DISSERTATION

Mikrosatelliten – Ein Werkzeug zur genetischen Analyse. Grundlagen und Anwendung zur Suche nach genetischer Adaption.

Ausgeführt zum Zwecke der Erlangung des akademischen Grades eines Doktors der technischen Wissenschaften unter der Leitung

von

Univ. Prof. Dipl.-Ing. Dr. techn. Christian P. Kubicek
E1665
Gentechnik und angewandte Biochemie

eingereicht an der Technischen Universität Wien
Fakultät für Technische Naturwissenschaften und Informatik

von

Dipl.-Ing. Daniel Dieringer
9326646
1200 Wien, Otto-Probststraße 25

Wien, am 17.05.04

Daniel Dieringer

1. Zusammenfassung Deutsch

Evolution ist schon seit jeher ein Thema, das die Menschen nicht nur fasziniert und gespalten hat, sondern das auch mit fast jeder Antwort neue Fragen aufwirft. Eines der großen Probleme, die die Erforschung der Entwicklung der Arten mit sich bringt ist es, daß der Forscher sich dieser Materie wie ein Archäologe nähern muß. Wie alle „historischen" Wissenschaften ist auch die Evolutionsbiologie darauf angewiesen die Prozesse, die sie zum Untersuchungsgegenstand hat, anhand der Spuren die diese hinterlassen haben zu analysieren und zu charakterisieren. Die zentrale Frage ist es von der basalen Einheit der Evolution, der Nukleotidsequenz, eine möglichst genaue Darstellung wie sich die Arten entwickelt haben zu extrahieren. Mit Hilfe dieses Wissens ist es dann möglich die Modelle zu testen, mit deren Hilfe Evolutionsvorgänge beschreiben werden und damit wiederum das Wissen über reale Entwicklung der Arten selber zu verbessern. Da unsere Modelle (fast) nur auf Daten der heutigen Arten aufbauen, werden Unsicherheiten über die Vorgänge in der Vergangenheit immer vorhanden sein.
Gerade dies mach die Faszination dieser Materie aus, die sich mit einem der zentralsten Themen des Lebens beschäftigt: Woher kommen wir und wohin gehen wir.
Die Auswirkung von natürlicher Selektion als ein Mechanismus in der Entwicklung von Spezies ist seit Kimura 1968 seine „neutrale Theorie der Evolution" formulierte sehr umstritten (Kimura 1983). Seither wird versucht mittels empirischer Beweise zu klären, ob neutrale Modelle den genetischen genetische Variabilität von Populationen und die genetische Differenz zwischen den Arten erklären können, oder ob dies Modelle besser können, die Selektion implementieren.
Eine Methode diese Frage zu Untersuchen ist es neutrale genetische Marker zu verwenden. Mikrosatelliten sind solche Marker, die sehr häufig verwendet werden, so auch in den in dieser Arbeit präsentierten Studien. Diese werden durch Wiederholung kurzer Sequenzteile gebildet (Charlesworth 1998), sind leicht zu detektieren und haben eine hohe Variabilität. Eine grundsätzliche Eigenschaft von Mikrosatelliten ist, daß sie als neutral betrachtet werden können (Charlesworth 1998; Schlötterer and Wiehe 1999). Zu erwähnen ist jedoch, daß dies nicht generell gilt, da Mikrosatelliten im Erbgut auch funktionelle Aufgaben haben können (Li *et al.* 2002a) und ebenso an der Entstehung von Erbkrankheiten beteiligt sind (Sutherland and Richards 1995). Aus diesen Gründen ist die Entstehung und der Mutationsmechanismus von Mikrosatelliten (siehe auch Kapitel 3.1) per se ein interessantes Studiengebiet und wir konnten zeigen, daß

Mikrosatelliten sich schon bei deren Entstehung nicht so neutral verhalten, wie bisher zumeist angenommen (Kapitel 4.1).
Die weiteren hier präsentierten Arbeiten untersuchen wie die kosmopolitische Fruchtfliege *Drosophila melanogaster* ausgehend von Afrika den Rest der Welt besiedelte. Derzeit wird allgemein angenommen, daß die Art *D. melanogaster* etwa vor 2,5 Mio. Jahren im äquatorialen Afrika entstanden ist und vor etwa 10.000 Jahren, assoziert mit neolithischen Ackerbauern, zunächst Asien und Europa und schließlich in historischer Zeit die Neue Welt besiedelte (David and Capy 1988; Lachaise *et al.* 1988). Zwei Fragestellungen die diese Besiedelung aufwirft, werden in dieser Arbeit näher beleuchtet. Erstens stellt sich die Frage wie sich die Art in Afrika selber entwickelt hat und wie sie mit den unterschiedlichen klimatischen Bedingungen und den bekannten historischen Klimaveränderungen innerhalb Afrikas fertig geworden ist (Kapitel 4.3) und zweites, wie sie sich an die veränderten Bedingungen außerhalb diese Kontinentes angepaßt hat (Kapitel 4.2). Bei dem Emigration aus Afrika ging durch die geringe Populationsgröße in den Gründerpopulationen ein Teil der afrikanischen Variabilität verloren. Jedoch hat auch die genetische Anpassung an die veränderten ökologischen Rahmenbedingungen ihre Spuren im Genom der nichtafrikanischen *D. melanogaster* Populationen hinterlassen (Kauer *et al.* 2002). Folglich ist es von Interesse genomische Regionen zu kartieren, die von positiver Selektion in den nichtafrikanischen Populationen der Fruchtfliege beeinflußt wurden. Im Vergleich zum Rest des Genoms, haben diese Regionen eine besonders niedere genetische Variabilität, bzw. ist der Verlust an Variabilität im Vergleich zu den „ursprünglichen" afrikanischen Populationen besonders groß (Kapitel 3.2.4; 4.2; und Harr *et al.* 2002). Wenn man diesen Unterschied als Signal für positive Selektion verwenden will, ist es wichtig, daß man auch die „neutralen" afrikanischen Populationen studiert. Die falsche Wahl einer Referenzpopulation könnte das Ergebnis in nicht vorhersagbarer Weise beeinflussen. Zu diesem Zwecke befaßt sich ein Teil der in dieser Arbeit präsentierten Studien mit der Populationsstruktur von *D. melanogaster* innerhalb Afrikas (Kapitel 4.3).

Summary:
Evolution is a topic, that fascinated the mankind, as it splits the opinions. The main problem in study the development of species is that the researcher has to do this like an archaeologist. Like all 'historical' science also the evolution biology has to analyse and characterise the processes from the traces they left behind. The central question is to extract from the basal unit – the sequence – an accurate description of the species evolution. This knowledge then can be used to test the models which describes evolutionary processes.
Since Kimura's 'neutral theory of evolution' is the impact of natural selection as a mechanism in the evolution of species hotly debated (Kimura 1983). Therefore it is one of the most important aims to come up with a empirical evidence if neutral or selection based models can explain the pattern of genetic variation within populations or the divergence between species.
An important tool to solve this quenstion are microsatellites as genetic markers. They are repetitive arranged short sequence elements (Charlesworth 1998) and have the advantage that they are easy to detect and highly variable. An additional advantage is that they are thought to be neutral (Charlesworth 1998; Schlötterer and Wiehe 1999), but it has to be mentioned that microsatellites are known to have also putative functions (Li *et al.* 2002a) and are the source of some human genetic diseases (Sutherland and Richards 1995). Altogether are microsatellites (see also chapter 3.1) itself not only a important scientific tool, but also a interesting scientific topic and it shows, that their origin is not as neutral, as it was thought before (chapter 4.1).
In the other studies presented here we analysed how the cosmopolitan fruit fly *Drosophila melanogaster* has spread from its African habitat the rest of the world. There is the general accepted opinion, that the fruitfly has originated in equatorial Africa ~2.5 Mio. years ago and has colonised approximately 10,000 years ago Asia and Europe and in historical times also the New World (David and Capy 1988; Lachaise *et al.* 1988). There are two questions arising due to this colonisation. First how this species has developed in Africa and how has it adapted to the different climates and known historical climate changes within Africa (chapter 4.3) and second what adaptations was necessary for successfully colonisation of the rest of the world (chapter 4.2). During this emigration out of Africa the genomes lost variability due to a population bottleneck. Also positive and negative selection left its traces in the genome of non-African *D. melanogaster* populations during this adaptation (Kauer *et al.* 2002). Therefore it is of further interest to map the genomic regions where positive selection has occurred in non-African populations of *D. melanogaster*. In comparison to

the rest of the genome is the loss of variability between 'ancestral' African and 'derived' European populations remarkable high (chapter 3.2.4; chapter 4.2 and Harr *et al.* 2002). Using this difference between African and 'derived' populations as a sign for positive selection, then the correct selection of the African population chosen as a reference is very important. If the used reference is biased, then the results are changed, possible in a unpredictable way. To check for this we analysed also the population structure of *D. melanogaster* within Africa (chapter 4.3).

2. Inhaltsverzeichnis

3. Einleitung

Mikrosatelliten sind Elemente im Erbgut, die durch eine wiederholte Sequenzabordnung von 2 zwei bis drei Basen charakterisiert werden (Charlesworth 1998). Mikrosatelliten ändern aufgrund dieser speziellen Anordnung die Anzahl ihrer Bausteine. Der dafür verantwortliche Prozeß wird ‚replication slippage' genannt, da während der DNS Vervielfältigung die DNS-Polymerase einzelne Segmente ausläßt oder doppelt repliziert. Die dadurch erfolgende Längenveränderung erfolgt mit einer um 2 bis 3 Potenzen höheren Mutationsrate als die normale Punktmutation (Charlesworth 1998). Da diese Elemente meist als neutral betrachtet werden können (Charlesworth 1998; Schlötterer and Wiehe 1999), eignen sie sich durch ihre hohe Variabilität und der Tatsache das Längenänderungen leicht zu detektieren sind sehr gut als genomischen Marker. Die Anwendung von Mikrosatelliten als genetische Marker ist weit verbreitet und erfolgt etwa in Bereichen wie der Stammbaumforschung, der forensischer Analyse und der Populationsgenetik (Charlesworth 1998; Payseur *et al.* 2002; Schlötterer 2002b; Weissenbach 1993). Mikrosatelliten haben aber auch eine medizinische Relevanz, da ihre Hypervariabilität Erbkrankheiten wie das Huntington Syndrom auslösen kann.
Obwohl die Anwendung von Mikrosatelliten weit verbreitet ist, so ist bisher vor allem das Verhalten von langen Mikrosatelliten untersucht worden, da diese für die verschiedenen Anwendungen entscheidend sind (zusammengefaßt in Eisen 1999 und Li *et al.* 2002a). Wenig untersucht ist dagegen die Frage, ab welcher Anzahl an Sequenzwiederholungen ein Mikrosatelliten dessen typischen Eigenschaften – wie etwa die stark erhöhte Mutationsrate – aufweist. Meist wurde bisher von einer zufällige Entstehung ausgegangen (Levinson and Gutman 1987; Messier *et al.* 1996; Rose and Falush 1998; Stephan and Cho 1994), bei der durch Basenmutation eine ausreichend große Anzahl an wiederholten Einheiten entsteht. Einige Untersuchungen haben jedoch gezeigt, daß wahrscheinlich bereits kurze Repeats (2-3 fach) die Tendenz haben ihre Länge zu verändern (Nishizawa and Nishizawa 2002; Zhu *et al.* 2000b). Genauere Untersuchungen an einem großen Datensatz der nicht durch seine Herkunft (Zhu *et al.* 2000b) oder seine Beschränkung auf Pseudogene (Nishizawa and Nishizawa 2002) von bedingter Aussagekraft sind, gibt es über dieses Thema nicht.
Eine wichtige Anwendung von Mikrosatelliten als genomische Marker besteht im lokalisieren von adaptiver Selektion im Genom. Dies ist einerseits für die Grundlagenforschung interessant, um besser verstehen zu können wie sich Organismen und vor allem deren Gene an veränderte

Bedingungen anpassen. Dies ist aber letztlich auch in der angewandten Forschung interessant, wenn es darum geht durch gezielte Züchtungen oder gar künstlicher Veränderung der Keimbahn wirtschaftlich interessante Organismen in ihren Eigenschaften zu optimieren. Für die Etablierung solcher Methoden sind Modellorganismen nützlich, die leicht zu züchten und zu halten sind, eine kurze Generationszeit aufweisen und die bereits genetisch und biologisch gut charakterisiert sind. Einer dieser Organismen ist die Fruchtfliege *Drosophila melanogaster*. Sie stammt ursprünglich aus dem tropischen Afrika und hat vor etwa 10.000 Jahren den Rest der Welt kolonisiert (David and Capy 1988; Lachaise *et al.* 1988). Es ist anzunehmen, daß *Drosophila melanogaster* sich im Zuge dieser Kolonialisierung an die vollkommen andere Umweltbedingungen in den gemäßigten Klimazonen angepaßt haben muß. Daher sind funktionelle Unterschiede zwischen den Genomen afrikanischer und nichtafrikanischer Fliegen zu erwarten. Zusätzlich zu demographischen Faktoren, wie etwa dem Gründereffekt, führt die Auslese vorteilhafter Genvarianten in nichtafrikanischen Populationen zu einer Reduktion der Variabilität im Erbgut. Im Gegensatz zur globalen Reduktion der Variabilität durch demographische Effekte, führt Selektion nur zu einer lokal begrenzt Variabilitätsreduktion. Grundsätzlich kann die Auslese nur die Frequenzen der vorteilhaften Allele verändern, doch da sich diese schnell über eine Population ausbreiten können, werden auch benachbarte neutrale Regionen und Marker beeinflußt (siehe Kapitel 3.2.2).

Diese Auswirkungen auf die Variabilität kann dadurch mittels mehrerer Mikrosatelliten im Genom lokalisiert werden. Mehrere Mikrosatelliten sind nicht nur notwendig um lokale, durch Selektion bedingte, Reduktion zu finden, sondern auch um diese von der globalen, durch Demographie hervorgerufene, Variabilitätsreduktion unterscheiden zu können.

Für die von uns gewählte Strategie ist es vorteilhaft eine möglichst ursprüngliche Vergleichspopulation zu besitzen. Wie bereits erwähnt, hinterläßt nicht nur die genetische Adaption ihre Spuren im Genom, sondern auch neutrale Ereignisse wie Populationsreduktion, Expansion oder Migration. Vor allem letztere ist durch menschlichen Einfluß unter Umständen stark gestiegen und daher ist es von Interesse zu wissen, ob sich etwa der adaptierte und in seiner Variabilität reduzierte kosmopolitische Typ mit dem ursprünglichen afrikanischen Typ vermischt hat. Sollte dies in zu großem Ausmaß geschehen sein, würde eine Untersuchung solcher Populationen zu verzerrten Ergebnissen führen. Daher haben wir in der Studie im Kapitel 3.3 europäische und afrikanische Populationen untersucht, um dieses Problem abschätzen zu können.

3.1. Mikrosatelliten Entstehung und Mutationsprozeß

Wie bereits oben erwähnt, bestehen Mikrosatelliten aus sich wiederholenden kurzen Sequenzenelementen (Tautz 1989; Tautz and Renz 1984) wie zum Beispiel dem Element GA. Wenn man die Anzahl aller Elemente einer gewissen Länge in einem Genom zählt, wird man ab einer gewissen Länge feststellen, daß diese Anzahl weit größer ist, als aufgrund der Nukleotidzusammensetzung zu erwarten wäre (Epplen *et al.* 1993; Pupko and Graur 1999; Rose and Falush 1998; Tautz and Renz 1984). Bell (1996) schlug vor, daß der Überschuß und die Längenverteilung der Mikrosatelliten von einem Prozeß hervorgerufen wird, der die Anzahl der sich wiederholenden Elemente der Mikrosatelliten zufällig um eins erhöht oder erniedrigt. Dieser Prozeß beruht darauf, daß die langen Mikrosatelliten durch Fehlpaarung der DNS Stränge bei der Vervielfältigung, (zusammengefaßt in Eisen 1999 und Li *et al.* 2002a), ihre Länge ändern ≙ ‚replication slippage'.
Ebenfalls ist bekannt, daß die Mutationsrate von der Länge abhängt und für kürzere Wiederholungen kleiner ist als für längere (Brinkmann *et al.* 1998; Kruglyak *et al.* 1998). Dies ist eine Erklärung dafür, daß es weit mehr lange Mikrosatelliten im Erbgut gibt als erwartet, soweit diese nicht durch Selektion wieder entfernt werden, wie es vornehmlich in den Genomen der Prokaryonten der Fall ist (Field and Wills 1998). Direkte Studien zeigten, daß Mutationen in Mikrosatelliten häufiger in langen Wiederholungen auftritt (Brinkmann *et al.* 1998). Bei einigen Studien wurde eine ‚optimale' Allellänge feststellen, was heißt, daß Verlängerung und Verkürzung nicht gleich wahrscheinlich sind, sondern daß ab einer bestimmten Länge die Mikrosatelliten mit höherer Wahrscheinlichkeit verkürzt werden als verlängert (Dermitzakis *et al.* 1998; Garza *et al.* 1995; Li *et al.* 2002b; Samadi *et al.* 1998).
Was geschieht aber mit der Mutationsrate eines Mikrosatelliten der immer kürzer wird? Entweder wird die (slippage) Mutationsrate μ_s so klein, daß sie nicht mehr detektierbar ist, aber immer noch die Dichte/Anzahl an Mikrosatelliten im Genom bestimmt, oder sie (μ_s) wird schon bei größeren Längen null und macht für diese Längenklasse von Mikrosatelliten das ‚replication slippage' unwirksam. Generell gilt, daß sehr kurze Mikrosatelliten nur wenig Polymorphismus aufweisen und dieser mit der Anzahl der Wiederholungen steigt (Strassmann *et al.* 1997; Weber 1990; Zhu *et al.* 2000a; Zhu *et al.* 2000b). Rose und Falush (1998) beobachteten in Hefe, daß unterhalb einer Anzahl von fünf Wiederholungen die erwartete Anzahl an Mikrosatelliten mit der beobachteten Anzahl übereinstimmt, während Pupko und Graur (1999) mit dem selben Datensatz zu einem

konträren Ergebnis kommen, nämlich daß auch sehr kurze Mikrosatelliten überrepräsentiert sind und es daher keine Längengrenze gibt, unterhalb derer Mikrosatelliten mit einem anderen Modell beschrieben werden müßten.
Weitere Untersuchungen verwendeten ein einfaches Modell, in dem Slippagemutationen die Mikrosatelliten um eine Wiederholung verlängern oder verkürzen und Punktmutationen diese zerstören/verkürzen je nachdem wo sie im Mikrosatelliten auftreten. Anhand dieses Modells konnte gezeigt werden, daß die Mutationsrate μ_s direkt proportional zur Länge des Mikrosatelliten ist (Kruglyak *et al.* 2000; Kruglyak *et al.* 1998). Es zeigte sich jedoch, daß dies nur auf längere Wiederholungen zutrifft, während die kürzeren Mikrosatelliten (unterhalb von etwa fünf Wiederholungen) deutlich von den Erwartungen des Modells abweichen. Mit einem erweiterten Modell, bei dem die Wahrscheinlichkeit der Expansion oder Kontraktion nicht gleich sein muß, kamen auch Lai und Sun (2003) zu dem Ergebnis, daß unterhalb eines Grenzwertes die Mikrosatelliten nicht mehr mittels ‚Slippagemutationen' ihre Länge ändern können. Zuvor haben Bachtrog *et al.* (1999) einen signifikanten, aber nicht hundertprozentigen Zusammenhang von der $(AT/TA)_n$-Mikrosatellitendichte mit dem AT Gehalt entdeckt, was ebenfalls darauf hinweist, daß die Mikrosatellitenentstehung ein zufälliger Prozeß ist, der vor allem von Punktmutationen dominiert ist.
Wenn der Prozeß des ‚replication slippage' eine Mindestanzahl an sich wiederholenden Einheiten benötigt, dann müssen andere Mechanismen für die Entstehung und frühe Weiterentwicklung von Mikrosatellitenloci unterhalb dieses Grenzwertes verantwortlich sein. Eine mögliche Hypothese ist, das andere zufällige Mutationen, wie etwa Basensubstitution, die nötige Anzahl an sich wiederholenden Einheiten bildet (Levinson and Gutman 1987; Messier *et al.* 1996; Rose and Falush 1998; Stephan and Cho 1994).
Trotz allem stellt sich die Frage, ob die Mutationsrate μ_s für einzelne Mikrosatelliten nicht einfach nur zu klein wird um sie direkt messen zu können, wenn die Anzahl an Wiederholungen klein wird. Auch wenn die Mutationsrate μ_s unter die Detektionsgrenze fällt, kann der Prozeß des ‚replication slippage' entscheidend sein bei der Mikrosatellitenentstehung, da die Anzahl der kurzen Wiederholungen sehr groß ist. Ein Hinweis wie man sich die ‚Geburt' von sich wiederholenden Elementen vorstellen kann, gibt eine Studie über Minisatelliten. Minisatelliten sind Elemente wie Mikrosatelliten, aber mit längeren DNS-Einheiten als diese. Bei den Minisatelliten ist es viel unwahrscheinlicher als bei den Mikrosatelliten, daß sich zufällig – also durch Basenmutation – eine ausreichend große Anzahl an sich perfekt wiederholenden Elementen bildet, was nach dem

bisherigen Modell der Mikrosatellitenentstehung notwendig wäre. Bei Minisatelliten zumindest scheint jedoch schon eine einzige nicht perfekte Wiederholung unter Umständen auszureichen, um einen Minisatelliten zu generieren (Taylor and Breden 2000), also entweder die Wiederholung nicht komplett ist, oder auch eine geringfügig andere Sequenz aufweist.
Durch die niedere postulierte Mutationsrate der kurzen Mikrosatelliten ist es ein Problem diese durch direkte Studien zu überprüfen, da bei der geringen Anzahl der mit akzeptablen Aufwand testbaren Individuen und Mikrosatelliten es nicht zu erwarten wäre brauchbare Resultate zu erhalten. Eine Möglichkeit dieses zu umgehen ist es die Variabilität innerhalb einer Art zu beobachten. Dies hat den Vorteil, daß man damit viele Generationen ‚überblickt' und damit die Wahrscheinlichkeit, daß selbst bei niedriger Mutationsrate Mutationen detektiert werden können. Zhu *et al.* (2000b) verwendeten zu diesem Zweck Daten der ‚Human Gene Mutation Database'. In dieser sind Sequenzen für viele Mutation die ein bekanntes Krankheitsbild hervorrufen gespeichert. Anhand diesen Datensatzes konnten sie zeigen, daß 70% aller Insertionen Verdoppelungen von benachbarten Basen sind. Weiters sind 55% aller Dinukleotid-, 68% aller Trinukleotid- und 92% aller Tetranukleotidinsertionen Kopien der angrenzenden Sequenz. Eine weitere Studie untersuchte die Sequenzen von Pseudogenen und analysierte mittels bayes'scher Statistik, welche von den Mutationsmechanismen, die die beobachteten Unterschiede bewirken hätten können, jene mit der höchsten Wahrscheinlichkeit sind (Nishizawa and Nishizawa 2002). Auch diese Studie kommt zum Schluß, daß bereits sehr kurze Mikrosatellitenelemente eine hohe Instabilität aufweisen. Bisher fehlt jedoch eine Studie, die diese Frage auf Basis der nunmehr erhältlichen ganzen Genomsequenzen zu beantworten sucht.

3.2. Positive Darwin'sche Selektion und Methoden um diese in molekularen Daten zu detektieren

3.2.1. Positive Darwin'sche Selektion

Mit Kimuras (1983) Theorie der neutralen Evolution liegt eine klare Erwartung über das Schicksal von neutralen Mutationen in Populationen vor. Die grundlegenden Prinzipien sind etwa, daß die Wahrscheinlichkeit, daß eine neutrale Mutation in einer Population fixiert wird, also das alle Individuen einer Population nur mehr dieses eine Allel besitzen, proportional zur ursprünglichen Allelfrequenz ist. Für alle durch Mutation in einer Population entstandenen Allele ist diese Frequenz 1/ (2 N), wobei dabei N die haploide Populationsgröße ist. Weiters ist die Menge an Substitutionen von neutralen Allelen pro Generation (K) gleich der neutralen Mutationsrate, also gleich der Anzahl an neutralen Mutationen die in der Population auftreten (μ). Die Zeit die man zwischen zwei Fixierungen von neutralen Allelen warten muß (t_w) ist indirekt proportional zu dieser Mutationsrate (μ) $t_W = 1 / \mu$ und die Zeit die eine neutrale Mutation braucht um fixiert zu werden ist direkt proportional zur effektiven Populationsgröße N_e $t_f = 4 N_e$.
Diese Informationen betreffen nur neutrale Varianten, d.h. Mutationen, die keine Auswirkung auf die Fitneß der betroffenen Individuen haben. Für die Suche nach nicht-neutralen, postitiven Mutationen (die einen selektiven Vorteil haben), ist es notwendig eine genaue Vorstellung darüber zu besitzen, wie sich neutrale Allele verhalten, da dieses als Nullmodell für den Vergleich mit den positiv selektieren Mutationen dienen kann. Durch Selektion wird die Fixierungszeit eines vorteilhaften Alleles in einer Population auf $t_f = 2 \ln (2 N_e) / s$ (Li 1997) verkürzt. Bei dem in dieser Studie verwendeten Modellorganismus *D. melanogaster*, kann die effektive Populationsgröße N_e auf etwa 10^6 abgeschätzt werden (Andolfatto and Przeworski 2000). Mit geschätzten fünf Generationen pro Jahr, braucht ein neutrales Allel etwa 800.000 Jahre, um in einer solchen Population fixiert zu werden. Ein Allel mit einen selektiven Vorteil von nur 1‰ besitzt braucht im Vergleich dazu nur etwa 5800 Jahre bis zur Fixierung.
Da dieser Prozeß so schnell geschieht wird im englischen Fachjargon davon gesprochen, daß diese Allele durch die Population ‚fegen' und man nennt diesen Prozeß daher ‚selectiv sweep'. Diese Art der Selektion wird auch gerichtete Selektion genannt im Gegensatz zur stabilisierenden oder balancierenden Selektion wo zwei oder mehrere Allele ihre Frequenz nicht verändern, da dieser Gleichgewichtszustand einem Optimum an Fitneß entspricht (Li 1997).

3.2.2. Der 'Hitchhiking' Effekt

Wie oben erwähnt geschieht die Fixierung eines positiv selektiertes Allel sehr rasch, doch durch diese Geschwindigkeit kommt es zu einem Nebeneffekt, den ‚Anhaltereffekt' (engl. hitchhiking). Dieses Bezeichnung ist korrekt, da nicht nur das selektierte Allel fixiert wird, sondern auch die dieses umgebende neutrale Sequenz (Kaplan *et al.* 1989; Maynard Smith and Haigh 1974). Dieser Effekt ist für die Detektion von ‚selective sweeps' von großer Bedeutung, da er das Phänomen beschreibt, daß Selektion nicht nur das positive Allel selbst fixiert oder zumindest zu einer hohen Frequenz bringt, sondern auch Allele neutraler Loci die in der Nachbarschaft liegen. Wie groß diese Region ist, hängt von der Rekombinationsrate, der Selektionsrate und der Zeit seit dem ‚selective sweep' ab (Kim and Stephan 2000; Wiehe and Stephan 1993). Die Frage wie schnell und wie weit sich die invariable Region rund um das positiv selektierte Allele ausbreiten kann, hängt also von der Zahl an ‚meiotisches crossing overs' ab und wie stark der Selektionsdruck ist. Jede Variabilität eines neutralen Locus, die nicht mit dem positiven Allele gelinked ist, wird entfernt, sofern es die Rekombination nicht gelingt genügend neutrale Varianten an das positive Allel anzuhängen. Daher ist die Region umso kleiner, je häufiger Rekombination auftritt und je schwächer der Selektionsdruck ist. Problematisch wird dieser Sachverhalt daher dann bei Regionen in denen keine oder nur sehr wenig Rekombination vorkommt, wie etwa in Mitochondrien, Chloroplasten, in Teilen des Y-Chromosomes, den Centromeren- oder den Telomeren-Regionen von Chromosomen, denn dort entfernt ‚hitchhiking' jeden Polymorphismus. Für diese Regionen, aber auch generell gilt, das deren Sequenz ident ist mit der, die das vorteilhafte Allel bei dessen Entstehung umgeben hat. Jeder Polymorphismus innerhalb dieser Sequenzteile muß daher erst nach dem ‚selective sweep' durch neue Mutationen entstanden sein. Diese neuen Mutationen entfernen daher mit der Zeit alle Anzeichen des ‚hitchhiking' aus der Sequenz (Kim and Stephan 2000; Nielsen 2001; Wiehe and Stephan 1993).
Solange jedoch die neue Mutationen den natürlichen Grad an neutralem Polymorphismus noch nicht wieder hergestellt haben, können neutrale Marker jeder Art (z.B.: Mikrosatelliten, SNPs, RFLPs) verwendet werden um diese ‚hitchhiking' Region im Genom zu detektieren.

3.2.3. Nachweis der positiven Selektion auf dem molekularen Niveau

Es gibt unterschiedliche Methoden um Selektion auf der molekularen Ebene nachzuweisen. Die Methoden können grob in zwei Gruppen eingeteilt werden:

- Direkte Tests. Diese testen direkt die möglichen Ziele von positiver Selektion, also vor allem für Aminosäuren codierende Regionen (Hurst 2002; Nei and Kumar 2000; Yang and Bielawski 2000)
- Indirekte Test, die auf dem oben beschriebenen ‚hitchhiking' Effekt aufbauen und die daher auch mit nicht-kodierende Regionen verwendet werden können (Andolfatto 2001a; Fu 1996; Nielsen 2001; Otto 2000; Schlötterer 2002a; Schlötterer 2002c; Wall and Hudson 2001)

Natürlich werden auch Aminosäuren, die in der Nähe der selektierten Position liegen ebenso durch ‚hitchhiking' beeinflußt wie die nicht-kodierenden Regionen und daher ist der Unterschied in der Praxis meist nicht so klar wie in der Theorie und daher die Unterscheidung vor allem eine Frage des Konzeptes und der Betrachtungsweise und für die Formulierung der theoretischen Grundlagen eines Nachweises von Bedeutung.
Für die hier betrachteten Studien sind Mikrosatelliten Loci verwendet worden, d.h. neutrale Marker die in nicht kodierenden Regionen liegen und daher sind hier nur die indirekten Methoden der Selektionsdetektion von Bedeutung. Natürlich muß nach der Entdeckung einer selektierten Region, diese mittels ‚Feinmapping' und der Verwendung von direkten Methoden auf Genebene bestätigt und näher untersucht werden (Harr *et al.* 2002).
Um nun mittels Polymorphismusdaten von nicht-kodierender DNS einen indirekten Test für Selektion zu entwickeln, der auf dem Effekt des ‚hitchhikings' basiert, ist es – wie bereits oben erwähnt – notwendig die neutrale Erwartung von der Erwartung unter Selektion unterscheiden zu können. Es existieren inzwischen verschiedene Theorien darüber, wie sich die Variablität eines neutralen DNS-Segment verändert, wenn dieses durch positive Selektion in einem benachbarten Locus beeinflußt wird, also durch ‚hitchhiking' (Andolfatto 2001a; Fu 1996; Nielsen 2001; Otto 2000; Schlötterer 2002c; Wall and Hudson 2001).
Die Erwartungen können grob in zwei Gruppen eingeteilt werden:

- Effekte unter positiver Selektion die das ganze Genom beeinflussen.
- Effekte die nur bestimmte Regionen in einem Genom betreffen.

Erstere entstehen durch wiederholte positiv selektierte Mutationen, die zufällig über das ganze Genom verteilt sind und es bildet sich dann einen Gleichgewichtszustand, unter dem folgendes erwartet werden kann:

- Autosomen und Geschlechtchromosome haben ein unterschiedliches Polymorphismusniveau (Aquadro *et al.* 1994; Begun and Whitley 2000; Orr and Betancourt 2001).
- Die Anzahl an Rekombinationsereignissen in einem Chromosomenabschnitt ist direkt proportional zu dem Polymorphismus der DNS (Begun and Aquadro 1992).

Für eine bestimmte Region, die unter Umständen nur von einer einzelnen vorteilhaften Mutation beeinflußt wird – siehe aber auch oben – gelten folgende Erwartungen:

- Erhöhtes ‚Linkage Disequilibrium' (Andolfatto and Przeworski 2000; Depaulis 1998; Pritchard and Przeworski 2001; Przeworski and Wall 2001; Przeworski *et al.* 2001).
- Erhöhte genetische Distanz (Akey *et al.* 2002; Charlesworth 1998).
- Das Spektrum der Allelfrequenzen weist eine Verschiebung auf:
  - Verschiebung zu einer größeren Anzahl von niederfrequenten Allelen. (Braverman *et al.* 1995; Fu and Li 1993; Simonsen *et al.* 1995; Tajima 1989)
  - Ebenfalls eine größere Anzahl an von der ursprünglichen Population abgeleiteten (‚derived') hochfrequenten Allele (Fay and Wu 2000; Przeworski 2002).
- Lokal ist der Grad an Polymorphismus reduziert
  - im Vergleich mit einer angrenzenden Region (Kim and Stephan 2002; Wiehe 1998).
  - im Vergleich zu einem Satz an Reverenzloci im gleichen Genom (Galtier *et al.* 2000; Lewontin and Krakauer 1973; Schlötterer 2002b).
  - im Vergleich zu dem Unterschied zwischen Spezies (‚divergence') (Hudson *et al.* 1987).

Alle der oben beschriebenen Möglichkeiten haben ihre Vorteile und Nachteile, die in einer Reihe von Zusammenfassungen beschreiben worden sind (Andolfatto 2001a; Nielsen 2001; Otto 2000; Przeworski 2002; Schlötterer 2002c). Bei der Verwendung von indirekten Neutralitätstests ist bei den Analysen immer das Problem zu berücksichtigen, wie die demographische Geschichte der Populationen berücksichtigt wird. Denn diese kann oft ebenfalls die unter Selektion erwarteten Effekte hervorrufen. Geringe Variabilität, wie sie durch einen positive Selektion erwartet wird, wird ebenfalls erzeugt, wenn die Population eine Zeit mit starker

Größenreduktion (‚bottleneck') erlebt hat. Ebenso kann schnelles Populationswachstum ebenso einen Überschuß an seltenen Allelen hervorrufen wie positive Selektion. Der Unterschied zwischen diesen beiden Möglichkeiten ist, daß Demographie immer das ganze Genom beeinflußt, Selektion aber immer nur eine eingeschränkte Region (siehe auch Kapitel 3.2.2). Daher kann man versuchen auf diese Weise die beiden Möglichkeiten zu unterscheiden, denn wenn eine große Anzahl an Loci die gleiche Abweichung von der Normalität aufweisen, dann kann dies als Evidenz für ein demographisches Ereignis angesehen werden. Wenn dagegen nur eine geringe Anzahl an Loci eine starke Abweichung von der Normalität (etwa Mangel an Variabilität oder einen Überschuß an seltenen Allelen) aufweisen, dann spricht dies dafür, daß diese durch Selektion beeinflußt wurden. Wie oben angeführt, sind daher eine Reihe von Tests entwickelt worden, die individuellen Loci gegen einen Satz an Referenzloci im Genom testen, um die Regionen zu finden, die auf Selektion hindeuten. Als Anmerkung soll noch einmal die bereits angeführte Möglichkeit erwähnt werden, daß mehrere unabhängige ‚Sweeps' auch ein ganzes Chromosom/Genom beeinflussen können, wenn die Anzahl groß genug ist und der Zeitraum ausreicht (siehe oben).

3.2.4. Erwartungen von positiver Selektion basierend auf Polymorphismusdaten

Die oben genannten Methoden wurden von diversen Studien verwendet um nach Selektion im Genom zu suchen. Wie in Kapitel 3.2.3 bereits erwähnt kann jedoch auch Demographie die Suche nach der Selektion beeinflussen. Vor allem wenn versucht wird den Anteil an Regionen des Erbgutes zu errechnen, die von positiver Selektion beeinflußt worden sind, gilt es die möglichen Verfälschungen zu beachten und wenn möglich zu korrigieren (Andolfatto 2001a; Schlötterer 2002c).
Beispielsweise zeigte eine Studie von Sequenzpolymorphismen (29 Loci) in afrikanischen *D. melanogaster* Populationen, daß in Regionen mit geringer Rekombination zu einer Verschiebung zu mehr seltenen Varianten gekommen ist (Andolfatto and Przeworski 2001). Dies kann am besten durch ‚hitchhiking' erlärt werden und zeigt eine wichtige Rolle von positiven Mutationen in *D. melanogaster* hin. Dies ist ähnlich zu den Ergebnissen einer vorhergehenden Studie, bei denen sowohl in 24 Genen von *D. melanogaster* also auch in 16 Genen von *D. simulans* ein Überschuß von ‚Linkage Disequilibrium' detektiert worden ist (im Vergleich zu der neutralen Erwartung) (Andolfatto and Przeworski 2000). Erklärt werden kann dies durch komplexe demographische Szenarien oder

durch einen genomweiten Einfluß von positiver Selektion. Auch in menschlichen Daten konnte ähnliches gefunden werden, wobei diese Studien eine weit größere Anzahl an Loci verwendeten. Mit 5048 Mikrosatelliten fanden Huttley *et al.* (1999) das in europäischen Populationen von Menschen ein Überschuß an kurzem ‚Linkage Disequilibrium' herrscht. Mit 5257 Mikrosatelliten, ebenfalls im menschlichen Erbgut, fanden Payseur *et al.* (2002) das sich das Frequenzspektrum in Richtung rare Allele verschiebt. Besonders häufig konnte dies in Regionen mit niederer Rekombinationsrate beobachtet werden, und dies kann nicht durch eine einfache Populationsexpansion erklärt werden. Dieser Überschuß an seltenen Allelen ist auf dem X Chromosom gehäuft zu beobachten und ebenfalls ist dort die Variabilität niederer als auf den Autosomen. Generell wird auf dem X-Chromosom eine niedere Variabilität im Mensch gefunden (Aquadro *et al.* 2001; Nachman 2001; Sachidanandam *et al.* 2001) ebenso wie in *D. melanogaster* (Andolfatto 2001b; Aquadro *et al.* 1994), *D. simulans* (Begun and Whitley 2000) und Mäusen (Hedrick and Parker 1997). In einer anderen Studie in *D. melanogaster* wurde für das X-Chromosom in nichtafrikanischen Populationen im Vergleich zu den ursprünglichen afrikanischen Populationen ein deutlich stärkerer Verlust an Variabilität gefunden als bei den Autosomen und der Verlust ist korreliert mit der Rekombinationsrate (Kauer *et al.* 2002). Wie bereits bei den anderen Studien, wurde diese Korrelation als Zeichen von positiver Selektion in dieser Spezies gewertet, in diesem Fall als ‚Begleitprozeß' der Kolonisation und Anpassung an neue Lebensräume.
Genomweite Untersuchung von positiver Selektion im menschlichen Erbgut identifizierten 174 Kandidatengene unter der Verwendung von Populations-Differenzierungsdaten (Akey *et al.* 2002) und 89 Gene unter der Verwendung von SNP (single nucletide polymorphism) Polymorphismusdaten (Diller *et al.* 2002). Unter der Verwendung der Methode des ‚hitchhiking mapping' (Harr *et al.* 2002) wurden zwei Kandidatenregionen für Selektion im Mensch gefunden (Schlötterer 2002b). In Kapitel 4.2 wurde die erste genomweite Untersuchung nach positiver Selektion in *D. melanogaster* präsentiert und 36 Regionen gefunden, die Kandidaten für positive Selektion sind.

3.3. Literatur

Akey, J. M., G. Zhang, K. Zhang, L. Jin and M. D. Shriver, 2002 Interrogating a high-density SNP map for signatures of natural selection. Genome Res **12:** 1805-1814.

Andolfatto, P., 2001a Adaptive hitchhiking effects on genome variability. Curr Opin Genet Dev **11:** 635-641.

Andolfatto, P., 2001b Contrasting Patterns of X-Linked and Autosomal Nucleotide Variation in *Drosophila melanogaster* and *Drosophila simulans*. Mol Biol Evol **18:** 279-290.

Andolfatto, P., and M. Przeworski, 2000 A genome-wide departure from the standard neutral model in natural populations of Drosophila. Genetics **156:** 257-268.

Andolfatto, P., and M. Przeworski, 2001 Regions of lower crossing over harbor more rare variants in African populations of Drosophila melanogaster. Genetics **158:** 657-665.

Aquadro, C. F., V. Bauer DuMont and F. A. Reed, 2001 Genome-wide variation in the human and fruitfly: a comparison. Curr Opin Genet Dev **11:** 627-634.

Aquadro, C. F., D. J. Begun and E. C. Kindahl, 1994 Selection, recombination, and DNA polymorphism in *Drosophila*, pp. 46-56 in *Non-neutral evolution*, edited by B. Golding. Chapman & Hall, London.

Bachtrog, D., S. Weiss, B. Zangerl, G. Brem and C. Schlotterer, 1999 Distribution of dinucleotide microsatellites in the Drosophila melanogaster genome. Mol Biol Evol **16:** 602-610.

Begun, D. J., and C. F. Aquadro, 1992 Levels of naturally occurring DNA polymorphism correlate with recombination rates in *D. melanogaster*. Nature **356:** 519-520.

Begun, D. J., and P. Whitley, 2000 Reduced X-linked nucleotide polymorphism in *Drosophila simulans*. Proc. Natl. Acad. Sci. USA **97:** 5960-5965.

Bell, G. I., 1996 Evolution of simple sequence repeats. Computational Chemistry **20:** 41-48.

Braverman, J. M., R. R. Hudson, N. L. Kaplan, C. H. Langley and W. Stephan, 1995 The hitchhiking effect on the sites frequency spectrum of DNA polymorphisms. Genetics **140:** 783-796.

Brinkmann, B., M. Klintschar, F. Neuhuber, J. Huhne and B. Rolf, 1998 Mutation rate in human microsatellites: influence of the structure and length of the tandem repeat. Am J Hum Genet **62:** 1408-1415.

Charlesworth, B., 1998 Measures of divergence between populations and the effect of forces that reduce variability. Molecular Biology and Evolution **15:** 538-543.

David, J. R., and P. Capy, 1988 Genetic variation of *Drosophila melanogaster* natural populations. Trends in Genetics **4:** 106-111.

Depaulis, F., 1998 Neutrality tests based on the distribution of haplotypes under an infinite-site model. Molecular Biology and Evolution **15:** 1788-1790.

Dermitzakis, E. T., A. G. Clark, C. Batargias, A. Magoulas and E. Zouros, 1998 Negative covariance suggests mutation bias in a two-locus microsatellite system in the fish *Sparus aurata*. Genetics **150:** 1567-1575.

Diller, K. C., W. A. Gilbert and T. D. Kocher, 2002 Selective sweeps in the human genome: a starting point for identifying genetic differences between modern humans and chimpanzees. Mol Biol Evol **19:** 2342-2345.

Eisen, J. A., 1999 Mechanistic basis for microsatellite instability, pp. 34-48 in *Microsatellites: evolution and applications*, edited by D. Goldstein and C. Schlötterer. Oxford University Press, Oxford.

Epplen, C., G. Melmer, I. Siedlaczck and e. al., 1993 *On the essence of 'meaningless' simple repetitive DNA in eukaryote genomes.* Birkhäuser Verlag, Basel Switzerland.

Fay, J. C., and C. I. Wu, 2000 Hitchhiking under positive Darwinian selection. Genetics **155:** 1405-1413.

Field, D., and C. Wills, 1998 Abundant microsatellite polymorphism in Saccharomyces cerevisiae, and the different distributions of microsatellites in eight prokaryotes and S. cerevisiae, result from strong mutation pressures and a variety of selective forces. Proc Natl Acad Sci U S A **95:** 1647-1652.

Fu, Y.-X., 1996 New statistical tests of neutrality for DNA samples from a population. Genetics **143:** 557-570.

Fu, Y.-X., and W.-H. Li, 1993 Statistical tests of neutrality of mutations. Genetics **133:** 693-709.

Galtier, N., F. Depaulis and N. H. Barton, 2000 Detecting bottlenecks and selective sweeps from DNA sequence polymorphism. Genetics **155:** 981-987.

Garza, J. C., M. Slatkin and N. B. Freimer, 1995 Microsatellite allele frequencies in humans and chimpanzees with implications for constraints on allele size. Molecular Biology and Evolution **12:** 594-603.

Harr, B., M. Kauer and C. Schlötterer, 2002 Hitchhiking mapping- a population-based fine mapping strategy for adaptive mutations in

Drosophila melanogaster. Proc Natl Acad Sci U S A **99:** 12949-12954.

Hedrick, P. W., and J. D. Parker, 1997 Evolutionary genetics and genetic variation of haplodiploids and X-linked genes. Annu. Rev. Ecol. Syst. **28:** 55-83.

Hudson, R. R., M. Kreitman and Aguadé, 1987 A test of neutral molecular evolution based on nucleotide data. Genetics **116:** 153-159.

Hurst, L., 2002 The Ka/Ks ratio: diagnosing the form of sequence evolution. Trends Genet **18:** 486.

Huttley, G. A., M. W. Smith, M. Carrington and S. J. O'Brien, 1999 A scan for linkage disequilibrium across the human genome. Genetics **152:** 1711-1722.

Kaplan, N. L., R. R. Hudson and C. H. Langley, 1989 The "hitchhiking effect" revisited. Genetics **123:** 887-899.

Kauer, M., B. Zangerl, D. Dieringer and C. Schlötterer, 2002 Chromosomal patterns of microsatellite variability contrast sharply in African and non-African populations of Drosophila melanogaster. Genetics **160:** 247-256.

Kim, Y., and W. Stephan, 2000 Joint effects of genetic hitchhiking and background selection on neutral variation. Genetics **155:** 1415-1427.

Kim, Y., and W. Stephan, 2002 Detecting a local signature of genetic hitchhiking along a recombining chromosome. Genetics **160:** 765-777.

Kimura, M., 1983 *The neutral theory of molecular evolution.* Cambridge University Press, Cambridge.

Kruglyak, S., R. Durrett, M. D. Schug and C. F. Aquadro, 2000 Distribution and abundance of microsatellites in the yeast genome can Be explained by a balance between slippage events and point mutations. Mol Biol Evol **17:** 1210-1219.

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 U S A **95:** 10774-10778.

Lachaise, D., M.-L. Cariou, J. R. David, F. Lemeunier, L. Tsacas *et al.*, 1988 Historical biogeography of the *Drosophila melanogaster* species subgroup. Evol. Biol. **22:** 159-225.

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.

Levinson, G., and G. A. Gutman, 1987 Slipped-strand mispairing: a major mechanism for DNA sequence evolution. Mol Biol Evol **4:** 203-221.

Lewontin, R. C., and J. Krakauer, 1973 Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. Genetics **74:** 175-195.

Li, W.-H., 1997 *Molecular Evolution.* Sinauer Associates, Sunderland.

Li, Y. C., A. B. Korol, T. Fahima, A. Beiles and E. Nevo, 2002a Microsatellites: genomic distribution, putative functions and mutational mechanisms: a review. Mol Ecol **11:** 2453-2465.

Li, Y. C., M. S. Roder, T. Fahima, V. M. Kirzhner, A. Beiles *et al.*, 2002b Climatic effects on microsatellite diversity in wild emmer wheat (Triticum dicoccoides) at the Yehudiyya microsite, Israel. Heredity **89:** 127-132.

Maynard Smith, J., and J. Haigh, 1974 The hitch-hiking effect of a favorable gene. Genet. Res. **23:** 23-35.

Messier, W., S.-H. Li and C.-B. Stewart, 1996 The birth of microsatellites. Nature **381:** 483.

Nachman, M. W., 2001 Single nucleotide polymorphisms and recombination rate in humans. Trends Genet **17:** 481-485.

Nei, M., and S. Kumar, 2000 Molecular Evolution and Phylogenetics. Oxford University Press **pp. 51 ff.**

Nielsen, R., 2001 Statistical tests of selective neutrality in the age of genomics. Heredity **86:** 641-647.

Nishizawa, M., and K. Nishizawa, 2002 A DNA sequence evolution analysis generalized by simulation and the markov chain monte carlo method implicates strand slippage in a majority of insertions and deletions. J Mol Evol **55:** 706-717.

Orr, H., and A. Betancourt, 2001 Haldane's sieve and adaptation from the standing genetic variation. Genetics **157:** 875-884.

Otto, S. P., 2000 Detecting the form of selection from DNA sequence data. Trends Genet **16:** 526-529.

Payseur, B. A., A. D. Cutter and M. W. Nachman, 2002 Searching for evidence of positive selection in the human genome using patterns of microsatellite variability. Mol Biol Evol **19:** 1143-1153.

Pritchard, J. K., and M. Przeworski, 2001 Linkage disequilibrium in humans: models and data. Am J Hum Genet **69:** 1-14.

Przeworski, M., 2002 The signature of positive selection at randomly chosen loci. Genetics **160:** 1179-1189.

Przeworski, M., and J. D. Wall, 2001 Why is there so little intragenic linkage disequilibrium in humans? Genet Res **77:** 143-151.

Przeworski, M., J. D. Wall and P. Andolfatto, 2001 Recombination and the frequency spectrum in Drosophila melanogaster and Drosophila simulans. Mol Biol Evol **18:** 291-298.

Pupko, T., and D. Graur, 1999 Evolution of microsatellites in the yeast Saccharomyces cerevisiae: rold of length and number of repeated units. J Mol Evol **48:** 313-316.

Rose, O., and D. Falush, 1998 A threshold size for microsatellite expansion. Mol Biol Evol **15:** 613-615.

Sachidanandam, R., D. Weissman, S. C. Schmidt, J. M. Kakol, L. D. Stein *et al.*, 2001 A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature **409:** 928-933.

Samadi, S., F. Erard, A. Estoup and P. Jarne, 1998 The influence of mutation, selection and reproductive systems on microsatellite variability: a simulation approach. Genet. Res. Camb. **1998:** 213-222.

Schlötterer, C., 2002a Hitchhiking mapping - functional genomics from the population genetics perspective, Pages 32-38. Trends in Genetics **19:** 32-38.

Schlötterer, C., 2002b A microsatellite-based multilocus screen for the identification of local selective sweeps. Genetics **160:** 753-763.

Schlötterer, C., 2002c Towards a molecular characterization of adaptation in local populations. Curr Opin Genet Dev **12:** 683-687.

Schlötterer, C., and T. Wiehe, 1999 Microsatellites, a neutral marker to infer selective sweeps, pp. 238-248 in *Microsatellites - evolution and applications*, edited by D. Goldstein and C. Schlötterer. Oxford University Press, Oxford.

Simonsen, K. L., G. A. Churchill and C. F. Aquadro, 1995 Properties of statistical tests of neutrality for DNA polymorphism data. Genetics **141:** 413-429.

Stephan, W., and S. Cho, 1994 Possible role of natural selection in the formation of tandem-repetitive noncoding DNA. Genetics **136:** 333-341.

Strassmann, J. E., K. Barefield, C. R. Solis, C. R. Hughes and D. C. Queller, 1997 Trinucleotide microsatellite loci for a social wasp, Polistes. Mol Ecol **6:** 97-100.

Sutherland, G. R., and R. I. Richards, 1995 Simple tandem DNA repeats and human genetic disease. Proc Natl Acad Sci U S A **92:** 3636-3641.

Tajima, F., 1989 Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics **123:** 585-595.

Tautz, D., 1989 Hypervariability of simple sequences as a general source for polymorphic DNA markers. Nucleic Acids Research **17:** 6463-6471.

Tautz, D., and M. Renz, 1984 Simple sequences are ubiquitous repetitive components of eukaryotic genomes. Nucleic Acids Res **12:** 4127-4138.

Taylor, J. S., and F. Breden, 2000 Slipped-strand mispairing at noncontiguous repeats in Poecilia reticulata: a model for minisatellite birth. Genetics **155:** 1313-1320.

Wall, J. D., and R. R. Hudson, 2001 Coalescent simulations and statistical tests of neutrality. Mol Biol Evol **18:** 1134-1135; discussion 1136-1138.

Weber, J. L., 1990 Informativeness of human $(dC\text{-}dA)_n \cdot (dG\text{-}dT)_n$ polymorphisms. Genomics **7:** 524-530.

Weissenbach, J., 1993 Microsatellite polymorphisms and the genetic linkage map of the genome. Curr. Op. Gen. Dev. **3:** 414-417.

Wiehe, T., 1998 The effect of selective sweeps on the variance of the allele distribution of a linked multi-allele locus-hitchhiking of microsatellites. Theoretical Population Biology **53:** 272-283.

Wiehe, T. H., and W. Stephan, 1993 Analysis of a genetic hitchhiking model, and its application to DNA polymorphism data from *Drosophila melanogaster*. Molecular Biology and Evolution **10:** 842-854.

Yang, Z., and J. P. Bielawski, 2000 Statistical methods for detecting molecular adaptation. Trends in Ecology & Evolution **15:** 496-503.

Zhu, Y., D. C. Queller and J. E. Strassmann, 2000a A phylogenetic perspective on sequence evolution in microsatellite loci. Journal of Molecular Evolution **50:** 324-338.

Zhu, Y., J. E. Strassmann and D. C. Queller, 2000b Insertions, substitutions, and the origin of microsatellites. Genet Res **76:** 227-236.

4. Ausgeführte Studien während der Dissertation

4.1. Two distinct modes of microsatellite mutation processes-evidence from the complete genomic sequences of nine species

Publiziert in Genome Research **13**:*2242-2251*

4.2. A microsatellite variability screen for positive selection associated with the 'out of Africa' habitat expansion of *Drosophila melanogaster*

Publiziert in Genetics **165**:*1137-1148*

4.3. Population structure in African *D. melanogaster* revealed by microsatellite analysis

Review bei Genetics

4.4. Allele excess at neutrally evolving microsatellites and the implications for tests of neutrality

Publiziert in Proc. R. Soc. B **271**: 869-874

Two Distinct Modes of Microsatellite Mutation Processes: Evidence From the Complete Genomic Sequences of Nine Species

Daniel Dieringer and Christian Schlötterer[1]

Institut für Tierzucht und Genetik, 1210 Wien, Austria

We surveyed microsatellite distribution in 10 completely sequenced genomes. Using a permutation-based statistic, we assessed for all 10 genomes whether the microsatellite distribution significantly differed from expectations. Consistent with previous reports, we observed a highly significant excess of long microsatellites. Focusing on short microsatellites containing only a few repeat units, we demonstrate that this repeat class is significantly underrepresented in most genomes. This pattern was observed across different repeat types. Computer simulations indicated that neither base substitutions nor a combination of length-dependent slippage and base substitutions could explain the observed pattern of microsatellite distribution. When we introduced one additional mutation process, a length-independent slippage (indel slippage) operating at repeats with few repetitions, our computer simulations captured the observed pattern of microsatellite distribution.

[Supplemental material is available online at www.genome.org.]

Microsatellites are short tandemly repeated sequence motifs consisting of 1–6 bp (Tautz 1993; Ellegren 2000b; Schlötterer 2000). Over the past decade, microsatellites have attracted considerable attention due to their involvement in some neurodegenerative diseases and their high polymorphism (Goldstein and Schlötterer 1999). Microsatellites gain and lose repeat units at high rates. The underlying mutation process has been termed "DNA replication slippage." It is assumed that during DNA synthesis, the nascent strand dissociates and realigns out of register. When DNA synthesis continues, the repeat number at the microsatellite is altered at the nascent strand (Ellegren 2000b; Schlötterer 2000). Interestingly, the DNA replication slippage rate seems to be dependent on the length of the microsatellite. Alleles with a high repeat number are less stable than those with a small repeat number. This trend has been seen in pedigree studies (Brinkmann et al. 1998; Brohede et al. 2002); in vitro experiments (Shinde et al. 2003), and in surveys of population variability (Goldstein and Clark 1995; Bachtrog et al. 2000). In addition to this dependence on repeat count, microsatellite stability also depends on the repeat motif (Schlötterer and Tautz 1992; Chakraborty et al. 1997; Bachtrog et al. 2000). The major drawback of experiments aiming to characterize microsatellite mutational dynamics from direct observations of microsatellite mutations (e.g., pedigree analyses) is their limitation to microsatellites with a high mutation rate (and thus with a high repeat count). Population surveys could in principle also use shorter microsatellites, but here the observed microsatellite variability is not only a reflection of the microsatellite mutation rate, but also population history (Stumpf and Goldstein 2001) and selection (Harr et al. 2002a).

With the steadily increasing number of genomic sequences, an alternative approach to study microsatellite evolution has become feasible (Jurka and Pethiyagoda 1995; Bell and Jurka 1997; Cox and Mirkin 1997; Field and Wills 1998). Assuming that the distribution of microsatellites in the genome reflects an equilibrium state between the operating evolutionary forces, information about the underlying mutation processes could be extracted from the analysis of genomic sequences. Comparing the observed distribution of microsatellites to the expectations based on the random distribution of nucleotides conditional on their frequencies, it was noted that the distribution of microsatellites deviates from this simple Bernoulli model. The most obvious deviation from expectations was an overrepresentation of long microsatellites. Bell and Jurka (1997) used an unbiased random walk model to analyze this overrepresentation of long microsatellites. Their model incorporated two opposing forces operating on microsatellite sequences: (1) length-dependent DNA replication slippage, which results in a growth of repeats, and (2) base substitutions interrupting the repeat structure. Using the unbiased random walk model, those authors obtained a better fit for long microsatellites than the Bernoulli model.

Applying a Markov chain model of microsatellite evolution to a number of eukaryotic organisms, Kruglyak et al. (1998) were able to quantify the rates of replication slippage. Similar to Bell and Jurka (1997), they assumed that the slippage rate increases with the length of a repeat unit, and they combined this with a constant rate of base substitutions. By fitting observed microsatellite length distributions to the stationary length distribution under their model, the authors inferred a species-specific rate of replication slippage (Kruglyak et al. 1998). Although the combination of base substitutions and DNA replication slippage seemingly accounted for the lack of very long microsatellites, a refined model indicated that additional mutational forces must be incorporated to describe genomic microsatellite distributions (Calabrese et al. 2001). Evidence from yeast, *D. melanogaster*, and humans suggests that long microsatellite alleles have a downward mutation spectrum, which could result in a size constraint of microsatellites (Schlötterer 1998; Ellegren 2000a; Harr and Schlötterer 2000; Xu et al. 2000; Harr et al. 2002b).

The genomic distribution of long microsatellites has been studied in relation to functionally different components of the genome. Tri-nucleotide repeats are overrepresented in coding sequences, but less frequent than mono- and di-nucleotide repeats in noncoding regions (Tóth et al. 2000; Morgante et al. 2002). Furthermore, the microsatellite distribution seems to differ between intergenic and intronic sequences (Tóth et al. 2000; Mor-

gante et al. 2002). In plants also, 5' and 3' UTRs differ in microsatellite density (Morgante et al. 2002). In *Arabidopsis thaliana*, low microsatellite densities are observed in the centromeric region containing many transposable elements (Schlötterer 2000). In *D. melanogaster*, the microsatellite distribution differs between X-chromosomes and autosomes (Bachtrog et al. 2000). These data suggest a significant heterogeneity in microsatellite distribution with respect to functionally different components of the genome.

Over the past years much has been learned about the complex mutational behavior of long microsatellites, but still very little is known about the mutational dynamics of short microsatellites. Early experimental results suggested that short microsatellites do not mutate by DNA replication slippage. In addition, the analysis of the complete genomic sequence of *Saccaromyces cerevisiae* found a close fit between the observed density of microsatellites below 8–10 bp and the density expected by the base composition (Rose and Falush 1998). Based on this close fit, the authors concluded that short microsatellites are highly stable (Rose and Falush 1998). Interestingly, another study used the same genomic sequence and concluded that even short microsatellites mutate by DNA replication slippage (Pupko and Graur 1999). Further evidence of instability of short microsatellites comes from a detailed analysis of the mutation spectrum in mice with a *lacI* transgene (Halangoda et al. 2001). Those authors observed an exponential increase of slippage mutations in mononucleotide runs consisting of 1–5 repeats, suggesting that short repeats mutate by DNA replication slippage. Interspecies comparisons also indicated that microsatellites with a small number of repeats also gain and lose repeat units (Schlötterer and Zangerl 1999; Webster et al. 2002). Further support for the gains and losses of repeat units has been provided in a seminal paper by Zhu et al. (2000). They analyzed the mutation spectrum in human genes resulting in a diseased phenotype. They found that new mutations were the most common events, which generated new microsatellites consisting of two repeat units. More important, however, they also observed that over 70% of the 2–4-bp insertions are duplications of adjacent sequences. Similar results were obtained from an analysis of pseudogenes. Nishizawa and Nishizawa (2002) found that between 50% and 80% of the indels are involved in a slippage-like process. Even short repeats, such as CCCC, have a 10–15-fold increased susceptibility to insertions and deletions compared to nonrepetitive sequences.

Thus, although some evidence suggests that short microsatellites are gaining/losing repeats, the current evidence is either derived from a biased sample set (Zhu et al. 2000, Nishizawa and Nishizawa 2002) or lacks statistical support (Pupko and Graur 1999). In the present study, we analyzed the genomic distribution of microsatellites and used a permutation procedure to evaluate the statistical significance of observed deviations from the expectation. Based on our analyses, we suggest that short microsatellites evolve by a slippage-like mutation process. In contrast to DNA replication slippage, this mutation process seems not to be length-dependent.

RESULTS

We determined the density of microsatellite repeats for nine different eukaryotes for which the complete genomic sequence is available. Four different microsatellite classes were analyzed, mono-, di-, tri-, and tetranucleotide repeats. Consistent with previous reports (Katti et al. 2001), we found dramatic differences among repeat classes within and between species (Table 1). Such differences in microsatellite composition may be caused by (1) species-specific differences in the DNA synthesis and repair machinery (Harr et al. 2002b), (2) selection (Nauta and Weissing 1996), or (3) base composition (see below).

Influence of Base Composition on Microsatellite Density

We evaluated the potential influence of variation in base composition on the observed microsatellite density, by varying the GC content in a sequence between 0.2 and 0.8. Figure 1 indicates that base composition has a dramatic influence on microsatellite density. Both mono- and dinucleotide sequences have the lowest density at a balanced GC content. The expected number of microsatellites strongly increases with a more biased GC content. This effect becomes more pronounced for longer microsatellites and opposite for very short ones (data not shown). Given that the base composition differs substantially among species and also

Table 1A. Mononucleotide Repeat Microsatellite Densities (Microsatellites per 10 kb) in Different Genomes for Each Repeat Length (bp)

Length	*H. sap.*	*M. mus.*	*F. rub.*	*D. mel.*	*C. ele.*	*A. tha.*	Rice 1[a]	Rice 2[b]	*C. alb.*	*S. cer.*
2	1311	1312	1287	1357	1264	1362	1378	1397	1446	1422
3	436	425	404	426	409	414	374	381	435	401
4	144	135	123	140	187	145	125	128	141	130
5	50.0	39.9	39.6	46.7	85.7	47.0	39.5	41.0	50.2	43.1
6	14.5	9.9	11.0	13.0	34.7	14.0	12.7	12.3	16.0	14.0
7	5.58	4.00	5.21	3.88	13.19	6.38	6.19	5.94	6.65	5.81
8	2.02	1.56	2.55	1.80	6.55	2.93	3.58	3.51	3.52	2.21
9	1.16	1.08	1.23	1.23	4.71	1.85	1.84	1.86	2.65	1.04
10	0.754	0.777	0.817	0.901	2.237	1.202	1.018	0.935	2.342	0.631
11	0.468	0.456	0.507	0.541	0.602	0.589	0.336	0.322	1.209	0.401
12	0.364	0.325	0.351	0.360	0.249	0.327	0.186	0.167	0.789	0.292
13	0.318	0.252	0.245	0.233	0.117	0.207	0.112	0.090	0.417	0.201
14	0.283	0.193	0.174	0.164	0.068	0.148	0.067	0.053	0.280	0.108
15	0.251	0.148	0.122	0.122	0.045	0.115	0.043	0.031	0.159	0.073
16	0.215	0.110	0.085	0.088	0.030	0.092	0.026	0.019	0.106	0.054
17	0.175	0.077	0.062	0.053	0.024	0.067	0.015	0.012	0.055	0.044
18	0.140	0.053	0.046	0.035	0.018	0.050	0.012	0.007	0.039	0.023
19	0.119	0.041	0.032	0.022	0.010	0.038	0.008	0.005	0.016	0.029
20	0.101	0.034	0.021	0.018	0.008	0.030	0.005	0.003	0.015	0.018

[a]*Oryza sativa L. ssp. Indica.*
[b]*Oryza sativa L. ssp. japonica.*

Table 1B. Dinucleotide Repeat Microsatellite Densities (Microsatellites per 10 kb) in Different Genomes for Each Repeat Length (bp)

Length	*H. sap.*	*M. mus.*	*F. rub.*	*D. mel.*	*C. ele.*	*A. tha.*	Rice 1[a]	Rice 2[b]	*C. alb.*	*S. cer.*
4	255	262	263	230	228	246	246	244	214	232
5	79.6	86.9	76.3	61.4	57.3	79.2	74.4	73.0	61.4	63.4
6	19.7	23.9	21.9	17.8	14.9	21.8	21.2	20.8	14.9	15.5
7	7.73	9.55	8.05	6.39	4.53	8.99	8.19	8.02	5.21	4.66
8	2.57	3.25	2.94	2.51	1.31	3.20	3.14	3.06	1.69	1.31
9	1.24	1.68	1.56	1.21	0.42	1.47	1.44	1.46	0.93	0.57
10	0.510	0.652	0.716	0.555	0.220	0.584	0.575	0.578	0.393	0.187
11	0.366	0.584	0.571	0.345	0.146	0.320	0.310	0.359	0.305	0.120
12	0.196	0.299	0.331	0.225	0.099	0.148	0.159	0.171	0.156	0.053
13	0.174	0.317	0.316	0.190	0.071	0.117	0.136	0.144	0.163	0.056
14	0.102	0.179	0.202	0.143	0.049	0.073	0.078	0.081	0.105	0.029
15	0.096	0.193	0.215	0.133	0.046	0.068	0.078	0.080	0.087	0.030
16	0.055	0.111	0.143	0.104	0.033	0.056	0.050	0.051	0.059	0.020
17	0.054	0.120	0.162	0.086	0.025	0.048	0.051	0.047	0.059	0.025
18	0.035	0.073	0.118	0.076	0.022	0.037	0.032	0.033	0.034	0.019
19	0.035	0.082	0.127	0.061	0.015	0.036	0.031	0.029	0.036	0.011
20	0.025	0.054	0.091	0.049	0.014	0.024	0.024	0.019	0.027	0.015

[a]*Oryza sativa L. ssp. Indica.*
[b]*Oryza sativa L. ssp. japonica.*

Table 1C. Trinucleotide Repeat Microsatellite Densities (Microsatellites per 10 kb) in Different Genomes for Each Repeat Length (bp)

Length	*H. sap.*	*M. mus.*	*F. rub.*	*D. mel.*	*C. ele.*	*A. tha.*	Rice 1[a]	Rice 2[b]	*C. alb.*	*S. cer.*
6	91.5	89.7	102.8	89.9	88.1	98.8	104.2	104.2	110.4	103.7
7	27.8	27.3	31.8	28.1	25.5	33.6	35.8	36.0	39.7	32.5
8	8.86	8.98	11.35	11.30	9.08	13.06	14.37	14.57	17.90	11.74
9	2.82	2.52	3.71	3.98	3.02	4.60	5.08	5.20	6.62	3.84
10	1.03	0.95	1.43	1.71	1.07	1.97	2.31	2.40	3.25	1.43
11	0.485	0.579	0.781	1.103	0.613	1.113	1.274	1.332	2.314	0.790
12	0.185	0.215	0.324	0.474	0.270	0.505	0.583	0.608	1.071	0.265
13	0.114	0.130	0.208	0.261	0.155	0.280	0.374	0.387	0.721	0.132
14	0.116	0.174	0.182	0.235	0.124	0.226	0.289	0.280	0.665	0.115
15	0.048	0.069	0.084	0.126	0.078	0.122	0.175	0.193	0.395	0.050
16	0.031	0.053	0.069	0.084	0.047	0.086	0.132	0.129	0.331	0.049
17	0.050	0.080	0.077	0.094	0.036	0.068	0.112	0.110	0.361	0.055
18	0.019	0.031	0.037	0.053	0.024	0.046	0.072	0.071	0.225	0.023
19	0.014	0.024	0.034	0.036	0.016	0.031	0.052	0.051	0.180	0.009
20	0.024	0.041	0.039	0.045	0.010	0.032	0.046	0.042	0.179	0.029

[a]*Oryza sativa L. ssp. Indica.*
[b]*Oryza sativa L. ssp. japonica.*

Table 1D. Tetranucleotide Repeat Microsatellite Densities (Microsatellites per 10 kb) in Different Genomes for Each Repeat Length (bp)

Length	*H. sap.*	*M. mus.*	*F. rub.*	*D. mel.*	*C. ele.*	*A. tha.*	Rice 1[a]	Rice 2[b]	*C. alb.*	*S. cer.*
8	27.0	30.0	26.4	28.1	22.9	27.1	29.3	29.1	27.2	22.4
9	9.00	9.95	8.23	9.36	6.91	9.03	9.33	9.31	9.10	6.54
10	3.23	4.17	2.78	3.59	2.33	3.23	3.52	3.49	3.50	2.19
11	1.57	2.13	1.45	1.66	1.03	1.33	1.53	1.52	1.62	0.74
12	0.515	0.795	0.473	0.608	0.356	0.467	0.564	0.563	0.652	0.197
13	0.277	0.394	0.234	0.317	0.172	0.220	0.261	0.259	0.338	0.073
14	0.187	0.281	0.146	0.202	0.084	0.111	0.153	0.150	0.235	0.038
15	0.297	0.442	0.191	0.138	0.052	0.070	0.114	0.104	0.181	0.019
16	0.098	0.154	0.075	0.069	0.029	0.031	0.062	0.056	0.122	0.013
17	0.071	0.096	0.050	0.048	0.022	0.019	0.037	0.036	0.079	0.011
18	0.067	0.087	0.037	0.036	0.016	0.013	0.029	0.028	0.069	0.010
19	0.120	0.165	0.069	0.025	0.007	0.007	0.027	0.024	0.059	0.008
20	0.0372	0.0587	0.0281	0.0201	0.0054	0.0038	0.0146	0.013	0.0442	0.0008

[a]*Oryza sativa L. ssp. Indica.*
[b]*Oryza sativa L. ssp. japonica.*

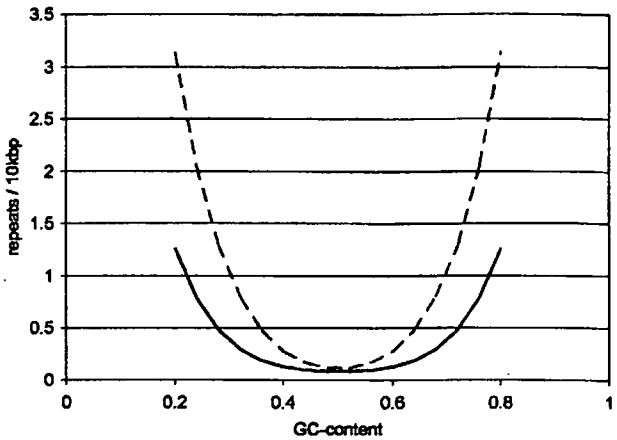

Figure 1 Correlation between expected microsatellite density and GC content for mononucleotide microsatellites (dashed line) and dinucleotide microsatellites (solid line). For both repeat classes, all microsatellites ≥9 bp were considered.

within genomes, any comparison of microsatellite densities also needs to account for local base composition.

Statistical Evaluation of Microsatellite Densities

Assuming a random distribution of nucleotides in a given sequence (mononucleotide space), the expected number of microsatellites could be calculated analytically. The number of $(CA)_4$ repeats in a 10-kb frequent with a balanced base composition would be $(1 - p_A) \times p_C^4 \times p_A^4 \times (1 - p_C) \times 10000$, where p_A and p_C are the frequency of A and C, respectively. When higher-order moments, such as a pronounced dinucleotide pattern (Gentles and Karlin 2001), are present in a sequence, the expectation for the number of microsatellites could be calculated as $(1 - p_A) \times p_{CA}^4 \times (1 - p_C) \times 10{,}000$. Although both approaches have been pursued, they are somewhat unsatisfactory as they provide only an expectation for the microsatellite density, and no variance. Hence, it is not possible to determine whether an observed microsatellite density differs significantly from the expectations.

To overcome these limitations, we permuted the genomic sequences in 20-kb intervals. This strategy was chosen to account for heterogeneity in base composition within species (Lander et al. 2001). A certain repeat type was considered to deviate from expectations when the observed number of microsatellites fell outside of the 95% confidence interval determined by our permutation procedure. Irrespective of whether our permutations were based on a mononucleotide or a dinucleotide space, we observed substantial differences among repeat types and species. The different permutation schemes (permuting either single nucleotides or two adjacent bases) provided qualitatively similar results; therefore, we limit our discussion to the results obtained by those permutations assuming a mononucleotide space (Table 2, Supplemental Tables A1 and A2, available online at www.genome.org). Results based on the dinucleotide space are given in the Supplemental Tables A3 and A4.

Overrepresentation of Microsatellites With a High Repeat Number

Previous results suggested that above a certain threshold repeat number, microsatellites are found more frequently in the genome than expected by chance (Rose and Falush 1998). Using our statistical approach, we were able to evaluate the statistical significance of such observations. Although some repeat motifs were overrepresented at all repeat numbers, most repeat types were only overrepresented beyond a certain threshold (Table 2; see Suppl. Tables A1 and A2 for more details). For some repeat types, such as $(GC)_n$, no long microsatellites were detected in species with a small genome. In species with a larger genome, such as human and mouse, however, these repeats were present and also overrepresented.

Absolute Length in Base Pairs Versus Repeat Number

As we were comparing microsatellites with different repeat lengths (i.e., mono-, di-, tri-, and tetranucleotide repeats), we were interested whether the microsatellite distribution is better analyzed by the absolute length of the repeat structure (in bp) or by the number of repeats. Figure 2 shows the ratio of the observed and expected microsatellite density for *H. sapiens* (for the other genomes see Suppl. Fig. A1). Figure 2A is scaled by absolute length in base pairs, and Figure 2B is scaled by repeat number. For

Table 2. Threshold of Overrepresentation[a] for Each Genome and Repeat Type

	Mono		di				tri									
	a/t	g/c	ac/gt	ag/ct	at/ta	cg/gc	aac/gtt	aag/ctt	aat/att	acc/ggt	acg/cgt	act/agt	agc/gct	agg/cct	atc/gat	ccg/cgg
H. sap.	3	9	≤4	≤4	8	13	11	≤6	10	≤6	15	16	≤6	≤6	9	11
M. mus.	7	10	≤4	≤4	12	12	13	13	14	11	14	14	12	11	13	12
F. rub.	3	5	≤4	≤4	10	13	≤6	≤6	8	≤6	14	14	≤6	≤6	≤6	7
D. mel.	3	6	5	9	6	13	≤6	8	8	≤6	9	14	≤6	≤6	10	≤6
C. ele.	3	10	10	10	15	10	14	≤6	16	≤6	12	14	≤6	≤6	13	≤6
A. tha.	3	11	7	≤4	8	12	≤6	≤6	16	≤6	12	14	≤6	≤6	≤6	7
Rice 1[b]	3	10	7	5	6	5	≤6	≤6	7	≤6	7	8	≤6	≤6	≤6	≤6
Rice 2[c]	3	4	7	5	6	5	≤6	≤6	7	≤6	7	8	≤6	≤6	≤6	≤6
C. alb.	≤2[d]	≤2[d]	7	9	11	never	≤6	≤6	10	≤6	11	9	≤6	≤6	≤6	11
S. cer.	3	11	12	13	13	11	13	≤6	16	≤6	12	14	≤6	13	≤6	11

[a]Long microsatellites are overrepresented. We determined at which size a given repeat motif was more frequently detected than expected by chance (see Methods). Here we provide the number of bases at which we start to observe a significant overrepresentation.
[b]*Oryza sativa L. ssp. Indica.*
[c]*Oryza sativa L. ssp. japonica.*
[d]Because our counting procedure was limited to a minimum of two repeats, we cannot make any inference about shorter repeats consisting of less than two repeats. For those cases in which we observed an overrepresentation at two repeat units, we used the "≤" symbol to indicate this uncertainty.

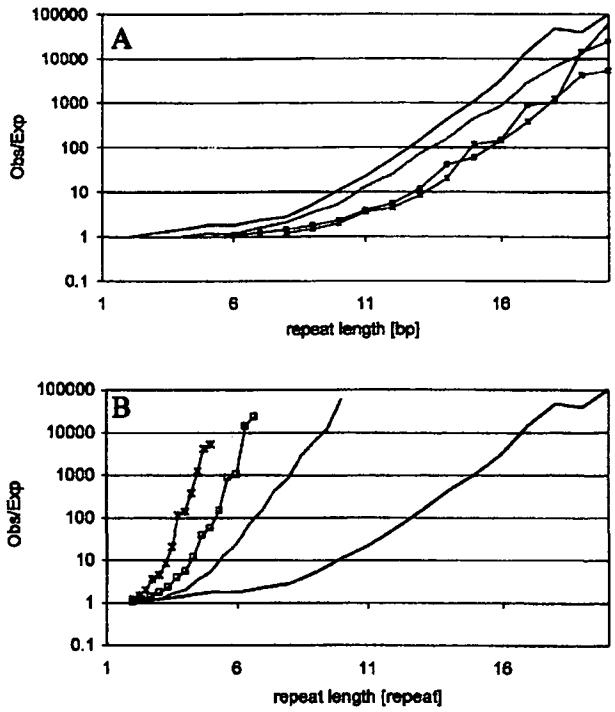

Figure 2 Ratio of the observed density and expected microsatellite density for *H. sapiens*. (*A*) Length scaled by absolute length (bp). (*B*) Length scaled by repeat number. Black lines, mononucleotide microsatellites; gray lines, dinucleotide microsatellites; (□) trinucleotide microsatellites, and (×) tetranucleotide microsatellites.

all species, the scaling by base pairs resulted in almost perfect parallel lines, and scaling by repeat number resulted in more pronounced differences. Thus, different microsatellites seem to show similar patterns of over- or underrepresentation when their absolute lengths in base pairs are compared. On the other hand, microsatellites with the same repeat number but different repeat lengths show drastically different patterns of over- or underrepresentation. For this reason, all of our representation plots are scaled by the size of the microsatellite stretch in base pairs.

Deviations of Observed and Expected Microsatellite Density Are Length-Dependent

Previous studies focused on microsatellites with a high repeat number, but we also analyzed microsatellites with only a small number of repeats. Although the patterns of over- and underrepresentation differed remarkably between repeat types and species, a general picture emerged (Fig. 3, Suppl. Fig. A2). In principle, three different zones could be distinguished in the representation plot. The first zone contains the very short microsatellites with the size of a few bp only. Microsatellites in this size class are often significantly underrepresented (see also Suppl. Tables A1 and A2). In the second zone, containing microsatellites of an intermediate size range (4–15 bp), the observed and expected microsatellite densities are either very similar or a slight overrepresentation is observed. Finally, the third zone is characterized by a marked overrepresentation of microsatellites. It is important to note that at the transition between zones two and three, the slope of the representation plot changes. For some repeat types, the transition to zone three is associated with an underrepresentation (Fig. 3, Suppl. Fig. A2). The overrepresentation of microsatellites in the longer size classes is consistent with previous studies and has been attributed to DNA replication slippage (Bell and Jurka 1997).

We used computer simulations to understand which genome dynamics could result in a microsatellite distribution that matches our observations in the three zones. Based on the large variation in microsatellite densities among repeat types and species, it is obvious that such computer simulations could only serve as an exploratory tool, and not as a systematic approach to estimate exact parameters. We assumed a random sequence with a balanced nucleotide composition, which experiences base substitutions (with equal transition probabilities) and slippage mutations. For computational simplicity we simulated a genome size of 2×10^7 bp and focused on mononucleotide repeats only, as this repeat type occurs at the highest density. The simplest model assumed only base substitutions (Table 3, Model 1). Hence, mutations could create new microsatellites and also destroy them.

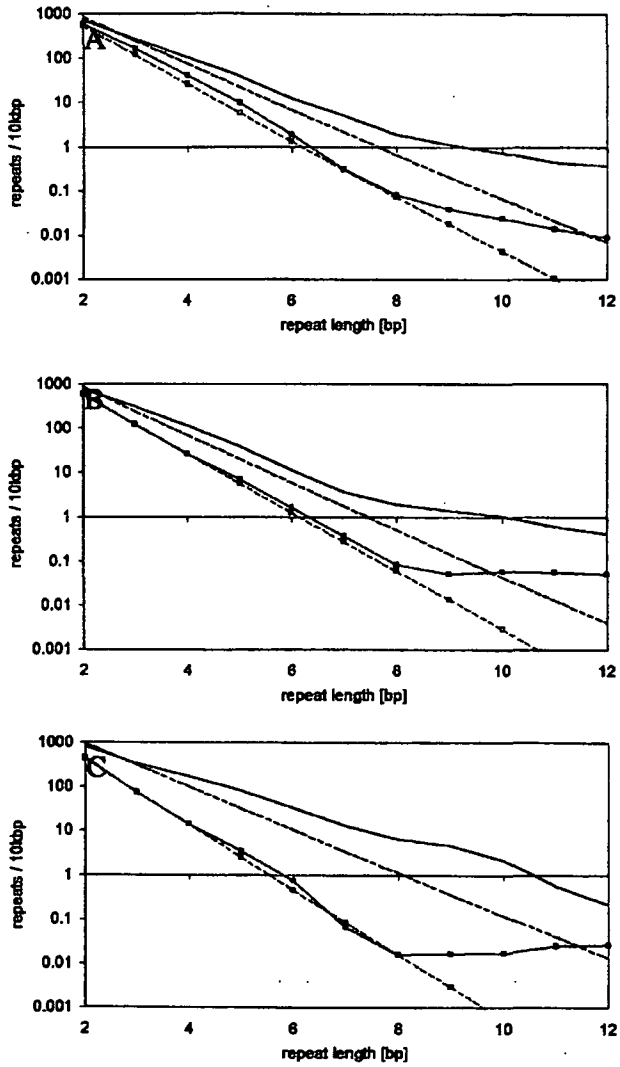

Figure 3 Representation plots for mononucleotide microsatellites. Solid lines, observed densities; dashed lines, the expected densities. Black lines, A/T-microsatellites; gray lines with □, G/C microsatellites. (*A*) *Homo sapiens*, (*B*) *Drosophila melanogaster*, (*C*) *Caenorhabditis elegans*.

Table 3. Simulation Parameters

	Minimum repeat number of μ_{slip}	μ_{indel}
Model 1	∞	0
Model 2	1	0
Model 3	6	0
Model 4	8	0
Model 5	8	$2.5\ 10^{-4}$
Model 6	8[a]	0

[a]Microsatellites with length 8 were not allowed to lose repeats by slipping. Thus, their slippage mutation rate was $0.5*\mu_{slip}$.

Figure 4A clearly indicates that this model corresponds to the expected microsatellite distribution and therefore fails to explain the observed distribution. The second evolutionary model incorporated base substitutions together with replication slippage using a length-dependent slippage rate (i.e., a linear increase with repeat number, Fig. 4B; Table 3, Model 2). This model did not match our observations in two aspects. First, the short microsatellites in the first zone were more underrepresented than in our genome survey. This result is consistent with previous publications, which also noted a large discrepancy in the observed distribution of short microsatellites (Kruglyak et al. 1998). Second, the slope of the representation graph does not change over the entire size range, whereas we observed at least one change between zone two and three. Therefore, this model does also not fully explain the observed genomic distribution of microsatellites. When this model was modified and slippage occurred only for microsatellites above a certain repeat number (Table 3, Models 3, 4, 6), a depletion of microsatellites shorter than the slippage boundary was detected (Fig. 4). One intuitive interpretation of this pattern is that microsatellites above the threshold have the tendency to grow, which results in the overrepresentation of microsatellites with a higher repeat number. At the boundary between slippage and no-slippage, this leads to an underrepresentation of this size class (as the threshold prevents the supply of shorter microsatellites). Nevertheless, this sequence evolution model still does not explain the underrepresentation of the size

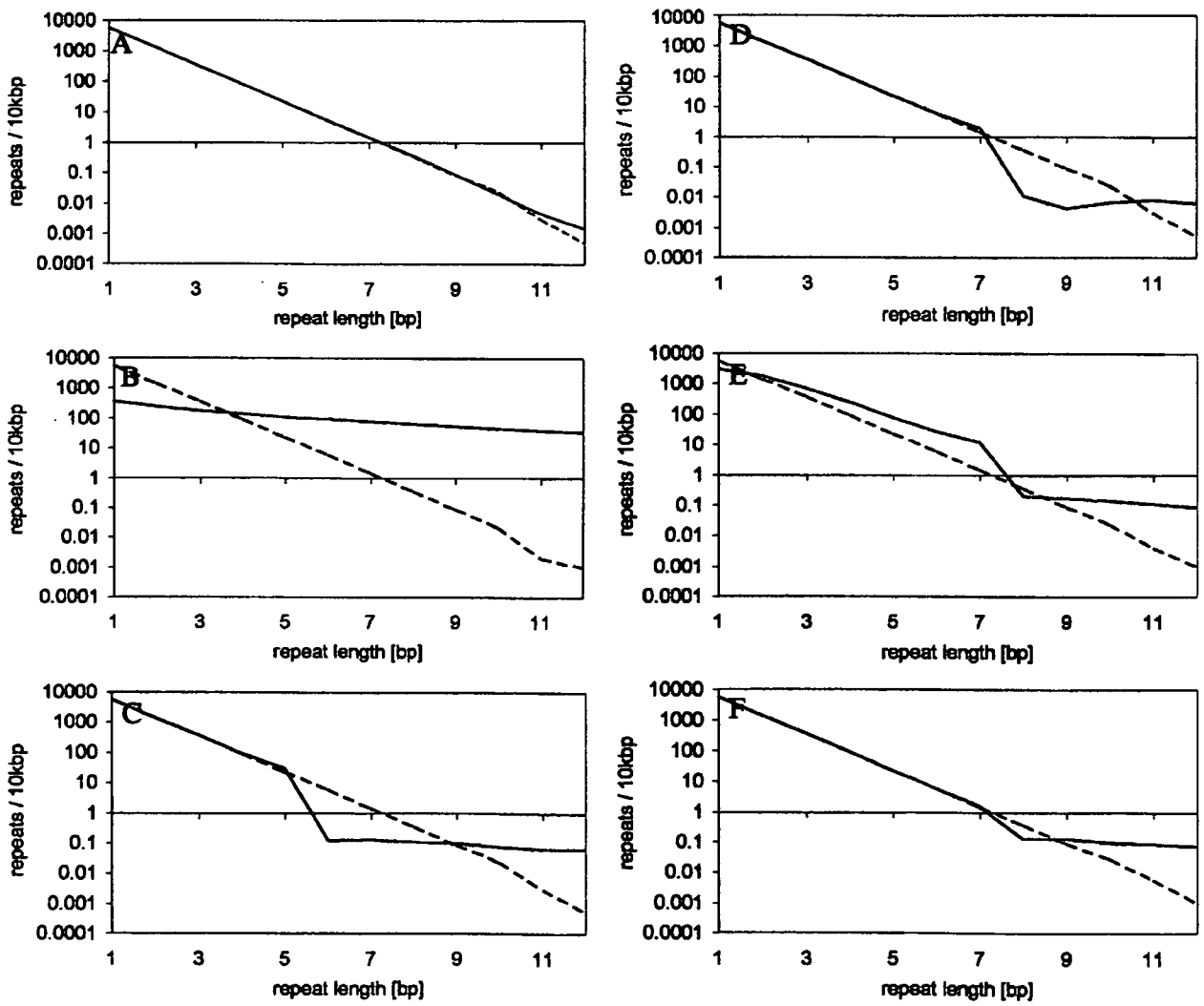

Figure 4 Representation plot for simulation data. Solid lines, simulated densities; dashed lines, the corresponding "expected" densities. (*A*) model 1, (*B*) model 2, (*C*) model 3, (*D*) model 4, (*E*) model 5, (*F*) model 6 (see Table 3 for details).

class of very short microsatellites. In an attempt to understand this, we introduced one additional mutation process. Rather than assuming a slippage process, the mutation rate of which is length-dependent, we introduced a process which we termed "indel slippage." Contrary to the slippage process, we assumed indel-slippage to be not length-dependent, but to occur at a constant rate (Table 3, Model 5). This model is justified by the observation that insertions often copy the flanking sequence, which creates a short microsatellite (Zhu et al. 2000; Nishizawa and Nishizawa 2002). The representation plots for these computer simulations show the same trends observed for the genomic data (Fig. 4E). The first zone with microsatellite underrepresentation could be recognized. The second zone has an overrepresentation of microsatellites, but it is still close to the expectations, and zone three is characterized by the well described overrepresentation. Most important, the simulated data capture not only the change in slope in the representation graph, but also show an underrepresentation at the transition.

Our computer simulations indicate clearly that base substitutions and DNA replication slippage alone are unlikely to explain the genomic distribution of microsatellites. More likely, other turnover mechanisms also need to be considered. Based on previous studies, we assumed an indel-slippage process and could obtain a qualitatively similar pattern. It may be needless to say that the dramatic divergence in pattern of microsatellite distribution also implies that different rates of the three processes and possibly other factors also shape the genomic distribution of microsatellites.

Microsatellite Distribution Differs Among Species

The base composition differs between species, and the number of expected microsatellites is dependent on the base composition. Thus, it is extremely difficult to compare microsatellite distributions across species. One obvious factor contributing to this difficulty is the base composition, which results in different expectations for the microsatellite density.

In an attempt to make the comparison of the mononucleotide microsatellite distribution across species more informative, we calculated for each species the difference between observed and expected microsatellite density, and we standardized this by the mean difference of all species analyzed. Figure 5 shows the relative distribution of mononucleotide repeat microsatellites between two and 20 bp. Most species show only moderate differences. *C. elegans* has significantly more microsatellites in the size class 3–10 bp than expected. *C. albicans* has a similar but less pronounced overrepresentation of microsatellites sized between 10 and 14 bp. The mononucleotide repeats longer than 14 bp were most abundant in humans. Interestingly, the microsatellite distribution in the two rice genomes was very similar and followed the same trends, but was nevertheless not identical. We also performed the same analysis for other repeat types and observed similar species-specific differences in microsatellite distribution (Suppl. Fig. A3).

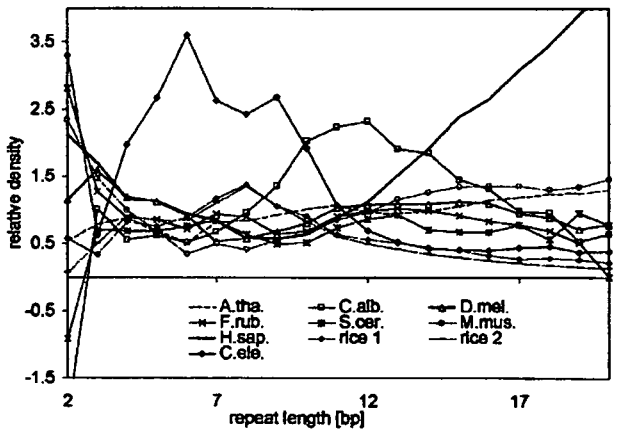

Figure 5 Relative density of mononucleotide microsatellites for all genomes plotted against repeat length. Relative densities were calculated by the comparison of a given species against the mean of all genomes (see text for details).

DISCUSSION

In our survey of microsatellite distribution in nine eukaryotic genomes, we did not discriminate between functionally different components of the genome. The reason for this is that several studies already demonstrated that the distribution of microsatellites differs substantially among exons, introns, 5', and 3' regions of a gene (e.g., Morgante et al. 2002). Our goal, however, was to obtain an overall pattern, rather than to distinguish between different parts of the genome.

Using our parameter-free permutation test, we showed that the pattern of microsatellite over- and underrepresentation differs among repeat types, microsatellite length classes, and species. To determine whether the observed microsatellite densities deviated significantly from expectations, we were very careful to determine the expectation correctly. We accounted for heterogeneity in base composition in a genomic sequence, by limiting our permutation test to small windows sized 20 kb. The simplest model to determine the expected number of microsatellites assumes independence of all nucleotides (Bernoulli model). Thus the expectation could be directly obtained from the frequencies of each nucleotide in the 20-kb window. In addition to this, we also considered higher-order interactions. Rather than permuting single bases, we permuted two neighboring bases. However, independently of the permutation procedure, we observed significant deviations from the expected microsatellite distribution, not only for long microsatellites, but also for the length class encompassing only two repeats. This observation suggested that other evolutionary forces must be operating, which determine the microsatellite distribution.

Consistent with previous studies, we found that long microsatellites were overrepresented. Interestingly, the length at which we observed a significant overrepresentation differed among repeat types and species. Assuming that this overrepresentation is caused by length-dependent DNA replication slippage, our observation may indicate that the rate of slippage differs among species and repeat types. Alternatively, the mutation pattern may differ among species (Harr and Schlötterer 2000; Harr et al. 2002b).

Although the slippage model assumes that long microsatellites are generated by a random walk of upward and downward mutations, which occasionally result in a high repeat number, alternative scenarios are possible, under which long repeats are generated instantaneously. An association of A-rich microsatellites with retrotransposable elements was observed (Nadir et al. 1996). Those authors suggested that A-rich microsatellites were generated by a 3' extension of retrotranscripts, similar to mRNA polyadenylation. Similarly, the de novo generation of $(GT)_n$ microsatellites during nonhomologous repair of double-strand breaks has been described (Liang et al. 1998). To what extent such processes shape the genomic microsatellite distribution remains open to further investigation.

Highly surprising was our observation that the density of short microsatellites also deviated from expectations. Rather than a consistent overrepresentation, which may have been consistent with DNA replication slippage operating at short repeats, we observed a rather complex pattern. Some short repeats were significantly underrepresented, and others were overrepresented. Even more interestingly, several repeat type microsatellites of intermediate length either showed no or moderate overrepresentation. Building on seminal work of Zhu et al. (2000), we considered one additional mutation process to explain this pattern. Zhu et al. (2000) showed that indel-like processes tend to duplicate short sequences, and thus we termed this process indel slippage. Using computer simulations we showed that the combination of indel slippage with base substitutions and length-dependent DNA replication slippage could result in a microsatellite distribution that closely resembled our observations.

The distribution of microsatellites in protein coding regions was studied recently (Borstnik and Pumpernik 2002). For alanine codons, those authors observed that the abundance of repeats follows a curve with two different slopes. This is very similar to the pattern we observed for many repeat types. In contrast to our model, Borstnik and Pumpernik assumed only length-dependent slippage and base substitutions. Whereas we assumed complete neutrality, those authors also accounted for the higher order of the protein coding sequence. Thus, similar results regarding the distribution of microsatellite repeat types could be obtained using different assumptions. This result underpins the overall difficulty of extracting precise information about the mutation process of genomic sequences from their genomic distribution. Nevertheless, the availability of sequences from more closely related species will greatly facilitate future attempts to understand the evolution of tandemly repeated sequences. Based on such sequences, it will be possible to compare orthologous microsatellites in fully sequenced genomes (Webster et al. 2002).

METHODS

Sequences

Genome sequences were obtained in the FASTA format, and non-sequence information as well as ambiguous sequence information (e.g., N) was removed.

Arabidopsis thaliana (*A. tha.*) chromosomes were downloaded as pseudomolecules, which contain all available nonredundant sequences: ftp://ftp.tigr.org/pub/data/a_thaliana/ath1/PUBLICATION_RELEASE/PSEUDOMOLECULES/ (21.12.2000)

Saccharomyces cerevisiae (*S. cer.*), *Drosophila melanogaster* (*D. mel.*), and *Caenorhabditis elegans* (*C. ele.*) pseudomolecules were obtained from: ftp://ncbi.nlm.nih.gov/genomes/ and ftp://ncbi.nlm.nih.gov/genbank (17.03.2000)

Homo sapiens (*H. sap.*) pseudomolecules were downloaded from: ftp://ncbi.nlm.nih.gov/genomes/ (09.07.2001)

We obtained draft sequences from two different rice varieties, *Oryza sativa L. ssp. indica* (*O.sat. indica* or Rice 1) from http://btn.genomics.org.cn/rice (30.4.2002) and *Oryza sativa L. ssp. japonica* (*O.sat. japonica* or Rice 2) from a CD distributed by Syngenta Biotechnology (formerly the Torrey Mesa Research Institute; 3.6.2002).

The *Fugu rubripes* (*F. rub.*) unannotated draft genome assembly was obtained from: www.fugu-sg.org (25.10.2002).

Mus musculus (*M. mus.*) sequences were obtained from: ftp://ftp.ensembl.org/pub/current_mouse/data/fasta/dna/ (5.12.2002)

Candida albicans (*C. alb.*) sequence was downloaded from: ftp://ncbi.nlm.nih.gov/genomes/ (17.04.2002).

Calculation of Expected Microsatellite Density From Base Composition

The expected distribution of mononucleotide repeats could be directly obtained from the genomic frequency p_A of the nucleotide forming the mononucleotide repeat:

$$p_{A,N} = (1 - p_A) \times p_A^N \times (1 - p_A) \quad (1)$$

where N is the number of repeats.

This formula could be extended to dinucleotide repeats consisting of two bases A and C, which occur at frequency p_A and p_C. The expected frequency of dinucleotide repeats with a length of N base pairs could be calculated as

$$p_{CA,N} = (1 - p_A) \times p_C^N \times p_A^N \times (1 - p_C) \quad (2)$$

for even N, and

$$p_{CA,N} = (1 - p_A) \times p_C^{N+0,5} \times p_A^{N-0,5} \times (1 - p_A) \quad (3)$$

for odd N. (Kruglyak et al. 1998; Rose and Falush 1998).

Microsatellite Counts

We determined the number of microsatellite repeats by counting the number of bases forming a consecutive sequence stretch consisting of a single repeat type. This number was divided by the size of the repeat unit. Note that we also included noninteger repeat units (e.g., $[CA]_{15.5}$). Higher-order motifs that could be decomposed into a lower-order motif were not considered (e.g., $[CACA]_7$ was counted as $[CA]_{14}$). Only those repeats consisting of a minimum of two repeats were counted. Interrupted repeat stretches were decomposed into multiple repeats without interruption. Juxtaposed microsatellite repeats consisting of two different repeat types were counted as two separate repeats. In cases in which the repeat units of such juxtaposed microsatellites share some bases, we resolved the repeat structure in the 5′ to 3′ direction (i.e., the sequence $[A]_{10}[AAT]_5$ would be resolved as two microsatellites $[A]_{12}$ and $[TAA]_{4.3}$).

Microsatellite densities are given in number of microsatellites per 10 kb. The C-code for counting microsatellite densities is available from the authors on request.

Permutation Tests

To evaluate whether the genomic microsatellite distribution deviates from the expectation based on a random genome composition, we used a permutation procedure. In nonoverlapping 20-kb windows, we permuted the original sequence and determined the microsatellite distribution. The permutation was limited to the 20-kb windows to account for heterogeneity in base composition within genomes. This procedure was chosen to account for the known local variation of the base composition over a genome (Lander et al. 2001). For each genome, 250 permutations were made. The genomic microsatellite distribution was considered to deviate significantly from randomness when an observed microsatellite density fell outside the 95% confidence interval obtained from 250 permutations. Note that for microsatellites with a low repeat count, this number of permutations is sufficient (data not shown). As microsatellites with a higher repeat count are much rarer, a higher number of permutations would be required to estimate the 95% confidence interval reliably.

The permutation procedure was modified to account for a higher-order structure in the genome (Gentles and Karlin 2001). Rather than permuting single bases, we permuted two adjacent bases.

Representation Plots

We used a representation plot to visualize the over-and under-representation of microsatellites over their entire size range. The expected microsatellite density and its confidence interval are obtained by permutations based on the observed mono- or di-

nucleotide space. Fewer than expected microsatellites fall below the diagonal, and more than expected microsatellites are located above the diagonal.

Genome Evolution Simulations

To determine the genomic microsatellite distribution when different mutational processes are operating, we analyzed the distribution of mononucleotide repeats in a sequence stretch of 20 Mb. The simulations were started with a random distribution of the four nucleotides all occurring at the same frequency. The entire genome was exposed to multiple rounds of mutation. The number of mutations occurring in each round is drawn from a Poisson distribution. The position of a mutation in the sequence is drawn from a uniform distribution. Transitions and transversions were equally likely.

DNA replication slippage mutations were added for each microsatellite length class separately. First the number of repeats (N) in each class was determined. The number of microsatellite mutations occurring in each length class was determined by a Poisson distribution with a mean of M.

$$M = \mu_{slip} N l. \quad (4)$$

μ_{slip} is the per repeat unit mutation rate, N is the number of repeats, and l is the repeat number of the corresponding size class. We assumed that in one round of mutation, no more than a single slippage mutation occurs at a given microsatellite. Thus, from the total number of microsatellites in a given size class, we randomly chose one microsatellite for each microsatellite mutation. Insertion and deletions of single repeat units were selected with equal probabilities.

Indel slippage mutations were generated in a manner similar to that used for the replication slippage mutations. Only the mean of the Poisson distribution differed: $M = M = \mu_{indel} N$, where N is the total number of microsatellites in the sequence stretch.

The simulations were discontinued after 10,000 rounds or when a stable microsatellite size distribution was reached. The C-code for these simulations is available from the authors on request.

Three different mutation processes were considered: base substitutions occurring at rate μ_{base} ($=10^{-4}$), length-dependent DNA replication slippage occurring at a rate μ_{slip}($=\mu_{slipbase}$ * repeat number, $\mu_{slipbase} = 5\ 10^{-3}$), and indel slippage with the mutation rate μ_{indel}, which is length-independent. Slippage mutations followed the strict stepwise mutation model, which results in the gain or loss of one repeat unit with equal probability. Table 3 provides an overview of the parameters used for the simulations

ACKNOWLEDGMENTS

We thank the members of the CS Lab and Claus Vogel for discussion and three anonymous reviewers for their helpful comments on a previous version of this manuscript. This work was supported by Fonds zur Förderung der wissenschaftlichen Forschung (FWF) grants to C.S.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Bachtrog, D., Agis, M., Imhof, M., and Schlötterer, C. 2000. Microsatellite variability differs between dinucleotide repeat motifs-evidence from *Drosophila melanogaster*. *Mol. Biol. Evol.* **17:** 1277–1285.

Bell, G.I. and Jurka, J. 1997. The length distribution of perfect dimer repetitive DNA is consistent with its evolution by an unbiased single-step mutation process. *J. Mol. Evol.* **44:** 414–421.

Borstnik, B. and Pumpernik, D. 2002. Tandem repeats in protein coding regions of primate genes. *Genome Res.* **12:** 909–915.

Brinkmann, B., Klintschar, M., Neuhuber, F., Huhne, J., and Rolf, B. 1998. Mutation rate in human microsatellites: Influence of the structure and length of the tandem repeat. *Am. J. Hum. Genet.* **62:** 1408–1415.

Brohede, J., Primmer, C.R., Moller, A., and Ellegren, H. 2002. Heterogeneity in the rate and pattern of germline mutation at individual microsatellite loci. *Nucleic Acids Res.* **30:** 1997–2003.

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

Chakraborty, R., Kimmel, M., Stivers, D.N., Davison, L.J., and Deka, R. 1997. Relative mutation rates at di-, tri-, and tetranucleotide microsatellite loci. *Proc. Natl. Acad. Sci.* **94:** 1041–1046.

Cox, R. and Mirkin, S.M. 1997. Characteristic enrichment of DNA repeats in different genomes. *Proc. Natl. Acad. Sci.* **94:** 5237–5242.

Ellegren, H. 2000a. Heterogeneous mutation processes in human microsatellite DNA sequences. *Nat. Genet.* **24:** 400–402.

Ellegren, H. 2000b. Microsatellite mutations in the germline: Implications for evolutionary inference. *Trends Genet.* **16:** 551–558.

Field, D. and Wills, C. 1998. Abundant microsatellite polymorphism in *Saccharomyces cerevisiae*, and the different distributions of microsatellites in eight prokaryotes and *S. cerevisiae*, result from strong mutation pressures and a variety of selective forces. *Proc. Natl. Acad. Sci.* **95:** 1647–1652.

Gentles, A.J. and Karlin, S. 2001. Genome-scale compositional comparisons in eukaryotes. *Genome Res.* **11:** 540–546.

Goldstein, D. and Schlötterer, C. 1999. *Microsatellites: Evolution and applications*. Oxford University Press, Oxford, UK.

Goldstein, D.B. and Clark, A.G. 1995. Microsatellite variation in North American populations of *Drosophila melanogaster*. *Nucleic Acids Res.* **23:** 3882–3886.

Halangoda, A., Still, J.G., Hill, K.A., and Sommer, S.S. 2001. Spontaneous microdeletions and microinsertions in a transgenic mouse mutation detection system: Analysis of age, tissue, and sequence specificity. *Environ. Mol. Mutagen.* **37:** 311–323.

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

Harr, B., Kauer, M., and Schlötterer, C. 2002a. Hitchhiking mapping—A population based fine mapping strategy for adaptive mutations in *D. melanogaster*. *Proc. Natl. Acad. Sci.* **99:** 12949–12954.

Harr, B., Todorova, J., and Schlötterer, C. 2002b. Mismatch repair driven mutational bias in *D. melanogaster*. *Mol. Cell* **10:** 199–205.

Jurka, J. and Pethiyagoda, C. 1995. Simple repetitive DNA sequences from primates: Compilation and analysis. *J. Mol. Evol.* **40:** 120–126.

Katti, M.V., Ranjekar, P.K., and Gupta, V.S. 2001. Differential distribution of simple sequence repeats in eukaryotic genome sequences. *Mol. Biol. Evol.* **18:** 1161–1167.

Kruglyak, S., Durret, R.T., Schug, M., and Aquadro, C.F. 1998. Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations. *Proc. Natl. Acad. Sci.* **95:** 10774–10778.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860–921.

Liang, F., Han, M., Romanienko, P.J., and Jasin, M. 1998. Homology-directed repair is a major double-strand break repair pathway in mammalian cells. *Proc. Natl. Acad. Sci.* **95:** 5172–5177.

Morgante, M., Hanafey, M., and Powell, W. 2002. Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes. *Nat. Genet.* **30:** 194–200.

Nadir, E., Margalit, H., Gallily, T., and Ben-Sasson, S.A. 1996. Microsatellites spreading in the human genome: Evolutionary mechanisms and structural implications. *Proc. Natl. Acad. Sci.* **93:** 6470–6475.

Nauta, M.J. and Weissing, F.J. 1996. Constraints on allele size at microsatellite loci: Implications for genetic differentiation. *Genetics* **143:** 1021–1032.

Nishizawa, N. and Nishizawa, K. 2002. A DNA sequence evolution analysis generalized by simulation and the Markov Chain Monte Carlo method implicates strand slippage in a majority of insertions and deletions. *J. Mol. Evol.* **55:** 706–717.

Pupko, T. and Graur, D. 1999. Evolution of microsatellites in the yeast *Saccharomyces cerevisiae*: Role of length and number of repeated units. *J. Mol. Evol.* **48:** 313–316.

Rose, O. and Falush, D. 1998. A threshold size for microsatellite expansion. *Mol. Biol. Evol.* **15:** 613–615.

Schlötterer, C. 1998. Are microsatellites really simple sequences? *Curr. Biol.* **8:** R132–R134.

Schlötterer, C. 2000. Evolutionary dynamics of microsatellite DNA. *Chromosoma* **109:** 365–371.

Schlötterer, C. and Tautz, D. 1992. Slippage synthesis of simple sequence DNA. *Nucleic Acids Res.* **20:** 211–215.

Schlötterer, C. and Zangerl, B. 1999. The use of imperfect microsatellites

for DNA fingerprinting and population genetics. In *DNA profiling and DNA fingerprinting* (eds. J.T. Epplen and T. Lubjuhn), pp. 153–165. Birkhäuser, Basel, Switzerland.
Shinde, D., Lai, Y., Sun, F., and Arnheim, N. 2003. Taq DNA polymerase slippage mutation rates measured by PCR and quasi-likelihood analysis: (CA/GT)n and (A/T)n microsatellites. *Nucleic Acids Res.* **31:** 974–980.
Stumpf, M.P. and Goldstein, D.B. 2001. Genealogical and evolutionary inference with the human Y chromosome. *Science* **291:** 1738–1742.
Tautz, D. 1993. Notes on the definition and nomenclature of tandemly repetitive DNA sequences. In *DNA fingerprinting: State of science* (eds. S.D.J. Pena, R. Chakraborty, J.T. Epplen, and A.J. Jeffreys), pp. 21–28. Birkhäuser Verlag, Basel, Switzerland.
Tóth, G., Gáspári, Z., and Jurka, J. 2000. Microsatellites in different eukaryotic genomes: Survey and analysis. *Genome Res.* **10:** 967–981.
Webster, M.T., Smith, N.G., and Ellegren, H. 2002. Microsatellite evolution inferred from human–chimpanzee genomic sequence alignments. *Proc. Natl. Acad. Sci.* **99:** 8748–8753.
Xu, X., Peng, M., and Fang, Z. 2000. The direction of microsatellite mutations is dependent upon allele length. *Nat. Genet.* **24:** 396–399.
Zhu, Y., Strassmann, J.E., and Queller, D.C. 2000. Insertions, substitutions, and the origin of microsatellites. *Genet. Res.* **76:** 227–236.

WEB SITE REFERENCES

ftp://ftp.tigr.org/pub/data/a_thaliana/ath1/PUBLICATION_RELEASE/PSEUDOMOLECULES/; TIGR ftp-site for *A. thaliana* sequence data.
ftp://ncbi.nlm.nih.gov/genomes/; NCBI genome data Web page.
ftp://ncbi.nlm.nih.gov/genbank; NCBI GenBank Web page.
http://btn.genomics.org.cn/rice; Genome database of Chinese Super Hybrid Rice.
www.fugu-sg.org; The IMCB—FUGU Genome Project Web page.
ftp://ftp.ensembl.org/pub/current_mouse/data/fasta/dna/; The Ensembl mouse genome project, current sequences.

Received April 9, 2003; accepted in revised form August 11, 2003.

A Microsatellite Variability Screen for Positive Selection Associated With the "Out of Africa" Habitat Expansion of *Drosophila melanogaster*

M. O. Kauer,[1] D. Dieringer[1] and C. Schlötterer[2]

Institut für Tierzucht und Genetik, 1210 Wien, Austria

Manuscript received October 9, 2002
Accepted for publication June 13, 2003

ABSTRACT

We report a "hitchhiking mapping" study in *D. melanogaster*, which searches for genomic regions with reduced variability. The study's aim was to identify selective sweeps associated with the "out of Africa" habitat expansion. We scanned 103 microsatellites on chromosome 3 and 102 microsatellites on the *X* chromosome for reduced variability in non-African populations. When the chromosomes were analyzed separately, the number of loci with a significant reduction in variability only slightly exceeded the expectation under neutrality—six loci on the third chromosome and four loci on the *X* chromosome. However, non-African populations also have a more pronounced average loss in variability on the *X* chromosomes as compared to the third chromosome, which suggests the action of selection. Therefore, comparing the *X* chromosome to the autosome yields a higher number of significantly reduced loci. However, a more pronounced loss of variability on the *X* chromosome may be caused by demographic events rather than by natural selection. We therefore explored a range of demographic scenarios and found that some of these captured most, but not all aspects of our data. More theoretical work is needed to evaluate how demographic events might differentially affect *X* chromosomes and autosomes and to estimate the most likely scenario associated with the out of Africa expansion of *D. melanogaster*.

THE "neutral theory of evolution" (KIMURA 1983) has dominated the study of molecular evolution for many years, but recent evidence suggests that beneficial mutations may be more abundant than previously assumed. For example, >40% of the amino acid replacements in Drosophila have been estimated to be driven by positive natural selection (FAY and WU 2001; FAY *et al.* 2002; SMITH and EYRE-WALKER 2002). Similarly it was suggested that amino acid replacements in *Drosophila melanogaster* are on average beneficial (BUSTAMANTE *et al.* 2002). High values of putatively positively selected amino acid replacements have also been estimated for humans (FAY *et al.* 2001). Hence, the application of suitable methods should make it feasible to systematically screen for beneficial mutations.

Recent advances in molecular biology allow the processing of multiple samples, which permits the analysis of multiple genetic markers in many individuals. Genome scans to test for the effect of directional selection rely on the concept of hitchhiking (MAYNARD SMITH and HAIGH 1974; KAPLAN *et al.* 1989), which describes the phenomenon that neutral variation flanking the selected site is also affected when beneficial mutations increase in frequency. Thus, a genome scan using a sufficiently high density of neutral markers could be used to identify genomic regions subjected to recent selective sweeps; this approach was recently termed hitchhiking mapping (HARR *et al.* 2002). Several genome scans based on microsatellite variation have already been performed for a range of organisms (HUTTLEY *et al.* 1999, 2000; PAYSEUR *et al.* 2002; SCHLÖTTERER 2002; VIGOUROUX *et al.* 2002; WOOTTON *et al.* 2002). A problem is to identify loci that differ from the rest of the genome, suggesting selection. Commonly used measures for inferring selection are increased linkage disequilibrium between loci (HUDSON *et al.* 1994; HUTTLEY *et al.* 2000; KOHN *et al.* 2000; WOOTTON *et al.* 2002), reduced polymorphism (NURMINSKY *et al.* 2001; HARR *et al.* 2002; KIM and STEPHAN 2002; SCHLÖTTERER 2002; WOOTTON *et al.* 2002) or a skewed allele frequency spectrum at individual loci (BRAVERMAN *et al.* 1995; PAYSEUR *et al.* 2002; VIGOUROUX *et al.* 2002).

Here we report a genome scan in *D. melanogaster*, specifically designed to identify genomic regions involved in adaptation to novel habitats. *D. melanogaster* originated in sub-Saharan Africa and colonized the rest of the world only ~10,000 years ago (DAVID and CAPY 1988). Previous studies have suggested that this habitat expansion involved the spread of beneficial mutations in non-African populations (BEGUN and AQUADRO 1993; KIRBY and STEPHAN 1996; KAUER *et al.* 2002) and furthermore African and non-African populations may also differ because of recent selection pressure imposed by man such as insecticide resistance (DABORN *et al.* 2001). By comparing putative ancestral African populations and derived European populations, we aimed to identify

[1]These authors contributed equally to this work.

genomic regions in *D. melanogaster* in which an allelic variant had been selected during/after the "out of Africa" colonization. In total, variability at 205 microsatellites was studied on the third and the *X* chromosome.

MATERIALS AND METHODS

Microsatellites: We surveyed 102 *X* chromosomal microsatellite loci and 103 microsatellites located on the third chromosome. For most loci, primers were designed using sequences available from the Drosophila genome project or the Drosophila whole-genome shotgun sequence (releases 1 and 2, http://flybase.bio.indiana.edu/). Microsatellites that were cloned in our lab were, like the above sequences, obtained from non-African flies. Only loci with an uninterrupted repeat structure longer than eight repeat units were chosen for primer design. All loci were typed in two African populations from Zimbabwe and various European populations. The main set of loci was typed without prior evidence for selection. After the first screen additional loci were genotyped in candidate regions for selection. A full list of the loci, populations, and basic statistics is available as online supplementary material (Table S1 at http://www.genetics.org/supplemental/). Loci that were also typed in the previous study of Harr *et al.* (2002) are indicated in Table S1. Primer sequences, annealing temperatures, repeat motifs, and cytological positions are available from the authors' web page (http://i122server.vu-wien.ac.at/). Microsatellite analysis followed standard protocols (Schlötterer and Zangerl 1999).

Fly strains: Zimbabwe flies were sampled from two locations, Sengwa Wildlife Reserve (ZS) and Harare, the capital of Zimbabwe (ZH), and were kindly provided by C. F. Aquadro and C.-I Wu. In previous studies we found different African populations, mainly from Kenya, to be very similar in variability levels to the populations from Zimbabwe (Harr *et al.* 2002; Kauer *et al.* 2002). Population structure is weak ($F_{ST} < 0.025$) and all populations share very high variability levels. The two Zimbabwean populations should, therefore, provide a reasonable sample for the ancestral population. To take into account inbreeding in these isofemale lines we calculated heterozygosity and variance in repeat number by averaging over 200 random data sets. In each data set, one allele was discarded from all inbred individuals. This procedure is implemented in the Microsatellite-Analyzer (MSA) software (Dieringer and Schlötterer 2003). European flies were from Poland (Katovice, 2000; collected by J. Gorczyca), Germany (Friedrichshafen, 1998 and Neustadt/Mannheim, 2000; collected by B. Harr, M. Kauer, and B. Zapfel, respectively), Switzerland (Nyon and Gotheron, 1998; collected by J. David), Russia (Moscow, 1998; collected by J. David), Austria (Vienna1, 1999 and Vienna2, 2000; collected by B. Harr and C. Schlötterer, respectively), Italy (Naples, 2001 and Rome, 1998; collected by C. Schlötterer), France (Prunay, 1998; collected by J. David), The Netherlands (Texel, 1999; collected by D. Slezak), and Denmark (Copenhagen, 1999; provided by V. Loeschke).

For each European population, 30 F_1 individuals were used. For the African populations a minimum of 20 individuals were typed for each locus.

Variability measures: Two measures of microsatellite variability were used: variance in repeat number (Goldstein and Clark 1995) and expected heterozygosity or gene diversity (Nei 1978). Both measures were corrected for small sample sizes by multiplying by $n/(n-1)$, where n is the number of chromosomes that were analyzed. For monomorphic loci, we assumed that one additional allele differed from the others (to avoid division by zero in the calculation of the ratio of European to African variabilities—see below); for variance in repeat number, we assumed that this allele differs from the other alleles by one mutation step (*e.g.*, 2 bp for dinucleotide repeats).

Measures to detect positive selection: Reductions in variability below neutral expectations at individual loci can be indicative of positive selection (Lewontin and Krakauer 1973; Maynard Smith and Haigh 1974; Galtier *et al.* 2000; Kim and Stephan 2002; Schlötterer 2002). We were interested in detecting such reductions in variability in non-African *D. melanogaster* populations, as these reductions may form the footprint of a selective sweep associated with the out of Africa habitat expansion. We used the ln RV and ln RH statistic to search for loci with levels of variability below neutral expectations. The test statistics consider the joint empirical distribution of all loci and identify loci that differ significantly in variability from the remainder of the genome. For each locus, the ratio of the genetic variabilities of two populations is calculated (Schlötterer 2002). Thus all loci have the same expectation irrespective of locus-specific mutation rates. Computer simulations indicate that, if the data do not contain a large number of invariant loci, the test statistics are relatively insensitive to demographic events such as bottlenecks, as demography affects all loci to a similar extent (Schlötterer 2002; C. Schlötterer and D. Dieringer, unpublished results).

The variance-based ln RV is calculated as

$$\ln[E(\mathrm{RV})] = \ln\left[E\left(\frac{\frac{1}{2}\theta_{\mathrm{Pop1}}}{\frac{1}{2}\theta_{\mathrm{Pop2}}}\right)\right] = \ln\left[E\left(\frac{(2N_{e_{\mathrm{Pop1}}}\mu)}{(2N_{e_{\mathrm{Pop2}}}\mu)}\right)\right] \cong \ln\left[\frac{E(V_{\mathrm{Pop1}})}{E(V_{\mathrm{Pop2}})}\right] \quad (1)$$

with $V = \theta/2$ (Moran 1975), where $\theta = 4N_e\mu$, N_e is the effective population size, and μ is the mutation rate. The corresponding equation for gene diversity is

$$\ln[E(\mathrm{RH})] = \ln\left[E\left(\frac{\theta_{\mathrm{Pop1}}}{\theta_{\mathrm{Pop2}}}\right)\right] = \ln\left[E\left(\frac{((1/1 - H_{\mathrm{Pop1}})^2 - 1)1/8\mu}{((1/1 - H_{\mathrm{Pop2}})^2 - 1)1/8\mu}\right)\right] \cong \ln\left[\frac{E((1/1 - H_{\mathrm{Pop1}})^2 - 1)}{E((1/1 - H_{\mathrm{Pop2}})^2 - 1)}\right], \quad (2)$$

where H is related to θ by the formula $H = 1 - (1/(1 + 2\theta)^{1/2})$ (Ohta and Kimura 1973).

For the remainder of the text, we use ln $R\theta$ for both ln RV and ln RH. Computer simulations indicate that under neutrality ln RV and ln RH values follow a Gaussian distribution (Schlötterer 2002; C. Schlötterer and D. Dieringer, unpublished results). Note that this assumption also holds when the number of loci used is smaller (*i.e.*, 120) than that of the original article (Schlötterer 2002; data not shown). Furthermore, we used computer simulations to verify that the assumption of normality for the ln $R\theta$ test statistic also holds when an ancestral and a recently derived population are compared (see web supplement, Table S2 at http://www.genetics.org/supplemental/). Given that ln $R\theta$ values are approximately normally distributed, the probability that a given locus deviates from neutrality can be obtained from the density function of a standard normal distribution. Hence, the observed ln $R\theta$ values need to be standardized by the mean and standard deviation of ln $R\theta$ values from putatively neutrally evolving loci typed in the same populations. The standardized distribution of ln $R\theta$ has therefore a mean of zero and a standard deviation of one. After standardization, 95% of the loci are expected to have values between 1.96 and −1.96. Those loci for which ln $R\theta$ values fall outside of this interval are considered as putatively selected loci. Coalescence-based

computer simulations (C. SCHLÖTTERER and D. DIERINGER, unpublished results) demonstrate a higher power for ln RH than for ln RV to detect selected loci, as ln RH has a smaller variance than ln RV. On the basis of the simulations, we mainly used the ln RH test statistic for the inference of positive selection. C. SCHLÖTTERER and D. DIERINGER (unpublished results) also noted that the type I error can be reduced two- to threefold when both test statistics, ln RH and ln RV, are considered jointly (*i.e.*, when the test is significant for both ln RV and ln RH).

We applied two methods to adjust significance levels of ln $R\theta$ for multiple testing, Bonferroni correction, and the combination of ln RV and ln RH (see above). While both methods are certainly valid for ruling out false positives, they are extremely conservative. The goal of this study, however, was to provide candidate loci for positive selection that deserve a more detailed analysis, and we therefore report all significant loci.

Identification of out of Africa sweeps: The main goal of this study was the identification of loci that show strongly reduced variability in European populations. To ensure the identification of a putative selective sweep associated with the habitat expansion of *D. melanogaster*, rather than local adaptation of a European population, we analyzed multiple European populations. For each locus, we took the arithmetic mean of variabilities over all populations in the two groups (European and African populations). The test statistics ln RV and ln RH are based on these averages. As we focused on positive selection in European populations, we concentrated on loci with significantly reduced variability.

To determine significance levels for the reduction of variability at individual loci, the empirical distribution of ln $R\theta$ values has to be standardized (see above). When most of the loci evolve neutrally and only a small number of loci are subject to directional selection, the mean and the standard deviation of the empirical ln $R\theta$ distribution can be used and the selected loci should fall into the lower tail of the distribution. When a substantial fraction of the analyzed loci have been affected by directional selection in the same population, this procedure is problematic because the whole distribution would be shifted to negative values and therefore only loci with the most extreme ln $R\theta$ values would fall into the lower tail of the distribution. Alternatively, a set of neutrally evolving loci could be used for standardizing. In this study we found the distribution of ln $R\theta$ values on the *X* chromosome to be shifted to negative values (see RESULTS). Previous studies also suggested that *X* chromosomal loci may be influenced by selection more than autosomal ones (ANDOLFATTO 2001b; KAUER *et al.* 2002). While the shift toward negative values of the *X* chromosomal ln $R\theta$ distribution could be also caused by demographic events (see DISCUSSION), we used two approaches to standardize the ln RV and ln RH distributions. First we standardized both chromosomal distributions by their own mean and standard deviation (standardization procedure 1). With this treatment demographic factors such as a bottleneck or differential reproductive success of males and females (CABALLERO 1994), which could potentially affect *X* and autosomes to a different extent (WALL *et al.* 2002), cannot bias the results. To account for the possibility that the *X* chromosome might have been more affected by selection than the autosome was, we also standardized the *X* chromosome with the mean and standard deviation of ln $R\theta$ of the third chromosome (standardization procedure 2). This second procedure, which *a priori* assumes more selection on the *X* chromosome, is not appropriate if the two types of chromosomes have been differentially affected by demographic events (*i.e.*, a bottleneck and/or differential reproductive success of males and females). Thus, the two methods of standardizing therefore provide a conservative and a nonconservative estimate of the number of candidate loci.

Using an analytical approach, we estimated whether demographic events could theoretically explain the difference between *X* and autosomal variation. Furthermore, we used coalescent simulations to estimate the influence of standardization procedure 2 on the number of false positives.

Analytical estimation of the relative variabilities of *X* chromosomes and autosomes: Ignoring new mutations, the genetic variability at time point T (θ_T) can be expressed as a function of the variability level at time point 0 in the past (θ_0), the new effective population size (N_e), which is assumed to remain constant, and the time (t) that elapsed between time points 0 and T:

$$\theta_T = \theta_0 \exp\frac{-t}{2N_e}. \tag{3}$$

This equation can be used to estimate the relative loss of variability on *X* chromosomes and autosomes after a bottleneck by taking into account the difference of (N_e) between the chromosomes. The population is not assumed to be in equilibrium, so that θ_0 is arbitrary.

Solving (3) for ($-t/2N_e$) yields

$$\frac{-t}{2N_e} = \ln\frac{\theta_T}{\theta_0} = \ln \text{RH}. \tag{4}$$

From (4) it follows that the expected ratio of ln θ_T/θ_0 on the *X* compared to an autosome is given by

$$\frac{(-t/2N_e)_X}{(-t/2N_e)_A} \approx \frac{(N_e)_A}{(N_e)_X} = k. \tag{5}$$

In Equation 5 t cancels out. Hence the relative loss of variability for *X* chromosomes and autosomes between time points 0 and T depends only on the ratio of the effective population sizes. For the same distribution of reproductive success for the two sexes the expectation is 1.33 irrespective of the time of the bottleneck (due to the absence of new mutations). Equations 4 and 5 offer the advantage that the loss of variability due to a bottleneck can be approximated by the ln $R\theta$ test statistic, which is easily obtained from experimental data. Thus, the ratio of ln $R\theta$ of the autosomes and *X* chromosomes is conservatively estimated by Equation 5.

The expected ratio of the effective population sizes of the chromosomes for a discrete-generation model can be calculated as

$$k = \frac{N_A}{N_X} = \frac{8(N_{ef} + 2N_{em})}{9(N_{ef} + N_{em})}, \tag{6}$$

where N_{ef} and N_{em} are the effective population sizes of females and males, respectively (CABALLERO 1994). This ratio in Equation 6 is bounded between 0.889 and 1.778 and equals 1.33 if $N_{ef} = N_{em}$.

Coalescent simulations based on population bottlenecks and differential effective population sizes of chromosomes: We used computer simulations to evaluate the consequences of various demographic scenarios. In a first set of simulations we assumed a constant ancestral effective population size of $N_e = 10^6$ for autosomes and 0.75×10^6 for *X* chromosomes. At time point t, a bottleneck instantaneously reduced the population size by a factor f. After the bottleneck the population increases exponentially in size until it reaches the current population size of 10^6 for autosomes and 0.75×10^6 for *X* chromosomes. The microsatellite mutation rate was set to $\mu = 5 \times 10^{-6}$ (SCHUG *et al.* 1998a; HARR and SCHLÖTTERER 2000; VAZQUEZ *et al.* 2000). The time point t of the bottleneck was scaled by $4N_e$, which differed for autosomal and *X* chromo-

TABLE 1

Mean microsatellite variabilities in European and African populations

	Heterozygosity		Variance in repeat no.	
	Europe	Africa	Europe	Africa
X chromosome	0.51 (0.2)	0.81 (0.13)	12.34 (18.96)	26.65 (42.67)
Chromosome 3	0.53 (0.18)	0.71 (0.14)	3.46 (4.55)	5.6 (8.58)

Standard deviations are shown in parentheses.

somal loci. To account for this we multiplied t for the X chromosomal simulations by the factor 1.33, which corresponds to the ratio of autosomal to X chromosomal population sizes assuming equal distributions of reproductive success for the two sexes. In a second set of simulations, we assumed that X chromosomes have a higher variability level ($\theta = 4N_e\mu$) than autosomes in the ancestral populations while the distribution of reproductive success was assumed to be the same for the two sexes after the population size reduction. Finally, we simulated scenarios where more variability is present on the X chromosomes in the ancestral population but the effective population size for females is lower than that of males after the bottleneck. These scenarios were simulated only for those variability levels that were most similar to the ones observed in the empirical data set (*i.e.*, $\theta_X = 3\theta_A$ in the ancestral population). Summary statistics for all simulations are shown in Table S3 at http://www.genetics.org/supplemental/. Coalescent simulations were performed with a modification of the Makesamples software (Hudson 2002), which incorporates a stepwise microsatellite mutation model (Ohta and Kimura 1973; T. Wiehe, unpublished results). The number of mutations occurring on a branch was converted into microsatellite mutations by adding or removing (with equal probability) one repeat unit for each mutation. To calculate ln RH, one set of data was generated using the ancestral settings without demography and one data set was generated with demography. Monomorphic loci were treated identically to experimental loci with no variability.

Coalescent simulations for evaluating the influence of nonstepwise mutations on ln RH and ln RV: We relied on a commonly used coalescent-based computer simulation algorithm (Hudson 1990), modified to take into account the stepwise mutation behavior of microsatellites (see above). In addition to stepwise changes in repeat number, we also simulated insertions/deletions (indels) occurring in the flanking sequence. The indel size for most simulations was taken from a uniform distribution between 1 and 20 repeat units; for a subset of simulations the indel size was taken from a uniform distribution between 1 and 10 repeat units. In our simulations we allowed for different mutation rates for microsatellites (slippage) and indels. We simulated different frequencies of nonstepwise mutations and also different maximum step sizes. In each simulation run, 1 locus out of 100 was subjected to directional selection, and 1000 replicas were simulated for each parameter combination. Selection was simulated as an instantaneous reduction in the population size. All simulation runs assumed a selective sweep, which occurred $0.05 \times 2N_e$ generations ago and reduced variability by a factor of 0.01. Summary statistics for all simulations are shown in Table S4 at http://www.genetics.org/supplemental/.

Allele excess: Allele excess was determined with the Bottleneck program (Cornuet and Luikart 1996). Bottleneck provides P values for single loci and also deviations from the strict stepwise mutation model (SMM) can be included [two-phase model (TPM)]. Note that the program Bottleneck calculates "heterozygote deficiency" as a measure of allele excess. Here, however, we use the term "allele excess" as a synonym for heterozygote deficiency.

Genetic distance and F_{ST}: Genetic distances (defined as 1 − proportion of shared alleles) and unbiased estimators of F_{ST} (Weir and Cockerham 1984) between populations were calculated with the software Microsatellite-Analyzer (Dieringer and Schlötterer 2003). Significance levels for F_{ST} values were calculated by permuting genotypes among populations (10,000 times) and were corrected for multiple tests (Sokal and Rohlf 1995).

Recombination rates: Recombination rates (in percentage of recombination per kilobase and generation) of genomic sequence and generation were calculated as outlined in Comeron *et al.* (1999) with a program kindly provided by J. M. Comeron. We did not include loci with recombination rates <0.0001% recombination per kilobase after adjusting for zero recombination in males (*i.e.*, multiplying by 0.67 for the X chromosome and by 0.5 for the third chromosome). The rationale for this was that, in genomic regions with low recombination rates, hitchhiking events affect very large regions, thus making the identification of the target of selection impossible (Schlötterer and Wiehe 1999). This selection criterion mainly excluded centromeric and telomeric regions.

RESULTS

Consistent with previous reports (Begun and Aquadro 1993; Andolfatto 2001b; Kauer *et al.* 2002), African flies were more variable than European ones (Table 1). Mean microsatellite variabilities were higher than recently reported (Kauer *et al.* 2002). Furthermore, no correlation between recombination rate and microsatellite variability was detected, irrespective of whether chromosomes were analyzed separately or jointly (data not shown). The discrepancy between the data reported here and our previous report (Kauer *et al.* 2002) can be attributed to the lack of microsatellites located in genomic regions with low recombination rates in this study (see materials and methods). The correlation of recombination rate and microsatellite variability that was recently found by Kauer *et al.* (2002) was mainly due to very low levels of variability in regions of very low recombination rate.

In our analysis, we averaged microsatellite variabilities across populations. As the set of populations analyzed differed among loci, this could have biased our analysis. Consistent with previous reports (Begun and Aquadro 1993; Caracristi and Schlötterer 2003), differentia-

TABLE 2

Genetic differentiation between populations on the *X* chromosome

D/F_{ST}[a]	Rome	Friedrichshafen	Gotheron	Nyon	Copenhagen	ZH	ZS
Rome		0.04	0.03	0.04	0.06	0.23	0.25
Friedrichshafen	0.22		0.04	0.05	0.07	0.23	0.24
Gotheron	0.21	0.22		0.04	0.07	0.23	0.24
Nyon	0.20	0.21	0.19		0.04	0.24	0.25
Copenhagen	0.22	0.25	0.25	0.22		0.24	0.25
ZH	0.65	0.66	0.66	0.67	0.67		−0.001
ZS	0.67	0.67	0.67	0.68	0.68	0.42	

Country origin of populations: Denmark (Copenhagen), France (Gotheron), Germany (Friedrichshafen), Italy (Rome), Switzerland (Nyon), Zimbabwe (ZH/Harare and ZS/Sengwa).

[a] Above diagonal, F_{ST} values (all values are significant, $P < 0.01$); below diagonal, genetic distance (1 − proportion of shared alleles).

tion among European populations was much lower than that between European and African populations (Tables 2 and 3). *X* chromosomal loci were more differentiated than loci on chromosome 3 between European and African populations. This difference cannot be attributed to the choice of populations, as different European populations gave similar results for loci located on the same chromosome (Tables 2 and 3). Variability levels were also very similar among European populations.

Influence of nonstepwise mutations and indel polymorphisms on ln RH and ln RV: Before applying the ln *R*θ test statistics to our data we wanted to examine a critical aspect of the two test statistics used, ln RH and ln RV: their robustness to deviations from the strict stepwise mutation model as described by OHTA and KIMURA (1973). C. SCHLÖTTERER and D. DIERINGER (unpublished results) showed that ln RH and ln RV are not perfectly correlated when a strict stepwise mutation model is assumed. The correlation between ln RH and ln RV in our data is lower than that found by C. SCHLÖTTERER and D. DIERINGER (unpublished results) for a strict stepwise mutation model (*X* chromosome, $r = 0.59$; chromosome 3, $r = 0.46$; simulation under SSM, $r \approx 0.7$–0.8, $P < 0.01$, Spearman rank correlation), suggesting some deviation from the strict stepwise mutation model. In simulations that allow for some mutations of multiple microsatellite repeat units (TPM; DI RIENZO *et al.* 1994), the correlation between ln RH and ln RV is lower than that for SSM. Thus, a two-phase microsatellite mutation model may be sufficient to explain the low correlation between ln RH and ln RV in our data. On the other hand indel polymorphisms in the flanking sequence of the microsatellite could also have a similar effect. Such indel polymorphisms are frequent in Drosophila (COLSON and GOLDSTEIN 1999), so we performed computer simulations to estimate the influence of indels in the flanking sequence of a microsatellite. Consistent with the results of C. SCHLÖTTERER and D. DIERINGER (unpublished results) for the two-phase model, we found that ln RH is quite insensitive to indel polymorphisms in the flanking sequence, whereas the power of ln RV drops (Table S4 at http://www.genetics.

TABLE 3

Genetic differentiation between populations on chromosome 3

D/F_{ST}[a]	Moscow	Texel	Prunay	Vienna1	Katovice	Naples	Neustadt	Vienna2	ZH	ZS
Moscow		0.05	0.03	0.08	—	—	—	—	0.12	0.16
Texel	0.27		0.04	0.08	—	—	—	—	0.12	0.16
Prunay	0.22	0.24		0.07	—	—	—	—	0.13	0.17
Vienna1	0.26	0.30	0.25		0.00	0.01	0.02	—	0.11	0.17
Katovice	—	—	—	0.21		0.04	0.02	0.01	0.14	0.18
Naples	—	—	—	0.28	0.18		0.01	0.04	0.12	0.17
Neustadt	—	—	—	0.35	0.16	0.15			0.14	0.16
Vienna2	—	—	—	—	0.17	0.21			0.11	0.17
ZH	0.47	0.48	0.50	0.46	0.42	0.39	0.40	0.40		0.01
ZS	0.51	0.55	0.55	0.53	0.48	0.45	0.42	0.48	0.28	

Country origin of populations: Austria (Vienna1 and Vienna2), France (Prunay), Germany (Neustadt), Italy (Naples), Netherlands (Texel), Poland (Katovice), Russia (Moscow), Zimbabwe (ZH/Harare and ZS/Sengwa).

[a] Above diagonal, F_{ST} values (all values are significant, $P < 0.01$); below diagonal, genetic distance (1 − proportion of shared alleles).

org/supplemental/). Interestingly, the indel mutation rate had almost no effect on the power of both ln RV and ln RH. However, the power of ln RV decreased with an increasing mean indel size (Table S4). The lower power of the ln RV test statistic is the outcome of an increased variance of ln RV values among loci.

Indel mutations become even more problematic when only a small number of loci are affected. As loci with indel mutations have a higher variance in ln RV, they are more frequently located in the tails of the distribution when analyzed jointly with loci varying only in microsatellite repeat number. Hence, indels in the flanking sequence not only reduce the power of ln RV to detect selective sweeps, but also increase the type I error rate. Therefore we relied mainly on ln RH for the identification of candidate loci for positive selection.

Identification of candidate loci: Consistent with previous computer simulations (SCHLÖTTERER 2002; C. SCHLÖTTERER and D. DIERINGER, unpublished results) we found that ln RV, and especially ln RH values, were approximately normally distributed. Nonstandardized ln $R\theta$ values show more negative values on the X than on the third chromosome (mean/SD of ln RH, X, $-2.37/1.37$; third chromosome, $-1.18/0.94$), indicating a larger loss of variability on the X chromosome than on the third chromosome (KAUER *et al.* 2002). Furthermore the X chromosomal distribution is broader, as indicated by the higher standard deviation (Figure 1).

As outlined in MATERIAL AND METHODS, we pursued two different approaches to identify candidate loci for positive selection. In standardization procedure 1 we treated each chromosome separately and standardized the ln $R\theta$ values by the mean and standard deviation from all loci mapping to the same chromosome. Using this approach, we identified a conservative set of candidate loci. In standardization procedure 2, the statistical significance of the ln $R\theta$ values of individual loci on the X chromosome was determined by standardizing with the distribution of ln $R\theta$ of the third chromosome. In the absence of demographic events, this procedure is not expected to bias the results and may be even favorable when a larger number of selective sweeps is expected on the X chromosome. Demographic events, however, may lead to a more pronounced loss in variability at X-linked loci, so that a larger number of false positives may be obtained.

Standardization procedure 1: When both chromosomes were standardized with their own distribution of ln RH and ln RV, seven loci on the X chromosome and eight loci on the third chromosome showed a significant reduction in variability by either ln RH or ln RV or both (Table 4, Figure 1). Using ln RH, four loci were located in the lower tail and two loci in the upper tail of the X chromosomal distribution. On the third chromosome, seven significant loci were identified, five of which were in the lower tail of the distribution. The ratios of significant loci in the lower and the upper tail of the ln RV distribution were 4:2 on the X chromosome and 3:1 for chromosome 3. No difference in the power of the two test statistics was observed and only one locus on the X chromosome was significant for both ln RH *and* ln RV tests. After adjusting the α-value for multiple testing by Bonferroni correction (*i.e.*, ln RH and ln RV < -3.67, Table 4) none of the loci in this set remained significant.

Standardization procedure 2 for the X chromosome: A total of 30 loci on the X were identified by either ln RH or ln RV using this method (Table 4, Figure 1). Because of the larger reduction of variability on the X chromosome, all 28 loci identified with ln RH were located in the lower tail and none in the upper tail of the third chromosomal ln RH distribution. Using ln RV 10 loci were found in the lower tail and 4 loci in the upper tail of the distribution. Eight loci identified with ln RH remained significant after Bonferroni correction.

Considering that deviations from the strict stepwise mutation model have a strong impact on the ln RV test statistic, and given that flanking sequence indels occur frequently in Drosophila (COLSON and GOLDSTEIN 1999), we considered only those loci with a significant ln RV as candidates for a selective sweep when ln RH also indicated a loss of variability. On the basis of this criterion we rejected two loci on the third chromosome (3L16575599gt and 3R5316419ta) as false positives. Both loci show large allele gaps in the African but not in the European population. In the absence of indel polymorphisms, the repeat number at any microsatellite allele can be determined by subtracting the flanking sequence length (obtained from the published genomic sequence of *D. melanogaster*) from the PCR product length. For these two loci we inferred a microsatellite length of 21 and 51 repeats. Given that long microsatellite alleles are rare in *D. melanogaster* (SCHUG *et al.* 1998b; BACHTROG *et al.* 1999; HARR and SCHLÖTTERER 2000), we regard it as likely that an insertion in the flanking sequence has occurred. For three other loci (X3642495gct, 66-95-3, and 3L2299865), which were significant by the ln RV test statistic, we also observed a loss of variability with ln RH (*i.e.*, ln RH < -1.48, Table 4). Despite the relatively weak support for these loci we included them as putative candidates for a selective sweep associated with the out of Africa habitat expansion of *D. melanogaster* to avoid type 2 error. As mentioned above the remaining eight loci were significant on both test statistics.

All candidate loci are shown in Figure 1, where the confidence limits for both standardization procedures are also drawn. Visual inspection suggests no obvious spatial clustering of significant loci on the third chromosome and of the nonconservative X chromosomal set. Three of the four significant X chromosomal loci (based on the conservative standardization procedure 1) are located in relatively close proximity to each other.

As microsatellite mutation rates are dependent on repeat length (HARR and SCHLÖTTERER 2000; SCHLÖTTERER 2000), an important prerequisite for the applica-

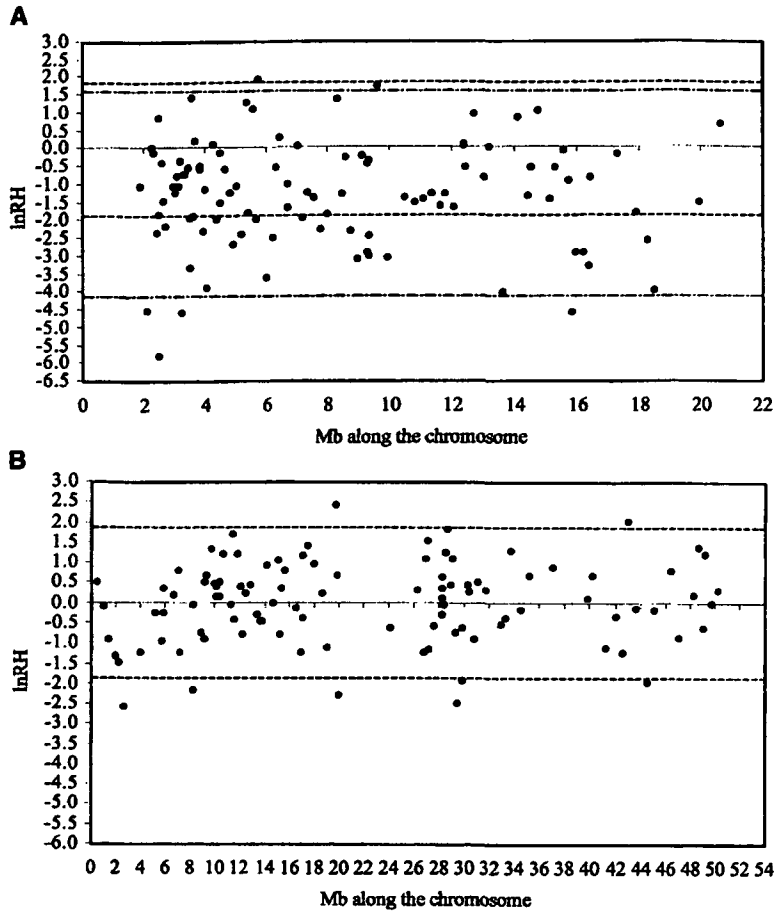

FIGURE 1.—Plot of ln RH along chromosomes. For all loci ln RH values are shown. (A) *X* chromosome. (B) Chromosome 3. Loci are aligned according to their chromosomal position; dashed lines indicate the limits of the 95% confidence interval (−1.96, +1.96) for standardization with loci from the third chromosome. The dashed-dotted lines in A indicate the 95% confidence interval for the standardization with *X* chromosomal loci.

tion of the ln RH and ln RV test statistic is that the repeat number does not differ between African and European *D. melanogaster* populations. For the candidate loci, the inferred mean repeat number of European alleles was not significantly different from the mean repeat number in the African populations ($P > 0.8$, sign test). Furthermore, the mean repeat length in European populations did not differ between the candidate loci and the others ($P > 0.2$, Mann-Whitney *U*-test).

Allele excess and genetic distance of candidate loci: An excess of rare alleles is often taken as evidence for selection (TAJIMA 1989; BRAVERMAN *et al.* 1995; PAYSEUR *et al.* 2002; VIGOUROUX *et al.* 2002). Computer simulations indicate that microsatellite loci with low gene diversity are biased toward an excess of rare alleles even under neutrality (C. SCHLÖTTERER and M. O. KAUER, unpublished results). Thus, we did not consider this test statistic to identify candidate loci, but compared it to our set of candidate loci on the basis of ln RH and ln RV results. Assuming a single SSM we found 12 loci with a significant excess of alleles. Eight of these loci were included in the nonconservative set of candidate loci and 1 in the conservative set. The mean allele excess of the nonconservative candidate loci is significantly higher than that for the rest of the *X* chromosomal loci ($P < 10^{-5}$ for SSM and $P < 10^{-4}$ for TPM, Mann-Whitney *U*-test). The same trend can be seen on the third chromosome, but power is very low because there are only 6 candidate loci ($P = 0.086$ for SSM and $P = 0.028$ for TPM, Mann-Whitney *U*-test).

A selective sweep removes allelic variation around a selected site. Thus the genetic distance between a selected and neutrally evolving population is increased at a locus affected by a selective sweep. As absolute genetic distances were found to be superior to relative measures of genetic distance (*e.g.*, F_{ST}) for the comparison of variability in selected and neutrally evolving regions (CHARLESWORTH 1998), we used the proportion of shared alleles as the genetic distance measurement. Mean genetic distances between European and African populations were higher for the nonconservative set of candidate loci than for the rest of the loci, but the difference is not statistically significant on the *X* chromosome (chromosome 3, $P = 0.008$; *X* chromosome, $P = 0.29$; Mann-Whitney *U*-test).

Analysis of the genomic region flanking a candidate locus: This study's purpose was to identify regions in the genome of *D. melanogaster* that are reasonable candidates for a thorough examination of their adaptive value in European populations. In the analysis above we pre-

TABLE 4

Candidate loci for positive selection

Locus[a]	Mb	Arm	Band	ln RH	ln RV	Heterozygosity[b] Africa	Heterozygosity[b] Europe	Variance in repeat no.[b] Africa	Variance in repeat no.[b] Europe	Allele excess (bottleneck)[b,d] TPM	Allele excess (bottleneck)[b,d] SSM	Genetic distance[b]: D
						X chromosome						
P3B02gt[c,e,f]	*2.51*	*X*	*3c*	*−5.81*	*−6.34*	*0.85*	*0.03*	*22.11*	*0.02*	Monomorphic	Monomorphic	*0.68*
X3219363gt[c,e]	*3.22*	*X*	*3d*	*−4.60*	*−2.24*	*0.94*	*0.31*	*30.98*	*1.61*	*−5.74****	*−8.48****	*0.81*
X15830711gt[c,e]	15.83	*X*	14a	−4.57	−0.54	0.92	0.24	153.69	46.75	−1.24	−1.84	0.76
X2102441ct[c,e]	2.10	*X*	2f	−4.56	−1.77	0.97	0.55	25.15	2.14	−0.07	−0.57	0.70
X13624957[c,f]	*13.62*	*X*	*12c*	*−4.02*	*−3.69*	*0.83*	*0.10*	*3.91*	*0.05*	*−0.47*	*−0.52*	*1.00*
X18472039ca[c,f]	*18.47*	*X*	*17d*	*−4.01*	*−3.91*	*0.76*	*0.06*	*8.33*	*0.08*	*−1.38*	*−1.55*	*0.51*
X4071888gt	4.07	*X*	4b	−3.88	−1.44	0.92	0.33	43.41	5.20	−3.53	−5.61*	0.68
X5973753gt	5.97	*X*	5d	−3.63	−1.80	0.94	0.51	32.25	2.65	−2.08	−3.71*	0.70
X3516772ga	*3.52*	*X*	*3f*	*−3.34*	*−2.37*	*0.92*	*0.43*	*8.40*	*0.38*	*−0.53*	*−1.03*	*0.67*
DS09020	16.39	*X*	15a	−3.28	−1.68	0.91	0.42	7.49	0.70	−1.57	−2.67*	0.81
X8956947gt	8.96	*X*	8d	−3.09	−1.54	0.96	0.70	69.87	7.58	−0.29	−1.43	0.83
X9928573gt	*9.93*	*X*	*9b*	*−3.06*	*−2.99*	*0.88*	*0.33*	*25.71*	*0.62*	*−1.51*	*−2.35*	*0.72*
P08E01gt	9.33	*X*	8e	−3.02	−0.65	0.83	0.22	6.80	1.85	−2.75*	−3.85*	0.52
X16203512gt	16.20	*X*	14d	−2.93	0.48	0.75	0.13	1.11	0.97	−1.46	−1.86	0.82
X9312943	9.31	*X*	8e	−2.93	0.26	0.87	0.34	3.89	2.72	−2.65*	−3.76*	0.95
X15959225ca	15.96	*X*	14c	−2.91	−1.07	0.75	0.13	4.73	0.83	−0.48	−0.61	0.64
X4944599ca	4.94	*X*	4d	−2.68	−1.63	0.91	0.50	11.51	1.13	−0.04	−0.44	0.65
X18283112ta	18.28	*X*	17c	−2.59	1.36	0.92	0.57	8.40	18.36	−0.18	−0.81	0.76
X6213328ca	6.21	*X*	5f	−2.51	−0.11	0.91	0.54	14.96	7.16	−2.45*	−4.21*	0.69
X9325355	9.33	*X*	8e	−2.43	−1.81	0.81	0.26	6.10	0.50	−0.52	−0.96	0.76
X5179712gt	5.18	*X*	4f	−2.39	−0.44	0.92	0.61	18.87	6.41	−1.72	−3.19*	0.71
P3B02 atc[c,f]	*2.42*	*X*	*3b*	*−2.37*	*−3.88*	*0.68*	*0.12*	*10.09*	*0.10*	*−0.78*	*−0.86*	*0.46*
X8756567gt	8.76	*X*	8c	−2.33	−1.30	0.84	0.36	9.06	1.26	−0.61	−1.02	0.52
DS00146	*3.96*	*X*	*4b*	*−2.32*	*−2.62*	*0.87*	*0.43*	*8.07*	*0.28*	*0.03*	*−0.30*	*0.50*
X7809164ca	7.81	*X*	7d	−2.27	−1.09	0.84	0.36	6.71	1.15	−1.88	−2.74	0.69
DS06335a	2.70	*X*	3c	−2.17	−0.16	0.93	0.67	11.65	5.26	−0.74	−1.82	0.66
X4364768gt	4.36	*X*	4c	−1.99	−0.83	0.87	0.47	3.57	0.81	−0.34	−1.00	0.54
DS00589	5.67	*X*	5c	−1.98	0.96	0.88	0.52	5.87	8.50	0.09	−0.31	0.74
X3642495gct	3.64	*X*	3f	−1.90	−2.03	0.83	0.39	11.61	0.75	−0.53	−1.17	0.65
66-95-3	2.66	*X*	3c	−1.49	−1.97	0.67	0.22	4.74	0.33	−0.86	−1.22	0.43
						Chromosome 3						
3L2674504gt	2.67	3L	63A	−2.56	−0.84	0.87	0.39	6.36	1.42	−1.32	−1.88	0.59
3R5511972ca	5.51	3R	85E	−2.49	−0.27	0.67	0.10	0.54	0.22	−1.25	−1.52	0.72
3L20028475ca	20.03	3L	77a	−2.29	−1.41	0.77	0.23	6.21	0.77	−0.87	−1.40	0.56
3L8253482ca	8.25	3L	66c	−2.15	−0.71	0.78	0.25	1.79	0.46	−1.50	−2.30	0.76
3R20604755ta	20.60	3R	96b	−1.98	−1.64	0.86	0.47	10.82	1.06	−0.94	−1.60	0.79
3L16575599gt	16.58	3L	73c	−0.12	−2.73	0.84	0.70	67.35	2.13	−0.05	−0.81	0.61
3R5316419ta	5.32	3R	85d	−0.74	−2.43	0.60	0.26	8.78	0.38	−1.39	−2.04	0.26
3L/2299865	2.30	3L	12b	−1.48	−2.23	0.67	0.22	3.81	0.20	−2.26	−3.26*	0.39

[a] Ordered by ln RH; loci that are significant with ln RH *and* ln RV are italic.
[b] Averages over all European populations.
[c] ln RH or ln RV significant after Bonferroni correction.
[d] Significant values are indicated as $*P < 0.05$, $**P < 0.01$, $***P < 0.001$.
[e] Significant with ln RH when standardized with ln RH distribution from the *X* chromosome, *i.e.*, "conservative" candidates on the *X*.
[f] Significant with ln RV when standardized with ln RV distribution from the *X* chromosome, *i.e.*, "nonconservative" candidates on the *X*.

sented both nonconservative and conservative estimations of the number of loci that may have been affected by selection. One approach to verifying a candidate region takes advantage of the fact that a selective sweep reduces variability in the genomic region flanking the selected site (Maynard Smith and Haigh 1974; Wiehe 1998; Kim and Stephan 2002). While some variation in variability among genomic regions is expected under neutrality, a selective sweep renders neighboring sites more correlated than under neutrality (Kaplan *et al.* 1989;

KIM and STEPHAN 2002). Therefore determining variability levels around candidate loci could provide a tool for distinguishing between a neutral and a selection scenario (NURMINSKY *et al.* 2001; HARR *et al.* 2002; KIM and STEPHAN 2002; WOOTTON *et al.* 2002). Recently, HARR *et al.* (2002) analyzed linked microsatellites in 850 kb of genomic sequence and identified three putative out of Africa sweeps. For each sweep the authors described a "valley" of reduced variability around the putative target of selection. On the basis of the results of HARR *et al.* (2002), we estimated that a genomic region of up to 100 kb is affected by a selective sweep. Thus, we include in Table S5 (http://www.genetics.org/supplemental/) all microsatellite loci that were characterized within a 100-kb region around a candidate locus. Some of the loci in these 100-kb regions were genotyped without knowledge of the selective sweep (*i.e.*, before the availability of the full genomic sequence of *D. melanogaster*), and we also specifically designed PCR primers for loci falling into the 100-kb region around candidate loci (Table 4).

Table S5 shows that some of the candidate loci fall into the same genomic region and so may indicate the same putative selective sweep. This clustering would reduce the number of independent candidate sweep regions on the *X* chromosome from 30 to 27 (nonconservative set before correction for multiple tests). Consistent with HARR *et al.* (2002) we observed several regions for which variability was significantly reduced for more than one locus (*e.g.*, regions 6, 20, 22, 23, 27, and 32 in Table S5). For some other regions we detected one significant candidate locus, which was flanked by microsatellites with reduced variability (*i.e.*, negative ln RH values) but lacking statistical significance (*e.g.*, regions 3 and 25).

Note that the regions around P3B02gt, 66-95-3, and 3L2299865gt have already been reported and analyzed in detail by HARR *et al.* (2002) but have been typed here using a different set of European populations. Interestingly, two other groups have independently inferred a putative selective sweep in the region around P3B02gt (J. POOL and C. AQUADRO, personal communication; D. DE LORENZO and W. STEPHAN, personal communication).

A detailed discussion and a list of genes in candidate regions can be found in the web supplement accompanying this article (Tables S5 and S6; http://www.genetics.org/supplemental/).

DISCUSSION

In this microsatellite variability screen, we have found a more pronounced reduction in variability in non-African *X* chromosomes than on autosomes. This unbalanced reduction of microsatellite variability could arise from a bottleneck associated with the habitat expansion of *D. melanogaster* (FAY and WU 1999; WALL *et al.* 2002), biased distributions of reproductive success between the sexes (CHARLESWORTH 2001), or from multiple selective sweeps (MAYNARD SMITH and HAIGH 1974), or a combination of these. While selective sweeps affect individual sites, the first two neutral scenarios are genome-wide effects, which affect variability levels at all loci with stochastic variation among them (HUDSON *et al.* 1987; GALTIER *et al.* 2000; ANDOLFATTO 2001a).

Influence of bottlenecks and skewed reproductive success: Due to the different reduction of microsatellite variability on the chromosomes, we identified very different numbers of candidate loci with our two standardization procedures. While all of the loci in the conservative set are good candidates for positive selection outside of Africa, in the nonconservative set there may be a higher number of false positives than indicated by the nominal α-value of 0.05. This number depends on the demographic scenario that was associated with the colonization of non-African habitats by *D. melanogaster*. Therefore, to evaluate whether our data could be explained under neutrality, we explored a range of demographic models analytically and with coalescence simulations.

Analytical approach: Using the analytical approach outlined in MATERIALS AND METHODS (Equations 3–6), we estimated whether the different behavior of *X* chromosomes and autosomes could be explained by a bottleneck and/or skewed sex ratios. Because of the problematic properties of ln RV for nonstepwise mutations (see RESULTS) we relied on ln RH. Assuming no sex differences in the distribution of reproductive success outside of Africa, it follows from Equation 5 that the expectation for ln RH is identical for both chromosomes when autosomal ln RH values are multiplied by 1.33. Importantly, this expectation is independent of the relative variability levels before the bottleneck (*i.e.*, the ratio of θ_0 for the chromosomes in Africa). Therefore multiplying autosomal ln RH values by 1.33 assumes an equal distribution of reproductive success for the two sexes only outside of Africa (between time points 0 and *T*). In contrast to this expectation, we found that the mean of ln RH values for the *X* chromosome are significantly more negative than the ln RH values of autosomal loci (*X*, −2.37; *A*, −1.57; $P < 0.0001$, *t*-test). Thus, relative to the autosome the *X* chromosome lost more variability than expected. This result is not affected by the different levels of variability on *X* chromosomes and autosomes in the African population (BEGUN and AQUADRO 1993; ANDOLFATTO 2001b; KAUER *et al.* 2002). Nevertheless, it applies only if males and females have the same distribution in reproductive success in non-African populations.

Similarly, to estimate the influence of our standardization procedure 2, we calculated the number of significant *X* chromosomal ln RH values when standardized with the mean and the standard deviation of ln RH values from the third chromosome multiplied by 1.33. Standardizing in this way yields twice as many significant

ln RH values on the X chromosome (11 loci) as on the autosome (5 loci). These 11 X chromosomal loci are the ones with the most negative ln RH values in Table 4.

Another factor that could influence the distribution of ln RH is differential reproductive success of males and females. An effective surplus of males in Europe (BOULETREAU 1978; CHARLESWORTH 2001; between time points 0 and T) would reduce the effective population size of the X relative to the autosome, therefore changing the expectation in Equation 5. We tested the influence of skewed effective population sizes of chromosomes by assuming effective male:female ratios of 5:1 and 10:1 (k = 1.63 and 1.69, respectively). The ratio of chromosomal variability levels (k) was calculated using Equation 6; k was then used as the expectation in Equation 5 and the ln RH values on autosomes were corrected accordingly. After correcting for a 5-fold excess of males the mean ln RH of the X is still significantly more negative than that on the autosome (X, −2.37; A, −1.92; P = 0.03, t-test); after correction for a 10-fold excess of males the difference is marginally significant (X, −2.37; A, −2.00; P = 0.08, t-test).

With the analytical analyses we explored in a simple way (ignoring new mutations) the combined effect of a bottleneck and skewed sex ratios on the relative loss of variability on the X and the autosome outside of Africa. Assuming that different levels of variability among X chromosomes and autosomes are caused by different distributions of reproductive success for males and females, it has to be noted that an inverse difference must be present among African and non-African populations, as in Africa X chromosomes are more variable. Note that the analytical analyses implicitly assumed a bottleneck outside of Africa, because otherwise no variability would have been lost. The results from these analyses indicate that our data can be explained under certain demographic scenarios.

Coalescence simulations: Despite the fact that *D. melanogaster* microsatellite mutation rates are low (SCHUG *et al.* 1997; SCHLÖTTERER 2000) and non-African variation appears to be a subset of African variation (ANDOLFATTO 2001b; SCHLÖTTERER and HARR 2002), we evaluated the impact of demographic scenarios, which included mutations after the bottleneck. As outlined in MATERIALS AND METHODS, for standardization procedure 2 we used ln RH values from the third chromosome to standardize X chromosomal ln RH values. This procedure can bias the test statistic toward a higher number of nonneutral loci in non-African populations when demographic events were associated with the habitat expansion of *D. melanogaster*. To quantify this effect, we performed coalescent simulations for X-linked and autosomal loci under a range of demographic scenarios. Table S3 summarizes these simulations and shows the ratio of false positives in the postbottleneck population on the basis of the standardization using autosomal loci relative to the standardization using X chromosomal loci; a ratio of 2 would indicate that standardization method 2 (with autosomes) leads to twice as many "significant" loci as the conservative standardization method 1.

Computer simulations that assumed the same distribution of reproductive success for males and females in Africa did not result in a large excess of false positives when standardization procedure 2 was used (Table S3, 1a–1h). Nevertheless, this set of simulations failed to capture the higher X chromosomal variability in Africa. Therefore, for another set of simulations we assumed different θ-values for X chromosomes and autosomes (Table S3, 2a–4g). The best fit to the observed African variation was obtained when θ of the X chromosome was 2–4 times as high as θ on the autosome (Table 1, Table S3). This is not surprising as the observed mean value of θ (based on heterozygosity) of the X chromosome in our data is ~2.5 times the one on the autosome in Zimbabwe [Table 1, where heterozygosity (H) is related to θ by the formula $H = 1 - (1/(1 + 2\theta)^{1/2})$ (OHTA and KIMURA 1973)]. For some demographic scenarios (*e.g.*, Table S3, 3g, 3c, and 4c) we found ln RH and the heterozygosity of the derived population to be very similar among simulated and experimental data (Table 1 and Table S3). Importantly, the number of false positives was strongly increased when we applied standardization procedure 2 to data sets generated under these demographic models. An aspect of the experimental data that these simulations could not reproduce is the variance of ln RH. While with a higher impact of selection on the X a higher variance of ln RH could be expected on the X chromosomes (KAUER *et al.* 2002), no difference could be noted between X chromosomes and autosomes under these models.

Finally, in simulations 5a–5h (Table S3) we combined a threefold excess of variation (θ) for African X chromosomes relative to autosomes (as for simulations 3a–3h in Table S3) with an unequal distribution of reproductive success of the two sexes in non-African populations (Table S3, 5a–5h). The effective population sizes of males and females in the postbottleneck population were set to 5:1. Three aspects could be highlighted in these simulations: (i) some parameter combinations closely matched the observed levels of variability in African and non-African chromosomes; (ii) the number of false positives increased when standardization procedure 2 was applied; and (iii) for some scenarios the variance in ln RH was increased on the X chromosome, although to a lesser extent than in the empirical data.

Analytical analyses and simulations indicated that the standardization of X chromosomal data with autosomal data may be associated with an error leading to an overestimation of the number of selected loci on the X chromosome. The magnitude of this error can be so large as to explain a large number of candidate loci we found on the X chromosome when using standardization procedure 2. The actual error that is made could,

however, be estimated only if the true demographic scenario was known. An exhaustive likelihood approach where the probability to observe the data assuming different demographic scenarios and also incorporating selection will be a worthwhile task for future analysis but is beyond the scope of this study. Another factor that our simulations may not have captured is a different mutation rate on *X* chromosome and autosome. This could explain the difference in microsatellite variability in Africa and could in principle bias the distribution of the variability reduction in non-African populations. A conservative estimation of the number of candidate loci on the *X* is given by standardization procedure 1.

Positive selection in non-African populations of *D. melanogaster*: An alternative explanation for the larger reduction of variability on the *X* chromosome is a higher impact of selection on the *X* chromosome. This could be the result of hemizygosity of the *X* chromosomes in males or the result of more beneficial alleles on the *X* chromosomes (AQUADRO *et al.* 1994; KAUER *et al.* 2002). In support of a nonneutral interpretation, KAUER *et al.* (2002) found the loss of variability outside of Africa to be most pronounced in regions of low recombination rate on the *X* chromosome but not on the autosomes. The higher variance of ln RH on the *X* chromosome than on autosomes that could not be reproduced by simulations could also be attributed to more selection on the *X* chromosome, as many selected loci would provide more extreme values in the distribution of ln RH (KAUER *et al.* 2002).

Although, as noted above, an exhaustive examination of demographic scenarios remains to be done, an empirical approach to disentangle the false positives from truly selected loci could be to gather more information about all candidate loci. As presented in RESULTS, a first step in this direction is a detailed analysis of variability in the genomic region flanking the candidate loci.

A question of great interest would be to extract the rate of adaptation of *D. melanogaster* to non-African habitats from our data. This goal is difficult to address even when the effects of demography are ignored, as the power of our approach is dependent on the impact of hitchhiking. This impact can be different for the *X* chromosome and autosomes (ORR and BETANCOURT 2001). Thus, apart from demographic effects, it is possible that the higher number of candidate loci on the *X* chromosome may be due to a higher power to detect hitchhiking on the *X* chromosome, whereas it is likely that both chromosomes carry an equal number of beneficial mutations. Furthermore, additional, elusive parameters, such as the size of the region affected by selection, are required to calculate the fraction of loci affected by selective sweeps. Hence, to extrapolate the rate of adaptation from our data could be misleading. Finally, demography could also inflate the number of selected loci, and different chromosomes may be differently affected by this.

We thank G. Muir, B. Payseur, C. Vogl, and members of the C.S. lab for helpful discussions on the manuscript. D. Charlesworth and three anonymous reviewers provided several helpful suggestions, which significantly improved our manuscript. T. Wiehe shared unpublished software. Many thanks also go to B. Görnet and J. Tordorova for help with microsatellite typing. This work was supported by Fond fur Forderung der Wissenschaftlichen Forschung grants to C.S.

LITERATURE CITED

ANDOLFATTO, P., 2001a Adaptive hitchhiking effects on genome variability. Curr. Opin. Genet. Dev. **11:** 635–641.

ANDOLFATTO, P., 2001b Contrasting patterns of *X*-linked and autosomal nucleotide variation in *Drosophila melanogaster* and *Drosophila simulans*. Mol. Biol. Evol. **18:** 279–290.

AQUADRO, C. F., D. J. BEGUN and E. C. KINDAHL, 1994 Selection, recombination, and DNA polymorphism in *Drosophila*, pp. 46–56 in *Non-Neutral Evolution*, edited by B. GOLDING. Chapman & Hall, London.

BACHTROG, D., S. WEISS, B. ZANGERL, G. BREM and C. SCHLÖTTERER, 1999 Distribution of dinucleotide microsatellites in the *Drosophila melanogaster* genome. Mol. Biol. Evol. **16:** 602–610.

BEGUN, D., and C. F. AQUADRO, 1993 African and North American populations of *Drosophila melanogaster* are very different at the DNA level. Nature **365:** 548–550.

BOULETREAU, J., 1978 Ovarian activity and reproductive potential in a natural population of *Drosophila melanogaster*. Oecologia **35:** 319–342.

BRAVERMAN, J. M., R. R. HUDSON, N. L. KAPLAN, C. H. LANGLEY and W. STEPHAN, 1995 The hitchhiking effect on the sites frequency spectrum of DNA polymorphisms. Genetics **140:** 783–796.

BUSTAMANTE, C. D., R. NIELSEN, S. A. SAWYER, K. M. OLSEN, M. D. PURUGGANAN *et al.*, 2002 The cost of inbreeding in Arabidopsis. Nature **416:** 531–534.

CABALLERO, A., 1994 Developments in the prediction of effective population size. Heredity **73:** 657–679.

CARACRISTI, G., and C. SCHLÖTTERER, 2003 Genetic differentiation between American and European *D. melanogaster* populations could be attributed to admixture of African alleles. Mol. Biol. Evol. **20:** 792–799.

CHARLESWORTH, B., 1998 Measures of divergence between populations and the effect of forces that reduce variability. Mol. Biol. Evol. **15:** 538–543.

CHARLESWORTH, B., 2001 The effect of life-history and mode of inheritance on neutral genetic variability. Genet. Res. **77:** 153–166.

COLSON, I., and D. B. GOLDSTEIN, 1999 Evidence for complex mutations at microsatellite loci in Drosophila. Genetics **152:** 617–627.

COMERON, J. M., M. KREITMAN and M. AGUADE, 1999 Natural selection on synonymous sites is correlated with gene length and recombination in Drosophila. Genetics **151:** 239–249.

CORNUET, J. M., and G. LUIKART, 1996 Description and power analysis of two tests for detecting recent population bottlenecks from allele frequency data. Genetics **144:** 2001–2014.

DABORN, P., S. BOUNDY, J. YEN, B. PITTENDRIGH and R. FFRENCH-CONSTANT, 2001 DDT resistance in Drosophila correlates with Cyp6g1 over-expression and confers cross-resistance to the neonicotinoid imidacloprid. Mol. Genet. Genomics **266:** 556–563.

DAVID, J. R., and P. CAPY, 1988 Genetic variation of *Drosophila melanogaster* natural populations. Trends Genet. **4:** 106–111.

DIERINGER, D., and C. SCHLÖTTERER, 2003 Microsatellite analyzer (MSA)—a platform independent analysis tool for large microsatellite data sets. Mol. Ecol. Notes **3:** 167–169.

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.

FAY, J. C., and C.-I WU, 1999 A human population bottleneck can account for the discordance between patterns of mitochondrial versus nuclear DNA variation. Mol. Biol. Evol. **16:** 1003–1005.

FAY, J. C., and C.-I WU, 2001 The neutral theory in the genomic era. Curr. Opin. Genet. Dev. **11:** 642–646.

FAY, J. C., G. J. WYCKOFF and C.-I WU, 2001 Positive and negative selection on the human genome. Genetics **158:** 1227–1234.

FAY, J. C., G. J. WYCKOFF and C.-I WU, 2002 Testing the neutral theory of molecular evolution with genomic data from Drosophila. Nature **415:** 1024–1026.

GALTIER, N., F. DEPAULIS and N. H. BARTON, 2000 Detecting bottlenecks and selective sweeps from DNA sequence polymorphism. Genetics **155:** 981–987.

GOLDSTEIN, D. B., and A. G. CLARK, 1995 Microsatellite variation in North American populations of *Drosophila melanogaster*. Nucleic Acids Res. **23:** 3882–3886.

HARR, B., and C. SCHLÖTTERER, 2000 Long microsatellite alleles in *Drosophila melanogaster* have a downward mutation bias and short persistence times, which cause their genome-wide underrepresentation. Genetics **155:** 1213–1220.

HARR, B., M. KAUER and C. SCHLÖTTERER, 2002 Hitchhiking mapping—a population-based fine mapping strategy for adaptive mutations in Drosophila melanogaster. Proc. Natl. Acad. Sci. USA **99:** 12949–12954.

HUDSON, R. R., 1990 Gene geneologies and the coalescent process. Oxf. Surv. Evol. Biol. **7:** 1–44.

HUDSON, R. R., 2002 Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics **18:** 337–338.

HUDSON, R. R., M. KREITMAN and M. AGUADÉ, 1987 A test of neutral molecular evolution based on nucleotide data. Genetics **116:** 153–159.

HUDSON, R. R., K. BAILEY, D. SKARECKY, J. KWIATOWSKI and F. J. AYALA, 1994 Evidence for positive selection in the superoxide dismutase (*Sod*) region of *Drosophila melanogaster*. Genetics **136:** 1329–1340.

HUTTLEY, G. A., M. W. SMITH, M. CARRINGTON and S. J. O'BRIEN, 1999 A scan for linkage disequilibrium across the human genome. Genetics **152:** 1711–1722.

HUTTLEY, G. A., S. EASTEAL, M. C. SOUTHEY, A. TESORIERO, G. G. GILES *et al.*, 2000 Adaptive evolution of the tumour suppressor BRCA1 in humans and chimpanzees. Australian breast cancer family study. Nat. Genet. **25:** 410–413.

KAPLAN, N. L., R. R. HUDSON and C. H. LANGLEY, 1989 The "hitchhiking effect" revisited. Genetics **123:** 887–899.

KAUER, M., B. ZANGERL, D. DIERINGER and C. SCHLÖTTERER, 2002 Chromosomal patterns of microsatellite variability contrast sharply in African and non-African populations of *Drosophila melanogaster*. Genetics **160:** 247–256.

KIM, Y., and W. STEPHAN, 2002 Detecting a local signature of genetic hitchhiking along a recombining chromosome. Genetics **160:** 765–777.

KIMURA, M., 1983 *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge, UK.

KIRBY, D. A., and W. STEPHAN, 1996 Multi-locus selection and the structure of variation at the *white* gene of *Drosophila melanogaster*. Genetics **144:** 635–645.

KOHN, M. H., H. J. PELZ and R. K. WAYNE, 2000 Natural selection mapping of the warfarin-resistance gene. Proc. Natl. Acad. Sci. USA **97:** 7911–7915.

LEWONTIN, R. C., and J. KRAKAUER, 1973 Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. Genetics **74:** 175–195.

MAYNARD SMITH, J., and J. HAIGH, 1974 The hitch-hiking effect of a favorable gene. Genet. Res. **23:** 23–35.

MORAN, P. A. P., 1975 Wandering distributions and electrophoretic profile. Theor. Popul. Biol. **8:** 318–330.

NEI, M., 1978 Estimation of average heterozygosity and genetic distance from a small number of individuals. Genetics **89:** 583–590.

NURMINSKY, D., D. D. AGUIAR, C. D. BUSTAMANTE and D. L. HARTL, 2001 Chromosomal effects of rapid gene evolution in Drosophila melanogaster. Science **291:** 128–130.

OHTA, T., and M. KIMURA, 1973 A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. Genet. Res. **22:** 201–204.

ORR, H., and A. BETANCOURT, 2001 Haldane's sieve and adaptation from the standing genetic variation. Genetics **157:** 875–884.

PAYSEUR, B. A., A. D. CUTTER and M. W. NACHMAN, 2002 Searching for evidence of positive selection in the human genome using patterns of microsatellite variability. Mol. Biol. Evol. **19:** 1143–1153.

SCHLÖTTERER, C., 2000 Evolutionary dynamics of microsatellite DNA. Chromosoma **109:** 365–371.

SCHLÖTTERER, C., 2002 A microsatellite-based multilocus screen for the identification of local selective sweeps. Genetics **160:** 753–763.

SCHLÖTTERER, C., and B. HARR, 2002 Single nucleotide polymorphisms derived from ancestral populations show no evidence for biased diversity estimates in Drosophila melanogaster. Mol. Ecol. **11:** 947–950.

SCHLÖTTERER, C., and T. WIEHE, 1999 Microsatellites, a neutral marker to infer selective sweeps, pp. 238–248 in *Microsatellites—Evolution and Applications*, edited by D. GOLDSTEIN and C. SCHLÖTTERER. Oxford University Press, Oxford.

SCHLÖTTERER, C., and B. ZANGERL, 1999 The use of imperfect microsatellites for DNA fingerprinting and population genetics, pp. 153–165 in *DNA Profiling and DNA Fingerprinting*, edited by J. T. EPPLEN and T. LUBJUHN. Birkhäuser, Basel, Switzerland.

SCHUG, M. D., T. F. C. MACKAY and C. F. AQUADRO, 1997 Low mutation rates of microsatellite loci in *Drosophila melanogaster*. Nat. Genet. **15:** 99–102.

SCHUG, M. D., C. M. HUTTER, K. A. WETTERSTRAND, M. S. GAUDETTE, T. F. MACKAY *et al.*, 1998a The mutation rates of di-, tri- and tetranucleotide repeats in *Drosophila melanogaster*. Mol. Biol. Evol. **15:** 1751–1760.

SCHUG, M. D., K. A. WETTERSTRAND, M. S. GAUDETTE, R. H. LIM, C. M. HUTTER *et al.*, 1998b The distribution and frequency of microsatellite loci in *Drosophila melanogaster*. Mol. Ecol. **7:** 57–69.

SMITH, N. G., and A. EYRE-WALKER, 2002 Adaptive protein evolution in Drosophila. Nature **415:** 1022–1024.

SOKAL, R. R., and F. J. ROHLF, 1995 *Biometry*. W. H. Freeman, New York.

TAJIMA, F., 1989 Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics **123:** 585–595.

VAZQUEZ, J., T. PEREZ, J. ALBORNOZ and A. DOMINGUEZ, 2000 Estimation of microsatellite mutation rates in *Drosophila melanogaster*. Genet. Res. **76:** 323–326.

VIGOUROUX, Y., M. MCMULLEN, C. T. HITTINGER, K. HOUCHINS, L. SCHULZ *et al.*, 2002 Identifying genes of agronomic importance in maize by screening microsatellites for evidence of selection during domestication. Proc. Natl. Acad. Sci. USA **99:** 9650–9655.

WALL, J. D., P. ANDOLFATTO and M. PRZEWORSKI, 2002 Testing models of selection and demography in *Drosophila simulans*. Genetics **162:** 203–216.

WEIR, B. S., and C. C. COCKERHAM, 1984 Estimating *F*-statistics for the analysis of population structure. Evolution **38:** 1358–1370.

WIEHE, T., 1998 The effect of selective sweeps on the variance of the allele distribution of a linked multi-allele locus-hitchhiking of microsatellites. Theor. Popul. Biol. **53:** 272–283.

WOOTTON, J. C., X. FENG, M. T. FERDIG, R. A. COOPER, J. MU *et al.*, 2002 Genetic diversity and chloroquine selective sweeps in Plasmodium falciparum. Nature **418:** 320–323.

Communicating editor: D. CHARLESWORTH

Population structure in African D. melanogaster revealed by microsatellite analysis

Daniel Dieringer, Viola Nolte and Christian Schlötterer

Institut für Tierzucht und Genetik, Josef-Baumann Gasse 1, 1210 Wien

Key words: admixture, Africa, Europe

Corresponding author:
Christian Schlötterer
Institut für Tierzucht und Genetik
Josef Baumann Gasse 1
1210 Wien
Austria
Tel.: +43-1-25077-5603
Fax: +43-1-25077-5693
christian.schloetterer@vu-wien.ac.at

ABSTRACT

Tropical Sub-Saharan regions are considered to be the geographic origin of *Drosophila melanogaster*. Starting from there the species colonized the rest of the world after the last glaciation about 10,000 years ago. Consistent with this demographic scenario, African populations have been shown to harbor higher levels of microsatellite and sequence variation than cosmopolitan populations. Nevertheless, very limited information is available on the genetic structure of African populations. We used X-chromosomal microsatellite variation to study the population structure of *D. melanogaster* populations using 14 sampling sites in North-, West- and East- Africa. These populations were compared to six European and one North-American population. Significant population structure was found among African *D. melanogaster* populations. Using a Bayesian method for inferring population structure we detected two distinct groups of populations among the African *D. melanogaster*. Interestingly, the comparison to cosmopolitan *D. melanogaster* populations indicated that one of the divergent African groups is closely related to cosmopolitan flies. Low, but significant levels of differentiation were observed for sub-Saharan *D. melanogaster* populations from West- and East-Africa. Furthermore, we note a temporal heterogeneity in African *D. melanogaster*. Populations have a tendency to group according to collection date, rather than geographic origin, indicating that both collection time and location need to be considered to make meaningful comparisons.

INTRODUCTION

It is generally accepted that *D. melanogaster* originated in Africa, with West-Africa being the most likely center of origin (DAVID and CAPY 1988; LACHAISE *et al.* 1988). African *D. melanogaster* are highly variable for chromosomal inversions (LEMEUNIER and AULARD 1992), sequence polymorphism (ANDOLFATTO 2001; BEGUN and AQUADRO 1993; GLINKA *et al.* 2003; HARR *et al.* 2002; LANGLEY *et al.* 2000), and microsatellite polymorphism (CARACRISTI and SCHLÖTTERER 2003; KAUER *et al.* 2002; KAUER *et al.* 2003b). Starting from Africa *D. melanogaster* recently expanded its habitat into more temperate regions. Given that such a change in habitat presumably required numerous adaptations to the novel environment, a comparison of African and non-African *D. melanogaster* could be used for the identification of genomic regions carrying mutations facilitating the habitat expansion (hitchhiking mapping (SCHLÖTTERER 2003)). Previous hitchhiking mapping studies aiming for the identification of mutations associated with the out of Africa habitat expansion were primarily comparing levels of microsatellite variability in the two groups (KAUER *et al.* 2002; KAUER *et al.* 2003b). Subsequent sequence analysis of genomic regions putatively carrying such a beneficial mutation indicated either no fixed difference or multiple mutations fixed between non-African and African populations (HARR *et al.* 2002).

Despite that fixed differences between African and non-African flies are good candidates for ecologically relevant mutations, their identification is complicated by i) population substructure in Africa and ii) recent admixture with cosmopolitan alleles in African populations (KAUER *et al.* 2003a). Pronounced population structure in African *D. melanogaster* would require the analysis of many African populations to determine the African allele distribution and the identification of fixed differences between African and non-African flies. However, previous results based on a small number of African populations suggested very limited population substructure in African *D. melanogaster* (AGUADÉ 1998; AGUADÉ 1999; BÉNASSI and VEUILLE 1995; CARACRISTI and SCHLÖTTERER 2003; KAUER *et al.* 2003a; SCHLÖTTERER *et al.* 1997). One further complication for the identification of fixed differences between African and non-African populations is the back-migration of non-African alleles into Africa. Recent studies provided strong evidence for the presence of cosmopolitan alleles in Africa, indicating that African populations are not genetically isolated from European flies (BÉNASSI and VEUILLE 1995). Kauer et al. (2003) demonstrated that an urban population collected in Harare showed a significant admixture of non-African alleles, particularly on the autosomes. An even more extreme case was described in Central-Africa, where an urban *D. melanogaster* population with non-African genotypes had been detected (CAPY *et al.* 2000).

Given the recent popularity of out of Africa hitchhiking mapping and its dependence on a reliable characterization of the presumably ancestral African variation, we characterized microsatellite variation in 14 African population samples and compared it to non-African *D. melanogaster*.

MATERIAL AND METHODS

Microsatellite loci: In a first data set we surveyed 17 X chromosomal microsatellite loci in 14 populations from Africa, six populations from Europe and one from North America. To confirm the results obtained from this moderate number of loci we typed most of the populations for additional 82 X chromosomal loci. Primer sequences, annealing temperatures, repeat motifs and cytological positions are provided in the supplementary material (Table A1). Except for the loci X651868ca, 56g7-AG, AE002566_gtc all loci used in this study are identical to those used in Kauer et al. (2003).

Fly strains: A complete list of the strains used, their origin and collection date (if known) is provided in Table A2.

Measures of genetic differentiation: As most of the African strains used were isofemale lines, which have been inbred for many generations, we had to account for the loss of allelic variation within isofemale lines. To obtain an unbiased estimator of variability we randomly selected one allele at each locus, calculated the corresponding statistic, and report the average over 200 replicates. This procedure is implemented in the MICROSATELLITE-ANALYZER (MSA 3.12) software (DIERINGER and SCHLÖTTERER 2003). Variance in repeat number, expected heterozygosity (gene diversity) (NEI 1987) and the proportion of shared alleles were calculated using MSA 3.12. Gene diversity was corrected for sample size by n/(n-1) where n is the number of analyzed chromosomes. Trees were calculated from genetic distance matrices using the NEIGHBOR JOINING algorithm (SAITOU and NEI 1987) provided with the PHYLIP software package (FELSENSTEIN 1991) and were graphically displayed with TREEVIEW software (PAGE 1996). The unbiased estimator of F_{ST} (WEIR and COCKERHAM 1984) between populations was calculated with MSA 3.12. Significance levels were determined by permuting genotypes 100,000among all population pairs. This conservative procedure does not assume Hardy-Weinberg equilibrium and allows for linkage among loci. We accounted for multiple testing by using the Bonferroni correction (SOKAL and ROHLF 1995).

Bayesian analysis of population structure: We used the program BAPS (CORANDER *et al.* 2003) to test for population differentiation. This program estimates the hidden population substructure using a Bayesian approach to test whether the allele frequencies between populations are significantly different. We used 10,000 updates with a burn in phase of 5,000 updates. In order to assure convergence we performed two runs and obtained very similar

results (data not shown). The MCMC was initialized assuming that each population represents a different cluster.

Bayesian analysis of past demographic events: We analyzed the populations Lake Kariba, Sengwa, Harare, Mali, Lamto, Rome, Crete and Copenhagen with the program msvar (BEAUMONT 1999). Using the stepwise mutation model msvar calculates the Bayesian posterior distribution of demographic and mutational parameters, using a Markov chain Monte Carlo (MCMC) method. The two demographic parameters r and t_f are of further interest for this study. The model used here expects that a population of the actual size N_0 has changed its size for t_A generations starting from a size N_1 in the past. t_f is defined as $t_f = t_A / N_0$.
The parameter r is defined as $r = N_0/N_1$, indicating population expansion if $r > 1$ and population decline if $r < 1$.
As the time since the demographic event is scaled by the effective population size, it is difficult to compare the results obtained for African and non-African populations. Given that non-African flies were derived from African ones, we assumed that both populations should have shared the same ancestral population size N_1. Scaling the time with this ancestral population size N_1 (instead of N_0 $t_C = t_A/N_1$) we obtained comparable estimates of the time passed since the demographic event for African and European populations.

In addition to the data set consisting of 17 loci, we selected for a subset of the populations 17 additional loci to test for locus specific effects in the inference of the demographic parameters. This later set of loci contained only those for which inspection by eye did not indicate large allele gaps (BEAUMONT 1999).
For all calculations we used the exponential growth model which is more suitable for modeling changes in population size on a shorter timescale (BEAUMONT 1999). $2x10^8$ updates were calculated for each population, and only the last 90% of the chains were used. Each population/locus set combination was run at least twice to test the general stability of the solution from the Markov chain.

RESULTS

Variability and population differentiation: We used a Bayesian approach to determine the number of genetically distinct groups in our sample of 14 African *D. melanogaster* populations. In total, only two distinct genetic groups could be detected using this approach (P>0.999). One group consisted of flies from Kenya, Zimbabwe, Uganda and Mali. The other group encompassed North African flies from Morocco and Tunisia. Interestingly, the two populations from the Ivory coast were not assigned to the same group. While the population from Abidjan

grouped with the other sub-Saharan populations, the population from Lamto grouped with the North-African populations. An analysis of genetic distance measured either by the proportion of shared alleles or F_{ST} provided similar results as the Bayesian analysis using BAPS. In addition to the two groups recognized by the BAPS analysis, pairwise F_{ST} values were also significant between Mali and ZH and Zlk. No significant difference was detected between the East-African populations from Kenya and Zimbabwe (Table 2). However, when Hardy-Weinberg equilibrium and no linkage were assumed more pairwise comparisons were found to be significant (see Table A3, online supplement).

We significantly expanded the number of analyzed loci to scrutinize the inferred population structure in African *D. melanogaster* populations. Interestingly, despite the expected gain in power, the model-based clustering method still provided the highest support for two separate population groups in Africa (P>0.999). For pairwise F_{ST} values, however, a higher proportion of pairwise comparisons resulted in a significant difference (Table 3 and Table A4, online supplement). When Hardy-Weinberg equilibrium and no linkage were assumed, all pairwise comparisons except Lamto/Marrakech and Sengwa/Harare indicated a significant differentiation . Consistent with the grouping obtained from BAPS, the pairwise F_{ST} value between Lamto and the North-African populations was substantially smaller than between Lamto and the other sub-Saharan populations.

We related the two African groups of populations to the cosmopolitan populations by including North-American and European flies in the analysis. After including these populations BAPS analysis still provided the highest posterior probability for two different groups. Interestingly, North-African populations and Lamto were clustered together with non-African populations, but the other group remained largely unaffected. This result holds irrespective whether we performed the analysis with the complete set of populations with 99 loci, P>0.999), or the reduced set of populations with 17 loci (P>0.999). Also pairwise F_{ST} analysis supported the close relationship between Lamto and the non-African flies (Table 3 and Table 4).

To provide further insight into the relationship of the African populations, we constructed a phylogenetic tree using the proportion of shared alleles as measurement of genetic divergence. Figure 1 and 2 show the trees for the two data sets. The close relationship of North-African and the Lamto population with the non-African populations is apparent in the trees from both data sets. One other interesting observation is that no grouping according to geographic origin of the sub-Saharan populations was detected. Our analysis included four populations from Zimbabwe, but only the Sengwa and Harare population group together. As these two populations have been sampled at the same time, this suggests that temporal heterogeneity in the sample may be more important than geographic origin of the flies. Even more evident is the importance of the temporal sampling by the grouping of populations from Zimbabwe and Kenya both collected in the same

year. Four sub-Saharan populations were sampled in 2001. These populations neither show a grouping according to geography nor to sampling year. Rather, we observe the grouping of two pairs (Zimbabwe (ZW) /Kenya (KYO) and Mali/Kisoro (Kis)). While the latter population pair was collected in the first quarter of 2001, the Zimbabwe/Kenya pair was collected in the third quarter. This observation suggests that not only the year, but also the season may have an important effect.

Inferring demographic history : We used a recent MCMC method to infer the demographic history of the African *D. melanogaster* populations. These calculations are extremely CPU intensive, therefore, we limited our analysis to a subset of populations using only a reduced set of microsatellite loci. Given the obvious dichotomy of African flies of one group closely resembling cosmopolitan populations and others presumably closer to the ancestral African populations, we analyzed the demographic history of populations representing both groups. All Zimbabwe populations as well as the population from Mali could be considered as representatives of the ancestral African populations. Thus, we expect that these populations should be the best approximation of an equilibrium population. Interestingly, for both sets of microsatellite loci we detected an unambiguous indication of a population size decline in all populations assumed to be in equilibrium (Table 4 and Table 5). The reduction in variability (r) was more pronounced in non-African populations. However, the estimated degree of reduction (r) seems to be more similar across populations using the same set of loci rather than across different sets of loci (Table 4 and Table 5). The population size change occurred more recently in the East-African populations, than in Lamto or the European populations.

DISCUSSION

Ancestral *D. melanogaster* populations: The presumed origin of *D. melanogaster* is sub-Saharan West-Africa (DAVID and CAPY 1988; LACHAISE *et al.* 1988). Nevertheless, a large number of molecular studies used populations from sub-Saharan East-Africa (Zimbabwe and Kenya) as representatives for the presumed ancestral African variation. While DNA sequencing studies failed to detect population differentiation between West- and East-African populations (AGUADÉ 1998; AGUADÉ 1999), a microsatellite study found significant differentiation (MICHALAKIS and EXCOFFIER 1996). Large differences in the frequency of the *ln(2L)t* inversion were also described between East- and West-African *D. melanogaster* (VEUILLE *et al.* 1998). Finally, a west-east differentiation has been described for the African polymorphic endemic inversions (AULARD *et al.* 2002). In this study, we included three different populations from West-Africa, two from the

[illegible]

[illegible]

[illegible]

[illegible]

Ivory Coast and one from Mali to address the genetic differentiation of West- and East-African *D. melanogaster*. A comparison of the three populations indicated that Mali and Abidjan were genetically very similar, but Lamto differed significantly from the two other populations. As we cannot exclude a contamination of our sample of the Lamto population (see below for further discussion), we limit our discussion to the other two West-African populations. Levels of differentiation between West- and East-African populations are slightly higher than among East-African populations (0.036 vs. 0.025 mean F_{ST} over all comparisons). Gene diversities and variance in repeat number were also very similar for West- and East-African flies (Table 1). Due to the different sample sizes for West- and East-African populations, we cannot compare the number of alleles or private alleles. Thus, our data provide no further insight on whether West- or East-African *D. melanogaster* harbor more ancestral variation and should thus be considered as the more ancestral ones.

Demography of African populations: The high similarity of sub-Saharan *D. melanogaster* populations suggests considerable levels of geneflow or a shared recent ancestry of the populations. Based on our available samples it is not possible to distinguish between the two hypotheses. Nevertheless, the fact that East-African *D. melanogaster* cluster by collection date rather than by geographic origin (Fig. 1) suggests that temporal heterogeneity may have to be considered to understand the genetic architecture of African *D. melanogaster* populations. Similarly, temporal heterogeneity has been observed in non-African *D. melanogaster* populations (LAZZARO and CLARK 2003; SCHLÖTTERER and AGIS 2002). Nevertheless, other studies failed to detect a temporal heterogeneity for French *D. melanogaster* populations (VEUILLE *et al.* 2004). Using a MCMC approach to estimate demographic history, we obtained an unambiguous signal of population size contraction for all analyzed sub-Saharan *D. melanogaster*. Given that the method makes strong assumptions (microsatellite mutation model, no population structure) it is not clear to what extent the population contraction signal is an artifact caused by deviations from the model assumptions. Using two different sets of microsatellites, we obtained pronounced differences in the estimates for r and t, which indicates a strong influence of the loci used for the analysis.

Hitchhiking mapping: Recently, the comparison of African and non-African *D. melanogaster* populations has been shown to be a very powerful tool for the identification of genomic regions subjected to recent selective sweeps putatively associated with the habitat expansion of *D. melanogaster* (GLINKA *et al.* 2003; HARR *et al.* 2002; KAUER *et al.* 2002; KAUER *et al.* 2003b). Microsatellite-based genome scans could be used to test for a significant reduction in non-African populations relative to African flies (SCHLÖTTERER 2002; SCHLÖTTERER and DIERINGER 2003). As this lnRθ test statistic only requires a population that has not been subjected to same sweep, it is not expected to be highly sensitive to the choice of the reference population. Nevertheless, the very similar levels of variability among the sub-Saharan populations further indicate that each of

the analyzed populations could serve as a reference in the lnRθ test statistic. On the other hand, the low, but significant differences among the sub-Saharan populations indicate that fixed differences between African and non-African populations (GLINKA *et al.* 2003; HARR *et al.* 2002) can only be identified by the comparison of multiple sub-Saharan and non-African populations.

Lamto: Our analysis included one population from Lamto. The Lamto sample differs in various aspects from the two other West-African populations-it harbors less variability and is highly differentiated from the other sub-Saharan populations included in our study. This result sharply contrasts with other studies using population samples from Lamto (AGUADÉ 1998; AGUADÉ 1999; MICHALAKIS and VEUILLE 1996; VEUILLE *et al.* 2004). While previous microsatellite studies indicate that the level of variability in the Lamto population is intermediate, falling between non-African and East-African populations (VEUILLE *et al.* 2004), on the sequence level, high levels of variability are observed. A previous study on two populations collected in Brazzaville indicated striking differences between the two samples. While one population closely resembled non-African flies, the other population harbored more African characteristics (CAPY *et al.* 2000). Unfortunately, the exact collection site and date for the Lamto sample analyzed in this study are not available. Thus, we cannot decide whether the Lamto sample in our study is also derived from the same collection as the Lamto sample used in other studies. The samples used in this study were propagated as isofemale lines until 1994 when the lines were received, but the other studies used frozen flies (some were isogenized before freezing). Thus, it may be possible that the lines were contaminated during propagation.

North-African *D. melanogaster*: The Bayesian analysis of population differentiation grouped North-African flies together with non-African flies. Also pairwise F_{ST} values suggested a close relationship of North-African flies to non-African ones. A slight trend for higher F_{ST} values for those comparisons involving the population from Agadir has to be noted, but the small sample size of this population prevents further conclusions. Also levels of variability were very similar for North-African and non-African flies. Two different hypotheses could account for the striking similarity between North-African and non-African flies. Either the major bottleneck after the habitat expansion from sub-Saharan Africa occurred already with the colonization of Northern Africa, or substantial migration between North-African and non-African *D. melanogaster* populations has resulted in the observed low differentiation.

ACKNOWLEDGMENTS

We are grateful to C. Aquadro, B. Ballard, P. Capy, J. David, A. Djiteye, V. Loeschke, and C.-I Wu for sharing flies. The work has been supported by Fonds zur Förderung der wissenschaftlichen Forschung grants to C.S..

LITERATURE CITED

AGUADÉ, M., 1998 Different forces drive the evolution of the *Acp26Aa* and *Acp26Ab* accessory gland genes in the *Drosophila melanogaster* species complex. Genetics **150:** 1079-1089.

AGUADÉ, M., 1999 Positive selection drives the evolution of the *Acp29AB* accessory gland protein in Drosophila. Genetics **152:** 543-551.

ANDOLFATTO, P., 2001 Contrasting Patterns of X-Linked and Autosomal Nucleotide Variation in *Drosophila melanogaster* and *Drosophila simulans*. Molecular Biology and Evolution **18:** 279-290.

AULARD, S., J. R. DAVID and F. LEMEUNIER, 2002 Chromosomal inversion polymorphism in Afrotropical populations of *Drosophila melanogaster*. Genet Res **79:** 49-63.

BEAUMONT, M. A., 1999 Detecting population expansion and decline using microsatellites. Genetics **153:** 2013-2029.

BEGUN, D., and C. F. AQUADRO, 1993 African and North American populations of *Drosophila melanogaster* are very different at the DNA level. Nature **365:** 548-550.

BÉNASSI, V., and M. VEUILLE, 1995 Comperative population structuring of molecular and allozyme variation of *Drosophila melanogaster Adh* between Europe, West Africa and East Africa. Genet. Res. **65:** 95-103.

CAPY, P., M. VEUILLE, M. PAILLETTE, J. M. JALLON, J. VOUIDIBIO *et al.*, 2000 Sexual isolation of genetically differentiated sympatric populations of *Drosophila melanogaster* in Brazzaville, Congo: the first step towards speciation? Heredity **84:** 468-475.

CARACRISTI, G., and C. SCHLÖTTERER, 2003 Genetic differentiation between American and European *Drosophila melanogaster* populations could be attributed to admixture of African alleles. Mol Biol Evol **20:** 792-799.

CORANDER, J., P. WALDMANN and M. J. SILLANPAA, 2003 Bayesian analysis of genetic differentiation between populations. Genetics **163:** 367-374.

DAVID, J. R., and P. CAPY, 1988 Genetic variation of *Drosophila melanogaster* natural populations. Trends in Genetics **4:** 106-111.

DIERINGER, D., and C. SCHLÖTTERER, 2003 Microsatellite analyzer (MSA) - a platform independent analysis tool for large microsatellite data sets. Molecular Ecology Notes **3:** 167-169.

FELSENSTEIN, J., 1991 PHYLIP, Version 3.57c (University of Washington, Seattle).

GLINKA, S., L. OMETTO, S. MOUSSET, W. STEPHAN and D. DE LORENZO, 2003 Demography and natural selection have shaped genetic variation in *Drosophila melanogaster*: a multi-locus approach. Genetics **165:** 1269-1278.

HARR, B., M. KAUER and C. SCHLÖTTERER, 2002 Hitchhiking mapping - a population based fine mapping strategy for adaptive mutations in *D. melanogaster*. Proc. Natl. Acad. Sci. USA **99:** 12949-12954.

KAUER, M., D. DIERINGER and C. SCHLÖTTERER, 2003a Nonneutral admixture of immigrant genotypes in African *Drosophila melanogaster* populations from Zimbabwe. Mol Biol Evol **20:** 1329-1337.

KAUER, M., B. ZANGERL, D. DIERINGER and C. SCHLÖTTERER, 2002 Chromosomal patterns of microsatellite variability contrast sharply in African and non-African populations of *Drosophila melanogaster*. Genetics **160:** 247-256.

KAUER, M. O., D. DIERINGER and C. SCHLÖTTERER, 2003b A microsatellite variability screen for positive selection associated with the "out of Africa" habitat expansion of *Drosophila melanogaster*. Genetics **165:** 1137-1148.

LACHAISE, D., M.-L. CARIOU, J. R. DAVID, F. LEMEUNIER, L. TSACAS *et al.*, 1988 Historical biogeography of the *Drosophila melanogaster* species subgroup. Evol. Biol. **22:** 159-225.

LANGLEY, C. H., B. P. LAZZARO, W. PHILLIPS, E. HEIKKINEN and J. M. BRAVERMAN, 2000 Linkage disequilibria and the site frequency spectra in the *su(s)* and *su(w^a)* regions of the *Drosophila melanogaster* X chromosome. Genetics **156:** 1837-1852.

LAZZARO, B. P., and A. G. CLARK, 2003 Molecular population genetics of inducible antibacterial peptide genes in *Drosophila melanogaster*. Mol Biol Evol **20:** 914-923.

LEMEUNIER, D., and S. AULARD, 1992 Inversion polymorphism in *Drosophila melanogaster*, pp. 339-405 in *Drosophila inversion polymorphism*, edited by C. B. KRIMBAS and J. R. POWELL. CRC Press, Cleveland.

MICHALAKIS, Y., and L. EXCOFFIER, 1996 A generic estimation of population subdivision using distances between alleles with special reference for microsatellite loci. Genetics **142:** 1061-1064.

MICHALAKIS, Y., and M. VEUILLE, 1996 Length variation of CAG/CAA trinucleotide repeats in natural populations of *Drosophila melanogaster* and its relation to the recombination rate. Genetics **143:** 1713-1725.

NEI, M., 1987 *Molecular evolutionary genetics*. Columbia University Press, New York.

PAGE, R. D. M., 1996 TREEVIEW: An application to display phylogenetic trees on personal computers. Computer Applications in the Biosciences **12:** 357-358.

SAITOU, R. K., and M. NEI, 1987 The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4:** 406-425.

SCHLÖTTERER, C., 2002 A microsatellite-based multilocus screen for the identification of local selective sweeps. Genetics **160:** 753-763.

SCHLÖTTERER, C., 2003 Hitchhiking mapping - functional genomics from the population genetics perspective. Trends Genet **19:** 32-38.

SCHLÖTTERER, C., and M. AGIS, 2002 Microsatellite analysis of *Drosophila melanogaster* populations along a microclimatic contrast at Lower Nahel Oren Canyon, Mount Carmel Israel. Molecular Biology and Evolution **19:** 563-568.

SCHLÖTTERER, C., and D. DIERINGER, 2003 A novel test statistic for the identification of local selective sweeps based on microsatellite gene diversity in *Selective Sweep*, edited by D. I. NURMINSKY. Landes Bioscience, Georgetown.

SCHLÖTTERER, C., C. VOGL and D. TAUTZ, 1997 Polymorphism and locus-specific effects on polymorphism at microsatellite loci in natural *Drosophila melanogaster* populations. Genetics **146:** 309-320.

SOKAL, R. R., and F. J. ROHLF, 1995 *Biometry*. W. H. Freeman and Company, New York.

VEUILLE, M., E. BAUDRY, M. COBB, N. DEROME and E. GRAVOT, 2004 Historicity and the population genetics of *Drosophila melanogaster* and D. *simulans*. Genetics **120:** 61-70.

VEUILLE, M., V. BENASSI, S. AULARD and F. DEPAULIS, 1998 Allele-specific population structure of *Drosophila melanogaster* alcohol dehydrogenase at the molecular level. Genetics **149:** 971-981.

WEIR, B. S., and C. C. COCKERHAM, 1984 Estimating *F*-statistics for the analysis of population structure. Evolution **38:** 1358-1370.

Fig 1 Phylogenetic tree of 21 D. melanogaster populations using 17 polymorphic microsatellite loci (proportion of shared alleles, neighbor joining)

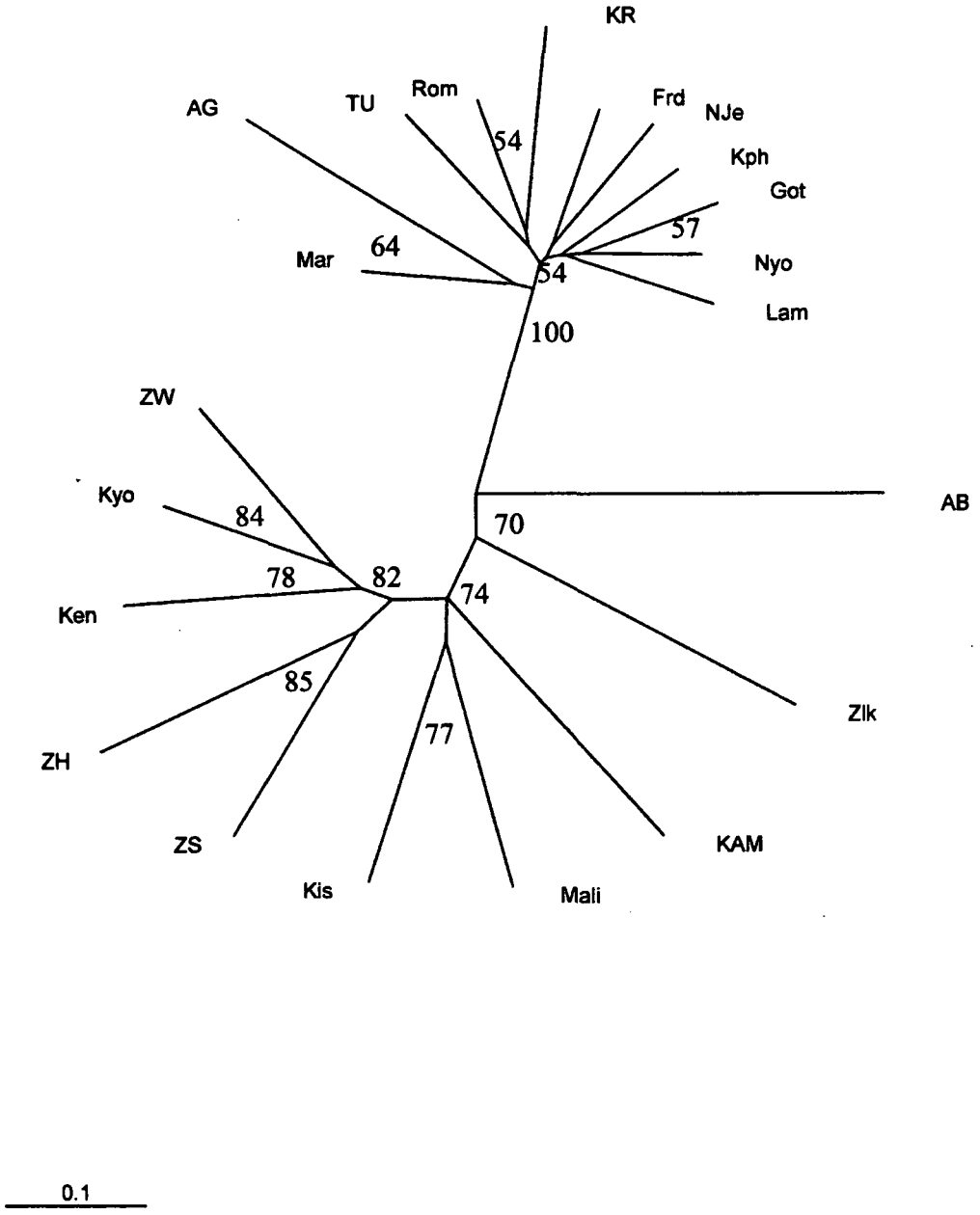

Fig 2 Phylogenetic tree of 13 D. melanogaster populations using 99 polymorphic microsatellite loci (proportion of shared alleles, neighbor joining)

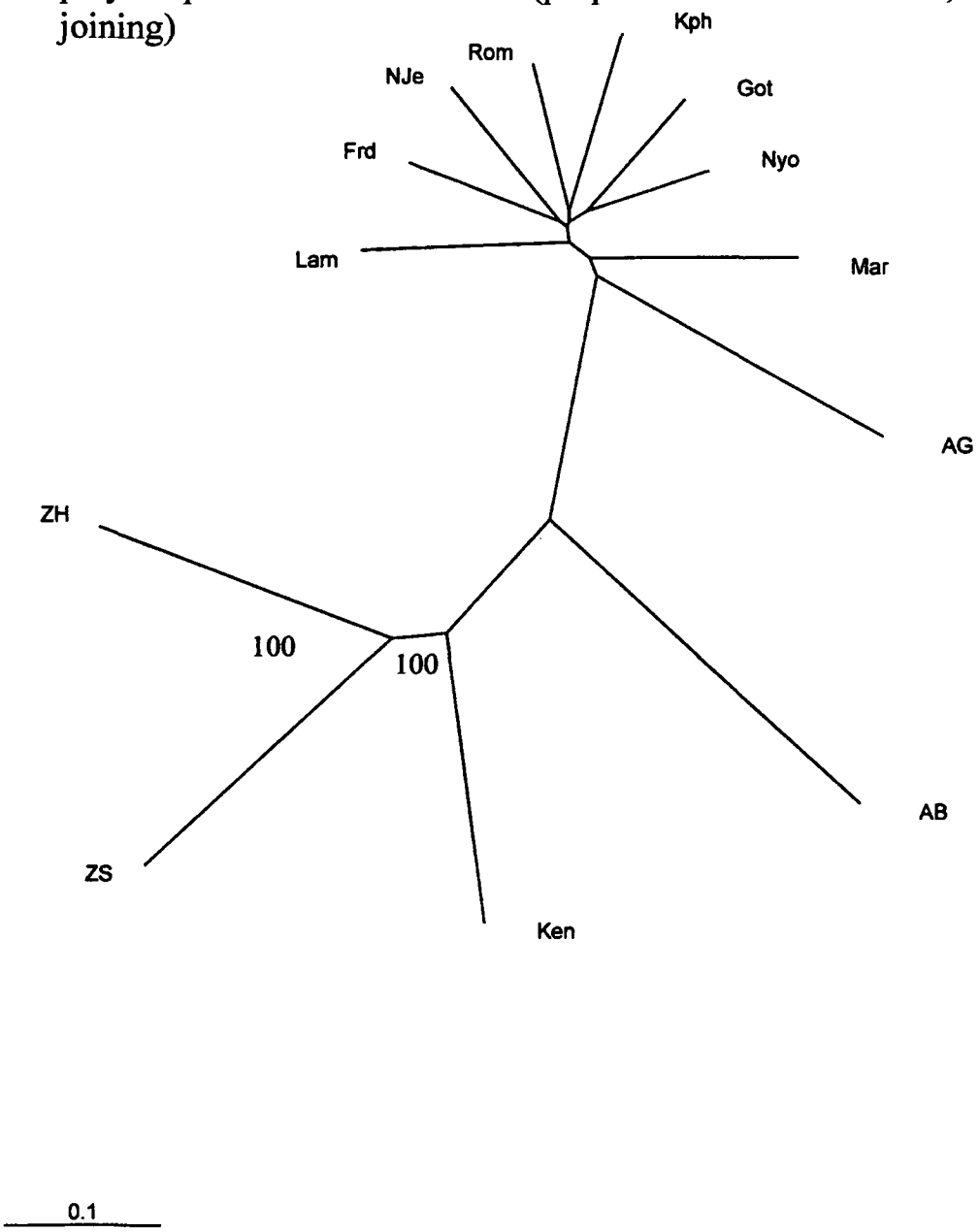

Table 1 Variability overview

	Het exp[1,3]	Het exp[1,4]	Var[2,3]	Var[2,4]
Rom	0.53	0.49	8.37	5.36
Frd	0.52	0.50	8.44	5.72
Got	0.55	0.49	8.54	5.70
Nyo	0.51	0.51	8.61	6.19
Kph	0.47	0.50	8.75	7.49
NJe	0.53	0.54	8.87	6.69
KR	0.48	n.d.	9.43	n.d.
TU	0.57	n.d.	9.02	n.d.
Mar	0.54	0.52	9.05	7.41
AG	0.51	0.55	9.08	6.11
Mali	0.75	n.d.	9.88	n.d.
Lam	0.55	0.51	9.99	6.56
AB	0.81	0.78	10.08	14.88
Ken	0.78	0.80	9.47	14.25
ZS	0.80	0.79	9.52	14.51
ZH	0.78	0.80	9.43	11.30
Zlk	0.74	n.d.	9.60	n.d.
Kyo	0.79	n.d.	9.36	n.d.
Kis	0.79	n.d.	9.65	n.d.
Kam	0.76	n.d.	9.78	n.d.
ZW	0.78	n.d.	7.97	n.d.

[1] Gene diversity
[2] Variance in repeat number
[3] using 17 polymorphic microsatellite loci
[4] using 99 polymorphic microsatellite loci

Table 2: pairwise F_{ST} and P-values for 21 *D. melanogaster* populations using 17 polymorphic microsatellite loci

	Rom	Frd	Got	Nyo	Kph	NJe	KR	TU	Mar	AG	Mali	Lam	AB	Ken	ZS	ZH	Zlk	Kyo	Kis	Kam	ZW
Rom		0.047	0.050	0.039	0.036	0.058	0.030	0.041	0.041	0.131	0.210	0.045	0.226	0.193	0.181	0.217	0.190	0.185	0.186	0.177	0.195
Frd	***		0.048	0.039	0.039	0.034	0.065	0.057	0.037	0.134	0.219	0.026	0.241	0.203	0.204	0.231	0.203	0.196	0.195	0.186	0.211
Got	***	***		0.020	0.046	0.053	0.086	0.054	0.031	0.092	0.196	0.012	0.213	0.194	0.186	0.225	0.197	0.186	0.176	0.183	0.195
Nyo	***	***	n.s.		0.040	0.050	0.063	0.066	0.023	0.113	0.231	0.007	0.232	0.227	0.220	0.253	0.229	0.220	0.205	0.210	0.229
Kph	***	***	***	***		0.062	0.076	0.063	0.053	0.148	0.252	0.024	0.295	0.242	0.231	0.268	0.239	0.232	0.233	0.230	0.243
NJe	***	***	***	***	***		0.055	0.045	0.038	0.136	0.213	0.027	0.228	0.190	0.207	0.232	0.190	0.190	0.187	0.188	0.200
KR	n.s.	***	***	**	***	***		0.031	0.073	0.170	0.236	0.046	0.232	0.178	0.163	0.198	0.198	0.183	0.214	0.194	0.194
TU	***	***	***	***	***	***	n.s.		0.040	0.124	0.181	0.043	0.198	0.167	0.158	0.191	0.172	0.158	0.180	0.161	0.172
Mar	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.		0.002	0.165	-0.021	0.144	0.156	0.147	0.175	0.138	0.155	0.132	0.160	0.171
AG	n.s.	***	n.s.	n.s.	n.s.	***	n.s.	n.s.	n.s.		0.220	0.048	0.162	0.166	0.157	0.205	0.194	0.199	0.168	0.215	0.215
Mali	***	***	***	***	***	***	***	***	***	n.s.		0.180	0.034	0.044	0.045	0.053	0.084	0.024	0.007	0.028	0.035
Lam	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	***		0.144	0.173	0.165	0.194	0.15	0.17	0.136	0.165	0.185
AB	***	***	***	***	***	***	n.s.	**	n.s.	n.s.	n.s.	n.s.		0.041	0.024	0.032	0.044	0.042	-0.019	0.048	0.043
Ken	***	***	***	***	***	***	***	***	***	**	n.s.	***	n.s.		0.01	0.017	0.051	0.001	0.018	0.024	0.013
ZS	***	***	***	***	***	***	***	***	***	n.s.	n.s.	***	n.s.	n.s.		-0.012	0.04	0.012	0.028	0.014	0.017
ZH	***	***	***	***	***	***	***	***	***	n.s.	**	***	n.s.	n.s.	n.s.		0.053	0.014	0.043	0.035	0.022
Zlk	***	***	***	***	***	***	***	***	**	n.s.	**	***	n.s.	n.s.	n.s.	n.s.		0.036	0.034	0.073	0.042
Kyo	***	***	***	***	***	***	***	***	***	***	n.s.	***	n.s.	n.s.	n.s.	n.s.	n.s.		0.009	0.016	-0.001
Kis	***	***	***	***	***	***	***	***	***	***	n.s.	***	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.		0.032	0.02
Kam	***	***	***	***	***	***	***	***	***	n.s.	n.s.	***	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.		0.025
ZW	***	***	***	***	***	***	***	***	***	***	n.s.	***	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	

n.s. is not significant ≥ 0.1, * ≥ 0.05, ** ≥ 0.01, *** <0.01

Table 3: pairwise F_{ST} and P-values for 13 *D. melanogaster* populations using 99 polymorphic microsatellite loci

	Rom	Frd	Got	Nyo	Kph	Nje	Mar	AG	Lam	AB	Ken	ZS	ZH
Rom		0.036	0.036	0.026	0.040	0.043	0.041	0.110	0.037	0.229	0.192	0.248	0.228
Frd	***		0.042	0.037	0.050	0.032	0.041	0.109	0.038	0.236	0.194	0.249	0.229
Got	***	***		0.022	0.055	0.048	0.055	0.116	0.028	0.232	0.199	0.247	0.232
Nyo	***	***	***		0.044	0.037	0.037	0.100	0.028	0.229	0.192	0.238	0.229
Kph	***	***	***	***		0.052	0.060	0.121	0.054	0.250	0.195	0.245	0.233
NJe	***	***	***	***	***		0.026	0.087	0.034	0.217	0.173	0.225	0.211
Mar	***	***	***	***	***	***		0.026	0.011	0.163	0.125	0.177	0.164
AG	***	***	***	***	***	***	n.s.		0.042	0.129	0.120	0.161	0.155
Lam	***	***	**	**	***	***	n.s.	n.s.		0.167	0.161	0.209	0.197
AB	***	***	***	***	***	***	n.s.	n.s.	n.s.		0.018	0.032	0.028
Ken	***	***	***	***	***	***	***	**	***	n.s.		0.020	0.012
ZS	***	***	***	***	***	***	***	***	***	n.s.	**		0.003
ZH	***	***	***	***	***	***	***	**	***	n.s.	n.s.	n.s.	

n.s. is not significant ≥ 0.1, * ≥ 0.05, ** ≥ 0.01, *** <0.01

Table 4: Bayesian estimation of population size parameter and 90% confidential interval using those 17 polymorphic microsatellite loci that were analyzed in all 21 *D. melanogaster* populations

	Rom	Kph	KR	Mali	Lam	Ken
r	-2.71	-2.93	-2.55	-2.57	-3.42	-2.88
	(-3.06 - -2.27)	(-3.52 - -2.50)	(-2.91 - -2.21)	(-3.25 - -1.96)	(-4.49 - -2.85)	(-3.21 - -2.33)
t_f	0.11	0.48	0.32	-0.25	0.46	-0.06
	(0.03 - 0.18)	(0.35 - 0.58)	(0.16 - 0.45)	(-0.38 - -0.11)	(0.33 - 0.61)	(-0.20 - 0.01)
t_c	-2.60	-2.45	-2.23	-2.82	-2.96	-2.94
	(-3.04 - -2.09)	(-3.17 - -1.92)	(-2.75 - -1.77)	(-3.63 - -2.07)	(-4.16 - -2.24)	(-3.41 - -2.32)
	ZS	ZH	Zlk	Kyo	Kis	Kam
r	-3.18	-2.88	-3.18	-3.37	-3.18	-3.26
	(-3.64 - -2.64)	(-3.34 - -2.41)	(-3.64 - -2.64)	(-3.77 - -2.10)	(-3.77 - -2.55)	(-3.48 - -2.47)
t_f	0.06	0.04	0.06	-0.10	-0.10	-0.15
	(-0.06 - 0.16)	(-0.10 - 0.16)	(-0.06 - 0.16)	(-0.29 - -0.01)	(-0.18 - -0.02)	(-0.25 - -0.06)
t_c	-3.12	-2.84	-3.12	-3.46	-3.28	-3.41
	(-3.69 - -2.48)	(-3.44 - -2.25)	(-3.69 - -2.48)	(-4.06 - -2.11)	(-3.95 - -2.56)	(-3.73 - -2.53)

Table 5: Bayesian estimation of population size parameter and 90% confidential interval from 17 randomly selected microsatellites differing from the 17 loci used in Table 4

	Rom	Kph	Lam	ZH	ZS
r	-1.50	-2.05	-2.19	-2.31	-2.06
	(-2.00 - -1.15)	(-2.43 - -1.61)	(-3.20 - -1.72)	(-2.84 - -1.37)	(-2.72 - -1.60)
t_f	0.03	0.43	0.20	-0.41	-0.28
	(-0.31 - 0.29)	(0.21 - 0.64)	(0.01 - 0.37)	(-0.68 - -0.20)	(-0.42 - -0.12)
t_C	-1.47	-1.62	-1.99	-2.72	-2.34
	(-2.31 - -0.86)	(-2.22 - -0.98)	(-3.19 - -1.34)	(-3.52 - -1.57)	(-3.14 - -1.71)

APPENDIX

Table A1: Microsatellite loci used

Locus	Data set 1[1]	Chromosome	Repeat Type[2]	Annealing temp.	Primer forward	Primer reverse
X651868ca		1C	(GT)10	48	TGT CTG AAG TTT TGT ATC GTC	TTA GTA TTA TAC CCT CTT TCG
56g7-AG		1E	(AG)9	49.8	ATT TGA TTA AAG TCC CAG TC	AAA TTC CCC AAG GTG GTC AC
AE002566_gtc		1F	(GTC)10	58	ATG TCG CCC ATT GCC AC	CCG CCA GCA CGA CGA G
X1883543gt		2D	(GT)12	52.4	ACG AAT ACG AAA ATC CGA C	GGG AGA AGT TGA AGT GGA G
X2102441ct		2F	(CT)15	55	AAA GAC CTC CTG TCT GGT AC	GGA ATG AGT TTG TGC CAC
X2307713ca	x	3A	(CA)16	58	CCC ATA CAG ACA GAC GCA CG	ATT GCC ACG CCC ACT TTA TC
P3B02 atc	x	3B	(ATC)8	51	CGA CAG TGA TGC GAG AG	AAA GAT GCC GAT GTA AAT G
X2609012gt		3B	(GT)19	54	TAG TGG ACT CAA AGA CAC ATA C	GTC AAG AGG TGT TGT CTG CC
DS06335a		3C	(GT)15	53	ACT GTA ATT GCT GTT CTA TGT	CGC ACA CTG GGA CAC AAA A
66-95-3		3C	(CA)12	57	GCA CAA TCA CAT CGT ATT CAC TCA GCC AGA	ATT GTT GTT GCT GCG ATT TTC AAA T(
95B7-AT		3C	(AT)9	50.6	ACT GGG AAC GCT GCT TGA TC	TGC CAA CTG TTT TGC TTG TC
X3026663gt		3C	(GT)17	55	AAT GTT TGA CGA CTG CCT CTC	GAT GGT CTA AGG GAG CAT CTG
X2986542ta		3C	(TA)14	45	GCC AAA TAG ATC ACT AAC TCG	CGC TCA AAT AAA CGT ATA TTG
AE002566_gt2		3D	(GT)20	52	TGG TCT TTG CCT CTG TTG	ATC TGT TGT GCT TGT GCT G
X3219363gt		3D	(GT)18	55	CAA ATC ATA ATG CCT AAT TCG	CAG TTA GAG CCG ATA AGG AGC
X3076173ca		3D	(CA)14	55	CAA TGT TCC TGA TGA GCT GTC	TGG GTA TTG GGT ATT GCT CG
X3202250ata		3D	(ATA)13	50	GTG CAA ATA GAA AAT AGC TG	ACA TTA TTT TGA TGG ACT TG
X3439769ca	x	3E	(CA)15	55	GCG AGT GAA GAG GGT ACG CAC	AAC ATG GCA AAT ACA CGG TCG
X3306698ca		3E	(CA)9G(CA)9TA(CA)5	55	GCA AGT ACC TCA CGA ATT TCC	CTC GTG GAA AAC TTT GCC AGC
X3343263ca		3E	(CA)13	58	GCG TAT GAG CAA TGC ACA AAC	GGA CCA ACT GCC CAC CTA TAC
X3525583ta	x	3F	(TA)15	50	TTG CCC GTC TGT GTG CTT TTC	AAT CCA CGT TAA CTT TCA TTG
X3550011ca		3F	(CA)16	52	GTT AAG TAT ATG CCG CTC AAC	ACA AAA GCC ACA CTT ACA TGC
X3655941ga		3F	(GA)12	55	CTG AGA ATC GAA TCG GAG TGC	GGG AGT TTG TCT TGT AGT AGG
X3829513gt		4A	(GT)19	58	TGC GAG TAT GTT ACC CAT GAC	CTC AAC CCT TTC ACA CAA CAG
DS00146		4B	(GT)10	53	GAG TCA ACG AGC CAG CAA AGT	AAC AAT ACA GAG CAG CAC ACG
X3999387ca		4B	(CA)15	56	CCA ACA ACT GGA ACA TAA TTG	AGG TGC GAG CAA CTA AAA GTG
X4071888gt		4B	(GT)16	55	CCA CGG CGA AAT CTT ATC AAA C	ACC ATC TCA GGG CTG GGG ACC
AE002566_ca		4C	(CA)25	53	TTA GCA GAG GCA AGA ACC	TAC TCG TTC GGT TGT AGT GG
X4275758gt	x	4C	(GT)12	55	CCC CCA TCA ATA CAT TTG TAT G	ATT TTT GCT GAA AAC TCG TGC G
X4500516ga		4C	(GA)13	58.5	TGG TGC TTC GCA GCT TCT C	GGA GCG AGC GAG ACG GCA G

X4631731ca		4C	(CA)13	55	TTA TGG CAC AAA ATA AAT TCC	TTC GAG TCT CTT GCT CTG C
X4814651ct		4D	(CT)12	2-Step: 68/94°C	GCA AAC GTG TGC CAA GCA GTG	ATG CTT AAC GCA GCG GCA GTG
X4944599ca		4D	(CA)14	55	GTC CTG CTG CGT TGA TTA AAC	TAG ACA CCC TGA ATG TGA ATG C
X5029944gt	x	4E	(GT)14	55	ACT GGT GAG TAC TGG TCG AAA G	ATG TCT TGA AGC TGG AAA TAA G
X5179712gt		4F	(GT)15	53	GAG TCA CCT AAC GAT TCT TGC	ATG TTG CAG GTT CTT ATG ATC
X5326452ct		5A	(CT)16	58	CCT GAT CGT TTC GTC CCA CTG	AAT TCT CCC ATC GTT ACA CTC G
X5408669ca		5A	(CA)14	52	CAA CGC TAC ACG AAT TTG TTA C	TTA CAA ACA CAT ATA ACA ACC G
X5592060ca		5B	(CA)14	2-Step:68/94°C	GAT GAA AGC GAG AGT GGG CAG C	ACC ATC GCC CAT TGT CCC ACT G
X5710427		5C	(TG)5(AG)3	53	GTT GAA TTG CGG CGG CCA AGT	AAA TGG AGA AAA CCT TCG AGC
DS00589		5C	(CA)11	52	CGT TTT TTA TTT GCG GGC AG	ACA TCC CTC TCT TTC GCT TC
X5973753gt		5D	(GT)17	52	ATC TTC AGC TTG CAG CCT TTG	GTG GCA TAA AAT AAA TAA ATG AC
X6213328ca	x	5F	(CA)13	55	CAC TTA TTT ATT ATT GGC CTG AC	GGG TTG CAG CAG CTT AAG GAC AC
X6325133ca		6A	(CA)12	51	GGA TGT TCA AAT GGT TCA AGG	CAT TTT CAT AAG ACG CTC AAC
DS06329		6C	(GT)12	54	CCT GGT TGC TCC CGC TGC	TTC CGA GAT CAC CTG AGA
X6694934gt		6D	(GT)14	55	CAT CAT CAT CGT GTG CTC TCC	CGA TCT GAT GTG GCC CAC TTC
DS04440		6E	(GT)9	55	TTC TCC CAC CGT AAC GCC CTA T	ACA CAA CAT CCG TTG CTG CTG T
X7028104ga		7A	(GA)16	55	CTG AAC TCC AGA GAG AAC TGC	GAC AAT GCT CCA CAG ATC CTG
X7192669ca		7B	(CA)12	52	TTT GTG AGG CTG TCA TGT GTC	TCG TGT AAC ATA AAA TCT TGT GG
X7340418ga	x	7B	(GA)12	55	ATT TCC TGG TAA TAA AAC AGA GAG	GCA GCG GCG ATA CGT TAG TC
X7586980ca		7C	(CA)14	55	CTG CAA ACT TGA CGA CAA AAG	CGT TTT TAG CCA ATT CCA ATG
X7809164ca		7D	(CA)13	55	AAA GAA CGT GTT ATT TAT GGT C	CGT TAG TTA TTA CTT GGC ATC
X8022709ca		7E	(CA)15	52	AAC ACT GGC AAC AAA TAA ACT C	ACG TTT TCA AGT CGA GTG TTT G
X8312980gt		7F	(GT)12	55	GAT GGT GAC GAA GAA GAA GAG C	TCA AAA AGT GAT CTC TTG AAG C
X8490060gt		8A	(GT)14	55	GTC TGT GTC CAT TGT TTT GTC G	CGC TCG TTC ACT TAC TCA CTT A
DS09021		8B	(GT)12	52	TTCCCGCATATGTGTGAG	TTTCGTGTACTTCTCGGTGC
X8756567gt		8C	(GT)12	55	TTG TGA AAT GCG GTC ATC TAC	ACA GAC AAT GCG AAC AAA GGC
AE002566_atc	x	8D	(ATC)14	52	TTT ACC AGA TTG CCG TTG	GCA GAT TGA TGA TGA GCC
X8956947gt		8D	(GT)13	55	AGC AGG AGT AAA GAA GAG CTT G	TTT GCG TTA AGT TTA CGT TAC G
X9312943		8E	(GT)14	55.4	ACC CCA TTT TTG CTA CGC CTC	GCA TTA GGG ATA ACG ATG TGT
X9325355		8E	(AC)2GC (AC)2TC (AC)6	52.3	CTG GCT ATT TCG TTC TTG AAG	GTG TGC AAG GCT GCG TAT ATG
DS01391		9A	(GT)10	57	GCCTGCTGCAGTCGCATGTG	CCAGCGGCATACGTGTAAAC
X9928573gt		9B	(GT)13.5	55	GTT GTG CCT CTG CCA GTC AGT C	GAA TTA TTT CAC GAT TAT CTT CAG G
X10509490gt		9E	(GA)13	52	CAA AGC AAT TTT TTG CGT TAG	CAA ACA AGC ACA CAA ACT CAC

X10809186ga		10A	(GA)13	58	GGC TAT TTG AGT GGC GAA AG	CAG CAG AGC AGA GCC AGC AC
X11087689ca		10B	(CA)14	55	GCC GCA ATT TGA AGT GCT AC	TCT CAT TTT CGC TTT TAT GC
X11347407ca		10C	(CA)12	58	CTG CCT GCT GTT CGT TGT GG	CCT ATA ACC ATT ATG CCC ACC CAC
X11601257ca		10F	(CA)10T (CA)11.5	55	TTT GTA AAA TTG ATA TCC TGC C	TAA CCA TTA AAT TCC AAC TGT GTG
X11804446gct		11A	(CTG)13	55	GGT GTA TGG GTA ATG TCC TTG C	TGG GGA AAG TGT CAA GTT AAT G
X12075563gt		11A	(GT)12.5	52	ATA TCC TTT TTC TCT CGG TGT G	ACG ACT TAG TTG ACT TTT GTG C
P11B01ta		11B	(TG)10TA (TG)6	50	AATGATCTGTCGCATATACC	TTTATGAAAACAACACATGC
P11B01tg	x	11B	(TG)11	56	GCTTTGCTAATGTCGTGTTGT	GTGTCCACTGTGCGGAG
P11B01ca		11B	(CA)9	51	ACATCGCCAGGATTCAC	TCCCGTTTACAGCAACAA
X12683051ct		11C	(CT)15	55	ACA TAG GCT CCA TCT CAT TC	AAA GCG ATT TGA AGT TGT GC
X13039889ca		11E	(CA)15	58	GCT TTA CGG GTC GGT CGG TC	ACT TGC CTT TCA AAT GGA TGG TG
X13203739gt		11F	(GT)20	55	GTG AGT GGG TGG CAA ATA CTG	GAG TTA CAA CCA AAT GAG TAA GAT G
X13631919ag		12C	(GA)10.5T A (AG)11	50	ATT TGT TTA CTT GGA GTG AGT GAG	TAA ATA CTC TAC TTG ATT ATC TTC
X14128213ca		12E	(CA)15	58	AAC TTT CGC CAC CAG TGT CTG	AAT ACC CAG CGG AAC GAG AAC
P12F01ca		12F	(CA)16	53	TTGTCGTCACCTGGGAAAG	GCCAAAGGGTCATCAGCA
X14425888gt	x	12F	(GT)13	58	TTT TGG TAT GTG AAT GCG GCT G	GAT GGA GGC TAA GTG CGG AAC G
X14765146ct		13A	(CT)12	55	AAC GAT AAT GGA GTG GGT TC	ATG TAA TGC GCA AGT TGC TG
X15146508gt		13C	(GT)11	52	TTG TGT GAG TGT AAG TGT GCG T	GTA AGT TTA TTC TCT TGC GTT C
X15279912atc	x	13E	(ATC)12	57	TGC CAG ACG CCA TAA TCA TCA C	AGT GCC TTG GTC ATT TGC CTC G
X15544500ata		13F	(ATA)10	50	GGA TCT GAA ACA GAC CGT GG	AAC TTA GCA CAC ACG AAC GC
3641.2		14A	(GGT)5	55	GAT TTT CTC GTT CAG CAC G	CGC TGT TCA AAG AAG CAC T
X15830711gt		14A	(GT)11	55	TCT CTC ACT TTT GAC ACT CTC C	ATT TTA TGT TGT GAA GAG CGA C
X15854539ta		14A	(TA)6	52	GAC TTC CTT CCT GTT TTA TCT G	ATT GTA TTG CCT GTG AAA TGA G
X15959225ca		14C	(CA)11	58	GCT GTG TGC TGT TGT TGT TTT C	TCC AGT CCA CTC TTC TCT CTG C
X16203512gt		14D	(GT)13	52	GTT CGG TAC TGT TGT CGC TTT	ATA TTA GAG GTT ATG TCT CAT TTT G
X16183204ga	x	14D	(GA)7.5	58	CGA CGG CGA CTT CAT CAT GCT G	TTC CAC TTT CCC ATC TCG TTT C
X16262195ga		14E	(GA)9.5	58	GTT GCG GTC CAG TTA GAT GGG TC	TTG AGC CCC AGC CAT CCA ATC
X16349316ga		14F	(GA)12.5	58	TCA CCC ACA AAT GCG AGC GAG AC	AAA GTA CGG AAA GAG GGC GAC AG
DS09020		15A	(GT)8(GC) 4	51	CTA AAC AGG ATG CAG GAC AAC	GCG ATC AAG GTT AAA TGG TTC
X16416576ta		15A	(TA)15	52	ATG AGT GTG ACA TTC TTT CG	AGA GGT GGA GGA ACT TGG C
X17276522ct		16B	(CT)19	55	GAT TCT TGT TTA TCC TCC AGC	CGT GTC CAG TTC TCA GAC TTG
X17869774gt	x	17A	(GT)14	58	CTG CGT GAG TGC GTG CGT GC	CAC TGT CCC CAT CCA CAT ACC G
X18374645ta		17C	(TA)16	50	AGT GGA CGG AAC ACA GAG TAA C	AAA TTC AGG GTC GCA CTG GAT G
X19942721gt	x	19C	(GT)18	57	CTG CTG CCA ACT CCA TTC TG	TTC TGC CAC ATT TGC GAT TC
X18478793ga	x	19D	(GA)10	55	ATG CCA TTA AGA GCC TAC CAT C	GGC TGC CTC AAG TGT ATT TTA C

X20639205ca	x	19E	(CA)27	55	GGT GAG CAA ATG GGT GGT AG	TTC TCC TTG GTC CTT TCT CT

[1] Loci marked mit x are used in data set 1

[2] the number of repeats is taken from the published *D. melanogaster* genomic sequence. Non-integer repeat numbers indicate that only a part of the repeat type was present (e.g.: (TA)2.5 corresponds to TATAT)

Table A2 *D. melanogaster* populations used[1]

	Location		Collected by	Provided by	Sampling date	Number of individuals	Isofemale line [2]
Rom	Rome	Italy	C.Schlötterer	C.Schlötterer	07/1998	30	no
Frd	Friedrichshafen	Germany	B. Harr	B. Harr	1998	30	no
Got	Gotheron	Switzerland	J.David	J.David	1998	30	no
Nyo	Nyon	Switzerland	J.David	J.David	1998	30	no
Kph	Copenhagen	Denmark		V. Loeschke	1999	30	no
Nje	Rockaway	New Jersey	S.Weiss	S.Weiss	1999	30	no
KR	Kreta	Greece	M. Hans	A. Hans	08/1998	8	no
TU	Medoun on Island Djerba	Tunisia	M. Puchinger	M. Puchinger	2001	24	no
Mar	Marrakech	Morocco	J.David	J.David	12/1998	8	yes
AG	Agadir	Morocco		P. Capy		4	yes
Mali	Moribabougou	Mali	A. Djiteye	A. Djiteye	01/2001	10	yes
Lam	Lamto	Ivory Coast	D. Lachaise	P. Capy		10	yes
AB	Abidjan	Ivory Coast	D. Lachaise	P. Capy		4	yes
Ken	various locations	Kenya		Drosophila stock center		18	yes
Kyo	Malindi New Market	Kenya	B. Ballard	B. Ballard	07/2001	21	yes
ZS	Sengwa. Wildlife Reserve	Zimbabwe		C.F. Aquadro and C.-I Wu	1990	13	yes
ZH	Harare. Capital City	Zimbabwe		C.F. Aquadro and C.-I Wu	1990	13	yes
Zlk	Lake Kariba	Zimbabwe		V. Bauer	1994	11	yes
Kis	Kisoro	Uganda	M. Imhof	M. Imhof	03/2001	15	no
Kam	Kampala	Uganda	K. Pieta	K. Pieta	05/2000	10	no
ZW	Victoria Falls	Zimbabwe	B. Ballard	L. Partridge lab	07/2001	28	yes

[1] blanks indicate that no information is available

[2] yes indicates that the line has been maintained for several generations by brother sister mating. No indicates that the flies were the offspring from freshly collected flies

Table A3: pairwise F_{ST} and P-values assuming Hardy-Weinberg equilibrium and unlinked loci (based on 17 loci)

	Rom	Frd	Got	Nyo	Kph	NJe	KR	TU	Mar	AG	Mali	Lam	AB	Ken	ZS	ZH	Zlk	Kyo	Kis	Kam	ZW
Rom		0.047	0.05	0.039	0.036	0.058	0.03	0.041	0.041	0.131	0.21	0.045	0.226	0.193	0.181	0.217	0.19	0.185	0.186	0.177	0.195
Frd	***		0.048	0.039	0.039	0.034	0.065	0.057	0.037	0.134	0.219	0.026	0.241	0.203	0.204	0.231	0.203	0.196	0.195	0.186	0.211
Got	***	***		0.02	0.046	0.053	0.086	0.054	0.031	0.092	0.196	0.012	0.213	0.194	0.186	0.225	0.197	0.186	0.176	0.183	0.195
Nyo	***	***	***		0.04	0.05	0.063	0.066	0.023	0.113	0.231	0.007	0.232	0.227	0.22	0.253	0.229	0.22	0.205	0.21	0.229
Kph	***	***	***	***		0.062	0.076	0.063	0.053	0.148	0.252	0.024	0.295	0.242	0.231	0.268	0.239	0.232	0.233	0.23	0.243
NJe	***	***	***	***	***		0.055	0.045	0.038	0.136	0.213	0.027	0.228	0.19	0.207	0.232	0.19	0.19	0.187	0.188	0.2
KR	n.s.	***	***	***	***	***		0.031	0.073	0.17	0.236	0.046	0.232	0.178	0.163	0.198	0.198	0.183	0.214	0.194	0.194
TU	***	***	***	***	***	***	n.s.		0.04	0.124	0.181	0.043	0.198	0.167	0.158	0.191	0.172	0.158	0.18	0.161	0.172
Mar	**	n.s.	n.s.	n.s.	***	**	**	n.s.		0.002	0.165	-0.021	0.144	0.156	0.147	0.175	0.138	0.155	0.132	0.16	0.171
AG	***	***	***	***	***	***	***	***	n.s.		0.22	0.048	0.162	0.166	0.157	0.205	0.194	0.199	0.168	0.215	0.215
Mali	***	***	***	***	***	***	***	***	***	***		0.18	0.034	0.044	0.045	0.053	0.084	0.024	0.007	0.028	0.035
Lam	***	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	**	n.s.	n.s.	***		0.144	0.173	0.165	0.194	0.15	0.17	0.136	0.165	0.185
AB	***	***	***	***	***	***	***	***	***	***	n.s.	***		0.041	0.024	0.032	0.044	0.042	-0.019	0.048	0.043
Ken	***	***	***	***	***	***	***	***	***	***	***	***	n.s.		0.01	0.017	0.051	0.001	0.018	0.024	0.013
ZS	***	***	***	***	***	***	***	***	***	***	***	***	n.s.	n.s.		-0.012	0.04	0.012	0.028	0.014	0.017
ZH	***	***	***	***	***	***	***	***	***	***	***	***	n.s.	n.s.	n.s.		0.053	0.014	0.043	0.035	0.022
Zlk	***	***	***	***	***	***	***	***	***	***	***	***	n.s.	***	***	***		0.036	0.034	0.073	0.042
Kyo	***	***	***	***	***	***	***	***	***	***	**	***	n.s.	n.s.	n.s.	n.s.	***		0.009	0.016	-0.001
Kis	***	***	***	***	***	***	***	***	***	***	n.s.	***	n.s.	n.s.	***	***	n.s.	n.s.		0.032	0.02
Kam	***	***	***	***	***	***	***	***	***	***	**	***	***	***	n.s.	***	***	n.s.	***		0.025
ZW	***	***	***	***	***	***	***	***	***	***	***	***	**	n.s.	**	***	***	n.s.	***	***	

n.s. is not significant ≥ 0.1. * ≥ 0.05. ** ≥ 0.01. *** <0.01

Table A4: pairwise F_{ST} and P-values assuming Hardy-Weinberg equilibrium and unlinked loci (based on 99 loci)

	Rom	Frd	Got	Nyo	Kph	Nje	Mar	AG	Lam	AB	Ken	ZS	ZH
Rom		0.036	0.036	0.026	0.040	0.043	0.041	0.110	0.037	0.229	0.192	0.248	0.228
Frd	***		0.042	0.037	0.050	0.032	0.041	0.109	0.038	0.236	0.194	0.249	0.229
Got	***	***		0.022	0.055	0.048	0.055	0.116	0.028	0.232	0.199	0.247	0.232
Nyo	***	***	***		0.044	0.037	0.037	0.100	0.028	0.229	0.192	0.238	0.229
Kph	***	***	***	***		0.052	0.060	0.121	0.054	0.250	0.195	0.245	0.233
NJe	***	***	***	***	***		0.026	0.087	0.034	0.217	0.173	0.225	0.211
Mar	***	***	***	***	***	***		0.026	0.011	0.163	0.125	0.177	0.164
AG	***	***	***	***	***	***	**		0.042	0.129	0.120	0.161	0.155
Lam	***	***	***	***	***	***	n.s.	***		0.167	0.161	0.209	0.197
AB	***	***	***	***	***	***	***	***	***		0.018	0.032	0.028
Ken	***	***	***	***	***	***	***	***	***	***		0.020	0.012
ZS	***	***	***	***	***	***	***	***	***	***	***		0.003
ZH	***	***	***	***	***	***	***	***	***	***	***	n.s.	

n.s. is not significant ≥ 0.1. * ≥ 0.05. ** ≥ 0.01. *** <0.01

Allele excess at neutrally evolving microsatellites and the implications for tests of neutrality

Christian Schlötterer*, Max Kauer and Daniel Dieringer

Institut für Tierzucht und Genetik, Veterinärmedizinische Universität Wien, Josef Baumann Gasse 1, 1210 Wien, Austria

Skews in the observed allele-frequency spectrum are frequently viewed as an indication of non-neutral evolution. Recent surveys of microsatellite variability have used an excess of alleles as a statistical approach to infer positive selection. Using neutral coalescent simulations we demonstrate that the mean numbers of alleles expected under the stepwise-mutation model and infinite-allele model deviate from the observed numbers of alleles. The magnitude of this difference is dependent on the sample size, mutation rates (θ-values) and observed gene diversities. Moreover, we show that the number of observed alleles differs among loci with the same observed gene diversity but different mutation rates (θ-values). We propose that a reliable test statistic based on allele excess must determine the confidence interval by computer simulations conditional on the observed gene diversity and θ-values. As the latter are notoriously difficult to obtain for experimental data, we suggest that other statistics, such as $\ln RV$, may be better suited to the identification of microsatellite loci subject to selection.

Keywords: neutrality test; microsatellites; allele excess; hitchhiking mapping

1. INTRODUCTION

Over the past decades several theories have been developed to explain the observed levels of variability in natural populations. One interesting subject, which has arisen from this research, is the inference of past selective events from extant natural variability. Several statistical tests have been developed that use sequence variation to distinguish between neutrally evolving genes and selected ones (reviewed in Otto 2000). With the recent progress in high-throughput technology, emphasis is shifting from single-locus studies to complete-genome scans aiming to detect genomic regions that recently acquired a beneficial mutation (Schlötterer 2003). The general idea of such a genome scan is that the fixation of a beneficial mutation in a population also affects sites linked to the target of selection. This phenomenon has been called hitchhiking (Maynard Smith & Haigh 1974; Barton 2000). Hence, screening a large number of markers is expected to identify those linked to a selected site, which therefore deviate from neutral expectations.

Given the high costs of a DNA-sequencing-based genome scan, other more cost-effective genetic markers are required. Microsatellites are highly polymorphic DNA regions distributed over the euchromatic part of the genome in all eukaryotic organisms (Ellegren 2000; Schlötterer 2000). The ease of microsatellite typing in combination with their predominantly neutral evolution renders microsatellites an excellent marker for genome scans (Schlötterer 2004). Nevertheless, in contrast to the analysis of DNA sequences, only a limited number of statistical tests are available to compare observed patterns of microsatellite variability with their neutral expectations.

Microsatellite mutations encompass gains and losses of repeat units (Ellegren 2000; Schlötterer 2000). This stepwise-mutation process was originally studied in the context of protein evolution. The statistical properties of this model have also been used to describe the mutation dynamics of microsatellites. Kimura & Ohta (1975) derived an analytical formula for the expected number of alleles based on the stepwise-mutation model (SMM). The observed number of alleles can be compared with expectations based on the observed genetic diversity at this locus. As the observed and expected numbers of alleles should not differ significantly under neutrality, a simple test statistic can be developed. Comparing the observed genetic diversity with the genetic diversity expected from the number of observed alleles has been used to infer deviations from the stepwise-mutation behaviour of microsatellites (Shriver *et al.* 1993; Estoup *et al.* 1995) and neutrality of microsatellite variability (Michalakis & Veuille 1996) and to identify single loci deviating from neutral expectations (Payseur *et al.* 2002; Vigouroux *et al.* 2002*b*). As the same discrepancy is tested in these scenarios, i.e. observed and expected genetic diversities conditional on the observed number of alleles or the observed and expected numbers of alleles conditional on the observed genetic diversity, we will refer to these statistical tests collectively as the allele-excess test statistic.

The analytical formulae were tested for low mutation rates only (Kimura & Ohta 1975). Recent computer simulations have demonstrated that the expected number of alleles is underestimated for loci with a high mutation rate (Shriver *et al.* 1993). Therefore, a recently developed test that compares observed and expected genetic diversities relies on computer simulations to determine the expected genetic diversity (Cornuet & Luikart 1996). In this report we focus on the suitability of the allele-excess test statistic for the inference of selection at individual microsatellite loci.

2. MATERIAL AND METHODS

We used a commonly employed coalescent-based computer simulation algorithm (Hudson 1990), which has been modified

to account for the stepwise-mutation behaviour of microsatellites. Rather than counting the number of mutations occurring on a branch, our simulations traced the allele length of a microsatellite locus. The number of mutations occurring on a branch was converted into microsatellite mutations by adding or removing (with equal probability) one repeat unit for each mutation. The accuracy of the code was checked by comparing the observed variance in repeat number with its expectation, $E(V)$, ($E(V) = \theta/2$, where θ is the mutation rate). Genetic diversities, H, were calculated as

$$H = \frac{n}{n-1}\left(1 - \sum_{i=1}^{m} x_i^2\right),$$

where m is the number of alleles, n is the number of analysed chromosomes and x is the allele frequency. Calculated genetic diversities were verified by using the simulated allele frequencies as input in the MSA software, which was independently written to compute microsatellite-specific statistics (Dieringer & Schlötterer 2003). Finally, we tested the code with a different microsatellite-evolution program, which uses a different algorithm to generate random numbers and Poisson deviates (kindly provided by T. Wiehe).

Unless otherwise noted, θ-values were drawn from a uniform distribution between 0.1 and 10.1 to account for heterogeneity in microsatellite mutation rates. We simulated 30 000 unlinked loci for each combination of parameters and a sample size of 100 chromosomes.

We also performed computer simulations accounting for the observed distribution of microsatellite variability in natural populations. Previous studies have demonstrated that the natural logarithm of the observed variance in a repeat number follows a normal distribution (Goldstein *et al.* 1996; Harr *et al.* 1998). Therefore, we used the mean (1.96) and standard deviation (1.28) of the natural logarithm of V observed in African *Drosophila melanogaster* populations (Caracristi & Schlötterer 2003) to describe the normal distribution from which we sampled the log $\theta/2$-values for our computer simulations.

The expected number of alleles under the SMM was calculated as described by Kimura & Ohta (1975):

$$n_{\text{expected}} = \frac{\theta + \beta}{\beta}\left\{1 - \prod_{i=0}^{2N-1}\left(\frac{i + \theta}{i + \theta + \beta}\right)\right\}, \tag{2.1}$$

where θ with

$$\theta = 4N_e\mu = \left(\frac{1}{H_o^2} - 1\right)\frac{1}{2} \tag{2.2}$$

and

$$\beta = \frac{\theta + 1 - \sqrt{1 + 8N_e\mu}}{\sqrt{1 + 8N_e\mu} - 1} = \frac{\theta + 1 - \frac{1}{H_o}}{\frac{1}{H_o} - 1} = \frac{H_o\theta + H_o - 1}{1 - H_o}. \tag{2.3}$$

The expected number of alleles under the infinite-allele model (IAM) was calculated as described by Watterson (1975):

$$n_{\text{expected}} = \sum_{i=1}^{2N}\frac{\theta}{\theta + i - 1}, \tag{2.4}$$

with

$$\theta = \frac{1 - H_o}{H_o}. \tag{2.5}$$

N_e is the effective diploid population size, N is the number of diploid individuals in the analysed sample, H_o is the expected homozygosity and μ is the microsatellite mutation rate.

Allele excess was determined as

$$\text{AE} = \frac{n_{\text{observed}} - n_{\text{expected}}}{n_{\text{expected}}}. \tag{2.6}$$

3. RESULTS

We simulated 30 000 microsatellite loci using a broad range of sample sizes and θ-values. For each dataset the mean allele excess was determined based on the SMM and the IAM. While for most simulations the SMM resulted in allele excess, the IAM indicated an allele deficiency (table 1). The closest fit to the expectation for both models was obtained for very low θ-values ($\theta = 0.05$). The SMM also provided a good fit to the expectations in the simulation based on small sample sizes ($n = 10$). Interestingly, the mean allele excess was strongly influenced by the sample size. For the SMM a larger sample size resulted in more pronounced allele excess, while a more extreme allele deficiency was observed at large sample sizes under the IAM. The same trend was observed when no correction for sample size was made to the genetic-diversity estimate (data not shown). We also found that θ-values were linked to allele excess. An increase in θ resulted in a higher allele excess under the SMM and a larger allele deficiency under the IAM. Deviations from the strict SMM that allowed for larger changes in repeat number (two-phase model; Di Rienzo *et al.* 1994) increased the allele excess under the SMM, but resulted in a less pronounced allele deficiency under the IAM (table 1).

As some tests of neutrality based on the excess of alleles focus on individual loci, we were interested in the distribution of allele excess over the range of the simulated data. To investigate this, we grouped data simulated from a broad range of θ-values into different (observed) genetic-diversity classes and determined the allele excess for each of the classes separately. For an unbiased test statistic the observed allele excess should be independent of the observed genetic diversity. As expected, no allele excess was observed for monomorphic loci (figure 1). Very strong allele excess was observed for loci with low levels of observed genetic diversity ($0 < H < 0.1$; figure 1). Interestingly, both the IAM and SMM resulted in almost the same very strong allele excess. Only for larger genetic diversities did the difference between the two mutation models become apparent (figure 1). Further computer simulations based on different sample sizes consistently indicated that the lowest-genetic-diversity class (excluding $H = 0$) had the most extreme allele excess (figure 2*a*,*b*). The same phenomenon was observed when computer simulations were performed with fixed θ-values. Irrespective of the θ-value used, the lowest-genetic-diversity class had the most pronounced allele excess (figure 3*a*,*b*). These results clearly indicate that the average allele excess and its confidence interval are not well suited to determining the significance of allele excess, as the mean allele excess differs substantially among different classes of observed genetic diversity. One further problem of the allele excess becomes apparent when comparing dif-

Table 1. Comparison of the allele excess based on the SMM and IAM using simulated microsatellite data. (Standard deviations of all means are given in brackets.)

θ	0.1–10.1	0.1–10.1	0.1–10.1	0.05	0.5	5.0	50.0	0.1–10.1[a]
sample size (N)	10	100	1000	100	100	100	100	100
mean θ	5.116 (±2.882)	5.097 (±2.884)	5.089 (±2.901)	0.050	0.500	5.000	50.000	5.092 (±2.881)
mean V	2.583 (±4.174)	2.532 (±3.670)	2.530 (±3.778)	0.025 (±0.072)	0.250 (±0.354)	2.491 (±2.863)	25.077 (±29.271)	5.276 (±7.913)
mean H	0.650 (±0.214)	0.648 (±0.186)	0.648 (±0.184)	0.047 (±0.120)	0.293 (±0.207)	0.699 (±0.097)	0.901 (±0.028)	0.684 (±0.188)
mean number of observed alleles	3.723 (±1.337)	5.753 (±2.203)	6.556 (±2.473)	1.238 (±0.458)	2.477 (±0.826)	5.998 (±1.489)	15.555 (±3.371)	7.176 (±2.830)
expected number of alleles SMM	3.904 (±1.403)	5.218 (±1.776)	5.356 (±1.661)	1.237 (±0.607)	2.511 (±1.104)	5.462 (±1.205)	12.710 (±2.867)	5.823 (±2.108)
expected number of alleles IAM	4.310 (±1.520)	9.418 (±4.089)	15.021 (±7.239)	1.338 (±0.934)	3.367 (±2.033)	9.981 (±2.914)	23.842 (±4.312)	10.789 (±4.744)
(o − e)/e SMM[b]	−0.031 (±0.163)	0.118 (±0.265)	0.232 (±0.290)	0.042 (±0.233)	0.104 (±0.416)	0.113 (±0.227)	0.237 (±0.177)	0.249 (±0.302)
(o − e)/e IAM[b]	−0.118 (±0.159)	−0.324 (±0.262)	−0.478 (±0.298)	0.023 (±0.255)	−0.047 (±0.480)	−0.371 (±0.172)	−0.346 (±0.088)	−0.269 (±0.275)

[a] In this column 30% of the microsatellite mutations encompassed more than a single repeat unit. The upper boundary for a change in repeat number was fixed to three repeats.
[b] (o − e)/e, difference between observed and expected expressed as a fraction of the expected number of alleles.

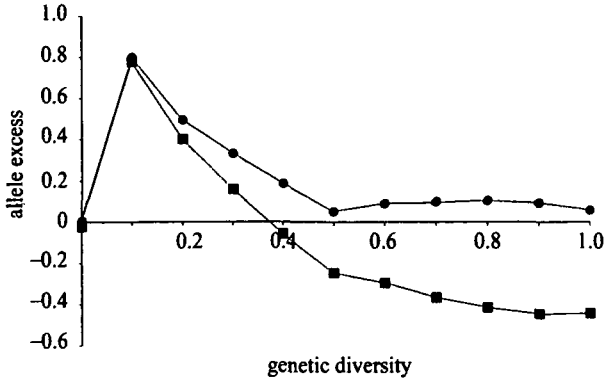

Figure 1. Dependence of allele excess on observed genetic diversity (squares, IAM; circles, SMM). Genetic diversities obtained from neutral coalescent simulations were grouped into 11 bins and for each bin the mean allele excess is shown. Parameters for the coalescent simulations were: $\theta = 0.1$–10.1, $N = 100$ chromosomes.

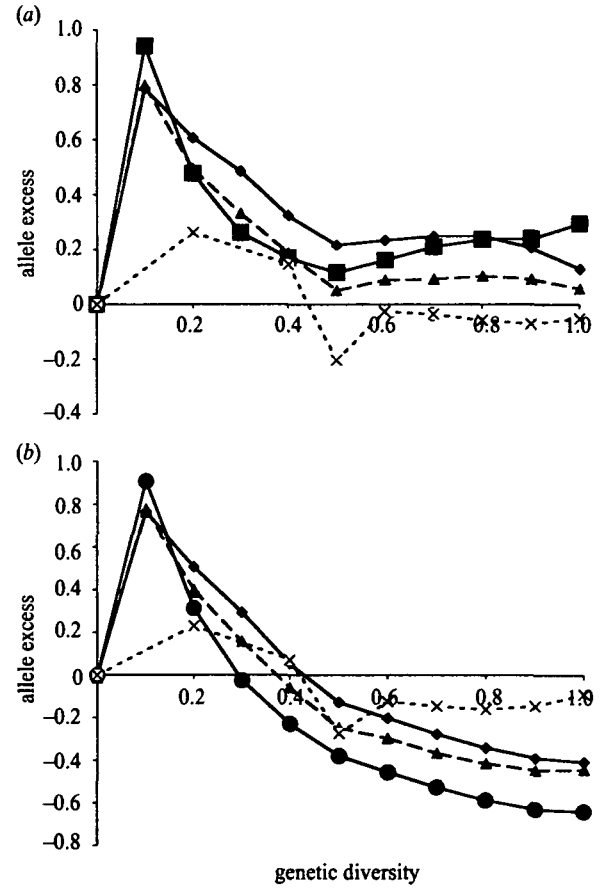

Figure 2. Allele excess and sample size. (*a*) SMM (diamonds, $N = 100$, two-phase; squares, $N = 1000$; triangles, $N = 100$; crosses, $N = 10$). (*b*) IAM (diamonds, $N = 100$, two-phase; circles, $N = 1000$; triangles, $N = 100$; crosses, $N = 10$). Genetic diversities obtained from neutral coalescent simulations were grouped into 11 bins and for each bin the mean allele excess is shown. For all simulations θ was drawn from a uniform distribution between 0.1 and 10.1.

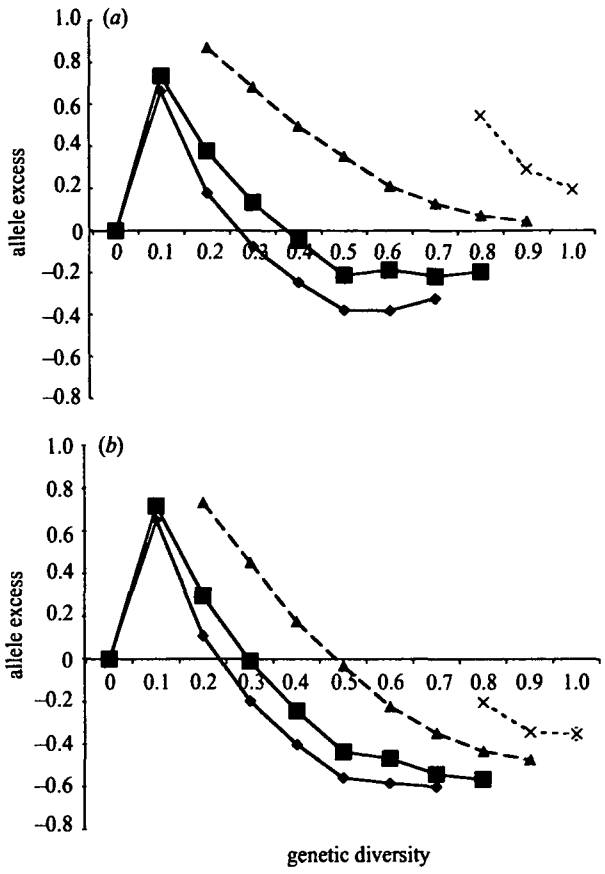

Figure 3. Allele excess and θ. (*a*) SMM (diamonds, $\theta = 0.05$; squares, $\theta = 0.5$; triangles, $\theta = 5$; crosses, $\theta = 50$). (*b*) IAM (diamonds, $\theta = 0.05$; squares, $\theta = 0.5$; triangles, $\theta = 5$; crosses, $\theta = 50$). Genetic diversities obtained from neutral coalescent simulations were grouped into 11 bins and for each bin the mean allele excess is shown. All simulations are based on a sample size of 100 chromosomes.

ferent θ-values (figure 3). As expected, simulations based on different θ-values generate overlapping distributions of genetic diversities. As a consequence, loci with genetic diversities between 0.4 and 0.5 showed a mean allele deficiency (AE = −0.21, SMM) when they were simulated with θ-values of 0.5. For simulations using an θ-value of 5.0, loci with a genetic diversity between 0.4 and 0.5 showed an allele excess (AE = 0.35, SMM). This difference is highly significant ($p < 0.0001$, Mann–Whitney *U*-test), indicating that, despite a similar genetic diversity, the number of alleles at a given locus is determined by its mutation rate.

As the analytical formula by Kimura & Ohta (1975) underestimates the expected number of alleles for loci with high mutation rates, the difference in allele excess could also result from this bias. Therefore, we compared the numbers of observed alleles for different θ-values (table 2), but the same trend could be recognized. Depending on the θ-value used for the computer simulation, we detected substantial differences in the mean number of alleles observed within a genetic-diversity class.

So far, we have considered only θ-values drawn from a uniform distribution. As the natural logarithm of the variance in repeat number follows a normal distribution (Goldstein *et al.* 1996; Harr *et al.* 1998), we also performed computer simulations using log θ-values drawn from a normal distribution where the mean and standard deviation were estimated from an African *D. melanogaster* population (Caracristi & Schlötterer 2003). As expected, we observed that simulation runs with low genetic diversity showed a pronounced allele excess (figure 4). We also examined the allele excess when $\theta = 4N_e\mu$ was determined by genetic diversity (equation (2.2)) or by the variance in repeat number ($\theta = 2V$). Both estimators showed the pronounced surplus of expected alleles for small genetic diversities. For larger genetic diversities, however, the variance-based estimator showed a more pronounced allele excess than did the genetic-diversity based one (figure 4).

4. DISCUSSION

Our simulations indicate that the mean numbers of observed alleles for small sample sizes and low θ-values are very similar to those predicted under the SMM (Kimura & Ohta 1975). For larger sample sizes and higher θ-values (in the range typical for microsatellites) we found a large discrepancy between the observed number of alleles and the expectation based on the analytical formulae (Kimura & Ohta 1975). Our observation is in qualitative agreement with that of a previous simulation study (Shriver *et al.* 1993).

Currently, allele excess is widely used for the identification of loci that are affected by natural selection (Estoup *et al.* 1995; Payseur *et al.* 2002; Vigouroux *et al.* 2002*b*). The challenge for such tests, particularly for genome scans, is that several evolutionary forces are influencing variability in natural populations. For example, low genetic diversity at a microsatellite locus may have different causes: (i) the microsatellite may have a low mutation rate, resulting in a lower expected genetic diversity than for loci with higher mutation rates; (ii) even loci with high mutation rates could have low levels of genetic diversity if the sampled alleles share a common ancestor in the past; and (iii) hitchhiking: if a microsatellite locus is closely linked to a genomic region that recently experienced a selective sweep, this microsatellite will have lower levels of variability. The task of any neutrality test is to interpret observed variation so as to distinguish the two neutral scenarios ((i) and (ii)) from the selection hypothesis. The underlying idea is that under neutrality the observed number of alleles should be consistent with the observed genetic diversity. If more alleles are observed than expected from the observed genetic diversity, this is regarded as evidence for selection. Our computer simulations indicated two possible complications in using this approach. First, the analytical formula underestimates the expected number of alleles, leading to allele excess. This problem could be solved by using computer simulations to predict the expected number of alleles conditional on the observed genetic diversity (or alternatively to determine the expected genetic diversity conditional on the observed number of alleles). Second, the average number of alleles at a locus with a given genetic diversity depends on their mutation rates (θ-values). Note that this observation is independent of the analytical formula used to determine the expected number of alleles. Thus computer simulations conditioning on the observed genetic diversity are

Table 2. Observed numbers of alleles in different classes of genetic diversity ($N = 100$).

genetic diversity	$\theta = 0.5$	$\theta = 5$	$\theta = 50$
0	1 (±0)	—	—
> 0–0.1	2.19 (±0.40)	—	—
> 0.1–0.2	2.43 (±0.54)	3.50 (±0.67)	—
> 0.2–0.3	2.55 (±0.59)	3.90 (±0.94)	—
> 0.3–0.4	2.62 (±0.64)	4.21 (±0.99)	—
> 0.4–0.5	2.60 (±0.67)	4.49 (±1.04)	—
> 0.5–0.6	3.08 (±0.64)	4.76 (±1.05)	—
> 0.6–0.7	3.57 (±0.65)	5.34 (±1.10)	—
> 0.7–0.8	4.39 (±0.55)	6.34 (±1.18)	10.18 (±1.94)
> 0.8–0.9	—	7.98 (±1.26)	13.13 (±2.28)
> 0.9–1	—	—	17.42 (±2.82)

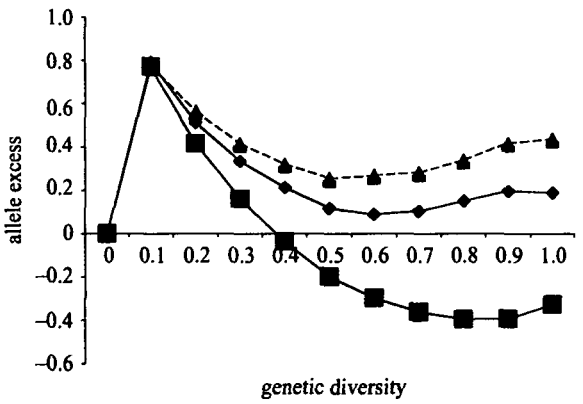

Figure 4. Allele excess for computer simulations based on a normal distribution of log θ-values. Genetic diversities obtained from neutral coalescent simulations were grouped into 11 bins and for each bin the mean allele excess is shown. All simulations are based on a sample size of 100 chromosomes. The expected number of alleles is based either on the observed genetic diversities (diamonds, SMM; squares, IAM) or on the observed variance in repeat number (triangles, SMM).

not well suited to addressing this discrepancy. An unbiased test statistic would require computer simulations conditional on the observed genetic diversity and the θ-values (a joint estimator of the microsatellite mutation rate and the effective population size). As mutation rates differ substantially among loci (Di Rienzo *et al.* 1994; Harr *et al.* 1998; Vigouroux *et al.* 2002*a*) in most experimental surveys of microsatellite variation, the required θ-values are not available. Consequently, it is extremely difficult to obtain an unbiased significance level for the allele-excess test statistic.

The outcome of the complex behaviour of the allele-excess test statistic is indicated in figure 1. Our computer simulations show that even under neutrality the mean allele excess was greatly elevated for loci with low observed genetic diversity. Consistent with this observation, those loci that were identified by the allele-excess test statistic as significant outliers had low genetic diversities (Vigouroux *et al.* 2002*b*). Similar results were obtained in a genome scan in *Drosophila*, which found that loci with a low genetic diversity were more likely to have an excess of alleles (Kauer *et al.* 2003).

5. CONCLUSION

We demonstrated a strong dependence of the allele-excess test statistic on both θ and the genetic diversity of each microsatellite locus. Given the difficulty in obtaining reliable locus-specific θ estimates, we suggest that results obtained with the allele-excess test statistic should be treated with caution. An alternative to the use of allele excess is a recently suggested statistic ($\ln RV$) that also uses microsatellite polymorphism for the inference of selective sweeps (Schlötterer 2002). Rather than contrasting observed and expected allele numbers, this test compares levels of variability for each locus in two populations. By calculating the ratio of the observed variances in repeat number, this statistic has an identical expectation for all loci, independent of their θ-values.

Finally, we note that similar problems will be encountered for DNA sequence data. Neutrality tests attempting to infer non-neutral evolution are based on θ-values estimated from polymorphism data. In contrast to the problem with microsatellites, this problem could be alleviated by the use of mutation-rate estimates from between-species divergence in combination with reliable estimates of population size.

We are grateful to T. Wiehe for sharing his code for the simulation of microsatellite data. R. Bürger, B. Harr and G. Muir provided helpful comments on the manuscript. This work has been supported by grants from the Fonds zur Förderung der wissenschaftlichen Forschung (FWF) and an EMBO young investigator award to C.S.

REFERENCES

Barton, N. H. 2000 Genetic hitchhiking. *Phil. Trans. R. Soc. Lond.* B 355, 1553–1562. (DOI 10.1098/rstb.2000.0716.)

Caracristi, G. & Schlötterer, C. 2003 Genetic differentiation between American and European *Drosophila melanogaster* populations could be attributed to admixture of African alleles. *Mol. Biol. Evol.* 20, 792–799.

Cornuet, J. M. & Luikart, G. 1996 Description and power analysis of two tests for detecting recent population bottlenecks from allele frequency data. *Genetics* 144, 2001–2014.

Dieringer, D. & Schlötterer, C. 2003 Microsatellite analyzer (MSA): a platform independent analysis tool for large microsatellite data sets. *Mol. Ecol. Notes* 3, 167–169.

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.

Ellegren, H. 2000 Microsatellite mutations in the germline: implications for evolutionary inference. *Trends Genet.* **16**, 551–558.

Estoup, A., Garnery, L., Solignac, M. & Cornuet, J.-M. 1995 Microsatellite variation in honey bee (*Apis mellifera* L.) populations: hierarchical genetic structure and test of the infinite allele and stepwise mutation models. *Genetics* **140**, 679–695.

Goldstein, D. B., Zhivotovsky, L. A., Nayar, K., Ruiz Lineares, A., Cavalli-Sforza, L. L. & Feldman, M. W. 1996 Statistical properties of the variation at linked microsatellite loci: implications for the history of human Y chromosomes. *Mol. Biol. Evol.* **13**, 1213–1218.

Harr, B., Zangerl, B., Brem, G. & Schlötterer, C. 1998 Conservation of locus specific microsatellite variability across species: a comparison of two *Drosophila* sibling species *D. melanogaster* and *D. simulans*. *Mol. Biol. Evol.* **15**, 176–184.

Hudson, R. R. 1990 Gene geneologies and the coalescent process. *Oxf. Surv. Evol. Biol.* **7**, 1–44.

Kauer, M. O., Dieringer, D. & Schlötterer, C. 2003 A microsite variability screen for positive selection associated with the 'out of Africa' habitat expansion of *Drosophila melanogaster*. *Genetics* **165**, 1137–1148.

Kimura, M. & Ohta, T. 1975 Distribution of allelic frequencies in a finite population under stepwise production of neutral alleles. *Proc. Natl Acad. Sci. USA* **72**, 2761–2764.

Maynard Smith, J. & Haigh, J. 1974 The hitch-hiking effect of a favorable gene. *Genet. Res.* **23**, 23–35.

Michalakis, Y. & Veuille, M. 1996 Length variation of CAG/CAA trinucleotide repeats in natural populations of *Drosophila melanogaster* and its relation to the recombination rate. *Genetics* **143**, 1713–1725.

Otto, S. P. 2000 Detecting the form of selection from DNA sequence data. *Trends Genet.* **16**, 526–529.

Payseur, B. A., Cutter, A. D. & Nachman, M. W. 2002 Searching for evidence of positive selection in the human genome using patterns of microsatellite variability. *Mol. Biol. Evol.* **19**, 1143–1153.

Schlötterer, C. 2000 Evolutionary dynamics of microsatellite DNA. *Chromosoma* **109**, 365–371.

Schlötterer, C. 2002 A microsatellite-based multilocus screen for the identification of local selective sweeps. *Genetics* **160**, 753–763.

Schlötterer, C. 2003 Hitchhiking mapping: functional genomics from the population genetics perspective. *Trends Genet.* **19**, 32–38.

Schlötterer, C. 2004 The evolution of molecular markers—just a matter of fashion? *Nat. Rev. Genet.* **5**, 63–69.

Shriver, M. D., Jin, L., Chakraborty, R. & Boerwinkle, E. 1993 VNTR allele frequency distributions under the stepwise mutation model: a computer simulation approach. *Genetics* **134**, 983–993.

Vigouroux, Y., Jaqueth, J. S., Matsuoka, Y., Smith, O. S., Beavis, W. D., Smith, J. S. & Doebley, J. 2002*a* Rate and pattern of mutation at microsatellite loci in maize. *Mol. Biol. Evol.* **19**, 1251–1260.

Vigouroux, Y., McMullen, M., Hittinger, C. T., Houchins, K., Schulz, L., Kresovich, S., Matsuoka, Y. & Doebley, J. 2002*b* Identifying genes of agronomic importance in maize by screening microsatellites for evidence of selection during domestication. *Proc. Natl Acad. Sci. USA* **99**, 9650–9655.

Watterson, G. A. 1975 On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.* **7**, 256–276.

Danksagung

Ich möchte mich zuallererst bei meinen Eltern bedanken. Wenn sie sich nicht so für mich eingesetzt hätten, wäre mein Leben wohl ganz anders verlaufen. Besonders erwähnen möchte ich meinen Vater, der diesen Moment nicht mehr erleben durfte, auch wenn ich weiß, daß er immer noch bei mir ist.
Weiters möchte ich mich auch bei dem Rest meiner Familie bedanken, die mir gezeigt hat, daß es nichts wichtigeres im Leben gibt als wenn man zusammenhält.

Meiner Frau möchte ich danken für die Hilfe in schweren Zeiten und für meine Tochter Teresa. Mag ich in meinem Studium viel nicht gelernt haben, so weiß ich nun mit Herz und Verstand, daß alles Wissen und aller Reichtum sinnlos ist ohne Kinder.

Ich möchte mich hier auch bei meinen Professoren an der TU bedanken, die mich überhaupt dazu brachten mich mit Genetik zu befassen, den eigentlich war dies das allerletzte was ich studieren wollte. Vor allem möchte ich Professor Kubicek erwähnen, der mir schon bei meiner Diplomarbeit geholfen hatte.
Das „CS-Lab“ mit all seinen Mittgliedern aus aller Welt war immer ein wichtiger Begleiter auf dem Weg durch die „Tiefen der Populationsgenetik“. Besonders bei Max Kauer möchte ich mich bedanken, an dessen Englisch ich bis heute zehre.
Ganz besonders möchte ich mich natürlich bei Christian Schlötterer bedanken, der mit Ausdauer, Begeisterung und einer Menge Energie aus mir weit mehr herausgeholt hat, als ich es eigentlich selber vorhatte. Wohin mich mein wissenschaftlicher Weg auch führen mag, er hat dazu einen wichtigen Eckstein gelegt.

LEBENSLAUF

Name:	Daniel DIERINGER
geboren am:	28. April 1973 Wien
Familienstand:	verheiratet
Tochter:	Teresa Maria, 18.05.2003
Präsenzdienst:	geleistet

1987-1992	HBLVA für chemische Industrie Rosensteingasse Ausbildungszweig Technische Chemie mit Matura
1.1993-8.1993	Präsenzdienst
10.1993-12.1999	Technische Chemie an der TU Wien, Studienzweig Biochemie Diplomarbeit zum Thema: „TupA, ein mögliches Element des Glukoserepressionssignalweges in *Aspergillus nidulans*", Institut für Biochemische Technologie und Mikrobiologie, Abteilung für Mikrobielle Biochemie (Prof. Kubicek)
12.1999	Diplomprüfung mit Auszeichnung
2000-2003	Doktorarbeit am für Tierzucht und Genetik. Arbeitsgruppe Dr. Schlötterer Veterinärmedizinische Universität „Selektive Sweeps in Drosophila"

Publikationsliste:

Hicks J, Lockington RA, Strauss J, Dieringer D, Kubicek CP, Kelly J and Keller N (2001) RcoA has pleiotropic effects on *Aspergillus nidulans* cellular development. *Mol Microbiol.* **39**: 1482-1493

Kauer M, Zangerl B, Dieringer D ans Schlötterer C (2002) Chromosomal patterns of microsatellite variability constrast sharply in African and non-African populations of *Drosophila melanogaster*. *Genetics* **160**: 247-256

Dieringer D and Schlötterer C (2003) MICROSATELLITE ANALYSER (MSA): a platform independent analysis tool for large microsatellite data sets. *Mol Ecol Notes*. **3**: 167-169

Kauer M, Dieringer D and Schlötterer C (2003) Nonneutral Admixture of Immigrant Geneotypes in African *Drosophila melanogaster* Populations from Zimbabwe. *Mol Biol Evol.* **20**: 1329-1337

Kauer M*, Dieringer D* and Schlötterer C (2003) A microsatellite variability screen for colonization associated positive selection in the genome of *D. melanogaster*. *Genetics*. **165**: 1137-1148
* both autors contributed equally

Schlötterer C, Kauer M and Dieringer D (2004) Allele excess at neutrally evolving microsatellites and the implication for tests of neutrality *Proc. R. Soc.* **271**: 869-874

Dieringer D and Schlötterer C (2003) Two distinct modes of microsatellite mutation processes - evidence from the complete genomic sequences of nine species. *Gen Res*. **13**: 2242-2251

Dieringer D and Schlötterer C (2003) Irregular Pattern of European Admixture in Sub-Saharan African Populations. *Genetics*. **Under review**