Stock maintenance, outcrossing, and backgrounds

The RNA-Seq based mapping method has thus far successfully mapped mutations in known genes (described in Miller et al. (accepted, Genome Research)) as well as several unknown (> 10) mutations. We have tried different backgrounds ("strict" *AB, map crosses (AB/WIK)) and generations (F3-F6), all with great success (...except one described below).

In the experiments described in the paper we used a strict *AB background. By strict *AB we mean that the *AB line was maintained by incrossing of *AB, therefore limiting polymorphism. Never-the-less we found plentiful SNPs with which to map the mutants (all correctly mapping to the known locus). In subsequent experiments where outcrosses were performed to polymorphic strainswe found an increase in the number of SNPs by ~2-fold compared to strict AB and this increased the confidence of the mapping analysis. This comes at the cost of having more SNPs to contend with when identifying mutants; however, by filtering against known WT SNPs this is not a great concern as most are naturally occurring polymorphisms. We have been able to map mutations from the F3-F6 generation, and see no significant changes in map size or candidate identification. The one mapping "failure" was in a *very* strict *AB background that had ~1/4 the number SNPs when compared to any of the other *AB only backgrounds. The experiment still linked the mutation to a chromosome but the confidence in the mapping was not high and we could not define a smaller region of mapping. Subsequent analysis has shown that the chromosome did carry the mutation, but clearly increasing the number of SNPs is important for resolving a small region of linkage.

Best practice: Perform an outcross. This will give better mapping data, and the excess SNPs will generally be filtered out by comparing to known WT SNPs.

Crosses

We have thus far only used single pair crosses for each experiment. This simplifies the SNP background, reducing the chance of complication due to multiple alleles at any given locus. However, recent work from our collaborators (Obholzer et al. (Development)) has shown that mixing embryos from several mating pairs works well. This could increase the speed of acquisition of mutant animals for mapping. https://wiki.med.harvard.edu/SysBio/Megason/MegaMapper

We have successfully mapped with as few as 8 mutant/8 wt embryos and as many as 80 mutant/8wt embryos. Increasing the number of mutants decreases the size of the linked region. However, there are diminishing returns given that the smaller the mapped region the less likely a recombinant will be found that reduces the linked region. We find that ~50 mutants is a good number to use giving a map size of about 4Mb, small enough that there will be few candidates.

Clean sorting of the phenotype is VERY important. While we have found that the method can tolerate wildtype contamination in the mutant pool, this decreases the allele frequency within the linage and therefore creates ambiguity. Be extremely strict in the selection of animals for the mutant pool. The wildtype pool is less

sensitive as there are always mutant alleles in the wildtype pool from the het (+/-) animals. Again, be careful with the mutant pool!

Best practice: Single pair cross and collect 50 mutant (and wildtype) embryos. But let us know if you try something else!

RNA extraction

Extract total RNA from mutant and wildtype pools separately using Trizol (Invitrogen). RNAlater could work but I find it "gunky" and annoying. Make sure to be careful because RNA is easily degraded by all the worldly RNAses. Use RNAse ZAP wipes and gloves and care. Good quality RNA is key to a good mapping.

RNA quality should be tested for quality using a spectrophotometer and the Agilent 2100 Bioanalzyer; RNA was only accepted if it was uncontaminated with phenol or guanidinium thiocyanate (both from Trizol), and the RNA Integrity Number (RIN) was greater than 9.0. Alternatively, an RNA gel can be run and a comparison of 28s to 18s can be used – this number should be ~2.0.

RNA prep, cDNA, and Illumina libraries

We have used our local genomics facility to go from RNA to Illumina libraries. However, the steps are not overly complex and are described in a kit from Illumina. (http://www.illumina.com/truseq.ilmn)
The basics are to use ~1.0 ug of total RNA to start (we have used anywhere from 0.2 - 1 ug). Perform a polyA selection for mRNA, chemically fragmented to ~200bp, and use random hexamer primers to create cDNA. Library preparation follows the TruSeq Illumina protocol with each individual library receiving a unique Illumina barcode, allowing for their identification after multiplexed sequencing. RNA-Seq was performed on an Illumina HiSeq 2000 machine with 6 pools multiplexed per lane using 50bp paired-end reads. This resulted in an average of 250 million reads per lane, with an average of 43 million reads per sample. Longer reads will allow for more multiplexing per lane.

However, the ultimate goal is to sequence the RNA from the sample, align those reads with the genome. The aligned sequencing file will then allow you to identify mapping SNPs in the wildtype siblings, and use this information to look at the SNP allele frequencies in mutant, thereby identifying regions of homozygosity (linkage) to your mutation. Therefore, any method that generates quality sequencing reads that can be aligned to the genome in the .bam format will work.

Methods from paper (Miller *et al.* Genome Research)

The *hoxb1b*^{1219}, *nhsl1b*^{fh131}, and *egr2b*^{fh227} mutations were generated in the *AB strain and maintained in either a *AB (*hoxb1b*^{1219}, *egr2b*^{fh227}) or a *AB/Tu background (*nhsl1b*^{fh131}). The *vangl*^{m209} was generated in the Tu strain and maintained in a *AB background. *hoxb1b*^{1219} and *egr2b*^{fh227} embryos were collected in the F3 generation, while *nhsl1b*^{fh131} and *vangl*^{m209} were outcrossed for greater than five generations. A single-pair of heterozygous carriers were crossed for each mutation and embryos were collected and sorted, based on morphological phenotypes, into mutant and wildtype pools. Total RNA was extracted from each pool separately using a standard acid guanidinium thiocyanate and phenol chloroform extraction (Trizol, Invitrogen). RNA was tested for quality using a spectrophotometer and an Agilent 2100 Bioanalzyer; RNA was only accepted if it was uncontaminated with phenol or guanidinium thiocyanate and the RNA Integrity Number (RIN) was greater than 9.0. Approximately 1.0 ug of total RNA was then polyA selected, chemically fragmented to ~200bp, and cDNA was created using random hexamer primers. Library preparation followed the TruSeq Illumina protocol with each individual library receiving a unique Illumina barcode, allowing for their identification after multiplexed sequencing. RNA-Seq was performed on an Illumina HiSeq 2000 machine with 6 libraries multiplexed per lane using 50bp paired-end reads. This resulted in an average of 250 million reads per lane, with an average of 43 million reads per sample.

Raw reads were aligned to the zebrafish genome (Zv9.63) using TopHat/bowtie, an intron and splice aware aligner (Trapnell *et al.* 2009). SNPs were identified using the SAMtools mpileup and bcftools (Li *et al.* 2009) variant caller requiring the map and nucleotide quality to be > 30 and, importantly, allowing for anomalous pairs to be mapped – these “anomalous” pairs being reads spanning large exons. For the purposes of mapping, SNPs were further filtered for quality based on expression level (at least 25x) and for high alternative allele frequency (at least 25%) using the custom R script RNAmappe.R. RNAmappe.R then assessed the mutant allele frequency at the positions of the high-quality wildtype SNP markers and averaged these frequencies using a sliding window of 50 neighboring markers. This average was then plotted across the genome and linkage was identified by analyzing the genome-wide mapping data for the region of highest average frequency. Subsequently the custom R script, RNAidentifie.R, was used to extract all SNPs within the region of linkage and filter the mutant SNPs against independently identified wildtype SNPs. A custom R script, VCFmerge.R, was written to combine VCF formatted SNPs from multiple sources, including RNA-Seq data from our in-house wildtype strains, recent WGS projects (Bowen *et al.* 2012; Obholzer *et al.* submitted), and standard community sites (dbSNP, Ensembl). SNPs remaining after filtering were then assessed for consequences to proteins using Ensembl’s Variant Effect Predictor (McLaren *et al.* 2010) or snpEff (Cingolani et al., 2012). The custom R script VEPsorte.R was

used to sort and prioritize SNP candidates from VEP. The Cufflinks package was used to assess differences in expression between the wildtype and mutant pools (Trapnell *et al.* 2012). The custom R script RNAeffecto.R was used to extract and identify genes with significant expression level changes within the linked region. IGV (Thorvaldsdóttir *et al.* 2011) was used to assess splice changes at intron/exon boundaries and also to visually assess each potential candidate mutation. All custom scripts were written in R and are available for download with an open-source BSD license (free for academic and commercial use).

To generate a user-friendly mapping platform, we developed RNAmapper*,* a downloadable or web-based bioinformatics pipeline based on Galaxy (http://galaxy.psu.edu). RNAmapper can be run on powerful, desktop workstation or on the Amazon cloud. We packaged RNAmapper and all associated required programs and reference data into a single bundle using VirtualBox (https://www.virtualbox.org/). We also created an Amazon Machine Image (AMI), to allow users to instantiate their own RNAmapper server on the Amazon Elastic Compute Cloud. The source code and virtual machines will be free to download.