DETECTING DE NOVO MICROSATELLITE MUTATIONS IN A POPULATION OF FAMILIES WITH SPORADIC AUTISM

MITCHELL AVICHAI BEKRITSKY

Ph. D.

M.A. BEKRITSKY

2014

TABLE OF CONTENTS

FIGURE INDEX

TABLE INDEX

1 List of Abbreviations

ADDM	Autism and Developmental Disabilities Monitoring
AGRE	Autism Genetic Resource Exchange
AIB1	amplified in breast cancer 1, also known as NCOA3, or nuclear receptor coactivator 3
ALS	amyotrophic lateral sclerosis
APC	adenomatous polyposis coli
API	application programming interface
AR	androgen receptor
ASD	autism spectrum disorder
ATN1	atrophin 1
ATXN1/2/3/7	ataxin-1/2/3/7
AUC	area under the curve
BAM	BGZF-compressed SAM format
BAX	BCL2-associated X protein
BLM	Bloom syndrome gene, also known as RECQL3, or RECQ protein-like 3
bp	base pairs
BRCA1	breast cancer 1 gene
BRCA2	breast cancer 2 gene
BWT	Burrows-Wheeler transform
C9ORF72	chromosome 9 open reading frame 72
CACNA1A	calcium channel, voltage-dependent, P/Q type, alpha-1A subunit
CCDS	consensus coding sequence
CEPH	Centre d'Etude du Polymorphisme Humain
CHD8	chromodomain helicase DNA binding protein 8
cM	centimorgan
CNP	copy number polymorphism
CNV	copy number variation
CODIS	Combined DNA Index System
DM1	dystrophia myotonica 1
DM2	dystrophia myotonica 2
DMPK	dystrophia myotonica protein kinase
DRPLA	dentatorubral-pallidoluysian atrophy
DSB	double-strand break
DSM-5	Diagnostic and Statistical Manual of Mental Disorder (5th ed.)
EL	glutamate/leucine
EM	expectation maximization

EPHB2	ephrin receptor ephB2
FALS	familial associated lateral sclerosis
FBI	Federal Bureau of Investigation
FIFO	first-in, first-out
FMR1	fragile X mental retardation 1
FMRP	fragile X mental retardation protein
FPR	false positive rate
FSM	finite state machine
FTD	frontotemporal dementia
FXN	frataxin
GB	gigabyte
GWAS	genome-wide association study
HFPMM	High Fidelity Phusion Master Mix
HGDP	Human Genome Diversity Project
HNPCC	hereditary non-polyposis colorectal cancer
HOXD13	homeobox D13
HRS	Haw River syndrome
HTS	high throughput sequencing
HTT	huntingtin
IAN	Interactive Autism Network
IGFIIR	insulin-like growth factor II receptor
IRLBA	implicitly restarted Lanczos bidiagonalization algorithm
LOH	loss of heterozygosity
Mb	megabase
MJD	Machado-Joseph disease
MLE	maximum likelihood estimate
MLH1	mutL homolog 1
MMP	microsatellite mutator phenotype
MMR	mismatch repair
MSH2/3/6	mutS homolog 2/3/6
MTF	move-to-front
NCBI	National Center for Biotechnology Information
NDIS	National DNA Index System
NFS	network file system
NHEJ	non-homologous end joining
NTNG1	netrin G1
OPMD	oculopharyngeal muscular dystrophy
PABPN1	polyadenylate binding nuclear protein 1, also known as PABP2, or polyadenylate binding protein 2
PAGE	poly-acrylimide gel electrophoresis
PCA	principal component analysis
PCR	polymerase chain reaction
PMS1/2	postmeitotic segregation increased 1/2

RAD50	homolog of S. cerevisiae RAD50
RMSE	root-mean-square error
ROC	receiver operating characteristic
RUCDR	Rutgers University Cell and DNA Repository
RUNX2	runt-related transcription factor 2
SAM	Sequence alignment/map file format
SBMA	spinal and bulbar muscular atrophy
SCA1/2/3/6/7	spinocerebellar ataxia 1/2/3/6/7
SFARI	Simons Foundation Autism Research Initiative
SGE	Sun grid engine
SNP	single nucleotide polymorphism
SNV	single nucleotide variant
SSC	Simons simplex collection
SSR	simple sequence repeat
STR	simple tandem repeat
SVD	singular value decomposition
TB	terabyte
TCF4	transcription factor 4
TGFβRII	transforming growth factor-beta receptor, type II
TPR	true positive rate
TRF	Tandem Repeats Finder
VCF	variant call format
VNTR	variable number tandem repeat
WUGSC	Washington University Genome Sequencing Center
ZNF9	zinc finger protein 9, also known as CNBP1, or cellular retroviral nucleic acid-binding protein 1

2 Acknowledgments

No amount of space is sufficient to express the debt of gratitude I owe to the people who have supported me during my graduate research. The mentorship and friendship I have experienced at CSHL will remain with me long after I have moved on. I have had the good fortune of learning from and getting to know some truly brilliant people, who have a passion for knowledge and unending curiosity. I hope that the lessons and encounters of these past six years will push me to excel during the rest of my research career.

I am exceedingly grateful for Mike Wigler's mentorship and support. In the course of working on an incredibly challenging project, his insight has been essential to the successes I have had. Mike has never shied from demanding my best efforts. His expectations have helped me develop as a critical thinker and scientist in ways that I never expected when I joined his lab. As I realize how far I have come since I first arrived at CSHL, it is clear to me that I owe any development and knowledge I have gained to his patience and wisdom.

Mike Schatz has been an exceptionally kind and devoted mentor who has helped me develop my computer programming and scientific research skills since he first came to CSHL. Throughout the past four years, Mike has served as a sort of second research mentor, whose passion and skill have been essential to my development as a scientist. Mike is an inspiring researcher to work with, whose thirst for knowledge and drive to succeed are a constant source of inspiration. Mike Schatz's former postdoc, Giuseppe Narzisi, has also been an incredible friend and colleague.

Dan Levy has been a friend and mentor that I have been able to rely on since my earliest days in the Wigler lab. He has led by example so many times, and has helped me think through some of the most complex aspects of my graduate research. Whenever I have been uncertain of what path to take, Dan has served as my scientific conscience, helping me (usually) make the right choices. Many of the insights I have had over the course of my research have had their mettle tested in conversations with Dan, and they are always better for his input.

Ivan Iossifov has been an incredible guide to the SSC dataset and has been an extraordinary resource when thinking through some of the challenges of calling *de novo* mutations. Mike Ronemus has provided keen insight into the state of autism research. The depth and breadth of his knowledge have been essential in navigating the autism literature and the SSC dataset.

Being a member of the Wigler lab has been an amazing experience. The people in the lab are uniformly among the smartest, funniest, kindest people I have had the pleasure of working with. In particular, I would like to thank Jen Troge for her support, friendship, and bench skills. Her hard work is what made the validation sequencing for this project possible. I would also like to thank the technicians in the Wigler lab who have done the hard work of preparing the SSC sequencing data that has been essential to my thesis—Inessa Hakker, Beicong Ma, and Julie Rosenbaum. Peter Andrews has been an incredible resource as I have learned how to write software that runs well, runs quickly, and runs without crashing a computing cluster. His patience, knowledge, and creativity have made helped me develop my software writing and data processing skills almost painlessly.

I would like to thank my thesis committee—Mickey Atwal, Dick McCombie, Scott Powers, Mike Schatz, Raffaella Sordella, and Mike Wigler—for their support and insight throughout the course of my project. They have suggested many useful ideas and analyses over the last six years, which have consistently been to the benefit of my project. I am especially grateful for their feedback regarding the presentation of my research. This project has not always lent itself to intuitive or concise graphical representations, but thanks to their input, I hope that I have come a long way in presenting complex ideas in a relatively straightforward fashion. I would like to specifically thank Mickey Atwal for serving as my thesis chair and for his support and encouragement. I would also like to thank Dick McCombie for his support as my academic mentor. He has provided some essential tools to help me better present complex data.

I have had the privilege of making many good friends in my time at CSHL. I have developed especially close friendships with some of my WSBS classmates. I would also like to thank my fellow Wigler/Hicks lab graduate students, Timour Baslan and Talitha Forcier, for their warmth and friendship. I would also like to thank Swagatam Mukhopadhyay for being an exceptional friend and for being someone who can always be relied upon for thoughtful analysis and support.

My wife, Rikki Stern, has been my rock and my role model since I started at CSHL. She deserves special thanks for maintaining her patience while I have written my dissertation. Despite my crankiness, distractedness, and somewhat concerning caffeine addiction, she has remained a patient, thoughtful, and unconditionally loving spouse, even when I have done nothing to deserve it. She has been by my side through the most difficult parts of these past six years, and has been a constant voice of support and encouragement. No matter what the challenges were, her belief in me has been essential to whatever success I can claim in the course of my doctoral research. My wife has set a standard of success and ambition that I can only hope to emulate. She is fearless and motivated, and has never settled for second-best. No matter the demands life places on her, she never falters. I can only hope to be half the person she is.

I would like to thank my family for their constant encouragement. Their curiosity and support has carried me through some of the most challenging moments in these past years. I could not ask for a warmer, more welcoming group of people to call or visit when I need a break from work. Their love and support has been unconditional, and has helped sustain me during my time at the Watson School.

I would like to thank the staff of the Watson School, past and present, for all their help and guidance. Their friendliness and sound advice have helped me navigate the academic landscape of CSHL. Their warmth has enriched my time in graduate school.

3 Abstract

Microsatellites are a ubiquitous feature of eukaryotic genomes. Simply defined, they are consecutive repeats of a short DNA motif. Microsatellites are among the least stable DNA regions in many genomes, primarily due to their unique mutation mechanism. Microsatellites are a useful tool in genetic analyses, a source of phenotypic variation, and a contributor to the etiology of many severe human diseases.

Until recently, microsatellite genotyping methods have been unable to leverage the dramatic advances in DNA sequencing technology to enable genome-wide genotyping. While there are some existing methods that call microsatellite genotypes from high throughput sequencing data, none have sufficient accuracy to identify *de novo* mutations (Gymrek et al. 2012; Highnam et al. 2013). This is particularly important in analyzing the role of microsatellite mutations in human diseases that may be caused by *de novo* mutations.

We have developed the uSeq pipeline to call microsatellite genotypes in sequencing data from large study populations. This pipeline detects, aligns and filters reads with microsatellites and then assembles microsatellite profiles for each individual in the study. By considering the wealth of information available from the population at a single microsatellite locus, uSeq infers individual-, locus-, and allele-specific parameters that allow for highly accurate genotypes and specific identification of *de novo* mutations.

We have applied the uSeq pipeline to exome sequencing data from the Simons Simplex Collection of families with autism. *De novo* mutations are known to play a significant role in the incidence of autism and autism spectrum disorders, particularly in the sporadic cases enriched for in the SSC (Sebat et al. 2007; Zhao et al. 2007; Levy et al. 2011;

Iossifov et al. 2012; Neale et al. 2012; O'Roak et al. 2012b; Sanders et al. 2012). We are able to identify microsatellite *de novo* mutations in this population with >90% specificity, which is on par with traditional microsatellite genotyping methods such as capillary electrophoresis or Sanger sequencing. In analyzing these mutations, we observe that children with autism are more likely to have *de novo* microsatellite insertions in exons than their siblings (p = 0.03). Although not statistically significant, we also observe several frameshift mutations in children with autism, while we observe none in their siblings. We have developed a novel and highly accurate microsatellite genotyping method and have used it to demonstrate a potential role for *de novo* microsatellite mutations in autism incidence. It is reasonable to expect that microsatellite polymorphism and *de novo* mutations play a much larger role in human disease than what is currently known.

4 Introduction

4.1 Microsatellites

4.1.1 Microsatellite traits

4.1.1.1 Basic microsatellite characteristics

Several features define any microsatellite locus. Every microsatellite is composed of a short, repeated motif, which is described by its composition and length. In most genome-wide microsatellite studies, all motif compositions are considered, and only motif length is limited. Typically, any repeated tract with a motif length between 1 and 6 bp is considered to be a microsatellite (Ellegren 2004). Motifs can be combined into "equivalence classes". A microsatellite equivalence class is defined by all circular shifts of a microsatellite motif and its reverse complement. Within this framework, an AC motif is equivalent to CA, TG, and GT motifs.

In addition to its constituent motif, a microsatellite tract has three other characteristics. The repeat number is the number of times a motif is repeated within a microsatellite tract, which can be a non-integer. The tract length is the length of the microsatellite tract in base pairs. This is simply obtained by multiplying the tract's motif length and repeat number. Finally, the microsatellite tract is defined by its sequence identity. An uninterrupted microsatellite tract contains no interrupting SNPs or indels and therefore has perfect sequence identity. As interruptions are introduced, the sequence identity of a microsatellite tract decreases.

Microsatellites can also be defined relative to the reference genome. The chromosome, start coordinate, and stop coordinate provide the location of a microsatellite within the reference genome. Microsatellite genomic coordinates can be compared to coordinates for other relevant genomic features—such as exon positions—to understand microsatellite tracts in context. In this work, when we refer to multiple alleles at the same locus, we specify a microsatellite tract's location by its chromosome and start coordinate.

4.1.1.2 A brief history of microsatellites

One of the first descriptions of microsatellites was inferred from amino acid sequences soon after the genetic code was completely defined (Khorana et al. 1966). By screening trypsin-digested T4 phage-borne lysozyme in bacteria for pseudowild double frameshift mutations, codons for the frameshift mutants could be inferred. In many double mutants examined, frameshift mutations occurred in a region with repeated 1 and 2 bp motifs by the insertion or deletion of a single motif. Although the mutated DNA sequence was inferred from amino acid codons, the authors suggested that codons containing repeated sequences, i.e. microsatellites, could be mutational hotspots (Streisinger et al. 1966).

Microsatellites were also observed in density gradient centrifugation experiments in the early 1970s (see, for example, Skinner and Beattie 1973). In a typical centrifugation experiment, cesium salt density gradients resolved “satellite” bands distinct from the bulk of genomic DNA. As different cesium salt-based solutions were studied, scientists realized that the behavior of satellite DNA bands were dependent on DNA sequence. Careful analysis led to the observation that some density gradients were indicative of large DNA polymers composed of repeated short sequences (Skinner and Beattie 1973). These initials observations are also the source of the term “microsatellite”. Throughout the literature, microsatellites have also been referred to as SSRs or STRs. Along with minisatellites, they also comprise VNTR loci.

In the 1980s, DNA sequencing of several loci—including the region between human δ and β globin genes and the cardiac muscle actin gene—led to first direct observations of AC microsatellites (Miesfeld et al. 1981; Hamada et al. 1982a). Concurrently, Southern blot hybridizations using AC fragments revealed that these repeats were ubiquitous in the human genome as well as many other eukaryotic genomes (Miesfeld et al. 1981; Hamada and Kakunaga 1982; Hamada et al. 1982b; Jeang and Hayward 1983). The prevalence of AC microsatellites throughout multiple genomes was initially assumed to be due to evolutionary conservation, which suggested a special function for these repeats (Hamada et al. 1982b). Some early studies sought to ascribe particular importance to AC microsatellites since they could adopt a Z-DNA conformation *in vitro*, as opposed to the more common B-DNA conformation. The apparent conservation of AC microsatellites and their ability to adopt novel conformations *in vitro* led some scientists to believe that AC microsatellites—and perhaps all sequences capable of adopting the Z-DNA conformation—played a role in mutational hotspots, recombination, and gene regulation (Hamada and Kakunaga 1982; Hamada et al. 1982b). Subsequent studies demonstrated that AC microsatellites modulated gene expression, which reinforced these assumptions. Although CG microsatellites could also adopt the Z-DNA conformation *in vitro*, they did not seem to modulate gene expression—this discrepancy was attributed to structural differences favoring interactions with AC microsatellites (Hamada et al. 1984).

While some researchers were attempting to attribute specific functions to AC microsatellites, others were beginning to realize that many microsatellite motifs were ubiquitous in eukaryotic genomes. Southern blots using A, G, AG, AC, and ACG microsatellite probes showed that each of these tracts could be found in phylogenetically diverse eukaryotic organisms. Based on these observations, microsatellite tracts were proposed to be a result of stochastic indel and recombination mutations. These mechanisms implied no conserved evolutionary role for microsatellites—ubiquitous AC tracts had no direct influence on gene expression (Tautz and Renz 1984).

As the decade progressed, evidence accumulated that microsatellites were an abundant DNA class within eukaryotic genomes. Microsatellite genomic frequency was significantly higher throughout eukaryotic genomes than would be expected by chance. The same was not true in prokaryotic, organellar, or viral genomes (Tautz et al. 1986). The abundance of eukaryotic microsatellite tracts necessitated a mechanism that could explain their emergence. Possible mechanisms for microsatellite emergence are discussed in the following section.

Microsatellites were also observed to undergo frequent frameshift mutations, pinpointing microsatellites as mutational hotspots within eukaryotic genomes. Among the first direct observations of this instability was in early studies of spontaneous frameshift mutations in the *Escherichia coli lacI* gene, which were found to cluster in a CTGG microsatellite (Farabaugh et al. 1978). A more in depth study of frameshift mutation characteristics in *E. coli* nearly a decade later studied an in-frame 40 bp AC microsatellite inserted in the *lacZ* gene. Frameshift mutations were observed in almost 1.2% of *E. coli* plaques. Mutation frequency rose in *E. coli* strains without functional methyl-directed MMR proteins, but remained unchanged in *E. coli* strains with impaired recombination machinery (Levinson and Gutman 1987a).

Evidence that microsatellites might be unstable in humans came soon after initial observations of microsatellite instability in *E. coli*. Once PCR techniques using thermostable DNA polymerase had been developed, the same tools were almost immediately adapted for studies of microsatellite loci (Saiki et al. 1988). A study of 10 AC microsatellite loci within a population of unrelated individuals and multi-generational families revealed polymorphism at every locus assayed (Weber and May 1989). Another study of the intronic microsatellite in the cardiac muscle actin gene first described by Hamada and Kakunga produced similar results in a multi-generational family pedigree and in unrelated individuals (Litt and Luty 1989).

The publications of Litt and Luty and Weber and May provided two important breakthroughs—microsatellites were polymorphic in human populations and they could be easily isolated for genotyping using PCR. These findings led to an explosion in microsatellite research. Scientists began studying the factors influencing microsatellite stability and their applications to various genomic analyses. Microsatellites are currently the DNA sequence of choice for forensic science; are still used in genetic linkage studies; and have proven invaluable in understanding the genetic history of many species. Microsatellite polymorphisms can also have dramatic phenotypic effects. Microsatellite mutations have been linked to several neurodegenerative diseases and cancer, and have been shown to affect behavior in prairie voles and facial morphology in domesticated dogs. In the rest of this section, we discuss the many facets of microsatellite research that emerged from these initial studies.

4.1.1.3 The birth of a microsatellite locus

Although microsatellites are prevalent throughout eukaryotic genomes, there does not appear to be a single process by which they emerge. One theory of microsatellite emergence relies on the observation that non-repetitive sequences between nearby direct repeats tend to be preferentially deleted via replication slippage, which would then create a tandemly repeated sequence. This tandem repeat could then undergo further replication slippage, creating a microsatellite locus. This could account for the microsatellite emergence in regions under no selective constraint (Levinson and Gutman 1987b).

Based on observations of two separate microsatellites in the primate η-globin pseudogene, microsatellite emergence was also suggested to derive from substitutions that create adjacent copies of a microsatellite motif in previously non-repetitive sequence (Messier et al. 1996). For example, the initial sequence ATGTGTGT could undergo transition mutations at its first or fifth base, leading to the sequences **G**TGTGTGT or ATGT**A**TGT, respectively. These substitution events would be exceedingly rare, and only observable on an evolutionary timescale. Once a mutation has generated an unstable microsatellite tract, species sharing a common ancestor with the initial destabilization event will have microsatellite polymorphism that would be observable on much shorter timescales (Messier et al. 1996).

Microsatellites may also emerge via rare replication slippage or indel events in regions with very short tandem repeat tracts or even in the absence of any repeated sequence (Schlötterer and Tautz 1992; Zhu et al. 2000a). Several different groups have proposed this theory with various forms of supporting evidence. In its first iteration, slippage was proposed to occur as an initiating event at very short repeat tracts (Schlötterer and Tautz 1992).

Several years later, an analysis of mutations found in the Human Gene Mutation Database synthesized the microsatellite emergence theories of Messier, et al. and Schlötterer and Tautz. In analyzing the database, researchers observed that both substitutions and insertions created short microsatellite tracts capable of undergoing replication slippage. Between 3% and 16% of observed substitutions created microsatellites with 2 repeats—the specific rate depends on the motif length. Substitutions rarely extended any microsatellite tract to three repeats. In contrast, >70% of 2 – 4 bp insertions were copies of adjacent sequence, the majority of which were not already part of a microsatellite tract of any length. Although no specific mechanism is proposed, the authors suggested that tandem insertions of non-microsatellite sequence might be due to a slippage-like process. Although substitution-derived short microsatellite tracts are much more common, the vast majority of insertion mutations create short microsatellite tracts (Zhu et al. 2000b).

The contribution of SNPs, insertions, and replication slippage to microsatellite birth has also been demonstrated by a comparative analysis of microsatellite loci in human and chimpanzee reference genomes. In this study, both insertions copying adjacent sequence and substitutions creating adjacent repeats were observed as a means of microsatellite emergence. This analysis produces a similar observation to Zhu et al.—microsatellite loci resulting from duplicative insertions of non-microsatellite sequence are common, particularly for short motif lengths. Leclercq et al. also suggest NHEJ as a mechanism for indel slippage—Leclercq et al.'s term for insertions of non-microsatellite sequence. In this model, indel slippage occurs after a DSB in DNA sequence either through stable base mispairings that are subsequently filled in, or after a DSB is blunt-ended and ligated back together (Leclercq et al. 2010).

4.1.2 Microsatellite instability

4.1.2.1 Mechanisms of microsatellite instability

4.1.2.1.1 Unequal crossing over

Recombination was initially thought to play a role in microsatellite instability through a mechanism described as unequal crossing over (Smith 1974). In this process, microsatellite regions can pair with each other and recombine due to microhomologies within their tracts. If the chromosomes do not pair exactly, the resulting recombinant chromosomes will have an expanded microsatellite allele, and a correspondingly contracted microsatellite allele (Figure 4.1, page 90). Heterozygous SNPs near the recombination site could serve as a means of determining whether an unequal crossing over event occurred. At least one early paper characterizing microsatellite sequences throughout eukaryotic genomes suggested unequal crossing over as a mechanism that could explain some variation at microsatellite loci (Tautz and Renz 1984).

4.1.2.1.2 Replication slippage

Replication slippage was proposed as a mechanism that explained microsatellite instability in one of the first papers that considered frameshift mutations in coding sequence (Streisinger et al. 1966). Replication slippage products at AT microsatellite tracts had previously been observed under very specific conditions, but this behavior had yet to be established as a general feature of cellular DNA (Kornberg et al. 1964). Replication slippage can occur when DNA polymerase dissociates from a nascent and template strand during

DNA replication. Once the polymerase dissociates, the two strands are then free to dissociate from one another as well. The free strands can re-anneal to one another out-of-phase, since complementary bases in the microsatellite tract that are in the incorrect position still allow for correct realignment. If the nascent and template strands re-anneal out-of-phase, a bulge will be present in one of the strands. If the bulge is in the nascent strand, the nascent microsatellite tract will be expanded relative to the template microsatellite tract. If the bulge is in the template strand, the nascent microsatellite tract will be contracted relative to the template microsatellite tract. Once the strands re-anneal and DNA polymerase re-associates, replication continues as usual (Figure 4.2, page 91). Replication slippage can occur in DNA sequences of any length, since it only relies on local dissociation between strands (Schlötterer and Tautz 1992). Replication slippage has been frequently been proposed as the primary means of microsatellite instability (see, for example, Tautz and Renz 1984; Schlötterer and Tautz 1992; Eichler et al. 1994).

There are several intriguing characteristics of microsatellite mutation via replication slippage. Replication slippage is not a recombinant process, and can therefore occur any time a cell undergoes DNA replication. This is in stark contrast to unequal crossing, which would primarily occur during meiosis (Andersen and Sekelsky 2010). Replication slippage would not affect nearby heterozygous SNPs. Replication slippage also provides a mechanism for microsatellite mutation for haploid chromosomes—this is especially relevant for microsatellites in sex chromosomes in men. Finally, replication slippage can only progress through insertions or deletions of complete motifs.

4.1.2.1.3 Microsatellite instability is primarily due to replication slippage

Several lines of evidence support replication slippage as the primary driver of microsatellite instability. In bacteriophage-borne AC microsatellites in *E. coli*, observed replication slippage mutations are specific to the microsatellite tract; are only observed as insertions or deletions of complete motifs, usually only a single motif; and the frameshift frequency is tract length-dependent (Levinson and Gutman 1987a). These characteristics are consistent with replication slippage, but they do not rule out unequal crossing over as another driver of microsatellite instability. Replication slippage was also demonstrated to be a primary driver of microsatellite instability in studies of *in vitro* microsatellite extension. PCR amplification of initial short microsatellite primers led to their expansion into much longer species in the absence of any means of recombination (Schlötterer and Tautz 1992). The only possible mechanism for microsatellite mutation in this study was replication slippage.

Evidence ruling out unequal crossing over as a primary driver of microsatellite instability came from studies of microsatellite tracts in *Saccharomyces cerevisiae.* In *S. cerevisiae*, mitotic recombination around regions of microhomology rarely occurs—the rate had been previously measured to be ~10^{-10} for a 26 bp non-microsatellite sequence. The observed rate of microsatellite mutation was six orders of magnitude higher than this mitotic recombination rate (Henderson and Petes 1992). In addition, mitotic recombination in *S. cerevisiae* largely depends on the RAD52 protein. The rates of microsatellite instability were identical in *S. cerevisiae* strains with and without a functional RAD52 protein, indicating that microsatellite mutation is independent of a primary driver of mitotic recombination in *S. cerevisiae* (Henderson and Petes 1992; Wierdl et al. 1997).

In several studies evaluating the role of MMR proteins on replication slippage repair, the mutation rates of microsatellite loci in mitotically and meiotically reproducing *S. cerevisiae* were compared. Despite dramatically higher recombination rates during meiosis, microsatellite mutation rates were identical to those of mitotically reproducing *S. cerevisiae* (Strand et al. 1993; Wierdl et al. 1997). This suggests that unequal crossing over plays almost no role in microsatellite instability.

Studies of microsatellite mutation rates of the human Y chromosome both reinforce replication slippage as the driver of microsatellite instability and suggest that there is no role for unequal crossing over. Two independent studies compared Y chromosome microsatellite mutation rates to established autosomal microsatellite mutation rates. Both studies found that Y chromosome microsatellite mutation rates were indistinguishable from autosomal microsatellite mutation rates, despite the near-complete absence of recombination on the Y chromosome (Heyer et al. 1997; Kayser et al. 2000). This provides further support that microsatellite mutation is largely, if not entirely, independent from recombination or unequal crossing over.

Population genetic evidence also supports replication slippage over unequal crossing over as the driver of microsatellite instability. In a larger study on the origin of modern humans, phylogenies based on microsatellite polymorphisms were considered using two mutation models—the stepwise mutation model and the infinite alleles model. The stepwise mutation model assumes that each mutation at a microsatellite locus consists of the insertion or deletion of one motif; therefore, identical alleles need not be derived from a single common ancestor. This behavior is consistent with replication slippage. The infinite alleles model assumes that each allele is derived from a single common ancestor, and is consistent with unequal crossing over. In comparing the separation of continental population groups, it is apparent that the phylogenies assuming a stepwise mutation model are more consistent with phylogenies generated from other polymorphism data than phylogenies assuming an infinite alleles model for microsatellite mutation (Jorde et al. 1995).

Initial studies of microsatellite instability demonstrate a marked bias towards microsatellite mutations in the paternal germline. Since paternal germlines undergo many more cycles of cell division than maternal germlines, this observation is consistent with replication slippage as the major driver of microsatellite instability (Weber and Wong 1993). Additional observations of somatic instability at microsatellite loci support the absence of a role for unequal crossing over, since somatic instability is almost completely independent of recombination (see, for example, Chong et al. 1995; Salipante and Horwitz 2006). Finally, an exhaustive study of more than 2,400 microsatellite loci in over 85,000 Icelanders failed to show a correlation between microsatellite mutation rates and local recombination rates (Sun et al. 2012).

4.1.2.2 Replication slippage repair

4.1.2.2.1 MMR proteins

Replication slippage intermediates are frequently repaired by post-replicative MMR machinery before they can be propagated to subsequent generations. Efficient repair of frameshift intermediates in newly replicated DNA was first demonstrated in *E. coli* using bacteriophage-borne frameshift intermediates. The *E. coli* MMR system is methyl-directed, which ensures that MMR targets newly synthesized DNA. Frameshift and substitution mutation intermediates were repaired with equal efficiency; and frameshift repair is impaired in *E. coli* cells with defective MutL (Dohet et al. 1986). Subsequent analysis of AC microsatellite slippage in *E. coli* confirmed that microsatellite mutation rates increase dramatically in the absence of MutL and MutS, the primary proteins involved in MMR (Levinson and Gutman 1987a).

Homologs for *E. coli* MutL and MutS were subsequently identified in *S. cerevisiae*. Initial studies of MMR mutations in *S. cerevisiae* demonstrated a dramatic increase in plasmid and chromosomal microsatellite instability in the absence of the MutL homologs PMS1 and MLH1, as well as in the absence of the MutS homolog MSH2. Some of these same studies demonstrated no further increase in microsatellite instability in PMS1/MLH1 double mutants, indicating that they represent the same MMR pathway (Strand et al. 1993; Greene and Jinks-Robertson 1997; Sia et al. 1997b; Tran et al. 1997). Mutations to MSH2, MLH1, PMS1, and PMS2 increase the rate of slippage mutations and substitutions (Strand et al. 1995; Sia et al. 1997b). The same is not true for mutations to the MutS homolog MSH3—substitution rates are largely unchanged in MSH3 mutants while slippage rates increase by up to 40 – 50 fold (Strand et al. 1995; Greene and Jinks-Robertson 1997; Sia et al. 1997b). Mutants of a third MutS homolog—MSH6—increases the rate of replication slippage and substitution, although slippage mutations are not as prevalent as they are in other MMR mutants. However, MSH3/MSH6 double mutants reproduce the dramatically increased slippage rate observed in MSH2, PMS1, and MLH1 mutants. This rate is maintained in MSH2/MSH3/MSH6, MSH2/MSH6, and MSH2/MSH3 mutants, indicating that MSH3 and MSH6 are epistatic to MSH2 mutations. This also indicates that there are probably at least two MMR complexes. While both complexes appear to contain PMS1, MLH1, and MSH2, either complex only contains MSH3 or MSH6 (Johnson et al. 1996; Greene and Jinks-Robertson 1997; Sia et al. 1997b). Although MMR appears to maintain its preference for repairing newly synthesized DNA in eukaryotes, the mechanism by which this is accomplished is unclear (Sia et al. 1997a).

Many human MMR proteins were characterized after rampant microsatellite instability was observed in HNPCC and other colorectal cancers (Aaltonen et al. 1993; Ionov et al. 1993; Parsons et al. 1993; Thibodeau et al. 1993; Perucho et al. 1994). The MutS homolog hMSH2 was the first MMR protein characterized in humans (Leach et al. 1993; Fishel et al. 1994). The following year the MutL homologs hMLH1, hPMS1 and hPMS2 were identified in HNPCC patients, suggesting these genes were involved in MMR in humans (Nicolaides et al. 1994; Papadopoulos et al. 1994). The MutS homolog hMSH6 was identified in complex with hMSH2 after it was determined that both proteins were necessary to restore proper MMR function to MMR-deficient tumor cells (Drummond et al. 1995; Palombo et al. 1995). Finally, hMSH3 was identified in studies of endometrial cancers exhibiting microsatellite instability, but without mutations to other known MMR genes (Risinger et al. 1996).

4.1.2.2.2 MMR efficiency and sensitivity

MMR efficiency and sensitivity depends on the size of the loop generated in a frameshift intermediate. In an extensive study of microsatellites of varying motif lengths in *S. cerevisiae,* the behavior of the MSH3 and MSH6 MMR complexes were readily distinguished. In MSH6 mutant *S. cerevisiae* cells, replication slippage rates were up to 30 times higher than wild-type for 1 and 2 bp loops, while no change was observed for longer loops. In contrast, MSH3 mutant cells exhibited replication slippage rates up to 130 times higher than wild-type for loops up to 8 bp long (Sia et al. 1997b). By comparing the replication slippage rates of MSH2 mutant and wild-type *S. cerevisiae* colonies, MMR appears to correct more than 99.9% of all 1 bp loops. MMR becomes less effective as loop size increases—MMR efficiency is ~80% for loops of up to 13 bp, while no MMR-dependent correction is observed in loops larger than 16 bp (Sia et al. 1997b). Interestingly, MSH3 complexes appear to be more efficient at repairing deletions than insertions, which suggests that it favors template loops to loops in newly synthesized DNA (Strand et al. 1995; Johnson et al. 1996; Sia et al. 1997b).

While MMR correction efficiency is highly dependent on loop size, and therefore motif length, it appears to be independent of tract length. In studies of plasmid-borne AC microsatellites in *S. cerevisiae* with tract lengths ranging from 15 to 99 bp, overall correction efficiency is above 97%, although deletions are corrected more often than insertions (Wierdl et al. 1997). MMR efficiency appears to have some dependency on motif composition and sequence context. In a study of slippage events in wild-type *S. cerevisiae*, mutations occurred more frequently in microsatellites with C motifs than those with A motifs, and similarly sized microsatellites with identical 1 bp motifs mutated at distinctly different frequencies (Greene and Jinks-Robertson 1997).

4.1.2.2.3 DNA proofreading

A role for DNA proofreading in replication slippage repair has also been suggested. In *S. cerevisiae* strains with mutant DNA-Pol δ or DNA-Pol ε, replication slippage at short A microsatellite tracts increased between 3 – 300 fold relative to wild-type slippage rates. Since both DNA-Pol δ and DNA-Pol ε have 3’→5’ exonucleolytic proofreading activity, this would imply a role for DNA polymerase-dependent replication slippage repair. In DNA-Pol ε/MSH2 double mutants, A microsatellites with tract lengths of 4 or 5 bp were significantly less stable than either single mutant for either gene (Tran et al. 1997). As microsatellite tract lengths increased, the role of exonucleolytic proofreading rapidly diminishes. DNA polymerase mutations do not appear to have an impact on replication slippage rates for microsatellite tracts longer than 8 – 10 bp (Strand et al. 1993; Tran et al. 1997). In general, DNA proofreading appears to play a very limited role in replication slippage repair.

4.1.2.3 Factors influencing microsatellite instability

4.1.2.3.1 Motif length

Microsatellite mutations typically insert or delete one complete motif from a microsatellite tract, although mutations involving more motifs are possible (see, for example, Levinson and Gutman 1987a; Henderson and Petes 1992; Strand et al. 1993; Brinkmann et al. 1998). Microsatellite instability also has a well-characterized dependence on motif length. Early studies in plasmid-borne C and AC microsatellites in *S. cerevisiae* demonstrated that microsatellites with 1 bp motifs had a slightly elevated mutation rate relative to those with 2 bp motifs (Henderson and Petes 1992). That same year, *in vitro* microsatellite primer extension assays demonstrated that microsatellites with 2 bp motifs grew more rapidly than those with 3 bp motifs, implying that 2 bp motifs are less stable (Schlötterer and Tautz 1992). Evidence for the motif length dependence of microsatellite stability in chromosomal DNA came later. Studies of polymorphism at microsatellite loci grouped by motif length within human populations demonstrated that in general, the mutation rate for microsatellites with 2 bp motifs is highest, followed by those with 3 bp motifs, and then those with 4 bp motifs. Disease-causing loci with 3 bp motifs were found to be the least stable (Chakraborty et al. 1997). A study of microsatellite instability in microsatellites with 2 and 4 bp motifs in MMR deficient cell lines produced similar results (Lee et al. 1999). This observation was reinforced by studies of *de novo* microsatellite mutations in *Drosophila melanogaster*, which showed that microsatellites with 2 bp motifs mutated ~6 times more often than those 3 bp motifs, and ~8 times more often than those with 4 bp motifs (Schug et al. 1998). Recent studies examining several hundred to over a hundred thousand microsatellite loci in human populations make it clear that this behavior is universal—microsatellites are more stable as their motif length increases (Pemberton et al. 2009; Gymrek et al. 2012). This behavior was also demonstrated in comparisons of the human and chimpanzee genomes at microsatellite loci with motif lengths from 1 to 4 bp (Kelkar et al. 2008). Some studies have a shown a higher mutation rate in highly polymorphic 4 bp marker loci as compared to highly polymorphic 2 bp marker loci (Weber 1990; Sun et al. 2012).

4.1.2.3.2 Tract length and repeat count

Tract length and repeat count were among the earliest factors demonstrated to effect microsatellite instability. In some of the first observations of spontaneous frameshift mutations of short A microsatellite loci in *E. coli*, the mutation rate increased dramatically with each additional motif (Streisinger and Owen 1985). In another early study of microsatellite instability, 40 bp AC microsatellite loci in *E. coli* were shown to be more than twice as unstable as 22 bp AC microsatellite loci (Levinson and Gutman 1987a). Studies in *S. cerevisiae* of chromosomally integrated microsatellites with 1 bp motifs also demonstrated decreased stability even for small changes in tract length—A microsatellites exhibit a 30-fold destabilization when they expand from 9 bp to 15 bp (Tran et al. 1997). This observation is reinforced in *S. cerevisiae* studies of plasmid-borne AC microsatellites with tract lengths ranging from 15 to 99 bp—in wild-type cells, mutation rates increased by more than two orders of magnitude within range of tract lengths studied (Wierdl et al. 1997). A separate study of microsatellites in the *S. cerevisiae* genome suggested that microsatellite instability was dependent on a microsatellite tract's length, but not its repeat count (Pupko and Graur 1999).

Studies of microsatellite variation in natural populations also demonstrate the relationship between repeat count or tract length and microsatellite instability. Early analyses of polymorphism at AC microsatellite loci in human populations demonstrated that loci with higher average repeat counts tended to have more alleles, indicating increased instability (Weber 1990). Later studies of microsatellite instability in human populations showed a positive correlation between the mutation rate of a microsatellite locus and its geometric mean, arithmetic mean, or maximum uninterrupted repeat count (Brinkmann et al. 1998; Pemberton et al. 2009). The same phenomena has been demonstrated in *D. melanogaster* populations—microsatellite loci with high maximum repeat counts have greater variance in their repeat length distributions (Goldstein and Clark 1995; Schug et al. 1998; Bachtrog et al. 2000). However, some studies were not able to reproduce this effect in worldwide collections of *D. melanogaster* or in human Y chromosome father-son pairs (Schlötterer et al. 1997; Kayser et al. 2000).

Phylogenetic studies of microsatellite loci in human populations also demonstrate a strong relationship between allele count or tract length and microsatellite tract instability (Jin et al. 1996). Studies estimating microsatellite polymorphism or mutation rates from large populations or sets of microsatellite loci demonstrate that longer microsatellites are more polymorphic and have higher mutation rates (Gymrek et al. 2012; Sun et al. 2012). Longer microsatellite tract lengths and higher repeat counts also clearly destabilize microsatellite loci in comparisons of human and chimpanzee genomes (Kelkar et al. 2008). The relationship between microsatellite tract length and variability has even been demonstrated in *Arabidopsis thaliana* (Cao et al. 2014).

There is some debate about whether replication slippage can only occur once a microsatellite locus has exceeded a threshold repeat count or tract length. Initial studies on microsatellite emergence suggested that the transition from non-repetitive sequence to microsatellite sequence occurs once substitutions introduce a minimum number of consecutive repeats (Messier et al. 1996). Subsequent studies of the over-representation of repeat tracts for each equivalence class in the *S. cerevisiae* genome suggested a minimum tract length of 8 bp (Rose and Falush 1998). A study of microsatellite polymorphism within the genomes of 179 individuals from the 1000 Genomes Project indicated that extreme instability at microsatellite loci was dependent on repeat count. A minimum of 9 repeats was necessary for replication slippage to occur for microsatellites with 1 bp motifs, 5 repeats for those with 2 bp motifs, and 4 repeats for those with motif lengths of 3 and 4 bp (Ananda et al. 2013).

Other studies suggest that there is no minimum threshold for replication slippage at microsatellite loci. In a study paralleling Rose and Falush, analysis of all repeats in the *S. cerevisiae* genome grouped by motif length indicated that microsatellites were over-represented for all repeat counts greater than two. As microsatellite repeat counts increase, the over-representation simply becomes more pronounced (Pupko and Graur 1999). Studies suggesting a base rate of replication slippage independent of threshold repeat counts would also support this perspective (Zhu et al. 2000b; Leclercq et al. 2010). The absence of a threshold for replication slippage at microsatellite loci is also consistent with studies of short microsatellite repeats in *S. cerevisiae*, as well as observations of frameshift mutations at short microsatellites with 1 bp motifs within the APC gene of cancer patients with and without MMP (Huang et al. 1996; Greene and Jinks-Robertson 1997).

4.1.2.3.3 Motif composition

Although microsatellite motif composition plays a role in microsatellite instability, the exact nature of this relationship is unclear. Early studies of microsatellite tract synthesis *in vitro* demonstrated markedly increased instability for AG microsatellites as compared to AC microsatellites, despite their equivalent GC content. The same is also true for microsatellites with 3 bp motifs. Although motifs with higher GC content tend to be more stable, this is not consistent for all motifs—AGG microsatellites are among the least stable 3 bp motifs (Schlötterer and Tautz 1992). Mutation rate estimates for microsatellites with 2 bp motifs in global populations of *D. melanogaster* conflict with the findings of Schlötterer and Tautz—AC microsatellites are the least stable, followed by AG microsatellites, then AT microsatellites. This observation confirms that the GC content of a microsatellite motif does not appear to affect microsatellite tract instability (Bachtrog et al. 2000). In contrast, studies using HTS data from 13 strains of *A. thaliana* suggested that GC content was inversely correlated with microsatellite stability (Cao et al. 2014).

The effect of motif composition on microsatellite instability is no clearer in studies of microsatellites in the human genome. In comparisons of microsatellite loci in the human and chimpanzee reference genomes, A repeats are less stable than C repeats. In contrast to what is observed in *D. melanogaster*, AT microsatellites are the least stable of all 2 bp motifs in the human genome. This study also demonstrates variability in microsatellite instability among 3 and 4 bp motifs, but the relationship is even murkier than it is for 1 and 2 bp motifs. Overall, the GC content of a motif did not appear to affect microsatellite instability (Kelkar et al. 2008). However, in another study of microsatellite loci in ~1,000 human DNA samples, GC content was found to affect microsatellite instability for 4 bp motifs—increased heterozygosity at microsatellite loci was correlated with higher GC content in microsatellite motifs (Pemberton et al. 2009). In a study of microsatellite variation at loci with motif lengths of at least 2 bp using HTS data, most variation occurred in microsatellites with AC or AT motifs (McIver et al. 2011).

The relationship between motif composition and microsatellite instability may also depend on repeat count. Although A and C microsatellites have different levels of instability for shorter tract lengths, they are nearly equally unstable at loci with more than 17 repeats. In addition, while AC microsatellites are less stable than AG microsatellites for repeat counts less than 12, they are more stable for repeat counts greater than 15 (Kelkar et al. 2008).

Although it is clear motif composition affects microsatellite instability, there is no consensus on how that effect is realized. Studies are inconsistent with regards to the effect motif GC content plays on microsatellite stability, if indeed it plays a role at all (Bachtrog et al. 2000; Kelkar et al. 2008; Pemberton et al. 2009; Cao et al. 2014). Studies of microsatellite loci in the same organism do not report consistent relative stabilities for motifs with the same length but different compositions. Among microsatellite with 2 bp motifs in humans, one study suggests that AT motifs are the least stable, while another observes more instability at AC motifs (Kelkar et al. 2008; McIver et al. 2011).

At least two theories attempt to explain the relationship between motif composition and microsatellite instability. These theories are not mutually exclusive. The first theory suggests that MMR efficiency may vary for different motifs in a species-dependent manner. This motif-dependent MMR efficiency may itself be due to variation in GC content in different species (Bachtrog et al. 2000). This phenomenon may also be due to the stability of frameshift intermediates of different motifs. Since AT loops may be more stable than AC or AG loops, AT microsatellites may be less stable (Kelkar et al. 2008).

4.1.2.3.4 Sequence identity

Microsatellite instability is highly dependent on sequence identity. In one of the first studies of microsatellite instability in humans, interrupted microsatellite loci were significantly more stable than uninterrupted microsatellite loci. Moreover, the best predictor of instability at an interrupted microsatellite locus was the length of its longest uninterrupted microsatellite tract (Weber 1990). In an analysis of *de novo* microsatellite mutations in parent/child trios, mutations in interrupted microsatellite alleles were less frequent than mutations at uninterrupted alleles. Akin to the results from Weber, the geometric mean of the longest uninterrupted repeat count at a microsatellite locus was the best predictor of its instability (Brinkmann et al. 1998). In subsequent analyses of *de novo* microsatellite mutations in father/son pairs on the human Y chromosome, the only mutations observed were in uninterrupted microsatellite tracts with at least 11 repeats (Kayser et al. 2000). Two later studies evaluating instability at thousands of microsatellite loci continued to demonstrate a marked difference in the stability of interrupted and uninterrupted microsatellites. Neither study sought to determine if instability was correlated with the tract length or repeat count of the longest uninterrupted portion of an interrupted microsatellite tract (McIver et al. 2011; Sun et al. 2012).

The dependence of microsatellite instability on sequence identity is reinforced in phylogenetic studies of microsatellite loci in the human genome. One study described two distinct subpopulations at a single microsatellite locus with drastically different levels of polymorphism. One subpopulation had an uninterrupted AC repeat and was highly polymorphic, while the other subpopulation had an interrupted repeat and was nearly monomorphic. This discrepancy was driven primarily by the sequence identity of the AC repeat, and not population genetic factors such as lineage age or a recent population expansion (Jin et al. 1996).

Studies of microsatellite loci known to play a role in disease incidence also demonstrate the dramatic increase in microsatellite stability at interrupted microsatellite alleles. In several studies of predisposing alleles in SCA1, ~98% of unaffected individuals possessed a stabilizing CAT variant motif in the CAG microsatellite tract whose expansion leads to the disorder, as did all unexpanded alleles in affected individuals (Chung et al. 1993; Jodice et al. 1994; Chong et al. 1995). All pathogenic expansions in individuals with SCA1 were at uninterrupted CAG microsatellites (Chung et al. 1993). A similar study of AGG variant repeats in the FMR1 CGG microsatellite tract whose expansion leads to Fragile X indicates that all pathogenic alleles have lost at least one variant repeat. Similar to Weber and Brinkmann et al., allele instability was related to the length of the uninterrupted CGG tract within the microsatellite locus (Eichler et al. 1994).

The relationship of microsatellite instability to sequence identity remains true in non-human species. In two studies of the role of variant repeats on AC microsatellite stability in *S. cerevisiae*, interrupted microsatellites are significantly more stable than uninterrupted repeats (Heale and Petes 1995; Petes et al. 1997). A population genetic analysis of microsatellite loci in *D. melanogaster* also demonstrated increased stability in interrupted microsatellite tracts. The same study also found that the longest uninterrupted microsatellite tract is more stable than would be expected based on the uninterrupted repeat count

(Goldstein and Clark 1995). Finally, an evaluation of HTS sequencing data from multiple *A. thaliana* strains demonstrates a strong correlation between sequence identity and microsatellite variability (Cao et al. 2014).

4.1.2.3.5 Insertions and deletions in microsatellite tracts

Several studies have reported conflicting results regarding a bias towards insertions or deletions at microsatellite loci. In a study of spontaneous frameshift mutations in short A microsatellite tracts in *E. coli*, deletions were 2 to 4 times as common as insertions (Streisinger and Owen 1985). In early studies of microsatellite instability in a bacteriophage-borne 40 bp AC microsatellite tract in *E. coli*, deletions were also 3 times as common as insertions (Levinson and Gutman 1987a). Later studies in *S. cerevisiae* demonstrated higher rates of deletions in a 10 bp A microsatellite tract than insertions in a longer 12 bp A microsatellite tract, suggesting a bias towards deletions at microsatellite loci (Tran et al. 1997).

A deletion bias—particularly at short microsatellite tracts—is inconsistent with the observed abundance of microsatellite tracts throughout many eukaryotic genomes. Fortunately, there is ample evidence of insertions being favored at microsatellite loci. In an early study of spontaneous frameshift mutations in a 12 bp CTGG tract in *E. coli*, insertions are more than 4 times as common as deletions (Farabaugh et al. 1978). In studies of plasmid-borne 33 bp AC microsatellites in *S. cerevisiae*, small insertions were significantly more common than small deletions. The same behavior is observed in *S. cerevisiae* colonies with mutant DNA-Pol δ and DNA-Pol ε (Strand et al. 1993; Petes et al. 1997). Further characterization of insertion biases in *S. cerevisiae* demonstrated a significant small insertion bias in wild-type strains, but a bias towards small deletions in strains with mutant MSH3

(Strand et al. 1995). Another study of tract instability in *S. cerevisiae* did not find any deletion bias in MSH3 mutant strains (Petes et al. 1997). A study of *de novo* microsatellite mutations in father/son pairs at human Y chromosomal loci also describes an insertion bias at several microsatellite loci with 2 and 4 bp motifs (Kayser et al. 2000).

Several other studies suggest that there is no bias towards insertions or deletions at human microsatellite loci. In a study of somatic mutations at 24 microsatellite loci in patients with colon adenocarcinoma, insertions and deletions occurred with the same frequency (Di Rienzo et al. 1998). In MMR defective human cells, insertions and deletions occur with equivalent frequencies. This result suggests that any bias in mutations rates observed—at least at human microsatellite loci—could be due to biased MMR efficiency in repairing template and nascent frameshift intermediates (Lee et al. 1999). In a study of *de novo* microsatellite mutations at 4 bp microsatellite tracts, there is similarly no observation of a bias in per-generation microsatellite mutation rates (Sajantila et al. 1999).

In an attempt to reconcile these conflicting observations, two studies examined microsatellite insertion/deletion biases at microsatellite tracts of varying lengths. In *S. cerevisiae*, a study of plasmid-borne AC microsatellite loci with lengths from 15 – 99 bp demonstrated that while the rate of single motif insertions rises as tract length increases, the rate of single motif deletions appears to decrease dramatically as tract length increases. In addition, large deletions are significantly more frequent for longer tract lengths (Wierdl et al. 1997). In a study of over 100 4 bp microsatellites in human families, deletions appeared to increase exponentially as tract length increased, while the rate of insertions appeared to be independent of tract length (Xu et al. 2000). Both of these studies suggest that reported insertion/deletion biases might be related to the balance of insertions and deletions specific to an observed tract length or subset of tract lengths, rather than an indication of a universal insertion or deletion bias.

4.1.2.3.6 Other factors affecting microsatellite instability

Microsatellite orientation could play a role in microsatellite instability if leading or lagging strand DNA replication were more prone to slippage or produced frameshift intermediates that were poorly recognized by MMR proteins. Plasmid-borne microsatellites in *S. cerevisiae* allow control of microsatellite tract orientation relative to the origin of replication. In two separate studies, instability at 2 bp microsatellite tracts did not appear to depend on the tract orientation relative to the origin of replication (Henderson and Petes 1992; Wierdl et al. 1997). A study of CAG microsatellites in *S. cerevisiae* suggested that lagging microsatellites near an origin of replication are unstable compared to their leading strand counterparts in a motif-dependent manner (Maurer et al. 1996). A similar study in CTG microsatellite tracts in *E. coli* demonstrated that insertions preferentially occurred when the microsatellite was in the leading strand, while deletions were more common if it was in the lagging strand (Kang et al. 1995). There does not appear to be any reported orientation-dependent effect on instability at human microsatellite loci.

Although microsatellites with similar properties can have very different mutation rates within a genome, the effects of surrounding sequence context on microsatellite stability are not well defined. An initial study of American alligator microsatellite loci suggested that high local GC content reduced allelic diversity at microsatellite loci (Glenn et al. 1996). However, an analysis of microsatellite loci in *D. melanogaster* showed no correlation between microsatellite stability and local GC content (Bachtrog et al. 2000). An extensive analysis of microsatellite loci in HGDP-CEPH individuals also failed to demonstrate any effect of surrounding GC content on microsatellite stability (Pemberton et al. 2009). GC content appeared to have a weak effect on microsatellite stability in comparisons of human and chimpanzee genomes (Kelkar et al. 2008). Similarly, an analysis of HTS data from 13 *A. thaliana* strains showed a significant correlation between surrounding GC content and microsatellite instability (Cao et al. 2014). Variability in microsatellite stability due to surrounding sequence is largely unexplained, although some studies attribute a role to flanking GC content.

Transcriptional activity may destabilize microsatellite loci. In a study of a chromosomal AC microsatellite adjacent to the GAL1-10 promoter in *S. cerevisiae*, instability increased in experimental conditions that favored high levels of transcription (Wierdl et al. 1996). Similar behavior is observed for plasmid-borne CAG microsatellites in *E. coli*, which seems to determine instability in concert with microsatellite orientation (Mochmann and Wells 2004). In a comparison of microsatellite loci in the human and chimpanzee genomes, instability does not appear to differ between intergenic and intronic microsatellite loci, although this might be due to the active transcription of intergenic portions of the genome (Kelkar et al. 2008).

4.1.2.3.7 Microsatellite instability summary

Although microsatellites are known to be highly unstable, the effects of extrinsic and intrinsic microsatellite locus characteristics on stability are inconsistently defined. It is clear that short microsatellite motifs are less stable. Many studies also suggest that longer microsatellite alleles are less stable as well. This observation is reinforced by several lines of evidence, including population genetic studies and direct observations of replication slippage in model organisms. It is also very clear that sequence identity affects microsatellite stability. Observations in human disease, studies of variant repeats in model organisms, and population genetic studies all suggest that uninterrupted microsatellite loci are significantly less stable than their interrupted counterparts. Moreover, many of these studies suggest that instability observed at interrupted microsatellite loci is highly correlated with various measures describing the length of the uninterrupted tract contained within the locus.

Other factors affect microsatellite instability in unclear ways. While it is clear that microsatellite motif composition has a significant effect on microsatellite stability, studies report contradictory results. Some of these contradictions may result from organism-specific microsatellite instability biases, but even within the same organism, conflicting results have been reported. Transcription, flanking GC content, and direction of replication may also play a role in microsatellite instability. Finally, insertions and deletions may have length-dependent biases as well.

All of this suggests that a significant proportion of microsatellite instability is unexplained, and that previous studies have not produced consistent results regarding various factors known to affect microsatellite stability. It might not be wise to generalize the same microsatellite mutation rates to similar microsatellite loci. Microsatellites appear to have very different rates of instability, even when considering similar loci. Until a better understanding of the interplay of various factors affecting microsatellite stability exists, it may be best to consider microsatellite mutation rates independently for each locus.

4.1.2.4 Literature-estimated human microsatellite mutation rates

Microsatellite mutation rates in humans have been estimated multiple times from different marker loci. In the earliest direct study of *de novo* microsatellite mutation at 28 microsatellite loci, the average mutation was determined to be 1.2×10^{-3} mutations per locus per gamete per generation, while individual mutation rates for each locus ranged from 0 to 8.0×10^{-3}. Evaluated by motif length, the mutation rate estimates were 2.1×10^{-3} for 4 bp motifs and 5.6×10^{-4} for 2 bp motifs (Weber and Wong 1993). These estimated mutation rate estimates are in agreement with estimates from two separate studies based on chromosome Y microsatellite loci, a study of autosomal microsatellite loci, and the average mutation rate of microsatellite loci from the first Généthon microsatellite linkage map (Weissenbach et al. 1992; Heyer et al. 1997; Brinkmann et al. 1998; Kayser et al. 2000). A later study demonstrated that two of the loci analyzed by Weber and Wong accounted for more than half of all 4 bp microsatellite mutations. If these loci were excluded, the average microsatellite mutation rate drops to 8.5×10^{-4} per locus per gamete per generation and 2 bp motifs are nearly twice as unstable as 4 bp motifs (Chakraborty et al. 1997). A study of microsatellite *de novo* mutations at 5 loci determined mutation rates ranging from 1.7×10^{-4} to $< 3.3 \times 10^{-3}$ per locus per gamete per generation, with an average mutation rate of 6.7×10^{-4} (Sajantila et al. 1999). A study of thousands of microsatellite loci in a large Icelandic cohort estimated mutation rates of 1.0×10^{-3} per locus per generation for 4 bp motifs and 2.7×10^{-4} for 2 bp motifs (Sun et al. 2012). All of these studies use marker microsatellite loci that are known to be highly polymorphic, so these estimates might be higher than mutation rates for non-marker microsatellite loci.

Several lines of evidence also demonstrate a clear bias for paternal transmission of *de novo* mutations in humans. In studies of SCA1-affected families, expanded pathogenic CAG alleles were inherited from the father significantly more often than the mother, particularly in cases of juvenile-onset disease (Chung et al. 1993; Orr et al. 1993). Analyses of *de novo* microsatellite mutations in pedigrees and families demonstrated that mutation events occur 3

– 5 times more often in the paternal germline as they do in the maternal germline (Weber and Wong 1993; Brinkmann et al. 1998; Henke and Henke 1999; Sajantila et al. 1999; Sun et al. 2012). Moreover, *de novo* microsatellite mutations derived from paternal germlines correlate strongly with paternal age—in one study, the germline microsatellite mutation frequency doubles in fathers from age 20 to 58 (Brinkmann et al. 1998; Sun et al. 2012).

4.1.2.5 Operationally defining microsatellites

We define microsatellites by their ability to undergo replication slippage. The thresholds we define are motivated by some of the better defined characteristics of microsatellite instability described above. We allow for any motif composition for microsatellite motif lengths between 1 and 6 bp. Even though slippage seems to occur at microsatellites of any repeat count or length, all studies agree that the rate of replication slippage increases proportionally with microsatellite tract length. We attempt to capture the greatest amount of microsatellite polymorphism in our study. Therefore, we require a microsatellite tract to have a repeat count of at least 3, and a tract length of at least 8 bp. These thresholds are in line with previously described minimum tract lengths and repeat counts for replication slippage at microsatellite loci (Rose and Falush 1998; Ananda et al. 2013). We also limit our study to uninterrupted microsatellite loci. Microsatellites could also be defined statistically, by considering whether adjacent, approximate copies of a repeat are closer together than might otherwise be expected (Benson 1999).

4.1.3 Microsatellite distribution in the human genome

Several studies have sought to characterize the microsatellite landscape of the reference human genome. In an exhaustive study of uninterrupted microsatellites with motif lengths from 1 to 6 bp and tract lengths of at least 12 bp in build 29 of the human reference genome, most chromosomes had a microsatellite density of ~12,000 to ~14,000 bp/Mb. Microsatellites with 1 bp motifs were uniformly distributed independent of genomic context, and exhibited a marked preference for A motifs than C motifs. Microsatellites with 2 bp motifs are more common in introns and intergenic regions than they are in exons. AC and AT motifs are the predominant 2 bp motifs, and CG motifs are exceedingly rare. Microsatellites with 3 bp motifs are more common in exons than they are in introns and intergenic regions. AAT and AAC motifs are the predominant 3 bp motifs. AAG and AGG motifs occur less frequently, but distinctly more often than ACC, AGC, or ATG motifs. There are very few CCG, ACG, or ACT motifs anywhere in the genome. 4 and 5 bp motifs are more common in introns and intergenic regions than they are in exons. AAAT, AAAG, AAAC, and AACC motifs are the predominant 4 bp motifs, and several other 4 bp motifs rarely occur in the genome. The most common 5 bp motifs are AAAAT and AAAAC. 6 bp motifs are more common in exons than they are in introns or intergenic regions, and the most common motifs are AAAAAT, AAAAAC, and AAAAAG (Subramanian et al. 2003).

Another study of microsatellites in build 36.1 of the human reference genome that used TRF to identify tract lengths of at least 12 bp and 90% identity found ~1.2 million total microsatellite loci. After filtering microsatellite loci in retrotransposons and those with 1 bp motifs, 376,685 loci remained. Of those loci, only 1.0% are in exons, 7.4% are in UTRs, 33% are in introns, and the remainder are in intergenic regions of the genome. Microsatellites are more commonly found in 5’ UTRs as compared to 3’ UTRs (McIver et al. 2011). Although the total length of the 5’ and 3’ UTRs in build 36.1 of the human reference genome is not reported, in build GRCh37 there are roughly 3 times as many bases in the 3’ UTR than there are in the 5’ UTR. This would imply a significant enrichment of microsatellites in 5’ UTR as compared to the 3’ UTR. A further discussion on the microsatellite distribution in GRCh37 as part of this project can be found in section 6.1.

4.1.4 Microsatellite relevance

4.1.4.1 Microsatellite diseases

4.1.4.1.1 Trinucleotide microsatellite expansion diseases

4.1.4.1.1.1 Fragile X Syndrome

The X-linked Fragile X syndrome was the first disease linked to a microsatellite expansion. The only portion of the region linked to Fragile X that varied within an initial study population was the length of a CGG microsatellite in the 5’ UTR of the FMR1 gene. In normal individuals, the microsatellite appeared to have repeat counts ranging from 15 to 65, while the microsatellite could not be accurately characterized in affected individuals due to its instability and length (Kremer et al. 1991). CGG expansions upstream of the FMR1 gene lead to methylation and transcriptional repression of the gene product in most affected individuals (Pieretti et al. 1991). Later studies established that the disease risk for different haplotypes depended on the presence of at least two stabilizing AGG variant repeats within the CGG microsatellite tract. In all cases, CGG microsatellite expansion was preceded by the loss of one or more variant repeats. Stabilizing variant repeats also offered an explanation of a previously described “grey zone” of instability, where CGG microsatellites with identical tract lengths appeared to be stable in some families and unstable in others (Eichler et al. 1994).

4.1.4.1.1.2 Fragile X and genetic anticipation

Genetic anticipation is a characteristic shared by fragile X and many other microsatellite-linked disorders: microsatellite alleles expand from one generation to the next, which leads to earlier disease onset and more severe symptoms. In fragile X, this was first studied in families where fathers with a fragile X pre-mutation allele appeared to be normal, as did their daughters, but their grandsons had the disease. The repeat count in the pre-mutation allele explains this phenomenon—since the CGG repeat becomes less stable as it expands, successive generations have increased risk of Fragile X incidence after a threshold length is exceeded (Fu et al. 1991).

4.1.4.1.1.3 SBMA

Microsatellite expansions also cause the X-linked disorder SBMA, alternatively known as Kennedy Disease. Investigations of a disease candidate region initially defined through linkage analysis revealed a CAG repeat in the first coding exon of the AR gene. In unrelated individuals diagnosed with SBMA from diverse ethnic backgrounds, the CAG repeat count varied from 40 to 52; while in 263 unaffected controls, including family members of affected individuals, the CAG repeat count varied from 17 to 26 (La Spada et al. 1991). This indicated that the expanded CAG allele segregates with SBMA incidence, and is therefore responsible for the disease phenotype.

4.1.4.1.1.4 Myotonic dystrophy

Myotonic dystrophy is caused by mutations to two separate genes, both of which are microsatellite expansions. The first form of the disease, DM1, is caused by an expansion of a CTG repeat in the 3' UTR of the DMPK gene (Brook et al. 1992; Buxton et al. 1992; Fu et al. 1992; Mahadevan et al. 1992; Bowcock et al. 1994). One reports link this mutation to

~98% of all cases of myotonic dystrophy (Fu et al. 1992). Both DM1 age-of-onset and severity are correlated with repeat length, which are hallmarks of genetic anticipation (Brook et al. 1992; Buxton et al. 1992; Fu et al. 1992). In normal individuals, the CTG repeat count ranges from 5 to 30, while in individuals with DM1, the repeat count is at least 50, but can exceed 1,000 in severely affected individuals (Brook et al. 1992; Fu et al. 1992; Mahadevan et al. 1992). The expanded allele segregates perfectly with disease incidence in unrelated DM1 patients, and is never found in unaffected individuals (Buxton et al. 1992; Fu et al. 1992; Mahadevan et al. 1992). In addition, the CTG repeat exhibits somatic instability in DM1 individuals and is transcribed (Brook et al. 1992; Fu et al. 1992). Transcription of the expanded allele suggests that DM1 is caused by a neofunctionalization of DMPK.

Nearly a decade later, the mutation responsible for DM2 was identified as a CCTG microsatellite expansion in the first intron of the ZNF9 gene. Although not a trinucleotide microsatellite disorder, DM2 is included here since it is related to DM1. In normal individuals, the CCTG repeat count is ~26, and is stabilized by two variant repeats. Expanded alleles had at least 75 to >11,000 repeats, although it is unclear if they also contain variant repeats. Unlike DM1, no correlation was observed between age-of-onset and repeat count, and there is no indication of genetic anticipation. Like the DMPK CTG microsatellite, the ZNF9 CCTG microsatellite is somatically unstable, which could complicate the interpretation of the variant repeat status of affected individuals and age-of-onset correlation. The expanded allele perfectly segregates with the DM2 phenotype (Liquori et al. 2001).

4.1.4.1.1.5 Huntington's Disease

After DM1, Huntington's disease was the next disease that was linked to microsatellite expansion. In normal individuals, a CAG microsatellite in the first exon of the

HTT gene has a repeat count between 11 and 34, while in affected individuals, repeat counts were at least 42. The expanded repeat always segregated with Huntington's disease incidence in patients from various ethnic backgrounds. Longer alleles are correlated with earlier onset of disease symptoms, which indicates that Huntington's disease exhibits classic signs of genetic anticipation. Huntington's disease is autosomal dominant and the mutant allele is expressed, which implies that the microsatellite expansion causes a neofunctionalization of the HTT gene (Macdonald et al. 1993).

4.1.4.1.1.6 Spinocerebellar ataxias

Many spinocerebellar ataxias are caused by microsatellite repeat expansions. SCA1 was the first spinocerebellar ataxia attributed to a microsatellite expansion. In normal individuals, a CAG microsatellite in the second coding exon of ATXN1 has a repeat count ranging from 6 to 39. Individuals with SCA1 have repeat counts from 41 to 81 (Chung et al. 1993; Orr et al. 1993; Banfi et al. 1994). SCA1 has at least one classic sign of genetic anticipation—longer alleles lead to early disease age-of-onset. Expanded alleles were most likely to be paternally transmitted, particularly in juvenile onset SCA1 (Chung et al. 1993; Orr et al. 1993). Two separate studies identified variant motifs that stabilized the CAG microsatellite in most normal individuals and in the normal alleles of SCA1 patients (Chung, et al. 1993; Chong, et al. 1995). 98% of normal individuals had at least one variant motif, while the remaining 2% had very short repeats (Chung et al. 1993). Every individual affected with SCA1 had an uninterrupted repeat (Chung et al. 1993; Chong et al. 1995). Disease incidence segregates with the expanded allele, and all affected individuals in one study were heterozygous for the expanded allele, indicating autosomal dominant behavior (Orr et al. 1993). The expanded SCA1 allele is expressed, which implies that a novel function is gained by ATXN1 in SCA1 patients (Orr et al. 1993; Banfi et al. 1994; Chong et al. 1995). Finally, uninterrupted SCA1 alleles exhibit significant somatic instability (Chong et al. 1995).

Many spinocerebellar ataxias have been attributed to microsatellite expansions since the pathogenic SCA1 allele has been characterized. SCA2 is caused by a CAG microsatellite expansion in the coding region of ATXN2. Normal individuals have repeat counts between 17 and 29, and variant repeats appear to stabilize the tract. Typical of genetic anticipation, SCA2 age-of-onset decreases as pathogenic allele length increases. Individuals affected with SCA2 have repeat counts between 36 and 52 in what appear to be uninterrupted microsatellite tracts, and the mutant allele is transcribed (Imbert et al. 1996; Pulst et al. 1996).

SCA3, also known as MJD, is caused by a CAG microsatellite expansion in the coding region of ATXN3. Normal individuals have repeat counts between 13 and 36, while individuals with SCA3 have repeat counts from 68 to 79. Unlike SCA1 and SCA2, variant alleles are present in both normal and affected individuals, although expansions in SCA3 patients are limited to the 3’ side of the last variant motif. Typical of other trinucleotide microsatellite disorders, SCA3 has classic hallmarks consistent with genetic anticipation, as repeat count is correlated with age-of-onset (Kawaguchi et al. 1994).

SCA6 is caused by a CAG microsatellite expansion in the coding region of several isoforms of CACNA1A. Normal individuals have a repeat counts from 4 to 16, while affected individuals have repeat counts from 21 to 27 (Zhuchenko et al. 1997). SCA7 is caused by a CAG microsatellite expansion in the coding region of ATXN7. Normal individuals have repeat counts ranging from 7 to 17, while affected individuals exhibit extreme variability, with repeat counts from 38 to 130. Expanded alleles are somatically unstable and are highly associated with paternal transmissions. Finally, SCA7 age-of-onset increases in successive generations, which is a hallmark of anticipation (David et al. 1997). According to the National Ataxia Foundation, there are now 36 types of spinocerebellar ataxia. Of these, ten appear to be caused by microsatellite expansions, including the five described above.

4.1.4.1.1.7 Friedreich ataxia

Friedreich ataxia is unusual among microsatellite-linked neurodegenerative disorders since it is autosomal recessive. 98% of Friedreich ataxia chromosomes are caused by an expansion of a GAA microsatellite in the first intron of the FXN gene. The remaining 2% of chromosomes studied had point mutations that were heterozygous with allele expansions—all other individuals with Friedreich ataxia were homozygous for the expanded microsatellite. In unaffected individuals, the GAA microsatellite has a repeat count between 7 and 22, while affected individuals had repeat counts of at least 200. Since Friedreich ataxia is autosomal recessive, it exhibits no signs of genetic anticipation. In fact, all tested parents of individuals affected with Friedreich ataxia were heterozygous for the expanded GAA microsatellite (Campuzano et al. 1996).

4.1.4.1.1.8 DRPLA/HRS

DRPLA, also known as HRS, is yet another neurodegenerative disorder cause by a microsatellite expansion. DRPLA is caused by an expansion of a CAG microsatellite in the coding region of ATN1. In unaffected individuals, the microsatellite repeat count ranges from 3 to 25; while in individuals with DRPLA, repeat counts range from 49 to >75 (Burke et al. 1994; Koide et al. 1994; Nagafuchi et al. 1994). Some studies suggest that longer CAG alleles are correlated with earlier age-of-onset and more severe forms of DRPLA, while the same may not be true for HRS, suggesting that anticipation may play a role in DRPLA (Koide et al. 1994; Nagafuchi et al. 1994). Most studies suggest that the expanded allele perfectly segregates with incidence of DRPLA/HRS (Burke et al. 1994; Koide et al. 1994). In the study that did not show perfect segregation, some unaffected individuals possessing the expanded allele later developed symptoms of DRPLA, implying that long-term disease incidence could segregate perfectly with the expanded allele (Nagafuchi et al. 1994). Expanded alleles are typically paternally transmitted, and affected individuals are all heterozygous for the expanded allele (Burke et al. 1994; Koide et al. 1994; Nagafuchi et al. 1994).

4.1.4.1.1.9 OPMD

OPMD has a few characteristics that distinguish it from other typical microsatellite expansion diseases where mutations occur in coding regions. While most diseases have expanded CAG microsatellites coding for polyglutamine tracts, OPMD has an expanded GCG microsatellite, which codes for a polyalanine tract. The OPMD GCG microsatellite is in the first exon of PABPN1. 98% of chromosomes in normal individuals have a repeat count of 6, while the remaining 2% have a repeat count of 7. Affected individuals have very short expansions—one individual homozygous for a repeat count of 7 had autosomal recessive OPMD, while other individuals in the study population have repeat counts from 8 to 13. Individuals with repeat counts above 8 had a dominant form of OPMD. All alleles observed in OPMD are comparatively stable (Brais et al. 1998).

4.1.4.1.1.10Synpolydactyly

Synpolydactyly is caused by expansions in a GCG microsatellite in the first exon of the HOXD13 gene. The GCG microsatellite is interrupted, and the sequenced microsatellite expansions possess the same interruptions. Affected individuals are heterozygous for insertions of 21, 24, and 30 bp, although some unaffected individuals also appear to be heterozygous for expanded alleles as well (Muragaki et al. 1996). Of all the microsatellite diseases reviewed here, synpolydactyly is the only one where expanded alleles possess variant repeats, and it also has the most complex penetrance pattern, since disease incidence does not appear to perfectly segregate with the expanded allele.

4.1.4.1.2 Microsatellites in cancer

4.1.4.1.2.1 HNPCC and MMP tumors

HNPCC, also known as Lynch syndrome, is defined by increased risk for a constellation of cancers. Individuals with HNPCC are extremely prone to colorectal cancers, and are more likely to be diagnosed with endometrial, gastric, pancreatic, and other cancers. HNPCC cancers typically have diploid tumor genomes. HNPCC has specific diagnostic criteria, known as the Amsterdam criteria, which distinguish it from other MMP tumors (reviewed in de la Chapelle and Peltomaki 1995). The MMP phenotype was first observed in a majority of HNPCC tumors studied, and in a subset of sporadic colorectal tumors (Aaltonen et al. 1993; Ionov et al. 1993; Parsons et al. 1993; Thibodeau et al. 1993; Perucho et al. 1994). A study of an HNPCC-derived tumor cell line demonstrated marked defects in the repair of microsatellite slippage mutations and substitutions (Parsons et al. 1993). This phenotype would indicate that while MMP may be a hallmark of some tumors, somatic substitutions might be the primary contributor to tumor development and progression. The link established by Parsons, et al. between HNPCC and MMR motivated researchers to identify human orthologs to *S. cerevisiae* and *E. coli* MMR proteins. A brief review of the various MMR genes in humans can be found in section 4.1.2.2.1.

MMP is not limited to HNPCC and sporadic colorectal tumors. In a study of 10 ovarian tumors without conclusive HNPCC diagnoses, 50% had MMP (Orth et al. 1994). However, other studies of sporadic ovarian tumors did not reveal extensive MMP (Han et al. 1993; Osborne and Leech 1994). One of the studies undermining the role of MMP in ovarian tumor development did demonstrate that a significant portion of sporadic gastric and pancreatic tumors had MMP (Han et al. 1993). A study of 33 primary small cell lung cancers also revealed that 45% had MMP (Merlo et al. 1994).

4.1.4.1.2.2 Frameshift mutations in MMR proteins

In addition to being a hallmark of MMP, microsatellite frameshift mutations can play a role in cancer incidence and progression. Several studies have identified frameshift mutations in genes that are components of the MMR machinery. An 8 bp A microsatellite in MSH3 was mutated in 39% of MMP colorectal tumors studied, while an 8 bp C microsatellite in MSH6 was mutated in 30% of the same tumors. The mutations were insertions or deletions of single motif, leading to frameshift mutations that could disrupt MMR function, particularly if they are complemented by nonsense or frameshift mutations in the second allele of the appropriate gene. These mutations are not limited to colorectal tumors, as they are found in other MMP tumors studied (Malkhosyan et al. 1996; Miyaki et al. 1997). Similar microsatellite tracts in other genes remain unchanged, indicating that these mutations are selected for in the course of tumor development (Malkhosyan et al. 1996).

4.1.4.1.2.3 Frameshift mutations in tumor suppressor genes

TGFβRII, a gene responsible for controlling cellular proliferation, was the first tumor suppressor gene to be identified as a common target of microsatellite frameshift mutations. In a study of 111 HNPCC tumors, 90% had insertions or deletions of 1 – 2 bp in a 10 bp A microsatellite, while no frameshift mutations were observed in non-MMP tumors (Markowitz et al. 1995; Parsons et al. 1995). Most tumors studied appeared to have frameshift mutations in both alleles. In sequence analysis of a selection of tumors heterozygous for the frameshift mutation, a complementary mutation was found in the second TGFβRII allele. As was the case for frameshift mutations in MSH3 and MSH6, the TGFβRII A microsatellite tract mutates more frequently than other similar microsatellites, suggesting that this mutation provides a selective growth advantage for the tumor (Parsons et al. 1995).

IGFIIR is another tumor suppressor gene that is mutated in some MMP tumors. In a study of a 92 diverse MMP tumors, 13% had mutations in IGFIIR, while no mutations were observed in non-MMP tumors. Most of these mutations occurred in an 8 bp C microsatellite, while one mutation occurred in an interrupted CT microsatellite. Mutations to IGFIIR and TGFβRII are nearly mutually exclusive—of 31 tumors with mutations in either IGFIIR or TGFβRII, only three had mutations in both. In total, more than one third of MMP tumors had a mutation in either gene. Once again, similar microsatellites in other genomic locations do not have recurrent mutations in MMP tumors, suggesting a selective growth advantage is provided by the IGFIIR frameshift mutation (Souza et al. 1996).

A study of the pro-apoptotic BAX gene revealed that >50% of MMP colorectal cancers studied had frameshift mutations in an exonic 8 bp C microsatellite. These mutations were typically insertions or deletions of a single motif. No BAX protein was detected in cell lines with BAX mutant microsatellites, indicating that these mutations were homozygous, hemizygous, or had a complementary mutation on the second BAX allele. BAX frameshift mutations did not occur in non-MMP tumors. Microsatellite tracts in other regions of the genome with composition similar to this BAX microsatellite do not have recurrent frameshift mutations, suggesting that the BAX frameshift mutation provides a selective advantage in tumor development (Rampino et al. 1997).

Frameshift mutations in other tumor suppressors may also provide selective growth advantages to MMP tumors. In a study of 49 MMP colorectal tumors, 39% of tumors had a 1 bp deletion in a 9 bp A microsatellite in TCF4. Similar microsatellite tracts did not have the same mutation frequency in MMP tumors (Duval et al. 1999). In two studies of MMP tumors, EPHB2 was found to have frequent frameshift mutations in a 9 bp A microsatellite—40% of MMP colorectal cancers had the mutation, as did 39% of MMP gastric tumors and 14% of endometrial tumors. All of the colorectal frameshift mutations were somatic; and both colorectal and gastric EPHB2 appeared to confer a selective advantage to the tumor (Alazzouzi et al. 2005; Davalos et al. 2007). In a study of 15 sporadic gastrointestinal MMP tumors, 1 bp deletions in a 9 bp A microsatellite in the BLM gene occurred in 27% of the cases and did not occur in any non-MMP gastrointestinal tumors (Calin et al. 1998; Calin et al. 2001). In a study of 75 MMP gastric and colorectal cancers, 31% were found to have frameshift mutations in a 9 bp A microsatellite in RAD50 and 21% had mutations in the above-mentioned BLM microsatellite. All BLM and RAD50 mutations were heterozygous (Kim et al. 2001). The roles of TGFβRII, BAX, MSH3, MSH6, IGFIIR, and BLM were reinforced in a study of 63 MMP colorectal tumors of various types, which had frequent recurrent frameshift microsatellite mutations in each gene. Many of these mutations, including the BLM frameshift, were correlated with specific clinical and pathological parameters, reinforcing their roles in tumor development and progression (Calin et al. 2000).

Microsatellite frameshift mutations are not limited to MMP tumors. The APC tumor suppressor gene is frequently mutated in both MMP and non-MMP tumors. In a study of 101 sequenced APC mutations in colorectal cancer, frameshift mutations at short microsatellite loci were identified in both MMP and non-MMP tumors. In MMP tumors, 81% of APC frameshift mutations were insertions or deletions of 1 – 2 bp at microsatellite loci with 1 and 2 bp motifs and tract lengths ranging from 3 bp to 10 bp. In non-MMP tumors, 37% of APC frameshift mutations were insertions or deletions of 1 – 4 bp at microsatellite loci with 1 and 2 bp motifs and tract lengths ranging from 3 bp to 13 bp. Every tumor screened had truncated APC protein product (Huang et al. 1996). While frameshift mutations at microsatellite loci are more common in MMR tumors, they may play a role in many different types of tumors.

There is also an apparent association between CAG microsatellite expansions in the AIB1 gene and breast cancer risk in BRCA1 carriers. The relative risk of breast cancer incidence was 1.29 in individuals the BRCA1 mutation and two AIB1 alleles with repeat counts of at least 29. The same effect is not observed in BRCA2 carriers, indicating a specific effect that distinguishes between these two high-risk breast cancer mutations (Kadouri et al. 2004).

4.1.4.1.3 FALS/FTD

FALS/FTD is the third most common neurodegenerative disease and is autosomal dominant. Many cases of FALS/FTD are due to an expansion of a GGGGCC microsatellite tract in C9ORF72. C9ORF72 has at least three transcript variants, two that contain the microsatellite in the first intron, and one that contains the microsatellite in the promoter (DeJesus-Hernandez et al. 2011). In normal individuals, the repeat count ranges from 0 to 23; while in affected individuals, the repeat count ranges from 30 to >1600 (DeJesus-Hernandez et al. 2011; Renton et al. 2011). In a Finnish cohort study, this expansion was observed in ~46% of familial FALS/FTD cases and 21% of sporadic cases (Renton et al. 2011). One third of FALS/FTD patients of European descent also have the expanded allele (Renton et al. 2011). The expansion almost perfectly segregates with the FALS/FTD disease phenotype—only 0.4% of unaffected individuals have the mutant allele (Renton et al. 2011). The expanded microsatellite is haplotype-specific, although normal individuals can have the risk haplotype without the expanded allele (DeJesus-Hernandez et al. 2011). The GGGGCC expansion in C9ORF72 is the most common cause of FALS/FTD that has been identified to date (DeJesus-Hernandez et al. 2011; Renton et al. 2011).

4.1.4.1.4 Microsatellite disease summary

The current model of the role of microsatellites in human disease is composed of two broad categories. The first is composed of neurodegenerative diseases caused by microsatellite expansions, typically at trinucleotide motifs. Microsatellite mutations in these diseases are highly penetrant and segregate strongly with disease incidence. These expansions occur in a variety of genomic contexts and do not have a single pathogenic mechanism. Many of these diseases exhibit genetic anticipation—age-of-onset becomes earlier in successive generations, and can be accompanied by increasingly severe disease symptoms. The second category of diseases is composed of frameshift mutations in tumor suppressor genes. These mutations typically occur in short microsatellites with 1 bp motifs. Many of them truncate protein products, implying that these mutations must either be homozygous, hemizygous, or complemented by a nonsense mutation in the second allele of the gene. Some of these tumor suppressor genes may also be dosage dependent. These categories are probably not exhaustive—there is no reason to assume that they are the only mechanisms by which microsatellite mutations cause disease. A summary of the microsatellite loci linked to human disease can be found in Table 4.1 on pages 92-93.

4.1.4.2 Microsatellites in forensic analysis

Forensic DNA analysis relies on microsatellite genotyping for DNA evidence in convictions, exonerations, and identifying human remains. Microsatellites are particularly useful for these purposes since they can be observed in short DNA fragments (< 300 bp) and can be amplified from samples with non-human contamination. These two factors are particularly important when unidentified human remains are found in unprotected gravesites or have decayed significantly. Microsatellites were first used to identify the remains of an 8-year old murder victim in 1991. This was accomplished by comparing the victim's microsatellite genotypes to her parents' at 6 microsatellite loci. Identity was established by calculating the likelihood that any person would have genotypes consistent with Mendelian inheritance by chance using established microsatellite allele frequencies for the loci used (Hagelberg et al. 1991). Microsatellite genotypes have also been used to identify the remains of victims after enormous disasters, including plane crashes (Olaisen et al. 1997).

Microsatellite analysis can occasionally aid in victim identification even when living parents are not available. In 1991, a shallow grave was discovered in Yekaterinburg, Russia with remains of what was thought to be the last Tsar and Tsarina of Russia, three of their children, and three servants, all of whom were murdered on July 17, 1918. Microsatellite genotyping of the 75-year old remains established that five of the victims at the gravesite were two parents and their children, and that the three other victims were unrelated to any of the others. Since living relatives of the Tsar's family were separated by several generations, the bloodlines of the Tsar and Tsarina were confirmed using mitochondrial DNA (Gill et al. 1994).

A 14-locus microsatellite panel for use in forensic investigations was first proposed in 1993, with a match probability of $< 1 \times 10^{-14}$ (Kimpton et al. 1993). As of 2004, there are 13 CODIS core loci in the FBI's NDIS, with a match probability $< 1 \times 10^{-13}$ (Jobling and Gill 2004). Federal, state, and local authorities use NDIS to identify suspects using DNA evidence isolated from crime scene samples. The same information can be used to exonerate wrongly convicted individuals. Although there are some initial attempts to genotype CODIS microsatellites using HTS data, forensic genotyping still relies on PCR and capillary electrophoresis (Bornman et al. 2012).

4.1.4.3 Paternity testing

Microsatellite genotypes are frequently used to establish paternity or patrilineal descent. One particularly interesting example was the use of Y chromosome microsatellite loci in combination with other DNA markers to establish that Thomas Jefferson likely fathered at least one child with his slave, Sally Hemings. In particular, male-line descendants of Thomas Jefferson's uncle and of Eston Hemings Jefferson—one of Sally Heming's purported sons with Thomas Jefferson—share the fairly unique Jefferson Y chromosome (Foster et al. 1998). Estimating the average rate of *de novo* microsatellite mutations in true descendants is essential when establishing paternity. Several studies have estimated microsatellite mutation rates on the Y chromosome for use when establishing paternity for male offspring (Heyer et al. 1997; Kayser et al. 2000).

The most successful application of HTS microsatellite genotypes to date has been surname recovery from chromosome Y microsatellite loci from whole genome sequencing data. By comparing microsatellite genotypes to those available from public genetic genealogy databases and assuming that chromosome Y microsatellites and surnames are both patrilineal, some family pedigrees and specific individuals could be identified. This was particularly true when other demographic information was available for a person, such as birth year and state of residency. Genetic privacy will continue to decrease as the amount of public microsatellite genotypes and other polymorphism data linked to surnames increases (Gymrek et al. 2013).

4.1.4.4 Microsatellite-based genetic linkage maps and identifying disease loci

One of the first studies describing extensive microsatellite polymorphism within a human population also suggested leveraging this polymorphism in linkage studies (Litt and Luty 1989; Weber and May 1989). This achievement was first realized by creating a genetic linkage map of the human genome with an average resolution of 5 cM (Weissenbach et al. 1992). Several years later, average resolution was increased to 1.6 cM, allowing highly accurate linkage mapping of both monogenic and complex human diseases (Dib et al. 1996). Many of the diseases described above in section 4.1.4.1.1 were first characterized by identifying particular haplotypes and genomic regions that had strong linkage with disease incidence in families (see, for example, Burke et al. 1994; Muragaki et al. 1996; Brais et al. 1998). With the development of SNP arrays and GWAS, linkage mapping in human diseases is less common than it once was. However, linkage maps remain a powerful means of relating phenotypes to genomic regions in organisms without published reference genomes.

4.1.4.5 Microsatellites in population genetics

Microsatellites have found several interesting applications within population genetics. Microsatellite polymorphism has been used to assign individuals from geographically distinct regions to their continents of origin with 88% accuracy. Within European, American, and Oceanian continental subgroups, individuals tend to form clusters corresponding to their population of origin. African populations have greater microsatellite allele diversity than other continental subgroups, supporting the hypothesis of an African origin for humans. This same study provided two other key insights regarding microsatellite diversity in population genetics. First, most microsatellite alleles are neither continent- nor population-specific—on average, individuals from different continental subpopulations share 27% of their alleles, while individuals from the same continental subpopulation share 36% of their alleles. Second, microsatellite genotypes are not informative enough to recover deeper lineages, such as primate lineages (Bowcock et al. 1994). This study suggests that microsatellites might be an effective means of resolving relatively short evolutionary timescales within a particular species (Jorde et al. 1997).

Several later studies reinforce the findings of Bowcock, et al., particularly with regards to recovering continental subgroups from microsatellite data and the enormous amount of microsatellite polymorphism relative to other genetic markers. In two studies on microsatellite diversity and the origin of modern humans, it was again suggested that there is little genetic differentiation between continental subpopulations—microsatellite diversity is nearly the same within and between population subgroups. Microsatellite instability is high enough that the same allele will emerge multiple times in disparate human populations, violating the infinite alleles model. Continental subgroups could still be resolved despite the high level of diversity (Jorde et al. 1995; Jorde et al. 1997). Many studies also echo the observation of Bowcock, et al. with regards to the resolution provided by microsatellite loci—they are most effective for recovering relatively recent lineages, and are not suited for inter-species evolutionary analyses (Heyer et al. 1997; Jorde et al. 1997; Schlötterer et al. 1997). Since microsatellite mutation rates are higher than mutation rates for other genetic markers and are highly variable among loci, care needs to be taken when considering them in analyses that require mutation rate estimates or that attempt to distinguish subpopulations based on the presence of unique alleles. One strategy that has been employed with some success it to use short microsatellite loci that are more likely to have population-specific alleles due to their lower mutation rates (Brinkmann et al. 1998). Another strategy is to employ statistical tests that do not implicitly assume a constant mutation rate (Jorde et al. 1997).

Population genetics studies using microsatellite loci have been used to infer interesting details about human history. The Kaingang people exhibit decreased microsatellite diversity as compared to other populations, which implies a population bottleneck in indigenous American populations around the time of the colonization of the Americas (Di Rienzo et al. 1998). This observation is reinforced by a genome-wide microsatellite study of indigenous American populations (Wang et al. 2007). Additionally, in comparisons of two models for human demographic history, microsatellite variation seems to imply that humans underwent rapid population growth since the human-chimpanzee speciation event (Di Rienzo et al. 1998). A later study estimated a microsatellite mutation rate from loci with 2 bp motifs, which was used to infer that the human-chimpanzee speciation event occurred 3.75 – 6.57 million years ago (Sun et al. 2012).

Population genetic insights from microsatellite loci are not limited to humans. A study of natural D. melanogaster populations lent additional support to the African origin hypothesis for the species. The same study was also able to identify at least one selective sweep within a specific D. melanogaster subpopulation, with weaker evidence for selective sweeps in other populations (Schlötterer et al. 1997).

4.1.4.6 Recovering cellular lineages using microsatellite loci

Several studies have leveraged the considerable somatic instability of some microsatellite loci to infer the cellular lineages, and by extent, developmental history, of animals and tumors. One of the first efforts to infer tumor development from microsatellite mutations evaluated cancer progression in HNPCC patients. Based on the recovered lineages inferred from microsatellite genotypes in hundred-cell samples, consecutive studies suggested that tumors exhibit behavior consistent with clonal evolution along multiple lineages (Tsao et al. 1998; Tsao et al. 1999).

Subsequent proof-of-concept studies sought to demonstrate that cellular resolution of tumor and developmental lineages could be obtained from somatic microsatellite mutations. In a study of 84 cells genotyped at 31 G microsatellites in a single mouse, sufficient resolution was obtained to suggest that liver development pattern occurred at a lobular level, although resolution was insufficient for developmental inferences in other organs (Salipante and Horwitz 2006). By creating artificial "clones" from cultured MMR deficient cells, a second study was able to recover accurate lineages based on a few hundred microsatellite loci. The same study suggested that in MMR proficient cells, hundreds of thousands to millions of microsatellite genotypes might be necessary to recover complete lineages (Frumkin et al. 2005).

Inferences based on microsatellite-based cellular lineages have provided some limited insight into organ and tumor development. A study of microsatellite phylogenies from a subset of murine cell types suggested that cell types are not derived from distinct, clonal lineages and that nearby myosatellite cells share longer developmental paths than distant myosatellite cells (Wasserstrom et al. 2008). Microsatellite lineages were also used to provide insight into the dynamics of murine colonic stem cell crypts: crypts are monoclonal, there is no evidence of an immortal strand within colonic stem cells, and proximal colonic crypts are more closely related to each other than distal crypts (Reizel et al. 2011). Microsatellite lineages allowed for tumor age estimation in a murine lung tumor, demonstrated that nearby tumor cells are more closely related to each other, and established the particular tumor as having a monoclonal origin (Frumkin et al. 2008).

4.1.4.7 Phenotypic variation due to microsatellites

4.1.4.7.1 Variation due to exonic microsatellite polymorphism

Phenotypic variability due to microsatellite polymorphism is not limited to human disease. In a study evaluating the effect of microsatellites coding for polyproline and polyglutamine stretches within GAL4 on transcriptional activity, the lengths of the respective amino acid tracts were correlated with expression. In HeLa cells, transcriptional activity was positively correlated with glutamine tracts of up to 40 units. In the same system, a series of 10 consecutive prolines was found to produce maximal transcriptional activity—subsequent proline tract expansions led to reduced activity (Gerber et al. 1994). In several studies, variation in adjacent glutamine and alanine repeats in RUNX2 accounted for a significant amount of variance in the craniofacial shape of different dog breeds, and in Carnivora more generally (Fondon and Garner 2004; Fondon and Garner 2007; Sears et al. 2007).

4.1.4.7.2 Variation due to non-exonic microsatellite polymorphism

Variation in non-coding microsatellites also contributes to phenotypic variation. Soon after the discovery that AC microsatellites were ubiquitous in eukaryotic genomes, AC microsatellites were inserted into enhancer sequences to determine their effect on gene expression. As compared to controls, transcription from an enhancer with an AC microsatellite was 2 – 10 times higher than control enhancers. Transcriptional activity decreased as the AC microsatellite length increased (Hamada et al. 1984). A more direct phenotypic effect of non-exonic microsatellite variation modulating gene expression was described in an analysis of the sociobehavioral effects of microsatellite variation in prairie voles. Differences in the distribution of the *avpr1a* gene product in the prairie vole brain depend on variation in several long microsatellite tracts in the 5' regulatory region of the gene. Males with longer microsatellite alleles spend more time caring for their pups and exhibit stronger monogamous tendencies than prairie voles with shorter alleles. There were no general changes in male anxiety that could explain these changes, indicating that tract length variation likely accounted for specific social behaviors (Hammock and Young 2005).

4.1.5 Microsatellite genotyping methods

4.1.5.1 Southern blot detection and genotyping of microsatellite loci

The Southern blot was essential to the discovery of widespread microsatellite sequences throughout mammalian genomes. Southern blots demonstrating strong hybridization between an AC microsatellite tract in *E. coli* and DNA from HeLa cells provided the first evidence that AC microsatellites were common throughout the human genome (Hamada and Kakunaga 1982). Subsequent blotting experiments using bacterial and viral DNA as probes provided the first evidence that AC microsatellites were a ubiquitous component of eukaryotic genomes (Hamada et al. 1982b; Jeang and Hayward 1983). Only a few years later, this observation was extended to many more microsatellite motifs in several eukaryotic genomes using the same methods (Tautz and Renz 1984).

Southern blotting has also been used to assess the variability of disease-associated microsatellites. In some of the initial experiments characterizing the mutations causing SCA1, candidate genes were screened for microsatellites by Southern blot using a CAG probe. Once ATXN1 was identified as a candidate gene for SCA1, the CAG repeat variability was established by Southern blotting after a restriction digest to isolate the appropriate DNA fragment (Orr et al. 1993). The same approach was taken to assess the instability of the Fragile X-linked CGG repeat (Kremer et al. 1991).

4.1.5.2 PCR amplification of microsatellite loci

PCR amplification of microsatellite loci is currently the most widespread method to isolate microsatellites for downstream analysis. PCR methods for isolating microsatellite loci were developed almost simultaneously by two separate groups (Litt and Luty 1989; Weber and May 1989). Almost every experiment described in this section that isolated microsatellite sequences for genotyping started with PCR amplification. Although PCR amplification can be followed by Southern blot genotyping, most current microsatellite genotyping methods involve one of the following three techniques.

4.1.5.2.1 PAGE genotyping of microsatellite loci

When primers specific to a particular microsatellite locus can be designed, PAGE is an effective means of calling genotypes at a locus. PAGE genotyping was used in the first applications of PCR to microsatellite loci (Litt and Luty 1989; Weber and May 1989). PAGE comparison of microsatellite genotypes has been used to analyze normal and mutant alleles at disease-associated microsatellite loci and in MMP tumors (see, for example, Fu et al. 1992; Mahadevan et al. 1992; Thibodeau et al. 1993). PAGE was also the initial technique of choice in early forensic analyses of microsatellite loci (Hagelberg et al. 1991; Gill et al. 1994; Olaisen et al. 1997). Initial studies of tumor evolution also relied on PAGE to call microsatellite genotypes (Tsao et al. 1998; Tsao et al. 1999).

4.1.5.2.2 Capillary electrophoretic microsatellite genotyping

Capillary electrophoresis allows for significant automation of microsatellite genotyping, and has higher throughput than PAGE or Southern blot genotyping methods. Current forensic microsatellite genotyping is typically performed using capillary electrophoresis (Jobling and Gill 2004). Many of the more recent microsatellite genotyping studies described earlier in this section called microsatellite genotypes using capillary electrophoretic data. With these techniques, studies of thousands of microsatellite loci have been possible on an enormous Icelandic cohort (Sun et al. 2012). Capillary electrophoresis has also been the genotyping technique of choice in microsatellite genotyping studies using single cells (see, for example, Frumkin et al. 2005; Salipante and Horwitz 2006).

4.1.5.2.3 Microsatellite genotyping from first-generation DNA sequencing data

Maxam-Gilbert sequencing led to the first direct observations of microsatellite sequences (see, for example, Miesfeld et al. 1981; Hamada et al. 1982a). Sanger sequencing has been an essential tool in analyzing disease-linked microsatellite loci. In particular, Sanger sequencing of selected affected and normal individuals in various disease studies led to the discovery of stabilizing variant repeats or otherwise interrupted microsatellite tracts (see, for example, Chung et al. 1993; Eichler et al. 1994; Muragaki et al. 1996).

4.1.5.3 HTS methods for microsatellite genotyping

4.1.5.3.1 HTS microsatellite genotyping challenges

In order to genotype microsatellite loci using HTS data, several issues must be addressed. Any microsatellite genotyper must first define what it considers a microsatellite locus. Every published method for HTS microsatellite genotyping so far appears to rely on TRF to define microsatellite loci. TRF is capable of identifying interrupted microsatellites with statistical signal for repetitiveness. However, some shorter microsatellite motifs that have a demonstrated ability to undergo slippage may go undetected. While this choice may be appropriate for some applications, it will be unsuitable for others.

With the exception of lobSTR, microsatellite genotypers do not directly identify microsatellites in sequencing data. Other genotypers rely on identifying microsatellites by comparing read positions to known microsatellite regions once alignment is complete. Direct microsatellite detection in sequencing data depends on the base-calling quality of the sequence within the microsatellite read. The only current microsatellite genotyper to consider base-calling quality is RepeatSeq. There does not appear to be a single algorithm that directly detects microsatellites in sequencing data and considers base-calling quality in detected microsatellites.

Microsatellite alignment is very sensitive to insertions and deletions. Alignment algorithms such as BWA have *ad hoc* rules for aligning insertions and deletions, as well as about sequence variation more broadly, that limit their ability to map microsatellites with large mutations relative to the reference genome (Li and Durbin 2009). Even methods that use more sophisticated mappers such as BWA-MEM or Bowtie2 may not have alignment parameters that are appropriate for microsatellite loci. Any microsatellite genotyper that relies on an alignment algorithm’s general indel detection framework to correctly map microsatellite reads may not be able to accurately map microsatellite reads with large slippage events.

A well-documented problem that is present any time a microsatellite locus is amplified is PCR slippage (Litt and Luty 1989; Weber and May 1989; Gymrek et al. 2012; Highnam et al. 2013). Furthermore, while the average rate of slippage may be modeled accurately by a few parameters describing the locus, slippage could vary considerably between similar microsatellite loci. Even with the characteristics affecting microsatellite instability described in section 4.1.2.3, microsatellite stability is still not completely understood. If this unpredictability is true for slippage *in vivo* and *in vitro*, any averaged PCR slippage estimate will be inaccurate for a significant portion of microsatellite loci. At the very least, metrics should be provided describing how well a proposed genotype fits with a genotyper’s error model. With this information, users can make informed choices about whether the proposed genotype is consistent with the slippage model proposed. Generalized error models could significantly impair the sensitivity and specificity of any microsatellite genotyper, which would be particularly damaging when attempting to identify *de novo* microsatellite mutations.

In certain systems, such as exome capture, microsatellite capture may be significantly impaired. Some studies have already described decreased capture efficiency for non-reference indel alleles in exome capture data (Iossifov et al. 2012). It is reasonable to expect that these same challenges will exist when considering microsatellite indels specifically. Moreover, in order to genotype microsatellite loci, alleles must be completely contained within sequencing reads. Therefore, longer microsatellite alleles may have lower coverage than shorter alleles at the same locus if read lengths are a limiting factor. Both of these concerns suggest that coverage biases among microsatellite alleles at a locus must be accounted for to obtain accurate microsatellite genotype calls.

Finally, regardless of how a genotyper defines a microsatellite, some true microsatellite mutations will escape detection. No published algorithm accounts for mutation events that move microsatellites outside their respective detectable ranges. Furthermore, microsatellite can exist in CNVs, which would invalidate any bi-allelic genotype call. In these situations, it is necessary to have some sense of what the expected coverage at a microsatellite locus should be. Only then can a genotyper assess whether coverage is sufficient to suggest that two alleles—and only two alleles—are being detected.

4.1.5.3.2 TRF-derived microsatellite genotypers

Since the development of HTS, several groups have designed microsatellite genotypers that leverage this data to call microsatellite genotypes either in great depth or across hundreds of thousand to millions of loci. Nearly every current HTS microsatellite genotyper relies on TRF to define reference microsatellite loci. Therefore, before describing the advantages and disadvantages of some of the most recent microsatellite genotyping methods, a brief overview of how TRF defines microsatellite loci is in order.

TRF uses a probabilistic model to detect two or more contiguous, approximate copies of the same motif. TRF detects motif lengths up to several hundred bp, so it will detect microsatellite, minisatellite, and satellite DNA. It accomplishes this by scanning an input sequence for matches of DNA kmers and comparing the distance between consecutive matches, as well as any intervening matches to other kmers. TRF looks to identify regions of the genome where a repeated pattern occurs more often than would be expected by chance, as characterized by several distributions. By setting a window size, kmer length, and a tolerance for mismatches and indels, candidate tandem repeats can be identified. If the putative tandem repeat contains at least two copies of the putative motif, TRF will report the repeat with information including the motif length, motif, and sequence identity. TRF performance on sample datasets demonstrates some inconsistency in detecting microsatellites—five previously annotated microsatellites in a human gene were not detected by TRF, while another 13 previously unknown microsatellites in the same gene were identified by TRF (Benson 1999).

4.1.5.3.2.1 lobSTR

The most successful HTS microsatellite genotyping algorithm to date, lobSTR splits genotyping into three stages: sensing, alignment, and allelotyping. lobSTR statistically senses microsatellites by measuring the sequence entropy of overlapping windows within a sequencing read using 2 bp kmers. A lower entropy score corresponds to more repetitive sequence. By setting a specific entropy threshold, lobSTR identifies both interrupted and uninterrupted putative microsatellite sequences with high specificity and sensitivity. lobSTR then determines the microsatellite motif length using a fast Fourier transform to identify strong periodic signals within the repetitive region. Finally, lobSTR sets the most common kmer of the appropriate motif length within the repetitive region as the consensus motif (Gymrek et al. 2012).

lobSTR uses BWA to align microsatellite reads to a customized reference set of microsatellite flanking sequences separated by motif equivalence class (Li and Durbin 2009). The reference microsatellite set is determined using a TRF scan of the reference genome. The minimum amount of flanking sequence on either side of a microsatellite typically required by lobSTR to accurately map a microsatellite locus is 8 – 9 bp. The resulting alignments determine the position and tract length of a microsatellite in a read, since the lobSTR sensing step cannot define exact microsatellite boundaries. After an initial alignment step, lobSTR realigns reads in the candidate region using the Needleman-Wunsch local realignment algorithm (Needleman and Wunsch 1970; Gymrek et al. 2012).

lobSTR calls "allelotypes" by considering the reads aligned at a particular microsatellite locus and a model of expected stutter noise. The expected stutter noise model can either be generated in the course of sample preparation, or a standardized model can be used. For each read, lobSTR calculates the likelihood that it represents a true genotype or whether it is a product of slippage noise, taking into account the distance between the observed tract length and the putative allele lengths. lobSTR's noise model is based on a logistic regression modeling of the slippage probability and a Poisson regression modeling of the distance of a read from a hemizygous reference. The logistic regression model uses four parameters when estimating slippage probability: motif length, tract length, flanking GC content, and sequence identity. However, lobSTR does not detect any relationship between tract length and noise rate. The Poisson regression on the distance of a read from a hemizygous reference allele assumes that errors frequently arise from slippage events, and occasionally arise from non-microsatellite indels and substitutions (Gymrek 2013). lobSTR noise models are inferred from hemizygous microsatellites on the Y chromosome in males. In addition to a noise model, lobSTR requires that each allele at a heterozygous locus have ≥20% of total coverage at a locus, that the best allelotype has ≥50% of total coverage at a locus, and that allelotypes on sex chromosomes only have one allele (Gymrek et al. 2012).

lobSTR was initially validated using biological replicates from the same individual. In separate blood and saliva samples from a single person, and limiting calls to loci with at least 21X coverage, genotype discordance was 3%, while allelotype discordance was 2%. Most discordant microsatellite loci had slippage-prone 2 bp motifs, so miscalls may be due to high error rates. Comparing Mendelian inheritance within a trio also validated lobSTR. Loci with at least 15X coverage each trio member had a Mendelian inheritance rate of 99%. However, this coverage threshold limited lobSTR genotypes to ~1% of all loci in their reference set (Gymrek et al. 2012). Even if we were to assume that every microsatellite locus had the highest microsatellite mutation rates reported in the literature, this de novo mutation rate is at least an order of magnitude too high. lobSTR was subsequently used with moderate success to recover surnames from sequencing data based on chromosome Y STRs (Gymrek et al. 2013).

lobSTR is a powerful tool for certain types of microsatellite analyses. By focusing on somewhat long and interrupted microsatellite loci, it can compare genotypes within a population at loci that have frequent SNPs or indels. The types of applications it has been applied to this far benefit greatly from the overall accuracy of the technique, since the focus has been on population information, or on Y chromosome microsatellites where there are no heterozygous calls to be made.

lobSTR is not without its caveats. Some versions consider microsatellites with 1 bp, while others do not. lobSTR also has very limited coverage of microsatellites in exons. The standard slippage model distributed with lobSTR was compiled by the Erlich lab from tens of thousands of Y chromosome microsatellite loci. Therefore, variance in microsatellite slippage rates due to specific experimental protocols will not be detected. The error model also assumes that locus instability can be estimated per locus by a logistic regression on four variables. This assumption may describe the average locus slippage rate based on broad parameters, but it will not capture the variance in microsatellite slippage rate among similar microsatellite loci. This variance can be stochastic, or it can be due to parameters not included in the model. Finally, lobSTR's own validation demonstrates that microsatellite allelotypes are correct ~99% of the time, which is sufficient for inferring "broad" information about microsatellites in the human genome. This could be population behavior at a microsatellite locus, or a sense of microsatellite variation within the lobSTR reference set. lobSTR's reported accuracy does not suggest it is capable of accurately identifying specific instances of microsatellite mutation or polymorphism, particularly *de novo* microsatellite mutations or microsatellite mutations that may be related to particular phenotypes or diseases.

4.1.5.3.2.2 RepeatSeq

The only other HTS microsatellite genotyper published that considers microsatellite slippage is RepeatSeq. RepeatSeq takes data that has already been aligned using BWA or Bowtie2, and has been processed using GATK's IndelRealigner. RepeatSeq then identifies reads that completely contain microsatellites in its TRF-derived reference set of microsatellites with motif lengths from 1 to 5 bp. RepeatSeq determines the most probable genotype at a locus by considering all possible allele combinations using Bayesian model selection, where each model is a possible genotype. The noise parameters for RepeatSeq are derived from sequencing data from >100 inbred fly genomes. The noise model is based on three parameters: reference microsatellite tract length, motif length, and average base-calling quality of the microsatellite tract. The model is binned into a 5 × 5 × 5 array, which is used to select an appropriate error rate for a locus. RepeatSeq requires a minimum of two reads at a locus to call a genotype (Highnam et al. 2013).

In early implementations of lobSTR and RepeatSeq, lobSTR appeared to have the most accurate microsatellite mappings, while RepeatSeq mapped more microsatellite loci. However, in their current implementations, lobSTR and RepeatSeq consider the same reference microsatellite set. When calling genotypes in a parent/child trio, RepeatSeq's best performance with regards to Mendelian inheritance is ~98% when minimum coverage for every individual is at least 17X. In comparisons of lobSTR and RepeatSeq genotypes from the same sequencing data, genotypes were found to be ~88% concordant at loci where both pipelines made a genotype call. At loci with discordant calls between the two pipelines, RepeatSeq was concordant with capillary electrophoresis validation ~62% of the time, while lobSTR was only concordant 10% of the time. lobSTR also occasionally mapped more reads to the locus (Highnam et al. 2013).

RepeatSeq has many of the same drawbacks as lobSTR when genotyping microsatellite loci. While both methods allow for improved microsatellite genotyping from HTS data as compared to methods that do not consider microsatellite slippage, neither comes close to the accuracy necessary to detect some of the most meaningful microsatellite mutations. Its error model is likely to be less accurate than lobSTR's error model, particularly when a microsatellite allele in an individual deviates significantly from the reference. Moreover, RepeatSeq is sensitive to whatever heuristics are employed by the upstream alignment algorithm, which might limit indel size and impair its ability to detect long microsatellite indels. This is certainly the case with BWA (Li and Durbin 2009).

4.1.5.3.2.3 Other methods

At least three other methods have been published that attempt to use HTS data to genotype microsatellites. One method used reads aligned by BWA that have been filtered using a set of *ad hoc* parameters to provide robust allelotypes at ~1,000 informative microsatellite loci that also conformed to Mendelian inheritance (McIver et al. 2011). This method did not have any error model accounting for slippage noise and was subject to the additional *ad hoc* parameters defined by BWA when detecting indels in sequencing data (Li and Durbin 2009). Another HTS microsatellite genotyper was published that was designed to specifically call genotypes at CODIS loci. This genotyper did not need an explicit error model since coverage was very high at the selected loci. In addition, the targeted loci were mapped to a reference set of CODIS loci with all possible alleles, rather than aligning reads to the reference genome, obviating the need for gapped alignment. In general, this method seemed to perform well as compared to typical CODIS genotyping via capillary electrophoresis, but it does not present a generalizable method for large-scale microsatellite genotyping (Bornman et al. 2012). Most recently, STRViper, a method for calling genotypes at TRF-defined microsatellite loci by inferring insertions or deletions from the insert sizes of paired-end reads was published. This method relies on high coverage and low variance in DNA fragment length, and has no noise model for stutter (Cao et al. 2014). While it is able to detect large indels, it provides no direct observation that proves an indel near a microsatellite locus is actually a microsatellite indel. In addition, it has no sensitivity to identify short microsatellite mutations whose indel magnitude is not distinguishable from insert size variance.

4.2 ASD

4.2.1 ASD symptoms

According to the DSM-5, ASD disorders are characterized by a constellation of symptoms. Individuals diagnosed with ASD have persistent deficits in social communication and interaction, including deficits in social and emotional reciprocity; deficits in verbal and nonverbal communicative behaviors; and deficits in developing and maintaining relationships. These individuals also have restricted patterns of behavior, interests, or activities. This can include repetitive movements or speech patterns; inflexible adherence to routines; fixated interests; and hyper- or hyporeactivity to sensory input (American Psychiatric Association 2013).

ASD symptoms present in early development and must cause clinically significant impairment in important areas of social functioning. ASD is distinct from intellectual disability or global developmental delay. The DSM-5 recognizes 3 levels of ASD severity, which could suggest the need for anywhere from some support to very substantial support (American Psychiatric Association 2013).

4.2.2 ASD diagnosis prevalence

According to the latest study from the CDC using data from 11 ADDM sites, one of every 68 8-year old children will be diagnosed with ASD. ASD incidence varies significantly by gender and ethnic affiliation. Boys are 4.5 times more likely to be diagnosed with ASD than girls. Non-Hispanic white children are ~30% more likely to be given an ASD diagnosis than non-Hispanic black children, and ~50% more likely to be given an ASD diagnosis than Hispanic children. The median age of earliest known ASD diagnosis was 53 months old, and did not vary significantly by gender or ethnic affiliation. ASD diagnoses also vary geographically. Some of the variation in diagnosis may be due to external factors such as access to healthcare and psychiatric services or differences in diagnostic practices (Centers for Disease Control and Prevention (CDC) 2014).

4.2.3 The SSC dataset

The SSC dataset was created in an effort to identify *de novo* genetic mutations that contribute to autism incidence. Its focus is on recruiting simplex autism families with unaffected parents and siblings. Simplex autism families have only child with autism, while multiplex families have more than one child with autism. Children with autism in the SSC have at least one unaffected sibling, and typically exhibit moderate to severe symptoms of autism with few signs of intellectual disability. The absence of recurrent autism phenotypes within SSC families suggest that this dataset is enriched for *de novo* mutations that are likely to play a role in autism incidence. Each family has had blood samples drawn and frozen at RUSDR for use in various genetic analyses (Fischbach and Lord 2010). In order to enrich the SSC for *de novo* mutations, simplex families are specifically recruited, and multiplex families are excluded from the dataset. However, it remains possible that there are transmitted polymorphisms within the dataset that cause autism (Levy et al. 2011; Iossifov et al. 2012). The SSC dataset is currently composed of 2,700 families.

4.2.4 The genetics of autism

ASD is among the most strongly heritable psychiatric illnesses. In studies of ASD incidence among monozygotic twins, concordance ranges from ~60% to ~90% (Rosenberg et al. 2009; Hallmayer et al. 2011). A well-supported genetic model for autism incidence suggests that there are two autism risk classes. The first class consists of high-risk families with transmitted causative mutations from affected parents to children with autism. These transmitted mutations would likely act in a dominant fashion. Based on the gender bias observed for autism, these mutations are primarily transmitted from unaffected mothers to their affected sons, and carry a ~50% risk of recurrence if a family has multiple sons. These carrier females were thought to comprise <1% of all women, and the transmitted mutations would be novel relative to the maternal grandparents. Based on autism incidence data from AGRE and IAN, the high-risk class comprises <1% of all families with autism (Zhao et al. 2007).

The second risk class consists of the remaining >99% of the population, where autism incidence is driven by *de novo* mutations in parental germlines. Based on the gender bias observed in autism, these *de novo* mutations are expected to have high penetrance in males and relatively poorer penetrance in females. Due to the role played by *de novo* mutations, parental age should be positively correlated with autism incidence. This model is consistent with self-reported autism incidence data from AGRE and IAN. This model suggests that mutations to a single allele of a gene frequently lead to autism, which implies that some of the genes where these mutations occur could be susceptible to haploinsufficiency (Zhao et al. 2007).

The conclusions of this model were reinforced by a concurrent study demonstrating frequent *de novo* CNV mutations in families. Most of the CNVs observed in children with autism were deletions, reinforcing the role of haploinsufficiency in autism etiology. Interestingly, none of the *de novo* mutations were seen more than twice in the dataset studied, suggesting that mutations at many loci throughout the genome contributed to autism incidence (Sebat et al. 2007). These observations were reinforced in a study of *de novo*

CNVs in the SSC dataset. In addition to reinforcing the observations of Sebat, et al., this study observed that CNVs in children with autism had higher gene content than CNVS in their siblings, and that deletions of genes only occurred in children with autism. Females with autism tended to have *de novo* CNVs with higher gene content than males with autism, suggesting that females can tolerate a higher mutational load before exhibiting disease symptoms. Once again, most *de novo* CNVs were unique, suggesting that there are many genes whose mutations can lead to autism (Levy et al. 2011).

A large study of *de novo* SNVs in or near exons in the SSC demonstrated that children with autism were twice as likely as their unaffected siblings to have *de novo* mutations that disrupt gene function, such as frameshift indels, nonsense SNVs, and splice-site SNVs. These mutations were most commonly transmitted from the paternal germline, which accumulated mutations with advancing paternal age. This study estimated that disruptive mutations to a set of 350 – 400 autism risk genes. Additionally, disruptive mutations were observed significantly more often in FMRP-associated genes than would be expected by chance. In unaffected children, disruptive mutations to FMRP-associated genes are exceedingly rare (Iossifov et al. 2012).

Several smaller exome sequencing studies—primarily using SSC families that were not studied by Iossifov et al.—further elucidated the genetics of autism. One study observed the same bias for paternal transmission and paternal age for *de novo* mutation incidence and estimated that mutations to any of 384 – 821 genes could lead to autism. Unlike Iossifov et al., missense mutations were significantly more common in children with autism than in their unaffected siblings. This study observed recurrent mutations in CHD8 and NTNG1, suggesting that mutations to these genes may be causal. Many of the most severe *de novo* mutations in this study occurred in an interconnected β-catenin/chromatin remodeling protein network, suggesting another pathway that could be enriched for genes whose mutations might lead to autism (O'Roak et al. 2012b). Two other studies, only one of which used the SSC dataset, observed recurrent mutations in CHD8 as well and an increased *de novo* mutation load in older fathers (Neale et al. 2012; Sanders et al. 2012). In the study that used the SSC dataset, *de novo* mutations in brain-expressed genes were more likely to occur in children with autism than in their siblings (Sanders et al. 2012).

A targeted sequencing study using the complete SSC cohort at the time found recurrent mutations in six genes of a 44-gene target set. Of the 27 *de novo* mutations observed, one third occurred in CHD8. Children with autism who had CHD8 mutations also tended to have larger head circumferences than people without a CHD8 mutation. Although this study identified specific loci that likely cause ASD, it was unable to determine whether mutations to these genes were sufficient to lead to autism incidence (O'Roak et al. 2012a).

Most recently, a review was published synthesizing the observations of the studies described above that reinforced the significant role of *de novo* mutations in autism incidence. By combining the observations of Iossifov, et al., and Sanders, et al., a significant enrichment of missense mutations in children with autism relative to their siblings was observed. The authors predict that there are 250 – 500 autism risk genes. As of the time since the review has been published, common variants have yet to be associated with autism, suggesting that autism incidence is primarily driven by the two-class model of Zhao et al. The authors also estimate that within the SSC dataset, 60% of autism cases arise from *de novo* mutations. *De novo* mutations have only been observed in ~35% of children with autism in the SSC, which means that many undetected mutations remain to be identified (Ronemus et al. 2014).

Based on the observations of frequent deletions in CNV studies, the authors suggested that genes involved in autism incidence are dosage sensitive. This dosage sensitivity of deleted genes, and by extension, disruptive SNVs and indels, could be realized in three ways. First, disrupted genes may be particularly sensitive to decreased expression, i.e. are haploinsufficient. Second, these genes may be monoallelically expressed, in which case, one disrupted copy could lead to many cells without an essential gene product. Finally, these mutations could lead to dominant mutant alleles that negate the function of their normal counterparts (Ronemus et al. 2014).

4.3 Summary

Microsatellites are widespread throughout eukaryotic genomes and have dynamic, but poorly understood, mutation processes. Their exceptional variability has made them invaluable as markers in various genomes, particularly in humans. Microsatellite mutations are known to play a role in several devastating monogenic neurodegenerative diseases. They are both a hallmark of MMP tumors and a contributory factor to tumor development in MMP and non-MMP tumors. There is no reason to assume that the roles of microsatellites in human disease are limited to these two disease models.

Microsatellite genotyping technology is at an impasse. Low throughput technologies such as Sanger sequencing or capillary electrophoresis produce highly accurate and reproducible genotype calls that are suitable for *de novo* mutation detection. However, their utility for studying microsatellite loci in an unbiased, genome-wide manner is limited since they require primers for specific loci. Although some algorithms have been designed to leverage HTS data to call genotypes at hundreds of thousands—even millions—of microsatellite loci, their accuracy is insufficient for specific detection of *de novo* microsatellite mutations. Based on their reported Mendelian inheritances, true *de novo* mutations will be swamped by false positives.

This is particularly unfortunate in the study of human diseases such as autism, where sporadic disease incidence seems to be driven by *de novo* mutations in parental germlines. Due to their well-documented instability, frequent mutations to microsatellite loci may play a significant role in disease incidence, particularly when they are uninterrupted by SNVs or indels. This project aims to bring microsatellite genotyping to the next level—by employing sophisticated genotyping methods and leveraging population-level information, we aim to develop sufficient specificity to accurately call *de novo* microsatellite mutations from HTS data. By demanding high specificity in our genotype calls, we expect that this approach will be useful for many applications beyond *de novo* mutation detection.

The uSeq pipeline is the realization of this goal. This pipeline takes sequencing data from thousands of individuals as input and detects uninterrupted microsatellites, which it then maps to a reference genome. After filtering reads and creating microsatellite profiles for each individual, uSeq learns parameters from the data that are specific to the experimental protocol being used. These parameters include per-allele, per-locus expected coverage estimates, locus-specific slippage error estimates, and allele-specific capture bias estimates. These estimates enable highly accurate and tailored genotype calling, which is accompanied by careful consideration of how each call fits with uSeq's proposed genotyping model. These genotypes are then used to identify possible *de novo* mutations, which are filtered to ensure that only high-confidence bi-allelic trios are considered.

There are several advantages to uSeq as compared to other HTS genotyping methods. uSeq considers vastly more microsatellite loci than the latest implementations of lobSTR and

RepeatSeq. Although uSeq does not employ the same detection strategy as TRF, only a small percentage of TRF-derived reference loci do not overlap a uSeq reference microsatellite locus. Over 99.9% of all trio genotype calls are consistent with Mendelian inheritance. Detected alleles are also consistent when comparing exome and whole genome sequencing coverage from the same individuals. Moreover, uSeq demonstrates excellent specificity when distinguishing true *de novo* mutations from noise, as determined by validation sequencing. We are able to leverage uSeq's ability to detect *de novo* microsatellite mutations to determine that children with autism are significantly more likely to have unique microsatellite *de novo* expansions than their siblings. Although not statistically significant, we also observe several frameshift mutations in children with autism. Many of these frameshift mutations are consistent with the protein networks and recurrent mutations described by other studies of sporadic autism. We do not observe a single microsatellite frameshift mutation in unaffected siblings. Other general indel detection pipelines applied to the SSC dataset were incapable of detecting the majority of uSeq *de novo* microsatellite mutations.

While the results of this study suggest a role for *de novo* microsatellite mutations in autism, there is a much broader point to be drawn from this project. Microsatellite mutations should be expected to play a role in many complex human diseases and other sources of phenotypic variation. They must be examined in an unbiased, genome-wide manner. uSeq provides the means to further elucidate the role *de novo* mutations and polymorphism may play in a variety of different diseases in HTS-based population studies.

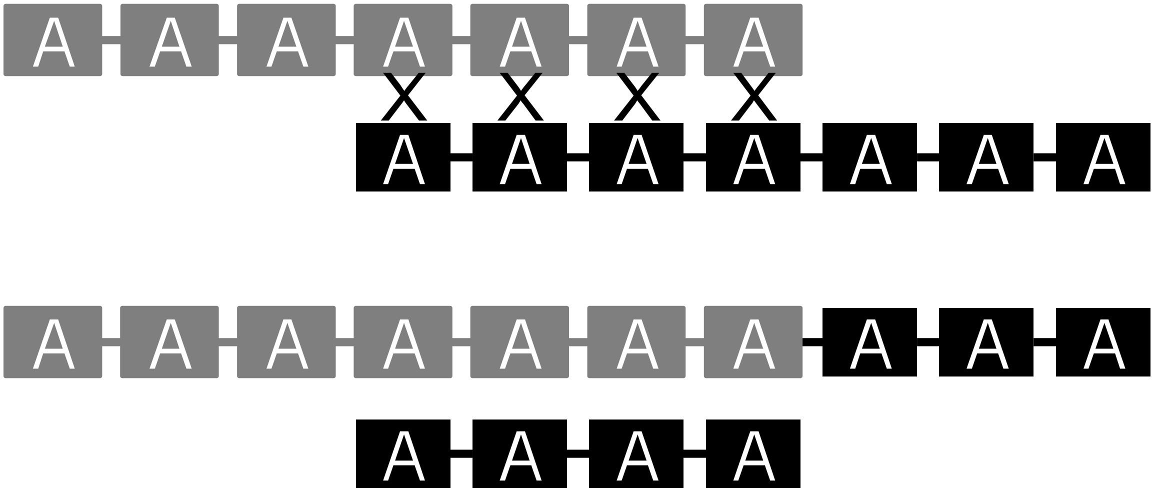


Figure 4.1: A schematic representation of unequal crossing over. Two chromosomes with equal numbers of identical tandem repeats align out-of-phase and undergo recombination. This results in one shorter repetitive region and one longer repetitive region in the resulting chromosomes.

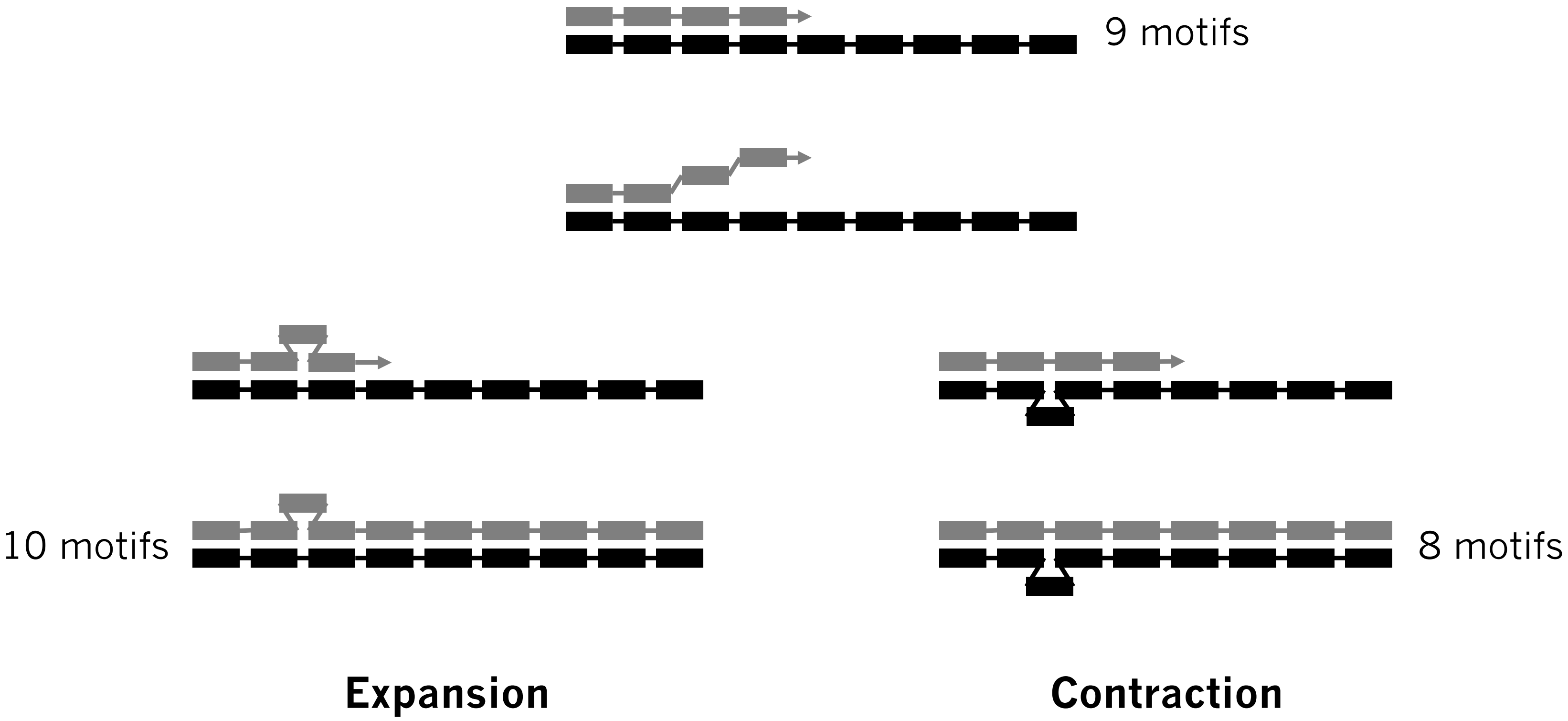


Figure 4.2: A schematic representation of replication slippage. Once DNA polymerase dissociates from a nascent and template strand, the strands can dissociate from each other. The strands can then re-anneal out of phase, creating a bulge in the nascent or template strand. If the bulge is in the nascent strand (grey), the nascent microsatellite will expand relative to the template microsatellite. If the bulge is in the template strand (black), the nascent microsatellite will contract relative to the template microsatellite.

Disease	Gene	Context	Primary motif	Genomic position (GRCh37)	Normal range[1]	Affected range[1]	Variant repeat
Fragile X syndrome	FMR1	5' UTR	CGG	chrX:146,993,569-146,993,629	15 - 65	200+	Y
SBMA	AR	Exon	CAG	chrX:66,765,159-66,765,227	17 - 26	40 - 52	N
DM1	DMPK	3' UTR	CTG	chr19:46,273,463-46,273,524	5 - 30	50+	N
DM2	ZNF9	Intron	CCTG	chr3:128,891,420-128,891,502	~26	75+	Y
Huntington's disease	HTT	Exon	CAG	chr4:3,076,604-3,076,662	11 - 34	42+	N
SCA1	ATXN1	Exon	CAG	chr6:16,327,865-16,327,955	6 - 39	41 - 81	Y
SCA2	ATXN2	Exon	CAG	chr12:112,036,754-112,036,823	17 - 29	36 - 52	Y
MJD/SCA3	ATXN3	Exon	CAG	chr14:92,537,355-92,537,378	13 - 36	68 - 79	Y
SCA6	CACNA1A	Exon	CAG	chr19:13,318,673-13,318,712	4 - 16	21 - 27	N
SCA7	ATXN7	Exon	CAG	chr3:63,898,361-63,898,392	7 - 17	38 - 130	N
Friedreich ataxia	FXN	Intron	GAA	chr9:71,652,201-71,652,220	7 - 22	200+	N
DRPLA/ HRS	ATN1	Exon	CAG	chr12:7,045,892-7,045,938	3 - 25	49+	N
OPMD	PABPN1	Exon	GCG	chr14:23,790,681-23,790,701	6 - 7	8 - 13	N
Synpolydactyly	HOXD13	Exon	GCG	chr2:176,957,786-176,957,827	NA	NA	Y
MMP cancers	MSH3	Exon	A	chr5:79,970,915-79,970,922	8	fs	N

Disease	Gene	Context	Primary motif	Genomic position (GRCh37)	Normal range[1]	Affected range[1]	Variant repeat
MMP cancers	MSH6	Exon	C	chr2:48,030,640-48,030,647	8	fs	N
HNPCC	TGFβRII	Exon	A	chr3:30,691,872-30,691,881	10	fs	N
MMP cancers	IGFIIR	Exon	G	chr6:160,485,488-160,485,495	8	fs	N
MMP cancers	IGFIIR	Exon	CT	chr6:160,505,170-160,505,181	12	fs	Y
MMP colorectal cancer	BAX	Exon	C	chr19:49,458,971-49,458,978	8	fs	N
MMP colorectal cancer	TCF4	Exon	A	chr18:52,894,880-52,894,888	9	fs	N
MMP cancers	EPHB2	Exon	A	chr1:23,240,246-23,240,254	9	fs	N
MMP cancers	BLM	Exon	A	chr15:91,304,139-91,304,147	9	fs	N
MMP cancers	RAD50	Exon	A	chr5:131,931,452-131,931,460	9	fs	N
BRCA1 breast cancer	AIB1	Exon	CAG	chr20:46,279,815-46,279,865	?	29+	Y
FALS/FTD	C9ORF72	Intron/ promoter	GGGGCC	chr9:27,573,522-27,573,544	0 - 23	30+	?

Table 4.1: Diseases caused by microsatellite mutations. Since there is no specific microsatellite locus in APC that is mutated, it is not shown here. References can be found in the main text.

[1] Affected and normal ranges are given as repeat counts unless otherwise specified.

5 Methods

5.1 uSeq pipeline overview

The uSeq pipeline can be broken down into two modules. The individual module of the uSeq pipeline processes sequencing data from individuals within the study population and consists of four stages (Figure 5.1, page 170):

1. Scan and condense microsatellites
2. Align condensed reads to a condensed reference genome using BWA
3. Convert file formats, reindex reads to the uncondensed reference genome, sort reads, and perform file management tasks
4. Merge sequencing data from multiple lanes, mark PCR duplicates and create profiles for each observed microsatellite locus

The first three steps of the individual module are run separately on every sequencing file provided for an individual.

The population module of the uSeq pipeline merges the microsatellite profiles of the entire study population, estimates genotyping parameters, and calls genotypes. This module also consists of four stages (Figure 5.2, page 171):

1. Merge microsatellite profiles for the entire study population
2. Identify loci with sufficient study-wide coverage for genotyping
3. Generate per-locus, per-person expected coverage estimators
4. Call genotypes while iteratively updating per-allele bias parameters and per-locus noise parameters using a population EM algorithm

Since the specific application discussed in this thesis is identifying *de novo* mutations, the genotypes are then used to identify strong Mendel violation candidates. A complete discussion of how uSeq identifies strong *de novo* mutation candidates can be found later in the chapter, in section 5.4.

uSeq also creates a condensed reference genome prior to processing sequencing data. The same microsatellite detection parameters must be used to scan the reference genome and the sequencing data that is aligned to it. Once microsatellite loci have been identified in the reference genome, any steps necessary to generate the proper files for a particular short-read alignment algorithm—such as BWA—must be performed.

5.2 uSeq Individual Module

5.2.1 Microsatellite detection

5.2.1.1 Microsatellite detection inputs

Microsatellite detection is done using an algorithm of our own design. The detection algorithm can take sequencing data in single- or paired-end FASTQ or BAM formats, or any other DNA sequence—such as a reference genome—in FASTA format. If sequencing data is provided, additional parameters can be specified that truncate subsequences with poor base-calling quality, limiting microsatellite detection to confidently called regions of sequencing reads. The algorithm can also take user-defined parameters specifying minimum tract length, number of repeated motifs, and upper and lower boundaries for motif length. If these parameters are not specified, uSeq uses default parameters. In this study, the minimum tract length was 8 nucleotides, the minimum repeat number is 3, and motif lengths between 1 and 6 bp were considered. For the sequencing data used, portions of reads with base-calling quality below 20 are trimmed using the method outlined in the original BWA manuscript (Li and Durbin 2009).

5.2.1.2 A rapid algorithm for perfect microsatellite detection

The detection algorithm is a FSM derived from the MTF transform, which was originally proposed as a locally adaptive data compression scheme, but whose properties enable it to detect locally repetitive patterns (Bentley et al. 1986). The FSM has two states: putative microsatellite (M) and non-microsatellite (N). As the algorithm proceeds along an input DNA sequence, *S*, the state machine transitions between states based on the local repetitiveness of S (Figure 5.3, page 172). A FIFO dictionary of the k most recent overlapping kmers tracks local repetitiveness, where k is the maximum microsatellite motif length. The FSM initializes with an empty dictionary in the N state. At each position in *S*, the current kmer is compared to the dictionary kmers. The dictionary then updates by adding the most recent kmer to the end of the dictionary, and removing the k^{th} kmer from the beginning of the dictionary. This dictionary update occurs regardless of the FSM state. If no match is found in the dictionary, the FSM remains in the N state, and the next kmer is compared to every kmer in the dictionary. If a match is found, the FSM transitions to the M state. While in the M state, the FSM only checks *i*—the index position of the match that initiated the transition from the N state to the M state—and increments the match count, *m*, by one for every match. Once the current kmer no longer matches the kmer at index *i* in the dictionary, the FSM reports the current repetitive sequence if it meets the specified microsatellite thresholds, and then returns to the N state. Upon this transition, the tract length, start position, and motif length are easily extracted from the state machine:

$$tract\ length\ (t): m + k + i - 1$$

$$motif\ length\ (u){:}\ k - i$$

$$start\ position\ (s){:}\ x - t + 1$$

In the start position calculation, x is the start position of the kmer in S that first matched a dictionary kmer. Repeat count can be readily calculated using t and m, and the microsatellite motif sequence is simply the substring of S starting at s with length m. Below is the pseudocode for the microsatellite detection algorithm.

Algorithm 5.1 Detecting perfect microsatellite sequences

```
Input: maximum kmer length $k$; input sequence $S$; detector thresholds $minMotifLength$, $maxMotifLength$,
    $minTractLength$, and $minRepeatCount$
$matchIndex \leftarrow \emptyset$        // If the detector has identified a putative microsatellite, $matchIndex$ defines the index in the kmerDictionary that the detector checks
$matchCount \leftarrow 0$
$kmerDictionary \leftarrow \emptyset$
for $i = 0$ to $k$ do        // Initialize $kmerDictionary$
    $kmerDictionary[i] \leftarrow \emptyset$
end for
for $i = 0$ to $|S| - k - 1$ do        // all overlapping kmers in $S$
    $currentKmer \leftarrow$ substring of $S$ starting at $i$ with length $k$
    if $matchIndex = \emptyset$ then        // not in a putative microsatellite tract
        for $j = 0$ to $k - 1$ do
            if $currentKmer = kmerDictionary[j]$ then
                $matchCount \leftarrow 1$
                $matchIndex \leftarrow j$
                $transitionStart \leftarrow i$
            end if
        end for
    else        // in a putative microsatellite
        if $currentKmer = kmerDictionary[matchIndex]$ then
            increment $matchCount$
        else        // end of the putative microsatellite
            $motifLength \leftarrow k - matchIndex$
            $tractLength \leftarrow matchCount + k + matchIndex - 1$
            $repeatCount \leftarrow tractLength/motifLength$
            $tractStart \leftarrow transitionStart - tractLength + 1$
            $motif \leftarrow$ substring of $S$ of $motifLength$ starting at $tractStart$
            if $minMotifLength \leq motifLength \leq maxMotifLength$ and $minTractLength \leq tractLength$ and
                $minRepeatCount \leq repeatCount$ then
                report $tractStart$, $motif$, and $tractLength$
            end if
            $matchIndex \leftarrow \emptyset$
        end if
    end if
    for $i = 1$ to $k - 1$ do        // shift all kmers, remove $k^{th}$ kmer
        $kmerDictionary[i - 1] \leftarrow kmerDictionary[i]$
    end for
    $kmerDictionary[k - 1] \leftarrow currentKmer$
end for
```

By using a smaller dictionary of the *k* most recent kmers instead of all possible kmers, we dramatically reduce the compute time necessary to find repetitive sequence. If we are considering microsatellite motif lengths less than or equal to some value of *k*, instead of scanning 4^k dictionary elements at every position in *S*, the algorithm scans 1 position in the M state or *k* in the N state, which reduces our computation time from ~O(4^k) to ~O(*k*).

This detection algorithm immediately and exactly determines microsatellite boundaries in a single step—unlike TRF or lobSTR—and is highly sensitive and specific within the parameters specified. The detection algorithm is rapid—on a single 3.3 GHz CPU, GRCh37 microsatellite detection completes in ~10 minutes, and detection of microsatellites in 1 million 101 bp Illumina paired-end reads takes ~24 seconds.

5.2.1.3 Microsatellite detection algorithm outputs

The uSeq detection algorithm produces four additional files when it is provided with FASTA data as input. The first file is a tab-delimited microsatellite database file that provides information for every microsatellite locus found in the FASTA file, and contains the following fields:

1. Coordinate/chromosome
2. Microsatellite tract start position
3. Microsatellite tract stop position
4. Tract length
5. Motif
6. 5’ flanking sequence
7. 3’ flanking sequence
8. Microsatellite snapshot

In addition to serving as a reference file for later steps in the uSeq pipeline, the microsatellite database provides an easy-to-navigate microsatellite reference for the input sequence. The second file provides a very brief summary of the scanning output—the number of sequences processed, the number of microsatellites found, and how many sequences were printed as output. The third file is a "condensed" version of the input sequence, with all detected microsatellites removed, which enables the input sequence to be used as a reference genome for microsatellite alignments from sequencing data. This algorithm is described in section 3.3. The final output file is the offset index, which enables the uSeq pipeline to reindex microsatellites from the condensed reference genome to the standard reference genome. If the microsatellite database is the only desired output, the condensed reference genome and offset index can be suppressed.

When provided with DNA sequencing data, three files are produced as output. The first file is a FASTQ-formatted read database of all the reads taken as input, with the sequence ID line containing summary data on any microsatellites that were detected. The microsatellite summary format is:

```
<sequence ID>[|<motif>:<start>:<length>:<base quality>(|…)]
```

The second file is a condensed FASTQ-formatted file, with all microsatellites removed, which is compatible to the condensed reference genome during alignment. The third file is a separate header file of microsatellite summaries described above, which uSeq adds back to the reads after alignment. This is necessary due to the variations in the formats produced by different versions of Illumina sequencing technology and in the formats taken as input by different alignment algorithms. By default, the uSeq detection algorithm outputs all reads it receives as input. This behavior can be suppressed, and only reads containing microsatellites—and their mate pairs in paired-end sequencing data—will be output for alignment and downstream processing. This can significantly reduce processing time.

5.2.1.4 Limitations of the detection algorithm

5.2.1.4.1 Detection of imperfect microsatellites

The uSeq detection algorithm does not detect microsatellites interrupted by point mutations or indels. As discussed in the introduction, convergent lines of evidence indicate that the vast majority of microsatellite instability takes place within uninterrupted microsatellite tracts (see, for example, Chung et al. 1993; Eichler et al. 1994; Chong et al. 1995; Goldstein and Clark 1995; Petes et al. 1997). Since our study is focused on identifying *de novo* mutations, our analysis is currently tailored to the detection, alignment, and genotyping of uninterrupted microsatellites.

While we do not expect interrupted microsatellite loci to contribute significantly to *de novo* mutations, such loci may play a larger role in other microsatellite analyses. In population genetic studies, certain haplotypes may have relatively stable interrupted microsatellites, which may serve as distinct markers for specific subpopulations (Jin et al. 1996). As shown in studies of microsatellites in disease, disease incidence is usually reduced in populations with stabilizing variant repeats (Chung et al. 1993; Eichler et al. 1994; Chong et al. 1995). uSeq can detect uninterrupted microsatellite subsequences within longer interrupted microsatellite loci, but comparisons among microsatellite tracts irrespective of any interruptions might streamline microsatellite mutation analyses within a population.

Interrupted microsatellite detection can be incorporated into future versions of uSeq. The simplest means of expanding detection would involve one of two strategies. The first would be somewhat similar to the TRF microsatellite detection strategy—after perfect microsatellites are identified, microsatellites with identical motifs that are closer to each other than would be expected by chance could be combined into longer microsatellite tracts. In addition, microsatellites could be extended if there are nearby motifs identical to those of the microsatellite tract separated by an SNV or short indel. This behavior could be controlled by a matching probability and indel probability, as they are in TRF. A model defining criteria for deciding whether nearby microsatellite tracts are parts of one contiguous microsatellite tract could be derived from the distribution of repetitive sequence within a reference genome, the observation of slippage at microsatellite loci within a population, or a combination of the two. This strategy would require a second step to identify interrupted microsatellite loci after the initial scan for uninterrupted microsatellite loci.

A second option for detecting interrupted microsatellite loci would be to allow for wildcard characters in the kmer dictionary used in uSeq’s microsatellite detection algorithm. This would allow the algorithm to detect near-perfect microsatellites with a tolerance for a certain number of mismatches per kmer. Subsequent processing would ensure that these interruptions are contained within the microsatellite itself and do not extend the locus into adjacent flanking sequence. Similar to the strategy above described in the previous paragraph, statistical analysis would be required to ensure that the tolerance for mismatches and indels maintains the essential biological traits of microsatellite loci.

The SSC dataset could be used to learn more about the traits that distinguish functional microsatellites from highly repetitive regions that are otherwise stable. This could be accomplished by comparing the frequency of *de novo* mutations at interrupted and uninterrupted microsatellite loci. Factors affecting the stability of interrupted and uninterrupted microsatellite loci could include sequence context, motif length, motif composition, and tract length. Using a large study population, such as the SSC, large populations with interrupted and uninterrupted alleles at the same individual locus could be evaluated to gain insight in to the mechanisms influencing their stability.

5.2.1.4.2 Microsatellites below detection threshold

The microsatellite detection method used by uSeq sets minimum thresholds for tract length and repeat number. Although, as discussed in the introduction, microsatellite slippage at shorter microsatellite loci appears to be rare, evidence from MMP tumors indicate that microsatellite mutations can occur in repetitive mononucleotide microsatellites with tract lengths as short as 2 bp (Huang et al. 1996). uSeq can accommodate these microsatellites by simply adjusting detection parameters for minimum tract length and repeat number. Since our study population is assumed to generally have intact MMR, we chose to set a threshold for minimum tract length at 8 bp. Short microsatellites could be particularly relevant when evaluating microsatellite loci with reference tract lengths just exceeding the minimum thresholds set by uSeq. In such cases, alleles below the detection threshold might be missed in some individuals. Additionally, slippage noise introduced during sample preparation will be missed, which would lead to underestimates of noise rates during genotyping. Both of these concerns are considered in detail with regards to the SSC dataset in sections 6.3.1 and 6.3.4.4. In general, these are issues worth considering when setting detection thresholds for microsatellite genotyping studies using uSeq.

5.2.1.4.3 Incompatible microsatellite detection parameters

In the current implementation of the uSeq microsatellite detection algorithm, certain combinations of input parameters are incompatible. When uSeq is provided with a minimum motif length and maximum motif length, *a* and *b*, respectively, the minimum detectable tract length for a microsatellite of motif length j, where $a \leq j \leq b$, is $b + j$. When the minimum detectable tract length for any motif length is longer than the desired minimum detectable tract length, t_{min}, uSeq will warn the user. uSeq will also provide the user the maximum value of b that would still enable detection of microsatellites with motif length j and a minimum tract length of t_{min}. In future versions of uSeq, it would be straightforward to accommodate any t_{min} by allowing uSeq to scan input sequences with multiple dictionaries of different kmer sizes. The current implementation of the uSeq detection algorithm does not have this feature implemented since the study-wide t_{min} exceeds the minimum detectable tract length for all specified motif lengths.

5.2.2 Microsatellite alignment

5.2.2.1 An alignment method robust to microsatellite indels

As mentioned in the previous section, microsatellites found in the reference genome and in sequencing data are condensed during microsatellite detection. The condensation approach aligns microsatellites solely based on their flanking sequence, which is intended to ensure that short read alignment is robust to microsatellite indels. This approach also allows uSeq to use most popular BWT-based alignment algorithms, which are optimized for speed- and memory-efficiency and use paired-end information to improve alignment accuracy, without relying on their indel detection methods. The uSeq alignment algorithm currently uses BWA; although any mapping algorithm with SAM/BAM formatted output can be used (Li and Durbin 2009). uSeq is therefore able to maintain flexibility as sequencing technologies and alignment algorithms continue to evolve.

The uSeq alignment algorithm is designed to complement popular short-read alignment algorithms—a standard alignment algorithm is used to account for non-microsatellite indels, while condensed microsatellite indels are considered separately by uSeq. Many common strategies for detecting indels are inappropriate for detecting microsatellite indels. Algorithms such as bwa aln may have thresholds designed to disfavor long or terminal indels—specifically, capping the possible number of indels per read and their lengths; not calling indels near the ends of reads; and allowing a maximum of *k* differences in an alignment, where *k* is determined as described in the BWA manuscript (Li and Durbin 2009). These heuristics may be appropriate for non-microsatellite indels, but they are incapable of accommodating the more complex and dynamic behavior of microsatellite indels.

Recently, newer algorithms applying a seed-and-extend approach to short-read alignment have emerged, including bowtie2 and bwa mem (Langmead and Salzberg 2012; Li 2013). These algorithms circumvent some of the limitations of earlier alignment algorithms by using a BWT to find exact matches within a sequencing read, which are then used to seed local realignments that can provide more accurate indel calls. However, the gap-opening and gap extension penalties used in local realignment of non-microsatellite indels may be inappropriate for microsatellite indels. Microsatellite indels are more likely to occur than non-microsatellites indels, and microsatellite stability is dependent on motif length, tract length, and motif composition, as discussed in section 4.1.2.3. Within the context of local realignment, a variable gap-opening penalty that depends on the context of the gap may be more appropriate. Microsatellites frequently have significantly divergent lengths from the reference genome and mutate in a step-wise fashion, which implies that standard gap extension penalties may need to be rethought for microsatellites. For example, although 6 bp insertions in microsatellites with 1, 2, 3, and 6 bp motifs tend to occur at different frequencies, a gapped alignment considering the length of an indel rather than the number of motifs inserted would score them equivalently.

5.2.2.2 Mapping algorithm sensitivity

The success of the uSeq's alignment approach depends on the amount of flanking sequence surrounding an observed microsatellite, as well as its uniqueness within a reference genome. To test the accuracy of the uSeq alignment algorithm, we evaluated the amount of flanking sequence necessary to map microsatellites found in GRCh37 with a BWA mapping quality score of at least 30, which would imply a high-quality match. Flanking sequences from 20 to 80 bp long in intervals of 5 bp were obtained from each reference microsatellite locus. The flank length of a reference microsatellite locus was split between its 5' and 3' flanking sequences in 5 bp increments. For example, 20 bp flanks were split into 5/15, 10/10, and 15/5 bp flanks for the 5' and 3' flanking sequence of a reference microsatellite. The flanks were mapped as single-end reads; therefore, they represent a lower bound on achievable mapping accuracy. More than 75% of microsatellites in the human genome can be mapped with 40 bp of flanking sequence (Figure 5.4, page 173). The SSC exome study uses 101 bp Illumina reads, which allows uSeq to accurately map most microsatellites with tract lengths shorter than 66 bp long. The mean microsatellite tract length in hg19 is 12 bp. Additionally, nearly ~90% of microsatellites can be mapped with 80 bp of flanking sequence in single-end reads. In the SSC exome study, uSeq uses paired-end data, which should provide significant improvements over this already considerable accuracy.

5.2.2.3 Mapping algorithm precision

uSeq will occasionally align reads to an incorrect position in the reference genome. To test the precision of the mapping algorithm, and to discern whether there are systematic effects behind incorrect alignments, we simulated 16 million 101 bp paired-end reads from chr22 in GRCh37 using wgsim v0.2.3 with default parameters. After reads were scanned and aligned to the entire condensed reference genome by uSeq, we obtained precision and recall metrics for different subsets of the data. Recall refers to the percentage of correctly aligned reads that exceed a mapping quality threshold. Precision is the percentage of reads above a quality threshold that are correctly aligned to the reference genome. When considering all reads, including those without microsatellites, uSeq has a recall of 82.2% for reads with mapping qualities above 30, and precision of 98.7% (Figure 5.5, page 174). Considering only microsatellite reads, uSeq has somewhat diminished precision and recall—for reads with mapping quality of at least 30, recall is 74.4% and precision is 96.8% (Figure 5.5, page 174). Most importantly for the SSC study, precision is 99.4% when analysis is limited to reference microsatellite loci, with recall of 79.1%, for reads with mapping quality scores of at least 30 (Figure 5.5, page 174). Since uSeq aligns reads based on flanking sequence only, we expect similar precision in real data that will remain robust to the frequency and magnitude of microsatellite indels. We would not expect standard alignment algorithms to maintain consistent precision and recall for unstable microsatellite loci.

This simulation does not include any microsatellite indels, so reads aligned with uSeq are expected to be slightly less accurate than reads aligned to an uncondensed genome using BWA. A comparison of mapping performance at microsatellite reference loci between uSeq and GATK using SSC exome sequencing data can be found in the section 6.2.7. GATK follows an initial BWA alignment with its own indel realignment algorithm. Therefore, any alignment improvement demonstrated for uSeq as compared to GATK would also represent an improvement for uSeq as compared to BWA.

5.2.2.4 Limitations of the alignment algorithm

5.2.2.4.1 Undetected terminal microsatellites

Undetected microsatellites at the 5’ or 3’ end of a sequencing read are a notable contributor to incorrect alignments in uSeq. Since these microsatellites cannot be detected by uSeq, they cannot be condensed (Figure 5.6, page 175). Successfully finding an alignment for a read with an undetected terminal microsatellite depends on the length of the undetected tract and its similarity to the microsatellite’s flanking sequence—some reads may align to an incorrect reference genome position and some may not align at all. Based on our analyses described above, this does not substantially reduce uSeq’s precision at reference microsatellite loci. The primary effects of undetected terminal microsatellites are slightly reduced coverage at some reference microsatellite loci and occasional spurious non-reference alignments, which would confound identification of novel, non-reference microsatellite loci.

Over 98% of incorrectly aligned reads with detected microsatellites are within 75 bp of their correct position in the reference genome (Figure 5.7, page 176). This suggests a potential refinement to our alignment strategy by implementing a local realignment step with a modified Smith-Waterman algorithm allowing for different gapping penalties at microsatellite loci in the region surrounding the candidate alignment position reported by BWA. uSeq does not incorporate a local realignment step since the current focus of the SSC project is on reference microsatellite loci, where precision is already very high. A local realignment approach would not improve precision for unmapped reads.

5.2.2.4.2 Other limitations

uSeq can neither detect nor align reads with interrupted microsatellites or microsatellites below its detection threshold. The current alignment approach should theoretically accommodate short and interrupted microsatellite loci as uSeq's detection capabilities evolve. However, unanticipated complications may arise which would require further development to maintain high precision and recall in uSeq's alignment algorithm.

5.2.3 Post-alignment processing

5.2.3.1 Preparing alignments for reindexing

After reads have been aligned and converted to an unsorted BAM file, alignment start positions are reindexed from the condensed genome to the uncondensed reference genome. Alignments are concurrently matched to their respective sequence IDs that were saved to a header file during detection to reincorporate any relevant microsatellite summaries. If an alignment maps to the reverse complement of the reference genome (i.e. SAM flag bit 0x10 is set to 1), the microsatellite summary is recalculated for the reverse complement of the originally observed tract. Once an alignment has been matched to any relevant microsatellite summaries and any necessary recalibration has been performed, the alignment can then be reindexed to the uncondensed reference genome. Before reindexing, the microsatellite summary from the uSeq detection algorithm is added to each alignment entry in the BAM file as an MU tag with the following format:

```
MU:Z:<motif>:<start>:<length>:<base quality>[|…]
```

Once any microsatellite information has been added to the alignment entry, the alignment length is recalculated. If a sequence ID in the BAM file and header file are not identical, reindexing is immediately terminated and the user is notified of the error. These steps are incorporated into Algorithm 5.4 on page 110.

5.2.3.2 Reindexing alignments to the uncondensed reference genome

Uncondensed coordinates are derived from the start positions provided by BWA and the reference genome offset index generated by uSeq's microsatellite detection algorithm. The offset index contains two vectors of equal length for each chromosome in the reference genome, which match positions in the condensed reference genome to the corresponding cumulative microsatellite bases detected in the uncondensed reference genome. For each alignment with a unique mapping to the reference genome, a modified binary search finds the microsatellite locus immediately preceding the alignment's start position. The pseudocode for the modified binary search is below.

Algorithm 5.2 A modified binary search to find the microsatellie locus immediately preceding an alignment start site

```
Input: $alignmentStart$, the start site reported for an alignment; $positions$, a vector of microsatellite start positions in
    the condensed reference genome for the alignment's coordinate
$first \leftarrow 0$
$last \leftarrow$ length of $positions$
while $first < last$ do
    $mid \leftarrow (last + first)/2$
    if $alignmentStart > positions[mid + 1]$ then        // alignment starts after $positions[mid + 1]$
        $first = mid + 1$
    else if $alignmentStart <= positions[mid]$ then        // alignment starts at or before $positions[mid]$
        $last = mid$
    else        // alignment starts between $positions[mid]$ and $positions[mid + 1]$
        return $mid$
    end if
end while
return $-1 \times (first + 1)$        // alignment starts before first element in $positions$ or after last element in $positions$
```

Once the nearest preceding microsatellite locus has been identified, the number of offset bases needed to reindex an alignment from the condensed reference genome to the uncondensed reference genome is calculated. If the alignment start position does not coincide with a microsatellite locus, the offset is simply the number of cumulative microsatellite bases preceding the read. If the alignment is adjacent to a microsatellite tract, the offset includes the bases of the next microsatellite locus. If an alignment starts within a microsatellite tract, any detected terminal microsatellites are factored into the offset calculation. Pseudocodes for the two algorithms needed to calculate the offset are below.

Algorithm 5.3 Calculating the number of microsatellite bases preceding an alignment start site

Input: $alignmentStart$, the start site reported for an alignment; $positionIndex$, the index of the microsatellite locus immediately preceding $alignmentStart$ (Algorithm 5.2); $offsets$, a vector of offsets for each microsatellite locus from the offset index file; $positions$, as described in Algorithm 5.2; $microsatelliteInfo$, the microsatellite information for any microsatellites in the alignment // Algorithm 5.2 guarantees that $alignmentStart$ will not be greater than the position at $positionIndex + 1$

```
if $alignmentStart < positions[positionIndex + 1]$ then    // $alignmentStart$ is between microsatellites
  return $offsets[positionIndex]$
  if $alignmentStart = positions[positionIndex + 1]$ then    // $alignmentStart$ intersects a microsatellite
    if $microsatelliteInfo \leftarrow \emptyset$ then
      return $offsets[positionIndex + 1]$
    else
      return $offsets[positionIndex + 1]+$ tract lengths of any microsatellites from the 5' end of alignment
    end if
  end if
end if
```

Algorithm 5.4 Finding the offset length for an alignment in a condensed reference genome

Input: BAM file of uSeq alignments; header file(s) with microsatellite information from uSeq microsatellite detection (Algorithm 5.1); $offsetIndex$, a dictionary with a vector of offsets for each coordinate of the condensed reference genome // the BAM file and header file(s) must have the same sequence ID order

```
$alignmentOffsets \leftarrow \emptyset$    // A vector storing the number of offset bases for every alignment
for all $alignment$ in BAM file and $headerSequenceID$ in header file(s) do
  $alignmentSequenceID \leftarrow$ sequence ID for $alignment$
  $alignmentPairInfo \leftarrow$ read pair information for $alignment$
  $headerPairInfo \leftarrow$ read pair information from $headerSequenceID$
  if $alignmentSequenceID \neq headerSequenceID$ or $alignmentPairInfo \neq headerPairInfo$ then
    report error, exit    // the header file and BAM file are not synchronized
  end if
  $alignmentFlag \leftarrow$ SAM bitwise flag for $alignment$
  $alignmentMappingQuality \leftarrow$ $alignment$ mapping quality
  if $alignmentFlag$ has true unmapped bit or $alignmentMappingQuality = 0$ then
    skip $alignment$, proceed to next
    // $alignment$ maps nowhere or to multiple positions, no reason to reindex
  end if
  $alignmentCoord \leftarrow$ coordinate for $alignment$    // chromosome where the alignment was found
  if $offsetIndex[alignmentCoord] = \emptyset$ then
    report error, exit    // $alignmentCoord$ was not found in $offsetIndex$, suggests $alignment$
                          and $offsetIndex$ are based on different reference genomes
  end if
  $microsatelliteInfo \leftarrow$ parsed microsatellite information from $headerSequenceID$
  if $alignmentFlag$ has true reverse strand bit then
    for all microsatellites in $microsatelliteInfo$ do    // reverse complement $microsatelliteInfo$
      reverse microsatellite start and stop coordinates
      update microsatellite motif
      reverse microsatellite tract base quality string
    end for
  end if
  $alignmentStart \leftarrow$ start position for $alignment$
  $positionIndex \leftarrow$ index of microsatellite locus immediately preceding $alignmentStart$ (Algorithm 5.2)
  if $positionIndex > 0$ then
    $currentOffset \leftarrow$number of microsatellite bases preceding $alignmentStart$ (Algorithm 5.3)
  else if $positionIndex = -1$ then
    $currentOffset \leftarrow 0$    // $alignmentStart$ precedes first microsatellite of a coordinate
  else
    $currentOffset \leftarrow$ total microsatellite bases in coordinate    // no microsatellites follow $alignmentStart$
  end if
  append $currentOffset$ to $readOffsets$
end for
return $alignmentOffsets$
```

5.2.3.3 Calculating read and microsatellite start positions

The start position in the uncondensed reference genome is added to the alignment's entry in the BAM file as an OP tag, which reports the position in the original genome as an integer. Once the alignment start position has been determined, the start positions for any microsatellites within the alignment are calculated. The start positions are added to the alignment entry as an MC tag with the following format:

```
MC:Z:<coord>:<motif>:<start>:<length>[|…]
```

Microsatellite start positions are reported for every microsatellite, including non-reference microsatellites and detected terminal microsatellites. Detected terminal microsatellites are distinguished from microsatellites completely contained within a read by the motif's case in the MC and MU tag—detected terminal microsatellite motifs are lowercase, while completely contained microsatellite motifs are uppercase. The pseudocode is below for calculating start positions for the alignment and any microsatellites it contains.

Algorithm 5.5 Converting alignment and microsatellite start positions from condensed to uncondensed genomic coordinates

Input: BAM file of uSeq alignments; *alignmentOffsets*, the number of microsatellite bases preceding every *alignmentStart* (Algorithm 5.4)

```
for all $alignment$ in BAM file do
    $cigarString \leftarrow$ cigar string for $alignment$
    $alignmentFlag \leftarrow$ SAM bitwise flag for $alignment$
    $adjustment \leftarrow 0$    // the alignment algorithm used reports start sites beginning after any soft-clipped sequence. This needs to be confirmed for each alignment algorithm
    if $alignmentFlag$ has false reverse strand bit and $cigarString$ begins with soft-clipping then
        $adjustment \leftarrow$ number of soft-clipped bases at beginning of $cigarString$
    else if $alignmentFlag$ has true reverse strand bit and $cigarString$ ends with soft-clipping then
        $adjustment \leftarrow$ number of soft-clipped bases at end of $cigarString$
    end if
    $alignmentStart \leftarrow$ start position for $alignment$
    $offset \leftarrow$ offset for read from $alignmentOffsets$
    $originalPosition \leftarrow alignmentStart + offset - adjustment + 1$
    $microsatelliteInfo \leftarrow$ microsatellite information for $alignment$
    for all $microsatellite$ in $microsatelliteInfo$ do
        $microsatelliteStartPosition \leftarrow$ start position of $microsatellite$ in $alignment$
        $genomeStartPosition \leftarrow alignmentStart + microsatelliteStartPosition + offset - adjustment$
    end for
    report $originalPosition$ and $genomeStartPosition$ for $alignment$
end for
```

5.2.3.4 Subsequent post-processing

Once alignments are combined with their microsatellite information and have been reindexed to the uncondensed reference genome, they are then sorted by position in the condensed reference genome. After sorting, read group IDs are added to each BAM file that uniquely identify reads from each sequencing file by the flowcell ID, lane number, and SSC individual ID. The only files generated in post-processing are the reindexed, sorted BAM file and a counts file from the reindexing program. At this point, any files remaining from detection and alignment are removed or compressed.

5.2.4 Marking PCR replicates and profiling microsatellite loci

5.2.4.1 Marking replicates

Rather than refer to read pairs coming from the same genomic fragment as PCR duplicates, we prefer the term PCR replicates. While most PCR replicate sets only contain two read pairs, more than two read pairs can be derived from one genomic fragment. Therefore, “PCR replicate” more accurately portrays the phenomenon being characterized.

Prior to marking replicates, BAM files from each lane of sequencing data for a single individual are merged. Merged BAM files are used to mark replicates since all lanes of sequencing data for any individual derive from a single library preparation. PCR replicates can therefore occur in separate lanes or flow cells from the same individual within the SSC study.

uSeq uses Picard’s MarkDuplicates to mark PCR replicates in the merged BAM file (http://picard.sourceforge.net). uSeq has modified criteria for replicate removal that differ from the approach used by samtools rmdup, which keeps the alignment with the highest mapping quality score. uSeq first checks to ensure that the PCR replicate alignments all report the same tract length. If all of the replicate alignments have concordant tract lengths, the alignment with the highest mapping quality is used for genotyping. If any replicate alignment has a discordant tract length, no alignments are reported. Since we assume that all alignments within a replicate came from the same genomic fragment, discordant tract lengths could only be observed if a subset of them underwent slippage during sample preparation or sequencing. Since it is impossible to determine the original genomic tract length, none of the replicate alignments are considered. PCR replicates represent another means of analyzing microsatellite slippage during sample preparation, which could be useful in future extensions of uSeq that focus on individuals or very small populations. Therefore, uSeq reports summaries for every PCR replicate set in every individual. The pseudocode for PCR replicate handling is below.

Algorithm 5.6 Specialized uSeq handling of PCR replicate sets

Input: For the N alignments for a microsatellite locus: *alignmentFlags*, their BAM flags; *alignmentStarts*, their alignment start positions; *mateStarts*, the alignment start positions for their mate pairs; *tractLengths*, their observed tract lengths; *mappingQualities*, their reported mapping qualities

```
$observedReplicatePositions \leftarrow \emptyset$          // a vector of paired alignment and mate start positions
for $i = 0$ to $N$ do
  if $alignmentFlags[i]$ has true duplicate flag bit and the pair ($alignmentStarts[i]$, $mateStarts[i]$) is not in
  $observedReplicatePositions$ then
    $replicateSetIndices \leftarrow \emptyset$
    add $i$ to $replicateSetIndices$
    for $j = 0$ to $N$ do          // find all other alignments with same alignment and mate start positions
      if $j \neq i$ then
        if ($alignmentStarts[i] = alignmentStarts[j]$ and $mateStarts[i] = mateStarts[j]$) or ($alignmentStarts[i] =$
        $mateStarts[j]$ and $mateStarts[i] = alignmentStarts[j]$) then
          add $j$ to $replicateSetIndices$
        end if
      end if
    end for
    if $tractLengths[replicateSetIndices]$ are all equivalent then
      report alignment with the maximum value in $mappingQualities[replicateSetIndices]$
      suppress reporting of all other alignments in $replicateSetIndices$
    else
      suppress reporting of all alignments in $replicateSetIndices$
    end if
    add ($alignmentStarts[i]$, $mateStarts[i]$) to $observedReplicatePositions$
  end if
end for
```

5.2.4.2 Managing overlapping read pairs

In the course of designing and evaluating uSeq, we have observed many instances of overlapping read pairs that report the same microsatellite locus in both fragments. uSeq has a specific set of rules it applies to these overlapping read pairs. Since both microsatellite observations are derived from one sequencing library fragment, they should have identical tract lengths. When this is true, the first observation of a microsatellite locus is reported from the overlapping read pair. When an overlapping read pair has discordant tract lengths, neither read is included in the microsatellite profile, and the read pair is added to a BAM file containing all discordant overlapping read pairs. The pseudocode for managing overlapping read pairs is below.

Algorithm 5.7 Specialized uSeq handling of microsatellites in overlapping read pairs

Input: For the N alignments for a microsatellite locus: $sequenceIDs$, their sequence IDs; $tractLengths$, their observed tract lengths; $motifs$, their observed microsatellite motifs; $genomeStarts$, the genome start positions for the microsatellite reported in each alignment (Algorithm 5.5)

```
$observedSequenceIDs \leftarrow \emptyset$
for $i = 0$ to $N$ do
  for $j = 0$ to $N$ do                                // look for any alignments with same sequence ID
    if $sequenceIDs[i] = observedSequenceIDs[j]$ then
      if $motifs[i] \neq motifs[j]$ or $tractLengths[i] \neq tractLengths[j]$ or $genomeStarts[i] \neq genomeStarts[j]$ then
        suppress reporting of both alignments in read pair
      else
        report first alignment observed from read pair, suppress second alignment
      end if
    end if
  end for
  if $sequenceIDs[i]$ was not in $observedSequenceIDs$ then
    add $sequenceIDs[i]$ to $observedSequenceIDs$
  end if
end for
```

5.2.4.3 Additional read filtering

In addition to marking PCR replicates and resolving overlapping read pairs, several other criteria are considered as reads are incorporated into microsatellite profiles. The following criteria are used to filter reads:

1. Unmapped reads are excluded (during reindexing)

2. Alignments must exceed a minimum mapping quality threshold—the default for uSeq is 30
3. Both alignments in a read pair must map to the same chromosome
4. Terminal microsatellites are excluded
5. Microsatellite motifs may only contain standard nucleotides (A, C, G, and T)
6. If an alignment maps to a reference microsatellite locus, its motif must match the reference motif

These filters are incorporated into Algorithm 5.8 on page 117.

5.2.4.4 Flanking sequences

When profiling microsatellite loci, uSeq also tracks the flanking sequence surrounding the microsatellite locus in each alignment. This is particularly useful when distinguishing slippage events from SNVs or non-microsatellite indels that extend or contract microsatellite sequence. If a microsatellite mutates via slippage, all alleles should have identical flanking sequence. If a microsatellite mutation is not due to slippage, the microsatellite flanking sequence is likely to differ (Figure 5.8, page 177). Each microsatellite profile reported by uSeq has at least two entries. The first entry reports the observed tract lengths for all flanking sequences observed at the locus. Subsequent entries report all observed tract lengths for every unique flanking species encountered at the microsatellite locus.

Each microsatellite profile entry has an associated flank sequence and flank ID. The first flank sequence is always "all" and its flank ID is always "0". Each flanking species is described by the 5 bp on either side of the observed microsatellite and its flank ID is determined by the total coverage within the species—flanks with more coverage are closer to

“0”. The flank sequence is padded with “#”s if a microsatellite starts or ends within 5 bp of a reads’ boundaries (e.g. ##AGAATGTT).

5.2.4.5 Profiling microsatellite loci

Once uSeq read filtering has completed, microsatellite profiles can then be assembled. The profiler software maintains a profile for every microsatellite locus covered by at least one alignment, even if the locus is not found in the reference genome. By default, uSeq will only report microsatellite profiles covered by at least two alignments. The microsatellite profile is a tab-delimited file with the following fields:

1. Coordinate/chromosome
2. Start position
3. Motif
4. Reference tract length (0 for non-reference microsatellites)
5. Total number of flanking species at a microsatellite locus
6. Current flank ID
7. Current flank sequence
8. Total coverage supporting the current flank
9. Comma-delimited read count distribution
10. Comma-delimited supporting sequence IDs

The read count distribution is one-indexed. Therefore, the number at a particular index in the distribution provides the coverage for the allele tract length given by that index. For example, if there are 5 reads supporting an 8 bp allele in a read count distribution, the 8th index position will have the number 5. The complete pseudocode for profiling microsatellite loci is below.

Algorithm 5.8 Creating profiles for microsatellite loci

Input: *BAM*, an input BAM file; *microsatelliteDatabase*, a database of reference microsatellite loci; *minCount*, the minimum number of reads a profile must have to be reported; *minMappingQuality*, the minimum mapping quality for a read; and *flankLength*, the amount of flanking sequence to record from each observed microsatellite

```
profileDictionary $\leftarrow \emptyset$
lastCoordinate $\leftarrow \emptyset$
for all alignment in BAM do
  currentCoordinate $\leftarrow$ coordinate of alignment
  alignmentFlag $\leftarrow$ SAM bitwise flag for alignment
  if alignmentFlag has true unmapped bit then
    skip alignment
  end if
  alignmentMappingQuality $\leftarrow$ mapping quality for alignment
  if alignmentMappingQuality <= minMappingQuality then
    skip alignment
  end if
  mateCoordinate $\leftarrow$ mate pair coordinate of alignment
  if currentCoordinate $\neq$ mateCoordinate then        // read pairs aligned to different chromosomes
    skip alignment
  end if
  microsatelliteInfo $\leftarrow$ all microsatellites observed in alignment
  for all microsatellite in microsatelliteInfo do
    microsatelliteMotif $\leftarrow$ motif from microsatellite
    if microsatelliteMotif contains any nucleotides aside from A, C, G, or T then
      skip alignment
    end if
    keyName $\leftarrow$ unique identifier for locus generated from microsatellite
    check if the mate pair of alignment was already added to this microsatellite locus (Algorithm 5.7)
    profileFlank $\leftarrow$ flankLength bp from each side of microsatellite
    if microsatellite is within flankLength of alignment boundary then
      pad flank with ”#”’s
    end if
    add microsatellite and profileFlank to profileDictionary[keyName]
  end for
  if lastCoordinate $\neq$ currentCoordinate then        // print all profiles when switching chromosomes
    for all microsatelliteProfiles in profileDictionary do
      if microsatelliteProfile matches a locus in microsatelliteDatabase then
        profileMotif $\leftarrow$ motif in microsatelliteProfile
        databaseMotif $\leftarrow$ databaseMotif motif for reference locus in microsatelliteDatabase
        if profileMotif $\neq$ databaseMotif then
          do not report microsatelliteProfile
        end if
      end if
      profileSize $\leftarrow$ number of alignments in microsatelliteProfile
      if profileSize $\geq$ minCount then
        remove PCR replicate sets (Algorithm 5.6)
        report microsatelliteProfile for locus as a whole
        report microsatelliteProfile for each unique profileFlank
      end if
    end for
    profileDictionary $\leftarrow \emptyset$
  end if
  lastCoordinate $\leftarrow$ currentCoordinate
end for
```

5.2.5 uSeq implementation for large datasets on a distributed computing cluster

5.2.5.1 uSeq component implementation

The current implementation of the uSeq pipeline is tailored to process SSC data using SGE in a distributed cluster environment where each node has its own RAM and disk space. uSeq must be able to process large numbers of samples simultaneously and efficiently, so it has been designed to maximize processing speed while minimizing the number of threads used per operation, RAM used by each process, and disk space requirements. Therefore, all of the sequence-processing components of the uSeq individual module are written in C++, which allows for effective memory management and rapid processing times. All uSeq programs are single-threaded to maximize the number of samples that can be run on each node. Users do have the option of running the alignment step using two threads for bwa aln, which is the current uSeq default. The pipeline manager is implemented in Perl and handles maintenance operations such as logging, error reporting, file management, and tracking the time used by pipeline components.

5.2.5.2 Data distribution

Since the SSC data was processed on a distributed memory cluster, the uSeq pipeline individual module is designed with special attention to data distribution. uSeq programs accessing a sequencing file are run on the node where the file is stored and most outputs are written locally. This cuts down on computation time by eliminating many costly NFS operations that would be necessary to move, read, or write files between nodes. Since data is stored throughout the cluster, uSeq maintains a master directory on one node, which primarily consists of symbolic links to files distributed throughout the cluster nodes. The only files that are ever copied between nodes by uSeq are small files—such as count or log files—and microsatellite profiles. All microsatellite profiles are stored on the same node since they must all be accessed simultaneously in the uSeq pipeline population module.

5.2.5.3 uSeq modifications for high-traffic computing clusters

The cluster used by uSeq to process the SSC dataset is under constant use by many individuals within the CSHL community. Therefore, uSeq must be controlled enough that it can run within a high-traffic cluster environment without wreaking havoc. In addition to the adjustments described above, three additional adjustments to uSeq have been implemented. First, each component of the individual module is submitted as a separate job to the SGE system that manages cluster access. This ensures that uSeq does not monopolize the cluster for hours at a time. Second, when running uSeq on large populations such as the SSC, the number of individuals being processed simultaneously is capped. This ensures that the number of uSeq-generated SGE jobs running on the cluster does not flood the system. Finally, uSeq limits the amount of virtual memory available to Picard's MarkDuplicates. In our experience, when this restraint is not specified, MarkDuplicates can take up significant amounts of RAM on a node.

5.2.5.4 Pipeline accessory files

The uSeq pipeline generates two accessory file types during execution that track the commands run by the pipeline and their respective outputs. The first accessory file is the log file, which tracks the parameters passed to the uSeq individual module and its components. The log file also reports the system calls, SGE job submission commands, and file management tasks performed by any pipeline component. If an error is encountered in a component, it will be reported in the appropriate log file. Log files dramatically simplify error and file tracking in the SSC. Errors that occur in one instance of uSeq in one individual can be easily found, their sources can be readily identified, and the offending commands can be reproduced and debugged. This allows uSeq to scale up to large populations while maintaining robust tracking of any errors.

uSeq also maintains a counts file for each component in the pipeline, which reports the number of input and output reads, as well as any other relevant counts. For instance, during reindexing, uSeq reports the number of reads that align to multiple genomic positions or that fail to align to any genomic positions. Tracking input and output reads between pipeline components is another effective way of ensuring that the uSeq pipeline is running smoothly.

5.3 uSeq population module

5.3.1 Creating a population profile summary table

In order to take specific advantage of the population level data from the SSC, uSeq must summarize and merge the microsatellite profiles of every individual within the population before making genotype calls. The summarizer takes the master directory mentioned in the previous section as input, finds all microsatellite profiles in the directory, and synchronously proceeds through the microsatellite profiles of the entire population. Currently, the summarizer only reports reference microsatellite loci. The output of the summarizer is a large tab-delimited matrix with as many rows as there are unique reference loci, and as many columns as there are individuals in the study population. Each entry reports the total coverage for one individual at one locus within the population. If a person has any coverage, the format of their entry is:

```
<# tract lengths observed>;<tract lengths>;<coverage>
```

The tract lengths and their coverage are both comma-delimited lists. If no coverage is observed for an individual at a locus, their entry is simply “0;;”. Currently, the population profile summary does not consider individual flanks. Due to the complexity of this information and the huge amount of data already being processed, we are still evaluating options to incorporate this information into the summary table.

In addition to the summary table, the uSeq summarizer produces two index files—a locus index for the table rows and a person index for the table columns. The locus index is a tab-delimited file with the following fields:

1. Coordinate/chromosome
2. Start position
3. Motif
4. Reference tract length
5. Number of people with any coverage
6. Maximum individual coverage
7. Aggregate coverage in population

The person index file is another tab-delimited file with the following fields:

1. SSC individual ID
2. SSC family ID
3. Relationship to proband

While the locus index would remain largely identical independent of study design, the person index file is study-specific. For instance, in a case-control study, it might simply contain an ID and affected status for every individual. Additional information can be added to either index file as needed.

The summarizer is currently implemented stably in Python using code generously supplied by Dan Levy. An implementation in C++ has been written, but has not been tested on the complete SSC population. The Python implementation can process the entire SSC population in several days to approximately one week. Recently, we have begun to compress profiles for each SSC individual to reduce their disk space footprint. This modification nearly doubles the Python summarizer's processing time. Re-implementing the summarizer in C++ is intended to reduce the profiler's running time, especially in larger study populations.

5.3.2 Identifying well-covered loci

Since uSeq relies on population information to make genotype calls, when there is insufficient study-wide coverage at a locus, uSeq does not attempt to call genotypes. This is particularly relevant in targeted capture studies such as the SSC exome study. Although sequencing coverage is targeted to exons, off-target coverage is frequently observed. A discussion of off-target and on-target coverage at reference microsatellite loci in the SSC dataset can be found in the section 6.4.1. In order to maximize uSeq's genotyping accuracy, we limit the loci at which genotypes are called by requiring that a locus must be observed in at least 60% of the study population and have a maximum individual coverage of at least 25. These are not hard thresholds, and can be modified to consider any subsample of the loci observed within the population.

5.3.3 Genotyping microsatellite loci

5.3.3.1 Parameters modeled by uSeq

The uSeq genotyper is designed to call high accuracy microsatellite genotypes with consideration for the specific challenges that complicate microsatellite analysis in exome sequencing data. Microsatellite slippage errors during sample preparation are one complication that has received considerable attention in other published HTS microsatellite genotypers (Gymrek et al. 2012; Highnam et al. 2013). In addition to estimating a per-locus error rate, uSeq also estimates per-allele, per-person coverage and per-allele bias. The genotyper consists of the following stages (Figure 5.9, page 178):

1. Estimate per-allele, per-locus coverage using SVD
2. Initialize estimates of allele bias and locus noise rate
3. Call genotypes using bias and noise estimates
4. Use EM to update locus noise rate and allele bias estimates
5. Calculate model likelihood with new estimates

Steps 3 through 5 are iterative, and final genotypes are derived from the EM once the model likelihood has converged, i.e. the estimates for allele bias and locus noise rate appear to describe the population at the locus as well as our model allows.

5.3.3.2 Per-allele, per-person expected coverage

5.3.3.2.1 Motivations for an expected coverage estimator

We recognized the need for an expected per-allele, per-person coverage estimator after our initial attempts to call microsatellite genotypes. This was motivated by our knowledge of the limits of uSeq’s microsatellite detection; observed coverage variability in sequencing data; and the possibility of microsatellite polymorphism occurring within CNVs. While uSeq may not detect interrupted or short microsatellites, that does not prevent them from occurring within the datasets being analyzed. Similarly, even when both alleles are detected by uSeq, insufficient coverage may reduce our confidence that both alleles have been observed in an individual at a locus. Finally, in its current implementation, uSeq is designed to make bi-allelic genotype calls. These calls can only be considered accurate if there is no evidence of CNVs at the microsatellite locus. Estimates of per-allele, per-person expected coverage capture our uncertainty about the observed coverage at a locus. They also allow uSeq to distinguish stochastic coverage variability from other effects that would severely curtail our ability to call genotypes accurately.

The need for a coverage estimator also became apparent during our initial attempts to call *de novo* mutations in the SSC dataset. Occasionally, coverage among family members at a locus would be extremely variable. In these situations, it is essential to distinguish whether this coverage variation is stochastic, systematic, or due to a CNV in one or more family members. If the coverage variation were due to a CNV, uSeq would not be able to reliably detect any *de novo* mutations that have occurred at the locus.

5.3.3.2.2 Designing a coverage estimator

In the initial publications discussing the Illumina sequencing technology, the genomic distribution of read coverage is random, and can be modeled as a Poisson distribution with a rate equal to the mean coverage (Bentley et al. 2008). Due to the overdispersion that is generally observed in sequencing data, a negative binomial model might be more appropriate (Anders and Huber 2010). However, both models assume that sequencing reads are sampled randomly from the sequencing library, and that each position in the genome is roughly equally likely to be sampled. Exome sequencing violates these assumptions—reads in or near exons are much more likely to be sequenced. Additionally, probe affinity will lead to non-uniform coverage among exons. Therefore, estimating expected coverage is not simply a matter of considering randomly sampled sequences from a biased sequencing library.

Our initial per-allele, per-person expected coverage was estimated by minimizing a least squares objective function for two vectors of parameters. The first parameter vector contained per-person terms, which described each person's average coverage. The second parameter vector contained per-locus multipliers reflecting the capture efficiency of each locus. While this approach had the benefit of simplicity, it did not perform well as an estimator. The poor performance of the two-parameter model was due in part to complex coverage variability—coverage is non-uniform in SSC individuals and in well-covered loci (Figure 5.10, page 179). This variation might be partially attributed to several sources, including variability in sample preparation or exome capture kits.

5.3.3.2.3 Estimating expected coverage using SVD

In an effort to better capture coverage variability at microsatellite loci in the SSC data, we used the SVD of the total allele coverage matrix to create a low-rank expected coverage estimator. The allele coverage matrix is a $m \times n$ matrix, where m is the number of loci being genotyped and n is the population size. Each entry in the matrix is the total coverage for an SSC individual at a microsatellite locus, split in half. The values are split in half to give a per-allele coverage estimate, and depend on the assumption that most loci in most individuals of the SSC have two copies of each locus. There is no reason to assume that SSC individuals have extensive CNPs.

SVD is a means of separating an input matrix into the orthogonal singular values that define its composition, akin to principal components in PCA. The greatest singular values contribute the most amount of variation to the matrix. Due to their orthogonal nature, each singular value describes a distinct axis of variation within the dataset. In our case, these singular values may represent any combination of a number of factors affecting the allele coverage matrix, including the number of reads sequenced in an individual, GC content at different loci, variations in sample preparation between technicians, or other systematic variability. As the singular values decrease in magnitude, their contributions to variability in overall allele coverage become less significant. These smaller contributions could be due to stochastic processes, such as Poisson sampling; or they could be due to biological effects that are specific to a small number of individuals within the population, such as CNPs. We estimate per-allele, per-person expected coverage by creating a k-rank approximation of the input allele coverage matrix. This will recover the k components that most strongly influence coverage variability, while limiting contributions from stochastic sampling or occasional biological variability.

More specifically, SVD is a factorization of a $m \times n$ matrix M of the following form:

$$M = U\Sigma V^T$$

U is a $m \times m$ matrix of the eigenvectors of MM^T; V is a $n \times n$ matrix of the eigenvectors of M^TM; and Σ is a $m \times n$ diagonal matrix of the square roots of the eigenvalues of U and V, which are called singular values. The singular values in Σ are in descending order, and correspond to the magnitude of the effect of an eigenvalue and corresponding eigenvectors to the values observed in M. All eigenvectors of U and V are orthogonal to each other. We derive $\widehat{M}$, a k-rank approximation of the input matrix M, by taking the first k eigenvalues of Σ and the first k eigenvectors of U and V.

5.3.3.2.4 Identifying significant singular values

The SVD contains every singular value that contributes to our observed matrix—if $\widehat{M}$ were composed of every eigenvalue and eigenvector from the SVD of M, the two matrices would be equivalent. Therefore, as we incorporate more eigenvalues into $\widehat{M}$, we increase the risk of over-fitting our low-rank approximation, and reduce its effectiveness as a coverage estimator. Determining the rank for our coverage estimator matrix requires an algorithm that identifies the threshold for significant eigenvalues in the SVD of any input matrix. We determine the rank of our expected coverage matrix by comparing the singular values of our observed allele coverage SVD to the singular values derived from the SVD of a randomized matrix containing the same values as the allele coverage matrix. Any singular values that exceed their randomized counterparts are considered significant. The pseudocode for identifying significant eigenvalues is below.

Algorithm 5.9 Picking significant eigenvalues from the SVD of an allele coverage matrix

```
Input: $N$, the population size; $M$, the number of loci; $totalCoverage$, an $N \times M$ matrix reporting the total coverage for
    each person at each locus; $nRand$, total randomizations performed to find significant eigenvalues; $nEigen$, maximum
    number of eigenvalues from SVD                                  // this algorithm requires the R package IRLBA
$alleleCoverage \leftarrow totalCoverage/2$                // this assumes most loci in $totalCoverage$ are bi-allelic
$alleleCoverageSVD \leftarrow$ SVD of $alleleCoverage$         // using IRLBA, taking $nEigen$ total eigenvalues
$alleleCoverageEigenvalues \leftarrow$ eigenvalues from $alleleCoverageSVD$
$shuffledEigenvalues \leftarrow nRand \times nEigen$ matrix to store eigenvalues from each matrix randomization
for $i = 1$ to $nRand$ do
    $shuffledCoverage \leftarrow$ sampled with replacement from $alleleCoverage$
    // same dimensions as $alleleCoverage$
    $shuffledSVD \leftarrow$ SVD of $shuffleCoverage$          // using IRLBA, taking $nEigen$ total eigenvalues
    $shuffledEigenvalues[i][1 : nEigen] \leftarrow$ eigenvalues from $shuffledSVD$
end for
for $i = 1$ to $nEigen$ do
    if $alleleCoverageEigenvalues[i] < shuffledEigenvalues[1 : nRand][i]$ then
        return $i - 1$        // all eigenvalues larger than their shuffled counterparts are significant
    end if
end for
```

5.3.3.2.5 Reducing expected coverage calculation time by partial SVD

A complete SVD factorization of a very large matrix such as the SSC allele coverage matrix would be time-consuming and pointless. Since we only need a *k*-rank approximation of the input matrix, we can cut down on the time needed to decompose our input matrix by performing a partial SVD. Fortunately, SVD is a common tool for the analysis of very large matrices, and very talented mathematicians have designed fast and accurate partial SVD algorithms. The uSeq genotyper is implemented in R, and therefore it is able to use IRLBA SVD decomposition (Baglama and Reichel 2014). IRLBA is most famous for its use in the winning Netflix Prize algorithm, which used the algorithm to create an improved predictive model for Netflix's movie recommendation and ratings system. In the context of uSeq, IRLBA is used to calculate the partial SVD of the original allele coverage matrix and the randomized coverage matrices. In the current implementation, uSeq only calculates the first 50 singular values.

5.3.3.2.6 Accurate recovery of expected coverage parameters

We designed a simulation to test the ability of our expected coverage estimation algorithm to accurately recover input coverage parameters in the presence of stochastic noise and biological variation. In our simulation, we took a k-rank matrix derived from a subsample of our real coverage data to be used as Poisson rate parameters. In addition to Poisson sampling noise, we introduced CNVs, which would increase or reduce coverage for our test individuals at random loci. We then compared the SVD-derived expected coverage estimators from our algorithm to the input Poisson rate parameters provided to the simulator (Figure 5.11, page 180). We also ensured that the rank of the SVD was the same as that of the input Poisson rate matrix.

Our simulations demonstrate the success of our expected coverage estimation algorithm. Even when 5% of loci in the simulation have CNVs, the expected coverage estimators are nearly equivalent to the input Poisson rate parameters. The Pearson correlation coefficient between the expected coverage estimators and the input coverage parameters is >0.99, and the RMSE is 0.52 (Figure 5.12A, page 181). This demonstrates the estimators' robustness to stochastic sampling noise. The relationship remains true for individuals at loci with CNVs—correlation is >0.99 and RMSE is 0.92 (Figure 5.12B, page 181). Therefore, the estimators are robust to considerable amounts of biological variation. In real data, we cannot compare our expected coverage estimators to input Poisson rate parameters, but we can compare them to our observed coverage. We observe a strong correlation between simulated observed coverage and the expected coverage estimators, but we also observe that RMSE is much larger than it is for input Poisson parameters and expected coverage estimators (Figure 5.13, page 182). In the context of the simulation, an increased RMSE is due to Poisson sampling. In real data, we might expect to see somewhat lower correlation and higher RMSE than we observe in simulated data.

There is one other demonstration of the power of our expected coverage estimation algorithm that came about due to a bug in an earlier version of the uSeq pipeline. If we were to divide an individual's observed coverage by their expected coverage estimator, we would expect to see that the resulting distribution would be distributed around 2, since most alleles have two copies plus some stochastic sampling variation. In an earlier version of the pipeline, we observed that some individuals had significant deviations from this expected distribution in contiguous blocks of microsatellite loci throughout their genomes. When we evaluated the sequencing data from individuals with these aberrant distributions, we found that uSeq had deleted contiguous chunks of sorted, aligned reads, which made it appear as though these individuals had enormous deletions at microsatellite loci. Due to the power of this expected coverage estimation algorithm, we were able to identify and correct the bug before it significantly damaged our analysis.

5.3.3.2.7 Limitations of the current expected coverage estimation algorithm

The expected coverage estimation algorithm relies on the assumption that the vast majority of the genomes in the sample are bi-allelic. This is a safe assumption in the SSC dataset, but it would need to be reconsidered in some other studies, such as copy number unstable tumors. The expected coverage estimation algorithm could be adapted to datasets with high frequency CNVs by incorporating data from an orthogonal copy number prediction algorithm. Since uSeq focuses on short, dispersed microsatellite loci, it is incapable of accurate copy number prediction on its own. The expected coverage estimator is designed to weigh the evidence supporting a bi-allelic state for every SSC individual at every locus, nothing more.

5.3.3.2.8 Additional coverage estimation algorithm implementation details

Currently, the coverage model estimates expected coverage for autosomes. Sex chromosomes are not evaluated since their copy number varies by gender, although this could be easily accommodated in future versions of uSeq. No coverage estimators are provided for the mitochondrial genome. There is no reasonable assumption to be made regarding the copy number of an individual's mitochondrial genome since the number of sequenced mitochondrial genome copies will vary among individuals.

Once the low-rank SVD coverage estimator matrix has been derived, uSeq eliminates coverage outliers within the study population. Families with at least one member who has total coverage more than two standard deviations below the population mean total coverage are excluded. In addition, families with at least one member who has poor correlation between their observed total coverage and their estimated expected coverage are excluded. The current threshold for poor correlation is 0.8. In practice, these rules only exclude a handful of SSC families from further analysis.

5.3.3.3 Per-allele capture bias and per-locus noise estimation

5.3.3.3.1 Motivations for per-locus noise rate estimation

Typically, HTS and sample preparation introduce noise in the form of base miscalls and more rarely, spurious indels. In addition, replication slippage is possible at a microsatellite locus any time DNA is replicated, even during *in vitro* replication such as PCR. While cellular DNA replication can correct slippage mutations via the MMR machinery, *in vitro* DNA replication has no equivalent mechanism for slippage repair. Therefore, microsatellite instability can be dramatically higher during sample preparation than it would be during cellular replication. The factors underlying cellular microsatellite instability discussed in section 4.1.2.3 also affect microsatellite instability during PCR. Accurate estimates of microsatellite instability rates during sample preparation are one essential component of a good microsatellite genotyper.

Microsatellite loci with similar properties can have variable stability during sample preparation. The most popular microsatellite genotypers estimate noise rates for a microsatellite locus based on its general properties using fixed models that are independent of experimental setup. This general approach might be adequate when genotyping microsatellites in individuals, but it can fall short when experimental conditions affect the slippage rate relative to the genotyper model; or when the properties accounted for by the model are insufficient to accurately capture the slippage behavior of a particular locus. The uSeq approach to modeling sequencing noise addresses both of these problems. In large datasets such as the SSC, the population size at any locus is sufficient to estimate locus-specific noise parameters. These estimates will be much more sensitive to slippage variability among similar loci. Additionally, uSeq always estimates noise rates using study-specific coverage data. Assuming consistent experimental protocols, uSeq will be able to estimate study-specific noise rates.

Per-locus noise estimates could be even more fine-grained. In theory, noise rates can vary among people at a locus. Currently, uSeq assumes that there is no significant intra-person variation in locus-wide noise. Alleles at the same locus may also have different noise rates. Since we have no clear means of distinguishing which allele produced a particular slippage product, uSeq only infers a locus-specific noise rate estimate, accounting for slippage mutations as well as base miscalls and spurious indels. In addition, allele-specific error rates could be hampered for alleles with few observations within the study population. Locus-specific noise rate estimates are sufficient to call accurate genotypes and identify *de novo* mutations.

5.3.3.3.2 Motivations for per-allele allele capture bias

In our initial attempts to call *de novo* microsatellite mutations in the SSC, we observed that exome capture probes favored reference alleles at many loci. This capture bias affected both the sensitivity and specificity of uSeq *de novo* calls. In allele scatter plots of some microsatellite loci, it is immediately obvious that different alleles at the same locus are consistently recovered and sequenced with varying efficiency (Figure 5.14, page 183). Per-allele capture bias has also been observed for indels in the broader SSC study as well

(Iossifov et al. 2012; see Table S2). Without an estimator for per-allele capture bias, accurate detection of *de novo* microsatellite mutations would be impossible.

Like the per-locus noise rate estimates, per-allele capture bias estimates might not have generalizable patterns. The bias can depend on any number of factors that are exceedingly complex or impossible to discern. Possible contributors to capture bias may include the position of a microsatellite locus relative to the exome capture probe; the difference between the capture probe microsatellite tract length and the library fragment microsatellite tract length; or the ability of a non-reference microsatellite allele to maintain a secondary structure favorable to a stable interaction with the exome capture probe. Capture bias may also be a general feature of microsatellite genotyping, independent of exome capture. For example, long microsatellite alleles will have less coverage than short microsatellite alleles at the same locus because the sequencing read has fewer opportunities to completely contain the longer microsatellite allele. This limitation would diminish as the read lengths possible in HTS become longer.

5.3.3.3.3 A custom EM to estimate allele capture bias and noise parameters while calling microsatellite genotypes

5.3.3.3.3.1 EM overview

The uSeq genotyper uses a custom EM algorithm that simultaneously solves for MLEs for per-locus noise and per-allele bias parameters. In a typical Gaussian EM, we assume that the observed data is derived from a mixture of several Gaussian distributions with unknown parameters. The identity of the distribution from which a point is drawn is defined by an unknown latent variable. Since we do not know which of the distributions a point is drawn from, we can only estimate the membership probability for each point in every distribution—ideally, a point's source distribution should have the highest membership probability. The EM requires a log-likelihood function relating the distribution parameters and membership probabilities to our observed data. This log-likelihood function provides MLEs for each unknown distribution parameter, which the EM iteratively updates, along with membership probabilities, until the model converges.

Before the EM algorithm can begin, we must specify initial estimates for membership probabilities and distribution parameters. Membership probabilities for each data point typically start as a discrete uniform distribution—in the absence of distribution parameter estimates, all distributions are equally likely to be the source of a data point. Initial distribution parameter estimates may sometimes be intuitive, while other times they might be randomly selected. The EM can also be repeated several times with different initial estimates for distribution parameters to ensure that their MLEs are global maxima, not local maxima.

A single iteration of the EM consists of two steps: the E—or expectation—step, and the M—or maximization—step. During the E step, membership probabilities for each data point are estimated. The membership probability is influenced by the latest distribution parameter MLEs, as well as each distribution's weight—the likelihood that any point is drawn from that distribution. A distribution with low weight is unlikely to produce many observations. During the M step, the distribution weights and MLEs for each distribution parameter are updated. The distribution weights are simply the mean membership probabilities for every distribution. The MLEs are calculated from the partial derivatives of the log-likelihood function for each distribution parameter. The derivation of the log-likelihood equation, MLEs, and membership probabilities will be elaborated on in our discussion of the custom EM we developed for the uSeq genotyper.

The EM algorithm will iterate until it converges. Sometimes, convergence can be determined by comparing MLEs in consecutive iterations; in which case the EM will cease once the difference between MLEs in consecutive iterations is below some threshold. Convergence can also be determined by comparing model likelihoods— once model likelihoods from consecutive iterations are within some threshold of each other, the EM will cease. EM algorithms can also be forced to stop prematurely if certain parameter thresholds are exceeded or if a maximum number of iterations has been reached.

5.3.3.3.3.2 An EM genotyper which incorporates sequencing noise and allele bias

Let $x = \{x_1, x_2, \dots, x_N\}$ be a set of non-negative integers describing the total observed coverage for a population of *N* people at a microsatellite locus. Let *A* be any integer greater than zero, which defines the total number of alleles observed at the locus. $x_a = \{x_{1a}, x_{2a}, \dots, x_{Na}\}$ is a set of non-negative integers that define the allele coverage throughout the population for any allele a of the *A* alleles observed at the locus. Finally, let $\lambda = \{\lambda_1, \lambda_2, \dots, \lambda_N\}$ be the per-allele expected coverage estimate for each individual within the population, which is the SVD-derived per-allele, per-locus coverage estimator described in section 5.3.3.2.3.

G is the set of possible genotypes at a locus. Each possible genotype is referred to by its constituent alleles. Since genotypes are bi-allelic, genotype annotation in the likelihood equation and MLEs will only refer to three alleles: a, b, and 0. The purpose of the 0, or null, allele will be explained in section 5.3.3.4.1. Alleles a and b refer to any two alleles in *A*. Therefore, the genotype g_{aa} refers to any homozygous genotype; g_{ab} refers to any heterozygous, non-null genotype; g_{a0} refers to any genotype containing one null allele; and

g_{00} refers to the genotype containing two null alleles. The number of genotypes in G is given as $A(A+1)/2$.

Let $c = \{c_1, c_2, \dots, c_N\}$ be a vector of vectors defining the copy number for all *A* alleles and all *N* individuals. Each element $c_i = \{c_{i1}, c_{i2}, \dots, c_{iA}\}$, provides the copy number of each allele for individual *i*. Elements of c_i are limited to $c_{ia} = [0,1,2]$, and $\sum_1^A c_{ia} \leq 2$. These restrictions enforce the assumption that any genotype is comprised of at most two alleles. If a genotype has a single null allele, $\sum_1^A c_{ia} = 1$. If it contains two null alleles, $\sum_1^A c_{ia} = 0$. An individual's copy number vector defines their genotype, which is the latent variable in our EM.

Let $t = \{t_1, t_2, \dots, t_N\}$ define a vector of membership probability vectors for all individuals in *N*. Each element in *t* is a vector defining the membership probabilities of a particular individual for all possible genotypes. The membership probability of individual *i* for genotype g is referred to as $t_{i,g}$. For any individual, $\sum_g^G t_{i,g} = 1$.

Each genotype has a weight describing its frequency at the locus. Like genotypes, genotype weights are also referenced by the alleles they contain. The vector w contains all the genotype weights for the locus, and $\sum_g^G w_g = 1$.

We choose to model each allele as an independent Poisson process—the Poisson allele coverage model. This decision was motivated by the observation that homozygous genotypes for biased alleles tend to have lower coverage than homozygous genotypes for unbiased alleles (Figure 5.14B). This would suggest that coverage at biased alleles is not reduced due to competition with unbiased alleles, but rather that the capture process itself is less effective. By modeling alleles as separate Poisson distributions with rates modified by per-allele capture bias, uSeq is able to account for this behavior. An alternative genotyping model could use a Poisson distribution to model the total coverage at any locus, followed by binomial sampling from that total coverage when assigning reads to both alleles. This model, referred to as the Poisson total coverage model, accounts for capture bias by modifying the binomial probability of a read being assigned to a particular allele. When considering heterozygous genotypes, a comparison of simulated allelic coverage for the two proposed models demonstrates that the model choice is probably irrelevant. The difference is dramatic when considering homozygous genotypes. Since binomial sampling in the Poisson total coverage model has no effect on total coverage, it would be impossible to account for the observed decrease in total coverage for homozygous genotypes of biased alleles. Of the two models, only the Poisson allele coverage model accounts for this behavior.

The likelihood function of a genotype in any individual is defined as the products of the Poisson probability of the observed coverage for any genotyped alleles multiplied by the binomial probability of the observed noise coverage given p. The Poisson rate for any allele a in individual i is defined as $c_{ia}\lambda_i\alpha_a$. Therefore, the likelihood of observing any genotype can be defined by the following equations:

$$P(g_{aa}|x_i) = \frac{(2\lambda_i\alpha_a)^{x_{ia}}}{x_{ia}!}e^{2\lambda_i\alpha_a}\binom{x_i}{x_i - x_{ia}}p^{x_i - x_{ia}}(1-p)^{x_{ia}}$$

$$P(g_{ab}|x_i) = \frac{(\lambda_i\alpha_a)^{x_{ia}}}{x_{ia}!}e^{\lambda_i\alpha_a}\frac{(\lambda_i\alpha_b)^{x_{ib}}}{x_{ib}!}e^{\lambda_i\alpha_b}\binom{x_i}{x_i - x_{ia} - x_{ib}}p^{x_i - x_{ia} - x_{ib}}(1-p)^{x_{ia}+x_{ib}}$$

$$P(g_{a0}|x_i) = \frac{(\lambda_i\alpha_a)^{x_{ia}}}{x_{ia}!}e^{\lambda_i\alpha_a}\binom{x_i}{x_i - x_{ia}}p^{x_i - x_{ia}}(1-p)^{x_{ia}}$$

$$P(g_{00}|x_i) = p^{x_i}$$

The purpose of the EM algorithm is to estimate the membership probabilities of each possible genotype in *G* for every individual in *N*, as well as the set of unknown parameters, $\theta = \{w, \alpha, p\}$.

We define the following likelihood function for our model, with parameters θ:

$$l(\theta) = \left(\prod_{a\in A} w_{aa}\sum_{i}^{N} t_{i,aa}\, P(g_{aa}|x_i)\right) \times \left(\prod_{a\in A}\prod_{b\in A, a\notin A} w_{ab}\sum_{i}^{N} t_{i,ab}\, P(g_{ab}|x_i)\right)$$

$$\times \left(\prod_{a\in A} w_{a0}\sum_{i}^{N} t_{i,a0}\, P(g_{a0}|x_i)\right) \times \left(w_{00}\sum_{i}^{N} t_{i,00}\, P(g_{00}|x_i)\right)$$

The log-likelihood function of $l(\theta)$, $L(\theta)$, will then be:

$$L(\theta) = \sum_{a\in A} \ln\left(w_{aa}\sum_{i}^{N} t_{i,aa}P(g_{aa}|x_i)\right) + \sum_{a\in A}\sum_{b\in A, a\notin A} \ln\left(w_{ab}\sum_{i}^{N} t_{i,ab}P(g_{ab}|x_i)\right)$$

$$+ \sum_{a\in A} \ln\left(w_{a0}\sum_{i}^{N} t_{i,a0}P(g_{a0}|x_i)\right) + \ln\left(w_{00}\sum_{i}^{N} t_{i,00}P(g_{00}|x_i)\right)$$

According to Jensen's inequality, $\ln\sum_i \beta_i y_i \geq \sum_i \beta_i \ln y_i$. Using this property, we can set a lower bound for our log-likelihood equation (Equation 1):

$$L(\theta) \geq \sum_{a\in A} w_{aa}\sum_{i}^{N} t_{i,aa}\ln\big(P(g_{aa}|x_i)\big) + \sum_{a\in A}\sum_{b\in A, a\notin A} w_{ab}\sum_{i}^{N} t_{i,ab}\ln\big(P(g_{ab}|x_i)\big)$$

$$+ \sum_{a\in A} w_{a0}\sum_{i}^{N} t_{i,a0}\ln\big(P(g_{a0}|x_i)\big) + w_{00}\sum_{i}^{N} t_{i,00}\ln\big(P(g_{00}|x_i)\big)$$

This can be written more concisely as:

$$L(\theta) \geq \sum_{g\in G} w_g \sum_{i}^{N} t_{i,g}\ln\big(P(g|x_i)\big)$$

In the rest of the discussion of the EM, $L(\theta)$ refers to the log-likelihood equation simplified by Jensen's inequality. After each iteration, the EM algorithm uses its current parameter estimates to calculate $L(\theta)$. Values of $L(\theta)$ for consecutive iterations can then be compared to determine whether the EM has converged.

5.3.3.3.3.2.1 MLEs for α and p

Deriving MLEs for the parameters α and p requires the logarithms of each genotype likelihood function, which are given here:

$$\ln P(g_{aa}|x_i) = x_{ia}\ln 2\lambda_i\alpha_a - \ln x_{ia}! + 2\lambda_i\alpha_a + \ln\binom{x_i}{x_i - x_{ia}} + (x_i - x_{ia})\ln p$$
$$+ x_{ia}\ln(1-p)$$

$$\ln P(g_{ab}|x_i) = x_{ia}\ln\lambda_i\alpha_a - \ln x_{ia}! + \lambda_i\alpha_a + x_{ib}\ln\lambda_i\alpha_b - \ln x_{ib}! + \lambda_i\alpha_b$$
$$+ \ln\binom{x_i}{x_i - x_{ia} - x_{ib}} + (x_i - x_{ia} - x_{ib})\ln p + (x_{ia} + x_{ib})\ln(1-p)$$

$$\ln P(g_{a0}|x_i) = x_{ia}\ln\lambda_i\alpha_a - \ln x_{ia}! + \lambda_i\alpha_a + \ln\binom{x_i}{x_i - x_{ia}} + (x_i - x_{ia})\ln p$$
$$+ x_{ia}\ln(1-p)$$

$$\ln P(g_{00}|x_i) = x_i\ln p$$

The partial derivative of $L(\theta)$ with respect to p can be solved for zero to provide its MLE:

$$\frac{dL(\theta)}{dp} = \frac{1}{p}\left(\sum_{a\in A} w_{aa}\sum_{i}^{N} t_{i,aa}(x_i - x_{ia}) + \sum_{a\in A}\sum_{b\in A, a\notin A} w_{ab}\sum_{i}^{N} t_{i,ab}(x_i - x_{ia} - x_{ib})\right.$$
$$\left. + \sum_{a\in A} w_{a0}\sum_{i}^{N} t_{i,a0}(x_i - x_{ia}) + w_{00}\sum_{i}^{N} t_{i,00}x_i\right)$$
$$- \frac{1}{1-p}\left(\sum_{a\in A} w_{aa}\sum_{i}^{N} t_{i,aa}x_{ia} + \sum_{a\in A}\sum_{b\in A, a\notin A} w_{ab}\sum_{i}^{N} t_{i,ab}(x_{ia} + x_{ib})\right.$$
$$\left. + \sum_{a\in A} w_{a0}\sum_{i}^{N} t_{i,a0}x_{ia}\right) = 0$$

Solving the above equation for p, our MLE is (Equation 2):

$$\hat{p} = \frac{A}{B}$$

$$A = \sum_{a \in A} w_{aa} \sum_{i}^{N} t_{i,aa}(x_i - x_{ia}) + \sum_{a \in A} \sum_{b \in A, a \notin A} w_{ab} \sum_{i}^{N} t_{i,ab}(x_i - x_{ia} - x_{ib})$$

$$+ \sum_{a \in A} w_{a0} \sum_{i}^{N} t_{i,a0}(x_i - x_{ia}) + w_{00} \sum_{i}^{N} t_{i,00} x_i$$

$$B = \sum_{a \in A} w_{aa} \sum_{i}^{N} t_{i,aa} x_i + \sum_{a \in A} \sum_{b \in A, a \notin A} w_{ab} \sum_{i}^{N} t_{i,ab} x_i + \sum_{a \in A} w_{a0} \sum_{i}^{N} t_{i,a0} x_i$$

$$+ \sum_{a \in A} w_{00} \sum_{i}^{N} t_{i,00} x_i$$

Although A and B look complicated, the intuition for the MLE $\hat{p}$ is straightforward. The numerator, A, is the sum of the total noise coverage for each genotype multiplied by each individual's membership probability for the respective genotype, weighted by the genotype frequency. Similarly, the denominator, B, is the sum of the total observed coverage for each individual multiplied by their membership probability for every genotype, weighted by the genotype frequency. In essence, the noise rate is related to the number of noise reads divided by the total coverage.

The MLEs for any α_a only depend on the partial derivative of $L(\theta)$ with respect to all genotypes that include the a allele. We obtain the MLE by setting the partial derivative to zero and solving for α_a:

$$\frac{dL(\theta)}{d\alpha_a} = \frac{1}{\alpha_a}\left(w_{aa}\sum_i^N t_{i,aa}x_{ia} + \sum_{b\in A,a\notin A} w_{ab}\sum_i^N t_{i,ab}x_{ia} + w_{a0}\sum_i^N t_{i,a0}x_{ia}\right)$$

$$- w_{aa}\sum_i^N t_{i,aa}2\lambda_i - \sum_{b\in A,a\notin A} w_{ab}\sum_i^N t_{i,ab}\lambda_i - w_{a0}\sum_i^N t_{i,a0}\lambda_i = 0$$

Solving the equation for α_a, provides the following MLE (Equation 3):

$$\widehat{\alpha_a} = \frac{w_{aa}\sum_i^N p_{i,aa}x_{ia} + \sum_{b\in A,a\notin A} w_{ab}\sum_i^N p_{i,ab}x_{ia} + w_{a0}\sum_i^N p_{i,a0}x_{ia}}{w_{aa}\sum_i^N p_{i,aa}2\lambda_i + \sum_{b\in A,a\notin A} w_{ab}\sum_i^N p_{i,ab}\lambda_i + w_{a0}\sum_i^N p_{i,a0}\lambda_i}$$

The intuition behind the MLE $\widehat{\alpha_a}$ is also meaningful. The numerator is the sum of each individual's observed allele coverage multiplied by their respective membership probability, weighted by the genotype frequency. The denominator is the sum of the expected allele coverage for each genotype in each individual, multiplied by their respective membership probability, weighted by the appropriate genotype frequency. In its simplest sense, the allele bias is the observed coverage divided by the expected coverage.

5.3.3.3.3.2.2 Initializing the EM

The EM initializes with the assumption that every allele is unbiased, i.e. $\alpha_a = 1$ for all alleles. All possible genotypes are assumed equally likely, and the noise rate for the locus is assumed 0.001. Currently, the initial genotype frequencies and allele biases are not modifiable, but the initial noise estimate can be changed.

5.3.3.3.3.2.3 E step

During the expectation step, the current model is used to calculate membership probabilities for each person *i* and each genotype g_j (Equation 4):

$$t_{i,g_j} = \frac{w_{g_j}P(g_j|x_i)}{\sum_{g\in G} w_g P(g|x_i)}$$

5.3.3.3.3.2.4 M step

Genotype frequencies are estimated as (Equation 5):

$$\hat{w}_g = \frac{1}{N}\sum_{i}^{N} t_{i,g} \; ; g \in G$$

The MLEs for α and p are as described in section 3.3.3.3.2.1.

5.3.3.3.3.2.5 Determining EM convergence

The EM converges when the current model log-likelihood is within some threshold of the log-likelihood for the model immediately preceding it. In the current uSeq implementation, this threshold is 10. The EM will stop prematurely under two conditions:

1. The EM has not converged after a threshold number of iterations. The default threshold is 25 iterations.
2. The noise rate has exceeded some upper limit. The default threshold is 0.4.

The EM is forced to stop once the noise rate is so high as to make confident genotype calling impossible. If 2 of every 5 reads may be the result of slippage noise, there is almost certainly no means of consistently calling accurate bi-allelic genotypes within the study population at the locus. A high estimated noise rate does not necessarily mean a locus is unstable during sample preparation—it could also suggest that a locus has frequent somatic variations or CNPs. The complete pseudocode for the EM genotyper is below.

Algorithm 5.10 Calling microsatellite genotypes with an EM algorithm estimating per-allele bias and per-locus noise rates

```
Input: $N$, the total number of people in the population; $A$, the number of alleles observed at a microsatellite locus; $alleleCoverage$, an $N \times A$ matrix with the coverage for all observed alleles throughout the population; $expectedCoverage$, a length $N$ vector of per-allele expected coverage estimators for the population; $initialBiasEstimate$, the initial bias estimates for the alleles; $initialNoiseEstimate$, the initial locus error rate estimate; $maxIterations$, the maximum number of EM iterations to perform; $maxNoiseRate$, the maximum noise rate to allow; $emThreshold$, the threshold for EM convergence
$G \leftarrow (A+1)(A+2)/2$    // number of possible genotypes, including null genotype calls
$emWeights \leftarrow \emptyset$
for $i = 1$ to $G$ do    // all genotypes are equally likely before the EM begins
    $emWeights[i] \leftarrow 1$
end for
$errorRateMLE \leftarrow initialNoiseEstimate$
$biasMLE \leftarrow initialBiasEstimate$
$modelLogLikelihood \leftarrow$ likelihood of observed data, given $alleleCoverage$, $expectedCoverage$, $initialBiasEstimate$, and $initialNoiseEstimate$    // equation 1, page 138
$oldLogLikelihood \leftarrow modelLogLikelihood$
$newLogLikelihood \leftarrow modelLogLikelihood + (2 \times emThreshold)$
$numIterations \leftarrow 0$
while $abs(oldLogLikelihood - newLogLikelihood) > emThreshold$ and $numIterations < maxIterations$ and $errorRateMLE < maxNoiseRate$ do
    increment $numIterations$
    $genotypeProbabilities \leftarrow$ genotype probabilities using $emWeights$, $alleleCoverage$, $expectedCoverage$, $biasMLE$, and $errorRateMLE$    // dimensions $N \times G$; Equation 4, page 141
    $emWeights \leftarrow$ genotype weights using $genotypeProbabilities$    // Equation 5, page 142
    $errorRateMLE \leftarrow$ estimated error rate from $emWeights$, $genotypeProbabilities$, and $alleleCoverage$    // Equation 2, page 140
    $biasMLE \leftarrow$ estimated biases from $emWeights$, $genotypeProbabilities$, $alleleCoverage$, and $expectedCoverage$    // Equation 3, page 141
    $oldLogLikelihood \leftarrow newLogLikelihood$
    $newLogLikelihood \leftarrow$ likelihood of observed data, given latest $biasMLE$ and $errorRateMLE$, as well as $alleleCoverage$ and $expectedCoverage$    // Equation 1, page 138
end while
$genotypes \leftarrow \emptyset$    // $N \times 2$ matrix reporting bi-allelic genotypes
$genotypeConfidence \leftarrow \emptyset$    // vector of length $N$
$genotypeCoverage \leftarrow \emptyset$
// $N \times 3$ matrix reporting coverage for genotyped alleles, as well as noise coverage
$noiseGoodnessOfFit \leftarrow \emptyset$    // vector of length $N$
$alleleGoodnessOfFit \leftarrow \emptyset$    // vector of length $N$
$marginalNullProbability \leftarrow \emptyset$    // vector of length $N$
for $i = 1$ to $N$ do
    $genotypeConfidence[i] \leftarrow$ maximum value in $genotypeProbabilities[i][1:G]$
    // find the likeliest genotype for every individual
    $genotypes[i] \leftarrow$ genotype of $genotypeConfidence[i]$
    $genotypeCoverage[i] \leftarrow$ coverage for genotyped alleles and any noise coverage
    $noiseGoodnessOfFit[i] \leftarrow$ one-sided binomial exact test given sum of $alleleCoverage[i]$ and noise coverage in $alleleCoverage[i]$    // see text, page 153
    $alleleGoodnessOfFit[i] \leftarrow$ two-sided Poisson exact test given $alleleCoverage[i]$ for all genotyped alleles, their values in $biasMLE$, and $expectedCoverage[i]$    // see text, page 154
    $marginalNullProbability[i] \leftarrow$ sum of $genotypeProbabilities[i]$ for all genotypes with a null allele
end for
return $genotypes$, $genotypeConfidence$, $genotypeCoverage$, $noiseGoodnessOfFit$, $alleleGoodnessOfFit$, $marginalNullProbability$
```

5.3.3.3.3.2.6 Interpreting EM results

Once the EM has converged or stopped prematurely at a locus, uSeq saves the noise rate MLE, a vector of allele capture bias MLEs, and an upper triangular matrix of the genotype weights with dimensions $A \times A$. The rows of the genotype weight matrix correspond to first allele of a genotype, and the columns correspond to the second allele of a genotype. Allele biases can take any non-negative value. If an allele has a bias of 1, it is captured with the efficiency predicted by the per-allele, per-person expected coverage estimators at that locus. A bias term below 1 indicates decreased capture efficiency for an allele relative to expected coverage estimators for that locus. A bias term above 1 indicates increased capture efficiency relative to the coverage estimators for that locus.

The EM also reports every individual's membership probabilities for any genotype at the locus. The complete membership probabilities for any locus are reported as a $N \times A \times A$ array of N upper triangular matrices—one for each individual. The complete membership probabilities are unwieldy and most genotype probabilities are very low. uSeq only reports the maximum membership probability for each individual, along with the alleles comprising the corresponding genotype. The maximum membership probability is reported as the genotype confidence. Genotype confidences close to 1 indicate confident genotype calls.

5.3.3.3.3.3 Benefits of the EM approach

An EM approach is appealing for several reasons. EM iteratively updates its distribution parameter estimates, as it improves its membership probabilities and distribution weights. This allows an EM to find MLEs even when initial parameter estimates are inaccurate. The EM membership probabilities not only report the likeliest source distribution source for each data point, they also indicate any uncertainty regarding the source distribution. This property is essential to uSeq's ability to call microsatellite genotypes and report their confidences. The EM convergence threshold can be set arbitrarily, which allows the user to balance accurate parameter estimates and additional computation time. Similarly, the total number of iterations and distribution parameters can be capped in an EM. These final two properties are particularly useful in uSeq, where a separate EM is required for each locus. Since there are so many loci, the current EM implementation favors rapid convergence by limiting the maximum number of iterations and noise rate.

5.3.3.3.4 Testing the uSeq EM genotyper

5.3.3.3.4.1 Simulating biased microsatellite coverage

We designed a single-locus simulation to test the uSeq EM genotyper's performance. The simulator takes the following parameters as inputs:

- Population size
- Per-person, per-allele expected coverage
- Number of alleles
- Per-allele capture bias
- Population allele frequencies
- Null allele frequency
- Locus-wide noise rate

For each individual in the population, the simulator picks a random genotype based on the provided allele frequencies. The simulator will then randomly remove alleles from individuals within the population based on the null allele frequency. After the genotypes for each individual have been established, coverage is assigned to each genotyped allele from a Poisson distribution with a rate parameter defined by each person's expected coverage, the capture biases for each allele, and each person's copy number vector. In every individual, the number of slippage events from each allele is assigned from a binomial distribution using each allele's coverage and the locus-wide noise rate. Slippage events are randomly assigned to any allele except for the allele from which they are derived. Simulated noise is independent of other typical sequencing errors since they are not expected to significantly affect the number of errors observed in any individual.

The EM genotyper is provided the observed locus coverage and the per-allele expected coverage for each individual. The simulator does not derive expected coverage parameters using the SVD described in section 5.3.3.2. We previously demonstrated that input Poisson parameters and SVD-derived per-allele, per-person expected coverage estimators are highly correlated in simulation; therefore, we do not expect the source of the EM genotyper's expected coverage estimators to affect its performance. Given this input data, the EM predicts each allele's capture bias, the locus-wide noise rate, and the genotypes of each individual. The estimated biases and noise rate are compared to their respective input parameters. The EM-derived genotypes are also compared to genotypes derived from a naïve genotyper that does not assume any capture bias.

5.3.3.3.4.2 Simulated EM genotyper performance—single parameter set

As a demonstration of the EM genotyper's performance, we simulated a locus with a population of 5,000 individuals with two alleles, A and B. Per-allele expected coverage for each individual is sampled from a Poisson distribution with a rate parameter of 10. The allele frequency for A was 0.9 and the frequency for B was 0.1. No null alleles were assigned. The capture bias for A was 1, and the capture bias for B was 0.5. As a reminder, the closer the allele bias is to 1, the less bias it has. Finally, we specify a locus-wide noise rate of 0.03.

The true genotypes and observed coverage assigned by the simulation can be found in Figure 5.15A on page 184, and genotypes called by the naïve genotyper are in Figure 5.15B on the same page. By comparing the two plots, the shortcomings of the naïve genotyper are readily apparent. Even when the correct locus-wide noise rate is provided, genotypes are consistently called incorrectly—almost every BB genotype, many AB genotypes, and even some AA genotypes are called incorrectly.

After applying the EM genotyper to the simulated data, we first compare the input slippage rates and bias to their EM-derived values. The comparison is very favorable—the EM-inferred noise rate is 0.01, compared to the input noise rate of 0.03; the EM-inferred capture bias for A is 0.99, compared to 1; and the EM-inferred capture bias for B is 0.51, compared to 0.5. The accuracy of the EM-inferred parameter estimates improves as the locus coverage and population size increase. The genotypes provided by the EM genotyper are also visibly more accurate than the naïve genotypes (Figure 5.16B, page 185).

A careful comparison demonstrates a clear improvement of the EM genotyper relative to the naïve genotyper. We can define a summary table for both genotypes, where each row corresponds to the original, simulated genotypes, and each column corresponds to a genotype call using any genotyping algorithm. This summary table reports a genotyper's precision for each category of original genotypes, as well as any typical errors it makes in its genotype calls. The summary tables for both genotypers, considering all genotypes, regardless of genotype confidence, can be found in Figure 5.17 on page 186. 87% of the naïve genotyper's calls are correct. It is most precise when the original genotype is AB and has a decreased precision for AA genotypes. The naïve genotyper's precision drops precipitously for BB genotypes—less than 10% of all BB genotypes are called correctly. In contrast, the EM genotyper calls >98% of genotypes correctly. Its precision is >99% for AA genotypes, and >92% for AB genotypes. The most dramatic improvement is for BB genotypes, which are called correctly >96% of the time.

If we vary genotype confidence thresholds from 0 to 1, we can measure the sensitivity and specificity of both genotypers for confidence cutoffs. We can do this graphically by comparing the ROC plots for both genotypers (Figure 5.18, page 187). The naïve genotyper is comparatively mediocre job at distinguishing true positive and false positives genotype calls—maintaining an FPR below 5% requires a minimum confidence threshold of 0.988, which limits sensitivity to <50%. In contrast, the EM genotyper requires a minimum confidence threshold of 0.996 to maintain an FPR below 5%, but TPR is a vastly improved 82%. A more general comparison of model performance can be obtained by comparing the AUC for both ROC curves. The AUC measures the chance that the genotyper will provide a higher confidence for a randomly selected correct genotype call than for a randomly selected incorrect genotype call. In this particular simulation, the AUC for the EM genotyper is 0.97, while the AUC for the naïve genotyper is 0.85, indicating a clear advantage in implementing the EM genotyper.

The FPR is calculated as the number of false positives divided by the total number of negative calls (true negatives and false positives). When the number of false positives and true negatives are low, the FPR can seem high even though the total number of false positives is low. In this particular simulation, of the 5,000 total calls made by the EM genotyper, just 76 are incorrect. Requiring an FPR of 5% in the EM genotyper allows for just three false positives, but we may be willing to be more permissive since the maximum false positive count is low. This behavior can be accurately captured using a precision-recall curve (Figure 5.19, page 188). This curve presents a more nuanced view of the precision measured by the summary tables in Figure 5.17. If we demand higher precision from the naïve genotyper, its recall drops dramatically—requiring at least 95% precision limits recall to <85%. In contrast, the minimum precision of the EM genotyper is 98% when recall is 100%.

5.3.3.3.4.3 Simulated EM genotyper performance—multiple parameter sets

While the EM genotyper's performance for a single simulation is very encouraging, extensive tests with a wide range of parameters are necessary to establish its effectiveness. The EM genotyper has been tested on two simulation classes totaling over 5,000 parameter combinations. The first simulation class explores every combination of the following parameters for a locus with 3 alleles (A, B, and C):

- Population sizes of 100; 1000; 3,500; and 10,000
- Mean per-allele, per-person coverage of 7, 12, 20, and 30
- Allele frequencies for alleles A, B, and C
    - 60%, 26%, and 14%
    - 41%, 41%, and 18%
    - 70%, 30%, and 0%
    - 100%, 0%, and 0%
- Null frequencies of 0, 0.01, and 0.001
- Capture biases for alleles A, B, and C
    - 1, 0.8, and 0.3
    - 1, 0.6, and 0.6
    - 1, 0.3, and 0.8
    - 1, 0.8, and 0.8

    - 1, 0.3, and 0.3

- Locus-wide slippage rates of 0.001, 0.01, 0.05, and 0.1

The second simulation class explores every combination of the following parameters for a locus with 5 alleles (A, B, C, D, and E):

- Population sizes of 100; 1000; 3,500; and 10,000
- Mean per-allele, per-person coverage of 7, 12, 20, and 30
- Allele frequencies for alleles A, B, C, D, and E
    - 43%, 19%, 19%, 9.5%, and 9.5%
    - 35%, 35%, 15%, 15%, and 0%
    - 52%, 24%, 24%, 0%, and 0%
    - 30%, 30%, 30%, 0%, and 0%
- Null frequencies of 0.01 and 0.001
- Capture biases for alleles A, B, C, D, and E
    - 1, 0.8, 0.3, 0.3, and 0.3
    - 1, 0.8, 0.8, 0.8, and 0.3
    - 1, 0.8, 0.8, 0.8, and 0.8
    - 1, 0.3, 0.3, 0.3, and 0.3
- Locus-wide slippage rates of 0.001, 0.01, 0.05, and 0.1

Each combination of the 3-allele class was simulated three times, while combinations of the 5-allele class were only simulated once. We compared the performances of the naïve and EM genotypers by measuring their ROC AUCs for each parameter combination. For the 3-allele class, the mean ROC AUC for each combination was reported.

The EM genotyper performs at least as well as the naïve genotyper for >92% of all parameter combinations in the 3-allele class (Figure 5.20A, page 189). When the EM genotyper underperforms relative to the naïve genotyper, the locus-wide slippage rate is typically 0.1 or the mean per-allele, per-person coverage is 7, particularly when the biases for alleles B and C is 0.3 and both alleles have non-zero frequencies. Within this set, the EM genotyper outperforms the naïve genotyper in ~75% of all combinations, despite frequent slippage, low coverage, and strong capture biases. The relative performance of the EM genotyper is not strongly affected by population size. These observations indicate that genotypes tend to be most accurate for common alleles and rare alleles with relatively weak bias without very high noise rates. A rare allele may be genotyped incorrectly if its capture bias is particularly strong or if the corresponding locus-specific noise rate is especially high.

Among all combinations in the 5-allele class, the EM genotyper performs at least as well as the naïve genotyper for >83% of all parameter combinations (Figure 5.20B, page 189). As was the case for the 3-allele class, the EM genotyper occasionally underperforms relative to the naïve genotyper when the noise rate is high or the mean per-allele, per-person coverage is low. This is particularly true when biases are strong and many alleles have non-zero frequencies.

In general, the performance of uSeq's EM genotyper is superior to the performance of a naïve genotyper. The limitations of the EM genotyper are unsurprising—low coverage and noisy loci make genotyping more difficult, particularly when capture bias is strong. When these characteristics are true at a locus, any genotyper would have trouble calling genotypes accurately. Additionally, comparing the AUCs of the simulated ROC curves demonstrate that genotype confidence is a generally meaningful measure of genotype precision.

5.3.3.4 Interpreting uSeq genotypes

5.3.3.4.1 The null allele

In the definition of the uSeq EM genotyping algorithm, we introduced the concept of the null allele. The null allele allows uSeq to indicate a lack of confidence that complete information has been observed at a locus, i.e. the observed coverage is less than expected, even when accounting for capture bias. The EM may have a high confidence null allele for several reasons. A null allele could be due to insufficient coverage because of stochastic processes. A null allele could also be due to an undetected microsatellite in an individual—perhaps they have a microsatellite allele whose tract length is below uSeq's detection threshold or is interrupted by an indel or SNP. A null allele could also be due to a deletion of the genomic region containing the microsatellite locus in an individual. The genotypes called by uSeq are held to high standards—there must be strong evidence supporting any observed alleles, and there should be no evidence of undetected alleles at a locus. In practice, many null alleles are the result of stochastic sampling. The marginal frequency of the null allele throughout the population at a locus can provide strong evidence of a common undetected microsatellite or deletion.

5.3.3.4.2 Measuring genotype goodness-of-fit with respect to uSeq genotype parameters

uSeq leverages the richness of the SSC population to estimate per-allele, per-person coverage estimators and per-locus noise estimates. These estimators allow uSeq to make highly accurate genotype calls, assuming an individual's true genotype is bi-allelic. If an individual does not have two alleles—either because they have a CNV or a somatic mutation—the per-locus noise rate may mistake a third or fourth microsatellite allele for sequencing errors. Even in the absence of extensive noise, higher-than-expected coverage could indicate the presence of a CNV within an individual. It is not enough for a genotype to have high confidence and low null probability; the allelic and noise coverage must be consistent with a bi-allelic genotype at a microsatellite locus in an individual in our current genotyping model.

The uSeq genotyper measures the evidence that a genotype is bi-allelic by calculating its goodness-of-fit with respect to the per-allele, per-person expected coverage estimator, the per-allele bias estimator, and the per-locus noise rate estimator. Fundamentally, these values address two important concerns about uSeq's final genotype calls—whether the allelic coverage is consistent with our expected coverage model and allele bias model, and whether the observed noise coverage is consistent with the total locus coverage and the per-locus noise rate.

Allelic goodness-of-fit is measured using a two-sided exact Poisson test. Since we expect the coverage for any allele to be sampled from a Poisson distribution, this test is appropriate. The null hypothesis of the test is that the per-allele, per-person expected coverage; the allele capture bias; and the allele copy number define the Poisson rate for allelic coverage in any genotype, i.e.:

$$H_0\colon \lambda_{obs} = \lambda_i \alpha_a c_{ia}$$

for some person i and some allele a. The null hypothesis is rejected when the observed coverage is too high or too low to assume that this rate defines the distribution that produces our observed coverage. The threshold to reject the null hypothesis depends on the study size. When allelic goodness-of-fit is calculated for a heterozygous genotype, the reported goodness-of-fit is the product of each allele's goodness-of-fit. A genotype's allelic goodness-of-fit provides an intuition regarding how well the uSeq genotyping model describes the coverage observed at the locus. A low allelic goodness-of-fit would indicate suspiciously high or low coverage, which could indicate the chance of a CNP or other genomic abnormality at the locus.

Noise goodness-of-fit is measured using a one-sided exact binomial test. We chose this test since the uSeq genotyper assumes that noise reads are sampled from a binomial distribution from a population of x_i reads with a rate of p, using the terms defined in the EM description in section 5.3.3.3.3.2. The test's null hypothesis is that the noise rate for an individual's genotype is less than or equal to the locus-wide noise rate. Since we do not want to penalize genotypes with fewer noise reads than expected, we do not use a two-sided test. The null hypothesis is only rejected if the observed noise coverage is too high to be explained by the locus-wide noise rate estimate. Similar to allelic goodness-of-fit, the threshold for rejecting the null hypothesis depends on the study size. A poor noise goodness-of-fit score would indicate that the best possible genotype for an individual at a locus is still too noisy. This could just be a matter of bad luck—perhaps the individual just happened to have excessive noise at the locus under consideration; but it could also indicate the presence of more than two alleles at the locus or the presence of a somatic mutation.

5.3.3.5 Genotyper implementation

The uSeq EM genotyper is currently implemented in R, which is in part due to the dependence of the expected coverage estimation algorithm on the R package IRLBA, and in part due to R's statistical and graphical capabilities (R Core Team 2013). For loops in R are notoriously slow, which affects the runtime of the uSeq genotyper. Future genotyper implementations may rely on further optimization to reduce the number of for loops; implementing R for loops in C++ using Rcpp; parallelization using one of several R parallel computing packages; or implementing the genotyper in an entirely different programming language, such as Python. In the meantime, uSeq reduces genotyper runtime by splitting genotypes into 10,000 locus chunks that are processed in parallel. This also maintains the genotyper's scalability, since R limits the size of array and matrices. The size of the SSC dataset would make it impossible to load genotype information for all people at all loci into R at once.

5.3.3.6 Standard genotyper output

Currently, genotyper output is stored separately for each chunk of loci in the RData format, which allows for storage of high-dimensional data structures such as lists of matrices or multi-dimensional arrays. The RData file for each chunk contains the following information:

- A vector of the number of EM iterations per locus
- A vector of the per-locus noise rate
- A list of vectors of per-allele capture biases per locus
- A list of matrices of the final locus-wide genotype frequencies from the EM
- A 3-dimensional genotype array

The first dimension of the genotype array has one entry per locus, and the second dimension has one entry per SSC individual. The third dimension consists of 11 fields, which report the following information for each genotype:

1. First genotyped allele
2. Second genotyped allele
3. Genotype confidence
4. Allelic goodness-of-fit

5. Noise goodness-of-fit
6. Allele 1 coverage
7. Allele 2 coverage
8. Noise coverage
9. Expected per-allele coverage estimator
10. Expected noise coverage
11. Marginal null probability

Further improvements to genotyper output could involve reimplementation of the output format in HDF5, which is a popular set of libraries and file formats suited for the storage of large, complex datasets (The HDF Group 1997-2014). HDF5 allows for efficient and flexible management of highly complex data, and has APIs for C++, Perl, Python, and R.

5.3.3.7 VCF format for individuals

Since uSeq considers genotypes on a population-level, its most intuitive format is the 3-dimensional array described in the previous section. To allow for comparisons with other genotypers, uSeq genotype calls for any individual can be obtained in a VCF version 4.1 file format (http://samtools.github.io/hts-specs/VCFv4.1.pdf). Each VCF entry has four field types: required, information, format, and uSeq fields. The format field describes the information found in the uSeq field. The VCF format requires the following fields for each microsatellite genotype:

1. Chromosome
2. Microsatellite start position
3. ID
4. Reference sequence

5. Alternate sequence
6. Phred-scaled genotype confidence
7. Filter

The ID and filter fields are not used in the uSeq pipeline, so they are both set to “.”. Depending on the genotype, the alternate sequence field can be set to “.”; it can contain a single non-reference sequence; or it can contain a comma-delimited list of each genotyped non-reference sequence. The Phred-scaled genotype confidence is given as $-10\log_{10}(1 - genotype\ confidence)$. A genotype confidence of 1 will produce an infinite Phred-scaled confidence. To avoid confusion, the Phred-scaled genotype confidence is set to 100 when this occurs.

The information field contains a semicolon-delimited list of short key-value pairs in a <key>=<value> format. The keys and their meanings are as follows:

1. END: microsatellite stop position
2. RL: microsatellite reference tract length
3. RU: microsatellite reference unit
4. POP: total number of people with coverage
5. SUM: total coverage throughout population
6. TOP: maximum total coverage observed in any individual
7. NR: $\log_{10}$ of the locus-wide noise rate

The format field describes the colon-delimited data found in the uSeq field. The format/uSeq fields provide the following information:

1. GT: genotype indices
2. GB: genotype alleles

3. GC: genotype confidence (not Phred-scaled)
4. EC: expected per-allele coverage
5. AF: allelic goodness-of-fit
6. NF: noise goodness-of-fit
7. AL1: allele 1 coverage
8. AL2: allele 2 coverage
9. ALN: noise coverage
10. PNULL: Phred-scaled marginal null likelihood
11. DP: total coverage
12. SEEN: genotyped allele summary

The genotype indices refer to the reference and alternate sequence fields in the VCF entry. The reference allele has an index of 0, and the alternate alleles start with an index of 1. The genotyped allele summary reports the unique non-null alleles observed.

5.4 Identifying potential *de novo* mutations

A *de novo* mutation is any violation of Mendelian inheritance between two parents and their child. A Mendelian inheritance pattern is any set of genotypes where a child inherits one allele from each parent. When Mendel inheritance appears unlikely, there is a chance that a novel mutation has occurred in the child relative to their parents. This mutation may have occurred in one of the parental germlines, or it may have occurred somatically during the child's development. Mendel violations can be classified into two categories: "commissions" and "omissions". A commission violation occurs when a child has an allele that is not observed in either parent. An omission violation occurs when a child has two copies of an allele that is only observed in one parent (Figure 5.21, page 190).

Omission violations are exceedingly rare for SNVs, but may not be as unusual at microsatellite loci. Omissions at SNV loci may be due to a missed allele from the second parent, an LOH event in the child, or an identical sporadic SNV occurring at the same position by chance. The first explanation is usually the most likely, if coverage is not exceptionally high and capture efficiency is the same for all parental alleles. LOH can be reinforced by nearby loci that are heterozygous between parents. It is exceedingly unlikely that the same mutation observed in a parent occurred sporadically in their child. Since mutation rates are high at some microsatellite loci, and microsatellite mutations are constrained to insertions or deletions of complete units, microsatellite omissions may occur more frequently. This would be due to the increased likelihood of an identical sporadic mutation occurring in a child relative to one of their parents. Therefore, omissions may represent true mutations at microsatellite loci more often than they do at SNVs.

5.4.1 Calculating per-trio Mendel obedience scores

The Mendel obedience score expresses the Mendelian behavior of any parent-child trio. It is calculated as the Phred-scaled sum of the joint probabilities for all possible Mendelian inheritance patterns:

$$-10\log_{10}\sum_{\substack{\text{All Mendelian}\\\text{trios}}} P(M=g_M)P(P=g_P)P(C=g_C)$$

In this equation, $P(M=g_M)$ refers to the probabilities that the mother has a particular genotype, g_M. Similarly, $P(P=g_P)$ and $P(C=g_C)$ refer to genotype probabilities in the father and child, respectively. The probabilities referred to in this equation are the membership probabilities for each genotype provided by the EM genotyping algorithm in section 5.3.3.3.3.2. Since the Mendel obedience score is Phred-scaled, the higher the score, the stronger the evidence that a particular trio contains a Mendel violation.

As is expressed in the Mendel obedience score, the quality of a Mendel violation depends on the confidence of the underlying genotypes. Even when the likeliest trio genotype violates Mendelian inheritance, other potential trio genotypes may still be consistent with Mendelian inheritance. These alternative trio genotypes are represented in the Mendel obedience score. Several alternative explanations exist for apparent violations of Mendelian inheritance that do not involve actual *de novo* mutations. In the case of commission violations, the putative novel allele may actually have been in either parent, but simply went unobserved. Novel alleles may also be the product of sample-induced slippage. In the case of omission violations, an allele that was only observed in one parent may have actually been in both parents, but it had insufficient coverage in one parent. Alternatively, the child may have in fact inherited one allele from each parent, but one parental allele was not observed with sufficient coverage in the child.

When a trio genotype appears to violate Mendelian inheritance, it is essential to distinguish a true *de novo* mutation from an artifact due to any of the several possibilities described above. The Mendel obedience score is an effort to make that distinction—as Mendelian genotypes patterns within a trio become less likely, the score rises, indicating a higher likelihood that a de novo mutation is being observed. The Mendel obedience score explicitly considers all possible Mendelian inheritance patterns, but it does not consider whether the most likely trio genotypes are bi-allelic and detectable. For this reason, we incorporate additional metrics assessing trio genotype quality in order to get the most accurate set of possible *de novo* mutations.

5.4.2 Assessing per-trio kinship

When identifying candidate *de novo* mutations, we assume that only one mutation event occurred within any trio at a locus. *De novo* mutations that introduce two novel alleles at a locus will be exceedingly unlikely when compared to alternative explanations. Possible alternative explanations are similar to those mentioned in the description of omission and commission violations described above.

In the current implementation of uSeq, per-trio kinship is calculated as the maximum swapped Mendel obedience score within a trio. In this calculation, the child in a trio is swapped with either parent and the Mendel obedience score is calculated. Since parents must pass one allele to their children, the swapped Mendel obedience score tests to ensure that the child contains at least one allele that is seen in either parent. The kinship score is not Phred-scaled, so the closer it is it to 1, the more likely it is that a child has inherited at least one allele from a parent. If a kinship score is close to 0, it is more likely that the child does not possess any alleles seen in their parents.

A probabilistic measure of trio kinship can also be calculated without the need for calculating two swapped Mendel obedience scores. This measure would be the sum of the joint probabilities for any parental genotype combination, multiplied by the child's marginal probabilities for any allele observed in either parent for a particular genotype combination. Kinship probabilities close to 1 would be a strong indicator that at least one allele has been inherited from a parent.

5.4.3 Other measures of trio genotype quality

In addition to the Mendel obedience score, we implement additional measures of trio genotype quality to identify the strongest candidate *de novo* mutations. They measure whether every trio member fits the uSeq expected coverage, bias, and slippage models, whether all trio genotypes have been called confidently, and whether anyone within the trio has any evidence of a null allele.

5.4.3.1 Trio genotype confidence

D*e novo* mutations can be filtered by trio genotype confidence. The trio genotype confidence score is calculated as the product of the genotype confidences for each trio member:

$$\prod_{i \in (M,F,C)} \max_{g_i \in G}\left(P(i = g_i)\right)$$

Low trio genotype confidence scores can be the result of low individual genotype confidences in any trio member. This score allows us to discern potential Mendel violations that do not have one clear trio genotype. Ambiguous trio genotypes can be excluded from further analysis, or can be studied separately to gain a better understanding of their unusual behavior.

5.4.3.2 Trio allelic goodness-of-fit

The trio allelic goodness-of-fit score is calculated as the product of each trio member's allelic goodness-of-fit:

$$\prod_{i \in (M,F,C)} D_{allele,i}$$

The allelic goodness-of-fit for each trio member is represented as $D_{allele,i}$. Trio allelic goodness-of-fit is a means of determining how well the trio as a whole is in accordance with their expected coverage parameters and applicable allele bias estimates. Low trio allelic goodness-of-fit could be an indication of a suspiciously high coverage in any trio member.

Trio genotypes with poor trio allelic goodness-of-fit can be used to filter *de novo* mutations, or they can be studied separately in an attempt to better understand why the expected coverage estimator does not fit the trio genotypes.

5.4.3.3 Trio noise goodness-of-fit

Like the trio allelic goodness-of-fit, the trio noise goodness-of-fit is calculated as the product of each trio member's noise goodness-of-fit:

$$\prod_{i \in (M,F,C)} D_{noise,i}$$

Each individual's noise goodness-of-fit is represented as $D_{noise,i}$. Trio noise goodness-of-fit is an effective means of identifying trio genotypes that appear to have at least one member whose coverage is either too noisy or appears to have more than two alleles at a locus. Trio genotypes with poor noise goodness-of-fit may be especially interesting to study further. Members of these trios may have novel CNPs or somatic mutations, or they may be derived from especially noisy loci. By studying these trios in-depth, perhaps some insight could be gained into loci with particularly complex genotypes.

5.4.3.4 Trio null probability

The trio null probability is the probability that any individual within the trio has a null allele, regardless of whether it is their most likely genotype. It is calculated as:

$$1 - \prod_{i \in (M,F,C)} \left(1 - P_{null,i}\right)$$

$P_{null,i}$ is the marginal probability that a trio member has a null allele. The trio null probability is a means of determining whether any trio member is likely to be missing any information at a locus. As discussed in section 5.3.3.4.1, low trio null probabilities are often the result of stochastically low sampling for some trio members. At loci with high null probabilities, these trios may represent good study cases for undetected microsatellite alleles.

5.4.4 Mendel obedience score implementation

De novo mutation detection is implemented in R, similar to the uSeq EM genotyper (R Core Team 2013). It is also split into separate 10,000 locus chunks to reduce the time needed to calculate Mendel obedience scores. It is subject to the same avenues for improvement as the EM genotyper. The only additional limitation placed on the uSeq EM genotyper is the number of alleles considered when calculating Mendel obedience scores for each trio. The Mendel obedience score is calculated for any combination of the six most common alleles within a trio. There are 1,281 possible Mendelian genotype trio combinations from six alleles, which are considered for every trio at every locus. As the number of alleles considered for each trio grows, the number of possible Mendelian genotype trios increases exponentially. The time taken to calculate Mendel obedience scores will dramatically increase with the number of alleles considered.

5.4.5 Mendel obedience output

Mendel obedience scores are saved as 3-dimensional arrays in the RData format for each genotyping chunk. The first dimension of each array contains one entry per locus, the second dimension contains one entry per family, and the third dimension contains 14 fields. The first eight fields contain the likeliest genotypes for each family member in order for the mother, father, proband, and unaffected sibling. The remaining six fields are:

9. Proband trio genotype confidence
10. Sibling trio genotype confidence
11. Proband Mendel obedience score

12. Sibling Mendel obedience score
13. Proband kinship score
14. Sibling kinship score

5.4.6 Identifying strong candidate *de novo* mutations

Potential *de novo* mutations can be filtered using any combination of the metrics defined in the preceding sections to produce a set of high-confidence *de novo* mutation calls. The appropriate thresholds for many of the metrics will be study dependent or can be arbitrarily set. Specific thresholds that yield very good specificity and positive predictive value are discussed in the section 6.6.1.

On occasion, sample mix-ups can occur within a dataset, which will make it appear as though a child has an excessively high number of *de novo* mutations relative to their parents. While there may be biological explanations for high *de novo* mutation rates, we currently take a conservative approach and exclude all families where either child has more than 20 *de novo* mutations. Families with excessive numbers of *de novo* mutation rates can be studied separately to determine whether the increased rate is due to a sample mix-up, impaired MMR efficiency, or some cause.

5.5 Experimental procedures

5.5.1 SSC exome sequencing protocol

SSC samples were collected at 13 clinical centers. The institutional review board at CSHL approved the SSC study, and written informed consent was obtained from all subjects by SFARI. Families selected for uSeq processing were solely composed of "quads"—two unaffected parents, a single proband, and an unaffected sibling. Blood samples from each family member were drawn and sent to RUCDR for DNA preparation. Exome sequencing data from 875 families (3,500 individuals) were selected for analysis in the current study. Of the initial 875 families, 173 were processed at the Genome Center at Washington University, while the remaining 702 families were processed at CSHL. In total, 807 families passed expected coverage filters and passed the filters described in Iossifov et al. in 2012.

A complete description of the exome capture and sequencing protocol can be found in (Iossifov et al. 2012). Briefly, exome capture was performed using NimbleGen SeqCap EZ Exome v2.0, which targets 36.0 Mb of the GRCh37 build of the reference human genome. The standard NimbleGen protocol was slightly modified. In particular, barcodes were ligated prior to capture to enable pooled sequencing of samples. Following sample preparation, 101bp paired-end sequencing was performed on Illumina's HiSeq 2000 platform.

5.5.2 Validation sequencing protocol

Candidate *de novo* mutations were confirmed via PCR amplification of loci with putative mutations, followed by pooled HTS on the Illumina MiSeq instrument. Steven Marks carried out all the validation primer design for this project. Primers were designed using BatchPrimer3 with the following conditions: primers must be between 18 – 27 bp; amplicon lengths are ~300 bp; and ideal primer melting temperatures were 62º C. Primers start within 400 bp of the *de novo* mutation, and were targeted to either the forward or reverse strand of the chromosome. Following primer design, primers were validated using BLAST. Some loci have putative *de novo* mutations in multiple families. These *de novo* mutations were processed in separate validation sets to avoid confusion. At the primer design stage, validation targets can be excluded for one of three reasons:

1. De novo mutation was previously submitted for validation after discovery in other pipelines
2. No DNA was available for at least one family member in the trio
3. Primers could not be designed for target locus

Jennifer Troge performed all sample preparation and sequencing for the microsatellite validation protocol. Mitchell Bekritsky and Jennifer Troge adapted the standard Wigler lab validation protocol for microsatellite loci. Primers were synthesized by Integrated DNA Technologies (Coralville, IA, USA). Primer pairs were resuspended to a 10 µM concentration. Template DNA from whole blood cells was obtained for each family member. Primer cocktails were prepared of 5 µL 2X HFPMM and 3 µL water in 25 5-person aliquots. 5 µL of the respective family primer pair was added to each 5-person aliquot. 1 µL of each individual's template DNA was added to 9 µL of each family-specific primer cocktail. PCR amplification of each individual was carried out under the following conditions:

- 1 cycle at 98º C for 10 seconds
- 24 cycles with the following conditions:
  - 10 seconds at 98º C
  - 15 seconds at 60º C
  - 10 seconds at 72º C
- 1 cycle at 72º C for 10 minutes
- Hold at 4º C

Following PCR amplification, samples were pooled by family member (fathers, mothers, probands, and siblings) and 100 µL of each pool was purified using Qiaquick (Qiagen). The sample concentration of purified samples was determined by Nanodrop (ThermoFisher Scientific), and size distribution was ascertained on a Bioanalyzer (Agilent Technologies). DNA ends were then polished, 3' adenylated, and ligated to barcoded Illumina sequencing adaptors using standard protocols. The barcodes for each adaptor were designed in the Wigler lab. Adaptor-ligated DNA was amplified for 5 cycles and purified. Sample concentration was measured once more by Nanodrop, diluted to 20 ng/μL, and quantitated (Bioanalyzer). After diluting to the appropriate concentration, pooled samples were loaded on an Illumina MiSeq and sequenced using 150bp paired-end reads.

Following sequencing, sample barcodes were deconvoluted, and each family member pool was processed via the individual component of the uSeq pipeline. Since loci were amplified via targeted sequencing, PCR replicate sets are indistinguishable from DNA sequences deriving from distinct genomic fragments. Therefore, we did not mark or filter our PCR replicates. Once profiles have been assembled from each individual, putative *de novo* mutations were manually validated for each family member (mother, father, *de novo* child).

A novel allele is observed if it is distinguishable from slippage noise at a locus in an individual's profile. A *de novo* mutation is considered valid if the novel allele is observed in the appropriate child and is not observed in either parent. If a putative novel allele is not observed in a child, or is observed in the child and either parent, it is considered invalid. If a novel allele is observed in a child, but has no clear interpretation in either parent, the validation status is unclear. Additionally, if any trio member appears to have more than two alleles, the validation status is unclear. Unclear validation results could be the result of somatic mutations in a trio member; sample mix-ups in exome or validation sequencing; or excessive slippage during sample preparation. Any trio with one or more individuals with no observed coverage or observed coverage below 200x was considered to have failed validation. In contrast, observed coverage at validated loci is generally >2000X.

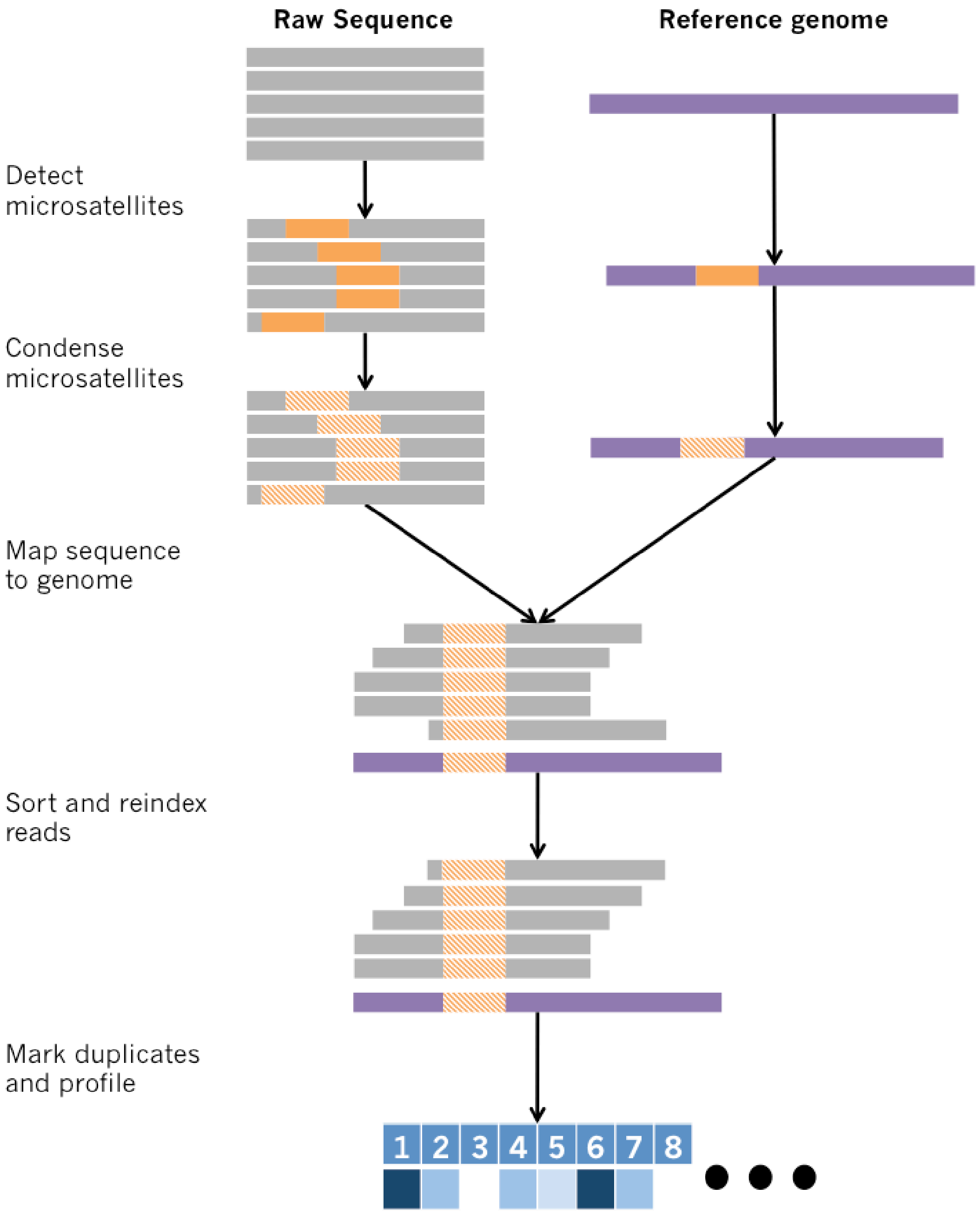
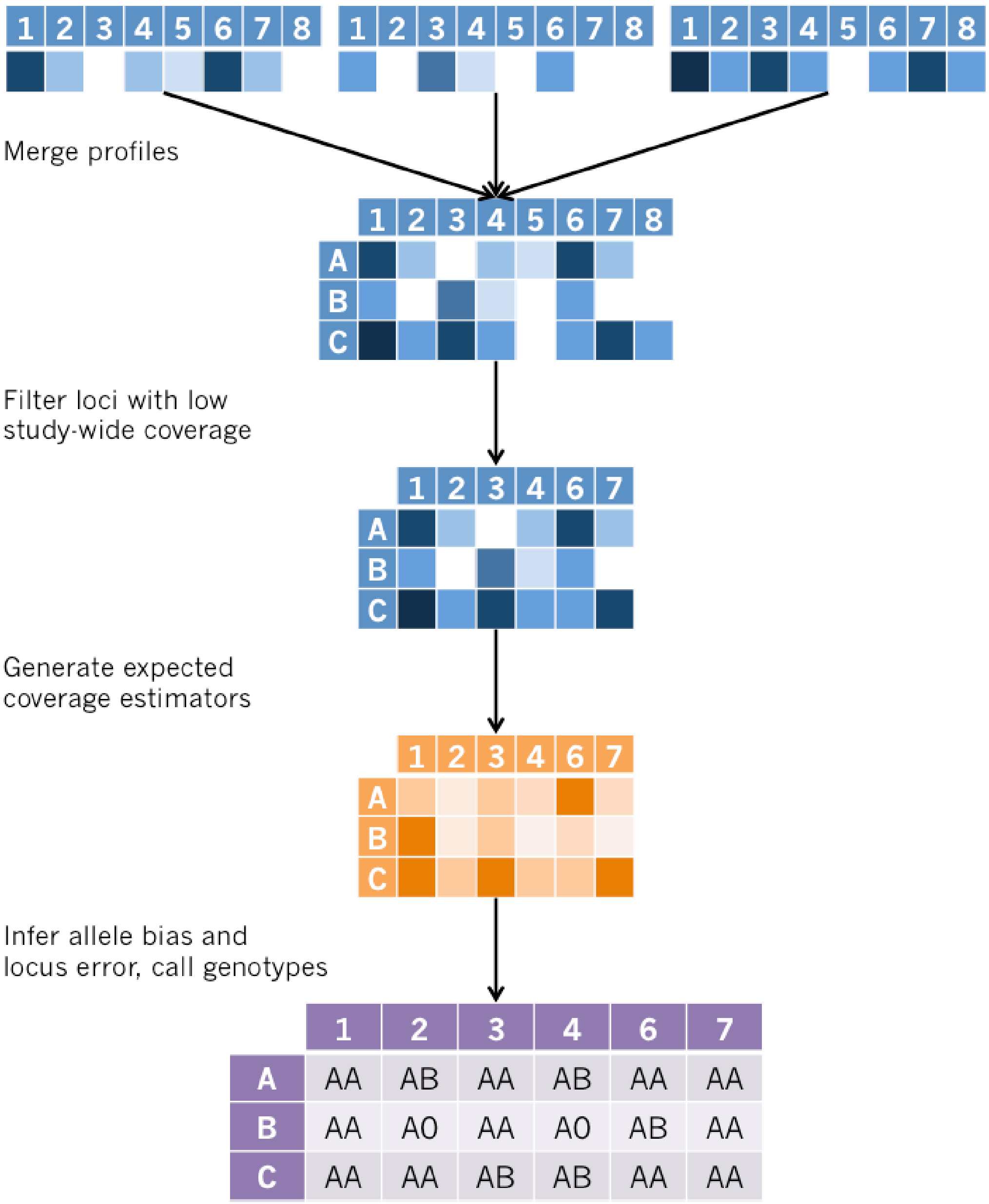


Figure 5.1: An outline of the individual component of the uSeq pipeline. Reads and a reference genome are taken as input. Both are scanned for microsatellites using user-defined parameters for a microsatellite locus. All detected microsatellite loci are removed from the reads and reference genome. Reads are aligned to the condensed reference genome and post-processed. Microsatellite profiles reporting the coverage for each allele at each detected locus are given as output.


1 2 3 4 5 6 7 8
1 2 3 4 5 6 7 8
1 2 3 4 5 6 7 8
Merge profiles
1 2 3 4 5 6 7 8
A
B
C
Filter loci with low study-wide coverage
1 2 3 4 6 7
A
B
C
Generate expected coverage estimators
1 2 3 4 6 7
A
B
C
Infer allele bias and locus error, call genotypes

	1	2	3	4	6	7
A	AA	AB	AA	AB	AA	AA
B	AA	A0	AA	A0	AB	AA
C	AA	AA	AB	AB	AA	AA

Figure 5.2: An outline of the population component of the uSeq pipeline. Microsatellite profiles from a study population are merged and loci with low population coverage are removed. Per-allele, per-person expected coverage parameters are estimated using SVD. After expected coverage estimators are derived, an EM algorithm estimates locus-wide error rates and allele-specific capture bias parameters from population data to call genotypes. Genotypes are called for every study individual at every locus, with genotype confidence scores and metrics assessing each genotype's fit to a bi-allelic genotype model.

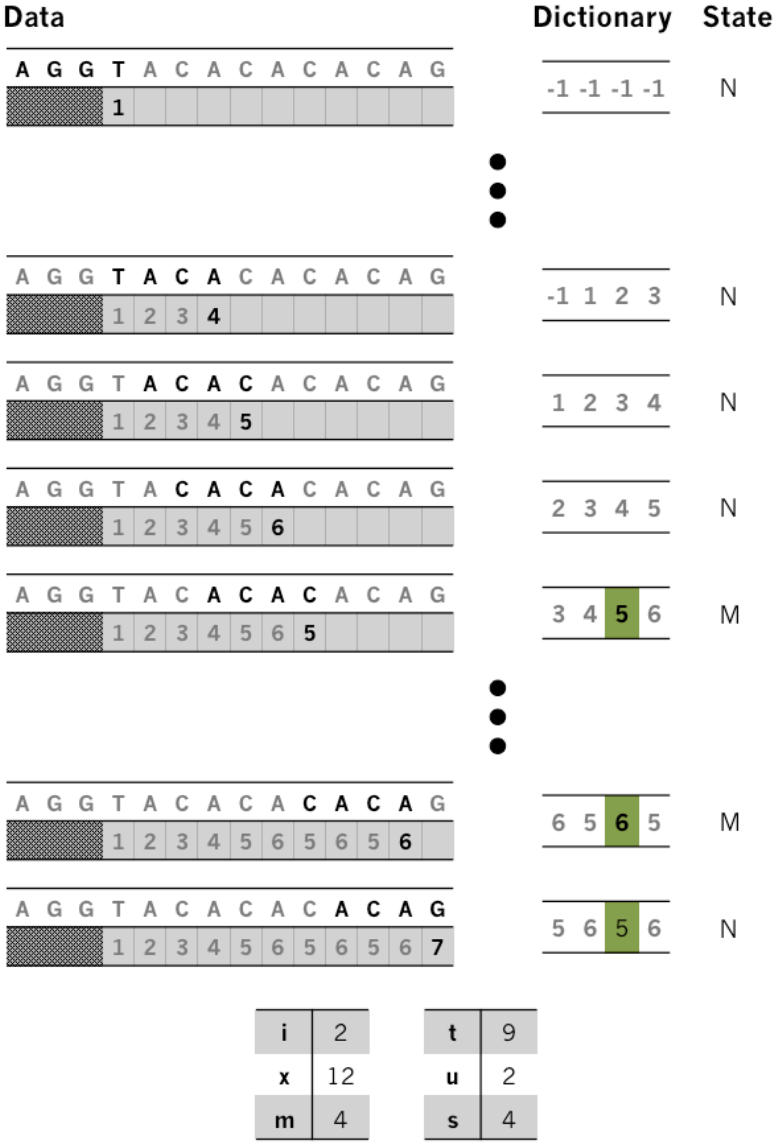


Figure 5.3: An example of the uSeq microsatellite detector FSM, with a maximum kmer size of 4 bp. Positions in the dictionary and in the input sequence are 0-indexed. At each step, the current kmer (bolded in the data column) is converted to a code, which is compared to the kmer codes in the dictionary. If the FSM is in the N state, no putative microsatellite has been detected, and all dictionary kmers are checked. If the FSM is in the M state, a putative microsatellite index has been detected, and the current code is only compared to the initial matching index (in green). Once the current code no longer matches the dictionary code at the matching index, the FSM returns to the N state, and the putative microsatellite is reported if it meets the detector thresholds. The variable names and their values correspond to the variable names and calculations described in section 5.2.1.2.

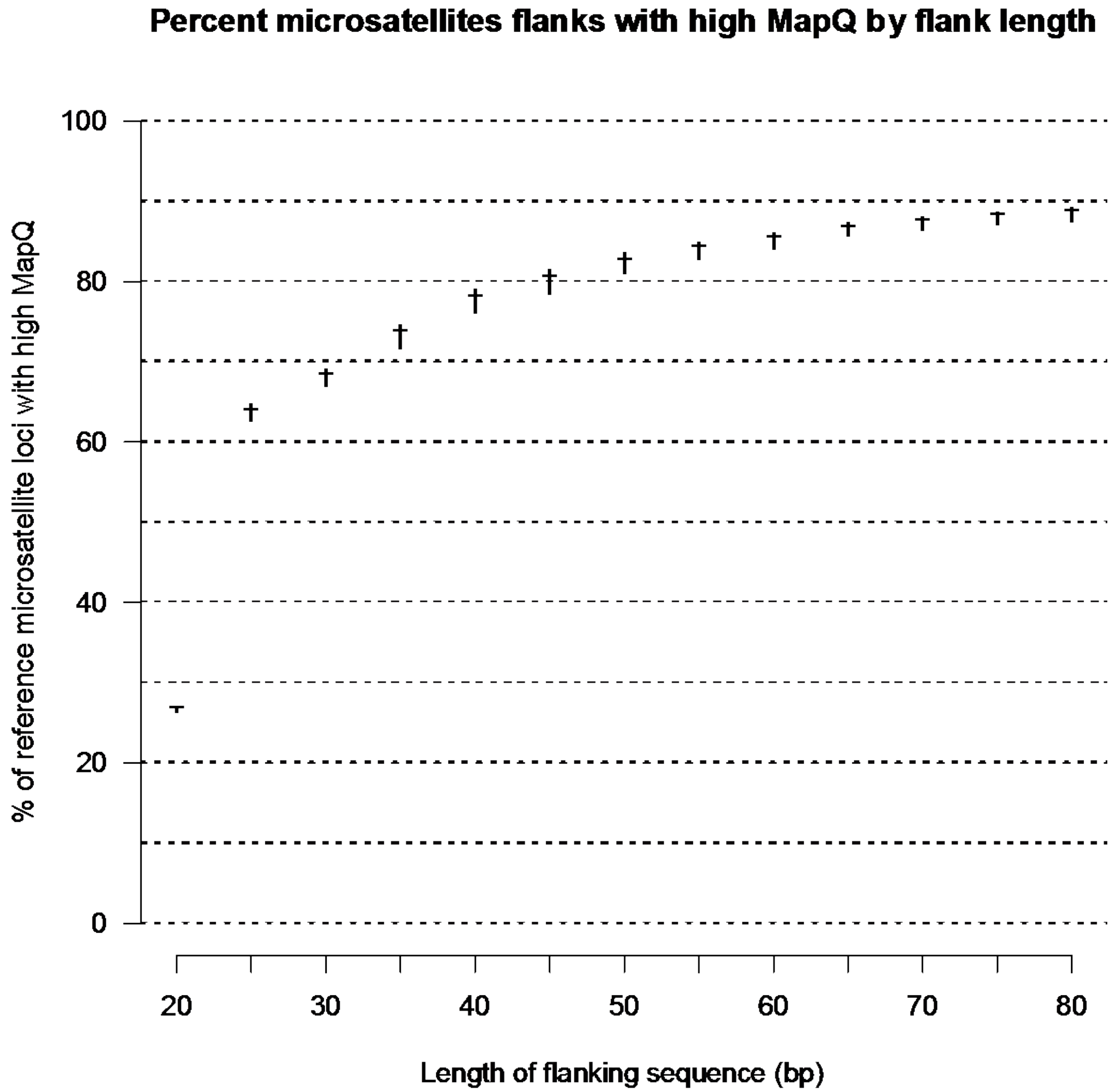


Figure 5.4: Percent of microsatellite flanks that map with mapping quality ≥30 by total flanking sequence length. Horizontal lines mark the mean percentage of high quality microsatellite alignments among all flank divisions for each flank length. Flank divisions are described in section 5.2.2.2. Vertical lines mark the range of high quality alignments among all flank divisions for each flank length. Most microsatellites map with high quality to the reference genome even when flanks are as short as 25 bp. For any flank length, the flank divisions do not significantly affect mapping success.

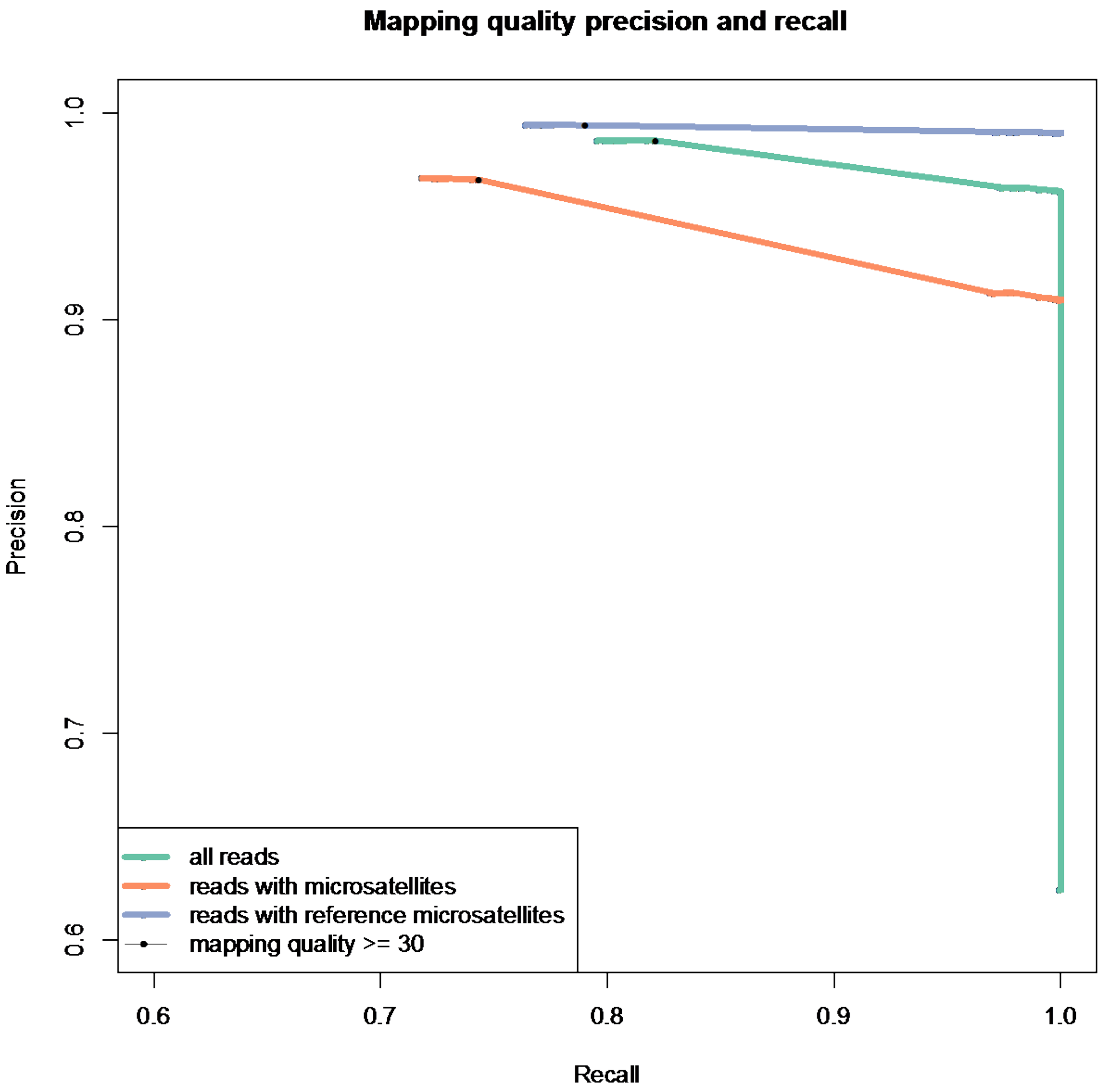


Figure 5.5: uSeq alignment mapping quality precision and recall. The lines represent mapping quality precision and recall for all reads; reads with microsatellites only; or reads with reference microsatellites only, respectively. The black dot indicates a threshold of mapping quality of at least 30. Precision and recall were calculated for simulated reads from chr22 that were condensed and aligned by uSeq. Correctly mapped reads either shared the start position of the simulated fragment, or were 100 bp less than the stop position of the simulated fragment. Each point represents a mapping quality threshold, which range from 0 to 60. For reference microsatellite loci, precision of over 0.99 can be obtained with recall near 0.80 for reads with mapping qualities of at least 30.

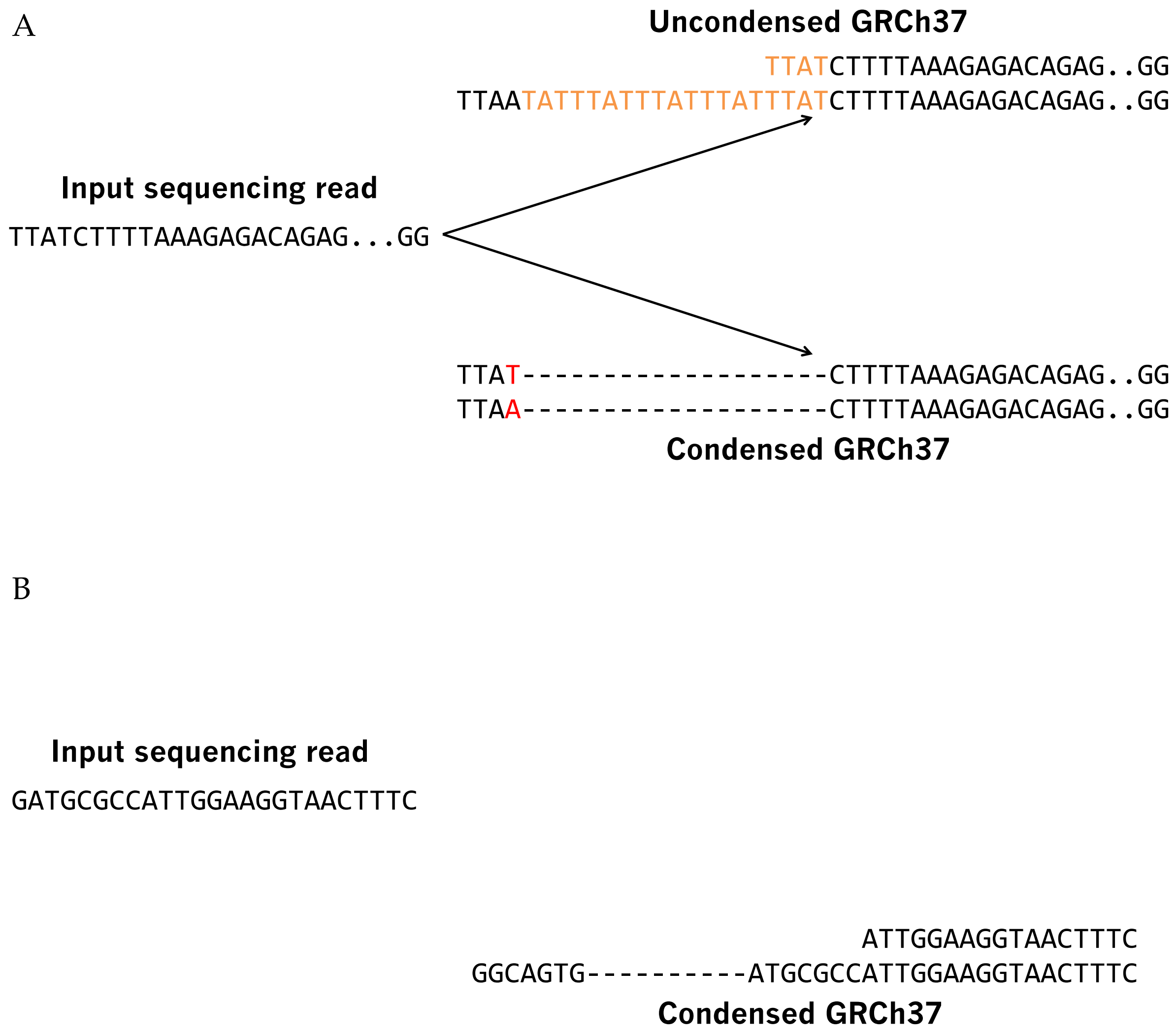

A
Uncondensed GRCh37
TTATCTTTTAAAGAGACAGAG..GG
TTAATATTTATTTATTTATTTATCTTTTAAAGAGACAGAG..GG
Input sequencing read
TTATCTTTTAAAGAGACAGAG...GG
TTAT-------------------CTTTTAAAGAGACAGAG..GG
TTAA-------------------CTTTTAAAGAGACAGAG..GG
Condensed GRCh37
B
Input sequencing read
GATGCGCCATTGGAAGGTAACTTTC
ATTGGAAGGTAACTTTC
GGCAGTG----------ATGCGCCATTGGAAGGTAACTTTC
Condensed GRCh37

Figure 5.6: Undetected terminal microsatellites can cause uSeq to misalign reads. Microsatellite tracts are highlighted in orange, and base mismatches in alignments are highlighted in red. In both examples, the read is aligned incorrectly to the condensed reference genome. In (**A**), a complete motif is contained in the read, which maps to the 5' flanking region of the microsatellite with a single mismatch. In (**B**), only one base in the read is from the microsatellite tract. The base is a perfect match to the last base preceding the microsatellite.

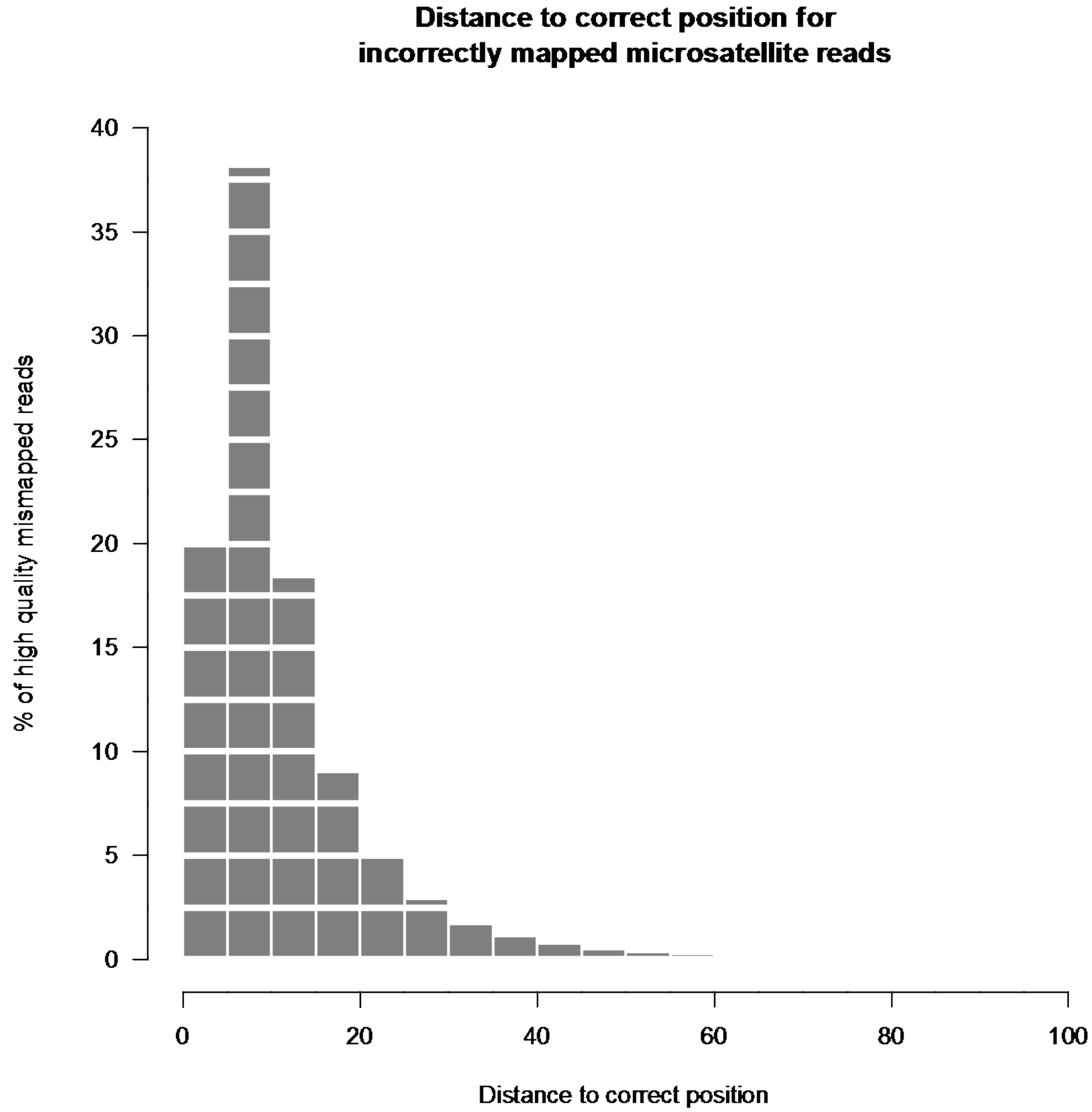


Figure 5.7: Distance of misaligned microsatellite reads to the correct genomic position. The maximum distance to the correct genomic position is >51 Mb. Distance to the correct position was measured as the minimum absolute distance of a start position reported by uSeq to the simulated fragment's start and stop positions. Distances to correct genomic positions are only reported for reads with mapping quality scores ≥30. Most of these misaligned reads with microsatellites are within 100 bp of their correct genomic positions.

A

Non-slippage mutation	Sequence	5’ flank	3’ flank
Original	GATACGAAAAAAAAAAGTCGAT	ATACG	GTCGA
Mutant	GATACGAAAAAAAAAAATCGAT	ATACG	ATACG

B

Slippage mutation	Sequence	5’ flank	3’ flank
Original	GATACG--AAAAAAAAAAGTCGAT	ATACG	GTCGA
Mutant	GATACGAAAAAAAAAAAGTCGAT	ATACG	GTCGA

Figure 5.8: Flanking sequence can distinguish between slippage and substitutions at microsatellite loci. In both (**A**) and (**B**), a microsatellite sequence has an additional motif that extends the sequence by 1 bp. In (**A**), the 3’ flanking sequences of the original microsatellite locus and the mutant microsatellite locus differ, suggesting the mutation occurred via a substitution immediately adjacent to the microsatellite. In (**B**), the 3’ flanking sequences of the original and mutant microsatellite loci are identical, suggesting that a slippage mutation extended the microsatellite locus.

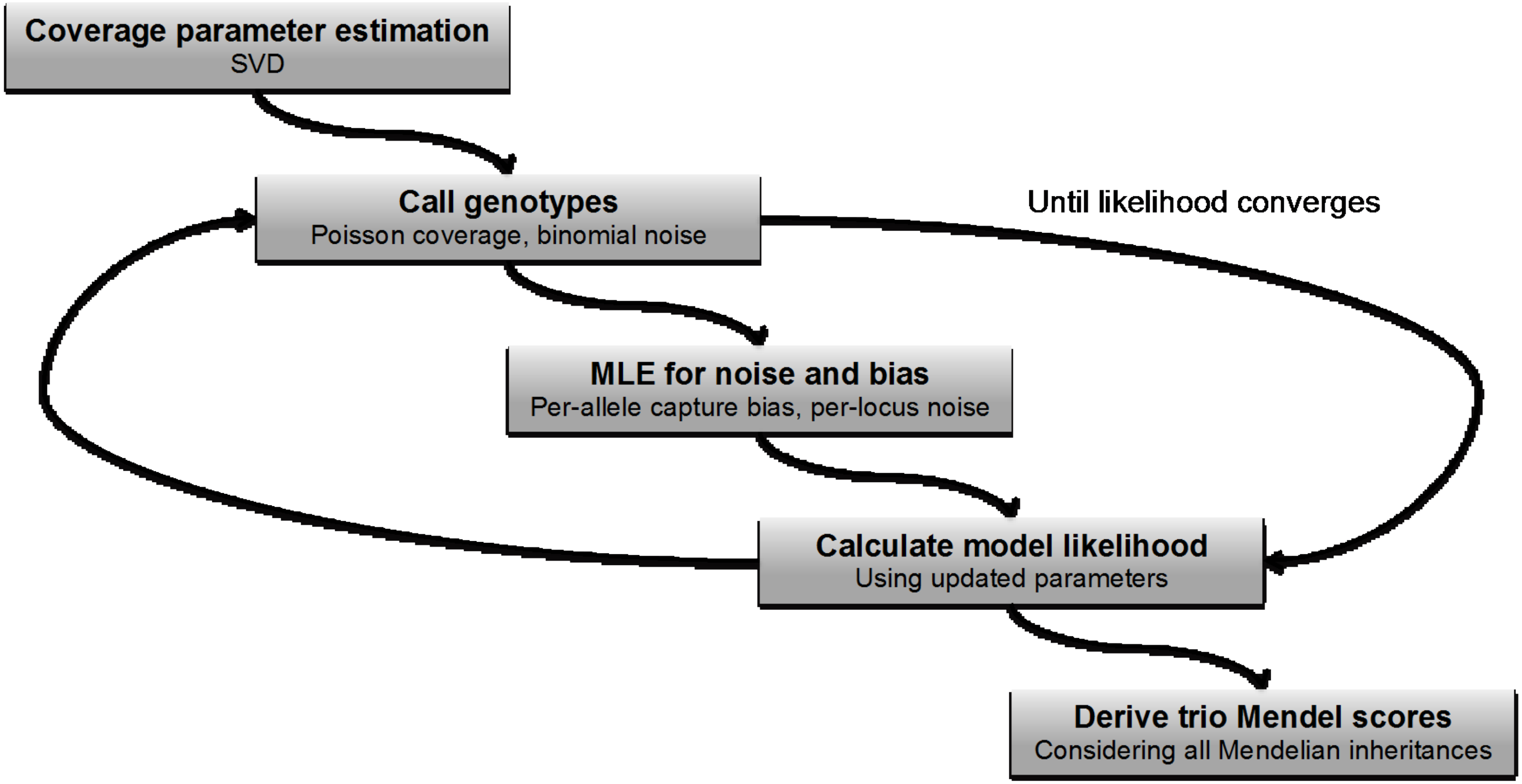

Coverage parameter estimation
SVD
Call genotypes
Poisson coverage, binomial noise
Until likelihood converges
MLE for noise and bias
Per-allele capture bias, per-locus noise
Calculate model likelihood
Using updated parameters
Derive trio Mendel scores
Considering all Mendelian inheritances

Figure 5.9: An outline of the uSeq genotyper, as described in section 5.3.3. First, expected per-allele coverage is estimated for each individual at each locus using SVD. Then, a custom EM iteratively calls genotypes and updates allele bias and locus error parameters until it reaches convergence. Finally, genotype probabilities are used to calculate Mendel obedience scores, which can be used to identify potential Mendel violations. Steps within the loop represent the EM genotyper algorithm.

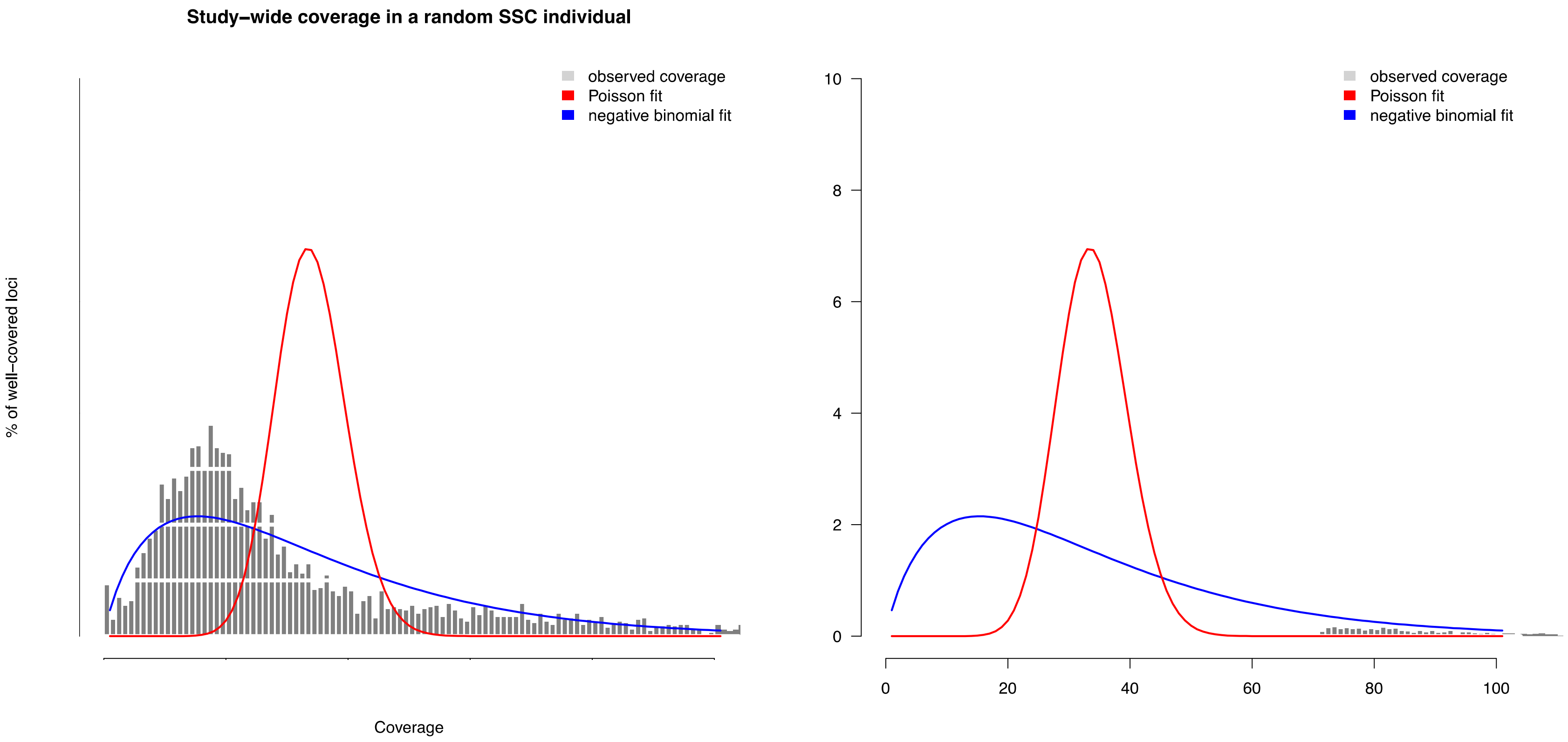


Figure 5.10: Coverage within SSC exome sequencing data is highly variable and complex. In (**A**), coverage variability is shown among all well-covered loci in a randomly selected SSC individual. In (**B**), coverage variability is shown among all SSC individuals at a single well-covered locus. The best-fitting negative binomial and Poisson distributions are shown for each distribution. Neither distribution adequately models the coverage variability throughout a locus or an individual. Parameters for the distributions were calculated using the 'fitdistr' function in the MASS package using R (Venables and Ripley 2002).

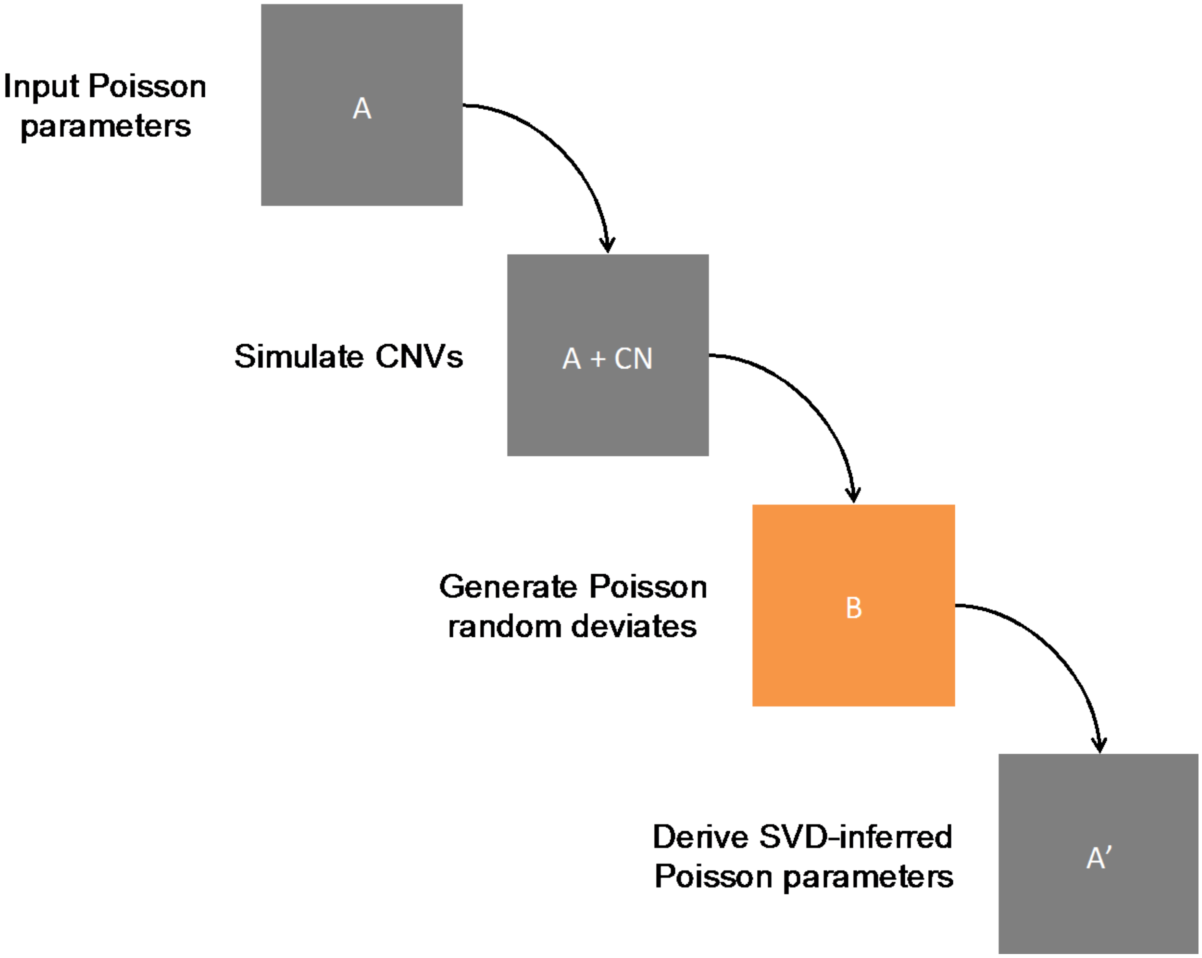

Input Poisson parameters
A
Simulate CNVs
A + CN
Generate Poisson random deviates
B
Derive SVD-inferred Poisson parameters
A’

Figure 5.11: Simulation test schematic for the SVD-derived expected coverage estimator. Poisson random deviates are generated from input Poisson parameters for per-allele coverage and CNVs randomly distributed throughout the entire test set. The SVD-inferred Poisson parameters (A’) are compared to the input Poisson parameters (A) to determine estimator accuracy.

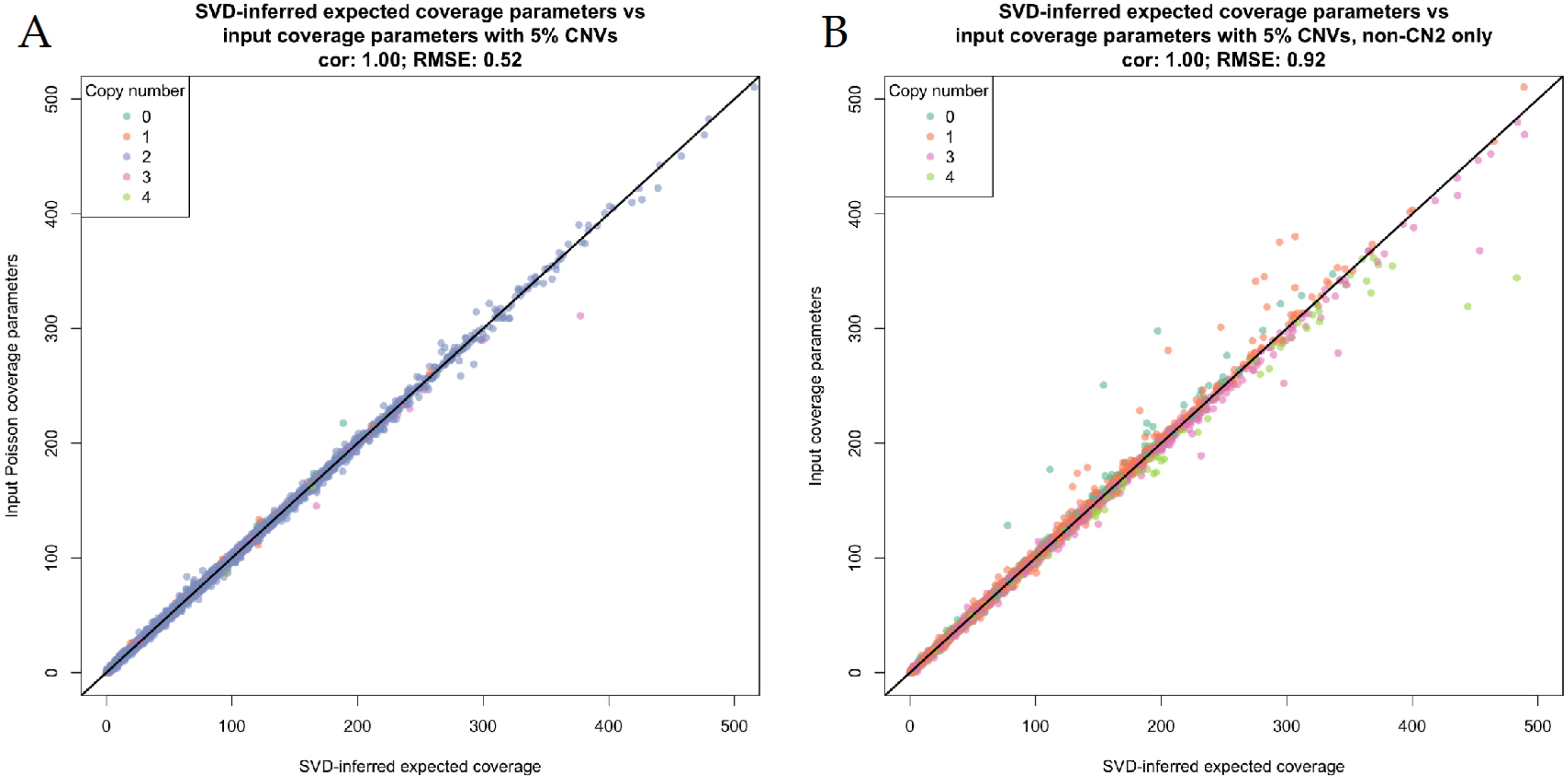


Figure 5.12: SVD-inferred coverage estimator performance on simulated data. In (**A**), a subset of 200,000 data points is shown from all 35,000,000 simulated data points. In (**B**), a subset 200,000 of data points from simulations where simulated copy number is not 2 is shown. In both plots, the line where input and expected coverage are equivalent is shown in black. Reported Pearson correlation coefficients and RMSE are given for all data in (**A**), and for all data that is not copy number 2 in (**B**). The high correlation and low RMSE of input Poisson parameters to SVD-inferred expected coverage estimators strongly suggests that our SVD algorithm is able to successfully approximate input Poisson parameters in from coverage data.

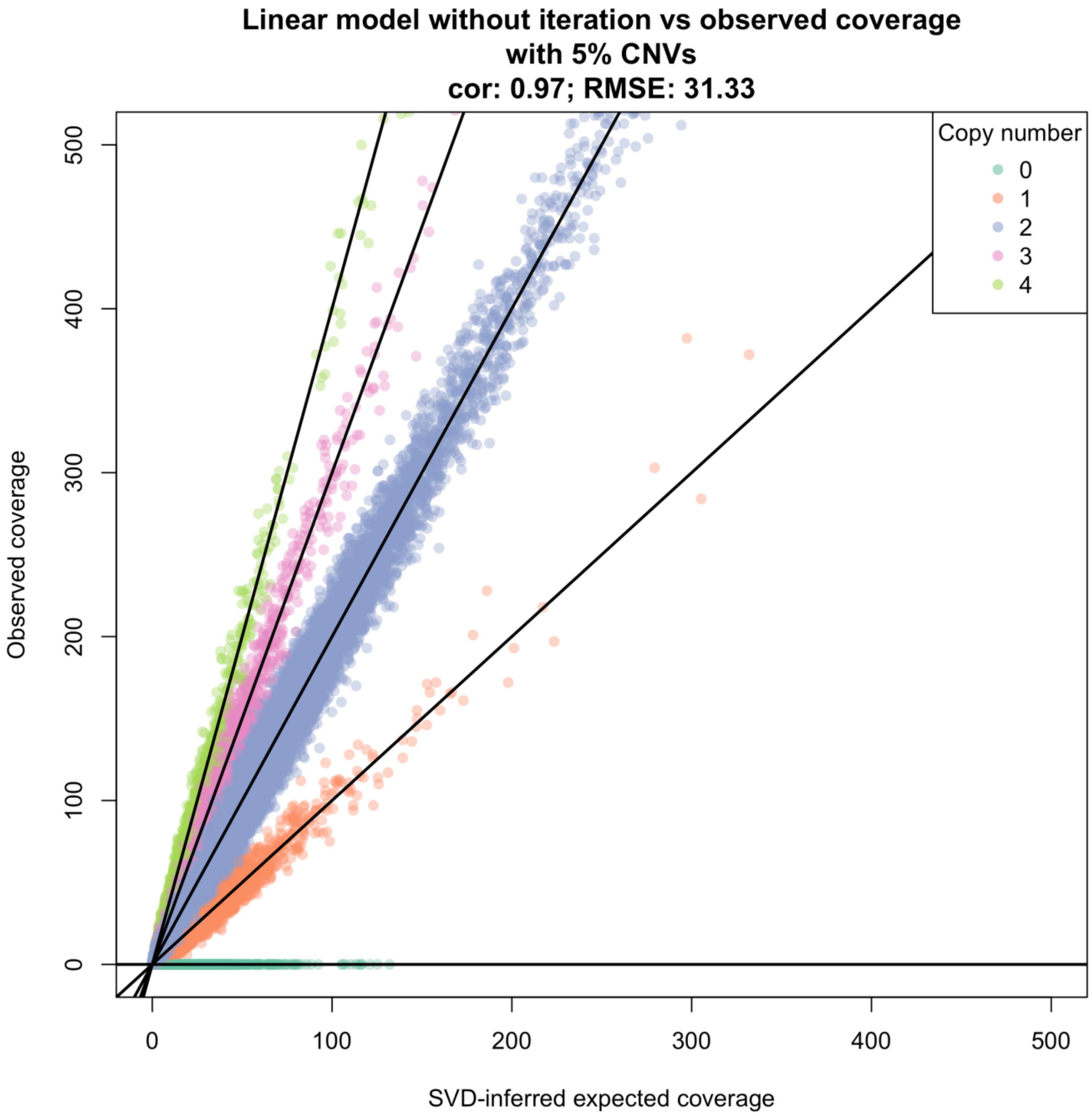


Figure 5.13: SVD-inferred coverage parameters compared to simulated observed coverage. Correlations and RMSE are shown for all data. When limited to copy number 2 simulated data only, Pearson's correlation is >0.99, while RMSE is 30. Each line represents an idealized copy state, where the observed coverage is the product of a locus's copy number state and its expected coverage. Although our SVD approximation algorithm accurately estimates input Poisson parameters, the RMSE for observed coverage and expected per-allele coverage can be large.

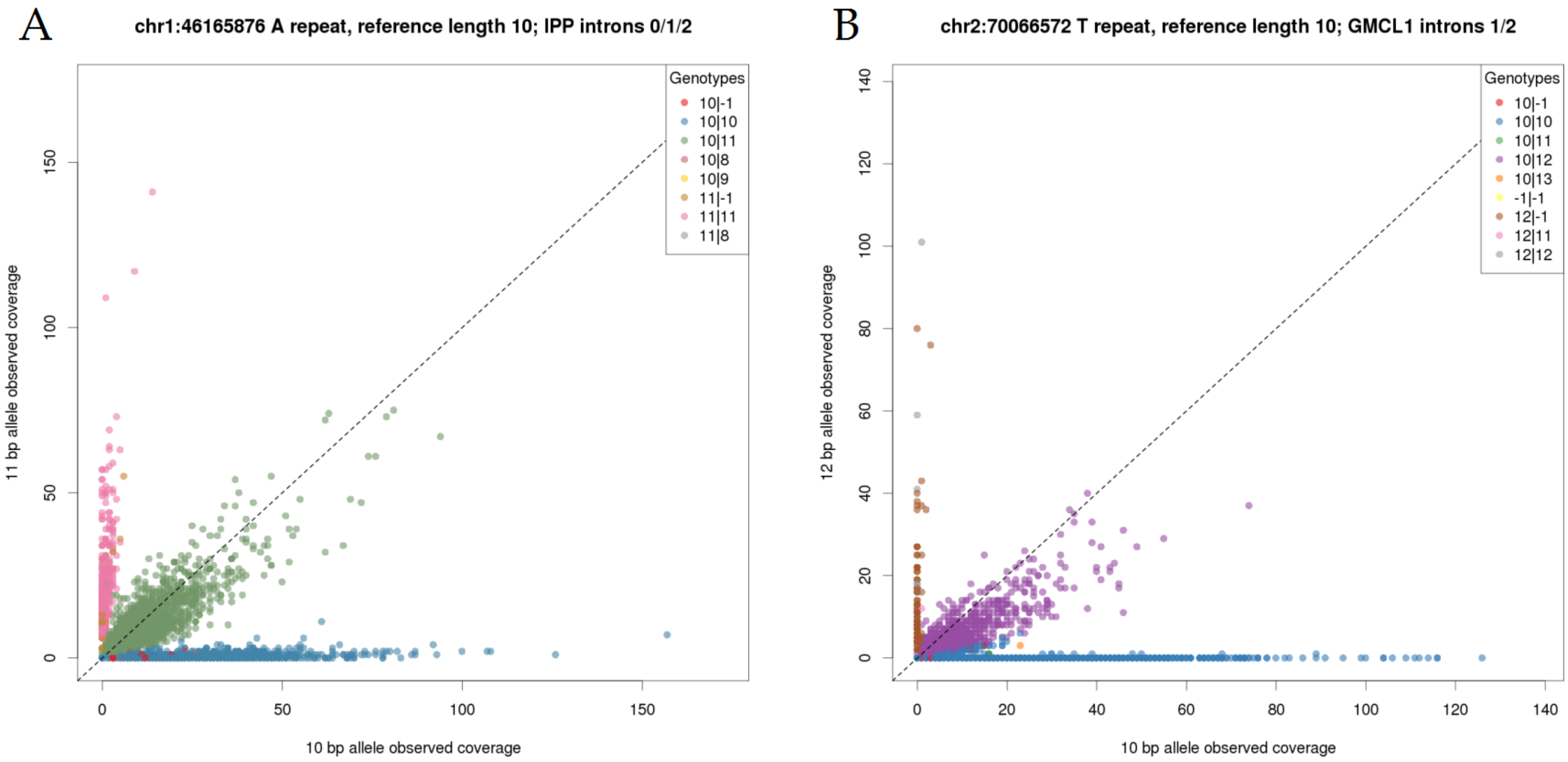


Figure 5.14: Biased and unbiased coverage in SSC data. In (**A**), a locus with two alleles and no discernible bias is shown. Points are plotted for each individual based on their coverage for both alleles, and are colored according to their most likely genotype assuming no allele bias. Both alleles have similar coverage. In (**B**), a locus with two alleles demonstrating significant bias is shown. Points are plotted as in (**A**), although genotypes have different colors. The 12 bp allele consistently has half the coverage of the 10 bp allele, which leads to extensive genotyping errors. Most noticeably, probable 12|12 genotypes are frequently called as 12|-1 (single null genotype), and many probable 10|12 genotypes are miscalled as 12|12.

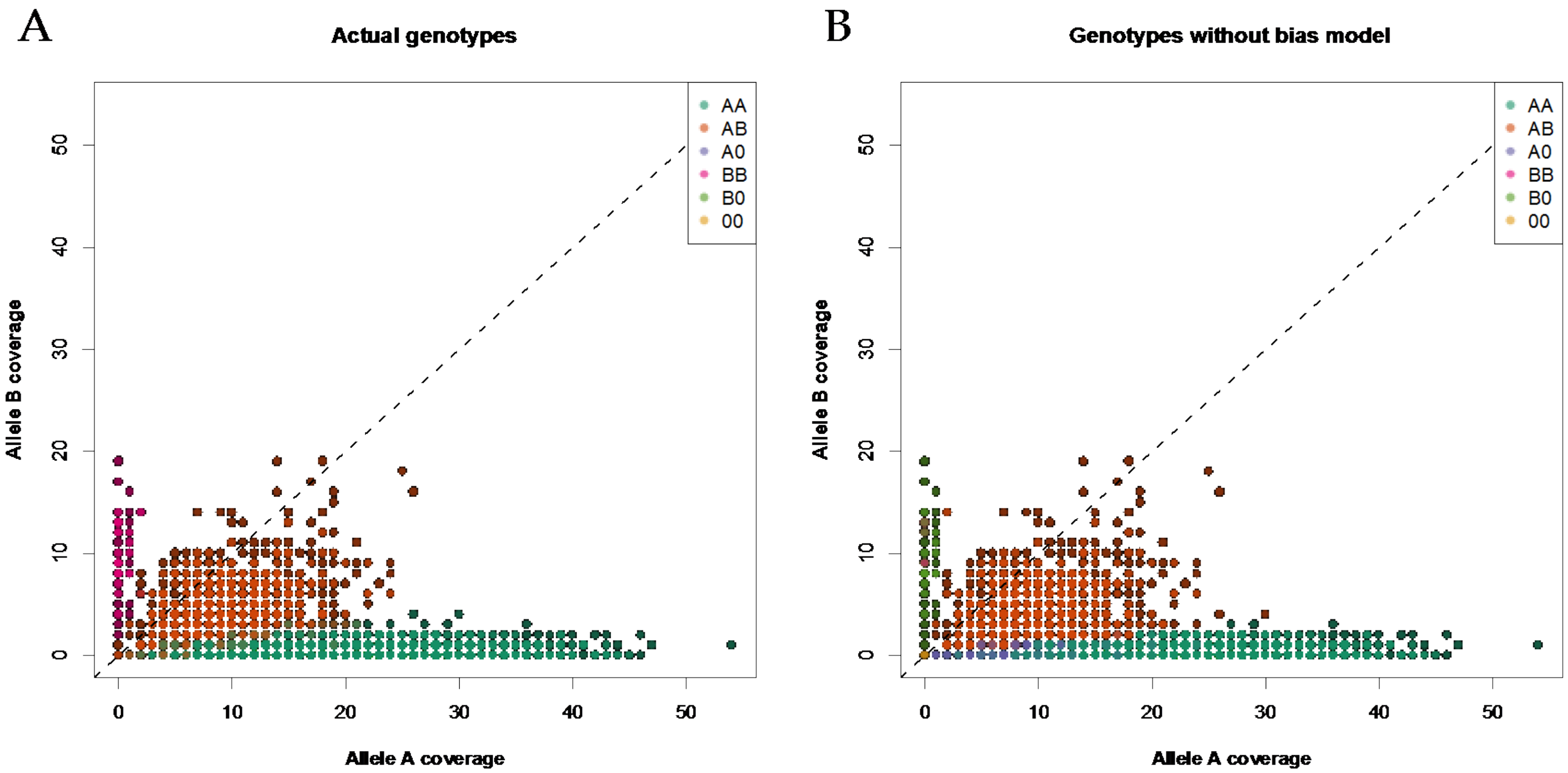


Figure 5.15: Comparing simulated genotypes to naïve genotypes. The genotypes in (**A**) are randomly assigned coverage in simulation using a per-allele expected coverage parameter and per-allele capture bias parameters. Each point is colored according to its assigned genotype. A naïve genotyper calls the genotypes in (**B**) without accounting for capture bias. Each point is colored according to the estimated genotype. Identically colored points in (**A**) and (**B**) are called correctly by the genotyper without a bias model. By comparing (**A**) and (**B**), it is apparent that the genotyper without a bias model makes frequent erroneous B0 and A0 genotype calls.

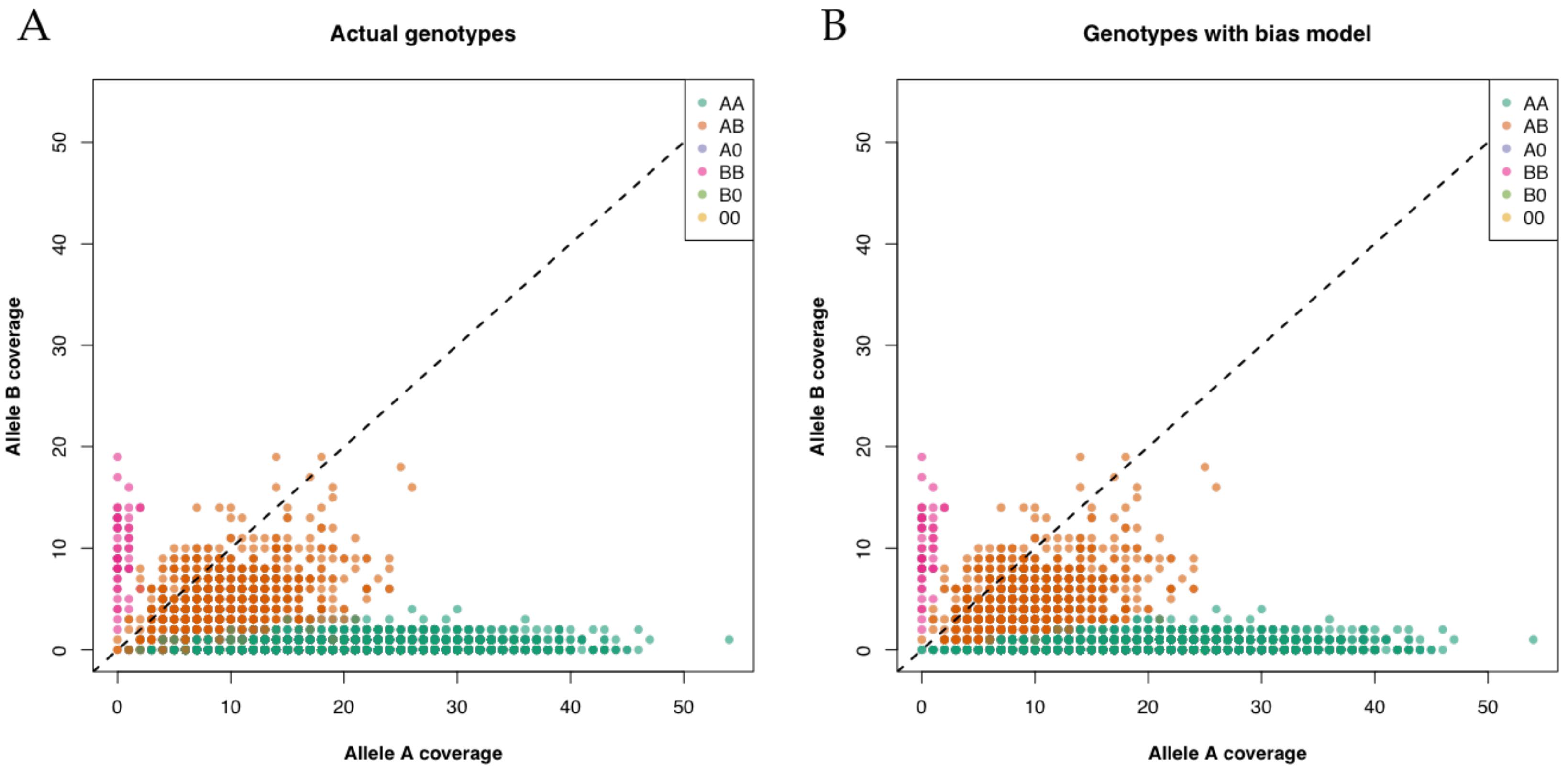


Figure 5.16 Comparing simulated genotypes to bias EM genotypes. The genotypes in (**A**) are identical to those in figure 15A. An EM genotyper that accounts for capture bias calls the genotypes in (**B**). Each point is colored according to the estimated genotype. As in Figure 5.15, identically colored points in (**A**) and (**B**) are called correctly by the genotyper with a bias model. By comparing (**A**) and (**B**), it is apparent that the genotyper incorporating a bias model calls more genotypes correctly than the genotyper without a bias model.

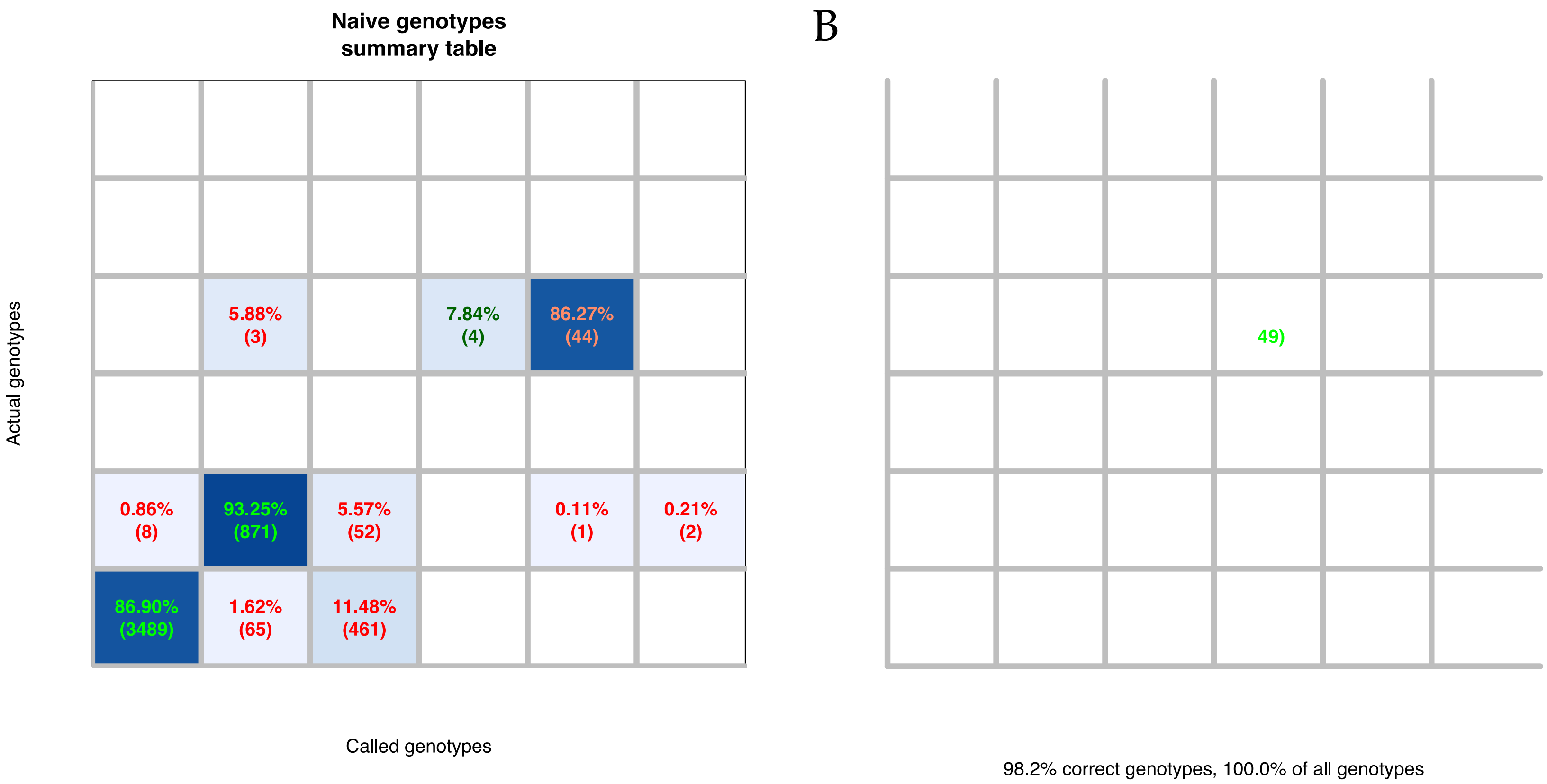


Figure 5.17: Summary table for the naïve and EM genotypers. Each row in the summary table corresponds to the genotype assigned in simulation, and each column represents a genotype called by either genotyper. All rows sum to 100%. The text in each cell is colored to reflect whether the genotypes called by a genotyper are in agreement with the simulation-assigned genotypes. Cells colored green represent correct calls, while cells colored in red represent incorrect calls. Naïve genotyper performance is presented in (**A**), while EM genotyper is presented in (**B**). The EM genotyper is considerably more accurate. Most notably in this example, the EM genotyper calls many more AA and BB genotypes correctly.

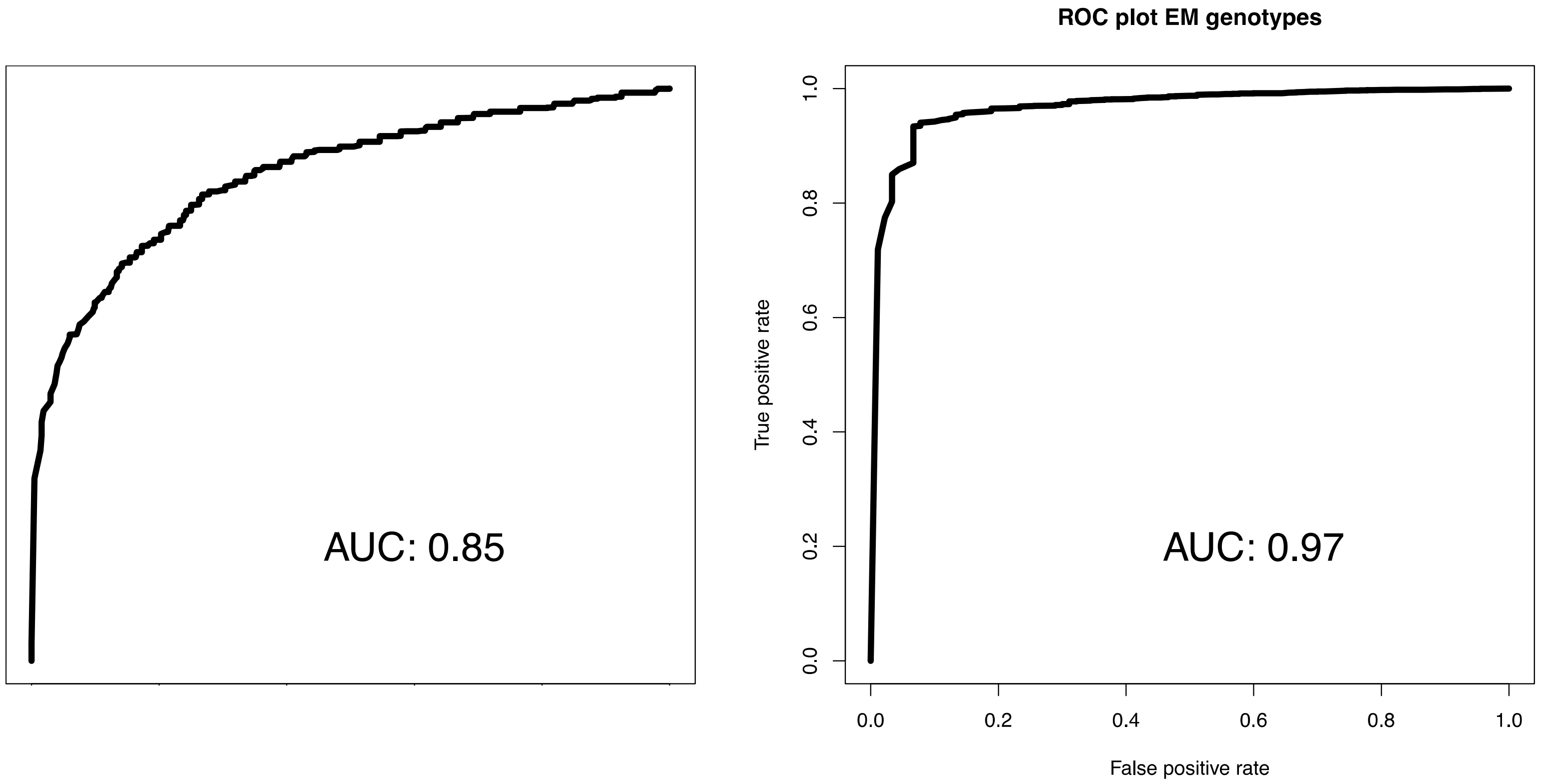


Figure 5.18: ROC plots for the naïve and EM bias model genotypers. Maintaining a low false positive rate is especially challenging in the naïve genotyper—the true positive rate drops dramatically even when the false positive rate is limited to just 0.2. The EM genotyper is able to maintain high sensitivity while maintaining a much lower false positive rate. This behavior is reflected in each genotyper's ROC AUC—the EM genotyper is more likely to give a correct genotype a higher confidence call than the naïve genotyper.

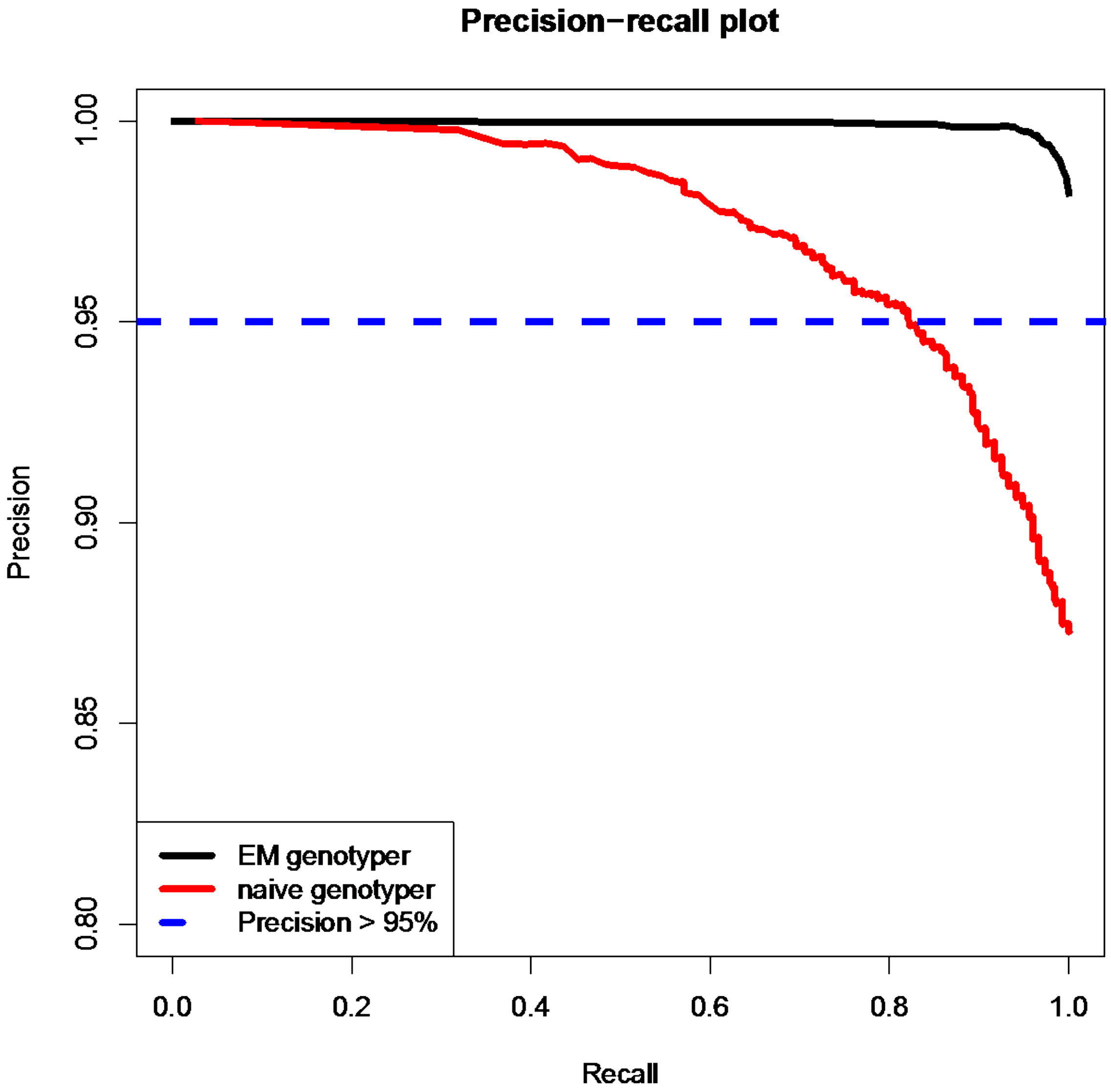


Figure 5.19: Precision-recall curves for naïve and EM genotypers. In this simulation, the EM genotyper has perfect recall while maintaining precision >0.95. The recall of the naïve genotyper is limited to ~0.8 if we demand precision >0.95. The naïve genotyper has recall of just ~0.5 if we want precision comparable to what we observe in the EM genotyper with total recall (~0.98).

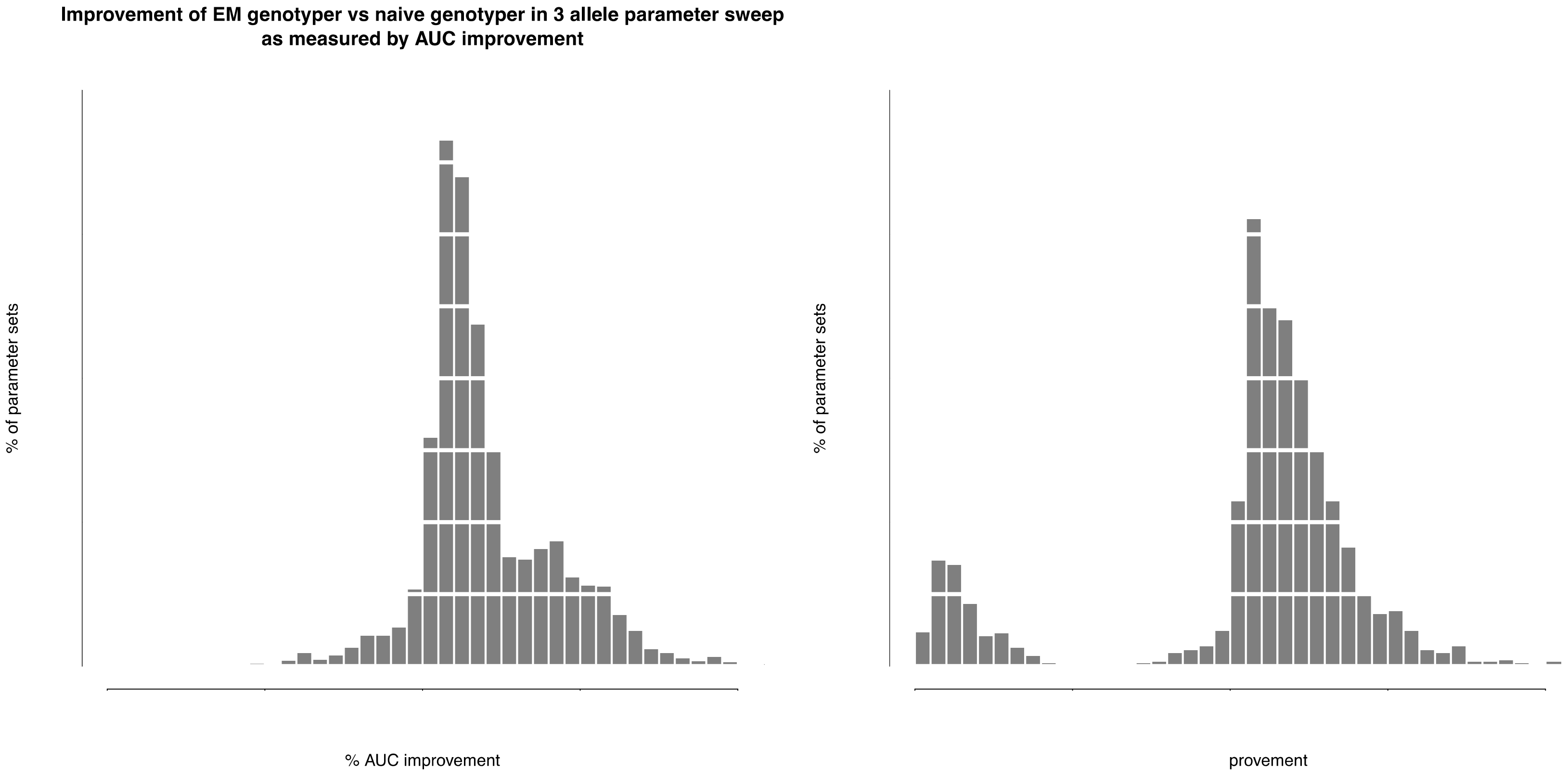


Figure 5.20: Performance of EM genotyper in parameter sweeps. Performance improvement is measured as the difference in ROC AUC between the EM and naïve genotyper divided by the ROC AUC of the naïve genotyper. A percent AUC improvement of at least 0% indicates that the EM genotyper performs at least as well as the naïve genotyper. A percent AUC improvement below 0% indicates that the naïve genotyper outperforms the EM genotyper. In (**A**), performance is shown for a range of parameter combinations at simulated loci containing three alleles. In (**B**), performance is shown for a range of parameter combinations at simulated loci containing five alleles. In both parameter sweeps, the EM genotyper consistently outperforms the naïve genotyper. The naïve genotyper occasionally has a higher ROC AUC when the locus error rate is high; expected coverage is low; and allele bias is strong.

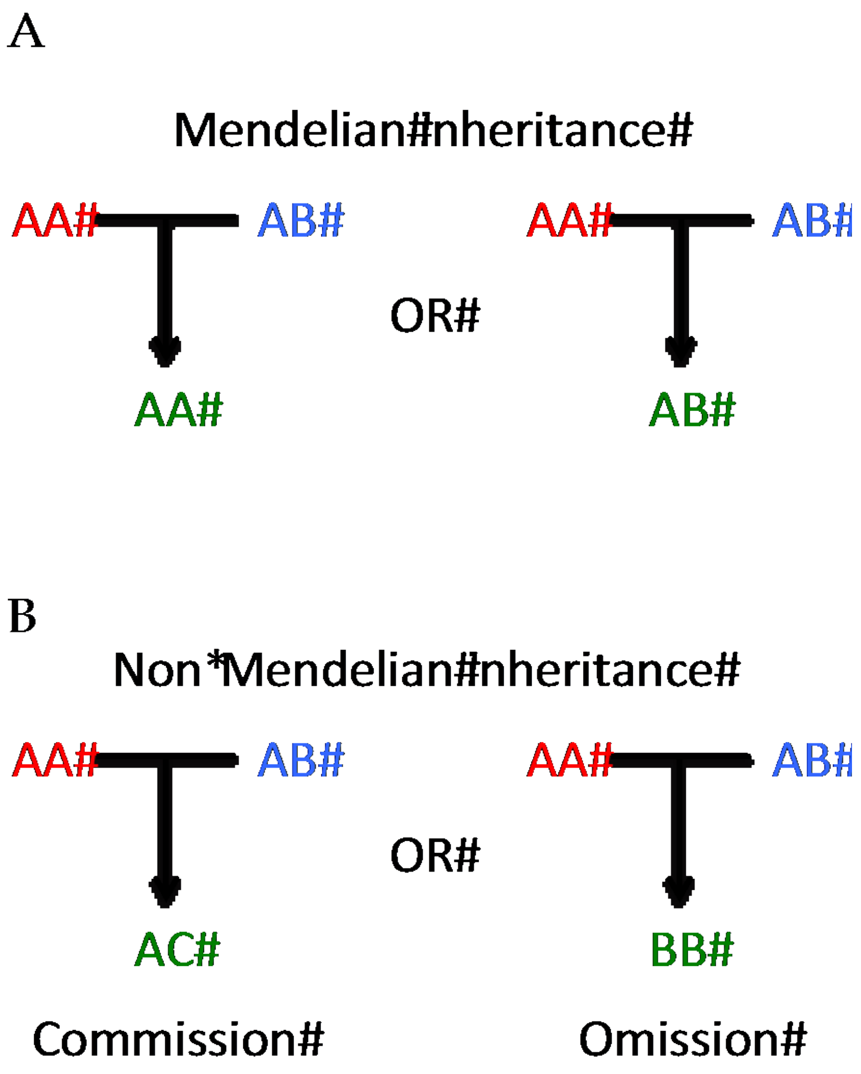


Figure 5.21: Patterns of Mendelian and non-Mendelian inheritance. In (**A**), one parent is homozygous for A, while the other is heterozygous with alleles A and B. The only possible Mendelian genotypes these parents can pass to their children are AA and AB. In (**B**), the parents have the same genotypes in (**A**), but the child either has a novel C allele not seen in either parent, or the child inherits two copies of the B allele that is only present in one parent. Both patterns represent different types of non-Mendelian inheritance.

6 Results

6.1 uSeq reference microsatellite loci

The uSeq detection algorithm describes a novel set of reference microsatellite loci. These loci differ significantly from the TRF-based reference microsatellite sets used by other microsatellite genotyping algorithms such as lobSTR or RepeatSeq (Gymrek et al. 2012; Highnam et al. 2013). uSeq detects perfect microsatellite loci with motif lengths between 1 and 6 bp, with minimum tract lengths of 8 bp, and minimum of 3 repeats of the same motif. With these parameters, uSeq identifies 5,784,968 microsatellite loci in release GRCh37 of the human reference genome, comprising ~2.4% of all the bases in the reference genome. Microsatellite frequency varies by genomic context (Figure 6.1, page 230). Of particular relevance to the exome sequencing performed as part of the SSC study, microsatellite frequency in exons is ~1.2%. This depletion of microsatellites within the exon might represent an evolutionary adaptation to minimize the effects of microsatellite instability on gene function. The microsatellite base frequency reported by uSeq is nearly double that of earlier microsatellite base frequencies (Subramanian et al. 2003).

The genome-wide mean reference microsatellite tract length is ~12 bp, and the genome-wide median tract length is 10 bp. While >98% of microsatellite loci have reference tract lengths ≤30 bp, some microsatellite loci have reference tract lengths >1,000 bp. Microsatellites in exons tend to be shorter than those in introns and intergenic regions—the longest exon microsatellites are just 69 bp (Figure 6.2, page 231). Exon mean and median reference tract lengths are shorter than other genomic contexts—the mean exon microsatellite tract length is ~10 bp, while the median is 9 bp. In GRCh37 as a whole, the mean is ~12 bp and the median is 10 bp (Table 6.1, page 289). Mean microsatellite tract lengths are 14% – 30% shorter in exons for motif lengths whose slippage products would induce frameshift mutations. In contrast, mean microsatellite tract lengths for 3 and 6 bp motifs are nearly indistinguishable from their genome-wide counterparts.

The most common genome-wide motif lengths are 1, 2, and 3 bp, which comprise 24% – 33% of all microsatellite loci in the reference set. Motif length distributions do not vary widely among most genomic contexts (Figure 6.3, page 232). There are two dramatic exceptions to this behavior: miRNAs are enriched for 2 bp motifs (21/35; 60%), and exons are enriched for 3 bp motifs (32,845/42,028; 78%). While the interpretation of the 2 bp motif enrichment at miRNA loci is unclear, the enrichment of 3 bp motifs in exons is somewhat intuitive. By favoring 3 bp motifs, most microsatellites in exons would maintain the proper coding frame if they underwent slippage mutations, minimizing the opportunities for highly deleterious frameshift mutations. Tandem amino acid repeats in a protein primary sequence could still be composed of degenerate codons to minimize the number of slippage-prone sites. However, many short microsatellites with 3 bp motifs still occur, suggesting that while the genome has selected against frameshift-prone exon microsatellites, there does not appear to be a similar effect on microsatellites that would induce in-frame indels. The mechanism by which 3 bp motifs are enriched in exons—as opposed to leveraging the degeneracy of the genetic code to minimize the occurrence of all microsatellite tracts—is unclear.

Throughout the entire genome, the most common motif equivalence is the A class, which represents >30% of all microsatellite tracts, and 98% of all 1 bp microsatellite motifs (Figure 6.4, page 233). Although A class microsatellites are among the 10 most common equivalence classes in exons, they are only ~2.5% of all microsatellite in exons (Figure 6.5, page 234). Of the ten most common equivalence classes in exons, seven have 3 bp motifs, however, AC, AG, and A class microsatellites still account for >16% of all microsatellites in exons. Overall, microsatellites with the potential to cause frameshift mutations compose ~21% of all microsatellites in exons.

C class motifs are much more common in exons than they are in the rest of the genome—nearly 15% of all exon microsatellites with 1 bp motifs have a G class motif (Table 6.2, page 289). In contrast, G class motifs comprise just 1.6% of genome-wide microsatellites with 1 bp motifs. AT, AC, and AG classes each comprise over 30% of all genome-wide microsatellites with 2 bp motifs, and only 0.7% of all microsatellites with 2 bp motifs are in the CG class (Table 6.3, page 289). The distribution of 2 bp microsatellite equivalence classes in exons differs markedly from their distribution genome-wide. CG class motifs are >1400% more common in exons than they are genome-wide, AG class motifs are ~71% more frequent in exons than they are genome-wide, and AT microsatellites are only found with 13% of the frequency as they are genome-wide. The most common genome-wide microsatellites with 3 bp motifs are from the AAT, ACC, AAG, AGG, and AAC classes, with frequencies from 25% to 10%. Once again, exons exhibit a markedly different distribution of microsatellite equivalence classes. The AGC, AGG, AAG, CCG, and ACC classes compose 25% to 10% of all exon microsatellites with 3 bp motifs, while the AAC and AAT classes are <5% of exon microsatellites with 3 bp motifs (Table 6.4, page 290).

There are 33 equivalence classes for motif lengths of at least 4 bp, which is somewhat unwieldy, so only a brief overview will be provided here. The most common genome-wide microsatellites with 4 bp motifs are from the AAAC, AAAG, AAAT, and AAGG equivalence classes, with frequencies ranging from 24% to 8%. The most common exon microsatellites with 4 bp motifs are from the AAAG, AGCC, AGGC, and CCCG classes, with frequencies ranging from 18% to 7%. While the CCCG class is common in exon microsatellites with 4 bp motifs, it is only present in 0.3% of genome-wide microsatellites with 4 bp motifs. The most common genome-wide microsatellites with 5 bp motifs are from the AAAAC, AAAAG, and AAAAT classes, with frequencies from 30% to 15%. No other motif has a frequency above 5%. There are only 116 exon microsatellites with 5 bp motifs. Of those, most are from the AGCCC, AAAAG, CCGCG, AAGAG, and CCCCG classes, with frequencies from 15% to 8%. Aside from the AAAAG class, the most common 5 bp motif equivalence classes in exons occur with frequencies <2% genome-wide. The most common genome-wide microsatellites with 6 bp motifs are from the AAAAAC, AAAAAG, and AAAAAT classes, with frequencies from 28% to 13%. Of the 413 exon microsatellites with 6 bp motifs, many occur with frequencies from 7% to 1%. Of the most common genome-wide microsatellites with 6 bp motifs, the AAAAAG class is the only one observed in exons. In general, motifs with high GC content appear to be more prevalent in exon microsatellites than they are in microsatellites from introns and intergenic regions.

6.2 SSC whole exome sequencing data

In this study, exome sequencing data from 875 SSC quads (3,500 individuals) was analyzed for *de novo* microsatellite mutations. No families with more or less than four individuals were considered. Within this dataset, 787 of children with autism are male, and only 81 are female, which is a nearly 10-fold enrichment of males versus females with autism. This is nearly double the sex bias reported by the CDC (Centers for Disease Control and Prevention (CDC) 2014). Siblings are nearly equally likely to be male or female—463 siblings are females, while 405 are male. Each member of the cohort was subject to exome capture using the NimbleGen SeqCap EZ Exome v2.0, then barcoded, pooled, and sequenced on the Illumina HiSeq 2000 platform using 101-bp paired end reads.

6.3 uSeq microsatellite mapping performance

6.3.1 General mapping statistics

An average of 9% of all read pairs contain at least one microsatellite in the SSC dataset (Figure 6.6, page 235). After aligning microsatellite reads from SSC individuals to the condensed reference genome, an average of 8% of reads failed to align (Figure 6.7, page 236). An additional 15% of reads had mapping quality scores below 30, which we have chosen as our mapping quality threshold (Figure 6.8, page 237). Only ~1% of all read pairs align to disparate chromosomes (Figure 6.9, page 238).

6.3.2 PCR replicate statistics

On average, 4% of read pairs are in PCR replicate sets (Figure 6.10, page 239). The number of distinct replicate sets varies by SSC individual, with an average of ~400,000 replicate sets (Figure 6.11, page 240). Most PCR replicate sets are very stable—only 0.6% of PCR replicate sets are discordant in the average SSC individual (Figure 6.12, page 241). The observed stability within the PCR replicate sets is an encouraging indication of microsatellite stability during the SSC sample preparation protocol. As discussed in section 5.2.4.1, we expect that most discordance in replicate sets would be due to slippage events occurring at microsatellite loci during PCR. The fact that PCR replicate set discordance is so low indicates that most microsatellite loci exhibit a remarkable amount of stability during sample preparation. Therefore, we would not expect PCR slippage noise to be a major complication when calling genotypes at most loci.

6.3.3 Overlapping read pairs

Nearly 7% of all read pairs in an average SSC individual overlap and report the same microsatellite locus (Figure 6.13, page 242). An overlapping read pair occurs when a sequencing fragment is shorter than twice the read length. The relatively high proportion of overlapping reads suggests that many read fragments in SSC individuals are less than 200 bp in length. Most overlapping reads report identical microsatellite tract lengths—only 0.3% of all overlapping read pairs are discordant in the average SSC individual (Figure 6.14, page 243). This provides further encouragement regarding the stability of most microsatellites detected in the SSC dataset. A read pair would be most likely to report discordant tract lengths due to base miscalls or other errors introduced by the sequencing instrument. The low read pair discordance rate indicates that the Illumina HiSeq 2000 instrument used to sequence microsatellite reads is not especially error-prone at most microsatellite loci.

6.3.4 SSC microsatellite profiles

Since the data from the SSC dataset is obtained from exome sequencing, we do not expect to capture anywhere near the complete complement of microsatellite loci in GRCh37. While we expect that loci within the exome target region will have the most consistent coverage in the SSC population, loci outside of the target region may have sporadic coverage in a subset of SSC individuals. We see an average of ~1,000,000 profiles in each SSC individual (Figure 6.15, page 244). Since there are only ~40,000 microsatellite loci in exons, this represents extensive off-target coverage throughout the genome.

Although uSeq is not limited to detecting microsatellite loci that are in its reference set, >96% of the reported microsatellite profiles in an average SSC individual are at reference microsatellite loci (Figure 6.16, page 245). This suggests that the reference microsatellite set defined by uSeq captures a significant portion of the full microsatellite complement of many human genomes. An abundance of non-reference microsatellite profiles could indicate that GRCh37 is a poor reference for microsatellite loci. Frequent non-reference profiles would imply a high rate of microsatellite emergence within human genomes or frequent mutations that affect a microsatellite's detectability. We see no evidence to support either assumption.

A non-reference microsatellite profile may not represent a truly novel microsatellite locus. As discussed in section 5.2.2.4.1, sequencing reads with incorrect alignments and undetected terminal repeats usually align near their true position within the reference genome. This error mode can produce significant coverage at spurious microsatellite loci. For this reason, we limit our SSC analysis to reference microsatellite loci only. uSeq may be able to identify novel microsatellite loci in the SSC if it incorporated a local realignment step.

6.3.5 Well-covered microsatellite loci

HTS data from populations of individuals is essential to uSeq's current genotyping strategy. As a means of guaranteeing accurate genotypes, "well-covered" loci are identified using ad hoc filters. In total, 5,453,739 reference microsatellite loci are observed in at least one SSC individual, which is ~94% of all reference loci. In simulations of single-ended microsatellite reads, just under 90% of microsatellite loci have a high quality alignment with 80 bp of flanking sequence. Paired-end sequencing data provides a clear improvement in mapping microsatellite loci—more reads containing microsatellites can be aligned to reference loci with less flanking sequence.

Most uSeq reference microsatellite loci are observed in a fraction of SSC individuals (Figure 6.17, page 246). The average number of SSC individuals covered at a reference microsatellite locus is ~650, while the median is 415. This indicates a heavily skewed distribution with most loci only observed in <30% of all SSC individuals. The majority of reference microsatellite loci reported in any SSC individual's profile by uSeq will be due to sporadic coverage outside the exome.

Microsatellite loci with higher maximum locus coverage tend to have coverage in more SSC individuals (Figure 6.18, page 247). This would indicate that most loci consistently observed in the SSC dataset are likely to be in or near exome target regions, and would accordingly have higher coverage. The maximum locus coverage is the highest observed coverage for any SSC individual at microsatellite locus, and is highly correlated both with the mean locus coverage and total locus coverage, which are other measures of how consistently a locus is captured within the SSC (all pairwise Pearson correlations >0.97).

To identify reference microsatellite loci that are consistently covered throughout the SSC population, we limit our analyses to loci detected in at least 60% of SSC individuals with maximum locus coverage of at least 25 (Figure 6.18, page 247). These filters limit our analyses to 133,300 reference microsatellite loci, or ~2.3% of uSeq reference microsatellite loci. 83% of all exon microsatellite loci in the uSeq reference set are observed in the SSC dataset, a dramatic enrichment as compared to other genomic contexts (Table 6.5, page 290). Introns comprise 57% of all well-covered loci, and another 26% of all well-covered loci are in exons. Well-covered loci span all genomic contexts (Figure 6.19, page 248).

The NimbleGen EZ Exome V2.0 capture kit has probes designed primarily to target NCBI RefSeq protein-coding regions and CCDS genes from September 2009; miRBase miRNAs from September 2009 (version 14); and additional customer inputs. Probes designed to target these regions cover a total of 44.1 Mb of the reference genome (Roche Nimblegen 2010). Of the 133,300 well-covered loci in the SSC dataset, ~81% are either directly targeted by a capture probe or are within 200 bp of a capture probe (Figure 6.20, page 249). The remaining 19% of well-covered loci are <3 Mb from the nearest probe. There is a clear relationship between the distance to the nearest EZ Exome capture probe and the mean locus coverage (Figure 6.21, page 250). Similarly, the median per-locus mean coverage is ~15X for all well-covered loci in the SSC dataset, while the median per-locus mean coverage is ~7X for well-covered loci that are more than 200 bp from the nearest capture probe. Well-covered loci that are far from specified capture regions may have consistent coverage within the SSC population due to weak homology of other genome regions to capture-targeted regions.

6.3.6 uSeq coverage in whole exome and whole genome sequencing data

In the interest of determining the reproducibility of uSeq microsatellite genotypes, whole genome sequencing data was obtained from two SSC families. This data can serve as a sort of technical replicate, since the biological material is the same, although the sequencing strategy is different. It is also useful in understanding the particular biases introduced by exome sequencing. Whole genome sequencing was obtained using the Illumina HiSeq 2000 platform using 101-bp paired end reads.

As a means of determining the consistency of uSeq genotypes in the same individuals with different experimental protocols, we compared the allele coverage for confidently genotyped alleles from whole genome and exome sequencing data. More then 97% of high-confidence homozygous genotypes have >90% of total coverage at the same genotyped allele in both datasets (Figure 6.22A, page 251). The same trend appears to be true for the first allele in high-confidence heterozygous genotypes (Figure 6.22B, page 251). These results confirm that microsatellite loci are consistently detected by uSeq. When a genotype is called by uSeq with high confidence, the pipeline can reproducibly detect sequencing data consistent with the genotype from the same biological sample independent of the experimental protocol used for sequencing.

6.3.7 Comparing uSeq coverage to standard laboratory pipeline

At reference microsatellite loci, uSeq coverage is generally as good or better than a GATK-based pipeline that realigns and recalibrates indels (Figure 6.23, page 252). This is true when the GATK-aligned reads are subject to the same filters as uSeq-aligned reads, i.e. only one overlapping read is counted towards coverage in overlapping read pairs, reads must have a minimum mapping quality score of 30, and the read must completely contain the microsatellite. Among all loci with at least 10X coverage in either pipeline, uSeq has ~6% more reads on average. This demonstrates that the uSeq alignment strategy is comparable to commonly used indel-detection pipelines at microsatellite loci. When spot-checked, the most common reasons for higher coverage in the GATK-based pipeline are SNVs or indels interrupting the microsatellite, which could be true genomic variation or sequencing artifacts. Additionally, reads may have different mapping quality scores in either pipeline, so reads that might be above the mapping quality threshold in one pipeline could fall below that threshold in the other.

There are some loci where uSeq reports no coverage while GATK reports coverage. While analysis has not been exhaustive, all cases analyzed so far where uSeq has no coverage appear to be due to microsatellite regions that are frequently duplicated in human genomes.

This can arise when two regions in the reference genome have identical flanking sequence but differing microsatellite tract lengths. In these situations, GATK-based pipelines may align reads uniquely to one of the regions due to the microsatellite tract length. Since the flanking sequence for both loci are identical, uSeq will consider any reads mapping to these regions as having multiple reference genome alignments, and will therefore not consider them.

6.4 uSeq genotyping

6.4.1 Per-allele, per-person expected coverage estimates

The uSeq genotyper relies on per-allele, per-person expected coverage estimates to assess whether a genotype has observed coverage consistent with a bi-allelic genotype model. As described in section 5.3.3.2.2, a high correlation between observed coverage and expected coverage estimators indicates that the SVD-estimated parameters are an accurate representation of the underlying, non-biological processes influencing coverage. The mean correlation between observed coverage and expected coverage for SSC individuals is >0.98, demonstrating the reliability of the SVD-derived expected coverage estimators in real sequencing data (Figure 6.24, page 253).

Of the 3,500 people in the SSC dataset, four people from three families had low correlations between expected and observed coverage (<0.8) and exceptionally low total coverage. Families containing these individuals were excluded from genotyping. One SSC individual had exceptionally low coverage in the dataset, and the family containing this individual was excluded as well. The median total read count in the SSC dataset is 3,177,601; every individual with poor correlation and/or bad coverage had a total read count between 11 and 309,753. Finally, all individuals in this study had DNA obtained from whole blood, except for individuals from three families. These families were also excluded from genotyping. Therefore, genotyping was performed on 3,472 SSC individuals.

When analyzing observed coverage in different genomic contexts, there is a clear difference in the coverage distribution at microsatellite in exons and outside of exons (Figure 6.25, page 254). The median coverage across all SSC individuals in all well-covered exon microsatellite loci is 33X, indicating robust coverage for exon microsatellites throughout the dataset. The exon coverage distribution is very overdispersed, with a long tail that extends well past 100X coverage. This behavior is reflective of the coverage variation among SSC individuals, as well as the variation in probe capture efficiency for different microsatellite loci. The median coverage among all SSC individuals at loci outside of exons is 10X, a dramatic decrease from exonic coverage. While the non-exon coverage distribution still exhibits some overdispersion, it is not nearly as dramatic as it is in the exon coverage distribution. Although the high median exon coverage is very favorable for *de novo* mutation detection, the median coverage for non-exon loci is less promising. It is likely that some *de novo* mutations outside of exons will be missed—low coverage in any trio member will lead to less certain genotype calls, which can deflate Mendel obedience scores, even for true *de novo* mutations.

6.4.2 Locus-specific error rate estimates

The median EM-estimated locus noise rate is ~0.001, while the mean noise rate is 0.013. Both summary statistics describe a noise rate distribution that suggests that noise will not be a confounding factor for genotyping at most microsatellite loci. Error rates are strongly influenced by microsatellite motif length (Figure 6.26, page 255). Although most microsatellite with motifs of at least 3 bp appear to be incredibly stable, microsatellites with 1 and 2 bp motifs often have error rates greater than 0.01. Most notably, many microsatellite loci with 1 bp motifs appear to be highly unstable—17% of 1 bp motifs have estimated error rates above 0.10. Only 0.9% of microsatellites with motifs longer than 2 bp have noise rates in excess of 0.01.

In addition to clear motif length dependence, microsatellites appear to be increasingly unstable for longer reference tract lengths (Figure 6.27, page 256). This is particularly apparent for microsatellites with 1 bp motifs—these microsatellites regularly have slippage rates above 0.10 for tract lengths as short as 11 or 12 bp (Figure 6.27A, page 256). As microsatellites with 1 bp motifs extend, slippage rates might be too high to be able to call genotypes accurately. If this slippage is due to the particular methods of sample preparation and not somatic variation, it is likely that protocols involving less PCR might make it possible to genotype longer microsatellites with 1 bp motifs. The increase in error rate with longer tract lengths for microsatellites with longer motifs is much less dramatic (Figure 6.27B, page 256). Microsatellites with 2 bp motifs regularly have error rates below 0.10 independent of reference tract length, and error rates above 0.10 are almost never observed for microsatellites with motifs of at least 3 bp.

The error rate distributions in Figure 6.26 and 6.27 (pages 255 – 256) have a distinct bimodal form that requires consideration. For all motif lengths larger than 1 bp, the first distribution has a mean of approximately 0.005, and the second distribution has a mean slightly greater than 0.001. When analyzing the tract length distributions for 1 and 2 bp motifs in Figure 6.27 on page 256, this first distribution is almost entirely attributable to microsatellites with the minimum detectable tract length for either motif length. This lower peak is likely due to slippage events at these microsatellites that are undetectable by uSeq, which suggests that error rate estimates for the minimum detectable tract lengths may be slightly underestimated. Even at longer tract lengths, error rates below 0.01 are regularly observed. This suggests that the shortest detected microsatellites are highly stable and can be genotyped accurately.

Even when accounting for motif length and reference tract length, the distribution of locus-specific error rate estimates varies considerably over several orders of magnitude. For instance, error rates for 2 bp microsatellite motifs with a reference tract length of 9 bp can vary from as low as ~0.0001 to as high as ~0.01. This indicates that error rate estimated from these characteristics might be highly inaccurate for any specific locus. Two other common HTS microsatellite genotypers estimate error rates for a particular microsatellite locus by considering motif length and tract length, as well as additional parameters. lobSTR also considers GC content and STR purity, while RepeatSeq only adds base-calling quality score to its error rate estimation (Gymrek et al. 2012; Highnam et al. 2013). While these models may describe the average error rates for a particular allele length or reference locus, they are almost certainly incapable of encompassing the full variation in error rates due to parameters not included in their models. uSeq is the only microsatellite pipeline to account for this variability, and it does so by considering population-level information at each locus. The uSeq genotyper's superior accuracy in locus error rate estimates contributes to more accurate genotype calls, which in turn allow for specific identification of *de novo* mutations.

The error rate estimates at each locus reflect the MLE error rate estimate derived from the EM, provided the uSeq genotyping model is correct for an individual and/or locus. If a locus is not bi-allelic in an individual, or the error rate cannot be described by a binomial distribution, the estimated error rate will not be accurate. These assumptions could be violated at loci with high levels of somatic instability. Although it is reasonable to assume that most loci are somatically stable, some of the least stable loci we observe could be a product of population-wide somatic instability. In addition, while the per-locus error rate describes the behavior of most individuals at the locus accurately, some individuals at the locus may have unexpectedly high error due to somatic mutation. In these cases, the noise goodness-of-fit test could flag individuals with potential somatic mutations at a locus.

6.4.3 Per-allele bias estimates

Per-allele capture bias differs dramatically for reference and non-reference microsatellite loci (Figure 6.28, page 257). The mean bias for reference alleles is 1.05, which tracks closely to the median of 1.01, indicating a relatively unskewed distribution. As a reminder, a bias estimate near 1 indicates that a microsatellite allele's coverage in the population is consistent with each individual's per-allele expected coverage. If an allele's bias is below 1, it is consistently observed with less than expected coverage throughout the population. Similarly, if an allele's bias is above 1, it is consistently observed with more coverage than expected within the population. The reference allele bias distribution is narrowly distributed around 1, which implies that reference alleles are usually observed with coverage consistent with their expected coverage estimators (Figure 6.28A, page 257).

Non-reference alleles have capture biases that vary widely (Figure 6.28B, page 257). The mean bias for non-reference alleles is 0.46, while the median is 0.26, which suggests a heavily skewed distribution. This is clearly demonstrated by the distribution of non-reference allele biases, which exhibits bimodal behavior. Many non-reference alleles have extreme biases—almost 44% of all non-reference alleles have biases ≤0.15. Less than 30%

of non-reference alleles have capture biases between 0.75 and 1.25. In contrast, 94% of capture bias estimates for reference microsatellite alleles fall within this range. This suggests that exome capture can be extremely inefficient for non-reference alleles. When interpreting microsatellite genotypes—and almost certainly indels in general—there is a considerable chance that many non-reference alleles are not consistently observed within a study population. Without a model for capture bias, any indel or microsatellite genotyper could have significantly compromised accuracy. Although not addressed in this study, this clear bias at microsatellite loci may be a general feature of indel and microsatellite sequencing studies independent of sample preparation protocol.

Allele capture bias will be especially problematic when attempting to call *de novo* mutations. Some true *de novo* mutations may be missed when a novel allele in a child is assumed noise due to extreme capture bias. More troubling, spurious microsatellite *de novo* mutations could be called when a parent has low coverage at a putative novel allele due to capture bias. By modeling allele capture bias, uSeq is able to consider *de novo* mutations with considerable specificity. In trios where a child has a novel allele with strong capture bias, the genotyper would require higher coverage to demonstrate that the allele is truly novel and not due to noise. Alternatively, when a child has an apparently novel allele that has extreme bias, any parental coverage for the same allele would be a strong indication that there is no *de novo* mutation within that trio at that locus. Additional consideration for parental coverage at biased alleles is especially important for improving the specificity of *de novo* mutation identification.

Allele capture bias is only capable of including an allele in a genotype if it is distinguishable from noise coverage at its locus. If a locus has a very high noise rates, an allele with very strong capture bias may not be detected since it would appear to be no different than noise at the locus. This would in turn lead to decreased sensitivity to detect *de novo* mutations for alleles with strong bias or high noise rates.

6.4.4 Genotype quality statistics

6.4.4.1 Genotype confidence

In total, 462,765,520 genotypes are called in the SSC dataset. Most genotypes are called confidently—the median genotype confidence for all loci and SSC individuals is >0.99 (Figure 6.29, page 258). No SSC individual has uniformly low genotype confidence. 82% of all well-covered loci have a median confidence above 0.9. This indicates that some loci do not have clear bi-allelic genotype calls that can be called confidently. The most probable explanations for low locus median confidence are low coverage, high levels of noise, or extremely biased alleles. Any of these factors could make it difficult to confidently distinguish among possible genotypes at a particular locus.

6.4.4.2 Allelic goodness-of-fit

Nearly 16% of all allelic goodness-of-fit p-values are 1, while the rest appear to be uniformly distributed between 0 and 0.9 (Figure 6.30, page 259). This bias can be explained by loci with both low expected and observed coverage. The two-sided Poisson exact test will return a p-value of 1 if the observed coverage is equal to the expected coverage or one less than the expected coverage. When expected coverage is low, there is a good chance of observing either of these two values by chance. For instance, if we apply our allelic goodness-of-fit test to a simulated dataset of individuals with low coverage sampled from a Poisson distribution with a rate of 10, the p-value distribution is strikingly similar to the one observed in the SSC dataset. As the median expected coverage increases, this distribution becomes increasingly uniform. This suggests that non-exonic loci in the SSC are likely to be the primary contributors to the unusual behavior of our observed distribution. When expected coverage is low, it will be difficult to confidently distinguish a bi-allelic genotype from a genotype with one or more null alleles. In these situations, a high marginal null probability can be particularly informative when determining whether a bi-allelic genotype model is supported by the data.

Median allelic goodness-of-fit scores for each SSC individual range between 0.3 and 0.6, which would indicate that there are no individuals with particularly poor overall fit to their expected coverage estimators and the appropriate allele bias estimates. 90% of locus median allelic goodness-of-fit scores are between 0.43 and 0.83, indicating that most loci do not appear to have generally poor allelic goodness-of-fit scores. Some loci do have uniformly low allelic goodness-of-fit scores. These loci likely represent locus-specific phenomena that cause significant deviations between expected and observed coverage. These could be biological in nature or they could represent some other systematic noise that did not contribute significantly to overall coverage, and therefore did not exceed the required threshold to be included in the expected coverage estimator matrix.

6.4.4.3 Noise goodness-of-fit

Noise goodness-of-fit primarily consists of two regimes—goodness-of-fit p-values near 1, and goodness-of-fit p-values near 0 (Figure 6.31, page 260). More than 85% of all genotypes have noise goodness-of-fit p-values of 1, while another 6% have goodness-of-fit p-values less than 0.01. This behavior is primarily due to the behavior of the one-sided binomial exact test used to calculate the p-values and the generally low error rates at microsatellite loci. In a one-sided binomial exact test, the goodness-of-fit p-value will be 1 any time the observed error coverage is less than the expected error coverage. Since the expected error rate for most loci is on the order of 0.001 and the expected coverage for most loci is <50, the expected error coverage for most people at most loci will be considerably less than 1. Therefore, most noise goodness-of-fit p-values will be 1. For these same loci, p-values for any individuals with some error coverage will be much closer to 0 than they are to 1. For instance, if an individual has one putative error read at a locus with 10X coverage and an error rate of 0.001, the goodness-of-fit p-value will be <0.01. If coverage in the SSC dataset was higher, or microsatellite loci were noisier, the noise goodness-of-fit p-values would likely be more uniformly distributed.

The median noise goodness-of-fit p-value for each individual in the SSC is 1, which indicates that no individuals are consistently noisier than expected. The median noise goodness-of-fit p-value for nearly 95% of all loci is 1, and all but 16 loci have median noise goodness-of-fit p-values above 0.01. This behavior seems to indicate most loci are generally well described by their respective locus-wide error rate estimates. The few loci that do not appear to be well modeled by the estimated error rates may have multiple copies in non-reference genomes or have frequent somatic mutations.

In essence, if the alternative error rate were only moderately higher than the estimated error rate, the noise goodness-of-fit p-value is underpowered to reject the null hypothesis that noise coverage in an SSC individual at a locus is drawn from a binomial distribution with the estimated locus-wide error rate. However, the noise goodness-of-fit p-value should still be capable of distinguishing genotypes with much higher apparent error rates than the locus-wide error rate.

6.4.4.4 Null calls and marginal null probability

The median single null call frequency among all well-covered loci is <0.004, which is roughly 13 genotypes with a single null allele called per locus (Figure 6.32A, page 261). Single null call frequencies for 90% of all loci are less than 0.1. The median double null call frequency among all well-covered loci is <0.0006, which is roughly two people with a double null call per locus. 88% of loci have a double null call frequency below 0.1, and 44% of all loci do not have any double null genotypes (Figure 6.32B, page 261). The low frequency of null genotype calls indicates that the uSeq genotyper detects most microsatellite alleles in the SSC dataset. Most genotype calls containing null alleles are likely to be due to stochastically lower than expected coverage. This is reinforced when evaluating coverage of null genotype calls at loci with low null frequencies in GATK-aligned reads—these spot-checked SSC null genotypes are not due to SNVs, indels, or slippage events that would make an allele undetectable. However, loci with high null frequencies could still indicate a SNV or non-microsatellite indel that affects a microsatellite locus.

Even when a null genotype is not the most likely genotype for an SSC individual, the marginal null probability for the locus can be an indication of the overall likelihood that the locus genotype contains a null allele (Figure 6.33, page 262). The median marginal null probability is 0.003, and nearly 80% of genotypes have null probabilities less than 0.1. The average median marginal null probability for SSC individuals is 0.004, indicating that most individual do not have consistently high null probabilities. Median marginal null probabilities for each locus are less consistent. The average median marginal null probability for all loci is 0.05, while the remaining 22% of loci have higher marginal null probabilities. Higher locus-wide marginal null probabilities are primarily due to low observed coverage at a locus, especially compared to their respective expected coverage estimators. High marginal null probabilities suggest that coverage at a locus is insufficient to strongly support a bi-allelic genotype model. While this may not be a large impediment to accurate microsatellite genotyping in general, it bears special consideration when calling *de novo* microsatellite mutations.

6.4.5 Polymorphism in genotyped loci

At 68% of well-covered loci studied, all SSC individuals without a null genotype call are homozygous for the reference. The remaining 32% of well-covered loci have highly variable non-null heterozygous genotype call frequency (Figure 6.34, page 263). Nearly 75% of heterozygous loci have ≤10 individuals with a heterozygous genotype. This would seem to indicate that most well-covered loci evaluated as part of this SSC study are generally stable. Low levels of heterozygous genotype calls may be due to low mutation rates at the shortest microsatellite loci.

Many of the loci that were genotyped in SSC individuals as part of this study have multiple detected alleles (Figure 6.35, page 264). Nearly 34% of all loci have at least two genotyped alleles, regardless of null genotype status. 27% of loci have two detected alleles within the SSC population, 4% have three alleles, and the remaining 2% have more than three alleles detected at a locus. This indicates that although many loci may have few heterozygous genotypes, many loci within the SSC dataset exhibit some degree of polymorphism. The combination of infrequent heterozygous genotype calls and high levels of polymorphism at microsatellite loci suggests that although these loci are generally stable, their mutation rates are sufficient to introduce rare novel alleles. As microsatellite genotyping studies expand to larger populations, we would expect to see more microsatellite loci with rare polymorphism and infrequent heterozygous genotype calls. Non-reference microsatellite alleles shared by SSC individuals from different families may not necessarily derive from the same initial mutation at a specific locus. Polymorphism in SSC individuals from different families could result from multiple independent founding mutations.

6.5 Comparing uSeq to other microsatellite genotypers

To determine the performance of uSeq as compared to RepeatSeq and lobSTR, SSC family auSSC14395 was processed by all pipelines. lobSTR version 2.0.3 was used for the analyses described here, along with its reference set. Paired-end reads were provided as BAM input to lobSTR, and reads were trimmed based on a quality score of 20. lobSTR was allowed to process reads in parallel using two threads, which is the maximum amount of processors used by the uSeq pipeline. All genotypes called used the genotyping model provided in version 2.0.3 of the lobSTR resource bundle. RepeatSeq version 0.8.2 was used for the analyses described here, along with the latest annotation file provided by the authors as of February 25, 2014. All confident sites were emitted using the -emitconfidentsites flag and reads flagged as repeats by BWA are excluded. A minimum mapping quality score of 30 was required, which is the same threshold used by uSeq. Reads were aligned using BWA, then realigned and recalibrated using GATK before submitted to RepeatSeq. The GATK resource bundle file (GATK_ResourceBundle_5777_b37_phiX174_chrAll.fa) was provided as RepeatSeq's required FASTA input.

6.5.1 Reference microsatellite sets

The uSeq reference microsatellite set is nearly 3.5 times larger than the reference sets used by lobSTR and RepeatSeq. There are 1,638,523 lobSTR reference microsatellite loci, while the uSeq reference microsatellite set contains 5,784,968 loci. All of the loci in the RepeatSeq reference microsatellite set are contained in the lobSTR reference microsatellite set, so comparisons will only be made between lobSTR and uSeq. The uSeq and lobSTR reference sets overlap at 1,576,477 lobSTR-defined reference loci (Figure 6.36, page 265). Only 62,198 reference microsatellite loci are unique to lobSTR, which is <4% of all lobSTR reference loci. All of these loci are interrupted microsatellite tracts, as determined by the TRF scores of microsatellite loci unique to lobSTR. In contrast, 3,679,234 microsatellite loci are unique to the uSeq reference set, which is a dramatic increase in the number of detectable loci.

Since lobSTR and RepeatSeq both detect interrupted microsatellite loci, it is possible that a single lobSTR reference microsatellite locus contains several uSeq microsatellite loci. 81% of lobSTR reference microsatellite loci overlap a single uSeq reference microsatellite locus (Figure 6.37, page 266). An additional 17% of lobSTR reference microsatellites contain two or three uSeq reference microsatellite loci, while the remaining 2% are composed of more than three uSeq reference loci. This indicates that although lobSTR—and by extension, TRF—detects interrupted microsatellite loci, the vast majority of these loci contain a core of at least one uninterrupted microsatellite repeat that is detectable by uSeq. When evaluating the TRF scores, 62% of all lobSTR reference microsatellite loci that overlap with a uSeq reference microsatellite have the maximum possible TRF score. The maximum possible TRF score for a locus indicates that the locus is uninterrupted and would be identical to the corresponding uSeq reference microsatellites.

Novel uSeq reference microsatellite loci are distributed throughout all genomic contexts (Figure 6.38, page 267). In introns and intergenic regions, 3,598,602 microsatellite loci are uniquely detected by uSeq. In UTRs and miRNAs, uSeq uniquely detects 51,320 microsatellite loci. Most importantly, uSeq detects 37,988 microsatellite loci in exons in addition to the 3,887 that are also detected by lobSTR and RepeatSeq—ten times more loci than other pipelines. In contrast, lobSTR and RepeatSeq have just 1,086 unique microsatellite loci in exons, UTRs, and miRNAs combined. Many microsatellite loci that can significantly affect protein function, particularly in exons, are only detectable by uSeq.

Nearly 93% of all microsatellite loci unique to uSeq have motif lengths of 1, 2, or 3 bp (Figure 6.39, page 268). Microsatellite loci with short tract lengths may be inconsistently detected by lobSTR or TRF. This is likely to be the case as 99% of reference microsatellite loci unique to uSeq have tract lengths less than 18 bp (Figure 6.40, page 269). While lobSTR appears to have some reference microsatellite loci with tract lengths as low as 8 bp, many short microsatellites observed by uSeq are not in its reference set. This suggests that uSeq would be the more reliable genotyping pipeline for short microsatellite loci, since lobSTR detects them inconsistently. Interestingly, a fraction of uSeq-specific reference microsatellite loci are up to 263 bp long. It is not clear why these microsatellites would not be in the lobSTR reference microsatellite set.

Loci unique to the uSeq pipeline exhibit significant polymorphism (Figure 6.41, page 270). Almost 60% of reference microsatellite loci shared by all pipelines are polymorphic, as determined by uSeq genotype calls within the SSC population. An additional 30% of reference microsatellite loci that are unique to uSeq are also polymorphic. This indicates that many of the microsatellite loci detected by uSeq are highly informative, and neither lobSTR nor RepeatSeq will call genotypes at these polymorphic microsatellite loci. Many of these loci have short reference tract lengths, since most of the additional microsatellites detected by uSeq are short. The level of polymorphism for short microsatellite loci unique to uSeq indicates that they should be included in microsatellite genotyping studies, as they are still prone to slippage mutations and polymorphism. lobSTR and RepeatSeq are not able to call genotypes at millions of potentially polymorphic microsatellite loci.

6.5.2 Microsatellite locus coverage

Pairwise coverage comparisons show that uSeq aligns microsatellite reads as well as RepeatSeq, and perhaps slightly better than lobSTR (Figure 6.42, page 271). Pairwise coverage was assessed for all shared loci in two pipelines for a trio family member. For all family members, lobSTR only reported even coverage, which was assumed to mean that lobSTR was unintentionally doubling coverage. This is borne out by spot-checking alignments in the BAM files produced by lobSTR—reported coverage is double the number of reads aligned to a microsatellite locus. With that assumption, all lobSTR coverage is halved. If lobSTR coverage is erroneously doubled, uSeq has roughly twice the coverage as lobSTR; if lobSTR reported coverage is accurate, then uSeq and lobSTR coverage is generally equivalent. The correlation between lobSTR and uSeq coverage is somewhat low, and uSeq appears to get consistently higher coverage than lobSTR for many microsatellite loci (Figure 6.42A, page 271). RepeatSeq also seems to get consistently higher coverage than lobSTR (Figure 6.42C, page 271). Higher coverage in uSeq is somewhat surprising, since lobSTR can detect interrupted microsatellites, which should allow it to pick up more reads that have sequencing errors that disrupt microsatellite tracts.

uSeq and RepeatSeq have very similar coverage for all common loci, and coverage is highly correlated between the two pipelines (Figure 6.42B, page 271). Although coverage generally appears to be nearly equivalent between the two pipelines, RepeatSeq usually has somewhat higher coverage than uSeq. This is most likely due to the GATK-based alignment strategy employed by RepeatSeq, which should allow it to detect and align interrupted microsatellite loci, unlike uSeq. In addition, RepeatSeq does not use the same rules as uSeq for filtering PCR replicate sets and overlapping reads. Some loci have higher coverage in uSeq than in RepeatSeq. One possible explanation for this observation is that a GATK-based pipeline does not effectively map microsatellite indels at these particular loci. This is consistent with the comparison of coverage at microsatellite loci in uSeq and a standard GATK-based pipeline, as discussed in section 6.2.7.

uSeq aligns nearly as well as RepeatSeq, and at least as well as lobSTR. This demonstrates that the uSeq alignment algorithm has accuracy comparable to these other microsatellite genotyping pipelines. While it is unclear why lobSTR's alignment algorithm seems to consistently underperform compared to uSeq and RepeatSeq, it is certainly an important consideration when choosing a microsatellite genotyper.

6.5.3 Genotyping comparison

Since uSeq only detects perfect microsatellites, it will often report different allele lengths than lobSTR or RepeatSeq. Rather than comparing genotypes or alleles directly, we chose to compare each pipeline's Mendel violation rate as a means of assessing genotype accuracy. We identified the loci shared by all three pipelines in either the proband or sibling trio, and calculated the Mendel violation frequency for different trio genotype confidence thresholds (Figure 6.43, page 273). Although all pipelines demonstrate increasing accuracy as trio genotype confidence increases, both lobSTR and RepeatSeq maintain very high *de novo* mutation rates relative to the highest per-locus microsatellite mutation rate estimates described in the literature (Sajantila et al. 1999). Since most previous microsatellite mutation rate estimates used highly polymorphic marker loci, the range of known microsatellite mutation rates is likely to be an overestimate relative to the mutation rate for non-marker loci. In contrast, uSeq calls Mendel violations with an almost 10-fold lower frequency.

Many of the Mendel violations called confidently by RepeatSeq or lobSTR are not called by uSeq. When evaluating Mendel violations called with a minimum trio genotype confidence threshold of at least 0.8, ~50% of lobSTR and RepeatSeq Mendel violations are not called by uSeq because at least one trio member has a null allele. However, the mean and median coverage is at least as high in uSeq as compared to lobSTR or RepeatSeq for these loci. This indicates that uSeq null allele calls are not driven by lower coverage of these potentially Mendel-violating loci in the uSeq pipeline. The remaining 50% of high trio genotype confidence Mendel violation calls by lobSTR or RepeatSeq have lower trio genotype confidence scores in uSeq. Therefore, it seems that the increased accuracy of uSeq—as measured by the Mendel violation rate—is due to the additional parameters modeled by its genotyper. By accounting for allele bias and expected coverage, and by having per-locus error rates, the Mendel violation false positive rate is greatly reduced.

Most, if not all, of the Mendel violations called by any pipeline within this family are false positives. Trio genotype confidence is only one means of identifying loci that can be screened for Mendel violations. Since all three pipelines report genotype confidence scores, we chose to use this metric to evaluate genotype accuracy. However, true *de novo* mutations require the thorough analyses described in section 5.4.

6.5.4 Run-time

Of all three pipelines, lobSTR takes the longest time to process each individual. Input sequence was provided to each pipeline as a BAM file, which lobSTR requires to be lexicographically sorted. This sort took 20 – 40 minutes per BAM file. The lobSTR alignment algorithm took 16 – 21 hours to complete using two threads. Once alignment has completed, post-processing and genotype calling takes an additional 5 – 7 minutes. Therefore, lobSTR used approximately 32 – 42 hours of computing time to call microsatellites in each individual.

RepeatSeq takes already aligned files as input, and reports all confident genotypes with additional "call" files in 14 – 17 minutes. However, in this study, RepeatSeq took BWA-aligned and GATK-realigned and recalibrated reads. The time taken to align the raw reads for each individual for RepeatSeq was 19 – 35 minutes, and post-processing took 45 – 145 minutes. Realignment was performed on a BAM file with all reads from the family, and took a total of ~27 hours. Recalibration took an additional 14 hours. An additional 2.5 hours were needed to merge realigned BAM files and index the recalibrated BAM file. Although RepeatSeq takes very little pipeline-specific time, preparing data for input took >45 hours for this family, or ~11 hours per family member.

For the same individuals in the uSeq pipeline, microsatellite detection took 10 – 30 minutes. Alignments took an additional 10 – 30 minutes using BWA with two threads to enable parallel processing. Post-processing, including BAM file production and reindexing, took 20 – 30 minutes. Merging and marking duplicates in BAM files took ~25 minutes, and profiling took another 17 – 25 minutes. Therefore, the individual component to the uSeq pipeline took between 1.5 hours and 2.3 hours to align and process each individual, not including time spent waiting for cluster access. The population component of the uSeq pipeline took ~7 days to merge over 5 million microsatellite loci in 3,500 individuals. Identifying well-covered loci, deriving an expected coverage model, genotyping, and finding *de novo* mutations took ~2.5 days, although genotyping was performed on 13 threads. The total amortized processing time per SSC individual for the population component is ~16 minutes of computer time and 4 minutes of wall clock time.

If the total amount of time taken to process reads and call genotypes is considered, and time is amortized for each individual genotyped, uSeq is the fastest of the three pipelines. When only considering pipeline-specific runtime, RepeatSeq is the fastest pipeline. lobSTR's processing time is primarily devoted to their alignment algorithm, although this does not appear to dramatically improve its coverage or genotyping accuracy as compared to uSeq or RepeatSeq. This would not justify lobSTR's extended alignment times, particularly when evaluating microsatellites in large populations. Each individual in the SSC population requires about 1.5 – 2.5 hours of time—including amortized population component time—to detect, align, and process microsatellite reads; infer model parameters from the population; call genotypes; and identify *de novo* mutations.

6.5.5 Summary

uSeq can detect more microsatellite loci, align them as effectively, and genotype them more accurately than other microsatellite genotyping pipelines in exome sequencing data. The amortized run time for uSeq is very competitive, and is faster than lobSTR or RepeatSeq when considering both read alignment and genotyping time. It is probable that uSeq's EM-based genotyping model would maintain its advantage in whole genome sequencing studies—although allele capture bias may play less of a role, coverage estimators and locus-specific error rates will still be relevant. The superiority of the uSeq pipeline is particularly important when microsatellite genotyping is undertaken to identify potentially disease-causing alleles or *de novo* mutations at microsatellite loci.

6.6 De novo microsatellite mutations

6.6.1 Trio genotype quality distributions

Over 99.9% of the 231,382,755 Mendel obedience scores calculated for the SSC dataset are below 40 (Figure 6.44, page 274). The vast majority of trio genotypes are Mendel obedient in the uSeq pipeline, which suggests that the genotypes being called are generally accurate and reliable. As the Mendel obedience score threshold rises, the number of candidate de novo mutations drops precipitously—just over 56,000 trio genotypes have obedience scores above 40, while less than 17,000 have obedience scores above 60, a decrease of nearly 70%. When identifying potential de novo mutations, we require that every person within the family has a definitive genotype, and therefore we do not allow for null alleles. At a threshold of 40, ~50% of all potential Mendel violations have a null genotype in a family member.

In preliminary analyses where we limited our analysis to families with no null genotypes and various Mendel obedience thresholds, it became immediately apparent that most candidate *de novo* mutations were likely to be false positives. 62 SSC families had unusually high *de novo* mutation frequencies for SNVs and microsatellite loci—these families were excluded from any further analysis. We assume that the "*de novo*" calls in these families are likely due to sample mix-ups.

In an effort to eliminate more spurious *de novo* calls, we introduced several other trio quality measures. While the Mendel obedience score specifically addresses the Mendelian behavior of a trio at a locus, these additional measures addressed the trio genotype quality of the likely *de novo* mutation. These measures are the kinship score; trio noise and allelic goodness-of-fit; trio marginal null probability; and trio genotype confidence. In addition, we filtered *de novo* microsatellite mutation calls from loci that appear to be very noisy, reasoning that they would be loci that are most prone to false positives.

In our putative *de novo* mutation analysis, we set a liberal threshold on the Mendel obedience in an effort to assess its reliability for *de novo* mutation detection. We chose to set conservative thresholds for the other trio genotype quality measures used to filter de novo mutations so that we could focus on the behavior of the Mendel obedience score. These thresholds were set by observing the appropriate distributions and attempting to find thresholds that would maintain high genotype quality while minimizing false positives.

Our first analyses demonstrated that many putative *de novo* microsatellite mutations had low kinship scores. A low kinship score would suggest that these mutations required multiple mutational events in the parental germlines or during development (Figure 6.45, page 275). As means of reducing the number of de novo mutations that required increasingly improbable kinship relationships, we set a minimum kinship score of 0.8.

Most microsatellite loci with recurrent *de novo* mutations will be enriched for false positives and can be eliminated without losing too many true *de novos*. Microsatellite loci with higher locus error rates are more likely to have recurrent *de novo* mutations (Figure

6.46, page 276). This suggests that some spurious microsatellite *de novo* mutations are simply due to difficulty calling genotypes accurately at these unstable loci. To minimize false positives in our analysis of the behavior of the Mendel obedience score, we limit our study to *de novo* mutations called at loci with estimated error rates less than 0.17.

Thresholds on the remaining metrics were set to ensure that highly confident *de novo* mutations were called (Figure 6.47, page 277). Minimum trio genotype confidence was set to 0.99 to ensure that the likeliest trio genotype was unambiguous (Figure 6.47A, page 277). The maximum trio marginal null probability was set to 0.01 to ensure that a trio had little chance of a null genotype (Figure 6.47B, page 277). Allele fits and noise fits were set to 0.00001 and 0.001, respectively, to ensure that no trio genotype being considered had suspiciously high coverage or was too noisy (Figure 6.47C-D, page 277). These thresholds were set in an ad hoc manner as a means of isolating the behavior of the Mendel obedience score. These thresholds performed successfully in this analysis, but a more detailed evaluation of the efficacy with which each parameter or parameter combination can distinguish true *de novo* mutations has not been undertaken. Such an analysis could increase the sensitivity with which uSeq is able identify true *de novo* mutations.

By setting these thresholds, we are able to narrow down our list of putative *de novo* microsatellite mutations from ~56,000 to 139. These candidate de novo mutations were then manually evaluated, compared to previous *de novo* calls from the GATK-based lab pipeline, and compared to calls made by the local reassembly pipeline Scalpel (Narzisi et al. 2013). A set of recurrent *de novo* mutations at one microsatellite was excluded from further analysis based on local reassembly evidence that demonstrated that the locus had a common indel polymorphism that removed a portion of the microsatellite locus and its flanking sequence.

Based on earlier analyses, an additional four de novo mutations were submitted for validation. These additional mutations had been excluded in the latest analysis due to low kinship scores. Candidate *de novo* mutations were submitted for validation sequencing using the protocol described in section 5.5.2. After sequencing, each de novo mutation was analyzed and scored as valid (*de novo* confirmed), invalid (no evidence of *de novo*), failed (no primers or no DNA), or unclear (ambiguous alleles). Of the 144 candidate de novo mutations, 22 were valid, 62 were invalid, 8 were unclear, and 52 failed.

6.6.2 Determining a threshold Mendel obedience score for microsatellite *de novo* mutations

The Mendel obedience scores of valid and invalid microsatellite *de novo* mutations are distinct and can be used to discern true events with high specificity (Figure 6.48, page 278). The ROC AUC for the Mendel obedience score is >0.95, which means that 95% of the time, the Mendel obedience score for a true *de novo* mutation will be higher than for a spurious *de novo* mutation. Therefore, with the thresholds used in this study, the Mendel obedience score is a very robust means of identifying true *de novo* microsatellite mutations. The ROC curve can also be used to calculate specificity and sensitivity within the validation set at a particular Mendel obedience score threshold. If we were to limit our false positive rate to <10% (6/62), we are still able to recover >91% (20/22) of all true positives if we set a minimum Mendel obedience score of 63. This performance is comparable to large-scale microsatellite *de novo* mutation genotyping using capillary electrophoresis, which reported a false positive rate of 7.2% (Sun et al. 2012). Our thresholds may exclude some de novo mutations, so the sensitivity described here is the Mendel obedience score's ability to correctly identify true positives within the validation set, not the complete SSC dataset.

6.6.3 Microsatellite de novo mutations

6.6.3.1 De novo mutation overview

There are a total of 45 *de novo* mutations in the dataset above a Mendel obedience score threshold of 63—20 were valid, 6 were invalid, 4 were unclear, and 14 failed (Table 6.6, page 294). Other pipelines do not observe 36 of the *de novo* mutations identified by uSeq in this study, and 12 of the valid mutations are unique to this analysis. *De novo* mutations are equally likely to occur in probands or their siblings in the complete set of candidate mutations, and in each context individually (Table 6.7, page 295). Nearly 60% of de novo mutations are in introns, and an additional 30% are in exons. All exon mutations were either valid or failed to validate (Table 6.8, page 295). Invalid *de novo* mutations occurred exclusively in introns and intergenic regions. 38 of the 45 *de novo* mutations are slippage mutations, while the rest appear to be point mutations inside or adjacent to microsatellite loci. Slippage insertions and deletions are equally common in probands and siblings (Table 6.9, page 295). Unique mutations also do not appear to be more common in children with autism or their siblings (Table 6.10, page 295). Omission and commission *de novo* mutations occur with equal frequency in children with autism and their siblings, and only three of the 45 *de novo* mutations are omissions (Table 6.11, page 296). It is clear that uSeq is capable of reliably identifying novel *de novo* mutations that are undetectable with other pipelines. However, there does not appear to be a single mutation characteristic that suggests a *de novo* mutation bias in children with autism or their siblings.

6.6.3.2 Unique slippage insertions occur exclusively in children with autism

De novo mutations are found in varying contexts and some of them have very clear effects. Recently, an analysis of *de novo* SNVs in exome sequencing data reported a link between FMRP-associated genes and autism incidence (Iossifov et al. 2012). Although we do not observe the same significant association in our current study, we do observe two frameshift mutations in FMRP-associated genes (Darnell et al. 2011). A mutation in a microsatellite with a 1 bp motif produces a frameshift mutation in a child with autism in the FMRP-associated gene KIF21A (Figure 6.49, page 279). This mutation has not been observed in any other *de novo* mutation detection pipeline applied to the SSC. This allele is also unique within SSC parents and is not found in dbSNP v138.

An additional mutation in a microsatellite with a 1 bp motif produces a frameshift mutation in a child with autism in CHD8 (Figure 6.50, page 281). CHD8 has been linked to autism incidence by several studies, both due to its association with FMRP and due to recurrent mutations in CHD8 identified in several studies (Darnell et al. 2011; Neale et al. 2012; O'Roak et al. 2012a; O'Roak et al. 2012b; Sanders et al. 2012). This same mutation was observed in a study of *de novo* mutations at targeted autism genes (O'Roak et al. 2012a). This mutation is also unique within the SSC parents and is not found in dbSNP v138.

A third frameshift mutation in a child with autism is found in B4GALNT1 (Figure 6.51, page 282). The mutant allele is unique within the SSC dataset and does not appear in dbSNP v138. It has been detected by other *de novo* mutation detection pipelines applied to the SSC dataset. A homozygous insertion mutation identical to the heterozygous mutation observed in this study was reported in an individual suffering from the autosomal recessive disorder spastic paraplegia 26. B4GALNT1 is involved in the synthesis of complex gangliosides, which are components of the synaptic plasma membrane (Boukhris et al. 2013). Some studies have suggested that the genes affecting synaptic plasticity may pay a role in autism incidence (Darnell et al. 2011; Iossifov et al. 2012).

A frameshift mutation also occurs in the HMMR gene, although it failed to validate (Figure 6.52, page 283). Other *de novo* detection pipelines do not find the mutation, no SSC parents have the mutations, and it is in dbSNP v138. There are no known associations of HMMR to autism. A final potential frameshift mutation occurs in an ATCT microsatellite of the GDPD4 gene (Figure 6.53, page 284). The microsatellite spans the junction between the last exon of the gene and its 3' UTR, so its effect is unclear. This mutation was valid and other *de novo* mutation detection pipelines have detected it. It is unique in the SSC dataset and is not found in dbSNP v138.

There are two in-frame slippage insertions mutations in children with autism in HEXIM1 and CD3EAP. Both mutations were valid, and other pipelines detected them. The HEXIM1 mutation adds an additional glutamate and leucine to an alternating poly-EL tract. The CD3EAP mutation adds one residue to a lysine tract. All four mutations are unique in the SSC and none are found in dbSNP v138. There are two non-slippage exon *de novo* mutations in children with autism detected at microsatellite loci in the CCDC27 and INSR genes. Other pipelines detected both mutations. While the CCDC27 mutation was valid, the INSR mutation failed to validate. Figures for these mutations—and all *de novo* microsatellite mutations in Table 6.6—can be found in Appendix 1.

Five *de novo* microsatellite mutations in exons are detected in siblings of children with autism, and all mutations are in-frame. A *de novo* microsatellite mutation in a sibling in CPSF1 is not detected by other pipelines, is unique in the SSC dataset, and is not found in dbSNP v138 (Figure 6.54, page 285). This mutation removes one residue from an aspartate tract. A second *de novo* mutation in a sibling removes a residue from a glutamate tract in VSIG10, which is seen in other SSC individuals, but is not seen in dbSNP v138. Other pipelines do not detect this mutation either. A third *de novo* mutation removes a residue from an aspartate tract in RNF6. This mutation is unique in the SSC dataset and is not seen in dbSNP v138, but other pipelines detect it. A *de novo* mutation in a sibling in a KDM8 exon removes a residue from a glutamate tract. Although this mutation is not seen within the SSC, it is commonly found in dbSNP v138, and other pipelines detect it. A final *de novo* mutation in a sibling is detected in MED15, which adds a residue to a glutamine tract. This mutation is common in the SSC dataset, although it is not seen dbSNP v138. Other pipelines do not detect it. This mutation is one of the four mutations that did not meet the criteria outlined for microsatellite *de novo* mutations. All of these mutations were valid in targeted sequencing.

All frameshift mutations detected in this study are in children with autism, and none are in their siblings. This bias is not statistically significant (p = 0.13), although it is suggestive of a trend towards microsatellite frameshift mutations in children with autism. An expanded analysis of the SSC dataset might reveal a statistically significant association. Considering all slippage mutations, there is no significant difference in the *de novo* mutation frequency in exons (7 in probands, 5 siblings). Although unique slippage mutations in exons are more common in children with autism than in their siblings, this bias is not significant (7 in probands, 2 in siblings, p = 0.18).

Unique microsatellite insertions are more likely to occur in children with autism—six validated slippage mutations found in children with autism are unique insertions, while there are none in their siblings (Table 6.12, page 296; p = 0.03). If we include the slippage insertion observed in HMMR that failed to validate, this bias become more significant (p = 0.016). Moreover, if we consider GDPD4 to be a *de novo* mutation in an exon, then the bias is for unique exonic slippage insertions in children with autism as compared to their siblings. No unique slippage insertions are observed in siblings in any context. There is one additional unique slippage insertion mutation in an intron of HYDIN. Although this mutation is above the Mendel obedience threshold, it is invalid. If we were to include this mutation, the bias towards unique insertions in children with autism would have a p-value <0.01.

6.6.3.3 A Robertsonian translocation on chromosome 22

The only two valid omission *de novo* mutations were observed in the same child on chromosome 22. The first omission *de novo* mutation occurs in an AGGGG microsatellite in an intron of SF3A1, which is common in the SSC cohort and dbSNP v138 (Figure 6.55, page 286). A second omission *de novo* occurs ~7 Mb downstream, in a GT microsatellite in an intron of IL2RB (Figure 6.56, page 287). This mutation is common in the SSC cohort, but is not observed in dbSNP v138. Other pipelines fail to detect either mutation. A third omission *de novo* was later found 3 Mb from the IL2RB mutation, in a GCA microsatellite in an exon of TNRC6B (Figure 6.57, page 288). It was not submitted for validation, but it is consistent with the other omission *de novo* mutations detected in the same child on chromosome 22.

At all loci, the parents are homozygous for two different alleles. Mendelian inheritance dictates that each child should inherit one allele from each parent, and would therefore be heterozygous. At all three loci, the unaffected sibling is heterozygous for the parental alleles. However, the child with autism is always homozygous for the maternal allele, indicating that the child did not inherit a paternal allele. Since two of the three mutations validated, we sought to determine if events occurred by chance, or if they were part of a larger chromosomal phenomenon. In an analysis of heterozygous SNVs from either parent on chromosome 22, the child with autism always had the maternal genotype, regardless of the paternal genotype. This suggests that the child with autism inherited both copies of maternal chromosome 22, and perhaps no paternal copies. This behavior is consistent with a homologous maternal Robertsonian translocation or maternal uniparental heterodisomy.

One possible mechanism leading to a Robertsonian translocation is the fusion of two acrocentric chromosomes at the centromere, followed by the loss of the short chromosome arms and the retention of both long arms. Since the short arm of chromosome 22 does not contain any unique genes, this translocation is not expected to have a phenotypic effect. However, this demonstrates that genome-wide *de novo* microsatellite genotyping could also be an effective means of identifying rare chromosomal events.

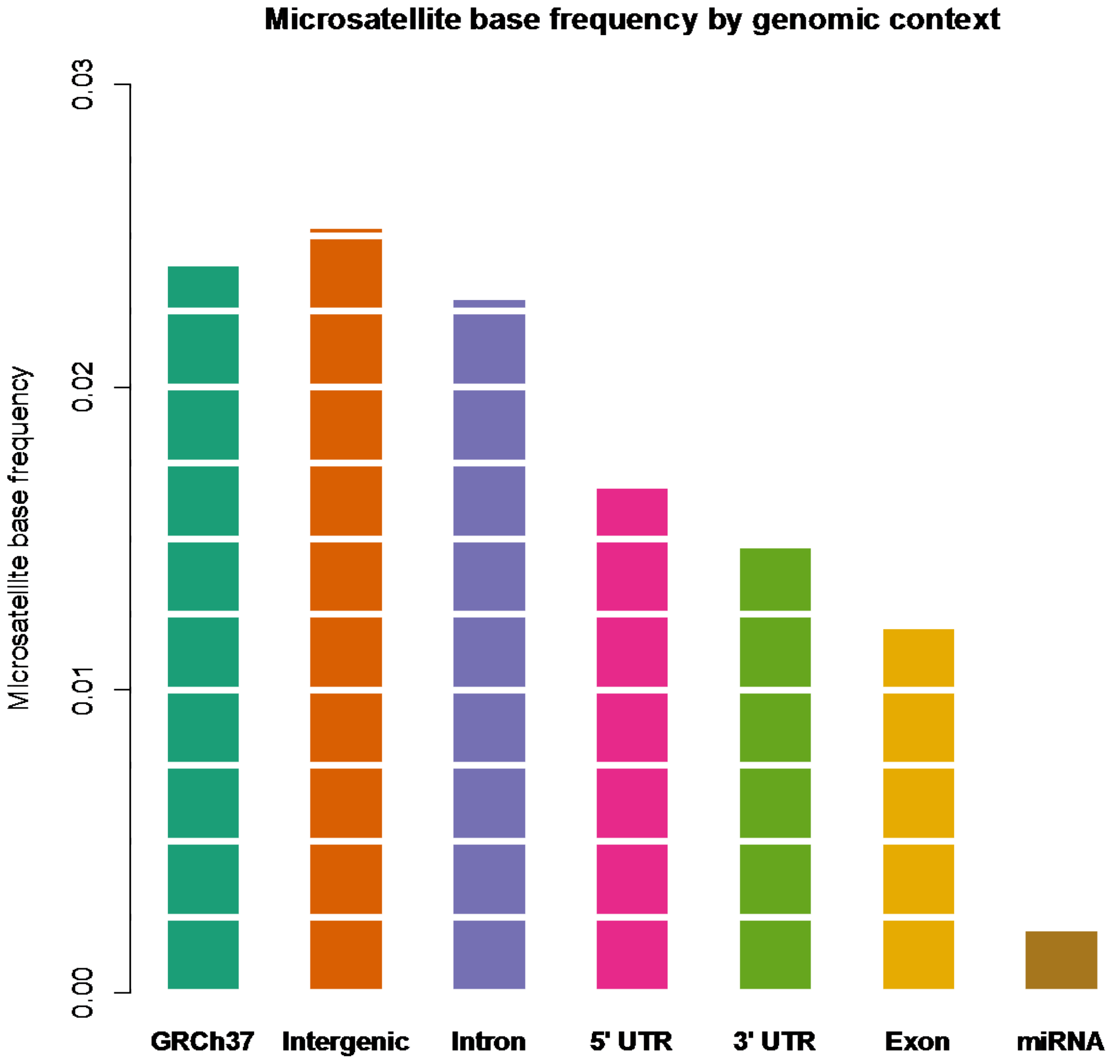


Figure 6.1: Microsatellite base frequency by context. Base frequency is the total number of microsatellite bases within a context divided by the total number of bases in that context. ~2.3% of bases in GRCh37 are in a microsatellite tract. Exons are depleted for microsatellite tracts compared to the genome as a whole—just ~1.2% of bases in exons are in a microsatellite tract.

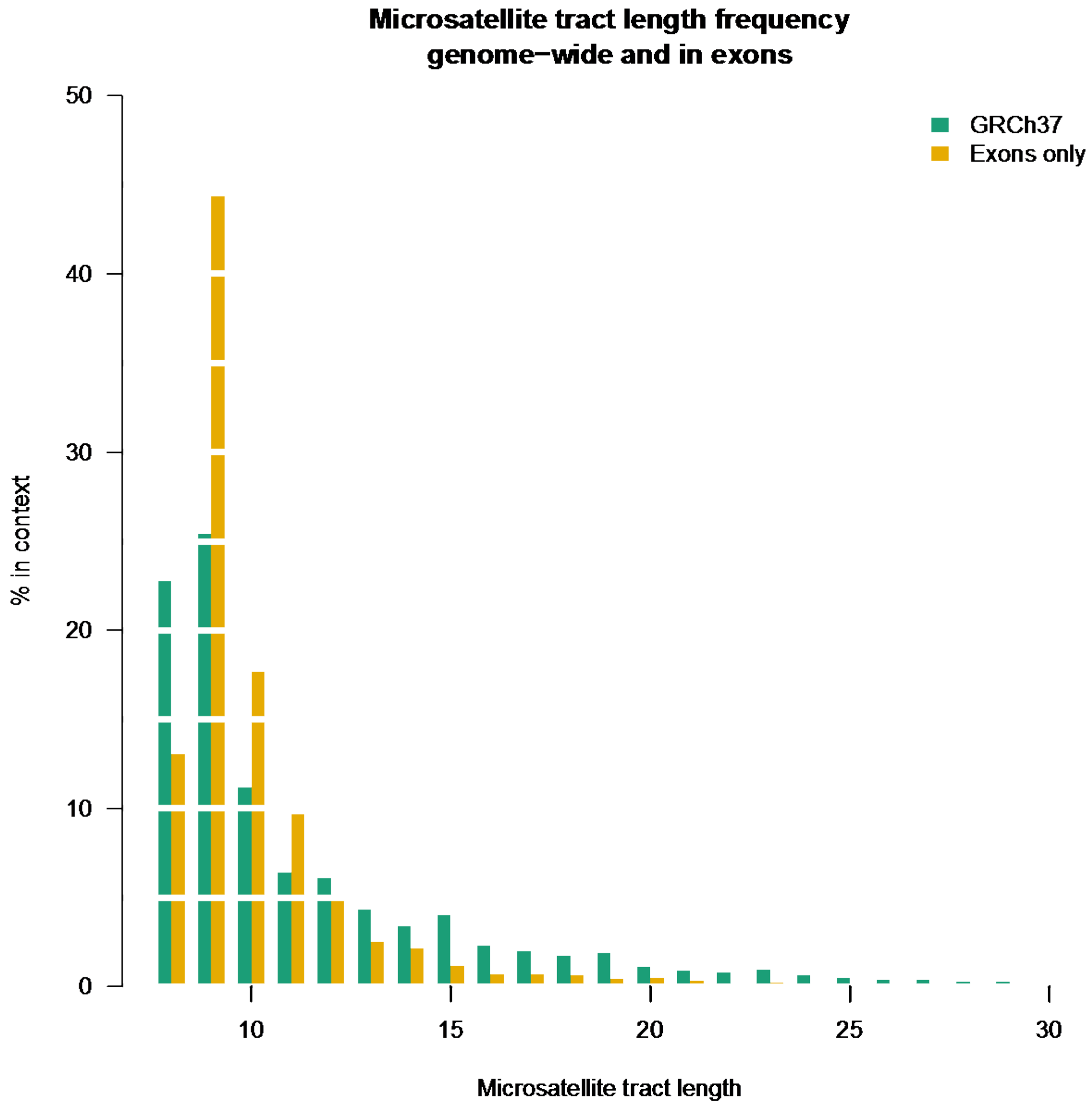


Figure 6.2: Microsatellite tract length distribution, genome-wide and in exons. The percentage in context refers to the percentage of all microsatellites in GRCh37 or in exons only with a particular tract length. Microsatellites in exons tend to be shorter than microsatellites in GRCh37 as a whole.

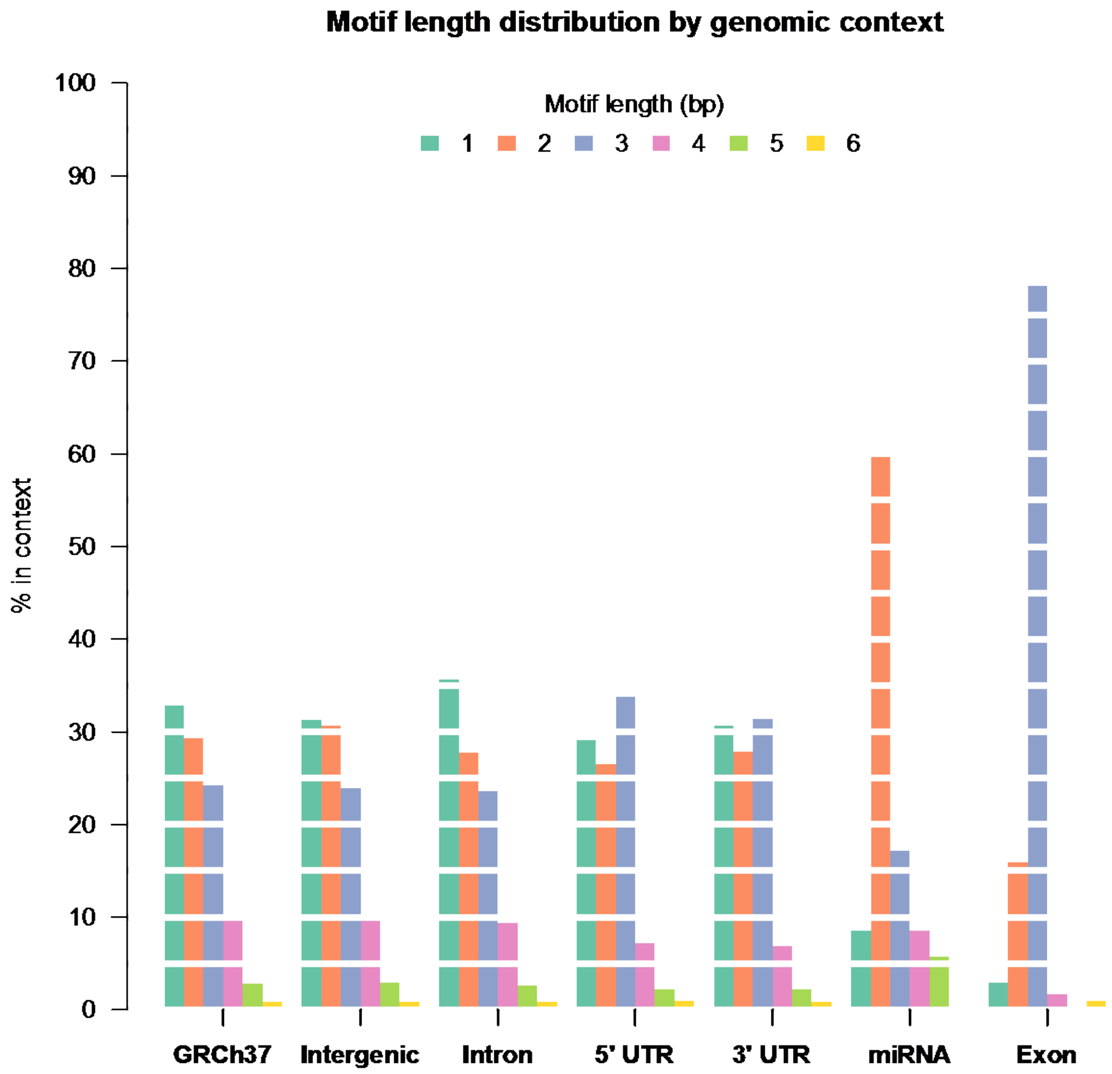


Figure 6.3: Microsatellite motif length distribution by genomic context. Percentages for each context sum to 100%. The frequencies of microsatellites with 1, 2, and 3 bp motifs are similar for GRCh37 as a whole, and in introns and intergenic regions. Both 5' and 3' UTRs are enriched for microsatellites with 3 bp motifs relative to GRCh37. Microsatellites with 2 bp motifs are enriched in miRNA, although the total number of microsatellites in miRNA is low. Microsatellites with 3 bp motifs are heavily favored in exons, which ensures that slippage mutations at most exon microsatellite loci do not disrupt the proper reading frame.

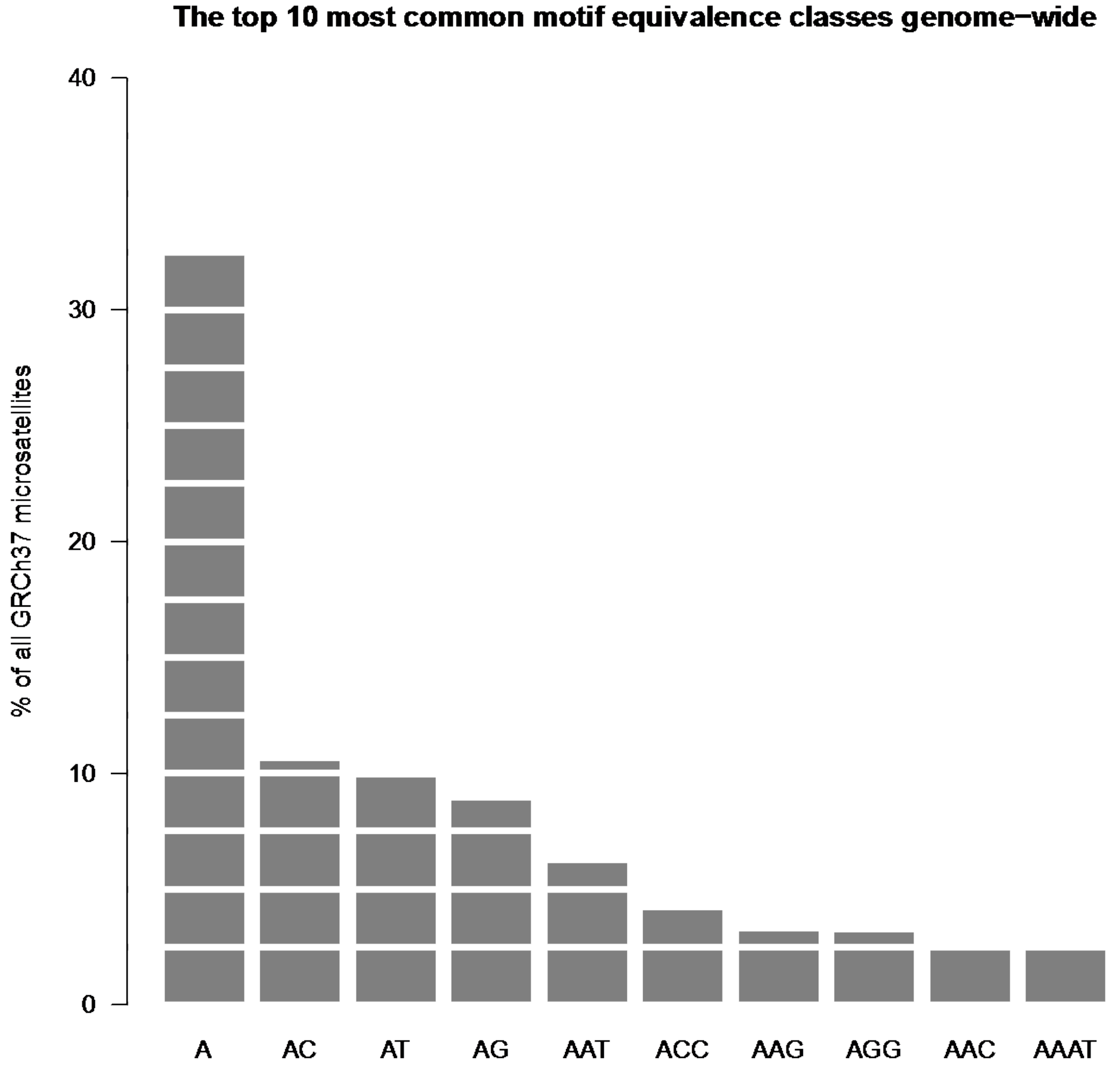


Figure 6.4: The most common motif equivalence classes throughout GRCh37. Microsatellites with A class motifs are the most common, while motifs composed exclusively of C and G nucleotides are more rare.

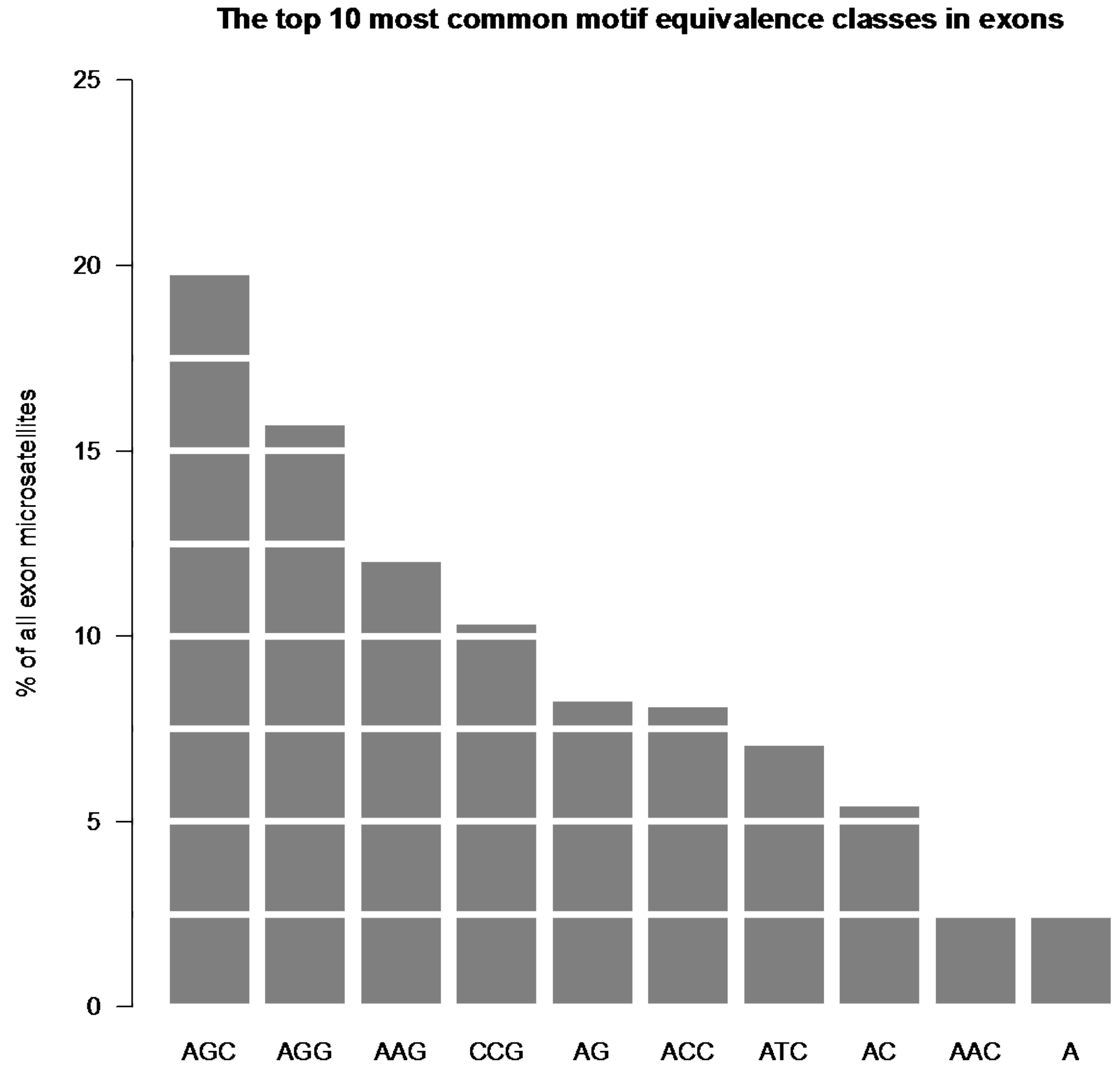


Figure 6.5: The most common equivalence classes in the exons of GRCh37. Microsatellite motifs composed exclusively of C and G nucleotides are more common in exons than they are in the rest of the genome. Although most exon microsatellites have 3 bp motifs, 3 of the most common microsatellite equivalence classes have motif lengths of 1 or 2 bp.

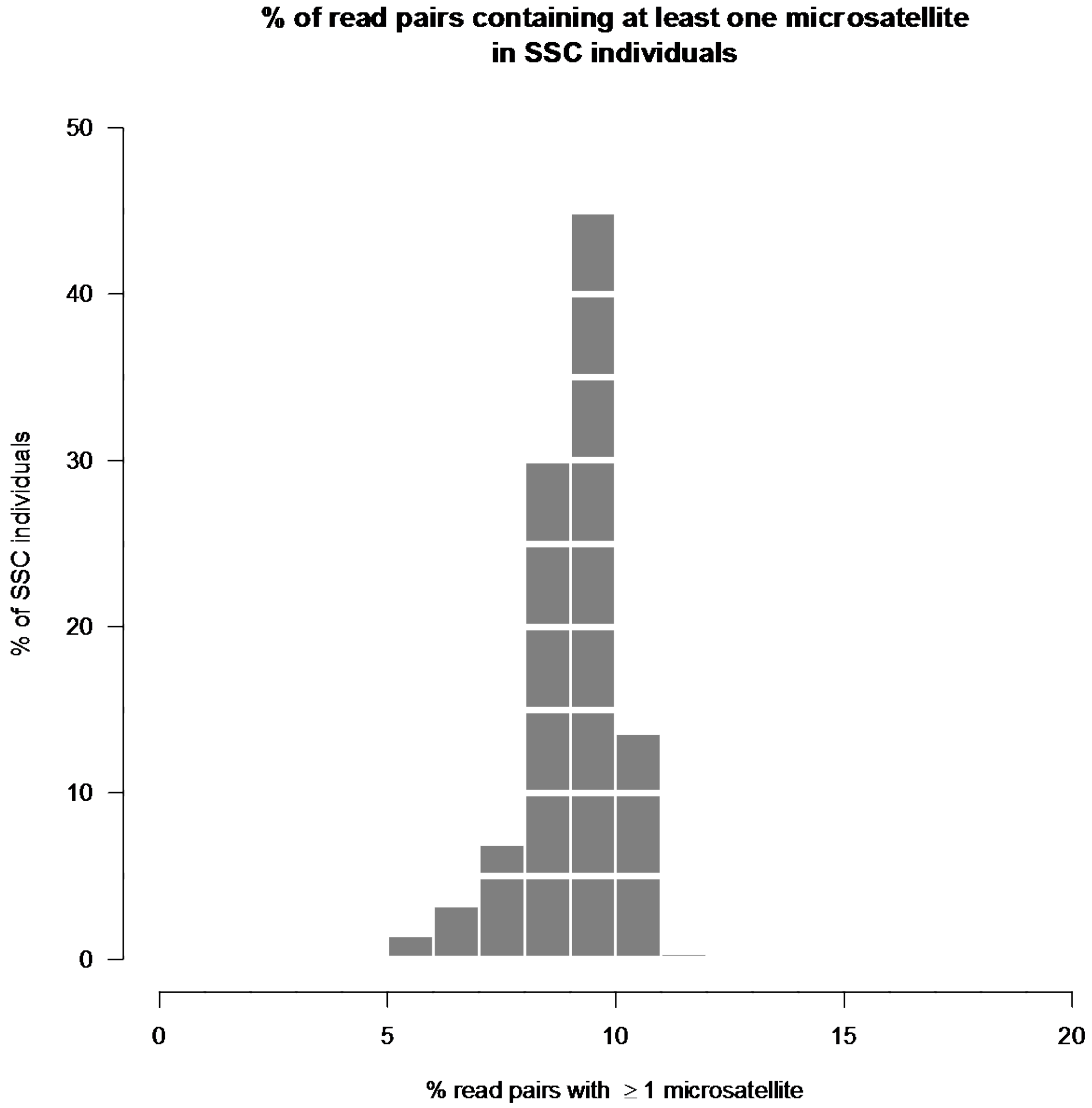


Figure 6.6: Percentage of read pairs containing at least one microsatellite in each SSC individual. Approximately 9% of all read pairs have at least one microsatellite in the average SSC individual.

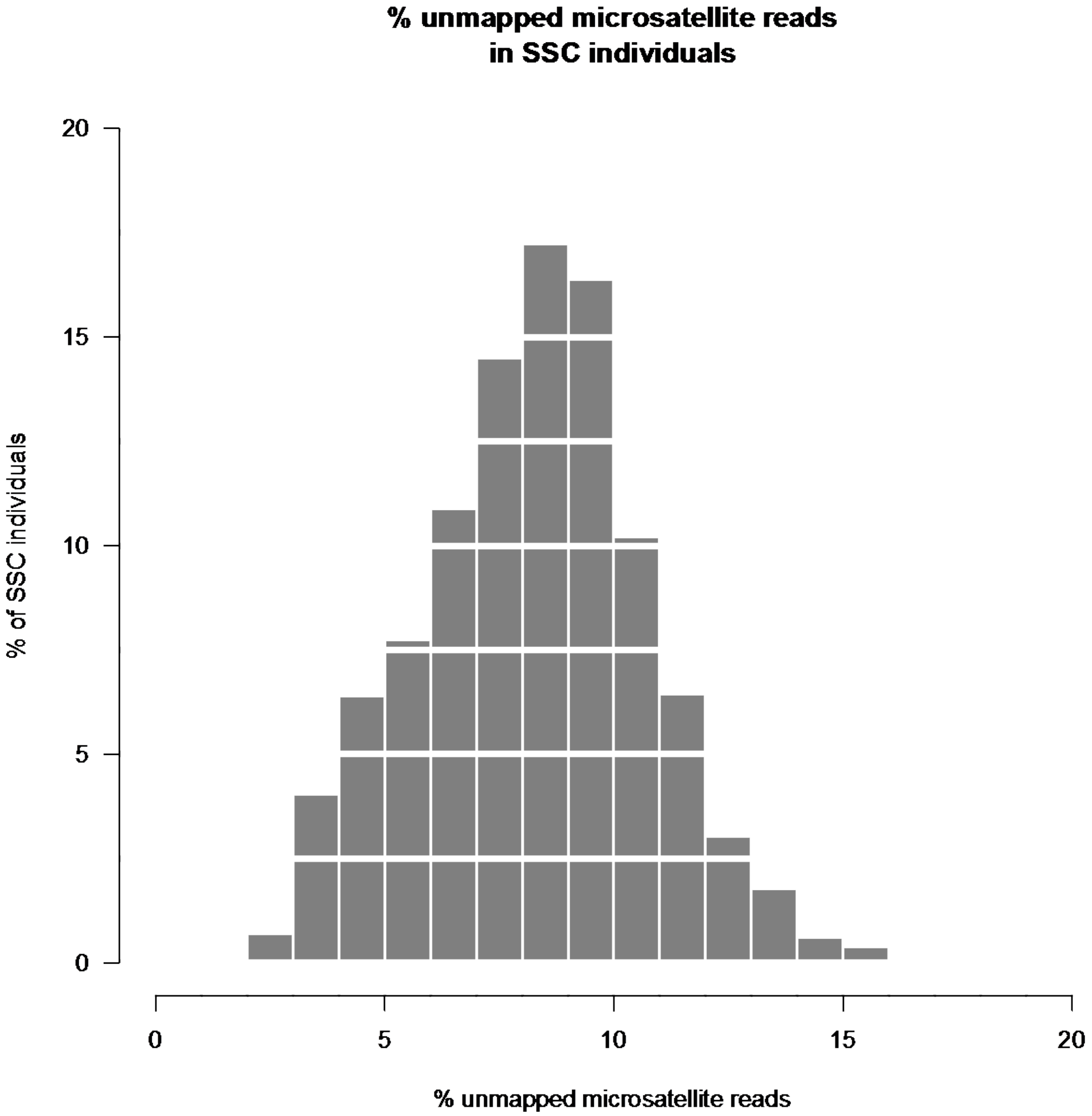


Figure 6.7: Percentage of unmapped microsatellite reads in each SSC individual. Approximately 8% of reads with microsatellites fail to map in the average SSC individual.

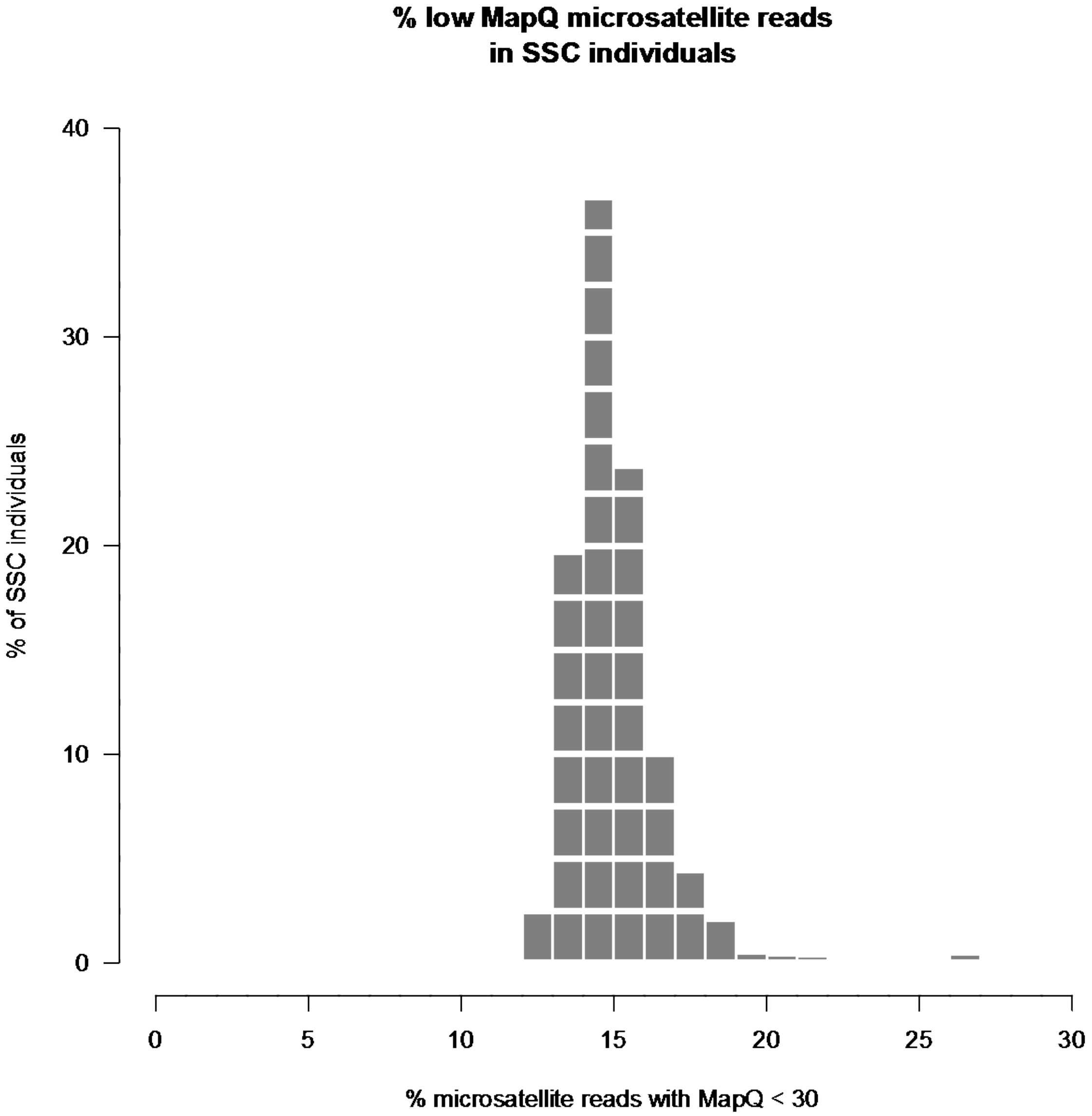


Figure 6.8: Percentage of microsatellite reads with low mapping quality scores for each SSC individual. Any mapping quality score <30 is considered low. 15% of microsatellites have an alignment with a mapping quality score <30 in the average SSC individual.

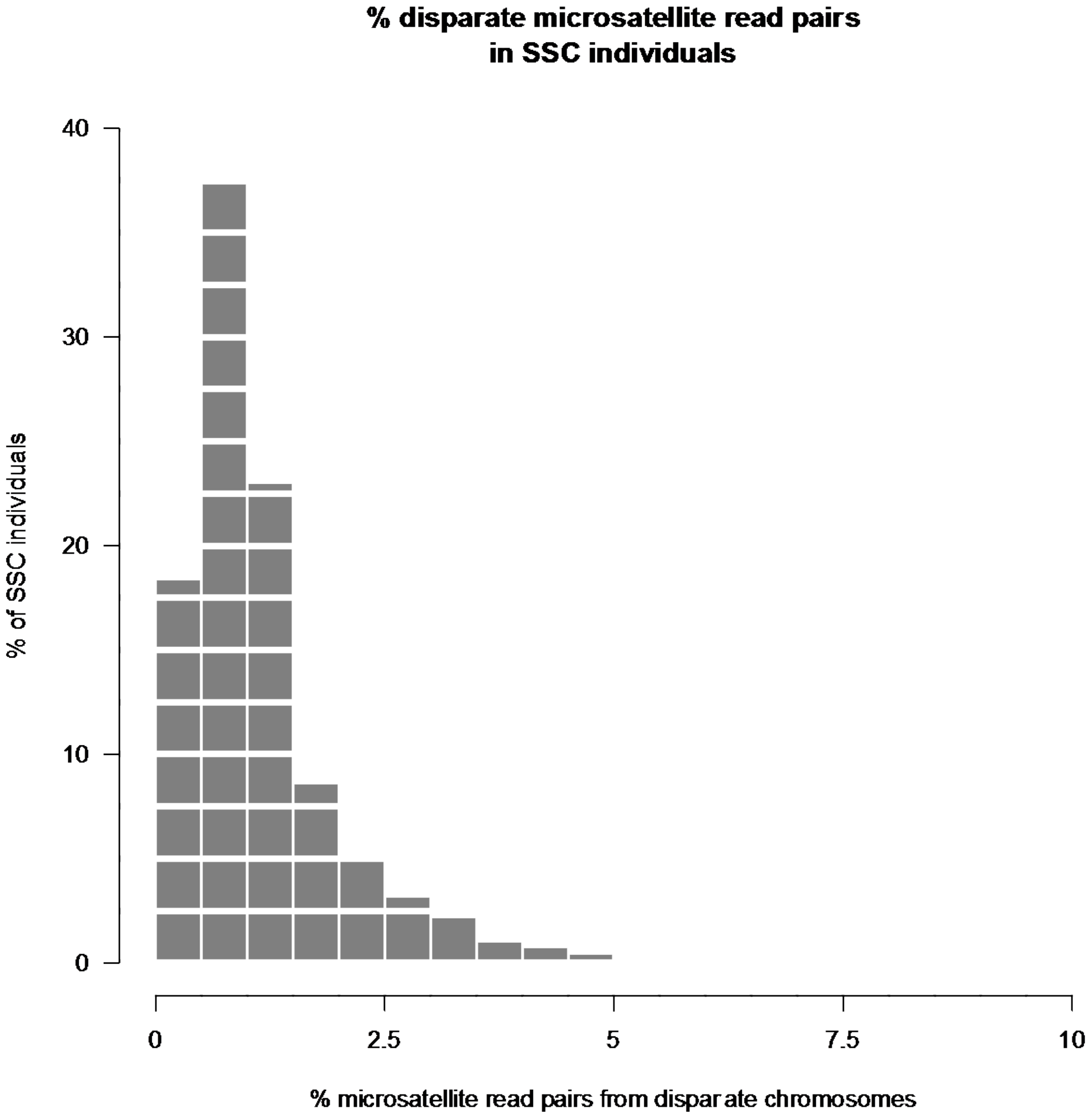


Figure 6.9: Percentage of microsatellite read pairs that map to disparate chromosomes in each SSC individual. Approximately 1% of read pairs map to disparate chromosomes in the average SSC individual.

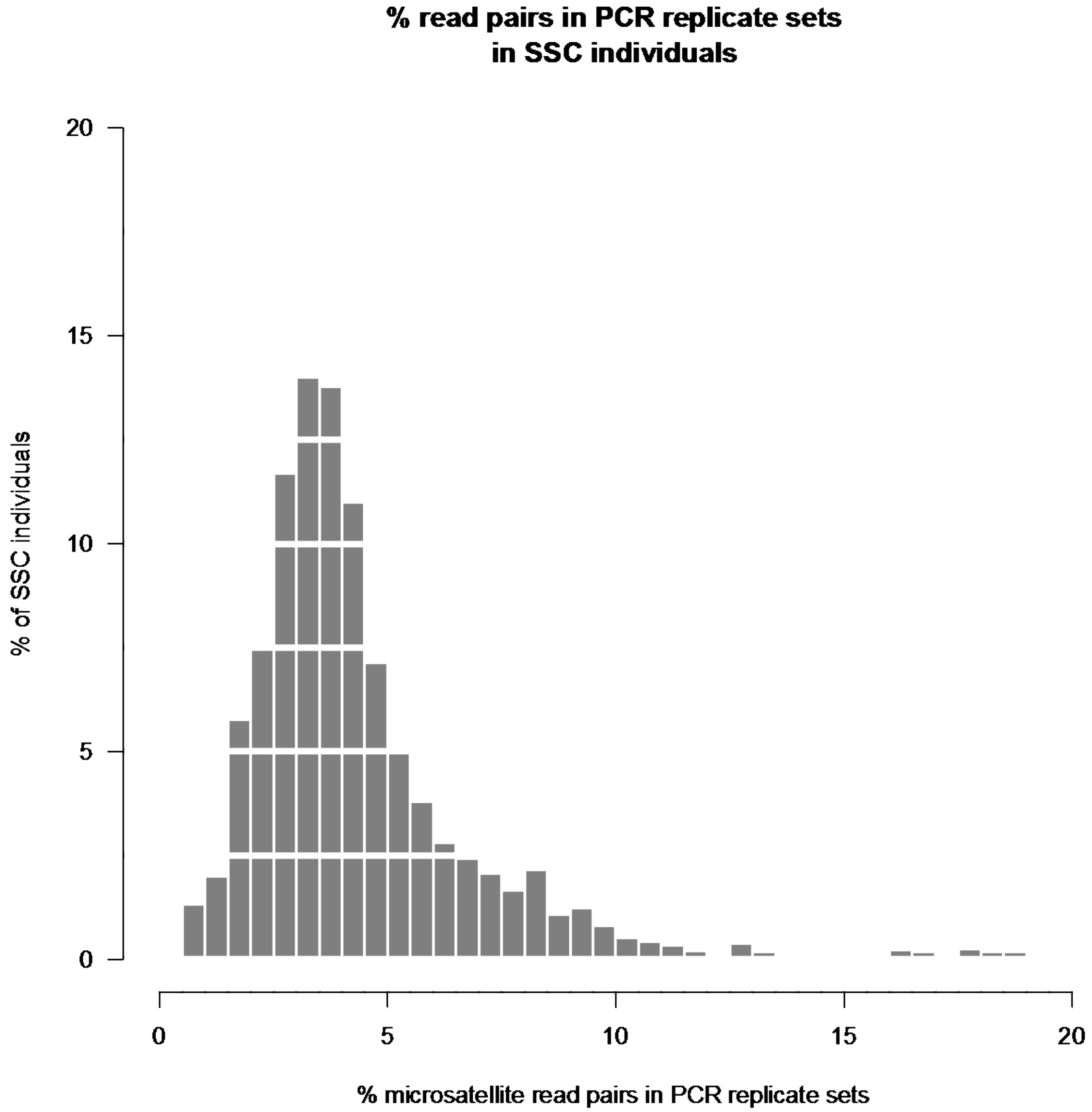


Figure 6.10: Percentage of microsatellite read pairs in PCR replicate sets in each SSC individual. Approximately 4% of read pairs are in PCR replicate sets in the average SSC individual.

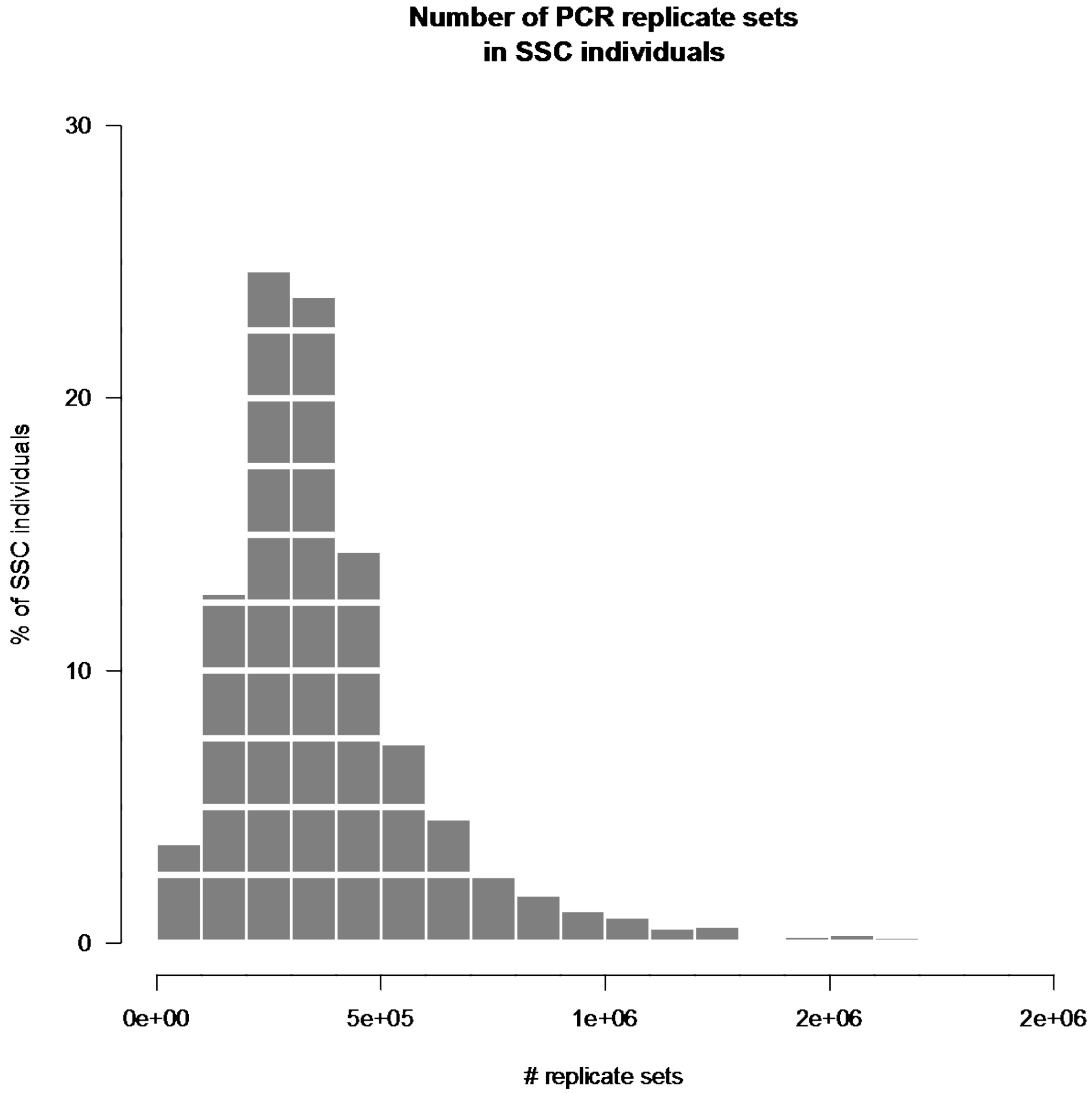


Figure 6.11: Number of PCR replicate sets in each SSC individual. The average SSC individual has ~400,000 replicate sets.

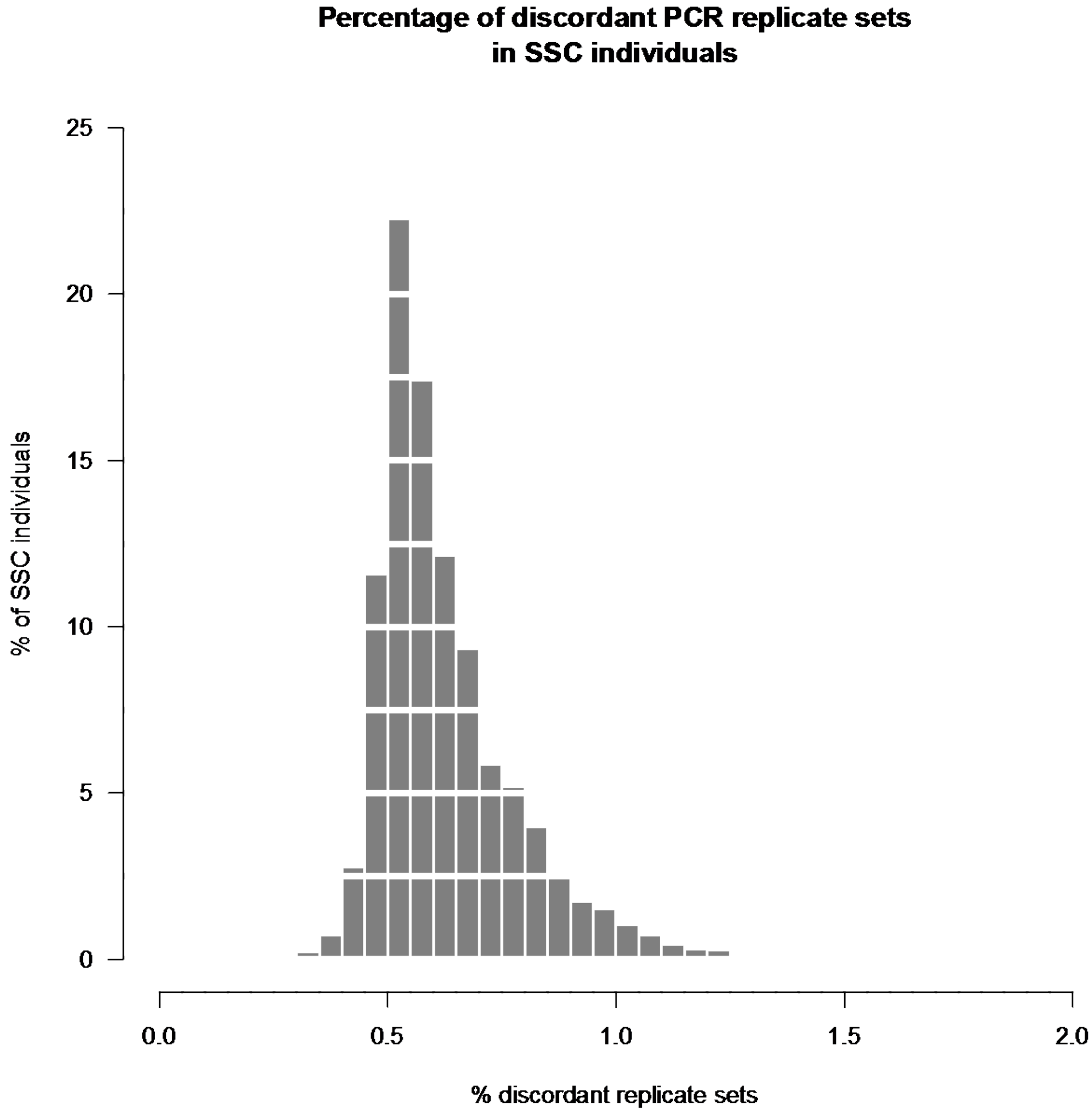


Figure 6.12: Percentage of discordant PCR replicate sets per SSC individual. A PCR replicate set is considered discordant if all reads in the set do not report identical microsatellite tract lengths. Only 0.6% of all PCR replicate sets are discordant in the average SSC individual.

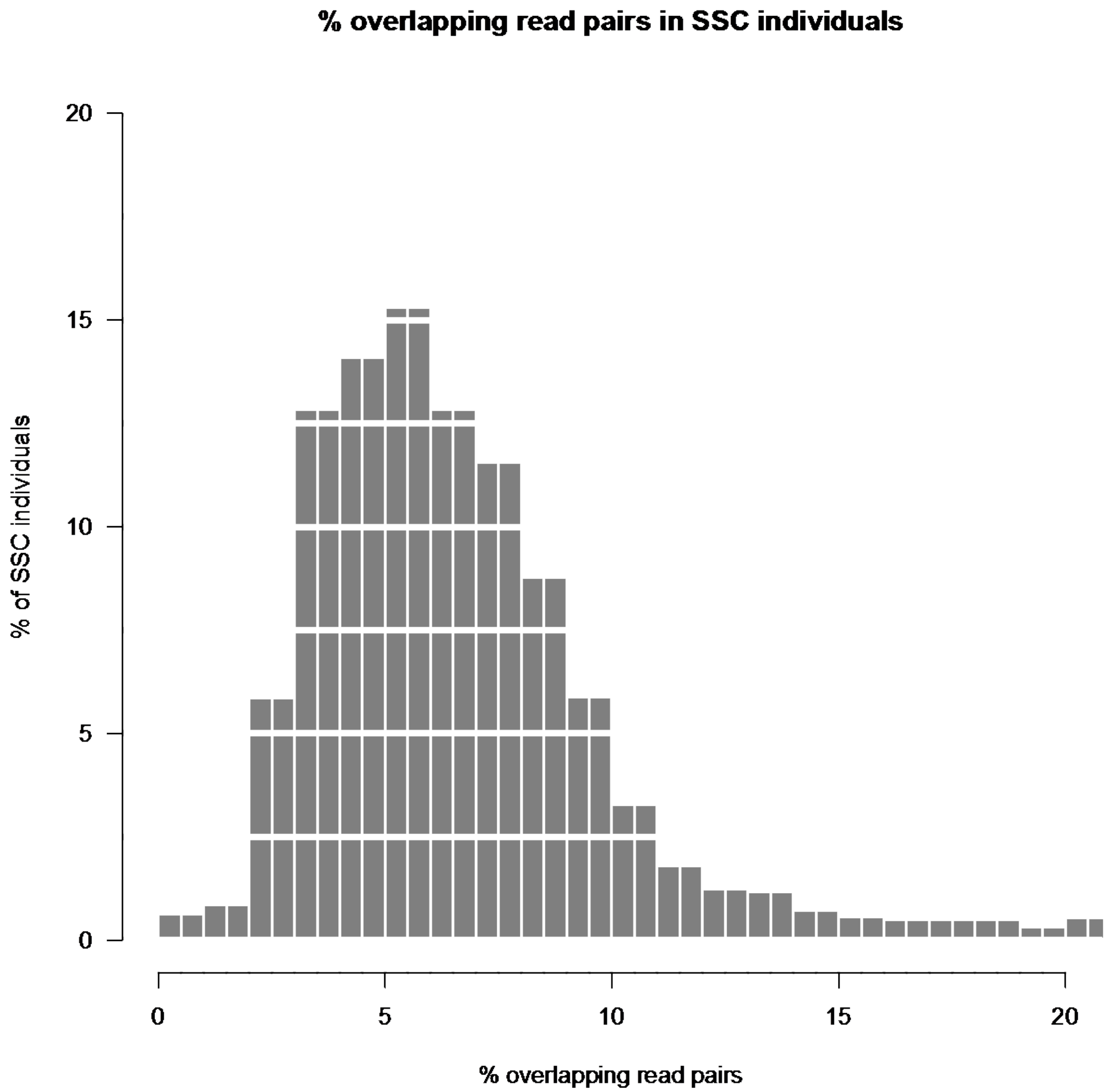


Figure 6.13: Percentage of overlapping read pairs in each SSC individual that cover the same microsatellite locus/loci. 7% of read pairs in the average SSC individual contain the same microsatellite tract in both reads.

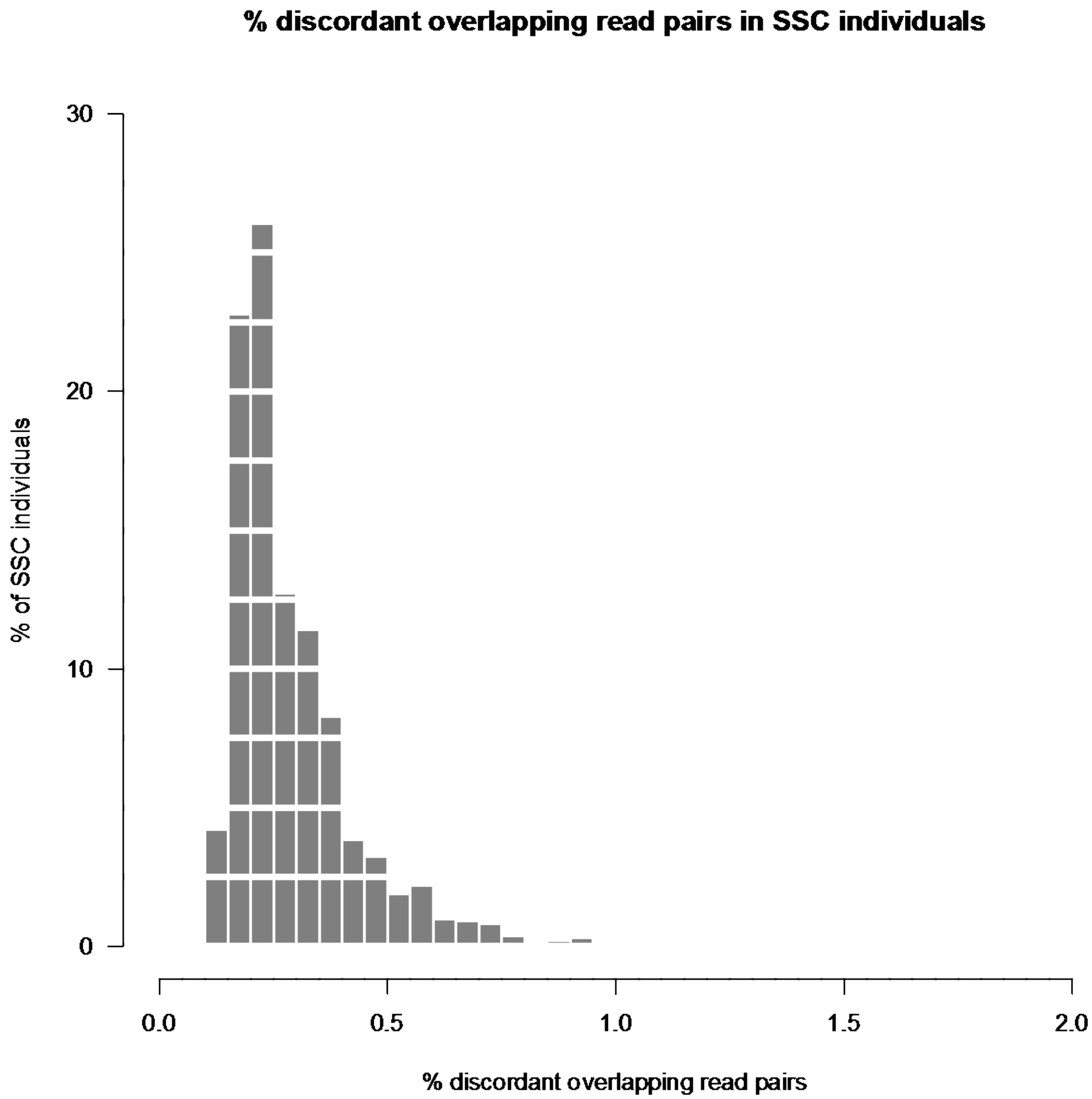


Figure 6.14: Percentage of overlapping read pairs reporting discordant microsatellite tract lengths. Only 0.3% of overlapping read pairs report discordant tract lengths for the same microsatellite locus in the average SSC individual.

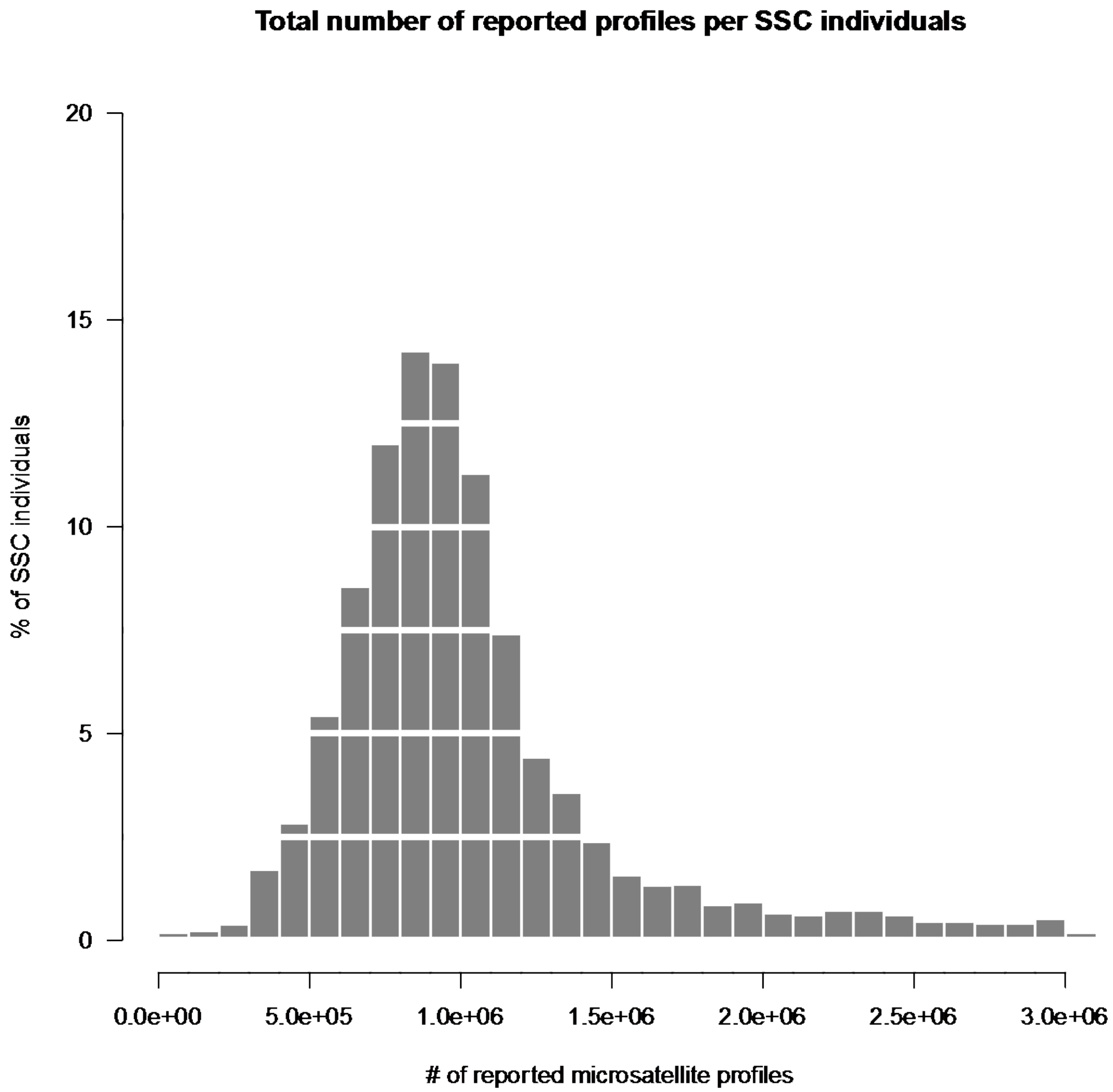


Figure 6.15: Number of microsatellite profiles observed per SSC individual. The average SSC individual has ~1 million profiles for distinct microsatellite loci.

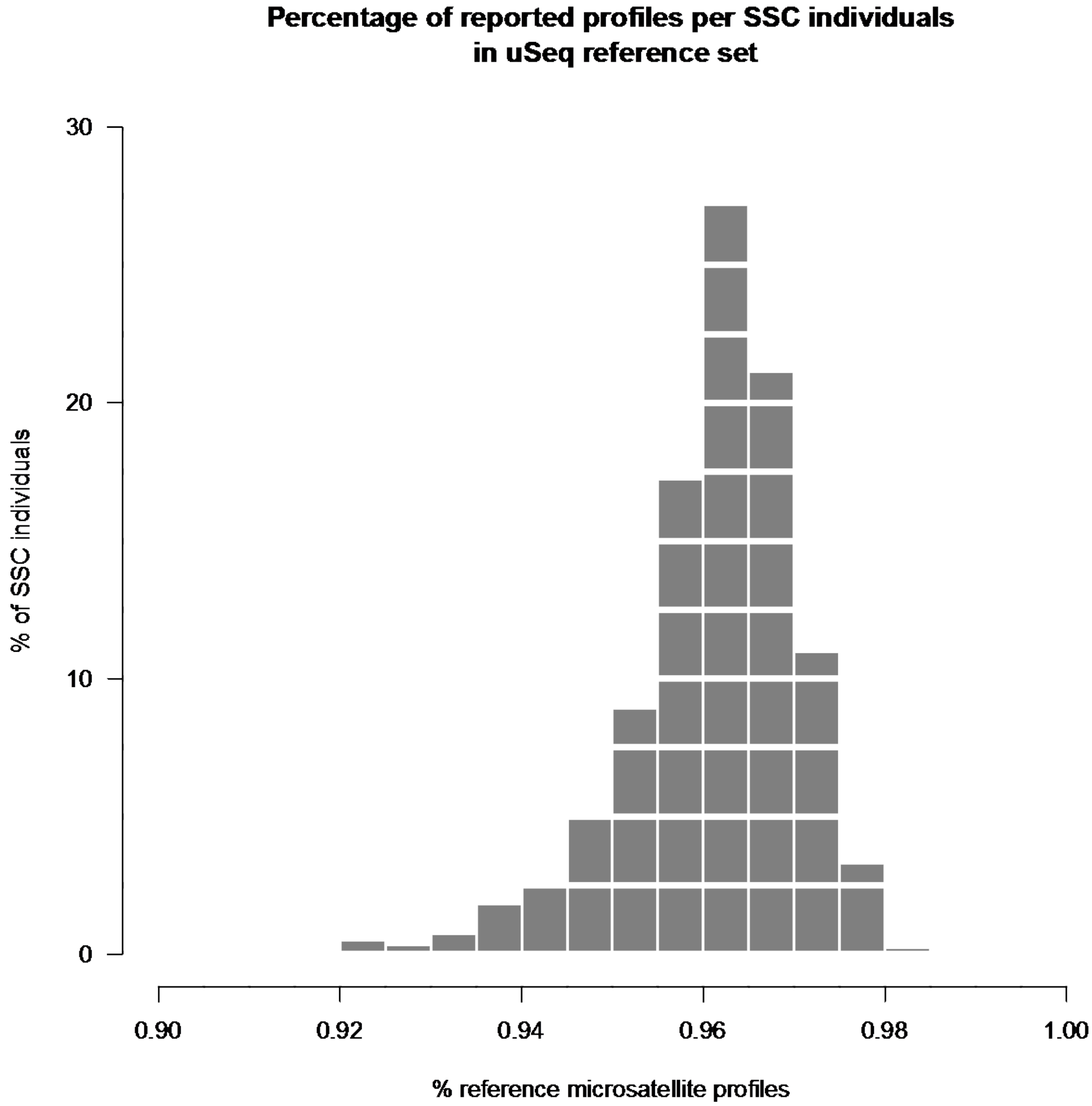


Figure 6.16: Percentage of reference microsatellite profiles per SSC individual. Note that the x-axis starts at 0.90, not 0.00. In the average SSC individual, >96% of all reported profiles are for uSeq reference microsatellite loci. Many non-reference profiles reported are likely due to consistent, but incorrect, alignments of sequencing reads due to terminal microsatellite repeats. The percentage of reported reference microsatellite profiles might increase if uSeq incorporated a local realignment step.

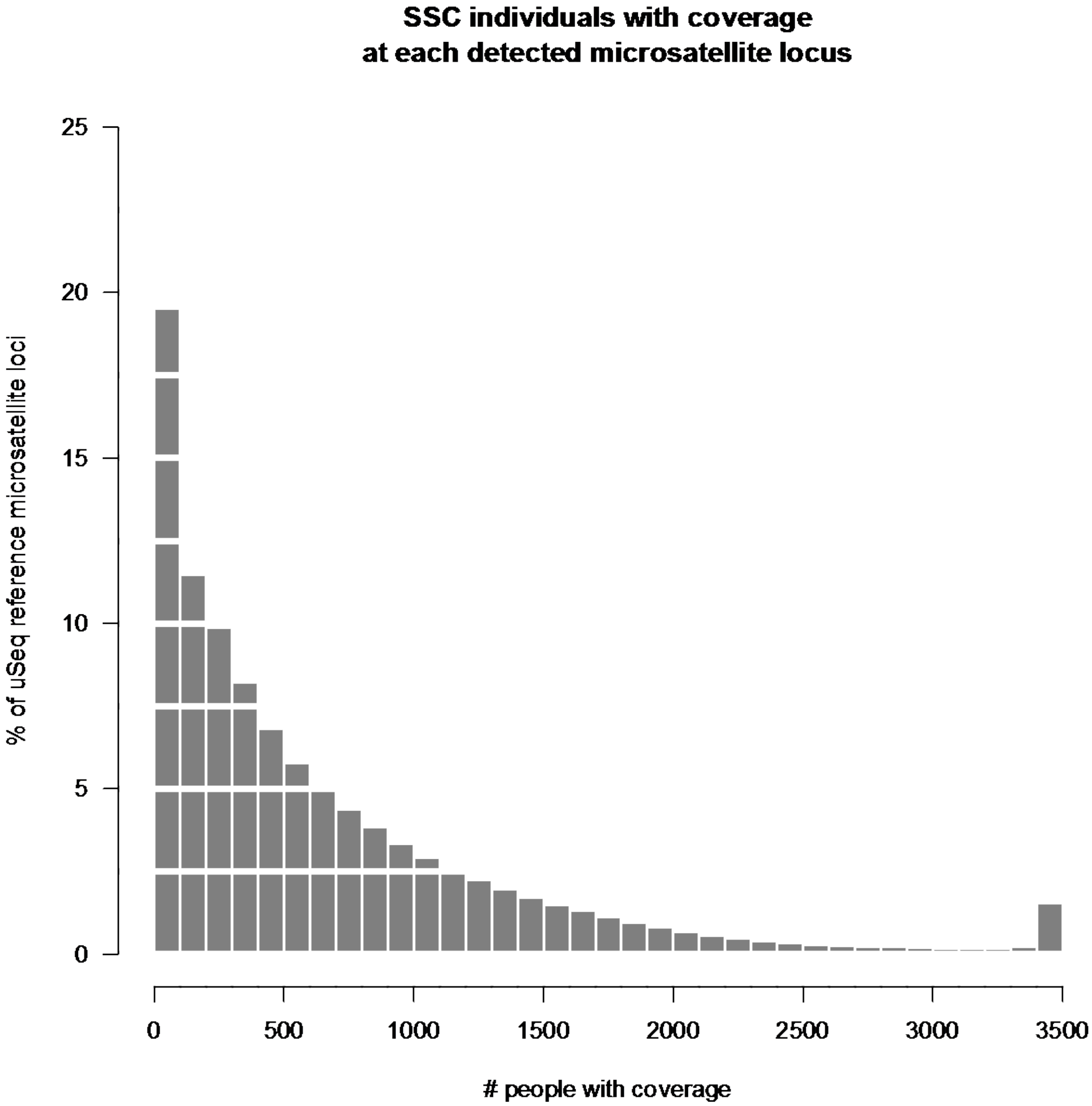


Figure 6.17: Number of SSC individuals with any coverage at each observed uSeq reference microsatellite locus. Over 92% of the reference microsatellite loci observed have coverage in less than half of all SSC individuals.

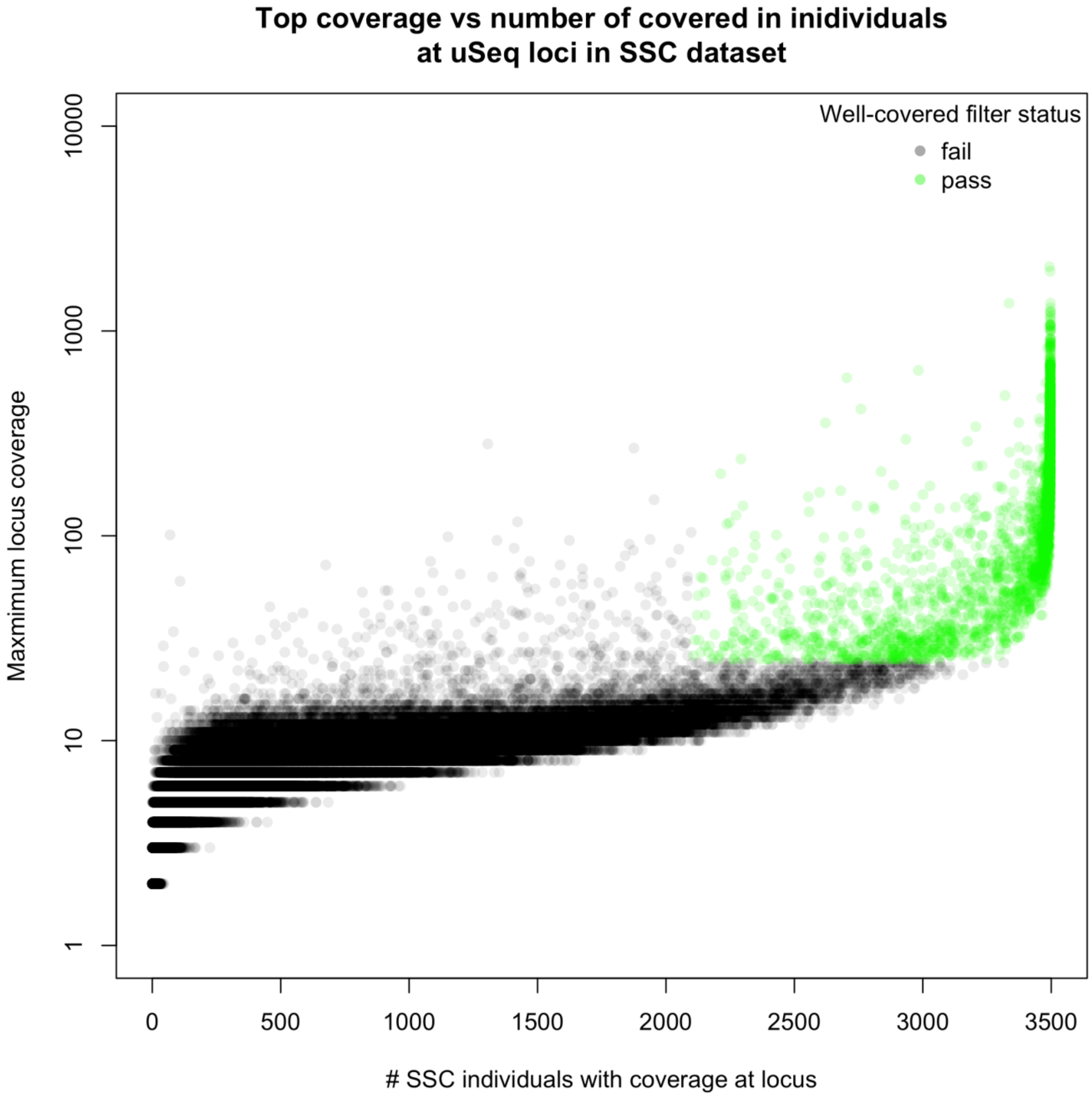


Figure 6.18: Loci with coverage in more SSC individuals have higher maximum locus coverage. Maximum locus coverage is the highest coverage in any SSC individual at a locus. Loci with higher maximum coverage are usually observed in more SSC individuals.

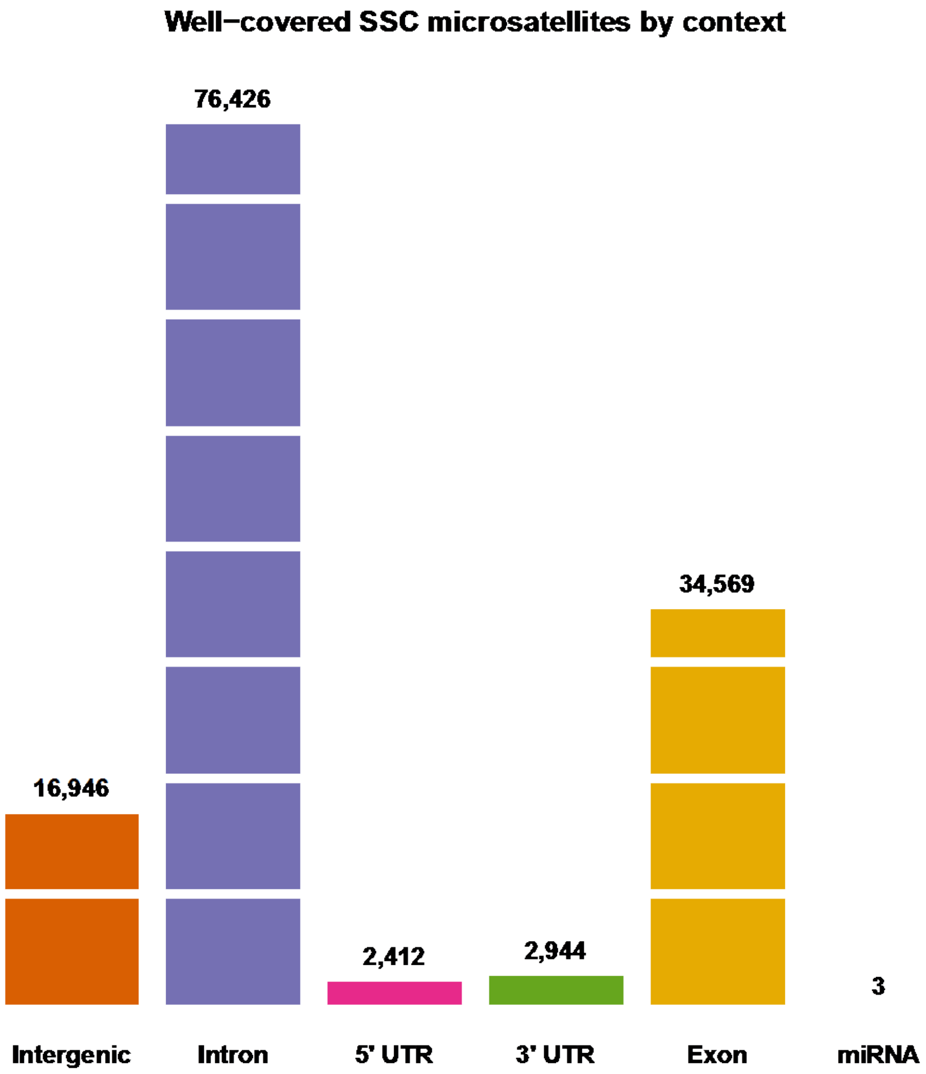


Figure 6.19: Well-covered uSeq reference microsatellite loci in SSC exome sequencing data. Most well-covered microsatellite loci are in introns, exons, or intergenic regions.

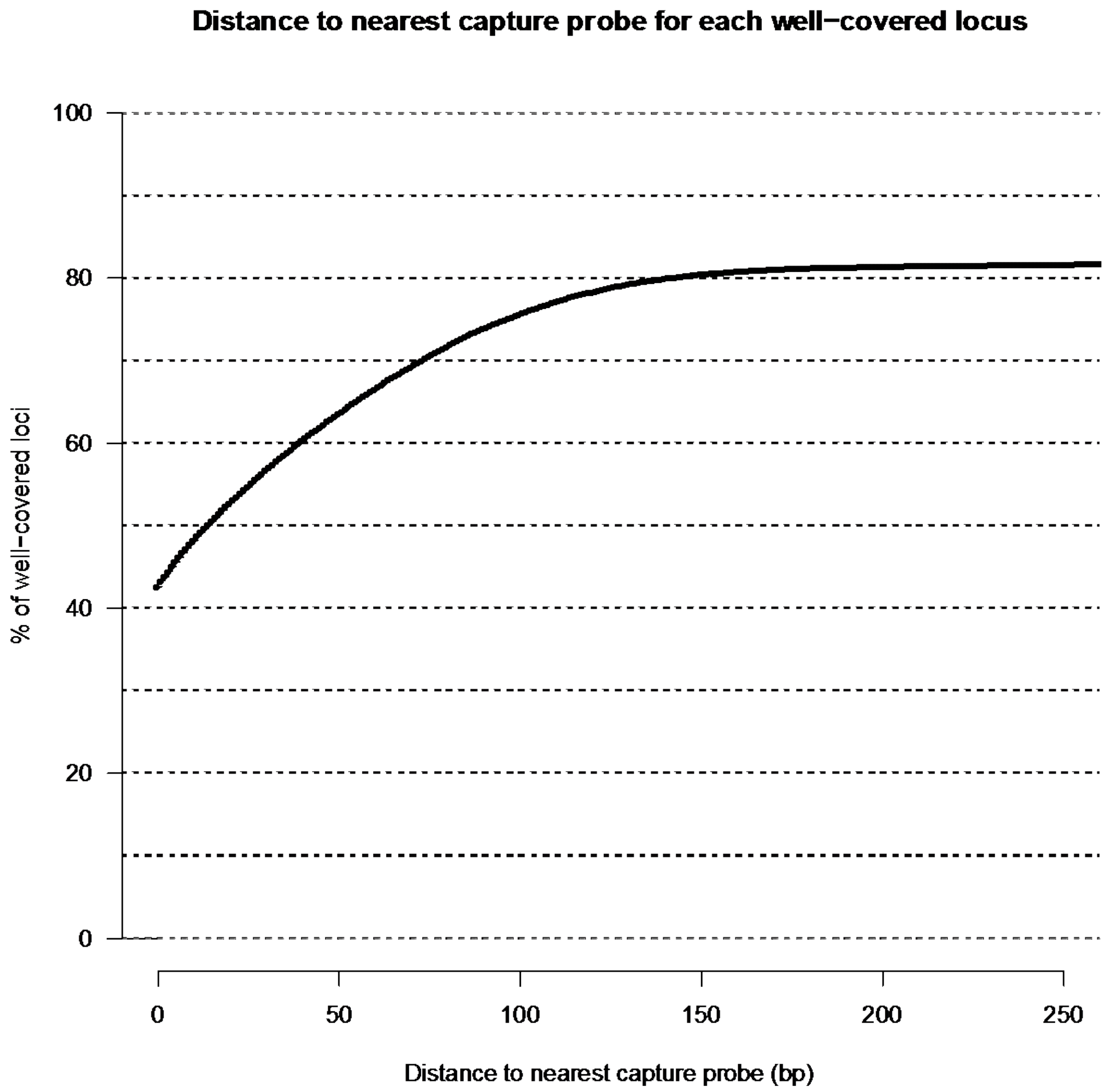


Figure 6.20: Distance of well-covered loci to nearest capture probe in NimbleGen EZ Exome V2.0 kit. Just over 80% of all well-covered loci are within 250 bp of a capture probe, or are targeted directly by a capture probe. The remaining ~20% of well-covered loci may have some degree of homology with the capture probe sequences.

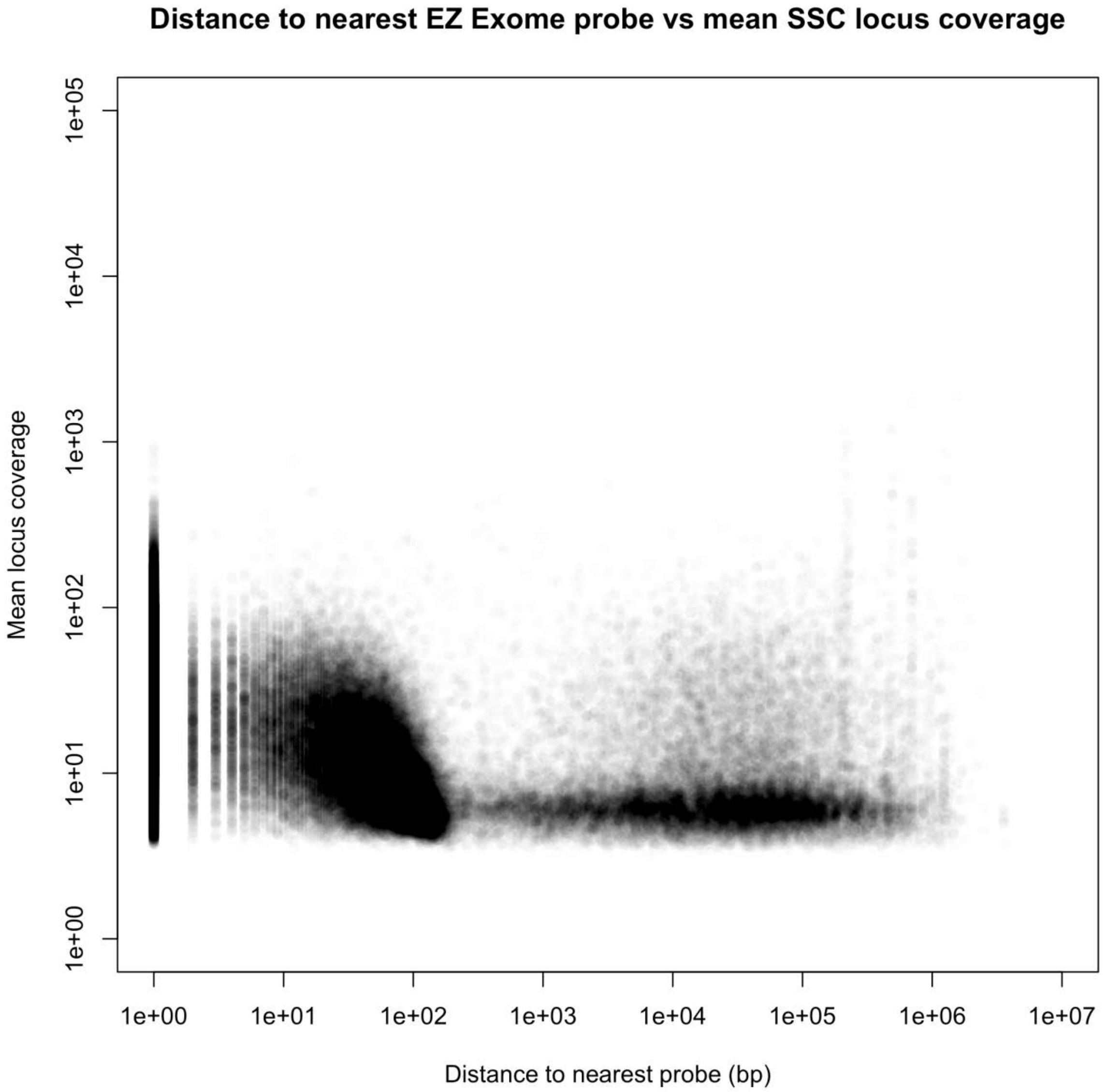


Figure 6.21: Mean locus coverage decreases for well-covered loci that are further from EZ Exome capture probes. Well-covered loci that are further away from an exome capture probe tend to have uniformly lower coverage than those that are either targeted directly or within ~250 bp of a probe.

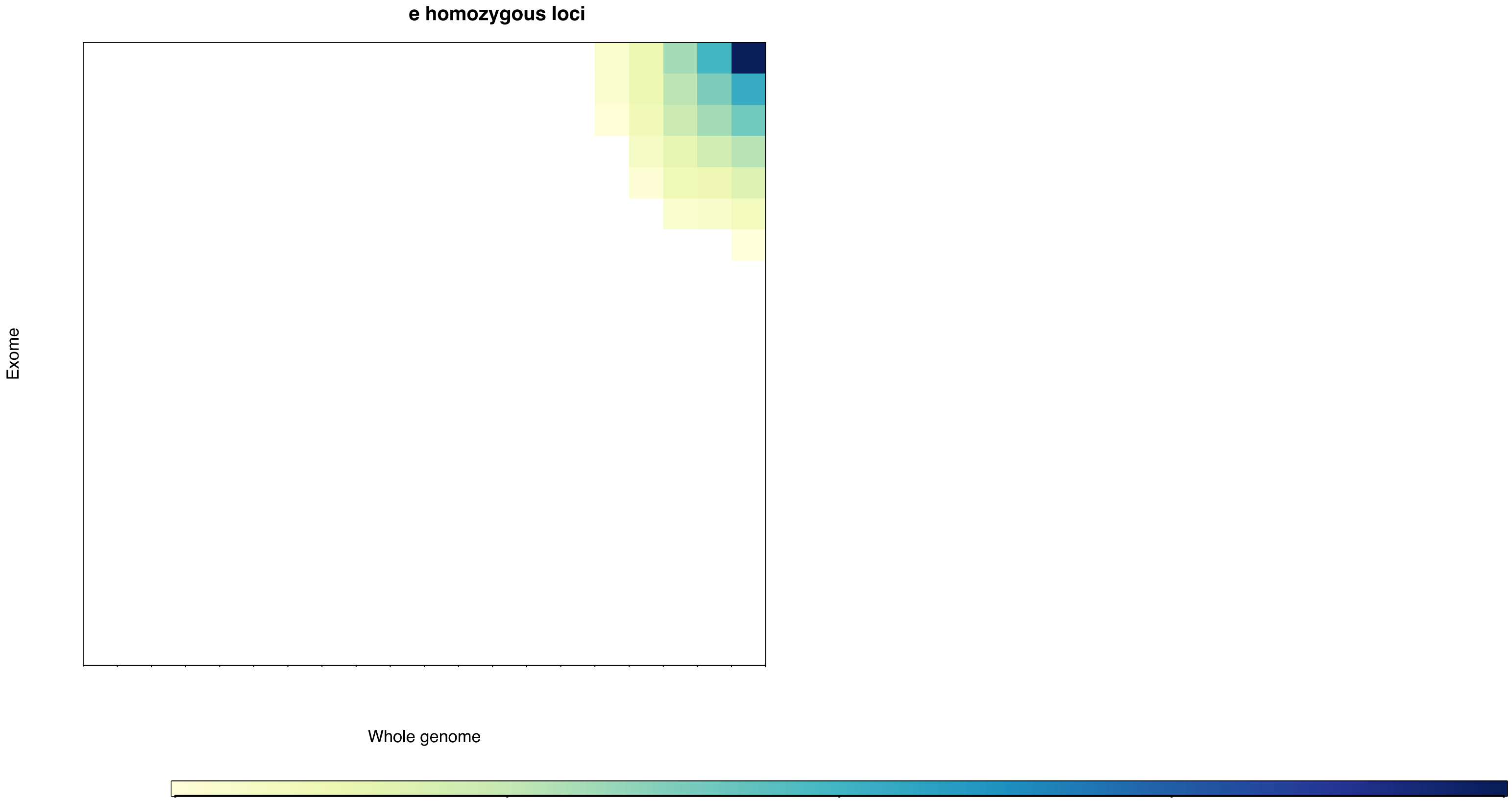


Figure 6.22: Comparing coverage of high confidence exome genotypes in whole genome and exome sequencing data. Analysis was performed on eight SSC individuals from two families. Exome genotypes have a minimum confidence of 0.9. The proportion of total coverage at allele 1 in whole genome and exome sequencing data for high-confidence heterozygous genotypes is the coverage for the first allele in either dataset, divided by the total locus coverage for the appropriate individual. Coverage comparisons are split by uSeq genotype call—(**A**) homozygous and (**B**) heterozygous genotypes. The proportion of coverage for the first allele in whole genome and whole exome sequencing are consistent, suggesting that high confidence uSeq microsatellite genotypes are reproducibly observed when using different sequencing protocols.

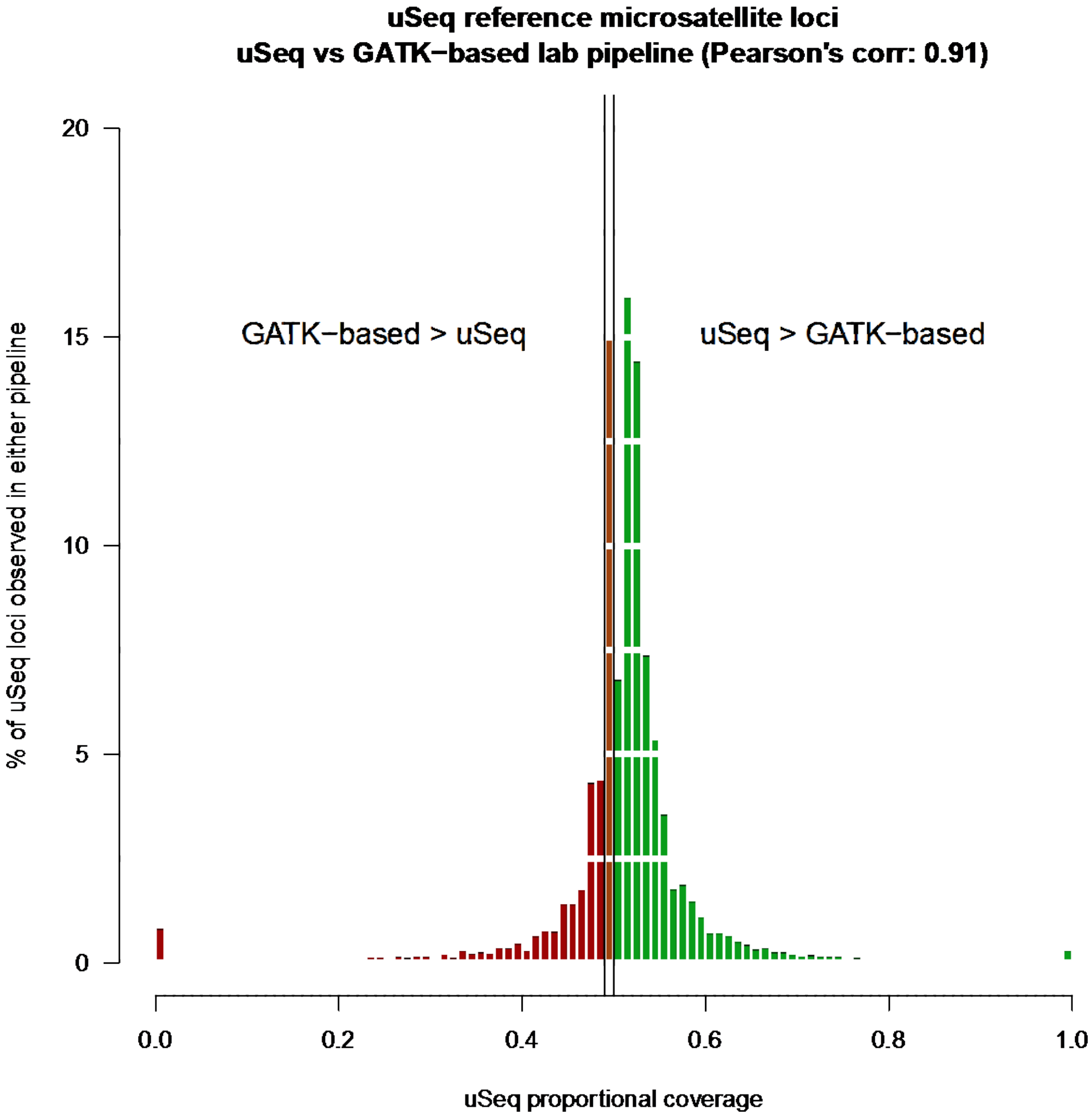


Figure 6.23: Proportional coverage in an SSC individual at microsatellite loci compared to GATK. Only reference microsatellite loci where either pipeline had at least 10X coverage were compared. Proportional coverage is calculated as (uSeq coverage)/(uSeq coverage + GATK-based coverage). Proportional coverage is shown for an SSC individual SSC08278. At the average locus observed by either pipeline, uSeq has 6% more reads than the GATK-based pipeline. Small discrepancies in coverage could be due to variable mapping qualities assigned to the same read in different pipelines or interrupted/short microsatellites only detected by the GATK-based pipeline. Loci unique to the GATK-based pipeline (uSeq proportional coverage of 0) tend to be microsatellite loci with identical flanking sequence, but different tract lengths.

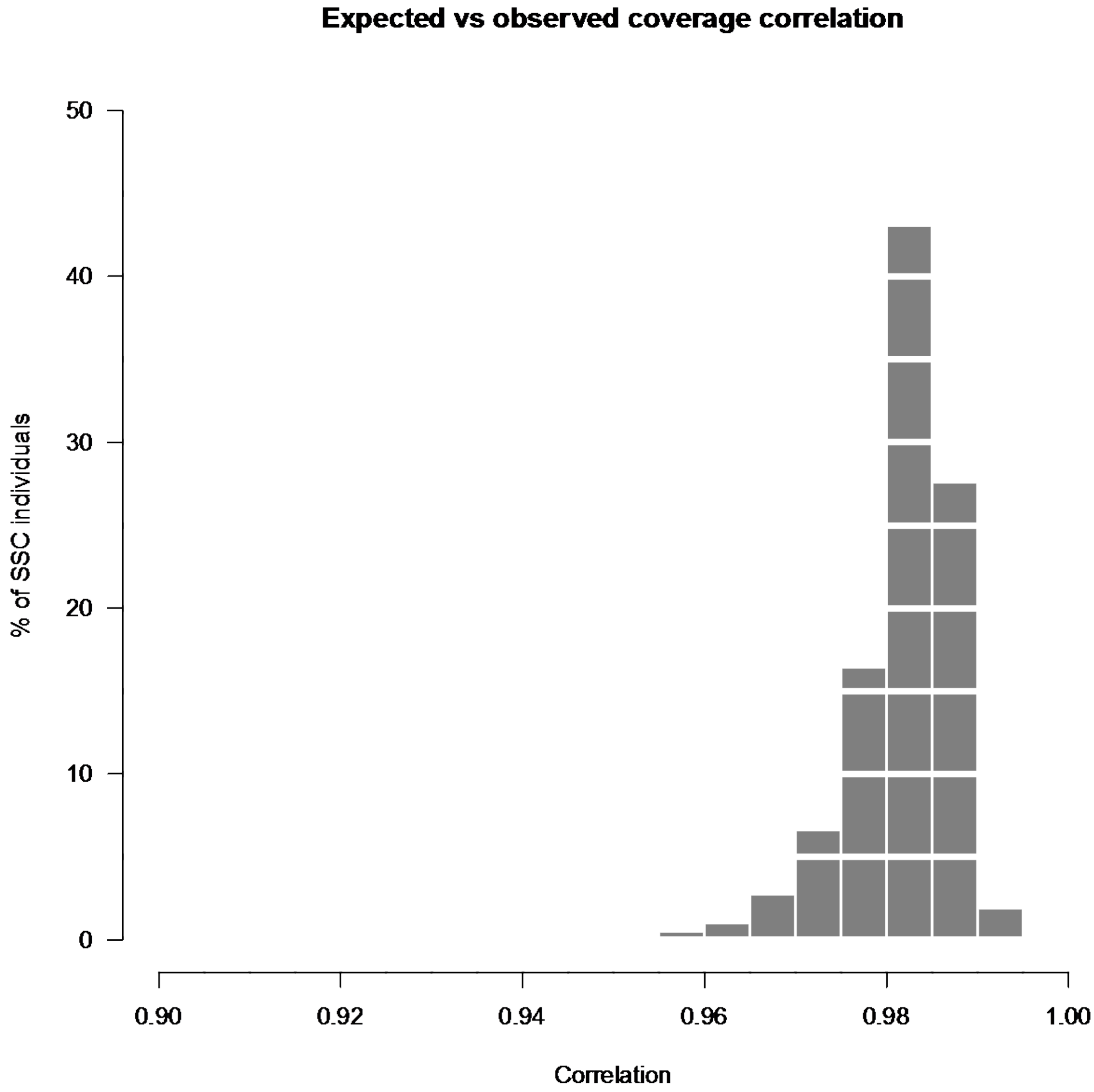


Figure 6.24: Correlation between observed coverage and SVD-inferred per-allele, per-person expected coverage estimators for each SSC individual. The average SSC individual has a correlation >0.98 between their expected and observed coverage, which is consistent with the simulation results from 5.3.3.2.6. This suggests that estimating per-allele expected coverage from observed coverage in the SSC dataset using SVD is performing as expected.

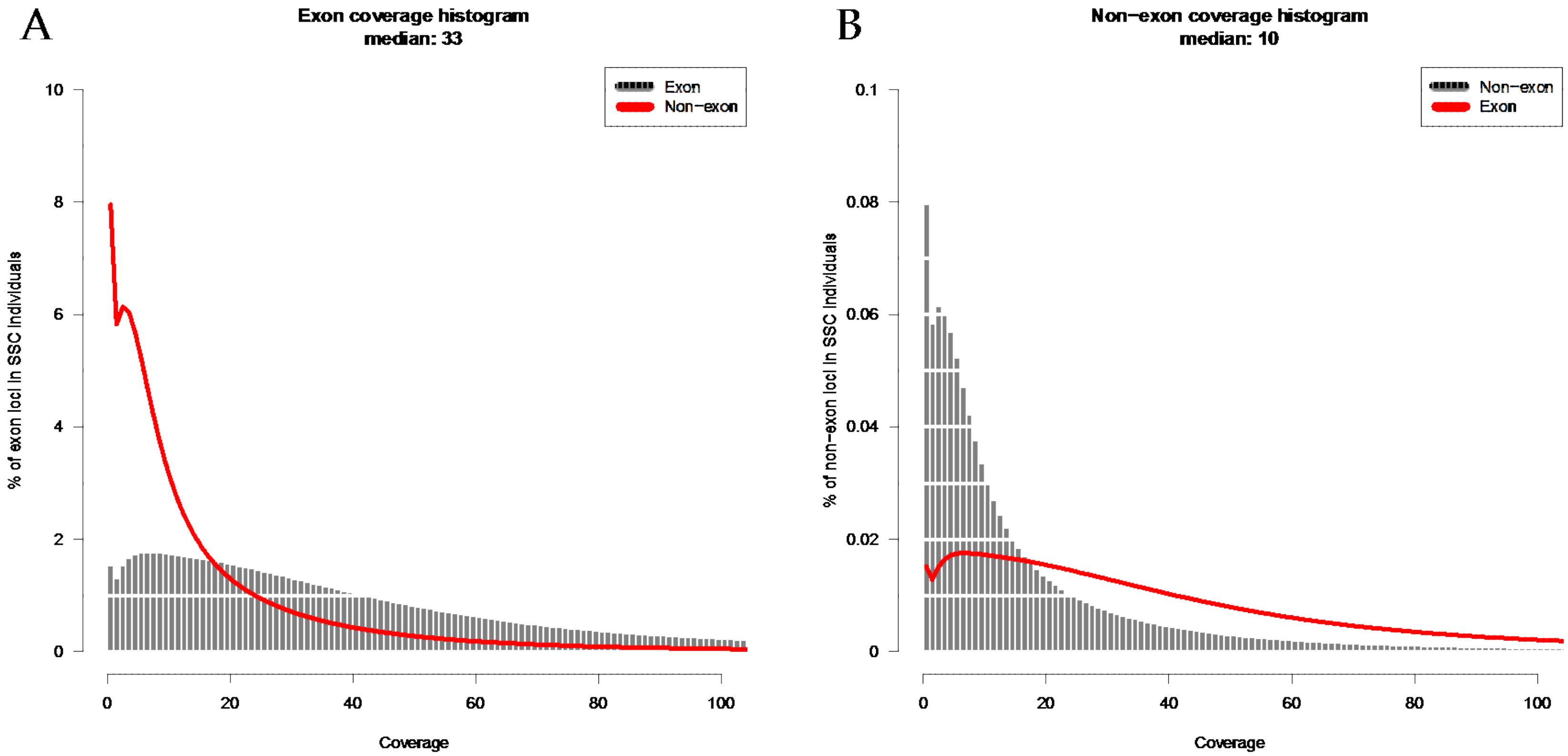


Figure 6.25: Coverage across all SSC individuals at well-covered loci (**A**) in exons and (**B**) outside of exons. In (**A**), the median coverage for all exon microsatellite loci and all SSC individuals is 33. The exon coverage distribution is highly overdispersed—many individuals have coverage higher than the median at many loci. Coverage at exon microsatellites is very good—over 70% of all exon loci in all SSC individuals have coverage ≥20. In (**B**), the median coverage for all non-exon microsatellite loci and all SSC individuals is just 10. Fewer than 40% of all non-exon loci and all SSC individuals have coverage ≥20.

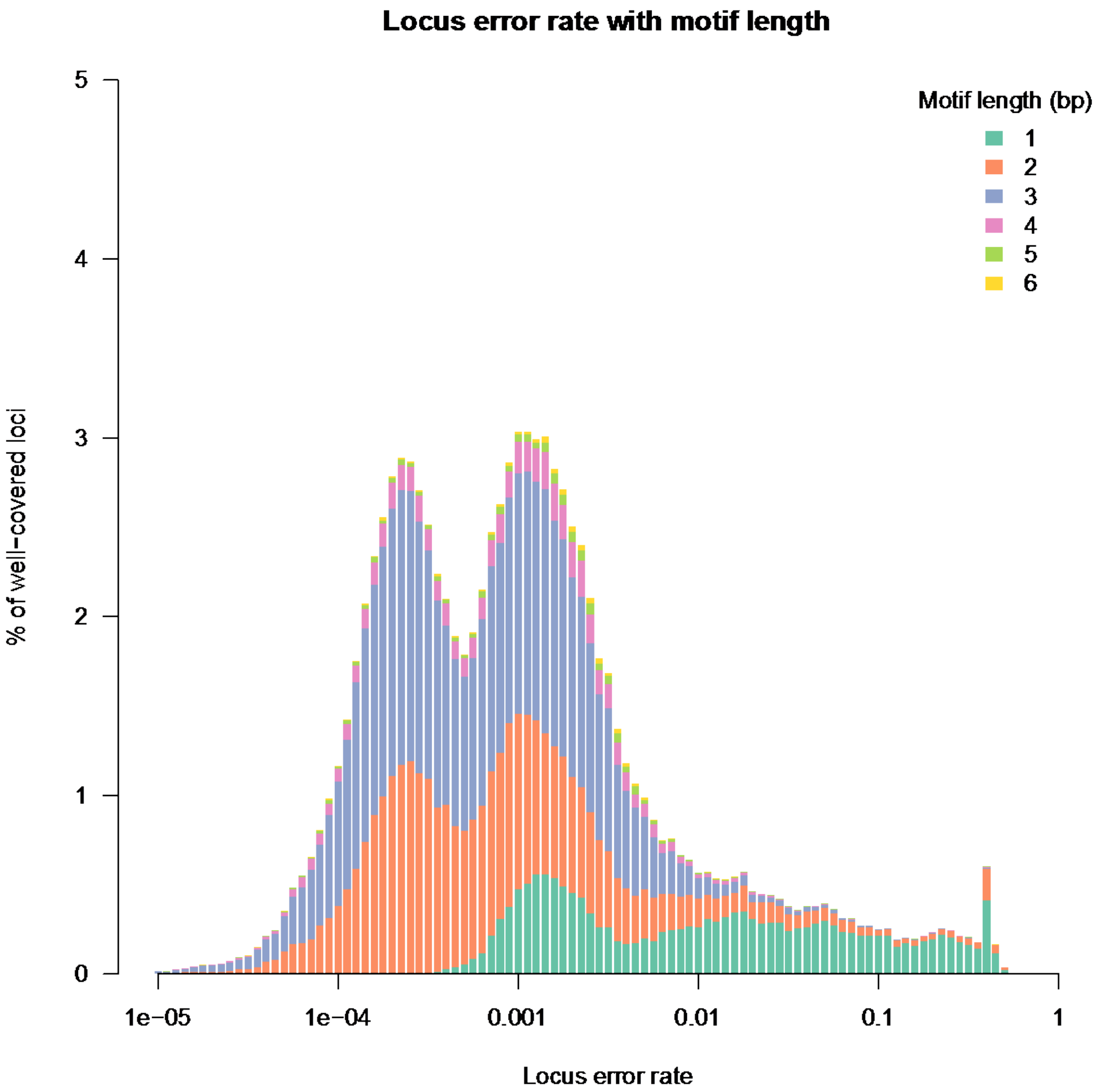


Figure 6.26: EM-estimated locus-specific error rates. Colors represent the proportion of loci in a particular error rate bin with a given motif length. Microsatellites with 1 and 2 bp motifs occasionally have considerably high error rates—in particular, 17% of microsatellites with 1 bp motifs have error rates ≥0.10. Microsatellites with motifs ≥3 bp tend to be much more stable—<1% of these loci have error rates above 0.01.

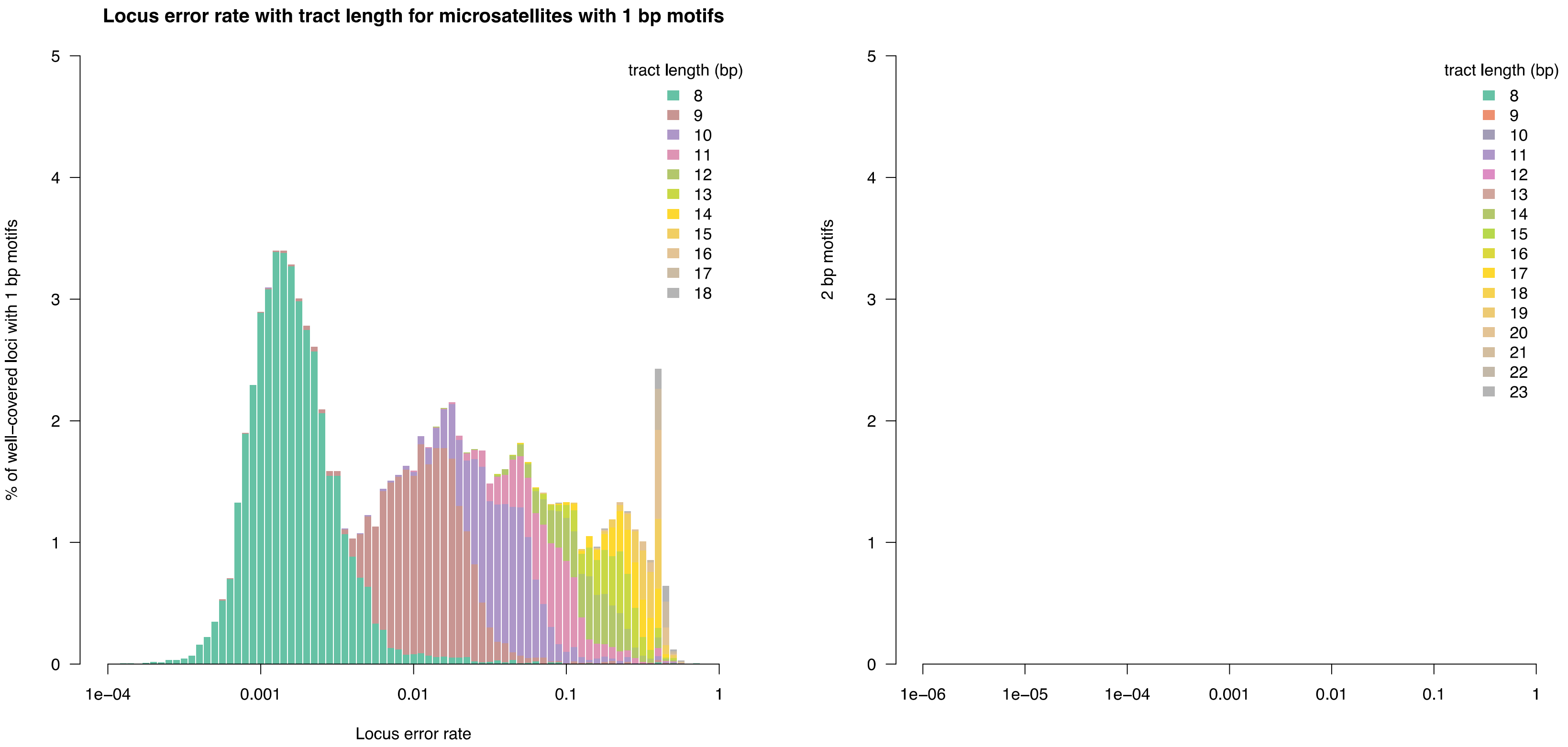


Figure 6.27: Locus-specific error rate estimates by reference tract length. Tract lengths are shown for individual motif lengths of (**A**) 1 bp and (**B**) 2 bp. Note that although the y-axes for both plots are identical, the x-axes differ—(**A**) ranges from 1e-4 to 1, while (**B**) ranges from 1e-6 to 1. The observed bimodality in both (**A**) and (**B**) is due to the minimum detectable tract length—some slippage errors at the shortest tract lengths might not be detectable by uSeq. For a given motif length and reference tract length, error rates can vary by as much as an order of magnitude, suggesting that these parameters—even when combined with base calling quality, flanking GC content, or sequence identity—may not be capable of explaining the variation in error rates among similar loci.

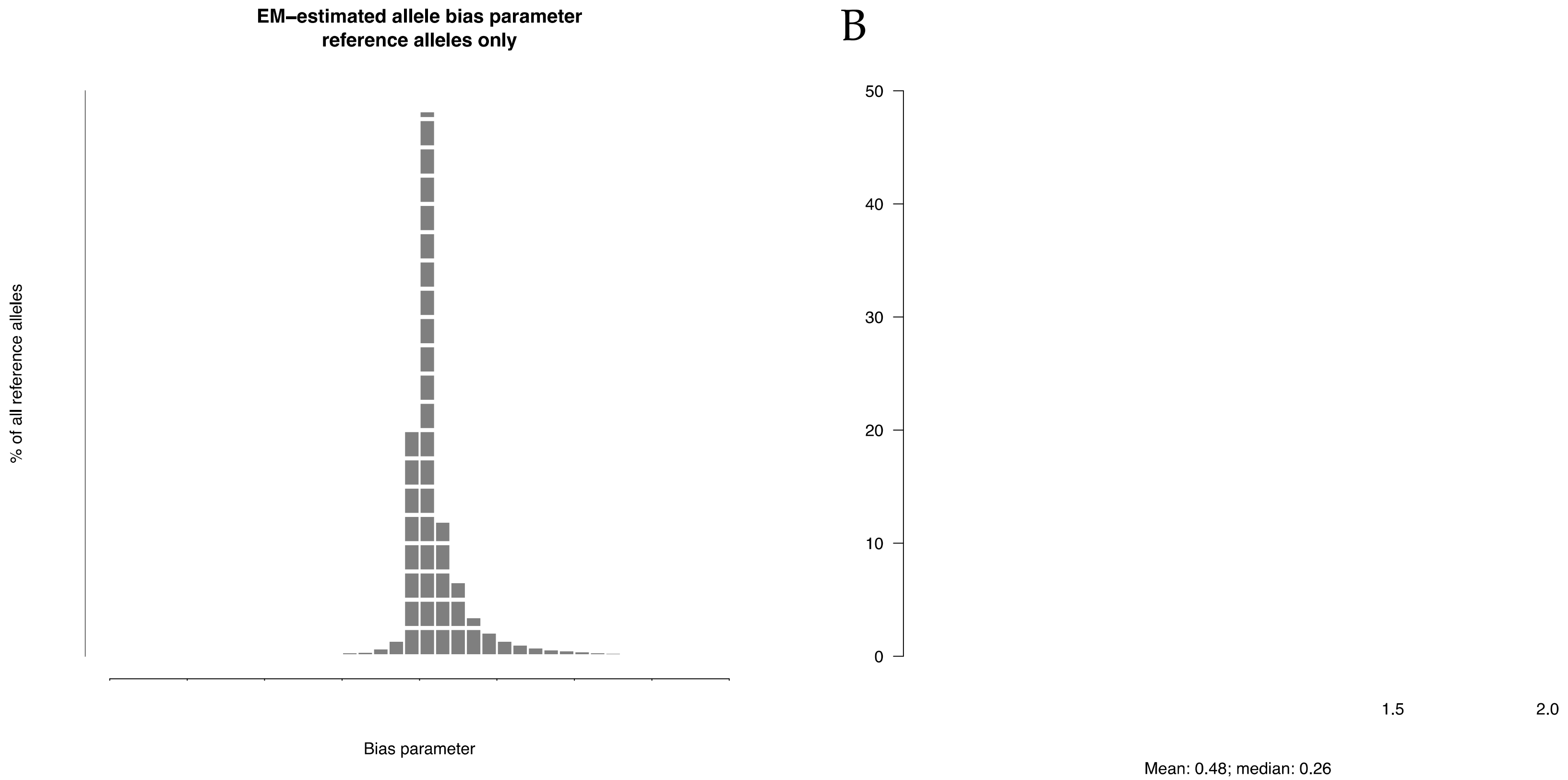


Figure 6.28: EM-estimated per-allele capture bias rates. Bias rates are shown separately for (**A**) alleles with reference tract lengths and (**B**) alleles with non-reference tract lengths. The vast majority of bias estimates for reference alleles range from 0.75 to 1.25, indicating that strong bias is generally not observed for reference alleles. In contrast, <30% of non-reference alleles have bias estimates from 0.75 to 1.25, and >44% have biases <0.15. This suggests that many non-reference alleles are poorly captured in this exome sequencing study. As a result, *de novo* mutations involving alleles with strong bias may not be detected.

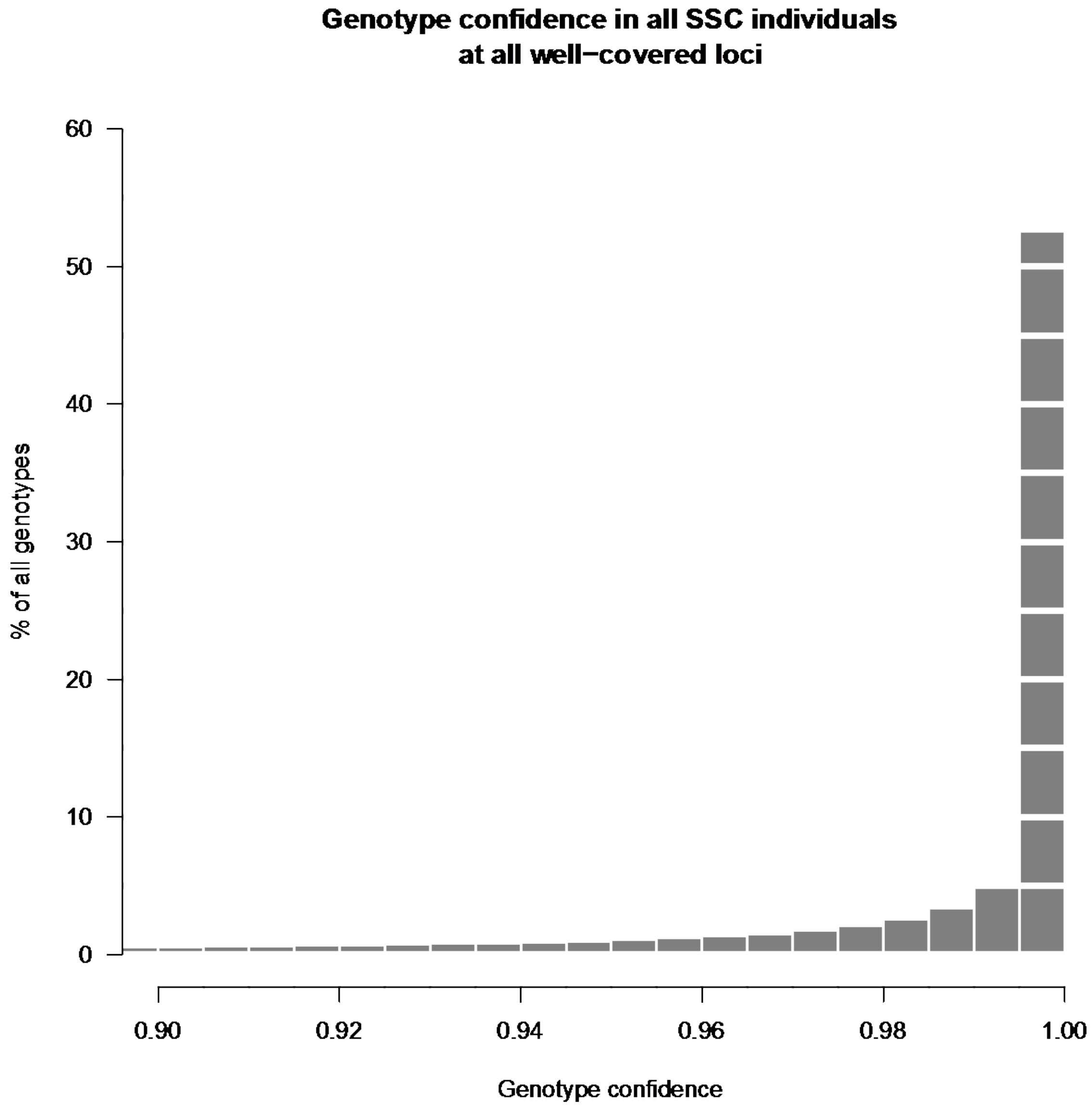


Figure 6.29: Genotype confidence for all SSC genotypes at well-covered loci. The median confidence is >0.99 for the more than 450 million genotypes called in this study.

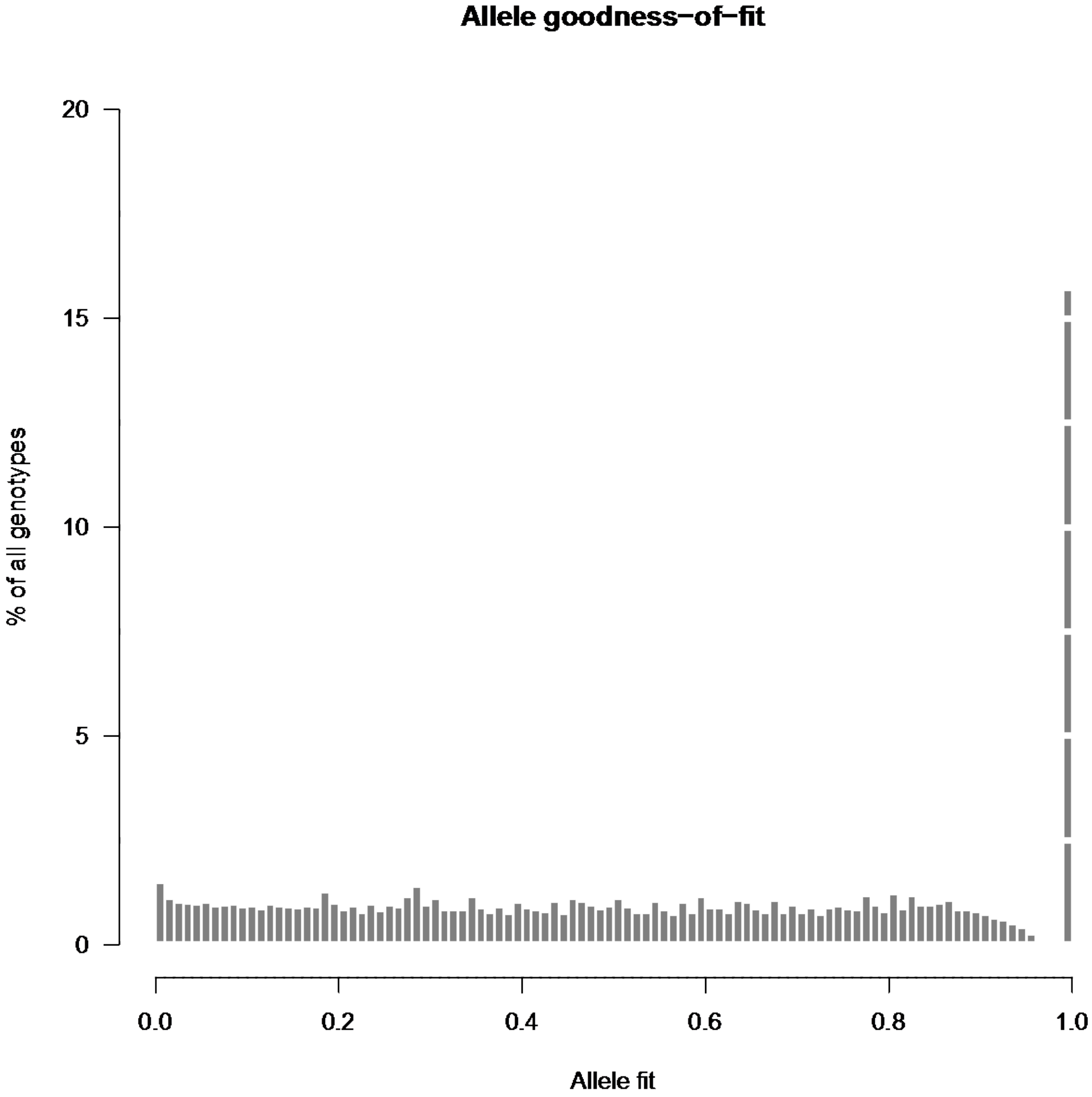


Figure 6.30: Allelic goodness-of-fit for all SSC genotypes at well-covered loci. The distribution of allelic goodness-of-fit p-values is not uniform due to low expected and observed coverage for many loci in the dataset, as discussed in the text.

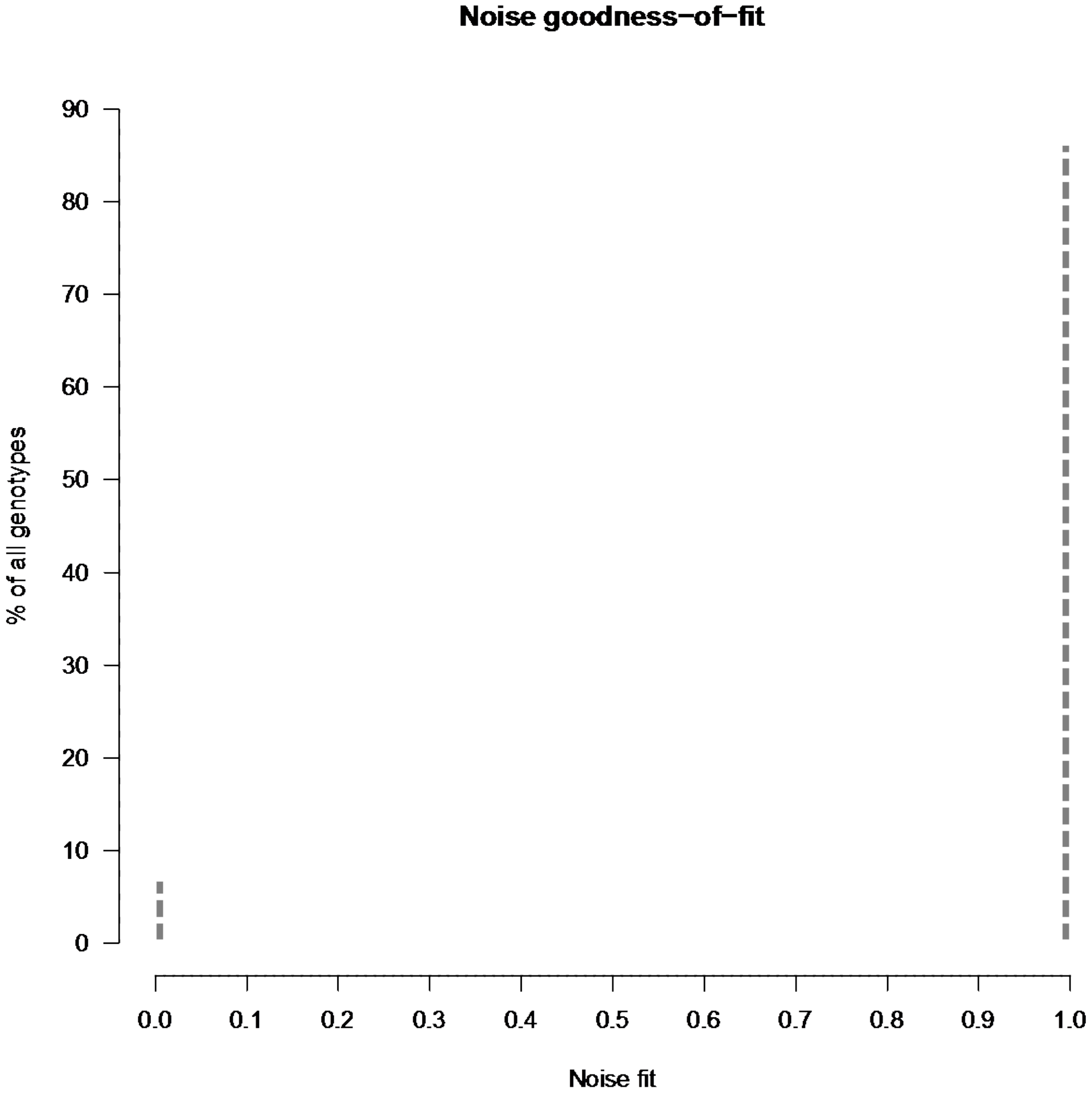


Figure 6.31: Noise goodness-of-fit for all SSC genotypes at all well-covered loci. The distribution of noise goodness-of-fit p-values is exceedingly non-uniform, due to the low expected error coverage at most well-covered loci in SSC individuals.

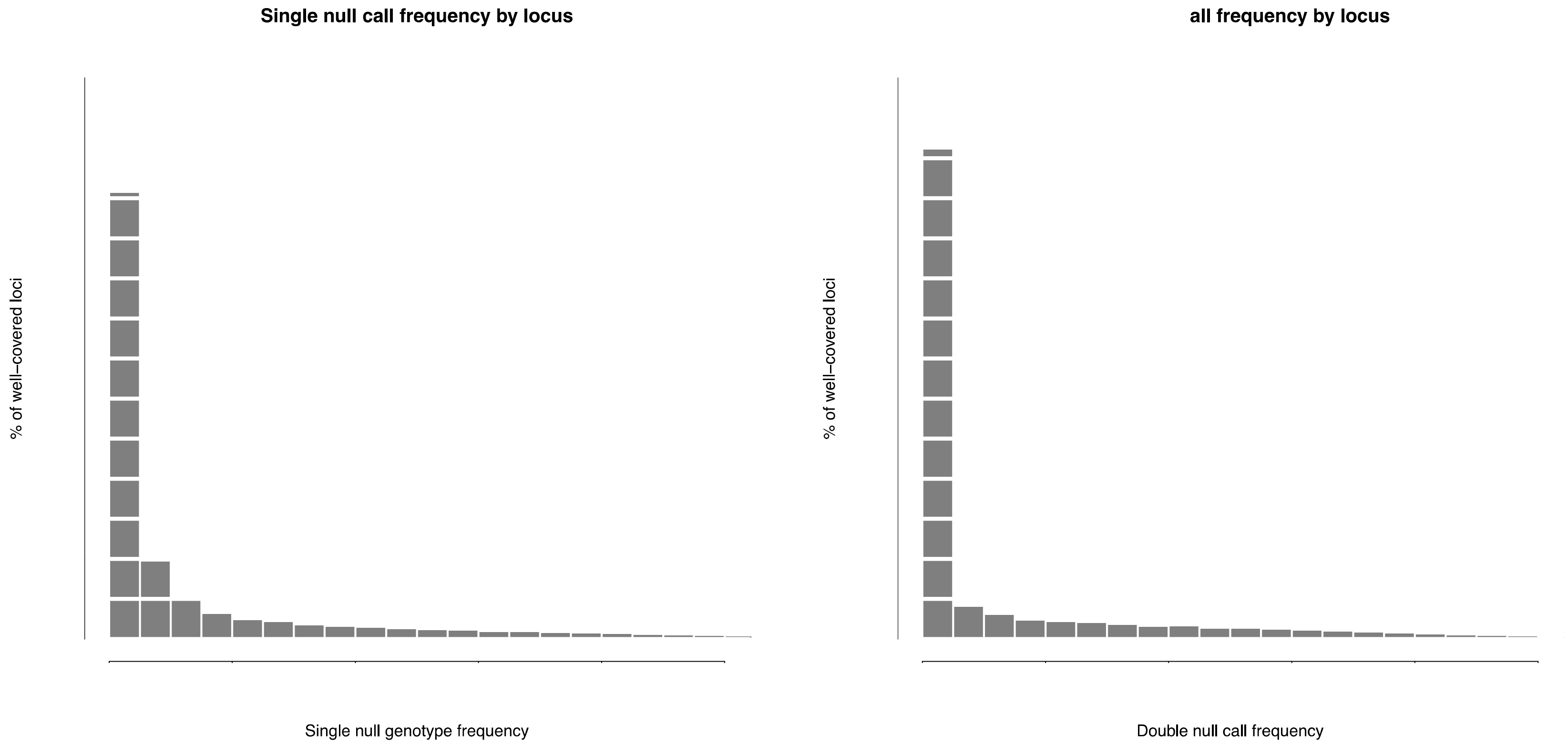


Figure 6.32: Null allele call frequencies in SSC genotypes. Shown are (**A**) single null call frequencies by locus and (**B**) double null call frequencies by locus. In (**A**), the median single null call frequency for all well-covered loci is <0.004, and 90% of all well-covered loci have single null call frequencies below 0.1. In (**B**), the median double null call frequency is < 0.0006, and 88% of all well-covered loci have double null call frequencies below 0.1. This suggests that most individuals at most loci have coverage consistent with bi-allelic genotypes. Many loci with higher single or double null call frequencies are likely due to uniformly low coverage throughout the SSC population at a locus, which would make it difficult to confidently distinguish bi-allelic and null genotype calls. Some loci with high single or double null call frequencies may have undetected microsatellite alleles or may be in a common CNV.

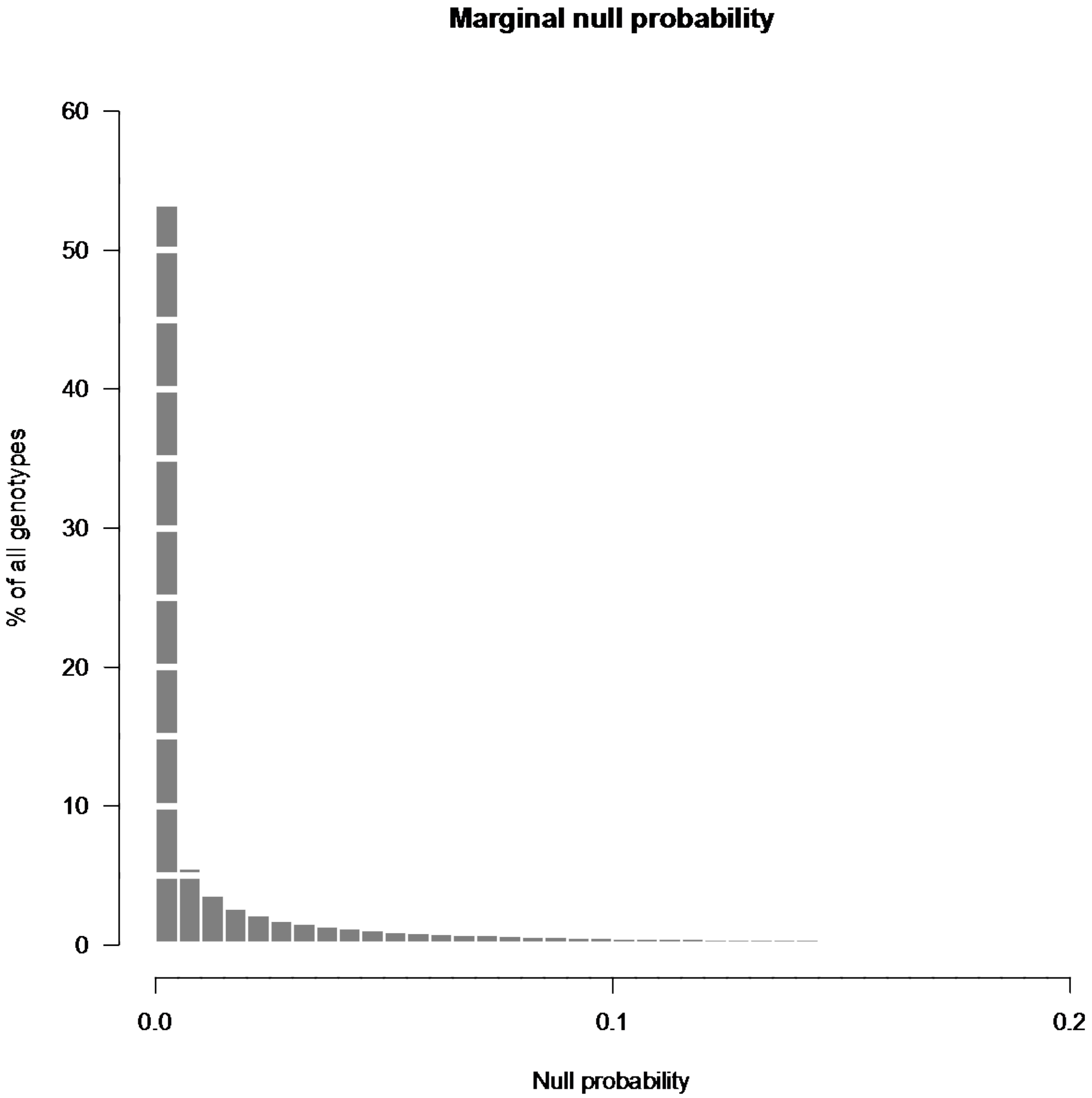


Figure 6.33: Marginal null probabilities for all SSC genotypes at all well-covered loci. The median marginal null probability for all genotypes is 0.003, and 80% of all genotypes called in this study had a marginal null genotype <0.10. This is consistent with the observation that most genotypes appear to be bi-allelic.

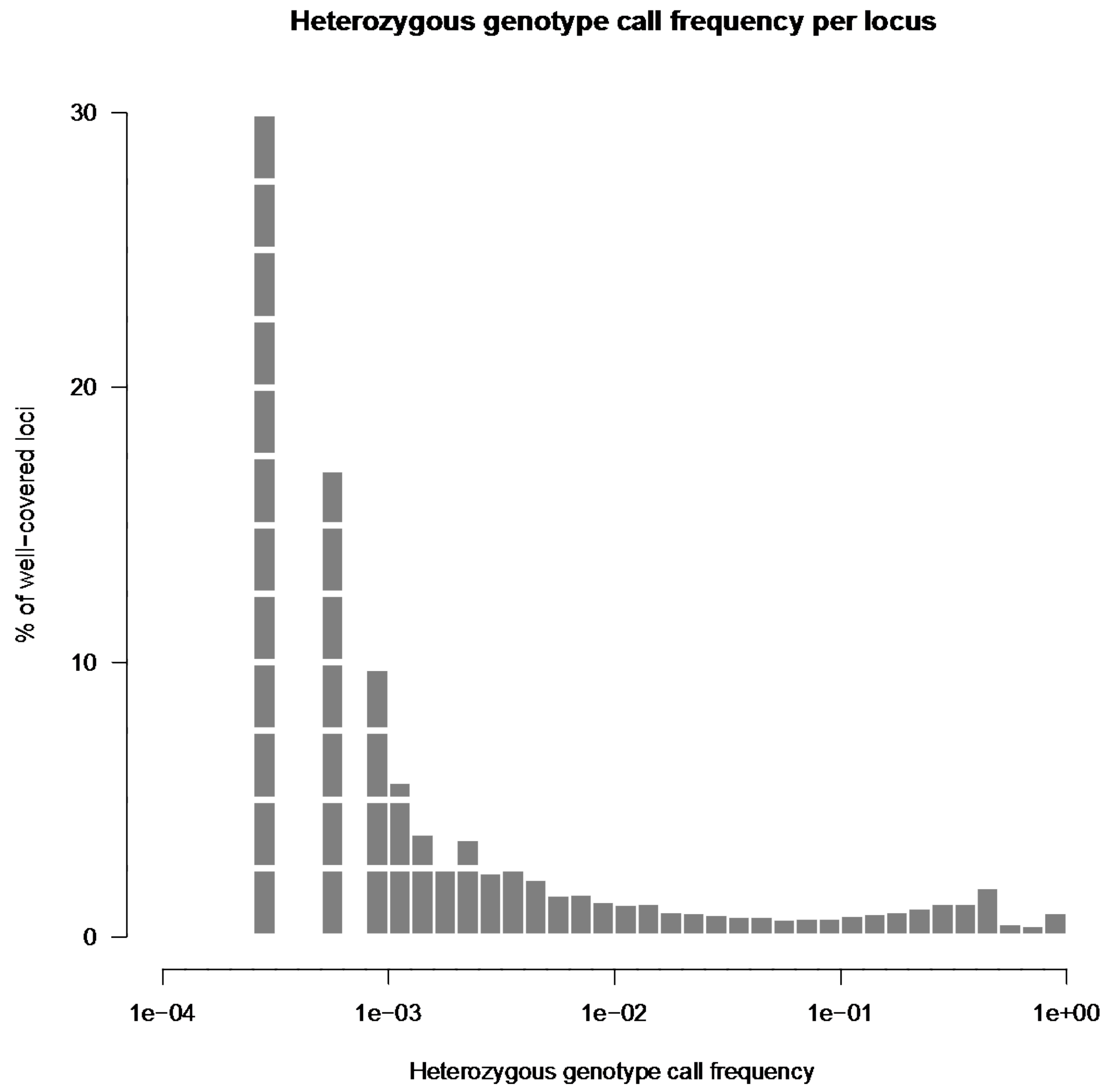


Figure 6.34: The frequency of heterozygous SSC genotype calls for each well-covered locus. Heterozygous genotype call frequency is calculated for non-null genotypes only. Approximately 32% of all loci in this study have at least one individual with a heterozygous genotype call. 75% of all heterozygous loci have heterozygous genotype calls in ≤10 SSC individuals.

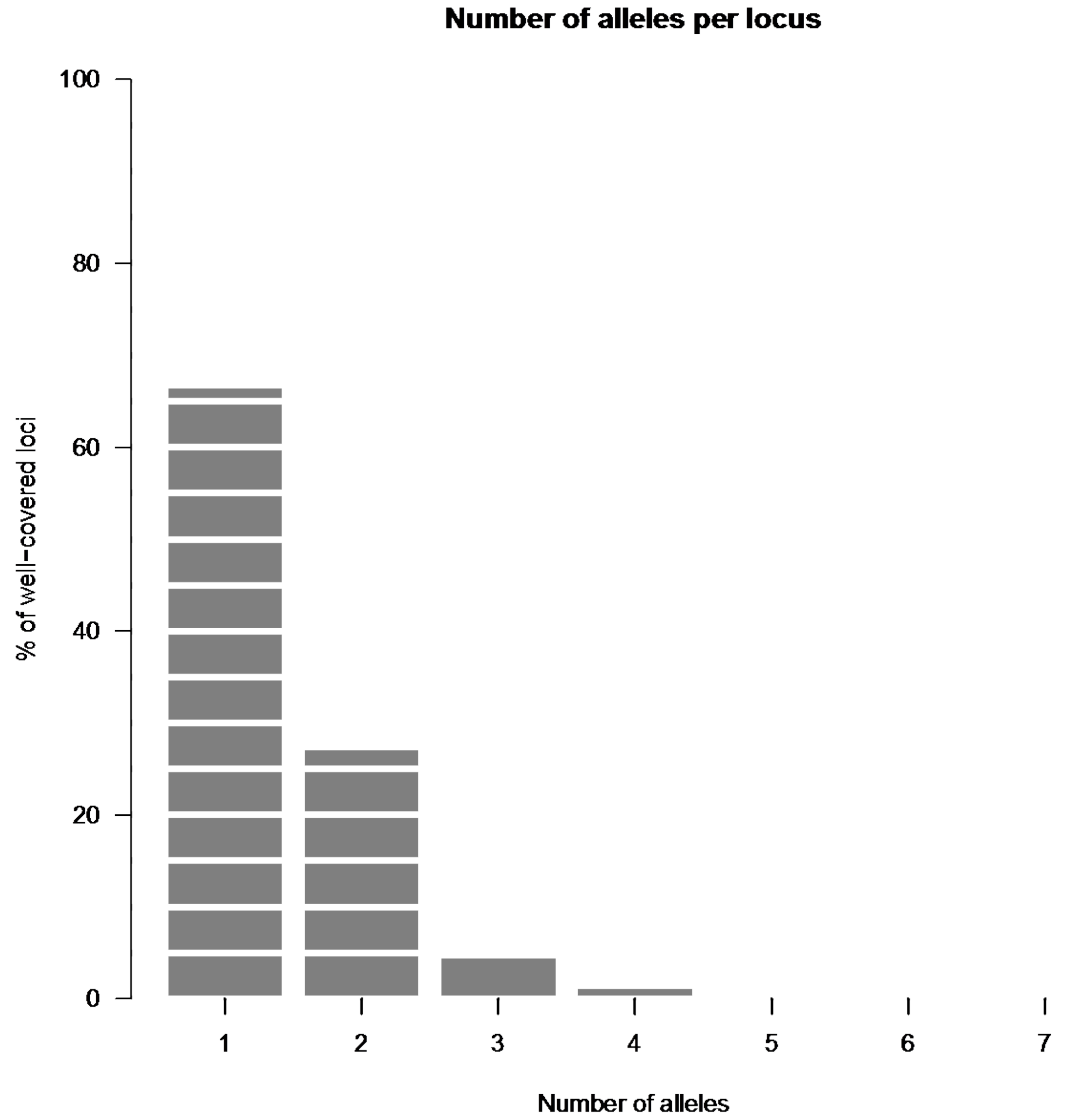


Figure 6.35: Number of alleles called in genotypes at all well-covered loci. Alleles called as part of single null genotypes are included. Nearly 34% of all well-covered loci have at least two genotyped alleles. 6% of all well-covered loci have ≥3 alleles, suggesting a large degree of polymorphism in many well-covered loci analyzed in this study.

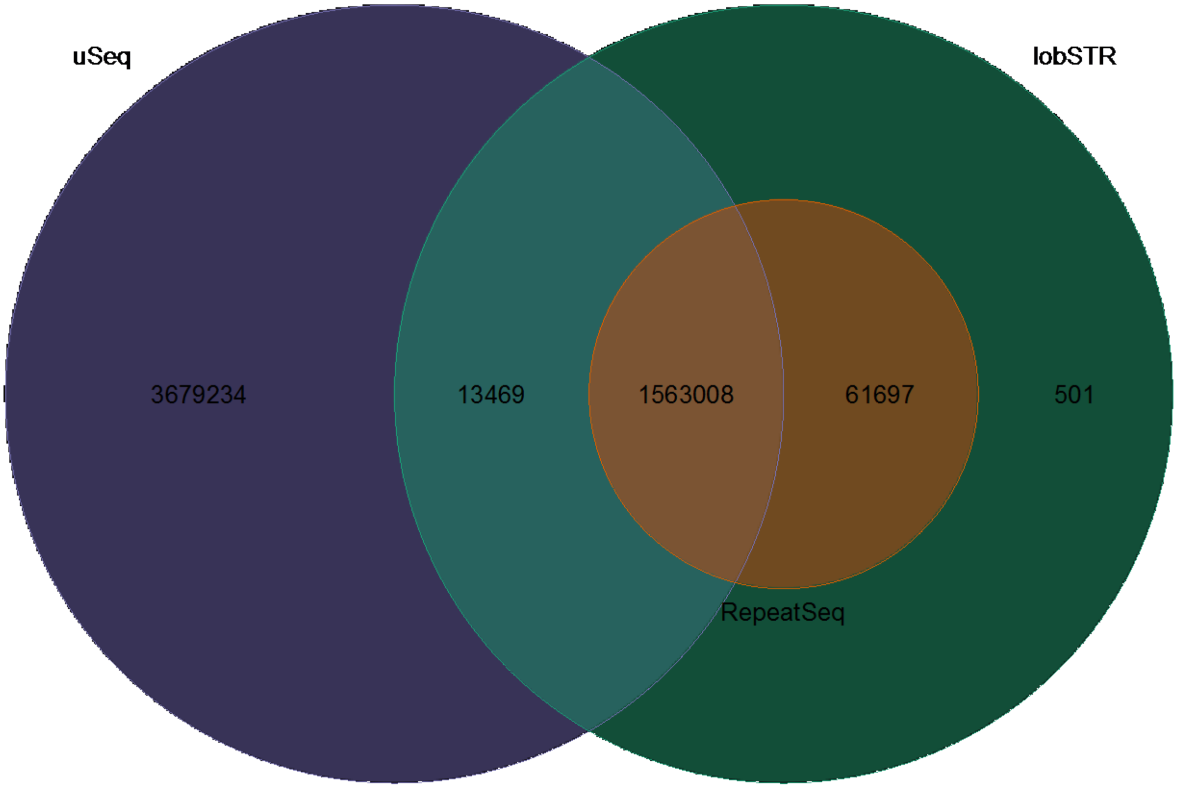


Figure 6.36: Venn diagram of the reference loci for uSeq, lobSTR, and RepeatSeq. Intersections between all reference microsatellite sets were determined using BedTools (Quinlan and Hall 2010). Of the 1.63 million reference microsatellite in lobSTR and/or RepeatSeq, >96% overlap at least one uSeq reference microsatellite locus. All 62,198 loci that are unique to RepeatSeq and/or lobSTR are interrupted. uSeq uniquely detects >3.6 million uninterrupted microsatellite loci.

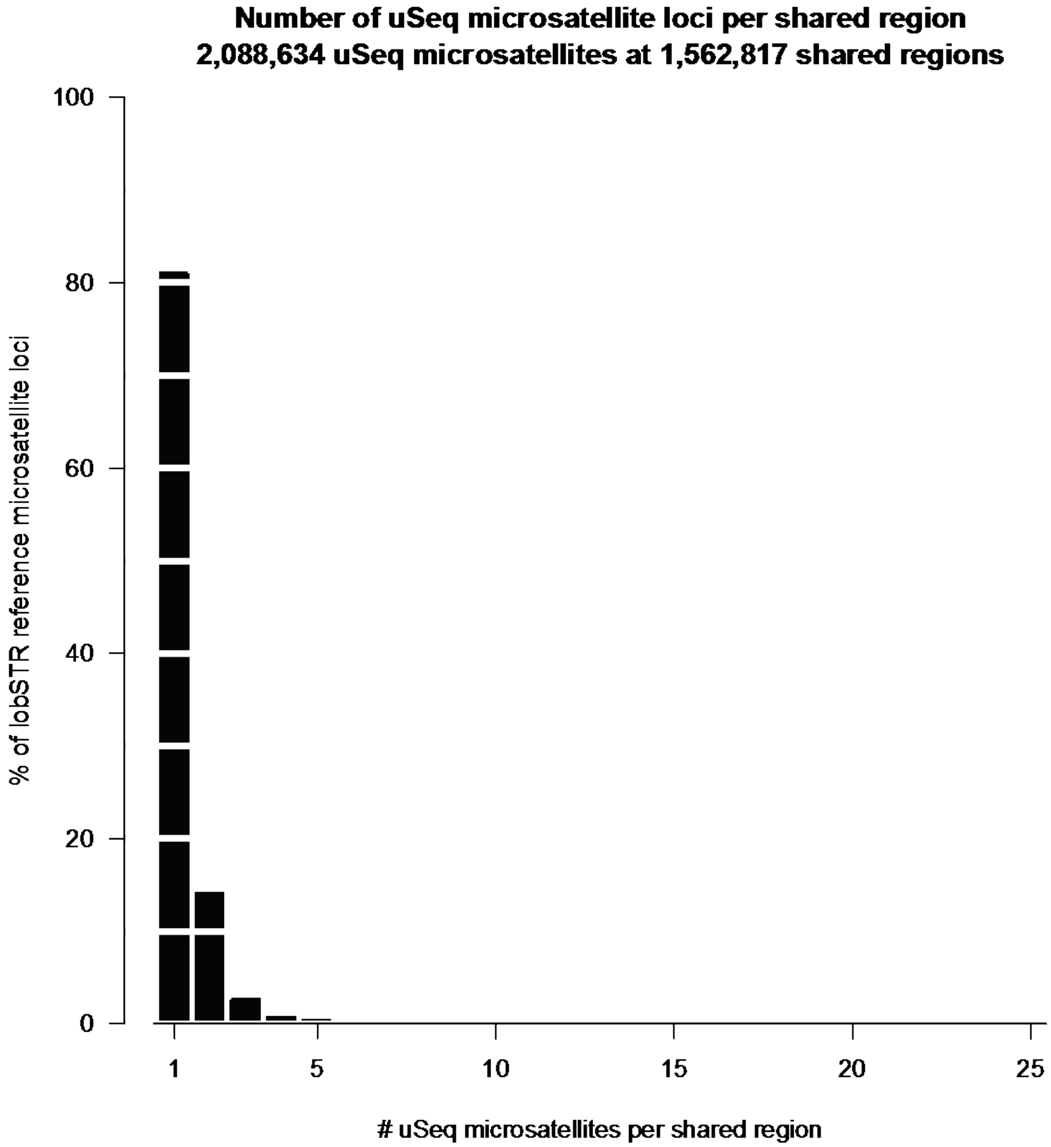


Figure 6.37: The number of uSeq reference microsatellite loci contained in the corresponding lobSTR/RepeatSeq reference microsatellite locus. The vast majority of larger interrupted microsatellite tracts detected by lobSTR have at ≥1 uninterrupted core microsatellite that can be detected by uSeq.

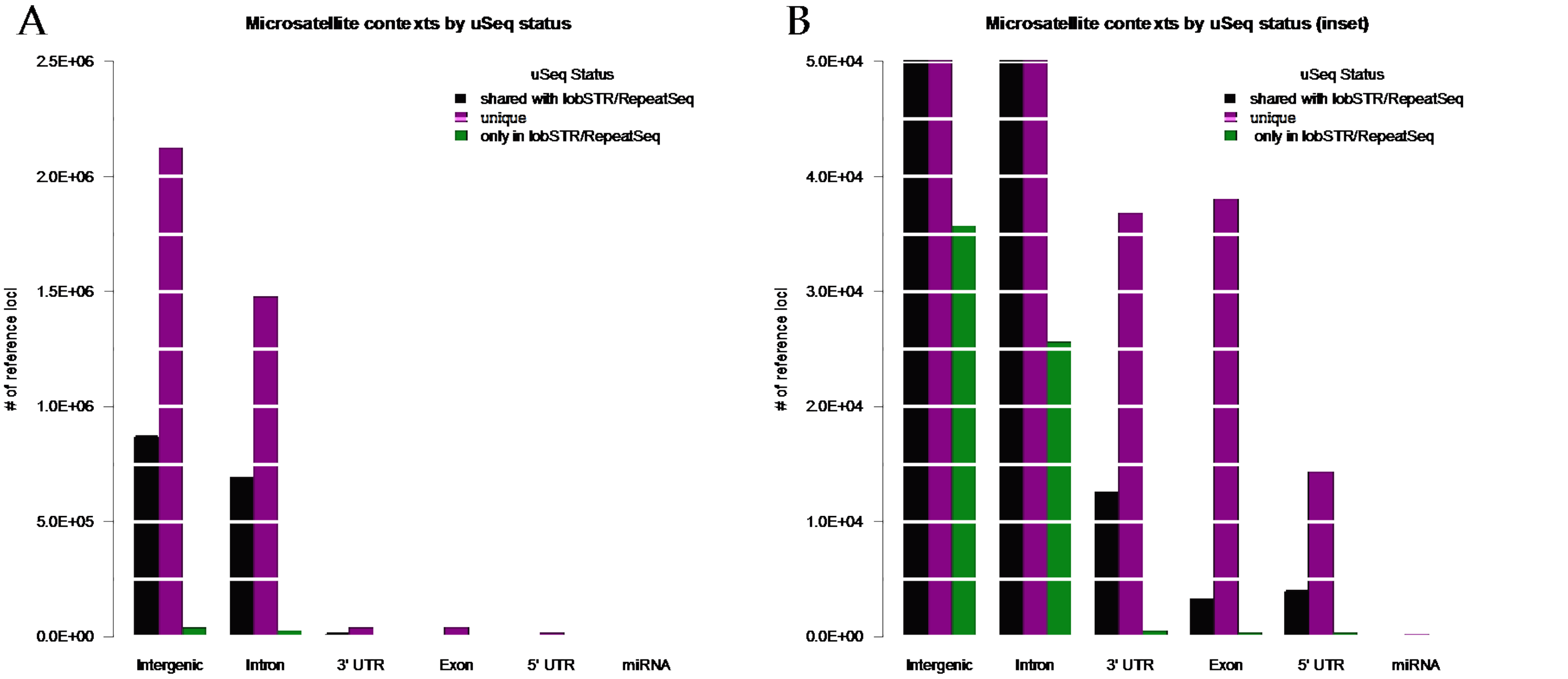


Figure 6.38: Pipeline loci by context. Loci are shown colored according to their uSeq status. (**A**) shows the number of loci across all contexts, (**B**) is an inset of (**A**) to visualize lobSTR/RepeatSeq-specific loci and unique uSeq contributions to exons and UTRs. The vast majority of reference microsatellite loci in all pipelines are in intron and intergenic regions, as are most loci unique to either uSeq or lobSTR. uSeq detects almost 38,000 microsatellite reference loci in exons that are not detected by lobSTR or RepeatSeq.

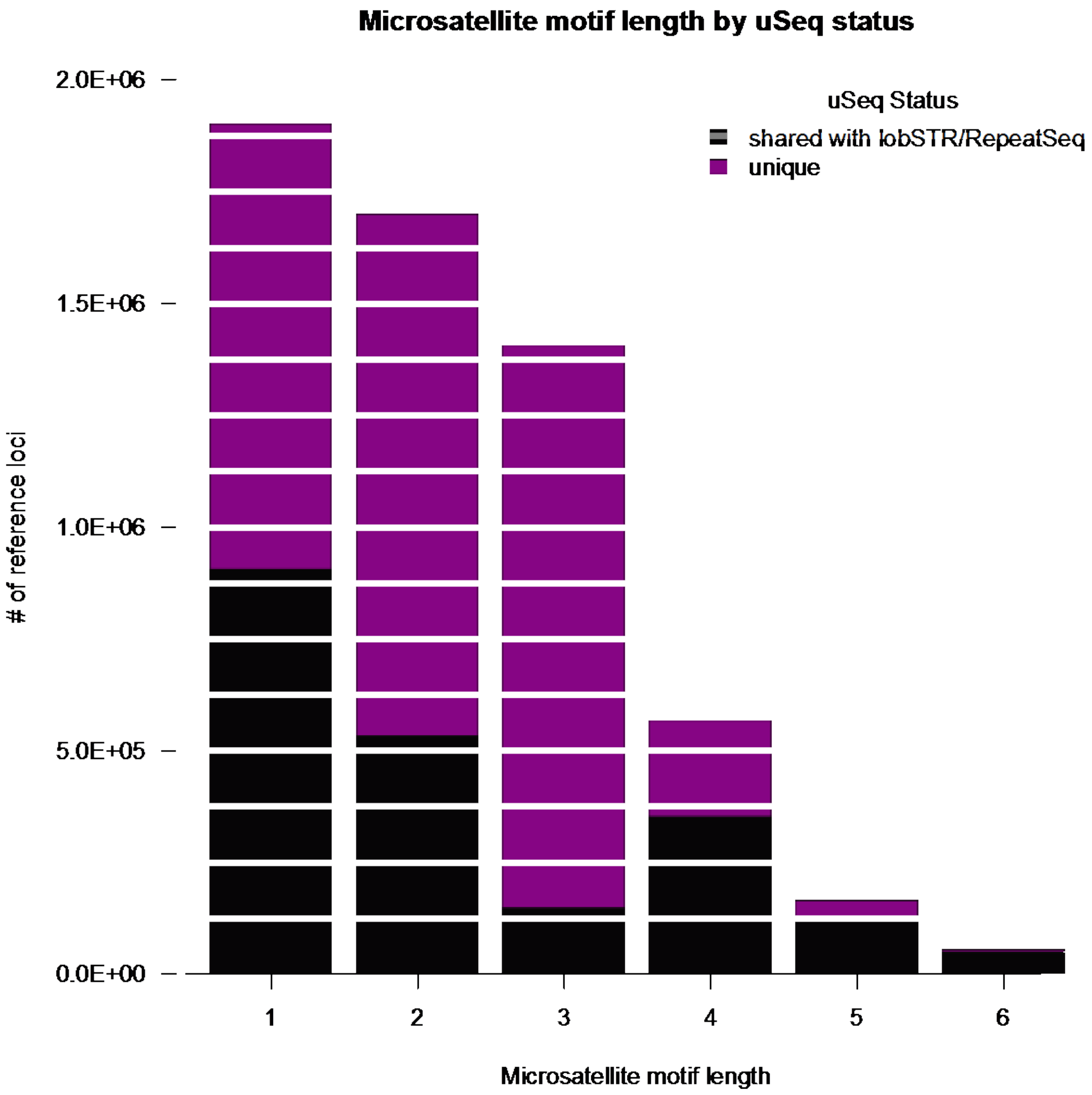


Figure 6.39: Motif lengths of reference microsatellite loci by uSeq status. Most of the reference microsatellite loci unique to uSeq have motif lengths ≤3 bp.

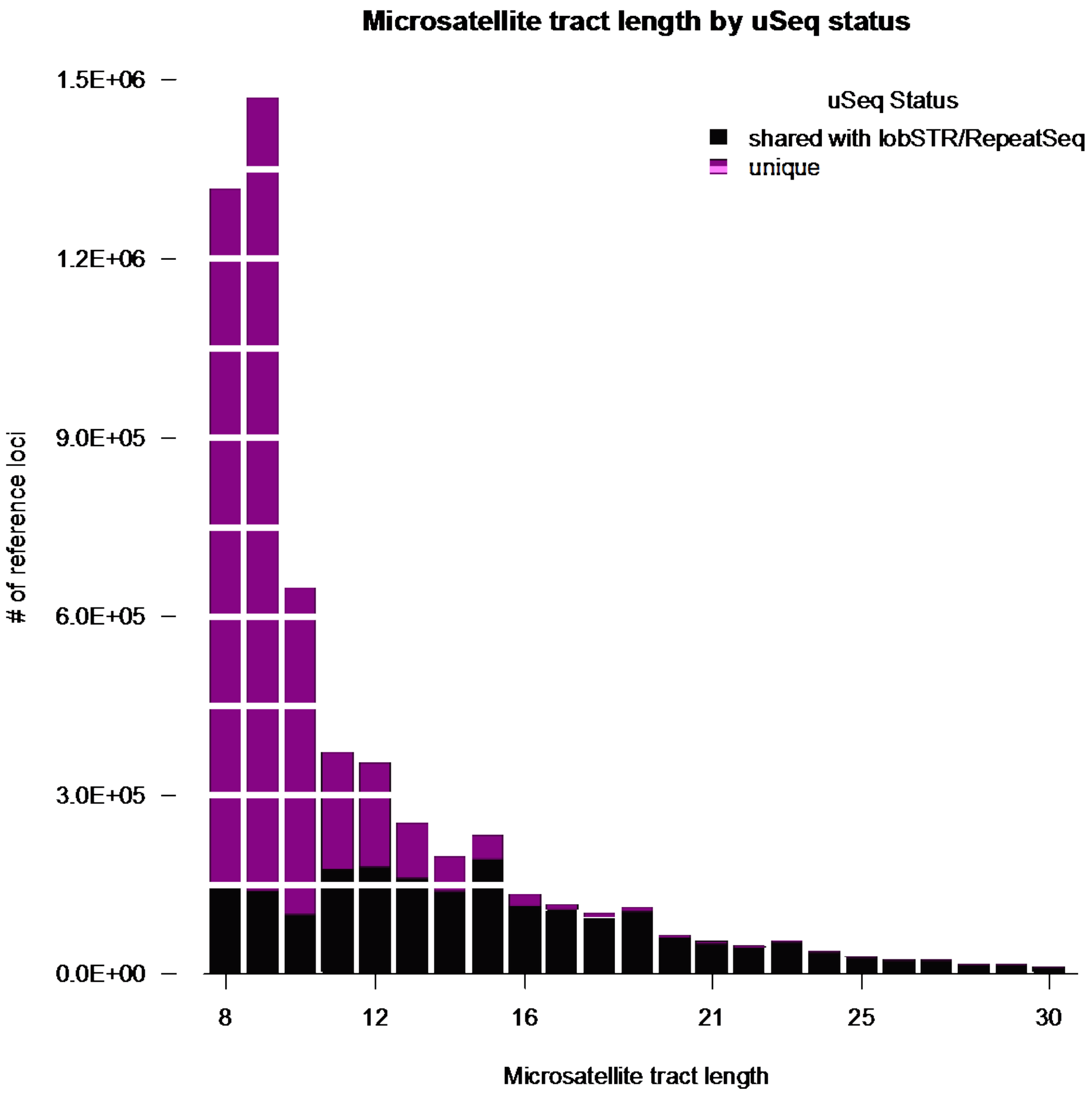


Figure 6.40: Tract lengths of reference microsatellite loci by uSeq status. Over 99% of reference microsatellite loci detected uniquely by uSeq have tract lengths ≤18 bp. Short microsatellite tracts are detected with greater consistency by uSeq than by lobSTR or RepeatSeq.

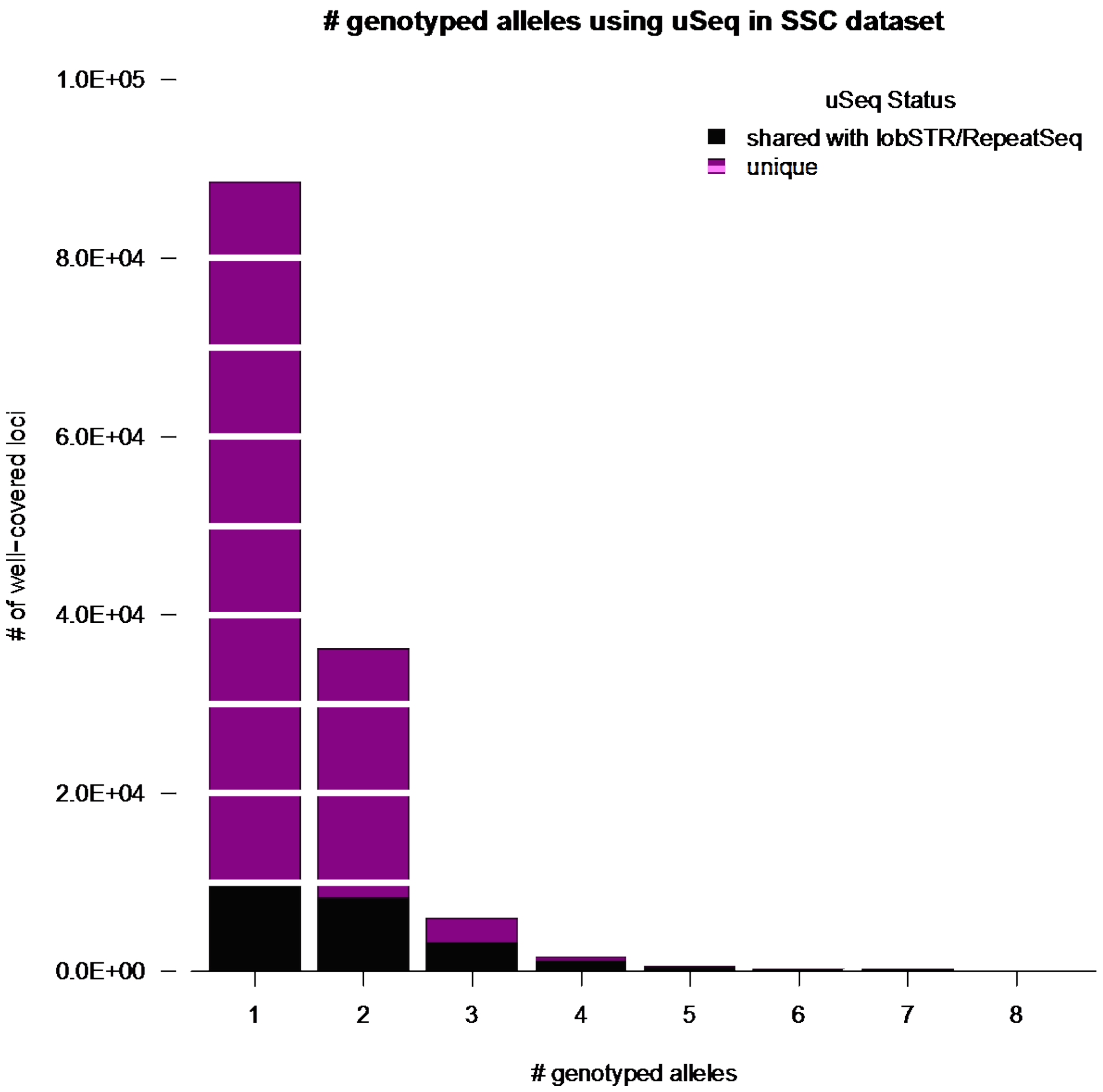
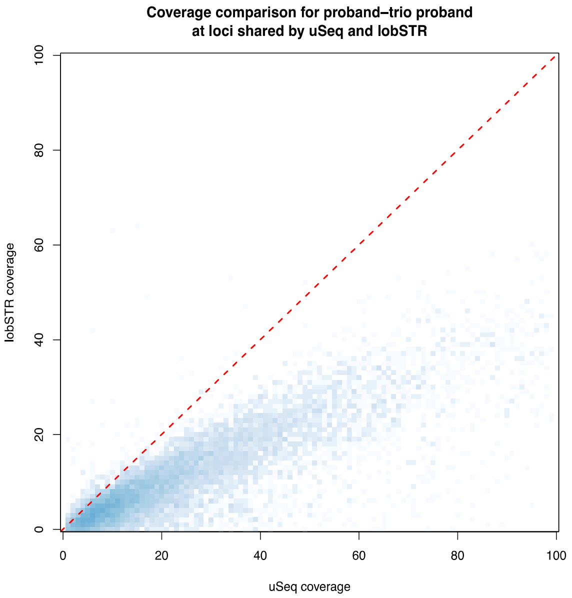
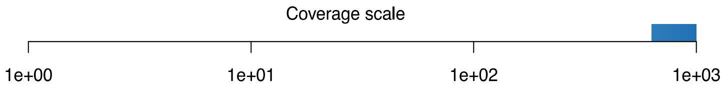


Figure 6.41: Locus polymorphism at reference microsatellite loci by uSeq status. Number of alleles is calculated as the number of non-null alleles called by the uSeq genotyper at each locus. In approximately 60% of loci detected by all three pipelines, uSeq called >1 allele in the SSC dataset. The SSC dataset had >1 allele called by uSeq in another 30% of loci unique to the uSeq reference microsatellite set. This suggests that uSeq is capable of detecting extensive polymorphism that would not be observed in other pipelines.

Figure 6.42: Pairwise pipeline coverage comparison in an SSC individual. Coverage was compared for all loci genotyped by both pipelines for each pairwise comparison. The red dashed line in each graph is the x = y line. Two pipelines with similar coverage will have most of their points centered on this line. Part (**A**) shows pairwise coverage for uSeq and lobSTR; (**B**) shows pairwise coverage for uSeq and RepeatSeq; and (**C**) shows coverage for lobSTR and RepeatSeq. The coverage scale is the same for all plots. Although only one individual is shown here, the trends are consistent for all individuals. Coverage for lobSTR is modified as described in the text. Among all three pipelines, both uSeq and RepeatSeq have consistently higher coverage than lobSTR at shared loci. RepeatSeq has comparable coverage to uSeq.



Coverage comparison for proband–trio proband
at loci shared by uSeq and lobSTR



Coverage scale

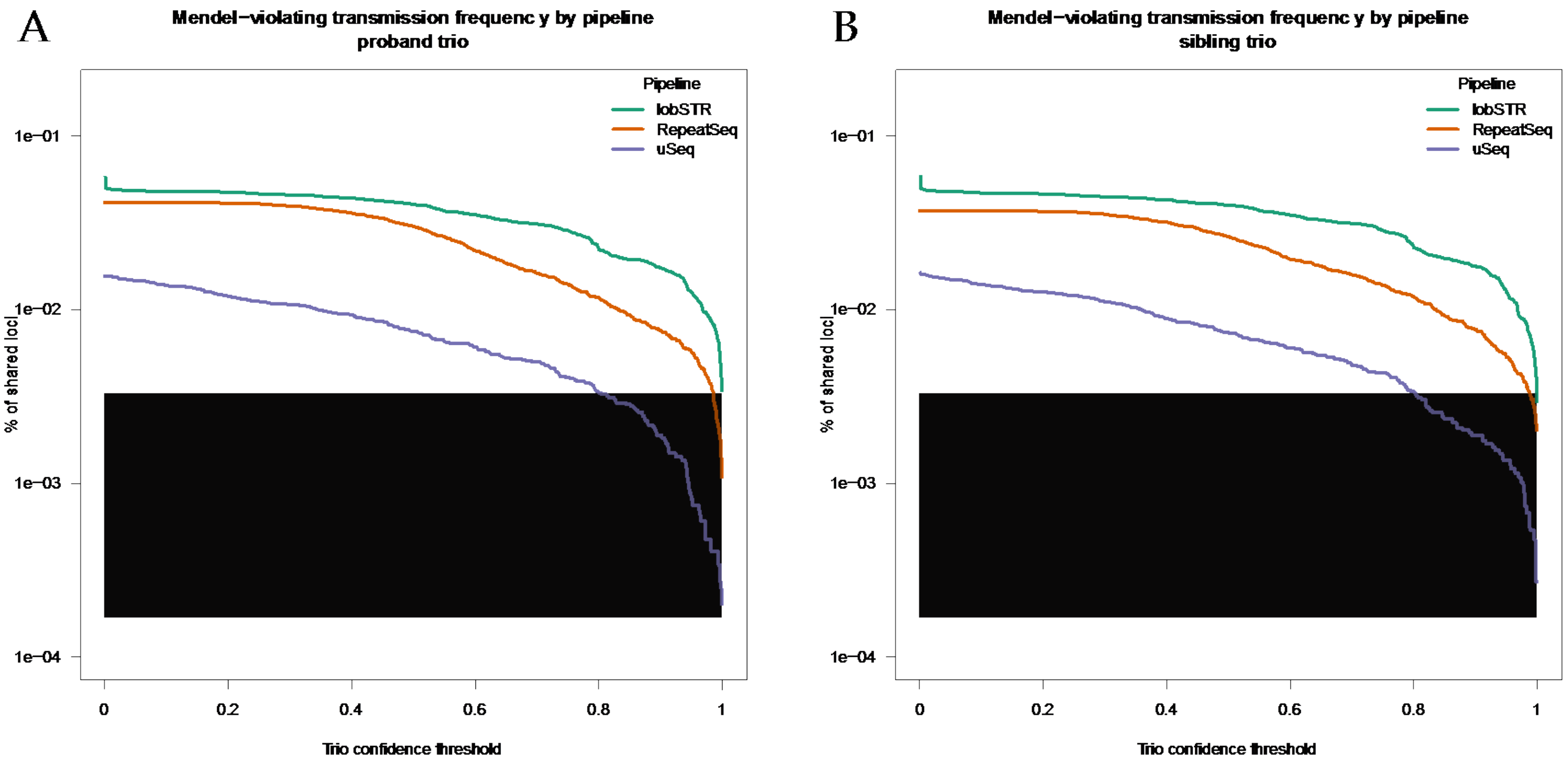


Figure 6.43: Mendel violation frequency for lobSTR, RepeatSeq, and uSeq pipelines in sample trios. Part (**A**) shows the Mendel violation frequency for the proband trio, (**B**) shows the Mendel violation frequency for the sibling trio. The shaded grey box shows the range of literature-estimated per-generation, per-gamete microsatellite mutation rates, as discussed in Introduction section 1.2.4. Trio genotype confidence is the product of each trio member's individual genotype confidence. uSeq has consistently lower Mendel violation rates than lobSTR or RepeatSeq, suggesting it has higher genotyping accuracy. The literature-reported microsatellite mutation rates are based on highly unstable and polymorphic marker microsatellite loci, and are likely an overestimate for most microsatellite loci in the genome. Therefore, Mendel violations at shared loci called by any pipeline are likely to be false positives.

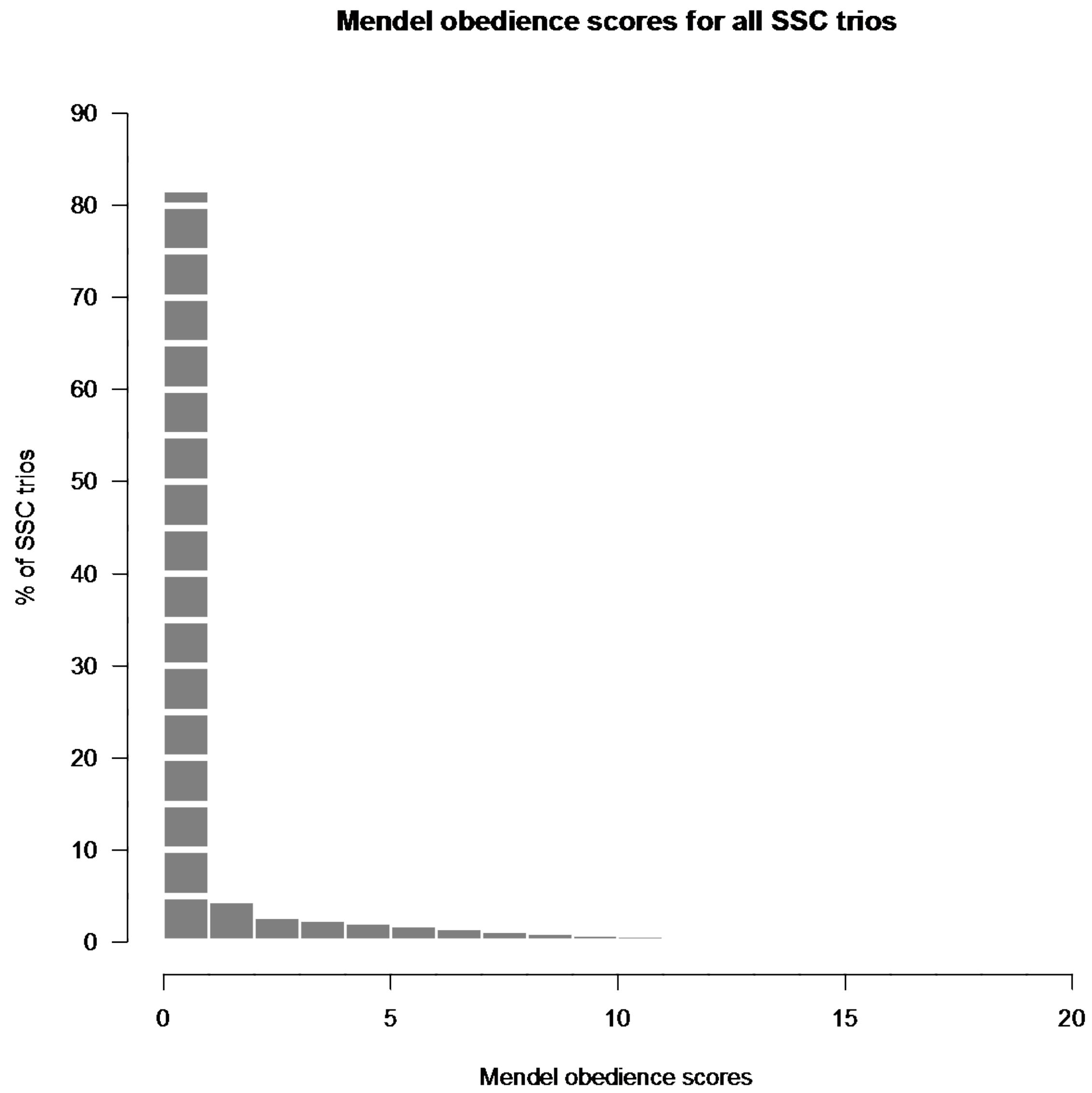


Figure 6.44: Mendel obedience scores for all SSC trios. Most trios have Mendel obedience scores ≤1, and >99.9% of all trios have Mendel obedience scores <40. This indicates that most trios genotyped at well-covered loci in the SSC dataset have inheritance patterns consistent with Mendelian inheritance.

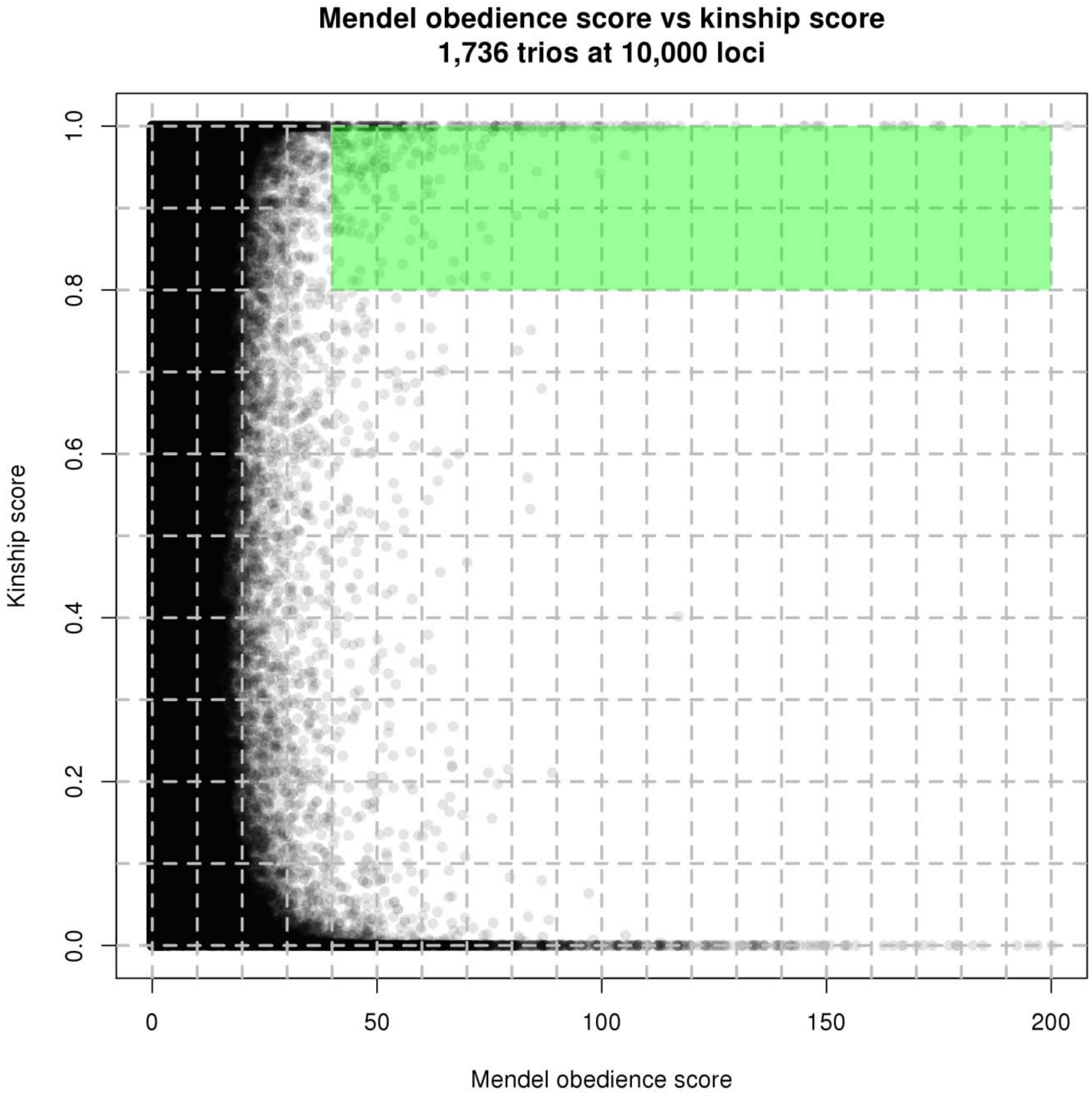


Figure 6.45: Comparing Mendel obedience scores and kinship scores for a subset of SSC trios. The green box represents the kinship and Mendel obedience thresholds used as thresholds for *de novo* microsatellite mutations. The minimum Mendel obedience score considered was 40, and the minimum kinship score was 0.80. We chose these threshold to limit our analysis to trios that had a high likelihood of violating Mendel obedience while also minimizing calls that have low kinship scores, which would suggest a likely false positive.

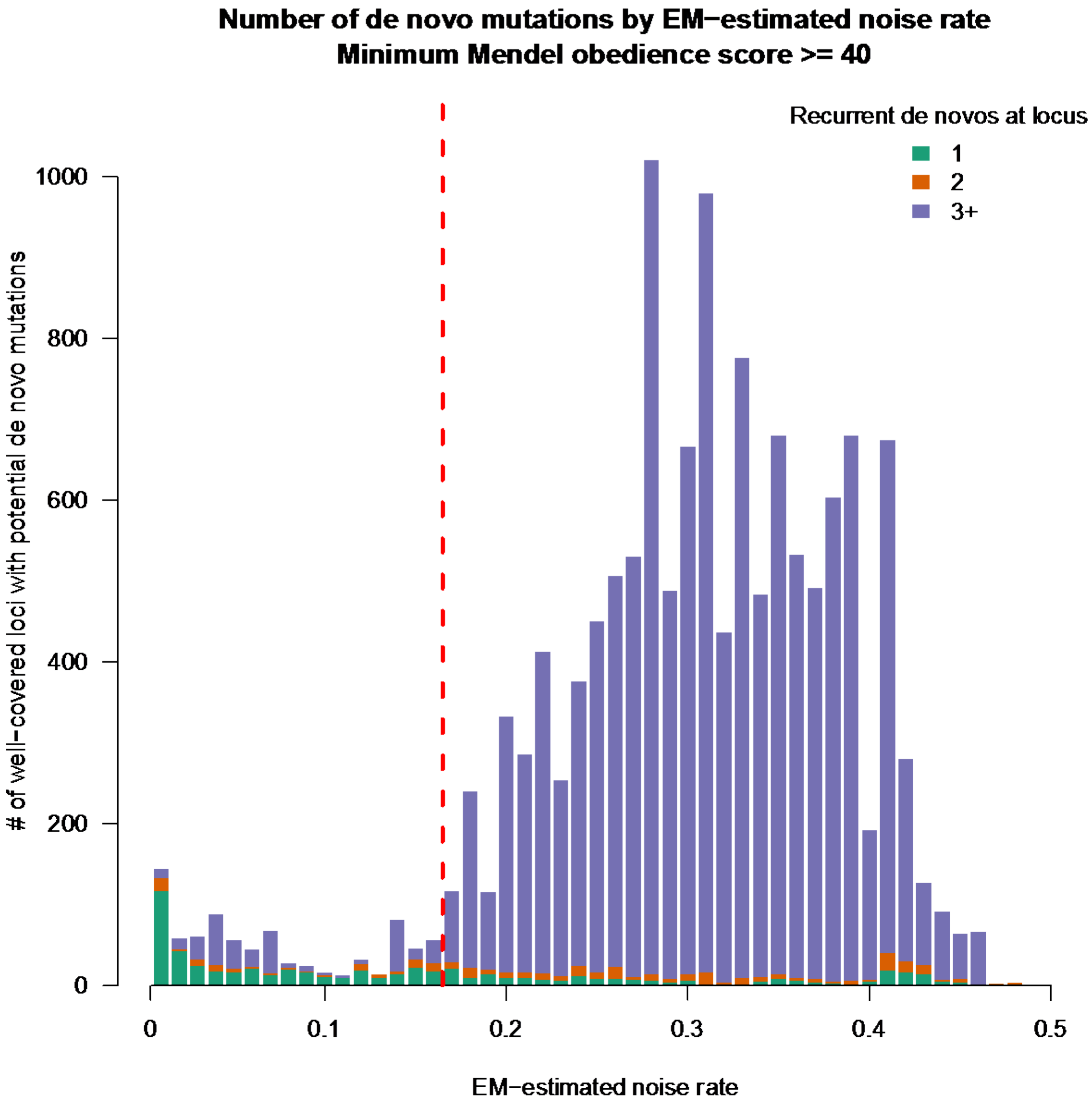


Figure 6.46: Recurrent *de novo* mutations are common when EM-estimated per-locus error rate is high. The red line is the maximum error rate used in our analysis of Mendel violations, which is set to 0.17. Any trio with a Mendel obedience score ≥40 and a kinship score ≥0.8 is considered for this analysis. As the estimated noise rate increases, recurrent *de novo* mutations at the same locus (in orange and blue) become more common. Recurrent *de novo* mutations at a locus are more likely to be false positives than truly independent mutation events.

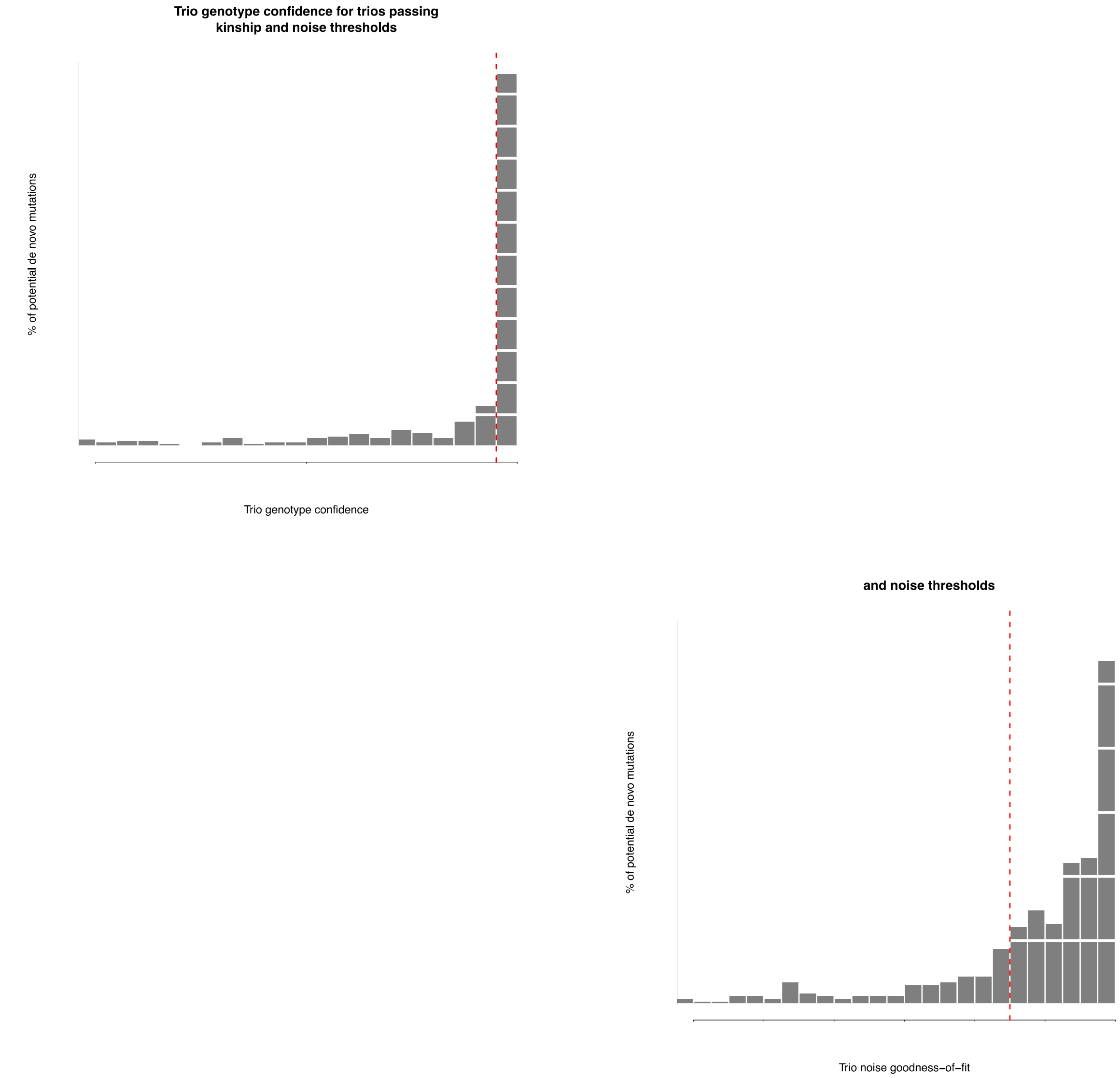


Figure 6.47: Trio genotype quality metrics used to filter *de novo* mutations. Trio genotype metrics were (**A**) trio genotype confidence, (**B**) trio null probability, (**C**) trio allelic goodness-of-fit, and (**D**) trio noise goodness-of-fit. The minimum trio genotype confidence considered was 0.99, minimum allelic-goodness-of-fit was 0.00001, minimum noise goodness-of-fit was 0.001, and maximum marginal null probability was 0.01. Before being filtered for genotype quality, potential de novo mutations had to have a Mendel obedience score ≥40, a kinship score ≥0.80, and an EM-estimated locus error rate ≤0.17.

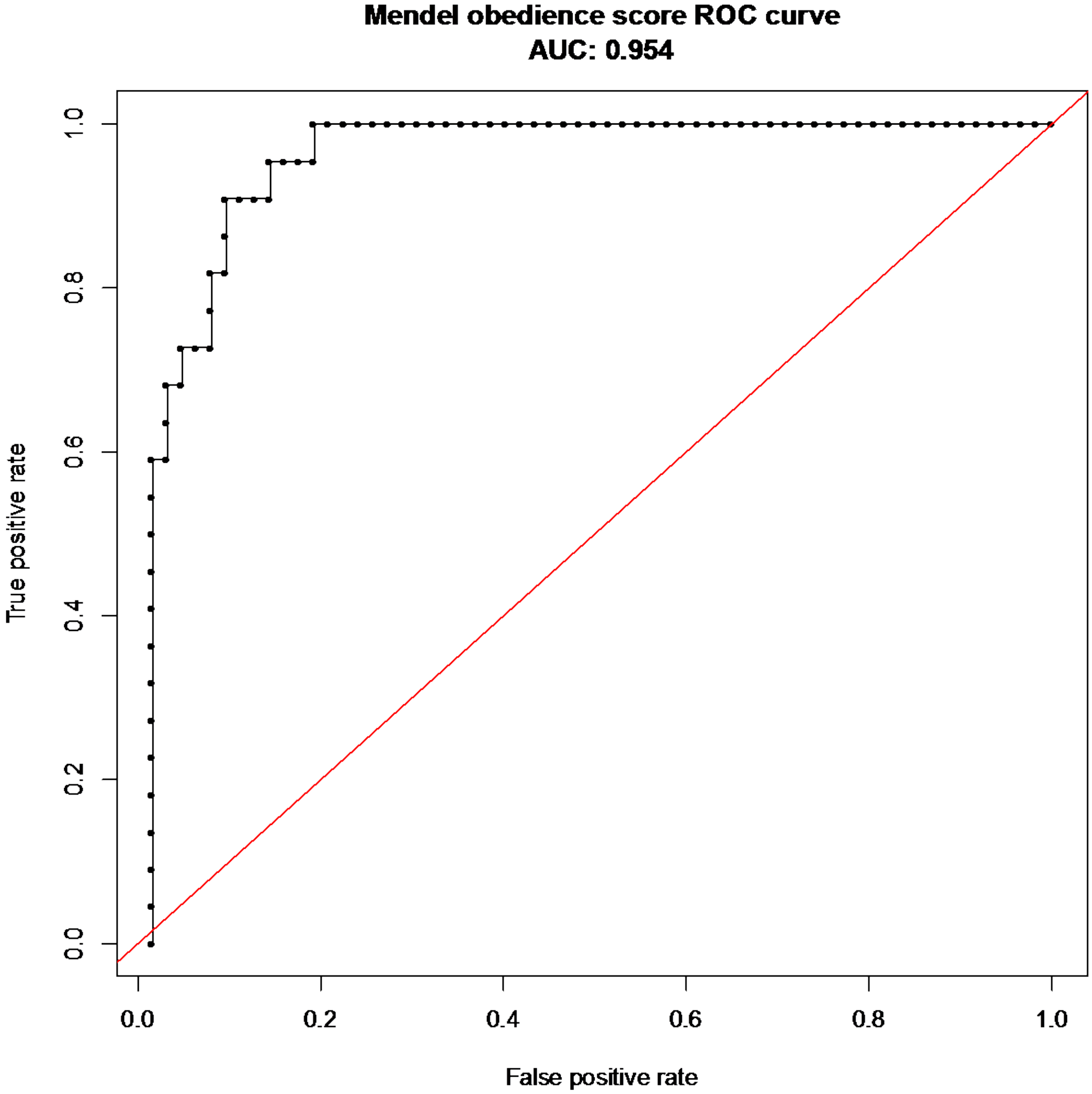
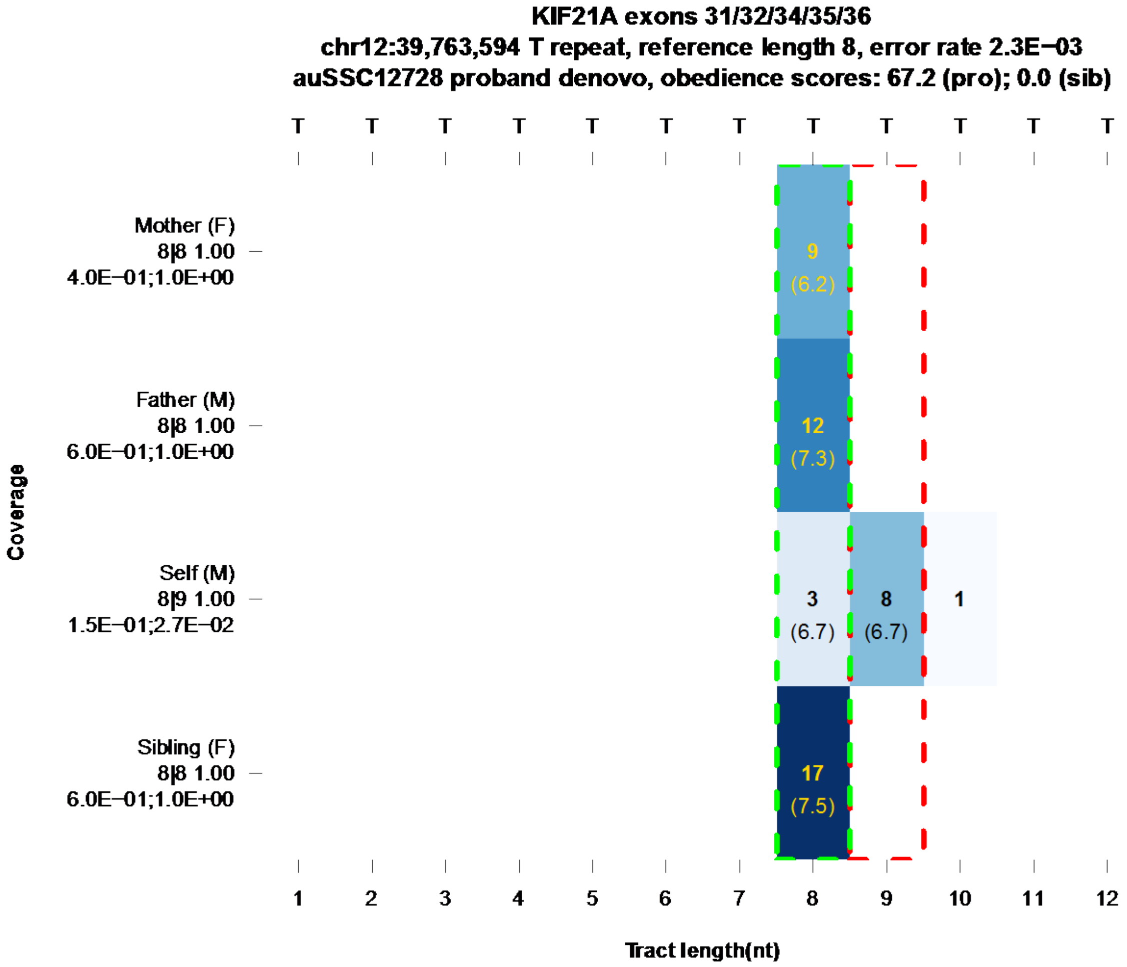


Figure 6.48: The Mendel obedience score is a very robust means of identifying true *de novo* microsatellite mutations. The red line is marks where the true positive and false positive rates are equivalent, which would provide a ROC AUC of 0.5. A ROC AUC of 0.5 would imply that a score is incapable of distinguishing true and false positives. A ROC AUC close to 1 indicates that the Mendel obedience score is able to robustly distinguish true and false positives. The ROC AUC for the Mendel obedience score is >0.95, indicating that with the thresholds specified in the text, the Mendel obedience score is capable of distinguishing between true and false positive *de novo* mutations.

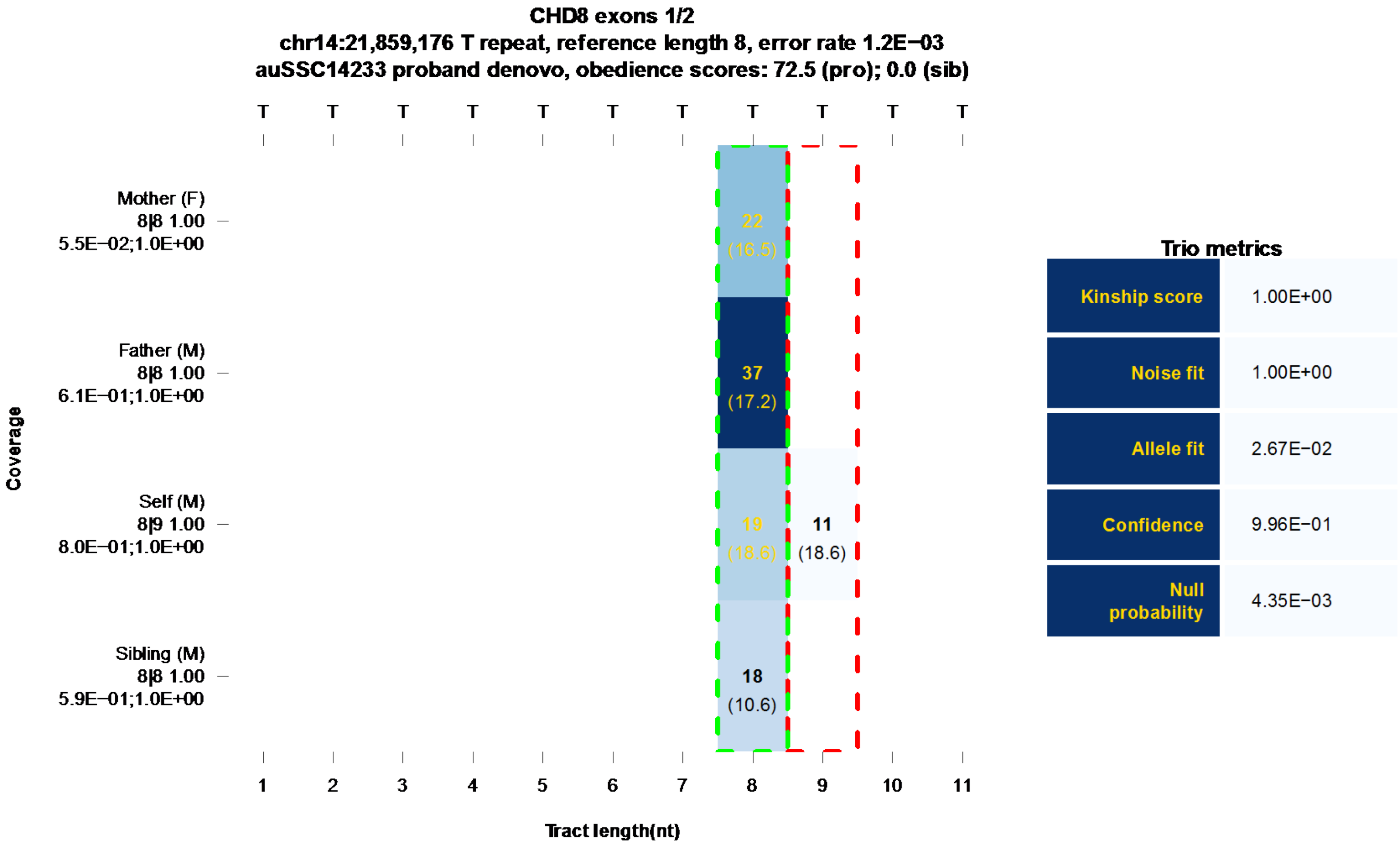
Figure 6.49: A validated *de novo* microsatellite mutation in a child with autism in an exon of KIF21A. The mutation is at a T microsatellite, so the mutation in the child with autism would produce a frameshift. The title provides the microsatellite locus context, gene name, location, noise rate, and family information. The labels on the y-axis correspond to each family member. The first line of the label defines the family member's relationship to the proband and their gender; the second line provides the member's genotype and confidence; and the last line provides the genotype allelic and noise goodness-of-fits. A green box surrounds the reference allele and a red box surrounds the novel allele. Within the table, each cell reports the observed coverage for a tract length in the corresponding individual. If the coverage corresponds to a genotyped allele, the unbiased expected coverage estimator is reported underneath the observed coverage in parentheses. A box with all trio metrics discussed in the manuscript is to the left of the coverage plot.



KIF21A exons 31/32/34/35/36
chr12:39,763,594 T repeat, reference length 8, error rate 2.3E−03
auSSC12728 proband denovo, obedience scores: 67.2 (pro); 0.0 (sib)

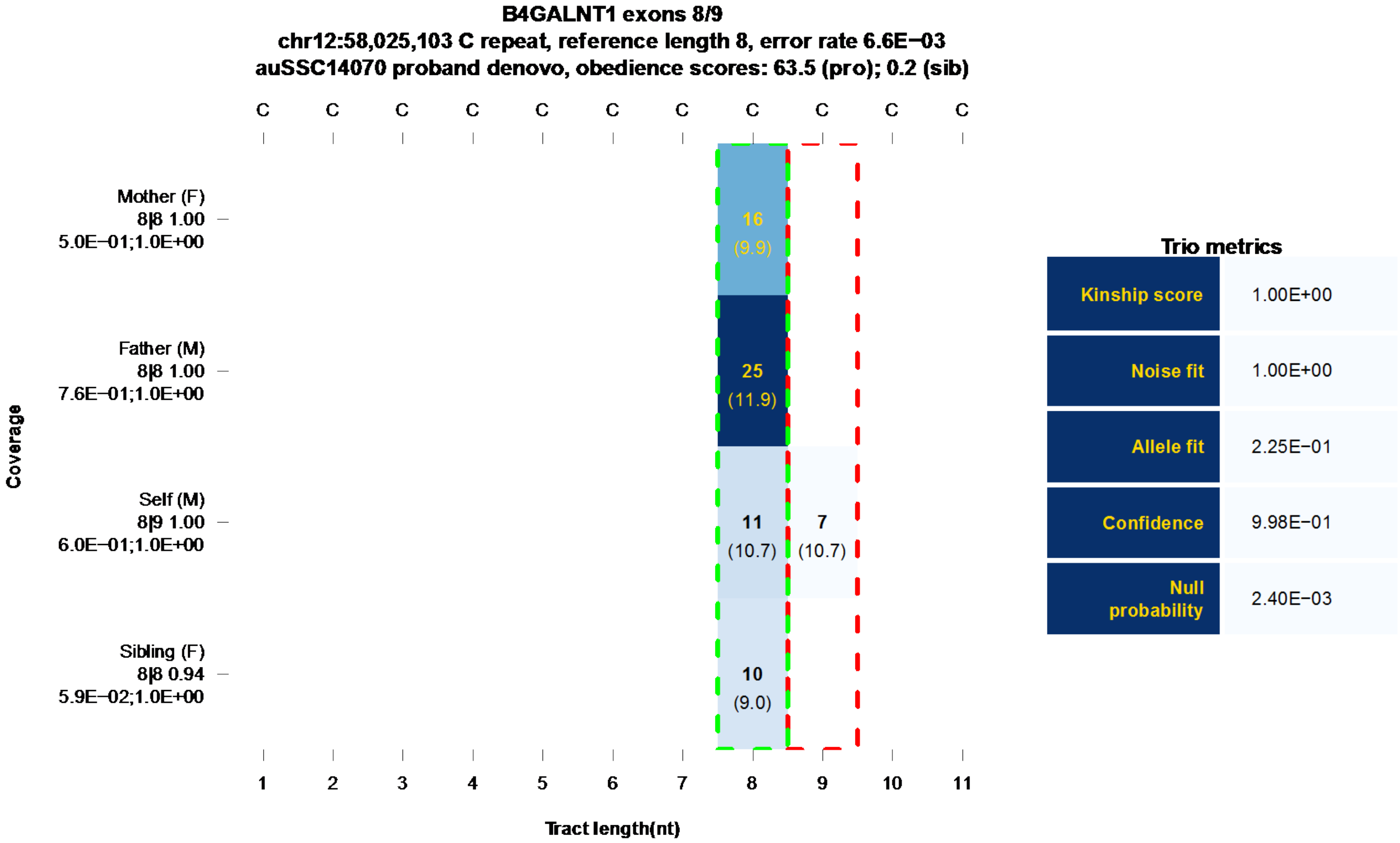
Trio metrics

Kinship score	9.99E−01
Noise fit	2.69E−02
Allele fit	3.62E−02
Confidence	9.96E−01
Null probability	4.20E−03



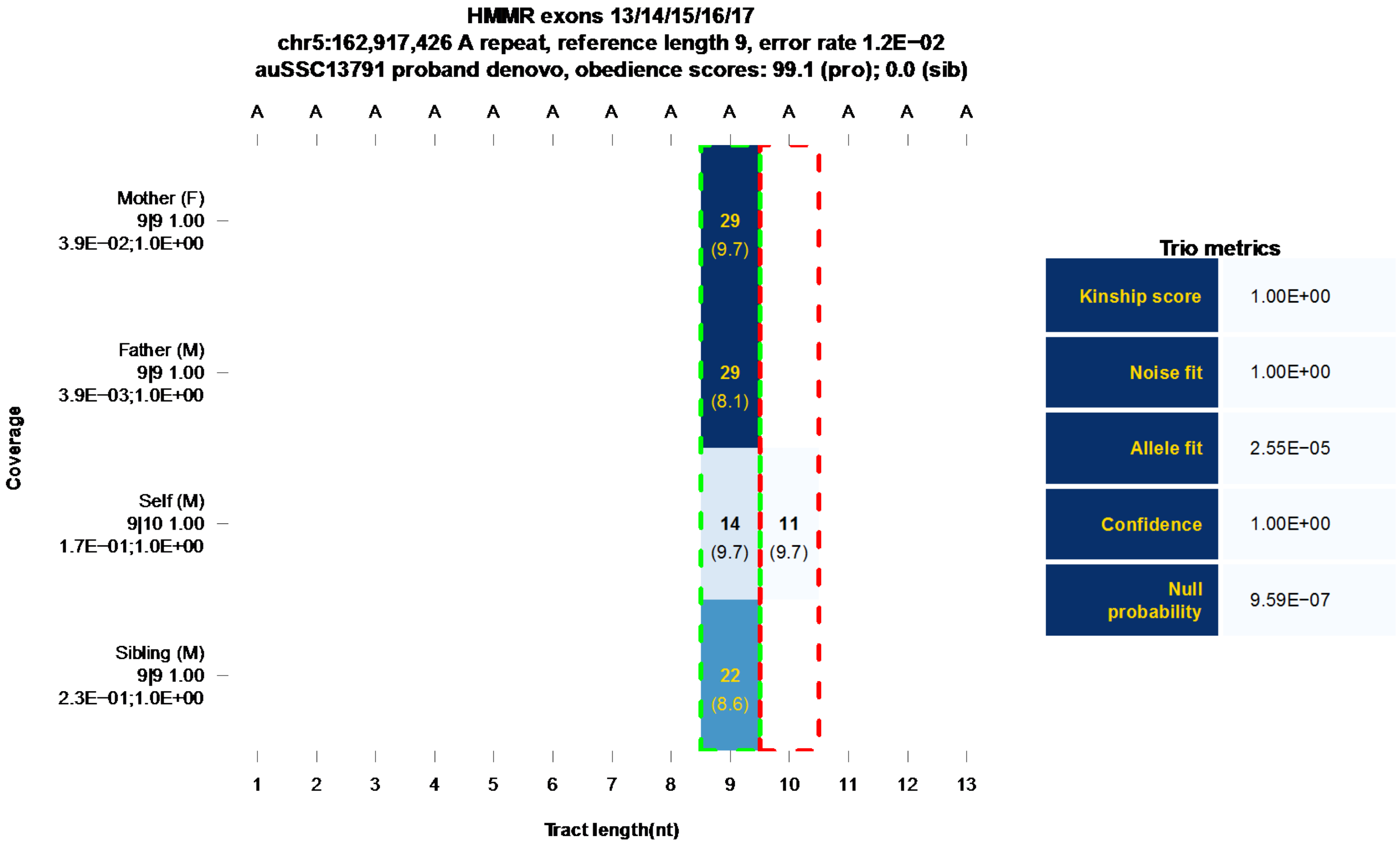
Trio metrics	
Kinship score	1.00E+00
Noise fit	1.00E+00
Allele fit	2.67E−02
Confidence	9.96E−01
Null probability	4.35E−03

Figure 6.50: A validated *de novo* microsatellite mutation in a child with autism in an exon of CHD8. This mutation introduces a frameshift mutation in a T microsatellite.



Kinship score	1.00E+00
Noise fit	1.00E+00
Allele fit	2.25E−01
Confidence	9.98E−01
Null probability	2.40E−03

Figure 6.51: A validated *de novo* microsatellite mutation in a child with autism in an exon of the B4GALNT1 gene. This mutation introduces a frameshift mutation in a C microsatellite.



Kinship score	1.00E+00
Noise fit	1.00E+00
Allele fit	2.55E−05
Confidence	1.00E+00
Null probability	9.59E−07

Figure 6.52: A *de novo* microsatellite mutation in a child with autism in an exon of the HMMR gene. This mutation introduces a frameshift mutation in an A microsatellite. This *de novo* mutation failed to validate.

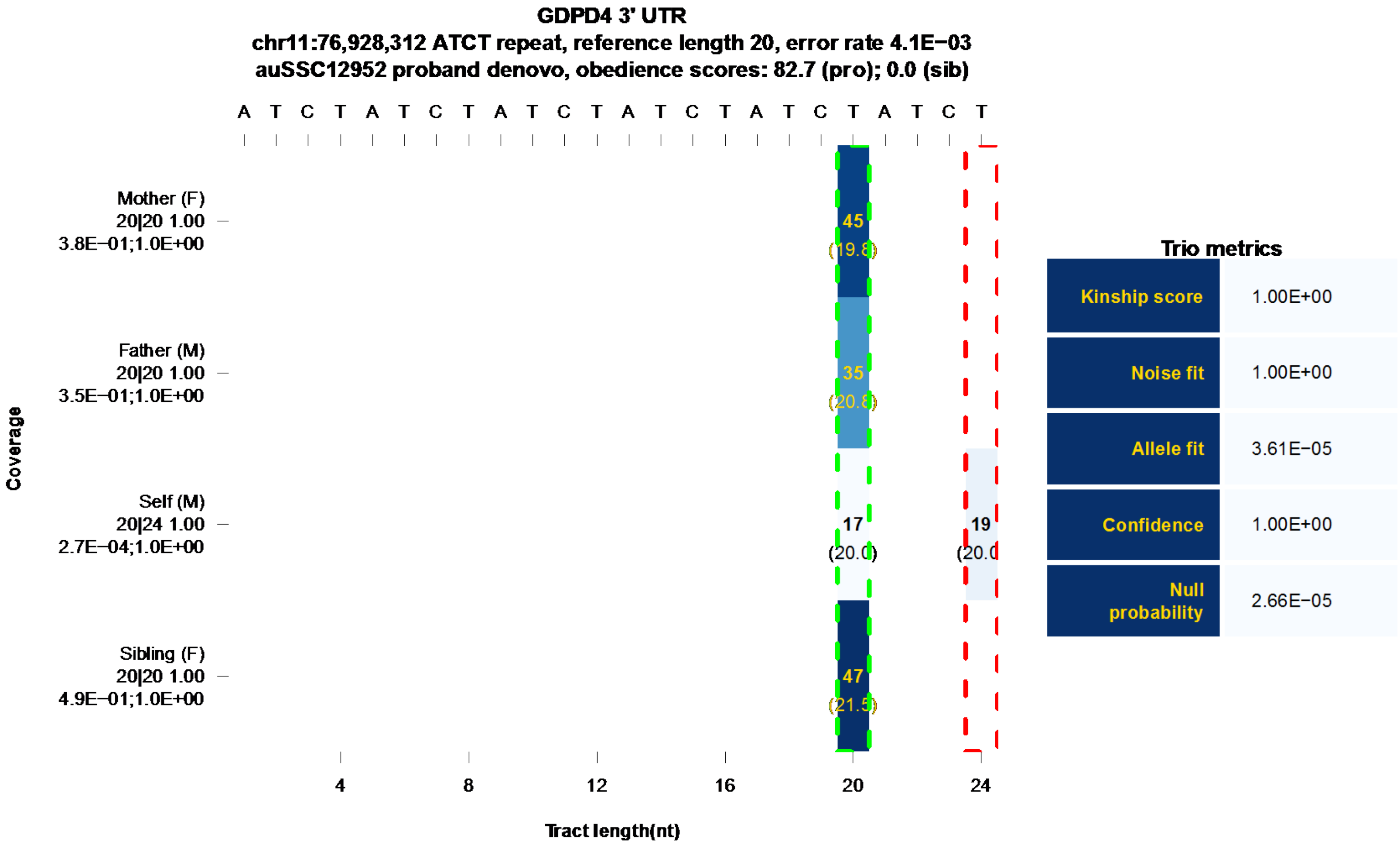


Figure 6.53: A validated *de novo* microsatellite mutation in a child with autism that spans the final exon and 3' UTR of the GDPD4 gene. This mutation might introduce a frameshift mutation.

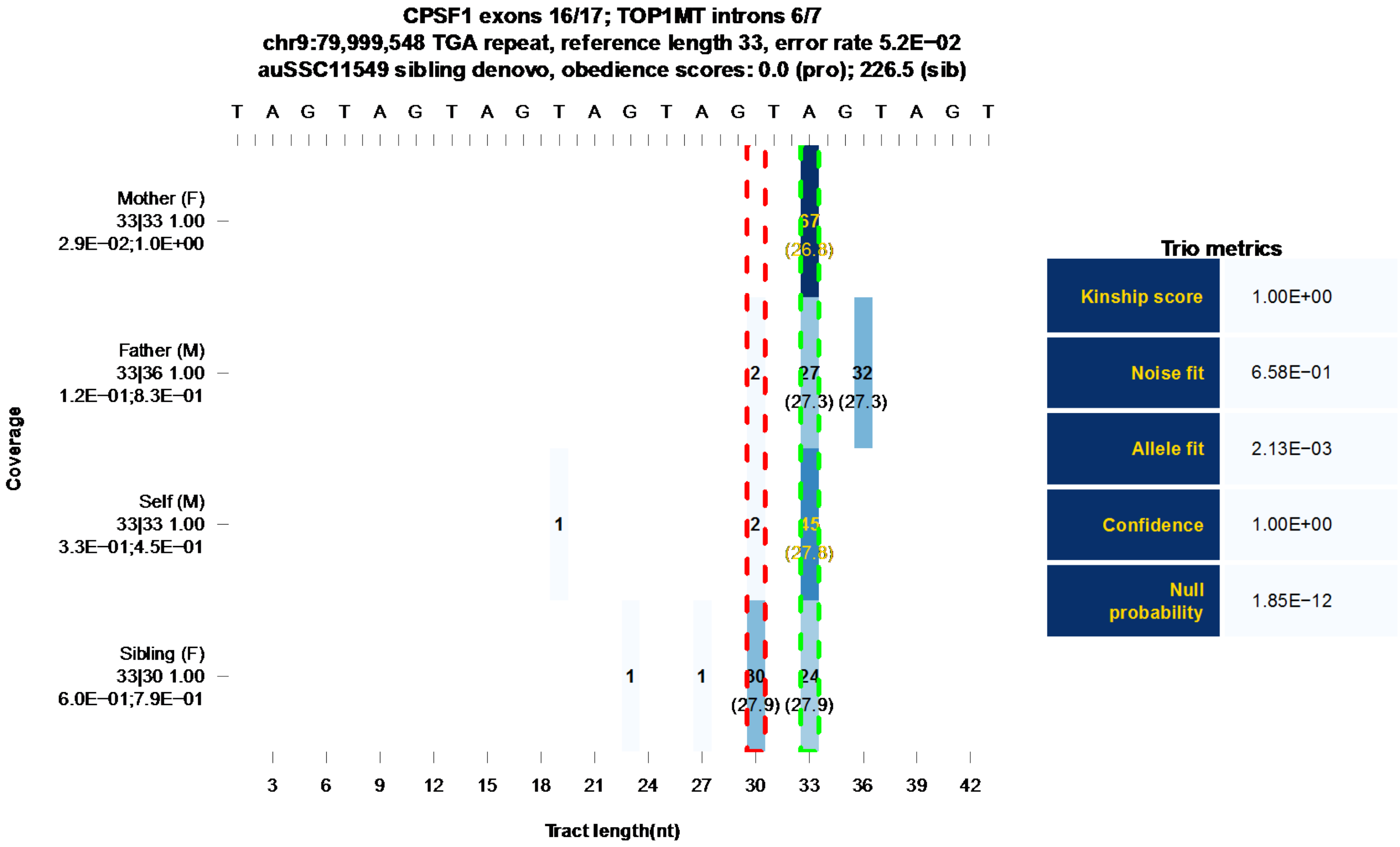
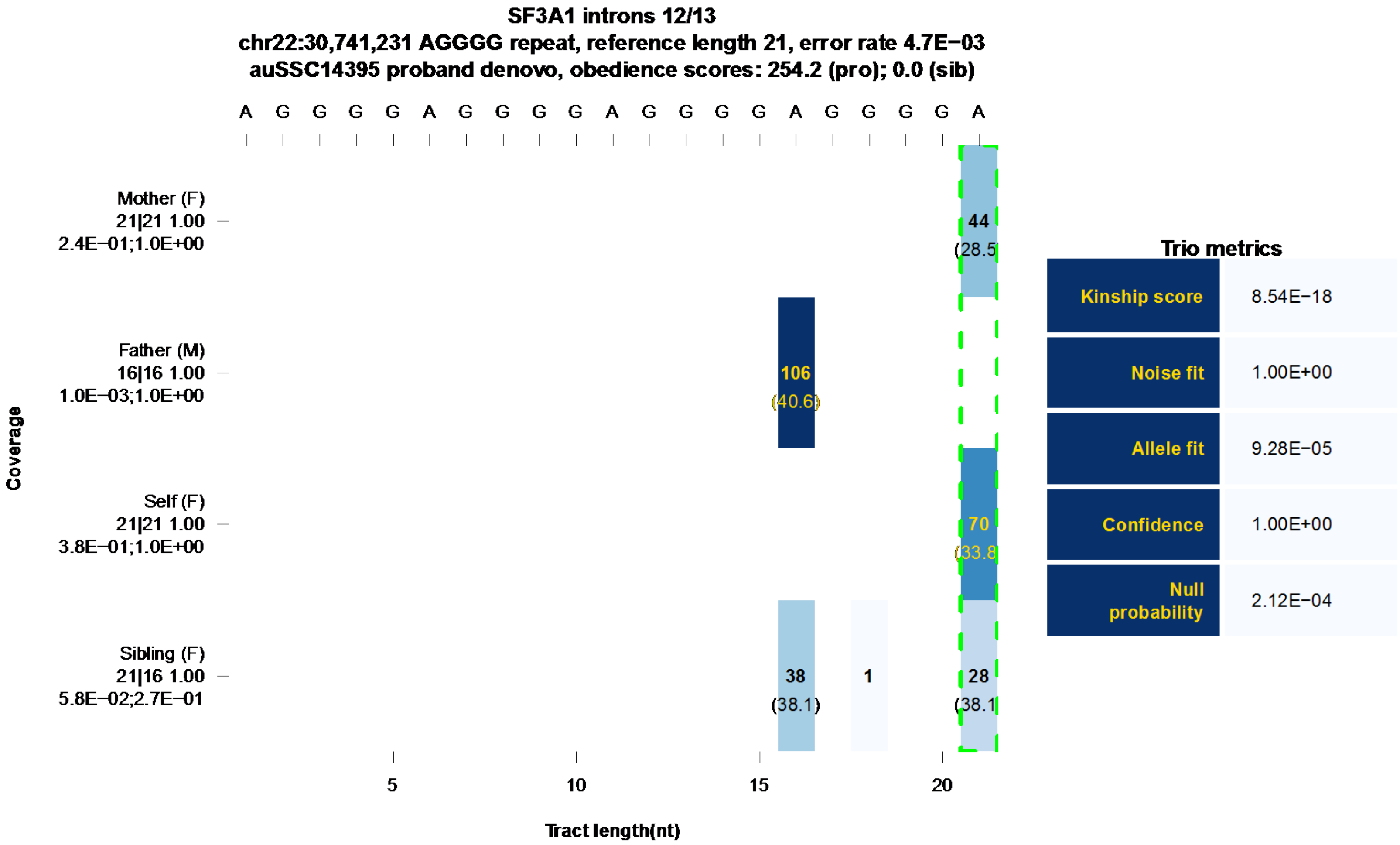


Figure 6.54: A validated in-frame microsatellite *de novo* mutation in CPSF1 in an unaffected sibling. Since the mutation is in-frame, it is unlikely to have a phenotypic effect.



Kinship score	8.54E−18
Noise fit	1.00E+00
Allele fit	9.28E−05
Confidence	1.00E+00
Null probability	2.12E−04

Figure 6.55: The first *de novo* mutation indicating a Robertsonian translocation on chromosome 22 in the proband of auSSC14395. The mother is homozygous for a 21 bp allele in an AGGGG microsatellite in an intron of SF3A1, while the father is homozygous for a 16 bp allele. Like the mother, the child with autism is homozygous for the 21 bp allele.

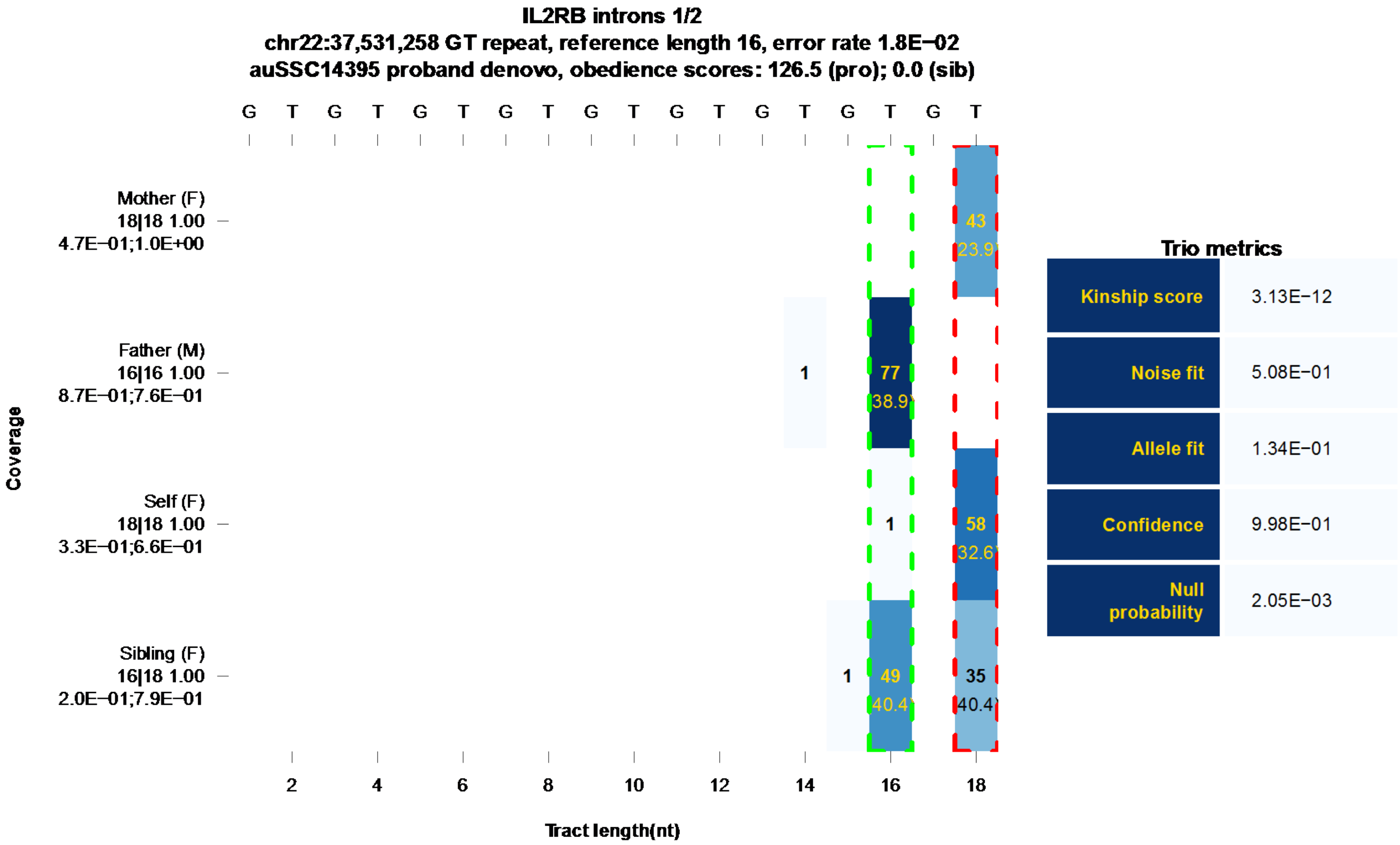
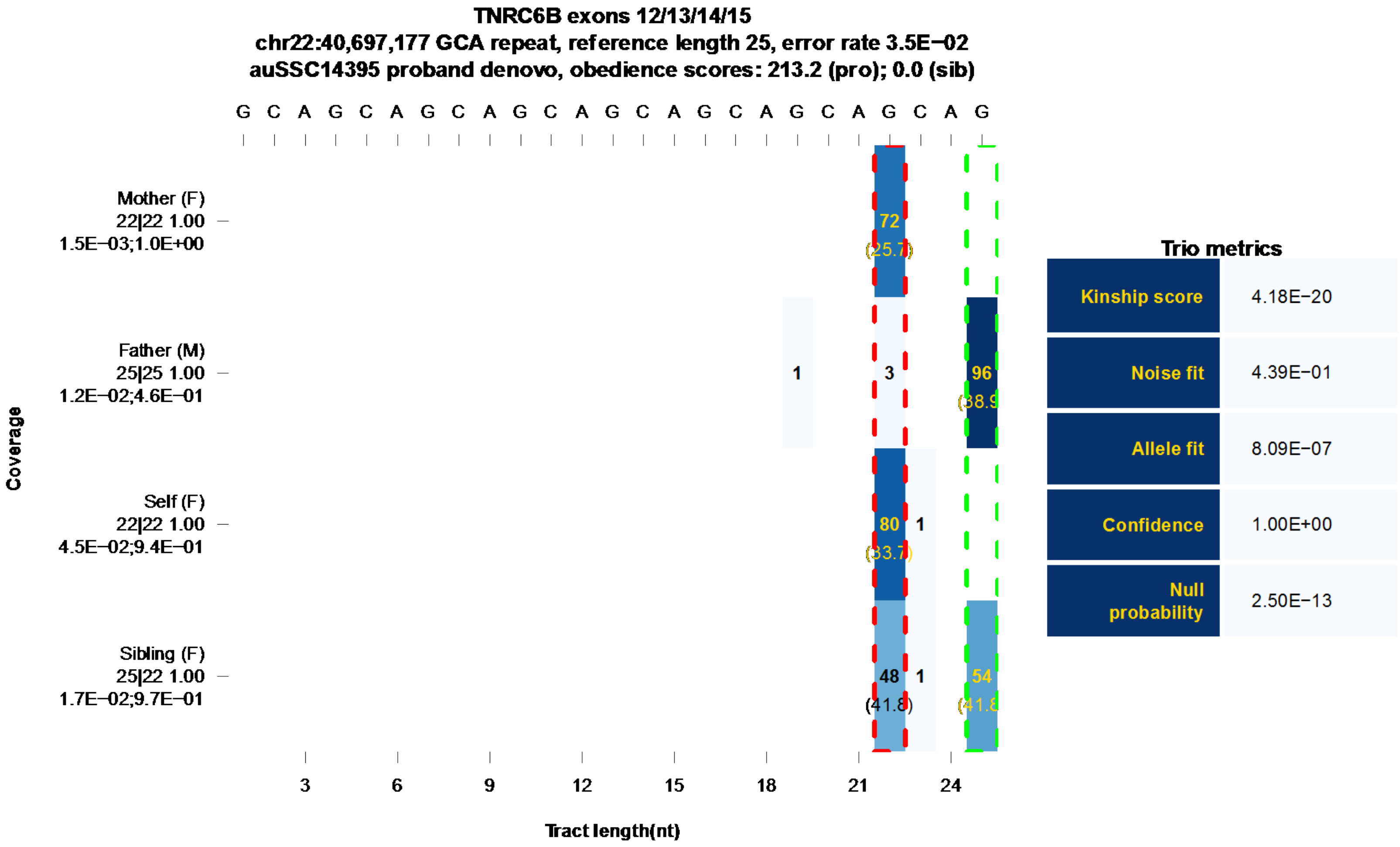


Figure 6.56: The second *de novo* mutation indicating a Robertsonian translocation on chromosome 22 in the proband of auSSC14395. The mother is homozygous for an 18 bp allele in a GT microsatellite in an intron of IL2RB, while the father is homozygous for a 16 bp allele. Like the mother, the child with autism is homozygous for the 18 bp allele.



Kinship score	4.18E−20
Noise fit	4.39E−01
Allele fit	8.09E−07
Confidence	1.00E+00
Null probability	2.50E−13

Figure 6.57: The third *de novo* mutation indicating a Robertsonian translocation on chromosome 22 in the proband of auSSC14395. The mother is homozygous for a 22 bp allele in a GCA microsatellite in an exon of TNRC6B, while the father is homozygous for a 25 bp allele. Like the mother, the child with autism is homozygous for the 22 bp allele.

	All	1	2	3	4	5	6
GRCh37	12.01	11.83	10.90	10.33	16.99	19.82	22.03
	(10)	(10)	(9)	(9)	(15)	(18)	(21)
Intergenic	12.15	11.71	10.92	10.34	17.19	19.95	22.20
	(10)	(10)	(9)	(9)	(15)	(18)	(21)
Intron	11.87	12.00	10.90	10.32	16.74	19.67	21.82
	(10)	(10)	(9)	(9)	(15)	(18)	(21)
5' UTR	11.17	11.09	10.42	10.62	14.56	18.30	21.78
	(9)	(9)	(9)	(9)	(13)	(17)	(20)
3' UTR	11.07	11.00	10.35	10.47	14.85	18.22	21.08
	(9)	(9)	(9)	(9)	(13)	(17)	(20)
Exon	10.14	8.33	8.48	10.38	12.85	15.94	21.05
	(9)	(8)	(8)	(9)	(12)	(15)	(20)
miRNA	10.88	11.00	10.00	11.33	16.33	15.00	
	(10)	(11)	(9)	(9.5)	(16)	(15)	n/a

Table 6.1: Mean microsatellite tract lengths for each genomic context and motif length. Medians are in parentheses. Each column is labeled by its motif length(s). Tract lengths in exons tend to be shorter than elsewhere in the genome for frameshift-inducing motif lengths.

	A	C
GRCh37	98.4%	1.6%
Intergenic	98.5%	1.5%
Intron	98.3%	1.7%
5' UTR	95.2%	4.8%
3' UTR	95.8%	4.2%
Exon	85.2%	14.7%
miRNA	100.0%	0.0%

Table 6.2**:** Distribution of 1 bp microsatellite equivalence classes by context. C microsatellites are more prevalent in exons than they are in the rest of the genome.

	AC	AG	AT	CG
GRCh37	35.8%	30.1%	33.4%	0.7%
Intergenic	34.3%	30.1%	35.0%	0.5%
Intron	38.0%	29.9%	31.5%	0.7%
5' UTR	38.8%	32.2%	21.9%	7.1%
3' UTR	41.7%	30.0%	23.0%	5.3%
Exon	33.9%	51.6%	4.1%	10.4%
miRNA	47.6%	0.0%	52.4%	0.0%

Table 6.3**:** Distribution of 2 bp microsatellite equivalence classes by context. AG microsatellites are more common in exons than they are elsewhere, while AT microsatellites are depleted in exons. CG microsatellites are more common in exons and UTRs than they are in the rest of the genome.

	AAC	AAG	AAT	ACC	ACG	ACT	AGC	AGG	ATC	CCG
GRCh37	10.4%	13.0%	25.2%	16.8%	0.1%	2.6%	9.3%	13.0%	8.1%	1.5%
Intergenic	10.8%	13.9%	26.9%	15.6%	0.1%	2.4%	8.6%	12.3%	8.5%	0.9%
Intron	10.4%	11.8%	24.6%	19.2%	0.1%	2.9%	9.1%	13.4%	7.6%	1.0%
5' UTR	6.5%	7.8%	10.4%	9.9%	0.4%	1.5%	14.9%	18.8%	3.9%	25.8%
3' UTR	7.2%	8.7%	12.3%	11.3%	0.3%	1.9%	15.2%	18.7%	4.9%	19.5%
Exon	3.3%	15.4%	1.3%	10.3%	1.2%	1.1%	25.3%	20.1%	9.0%	13.2%
miRNA	16.7%	0.0%	16.7%	16.7%	0.0%	0.0%	0.0%	33.3%	16.7%	0.0%

Table 6.4: Distribution of 3 bp microsatellite equivalence classes by context. Most notably, AAT microsatellites are much less common in exons than they are in the rest of the genome; while AGC, AGG, and CCG microsatellites are more common in exons and UTRs than they are in the rest of the genome.

	Count	Percent of well-covered loci	Percent of reference microsatellite loci
GRCh37	133,300	100.0%	2.3%
Intergenic	16,946	12.7%	0.5%
Intron	76,426	57.3%	3.3%
5' UTR	2,412	1.8%	15.2%
3' UTR	2,944	2.2%	7.1%
Exon	34,569	25.9%	83.3%
miRNA	3	0.0%	8.8%

Table 6.5: Summary of well-covered loci in SSC exome-capture dataset. "Percent of well-covered loci" refers to the proportion of microsatellites from a particular context within the well-covered set. “Percent of reference microsatellite loci” refers to the proportion of microsatellites from a particular context within the SSC well-covered set. Although introns make up a large portion of the well-covered set, the vast majority of reference intron microsatellite loci are not observed. Most exon reference microsatellite loci are observed in the set of well-covered loci.

ID	Chr	Pos[1]	Motif	Gene name	Genotype[2]	Score[3]	Code[4]	Novel allele	Unique[5]	Seen[6]	Status[7]
13602	chr1	3679939	GAG	CCDC27	9 9 9 9 9 9 10 9	131.61	EPICp	10	UR	Y	Y
14559	chr1	16360072	C	CLCNKA	8 8 8 8 9 9 9 10	72.64	ISICs	10	CR	N	F
11138	chr1	77042616	AAAG	ST6GALNAC3	19 19 19 19 15 19 15 23	87.14	ISICs	23	CD	N	F
12892	chr2	72742084	AC	EXOC6B	24 24 24 24 26 26 26 22	71.08	ISDCs	22	CR	N	F
12424	chr4	4115618	A	NA	8 8 10 8 10 10 9 10	67.06	IPICs	9	CR	N	F
11800	chr4	70825805	A	CSN2	9 9 9 9 9 9 9 8	126.89	ISDCs	8	UR	N	Y
13791	chr5	162917426	A	HMMR	9 9 9 9 9 9 10 9	99.07	EPICs	10	UR	N	F
12579	chr6	49701625	AC	CRISP3	12 12 12 12 12 12 12 10	79.56	ISDCs	10	UR	N	Y
12473	chr6	69785876	T	BAI3	9 9 9 9 9 9 9 10	278.18	ISICs	10	CR	N	N
14393	chr6	109954066	ATAG	AKD1	45 41 45 49 37 53 41 37	111.59	3SICs	49	CR	N	*Y
13412	chr7	141315261	C	AGK	8 8 8 8 8 8 9 8	155.13	IPICs	9	CR	N	Y
13579	chr8	97245323	A	GTPBP4	11 11 11 11 11 11 10 11	101.00	IPDCs	10	CR	N	Y
13418	chr9	27217784	T	EIF3H	8 8 8 8 8 8 9 8	84.71	IPICp	9	CR	N	N
11549	chr9	79999548	TGA	CPSF1	33 33 33 33 33 36 33 30	226.46	ESDCs	30	CR	N	Y

ID	Chr	Pos[1]	Motif	Gene name	Genotype[2]	Score[3]	Code[4]	Novel allele	Unique[5]	Seen[6]	Status[7]
12497	chr10	70056620	A	PBLD	13 12 12 12 14 12 14 12	66.25	ISNOs	12	CR	N	*F
14573	chr11	379842	C	B4GALNT4	8 8 8 10 10 10 10 9	67.12	ISICs	9	CR	N	F
11657	chr11	836759	C	CD151	8 8 8 8 8 8 8 9	69.27	ISICs	9	CR	N	N
14333	chr11	26692595	TG	SLC5A12	17 17 17 31 31 31 17 29	113.15	3SICs	29	CR	N	?
13529	chr11	71344184	TTG	NA	20 20 20 20 20 20 17 20	73.97	NPDCs	17	CR	N	N
12952	chr11	76928312	ATCT	GDPD4	20 20 20 20 20 20 24 20	82.67	(E3)PICs	24	UR	Y[†]	Y
12728	chr12	39763594	T	KIF21A	8 8 8 8 8 8 9 8	67.24	EPICs	9	UR	N	Y
14070	chr12	58025103	C	B4GALNT1	8 8 8 8 8 8 9 8	63.52	EPICs	9	UR	Y	Y
11552	chr12	118506328	TCC	VSIG10	26 26 26 26 20 26 26 23	192.16	ESDCs	23	CR	N	Y
13508	chr12	124362257	G	DNAH10	8 8 9 8 9 9 11 8	65.05	IPICp	11	CR	N	F
11138	chr13	26788216	ATC	RNF6	13 13 13 13 13 13 13 10	274.19	ESDCs	10	UR	Y	Y
14233	chr14	21859176	T	CHD8	8 8 8 8 8 8 9 8	72.54	EPICs	9	UR	Y[‡]	Y
14021	chr16	8858701	CA	ABAT	22 22 22 22 24 24 24 20	71.60	ISDCs	20	CR	N	?
11376	chr16	8858701	CA	ABAT	22 22 22 22 24 24 24 20	75.03	ISDCs	20	CR	N	?

ID	Chr	Pos[1]	Motif	Gene name	Genotype[2]	Score[3]	Code[4]	Novel allele	Unique[5]	Seen[6]	Status[7]
14231	chr16	27230338	GGA	KDM8	12 12 12 12 12 12 12 9	181.33	ESDCs	9	UM	Y	Y
12997	chr16	28631315	TG	SULT1A1	25 25 25 25 19 19 19 21	80.64	ISICs	21	CR	N	F
12864	chr16	70993832	T	HYDIN	8 8 8 8 8 8 9 8	65.98	IPICs	9	UR	N	N
14573	chr16	72089985	TG	HP	10 10 10 10 10 10 10 11	144.64	ISICp	11	UR	N	F
13166	chr16	88585146	TGGA	ZFPM1	38 38 38 38 38 38 34 38	199.65	IPDCs	34	CD	N	F
13692	chr16	88585519	TGGA	ZFPM1	19 19 19 19 22 22 22 26	99.50	ISICs	26	CR	N	F
12493	chr17	42937082	AC	EFTUD2	11 11 11 11 11 11 12 11	231.89	IPICp	12	UR	N	Y
14066	chr17	43227526	GGAGCT	HEXIM1	21 21 21 21 21 21 27 21	66.40	EPICs	27	UR	Y	Y
13807	chr19	3910998	TG	ATCAY	12 12 12 12 12 12 10 12	90.06	IPDCs	10	CR	N	F
14679	chr19	7184372	TTG	INSR	10 10 10 10 10 10 11 10	168.01	EPICp	11	UR	Y	F
11436	chr19	45912490	AAG	CD3EAP	20 20 20 20 20 20 23 20	111.84	EPICs	23	UD	Y	Y
12723	chr21	9907174	CA	NA	13 13 13 13 13 13 12 12	68.86	NPDCp	12	CR	N	?
12861	chr21	10969895	ATGG	TPTE	32 32 32 32 32 32 36 32	76.35	IPICs	36	CR	N	F
13508	chr21	14771784	AC	NA	9 9 9 9 9 9 9 11	70.92	NSICs	11	CI	N	N

ID	Chr	Pos[1]	Motif	Gene name	Genotype[2]	Score[3]	Code[4]	Novel allele	Unique[5]	Seen[6]	Status[7]
13620	chr22	20920814	CAG	MED15	36 36 36 36 36 36 36 39	63.72	ESICs	39	CR	N	*Y
14395	chr22	30741231	AGGGG	SF3A1	21 16 21 21 21 16 21 16	254.22	IPNOs	21	CD	N	*Y
14395	chr22	37531258	GT	IL2RB	18 16 18 16 18 16 18 18	126.51	IPIOs	18	CR	N	*Y

Table 6.6: De novo microsatellite mutations above the validation threshold. [1]Microsatellite start positions are provided in the GRCh36 human genome assembly. [2]”Genotype” provides the family genotype at the microsatellite locus. There is one line for each allele. In each line, the numbers correspond to the microsatellite tract lengths in the mother, father, proband, and sibling, respectively. [3]Mendel obedience scores for the trio containing the *de novo* mutation. [4]This code describes the *de novo* mutation. The first character is “E”, “I”, “3”, “5”, or “N” for a de novo mutation in an exon, intron, 3’ UTR, 5’ UTR, or intergenic region, respectively. If a locus spans multiple contexts, they are coded in parentheses. The second character is either “P” or “S” for a *de novo* mutation in a proband or sibling. The third character describes the *de novo* indel status—“I” is an insertion, “D” a deletion, and “N” a mutation that is identical to the reference tract length. The fourth character describes the mutation as a commission (C) or an omission (O). The final character indicates whether the mutation is due to a slippage event (s), or point mutation (p). [5]”unique” describes whether the mutation is unique in the SSC cohort or dbSNP v138. The first character describes whether the mutation is unique (U) or common (C) in the SSC cohort. The second character describes whether a mutation in the microsatellite is found in dbSNP v138. “R” is a mutation that is not found in dbSNP, “I” is a dbSNP insertion, “D” is a dbSNP deletion, and “M” is a dbSNP missense mutation. [6]The “Seen” field describes whether the de novo mutation has been detected previously. Unless noted otherwise, mutations were detected in unpublished results using a GATK-based pipeline or Scalpel. [7]”Status” describes the validation status of each *de novo* mutation. “Y” and “N” define valid and invalid mutations. “F” is failed validation, and “?” is an ambiguous validation result. Finally, the “*” marks *de novo* mutations that were submitted for validation despite not passing the appropriate thresholds.

[†]Iossifov, et al. *Neuron* 2012

[‡]O’Roak, et al. *Science* 2012

	Exon	Intergenic	Intron	3' UTR	Total
Proband	9	3	11	0	23
Sibling	5	1	15	1	22
p	0.42	0.63	0.55	1	1

Table 6.7: Number of *de novo* mutations in children with autism and their siblings by context. P-values are derived from a two-sided binomial exact text with an expected probability of 0.5. No bias in the rate of *de novo* mutations between children with autism and their siblings is observed for any context.

	Exon	Intergenic	Intron	3' UTR	Total
Valid	12	0	8	0	20
Invalid	0	2	4	0	6
Unclear	0	1	2	1	4
Failed	2	1	12	0	15

Table 6.8: Validation status by context for *de novo* mutations with a Mendel obedience score ≥63. Most of these *de novo* mutations validate or failed to validate.

	Deletion	Insertion	NA	Total
Proband	4	12	1	17
Sibling	9	11	1	21
p	0.26	1	1	0.63

Table 6.9: Insertions and deletions by child. P-values are derived from a two-sided binomial exact text with an expected probability of 0.5. A value of "NA" means that the novel allele has the same tract length as the reference allele. No bias is observed in the rate of *de novo* mutations between children with autism and their siblings for insertions or deletions.

	Unique	Not unique	Total
Proband	8	9	17
Sibling	4	17	21
p	0.39	0.16	0.63

Table 6.10: Novel allele unique status by child. P-values are derived from a two-sided binomial exact text with an expected probability of 0.5. No bias is observed in the rate of *de novo* mutations between children with autism and their siblings for unique or common mutations.

	Commission	Omission	Total
Proband	15	2	17
Sibling	20	1	21
p	0.5	1	0.63

Table 6.11: Commissions and omissions by child. P-values are derived from a two-sided binomial exact text with an expected probability of 0.5. No bias is observed in the rate of *de novo* mutations between children with autism and their siblings for commissions or omission.

	Deletion	Insertion	Total
Proband	0	6	6
Sibling	2	0	2
p	0.5	**0.03**	0.29

Table 6.12: Validated unique *de novo* mutations. P-values are derived from a two-sided binomial exact text with an expected probability of 0.5. Significant p-values are in bold. Children with autism are significantly more likely to have a unique *de novo* microsatellite expansion than their siblings (p=0.03).

7 Conclusions and Perspectives

7.1 A role for microsatellite mutation in autism incidence

7.1.1 The role of *de novo* microsatellite mutations in autism

This work presents the first evidence suggesting that microsatellite mutations may play a role in autism, and is among the first evidence of microsatellite mutations playing a role in a complex human disease. Children with autism are significantly enriched for unique *de novo* insertions as compared to their siblings, and while not statistically significant, we only observe microsatellite frameshift mutations in children with autism.

While frameshift mutations have an obvious effect on protein function, the effects of short in-frame insertions are less obvious. If these mutations do have a role in the autism phenotype, perhaps the strongest parallel is to the disease etiology of OPMD. As discussed in the introduction, small expansions of the GCG trinucleotide repeat in the first exon of PABPN1, which code for a polyalanine stretch, lead to the occurrence of OPMD (Brais, et al. 1998). Just 1 – 3 extra GCG repeats distinguish healthy individuals from those affected by OPMD. These small expansions lead to misfolding and aggregation of the PABPN1 protein, which impair the function of the ubiquitin-proteasome pathway and inhibit molecular chaperone function (Calado et al. 2000; Abu-Baker et al. 2003). While many short in-frame microsatellite mutations may be harmless, some of these mutations can have significant phenotypic effects.

7.1.2 Expanding the analysis of the SSC

The current study exploited exome sequencing data from a significant portion of the SSC—875 families—but can be strengthened by incorporating data from 118 additional SSC quads and 257 SSC trios which have been processed at CSHL and WUGSC. We would expect this data to lend more support to our observations of a unique insertion bias in children with autism. With these additional families, we may observe a statistically significant bias for *de novo* frameshift mutations in children with autism. Previous studies of the SSC exome data in the Wigler lab demonstrated a significant enrichment of *de novo* mutations in FMRP-associated genes (Iossifov et al. 2012). Although an expanded analysis could reveal a similar trend for *de novo* microsatellite mutations, there is no statistical enrichment for FMRP-associated genes in our study.

Collaborators of the Wigler lab have also sequenced SSC exome datasets, which are available for analysis. Sequencing performed in the Eichler lab on the SSC collection would add 158 parent-proband trios and 31 quads. The data generated by the Eichler lab has a mix of 50- and 76-bp paired end reads—shorter reads may lead to fewer detected microsatellites in this dataset (O'Roak et al. 2012b). Sequencing performed in the State lab would add 225 quads and 25 trios to our analysis. As with the Eichler sequencing data, care must be taken when comparing results among datasets—in addition to having performed sequencing on Illumina's GAIIx and HiSeq 2000 sequencing platforms with 76- or 101-bp paired-end reads, the State lab also used two different exome capture kits. One is the NimbleGen EZExome V2.0 that has been used by the other studies, while the other is a custom array designed by the State lab (Sanders et al. 2012). Shorter read lengths will hamper the observation of some microsatellite tracts, and each exome capture kit will have its own unique biases when capturing alleles of varying lengths. While uSeq should be sufficiently robust to call genotypes using HTS data generated from diverse experimental protocols, it is probably wise to call genotypes separately for data from each lab. This will ensure that the per-allele expected coverage parameters, allele bias parameters, and locus noise parameters are protocol-specific.

7.1.3 Transmitted microsatellite variants in autism

In the current study, we have not evaluated the role of transmitted microsatellite polymorphisms in the autism incidence. Given the size of our study population and the accuracy of our genotyping, this analysis is possible. An initial analysis might leverage a statistic similar to the transmission disequilibrium test to identify microsatellite alleles that are more common in children with autism than their siblings. Given the gender bias observed in autism, it would be reasonable to extend the analysis to the evaluation of gender-specific biased variant transmission. This analysis could identify alleles that have a sex-specific effect on autism risk. The tools for this analysis have yet to be developed, but the statistical tests may be simpler than those used to identify strong *de novo* microsatellite mutation candidates in this study.

7.2 An expanded role for microsatellites in human disease

7.2.1 Studying microsatellite mutations in cancer

If past research is any indication, there is a broad role for microsatellite mutations in human disease etiology. Most studies of microsatellite mutations in cancer describe frameshift mutations in mononucleotide microsatellites that disrupt the function of tumor suppressor genes (see, for example, Markowitz et al. 1995; Parsons et al. 1995; Malkhosyan et al. 1996; Miyaki et al. 1997; Rampino et al. 1997). The exception to this category is AIB1—the mutations are in-frame and the mechanism by which they increase breast cancer risk remains unclear (Kadouri et al. 2004). Many mutations in tumor suppressor genes occur in short microsatellite tracts with 1 bp motifs—the mutations inactivating TGFβRII and BAX occur in a 10 bp A microsatellite tract and an 8 bp G microsatellite tract, respectively (Parsons et al. 1995; Rampino et al. 1997). These mutations have only been characterized in tumors exhibiting MMP, implying that these short tracts may be generally stable, unless components of the MMR pathway are inactivated. Of the *de novo* mutations we observe in our study, many are at loci whose reference lengths are between 8 and 10 bp. This is despite the fact that there is no reason to assume that MMP is prevalent within our study cohort. Therefore, *de novo* mutations at otherwise stable short microsatellite loci may play a more significant role in cancer incidence than previous studies have indicated. The effect of a particular microsatellite mutation can also depend on the genetic background of an affected individual—for instance, CAG expansions in AIB1 increase the risk of breast cancer in BRCA1 carriers, but not BRCA2 carriers (Kadouri et al. 2004). The extent to which microsatellite mutations play a role in cancer is far from being completely defined.

7.2.2 Studying microsatellite mutations in neurological diseases and other genetic disorders

In neurological disorders—such as Huntington’s disease, Fragile X syndrome, or ALS—monogenic, highly penetrant microsatellite mutations act in a variety of ways to disrupt normal protein function (Kremer et al. 1991; Pieretti et al. 1991; Macdonald et al. 1993; DeJesus-Hernandez et al. 2011; Renton et al. 2011). Many neurological disorders, including myotonic dystrophy and SCA1, are autosomal dominant, which could indicate a disease model for other heterozygous microsatellite mutations (see, for example, Mahadevan et al. 1992; Orr et al. 1993). Disease-causing microsatellite mutations are not limited to exons—Fragile X is caused by a CGG repeat expansion in the 5' UTR of FMR1 and Friedreich ataxia is caused by a GAA repeat expansion in the first intron of FXN (Kremer et al. 1991; Campuzano et al. 1996). Microsatellite mutations in non-exonic regions may have a variety of phenotypic effects—for example, they can alter spacing between cis-regulatory elements such as transcription factor binding sites or splicing enhancers, or they can provide additional binding sites for regulatory molecules. Sequence identity within a microsatellite tract in a particular haplotype can dramatically affect tract stability, as has been shown in FMR1 and SCA1 (Chung et al. 1993; Eichler et al. 1994; Chong et al. 1995). The introduction of variant repeats into an ancient microsatellite tract in some genetic backgrounds, or their removal in others, is almost certain to play a role in disease incidence in uncharacterized microsatellite diseases. Microsatellite mutations may also affect gene function in novel, yet unimagined ways. It is almost certain that the full complement of diseases influenced by microsatellite mutations have yet to be characterized.

7.2.3 Somatic microsatellite mutations

Microsatellite instability is not limited to the germline. Somatic microsatellite instability plays an obvious role in MMP cancers, but intact MMR machinery does not guarantee somatic microsatellite stability. Somatic microsatellite instability has also been observed at disease loci in male germline cells, peripheral blood, and within specific tissues in the brain (see, for example, Chong et al. 1995). Any disease caused by a mutant microsatellite allele can be caused by somatic mutation. The elevated mutation rate suggested by many early studies of microsatellite loci rate at microsatellite loci suggests that these loci may play a significant role in sporadic disease incidence and require careful study.

Somatic microsatellite mutations are also a rich source of intracellular variation that can be used to recover high-resolution developmental lineages. Microsatellite mutations at a few hundred loci have been sufficient to distinguish developmental lineages in normal mice (Salipante and Horwitz 2006). Somatic microsatellite mutations may be the only source of mutations sufficiently frequent enough to create measurable intracellular variation. As has been made apparent in this study, microsatellite instability varies enormously among loci, so care needs to be taken that a set of loci capable of generating sufficient polymorphism are observed in the cells being analyzed.

7.3 Understanding microsatellite stability

Understanding the role of microsatellite mutations in human disease depends on in-depth knowledge of the factors influencing the stability of individual microsatellite loci. Several studies have already demonstrated that the per-generation, per-gamete microsatellite mutation rate can vary widely between loci, but many of the factors affecting microsatellite mutation remain poorly understood. Microsatellite motif length is inversely correlated with mutation rates. Motif composition clearly has an effect on microsatellite mutation rate, but that relationship is still poorly defined. While it is clear that variant repeats stabilize microsatellite tracts, many studies identify the longest uninterrupted microsatellite sequence as the primary determinant of microsatellite stability for these loci, which could imply that the vast majority of slippage mutations occur in uninterrupted microsatellite sequence. Additionally, in our study, microsatellite loci with identical composition and identical reference tract lengths have vastly different levels of polymorphism within our study population. These differences may be due to the pressure of selective forces on particular microsatellite loci, stochastic variation between loci, or factors affecting microsatellite stability that have yet to be defined.

Leveraging uSeq and current sequencing technology, we can thoroughly examine these factors in ways that were impossible until now. Using the SSC study described here—and other population-scale studies like it—we can investigate microsatellite polymorphism and mutation throughout the human genome. These studies will allow us to better identify microsatellite loci that are likely to mutate, as well as aid in identifying haplotypes that are particularly stable or unstable. This information will be invaluable in understanding the roles microsatellites may play in human disease, and will serve as an important tool in furthering our understanding of the dynamics of microsatellite instability.

7.4 Population genetics of microsatellites

Population genetics has been an invaluable tool in the advancement of the scientific community's understanding of microsatellite loci. Some of the early estimates of the per-generation, per-locus microsatellite mutation rate were made based on extensions of population genetics to microsatellite loci (Brinkmann et al. 1998). These same analyses may refine our estimates of microsatellite mutation rates based on polymorphism in the SSC dataset, and the same can be done in future studies. Population genetics may also identify microsatellite loci that violate the neutral theory of natural selection. This could be invaluable in identifying evidence of purifying selection at microsatellite loci, particularly with respect to coding regions.

As our understanding of microsatellite polymorphism deepens, the population genetic analyses possible using microsatellite loci will also expand. Population genetic analyses of microsatellite loci have already enabled many fascinating studies, including support for the out of Africa hypothesis on the origin of modern humans and selective sweeps in natural populations of *Drosophila melanogaster* (Bowcock et al. 1994; Jorde et al. 1995; Jorde et al. 1997; Schlötterer et al. 1997; Di Rienzo et al. 1998). In tandem with other sources of genetic information, microsatellites can further develop our knowledge of the genetic histories of humans, other species, and complex interplay of humans with the biological systems with which they interact.

The ability to genotype microsatellite loci throughout the genome in large populations enables a shift towards broader strategies for identifying genetic associations between microsatellite mutations and various phenotypes, including the family-based study design demonstrated in this project as well as GWAS. However, the assumptions of linkage disequilibrium that are essential to GWAS studies need to be revisited for microsatellite loci. Alleles at polymorphic microsatellite loci are shared by geographically distinct populations, which might imply that microsatellite loci must be studied directly to call genotypes, and cannot be inferred from nearby SNPs in linkage disequilibrium (Bowcock et al. 1994).

7.5 The importance of genome-wide microsatellite genotyping

7.5.1 uSeq can detect arbitrarily small microsatellites

Genome-wide microsatellite analyses are only possible with the recent development of algorithms and pipelines that leverage high-throughput sequencing data and address specific algorithmic challenges posed by microsatellite loci. uSeq is a potent and powerful method for these types of analyses, and has several unique characteristics that make it superior for studying microsatellite mutations in large populations. uSeq can detect microsatellites with arbitrary tract length and repeat number thresholds—while our current study is limited to microsatellites with a tract length of at least 8 bp and 3 repeats of the same motif, microsatellite mutations can still occur in shorter tracts (Huang et al. 1996; Greene and Jinks-Robertson 1997). Although the uSeq detection algorithm may be limited by the parameters for minimum tract length and motif length range, future implementations of uSeq can address this challenge easily. Detecting short microsatellite tracts is particularly important in cancer—microsatellite mutations have been observed in tracts as short as 2 bp in the APC gene in both MMP and non-MMP tumors and mutations to 8 bp mononucleotide tracts in MSH3 and MSH6 are frequently deletions (Huang et al. 1996; Malkhosyan et al. 1996). In comparison to other microsatellite genotyping algorithms, uSeq can survey many more loci, and is the only pipeline to leverage population-level information to call highly accurate genotypes and identify strong *de novo* mutation candidates.

7.5.2 uSeq is the only microsatellite pipeline capable of identifying *de novo* microsatellite mutations

The ability to identify these *de novo* mutations is the result of the algorithms we have developed and implemented as part of the uSeq pipeline. Although lobSTR and RepeatSeq both model microsatellite slippage errors as part of their genotyping methods, only uSeq has a locus-specific error model, which provides superior genotype-calling accuracy. As is evident in our study, perfect microsatellites with identical motif and tract lengths can have widely varying error rates. The logistic regression error model used by lobSTR and the empirically derived Bayesian model selection error model used by RepeatSeq are simply not equipped to address this variability. Indeed, when comparing Mendelian inheritance violations among pipelines, many of the violations called by lobSTR and RepeatSeq are at loci that the uSeq population EM identifies as being particularly noisy. uSeq is also the only genotyper to define a per-allele, per-person expected coverage, which is essential for determining whether the observed coverage in an individual is sufficient to call a confident bi-allelic genotype. uSeq is the only genotyper to consider allele bias, which is an important tool when distinguishing between noise and signal at microsatellite loci, and is particularly important when identifying *de novo* mutations. While uSeq's allele bias model was designed to address allele biases due to varying affinities of exome capture probes for microsatellite indels, it should be able to model allele bias arising from any number of sources, such as tract length or GC content. uSeq's superior genotyping accuracy is evident in its ability to identify *de novo* microsatellite mutations with high specificity and good precision. While RepeatSeq and lobSTR have admirable genotyping accuracy, neither pipeline is capable of accurately identifying *de novo* microsatellite mutations (Gymrek et al. 2012; Highnam et al. 2013).

7.5.3 uSeq is fast, requires relatively little disk space, and is highly parallelizable

uSeq is rapid and efficient enough to use on large datasets, as demonstrated in its application to a 3,500-person exome sequencing study on a large portion of the SSC. The complete analysis of the SSC dataset took less than 2 weeks on a 14-node computer cluster with 24 cores per node. Less than one week of that time was spent on detection and alignment. If we were to attempt the same analysis with lobSTR, which took ~19 hours on average to detect and align microsatellites per individual in one family using 2 cores, it would take more than 2 weeks of dedicated time on a similar cluster just to perform detection and alignments. If a dedicated cluster were used to sort the alignments and call allelotypes, only 2 hours of time would be needed. Since RepeatSeq does not perform its own alignments, it would only take approximately 3 hours of dedicated cluster time to obtain genotypes, although the alignment algorithms recommended in the RepeatSeq manuscript can take 1 – 2 days to map just 43 million reads (Highnam et al. 2013). uSeq output is not overly burdensome—approximately 2.1 GB of data is kept for each SSC individual processed (~7 TB total for 3,500 people), and the genotyper output for 3,500 individuals takes an additional ~25 GB. uSeq performs well on large populations since it has low memory overhead and is primarily single-threaded, which allows for a highly parallelizable pipeline appropriate for use on large populations. Neither RepeatSeq nor lobSTR have been tested in large populations, and the memory or CPU demands of either pipeline may limit either pipeline's ability to efficiently process sequencing data from very large datasets.

7.5.4 Improvements to uSeq

uSeq can still be improved and expanded in several meaningful ways. Among the most obvious improvements would be enabling detection of interrupted microsatellites and improving mapping accuracy for reads with 5' or 3' truncated microsatellites, as described in the methods section. Adding imperfect microsatellite detection may increase the number of observable microsatellite loci, but based on previous work relating microsatellite stability to microsatellite sequence identity, these loci may not add much polymorphism (see, for example, Weber 1990; Henderson and Petes 1992; Eichler et al. 1994; McIver et al. 2011). Improving microsatellite mapping is not expected to dramatically alter coverage at reference microsatellite loci. However, improved mapping would enable detection of novel microsatellite alleles not present in reference genomes, which could reveal population-specific microsatellite loci and could help elucidate the mechanisms of microsatellite emergence in the genome.

uSeq's *de novo* mutation detection is currently geared towards quads, since this is the pedigree structure of the SSC data analyzed. As we expand our analysis of the uSeq dataset to include trios, *de novo* detection will be modified to allow for arbitrary pedigree structures, which will expand uSeq's utility in disease and population genetic studies. Similarly, if uSeq is applied to matched tumor-normal data in cancer studies or sequencing data from different organs within one individual, other arbitrary structures defining the relationships between samples can be implemented.

uSeq is currently not publicly available, although it can be made available on request. In the very near term, uSeq will be made available for download, which will encourage new development and applications. Importantly, although uSeq has only been tested on human sequencing data, in theory, its function is independent of organism. Applications to non-human data will help identify any species-dependence that may need to be resolved.

7.6 Leveraging population data to call genotypes

One of the primary focuses of this project has been the development of methods for accurate microsatellite genotyping by leveraging population-level data to estimate per-allele, per-person expected coverage; per-allele bias; and per-locus noise parameters. These considerations are not limited to microsatellite loci.

Expected coverage is a relevant genotyping parameter in many experimental setups. In any capture experiment, expected coverage will vary among loci due to many factors, including the location of a target site relative to its capture probe; the capture probe affinity for its target region; and variability between experiments, even if they use the same protocol. In whole-genome studies, expected coverage can vary among loci depending on the uniqueness of a particular genomic region, its GC content, and experimental variability.

Genotyping algorithms must consider the likelihood that a particular locus is bi-allelic, which is made possible by having an estimator for expected coverage. Even if a genotyper is not attempting to call CNPs, a bi-allelic genotype only makes sense if there is sufficient evidence that there are only two alleles at a locus. In particular, when the detection of a particular genomic feature is not trivial, as is the case with indels, a good algorithm should consider the likelihood that it missed relevant data. In uSeq, this uncertainty is reflected in the null genotype probability. If expected per-allele coverage can be calculated accurately, genotypes with uncharacteristically low coverage can considered with the appropriate level of skepticism.

As demonstrated in this thesis, indels can have a significant, unpredictable effect on observed coverage in targeted sequencing experiments. Failure to consider allele bias will lead to inaccurate genotype calls, and will reduce the sensitivity and specificity of any downstream analyses. For example, before implementing the uSeq allele bias model, many *de novo* calls were made in trios where a parent had lower than expected coverage for the novel allele. In many of these cases, the low parental coverage was due to inefficient capture of the novel allele, and was readily apparent upon examination of that allele throughout the SSC study population. Similarly, truly novel alleles can be missed when bias is not considered, particularly if that allele has less coverage than expected. If the approach taken in uSeq is any indication, a robust allele bias model depends on having an accurate expected coverage model. Allele bias may be an important consideration in whole-genome sequencing studies as well. Even without the additional biases contributed by targeted capture approaches, the efficiency with which different indels are observed in sequencing data might be expected to vary. This could be due to allele length, or other factors that have yet to be characterized.

Locus-specific error rates have a clear mechanism explaining their variability when considering microsatellite loci, but that does not limit their utility to microsatellite genotyping. Sequencing technologies have systematic errors that are locus- or sequence-specific—for instance, a 2011 study characterized two sources of sequence-specific error on Illumina's Genome Analyzer II (Nakamura et al. 2011). These systematic errors can vary between platforms, and could vary depending on the particular chemistry or protocol used to generate sequencing data. Similarly, particular regions of the genome may have increased rates of mutation. This is most apparent in the case of cytosine deamination. Rather than defining locus-specific error models empirically for each combination of protocol, platform, and chemistry, the general approach used here to estimate locus-specific error rates given population-level sequencing data may prove to be simpler to implement and more accurate.

7.7 Conclusion

In the course of this thesis, we have developed a novel microsatellite pipeline that can readily detect and map reads with microsatellite indels and accurately call genotypes using a sophisticated genotyping model that leverages the richness of the data we have from our study population. This approach has proven accurate enough to identify *de novo* mutations in families with sporadic autism. Previously, the only way to identify *de novo* mutations at microsatellite loci was to target specific loci for analysis. This approach was only feasible on a few thousand loci at most, and made it impossible to look for *de novo* microsatellite mutations in the complete exome, let alone genome-wide. This is the first study that has been able to survey microsatellites on such a scale in an unbiased manner and accurately identify *de novo* mutations.

Using this ability, we have been able to identify a role for microsatellite mutations in autism incidence. We expect the evidence for this association to be strengthened as we expand our analysis. More broadly, the contribution of microsatellites to human disease and general phenotypic variation has been hampered for the last ten years by the absence of highly accurate HTS microsatellite genotyping methods. uSeq is the most appropriate microsatellite genotyping pipeline for these analyses. It provides significant advantages for microsatellite genotyping, both in terms of its accuracy and the scope of the genomic microsatellite landscape it can cover. The door is now open to a much greater and nuanced understanding of microsatellite function. Given the variety of effects attributed to microsatellite mutations over the last 20 years, it is exciting to consider what the future holds in store as we enter a new age of genome-wide microsatellite studies.

8 References

Aaltonen LA, Peltomaki P, Leach FS, Sistonen P, Pylkkanen L, Mecklin JP, Jarvinen H, Powell SM, Jen J, Hamilton SR et al. 1993. Clues to the pathogenesis of familial colorectal cancer. *Science* **260**: 812-816.

Abu-Baker A, Messaed C, Laganiere J, Gaspar C, Brais B, Rouleau GA. 2003. Involvement of the ubiquitin-proteasome pathway and molecular chaperones in oculopharyngeal muscular dystrophy. *Hum Mol Genet* **12**: 2609-2623.

Alazzouzi H, Davalos V, Kokko A, Domingo E, Woerner SM, Wilson AJ, Konrad L, Laiho P, Espin E, Armengol M et al. 2005. Mechanisms of inactivation of the receptor tyrosine kinase EPHB2 in colorectal tumors. *Cancer Res* **65**: 10170-10173.

American Psychiatric Association APADSMTF. 2013. Diagnostic and statistical manual of mental disorders : DSM-5.

Ananda G, Walsh E, Jacob KD, Krasilnikova M, Eckert KA, Chiaromonte F, Makova KD. 2013. Distinct mutational behaviors differentiate short tandem repeats from microsatellites in the human genome. *Genome Biol Evol* **5**: 606-620.

Anders S, Huber W. 2010. Differential expression analysis for sequence count data. *Genome Biol* **11**: R106.

Andersen SL, Sekelsky J. 2010. Meiotic versus mitotic recombination: two different routes for double-strand break repair: the different functions of meiotic versus mitotic DSB repair are reflected in different pathway usage and different outcomes. *BioEssays* **32**: 1058-1066.

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

Baglama J, Reichel L. 2014. irlba: Fast partial SVD by implicitly-restarted Lanczos bidiagonalization.

Banfi S, Servadio A, Chung MY, Kwiatkowski TJ, Jr., McCall AE, Duvick LA, Shen Y, Roth EJ, Orr HT, Zoghbi HY. 1994. Identification and characterization of the gene causing type 1 spinocerebellar ataxia. *Nat Genet* **7**: 513-520.

Benson G. 1999. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* **27**: 573-580.

Bentley DR Balasubramanian S Swerdlow HP Smith GP Milton J Brown CG Hall KP Evers DJ Barnes CL Bignell HR et al. 2008. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**: 53-59.

Bentley JL, Sleator DD, Tarjan RE, Wei VK. 1986. A locally adaptive data compression scheme. *Commun ACM* **29**: 320-330.

Bornman DM, Hester ME, Schuetter JM, Kasoji MD, Minard-Smith A, Barden CA, Nelson SC, Godbold GD, Baker CH, Yang B et al. 2012. Short-read, high-throughput sequencing technology for STR genotyping. *BioTechniques* **0**: 1-6.

Boukhris A, Schule R, Loureiro JL, Lourenco CM, Mundwiller E, Gonzalez MA, Charles P, Gauthier J, Rekik I, Lebrigio RFA et al. 2013. Alteration of ganglioside biosynthesis responsible for complex hereditary spastic paraplegia. *Am J Hum Genet* **93**: 118-123.

Bowcock AM, Ruiz-Linares A, Tomfohrde J, Minch E, Kidd JR, Cavalli-Sforza LL. 1994. High resolution of human evolutionary trees with polymorphic microsatellites. *Nature* **368**: 455-457.

Brais B, Bouchard JP, Xie YG, Rochefort DL, Chretien N, Tome FMS, Lafreniere RG, Rommens JM, Uyama E, Nohira O et al. 1998. Short GCG expansions in the PABP2 gene cause oculopharyngeal muscular dystrophy. *Nat Genet* **18**: 164-167.

Brinkmann B, Klintschar M, Neuhuber F, Huhne J, 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.

Brook JD, Mccurrach ME, Harley HG, Buckler AJ, Church D, Aburatani H, Hunter K, Stanton VP, Thirion JP, Hudson T et al. 1992. Molecular basis of myotonic dystrophy: expansion of a trinucleotide (CTG) repeat at the 3' end of a transcript encoding a protein-kinase family member. *Cell* **68**: 799-808.

Burke JR, Wingfield MS, Lewis KE, Roses AD, Lee JE, Hulette C, Pericakvance MA, Vance JM. 1994. The Haw River syndrome: Dentatorubropallidoluysian atrophy (DRPLA) in an African-American family. *Nat Genet* **7**: 521-524.

Buxton J, Shelbourne P, Davies J, Jones C, Vantongeren T, Aslanidis C, Dejong P, Jansen G, Anvret M, Riley B et al. 1992. Detection of an unstable fragment of DNA specific to individuals with myotonic dystrophy. *Nature* **355**: 547-548.

Calado A, Tome FM, Brais B, Rouleau GA, Kuhn U, Wahle E, Carmo-Fonseca M. 2000. Nuclear inclusions in oculopharyngeal muscular dystrophy consist of poly(A) binding protein 2 aggregates which sequester poly(A) RNA. *Hum Mol Genet* **9**: 2321-2328.

Calin G, Herlea V, Barbanti-Brodano G, Negrini M. 1998. The coding region of the bloom syndrome BLM gene and of the CBL proto-oncogene is mutated in genetically unstable sporadic gastrointestinal tumors. *Cancer Res* **58**: 3777-3781.

Calin G, Ranzani GN, Amadori D, Herlea V, Matei I, Barbanti-Brodano G, Negrini M. 2001. Somatic frameshift mutations in the Bloom syndrome BLM gene are frequent in sporadic gastric carcinomas with microsatellite mutator phenotype. *BMC Genet* **2**.

Calin GA, Gafa R, Tibiletti MG, Herlea V, Becheanu G, Cavazzini L, Barbanti-Brodano G, Nenci I, Negrini M, Lanza G. 2000. Genetic progression in microsatellite instability high (MSI-H) colon cancers correlates with clinico-pathological parameters: a study of the TGF beta RII, BAX, HMSH3, HMSH6, IGFIIR and BLM genes. *Int J Cancer* **89**: 230-235.

Campuzano V, Montermini L, Molto MD, Pianese L, Cossee M, Cavalcanti F, Monros E, Rodius F, Duclos F, Monticelli A et al. 1996. Friedreich's ataxia: autosomal recessive disease caused by an intronic GAA triplet repeat expansion. *Science* **271**: 1423-1427.

Cao MD, Tasker E, Willadsen K, Imelfort M, Vishwanathan S, Sureshkumar S, Balasubramanian S, Boden M. 2014. Inferring short tandem repeat variation from paired-end short reads. *Nucleic Acids Res* **42**: e16.

Centers for Disease Control and Prevention (CDC). 2014. Prevalence of autism spectrum disorder among children aged 8 years — Autism and Developmental Disabilities Monitoring Network, 11 Sites, United States, 2010. *MMWR Morbidity and Mortality Weekly Reports* **63**: 1-21.

Chakraborty R, Kimmel M, Stivers DN, Davison LJ, Deka R. 1997. Relative mutation rates at di-, tri-, and tetranucleotide microsatellite loci. *P Natl Acad Sci USA* **94**: 1041-1046.

Chong SS, Mccall AE, Cota J, Subramony SH, Orr HT, Hughes MR, Zoghbi HY. 1995. Gametic and somatic tissue-specific heterogeneity of the expanded SCA1 CAG repeat in spinocerebellar ataxia type-1. *Nat Genet* **10**: 344-350.

Chung MY, Ranum LPW, Duvick LA, Servadio A, Zoghbi HY, Orr HT. 1993. Evidence for a mechanism predisposing to intergenerational CAG repeat instability in spinocerebellar ataxia type-I. *Nat Genet* **5**: 254-258.

Darnell JC, Van Driesche SJ, Zhang CL, Hung KYS, Mele A, Fraser CE, Stone EF, Chen C, Fak JJ, Chi SW et al. 2011. FMRP stalls ribosomal translocation on mRNAs linked to synaptic function and autism. *Cell* **146**: 247-261.

Davalos V, Dopeso H, Velho S, Ferreira AM, Cirnes L, Diaz-Chico N, Bilbao C, Ramirez R, Rodriguez G, Falcon O et al. 2007. High EPHB2 mutation rate in gastric but not endometrial tumors with microsatellite instability. *Oncogene* **26**: 308-311.

David G, Abbas N, Stevanin G, Durr A, Yvert G, Cancel G, Weber C, Imbert G, Saudou F, Antoniou E et al. 1997. Cloning of the SCA7 gene reveals a highly unstable CAG repeat expansion. *Nat Genet* **17**: 65-70.

de la Chapelle A, Peltomaki P. 1995. Genetics of hereditary colon cancer. *Annu Rev Genet* **29**: 329-348.

DeJesus-Hernandez M, Mackenzie IR, Boeve BF, Boxer AL, Baker M, Rutherford NJ, Nicholson AM, Finch NA, Flynn H, Adamson J et al. 2011. Expanded GGGGCC hexanucleotide repeat in noncoding region of C9ORF72 causes chromosome 9p-linked FTD and ALS. *Neuron* **72**: 245-256.

Di Rienzo A, Donnelly P, Toomajian C, Sisk B, Hill A, Petzl-Erler ML, Haines GK, Barch DH. 1998. Heterogeneity of microsatellite mutations within and between loci, and implications for human demographic histories. *Genetics* **148**: 1269-1284.

Dib C, Faure S, Fizames C, Samson D, Drouot N, Vignal A, Millasseau P, Marc S, Hazan J, Seboun E et al. 1996. A comprehensive genetic map of the human genome based on 5,264 microsatellites. *Nature* **380**: 152-154.

Dohet C, Wagner R, Radman M. 1986. Methyl-directed repair of frameshift mutations in heteroduplex DNA. *P Natl Acad Sci USA* **83**: 3395-3397.

Drummond JT, Li GM, Longley MJ, Modrich P. 1995. Isolation of an hMSH2-P160 heterodimer that restores DNA mismatch repair to tumor cells. *Science* **268**: 1909-1912.

Duval A, Gayet J, Zhou XP, Iacopetta B, Thomas G, Hamelin R. 1999. Frequent frameshift mutations of the TCF-4 gene in colorectal cancers with microsatellite instability. *Cancer Res* **59**: 4213-4215.

Eichler EE, Holden JJA, Popovich BW, Reiss AL, Snow K, Thibodeau SN, Richards CS, Ward PA, Nelson DL. 1994. Length of uninterrupted CGG repeats determines instability in the FMR1 gene. *Nat Genet* **8**: 88-94.

Ellegren H. 2004. Microsatellites: simple sequences with complex evolution. *Nat Rev Genet* **5**: 435-445.

Farabaugh PJ, Schmeissner U, Hofer M, Miller JH. 1978. Genetic studies of the lac repressor. VII. On the molecular nature of spontaneous hotspots in the lacI gene of Escherichia coli. *J Mol Biol* **126**: 847-857.

Fischbach GD, Lord C. 2010. The Simons Simplex Collection: a resource for identification of autism genetic risk factors. *Neuron* **68**: 192-195.

Fishel R, Lescoe MK, Rao MR, Copeland NG, Jenkins NA, Garber J, Kane M, Kolodner R. 1994. The human mutator gene homolog MSH2 and its association with hereditary nonpolyposis colon cancer. *Cell* **77**: 1027-1038.

Fondon JW, 3rd, Garner HR. 2007. Detection of length-dependent effects of tandem repeat alleles by 3-D geometric decomposition of craniofacial variation. *Dev Genes Evol* **217**: 79-85.

Fondon JW, Garner HR. 2004. Molecular origins of rapid and continuous morphological evolution. *P Natl Acad Sci USA* **101**: 18058-18063.

Foster EA, Jobling MA, Taylor PG, Donnelly P, de Knijff P, Mieremet R, Zerjal T, Tyler-Smith C. 1998. Jefferson fathered slave's last child. *Nature* **396**: 27-28.

Frumkin D, Wasserstrom A, Itzkovitz S, Stern T, Harmelin A, Eilam R, Rechavi G, Shapiro E. 2008. Cell lineage analysis of a mouse tumor. *Cancer Res* **68**: 5924-5931.

Frumkin D, Wasserstrom A, Kaplan S, Feige U, Shapiro E. 2005. Genomic variability within an organism exposes its cell lineage tree. *PLOS Comput Biol* **1**: 382-394.

Fu YH, Kuhl DPA, Pizzuti A, Pieretti M, Sutcliffe JS, Richards S, Verkerk AJMH, Holden JJA, Fenwick RG, Warren ST et al. 1991. Variation of the CGG repeat at the fragile X site results in genetic instability: resolution of the Sherman paradox. *Cell* **67**: 1047-1058.

Fu YH, Pizzuti A, Fenwick RG, King J, Rajnarayan S, Dunne PW, Dubel J, Nasser GA, Ashizawa T, Dejong P et al. 1992. An unstable triplet repeat in a gene related to myotonic muscular dystrophy. *Science* **255**: 1256-1258.

Gerber HP, Seipel K, Georgiev O, Hofferer M, Hug M, Rusconi S, Schaffner W. 1994. Transcriptional activation modulated by homopolymeric glutamine and proline stretches. *Science* **263**: 808-811.

Gill P, Ivanov PL, Kimpton C, Piercy R, Benson N, Tully G, Evett I, Hagelberg E, Sullivan K. 1994. Identification of the remains of the Romanov family by DNA analysis. *Nat Genet* **6**: 130-135.

Glenn TC, Stephan W, Dessauer HC, Braun MJ. 1996. Allelic diversity in alligator microsatellite loci is negatively correlated with GC content of flanking sequences and evolutionary conservation of PCR amplifiability. *Mol Biol Evol* **13**: 1151-1154.

Goldstein DB, Clark AG. 1995. Microsatellite variation in North American populations of Drosophila melanogaster. *Nucleic Acids Res* **23**: 3882-3886.

Greene CN, Jinks-Robertson S. 1997. Frameshift intermediates in homopolymer runs are removed efficiently by yeast mismatch repair proteins. *Mol Cell Biol* **17**: 2844-2850.

Gymrek M. 2013. Modeling PCR stutter noise for accurate calling of STRs from short reads.

Gymrek M, Golan D, Rosset S, Erlich Y. 2012. lobSTR: A short tandem repeat profiler for personal genomes. *Genome Res* **22**: 1154-1162.

Gymrek M, McGuire AL, Golan D, Halperin E, Erlich Y. 2013. Identifying personal genomes by surname inference. *Science* **339**: 321-324.

Hagelberg E, Gray IC, Jeffreys AJ. 1991. Identification of the skeletal remains of a murder victim by DNA analysis. *Nature* **352**: 427-429.

Hallmayer J, Cleveland S, Torres A, Phillips J, Cohen B, Torigoe T, Miller J, Fedele A, Collins J, Smith K et al. 2011. Genetic heritability and shared environmental factors among twin pairs with autism. *Arch Gen Psychiat* **68**: 1095-1102.

Hamada H, Kakunaga T. 1982. Potential Z-DNA forming sequences are highly dispersed in the human genome. *Nature* **298**: 396-398.

Hamada H, Petrino MG, Kakunaga T. 1982a. Molecular structure and evolutionary origin of human cardiac muscle actin gene. *P Natl Acad Sci USA* **79**: 5901-5905.

-. 1982b. A novel repeated element with Z-DNA-forming potential is widely found in evolutionarily diverse eukaryotic genomes. *P Natl Acad Sci USA* **79**: 6465-6469.

Hamada H, Seidman M, Howard BH, Gorman CM. 1984. Enhanced gene expression by the poly(dT-dG).poly(dC-dA) sequence. *Mol Cell Biol* **4**: 2622-2630.

Hammock EA, Young LJ. 2005. Microsatellite instability generates diversity in brain and sociobehavioral traits. *Science* **308**: 1630-1634.

Han HJ, Yanagisawa A, Kato Y, Park JG, Nakamura Y. 1993. Genetic instability in pancreatic cancer and poorly differentiated type of gastric cancer. *Cancer Res* **53**: 5087-5089.

Heale SM, Petes TD. 1995. The stabilization of repetitive tracts of DNA by variant repeats requires a functional DNA mismatch repair system. *Cell* **83**: 539-545.

Henderson ST, Petes TD. 1992. Instability of simple sequence DNA in Saccharomyces cerevisiae. *Mol Cell Biol* **12**: 2749-2757.

Henke J, Henke L. 1999. Mutation rate in human microsatellites. *Am J Hum Genet* **64**: 1473-1474.

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

Highnam G, Franck C, Martin A, Stephens C, Puthige A, Mittelman D. 2013. Accurate human microsatellite genotypes from high-throughput resequencing data using informed error profiles. *Nucleic Acids Res* **41**: e32.

Huang J, Papadopoulos N, McKinley AJ, Farrington SM, Curtis LJ, Wyllie AH, Zheng S, Willson JK, Markowitz SD, Morin P et al. 1996. APC mutations in colorectal tumors with mismatch repair deficiency. *P Natl Acad Sci USA* **93**: 9049-9054.

Imbert G, Saudou F, Yvert G, Devys D, Trottier Y, Garnier JM, Weber C, Mandel JL, Cancel G, Abbas N et al. 1996. Cloning of the gene for spinocerebellar ataxia 2 reveals a locus with high sensitivity to expanded CAG/glutamine repeats. *Nat Genet* **14**: 285-291.

Ionov Y, Peinado MA, Malkhosyan S, Shibata D, Perucho M. 1993. Ubiquitous somatic mutations in simple repeated sequences reveal a new mechanism for colonic carcinogenesis. *Nature* **363**: 558-561.

Iossifov I, Ronemus M, Levy D, Wang Z, Hakker I, Rosenbaum J, Yamrom B, Lee YH, Narzisi G, Leotta A et al. 2012. De novo gene disruptions in children on the autistic spectrum. *Neuron* **74**: 285-299.

Jeang KT, Hayward GS. 1983. A cytomegalovirus DNA sequence containing tracts of tandemly repeated CA dinucleotides hybridizes to highly repetitive dispersed elements in mammalian cell genomes. *Mol Cell Biol* **3**: 1389-1402.

Jin L, Macaubas C, Hallmayer J, Kimura A, Mignot E. 1996. Mutation rate varies among alleles at a microsatellite locus: phylogenetic evidence. *P Natl Acad Sci USA* **93**: 15285-15288.

Jobling MA, Gill P. 2004. Encoded evidence: DNA in forensic analysis. *Nat Rev Genet* **5**: 739-751.

Jodice C, Malaspina P, Persichetti F, Novelletto A, Spadaro M, Giunti P, Morocutti C, Terrenato L, Harding AE, Frontali M. 1994. Effect of trinucleotide repeat length and parental sex on phenotypic variation in spinocerebellar ataxia I. *Am J Hum Genet* **54**: 959-965.

Johnson RE, Kovvali GK, Prakash L, Prakash S. 1996. Requirement of the yeast MSH3 and MSH6 genes for MSH2-dependent genomic stability. *J Biol Chem* **271**: 7285-7288.

Jorde LB, Bamshad MJ, Watkins WS, Zenger R, Fraley AE, Krakowiak PA, Carpenter KD, Soodyall H, Jenkins T, Rogers AR. 1995. Origins and affinities of modern humans: a comparison of mitochondrial and nuclear genetic data. *Am J Hum Genet* **57**: 523-538.

Jorde LB, Rogers AR, Bamshad M, Watkins WS, Krakowiak P, Sung S, Kere J, Harpending HC. 1997. Microsatellite diversity and the demographic history of modern humans. *P Natl Acad Sci USA* **94**: 3100-3103.

Kadouri L, Kote-Jarai Z, Easton DF, Hubert A, Hamoudi R, Glaser B, Abeliovich D, Peretz T, Eeles RA. 2004. Polyglutamine repeat length in the AIB1 gene modifies breast cancer susceptibility in BRCA1 carriers. *Int J Cancer* **108**: 399-403.

Kang S, Jaworski A, Ohshima K, Wells RD. 1995. Expansion and deletion of CTG repeats from human disease genes are determined by the direction of replication in E. coli. *Nat Genet* **10**: 213-218.

Kawaguchi Y, Okamoto T, Taniwaki M, Aizawa M, Inoue M, Katayama S, Kawakami H, Nakamura S, Nishimura M, Akiguchi I et al. 1994. CAG expansions in a novel gene for Machado-Joseph disease at chromosome 14q32.1. *Nat Genet* **8**: 221-228.

Kayser M, Roewer L, Hedman M, Henke L, Henke J, Brauer S, Kruger C, Krawczak M, Nagy M, Dobosz T et al. 2000. Characteristics and frequency of germline mutations at microsatellite loci from the human Y chromosome, as revealed by direct observation in father/son pairs. *Am J Hum Genet* **66**: 1580-1588.

Kelkar YD, Tyekucheva S, Chiaromonte F, Makova KD. 2008. The genome-wide determinants of human and chimpanzee microsatellite evolution. *Genome Res* **18**: 30-38.

Khorana HG, Buchi H, Ghosh H, Gupta N, Jacob TM, Kossel H, Morgan R, Narang SA, Ohtsuka E, Wells RD. 1966. Polynucleotide synthesis and the genetic code. *Cold Spring Harb Sym* **31**: 39-49.

Kim NG, Choi YR, Baek MJ, Kim YH, Kang H, Kim NK, Min JS, Kim H. 2001. Frameshift mutations at coding mononucleotide repeats of the hRAD50 gene in gastrointestinal carcinomas with microsatellite instability. *Cancer Res* **61**: 36-38.

Kimpton CP, Gill P, Walton A, Urquhart A, Millican ES, Adams M. 1993. Automated DNA profiling employing multiplex amplification of short tandem repeat loci. *PCR Meth Appl* **3**: 13-22.

Koide R, Ikeuchi T, Onodera O, Tanaka H, Igarashi S, Endo K, Takahashi H, Kondo R, Ishikawa A, Hayashi T et al. 1994. Unstable expansion of CAG repeat in hereditary dentatorubral-pallidoluysian atrophy (DRPLA). *Nat Genet* **6**: 9-13.

Kornberg A, Bertsch LL, Jackson JF, Khorana HG. 1964. Enzymatic synthesis of deoxyribonucleic acid, XVI. Oligonucleotides as templates and the mechanism of their replication. *P Natl Acad Sci USA* **51**: 315-323.

Kremer EJ, Pritchard M, Lynch M, Yu S, Holman K, Baker E, Warren ST, Schlessinger D, Sutherland GR, Richards RI. 1991. Mapping of DNA instability at the fragile X to a trinucleotide repeat sequence P(CCG)n. *Science* **252**: 1711-1714.

La Spada AR, Wilson EM, Lubahn DB, Harding AE, Fischbeck KH. 1991. Androgen receptor gene-mutations in X-linked spinal and bulbar muscular-atrophy. *Nature* **352**: 77-79.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**: 357-359.

Leach FS, Nicolaides NC, Papadopoulos N, Liu B, Jen J, Parsons R, Peltomaki P, Sistonen P, Aaltonen LA, Nystromlahti M et al. 1993. Mutations of a mutS homolog in hereditary nonpolyposis colorectal cancer. *Cell* **75**: 1215-1225.

Leclercq S, Rivals E, Jarne P. 2010. DNA slippage occurs at microsatellite loci without minimal threshold length in humans: a comparative genomic approach. *Genome Biol Evol* **2**: 325-335.

Lee JS, Hanford MG, Genova JL, Farber RA. 1999. Relative stabilities of dinucleotide and tetranucleotide repeats in cultured mammalian cells. *Hum Mol Genet* **8**: 2567-2572.

Levinson G, Gutman GA. 1987a. High-frequencies of short frameshifts in poly-CA/TG tandem repeats borne by bacteriophage-M13 in Escherichia coli K-12. *Nucleic Acids Res* **15**: 5323-5338.

-. 1987b. Slipped-strand mispairing: a major mechanism for DNA sequence evolution. *Mol Biol Evol* **4**: 203-221.

Levy D, Ronemus M, Yamrom B, Lee YH, Leotta A, Kendall J, Marks S, Lakshmi B, Pai D, Ye K et al. 2011. Rare de novo and transmitted copy-number variation in autistic spectrum disorders. *Neuron* **70**: 886-897.

Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv:13033997*.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**: 1754-1760.

Liquori CL, Ricker K, Moseley ML, Jacobsen JF, Kress W, Naylor SL, Day JW, Ranum LPW. 2001. Myotonic dystrophy type 2 caused by a CCTG expansion in intron 1 of ZNF9. *Science* **293**: 864-867.

Litt M, Luty JA. 1989. A hypervariable microsatellite revealed by invitro amplification of a dinucleotide repeat within the cardiac muscle actin gene. *Am J Hum Genet* **44**: 397-401.

Macdonald ME, Ambrose CM, Duyao MP, Myers RH, Lin C, Srinidhi L, Barnes G, Taylor SA, James M, Groot N et al. 1993. A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. *Cell* **72**: 971-983.

Mahadevan M, Tsilfidis C, Sabourin L, Shutler G, Amemiya C, Jansen G, Neville C, Narang M, Barcelo J, O'Hoy K et al. 1992. Myotonic dystrophy mutation: an unstable CTG repeat in the 3' untranslated region of the gene. *Science* **255**: 1253-1255.

Malkhosyan S, Rampino N, Yamamoto H, Perucho M. 1996. Frameshift mutator mutations. *Nature* **382**: 499-500.

Markowitz S, Wang J, Myeroff L, Parsons R, Sun LZ, Lutterbaugh J, Fan RS, Zborowska E, Kinzler KW, Vogelstein B et al. 1995. Inactivation of the type-II TGF-beta receptor in colon-cancer cells with microsatellite instability. *Science* **268**: 1336-1338.

Maurer DJ, O'Callaghan BL, Livingston DM. 1996. Orientation dependence of trinucleotide CAG repeat instability in Saccharomyces cerevisiae. *Mol Cell Biol* **16**: 6617-6622.

McIver LJ, Fondon JW, Skinner MA, Garner HR. 2011. Evaluation of microsatellite variation in the 1000 Genomes Project pilot studies is indicative of the quality and utility of the raw data and alignments. *Genomics* **97**: 193-199.

Merlo A, Mabry M, Gabrielson E, Vollmer R, Baylin SB, Sidransky D. 1994. Frequent microsatellite instability in primary small cell lung cancer. *Cancer Res* **54**: 2098-2101.

Messier W, Li SH, Stewart CB. 1996. The birth of microsatellites. *Nature* **381**: 483.

Miesfeld R, Krystal M, Arnheim N. 1981. A member of a new repeated sequence family which is conserved throughout eucaryotic evolution is found between the human delta and beta globin genes. *Nucleic Acids Res* **9**: 5931-5947.

Miyaki M, Konishi M, Tanaka K, Kikuchi-Yanoshita R, Muraoka M, Yasuno M, Igari T, Koike M, Chiba M, Mori T. 1997. Germline mutation of MSH6 as the cause of hereditary nonpolyposis colorectal cancer. *Nat Genet* **17**: 271-272.

Mochmann LH, Wells RD. 2004. Transcription influences the types of deletion and expansion products in an orientation-dependent manner from GAC*GTC repeats. *Nucleic Acids Res* **32**: 4469-4479.

Muragaki Y, Mundlos S, Upton J, Olsen BR. 1996. Altered growth and branching patterns in synpolydactyly caused by mutations in HOXD13. *Science* **272**: 548-551.

Nagafuchi S, Yanagisawa H, Ohsaki E, Shirayama T, Tadokoro K, Inoue T, Yamada M. 1994. Structure and expression of the gene responsible for the triplet repeat disorder, dentatorubral and pallidoluysian atrophy (DRPLA). *Nat Genet* **8**: 177-182.

Nakamura K, Oshima T, Morimoto T, Ikeda S, Yoshikawa H, Shiwa Y, Ishikawa S, Linak MC, Hirai A, Takahashi H et al. 2011. Sequence-specific error profile of Illumina sequencers. *Nucleic Acids Res* **39**: e90.

Narzisi G, O''Rawe JA, Iossifov I, Lee Y-h, Wang Z, Wu Y, Lyon GJ, Wigler M, Schatz MC. 2013. Accurate detection of de novo and transmitted indels within exome-capture data using micro-assembly. *bioRxiv*.

Neale BM, Kou Y, Liu L, Ma'ayan A, Samocha KE, Sabo A, Lin CF, Stevens C, Wang LS, Makarov V et al. 2012. Patterns and rates of exonic de novo mutations in autism spectrum disorders. *Nature* **485**: 242-U129.

Needleman SB, Wunsch CD. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol* **48**: 443-453.

Nicolaides NC, Papadopoulos N, Liu B, Wei YF, Carter KC, Ruben SM, Rosen CA, Haseltine WA, Fleischmann RD, Fraser CM et al. 1994. Mutations of two PMS homologues in hereditary nonpolyposis colon cancer. *Nature* **371**: 75-80.

O'Roak BJ, Vives L, Fu WQ, Egertson JD, Stanaway IB, Phelps IG, Carvill G, Kumar A, Lee C, Ankenman K et al. 2012a. Multiplex targeted sequencing identifies recurrently mutated genes in autism spectrum disorders. *Science* **338**: 1619-1622.

O'Roak BJ, Vives L, Girirajan S, Karakoc E, Krumm N, Coe BP, Levy R, Ko A, Lee C, Smith JD et al. 2012b. Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. *Nature* **485**: 246-250.

Olaisen B, Stenersen M, Mevag B. 1997. Identification by DNA analysis of the victims of the August 1996 Spitsbergen civil aircraft disaster. *Nat Genet* **15**: 402-405.

Orr HT, Chung MY, Banfi S, Kwiatkowski TJ, Jr., Servadio A, Beaudet AL, McCall AE, Duvick LA, Ranum LP, Zoghbi HY. 1993. Expansion of an unstable trinucleotide CAG repeat in spinocerebellar ataxia type 1. *Nat Genet* **4**: 221-226.

Orth K, Hung J, Gazdar A, Mathis M, Bowcock A, Sambrook J. 1994. Ovarian tumors display persistent microsatellite instability caused by mutation in the mismatch repair gene hMSH-2. *Cold Spring Harb Sym* **59**: 349-356.

Osborne RJ, Leech V. 1994. Polymerase chain reaction allelotyping of human ovarian cancer. *Brit J Cancer* **69**: 429-438.

Palombo F, Gallinari P, Iaccarino I, Lettieri T, Hughes M, D'Arrigo A, Truong O, Hsuan JJ, Jiricny J. 1995. GTBP, a 160-kilodalton protein essential for mismatch-binding activity in human cells. *Science* **268**: 1912-1914.

Papadopoulos N, Nicolaides NC, Wei YF, Ruben SM, Carter KC, Rosen CA, Haseltine WA, Fleischmann RD, Fraser CM, Adams MD et al. 1994. Mutation of a mutL homolog in hereditary colon cancer. *Science* **263**: 1625-1629.

Parsons R, Li GM, Longley MJ, Fang WH, Papadopoulos N, Jen J, de la Chapelle A, Kinzler KW, Vogelstein B, Modrich P. 1993. Hypermutability and mismatch repair deficiency in RER+ tumor cells. *Cell* **75**: 1227-1236.

Parsons R, Myeroff LL, Liu B, Willson JK, Markowitz SD, Kinzler KW, Vogelstein B. 1995. Microsatellite instability and mutations of the transforming growth factor beta type II receptor gene in colorectal cancer. *Cancer Res* **55**: 5548-5550.

Pemberton TJ, Sandefur CI, Jakobsson M, Rosenberg NA. 2009. Sequence determinants of human microsatellite variability. *BMC Genomics* **10**: 612.

Perucho M, Peinado MA, Ionov Y, Casares S, Malkhosyan S, Stanbridge E. 1994. Defects in replication fidelity of simple repeated sequences reveal a new mutator mechanism for oncogenesis. *Cold Spring Harb Sym* **59**: 339-348.

Petes TD, Greenwell PW, Dominska M. 1997. Stabilization of microsatellite sequences by variant repeats in the yeast Saccharomyces cerevisiae. *Genetics* **146**: 491-498.

Pieretti M, Zhang FP, Fu YH, Warren ST, Oostra BA, Caskey CT, Nelson DL. 1991. Absence of expression of the FMR-1 gene in fragile X syndrome. *Cell* **66**: 817-822.

Pulst SM, Nechiporuk A, Nechiporuk T, Gispert S, Chen XN, Lopes-Cendes I, Pearlman S, Starkman S, Orozco-Diaz G, Lunkes A et al. 1996. Moderate expansion of a normally biallelic trinucleotide repeat in spinocerebellar ataxia type 2. *Nat Genet* **14**: 269-276.

Pupko T, 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.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**: 841-842.

R Core Team. 2013. R: A Language and Environment for Statistical Computing. Vienna, Austria.

Rampino N, Yamamoto H, Ionov Y, Li Y, Sawai H, Reed JC, Perucho M. 1997. Somatic frameshift mutations in the BAX gene in colon cancers of the microsatellite mutator phenotype. *Science* **275**: 967-969.

Reizel Y, Chapal-Ilani N, Adar R, Itzkovitz S, Elbaz J, Maruvka YE, Segev E, Shlush LI, Dekel N, Shapiro E. 2011. Colon stem cell and crypt dynamics exposed by cell lineage reconstruction. *PLOS Genet* **7**.

Renton AE, Majounie E, Waite A, Simon-Sanchez J, Rollinson S, Gibbs JR, Schymick JC, Laaksovirta H, van Swieten JC, Myllykangas L et al. 2011. A hexanucleotide repeat expansion in C9ORF72 is the cause of chromosome 9p21-linked ALS-FTD. *Neuron* **72**: 257-268.

Risinger JI, Umar A, Boyd J, Berchuck A, Kunkel TA, Barrett JC. 1996. Mutation of MSH3 in endometrial cancer and evidence for its functional role in heteroduplex repair. *Nat Genet* **14**: 102-105.

Roche Nimblegen. 2010. NimbleGen SeqCap EZ Exome Library v2.0 – Design and Annotation Files: Release Notes. Roche NimbleGen, Inc.

Ronemus M, Iossifov I, Levy D, Wigler M. 2014. The role of de novo mutations in the genetics of autism spectrum disorders. *Nat Rev Genet* **15**: 133-141.

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

Rosenberg RE, Law JK, Yenokyan G, McGready J, Kaufmann WE, Law PA. 2009. Characteristics and concordance of autism spectrum disorders among 277 twin pairs. *Arch Pediat Adol Med* **163**: 907-914.

Saiki RK, Gelfand DH, Stoffel S, Scharf SJ, Higuchi R, Horn GT, Mullis KB, Erlich HA. 1988. Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science* **239**: 487-491.

Sajantila A, Lukka M, Syvanen AC. 1999. Experimentally observed germline mutations at human micro- and minisatellite loci. *Eur J Hum Genet* **7**: 263-266.

Salipante SJ, Horwitz MS. 2006. Phylogenetic fate mapping. *P Natl Acad Sci USA* **103**: 5448-5453.

Sanders SJ, Murtha MT, Gupta AR, Murdoch JD, Raubeson MJ, Willsey AJ, Ercan-Sencicek AG, DiLullo NM, Parikshak NN, Stein JL et al. 2012. De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature* **485**: 237-241.

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

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

Schug MD, Hutter CM, Wetterstrand KA, Gaudette MS, Mackay TF, Aquadro CF. 1998. The mutation rates of di-, tri- and tetranucleotide repeats in Drosophila melanogaster. *Mol Biol Evol* **15**: 1751-1760.

Sears KE, Goswami A, Flynn JJ, Niswander LA. 2007. The correlated evolution of Runx2 tandem repeats, transcriptional activity, and facial length in Carnivora. *Evol Dev* **9**: 555-565.

Sebat J, Lakshmi B, Malhotra D, Troge J, Lese-Martin C, Walsh T, Yamrom B, Yoon S, Krasnitz A, Kendall J et al. 2007. Strong association of de novo copy number mutations with autism. *Science* **316**: 445-449.

Sia EA, Jinks-Robertson S, Petes TD. 1997a. Genetic control of microsatellite stability. *Mutat Res* **383**: 61-70.

Sia EA, Kokoska RJ, Dominska M, Greenwell P, Petes TD. 1997b. Microsatellite instability in yeast: dependence on repeat unit size and DNA mismatch repair genes. *Mol Cell Biol* **17**: 2851-2858.

Skinner DM, Beattie WG. 1973. Cs2S04 gradients containing both Hg2+ and Ag+ effect the complete separation of satellite deoxyribonucleic acids having identical densities in neutral CsCl gradients. *P Natl Acad Sci USA* **70**: 3108-3110.

Smith GP. 1974. Unequal crossover and the evolution of multigene families. *Cold Spring Harb Sym* **38**: 507-513.

Souza RF, Appel R, Yin J, Wang S, Smolinski KN, Abraham JM, Zou TT, Shi YQ, Lei J, Cottrell J et al. 1996. Microsatellite instability in the insulin-like growth factor II receptor gene in gastrointestinal tumours. *Nat Genet* **14**: 255-257.

Strand M, Earley MC, Crouse GF, Petes TD. 1995. Mutations in the MSH3 gene preferentially lead to deletions within tracts of simple repetitive DNA in Saccharomyces cerevisiae. *P Natl Acad Sci USA* **92**: 10418-10421.

Strand M, Prolla TA, Liskay RM, Petes TD. 1993. Destabilization of tracts of simple repetitive DNA in yeast by mutations affecting DNA mismatch repair. *Nature* **365**: 274-276.

Streisinger G, Okada Y, Emrich J, Newton J, Tsugita A, Terzaghi E, Inouye M. 1966. Frameshift mutations and the genetic code. *Cold Spring Harb Sym* **31**: 77-84.

Streisinger G, Owen J. 1985. Mechanisms of spontaneous and induced frameshift mutation in bacteriophage T4. *Genetics* **109**: 633-659.

Subramanian S, Mishra RK, Singh L. 2003. Genome-wide analysis of microsatellite repeats in humans: their abundance and density in specific genomic regions. *Genome Biol* **4**: R13.

Sun JX, Helgason A, Masson G, Ebenesersdottir SS, Li H, Mallick S, Gnerre S, Patterson N, Kong A, Reich D et al. 2012. A direct characterization of human mutation based on microsatellites. *Nat Genet* **44**: 1161-1165.

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

Tautz D, Trick M, Dover GA. 1986. Cryptic simplicity in DNA is a major source of genetic variation. *Nature* **322**: 652-656.

The HDF Group. 1997-2014. Hierarchical Data Format, version 5.

Thibodeau SN, Bren G, Schaid D. 1993. Microsatellite instability in cancer of the proximal colon. *Science* **260**: 816-819.

Tran HT, Keen JD, Kricker M, Resnick MA, Gordenin DA. 1997. Hypermutability of homonucleotide runs in mismatch repair and DNA polymerase proofreading yeast mutants. *Mol Cell Biol* **17**: 2859-2865.

Tsao JL, Tavare S, Salovaara R, Jass JR, Aaltonen LA, Shibata D. 1999. Colorectal adenoma and cancer divergence. Evidence of multilineage progression. *Am J Pathol* **154**: 1815-1824.

Tsao JL, Zhang JS, Salovaara R, Li ZH, Jarvinen HJ, Mecklin JP, Aaltonen LA, Shibata D. 1998. Tracing cell fates in human colorectal tumors from somatic microsatellite mutations: evidence of adenomas with stem cell architecture. *Am J Pathol* **153**: 1189-1200.

Venables WN, Ripley BD. 2002. *Modern Applied Statistics with S*. Springer, New York.

Wang S, Lewis CM, Jakobsson M, Ramachandran S, Ray N, Bedoya G, Rojas W, Parra MV, Molina JA, Gallo C et al. 2007. Genetic variation and population structure in native Americans. *PLOS Genet* **3**: e185.

Wasserstrom A, Adar R, Shefer G, Frumkin D, Itzkovitz S, Stern T, Shur I, Zangi L, Kaplan S, Harmelin A et al. 2008. Reconstruction of cell lineage trees in mice. *PLOS ONE* **3**.

Weber JL. 1990. Informativeness of human (dC-dA)n.(dG-dT)n polymorphisms. *Genomics* **7**: 524-530.

Weber JL, May PE. 1989. Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. *Am J Hum Genet* **44**: 388-396.

Weber JL, Wong C. 1993. Mutation of human short tandem repeats. *Hum Mol Genet* **2**: 1123-1128.

Weissenbach J, Gyapay G, Dib C, Vignal A, Morissette J, Millasseau P, Vaysseix G, Lathrop M. 1992. A second-generation linkage map of the human genome. *Nature* **359**: 794-801.

Wierdl M, Dominska M, Petes TD. 1997. Microsatellite instability in yeast: dependence on the length of the microsatellite. *Genetics* **146**: 769-779.

Wierdl M, Greene CN, Datta A, Jinks-Robertson S, Petes TD. 1996. Destabilization of simple repetitive DNA sequences by transcription in yeast. *Genetics* **143**: 713-721.

Xu X, Peng M, Fang Z. 2000. The direction of microsatellite mutations is dependent upon allele length. *Nat Genet* **24**: 396-399.

Zhao X, Leotta A, Kustanovich V, Lajonchere C, Geschwind DH, Law K, Law P, Qiu S, Lord C, Sebat J et al. 2007. A unified genetic theory for sporadic and inherited autism. *P Natl Acad Sci USA* **104**: 12831-12836.

Zhu Y, Queller DC, Strassmann JE. 2000a. A phylogenetic perspective on sequence evolution in microsatellite loci. *J Mol Evol* **50**: 324-338.

Zhu Y, Strassmann JE, Queller DC. 2000b. Insertions, substitutions, and the origin of microsatellites. *Genet Res* **76**: 227-236.

Zhuchenko O, Bailey J, Bonnen P, Ashizawa T, Stockton DW, Amos C, Dobyns WB, Subramony SH, Zoghbi HY, Lee CC. 1997. Autosomal dominant cerebellar ataxia (SCA6) associated with small polyglutamine expansions in the alpha 1A-voltage-dependent calcium channel. *Nat Genet* **15**: 62-69.