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

Nikolaus Obholzer1,*, Ian A. Swinburne1,*, Evan Schwab1, Alex V. Nechiporuk3, Teresa Nicolson2 and Sean G. Megason1,±

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. 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 \( \text{cdh23} \) and \( \text{lmx1b} \) cause abnormal ear morphogenesis. With HMFseq we validated that the disruption of \( \text{jagged1b} \) abolishes the ear’s sensory functions and identified a candidate lesion in \( \text{lhfp5a} \) 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; Ernst 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,
Cloning by sequencing

**Materials and Methods**

**Zebrafish husbandry**

The $jj39$ and $jj410$ mutants, generated in the AB background (Schibber and Malicki, 2007), were obtained from ZIRC (Eugene, OR, USA) and crossed into SJD for mapping. The minicells were generated in AB, then hybridized with TU/TLF. We sequenced mapping F2s, but used homozygosity instead of haplotype to map the mutation. By contrast, the zebrafish 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 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 library were obtained, for average genome coverage of 6-14×.

**MegaMapper pipeline**

Sequencing reads were mapped to the unmapped 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 reads with a mismatch rate of >0.2. Mapping 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 genome-wide 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 remaining. 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 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, Perl 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/MegaGen/ 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 1X 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 1X 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 Technologies).

**Zebrafish husbandry**

The $jj39$ 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. The $jj39$ 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, the $jj410$ 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 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 library were obtained, for average genome coverage of 6-14×.
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; Hildebrant 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/+ or m/m) by in-crossing and checking for phenotypically mutant progeny (+/m or 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 tedious 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 outcroses to prevent low fecundity. AB and TU are not highly inbred strains; therefore, they were expected to contain intrastain 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 for at least 3 hours at 4°C, rinsed again and finally mounted and imaged on a ZEISS LSM710 confocal microscope.

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 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 across 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, *jj*59, which gave similar results (supplementary material Fig. S2).

**Homozygosity mapping**

The accuracy of the heterozygous SNP density analyses and the presence of a sizable intrastain 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
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...
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 \textit{cadherin 23} (\textit{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 \textit{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

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

In order to determine the impact of wild-type contamination on the mapping results, we calculated the standard deviation (s.d.) for the distance from identified lesions in the four mutants. The standard deviation ranged from 1.4 Mb to 1.7 Mb, indicating that the filters were effective in reducing the number of false positive lesions.

### The zebrafish SNP universe

The zebrafish SNP universe is a comprehensive database of single nucleotide polymorphisms (SNPs) that have been identified in zebrafish. This database is derived from a combination of traditional sequencing approaches and high-throughput sequencing technologies. The zebrafish SNP universe is continuously updated with new data as they become available, ensuring that it remains one of the most comprehensive resources for genetic research in zebrafish.

### Power of filtering to identify causative lesions

The power of filtering to identify causative lesions is a critical aspect of genetic mapping. By applying filters such as heterozygosity minimum and SNP density ratio fit, the number of candidate SNPs can be significantly reduced, allowing for a more accurate identification of the causative lesion. The compromise fit is the average of the heterozygosity minimum and the SNP density ratio fit, providing a balanced approach to filtering.

### 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). We used traditional methods to map these lesions, including PCR amplification, sequencing, and Southern blotting.

We experimentally validated the other three predicted mutations. For jj410 and jj59, the allele was shorter in the anterior-posterior axis compared with wild type. For tm290d, the otic vesicle was shorter in the anterior-posterior axis. All putative lesions were re-sequenced using traditional Sanger sequencing of pools and individuals.

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.
and the remaining 5/9 heterozygous for \textit{jj410}, and 4/10 were homozygous wild type and the remaining 6/10 heterozygous for \textit{jj59}. Additional validation came from the expression patterns of \textit{lmx1b} and \textit{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 \textit{lmx1b} or \textit{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 \textit{jj59} mutants were similar to those seen in the \textit{mind bomb} zebrafish mutant in which Notch signaling is globally disrupted (supplementary material Fig. S7).

To validate our prediction of the mutation in \textit{cdh23} in \textit{an158-3}, we performed a complementation cross with a previously published allele of \textit{cdh23}, \textit{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 locus using Sanger sequencing in the mutant population (Fig. 4S). We confirmed linkage by genotyping individuals: 4/4 \textit{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 \textit{jj410}, \textit{jj59} and \textit{an158-3} are caused by mutations in \textit{lmx1b}, \textit{jag1b} and \textit{cdh23}, respectively. In accordance with ZFIN conventions, these alleles are now \textit{lmx1b}\textit{jj410}, \textit{jag1b}\textit{jj59} and \textit{cdh23}.

**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.](image)

**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; Post., posterior projection; Vnt., ventral projection. (B) Semicircular canal projections are abnormally shaped in \textit{jj410} ears at 65 hpf. (C) \textit{lmx1b} morpholino knockdown recapitulates \textit{jj410} ear phenotype at 65 hpf. (D) \textit{lmx1b} antisense in situ probe marks the inner ear and concentrates in semicircular canal projections. (E) Severe reduction of \textit{lmx1b} transcript levels in a \textit{jj410} mutant larval inner ear at 65 hpf. (F) Sequencing trace of the \textit{jj410} genomic lesion in \textit{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) \textit{jj59} inner ears are shorter in the anterior-posterior axis and semicircular canal projections are malformed at 65 hpf. (I) \textit{jag1b} morpholino knockdown recapitulates \textit{jj59} ear phenotype at 65 hpf. (J) \textit{jag1b} antisense in situ probe marks the inner ear and concentrates in semicircular canal projections. (K) Severe reduction of \textit{jag1b} transcript levels in a \textit{jj59} mutant larval inner ear at 65 hpf. (L) Sequencing trace of the \textit{jj59} genomic lesion in \textit{jag1b}; arrow denotes causative mutation. (M) YO-PRO1 uptake by neuromast hair cells in wild-type larvae. (N) \textit{an158-3(cdh23)nl9} mutant larval neuromast hair cells fail to take up YO-PRO1. (O) Failure to complement YO-PRO1 uptake in \textit{an158-3} x \textit{tc317e} larval hair cells. (P) Intact inner ear hair cell bundles in the wild type. (Q) Hair cell bundles are splayed in \textit{an158-3(cdh23)nl9} mutant inner ear hair cells. (R) Hair cell bundles show splayed phenotype and thus fails to complement in \textit{an158-3} x \textit{tc317e} mutant inner ear. (S) Sequencing trace of the \textit{an158-3(cdh23)nl9} genomic lesion in \textit{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 fewMb) 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 j410 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 Abrecht 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 HL097599], 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.
References


