Identification of EMS-induced mutations in *Drosophila melanogaster* by whole genome sequencing

Justin P. Blumenstiel*,†, Aaron C. Noll†, Jennifer A. Griffiths†, Anoja G. Perera†, Kendra N. Walton†, William D. Gilliland†, R. Scott Hawley†‡ and Karen Staehling-Hampton†,1

* Department of Ecology & Evolutionary Biology, University of Kansas, Lawrence, Kansas, 66045, † Stowers Institute for Medical Research, Kansas City, Missouri, 64110 and ‡ Department of Physiology, Kansas University Medical Center, Kansas City, Kansas, 66160.

1 Corresponding Author

Sequence data from this article have been deposited with the Short Read Archive (SRA) at NCBI under accession number SRA008255.
Running Title: Mutation Mapping by Sequencing

Key words: EMS, gene mapping, whole genome sequencing, *Drosophila melanogaster*, encore

1Corresponding Author:

Karen Staehling-Hampton

1000 East 50th Street

Kansas City, MO 64110

Phone: 816-926-4416

Fax: 816-926-2053

KSH@Stowers.org
ABSTRACT

Next generation methods for rapid whole genome sequencing enable the identification of single base pair mutations in Drosophila by comparing a chromosome bearing a new mutation to the un-mutagenized sequence. To validate this approach, we sought to identify the molecular lesion responsible for a recessive EMS-induced mutation affecting egg shell morphology by using Illumina next generation sequencing. After obtaining sufficient sequence from larvae that were homozygous for either the wildtype or mutant chromosomes, we obtained high quality reads for base pairs comprising ~70% of the 3rd chromosome of both DNA samples. We verified 103 single base changes between the two chromosomes. Nine changes were non-synonymous mutations and two were nonsense mutations. One nonsense mutation was in a gene, encore, whose mutations produce an egg shell phenotype similarly observed in progeny of homozygous mutant mothers. Complementation analysis revealed that the chromosome carried a new functional allele of encore, demonstrating that one round of next generation sequencing can identify the causative lesion for a phenotype of interest. This new method of whole genome sequencing represents great promise for mutant mapping in flies, potentially replacing conventional methods.
Standard practices of genetic mapping typically occur in three phases. First, polymorphisms must be identified that distinguish the chromosome carrying the mutation to be mapped from that of the homolog bearing a wildtype allele of that gene. Second, by genotyping recombinant chromosomes that do or do not carry the mutation of interest, an association between polymorphisms and the mutation can be identified which can thus be used to pinpoint the location of the relevant mutation. Finally, candidate genes within the interval must be identified and regions sequenced to find the causative mutation. Often, these three steps are performed iteratively. In situations where there are few polymorphic markers or candidate genes this process can be arduous and, depending on the organism, can consume months to years.

New genome sequencing technologies (Barski et al. 2007; Bentley 2006; Margulies et al. 2005; Sarin et al. 2008; Smith et al. 2008; Valouev et al. 2008) show tremendous promise in reducing the time needed to identify causative mutations. Using these approaches one may be able to directly identify causative lesions by comparing the nucleotide sequences of wildtype and mutant genomes. Indeed, we have conducted a proof of principal experiment to determine the feasibility of such an approach in Drosophila melanogaster. In the course of an EMS-based genetic screen, we identified a chromosome, designated 791, which displayed a fused dorsal appendage phenotype in embryos of homozygous mothers. Such phenotypes usually arise from a defect in the maternal establishment of the dorso-ventral axis. To identify the mutated gene that gives rise to this phenotype, we used a next generation sequencing platform to directly compare the nucleotide sequence of the original and mutagenized chromosomes. Because this phenotype is well studied and our mutation is recessive, we could use complementation
analysis to test the causative nature of any candidate lesions. However, even if other mutants with similar phenotypes weren’t already known, the small number of candidate loci identified could have been easily tested by transformation rescue. Importantly, this approach also improved our understanding of the global effects of EMS mutagenesis. Here we demonstrate how whole genome sequencing technologies can be used to discover causative mutations and how these technologies can shed light on processes such as EMS mutagenesis and gene conversion at a genomic level.

**MATERIALS AND METHODS**

**DNA preparation for sequencing:**

DNA for sequencing was prepared from wandering 3rd instar larvae that were homozygous for either *A15* (the target chromosome) or *79I* (the mutagenized chromosome). Homozygosity was determined by selection against *TM6b,Tb* balancer chromosomes. Wandering 3rd larvae were chosen for three reasons: first, at this stage they have begun gut evacuation thus minimizing contaminating DNA from the yeast food source; second, they can be easily bleached to remove surface contamination; and finally, larval salivary glands contain polytene chromosomes that are enriched for euchromatic over heterochromatic sequences. Since heterochromatic sequences are not easily assembled, especially for the short read lengths generated by Illumina sequencing, we favored minimizing their contribution to the sequencing runs.

DNA was prepared from 10 larvae that had been briefly rinsed in 50% bleach followed by water and frozen at -80° for at least one hour. Larvae were then homogenized
in 500 ul of 10 mM Tris-HCl (pH 8.0), 20 mM EDTA, 0.1% SDS and 5 ug of RNase A and incubated at room temperature for 10 minutes. 5 ul of Proteinase K (20 mg/ml) and 40 ul of 10% SDS were then added and the homogenate was incubated at 65° for one hour, followed by 95° for 5 minutes. 125 ul of 5 M ammonium acetate were added, tubes were incubated on ice for 10 minutes, spun for 10 minutes and supernatant was collected and extracted once with Phenol:Chloroform:Isoamyl Alcohol (25:24:1) and once with Chloroform. DNA was precipitated by the addition of 2X volumes of cold ethanol, and the pellet was rinsed once with 70% ethanol. The pellet was resuspended in 50 ul of 10 mM Tris-HCl, pH 8.5.

**Illumina whole genome sequencing:**

5 ug of genomic DNA from either A15 or 791 homozygous larvae was sheared to ~800bps using sonication. We then performed end repair, added ‘A’ bases to the 3’ end of the DNA fragments, ligated adapters, and purified and size selected ligated products. Clusters were generated on the Illumina cluster station according to the manufacturer’s protocol. Single read sequencing was done for 36 cycles (36bp) on an Illumina Genome Analyzer I instrument. One flow cell was run for each library. Seven lanes were run for the A15 background strain, and seven lanes were run for the 791 mutant. The eighth lane of each flow cell was used for a Phi-X control.
**Illumina Data Analysis and SNP Detection:**

Data analysis was done using a combination of commercially available software, open source software and custom programs. Images from the Illumina Genome Analyzer were processed using the Illumina Analysis Pipeline version 0.3.0 (Firecrest, Bustard) to generate FASTQ sequence files. 36bp reads that passed the Gerald chastity filter were aligned uniquely to the reference genome sequence using the eland alignment tool. All quality filtered and uniquely aligning reads were provided to the MAQ package (Li et al. 2008, http://maq.sourceforge.net) using default settings. MAQ was used to align reads to the ensembl 49.44 release of the *D. melanogaster* genome (http://mar2008.archive.ensembl.org/Drosophila_melanogaster). A15 and 791 consensus sequences from MAQ for the 3rd chromosome were then compared in a pair-wise fashion. Criteria used when comparing references were a minimum read depth of four, a homozygous consensus call and a minimum consensus quality score of 22. Non-matching, threshold passing pairs were then annotated. When a pair’s chromosomal position was determined to land in a transcript and the resulting translated protein change was non-synonymous, the SIFT program (Ng and Henikoff 2002) was used to predict the impact as deleterious or tolerated. All subsequent secondary analysis was performed using custom scripts and the R programming language.

**Sanger Sequencing Validation:**

Primers of length 18-27 bp and Tm 57-63° were designed to amplify approximately 700 bp products including at least 350 bp on either side of the putative SNP. M13 universal primer tags were appended to the 5’ end of each primer to aid in sequencing
reaction setup (forward M13 primer: TGT AAA ACG ACG GCC AGT; reverse M13 primer: AAC AGC TAT GAC CAT G). 15 µl PCR reactions for each pair of primers were set up in 384-well plates, using TAQ Gold (Applied Biosystems). A standard PCR protocol was used for all regions: 10 min at 95°; 30 s at 95°, 30 s at 60°, 1 min at 72° (for a total of 30 cycles) and then 10 min at 72° followed by an 8° hold. Unincorporated primers, nucleotides and salts were removed on a Biomek FX using AMPure cleanup (Agencourt). In a 384-well plate, 2 µl of each eluted PCR product was added to 8 µl of a Big Dye Terminator v3.1 sequencing cocktail (Applied Biosystems), including either the forward or reverse M13 sequencing primer. The same sequencing PCR cycle was used for all regions: 10 s at 96°, 5 s at 50° and 4 min at 60°, followed by an 8° hold. Reactions were purified on the Biomek FX using CleanSEQ cleanup (Agencourt) and sequenced on an Applied Biosystems 3730xl sequencer. Vector NTI software (Invitrogen) was used to assemble, view the data and detect SNPs.

**RESULTS**

Illumina fragment libraries were made from genomic DNA isolated from homozygous larvae, carrying either the original 3\textsuperscript{rd} chromosome (designated A15) or the EMS-mutagenized 3\textsuperscript{rd} chromosome (designated 791), which carries a lesion that causes a fused dorsal appendage phenotype. Each library was run on a single flow cell on an Illumina Genome Analyzer using the single read protocol. Approximately 30 million filtered and uniquely aligning reads of 36bp length were generated for each sample. This produced 1.1 Gb of sequence for the original stock and 1.0 Gb of sequence for 791, giving
8.7 and 8.3 X genome coverage, respectively (Table 1). From this set of data, we limited our analysis to the 3rd chromosome since this chromosome was the target of mutagenesis. Sequencing coverage was not Poisson distributed. Instead, the variance in the distribution of coverage was greater than predicted by a Poisson distribution and there was an excess of zero coverage bases (Figure 1A) for both sequence runs. If these deviations were due to an underlying random process (albeit a non-Poisson process) that was independent across samples, we would expect to see little correlation in sequence depth between the two samples. However, this was clearly not the case. Figure 1A shows a frequency heat map for pair wise coverage across the two samples. There is a clear correlation between the sequencing depth at any particular base in one run with the sequencing depth in the other. Furthermore, the zero coverage class in one sample is quite coincident with the zero coverage class in the other sample. This correlation indicates that sequence depth is non-independent across samples, suggesting a certain bias for some parts of the genome to not be sampled in each sequencing run. This bias could be due to bias in the sequencing process or due to the contribution of polytene chromosomes to the pool of DNA collected from third instar larvae. Polytene chromosomes likely vary in the extent to which they contribute sequences from different genomic regions. If coverage were truly independent across samples, the removal of low coverage data from the analysis would also be independent across samples. This would multiply the false negative rate since a site with reasonable coverage in one sample would frequently have low coverage in another and thus be eliminated from analysis.

Considering only sites that had non-ambiguous consensus calls and a minimum sequence depth of four (since consensus quality scores are expected to be less meaningful
for low coverage bases), we also characterized the distribution of MAQ quality scores for each nucleotide of the consensus. For both samples, the distribution of quality scores is variable at or below a score of approximately 20 (Figure 1B). However, above this threshold, the distribution of quality scores appears more continuous (aside from the fact that the MAQ consensus quality score algorithm appears to give a somewhat punctate distribution of values). As with coverage depth, consensus quality scores are somewhat correlated across bases (Figure 1B). A very high quality consensus base in one sample is more likely to be higher quality in the other sample. This again indicates non-independence in quality across runs – some bases are more likely to be read as high quality by Illumina sequencing than others. It is also expected when coverage across bases is correlated between runs and when bases with higher coverage have higher quality scores. For the same reason as with coverage, this non-independence makes a threshold quality cutoff for SNP determination less likely to have drastic influence on the false negative rate.

In order to identify EMS-induced mutations we chose an approach of directly comparing the consensus sequences of the two chromosomes generated using the MAQ software program. An alternative approach would be to identify all SNPs relative to the completely sequenced reference for each genome, and identify the EMS-induced mutations based on the comparison of these two lists of SNPs. This method is problematic, however, since there is a great deal of natural variation that is expected to distinguish the unmutagenized chromosome from the reference genome. Even a very small false negative or false positive rate of SNP identification for each genome relative to
the reference would lead to a large excess of putative SNPs unique to one genome that would, in fact, not be SNPs between A15 and 791.

Using a threshold which only considered non-ambiguous consensus bases from both chromosomes that had a minimum read depth of four and quality score of 22, we covered 70.9% of both 3rd chromosomes. Table 1 shows that this fraction is greater than expected if the chance of a given nucleotide passing this filter is independent across samples. Furthermore, since low coverage bases are more likely to be low quality bases (data not shown), a portion of the genome that we removed from analysis is enriched for bases that are predisposed to being low quality. This is supported by the fact that bases excluded from analysis are enriched for repeats. While 7.6% of the 3rd chromosome is masked by RepeatMasker, 20.6 % of the bases not meeting threshold is masked by RepeatMasker. Thus, a portion of the 3rd chromosome that is not included in the analysis is, due to repetitiveness, unlikely to contribute to any whole genome sequencing SNP detection approach in Drosophila, even in the face of greater sequencing depth. Furthermore, since the primary goal is to identify mutations in genes with unique function, unidentified SNPs located within repeat sequences, such as transposons, are not likely to be SNPs of interest. The portion of the 3rd chromosome not included in the analysis due to low coverage could have been decreased by running additional lanes of sample. When we add the results of a test run from another full flow cell of 791 reads (with somewhat lower quality and not included in this analysis) we increased the threshold shared coverage between A15 and 791 from 70.9 to 74.8 percent. Thus, the percentage coverage does not increase drastically with data from additional flow cells. This is not surprising since only 80% of a complex eukaryotic genome can be uniquely mapped with short 36bp reads.
Using this threshold, we identified 165 candidate SNPs that distinguished the mutagenized 3rd chromosome from the unmutagenized chromosome. We successfully performed Sanger sequencing on 125 of these SNPs and verified 103, giving a false positive rate of 17.6%. For a complete list of all 103 verified SNPs, see supplemental material. Visual inspection indicated that a number of false positives were in low complexity and repeated regions while others were likely due to sequencing errors or potential PCR amplification errors during library preparation. If we apply this respective false positive rate to the entirety of the 165 candidate SNPs, we would yield ~136 true SNPs for the 70.9% of the genome covered after filtering. This yields ~ one mutation per 273 kb. Considering the fact that a 45 mM dose of EMS was used, this is consistent with previous reports of a ~1/380 kb and ~1/480 kb found with a 25 mM dose of EMS (COOPER et al. 2008).

We found that the verified SNPs could be placed in two different categories (Fig 2A). The first category was designated “standard” for SNPs that distinguished the unmutagenized and mutagenized chromosome, and for which the nucleotide on the mutagenized chromosome differed from the reference sequence. Seventy-five nucleotides fell into this class. The second category was designated “anomalous” for SNPs that differed between the unmutagenized and mutagenized chromosomes, but for which the mutagenized chromosome had the same sequence as the reference genome. Twenty-eight nucleotides fell into this class. Interestingly, the false positive rate was much higher for this class of SNPs (37.8%) than for the ‘standard’ class (6.25%). The probability of a nucleotide differing between the unmutagenized chromosome and the reference sequence reverting to the reference sequence is exceedingly small, therefore the verified
‘anomalous’ SNPs warranted further investigation (see section on anomalous SNPs below).

Of the 75 verified “standard” class lesions, 80% were G/C to A/T transitions, which are known to arise from EMS-mediated alkylation of guanine. This is consistent with the proportion observed in other comprehensive analyses of EMS-induced mutations in Drosophila: 70-76% (COOPER et al. 2008); 100% (BENTLEY et al. 2000); and 84% (WINKLER et al. 2005). It also confirms the observation that the mutation profile under EMS dramatically differs from Arabidopsis, which shows > 99% G/C to A/T transitions (COOPER et al. 2008; GREENE et al. 2003). Finally, annotation of these 75 verified “standard” SNPs indicated that 58 were in non-coding regions, nine were non-synonymous, two were nonsense and the remaining were silent (Table 2). The two nonsense mutations were in the genes encore and His2AV. Non-synonymous mutations were found in the following genes: CG5146, CG3996, prospero. Spt3, CG7839, CG32091, CG32425, Cad99C, RhoGAP100F.

Importantly, one of the EMS-induced mutations was a nonsense mutation in the gene encore (Fig 3A). encore plays a role in the regulation of Cyclin E during oogenesis and encodes for a protein which is 1823 amino acids in length (PA isoform) (HAWKINS et al. 1996; HAWKINS et al. 1997; OHLMEYER and SCHUPBACH 2003; VAN BUSKIRK et al. 2000). The lesion we identified, designated encore791, results in the replacement of glutamine 1353 with a stop codon (Figure 3A). Mutations in the gene encore are known to have an effect on dorsal appendage formation similar to that observed in the embryos of 79I homozygous mothers. A complementation test performed with mothers raised at the sensitive temperature of 18° revealed that the 79I chromosome failed to complement the
encore$^{R1}$ allele for the fused dorsal appendage defect. This reveals that the 791 mutation is a new hypomorphic allele of encore (Figure 3B). Thus, using a whole genome sequencing approach, we have identified the causative mutation underlying the fused dorsal appendage phenotype associated with the 791 chromosome.

‘Anomalous’ class of SNPs result from gene conversions with the balancer chromosomes:

Strikingly, the mutation profile differed dramatically between the verified “standard” class SNPs and those which were verified and classified ‘anomalous’ (Fig. 2A). Only 42.9% of the latter class was G/C to A/T transitions. Moreover, we noticed that the “anomalous” SNPs were highly clustered. Defining clusters as SNPs that are less than 500 bp apart from one another, 21 of 28 verified “anomalous” SNPs resided in a total of eight clusters (Fig. 2B,C). Annotation of the verified SNPs also indicated a strong difference in the spectrum of impact between the two classes (Table 2). Unlike the verified "standard" class SNPs, none of the verified "anomalous" SNPs changed protein function – all were either in non-coding regions or silent. This difference in impact is significant between the two classes (Fisher’s Exact Test, p < 0.05). In aggregate, these data indicate the most likely source of the “anomalous” lesions is gene conversion off a segregating balancer. This is the most parsimonious explanation as gene conversion is expected to produce what appear to be continuous tracts of mutations that are not canonical G/C to A/T EMS-induced transitions, but rather apparent “reversions” to an alternate sequence.

To determine whether or not these clusters of mutations were in fact due to gene conversion events, using the segregating balancer chromosome ($TM6b,Tb$ and $TM3,Sb$) as
a donor, we sequenced the cluster regions on the balancer chromosomes in heterozygous adults. In addition, we performed Sanger sequencing of flies homozygous for a third chromosome designated 3RT from which 791 and A15 had been generated. With these additional data, we found that all of the clustered “mutations” were in fact identical to a corresponding set of polymorphisms that distinguished a segregating balancer chromosome from the original 3RT chromosome (Fig. 2B). This included five SNPs which were originally classified as ‘standard’ but also resided within the clusters. Moreover, we found that where either A15 or 791 possessed a unique sequence, this sequence corresponded to the balancer it had been maintained over, namely TM6b,Tb in the case of A15 and TM3,Sb in the case of 791. Thus, we conclude that these lesions arose from gene conversion events that transferred sequence information from balancer to balanced chromosomes. The minimal length of these gene conversion tracts ranged from 12 to 724 bp, with a mean of 245 bp. Since 21 of the 21 clustered ‘anomalous’ mutations arose from apparent gene conversion events with balancer chromosomes, we conclude this to be the most parsimonious explanation for the ‘reversion’ of anomalous lesions to the reference sequence. Considering the entire set of 103 differences between the A15 and 791 chromosomes, 33 (28 anomalous mutations + five standard mutations residing in the clusters) can thus be attributed to gene conversion occurring within either of the balanced stocks.
DISCUSSION

New technologies for whole genome sequencing have tremendous potential in aiding the search for mutations of interest. By identifying, in one round of sequencing, *encore* as the gene whose defect caused the fused dorsal appendage phenotype associated with the *791* chromosome, we have demonstrated a proof of concept that next generation sequencing can be a powerful method for identifying lesions that produce phenotypes of interest. This study was done on an Illumina Genome Analyzer I (GAI) with the single read 36bp protocol using the original chemistry and version 0.3.0 of the Illumina Analysis Pipeline. The current version of the Illumina Genome Analyzer platform (GAII with paired end module, analysis pipeline v1.3 and chemistry v3) is capable of much longer reads of over 100bp and can generate almost 20 gigabases per run compared with ~1 gigabase per run reported in this study. In addition, paired-end reads make reading through repetitive regions possible. Based on the current performance statistics of the platform, we predict that >90% of the Drosophila genome can be sequenced to greater than 20X coverage with just several lanes of a flowcell. This economy of scale makes large throughput whole genome sequencing in flies economically feasible for most Drosophila researchers.

It is important to note that this approach unifies several different aspects of genetics research. Historically, fine scale mapping was done in an iterative process that required narrowing down a region of interest and identifying new markers that could identify recombination events within successively smaller regions. However, using the approach outlined here, one may be able to identify candidate lesions that can immediately
be tested for their role in a given phenotype. Even without alleles that enable the complementation test, overlapping deficiencies and transformation rescue experiments can be used to identify causative lesions. We expect that additional confirmation by these methods will be fairly straightforward since, with a 45 mM dose of EMS, we recovered only 11 lesions that affected coding sequence, 10 of which were obviously EMS-induced candidates with one lesion in the Cad99C gene likely resulting from a gene conversion event off a balancer. Moreover, even in the face of no obvious causative lesion, future researchers will be able to use the EMS-induced SNPs themselves as mapping markers. This will eliminate the need to recombine mutations of interest onto chromosomes with previously defined SNP markers.

A second aspect of genetics research that is unified with this approach is the generation of new alleles. In the past, an EMS screen would be used to identify genes with a particular phenotype of interest. In this process, however, countless other lesions that might have been of interest to others would be ignored due to lack of an effect on the relevant phenotype. In one iteration of this process, we have identified a total of 82 non-coding mutations, nine new non-synonymous alleles (one of which was attributed to gene conversion) and two new nonsense alleles. Thus, using a next generation sequencing approach, future geneticists will effectively be able to merge marker discovery, mapping and targeted mutagenesis.

But beyond using next generation sequencing as a genetics tool, it also allows deeper insight into fundamental biological processes. The spectrum of EMS-induced lesions is known to differ between flies and other organisms, but the mechanism underlying this difference is not clear. It has been suggested that the mechanism of DNA
repair may differ enough between species to explain this difference. We have found evidence that a significant fraction of non-canonical EMS mutations in flies are found in clusters that likely arise through gene conversion. Thus, part of the difference in the mutational spectrum during treatment with EMS may lie in the false attribution of gene conversion events as being induced by EMS. This false inference will be more common with an increasing likelihood of gene conversion off a homolog with distinguishing variants. Drosophila and broader dipterans are especially known for their efficiency in homolog pairing. Even though balancers inhibit crossing over through their multiple inversions, they pair surprisingly well (Gong et al. 2005). Furthermore, there is strong evidence that gene conversion events can occur from balancers to balanced chromosomes (Cooper et al. 2008). Thus, one possible explanation for the difference in mutational profiles after EMS treatment in Arabidopsis and Drosophila is that while mutagenesis in Arabidopsis typically makes use of inbred lines for which gene conversion will not carry distinguishing variants between homologs, mutagenesis in Drosophila is typically performed using males which are mated to females carrying a balancer chromosome. A gene conversion event off the balancer chromosome within the stock would appear to be an “induced lesion” that is not a canonical EMS-induced mutation.

**ACKNOWLEDGEMENTS**

We would like to thank Casey Wimberly, Keith Smith and Michael Peterson for Sanger sequencing and help with data analysis. We would also like to thank Madelaine Gogol for Sanger sequencing primer design. We are grateful to Dorothy Stanley for editing of the manuscript and help with submission for publication. This research was supported by the
Stowers Institute for Medical Research, an American Cancer Society Postdoctoral Fellowship to JPB and an American Cancer Society Professorship to RSH.
LITERATURE CITED


### TABLES

#### Table 1. Run statistics

<table>
<thead>
<tr>
<th></th>
<th># reads millions</th>
<th>Basepairs millions</th>
<th>Genome Coverage</th>
<th>% Error rate</th>
<th>bp chrom 3 pass filter (%)</th>
<th>Both (%)</th>
<th>[Expected]</th>
</tr>
</thead>
<tbody>
<tr>
<td>A15</td>
<td>30</td>
<td>1080</td>
<td>8.7 X</td>
<td>0.84+/-0.05</td>
<td>39,604,870 (75.5)</td>
<td>37,165,510 (70.9)</td>
<td></td>
</tr>
<tr>
<td>791</td>
<td>29</td>
<td>1040</td>
<td>8.3 X</td>
<td>1.14+/-0.07</td>
<td>42,910,551 (81.8)</td>
<td>[61.8]</td>
<td></td>
</tr>
</tbody>
</table>

Statistics for A15 and 791 Illumina Genome Analyzer runs. The last column indicates the number of bases of the 3rd chromosome that pass the quality filter from both runs. The percentage of coverage expected, given independence between runs for nucleotides to pass the filter, is given in the last column.

#### Table 2. Annotation of verified nucleotide changes

<table>
<thead>
<tr>
<th></th>
<th>Total</th>
<th>Non-coding</th>
<th>Synonymous</th>
<th>Non-synonymous</th>
<th>Nonsense</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>103</td>
<td>82</td>
<td>10</td>
<td>9</td>
<td>2</td>
</tr>
<tr>
<td>‘Standard’</td>
<td>75</td>
<td>58</td>
<td>6</td>
<td>9</td>
<td>2</td>
</tr>
<tr>
<td>‘Anomalous’</td>
<td>28</td>
<td>24</td>
<td>4</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Lesions classified as ‘standard’ are more likely to have arisen by EMS mutagenesis and have functional consequence. Lesions classified as anomalous are more likely to have arisen by gene conversion and not have functional consequence. Five standard (3 non-coding, 1 synonymous and 1 non-synonymous mutation in the *Cad99C gene*) were located in the gene conversion clusters (see fig 2B,C) and have been shown to be gene conversion events.
1. Coverage and quality analysis of the 3rd chromosome from A15 and 791 runs. A) Distribution of nucleotide coverage depth for the original A15 3rd chromosome and for the 791 mutagenized 3rd chromosome. The heat map indicates pair wise coverage. B) Distribution of MAQ consensus nucleotide quality scores for A15 and 791 for nucleotides of the 3rd chromosome. Scores are only shown for consensus nucleotides that were not ambiguous and had a depth of at least 4. Heat map indicates pair wise quality.

2. Analysis of SNPs between the original A15 and mutagenized 791 chromosomes. A) Classification, verification and confirmation information for initial set of 165 candidate SNPs. B) Gene conversion clusters. For each SNP cluster, the 3RT nucleotide is shown above, the A15 is shown in the middle, and the 791 nucleotide is shown below. Yellow indicates identity with the 3RT nucleotide, red indicates a nucleotide that is different from the 3RT nucleotide. Relevant balancer sequence is shown to the right of each cluster, with inferred gene conversion event indicated with a red arrow. Relative spacing of SNPs is shown with a scale bar. C) Distribution of verified variants along the 3rd chromosome, EMS Canonical G/C to A/T differences above, Non-canonical EMS differences below. Gene conversion cluster of mutations are indicated by red stars.

3. Annotation of encore. A) A “C to T” transition turns the 1353 glutamine codon to a pre-mature stop. B) Complementation test of encore791 lesion. Embryos of mothers raised at 18 C were assayed for the fused dorsal appendage phenotype.
3RT indicates the target chromosome from which the \textit{A15} chromosome was derived.
A. 

**A15**

- Density vs Depth
- Depth vs Density

**791**

- Density vs Depth
- Depth vs Density

B. 

**A15**

- Log10(Count) vs Consensus Qual
- Consensus Qual vs Log10(Count)

**791**

- Log10(Count) vs Consensus Qual
- Consensus Qual vs Log10(Count)
A. Candidate SNPs 165
   - Checked by Sanger Sequencing Both Strains 125
   - Confirmed by Sanger 103

   'standard' 791 nucleotide matches ref. seq.
   75 (5 in clusters) (FP Rate: 6.25 %)

   'anomalous' 791 nucleotide doesn't match ref. seq.
   28 (21 in clusters) (FP Rate: 37.8 %)

   Canonical EMS 60 (80 %)
   Canonical EMS 12 (42.9 %)

B. 100 nt

   3RT
   A15
   791

   Canonical EMS 60 (80 %)
   Canonical EMS 12 (42.9 %)

   Confirmed by Sanger 103

   TM3
   C A A

   TM3
   G T T C

   TM3
   A G

   TM3
   A G A

   TM6
   G G A

   TM6
   G G A

   TM6
   G C A

C. cluster

   EMS Canonical
   EMS Non-canonical

   61C6 80D4 82F1 100D4
A. \textit{encore}^{791}

B. \begin{align*}
\text{N=} & \quad 1008 \quad 1064 \quad 796 \quad 1021 \quad 774 \quad 102 \quad 251 \\
\% \text{ fused} & \quad 100 \quad 80 \quad 60 \quad 40 \quad 20 \quad 0
\end{align*}