REVIEW ARTICLE

Whole genome sequencing for lung cancer

Marissa Daniels[1,2], Felicia Goh[1,2], Casey M Wright[1,2], Krishna B Sriram[1,2], Vandana Relan[1,2], Belinda E Clarke[3], Edwina E Duhig[3], Rayleen V Bowman[1,2], Ian A Yang[1,2], Kwun M Fong[1,2]

[1]Thoracic Research Laboratory, Department of Thoracic Medicine, The Prince Charles Hospital, Brisbane, Australia; [2]UQ Thoracic Research Centre, School of Medicine, The University of Queensland, Brisbane, Australia; [3]Pathology Queensland, Department of Anatomical Pathology, The Prince Charles Hospital; Brisbane, Australia

ABSTRACT

Lung cancer is a leading cause of cancer related morbidity and mortality globally, and carries a dismal prognosis. Improved understanding of the biology of cancer is required to improve patient outcomes. Next-generation sequencing (NGS) is a powerful tool for whole genome characterisation, enabling comprehensive examination of somatic mutations that drive oncogenesis. Most NGS methods are based on polymerase chain reaction (PCR) amplification of platform-specific DNA fragment libraries, which are then sequenced. These techniques are well suited to high-throughput sequencing and are able to detect the full spectrum of genomic changes present in cancer. However, they require considerable investments in time, laboratory infrastructure, computational analysis and bioinformatic support. Next-generation sequencing has been applied to studies of the whole genome, exome, transcriptome and epigenome, and is changing the paradigm of lung cancer research and patient care. The results of this new technology will transform current knowledge of oncogenic pathways and provide molecular targets of use in the diagnosis and treatment of cancer. Somatic mutations in lung cancer have already been identified by NGS, and large scale genomic studies are underway. Personalised treatment strategies will improve care for those likely to benefit from available therapies, while sparing others the expense and morbidity of futile intervention. Organisational, computational and bioinformatic challenges of NGS are driving technological advances as well as raising ethical issues relating to informed consent and data release. Differentiation between driver and passenger mutations requires careful interpretation of sequencing data. Challenges in the interpretation of results arise from the types of specimens used for DNA extraction, sample processing techniques and tumour content. Tumour heterogeneity can reduce power to detect mutations implicated in oncogenesis. Next-generation sequencing will facilitate investigation of the biological and clinical implications of such variation. These techniques can now be applied to single cells and free circulating DNA, and possibly in the future to DNA obtained from body fluids and from subpopulations of tumour. As costs reduce, and speed and processing accuracy increase, NGS technology will become increasingly accessible to researchers and clinicians, with the ultimate goal of improving the care of patients with lung cancer.

KEY WORDS

High-throughput nucleotide sequencing; DNA sequence analysis; lung neoplasms; non-small cell lung carcinoma; small cell lung carcinoma

J Thorac Dis 2012;4(2):155-163. DOI: 10.3978/j.issn.2072-1439.2012.02.01

Introduction

Lung cancer is the most common cause of cancer death globally (1,2), and was responsible for 1.4 million deaths in 2008 (3). Five year survival rates remain around 15% (1,2), with the majority of cases at an advanced stage at the time of diagnosis (1). Improved understanding of the pathogenesis of lung cancer is required to aid timely diagnosis, selection of cancer treatment and development of new therapeutic modalities in order to improve patient outcomes. The search for the genomic basis of cancer has expanded exponentially since the introduction of DNA sequencing techniques in the late 1970s (4,5) and

No potential conflict of interest.

Corresponding to: Marissa Daniels. Department of Thoracic Medicine, The Prince Charles Hospital, Rode Rd, Chermside 4032, Australia. Tel: +61 7 3139 4000; Fax: +61 7 3139 4510. Email: M.Daniels@uq.edu.au.

Submitted Jan 2, 2012. Accepted for publication Feb 01, 2012.
Available at www.jthoracdis.com

ISSN: 2072-1439 ©

Table 1. Next-generation sequencing techniques available for whole genome sequencing.

Read length (bases)/ read time	Data processing capacity	Advantages	Disadvantages	Primary applications
Illumina HiSeq/Genome Analyser (22): optical detection of fluorescence-labelled nucleotides.				
Up to 100/1.5-11 days	55 Gb/day*	Extensive international experience	Relatively low multiplex capacity	Whole genome-, exome-, and transcriptome sequencing; single nucleotide polymorphism detection; epigenetic studies
454 FLX Pyrosequencer (23): optical detection of fluorescence triggered by pyrophosphate release during base incorporation.				
Up to 1000/10-23 hours	700 Mb/genome	Speed, read length useful for confirmatory sequencing	Errors in homopolymer repeats	Whole genome-, and transcriptome sequencing; targeted resequencing
SOLiD™ (24): optical detection of fluorescence-labelled nucleotide octamers.				
Up to 75/up to 7 days	7-9 Gb/day	Improved accuracy with two base encoding	Slow processing of short sequences	Whole genome and exome sequencing
Ion Torrent (25): semiconductor detection of hydrogen ions released during nucleotide incorporation.				
Up to 200/up to 4.5 hours	1 Gb/genome	Speed, potential for technical improvements	Error rates	Targeted sequencing

*For a dual flow cell system performing 100 base pair, paired end read.

identification of the first naturally occurring human cancer-causing somatic mutation in the early 1980s (6,7). Following these discoveries, it quickly became apparent that cancer gene discovery could either continue in a piecemeal fashion or the whole cancer genome could be sequenced (8). In 2004, further technological advances led to sequencing of the first, 99.7% complete, human genome (9,10). The multi-centre, resource-intensive Human Genome Project took many years to complete using first-generation sequencing techniques, at an estimated cost of $10-25 million, and generated much debate about the efficiency of existing technology (11).

Substantial advances made since 2004 have had a major impact on our ability to explore the genome. Traditional techniques such as capillary electrophoresis-based DNA sequencing, genome-wide analysis of amplifications and deletions using array-based technology, and gene expression arrays have been used to identify genomic drivers of lung cancer (12-14) that can be targeted by directed therapies (12,15). However, with second-generation or next-generation sequencing (NGS) technology, it is now possible to sequence complete genomes (16), exomes (17) and transcriptomes (18). These technologies are revolutionising how we explore the genome and exponentially increasing scope for investigation of cancer pathogenesis, diagnosis and treatment (19-21).

Next-generation sequencing techniques

Next-generation DNA technology has been commercially available since 2004. These massively parallel sequencing platforms share many advantages over traditional techniques. Sample preparation for sequencing is streamlined, and the yield of sequence reads is considerably greater than that possible using capillary sequencers. Each method requires significant investments in laboratory infrastructure and computational support, but these massively parallel techniques provide exponentially greater sequencing capabilities than first-generation technology (19).

Many available sequencing techniques share conceptual similarities. The Roche/454 Pyrosequencer, Illumina HiSeq and Genome Analyser, and Applied Biosystems SOLiD™ require a DNA fragment library, obtained by annealing platform specific linkers to DNA fragments generated from the genome of interest. Each strand in a fragment library is amplified by PCR prior to performing sequencing reactions on the amplified strands (19). The utility of NGS platforms for various applications depends on the nature of DNA reads obtained, read time and error rates, cost, data processing capacity and computational requirements for analysis (Table 1).

Driven by the potential to translate biological discovery into improvements in patient care, NGS techniques are constantly improving and expanding. Newly developed methods include the non-optical Ion Torrent platform (26), and combinatorial probe-anchor ligation sequencing, in which fluorescence-labelled nucleotides are incorporated into sequences generated from template DNA that has been aggregated into nanoballs (27).

Illumina hiSeq and genome analyser systems

The massively parallel platform ultimately acquired by Illumina after a sequence of company mergers was based on the second

or next-generation DNA sequencing technique. After the creation of a DNA fragment library from the target genome sequence, both ends of each fragment are ligated to an Illumina specific adaptor. A flow cell containing eight individual lanes populated with capture oligonucleotide anchors hybridises the modified DNA fragments. DNA amplification then takes place by bridge amplification, in which the DNA template arches over and hybridises to an adjacent oligonucleotide primer anchor complementary to the adaptor sequence attached to the template DNA. After amplification, sequencing commences with the addition of DNA polymerase and a mixture of four differently coloured fluorescent reversible dye terminators in the flow cell. DNA fragments are extended one nucleotide at a time. Digital images are acquired after each addition to record the location of the labelled nucleotide. Finally, dye and terminal 3' blockers are removed prior to the next cycle of nucleotide coupling (19,21).

Although subject to nucleotide incorporation errors and imperfect polymerase activity, advances in bioinformatic analysis techniques have improved the accuracy of sequence reads (28,29). Illumina provides the most widely used NGS technology (21), and this has found application in whole genome (30), and whole exome sequencing (31) including The Cancer Genome Project (32).

454 FLX pyrosequencer

This was the first massively parallel DNA sequencing technique to be introduced commercially. DNA library fragments are attached to 454-specific adaptor sequences, then mixed with agarose beads carrying oligonucleotides complementary to the adapter sequence. Each DNA fragment anneals to one bead, and individual fragment: bead complexes are placed in an oil: water micelle containing PCR reactants. Using thermal cycling, one million copies of each DNA fragment are formed on each bead. Beads are then placed in a picotitre plate and each remains in a fixed location for sequencing. After addition of enzymes to catalyse the pyrosequencing reactions, pure single nucleotide solutions are sequentially introduced. Pyrophosphate released by DNA polymerase at each nucleotide incorporation initiates a series of reactions ultimately producing an amount of light proportional to the number of nucleotides incorporated. Image acquisition after each nucleotide incorporation step permits analysis of the DNA sequence (19,33).

Pyrosequencing has been used in the detection of single nucleotide polymorphisms (34) and targeted capture sequencing (35). Most inaccuracies are due to limitations of the PCR technique used for DNA amplification, and errors in sequencing homopolymer repeats greater than seven bases in length (33). However, this technique is able to generate longer reads in less time than other NGS, and improvements in experimental and analytical techniques are providing ways to improve sequence accuracy (33).

SOLiD™ sequencing

Supported Oligonucleotide Ligation and Detection relies on PCR amplification of template DNA fragments on the surface of 1μm magnetic beads. The fragments are deposited onto a flow cell slide, with two flow cells per instrument run, each able to contain four DNA libraries. Primer is annealed to the adaptor sequences on each amplified DNA fragment. Specific fluorescent octamers whose fourth and fifth bases are encoded by an attached fluorescent label are then incorporated into the DNA sequence by DNA ligase. Di-base probes interrogate every first and second base in each ligation reaction. The ligation is followed by fluorescence detection, removal of the fluorescent group, then a further round of ligation. After a series of five ligation cycles, the extension product is removed and prepared for further ligation cycles with a primer complementary to the n-1 position (19,21,36).

SOLiD™ technology has similar applications to the Illumina platform. While base calling errors are reduced by two base encoding, this is at the expense of test speed and analytical simplicity (21,36).

Ion torrent

This technique eliminates the complexity of the optical detection systems of many NGS methods. DNA polymerase sequencing with unmodified nucleotides occurs, and hydrogen ions released during each cycle of polymerisation are detected using a semiconductor (26).

As this technique has low capacity for parallel sequencing, it is primarily used for sequencing short genomic segments of interest. Although there have been issues with error rates, this is the fastest and lowest cost NGS available (21), and sequencing capacity using this platform is increasing exponentially.

Next-generation sequencing *vs.* sanger sequencing

Next-generation sequencing techniques provide many potential advantages over traditional first-generation sequencing techniques such as Sanger sequencing. However, Sanger sequencing remains the method of choice for detecting genetic changes in a patient's genome for the purposes of guiding therapy, both in lung cancer (37) and other malignancies (38). Sanger sequencing is used to detect *EGFR* mutations in non-small cell lung cancer (NSCLC) patients to select those who may benefit from targeted therapy (37). Similarly, fluorescence in-situ hybridisation has been used to detect *EML4-ALK* translocation in trials of Crizotinib in lung cancer (15).

Sanger's chain termination method of DNA sequencing (4) is based on automatic detection of fluorescence-labelled nucleotide

Table 2. Next-generation sequencing of lung cancer.

First Author	Year	Genomes	Platform	Main Findings
Campbell (30)	2008	2 (SCLC, NE) cell lines	Illumina†	103 somatic rearrangements
Lee (27)	2010	2 (paired NSCLC/normal lung) tissue samples	cPAL; WGS	>50,000 single nucleotide polymorphisms; 392 in coding regions. 43 structural changes.
Pleasance (43)	2010	1 SCLC cell line	SOLiD™; WGS	22,910 somatic mutations; 134 in coding regions. 58 structural changes; 18 deletions, 9 tandem duplications. Tobacco-associated mutation signatures.
Ju (44)	2011	2 (paired liver metastasis from NSCLC/normal lung) tissue samples	Illumina; WGS, WTS	10,724 single nucleotide variations; 334 in coding regions. 52 fusion genes; novel *KIF5B*: RET proto-oncogene.

†Structural variants only sequenced. SCLC: Small-cell lung cancer; NE: Neuroendocrine cell carcinoma; cPAL: combinatorial probe anchor ligation; WGS: whole genome sequencing; WTS: whole transcriptome sequencing.

sequences. DNA elongation occurs along single-stranded DNA templates and is randomly terminated by incorporation of fluorescent dideoxynucleotide chain terminators. DNA sequences of increasing length are detected by capillary electrophoresis. This technique is unable to detect structural changes such as translocations, or gene copy number changes, and multiplexing is difficult and costly. On the other hand, NGS is well suited to large scale gene sequencing and can provide this at lower cost per base than traditional techniques (21). Illumina and SOLiD™ techniques are able to yield tens of millions of reads, and the Roche/454 platform several hundred thousand, compared with the 96 reads produced by a capillary sequencer run (19).

While cost per base is lower for massively parallel sequencing, the cost per test remains greater than for traditional methods. Longer run times are required for NGS due to the need for more frequent image acquisition than Sanger sequencing, and reads are much shorter (21). Platform specific investments in laboratory information management, computational analysis and bioinformatic support are required to produce and interpret final sequence reads, taking into account the unique error model of each platform (19).

Next genome sequencing is necessitating a paradigm shift in the organisation required for genomic sequencing and the information technology and laboratory systems required to support it. Application of these techniques to the study of not only the whole genome, but also exomes (39), transcriptomes (40), and epigenetics (41) will extend their scope for scientific discovery (19).

Next-generation sequencing in lung cancer

Prior to the introduction of NGS, candidate gene studies (12,42) had begun to provide insight into the genomic drivers of lung cancer, such as mutations in *BRAF* (42) and *EGFR* (37). The introduction of massively parallel sequencing, however, has seen further discoveries of somatic mutations in lung cancer, with scope for an exponential increase in understanding of oncogenesis (Table 2).

The first application of massively parallel sequencing in lung cancer research was published in 2008. Examination of genomes from two lung cancer cell lines characterised 306 germline structural changes and 103 somatic rearrangements with single base resolution, including four associated with changes in gene expression (30). In 2010, Lee *et al*. published whole genome studies of tumour and normal lung specimens obtained from a patient with an adenocarcinoma. A large set of new somatic mutations were discovered, at a rate of around 17.7 per megabase of template DNA sequenced. Mutations were located in non-expressed genes and promoter regions up to five kilobases upstream of coding proteins rather than within expressed genes. Five hundred and thirty somatic single nucleotide variants were found, including one in the *KRAS* proto-oncogene, and 391 in other coding regions. Forty-three structural variations were detected (27).

Pleasance *et al*. published the results of massively parallel sequencing of a small cell lung cancer cell line the same year. Of the 22,910 somatic mutations found, 134 were detected in coding regions. A tandem duplication of exons 3-8 of *CHD7* in frame was identified, and mutation signatures associated with tobacco exposure were documented (43).

Most recently, examination of paired NSCLC and normal lung tissue from a never smoking patient with adenocarcinoma by Ju *et al*. led to the discovery of a novel fusion gene, *KIF5B-RET*. Correlation between whole genome and transcriptome sequences from this tumour demonstrated overexpression of chimeric *RET* receptor tyrosine as well as the chromosomal

inversion (44).

These results, including the identification of single nucleotide mutations as well as large structural changes, testify to the utility of NGS in comprehensively characterising the genomic changes present in cancers. However, most mutations present in tumour DNA are merely passengers and do not contribute to oncogenesis (45). Successful identification of driver mutations amongst clusters of random passenger mutations requires the power of large sample sizes.

The cancer genome atlas

An international effort, led by the International Cancer Genome Consortium (ICGC), is underway to comprehensively catalogue the genomic and epigenomic changes in cancer (46). The ICGC is coordinating cancer genome studies in 50 different cancer types that are of global clinical and societal importance. A contributor to this project, The Cancer Genome Atlas (TCGA), is a multinational, collaborative project that seeks to apply NGS technology to further scientific knowledge of the biology of cancer, with the aim of improving cancer care (47). It is coordinated by a consortium between the National Cancer Institute, the National Human Genome Research Institute and several international specimen source centres and analytical sites across the globe. More than 20 cancer types have been selected for study by the TCGA to contribute to the ICGC, with up to 500 samples to be collected for each tumour type, including lung adenocarcinoma and squamous cell carcinomas.

This large scale project requires the coordinated efforts of several specialised TCGA research network components. To ensure nucleotide sequencing of the highest quality, paired tumour and normal control samples are examined by the Biospecimen Core Resource laboratories to confirm specimens meet stringent standards for tumour quality and quantity. Extracted DNA and RNA are sequenced, the results analysed and genome changes then interpreted in the context of clinical data at Genome Characterisation, Sequencing, and Data Coordinating Centres. Data are made publically available to the international community to facilitate further data analysis and validation of discoveries (47).

The Cancer Genome Project has provided impetus for advances in experimental, computational analysis and bioinformatic research methods to ensure efficient processing of the large number of samples required and data obtained (48).

Clinical applications of next-generation sequencing

The ultimate goal of elucidating the mechanisms of cancer pathogenesis is to improve strategies for diagnosis, prognostication and treatment of patients with cancer. Detailed molecular subtyping at the time of diagnosis permits selection of personalised therapies for patients who are most likely to benefit. This benefits not only the individual receiving treatment but spares others, and the community, the cost and morbidity of futile intervention (49,50). Next-generation sequencing may also be incorporated into patient selection algorithms for clinical trials. Improved techniques for tumour categorisation in clinical trials will increase statistical power to detect clinically important results (50).

A thorough understanding of the molecular drivers of cancer may ultimately allow prediction of an individual's risk of developing cancer on the basis of their constitutional genome sequence. The presence of genetic variants and non-genetic contributors to cancer phenotypes will limit the precision of such prediction. Nevertheless, discovery of mutations and chromosomal changes predictive of disease may permit risk reduction and prevention strategies for some cancers (50).

The widespread application of NGS, both in research and clinical settings, raises many practical and ethical issues. Experiments are expensive and the results, to some extent, unpredictable. Institutional ethical approval, patient consent, intellectual property and data release all require careful consideration. Given the costly and demanding nature of NGS, coordinated applications such as TCGA are critical in order to maximise use of resources (51).

Issues for clinical translation

Sample processing

Translating the discoveries obtained by whole genome sequencing into clinical practice will require thoughtfully designed clinical trials that incorporate genomic studies into the patient care algorithm. Procurement of high quality, tumour-rich, fresh frozen specimens with matched blood samples from consented patients is critical, as is optimal specimen processing to ensure the absence of artefacts in the sequencing data obtained.

Specimens require storage under conditions which maintain the integrity of nucleic acids (52).

Tissue samples are often formalin-fixed and paraffin-embedded (FFPE). While DNA extracted from such specimens may still be suitable for whole genome sequencing, FFPE processing, especially fixative concentration and pH, and duration of FFPE storage may disturb DNA integrity (53,54).

The technical challenges of using archived FFPE samples for direct DNA sequencing were highlighted during reanalysis of the landmark BR.21 study, which used direct DNA sequencing to screen for *EGFR* mutations (55). Mutational analysis was attempted on 197 samples, of which 40 (20.3%) required microdissection to obtain sufficient tumour for analysis. Due to

technical difficulties, mutation analysis was not possible in 20 (10%) samples (55).

Whole genome sequencing must overcome challenges posed by clinical samples, which may be of suboptimal size, contain necrotic tumour, have high stromal or non-tumoral content, or yield fragmented DNA. These limitations have prompted the development of novel experimental approaches and computational methods in order to obtain accurate second-generation sequencing data from FFPE samples (56,57).

Tumour content in diagnostic samples

Whole genome sequencing using NGS is typically performed on surgically resected tumour samples (58,59). However, as most patients are diagnosed at an advanced stage, such specimens are obtained in only a minority of patients with NSCLC (60). In situations where surgically resected specimens are available for whole genome sequencing, DNA quality from fresh/frozen samples is better, yielding more accurate sequencing data than tissue processed by FFPE (61,62).

In patients with advanced lung cancer, only diagnostic biopsy samples that can be obtained with minimal morbidity are available for molecular testing. Diagnosis for these patients is often based on small biopsy samples obtained during bronchoscopy or percutaneous needle aspiration biopsy procedures. Limited tumour tissue is obtained in such samples. A recent study used computer-aided morphometry to measure tumour area in morphologic sections of 100 bronchial biopsy samples and found that only 48% of biopsy samples contained tumour (63). Hence if bronchial biopsy samples are to be used for molecular testing, they should be first stained with haematoxylin and eosin (H&E) and the percentage of tumour cell content estimated by a pathologist. If the tumour content is less than 40%, the sample may require microdissection to augment the tumour cell proportion, a technique likely to be impractical in the clinical setting (64,65). Such additional processing steps are constrained by the amount of sample available.

The nature and purity of tumour in samples used as a source of DNA for NGS has the potential to profoundly influence interpretation of the sequencing output. Refinement of bioinformatic techniques to assess the significance of a somatic mutation in the context of the factors such as the background mutation rate, tumour ploidy and stromal contamination for primary cancers remain a major challenge of NGS (20).

Tumour heterogeneity

Not only is the tumour content of specimens used to source DNA of great importance, but heterogeneity within tumour samples can also have a profound impact on NGS results (20).

Intertumoral heterogeneity has long been recognised, as reflected in tumour classification systems based on morphology and immunohistochemical profiles. Cognisant of intratumoral heterogeneity, pathologists inspect several tumour sections when thoroughly examining a specimen. Intratumoral heterogeneity is particularly prevalent in pulmonary adenocarcinoma, leading to the recent recommendation by an international expert panel of quantitation of the predominant and lesser histological patterns in pathology reports to better predict prognosis and guide subsequent molecular investigation (66). DNA sequencing studies have provided insight into the molecular basis of this tumour heterogeneity. In a study of brain metastases from breast cancer by Ding *et al.* (67), similar sets of coding mutations were identified in the primary tumour and the metastasis, with gross differences in allelic frequency. This finding was suggestive of the presence of small subpopulations of cells with metastatic potential in the primary tumour.

The presence of subpopulations adds further complexity to cancer genome sequencing, and requires further investigation (68). Intratumoral heterogeneity may influence tumour aggressiveness, treatment responsiveness and resistance (69). Techniques such as macro-dissection or laser capture microdissection may permit isolation of subpopulations of cells prior to NGS, thus increasing power to detect mutations present in only part of the tumour (68). It is now possible to sequence amplified DNA sequences from single cells using NGS (70). Using this technology, it will become plausible to examine DNA obtained in samples of body fluid such as blood, urine or pleural fluid to diagnose cancer, and monitor for treatment resistance or recurrence.

Summary

Next-generation technology has the potential to transform our understanding of oncogenesis, techniques available for cancer research and the paradigm of cancer care. Whole genome sequencing is a powerful technique with which to examine changes present in the cancer genome, but requires considerable investments in sample collection and processing, computational analysis and bioinformatic interpretation of results. Large scale, international collaborative efforts are currently underway to apply this technology to the study of many cancers of public health importance.

Ongoing advances in NGS will lower the cost and increase speed and accuracy of processing, enable its application to smaller specimens and increase its accessibility to researchers. Ultimately, NGS may be available to clinicians in the diagnosis, treatment and follow-up of patients with lung cancer.

Funding sources

This work was supported by National Health and Medical Research Council (NRMHC) project grants (KF); NHMRC

Practitioner Fellowship (KF); NHMRC Career Development Fellowship (IY); NHMRC Biomedical Scholarship (CW); Cancer Council Queensland PhD Scholarship (MD); Cancer Council Queensland Senior Research Fellowship (KF); Cancer Council Queensland project grants; Queensland Smart State project grants; Office of Health and Medical Research (OHMR) project grants; The Prince Charles Hospital Foundation, Australian Lung Foundation/Boehringer Ingelheim COPD Research Fellowship (IY); NHMRC Postgraduate Medical Scholarship (KBS); University of Queensland (UQ) PhD Scholarship (KBS) and UQ Early Career Researcher Fellowship (VR).

Acknowledgements

We thank the patients and staff of The Prince Charles Hospital, for their involvement and contribution to the TPCH lung research program.

References

1. Howlader N, Noone AM, Krapcho M, Neyman N, Aminou R, Waldron W, et al. SEER Cancer Statistics Review, 1975-2008. Bethesda: National Cancer Institute. 2011 [cited 2012 Jan 16]. Available online: http://seer.cancer.gov/csr/1975_2008
2. Australian Institute of Health and Welfare, Cancer Australia, Australasian Association of Cancer Registries. Cancer survival and prevalence in Australia: cancers diagnosed from 1982 to 2004. Canberra: AIHW; 2008.
3. Globocan 2008. Lyon: International Agency for Research on Cancer. c2010 [cited 2012 Jan 15]. Available online: http://globocan.iarc.fr/
4. Sanger F, Coulson AR. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. J Mol Biol 1975;94:441-8.
5. Maxam AM, Gilbert W. A new method for sequencing DNA. Proc Natl Acad Sci U S A 1977;74:560-4.
6. Reddy EP, Reynolds RK, Santos E, Barbacid M. A point mutation is responsible for the acquisition of transforming properties by the T24 human bladder carcinoma oncogene. Nature 1982;300:149-52.
7. Tabin CJ, Bradley SM, Bargmann CI, Weinberg RA, Papageorge AG, Scolnick EM, et al. Mechanism of activation of a human oncogene. Nature 1982;300:143-9.
8. Dulbecco R. A turning point in cancer research: sequencing the human genome. Science 1986;231:1055-6.
9. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature 2001;409:860-921.
10. Green P. Against a whole-genome shotgun. Genome Res 1997;7:410-7.
11. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. Nature 2004;431:931-45.
12. Paez JG, Jänne PA, Lee JC, Tracy S, Greulich H, Gabriel S, et al. EGFR mutations in lung cancer: correlation with clinical response to gefitinib therapy. Science 2004;304:1497-500.
13. Soda M, Choi YL, Enomoto M, Takada S, Yamashita Y, Ishikawa S, et al. Identification of the transforming EML4-ALK fusion gene in non-small-cell lung cancer. Nature 2007;448:561-6.
14. Weir BA, Woo MS, Getz G, Perner S, Ding L, Beroukhim R, et al. Characterizing the cancer genome in lung adenocarcinoma. Nature 2007;450:893-8.
15. Kwak EL, Bang YJ, Camidge DR, Shaw AT, Solomon B, Maki RG, et al. Anaplastic lymphoma kinase inhibition in non-small-cell lung cancer. N Engl J Med 2010;363:1693-703.
16. Ley TJ, Mardis ER, Ding L, Fulton B, McLellan MD, Chen K, et al. DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. Nature 2008;456:66-72.
17. Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, Lee C, et al. Targeted capture and massively parallel sequencing of 12 human exomes. Nature 2009;461:272-6.
18. Maher CA, Kumar-Sinha C, Cao X, Kalyana-Sundaram S, Han B, Jing X, et al. Transcriptome sequencing to detect gene fusions in cancer. Nature 2009;458:97-101.
19. Mardis ER. Next-generation DNA sequencing methods. Annu Rev Genomics Hum Genet 2008;9:387-402.
20. Meyerson M, Gabriel S, Getz G. Advances in understanding cancer genomes through second-generation sequencing. Nat Rev Genet 2010;11:685-96.
21. Ross JS, Cronin M. Whole cancer genome sequencing by next-generation methods. Am J Clin Pathol 2011;136:527-39.
22. Illumina. Illumina Inc.; c2012 [cited 2012 Jan 20]. Available online: http://www.illumina.com
23. 454 Sequencing. Roche Diagnostics Corporation; c1996-2012 [cited 2012 Jan 20]. Available online: http://www.454.com
24. Life Technologies: Applied Biosystems. Life Technologies; c2011 [cited 2012 Jan 20]. Available online: http://www.appliedbiosystems.com.au
25. Life Technologies: Ion Torrent. Ion Torrent Systems Inc.; [cited 2012 Jan 20]. Available online: http://www.iontorrent.com
26. Rusk N. Torrents of sequence. Nat Methods 2011;8:44.
27. Lee W, Jiang Z, Liu J, Haverty PM, Guan Y, Stinson J, et al. The mutation spectrum revealed by paired genome sequences from a lung cancer patient. Nature 2010;465:473-7.
28. Erlich Y, Mitra PP, delaBastide M, McCombie WR, Hannon GJ. Alta-Cyclic: a self-optimizing base caller for next-generation sequencing. Nat Methods 2008;5:679-82.
29. Ding L, Wendl MC, Koboldt DC, Mardis ER. Analysis of next-generation genomic data in cancer: accomplishments and challenges. Hum Mol Genet 2010;19:R188-96.
30. Campbell PJ, Stephens PJ, Pleasance ED, O'Meara S, Li H, Santarius T, et al. Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. Nat Genet 2008;40:722-9.
31. Wang K, Kan J, Yuen ST, Shi ST, Chu KM, Law S, et al. Exome sequencing identifies frequent mutation of ARID1A in molecular subtypes of gastric cancer. Nat Genet 2011;43:1219-23.
32. Quail MA, Kozarewa I, Smith F, Scally A, Stephens PJ, Durbin R, et al. A large genome center's improvements to the Illumina sequencing system.

Nat Methods 2008;5:1005-10.
33. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, et al. Genome sequencing in microfabricated high-density picolitre reactors. Nature 2005;437:376-80.
34. Fakhrai-Rad H, Pourmand N, Ronaghi M. Pyrosequencing: an accurate detection platform for single nucleotide polymorphisms. Hum Mutat 2002;19:479-85.
35. Borràs E, Jurado I, Hernan I, Gamundi MJ, Dias M, Martí I, et al. Clinical pharmacogenomic testing of KRAS, BRAF and EGFR mutations by high resolution melting analysis and ultra-deep pyrosequencing. BMC Cancer 2011;11:406.
36. Metzker ML. Sequencing technologies - the next generation. Nat Rev Genet 2010;11:31-46.
37. Kobayashi S, Boggon TJ, Dayaram T, Jänne PA, Kocher O, Meyerson M, et al. EGFR mutation and resistance of non-small-cell lung cancer to gefitinib. N Engl J Med 2005;352:786-92.
38. Monzon FA, Ogino S, Hammond ME, Halling KC, Bloom KJ, Nikiforova MN. The role of KRAS mutation testing in the management of patients with metastatic colorectal cancer. Arch Pathol Lab Med 2009;133:1600-6.
39. Sjöblom T, Jones S, Wood LD, Parsons DW, Lin J, Barber TD, et al. The consensus coding sequences of human breast and colorectal cancers. Science 2006;314:268-74.
40. Yassour M, Kaplan T, Fraser HB, Levin JZ, Pfiffner J, Adiconis X, et al. Ab initio construction of a eukaryotic transcriptome by massively parallel mRNA sequencing. Proc Natl Acad Sci U S A 2009;106:3264-9.
41. Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, et al. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature 2007;448:553-60.
42. Brose MS, Volpe P, Feldman M, Kumar M, Rishi I, Gerrero R, et al. BRAF and RAS mutations in human lung cancer and melanoma. Cancer Res 2002;62:6997-7000.
43. Pleasance ED, Stephens PJ, O'Meara S, McBride DJ, Meynert A, Jones D, et al. A small-cell lung cancer genome with complex signatures of tobacco exposure. Nature 2010;463:184-90.
44. Ju YS, Lee WC, Shin JY, Lee S, Bleazard T, Won JK, et al. A transforming KIF5B and RET gene fusion in lung adenocarcinoma revealed from whole-genome and transcriptome sequencing. Genome Res 2012;Genome Res 2012;22:436-45.
45. Greenman C, Stephens P, Smith R, Dalgliesh GL, Hunter C, Bignell G, et al. Patterns of somatic mutation in human cancer genomes. Nature 2007;446:153-8.
46. International Cancer Genome Consortium. International Cancer Genome Consortium; c2011 [updated 2011 Dec 12, cited 2012 Jan 21]. Available online: http://www.icgc.org
47. The Cancer Genome Project. Bethesda: National Health Institute; [cited 2012 Jan 16]. Available online: http://cancergenome.nih.gov
48. Hanauer DA, Rhodes DR, Sinha-Kumar C, et al. Bioinformatics approaches in the study of cancer. Curr Mol Med 2007;7:133-41.
49. Jackson DB, Sood AK. Personalized cancer medicine--advances and socio-economic challenges. Nat Rev Clin Oncol 2011;8:735-41.
50. Lander ES. Initial impact of the sequencing of the human genome. Nature 2011;470:187-97.
51. Stratton MR, Campbell PJ, Futreal PA. The cancer genome. Nature 2009;458:719-24.
52. Wang F, Wang L, Briggs C, Sicinska E, Gaston SM, Mamon H, et al. DNA degradation test predicts success in whole-genome amplification from diverse clinical samples. J Mol Diagn 2007;9:441-51.
53. Srinivasan M, Sedmak D, Jewell S. Effect of fixatives and tissue processing on the content and integrity of nucleic acids. Am J Pathol 2002;161:1961-71.
54. Hewitt SM, Lewis FA, Cao Y, Conrad RC, Cronin M, Danenberg KD, et al. Tissue handling and specimen preparation in surgical pathology: issues concerning the recovery of nucleic acids from formalin-fixed, paraffin-embedded tissue. Arch Pathol Lab Med 2008;132:1929-35.
55. Tsao MS, Sakurada A, Cutz JC, Zhu CQ, Kamel-Reid S, Squire J, et al. Erlotinib in lung cancer - molecular and clinical predictors of outcome. N Engl J Med 2005;353:133-44.
56. Gallegos Ruiz MI, Floor K, Rijmen F, Grünberg K, Rodriguez JA, Giaccone G. EGFR and K-ras mutation analysis in non-small cell lung cancer: comparison of paraffin embedded versus frozen specimens. Cell Oncol 2007;29:257-64.
57. Lee W, Jiang Z, Liu J, Haverty PM, Guan Y, Stinson J, et al. The mutation spectrum revealed by paired genome sequences from a lung cancer patient. Nature 2010;465:473-7.
58. Campbell PJ, Stephens PJ, Pleasance ED, O'Meara S, Li H, Santarius T, et al. Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. Nat Genet 2008;40:722-9.
59. Hoffman PC, Mauer AM, Vokes EE. Lung cancer. Lancet 2000;355:479-85.
60. Marchetti A, Felicioni L, Buttitta F, Tsao MS, Kamel-Reid S, Shepherd FA. Assessing EGFR Mutations. N Engl J Med 2006;354:526-8.
61. Tsao MS. Should mutational analyses of tumor samples bypass histopathology? J Thorac Oncol 2007;2:375-6.
62. Wood HM, Belvedere O, Conway C, Daly C, Chalkley R, Bickerdike M, et al. Using next-generation sequencing for high resolution multiplex analysis of copy number variation from nanogram quantities of DNA from formalin-fixed paraffin-embedded specimens. Nucleic Acids Res 2010;38:e151.
63. Coghlin CL, Smith LJ, Bakar S, Stewart KN, Devereux GS, Nicolson MC, et al. Quantitative analysis of tumor in bronchial biopsy specimens. J Thorac Oncol 2010;5:448-52.
64. Eberhard DA, Giaccone G, Johnson BE; Non-Small-Cell Lung Cancer Working Group. Biomarkers of response to epidermal growth factor receptor inhibitors in Non-Small-Cell Lung Cancer Working Group: standardization for use in the clinical trial setting. J Clin Oncol 2008;26:983-94.
65. John T, Liu G, Tsao MS. Overview of molecular testing in non-small-cell lung cancer: mutational analysis, gene copy number, protein expression and other biomarkers of EGFR for the prediction of response to tyrosine kinase inhibitors. Oncogene 2009;28:S14-23.
66. Travis WD, Brambilla E, Noguchi M, Nicholson AG, Geisinger KR, Yatabe Y, et al. International association for the study of lung cancer/

american thoracic society/european respiratory society international multidisciplinary classification of lung adenocarcinoma. J Thorac Oncol 2011;6:244-85.

67. Ding L, Ellis MJ, Li S, Larson DE, Chen K, Wallis JW, et al. Genome remodelling in a basal-like breast cancer metastasis and xenograft. Nature 2010;464:999-1005.
68. Russnes HG, Navin N, Hicks J, Borresen-Dale AL. Insight into the heterogeneity of breast cancer through next-generation sequencing. J Clin Invest 2011;121:3810-8.
69. Merlo LM, Shah NA, Li X, Blount PL, Vaughan TL, Reid BJ, et al. A comprehensive survey of clonal diversity measures in Barrett's esophagus as biomarkers of progression to esophageal adenocarcinoma. Cancer Prev Res (Phila) 2010;3:1388-97.
70. Navin N, Kendall J, Troge J, Andrews P, Rodgers L, McIndoo J, et al. Tumour evolution inferred by single-cell sequencing. Nature 2011;472:90-4.

Cite this article as: Daniels M, Goh F, Wright GM, Sriram KB, Relan V, Clarke BE, Duhig EE, Bowman RV, Yang IA, Fong KM. Whole genome sequencing for lung cancer. J Thorac Dis 2012;4(2):155-163. DOI: 10.3978/j.issn.2072-1439.2012.02.01