Published in final edited form as:

Mamm Genome. 2012 October ; 23(0): 499–513. doi:10.1007/s00335-012-9424-0.

High throughput sequencing approaches to mutation discovery in the mouse

Michelle M Simon¹, Ann-Marie Mallon¹, Gareth R. Howell², and Laura G. Reinholdt²

¹MRC Harwell, Harwell Science and Innovation Campus, Oxfordshire OX11 0RD, UK

²The Jackson Laboratory, 600 Main Street, Bar Harbor, ME 04660

Abstract

Phenotype driven approaches in mice are powerful strategies for the discovery of genes, gene functions and for unravelling complex biological mechanisms. Traditional methods for mutation discovery are reliable and robust, but they can also be laborious and time consuming. Recently, high throughput sequencing (HTS) technologies have revolutionised the process of forward genetics in mice by paving the way to rapid mutation discovery. However, successful application of HTS for mutation discovery is relies heavily on the sequencing approach employed and strategies for data analysis. Here, we review current HTS applications and resources for mutation discovery and provide an overview of the practical considerations for HTS implementation and data analysis.

INTRODUCTION

The sequencing of human, mouse and other genomes has revolutionised our understanding of mammalian biology, disease and evolution. Approximately 99% of human protein-coding genes, and many non-coding RNAs, have an equivalent gene in the mouse genome making mice an ideal system for modelling human disease (Waterston et al. 2002). Human genomic variation, including single nucleotide variants (SNVs), inversions, insertions and deletions (INDELS), large chromosomal rearrangements and copy number variations cause heritable diseases, contribute significantly to the etiology of complex diseases, and are the basis of personalized medicine (Ng et al. 2012). Mice that carry equivalent or similar variations can recapitulate key features of human diseases. Variations can be introduced into mice using targeted, reverse genetic approaches (to directly model human disease variation) or at random (induced or spontaneous) to uncover novel disease-causing mutations. Studying the resulting strains of mice provide powerful animal models that can be used to answer basic biological questions relating to gene function, disease mechanisms and drug development (Ledford 2012).

Prior to the advent of genetic engineering and reverse genetic approaches in the 1980s, the primary approach for studying gene function in the mouse was forward genetics, whereby gene discovery follows the observation of a clinical phenotype. The earliest mouse mutants arose spontaneously or were induced by radiation-based strategies at mouse stock centers around the world, like the Medical Research Council Harwell (UK), The Jackson Laboratory (USA), Oakridge National Laboratory mouse genetics program (USA), and National Institute of Genetics Mammalian Genetics Laboratory (Japan). Spontaneous mutant mice remain an important tool for the discovery of novel genes and many spontaneous mutants have become important disease models in a variety of areas including obesity (*Lep^{ob}*)

(Zhang et al. 1994), diabetes (*Lepr^{db}*) (Chen et al. 1996), Duchenne muscular dystrophy (*Dmd^{mdx}*) (Sicinski et al. 1989) and glaucoma (DBA/2J) (John et al. 1998).

The generation of spontaneous mutants depends on natural mutation rates (which for mice are estimated to be 38×10^{-9} per base per generation (Lynch 2010)). Therefore, large colonies are typically the source of new spontaneous mutant collections. Mutation rates can be significantly increased using mutagenesis. Through the 1940's, radiation was the primary tool for mutagenesis in mice. Later, in the 1980's and 90's, chemical mutagenesis became more widely used and remains the mutagen of choice for many large-scale mouse mutagenesis programs (Brown and Nolan 1998; Hrabe de Angelis et al. 2000; Nolan et al. 2000; O'Brien and Frankel 2004, Brown, 1996 #57). Collectively, these programs have generated large collections of mutagenised mice and have developed standardised protocols for phenotyping. Chemical mutagens commonly used include ethylmethanesulfate (EMS) to induce mutations in ES cells or, *N-ethyl-N-nitrosourea* (ENU) to induce single point mutations randomly in the spermatogonia of male mice, increasing mutation rates to $\sim 1 \times 10^{-6}$ per base (Concepcion et al. 2004). There are now more than twenty ENU mouse mutagenesis programs worldwide (Gondo 2008; Oliver et al. 2007) and at least three centers which have large (>10,000) DNA and sperm mouse archive (e.g. The Jackson Laboratory Mutant Mouse Resource; RIKEN BioResource Center, MRC Harwell). Induced mutants from these programs continue to provide the critical starting material for high impact discoveries of novel gene function in the mammalian genome (e.g. (Klus et al. 2012; Kurapati et al. 2012; Qian et al. 2011)),

The phenotype driven approach applied to the discovery of spontaneous or induced mutations is unbiased and is therefore a powerful tool for the discovery of novel genes and novel functions. These mutations provide a full range of alleles including null (loss of function), hypomorphic (reduced function), and neomorphic (altered function); and are therefore more similar to mutations found in the human genome. Moreover, these mutations can reveal gene functions that would not have been discovered through the analysis of null alleles alone (e.g. (Qian et al. 2011)). The offspring of mutagenised mice are subjected to a phenotype screen, which is designed to detect either dominant or recessive phenotypes. Phenotyping can be very broad, based on clinical phenotypes that are immediately apparent or very specific, relying on specialised equipment, expertise or strains (e.g. (Blewitt et al. 2005; Kumar et al. 2011; Reinholdt et al. 2004; Shima et al. 2003; Tchekneva et al. 2007)).

When a heritable phenotype is discovered, positional cloning of the underlying mutation follows. Traditionally, this process involves the establishment of linkage between the phenotype and a chromosome, followed by fine mapping. Until recently, fine mapping has required the introduction of a polymorphic strain background, followed by genotyping with genome wide sets of polymorphic microsatellite DNA markers. Accordingly, the size of the interval depends on the number of recombination events screened and the number of known polymorphic markers within the interval. When the region is small enough to contain a manageable number of candidate genes, each gene is amplified by PCR (usually exon by exon) and individual PCR products are sequenced and analyzed for the presence of a putative causative mutation. While this approach to mutation discovery is robust, it can be laborious and time consuming. Moreover, the mapping process is reliant on recombination that is not uniform throughout the genome, and can be complicated by phenotypic changes that can result from introduction of the divergent strain background required for the mapping cross.

Recently, HTS technologies have supplanted fine mapping and sequencing of individual PCR products, drastically changing the process by which mutations are discovered and

rapidly accelerating forward genetic approaches. In this review we discuss the impact, application and challenges of HTS technologies for the discovery of novel mouse mutations.

HTS TECHNOLOGIES

Genome sequencing for identification of mutations has, until recently, relied on Sanger sequencing (Sanger et al. 1977), or ‘First Generation Sequencing’. In 2005 massively parallel DNA sequencing technologies emerged. These technologies are collectively referred to as ‘Second Generation Sequencing’ or ‘High Throughput Sequencing – HTS’ (Margulies et al. 2005; Shendure et al. 2005). These methods produce millions to billions of reads several orders of magnitude faster than traditional methods (Pettersson et al. 2009), providing massively parallel generation of sequencing data at unprecedented speed and depth.

Next generation sequencing technologies have led to a vast increase in the amount of genomic data generated per base coupled with precipitously decreased cost (Metzker 2010). Currently, HTS technologies can generate a mammalian genome at 20X theoretical coverage for less than US\$5000. HTS sequencing of the mouse genome has been accomplished with a number of commercially available technologies (Arnold et al. 2011; Fairfield et al. 2011; Hilton et al. 2011; Keane et al. 2011; Sun et al. 2012), including Roche/454 (Margulies et al. 2005), Illumina/Solexa (Bentley et al. 2008) and AB SoLiD (Shendure et al. 2005) (Table 1). These platforms are based on clonally amplified templates generated from fragments of single DNA molecules and can produce millions of sequence bases (Roche/454 GS FLX) to billions of bases (Illumina HiSeq) in a single sequencing cycle/run (Fedurco et al. 2006; Metzker 2010). A disadvantage of clonal amplification by PCR is the possible introduction of sequence errors or bias. Hence, ‘third generation sequencing’ (Gupta 2008; Schadt et al. 2010), single molecule sequencing technologies are being developed to bypass the amplification step. Some examples of third generation sequencing platforms include Pacific Biosciences (Eid et al. 2009) and Oxford Nanopore (Maglia et al. 2008) (Table 1). While these technologies offer improved error rates and in some cases, very long reads, they are not yet widely used because they are expensive and relatively low throughput. Another disadvantage of both second and third generation sequencing technologies is that both rely on detection of fluorescence at every cycle, which requires storage and processing of large image files. To overcome these limitations, emerging technologies rely on alternative means of signal detection. For example, the Ion Torrent (Rothberg et al. 2011) platform detects release of hydrogen ions through an ion sensor. This allows for unprecedented speed at prices that are an order of magnitude lower than second and third generation sequencing platforms. However, technologies like the Ion Torrent have yet to achieve the throughput required for routine whole genome sequencing of the mouse genome.

HTS APPROACHES FOR MUTATION DISCOVERY

There are a variety of HTS approaches and bioinformatics tools that are potentially useful for mutation discovery in the mouse genome. When determining the best approach, two major considerations need to be taken into account; first, whether to perform targeted, whole exome or whole genome sequencing, and second, which analysis tools are most likely to reveal causative mutations.

Targeted Sequencing

Targeted sequencing is the method used to sequence select regions of the genome either by PCR amplification of target genomic regions or by hybridization of genomic DNA to probes specific for the target genomic regions (Albert et al. 2007; Okou et al. 2007; Tewhey et al. 2009). These methods effectively enrich for the genomic region(s) of interest, producing

very high read depth and are often used to sequence select genomic regions across large sample sets. Targeted sequencing bypasses the necessity for large DNA library preparations, costly bioinformatics infrastructure and extensive downstream analysis. This approach has been used with great success for the identification of mutations in the mouse genome (Boles et al. 2009; D'Ascenzo et al. 2009; Fairfield et al. 2011; Kurapati et al. 2012). It requires both a known genomic interval and the design and acquisition of custom probes or primers. Therefore, it is most applicable to mutants with significant mapping data or situations where sufficient funds are available to cover the cost of large custom probe pools. As the cost of sequencing continues to drop, the cost benefit of purchasing custom probe pools must be considered. However, if the same probe pool can be used across multiple mutant samples, targeted sequencing is a highly efficient approach.

Whole Exome Sequencing

Exome sequencing is a targeting sequencing strategy where the vast majority of known coding and some non-coding regions of the genome are enriched by hybridization to DNA or RNA probes and the resulting enriched genomic DNA is then sequenced by HTS (Hodges et al. 2007). Pre-designed probe pools for mouse exome sequencing are commercially available (see Resources box). The coding gene models and non-coding RNAs used to guide the design of both commercial probe sets are more comprehensive than the mouse CCDS (collaborative consensus coding sequence) gene set (Pruitt et al. 2009), which is restricted to proteins that are common between human and mouse. Instead, non-redundant gene models from the Mouse Gene Catalog (Blake et al. 2011), which incorporates gene annotations from CCDS, National Center for Biotechnology Information (NCBI), Ensembl and The Vertebrate Genome Annotation (VEGA) database were used. While both commercial designs are based on the Mouse Gene Catalog, there are differences in probe distribution, so a careful comparison of each design is advisable before choosing a platform. Whole exome sequencing has been shown to provide high coverage of coding sequence, flanking intron sequence (50-100 bp) and to a limited degree untranslated regions (UTRs). Application of this approach to large mutant collections has resulted in ~40-60% success rate for putative mutation discovery (Andrews et al. 2012; Boles et al. 2009; Fairfield et al. 2011) depending on the type of mutation (spontaneous or ENU), at a fraction (~1/3) of the cost of whole genome (see below), coupled with significant reduction in storage and compute requirements for analysis. One limitation of this approach is that ~5-15% of regions are inefficiently enriched depending on the platform used (Andrews et al. 2012; Fairfield et al. 2011). An additional limitation of this approach is that it is only applicable to the discovery of mutations within the enriched regions. Mutations in poorly or recently annotated coding sequence or in non-coding sequence will not be found. Historical data on non-coding mutations are impacted by discovery bias (because these mutations have been harder to find and to prove) and to date, relatively few known *causative* mutations affect non-coding sequence. However, increasing application of next generation sequencing approaches to mutation discovery have revealed that a significant percentage of ENU induced mutations reside in UTR's, but the impact (or causality) of these mutations remains unclear (Boles et al. 2009). Non-coding sequences that will not be assessed by exome sequencing include much of the UTRs, many non-coding RNA genes and regulatory sequences such as transcription factor binding sites and enhancers (Semina et al. 2000; Zuniga et al. 2004 {Boles, 2009 #245}). However, these non-coding functional elements often reside in regions of the genome that are conserved between species. Therefore, the limitations of exome sequencing could be overcome if an additional set of probe pools, designed to the non-coding conserved elements (e.g. predicted by Phast, (Hubisz et al. 2011)) were included in the selection. To this end, custom designed probe pools are commercially available and can be used to augment exome probe pools (see Resources box).

Whole Genome Sequencing

The clear advantage of whole genome sequencing is that all regions of the genome are sequenced. In theory, any mutation, including copy number variations and chromosomal rearrangements such as inversions and translocations, could be detected by this approach. However, mutation discovery by whole genome sequencing can be limited by budget constraints, access to data analysis tools and data management infrastructure. One important consideration when implementing whole genome sequencing is coverage. The current cost of generating sufficient sequence to cover the mouse genome at 20X theoretical coverage is \$3,000-\$4,000 (USD) for 2×100 bp sequencing using the Illumina HiSeq platform. This number can certainly vary depending on the sequencing facility and the sequencing platform employed. Whether 20X theoretical coverage provides sufficient minimum read depth for mutation detection depends on a number of factors, including analysis approach (reference based mapping or *de novo assembly*), the genotype of the sequenced sample, and the type of mutation (SNV / INDEL or structural variant). For reference based mapping (discussed below), sequence depth of just 5 reads coverage over a given nucleotide is considered sufficient read depth to confidently call homozygous variants and this minimum read depth is usually obtained (for those regions that are accessible to HTS and standard reference based alignment) with 20 to 30X whole genome coverage (Keane et al. 2011). However, a read depth of 15 reads is often preferable to call heterozygous variants, especially in cases where the underlying mutation itself introduces allelic bias in the alignment data (e.g. a small deletion that may interfere mapping of reads from the mutant allele), therefore, higher coverage is desirable. Finally, the data storage and compute requirements for whole genome sequencing are significant and often underestimated, therefore the availability of data management and analysis resources must be assessed.

Transcriptome sequencing

An alternative exome-based approach is transcriptome analysis (commonly referred to as RNA-seq). In this case, cDNA generated from RNA from a tissue of interest is sequenced. The advantage of this approach is that mutations that cause a change in transcript abundance or that create novel splice variants can be captured. When these mutations occur in expressed genes, they can be identified using the same analysis approaches that are applied to exome data sets. However, when these mutations reside in unexpressed regulatory regions or introns, they will not be represented in the transcriptome data and will require subsequent DNA sequencing for discovery. However, since candidate affected genes are revealed by the transcriptome data, analysis is vastly simplified. The applicability of RNA-seq can be limited by the availability and/or appropriate selection of tissue. Also, a mutation in one gene can change the expression levels of many other genes in the affected tissue, confounding the mutation discovery process.

ANALYSIS OF HTS DATA FOR MUTATION DISCOVERY

Identifying variations from next generation sequence data is challenging and a single, versatile method has not yet been defined. Many different approaches are possible, each with their own strengths and weaknesses. The development of HTS analysis tools is a fast evolving field and a comprehensive review would be quickly out of date. Therefore, here we summarize the general steps taken for mutation detection analysis, highlighting some of the well-established tools for mutation discovery by HTS (Table 1).

Sequence analysis to discover mutations requires three basic steps - (i) alignment to a reference genome, (ii) variant prediction and (iii) variant annotation. Currently, tools are optimized for detecting SNVs and smaller indels. Accurate prediction of larger insertions

and deletions, other structural variations and copy number variation is more challenging and is discussed as a separate section.

Alignment

Raw sequence data are usually made available in pre-formatted, text-based FASTQ files that contain the sequence information organized by individual reads along with quality scores. After ensuring the sequence is of high quality and sufficient coverage, the raw sequences (either single end but more commonly paired end) are aligned to a reference genome. The current reference genome for mouse is of C57BL/6J (B6) origin (current build NCBI37, mm9), however a new assembly of the mouse reference genome (mm10) will soon be available from UCSC Genome Bioinformatics (see Resources Box). Due to natural genetic drift between mouse colonies, it is not uncommon to find variations when comparing sequence from a B6 mouse from one colony to the B6 reference genome. When sequencing mutations on non-B6 genetic backgrounds, detection of the causative mutation(s) becomes more challenging because large numbers of strain-specific variations are identified in addition to the phenotype-causing mutations. Strategies that can be used to discern strain-specific variations from putative causative mutations are discussed below.

The challenge with aligning short sequence reads to a reference genome is “accuracy” versus “compute time/power”. It is possible to very accurately align sequences to the reference genome but optimal accuracy is slow and computationally intensive. Therefore, compromises are made to ensure alignments are completed in a reasonable time. A wide variety of commercial or publically available alignment tools are available (Table 2). Historically, aligners such as Maq (Li et al. 2008a) and ELAND (Cox unpublished) were the first to lead the field, but have been superseded by more accurate aligners such as BWA and Bowtie (Li and Durbin 2010).

For mutation detection, optimum alignment parameters are required to ensure sufficient genome coverage, maximize placement of uniquely mappable reads and reduced misplaced reads. Alignment parameters are often set as defaults for the publicly available aligners and should only be altered for specific purposes by experienced users. Importantly, the quality and number of predicted variants will change depending on the quality and parameter settings of the alignment algorithm. This is particularly true for INDELs, where false negative rates can be relatively high (Keane et al. 2011). In some cases, it may be useful to run more than one alignment tool to take advantage of the strengths found in different tools. For instance, one strategy is to run BWA to efficiently align the majority of sequence reads to the reference genome, and subsequently align the resulting unmapped reads using a slower but more accurate aligner such as Novoalign (<http://www.novocraft.com/main/index.php>). Also, as targeted sequencing and exome sequencing have DNA amplification (PCR) steps prior to sequencing, it is necessary to remove PCR duplicates as PCR-induced errors can introduce false positives.

SNVs and INDEL prediction

Once alignments are generated, putative variants are called using a variant caller or assembler (Table 2). Variant callers will initially predict thousands of SNVs and INDELs and false positives are abundant due to sequencing errors (Nielsen et al. 2011), low read depth, and /or alignment errors. Filtering strategies are employed to minimize aberrant variant calls. First, variant quality scores are generated and variants that fall below a quality threshold, generally <20-30 (Phred scale), are discarded. Second, variants that are predicted in regions covered by less than a pre-determined number of reads are also discarded. Acceptable read depth varies between different projects and users. For example, for homozygous variants, read depth as low as 3-6 reads is considered acceptable, for

heterozygous variants it may be between 6-15 reads. Third, strand bias is taken into account. Genuine variants are expected to occur with equal frequency on forward and reverse strands. In contrast, erroneous variants can show a strand bias. Fourth, some variant callers discard SNVs that are adjacent to INDELs. The remaining high quality variants can then be filtered by allele ratio, depending on the presumed genotype of the sample with respect to the mutation. Homozygous variants are predicted to have allele ratios of 0.7 - 0.9 or higher, heterozygous variants can have allele ratios 0.2 - 0.8, depending on the type of variant and whether it influences read mapping. If a control sample has been sequenced, all common variants between the affected and control sample can be excluded from consideration (unless, of course, the control is heterozygous). Finally, variants can also be filtered based on map position if mapping data are available.

SAMtools (Li et al. 2009) and The Genome Analysis Toolkit (GATK, (DePristo et al. 2011)) are among the more commonly used variant callers. For SAM tools, “Pileup/Filter pileup” were originally developed, but “mPileup” is now the optimum variant caller due to its improvement in predicting INDELs. As the name suggests, GATK is a suite of programs that predicts and annotates variants. Among them, “Unified Genotyper” is capable of predicting SNPs and INDELs in individual or multiple samples. As with sequence alignment, prediction of variants using at least two variant callers is an advisable strategy. Many variants will be common to datasets generated by different callers but each caller is also likely to predict unique variants. Considering only variants predicted by two or more callers is one way to reduce false positives, but it's important to recognize that variants predicted by only one caller need not be false. In addition, standard alignment and SNV/INDEL calling is not effective for larger structural mutations (>20-30 bp) and these classes of variants are considered separately below.

Annotating and prioritising variants

Providing functional context for variants allows the potential impact of the variant to be determined and subsequent filtering and prioritization of variants for further testing. Variants are typically annotated based on their position with respect to genes, coding exons, UTRs, splice sites, conserved sequences, repetitive sequence and their position with respect to known variants. Annotation of known variants can be challenging depending on the strain background involved and the quality of existing variant data. Variant data in dbSNP is one important source of known variant information, with the majority of strain variation data coming from large SNP discovery efforts like the Perlegen mouse re-sequencing / SNP discovery project (Frazer et al. 2007) and more recent large scale whole genome re-sequencing efforts like the Mouse Genomes Project at The Sanger Institute, which has re-sequenced 17 inbred strains and have made available a variation data set of unprecedented scale for the laboratory mouse (Keane et al. 2011). The 17 inbred strains selected for this project include the progenitors of major inbred laboratory strains and represent 4 major taxa; *M.m. castaneus*, *M.m. musculus*, *M.m. domesticus* and *Mus spretus*. The variation data are available from dbSNP and from The Sanger Institute (<http://www.sanger.ac.uk/resources/mouse/genomes/>). One limitation of these resources is that they are heavily biased towards a particular set of inbred strains and therefore will not capture all strain and stock / colony specific variation. Therefore, as mentioned above, sequencing of an unaffected littermate control, while not necessary in all cases, is the ideal way to distinguish between potential mutations and background specific variation.

In addition to the genome build used for reference-based mapping, it is important to consider which databases inform variant annotation. Commonly used, high quality gene models for exon annotation include RefSeq, Ensembl, and Vega. Comprehensive annotation from multiple sources is ideal since not all gene models are in agreement, especially for exons that have little supporting evidence or for UTRs, which are notoriously difficult to annotate.

To that end the Mouse Genome Database maintains a curated gene catalog containing non-redundant gene predictions from RefSeq, Ensembl and Vega; this gene catalog is quickly emerging as the gold standard for mouse gene annotation (Blake et al. 2011). Additional annotation information includes known and common strain variants (e.g. dbSNP and Sanger Mouse Genomes Project), conserved sequences (e.g. predicted by Phast (Hubisz et al. 2011)), and repetitive sequences (e.g. RepeatMasker (Jurka et al. 2005) and Tandem Repeats Finder (Benson 1999)). For manual annotation, all of the data required for variant annotation are freely available from public genome browsers like UCSC Genome Bioinformatics and Ensembl in file formats that can be directly used variant annotation.

A variety of tools are available for variant annotation, consequence prediction and data visualisation (Table 2), including GATK, Annovar (Wang et al. 2010), SNPeff (Minevich et al. unpublished), Panther cSNP (Thomas et al. 2006) and NGS-SNP (Grant et al. 2011). For many tools that predict variant consequences, a probability score is generated reflecting the likelihood the mutation is deleterious. However such scores are inconsistent amongst the variation algorithms, suggesting an integrated approach may be required to determine a more reliable prediction. Predicting the potential consequence of the variant allows variants to be prioritized for further study based on the predicted impact of the variant. For instance, variants that greatly affect the protein-coding region (introduction of a stop codon, changing a splice site) may be more likely to be the causative variant than one that is not predicted to change the amino acid (synonymous change). For human mutation detection, additional strategies for prioritising variants are available. For instance, programs such as Polyphen predict the likelihood a non-synonymous variant is damaging. Also, assessing conservation (e.g. Genome Evolutionary Rate Profiling, GERP) of the altered nucleotide(s) in mammals/vertebrates can be used as predictor of consequence. Although these strategies are well established for human mutation detection, mouse-specific methods are emerging (Andrews et al. 2012).

In summary, comprehensive variant calling and annotation provides the means to generate a “short list” of the most likely candidate mutations. This heuristic approach involves the application of a series of filters based on the assumption that the causative mutation is most likely to be a unique variant, located in a known functional region of the genome and that its allele ratio reflects the expected genotype of the sample. As a note of caution, due to the flexible nature of the parameters incorporating some of the filtering steps it is possible to inadvertently remove causative mutations that masquerade as sequencing errors. For example, a heterozygous mutation may be discarded if present at low coverage and/or at low base quality. Similarly a mutation may be missed if present within a poorly annotated region or in a region of the genome that is inaccessible; regions like this account for 13% to 23% of the mouse genome (Keane et al. 2011). Such discrepancies will be minimised by improvements in the reference sequence, gene annotation, and variant detection methods. Ultimately, any variant data set requires quality control, for example, by assessing false negative rates using known variants. Furthermore putative mutations require manual examination, including visualization of the surrounding alignment using visualization tools (e.g. IGV, Table 1) and wet bench validation using independent methods in additional mutant and control samples (if available).

Detection of structural mutations

Unlike SNVs and short INDELs, structural mutations are mutations that involve more than 50 nucleotides (Alkan et al. 2011). Structural mutations, including duplications, deletions, insertions, inversions and translocations frequently underlie mutant phenotypes, especially phenotypes that arise spontaneously or are induced by radiation. However, spontaneously occurring mutations can arise at any time and in any colony, including ENU colonies (Bannister et al. 2007; Yang et al. 2003). Therefore, it's important to consider the likelihood

of a structural mutation and to apply the appropriate methodological approach. Array based platforms, including comparative genome hybridization and SNP microarrays are popular tools for discovery of structural variation. However, these platforms have limited resolution (~2kb or higher depending on the density of the array) and are not applicable to all types of structural variants. HTS is quickly replacing microarray based approaches for structural mutation discovery, however, a unified approach for library construction, sequencing and analysis that will be applicable to all types of structural variants has not yet been identified. Therefore, to fully harness the power of HTS for the discovery of structural mutations, multiple approaches should be considered.

Alternative analysis approaches for discovery of structural variations are accomplished by analysis of read depth or by analysis of read mapping patterns (also known as read pair and split-read approaches). In the former, regions with unusually high or unusually low (or no) coverage are identified using algorithms that count mapped reads and assess read depth with respect to random (or Poisson) distribution (Bailey et al. 2002). Comparison to control data sets or unrelated data sets of identical size and read length increases the power of this approach. These approaches work well for unbalanced structural variations like duplications and deletions and other CNVs. Structural variations can also cause unexpected or discordant read mapping patterns. In this approach, paired end sequencing data are essential. Paired end data can be generated by several methods depending on the desired fragment size. These methods include paired end mapping (PEM, up to 700 bp), which has been successfully used to discover structural mutations and variations in the mouse genome (Flaherty et al. 2011) (Yalcin et al. 2012), mate pair approaches (up to 5 kb) (Korbel et al. 2007) (Fullwood et al. 2009), which are ideal for the discovery of inversions and translocations, particularly in genomes that are rich in repetitive sequence and large insert BAC or fosmid libraries, which are usually confined to experiments requiring *de novo* assembly.

Individual reads that span the junctions of structural rearrangements also provide critical information for structural mutation discovery. These reads are collectively referred to as 'split reads'. Successful mapping of split reads depends on the mapping algorithm used, the penalties incurred for sub-optimal alignment and the position of the read with respect to the junction of the structural rearrangement. Therefore, reads that directly map to structural rearrangement breakpoints tend to be underrepresented among mapped reads. To enrich for these rare reads, it's often advantageous to obtain higher coverage, by targeted sequencing (Reinholdt et al. 2011; Sobreira et al. 2011) or by oversampling the whole genome (>30X). Table 2 provides some examples of popular analysis tools for structural variant discovery by HTS. For more extensive explanation of structural variant discovery approaches by HTS and additional approaches used for discovery of complex structural variation there are several excellent reviews, including (Alkan et al. 2011) and (Quinlan and Hall 2012).

Now that HTS platforms are evolving towards longer reads, *de novo* assembly is increasingly possible. According to Gnerre et al., genome wide *de novo* assembly is possible given read lengths 100 bp or longer and a combination of paired end sequences (small fragment sizes, less than 400 bp) and mate pair libraries with ~3 kb inserts (Earl et al. 2011; Gnerre et al. 2011). However, for the purpose of discovering structural variants in defined regions (e.g. if a mutation has been mapped to a particular region in or near a gene), local *de novo* assemblies using paired end sequencing data are frequently possible without mate pair libraries. The resulting contigs can then be compared to the reference genome. The advantage of local *de novo* assembly is that reads containing novel sequence (insertions) or reads that span unique structural rearrangement junctions, that would normally fail to align to reference, are captured in a *de novo* assembly by virtue of overlapping segments of homology with neighbouring sequences.

DATA MANAGEMENT AND COMPUTATIONAL RESOURCES

Although sequencing costs remain challenging many laboratories, the so-called \$1000 genome is fast approaching (Schadt et al. 2010). The bottleneck of using HTS for mutation discovery is in data management and processing, which requires significant storage and compute infrastructure. On average a sequencing run can create terabytes of raw sequencing data (which are generally too large to store) which are usually immediately processed into the more manageable text based FASTQ files, the deliverable provided by most sequencing cores. As these files are large, storage of the resultant alignment files after sequencing reads are mapped to the reference genome is often preferable. Most alignment tools generate a file in the universal SAM format (Li et al. 2009) that can then be converted to a compressed binary file, BAM. In fact, this compressed alignment file is emerging as the preferred file format for data storage. The speed with which a BAM file can be generated depends heavily upon the infrastructure available for processing the data. Alignment and assembly can be improved by parallelisation hence many of the modern software packages can now run on multiple processors. Due to the high demand for processing power, future algorithms are implementing new technologies such as highly parallel graphics processing units (GPU) (Klus et al. 2012).

Many institutions offer bioinformatic support and IT infrastructure for HTS data management and analysis. However, for biologists who do not have access to professional bioinformatic support and who are not familiar with working in the command line interface, there are a growing number of freely and commercially available tools that are designed for management and analysis of HTS data through graphical user interfaces (Table 3). Galaxy is a popular open, web-based platform that provides access to many of the alignment, assembly and annotation tools covered in this review (Goecks et al. 2010). The Galaxy project provides user support and tutorials through Wiki pages, screencasts and mailing lists. Moreover, Galaxy can be installed locally. However, local implementation of Galaxy and similar software packages does require local IT support and infrastructure, which may not be available at all institutions. Alternatively, cloud infrastructure based storage and compute resources can be purchased for implementation of HTS analysis tools. Cloud infrastructure is built upon remote servers that are accessible for a fee through the web, providing a means to manage and analyze HTS data that is independent of institutional IT infrastructure. In addition to free, open source HTS data analysis packages, there are also popular commercial options that rely on either cloud or local storage and compute infrastructure (Table 3).

Conclusions

Which is the best HTS approach for mutation discovery? The choice of approach depends primarily on the type of mutation(s) to be discovered and the amount of mapping data available, but also on other factors including; budget, the materials available for sequencing, the tools and expertise to analysis the data and the strain background of the sample used. For ENU-induced mutations, single nucleotide substitutions (mainly transitions or transversions at A.T base pairs (Noveroske et al. 2000) in coding sequence or splice acceptor donor sites are the most common lesions, followed by mutations in highly conserved sequences (Boles et al. 2009). Therefore, exome sequencing is likely to identify phenotype-causing ENU-induced mutations in many cases. However, a complicating factor is that the number of ENU-induced mutations identified will depend on the origin of the sample sequenced. ENU induces approximately 1 mutation per Mb (Justice et al. 2000; Keays et al. 2006; Quwailid et al. 2004) and mice from early generations (e.g. G1) will contain more mutations than later generations (e.g. G2 or G3). The higher the number of mutations identified, the greater the burden to identify the phenotype-causing mutation. For spontaneous mutations, larger structural mutations occurring in non-exonic regions (such as introns) are as likely to be

causal as SNVs or INDELS. Therefore, sequencing exons, regulatory, and non-coding sequence needs to be considered. This can either be achieved by whole genome sequencing or targeted sequencing. If multiple mutations are likely to be present (as in an ENU-treated genome) the validation burden post-sequencing can be high. In this regard, establishing a chromosomal linkage by genetic mapping can be tremendously valuable and is currently easily accomplished using small panels of animals and closely related strain backgrounds (Xia et al. 2010). If mapping is not possible, then sequencing a sample from an unaffected but closely related mouse (preferably an unaffected littermate) can reduce validation burden. Caution using this approach needs to be taken if the phenotype of interest is not fully penetrant.

The power of HTS is in closing the gap between mutant phenotype and putative mutation(s). However, final proof of causation for any putative mutation requires significant follow up biology often including complementation testing or gene targeting. Implementation of HTS to mutation discovery in the mouse genome has already increased the rate of mutation discovery up to ten fold ((Fairfield et al. 2011), and unpublished data from the Mutant Mouse Resource, The Jackson Laboratory and MRC Harwell). As analysis tools, sequencing throughput and accuracy improve, this rate will continue to grow.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

References

- Albert TJ, Molla MN, Muzny DM, Nazareth L, Wheeler D, Song X, Richmond TA, Middle CM, Rodesch MJ, Packard CJ, Weinstock GM, Gibbs RA. Direct selection of human genomic loci by microarray hybridization. *Nat Methods*. 2007; 4:903–905. [PubMed: 17934467]
- Alkan C, Coe BP, Eichler EE. Genome structural variation discovery and genotyping. *Nat Rev Genet*. 2011; 12:363–376. [PubMed: 21358748]
- Andrews TD, Whittle B, Field MA, Balakishnan B, Zhang Y, Shao Y, Cho V, Kirk M, Singh M, Xia Y, Hager J, Winslade S, Sjollem G, Beutler B, Enders A, Goodnow CC. Massively parallel sequencing of the mouse exome to accurately identify rare, induced mutations: an immediate source for thousands of new mouse models. *Open Biology*. 2012
- Arnold CN, Xia Y, Lin P, Ross C, Schwander M, Smart NG, Muller U, Beutler B. Rapid identification of a disease allele in mouse through whole genome sequencing and bulk segregation analysis. *Genetics*. 2011; 187:633–641. [PubMed: 21196518]
- Bailey JA, Gu Z, Clark RA, Reinert K, Samonte RV, Schwartz S, Adams MD, Myers EW, Li PW, Eichler EE. Recent segmental duplications in the human genome. *Science*. 2002; 297:1003–1007. [PubMed: 12169732]
- Bannister LA, Pezza RJ, Donaldson JR, de Rooij DG, Schimenti KJ, Camerini-Otero RD, Schimenti JC. A dominant, recombination-defective allele of Dmc1 causing male-specific sterility. *PLoS Biol*. 2007; 5:e105. [PubMed: 17425408]
- Bateman A, Quackenbush J. Bioinformatics for next generation sequencing. *Bioinformatics*. 2009; 25:429. [PubMed: 19202193]
- Benson G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res*. 1999; 27:573–580. [PubMed: 9862982]
- Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR, Boutell JM, Bryant J, Carter RJ, Keira Cheatham R, Cox AJ, Ellis DJ, Flatbush MR, Gormley NA, Humphray SJ, Irving LJ, Karbelashvili MS, Kirk SM, Li H, Liu X, Maisinger KS, Murray LJ, Obradovic B, Ost T, Parkinson ML, Pratt MR, Rasolonjatovo IM, Reed MT, Rigatti R, Rodighiero C, Ross MT, Sabot A, Sankar SV, Scally A, Schroth GP, Smith ME, Smith VP, Spiridou A, Torrance PE, Tzonev SS, Vermaas EH, Walter K, Wu X, Zhang L, Alam MD, Anastasi C, Aniebo IC, Bailey DM, Bancarz IR, Banerjee S, Barbour SG, Baybayan PA,

Benoit VA, Benson KF, Bevis C, Black PJ, Boodhun A, Brennan JS, Bridgham JA, Brown RC, Brown AA, Buermann DH, Bundu AA, Burrows JC, Carter NP, Castillo N, Chiara ECM, Chang S, Neil Cooley R, Crake NR, Dada OO, Diakoumakos KD, Dominguez-Fernandez B, Earnshaw DJ, Egbujor UC, Elmore DW, Etchin SS, Ewan MR, Fedurco M, Fraser LJ, Fuentes Fajardo KV, Scott Furey W, George D, Gietzen KJ, Goddard CP, Golda GS, Granieri PA, Green DE, Gustafson DL, Hansen NF, Harnish K, Haudenschield CD, Heyer NI, Hims MM, Ho JT, Horgan AM, Hoschler K, Hurwitz S, Ivanov DV, Johnson MQ, James T, Huw Jones TA, Kang GD, Kerelska TH, Kersey AD, Khrebtukova I, Kindwall AP, Kingsbury Z, Kokko-Gonzales PI, Kumar A, Laurent MA, Lawley CT, Lee SE, Lee X, Liao AK, Loch JA, Lok M, Luo S, Mammen RM, Martin JW, McCauley PG, McNitt P, Mehta P, Moon KW, Mullens JW, Newington T, Ning Z, Ling Ng B, Novo SM, O'Neill MJ, Osborne MA, Osnowski A, Ostadan O, Paraschos LL, Pickering L, Pike AC, Chris Pinkard D, Pliskin DP, Podhasky J, Quijano VJ, Raczy C, Rae VH, Rawlings SR, Chiva Rodriguez A, Roe PM, Rogers J, Rogert Bacigalupo MC, Romanov N, Romieu A, Roth RK, Rourke NJ, Ruediger ST, Rusman E, Sanches-Kuiper RM, Schenker MR, Seoane JM, Shaw RJ, Shiver MK, Short SW, Sizto NL, Sluis JP, Smith MA, Ernest Sohna Sohna J, Spence EJ, Stevens K, Sutton N, Szajkowski L, Tregidgo CL, Turcatti G, Vandevondele S, Verhovsky Y, Virk SM, Wakelin S, Walcott GC, Wang J, Worsley GJ, Yan J, Yau L, Zuerlein M, Mullikin JC, Hurles ME, McCooke NJ, West JS, Oaks FL, Lundberg PL, Klenerman D, Durbin R, Smith AJ. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*. 2008; 456:53–59. [PubMed: 18987734]

- Blake JA, Bult CJ, Kadin JA, Richardson JE, Eppig JT. The Mouse Genome Database (MGD): premier model organism resource for mammalian genomics and genetics. *Nucleic Acids Res*. 2011; 39:D842–848. [PubMed: 21051359]
- Blewitt ME, Vickaryous NK, Hemley SJ, Ashe A, Bruxner TJ, Preis JJ, Arkell R, Whitelaw E. An N-ethyl-N-nitrosourea screen for genes involved in variegation in the mouse. *Proceedings of the National Academy of Sciences of the United States of America*. 2005; 102:7629–7634. [PubMed: 15890782]
- Boles MK, Wilkinson BM, Wilming LG, Liu B, Probst FJ, Harrow J, Grafham D, Hentges KE, Woodward LP, Maxwell A, Mitchell K, Risley MD, Johnson R, Hirschi K, Lupski JR, Funato Y, Miki H, Marin-Garcia P, Matthews L, Coffey AJ, Parker A, Hubbard TJ, Rogers J, Bradley A, Adams DJ, Justice MJ. Discovery of candidate disease genes in ENU-induced mouse mutants by large-scale sequencing, including a splice-site mutation in nucleoredoxin. *PLoS Genet*. 2009; 5:e1000759. [PubMed: 20011118]
- Brown SD, Nolan PM. Mouse mutagenesis-systematic studies of mammalian gene function. *Hum Mol Genet*. 1998; 7:1627–1633. [PubMed: 9735384]
- Chaisson MJ, Brinza D, Pevzner PA. De novo fragment assembly with short mate-paired reads: Does the read length matter? *Genome Res*. 2009; 19:336–346. [PubMed: 19056694]
- Chen H, Charlat O, Tartaglia LA, Woolf EA, Weng X, Ellis SJ, Lakey ND, Culpepper J, Moore KJ, Breitbart RE, Duyk GM, Tepper RI, Morgenstern JP. Evidence that the diabetes gene encodes the leptin receptor: identification of a mutation in the leptin receptor gene in db/db mice. *Cell*. 1996; 84:491–495. [PubMed: 8608603]
- Chen K, Wallis JW, McLellan MD, Larson DE, Kalicki JM, Pohl CS, McGrath SD, Wendl MC, Zhang Q, Locke DP, Shi X, Fulton RS, Ley TJ, Wilson RK, Ding L, Mardis ER. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat Methods*. 2009; 6:677–681. [PubMed: 19668202]
- Concepcion D, Seburn KL, Wen G, Frankel WN, Hamilton BA. Mutation rate and predicted phenotypic target sizes in ethylnitrosourea-treated mice. *Genetics*. 2004; 168:953–959. [PubMed: 15514066]
- Cox A. ELAND: Efficient Local Alignment of Nucleotide Data. unpublished.
- D'Ascenzo M, Meacham C, Kitzman J, Middle C, Knight J, Winer R, Kukricar M, Richmond T, Albert TJ, Czechanski A, Donahue LR, Affourtit J, Jeddelloh JA, Reinholdt L. Mutation discovery in the mouse using genetically guided array capture and resequencing. *Mamm Genome*. 2009; 20:424–436. [PubMed: 19629596]
- DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell TJ, Kernysky AM, Sivachenko AY, Cibulskis K,

- Gabriel SB, Altshuler D, Daly MJ. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature genetics*. 2011; 43:491–498. [PubMed: 21478889]
- Earl D, Bradnam K, St John J, Darling A, Lin D, Fass J, Yu HO, Buffalo V, Zerbino DR, Diekhans M, Nguyen N, Ariyaratne PN, Sung WK, Ning Z, Haimel M, Simpson JT, Fonseca NA, Birol I, Docking TR, Ho IY, Rokhsar DS, Chikhi R, Lavenier D, Chapuis G, Naquin D, Maillet N, Schatz MC, Kelley DR, Phillippy AM, Koren S, Yang SP, Wu W, Chou WC, Srivastava A, Shaw TI, Ruby JG, Skewes-Cox P, Betegon M, Dimon MT, Solovyev V, Seledtsov I, Kosarev P, Vorobyev D, Ramirez-Gonzalez R, Leggett R, MacLean D, Xia F, Luo R, Li Z, Xie Y, Liu B, Gnerre S, MacCallum I, Przybylski D, Ribeiro FJ, Yin S, Sharpe T, Hall G, Kersey PJ, Durbin R, Jackman SD, Chapman JA, Huang X, DeRisi JL, Caccamo M, Li Y, Jaffe DB, Green RE, Haussler D, Korf I, Paten B. Assemblathon 1: a competitive assessment of de novo short read assembly methods. *Genome Res*. 2011; 21:2224–2241. [PubMed: 21926179]
- Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P, Bettman B, Bibillo A, Bjornson K, Chaudhuri B, Christians F, Cicero R, Clark S, Dalal R, Dewinter A, Dixon J, Foquet M, Gaertner A, Hardenbol P, Heiner C, Hester K, Holden D, Kearns G, Kong X, Kuse R, Lacroix Y, Lin S, Lundquist P, Ma C, Marks P, Maxham M, Murphy D, Park I, Pham T, Phillips M, Roy J, Sebra R, Shen G, Sorenson J, Tomaney A, Travers K, Trulson M, Vieceli J, Wegener J, Wu D, Yang A, Zaccarin D, Zhao P, Zhong F, Korlach J, Turner S. Real-time DNA sequencing from single polymerase molecules. *Science*. 2009; 323:133–138. [PubMed: 19023044]
- Emde AK, Schulz MH, Weese D, Sun R, Vingron M, Kalscheuer VM, Haas SA, Reinert K. Detecting genomic indel variants with exact breakpoints in single- and paired-end sequencing data using SplazerS. *Bioinformatics*. 2012; 28:619–627. [PubMed: 22238266]
- Fairfield H, Gilbert GJ, Barter M, Corrigan RR, Curtain M, Ding Y, D'Ascenzo M, Gerhardt DJ, He C, Huang W, Richmond T, Rowe L, Probst FJ, Bergstrom DE, Murray SA, Bult C, Richardson J, Kile BT, Gut I, Hager J, Sigurdsson S, Mauceli E, Di Palma F, Lindblad-Toh K, Cunningham ML, Cox TC, Justice MJ, Spector MS, Lowe SW, Albert T, Donahue LR, Jeddelloh J, Shendure J, Reinholdt LG. Mutation discovery in mice by whole exome sequencing. *Genome Biol*. 2011; 12:R86. [PubMed: 21917142]
- Fedurco M, Romieu A, Williams S, Lawrence I, Turcatti G. BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies. *Nucleic Acids Res*. 2006; 34:e22. [PubMed: 16473845]
- Flaherty JP, Fairfield HE, Spruce CA, McCarty CM, Bergstrom DE. Molecular characterization of an allelic series of mutations in the mouse *Nox3* gene. *Mamm Genome*. 2011; 22:156–169. [PubMed: 21161235]
- Frazer KA, Eskin E, Kang HM, Bogue MA, Hinds DA, Beilharz EJ, Gupta RV, Montgomery J, Morenzoni MM, Nilsen GB, Pethiyagoda CL, Stuve LL, Johnson FM, Daly MJ, Wade CM, Cox DR. A sequence-based variation map of 8.27 million SNPs in inbred mouse strains. *Nature*. 2007; 448:1050–1053. [PubMed: 17660834]
- Fullwood MJ, Wei CL, Liu ET, Ruan Y. Next-generation DNA sequencing of paired-end tags (PET) for transcriptome and genome analyses. *Genome Res*. 2009; 19:521–532. [PubMed: 19339662]
- Gnerre S, Maccallum I, Przybylski D, Ribeiro FJ, Burton JN, Walker BJ, Sharpe T, Hall G, Shea TP, Sykes S, Berlin AM, Aird D, Costello M, Daza R, Williams L, Nicol R, Gnirke A, Nusbaum C, Lander ES, Jaffe DB. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proceedings of the National Academy of Sciences of the United States of America*. 2011; 108:1513–1518. [PubMed: 21187386]
- Goecks J, Nekrutenko A, Taylor J. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol*. 2010; 11:R86. [PubMed: 20738864]
- Gondo Y. Trends in large-scale mouse mutagenesis: from genetics to functional genomics. *Nat Rev Genet*. 2008; 9:803–810. [PubMed: 18781157]
- Grant JR, Arantes AS, Liao X, Stothard P. In-depth annotation of SNPs arising from resequencing projects using NGS-SNP. *Bioinformatics*. 2011; 27:2300–2301. [PubMed: 21697123]
- Gupta PK. Single-molecule DNA sequencing technologies for future genomics research. *Trends Biotechnol*. 2008; 26:602–611. [PubMed: 18722683]

- Hilton JM, Lewis MA, Grati M, Ingham N, Pearson S, Laskowski RA, Adams DJ, Steel KP. Exome sequencing identifies a missense mutation in *Isl1* associated with low penetrance otitis media in dearisch mice. *Genome Biol.* 2011; 12:R90. [PubMed: 21936904]
- Hodges E, Xuan Z, Balija V, Kramer M, Molla MN, Smith SW, Middle CM, Rodesch MJ, Albert TJ, Hannon GJ, McCombie WR. Genome-wide in situ exon capture for selective resequencing. *Nature genetics.* 2007; 39:1522–1527. [PubMed: 17982454]
- Hrabe de Angelis MH, Flaswinkel H, Fuchs H, Rathkolb B, Soewarto D, Marschall S, Heffner S, Pargent W, Wuensch K, Jung M, Reis A, Richter T, Alessandrini F, Jakob T, Fuchs E, Kolb H, Kremmer E, Schaeble K, Rollinski B, Roscher A, Peters C, Meitinger T, Strom T, Steckler T, Holsboer F, Klopstock T, Gekeler F, Schindewolf C, Jung T, Avraham K, Behrendt H, Ring J, Zimmer A, Schughart K, Pfeffer K, Wolf E, Balling R. Genome-wide, large-scale production of mutant mice by ENU mutagenesis. *Nature genetics.* 2000; 25:444–447. [PubMed: 10932192]
- Hubisz MJ, Pollard KS, Siepel A. PHAST and RPHAST: phylogenetic analysis with space/time models. *Brief Bioinform.* 2011; 12:41–51. [PubMed: 21278375]
- John SW, Smith RS, Savinova OV, Hawes NL, Chang B, Turnbull D, Davisson M, Roderick TH, Heckenlively JR. Essential iris atrophy, pigment dispersion, and glaucoma in DBA/2J mice. *Invest Ophthalmol Vis Sci.* 1998; 39:951–962. [PubMed: 9579474]
- Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res.* 2005; 110:462–467. [PubMed: 16093699]
- Justice MJ, Carpenter DA, Favor J, Neuhauser-Klaus A, Hrabe de Angelis M, Soewarto D, Moser A, Cordes S, Miller D, Chapman V, Weber JS, Rinchik EM, Hunsicker PR, Russell WL, Bode VC. Effects of ENU dosage on mouse strains. *Mamm Genome.* 2000; 11:484–488. [PubMed: 10886010]
- Keane TM, Goodstadt L, Danecek P, White MA, Wong K, Yalcin B, Heger A, Agam A, Slater G, Goodson M, Furlotte NA, Eskin E, Nellaker C, Whitley H, Cleak J, Janowitz D, Hernandez-Pliego P, Edwards A, Belgard TG, Oliver PL, McIntyre RE, Bhomra A, Nicod J, Gan X, Yuan W, van der Weyden L, Steward CA, Bala S, Stalker J, Mott R, Durbin R, Jackson IJ, Czechanski A, Guerra-Assuncao JA, Donahue LR, Reinholdt LG, Payseur BA, Ponting CP, Birney E, Flint J, Adams DJ. Mouse genomic variation and its effect on phenotypes and gene regulation. *Nature.* 2011; 477:289–294. [PubMed: 21921910]
- Keays DA, Clark TG, Flint J. Estimating the number of coding mutations in genotypic- and phenotypic-driven N-ethyl-N-nitrosourea (ENU) screens. *Mamm Genome.* 2006; 17:230–238. [PubMed: 16518690]
- Klus P, Lam S, Lyberg D, Cheung MS, Pullan G, McFarlane I, Yeo G, Lam BY. BarraCUDA - a fast short read sequence aligner using graphics processing units. *BMC Res Notes.* 2012; 5:27. [PubMed: 22244497]
- Koboldt DC, Chen K, Wylie T, Larson DE, McLellan MD, Mardis ER, Weinstock GM, Wilson RK, Ding L. VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics.* 2009; 25:2283–2285. [PubMed: 19542151]
- Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, Miller CA, Mardis ER, Ding L, Wilson RK. VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* 2012; 22:568–576. [PubMed: 22300766]
- Korbel JO, Urban AE, Affourtit JP, Godwin B, Grubert F, Simons JF, Kim PM, Palejev D, Carriero NJ, Du L, Taillon BE, Chen Z, Tanzer A, Saunders AC, Chi J, Yang F, Carter NP, Hurles ME, Weissman SM, Harkins TT, Gerstein MB, Egholm M, Snyder M. Paired-end mapping reveals extensive structural variation in the human genome. *Science.* 2007; 318:420–426. [PubMed: 17901297]
- Kumar V, Kim K, Joseph C, Thomas LC, Hong H, Takahashi JS. Second-generation high-throughput forward genetic screen in mice to isolate subtle behavioral mutants. *Proceedings of the National Academy of Sciences of the United States of America.* 2011; 108(Suppl 3):15557–15564. [PubMed: 21896739]
- Kurapati R, McKenna C, Lindqvist J, Williams D, Simon M, Leproust E, Baker J, Cheeseman M, Carroll N, Denny P, Laval S, Lochmuller H, Ochala J, Blanco G. Myofibrillar myopathy caused

- by a mutation in the motor domain of mouse MyHC IIb. *Hum Mol Genet.* 2012; 21:1706–1724. [PubMed: 22199023]
- Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods.* 2012; 9:357–359. [PubMed: 22388286]
- Ledford H. Mice guide human drug trial. *Nature.* 2012; 483:389. [PubMed: 22437589]
- Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics.* 2009; 25:1754–1760. [PubMed: 19451168]
- Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics.* 2010; 26:589–595. [PubMed: 20080505]
- Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. The Sequence Alignment/Map format and SAMtools. *Bioinformatics.* 2009; 25:2078–2079. [PubMed: 19505943]
- Li H, Ruan J, Durbin R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.* 2008a; 18:1851–1858. [PubMed: 18714091]
- Li R, Li Y, Kristiansen K, Wang J. SOAP: short oligonucleotide alignment program. *Bioinformatics.* 2008b; 24:713–714. [PubMed: 18227114]
- Lin H, Zhang Z, Zhang MQ, Ma B, Li M. ZOOM! Zillions of oligos mapped. *Bioinformatics.* 2008; 24:2431–2437. [PubMed: 18684737]
- Lynch M. Evolution of the mutation rate. *Trends Genet.* 2010; 26:345–352. [PubMed: 20594608]
- Maccallum I, Przybylski D, Gnerre S, Burton J, Shlyakhter I, Gnirke A, Malek J, McKernan K, Ranade S, Shea TP, Williams L, Young S, Nusbaum C, Jaffe DB. ALLPATHS 2: small genomes assembled accurately and with high continuity from short paired reads. *Genome Biol.* 2009; 10:R103. [PubMed: 19796385]
- Maglia G, Restrepo MR, Mikhailova E, Bayley H. Enhanced translocation of single DNA molecules through alpha-hemolysin nanopores by manipulation of internal charge. *Proceedings of the National Academy of Sciences of the United States of America.* 2008; 105:19720–19725. [PubMed: 19060213]
- Manske HM, Kwiatkowski DP. LookSeq: a browser-based viewer for deep sequencing data. *Genome Res.* 2009; 19:2125–2132. [PubMed: 19679872]
- Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH, Irzyk GP, Jando SC, Alenquer ML, Jarvie TP, Jirage KB, Kim JB, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J, Lohman KL, Lu H, Makhijani VB, McDade KE, McKenna MP, Myers EW, Nickerson E, Nobile JR, Plant R, Puc BP, Ronan MT, Roth GT, Sarkis GJ, Simons JF, Simpson JW, Srinivasan M, Tartaro KR, Tomasz A, Vogt KA, Volkmer GA, Wang SH, Wang Y, Weiner MP, Yu P, Begley RF, Rothberg JM. Genome sequencing in microfabricated high-density picolitre reactors. *Nature.* 2005; 437:376–380. [PubMed: 16056220]
- Metzker ML. Sequencing technologies - the next generation. *Nat Rev Genet.* 2010; 11:31–46. [PubMed: 19997069]
- Minevich G, Park D, Poole RJ, Hobart O. A Cloud-based Pipeline for Analysis of Mutant Genome Sequences. unpublished.
- Ng KP, Hillmer AM, Chuah CT, Juan WC, Ko TK, Teo AS, Ariyaratne PN, Takahashi N, Sawada K, Fei Y, Soh S, Lee WH, Huang JW, Allen JC Jr, Woo XY, Nagarajan N, Kumar V, Thalamuthu A, Poh WT, Ang AL, Mya HT, How GF, Yang LY, Koh LP, Chowbay B, Chang CT, Nadarajan VS, Chng WJ, Than H, Lim LC, Goh YT, Zhang S, Poh D, Tan P, Seet JE, Ang MK, Chau NM, Ng QS, Tan DS, Soda M, Isobe K, Nothen MM, Wong TY, Shahab A, Ruan X, Cacheux-Rataboul V, Sung WK, Tan EH, Yatabe Y, Mano H, Soo RA, Chin TM, Lim WT, Ruan Y, Ong ST. A common BIM deletion polymorphism mediates intrinsic resistance and inferior responses to tyrosine kinase inhibitors in cancer. *Nat Med.* 2012; 18:521–528. [PubMed: 22426421]
- Nielsen R, Paul JS, Albrechtsen A, Song YS. Genotype and SNP calling from next-generation sequencing data. *Nat Rev Genet.* 2011; 12:443–451. [PubMed: 21587300]
- Ning Z, Cox AJ, Mullikin JC. SSAHA: a fast search method for large DNA databases. *Genome Res.* 2001; 11:1725–1729. [PubMed: 11591649]

- Nolan PM, Peters J, Vizor L, Strivens M, Washbourne R, Hough T, Wells C, Glenister P, Thornton C, Martin J, Fisher E, Rogers D, Hagan J, Reavill C, Gray I, Wood J, Spurr N, Browne M, Rastan S, Hunter J, Brown SD. Implementation of a large-scale ENU mutagenesis program: towards increasing the mouse mutant resource. *Mamm Genome*. 2000; 11:500–506. [PubMed: 10886012]
- Noveroske JK, Weber JS, Justice MJ. The mutagenic action of N-ethyl-N-nitrosourea in the mouse. *Mamm Genome*. 2000; 11:478–483. [PubMed: 10886009]
- O'Brien TP, Frankel WN. Moving forward with chemical mutagenesis in the mouse. *J Physiol*. 2004; 554:13–21. [PubMed: 14678486]
- Okou DT, Steinberg KM, Middle C, Cutler DJ, Albert TJ, Zwick ME. Microarray-based genomic selection for high-throughput resequencing. *Nat Methods*. 2007; 4:907–909. [PubMed: 17934469]
- Oliver PL, Bitoun E, Davies KE. Comparative genetic analysis: the utility of mouse genetic systems for studying human monogenic disease. *Mammalian Genome*. 2007; 18:412–424. [PubMed: 17514509]
- Pettersson E, Lundeberg J, Ahmadian A. Generations of sequencing technologies. *Genomics*. 2009; 93:105–111. [PubMed: 18992322]
- Pruitt KD, Harrow J, Harte RA, Wallin C, Diekhans M, Maglott DR, Searle S, Farrell CM, Loveland JE, Ruef BJ, Hart E, Suner MM, Landrum MJ, Aken B, Ayling S, Baertsch R, Fernandez-Banet J, Cherry JL, Curwen V, Dicuccio M, Kellis M, Lee J, Lin MF, Schuster M, Shkeda A, Amid C, Brown G, Dukhanina O, Frankish A, Hart J, Makedak BL, Mudge J, Murphy MR, Murphy T, Rajan J, Rajput B, Riddick LD, Snow C, Steward C, Webb D, Weber JA, Wilming L, Wu W, Birney E, Haussler D, Hubbard T, Ostell J, Durbin R, Lipman D. The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes. *Genome Res*. 2009; 19:1316–1323. [PubMed: 19498102]
- Qian L, Mahaffey JP, Alcorn HL, Anderson KV. Tissue-specific roles of Axin2 in the inhibition and activation of Wnt signaling in the mouse embryo. *Proceedings of the National Academy of Sciences of the United States of America*. 2011; 108:8692–8697. [PubMed: 21555575]
- Quinlan AR, Hall IM. Characterizing complex structural variation in germline and somatic genomes. *Trends Genet*. 2012; 28:43–53. [PubMed: 22094265]
- Quwailid MM, Hugill A, Dear N, Vizor L, Wells S, Horner E, Fuller S, Weedon J, McMath H, Woodman P, Edwards D, Campbell D, Rodger S, Carey J, Roberts A, Glenister P, Lalanne Z, Parkinson N, Coghill EL, McKeone R, Cox S, Willan J, Greenfield A, Keays D, Brady S, Spurr N, Gray I, Hunter J, Brown SD, Cox RD. A gene-driven ENU-based approach to generating an allelic series in any gene. *Mamm Genome*. 2004; 15:585–591. [PubMed: 15457338]
- Reinholdt L, Ashley T, Schimenti J, Shima N. Forward genetic screens for meiotic and mitotic recombination-defective mutants in mice. *Methods Mol Biol*. 2004; 262:87–107. [PubMed: 14769957]
- Reinholdt LG, Ding Y, Gilbert GT, Czechanski A, Solzak JP, Roper RJ, Johnson MT, Donahue LR, Lutz C, Davisson MT. Molecular characterization of the translocation breakpoints in the Down syndrome mouse model Ts65Dn. *Mamm Genome*. 2011; 22:685–691. [PubMed: 21953412]
- Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. Integrative genomics viewer. *Nat Biotechnol*. 2011; 29:24–26. [PubMed: 21221095]
- Rothberg JM, Hinz W, Rearick TM, Schultz J, Mileski W, Davey M, Leamon JH, Johnson K, Milgrew MJ, Edwards M, Hoon J, Simons JF, Marran D, Myers JW, Davidson JF, Branting A, Nobile JR, Puc BP, Light D, Clark TA, Huber M, Branciforte JT, Stoner IB, Cawley SE, Lyons M, Fu Y, Homer N, Sedova M, Miao X, Reed B, Sabina J, Feierstein E, Schorn M, Alanjary M, Dimalanta E, Dressman D, Kasinskas R, Sokolsky T, Fidanza JA, Namsaraev E, McKernan KJ, Williams A, Roth GT, Bustillo J. An integrated semiconductor device enabling non-optical genome sequencing. *Nature*. 2011; 475:348–352. [PubMed: 21776081]
- Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*. 1977; 74:5463–5467. [PubMed: 271968]
- Schadt EE, Turner S, Kasarskis A. A window into third-generation sequencing. *Hum Mol Genet*. 2010; 19:R227–240. [PubMed: 20858600]

- Semina EV, Murray JC, Reiter R, Hrstka RF, Graw J. Deletion in the promoter region and altered expression of Pitx3 homeobox gene in aphakia mice. *Hum Mol Genet.* 2000; 9:1575–1585. [PubMed: 10861284]
- Shendure J, Porreca GJ, Reppas NB, Lin X, McCutcheon JP, Rosenbaum AM, Wang MD, Zhang K, Mitra RD, Church GM. Accurate multiplex polony sequencing of an evolved bacterial genome. *Science.* 2005; 309:1728–1732. [PubMed: 16081699]
- Shima N, Hartford SA, Duffy T, Wilson LA, Schimenti KJ, Schimenti JC. Phenotype-based identification of mouse chromosome instability mutants. *Genetics.* 2003; 163:1031–1040. [PubMed: 12663541]
- Sicinski P, Geng Y, Ryder-Cook AS, Barnard EA, Darlison MG, Barnard PJ. The molecular basis of muscular dystrophy in the mdx mouse: a point mutation. *Science.* 1989; 244:1578–1580. [PubMed: 2662404]
- Sobreira NL, Gnanakkan V, Walsh M, Marosy B, Wohler E, Thomas G, Hoover-Fong JE, Hamosh A, Wheelan SJ, Valle D. Characterization of complex chromosomal rearrangements by targeted capture and next-generation sequencing. *Genome Res.* 2011; 21:1720–1727. [PubMed: 21890680]
- Sun M, Mondal K, Patel V, Horner VL, Long AB, Cutler DJ, Caspary T, Zwick ME. Multiplex Chromosomal Exome Sequencing Accelerates Identification of ENU-Induced Mutations in the Mouse. *G3.* 2012; 2:143–150. [PubMed: 22384391]
- Tchekneva EE, Rinchik EM, Polosukhina D, Davis LS, Kadkina V, Mohamed Y, Dunn SR, Sharma K, Qi Z, Fogo AB, Breyer MD. A sensitized screen of N-ethyl-N-nitrosourea-mutagenized mice identifies dominant mutants predisposed to diabetic nephropathy. *J Am Soc Nephrol.* 2007; 18:103–112. [PubMed: 17151334]
- Tewhey R, Warner JB, Nakano M, Libby B, Medkova M, David PH, Kotsopoulos SK, Samuels ML, Hutchison JB, Larson JW, Topol EJ, Weiner MP, Harismendy O, Olson J, Link DR, Frazer KA. Microdroplet-based PCR enrichment for large-scale targeted sequencing. *Nat Biotechnol.* 2009; 27:1025–1031. [PubMed: 19881494]
- Thomas PD, Kejariwal A, Guo N, Mi H, Campbell MJ, Muruganujan A, Lazareva-Ulitsky B. Applications for protein sequence-function evolution data: mRNA/protein expression analysis and coding SNP scoring tools. *Nucleic Acids Res.* 2006; 34:W645–650. [PubMed: 16912992]
- Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* 2010; 38:e164. [PubMed: 20601685]
- Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P, Antonarakis SE, Attwood J, Baertsch R, Bailey J, Barlow K, Beck S, Berry E, Birren B, Bloom T, Bork P, Botcherby M, Bray N, Brent MR, Brown DG, Brown SD, Bult C, Burton J, Butler J, Campbell RD, Carninci P, Cawley S, Chiaromonte F, Chinwalla AT, Church DM, Clamp M, Clee C, Collins FS, Cook LL, Copley RR, Coulson A, Couronne O, Cuff J, Curwen V, Cutts T, Daly M, David R, Davies J, Delehaunty KD, Deri J, Dermitzakis ET, Dewey C, Dickens NJ, Diekhans M, Dodge S, Dubchak I, Dunn DM, Eddy SR, Elnitski L, Emes RD, Eswara P, Eyraas E, Felsenfeld A, Fewell GA, Flicek P, Foley K, Frankel WN, Fulton LA, Fulton RS, Furey TS, Gage D, Gibbs RA, Glusman G, Gnerre S, Goldman N, Goodstadt L, Grafham D, Graves TA, Green ED, Gregory S, Guigo R, Guyer M, Hardison RC, Haussler D, Hayashizaki Y, Hillier LW, Hinrichs A, Hlavina W, Holzer T, Hsu F, Hua A, Hubbard T, Hunt A, Jackson I, Jaffe DB, Johnson LS, Jones M, Jones TA, Joy A, Kamal M, Karlsson EK, Karolchik D, Kasprzyk A, Kawai J, Keibler E, Kells C, Kent WJ, Kirby A, Kolbe DL, Korf I, Kucherlapati RS, Kulbokas EJ, Kulp D, Landers T, Leger JP, Leonard S, Letunic I, Levine R, Li J, Li M, Lloyd C, Lucas S, Ma B, Maglott DR, Mardis ER, Matthews L, Mauceli E, Mayer JH, McCarthy M, McCombie WR, McLaren S, McLay K, McPherson JD, Meldrim J, Meredith B, Mesirov JP, Miller W, Miner TL, Mongin E, Montgomery KT, Morgan M, Mott R, Mullikin JC, Muzny DM, Nash WE, Nelson JO, Nhan MN, Nicol R, Ning Z, Nusbaum C, O'Connor MJ, Okazaki Y, Oliver K, Overton-Larty E, Pachter L, Parra G, Pepin KH, Peterson J, Pevzner P, Plumb R, Pohl CS, Poliakov A, Ponce TC, Ponting CP, Potter S, Quail M, Reymond A, Roe BA, Roskin KM, Rubin EM, Rust AG, Santos R, Sapojnikov V, Schultz B, Schultz J, Schwartz MS, Schwartz S, Scott C, Seaman S, Searle S, Sharpe T, Sheridan A, Shownkeen R, Sims S, Singer JB, Slater G, Smit A, Smith DR, Spencer B, Stabenau A, Stange-Thomann N, Sugnet C, Suyama M, Tesler G, Thompson J, Torrents D, Trevaskis E, Tromp J, Ucla C, Ureta-Vidal A, Vinson JP, Von Niederhausern AC, Wade CM, Wall M, Weber RJ, Weiss RB, Wendl MC, West AP, Wetterstrand

- K, Wheeler R, Whelan S, Wierzbowski J, Willey D, Williams S, Wilson RK, Winter E, Worley KC, Wyman D, Yang S, Yang SP, Zdobnov EM, Zody MC, Lander ES. Initial sequencing and comparative analysis of the mouse genome. *Nature*. 2002; 420:520–562. [PubMed: 12466850]
- Wong K, Keane TM, Stalker J, Adams DJ. Enhanced structural variant and breakpoint detection using SVMerge by integration of multiple detection methods and local assembly. *Genome Biol*. 2010; 11:R128. [PubMed: 21194472]
- Xia Y, Won S, Du X, Lin P, Ross C, La Vine D, Wiltshire S, Leiva G, Vidal SM, Whittle B, Goodnow CC, Koziol J, Moresco EM, Beutler B. Bulk segregation mapping of mutations in closely related strains of mice. *Genetics*. 2010; 186:1139–1146. [PubMed: 20923982]
- Yalcin B, Wong K, Bhomra A, Goodson M, Keane TM, Adams DJ, Flint J. The fine-scale architecture of structural variants in 17 mouse genomes. *Genome Biol*. 2012; 13:R18. [PubMed: 22439878]
- Yang Y, Beyer BJ, Otto JF, O'Brien TP, Letts VA, White HS, Frankel WN. Spontaneous deletion of epilepsy gene orthologs in a mutant mouse with a low electroconvulsive threshold. *Hum Mol Genet*. 2003; 12:975–984. [PubMed: 12700166]
- Ye K, Schulz MH, Long Q, Apweiler R, Ning Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics*. 2009; 25:2865–2871. [PubMed: 19561018]
- Zerbino DR, McEwen GK, Margulies EH, Birney E. Pebble and rock band: heuristic resolution of repeats and scaffolding in the velvet short-read de novo assembler. *PLoS One*. 2009; 4:e8407. [PubMed: 20027311]
- Zhang Y, Proenca R, Maffei M, Barone M, Leopold L, Friedman JM. Positional cloning of the mouse obese gene and its human homologue. *Nature*. 1994; 372:425–432. [PubMed: 7984236]
- Zhang Z, Lin H, Ma B. ZOOM Lite: next-generation sequencing data mapping and visualization software. *Nucleic Acids Res*. 2010; 38:W743–748. [PubMed: 20530531]
- Zuniga A, Michos O, Spitz F, Haramis AP, Panman L, Galli A, Vintersten K, Klasen C, Mansfield W, Kuc S, Duboule D, Dono R, Zeller R. Mouse limb deformity mutations disrupt a global control region within the large regulatory landscape required for Gremlin expression. *Genes Dev*. 2004; 18:1553–1564. [PubMed: 15198975]

Fig 1.

Example of a HTS analysis pipeline to detect sequence variations in mice.

The first step aligns the sequence reads to the mouse reference genome. The resultant BAM file is used as input for mutation detection software. Putative sequence variations are identified following filtration of known inbred variants and other variables. Thereafter the map position, annotation and variation functional effect can be determined. Some steps may be repeated with multiple tools to confirm results and reduce false positives; these steps are denoted with a red star. Some examples of software to use in each section can be found in table 3.

★ Repeat these steps with different software packages or algorithms to reduce false positives or misleading results.

Next Generation Sequencing Platforms. Examples of currently available HTS platforms and comparison of their salient features, including read length and throughput.

Table 1

Technologies	Platforms	Sequencing method	Read Length (bp)	Gb per run	Notes
Roche/454	GS FLX Titanium GS Junior	Pyrosequencing	400 - 600	500	Long reads, fast run times, high error rate in homopolymer tracts
Illumina/Solexa	HiSeq 2000	Reversible terminator	75 or 100	200	Currently the most widely used platform.
Life Technologies/Applied Biosystems	ABI SOLiD	Sequencing by ligation	35 – 100	100	Low error rates, data format is CSFASTQ, requires appropriate analysis tools.
Polonator	Polonator G.007	Sequencing by ligation	26	10	Least expensive, open source users can adapt reagents.
Pacific BioSciences	PacBio RS	Single Molecule real time (SMRT)	>1000	0.45	Large read lengths, fast sample preparation and runs
Oxford Nanopore	GridION	Sequencing through a protein nanopore	Potentially >10,000bp	N/A	Large read lengths but high error rate
Ion Torrent	Ion Proton	Semiconductor sequencing	200	Up to 1	Comparatively inexpensive and fast run times. High error rate in homopolymer tracts

Bioinformatic tools for HTS analysis. An overview of some commonly used tools for alignment, variant calling and variant annotation. For a more comprehensive review of HTS analysis tools see (Bateman and Quackenbush 2009) and the resources box.

Table 2

Alignment - SNVs/INDELs			
BWA	http://bio-bwa.sourceforge.net/	Burrows-Wheeler gapped alignment, short read	(Li and Durbin 2009)
BOWTIE 2	http://bowtie-bio.sourceforge.net/index.shtml	gapped alignment, short read	(Langmead and Salzberg 2012)
NOVOALIGN	http://www.novocraft.com/main/index.php	gapped alignment, Needleman-Wunsch algorithm	NA
EULER	http://euler-assembler.ucsd.edu/portal	<i>de novo</i> assembly, combined reads from different HTS technologies.	(Chaisson et al. 2009)
MOSAIC	http://bioinformatics.bc.edu/marhlab/Mosaik	performs pairwise alignment of reads	NA
MAQ	http://maq.sourceforge.net/	Ungapped alignment and assembly (SNV calling)	(Li et al. 2008a)
ZOOM & ZOOM LITE	http://www.bioinformaticsolutions.com/all-products/zoom	Used for reference guided alignments	(Lin et al. 2008; Zhang et al. 2010)
SSAHA2	http://www.sanger.ac.uk/resources/software/ssaha2/	pairwise alignment, long reads, 454 data	(Ning et al. 2001)
SOAP	http://soap.genomics.org.cn/soap.snp.html	Provides a full solution to HTS analysis, incorporates GPU-based technology for fast analysis	(Li et al. 2008b)

Variant calling – SNVs/INDELs			
SAMTOOLS	http://samtools.sourceforge.net/	A variety of utilities for post-processing alignments, including variant calling	(Li et al. 2009)
VARSCAN 2	http://varscan.sourceforge.net/	SNV and CNV calling	(Koboldt et al. 2009; Koboldt et al. 2012)
GATK	ftp://ftp.broadinstitute.org/pub/gsa/GenomeAnalysisTK/	A variety of utilities for post-processing alignments, including variant calling	(DePristo et al. 2011)
Alignment and Variant calling - Structural Variants			
BREAKDANCER	http://breakdancer.sourceforge.net/	Read pair analysis, short inserts (<1kb)	(Chen et al. 2009)
SPLAZERS	http://www.seqan.de/projects/splazers/	split read analysis	(Ende et al. 2012)
VARSCAN	http://varscan.sourceforge.net/copy-number-calling.html	Variant calling and copy number analysis	(Koboldt et al. 2012)
VELVET	http://www.ebi.ac.uk/~zerbino/velvet/	<i>de novo</i> assembly	(Zerbino et al. 2009)
ALLPATHS	http://www.broadinstitute.org/scientific-community/science/genome-sequencing-and-analysis/computational-rd/computational-denovo	<i>de novo</i> assembly	(Maccallum et al. 2009)
SVMERGE	http://svmerge.sourceforge.net/	comprehensive pipeline for SV discovery, including multiple tools like Breakdancer, Pindel (Ye et al. 2009), RetroSeq (Wong et al. 2010), etc.	(Wong et al. 2010)
Annotation and Visualisation Tools			
ANNOVAR	http://www.openbioinformatics.org/annovar/	Annotates sequence variations for a variety of genomes. Uses	(Wang et al. 2010)

NGS-SNP	http://www.ualberta.ca/~stothard/downloads/NGS-SNP/	RefSeq, UCSC, ENSEMBL and Gencode genes.	(Grant et al. 2011)
SNPEFF	http://snpeff.sourceforge.net/	Annotates sequence variation using information primarily stored in Ensembl	(Minevich et al. unpublished)
PANTHER CSNP	http://www.pantherdb.org/tools/csnpscoreForm.jsp	Sequence variant annotation and effect prediction tool.	(Thomas et al. 2006)
LOOKSEQ	http://www.sanger.ac.uk/resources/software/lookseq/	Evolutionary analysis of coding variations	(Manske and Kwiatkowski 2009)
IGV	http://www.broadinstitute.org/igv/	Tool to visualise read alignments	(Robinson et al. 2011)
		Tool to visualise multiple alignments and small sequence variations simultaneously.	

Table 3

Open source and commercial tool suites for HTS data management and analysis

Tool	Link	Implementation	Reference
GALAXY	http://galaxy.psu.edu/	Web-based interface, local, remote, or cloud	(Goecks et al. 2010)
GENOMEQUEST	http://www.genomequest.com/	Web-based interface, local, or cloud	N/A *
CLC GENOMICS	http://snpeff.sourceforge.net/	Software, local	N/A *
DNAEXUS	https://dhanexus.com/	Web-based, cloud	N/A *

* Commercially available