A Cloud-Based Genotyping Analysis Pipeline

GEORGE GRANT, JASON CAREY, KATHLEEN TIBBETTS

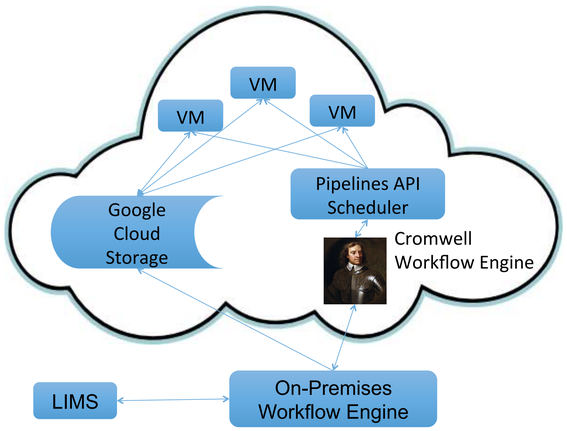
Broad Institute Genomics, Cambridge, MA.

Introduction

The use of genotyping arrays remains a cost-effective and valuable technique for validation of sequencing results and for running large-scale GWAS studies. At the Broad Institute we have developed a highly scalable, cloud-based genotyping array data analysis pipeline to facilitate our continued use of Illumina Sentrix genotyping arrays. This workflow follows on to and uses tools and strategies developed for our cloud-based Whole Genome Shotgun (WGS) workflow.

On-Premises to Cloud Workflow

The core of the genotyping array data analysis pipeline is a single-sample workflow, written in WDL, the Workflow Definition Language. This workflow is run on Cromwell (a Broad-developed workflow-execution engine) using Google's Pipelines API as its back-end.



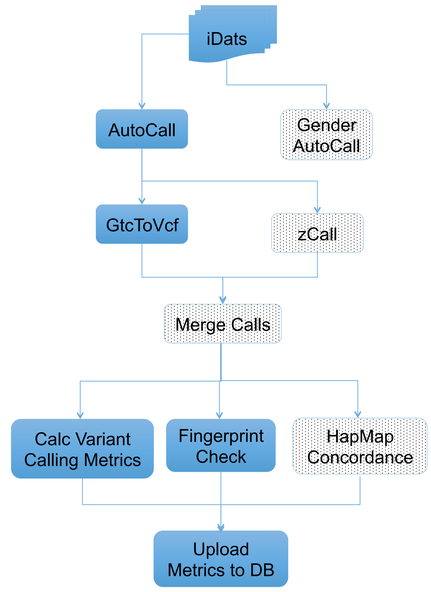
The single-sample workflow is kicked off automatically by a message from the Broad LIMS which contains information about the segment of the chip to be analyzed. This message is received and processed by a legacy on-premises workflow execution system (Zamboni) that does initial validation and processing and then copies appropriate data files (Illumina IDAT files, chip and sample metadata) to Google Cloud Storage (GCS) in order for them to be used by the cloud-based workflow.

The Single-Sample Genotyping Workflow

The single-sample genotyping array workflow is currently implemented for Illumina Sentrix chips and performs genotype calling using Illumina's AutoCall genotype calling software. The output of this is then run through several other tools (see workflow diagram) until the final output, a fully annotated, single-sample VCF is generated. This VCF contains complete information about the variants called by AutoCall and zCall, plus additional metadata (e.g. normalized intensity, quality scores, clustering information) that can be used by other tools.

Key Steps in the Single-Sample Arrays Workflow:

- AutoCall. An Illumina-provided genotype calling algorithm. This is the same tool that is used in Illumina's LIMS. It is used in our workflow for primary calling and (optionally, with a different cluster file) to determine gender on chips with high rare variant content on the gender chromosomes.
- GtcToVcf. An internal tool to convert the GTC files output by Autocall to VCF format. This tool uses information from the Illumina Manifest file, cluster file and dbSnp database to create a highly annotated file in the standard VCF format with information about all validated probes on the chip.
- zCall. A tool developed at the Broad to optimize genotype calls on chips with significant rare variant content (optional).
- MergeCalls. An internal tool to combine AutoCall and zCall generated genotype calls into a standard format in the VCF (optional).
- Calculate Variant Calling Metrics. An internal set of tools to generate QC and variant calling metrics from the Single-Sample VCF and upload these to a cloud SQL database.

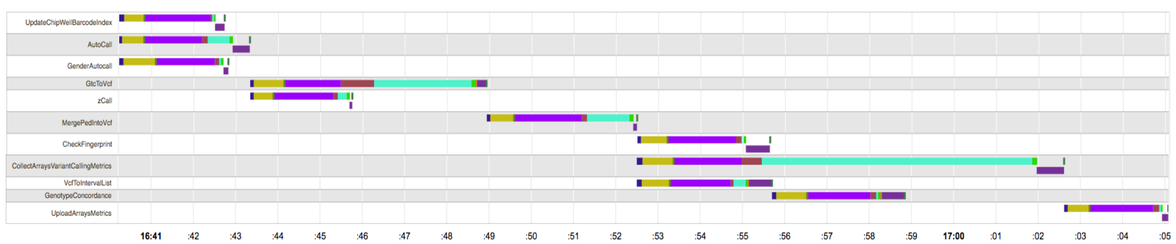


Single Sample Arrays Workflow

Running Tasks on the Cloud

Our workflows are written in WDL, which has allowed us to break down the steps of the single-sample workflow into a series of interdependent tasks with distinct inputs, outputs, resource requirements and command line invocations. These tasks are run in docker images running on Google VMs (virtual machines).

This architecture allows us to more easily scale our use of compute and disk to match the demands of the production environment. In addition, we can easily designate VMs with different resources (i.e. more cores, memory or disk) for hungrier tasks while assigning other tasks to lighter, more cost-effective VMs.



A Cromwell Timing Diagram of the Single Sample Arrays Workflow. Different colored segments denote different portions of a workflow task's execution process. i.e. Cyan = 'Running Docker'

Data Delivery – Multi-Sample VCFs to PLINK Ready Inputs

In order to facilitate the delivery of large data sets, a second cloud workflow is used to combine the single-sample VCFs into a multi-sample VCF, as run through the analysis pipeline. In addition to this VCF, metadata files are generated containing additional sample information, thus giving analysts all pertinent information.

These files can be run through VCFtools, either on-premises, or through FireCloud in order to generate bed/bim/fam files for input to PLINK.

The final outputs of the workflows are stored in the cloud. Thus delivery of data to customer accounts on the cloud is extremely straightforward. Simply use existing gcloud tools to copy the files to configured FireCloud Workspaces.

Future Improvements

Currently our cloud workflow implementation uses Google Cloud Services, we plan to extend this to other cloud resource providers, including Amazon Web Services.

The VCF generated by the single-sample workflow contains all sample-specific genotyping information in addition to information about the genotyping assay, as called by AutoCall, this includes normalized and raw intensity, B allele frequency and log R ratio. We plan to make use of this data – especially for copy number calling in future workflows.

References

WDL: https://github.com/broadinstitute/wdl

Cromwell: https://github.com/broadinstitute/cromwell

zCall: https://github.com/jigold/zCall

Acknowledgments

Thanks to:

Ben Neale and Jackie Goldstein for invaluable suggestions for data content and format.

The Green Team in DSDE (Jose Soto, Dave Shiga, Kylee Degatono, Brad Taylor and Nicole Bolinger).