DECODING CANCER GENOMES

Nicole Cloonan, Peter Wilson and Sean Grimmond*
Queensland Centre for Medical Genomics, Institute for Molecular Bioscience,
University of Queensland, St Lucia, QLD 4072
*Corresponding author: s.grimmond@uq.edu.au

Introduction

Massively parallel sequencing technology has progressed to the point where entire human genomes can be sequenced rapidly and comprehensively. Large-scale international collaborations are making use of this technology to sequence the genome, epigenome, and transcriptomes of cancers and their matched normal DNA – heralding an era of personalised medical genomics. This article explores the use of 'next-generation' sequencing in the specific context of mutation detection in cancer at single nucleotide resolution.

Cancer as a Genomic Disease

Throughout life, individuals accumulate somatic DNA changes in response to environmental and lifestyle exposure to mutagens. Inherited variations (germline mutations) may also contribute to tumourigenesis, either directly (mutations in oncogenes or tumour suppressors) or indirectly (mutations affecting chromosomal stability or DNA repair, leading to increased mutation rates) (reviewed in (1)).

There are a number of different mutation types that contribute to the onset of cancer (**Fig. 1**). Simple nucleotide variations (SNVs; substitutions and small insertions or deletions) (2,3), copy number variations (CNVs; amplification or loss) (4,5), structural variations (SVs; insertions, deletions, inversions, and translocations) (1,6-8), and epigenomic lesions (changes in DNA methylation or histone modification signatures) (9,10) can all contribute to tumourigenesis, either alone or in combination, and the consequences of these mutations will often be manifested in the transcriptome. SVs can result in fusion genes or place an intact gene under new regulatory elements, leading to dysregulation of gene expression. Epigenomic changes and CNVs, as well as SNVs in regulatory regions, can also lead to gene dysregulation. Activating mutations in oncogenes or loss of function mutations in tumour suppressors will often have secondary transcriptomic consequences, and changes in regulatory RNAs (such as miRNAs) can also dramatically affect gene regulation networks (11).

While some of these variations will contribute to tumourigenesis (driver mutations) and some will not (passenger mutations), the full repertoire of mutations will determine the severity of the disease, the response to therapeutics, and the propensity for metastasis in individual patients (1). The wide variety of possible mutations in all the different cancer types virtually guarantees that every cancer is unique; however, the presence of classifiable subtypes (such as basal vs luminal breast cancer (12)) suggests that there are common pathways affected, and that the identification of common driver mutations is likely to provide new therapeutic targets.

Comprehensive sequencing of all genomic dimensions will ensure the identification of potential oncogenes and tumour suppressors, but studies using hundreds of tumours are required to discriminate between driver and passenger mutations (13). Additionally, due to heterogeneity of human populations, both the tumour and normal genomic dimensions for individual patients need to be determined before we can distinguish between germline and somatic variations. Due to the scale and scope of the problem, funding bodies have invested significant amounts of money into large-scale genome sequencing centres.

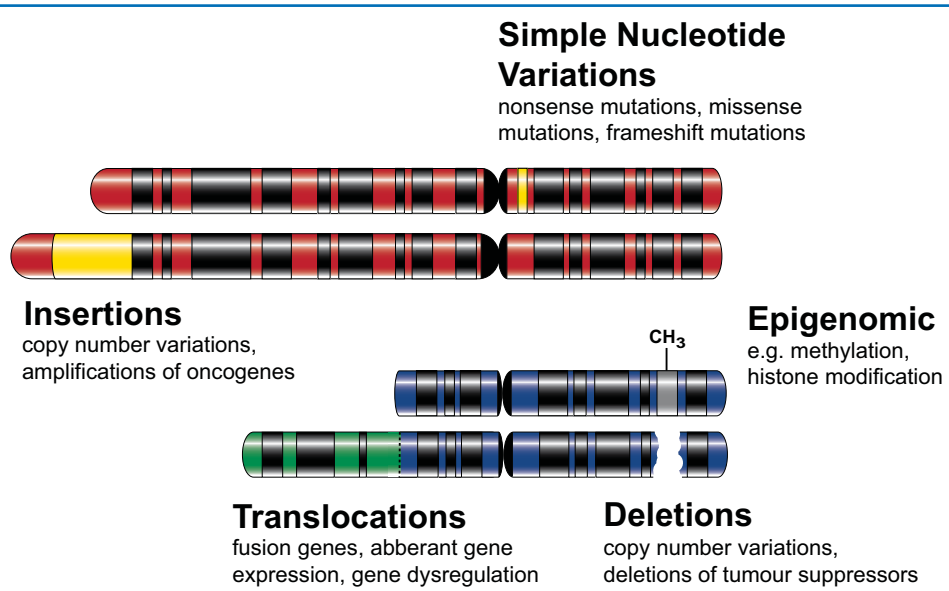

Fig. 1. Schematic depicting the molecular events underlying tumourigenesis and their consequences. Chromosome pairs are shown as the same colour. Although each event may only happen to one allele of the diploid pair, multiple event types can combine to affect a gene homozygously (e.g., DNA methylation silencing one allele, and a deletion event removing the other allele).

Large-Scale Sequencing for Cancer Genomics

The International Cancer Genome Consortium (ICGC; http://www.icgc.org) was launched to coordinate a large number of research projects with the goal of understanding the genetic changes that lead to cancer (reviewed in (13)). The ICGC will produce a catalogue of genomic abnormalities including somatic mutations, abnormal expression of genes and epigenetic modifications in 50 different tumour types. To be part of the consortium, the ICGC has mandated that data must be freely available and not restricted by intellectual property constraints – a policy that will ultimately accelerate cancer research worldwide. Ten countries and two European consortia have so far initiated 21 cancer genome projects under the umbrella of the ICGC.

In addition to the ICGC, there is also The Cancer Genome Atlas (TCGA; http://tcga.cancer.gov/) project – a comprehensive program in cancer genomics funded by the National Cancer Institute and the National Human Genome Research Institute (USA). The pilot work for TCGA focused on glioblastoma multiforme, serous ovarian cancer, and lung squamous carcinoma. Further investment of $275 million means that TCGA can now expand their effort to examine ten additional tumour types over the next two years. TCGA has played a role in launching the ICGC effort and has been involved in aligning the outputs of the two efforts so that data generation, data analysis and interpretation will be comparable between the two groups.

Large-Scale Sequencing Initiatives in Australia

In 2009, Australia commenced an ICGC program with the announcement of large-scale cancer genome sequencing projects into pancreatic and ovarian cancer (a total of 500 patients). Pancreatic cancer is the fourth most common cause of cancer death in the developed world, with the average patient living for less than six months after diagnosis. Ovarian tumours occur less frequently, but still rank sixth as the most common cause of death from cancer in developed countries. The program is based at the Queensland Centre for Medical Genomics (http://www.qcmg.org) at the Institute for Molecular Bioscience, University of Queensland. The pancreatic cancer project is carried out in partnership with the Garvan Institute of Medical Research, Ontario Institute of Cancer Research and John Hopkins University. The ovarian cancer project is performed in collaboration with the Peter MacCallum Cancer Centre. The program seeks to sequence both normal and tumour genomic DNA and RNA from all 500 patients and will map out tumour-associated changes to the genome, transcriptome and methylome. This project has seen the installation of a large fleet of genome sequencers (currently 11 version 4 SOLiD sequencers), giving the capacity to decode up to 600 billion nucleotides of DNA sequence every week.

Cancer Sequencing Strategies

Although each strategy for sequencing differs in the specifics, conceptually the protocols are highly similar: (i) fragment the nucleic acids to the appropriate size; (ii) capture the fragments between adaptors of a defined sequence; (iii) clonally amplify single molecules onto a template; and (iv) sequence the library using the known sequence in the adaptors. Different library-making strategies in the context of cancer mutation detection are discussed below.

Sequencing the Genome

By far the simplest library preparation protocol is a 'fragment' library, where the DNA is sheared to slightly larger than the tag read length (50–100bp). This strategy provides the best input DNA to sequenced tag ratio, can be used to determine simple nucleotide variations and copy number variations, and is often used to make libraries from 'exome' microarray-captured DNA. *De novo* detection of structural variations using fragment data alone is difficult and computationally intensive, although once breakpoints have been identified, fragment data can be aligned to these sequences directly.

'Paired-end' libraries are fragment-style libraries where short tags are read from a single fragment (typically 200–300nt long). A critical step in this protocol is size-selecting the DNA fragments, so that the physical distance between the paired tags is known. Once the pairs are aligned to a reference genome, the distance between the pairs can be compared with the expected distance between the pairs; discordant distances identify structural variations (**Fig. 2**). The vast majority of genome re-sequencing data comes from paired-end libraries, due to their relatively low requirement for input DNA and the simplicity of the library preparation.

The 'mate-pair' strategy (also known as 'di-tag') is less commonly used for large-scale genomic sequencing, but offers greater sensitivity to detect structural variations. The protocol involves breaking the DNA into large fragments (often between 1–10kb, allowing tag pairs to straddle repetitive regions) and, rather than sequence the entire fragment, each end is captured and sequenced as a pair of short tags. Although sensitive, this protocol is more involved and requires significantly more DNA – typically 10–50X more than paired-end libraries, depending on the sizes involved.

Sequencing the Transcriptome

While genome sequencing will provide comprehensive information about the mutations present in the genome, their impact on the transcriptome is best studied as another sequencing dimension. Whole transcriptome RNAseq can provide accurate and sensitive gene expression data, but also information on novel exons, expressed mutations, alternative splicing, and fusion gene identification (**Fig. 3**) (14). RNAseq data can be prepared as fragment or paired-end libraries, analogous to those described above – and the identification of exons and introns parallels the identification of deletions in genomic DNA. Fragment library sequencing can easily detect known exon-exon junction sequences, but the sensitivity for novel events is low. Paired-end libraries are far more sensitive, as they do not rely on a tag to cross an exon-exon boundary and can also detect the relationship between novel exonic sequence and the transcriptional framework. Once the framework is assembled, the presence of individual transcripts can be

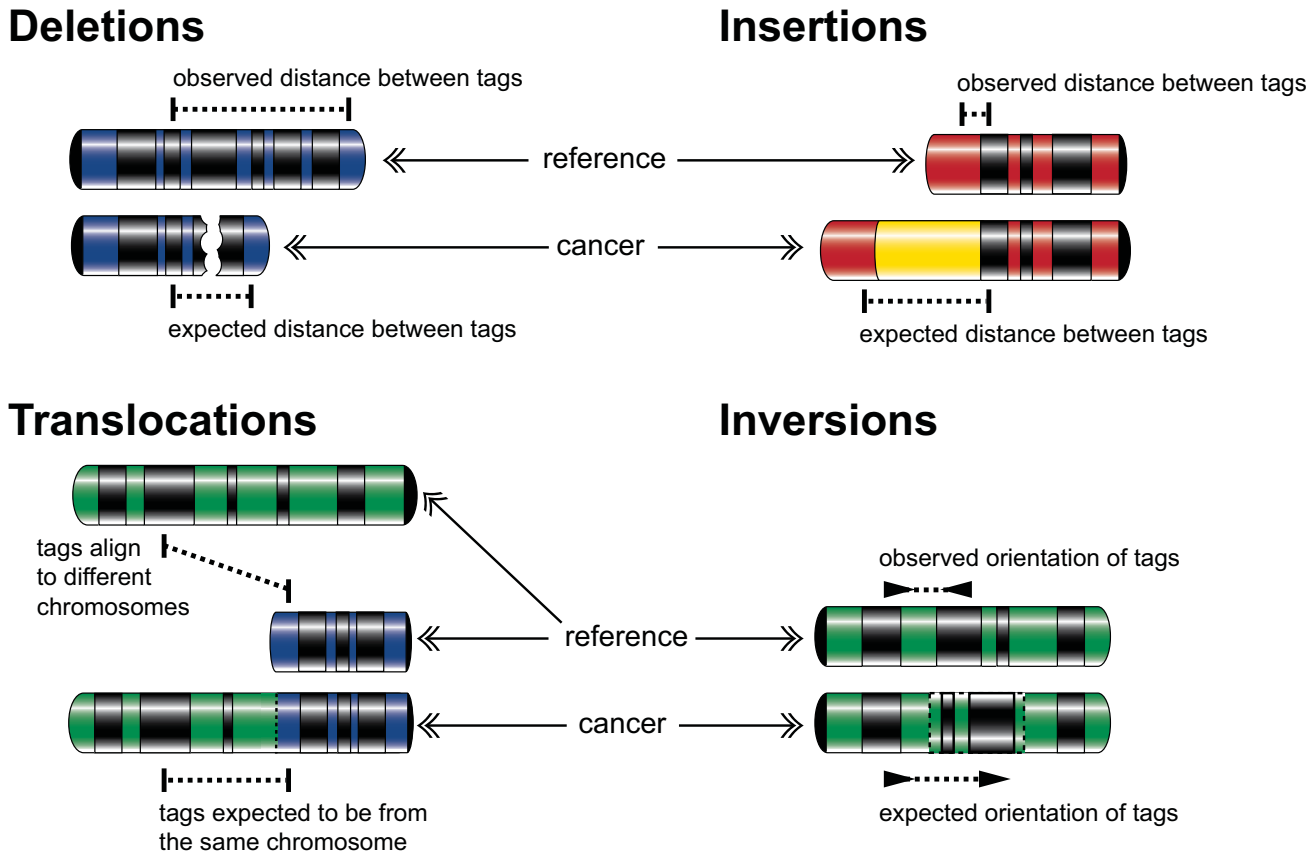

Fig. 2. Detecting structural variations using mate-pair sequencing. Tags from 'paired-end' or 'mate-pair' libraries are aligned to a reference genome, and the distance between the two tags is used to detect specific structural variants. Where the alignment distance is concordant with the size of the DNA fragments, no structural variation can be inferred. Discordant mapping distances and orientations between the two tags can identify specific molecular events. **Deletions** are detected when the observed alignment distance for paired-end reads is much larger than what would be expected from the library. **Insertions** can be inferred when the observed alignment difference is much smaller than the expected alignment distance. **Translocations**: Inter-chromosomal translocations can be detected when the paired ends of a DNA fragment map to different chromosomes. Intra-chromosomal events are often seen as pairs of insertion and deletion events. **Inversions** are identified when the orientation of the observed tags is different to the expected orientation, and depending on the size of the inversion, may also be coupled with discordant pairs detecting insertion and deletion events.

modelled and quantified (15,16). Short-tag sequencing is ideally suited to detecting microRNAs (~22nt) and other small RNA populations, as these molecules do not need to be fragmented prior to library preparation, and miRNA counts translate directly to gene expression levels.

Sequencing the Epigenome

There are a number of strategies for capturing epigenomic signatures. Antibody-based capture protocols, including ChIP (Chromatin ImmunoPrecipitation) and MeDIP (Methylated DNA ImmunoPrecipitation), are easily combined with fragment library-making strategies described above. Similarly, capture using methyl-CpG binding domains (MBDs) is also well suited for traditional fragment library sequencing; however, neither MeDIP nor MDB strategies are able to precisely determine the methylation status of each cytosine within the capture fragment. Bisulfite sequencing (a protocol that converts unmethylated cytosines to uracils) provides this resolution, but is difficult to combine with short-tag sequencing due to the ambiguity in aligning bisulfite-converted short tags.

The Future of Cancer Sequencing

The 'Holy Grail' of cancer genome sequencing is to summarise all somatic aberrations in a patient's tumour and use these data to improve cancer diagnosis and instruct optimal therapeutic intervention. In order to reach these goals, genome sequencing technology and the computational approaches used to detect mutations and put them in a biological context need to progress far beyond what is currently in use.

The prospect of technology improvements is very good, with all current sequencing platforms continuing to drive higher throughputs, longer reads, and improved accuracy from their current approaches. It is predicted that throughputs will reach 200–300 gigabases per run by the end of 2010, making cancer genome sequencing more affordable. The sequencing community is also eagerly awaiting the commercial release of single molecule sequencing (SMS) technology. Early access users have demonstrated that with SMS, it is possible to generate sequence reads in excess of 1000bp, in real time, from minute amounts of template. It is anticipated that the likes

of SMS will complement current sequencing approaches in the short to intermediate term, but should ultimately overtake current sequencing platforms as the technology of choice.

Improving the computational aspects of genome re-sequencing and interpreting the data is far more challenging. To make sense of what is actively driving individual tumours, it is clear that more than just the DNA needs to be interrogated, and these data need to be interpreted in terms of how perturbations in each '-ome' impact on important biological pathways. These monumental tasks are being actively pursued by the ICGC and TCGA communities. Their focus on sharing both methodology and genome data will be critical in advancing cancer research and the prospect of personal genomics in the years to come.

Conclusions

Even in its pioneering stages, genome sequencing is driving a revolution in medical research. The efforts of the ICGC have recently demonstrated that systematic re-sequencing of cancer genomes is feasible, and they have created the framework for the international research community to contribute to large-scale systematic analysis of common cancers. Over the next five years, more than 25,000 cancer patients will have their tumour and matched normal DNAs and RNAs sequenced in their entirety by ICGC members. These data will be accessible to the entire research community and stand to redefine how cancer research will be performed for the decade to come. The ICGC's co-ordinated international effort into cancer genome informatics also stands to propel this research forward faster than any single laboratory or country could achieve on its own and is paving the way for more clinical applications of this technology in the years to come.

References

1. Stratton, M.R., Campbell, P.J., and Futreal, P.A. (2009) *Nature* **458**, 719-724
2. Frohling, S., Scholl, C., Levine, R.L., *et al.* (2007) *Cancer Cell* **12**, 501-513
3. Greenman, C., Stephens, P., Smith, R., *et al.* (2007) *Nature* **446**, 153-158
4. Beroukhim, R., Mermel, C.H., Porter, D., *et al.* (2010) *Nature* **463**, 899-905
5. Pinkel, D., and Albertson, D.G. (2005) *Nat. Genet.* **37 Suppl**, S11-17
6. Futreal, P.A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., Rahman, N., and Stratton, M.R. (2004) *Nat. Rev. Cancer* **4**, 177-183
7. Mitelman, F., Johansson, B., and Mertens, F. (2007) *Nat. Rev. Cancer* **7**, 233-245
8. Stephens, P.J., McBride, D.J., Lin, M.L., *et al.* (2009) *Nature* **462**, 1005-1010
9. Baylin, S.B., and Ohm, J.E. (2006) *Nat. Rev. Cancer* **6**, 107-116
10. Esteller, M. (2007) *Nat. Rev. Cancer* **8**, 286-298
11. Olson, P., Lu, J., Zhang, H., Shai, A., Chun, M.G., Wang, Y., Libutti, S.K., Nakakura, E.K., Golub, T.R., and Hanahan, D. (2009) *Genes Dev.* **23**, 2152-2165
12. Perou, C.M., Sorlie, T., Eisen, M.B., *et al.* (2000) *Nature* **406**, 747-752
13. The International Cancer Genome Consortium (2010) *Nature* **464**, 993-998
14. Cloonan, N., and Grimmond, S.M. (2008) *Genome Biol.* **9**, 234
15. Guttman, M., Garber, M., Levin, J.Z., Donaghey, J., Robinson, J., Adiconis, X., Fan, L., Koziol, M.J., Gnirke, A., Nusbaum, C., Rinn, J.L., Lander, E.S., and Regev, A. *Nat. Biotechnol.* **28**, 503-510
16. Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J., and Pachter, L. *Nat. Biotechnol.* **28**, 511-515

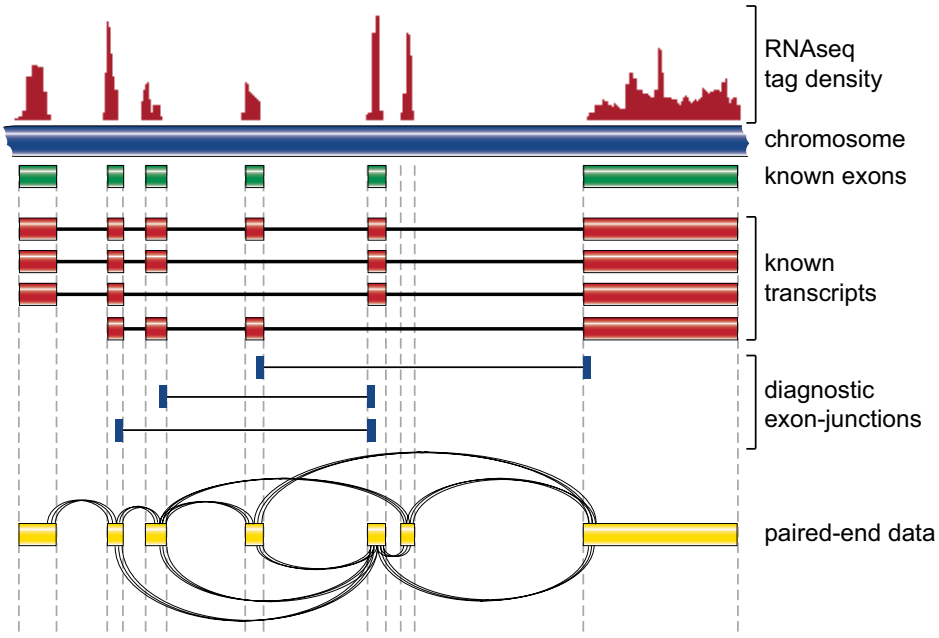

Fig. 3. A schematic diagram examining hypothetical RNAseq data in a genomic context. Both paired-end and fragment-based data can contribute to coverage or 'wiggle' plots of tag density, identifying both known and novel exons and their level of expression. For fragment libraries, tags that span exon-exon junctions are matched to known exon combinations to infer the transcripts used by an individual locus. Paired-end data have higher sensitivity to detect exon combinations, as the individual tags from each pair can reside anywhere within an exon, and not just across a boundary. Paired-end RNAseq also has the advantage of being able to link novel exonic sequence into the transcriptomic framework used by an exon. The dashed lines are visual guides for alignment.