Development 139, 4280-4290 (2012) doi:10.1242/dev.083931

Rapid positional cloning of zebrafish mutations by linkage and homozygosity mapping using whole-genome sequencing

Nikolaus Obholzer[1,*], Ian A. Swinburne[1,*], Evan Schwab[1], Alex V. Nechiporuk[3], Teresa Nicolson[2] and Sean G. Megason[1,‡]

SUMMARY

Forward genetic screens in zebrafish have identified >9000 mutants, many of which are potential disease models. Most mutants remain molecularly uncharacterized because of the high cost, time and labor investment required for positional cloning. These costs limit the benefit of previous genetic screens and discourage future screens. Drastic improvements in DNA sequencing technology could dramatically improve the efficiency of positional cloning in zebrafish and other model organisms, but the best strategy for cloning by sequencing has yet to be established. Using four zebrafish inner ear mutants, we developed and compared two approaches for 'cloning by sequencing': one based on bulk segregant linkage (BSFseq) and one based on homozygosity mapping (HMFseq). Using BSFseq we discovered that mutations in *lmx1b* and *jagged1b* cause abnormal ear morphogenesis. With HMFseq we validated that the disruption of *cdh23* abolishes the ear's sensory functions and identified a candidate lesion in *lhfpl5a* predicted to cause nonsyndromic deafness. The success of HMFseq shows that the high intrastrain polymorphism rate in zebrafish eliminates the need for time-consuming map crosses. Additionally, we analyzed diversity in zebrafish laboratory strains to find areas of elevated diversity and areas of fixed homozygosity, reinforcing recent findings that genome diversity is clustered. We present a database of >15 million sequence variants that provides much of this approach's power. In our four test cases, only a single candidate single nucleotide polymorphism (SNP) remained after subtracting all database SNPs from a mutant's critical region. The saturation of the common SNP database and our open source analysis pipeline MegaMapper will improve the pace at which the zebrafish community makes unique discoveries relevant to human health.

KEY WORDS: Deafness, Mutant, Next-generation sequencing, Positional cloning, Whole-genome sequencing, Zebrafish

INTRODUCTION

The rise of the zebrafish as one of the pre-eminent model systems began two decades ago with the application of genetic approaches for understanding embryonic development in vertebrates (Kimmel, 1989; Kimmel et al., 1989; Kimmel et al., 1995; Driever et al., 1996; Haffter et al., 1996). Zebrafish shares not only many of the same genes and genetic pathways with humans, but also similar cell types, tissue types, organs and developmental mechanisms that contribute to adult anatomy and physiology. For example, genetic screens in zebrafish identified mutations affecting the morphogenesis of the ear, regeneration of sensory hair cells and transduction of sound (Granato et al., 1996; Malicki et al., 1996; Whitfield et al., 1996; Ernest et al., 2000; Sidi et al., 2003; Solomon et al., 2003; Kappler et al., 2004; Söllner et al., 2004; Starr et al., 2004; Nicolson, 2005; Seiler et al., 2005; Asai et al., 2006; López-Schier and Hudspeth, 2006; Schibler and Malicki, 2007; Obholzer et al., 2008; Abbas and Whitfield, 2009; Behra et al., 2009; Dutton et al., 2009; Gleason et al., 2009; Millimaki et al., 2010; Sweet et al., 2011). These discoveries increased our understanding and potentially our ability to treat human patients suffering from loss of hearing and balance.

Although the screens were successful in finding many mutant lines, positional cloning of zebrafish genes remains laborious and time consuming. As cloning is the first step towards understanding the molecular origin of a phenotype, this bottleneck of uncloned mutants represents a significant investment of research time, labor and money that has not yet reached fruition. Zebrafish Information Network (ZFIN), the model organism database for zebrafish (Bradford et al., 2011), contains ~9900 mutant lines of which 62% of these chemically induced lines remain uncloned. Presumably, both the actual number and percentage of uncloned mutants is even higher than this figure because many researchers wish to clone mutants they find in screens before publishing and submission to ZFIN.

The approach currently used by most zebrafish laboratories for positional cloning has changed relatively little in over a decade (Zhou and Zon, 2011). A typical method involves performing hundreds, even thousands, of PCR reactions on each of hundreds of individual embryos to establish a map position, followed by piecemeal molecular analysis of all nearby genes in search of the mutation. Within the last few years, next-generation sequencing (NGS) platforms from several companies have matured to the point at which they are now a disruptive technology (Pareek et al., 2011). Rather than simply making previous approaches less expensive, NGS is making previous approaches obsolete and replacing them with new experimental paradigms. With regards to positional cloning, researchers have used genome-sequencing technologies to identify mutants in model organisms such as *Arabidopsis*,

[1]Department of Systems Biology, Harvard Medical School, 200 Longwood Avenue, Boston, MA 02115, USA. [2]Oregon Hearing Research Center and Vollum Institute, Howard Hughes Medical Institute, 3181 SW Sam Jackson Park Road, Oregon Health and Science University, Portland, OR 97239 USA. [3]OHSU School of Medicine, Department of Cell and Developmental Biology, 3181 SW Sam Jackson Park Road, Basic Sciences L215, Portland, OR 97239, USA.

*These authors contributed equally to this work
‡Author for correspondence (megason@hms.harvard.edu)

Accepted 28 August 2012

Caenorhabditis elegans, *Drosophila*, mouse and various fungi (Smith et al., 2008; Blumenstiel et al., 2009; Irvine et al., 2009; Schneeberger et al., 2009; Cuperus et al., 2010; Doitsidou et al., 2010; Zuryn et al., 2010; Austin et al., 2011; Fairfield et al., 2011; Uchida et al., 2011). However, these approaches are not universally applicable owing to differences in genome sizes, reference genome qualities and availability of inbred lines. Recently, several zebrafish groups have reported using novel sequencing strategies to map various mutants (Gupta et al., 2010; Bowen et al., 2012; Leshchiner et al., 2012; Voz et al., 2012).

Here, we describe an approach for positional cloning by whole-genome sequencing. Although we focus on the inner ear of zebrafish, this approach could be applied to mutants affecting any tissue in zebrafish as well as to additional organisms. We describe the identification of the causative genetic lesion for four inner ear mutations using whole-genome sequencing. We identified the first two mutations with what we term 'bulk segregant-based linkage analysis followed by bioinformatic filtering for mutagenic polymorphisms in whole-genome sequencing' (Fig. 1A, BSFseq). Mapping crosses followed by bulk segregant analysis (BSA) can take up to six months (two generations) of time and labor. In order to reduce the amount of time and labor needed to identify the causative lesion, we developed a second strategy called 'homozygosity mapping followed by bioinformatic filtering for mutagenic polymorphisms in whole-genome sequencing' (Fig. 1B, HMFseq) based on homozygosity mapping (Lander and Botstein, 1987) of pooled F1 embryos. Utilizing the large intrastrain variation in zebrafish laboratory strains combined with the aid of filtering previously known single nucleotide polymorphisms (SNPs) and mutant-effect prediction, we were able to identify mutants using this accelerated pipeline. To simplify the sequence analysis, we developed an open source and freely accessible software pipeline called MegaMapper based on the Galaxy toolkit (Giardine et al., 2005; Blankenberg et al., 2010; Goecks et al., 2010). Starting with raw sequence reads from BSFseq or HMFseq, MegaMapper automatically performs haplotype calling, homozygosity measurement, filtering of previously known wild-type variants, effect prediction of variants, and visualization. We estimate that by using the experimental approaches presented here, our extensive wild-type variant database and our software pipeline, mutants can now be positionally cloned one to two orders of magnitude faster than with standard approaches with the added benefit of being less costly: currently, our approach takes two weeks and costs $2000 per mutant.

MATERIALS AND METHODS

Zebrafish husbandry

The *jj59* and *jj410* mutants, generated in the AB background (Schibler and Malicki, 2007), were obtained from ZIRC (Eugene, OR, USA) and crossed into SJD for mapping. *an158-3/cdh23nl9* was generated in AB, then hybridized with TU/TLF. We sequenced mapping F2s, but used homozygosity instead of haplotype to map the mutation. By contrast, *astronaut/tm290d* was mutagenized in TU (Tubingen screen of 1996) and maintained in TLF for several generations; thus, a typical mapping generation was not used. Mutant zebrafish larvae were sorted for phenotype, euthanized by aquatic exposure to Tricaine, and flash frozen in liquid nitrogen for processing.

Library preparation for next-generation sequencing

For each library, we pooled 100-200 individuals at 120 hours post-fertilization (hpf). After isolation of genomic DNA, ~5 μg was sheared to 200-250 bp fragment size using a Covaris focused acoustic sonicator (Covaris, Woburn, MA, USA). After size selection of fragments by agarose gel electrophoresis (2% gel), we constructed paired-end libraries with Illumina adaptors using 1 μg of sheared input DNA and the NEXTflex DNA Sequencing Kit (BIOO Scientific, Austin, TX, USA). We subjected libraries to another round of size selection on a 2% low-melt agarose gel and DNA quality control using an Agilent 2100 Bioanalyzer. We performed sequencing as paired-end 50 bp runs on an Illumina HiSeq 2000 machine. Between 100 million and 220 million reads per end per library were obtained, for average genome coverage of 6-14×.

MegaMapper pipeline

Sequencing reads were mapped to the unmasked Zv9.60/danRer7 genome reference using BWA within Galaxy. Unmapped or unpaired reads and reads with a map quality of <30 (uniquely mapped with high confidence) were discarded, as were read duplicates (using rmdup, SAMtools). SNPs were then called using mpileup (SAMtools). We required a threefold minimum and 32-fold maximum coverage with a Q-score >30 for SNP calls. To remove potential sequencing errors, we removed 'heterozygous' SNPs that showed only a single base divergent with the reference. In addition, we required that mapping SNPs show at least one read in each direction. SNPs were called as heterozygous if their non-reference allele occurred with a frequency [nonref. reads/(ref. + nonref. reads)] of 0.2-0.9. SNPs were called homozygous if their non-reference allele frequency was 1 (i.e. no reference reads). For BSFseq mapping, the resulting list of mutant library SNPs was intersected with a list of pre-existing wild-type mutagenesis strain SNPs. The pre-existing list of mapping strain SNPs was then subtracted from this intersect. For each remaining SNP, the allele frequency (AF) was calculated. The average AF of each chromosome was determined, and the chromosome with the highest AF was chosen for further analysis. The candidate chromosome was divided into 100 bins and the average AF for each bin was calculated. A locally weighted scatter-plot smoothing (LOESS) fit was performed on the resulting values, and the bin with the maximum AF chosen as first critical interval center prediction. For the same bins, the frequency of heterozygous SNPs was calculated and another LOESS fit performed. All bins containing <20% of the maximum SNP value were considered to be part of the critical interval. The midpoint of this critical interval was then averaged with the predicted position of the highest AF, and this average was used as midpoint to establish a compromise critical interval of 6 Mb total. Next, previously known wild-type SNPs were subtracted from all homozygous SNPs within the critical interval, and a variant effect prediction was performed on the remainder. Non-synonymous SNPs and SNPs in splice sites were listed separately from synonymous SNPs.

For HMFseq mapping, the fraction of total homozygous over total SNPs per chromosome determined the candidate chromosome. The candidate chromosome was then divided into 100 bins the size of which was determined by correlating the MGH recombination map with physical distance. For each bin, the ratio between homozygous over heterozygous SNPs+0.1 determined the map score. The maximum of a LOESS fit over the map scores marked the first position. As in BSFseq, the minimum heterozygosity bin gave position two. The average of these positions marked the center of the HMFseq critical interval, and candidate SNPs were filtered and listed as for BSFseq.

All scripts were written in Python, Pearl or R and are available for download with an open-source BSD license (free for academic and commercial use) from https://wiki.med.harvard.edu/SysBio/Megason/MegaMapper. For data flow, see supplementary material Fig. S1.

Morpholinos, RT-PCR and genotyping primers

Morpholinos were obtained from Gene Tools (Philomath, OR, USA) and oligonucleotides from IDT (Coralville, Iowa, USA; for sequences, see supplementary material Table S1).

Assays for hair cell function and integrity

Dye uptake assays were performed by incubating live larvae in 10 μM YO-PRO1 (Molecular Probes/Life Technologies) in 1× Danieau's solution for 1 minute, then rinsing with Danieau's solution and imaging. To test hair bundle integrity, 120 hpf larvae were anesthetized in 1× Tricaine and fixed in 4% paraformaldehyde (PFA) in PBS at 4°C overnight. Larvae were washed three times in PBS and incubated in 2% Triton X-100 in PBS at 4°C overnight. After three further washes in PBS, larvae were then incubated with 2 mM Phalloidin-A488 (Molecular Probes/Life

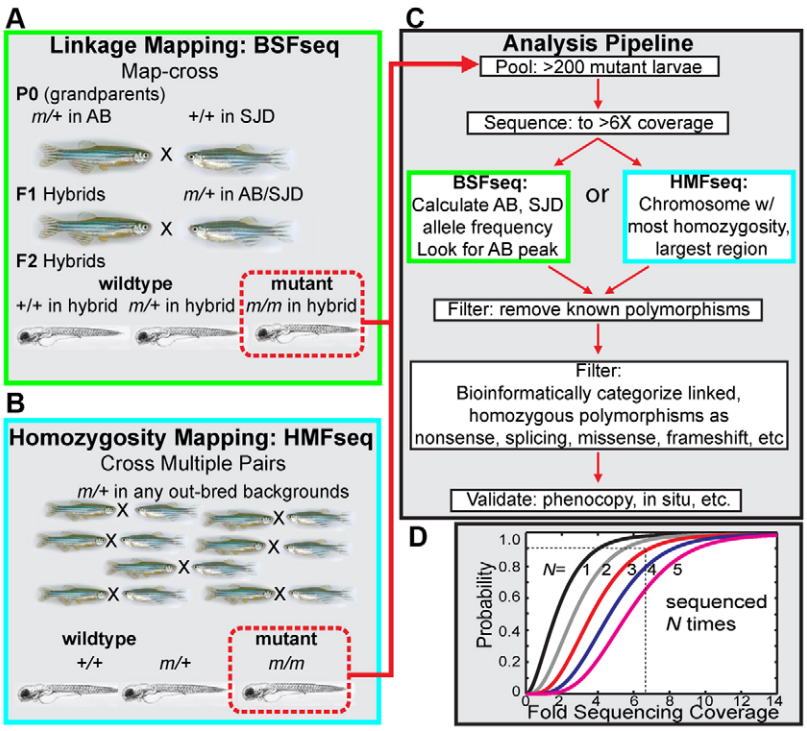

Fig. 1. Overview of bulk segregant linkage mapping and homozygosity mapping strategies. (**A**) In the bulk segregant linkage mapping (BSFseq) strategy, a carrier of the mutant (AB) was out-crossed to a different wild-type strain (SJD). We repeatedly crossed a pair of F1 hybrids to produce >200 mutant larvae, which were pooled to produce a bulk segregant sequencing library. (**B**) The homozygosity mapping strategy (HMFseq) is faster because of the absence of a mapping cross. Multiple pairs of carrier fish are crossed and mutants are collected (again, >200 mutant larvae). More than 3 months are saved in generation time and screening time. (**C**) The pooled libraries are sequenced to greater than sixfold coverage and either analyzed by linkage using known haplotypes from the outcross in BSFseq or analyzed for homozygosity. The outcome of both is a critical region in which the mutant lies. The next step is to filter out all known SNPs and then use genome annotations to categorize and rank the strength of the alleles' effects. The final step is validation by phenocopy and in situ. (**D**) Probability that a SNP is sequenced *N* times as a function of sequencing depth. We plotted Poisson cumulative distribution functions for different event frequencies (*N*) to determine the relationship between sequencing depth and the probability of sequencing a locus. For a 90% chance of sequencing a SNP at least three times, the genomes must be sequenced to at least 6.6-fold coverage.

Technologies) for at least 3 hours at 4°C, rinsed again and finally mounted and imaged on a ZEISS LSM710 confocal microscope.

RESULTS

Two strategies for genetic mapping by genome sequencing

There are a number of different potential strategies for positional cloning by whole-genome sequencing. These strategies differ in: (1) the amount of time and labor required for genetic crosses; (2) the cost of sequencing; and (3) the probability of identifying the causative mutation. We compared mapping based on bulk segregant analysis with linkage to an approach based on homozygosity (BSFseq and HMFseq; Fig. 1A,B). HMFseq is faster and less expensive, but may be less foolproof depending on the extent of consanguinity in the fish (Lander and Botstein, 1987; Hildebrandt et al., 2009). The approaches both involve sequencing a single pool of mutants and analyzing the sequence for causative mutations using a similar bioinformatic analysis pipeline, but they differ in the genetic approach used for mapping (Fig. 1C).

We performed standard map crosses to generate F2s with segregating polymorphisms. In order to maximize the number of polymorphisms for use in linkage analysis, the mutant lines were out-crossed to the divergent zebrafish strain SJD (Johnson et al., 1995; Guryev et al., 2006; Bradley et al., 2007). We raised these F1 hybrid embryos to adults and identified mutant carriers (*m*/+) by in-crossing and checking for phenotypically mutant progeny (*m*/*m*). We in-crossed F1 heterozygotes to generate F2 clutches, which we sorted by phenotype into wild-type and mutant pools to collect at least 200 mutant embryos. We used only a single P0 pair to simplify the analysis by ensuring that a maximum of four alleles were present for all loci, and more typically just two, although the later success of HMFseq indicates that this was not necessary.

Typically, much time and effort is put into defining the smallest critical interval containing the mutation because of the tedium of sequencing numerous candidate genes using traditional molecular biology approaches. However, with the mutant's genome sequence available, sifting through a larger candidate region can be done efficiently using bioinformatics. We thus tested an alternative approach called HMFseq based on homozygosity mapping and bioinformatic filtering. In this approach, a time-consuming mapcross was *not* performed (compare Fig. 1A and 1B). Instead, we collected mutants from approximately eight pairs of carriers that had been maintained through outcrosses to prevent low fecundity. AB and TU are not highly inbred strains; therefore, they were expected to contain intrastrain SNPs at an average frequency of one per 2000 bp (compared with an interstrain SNP rate of one per 500) (Guryev et al., 2006; Bradley et al., 2007). We again pooled and sequenced 100-200 phenotypically mutant embryos. But, rather than perform linkage analysis, we identified 500 kb regions that contained an elevated homozygosity as measured by the ratio between SNPs appearing homozygous versus heterozygous. Homozygosity mapping should pull out the mutant locus, but it might also pull out a number of nonlinked loci owing to partial inbreeding in the population and the presence of SNP deserts. The use of eight pairs of fish for collecting embryos minimized this problem. Furthermore, even if more than one region of homozygosity is found in HMFseq, our bioinformatic analysis pipeline can still generate a short prioritized list of candidates. We acknowledge that HMFseq is not as failsafe as BSFseq, but the time savings from a reduced number of generations make it very attractive. We expect that under optimal circumstances HMFseq will allow a mutant to be affordably cloned in less than two weeks. We anticipate that the cost (currently ~$2,000) and accessibility of HMFseq and BSFseq will steadily improve with sequencing technologies.

Sequencing coverage required for identifying causative lesions

Much of genome sequencing's power for identifying mutations arises from bioinformatic filtering and analysis. Ultimately, the lesion must be covered by enough sequencing reads to provide

confidence that the variation is not due to a sequencing error and is homozygous. Our targeted coverage of a mutant genome should result in a candidate polymorphism being sequenced at least three times, with a 12.5% or less chance that the allele is a false positive. We analyzed the distribution functions of random sampling processes to determine the theoretical sequencing depth needed to identify mutants. For a 90% chance of sequencing a SNP or any locus at least three times, the pooled genomes must be sequenced to at least 6.6-fold coverage (Fig. 1D, dotted line). This sequencing depth can be easily achieved for zebrafish using one lane of an Illumina HiSeq 2000. At the beginning of this project, the machine generated ~100 million reads per lane. Using one lane of 2 × 50 bp reads, this resulted in 6.6-fold coverage of the 1.5 Gb zebrafish genome. During the course of this research, Illumina improved the machine's chemistry and software. Currently, a single lane produces greater than tenfold coverage.

MegaMapper: pipeline of fast mapping, filtering, and effect predictor identifies candidates for mutations

To generate a user-friendly platform, we developed MegaMapper, a web-based bioinformatic analysis tool for HMFseq and BSFseq. MegaMapper is based on Galaxy (http://galaxy.psu.edu). Running MegaMapper and analyzing large quantities of data can be done on a standard powerful workstation or using cloud computing. Using VirtualBox (https://www.virtualbox.org/) we packaged MegaMapper, its dependencies, reference sequences, and Galaxy itself into a single bundle that is ready for data analysis on a local workstation as a virtual machine. We also created an Amazon Machine Image (AMI), to allow users to instantiate their own MegaMapper server on the Amazon Elastic Compute Cloud. Notwithstanding the turnkey packaging of MegaMapper, it remains open to specific parameter alterations at any step to optimize the pipeline for future changes in library specifics and to accommodate the particular needs of the individual researcher. The full source code, the ready-to-use VirtualBox, and the AMI are free to download (www.digitalfish.org/MegaMapper).

Presented in supplementary material Fig. S1 are the data flow diagrams for two pipelines in MegaMapper, one for BSFseq and one for HMFseq (described in Materials and methods). After cleaning up and aligning high quality reads to the reference genome, the pipeline forks into mutant mapping and a candidate prediction segment. The mapping segment groups SNPs into strain-specific SNPs (BSFseq) or homozygous and heterozygous SNPs (HMFseq). Mapping then establishes the maximum divergence between SNP groups. We expected and later confirmed that the chromosome with the greatest overall divergence harbors the mutation. The genomic interval with the greatest local divergence on this chromosome establishes the critical interval.

Meanwhile, the candidate prediction pipeline subtracts all known wild-type variants from the variant list [supplementary material Fig. S1A,B, the right pipelines (gray) are the same for BSFseq and HMFseq]. MegaMapper feeds the remaining variants into SnpEff, a variant effect prediction tool (Cingolani, 2012). MegaMapper filters variant effects by zygosity, category and strength. Finally, MegaMapper intersects the filtered variant effect list with the critical interval it has established and outputs a prioritized shortlist of candidates for consideration showing the gene name, mutation, predicted effect, and coverage as well as a graphical display of mapping results for all 25 chromosomes (all panels in Fig. 2 and Fig. 3 were automatically generated by MegaMapper). The user can explore the shortlist of candidates using other available databases and tools as well as their prior knowledge of the biology to pick candidates of interest for further validation.

Bulk segregant linkage mapping

To begin mapping ear morphogenesis mutants, we obtained mutants, *jj59* (*hako mimi*, *kmi*) and *jj410* (*ale ucho*, *alo*), that originated in a prior mutagenesis screen (Schibler and Malicki, 2007). We inbred the F1 hybrids (AB/SJD) to identify a pair of heterozygous hybrid carriers and then repeatedly crossed this pair to generate >200 mutant larvae for sequencing libraries (see Materials and methods). The sequencing reads were filtered and mapped using MegaMapper. Using our own (for SJD) and publicly available (for AB) sequencing data to generate reference SNP databases unique to either AB or SJD, we calculated and plotted the allele frequency of the two contributing haplotypes (Clark et al., 2011). Fig. 2A shows an example of these BSFseq mapping plots for *jj410* in which the red trace is the LOESS of binned AB allele frequency (Fig. 2A, shown in orange). In general, we found that the chromosome with the greatest proportion of mutagenic strain SNPs (AB) reliably identifies the correct chromosome (Fig. 2A, bar adjacent to each chromosome's plot). We then looked more closely at the putatively linked chromosome by locating the peak of its AB allele frequency plot (Fig. 2B). For *jj410*, this peak is at 33,250,000 on chromosome 8. Because allele frequency might be susceptible to imprecision originating in local deserts of AB SNP density, gene conversion, large deletions, structural variation and copy-number variation in both AB and SJD backgrounds, we chose to analyze the putative chromosome's SNP profile using an approach that does not rely on a priori knowledge of the mapping strains. The plot in Fig. 2C shows the heterozygous SNP density along chromosome 8 of *jj410* (Fig. 2C). This analysis identified a locus, the plot's minimum, at 36,750,000. For the BSFseq pipeline, we relied on a compromise fit that used the average of these two positions, 35,000,000, to minimize the different biases of each approach. Based on the expected recombination frequency, the number of individuals in the pool, and average sequencing depth, we estimated our mapping accuracy to be within 3 Mb of this position. The resultant critical region of the compromise fit stretched from 32,000,000 to 38,000,000 on chromosome 8. Our bioinformatics analysis pipeline identified a single candidate lesion in *lmx1b* (ENSDARG00000068365) that introduced a premature termination codon at 34,129,111 on chromosome 8, only 871 kb from the midpoint of the compromise fit's critical region. To validate BSFseq further, we mapped an additional ear morphogenesis mutant, *jj59*, which gave similar results (supplementary material Fig. S2).

Homozygosity mapping

The accuracy of the heterozygous SNP density analyses and the presence of a sizable intrastrain SNP variation suggested that mapping could be performed without defined haplotype linkage. We pursued a homozygosity mapping strategy (HMFseq) with two auditory/vestibular mutants, *tm290d* and *an158-3*, from independent mutagenesis screens (Granato et al., 1996; Nicolson et al., 1998) (T.N. and A.N., unpublished). To test our homozygosity mapping approach, we sequenced the previously unmapped N-ethyl-N-nitrosourea (ENU) mutant *an158-3* using a library prepared from 200 pooled phenotypically mutant F1 progeny to at least sixfold coverage (Fig. 1B). *tm290d* had previously been mapped and will be discussed in more detail later (see supplementary material Fig. S3 for the mapping results for *tm290d*). To map *an158-3*, we first plotted the density of homozygous and heterozygous SNPs for each

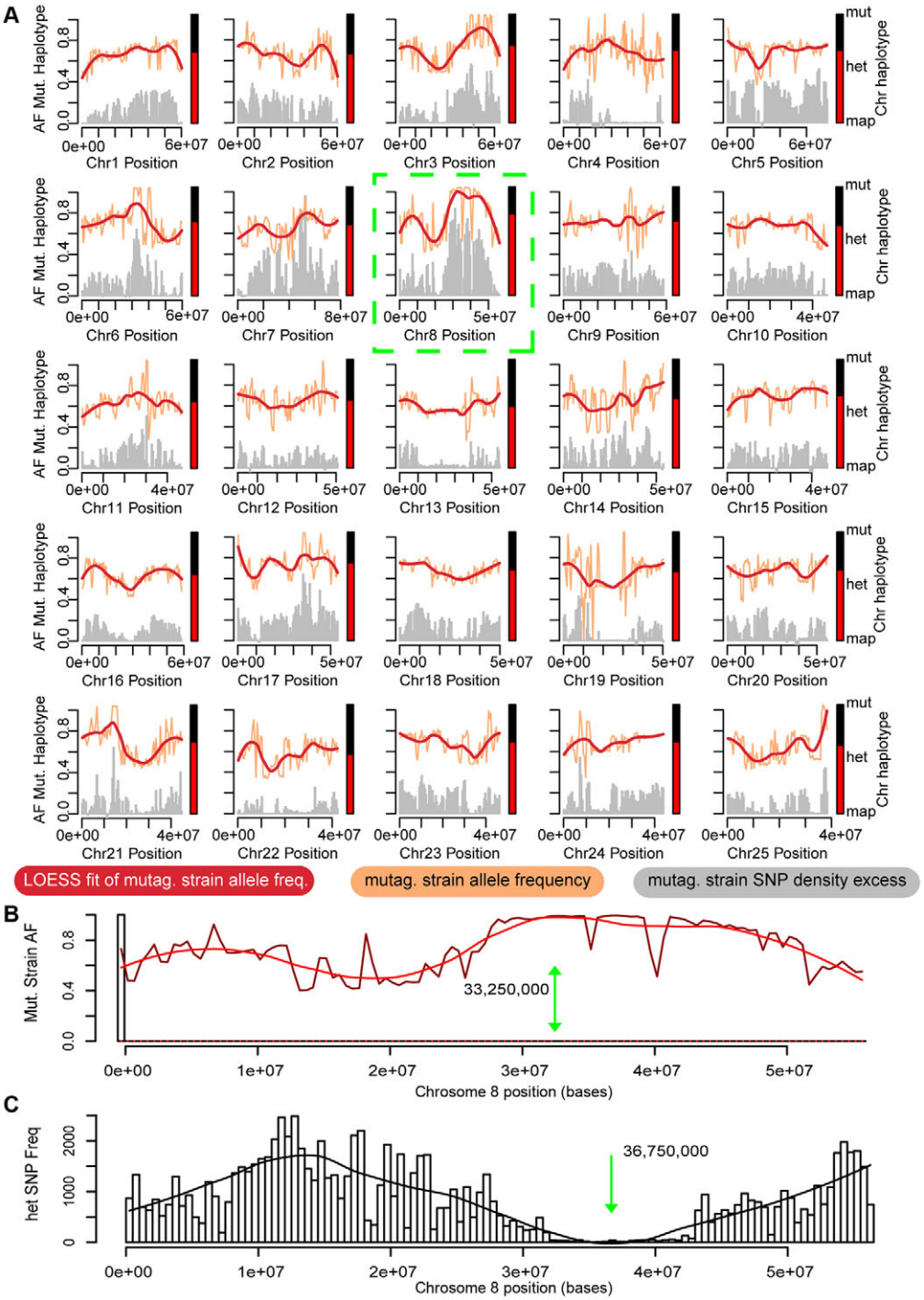

Fig. 2. Bulk segregant linkage mapping and filtering of *jj410/lmx1b*. (**A**) Mapping of regional AB haplotype allele frequencies onto individual chromosomes (AF Mut. Haplotype). Chromosomal position intervals of ~200 kb are plotted in orange. A LOESS fit of the haplotype interval data is plotted in red. Histogram data show the measured local SNP density of AB-haplotype SNPs (gray). Bars next to individual chromosomes show averaged chromosomal mutagenesis strain allele frequency (red, vs mapping strain AF in black). (**B**) Allele frequency scan of the candidate chromosome 8 (highest bulk allele frequency, green box in A) position by AB haplotype allele frequency (dark red) and the corresponding fit (red line). The maximum of the fit is indicated (green arrow). (**C**) SNP density histogram of all detected heterozygous SNPs on the chromosome. Loss of heterozygosity is calculated by fitting to the density histogram (black line). The minimum of the fit is indicated (green bar). After filtering of known wild-type SNPs, only one homozygous SNP remains that causes a stop gain in all three known splice isoforms of *lmx1b.1* (Table 3).

chromosome (Fig. 3A, red and black plots). To identify the putative chromosome, we calculated the fraction of homozygous SNPs out of total SNPs per chromosome (Fig. 3A, bar to the right of each chromosome's density plot). In plots of SNP density for the genome sequencing data of *an158-3*, most chromosomes showed a similar number of heterozygous and homozygous SNPs or an excess of heterozygous over homozygous SNPs. We ranked chromosomes by their chromosomal homozygosity value and chose the chromosome with the largest value, chromosome 13, as the candidate for harboring the mutation. We also divided each chromosome into 300 equally sized bins that each corresponded to ~100-300 kb in physical distance, depending on chromosome size. We then counted the number of homozygous (red) or heterozygous (black) SNPs per bin. We determined the ratio of homozygous to heterozygous SNPs per bin and performed a LOESS fit to visualize the trends (light blue line). We also plotted the average SNP coverage per bin to help distinguish putative SNP deserts from gaps in coverage (example shown in supplementary material Fig. S4).

We then chose the bin with the peak homozygous to heterozygous ratio at position 46,419,377 on the candidate chromosome 13 as the SNP density mapping prediction (Fig. 3B). We averaged this prediction with the minimum of the fit to the heterozygous SNP densities at position 44,500,000 to obtain the compromise fit peak at 45,459,500, which we used to define a critical interval from

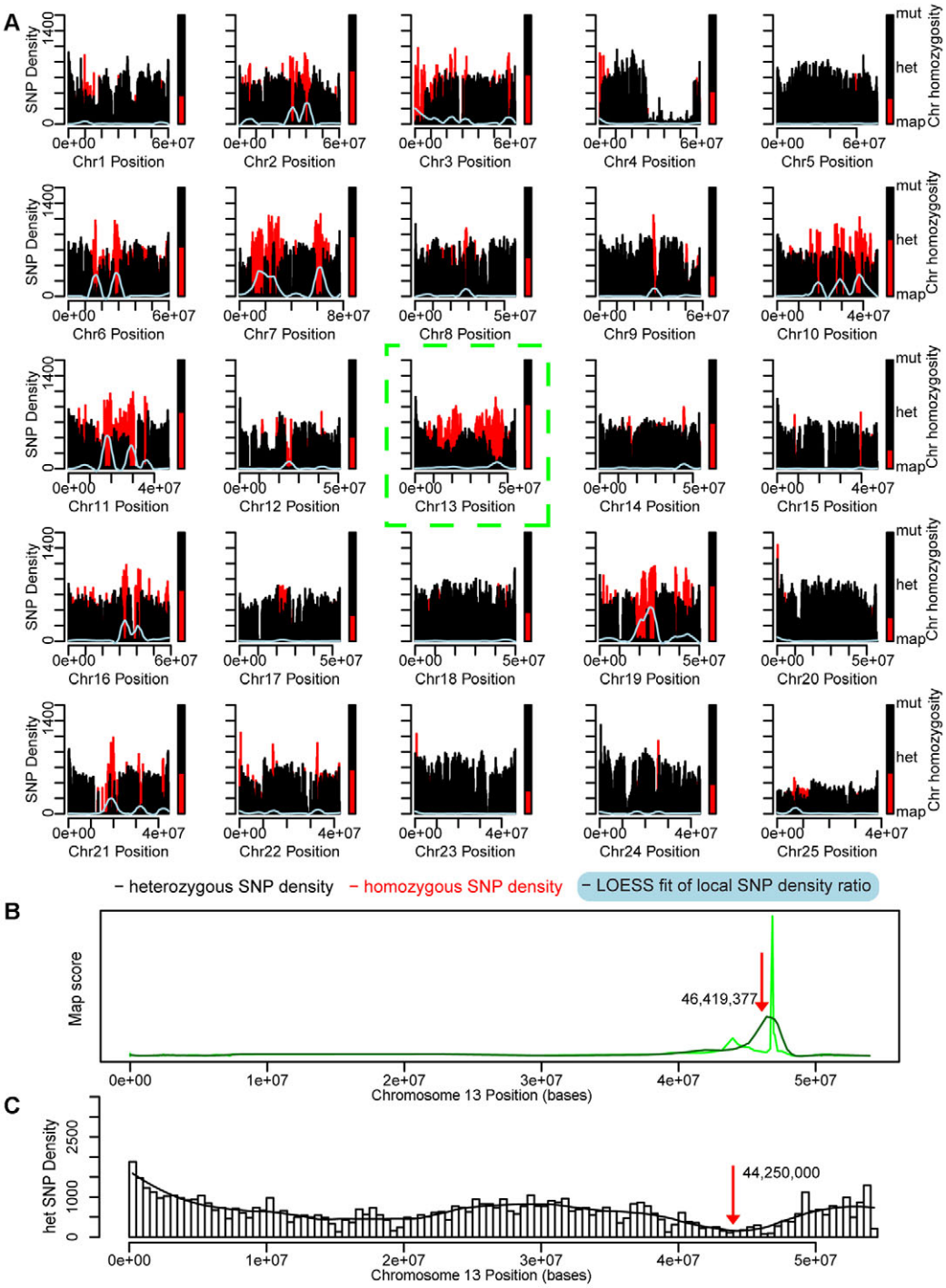

Fig. 3. Homozygosity mapping and filtering of *an158-3/cdh23^{nl9}*. (**A**) Chromosomal SNP density and homozygosity map of *an158-3/cdh23^{nl9}*. Red bars are homozygous SNP counts, black bars are heterozygous SNP counts. Homozygous-to-heterozygous ratios were fitted for trend recognition using local regression (LOESS; light blue). Bars next to individual chromosomes show the degree of homozygosity (red bar, vs heterozygosity in black). Chromosome 13 shows the highest degree of homozygosity and thus probably harbors the mutation. (**B**) Scan of chromosome 13 for a peak in local homozygosity and fit peaks at 46,419,377 (red arrow). (**C**) A SNP density histogram of all detected heterozygous SNPs on chromosome 13. Loss of heterozygosity is calculated by fitting to the density histogram (black line). The minimum of the fit is indicated (red arrow). The average of the resulting values from the graphs shown in B and C was used to position the center of the critical interval. Variant effect prediction of all detected homozygous SNPs within the critical interval after filtering reveals one candidate SNP. SNP 13_43745899_G/A is located in the essential part of a splice acceptor site for 'exon 7' (ENSDARE00000862400) of *cdh23*, leading to a frameshift by mis-splicing and truncation within the extracellular domain (Table 3).

40,000,000 to 48,000,000 in the same way as with our BSFseq strategy (Fig. 3C). Our bioinformatics pipeline again identified a single candidate in the critical interval at 43,745,899, effecting the essential invariant intronic portion of an acceptor splice site of *cadherin 23* (*cdh23*; ENSDARG00000007561), a known deafness gene that encodes a component of hair cell tip links (Bolz et al., 2001; Bork et al., 2001).

Comparison of accuracies for various BSFseq and HMFseq positioning methods

In our four mapping datasets we found that no single positioning method was consistently the most accurate. A compound approach that incorporated multiple types of positioning methods, as described above for both BSFseq and HMFseq, was the most accurate overall for our datasets (Table 1, compromise fit). We expect that mapping accuracy will improve with an improved genome assembly, which, in addition to general refinement, will require an understanding of large rearrangements between strains, precise cataloging of copy-number variation and proper placing of repetitive regions.

We also tested how robust our mapping strategy was to accidental mis-sorting of wild-type embryos into the phenotypic pool, which can be a problem with subtle phenotypes. For *jj410*, we sequenced both mutants as well as wild-type siblings allowing us to simulate

Table 1. Accuracies of various HMFseq and BSFseq positioning methods

MegaMapper pipeline	Method used for identification of mutation	Mean distance (Mb)	s.d. (Mb)
BSFseq	Mutant strain allele frequency maximum	1.9	1.6
	Heterozygosity minimum	4.4	6.4
	SNP density ratio fit	3.2	3.9
	Compromise fit	1.4	0.9
HMFseq	MGH SNP density ratio fit maximum	2.6	2.6
	Heterozygosity minimum	1.2	0.7
	SNP density ratio fit	3.2	3.5
	Compromise fit	1.8	1.4

Unique to BSFseq, we used the maximum of the mutagenic strain allele frequency to position the critical region. Unique to HMFseq, we used the ratio of homozygosity to heterozygosity as a function of genetic distance (MGH SNP density ratio fit maximum). For both pipelines, we used heterozygosity minimum and SNP density ratio as a function of absolute distance, as described in the main text. For BSFseq, the compromise fit is the average of the mutant strain allele frequency maximum and the heterozygosity minimum. For HMFseq, the compromise fit is the average of the SNP density ratio fit of the homozygosity maximum and the heterozygosity minimum. s.d., standard deviation for the distance from identified lesions in the four mutants.

computationally the effect of contaminating the mutant sequence with an additional 10% wild-type reads. We found that this level of wild-type contamination did not significantly affect mapping position, although as predicted the mutant locus may no longer appear homozygous (supplementary material Table S4).

The zebrafish SNP universe

In order to be able to map genomic lesions by haplotype, we set out to determine strain-specific SNP markers. To this end, we re-analyzed publicly accessible data of wild-type strains (AB, TU) deposited in the Sequence Read Archive (SRA; http://www.ncbi.nlm.nih.gov/sra), data published by Bowen et al. (Bowen et al., 2012) (AB,TU,WIK,TL) as well as our own sequencing data (AB, TU, TL, SJD, mixed strains). We defined 'strain-specific' as any SNP marker in a particular strain that we could not detect in any other wild-type SNP set (supplementary material Tables S2, S3). These strain-specific markers form one of the foundations of our BSFseq mapping strategy. Our final combined SNP dataset of 15 million unique SNPs represented the basis for both haplotype mapping and filtering of known wild-type variants in MegaMapper. This variant database appears to comprise a large part, perhaps the majority, of SNPs that are commonly present in varying combinations and frequencies in the common zebrafish laboratory strains (analysis shows that the distribution of SNPs is similar to other organisms and that our database is approaching saturation; supplementary material Figs S5, S6). We have summarized our SNP dataset (supplementary material Table S5) and it can be downloaded in VCF4.1 format from http://digitalfish.org.

Power of filtering to identify causative lesions

For the four mutants we mapped, 891, 121, 113 and 46 total homozygous SNPs remained within the mutant's critical region (Table 2). To reduce the number of candidates, we filtered out all other wild-type SNPs in our zebrafish SNP database. Notably, only a single candidate SNP remained for all four mutants after this subtraction (Table 2) and, as shown below, further evidence supported this single candidate as being the correct prediction in all four cases.

Identification of lesions in *lmx1b, jag1b and cdh23*

By applying MegaMapper with filtering to sequencing data from ear mutants, we identified the causative lesions of three mutants (*jj410*, *jj59* and *an158*-3) and a strong candidate for a fourth mutant (*tm290d*; Table 3). *tm290d* had previously been mapped by traditional methods between markers z1181 (11:1,745,143-1,749,357) and z8214 (11:3925407-3925550) (T.N., unpublished). We identified a homozygous, unique SNP in that critical region at 2,815,634 on chromosome 11 that causes a premature stop codon at residue 80 of Lhfpl5a (ENSDARG00000045023). The homolog of *lhfpl5a*, *LHFPL5*, causes nonsyndromic deafness in humans and the *hurry scurry* (*hscy*) mouse, thus making it a strong candidate (Longo-Guess et al., 2005; Shabbir et al., 2006; Longo-Guess et al., 2007). However, we could not validate this mutation by morpholino phenocopy, possibly because *lhfpl5a* functions at a late stage of development by which time the morpholinos have been degraded.

We experimentally validated the other three predicted mutations. At 65 hpf, the topology change of the zebrafish otic vesicle is well underway and the fusion of the semicircular canal's lateral projection with both the anterior and posterior projection is visibly complete (Fig. 4A, projections labeled). In *jj410* mutants, the lateral, anterior and posterior projections were abnormally shaped compared with wild type (Fig. 4B). In *jj59* mutants, the otic vesicle was shorter in the anterior-posterior axis (Fig. 4H). All putative lesions were re-sequenced using traditional Sanger sequencing of pools and individuals. For both *jj410* and *jj59*, the allele was completely homozygous in the pooled mutants (Fig. 4F,L, arrows). We confirmed linkage by genotyping individuals: 17/17 *jj410* and 10/10 *jj59* mutant individuals were homozygous for mutant alleles whereas for wild-type individuals 4/9 were homozygous wild type

Table 2. Power of filtering to identify mutation

MegaMapper pipeline	SNP class	*jj410*	*jj59*	*an158-3*	*tm290d*	Average
BSFseq	Unfiltered SNPs	891	121	105	40	289
	Filtered SNPs	1	1	1	1	1
HMFseq	Unfiltered SNPs	374	93	113	46	157
	Filtered SNPs	1	1	1	1	1

Within each mutant's 6 Mb critical region there were 40-891 homozygous SNPs before filtering. After filtering all variants in our SNP universe, there remained a single homozygous SNP in each mutant's critical region.

Table 3. Molecular lesions identified by MegaMapper

Allele	Chromosome	Position	Gene	Change	Type	Effect	Amino acid change	Residue number
jj410	8	34,129,111	*lmx1b.1*	T>A	STOP_GAINED	Truncation	R>*	147/192[§]
jj59	13	35,593,144	*jag1b*	G>A	SPLICE_DONOR	Frameshift	–	–
an158-3	13	43,745,899	*cdh23*	G>A	SPLICE_ACCEPTOR	Frameshift	–	–
tm290d[‡]	11	2,815,634	*lhfpl5a*	T>A	STOP_GAINED	Truncation	K>*	80

For the four mutants, MegaMapper identified lesions in these four genes. This table summarizes the locations of the lesions, the base change introduced by mutation, the consequence of the mutation, and the effect the mutation has on the gene's transcript or protein product.
[‡]The cloning of the *lhfpl5a* as the cause of *tm290d*'s phenotype remains provisional because of complications in its validation.
[§]Both of the known *lmx1b.1* splice isoforms are affected by the stop gain.

and the remaining 5/9 heterozygous for *jj410*, and 4/10 were homozygous wild type and the remaining 6/10 heterozygous for *jj59*. Additional validation came from the expression patterns of *lmx1b* and *jag1b* (ENSDARG00000013168), which are both expressed in the otic vesicle and canal projections at 65 hpf in wild-type embryos (Fig. 4D,J). By comparison, there was a strong reduction of in situ signal for the respective genes in the two morphogenesis mutants, presumably caused by nonsense-mediated decay (Fig. 4E,K). Furthermore, injection of 3 ng of morpholino targeting either *lmx1b* or *jag1b* phenocopied their respective mutants at 65 hpf (Fig. 4G,C,I; supplementary material Fig. S7). Finally, the morphogenetic defects present in the ears of *jj59* mutants were similar to those seen in the *mind bomb* zebrafish mutant in which Notch signaling is globally disrupted (supplementary material Fig. S7).

To validate our prediction of the mutation in *cdh23* in *an158-3*, we performed a complementation cross with a previously published allele of *cdh23*, *tc317e* (Söllner et al., 2004). The two alleles failed to complement, and compound mutants lacked an acoustic startle response (data not shown), did not label with the vital dye YO-PRO-1 (Fig. 4M-O) and had splayed hair bundles (Fig. 4P-R). We also re-sequenced the lesion using Sanger sequencing in the mutant population (Fig. 4S). We confirmed linkage by genotyping individuals: 4/4 *an158-3* mutant individuals carried homozygous mutant alleles whereas sequencing of individual wild-type siblings revealed that they were either heterozygous (7/10) or homozygous wild type (3/10). Because the lesion is situated in a splice acceptor site, we validated its effect further by RT-PCR and sequencing of the mis-spliced mutant transcript (supplementary material Fig. S7). In the mutant, a cryptic splice site in the downstream exon is used instead of the default site, leading to a frameshift and early truncation (supplementary material Fig. S7). These results phenotypically and molecularly validate that the mutant phenotypes of *jj410*, *jj59* and *an158-3* are caused by mutations in *lmx1b*, *jag1b* and *cdh23*, respectively. In accordance with ZFIN conventions, these alleles are now *lmx1b*^{jj410}, *jag1b*^{jj59} and *cdh23*^{nl9}.

DISCUSSION

Since the revolution caused by its introduction in the 1970s (Wensink et al., 1974), positional cloning has made incremental

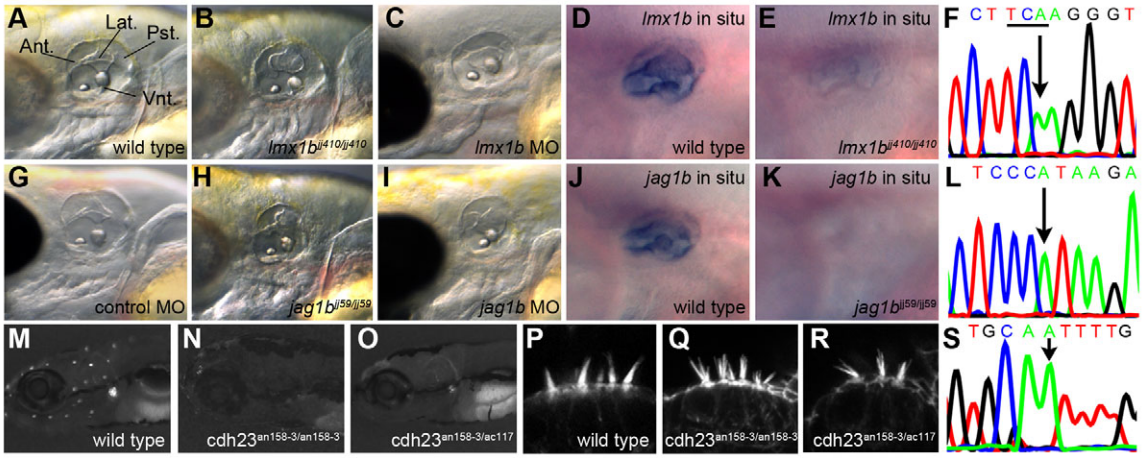

Fig. 4. Validation of the identified ear mutants. (**A**) Lateral view of a wild-type ear at 65 hpf (fused projections labeled). Ant., anterior projection; Lat., lateral projection; Pst., posterior projection; Vnt., ventral projection. (**B**) Semicircular canal projections are abnormally shaped in *jj410* ears at 65 hpf. (**C**) *lmx1b* morpholino knockdown recapitulates *jj410* ear phenotype at 65 hpf. (**D**) *lmx1b* antisense in situ probe marks the inner ear and concentrates in semicircular canal projections. (**E**) Severe reduction of *lmx1b* transcript levels in a *jj410* mutant larval inner ear at 65 hpf. (**F**) Sequencing trace of the *jj410* genomic lesion in *lmx1b*; arrow denotes causative mutation that introduces a premature stop codon TGA (underlined, antisense strand was sequenced). (**G**) Control morpholino injection recapitulates wild type. (**H**) *jj59* inner ears are shorter in the anterior-posterior axis and semicircular canal projections are malformed at 65 hpf. (**I**) *jag1b* morpholino knockdown recapitulates *jj59* ear phenotype at 65 hpf. (**J**) *jag1b* antisense in situ probe marks the inner ear and concentrates in semicircular canal projections. (**K**) Severe reduction of *jag1b* transcript levels in a *jj59* mutant larval inner ear at 65 hpf. (**L**) Sequencing trace of the *jj59* genomic lesion in *jag1b*; arrow denotes causative mutation. (**M**) YO-PRO1 uptake by neuromast hair cells in wild-type larvae. (**N**) *an158-3(cdh23nl9)* mutant larval neuromast hair cells fail to take up YO-PRO1. (**O**) Failure to complement YO-PRO1 uptake in *an158-3* x *tc317e* larval hair cells. (**P**) Intact inner ear hair cell bundles in the wild type. (**Q**) Hair cell bundles are splayed in *an158-3(cdh23nl9)* mutant inner ear hair cells. (**R**) Hair cell bundles show splayed phenotype and thus fails to complement in *an158-3* x *tc317e* mutant inner ear. (**S**) Sequencing trace of the *an158-3/cdh23nl9* genomic lesion in *cdh23* (arrow).

improvements. In the recent past, researchers have made efforts to scale up the cloning of zebrafish mutants but the task still remains inefficient, costly or laborious (Geisler et al., 2007). Previous approaches have used lower coverage at the risk of missing the mutation (Bowen et al., 2012). We have highlighted the numerical basis of the risk (Fig. 1D). Others have used genome enrichment that introduces bias in coverage owing to differences in efficiency of enrichment for different loci and requires additional cost and effort for the enrichment itself (Gupta et al., 2010). Facilitating positional cloning through new sequencing technologies carries the promise of a paradigm shift that is important for realizing a deeper and more systematic understanding of the connection between genotype and phenotype.

In this study, we developed and validated genetic and computational pipelines for identifying the causes of zebrafish mutant phenotypes. We packaged the computation pipeline as MegaMapper, which includes a Galaxy-based data flow going from bulk-sequenced mutant genomes with >6.6-fold coverage to a single candidate mutagenic lesion (supplementary material Fig. S1). MegaMapper can be easily and securely deployed as a virtual machine using either VirtualBox for local computing or an Amazon Machine Image for cloud computing (http://digitalfish.org/MegaMapper). The virtual machine contains our SNP databases, an example dataset, and the current zebrafish reference genome and assembly so that one may plug in any current or future sequencing data for analysis. By using Galaxy, many parameters are accessible for customization as well as many other tools.

Another significant bottleneck in current approaches to positional cloning is the collection of large numbers of mutants, which is required for high resolution (<100 kb) mapping. In hindsight, pools of 40 mutants or fewer for BSFseq would have been sufficient as only medium resolution mapping (a few Mb) is required given how well the bioinformatic analyses work. The ability to use fewer mutants for analysis represents significant savings in labor. For HMFseq, it is still advantageous to use mutants pooled from multiple pairs with different parents and grandparents to reduce the chance appearance of blocks of homozygosity. Even if spurious homozygous regions are found, computational subtraction of wild-type polymorphisms can eliminate these regions from consideration. Additionally, the predicted linkage of the mutation with any given chromosome or haplotype block can be confirmed by genotyping the locus with a small number of PCR reactions.

In designing the BSFseq and HMFseq strategies, we also considered sequencing the wild-type siblings. For the BSFseq mapping of *jj410* we did sequence wild-type siblings and the subtraction of its homozygous SNPs from the mutants did clean up spurious peaks, but this additional sequencing was not necessary for mapping or positional cloning (supplementary material Fig. S8). Moving forward, we decided not to sequence wild-type siblings because this would double the cost and provide little additional information to that in our SNP database. We expect that haplotype calls will become more accurate as more data accumulates and that this will reduce any ambiguity arising from secondary mapping peaks so that sequencing wild-type siblings is not worthwhile in well-studied model systems. Additional peaks can also be suppressed by increasing the minimum coverage (e.g. ten times) required for using a marker, or by plotting a percentile (e.g. 10%) of marker frequency rather than the mean; both of these options are user-selectable in MegaMapper. Additionally, many of the spurious peaks are real and due to shared haplotype blocks in the parents (particularly in HMFseq). Although our algorithm generally overcomes these signals, we recommend a visual inspection of the signal's shape. Shared haplotype blocks tend to have clear and sharp borders whereas the mutant locus will have gradual borders due to recombination. If these approaches were to be applied to non-model systems in which a reference genome and SNP database are not available, then sequencing the wild-type siblings could substitute. Also, for dominant mutations one could sequence the phenotypically wild-type individuals from a heterozygous in-cross to map the mutation and sequence the phenotypically mutant siblings to identify the molecular lesion.

All of the mapped genes, *jag1b*, *lmx1b*, *cdh23* and the candidate *lhfpl5a*, that cause ear phenotypes in zebrafish are associated with diseases in humans. Dominant mutations in the human orthologs of *jag1b* and *lmx1b* cause Alagille and Nail-Patella syndrome, respectively (Li et al., 1997; Oda et al., 1997). Alagille syndrome pathology also includes inner ear dysplasia (Koch et al., 2006). In contrast to syndromic phenotypes, autosomal recessive mutations in the human orthologs of both *cdh23* and *lhfpl5a* cause nonsyndromic deafness (Bolz et al., 2001; Bork et al., 2001; Longo-Guess et al., 2005; Shabbir et al., 2006; Longo-Guess et al., 2007). Along with Protocadherin 15, Cdh23 has recently been identified as one of the components of the tip links of hair-cell bundles, where mechanical forces are transduced into electrical potentials (Kazmierczak et al., 2007; Sakaguchi et al., 2009). *LHFPL5*, the mammalian homolog of zebrafish *lhfpl5a*, has also been identified as a component of hair bundles (Kalay et al., 2006; Shabbir et al., 2006). The specific function of *lhfpl5a* in zebrafish will require further validation, such as by transgenic rescue, that is beyond the scope of this paper. Although our sample size is small, the relevance of all four mutants to human health emphasizes the great impact lying dormant in the thousands of unmapped zebrafish mutants.

Much of the power of the BSFseq and HMFseq approaches comes from filtering out known zebrafish polymorphisms. For all the mutants we mapped, we identified a single candidate SNP after removing all wild-type SNPs in our combined database. The power of filtering to identify the causative lesion will only improve as genome annotation and our knowledge of the zebrafish SNP universe continues to increase. Already, in the work presented here the level of saturation of our zebrafish SNP database is evident. We expect that as more groups combine their genome re-sequencing data, the collective residual polymorphisms will be sufficient to know the majority of polymorphisms existing in laboratory zebrafish strains. In the near future we anticipate that sequencing the genomes of pooled mutants to tenfold or greater coverage, subtracting all known polymorphisms, and then identifying the mutant-causing lesion will be possible within a few days, which will greatly increase the potential of forward genetics in model organisms such as zebrafish as well as in non-model organisms.

Acknowledgements

We thank David Reich and members of the Megason Laboratory for discussions and comments; Bob Freeman for technical advice; and Andrea Albrecht for her assistance with the rough mapping of *tm290d*.

Funding

This work was supported by the National Institutes of Health (NIH) [DC010791 and DC012097]. I.A.S. was supported by a National Research Service Award [F32 FHL097599], T.N. was supported by the NIH [DC006800] and the Howard Hughes Medical Institute, and A.N. was supported by the NIH [HD07284401]. Deposited in PMC for release after 12 months.

Competing interests statement

The authors declare no competing financial interests.

Author contributions

I.A.S., N.O. and S.G.M. conceived and designed the experiments; I.A.S., N.O. and E.S. performed the experiments; N.O. processed the data and developed *MegaMapper*; I.A.S., N.O. and S.G.M. analyzed the data; A.V.N. and T.N. provided the hearing mutants; I.A.S., N.O. and S.G.M. wrote the paper.

Supplementary material

Supplementary material available online at
http://dev.biologists.org/lookup/suppl/doi:10.1242/dev.083931/-/DC1

References

Abbas, L. and Whitfield, T. T. (2009). Nkcc1 (Slc12a2) is required for the regulation of endolymph volume in the otic vesicle and swim bladder volume in the zebrafish larva. *Development* **136**, 2837-2848.

Asai, Y., Chan, D. K., Starr, C. J., Kappler, J. A., Kollmar, R. and Hudspeth, A. J. (2006). Mutation of the atrophin2 gene in the zebrafish disrupts signaling by fibroblast growth factor during development of the inner ear. *Proc. Natl. Acad. Sci. USA* **103**, 9069-9074.

Austin, R. S., Vidaurre, D., Stamatiou, G., Breit, R., Provart, N. J., Bonetta, D., Zhang, J., Fung, P., Gong, Y., Wang, P. W. et al. (2011). Next-generation mapping of Arabidopsis genes. *Plant J.* **67**, 715-725.

Behra, M., Bradsher, J., Sougrat, R., Gallardo, V., Allende, M. L. and Burgess, S. M. (2009). Phoenix is required for mechanosensory hair cell regeneration in the zebrafish lateral line. *PLoS Genet.* **5**, e1000455.

Blankenberg, D., Von Kuster, G., Coraor, N., Ananda, G., Lazarus, R., Mangan, M., Nekrutenko, A. and Taylor, J. (2010). Galaxy: a web-based genome analysis tool for experimentalists. *Curr. Protoc. Mol. Biol.* **19**, 1-21.

Blumenstiel, J. P., Noll, A. C., Griffiths, J. A., Perera, A. G., Walton, K. N., Gilliland, W. D., Hawley, R. S. and Staehling-Hampton, K. (2009). Identification of EMS-induced mutations in Drosophila melanogaster by whole-genome sequencing. *Genetics* **182**, 25-32.

Bolz, H., von Brederlow, B., Ramírez, A., Bryda, E. C., Kutsche, K., Nothwang, H. G., Seeliger, M., del C-Salcedó Cabrera, M., Vila, M. C., Molina, O. P. et al. (2001). Mutation of CDH23, encoding a new member of the cadherin gene family, causes Usher syndrome type 1D. *Nat. Genet.* **27**, 108-112.

Bork, J. M., Peters, L. M., Riazuddin, S., Bernstein, S. L., Ahmed, Z. M., Ness, S. L., Polomeno, R., Ramesh, A., Schloss, M., Srisailpathy, C. R. et al. (2001). Usher syndrome 1D and nonsyndromic autosomal recessive deafness DFNB12 are caused by allelic mutations of the novel cadherin-like gene CDH23. *Am. J. Hum. Genet.* **68**, 26-37.

Bowen, M. E., Henke, K., Siegfried, K. R., Warman, M. L. and Harris, M. P. (2012). Efficient mapping and cloning of mutations in zebrafish by low-coverage whole-genome sequencing. *Genetics* **190**, 1017-1024.

Bradford, Y., Conlin, T., Dunn, N., Fashena, D., Frazer, K., Howe, D. G., Knight, J., Mani, P., Martin, R., Moxon, S. A. et al. (2011). ZFIN: enhancements and updates to the Zebrafish Model Organism Database. *Nucleic Acids Res.* **39**, D822-D829.

Bradley, K. M., Elmore, J. B., Breyer, J. P., Yaspan, B. L., Jessen, J. R., Knapik, E. W. and Smith, J. R. (2007). A major zebrafish polymorphism resource for genetic mapping. *Genome Biol.* **8**, R55.

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

Clark, M. D., Guryev, V., Bruijn, E., Nijman, I. J., Tada, M., Wilson, C., Deloukas, P., Postlethwait, J. H., Cuppen, E. and Stemple, D. L. (2011). Single nucleotide polymorphism (SNP) panels for rapid positional cloning in zebrafish. *Methods Cell Biol.* **104**, 219-235.

Cuperus, J. T., Montgomery, T. A., Fahlgren, N., Burke, R. T., Townsend, T., Sullivan, C. M. and Carrington, J. C. (2010). Identification of MIR390a precursor processing-defective mutants in Arabidopsis by direct genome sequencing. *Proc. Natl. Acad. Sci. USA* **107**, 466-471.

Doitsidou, M., Poole, R. J., 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, A. F., Neuhauss, S. C., Malicki, J., Stemple, D. L., Stainier, D. Y., Zwartkruis, F., Abdelilah, S., Rangini, Z. et al. (1996). A genetic screen for mutations affecting embryogenesis in zebrafish. *Development* **123**, 37-46.

Dutton, K., Abbas, L., Spencer, J., Brannon, C., Mowbray, C., Nikaido, M., Kelsh, R. N. and Whitfield, T. T. (2009). A zebrafish model for Waardenburg syndrome type IV reveals diverse roles for Sox10 in the otic vesicle. *Dis. Model. Mech.* **2**, 68-83.

Ernest, S., Rauch, G. J., Haffter, P., Geisler, R., Petit, C. and Nicolson, T. (2000). Mariner is defective in myosin VIIA: a zebrafish model for human hereditary deafness. *Hum. Mol. Genet.* **9**, 2189-2196.

Fairfield, H., Gilbert, G. J., Barter, M., Corrigan, R. R., Curtain, M., Ding, Y., D'Ascenzo, M., Gerhardt, D. J., He, C., Huang, W. et al. (2011). Mutation discovery in mice by whole exome sequencing. *Genome Biol.* **12**, R86.

Geisler, R., Rauch, G. J., Geiger-Rudolph, S., Albrecht, A., van Bebber, F., Berger, A., Busch-Nentwich, E., Dahm, R., Dekens, M. P., Dooley, C. et al. (2007). Large-scale mapping of mutations affecting zebrafish development. *BMC Genomics* **8**, 11.

Giardine, B., Riemer, C., Hardison, R. C., Burhans, R., Elnitski, L., Shah, P., Zhang, Y., Blankenberg, D., Albert, I., Taylor, J. et al. (2005). Galaxy: a platform for interactive large-scale genome analysis. *Genome Res.* **15**, 1451-1455.

Gleason, M. R., Nagiel, A., Jamet, S., Vologodskaia, M., López-Schier, H. and Hudspeth, A. J. (2009). The transmembrane inner ear (Tmie) protein is essential for normal hearing and balance in the zebrafish. *Proc. Natl. Acad. Sci. USA* **106**, 21347-21352.

Goecks, J., Nekrutenko, A., Taylor, J. and the Galaxy Team (2010). Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol.* **11**, R86.

Granato, M., van Eeden, F. J., Schach, U., Trowe, T., Brand, M., Furutani-Seiki, M., Haffter, P., Hammerschmidt, M., Heisenberg, C. P., Jiang, Y. J. et al. (1996). Genes controlling and mediating locomotion behavior of the zebrafish embryo and larva. *Development* **123**, 399-413.

Gupta, T., Marlow, F. L., Ferriola, D., Mackiewicz, K., Dapprich, J., Monos, D. and Mullins, M. C. (2010). Microtubule actin crosslinking factor 1 regulates the Balbiani body and animal-vegetal polarity of the zebrafish oocyte. *PLoS Genet.* **6**, e1001073.

Guryev, V., Koudijs, M. J., Berezikov, E., Johnson, S. L., Plasterk, R. H., van Eeden, F. J. and Cuppen, E. (2006). Genetic variation in the zebrafish. *Genome Res.* **16**, 491-497.

Haffter, P., Granato, M., Brand, M., Mullins, M. C., Hammerschmidt, M., Kane, D. A., Odenthal, J., van Eeden, F. J., Jiang, Y. J., Heisenberg, C. P. et al. (1996). The identification of genes with unique and essential functions in the development of the zebrafish, Danio rerio. *Development* **123**, 1-36.

Hildebrandt, F., Heeringa, S. F., Rüschendorf, F., Attanasio, M., Nürnberg, G., Becker, C., Seelow, D., Huebner, N., Chernin, G., Vlangos, C. N. et al. (2009). A systematic approach to mapping recessive disease genes in individuals from outbred populations. *PLoS Genet.* **5**, e1000353.

Irvine, D. V., Goto, D. B., Vaughn, M. W., Nakaseko, Y., McCombie, W. R., Yanagida, M. and Martienssen, R. (2009). Mapping epigenetic mutations in fission yeast using whole-genome next-generation sequencing. *Genome Res.* **19**, 1077-1083.

Johnson, S. L., Africa, D., Horne, S. and Postlethwait, J. H. (1995). Half-tetrad analysis in zebrafish: mapping the ros mutation and the centromere of linkage group I. *Genetics* **139**, 1727-1735.

Kalay, E., Li, Y., Uzumcu, A., Uyguner, O., Collin, R. W., Caylan, R., Ulubil-Emiroglu, M., Kersten, F. F., Hafiz, G., van Wijk, E. et al. (2006). Mutations in the lipoma HMGIC fusion partner-like 5 (LHFPL5) gene cause autosomal recessive nonsyndromic hearing loss. *Hum. Mutat.* **27**, 633-639.

Kappler, J. A., Starr, C. J., Chan, D. K., Kollmar, R. and Hudspeth, A. J. (2004). A nonsense mutation in the gene encoding a zebrafish myosin VI isoform causes defects in hair-cell mechanotransduction. *Proc. Natl. Acad. Sci. USA* **101**, 13056-13061.

Kazmierczak, P., Sakaguchi, H., Tokita, J., Wilson-Kubalek, E. M., Milligan, R. A., Müller, U. and Kachar, B. (2007). Cadherin 23 and protocadherin 15 interact to form tip-link filaments in sensory hair cells. *Nature* **449**, 87-91.

Kimmel, C. B. (1989). Genetics and early development of zebrafish. *Trends Genet.* **5**, 283-288.

Kimmel, C. B., Kane, D. A., Walker, C., Warga, R. M. and Rothman, M. B. (1989). A mutation that changes cell movement and cell fate in the zebrafish embryo. *Nature* **337**, 358-362.

Kimmel, C. B., Ballard, W. W., Kimmel, S. R., Ullmann, B. and Schilling, T. F. (1995). Stages of embryonic development of the zebrafish. *Dev. Dyn.* **203**, 253-310.

Koch, B., Goold, A., Egelhoff, J. and Benton, C. (2006). Partial absence of the posterior semicircular canal in Alagille syndrome: CT findings. *Pediatr. Radiol.* **36**, 977-979.

Lander, E. S. and Botstein, D. (1987). Homozygosity mapping: a way to map human recessive traits with the DNA of inbred children. *Science* **236**, 1567-1570.

Leshchiner, I., Alexa, K., Kelsey, P., Adzhubei, I., Austin-Tse, C. A., Cooney, J. D., Anderson, H., King, M. J., Stottmann, R. W., Garnaas, M. K. et al. (2012). Mutation mapping and identification by whole-genome sequencing. *Genome Res.* **22**, 1541-1548.

Li, L., Krantz, I. D., Deng, Y., Genin, A., Banta, A. B., Collins, C. C., Qi, M., Trask, B. J., Kuo, W. L., Cochran, J. et al. (1997). Alagille syndrome is caused by mutations in human Jagged1, which encodes a ligand for Notch1. *Nat. Genet.* **16**, 243-251.

Longo-Guess, C. M., Gagnon, L. H., Cook, S. A., Wu, J., Zheng, Q. Y. and Johnson, K. R. (2005). A missense mutation in the previously undescribed gene

Tmhs underlies deafness in hurry-scurry (hscy) mice. *Proc. Natl. Acad. Sci. USA* **102**, 7894-7899.
Longo-Guess, C. M., Gagnon, L. H., Fritzsch, B. and Johnson, K. R. (2007). Targeted knockout and lacZ reporter expression of the mouse Tmhs deafness gene and characterization of the hscy-2J mutation. *Mamm. Genome* **18**, 646-656.
López-Schier, H. and Hudspeth, A. J. (2006). A two-step mechanism underlies the planar polarization of regenerating sensory hair cells. *Proc. Natl. Acad. Sci. USA* **103**, 18615-18620.
Malicki, J., Schier, A. F., Solnica-Krezel, L., Stemple, D. L., Neuhauss, S. C., Stainier, D. Y., Abdelilah, S., Rangini, Z., Zwartkruis, F. and Driever, W. (1996). Mutations affecting development of the zebrafish ear. *Development* **123**, 275-283.
Millimaki, B. B., Sweet, E. M. and Riley, B. B. (2010). Sox2 is required for maintenance and regeneration, but not initial development, of hair cells in the zebrafish inner ear. *Dev. Biol.* **338**, 262-269.
Nicolson, T. (2005). The genetics of hearing and balance in zebrafish. *Annu. Rev. Genet.* **39**, 9-22.
Nicolson, T., Rüsch, A., Friedrich, R. W., Granato, M., Ruppersberg, J. P. and Nüsslein-Volhard, C. (1998). Genetic analysis of vertebrate sensory hair cell mechanosensation: the zebrafish circler mutants. *Neuron* **20**, 271-283.
Obholzer, N., Wolfson, S., Trapani, J. G., Mo, W., Nechiporuk, A., Busch-Nentwich, E., Seiler, C., Sidi, S., Söllner, C., Duncan, R. N. et al. (2008). Vesicular glutamate transporter 3 is required for synaptic transmission in zebrafish hair cells. *J. Neurosci.* **28**, 2110-2118.
Oda, T., Elkahloun, A. G., Pike, B. L., Okajima, K., Krantz, I. D., Genin, A., Piccoli, D. A., Meltzer, P. S., Spinner, N. B., Collins, F. S. et al. (1997). Mutations in the human Jagged1 gene are responsible for Alagille syndrome. *Nat. Genet.* **16**, 235-242.
Pareek, C. S., Smoczynski, R. and Tretyn, A. (2011). Sequencing technologies and genome sequencing. *J. Appl. Genet.* **52**, 413-435.
Sakaguchi, H., Tokita, J., Müller, U. and Kachar, B. (2009). Tip links in hair cells: molecular composition and role in hearing loss. *Curr. Opin. Otolaryngol. Head Neck Surg.* **17**, 388-393.
Schibler, A. and Malicki, J. (2007). A screen for genetic defects of the zebrafish ear. *Mech. Dev.* **124**, 592-604.
Schneeberger, K., Ossowski, S., Lanz, C., Juul, T., Petersen, A. H., Nielsen, K. L., Jørgensen, J. E., Weigel, D. and Andersen, S. U. (2009). SHOREmap: simultaneous mapping and mutation identification by deep sequencing. *Nat. Methods* **6**, 550-551.
Seiler, C., Finger-Baier, K. C., Rinner, O., Makhankov, Y. V., Schwarz, H., Neuhauss, S. C. and Nicolson, T. (2005). Duplicated genes with split functions: independent roles of protocadherin15 orthologues in zebrafish hearing and vision. *Development* **132**, 615-623.
Shabbir, M. I., Ahmed, Z. M., Khan, S. Y., Riazuddin, S., Waryah, A. M., Khan, S. N., Camps, R. D., Ghosh, M., Kabra, M., Belyantseva, I. A. et al. (2006). Mutations of human TMHS cause recessively inherited non-syndromic hearing loss. *J. Med. Genet.* **43**, 634-640.
Sidi, S., Friedrich, R. W. and Nicolson, T. (2003). NompC TRP channel required for vertebrate sensory hair cell mechanotransduction. *Science* **301**, 96-99.
Smith, D. R., Quinlan, A. R., Peckham, H. E., Makowsky, K., Tao, W., Woolf, B., Shen, L., Donahue, W. F., Tusneem, N., Stromberg, M. P. et al. (2008). Rapid whole-genome mutational profiling using next-generation sequencing technologies. *Genome Res.* **18**, 1638-1642.
Söllner, C., Rauch, G. J., Siemens, J., Geisler, R., Schuster, S. C., Müller, U., Nicolson, T. and the Tübingen, 2000, Screen Consortium (2004). Mutations in cadherin 23 affect tip links in zebrafish sensory hair cells. *Nature* **428**, 955-959.
Solomon, K. S., Kudoh, T., Dawid, I. B. and Fritz, A. (2003). Zebrafish foxi1 mediates otic placode formation and jaw development. *Development* **130**, 929-940.
Starr, C. J., Kappler, J. A., Chan, D. K., Kollmar, R. and Hudspeth, A. J. (2004). Mutation of the zebrafish choroideremia gene encoding Rab escort protein 1 devastates hair cells. *Proc. Natl. Acad. Sci. USA* **101**, 2572-2577.
Sweet, E. M., Vemaraju, S. and Riley, B. B. (2011). Sox2 and Fgf interact with Atoh1 to promote sensory competence throughout the zebrafish inner ear. *Dev. Biol.* **358**, 113-121.
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 Physiol.* **52**, 716-722.
Voz, M. L., Coppieters, W., Manfroid, I., Baudhuin, A., Von Berg, V., Charlier, C., Meyer, D., Driever, W., Martial, J. A. and Peers, B. (2012). Fast homozygosity mapping and identification of a zebrafish ENU-induced mutation by whole-genome sequencing. *PLoS ONE* **7**, e34671.
Wensink, P. C., Finnegan, D. J., Donelson, J. E. and Hogness, D. S. (1974). A system for mapping DNA sequences in the chromosomes of Drosophila melanogaster. *Cell* **3**, 315-325.
Whitfield, T. T., Granato, M., van Eeden, F. J., Schach, U., Brand, M., Furutani-Seiki, M., Haffter, P., Hammerschmidt, M., Heisenberg, C. P., Jiang, Y. J. et al. (1996). Mutations affecting development of the zebrafish inner ear and lateral line. *Development* **123**, 241-254.
Zhou, Y. and Zon, L. I. (2011). The zon laboratory guide to positional cloning in zebrafish. *Methods Cell Biol.* **104**, 287-309.
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**, 427-430.

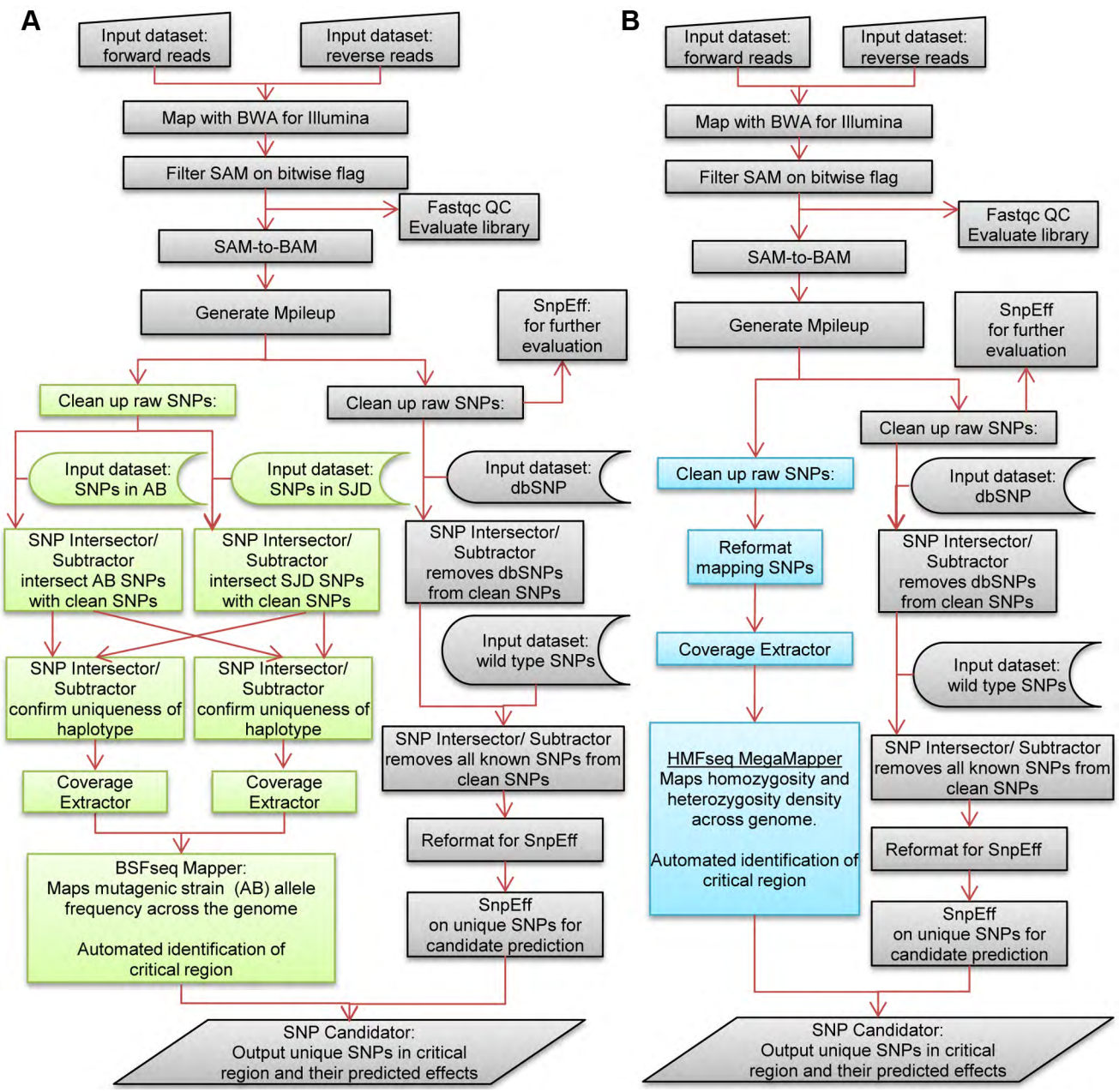

Fig. S1. MegaMapper, a zebrafish community bioinformatics resource for positional cloning by BSFseq or HMFseq. (**A**) BSFSeq flowchart. After read processing and mapping to the reference genome, BSFSeq intersects the candidate SNP library with a strain-specific SNP list to determine haplotype frequency in intersecting positions. (**B**) HMFSeq flowchart. After read processing and mapping to the reference genome, HMFSeq intersects homozygous and heterozygous SNP densities within the candidate library to determine the degree of homozygosity (hom/het ratio) in intersected positions.

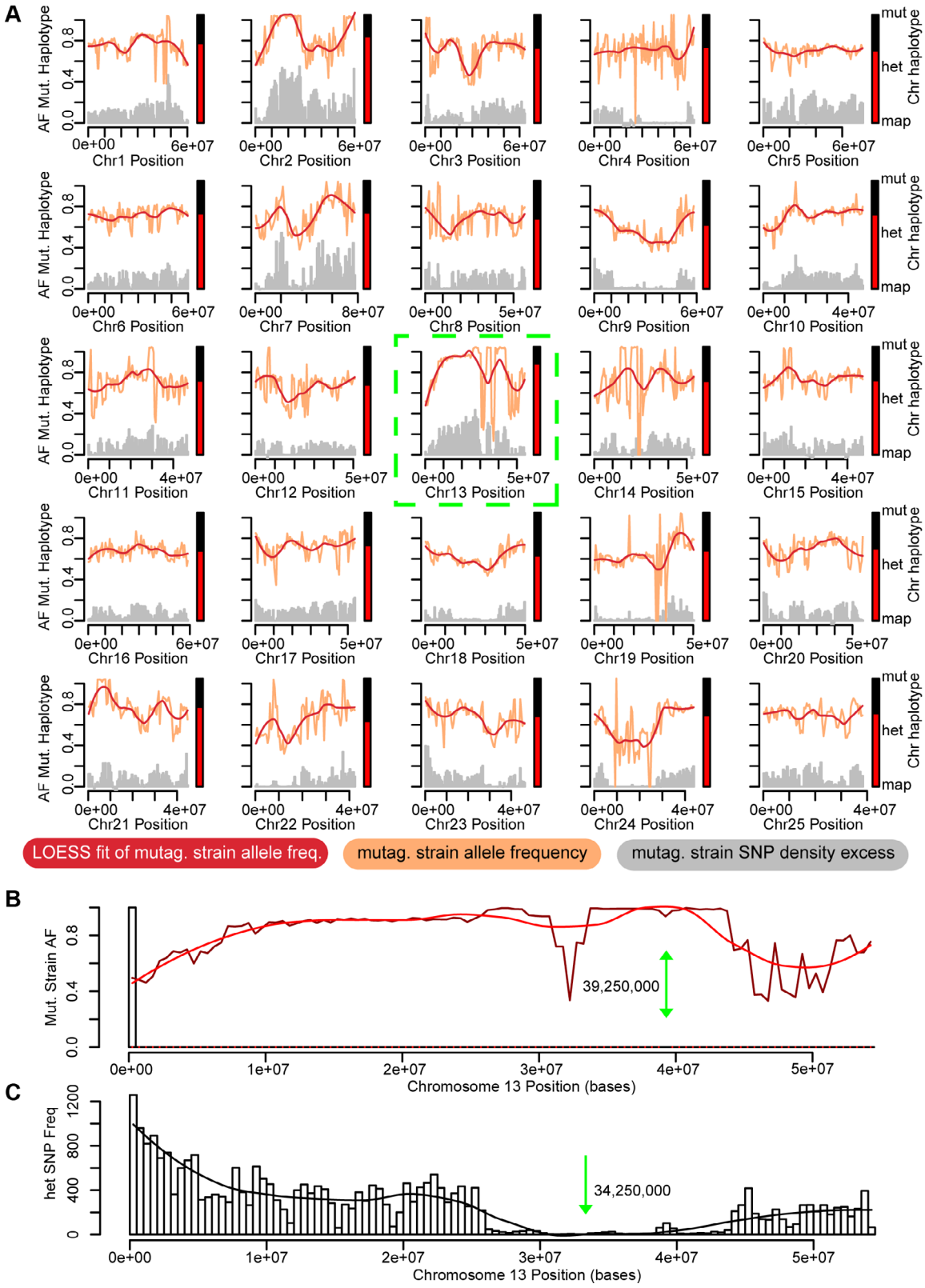

Fig. S2. Bulk segregant linkage mapping and filtering of *jj59/jagged1b*. (**A**) Mapping of regional AB haplotype allele frequencies onto individual chromosomes (AF Mut. Haplotype). Chromosomal position intervals of ~200 kb are plotted in orange. A LOESS fit of the haplotype interval data is plotted in red. Histogram data show the measured local SNP density of AB-haplotype SNPs (gray). Bars next to individual chromosomes show averaged chromosomal mutagenesis strain allele frequency (red, vs mapping strain AF in black). (**B**) Allele frequency scan of the candidate chromosome 13 (highest bulk allele frequency, green box in A) position by AB haplotype allele frequency (dark red) and the corresponding fit (red line). The maximum of the fit is indicated (green arrow). (**C**) The panel shows a SNP density histogram of all detected heterozygous SNPs on chromosome. Loss of heterozygosity is calculated by fitting to the density histogram (black line). The minimum of the fit is indicated (green arrow). After filtering of known wild-type SNPs, only one homozygous SNP remains. The predicted effects of this SNP denote a disrupted SPLICE_DONOR for *jagged1b* (Table 3).

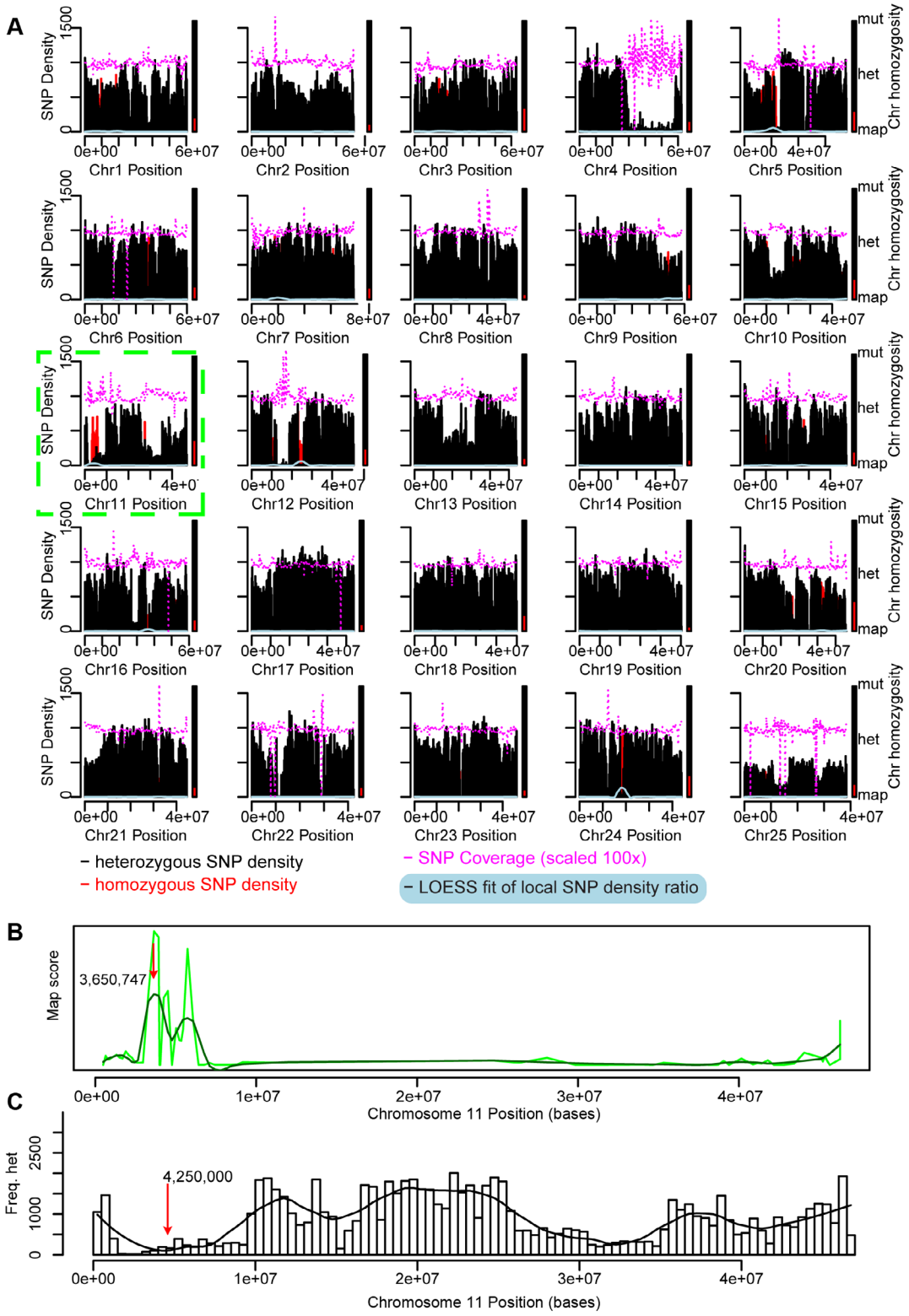

Fig. S3. Homozygosity mapping and filtering of *tm290d/lhfpl5a*. (**A**) Chromosomal SNP density and homozygosity map of *tm290d/lhfpl5a*. Red bars are homozygous SNP counts, black bars are heterozygous SNP counts. Homozygous-to-heterozygous ratios were fitted for trend recognition using local regression (LOESS; light blue). Bars next to individual chromosomes show the degree of homozygosity (red bar, vs heterozygosity in black). Chromosome 11 shows the highest degree of homozygosity and thus probably harbors the mutation. (**B**) Scan of chromosome 11 for a peak in local homozygosity and fit peaks at 3,650,747 (red arrow). (**C**) A SNP density histogram of all detected heterozygous SNPs on chromosome 11. Loss of heterozygosity is calculated by fitting to the density histogram (black line). The minimum of the fit is indicated (red arrow). The average of the resulting values of B and C were used to position the center of the critical interval. Variant effect prediction of all detected homozygous SNPs within the critical interval after filtering reveals a shortlist of only one candidate SNP remains. SNP 11_2815634_T/A introduces a stop codon at residue 80 of Lhfpl5a (Table 3).

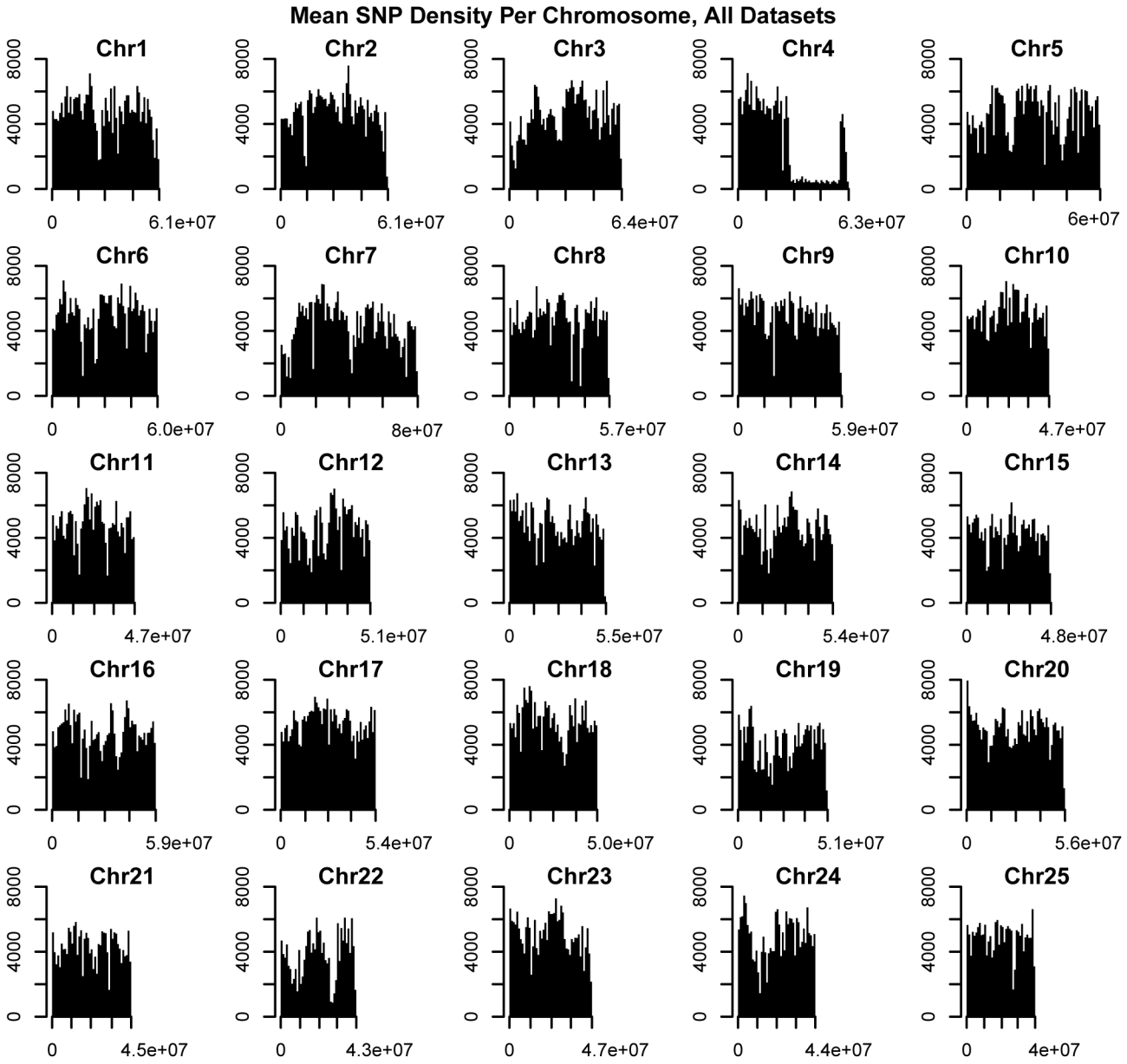

Fig. S4. Mean SNP density per chromosome revealed uneven genome-wide distribution of SNPS. The SNPs from all datasets were combined and plotted across each chromosome as SNP densities, the number of SNPs falling within the bin. Although there is a very large SNP desert on chromosome 4, every chromosome has regions of low SNP density and regions of higher SNP density that probably distort the accuracy of all mapping efforts.

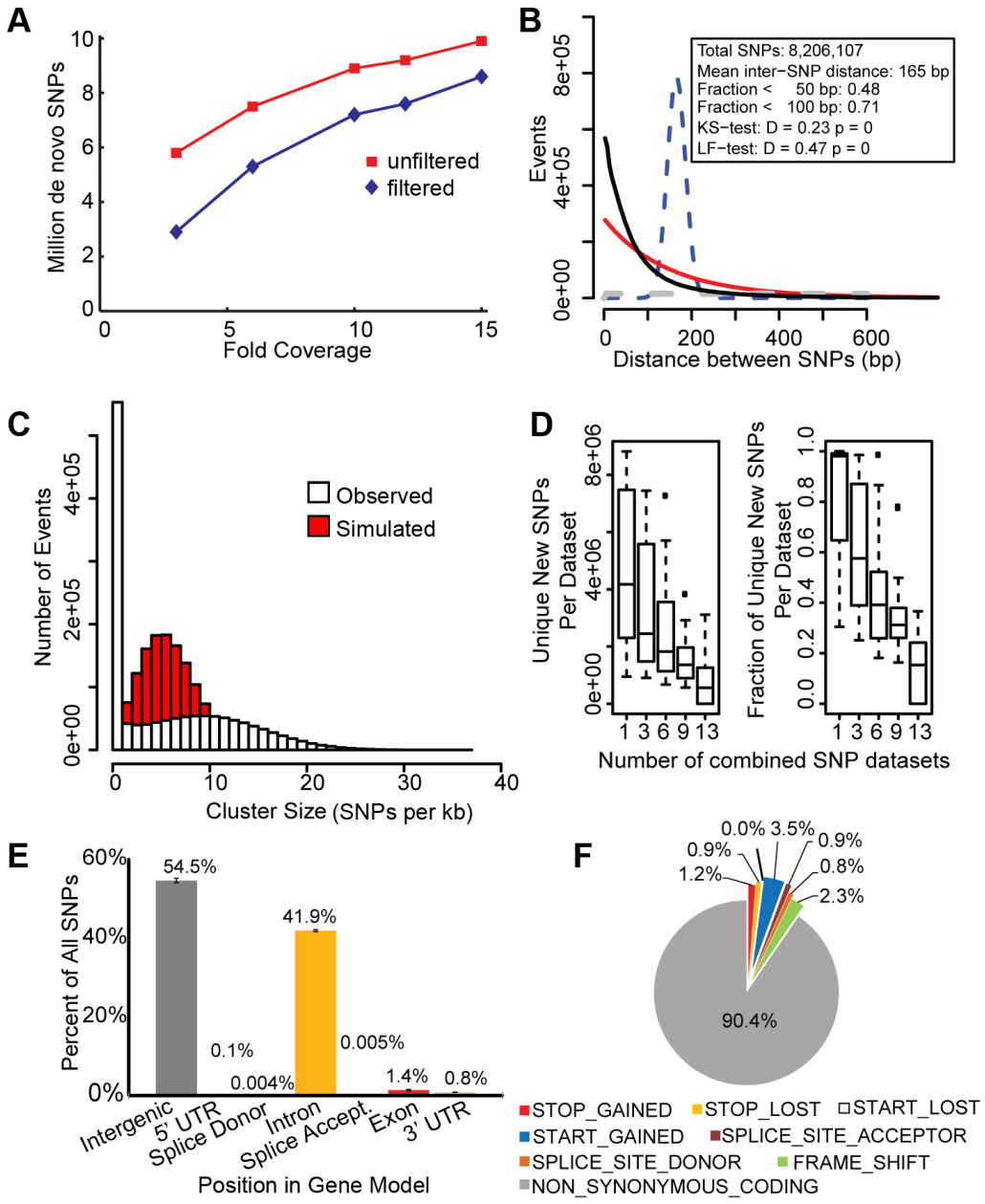

Fig. S5. Towards a saturating analysis of the zebrafish SNP universe. (**A**) Saturation of detected de novo SNPs per library at high coverage (*n*=14 libraries). (**B**) Distribution of inter-SNP distance (ISD) in sequencing data (black line) is skewed towards shorter ISD with a mean of 165 bp. Simulations (*n*=10) of the randomized distribution of an equal number of SNPs (red line) differs from real data (D=0.23, *P*=0), tight normal distribution around the mean of the measured data with an artificial standard deviation of one-eighth of the mean (blue dotted line), normal distribution around 165 bp using the measured s.d. (gray dashed line). (**C**) Histogram of observed and simulated SNP clusters (SNPs per kb) indicates that SNPs cluster more than they would by chance to produce regions of high SNP density and SNP deserts. (**D**) The number of new SNPs per genome decreases as a function of the number of genomes considered, indicating saturation. Left, absolute number of newly discovered SNPs versus number of combined datasets. Right, fraction of newly discovered SNPs versus number of combined datasets. (**E**) The distribution of detected SNPs is biased against exons and their splice site, supporting Zv9 gene annotation and SNP ascertainment (see also supplementary material Fig. S6). (**F**) Average distribution (*n*=10 datasets) of predicted non-silent SNP effects shows a small but reproducible number of deleterious effects in wild-type and mutant genomes alike according to Zv9 gene annotations.

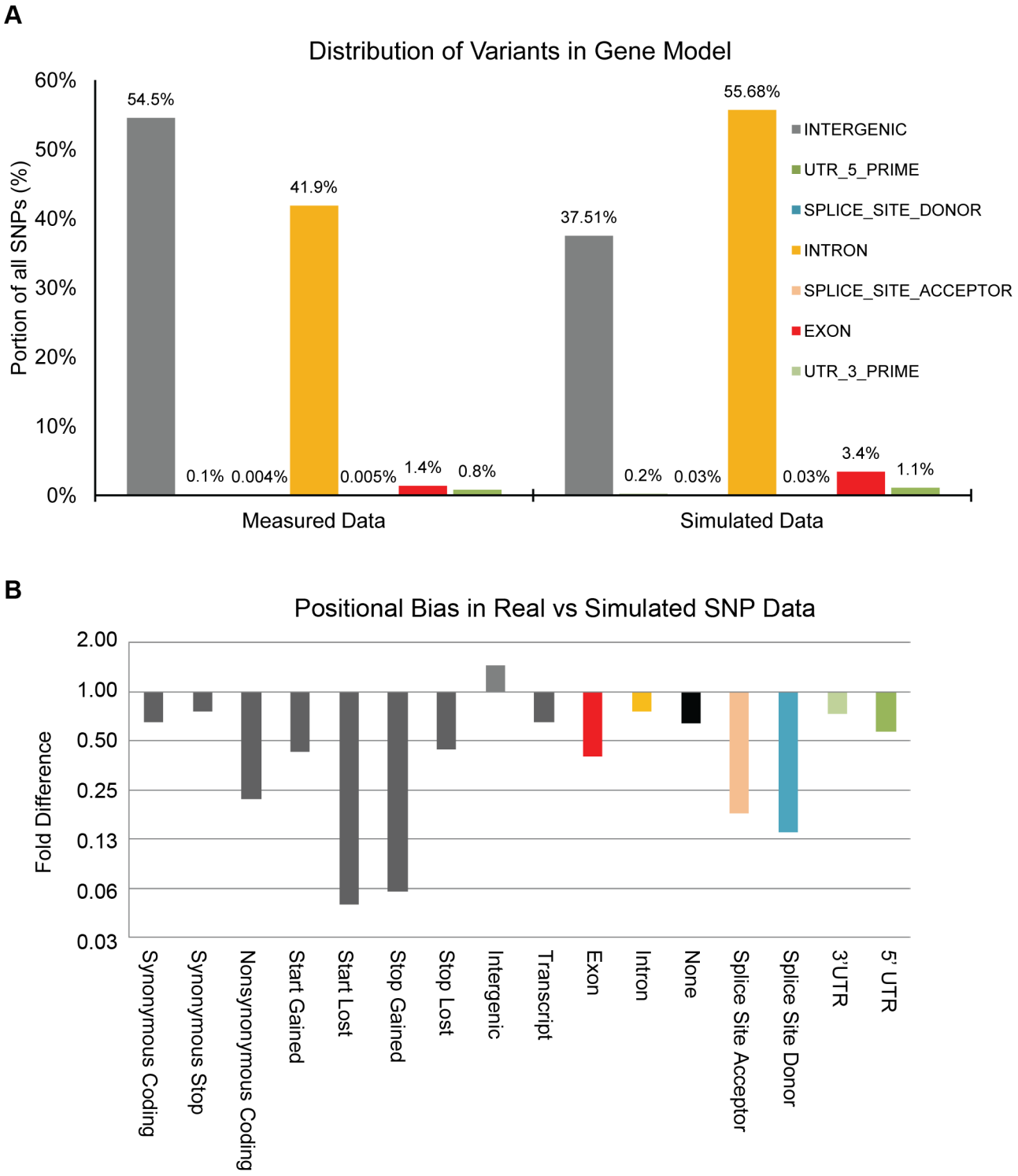

Fig. S6. A bias for intergenic regions in the distribution of variants with respect to gene architecture. We compared the distribution of the total observed SNPs to a size matched distribution of randomly simulated loci. (**A**) We observed a greater than random portion of SNPs within intergenic regions of the genome. By contrast, all genic regions were under-represented compared with our simulated set of loci. (**B**) The non-random distribution SNPs in our datasets produced a negative enrichment for variants that would alter gene expression or function.

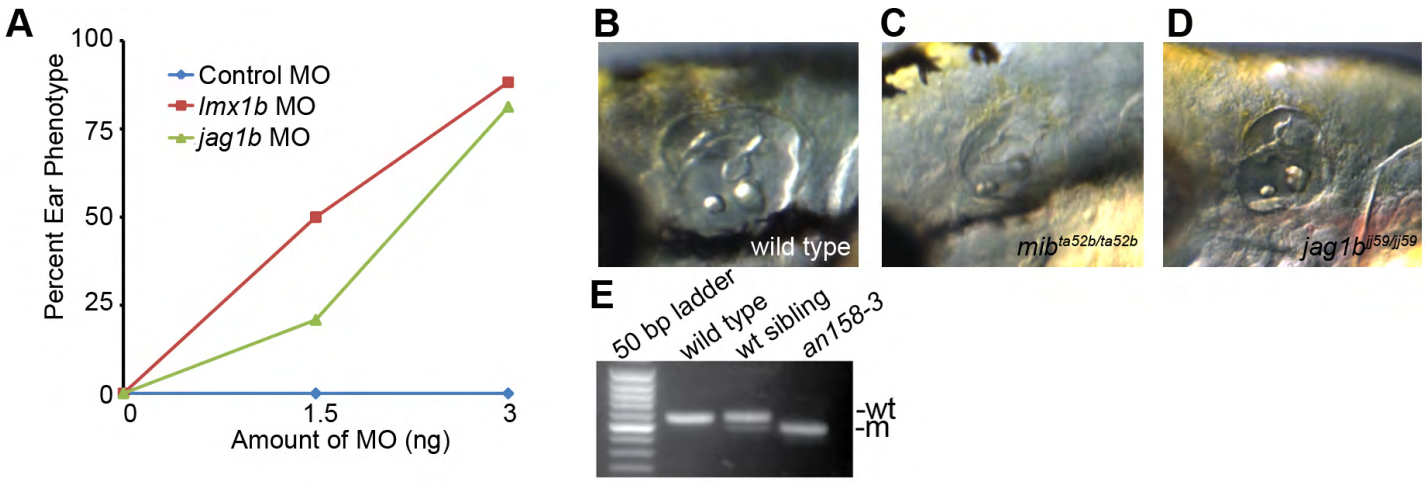

Fig. S7. Additional validation of mutants supported their computed identities. (**A**) We observed a dose-dependent phenocopy of the mutant phenotypes with morpholinos against *lmx1b* and *jag1b*. (**B-D**) Compared with wild type (B), the ear phenotypes of *mib*$^{ta52b/ta52b}$ (C) *and jag1b*$^{j59/jj59}$ (D) resemble one another and support the overlapping roles for both gene products in Notch signaling. (**E**) RT-PCR of *cdh23* transcript in wild type, sibling and mutant reveals mis-splicing in *an158-3(cdh23nl9)* transcripts.

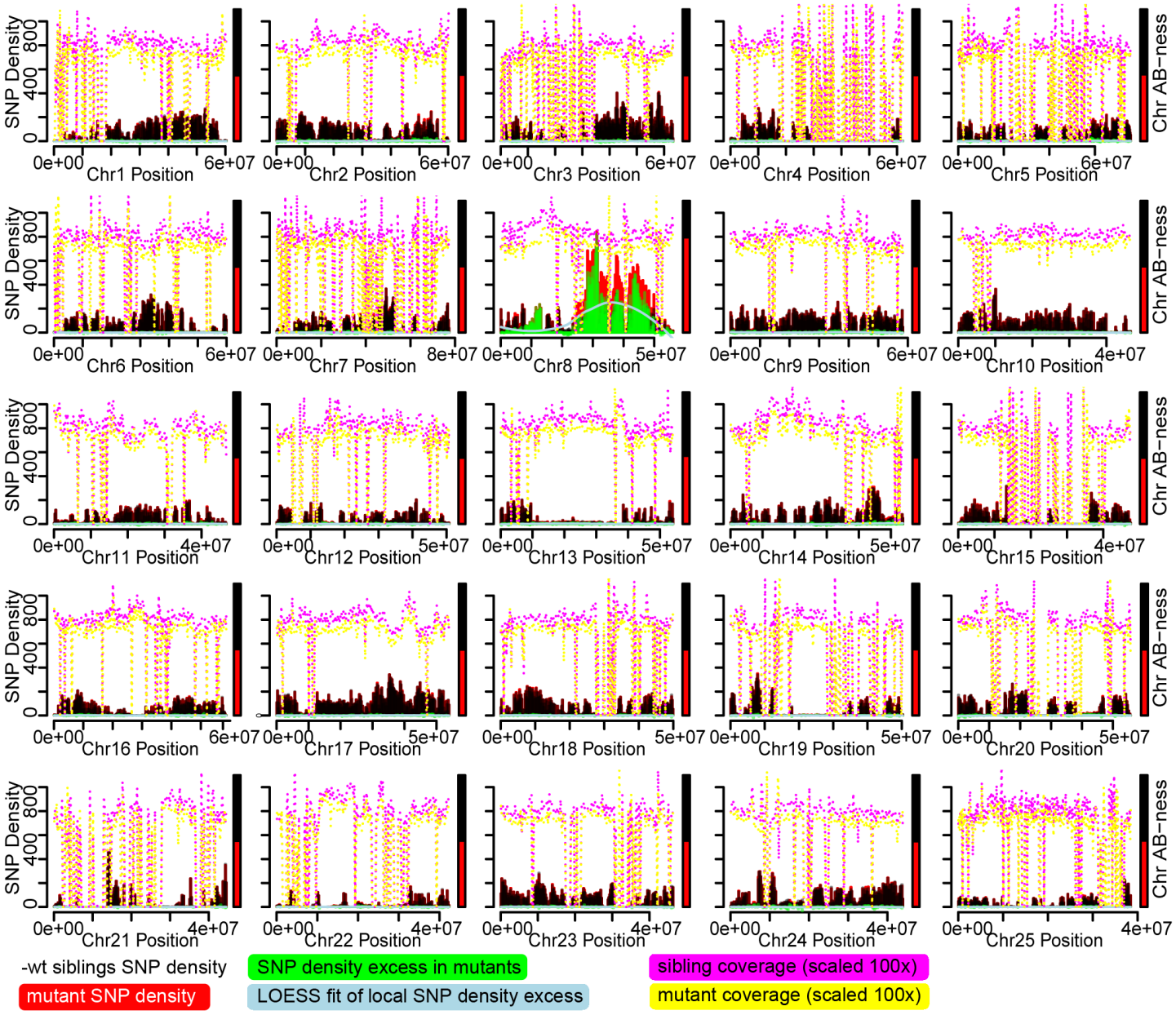

Fig. S8. Subtracting SNPs of wild-type siblings from the *jj410* mutant dataset cleaned up the mapping noise for *jj410*, but was not necessary. SNP density was plotted across the chromosomes for wild-type (SJD, black) and mutant (AB, red) SNP density. After subtraction, only chromosome 8 had a total AB density much higher than 50% (red and black bars to the right of each chromosomal plot, compare to Fig. 2A). Additionally, SNP density excess of AB (green) was only significant on chromosome 8. We also included plots of wild-type sibling coverage (magenta) and mutant coverage (yellow) to demonstrate that both libraries shared similar coverage biases. For instance, chromosome 4, which contains a large SNP desert, also has an overlapping region of intermittent coverage. The reason(s) for this are unknown, but may have something to do with biases against heterochromatin and/or elevated representation of repeats or difficult to map sequence.

Table S1. Morpholino and DNA oligo sequences.

Description	Sequence (5′ to 3′)
Standard control morpholino	CCTCTTACCTCAgTTACAATTTATA
lmx1b.1 splice site morpholino	TTGAAGGACTTACCGAGCATAACTC
jag1b splice site morpholino	AATAGTCTTTCTTACGGGAGTGGC
lmx1b.1 in situ probe forward	CCATCGACCGACTCTACTCCATGCAG
lhfpl5a ATG morpholino	CAGATAGCATTTTCGCCATGTTTGC
lhfpl5a splice site morpholino	GCTGGAGATGAAAAACACACTCAAT
lmx1b.1 in situ probe reverse	TAATACGACTCACTATAGGGAGACATCGTCTTTCAGAAGGGCTAAAC
jag1b in situ probe forward	CGAGGGCAAGAACTCCATCATTGC
jag1b in situ probe reverse	TAATACGACTCACTATAGGGAGACTCCTTTCCGACAATTGCTGTGGTG
lhfpl5a in situ probe forward	CCTGAAGATTTAGGTGACACTATAGAAGAGAAATGCTATCTGCCC AAGAGGCTGCCAAGA
lhfpl5a in situ probe reverse	GCCTGAATAATACGACTCACTATAGGGAGAACCGTACAGTTGCCC AGCGTGTATTTGTCT
lhfpl5a RT-PCR/ genotyping F	AAATGCTATCTGCCCAAGAGGCTGCCAAGA
lhfpl5a RT-PCR/ genotyping R	ACCGTACAGTTGCCCAGCGTGTATTTGTCT
an158-3/cdh23nl9 genotyping F	TGGTGCTGTAACAATGGCCCTTCA
an158-3/cdh23nl9 genotypingR	AGCTCTGCATTTAGACCCGCGTCATT
an158-3/cdh23nl9 RT-PCR For	TTCCTATTGGATCGTCTGTGTTCAGGGTGCAAGTCC
an158-3/cdh23nl9 RT-PCR R	GCCCGCTGGACCGTTATCGATGGCTTC
lmx1b.1 genotyping forward	GAAGGCTCGTCTCTGCTGTGTGGTG
lmx1b.1 genotyping reverse	CGTTATGGATGCGCTGAGACTGAATACC
jag1b genotyping forward	GACATAGATGACTGCAGCTTGAACC
jag1b genotyping reverse	GCATGTAGCTTCGTCACACTGGC

Table S1. Morpholino and DNA oligo sequences.

Table S2. Strain-specific SNP markers

Genome	SNPs vs REF	HOM SNPs vs REF	Unique (all)	Unique (hom)	Coverage (mean)
AB	13.6 E+06	9.0 E+06	4.1 E+06	2.5 E+06	12
TU	9.3 E+06	2.1 E+06	1.8 E+06	0.5 E+06	5
WIK	13.8 E+06	4.4 E+06	4.5 E+06	1.0 E+06	5
SJD	3.0 E+06	2.3 E+06	0.7 E+06	0.4 E+06	4
TLF	6.9 E+06	3.4 E+06	1.6 E+06	0.6 E+06	5
COMBINED	25.4 E+06	14.8 E+06	NA	NA	9
UNIVERSAL	0.5 E+06	0.4 E+06	NA	NA	NA
jj59	6.0 E+06	2.3 E+06	1.0 E+06	0.4 E+06	7
*jj410*_WT	9.2 E+06	2.7 E+06	1.6 E+06	0.4 E+06	8
*jj410*_Mut	8.9 E+06	2.70E+06	1.5 E+06	0.4 E+06	8
a158-3	4.2 E+06	1.6 E+06	0.6 E+06	0.3 E+06	5
tm290D	7.5 E+06	1.1 E+06	1.3 E+06	0.2 E+06	8
Average	8.3 E+06				
Standard deviation	3.5 E+05				

NA, not applicable.

Table S2. Strain-specific SNP markers. This table summarizes the SNPs that we identified in the wild-type and mutant strains that we sequenced and/or analyzed. SNPs were identified relative to the Zv9 genome assembly from Sanger, which is based on TU. Homozygous SNPs were defined as having at least three mapped sequences at the loci and all sequences having the same variant. Unique variants were defined as SNPs present only in that particular dataset and absent from all other dataset. The mean coverage is an effective coverage calculated by the Snpeff tool.

Table S3. Number of detected SNPs per genome

Library	DP3.filter.SNPs	ab.initio.calls	PF.Reads	cov
*AB*_SANGER	8,637,221	9,855,280	270,000,000	15
*AB*_HARRIS	7,579,548	10,141,300	90,000,000	4.6
*TU*_SANGER	2,378,304	3,946,057	186,000,000	11
*TUB*_HARRIS	5,550,176	7,284,290	97,000,000	5.1
*TUG*_HARRIS	3,954,060	6,035,650	81,000,000	4.1
*TLF*_HARRIS	6,845,445	9,862,938	91,000,000	3.8
*WKB*_HARRIS	7,720,936	10,757,717	96,000,000	4.1
*WKG*_HARRIS	7,565,379	10,725,557	81,000,000	4
*SJD*_MEGASON	3,184,219	8,481,074	97,670,816	6
*jj410_sjd*_sib_MEGASON	9,224,092	11,808,490	141,622,193	8
*jj410_sjd*_mut_MEGASON	8,926,680	11,593,392	131,531,446	8
*jj59*_MEGASON	6,017,281	11,843,236	117,941,618	7
*tm290d*__MEGASON	7,520,301	9,748,539	129,792,188	7
*a158-3*__MEGASON	4,167,428	8,260,138	126,796,756	7

Table S3. Number of detected SNPs per genome. This table summarizes each dataset used in the study. DP3 filter SNPs were sequenced at least three times in the particular library. The number of SNPs prior to filtering for number of times sequenced (depth) was the ab.initio.calls. PF reads passed our sequencing and mapping quality filters. The coverage (cov) is an effective coverage calculated by the Snpeff tool.

Table S4. Effect of 10% contamination of wild-type sequences in the mutant pool

Sample	Map mode	Chr	Mapping position	Distance from lesion
Control	Het_Valley	8	36,000,000	–1,870,889
Control	Density_ratio	8	34,726,299	–597,188
Control	Compromise	8	35,363,150	–1,234,039
Trial 1	Het_Valley	8	37,000,000	–2,870,889
Trial 1	Density_ratio	8	34,726,299	–597,188
Trial 1	Compromise	8	35,863,150	–1,734,039
Trial 2	Het_Valley	8	36,250,000	–2,120,889
Trial 2	Density_ratio	8	34,726,299	–597,188
Trial 2	Compromise	8	35,488,150	–1,359,039
Trial 3	Het_Valley	8	36,750,000	–2,620,889
Trial 3	Density_ratio	8	34,726,299	–597,188
Trial 3	Compromise	8	32,738,150	1,390,961

Original measurement (no contamination)				
Map Mode	**POS**	**STDEV**	**DIST.F.LESION**	**SHIFT (bp)**
Het_Valley	36,250,000	NA	–1,870,889	NA
Density_ratio	34,726,299	NA	–597,188	NA
Compromise	34,707,856	NA	–1,234,039	NA
10% contamination, simulated (three trials)				
Map Mode	**AVG**	**STDEV**	**DIST.F.LESION**	**SHIFT (bp)**
Het_Valley	36,666,667	381,881	–2,537,556	–666,667
Density_ratio	34,726,299	0	–597,188	0
Compromise	34,696,483	1,706,300	–567,372	666,667

SNP effect

Percentage contamination	Number	Chr	Pos	REF	ALT	Qual	DP	AF	Type
0	0	8	34129111	T	A	222	11	1.00	hom
10	1	8	34129111	T	A	224	13	0.85	het
10	2	8	34129111	T	A	222	11	1.00	hom
10	3	8	34129111	T	A	225	17	0.71	het

Table S4. Effect of 10% contamination of wild-type sequences in the mutant pool. We computationally simulated mis-sorting of embryos by taking adding 10% additional randomly chosen paired end reads from the wild-type *jj410* library into the mutant *jj410* library. Simulation was repeated three times.

Table S5. SNP effects by category and genome. This table summarizes the effects of all homozygous SNPs for each genome library we analyzed. **Download**