Spontaneous mutation rates and spectra with and without the influence of natural selection in *Daphnia pulex*

Jullien Flynn
Department of Biology
McGill University, Montreal

July 2016

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Master of Science

TABLE OF CONTENTS

ABSTRACT

Spontaneous mutations are the ultimate source of genetic variation, which can generate phenotypic variation upon which natural selection can act. Understanding the rates, patterns, and fitness effects of mutation is essential to many fields of biology, thus several studies have attempted to investigate this fundamental phenomenon over the years. However, knowledge is still limited regarding the mutation rates in most organisms as well as the way selection acts on new mutations in a population. My thesis seeks to increase the understanding of the evolutionary phenomena of mutation and selection by analysing the genomes of mutation accumulation (MA) lines of *Daphnia pulex* maintained under selection-minimized conditions for many generations as well as isolates from a laboratory population that was founded with the same asexual progenitor and was maintained with selection acting throughout the course of the experiment. This unique experimental setup allows comparison of the rates, types, and patterns of mutations accumulated in conditions with and without selection. The *Daphnia* were propagated asexually, which allowed the detection of new mutations in a heterozygous state as well as large-scale mutations that result in loss of heterozygosity (LOH). Whole genome sequencing of 24 MA lines facilitated the detection of 477 single nucleotide mutations, and I found that the overall mutation rate in *Daphnia* is similar to that of other metazoans. One MA line experienced a massive LOH event that caused complete homozygosity across an entire chromosome (3% of the genome), resulting from a large gene conversion event. I also sequenced six isolates from the laboratory population and found fewer mutations than expected, demonstrating that purifying selection was acting strongly in order to purge harmful mutations that decrease fitness. Surprisingly though, the population maintained a high level of genetic diversity, with four distinct lineages from only six individuals. This observed pattern of high diversity was likely driven by balancing selection. My work challenges the assumption that selection is inefficient in asexual populations, provides an example of high diversity maintenance, and provides insight into the entire spectrum and implications of mutation in *Daphnia*.

RÉSUMÉ

Les mutations spontanées sont la source ultime de toute la diversité génétique et fournissent de potentielles modifications phénotypiques, sur laquelle la sélection naturelle peut agir. La compréhension des taux, des types de mutations et de leurs effets sur la valeur sélective est essentielle pour beaucoup des domaines de la biologie. Ainsi, quelques études ont cherché à faire la lumière sur ce phénomène fondamental au cours des dernières années. Cependant, les connaissances sont limitées au sujet des taux de mutations chez la plupart des organismes, et la façon dont la sélection naturelle se comporte sur de nouvelles mutations dans une population. Ma thèse tente de combler ces insuffisances en analysant les génomes de lignées de mutation accumulation (MA) de *Daphnia pulex* qui étaient maintenues dans des conditions avec un minimum de sélection, et de plus des individus d'une population qui était fondée par le même ancêtre et maintenue en laboratoire avec sélection pendant la duration de l'expérimente. Ce protocole expérimental permet de comparer les taux, types et distribution des mutations accumulées dans des conditions sans sélection, par rapport aux conditions avec sélection. Les daphnies ont été élevées pour qu'elles se reproduisent de manière asexuée, ce qui a permis la détection des nouvelles mutations étant hétérozygote et également des mutations à grande échelle qui causent une perte d'hétérozygotie (PDH). En faisant la séquençage de 24 lignées MA, j'ai découvert 477 mutations ponctuelles et établi que le taux de mutations des daphnies ressemble à cela des autres métazoaires. Une lignée MA a subi un nombre énorme de PDH d'un chromosome complet (3 % du génome). En plus, j'ai séquencé six individus de la population et découvert moins de mutations que prévu, ce qui démontre que la sélection purificatrice était forte afin d'éliminer des mutations nuisibles qui diminuent la valeur sélective. Cependant, étonnamment, la population a maintenu un niveau élevé de diversité génétique, avec quatre lignées indépendantes entre six individus. Cela a probablement été provoqué par la sélection diversificatrice. Les conclusions de cette thèse contestent l'hypothèse que la sélection naturelle est inefficace dans les populations asexuées, fournissent un cas du maintien de la diversité, et fournissent un aperçu de la diversité et des conséquences des mutations chez la daphnie.

ACKNOWLEDGEMENTS

I first and foremost would like to thank my supervisor, Melania Cristescu for her support, guidance, and confidence in me. She initially took me in as an enthusiastic undergraduate student volunteer, and later provided me with opportunities as a research assistant, Honour's student, and Master's student. She helped to nurture my passion for evolutionary genomics and I am very grateful for my experiences I had in her lab. I will truly miss working with her as I move on with my academic career. I am also grateful to my committee member and co-author Daniel Schoen for his willingness to help with this project, his always constructive feedback, and for his encouraging advice. I would also like to thank my other committee member Graham Bell for useful discussions and advice on the direction of this project. Additionally, I thank my fellow lab members, who provided company, support, and help with anything I ever asked for: Genelle Harrison, Alessandra Loria, Katie Millette, Sarah Finlayson, Tiffany Chin, Guang Zhang, and James Bull. I would especially like to thank post-doc, co-author, and mentor Frederic Chain who helped me to become comfortable with the bioinformatics-heavy project I took on, for always providing suggestions when I felt stuck, and for giving me advice not only about technical and conceptual aspects of the research but on general advice about academic and career decisions.

This thesis would have not been possible without the help of many undergraduate student volunteers and work study students who helped maintain the MA lines and assisted with laboratory work, including Anna Abraham, Hadi Assemani, Elaine Bai, Hailee Beckenstien, Monica Bhuller, Jasmin Chahal, Isaac Eckert, Vanessa Fulginiti, Ghizlane Gzili, Kirusanthy Kaneshwaran, Daniel Kapustin, Jonathan Leung, Tudor Mihai, Assil Tabbara, and Ivan Zhang. Thank you to Jeremy Briand and Marc-Olivier Beausoleil for helping edit the French version of the abstract.

I would also like to thank important people in my life external to the academic environment for always providing love, support, and other fun activities to maintain a balance in my life. I thank my family, particularly my mom, for her continued support and genuine interest and enthusiasm about my research and chosen career path. I thank my partner for listening to me ramble about my research, listening to me practice academic presentations totally outside of his field, and for providing love, support, and happiness throughout the completion of my degree.

Finally, I thank my fellow teammates (i.e. best friends) and coach from the McGill varsity cross country and track teams, for providing valuable friendships and also adding balance to my life – throughout not only my Masters degree but also for the preceding years that led me here.

CONTRIBUTION OF AUTHORS

JF designed the project, performed the laboratory work, performed the bioinformatics analysis (designed and troubleshooted the workflow, wrote scripts), and wrote the thesis. FC helped with the bioinformatics analysis including writing scripts and troubleshooting the workflow, and also provided feedback on various stages of the thesis. DS provided feedback on the bioinformatics workflow, and on various stages of the thesis. MC helped design the project, provided feedback on the bioinformatics workflow, provided feedback on various stages of the thesis, and provided funding for the project.

GENERAL INTRODUCTION

Mutations are modifications in the genome sequence or structure and are ubiquitous across all organisms and agents with replicating genetic material, from viruses to mammals. Spontaneous mutations can originate from a variety of causes such as replication errors, damage repair failure, oxidative stress, and other regular cellular processes (Smith 1992). Mutations occur throughout the lifetime of an organism and in all cells, but it is the ones in the germ line that are transmitted from one generation to the next and have implications for population fitness and evolution (Lynch 2010). All the genetic variation we see among all organisms that exist today ultimately originated from mutation, thus properties and theories related to mutation carry a high level of importance in many fields of biology. For example, concepts like understanding the genetic basis of evolutionary transitions (Ronshaugen et al. 2002; Boggs et al. 2009), estimating population genetic parameters such as effective population size (Lynch and Conery 2003), and understanding the mutational decline associated with small populations (Lynch et al. 1995), are all rooted in mutation. Other applications include disease genetics (de Ligt et al. 2013), estimating divergence time between species (Kumar 2005), and the evolutionary origin of sexual reproduction (Kondrashov 1988). For years, researchers have been interested in studying the rates that mutations occur, their effects on fitness, and implications for evolution (Haldane 1937; Mukai 1964; Crow and Simmons 1983; Lynch 1988; Huang et al. 2016).

The experimental nature of studies investigating mutation rates has evolved over the years as methodology and technology have progressed. Early studies used “reporter genes” to estimate mutation rates mainly in microbial species, whereas many individuals were screened for mutations in specific genes that caused a known phenotypic change (reviewed in Drake et al. 1998). However, this approach has limitations since the rate of mutation in a single gene is not necessarily representative of the global rate across the genome, and synonymous mutations or substitutions to a similar amino acid may be missed (Baer et al. 2007). Perhaps the most widespread approach is that of mutation accumulation (MA, reviewed in Halligan and Keightley 2009), where a single genotype of the species of interest is propagated for many generations with frequent bottlenecks, thereby minimizing natural selection. In this way almost all mutations are allowed to accumulate, except for lethality or sterility inducing mutations. Rates can then be estimated based on changes in the phenotype or DNA sequence among sublines. Since most mutations have deleterious effects (Mukai 1964; Keightley and Lynch 2003), the average fitness of MA lines tends to decline over time (Bataillon 2000; Eyre-Walker and Keightley 2007; Schaack et al. 2013). Methods have been developed that estimate the mutation rate based on phenotypic measurements of fitness and variance in fitness among lines (e.g. the Bateman-Mukai method; Bateman 1959; Mukai 1964). These methods were used commonly in early MA studies, and even still today (Houle et al. 1992; Keightley and Caballero 1997; Hall et al. 2013). Direct estimates of mutation rates became possible as DNA sequencing emerged, first by sequencing portions of the genomes, such as mitochondrial genomes (Denver et al. 2000), microsatellite loci (Seyfert et al. 2008), or even many loci across the nuclear genome (Denver et al. 2004). In the current era of high-throughput sequencing, whole-genome sequencing (WGS) of MA lines can be used to quantify the genome-wide accumulated mutations and estimate mutation rates. An extension of the MA-WGS approach is to perform parent-offspring (or pedigree) sequencing, in which the parents and several progeny are sequenced in order to detect *de novo* mutations in the offspring (e.g., Roach et al. 2010; Keightley et al. 2014; Keightley et al. 2015). Since this approach only involves one generation of propagation, it avoids some limitations that exist in MA studies. Parent-offspring sequencing can be carried out on organisms that are not simple to maintain in the laboratory for long periods of time. Additionally, it avoids biases in the types of mutations that can be observed with inbreeding, which is used with MA lines to propagate sexual organisms for multiple generations (Keightley et al. 2015). However, in one generation it is unlikely to observe rare mutation types in the offspring, such as large-scale chromosomal changes, which are more likely to be detected in MA experiments that last hundreds of generations (Schrider et al. 2013). Moreover, both of these WGS approaches have generated mutation rate estimates for a variety of species (Table i, Appendix I).

New mutations arrive in a population at a rate proportional to the fundamental mutation rate, but their fate – i.e. whether the mutation is lost, maintained, or increases in frequency – is mainly driven by the effective population size (genetic drift) and the effect the mutation has on the organism's fitness (selection) (Hartl and Clark 1997). It is known that the fitness effects of mutations occupy a distribution ranging from deleterious to neutral to beneficial, however the shape of this distribution is not known for most species (Eyre-Walker and Keightley 2007). More importantly, empirical work on the action of selection on newly arising mutations has been limited, although theory has been developed for decades (Fisher 1930; Orr 2005). One way to obtain a better understanding of selection on new mutations would be to compare the rate that mutations occur versus the "realized rate" that make it through selection in a population. This comparison is difficult to do with both laboratory and natural populations. Laboratory experimental evolution studies – which start with multiple replicate populations from the same ancestral genotype and impose an environmental stress on the populations – can allow genomic changes to be observed through a period of strong selection (reviewed in Lang and Desai 2014). Although these studies can offer insight into the evolutionary dynamics of mutations, since these populations are typically exposed to high levels of stress, the studies are focused on the fixation dynamics of beneficial mutations (Tenaillon et al. 2016). On the other hand, natural populations that are generally assumed to be well adapted to their current environment can be used to understand the selection dynamics of new mutations with deleterious effects (purifying selection). The mutation rate and spectrum of a particular species (derived from MA lines) can be compared to the genetic variants in natural populations of the same species, and purifying selection is inferred if less genetic variation is observed than expected based on the mutation rate. Keith et al. (2015) used this approach and found a strong signature of purifying selection on large-scale deletions and duplications in natural populations of *Daphnia*. Similarly, Huang et al. (2016) studied genetic and phenotypic variation in *Drosophila* MA lines compared to natural isolates and found that selection was maintaining low variance in these traits in the wild.

Both of the studies on natural populations of metazoans mentioned above demonstrated signatures of selection reducing the genetic variation in the population. In fact, both the processes of genetic drift and natural selection tend to remove variation from populations, however some populations maintain high levels of genetic variation (Hartl and Clark 1997; Charlesworth and Hughes 2000; Fitzpatrick et al. 2009). Genetic variants can be maintained by some mechanism of balancing selection, such as negative frequency-dependent selection, in which the relative fitness of a genotype with a distinct role decreases if its frequency in the population increases (Mitchell-Olds *et al*. 2007). Frequency-dependent selection has been reported as a mechanism to maintain variation in both experimental evolution studies (Elena and Lenski 1997; Kazancioglu and Arnqvist 2014) and natural populations (Fitzpatrick et al. 2009).

A factor that affects the efficiency of selection to remove deleterious mutations and fix beneficial ones is recombination (or lack of). Because recombination can break up genomic linkage between new mutations and the rest of the genome to allow selection to act independently on different loci, selection should be more efficient in recombining populations (Hill and Robertson 1966). Therefore, theory predicts that the efficiency of selection is reduced in asexual populations, where recombination associated with meiosis does not occur (Kondrashov 1988; Otto and Lenormand 2002). Some empirical studies have in fact found that asexual lineages have a higher mutation load than lineages of the same or related species that reproduce sexually (Paland and Lynch 2006). However, it is difficult to make confident inferences based on natural populations when the initial genetic background and selective history of the population are not known and the *de novo* mutations are not observed in real time.

The understanding of various topics in evolutionary biology would be aided by knowledge of the characteristics of mutation – the fundamental source of genetic variation. The objectives of this thesis are to (i) quantify the mutation rates and describe the mutation spectra and patterns in *Daphnia pulex*, and (ii) explore the effects of selection on new mutations – specifically to make inferences on the fitness effects of mutation, the strength of selection, the patterns of genetic diversity, and the implications of mutation and selection in asexual species. In order to do this, *Daphnia* mutation accumulation lines were propagated for many generations without selection, and a non-MA population founded by the same asexual progenitor clone was maintained for over five years with selection acting. I performed whole-genome sequencing on 30 genomes including both MA lines as well as isolates from the non-MA population, and compared the rates and spectra of mutations between these two environments. Additionally, I sequenced four additional MA lines that had obvious declines in fitness to investigate whether the types and patterns of mutations could help explain their decline in fitness.

References

Baer CF, Miyamoto MM, Denver DR. 2007. Mutation rate variation in multicellular eukaryotes: causes and consequences. Nat Rev Genet. 8:619-631.

Bataillon T. 2000. Estimation of spontaneous genome-wide mutation rate parameters: whither beneficial mutations? Heredity 84:497-501.

Boggs NA, Nasrallah JB, Nasrallah ME. 2009. Independent S-Locus Mutations Caused Self-Fertility in Arabidopsis thaliana. Plos Genetics 5:e1000426.

Charlesworth B, Hughes KA. 2000. The maintenance of genetic variation in life-history traits. In: Evolutionary Genetics: from Molecules to Morphology (eds. Singh, R.S. & Krimbas, C.B.), Cambridge University Press, Cambridge, UK, pp. 369–392.

Crow J, Simmons M. 1983. The mutation load in Drosophila. The Genetics and Biology of Drosophila. 3:1-35.

de Ligt J, Veltman JA, Vissers LELM. 2013. Point mutations as a source of de novo genetic disease. Curr Opin Genet Dev. 23:257-263.

Denver DR, Morris K, Lynch M, Thomas WK. 2004. High mutation rate and predominance of insertions in the Caenorhabditis elegans nuclear genome. Nature 430:679-682.

Denver D, Morris K, Lynch M, Vassilieva L, Thomas W. 2000. High direct estimate of the mutation rate in the mitochondrial genome of *Caenorhabditis elegans*. Science 289:2342-2344.

Drake JW, Charlesworth B, Charlesworth D, Crow JF. 1998. Rates of spontaneous mutation. Genetics 148:1667-1686.

Elena S, Lenski R. 1997. Long-term experimental evolution in *Escherichia coli* .7. Mechanisms maintaining genetic variability within populations. Evolution 51:1058-1067.

Eyre-Walker A, Keightley PD. 2007. The distribution of fitness effects of new mutations. Nat Rev Genet. 8:610-618.

Fisher RA. 1930. The genetical theory of natural selection: a complete variorum edition. Oxford University Press.

Haldane J. 1937. The effect of variation of fitness. Am Nat. 337-349.

Hall DW, Fox S, Kuzdzal-Fick JJ, Strassmann JE, Queller DC. 2013. The rate and effects of spontaneous mutation on fitness traits in the social amoeba, *Dictyostelium discoideum*. G3 (Bethesda). 3:1115-1127.

Halligan DL, Keightley PD. 2009. Spontaneous mutation accumulation studies in evolutionary genetics. Ann. Rev. Ecol. Evol. Syst. 40:151-172.

Hartl DL, Clark AG, Clark AG. 1997. Principles of population genetics. Sinauer associates Sunderland.

Houle D, Hoffmaster D, Assimacopoulos S, Charlesworth B. 1992. The Genomic Mutation-Rate for Fitness in Drosophila. Nature 359:58-60.

Huang W, Lyman RF, Lyman RA, Carbone MA, Harbison ST, Magwire MM, Mackay TF. 2016. Spontaneous mutations and the origin and maintenance of quantitative genetic variation. Elife 5:e14625

Kazancıoğlu E, Arnqvist G. 2014. The maintenance of mitochondrial genetic variation by negative frequency‐dependent selection. Ecol Lett.17:22-27.

Keightley PD, Lynch M. 2003. Toward a realistic model of mutations affecting fitness. Evolution 57:683-685.

Keightley PD, Caballero A. 1997. Genomic mutation rates for lifetime reproductive output and lifespan in *Caenorhabditis elegans*. Proc Natl Acad Sci USA. 94:3823-3827.

Keightley PD, Ness RW, Halligan DL, Haddrill PR. 2014. Estimation of the spontaneous mutation rate per nucleotide site in a *Drosophila melanogaster* full-sib family. Genetics 196:313-320.

Kondrashov A. 1988. Deleterious mutations and the evolution of sexual reproduction. Nature 336:435-440.

Kumar S. 2005. Molecular clocks: four decades of evolution. Nat Rev Genet. 6:654-662.

Lang GI, Desai MM. 2014. The spectrum of adaptive mutations in experimental evolution. Genomics 104:412-416.

Lynch M. 1988. The rate of polygenic mutation. Genet Res. 51:137-148.

Lynch M, Conery JS. 2003. The origins of genome complexity. Science 302:1401-1404.

Lynch M, Conery J, Burger R. 1995. Mutation accumulation and the extinction of small populations. Am Nat. 146:489-518.

Lynch M. 2010. Rate, molecular spectrum, and consequences of human mutation. Proc Natl Acad Sci USA. 107:961-968.

Mitchell-Olds T, Willis JH, Goldstein DB. 2007. Which evolutionary processes influence natural genetic variation for phenotypic traits? Nature Rev. Gen. 8:845-856.

Mukai T. 1964. The genetic structure of natural populations of *Drosophila melanogaster*. I. Spontaneous mutation rate of polygenes controlling viability. Genetics 50:1-19.

Orr HA. 2005. The genetic theory of adaptation: a brief history. Nature Rev. Gen. 6:119-127.

Otto SP, Lenormand T. 2002. Resolving the paradox of sex and recombination. Nature Rev. Gen. 3:252-261.

Paland S, Lynch M. 2006. Transitions to asexuality result in excess amino acid substitutions. Science 311:990-992.

Roach JC, Glusman G, Smit AF, Huff CD, Hubley R, Shannon PT, Rowen L, Pant KP, Goodman N, Bamshad M et al. 2010. Analysis of genetic inheritance in a family quartet by whole-genome sequencing. Science 328:636-639.

Ronshaugen M, McGinnis N, McGinnis W. 2002. Hox protein mutation and macroevolution of the insect body plan. Nature. 415:914-917.

Seyfert AL, Cristescu ME, Frisse L, Schaack S, Thomas WK, Lynch M. 2008. The rate and spectrum of microsatellite mutation in *Caenorhabditis elegans* and *Daphnia pulex*. Genetics 178:2113-2121.

Smith KC. 1992. Spontaneous mutagenesis: experimental, genetic and other factors. Mutation Research/Reviews in Genetic Toxicology. 277:139-162.

Tenaillon O, Barrick JE, Ribeck N, Deatherage DE, Blanchard JL, Dasgupta A, Wu GC, Wielgoss S, Cruveiller S, Medigue C. 2016. Tempo and mode of genome evolution in a 50,000-generation experiment. Biorxiv 036806.

Spontaneous mutation accumulation in *Daphnia pulex* in selection-free versus competitive environments

Jullien M. Flynn, Frédéric J.J. Chain, Daniel J. Schoen, and Melania E. Cristescu

McGill University, Department of Biology, Montreal, Quebec, Canada, H3A 1B1

Running head: Mutation and selection in *Daphnia*

Keywords: mutation rates, selection, loss of heterozygosity, gene conversion, read mismapping

Corresponding author:

Jullien M. Flynn

1205 ave Docteur Penfield rm N6/9

Montreal, Quebec, Canada, H3A 1B1

514-398-1622

jullien.flynn@mail.mcgill.ca

This manuscript was submitted to the journal *Molecular Biology and Evolution* on June 30, and is currently being peer reviewed.

ABSTRACT

Understanding the rates, spectra, and fitness effects of spontaneous mutations is fundamental to answering key questions in evolution, molecular biology, disease genetics and conservation biology. To estimate mutation rates and evaluate the effect of selection on new mutations, we propagated mutation accumulation (MA) lines of *Daphnia pulex* for more than 82 generations and maintained a non-MA population that experienced selection. Both experiments were seeded with the same obligate asexual progenitor clone. By sequencing 30 genomes and implementing a series of validation steps that informed the bioinformatic analyses, we identified a total of 477 single nucleotide mutations (SNMs) in the MA lines, corresponding to a rate of 2.30×10^{-9} (95% CI $1.90 – 2.70 \times 10^{-9}$) per nucleotide per generation. The high overall loss of heterozygosity (LOH) rate of 4.82×10^{-5} per site per generation was mainly due to a massive gene conversion event spanning an entire chromosome (~6 Mb). In the non-MA population, we found significantly fewer mutations than expected based on the rate derived from the MA experiment, indicating purifying selection was likely acting to remove new deleterious mutations. We additionally observed a surprisingly high level of genetic variability in the non-MA population, which we propose to be driven by balancing selection. Our findings suggest that both positive and negative selection on new mutations is powerful and effective in a strictly clonal population.

INTRODUCTION

Mutation is the ultimate source of genetic variation and a fundamental component of evolution. Knowledge of the rates, types, and genome-wide patterns of mutations in a species is essential to understanding many biological phenomena, such as the nature of genetic diversity in populations (Johnson and Barton 2005), genetic diseases (de Ligt et al. 2013), adaptation to changing conditions (Latta et al. 2013), divergence time between related species (Kumar 2005), and the evolution and maintenance of sexual reproduction (Kondrashov 1988). Although it is generally accepted that most mutations that occur in functional regions of the genome are deleterious (Keightley and Lynch 2003), the relationship between mutation, fitness, and selection is poorly understood. Especially missing is information about the rate of spontaneous mutations as well as their fate when exposed to natural selection. These factors influence whether new variants disappear, persist, or increase in frequency in a population, for example leading to adaptation to new challenges, maintenance, or mutational decline in fitness.

Mutations are inherently difficult to study firstly because they are rare and secondly because deleterious mutations are often purged by selection in natural populations (Kondrashov and Kondrashov 2010). For these reasons, mutation accumulation (MA) lines have often been used to study spontaneous mutation (Halligan and Keightley 2009). MA experiments begin with a single progenitor or inbred lineage that is replicated among independent lines that are propagated forward for many generations. The effects of selection are greatly reduced due to population bottlenecks imposed on each line every generation, drastically reducing the effective population size and allowing all but the lethal and sterility-causing mutations to accumulate. Past studies have relied upon phenotypic decline and divergence (e.g. Houle et al. 1992) or sequencing of selected portions of the genome (e.g. Denver et al. 2004) of MA lines to estimate mutation rates, but these methods have limitations and require potentially problematic assumptions (Keightley and Eyre-Walker 1999; Lynch et al. 2008). With the application of whole-genome sequencing (WGS) technology, we can now obtain a more detailed view of the mutational process to examine the mutational divergence of the whole genomes of MA lines. Several studies using this approach have provided single nucleotide mutation (SNM) estimates, for example from *Dictostelium*, 2.9 x 10^{-11} per nucleotide per generation (Saxer et al. 2012) to mouse, 5.4 x 10^{-9} (Uchimura et al. 2015). In addition, such studies have started to provide indications regarding the extent to which mutation rates vary across divergent taxa (Lynch et al. 2008) and different genetic backgrounds (Ness et al. 2015), as well as selection on the mutation rate itself (Saxer et al. 2012; Sung et al. 2012).

Despite recent progress in the field, a number of limitations in studying the full spectrum of mutations still remain. A major limitation stems from the inaccessibility of full genome assemblies for non-model organisms as well as challenging bioinformatic analysis required for WGS studies in cases where there are few genomic resources. Accurate mutation rate estimates are difficult to obtain without thorough analysis of the data combined with validation approaches (Li 2011, Keightley et al. 2015). Moreover, most MA studies have been conducted on inbred lines (e.g. *Mus, Caenorhabditis, Drosophila*, *Arabidopsis*), haploid organisms (e.g. *Saccharomyces*, *Schizosaccharomyces, Chlamydomonas*, *Dictyostelium*), or homozygous progenitors in which the only observable mutations are to a heterozygous state (*Saccharomyces*, Zhu et al. 2014). Mutational biases can exist when diploid naturally outcrossing organisms are forced to inbreed. As well, recessive lethal mutations cannot be observed in MA studies that use inbreeding to propagate lines. Furthermore, large-scale mutations associated with loss of heterozygosity (LOH) cannot be observed in homozygous genomes. LOH can arise from either hemizygous deletion, which results in only one copy of the allele across the deletion tract, or gene conversion, which results in two identical copies of an allele across the converted tract (Keith et al. 2015). Gene conversion occurs when genetic material is transferred from a donor region to a homologous acceptor region and this can take place between homologous chromosomes during meiotic crossing over or in the repair of a double-stranded break (Chen et al. 2010). Understanding the rates of LOH may be particularly important because this neglected class of mutations has been suggested to be an important component of the mutational process in the evolution of various organisms such as fungi and *Daphnia* (Omilian et al. 2006; Forche et al. 2011; Tucker et al. 2013) and is prevalent in human genetic diseases (Lemeta et al. 2004).

Even when an accurate mutation rate estimate can be made, further investigation is required to understand the fitness consequences of new mutations and how selection acts in a population to remove or maintain them. This can be achieved by comparing the rate and spectrum of mutations in selection-limited environments (e.g. MA line propagation conditions) to populations experiencing selection. However, such studies are very limited. Keith et al. (2015) compared the large-scale deletion and duplication rates from MA lines to genetic diversity in a natural population and inferred that selection purged many of these mutations from the population. Ideally though, the *de novo* mutations that make it through the "filter" of selection (the "realized mutation rate") should be compared to MA lines with the same genetic background, since the genetic background can influence mutation rates (Ness et al. 2015). Nevertheless, Keith et al. (2015) provided a valuable snapshot of the current standing variation as evidence of selection. Still missing is a study comparing mutations in selection-minimized MA lines to a population that was founded by the same progenitor and experienced selection – so new mutations can be identified and directly compared between conditions with and without selection, without the potential biases of different genetic backgrounds.

The microcrustacean *Daphnia,* commonly known as the water flea, is a suitable organism for studying mutation. It is simple to propagate in the laboratory and has a reference genome sequence available. Its moderate to high natural heterozygosity levels and ability to reproduce asexually allow the concurrent study of both point mutations in a heterozygous state as well as large-scale LOH events. Moreover, the system is ideal for drawing inferences about mutation patterns and their implications for the long-term persistence of asexual lineages. Although most

Daphnia are capable of cyclical parthenogenesis (alternating between asexual and sexual phases), some lineages are obligate asexual, having lost the ability to reproduce sexually because of meiosis-suppressing elements that cause meiosis to abort in females (Innes and Hebert 1988; Hiruta et al. 2010). Past studies have used *Daphnia* MA lines to investigate LOH and its importance in asexual species by genotyping microsatellite loci (Omilian et al*.* 2006; Xu et al. 2011), and by using WGS (Keith et al. 2015).

In this study we estimate the mutation rates for a broad spectrum of mutations in *Daphnia pulex,* and evaluate the influence of natural selection on the fate of *de novo* mutations in a population. To do this, we carried out an MA experiment, and in parallel maintained a non-MA population (non-bottlenecked population where selection could operate) that was founded by the same MA line progenitor. We sequenced a total of 30 *D. pulex* genomes and applied a data analytical approach using strict filtering alongside Sanger sequencing validation of putative mutations to address problems associated with mapping short reads to the reference genome that has high duplication levels. We estimate the mutation rates for base substitutions, insertions and deletions, and loss of heterozygosity events via large-scale deletion and gene conversion. Finally, we analyse the mutation rate and spectrum in our MA lines and compare it to the non-MA isolates, as well as to studies of mutation in *Daphnia* and other organisms.

RESULTS

Sequencing of 30 genomes: We sequenced and analyzed 30 *Daphnia pulex* genomes: 24 MA lines randomly selected from 50 lines (C01-C50) propagated asexually for an average of 82 generations, and six isolates (CC3, CC4, CC6, CC7, CC8, CC9) from a laboratory-maintained population. The population was initiated from the same asexual MA progenitor as the MA lines and was maintained for the duration of the MA experiment (over five years) in a 15L tank (N= 100-250 individuals) where selection was allowed to act (i.e. no single-progeny bottlenecks every generation). We obtained a total of 130 Gbp of sequencing data from the 30 genomes analyzed. Approximately 85% of reads successfully mapped to the reference genome. The average coverage per MA line and non-MA isolate ranged from 7.44 – 14.4x, except for two MA lines that were intentionally sequenced to a higher depth (19.0 and 20.0x). The increase in sequence depth was conducted to test the effect of higher sequence coverage on the mutation rate estimate. To increase the power of detecting mutations while minimizing false positives, we focused on high-confidence sites based on several filtering criteria. After filtering, we retained a total of 52,530,668 sites with an average coverage of 13x for each of the MA lines. For the non-MAs, we analyzed a total of 41,697,845 sites for each isolate with an average depth of 14x. We expected to observe three possible genotype changes for mutations: Hom-Het, Het-Het, and Het-Hom, where "Hom" represents homozygous, "Het" represents heterozygous, and "-" represents a change in genotype of a site from the ancestral state due to a mutation. Hom-Het and Het-Het mutations reflect changes to a single genomic position, while Het-Hom mutations reflect LOH and are expected to occur in consecutive stretches. To quantify the mutations accumulated since the common ancestor and estimate mutation rates, we used strict filtering of variant calls informed by several validation steps.

MUTATIONS IN MA LINES

Single nucleotide mutations and indels: We found a total of 477 SNMs (single nucleotide mutations) and 6 indels (insertions/deletions) across all lines (Table S2 and Table S3). Each MA line contained 7-36 SNMs and 0-1 indels (Table 1). This includes only Hom-Het and Het-Het mutations, which are the types of changes that we expect from mutations affecting a single genomic site. We only found one Het-Het mutation after filtering and manual inspection– this is not surprising given that ancestrally heterozygous sites comprise less than 1% of the genome. The overall SNM rate was 2.30 x 10^{-9} (95% CI 1.90 – 2.70 x 10^{-9}) per site per generation and the overall indel rate was 2.90 x 10^{-11} per site per generation. The SNM rate obtained in our study may be a slight overestimate because our Sanger sequencing validation of 19 mutations had a false positive rate of 21%, which is likely attributable to mapping errors that were not detected in our pipeline. Therefore, if one assumes that 21% of our estimated SNMs are false positives, a more accurate mutation rate estimate is 1.80 x 10^{-9} per site per generation. There were five sets of multinucleotide mutations, (MNMs, mutations occurring within 50 bp of each other) observed after filtering (26 sets passed our filtering algorithm but 21 of these were excluded after manual inspection, see Materials and Methods). Three of the five MNMs were in a single MA line, C08, and the others were in MA lines C34 and C37. All identified MNMs were pairs or triplets of SNMs either immediately adjacent or two to three nucleotides apart. Of the small indels detected, two were insertions and four were deletions. Across all lines, a total of 9 bp were inserted by a 2 and 7 bp insertion and a total of 22 bp were deleted by events 1-13 bp in size. We also obtained interpretable Sanger sequences for two indels and two sets of MNMs, and all were confirmed.

The most frequent base substitution was G:C → T:A (where ":" represents Watson-Crick base pairing), both in absolute number and conditional rate (Table 2). The second most frequent substitution was G:C → A:T, making the G|C → A|T mutation rate almost four times that of A|T → C|G (where "|" represents "or"). This made the equilibrium A+T genome composition (based on mutation alone) much higher than the observed A+T composition (80.9% versus 58%). We also detected a transition bias with transitions being 1.6 times more frequent than the null expectation in our Hom-Het SNMs; the transition to transversion ratio (Ts/Tv) was 0.81 compared to the null expectation of 0.5 if each base mutated to another with equal probability.

Distribution of mutations across different genomic regions: The number of mutations in intergenic, exonic and intronic regions was proportional to the composition of the genome (Fig 1). However, there was a *ca.* 5% enrichment of SNMs in intergenic regions and a matching deficit in exons, and this difference was statistically significant ($\chi^2 = 8.4$, df = 2, p = 0.015). There were 29 synonymous mutations and 93 missense substitutions, five substitutions that created stop codons, and one substitution that led to loss of a start codon. Indels were in intron or intergenic regions except for two 1-2 bp indels causing frameshift mutations. Ancestral heterozygosity among the 12 chromosomes varied between 0.54 and 1.22% after normalization based on the number of sites that mapped to each chromosome, with an average of 0.69% across the entire genome. The distribution of SNMs across chromosomes varied up to 2.5-fold, which was significantly more than expected from a random distribution (χ^2 = 22.96, df = 11, p < 0.018), with chromosome 9 exhibiting the highest mutation rate and the greatest deviation from random expectations. Regions containing ancestrally duplicated loci, indicated by sites that contain two alleles but not in the 1:1 ratio (as expected for a single heterozygous locus), were also most frequent on chromosome 9.

Loss of heterozygosity: We were conservative in specifying mutational LOH, defined here as regions containing multiple Het-Hom sites spanning > 100 bp with minimal interruption of heterozygous sites. The overall rate of LOH was 4.82 x 10^{-5} per heterozygous site per generation. However, this relatively high rate was driven by one large gene conversion event (see below), making the gene conversion rate several orders of magnitude higher than the hemizygous deletion rate (4.80 x 10^{-5} versus 8.21 x 10^{-8}). There were five moderately sized LOH events in four of the 24 MA lines (Table 3). We inferred that all moderately sized events were caused by deletions, indicated by lower normalized read depth in the affected line compared to the other lines (Table 3, Fig 2A). The minimum size of these events, defined as the region spanning the Het-Hom sites ranged from 181 – 2078 bp. The maximum size of these events, defined as the distance between the two heterozygous sites flanking the Het-Hom sites, ranged from 503 – 4012 bp. If one assumes that the deletion breakpoints occurred halfway between the minimum and maximum size, the average deletion size was 1892 bp. Four of these deletions affected genes. We checked one of the moderately sized LOH events with Sanger sequencing (in C08, Table 3, Table S1) and confirmed that there were no heterozygous sites across the region in the mutant MA line, where two other MA lines used as controls had many.

One MA line, C40, had an extremely high LOH rate of 1.19 x 10^{-3} per heterozygous site per generation. This line shows one large LOH tract, spanning a total of over 6 Mb on chromosome 11 (Fig 2B). The seven scaffolds that map to chromosome 11 span the entire linkage group (Xu et al. 2015a, Fig 3) and were each comprised of Het-Hom sites in C40 for almost all the ancestral heterozygous sites. We positively verified five regions spanning approximately 900 bp each using Sanger sequencing, including most physical regions on chromosome 11 (Fig 3). This confirmed that C40 contained no heterozygous sites across these regions, both in the generation that was used for whole-genome sequencing and also from an individual taken from 5 generations later in the experiment. The two control MA lines used for validation contained several to dozens of heterozygous sites across the same regions. Normalized read depth in C40 across chromosome 11 was not lower than the average of the other 23 lines

(Fig 2B), suggesting that this event resulted in homozygosity as opposed to hemizygosity, likely caused by ameiotic recombination (gene conversion). The same pattern of equal read depth was found in the 13 unmapped scaffolds that experienced LOH in C40 as well, so we infer that these scaffolds also belong to chromosome 11 and resulted from the same conversion event (supplementary text S1, Supplementary Material online). Additionally, complete homozygous deletions in C40 occurred in at least two locations near the middle of chromosome 11, totalling ~100 kb (Fig 2B, Fig 3). The large LOH event in C40 affected an estimated total of 1457 genes. No other LOH event associated with a different chromosome was found in C40.

Resequencing of two MA lines: We found that increasing the coverage to an average of ~20x (28x for the sites used for analysis) in two randomly-selected MA lines did not affect the estimated mutation rate of SNMs. Both before and after doubling sequence read depth, the SNM rate we estimated for C01 and C35 was similar to the average of the other lines (Table 1). This provides justification that the depth of coverage in the other 22 lines (~15x per line for the variant sites under analysis) was sufficient for accurately estimating mutation rates. Sequencing the two lines at higher depth also justified our exclusion of stand-alone Het-Hom sites and consecutive Het-Hom sites spanning a region <100 bp (see Materials and Methods); such Het-Hom tracts were found to be mostly read sampling artefacts associated with low depth of coverage when C01 and C35 were sequenced to a greater depth. The number of putative stand-alone Het-Hom sites decreased with higher depth from 446 to 363 in C01 and 727 to 352 in C35. Similarly, the number of small regions (<100 bp) consisting of consecutive Het-Hom sites decreased with higher depth from 27 to 14 in C01 and 76 to 7 in C35.

NON-MA POPULATION

Population diversity and realized SNMs: We found a total of 20 independent SNMs among the six isolates from the non-MA population. We identified four distinct lineages with no shared mutations among all of these lineages, although one lineage contained three related isolates with shared mutations (Fig 4). Since we cannot confidently estimate the number of generations of propagation that lineages within this population underwent, we calculated the realized SNM rate based on a range of possible generation numbers. The MA lines had undergone 101 generations at the time the non-MA isolates were collected, but because of overlapping generations and possible life history differences in a population setting, it is possible that the non-MAs underwent fewer generations. We calculated the minimum number of generations the non-MA lineages would have had to undergo in order for their estimated realized

SNM rate to be significantly lower than the MA lines. We found that for generation numbers as low as 40, the non-MA realized SNM rate was below the 0.05 quantile of the rate distribution calculated from permutations of MA lines (Fig 5). Even if the rate of false positive SNMs in the MA lines are taken into account, the non-MA rate would still be below the 0.05 quantile at 50 generations. It is unlikely that the non-MA population would have progressed only half the number of generations as the MA lines in 5 years, thus it is reasonable to assume that the non-MA population actually underwent more generations than 50 and had a significantly lower realized mutation rate.

Mutation spectrum and distribution: The Ts/Tv ratio of the non-MAs was 2.3, and this is significantly greater than that of the MA lines at 0.81 (p = 0.019). This is attributed to the proportionally higher number of C:G → T:A transitions opposed to C:G → A:T transversions (Table 2). The non-MAs had a slightly smaller proportion of SNMs in exons and a slightly greater proportion in intergenic regions than the MA lines (Fig 1). However, the difference in the distribution of SNMs among exons, introns, and intergenic regions between the non-MAs and the MA lines was not significant (p = 0.637, 0.361, 0.760, respectively), although there was limited statistical power given only 20 SNMs were detected in the non-MAs. Of the six mutations that occurred in exons, three were synonymous, two were missense, and one was nonsense. All of the three shared mutations (Fig 4) were annotated as upstream or downstream gene variants, with two being intronic and one being intergenic. We did not detect any indels or MNMs in the non-MAs, but this was not a significant deficit considering the low rate of these types of mutation (p = 0.995 and p = 0.991, respectively). We detected LOH in only one isolate, and this event was a deletion on chromosome 9. This event spanned a total of about 37 kb (not including the interrupting regions possibly due to scaffold misassembly) (Table 3). This deletion affected 12 protein-coding genes, and one pseudogene. The same isolate had another small deletion (~1 kb) on a scaffold that has not been mapped to a chromosome, so this may or may not be part of the same event (Table 3).

DISCUSSION

In this study, we quantified the rates and described the spectra of spontaneous mutations, including large-scale LOH mutations that are rarely detectable in MA experiments that use homozygous and inbred lines. We also investigated how selection acts on new mutations by comparing these results with the observed mutations in non-MA individuals from a non-bottlenecked population with the same genetic background. Our results support the assumption that selection is nearly absent during MA line propagation, with only a slight deficit of mutations in exons compared to intergenic regions. This slight deficit can be explained by mutations that cause sterility or mortality, which cannot be propagated. In comparison, the non-MA population exhibited strong signatures of both diversifying and purifying selection.

Point mutation rate and spectrum in MA lines: The estimated SNM rate of 2.30 x 10^{-9} per site per generation is similar but slightly lower than that found in a recent MA study of *Daphnia* (~4 x 10^{-9}, Keith et al. 2015), which may be explained by the strict mutation filtering pipeline we used. Our estimate is also close to estimates from the other metazoan MA studies including *Drosophila* (3.5 x 10^{-9}, Keightley et al. 2009; 5.49 x 10^{-9}, Schrider et al. 2013), *Caenorhabditis* (2.5 x 10^{-9}, Denver et al. 2009), and *Mus* (5.4 x 10^{-9}, Uchimura et al. 2015). The strikingly similarity of these SNM estimates suggests that the mutation rate is robust across the animal kingdom and is potentially under tight selection. Our indel rate of 2.89 x 10^{-11} per site per generation was lower than some estimates in other organisms (e.g. 2.35 x 10^{-10} in *Drosophila,* Schrider et al. 2013). However, it is possible that we may have underestimated the rate of small indels due to the relatively low detection power in variant calling software (Fang et al. 2014) and our particularly strict filtering regime.

We found five sets of multinucleotide mutations (MNMs), mutations closely clustered together that almost certainly represent a single event in which multiple bases are changed at once. They have commonly been found to occur at rates higher than expected if they were independent mutations (Schrider et al. 2013; Zhu et al. 2014; Keith et al. 2015). Several mechanisms have been proposed to cause MNMs, including error-prone polymerases or DNA repair machinery, or even one mutation causing a second (Schrider et al. 2011). Interestingly, three out of the five MNMs we found occurred in a single line (C08), suggesting this line may have incurred a mutation making it more susceptible to MNMs. MNMs accounted for 2.3% of SNMs in our study, similar to Schrider et al. (2013), which found 2.79% of SNMs in *Drosophila* to be accounted for by MNMs (Fig 6A). In contrast, Keith et al. (2015) found a very high rate of MNMs, accounting for 16% of SNMs in their asexual *Daphnia* MA lines (even after excluding the MA line they found to be an outlier) (Fig 6A). Because MNMs involve multiple mutations occurring in the same read, mapping artefacts can inflate MNM estimates when reads from distinct but similar loci map to the same position in one line. We manually inspected the BAM alignment files to remove artefactual MNMs, but before doing this we found a similarly high amount of MNMs as Keith et al. (2015) (Fig 6A).

We found the substitution spectrum to be highly A+T biased. There were high rates of both C:G → A:T transversions and C:G → T:A transitions occurring at almost equal rates. C:G → T:A transitions are often suggested to be caused by the spontaneous deamination of 5-methyl cytosine (Keith et al. 2015), but this bias has also been observed in eukaryotes without a methylation system (Behringer and Hall 2015). The even slightly higher rate of C:G → A:T transversions in the MA lines suggests that deamination of methylated cytosine is not the only cause of A+T substitution bias we observed here. The equilibrium A+T genome composition in the MA lines (from substitution alone) is estimated to be 80.9%, while the actual genome composition of the MA progenitor is 58% A+T. Therefore factors other than substitution must have shaped the genome composition of *Daphnia* populations in nature. A similar A+T bias was also found in the Keith et al. (2015) study, and is common in MA studies of other organisms (e.g. Ossowski et al. 2010). Possible explanations for this include selection against A+T increase or GC-biased gene conversion in nature (Galtier et al. 2001). We found a slight transition bias with a Ts/Tv of 0.81, close to what was found in yeast (Zhu et al. 2014) and *Drosophila* (Keightley et al. 2009), but less than the Ts/Tv of 1.52 found in the previous *Daphnia* study (Keith et al. 2015). This discrepancy is due to our higher rate of C:G → A:T transversions.

Loss of heterozygosity: One MA line (C40) experienced a massive LOH event across the entire chromosome 11. Het-Hom sites in C40 were found across all scaffolds that mapped to this chromosome (Xu et al. 2015a), spanning over 6 Mb. The large LOH event in C40 likely resulted from an ameiotic recombination event, along with homozygous deletions totalling approximately 100 kb on the affected chromosome. Generally such a large-scale event is expected to have a high impact on the fitness of the organism. However, at the time of sequencing, C40 had a generation number equal to the median of that of all lines. Eventually, C40 did go extinct (after a total of 109 generations), 29 generations after it was sequenced. Its eventual extinction may have been influenced by its complete homozygosity across an entire chromosome, possibly because of interaction effects of fixed deleterious alleles with new mutations or another unknown cause.

In other MA lines besides C40, LOH was less common with five smaller LOH events detected in four of the other 23 lines, all of which were likely due to hemizygous deletions. Because the LOH event in C40 was so large, our per site estimated rate of gene conversion was higher than the rate of hemizygous deletion, contrary to past studies (Xu et al. 2011, Keith et al.

2015). The hemizygous deletion rate we found is orders of magnitude lower than past studies (10^{-7} versus 10^{-5}). This could reflect biological differences in the divergent lineages of *Daphnia* studied, or inaccuracies stemming from various methodological approaches in the different studies. However, the overall LOH rate is strikingly similar to estimates derived from past studies in divergent lineages of *Daphnia* both by genotyping microsatellites (Omilian et al. 2006, Xu et al. 2011) and whole genome sequencing (Keith et al. 2015) (Fig 6B). Omilian et al. (2006) reported a similar sized and structured LOH event in their MA study of a divergent *D. pulex* lineage as we did in C40, which encompassed over half of a chromosome and was caused by gene conversion followed by internal deletions (Xu et al. 2011). We observed different patterns of LOH than Keith et al*.* (2015), which found many short gene conversion events (dozens of base pairs) and only one MA line with homozygous LOH tracts larger than 1 kb. We found that small regions of Het-Hom sites spanning <100 bp were not reproducible by Sanger sequencing, so we did not include them in our LOH rate estimates (see Materials and Methods for details), but estimates including them would still be comparable to past studies (Fig 6B).

LOH and the evolution of asexual species: Previous studies in *Daphnia* have found the chromosomes containing the meiosis-suppressing elements in obligate asexuals (chromosomes 8 and 9) to contain the highest levels of heterozygosity (Tucker et al. 2013) and *de novo* copy number variants (Keith et al. 2015). We found chromosome 9 to have the highest SNM rate, possibly explained by highly heterozygous regions exhibiting an elevated mutation rate (Yang et al. 2015). We also found that chromosomes 8 and 9 contained high levels of historical duplication, indicated by the highest proportion of sites with signatures of duplicated loci mapping; i.e. two alleles mapping in ratios not supporting a single heterozygous locus (e.g., 200 reads A and 40 reads C). The spread of asexuality is hypothesized to occur through backcrossing of males of obligate asexual lineages (meiosis is suppressed in females only), and requires transmission of the entire intact haplotype that contains the meiosis-suppressing elements (Xu et al. 2015b). Our evidence, along with supporting evidence from Keith et al. (2015), who investigated large deletions and duplications, suggest that chromosomes 8 and 9 have been prone to copy number variation throughout the history of evolution of obligate asexual *Daphnia pulex.* It is possible that the duplications present on these chromosomes may increase the chances that all the meiosis-suppressing elements on the haplotype are transmitted together, thus improving the efficiency of the spread of asexuality (Xu et al. 2015b). A hemizygous deletion occurred on chromosome 9 in a non-MA isolate, but it was not detected to be part of a duplicated region, and it did not affect any of the sites found to be associated with obligate asexuality (Xu et al. 2015b).

Past studies have noted the potential importance of LOH in asexual species and suggested that ameiotic recombination may occur at significant levels in obligate asexual lineages, possibly reducing linkage between alleles and allowing deleterious or beneficial alleles to be lost or fixed (Omilian et al. 2006, Xu et al. 2011). Our findings support this view, and we suggest that LOH is an important mutational input in the evolution of asexual lineages, but most of these events are likely neutral or deleterious. We found one massive gene conversion event, and considering it caused complete homozygosity across 6 Mb, and that the MA line harbouring it eventually went extinct, this was likely not a beneficial mutational event. The other LOH events we found were deletions resulting in hemizygosity, and these are estimated to not be beneficial because of loss of complementation of existing or future deleterious alleles on the hemizygous haplotype (Archetti 2004). The hypothesis that LOH events are often deleterious and removed by selection was supported by Keith et al. (2015), who found that the frequency of large scale deletions was lower in natural populations than what would be expected based on their rate estimates derived from MA lines. On the other hand, our observation of a hemizygous deletion in the one non-MA isolate that experienced selection suggests that these events can also have negligible effects on fitness, especially in a species that contains many duplicated genes. We note, however, that it is possible that the deletion in the non-MA isolate occurred recently, before selection could remove it.

Mutation accumulation with and without selection: The experimental design of this study, based on comparing mutations accumulated in the same initial genotype in a selection-minimized MA experiment versus a non-MA environment allowed us to study the effect of selection on new mutations. We found evidence for purifying selection in the non-MA population, with a significantly lower realized mutation rate than the rate observed in the MA lines. The significantly greater transition bias found in the non-MA isolates is of interest, as it was mainly driven by SNMs in noncoding regions. Typically a transition bias is assumed to occur because transitions have a lower probability of causing amino-acid changes than transversions (Wakeley 1996), but only one synonymous mutation occurred in the non-MAs, and this was caused by a transversion. The higher Ts/Tv was driven by an altered mutation spectrum in the non-MAs, with 70% of SNMs being C:G →T:A transitions, compared to 32% in the MA lines. The altered spectrum could therefore be caused by different environmental conditions influencing the mechanism of mutation (e.g. Jiang et al. 2014), or by the influence of selection on non-coding regions. Since the conditions in the non-MA tank and the MA lines were almost identical, we suggest that it is not the environment that altered the mutation spectrum but instead is a result of selection on noncoding cis-acting regulatory sequences. Bergman and Kreitman (2001) found that ~25% of non-coding sequences were under selective constraints in *Drosophila*, and a 2-fold transition bias was observed in these regions.

High diversity in the non-MA population: An unexpectedly high level of diversity was detected in the non-MA population, with four distinct lineages found among the six sampled individuals (Fig 4). The lack of shared mutations among the four lineages argues against a recent bottleneck, suggesting that the genetic diversity generated relatively early in the experiment was maintained. One might find this surprising, assuming purifying selection was the only dominant mode of selection acting. Such an observed pattern of diversity could have been caused by: (a) most mutations being neutral and therefore selection being negligible; (b) reduced effectiveness of selection in the obligate asexual population (Paland and Lynch 2006; Tucker et al. 2013); and/or (c) selection favouring divergence between clones. Since the significantly lower realized mutation rate in the non-MAs compared to the MAs suggests that purifying selection against newly arising mutations was acting, factors (a) and (b) are not likely driving the observed pattern. Additionally, if balancing selection was not acting, it is unlikely that distinct lineages would be maintained with minimal branching within the lineages. We suggest that initial diversifying selection among clones was responsible for generating distinct lineages, and negative frequency-dependent selection maintained these lineages. Within the lineages, we suggest that purifying selection and clonal interference are likely acting against additional new mutations that would branch the lineages further. Our findings of balancing selection are contrasting from those from a recent study of quantitative trait variation in *Drosophila.* Huang et al. (2016) found less genetic and phenotypic variation in wild isolates than expected based on MA lines, and from this inferred strong stabilizing selection.

Selection without recombination: It has been suggested that selection is inefficient in asexual populations, since recombination does not occur with sexual reproduction at every generation, thus new mutations, deleterious and beneficial, are effectively linked to the rest of the genome (Muller 1964). Tucker et al. (2013) found no evidence for purifying selection on amino-acid-altering substitutions in natural populations of *Daphnia*. In contrast, we not only found evidence for purifying selection, but also positive diversifying selection in our non-MA population. With a census population size of approximately 100-250, and an effective population size likely lower since the population was founded by a single asexual clone, selection would have to have been quite strong in order to shape the patterns we observed. Additionally, we did not detect ameiotic recombination in the non-MA isolates, which has the potential to make selection more efficient by fixing or purging deleterious or beneficial alleles. Given that we did not impose any specific environmental stressors on the population, the main source of selection may stem from competition for food. Strong selection implies that a significant proportion of mutations have an effect on fitness in *D. pulex.* This may be surprising considering this species contains high amounts of duplicated genes, which may buffer the fitness effects of new mutations (Gu et al. 2003; Conant and Wagner 2004), albeit this subject is controversial (Su et al. 2014). Our findings suggest that selection is strong enough on fitness-affecting mutations in order to purge deleterious ones and select for beneficial ones in an asexual population.

MATERIALS AND METHODS

Mutation accumulation lines: The *Daphnia pulex* progenitor of the MA lines was collected from Canard Pond (Lat. 42°12", Long. -82°98") located in Windsor, ON, Canada. Fifty replicate lines were derived from this individual and were cultured in 20 ml of FLAMES soft-water media (Celis-Salgado 2008). Lines were fed a mixture of three species of algae, *Ankistrodesmus sp., Scenedesmus sp.* and *Pseudokirchneriella sp* twice per week. The environment was kept at a constant temperature of 18° C, a humidity of 70%, and a lighting regime of 12 hours of light and 12 hours of dark. A single progeny individual from each of the lines was transferred to fresh media every generation (*ca* 11-13 days). Backup lines were kept in case of mortality or sterility of the focal individual. These were used in 6% of transfers, corresponding to an average of once every 16 generations per line. A total of 24 MA lines were randomly selected for sequencing after an average of 82 generations of propagation (ranging between 72 to 88 generations).

Non-MA population: At approximately the same time as the initiation of the MA lines, a population was founded by the same obligate asexual *D. pulex* clone that was the progenitor of the MA lines. These non-MA *Daphnia* were maintained in a 15 L tank with identical media, temperature and lighting conditions. The tank was cleaned monthly to remove debris and partially refresh the media. The tank was fed twice weekly with 50 mL of the same mixture of three species of algae as to the MA lines. At the time the MA lines had undergone an average of 101 generations, six individuals from random locations in the tank (after stirring) were isolated and sequenced. Since overlapping generations occur in the tank, the non-MA isolates likely underwent fewer generations than the MA lines. The non-MA isolates were prepared, sequenced and processed following the same procedures as the MA lines.

Sequencing and variant calling: Sample preparation, sequencing, and preprocessing steps are described in Supplementary Material online (supplementary text S2). Calling and processing of variants was done with GATK v.3.3.0 (DePristo et al. 2011). As in GATK's recommended procedures, we first used HaplotypeCaller to assign a putative genotype to each site along the genome for each line separately, followed by genotyping all lines simultaneously with GenotypeGVCFs. A file containing all variant calls was produced using GATK SelectVariants and was used for subsequent filtering and evaluation of the putative variants. Only the nuclear genome was considered, and ribosomal DNA was excluded because of the known complication of many copy numbers (Crease and Lynch 1991). We excluded regions of the genome that were annotated as repeat regions (Ensembl v.23), as these are prone to mapping errors. We also removed sites with an overall depth across all MA 24 lines >620, as these sites have approximately double the expected average coverage likely because of mapping of multiple loci to the same reference position.

Mutation filtering in the MA lines informed by Sanger sequencing validations: We applied rigorous filtering on the called mutations since characteristics of the *Daphnia* genome pose a number of bioinformatic challenges (Li 2011; Ribeiro et al. 2015). The current reference genome assembly is derived from the sister species *Daphnia arenata*, is comprised of over 5000 scaffolds, and contains high levels of duplication (Colbourne et al. 2011) as well as some misassembled regions (Xu et al. 2015a). Single nucleotide mutations (SNMs) and indels were first filtered using a combination of GATK and vcftools v0.1.11 (http://vcftools.sourceforge.net) in the manner described below. We considered only sites that had a genotype call for all lines a minimum average depth of 6 per line. This allowed us to maintain high sensitivity to detect new mutations in a heterozygous state and avoid the false detection of loss of heterozygosity due simply to sampling error at low-depth sites. Initially, putative SNMs and indels were separated and filtered by the parameters recommended by GATK based on read and mapping quality, strand bias, and location in read (supplementary text S3, Supplementary Material online). After this initial filtering of genotype calls, approximately 20,000 putative mutations remained in the MA lines.

We used Sanger sequencing to: (a) validate putative mutations identified at different stages of filtering in order to decide on our filtering regime (Table S1, Supplementary Material online) and design a custom algorithm to remove false positives; and (b) to validate a subset of our final confident set of mutations. Primers were designed around the mutations using Primer3 (Rozen and Skaletsky 1999) or PrimerView (O'Halloran 2015) in order to amplify a 600-900 bp region. For each putative mutation, we amplified the region from: (1) the exact genomic DNA that the libraries were prepared with for the putative mutant line; (2) genomic DNA collected from ~5 generations later in the putative mutant line; and (3) genomic DNA from two other independent lines as negative controls. Sequencing reactions were completed using BigDye Terminator v3.1 (ThermoFisher) and analyzed by Genome Quebec Innovation Centre at McGill University with a 3730xl DNA Analyzer (Applied Biosystems). Electropherograms were analyzed using CodonCode Aligner (CodonCode Corporation).

Sanger sequencing of a subset of mutations that passed the initial filters and had allele ratios indicative of true mutations (supplementary text S4, Supplementary Material online) revealed that none of the 19 mutations for which we obtained interpretable sequences for were reproducible (Table S1, Supplementary Material online). Further filtering was therefore implemented to reduce the false signals of mutation introduced by sequencing and mapping errors. We applied a custom algorithm to re-evaluate the genotype calls, informed by the Sanger sequencing results. The algorithm performed binomial tests at each site to determine (i) the ancestral genotype using allele depths across all lines; followed by (ii) the genotype of each MA line based on the expectation of the ancestral state. To implement the algorithm, we extracted the read information from the sites that passed our initial filters from the BAM alignment files (only reads with mapping quality of at least 20 and sites with base quality of at least 10 were used) using mpileup in samtools v0.1.19 (Li 2011). We then used read depth information for every line at each putative variant site that passed our initial filters to infer both: (i) the genotype of the MA line progenitor (referred to here as the "ancestral genotype"); and (ii) the genotype of every individual MA line. Only bi-allelic sites that had strong support for either homozygosity or heterozygosity in all MA lines based on allele depth were kept for further analysis (supplementary text S5, Supplementary Material online). Sanger sequencing of another 10 mutations was done at this stage, and still 90% of them were not reproducible (Table S1, Supplementary Material online). Inspection of the BAM files of these false mutations suggested that many of them were caused by mismapping of multiple loci to a single position on the reference genome; for example, reads with many unique differences, reads originating from three distinct alleles, or low mapping quality reads containing the "mutation" that map to multiple other MA lines. This prompted us to remove putative mutations that were also detected at low frequency in reads mapping to other MA lines, and globally, regions of the genome prone to mapping issues (supplementary text S6, Supplementary Material online). SNMs immediately adjacent to indels (i.e. 1 bp away) were also removed because these were typically caused by alignment errors. For Hom-Het mutations, we only considered sites whose ancestral genotype was homozygous for the allele of the reference genome because homozygous alternate sites were prone to mapping biases (supplementary text S7, Supplementary Material online). For some mutation types, we could not apply specific filters that would consistently eliminate false positives, so we manually inspected BAM alignment files with Integrative Genomics Viewer (IGV, Thorvalsdottir et al. 2013). We did this for all multinucleotide mutations (MNMs, multiple SNMs occurring within 50 bp of each other, Schrider et al. 2011), Het-Het mutations, and indels. We excluded those putative mutant sites that did not show correct linkage to existing polymorphisms within the reads, that had more than two alleles mapping to the region, and that had other MA lines with similar "mutant" reads (Fig S1, Supplementary Material online). The mutations that passed these described filtering steps were considered as our final set. We validated a subset of these with Sanger sequencing, including 19 SNMs, 2 indels, and 6 LOH regions (Table S1, Supplementary Material online). The false positive rates were 21%, 0%, and 0% for SNMs, indels, and LOH regions respectively. We did not remove mutations from our final set that were likely false positives in order to not introduce bias since not all mutations were checked with Sanger and/or manual inspection. We also confirmed that validating mutations by manually inspecting BAM files in all cases resulted in concordant conclusions with Sanger sequencing.

Scans for large-scale loss of heterozygosity: The genomic signature expected from a large-scale LOH event is a long homozygous region in which the ancestral state includes heterozygous sites (i.e. Het-Hom sites). We estimate that 0.7% of the sites (that we obtained sufficient data for) are heterozygous in the MA progenitor *Daphnia*. Therefore we expected to observe a heterozygous site every ~145 bp on average assuming that heterozygous sites are uniformly distributed across the genome. This gave us sufficient resolution to search for LOH events, since a large deletion or gene conversion tract spans multiple ancestral heterozygous sites (2-30 kb tracts, Xu et al. 2011). To search for LOH regions, we identified regions having consecutive Het-Hom sites minimally interrupted by heterozygous sites. We found many LOH "regions" that consisted of a single Het-Hom site ("stand-alone") or only a few Het-Hom sites clustered spanning a short stretch <100 bp (Fig S2, Supplementary Material online). We checked 16 of these with Sanger sequencing (8 of them found only prior to implementing our custom filtering algorithm) to determine if they were true Het-Hom sites. We found that none of these were reproducible suggesting that they were artefacts of either sampling error or mismapping (Table S1, Supplementary Material online). Therefore, we considered only multiple Het-Hom sites that spanned at least 100 bp. To determine whether the LOH region was likely due to a deletion event, we assessed whether the normalized depth of coverage among MA lines differed in these regions, with the expectation that a hemizygous deletion would cause the mutant line to have approximately half the coverage of all other lines.

Mutation rate calculation and genome equilibrium: All the mutations that remained after the filters and/or manual inspections described above were considered as our final set of mutations and were used for estimating mutation rates. We calculated the per site per generation mutation rate using the formula $\mu = m/(2nTg)$, where m is the total number of mutations detected, n is the number of sites analyzed, T is the total number of lines analyzed, and g is the average number of generations of MA (Keith et al. 2015). The denominator was multiplied by two (the diploid rate) to facilitate comparison against other MA studies that used inbred lines where one allele of newly-arising heterozygous genotype allele gets fixed within a few generations. For n, we used the total number of reliable sites that were not removed by filtering, which was 52,530,668. For calculating loss of heterozygosity rates, $2n$ was the total number of sites inferred to be heterozygous in the progenitor, which was 345,493. Confidence intervals around the mutation rate were calculated by bootstrapping with replacement the mutations from the 24 MA lines 1000 times (Efron and Tibshirani 1993).

We calculated the conditional mutation rates for all six substitution possibilities: A:T → T:A, A:T → C:G, A:T → G:C, G:C → C:G, G:C → A:T, and G:C → T:A using the number of ancestral sites of each nucleotide that we considered in our analysis. We also calculated the G|C → A|T (v) and A|T → G|C (u) mutation rates using the number of G+C or A+T ancestral sites in the regions of the genome we consider. We used the following equation to calculate the equilibrium A + T genome composition based on nucleotide substitution, p (Lynch and Walsh 2007): $p = \frac{v}{u+v}$. The linkage map produced by Xu *et al*. (2015a) was used to map mutations to the 12 chromosomes of *D. pulex* (Cristescu *et al*. 2006).

Effects of mutations: We determined the functional region (exonic, intronic, intergenic) in which each mutation occurred by using information from the *Daphnia* genome annotation in Ensembl (version 21). We compared these proportions with the proportion of the total sites in each of the functional regions we used for variant calling. We used snpEFF (Cingolani et al. 2012) to estimate the predicted effects of the mutations (e.g. missense, synonymous, frameshift).

Resequencing of two MA lines: To facilitate comparison to previous studies and to increase confidence in our results, we resequenced two random MA lines (C01 and C35) to obtain higher coverage (reaching 19 and 20 times coverage, respectively, from the original ~10x coverage). We combined the additional reads with the data from the first run to assess whether the additional depth affected the mutations discovered and the estimated mutation rates.

Mutation filtering in the non-MA isolates: The mutation detection and filtering pipeline of non-MA isolates was identical to that of the MA lines, except manual inspections were done for all putative mutations. This was necessary since we did not have the expectation that all legitimate mutations would be unique to a single isolate so we could not apply the usual filter of removing putative mutations that have mutant reads mapping in multiple individuals. In order to test for significant differences in mutation rates between the non-MAs and the MAs, we sampled with replacement the SNM rate from four random MA lines of the 24 and calculated the average SNM rate. We performed 10,000 permutations of this, generating a distribution of average SNM rates, which we then compared to the average SNM rate of the four non-MA isolates that were found to be independent (CC3, CC4, CC8, CC9). We calculated the estimated mutation rate based on a range of possible generations, from the lowest possible to be equal to the 0.05 percentile of the MA line permutation distribution (40 generations) up to approximately the generation the MA lines were at when the non-MAs were collected (100 generations). For calculating differences in Ts/Tv and the distribution of mutations across functional regions, we did 10,000 permutations randomly selecting the same number of mutations found in the non-MAs (20) from the set of mutations in the MA lines. Each time, the value was calculated based on the 20 mutations randomly sampled, and then a distribution of values was calculated.

Supplementary Material

Upon publication of this manuscript, the genome sequence data from this study will be made available in NCBI sequence read archive (SRA) under accession number xxx. Supplementary information including supplementary text S1-S7, Table S1, FigS1 and S2 are also in Supplementary Material online. Tables S2 and S3 contain information on the genomic location, ancestral genotype, mutant, derived genotype, and the snpEff predicted effect for all identified mutations identified in the MA lines and non-MAs and are available in Supplementary Material online. Table S2 contains all SNMs and Table S3 contains all indels.

ACKNOWLEDGEMENTS

We would like to thank M. Dukic and the lab of D. Ebert for advice on the protocol for reducing contamination, S. Kryazhimskiy for recommendations on the protocol for tagmentation library preparation, N Yan and M. Celis-Sagado for advice on Daphnia rearing. We also thank G. Bell, J. Bull, R. Vergilino and S. Xu for helpful discussions, and the numerous undergraduate students and research assistants for assisting in the maintenance of the MA lines over the years. This work was supported by the NSERC CREATE training program in Aquatic Ecosystem Health (397997-2011) and by NSERC Discovery Grant (341423-2012) to MEC. JMF was supported by an NSERC Undergraduate Summer Research Award.

LITERATURE CITED

Archetti M. 2004. Recombination and loss of complementation: a more than two-fold cost for parthenogenesis. J Evol Biol. 17:1084-1097.

Behringer MG, Hall DW. 2015. Genome-Wide Estimates of Mutation Rates and Spectrum in *Schizosaccharomyces pombe* Indicate CpG Sites are Highly Mutagenic Despite the Absence of DNA Methylation. G3 (Bethesda). 6:149-160.

Bergman CM, Kreitman M. 2001. Analysis of conserved noncoding DNA in Drosophila reveals similar constraints in intergenic and intronic sequences. Genome Res. 11:1335-1345.

Celis-Salgado MP, Cairns A, Kim N, Yan ND. 2008. The FLAMES medium: a new, soft-water culture and bioassay medium for Cladocera. Verh. Int. Ver. Theor. Angew. Limnol. 30:265-271.

Chen J, Cooper DN, Chuzhanova N, Ferec C, Patrinos GP. 2007. Gene conversion: mechanisms, evolution and human disease. Nat Rev Genet. 8:762-775.

Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. Fly. 6:80-92.

Colbourne JK, Pfrender ME, Gilbert D, Thomas WK, Tucker A, Oakley TH, Tokishita S, Aerts A, Arnold GJ, Basu MK et al. 2011. The ecoresponsive genome of *Daphnia pulex*. Science 331:555-561.

Conant GC, Wagner A. 2004. Duplicate genes and robustness to transient gene knock-downs in *Caenorhabditis elegans*. Proc Biol Sci. 271:89-96.

Crease TJ, Lynch M. 1991. Ribosomal DNA variation in *Daphnia pulex*. Mol Biol Evol. 8:620-640.

Cristescu ME, Colbourne JK, Radivojac J, Lynch M. 2006. A microsatellite-based genetic linkage map of the waterflea, *Daphnia pulex*: On the prospect of crustacean genomics. Genomics 88:415-430.

de Ligt J, Veltman JA, Vissers LELM. 2013. Point mutations as a source of de novo genetic disease. Curr Opin Genet Dev. 23:257-263.

Denver DR, Morris K, Lynch M, Thomas WK. 2004. High mutation rate and predominance of insertions in the *Caenorhabditis elegans* nuclear genome. Nature 430:679-682.

Denver DR, Dolan PC, Wilhelm LJ, Sung W, Lucas-Lledo JI, Howe DK, Lewis SC, Okamoto K, Thomas WK, Lynch M et al. 2009. A genome-wide view of *Caenorhabditis elegans* base-substitution mutation processes. Proc Natl Acad Sci USA. 106:16310-16314.

DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, Del Angel G, Rivas MA, Hanna M. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 43:491-498.

Efron B, Tibshirani RJ. 1994. An introduction to the bootstrap. CRC press.

Eyre-Walker A, Keightley PD. 2007. The distribution of fitness effects of new mutations. Nat Rev Genet. 8:610-618.

Fang H, Wu Y, Narzisi G, O'Rawe JA, Jimenez Barron LT, Rosenbaum J, Ronemus M, Iossifov I, Schatz MC, Lyon GJ. 2014. Reducing INDEL calling errors in whole genome and exome sequencing data. Genome Med. 6:89.

Farlow A, Long H, Arnoux S, Sung W, Doak TG, Nordborg M, Lynch M. 2015. The spontaneous mutation rate in the fission yeast *Schizosaccharomyces pombe*. Genetics 201:737-744.

Forche A, Abbey D, Pisithkul T, Weinzierl MA, Ringstrom T, Bruck D, Petersen K, Berman J. 2011. Stress alters rates and types of loss of heterozygosity in *Candida albicans*. Mbio. 2:e00129-11.

Galtier N, Piganeau G, Mouchiroud D, Duret L. 2001. GC-content evolution in mammalian genomes: the biased gene conversion hypothesis. Genetics 159:907-911.

Gu Z, Steinmetz LM, Gu X, Scharfe C, Davis RW, Li W. 2003. Role of duplicate genes in genetic robustness against null mutations. Nature 421:63-66.

Halligan DL, Keightley PD. 2009. Spontaneous mutation accumulation studies in evolutionary genetics. Ann. Rev. Ecol. Evol. Syst.40:151-172.

Henson S, Bishop RP, Morzaria S, Spooner PR, Pelle R, Poveda L, Ebeling M, Kung E, Certa U, Daubenberger CA et al. 2012. High-resolution genotyping and mapping of recombination and gene conversion in the protozoan *Theileria parva* using whole genome sequencing. BMC Genomics. 13:503..

Hiruta C, Nishida C, Tochinai S. 2010. Abortive meiosis in the oogenesis of parthenogenetic *Daphnia pulex*. Chromosome Res.18:833-840.

Houle D, Hoffmaster D, Assimacopoulos S, Charlesworth B. 1992. The genomic mutation-rate for fitness in Drosophila. Nature 359:58-60.

Huang W, Lyman RF, Lyman RA, Carbone MA, Harbison ST, Magwire MM, Mackay TF. 2016. Spontaneous mutations and the origin and maintenance of quantitative genetic variation. Elife 5:e14625.

Innes DJ, Hebert PD. 1988. The origin and genetic basis of obligate parthenogenesis in *Daphnia pulex*. Evolution 1024-1035.

Jiang C, Mithani A, Belfield EJ, Mott R, Hurst LD, Harberd NP. 2014. Environmentally responsive genome-wide accumulation of de novo *Arabidopsis thaliana* mutations and epimutations. Genome Res. 24:1821-1829.

Jiang Y, Turinsky AL, Brudno M. 2015. The missing indels: an estimate of indel variation in a human genome and analysis of factors that impede detection. Nucleic Acids Res. 43:7217-7228.

Keightley PD, Lynch M. 2003. Toward a realistic model of mutations affecting fitness. Evolution 57:683-685.

Keightley PD, Eyre-Walker A. 1999. Terumi Mukai and the riddle of deleterious mutation rates. Genetics 153:515-523.

Keightley PD, Pinharanda A, Ness RW, Simpson F, Dasmahapatra KK, Mallet J, Davey JW, Jiggins CD. 2015. Estimation of the spontaneous mutation rate in *Heliconius melpomene*. Mol Biol Evol. 32:239-243.

Keightley PD, Trivedi U, Thomson M, Oliver F, Kumar S, Blaxter ML. 2009. Analysis of the genome sequences of three *Drosophila melanogaster* spontaneous mutation accumulation lines. Genome Res. 19:1195-1201.

Keith N, Tucker AE, Jackson CE, Sung W, Lucas Lledo JI, Schrider DR, Schaack S, Dudycha JL, Ackerman M, Younge AJ et al. 2016. High mutational rates of large-scale duplication and deletion in *Daphnia pulex*. Genome Res. 26:60-69.

Kondrashov A. 1988. Deleterious mutations and the evolution of sexual reproduction. Nature 336:435-440.

Kumar S. 2005. Molecular clocks: four decades of evolution. Nat Rev Genet. 6:654-662.

Latta LC, Morgan KK, Weaver CS, Allen D, Schaack S, Lynch M. 2013. Genomic background and generation time influence deleterious mutation rates in Daphnia. Genetics 193:539-544.

Lemeta S, Pylkkanen L, Sainio M, Niemela M, Saarikoski S, Husgafvel-Pursiainen K, Bohling T. 2004. Loss of heterozygosity at 6q is frequent and concurrent with 3p loss in sporadic and familial capillary hemangioblastomas. J Neuropathol Exp Neurol. 63:1072-1079.

Li H. 2011. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics 27:2987-2993.

Lynch M, Walsh B. 2007. The origins of genome architecture. Sinauer Associates Sunderland.

Lynch M, Sung W, Morris K, Coffey N, Landry CR, Dopman EB, Dickinson WJ, Okamoto K, Kulkarni S, Hartl DL et al. 2008. A genome-wide view of the spectrum of spontaneous mutations in yeast. Proc Natl Acad Sci USA. 105:9272-9277.

Muller H. 1964. The Relation of Recombination to Mutational Advance. Mutat Res. 1:2-9.

Ness RW, Morgan AD, Vasanthakrishnan RB, Colegrave N, Keightley PD. 2015. Extensive de novo mutation rate variation between individuals and across the genome of *Chlamydomonas reinhardtii*. Gen Res. 25:1739-1749.

O'Halloran DM. 2015. PrimerView: high-throughput primer design and visualization. Source Code for Biology and Medicine. 10:1.

Omilian AR, Cristescu ME, Dudycha JL, Lynch M. 2006. Ameiotic recombination in asexual lineages of Daphnia. Proc Natl Acad Sci USA. 103:18638-18643.

Ossowski S, Schneeberger K, Lucas-Lledo JI, Warthmann N, Clark RM, Shaw RG, Weigel D, Lynch M. 2010. The rate and molecular spectrum of spontaneous mutations in *Arabidopsis thaliana*. Science 327:92-94.

Ribeiro A, Golicz A, Hackett CA, Milne I, Stephen G, Marshall D, Flavell AJ, Bayer M. 2015. An investigation of causes of false positive single nucleotide polymorphisms using simulated reads from a small eukaryote genome. BMC Bioinformatics 16:1.

Rozen S, Skaletsky H. 1999. Primer3 on the WWW for general users and for biologist programmers. Bioinformatics Methods and Protocols 365-386.

Saxer G, Havlak P, Fox SA, Quance MA, Gupta S, Fofanov Y, Strassmann JE, Queller DC. 2012. Whole genome sequencing of mutation accumulation lines reveals a low mutation rate in the social amoeba *Dictyostelium discoideum*. PloS One 7:e46759.

Schrider DR, Houle D, Lynch M, Hahn MW. 2013. Rates and genomic consequences of spontaneous mutational events in *Drosophila melanogaster*. Genetics 194:937-954.

Schrider DR, Hourmozdi JN, Hahn MW. 2011. Pervasive multinucleotide mutational events in eukaryotes. Curr Biol. 21:1051-1054.

Su Z, Wang J, Gu X. 2014. Effect of duplicate genes on mouse genetic robustness: an update. Biomed Res Int. 2014:758672.

Sung W, Ackerman MS, Miller SF, Doak TG, Lynch M. 2012. Drift-barrier hypothesis and mutation-rate evolution. Proc Natl Acad Sci USA. 109:18488-18492.

Thorvaldsdottir H, Robinson JT, Mesirov JP. 2013. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform. 14:178-192.

Tucker AE, Ackerman MS, Eads BD, Xu S, Lynch M. 2013. Population-genomic insights into the evolutionary origin and fate of obligately asexual *Daphnia pulex*. Proc Natl Acad Sci USA. 110:15740-15745.

Uchimura A, Higuchi M, Minakuchi Y, Ohno M, Toyoda A, Fujiyama A, Miura I, Wakana S, Nishino J, Yagi T. 2015. Germline mutation rates and the long-term phenotypic effects of mutation accumulation in wild-type laboratory mice and mutator mice. Genome Res. 25:1125-1134.

Wakeley J. 1996. The excess of transitions among nucleotide substitutions: New methods of estimating transition bias underscore its significance. Trends Ecol Evol. 11:158-163.

Xu S, Omilian AR, Cristescu ME. 2011. High rate of large-scale hemizygous deletions in asexually propagating Daphnia: implications for the evolution of sex. Mol Biol Evol. 28:335-342.

Xu S, Ackerman MS, Long H, Bright L, Spitze K, Ramsdell JS, Thomas WK, Lynch M. 2015a. A male-specific genetic map of the microcrustacean *Daphnia pulex* based on single-sperm whole-genome sequencing. Genetics 201:31-38.

Xu S, Spitze K, Ackerman MS, Ye Z, Bright L, Keith N, Jackson CE, Shaw JR, Lynch M. 2015b. Hybridization and the origin of contagious asexuality in *Daphnia pulex*. Mol Biol Evol. 32:3215-3225.

Yang S, Wang L, Huang J, Zhang X, Yuan Y, Chen J, Hurst LD, Tian D. 2015. Parent-progeny sequencing indicates higher mutation rates in heterozygotes. Nature 523:463-467.

Zhu YO, Siegal ML, Hall DW, Petrov DA. 2014. Precise estimates of mutation rate and spectrum in yeast. Proc Natl Acad Sci USA. 111:E2310-E2318.

TABLES

Table 1. Summary of point mutations across MA lines. SNM rate is per site per generation. For indel size, (+) indicates an insertion and (-) a deletion. SE indicated the standard error.

Line	SNMs	Indels (size)	SNM rate
C01	11		1.19×10^{-9}
C02	9	1 (+2)	1.01×10^{-9}
C03	10		1.19×10^{-9}
C06	15		1.83×10^{-9}
C07	15		1.72×10^{-9}
C08	21	1 (+7)	2.50×10^{-9}
C12	7		7.75×10^{-10}
C13	25		2.80×10^{-9}
C14	15		1.85×10^{-9}
C16	8		9.64×10^{-10}
C17	12		1.33×10^{-9}
C18	16		1.93×10^{-9}
C20	36		4.23×10^{-9}
C21	24	1 (-1)	2.60×10^{-9}
C24	16		2.09×10^{-9}
C25	35	1 (-13)	3.83×10^{-9}
C34	18		2.14×10^{-9}
C35	19		2.06×10^{-9}
C36	17		1.95×10^{-9}
C37	33	1 (-1)	3.61×10^{-9}
C38	31		3.51×10^{-9}
C39	23	1 (-7)	2.58×10^{-9}
C40	32		3.81×10^{-9}
C44	29		3.83×10^{-9}

Table 2. Conditional mutation rates of all 6 possible substitutions. The number not in parenthesis in the frequency column indicates the number of each type of mutation found in the MAs, and in parenthesis the number in the non-MAs. The conditional rate was calculated from the MA lines.

	Substitution type	Frequency	Conditional rate (per nucleotide per generation)
Transitions	A:T → G:C	58 (0)	4.77×10^{-10}
	C:G → T:A	156 (13)	1.75×10^{-9}
Transversions	A:T → T:A	39 (1)	3.21×10^{-10}
	A:T → C:G	19 (1)	1.56×10^{-10}
	C:G → A:T	187 (3)	2.10×10^{-9}
	C:G → G:C	18 (1)	2.02×10^{-10}

Table 3. Summary of LOH regions in MA lines other than C40. Minimum boundaries are defined as the distance between the outermost Het-Hom sites of the LOH region (a minimum min range of 100 bp was required). Maximum boundaries are the distances between the heterozygous sites flanking the LOH region. The normalized depth ratio is the ratio of normalized read depth of the focal line to the average of the other lines.

Scaffold, chr	Min boundaries (min size)	Max boundaries (max size)	Line	Het-Hom sites	Normalized depth ratio, type
12, chr5	1029411 – 1031106 (1695 bp)	1028341 – 1032353 (4012 bp)	C20	6	0.63, deletion
20, chr5	1063860 – 1065038 (1178 bp)	1063351 – 1065479 (2128 bp)	C01	19	0.64, deletion
213, chr5	67562 – 68475 (913 bp)	66286 – 68643 (2357 bp)	C02	2	0.70, deletion
84, chr12	355063 – 355244 (181 bp)	355031 – 355534 (503 bp)	C08	8	0.65, deletion
48, chr7	123641 – 125719 (2078 bp)	122974 – 126858 (3884 bp)	C08	21	0.65, deletion
Various, chr11			C40	32835	~1.0, gene conversion
51, chr9	752324-763231 (10,907 bp)	732650-770446 (37,796 bp)	CC9	67	0.54, deletion
51, chr9	792563-793686 (1123 bp)	777732-796805 (19,073 bp)	CC9	24	0.67, deletion
51, chr9	823236-848675 (25,439 bp)	820242-848891 (28,649 bp)	CC9	26	0.46, deletion
180, unknown	180632-181514 (882 bp)	178834-181818 (2984 bp)	CC9	22	0.54, deletion

FIGURES

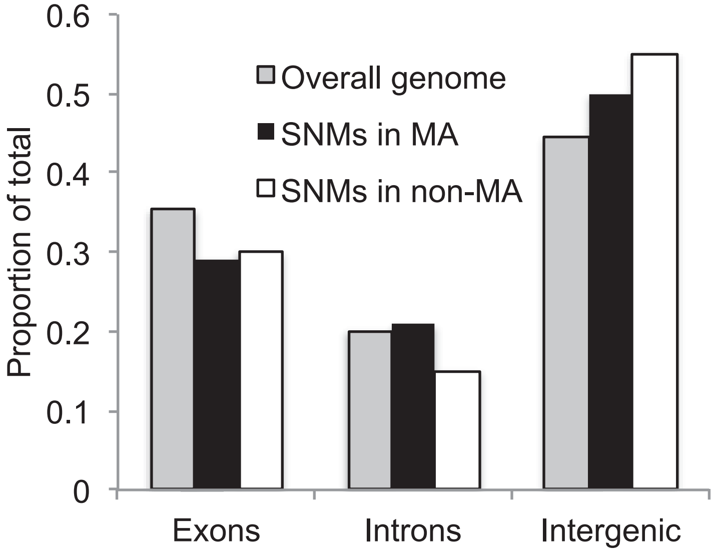


Figure 1. Proportion of single nucleotide mutations (SNMs) within the various functional regions in the MA lines and non-MA isolates compared to the overall composition of the genome.

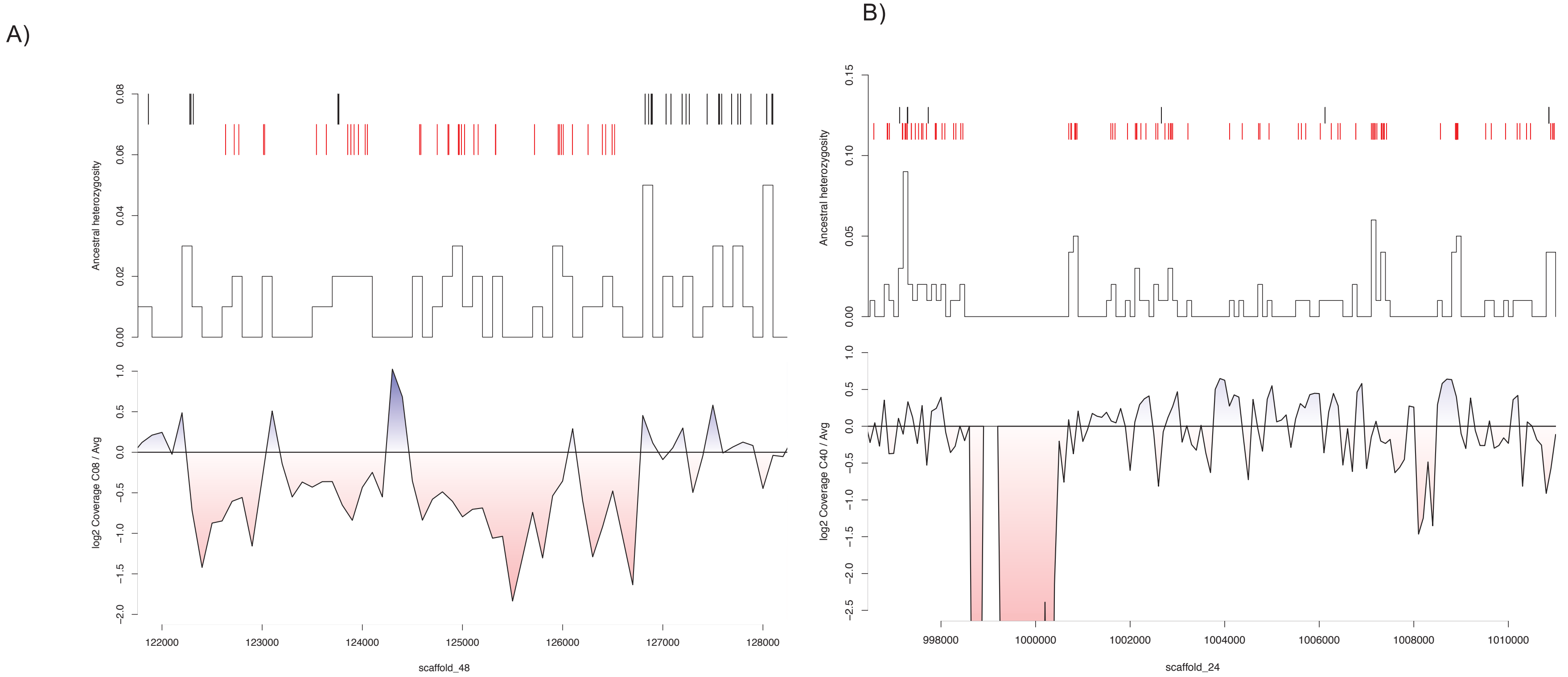


Figure 2. Genome tracks illustrating LOH. On the top track, black lines represent sites that are heterozygous in all MA lines (some may be found even in LOH regions due to mapping/genotyping errors or mutations occurring after the LOH event), and red lines represent Het-Hom sites in the mutant line. The middle track shows the ancestral heterozygosity levels across the genomic region. The bottom track shows the normalized read depth, with the expected value for a hemizygous deletion being -1, infinitely negative for a homozygous deletion, and 0 for a gene conversion. (A) The hemizygous deletion in C08. (B) A portion of the gene conversion event in C40 including homozygous gene conversion regions flanking a homozygous deletion.

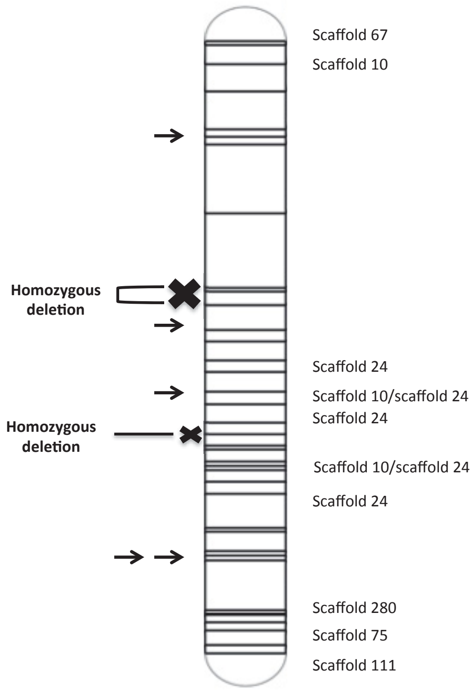


Figure 3. Map of complex LOH on chromosome 11 in C40. The physical locations of the scaffolds mapping to chromosome 11 are on the right (as illustrated some scaffolds have been misassembled). Each arrow on the left indicates a ~900 bp fragment that was amplified and confirmed with Sanger sequencing to have ancestrally heterozygous sites but only homozygous sites in C40. Homozygous deletions that have been able to be mapped to physical locations are also shown.

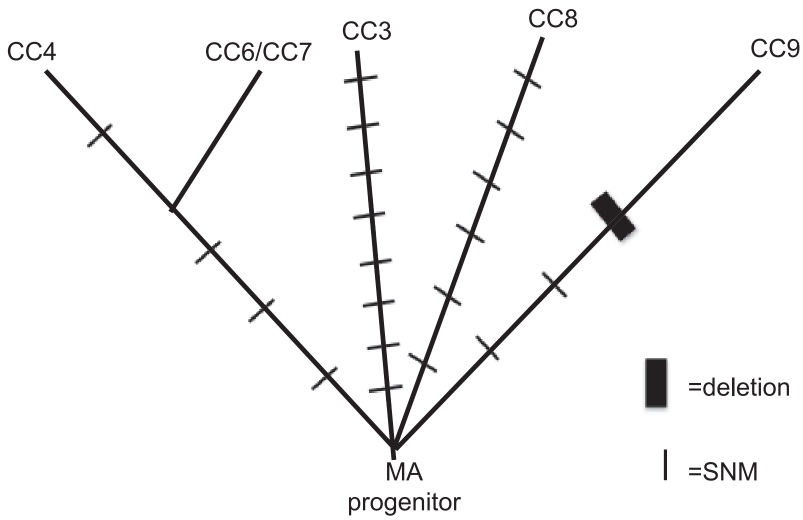


Figure 4. Manually-drawn evolutionary relationship among the six non-MA isolates based on mutations detected.

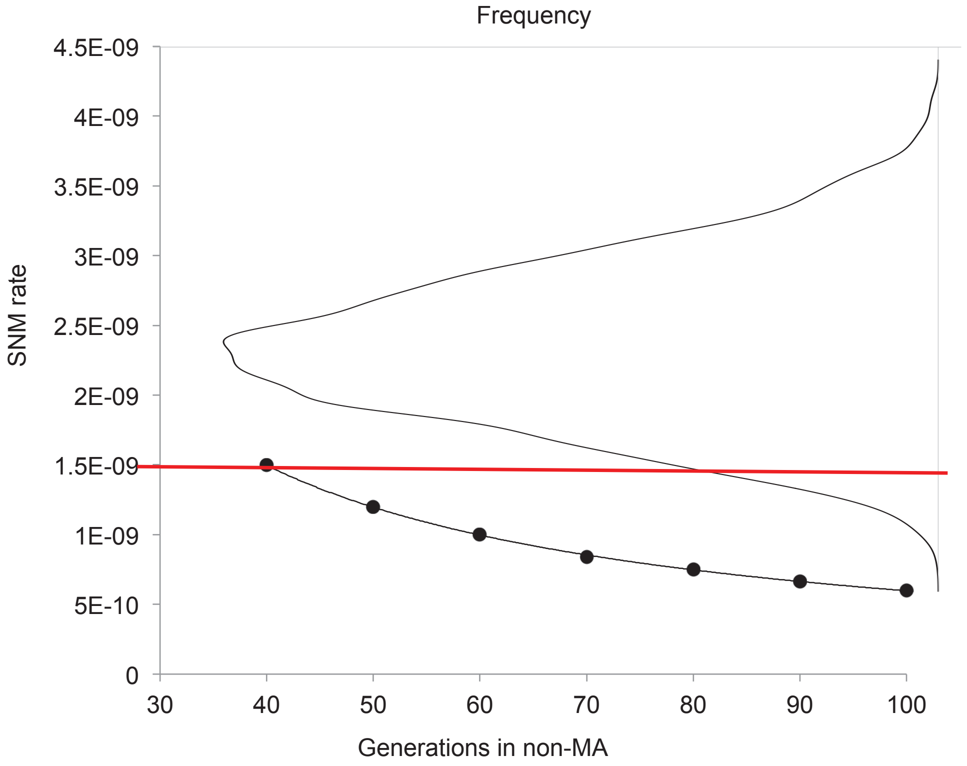


Figure 5. Mutation rates in non-MA isolates calculated based on a range of generations compared to the MA lines. The approximately normal distribution (thin black line) is based on 10,000 permutations of sampling with replacement of the mutation rates from four random MA lines. The red line represents the p = 0.05 quantile of the MA distribution. Only the four independent non-MA lineages were used to calculate the SNM rate at various generation intervals (connected black dots).

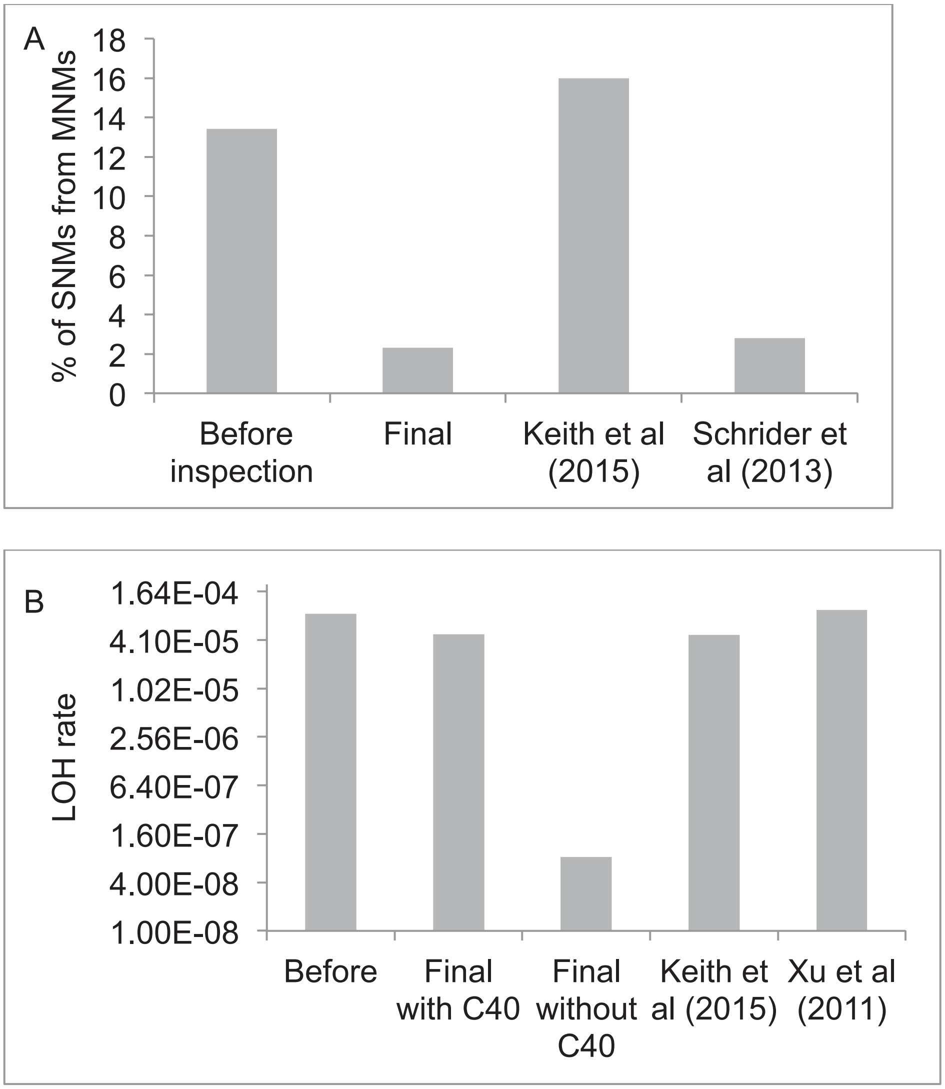


Figure 6. Mutation types before and after filtering. (A) Percentage of SNMs accounted for by MNMs (multinucleotide mutations) before and after manual inspection. "Before inspection" refers to this study after all filtering steps except manual inspection of BAM alignment files. "Final" refers to this study after manual inspection of BAM files. The estimate from Keith *et al*. (2015) is from the asexual lineage of *Daphnia pulex* MA lines after excluding their hypermutator outlier. The Schrider *et al.* (2013) study was conducted on *Drosophila melanogaster*. (B) LOH rates in our study: "before" is before removing stand-alone Het-Hom sites and regions of consecutive Het-Hom sites spanning <100 bp, "final with C40" is after implementing this filter and including the event in C40, "final without C40" does not include C40's event in the calculation. The estimate from Keith *et al.* (2015) is the sum of the rate of deletion and gene conversion in the asexual *D. pulex* MA lines, and the estimate from Xu *et al.* (2011) was also from *D. pulex* derived from genotyping microsatellites.

ADDITIONAL DATA

Sequencing of four low-fitness MA lines

Some MA lines demonstrated a noticeable deterioration in fitness during the course of the experiment. I sequenced and analyzed the genomes of four such lines (C23, C27, C43, C49) with the objective of drawing inferences on the patterns (numbers and types) of mutation correlated with fitness decline. These MA lines had progressed fewer generations than the average of all the MA lines (on average 2.6 standard deviations fewer), indicating higher mortality and/or slower generation time, and required backup lines to be used more frequently. Backup lines are used when the focal individual does not produce offspring or experiences mortality. One MA line, C43, additionally demonstrated a phenotypic abnormality of a visibly different and seemingly less efficient swimming pattern.

Sequencing and analysis were carried out with identical procedures as the other MA lines and non-MA isolates, as indicated in the manuscript. A total of 41,697,845 sites were analyzed with an average coverage of 13x per line. Among the four new MA lines, I found a total of 20 SNMs, corresponding to a rate of 7.27×10^{-10} per nucleotide per generation (Table ii, Appendix III). This was significantly lower than the SNM rate of 2.30×10^{-9} of the other 24 MA lines that were sequenced (permutation test, $p < 0.0001$). Only a single indel, a one base pair deletion in an intronic region, was found in C49. This corresponds to an indel rate of 3.63×10^{-11} per site per generation, which is close to the rate of 2.90×10^{-11} in the other 24 MA lines, although with so few indels the comparison is not particularly meaningful. However, the SNM results support the hypothesis that the absolute number of mutations is not necessarily correlated to the level of fitness decline, rather it is a few large effect mutations that cause deleterious effects (Dillon and Cooper 2016). Although the number of mutations observed suggests a lower mutation rate, the biological mutation rate of these four MA lines was not necessarily lower: it is possible that because of an early deleterious mutation, some new mutations could not accumulate on this background without severe fitness effects, so they were not propagated. For example, a severe deleterious mutation could have occurred relatively early in the experiment (e.g. a large deletion, see below), and then subsequent mutations had negative interactions with the existing deleterious mutation and so the MA line did not propagate that generation and backup lines had to be used.

There were a total of 140,356 ancestrally heterozygous sites that were used for the loss of heterozygosity (LOH) scan. I found LOH events in three of these four MA lines, and all of these events resulted from hemizygous deletion (Table iii, Appendix III). This corresponded to a rate of 3.77 x 10^{-5} per heterozygous site per generation, which is close to the average overall LOH rate in the 24 MA lines of 4.82 x 10^{-5}. This finding supports the finding of robust overall LOH rates found in the manuscript. However, although the overall rate is similar, the contributions of hemizygous deletion versus gene conversion are vastly different between the 24 MA lines and the four lower fitness lines. No gene conversion but high rates of hemizygous deletion were found in the lower fitness MA lines, and high rates of gene conversion and low rates of deletion were found in the 24 MA lines. In fact, the hemizygous deletion rate is orders of magnitude higher in the lower fitness MA lines compared to the 24 others (3.77 x 10^{-5} versus 8.21 x 10^{-8}) and this difference is statistically significant (permutation test, $p < 0.0001$). This qualitative correlation between fitness and hemizygous deletion supports the hypothesis that deletions resulting in hemizygosity are often deleterious – which complements the results in the main text of the manuscript. MA lines C27 and C49 incurred deletions spanning over 3 kb and 16 kb, respectively. MA line C23 experienced a particularly large deletion, spanning approximately 700 kb (Table iii, Appendix III). Interestingly, this event occurred on chromosome 11, the same chromosome that experienced the massive gene conversion in C40. This finding suggests that chromosome 11 may be prone to DNA breakage, leading to repair mechanisms that result in deletions or gene conversions (Preston et al. 2006).

References

Dillon MM, Cooper VS. 2016. The fitness effects of spontaneous mutations nearly unseen by selection in a bacterium with multiple chromosomes. bioRxiv 060483.

Preston CR, Flores CC, Engels WR. 2006. Differential usage of alternative pathways of double-strand break repair in Drosophila. Genetics 172:1055-1068.

GENERAL CONCLUSIONS

The main purpose of this thesis was to provide mutation rate estimates for *Daphnia pulex*, and also to demonstrate the action of selection on new mutations. Being one of the largest whole genome sequencing studies of metazoan MA lines to date, I present a single nucleotide mutation rate of 2.30 x 10^{-9} per nucleotide per generation, which is similar to findings in other metazoans. I show that loss of heterozygosity events occur at a high rate, and mostly result in neutral or deleterious effects, with an ameiotic recombination event encompassing an entire chromosome, and several deletions resulting in hemizygosity. MA lines that experienced greater declines in fitness had significantly higher deletion rates. I also found that diversity can be selected for and maintained in a population, as four distinct lineages were observed from the sampling of six individuals from a population founded by a single asexual clone. Purifying selection was also acting in the population, indicated by fewer mutations than expected and a significantly higher transition bias than the MA lines. Both positive and negative selection on mutations was strong in *Daphnia*, even though our population was not imposed to any particular stress, the genome contains many duplicate genes, and the population was asexual with no recombination occurring. This work appeals to broad fields in biology because I present results about the maintenance of diversity in populations, the strength and efficiency of selection, the implications of LOH in asexual species, and also provide solutions to methodological issues especially pertinent to researchers working with poor genome assemblies or genomes with high duplication levels.

The conclusions I present here and the knowledge I gained through this work provoke me to make predictions and recommendations for future work in the field. Firstly, more mutation rate studies using different organisms from a wider variety of taxa should continue to be conducted. The relatively stable mutation rate across metazoans suggests tight selection on mutation rates. However, due to practical limitations, mutation rate studies have been carried out in a fairly narrow range of metazoan taxa (mostly arthropods and nematodes), exhibiting large effective population sizes (Ne) in nature and similar genome sizes within the range of 100-300 Mb. This limitation makes it difficult to draw strong conclusions on the selective regime on the mutation rate since Ne and genome size may be important factors influencing the mutation rate (Sung et al. 2012). Secondly, further work needs to be done in order to understand the fitness effects of different types and even individual mutations; that is, drawing a direct link from genotype to phenotype. This has been a long-standing challenge in the field of genetics and is now becoming possible with large amounts of genomic data from many species becoming available (Joly-Lopez et al. 2016). Although we do not show the direct fitness effects of specific mutations, we infer that a significant proportion of mutations must have high effects on fitness in order to observe the patterns of we did: both positive effects – to observe the high levels of diversity (selection for diversity); and negative effects – to observe fewer mutations than expected based on the mutation rate (selection purging deleterious mutations).

Our approach of comparing MA lines to a population experiencing selection allowed a direct comparison of the rates, types, and patterns of mutations accumulated in conditions with and without selection. A similar approach can be taken with different species, as it would be interesting to investigate whether the signature of selection would be as strong in other species with different genome properties, as we inferred in our study in *Daphnia*. A large portion of the *Daphnia* genome is coding (~15%) and many genes are duplicated (Colbourne et al. 2011), and this could affect the proportion of mutations that have strong fitness effects. Studying LOH in other asexual species would also be interesting – to investigate if the high rates we find are unique to *Daphnia* or are common across many asexual species. All studies investigating LOH in mutation accumulation lines thus far have used *Daphnia* (Omilian et al. 2006, Xu et al. 2011, Keith et al. 2015, this study).

My study also highlights the need for thorough bioinformatic analyses for calling mutations in whole-genome sequencing data, and that identical procedures may not be appropriate for all organisms. Existing studies have used a variety of different methods for calling mutations, including different genome mapping and variant calling software and different thresholds for filtering mutations (see references within Table i, Appendix I). In fact, it can be difficult to standardize procedures for all organisms – which have reference genome assemblies of varying quality and different genome properties (e.g. heterozygosity and gene duplication levels). For example, organisms with lower quality reference genomes and genomes with high amounts of duplicate genes will tend to have more issues with read mismapping (Li 2011), and different filtering thresholds will have to be used when mutations are expected to be heterozygous versus homozygous (Zhu et al. 2014, Keith et al. 2015). Mutation filtering procedures on newly studied organisms should be well informed by building on methods from existing studies combined with validation procedures. Comparisons of mutation rates within the same order of magnitude between species may be difficult or even unmeaningful because independent research groups use different data processing and mutation filtering procedures. To remedy this, independent replicate experiments of the same species can be carried out, or data can be re-analyzed by different research groups using different methods. This further investigation would allow us to obtain an understanding of the robustness of estimates to different analytic procedures, and therefore the meaningfulness of differences in mutation rates among species.

Additionally, more diverse classes of mutations will likely be studied thoroughly in the future. The focus of past studies has been on SNMs, however the accuracy and sensitivity of the detection of indels and large-scale structural changes – such as copy number variants, inversions, and transpositions, will likely improve. This may be aided by new technologies allowing sequencing of very long DNA strands (Ambardar et al. 2016). Overall, the results of this project contribute to the understanding of the fundamental phenomenon of mutation as well as the impact of selection. My work will be useful for further research of mutation rates and genomic analyses of species with high amounts of duplicate genes, and will also spark research about differences in the fitness-effect spectrum of mutations and the strength and patterns of selection.

References

Ambardar S, Gupta R, Trakroo D, Lal R, Vakhlu J. 2016. High Throughput Sequencing: An Overview of Sequencing Chemistry. Indian J Microbiol. :1-11.

Colbourne JK, Pfrender ME, Gilbert D, Thomas WK, Tucker A, Oakley TH, Tokishita S, Aerts A, Arnold GJ, Basu MK et al. 2011. The ecoresponsive genome of *Daphnia pulex*. Science. 331:555-561.

Joly-Lopez Z, Flowers JM, Purugganan MD. 2016. Developing maps of fitness consequences for plant genomes. Curr Opin Plant Biol. 30:101-107.

Keith N, Tucker AE, Jackson CE, Sung W, Lucas Lledo JI, Schrider DR, Schaack S, Dudycha JL, Ackerman M, Younge AJ et al. 2016. High mutational rates of large-scale duplication and deletion in *Daphnia pulex*. Genome Res. 26:60-69.

Li H. 2011. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics 27:2987-2993.

Omilian AR, Cristescu ME, Dudycha JL, Lynch M. 2006. Ameiotic recombination in asexual lineages of Daphnia. Proc Natl Acad Sci USA. 103:18638-18643.

Schrider DR, Houle D, Lynch M, Hahn MW. 2013. Rates and genomic consequences of spontaneous mutational events in *Drosophila melanogaster*. Genetics 194:937-954.

Sung W, Ackerman MS, Miller SF, Doak TG, Lynch M. 2012. Drift-barrier hypothesis and mutation-rate evolution. Proc Natl Acad Sci USA. 109:18488-18492.

Xu S, Omilian AR, Cristescu ME. 2011. High rate of large-scale hemizygous deletions in asexually propagating Daphnia: implications for the evolution of sex. Mol Biol Evol. 28:335-342.

Zhu YO, Siegal ML, Hall DW, Petrov DA. 2014. Precise estimates of mutation rate and spectrum in yeast. Proc Natl Acad Sci USA. 111:E2310-E2318.

Appendix I

Table i. Examples of mutation rate estimates using the WGS approach.

Study	Organism	SNM rate	Method
Denver et al. 2009	*Caenorhabditis elegans*	2.5 x 10^{-9}	MA
Keightley et al. 2009	*Drosophila melanogaster*	3.5 x 10^{-9}	MA
Ossowski et al. 2010	*Arabidopsis thaliana*	7.0 x 10^{-9}	MA
Roach et al. 2010	*Homo sapiens*	1.1 x 10^{-8}	Parent-offspring
Saxer et al. 2012	*Dictyostelium discoideum*	2.9 x 10^{-11}	MA
Ness et al. 2012	*Chlamydomonas reinhardtii*	2.08 x 10^{-10}	MA
Zhu et al. 2014	*Saccharomyces cerevisiae*	1.67 x 10^{-10}	MA
Keightley et al. 2014	*Drosophila melanogaster*	2.8 x 10^{-9}	Parent-offspring
Keightley et al. 2015	*Heliconius melpomene*	2.9 x 10^{-9}	Parent-offspring

Appendix II – Manuscript supplementary material

Supplementary text S1-S7
Table S1
Figs S1, S2

Supplementary text
S1:

Scaffolds associated with chromosome 11. Scaffolds with an asterisk we inferred to be associated with chromosome 11 because of the gene conversion event in C40. Others were previously identified in Colbourne et al. (2011) and Xu et al. (2015).

10, 24, 67, 75, 81*, 85*, 102*, 111, 116*, 141*, 166*, 177*, 193*, 260*, 280, 283*, 315*, 327*,743, 4243*

S2:

Sample preparation**:** Since *Daphnia* are known to carry microparasites and symbionts (Qi et al. 2009), individuals were subjected to a procedure designed to reduce the amount of foreign DNA prior to harvesting tissue for DNA extraction (Fields et al. 2015). Adult *Daphnia* were placed in an antibiotic solution of 30 mg/L tetracycline, 50 mg/L streptomycin, and 50 mg/L ampicillin for 48 hours. During this time they were fed frequently with sterile 50 µm Sephadex G-25 beads to clear out gut contents. After this procedure, organisms were harvested (1-5 adult individuals per line) and DNA was extracted with the cetyltrimethylammonium bromide method (Doyle 1987). DNA concentrations were quantified with PicoGreen Quant-IT (Invitrogen), and all samples were diluted to 2.5 ng/µl as required for library preparation.

Library preparation and data processing: Libraries were prepared via a tagmentation procedure, which consists of simultaneous fragmentation and tagging (adding adapters) of genomic DNA with a modified transposase enzyme (Adey et al. 2010). To optimize for efficiency and reduce the input DNA we used a protocol modified from the Illumina Nextera approach (Baym et al. 2015). Libraries were cleaned and short products removed with AMPure XP beads (Beckman Coulter) before being normalized and pooled. The final libraries for the MA lines were run on three lanes of Illumina HiSeq 100 bp paired-end reads at the Genome Quebec Innovation Centre at McGill University. The non-MA libraries were sequenced on a separate lane, along with four other *D. pulex* samples (data not in this manuscript). For each sample and each lane of sequencing, reads were then mapped against the *Daphnia pulex* reference genome (wfleabase.org) using Burrows-Wheeler Aligner v0.7.10 (Li and Durbin 2009). Adapter sequences were removed and overlapping sequences merged with SeqPrep (https://github.com/jstjohn/SeqPrep). After alignment, the resulting SAM files were cleaned and sorted, and duplicates were removed with Picard tools v1.123 (http://broadinstitute.github.io/picard). Local realignments of reads around insertions and deletions (indels) were performed using Broad Institute's Genome Analysis Tool Kit (GATK) algorithms v.3.3.0 (DePristo et al. 2011, https://www.broadinstitute.org/gatk/index.php). All lanes for the same MA line were combined, and then reprocessed for duplicates and local realignment.

S3:

For SNMs, all sites were removed by VariantFiltration that had QualByDepth (QD) < 2.0, FisherStrand (FS) > 60.0, Root Mean Square Mapping Quality (MQ) < 20.0, Quality (QUAL) < 30.0 , MappingQualityRankSum (MQRankSum) < -12.5, or ReadPosRankSum < -8.0. Similarly, for indels, all sites that had parameters QD < 2.0, FS > 200.0, QUAL< 30.0, or ReadPosRankSum < -20.0 were removed, as recommended by GATK (https://www.broadinstitute.org/gatk/index.php).

S4:

We initially tried a GATK-based filtering regime in order to filter the various types of mutations (Hom-Het, Het-Het, Het-Hom). For Hom-Het and Het-Het mutations, the depth of the alternative allele was required to be at least 3 with allele ratios between 0.25 and 0.75 in the mutant line. For Het-Hom mutations, at least 90% of the reads were required to support the homozygous call. After testing 13 Hom-Het or Het-Het mutations and 6 Het-Hom mutations and finding that all were not reproducible, we decided against using this method and instead designed our own algorithm.

S5:

We extracted read information using mpileup in samtools v0.1.19 (Li 2011) to infer the ancestral state of each putatively variant site and also test if any of the MA lines had a genotype that was statistically supported to be different than the ancestral state. To infer the ancestral state at a site, a Z-test (approximating the normal distribution) was used to test if the observed allele depths across all 24 MA lines could be explained by the ancestral state being homozygous or heterozygous. We then scanned the allele depths of each MA line and used a binomial test to test if the line deviated from the inferred ancestral genotype. We used strict thresholds to filter true mutants from mapping and sequencing errors: with Hom-Het and Het-Het mutations requiring an alternative allele depth of at least two, and Het-Hom mutations requiring zero reads of the alternative allele. For Hom-Het and Het-Het mutations, no more than one read containing the mutant allele was allowed to occur in an MA line other than the one that the mutation was being called in. After finding possible mutant lines, we excluded their reads to re-infer the ancestral genotype to ensure it was correct. Sites that were rejected as being ancestrally homozygous or heterozygous (e.g. if they have two alleles mapping but statistically different from a 1:1 ratio) are probably due to mapping errors and were therefore excluded from the analysis.

S6:

Upon manual inspection of BAM files of putative mutations with IGV, we found that many were signatures of false positives caused by mapping artefacts. Reads from a different locus and containing a different variant would be mapping to a locus, and reads like this would be found in all or most MA lines. Our thresholds for removing reads with low mapping quality should get rid of most reads; however, in some cases, enough reads were retained, in a single line, in order to call a mutation in that line. Therefore, we had to use information from the raw unfiltered BAM files in order to ensure we removed these artefactual mutations and the error-prone regions that contained them. We removed error-prone regions of the genome with the following criteria: (1) the region contained multiple alleles mapping in multiple MA lines in the raw mapped data; and (2) the region also contained either a Hom-Het site (that was detected as a false mutation because of mismapping) or a site with allele depth ratios not representative of a true heterozygous locus (e.g. 200 reads A, 40 reads T). We ensured that all regions that we determined were part of a loss of heterozygosity region we determined with our scan were not removed.

S7:

Sites whose ancestral state was homozygous alternate from the reference genome had a mutation rate one order of magnitude higher than sites that were homozygous for the same allele as the reference genome. Upon inspection of BAM files, this appeared to be because reads originating from a different locus but had the reference base at that position mapped there preferentially. Because of these reasons and that homozygous alternate sites only represented *ca*. 1% of our data, only sites that were homozygous for the reference allele were included in our mutation rate estimates.

References

Adey A, Morrison HG, Asan, Xun X, Kitzman JO, Turner EH, Stackhouse B, MacKenzie AP, Caruccio NC, Zhang X et al. 2010. Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition. Genome Biol. 11:R119.

Baym M, Kryazhimskiy S, Lieberman TD, Chung H, Desai MM, Kishony R. 2015. Inexpensive multiplexed library preparation for megabase-sized genomes. PLoS one 10: e0128036.

Colbourne JK, Pfrender ME, Gilbert D, Thomas WK, Tucker A, Oakley TH, Tokishita S, Aerts A, Arnold GJ, Basu MK et al. . 2011. The ecoresponsive genome of Daphnia pulex. Science. 331:555-561.

DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, Del Angel G, Rivas MA, Hanna M. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 43:491-498.

Doyle JJ. 1987. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. Phytochem Bull. 19:11-15.

Fields PD, Reisser C, Dukić M, Haag CR, Ebert D. 2015. Genes mirror geography in Daphnia magna. Mol Ecol. 24:4521-4536.

Li H. 2011. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics. 27:2987-2993.

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

Liu X, Han S, Wang Z, Gelernter J, Yang B. 2013. Variant Callers for Next-Generation Sequencing Data: A Comparison Study. PLoS One. 8.

Pirooznia M, Kramer M, Parla J, Goes FS, Potash JB, McCombie WR, Zandi PP. 2014. Validation and assessment of variant calling pipelines for next-generation sequencing. Hum Genomics. 8:14.

Qi W, Nong G, Preston JF, Ben-Ami F, Ebert D. 2009. Comparative metagenomics of Daphnia symbionts. BMC Genomics. 10:1.

Xu S, Ackerman MS, Long H, Bright L, Spitze K, Ramsdell JS, Thomas WK, Lynch M. 2015. A male-specific genetic map of the microcrustacean *Daphnia pulex* based on single-sperm whole-genome sequencing. Genetics. 201:31-38.

Table S1. Putative mutations we tested with Sanger sequencing and how this influenced our mutation processing and filtering. "GATK specific filtering" refers to the initial analysis we performed on the data using only GATK algorithms and performing filtering based on GATK's recommended procedures and expected allele frequencies (see Supplementary Note S2 for details). "GATK + binomial test" refers to filtering variant sites with GATK and then performing a binomial test with the read depth information to determine mutations. "Final filtering" refers to the final filtering regime we implemented: GATK + binomial test with removing regions prone to mismapping and not considering sites whose ancestral state is homozygous alternate from the reference genome (see main text Materials and Methods for details).

Mutation coordinates	**Expected mutation; filtering stage**	**Result**	**Conclusions**
Scaffold_166: 125659	Hom-Hom deletion in C07; GATK specific filtering	No deletion found	Remove regions of the genome that are prone to mismapping and ensure new mutations are statistically supported by read counts
Scaffold_1: 1139609	Hom-Het insertion in C37; GATK specific filtering	No insertion found	Remove regions of the genome that are prone to mismapping and ensure new mutations are statistically supported by read counts
Scaffold_52: 56793	Hom-Hom SNM in C21; GATK specific filtering	C21 actually homozygous for the ancestral state	Remove regions of the genome that are prone to mismapping and ensure new mutations are statistically supported by read counts
Scaffold_1: 1972261	Hom-Het deletion in C18; GATK specific filtering	C18 actually homozygous for ancestral state	Remove regions of the genome that are prone to mismapping and ensure new mutations are statistically supported by read counts
Scaffold_149: 47783	Hom-Het deletion in C37; GATK + binomial test	C37 actually homozygous for ancestral state	Remove regions of the genome that are prone to mismapping
Scaffold_5: 502034	Hom-Hom SNM in C13; GATK specific filtering	C13 actually homozygous for the ancestral state, which is different from the allele of the reference genome	Do not consider sites that are homozygous alternate from the reference genome
Scaffold_48: 5826	Hom-Het SNM in C07; GATK specific filtering	C07 actually homozygous for the ancestral state	Remove regions of the genome that are prone to mismapping
Scaffold_56: 692559	Het-Het SNM in C35; GATK specific filtering	Ancestral state is actually homozygous	Do not consider sites that are homozygous alternate from the reference genome

Scaffold_6: 1357350	Het-Het SNM in C13; GATK specific filtering	Ancestral state is actually homozygous, no het sites detected across the entire sequenced locus	Remove regions of the genome that are prone to mismapping and ensure new mutations are statistically supported by read counts
Scaffold_18: 471250	Het-Het SNM in C13; GATK specific filtering	Ancestral state is actually homozygous	Remove regions of the genome that are prone to mismapping and ensure new mutations are statistically supported by read counts
Scaffold_42: 500634	Het-Het SNM in C02; GATK specific filtering	C02 actually Het with the same alleles as the ancestor	Remove regions of the genome that are prone to mismapping and ensure new mutations are statistically supported by read counts, do not consider mutations that are claiming to be towards the allele of the reference genome
Scaffold_39: 1357350	Het-Het SNM in C39; GATK specific filtering	Ancestral state is actually homozygous	Remove regions of the genome that are prone to mismapping and ensure new mutations are statistically supported by read counts
Scaffold_3: 3405632	Het-Het SNM in C35; GATK + binomial test	Ancestral state is actually homozygous: no Het sites across the region sequenced	Remove regions of the genome that are prone to mismapping and ensure new mutations are statistically supported by read counts
Scaffold_22: 934589	Hom-Het SNM in C38; GATK specific filtering	C38 actually homozygous for the ancestral state	Do not consider sites that are homozygous alternate from the reference genome
Scaffold_88: 16240	Hom-Het SNM in C03; GATK + binomial test	C03 actually homozygous for the ancestral state	Remove regions of the genome that are prone to mismapping
Scaffold_54: 519150	Hom-Het SNM in C40; GATK + binomial test	C40 actually homozygous for the ancestral state	Remove regions of the genome that are prone to mismapping
Scaffold_114: 211,320	Hom-Het SNM in C21; GATK + binomial test	C21 actually homozygous for the ancestral state	Do not consider sites that are homozygous alternate from the reference genome
Scaffold_114: 211,366	Hom-Het SNM in C21; GATK + binomial test	C21 actually homozygous for the ancestral state	Do not consider sites that are homozygous alternate from the reference genome
Scaffold_12: 1261434	Hom-Het SNM in C25; GATK + binomial test	C25 actually homozygous for the ancestral state	Do not consider sites that are homozygous alternate from the reference genome, remove regions of the genome that are prone to mismapping
Scaffold_12: 1261622	Hom-Het SNM in C25; GATK + binomial test	C38 actually homozygous for the ancestral state	Remove regions of the genome that are prone to mismapping
Scaffold_43: 204656	Hom-Het SNM in C13; GATK + binomial test	Ancestral state actually heterozygous	Remove regions of the genome that are prone to mismapping or not clear if multiple loci or heterozygous
Scaffold_23: 832332	Hom-Het SNM in C07; GATK specific filtering	C07 actually homozygous for the ancestral state	Do not consider sites that are homozygous alternate from the reference genome
Scaffold_211: 25176	Stand alone Het-Hom in C39; GATK specific filtering	C39 actually has both alleles as in the ancestral state	Stand alone Het-Hom mutations are likely sampling error
Scaffold_118: 165343	Stand alone Het-Hom C44; GATK +	Ancestral state actually homozygous: no Het	Stand alone Het-Hom mutations are likely sampling error

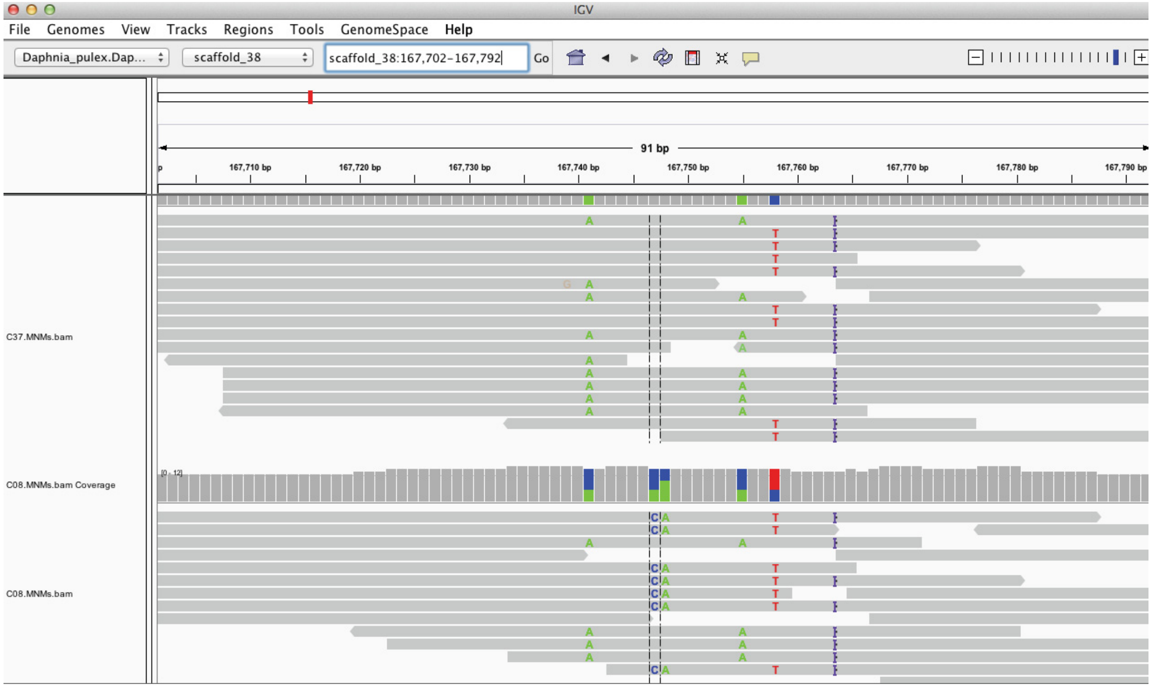
	binomial test	sites across the region sequenced	
Scaffold_20: 1266296	Stand alone Het-Hom in C20; GATK + binomial test	C20 actually has both alleles as in the ancestral state	Stand alone Het-Hom mutations are likely sampling error
Scaffold_27: 89593	Stand alone Het-Hom in C34; GATK + binomial test	Ancestral state actually homozygous: no Het sites across the region sequenced	Stand alone Het-Hom mutations are likely sampling error
Scaffold_5: 1800800	Stand alone Het-Hom in C06; GATK + binomial test	Ancestral state actually homozygous: no Het sites across the region sequenced	Stand alone Het-Hom mutations are likely sampling error
Scaffold_191: 126926	Stand alone Het-Hom in C39; GATK specific filtering	C39 actually has both alleles as in the ancestral state	Stand alone Het-Hom mutations are likely sampling error
Scaffold_87: 507336	Stand alone Het-Hom in C21; GATK specific filtering	C21 actually has both alleles as in the ancestral state	Stand alone Het-Hom mutations are likely sampling error
scaffold_198: 79727	Stand alone Het-Hom in C40; GATK + binomial test	C40 actually has both alleles as in the ancestral state	Stand alone Het-Hom mutations are likely sampling error
scaffold_82: 525708	Stand alone Het-Hom in C02; GATK + binomial test	C02 actually has both alleles as in the ancestral state	Stand alone Het-Hom mutations are likely sampling error
scaffold_39: 28501	Stand alone Het-Hom in C35; GATK + binomial test	C35 actually has both alleles as in the ancestral state	Stand alone Het-Hom mutations are likely sampling error. After adding more depth, this site is no longer called as Het-Hom in C35
Scaffold_62: 446482 - 446792	LOH region in C39 spanning a small region; GATK specific filtering	Ancestral state actually homozygous: no Het sites across the region sequenced	Require Het-Hom sites of LOH region to span at least 100 bp, remove regions of the genome that are prone to mismapping, ensure new mutations are statistically supported by read counts
Scaffold_174: 89183, 89189	LOH region in C38 (spanning a small region); GATK + binomial test	All 4 Het-Hom sites are actually Het in C38	Require Het-Hom sites of LOH region to span at least 100 bp
Scaffold_6: 2256833 - 2256876	LOH region in C34 (spanning a small region); GATK + binomial test	All 5 Het-Hom sites are actually Het in C34	Require Het-Hom sites of LOH region to span at least 100 bp
Scaffold_3: 211785 - 211853	LOH region in C02 (spanning a small region); GATK specific filtering	Ancestral state actually homozygous: no Het sites across the region sequenced	Require Het-Hom sites of LOH region to span at least 100 bp, remove regions of the genome that are prone to mismapping, ensure new mutations are statistically supported by read counts
Scaffold_5: 1109831 - 1109862	LOH region in C20 (spanning a small region); GATK specific filtering	Ancestral state actually homozygous: no Het sites across the region sequenced	Require Het-Hom sites of LOH region to span at least 100 bp, ensure new mutations are statistically supported by read counts
Scaffold_6:	LOH region in C44	Ancestral state actually	Require Het-Hom sites of LOH region to span at

583933 - 583951	(spanning a small region); GATK specific filtering	homozygous: no Het sites across the region sequenced	least 100 bp, ensure new mutations are statistically supported by read counts
Scaffold_48: 123641 - 125331	LOH region in C08; GATK specific filtering and final filtering	Confirmed	
Scaffold_10: 222369-223369	LOH region in C40; GATK specific filtering and final filtering	Confirmed	
Scaffold_10: 401165-401836	LOH region in C40; GATK specific filtering and final filtering	Confirmed	
Scaffold_10: 1294965-1295629	LOH region in C40; GATK specific filtering and final filtering	Confirmed	
Scaffold_24: 166129-166566	LOH region in C40; GATK specific filtering and final filtering	Confirmed	
Scaffold_24: 90076-90887	LOH region in C40; GATK specific filtering and final filtering	Confirmed	
Scaffold_82: 525788	Hom-Het SNM in C07; GATK + binomial test and final filtering	Confirmed	
scaffold_65: 682736	Hom-Het SNM in C14; final filtering	Confirmed	
scaffold_6: 1158702	Hom-Het SNM in C24; final filtering	Confirmed	
scaffold_128: 63511	Hom-Het SNM in C18; final filtering	Confirmed	
scaffold_1: 3073236	Hom-Het SNM in C14; final filtering	High confidence*	
scaffold_20: 250628	Hom-Het SNM in C21; final filtering	Confirmed	
scaffold_4: 2688514	Hom-Het SNM in C40; final filtering	C40 actually does not have the mutation but nearby Het sites were confirmed	False positives due to rare mapping artefacts will still occur to some degree
scaffold_32: 668996	Hom-Het SNM in C20; final filtering	High confidence*	
scaffold_5: 489891	Hom-Het SNM in C38; final filtering	Confirmed	
scaffold_6: 453049	Hom-Het SNM in C20; final filtering	C20 actually does not have the mutation but nearby Het sites were confirmed	False positives due to rare mapping artefacts will still occur to some degree
scaffold_89: 89878	Hom-Het SNM in C39; final filtering	C39 actually does not have the mutation but	False positives due to rare mapping artefacts will still occur to some degree

		nearby Het sites were confirmed	
scaffold_29: 167476	Hom-Het SNM in C21; final filtering	C21 actually does not have the mutation but nearby Het sites were confirmed	False positives due to rare mapping artefacts will still occur to some degree
scaffold_38: 167747, 167748	MNM in C08; final filtering	High confidence*	
scaffold_4: 264803, 264805	MNM in C37; final filtering	Confirmed	
scaffold_2: 3537734	SNM in C12; final filtering	High confidence*	
scaffold_8: 2035679	SNM in C16; final filtering	Confirmed	
scaffold_38: 693601	SNM in C36; final filtering	Confirmed	
scaffold_95: 197858	13 bp deletion in C25; final filtering	Confirmed	
scaffold_14: 1376349	1 bp deletion in C21; final filtering	Confirmed	

*The high confidence mutations were validated with inspection and not refuted with Sanger sequencing. Only 1 of 2 alleles were sequenced with Sanger sequencing resulting from allele-specific amplification. Upon inspection of BAM files, mutations were present with high confidence with no signals of mapping error. Primers were found to contain heterozygous sites, potentially introducing the amplification bias. For all mutations that were analyzed with both Sanger sequencing and by inspecting the BAM files, there was 100% concordance between these methods of validation

A)


IGV
File
Genomes
View
Tracks
Regions
Tools
GenomeSpace
Help
Daphnia_pulex.Dap...
scaffold_38
scaffold_38:167,702-167,792
Go
91 bp
167,710 bp
167,720 bp
167,730 bp
167,740 bp
167,750 bp
167,760 bp
167,770 bp
167,780 bp
167,790 bp
C37.MNMs.bam
C08.MNMs.bam Coverage
[0 - 12]
C08.MNMs.bam

B)

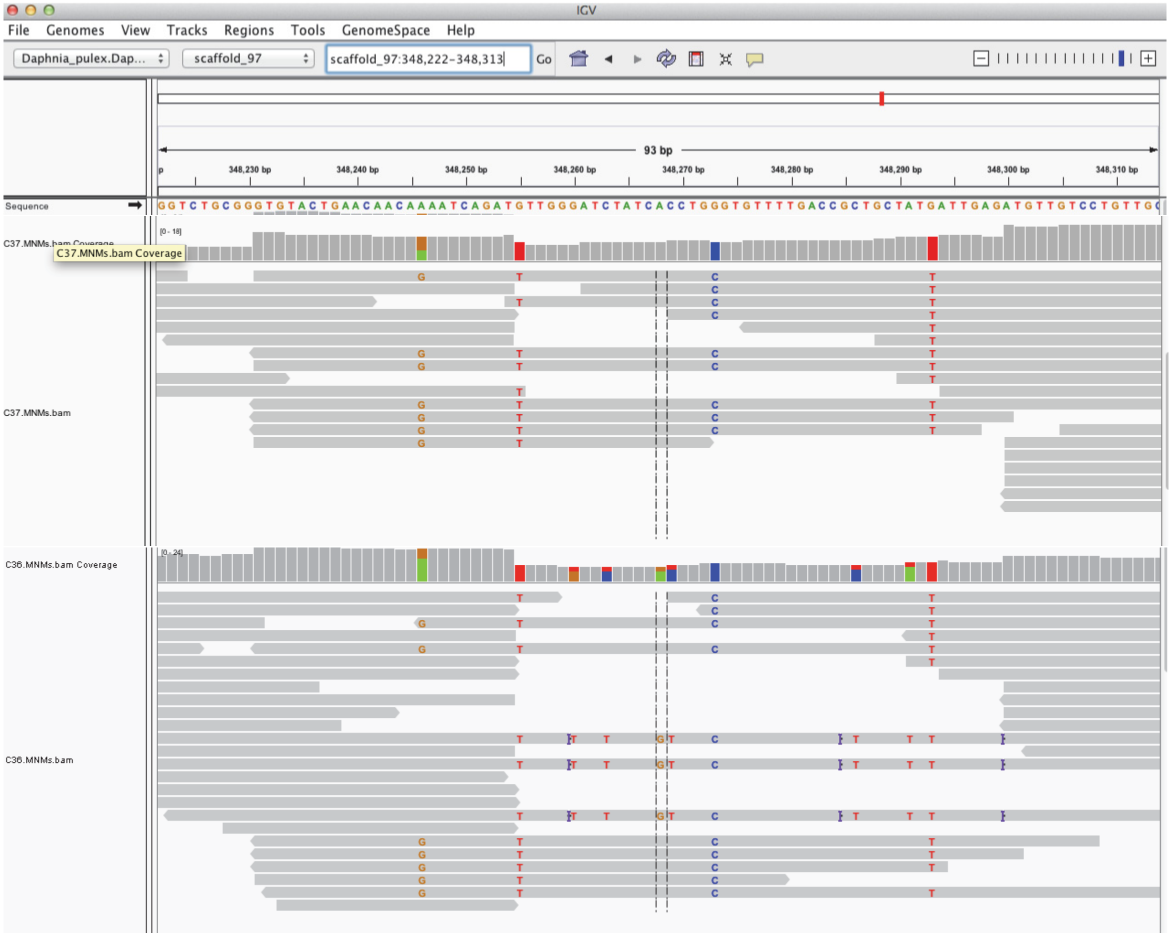

IGV
File
Genomes
View
Tracks
Regions
Tools
GenomeSpace
Help
Daphnia_pulex.Dap...
scaffold_97
scaffold_97:348,222-348,313
Go
93 bp
348,230 bp
348,240 bp
348,250 bp
348,260 bp
348,270 bp
348,280 bp
348,290 bp
348,300 bp
348,310 bp
Sequence
GGTCTGCGGGTGTACTGAACAACAAAATCAGATGTTGGGATCTATCACCTGGGTGTTTTGACCGCTGCTATGATTGAGATGTTGTCCTGTTG
C37.MNMs.bam Coverage
[0 - 18]
C37.MNMs.bam Coverage
C37.MNMs.bam
C36.MNMs.bam Coverage
[0 - 24]
C36.MNMs.bam

Fig S1: Screenshots of putative multinucleotide mutations (MNMs) viewed in IGV. Gray horizontal bars represent reads, and coloured bases are those different from the reference genome. (A) An example of a real multinucleotide mutation. The top panel shows the reads mapping to a MA line not containing the MNM, C37. The bottom panel shows the reads mapping to the MA line harbouring the MNM, C08. The first novel SNM in C08 is between the vertical lines, and the second one is immediately adjacent. The new mutations are linked to the existing polymorphism in the allele upon which the MNM arose, and this existing polymorphism is also present in the other lines not containing the MNM. (B) An example of an artifactual MNM. The top panel shows the reads mapping to a MA line not containing the MNM, C37. The bottom panel shows the reads mapping to the line harbouring the putative MNM, C36. The reads containing the MNM also contain many other variants (including large insertions) that are not present in the other MA lines, suggesting that these reads originated from a different locus. Another read exactly like this also mapped to a different line (not shown), further supporting that this is a mapping error.

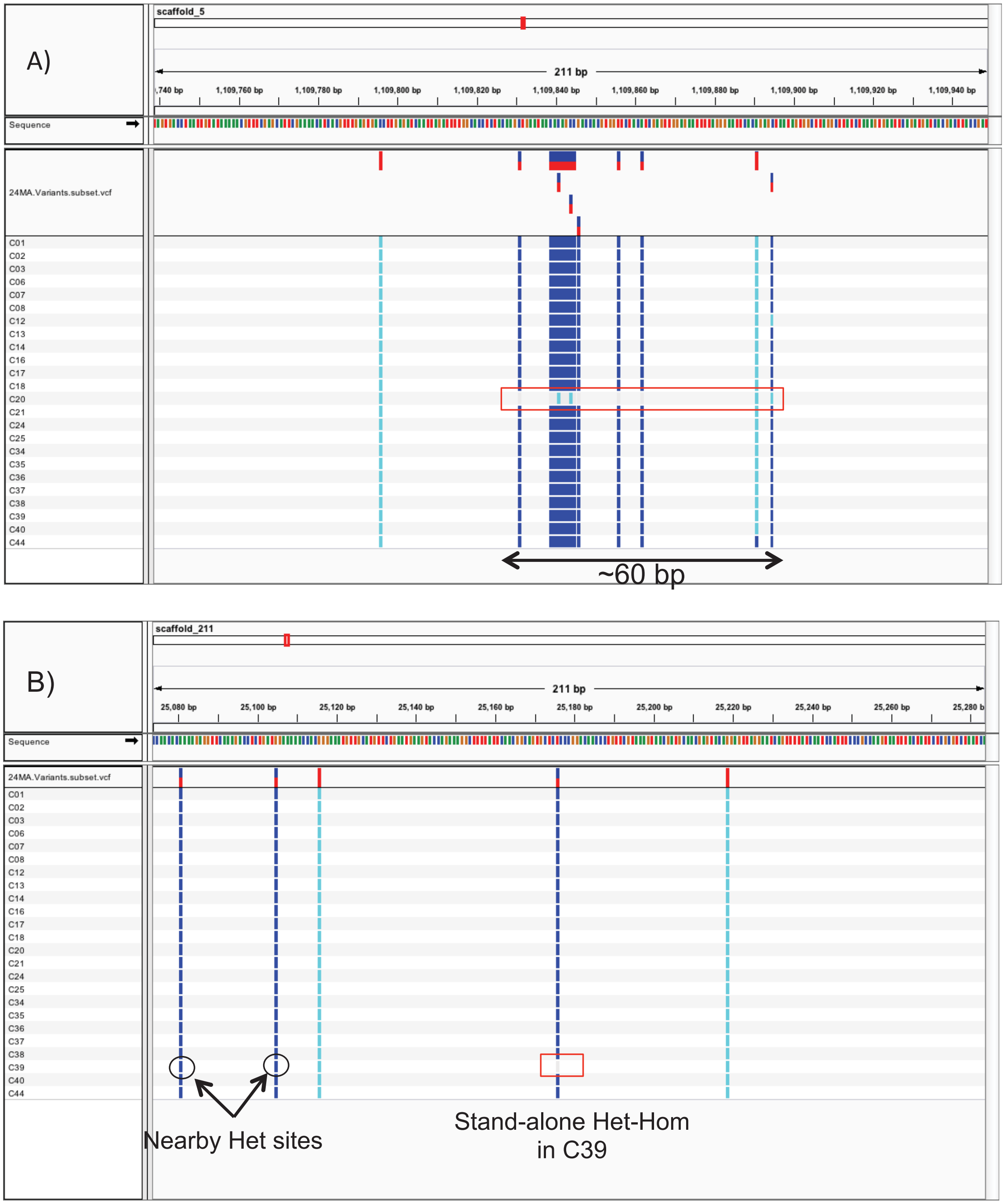

A)
scaffold_5
211 bp
,740 bp
1,109,760 bp
1,109,780 bp
1,109,800 bp
1,109,820 bp
1,109,840 bp
1,109,860 bp
1,109,880 bp
1,109,900 bp
1,109,920 bp
1,109,940 bp
Sequence
24MA.Variants.subset.vcf
C01
C02
C03
C06
C07
C08
C12
C13
C14
C16
C17
C18
C20
C21
C24
C25
C34
C35
C36
C37
C38
C39
C40
C44
~60 bp
B)
scaffold_211
211 bp
25,080 bp
25,100 bp
25,120 bp
25,140 bp
25,160 bp
25,180 bp
25,200 bp
25,220 bp
25,240 bp
25,260 bp
25,280 b
Sequence
24MA.Variants.subset.vcf
C01
C02
C03
C06
C07
C08
C12
C13
C14
C16
C17
C18
C20
C21
C24
C25
C34
C35
C36
C37
C38
C39
C40
C44
Nearby Het sites
Stand-alone Het-Hom
in C39

Fig S2

Fig S2: Screenshots of types of putative loss of heterozygosity that we excluded. Colored vertical bars represent the genotype of a given MA line at the given genomic position. Gray represents homozygous for the reference allele, blue represents heterozygous, and cyan represents homozygous for the alternate allele. (A) A LOH region that only spans a small stretch in one MA line and is likely due to mismapped reads or sampling bias of this allele in this region. This 60 bp region contains 8 clustered Het-Hom sites and there are no other Het-Hom sites in the flanking region. (B) A single "stand-alone" Het-Hom site in one MA line, where there are sites nearby that are heterozygous.

Table S2. Information for all the SNMs (single nucleotide mutations) that were detected and passed filtering. "Anc" refers to the ancestral genotype of the MA lines at that particular site, "Mut" refers to the mutation genotype and MA line, "Chr" refers to the chromosome location, "Reg" the functional region, and "Ann" the snpEff annotation. The "-" character is used when the information is not available or applicable.

Scaffold	Start	End	Anc	Mut	Chr	Reg	Gene	Ann
scaffold_1	1197114	1197115	GG	C20:GT:6-2	chr2	intergenic	-	intergenic_region
scaffold_1	1576516	1576517	CC	C24:CA:5-2	chr2	intronic	DAPPUDRAFT_300056	upstream_gene_variant
scaffold_1	1600835	1600836	CC	C37:TC:4-4	chr2	exonic	DAPPUDRAFT_94124	missense_variant
scaffold_1	1913702	1913703	GG	C18:GA:6-5	chr2	exonic	DAPPUDRAFT_309722	synonymous_variant
scaffold_1	1948270	1948271	CC	C21:CA:6-3	chr2	exonic	DAPPUDRAFT_232322	missense_variant
scaffold_1	2229649	2229650	GG	C44:AG:5-3	chr2	intronic	DAPPUDRAFT_39998	splice_acceptor_variant
scaffold_1	2337170	2337171	GG	C39:GT:8-2	chr2	exonic	DAPPUDRAFT_309815	synonymous_variant
scaffold_1	2476249	2476250	CC	C02:AC:4-4	chr2	intronic	DAPPUDRAFT_309841	upstream_gene_variant
scaffold_1	2498145	2498146	GG	C37:GT:5-2	chr2	exonic	DAPPUDRAFT_20910	missense_variant
scaffold_1	2733144	2733145	CC	C38:CG:5-3	chr2	exonic	DAPPUDRAFT_309887	missense_variant
scaffold_1	2755028	2755029	CC	C06:CT:11-5	chr2	intergenic	-	downstream_gene_variant
scaffold_1	2796469	2796470	GG	C14:AG:8-6	chr2	intergenic	-	upstream_gene_variant
scaffold_1	3073235	3073236	GG	C14:AG:8-7	chr2	exonic	DAPPUDRAFT_94409	synonymous_variant
scaffold_1	3247481	3247482	TT	C38:TC:5-2	chr2	intronic	DAPPUDRAFT_299950	upstream_gene_variant
scaffold_1	3337524	3337525	GG	C03:GA:11-4	chr2	intergenic	-	upstream_gene_variant
scaffold_1	3454480	3454481	CC	C44:AC:9-6	chr2	intergenic	-	intergenic_region
scaffold_1	3535008	3535009	GG	C03:GA:13-8	chr2	intergenic	-	intergenic_region
scaffold_1	3594258	3594259	CC	C13:CA:3-2	chr2	intergenic	-	intergenic_region
scaffold_1	3763031	3763032	GG	C20:GT:3-2	chr2	exonic	DAPPUDRAFT_232685	missense_variant
scaffold_1	3804740	3804741	GG	C20:GT:6-2	chr2	exonic	DAPPUDRAFT_232692	missense_variant
scaffold_1	3967242	3967243	AA	C13:TA:7-4	chr2	intronic	DAPPUDRAFT_205517	intron_variant
scaffold_1	45676	45677	TT	C06:TC:10-7	chr2	intronic	DAPPUDRAFT_231944	intron_variant
scaffold_1	838076	838077	CC	C25:CG:9-4	chr2	intergenic	-	intergenic_region
scaffold_10	1151913	1151914	CC	C06:CA:4-2	chr11	exonic	DAPPUDRAFT_314053	3_prime_UTR_variant
scaffold_10	116025	116026	TT	C13:TC:6-2	chr11	intergenic	-	upstream_gene_variant
scaffold_10	1201744	1201745	GG	C25:TG:5-3	chr11	exonic	DAPPUDRAFT_98857	missense_variant
scaffold_10	1341715	1341716	CC	C20:CA:11-3	chr11	intergenic	-	upstream_gene_variant
scaffold_10	1581610	1581611	AA	C36:AT:6-2	chr11	intronic	DAPPUDRAFT_314176	upstream_gene_variant
scaffold_10	1612754	1612755	GG	C38:GT:3-2	chr11	exonic	DAPPUDRAFT_237981	missense_variant
scaffold_10	1671942	1671943	TT	C25:TC:6-5	chr11	intronic	DAPPUDRAFT_20190	intron_variant
scaffold_10	1849931	1849932	GG	C21:AG:4-3	chr11	exonic	DAPPUDRAFT_300453	missense_variant
scaffold_10	24614	24615	AA	C34:GA:8-3	chr11	intronic	DAPPUDRAFT_313768	splice_donor_variant
scaffold_10	525532	525533	TT	C34:TC:5-3	chr11	intergenic	-	upstream_gene_variant
scaffold_10	525536	525537	CC	C34:CT:5-3	chr11	intergenic	-	upstream_gene_variant
scaffold_10	738238	738239	GG	C14:GT:8-2	chr11	intergenic	-	upstream_gene_variant
scaffold_10	897237	897238	AA	C39:AT:6-6	chr11	exonic	DAPPUDRAFT_300387	missense_variant
scaffold_10	930023	930024	TT	C36:TG:10-6	chr11	intergenic	-	upstream_gene_variant

scaffold_100	194237	194238	GG	C40:GA:3-2	chr10	exonic	DAPPUDRAFT_112805	synonymous_variant
scaffold_100	432944	432945	CC	C44:CA:4-2	chr10	intergenic	-	upstream_gene_variant
scaffold_102	398700	398701	GG	C03:GA:6-2	chr10	exonic	DAPPUDRAFT_112977	synonymous_variant
scaffold_103	323234	323235	CC	C25:CA:3-2	chr10	intergenic	-	upstream_gene_variant
scaffold_104	17349	17350	GG	C38:TG:2-2	chr9	intergenic	-	upstream_gene_variant
scaffold_105	197752	197753	GG	C20:GT:3-2	chr10	intergenic	-	intergenic_region
scaffold_105	201134	201135	GG	C36:GA:6-2	chr10	intergenic	-	upstream_gene_variant
scaffold_105	97548	97549	GG	C14:GT:5-2	chr10	intergenic	-	upstream_gene_variant
scaffold_106	208044	208045	AA	C18:AG:5-2	chr1	exonic	DAPPUDRAFT_328636	missense_variant
scaffold_106	235839	235840	GG	C39:AG:2-2	chr1	intergenic	-	upstream_gene_variant
scaffold_106	447892	447893	CC	C21:CA:3-2	chr1	intergenic	-	intergenic_region
scaffold_107	189128	189129	GG	C01:AG:9-7	chr9	exonic	DAPPUDRAFT_300674	3_prime_UTR_variant
scaffold_107	228439	228440	CC	C38:CA:9-2	chr9	intergenic	-	upstream_gene_variant
scaffold_109	273887	273888	CC	C35:GC:11-6	chr8	exonic	DAPPUDRAFT_328806	missense_variant
scaffold_11	1092793	1092794	CC	C38:CA:4-2	chr4	intronic	DAPPUDRAFT_209250	downstream_gene_variant
scaffold_11	122229	122230	GG	C20:AG:4-2	chr4	intergenic	-	downstream_gene_variant
scaffold_11	2091333	2091334	GG	C06:GT:5-2	chr4	unknown	-	-
scaffold_11	2143090	2143091	GG	C40:GT:4-2	chr4	intergenic	-	intergenic_region
scaffold_11	294479	294480	CC	C18:TC:7-4	chr4	intronic	DAPPUDRAFT_238167	intron_variant
scaffold_11	443853	443854	AA	C14:AG:9-9	chr4	intronic	DAPPUDRAFT_238184	intron_variant
scaffold_11	811912	811913	AA	C08:GA:6-5	chr4	exonic	DAPPUDRAFT_314379	missense_variant
scaffold_11	850029	850030	AA	C14:GA:12-8	chr4	intronic	DAPPUDRAFT_46327	upstream_gene_variant
scaffold_111	245350	245351	CC	C08:CT:9-4	chr11	exonic	DAPPUDRAFT_328913	missense_variant
scaffold_111	271179	271180	GG	C39:AG:4-4	chr11	intronic	DAPPUDRAFT_202746	intron_variant
scaffold_113	134668	134669	AA	C40:GA:4-3	chr2	intergenic	-	upstream_gene_variant
scaffold_114	249785	249786	TT	C37:GT:8-6	chr3	exonic	DAPPUDRAFT_30479	missense_variant
scaffold_114	252782	252783	GG	C13:AG:6-5	chr3	intergenic	-	upstream_gene_variant
scaffold_116	122470	122471	CC	C39:CT:5-2	-	intergenic	-	downstream_gene_variant
scaffold_116	299591	299592	GG	C20:GT:4-2	-	unknown	-	-
scaffold_118	226409	226410	TT	C36:TA:8-2	chr9	exonic	DAPPUDRAFT_113862	missense_variant
scaffold_12	1481607	1481608	CC	C25:CA:6-2	chr5	exonic	DAPPUDRAFT_238845	synonymous_variant
scaffold_12	162205	162206	CC	C25:TC:8-5	chr5	intergenic	-	upstream_gene_variant
scaffold_12	2111993	2111994	CC	C25:CA:4-2	chr5	exonic	DAPPUDRAFT_314929	stop_gained
scaffold_12	2113698	2113699	GG	C21:GA:8-2	chr5	intergenic	-	upstream_gene_variant
scaffold_12	310737	310738	TT	C34:TC:6-2	chr5	intronic	DAPPUDRAFT_314623	upstream_gene_variant
scaffold_12	598341	598342	GG	C37:GA:3-2	chr5	intronic	DAPPUDRAFT_314679	upstream_gene_variant
scaffold_12	605118	605119	CC	C44:CA:7-3	chr5	exonic	DAPPUDRAFT_99499	missense_variant
scaffold_12	610351	610352	AA	C40:GA:6-5	chr5	exonic	DAPPUDRAFT_46968	synonymous_variant
scaffold_12	760768	760769	AA	C07:AG:6-4	chr5	exonic	DAPPUDRAFT_314699	missense_variant
scaffold_12	761918	761919	CC	C02:CA:8-3	chr5	exonic	DAPPUDRAFT_238698	synonymous_variant
scaffold_12	906483	906484	CC	C01:TC:10-8	chr5	exonic	DAPPUDRAFT_209520	synonymous_variant
scaffold_12	974378	974379	GG	C25:GT:5-2	chr5	exonic	DAPPUDRAFT_314736	missense_variant
scaffold_121	374253	374254	CC	C25:CA:5-2	chr10	exonic	DAPPUDRAFT_228840	missense_variant
scaffold_1226	6160	6161	AA	C35:AT:8-2	-	intergenic	-	upstream_gene_variant
scaffold_124	327886	327887	GG	C08:GA:11-10	-	intergenic	-	downstream_gene_variant

scaffold_124	371883	371884	CC	C21:TC:7-6	-	intergenic	-	upstream_gene_variant
scaffold_128	63510	63511	CC	C18:AC:7-3	chr6	intronic	DAPPUDRAFT_259734	intron_variant
scaffold_129	45809	45810	CC	C24:CT:9-3	chr7	intergenic	-	upstream_gene_variant
scaffold_13	1174026	1174027	CC	C34:CT:5-2	chr9	intergenic	-	downstream_gene_variant
scaffold_13	1282332	1282333	GG	C03:GT:7-2	chr9	exonic	DAPPUDRAFT_315172	missense_variant
scaffold_13	1284333	1284334	GG	C35:AG:8-6	chr9	intergenic	-	upstream_gene_variant
scaffold_13	1369951	1369952	GG	C40:GA:6-3	chr9	intronic	DAPPUDRAFT_239330	intron_variant
scaffold_13	1690513	1690514	CC	C18:CA:7-2	chr9	intergenic	-	upstream_gene_variant
scaffold_13	1716997	1716998	CC	C39:TC:5-2	chr9	exonic	DAPPUDRAFT_239397	synonymous_variant
scaffold_13	196442	196443	CC	C24:CA:6-2	chr9	intronic	DAPPUDRAFT_194528	intron_variant
scaffold_13	426243	426244	CC	C24:CA:5-2	chr9	intergenic	-	upstream_gene_variant
scaffold_13	500777	500778	CC	C40:CA:5-2	chr9	intergenic	-	upstream_gene_variant
scaffold_133	125016	125017	GG	C24:GT:5-2	chr7	exonic	DAPPUDRAFT_330083	missense_variant
scaffold_137	304458	304459	GG	C25:GT:3-2	-	exonic	DAPPUDRAFT_330287	missense_variant
scaffold_14	1507208	1507209	GG	C03:GA:6-3	chr9	intronic	DAPPUDRAFT_315602	upstream_gene_variant
scaffold_14	365924	365925	GG	C37:GT:8-2	chr9	intergenic	-	upstream_gene_variant
scaffold_14	431438	431439	TT	C39:TC:2-2	chr9	intergenic	-	intergenic_region
scaffold_14	849228	849229	GG	C20:GT:10-6	chr9	intergenic	-	upstream_gene_variant
scaffold_141	55825	55826	AA	C01:AG:11-7	-	intergenic	-	upstream_gene_variant
scaffold_143	193711	193712	GG	C39:GA:6-2	-	exonic	DAPPUDRAFT_330565	missense_variant
scaffold_144	207945	207946	GG	C25:GA:4-2	-	intronic	DAPPUDRAFT_330596	upstream_gene_variant
scaffold_144	260310	260311	GG	C21:GT:5-2	-	intergenic	-	upstream_gene_variant
scaffold_144	260312	260313	AA	C21:AT:5-2	-	intergenic	-	upstream_gene_variant
scaffold_145	141473	141474	CC	C02:CT:3-2	chr7	exonic	DAPPUDRAFT_229110	missense_variant
scaffold_145	23692	23693	TT	C38:TA:5-2	chr7	exonic	DAPPUDRAFT_330617	missense_variant
scaffold_147	278113	278114	GG	C20:GT:5-2	-	intergenic	-	upstream_gene_variant
scaffold_149	63606	63607	CC	C37:CT:12-8	-	intronic	DAPPUDRAFT_115247	upstream_gene_variant
scaffold_1496	8968	8969	TT	C13:TA:9-7	-	intergenic	-	upstream_gene_variant
scaffold_1499	2842	2843	CC	C08:TC:2-2	-	exonic	DAPPUDRAFT_301818	synonymous_variant
scaffold_15	116126	116127	GG	C40:GT:4-2	chr5	intergenic	-	upstream_gene_variant
scaffold_15	1211667	1211668	TT	C18:TC:3-2	chr5	intergenic	-	upstream_gene_variant
scaffold_15	1345126	1345127	GG	C37:GT:6-2	chr5	intronic	DAPPUDRAFT_223446	downstream_gene_variant
scaffold_15	285257	285258	GG	C40:AG:5-4	chr5	unknown	-	-
scaffold_15	513423	513424	GG	C34:AG:4-4	chr5	intergenic	-	upstream_gene_variant
scaffold_15	940602	940603	CC	C40:CA:5-2	chr5	intergenic	-	downstream_gene_variant
scaffold_151	224605	224606	TT	C24:TC:8-5	-	intergenic	-	intergenic_region
scaffold_158	27473	27474	CC	C06:TC:13-10	chr12	exonic	DAPPUDRAFT_115522	synonymous_variant
scaffold_16	1065361	1065362	GG	C44:GT:7-2	chr3	intergenic	-	upstream_gene_variant
scaffold_16	1305646	1305647	CC	C08:CT:11-7	chr3	intergenic	-	upstream_gene_variant
scaffold_16	1458639	1458640	CC	C21:CA:6-2	chr3	intergenic	-	upstream_gene_variant
scaffold_16	430889	430890	AA	C14:AG:9-2	chr3	intronic	DAPPUDRAFT_1297	intron_variant
scaffold_16	489122	489123	CC	C44:CA:9-2	chr3	exonic	DAPPUDRAFT_100839	missense_variant
scaffold_160	130654	130655	CC	C13:CT:12-5	chr5	intergenic	-	intergenic_region
scaffold_166	172739	172740	CC	C36:TC:7-5	-	intergenic	-	upstream_gene_variant
scaffold_17	1119949	1119950	CC	C16:TC:11-8	chr10	exonic	DAPPUDRAFT_48987	missense_variant

scaffold_17	1320096	1320097	CC	C37:CA:5-2	chr10	intergenic	-	upstream_gene_variant
scaffold_17	194663	194664	CC	C39:CA:4-2	chr10	intronic	DAPPUDRAFT_48686	downstream_gene_variant
scaffold_17	579412	579413	CC	C17:CA:6-2	chr10	intronic	DAPPUDRAFT_240818	splice_acceptor_variant
scaffold_17	857229	857230	GG	C21:GT:3-2	chr10	intergenic	-	upstream_gene_variant
scaffold_170	31447	31448	GG	C39:GT:5-2	chr2	intergenic	-	upstream_gene_variant
scaffold_173	134058	134059	CC	C25:CA:11-3	chr10	intergenic	-	upstream_gene_variant
scaffold_178	162035	162036	GG	C40:GT:5-2	chr3	exonic	DAPPUDRAFT_331990	synonymous_variant
scaffold_1780	1418	1419	GG	C37:GT:5-2	-	exonic	DAPPUDRAFT_337785	missense_variant
scaffold_18	1133997	1133998	GG	C06:GT:9-2	chr7	exonic	DAPPUDRAFT_241460	missense_variant
scaffold_18	173083	173084	CC	C17:CT:5-2	chr7	intergenic	-	upstream_gene_variant
scaffold_18	429259	429260	TT	C13:TG:9-7	chr7	intergenic	-	downstream_gene_variant
scaffold_18	604112	604113	CC	C37:AC:2-2	chr7	exonic	DAPPUDRAFT_316737	stop_gained
scaffold_184	118789	118790	TT	C16:TG:8-2	-	intergenic	-	downstream_gene_variant
scaffold_184	141751	141752	GG	C07:TG:11-5	-	intergenic	-	downstream_gene_variant
scaffold_184	161323	161324	GG	C34:GA:5-2	-	intergenic	-	intergenic_region
scaffold_185	21428	21429	CC	C17:CA:7-2	-	intergenic	-	upstream_gene_variant
scaffold_186	146613	146614	CC	C08:TC:6-3	chr4	intergenic	-	intergenic_region
scaffold_187	137207	137208	CC	C20:CA:3-2	chr6	intergenic	-	downstream_gene_variant
scaffold_187	176438	176439	GG	C25:GT:5-2	chr6	exonic	DAPPUDRAFT_302747	3_prime_UTR_variant
scaffold_19	1041496	1041497	TT	C39:TA:7-4	chr2	intronic	DAPPUDRAFT_210921	intron_variant
scaffold_19	1361039	1361040	TT	C35:TC:9-8	chr2	intronic	DAPPUDRAFT_317166	intron_variant
scaffold_19	952392	952393	CC	C17:CA:9-3	chr2	unknown	-	-
scaffold_198	148678	148679	GG	C44:GT:7-2	chr4	intergenic	-	intergenic_region
scaffold_2	1190346	1190347	CC	C38:CT:6-3	chr3	intronic	DAPPUDRAFT_233047	upstream_gene_variant
scaffold_2	1376001	1376002	GG	C35:GA:14-11	chr3	intergenic	-	downstream_gene_variant
scaffold_2	1511202	1511203	AA	C08:TA:8-6	chr3	intergenic	-	upstream_gene_variant
scaffold_2	1931741	1931742	CC	C21:CT:4-2	chr3	intronic	DAPPUDRAFT_233164	intron_variant
scaffold_2	19333	19334	AA	C35:CA:12-11	chr3	exonic	DAPPUDRAFT_310073	synonymous_variant
scaffold_2	2201612	2201613	GG	C01:GC:17-7	chr3	exonic	DAPPUDRAFT_40742	missense_variant
scaffold_2	2680059	2680060	AA	C01:AG:5-2	chr3	intergenic	-	upstream_gene_variant
scaffold_2	2813068	2813069	AA	C35:TA:12-7	chr3	exonic	DAPPUDRAFT_40462	missense_variant
scaffold_2	3008317	3008318	TT	C38:CT:7-5	chr3	intergenic	-	downstream_gene_variant
scaffold_2	3195392	3195393	CC	C37:AC:2-2	chr3	exonic	DAPPUDRAFT_233379	missense_variant
scaffold_2	3238321	3238322	CC	C03:CA:5-2	chr3	exonic	DAPPUDRAFT_303225	missense_variant
scaffold_2	3537733	3537734	GG	C12:AG:7-6	chr3	exonic	DAPPUDRAFT_220651	3_prime_UTR_variant
scaffold_2	431097	431098	CC	C25:CA:6-2	chr3	exonic	DAPPUDRAFT_310145	missense_variant
scaffold_2	471561	471562	GG	C36:GT:10-3	chr3	intronic	DAPPUDRAFT_232903	intron_variant
scaffold_2	605587	605588	GG	C14:GA:11-4	chr3	exonic	DAPPUDRAFT_303118	synonymous_variant
scaffold_20	1075015	1075016	GG	C06:GT:3-2	chr5	intronic	DAPPUDRAFT_49977	intron_variant
scaffold_20	1077828	1077829	AA	C07:AC:5-2	chr5	intronic	DAPPUDRAFT_49977	downstream_gene_variant
scaffold_20	115782	115783	GG	C24:GT:4-2	chr8	intergenic	-	upstream_gene_variant
scaffold_20	1346339	1346340	GG	C44:GA:7-3	chr5	exonic	DAPPUDRAFT_317413	missense_variant
scaffold_20	250627	250628	CC	C21:TC:7-5	chr5	intergenic	-	upstream_gene_variant
scaffold_20	916467	916468	AA	C02:AG:6-3	chr5	exonic	DAPPUDRAFT_224138	missense_variant
scaffold_202	163442	163443	CC	C39:CA:4-2	-	intergenic	-	upstream_gene_variant

scaffold_205	147712	147713	CC	C02:CT:10-3	-	intergenic	-	intergenic_region
scaffold_206	52211	52212	GG	C17:GC:8-7	chr2	intergenic	-	upstream_gene_variant
scaffold_21	1059608	1059609	CC	C37:CA:9-3	chr3	exonic	DAPPUDRAFT_303428	missense_variant
scaffold_21	1132826	1132827	CC	C40:CA:5-2	chr3	intronic	DAPPUDRAFT_188084	upstream_gene_variant
scaffold_21	1205119	1205120	GG	C34:GA:5-2	chr3	intronic	DAPPUDRAFT_317679	intron_variant
scaffold_21	546439	546440	GG	C36:GT:3-2	chr3	intergenic	-	upstream_gene_variant
scaffold_21	552303	552304	GG	C37:GT:6-2	chr3	intergenic	-	upstream_gene_variant
scaffold_210	82276	82277	CC	C20:CA:5-2	-	intronic	DAPPUDRAFT_303518	upstream_gene_variant
scaffold_219	62155	62156	GG	C25:GA:8-2	-	intronic	DAPPUDRAFT_117170	intron_variant
scaffold_22	474703	474704	TT	C39:TG:2-2	chr8	intergenic	-	downstream_gene_variant
scaffold_220	63716	63717	TT	C44:CT:7-3	chr7	intronic	DAPPUDRAFT_264571	upstream_gene_variant
scaffold_220	83268	83269	GG	C44:TG:2-2	chr7	exonic	DAPPUDRAFT_333111	missense_variant
scaffold_221	478	479	AA	C02:AG:8-4	chr3	intergenic	-	downstream_gene_variant
scaffold_223	88372	88373	GG	C36:GA:18-8	chr10	intergenic	-	upstream_gene_variant
scaffold_228	43838	43839	GG	C25:GT:8-3	chr2	intergenic	-	intergenic_region
scaffold_228	61723	61724	GG	C37:GT:5-2	chr2	intergenic	-	intergenic_region
scaffold_228	63872	63873	CC	C24:TC:3-2	chr2	intergenic	-	intergenic_region
scaffold_23	1131200	1131201	GG	C44:GT:4-2	chr6	intronic	DAPPUDRAFT_318159	intron_variant
scaffold_23	198416	198417	GG	C12:GA:14-5	chr6	exonic	DAPPUDRAFT_317988	synonymous_variant
scaffold_23	72724	72725	CC	C07:TC:8-8	chr6	intergenic	-	upstream_gene_variant
scaffold_235	94477	94478	GG	C35:TG:11-9	-	exonic	DAPPUDRAFT_333422	missense_variant
scaffold_236	93345	93346	TT	C12:GT:14-6	-	intronic	DAPPUDRAFT_333439	upstream_gene_variant
scaffold_237	112087	112088	TT	C06:TC:7-2	chr2	intergenic	-	downstream_gene_variant
scaffold_237	70063	70064	CC	C38:CA:5-2	chr2	intronic	DAPPUDRAFT_265088	downstream_gene_variant
scaffold_24	1072246	1072247	CC	C03:CT:4-2	chr11	exonic	DAPPUDRAFT_318379	synonymous_variant
scaffold_24	1110391	1110392	AA	C20:AT:12-7	chr11	intronic	DAPPUDRAFT_103198	intron_variant
scaffold_24	1135134	1135135	AA	C07:AT:7-2	chr11	intergenic	-	upstream_gene_variant
scaffold_24	1282693	1282694	CC	C17:CT:7-5	chr11	exonic	DAPPUDRAFT_318416	synonymous_variant
scaffold_24	467015	467016	GG	C16:GT:5-2	chr11	intergenic	-	upstream_gene_variant
scaffold_24	607321	607322	CC	C20:CA:5-2	chr11	intronic	DAPPUDRAFT_318263	intron_variant
scaffold_241	79228	79229	GG	C17:GA:7-6	-	intergenic	-	downstream_gene_variant
scaffold_244	89439	89440	GG	C06:GA:16-6	-	intronic	DAPPUDRAFT_333566	downstream_gene_variant
scaffold_245	36016	36017	GG	C37:AG:5-5	-	exonic	DAPPUDRAFT_65757	missense_variant
scaffold_25	522220	522221	GG	C40:GT:3-2	chr8	intergenic	-	upstream_gene_variant
scaffold_25	74402	74403	CC	C44:CA:5-2	chr8	intergenic	-	upstream_gene_variant
scaffold_25	966213	966214	CC	C25:TC:7-2	chr8	intergenic	-	downstream_gene_variant
scaffold_252	39692	39693	AA	C01:TA:11-8	-	intronic	DAPPUDRAFT_333654	upstream_gene_variant
scaffold_26	105113	105114	GG	C37:AG:4-4	chr3	intergenic	-	upstream_gene_variant
scaffold_26	146745	146746	CC	C38:CT:9-4	chr3	intergenic	-	upstream_gene_variant
scaffold_26	529769	529770	CC	C08:TC:8-6	chr3	intergenic	-	upstream_gene_variant
scaffold_26	698585	698586	AA	C07:AT:11-5	chr3	exonic	DAPPUDRAFT_304178	missense_variant
scaffold_27	1163698	1163699	CC	C20:CA:4-2	chr12	exonic	DAPPUDRAFT_212220	missense_variant
scaffold_27	136171	136172	TT	C06:AT:3-2	chr12	intergenic	-	upstream_gene_variant
scaffold_27	474288	474289	CC	C21:TC:4-3	chr12	intergenic	-	upstream_gene_variant
scaffold_27	518343	518344	GG	C24:GC:4-2	chr12	intergenic	-	downstream_gene_variant

scaffold_275	36801	36802	CC	C03:CT:9-5	chr5	intergenic	-	upstream_gene_variant
scaffold_28	510874	510875	GG	C20:GT:5-2	chr6	intronic	DAPPUDRAFT_319191	intron_variant
scaffold_28	595217	595218	AA	C34:AC:6-3	chr6	intergenic	-	upstream_gene_variant
scaffold_28	604150	604151	GG	C37:GT:7-2	chr6	exonic	DAPPUDRAFT_188248	missense_variant
scaffold_28	951474	951475	GG	C38:GT:6-2	chr6	intergenic	-	downstream_gene_variant
scaffold_283	25785	25786	GG	C25:GT:4-2	-	intergenic	-	intergenic_region
scaffold_289	10589	10590	AA	C34:AT:9-8	chr9	intergenic	-	upstream_gene_variant
scaffold_289	65185	65186	AA	C24:AG:6-3	chr9	intergenic	-	upstream_gene_variant
scaffold_29	1201562	1201563	GG	C36:GA:19-10	chr10	intergenic	-	upstream_gene_variant
scaffold_29	167475	167476	GG	C21:GA:6-2	chr10	exonic	DAPPUDRAFT_304469	synonymous_variant
scaffold_29	471251	471252	CC	C37:CT:14-8	chr10	exonic	DAPPUDRAFT_244673	missense_variant
scaffold_29	782883	782884	GG	C07:GC:5-3	chr10	intergenic	-	downstream_gene_variant
scaffold_3	1326982	1326983	GG	C02:GT:6-3	chr1	intergenic	-	downstream_gene_variant
scaffold_3	1337934	1337935	GG	C38:GT:5-2	chr1	intronic	DAPPUDRAFT_304684	intron_variant
scaffold_3	1472966	1472967	AG	C37:A,T:8,2	chr1	intergenic	-	upstream_gene_variant
scaffold_3	1610643	1610644	CC	C08:CT:9-3	chr1	exonic	DAPPUDRAFT_191143	stop_gained
scaffold_3	3139058	3139059	CC	C14:GC:12-4	chr1	intergenic	-	upstream_gene_variant
scaffold_3	3449333	3449334	AA	C07:AC:7-6	chr1	intronic	DAPPUDRAFT_304786	upstream_gene_variant
scaffold_3	468686	468687	AA	C18:AG:5-2	chr1	exonic	DAPPUDRAFT_95321	missense_variant
scaffold_3	836431	836432	AA	C38:AT:8-2	chr1	intronic	DAPPUDRAFT_310774	upstream_gene_variant
scaffold_3	836432	836433	GG	C38:GA:8-2	chr1	intronic	DAPPUDRAFT_310774	upstream_gene_variant
scaffold_30	141929	141930	TT	C06:TC:7-3	chr2	intergenic	-	intergenic_region
scaffold_30	347216	347217	GG	C36:AG:10-9	chr2	intronic	DAPPUDRAFT_244900	downstream_gene_variant
scaffold_30	534017	534018	CC	C12:CA:11-3	chr2	exonic	DAPPUDRAFT_319599	synonymous_variant
scaffold_31	1023897	1023898	GG	C14:GA:3-2	chr4	exonic	DAPPUDRAFT_2876	missense_variant
scaffold_31	366948	366949	TT	C39:TC:6-2	chr4	exonic	DAPPUDRAFT_188325	synonymous_variant
scaffold_31	613492	613493	GG	C36:GA:10-4	chr4	exonic	DAPPUDRAFT_319776	missense_variant
scaffold_31	850818	850819	GG	C44:GT:5-2	chr4	intergenic	-	intergenic_region
scaffold_32	38370	38371	GG	C38:GA:9-3	chr6	intergenic	-	upstream_gene_variant
scaffold_32	498664	498665	GG	C40:GT:4-2	chr6	intronic	DAPPUDRAFT_245388	intron_variant
scaffold_32	668995	668996	TT	C20:TC:7-4	chr6	exonic	DAPPUDRAFT_305004	missense_variant
scaffold_32	680913	680914	AA	C37:AT:11-3	chr6	intronic	DAPPUDRAFT_4209	upstream_gene_variant
scaffold_32	867759	867760	GG	C20:GT:6-3	chr6	intergenic	-	intergenic_region
scaffold_32	869690	869691	CC	C34:CA:5-2	chr6	intergenic	-	intergenic_region
scaffold_33	1061324	1061325	CC	C38:CA:3-2	chr8	intergenic	-	upstream_gene_variant
scaffold_33	1096983	1096984	CC	C20:CA:4-2	chr8	intergenic	-	intergenic_region
scaffold_33	289464	289465	AA	C03:CA:7-3	-	intergenic	-	upstream_gene_variant
scaffold_33	39117	39118	GG	C08:GT:3-2	chr3	intergenic	-	upstream_gene_variant
scaffold_34	210471	210472	CC	C38:CA:11-3	chr8	intronic	DAPPUDRAFT_105016	downstream_gene_variant
scaffold_35	150941	150942	TT	C34:TC:8-3	chr10	intronic	DAPPUDRAFT_24785	intron_variant
scaffold_35	166675	166676	GG	C18:TG:8-7	chr10	intergenic	-	upstream_gene_variant
scaffold_35	18769	18770	GG	C44:GT:5-2	chr10	intronic	DAPPUDRAFT_320348	downstream_gene_variant
scaffold_35	193609	193610	CC	C40:TC:2-2	chr10	intergenic	-	intergenic_region
scaffold_35	314140	314141	CC	C44:CA:3-2	chr10	intronic	DAPPUDRAFT_105173	upstream_gene_variant
scaffold_35	495758	495759	GG	C44:GT:6-2	chr10	intergenic	-	upstream_gene_variant

scaffold_35	619037	619038	TT	C06:TC:8-7	chr10	intergenic	-	upstream_gene_variant
scaffold_35	953106	953107	CC	C20:CA:4-2	chr10	intergenic	-	downstream_gene_variant
scaffold_355	65004	65005	TT	C38:CT:8-7	-	intergenic	-	upstream_gene_variant
scaffold_356	12032	12033	AA	C07:AT:5-2	-	intergenic	-	upstream_gene_variant
scaffold_36	399322	399323	GG	C20:GA:8-6	chr10	intergenic	-	upstream_gene_variant
scaffold_36	465642	465643	CC	C21:CA:5-2	chr10	exonic	DAPPUDRAFT_20600	missense_variant
scaffold_36	638452	638453	CC	C18:TC:10-8	chr10	exonic	DAPPUDRAFT_225476	5_prime_UTR_variant
scaffold_36	833518	833519	CC	C37:CT:4-3	chr10	exonic	DAPPUDRAFT_320673	missense_variant
scaffold_36	888726	888727	AA	C21:AG:9-2	chr10	exonic	DAPPUDRAFT_53488	missense_variant
scaffold_36	888727	888728	GG	C21:GA:9-2	chr10	exonic	DAPPUDRAFT_53488	synonymous_variant
scaffold_37	166654	166655	GG	C17:GA:7-2	chr5	intronic	DAPPUDRAFT_320781	upstream_gene_variant
scaffold_37	213999	214000	GG	C01:GA:21-13	chr5	intergenic	-	upstream_gene_variant
scaffold_37	959525	959526	CC	C44:CA:3-2	chr5	intronic	DAPPUDRAFT_320914	intron_variant
scaffold_37	966080	966081	CC	C25:CA:5-2	chr5	exonic	DAPPUDRAFT_320914	missense_variant
scaffold_38	167746	167747	AA	C08:CA:7-4	chr5	intergenic	-	intergenic_region
scaffold_38	167747	167748	CC	C08:AC:7-4	chr5	intergenic	-	intergenic_region
scaffold_38	524863	524864	GG	C36:GT:4-2	chr5	intergenic	-	intergenic_region
scaffold_38	682282	682283	CC	C40:CA:3-2	chr5	intergenic	-	upstream_gene_variant
scaffold_38	693600	693601	GG	C36:AG:7-7	chr5	intergenic	-	downstream_gene_variant
scaffold_39	389354	389355	GG	C37:AG:13-7	chr5	intergenic	-	intergenic_region
scaffold_39	921769	921770	TT	C13:GT:13-10	chr5	intronic	DAPPUDRAFT_321218	upstream_gene_variant
scaffold_4	1707800	1707801	CC	C44:CA:3-2	chr7	intronic	DAPPUDRAFT_42004	upstream_gene_variant
scaffold_4	2066952	2066953	CC	C44:CT:6-2	chr7	intergenic	-	upstream_gene_variant
scaffold_4	2177513	2177514	CC	C36:CA:5-2	chr7	exonic	DAPPUDRAFT_92028	missense_variant
scaffold_4	221748	221749	GG	C01:GC:15-11	chr7	exonic	DAPPUDRAFT_230293	missense_variant
scaffold_4	2226565	2226566	GG	C37:GA:6-2	chr7	intronic	DAPPUDRAFT_41891	upstream_gene_variant
scaffold_4	264802	264803	TT	C37:CT:9-7	chr7	exonic	DAPPUDRAFT_305574	synonymous_variant
scaffold_4	264804	264805	CC	C37:AC:8-8	chr7	exonic	DAPPUDRAFT_305574	stop_gained
scaffold_4	2688513	2688514	CC	C40:CA:11-3	chr7	intronic	DAPPUDRAFT_311743	intron_variant
scaffold_4	3025458	3025459	TT	C35:TC:19-9	chr7	exonic	DAPPUDRAFT_192054	missense_variant
scaffold_4	718243	718244	CC	C25:GC:7-5	chr7	intronic	DAPPUDRAFT_311351	downstream_gene_variant
scaffold_4	760810	760811	GG	C37:GT:11-3	chr7	intronic	DAPPUDRAFT_305606	intron_variant
scaffold_4	90959	90960	TT	C08:TA:12-4	chr7	intergenic	-	intergenic_region
scaffold_4	90961	90962	TT	C08:TC:12-4	chr7	intergenic	-	intergenic_region
scaffold_4	90965	90966	CC	C08:CT:12-4	chr7	intergenic	-	intergenic_region
scaffold_40	198251	198252	AA	C25:AG:8-5	chr7	exonic	DAPPUDRAFT_106003	missense_variant
scaffold_40	221478	221479	GG	C13:GA:4-2	chr7	exonic	DAPPUDRAFT_106011	missense_variant
scaffold_40	67269	67270	GG	C06:GT:3-2	chr7	exonic	DAPPUDRAFT_321261	missense_variant
scaffold_405	33024	33025	TT	C07:TC:13-5	-	intergenic	-	downstream_gene_variant
scaffold_41	218851	218852	GG	C14:GT:5-2	chr9	exonic	DAPPUDRAFT_305827	synonymous_variant
scaffold_41	357360	357361	AA	C25:AG:3-3	chr9	intergenic	-	upstream_gene_variant
scaffold_41	451018	451019	CC	C13:CA:4-2	chr9	intergenic	-	upstream_gene_variant
scaffold_41	710010	710011	TT	C40:TC:8-2	chr9	intronic	DAPPUDRAFT_106242	upstream_gene_variant
scaffold_42	917965	917966	GG	C44:GA:10-7	chr1	exonic	DAPPUDRAFT_321748	stop_gained
scaffold_42	964398	964399	GG	C39:AG:2-2	chr1	intergenic	-	intergenic_region

scaffold_4245	3276	3277	GG	C18:GT:5-2	-	exonic	DAPPUDRAFT_340400	missense_variant
scaffold_43	301110	301111	GG	C44:GC:3-2	chr4	unknown	-	-
scaffold_43	607333	607334	CC	C44:CA:8-3	chr4	intergenic	-	upstream_gene_variant
scaffold_44	394411	394412	CC	C06:CT:13-5	chr3	intergenic	-	upstream_gene_variant
scaffold_44	423384	423385	AA	C40:AG:9-8	chr3	intronic	DAPPUDRAFT_321968	upstream_gene_variant
scaffold_44	502276	502277	CC	C20:CG:8-4	chr3	intergenic	-	upstream_gene_variant
scaffold_44	758423	758424	GG	C25:GT:5-2	chr3	intergenic	-	upstream_gene_variant
scaffold_44	785244	785245	TT	C18:GT:7-2	chr3	intergenic	-	upstream_gene_variant
scaffold_44	825910	825911	GG	C20:TG:2-2	chr3	intronic	DAPPUDRAFT_248586	upstream_gene_variant
scaffold_45	490173	490174	CC	C35:GC:7-6	-	exonic	DAPPUDRAFT_248756	missense_variant
scaffold_452	7587	7588	AA	C24:CA:8-6	-	intergenic	-	upstream_gene_variant
scaffold_46	295556	295557	GG	C40:GA:11-7	chr7	exonic	DAPPUDRAFT_231196	missense_variant
scaffold_46	504356	504357	GG	C38:GT:4-2	chr7	exonic	DAPPUDRAFT_55080	missense_variant
scaffold_46	878153	878154	CC	C21:CT:5-4	chr7	intergenic	-	upstream_gene_variant
scaffold_47	15470	15471	GG	C40:TG:8-8	chr6	intronic	DAPPUDRAFT_249084	intron_variant
scaffold_47	678945	678946	GG	C38:GT:6-2	chr6	exonic	DAPPUDRAFT_226110	missense_variant
scaffold_47	98163	98164	AA	C40:AG:3-2	chr6	intronic	DAPPUDRAFT_249107	upstream_gene_variant
scaffold_48	284366	284367	TT	C35:TA:11-3	chr7	intergenic	-	downstream_gene_variant
scaffold_48	336810	336811	CC	C37:CA:4-2	chr7	intergenic	-	upstream_gene_variant
scaffold_48	469699	469700	GG	C18:AG:8-8	chr7	exonic	DAPPUDRAFT_322596	missense_variant
scaffold_48	587930	587931	GG	C38:GT:4-2	chr7	exonic	DAPPUDRAFT_107269	start_lost
scaffold_48	633311	633312	TT	C21:TC:6-2	chr7	intergenic	-	downstream_gene_variant
scaffold_49	194648	194649	GG	C38:GA:9-5	chr10	intergenic	-	upstream_gene_variant
scaffold_49	270378	270379	GG	C37:GT:5-2	chr10	intronic	DAPPUDRAFT_107345	splice_region_variant
scaffold_49	305853	305854	AA	C40:AT:6-4	chr10	exonic	DAPPUDRAFT_322732	3_prime_UTR_variant
scaffold_49	434026	434027	CC	C38:TC:8-8	chr10	intergenic	-	downstream_gene_variant
scaffold_49	674903	674904	GG	C40:GT:3-2	-	intergenic	-	intergenic_region
scaffold_5	102492	102493	GG	C14:GT:6-4	chr12	exonic	DAPPUDRAFT_311854	missense_variant
scaffold_5	1398946	1398947	CC	C20:CA:5-2	chr12	intergenic	-	upstream_gene_variant
scaffold_5	1634937	1634938	CC	C21:CT:9-4	chr12	intergenic	-	intergenic_region
scaffold_5	2334164	2334165	CC	C20:CA:8-3	chr12	exonic	DAPPUDRAFT_306572	synonymous_variant
scaffold_5	2366283	2366284	GG	C13:TG:4-4	chr12	intergenic	-	upstream_gene_variant
scaffold_5	2417828	2417829	GG	C39:AG:9-3	chr12	exonic	DAPPUDRAFT_312180	3_prime_UTR_variant
scaffold_5	464061	464062	TT	C40:TC:9-4	chr12	intergenic	-	upstream_gene_variant
scaffold_5	489890	489891	GG	C38:AG:14-8	chr12	intronic	DAPPUDRAFT_42745	intron_variant
scaffold_5	874976	874977	TT	C12:TA:9-2	chr12	intronic	DAPPUDRAFT_306499	upstream_gene_variant
scaffold_51	442022	442023	CC	C39:TC:7-7	chr9	intergenic	-	upstream_gene_variant
scaffold_52	457586	457587	CC	C25:CA:3-2	chr8	intergenic	-	upstream_gene_variant
scaffold_52	479362	479363	CC	C24:CT:10-5	chr8	intergenic	-	downstream_gene_variant
scaffold_53	296941	296942	CC	C18:CT:10-8	chr1	exonic	DAPPUDRAFT_306791	missense_variant
scaffold_53	776441	776442	AA	C35:AG:19-15	chr1	intergenic	-	upstream_gene_variant
scaffold_53	797361	797362	CC	C20:CA:5-2	chr1	intronic	DAPPUDRAFT_199461	intron_variant
scaffold_53	83832	83833	GG	C35:AG:14-13	chr1	exonic	DAPPUDRAFT_306783	synonymous_variant
scaffold_54	519149	519150	GG	C40:GT:8-2	chr7	exonic	DAPPUDRAFT_226406	missense_variant
scaffold_56	133418	133419	CC	C13:TC:7-7	chr10	intergenic	-	upstream_gene_variant

scaffold_56	330927	330928	GG	C16:GA:7-5	chr10	intronic	DAPPUDRAFT_306917	intron_variant
scaffold_56	463482	463483	GG	C13:AG:7-5	chr10	intronic	DAPPUDRAFT_226482	upstream_gene_variant
scaffold_56	649690	649691	GG	C38:GT:5-2	chr10	exonic	DAPPUDRAFT_306904	missense_variant
scaffold_57	213867	213868	AA	C38:AT:5-2	chr8	intergenic	-	upstream_gene_variant
scaffold_57	270916	270917	CC	C07:TC:5-4	chr8	intergenic	-	upstream_gene_variant
scaffold_57	384185	384186	CC	C25:CT:5-3	-	intergenic	-	downstream_gene_variant
scaffold_574	35304	35305	AA	C39:AG:3-2	-	intergenic	-	upstream_gene_variant
scaffold_58	397448	397449	AA	C20:TA:6-4	chr12	intronic	DAPPUDRAFT_30674	intron_variant
scaffold_58	469447	469448	CC	C13:TC:7-7	chr12	intronic	DAPPUDRAFT_306989	upstream_gene_variant
scaffold_58	79088	79089	GG	C20:GT:5-2	chr12	exonic	DAPPUDRAFT_56709	missense_variant
scaffold_59	610959	610960	AA	C34:AT:9-2	chr12	intronic	DAPPUDRAFT_56750	downstream_gene_variant
scaffold_6	1158701	1158702	AA	C24:GA:4-3	chr10	exonic	DAPPUDRAFT_307126	missense_variant
scaffold_6	1339044	1339045	CC	C37:CA:3-2	chr10	exonic	DAPPUDRAFT_43397	missense_variant
scaffold_6	1363344	1363345	TT	C44:TA:7-2	chr10	intergenic	-	intergenic_region
scaffold_6	1844064	1844065	TT	C13:TC:9-2	chr10	intronic	DAPPUDRAFT_97337	upstream_gene_variant
scaffold_6	1952206	1952207	GG	C21:GC:3-2	chr10	intergenic	-	upstream_gene_variant
scaffold_6	1952373	1952374	CC	C14:CT:4-2	chr10	intergenic	-	upstream_gene_variant
scaffold_6	197410	197411	GG	C21:GT:6-2	chr10	intergenic	-	downstream_gene_variant
scaffold_6	1980189	1980190	CC	C44:CT:8-5	chr10	exonic	DAPPUDRAFT_187429	missense_variant
scaffold_6	2320053	2320054	TT	C25:TG:6-4	chr10	intergenic	-	downstream_gene_variant
scaffold_6	453048	453049	GG	C20:GT:5-2	chr10	exonic	DAPPUDRAFT_97016	missense_variant
scaffold_6	603172	603173	GG	C36:GA:5-2	chr10	intronic	DAPPUDRAFT_312326	splice_region_variant
scaffold_6	659647	659648	CC	C20:CA:3-2	chr10	exonic	DAPPUDRAFT_312340	missense_variant
scaffold_6	681913	681914	CC	C34:TC:9-2	chr10	exonic	DAPPUDRAFT_192569	missense_variant
scaffold_6	971589	971590	CC	C21:AC:6-2	chr10	exonic	DAPPUDRAFT_312411	missense_variant
scaffold_60	160598	160599	CC	C20:CA:5-2	chr5	intronic	DAPPUDRAFT_199801	intron_variant
scaffold_60	701935	701936	GG	C37:GA:4-2	chr5	intronic	DAPPUDRAFT_108778	upstream_gene_variant
scaffold_61	569667	569668	GG	C18:GT:6-2	chr1	intergenic	-	upstream_gene_variant
scaffold_62	43502	43503	CC	C25:TC:9-4	chr3	intergenic	-	intergenic_region
scaffold_62	701394	701395	GG	C13:GT:10-9	chr3	intergenic	-	upstream_gene_variant
scaffold_63	620676	620677	GG	C13:GT:4-2	chr12	intronic	DAPPUDRAFT_57582	downstream_gene_variant
scaffold_63	645199	645200	TT	C25:TA:5-3	chr12	intergenic	-	intergenic_region
scaffold_63	92522	92523	GG	C34:GA:6-2	chr12	exonic	DAPPUDRAFT_215208	3_prime_UTR_variant
scaffold_64	491817	491818	AA	C25:TA:7-5	chr10	intergenic	-	downstream_gene_variant
scaffold_65	243714	243715	GG	C12:AG:5-5	chr7	intergenic	-	upstream_gene_variant
scaffold_65	372776	372777	CC	C38:CT:11-5	chr7	exonic	DAPPUDRAFT_57725	missense_variant
scaffold_65	468713	468714	GG	C35:GT:8-3	chr7	intergenic	-	downstream_gene_variant
scaffold_65	682735	682736	CC	C14:CT:7-4	chr7	intronic	DAPPUDRAFT_307566	intron_variant
scaffold_657	7813	7814	GG	C35:GT:14-10	-	intergenic	-	downstream_gene_variant
scaffold_66	380005	380006	CC	C20:CA:4-2	chr1	intergenic	-	upstream_gene_variant
scaffold_66	656372	656373	GG	C14:GT:9-2	chr1	intronic	DAPPUDRAFT_307583	upstream_gene_variant
scaffold_66	8803	8804	AA	C08:AG:2-2	chr1	intronic	DAPPUDRAFT_324913	upstream_gene_variant
scaffold_68	396217	396218	CC	C36:CT:11-7	chr7	intronic	DAPPUDRAFT_253101	upstream_gene_variant
scaffold_69	364929	364930	GG	C44:GA:4-2	-	intergenic	-	downstream_gene_variant
scaffold_69	439136	439137	CC	C38:CA:5-2	-	intergenic	-	upstream_gene_variant

scaffold_7	1213999	1214000	GG	C13:GA:11-4	chr8	intronic	DAPPUDRAFT_236362	upstream_gene_variant
scaffold_7	1216680	1216681	GG	C25:GT:5-2	chr8	exonic	DAPPUDRAFT_236363	missense_variant
scaffold_7	1479786	1479787	GG	C20:GT:5-2	chr8	intergenic	-	upstream_gene_variant
scaffold_7	1824735	1824736	CC	C02:CA:6-2	chr8	intronic	DAPPUDRAFT_307889	intron_variant
scaffold_7	2198145	2198146	CC	C37:CA:10-3	chr8	intronic	DAPPUDRAFT_44451	downstream_gene_variant
scaffold_7	397513	397514	GG	C13:GT:5-2	chr8	unknown	-	-
scaffold_7	842054	842055	CC	C20:CA:5-2	chr8	exonic	DAPPUDRAFT_307840	missense_variant
scaffold_7	900157	900158	GG	C39:GA:7-5	chr8	exonic	DAPPUDRAFT_312824	synonymous_variant
scaffold_705	38880	38881	CC	C37:CG:5-2	-	intergenic	-	upstream_gene_variant
scaffold_72	301983	301984	CC	C12:TC:2-2	chr10	exonic	DAPPUDRAFT_110207	missense_variant
scaffold_73	265213	265214	TT	C13:CT:6-3	chr2	intergenic	-	upstream_gene_variant
scaffold_74	181366	181367	CC	C08:CT:5-4	chr2	intergenic	-	intergenic_region
scaffold_74	408943	408944	GG	C40:GT:5-2	chr2	intronic	DAPPUDRAFT_58603	upstream_gene_variant
scaffold_74	512568	512569	CC	C18:TC:4-4	chr2	intergenic	-	upstream_gene_variant
scaffold_76	177645	177646	CC	C24:CA:6-2	chr8	intergenic	-	upstream_gene_variant
scaffold_76	324756	324757	GG	C01:GA:8-5	chr8	intronic	DAPPUDRAFT_326110	intron_variant
scaffold_76	67710	67711	TT	C01:AT:20-14	chr8	exonic	DAPPUDRAFT_326077	missense_variant
scaffold_77	337355	337356	CC	C39:CT:5-2	chr8	exonic	DAPPUDRAFT_254414	missense_variant
scaffold_77	374066	374067	GG	C08:AG:8-4	chr8	intergenic	-	downstream_gene_variant
scaffold_77	469788	469789	GG	C34:GT:8-2	chr8	intergenic	-	intergenic_region
scaffold_77	85487	85488	CC	C40:CA:5-2	chr8	intergenic	-	upstream_gene_variant
scaffold_78	103489	103490	CC	C20:CA:4-2	chr2	exonic	DAPPUDRAFT_326254	missense_variant
scaffold_78	205380	205381	GG	C17:GT:7-2	chr2	intergenic	-	upstream_gene_variant
scaffold_78	408147	408148	GG	C40:TG:2-2	chr2	intergenic	-	upstream_gene_variant
scaffold_78	456310	456311	CC	C16:TC:11-4	chr2	exonic	DAPPUDRAFT_110863	missense_variant
scaffold_78	503084	503085	GG	C17:GT:11-3	chr2	intergenic	-	upstream_gene_variant
scaffold_788	12101	12102	TT	C24:TA:4-2	-	intergenic	-	downstream_gene_variant
scaffold_79	463547	463548	TT	C07:TA:5-2	chr10	intergenic	-	upstream_gene_variant
scaffold_8	1254722	1254723	GG	C44:GA:14-9	chr4	exonic	DAPPUDRAFT_313219	missense_variant
scaffold_8	1303506	1303507	GG	C37:GT:5-2	chr4	intronic	DAPPUDRAFT_236855	upstream_gene_variant
scaffold_8	1354537	1354538	CC	C01:CT:15-9	chr4	exonic	DAPPUDRAFT_98029	missense_variant
scaffold_8	1600606	1600607	CC	C39:CG:6-6	chr4	intergenic	-	upstream_gene_variant
scaffold_8	1767632	1767633	GG	C13:GT:5-2	chr4	intergenic	-	downstream_gene_variant
scaffold_8	1913339	1913340	CC	C34:CA:8-7	chr4	intronic	DAPPUDRAFT_308512	upstream_gene_variant
scaffold_8	2035678	2035679	GG	C16:GA:9-8	chr4	exonic	DAPPUDRAFT_308524	missense_variant
scaffold_8	2182227	2182228	CC	C35:CT:9-2	chr4	intergenic	-	downstream_gene_variant
scaffold_8	462881	462882	CC	C38:CA:8-3	chr4	intergenic	-	intergenic_region
scaffold_8	686666	686667	GG	C07:TG:2-2	chr4	exonic	DAPPUDRAFT_44942	missense_variant
scaffold_80	16734	16735	CC	C35:CT:10-8	chr6	intergenic	-	downstream_gene_variant
scaffold_80	413024	413025	GG	C21:AG:2-2	chr6	intergenic	-	upstream_gene_variant
scaffold_81	384105	384106	AA	C25:TA:7-5	-	intergenic	-	intergenic_region
scaffold_81	92938	92939	GG	C16:GT:6-2	-	intergenic	-	upstream_gene_variant
scaffold_82	525787	525788	CC	C07:TC:5-5	chr7	exonic	DAPPUDRAFT_308628	missense_variant
scaffold_83	116416	116417	TT	C40:TA:4-2	chr8	exonic	DAPPUDRAFT_308639	5_prime_UTR_variant
scaffold_83	189143	189144	GG	C21:GT:8-2	chr8	intergenic	-	upstream_gene_variant

scaffold_83	575487	575488	GG	C44:GT:6-2	chr8	intergenic	-	upstream_gene_variant
scaffold_84	489958	489959	CC	C39:CG:4-2	chr12	intronic	DAPPUDRAFT_326847	upstream_gene_variant
scaffold_85	112263	112264	CC	C07:CT:5-2	-	intergenic	-	upstream_gene_variant
scaffold_87	11086	11087	GG	C24:GA:12-8	chr7	intergenic	-	downstream_gene_variant
scaffold_89	552493	552494	CC	C36:CT:6-2	chr5	exonic	DAPPUDRAFT_111902	synonymous_variant
scaffold_89	559160	559161	GG	C44:GT:5-2	chr5	intergenic	-	intergenic_region
scaffold_89	569789	569790	AA	C03:AT:4-2	chr5	intergenic	-	intergenic_region
scaffold_89	57081	57082	TT	C17:TC:7-2	chr5	intergenic	-	upstream_gene_variant
scaffold_89	636399	636400	CC	C08:TC:5-2	chr5	intergenic	-	intergenic_region
scaffold_89	636401	636402	TT	C08:AT:5-2	chr5	intergenic	-	intergenic_region
scaffold_89	89877	89878	GG	C39:GA:8-4	chr5	intergenic	-	upstream_gene_variant
scaffold_9	1255620	1255621	GG	C35:GA:17-7	chr9	intergenic	-	upstream_gene_variant
scaffold_9	1263920	1263921	CC	C02:CG:11-5	chr9	exonic	DAPPUDRAFT_98439	missense_variant
scaffold_9	1534047	1534048	GG	C20:GT:6-2	chr9	intergenic	-	upstream_gene_variant
scaffold_9	1605105	1605106	GG	C13:GA:6-3	chr9	exonic	DAPPUDRAFT_237423	missense_variant
scaffold_9	1772304	1772305	CC	C38:CA:6-2	chr9	intergenic	-	upstream_gene_variant
scaffold_9	1836863	1836864	CC	C13:CT:8-3	chr9	intergenic	-	upstream_gene_variant
scaffold_9	2037340	2037341	GG	C13:GA:7-2	chr9	intronic	DAPPUDRAFT_21055	intron_variant
scaffold_9	263705	263706	GG	C25:GT:5-2	chr9	intergenic	-	upstream_gene_variant
scaffold_9	647697	647698	AA	C17:CA:5-3	chr9	intergenic	-	downstream_gene_variant
scaffold_9	761999	762000	CC	C40:CT:3-2	chr9	intergenic	-	upstream_gene_variant
scaffold_9	826523	826524	GG	C35:GT:4-2	chr9	exonic	DAPPUDRAFT_313559	synonymous_variant
scaffold_90	297555	297556	CC	C44:CA:8-2	chr2	exonic	DAPPUDRAFT_111959	missense_variant
scaffold_91	368976	368977	AA	C16:AG:9-2	chr7	intergenic	-	upstream_gene_variant
scaffold_93	279751	279752	AA	C06:TA:4-3	chr7	intergenic	-	downstream_gene_variant
scaffold_95	146079	146080	CC	C08:TC:6-2	chr9	intergenic	-	upstream_gene_variant
scaffold_95	315013	315014	TT	C40:TC:7-2	chr9	intergenic	-	upstream_gene_variant
scaffold_96	271988	271989	TT	C25:TA:4-2	chr12	intergenic	-	upstream_gene_variant
scaffold_96	274105	274106	AA	C40:AC:8-2	chr12	intergenic	-	upstream_gene_variant
scaffold_97	329591	329592	AA	C34:AG:7-4	chr3	exonic	DAPPUDRAFT_309391	3_prime_UTR_variant
scaffold_98	366429	366430	AA	C13:AG:6-5	chr5	intergenic	-	upstream_gene_variant
scaffold_99	208098	208099	GG	C25:GA:6-4	chr9	intronic	DAPPUDRAFT_309440	upstream_gene_variant
scaffold_99	388409	388410	GG	C20:GT:4-2	chr9	intronic	DAPPUDRAFT_328360	splice_region_variant
scaffold_99	403975	403976	AA	C18:AC:16-6	chr9	intergenic	-	intergenic_region
scaffold_99	437936	437937	AA	C25:AG:11-10	chr9	intergenic	-	upstream_gene_variant
scaffold_923	9233	9234	GG	CC4:7-11,CC6:7-10,CC7:17,10:GT	-	intergenic	-	downstream_gene_variant
scaffold_56	709585	709586	CC	CT:CC3:4-7	chr10	intergenic	-	upstream_gene_variant
scaffold_32	508809	508810	CC	CC3:CA:8-2	chr6	exonic	DAPPUDRAFT_245388	missense_variant
scaffold_31	355262	355263	GG	CC3:GA:7-7	chr4	intergenic	-	upstream_gene_variant
scaffold_23	581135	581136	GG	CC4:9-3,CC6:7-6,CC7:8-8:GA	chr6	intronic	DAPPUDRAFT_303742	downstream_gene_variant
scaffold_13	800030	800031	CC	CC4:CT:10-5	chr9	exonic	DAPPUDRAFT_315087	missense_variant
scaffold_2	950324	950325	CC	CC4:13-	chr3	intronic	DAPPUDRAFT_233011	upstream_gene_variant

				10,CC6:13-10,CC7:9-11:CT				
scaffold_1	1076197	1076198	GG	CC3:GA:7-8	chr2	intergenic	-	intergenic_region
scaffold_61	148746	148747	GG	CC8:GA:5-12	chr1	intergenic	-	downstream_gene_variant
scaffold_44	85912	85913	GG	CC8:GA:2-2	chr3	intergenic	-	upstream_gene_variant
scaffold_41	466990	466991	TT	CC9:TG:9-7	chr9	intergenic	-	upstream_gene_variant
scaffold_27	156600	156601	AA	CC3:AT:10-4	chr12	intronic	DAPPUDRAFT_318955	upstream_gene_variant
scaffold_21	652849	652850	CC	CC9:CG:9-6	chr3	exonic	DAPPUDRAFT_102408	synonymous_variant
scaffold_16	1425354	1425355	CC	CC8:CT:5-8	chr3	intergenic	-	intergenic_region
scaffold_7	981708	981709	GG	CC8:GA:11-4	chr8	intergenic	-	upstream_gene_variant
scaffold_3	358640	358641	GG	CC8:GA:13-11	chr1	exonic	DAPPUDRAFT_30734	stop_gained
scaffold_41	464983	464984	CC	CC3:CA:3-2	chr9	intergenic	-	downstream_gene_variant
scaffold_18	628951	628952	CC	CC3:CT:9-14	chr7	intergenic	-	upstream_gene_variant
scaffold_9	340058	340059	GG	CC8:GA:5-2	chr9	exonic	DAPPUDRAFT_309044	synonymous_variant
scaffold_5	689830	689831	GG	CC3:GA:5-5	chr12	exonic	DAPPUDRAFT_311909	synonymous_variant

Table S3. Information for all the indels (insertions and deletions) that were detected and passed filtering. All columns are the same as Table S2 except for the addition of the "Size" column, which indicates the size of the indel in base pairs: (+) represents and insertion and (-) represents a deletion.

Scaffold	Start	End	Anc	Mut	Size	Chr	Reg	Gene	Ann
scaffold_95	197858	197872	CTTGTTTTCACCAA/CTTGTTTTCACCAA	C25:CTTGTTTTCACCAA/C:6-3	-13	chr9	intergenic	-	upstream_gene variant
scaffold_40	754723	754724	T/T	C08:T/TGGCAACT:7-4	+7	chr5	intergenic	-	upstream_gene variant
scaffold_26	864716	864724	ATAGGTAT/ATAGGTAT	C39:ATAGGTAT/A:7-2	-7	chr3	intronic	DAPPUDRAFT_318871	upstream_gene variant
scaffold_14	1376349	1376351	TC/TC	C21:TC/T:13-8	-1	chr9	intergenic	-	downstream_gene variant
scaffold_12	891479	891480	T/T	C02:T/TTC:4-3	+2	chr5	exonic	DAPPUDRAFT_314726	frameshift_variant
scaffold_5	76129	76131	AC/AC	C37:AC/A:6-4	-1	chr12	exonic	DAPPUDRAFT_311851	frameshift_variant

Appendix III

Table ii. SNMs and indels in the 4 MA lines with lower fitness. “Anc” refers to the ancestral genotype of the MA lines at that particular site.

Scaffold	Start	End	Anc	Mutation	Chr	Functional Region	gene (DAPPU DRAFT)	snpEff annotation
187	164767	164768	GG	C49:GA:8,8	6	intronic	302757	upstream gene variant
182	109254	109255	CC	C43:CT:9,5	10	intergenic	-	upstream gene variant
95	167461	167462	TT	C27:TA:14,8	9	intergenic	-	downstream gene variant
76	382265	382266	TT	C23:TG:12,9	8	intergenic	-	intron variant
83	561843	561844	CC	C49:CT:12,6	8	exonic	308666	missense
66	426118	426119	GG	C49:GA:20,12	1	intergenic	-	intergenic
44	373771	373772	GG	C43:GA:14,13	3	intergenic	-	intron variant
34	718378	718379	AA	C27:AT:13,7	8	intergenic	-	upstream gene variant
34	885717	885718	AA	C23:AG:6,5	8	intronic	246016	splice acceptor variant
32	842057	842058	GG	C27:GA:7,5	6	intergenic	-	intergenic
32	727693	727694	GG	C49:GC:8,4	6	intergenic	-	downstream gene variant
28	244972	244973	GG	C49:GA:13,11	6	intergenic	-	upstream gene variant
23	874495	874496	TT	C43:TC:4,4	6	exonic	303760	synonymous
20	720778	720779	AA	C43:AT:5,5	5	exonic	102163	stop gained
10	1879127	1879128	AA	C49:AC:11,7	11	intronic	45893	upstream gene variant
9	639774	639775	GG	C49:GT:10,6	9	intergenic	-	upstream gene variant
9	1060091	1060092	AA	C49:AC:11,7	9	intergenic	-	intergenic region
4	2497275	2497276	AA	C49:AT:9,2	7	exonic	96359	missense
4	2723236	2723237	TT	C23:TA:7,2	7	intronic	305719	intron variant
2	2224692	2224693	GG	C43:GA:9,6	3	exonic	310414	synonymous
8	1323798	13237800	GC/GC	C49:GC/G:8,5	4	intronic	236859	upstream gene variant

Table iii. LOH events in the 4 MA lines with lower fitness.

Scaffold, chr	Min boundaries (min size)	Max boundaries (max size)	Line	Het-Hom sites	Normalized depth ratio, type
scaffold_260, chr 11*	41947 – 42498 (551 bp)	28878 – 94546 (65,668 bp)	C23	5	0.84, conversion
scaffold_193, chr11*	75854 – 88329 (12,475 bp)	72392 – 88398 (16,006 bp)	C23	8	0.51, deletion
scaffold_193, chr11*	150939 – 155342 (4403 bp)	133169 – 155916 (22,747 bp)	C23	6	0.61, deletion
scaffold_166, chr11*	171117 – 246395 (75,278 bp)	168710 – 258489 (89,779 bp)	C23	46	0.48, deletion
scaffold_111, chr 11	165388 – 366867 (201,479 bp)	163734 – 369608 (205,874 bp)	C23	74	0.54, deletion
scaffold_116, chr 11*	36227 – 46334 (10,107 bp)	23711 – 47659 (23,948 bp)	C23	10	0.53, deletion
scaffold_116, chr11*	125971 – 229840 (103,869 bp)	122289 – 231565 (109,276 bp)	C23	17	0.53, deletion
scaffold_116, chr11*	249330 – 283872 (34,542 bp)	247645 – 286124 (38,479 bp)	C23	23	0.51, deletion
scaffold_116, chr11*	299492 – 374626 (75,134 bp)	297907 – 396221 (98,314 bp)	C23	26	0.55, deletion
scaffold_97, chr3	235150 – 238192 (3042 bp)	234698 – 238309 (3611 bp)	C27	27	0.55, deletion
scaffold_85, chr11*	190753 – 214279 (23,526 bp)	172052 – 222021 (49,969 bp)	C23	10	0.49, deletion
scaffold_85, chr11*	313070 – 454673 (141,603 bp)	254698 – 456944 (202,246 bp)	C23	33	0.51, deletion
scaffold_85, chr11*	530091 – 542515 (12,424 bp)	500276 – 551187 (50,911 bp)	C23	12	0.58, deletion
scaffold_58, chr12	468285 – 470265 (1980 bp)	467697 – 470827 (3130 bp)	C23	15	0.58, deletion
scaffold_29, chr10	1166425 – 1167046 (621 bp)	1166261 – 1167058 (797 bp)	C49	10	0.60, deletion
scaffold_3, chr1	2850295 – 2858562 (8267 bp)	2843066 – 2858894 (15,828 bp)	C49	52	0.70, deletion
scaffold_3, chr1	2859541 – 2866761 (7220 bp)	2859178 – 2866916 (7738 bp)	C49	63	0.65, deletion

*chromosome is inferred from scaffold linkage information we obtained from the chromosome 11-wide gene conversion event in the MA line C40