Implementing an NGS Bioinformatics Pipeline: Making the Transition From Research to Clinical

L. Watkins, K. Hetrick, H. Ling, S. Griffith, E. Hsu, G. Lowe, K. Roberts
D. Snyder, M. Mawhinney, B. Craig, J. Romm, K. Doheny

Center for Inherited Disease Research (CIDR), IGM, JHU-SOM, Baltimore, MD

Introduction

The Center for Inherited Disease Research (CIDR) provides high quality next-generation sequencing (NGS), genotyping and statistical genetics consultation to investigators working to discover genes that contribute to disease. The CIDR NGS bioinformatics analysis pipeline, continually developed since 2009 to keep pace with the rapid changes in the predominant sequencing analysis tools, is designed and tuned for large-scale research projects with huge numbers of analyses that must be done in parallel and fault-tolerant fashion. In late 2013-14 CIDR partnered with a molecular diagnostic lab at our institution to migrate a traditional (Sanger/MLPA) 52-gene test panel to NGS on Illumina MiSeq, wherein CIDR would run the NGS bioinformatics analysis and the clinical lab would handle everything else. Despite a limited budget and tight timeline the effort was successful, with the first clinical results reported out in March 2014 and in continuous use since then. In some regards the requirements for a clinical bioinformatics pipeline are similar to those of production research (e.g., the need for reliability and strict quality control) while others are quite different (e.g., small numbers of samples, absolute need for quick turn-around time, strict adherence to well-defined regions of interest, an imperative to never report a false positive, specific and limited relevant annotations). A major difference is the requirement to lock down all aspects of the pipeline after validation and to document it in detail and track any associated changes in accordance with new CAP guidelines for NGS bioinformatics analysis. This militates against sharing bioinformatics infrastructure and pipelines between research and clinical, so the decision was made to establish separate systems for clinical NGS work. This poster focuses on the informatics and other practical aspects and related considerations of this initial effort.

Results

CIDR lab and informatics staff worked closely with diagnostic lab staff to answer questions such as: how much input DNA should be used with which polymerase and could we use DNA extracted by different methods; could we detect contamination and a variety of mutation types, and which aligner and caller perform most accurately with what combination of parameters and values. Once these questions were answered the bioinformatics pipeline was iteratively adjusted to achieve the desired sensitivity and specificity. The pipeline was then validated on 64 DNA samples, 1-4 genes/sample, assessing over 374,000 nucleotides to confirm >99% sensitivity/specificity across a set of 10 genes selected for initial implementation from a total NGS panel of 83 genes (plus 95 barcode SNPs included in the target design). A simplified finalized analysis workflow is shown in Figure 1. The full list consists of >30 separate steps, almost every one with multiple parameters set to specific values based on our iterative testing, which relies on 20 input files besides the initial fastq files, utilizing 10 separate software packages including BWA 0.7.5a and GATK 2.7-4. A file structure consisting of eight directories was defined to capture and store (as per CAP guidelines) all input and reference files (e.g., fastq, bed, sample sheets), submission and pipeline scripts, logs from any program that produces them, as well as all output

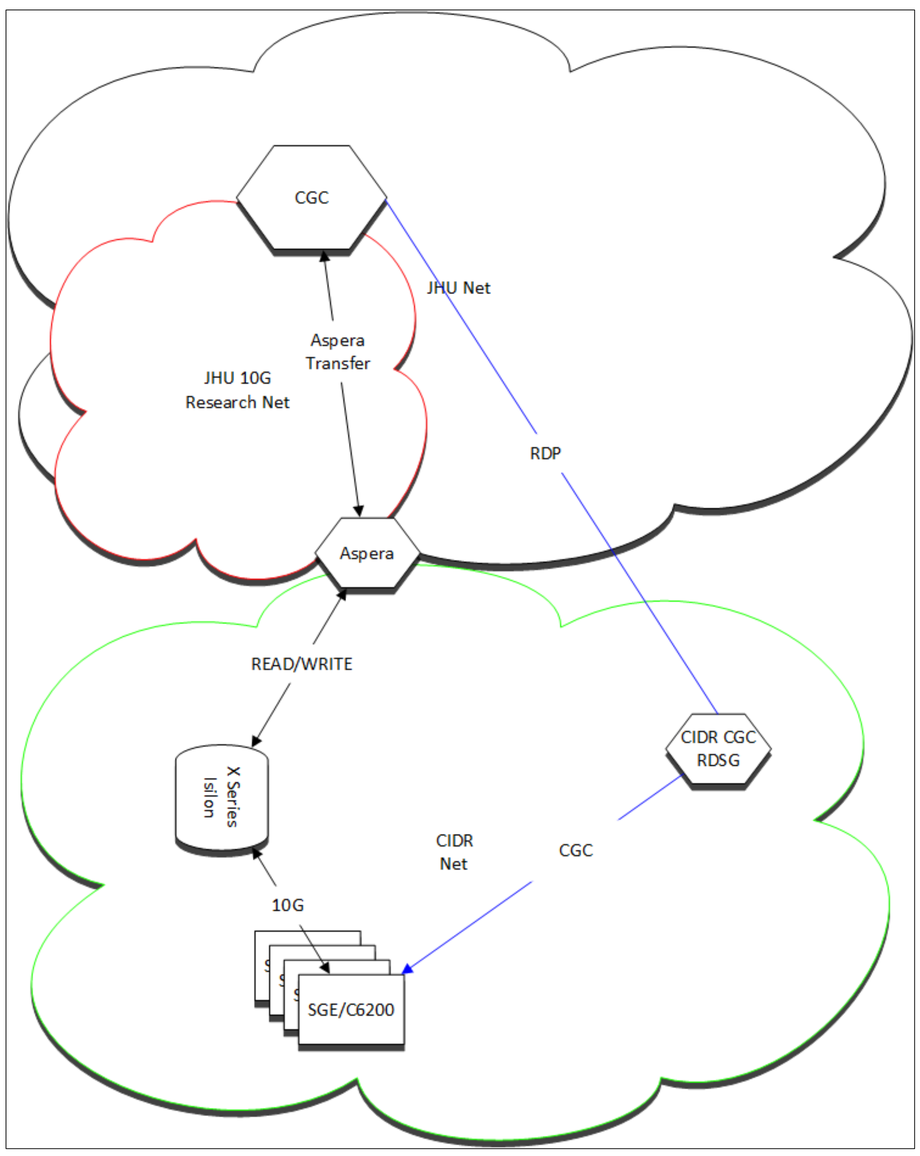

Figure 2: Network and system diagram showing data transfer paths, processing and storage.

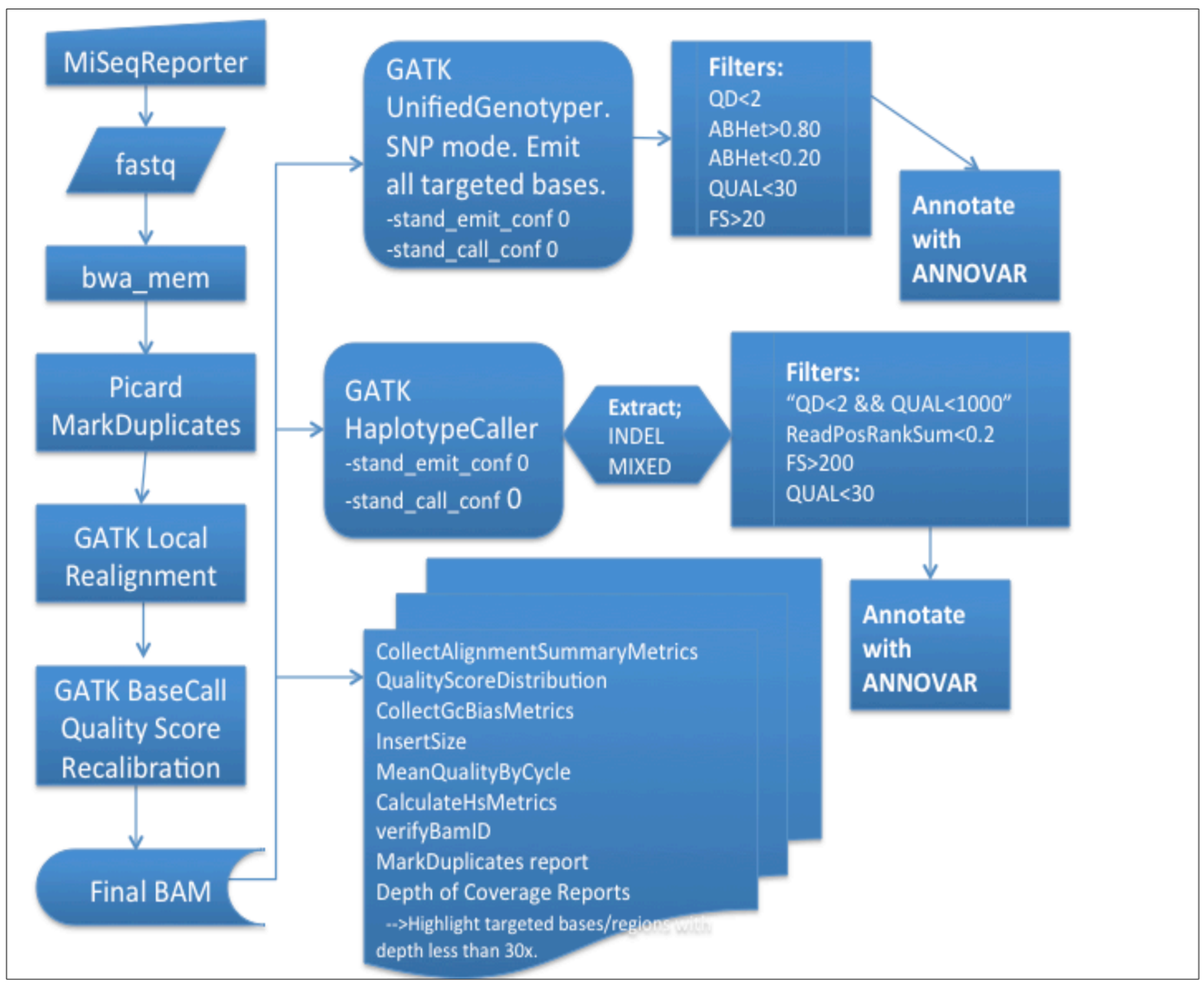

Figure 1: Final clinical diagnostic lab NGS bioinformatic analysis workflow.

data (e.g., bam, vcf) and reports in subdirectories dynamically created and named according to the run and sample name (SM_tag). Figure 2 shows the informatics infrastructure subsequently put in place, structured in accordance with CAP data security and confidentiality requirements, e.g., all data transfer is via a secure Aspera connection over an internal Hopkins network that is not exposed to the Internet. Every data transfer is accompanied by MD5 checks to assure data integrity; this proved critically important as during the pre-certification phase we encountered at least one case of spurious results found to be caused by a partially-corrupt BAM file. Diagnostic lab staff gain access to the CIDR LAN via secure remote desktop connection to a terminal server, from whence they connect to a Linux server to launch analysis jobs from the command line. While unsophisticated, this approach enabled a short time to completion and is simple, secure and fully adequate for start-up, but offers several opportunities for improvement in the next iteration.

Sequencer:	Illumina MiSeq; Agilent SureSelect capture - 433Kb targeted, ~15,000 probes
Storage:	Isilon 7200X, 4 node, 200TB
Compute:	Dell C6200, 4 x 64 core, 96GB; SGE
IT Costs:	$175,000 (approximate)
Staff Hours:	>1,000 over 4 months
Roles:	50% Bioinformatics, 30% Leadership, 20% IT
Cost:	$75,000 (approximate)

Table 1: IT, staff resources applied to clinical NGS start-up effort

Table 2 breaks down approximate informatics and staff costs that CIDR incurred implementing clinical NGS in partnership with the diagnostic lab. All of the IT equipment was repurposed from other uses; costs given are amortized value plus any relevant annual maintenance/support. Staff costs were tracked over the 4-month period Oct. 2013 through Jan. 2014 and do not include clinical staff. Overall, bioinformatics engineering and software development alone accounted for over three times as much effort expended as any other single category of effort. The proportions reflect the fact that this was an initial start-up effort and as such do not characterize ongoing effort.

Conclusion

A collaborative approach combining high-throughput production research with clinical diagnostic expertise is an effective and expeditious way to initially implement clinical NGS.