Exome sequencing reveals pathogenic mutations in 91 strains of mice with Mendelian disorders

Heather Fairfield,[1] Anuj Srivastava,[1,8] Guruprasad Ananda,[1,8] Rangjiao Liu,[2,8] Martin Kircher,[3] Anuradha Lakshminarayana,[2] Belinda S. Harris,[1] Son Yong Karst,[1] Louise A. Dionne,[1] Coleen C. Kane,[1] Michelle Curtain,[1] Melissa L. Berry,[1] Patricia F. Ward-Bailey,[1] Ian Greenstein,[1] Candice Byers,[1] Anne Czechanski,[1] Jocelyn Sharp,[1] Kristina Palmer,[1] Polyxeni Gudis,[1] Whitney Martin,[1] Abby Tadenev,[1] Laurent Bogdanik,[1] C. Herbert Pratt,[1] Bo Chang,[1] David G. Schroeder,[1] Gregory A. Cox,[1] Paul Cliften,[4] Jeffrey Milbrandt,[4] Stephen Murray,[1] Robert Burgess,[1] David E. Bergstrom,[1] Leah Rae Donahue,[1] Hanan Hamamy,[5] Amira Masri,[6] Federico A. Santoni,[5] Periklis Makrythanasis,[5,7] Stylianos E. Antonarakis,[5,7] Jay Shendure,[3,8] and Laura G. Reinholdt[1,8]

[1]The Jackson Laboratory, Bar Harbor, Maine 04609, USA; [2]The Jackson Laboratory for Genomic Medicine, Farmington, Connecticut 06032, USA; [3]Department of Genome Sciences, University of Washington, Seattle, Washington 98195, USA; [4]Department of Genetics, Washington University, St. Louis, Missouri 63130, USA; [5]Department of Genetic Medicine and Development, University of Geneva Medical School, 1211 Geneva 4, Switzerland; [6]Pediatric Department, The University of Jordan, Amman 11942, Jordan; [7]Service of Genetic Medicine, University Hospitals of Geneva, 1211 Geneva 4, Switzerland

Spontaneously arising mouse mutations have served as the foundation for understanding gene function for more than 100 years. We have used exome sequencing in an effort to identify the causative mutations for 172 distinct, spontaneously arising mouse models of Mendelian disorders, including a broad range of clinically relevant phenotypes. To analyze the resulting data, we developed an analytics pipeline that is optimized for mouse exome data and a variation database that allows for reproducible, user-defined data mining as well as nomination of mutation candidates through knowledge-based integration of sample and variant data. Using these new tools, putative pathogenic mutations were identified for 91 (53%) of the strains in our study. Despite the increased power offered by potentially unlimited pedigrees and controlled breeding, about half of our exome cases remained unsolved. Using a combination of manual analyses of exome alignments and whole-genome sequencing, we provide evidence that a large fraction of unsolved exome cases have underlying structural mutations. This result directly informs efforts to investigate the similar proportion of apparently Mendelian human phenotypes that are recalcitrant to exome sequencing.

[Supplemental material is available for this article.]

Causative mutation discovery provides the foundation for understanding the pathophysiology of genetic disorders. It also enables development of diagnostic assays and specifies therapeutic targets. Since the early 20th century (Cuenot 1905; Castle and Little 1910), the laboratory mouse has served as the primary model organism for understanding human Mendelian disorders, and in the era of genetic engineering it remains the most economical, genetically tractable model organism for both mechanistic studies and the development of therapeutics. With the convergence of massively parallel DNA sequencing and genome editing technologies, we are poised to enter a new era of disease gene discovery and parallel modeling between man and mouse.

In the 5 years since the first demonstrations of whole-exome sequencing (WES) in the context of Mendelian disorders (Choi et al. 2009; Ng et al. 2009), more than 100 underlying causative genes have been discovered using this approach. Similarly, pilot studies in the mouse demonstrated that implementation of WES could significantly increase the rate of Mendelian disease gene discovery in spontaneous mutant strains (Fairfield et al. 2011). These technological advances in mutation discovery have a significant impact in functional genomics since spontaneously arising alleles and allelic series provide more complete recapitulation of disease gene function than can be provided by null alleles alone (Antonarakis and Beckmann 2006).

Disease gene discovery by WES has been most successful for rare Mendelian disorders where there is limited locus heterogeneity and, often, supporting genetic data and evidence for causation

[8]These authors contributed equally to this work.
Corresponding author: laura.reinholdt@jax.org
Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.186882.114.

© 2015 Fairfield et al. This article is distributed exclusively by Cold Spring Harbor Laboratory Press for the first six months after the full-issue publication date (see http://genome.cshlp.org/site/misc/terms.xhtml). After six months, it is available under a Creative Commons License (Attribution-NonCommercial 4.0 International), as described at http://creativecommons.org/licenses/by-nc/4.0/.

25:948–957 Published by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/15; www.genome.org

(e.g., trio-sequencing for de novo mutations or multiple pedigrees for linkage analysis). Success rates have steadily improved as resources for human genetic variation have expanded (The 1000 Genomes Project Consortium 2012; Fu et al. 2013), providing deep reference data for filtering common genetic variation that drives causative gene discovery. However, causative gene discovery in Mendelian disorders still suffers from limitations as evidenced by its <50% success rate (Gilissen et al. 2012; Beaulieu et al. 2014). Possible causes for failed discovery by WES include poor or incomplete gene annotation, inefficient or incomplete exon capture, shortcomings of variant calling tools (particularly with respect to insertions/deletions [indels] and structural variation), insufficient ancillary information to successfully narrow catalogs of potentially causative variation, inaccurate phenotyping, or sample errors. Moreover, regulatory mutations that reside outside of coding regions will escape detection by WES.

We previously reported the development and application on a pilot scale of WES for discovery of spontaneous mutations for Mendelian disorders in the laboratory mouse (Fairfield et al. 2011). In contrast to human disease gene discovery, disease gene discovery in the mouse is highly powered by selective breeding, large consanguineous pedigrees, and genetically defined inbred strain backgrounds, each of which minimizes genetic heterogeneity. Moreover, causation can be readily supported through bulk segregation analysis and ultimately proven through complementation testing and/or genetic engineering.

Here we report a large-scale effort to identify the causative mutations for 172 distinct Mendelian disorders in laboratory mouse strains with clinically relevant phenotypes. This effort distinguishes itself from other large-scale functional genomic efforts in mice (e.g., The Knockout Mouse Project, KOMP) because it is phenotype driven, and unlike phenotype-driven saturation ENU mutagenesis projects, the molecular nature of spontaneous mutations is directly comparable to naturally occurring mutations in the human genome.

Results

We collected DNA samples from individuals representing 172 unique strains of spontaneous mutant mice, maintained by the Mouse Mutant Resource at The Jackson Laboratory and representative of a diverse spectrum of Mendelian disorders, including spondylocostal dysplasia, Hermansky-Pudlak syndrome, spinocerebellar ataxia, congenital myopathy, and many others. These strains were identified by animal care technicians on the basis of visibly apparent deviations from standard strain characteristics within a production scale vivarium that houses a population of nearly 1 million mice (Supplemental Table 1). The range of phenotypes found in our cohort of mice, therefore, is limited to phenotypes that are readily detectable in a vivarium setting and to any secondary, co-morbid phenotypes identified after further study. The major phenotypes include defects in behavioral/neurological function (27%), integument (10%), growth/size (9%), lifespan (9%), craniofacial development (8%), and skeletal morphology (8%) (Fig. 1).

Because chromosomal linkage data were available for the majority of the strains in our study, we selected a single affected individual for whole-exome sequencing. For the 17 strains lacking linkage data, we included an additional unaffected sibling control sample. Ten samples representing the most common inbred strain backgrounds were also included to maximize our ability to filter strain-specific genetic variation. Finally, eight of the 172 mutant strains were from colonies of ENU- (ethylnitrosourea) mutagenized mice.

We developed a pipeline for mouse exome analysis that takes into account (1) strain background by using high quality inbred strain-specific SNPs from the Sanger Mouse Genomes Project (Keane et al. 2011) for base quality recalibration through the Genome Analysis Toolkit (GATK) (McKenna et al. 2010; DePristo et al. 2011), and (2) custom variant filters based on coverage, variant quality, mapping quality, presence/absence in dbSNP, overlap with simple repeats, and observed variant frequency, stemming from the accumulated false discovery data from our pilot exome sequencing efforts (Fairfield et al. 2011). Genomic annotations were assigned to the variant calls and summary sample metadata (e.g., strain, phenotype, phenotype status, mode of inheritance, read coverage, etc.) were compiled for each sample.

Mouse exome variation

The total number of raw variants called per sample depended on the relatedness of each strain to the laboratory mouse genome reference strain, C57BL/6J, *Mus musculus domesticus*. As expected, exome data sets from strains within the same *M. m. domesticus* subspecies contained ~1500 to ~120,000 variants per exome, while data sets from the two other major subspecies represented in our sample set, *M. m. musculus* and *M. m. castaneous,* contained more than 300,000 variants per exome. As expected, samples with the

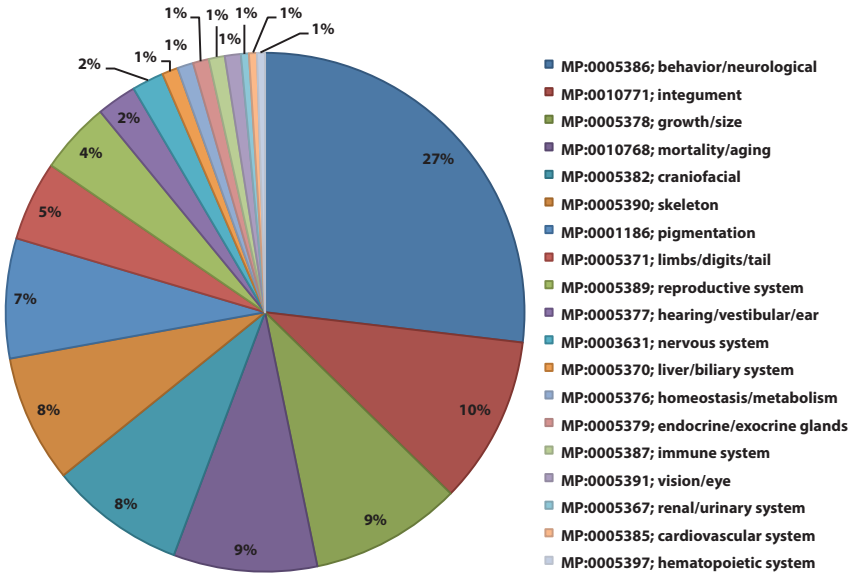

Figure 1. Phenotypic distribution of spontaneous mutant strains in the study. The cohort of mutant strains selected for exome sequencing represents various phenotypes that have an observable common characteristic during the course of normal breeding and husbandry. System level or tissue level mammalian phenotype (MP) terms were assigned on the basis of primary phenotypes. Phenotypes are arranged clockwise from largest to smallest group and similarly from *top* to *bottom* in the key.

fewest exome variants were from the same inbred strain background as the mouse reference genome, C57BL/6J. For example, in a C57BL/6J control exome, approximately 18 inbreeding generations removed from reference (see Discussion), we found a total of 1547 unique variants. Seventy percent (1083) of these passed quality filters and of those, there were 117 homozygous calls that were shared between other C57BL/6J mutant strains in our study as well as mixed background strains with C57BL/6J contributions. There were six private (not found in any other sample in our study) calls found in this sample. Of these six variants, two were heterozygous calls, one was homozygous, and the remaining three had low allele ratios (<0.3) indicative of false positives.

There were 943 variants (SNPs/indels) common across 75% or more of the samples in our study regardless of strain background. To investigate the origin of these variant calls, we examined a subset (272 that were found in 100% of the samples) and found that the majority (92%) were clustered (more than two variants within 1 kb) in the genome. To test the idea that these could be false calls arising from the underlying genome reference assembly, we compared the variant positions to the coordinates of known assembly issues cataloged by the Genome Reference Consortium (GRC) incident database (http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/index.shtml; data available for FTP at ftp://ftp.ncbi.nlm.nih.gov/pub/grc/). We found that most (151 variants) overlap with reported errors in the reference sequence assembly (empirical P-value = 9.999×10^{-5}; mean = 28.74, SD = 5.02). The remaining common variants (121 variants) are likely due to underlying assembly issues that are as yet unreported.

Genetic variation among inbred strains heavily influences disease susceptibility and phenotypic variability; failure to consider genetic background is a major contributor to irreproducibility of mouse studies (Gerlai 2001; Linder 2001; Wolfer et al. 2002; Perrin 2014). The Sanger Mouse Genomes Project (MGP) maintains a growing resource of variation data from inbred strains and for 17 of these strains, the sequenced individuals were females sourced from foundation stocks at The Jackson Laboratory (Keane et al. 2011). We compared variant data from our inbred strain control samples (males, also sourced from The Jackson Laboratory) to the MGP variant data and found that the vast majority (>~93%) of variant calls were shared between the data sets. Unique calls were primarily due to differences in sequencing coverage, minimum coverage requirements for SNP/indel calling, and differences in quality filters between the two projects. This also explains the smaller proportion of common indel calls (>~80%), as indels are lower quality overall. A much smaller proportion of the unique calls were due to differences between the samples (sex, inbreeding generation) (Supplemental Table 2). To further examine inbred strain background and relatedness across our samples, we used a hierarchical clustering method. Using this method, we successfully identified the strain origin for nine samples with "unknown" strain origin. We also identified six samples for which strain background was incorrectly assigned (Supplemental Fig. 1) 107:S10, 18:S06, 23:S02, 76:S02, 78:S02, and 102:S00).

Discovery and validation of putative pathogenic mutations

We developed the Mouse Mutant Resource Database (https://mmrdb.jax.org) to host annotated variant calls and sample metadata and to facilitate data sorting, filtering, querying, and sharing. The database employs an algorithm for variant prioritization. The algorithm makes the following assumptions about causative variants: they will be rare (<3%) in the database, the allele ratio of the variant in the sample will fall within expectations for the sample genotype (>0.9 homozygous; 0.2–0.8 heterozygous), and the chromosomal position of the variant will be in agreement with chromosomal linkage data. We optimized the algorithm iteratively by reanalyzing exome data sets with previously confirmed, known mutations (Fairfield et al. 2011).

The mutation candidate algorithm flagged 8360 variants across 172 exomes and 1918 of these were variants with high (e.g., frameshift, exon deleted, start/stop lost, splice site, rare amino acid change) or moderate impact (e.g., codon change, deleted UTR) annotations. Functional annotation of variants was accomplished using SnpEff and ANNOVAR, each of which generate functional predictions on the basis of the genomic location of a variant with respect to coding sequence and the type of amino acid change that is predicted (if any). These tools are limited in their capacity to predict pathogenicity on the basis of amino acid conservation, RNA processing, transcriptional regulation or translation, and post-translational modification. Because our typical mapping panel consisted of ~12 affected and 12 unaffected individuals from a pedigree, we could further refine our map positions to subchromosomal intervals of 30–60 cM. This information contributed to a significant reduction in our candidate list from 1198 to 108 putative pathogenic mutations. Mutations were validated using Sanger sequencing of PCR amplicons and in some cases, RT-PCR to genotype affected and unaffected individuals from each pedigree. Using this approach, we validated putative pathogenic mutations in 78 strains. These mutations were in 62 genes, each of which has a single human ortholog (Table 1; Supplemental Tables 3, 4). Six of these were previously reported in our pilot study and used here as validated data sets to support the development of our pipeline (Fairfield et al. 2011). The plurality of the mutations discovered were missense mutations (43%) followed by nonsense mutations (21%), single nucleotide mutations in canonical splice sites (12%), and small indels (13%) (Fig. 2A). Using whole-genome sequencing and manual analysis of exome alignment data, we identified "exome-recalcitrant" structural mutations (insertions, deletions, duplications >50 base pairs [bp]) in an additional 13 strains, representing 11 additional genes (Fig. 2A; Supplemental Tables 3, 4) (see "Exome-recalcitrant mutations" below). Taken together, we found 89 putative pathogenic mutations in 73 genes across 91 mutant strains.

Approximately 11% (10/89) of the mutations discovered were in genes that have yet to be associated with a mouse phenotype; these novel genes were *4732456N10Rik*, *4930453N24Rik*, *Ddx10*, *Kntc1*, *Rpl31* (two mutations), *Myo10*, *Fdxr*, *Otop3*, and *Golga1* (Table 1). While little is known about the function of these genes in mice, all are well conserved in vertebrates.

In addition to the novel genes, 37% (33/89) are new alleles of mouse genes that have not yet been associated with a human Mendelian disease. To determine if our data could be used to inform unsolved human exome projects, we used the GeneMatcher tool (http://www.genematcher.org/) to compare our gene list to candidate gene lists from unsolved human exome sequencing projects. This search resulted in a corresponding match (1/73 genes) for *Ap3b2*/*AP3B2*, where a splice site variant (NM_004644.3: c.588 + 1G > T, Chr 15: 83349863 C > A, in hg19) was found in two individuals from a consanguineous family. These individuals were diagnosed with hypotonia, developmental delay, tonic-clonic seizures, and visual impairments. Similar to the clinical symptoms reported in this family, our recessive *Ap3b2* mutation is associated with behavioral and neurological phenotypes including tonic-clonic seizures (Table 1). Moreover, an engineered knockout allele

Table 1. A subset of the pathogenic mutations discovered by exome sequencing in mice with Mendelian disorders

Gene symbol	Allele name (symbol)	Human ortholog	Human disease association(s)/OMIM	Molecular description (Annovar)	Mutation category	Inheritance	Primary phenotype
4930453N24Rik[a]	dense incisors *(din)*	*C3orf38*		*4930453N24Rik*: NM_026273: exon3: c.T729A:p.C243X	Nonsense	Recessive	Craniofacial
Ap3b2	mutation 2 Jackson *(m2J)*	*AP3B2*		*Ap3b2*:NM_021492: exon12:c.C1303T:p. R435X	Nonsense	Recessive	Behavior; neurological
Ddx10[a]	mutation 1 Jackson *(m1J)*	*DDX10*		*Ddx10*:NM_029936: exon15:c.C2208A:p. D736E	Missense	Recessive	Craniofacial
Fdxr[a]	mutation 1 Jackson *(m1J)*	*FDXR*		*Fdxr*:NM_007997: exon10:c.G1166A:p. R389Q	Missense	Recessive	Behavioral; neurological
Kntc1[a]	jagged tail like *(jgl)*	*KNTC1*		*Kntc1*:NM_001042421: exon30:c.C2596T:p. R866X	Nonsense	Recessive	Reproductive; skeletal
Myo10[a]	mutation 1 Jackson *(m1J)*	*MYO10*		*Myo10*:NM_019472: exon25: c.2845_2853A	Small deletion	Recessive	Pigmentation, skeletal
Otop3[a]	mutation 1 Jackson *(m1J)*	*OTOP3*		5′ UTR; NM_172801: c.190G>A	5′ UTR	Recessive	Behavior; neurological
Rpl31[a]	dominant tail short *(Dts)*	*RPL31*	Diamond Blackfan anemia	3′ UTR; NM_001252218: c.12_16delinsC, NM_001252219: c.12_16delinsC, NM_053257: c.12_16delinsC	3′ UTR	Dominant	Limbs/digits/tail
Rpl31[a]	dominant tail short 2 Jackson *(Dts-2J)*	*RPL31*	Diamond Blackfan anemia	3′ UTR; NM_001252218: c.13_26delinsA, NM_001252219: c.13_26delinsA, NM_053257: c.13_26delinsA	3′ UTR	Dominant	Skeletal; limbs/ digits/tail
Tshr	hypothyroid 3 Jackson *(hyt-3J)*	*TSHR*	Hyperthyroidism, nonautoimmune	Large deletion; ~200 bp	Large deletion	Recessive	Growth/size
Wnt7a	postaxial hemimelia Jackson *(px-J)*	*WNT7A*	Fibular aplasia or hypoplasia, femoral bowing and poly-, syn-, and oligodactyly	Splice donor; deletion > 20 kb	Large deletion	Recessive	Skeletal; limbs/ digits/tail

For each confirmed mutant allele, the mouse gene symbol, the human ortholog, associated human disease where known, inheritance, allele annotation (ANNOVAR) as predicted by our pipeline and the primary clinical phenotype by Mammalian Phenotype terms (system or tissue level) are provided. If the same mutation was found in more than one strain, that allele is represented only once in this table, as is only one associated human disease per allele. For the full table, see Supplemental Table 3.
[a]Novel mouse genes for which we report the first allele and phenotype.

of *Ap3b2* showed identical phenotypes (Seong et al. 2005), consistent with the conclusion that our newly discovered nonsense mutation is likely a null allele and that the splice site mutation found in the human exome data is likely to be pathogenic. This is the first pathogenic mutation reported for human *AP3B2* and the first association of this gene with a Mendelian disease.

Chromosomal linkage and evidence for causation

There were 17 strains in our study for which chromosome linkage data were not available and only one of them was solved (MMR_1370, $Ddx10^{m1J}$). The remainder had far too many plausible candidates and in one case a significant false discovery was made. A candidate mutation was found in *Gm15448*, which is one of several orthologs to human *LILRA6* (leukocyte immunoglobulin-like receptor [LIR], subfamily A, member 6), that encodes a poorly characterized LIR receptor expressed in monocytes and capable of associating with an FcR gamma protein family member (Bashirova et al. 2014). The inflammatory disease phenotypes (dermatitis, enlarged lymph node and spleen, arthritis, osteomyelitis, etc.) of affected mice (dwss [dermatitis with small size]) (Supplemental Table 1) further supported the case for this variant. Co-segregation analysis in a limited number of individuals from the pedigree (approximately three affected and three unaffected) also supported pathogenicity. However, chromosomal linkage mapping was completed during our study and the *Gm15448* variant was ultimately excluded. We also found one case where existing chromosome linkage data were insufficient. Two linked candidate mutations (<1 Mb apart) were identified for a single ENU strain (ENU strain MMR_1400716, $Fdxr^{m1J}$ and $Otop3^{m1J}$), which is not unusual given the high mutation rates induced by ENU mutagenesis. These strains often require more extensive mapping or bulk co-segregation analysis. When we retrospectively examined the impact of chromosomal linkage data, we found it reduces the total number of mutation candidates by up to two orders of magnitude (Fig. 2B). Therefore, in addition to minimizing false positives,

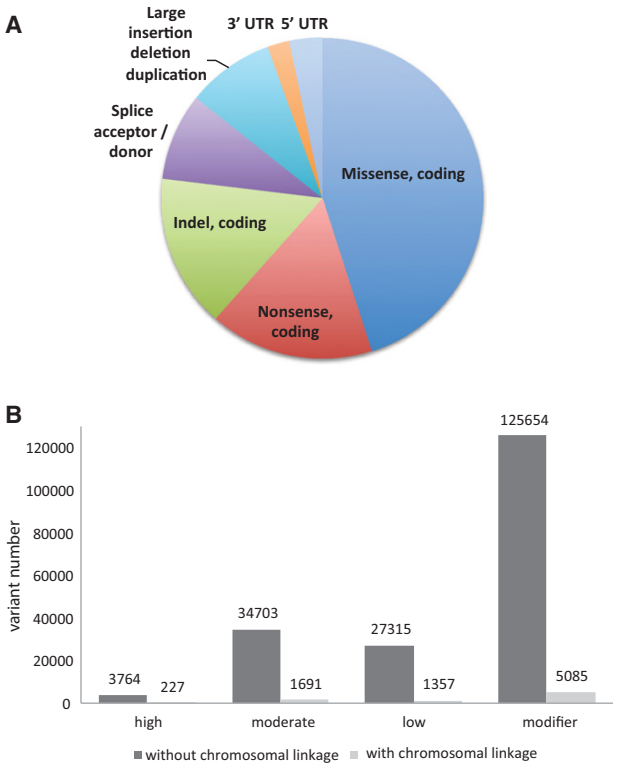

Figure 2. Distribution of pathogenic mutation types (*A*) and the value of chromosomal linkage (*B*) for mutation discovery in spontaneous mouse models of Mendelian disease. Pathogenic mutations consisted of a variety of lesions, the majority of which were single nucleotide substitutions. Due to ascertainment bias, copy number variants and structural mutations (>50 bp) were more rare (*A*). Chromosomal linkage data had a significant impact on the validation burden and a potential for false positive mutation calls. The largest effect (two orders of magnitude) was on potentially low-impact (modifier) variant calls. Variant calls were categorized by predicted impact according to SnpEff impact annotations (*B*) (see Methods and Supplemental File 4).

chromosomal linkage data significantly reduce validation burden for the price of a relatively small mapping cross, and importantly, these data strengthen the evidence for causation.

Exome-recalcitrant mutations

For nearly half of the samples in our study, the pipeline failed to predict a single causative mutation. A subset of these had strong candidate genes within mapped intervals, so we examined any variant calls within these genes. We found that many of these variants were rare but had unexpected allele ratios given sample genotype (e.g., <0.8 when a homozygous variant was expected). Moreover, our attempts to validate these variants often resulted in failed PCR reactions. When we examined the local alignments around the variant calls, we found evidence for large copy number variants or structural rearrangements (deletions, duplications, or insertions). Three examples are shown in Figure 3; two are deletions in *Tshr* and *Wnt7a*, and one is an insertion in *Myo5a* (Table 1; Supplemental Tables 3, 4). All three of these genes have previously reported mouse alleles with identical phenotypes, lending support to the pathogenicity of the structural lesions reported here (Stein et al. 1994; Huang et al. 1998a,b; Parr et al. 1998).

Based on these data, we surmised that at least one category of exome-recalcitrant mutation might be larger copy number variations or structural rearrangements. Alternatively, exome-recalcitrant mutations may simply be regulatory mutations that reside outside of the exome or are coding mutations occurring in poorly annotated regions. To explore this, we used whole-genome sequencing (WGS). We selected five strains for which our initial attempts failed to identify causative mutations. For each of these mutant strains a control with matched strain background was also included. In addition to standard BWA alignment and SNP/indel calling, we used Pindel (Ye et al. 2009; Handsaker et al. 2011) for detection of large structural mutations. We discovered and validated causative mutations in four of the five mutant samples (Table 2). All four lesions involved coding sequence from a single gene and represented a variety of mutation types; *sunk* was an ~2.4-kb deletion, *hstp* was an ~300-kb duplication, *whnl* was a small 7-bp insertion, and *bucp* was a SNP that went undetected by whole-exome sequencing due to poor coverage. Taken together, our analysis of 13 exome-recalcitrant mutations revealed that most are large copy number variants or structural mutations involving coding sequence (Supplemental Fig. 2).

Discussion

We sought to identify putative pathogenic mutations in 172 strains of mice exhibiting a variety of clinically relevant, Mendelian disease phenotypes. Using an optimized pipeline for analysis of mouse exome data and a database that integrates sample and variant data across strains (http://mmrdb.jax.org), we nominated and validated putative pathogenic mutations for 91 (53%) of our sequenced strains and these mutations were in 73 genes. Our complete data set consists of more than 4 million exonic variants and is a coding variation resource of extraordinary depth and breadth for laboratory mouse strains. Our publicly available database provides tools for reproducible exome data analysis and mining through user-defined searches, as well as candidate calling for causative mutations that is informed by sample metadata and relevant variant features. The total number of raw variants found per strain per exome ranged from ~1500 to more than 300,000 depending on the relatedness of the strain to the C57BL/6J reference. By comparison, a typical human exome data set contains ~25,000, highlighting the allelic diversity across laboratory mouse strains and the genetic consequences of both inbreeding and short generation times.

Inbred strains and genetic drift

Ninety-seven percent of the variants that we found in any given C57BL/6J-related sample were also found in other C57BL/6J samples, but not (by definition) in the C57BL/6J reference. The mouse reference genome is a haploid assembly consisting of sequence calls from multiple individuals spanning several generations of inbreeding (approximately F208–F214) at The Jackson Laboratory (Mouse Genome Sequencing Consortium 2002; The Jackson Laboratory, pers. comm.). The origin of these variants could be "residual" heterozygosity at F208–F212, not captured in the final haploid assembly.

Mutation is inexorable and in an inbred strain it is the underlying cause of heterozygosity in the absence of heterozygous selection and genetic contamination. We could trace the origins of the C57BL/6J and C57BL/6J-related strains in our study to generations >F226, the generation of the archived C57BL/6J embryo stock that is now used to maintain the C57BL/6J production colonies at The Jackson Laboratory (Genetic Stability Program [Taft et al. 2006]).

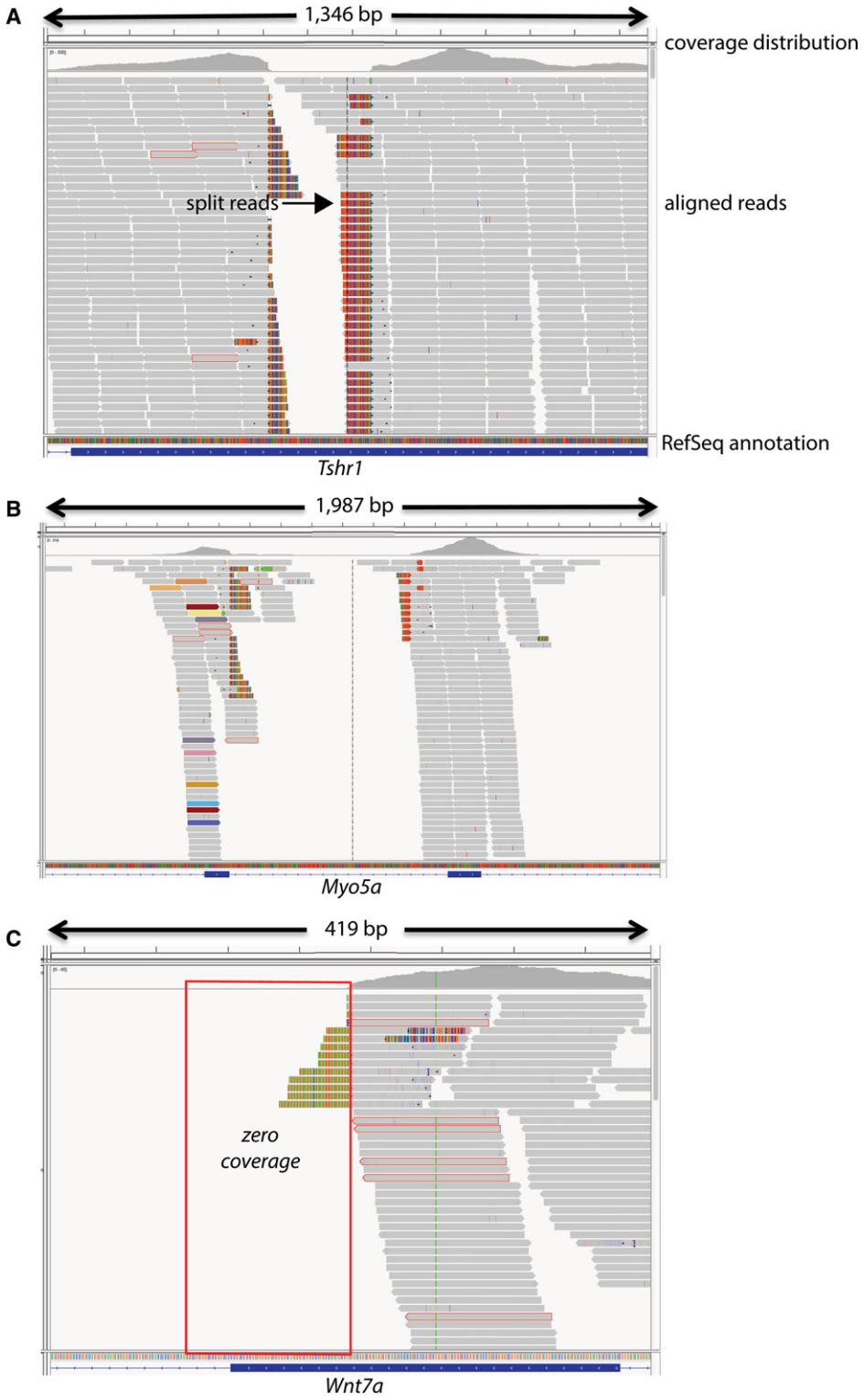

Figure 3. Graphical view of alignments across *Tshr* (*A*), *Myo5a* (*B*), and *Wnt7a* (*C*). Graphical views of the alignments were generated using the Integrative Genomics Viewer (IGV) and RefSeq exon annotations are shown. In each case, split reads (arrow) span the junctions of copy number variations and structural rearrangements. In *Tshr*, a cluster of four single nucleotide variants (SNVs) with unexpected allele frequencies of 0.3–0.63 was called in a homozygous sample; three of the four SNVs were soft filtered as a SNP cluster by GATK. Manual analysis of the alignment revealed a homozygous deletion in the final exon of this gene (*A*). In another example, a heterozygous SNV was called in a splice donor site of myosin VA (*Myo5a*) in a sample. In the alignment surrounding the SNV call there were split reads, as well as flagged reads (*B*, colored reads) with mates mapping throughout the genome, providing evidence of a retroviral or intra-cisternal A-particle (IAP) insertion in exon 3 (*B*). In a third example, a SNV call was flagged by our algorithm as a mutation candidate but could not be validated due to multiple failed PCR assays. The SNV was in wingless-related MMTV integration site 7a (*Wnt7a*) in an affected sample from a pedigree with recessive skeletal abnormalities. Manual analysis of the alignment surrounding the SNV call revealed two clusters of flagged reads flanking an ~23-kb region, spanning intron 3, the 5′ splice site, and a portion of exon 3. Moreover, there was zero coverage across exon 3 and the 5′ splice site of intron 3, regions that are normally covered by WES (*C*).

Table 2. Exome-recalcitrant mutations discovered by whole-genome sequencing

Allele name	Gene	Lesion	Chromosome location (GRCm38/mm10)
sunken *(sunk)*	*Samd4*	Large deletion	Chr 14: 46,882,440-46,884,934
highstepper *(hstp)*	*Rorb*	Large duplication	Chr 19: 19,010,566-19,336,743
buttercup *(bucp)*	*Slc6a19*	SNP/indel	Chr 13: 73819606
atypical hair loss *(aphl)*	Unknown	Unknown	Unknown
witchnails *(whnl)*	*4732456N10Rik*	7-bp insertion	Chr 15: 101,553,337

Five strains for which mutations were not discovered by exome sequencing alone were selected for whole-genome sequencing and mutations were found in four. The mutations impacted coding sequence in every case but were not originally discovered by WES due to the nature of the mutation (larger structural mutations) or due to poor coverage.

Published rates of spontaneous mutations in mammals range from 10^{-5} to 10^{-6} per locus per gamete (Schlager and Dickie 1967; Bailey 1978), based on breeding and specific locus testing, or 10^{-8} per bp per generation, based on sequencing data (Segurel et al. 2014). Given a per locus per gamete mutation rate of 10^{-5}, the expected number of spontaneously arising genic mutations that become fixed through inbreeding is one in five generations, according to Bailey (1978); therefore, three to four fixed homozygous coding variants are expected to have occurred by F226 compared to the reference (F208–F212). Similarly, given the estimated mutation rate of $0.5–3 \times 10^{-8}$ per base pair per generation (Segurel et al. 2014), ~12–60 mutations are expected to arise through genetic drift, genome-wide, per generation or approximately one coding mutation per generation (assuming that ~2% of the mouse genome is protein coding [Mouse Genome Sequencing Consortium 2002]). Assuming a rate of fixation similar to the above predictions, approximately three to four coding homozygous mutations are expected between F226 and F208–F212. Our observation of one homozygous "exome" mutation in a C57BL/6J sample at F230 is slightly lower, but this is not unexpected since this analysis was limited to single nucleotide variants and small indels.

Novel genes

Approximately 11% of the mutations we discovered were in novel genes—genes for which phenotypes have not previously been associated. While this is not a substantial fraction of our data set, it demonstrates that novel gene discoveries remain after more than 90 years of spontaneous mutation research at The Jackson Laboratory. Some examples include *Golga1*, *Rpl31* (two mutations), and *Ddx10*.

Golga1 (golgi autoantigen, golgin sumbfamily a, 1) is a gene that encodes Golgin 97, a component of the *trans*-golgi network. Golgin 97 is a protein that is poorly characterized at a functional level but nonetheless is a commonly used marker for imaging the *trans*-golgi network. Mice homozygous for the recessive *awag* (ages with abnormal gait) allele carry a homozygous deletion in *Golga1* and exhibit a late onset abnormal gait with a pronounced tremor, consistent with a neuromuscular defect. As in many other cell types, vesicular trafficking is critical to normal neuron function and has been implicated in a variety of neurodegenerative diseases; the *awag* mouse model provides the first opportunity to determine the function of *Golga1* in this context. While there is a human ortholog of *Golga1*, this gene has yet to be associated with a human disease.

We also found two dominant mutations (*Dsht* and *Dsht-2J*) in *Rpl31*, which encodes the ribosomal protein L31. Heterozygous mice have short, kinked tails, and homozygous mice die during embryogenesis. Recently, a mutation in the human homolog of this gene was implicated in Diamond Blackfan anemia (Farrar et al. 2014). While the $Rpl31^{Dsht}$ and $Rpl31^{Dsht\text{-}2J}$ heterozygous mice have not been tested for anemia, the phenotype of these mice is strikingly similar to mice carrying mutations in the related ribosomal subunit, *Rpl38*. $Rpl38^{Ts}$ mutant mice also have dominant skeletal defects (short tail), recessive embryonic lethality, and hematopoietic defects (Kondrashov et al. 2011). Therefore, more extensive phenotyping of mice carrying the $Rpl31^{Dsht}$ and $Rpl31^{Dsht\text{-}2J}$ alleles will provide in vivo validation of the role of *Rpl31* in Diamond Blackfan anemia and provide animal models for mechanistic studies.

Ddx10 encodes a DEAD box containing ATP-dependent RNA/DNA helicase that is likely to be involved in RNA biogenesis. $Ddx10^{m1J}$ is a missense mutation adjacent to the coiled-coil domain of the protein that is likely to be involved in protein-protein interactions. Mice homozygous for this mutation have craniofacial defects, including short and split nose/maxilla. While additional data are needed to demonstrate the functional consequences of the amino acid change and its potential role in craniofacial development, an interesting association was recently reported in a child with intellectual disabilities and craniofacial abnormalities. In this patient, array comparative genome hybridization revealed a 170-kb microdeletion at 11q22.3, a region containing only a single gene, *DDX10* (Kashevarova et al. 2014). The clinical features described in the patient included absence of speech, ADHD, convergent strabismus (paralytic strabismus), lateral nystagmus, hypermetropic astigmatism, low-set ears, pear-shaped nose, and narrow face. While the study reports lack of parental DNA to support the pathogenicity of the microdeletion in the case study, our data provide new and compelling evidence to support pathogenicity (Kashevarova et al. 2014).

New alleles and new disease associations

In addition to novel genes, 37% of the mutations we found were new alleles of genes that have not yet been associated with a human disease. One caveat is that these associations rely on existing annotation and curation of published data. Clinical exome sequencing projects are now generating disease variant data at unprecedented rates and very little of it is published. To compare our data to unpublished human exome variant data, we used a publicly available exome data sharing resource for Mendelian disease (i.e., GeneMatcher, http://www.genematcher.org), and we were able to identify a match between our relatively short list of new mouse alleles lacking previous human disease associations and candidate genes from unsolved human exome projects. Extrapolating these results to the full set of phenotype-associated mouse genes that have human orthologs but do not have human disease associations (~2000, data from MouseMine, Mouse Genome Database), we predict that nearly 100 novel associations could be made through simple data integration with the existing gene lists available through GeneMatcher. While this integration alone is not sufficient to establish causality, it provides critical evidence as well as potential disease models with little additional effort.

Structural mutations

The vast majority of the mutations we discovered were single nucleotide changes or small insertions/deletions (indels). However, this is clearly an ascertainment bias. Our analysis of exome-recalcitrant mutations revealed that the vast majority were larger CNVs (duplications, deletions > 50 bp) and structural rearrangements (insertions) that escape detection by standard WES pipelines. The mouse genome harbors active endogenous retroviral elements (ERVs) that underlie the vast majority of de novo insertions as well as mutant allele reversion events, especially in strains with particularly high transposition activity, like C3H (>50% compared to C57BL/6J) (Maksakova et al. 2006). In fact, of the 52% of exome cases that we solved, the C3H inbred strain background was underrepresented.

All of the deletions, duplications, and insertions that were validated involved coding sequence, and many were below the detection limit for commercial array-based CNV discovery tools (<20 kb). A significant fraction of the exome failures in our study could be resolved by optimizing the repertoire of analytical tools available for discovery of structural exome variants for WES and/or by transitioning to WGS. This result also directly informs efforts to investigate the similar proportion of apparently Mendelian human phenotypes that are recalcitrant to exome sequencing.

Conclusions

High-throughput sequencing technologies have revolutionized the process of Mendelian disease gene discovery. Our efforts to identify putative pathogenic mutations in mice using high-throughput sequencing demonstrate the power of this technology, as well as its limitations. Understanding the mechanism of Mendelian disease genes is essential to the development of therapeutics, and while some mechanistic studies can be accomplished in patient-derived cell lines, most require a physiological context that can only be provided by a mammalian animal model. Moreover, human development is complex and many Mendelian disease genes have essential roles during development that can only be recapitulated in vivo. We anticipate that mouse models of human disease, whether engineered or spontaneously occurring, will continue to provide essential tools for mechanistic studies as well as preclinical research.

Methods

Exome sequencing

Spleen and tail tissue were collected from 172 affected mice, each representing a unique pedigree with a Mendelian disorder. An additional 27 samples were collected from unaffected siblings and from strain background controls. Samples and linkage data were provided by the Mouse Mutant Resource at The Jackson Laboratory. Genomic DNA was extracted by phenol chloroform extraction of nuclear pellets or by using a Qiagen DNeasy Blood and Tissue kit (Qiagen). Illumina paired-end (PE) libraries (2×76 or 2×100) and liquid phase sequence capture were performed as previously described using the Roche NimbleGen SeqCap EX Mouse Exome Design (110624_MM9_Exome L2R_D02_EZ_HX1, #99990-42611) (Fairfield et al. 2011). Enriched libraries were sequenced on the Illumina GAIIx or the Illumina HiSeq (Illumina).

Mapping and variant analysis

Sequencing reads were subjected to quality control using NGS QC Toolkit v2.3, and reads with base qualities ≥ 30 and >70% of read length were used in the downstream analysis (Patel and Jain 2012). High-quality reads were mapped to the mouse genome (GRCm38, mm10) using BWA v0.5.10-tpx (Li et al. 2009) with default parameters. The resulting alignment was sorted by coordinates and further converted to binary alignment map (BAM) format by Picard v1.95 SortSam utility (http://picard.sourceforge.net). The Picard MarkDuplicates module was used to remove the duplicates from the data. The Genome Analysis Tool Kit (GATK) v2.2-16 (McKenna et al. 2010; DePristo et al. 2011) module IndelRealigner and BaseRecalibrator were used to preprocess the alignments. During base quality recalibration, dbSNP variants were used as known sites, and only variants obtained from the closest strain (based on the strain background of the sample) were used for training. If the closest strain information was not available or unknown, then the base recalibration step was skipped. Target-capture efficiency metrics were determined using Picard HsMetrics. The realigned and recalibrated BAM file was used as an input to GATK UnifiedGenotyper at parameters, -stand_call_conf 50.0, -stand_emit_conf 30.0, -dt NONE. Variant calls were restricted to the target regions (110624_MM9_Exome L2R_D02_EZ_HX1, #99990-42611). If the BQSR step was skipped during preprocessing, then the –baq RECALCULATE parameter was turned on during variant calling. Finally, raw variant calls were soft filtered using GATK VariantFiltration based on the following parameters: LowCoverage (DP < 5), LowQual (30 < Q < 50), VeryLowQual (Q < 30), StrandBias (FS *P*-value > 60), SNV cluster (three or more SNVs within 10 bp), Poor Mapping quality (>10% of reads have non-unique alignments). Variants were annotated by SnpEff v2.0.5 (Cingolani et al. 2012) and ANNOVAR (Wang et al. 2010). Those of highest impact were reported by GATK VariantAnnotator.

Hierarchical clustering

Within each sample, genes were coded "1" or "0" depending on the presence or absence, respectively, of high-impact mutations —comprising frameshift, stop gain/loss, start loss, splice site acceptor/donor variation, and rare amino acid change. The distance matrix of similarity between samples based on binary-coded gene values was computed using dist.binary() function from the "ade4" R package (Dray and Dufour 2007)—the Sokal & Michener distance metric, which computes the proportion of genes that are co-present (1 in both) and co-absent (0 in both) between samples, was considered. Hierarchical clustering on the distance matrix was performed using the hclust() function. Significance of the resulting clusters was assessed via multiscale bootstrap resampling (with 1000 iterations) using functions from the R "Pvclust" package (Suzuki and Shimodaira 2006). The resulting approximately unbiased (AU) *P*-values were incorporated into the clustering dendrogram: Clusters with AU *P*-value > 90% can be considered as stable clusters at a 10% level of significance.

False negative rate (FNR) and false discovery rate (FDR) calculations

The FNR for SNP calling was calculated for six of the inbred strains in our study (AKR/J, BALB/cJ, NOD/ShiLtJ, NZO/HILtJ, PWK/PhJ, C3H/HeJ). The Broad2 (Kirby et al. 2010) and dbSNP data obtained from the Mouse Phenome Database (MPD) (http://phenome.jax.org) (Kirby et al. 2010) were used as a truth set. Variant calls

from MPD and background samples were restricted to the target region. Regions with no aligned reads (determined by obtaining regions with zero median coverage across 60 randomly selected strains) were also removed during analysis. Further, only homozygous variants were selected for analysis. Variants present in both background strain and C57BL/6J in MPD were considered as true positives, and variants unique to C57BL/6J were considered as false negatives, i.e., considered missed by our analytical pipeline.

The FDR for SNP calling was calculated by regenotyping SNPs across the same six inbred strains using Sequenom MASSArray iPlex technology (Gabriel et al. 2009). Six hundred SNVs/indels were randomly selected for assay design. Approximately 10% of this set consisted of variant calls that were soft filtered. Of these, 27 had design issues due to local repetitive sequence. Of the remaining 573 assays, 488 had a success rate of >66.7% (at least 4/6 samples with calls). The FDR was calculated (coverage ≥ 5) by dividing the number of incorrect calls (false positives, FPs) by the total number of calls (true positives, TPs, plus FPs). The FDR and FNR data are shown in Supplemental Table 5.

Database development

The database was developed as a research module of the JAX Comprehensive Genome Analytics (CGA) system. The CGA system integrates analytical genomics pipelines, with a relational database of genomic variants and expression profiles organized into molecular signatures that are associated with biological metadata and specific sample records. The Mouse Mutant Resource (MMR) module correlates molecular data with mouse strains and phenotypes. The database backend runs on a MySQL server relational database management system. A custom data loading process parses variant call format (VCF) files and calls the mutation identification algorithms described above while loading the variant records. The database user interface is a Java Server Faces (JSF) web application hosted in Apache Tomcat 7.0. The system is tested to run on Mozilla Firefox browsers on both Macintosh and Windows PC computers. The web application is accessible on the Internet at https://mmrdb.jax.org. User accounts require pre-approval to search on non-public samples and variants.

Mutation validation

PCR primers flanking each candidate mutation were designed using Primer3 software, and PCR amplicons were Sanger-sequenced on an ABI 3730×l DNA sequencer (Applied Biosystems) (Supplemental Table 6). Sanger sequencing of PCR amplicons was performed on the originally sequenced sample as well as up to five additional affected and unaffected individuals from the same pedigree. Sequencing data were analyzed using Sequencher 5.0 (Gene Codes Corp.).

Whole-genome sequencing

Six samples, three positive controls and three background samples, were sequenced on two full HiSeq flow cells with 101,101 paired-end reads and a 7-bp sample barcode read. Reads in each lane were split by barcode, allowing errors in barcode identification but enforcing a minimum quality score of 10 (Meyer and Kircher 2010; Kircher et al. 2012). Sequencing adapters were trimmed using information from the paired-end reads; resulting reads shorter than 50 bp were removed and all other reads were quality filtered (removing reads with more than five bases below a Phred quality score of 10) (Kircher et al. 2011). Processed reads were aligned using BWA to the mouse reference genome (NCBIm37, mm9) with default parameters; about 12×–16× average coverage was obtained (assuming an accessible genome size of 2.7 Gb). The BWA output was directly BAM-converted and sorted using SAMtools (Li et al. 2009); read groups were added using pysam/SAMtools API (Li et al. 2009) and BAM files merged by sample and subjected to a GATK v2.2-8 realignment process. SNVs and indels were called for all sites in the genome using GATK's UnifiedGenoTyper (without filtering for quality of the variants called to reduce the number of false negative candidates). The obtained VCF files were restricted to the previously mapped target regions and annotated using Ensembl Variant Effect Predictor v67 (VEP) (McLaren et al. 2010). Realigned BAM files were used with Pindel (Ye et al. 2009) to detect breakpoints of large deletions and insertions, assuming an average insert size of 220 bp. All samples were used together for calling structural variants by Pindel. Pindel variants were overlapped with Ensembl v66 annotation for prioritization.

Data access

The raw sequencing data for all of the mutant and inbred strains reported here have been submitted to the NCBI Sequence Read Archive (SRA; http://www.ncbi.nlm.nih.gov/sra) under accession number SRP053040. Variant data are available through the Mouse Mutant Resource Database (https://mmrdb.jax.org); all strains are available through The Jackson Laboratory Mouse Mutant Resource (www.jax.org), and mutant allele/phenotype information is available through the Mouse Genome Database (search by allele, www.informatics.jax.org).

Acknowledgments

We thank the Mouse Mutant Resource and the highly trained animal care technicians at The Jackson Laboratory for their hard work and dedication; they are collectively responsible for the original discovery of the mutant strains used in this study. We also thank the staff of the Genome Technologies core services at The Jackson Laboratory who provided high-throughput and Sanger sequencing support, as well as SNP genotyping and DNA archiving. We also thank Drs. Nara Lygia Sobreira and David Valle of the Centers for Mendelian Genomics for providing access to Gene Matcher and for their assistance in comparing our gene list to unpublished human Mendelian disease candidate genes. This project was funded in part by the Office of Extramural Research, National Institutes of Health grants OD011163 and OD010972 to L.G.R., D. B., and L.R.D.; National Institutes of Health, National Eye Institute grant EY015073 to L.R.D.; National Institutes of Health, National Institute of Dental and Craniofacial Research grant DE020052 to S.M. and L.R.D.; and National Institutes of Health, National Eye Institute grant R01EY019943 to B.C. C.H.P was supported by a research grant from the Council for Nail Disorders and the Cicatricial Alopecia Research Foundation. The S.E.A. laboratory is supported by grants from the Swiss National Science Foundation (SNF 144082), the European Research Council (ERC 249968), and the Gebert and Childcare Foundations. P.M. was supported by a grant from the vonMeissner Foundation. J.S. and M.K. were supported by National Institutes of Health, National Human Genome Research Institute grant HG006493.

References

The 1000 Genomes Project Consortium. 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491:** 56–65.

Antonarakis SE, Beckmann JS. 2006. Mendelian disorders deserve more attention. *Nat Rev Genet* **7:** 277–282.
Bailey DW. 1978. *Sources of subline divergence and their relative importance for sublines of six major inbred strains of mice*. Academic Press, Waltham, MA.
Bashirova AA, Apps R, Vince N, Mochalova Y, Yu XG, Carrington M. 2014. Diversity of the human LILRB3/A6 locus encoding a myeloid inhibitory and activating receptor pair. *Immunogenetics* **66:** 1–8.
Beaulieu CL, Majewski J, Schwartzentruber J, Samuels ME, Fernandez BA, Bernier FP, Brudno M, Knoppers B, Marcadier J, Dyment D, et al. 2014. FORGE Canada Consortium: outcomes of a 2-year national rare-disease gene-discovery project. *Am J Hum Genet* **94:** 809–817.
Castle WE, Little CC. 1910. On a modified Mendelian ratio among yellow mice. *Science* **32:** 868–870.
Choi M, Scholl UI, Ji W, Liu T, Tikhonova IR, Zumbo P, Nayir A, Bakkaloglu A, Ozen S, Sanjad S, et al. 2009. Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proc Natl Acad Sci* **106:** 19096–19101.
Cingolani P, Platts A, Wang le L, 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 (Austin)* **6:** 80–92.
Cuenot L. 1905. Les races pures et leurs combinaisons chez les souris. *Arch Zool Exp Gen* **3:** cxxiii–cxxxii.
DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, et al. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* **43:** 491–498.
Dray S, Dufour AB. 2007. The ade4 package: implementing the duality diagram for ecologists. *J Stat Softw* **22:** 1–20.
Fairfield H, Gilbert GJ, Barter M, Corrigan RR, Curtain M, Ding Y, D'Ascenzo M, Gerhardt DJ, He C, Huang W, et al. 2011. Mutation discovery in mice by whole exome sequencing. *Genome Biol* **12:** R86.
Farrar JE, Quarello P, Fisher R, O'Brien KA, Aspesi A, Parrella S, Henson AL, Seidel NE, Atsidaftos E, Prakash S, et al. 2014. Exploiting pre-rRNA processing in Diamond Blackfan anemia gene discovery and diagnosis. *Am J Hematol* **89:** 985–991.
Fu W, O'Connor TD, Jun G, Kang HM, Abecasis G, Leal SM, Gabriel S, Rieder MJ, Altshuler D, Shendure J, et al. 2013. Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature* **493:** 216–220.
Gabriel S, Ziaugra L, Tabbaa D. 2009. SNP genotyping using the Sequenom MassARRAY iPLEX platform. *Curr Protoc Hum Genet* 60: 2.12.1–2.12.16.
Gerlai R. 2001. Gene targeting: technical confounds and potential solutions in behavioral brain research. *Behav Brain Res* **125:** 13–21.
Gilissen C, Hoischen A, Brunner HG, Veltman JA. 2012. Disease gene identification strategies for exome sequencing. *Eur J Hum Genet* **20:** 490–497.
Handsaker RE, Korn JM, Nemesh J, McCarroll SA. 2011. Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. *Nat Genet* **43:** 269–276.
Huang JD, Cope MJ, Mermall V, Strobel MC, Kendrick-Jones J, Russell LB, Mooseker MS, Copeland NG, Jenkins NA. 1998a. Molecular genetic dissection of mouse unconventional myosin-VA: head region mutations. *Genetics* **148:** 1951–1961.
Huang JD, Mermall V, Strobel MC, Russell LB, Mooseker MS, Copeland NG, Jenkins NA. 1998b. Molecular genetic dissection of mouse unconventional myosin-VA: tail region mutations. *Genetics* **148:** 1963–1972.
Kashevarova AA, Nazarenko LP, Skryabin NA, Salyukova OA, Chechetkina NN, Tolmacheva EN, Sazhenova EA, Magini P, Graziano C, Romeo G, et al. 2014. Array CGH analysis of a cohort of Russian patients with intellectual disability. *Gene* **536:** 145–150.
Keane TM, Goodstadt L, Danecek P, White MA, Wong K, Yalcin B, Heger A, Agam A, Slater G, Goodson M, et al. 2011. Mouse genomic variation and its effect on phenotypes and gene regulation. *Nature* **477:** 289–294.
Kirby A, Kang HM, Wade CM, Cotsapas C, Kostem E, Han B, Furlotte N, Kang EY, Rivas M, Bogue MA, et al. 2010. Fine mapping in 94 inbred mouse strains using a high-density haplotype resource. *Genetics* **185:** 1081–1095.
Kircher M, Heyn P, Kelso J. 2011. Addressing challenges in the production and analysis of Illumina sequencing data. *BMC Genomics* **12:** 382.
Kircher M, Sawyer S, Meyer M. 2012. Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. *Nucleic Acids Res* **40:** e3.
Kondrashov N, Pusic A, Stumpf CR, Shimizu K, Hsieh AC, Xue S, Ishijima J, Shiroishi T, Barna M. 2011. Ribosome-mediated specificity in Hox mRNA translation and vertebrate tissue patterning. *Cell* **145:** 383–397.
Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25:** 2078–2079.
Linder CC. 2001. The influence of genetic background on spontaneous and genetically engineered mouse models of complex diseases. *Lab Anim (NY)* **30:** 34–39.
Maksakova IA, Romanish MT, Gagnier L, Dunn CA, van de Lagemaat LN, Mager DL. 2006. Retroviral elements and their hosts: insertional mutagenesis in the mouse germ line. *PLoS Genet* **2:** e2.
McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, et al. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20:** 1297–1303.
McLaren W, Pritchard B, Rios D, Chen Y, Flicek P, Cunningham F. 2010. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics* **26:** 2069–2070.
Meyer M, Kircher M. 2010. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harb Protoc* doi: 10.1101/pdb.prot5448.
Mouse Genome Sequencing Consortium. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420:** 520–562.
Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, Lee C, Shaffer T, Wong M, Bhattacharjee A, Eichler EE, et al. 2009. Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* **461:** 272–276.
Parr BA, Avery EJ, Cygan JA, McMahon AP. 1998. The classical mouse mutant postaxial hemimelia results from a mutation in the Wnt 7a gene. *Dev Biol* **202:** 228–234.
Patel RK, Jain M. 2012. NGS QC Toolkit: a toolkit for quality control of next generation sequencing data. *PLoS One* **7:** e30619.
Perrin S. 2014. Preclinical research: make mouse studies work. *Nature* **507:** 423–425.
Schlager G, Dickie MM. 1967. Spontaneous mutations and mutation rates in the house mouse. *Genetics* **57:** 319–330.
Segurel L, Wyman MJ, Przeworski M. 2014. Determinants of mutation rate variation in the human germline. *Annu Rev Genomics Hum Genet* **15:** 47–70.
Seong E, Wainer BH, Hughes ED, Saunders TL, Burmeister M, Faundez V. 2005. Genetic analysis of the neuronal and ubiquitous AP-3 adaptor complexes reveals divergent functions in brain. *Mol Biol Cell* **16:** 128–140.
Stein SA, Oates EL, Hall CR, Grumbles RM, Fernandez LM, Taylor NA, Puett D, Jin S. 1994. Identification of a point mutation in the thyrotropin receptor of the hyt/hyt hypothyroid mouse. *Mol Endocrinol* **8:** 129–138.
Suzuki R, Shimodaira H. 2006. Pvclust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics* **22:** 1540–1542.
Taft RA, Davisson M, Wiles MV. 2006. Know thy mouse. *Trends Genet* **22:** 649–653.
Wang K, Li M, Hakonarson H. 2010. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* **38:** e164.
Wolfer DP, Crusio WE, Lipp HP. 2002. Knockout mice: simple solutions to the problems of genetic background and flanking genes. *Trends Neurosci* **25:** 336–340.
Ye K, Schulz MH, Long Q, Apweiler R, Ning Z. 2009. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* **25:** 2865–2871.

Received November 10, 2014; accepted in revised form April 20, 2015.