Mutation mapping and identification by whole genome sequencing

Ignat Leshchiner, Kristen Alexa, Peter Kelsey, et al.

Genome Res. published online May 3, 2012
Access the most recent version at doi:10.1101/gr.135541.111

Mutation mapping and identification by whole genome sequencing

Ignaty Leshchiner [1], Kristen Alexa [1], Peter Kelsey [1], Ivan Adzhubei [1], Christina A. Austin [2], Jeffrey D. Cooney [3], Heidi Anderson [3], Matthew J. King [3], Rolf Stottmann [1], Seungshin Ha [1], Iain A. Drummond [2], Barry H. Paw [3], Trista E. North [4,5], David R. Beier [1], Wolfram Goessling [1,5,6,7 #], Shamil Sunyaev [1,6 #]

[1]Genetics Division, Brigham and Women's Hospital, Harvard Medical School, NRB4, 77 Ave Louis Pasteur, Boston, MA 02115, USA; [2]Department of Genetics, Massachusetts General Hospital, Boston, MA 02114, USA; [3]Hematology Division, Brigham and Women's Hospital, Harvard Medical School, Karp Family Research Building, 1 Blackfan Circle, Boston, MA 02114, USA; [4]Department of Pathology, Beth Israel Deaconess Medical Center, Harvard Medical School, CLS636, 3 Blackfan Circle, Boston, MA 02115, USA; [5]Harvard Stem Cell Institute, Cambridge, MA, USA; [6]Harvard/MIT Division of Health Sciences and Technology, 77 Massachusetts Avenue, Cambridge MA 02139, [7]Gastrointestinal Cancer Center, Dana-Farber Cancer Institute, 450 Brookline Ave, Boston, MA 02115, USA.

To whom correspondence should be addressed:

Shamil Sunyaev (ssunyaev@rics.bwh.harvard.edu)

or

Wolfram Goessling (wgoessling@partners.org)

Genetics Division

Brigham and Women’s Hospital

NRB 4

77 Ave Louis Pasteur

Boston, MA 02115

ABSTRACT

Genetic mapping of mutations in model systems has facilitated the identification of genes contributing to fundamental biological processes, including human diseases. However, this approach has historically required the prior characterization of informative markers. Here, we report a fast and cost-effective method for genetic mapping using Next Generation Sequencing that combines single nucleotide polymorphism discovery, mutation localization, and potential identification of causal sequence variants. In contrast to prior approaches, we have developed a Hidden Markov Model to narrowly define the mutation area by inferring recombination breakpoints of chromosomes in the mutant pool. In addition, we created an interactive online software resource to facilitate automated analysis of sequencing data and demonstrate its utility in the zebrafish and mouse models. Our novel methodology and online tools will make Next Generation Sequencing an easily applicable resource for mutation mapping in all model systems.

INTRODUCTION

There can be little argument that genetic mapping has made a substantial contribution to our understanding of biology. For many years these studies employed phenotypically defined markers, such as those used by Morgan in *Drosophila* and Haldane in mice (Morgan 1911; Haldane et al. 1915). The modern era of genetic analysis was heralded by the recognition that variation in genomic DNA sequence itself could be used as a facile assay for mapping (Botstein et al. 1980). This was initially accomplished using analysis of restriction fragment length polymorphisms, which were later replaced by microsatellites and subsequently by single nucleotide polymorphisms (SNPs). Despite the remarkable technological advances, these approaches hold in common with those of Morgan and Haldane the utilization of pre-specified markers. Next generation sequencing (NGS) technology enables simultaneous discovery of very dense sets of informative markers and actual gene mapping in the same experiment. Here, we present a strategy and computational tools to map genes in model organisms using sequencing of pooled samples. The approach can be applied to any model organism with a characterized genome and to both spontaneous and induced mutants. We demonstrate the utility of the strategy and efficiency of the computational approach by mapping spontaneous and ethylnitrosourea (ENU)-induced developmental mutants in zebrafish and mouse.

Large-scale forward mutagenesis screens in zebrafish have been used with success to investigate fundamental developmental processes. While the recent completion of the zebrafish genome has greatly aided in the identification of genes, mapping analyses

continue to rely on the use of traditional microsatellite markers. However, the utilization of SNPs for mapping of zebrafish mutants was proposed almost a decade ago (Stickney et al. 2002), large numbers of SNPs have been identified (Guryev et al. 2006; Bradley et al. 2007), and the application of next generation sequencing (NGS) for SNP discovery and mutation mapping has been demonstrated (Baird et al. 2008), indicating the feasibility of modern genomic approaches.

The reasons SNP analysis has not been routinely employed in zebrafish genetics may include the fact that most zebrafish strains employed for genetic studies are not inbred and show abundant intrastrain variability, in addition to profuse interstrain haplotype sharing (Guryev et al. 2006). Furthermore, no well-characterized panel of SNPs amenable to automated analysis has been developed, in a manner similar to that which we developed for mice (Moran et al. 2006). Also, it would be desirable to develop methodology that can take advantage of the fact that large numbers of mutant zebrafish progeny can be generated and analyzed as pools (bulk segregant analysis), and, while hybridization-based assays can be used to infer allele distributions in a population (Bradley et al. 2007), this is technically challenging and imprecise.

Fortuitously, the rapid progression in NGS technology makes low-pass whole genome sequencing a practical method for genetic mapping and positional cloning. Firstly, powerful informatics tools enable SNP identification with great efficiency and reliability, such that *a priori* knowledge of SNP variation, while useful, is not required. Also, because individual reads can be counted, the underlying distribution of alleles in a pool

can be enumerated, and the sites of recombination breakpoints assessed. While these could be specifically identified by indexing ("bar-coding") DNA from individual genomes, our method makes this additional step unnecessary. Finally, this approach has the obvious utility of potentially identifying the causal mutation directly, with a likelihood that is proportional to the depth of sequence coverage.

Indeed, NGS analysis of SNP variation has been used for mutation discovery in *C. elegans* (Sarin et al. 2008; Doitsidou et al. 2010) and mice (Arnold et al. 2011). Furthermore, Schneeberger (Schneeberger et al. 2009) and Mokry (Mokry et al. 2011) compared heterozygous and homozygous areas in *Arabidopsis* to identify mutation regions; however, this approach has not been successfully applied to larger vertebrate genomes. Our aim in this report, however, was to determine the map position by integrating SNP variant detection informatics with an algorithm that compares allele distribution between mutant and wild-type pools. This automated method indicates a maximum likelihood interval within which the mutation should reside. Our initial approach employed a simple comparative allele-counting method similar to what has been previously reported (Schneeberger et al. 2009; Mokry et al. 2011). However, this analysis can be undermined by the fact that SNP variation between strains is not randomly distributed, and large non-informative regions can be present. Our method based on maximization of log likelihood computed by a Hidden Markov Model is much more precise, as it integrates allele distribution inferences over large intervals. Moreover, it is tolerant of ambiguity and errors, such as sequencing artifacts and misscoring of mutant embryos.

RESULTS

Genetic mapping experiments for recessive mutations rely on identifying regions of homozygosity in the mutant pool. Mutant pools lack diversity in the genomic region harboring the mutation and the diversity gradually increases with distance from the mutation position due to recombination. Given the incomplete characterization of SNP variants in zebrafish, sequencing of the pool of unaffecteds is required to identify the complete spectrum of informative variants. This is in contrast to previously reported methods in other species, where only affected mutants were sequenced and compared to the reference genome (Sarin et al. 2008). This mapping process can be represented by a Hidden Markov Model (HMM) with the architecture presented in **Figure 1**. Specifically, the number of recombination events from the mutation position can be considered a hidden state. What is observable, however, are sequencing reads covering informative variants in the mutant pool. The number of recombination events from the mutant position determines allele frequency in the sample. If none of the fish chromosomes recombined (HMM state 0), all informative SNPs would be homozygous in the mutant pool. If one recombination occurs (state 1), major allele frequency of an informative SNP in the pool of N chromosomes depends on whether both parents were heterozygous for this SNP or one of the parents was homozygous. If both parents were heterozygous, the major allele frequency is $(N-1)/N$. If one of the parents is homozygous, there is an equal chance that either the mutant pool will be homozygous for that SNP or it will have the same frequency as in case of two heterozygous parents (depending on the parental origin of the recombination event). The same logic applies to additional recombinations,

corresponding to higher HMM states. Short read sequencing can then be modeled as a binomial sampling. Overall, the emission probability for state ***k*** (the probability to observe ***i*** reads supporting major allele out of total ***m*** reads in state ***k***) is given by:

$$P_k(i \mid m) = p(het) \cdot C_m^i \left(\frac{N-k}{N}\right)^i \left(\frac{k}{N}\right)^{m-i} + p(het) \cdot C_m^i \left(\frac{N-k}{N}\right)^{m-i} \left(\frac{k}{N}\right)^i + \\ + p(\text{hom}) \sum_{j=0}^{k} C_k^j \left(\frac{1}{2}\right)^k C_m^i \left(\frac{N-j}{N}\right)^i \left(\frac{j}{N}\right)^{m-i} + p(\text{hom}) \sum_{j=0}^{k} C_k^j \left(\frac{1}{2}\right)^k C_m^i \left(\frac{N-j}{N}\right)^{m-i} \left(\frac{j}{N}\right)^i \quad \textbf{(1)}$$

Here ***p(het)*** and ***p(hom)*** are probabilities that both parents are heterozygous or one parent is homozygous for any given SNP. In case of the outcross of inbred lines, both heterozygous parents should always be expected. In case of an incross, having one homozygous parent is more probable. These probabilities can be easily estimated from the data. Each individual term in the equation represents the contribution of each type of SNP-event to the overall probability, i.e. reference allele is major and either both parents are heterozygous or one parent is homozygous, non-reference allele is major and either both parents are heterozygous or one parent is homozygous.

Transition probabilities between the mentioned HMM states are determined by the recombination rate. Recombination rate is assumed to be constant along the genome. This assumption greatly facilitates computation and is justified by robustness of the method with respect to local variation of recombination rate determined in simulation experiments.

Starting at the mutation position, at which point there are assumed to be no recombination events, the model can successively move to states with increasing number of recombinants. We also allow the model to leave the identity by descent (IBD) region and move to the state representing a "random" genome location or to a state corresponding to sequencing or mapping error.

We compute the log likelihood of the HMM using the Viterbi algorithm starting from discrete locations on the chromosome given that the starting position corresponds to the location of mutation (Nielsen and Sand 2011). The rationale for this approach, as opposed to simply estimating positions of recombination events using HMM, is that the data carry additional information about the mutation location within the completely homozygous region, so the log likelihood is not flat within this homozygous sequence region. That is, the likelihood peak should correspond to the position of the mutation.

For simplicity and computational efficiency, we initially map the mutation to a chromosome by calculating the homozygosity score, expressed as a ratio of heterozygous SNP calls between mutant and control pools multiplied by the number of informative homozygous SNP calls in the mutant pool. This score surveys the entire genome in 10kb windows and can identify an isolated high scoring region on a single chromosome. Then, we fine-map the mutation position using the algorithm outlined above.

We tested the method in a series of simulation experiments. We varied pool sizes, defined as the number of individual mutants or unaffected siblings analyzed together, and

coverage depths to assess the expected maximal likelihood interval size as well as minimum coverage required to roughly map the mutation region (**Supplementary Figures S1-S4**). These simulations demonstrated that larger pools may result in narrower mapping intervals but would require higher coverage depth. Computational experiments with 5-7x coverage for a pool size of 20 individuals resulted in the expected size of the mapping interval of ~2-3 Mb. Increasing pool size to 40 individuals narrowed the expected interval to ~1 Mb (**Supplementary Figures S2-S4**). The estimates obtained by simulations may be optimistic because simulations did not take into account the reference genome quality, the density of errors due to read misalignments and heterogeneity of SNP density along the genome. We also analyzed simulated datasets made by downsampling coverage of obtained experimental results (described below) and producing coverage as low as 1-1.5x per pool. The homozygosity score can still detect the correct region amid increased noise level on the whole genome plot (**Supplementary Figure S1**). However, the HMM score suffers from the inherently low signal and a genomic coverage of at least 2-3x is recommended.

To test the feasibility and accuracy of our approach in a real life experiment, we analyzed two distinct alleles of the same mutant phenotype, *cloche (clo)*. This mutant lacks endothelial and hematopoietic cells (Stainier et al. 1995). One allele, *clo* ^{m39}, is a spontaneous deletion mutant originally described in a semi-wild population from an Indonesian fish farm (Stainier et al. 1995; Stainier et al. 1996), while another one, *clo* ^{s5}, is caused by a presumed point mutation induced by ENU chemical mutagenesis. The *cloche* mutation has been previously mapped to the telomeric region of chromosome 13

(Liao et al. 2000). *Lysocardiolipin acyltransferase* (*lycat*) has been shown to be located in the *clo* ^{m39} deletion region and described as a candidate gene responsible for the *cloche* phenotype (Xiong et al. 2008). For sequencing, 20 pooled mutants and 20 unaffected siblings were obtained from incrosses of *clo* ^{s5} genotype and pools of 160 individuals each for the *clo* ^{m39}. Sequencing each pool on a separate lane of Illumina HiSeq2000 platform using 100 base pair (bp) single-end reads gave between 90 and 110 million reads, resulting in ~6-7-fold genomic coverage (see **Supplementary Table ST1**). Homozygosity scores for both *clo* ^{s5} and *clo* ^{m39} revealed corresponding peaks on the telomeric region of chromosome 13 (**Figure 2a**). The *clo* ^{m39} peak borders an area where no sequence reads were returned for the mutant pool (**Figure 2b**), corresponding to the previously described deletion area (Xiong et al. 2008). Log likelihood analysis with HMM revealed a well-defined ~800kb interval for both *clo* alleles, overlapping with the area indicated by the homozygosity score (**Figure 2c**). Analysis of the *clo* ^{s5} allele revealed a log likelihood peak aligned with the deletion interval determined for the *clo* ^{m39} mutant. The area under the peak that spans a genomic interval of ~800kb contains about 22 presumptive genes, including *lycat.* No homozygous non-synonymous SNPs in the *lycat* coding region or splice junctions were found in *s5* mutants. In our analysis with 20 and even 160 pooled embryos the region of maximal likelihood spanned several genes. This is consistent with the results of computer simulations described above.

It is important to note that identification of all ENU-induced or novel variants within the mutation region can reveal causal mutations directly, given sufficient sequencing coverage. As an example, we performed NGS-based SNP analysis of a spontaneous

zebrafish cilial mutant *ca1* and localized the likely recombinant interval to chromosome 3 by homozygosity score (**Figure 3a**). HMM log likelihood analysis further defined this region to a ~7 Mb interval, while the homozygosity score missed the mutation region due to very low SNP frequency in some areas of the interval (**Figure 3b**). Targeted PCR-sequencing of candidate SNPs revealed a premature stop codon in exon 6 of the *fleer* (*flr)* gene, identifying this *ca1* mutant line as a new allele of *fleer* (Pathak et al. 2007) (**Figure 3c**). In a similar fashion, NGS of the red blood cell mutant *malbec (mlb* ^{bw306}*)* resulted in 8-9-fold genome coverage (see Supplement Table ST1) and identified a mutation region on chromosome 7 (**Figure 4a, b**), consistent with data obtained from prior microsatellite-based mapping analysis of 5440 meioses (**Figure 4c**, BH Paw, personal communication). Here, instead of comparing the mutant genome to unaffected siblings, wild-type parental siblings (aunts/uncles) were used. Further analysis of the SNPs within the mutation region revealed a nonsense mutation in the gene (**Figure 4d**). Allele-specific oligonucleotide hybridization with primers corresponding to the respective wild type and nonsense mutant alleles showed complete linkage with the *mlb* phenotype (**Supplementary Figure S5**). These examples demonstrate that our method enables direct identification of mutation-causing SNPs.

One surprising outcome of our analysis was the observation of unexpected patterns in homozygous allele distribution, which appeared to be caused by reference genome misassembly (**Figure 5a**). Areas with absolute homozygosity suddenly were interrupted by areas of high heterozygosity in the mutant pool followed again by regions of complete homozygosity. The HMM score also showed non-monotonous behavior with severe

drops and following hikes back (**Figures 2c, 5b**). Inspection of these aberrant patterns and their position with respect to scaffolds showed that they were perfectly delineated by the scaffold boundaries, suggesting the misplacement of several scaffolds in the assembly. Removal of erroneously placed scaffolds (**Figure 5b**) resulted in the improvement of the HMM log likelihood curve. Similarly, the incorporation of unplaced scaffolds into the homozygous region could be accomplished by means of the homozygosity score (**Supplementary Figure S6**).

To demonstrate the utility of the method across vertebrate organisms, we applied it to mapping an ENU-induced mutation in a mouse with a holoprosencephaly phenotype (Line 27SH). Craniofacial defects in affected mice include cleft palate, abnormal snout shape, round forehead shape, and abnormal or absent eyes. Gross morphological defects of the brain, smaller or single olfactory bulb and abnormal interhemispheric fissure, were often observed. Histological analysis revealed that the mutants failed to form a nasal septum or to develop normal forebrain with bilateral hemispheres, and often had fused lateral ventricles (**Supplementary Figure S7**). A pool of genomic DNA obtained from nine affected Line 27SH mice was analyzed by NGS, resulting in ~131 million 75 bp paired-end reads (~7.6x genomic coverage)**,** which was subsequently analyzed by homozygosity score and HMM. Fully characterized genetic variation in parental strains available from public databases was used instead of sequencing unaffected siblings. This analysis identified a single peak (approximately 5.5 Mb) confirming and narrowing the independent mapping using SNP genotyping (**Figure 6a, b**). We observed one non-synonymous change that distinguishes mutants from both parental strains in gene *Lrp2*.

Lrp2 is an excellent biological candidate as a targeted knock-out mutation results in a holopropsencephaly phenotype similar to that observed in our ENU-induced mutant line (Willnow et al. 1996). These results demonstrate full applicability of our method to large mammalian genomes.

We implemented the entire variant calling pipeline together with the gene mapping method in a web-based software tool, SNPtrack. The SNPtrack resource is available online at http://genetics.bwh.harvard.edu/snptrack/. Paired files containing sequencing reads (in fastq format) are uploaded as SNPtrack input. Sequencing reads are then automatically aligned by BWA (Li and Durbin 2009) and sequence variants are called using GATK (DePristo et al. 2011). The log likelihood score is then returned to remotely located users. The interactive software enables visualization of the mutation region and direct examination of individual SNPs in the interval. It directly links to the UCSC and EnsEMBL genome browsers to readily highlight annotated genes in the mutation region, which enables rapid biological assessment and potential validation of putative causal loci by PCR sequencing, cRNA complementation, and morpholino-mediated knockdown.

DISCUSSION

SNPtrack enables rapid and accurate mapping of mutants in model organisms and facilitates immediate identification of causative mutations. In the analysis of fish data, where strains maintain natural genetic variation, we employed direct comparison of mutant and unaffected siblings, increasing the density of informative SNPs and reducing mapping errors due to fixed regions of homozygosity in the parental strains. In contrast to

prior approaches, we developed a robust and precise methodology by calculating the log likelihood based on a Hidden Markov Model of recombination breakpoints to define the mutation region. The frequency of intrastrain SNP variation in zebrafish combined with the sensitivity of the analysis obviates the need for outcrossing into polymorphic strains, thereby saving considerable time and resources in the mutant mapping process. We demonstrated that SNPtrack is equally applicable to other species, including mice where the availability of extensively sequenced reference strains results in the complete knowledge of informative SNPs, allowing sequencing of the mutant strains alone. In addition to facilitating gene mapping online, SNPtrack offers the opportunity to compile and compare a large number of sequenced genomes centrally and will serve as an additional resource to further improve and annotate the existing zebrafish genome assembly: this method can improve the assembly of until now not fully completed genomes, by enabling the identification of misassembled contigs through the homozygosity score calculated as part of the mapping process. Use of the SNPtrack software by multiple users will allow the accumulation of these informative data to enhance the accuracy of existing assemblies.

This method also allows mapping of mutations when it is not feasible to provide hundreds or thousands of mutants typically required for classical mapping panels, as we only used as low as 5 or 20 embryos per pool in our studies; this will be of particular interest for hard-to-score phenotypes such as behavioral phenotypes or adult phenotypes, such as cancer susceptibility. Importantly, the presumptive recombination break points are inferred with high resolution, which maximizes the informative content of even

limited numbers of characterized meioses. In addition, our method can support innovative strategies to elucidate oligogenic traits, such as modifier or synthetic lethal screens. Also, it is possible to envision pursuing mutagenesis on fully inbred backgrounds (i.e., without a cross to a mapping strain), using the mutagen-induced sequence variants as informative markers, which will be directly applicable to mice. Further studies are needed to systematically determine the optimal mapping protocol and optimize parameters of the experiment including number of individuals per pool, sequencing coverage depth and length of sequence reads. Sequencing technology is rapidly evolving and will lead to faster and cheaper sequencing of longer reads, further enhancing the impact and accuracy of our approach.

METHODS

DNA from pooled zebrafish embryos and from mutant mice was isolated using DNeasy Blood+Tissue kits (QIAGEN) and sheared (Covaris) to ~150 bp (for single-end reads) and ~350 base length (for paired-end reads). Libraries were constructed using Illumina PE-adapter, NEBNext DNA Library Prep kits and standard Illumina protocols. Invitrogen 2% SizeSelect E-gel was used to purify and size select the library after the adapter ligation step and as the final post-PCR purification step. 1 µg of DNA per pool was sufficient, indicating the applicability of this method for sequencing of zebrafish embryos as early as 2-3 dpf; the amount can further be lowered by special library preparation protocols. NGS was performed on an Illumina HiSeq2000 apparatus using one lane per pool with 100 bp single-end or 75 bp paired-end reads. Returned read number varied between 90 and 132 million, resulting in about 6-10-fold coverage of the genome (see

Supplementary Table ST1). Sequence read files in fastq format were aligned with the zebrafish and mouse reference genome using the Burrows-Wheeler Aligner (BWA) (Li and Durbin 2009), and SNPs were identified by the Genome Analysis Tool Kit (GATK) (DePristo et al. 2011). Analysis of read alignments and subsequent visualization of scores was accomplished using the Integrative Genomics Viewer (IGV) (Robinson et al. 2011). The crude SNP sets were further quality controlled on mapping accuracy and genomic coverage as well as controlled for possible PCR-errors prior to running HMM analysis. Initial homozygosity scoring was performed on the whole genome scale as described above and the HMM algorithm that used the parredhmmlib (Nielsen and Sand 2011) library was run on the highest scoring chromosome giving an accurate measure of the mutation position.

The HMM architecture is shown in **Figure 1**. Transition probabilities between states ***k*** and ***k+1*** within the IBD region are determined by recombination rate and are given by:

$$P_k\left(k+1 \mid k\right) = \left(N-k\right)\cdot l \cdot r \qquad \textbf{(2)}$$

where ***r*** is per nucleotide recombination rate, assumed to be constant across the genome, and ***l*** is the distance between SNPs. For computational feasibility instead of computing transition probability for each and every SNP-to-SNP distance it is valid to take an average fixed distance (currently 100 bp) and if there is more than one informative SNP inside interval ***l*** - dispose the rest, when there is none - an empty SNP (with coverage 0) should be used to maintain the correct distance from mutation. It is worth noting that during this process a small amount of informative SNPs can be lost, though a lot of the SNPs clustered in less than 100bp are misalignment errors which are beneficial to avoid.

Transition probabilities to the "random" state are chosen so the length of the IBD region would be approximately 20Mb. Certainly the size of IBD region significantly depends on the breeding history of a specific strain, but this value will not affect the HMM behavior greatly if set somewhat incorrectly (the main information comes from the emission of events). Transition probability to the error state was set to 10^{-5} and the probability to stay in the error state was set to 0.01 to account for clustered errors. Emission probabilities for the states within the IBD region are given by **Equation 1**. Emission probabilities for the "random" state outside the IBD region are given by a randomized binomial distribution.

Further analysis of variants in the region was performed by the use of Variant Effect Predictor (EnsEMBL) and snpEff tools. All computational steps are fully automated in the web-based SNPtrack software.

Prior to the mouse mapping experiment, A/J male mice were mutagenized using ENU, and outcrossed with C57BL6/N females to generate the first generation (G1) progeny. The line 127 G1 male was outcrossed with C57BL/6N females to obtain G2 progeny; G2 females were backcrossed with the line 127 G1 male and mutant G3 progeny with a holoprosencephaly phenotype were identified. Mutants used for WGS were obtained from three different intercrosses using G2 and G3 parents. Genomic DNA was prepared from embryonic liver or skin of a total of nine mutants. NGS and computational analysis was done as described above.

DATA ACCESS

Next Generation Sequencing data used in this work have been submitted to the NCBI Sequence Read Archive (SRA) under accession no. SRA051382.1.

ACKNOWLEDGEMENTS

This work was supported by a Junior Faculty Grant from the Harvard Stem Cell Institute and institutional start-up funds to W.G., American Heart Association grants to J.D.C. and M.J.K., Sigrid Juselius Foundation grant to H.A., March of Dimes grant to B.H.P., and NIH grants 1R01DK090311 to W.G., 5R01MH084676 to S.S., R01HD036404 and R01MH081187 to D.B, R01DK070838 and P01HL032262 to B.H.P.

AUTHOR CONTRIBUTIONS

I.L. and S.S designed and I.L performed the bioinformatic analysis and created the webtool. K.A. and P.K. isolated mutants and performed DNA isolation and library construction. I.A. assisted with the bioinformatic analysis. C.A.A. and I.A.D. isolated and performed manual sequencing analysis of the *fleer* mutant allele. J.D.C., M.J.K., H.A. and B.H.P isolated and performed confirmation analysis of the *malbec* (*mlb* ^{bw306}) mutant. R.S. and S.H. identified, characterized and provided DNA for the mouse mutant. T.E.N. provided m39 and s5 *cloche* mutants and DNA and contributed to study design. D.R.B, S.S., and W.G. conceived and supervised the project and wrote the manuscript.

REFERENCES

Arnold CN, Xia Y, Lin P, Ross C, Schwander M, Smart NG, Muller U, Beutler B. 2011. Rapid identification of a disease allele in mouse through whole genome sequencing and bulk segregation analysis. *Genetics* **187**(3): 633-641.

Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA, Selker EU, Cresko WA, Johnson EA. 2008. Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS One* **3**(10): e3376.

Botstein D, White RL, Skolnick M, Davis RW. 1980. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am J Hum Genet* **32**(3): 314-331.

Bradley KM, Elmore JB, Breyer JP, Yaspan BL, Jessen JR, Knapik EW, Smith JR. 2007. A major zebrafish polymorphism resource for genetic mapping. *Genome Biol* **8**(4): R55.

DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M et al. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* **43**(5): 491-498.

Doitsidou M, Poole RJ, Sarin S, Bigelow H, Hobert O. 2010. C. elegans mutant identification with a one-step whole-genome-sequencing and SNP mapping strategy. *PLoS One* **5**(11): e15435.

Guryev V, Koudijs MJ, Berezikov E, Johnson SL, Plasterk RH, van Eeden FJ, Cuppen E. 2006. Genetic variation in the zebrafish. *Genome Res* **16**(4): 491-497.

Haldane JBS, Sprunt AD, Haldane NM. 1915. Reduplication in mice. *Journal of Genetics* **5**: 133-135.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**(14): 1754-1760.

Liao W, Ho CY, Yan YL, Postlethwait J, Stainier DY. 2000. Hhex and scl function in parallel to regulate early endothelial and blood differentiation in zebrafish. *Development.* **127** (20): 4303-4313.Mokry M, Nijman IJ, van Dijken A, Benjamins R, Heidstra R, Scheres B, Cuppen E. 2011. Identification of factors required for meristem function in Arabidopsis using a novel next generation sequencing fast forward genetics approach. *BMC Genomics* **12**: 256.

Moran JL, Bolton AD, Tran PV, Brown A, Dwyer ND, Manning DK, Bjork BC, Li C, Montgomery K, Siepka SM et al. 2006. Utilization of a whole genome SNP panel for efficient genetic mapping in the mouse. *Genome Res* **16**(3): 436-440.

Morgan TH. 1911. Random Segregation Versus Coupling in Mendelian Inheritance. *Science* **34**(873): 384.

Nielsen J, Sand A. 2011. Algorithms for a parallel implementation of Hiiden Markov Models with a small state space. *IEEE International Parallel & Distributed Processing Symposium*: 447-454.

Pathak N, Obara T, Mangos S, Liu Y, Drummond IA. 2007. The zebrafish fleer gene encodes an essential regulator of cilia tubulin polyglutamylation. *Mol Biol Cell* **18**(11): 4353-4364.

Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. 2011. Integrative genomics viewer. *Nat Biotechnol* **29**(1): 24-26.

Sarin S, Prabhu S, O'Meara MM, Pe'er I, Hobert O. 2008. Caenorhabditis elegans mutant allele identification by whole-genome sequencing. *Nat Methods* **5**(10): 865-867.

Schneeberger K, Ossowski S, Lanz C, Juul T, Petersen AH, Nielsen KL, Jorgensen JE, Weigel D, Andersen SU. 2009. SHOREmap: simultaneous mapping and mutation identification by deep sequencing. *Nat Methods* **6**(8): 550-551.

Stainier DY, Fouquet B, Chen JN, Warren KS, Weinstein BM, Meiler SE, Mohideen MA, Neuhauss SC, Solnica-Krezel L, Schier AF, Zwartkruis F, Stemple DL, Malicki J, Driever W, Fishman MC. 1996. Mutations affecting the formation and function of the cardiovascular system in the zebrafish embryo. *Development* **123**: 285-292.

Stainier DY, Weinstein BM, Detrich HW, 3rd, Zon LI, Fishman MC. 1995. Cloche, an early acting zebrafish gene, is required by both the endothelial and hematopoietic lineages. *Development* **121**(10): 3141-3150.

Stickney HL, Schmutz J, Woods IG, Holtzer CC, Dickson MC, Kelly PD, Myers RM, Talbot WS. 2002. Rapid mapping of zebrafish mutations with SNPs and oligonucleotide microarrays. *Genome Res* **12**(12): 1929-1934.

Willnow TE, Hilpert J, Armstrong SA, Rohlmann A, Hammer RE, Burns DK, Herz J. 1996. Defective forebrain development in mice lacking gp330/megalin. *Proc Natl Acad Sci U S A* **93**(16): 8460-8464.

Xiong JW, Yu Q, Zhang J, Mably JD. 2008. An acyltransferase controls the generation of hematopoietic and endothelial lineages in zebrafish. *Circ Res* **102**(9): 1057-1064.

FIGURE LEGENDS

Figure 1

Schematic representation of mutation analysis by NGS and SNP mapping.

Figure 2

Mapping of two alleles of the *cloche* mutant. a – Homozygosity scoring of the *clo* ^{s5} and *clo* ^{m39} alleles reveals corresponding overlapping peaks on chromosome 13, where the mutation has been previously mapped. b – Sequence pileup of the *clo* ^{m39} allele confirms the previously established deletion and reveals the boundaries of the deletion interval. c – Log likelihood scoring of the *clo* ^{s5} alleles achieves a higher resolution of the mutation region with a narrower interval than the homozygosity scoring method.

Figure 3

Mapping of *ca1*. a – Homozygosity scoring reveals a single peak on chromosome 3. b – Log likelihood analysis (red line) returns a ~7 Mb interval for the presumptive mutation, while the homozygosity score (green graph) misses the mutation region due to very low SNP frequency in some areas of the interval. c – Sequencing of exon 6 of the *flr* gene reveals a premature stop codon.

Figure 4

Mapping of *malbec*. a – Homozygosity scoring maps the mutation to chromosome 7. b – Log likelihood calculation reveals a ~23 Mb high-scoring interval. c - Meiotic map of the

mlb locus on chromosome 7. The tightly linked genetic marker *snp1* was identified by a traditional chromosomal walk. d – NGS sequence analysis of *mlb* and control embryos identified a nonsense mutation in a candidate gene disrupted in *mlb*. This nonsense mutation is non-recombinant with the mutant phenotype, and resolved all remaining 12 genetic recombinants to a resolution <1/5440 meioses. .

Figure 5:

Improving the genome assembly using NGS-based SNP mapping. a – Analysis of homozygous and heterozygous SNPs in the *clo* ^{m39} mutation region demonstrates abrupt changes in the frequency of heterozygous SNPs, corresponding with 3 distinct scaffolds. b – Removal of the scaffolds indicated above improves the HMM log likelihood score.

Figure 6:

Application to ENU-induced mouse mutant a – Homozygosity scoring reveals a single peak on chromosome 2. b – Log likelihood analysis highlights a 5.5 Mb interval harboring a single non-synonymous change in the candidate gene *Lrp2*.

Figure 1

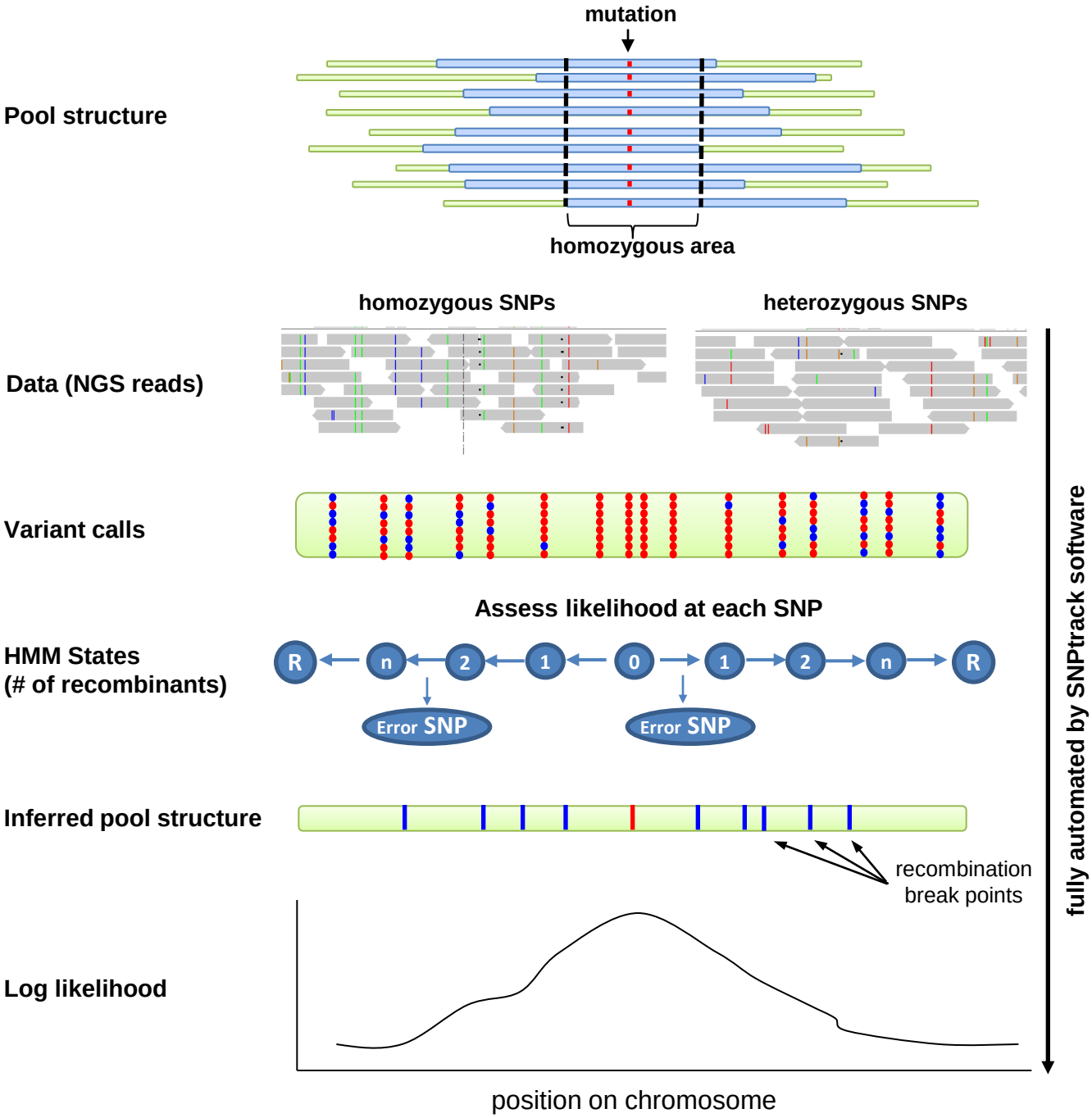
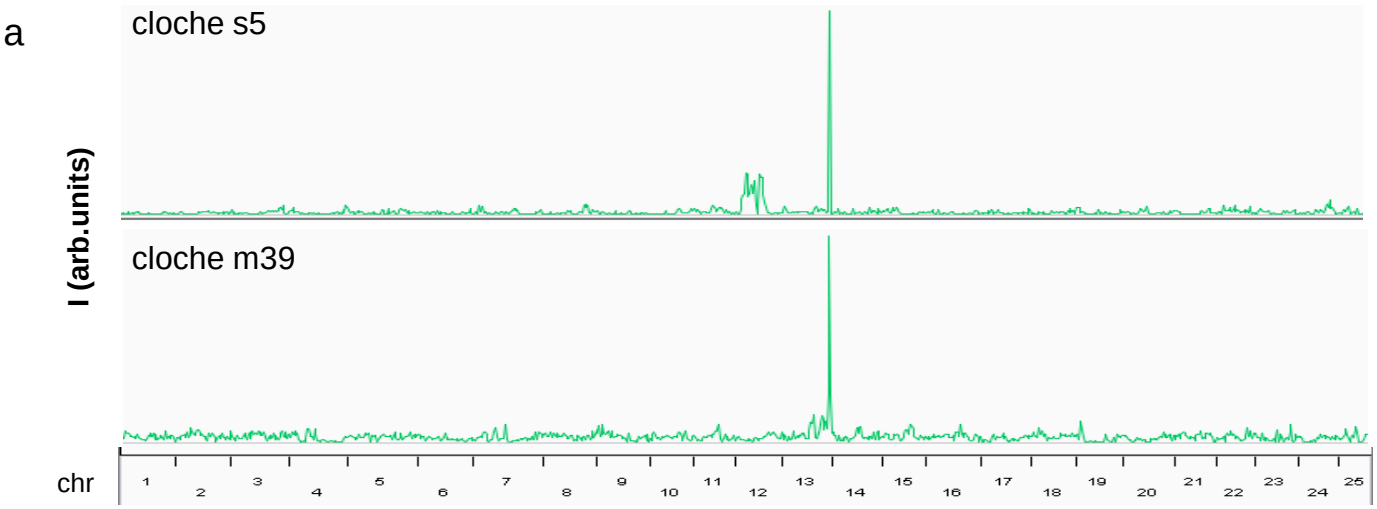
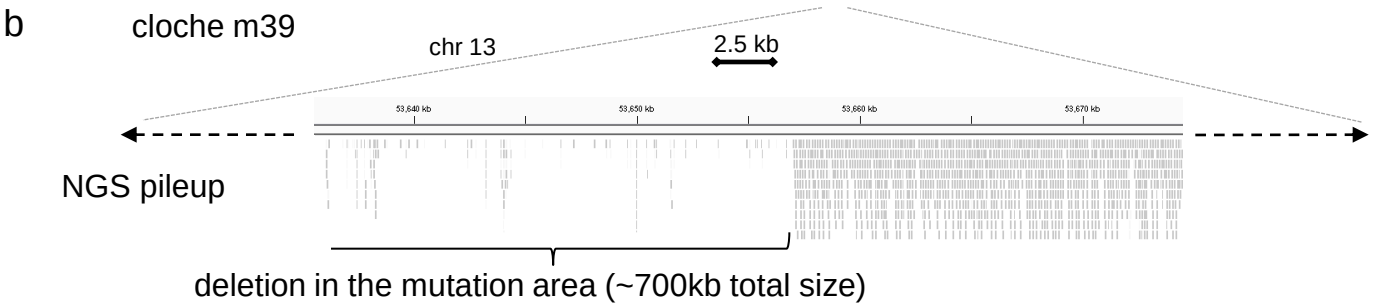

mutation
Pool structure
homozygous area
homozygous SNPs
heterozygous SNPs
Data (NGS reads)
Variant calls
Assess likelihood at each SNP
HMM States
(# of recombinants)
R
n
2
1
0
1
2
n
R
Error SNP
Error SNP
Inferred pool structure
recombination
break points
Log likelihood
position on chromosome
fully automated by SNPtrack software

Figure 2


a
cloche s5
I (arb.units)
cloche m39
chr
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25


b
cloche m39
chr 13
2.5 kb
53,640 kb
53,650 kb
53,660 kb
53,670 kb
NGS pileup
deletion in the mutation area (~700kb total size)

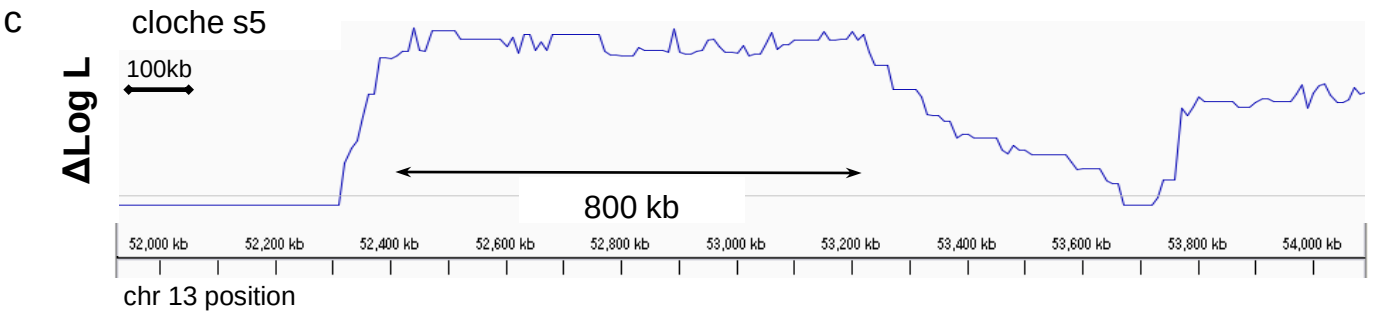
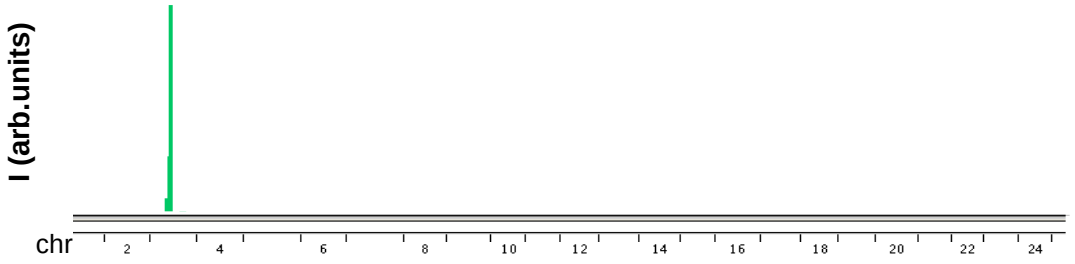
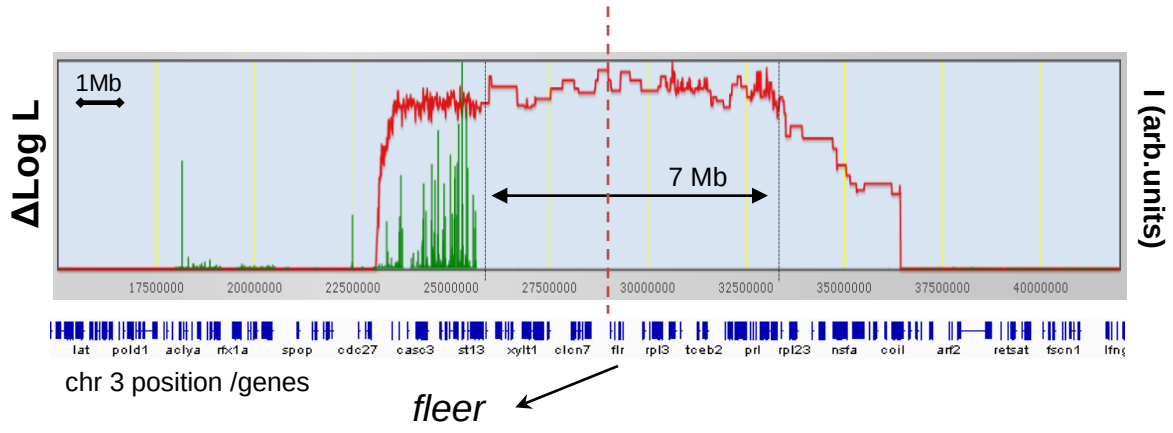

c
cloche s5
ΔLog L
100kb
800 kb
52,000 kb
52,200 kb
52,400 kb
52,600 kb
52,800 kb
53,000 kb
53,200 kb
53,400 kb
53,600 kb
53,800 kb
54,000 kb
chr 13 position

Figure 3

a


I (arb.units)
chr
2
4
6
8
10
12
14
16
18
20
22
24

b


ΔLog L
1Mb
7 Mb
I (arb.units)
17500000
20000000
22500000
25000000
27500000
30000000
32500000
35000000
37500000
40000000
lat
pold1
aclya
rfx1a
spop
cdc27
casc3
st13
xylt1
clcn7
flr
rpl3
tceb2
prl
rpl23
nsfa
coil
arf2
retsat
tscn1
chr 3 position /genes
fleer

c

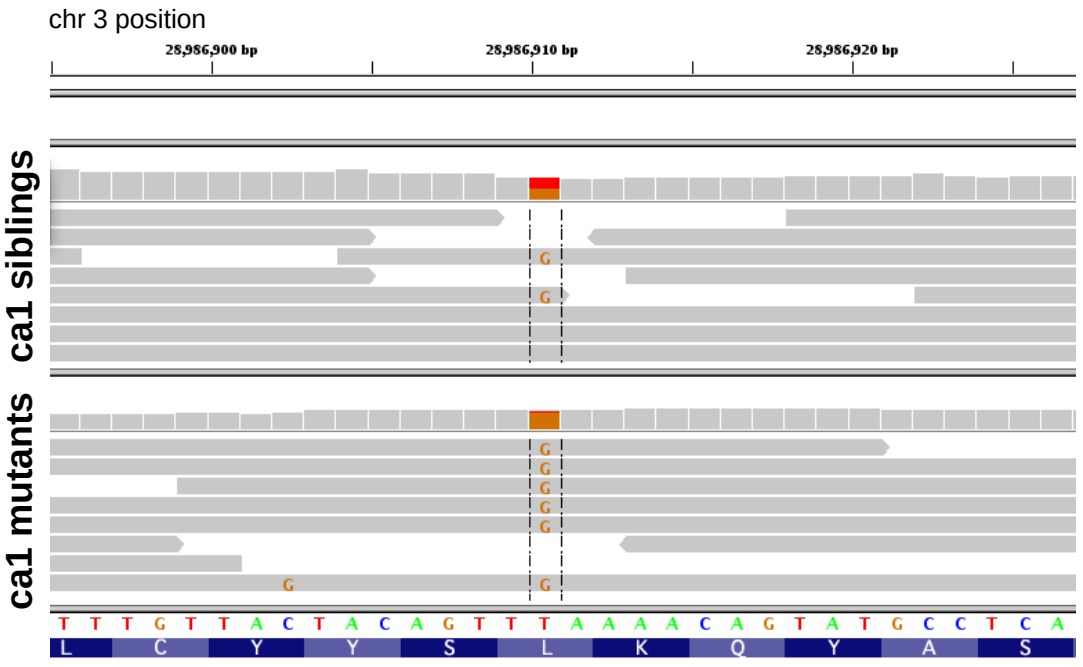
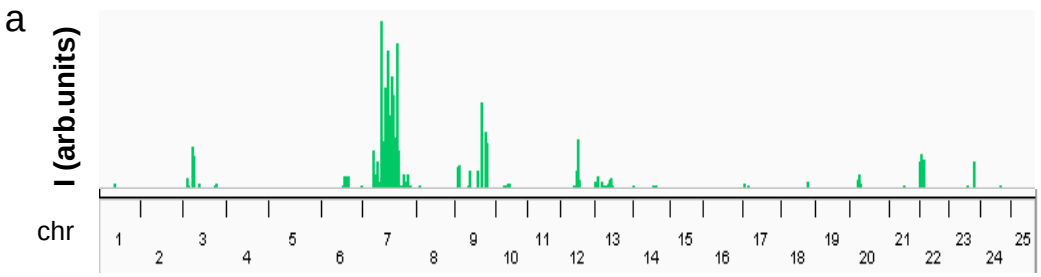
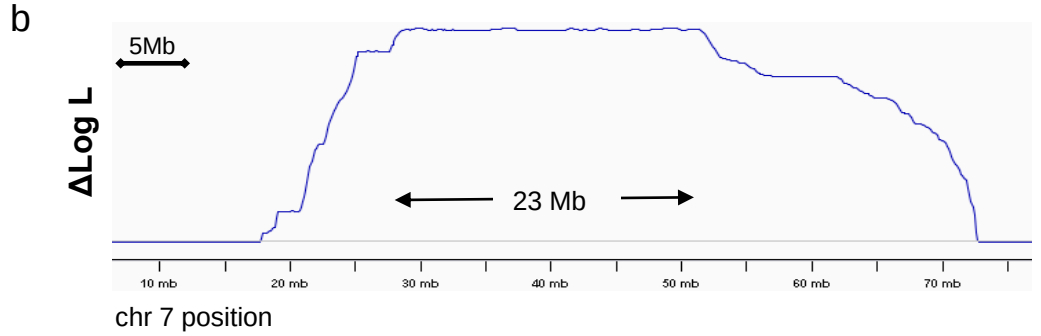
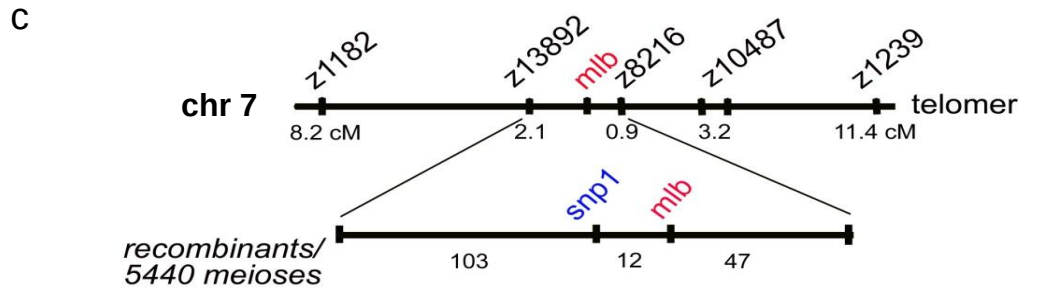

chr 3 position
28,986,900 bp
28,986,910 bp
28,986,920 bp
ca1 siblings
ca1 mutants
T T T G T T A C T A C A G T T T A A A A C A G T A T G C C T C A
L C Y Y S L K Q Y A S

Figure 4


a
I (arb.units)
chr
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25


b
5Mb
ΔLog L
23 Mb
10 mb
20 mb
30 mb
40 mb
50 mb
60 mb
70 mb
chr 7 position


c
chr 7
z1182
z13892
mlb
z8216
z10487
z1239
telomer
8.2 cM
2.1
0.9
3.2
11.4 cM
snp1
mlb
recombinants/
5440 meioses
103
12
47

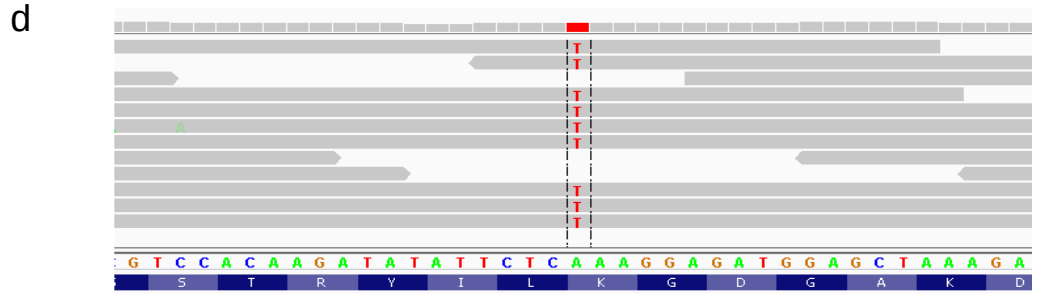
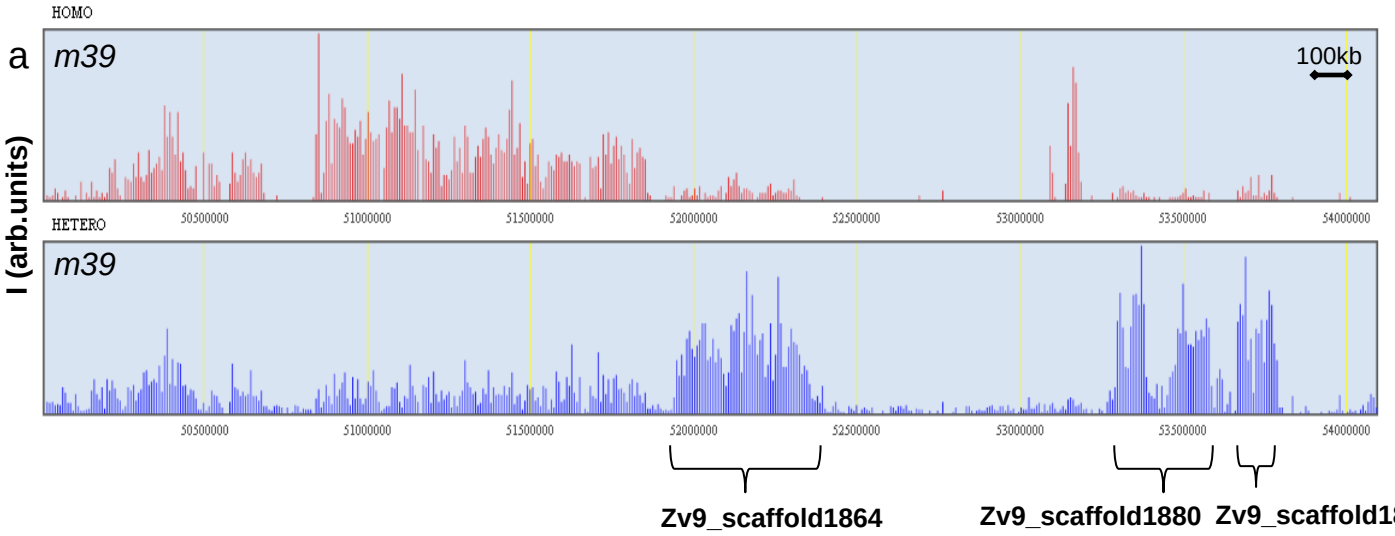

d

Figure 5


a
HOMO
m39
100kb
HETERO
m39
I (arb.units)
50500000
51000000
51500000
52000000
52500000
53000000
53500000
54000000
Zv9_scaffold1864
Zv9_scaffold1880
Zv9_scaffold1882

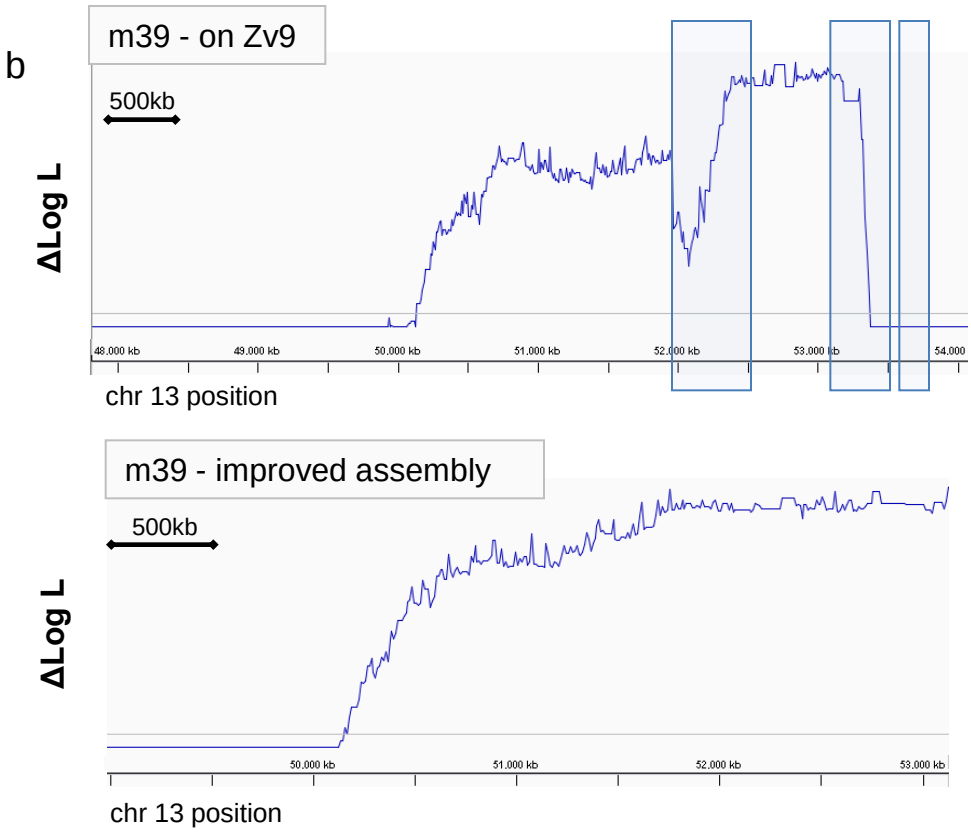
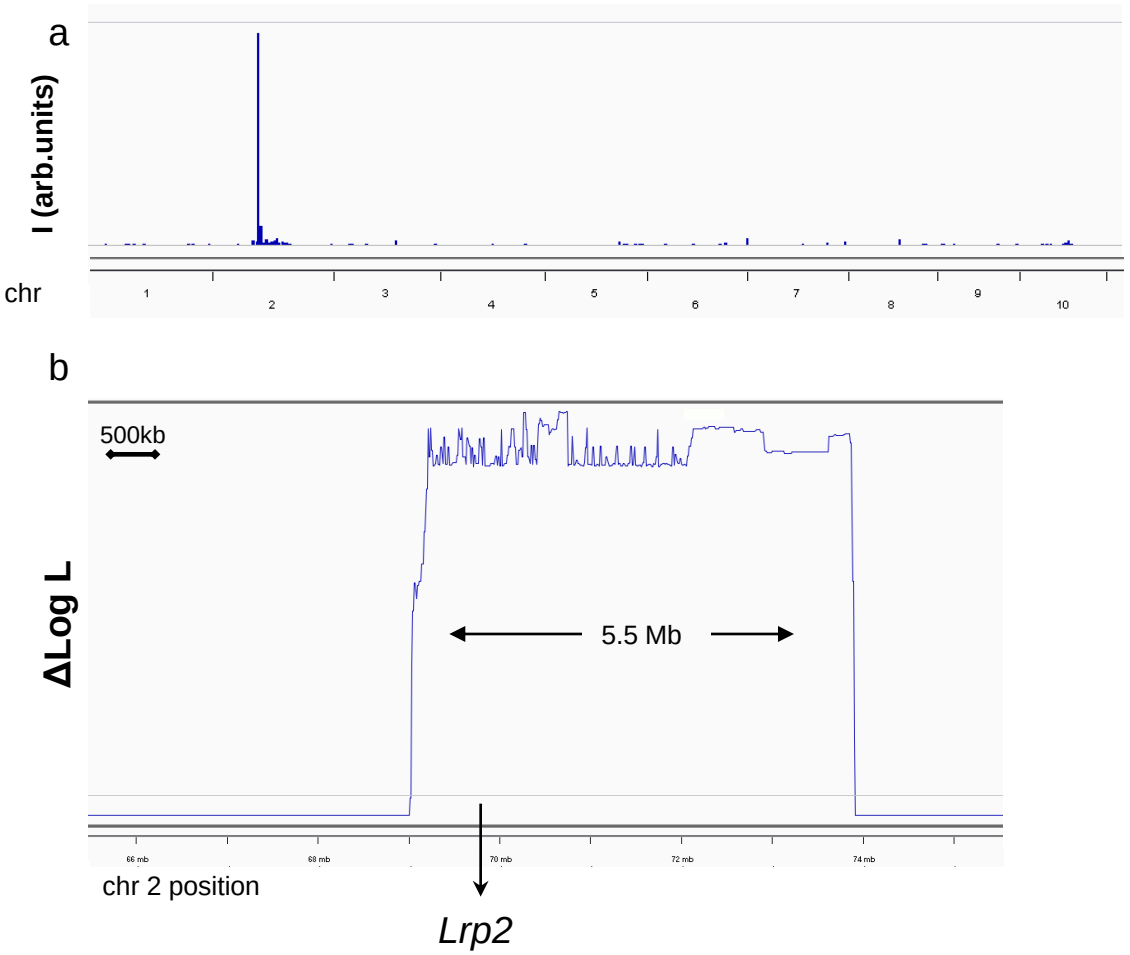

b
m39 - on Zv9
500kb
ΔLog L
48.000 kb
49.000 kb
50.000 kb
51.000 kb
52.000 kb
53.000 kb
54.000
chr 13 position
m39 - improved assembly
500kb
ΔLog L
50.000 kb
51.000 kb
52.000 kb
53.000 kb
chr 13 position

Figure 6


a
I (arb.units)
chr
1
2
3
4
5
6
7
8
9
10
b
500kb
ΔLog L
5.5 Mb
66 mb
68 mb
70 mb
72 mb
74 mb
chr 2 position
Lrp2