Generation II DNA sequencing technologies

potential impact on drug discovery and development

Generation II DNA sequencing has been widely heralded as a disruptive technology, generating tens of millions of random short sequences at efficiencies up to 20,000-fold greater than Generation I (Sanger) sequencing. This review addresses the technical specifications of the leading Generation II sequencing technologies and current applications of relevance to drug discovery and development.

By Dr Stephen F. Kingsmore, Dr Jenny C. van Velkinburgh, Dr Joann Mudge, and Dr Gregory D. May

Before discussing specific areas of drug discovery and development that we anticipate will be profoundly impacted by Generation II (or NextGen) sequencing, we first review current knowledge of nucleic acid variation, its functional correlates, and the current technical capabilities of Generation II sequencers.

Deoxyribonucleic acid (DNA) molecules are composed of four nucleotide bases (the purines Adenine and Guanine, and the pyrimidines Thymine and methylated and non-methylated Cytosine). 3'-5' phosphodiester bonds join the three billion bases of the human genome in 23 pairs of linear polynucleotide chromosomes (Chr). Native DNA molecules are redundant, being formed through specific base pairing between purines and pyrimidines and resulting in two antiparallel, complementary strands. The two distinct copies of each Chr present in each diploid human cell form the basis for inheritance of traits and diseases. Thus, diploid human genomes contain six billion base pairs[1]. Inherited variation between individuals has two principal components – approximately four million single nucleotide variants (single nucleotide substitutions [SNPs] and deletion insertion polymorphisms [DIPs], in a 10:1 ratio, comprising 1% of the genome) and approximately 10,000 structural variants (DNA stretches that are inserted, deleted, inverted or translocated [copy number variations, CNVs], comprising up to 10% of the genome)[2,3]. Cells in an individual also develop genome variation, primarily somatic mutations (which are the primary driver in neoplasia)[4], viral genome insertions and cytosine methylation (epigenetics)[5]. The human genome references which have been established comprise approximately 90% of the genome which can be cloned. Catalogues of somatic and meiotic genome variation remain very incomplete and understanding of the functional consequences of genetic variants is in its infancy. One of the first applications of Generation II sequencing has been maturation of reference genome sequences and variation catalogues, with compilation variant frequencies by sequencing of many human genomes (resequencing)[116-119]. These efforts are of critical importance for drug discovery and development since they form the reference sets against which disease- or drug-related changes are measured.

Approximately 25,000 protein-coding genes are interspersed throughout the genome as discontinuous sets of exons. Expression of genes occurs through transcription – the assembly of continuous, single stranded, messenger ribonucleic acid (mRNA) copies of exons, which are translated into proteins. Protein diversity largely results from multiple, alternative ways in which exons are assembled into mRNAs (alternative splicing)[120]. While exonic DNA, which accounts for 1-2% of the genome, is the code of life, it is transcription that is the master controller of conversion of that code into cellular, organ and organismal activity. The emerging, primary function of the remaining 98% of the genome is to control gene transcription precisely and ensure faithful copying of genomes during cell replication. It has recently been discovered that most of the genome is transcribed[6-9]. The non-exonic component of the genome accomplishes control primarily by transcription into a plethora of regulatory RNA types (including smRNA, ncRNA, tRNA, rRNA and antisense transcription) or by containing structural genomic DNA features that influence transcription and replication (both control elements and DNA methylation). The set of RNA molecules in a cell (transcriptome) is highly complex, dynamic and unique to that cell and its state at that moment. The activity of a gene in a particular cell and state can be inferred by measuring and integrating the abundance of its transcripts and regulatory RNAs. Since genes usually act in networks and pathways, integration of measurements of the abundance of sets of transcripts and regulatory RNAs is more biologically useful. Given the enormous complexity and dynamism of transcription, this is typically informative only when comparisons are performed on groups of samples that differ in only a few parameters. Thus experimental designs are a critical determinant of the translation of transcriptional measurements into biological knowledge. Another early application of Generation II sequencing is measurements of transcriptome components (digital transcript expression, DTE) with greater comprehensiveness, flexibility, sensitivity and precision than possible with array hybridisation by counting millions of random sequence tags[121]. DTE is emerging as an important area for drug discovery and development by identification transcriptional changes that were not apparent by array hybridisation, identification of novel transcript isoforms and identification of novel mechanisms of transcriptional control (such as small RNA), which may be amenable to intervention[120,121].

Finally, given the current state of genome tools and knowledge, an immense current need is improved understanding of the functional consequences of genome and transcriptome features and variation (annotation). The third promise of Generation II sequencing is high throughput functional annotation by combining NextGen sequencing applications, such as genome resequencing and DTE, or methylome sequencing and DTE. For example, genome-wide association of SNP genotypes and mRNA sequencing enables identification cis- and trans- acting nucleotide variants that affect gene expression or splicing (eSNPs)[10-20] (Kingsmore et al, unpublished). Identification of eSNPs is anticipated to be an important approach for translation of genome-wide association signals into knowledge of disease gene perturbations. In toto, these applications of Generation II sequencing are anticipated to have an immense impact on molecular diagnostics, pharmacogenetics, drug discovery and development[21-25].

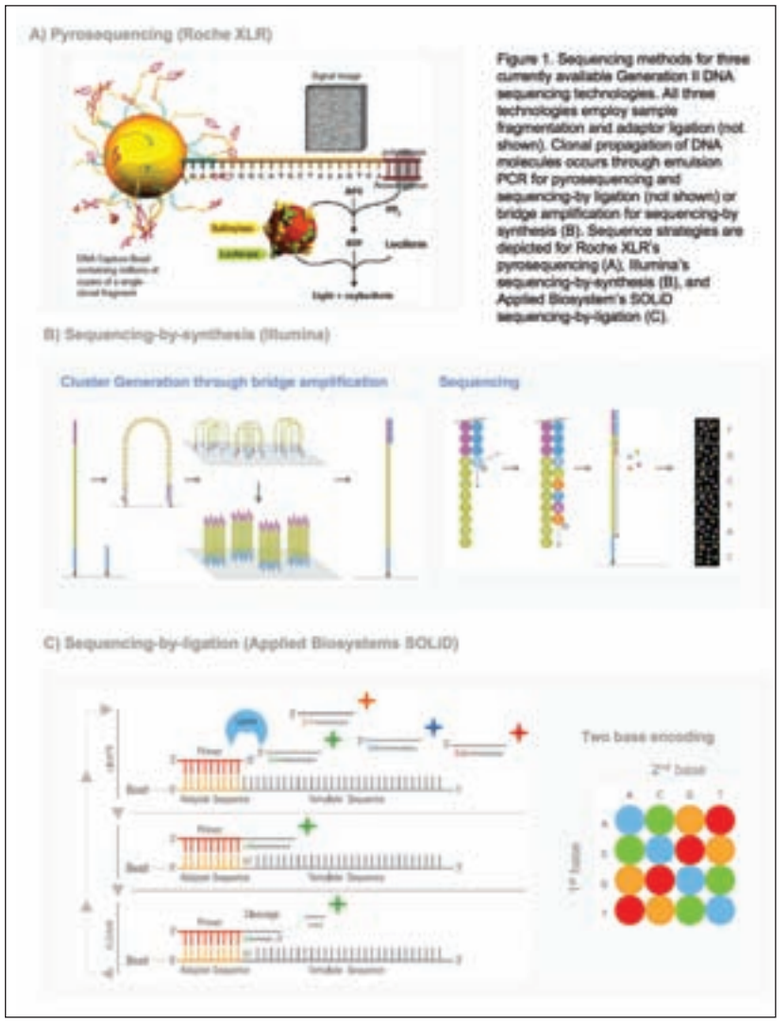

Figure 1. Sequencing methods for three currently available Generation II DNA sequencing technologies. All three technologies employ sample fragmentation and adaptor ligation (not shown). Clonal propagation of DNA molecules occurs through emulsion PCR for pyrosequencing and sequencing-by ligation (not shown) or bridge amplification for sequencing-by synthesis (B). Sequence strategies are depicted for Roche XLR's pyrosequencing (A), Illumina's sequencing-by-synthesis (B), and Applied Biosystem's SOLiD sequencing-by-ligation (C).

Generation I DNA sequencing technology

Generation I DNA sequencing uses Sanger chemistry and electrophoretic product separation to decode the primary structure (linear order of bases) of isolated (cloned) DNA molecules[26-28]. It involves cycled, primer extension by incorporation of deoxynucleotide triphosphates (dNTPs) complementary to thousands of cloned, identical single-stranded DNA fragments, with reaction termination by dideoxynucleotide triphosphates (ddNTPs) lacking the 3'-hydroxyl group necessary to bond the 5'-amino group of the next dNTP. Thermostable DNA polymerase lacking 5'→3' exonuclease activity allows cycled primer elongation without degradation of templates or oligonucleotide products. Reactions use optimised ratios of dNTPs and four ddNTPs, each with a different label to generate nested sets of copies of the template of varying length that are labelled according to the terminal nucleotide. Electrophoretic separation on the basis of fragment length, together with label identification, allows determination of base composition. Generation I DNA sequencing instruments generate 96,800bp, high quality sequences per run at a cost of $1 per kb. Highly automated, Sanger production sequencing generates up to 2mb of sequence per instrument per day. Cost and the need to isolate individual clonal DNA templates (which greatly limits tag counting applications) are the principal limitations of high throughput Generation I sequencing. Sanger sequencing remains, however, the technology of choice for targeted, low and medium throughput, long, high quality sequencing.

Generation II DNA sequencing technologies

Currently, four Generation II sequencing approaches dominate the market and are established in the scientific literature and additional platforms are continuing to be released (**Table 1**, **Figure 1**). They include three sequencing-by-synthesis (SBS) platforms (pyrosequencing from Roche Applied Science XLR, the Danaher-Motion Polonator and the Illumina GA IIx) and the Applied Biosystems SOLiD 3.0 instrument, featuring sequencing-by-ligation (reviewed in [29-31]). Generation II sequencing technologies have several features in common: First, all have similar DNA or cDNA template preparation procedures that obviate cloning vectors, propagation in a bacterial host, and clone isolation. Instead, DNA or cDNA templates are randomly fragmented, ligated to application-specific (and proprietary) adapters, and clonally amplified on a solid phase. This results in gigabase-per-day scalability and vastly improved ease-of-use. It also results in production of random (shotgun), rather than directed, sequences. Second, all sequencing reagents flow over a fixed array of template fragments, permitting digital image capture of fluorescent or chemiluminescent signals. This obviates the need for electrophoresis, another limitation of traditional sequencing. Third, they generate somewhat shorter, often paired reads with higher error rates than traditional sequencing. This increases the difficulty of accurate read alignment, assembly and variant detection (see below).

Pyrosequencing using emulsion PCR

Pyrosequencing was the first Generation II sequencing technology to become commercially available and, therefore, boasts the largest number of peer-reviewed application manuscripts[32-59] (**Figure 1A**). Templates are fragmented, adapters ligated and clonal amplification is performed by capture of single fragments on beads containing primers complementary to adapters and emulsion PCR. Following emulsion breaking, beads are deposited into microwells in a 1:1 bead:well stoichiometry. dNTPs are sequentially flowed over the wells, one at a time, and pyrophosphate, released when a nucleotide is incorporated by primer extension into the strand complementary to the template, is converted to ATP by sulfurylase, resulting in luciferase-generated chemiluminescence. Unincorporated nucleotides are degraded by apyrase and the dNTP flow sequence is repeated. Chemiluminescence intensities are algorithmically translated into base-calls. Sequential, identical bases produce an incremental signal increase. Advantages of pyrosequencing are relatively longer read lengths (500bp), availability of long paired reads (>2kb span) and applications reported in more than 300 peer-reviewed publications. Long paired reads are generated by circularisation of long DNA fragments with a biotin-labelled segment separating template ends. Following circle fragmentation and biotin-containing fragment capture, emulsion PCR is performed. The Roche XLR instrument currently generates 500mb per run; however, pyrosequencing has a higher cost per gigabase than some other Generation II approaches, which limits its use to smaller projects, *de novo* sequencing, metagenomics, identification of genomic structural variants and targeted resequencing (in combination with Nimblegen capture arrays). In contrast, the Polonator (introduced in 2008) is a substantially lower cost platform, generating much shorter reads at low cost and using generic reagents, flexible chemistry and open-source informatics[72,82-84].

Table I: Comparison of Gen I (AB 3730xl) and current Gen II sequencing technologies

	SANGER	AB SOLID 3.0	ILLUMINA GA IIX	ROCHE XLR
Max read length	1200	2 x 50	2 x 125	600
Max reads per slide/plate	96	200 million	100 million	1 million
Max bases per slide/plate	86,000	20 billion	25 billion	1 billion
Run time for max bases	60	1	10 days	10 hours
Cost per GB	NA	$	$	$$
Accuracy	99.9%	99.9%	99%	99.5%
Ease of Use	+	++	++++	++
Peer reviewed manuscripts	++++	++	+++	++++
Notable Applications	Targeted & de novo sequencing	Resequencing, mRNA-seq	Resequencing, mRNA-Seq, methylation, ChipSeq	*De novo* sequencing, metagenomics, targeted resequencing

Sequencing-by-synthesis using bridge amplification (Illumina GA IIx)

The second commercially successful NextGen technology is marketed by Illumina, Inc and employed unique methods of solid-phase polony amplification ('bridge' PCR, described below) and fluorescently-labelled reversible chain terminators[8,9,60-79] (**Figure 1B**). Templates are fragmented, adapters ligated and bridge PCR is performed to amplify polonies (polymerase colonies) anchored directly on a flowcell containing primers complementary to both adapters. Sequencing of resultant polonies is carried out in the flowcell using a primer complement oligonucleotide and four dNTPS with cleavable fluorescent labels and reversible terminators. Following the first cycle of SBS, images are captured, the blocked 3'-terminus and fluorescent tag are removed, and further cycles are performed. Paired reads are generated by flowcell denaturation following SBS using the first primer, followed by SBS using a primer complementary to the adapter at the other end of the polony fragment. The Illumina GA IIx instrument generates 420gb per paired run, dependent upon read length (36-106bp). Advantages of Illumina's SBS are lower cost per gigabase, large number of sequence tags (80 million per flowcell) for counting purposes, ease-of-use and practical applications established in more than 200 peer-reviewed publications. Principal applications of Illumina SBS relevant to drug discovery include genome resequencing and DTE.

Sequencing-by-ligation (Applied Biosystems' SOLiD)

Originally described by Syndney Brenner[85], sequencing-by-ligation (SBL) has been successfully commercialised by Applied Biosystems. The SOLiD 2.0 system (**Figure 1C**) employs nearly identical sample preparation as the Roche FLX. Following emulsion breaking, however, beads are attached to a flowcell similar to the method described for the Illumina GA IIx. SBL is performed by template-directed ligation of eight-base primers, labelled in one position and comprised of all possible sequence compositions, to the sequencing primer. Following image acquisition, the ligated primer is cleaved internally, exposing a 5' phosphate, and four further rounds of ligation are performed. The sequencing primer and ligation product are then removed and a second, offset sequencing primer is annealed. This process is repeated, generating 35-50bp paired or singleton reads, respectively. The SOLiD instrument generates ~20gb per pair of slides. Advantages of SBL are low cost per gigabase, large number of sequence tags (120 million per slide) for counting applications, and very high accuracy (since each base is interrogated twice). Current limitations are the need to align to a reference database for basecalling (preventing use in *de novo* sequencing), and short read lengths. Emerging applications of SBL are genome resequencing and DTE[86].

Emerging generation III DNA sequencing technologies

Helicos Biosciences' HeliScope instrument[80,81] and Pacific Biosciences Single Molecule Real Time (SMRT) technologies use a similar approach to the Illumina GA IIx but represent the first single molecule sequencing technologies, avoiding amplification-induced artifacts[122,123]. By means of a melt-and-resequence method, Helicos eliminates an amplification step but requires a more expensive detection system, and is capable of sequencing multiple flowcells/slides simultaneously. The cost of the HeliScope and rapid advances in competitor platforms have deterred most customers to date. SMRT technology promises read-lengths that vastly exceed traditional Sanger sequencing, but is not yet commercially available.

Computational and bioinformatic implications of Generation II DNA sequencing

In Generation II resequencing and DTE applications, the sequence of short, random, clonal DNA fragments is decoded and either aligned to a reference database and/or overlapping fragment sequences are assembled into contiguous stretches (contigs). This is performed millions of times for each sample, in order to provide comprehensive coverage of all DNA species in that sample. Alignment and assembly are complicated by several technical considerations: Firstly, Generation II sequencing is relatively error prone (raw sequence accuracy of 98-99.5%), and, therefore, a consensus of redundant sequences must be determined if nucleotide variants are to be distinguished from basecalling errors[124]. In addition, the quality score of each nucleotide call is an important qualification of nucleotide variants. As a result, the cost effectiveness of Generation II sequencing is partly offset by the need for redundant coverage. Secondly, eukaryotic DNA contains a considerable amount of polymorphic, repetitive, paralogous and low complexity sequence. Since Generation II sequences are short, they may not bridge from such regions to adjacent unique sequences. Thus, unambiguous assignment is not possible for all sequences, particularly when nucleotide variation may also be present. This is particularly troublesome for highly polymorphic or large paralogous gene families or for alignment of short cDNA sequences to genomes[124]. For demanding applications, such as methylome or structural variant discovery in whole genome shotgun resequencing, long reads or paired sequences (sequences derived from both ends of DNA fragments of known length) are necessary. These considerations are driving the improvement of Generation II sequencing instruments and development of new assembly/alignment algorithms. Thirdly, Generation II sequencing generates millions of random ('shotgun') DNA fragment sequences from genomic DNA or RNA samples. As a result, Generation II instruments are ideally suited for tag counting applications, such as DTE and metagenomics, providing molecule counts that are intrinsically superior to array hybridisation signals. However, they are ill-suited for targeted resequencing of thousands of samples. As a result, substantial technological innovation is occurring in DNA barcoding (to enable differentiation within multiplexed sample sequences) and in target capture methods (to enable multiplexed targeted sequencing)[32,64,71].

Following optimal alignment or assembly, two types of additional computation are performed. Either nucleotide variants are identified (substitutions, insertions, deletions, inversions or repetitions) or the number of sequence tags aligned to each reference gene, transcript or feature is enumerated and converted to a read frequency (a measure of the mass abundance of that species in the original sample). Finally, sequences and metadata (quality scores, alignments, variants, frequencies) are stored in a relational database and, typically, queried for marker-phenotype associations or correlations[124]. A typical Generation II sequencing instrument generates several terabytes of raw data per run and a typical Generation II sequencing project accrues tens to hundreds of gigabases of sequence. To put this in perspective, this corresponds to the entire DNA sequence in all public databases maintained by the National Center for Biotechnology Information (NCBI) as of April 2006. While the output of Generation II sequencing instruments is approximately quadrupling each year, similar gains in efficiency of data storage, computation and analysis have not occurred. Consequently, computational, software engineering and bioinformatic expertise and resources are rate limiting at present[86-89].

Current applications of second generation sequencing technologies

Currently, DNA sequencing is used in four principal applications: 1) *De novo* sequencing to create near complete, first reference sets of sequences that render a species tractable to genomic investigation; 2) Resequencing in which genes, genome segments (eg quantitative trait locus intervals, chromatin immunoprecipitates), exomes (all exons), transcriptomes, genomes, methylomes (methylated cytosines) are sequenced in many samples, either to establish catalogs of variation or to undertake

Figure 2.A web-based software system, Alpheus®, developed initially for use in a malignant pleural mesothelioma (MPM) resequencing project. A) The variant query and filter page allows querying of an underlying database. This page allows detection of high confidence variants filtered by quality, consistency, segregation ratio, and coverage thresholds as well as by variant type (nonsynonymous SNP, synonymous SNP, indel, stop codon). In addition, genes with variants can be compared between two groups of samples to find those more prevalent in one group (ie. cases vs. controls). B) Report of variants identifying variant position, nucleotide and amino acid change, the NCBI variant id, if any, the strand orientation of the identifying read, and the fraction of reads containing the variant. C) Pileup view showing a closeup of Illumina sequencing-by-synthesis reads (green arrows) aligned to the exons (green boxes) within the blue box. The back figure has no variant filtering applied.

marker – trait association studies[90]; 3) RNA profiling (DTE), in which specific RNA species (small or large, coding or non-coding, sense or antisense) in sets of samples are converted to complementary DNA (cDNA) and sequenced in order to determine the composition and/or abundance of transcripts within samples (by tag counting) for marker – trait correlations; and 4) Metagenomics, in which DNA from samples containing more than one species are sequenced in order to determine the composition and/or abundance of species within samples (by tag counting) for environmental correlations[91-94]. Of these applications, resequencing and DTE are anticipated to have the most profound and imminent impact on drug discovery and development and are the focus of the remaining discussion.

Resequencing

Resequencing refers to sequencing specific transcripts, genes, genomic regions or genomes from a number of specimens, usually in order to evaluate association between genotype (or haplotype) and a phenotype or trait. As with other systems biology experiments, it is divided into 'Discovery' and 'Validation' phases[95]. In the Discovery phase a relatively small number of specimens are sequenced comprehensively and candidate associations are identified. In the validation phase, a few candidate associations are sequenced in many specimens. This design confers power for evaluation of statistical significance in the validation phase, provided attention is given to population structure[96]. Hitherto, discovery studies were limited by cost and throughput to sequencing of small numbers of candidate genes or to genotyping of larger numbers of genes. A weakness of a targeted discovery phase is that experimentation is limited to the prevailing molecular knowledge (hypothesis-testing), rather than being systematic, and hypothesis-informing. During the past two years, genome-wide genotyping arrays have shown the immense power of systematic, hypothesis-informing approaches for identifying genetic associations in common, complex disorders (genome-wide association studies, GWAS, reviewed in [97]). Resequencing applications of Generation II technologies are anticipated to further our understanding of the molecular basis of common diseases and pharmacophore response in at least three ways:

First, Generation II resequencing is being used to improve reference catalogs of genes, transcripts, nucleotide variants and structural variants. The ENCODE project, for example, is an NIH-sponsored initiative to catalog all human transcribed elements. The ENCODE pilot project and similar efforts have indicated that alternative splicing may result in as many as 500,000 mRNA species in mammals[6-9,82,85,98-103]. Of these, only one-tenth are currently annotated (present in RefSeq transcript or AceView databases), and 10,658 of the latter are putative or provisional genes that have been identified informatically from the human genome sequence ('LOC', 'FLJ' and 'C_orf' genes)[104]. These projects have also identified new classes of transcriptionally active regions (TARS) of the human genome that are poorly described or understood remain undefined[3,4]. Generation II mRNA or rRNA-depleted RNA resequencing enables comprehensive identification of splice isoforms and TARS[8,9,33,34,37,39,41,45,48-50,52,53,56,58,69,78,85]. Incomplete or inaccurate reference catalogues can lead to erroneous conclusions that impact drug discovery and development. For example, numerous studies have sought associations between behavioural phenotypes and tryptophan hydroxylase (TPH), which is important in serotonin metabolism. It is now known that the gene (TPH1) examined in many of these studies is not expressed in serotonergic neurons[105,106]. Examples of large Generation II resequencing efforts to improve reference catalogs of nucleotide variants and structural variants are the 1000 Genomes Project[107] and the Personal Genome Project[108].

Second, Generation II resequencing is being used

in drug discovery to translate signals identified in GWAS into drug targets. Array-based GWAS are predicated on identification of associations between complex traits and ~1 million random, common (minor allele frequency >5%) SNP genotypes or imputed haplotypes (by comparison with genome maps of SNP-tagged haplotype blocks, 108) in thousands of cases and controls (reviewed in [97]). During the past two years, the effectiveness of GWAS for identification of replicable, common, susceptibility variants in complex diseases has been unequivocally established[97]. However, many GWAS signals are located in genomic regions lacking annotated genes or with multiple genes. Even GWAS signals located adjacent to a single gene do not usually indicate the causal variant or functional consequence of the variant. Generation II resequencing is being used, in combination with target enrichment strategies, for GWAS interval resequencing and for high throughput functional annotation of nucleotide variants. For example, genome wide association of SNP genotypes and mRNA sequencing enables identification of cis- and trans- acting nucleotide variants that affect gene expression or splicing (eSNPs)[10-20] (Kingsmore et al, unpublished). Identification of eSNPs is anticipated to be a principal approach for translation of genome-wide association signals into knowledge of disease gene perturbations. This approach, termed Genome-Tissue Expression Analysis (GTEx), is likely to be the subject of a large NIH effort in the near future.

Third, Generation II resequencing is being used in drug discovery to directly identify variants associated with common traits by non-hypothesis-directed sequencing of many human genomes (resequencing). For example, Generation II mRNA resequencing has shown utility for identification of somatic mutations in cancers[38,54,55,59,104]. One recent project sought to identify novel, putatively damaging, somatic mutations expressed in malignant pleural mesothelioma (MPM) surgical specimens using Generation II resequencing[104]. MPM is a prototypic environmentally-induced cancer (due to asbestos) and does not exhibit mutations in established oncogenes. 1.7gb of cDNA sequence was generated from six surgical specimens (four microdissected MPMs, one lung adenocarcinoma and one normal pleura) using Roche pyrosequencing. A web-based software system was used for sequence management, pipelining, visualisation, and statistical analysis (**Figure 2**). Singleton pyrosequencing reads aligned to ~16,000 genes and identified ~1,300 cSNPs in each sample, of which ~1,000 were known, inherited, common variants. 69 novel, putatively damaging mutations were identified, of which 15 nsSNPs were found in tumour but not normal coisogenic tissue, reflecting somatic mutations, RNA editing, LOH due to chromosomal deletions and epigenetic silencing, including Chr X inactivation. Three of seven somatic nsSNP mutations were identified in additional MPM tumours. In response to proofs-of-concept, such as this, the International Cancer Genome Consortium proposes to generate genome sequences from 500 patients with each of 50 types of cancer in the next decade[109]. For inherited diseases and traits, array-based GWAS alone are insufficient to fully elucidate their genetic basis. Specifically, array based GWAS have limited effectiveness at recombination hotspots (~20% of the genome) or recent or 'private' mutations (which are anticipated to be important in the 'common disease – rare alleles' hypothesis[110]). For example, recent mutations and genetic heterogeneity have been suggested to be important in causality of schizophrenia[111], prompting our group to launch the Schizophrenia Genome Project[112]. This rationale is also driving the EUVADIS project, which proposes to sequence the human genomes of 1,000 European citizens with 10 common disorders[113]. Since 20-fold coverage of the haploid human genome (60gb) still costs ~$100,000 with Generation II sequencing technology, a major current focus of resequencing is on mRNA or comprehensive exon (exome) selection, which comprise approximately 2% of the genome but conceptually enables delineation of the vast majority of biologically relevant nucleotide variation in human cohorts[32,64,71]. Exome sequencing is being pursued in several disorders as part of the NIH medical resequencing programme and by the Psychiatric Genomics Center at Cold Spring Harbor Laboratory[114,115].

The pharmaceutical industry is becoming increasingly involved in these type of projects, particularly in therapeutic areas of greatest interest to individual organisations. Ultimately, Generation II (and emerging Generation III) sequencing is anticipated to play an important role in identification of novel drug targets, establishing genetic profiles related to drug response, confirming pharmacogenetic associations and accelerating the development of new drugs and companion diagnostic tests.

RNA profiling

RNA profiling is a well established and ubiquitous tool for target discovery, target validation and biomarker development. Commonly profiled RNA types are polyadenylated messenger RNA and small RNA molecules (smRNAs). The latter are 20-30 nucleotide molecules that block translation or induce

degradation of target mRNAs. smRNAs include microRNAs (miRNAs) and trans-acting small interfering RNAs (tasiRNAs) that regulate mRNA stability and translation, and small interfering (si) RNAs that cause post-transcriptional gene silencing and are important in cytosine DNA methylation. Gene silencing is important to proper cellular development and proliferation. As an epigenetic modification, RNAi may be reversible and is a promising therapeutic target. The goal of expression profiling experiments is typically to understand the dynamics of transcript network and pathway abundance between states or temporal events. Usually, this involves the identification of sets of transcripts whose expression differs as an external parameter is varied (time, treatment, dose, genotype, etc). The technology specifications for comprehensive gene expression profiling experiments are well established:

- Ability to measure most transcripts simultaneously (requiring a six log10 dynamic range).
- Small mass of RNA sample input (<1µg total RNA).
- High throughput (hundreds of samples).
- Precision (coefficients of variation of <10%).
- Moderate price.

The most prevalent technology for transcript profiling is array hybridisation. It has limitations of closed architecture (features are constrained by available exon sequences at time of array design), ratiometric measurement (rather than absolute mass measurement), limited sensitivity, limited dynamic range and relatively high imprecision. It is becoming clear that the ~25,000 mammalian[107] genes encode more than an order of magnitude greater number of transcripts via alternative initiation, splicing and termination[6-9,82,85,98-103].

Generation II sequencing achieves RNA profiling based on random sampling of molecules within a template mixture and counting the number of reads corresponding to each transcript within a sample[8,9,33-35,37,39-41,45,46,48-50,52,53,56,58]. Thus Generation II sequence-based RNA profiling provides absolute measurement of gene abundance with a sensitivity that is determined by the number of reads generated per sample. An economically practical level of sensitivity is a single channel of Illumina SBS or AB SBL (ie 10 million reads). Generation II sequencing is approximately 2.5 x log10 more sensitive than array hybridisation[120], detecting up to one RNA molecule per 30 cells (Hayashizake et al, personal communication). A variety of Generation II RNA profiling approaches have been described, including random mRNA sequencing, rRNA-depleted total RNA sequencing, smRNA profiling, 3' end tagging, sense- and antisense-5' cap tagging, paired read profiling[8,9,33-35,37,39-41,45,46,48-50,52,53,56,58]. These approaches allow profiling of specific transcript subsets or testing of specific hypotheses. Other potential benefits of Generation II sequencing for RNA profiling include: Sequence verification for each measurement, allowing discrimination of paralogs with high sequence similarity; Detection of all transcripts and isoforms, known and novel; utility in any species, whether sequenced or not; low run-to-run imprecision (~3.5%); absence of interference from abundant transcripts (eg globin); linear dose-response characteristics; decreased dependence on RNA integrity and, extensibility to concomitant detection of nucleotide and structural variation[104,120,121]. The only drawback to Generation II-based RNA profiling is approximately a three-fold increase in cost. As this cost differential decreases, it is anticipated that Generation II sequencing will largely replace array hybridisation as the technology of choice for RNA profiling. RNA profiling is the first Generation II sequencing application which is a mainstream component of pharmaceutical discovery and development and most pharmaceutical companies are evaluating and implementing Generation II sequencing technologies for this application at present. Excitingly, pilot studies are identifying many biologically relevant gene expression differences that are undetectable by array hybridisation. Proofs-of-concept in drug discovery applications are in preparation for publication.

Conclusions

Several Generation II sequencing technologies have become available within the last three years, marketed currently by Roche Applied Science, Illumina, Applied Biosystems Danaher Motion, and Helicos. Generation II technologies are largely distinguishable from Generation I sequencing by enabling a variety of functional genomics applications and by up to 20,000-fold greater cost-effectiveness. Established applications of Generation II sequencing are *de novo* genome and transcript sequencing, resequencing of genes, genomic segments, exons (exome), transcripts (transcriptome) and genomes, digital expression profiling of several classes of transcripts (including mRNA, 3' mRNA end-tags and smRNA) and methylome and chromatin immunoprecipitation sequencing. Currently, Generation II sequencing technologies are having a profound impact on basic and translational research, creating improved understanding of human genome variation and transcriptome complexity, improving catalogs of genes, splice isoforms and functional nucleotide variants and identifying

new molecular mechanisms underpinning disease development and drug response. The pharmaceutical industry is evaluating the utility of Generation II sequencing for digital transcript expression and resequencing related to target discovery and validation and biomarker development, particularly related to oncology and biotherapeutic development.

The advent of molecular cloning and Generation I sequencing created the biotechnology and biopharma industries. Generation II sequencing technologies are becoming established in several applications and enabling mRNA, exome and genome resequencing projects of unparalleled scale. Generation II sequencing is anticipated, in combination with GWAS, to lead to a new level of understanding of molecular mechanisms, molecular staging and subsetting of common diseases and drug responses. The pharmaceutical industry is starting to evaluate the potential impact of these technologies on drug discovery, drug development and personalised medicine. Within the next two years, Generation II sequencing technologies are anticipated to replace array hybridisation as the gold standard technology for RNA profiling. In the next five years, the fruits of these efforts will be novel drug targets for common diseases, identification of molecular biomarkers that inform clinical trials of investigational new drugs and discovery of companion molecular diagnostic tests that allow altered reimbursement of targeted therapies to patient subsets.

In the longer term, Generation III sequencing technologies are anticipated to decrease the cost of genome sequencing to $1,000. Generation III sequencing technologies will allow genome sequencing of 5% of the US population. Personal genome sequences will establish personalised medicine in which medical care will routinely include genome-based inference. In tandem, molecular diagnostic testing is anticipated to continue to be the largest growth segment of the diagnostics industry. Generation II and Generation III sequencing technologies are thus anticipated to become key enabling technologies for drug approvals that reflect segmented efficacy for groups of patients with labels referring to the appropriate companion diagnostic. Such targeted therapies are anticipated to reduce attrition of new, innovative medicines and improve reimbursement.

Acknowledgements

A Deo lumen, ab amicis auxiliam (to God for illumination, to friends for help). This work was partially supported by National Institutes of Health grants N01A000,064 and U01AI066,569, and by National Science Foundation grant 0524,775. DDW

Dr Stephen F. Kingsmore *is President/CEO of NCGR, Santa Fe, NM, a non-profit, research institute, whose mission is to improve human health and nutrition by genome sequencing and analysis. Previously, Dr Kingsmore was Chief Operating Officer of Molecular Staging Inc, and Vice President of Research of CuraGen Corporation, New Haven, CT. Before that, Dr Kingsmore was an Assistant Professor at the University of Florida. He received a BSc in medical microbiology with first class honors in 1982 and MB, ChB, BAO in 1985, both from the Queen's University of Belfast. He completed residency in Internal Medicine, and a fellowship in Rheumatology, at Duke University Medical Center, Durham, NC.*

Dr Jennifer C van Velkinburgh *received her BSc (1997, Presidential Scholar) at St Mary's University and her PhD (2005) from Vanderbilt University in Molecular Physiology and Biophysics. She has been the recipient of NIH, HHMI and DOE fellowship awards for research in diabetes and infectious disease. She currently performs post-doctoral genomic research at NCGR and is a reviewer and editorial board member for several international scientific journals.*

Dr Joann Mudge *is a Senior Scientist at NCGR specialising in bioinformatics and analysis of next generation sequence data. Before that Dr Mudge was a research scientist at the University of Minnesota working on the Medicago truncatula genome project. Previously, she received her PhD at the University of Minnesota in plant genetics in 1999, and her BSc (1993) and MS (1995) degrees in botany, both at Brigham Young University where she graduated cum laude and was a National Merit Scholar.*

Dr Gregory D. May *is Vice President and Genome Center Director of the National Center for Genome Resources, Santa Fe, NM. His research focuses on crop improvement applications of new sequencing technologies. Previously, he was an Associate Scientist and Head of the Genomics Program at the Samuel Roberts Nobel Foundation, Ardmore, Oklahoma. He received a BS in Biology from Southeast Missouri State University in 1987 and received his PhD in Plant Physiology from the Biochemistry and Biophysics Department at Texas A&M University in 1992. He completed his post-doctoral fellowship at the Institute of Bioscience and Technology (IBT), Houston, Texas and has served on the faculty of IBT, the Boyce Thompson Institute, Ithaca, New York and Cornell University.*

References

1 Human Genome Project. [cited; Available from: http://www.ornl.gov/sci/techresources/Human_Genome/home.shtml.

2 Frazer, KA, Ballinger, DG, Cox, DR, Hinds, DA, Stuve, LL, Gibbs, RA et al. A second generation human haplotype map of over 3.1 million SNPs. Nature 2007 Oct 18;449(7164):851-61.

3 Korbel, JO, Urban, AE, Affourtit, JP, Godwin, B, Grubert, F, Simons, JF et al. Paired-end mapping reveals extensive structural variation in the human genome. Science New York, NY 2007 Oct 19;318(5849):420-6.

4 Kunz, M. Genomic signatures for individualized treatment of malignant tumors. Current drug discovery technologies 2008 Mar;5(1):9-14.

5 Doerfler, W. De novo methylation, long-term promoter silencing, methylation patterns in the human genome, and consequences of foreign DNA insertion. Current topics in microbiology and immunology 2006;301:125-75.

6 Birney, E, Stamatoyannopoulos, JA, Dutta, A, Guigo, R, Gingeras, TR, Margulies, EH et al. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature 2007 Jun 14;447(7146):799-816.

7 Kapranov, P, Cheng, J, Dike, S, Nix, DA, Duttagupta, R, Willingham, AT et al. RNA maps reveal new RNA classes and a possible function for pervasive transcription. Science New York, NY 2007 Jun 8;316(5830):1484-8.

8 Lister, R, O'Malley, RC, Tonti-Filippini, J, Gregory, BD, Berry, CC, Millar, AH et al. Highly integrated single-base resolution maps of the epigenome in Arabidopsis. Cell 2008 May 2;133(3):523-36.

9 Nagalakshmi, U, Wang, Z, Waern, K, Shou, C, Raha, D, Gerstein, M et al. The transcriptional landscape of the yeast genome defined by RNA sequencing. Science New York, NY 2008 Jun 6;320(5881):1344-9.

10 Butz, JA, Yan, H, Mikkilineni, V, Edwards, JS. Detection of allelic variations of human gene expression by polymerase colonies. BMC genetics 2004 Feb 16;5:3.

11 Duan, S, Huang, RS, Zhang, W, Bleibel, WK, Roe, CA, Clark, TA et al. Genetic architecture of transcript-level variation in humans. American journal of human genetics 2008 May;82(5):1101-13.

12 Emilsson, V, Thorleifsson, G, Zhang, B, Leonardson, AS, Zink, F, Zhu, J et al. Genetics of gene expression and its effect on disease. Nature 2008 Mar 27;452(7186):423-8.

13 Ge, B, Gurd, S, Gaudin, T, Dore, C, Lepage, P, Harmsen, E et al. Survey of allelic expression using EST mining. Genome research 2005 Nov;15(11):1584-91.

14 Kwan, T, Benovoy, D, Dias, C, Gurd, S, Provencher, C, Beaulieu, P et al. Genome-wide analysis of transcript isoform variation in humans. Nature genetics 2008 Feb;40(2):225-31.

15 Kwan, T, Benovoy, D, Dias, C, Gurd, S, Serre, D, Zuzan, H et al. Heritability of alternative splicing in the human genome. Genome research 2007 Aug;17(8):1210-8.

16 Myers, AJ, Gibbs, JR, Webster, JA, Rohrer, K, Zhao, A, Marlowe, L et al. A survey of genetic human cortical gene expression. Nature genetics 2007 Dec;39(12):1494-9.

17 Pastinen, T, Ge, B, Hudson, TJ. Influence of human genome polymorphism on gene expression. Human molecular genetics 2006 Apr 15;15 Spec No 1:R9-16.

18 Schadt, EE, Molony, C, Chudin, E, Hao, K, Yang, X, Lum, PY et al. Mapping the genetic architecture of gene expression in human liver. PLoS biology 2008 May 6;6(5):e107.

19 Stranger, BE, Forrest, MS, Dunning, M, Ingle, CE, Beazley, C, Thorne, N et al. Relative impact of nucleotide and copy number variation on gene expression phenotypes. Science New York, NY 2007 Feb 9;315(5813):848-53.

20 Stranger, BE, Nica, AC, Forrest, MS, Dimas, A, Bird, CP, Beazley, C et al. Population genomics of human gene expression. Nature genetics 2007 Oct;39(10):1217-24.

21 Million, RP. Impact of genetic diagnostics on drug development strategy. Nature reviews 2006 Jun;5(6):459-62.

22 Phillips, KA, Van Bebber, S, Issa, AM. Diagnostics and biomarker development: priming the pipeline. Nature reviews 2006 Jun;5(6):463-9.

23 Phillips, KA, Van Bebber, SL. Measuring the value of pharmacogenomics. Nature reviews 2005 Jun;4(6):500-9.

24 Roses, A. Pharmacogenetics in Drug Discovery and Development: a translational perspective. Nature reviews 2008;In press.

25 Stoughton, RB, Friend, SH. How molecular profiling could revolutionize drug discovery. Nature reviews 2005 Apr;4(4):345-50.

26 Sanger, F, Nicklen, S, Coulson AR. DNA sequencing with chain-terminating inhibitors. 1977. Biotechnology Reading, Mass 1992;24:104-8.

27 Smith, LM, Sanders, JZ, Kaiser, RJ, Hughes, P, Dodd, C, Connell, CR et al. Fluorescence detection in automated DNA sequence analysis. Nature 1986 Jun 12-18;321(6071):674-9.

28 Tabor, S, Richardson, CC. Selective inactivation of the exonuclease activity of bacteriophage T7 DNA polymerase by in vitro mutagenesis. The Journal of biological chemistry 1989 Apr 15;264(11):6447-58.

29 Fan, JB, Chee, MS, Gunderson, KL. Highly parallel genomic assays. Nat Rev Genet 2006 Aug;7(8):632-44.

30 Shendure, J, Mitra, RD, Varma, C, Church, GM. Advanced sequencing technologies: methods and goals. Nat Rev Genet 2004 May;5(5):335-44.

31 Shendure, JA, Porreca, GJ, Church, GM. Overview of DNA sequencing strategies. Current protocols in molecular biology. Edited by Frederick M Ausubel, et al 2008 Jan;Chapter 7:Unit 7 1.

32 Albert, TJ, Molla, MN, Muzny, DM, Nazareth, L, Wheeler, D, Song, X et al. Direct selection of human genomic loci by microarray hybridization. Nature methods 2007 Nov;4(11):903-5.

33 Bainbridge, MN, Warren, RL, Hirst, M, Romanuik, T, Zeng, T, Go, A et al. Analysis of the prostate cancer cell line LNCaP transcriptome using a sequencing-by-synthesis approach. BMC genomics 2006;7:246.

34 Barbazuk, WB, Emrich, SJ, Chen, HD, Li, L, Schnable, PS. SNP discovery via 454 transcriptome sequencing. Plant J 2007 Sep;51(5):910-8.

35 Berezikov, E, Thuemmler, F, van Laake, LW, Kondova, I, Bontrop, R, Cuppen, E et al. Diversity of microRNAs in human and chimpanzee brain. Nature genetics 2006 Dec;38(12):1375-7.

36 Binladen, J, Gilbert, MT, Bollback, JP, Panitz, F, Bendixen, C, Nielsen, R et al. The use of coded PCR primers enables high-throughput sequencing of multiple homolog amplification products by 454 parallel sequencing. PLoS ONE 2007;2(2):e197.

37 Cheung, F, Haas, BJ, Goldberg, SM, May, GD, Xiao, Y, Town, CD. Sequencing Medicago truncatula expressed sequenced tags using 454 Life Sciences technology. BMC genomics 2006;7:272.

38 Dahl, F, Stenberg, J, Fredriksson, S, Welch, K, Zhang, M, Nilsson, M et al. Multigene amplification and massively parallel sequencing for cancer mutation discovery. Proceedings of the National Academy of Sciences of the United States of America 2007 May 29;104(22):9387-92.

39 Emrich, SJ, Barbazuk, WB, Li, L, Schnable, PS. Gene discovery and annotation using LCM-454 transcriptome sequencing. Genome research 2007 Jan;17(1):69-73.

40 Girard, A, Sachidanandam, R, Hannon, GJ, Carmell, MA. A germline-specific class of small RNAs binds mammalian Piwi proteins. Nature 2006 Jul 13;442(7099):199-202.

41 Gowda, M, Li, H, Alessi, J, Chen, F, Pratt, R, Wang, GL. Robust analysis of 5'-transcript ends (5'-RATE): a novel technique for transcriptome analysis and genome annotation. Nucleic acids research 2006;34(19):e126.

42 Hiller, NL, Janto, B, Hogg, JS, Boissy, R, Yu, S, Powell, E et al. Comparative genomic analyses of seventeen Streptococcus pneumoniae strains: insights into the pneumococcal supragenome. Journal of bacteriology 2007 Nov;189(22):8186-95.

43 Hoffmann, C, Minkah, N, Leipzig, J, Wang, G, Arens, MQ, Tebas, P et al. DNA bar coding and pyrosequencing to identify rare HIV drug resistance mutations. Nucleic acids research 2007;35(13):e91.

44 Hogg, JS, Hu, FZ, Janto, B, Boissy, R, Hayes, J, Keefe, R et al. Characterization and modeling of the Haemophilus influenzae core and supragenomes based on the complete genomic sequences of Rd and 12 clinical nontypeable strains. Genome biology 2007;8(6):R103.

45 Jones-Rhoades, MW, Borevitz, JO, Preuss, D. Genome-wide expression profiling of the Arabidopsis female gametophyte identifies families of small, secreted proteins. PLoS genetics 2007 Oct;3(10):1848-61.

46 Lau, NC, Seto, AG, Kim, J, Kuramochi-Miyagawa, S, Nakano, T, Bartel, DP et al.

Characterization of the piRNA complex from rat testes. Science New York, NY 2006 Jul 21;313(5785):363-7.
47 Margulies, M, Egholm, M, Altman, WE, Attiya, S, Bader, JS, Bemben, LA et al. Genome sequencing in microfabricated high-density picolitre reactors. Nature 2005 Sep 15;437(7057):376-80.
48 Ng, P, Tan, JJ, Ooi, HS, Lee, YL, Chiu, KP, Fullwood, MJ et al. Multiplex sequencing of paired-end ditags (MS-PET): a strategy for the ultra-high-throughput analysis of transcriptomes and genomes. Nucleic acids research 2006;34(12):e84.
49 Nielsen, KL, Hogh, AL, Emmersen, J. DeepSAGE—digital transcriptomics with high sensitivity, simple experimental protocol and multiplexing of samples. Nucleic acids research 2006;34(19):e133.
50 Ohtsu, K, Smith, MB, Emrich, SJ, Borsuk, LA, Zhou, R, Chen, T et al. Global gene expression analysis of the shoot apical meristem of maize (Zea mays L.). Plant J 2007 Nov;52(3):391-404.
51 Poly, F, Read, T, Tribble, DR, Baqar, S, Lorenzo, M, Guerry, P. Genome sequence of a clinical isolate of Campylobacter jejuni from Thailand. Infection and immunity 2007 Jul;75(7):3425-33.
52 Robb, SM, Ross, E, Sanchez Alvarado, A. SmedGD: the Schmidtea mediterranea genome database. Nucleic acids research 2008 Jan;36(Database issue):D599-606.
53 Ruan, Y, Ooi, HS, Choo, SW, Chiu, KP, Zhao, XD, Srinivasan, KG et al. Fusion transcripts and transcribed retrotransposed loci discovered through comprehensive transcriptome analysis using Paired-End diTags (PETs). Genome research 2007 Jun;17(6):828-38.
54 Thomas, RK, Baker, AC, Debiasi, RM, Winckler, W, Laframboise, T, Lin, WM et al. High-throughput oncogene mutation profiling in human cancer. Nature genetics 2007 Mar;39(3):347-51.
55 Thomas, RK, Nickerson, E, Simons, JF, Janne, PA, Tengs, T, Yuza, Y et al. Sensitive mutation detection in heterogeneous cancer specimens by massively parallel picoliter reactor sequencing. Nature medicine 2006 Jul;12(7):852-5.
56 Toth, AL, Varala, K, Newman, TC, Miguez, FE, Hutchison, SK, Willoughby, DA et al. Wasp gene expression supports an evolutionary link between maternal behavior and eusociality. Science New York, NY 2007 Oct 19;318(5849):441-4.
57 Wang, C, Mitsuya, Y, Gharizadeh, B, Ronaghi, M, Shafer, RW. Characterization of mutation spectra with ultra-deep pyrosequencing: application to HIV-1 drug resistance. Genome research 2007 Aug;17(8):1195-201.
58 Weber, AP, Weber, KL, Carr, K, Wilkerson, C, Ohlrogge, JB. Sampling the Arabidopsis transcriptome with massively parallel pyrosequencing. Plant physiology 2007 May;144(1):32-42.
59 Weir, BA, Woo, MS, Getz, G, Perner, S, Ding, L, Beroukhim, R et al. Characterizing the cancer genome in lung adenocarcinoma. Nature 2007 Dec 6;450(7171):893-8.
60 Chen, W, Kalscheuer, V, Tzschach, A, Menzel, C, Ullmann, R, Schulz, MH et al. Mapping translocation breakpoints by next-generation sequencing. Genome research 2008 May 21.
61 Ghildiyal, M, Seitz, H, Horwich, MD, Li, C, Du, T, Lee, S et al. Endogenous siRNAs derived from transposons and mRNAs in Drosophila somatic cells. Science New York, NY 2008 May 23;320(5879):1077-81.
62 Hafner, M, Landgraf, P, Ludwig, J, Rice, A, Ojo, T, Lin, C et al. Identification of microRNAs and other small regulatory RNAs using cDNA library sequencing. Methods (San Diego, Calif 2008 Jan;44(1):3-12.
63 Hillier, LW, Marth, GT, Quinlan, AR, Dooling, D, Fewell, G, Barnett, D et al. Whole-genome sequencing and variant discovery in C. elegans. Nature methods 2008 Feb;5(2):183-8.
64 Hodges, E, Xuan, Z, Balija, V, Kramer, M, Molla, MN, Smith, SW et al. Genome-wide in situ exon capture for selective resequencing. Nature genetics 2007 Dec;39(12):1522-7.
65 Ibarra, I, Erlich, Y, Muthuswamy, SK, Sachidanandam, R, Hannon, GJ. A role for microRNAs in maintenance of mouse mammary epithelial progenitor cells. Genes & development 2007 Dec 15;21(24):3238-43.
66 Mi, S, Cai, T, Hu, Y, Chen, Y, Hodges, E, Ni, F et al. Sorting of small RNAs into Arabidopsis argonaute complexes is directed by the 5' terminal nucleotide. Cell 2008 Apr 4;133(1):116-27.
67 Montgomery, TA, Howell, MD, Cuperus, JT, Li, D, Hansen, JE, Alexander, AL et al. Specificity of ARGONAUTE7-miR390 interaction and dual functionality in TAS3 trans-acting siRNA formation. Cell 2008 Apr 4;133(1):128-41.
68 Morin, RD, O'Connor, MD, Griffith, M, Kuchenbauer, F, Delaney, A, Prabhu, AL et al. Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells. Genome research 2008 Apr;18(4):610-21.
69 Mortazavi, A, Williams, BA, McCue, K, Schaeffer, L, Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nature methods 2008 May 30.
70 Olson, M. Enrichment of super-sized resequencing targets from the human genome. Nature methods 2007 Nov;4(11):891-2.
71 Porreca, GJ, Zhang, K, Li, JB, Xie, B, Austin, D, Vassallo, SL et al. Multiplex amplification of large sets of human exons. Nature methods 2007 Nov;4(11):931-6.
72 Shendure, J, Porreca, GJ, Reppas, NB, Lin, X, McCutcheon, JP, Rosenbaum, AM et al. Accurate multiplex polony sequencing of an evolved bacterial genome. Science New York, NY 2005 Sep 9;309(5741):1728-32.
73 Stark, A, Bushati, N, Jan, CH, Kheradpour, P, Hodges, E, Brennecke, J et al. A single Hox locus in Drosophila produces functional microRNAs from opposite DNA strands. Genes & development 2008 Jan 1;22(1):8-13.
74 Stark, A, Kheradpour, P, Parts, L, Brennecke, J, Hodges, E, Hannon, GJ et al. Systematic discovery and characterization of fly microRNAs using 12 Drosophila genomes. Genome research 2007 Dec;17(12):1865-79.
75 Stark, A, Lin, MF, Kheradpour, P, Pedersen, JS, Parts, L, Carlson, JW et al. Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures. Nature 2007 Nov 8;450(7167):219-32.
76 Tyler, DM, Okamura, K, Chung, WJ, Hagen, JW, Berezikov, E, Hannon, GJ et al. Functionally distinct regulatory RNAs generated by bidirectional transcription and processing of microRNA loci. Genes & development 2008 Jan 1;22(1):26-36.
77 Wakaguri, H, Yamashita, R, Suzuki, Y, Sugano, S, Nakai, K. DBTSS: database of transcription start sites, progress report 2008. Nucleic acids research 2008 Jan;36(Database issue):D97-101.
78 Wilhelm, BT, Marguerat, S, Watt, S, Schubert, F, Wood, V, Goodhead, I et al. Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. Nature 2008 May 18.
79 Wold, B, Myers, RM. Sequence census methods for functional genomics. Nature methods 2008 Jan;5(1):19-21.
80 Harris, TD, Buzby, PR, Babcock, H, Beer, E, Bowers, J, Braslavsky, I et al. Single-molecule DNA sequencing of a viral genome. Science New York, NY 2008 Apr 4;320(5872):106-9.
81 Milos, P. Helicos BioSciences. Pharmacogenomics 2008 Apr;9(4):477-80.
82 Kim, JB, Porreca, GJ, Song, L, Greenway, SC, Gorham, JM, Church, GM et al. Polony multiplex analysis of gene expression (PMAGE) in mouse hypertrophic cardiomyopathy. Science New York, NY 2007 Jun 8;316(5830):1481-4.
83 Lindell, D, Jaffe, JD, Coleman, ML, Futschik, ME, Axmann, IM, Rector, T et al. Genome-wide expression dynamics of a marine virus and host reveal features of co-evolution. Nature 2007 Sep 6;449(7158):83-6.
84 Porreca, GJ, Shendure, J, Church, GM. Polony DNA sequencing. Current protocols in molecular biology. Edited by Frederick M Ausubel et al 2006 Nov;Chapter 7:Unit 7 8.
85 Brenner, S, Johnson, M, Bridgham, J, Golda, G, Lloyd, DH, Johnson, D, Luo, S et al. Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. Nat Biotechnol 2000;18:1021.
86 Cloonan, N, Forrest, AR, Kolle, G, Gardiner, BB, Faulkner, GJ, Brown, MK et al. Stem cell transcriptome profiling via massive-scale mRNA sequencing. Nature methods 2008 May 30.
87 Chaisson, M, Pevzner, P, Tang, H. Fragment assembly with short reads. Bioinformatics (Oxford, England) 2004 Sep 1;20(13):2067-74.
88 Chiu, KP, Wong, CH, Chen, Q, Ariyaratne, P, Ooi, HS, Wei, CL et al. PET-Tool: a software suite for comprehensive processing and managing of Paired-End diTag (PET) sequence data. BMC bioinformatics 2006;7:390.
89 Wu, TD, Watanabe, CK. GMAP: a genomic mapping and alignment program for mRNA and

EST sequences. Bioinformatics (Oxford, England) 2005 May 1;21(9):1859-75.
90 Zerbino, DR, Birney, E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome research 2008 May;18(5):821-9.
91 Risch, N, Merikangas, K. The future of genetic studies of complex human diseases. Science New York, NY 1996 Sep 13;273(5281):1516-7.
92 Edwards, RA, Rodriguez-Brito, B, Wegley, L, Haynes, M, Breitbart, M, Peterson, DM et al. Using pyrosequencing to shed light on deep mine microbial ecology. BMC genomics 2006;7:57.
93 Krause, L, Diaz, NN, Bartels, D, Edwards, RA, Puhler, A, Rohwer, F et al. Finding novel genes in bacterial communities isolated from the environment. Bioinformatics (Oxford, England) 2006 Jul 15;22(14):e281-9.
94 Sogin, ML, Morrison, HG, Huber, JA, Mark Welch, D, Huse, SM, Neal, PR et al. Microbial diversity in the deep sea and the underexplored "rare biosphere". Proceedings of the National Academy of Sciences of the United States of America 2006 Aug 8;103(32):12115-20.
95 Tringe, SG, von Mering, C, Kobayashi, A, Salamov, AA, Chen, K, Chang, HW et al. Comparative metagenomics of microbial communities. Science New York, NY 2005 Apr 22;308(5721):554-7.
96 Kingsmore, SF. Multiplexed protein measurement: technologies and applications of protein and antibody arrays. Nature reviews 2006 Apr;5(4):310-20.
97 Devlin, B, Roeder, K, Bacanu, SA. Unbiased methods for population-based association studies. Genetic epidemiology 2001 Dec;21(4):273-84.
98 Kingsmore, SF, Lindquist, IE, Mudge, J, Gessler, DD, Beavis, WD. Genome-wide association studies: progress and potential for drug discovery and development. Nature reviews 2008 Mar;7(3):221-30.
99 Camargo, AA, Samaia, HP, Dias-Neto, E, Simao, DF, Migotto, IA, Briones, MR et al. The contribution of 700,000 ORF sequence tags to the definition of the human transcriptome. Proceedings of the National Academy of Sciences of the United States of America 2001 Oct 9;98(21):12103-8.
100 Carninci, P, Kasukawa, T, Katayama, S, Gough, J, Frith, MC, Maeda, N et al. The transcriptional landscape of the mammalian genome. Science New York, NY 2005 Sep 2;309(5740):1559-63.
101 Gustincich, S, Sandelin, A, Plessy, C, Katayama, S, Simone, R, Lazarevic, D et al. The complexity of the mammalian transcriptome. The Journal of physiology 2006 Sep 1;575(Pt 2):321-32.
102 Saha, S, Sparks, AB, Rago, C, Akmaev, V, Wang, CJ, Vogelstein, B et al. Using the transcriptome to annotate the genome. Nature biotechnology 2002 May;20(5):508-12.
103 Schadt, EE, Edwards, SW, GuhaThakurta, D, Holder, D, Ying, L, Svetnik, V et al. A comprehensive transcript index of the human genome generated using microarrays and computational approaches. Genome biology 2004;5(10):R73.
104 Scott, H. What transcripts are found in a human cell? Genome biology 2001;1(1):reports031.
105 Sugarbaker, DJ, Richards, WG, Gordon, GJ, Dong, L, De Rienzo, A, Maulik, G et al. Transcriptome sequencing of malignant pleural mesothelioma tumors. Proceedings of the National Academy of Sciences of the United States of America 2008 Mar 4;105(9):3521-6.
106 Walther, DJ, Peter, JU, Bashammakh, S, Hortnagl, H, Voits, M, Fink, H et al. Synthesis of serotonin by a second tryptophan hydroxylase isoform. Science New York, NY 2003 Jan 3;299(5603):76.
107 Zhang, X, Beaulieu, JM, Gainetdinov, RR, Caron, MG. Functional polymorphisms of the brain serotonin synthesizing enzyme tryptophan hydroxylase-2. Cell Mol Life Sci 2006 Jan;63(1):6-11.
108 Kuo, WP, Liu, F, Trimarchi, J, Punzo, C, Lombardi, M, Sarang, J et al. A sequence-oriented comparison of gene expression measurements across different hybridization-based technologies. Nature biotechnology 2006 Jul;24(7):832-40.
109 Scott, H. What transcripts are found in a human cell? Genome biology 2000;1(1):reports031.
110 1000 genomes project. [cited; Available from: http://www.1000genomes.org/page.php?page=home
111 Pritchard, JK, Cox, NJ. The allelic architecture of human disease genes: common disease-common variant...or not? Human molecular genetics 2002 Oct 1;11(20):2417-23.
112 McClellan, JM, Susser, E, King, MC. Schizophrenia: a common disease caused by multiple rare alleles. Br J Psychiatry 2007 Mar;190:194-9.
113 Schizophrenia Genome Project. [cited; Available from: http://psychcentral.com/blog/archives/2006/04/05/schizophrenia-genome-project/
114 EUVADIS. [cited; Available from: http://www.esf.org/activities/eurobiofund/eurobioforum-2008/euvadis.html
115 NIH Medical Resequencing Program. [cited; Available from: http://genome.gov/15014882
116 Cold Spring Harbor Laboratory Psychiatric Genomics Center. [cited; Available from: http://www.cshl.edu/public/releases/07_pgc.html
117 Kim, J-I, Ju, YS, Park, H, Kim, S, Lee, S et al. Completing the human phyletic tetrad using a highly-annotated, whole-genome sequence of an Altaic Asian Individual. Submitted.
118 Wang, J et al. The diploid genome sequence of an Asian individual. Nature 456 (7218), 60-65 (2008).
119 Bentley, DR et al. Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456 (7218), 53-59 (2008).
120 Wheeler, DA et al. The complete genome of an individual by massively parallel DNA sequencing. Nature 452 (7189), 872-876 (2008).
121 Wang, ET, Sandberg, R, Luo, S, Khrebtukova, I, Zhang, L et al. Alternative isoform regulation in human tissue transcriptomes. Nature. 2008 Nov 27;456(7221):470-6.
122 Mudge, J, Miller, NA, Khrebtukova, I, Lindquist, IE, May, GD et al. Genomic convergence analysis of schizophrenia: mRNA sequencing reveals altered synaptic vesicular transport in post-mortem cerebellum. PLoS ONE. 2008;3(11):e3625.
123 Eid, J, Fehr, A, Gray, J, Luong, K, Lyle, J et al. Real-time DNA sequencing from single polymerase molecules. Science. 2009 Jan 2;323(5910):133-8.
124 Harris, TD, Buzby, PR, Babcock, H, Beer, E, Bowers, J et al. Single-molecule DNA sequencing of a viral genome. Science. 2008 Apr 4;320(5872):106-9.
125 Miller, NA, Farmer, AD, Kingsmore, SF, Langley, RJ, Schilkey, FD et al. Management of High-Throughput DNA Sequencing Projects: Alpheus. JCSB 1: 132-148(2008).