Mouse Exome Sequencing Advisory

April 2012

1. Purpose

Recent experience at the Australian Phenomics Facility (APF) has shown us that there can be considerable variation in quality and cross-purposes in project design when researchers first start conducting Next Generation Sequencing (NGS) projects. It is a complex world of terminology and constantly changing technology. It is hoped that with the transfer of some background knowledge and a few tips, researchers might more accurately direct the sequencing facility as to their needs. Therefore the purpose of this advisory is to highlight options for mouse researchers in terms of exome capture and sequencing capabilities (section 2), and then discusses data output standards and expectations to help guide your analysis (section 3). Supporting this is a glossary of key terms provided in the appendix.

Advisory Summary – Some of the questions considered:

- What technologies are used for exome capture and sequencing?
- Should exomes be indexed pre- or post-capture?
- How many exomes can be indexed per lane?
- Is it preferable to request Gb per exome over raw coverage?
- How to interpret quality assurance measures?
- What methods are used for sequence alignment and SNV detection?
- To what extent are SNV lists filtered or quality-checked?
- What information might be usefully contained in a sequencing report?
- How can SNVs be validated?

Australian National University APF - Australian Phenomics Facility

agrf Australian Genome Research Facility Ltd

EMBL Australia

2. Exome sequencing of mouse samples for SNV detection

Why Exomes?

The APF have been identifying phenotype-changing genetic variants in mice for the past 10 years and observed that highly penetrant, Mendelian-inherited phenotypes are almost always the result of either coding or splice-site mutations. This encouraged us to design our variant detection processes around sequencing the mouse exome rather than the whole mouse genome for scientific as well as cost reasons. Over the past couple of years the APF has also developed a bioinformatic methodology for accurately isolating rare ENU-induced protein-changing variants in mice with heritable phenotypes.

The APF has accumulated a substantial body of knowhow and expertise around mouse exome sequencing and the bioinformatics associated with detecting Single Nucleotide Variants (SNVs). The expertise has been built exclusively on Illumina sequence data processed at a number of Australian and international sequencing facilities. In particular the bioinformatic team has benefitted from being able to conduct numerous replication trials cross checking the performance of the various platform configurations and SNV filters and thus ensuring robust detection programmes were developed.

Details around how best to utilise exome sequencing including the combinations of technologies used, the differences between the available kits and platforms, the requirements of bioinformatic analysis, and what to expect from the data is provided below.

Exome capture options – Agilent and Nimblegen

There are currently two solution based exome capture technologies available for mouse that supports both SOLiD and Illumina sequencing platforms. The first is the Agilent SureSelect XT Mouse All Exon Kit, whose 50 Mb capture covers the complete mouse exome and spans over 221,784 exons and 24,306 genes. The content was based on the current University of California, Santa Cruz (UCSC) mm9 mouse genome build. The second is the Nimblegen SeqCap EZ Mouse Exon Kit, with a 54Mb capture size designed to cover exons annotated from a range of available databases. The Nimblegen design does not appear to target the CCDS exome as well as the Agilent kit largely due to the removal of all olfactory and vomeronasal genes from the design.

Although the technologies are very similar there are a number of differences that need to be considered. The key difference between them is that Agilent baits are RNA and Nimblegen baits are DNA. The advantage of the Agilent RNA option is that they provide for stronger target interactions than the Nimblegen DNA baits. This means that hybridisation times are shorter using Agilent technology. The other key difference is the bait design. Agilent have a bait length of 120bp with the baits sitting end-to-end across a target region. In comparison, Nimblegen baits are shorter at 55-105bp sequences and are overlapped densely across the target (see Figure 1). Nimblegen uses a balancing technique when designing their baits that results in more baits across regions that do not hybridise as well and less across those that do. The result of this is a better uniformity of coverage.

Logistically the difference between the two kits is Agilent provides an „all in one kit" including library preparation reagents, whereas the user has to supply some of the oligos and library preparation reagents for the Nimblegen kit. Nimblegen recommends the Illumina TruSeq library preparation kit. This difference has an impact in terms of time, cost and reagent efficiency.

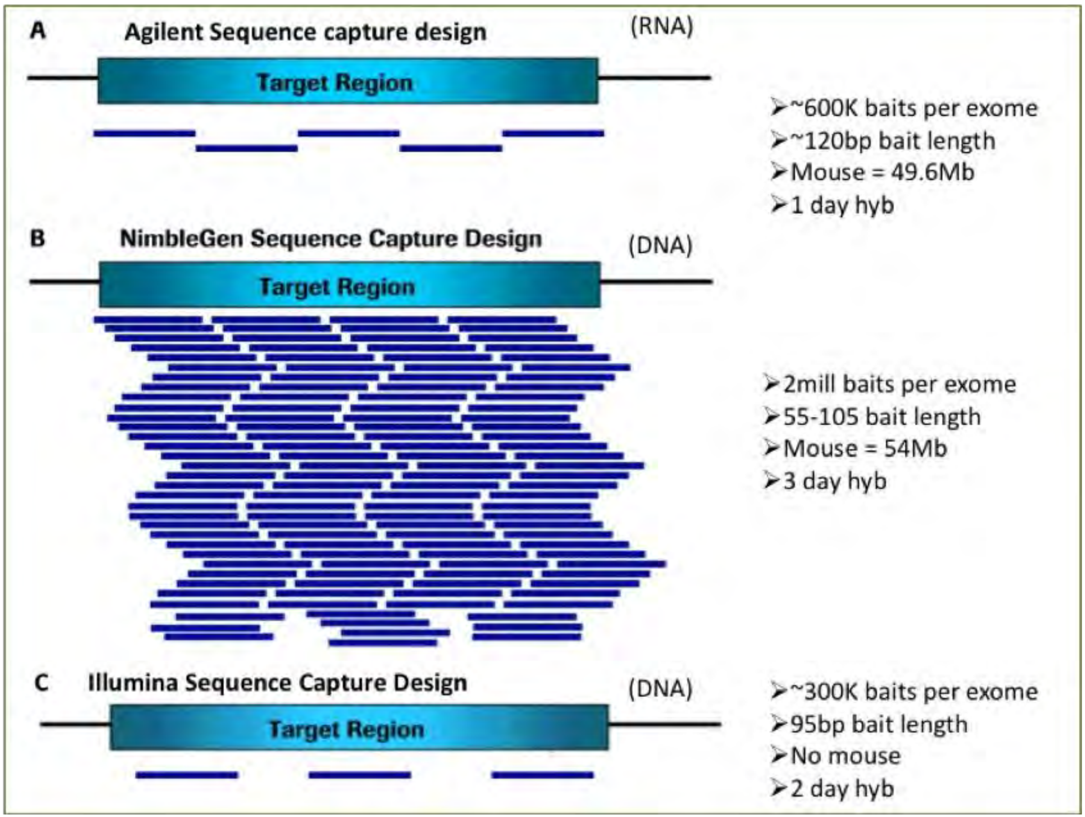

Figure 1. Comparison of exome capture bait tiling designs of the three technologies. Illumina do not currently offer a mouse exome capture kit

A plus for the Nimblegen technology is that it offers an indexing prior to capture protocol allowing users to index up to 12 samples per capture. Although Agilent currently do not support indexing prior to capture (however have announced a protocol release early in 2012), the APF have developed an in-house Agilent preindexing protocol to index 6 samples prior to capture and run in one lane as for the Nimblegen. The APF have started indexing 6 samples per capture in order to deliver sufficient coverage per sample, and run this in 1 lane of an Illumina HiSeq2000 using version 3 chemistry.

Running your sample - The Illumina HiSeq platform using TruSeq chemistry

The Illumina TruSeq chemistry uses an array technique to achieve cloning-free DNA amplification. Based on reversible terminator chemistry it provides massively parallel sequencing of millions of DNA fragments. The clonal bridge PCR amplification can create an ultra-high density sequencing flow cell containing hundreds of millions of clusters that, in turn, contain some 1000 copies of the same DNA template. These templates are finally sequenced through the sequencing-by-synthesis technique that applies reversible terminators with removable fluorescent dyes.

Illumina is continuously increasing the output of the HiSeq platform. At present, a 100bp, paired-end run using both flow cells produces about 600Gb of data. In terms of data volume, 100 million paired-end reads produces two uncompressed text output files (in fastq format) of 7-10Gb each. These compress (with bunzip2 or gzip utilities) to around 2Gb each. Once aligned to the genome and stored as a BAM format file, the total data size is reduced to around 4Gb. Our current variant call results spreadsheet generated from such data can have several thousand rows, but are generally small enough to be emailed (< 5MB).

There are two standard methods for moving Fastq files between sequence and bioinformatic facilities. The first is by an FTP site however the sequencing facility needs to have this option available. The second is simply by hard disk in the post.

3. Data output, quality indicators and how to interpret your data

Post-sequencing, the mouse exome sequence data returned to you should be associated with a number of quality and mapping statistics (or reports). The following section outlines what we have found to be the optimal characteristics of sequence data – in terms of quality, read numbers, lengths and mapping results – for accurate SNV detection. Standard mapping read-outs are explained and some guidance on their interpretation is provided. First, a key point of confusion will be discussed – exon coverage.

What is covered and what is not covered?

There are two points to raise when discussing how much of the exome has been covered. The first relates to what constitutes the „mouse exome", and the second to the design of exome capture technology:

1. **Defining the exome.** There has been much debate as to which sequences of the mouse or human genome are truly protein coding. Although the commercial bait designs contain large overlaps, they target different regions of the genome. To be able to measure effective coverage the APF use the Mouse CCDS as the reference target and ignore data from poorly annotated genes that may be included in the commercial exome bait designs. As a result, what the APF represent as „on target efficiency" or „percentage on target" is lower than what is commonly reported by Nimblegen or Agilent, as they report on the alignment of mapped data to the sequence used in the bait design. The difference being >70% of reads reported on target by mapping to the bait design, compared to 40-60% on target when mapping to the CCDS. For rare SNV detection, the advantage of this mapping difference is that it limits the number of false positive SNV calls arising from mis-mapping of reads. The disadvantage is that it does not record SNVs in poorly annotated, uncharacterised genes that may nevertheless still be involved in disease.

2. **Capturing the exome.** The weaknesses of exome sequencing coverage that one needs to be aware of: First our knowledge of all truly protein-coding exons in the genome is still incomplete, so current capture probes can only target exons that have been identified so far. Second, the efficiency of capture probes varies considerably, and some sequences fail to be targeted by capture probe design altogether. Third, not all templates are sequenced with equal efficiency, and fourth not all sequences can be aligned to the reference genome so as to allow base calling. Indeed, the effective coverage of exons using currently available commercial kits does vary.

What coverage do I need?

The APF have evaluated the amount of sequencing required to produce optimal coverage for accurate SNV detection. To evaluate how deeply an exome should be sequenced, we simulated an exome sequencing experiment where incremental proportions of one lane (of an Illumina GAIIx) of exome sequence reads were randomly sampled from a full lane of exome data. While reliable homozygous variant calls were made at even shallow read depths, a substantially greater depth was required for reliable heterozygous variant calls. True-positive heterozygous variant calls increased dramatically with increasing depth up to half a lane of Illumina GAIIx sequencing data or 30M reads, and 93% of true positive mutations were detected with 35-40M reads (22-26x median depth). With increasing read depth beyond this value, relatively few additional true positives were called but the number of false positive heterozygous SNV calls doubled. This indicates that four multiplexed samples could optimally be sequenced in a full lane of a HiSeq 2000 instrument.

The amount sequence required is a function of the average read depth necessary, which is in itself a function of the efficiency of the exome capture (which in turn can be affected by the quality of the input DNA). As a guide, for calling heterozygous variants we recommend a median read depth over the exonic portion of the genome of greater than 30x. This generally requires sequencing 80-100 million reads total, which is approximately a quarter of a lane on the HiSeq 2000. As a general rule, about 30 million of these need to align uniquely to the coding portion of the genome to allow sufficient sensitivity for variant detection. Fewer reads will still be useful, though at a loss of sensitivity (increased false negative SNV call rate).

Table 1. Mapping statistics

Gb per sample:	~**5Gb** data/exome if **6** samples have been run per lane of an Illumina Hiseq 2000 **or** ~**7Gb** for **4** exomes/lane (4 exomes/lane is recommended for low false negative results)
Number of reads per sample:	~**50M** reads (100bp paired end) for **6** exomes/lane **or** ~**80M** reads for **4** exomes/lane
Number of reads mapping to the genome:	**>95%** of your reads should map to the genome
Number of reads mapping to the exome:	We have seen on-target capture efficiencies that range **from 30% to better than 60%.** *Improvements in capture technology have seen a steady rise in these capture efficiencies; however run-to-run variation in the capture efficiency can be large. Recently, introduced protocols using pre-capture indexing has resulted in a drop back to the lower end of capture efficiency. This is largely due to the quantity of each sample pooled in the capture.*

Coverage reports

Coverage reports come in a variety of formats and content. The APF report format provides clients with a file listing exon coordinates and coverage over these regions. This can be used to check what has been covered adequately and what has not.

Most sequencing facilities report on the average coverage over the target exome. This can be a poor indicator of data quality and the success of the sequence run. For example, if 50% of the exome is covered at 5x, and 50% is covered at 55x the average coverage will be reported as 30x. A more useful report will provide a table or graph showing the distribution of coverage across a reference such as the CCDS. Figure 2 shows an example of a per chromosome CCDS coverage report from an analysis run at the APF.

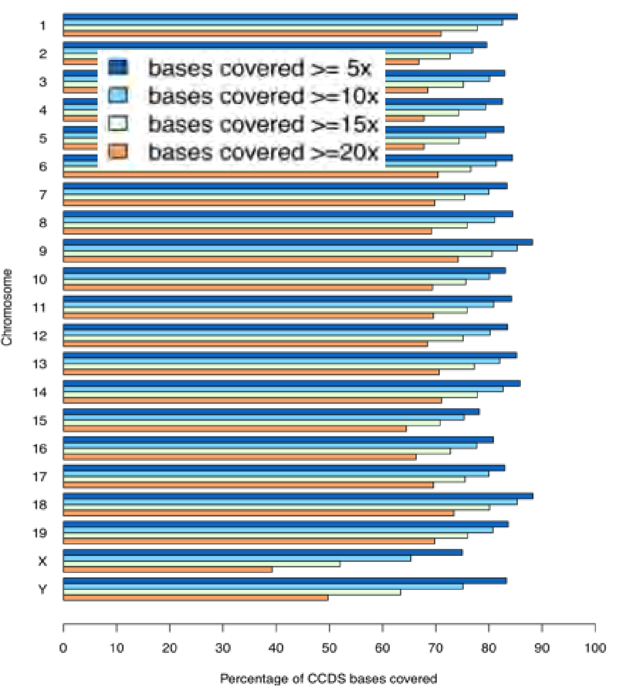

Figure 2. CCDS Coverage per chromosome at increasing levels of minimum read depth. This data was generated using Agilent capture and was processed using the APF analysis pipeline.

Quality reports

Quality scores (Q scores) can be given for (i) individual reported SNVs, (ii) median/average quality score across all SNVs, or (iii) median/average quality across all sequence data. When run on the HiSeq 2000 with version 3 chemistry, you should expect paired-end 100 bp runs to produce Q scores of above 30 for >80% of bases (rising to >85% of bases for paired 50bp runs).

SNV reports

SNV reports are often listed as standard outputs for various sequencing and bioinformatic providers. However, as outlined below, these raw SNV lists can be disappointing. The APF has developed an analysis pipeline that can successfully identify true unique SNVs from the several thousand typically generated from whole exome, short read sequencing.

Number of total SNVs

These are all SNVs determined prior to filtering. There will be thousands of these but many will be false positives (see APF Bioinformatic Pipeline below). Be wary that this is the only SNV list that you may get from most sequencing/bioinformatic facilities.

APF Bioinformatic Pipeline

When raw sequence datasets are submitted to the APF for SNV analysis, certain accompanying information is requested. The relationship of the samples (i.e. mice) with other samples is necessary. If any of the samples share rare genotypes - due to inheritance - these will be filtered out by the method used by the APF to remove recurring variation. The strain background of the mouse is also useful to know as strain-specific data sources exist to partially filter strain variation.

In terms of bioinformatics tools, the computational workflow comprises sequence alignment, variant calling, filtering, annotation and reporting steps (figure 3). Currently the APF uses BWA for alignment and Samtools for calling raw variants. We use a suite of in-house tools for filtering known variation, including an in-house catalogue of previously observed polymorphisms. Following SNV filtering, annotation steps assign the variants as synonymous, non-synonymous or splicing classes and tags each SNV with information about known phenotypes, gene expression pattern and Polyphen score. These results are compiled in a spreadsheet report that lists high-confidence SNVs that pass all filters at the top. Failing SNVs are also included in this report along with the criteria on why they were filtered out to allow researchers to perform their own custom filtering, should they wish this.

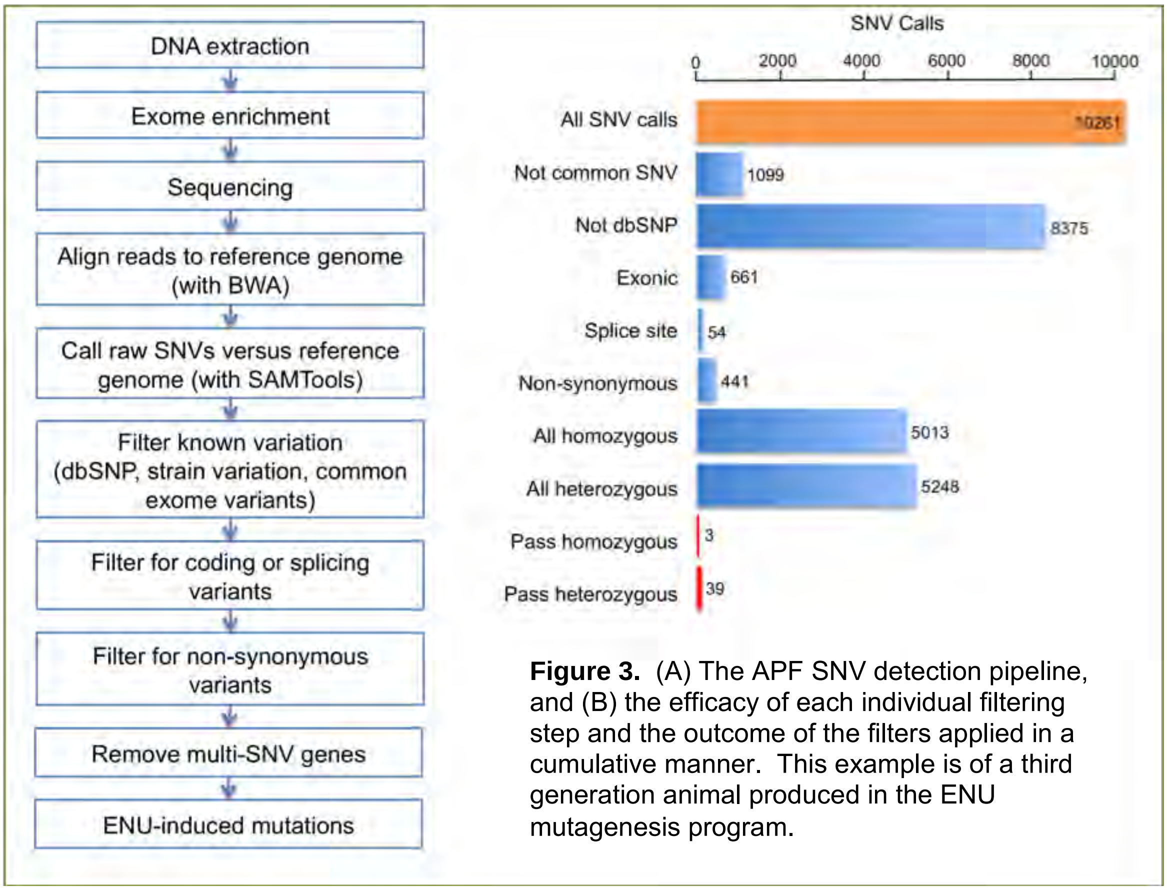

Figure 3. (A) The APF SNV detection pipeline, and (B) the efficacy of each individual filtering step and the outcome of the filters applied in a cumulative manner. This example is of a third generation animal produced in the ENU mutagenesis program.

Genes with multiple SNVs

This is a list of genes carrying more than one SNV. ENU delivered at the dosage we use induces ~1 SNV per Mb, which suggests that any genes with more than one SNV are likely to be a false positive due to mis-mapping. You will not find this in your report files from most facilities, but is included in reports from the APF.

What % of filtered SNVs will be true positives?

Every bioinformatic SNV pipeline will have different average false positive and false negative rates. With the APF SNV pipeline we are achieving false positive and negative rates of <20%. The false positive rate can be reduced by further filtering using the following cut-offs:

i. Quality: removal of all SNVs with a Q score of <30,
ii. Read depth: removal of heterozygous SNVs with <16x read depth,
iii. G/C transversions: removal of any G/C transversions unlikely to be generated by ENU.

This additional filtering may also remove some true positives and should be used with caution.

Other Bioinformatic Information

- For each ENU-mutated mouse we sequence, we expect in the range of 40-60 mutations to be found in the exonic portion of the genome. Our SNV calling pipeline usually finds 40-60 variants per genome, noting that our ascertained false positive call rate roughly equals our estimated false negative call rate. Depending on the generation of mice and the breeding undertaken, about 10% of these mutations are found in the homozygous form.

- When sequencing inbred mice that have been ENU-mutated, unless any two mice are related to the founding G0 male from which they were derived, any recurring variants can be considered unlikely to be ENU-induced. We filter this kind of variation and have compiled a large catalogue of recurring variation. This catalogue will include local, strain-specific variants, sequencing platform-specific variants and other common sources of false positive calls, such as common alignment errors.

- When sequencing mice from a non-reference background strain (i.e. not C57BL/6 mice) the majority of variants called will be strain-specific, as the reference mouse genome sequence is used to align with the sequenced reads. It is common to derive a list of strain-specific variants by sequencing an un-mutated mouse of this strain. However, this has proven to be not necessary - or at least, not the most cost-effective strategy. If the induced mutations within any given mouse can be assumed to never coincide with the mutations in an unrelated mouse, sequencing of multiple mutant mice from this strain can be used to compile a list of recurring variants, which can be filtered from the final list of presumed induced variants.

- We have found, especially for mice from a mixed strain background, that it can take up to twenty different exome sequences to capture all of the common variation of a non-reference strain (or indeed, to observe all of the strain variation in local colonies of reference C57BL/6 mice). Compiling a list of commonly observed variation aids the filtering of strain-specific variation and shortens the lists of candidate induced mutations that we generate. Sequencing five mice from the same strain background greatly reduces the calling of rare strain variants as induced mutations. After twenty mice from the same background are sequenced almost all strain-specific variants have been observed.

4. Validation of variants

Even the most advanced variant detection pipeline is not perfect, and eventually you will need to confirm an SNV as a true mutation, rather than a product of platform or software noise. There are a number of options available, and different research teams use different approaches. Some examples of validation techniques include Sanger resequencing and amplicon sequencing using one of the benchtop NGS instruments such as the Ion Torrent or MiSeq. The APF uses a fluorescent-based PCR with primers specific for the SNV (Amplifluor technology). We find this approach convenient as it allows us to validate the SNV and conduct broader genotyping using a single assay.

Validation of our SNV calls indicate a specificity of 80%, which is high given the propensity for false positive SNV calls from analysis of short-read data. From the read depth profile we see for exome capture sequencing, we can estimate the sensitivity of our SNV calls to be also a little lower than 80%. The missing calls are explained by the approximately 15% of the exome that is currently not efficiently targeted by any capture technology.

5. Appendix: Terminology and concepts

Baits
These are DNA or RNA fragments that have been designed to the target region and are used to “capture” the region of interest from sheared genomic DNA. The baits are biotinylated to allow “pull down” using streptavidin coated beads.

CCDS
See Exome

Coverage
Coverage can have many different meanings and can be theoretical or empirical. The term is often misunderstood or misused. Fold and depth of coverage are not the same because of sequencing error & unmappable regions of the genome.
The result often being that clients of large sequencing facilities often do not get the amount of data they are expecting. This is discussed further in the „recommendations‟ section of this document.

- **Theoretical "fold-coverage"**
  Often seen as X – coverage. Given the number of raw bases sequenced how many times (X) does the sequencing potentially cover the exome:
  *equals (number of reads * read length) / target size*

- **Theoretical or empirical "breadth-of-coverage" of an assembly**
  Often seen as % coverage it illustrates how well the genome is actually covered after all mapping and assembly is done:
  equals assembly size / target size

- **Empirical average or median "depth-of-coverage":**
  *equals (number of reads * read length) / assembly size*

Depth
Sequencing depth refers to the number of nucleotides contributing to a region of an assembly (e.g. base, gene, exome or chromosome). On an exome basis, it means that, on average, each base has been sequenced a certain number of times (10X, 20X etc). For a specific nucleotide, it represents the number of sequences that added information about that nucleotide. Such depth varies quite a lot depending on the exomic region. In consequence, an average sequencing depth of 30X leaves a lot of small portions of a genome unsequenced while others receive a lot more sequences. This can also be represented as “depth of coverage” (see Coverage). Median (rather than mean) depth is often used as a better indicator of coverage.

Duplicate reads
Duplicate reads are reads of the same sequence that will align/map to the same place in the exome. Duplicate reads are a result of a PCR amplification step in the library preparation protocol. Facilities with more experience usually reduce the number of PCR cycles to prevent or lower the percentage of duplicates obtained. Many bioinformatic pipelines remove the duplicate reads in the sequence data that may or may not reduce the percentage of reads that map to the target (exome).

Exome
The exome is the coding part of the genome and is estimated to comprise 1.5% of the total mouse genome. There is considerable uncertainty around defining which sequences of the mouse genome are truly protein coding. Each capture development company is choosing to use one or more of the below annotations to design their exome baits to:

- **CCDS -** The Consensus CDS project is a collaborative effort to identify a core set of protein coding regions that are consistently annotated and of high quality.

- **Genbank -** The GenBank archival sequence database includes publicly available DNA sequences submitted from individual laboratories and large-scale sequencing projects. It is part of the International Nucleotide Sequence Database Collaboration (INSDC) along with the European Nucleotide Archive and the DNA Data Bank of Japan (DDBJ). Submitted sequence data is exchanged daily between the three collaborators to achieve comprehensive worldwide coverage. As an archival database, GenBank can be very redundant for some loci.

- **RefSeq -** The Reference Sequence (RefSeq) database is an open access, annotated and curated collection of publicly available nucleotide sequences (DNA, RNA) and their protein products. This database is built by the National Center for Biotechnology Information (NCBI), and, unlike GenBank, provides only single record for each DNA molecule. The collection aims to provide a comprehensive, integrated, non-redundant, well-annotated set of sequences, including genomic DNA, transcripts, and proteins.

- **Ensembl -** The Ensembl gene set is based on protein and mRNA evidence in UniProtKB and RefSeq databases, along with manual annotation from the VEGA/Havana group.

Flow cell
Single use slide with 8 lanes (channels) used by Illumina. Two flow cells can be run simultaneously on the Illumina HiSeq 2000 instrument.

Index
A short, unique sequence of DNA that is attached to DNA fragments so that they can be pooled and processed in parallel. The sequence data for each sample can then be deconvoluted using the index sequence. Indexing can either be performed

following capture to allow multiple samples to be run in 1 lane of an Illumina slide, or prior to capture allowing you to also reduce the labour and cost of capture. It has become standard to index samples prior to sequencing due to the improvements in sequence instrument outputs. Currently for exome capture Illumina and Nimblegen support indexing prior to capture to reduce the costs further. Agilent have planned a release date for their protocols for early 2012. Although indexing prior to capture improves the cost and processing time of exome libraries it does have some effect on capture efficiency.

Lanes
One of 8 channels of an Illumina flow cell. May hold a single sample or a group of indexed samples. One lane on each flow cell is always used to run a PhiX control library so only 7 lanes are available for samples.

Mappable reads (in the context of exome seq)
Reads that can be determined to originate from a single location in the exome. With exome sequencing, bioinformatic cores often give you the number and percentage of reads that map/align to the whole genome and then the percentage of reads that map/align to the exome.

Paired-end reads
DNA sequences from each end of a DNA template. "Paired end" refers to how the library is made, and then how it is sequenced. Both are methodologies that, in addition to the sequence information, give you information about the physical distance between the two reads in your exome.

If the library preparation process includes the selection of 500bp fragments, paired end sequencing will sequence 75 or 100bp from each end. Through this process you will have three pieces of information:

i. 75-100bp sequence 1
ii. 75-100bp sequence 2
iii. The distance separating the two sequences

This allows mapping to a reference using the distance information and helps dramatically to assemble across repetitive regions, significantly improving coverage and going some way to ameliorating the limitations of short read technology

Quality score (Q score)
Reliable, validated base quality scores (also known as Q scores) provide a standard with which to compare data. Originating with the Phred program used for Sanger sequencing, the Q Score is universally used and has become a critical tool for comparing results. Phred-like quality scores compress a variety of types of information about the quality of base calls into a readily usable probability-of-error value. Many analysis tools and virtually all assemblers require quality score input to deliver accurate results.

Reads/Read length
A read refers to a short DNA sequence that is the output of a Next Gen sequencing instrument. Typically for Illumina, data reads are either 75bp or 100bp. These are considered to be short reads as compared to data from instruments such as the Roche 454. The choice of read length for a sequence run is dependent on which capture technology is used.

Target
The target is the genomic regions that the capture baits have been designed to. In the case of exome capture, the target may be defined in several ways (see Exome).

6. Reference and further reading

M.J. Clark, R. Chen, H.Y.K. Lam, K.J. Karczewski, R. Chen, G. Euskirchen, A.J. Butte & M. Snyder. **Performance comparison of exome DNA sequencing technologies.** *Nature Biotechnology* 2011, 29:908-914. PMID: 21947028

Manuscript submitted:
T. D. Andrews, B. Whittle, M. Field, Y. Zhang, Y. Shao, B. Balakishnan, E. Cho, Y. Xia, M. Kirk, Jörg Hager, G. Sjollema, B. Beutler, S. Winslade, A. Enders & C. C. Goodnow. **Massively parallel sequencing of the mouse exome to accurately identify rare, induced mutations: an immediate source for thousands of new mouse models.**

For more information on the APF sequencing and bioinformatics services visit the APF website at http://apf.anu.edu.au.

For more information on the Illumina HiSeq systems and the TruSeq chemistry, visit the Illumina websites below:
http://www.illumina.com/systems/hiseq_systems.ilmn
http://www.illumina.com/truseq.ilmn

For more information on the Agilent, Nimblegen and Illumina exome capture technologies, visit:
http://www.halogenomics.com/sureselect/how-it-works
http://www.nimblegen.com/seqcap/
http://www.illumina.com/products/truseq_exome_enrichment_kit.ilmn

For more information on the Amplifluor SNP detection technology, visit:
http://www.biosearchtech.com/store/product.aspx?catid=224,171,158