



MMAPPR: Mutation Mapping Analysis Pipeline for Pooled RNA-seq

Jonathon T Hill, Bradley L Demarest, Brent W Bisgrove, et al.

Genome Res. published online January 8, 2013

Access the most recent version at doi:[10.1101/gr.146936.112](https://doi.org/10.1101/gr.146936.112)

P<P	Published online January 8, 2013 in advance of the print journal.
Accepted Manuscript	Peer-reviewed and accepted for publication but not copyedited or typeset; accepted manuscript is likely to differ from the final, published version.
Open Access	Freely available online through the <i>Genome Research</i> Open Access option.
Creative Commons License	This manuscript is Open Access. This article is distributed exclusively by Cold Spring Harbor Laboratory Press for the first six months after the full-issue publication date (see http://genome.cshlp.org/site/misc/terms.xhtml). After six months, it is available under a Creative Commons License (Attribution-NonCommercial 3.0 Unported License), as described at http://creativecommons.org/licenses/by-nc/3.0/ .
Email Alerting Service	Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or click here .

Comprehensive immune receptor profiling.
Discover the **DriverMap™ AIR Assay** difference.

LEARN
MORE



To subscribe to *Genome Research* go to:
<https://genome.cshlp.org/subscriptions>

MMAPPR: Mutation Mapping Analysis Pipeline for Pooled RNA-seq

Jonathon T. Hill, Bradley L. Demarest, Brent W. Bisgrove, Bushra Gorski, Yi-Chu Su,

H. Joseph Yost

Department of Neurobiology and Anatomy, University of Utah Molecular Medicine Program

Corresponding Author:

Dr. H. Joseph Yost, PhD

University of Utah

Eccles Institute of Human Genetics

15 North 2030 East, Room 3160

Salt Lake City, Utah 84112

Phone: 801-585-6110

Email: jyost@genetics.utah.edu

Running Title: Mutation mapping by RNA-seq

Keywords: Genetic mapping, RNA-seq, zebrafish

ABSTRACT

Forward genetic screens in model organisms are vital for identifying novel genes essential for developmental or disease processes. One drawback of these screens is the labor-intensive and sometimes inconclusive process of mapping the causative mutation. In order to leverage high-throughput techniques to improve this mapping process, we have developed a Mutation Mapping Analysis Pipeline for Pooled RNA-seq (MMAPPR) that works *without* parental strain information or requiring a pre-existing SNP map of the organism, and adapts to differential recombination frequencies across the genome. MMAPPR accommodates the considerable amount of noise in RNA-seq datasets, calculates allelic frequency by Euclidean distance followed by Loess regression analysis, identifies the region where the mutation lies and generates a list of putative coding region mutations in the linked genomic segment. MMAPPR can exploit RNA-seq datasets from isolated tissues or whole organisms that are utilized for gene expression and transcriptome analysis in novel mutants. We tested MMAPPR on two known mutant lines in zebrafish, *nkx2.5* and *tbx1*, and used it to map two novel ENU-induced cardiovascular mutants, with mutations found in the *ctr9* and *cds2* genes. MMAPPR can be directly applied to other model organisms, such as *Drosophila* and *C. elegans*, that are amenable to both forward genetic screens and pooled RNA-seq experiments. Thus, MMAPPR is a rapid, cost-efficient, and highly automated online pipeline, available to perform mutant mapping in any organism with a well assembled genome.

INTRODUCTION

Forward genetic screens in zebrafish have identified a large number of genes essential for organogenesis (Haffter et al. 1996; Driever et al. 1996), laterality (Chen et al. 2001), axon guidance (Xiao et al. 2005) and cancer development (Moore et al. 2006), many of which have been linked to human disease. Similar large-scale forward-genetic screens have been and continue to be performed in mice (Yu et al. 2004; García-García et al. 2005), *Drosophila* (Nüsslein-Volhard and Wieschaus 1980; Medina et al. 2006) and *C. elegans* (Brenner 1974; Hughes et al. 2011). However, mapping mutants in many species traditionally has been labor-intensive and often inconclusive, especially in organisms with relatively complex genomes.

Several methods exist to expedite genetic mapping. For example, genotyping DNA pooled from phenotype-sorted individuals (bulk segregant analysis) has long been a standard method for low-resolution genetic mapping. Bulk segregant analysis provides a qualitative estimate of the linkage between a given marker and the mutant locus, while greatly reducing the time and expense of genotyping. However, this method is still labor-intensive because it requires that each marker be analyzed individually.

The development of techniques using genotyping arrays (Tabernero et al. 2012), genomic re-sequencing of individuals (Liling L Warren et al. 2012) and exome-capture sequencing (Xi Lin et al. 2012) have made mapping mutations much more rapid in human populations by allowing multiple markers to be analyzed simultaneously, but they have been less widely adopted in many model organisms because of incomplete genomic annotation, high polymorphism rates and the costs associated with performing these analyses on large numbers of individuals. Recently, several methods to use whole genome sequencing techniques to model organisms have been proposed for *A. thaliana*

(Schneeberger et al. 2009; Cuperus et al. 2010; Ryan S Austin et al. 2011; Uchida et al. 2011), zebrafish (Voz et al. 2012; Bowen et al. 2012; Leshchiner et al. 2012), mice (Arnold et al. 2011) and *C. elegans* (Zuryn et al. 2010; Doitsidou et al. 2010).

An alternative to whole genome sequencing (WGS) is RNA-seq, which is less expensive because the transcriptome is smaller than the genome, allowing greater read depth to be achieved with fewer reads. The utility of RNA-seq analysis for mapping has been demonstrated recently in self-pollinated individuals derived from inbred mapping strains in maize (Liu et al. 2012), but this has not been tested in more noisy datasets from outbred animal populations. In addition to mapping, RNA-seq is becoming a standard analysis method in model organisms for determining the gene expression and splicing changes underlying phenotypes derived from both forward and reverse genetics (Vesterlund et al. 2011; Aanes et al. 2011; Rösel et al. 2011).

As RNA-seq of pooled individuals can be used for differential expression analysis to further understand the phenotypes of novel mutants from forward genetic-screens, we sought to develop a method to use this data to identify the causative mutation underlying the observed phenotype, thus creating an inexpensive and rapid alternative to traditional mapping procedures. We have designed our method, which we call MMAPPR (Mutant Mapping Analysis Pipeline for Pooled RNA-seq), to directly utilize the data and experimental design typical in RNA-seq based transcriptome experiments. Although this paper goes through the principles and optimization of MMAPPR analysis as well as the details of successfully mapping four mutants, the average user will not be required to have this level of expertise and can simply process their datasets through our program (available at <http://yost.genetics.utah.edu/software.php>) in order to identify their mutant genes. We have validated MMAPPR on two known mutants, *nkx2.5*

(Targoff, KV, unpublished) and *tbx1* (Piotrowski et al. 2003), and two unknown mutant lines, *zy13* and *zy14*, identified in an ENU screen performed in our lab. MMAPPR was then used to identify a genomic region containing the mutation and generate a list of nonsynonymous mutations that serve as candidates for the gene encoding the causative mutation. In each case, the identified causative mutation was less than 1 cM from the maximum score generated by MMAPPR, indicating that MMAPPR is able to successfully and accurately identify mutations derived from a forward genetic screen in zebrafish. In addition to zebrafish, MMAPPR can be directly applied to other organisms, such as *D. melanogaster* and *C. elegans*, where both forward genetic screens and pooled RNA-seq experiments are common, thus removing a significant barrier for performing mutagenesis screens in model organisms.

RESULTS

We developed a novel method, MMAPPR, for identifying recessive mutations identified in forward genetic screens (outlined in Figure 1). Briefly, MMAPPR uses RNA-seq data from F2 embryos separated by phenotype into two pools: wild-type phenotype (which includes homozygous wild-type and heterozygotes) and mutant phenotype (which includes homozygous mutants). Candidate molecular mutations are then identified based on three criteria: physical location in the linked region, expression at the time of tissue collection, and effect on protein amino-acid sequence.

RNA-seq data contains many thousands of single nucleotide polymorphic markers (SNPs) spread across the entire genome, making it an ideal source for high-throughput mapping. However, these datasets are extremely noisy due to the variable expression levels of individual genes across the genome at the time of RNA collection,

PCR amplification artifacts, sequencing errors, mapping errors and genome annotation errors. MMAPPR compensates for the noise inherent in RNA-seq datasets and uses it to map mutations. MMAPPR encompasses five steps: (1) RNA sequencing and mapping, (2) allele frequency distance calculation, (3) signal processing, (4) candidate single nucleotide polymorphism (SNP) identification and (5) candidate confirmation. We have created a MMAPPR software package to perform steps 2-4, while steps 1 and 5 involve bench work and pre-existing software packages; however, because all five steps are integral to the process, each one is covered in more detail below.

RNA sequencing and mapping

Any pool of individuals derived from a cross between two heterozygous carriers of a SNP will contain a Mendelian distribution of genotypes (expected frequencies 0.25 AA, 0.5 Aa, and 0.25 aa) at every SNP locus where both parents were heterozygous. Similarly, the expected allele frequencies of such a SNP are $f_A=0.5$ and $f_a=0.5$. However, when the pool of individuals is subdivided into two pools based on a mutant phenotype, the expected allele frequencies for any SNP depends on its linkage to the mutation causing the phenotype. For example, a SNP located 10 cM away from the causative mutation has an expected allele frequency of 90% for the allele linked to the causative mutation and 10% for the complementary allele.

MMAPPR uses this principle by selecting polymorphic SNPs from mapped RNA-seq reads and calculating the SNP allele frequencies in each pool, and then uses this information to estimate the location of the causative mutation (described below). Therefore, the theoretical resolution of these SNPs is a function of both population size and read depth. It is expected that increasing these factors would improve results until other factors such as dataset noise and SNP density in the genome become limiting

factors. In order to test the effect of population size, we crossed *zy14* heterozygotes and prepared RNA-seq libraries from twenty phenotypically mutant and twenty phenotypically wild-type sibling embryos. We did not adjust this parameter to fewer than 20 individuals because of the amount of material typically required to build libraries for RNA-seq. This was performed three times, generating six libraries of three separate biological replicates, for the *zy14* line. We then used MMAPPR to analyze the 20-embryo RNA-seq datasets pairwise (matched sets of mutant and wild-type pools), and then combined the RNA-seq datasets from the three independent biological repeats to create a single dataset representing 60 embryos. Surprisingly, increasing the number of individuals from 20 to 60 resulted in only a marginal decrease in the width of the detected peak (from 7.1 MB to 5.8 MB for the *zy14* dataset). It is important to note that improvements in peak size and shape are limited by data noise and SNP density. Therefore, the observations that the peak was only marginally improved by increasing numbers of embryos in the pools and that the peak maximum is less than 1 cM from the causative mutation indicate that we may have reached these limits, thus masking any improved resolution gained by increasing the number of individuals in each pool. However, the increased number of individuals did increase the height of the peak from an average of 0.73 in the individual datasets to 1.29 in the combined dataset. Together, this suggests that RNA-seq libraries generated from twenty individuals should be sufficient for mapping.

An important question is how many reads are required in an RNA-seq dataset for MMAPPR. We analyzed the effects of read count on peak width and height. Read depth across the genome in RNA-seq datasets is widely variable, with most regions having very low coverage (Figure S1). Due to the differences in gene expression levels at different

developmental stages or in different tissues, SNP read coverage in any RNA-seq data is skewed and has a very high variance. Thus, the ‘mean coverage’ across the genome (which is important for WGS analysis) is not a stable or intuitive summary statistic for RNA-seq based analysis. Therefore, we utilized ‘total reads per pool’ as a statistic that is applicable across a variety of RNA-seq experiments. Twenty datasets were created by randomly downsampling reads from the combined (60 embryo) dataset, generating datasets that contained from 1% to 100% of the reads from the original data. The 1% dataset (containing 835,714 and 861,895 reads in the wild-type and mutant pools, respectively) did not contain enough SNPs meeting thresholds to identify a linked region. The next larger dataset (containing 5,008,193 and 5,165,185 reads in the wild-type and mutant pools, respectively) generated a 10.1 megabase peak and all remaining, larger datasets generated peaks between 5.7 and 6.9 megabases wide (Figure 2A). Peak maxima were also more volatile in the smaller datasets, but did not trend towards a higher or lower score as the number of reads was increased (Figure 2B). Therefore, both peak width and maximum score are robust at read depths greater than 10 million total reads per pool, which is within the normal range of RNA-seq datasets. However, increasing the read depth is expected to improve sensitivity for including genes in the final candidate list, which may significantly impact identification of low expressed genes. This information will be useful for strategizing how many barcoded samples can be included in a single lane of sequencing.

Two other important considerations when designing RNA-seq experiments that will be utilized in MMAPP are RNA collection time and tissue. It is important to point out that MMAPP will identify the linked interval regardless of time point or tissue utilized for RNA collection. For example, SNPs in neighboring genes that are expressed

at the time of RNA collection should be sufficient to identify the linked interval, even if the mutant gene is not expressed at that time. Obviously, discovery of a SNP within the mutant gene requires that the gene be expressed to be part of the RNA-seq dataset. We therefore suggest isolation of the affected tissue when feasible. However, as exemplified by mapping and identification of previously uncharacterized *zy13* and *zy14* mutants from our lab (see below), the cell lineage that is perturbed in a mutant is not always known or readily isolated, so in some cases, RNA isolated from whole embryos can be used. RNA should also be collected as soon as the embryos can be reliably segregated by phenotype to increase the likelihood of the gene being expressed.

After RNA collection and sequencing, reads are mapped to a reference genome. We used Novoalign (Novocraft) for read mapping. Although the algorithm should be software independent, selection of optimized alignment parameters may differ due to small differences in alignment and scoring algorithms in other software. MMAPP contains several signal processing steps (see below) to compensate for noise, but selecting the proper alignment parameters also improves results. One common filtering criterion for genomic mapping is the minimum mapping quality score (mapq). This score is a phred scaled probability that the read is misaligned, and is affected by the quality of the sequencing read, the number of mismatches/gaps that occur at the putative mapping location and the uniqueness of the sequence in the genome—all factors that can affect SNP calling. In order to identify the optimized mapq, we measured the distance from the peak maximum to the known mutation in the *nkx2.5* line using various mapq cutoffs (Figure 2C). This analysis showed that a cutoff of 30 resulted in a peak maximum nearest to the known mutation. Retrospectively, this cutoff was also the ideal for all of the lines tested, although we did not optimize the other lines

before analysis. Changing the score threshold did not have a large effect on any of the tested lines (largest range = 0.61 MB), indicating that the algorithm is robust against the effects of low quality reads on SNP calling.

Euclidean Distance calculation

After genomic mapping, two BAM formatted alignment files are submitted to the MMAPPR software package. MMAPPR then generates a *de novo* catalog of informative SNPs by identifying genomic positions at which there is a mixture of alleles in the phenotypically wild-type pool (containing both wild-type and heterozygous individuals). MMAPPR does not consider whether the base calls at these catalogued positions match the reference genome. This allows the algorithm to work regardless of parental background or the strain from which the genome build was derived. Small insertions or deletions (indels) and other genomic changes are not used due to the current limits of indel-calling algorithms. However, once these algorithms are reliable, they will be added to MMAPPR. The resulting dataset is then processed to minimize the effects of noise.

RNA-seq datasets are typically noisy due to a combination of several factors including library/PCR artifacts, sequencing error rates, mapping errors, etc. Therefore, measuring the relative allele frequencies between the mutant and control RNA-seq datasets requires a metric that is not susceptible to this noise. Traditional mapping methods typically use the log odds ratio (LOR) to measure linkage disequilibrium as shown in the equation:

$$LOR = \left| \log \left(\frac{a_{mut}/A_{mut}}{a_{wt}/A_{wt}} \right) \right|$$

where ‘ a ’ is the number of reads containing a non-reference allele at the position and ‘ A ’ is the number of reads containing the reference allele at the position. This metric has several characteristics that make it well suited for mapping analyses. Among these is the fact that it is asymptotic with its limit corresponding with the actual mutation, increasing its sensitivity near the segregating locus. However, its asymptotic nature also makes LOR extremely sensitive to noise generated from inaccurate measures of allele frequency and to situations where there is a zero in the denominator, both of which are common in relatively low coverage RNA-seq data. The combination of these factors resulted in a bimodal distribution of LOR scores with high LOR scores representing SNPs where low coverage stochastically resulted in a zero in the denominator. These locations are spread across the entire genome, masking any SNPs linked to the causative mutation (Figure S2).

Recent genome resequencing techniques in plants (Schneeberger et al. 2009) and *C. elegans* have (Doitsidou et al. 2010) used the density of SNP markers from the non-mutant line (based on previously generated SNP maps for parental strains) as an alternative metric to identify a putative linked region. However, RNA-seq datasets are often obtained as mutant and control pairs and comparable sets are not obtained from parental lines, so this metric was not suited for our needs, as MMAPPR is specifically designed to work *without* parental strain information. Another potential method was to identify regions with homozygous SNPs in the mutant pool with or without incorporating this data into statistical models (Liu et al. 2012). However, RNA-seq is susceptible to false positives in low coverage regions, resulting in many regions throughout the genome showing a large number of homozygous SNPs and this method

is susceptible to imperfect phenotypic identification or penetrance as non-homozygous candidates are automatically excluded.

Given the reasons for rejecting the approaches outlined above, we chose to measure allele segregation using Euclidean Distance (ED), as a metric that does not require parental strain information and is resistant to noise, using the equation:

$$ED = \sqrt{(A_{mut} - A_{wt})^2 + (C_{mut} - C_{wt})^2 + (G_{mut} - G_{wt})^2 + (T_{mut} - T_{wt})^2}$$

where each letter (A, C, G, T) corresponds to the frequency of its corresponding DNA nucleotide. In practice, SNP loci with more than two variants are extremely rare, so two of the terms will be zero. Frequencies are used in place of raw read counts to compensate for read coverage differences for loci across the genome and between mutant and wild-type pools. Because frequencies cannot be accurately measured at low read counts, a minimum cutoff of 10 reads is used. ED is advantageous because it is linear, making it less prone to errors in allelic frequency analysis, and is able to subtract out sequence-specific errors, an artifact of Illumina sequencing technology (Nakamura et al. 2011) that is assumed to be equally present in both samples. In contrast, sequence-specific artifacts can have a large effect on LOR, especially near a segregating mutant due to its asymptotic nature.

As an example of the combined effects of noise on LOR and ED, we calculated the theoretically expected and experimentally observed LOR and ED scores for the *nkx2.5* mutation. Expected scores were calculated using the read coverage numbers from the *nkx2.5* RNA-seq dataset and assuming the ideal frequencies of 100% mutant allele in the MUT pool and 33% mutant allele in the WT pool. The expected LOR for a completely linked SNP is 8.01, but the observed LOR score for the previously identified *nkx2.5* SNP was 1.20, which is only 15% of the expected value and within one standard deviation

from the median. In contrast, using the same RNA-seq datasets, the expected ED was 0.89 and the observed value was 0.71—80% of the expected value and several standard deviations above the median. Therefore, using the ED robustly compensates for the noise found in RNA-seq datasets.

Signal processing

Although measurement of allelic segregation between mutant and phenotypically wild-type (WT) pools by ED was greatly improved and did not show the bimodal distribution of SNP enrichments seen with the LOR analysis (Figure S2), it still showed considerable noise (Figure 3A). Therefore, MMAPPR employs two signal-processing steps to identify the linked genomic sequence: raising the distance measurement to a power (ED^x) to decrease noise created by small variations in the allelic frequency estimations (Figure 2B) and local linear regression (loess fit with a polynomial exponent of 1)(Cleveland 1988) of the EDs with a span automatically chosen by minimizing the corrected Akaike Information Criterion (AICc)(Hurvich et al. 1998) (Figure 2C).

Our method assumes that the Euclidean distance (ED) between allele frequencies in the mutant and WT pools decreases proportionally to the genetic distance between a given SNP and the causative mutation (see Figure 1B). Therefore, small EDs are a mixture of noise and signal from relatively distant markers. As the ED increases, it becomes increasingly likely that a given distance measurement is signal and less likely that it is noise, while simultaneously indicated stronger linkage. Therefore, we raise the allele frequency ED to a power to increase the effect of large ED measurements and decrease the effects of low ED measurements/noise (Figure 3B). The effect of raising an ED to increasingly greater powers minimizes the effects of increasing portions of the data. Consequently, raising the distance to too large of a power results in the shape of

the loess curve being dominated by a few outlying points. In order to determine the optimized power for our datasets, we ran MMAPPR on the *nkx2.5* RNA-seq data at various powers and measured the width of the peak at its base and the proximity of the top of the peak to the mutation. Our results show that a power of 4 resulted in the best fit (Figure 3D). A power of 4 also worked well on the three other mutants shown here.

A complimentary method for reducing the effects of noise is to fit a curve to the raw data. However, there is not a readily apparent model that the ED data is expected to fit, excluding the use of parametric methods. We also found that smoothing methods that utilized a fixed window were sensitive to skewing in regions with low expressed SNP density, so we chose to fit the data using loess regression. Loess regression is a non-parametric method for fitting curves based on a weighted-average of points within a given span. In linkage analysis, SNPs located near each other should have the same ED because they are genetically linked while SNPs located progressively farther apart are progressively less linked and thus the difference in their EDs is less restricted. Loess regression linearly fits the data using a set number of points while weighting by the distance between each of the points. Points close to each other, which all should be genetically linked together, are weighted heavily while points farther apart carry less weight. Using loess regression allows for and compensates for regions with low SNP density. In contrast, fixed window methods, which use a fixed number of points with equal weight, would result in windows that span several centimorgans in low SNP density regions, and fixed distance windows would have few points in low SNP density regions. Because MMAPPR is data driven, it also accounts for local recombination frequencies without requiring a previously generated genetic map (Bowen et al. 2012) or requiring one to assume that the recombination frequency is the same across the

genome, as required by HMM based models (Leshchiner et al. 2012). We allow the span to be chosen by the software using AICc, which is commonly used for optimizing loess fit curves (Hurvich et al. 1998). The resulting loess curve shows a very distinct peak (Figure 3C) at the shared segment that was not readily visible in the raw data (Figure 3A).

Candidate Selection

In addition to providing mapping information, RNA-seq datasets also contain mutation data for genes expressed at the developmental time-point when the tissue is collected. In order to identify putative causative mutations, MMAPPR selects SNPs within the identified peak(s) that have an ED above a threshold and have a high allele frequency in the mutant pool. Because of the noise inherent in the dataset, we set these cutoffs conservatively with a minimum distance of 0.5 and a minimum mutant pool allele frequency of 0.75. Identified SNPs are then analyzed using the Alleler program (part of the Useq package (Nix et al. 2008)), which uses annotated genes to determine whether a given mutation is non-synonymous. The result provides a list of putative mutations that are confined to the identified region and may impact the gene's protein product.

The Alleler program is not able to identify causative mutations that are in genomic sequences outside the RNA dataset, such as mutations in transcriptional regulatory regions, in genes that undergo nonsense-mediated decay or mutations that consist of large deleted regions. Therefore, in addition to this analysis we recommend that the datasets in hand be utilized for differential gene expression analysis using available software (for example, USeq, Cufflinks, Bioconductor) to identify genes in the identified region that have significantly different gene expression profiles.

Validation

We have verified MMAPPR using four mutants: two previously identified mutations, *nkx2.5* and *tbx1^{tm208}*, using RNA-seq from isolated heart tissue, and two unpublished mutations from an ENU screen performed in our lab, *zy13* and *zy14*, using RNA-seq from whole embryos. Results from these experiments are summarized in Table 1. First, we developed our method and optimized the method's parameters using the previously mapped *nkx2.5* line as shown in Figures 2B and 3. The final results for this line using the optimized parameters are shown in Figure 4A and Supplementary File 1. We next used these optimized parameters to analyze the previously mapped *tbx1^{tm208}* mutation (Figure 4B and Supplementary File 2). In both cases, MMAPPR identified a stop codon within 0.5 MB of the maximum fitted peak value corresponding with the previously identified mutations in each line. This shows that the optimized MMAPPR parameters were not specific to the *nkx2.5* line, but provided a strong starting point for novel mutation analysis.

We next used MMAPPR on two mutants, *zy13* and *zy14*, from an ENU screen for cardiovascular development mutants performed in our lab. The *zy13* mutant has severe pericardial edema and lacks melanophores and other migratory neural crest cells. Unlike the *nkx2.5* and *tbx1^{tm208}* lines, the phenotype of this mutant did not clearly indicate a dysfunction in a readily accessible tissue, which will likely be a common situation in the analysis of novel mutants. Therefore, we asked whether MMAPPR was able to identify a mutation using RNA-seq from whole embryos (Figure 5A and B) and compared this approach to the RAD-seq method (Baird et al. 2008) used by the commercial mapping service Floragenex (Figure 5C). Interestingly, both methods identified a large region spanning approximately half of chromosome 7, although the

peak found by MMAPPR was more informative because it had a maximum at approximately 67MB while the RAD-seq region was flat, without a peak.

In order to investigate possible causes for the large region on chromosome 7, we compared the published zebrafish genetic map (Bradley et al. 2011) to physical positions in the Zv9 genome assembly. In Figure 5D, lines between the two x-axes connect the positions of genetic mapping markers (in cM) and their corresponding genome assembly locations (in Mb). Crossing lines indicate locations where the two maps disagree on the genomic order of the markers. Differences between the relative distances between markers show localized differential recombination rates in the region. This analysis showed that this region falls very near the centromere (indicated by a black dot on the top and bottom x-axes), a region with very little recombination. There is also at least one pair of mapping markers that do not match the order in the physical genome assembly, indicating that there are possible errors in this region of the genomic build that may also affect the mapping results. Thus, it is likely that the large identified region was due to genomic characteristics that repress recombination around the centromere and inaccuracies in the genomic build.

Unlike the RADseq result shown in Figure 5C (which was relatively flat throughout the region), MMAPPR showed a discernible peak at approximately 67 MB. Candidate selection did not identify any stop codons in the region, but differential gene expression analysis of the RNA-seq data identified two genes near the peak maximum with greatly reduced expression levels in the mutant pool (Table 1). PCR-based DNA sequencing confirmed the presence of a nonsense mutation in the *ctr9* gene that appears to cause nonsense mediated decay, and the ability to rescue the mutant phenotype by injection of wild-type *ctr9* mRNA confirmed *ctr9* as the causative

mutation (Jurynek et al., submitted). These results show that MMAPPR generates an accurate estimate of the mutant location even in centromeric regions or genomic regions that are not perfectly represented in current genetic maps and genome assemblies.

The other mutant, *zy14*, has a complex phenotype, including failure to form intersomitic vessels and head vasculature. MMAPPR results for this line are shown in Figure 6. The identified region was 6.6 MB, similar to the regions identified for *nkx2.5* and *tbx1* and much smaller than the region identified for *zy13*. Several non-synonymous mutations segregating with the phenotype were identified within this region, but only a single nonsense mutation was found in the *cds2* gene (Table 1). Because MMAPPR works with RNA-seq libraries generated from 20 embryos, we analyzed these libraries for differential expression to see if the candidate was also identified. However, no genes with significantly different gene expression profiles were identified within the linked region using single sets of RNA-seq libraries. Since differential expression analysis using a single RNA-seq replicate can be limited because the variance cannot be accurately estimated (Anders and Huber 2010), we tested differential expression analysis using three separate biological replicates, each derived from 20 different embryos per mutant and phenotypically wild-type pool. This analysis showed that expression of *cds2* RNA containing the nonsense mutation is greatly reduced in the mutant pool, likely due to nonsense mediated decay, providing further evidence for its role in the observed phenotype. Based on this result and the optimization studies described above (Figure 2), in the limited cases where the differential expression analysis component of MMAPPR is required, we recommend utilizing at least three biological replicate RNA-seq libraries, as is common for differential expression analysis by RNA-seq, from pools each of ten to twenty wild-type and mutant embryos. Including at least ten individuals in each

replicate will yield 30 to 60 individuals for mapping by combining the datasets, which provides a number of individuals that is above the minimum required (Figure 2).

We first confirmed the putative nonsense mutation in *cds2* by Sanger sequencing mutants and their siblings to show that it segregated as expected in the population (Figure 6C and D). Next, we successfully rescued the mutant phenotype by injecting wild-type *cds2* RNA injection into one-cell stage embryos (Figure 6E-G). Mutant *cds2* mRNA, with the single base change, was unable to rescue the phenotype (Figure 6H). After we conducted our confirmation experiments, another group published a paper showing that a different mutation in the same gene gave a similar phenotype (Pan et al. 2012). Together, these experiments indicate that MMAPP correctly identified *cds2* as the causative mutant in the *zy14* line.

DISCUSSION

The MMAPP method described here is able to identify candidate mutations without any parental strain or genotype information, without previously identified SNP map databases, and without data from separate individuals. By using only single RNA-seq libraries from a small number of pooled mutant individuals and their phenotypically wild-type siblings, MMAPP requires fewer animals than is normally required for traditional mapping and less sequencing data than is required for whole genome sequence mapping. In addition, unlike whole genome sequencing, the same RNA-seq datasets can also be used for transcriptome analyses of mutants. Furthermore, MMAPP will allow identification of X-linked mutations as mutant allele frequency in the mutant pool will be 1, while the mutant allele frequency in the wild-type pool will be 0.2, creating a Euclidean distance between populations that is actually greater than the

autosomal recessive case. A linked region can also be identified for dominant mutations if they are fully penetrant using the Euclidean distance equation here, although modifications to the MMAPPR program would have to be made for candidate SNP identification. Together, these attributes make MMAPPR an efficient and cost effective means to identify spontaneous or induced mutations.

Using RNA-seq also allows candidate genes within the linked region to be identified by three different bioinformatics approaches. First, SNPs can be analyzed by their effect on the protein. We have integrated this analysis into MMAPPR because it directly uses the data generated for the linkage analysis. Second, gene expression analysis can be done using a number of available tools (USeq, Cufflinks, Bioconductor) to identify putative mutations that reduce mRNA levels, due either to mutations in the coding regions that lead to non-sense mediated decay or to mutations in gene regulatory regions. Finally, differential splicing can be analyzed using several tools (USeq, SpliceGrapher, SpliceSeq, KISSPLICE). By finding the intersection between the lists generated by these tools and the MMAPPR identified region, one can generate a robust list of candidate mutations underlying the phenotype in question. The use of these lists will differ in a case-by-case basis.

MMAPPR can be used for a wide variety of model and non-model organisms. For any organism, the criteria are a moderately well assembled genome, a sufficient level of sequence polymorphism (typical of most model organisms that are not highly inbred), and a sufficient number (approximately 20) of F2 offspring that can be pooled by phenotype for RNA-seq. These F2 offspring do not have to be siblings, but can be generated from crosses of multiple F1 carrier siblings, as we did with the *nkx2.5* and *tbx1* lines. In organisms that have a genome assembly but not a strong transcriptome

annotation, tools are available to build a transcriptome from RNA-seq (Grabherr et al. 2011).

Although MMAPPR worked well for the four examples shown here, it has several limitations. As currently implemented, it is unable to directly identify indels that are small enough to not affect overall gene expression levels. MMAPPR is capable of mapping larger deletions (data not shown). It is unable to identify genes that are missing from the reference build or are incorrectly annotated. Finally, it is unable to directly identify the causative lesion if the pooled samples are collected after the gene is no longer expressed or the mutation lies in untranscribed genomic regions. Nonetheless, it is important to note that in each of these cases MMAPPR *will* identify the genomic region containing the lesion and in some cases the affected gene can be identified using differential gene expression or splicing analysis, as described above. Any RNA-seq based mapping method also cannot be used in cases where the ability to isolate RNA for library construction is destroyed by tissue fixation or other processes necessary to identify the mutant phenotype.

We suggest several experimental design decisions to help increase one's odds of success. First, RNA should be extracted as soon as possible after onset of the phenotype to increase the likelihood that the causative gene is expressed and captured in the RNA-seq libraries. Second, we found that MMAPPR works with RNA isolated from whole animals or from tissue. Of note, the peaks for *tbx1* and *zy14*, which both fall near each other on Chromosome 5, were similar in size, even though *Tbx1* was identified from isolated tissue RNA-Seq and *zy14* was identified from whole embryo RNA-seq. However, we recommend when feasible that RNA isolated from the cells or tissues of interest be used for RNA-seq. In this case, MMAPPR will provide a smaller candidate

pool because it only identifies candidates in the genomic region that are expressed at the right time in the right tissue. Here, the *nkx2.5* and *tbx1* datasets were derived from hearts dissected from embryonic zebrafish. This tissue still provided a sufficient number of expressed SNPs for accurate mapping and, concurrently, a very small list of candidates within the mapped region (Table 1). Finally, increasing the read depth will increase the likelihood of detecting causative mutations in low expressed genes. It is commonly recommended to use at least three replicates for accurate differential gene analysis. Based on our results, three biological replicates containing at least 10 to 20 individuals each should be sufficient for this analysis. These replicates can be barcoded and multiplexed to reduce sequencing costs.

In conclusion, MMAPPR is a robust and effective way to map mutations generated from forward genetic screens or that spontaneously arise in a population. Adoption of this method might remove many of the barriers for researchers hesitant to conduct large-scale mutagenesis screens due to the labor-intensive and tedious mapping process.

MATERIALS AND METHODS

Animal care and mating

All fish were kept in the University of Utah Centralized Zebrafish Animal Resource facility or in the Yost Lab Zebrafish facility according to IACUC-approved protocols. For the *nkx2.5* and *tbx1^{tm208}* line, fish were received from the Yelon lab and the Trede lab, respectively, and mated into the *cmhc2:GFP* line maintained on an AB background. Offspring carrying both the *cmhc2:GFP* and the appropriate mutation were selected and mated. The *zy13* and *zy14* lines were identified from a standard F2

mutagenesis screen carried out in the AB strain. Mutant lines were subsequently maintained by outcrossing to the Wik zebrafish line (*zy13*, *zy14*) and the Tg(*fli1*-EGFP) and Tg(*kdr1*-EGFP) lines (*zy14*)(Lawson and Weinstein 2002; Beis et al. 2005).

RNA collection and sequencing

Offspring from zebrafish matings were raised to 30 hpf (*zy13* and *zy14*), 48 hpf (*nkx2.5*) or 72 hpf (*tbx1^{tm208}*) under standard conditions. These time points represent the earliest stage where we could confidently identify the phenotypes. Embryos were then segregated into mutant and phenotypically wild-type groups based on morphological phenotype as follows: *Nkx2.5*—enlarged atrium and diminished ventricle; *Tbx1*—loss of heart looping; *zy13*—pericardial edema; *zy14*—loss of intersegmental vessels. Pools of 20 whole embryos were collected for the *zy13* and *zy14* lines. For the other two lines, approximately 500 hearts were isolated as previously described (Geoffrey Burns and MacRae 2006) and placed in TRIzol. Both whole embryos and isolated hearts were processed using TRIzol extraction followed by the Qiagen RNeasy Mini kit (Qiagen). Isolated RNA was run on a Bioanalyzer 2100 Pico Chip (Agilent) to confirm RNA quantity and quality and then used to generate cDNA libraries as previously published (Christodoulou et al. 2011) at the Harvard Biopolymers Facility or using the Illumina Truseq kit (Illumina) at the University of Utah Microarray and Genomic Analysis Shared Resource. Samples were barcoded as WT vs. Mutant pairs (2 barcodes per lane) and single end 50 bp reads were generated on a Hiseq 2000 machine at the University of Utah Microarray and Genomic Analysis Shared Resource. Followed by processing using the Cassava 1.6 pipeline. Mapping was done using Novoalign (Novocraft) with default parameters except output was set to Sam format and FASTQ scoring to ILMFQ. Mapping was done using the Zv9 zebrafish build with splice

junctions derived from the UCSC Refseq refflat gene table. Data on the number of reads obtained from the RNA-seq datasets is summarized in Table 1.

Data Processing

A software implementation of MMAPPR was created using a combination of Python 3 and R and is available at <http://yost.genetics.utah.edu/software>. Briefly, the software package performs the following steps. First, Bam files are passed through the mpileup tool in the Samtools package (Heng Li et al. 2009) to create a pileup file. The pileup format converts the file to a position-based format showing the bases sequenced at each position. Reads at each position are filtered by the minimum base quality and minimum mapping quality set by the user and then the frequency of each allele is calculated. This data is subsequently passed to R for signal processing and peak identification. First, the Euclidean distance is calculated at each SNP location using the equation:

$$ED = \sqrt{(A_{mut} - A_{wt})^2 + (C_{mut} - C_{wt})^2 + (G_{mut} - G_{wt})^2 + (T_{mut} - T_{wt})^2}$$

where the letters (A, C, G, T) represent their corresponding bases. This distance is then raised to a power set by the user. Next, the data is fit using a Loess curve with a polynomial exponent of 1 and a span parameter determined by minimizing the AICc. Peak regions are defined as regions where the loess fitted values are greater than 3 standard deviations above the genome-wide median. R then plots the loess fits and returns a list of SNPs within the identified region(s) that are enriched in the mutant pool (an allele frequency > 0.75 and a Euclidean distance > 0.5). The identified SNPs are then passed to the Alleler program (part of the USeq package) to filter for non-synonymous SNPs using the provided gene annotation. The position and effect of these SNPs is finally exported to an output file. MMAPPR uses the optimized values reported

here as defaults, but allows the following variables to be modified by the user: mapping quality (default = 30), base quality (default = 20), minimum read depth (default = 10), power that ED is raised to (default = 4) and whether repetitive regions are masked (default = not masked).

Causative allele confirmation

Ctr9 was chosen from the list of *zy13* candidates based on its position relative to the peak of the mapped region and its known function and expression pattern. Because it was identified as a mutation that might lead to nonsense mediated decay, it was first sequenced to identify a G-A mutation resulting in a nonsense mutation at amino acid 580. Functional significance of the mutation was confirmed by phenotypic rescue using wild-type *ctr9* RNA (Jurynek et al., submitted). *Cds2* was chosen as the primary candidate for the *zy14* line based on both SNP and expression analysis (see results). The mutation was confirmed by Sanger sequencing of a 327 bp genomic DNA fragment amplified using primers flanking the mutation (Fwd: TGCAGACTTCTTTGCAAGTAAAC and Rev: TTTGGACACCCCTGCTTTAT). For RNA rescue confirmation, full-length wild-type and mutant cDNAs were amplified by PCR (Fwd: CCAGGCCTCTATTTTCACCA and Rev: CCTGGTGGTCCCAGAAGTTA) and inserted into pCS2+. Capped RNAs were synthesized with the mMessage Machine SP6 transcription kit (Ambion). Embryos derived from matings of *zy14*^{+/-};Tg (kdr1-EGFP)^{+/-} double heterozygous parents were injected at the 1-2 cell stage with 50 pg of wild-type or mutant *cds2* RNA and scored under epifluorescent illumination at 40 hpf for rescue of the mutant intersegmental vessel phenotype. As rescued embryos were phenotypically indistinguishable from their wild-type siblings, mutant genotypes were confirmed by Sanger sequencing.

Differential gene expression analysis

Gene expression analysis was done using the Useq 7.8.1 software package (Nix et al. 2008). A single RNA-seq replicate was used for the *nkx2.5*, *tbx1^{tm208}* and *zy13* lines. For the *zy14* line, both 3 biological replicates and 1 replicate were run for comparison. Differentially expressed genes were defined as genes with a false discovery rate less than 0.05.

Genetic Distance Estimation

Genetic distances between peak maxima and causative mutations were estimated using the SNP/str combined genetic map and the Zv9 build (Bradley et al. 2011). Two genetic markers surrounding the region and their corresponding Zv9 location were first selected and used to create a linear model of genetic vs. physical distance between the markers. This model was then used to estimate the genetic positions of each lesion and the corresponding peak maximum to calculate the genetic distance between them. Centromere positions were determined using data from previously published genetic maps (M A Mohideen et al. 2000; Shimoda et al. 1999).

DATA ACCESS

Raw Fastq files and aligned BAM files are available at the NHLBI Bench-to-Bassinet Consortium data-sharing hub (<https://b2b.hci.utah.edu:443/gnomex/gnomexFlex.jsp?topicNumber=27>). A software package for MMAPP analysis and instructions for its use can be accessed from <http://yost.genetics.utah.edu/software.php>.

ACKNOWLEDGEMENTS

We would like to thank D. Yelon and N. Trede for providing us with the *nkx2.5* and *tbx1* fish lines, respectively, and D. Nix, A. Holloway, N. Trede and M. Condic for their helpful insights on the manuscript. This study was funded by NHLBI Bench-to-Bassinet Consortium (<http://www.benchtobassinet.com>) grant to HJY (U01HL0981) and a core facilities support grant to New England Research Institute (U01 HL098188).

AUTHOR CONTRIBUTIONS

J.H. designed and implemented MMAPPR, analyzed data and wrote the manuscript. B.D. assisted in designing and coding the pipeline and analyzed data. B.B. collected and processed the RNA from the *zy13* and *zy14* lines and performed experiments to confirm the putative candidates. B.G. and J.H. collected and processed the RNA from the *Nkx2.5* and *Tbx1* lines. Y.C.S performed experiments to confirm the *zy14* candidate. H.J.Y supervised and guided the project and assisted with manuscript preparation.

DISCLOSURE DECLARATION

The authors do not have any conflicts of interest to disclose.

FIGURE LEGENDS

Figure 1: Schematic overview of MMAPPR. A) Mating scheme. Fish from the mutant line are outcrossed and then F1 progeny, heterozygous for the mutation, are crossed. Pools of phenotypically mutant and phenotypically wild-type fish are then sorted. B) Schematic representation of allelic segregation between wild-type and mutant pools. Black indicates genomic regions inherited from the mutant carrier line and grey

represents regions inherited from the outcross line. The bottom panel shows a plot of the expected Euclidean distances calculated from the allele frequencies in the two pools. C) Flow chart of the analysis steps incorporated in the MMAPPR algorithm. Each phase of the pipeline is shown by grey boxes and the portion of the pipeline processed by the MMAPPR software package is shown by a vertical grey bar.

Figure 2: Effect of various parameters on peak width, height and location. Reads were randomly sampled from the combined *zy14* dataset to create twenty datasets with 1% to 100% of the original data. (A) shows the width of the peak in each dataset. (B) shows the height of the peak for the same datasets. The 1% dataset did not have a sufficient number of SNPs to identify a peak (X in A and B). All other datasets are shown by diamonds. (C) shows the effect of different mapping quality scores on peak location (distance of the peak maximum from the identified mutation). Mapping quality had little effect on peak position, but the best overall position for all datasets was at a phred scaled mapping quality score threshold of 30.

Figure 3: Data progression through noise reduction steps. A) Raw Euclidean distance scores across the genome. For all panels, vertical grey lines delineate chromosome edges and chromosome widths represent the relative number of SNPs on the chromosome. B) Euclidean distance raised to the 4th power. C) Loess fit curve calculated using the data shown in B. D) Effect of raising the Euclidean distance to different powers on the Loess fit curve. All data are from the *nkx2.5* mutant and phenotypically wild-type RNA-seq datasets. Red asterisks mark the mutation location.

Figure 4: Linked region peaks identified by MMAPPR for the *nkx2.5* and *tbx1^{tm208}* lines. A) Genome-wide (top panel) and Chromosome 14 (bottom panel) loess fit curves for SNP allele frequency Euclidean distance raised to the 4th power for the *nkx2.5* line. B) Genome-wide (top panel) and Chromosome 5 (bottom panel) loess fit curves for SNP allele frequency Euclidean distance raised to the 4th power for the *tbx1* line. Asterisks mark the mutation location. Centromeres are marked with a large black dot on the x-axis.

Figure 5: Results for MMAPPR mapping of the *zy13* line. A-B) Genome-wide (A) and Chromosome 7 (B) loess fit results. Location of the mutation shown by an asterisk. C) RADseq mapping results for Chromosome 7. D) Location of SNP and microsatellite genetic map markers on the zv9 genome-build. Lines connect the genetic map position (bottom axis) to the physical map position (top axis) of each marker. Centromere location is marked in each panel with a black dot on the x-axis.

Figure 6: MMAPPR results for *zy14*. A-B) Genome-wide (A) and Chromosome 5 (B) loess fit results from mapping the *zy14* line. Centromere location is marked with a large black dot and location of the mutation is shown by an asterisk in B. C and D) Sanger sequencing traces from a wild-type sibling and a *zy14* mutant, respectively. Arrow indicates single base changed in the mutant, asterisk and underline indicates the resulting stop codon. E-H) Fluorescent images of the vascular phenotype visualized using the Tg(*kdr1*-EGFP) line at 40hpf, which marks vasculature. Panels are as follows: E) wild-type *zy14* sibling F) *zy14* mutant G) *zy14* mutant rescued by injection of wild-

type *cds2* mRNA, H) failure of *cds2* mutant mRNA injection to rescue *zy14* mutant embryo. In three experiments, uninjected embryos from heterozygote crosses had wild-type phenotypes in 39/51, whereas 109/109 embryos injected with wild-type *cds2* mRNA had wild-type phenotype, indicative of rescue. In contrast, mutant *cds2* mRNA with the single-base change seen in *zy14* was not able to rescue: 18/26 injected embryos had wild-type phenotype, compared to 14/20 uninjected siblings.

REFERENCES

- Aanes H, Winata CL, Lin CH, Chen JP, Srinivasan KG, Lee SGP, Lim AYM, Hajan HS, Collas P, Bourque G, et al. 2011. Zebrafish mRNA sequencing deciphers novelties in transcriptome dynamics during maternal to zygotic transition. *Genome research* 1328–1338.
- Anders S, and Huber W. 2010. Differential expression analysis for sequence count data. *Genome biology* 11: R106.
- Arnold CN, Xia Y, Lin P, Ross C, Schwander M, Smart NG, Müller U, and Beutler B. 2011. Rapid identification of a disease allele in mouse through whole genome sequencing and bulk segregation analysis. *Genetics* 187: 633–41.
- Austin RS, Vidaurre D, Stamatiou G, Breit R, Provart NJ, Bonetta D, Zhang J, Fung P, Gong Y, Wang PW, et al. 2011. Next-generation mapping of Arabidopsis genes. *The Plant journal* □ : for cell and molecular biology 67: 715–25.
- Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA, Selker EU, Cresko WA, and Johnson EA. 2008. Rapid SNP discovery and genetic mapping using sequenced RAD markers. ed. J.C. Fay. *PloS one* 3: e3376.
- Beis D, Bartman T, Jin S-W, Scott IC, D’Amico L a, Ober E a, Verkade H, Frantsve J, Field H a, Wehman A, et al. 2005. Genetic and cellular analyses of zebrafish atrioventricular cushion and valve development. *Development (Cambridge, England)* 132: 4193–204.
- Bowen ME, Henke K, Siegfried KR, Warman ML, and Harris MP. 2012. Efficient mapping and cloning of mutations in zebrafish by low-coverage whole-genome sequencing. *Genetics* 190: 1017–24.

- Bradley KM, Breyer JP, Melville DB, Broman KW, Knapik Ela W, and Smith JR. 2011. An SNP-Based Linkage Map for Zebrafish Reveals Sex Determination Loci. *G3 (Bethesda, Md.)* **1**: 3–9.
- Brenner S. 1974. The genetics of *Caenorhabditis elegans*. *Genetics* 71–94.
- Chen JN, van Bebber F, Goldstein AM, Serluca FC, Jackson D, Childs S, Serbedzija G, Warren KS, Mably JD, Lindahl P, et al. 2001. Genetic steps to organ laterality in zebrafish. *Comparative and functional genomics* **2**: 60–8.
- Christodoulou DC, Gorham JM, Herman DS, and Seidman JG. 2011. Construction of normalized RNA-seq libraries for next-generation sequencing using the crab duplex-specific nuclease. *Current protocols in molecular biology / edited by Frederick M. Ausubel ... [et al.]* **Chapter 4**: Unit4.12.
- Cleveland W. 1988. Locally weighted regression: an approach to regression analysis by local fitting. *Journal of the American Statistical Association* **83**: 596–610.
- Cuperus JT, Montgomery TA, Fahlgren N, Burke RT, Townsend T, Sullivan CM, and Carrington JC. 2010. Identification of MIR390a precursor processing-defective mutants in *Arabidopsis* by direct genome sequencing. *Proceedings of the National Academy of Sciences of the United States of America* **107**: 466–71.
- Doitsidou M, Poole RJ, Sarin S, Bigelow H, and Hobert O. 2010. *C. elegans* mutant identification with a one-step whole-genome-sequencing and SNP mapping strategy. *PLoS one* **5**: e15435.
- Driever W., Solnica-Krezel L, Schier AF, Neuhauss SC, Malicki J, Stemple DL, Stainier DY, Zwartkruis F, Abdelilah S, Rangini Z, et al. 1996. A genetic screen for mutations affecting embryogenesis in zebrafish. *Development* **123**: 37–46.
- García-García MJ, Eggenschwiler JT, Caspary T, Alcorn HL, Wyler MR, Huangfu D, Rakeman AS, Lee JD, Feinberg EH, Timmer JR, et al. 2005. Analysis of mouse embryonic patterning and morphogenesis by forward genetics. *Proceedings of the National Academy of Sciences of the United States of America* **102**: 5913–9.
- Geoffrey Burns C, and MacRae C. 2006. Purification of hearts from zebrafish embryos. *BioTechniques* **40**: 278–282.
- Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, et al. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature biotechnology* **29**: 644–52.
- Haffter P, Granato M, Brand M, Mullins M, Hammerschmidt M, Kane D, Odenthal J, van Eeden F, Jiang Y, Heisenberg C, et al. 1996. The identification of genes with unique and essential functions in the development of the zebrafish, *Danio rerio*. *Development* **123**: 1–36.

- Hughes SE, Huang C, and Kornfeld K. 2011. Identification of mutations that delay somatic or reproductive aging of *Caenorhabditis elegans*. *Genetics* **189**: 341–56.
- Hurvich CM, Simonoff JS, and Tsai C-L. 1998. Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **60**: 271–293.
- Lawson ND, and Weinstein BM. 2002. In vivo imaging of embryonic vascular development using transgenic zebrafish. *Developmental biology* **248**: 307–18.
- Leshchiner I, Alexa K, Kelsey P, Adzhubei I, Austin C, Cooney J, Anderson H, King M, Stottman R, Ha S, et al. 2012. Mutation mapping and identification by whole genome sequencing. *Genome Research*.
- Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, and Durbin R. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)* **25**: 2078–9.
- Lin X, Tang W, Ahmad S, Lu J, Colby CC, Zhu J, and Yu Q. 2012. Applications of targeted gene capture and next-generation sequencing technologies in studies of human deafness and other genetic disabilities. *Hearing research*.
- Liu S, Yeh C-T, Tang HM, Nettleton D, and Schnable PS. 2012. Gene mapping via bulked segregant RNA-Seq (BSR-Seq). ed. S.P. Hazen. *PloS one* **7**: e36406.
- Medina PMB, Swick LL, Andersen R, Blalock Z, and Brenman JE. 2006. A novel forward genetic screen for identifying mutations affecting larval neuronal dendrite development in *Drosophila melanogaster*. *Genetics* **172**: 2325–35.
- Mohideen MA, Moore J L, and Cheng K C. 2000. Centromere-linked microsatellite markers for linkage groups 3, 4, 6, 7, 13, and 20 of zebrafish (*Danio rerio*). *Genomics* **67**: 102–6.
- Moore Jessica L, Rush LM, Breneman C, Mohideen M-APK, and Cheng Keith C. 2006. Zebrafish genomic instability mutants and cancer susceptibility. *Genetics* **174**: 585–600.
- Nakamura K, Oshima T, Morimoto T, Ikeda S, Yoshikawa H, Shiwa Y, Ishikawa S, Linak MC, Hirai A, Takahashi H, et al. 2011. Sequence-specific error profile of Illumina sequencers. *Nucleic acids research* **39**: e90.
- Nix DA, Courdy SJ, and Boucher KM. 2008. Empirical methods for controlling false positives and estimating confidence in ChIP-Seq peaks. *BMC bioinformatics* **9**: 523.
- Nüsslein-Volhard C, and Wieschaus E. 1980. Mutations affecting segment number and polarity in *Drosophila*. *Nature* **287**: 795–801.

- Pan W, Pham VN, Stratman AN, Castranova D, Kamei M, Kidd KR, Lo BD, Shaw KM, Torres-Vazquez J, Mikelis CM, et al. 2012. CDP-diacylglycerol synthetase-controlled phosphoinositide availability limits VEGFA signaling and vascular morphogenesis. *Blood* **120**: 489–98.
- Piotrowski T, Ahn D, Schilling TF, Nair S, Ruvinsky I, Geisler R, Rauch G-J, Haffter Pascal, Zon LI, Zhou Y, et al. 2003. The zebrafish van gogh mutation disrupts *tbx1*, which is involved in the DiGeorge deletion syndrome in humans. *Development (Cambridge, England)* **130**: 5043–52.
- Rösel TD, Hung L-H, Medenbach J, Donde K, Starke S, Benes V, Rätsch G, and Bindereif A. 2011. RNA-Seq analysis in mutant zebrafish reveals role of U1C protein in alternative splicing regulation. *The EMBO journal* **30**: 1965–76.
- Schneeberger K, Ossowski S, Lanz C, Juul T, Petersen AH, Nielsen KL, Jørgensen J-E, Weigel D, and Andersen SU. 2009. SHOREmap: simultaneous mapping and mutation identification by deep sequencing. *Nature methods* **6**: 550–1.
- Shimoda N, Knapik E W, Ziniti J, Sim C, Yamada E, Kaplan S, Jackson D, de Sauvage F, Jacob H, and Fishman MC. 1999. Zebrafish genetic map with 2000 microsatellite markers. *Genomics* **58**: 219–32.
- Taberner MD, Maíllo A, Nieto AB, Díez-Tascón C, Lara M, Sousa P, Otero A, Castrillo A, Patino-Alonso MDC, Espinosa A, et al. 2012. Delineation of commonly deleted chromosomal regions in meningiomas by high-density single nucleotide polymorphism genotyping arrays. *Genes, chromosomes & cancer*.
- Uchida N, Sakamoto T, Kurata T, and Tasaka M. 2011. Identification of EMS-induced causal mutations in a non-reference *Arabidopsis thaliana* accession by whole genome sequencing. *Plant & cell physiology* **52**: 716–22.
- Vesterlund L, Jiao H, Unneberg P, Hovatta O, and Kere J. 2011. The zebrafish transcriptome during early development. *BMC developmental biology* **11**: 30.
- Voz ML, Coppieters W, Manfroid I, Baudhuin A, Von Berg V, Charlier C, Meyer D, Driever Wolfgang, Martial JA, and Peers B. 2012. Fast homozygosity mapping and identification of a zebrafish ENU-induced mutation by whole-genome sequencing. *PloS one* **7**: e34671.
- Warren LL, Li L, Nelson MR, Ehm MG, Shen J, Fraser DJ, Aponte JL, Nangle KL, Slater AJ, Woollard PM, et al. 2012. Deep Resequencing Unveils Genetic Architecture of ADIPOQ and Identifies a Novel Low-Frequency Variant Strongly Associated With Adiponectin Variation. *Diabetes* db11–0985–.
- Xiao T, Roeser T, Staub W, and Baier H. 2005. A GFP-based genetic screen reveals mutations that disrupt the architecture of the zebrafish retinotectal projection. *Development (Cambridge, England)* **132**: 2955–67.

Yu Q, Shen Y, Chatterjee B, Siegfried BH, Leatherbury L, Rosenthal J, Lucas JF, Wessels A, Spurney CF, Wu Y-J, et al. 2004. ENU induced mutations causing congenital cardiovascular anomalies. *Development (Cambridge, England)* **131**: 6211–23.

Zuryn S, Le Gras S, Jamet K, and Jarriault S. 2010. A strategy for direct mapping and identification of mutations by whole-genome sequencing. *Genetics* **186**: 427–30.

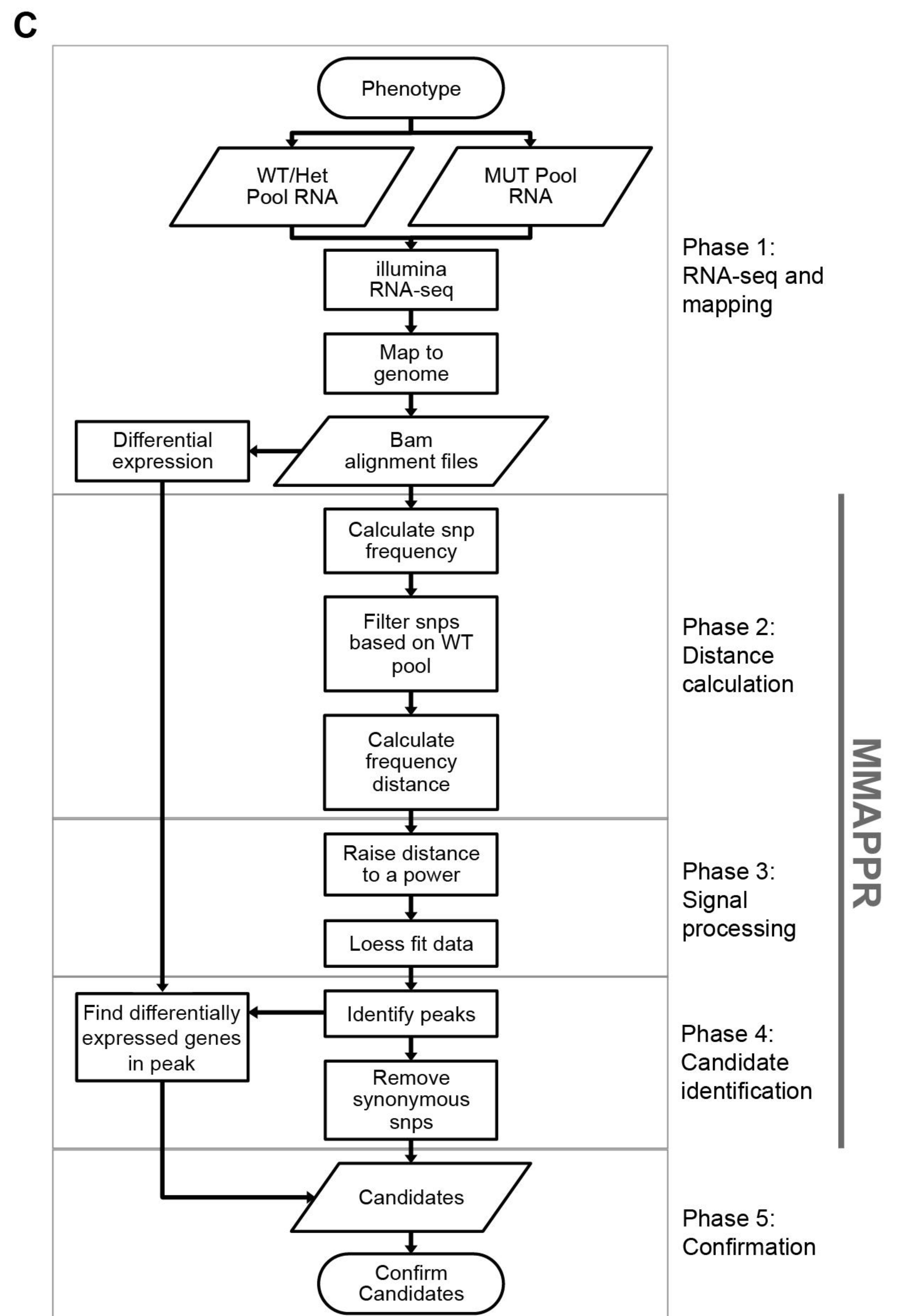
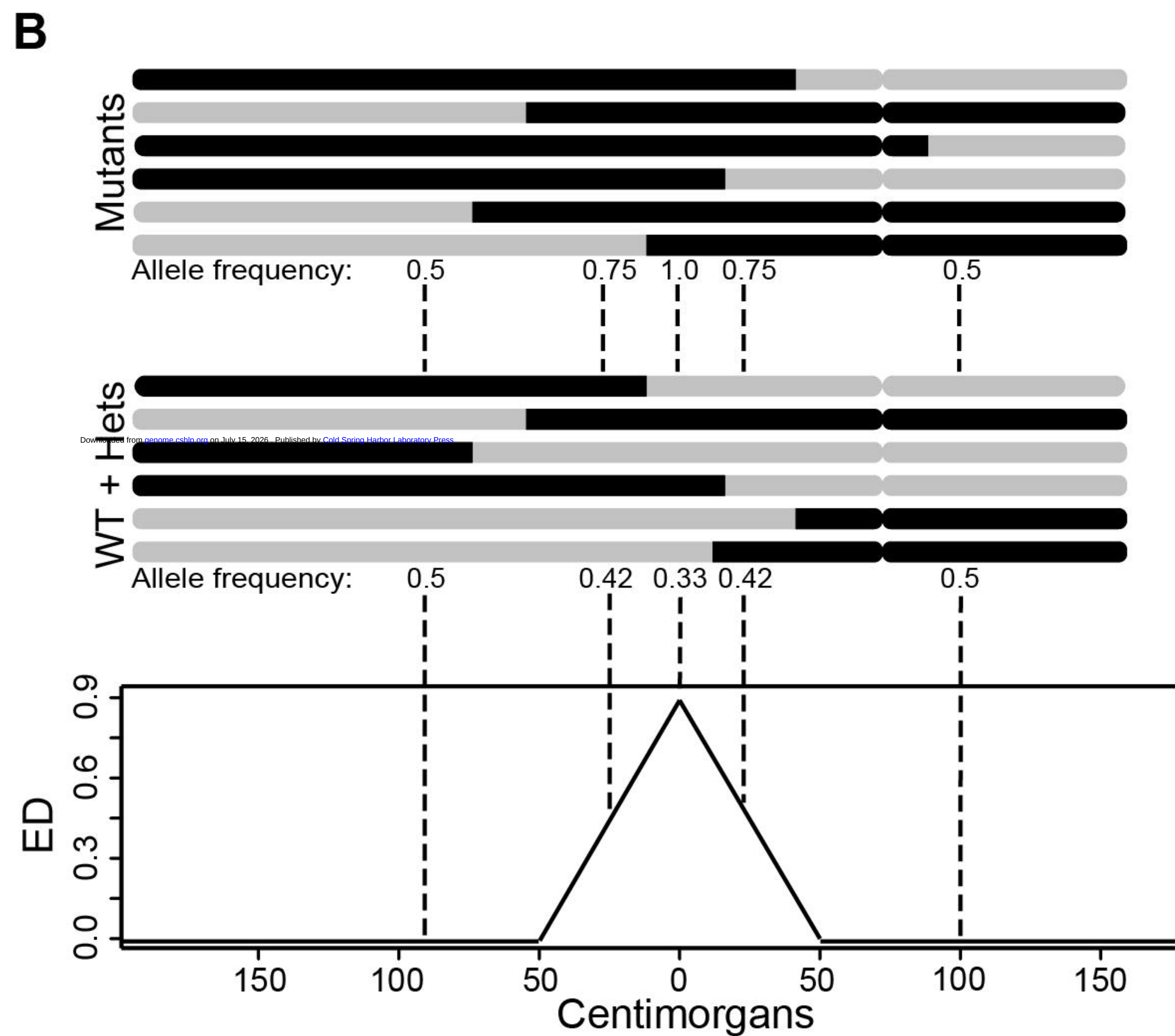
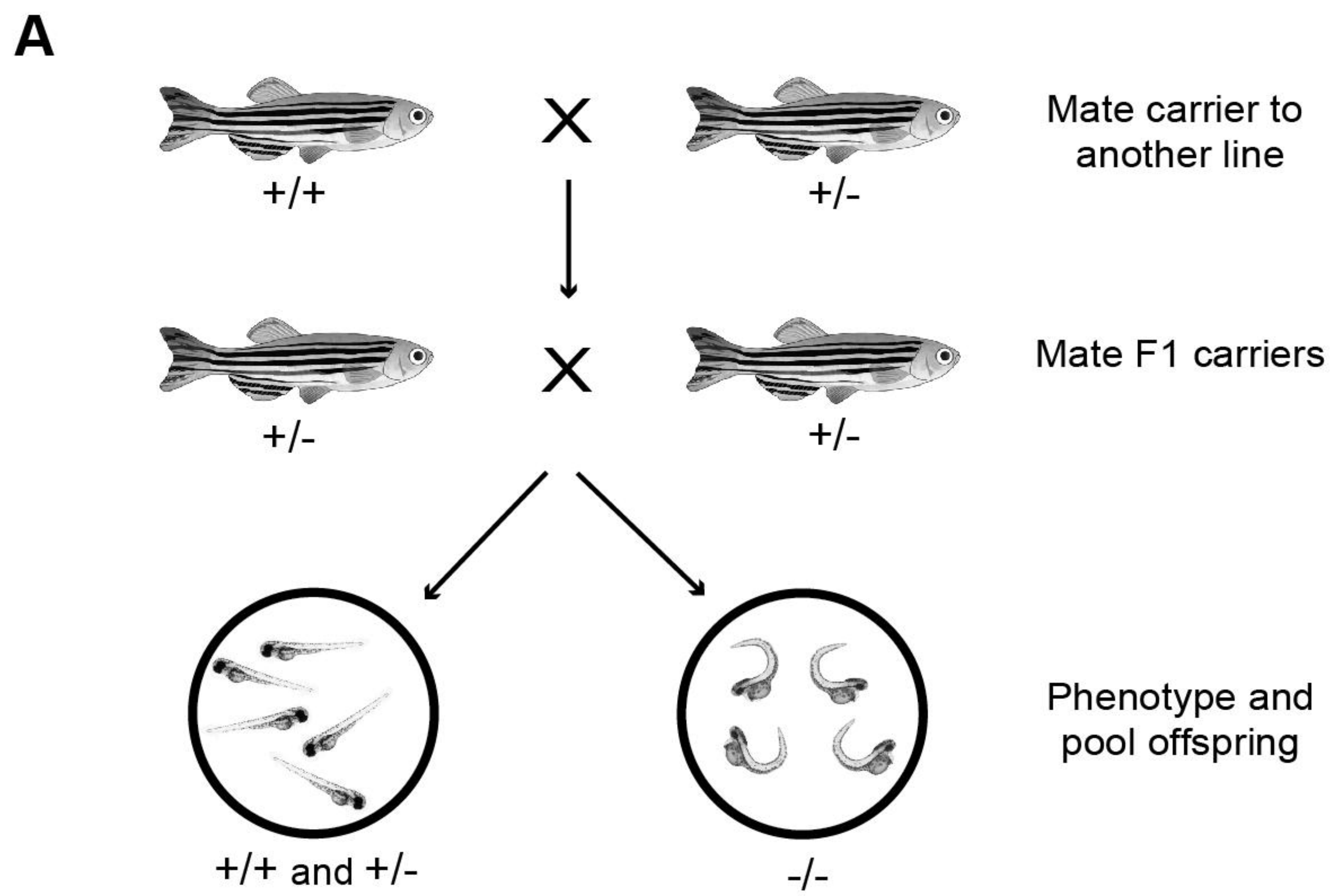
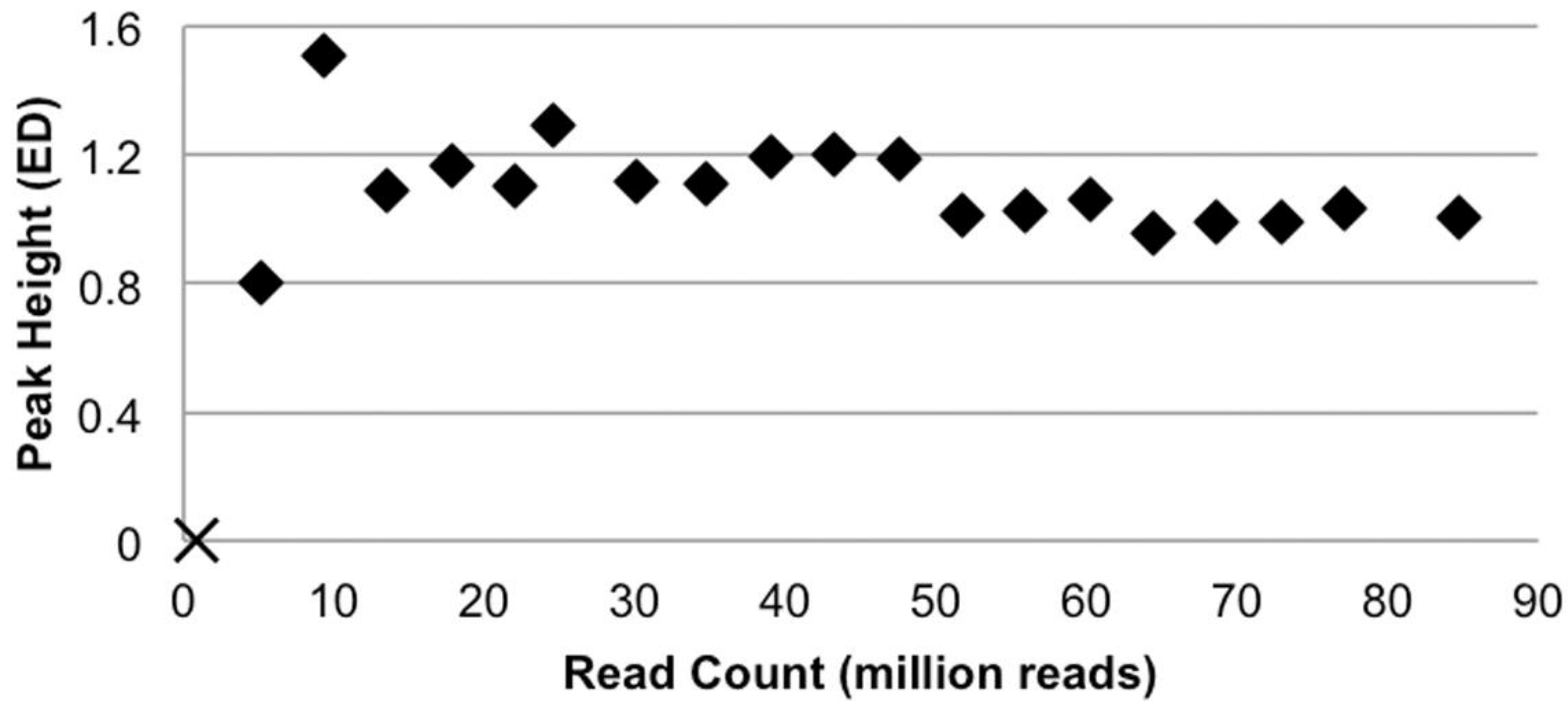
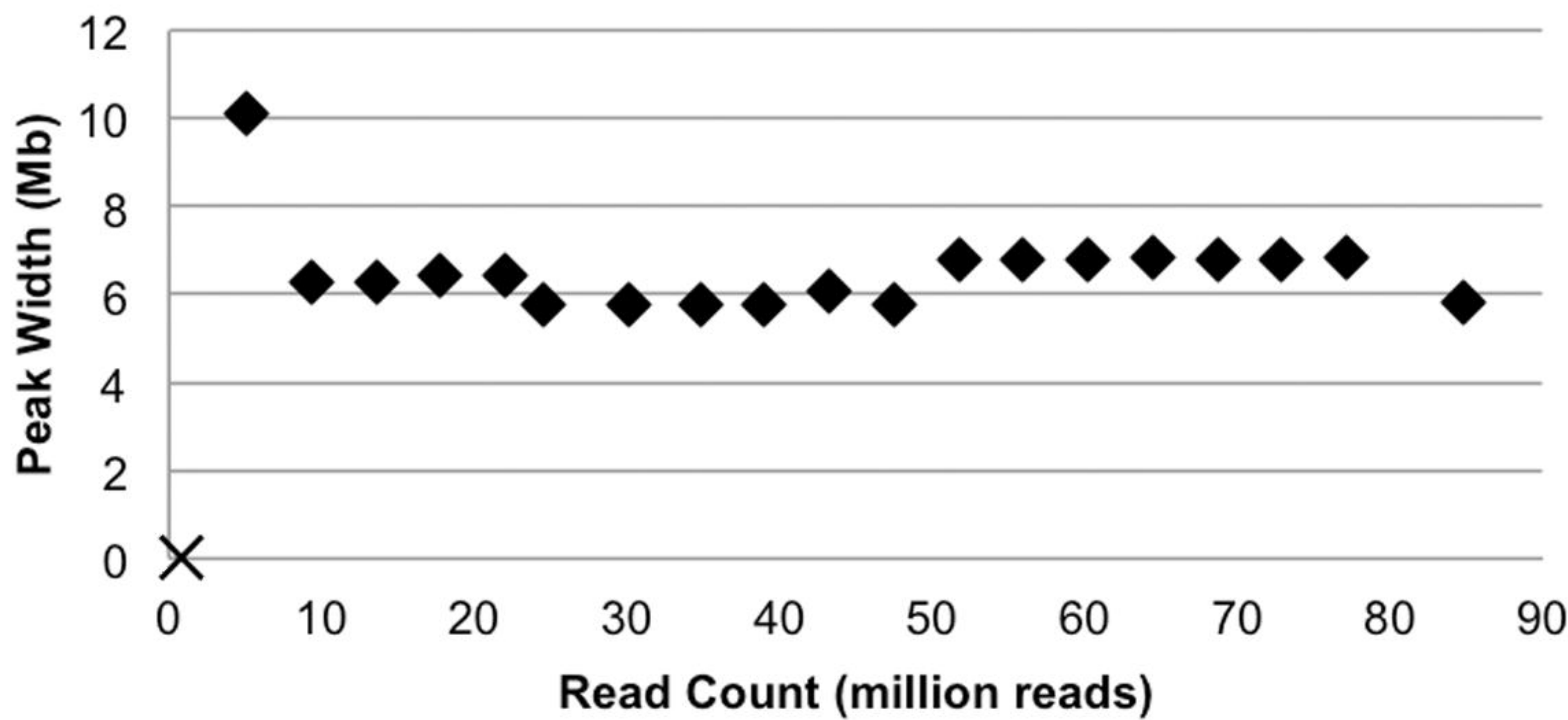


Figure-1

A



B



C

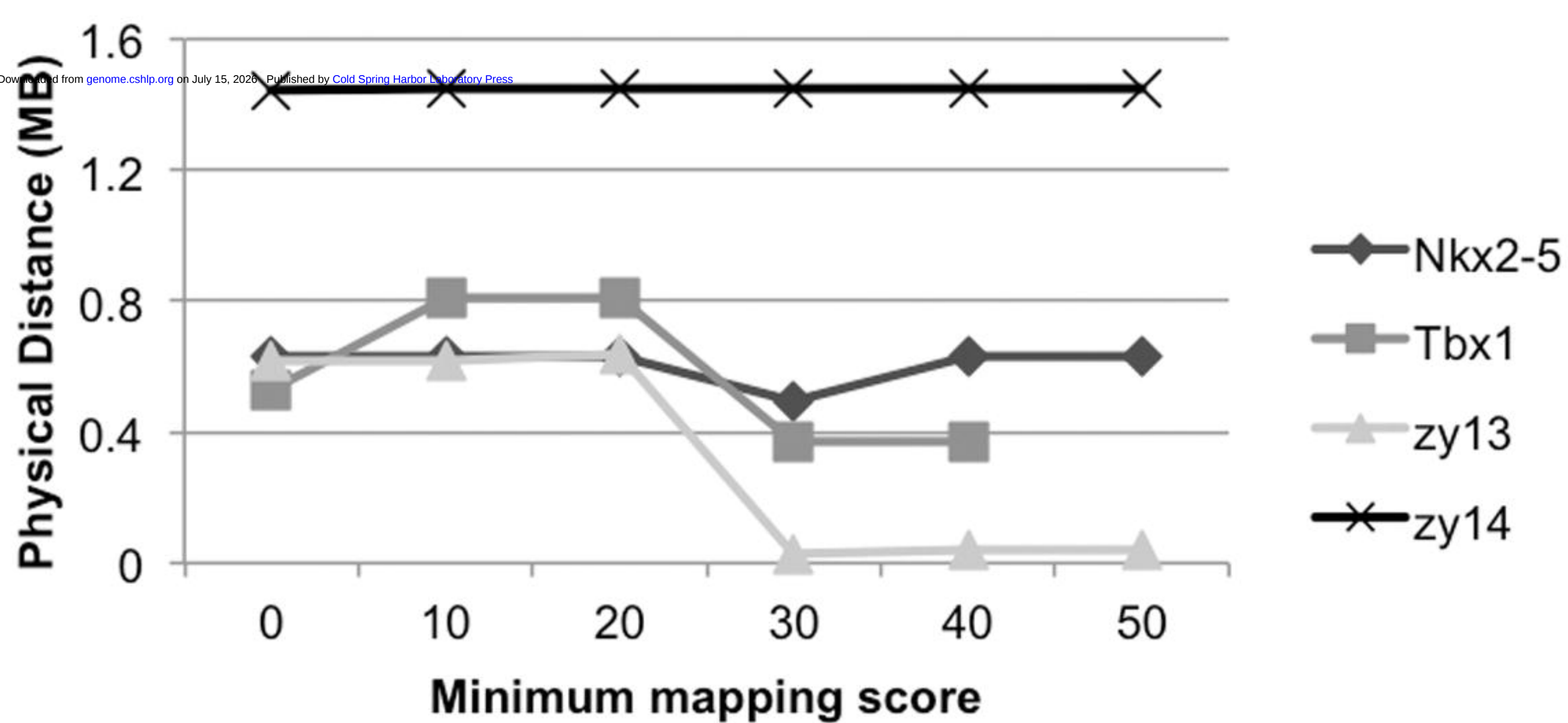


Figure-2

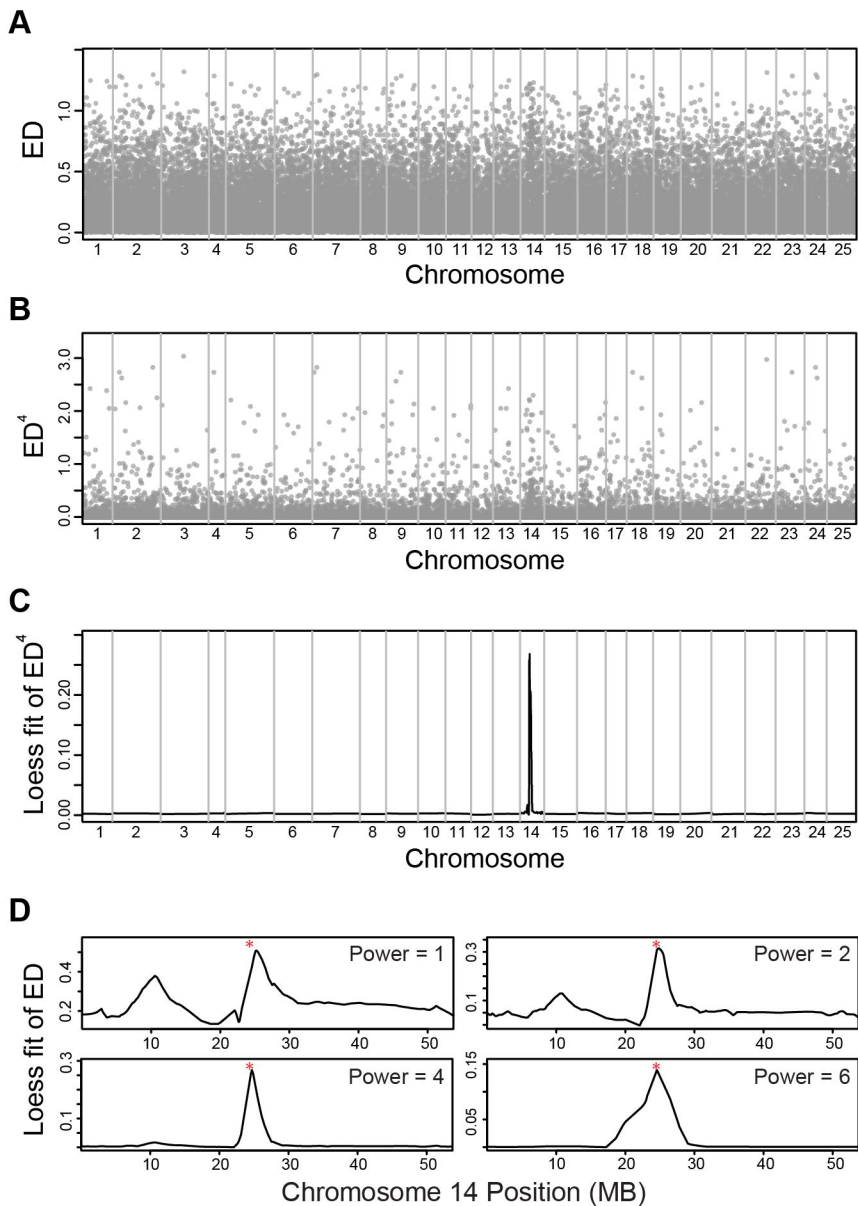


Figure-3

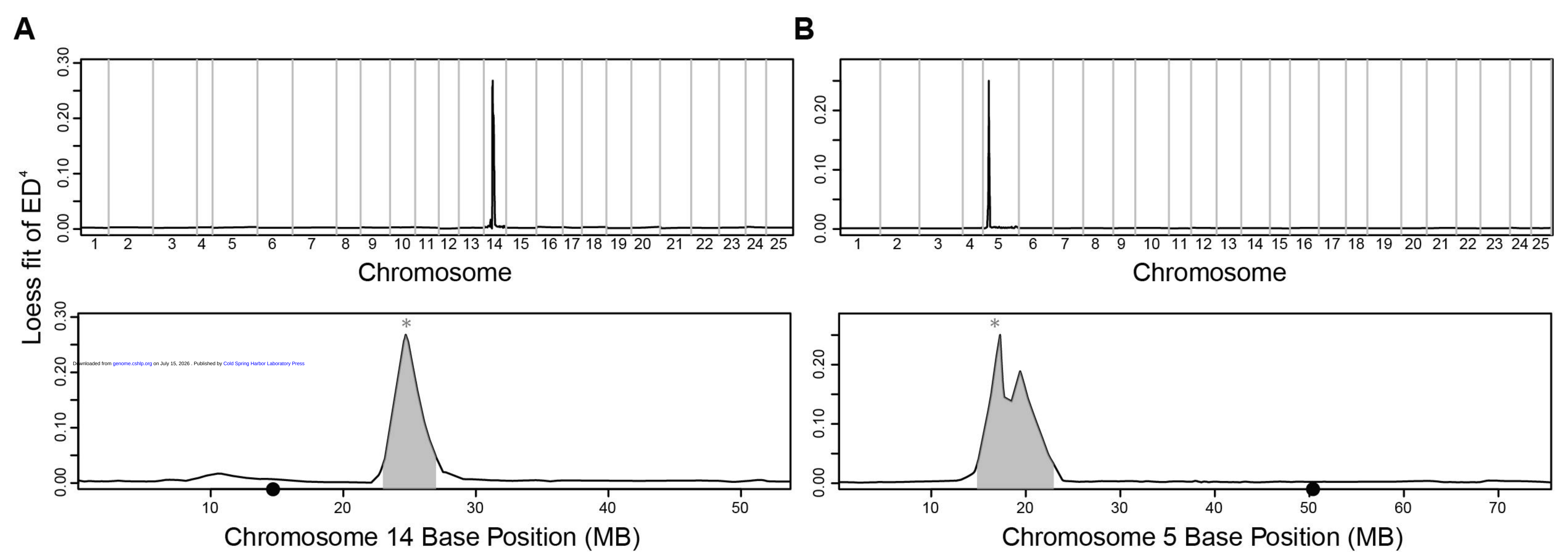
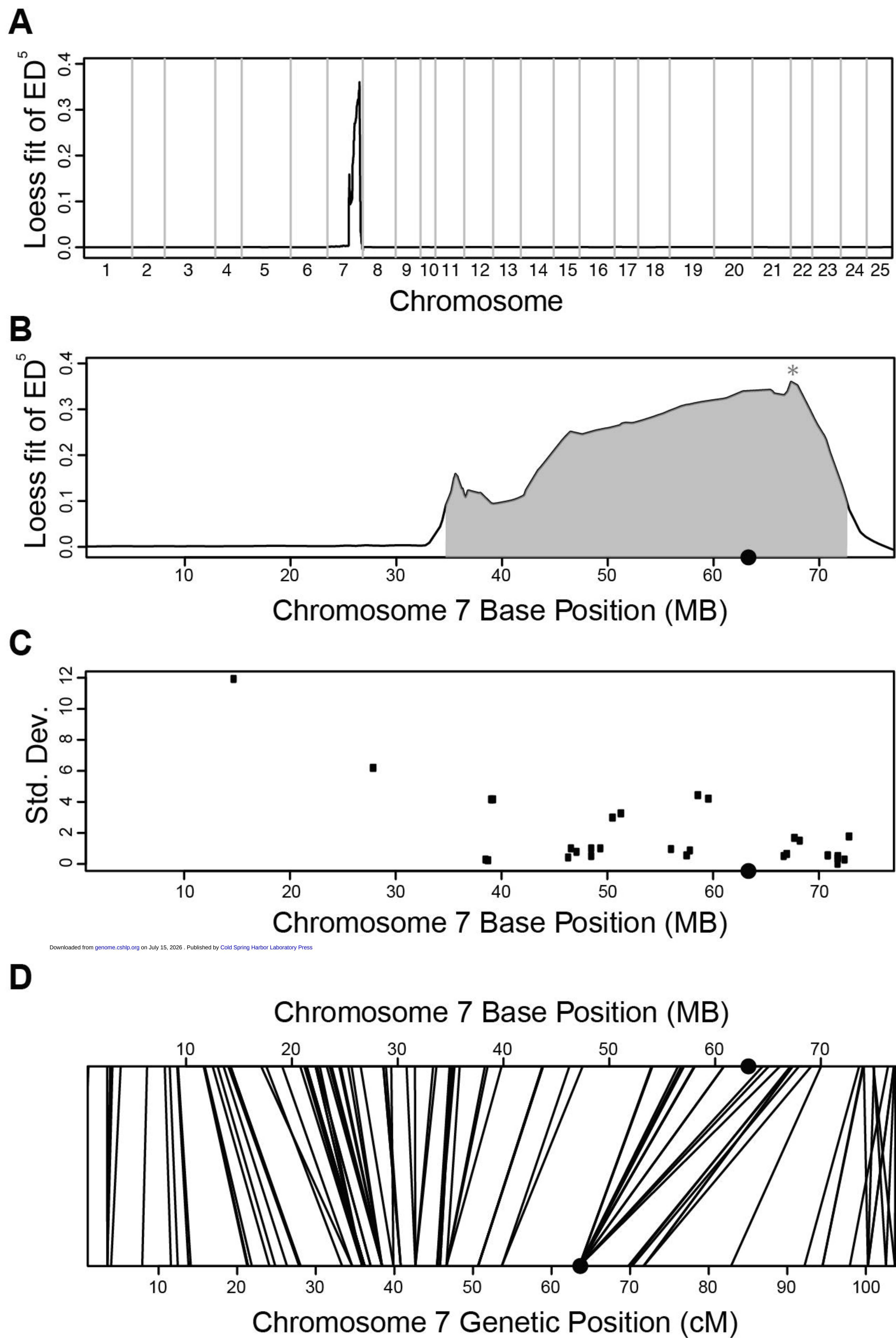


Figure-4



Downloaded from genome.cshlp.org on July 15, 2026 . Published by Cold Spring Harbor Laboratory Press

Figure-5

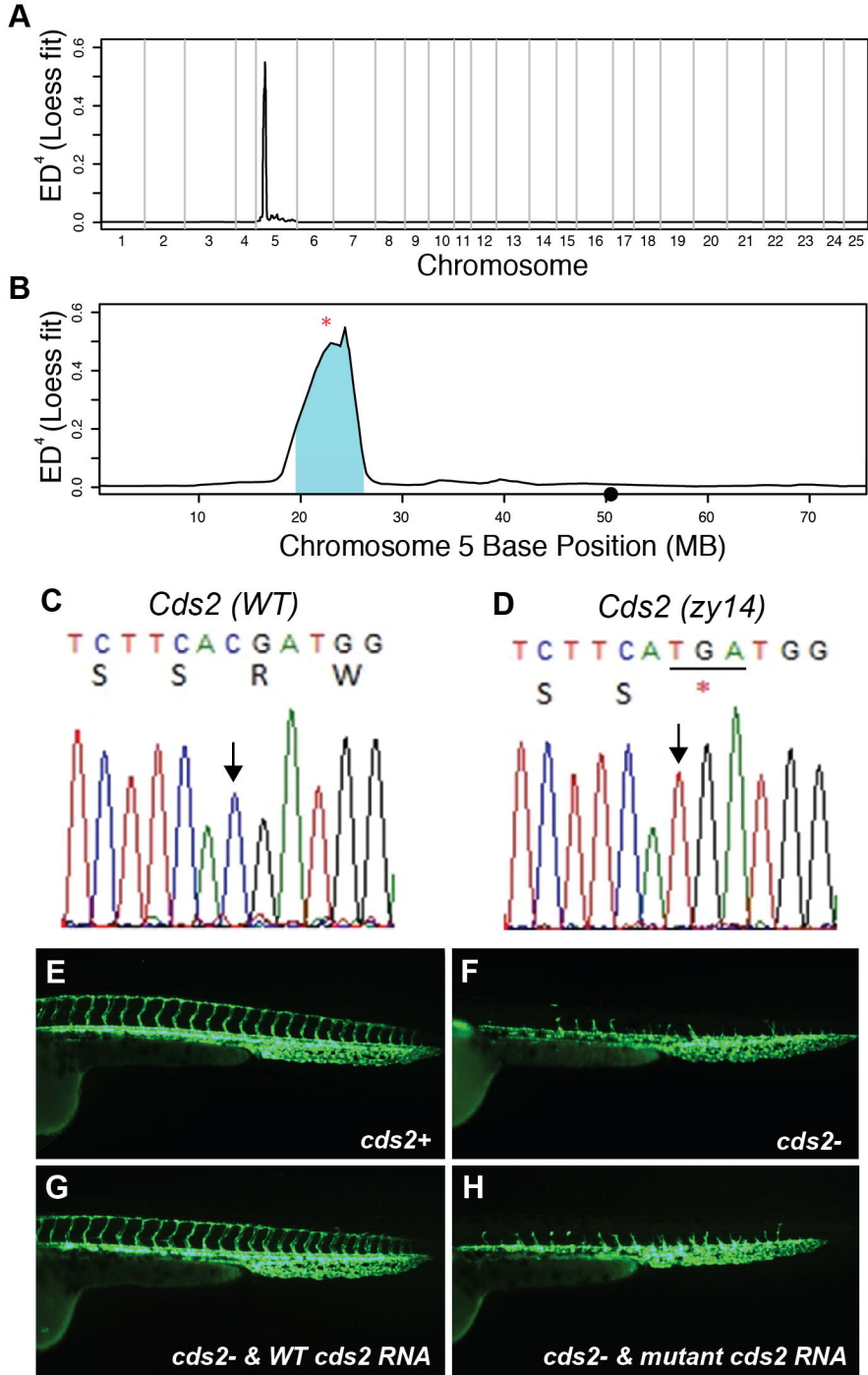


Figure-6

TABLE**Table 1: Summary of MMAPPR results.**

Line	Peak Width	Snps in Peak	Non-synonymous Snps	Genes in Peak	Genes Expressed	Diff. Exp. Genes	Confirmed Mutation	Distance from Peak max
<i>nkx2.5</i>	5.2 MB	181	1 (1 Nonsense)	35	15	0	NS in <i>nkx2.5</i>	0.5 MB (0.13 cM)
<i>tbx1^{tm208}</i>	8.1 MB	417	6 (1 Nonsense)	63	27	0	NS in <i>tbx1</i>	0.3 MB (0.22 cM)
<i>zy13</i>	37.9 MB	1680	22 (0 Nonsense)	299	116	4	NS in <i>ctr9</i>	0.04 MB (0.07 cM)
<i>zy14</i>	6.6 MB	518	19 (1 Nonsense)	111	45	2	NS in <i>cds2</i>	1.40 MB (0.98 cM)