Bioinformatics, YYYY, 0–0
doi: 10.1093/bioinformatics/xxxxx
Advance Access Publication Date: DD Month YYYY
Manuscript Category

Subject Section

VCPA: genomic variant calling pipeline and data management tool for Alzheimer's Disease Sequencing Project

Yuk Yee Leung[1,2]*, Otto Valladares[1,2], Yi-Fan Chou[1,2], Han-Jen Lin[1,2], Amanda B Kuzma[1,2], Laura Cantwell[1,2], Liming Qu[1,2], Prabhakaran Gangadharan[1,2], Alzheimer's Disease Sequencing Project (ADSP), William J Salerno[3], Gerard D. Schellenberg[1,2] and Li-San Wang[1,2]*

[1]Penn Neurodegeneration Genomics Center, Perelman School of Medicine at the University of Pennsylvania Richards Building, D101 3700 Hamilton Walk, Philadelphia, PA, USA.

[2]Department of Pathology and Laboratory Medicine, Perelman School of Medicine at the University of Pennsylvania, PA, USA.

[3]Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX, USA.

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Summary: We report VCPA, our SNP/Indel Variant Calling Pipeline and data management tool used for analysis of whole genome and exome sequencing (WGS/WES) for the Alzheimer's Disease Sequencing Project. VCPA consists of two independent but linkable components: pipeline and tracking database. The pipeline is coded in Workflow Description Language and is fully optimized for the Amazon elastic compute cloud environment. This includes steps for processing raw sequence reads including read alignment, and all the way up to variant calling using GATK. The tracking database allows users to dynamically view the statuses of jobs running and the quality metrics reported by the pipeline. Users can thus monitor the production process and diagnose if any problem arises during the procedure. All quality metrics (>100 collected per processed genome) are stored in the database, thus facilitating users to compare, share and visualize the results. To summarize, VCPA is functional equivalent to the CCDG/TOPMed pipeline. Together with the dockerized database (also available as Amazon Machine Image), users can easily process any WGS/WES data on Amazon cloud with minimal installation.

Availability: VCPA is released under the MIT license and is available for academic and nonprofit use for free. The pipeline source code and step-by-step instructions are available from the National Institute on Aging Genetics of Alzheimer's Disease Data Storage Site (http://www.niagads.org/VCPA).

Contact: yyee@pennmedicine.upenn.edu or lswang@pennmedicine.upenn.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The Alzheimer's Disease Sequencing Project (ADSP) is an integral component of the National Alzheimer's Project Act (NAPA) towards a cure of Alzheimer's Disease. ADSP will eventually analyze whole-genome sequencing (WGS) and whole-exome sequencing (WES) data from more than 20,000 late-onset Alzheimer's Disease (AD) patients and cognitively normal elderly to finding new genetic variants associated with disease risk.

To ensure all sequencing data are processed following the best practices with consistency and efficiency, a common workflow was developed by the Genome Center for Alzheimer's Disease (GCAD) in collabo-

ration with ADSP. The workflow "Variant Calling Pipeline and data management tool" (VCPA), is used to process all the ADSP sequencing data. VCPA 1) is optimized for large-scale production of WGS and WES data, 2) includes a tracking database with web frontend for the user to track production process and review quality metrics; 3) is implemented using the Workflow Description Language (WDL) for easier deployment and maintenance, 4) designed for the latest human reference genome build (GRCh38/hg38, version GRCh38DH) and follows best practices for WGS analysis with input from TOPMed (Trans-Omics for Precision Medicine) and CCDG (Centers for Common Disease Genomics), two other large sequencing programs supported by National Institutes of Health (NIH).

VCPA is composed of two independent but interoperable components: a tracking database (Figure 1A) with a web frontend (Figure 1B) and a SNP/indel calling pipeline (Figure 1C). The pipeline (available as an Amazon Machine Images (AMI): ami-82aa60f8) was optimized for automatic processing WGS/WES data in various file formats, from mapping sequence reads to the latest human reference genome (GRCh38/hg38) and variant calling. The tracking database (available as AMI: ami-acc840d3) was designed for monitoring the job status and recording quality metrics for each processed sample (Figure 1B). With a dynamic web interface of the database, researchers can easily compare, share and visualize all these individual level quality metrics.

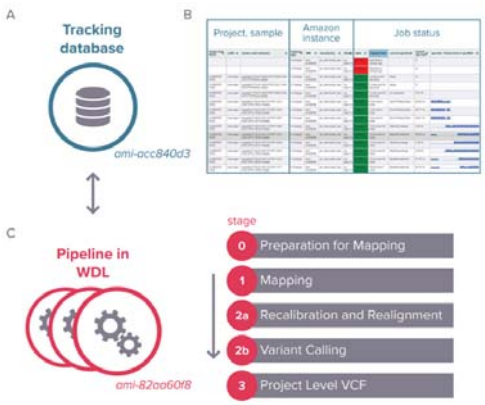

Figure 1: A) VCPA tracking database; B) Dynamic view of job status; C) VCPA Pipeline overview.

2 SNP/INDEL CALLING PIPELINE

The variant calling pipeline for the WGS (stages 1 and 2a) was developed in coordination with CCDG/TOPMed functional equivalent pipelines [2] and follows best practices of Germline Single Nucleotide Polymorphisms (SNPs) & Insertion/deletion (indel) Discovery for Genomic Analysis Toolkit (GATK) v3.7 [3]. VCPA pipeline is unique, as it accepts WGS or WES pair-end reads in FASTQ, BAM (binary sequence alignment map format), or CRAM (compressed BAM) formats with flow cell information and genomic regions for exome sequencing enrichment/capture kits. The processing workflow is modularized and consists of four stages (Figure 1C). If needed, the workflow can be configured to skip specific stages to reduce time and cost.

(1) Stage 0 includes preparation steps for read mapping. For samples already mapped previously, PICARD [8] is used to roll back BAM files to uBAM (unaligned BAM) files.

(2) Stage 1 generates BAM files. First, reads are mapped to GRCh38/hg38 using BWA-MEM [9], and duplicate reads are marked by BamUtil [10]. Next, BAM files are processed by Samblaster (adding MC and MQ tags to pair-end reads) [11] and sorted by genomic coordinates using SAMtools [12]. Finally, coverage statistics are computed using Sambamba [13].

(3) Stage 2A performs local realignment near known indel sites (1000 Genome indels) and recalibration of base call quality scores using GATK v3.7 [14].

(4) Stage 2B implements the GATK best practices steps for variant calling and annotation on SNPs and indels and generates genotype call files in genomic Variant Call Format (gVCF) for each sample individually. Quality metrics of called variants are computed using GATK [15].

(5) Stage 3 combines gVCF files from multiple samples and performs joint genotype calling using GATK best practices. A project-level VCF file is generated.

For each project, the user starts by preparing a manifest file of sample IDs and sequencing read file locations. The manifest file is then uploaded into the tracking database at a head node server (the database can track multiple projects), and the user can submit processing jobs for the samples individually or by batches by command line from the head node. Job dependency and error checking are implemented in the workflow. Whenever multi-threading is supported by the third-party programs, jobs are run in parallel to expedite the process. A complete run of the pipeline produces ~200 files including analysis-ready quality scored binned read alignment BAM/CRAM files, annotated gVCF files, and more than 140 log files with data quality metrics and run information.

3 TRACKING DATABASE

The tracking database enables the user to monitor production status (Figure 1B) and review sequencing quality such as mapping percentage, depth coverage, and quality of called variants. The quality metrics pages are defined by projects and pipeline stages. All 113 quality metrics are collected during the pipeline execution and imported into the database, and are viewable through an interactive web frontend display.

The tracking database is built on a LAMP (Linux, Apache Httpd, MySQL and PHP) application stack using the SLIM-PHP framework. The application has a small memory and storage footprint, provides a RESTful API interface to the MySQL back-end, and supports password protection to restrict access. The tracking database is dockerized and can be installed on-site (off the cloud) if preferred.

4 USING VCPA ON AMAZON EC2

We evaluated our pipeline using the NA12878 sample from the Genome In A Bottle project using the hg38 high confident set [15]. The gVCF from running VCPA on GIAB sample were compared against the GIAB truth variant calls using hap.py [16]. Sensitivity/precision of VCPA calls were 0.999/0.994 for SNPs and 0.985/0.987 for indels respectively and comparable to TOPMed/CCDG workflows [2]. Using Amazon EC2 instance type r3.8xlarge (244.0 GiB, 32 vCPUs), we benchmarked running time on WGS data of 9 ADSP Discovery and Discovery Extension phase samples (average 78 million reads) with different paired-read length and file types: 1) three BAMs of 100 bp reads; 2) three BAMs of 150bp reads and 3) three CRAMs of 150bp reads. Average processing time per genome was 26.43, 22.68 and 21.20 core-hours respectively for the three configurations/file-types.

To conclude, VSPA is an efficient pipeline for processing high quality WGS/WES data on Amazon EC2 environment. VCPA is used for ADSP production and can track and receive information from >1,000 genome analysis runs simultaneously. Future plans include incorporating other variant calling pipelines such as xAtlas [17] and GATK4.

Acknowledgements

We thank members of the Schellenberg and Wang labs, members from the GCAD and ADSP for their constructive input.

Funding

Approved users agree that the acknowledgment shall include the dbGaP accession number to the specific version of the dataset(s) analyzed. A sample statement for the acknowledgment of the ADSP dataset(s) follows: The Alzheimer's Disease Sequencing Project (ADSP) is comprised of two Alzheimer's Disease (AD) genetics consortia and three National Human Genome Research Institute (NHGRI) funded Large Scale Sequencing and Analysis Centers (LSAC). The two AD genetics consortia are the Alzheimer's Disease Genetics Consortium (ADGC) funded by NIA (U01 AG032984), and the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) funded by NIA (R01 AG033193), the National Heart, Lung, and Blood Institute (NHLBI), other National Institute of Health (NIH) institutes and other foreign governmental and non-governmental organizations. The Discovery Phase analysis of sequence data is supported through UF1AG047133 (to Drs. Schellenberg, Farrer, Pericak-Vance, Mayeux, and Haines); U01AG049505 to Dr. Seshadri; U01AG049506 to Dr. Boerwinkle; U01AG049507 to Dr. Wijsman; and U01AG049508 to Dr. Goate and the Discovery Extension Phase analysis is supported through U01AG052411 to Dr. Goate, U01AG052410 to Dr. Pericak-Vance and U01 AG052409 to Drs. Seshadri and Fornage. Data generation and harmonization in the Follow-up Phases is supported by U54AG052427 (to Drs. Schellenberg and Wang). The ADGC cohorts include: Adult Changes in Thought (ACT), the Alzheimer's Disease Centers (ADC), the Chicago Health and Aging Project (CHAP), the Memory and Aging Project (MAP), Mayo Clinic (MAYO), Mayo Parkinson's Disease controls, University of Miami, the Multi-Institutional Research in Alzheimer's Genetic Epidemiology Study (MIRAGE), the National Cell Repository for Alzheimer's Disease (NCRAD), the National Institute on Aging Late Onset Alzheimer's Disease Family Study (NIA-LOAD), the Religious Orders Study (ROS), the Texas Alzheimer's Research and Care Consortium (TARC), Vanderbilt University/Case Western Reserve University (VAN/CWRU), the Washington Heights-Inwood Columbia Aging Project (WHICAP) and the Washington University Sequencing Project (WUSP), the Columbia University Hispanic- Estudio Familiar de Influencia Genetica de Alzheimer (EFIGA), the University of Toronto (UT), and Genetic Differences (GD). The CHARGE cohorts are supported in part by National Heart, Lung, and Blood Institute (NHLBI) infrastructure grant HL105756 (Psaty), RC2HL102419 (Boerwinkle) and the neurology working group is supported by the National Institute on Aging (NIA) R01 grant AG033193. The CHARGE cohorts participating in the ADSP include the following: Austrian Stroke Prevention Study (ASPS), ASPS-Family study, and the Prospective Dementia Registry-Austria (ASPS/PRODEM-Aus), the Atherosclerosis Risk in Communities (ARIC) Study, the Cardiovascular Health Study (CHS), the Erasmus Rucphen Family Study (ERF), the Framingham Heart Study (FHS), and the Rotterdam Study (RS). ASPS is funded by the Austrian Science Fond (FWF) grant number P20545-P05 and P13180 and the Medical University of Graz. The ASPS-Fam is funded by the Austrian Science Fund (FWF) project I904),the EU Joint Programme - Neurodegenerative Disease Research (JPND) in frame of the BRIDGET project (Austria, Ministry of Science) and the Medical University of Graz and the Steiermärkische Krankenanstalten Gesellschaft. PRODEM-Austria is supported by the Austrian Research Promotion agency (FFG) (Project No. 827462) and by the Austrian National Bank (Anniversary Fund, project 15435. ARIC research is carried out as a collaborative study supported by NHLBI contracts (HHSN268201100005C, HHSN268201100006C, HHSN268201100007C, HHSN268201100008C, HHSN268201100009C, HHSN268201100010C, HHSN268201100011C, and HHSN268201100012C). Neurocognitive data in ARIC is collected by U01 2U01HL096812, 2U01HL096814, 2U01HL096899, 2U01HL096902, 2U01HL096917 from the NIH (NHLBI, NINDS, NIA and NIDCD), and with previous brain MRI examinations funded by R01-HL70825 from the NHLBI. CHS research was supported by contracts HHSN268201200036C, HHSN268200800007C, N01HC55222, N01HC85079, N01HC85080, N01HC85081, N01HC85082, N01HC85083, N01HC85086, and grants U01HL080295 and U01HL130114 from the NHLBI with additional contribution from the National Institute of Neurological Disorders and Stroke (NINDS). Additional support was provided by R01AG023629, R01AG15928, and R01AG20098 from the NIA. FHS research is supported by NHLBI contracts N01-HC-25195 and HHSN268201500001I. This study was also supported by additional grants from the NIA (R01s AG054076, AG049607 and AG033040 and NINDS (R01 NS017950). The ERF study as a part of EUROSPAN (European Special Populations Research Network) was supported by European Commission FP6 STRP grant number 018947 (LSHG-CT-2006-01947) and also received funding from the European Community's Seventh Framework Programme (FP7/2007-2013)/grant agreement HEALTH-F4-2007-201413 by the European Commission under the programme "Quality of Life and Management of the Living Resources" of 5th Framework Programme (no. QLG2-CT-2002-01254). High-throughput analysis of the ERF data was supported by a joint grant from the Netherlands Organization for Scientific Research and the Russian Foundation for Basic Research (NWO-RFBR 047.017.043). The Rotterdam Study is funded by Erasmus Medical Center and Erasmus University, Rotterdam, the Netherlands Organization for Health Research and Development (ZonMw), the Research Institute for Diseases in the Elderly (RIDE), the Ministry of Education, Culture and Science, the Ministry for Health, Welfare and Sports, the European Commission (DG XII), and the municipality of Rotterdam. Genetic data sets are also supported by the Netherlands Organization of Scientific Research NWO Investments (175.010.2005.011, 911-03-012), the Genetic Laboratory of the Department of Internal Medicine, Erasmus MC, the Research Institute for Diseases in the Elderly (014-93-015; RIDE2), and the Netherlands Genomics Initiative (NGI)/Netherlands Organization for Scientific Research (NWO) Netherlands Consortium for Healthy Aging (NCHA), project 050-060-810. All studies are grateful to their participants, faculty and staff. The content of these manuscripts is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or the U.S. Department of Health and Human Services. The three LSACs are: the Human Genome Sequencing Center at the Baylor College of Medicine (U54 HG003273), the Broad Institute Genome Center (U54HG003067), and the Washington University Genome Institute (U54HG003079). Biological samples and associated phenotypic data used in primary data analyses were stored at Study Investigators institutions, and at the National Cell Repository for Alzheimer's Disease (NCRAD, U24AG021886) at Indiana University funded by NIA. Associated Phenotypic Data used in primary and secondary data analyses were provided by Study Investigators, the NIA funded Alzheimer's Disease Centers (ADCs), and the National Alzheimer's Coordinating Center (NACC, U01AG016976) and the National Institute on Aging Genetics of Alzheimer's Disease Data Storage Site (NIAGADS, U24AG041689) at the University of Pennsylvania, funded by NIA, and at the Database for Genotypes and Phenotypes (dbGaP) funded by NIH. This research was supported in part by the Intramural Research Program of the National Institutes of health, National

Library of Medicine. Contributors to the Genetic Analysis Data included Study Investigators on projects that were individually funded by NIA, and other NIH institutes, and by private U.S. organizations, or foreign governmental or non-governmental organizations.

Conflict of Interest: none declared.

References

Hebert LE1, Weuve J, Scherr PA, Evans DA. Alzheimer disease in the United States (2010-2050) estimated using the 2010 census. Neurology. 2013 May 7;80(19):1778-83. doi: 10.1212/WNL.0b013e31828726f5. PMID: 23390181 PMCID: PMC3719424.

Allison A Regier, Yossi Farjoun, David Larson, Olga Krasheninina, Hyun Min Kang, Daniel P Howrigan, Bo-Juen Chen, Manisha Kher, Eric Banks, Darren C Ames, Adam C English, Heng Li, The NHLBI Trans-Omics for Precision Medicine (TOPMed) Program, Goncalo R Abecasis, Will Salerno, Michael C Zody, Benjamin M Neale, Ira M Hall Functional equivalence of genome sequencing analysis pipelines enables harmonized variant calling across human genetics projects doi: https://doi.org/10.1101/269316 https://www.biorxiv.org/content/early/2018/02/22/269316

DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell TJ, Kernytsky AM, Sivachenko AY, Cibulskis K, Gabriel SB, Altshuler D, Daly MJ. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011 May;43(5):491-8. doi: 10.1038/ng.806. PMID: 21478889 PMCID: PMC3083463.

https://gatkforums.broadinstitute.org/gatk/discussion/11008/ubam-unmapped-bam-format

https://samtools.github.io/hts-specs/CRAMv3.pdf

Voss K, Gentry J and Van der Auwera G. Full-stack genomics pipelining with GATK4 + WDL + Cromwell [version 1; not peer reviewed]. F1000Research 2017, 6(ISCB Comm J):1379 (poster) (doi: 10.7490/f1000research.1114631.1)

https://storage.cloud.google.com/genomics-public-da-ta/resources/broad/hg38/v0/

http://broadinstitute.github.io/picard

Li H. (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997v1 [q-bio.GN].

https://genome.sph.umich.edu/wiki/BamUtil

Faust GG, Hall IM. SAMBLASTER: fast duplicate marking and structural variant read extraction. Bioinformatics. 2014 Sep 1;30(17):2503-5. doi: 10.1093/bioinformatics/btu314. PMID: 24812344 PMCID: PMC4147885.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R; 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009 Aug 15;25(16):2078-9. doi: 10.1093/bioinformatics/btp352. PMID: 19505943 PMCID: PMC2723002.

Tarasov A, Vilella AJ, Cuppen E, Nijman IJ, Prins P. Sambamba: fast processing of NGS alignment formats. Bioinformatics. 2015 Jun 15;31(12):2032-4. doi: 10.1093/bioinformatics/btv098. PMID: 25697820 PMCID: PMC4765878.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010 Sep;20(9):1297-303. doi: 10.1101/gr.107524.110. PMID: 20644199 PMCID: PMC2928508.

ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/release/NA12878_HG001/latest/GRCh38/

https://github.com/Illumina/hap.py

Challis D, Yu J, Evani US, Jackson AR, Paithankar S, Coarfa C, Milosavljevic A, Gibbs RA, Yu F. An integrative variant analysis suite for whole exome next-generation sequencing data. BMC Bioinformatics. 2012 Jan 12;13:8. doi: 10.1186/1471-2105-13-8. PMID: 22239737 PMCID: PMC3292476.