**Cancer Research Division**

**PhD STUDENT COMMITTEE REPORT**

|  |  |
| --- | --- |
| **STUDENT NAME: Kenneth Doig** | |
| **MEETING DATE: 2 April 2014** | **VENUE: Jack Brockhoff Theatre, Lev. 3** |
| **COMMITTEE:** | |
| **SUPERVISORS: Prof. Stephen Fox and A/Prof. Tony Papenfuss** | |
| **MENTOR: Prof. Ricky Johnston** | |
| **PROJECT TITLE: Addressing the barriers to translating genetic sequencing into a clinical oncology setting.** | |
| **FIELD OF RESEARCH CODE: 111203 Cancer Genetics** | |

**PhD Committee Confirmation meeting**

# Abstract

*Researchers are increasingly using Genomics, but its uptake into the clinical environment has been slow, mainly because of the complexities of bioinformatic analysis. It is in hospitals that the high diagnostic and discriminative ability of genomics could have positive health impacts for individual patients and for infection control programs. Hospital pathologists and microbiologists have been reluctant to embrace genomic approaches because of issues of sequencing costs and the complexity of data analysis. The former issue is being addressed by rapid technology improvements in NGS equipment. However, data complexity and problems associated with data analysis remain a serious impediment to the wider use of genomics in clinical pathology.*

*Over the last few years, the commodisation of genetic sequencing platforms has brought nucleotide level genetic analysis within the economic reach of most hospitals and pathology labs. This has allowed these institutions to have unprecedented access to genetic diagnostic information about their patients and diseases. But the volume of data has left a gap between the availability of genetic information and the institute’s ability to process and interpret the data. This PhD proposes to address this gap and address the barriers to clinical adoption of high throughput sequencing diagnostics.*

# Report

Clinical cancer molecular diagnostics is undergoing a transformation. It is a result of a larger paradigm shift in the medical world from empirical treatment to the use of targeted therapies that are matched to mutations, coupled with the widespread availability of affordable high throughput sequencers. It holds the promise of a step change in our understanding of fundamental biology with an impact on disease and patient care. To harness the promise there is a concomitant requirement for diagnostic labs to adopt many of the methods of engineering, such as production line automation and data driven processing but this realisation is currently limited by the level of technical ability, our biological understanding of gene variants, logistical and educational needs. This PhD research considers these roadblocks and how they might be addressed.

Diagnostic genetic testing of patient cancer samples has been widely available for some time using Sanger sequencing[1](#_ENREF_1), the workhorse technology of the Human Genome Project2, 3[. In contrast, the research community has largely switched to high throughput sequencing (HTS), driven by the benefits of low cost and the wide range of genetic insights revealed](#_ENREF_3)[4-8](#_ENREF_4). Recognizing these advantages, numerous diagnostic labs are adopting HTS, but what has yet to occur is the widespread availability of the decision support tools to process, filter and report on sequencer output. HTS generates large volumes of data which produces up to 50 variants for every 1,000 bases sequenced. It is the nature of HTS that it has high error rates resulting in more false positives than true biological variants. It is the task of sophisticated bioinformatic pipelines, statistical methods and clinical informatics to sift the sequencer data and present the diagnostic scientist with a biologically relevant and meaningful set of mutations for analysis. It is this analysis that presents some of the greatest challenges for diagnostic labs entering the field. Further, the accreditation processes are struggling to adapt to the rapidly evolving landscape due the limited number of experts in the area. The barriers to entry centre on the complexity of translating the informatics needed to perform clinical quality assays from the research environment. The bioinformatics approaches of research do not scale to the production environment required for molecular pathology. Also, the typical skill sets of all but the largest clinical and research labs are unlikely to encompass the qualifications needed to support HTS. In addition to traditional wet lab skills, HTS requires people with skills in handling large volumes of data, bioinformaticians and molecular scientists capable of discriminating the nature of the large numbers of variants generated.

## Cancer Biology

Categorising a clinically actionable set of variants is the primary goal of cancer diagnostics but the intrinsic nature of cancer creates many challenges for sequencing. Each patient presents a unique genotype with benign polymorphisms and an unknown number of cancer predisposing variants. Overlaying this, somatic mutations are acquired of which a small subset drive the cellular proliferation of the tumour. To further complicate matters, the allele frequency of driver mutations may also impact patient response to therapies.

A germline sample contains homozygous and heterozygous variants, which are sampled at allele frequencies of around 100% and 50%. These values are well above the allele frequency of most sequencing false positives. In contrast, a tumour biopsy contains a largely unknown mix of stromal and tumour cells (tumour purity). Additionally, tumour heterogeneity further dilutes the mutational signal of variants present in the tumour. It’s been shown that the cellular mix of a tumour can change significantly[7](#_ENREF_7) as the result of the selective pressure from cancer interventions such as chemotherapy or targeted therapies, which may result in clonal evolution. The presence of deleterious mutations, not previously detected, in relapsed patients suggest that there are biological limits to what can be detected by HTS.

Many variants identified by HTS are of unknown significance. These may be benign personal germline mutations of the patient or passenger mutations that are fixed in tumour cells as they are selected for at tumourigenesis or later in disease progression. A lack of understanding of tumour biology, especially in non-coding regions, will mean these variants will remain of unknown significance without extensive lab work and functional testing. There is a long tail of cancer-implicated genes that have either a small effect in the cancer phenotype or occur rarely in patient cohorts[9](#_ENREF_9) and only a fraction these have diagnostic, prognostic or therapeutic utility. It is the aim of cancer diagnostics to reliably identify these variants and translate their presence into a meaningful clinical report for the treating oncologist.

## HTS Issues

A robust process for variant classification along a number of dimensions is needed to determine if and how they should be reported. These include determination whether the variant is an artefact from the sample preparation process (e.g. FFPE artefacts[10](#_ENREF_10)) or from sequencing. Furthermore, each targeted sequencing panel or capture technology has their own specific limitations, which must be identified and controlled by the bioinformatics software and its configuration. This is currently a largely empirical process to ‘tune’ the assay and bioinformatics pipeline for specificity and sensitivity. The laboratory needs to establish whether the variant is pathogenic in the patient’s disease context. If the variant has been previously identified it might be able to be classified using databases and the literature but as these are widely distributed and vary enormously in quality, accessibility and uniformity different laboratories may report variants differently. If the variant is undocumented, there are many ‘in-silico’ classification algorithms to infer pathogenicity[11](#_ENREF_11). These use a variety of biochemical, sequence conservation and locus attributes to assess the variant. However, no single classifier can be solely relied upon but collectively they can indicate the importance of a variant. Thus clinical labs must rely on their own research capabilities to identify variants seen within their own samples or catalogued in external databases. Although this may be mitigated in the future by a number of initiatives it will be some years, before a usable trusted clinical variant database emerges that aggregates variants together with patient phenotype data, a crucial step to identifying variants, especially those of low penetrance. Until then, the trustworthiness of external sources is a key issue for the confident assessment of a variant.

A key question is whether the variant is clinically actionable? Only a small proportion of identified pathogenic variants in patients will have an appropriate therapy for their inferred disease action. This number is increasing as new targeted drugs become available. The identification of therapies for a patient requires the diagnostic lab to access and interpret external databases for a match between the variant/gene/pathway and the currently approved therapies or clinical trials. Again, this is an informatics task outside the ability of most diagnostic laboratories.

The common requirements for the above issues is the need for sophisticated informatics tools to manage and manipulate the HTS data and create a clinically defensible clinical report. There is currently no consensus on how to perform many of the informatics tasks in HTS let alone the entire process. The components required to achieve HTS clinical diagnostics vary significantly from assay to assay and typically encompass a collection of open source and commercial systems joined together in an ad hoc fashion. The pedigree for these systems is typically the life science research community rather than commercial software vendors. Until unifying commercial software systems are developed, the ad-hoc nature of HTS will prevail and restrict clinical diagnostic HTS to organisations capable of supporting in-house bioinformaticians that can build the required pipelines and clinical informatics systems. Of these organisations, even fewer have in-house experts with the professional software engineering skills needed to build high quality systems that have been adequately tested using contemporary software engineering methodologies.

## Sequencing Technology

HTS sequencers continue to deliver improved capabilities at lower cost. Longer reads and more reads can be delivered faster and at a lower price points than ever before[12](#_ENREF_12). This progress brings sequencing within the reach of most diagnostic labs, but as with the IT revolution, the development of software has not matched the pace of hardware and this has become increasingly apparent in HTS processing [13](#_ENREF_13).

Each sequencing technology has limitations (see Table 1) and distinctive error characteristics that must be understood in the context of an assay. Sequencing errors from low complexity genome regions or regions with extreme GC content may require some regions to be excluded from the assay. Also, amplicon or capture methods may consistently bring up certain loci as false positives. This may be due to homopolymer runs in the capture region, off target effects from pseudogenes or poor assay design. These can be identified by observing bases that occur in multiple samples more frequently than appearing in polymorphism databases such as 1000g[14](#_ENREF_14) and ESP[15](#_ENREF_15). These types of variants must be identified on a per assay basis and require stringent validation of the assay prior to clinical use. On going monitoring of controls should ascertain that thresholds to filter out low variant allele frequency (VAF) sequencing noise are working correctly. There are many trade-offs with the technologies available for clinical diagnostics.

By nature, HTS has high error rates creating a statistical problem of discriminating signal (variants) from noise (sequencing artefacts). There is also a data-filtering problem of reducing gigabytes of sequencing data to a handful of filtered variants that can be reliably rendered into a diagnostic report in a clinically relevant timeframe. Both these tasks are carried out by a set of data manipulation tools collectively called a pipeline. Pipelines require careful software engineering to validate processes with a structured testing framework combined with at least three types of validation data.

## Variant Identification

A bioinformatics pipeline takes the raw output of short read DNA fragments from a sequencer and distils a set of variants representing the changes in a patients DNA relative to a known reference genome (usually GRCh37, HG19). There is currently little consensus on the tools that should be used, the parameters for the tools or even how the tools should be combined. Further, the rapid pace of sequencer and assay development has precluded a stable environment in which to create validated clinical pipelines. Different pipelines are required for each assay technology such as amplicon panels, capture panels, WES and WGS. Often multiple different pipelines are required for each technology to obtain acceptable results. The internals of a pipeline are closely linked to the sequencer and the wet lab processes of the assay. The pipeline should be responsible for identifying artifacts within the assay. Ideally, it would incorporate a model that assigns a probability that a variant is a sequencing artifact. This would require sophisticated modeling of the assay characteristics and a history of many sequencing runs to determine model parameters. The current methods of identifying artefactual variants varies widely across pipelines and may involve tens of QC parameters such as read depth, variant depth, uncallable regions of the assay and strand bias. The pedigree of the software tools making up todays HTS pipelines varies widely but is often the result of work done in the research environment that has found application more widely. Pipeline components are typically built as a concatenation of various open source and research derived tools. These tools are often built by researchers, (rather than software engineers), as a lab-centric solution to a local problem and are; not robust with respect to novel input data, poorly supported, poorly documented and often unmaintainable for a production environment or by groups other than the original developers. The development environment in which the tools were tested is usually up to the individual developers and often not of an appropriate standard for a clinical diagnostic tool. Clinical HTS diagnostics urgently needs a pipeline benchmarking framework together with a suite of tests and corresponding datasets and results to validate pipelines. Proposed requirements of clinical grade pipelines are suggested in see figure:

Although there will be different processing within pipelines, there needs to be a common framework across all pipelines covering common features such as; logging, restarting, auditing and documentation [16](#_ENREF_16). This allows pipeline behaviour to be reproducible between runs and allows sub tasks and common bioinformatic idioms to be modularised facilitating replacement of modules while maintaining overall pipeline behaviour. Clinical pipelines should also be structured to present a consistent interface for downstream processing by later components of molecular diagnostic processes. All variants should be formatted consistently using genomic coordinates to avoid transcript ambiguities. Regularising variants to a unique representation facilitates data matching with internal and external databases such as Clinvar and Cosmic.

Clinical pipelines need to undergo stringent test programs to mitigate the unknown quality of individual tools making up the pipeline. Functional end-to-end testing of pipelines offers the only realistic way to validate the behaviour and the interaction of components that may be written in multiple languages from different environments that need to cooperatively achieve a complex clinical diagnostic result.

Although there are “best practice” guides emerging for some suites of software, they are typically just the opinion of one institute rather than broadly agreed guidelines or rules sanctioned by a professional body. There is an urgent need to provide a validation framework for the benchmarking of clinical pipelines. This would achieve a consensus across a wide range of clinical applications and define standards for pipeline performance, behaviour, sensitivity and specificity. This is analogous to the development of computer CPU performance benchmarks in 70’s which allowed disparate claims made by competing computer vendors to be compared objectively. As with any complex software, rigorous testing in a controlled environment is needed.

The test data should contain comprehensive datasets comprising known gold standard samples, public read datasets, synthetic read datasets all with matching results to compare with pipeline output.

The discipline of software engineering has developed a number of methodologies to improve the quality of software. These include; test driven development to improve the testability of code, agile development to improve the alignment of code function with user requirements, continuous integration to identify faults early in the development lifecycle and automated web testing to improve the operation of user interfaces. These methodologies are rarely applied in a research setting but are essential for clinical software with patient care consequences.

The limitations of assay technology mean that multiple assays may need to be run to illuminate the various dimensions of tumour characteristics. While DNA sequencing identifies genomic changes, RNA sequencing can highlight which changes are expressed and how strongly. RNAseq can also identify which transcripts are present of the multiple possible gene isoforms. Methylation assays can also show regions of the genome that are suppressed through gene regulation mechanisms and altered protein function.

## Curation

Identifying polymorphisms present in a population including many ethnic groups is relatively straight forward with the availability of two well curated databases, 1000g[17](#_ENREF_17), ExAC and ESP[15](#_ENREF_15). Identifying cancer driver mutations has been a major research initiative over the last decade and has resulted in a number of large well-curated databases substantiated with documented case/control cohort data, although as outlined above, there is a long tail of genes with variants that occur with low frequency the pathogenicity of which is unclear.

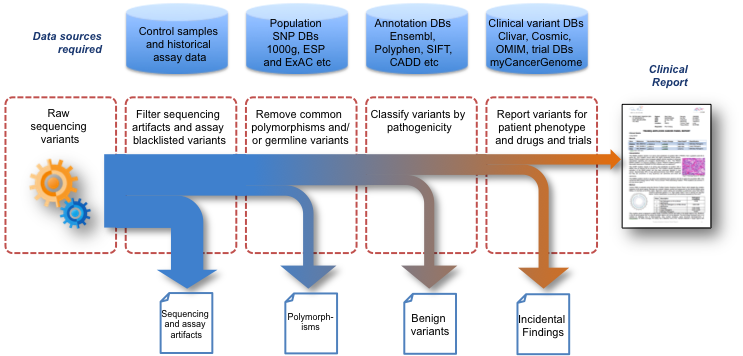


Figure 1 Variant Filtering

However, there is no single database aggregating all known pan-cancer variants together with pathogenicity status suitable for clinical curation purposes. Curation must rely on the complex task of federating multiple data sources including:

* Defining a uniform schema and semantics to store variants, pathogenicity, disease context and clinical case/control evidence
* Retrieval of data sources via the internet via FTP, web sites, subscription services etc.
* Transforming the data source variants to regularised HGVS genomic coordinates to enable data matching[18](#_ENREF_18)
* Transforming the pathogenicity status to a uniform semantics. eg C1-C5 classification[19](#_ENREF_19).
* Transforming the disease context (if any) of the data source variant to a uniform semantics
* Screening for duplicate reporting of the same variant/patient from different data sources
* Incorporating in-house curated variants into a consistent format.
* Refreshing data periodically to maintain currency.

There is a wide range of quality in variant databases with problems such as; ambiguous semantics of data fields, missing transcripts, ambiguous variant nomenclature and uncertain clinical provenance of variants. The error rates within database are usually unknown with no clear means of validation. These issues push labs into relying on the concordance between databases as a proxy for validation but this problematic unless the data has been sanitised and curated by data and curation experts.

The integration of HTS data with assay and pipeline is a complex task that must be easy for diagnostic scientists to use to minimise the effort required to be applied for each sample but robust and computationally correct to eliminate any misdiagnosis through software error.

## Reporting

As with cancer variants, drug and trial options need to be integrated in the HTS reporting system. Databases of drug and trials must be federated at the health system level (typically national). The therapeutic options need to be updated regularly as new drugs and trials become available or old drugs are repurposed for a broader range of cancer types. Such databases need to include detailed clinical information such the context where drugs can be used and trial enrolment criteria. However, as data volumes increase and the breadth of assayed genes expands, the crafting of recommendations will become a bottleneck in the turn around time for high volume labs. The skill mix required for HTS diagnostic reports are not common enough that this task can be scaled up to match the potential data volumes over the next few years.

Tumour boards are one approach to enable a recommendations but is not scalable to more than a handful of cases per meeting. Further, these meetings are often just for local or regional groups of experts meeting face-to-face constraining the frequency and convenience of such boards. We are in the process of implementing a Virtual Tumour Board (VTB) where a distributed group of cancer professionals that can subscribe to any combination of variants, genes, patient cohorts or cancer type within the curation system receive periodic notifications of cases via their medium of choice such as twitter, email, facebook or other messaging medium. The VTB concept is also readily scalable to large numbers of experts that can be located anywhere, even in different time zones. Harnessing social networking also offers novel options for the experts to interact to determine options. Similar ideas have been put forward for general practice in regional and remote areas.

## Human Factors

The current mix of skills within a traditional diagnostic lab focus are not necessarily appropriate for a lab transitioning to high volumes of data generated by HTS assays that centre on large data management, software customisation, automation of wet lab processing and genomic analysis. This mix is often more aligned with the discipline of software engineering and process automation than wet lab expertise. Existing labs are faced with the prospect of building up a team of software engineers and bioinformaticians or outsourcing these roles to larger regional labs that have the scale to create such teams. The other alternative is to use commercial off the shelf systems to process and analyse their in-house sequenced data. This path however carries considerable risk at present. Without the necessary software engineering and bioinformatics skills, validating an HTS process end-to-end will rely on systems vendors assurances which may not apply to the specific labs context or assays.

## Addressing the Barriers to Clinical Sequencing

This PhD aims to address the barriers to entry by developing the methodologies and software engineering systems necessary for the processing of HTS data from sequencer through to a clinical report. The following systems are in the process of being developed as exemplars of ways to address sequencing and analysis issues:

**Variant curation**: Path-OS is the in-house curation system used within the Peter MacCallum Molecular pathology department. It was created to meet the need of a system that could store, annotate and report on HTS variants for clinical diagnostics. The lack of commercial or public domain software to address this need meant that the system needed to be built as a new web application. I selected Grails (http://grails.org) as a rapid prototyping framework and developed the system which has been actively used for the past fifteen months.

**Variant Annotation Cache**: A major blocker to the widespread adoption of sequencing in the clinic is the access to reliable data repositories on the clinical nature of a variant. Currently molecular diagnostic labs rely on an ad-hoc collection of web sites, tools and processes to decide how to report a patients variants. I have created a site wide variant cache containing all variants and their corresponding annotations for all variants generated through HTS For each new sequencing run, the bioinformatics pipeline calls thousands of variants of which only approximately 10% are novel, the remainder having been called in previous runs and already annotated. This results in the saving of most of the computationally expensive annotation steps thereby improving pipeline times and reducing overall turn around times for patient diagnostics.

**Pipeline Validation**: NGS pipelines are often built by researchers, (rather than software engineers), as a lab-centric solution to a local problem. As such, they often share some or all of the following characteristics: not robust with respect to novel input data, poorly supported, poorly documented and often unmaintainable for a production environment. To mitigate the shortcomings of existing pipelines, I have developed PipeCleaner, a testing framework for pipeline validation. PipeCleaner generates known input reads for a pipeline, then repeatedly runs the pipeline under test and finally collects the actual variants found by the pipeline and compares them to the expected variants in the input reads.

**Variant Viewer**: VCF files have become the lowest common denominator for describing the variants generated by HTS. Despite their shortcomings and initiatives to enhance or replace them, they will remain the medium of exchange between HTS pipelines, downstream analysis tools and sequencing operators for some time to come. While formatted as a text file, VCF files are not readily manipulated by humans and so risk being misinterpreted. In addition, they struggle to fully represent the nuances of genetic changes that may be gleaned from downstream processes. To address these shortcomings I have built VariantViewer, a light-weight web application that can load multiple VCF files and display them in a tabular grid that allows merging, filtering and exporting. Additionally, the variants can be viewed and explored by an embedded genomic browser.

**Streaming Analysis**: Pipelines are often characterised by complex tuning of parameters, multiple software dependencies, lack of portability and the need for a computing environment with a large amounts of memory. I have developed Canary as a lightweight software application that bypasses the need for a pipeline and can generate a draft diagnostic report directly from a FASTQ file. Speed is achieved as a result of the observation that a diagnostic assay report may only need to report a handful of known actionable mutations lying within the capture region of the assay. Amplicon reads are effectively ‘pre-aligned’ by virtue of the DNA primer action and this removes the need for a computationally expensive alignment process.

The above systems and web applications are an initial approach to demonstrate potential ways to improve the end to end process of taking sequencer output and analysing the data to generate robust and meaningful diagnostic reports in a clinically relevant timeframe. These approaches will be validated and enhanced over the course of this PhD to determine their appropriateness for clinical oncological diagnostics.

# Published papers, submitted papers and books:

1. *Massively-parallel sequencing assists the diagnosis and guided treatment of cancers of unknown primary,* Richard W Tothill, Jason Li, Linda Mileshkin, **Ken Doig**, Terence Siganakis, Prue Cowin, Andrew Fellowes, Timothy Semple, Stephen Fox, Keith Byron, Adam Kowalczyk, David Thomas, Penelope Schofield, David D Bowtell, J Pathol. 2013 Dec;231(4):413-23. doi: 10.1002/path.4251.
2. *Bioinformatics Pipelines for Targeted Resequencing and Whole-Exome Sequencing of Human and Mouse Genomes: A Virtual Appliance Approach for Instant Deployment.*

Jason Li, Maria A. Doyle, Isaam Saeed, Stephen Q. Wong, Victoria Mar, David L. Goode, Franco Caramia, **Ken Doig**, Georgina L. Ryland, Ella R. Thompson, Sally M. Hunter, Saman K. Halgamuge, Jason Ellul, Alexander Dobrovic, Ian G. Campbell, Anthony T. Papenfuss, Grant A. McArthur, Richard W. Tothill *PLoS ONE 9(4): e95217. doi:10.1371/journal.pone.009521*

1. *Sequence artefacts in a prospective series of formalin-fixed tumours tested for mutations in hotspot regions by massively parallel sequencing.*

Stephen Q Wong, Jason Li, Angela Y-C Tan, Ravikiran Vedururu, Jia-Min B Pang, Hongdo Do, Jason Ellul, **Ken Doig**, Anthony Bell, Grant A MacArthur, Stephen B Fox, David M Thomas, Andrew Fellowes, John Parisot and Alexander Dobrovic*. BMC Medical Genomics*

1. *Assessing the clinical value of targeted massively parallel sequencing in a longitudinal, prospective population-based study of cancer patients. British Journal of Cancer (2015), 1–10 | doi: 10.1038/bjc.2015.80.*

Stephen Q. Wong1,2, Andrew Fellowes1, **Ken Doig3**,4, Jason Ellul3, Trent Bosma1, Darryl Irwin5, Ravikiran Vedururu1, Angela Y-C Tan1, Jonathan Weiss6, Kian Sing Chan7, Mark Lucas8, David M. Thomas2,4,9, Alexander Dobrovic1,2,4,6,10, John P Parisot2,4, Stephen B Fox

1. *Clinical Bioinformatics 2nd edition, Book Chapter*

Maria A. Doyle, Jason Li, **Ken Doig**, Andrew Fellowes and Stephen Q. Wong

1. *The Cancer 2015 Cohort: A Unique Prospective and Longitudinal, Population-based Cancer Genomics Cohort - Clinical and Economic Rationale for Personalized Medicine. Submitted for publication.*

John P Parisot, Heather Thorne, Andrew Fellowes, **Ken Doig**, Mark Lucas, John McNeil, Brett Doble, Alexander Dobrovic, Paul James, Lara Lipton, David Ashley, Theresa Hayes, Paul McMurrick, Gary Richardson, Paula Lorgelly, Stephen B Fox, David M Thomas.

# Conferences, Seminars and Working Groups

* Oral presentation. WEHI bioinformatics seminar, May 2014
* Oral presentation and Poster. Human Variome Project 5th Biennial meeting Paris, May 2014
* Oral presentation. VLSCI
* Oral presentation. Laby Foundation November/December 2014
* Oral presentation. Garvan Institute Sydney, November 2014.
* Oral presentation. 10th Australasian Mutation Detection Meeting, August 2014
* Upcoming Oral presentation. 13th International Symposium on Mutation in the Genome: detection, genome sequencing & interpretation, May 2015
* Invited oral presentation: VCCC Molecular Tumour Board, Sep 2013
* Invited participant. RCPA workshop - Standards for Clinical Databases of Genetic Variants.
* Invited participant. Melbourne Genomics Health Alliance.
* Invited participant. NHMRC TCR in Genomics: Federated Data Sharing Workshop. Mar 2015
* Invited participant. RCPA Pathwiki editorial group. Feb-Mar 2015

# Summary of progress to date

The following sections summarise the progress in the main areas of focus in this PhD. All systems have been built to address the various limitations of current practices or to demonstrate a practical solution to the aims of this PhD.

Variant Curation Systems: PathOS

PathOS is an operational system <http://bioinf-pathos/PathOS/> that has been used for curating Molecular Pathology samples since mid 2014. Prior to this it was developed to curate the Cancer 2015 data, a longitudinal, prospective population-based study of cancer patients. This system was created to address the need to replace laborious and error prone work practices in variant curation that were centred on spreadsheets and MS-word templates.

The current web pages for Path-OS outline the current state of PathOS and features in released versions see [*PathOS Confluence Page*](https://115.146.86.118/confluence/display/PVS/Path-OS+Variant+System). The Jira issue management site has a detailed list of completed features at [*PathOS Jira Issues*](https://115.146.86.118/jira/browse/PATHOS).

A paper has been prepared covering PathOS and is in the final stages of editing prior to submission.

Variant Annotation Cache

The current release of PathOS uses a variant cache to speed up annotation. The cache currently contains all sequenced NGS variants seen in Molecular Pathology panels since 2012. Only normalised variants that have been validated against the Mutalyzer web site are cached. The next release of PathOS will also cache all Ensembl, Annovar and Oncotator annotations.

VariantViewer

Variant Viewer[*Variant Viewer Confluence Page*](https://115.146.86.118/confluence/display/VV/Variant+Viewer) has been implemented as a demonstration web application to allow VCF variant files to loaded, annotated, merged and exported. It is fully functional but needs some further ‘hardening’ for public use. See <http://bioinf-pathos:8080/VcfViewer/>

Pipeline Validation: PipeCleaner

PipeCleaner is a validation and testing framework for independently verifying the capabilities and functional correctness of bioinformatics pipelines. A proof of concept implementation has been built and is used for regression testing PathOS releases.

Streaming Analysis Reporter: Canary

Canary has been built as a high speed variant reporter that operates directly on FASTQ files to identify specific variants. The current implementation can identify SNPs and INDELs from 380Mb short read files in 7 seconds.

# Proposed Schedule and Timeline

The following diagram outlines the major milestones to be achieved up to thesis submission.



For each of the tasks in the above diagram the following goals are proposed.

Variant Curation Systems: PathOS

* Integrate drug databases for reporting.
* Integrate trial databases for reporting
* Implement ACMG[20](#_ENREF_20) germline curation guidelines.
* Generalise the quality control information from pipelines to allow interchangable pipelines to be used upstream of PathOS
* Add support for CNVs and fusion events.
* Implement an integrated Circos viewer for both panel summarisation and individual sample navigation.
* Integrate genome browser to display sample variants and external database context
* Implement web driven automated functional testing to improve system robustness and validate fuctional operation.
* Improve application test coverage.
* Add support for social curation - Virtual Tumour Board
* Engage with external labs to encourage non Peter Mac use of PathOS.
* Automate the identification of sequencing artefacts using historical panel data.

Variant Annotation Cache

* Extend the current variant cache to cover all relevant cancer annotation sources, eg COSMIC, Clinvar, Oncotator, InSight, kConFab etc
* Investigate publication mining to access recently published literature as an annotation data source.
* Integrate all annotation sources into PathOS in a way that meaningfully prioritises data.
* Add a web interface to allow searching for any variant or its attributes
* Add a JSON API to allow any application in any language to query the cache

VariantViewer

* Gain feedback within Peter Mac researchers on the utility of Variant Viewer
* Harden Variant Viewer for common use cases of VCF manipulation.
* Deploy on NECTAR for public use.
* Establish a forum for user feedback on the public site.
* Add support for exporting VCF/MAF files.
* Add on-demand automation of VCF files using Variant Annotation Cache.

Pipeline Validation: PipeCleaner

* Develop comprehensive test suites for in-house pipelines covering, all functional, performance and regression dimensions of pipeline validation.
* Compare in-house pipelines with external pipelines to characterise differences and shortcomings.
* Undertake a full comparison of Illuminas standard Basespace workflow with an in-house pipeline.

Streaming Analysis Reporter: Canary

* Complete initial release for amplicon panels that identifies the VAF of a given set of variants.
* Validate Canary against the three years of sequencing held in PathOS.
* Calculate the specificity and sensitivity of the utility.
* Extend functionality for discovery analysis where variants are not provided.
* Understand the limits of applicability for non amplicon diagnostic assays.
* Integrate Canary into routine sequencer operation to give early warning of known deleterious mutations prior to sequencer completion.

Proposed Papers

The following first author papers and submission dates are planned.

* *The Clinical Bottleneck of Next Generation Cancer Diagnostics. Review paper for clinical oncologists (April 2015).*
* *Path-OS: A variant curation system for high throughput clinical sequencing. Technical Note (April 2015).*
* *Variant Viewer: A lightweight web application for managing VCF files. Technical Note (June 2015).*
* *A federated variant database for clinical sequencing curation. Technical Note (July 2015).*
* *PipeCleaner: A framework for the testing and evaluation of bioinformatics pipelines. Technical Note (November 2015).*
* *Canary: An ultra fast diagnostic reporter. Technical Note (Jan 2016).*

# Record of attendance



# References

1. Sanger, F. & Coulson, A.R. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J Mol Biol* **94**, 441-8 (1975).

2. Lander, E.S. et al. Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921 (2001).

3. Venter, J.C. et al. The sequence of the human genome. *Science* **291**, 1304-51 (2001).

4. Biesecker, L.G., Burke, W., Kohane, I., Plon, S.E. & Zimmern, R. Next-generation sequencing in the clinic: are we ready?(APPLICATIONS OF NEXT-GENERATION SEQUENCING--VIEWPOINT)(Report). *Nature Reviews Genetics*, 818 (2012).

5. Xuan, J., Yu, Y., Qing, T., Guo, L. & Shi, L. Next-generation sequencing in the clinic: promises and challenges. *Cancer Lett* **340**, 284-95 (2013).

6. Gullapalli, R.R., Desai, K.V., Santana-Santos, L., Kant, J.A. & Becich, M.J. Next generation sequencing in clinical medicine: Challenges and lessons for pathology and biomedical informatics. *J Pathol Inform* **3**, 40 (2012).

7. Lawrence, M.S. et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214-8 (2013).

8. Garraway, L.A. & Lander, E.S. Lessons from the cancer genome. *Cell* **153**, 17-37 (2013).

9. Vogelstein, B. et al. Cancer Genome Landscapes. *Science* **339**, 1546-1558 (2013).

10. Van Allen, E.M. et al. Whole-exome sequencing and clinical interpretation of formalin-fixed, paraffin-embedded tumor samples to guide precision cancer medicine. *Nat Med* **20**, 682-8 (2014).

11. Kircher, M. et al. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* **46**, 310-5 (2014).

12. Bentley, D.R. et al. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53-9 (2008).

13. Mardis, E.R. The $1,000 genome, the $100,000 analysis? *Genome Med* **2**, 84 (2010).

14. Abecasis, G.R. et al. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56-65 (2012).

15. Tennessen, J.A. et al. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* **337**, 64-9 (2012).

16. Sadedin, S.P., Pope, B. & Oshlack, A. Bpipe: a tool for running and managing bioinformatics pipelines. *Bioinformatics* **28**, 1525-6 (2012).

17. Abecasis, G.R. et al. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061-73 (2010).

18. Wildeman, M., van Ophuizen, E., den Dunnen, J.T. & Taschner, P.E. Improving sequence variant descriptions in mutation databases and literature using the Mutalyzer sequence variation nomenclature checker. *Hum Mutat* **29**, 6-13 (2008).

19. Thompson, B.A. et al. Application of a 5-tiered scheme for standardized classification of 2,360 unique mismatch repair gene variants in the InSiGHT locus-specific database. *Nat Genet* **46**, 107-15 (2014).

20. Richards, S. et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med* (2015).