RNA-seq–based mapping and candidate identification of mutations from forward genetic screens

Adam C. Miller,1,3 Nikolaus D. Obholzer,2 Arish N. Shah,1 Sean G. Megason,2 and Cecilia B. Moens1,3

1Division of Basic Sciences, Fred Hutchinson Cancer Research Center, Seattle, Washington 98109, USA; 2Department of Systems Biology, Harvard Medical School, Boston, Massachusetts 02115, USA

Forward genetic screens have elucidated molecular pathways required for innumerable aspects of life; however, identifying the causal mutations from such screens has long been the bottleneck in the process, particularly in vertebrates. We have developed an RNA-seq–based approach that identifies both the region of the genome linked to a mutation and candidate lesions that may be causal for the phenotype of interest. We show that our method successfully identifies zebrafish mutations that cause nonsense or missense changes to codons, alter transcript splicing, or alter gene expression levels. Furthermore, we develop an easily accessible bioinformatics pipeline allowing for implementation of all steps of the method. Overall, we show that RNA-seq is a fast, reliable, and cost-effective method to map and identify mutations that will greatly facilitate the power of forward genetics in vertebrate models.

[Supplemental material is available for this article.]
Results

RNA-seq sample preparation and sequencing for BSA

RNA-seq–based and WGS-based BSA mapping and candidate identification are similar except that with RNA-seq the sample being sequenced is limited to the genes expressed at the time of RNA extraction. To test the applicability of RNA-seq to BSA mapping, we performed analysis on a number of known, independent, N-ethyl-N-nitrosourea (ENU)-induced mutations in the zebrafish: two nonsense mutations, hoxb1b1b1219 (data not shown) and nhsl1b1b131 (Walsh et al. 2011); a mutation that causes splicing defects, vangl2m209 (Jessen et al. 2002); and a nonsense mutation causing nonsense-mediated decay, egr2b2227 (Monk et al. 2009). For each mutation, individual pairs of heterozygous mutants were crossed, and mutant progeny were selected based on their known phenotypes, just as would be done in the case of an unknown mutant. Separate pools with equal numbers (from eight to 80) (Table 1) of mutant (+/-C0/C0) and wild-type (+/+/+ or +/+) siblings were made (Supplemental Fig. 1). We reasoned that RNA extracted from embryos soon after the first appearance of the mutant phenotype would have the best chance of including transcripts carrying the causal mutation. We therefore performed RNA extractions on mutant and sibling pools directly after scoring for each phenotype of interest (stages ranging from 2 to 5 days post-fertilization, depending on the phenotype; see Methods). Sequencing libraries were prepared using standard procedures. Briefly, total RNA was extracted, mRNAs were polyA selected and chemically fragmented, cDNA was prepared, and sequencing libraries were created for mutant and wild-type pools. Each pool was uniquely barcoded, allowing for multiplexing samples from several different mutants during sequencing. Sequencing was performed on an Illumina HiSeq 2000 machine with six libraries (three sibling/mutant pairs) per lane, resulting in an average of 43 million 50-bp paired-end reads per sample (Table 1).

RNA-seq data processing and linkage mapping

Reads from each sibling/mutant pair were independently aligned to the zebrafish genome (Zv9.63) using TopHat/Bowtie, an intron and splice aware aligner (Supplemental Fig. 2; Trapnell et al. 2009). From the aligned data sets, we first identified SNPs within the wild-type sequence (using SAMtools mpileup and bcftools) (Li et al. 2009) that would serve as useful markers to test linkage within the mutant data (Supplemental Figs. 1C, 2). Within any genome, SNPs

Table 1. Characteristics of RNA-seq–based mapping experiments

| Mutation type/effect | hoxb1b | nhsl1b | vangl2 | egr2b |
|----------------------|--------|--------|--------|--------|
| No. of embryos in each pool | 20 | 40 | 80 | 8 |
| No. of 50-bp wild-type reads greater than Q30 (million) | 48.5 | 41.5 | 35.4 | 47.9 |
| No. of wild-type bases (Gb) | 2.4 | 2.1 | 1.8 | 2.4 |
| No. of 50-bp mutant reads greater than Q30 (million) | 46.7 | 36.0 | 38.1 | 24.1 |
| No. of mutant bases (Gb) | 2.3 | 1.8 | 1.9 | 1.2 |
| No. of SNPs wild-type greater than two reads | 630,994 | 564,915 | 506,331 | 508,872 |
| No. of SNPs mutant greater than two reads | 636,480 | 523,281 | 545,295 | 333,315 |

Mapping

| No. of markers (greater than 25-fold/25%) | 40,203 | 33,987 | 27,945 |
| Linked region (Mb) | 11.6 | 7.9 | 6.5 |

Candidate identification

| No. of 50-bp mutant reads greater than Q30 (thousand) | 424 | 245 | 205 |
| Average coverage of genes | 32 | 33 | 40 |
| Mode coverage of genes | 2 | 3 | 4 |
| % Genes covered at greater than fourfold | 71 | 79 | 84 |
| % Genes covered two- to fourfold | 26 | 19 | 14 |
| % Genes covered less than twofold | 3 | 2 | 2 |
| No. of homozygous SNPs greater than two reads | 3,851 | 2,175 | 1,922 |

No. of SNPs left after filtering | 662 | 317 | 283 |
| No. of SNPs affecting coding | 46 | 18 | 19 |
| No. of SNPs | 1 | 1 | 1 |
| No. of Missense | 6 | 1 | 1 |
| No. of Isoforms altered | 1 | 0 | 0 |
| No. of expression levels altered | 0 | 0 | 0 |

(NMD) Nonsense mediated decay; (SNP) single nucleotide polymorphisms; (nd) not determined; (Q30) quality score with an accuracy of 99.9%.

At least one forward alternative and one reverse alternative read.

aGreater than 25-fold coverage at SNP with >25% heterozygosity.

bRegion defined as having an average marker frequency within 1% of peak marker frequency.

cValues from mutant RNA-seq data within the linked region.

dKnown wild-type SNPs removed from further consideration.

ePredicted by Variant Effect Predictor from SNPs.

fAssessed manually using Integrative Genomics Viewer.

Known lesion included.

Predicted by Cuffdiff from aligned reads, greater than twofold change.
are most likely to be present in intergenic/intronic regions that are not represented in RNA-seq data. However, we find an average of ~500,000 SNPs within each of our transcriptome pools (Table 1), most of these residing within UTRs. In contrast to the relatively low coverage of WGS, the highly expressed genes within the transcriptome allow for the identification of high-quality SNP markers directly in the parental background under investigation; these markers provide highly reliable mapping information. We therefore identified SNPs within each wild-type sibling pool covered by at least 25 reads, of which at least 25% of the calls represented an alternative allele (using the custom R script RNAmappe.R) (Supplemental Fig. 2). This resulted in an average of 40,000 high-confidence markers per experiment (Table 1) that were then used to interrogate the mutant RNA-seq data for regions of the genome linked to the mutation of interest (Supplemental Fig. 1).

In bulk RNA extracted from a pool of many animals with the same mutant phenotype, the mutation underlying the phenotype of interest, as well as linked regions of the genome, will be homozygous. In contrast, due to recombination during meiosis and independent chromosome assortment, regions unlinked to the mutation (both on the same and independent chromosomes) will be heterozygous (light and dark gray bars in Supplemental Fig. 1B). Thus the SNP marker frequency at and near the mutation will be 1 (all alleles are the same), and this frequency will gradually decline with increasing genetic distance from the mutation (Supplemental Fig. 1B–D). We therefore calculated the allele frequency within the mutant RNA-seq data at the positions identified as high-quality markers and then used a sliding window of 50 neighboring SNPs (average window size of 1.9 Mb, average step size of 37.5 kb; see Methods) to average this frequency and plot it against chromosome position. This allowed for the identification of regions of linkage (Supplemental Fig. 1D; Fig. 1).

Validation of RNA-seq-based mutation mapping

We tested our RNA-seq-based mapping strategy on four known, independent mutations (hoxb1bb1219, nhsl1bb131, vangl2m209, egr2bf h227). We found that in each case the peak of highest allele frequency was centered on the known mutation (Fig. 1). For each experiment, the average allele frequency on the linked chromosome steadily rose until reaching its highest frequency surrounding the known locus (Fig. 1B). In most experiments, the average allele frequency reached greater than 0.98 (homozygous = 1) surrounding the known locus (Fig. 1). In contrast, the highest frequency on unlinked chromosomes never exceeded 0.89, and the average was ~0.65 (the average frequency of unlinked SNPs is higher than 0.5 because some SNPs are heterozygous in both parents, giving an allele frequency of 0.5, while others are homozygous in one of the two parents, giving an allele frequency of 0.75) (Fig. 1A). In one case, egr2bf h227, the highest peak frequency in the genome was 0.93, yet this peak surrounded the egr2b locus (Fig. 1B). This lower peak allele frequency is likely due to the missorting of wild-type animals into the mutant pool. We examined this missorting idea directly by computationally adding reads from the wild-type pool into the mutant data and then performing the mapping experiment. We used the hoxb1bb1219 experiment because it contained a clear region of homozygosity. The method still provided a single mapping peak even after being “contaminated” with up to 30% of wild-type reads within the mutant pool (Supplemental Fig. 3). Thus even some limited missorting of wild-type individuals into the mutant pool can be tolerated.

The size of the linked region from a mapping experiment is expected to decrease with an increase in the number of mutant animals pooled due to an increased likelihood of recombination between the causative mutation and nearby SNP markers. We tested this prediction by sequencing pools of 20, 40, and 80

Figure 1. RNA-seq-based mapping identifies single peaks of linkage centered on the known mutations in all experiments. (A,B) Rows represent individual experiments and are labeled by genotype and the number of mutant embryos used for mapping. (A) Genome-wide mapping data. The average frequency of mutant markers (black marks) is plotted against genomic position. In each case, a single region emerges with an allele frequency near one (red arrow). Each chromosome is separated by vertical lines and labeled at the bottom. (B) Detail of the chromosome containing the linked interval for a given experiment (row). The average frequency of mutant markers (green discs) is plotted against chromosomal position. A red box marks each region of linkage, and a red line marks the position of the known mutation. Each tick mark on the x-axis represents 10 Mb. Each y-axis is the same as in A, first row.
input embryos, we observed smaller homozygous intervals than
of embryos used in each RNA-seq–based mapping experiment. The
most'' and ''rightmost'' position with an average mutant allele
7.9, and 6.5 Mb, respectively (linkage was defined by the ''left-
Table1). At the
32%, while from 40 to 80, there was an 18% decrease (Fig. 2;
creasing from 20 to 40 embryos decreased the linked region by
map mutations, pooling more mutants is advisable to minimize
hoxb1bb1219 mutant embryos and found that, as expected, in-
creasing the number of embryos decreased the size of linkage: 11.6,
7.9, and 6.5 Mb, respectively (linkage was defined by the ''left-
most'' and ''rightmost'' position with an average mutant allele
frequency within 1% of the peak frequency). For hoxb1b1219, in-
creasing from 20 to 40 embryos decreased the linked region by
32%, while from 40 to 80, there was an 18% decrease (Fig. 2; Table1). At the
vangl2m209 and egr2b"h227" loci, given the number of input embryos, we observed smaller homozygous intervals than
would be predicted by the hoxb1b1219 experiments (37 and 30
embryos, 2.3 and 6.6 Mb, respectively) (Fig. 2); this likely reflects
the nonhomogeneous rate of recombination across the zebrafish
genome (Bowen et al. 2012). Since in some cases it is difficult to
acquire high numbers of mutant embryos, we also tested whether
mutations could be mapped using a small number of mutant
embryos. For rhs1b"m111", we used pools of eight mutant and eight
wild-type embryos with the mapping experiment resulting in
a single, large region of linkage (~40 Mb) surrounding the known
locus (Fig. 1). So while very few embryos can be used to accurately
map mutations, pooling more mutants is advisable to minimize
the size of the linked region. However, we found that increasing
the number of embryos beyond 40 led to diminishing returns in
reducing the mapped region size (Fig. 2B).
While each mapping experiment identified a single region of
linkage that was centered on the known mutation, as previously
described (Leshchiner et al. 2012), we found unexpected de-
flections from homozygosity within several linked regions; these
are likely due to misplaced contigs in the current assembly of the
zebrafish genome, which place unlinked SNPs that have an allele
frequency less than 1.0 into the region of homozygosity. Addi-
tionally, while our mapping produced single mapping peaks, it is
possible to identify regions of homozygosity due to shared lineage
instead of due to linkage to the mutation, particularly in inbred
backgrounds (Bowen et al. 2012); thus mutations are often out-
crossed to mapping strains. Our mutants were maintained in a
variety of backgrounds (see Methods), but we note that the
hoxb1b1219 and egr2b"h227" alleles were generated in a *AB back-
ground, were maintained through outcrosses to the *AB back-
ground, and were in the F3 generation post-mutagenesis; in species
with high intrastrain polymorphism like the zebrafish, it is there-
fore possible to use RNA-seq to map mutations from forward
screens without outcrossing to mapping strains, although out-
crossing does provide a higher frequency of high-quality markers
(Supplemental Fig. 4). Overall, the RNA-seq mapping strategy
provided robust mapping of mutations to correct regions of the
genome.

Identification of candidate deleterious SNP mutations
The most powerful aspect of WGS-based mapping is that it has the
potential to directly identify causal mutations within the homo-
zygous interval. After identifying a region of linkage, we revisited
the mutant RNA-seq data and extracted all SNPs within the region
(using the custom R script RNAIdentifie.R) (Supplemental Fig. 2). A
concern in using RNA-seq data is that it may not sequence the
mutant transcript of interest given that only genes expressed at the
time of RNA extraction are captured. We found that, on average,
62% of genes within our homozygous intervals are covered by
greater than four sequencing reads, 32% are covered by two to
four reads at each nucleotide (average mode = 2.17), while the
remaining 8% of genes are covered at levels below twofold (Ta-
ble1). Thus most genes are sequenced at levels that allow for the
identification of candidate mutations. Furthermore, since we iso-
late RNA at the time the mutant phenotype first emerges, it is
likely that the transcript carrying the lesion of interest will be
detected. In support of this idea, we found either the altered
transcript or the effect of the mutation on transcript levels directly
in the RNA-seq data for each of the known mutations (Fig. 3, see
below).

Within the linked region of each mapping experiment, we
used the RNA-seq data to analyze the number of SNP changes that
could represent mutations of interest if mapping unknown mu-
tants. We found that, on average, 275 alternative SNPs became
detected. In support of this idea, we found either the altered
transcript or the effect of the mutation on transcript levels directly
in the RNA-seq data for each of the known mutations (Fig. 3, see
below).

![Image](image-url)

**Figure 2.** Increasing the number of embryos in an RNA-seq-based
mapping experiment decreases the linkage size of the mapped region.
(A) Detail of chromosome 12 containing the linked interval for each
hoxb1b1219 mapping experiment. Each row is labeled with the number
of embryos used in the experiment. The average frequency of mutant
markers (green discs) is plotted against chromosomal position. A red box
marks each region of linkage, and a red line marks the position of the
hoxb1b gene; linkage was defined as the region between the "leftmost"
and "rightmost" positions within 1% of homozygosity. Each y-axis is the
same as in the first row. (B) Comparison of linked regions to the number
of embryos used in each RNA-seq–based mapping experiment. The
hoxb1b1219 experiments are labeled in red; rhs1b"m111", in green;
vangl2m209, in magenta; and egr2b"h227", in blue; and unknown mutations
mapped using this method, in cyan. Increasing the number of embryos
decreases the linked region with diminishing returns.
changes using the Ensembl tool Variant Effect Predictor (McLaren et al. 2010) and visually confirmed each of these using the Integrative Genomics Viewer (IGV) (Thorvaldsdóttir et al. 2012). Each hoxb1b\textsuperscript{p1219} mapping experiment (80, 40, and 20 embryos) contained exactly one nonsense change in the linked interval, which was the known lesion (Fig. 3A; Table1). There was one missense change in each 80- and 40-embryo hoxb1b\textsuperscript{p1219} linked region. The larger linkage region in the hoxb1b\textsuperscript{p1219} 20-embryo pool included five additional missense mutations (Table1). In the case of nhs1b\textsuperscript{p1131}, which mapped to an interval of ~40 Mb due to the small mutant pool size of eight mutant embryos, there were only two nonsense mutations detected, one of which was the known lesion, and 20 missense mutations (Fig. 3B, Table1); thus even when using a limited number of embryos, the RNA-seq approach provides a manageable number of SNP candidates that might be causative. In the vangl\textsuperscript{p2009} and egr2\textsuperscript{b/h227} intervals, there were zero nonsense mutations and two and four missense mutations detected, respectively. This methodology allows for the identification of a very small number of high priority nonsense and missense candidates underlying a phenotype of interest (Table1).

### Identification of candidate mutations that affect splicing

Many mutations from forward genetic screens alter the splicing of transcripts either by abolishing endogenous splice donor or acceptor sites or by creating new ones, as is the case of vangl\textsuperscript{p2009}. While WGS approaches could detect such mutations as homozygous SNPs within the linked interval, the effect of such changes can be difficult to predict, particularly in the case of the creation of a new splice acceptor or donor in the case of unannotated exons. We analyzed how many splicing variants were identified within the mapped regions by using IGV to visually assess the transcripts with defects in splicing. Within the linked intervals, we identified very few alterations to splicing patterns: In the 80- and 40-embryo hoxb1b\textsuperscript{p1219} and egr2\textsuperscript{b/h227} experiments, there were zero splicing alterations. In the 20-embryo hoxb1b\textsuperscript{p1123} and vangl\textsuperscript{p2009} pools, there was one alteration to the splicing of a transcript, with the vangl\textsuperscript{p2009} change being the known lesion (the splice alteration identified in the hoxb1b\textsuperscript{p1123} 20-embryo pool is outside the region of linkage obtained from the 80- and 40-embryo experiments) (Fig. 3C; Table1). The ability to directly identify and assess the consequences of splice-altering mutations is thus a benefit of the RNA-seq-based approach.

### Identification of candidate mutations affecting gene expression levels

A mutation can alter the level of expression of genes by (1) creating a nonsense change that results in the elimination of mutant transcripts by nonsense-mediated decay (NMD), as is the case of egr2\textsuperscript{b/h227}; (2) affecting transcription by disrupting regulatory elements; or (3) secondarily altering the expression of downstream target genes that could be in the homozygous interval. Although WGS data can detect mutations that disrupt gene expression, it does not include information about the effects of these mutations, making them hard to recognize as causal. We used Cufflinks (Trapnell et al. 2012) and the custom R script RNAeffecto.R (Supplemental Fig. 1) to identify the genes within linked regions whose expression levels are different in the wild-type and mutant pools. We found that only two of the linked regions contained significant expression level changes of greater than twofold: In the large, ~40 Mb, nhs1b\textsuperscript{p1131} region of linkage, there were three genes affected, and in the egr2\textsuperscript{b/h227} mutant pool, there was one gene affected, egr2b itself (Fig. 3D; Table1). Differentiating between the different possible causes of down-regulation is challenging; however, in the case of NMD the nonsense transcript might be captured and sequenced at low frequency. Indeed, in the egr2\textsuperscript{b/h227} case, there was one read in the wild-type pool carrying the nonsense change (data not shown). In the case of regulatory mutations, sequencing of
genomic DNA surrounding the candidate gene would allow for the identification of mutations in conserved enhancer regions that could be responsible for down-regulation. The direct identification of genes with altered expression levels is another powerful benefit of the RNA-seq–based approach.

Overall our bioinformatics pipeline identified a limited number of high priority candidate mutations within the linked region of each experiment (Table 1) and, in each case, identified the known lesion. To facilitate the usefulness of the technique, we have developed an integrated, bioinformatics pipeline called RNAmapper running on the Galaxy platform (Giardine et al. 2005; Blankenberg et al. 2010; Goecks et al. 2010). This package integrates all of the tools used in the pipeline and can be downloaded and used locally or run in the cloud by creating an Amazon Machine Image. RNAmapper and its documentation can be found at www.RNAmapper.org and make the RNA-seq–based mapping approach accessible.

Discussion
Together our results show that RNA-seq–based mapping and candidate gene identification is a powerful approach allowing for the rapid and inexpensive identification of mutations from forward genetic screens. Within this article, for validation purposes, we applied the technology to only known mutations. We have also used this approach to map several unknown mutants to unique regions of the genome (Supplemental Fig. 4) and to identify a limited number of candidate lesions for each (including nonsense, missense, and splice altering lesions) (data not shown). After mapping of unknown mutations and candidate identification, regardless of linkage size and candidate number, further experiments are necessary to link a potential lesion to the phenotype definitively (e.g., demonstrating linkage of the candidate lesion to the phenotype in a large number of individuals, phenocopy via candidate knockdown/removal, and/or rescue via exogenous candidate expression). However, the short list of candidates generated using RNA-seq is likely to yield results quickly. While we have focused on zebrafish, the use of RNA-seq–based mapping is applicable to other systems. A similar approach was recently used to map a mutation in maize (Liu et al. 2012); thus the technique is applicable broadly to many genetic systems with a sequenced genome. There are several other vertebrate model systems that are commonly used to perform forward genetic screens, and their genomes are of a similar size to that of the zebrafish (zebrafish, ∼1.5 Gb; Xenopus tropicalis, ∼1.4 Gb; Mus musculus and Rattus norvegicus, ∼2.7 Gb), suggesting the RNA-seq–based mapping described here can be used to identify candidate mutations in these organisms. For the larger, less polymorphic genome of the mouse, a map cross will be essential to ensure sufficient markers are identified for mapping the mutation. Additionally, because both mouse and rat have larger genomes, but have ∼3000 fewer genes than zebrafish, the mapping resolution will be reduced; however, such a reduction would still provide a relatively small linked region of mapping and a small number of candidate mutations. Our RNA-seq–based approach is simple, using common laboratory procedures and free bioinformatics programs packaged into the RNAmapper program (http://www.RNAmapper.org).

While it is encouraging that we found the known lesion in each of our test cases, a consideration when using RNA-seq for positional cloning is that the mutant transcript of interest could be missed due to low or zero expression. We found that within our mapped intervals, from 16% to 56% of genes had low levels of expression (less than fourfold) that would make it difficult to identify candidate mutations (these numbers were highly variable in the different linkage regions) (Table 1). To mitigate this concern, in the case of an unknown mutation, we suggest extracting RNA at a timepoint as close to the first emergence of the mutant phenotype as possible, as this increases the likelihood that the transcript carrying the causal lesion will be expressed. While such early selection of the phenotype might lead to increased misphenotyping and thereby inclusion of wild-type embryos in the mutant pool, we have found our method to be surprisingly robust against such contamination (Supplemental Fig. 3). Alternatively, RNA could be extracted from a number of different developmental stages or, in the case of mapping an adult phenotype, a number of different tissues. This will increase the breadth of transcripts captured and the likelihood of sequencing the mutant transcript itself. In the worst-case scenario—where the transcript is missed—our RNA-seq approach will still provide a mapping interval due to linked SNPs from neighboring transcripts becoming homozygous. Additionally, the sequencing of transcribed genes within the interval will allow many (44%–84% of genes have greater than fourfold coverage) (Table 1) to be ruled out as candidates, narrowing the search to a limited number of genes. While the possibility of missing the causal mutation using an RNA-seq–based approach remains, care in experimental setup is likely to make this concern minimal, and the mapping will identify a region of linkage with a small list of candidates to validate in subsequent experiments.

Recently developed WGS-based BSA approaches effectively map mutation, and identify 10-fold greater SNPs in each experiment than our RNA-seq–based approach (Bowen et al. 2012; Obholzer et al. 2012). However, we find that our RNA-seq methodology maps mutations to intervals of similar size compared with WGS methods (Supplemental Fig. 5), and both approaches are able to identify nonsense and missense mutations. The RNA-seq–based approach offers three main advantages: First, sequencing the transcriptome allows for the visualization of annotated and unannotated intron/exon boundaries, allowing for the direct identification of mutations affecting splicing. WGS approaches may identify changes that alter known splice acceptor/donors but would fail to directly detect mutations affecting nonannotated isoforms or creating new splice acceptor/donors. Second, sequencing of the mutant and wild-type transcriptomes allows for the identification of candidate genes whose expression is affected by regulatory mutations. While WGS would detect the mutations themselves, other cosegregating, noncoding polymorphisms could mask the identity of the causal lesion. By providing a direct comparison of expression levels, RNA-seq identifies the effects of such mutations, but in the case of noncoding regulatory mutations, it will not detect the mutation itself. In this case, further targeted genomic sequencing would be necessary to identify the causative mutation (Gupta et al. 2010); however, the expression data would focus the search for causal noncoding mutations to those surrounding the candidate whose expression was affected. Here we compared only a single mutant to a single wild-type transcriptome; additional biological replicates would increase the significance of any expression differences between mutant and wild-type pools. Finally, RNA-seq comes at a significantly reduced cost compared with WGS approaches. Currently, WGS approaches require one to two lanes on an Illumina HiSeq 2000 for each mutant (Bowen et al. 2012; Leshchiner et al. 2012; Voz et al. 2012). In contrast, we have multiplexed six samples (three mutant/wild-type pairs) in a single lane. The RNA-seq approach thus incurs one/third to one/sixth the expense of equivalent WGS approaches.
Additionally, we found that computationally decreasing the number of reads by half (to ~20 million 50-bp reads per sample) still allowed for mapping to a small region but came at the cost of reducing the number of reads at the lesion site (Supplemental Fig. 6; Supplemental Table 1). Thus doubling the number of samples multiplexed would further decrease the cost of RNA-seq–based mapping, but these savings would come at the expense of identifying candidate mutations. As sequencing costs fall, it will become feasible to use both WGS and RNA-seq approaches, which would confirm and complement one another powerfully. Currently, RNA-seq offers many advantages at reduced cost.

Methods

The hoxb1b1219, nhsl1bfh131, and egr2bf0227 mutations were generated in the *AB strain and maintained in either a *AB (hoxb1b1219, egr2bf0227) or a *AB/Tu background (nhsl1bfh131). The vanglm209 was generated in the Tu strain and maintained in a *AB background.

We monitored a number of mutants (including nhsl1bfh131 and egr2bf0227) that resulted in a shorter anterior/posterior axis (Jessen et al. 2009). SNPs were identified using the SAMtools mpileup and snpEff (Cingolani et al. 2012) to identify genes with significant expression level changes within the linked region. IGV (Thorvaldsdóttir et al. 2012) was used to assess each potential candidate mutation. All custom scripts and programs developed here can be found at www.RNAmapper.org.
preparation, and sequencing; and Ryan Basom for help in data processing. Funding was provided by the National Institutes of Health, R01HD037909 and R01HG002995 to C.B.M., the NRSF fellowship F32NS074839 to A.C.M., and NIDCD R21DC012097 to S.G.M.

Author contributions: A.C.M. and C.B.M. performed the in vivo experiments and sample preparation. A.C.M., A.N.S., and N.D.O. wrote the custom R scripts. N.D.O. created the Galaxy implementation of the bioinformatics pipeline. A.C.M. and C.B.M. wrote the manuscript. All authors edited the manuscript.

References
Blankenberg D, Von Kuster G, Coraor N, Ananda G, Lazarus R, Mangan M, Luley V, Taylor J. 2010. Galaxy: A web-based genome analysis tool for experimentalists. *Curr Protoc Mol Biol* 89: 19.10.1–19.10.21.

Blumenstiel JP, Noll AC, Griffiths JA, Perera AG, Walton KN, Gilliland WD, Hawley RS, Staehling-Hampton K. 2009. Identification of EMS-induced mutations in *Drosophila melanogaster* by whole-genome sequencing. *Genetics* 182: 25–32.

Bowen ME, Henke K, Siegfried KR, Warman ML, Harris MP. 2012. Efficient mapping and cloning of mutations in zebrafish by low coverage whole genome sequencing. *Genetics* 190: 1017–1024.

Brenner S. 1974. The genetics of *Caenorhabditis elegans*. *Genetics* 77: 71–94.

Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. 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* 6: 80–92.

Crick FH, Barnett L, Brenner S, Watts-Tobin RJ. 1961. General nature of the genetic code for proteins. *Nature* 192: 1227–1232.

Gehrke R, Rauch G, Geiger-Rudolph S, Albrecht A, van Bebber F, Berger A, Busch-Nentwich E, Dahm R, Dekens MPS, Dooley C, et al. 2007. Large-scale mapping of mutations affecting zebrafish development. *BMC Genomics* 10: 865–867.

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

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

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

Haffter P, Granato M, Brand M, Mullins MC, Hammerschmidt M, Kane DA, Odenthal J, van Esden MJ, Jiang Y, Heisenberg CP, et al. 1996. The identification of genes with unique and essential functions in the development of the zebrafish, *Danio rerio*. *Development* 123: 1–36.

Jessen JR, Topczewski J, Bingham S, Sepich DS, Marlow J, Chandrasekhar A, Solntca-Krezel L. 2002. Zebrafish trilobite identities new roles for Strabismus in gastrulation and neuronal movements. *Nat Cell Biol* 4: 610–615.

Leshchiner I, Kafri O, Aziziov A, Reshef D, Zuckerman A, Cooney CA, Anderson H, King M, Stottmann R, Ha S, et al. 2012. Mutation mapping and identification by whole genome sequencing. *Genome Res* 22: 1541–1548.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 1000 Genome Project Data Processing Subgroup. 2009. The sequence alignment/map (SAM) format and SAMtools. *Bioinformatics* 25: 2078–2079.

Liu S, Yeh CT, Tang HM, Nettleton D, Schnable PS. 2012. Gene mapping via bulked segregant RNA-Seq (BSR-Seq). *PLoS ONE* 7: e36406.

Mclaren W, Pritchard B, Rios D, Chen Y, Fitzek P, Cunningham F. 2010. Deriving the consequences of genomic variants with the Ensemble API and SNP Effect Predictor. *BMC Bioinformatics* 26: 2069–2079.

Meyerowitz EM, Pruitt RE. 1985. *Arabidopsis thaliana* and plant molecular genetics. *Science* 229: 1214–1218.

Monk KR, Naylor SG, Glenn TD, Mercurio S, Perlin JR, Dominguez C, Moens CB, Talbot WS. 2009. A G protein-coupled receptor is essential for Schwann cells to initiate myelination. *Science* 325: 1402–1405.

Nolan PM, Peters J, Strivens M, Rogers D, Hagan J, Spurr N, Gray IC, Vizor L, Brooker D, Whitehill E, et al. 2000. A systematic, genome-wide, phenotype-driven mutagenesis programme for gene function studies in the mouse. *Nat Genet* 25: 440–443.

Nüsslein-Volhard C, Wieschaus E. 1980. Mutations affecting segment number and polarity in *Drosophila*. *Nature* 287: 795–801.

Obholzer N, Swinburne JA, Schwab E, Nechiporuk AV, Nicolson T, Megason SG. 2012. Rapid positional cloning of zebrafish mutations by linkage and homozygosity mapping using whole-genome sequencing. *Development* 139: 4280–4290.

Sarín S, Prabhu S, O’Meara MM, Pérez I, Hobert O. 2008. *Caenorhabditis elegans* mutant allele identification by whole-genome sequencing. *Nat Methods* 5: 865–867.

Thorvaldsdóttir H, Robinson JT, Mardis ER. 2012. Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration. *Brief Bioinform* doi: 10.1093/bib/bbs017.

Trapnell C, Pachter L, Salzberg SL. 2009. TopHat: Discovering splice junctions with RNA-Seq. *Bioinformatics* 25: 1105–1111.

Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L. 2012. Differential gene and transcript expression analysis of RNA-Seq experiments with TopHat and Cufflinks. *Nat Protoc* 7: 562–578.

Voz ML, Coppeler W, Manfredi I, Baudhuin A, Von Berg V, Charlier C, Meyer D, Driever W, Martial JA, Peers E, et al. 2002. Fast homozygosity mapping and identification of a zebrafish ENU-induced mutation by whole-genome sequencing. *PLoS ONE* 7: e34671.

Walsh GS, Grant PK, Morgan JA, Moens CB. 2011. Planar polarity pathway and Nance-Horan syndrome-like 1b have essential cell-autonomous functions in neuronal migration. *Development* 138: 3033–3042.

Zhou Y, Zon LI. 2011. The zon laboratory guide to positional cloning in zebrafish. *Methods Cell Biol* 104: 287–309.

Received August 3, 2012; accepted in revised form December 18, 2012.