MOLECULAR ECOLOGY
RESOURCES
Molecular Ecology Resources (2015) doi: 10.1111/1755-0998.12434

simMSG: an experimental design tool for high-throughput genotyping of hybrids

MOLLY SCHUMER,*†[1] RONGFENG CUI,†‡§[1] GIL G. ROSENTHAL†‡ and PETER ANDOLFATTO*¶

*Department of Ecology and Evolutionary Biology, Princeton University, Princeton, NJ 08544, USA, †Centro de Investigaciones Científicas de las Huastecas 'Aguazarca', Calnali, Hidalgo, Mexico, ‡Department of Biology, Texas A&M University, TAMU, College Station, TX, USA, §Max Planck Institute for the Biology of Aging, Cologne, Germany, ¶Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, NJ 08544, USA

Abstract

Hybridization between closely related species, whether naturally occurring or laboratory generated, is a useful tool for mapping the genetic basis of the phenotypic traits that distinguish species. The development of next-generation sequencing techniques has greatly improved our ability to assign ancestry to hybrid genomes. One such next-generation sequencing technique, multiplexed shotgun genotyping (or MSG), can be a powerful tool for genotyping hybrids. However, it is difficult a priori to predict the accuracy of MSG in natural hybrids because accuracy depends on ancestry tract length and number of ancestry informative markers. Here, we present a simulator, 'simMSG', that will allow researchers to design MSG experiments and show that in many cases MSG can accurately assign ancestry to hundreds of thousands of sites in the genomes of natural hybrids. The simMSG tool can be used to design experiments for diverse applications including QTL mapping, genotyping introgressed lines or admixture mapping.

Keywords: hybridization, multiplexed shotgun genotyping, natural hybrids, reduced representation genotyping

Received 22 March 2015; revision received 19 May 2015; accepted 22 May 2015

Introduction

Hybridization between closely related species or populations is a common evolutionary process in both plants and animals (Mallet 2005). Our understanding of the scope of hybridization has been greatly advanced by next-generation sequencing techniques that facilitate the sequencing of large numbers of loci, allowing researchers to characterize gene flow throughout the genome (Twyford & Ennos 2012). Reduced representation techniques such as restriction-site-associated DNA markers (RAD-tags) allow researchers to genotype large numbers of individuals for population genetic studies or studies of hybridization (Davey *et al.* 2011; Peterson *et al.* 2012).

Among these is a low-cost technique for genomewide genotyping of laboratory hybrids called multiplexed shotgun genotyping or MSG (Andolfatto *et al.* 2011). MSG is a powerful tool for accurately genotyping large numbers of hybrid individuals at hundreds of thousands of sites throughout the genome. Although most studies so far have applied MSG to laboratory generated hybrids (Andolfatto *et al.* 2011; Cande *et al.* 2012; Slotte *et al.* 2012), it can also be used to efficiently genotype naturally occurring hybrids (Schumer *et al.* 2014).

Multiplexed shotgun genotyping is distinct from other genotyping approaches because it leverages information at many ancestry informative markers throughout the genome to infer local ancestry. While RAD is an approach for assigning genotypes at single nucleotide polymorphisms (SNPs), MSG uses SNPs to assign ancestry to particular genomic regions using a hidden Markov model (HMM). Because of this analytical approach, the same sites need not be sampled in multiple individuals to genotype the same region, allowing for lower-depth sequencing and higher levels of multiplexing. A further advantage is that ancestry calls are given as posterior probabilities, accommodating uncertainty in genotype due to variation in coverage among other factors. In addition, the current MSG protocol (Appendix S1, Supporting information) can be completed in just 2 days of bench work and requires only 10–25 ng of input DNA (compared to ~1 μg for RAD and ezRAD; Etter *et al.*

Correspondence: Molly Schumer, Fax: (609) 258-1712; E-mail: schumer@princeton.edu, and Rongfeng Cui, Fax: +49 (0) 221 37970-800; E-mail: rcui@age.mpg.de

[1]Denotes equal contribution.

2011; Toonen *et al.* 2013; ~200 ng for GBS Elshire *et al.* 2011; ~100 ng for ddRAD; Peterson *et al.* 2012). MSG also compares favourably to RAD approaches in the number of sites at which ancestry can be assigned throughout the genome (Table S1, Supporting information), improving the resolution of mapping studies. For example, approximately 30X more sites are surveyed between two fish species using MSG (*Xipophorus birchmanni – X. malinche*) than between another two fish species (*Oncorhynchus mykiss – O. clarkii lewisi*) using a RAD-tag approach, despite greater divergence time in the RAD-surveyed species (see Table S1, Supporting information, Hand *et al.* 2015; Schumer *et al.* 2014; Wilson & Turner 2009). Because the MSG pipeline accommodates genomic data in addition to reduced representation data, marker density in MSG can be further improved by collecting low-coverage genomic data.

Multiplexed shotgun genotyping has not been commonly used to genotype natural hybrids because the coverage and marker density required to determine the ancestry of short ancestry tracts with this method is unknown. Sensitivity to short ancestry tracts will ultimately be limited by coverage and ancestry informative marker (AIM) density, both of which will vary depending on the choice of biological system. In this manuscript, we describe a new simulation program called simMSG that will allow researchers to determine how accurately MSG will perform on individuals from hybrid populations with user-specified characteristics, such as number of generations since initial hybridization and admixture proportions. We apply this tool to hybrids simulated under a range of parameters and show that the expected accuracy of MSG is high when hybridization is recent (<50 generations) or divergence between species is high (~0.5% of sites AIMs). Although one of the main advantages of simMSG is its ability to predict MSG accuracy in natural hybrid populations, it also has useful applications to laboratory generated hybrids and introgressed lines (Andolfatto *et al.* 2011). For example, simMSG can be used to determine how much coverage is required for accurate genotyping of a particular laboratory hybrid cross or whether MSG is sensitive enough to detect an introgressed region of a particular size. This simulator is a useful tool for designing MSG experiments that will establish feasibility and lower experimental costs for researchers working with both laboratory and natural hybrids.

Materials and methods

Description of computational pipeline

The following sections describe the user input and steps followed in the simMSG computational pipeline (summarized in Fig. 1):

Input parameters. Users set parameters for the simulation in the simMSG configuration file before running the program from the command line (see Appendix S2, Fig. 1A). Most parameters have suggested default values (Table

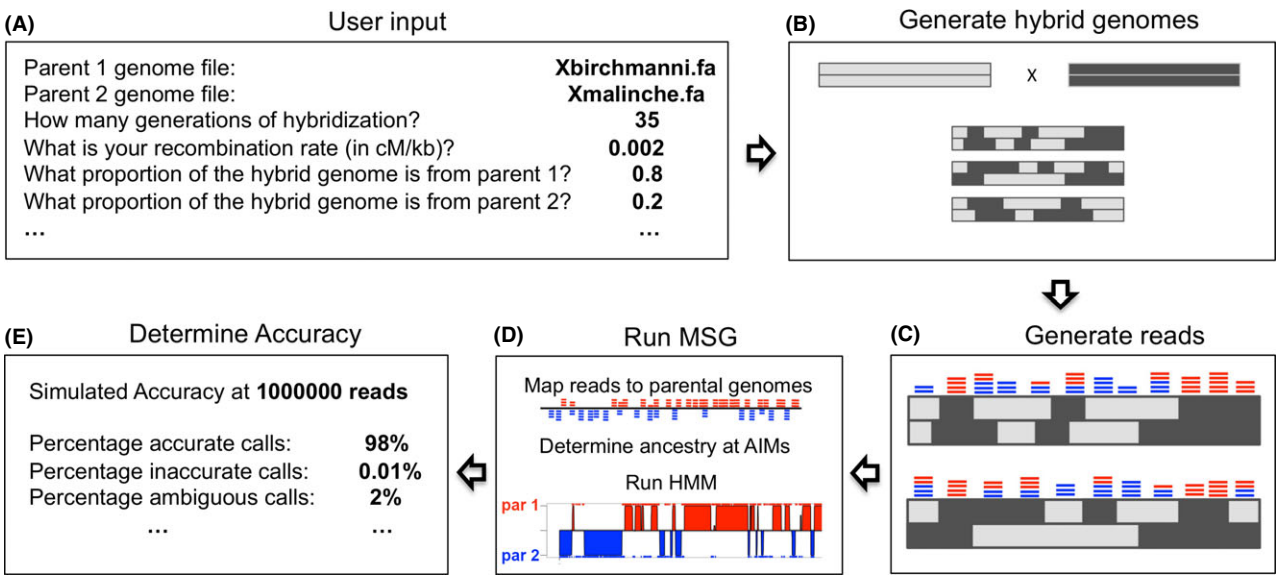

Fig. 1 Schematic of simMSG pipeline. (A) Users input hybrid-population-specific information and other parameters; see Table S2 (Supporting information) for complete list. (B) Hybrid chromosomes are generated by drawing blocks from the parental ancestry tract length distributions (Fig. S1, Supporting information) and polymorphism is introduced (Fig. S2, Supporting information). (C) Reads are generated at restriction sites or randomly throughout the chromosome. (D) The MSG pipeline is run and (E) accuracy is summarized.

S2, Supporting information) but several require user input. These parameters are the reference genomes of the two parental species (see next paragraph), the sex-averaged recombination rate, the number of generations of recombination, estimates of parental polymorphism and the proportion of the hybrid genome derived from each parent (Table 1). Although many of these parameters may be unknown for nonmodel species, users can simulate several scenarios to investigate whether MSG is likely to be successful in their system under a broad range of parameters (Table 1).

The two parental genomes used by simMSG must have the same coordinates and scaffold names. If parental genomes with the same coordinates are not available for the species of interest, there are a number of possibilities for generating these reference sequences including using map-to-reference approaches to generate a new reference sequence based on the coordinates for one parental species (e.g. see Schumer *et al.* 2013). Alternately, coalescent simulators (e.g. Chen *et al.* 2009) can be used to generate reference sequences with parameters appropriate for the species of interest, although simulated sequences will lack many of the complexities of real sequences.

Generate hybrid genomes – ancestry tract lengths. To predict MSG's accuracy on a particular hybrid population, the program must first generate hybrid chromosomes made up of ancestry tracts from both parents (Fig. 1B). Based on information input into simMSG about the recombination rate, the proportion of the genome derived from each parent, and the number of generations of recombination (Table 1), the program calculates the expected tract length distribution for each parent (Gravel 2012). Using these distributions, the program randomly samples ancestry tract lengths from these parental distributions to generate each hybrid chromosome (requiring a change in ancestry from the previously drawn tract; Fig. S1, Supporting information, Gravel 2012). This process continues until the hybrid chromosome is the required length; if an added tract causes the simulated chromosome to exceed the chromosome length, the tract is truncated at the length of the chromosome (Fig. S1, Supporting information). It should be emphasized that simMSG models neutral admixture in large populations. Selection and demographic processes (e.g. migration and changes in population size) may change the ancestry tract length distribution relative to that modelled by simMSG. Among these, migration from parental populations will generally lead to longer tracts and better performance than predicted. A population bottleneck may have a more complicated effect by increasing the mean tract length but also increasing the variance, implying that performance will be better than predicted except for the smaller proportion of the genome represented by small tract lengths.

Generating hybrid chromosomes as described in the previous paragraph is the most appropriate method for simulating late-generation hybrids (>10 generations, see Gravel 2012). For special types of hybrid crosses, such as

Table 1 Parameters that must be provided by the user to simulate a specific biological system. More detail on these parameters and default parameters can be found in Table S2 (Supporting information). Most required parameters can be tested at a range of values if there is uncertainty

Parameter	Description	Possible to test a range of values?
Parental genomes	Users must supply the parental genomes in fasta format. These sequences will be used to simulate hybrid genome sequences.	No
Sex-averaged recombination rate	Users supply a species-specific recombination rate that allows the program to estimate ancestry tract length distributions.	Yes
Number of generations of recombination	As the number of generations of recombination between parental genomes increases, the average length of ancestry tracts for each parent becomes shorter. If users have estimates of the number of generations of hybridization for their population of interest, this should be input, or alternately, users can simulate a range of generations.	Yes
Proportion parent 1 and parent 2	The proportion of the hybrid genome derived from each parental species is important in determining the ancestry tract length distribution. If users have estimates of the proportion of the hybrid genome derived from each parent, this information can be input, or a range of values can be simulated.	Yes
Polymorphism estimates	simMSG simulates private and shared polymorphism between the parental species when generating hybrid genome sequences (see Appendix S3, Supporting information). To estimate these parameters, the program uses the population mutation rate (θ) for each species. This parameter can be estimated from sequence data based on the proportion of polymorphic sites in a sample of sequences (θ_s) or the proportion of heterozygous sites in an a diploid individual (θ_π).	Yes

F_2 crosses or introgressed lines, users can specify a prior distribution of haplotype lengths for each parent instead of using the simMSG -generated distribution (see details in Table S2, Supporting information). The prior distribution files can be set to correspond to the expected number of breakpoints per chromosome. For example, to simulate an F_2 hybrid, tract lengths can be set so that there will be on average one to two recombination breakpoints per chromosome. Finally, users can indicate in the configuration file if they wish to simulate backcross hybrids (Table S2, Supporting information), which lack homozygous regions for one parent. Figure 2 shows MSG ancestry plots for hybrid individuals simulated under each of these scenarios.

Generate hybrid genomes – adding polymorphism. Individuals sampled in an actual experiment will have polymorphisms not present in the reference sequences, and some sites that are ancestry informative between the two reference sequences will not be fixed in the sampled individuals (Fig. 1B). To realistically simulate this, we calculate the expected number of ancestry informative sites, private polymorphisms and shared polymorphisms based on input parameters (Table 1) following Wakeley and Hey (see Appendix S3, Supporting information, Wakeley & Hey 1997). To integrate this information with sites in the reference genomes, ancestry informative sites are randomly selected from the set of sites that differ between the two reference genomes, while private and shared polymorphic sites are assigned to the remaining sites that differ between the two reference genomes and to randomly selected sites (see complete explanation in Appendix S3 and Fig. S2, Supporting information).

When generating hybrid chromosomes (as above), polymorphisms are added to each ancestry tract. The frequency of each polymorphism is drawn randomly from the neutral site frequency distribution. To determine which allele is the major allele, we treat this frequency as the probability that the reference allele is the minor allele. As an ancestry tract is drawn, the genotype at each polymorphic site is drawn from a multinomial distribution based on the frequency of the major and minor alleles. Each shared polymorphism is conservatively assumed to be at frequency = 0.5 in both parental species.

Optionally, users can tell simMSG to mask all shared polymorphisms in the reference genomes by specifying this option in the configuration file (Table S2, Supporting information). We use the program seqtk (https://github.com/lh3/seqtk) to mask polymorphisms within the simMSG pipeline. This masking step allows researchers to predict how much the performance of MSG would improve if shared polymorphisms present in the parental populations were masked in the reference genome, and is recommended for high polymorphism or recently diverged species.

Generate reads. Following the generation of hybrid chromosomes described above, simMSG generates reads from these chromosomes (Fig. 1C). Reads are generated at MseI restriction sites to simulate MSG library preparation (automatically recognized from the input reference genome; default recognition sequence TTAA can be changed in the configuration file) or randomly throughout the chromosome to simulate low-coverage genome sequencing (see Table S2, Supporting information). Reads are only generated from fragments 250–500 base pairs to simulate size selection in MSG library preparation (size range can be changed, see Table S2, Supporting information). To simulate variation in read length introduced by quality trimming, the length of each read is drawn from a random uniform distribution ranging from

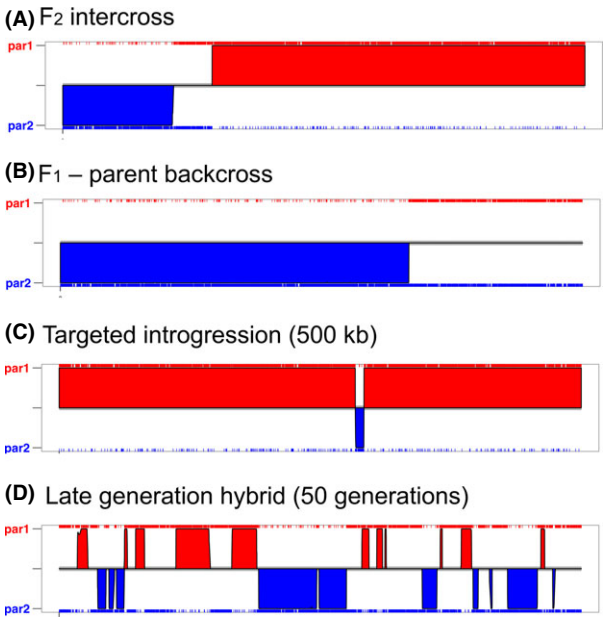

Fig. 2 Ancestry plots of different types of hybrids generated using simMSG. Results show an F2 intercross (A), F1-parent backcross (B), introgressed line (C) and late generation hybrid (D). Regions of solid red shading indicate regions homozygous for parent 1, regions of solid blue shading indicate regions homozygous for parent 2, and regions that are unshaded indicate regions that are heterozygous. The height of each shaded region on the y-axis represents the posterior probability of the ancestry call (see Andolfatto *et al.* 2011). Tick marks for the opposite colour in a solid red or blue region indicate individual sites that match the opposite parent species caused by simulated error (see Methods). Simulation parameters: 1000 reads/Mb, parental genomes 0.5% diverged. In D, 50 generations of 15 recombination at 50–50 admixture proportions were simulated.

75% to 100% of the user-specified read length. Base quality is arbitrarily set to 25 for all sites. The sequencing error rate is set by the user (Table S2, Supporting information); if a read contains sequencing errors, their position from the end of the read is determined by drawing from an exponential distribution (with a mean of read length/10).

The MSG error parameters 'deltapar1' and 'deltapar2' describe the probability that a SNP for parent 1 will be detected in a region homozygous for parent 2 (and vice versa). These parameters can be empirically estimated from real data during the MSG run by the proportion of ancestry informative sites that match the wrong parent in a region inferred to be homozygous (Andolfatto *et al.* 2011). Note that these error parameters are distinct from the errors in ancestry assignment that simMSG reports. In general, when performing MSG genotyping, we see empirical MSG error rates of around 3–5%. In actual MSG data, there are many sources of this error, including mismapped reads, genome incompleteness and contamination. Because these sources do not exist in the simulation pipeline, our initial simulations yielded MSG error estimates of <1%. To incorporate additional error and thus make simulations more realistic, we added a user-specified MSG error parameter to the pipeline (see Table S2, Supporting information), which adds contamination by simulating a proportion of reads from the parental genomes instead of the hybrid genome. This in silico contamination from the parental genomes generates error profiles matching intended MSG error rates (Fig. S3, Supporting information) although it is important to keep in mind that this simulated error is only partially comparable to real MSG error.

Run MSG pipeline (Fig. 1D) and determine accuracy given the simulated parameters (Fig. 1E). After reads are generated, an abridged version of the MSG pipeline is run on the simulated hybrid genomes to generate MSG ancestry calls (Fig. 1D). Following the completion of the MSG pipeline, genotype calls output by MSG are directly compared to true ancestry for each simulated individual and a summary results file is generated (Fig. 1E). Because MSG outputs genotype calls in the form of posterior probabilities, to summarize accuracy we treat posterior probabilities of >0.95 for the true genotype as an accurate call, posterior probabilities of >0.95 for the incorrect genotype as inaccurate calls and posterior probabilities of <0.95 as ambiguous calls. The results file summarizes accuracy, including information on overall accuracy, accuracy of genotype calls for each parent, marker density and the frequency of errors of different types. An example results file can be found in Table S3 (Supporting information). The pipeline also outputs two plots to visualize the simulation results, showing individual accuracy and hybrid index distributions ('accuracy_distribution.pdf' and 'ancestry_distribution.pdf', Fig. S4, Supporting information).

Example simulations

To explore the conditions under which MSG can accurately genotype hybrids, we use simMSG to investigate the effects of divergence between the parental species, number of generations of hybridization, error parameters and number of reads on MSG accuracy (Appendix S3, Supporting information). We performed these simulations using both a representative invertebrate genome (*Drosophila simulans* – 4 chromosomes, 124.5 Mb, 2.5 cM/Mb; Fiston-Lavier *et al.* 2010; Hu *et al.* 2013) and a vertebrate genome (*Xiphophorus malinche* – 24 chromosomes, 730 Mb, 1.8 cM/Mb; Cui *et al.* 2013; Schartl *et al.* 2013). See Appendix S3 (Supporting information) for a full description of these simulations. In addition, we compare results from simMSG to real data for two hybrid crosses (*D. simulans x sechellia* F_1-backcross hybrids and *X. malinche x birchmanni* natural hybrids, Appendix S3, Supporting information) and investigate MSG's performance on short contigs (Appendix S3, Supporting information).

Results

Example simulations

Simulations using simMSG demonstrate that the accuracy of MSG is primarily dependent on divergence between the two parental species and the length of ancestry tracts (Fig. 3). If species are sufficiently diverged (~0.5% of sites AIMs – equivalent to 1% divergence in our simulations), MSG is accurate for a broad range of hybrid populations, even populations that have short ancestry tracts (e.g. <500 kb, Fig. S5, Supporting information). On the other hand, even hybrids between species that have very low genetic differentiation (<0.1% of sites AIMs) can be accurately genotyped if ancestry tracts are long (few generations of hybridization, Fig. S5). Our simulations also show that increasing coverage and tuning parameters can further improve accuracy, even for late-generation hybrids between species with low divergence (Fig. 3, Appendis S3C–D, Supporting information).

In addition, our simulations show that MSG accuracy decreases as within-species polymorphism levels increase, but that this effect can be mitigated by masking polymorphism in the parental genomes (Fig. S6, Supporting information). High parental polymorphism levels decrease accuracy because fewer sites will be ancestry informative in the sampled individuals and

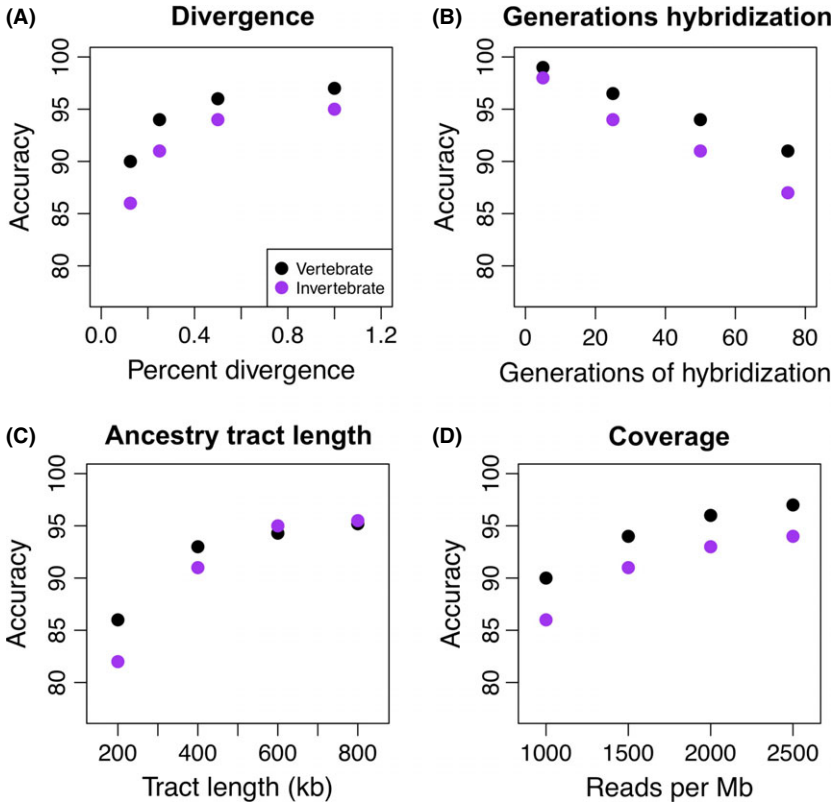

Fig. 3 Simulations using simMSG show that genotyping accuracy is most dependent on divergence between the parental genomes (A) and the number of generations of hybridization (B), which is directly related to ancestry tract length (C). Even when accuracy is low, as for simulations of 0.125% divergent genomes, increasing coverage can dramatically increase accuracy (D). To examine the effects of these parameters in two different contexts, we performed simulations based on a representative vertebrate and invertebrate genome (*X. malinche* and *D. simulans*, see Appendix S3, Supporting information) to capture variation in genome size (*X. malinche*: 730 Mb, *D. simulans*: 124.5 Mb), number of chromosomes (*X. malinche*: 24, *D. simulans*: 4) and recombination rate (*X. malinche*: 1.8 cM/Mb, *D. simulans*: 2.5 cM/Mb). Simulation parameters: (A) 50–50 admixture proportions, 25 generations of hybridization, 1000 reads/Mb; (B, C) 50–50 admixture proportions, 0.5% divergence, 1000 reads/Mb; (D) 50–50 admixture proportions, 25 generations of hybridization, 0.125% divergence.

because undetected shared polymorphisms between parental species contribute to errors in MSG genotyping. However, masking parental polymorphisms (mask_poly option, see Table S2, Supporting information) dramatically improves MSG's accuracy in high polymorphism scenarios. This can be performed experimentally by collecting MSG data for parental individuals (see e.g. Schumer *et al.* 2014).

We also used simMSG to generate data corresponding to hybrid crosses for which we had real MSG data (*Xiphophorus* and *Drosophila* hybrids, see Appendix S3, Supporting information). This allowed us to compare predicted accuracy to results from actual hybrids by subsampling the real data to generate coverage variation. Results from subsampling are consistent with predictions of simMSG (Appendix S3, Table S4, Supporting information), suggesting that simMSG is realistically simulating changes in accuracy as a result of variation in coverage.

Finally, we show using simMSG that MSG can be successfully applied to more fragmented references, as might be expected in draft genomes. Although accuracy will be dependent divergence levels, in the scenarios we simulate simMSG predicts >95% accuracy on scaffolds >250 kb (Fig. S7, Supporting information), demonstrating that MSG can be used even when high-quality reference sequences are unavailable.

Types of errors. In all simulations, short ancestry tracts had the highest error rate (see e.g. Fig. 3) and error in ancestry assignment calls was not random. Eighteen per cent of errors in our simulations were heterozygous regions incorrectly assigned as homozygous, while 82% of errors were homozygous regions incorrectly called as heterozygous (see for e.g. Table S3, Supporting information). Few errors involved assignment of a region homozygous for one parent as homo-

zygous for the other parent (<0.2% of errors in our simulations).

Benchmark. All off-cluster performance measurements were made on a computer running a Xeon E5420 CPU at 2.50 GHz, and 32 GB of RAM. All cluster performance measures were performed on Princeton's della cluster (http://www.princeton.edu/researchcomputing) using one node and one processor per node for all steps but msgRun2 for which we used one node and four processors per node (2.67 GHz Westmere nodes with 4 GB of RAM). Runtime increases linearly with the number of individuals and reads simulated (Fig. S8, Supporting information). The maximum per individual disc space footprint during the simulation run was 0.4 GB. The default cluster settings for simMSG are written for a SLURM system and were tested on Princeton's della cluster. However, we include templates that users can modify to adapt commands to other resource management systems (see Appendix S2).

Discussion

Recently, there has been a surge of interest in inexpensive whole-genome genotyping for a number of applications. Multiplexed shotgun genotyping (MSG) is a powerful tool for cost-effective genome-wide genotyping of hybrids. We present a new simulator tool, simMSG, that will allow researchers to predict MSG's accuracy and design MSG experiments for natural hybrid populations, laboratory generated hybrids and introgressed lines. Specifically, researchers can determine whether (i) MSG will be an effective tool for genotyping the hybrid population of interest and (ii) how much sequencing effort will be required to accurately genotype individuals (see e.g. Fig. 3). This tool will both establish feasibility and lower costs of planned MSG experiments.

Advantages and disadvantages of the MSG genotyping approach

The MSG ancestry inference approach affords researchers working on hybrid populations a lower-cost, higher-resolution alternative to SNP-genotyping-based techniques such as RAD-tag. Because MSG relies on a hidden Markov model for local ancestry inference from ancestry informative markers, the same region can be genotyped in a set of individuals even if the same SNPs are not sampled in these individuals. MSG typically outperforms SNP-genotyping approaches in the number of ancestry informative markers that can be genotyped in hybrids (Table S1, Supporting information). Another advantage of MSG is that data generated from both reduced representation and randomly sheared libraries (e.g. generated by sonication – Illumina technology, Tn5 – Nextera, dsDNA shearase – Zymo or NEB fragmentase) can be used in the analysis pipeline. In addition, the protocol for reduced representation library preparation is simpler than other reduced representation protocols and can be completed in approximately 2 days of bench time (see Appendix S1, Supporting information). However, there are also several drawbacks of the MSG approach. Because analysis is limited to ancestry informative markers, MSG results cannot be used in population genetic analyses of sites that vary within species, although issues have also been raised with using RAD data for population genetics (Arnold *et al.* 2013).

In addition, MSG requires parental reference genomes. This is because the HMM integrates SNP information over multiple ancestry informative markers, which is not possible without a reference genome to define the spatial relationship among markers. This means that although MSG is a fast and cost-effective method to genotype hybrids, this is only the case when a reference sequence is available for at least one parental species, given the cost and time involved in generating a genome assembly.

Finally, MSG performs better on reference genomes with longer average scaffold lengths, but can also be used to improve fragmented assemblies and correct errors in the reference sequence. MSG performs well on relatively short contigs (~500 kb), as might be present in draft genomes, but for very short contigs (e.g. <200 kb), the HMM is less accurate at inferring ancestry (Fig. S7, Supporting information). This performance depends in part on the level of divergence between species. However, one application of MSG it to identify associated scaffolds through linkage disequilibrium (LD), which can be used to improve the genome assembly (Andolfatto *et al.* 2011). In addition, MSG can be used to identify errors in genome assemblies such as inverted regions through particular patterns of LD in the data (Andolfatto *et al.* 2011; Cande *et al.* 2012; Slotte *et al.* 2012). If a site is not in LD with neighbouring sites but is in strong LD with sites in another region, this is a signal of either an inversion or a translocation in the assembly.

Using simMSG to predict MSG's accuracy on hybrid populations

When can MSG be used to genotype hybrids? Our simulations demonstrate that MSG performance is most sensitive to variation in the number of generations of recombination (because of its effect on ancestry tract length, Gravel 2012) and the density of ancestry informative markers. MSG consistently performs well on hybrids between parental species with sufficient divergence (~0.5% of sites AIMs) and also performs well on early-generation hybrids between species with few ancestry

informative sites (even <0.1% in our simulations, Fig. S5, Supporting information).

Although researchers will typically have some information on divergence rates between species, predicting how many generations of recombination have occurred may be difficult unless the hybrid population formed recently (e.g. Rosenthal *et al.* 2003; Nolte *et al.* 2009). In addition, even if a hybrid population has existed for many generations, selection against hybrids may result in an excess of recent hybrids (e.g. Rieseberg *et al.* 1999; Alexandrino *et al.* 2005). Analysis of ancestry tract length distributions (Gravel 2012) or the decay in linkage disequilibrium (Hellenthal *et al.* 2014) in a subset of individuals can be used to estimate the number of generations as hybridization. Alternately, users can simulate a range of parameter values if there is uncertainty.

Using simMSG to design MSG experiments

A primary goal of simMSG was to help researchers decide how much and what kind of data to collect for MSG studies. Although there are some hybridization scenarios in which MSG will never be highly accurate, in most cases increasing the number of reads (Fig. 3) can significantly increase accuracy. Increasing the number of reads increases accuracy for two reasons. The major benefit is in sampling a greater number of ancestry informative sites, although increased depth at individual sites marginally increases accuracy by increasing confidence of ancestry calls at those sites. By running multiple simulations varying coverage, researchers can determine the minimum coverage required for accurate genotyping. Similarly, when levels of parental polymorphism are high, masking polymorphism significantly increases accuracy (Fig. S6, Supporting information). simMSG can help users decide whether the increase in accuracy from masking parental polymorphism is worth the cost of collecting MSG data from a panel of parental individuals (e.g. Schumer *et al.* 2014).

In addition, resolution can be improved by collecting low-coverage whole-genome sequence data instead of MseI-associated data. The MSG analysis pipeline is compatible with data generated both by restriction enzyme digest and by random shearing of genomic DNA. Based on our simulations (Appendix S3), a random shearing approach dramatically increases the number of markers sampled (Fig. S9, Supporting information) and can provide increased resolution of recombination breakpoints (Rowan *et al.* 2015). However, preparing libraries using a random shearing approach is currently more expensive. simMSG results will help researchers assess whether the increased resolution from this approach is worth the increased cost of library preparation.

Differences between simulated and real data

Although simMSG will give researchers an idea of how MSG will perform in their system, it is important to point out several ways in which the simulated data differs from real data. Reads are simulated randomly from the input genome sequence, but in reality, there may be overdispersion in coverage. In addition, read mis-mapping, PCR errors, and errors or incompleteness in the reference sequence may generate high errors in genotyping in particular regions as opposed to the uniform error rate we simulate.

In actual MSG experiments, users can tune different parameters in the HMM to improve accuracy. For example, priors for each genotype (parent1, heterozygous and parent2) can be updated using the genotype calls from an initial run. Similarly, the MSG error rate parameters can be informed by empirical estimates. Another parameter, theta, can be modified in analysis of actual MSG data to account for nonindependence between reads caused by PCR duplication. While the effects of tuning some parameters can be investigated using simMSG (see Appendix S3, Supporting information), others, such as the MSG error parameters and theta, cannot be tuned without actual sequence data. However, for several parameters users can simulate a range of values to explore how accuracy is predicted to change as a result of variation in these parameters (Appendix S3, Supporting information).

Conclusions

The ability to use MSG to genotype hybrids has applications ranging from analysing patterns of introgression in natural populations to QTL or admixture mapping. The simMSG software will make it easy for users to decide whether MSG can be used to accurately genotype hybrids in their system and determine how much sequencing effort is required.

Acknowledgements

This work was supported in part by NSF GRFP (DGE0646086) and DDIG to MS (DEB-1405232) and NSF IOS-0923825 to GGR. We thank Maria Gutin for help collecting references and Ying Zhen, Bridgett vonHoldt and three anonymous reviewers for comments on earlier versions of this manuscript.

References

Alexandrino J, Baird SJE, Lawson L *et al.* (2005) Strong selection against hybrids at a hybrid zone in the *Ensatina* ring species complex and its evolutionary implications. *Evolution*, **59**, 1334–1347.

Andolfatto P, Davison D, Erezyilmaz D *et al*. (2011) Multiplexed shotgun genotyping for rapid and efficient genetic mapping. *Genome Research*, **21**, 610–617.

Arnold B, Corbett-Detig RB, Hartl D, Bomblies K (2013) RADseq underestimates diversity and introduces genealogical biases due to nonrandom haplotype sampling. *Molecular Ecology*, **22**, 3179–3190.

Cande J, Andolfatto P, Prud'homme B, Stern DL, Gompel N (2012) Evolution of multiple additive loci caused divergence between *Drosophila yakuba* and *D. santomea* in wing rowing during male courtship. *PLoS ONE*, **7**, e43888.

Chen GK, Marjoram P, Wall JD (2009) Fast and flexible simulation of DNA sequence data. *Genome Research*, **19**, 136–142.

Cui R, Schumer M, Kruesi K *et al*. (2013) Phylogenomics reveals extensive reticulate evolution in *Xiphophorus* fishes. *Evolution*, **67**, 2166–2179.

Davey JW, Hohenlohe PA, Etter PD, *et al*. (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics*, **12**, 499–510.

Elshire RJ, Glaubitz JC, Sun Q *et al*. (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS ONE*, **6**, e19379.

Etter PD, Bassham S, Hohenlohe PA, Johnson EA, Cresko WA (2011) SNP discovery and genotyping for evolutionary genetics using RAD sequencing. *Molecular Methods for Evolutionary Genetics*, **772**, 157–178.

Fiston-Lavier AS, Singh N, Lipatov M, *et al*. (2010) Drosophila melanogaster recombination rate calculator. *Gene*, **463**, 18–20.

Gravel S (2012) Population genetics models of local ancestry. *Genetics*, **191**, 607–619.

Hand BK, Hether TD, Kovach RP *et al*. (2015) Genomics and introgression: discovery and mapping of thousands of species-diagnostic SNPs using RAD sequencing. *Current Zoology*, **61**, 146–154.

Hellenthal G, Busby GBJ, Band G *et al*. (2014) A genetic atlas of human admixture history. *Science*, **343**, 747–751.

Hu TT, Eisen MB, Thornton KR, Andolfatto P (2013) A second-generation assembly of the *Drosophila simulans* genome provides new insights into patterns of lineage-specific divergence. *Genome Research*, **23**, 89–98.

Mallet J (2005) Hybridization as an invasion of the genome. *Trends in Ecology & Evolution*, **20**, 229–237.

Nolte AW, Gompert Z, Buerkle CA (2009) Variable patterns of introgression in two sculpin hybrid zones suggest that genomic isolation differs among populations. *Molecular Ecology*, **18**, 2615–2627.

Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE (2012) Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PLoS ONE*, **7**, e37135.

Rieseberg LH, Whitton J, Gardner K (1999) Hybrid zones and the genetic architecture of a barrier to gene flow between two sunflower species. *Genetics*, **152**, 713–727.

Rosenthal GG, de la Rosa Reyna XF, Kazianis S *et al*. (2003) Dissolution of sexual signal complexes in a hybrid zone between the swordtails *Xiphophorus birchmanni* and *Xiphophorus malinche* (Poeciliidae). *Copeia*, **2003**, 299–307.

Rowan BA, Patel V, Weigel D, Schneeberger K (2015) Rapid and inexpensive whole-genome genotyping-by-sequencing for crossover localization and fine-scale genetic mapping. *G3 (Bethesda)*, **5**, 385–398.

Schartl M, Walter RB, Shen Y *et al*. (2013) The genome of the platyfish, *Xiphophorus maculatus*, provides insights into evolutionary adaptation and several complex traits. *Nature Genetics*, **45**, 567–572.

Schumer M, Cui R, Boussau B *et al*. (2013) An evaluation of the hybrid speciation hypothesis for *Xiphophorus clemenciae* based on whole genome sequences. *Evolution*, **67**, 1155–1168.

Schumer M, Cui R, Powell D *et al*. (2014) High-resolution mapping reveals hundreds of genetic incompatibilities in hybridizing fish species. *eLife*, **3**, e02535.

Slotte T, Hazzouri KM, Stern D, Andolfatto P, Wright SI (2012) Genetic architecture and adaptive significance of the selfing syndrome in *Capsella*. *Evolution*, **66**, 1360–1374.

Toonen RJ, Puritz JB, Forsman ZH *et al*. (2013) ezRAD: a simplified method for genomic genotyping in non-model organisms. *PeerJ*, **1**, e203.

Twyford AD, Ennos RA (2012) Next-generation hybridization and introgression. *Heredity*, **108**, 179–189.

Wakeley J, Hey J (1997) Estimating ancestral population parameters. *Genetics*, **145**, 847–855.

Wilson WD, Turner TF (2009) Phylogenetic analysis of the Pacific cutthroat trout (*Oncorhynchus clarki* ssp.: Salmonidae) based on partial mtDNA ND4 sequences: a closer look at the highly fragmented inland species. *Molecular Phylogenetics and Evolution*, **52**, 406–415.

M.S., R.C and P.A. wrote the pipeline, M.S. and R.C. tested the pipeline, P.A. and G.G.R. supervised the project, all authors wrote the manuscript.

Data accessibility

The simMSG program can be downloaded at https://github.com/melop/simMSG, along with an example data set.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Table S1 Number of ancestry informative markers sampled in studies using MSG versus studies using other methods to investigate hybridization.

Table S2 Details of input parameters specified in the configuration file for simMSG.

Table S3 Example output statistics from a simMSG run with the following parameters: 50-50 admixture proportions, 50 generations of hybridization, 0.5% divergent genomes, 25% of divergent sites polymorphic.

Table S4 Correspondence between real data at different coverage levels to real data at full coverage in *Xiphophorus* and *Drosophila* hybrid individuals (see Appendix S3).

Table S5 Using naïve priors (par1 = 0.33,par1, par2 = 0.33, par2 = 0.33) in the simMSG pipeline demonstrates that accurate priors can be estimated from the MSG data, which is a useful feature for users without prior knowledge of the ancestry proportions of their hybrid population.

Appendix S1 Supporting MSG protocol.

Appendix S2 simMSG user manual.

Appendix S3 Supporting text and simulations.

Fig. S1 Schematic of the process used to generate hybrid haplotypes in simMSG.

Fig. S2 Schematic showing procedure for adding expected number of polymorphic sites to hybrid genomes.

Fig. S3 MSG error rate distributions generated by simMSG.

Fig. S4 Example output files produced by simMSG.

Fig. S5 Effect of the number of generations of hybridization on MSG accuracy.

Fig. S6 MSG accuracy decreases with increasing levels of within-species polymorphism.

Fig. S7 MSG accuracy as a function of contig length.

Fig. S8 Performance of simMSG program

Fig. S9 MSG accuracy and number of markers with low coverage genome sequence data.

Fig. S10 Simulations varying the MSG error parameter demonstrate that an increase in the MSG error rate can decrease overall accuracy.

Fig. S11 Simulations varying the recombination rate scaling factor (rfac) show that increasing rfac can decrease accuracy.