider manchmal fehlt, ist die genaue Angabe der Allele aller Vater-Sohn-Paare, in denen keine Mutation beobachtet wurde. Viele der von uns genutzten Studien befassen sich in erster Linie mit der Haplotypverteilung in einer bestimmten Population. Die Angabe von beobachteten Mutationen ist dann nur ein zusätzlicher Aspekt, und es ist wohl nicht offensichtlich, dass auch nicht mutierte Allele eine Information darstellen (*Stasis is data!* [82]). Konkret werden diese Informationen im von uns verwendeten Ansatz benötigt, um die Likelihood (eines Parametervektors) zu berechnen, denn diese ist das Produkt über alle Vater-Sohn-Paare, auch solche ohne Mutation (siehe Publikation (ii), Gleichung (1)). Die ISFG erkannte das Problem und empfahl in ihren Richtlinien 2006, immer alle Allele anzugeben [88]. Zwar halten sich immer noch nicht alle Autoren daran, aber es gab seitdem eine deutliche Verbesserung.

4.1.2 Mutieren Y-STRs wie autosomale STRs?

In Publikation (ii) haben wir ausschließlich MSY-Loci verwendet und konnten somit die Mutationen direkt beobachten, ohne potentielle Rekombinationsereignisse berücksichtigen zu müssen. Wir argumentieren dort, dass unsere qualitativen Ergebnisse dennoch auch auf autosomale Loci übertragbar sind. Dieses Argument soll hier noch etwas untermauert werden. Die entscheidenden Fragen lauten:

(i) Spielt Rekombination bei der Mutation von autosomalen STRs eine wesentliche Rolle?

(ii) Spielen Y-Chromosom-spezifische Mutationsmechanismen bei der Mutation von Y-STRs eine wesentliche Rolle?

Frage (i) wurde bereits in der Einleitung mit nein beantwortet. Wie dort in Abschnitt 1.4 erläutert wurde, gibt es viele Hinweise darauf, dass *Slipped-Strand Mispairing* (SSM) während der DNA-Replikation für STR-Mutationen verantwortlich ist. Einer dieser Hinweise, der hier noch nicht explizit als solcher erwähnt wurde, ist die Beobachtung, dass es bei der PCR-Amplifikation von STRs oft zur Bildung sogenannter Stotter-Fragmente kommt (siehe Abschnitt 1.6.1). Diese lassen sich als Resultat von SSM-Ereignissen in vitro, während der Replikation im Rahmen der PCR, erklären [61, Figure 2b]. Dies ist, wie wir in Publikation (ii) betonen, einer der wesentlichen Unterschiede zwischen Mikro- und Minisatelliten, denn bei letzteren spielt die Rekombination in der Tat eine wesentliche Rolle im Mutationsprozess [21].

Frage (ii) ist weniger gut untersucht, aber auch hier deutet bisher alles auf eine negative Antwort hin. Parallel zur Veröffentlichung der ersten MSY-Sequenz [208] erschien ein Artikel [199] derselben Arbeitsgruppe, in dem anhand eines Vergleiches mit den entsprechenden Sequenzabschnitten anderer Hominidae gezeigt wurde, dass für die Evolution der Palindrom-Bereiche die Genkonversion eine

wichtige Rolle spielt (siehe Abschnitt 1.6.3). Einige forensisch genutzte Y-STRs liegen in Palindromen, wie z.B. DYS385 a/b, und für diese ist nicht auszuschließen, dass die Genkonversion für ihr Mutationsverhalten wesentlich ist. Aber für die weitaus größere Gruppe aller Y-STRs, die nicht in Palindromen liegen (inclusive der sechs Loci, auf die wir uns in Publikation (ii) konzentriert haben), gibt es keine Hinweise darauf, dass Genkonversion zu ihrem Mutationsverhalten beiträgt [111, S. 603].

4.2 Ausblick

Wie in Abschnitt 1.4.2 dargelegt wurde, herrscht kein Mangel an Vorschlägen von Modellen, die den STR-Mutationsprozess beschreiben sollen (siehe auch [27]). Das in Publikation (i) zugrundegelegte SMM ist wie dort erwähnt in seiner Einfachheit natürlich nur eine erste Annäherung an die realen Mechanismen der STR-Mutation. Es stellt sich daher die Frage, inwieweit unsere Erkenntnisse über das qualitative Verhalten des Allelprozesses auch für komplexere Modelle Gültigkeit besitzen. In Publikation (ii) haben wir zwar komplexere Modelle verwendet, aber aufgrund der Natur der Daten waren einige vielversprechende Modelle nicht anwendbar. In diesem Abschnitt greifen wir einige dieser Ansätze auf und geben Hinweise für zukünftige Forschung.

4.2.1 Weitere Mutationsmodelle

Endlicher Zustandsraum

In Publikation (i) sind wir von einem SMM mit Zustandsraum $\mathbb{Z}$ ausgegangen. Offensichtlich sind aber negative Allel-Längen in der Realität unmöglich. Auch gibt es praktische Grenzen für die maximale Länge eines STRs. Deshalb liegt der Wunsch nahe, ein realistischeres Modell zu verwenden, in dem der Zustandsraum von X, also die Menge der möglichen Allele (evtl. verschoben um eine Konstante m), beschränkt ist. Für ein solches Modell stellt sich dann die Frage, welche Übergangswahrscheinlichkeiten in den Grenzzuständen gelten sollen. Eine Möglichkeit sind reflektierende Grenzen, d.h. für das längstmögliche Allel ist die Aufwärts-Mutationswahrscheinlichkeit 0 und dafür die Abwärts-Mutationswahrscheinlichkeit entsprechend größer, während für das kleinstmögliche Allel die Abwärts-Mutationswahrscheinlichkeit 0 und dafür die Aufwärts-Mutationswahrscheinlichkeit entsprechend größer ist. Wir hätten dann also eine Markov-Kette mit endlichem Zustandsraum. Da solche Markov-Ketten unter den gegebenen Bedingungen immer gegen eine stationäre Verteilung konvergieren, gilt dies sogar für den nicht normierten Allelprozess X. Dennoch ist selbst für solche Modelle ein Unterschied zwischen globalem und lokalem Verhalten möglich, der sich z.B. in unterschiedlichen Konvergenzgeschwindigkeiten oder unterschiedlichen stationären Verteilungen manifestieren kann.

Berücksichtigung von Punktmutationen

In [143] wurde ein Modell entwickelt, das das SMM um die zusätzliche Berücksichtigung von Punktmutationen ergänzt. Es wurde gezeigt, dass dann der Allelprozess ohne Normierung konvergiert [53], was dieses Modell sehr interessant macht. Aber beim Einsatz der in der Forensik verwendeten Multiplex-Kits (siehe Abschnitt 1.6.1) wirken sich Punktmutationen innerhalb des STRs in der Regel gar nicht auf das Ergebnis aus [88]. Insofern ist dieses Modell auf unsere typischen Vater-Sohn-Daten leider nicht anwendbar.

Außerdem ist zu bedenken, dass die Punktmutations-Wahrscheinlichkeit im Vergleich zur STR-Mutationswahrscheinlichkeit sehr gering ist. Genauer gesagt wird die Punktmutations-Wahrscheinlichkeit beim Menschen anhand von Eltern-Kind-Sequenzvergleichen im genomweiten Durchschnitt auf etwa $1,2 \times 10^{-8}$ pro Basenpaar und Generation geschätzt [205]. Demnach beträgt die Wahrscheinlichkeit für (mindestens) eine Punktmutation innerhalb von z.B. DYS19, dessen häufigstes Allel eine Länge von $14 \times 4 = 56$ Basenpaaren hat (Tabelle 1.3), $1-(1-1,2\times 10^{-8})^{56} \approx 56 \times 1,2 \times 10^{-8} \approx 6,7 \times 10^{-7}$ pro Generation. Dies steht im Kontrast zur STR-Mutationswahrscheinlichkeit von etwa $2,3 \times 10^{-3}$ an diesem Locus (Tabelle 1.3). Die im vorherigen Absatz erwähnte Konvergenz ist deshalb sehr langsam und für die meisten praktischen Anwendungen nicht relevant.

Mehrschritt-Mutationen

Einschritt-Mutationsmodelle sind zwar eine gute Approximation, aber es ist schon länger bekannt, dass an STR-Loci gelegentlich auch Mehrschritt-Mutationen vorkommen [97]. So haben wir in unseren akkumulierten Daten unter 163 Mutationen drei Doppelschritt-Mutationen (und keine größeren) beobachtet (Publikation (ii), Table 3). Solange die $Z_n(i)$ (mit $n \in \mathbb{N}$ und $i \in \{1, 2, \dots N\}$) unabhängig sind, was biologisch plausibel ist, bleibt unser zentrales Resultat (Theorem 11) aus Publikation (i) auch für eine Verallgemeinerung des SMM auf Mehrschritt-Mutationen gültig. Wenn die zu Z gehörige Irrfahrt nullrekurrent ist, gilt auch weiterhin Theorem 5.

In Publikation (ii) haben wir die beobachteten Mutationen in Auf- oder Abwärtsmutationen dichotomisiert und nur Einschritt-Modelle ausgewertet. Es wird in der Praxis sehr schwer werden, mit einem vergleichbaren Ansatz genug Mehrschritt-Mutationen zu beobachten, um auch Mehrschritt-Modelle sinnvoll anpassen zu können.

4.2.2 Migrationsmodelle

In der Populationsgenetik hat die Modellierung von Migration bereits Tradition, und es existiert eine Vielzahl verschiedener Modelle. Für entsprechende Übersichten siehe [17], [251, Chapter 12], [209], [110, Section 5.5], [95, Chap-

ter 9], [94, Section 6.5] und [33, Section 4.1]. Genauer gesagt geht es um *Gene Flow*, denn Migration kann natürlich auch ohne genetische Auswirkungen stattfinden [64]. Einige klassische Beispiele sind das allgemeine *Migration Matrix Model* [17], das *Continent-Island Model* und das *Island Model*, beide von Sewall Wright [247, 248, 250], ebenso wie das *Neighborhood Model*, auch bekannt als *Isolation by Distance* [248, 249, 250], und schließlich das *Stepping-Stone Model* von Motoo Kimura [126, 129] bzw. Gustave Malécot [161, 162]. Es ist erstrebenswert, derartige Ansätze in die Modellierung von STR-Markern in Zukunft mit einzubinden.

4.2.3 Alter des Vaters

Bei Punktmutationen hat das Alter des Vaters einen großen Einfluss auf die Mutationswahrscheinlichkeit [138]. Das ist biologisch plausibel, denn je älter der Vater ist, desto mehr Zellteilungen haben seine Gameten durchlaufen. Es ist also naheliegend, dass auch für STR-Mutationen das Alter des Vaters von Bedeutung ist. Uns liegen bisher noch nicht ausreichend geeignete Daten vor, um dies näher zu untersuchen, denn in den forensischen Publikationen von Vater-Sohn-Daten wird leider selten das Alter der Väter angegeben. Diese Situation beginnt sich aber zu verbessern, seit die ISFG eine entsprechende Empfehlung ausgesprochen hat [88].

4.3 Haben STRs eine Funktion?

Ob STRs auch in nichtcodierenden Bereichen des Genoms eine Funktion im klassischen evolutionsbiologischen Sinne besitzen, d.h. ob ihre Anwesenheit bzw. Abwesenheit, Anzahl und individuelle Länge die Fitness des Organismus beeinflussen, ist noch nicht abschließend geklärt [157]. Oft wird aber bei der Anwendung von STRs als Marker ihre Neutralität vorausgesetzt. So auch in der Forensik, denn z.B. in Deutschland dürfen nach einem Urteil des Bundesgerichtshofs [13] nur neutrale Marker für Identifikationszwecke verwendet werden [142, S. 206]. Auch in den USA war das ein wichtiges Kriterium bei der Wahl der STR-Loci, die in der FBI-Datenbank CODIS (siehe Abschnitt 1.6.2) erfasst werden:

> „the DNA profiles retained in the system [CODIS] are sanitized 'genetic fingerprints' that can be used to identify an individual uniquely, but do not disclose an individual's traits, disorders, or dispositions. The rules governing the operation of CODIS reflect its function as a tool for law enforcement identification, and do not allow DNA information within the scope of the system to be used to derive information concerning sensitive genetic matters.“ [44]

Nichtcodierende STRs wären demnach ein Teil der sogenannten *Junk-DNA*, über deren Anteil am Genom, insbesondere dem des Menschen, es seit der Einführung

des Konzepts 1972 [36, S. 313–323], siehe auch [181, 182], eine anhaltende Debatte gibt.

Dass diese Debatte keineswegs beendet ist, zeigte sich zuletzt seit September 2012, als eine Reihe von 30 koordinierten Artikeln im Rahmen der *Encyclopedia of DNA Elements* (ENCODE) erschien. Die zentrale Publikation enthielt die zusammenfassende Behauptung, zu 80% der Humangenomsequenz ließen sich „biochemische Funktionen“ zuordnen [63]. In derselben *Nature*-Ausgabe wurde in einem journalistischen Beitrag daraus „some sort of function“[160], während einige der beteiligten Wissenschaftler in der „News & Views“-Sektion [55] explizit die Schlussfolgerung zogen, dass der Großteil des menschlichen Genoms kein „Junk“ sein kann:

> „One of the more remarkable findings described in the consortium's 'entrée' paper [63] is that 80% of the genome contains elements linked to biochemical functions, dispatching the widely held view that the human genome is mostly 'junk DNA'.“ (Joseph R. Ecker)
>
> „The vast majority of the human genome does not code for proteins and, until now, did not seem to contain defined gene-regulatory elements. Why evolution would maintain large amounts of 'useless' DNA had remained a mystery, and seemed wasteful. It turns out, however, that there are good reasons to keep this DNA.“ (Inês Barroso)

Zugleich war z.B. in einer Pressemitteilung der NIH [164] von mehr als 80% des Humangenoms mit „spezifischer biologischer Funktion“ die Rede, während im Magazinteil von *Science* [190] und auch in diversen Massenmedien [20, 137, 108, 16, 230, und viele weitere] gleich kategorisch der „Tod der Junk-DNA“ ausgerufen wurde. Dies war keineswegs nur ein Problem der Kommunikation zwischen Wissenschaftlern und Journalisten. Selbst zwei leitende Wissenschaftler äußerten sich ähnlich:

> Die BBC zitierte Ewan Birney, den Leiter dieser Phase des ENCODE-Projekts, mit den Worten: „The term junk DNA must now be junked.“ [227]
>
> In einer NPR-Sendung [212] sagte John Stamatoyannopoulos, einer der ENCODE-Leiter [63]: „Most of the human genome is out there to control the genes, in other words it sort of [is] the programming of the genome or sort of its operating system if you will.“

Die Zahl von 80% [63] beruhte jedoch auf einer höchst eigenartigen und irreführenden Definition von Funktion [56, 179, 84, 51]. Insbesondere wurde ein Genomabschnitt als funktionell klassifiziert, wenn in irgendeiner Zelle, zu irgendeinem Zeitpunkt ein Transkriptionsfaktor auch nur ein einziges Mal daran gebunden beobachtet wurde [63]. Diese Vorgehensweise ist indes wenig erkenntnisfördernd, denn

es ist zu erwarten, dass auch zufällige, bedeutungslose Sequenzen gelegentlich an einen Transkriptionsfaktor anbinden [57].

Warum also wurde diese Definition von Funktion überhaupt verwendet? Ewan Birney schrieb in einem Blogeintrag [15], dass nach einer klassischen Definition der von ENCODE definitiv als funktionell erkannte Anteil des Genoms 9% [!] wäre:

> „However, on the other end of the scale — using very strict, classical definitions of 'functional' like bound motifs and DNaseI footprints; places where we are very confident that there is a specific DNA:protein contact, such as a transcription factor binding site to the actual bases — we see a cumulative occupation of 8% of the genome. With the exons (which most people would always classify as 'functional' by intuition) that number goes up to 9%."

Er räumte ein, dass er und seine Koautoren die wenig belastbare Zahl von 80% gezielt herausgestellt hatten, um die Medienberichterstattung über die ENCODE-Resultate zu stimulieren [15]:

> „We use the bigger number because it brings home the impact of this work to a much wider audience."

Ohne hier auf die ethischen Implikationen dieser Strategie eingehen zu wollen, wenden wir uns der Frage zu, wie groß der Anteil funktioneller Sequenzen (nach einer sinnvollen Definition, siehe dazu z.B. [51]) am menschlichen Genom wirklich ist. Die Junk-DNA-Hypothese wird im Rahmen von Widerlegungsversuchen oft dahingehend charakterisiert, dass sie auf einer naiven A-priori-Annahme beruhe. Demnach werde Funktionslosigkeit einfach postuliert, wenn eine spezifische Funktion (noch) nicht bekannt ist. Aber diese Charakterisierung ist falsch, denn es gab 1972 — und es gibt heute — eine Reihe guter Gründe für die Hypothese, dass ein großer Teil des Genoms von Säugetieren „Junk" ist:

(i) Schon lange bevor das erste Genom sequenziert wurde, war aufgrund von Gewichtmessungen bekannt, dass es große inter-Spezies Unterschiede in der Genomgröße gibt [166]. Die Genomgröße korreliert dabei nur schwach mit dem menschlichen Eindruck der Komplexität des jeweiligen Organismus [36, S. 313-314], [181, S. 366]. Selbst die Genome nahe verwandter Wirbeltier-Arten können sich um eine Größenordnung unterscheiden, wie das klassische Beispiel der Salamander zeigt (aktuelle Rekordhalter [86]: *Gyrinophilus porphyriticus* mit 10,12 pg [80] versus *Necturus lewisi* mit 120,60 pg [185]).

(ii) Ebenfalls schon lange bekannt ist die Beobachtung, dass ein vollständig funktionelles Genom in Verbindung mit der geschätzten Mutationsrate beim Menschen zu einer untragbaren Akkumulation von schädlichen Mutationen

führen würde [41]. Insbesondere wurde bereits Mitte des vorigen Jahrhunderts anhand von Überlegungen zur Häufigkeit lethaler Mutationen die Zahl der menschlichen Gene bemerkenswert gut abgeschätzt [211], siehe auch [213, S. 29]. 1950 prägte Hermann Muller den Begriff „genetische Bürde“ (*genetic load*) für die relative (im Vergleich zu einer Referenzpopulation ohne den jeweiligen Faktor) Auswirkung eines bestimmten Faktors (z.B. Mutation, Migration usw.) auf die durchschnittliche Fitness einer Population [172]. Auch schon Haldanes erstmals 1937 veröffentlichtes Konzept der „Kosten der natürlichen Selektion“ [89, 90] kann als eine Art von genetischer Bürde angesehen werden [42, S. 297]. Ende der 1960-er Jahre war die mit einem vollständig funktionellen menschlichen Genom einhergehende untragbar hohe Mutationsbürde dann ein entscheidender Grund für die Formulierung der Neutralitätstheorie der molekularen Evolution [127, 130].

(iii) Das gleichmäßige Ticken der molekularen Uhr [144] ist ein deutlicher Hinweis auf die Neutralität vieler Sequenzen. Insbesondere hat sich herausgestellt, dass auch STRs als molekulare Uhren verwendet werden können [217].

(iv) Die evolutionäre Geschichte und daraus resultierende Funktionslosigkeit bestimmter Klassen von DNA ist gut bekannt. Darunter fallen z.B. Pseudogene [222]. Wie erwartet zeigt nur ein geringer Anteil aller Pseudogene im Humangenom Evidenz für Funktion [122].

(v) Die Frage, welcher Anteil des Genoms überhaupt jemals transkribiert wird ist zwar umstritten [224, 34, 49], aber für die Junk-DNA-Debatte nicht entscheidend. Denn gelegentliche Transkription impliziert wie schon erwähnt nicht Funktion [215, 252]. Auch das ist im Prinzip schon lange bekannt [36, S. 317].

(vi) Insbesondere für STRs gilt, dass krankheitsassoziierte Loci im menschlichen Genom in der Regel in proteincodierenden Abschnitten, RNA-Genen oder in regulatorischen Regionen liegen. Siehe dazu die Beispiele in Tabelle 1.2, die sich alle jeweils einem Gen zuordnen lassen.

(vii) Experimentelle Studien (z.B. [175]) haben gezeigt, dass sich bei Säugetieren der Verlust großer nichtcodierender Genom-Abschnitte nicht erkennbar auf den Phänotyp auswirken muss.

(viii) Der Vergleich individueller Humangenome zeigt, dass auch in unserer Spezies große (mehrere 100 kb) Genomregionen existieren, die offenbar unnötig sind, da manche gesunde Menschen sie gar nicht besitzen [221].

In der Berichterstattung über die ENCODE-Resultate wurden diese Gründe oft ignoriert, und stattdessen die Geschichte der Junk-DNA-Debatte karikiert, auch

von Wissenschaftlern. So sagte Eric Green, Direktor des National Human Genome Research Institute (NHGRI, eines der National Institutes of Health, USA):

> „During the early debates about the Human Genome Project, researchers had predicted that only a few percent of the human genome sequence encoded proteins, the workhorses of the cell, and that the rest was junk. We now know that this conclusion was wrong. ENCODE has revealed that most of the human genome is involved in the complex molecular choreography required for converting genetic information into living cells and organisms.“ [164]

Beide Teile dieser Aussage sind falsch: Dass ENCODE nichts dergleichen gezeigt hat, wurde oben bereits festgestellt. Außerdem impliziert Green hier, dass alle nichtcodierende DNA als Junk angesehen wurde. Das war aber niemals der Fall. So bezog sich z.B. Ohno [181, 182] mit dem Begriff „Junk“ im Titel dieser Arbeiten ausschließlich auf funktionslose Pseudogene, wie aus seiner Diskussion ersichtlich ist. Zudem nahm er regulatorische Sequenzen, die ja Konservierung zeigen, explizit aus [182, S. 172]. Auch die Existenz von Genen für nicht-codierende RNA mit direkter Funktion, z.B. Transfer-RNA, war schon lange vor Beginn des Humangenomprojekts bekannt [114, S. 34-44].

Vielleicht rührt die Verwirrung daher, dass viele Autoren offenbar nicht mit der Neutralitätstheorie der molekularen Evolution [127, 130, 128] vertraut sind [158]. Auch einige Befürworter der *Junk*-Theorie haben zu dieser Verwirrung beigetragen. So schreibt Comings [36, S. 316]:

> „These considerations suggest that up to 20% of the genome is actively used and the remaining 80+% is junk. But being junk doesn't mean it is entirely useless. Common sense suggests that anything that is completely useless would be discarded. There are several possible functions of junk DNA.“

Eine solche Definition von Junk-DNA steht wie oben erwähnt im Widerspruch zu Ohnos Verwendung des Begriffes. Darüber hinaus ist die zitierte „Common sense“-Annahme für Menschen in Anbetracht unserer langfristig geringen effektiven Populationsgröße äußerst fragwürdig [83, 128, 158].

In den USA führte die Fehlinterpretation der ENCODE-Daten dazu, dass die Rechtmäßigkeit von CODIS in Frage gestellt wurde [23, 117]. Deshalb recherchierten Katsanis und Wagner [117], was über eventuelle Funktionen der 13 Kernloci und 11 weiterer STRs, die für die Erweiterung von CODIS vorgeschlagen sind [92], bekannt ist. Sie stellten fest, das einiger dieser STRs etwas mit Genregulation zu tun haben könnten, aber einzelne Allele bzw. Genotypen für diese Loci zur Zeit keinerlei Vorhersagen über Phänotypen ermöglichen [117].

Einige Beispiele für phänotypassoziierte STRs haben wir bereits in Abschnitt 1.2.2, insbesondere Tabelle 1.2 gegeben. Wie erwartet liegen die meisten, aber

nicht alle dieser krankheitsassoziierten STRs in codierenden Sequenzen. Ein Beispiel für einen codierenden STR, der früher forensisch genutzt wurde, ist HumARA [26, Box 8.1]. Ebenso ist zu erwarten, dass gelegentlich auch nichtcodierende STRs funktionelle Charakteristika zeigen, etwa wenn sie in einem regulatorischen Abschnitt liegen. Und auch das ist der Fall [75].

Selbst wenn ein Locus funktionslos ist, kann er natürlich trotzdem mit bestimmten Phänotypen korreliert sein. Eine mögliche Ursache dafür ist Confounding durch Populationsstratifikation. Ein anderer möglicher Grund ist die Existenz eines entsprechend kausalen Locus, der im Kopplungsungleichgewicht mit dem an sich funktionslosen Locus steht. Einige an sich neutrale STRs liegen sogar so dicht an einem Gen (bzw. die Rekombinationsrate zwischen STR und Gen ist so gering), dass sie nahezu perfekt mit dem Gen gekoppelt sind und sich somit gut als Marker für das Gen verwenden lassen. Historisch gesehen war diese Anwendung übrigens die treibende Kraft hinter der Entwicklung des genetischen Fingerabdrucks (zunächst mit Mini- anstelle von Microsatelliten, siehe Abschnitt 1.6.1). Dies wird z.B. im Begleitartikel von Thomas Caskey [29] zur ersten Veröffentlichung der Technik von Alec Jeffreys und Kollegen [106] deutlich. Caskey schreibt darin über die Kartierung von Genen, insbesondere in Bezug auf genetische Erkrankungen, ohne das forensische Potential der neuen Technik überhaupt zu erwähnen.

Zusammenfassend kann man sagen, dass die Hypothese, ein großer Teil des menschlichen Genoms habe keine Funktion, durch viele verschiedene Beobachtungen gestützt wird. Wie groß der Anteil funktioneller Sequenzen genau ist, wird die Zukunft zeigen, aber er liegt vermutlich eher bei 10%–20% als bei 80%. Aufgrund der Beobachtung, dass STRs gleichmäßig über das Genom verteilt sind, ist diese Zahl zugleich eine obere Grenze für den Anteil der STRs, die eine direkte kausale Assoziation mit einem Phänotyp aufweisen. Tatsächlich wird dieser Anteil noch geringer sein, denn in Anbetracht der hohen STR-Mutationsraten ist zu erwarten, dass die in funktionellen Bereichen des Genoms allgegenwärtige negative Selektion tendenziell gegen STRs wirkt.

Literaturverzeichnis

[1] M. M. Andersen, P. S. Eriksen und N. Morling. The discrete Laplace exponential family and estimation of Y-STR haplotype frequencies. *J. Theor. Biol.* 329: 39–51 (2013).

[2] M. A. Andrade und P. Bork. HEAT repeats in the Huntington's disease protein. *Nat. Genet.* 11(2): 115–116 (1995).

[3] S. E. Andrew, Y. P. Goldberg, B. Kremer, H. Telenius, J. Theilmann, S. Adam et al. The relationship between trinucleotide (CAG) repeat length and clinical features of Huntington's disease. *Nat. Genet.* 4(4): 398–403 (1993).

[4] C. T. Ashley und S. T. Warren. Trinucleotide repeat expansion and human disease. *Annu. Rev. Genet.* 29(1): 703–728 (1995).

[5] J. C. Avise. *Inside the Human Genome: A Case for Non-Intelligent Design.* Oxford University Press (2010).

[6] D. Bachtrog. Y-chromosome evolution: Emerging insights into processes of Y-chromosome degeneration. *Nat. Rev. Genet.* 14(2): 113–124 (2013).

[7] K. N. Ballantyne. Rapidly mutating Y-chromosomal STRs — a multi-center assessment of global male lineage and relative differentiation. Vortrag, 25th World Congress of the International Society for Forensic Genetics, Melbourne (5.9.2013).

[8] K. N. Ballantyne, M. Goedbloed, R. Fang, O. Schaap, O. Lao, A. Wollstein et al. Mutability of Y-chromosomal microsatellites: Rates, characteristics, molecular bases, and forensic implications. *Am. J. Hum. Genet.* 87(3): 341–353 (2010).

[9] K. N. Ballantyne, V. Keerl, A. Wollstein, Y. Choi, S. B. Zuniga, A. Ralf et al. A new future of forensic Y-chromosome analysis: Rapidly mutating Y-STRs for differentiating male relatives and paternal lineages. *Forensic Sci. Int. Genet.* 6(2): 208–218 (2012).

[10] G. P. Bates. The molecular genetics of Huntington disease — a history. *Nat. Rev. Genet.* 6(10): 766–773 (2005).

[11] E. Battaglia. The chromosome satellite [Navashin's sputnik or satelles]: A terminological comment. *Acta Biologica Cracoviensia, Series Botanica* 41: 15–18 (1999).

[12] H. Bauer. *Wahrscheinlichkeitstheorie*. W. de Gruyter, 4. Auflage (1991).

[13] BGH. Zulässigkeit der Gentechnik zur Überführung eines Täters. Urteil, Aktenzeichen 5 StR 145/90 (21.8.1990).

[14] P. R. Billings. *DNA on Trial: Genetic Identification and Criminal Justice*. Cold Spring Harbor Laboratory Press (1992).

[15] E. Birney. ENCODE: My own thoughts. Blog-Eintrag, URL: http://genomeinformatician.blogspot.de/2012/09/encode-my-own-thoughts.html (5.9.2012).

[16] K. Blawat. Menschliche DNA: Erbgut-Müll hat doch einen Nutzen. *Sueddeutsche.de* (6.9.2012).

[17] W. F. Bodmer und L. L. Cavalli-Sforza. A migration matrix model for the study of random genetic drift. *Genetics* 59(4): 565–592 (1968).

[18] C. H. Brenner. Fundamental problem of forensic mathematics — the evidential value of a rare haplotype. *Forensic Sci. Int. Genet.* 4(5): 281–291 (2010).

[19] A. H. Brown, D. R. Marshall und L. Albrecht. Profiles of electrophoretic alleles in natural populations. *Genetical Research* 25(2): 137–143 (1975).

[20] D. Brown und H. Boytchev. 'Junk DNA' concept debunked by new analysis of human genome. *The Washington Post* (5.9.2012).

[21] J. Buard, A. Shone und A. Jeffreys. Meiotic recombination and flanking marker exchange at the highly unstable human minisatellite CEB1 (D2S90). *Am. J. Hum. Genet.* 67(2): 333–344 (2000).

[22] J. Buckleton, M. Krawczak und B. Weir. The interpretation of lineage markers in forensic DNA testing. *Forensic Sci. Int. Genet.* 5(2): 78–83 (2011).

[23] B. Budowle. ENCODE and its first impractical application. *Investigative Genetics* 4(1): 4 (2013).

[24] J. M. Butler. Genetics and genomics of core short tandem repeat loci used in human identity testing. *J. Forensic Sci.* 51(2): 253–265 (2006).

[25] J. M. Butler. *Fundamentals of Forensic DNA Typing*. Elsevier Academic Press (2010).

[26] J. M. Butler. *Advanced Topics in Forensic DNA Typing: Methodology*. Elsevier Academic Press (2012).

[27] P. Calabrese und R. Sainudiin. Models of microsatellite evolution. In: R. Nielsen (Hrsg.), *Statistical Methods in Molecular Evolution*, S. 290–305. Springer (2005).

[28] P. P. Calabrese, R. T. Durrett und C. F. Aquadro. Dynamics of microsatellite divergence under stepwise mutation and proportional slippage/point mutation models. *Genetics* 159(2): 839–852 (2001).

[29] C. T. Caskey. New aid to human gene mapping. *Nature* 314(6006): 19 (1985).

[30] A. L. Castel, J. D. Cleary und C. E. Pearson. Repeat instability as the basis for human diseases and as a potential target for therapy. *Nat. Rev. Mol. Cell Biol.* 11(3): 165–170 (2010).

[31] R. Chakraborty und K. K. Kidd. The utility of DNA typing in forensic work. *Science* 254(5039): 1735–1739 (1991).

[32] G. K. Chambers und E. S. MacAvoy. Microsatellites: Consensus and controversy. *Comp. Biochem. Physiol. B: Biochem. Mol. Biol.* 126(4): 455–476 (2000).

[33] B. Charlesworth und D. Charlesworth. *Elements of Evolutionary Genetics*. Roberts & Co. (2010).

[34] M. B. Clark, P. P. Amaral, F. J. Schlesinger, M. E. Dinger, R. J. Taft, J. L. Rinn et al. The reality of pervasive transcription. *PLoS Biology* 9(7): e1000625 (2011).

[35] A. Collins und N. E. Morton. Likelihood ratios for DNA identification. *Proc. Natl. Acad. Sci. USA* 91(13): 6007–6011 (1994).

[36] D. E. Comings. The structure and function of chromatin. In: H. Harris und K. Hirschhorn (Hrsg.), *Advances in Human Genetics 3*, S. 237–431. Plenum Press (1972).

[37] N. R. C. Committee on DNA Forensic Science. *The Evaluation of Forensic DNA Evidence: An Update*. The National Academies Press (1996).

[38] N. R. C. Committee on DNA Technology in Forensic Science. *DNA Technology in Forensic Science*. The National Academies Press (1992).

[39] N. R. C. Committee on Identifying the Needs of the Forensic Sciences Community. *Strengthening Forensic Science in the United States: A Path Forward*. The National Academies Press (2009).

[40] G. Cooper, N. J. Burroughs, D. A. Rand, D. C. Rubinsztein und W. Amos. Markov Chain Monte Carlo analysis of human Y-chromosome microsatellites provides evidence of biased mutation. *Proc. Natl. Acad. Sci. USA* 96(21): 11916–11921 (1999).

[41] J. F. Crow. Genetic load. In: E. Fox Keller und E. A. Lloyd (Hrsg.), *Keywords in Evolutionary Biology*, S. 132–136. Harvard University Press (1992).

[42] J. F. Crow und M. Kimura. *An Introduction to Population Genetics Theory*. Harper & Row (1970).

[43] C. J. Cummings und H. Y. Zoghbi. Fourteen and counting: Unraveling trinucleotide repeat diseases. *Hum. Mol. Genet.* 9(6): 909–916 (2000).

[44] Department of Justice (USA). DNA-sample collection and biological evidence preservation in the federal jurisdiction. *Federal Register* 73(238): 74932–74943 (2008).

[45] B. Devlin, N. Risch und K. Roeder. Statistical evaluation of DNA fingerprinting: A critique of the NRC's report. *Science* 259(5096): 748–749 (1993).

[46] A. Di Rienzo, A. C. Peterson, J. C. Garza, A. M. Valdes, M. Slatkin und N. B. Freimer. Mutational processes of simple-sequence repeat loci in human populations. *Proc. Natl. Acad. Sci. USA* 91(8): 3166–3170 (1994).

[47] C. Dib, S. Fauré, C. Fizames, D. Samson, N. Drouot, A. Vignal et al. A comprehensive genetic map of the human genome based on 5,264 microsatellites. *Nature* 380(6570): 152–154 (1996).

[48] D. Dieringer und C. Schlötterer. Two distinct modes of microsatellite mutation processes: Evidence from the complete genomic sequences of nine species. *Genome Res.* 13(10): 2242–2251 (2003).

[49] S. Djebali, C. A. Davis, A. Merkel, A. Dobin, T. Lassmann, A. Mortazavi et al. Landscape of transcription in human cells. *Nature* 489(7414): 101–108 (2012).

[50] P. Donnelly und S. Tavaré. Coalescents and genealogical structure under neutrality. *Annu. Rev. Genet.* 29: 401–421 (1995).

[51] W. F. Doolittle. Is junk DNA bunk? A critique of ENCODE. *Proc. Natl. Acad. Sci. USA* 110(14): 5294–5300 (2013).

[52] W. F. Doolittle und C. Sapienza. Selfish genes, the phenotype paradigm and genome evolution. *Nature* 284(5757): 601–603 (1980).

[53] R. Durrett und S. Kruglyak. A new stochastic model of microsatellite evolution. *J. Appl. Probab.* 36(3): 621–631 (1999).

[54] M. Duyao, C. Ambrose, R. Myers, A. Novelletto, F. Persichetti, M. Frontali et al. Trinucleotide repeat length instability and age of onset in Huntington's disease. *Nat. Genet.* 4(4): 387–392 (1993).

[55] J. R. Ecker, W. A. Bickmore, I. Barroso, J. K. Pritchard, Y. Gilad und E. Segal. Genomics: ENCODE explained. *Nature* 489(7414): 52–55 (2012).

[56] S. R. Eddy. The C-value paradox, junk DNA and ENCODE. *Current Biology* 22(21): R898–R899 (2012).

[57] S. R. Eddy. The ENCODE project: Missteps overshadowing a success. *Current Biology* 23(7): R259–R261 (2013).

[58] A. Edwards, A. Civitello, H. A. Hammond und C. T. Caskey. DNA typing and genetic mapping with trimeric and tetrameric tandem repeats. *Am. J. Hum. Genet.* 49(4): 746–756 (1991).

[59] A. Edwards, H. A. Hammond, L. Jin, C. Caskey und R. Chakraborty. Genetic variation at five trimeric and tetrameric tandem repeat loci in four human population groups. *Genomics* 12(2): 241–253 (1992).

[60] J. A. Eisen. Mechanistic basis for microsatellite instability. In: D. B. Goldstein und C. Schlötterer (Hrsg.), *Microsatellites: Evolution and Applications*, S. 34–48. Oxford University Press (1999).

[61] H. Ellegren. Microsatellites: Simple sequences with complex evolution. *Nat. Rev. Genet.* 5(6): 435–445 (2004).

[62] H. Ellegren. Sex-chromosome evolution: Recent progress and the influence of male and female heterogamety. *Nat. Rev. Genet.* 12(3): 157–166 (2011).

[63] ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* 489(7414): 57–74 (2012).

[64] J. A. Endler. *Geographic Variation, Speciation, and Clines.* Princeton University Press (1977).

[65] J. T. Epplen und A. D. Akkad. Microsatellites. In: *Encyclopedia of Life Sciences (eLS).* J. Wiley & Sons (2011).

[66] W. J. Ewens. *Mathematical Population Genetics: I. Theoretical Introduction*. Springer (2004).

[67] D. Falush und Y. Iwasa. Size-dependent mutability and microsatellite constraints. *Mol. Biol. Evol.* 16(7): 960–966 (1999).

[68] M. W. Feldman, A. Bergman, D. D. Pollock und D. B. Goldstein. Microsatellite genetic distances with range constraints: Analytic description and problems of estimation. *Genetics* 145(1): 207–216 (1997).

[69] R. A. Fisher. *The Genetical Theory of Natural Selection*. Oxford University Press. Variorum-Ausgabe (1999).

[70] P. Francalacci, L. Morelli, A. Angius, R. Berutti, F. Reinier, R. Atzeni et al. Low-pass DNA sequencing of 1200 Sardinians reconstructs European Y-chromosome phylogeny. *Science* 341(6145): 565–569 (2013).

[71] Y.-X. Fu und W.-H. Li. Coalescing into the 21st century: An overview and prospects of coalescent theory. *Theor. Popul. Biol.* 56(1): 1–10 (1999).

[72] F. Galton. *Finger Prints*. Macmillan & Co. (1892).

[73] J. Garza, M. Slatkin und N. Freimer. Microsatellite allele frequencies in humans and chimpanzees, with implications for constraints on allele size. *Mol. Biol. Evol.* 12(4): 594–603 (1995).

[74] J. R. Gatchel und H. Y. Zoghbi. Diseases of unstable repeat expansion: Mechanisms and common principles. *Nat. Rev. Genet.* 6(10): 743–755 (2005).

[75] R. Gemayel, M. D. Vinces, M. Legendre und K. J. Verstrepen. Variable tandem repeats accelerate evolution of coding and regulatory sequences. *Annu. Rev. Genet.* 44(1): 445–477 (2010).

[76] P. Gill, C. Brenner, B. Brinkmann, B. Budowle, A. Carracedo, M. A. Jobling et al. DNA Commission of the International Society of Forensic Genetics: Recommendations on forensic analysis using Y-chromosome STRs. *Int. J. Legal Med.* 114(6): 305–309 (2001).

[77] J. Gitschier. The eureka moment: An interview with Sir Alec Jeffreys. *PLoS Genetics* 5(12): e1000765 (2009).

[78] D. W. Gjertson, C. H. Brenner, M. P. Baur, A. Carracedo, F. Guidet, J. A. Luque et al. ISFG: Recommendations on biostatistics in paternity testing. *Forensic Sci. Int. Genet.* 1(3–4): 223–231 (2007).

[79] H. Goehler, M. Lalowski, U. Stelzl, S. Waelter, M. Stroedicke, U. Worm et al. A protein interaction network links GIT1, an enhancer of huntingtin aggregation, to Huntington's disease. *Molecular Cell* 15(6): 853–865 (2004).

[80] O. B. Goin, C. J. Goin und K. Bachmann. DNA and amphibian life history. *Copeia* 1968(3): 532–540 (1968).

[81] D. B. Goldstein, G. W. Roemer, D. A. Smith, D. E. Reich, A. Bergman und R. K. Wayne. The use of microsatellite variation to infer population structure and demographic history in a natural model system. *Genetics* 151(2): 797–801 (1999).

[82] S. J. Gould und N. Eldredge. Punctuated equilibrium comes of age. *Nature* 366(6452): 223–227 (1993).

[83] S. J. Gould und R. C. Lewontin. The spandrels of San Marco and the Panglossian paradigm: A critique of the adaptationist programme. *Proc. R. Soc. Lond. B* 205(1161): 581–598 (1979).

[84] D. Graur, Y. Zheng, N. Price, R. B. R. Azevedo, R. A. Zufall und E. Elhaik. On the immortality of television sets: "Function" in the human genome according to the evolution-free gospel of ENCODE. *Genome Biol. Evol.* 5(3): 578–590 (2013).

[85] T. R. Gregory. Quotes of interest — satellite DNA. Blog-Eintrag, URL: http://www.genomicron.evolverzone.com/2008/02/quotes-of-interest-satellite-dna (17.2.2008).

[86] T. R. Gregory. Animal Genome Size Database. Website, URL: www.genomesize.com.

[87] J. F. Gusella, N. S. Wexler, P. M. Conneally, S. L. Naylor, M. A. Anderson, R. E. Tanzi et al. A polymorphic DNA marker genetically linked to Huntington's disease. *Nature* 306(5940): 234–238 (1983).

[88] L. Gusmão, J. Butler, A. Carracedo, P. Gill, M. Kayser, W. Mayr et al. DNA Commission of the International Society of Forensic Genetics (ISFG): An update of the recommendations on the use of Y-STRs in forensic analysis. *Forensic Sci. Int.* 157(2-3): 187–197 (2006).

[89] J. B. S. Haldane. The effect of variation on fitness. *The American Naturalist* 71(735): 337–349 (1937).

[90] J. B. S. Haldane. The cost of natural selection. *Journal of Genetics* 55(3): 511–524 (1957).

[91] G. Hampikian, E. West und O. Akselrod. The genetics of innocence: Analysis of 194 U.S. DNA exonerations. *Annu. Rev. Genomics Hum. Genet.* 12(1): 97–120 (2011).

[92] D. R. Hares. Expanding the CODIS core loci in the United States. *Forensic Sci. Int. Genet.* 6(1): e52–e54 (2012).

[93] B. Harper. Huntington disease. *J. R. Soc. Med.* 98(12): 550 (2005).

[94] D. L. Hartl und A. G. Clark. *Principles of Population Genetics*. Sinauer Associates, 4. Auflage (2007).

[95] P. W. Hedrick. *Genetics of Populations*. Jones & Bartlett, 3. Auflage (2005).

[96] J. Hein, M. H. Schierup und C. Wiuf. *Gene Genealogies, Variation and Evolution: A Primer in Coalescent Theory*. Oxford University Press (2005).

[97] Q.-Y. Huang, F.-H. Xu, H. Shen, H.-Y. Deng, Y.-J. Liu, Y.-Z. Liu et al. Mutation patterns at dinucleotide microsatellite loci in humans. *Am. J. Hum. Genet.* 70(3): 625–634 (2002).

[98] J. F. Hughes und S. Rozen. Genomics and genetics of human and primate Y chromosomes. *Annu. Rev. Genomics Hum. Genet.* 13(1): 83–108 (2012).

[99] J. F. Hughes, H. Skaletsky, L. G. Brown, T. Pyntikova, T. Graves, R. S. Fulton et al. Strict evolutionary conservation followed rapid gene loss on human and rhesus Y chromosomes. *Nature* 483(7387): 82–86 (2012).

[100] J. F. Hughes, H. Skaletsky, T. Pyntikova, T. A. Graves, S. K. M. van Daalen, P. J. Minx et al. Chimpanzee and human Y chromosomes are remarkably divergent in structure and gene content. *Nature* 463(7280): 536–539 (2010).

[101] Huntington's Disease Collaborative Research Group. A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. *Cell* 72(6): 971–983 (1993).

[102] G. Imbert, C. Kretz, K. Johnson und J.-L. Mandel. Origin of the expansion mutation in myotonic dystrophy. *Nat. Genet.* 4(1): 72–76 (1993).

[103] A. J. Jeffreys. Naming "DNA fingerprinting". Undatiertes Video, URL: www.dnalc.org/view/15106-Naming-DNA-fingerprinting-.html.

[104] A. J. Jeffreys, J. F. Y. Brookfield und R. Semeonoff. Positive identification of an immigration test-case using human DNA fingerprints. *Nature* 317(6040): 818–819 (1985).

[105] A. J. Jeffreys, V. Wilson, R. Neumann und J. Keyte. Amplification of human minisatellites by the polymerase chain reaction: Towards DNA fingerprinting of single cells. *Nucleic Acids Res.* 16(23): 10953–10971 (1988).

[106] A. J. Jeffreys, V. Wilson und S. L. Thein. Hypervariable 'minisatellite' regions in human DNA. *Nature* 314(6006): 67–73 (1985).

[107] A. J. Jeffreys, V. Wilson und S. L. Thein. Individual-specific 'fingerprints' of human DNA. *Nature* 316(6023): 76–79 (1985).

[108] A. Jha. Breakthrough study overturns theory of 'junk DNA' in genome. *The Guardian* (5.9.2012).

[109] M. A. Jobling und P. Gill. Encoded evidence: DNA in forensic analysis. *Nat. Rev. Genet.* 5(10): 739–751 (2004).

[110] M. A. Jobling, M. Hurles und C. Tyler-Smith. *Human Evolutionary Genetics: Origins, Peoples and Disease*. Garland (2004).

[111] M. A. Jobling und C. Tyler-Smith. The human Y chromosome: An evolutionary marker comes of age. *Nat. Rev. Genet.* 4(8): 598–612 (2003).

[112] L. B. Jorde, M. J. Bamshad, W. S. Watkins, R. Zenger, A. E. Fraley, P. A. Krakowiak et al. Origins and affinities of modern humans: A comparison of mitochondrial and nuclear genetic data. *Am. J. Hum. Genet.* 57(3): 523–538 (1995).

[113] L. B. Jorde, A. R. Rogers, M. Bamshad, W. S. Watkins, P. Krakowiak, S. Sung et al. Microsatellite diversity and the demographic history of modern humans. *Proc. Natl. Acad. Sci. USA* 94(7): 3100–3103 (1997).

[114] T. H. Jukes. *Molecules and Evolution*. Columbia University Press (1966).

[115] T. M. Karafet, F. L. Mendez, M. B. Meilerman, P. A. Underhill, S. L. Zegura und M. F. Hammer. New binary polymorphisms reshape and increase resolution of the human Y chromosomal haplogroup tree. *Genome Res.* 18(5): 830–838 (2008).

[116] Y. Kashi und D. G. King. Simple sequence repeats as advantageous mutators in evolution. *Trends Genet.* 22(5): 253–259 (2006).

[117] S. H. Katsanis und J. K. Wagner. Characterization of the standard and recommended CODIS markers. *J. Forensic Sci.* 58(S1): S169–S172 (2013).

[118] M. Kayser, A. Caglià, D. Corach, N. Fretwell, C. Gehrig, G. Graziosi et al. Evaluation of Y-chromosomal STRs: A multicenter study. *Int. J. Legal Med.* 110(3): 125–133, 141–149 (1997).

[119] Y. D. Kelkar, S. Tyekucheva, F. Chiaromonte und K. D. Makova. The genome-wide determinants of human and chimpanzee microsatellite evolution. *Genome Res.* 18(1): 30–38 (2008).

[120] H. Kesten. The number of alleles in electrophoretic experiments. *Theor. Popul. Biol.* 18(2): 290–294 (1980).

[121] H. Kesten. The number of distinguishable alleles according to the Ohta-Kimura model of neutral mutation. *J. Math. Biol.* 10(2): 167–187 (1980).

[122] A. N. Khachane und P. M. Harrison. Assessing the genomic evidence for conserved transcribed pseudogenes under selection. *BMC Genomics* 10(1): 435 (2009).

[123] M. Kimmel. Population dynamics coded in DNA: Genetic traces of the expansion of modern humans. *Physica A* 273(1-2): 158–168 (1999).

[124] M. Kimmel, R. Chakraborty, J. P. King, M. Bamshad, W. S. Watkins und L. B. Jorde. Signatures of population expansion in microsatellite repeat data. *Genetics* 148(4): 1921–1930 (1998).

[125] M. Kimmel, R. Chakraborty, D. N. Stivers und R. Deka. Dynamics of repeat polymorphisms under a forward-backward mutation model: Within- and between-population variability at microsatellite loci. *Genetics* 143(1): 549–555 (1996).

[126] M. Kimura. "Stepping Stone" model of population. *Annu. Rep. Nat. Inst. Genet. (Japan)* 3: 62–63 (1953).

[127] M. Kimura. Evolutionary rate at the molecular level. *Nature* 217(5129): 624–626 (1968).

[128] M. Kimura. *Die Neutralitätstheorie der molekularen Evolution.* Paul Parey (1987).

[129] M. Kimura und G. H. Weiss. The stepping stone model of population structure and the decrease of genetic correlation with distance. *Genetics* 49(4): 561–576 (1964).

[130] J. L. King und T. H. Jukes. Non-Darwinian evolution. *Science* 164(3881): 788–798 (1969).

[131] J. Kingman. A note on multidimensional models of neutral mutation. *Theor. Popul. Biol.* 11(3): 285–290 (1977).

[132] J. Kingman. The coalescent. *Stochastic Processes and their Applications* 13(3): 235–248 (1982).

[133] J. Kingman. Exchangeability and the evolution of large populations. In: G. Koch und F. Spizzichino (Hrsg.), *Exchangeability in Probability and Statistics*, S. 97–112. Elsevier (1982).

[134] J. F. C. Kingman. Coherent random walks arising in some genetical models. *Proc. R. Soc. Lond. A* 351(1664): 19–31 (1976).

[135] J. F. C. Kingman. On the genealogy of large populations. *J. Appl. Probab.* 19: 27–43 (1982).

[136] J. F. C. Kingman. Origins of the coalescent: 1974–1982. *Genetics* 156(4): 1461–1463 (2000).

[137] G. Kolata. Far from 'junk,' DNA dark matter proves crucial to health. *The New York Times* (5.9.2012).

[138] A. Kong, M. L. Frigge, G. Masson, S. Besenbacher, P. Sulem, G. Magnusson et al. Rate of de novo mutations and the importance of father's age to disease risk. *Nature* 488(7412): 471–475 (2012).

[139] M. Krawczak. Forensic evaluation of Y-STR haplotype matches: A comment. *Forensic Sci. Int.* 118(2-3): 114–115 (2001).

[140] M. Krawczak und J. Schmidtke. *DNA Fingerprinting.* BIOS Scientific – Springer, 2. Auflage (1998).

[141] B. Kremer, E. Almqvist, J. Theilmann, N. Spence, H. Telenius, Y. P. Goldberg et al. Sex-dependent mechanisms for expansions and contractions of the CAG repeat on affected Huntington disease chromosomes. *Am. J. Hum. Genet.* 57(2): 343–350 (1995).

[142] S. Krimsky und T. Simoncelli. *Genetic Justice: DNA Data Banks, Criminal Investigations, and Civil Liberties.* Columbia University Press (2011).

[143] S. Kruglyak, R. T. Durrett, M. D. Schug und C. F. Aquadro. Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations. *Proc. Natl. Acad. Sci. USA* 95(18): 10774–10778 (1998).

[144] S. Kumar. Molecular clocks: Four decades of evolution. *Nat. Rev. Genet.* 6(8): 654–662 (2005).

[145] A. R. La Spada und J. P. Taylor. Repeat expansion disease: Progress and puzzles in disease pathogenesis. *Nat. Rev. Genet.* 11(4): 247–258 (2010).

[146] Y. Lai und F. Sun. The relationship between microsatellite slippage mutation rate and the number of repeat units. *Mol. Biol. Evol.* 20(12): 2123–2131 (2003).

[147] E. S. Lander. DNA fingerprinting on trial. *Nature* 339(6225): 501–505 (1989).

[148] E. S. Lander. Invited editorial: Research on DNA typing catching up with courtroom application. *Am. J. Hum. Genet.* 48(5): 819–823 (1991).

[149] E. S. Lander und B. Budowle. DNA fingerprinting dispute laid to rest. *Nature* 371(6500): 735–738 (1994).

[150] E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin et al. Initial sequencing and analysis of the human genome. *Nature* 409(6822): 860–921 (2001).

[151] A. Laouina, S. Nadifi, R. Boulouiz, M. El Arji, J. Talbi, B. El Houate et al. Mutation rate at 17 Y-STR loci in "father/son" pairs from Moroccan population. *Legal Medicine* 15(5): 269–271 (2013).

[152] J.-M. Lee, E. Ramos, J.-H. Lee, T. Gillis, J. Mysore, M. Hayden et al. CAG repeat expansion in Huntington disease determines age at onset in a fully dominant fashion. *Neurology* 78(10): 690–695 (2012).

[153] G. Levinson und G. A. Gutman. High frequencies of short frameshifts in poly-CA/TG tandem repeats borne by bacteriophage M13 in Escherichia coli K-12. *Nucleic Acids Res.* 15(13): 5323–5338 (1987).

[154] G. Levinson und G. A. Gutman. Slipped-strand mispairing: A major mechanism for DNA sequence evolution. *Mol. Biol. Evol.* 4(3): 203–221 (1987).

[155] R. C. Lewontin. Let the DNA fit the crime. *The New York Review of Books* (23.2.2012).

[156] R. C. Lewontin und D. L. Hartl. Population genetics in forensic DNA typing. *Science* 254(5039): 1745–1750 (1991).

[157] Y.-C. Li, A. B. Korol, T. Fahima, A. Beiles und E. Nevo. Microsatellites: Genomic distribution, putative functions and mutational mechanisms: A review. *Mol. Ecol.* 11(12): 2453–2465 (2002).

[158] M. Lynch. *The Origins of Genome Architecture.* Palgrave Macmillan (2007).

[159] M. E. MacDonald. Huntingtin: Alive and well and working in middle management. *Science Signaling* 2003(207): pe48 (2003).

[160] B. Maher. ENCODE: The human encyclopaedia. *Nature* 489(7414): 46–48 (2012).

[161] G. Malécot. Quelques schémas probabilistes sur la variabilité des populations naturelles. *Ann. Univ. Lyon, Sciences, Section A* 13: 37–60 (1950).

[162] G. Malécot. Migration et parenté génétique moyenne. In: J. Sutter (Hrsg.), *Les Déplacements Humains: Aspects méthodologiques de leur mesure*, S. 205–212. Diffusion Hachette (1963).

[163] K. A. Mayntz-Press und J. Ballantyne. Performance characteristics of commercial Y-STR multiplex systems. *J. Forensic Sci.* 52(5): 1025–1034 (2007).

[164] O. McCrimmon. ENCODE data describes function of human genome. *NIH News* URL: http://www.genome.gov/27549810 (5.9.2012).

[165] L. McIver, J. Fondon III, M. Skinner und H. Garner. Evaluation of microsatellite variation in the 1000 Genomes Project pilot studies is indicative of the quality and utility of the raw data and alignments. *Genomics* 97(4): 193–199 (2011).

[166] A. E. Mirsky und H. Ris. The desoxyribonucleic acid content of animal cells and its evolutionary significance. *J. Gen. Physiol.* 34(4): 451–462 (1951).

[167] J. L. Mnookin. Fingerprint evidence in an age of DNA profiling. *Brooklyn Law Review* 67(1): 13–70 (2001).

[168] P. A. P. Moran. Wandering distributions and the electrophoretic profile. *Theor. Popul. Biol.* 8(3): 318–330 (1975).

[169] P. A. P. Moran. Wandering distributions and the electrophoretic profile II. *Theor. Popul. Biol.* 10(2): 145–149 (1976).

[170] N. Morris. A 'chilling' proposal for a universal DNA database. *The Independent* (6.9.2007).

[171] J. J. Mulero, C. W. Chang, L. M. Calandro, R. L. Green, Y. Li, C. L. Johnson et al. Development and validation of the AmpFlSTR Yfiler PCR amplification kit: A male specific, single amplification 17 Y-STR multiplex system. *J. Forensic Sci.* 51(1): 64–75 (2006).

[172] H. J. Muller. Our load of mutations. *Am. J. Hum. Genet.* 2(2): 111–176 (1950).

[173] K. B. Mullis und F. A. Faloona. Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction. *Methods in Enzymology* 155: 335–350 (1987).

[174] M. J. Nauta und F. J. Weissing. Constraints on allele size at microsatellite loci: Implications for genetic differentiation. *Genetics* 143(2): 1021–1032 (1996).

[175] M. A. Nóbrega, Y. Zhu, I. Plajzer-Frick, V. Afzal und E. M. Rubin. Megabase deletions of gene deserts result in viable mice. *Nature* 431(7011): 988–993 (2004).

[176] D. Nelson, H. Orr und S. Warren. The unstable repeats — three evolving faces of neurological disease. *Neuron* 77(5): 825–843 (2013).

[177] C. Neumann. Fingerprints at the crime-scene: Statistically certain, or probable? *Significance* 9(1): 21–25 (2012).

[178] D. H. P. Nguyen. Morbus Huntington und Huntington-ähnliche Erkrankungen. *Medizinische Genetik* 25(2): 221–227 (2013).

[179] D.-K. Niu und L. Jiang. Can ENCODE tell us how much junk DNA we carry in our genome? *Biochem. Biophys. Res. Commun.* 430(4): 1340–1343 (2013).

[180] M. Nordborg. Coalescent theory. In: D. J. Balding, M. J. Bishop und C. Cannings (Hrsg.), *Handbook of Statistical Genetics*, S. 179–212. J. Wiley & Sons (2001).

[181] S. Ohno. So much "junk" DNA in our genome. *Brookhaven Symposia in Biology* 23: 366–370 (1972).

[182] S. Ohno. Evolutional reason for having so much junk DNA. In: R. Pfeiffer (Hrsg.), *Modern Aspects of Cytogenetics: Constitutive Heterochromatin in Man*, S. 169–173. Schattauer (1973).

[183] T. Ohta und M. Kimura. A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. *Genetics* 22(2): 201–204 (1973).

[184] T. Ohta und M. Kimura. Simulation studies on electrophoretically detectable genetic variability in a finite population. *Genetics* 76(3): 615–624 (1974).

[185] E. Olmo. Quantitative variations in the nuclear DNA and phylogenesis of the Amphibia. *Caryologia* 26: 43–68 (1973).

[186] L. E. Orgel und F. H. Crick. Selfish DNA: The ultimate parasite. *Nature* 284(5757): 604–607 (1980).

[187] H. T. Orr und H. Y. Zoghbi. Trinucleotide repeat disorders. *Annu. Rev. Neurosci.* 30(1): 575–621 (2007).

[188] G. B. Panigrahi, R. Lau, S. E. Montgomery, M. R. Leonard und C. E. Pearson. Slipped (CTG)•(CAG) repeats can be correctly repaired, escape repair or undergo error-prone repair. *Nat. Struct. Mol. Biol.* 12(8): 654–662 (2005).

[189] C. E. Pearson, K. N. Edamura und J. D. Cleary. Repeat instability: Mechanisms of dynamic mutations. *Nat. Rev. Genet.* 6(10): 729–742 (2005).

[190] E. Pennisi. ENCODE project writes eulogy for junk DNA. *Science* 337(6099): 1159–1161 (2012).

[191] G. D. Poznik, B. M. Henn, M.-C. Yee, E. Sliwerska, G. M. Euskirchen, A. A. Lin et al. Sequencing Y chromosomes resolves discrepancy in time to common ancestor of males versus females. *Science* 341(6145): 562–565 (2013).

[192] J. K. Pritchard und M. W. Feldman. Statistics for microsatellite variation based on coalescence. *Theor. Popul. Biol.* 50(3): 325–344 (1996).

[193] L. Roewer. Populationsgenetik des Y-Chromosoms. *Medizinische Genetik* 20(3): 288–292 (2008).

[194] L. Roewer. Y chromosome STR typing in crime casework. *Forensic Sci. Med. Pathol.* 5(2): 77–84 (2009).

[195] L. Roewer, M. Kayser, P. de Knijff, K. Anslinger, A. Betz, A. Caglià et al. A new method for the evaluation of matches in non-recombining genomes: Application to Y-chromosomal short tandem repeat (STR) haplotypes in European males. *Forensic Sci. Int.* 114(1): 31–43 (2000).

[196] O. Rose und D. Falush. A threshold size for microsatellite expansion. *Mol. Biol. Evol.* 15(5): 613–615 (1998).

[197] N. A. Rosenberg und M. Nordborg. Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. *Nat. Rev. Genet.* 3(5): 380–390 (2002).

[198] C. A. Ross und S. J. Tabrizi. Huntington's disease: From molecular pathogenesis to clinical treatment. *The Lancet Neurology* 10(1): 83–98 (2011).

[199] S. Rozen, H. Skaletsky, J. D. Marszalek, P. J. Minx, H. S. Cordum, R. H. Waterston et al. Abundant gene conversion between arms of palindromes in human and ape Y chromosomes. *Nature* 423(6942): 873–876 (2003).

[200] D. C. Rubinsztein. Trinucleotide expansion mutations cause diseases which do not conform to classical Mendelian expectations. In: D. B. Goldstein und C. Schlötterer (Hrsg.), *Microsatellites: Evolution and Applications*, S. 80–97. Oxford University Press (1999).

[201] D. C. Rubinsztein, J. Leggo, M. Chiano, A. Dodge, G. Norbury, E. Rosser et al. Genotypes at the GluR6 kainate receptor locus are associated with variation in the age of onset of Huntington disease. *Proc. Natl. Acad. Sci. USA* 94(8): 3872–3876 (1997).

[202] A. Ruiz-Linares. Microsatellites and the reconstruction of the history of human populations. In: D. B. Goldstein und C. Schlötterer (Hrsg.), *Microsatellites: Evolution and Applications*, S. 183–197. Oxford University Press (1999).

[203] R. K. Saiki, D. H. Gelfand, S. Stoffel, S. J. Scharf, R. Higuchi, G. T. Horn et al. Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science* 239(4839): 487–491 (1988).

[204] R. K. Saiki, S. Scharf, F. Faloona, K. B. Mullis, G. T. Horn, H. A. Erlich et al. Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science* 230(4732): 1350–1354 (1985).

[205] A. Scally und R. Durbin. Revising the human mutation rate: Implications for understanding human evolution. *Nat. Rev. Genet.* 13(10): 745–753 (2012).

[206] C. Schlötterer und D. Tautz. Slippage synthesis of simple sequence DNA. *Nucleic Acids Res.* 20(2): 211–215 (1992).

[207] M. D. Shriver und R. A. Kittles. Genetic ancestry and the search for personalized genetic histories. *Nat. Rev. Genet.* 5(8): 611–618 (2004).

[208] H. Skaletsky, T. Kuroda-Kawaguchi, P. J. Minx, H. S. Cordum, L. Hillier, L. G. Brown et al. The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature* 423(6942): 825–837 (2003).

[209] M. Slatkin. Gene flow in natural populations. *Annu. Rev. Ecol. System.* 16: 393–430 (1985).

[210] R. G. Snell, J. C. MacMillan, J. P. Cheadle, I. Fenton, L. P. Lazarou, P. Davies et al. Relationship between trinucleotide repeat expansion and phenotypic variation in Huntington's disease. *Nat. Genet.* 4(4): 393–397 (1993).

[211] J. N. Spuhler. On the number of genes in man. *Science* 108(2802): 279–280 (1948).

[212] R. Stein. Scientists unveil 'Google maps' for human genome. *NPR: All Things Considered* Radiosendung, URL:

http://www.npr.org/blogs/health/2012/09/05/160599136/scientists-unveil-google-maps-for-human-genome (5.9.2012).

[213] C. Stern. *Principles Of Human Genetics*. W. H. Freeman, 2. Auflage (1960).

[214] S. M. Stigler. Galton and identification by fingerprints. *Genetics* 140(3): 857–860 (1995).

[215] K. Struhl. Transcriptional noise and the fidelity of initiation by RNA polymerase II. *Nat. Struct. Mol. Biol.* 14(2): 103–105 (2007).

[216] J. X. Sun, A. Helgason, G. Masson, S. S. Ebenesersdóttir, H. Li, S. Mallick et al. A direct characterization of human mutation based on microsatellites. *Nat. Genet.* (2012).

[217] J. X. Sun, J. C. Mullikin, N. Patterson und D. E. Reich. Microsatellites are molecular clocks that support accurate inferences about history. *Mol. Biol. Evol.* 26(5): 1017–1027 (2009).

[218] H. Takano und J. F. Gusella. The predominantly HEAT-like motif structure of huntingtin and its association and coincident nuclear entry with dorsal, an NF-kB/Rel/dorsal family transcription factor. *BMC Neuroscience* 3(1): 15 (2002).

[219] D. Tautz. Notes on the definition and nomenclature of tandemly repetitive DNA sequences. In: *DNA Fingerprinting: State of the Science* S. 21–28. Birkhäuser (1993).

[220] S. Tavaré. Line-of-descent and genealogical processes, and their applications in population genetics models. *Theor. Popul. Biol.* 26(2): 119–164 (1984).

[221] Thousand Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature* 491(7422): 56–65 (2012).

[222] D. Torrents, M. Suyama, E. Zdobnov und P. Bork. A genome-wide survey of human pseudogenes. *Genome Res.* 13(12): 2559–2567 (2003).

[223] P. A. Underhill und T. Kivisild. Use of Y chromosome and mitochondrial DNA population structure in tracing human migrations. *Annu. Rev. Genet.* 41: 539–564 (2007).

[224] H. van Bakel, C. Nislow, B. J. Blencowe und T. R. Hughes. Most "dark matter" transcripts are associated with known genes. *PLoS Biology* 8(5): e1000371 (2010).

[225] J. C. Venter, M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton et al. The sequence of the human genome. *Science* 291(5507): 1304–1351 (2001).

[226] B. Veytsman und L. Akhmadeyeva. Simple mathematical model of pathologic microsatellite expansions: When self-reparation does not work. *J. Theor. Biol.* 242(2): 401–408 (2006).

[227] F. Walsh. Detailed map of genome function. *BBC News* URL: http://www.bbc.co.uk/news/health-19202141 (5.9.2012).

[228] J. Wambaugh. *The Blooding: The True Story of the Narborough Village Murders*. W. Morrow & Co. (1989).

[229] J. C. Watkins. Microsatellite evolution: Markov transition functions for a suite of models. *Theor. Popul. Biol.* 71(2): 147–159 (2007).

[230] N. Weber. Encode-Projekt entschlüsselt Geheimnisse der Junk-DNA. *Spiegel Online* (5.9.2012).

[231] B. S. Weir. Population genetics in the forensic DNA debate. *Proc. Natl. Acad. Sci. USA* 89(24): 11654–11659 (1992).

[232] B. S. Weir. *Genetic Data Analysis*. Sinauer Associates, II. Auflage (1996).

[233] B. S. Weir. Forensics. In: D. J. Balding, M. J. Bishop und C. Cannings (Hrsg.), *Handbook of Statistical Genetics*, S. 721–739. J. Wiley & Sons (2001).

[234] J. Weissenbach. A second generation linkage map of the human genome based on highly informative microsatellite loci. *Gene* 135(1-2): 275–278 (1993).

[235] W. Weng, H. Liu, S. Li, J. Ge, H. Wang und C. Liu. Mutation rates at 16 Y-chromosome STRs in the South China Han population. *Int. J. Legal Med.* 127(2): 369–372 (2013).

[236] N. S. Wexler. Venezuelan kindreds reveal that genetic and environmental factors modulate Huntington's disease age of onset. *Proc. Natl. Acad. Sci. USA* 101(10): 3498–3503 (2004).

[237] N. S. Wexler. Huntington's disease: Advocacy driving science. *Annu. Rev. Med.* 63: 1–22 (2012).

[238] N. S. Wexler, A. B. Young, R. E. Tanzi, H. Travers, S. Starosta-Rubinstein, J. B. Penney et al. Homozygotes for Huntington's disease. *Nature* 326(6109): 194–197 (1987).

[239] J. C. Whittaker, R. M. Harbord, N. Boxall, I. Mackay, G. Dawson und R. M. Sibly. Likelihood-based estimation of microsatellite mutation rates. *Genetics* 164(2): 781–787 (2003).

[240] A. J. Williams und H. L. Paulson. Polyglutamine neurodegeneration: Protein misfolding revisited. *Trends Neurosci.* 31(10): 521–528 (2008).

[241] H. Willmann. *Langenscheidts Großwörterbuch Englisch, Teil I: Englisch-Deutsch*. Langenscheidt (2001).

[242] S. Willuweit, A. Caliebe, M. M. Andersen und L. Roewer. Y-STR Frequency Surveying Method: A critical reappraisal. *Forensic Sci. Int. Genet.* 5(2): 84–90 (2011).

[243] S. Willuweit und L. Roewer. Y chromosome haplotype reference database (YHRD): Update. *Forensic Sci. Int. Genet.* 1(2): 83–87 (2007).

[244] I. J. Wilson und D. J. Balding. Genealogical inference from microsatellite data. *Genetics* 150(1): 499–510 (1998).

[245] I. J. Wilson, M. E. Weale und D. J. Balding. Inferences from DNA data: Population histories, evolutionary processes and forensic match probabilities. *J. R. Stat. Soc. Ser. A* 166(2): 155–188 (2003).

[246] S. Wright. Evolution in Mendelian populations. *Genetics* 16(2): 97–159 (1931).

[247] S. Wright. Breeding structure of populations in relation to speciation. *The American Naturalist* 74(752): 232 (1940).

[248] S. Wright. Isolation by distance. *Genetics* 28(2): 114–138 (1943).

[249] S. Wright. Isolation by distance under diverse systems of mating. *Genetics* 31(1): 39–59 (1946).

[250] S. Wright. The genetical structure of populations. *Annals of Eugenics* 15: 323–354 (1951).

[251] S. Wright. *Evolution and the Genetics of Populations, Volume 2: The Theory of Gene Frequencies*. University of Chicago Press. Taschenbuch-Ausgabe (1984).

[252] F. Wyers, M. Rougemaille, G. Badis, J.-C. Rousselle, M.-E. Dufour, J. Boulay et al. Cryptic Pol II transcripts are degraded by a nuclear quality control pathway involving a new poly(A) polymerase. *Cell* 121(5): 725–737 (2005).

[253] X. Xu, M. Peng, Z. Fang und X. Xu. The direction of microsatellite mutations is dependent upon allele length. *Nat. Genet.* 24(4): 396–399 (2000).

[254] L. A. Zhivotovsky, M. W. Feldman und S. A. Grishechkin. Biased mutations and microsatellite variation. *Mol. Biol. Evol.* 14(9): 926–933 (1997).

Abkürzungsverzeichnis

Die folgende Liste enthält die in dieser Arbeit verwendeten Abkürzungen, ihre Bedeutung und gegebenenfalls die Seite, auf der sie eingeführt oder erläutert werden.

A	Adenin	
BBC	British Broadcasting Corporation	
bp	(DNA) base pairs	
C	Cytosin	
cM	centi (10^{-2}) Morgan	
CODIS	Combined DNA Index System	24
DNA	Desoxyribonucleic Acid	
DYS	DNA Y-Segment (Locusnamen-Präfix)	29
ECHR	European Court of Human Rights	25
ENCODE	Encyclopedia of DNA Elements	90
FBI	Federal Bureau of Investigation (USA)	
FMR	Fragile-X Mental Retardation	5
G	Guanin	
Gb	Giga (10^9) base pairs	
HD	Huntington Disease	6
ISFG	International Society of Forensic Genetics	1
kb	kilo (10^3) base pairs	
ku	kilo (10^3) units (atomare Masseneinheit)	
Mb	Mega (10^6) base pairs	
MRCA	Most Recent Common Ancestor	9
MSY	Male-Specific region of the Y chromosome	26
μ	Mutationswahrscheinlichkeit pro Locus und Generation	
mya	million years ago (10^6 Jahre)	
n	Nummer der Generation (in Modellen mit diskreter Zeit)	
n_e	effektive Anzahl neutraler Allele (in einer Population)	18
N	Populationsgröße	
$\mathbb{N}$	Menge der natürlichen Zahlen	
NIH	National Institutes of Health (USA)	
NIST	National Institute of Standards and Technology (USA)	1
NPR	National Public Radio (US-amerikanisches Syndikat)	

NRC	National Research Council (USA)	22
PCR	Polymerase Chain Reaction	22
pg	piko (10^{-12}) Gramm	
RFLP	Restriction Fragment Length Polymorphism	22
RMP	Random Match Probability	25
RM-Y-STR	Rapidly Mutating Y-STR	28
S	State Space (Zustandsraum einer Markov-Kette)	
SMM	Stepwise Mutation Model	12
SNP	Single Nucleotide Polymorphism	
SNV	Single Nucleotide Variant	
SRY	Sex-determining Region Y	27
SSM	Slipped-Strand Mispairing	10
SSR	Simple/Short Sequence Repeat	1
STR	Short Tandem Repeat	1
STRP	Short Tandem Repeat Polymorphism	1
T	Thymin	
θ	$2N \cdot \mu$	
UK	United Kingdom (Großbritannien und Nordirland)	
USA	United States of America	
V	normalisierter Allelprozess	34
VNTR	Variable Number of Tandem Repeats	1
X	Allelprozess	34
YHRD	Y Haplotype Reference Database	28
Y-STR	STR-Locus auf dem Y-Chromosom	25
Z	Mutationsprozess	20
$\mathbb{Z}$	Menge der ganzen Zahlen	

Danksagung

Viele Menschen haben zum Gelingen dieser Arbeit beigetragen. Ich möchte mich insbesondere bei den folgenden Personen bedanken.

Prof. Dr. Michael Krawczak hat meine Arbeit vorbildlich betreut und für hervorragende Arbeitsbedingungen im Institut für Medizinische Informatik und Statistik (IMIS) gesorgt. Prof. Dr. Uwe Rösler hat mich ermutigt, diese Arbeit zu beginnen, und er hatte die Grundidee für den in Abschnitt 3.1.2 ausgeführten alternativen Beweis der Existenz der invarianten Verteilung. Dr. Amke Caliebe hat das gesamte Manuskript gelesen und mir viele wertvolle Hinweise gegeben. Ich danke auch Prof. Dr. Manuela Dittmar für die gute Beratung und für die Bereitschaft, das Zweitgutachten zu erstellen.

Die *Y Chromosome Haplotype Reference Database* [243], implementiert und gepflegt von Prof. Dr. Lutz Roewer und Sascha Willuweit, war eine große Hilfe bei der Arbeit an den Publikationen (ii) und (iii).

Ich danke Mikkel Meyer Andersen und allen bereits genannten Koautoren der drei eingeschlossenen Publikationen für die stets konstruktive Zusammenarbeit. Dr. Raazesh Sainudiin und die anonymen Gutachter der drei Publikationen trugen mit ihrer Kritik und ihren Anregungen zur Verbesserung der Manuskripte bei.

Allen meinen Kollegen aus dem IMIS und dem Zentrum für Klinische Studien Kiel danke ich für die gute Zusammenarbeit und die angenehme Atmosphäre. Insbesondere danke ich Elfriede Fritzer, die in einer für mich kritischen Phase Lehrverpflichtungen übernommen hat, Olaf Junge, der jedes meiner technischen Probleme lösen konnte, sowie Petra Neumann, die in allen administrativen Fragen stets den Überblick behalten hat.

Meine Eltern Ursula Jochens und Holger Jochens haben mich immer unterstützt, dafür bin ich ihnen sehr dankbar. Und schließlich gilt mein ganz besonderer Dank Vera Elisabeth Jochens und Dr. Elén Jochens, die so viel Geduld und Verständnis gezeigt haben.

Eidesstattliche Erklärung

Ich versichere an Eides statt, dass die vorliegende Dissertation — abgesehen von der Beratung durch Herrn Prof. Dr. Krawczak — nach Inhalt und Form meine eigene Arbeit ist. Die Arbeit hat weder ganz noch zum Teil einer anderen Stelle im Rahmen eines Prüfungsverfahrens vorgelegen, und sie wurde — abgesehen von den drei gekennzeichneten Publikationen — noch nicht anderweitig veröffentlicht oder zur Veröffentlichung eingereicht. Die Arbeit ist unter Einhaltung der Regeln guter wissenschaftlicher Praxis der Deutschen Forschungsgemeinschaft entstanden.

_____

Arne Jochens