

VAR-Seq Workflow Template

First/last name (first.last@ucr.edu)

Last update: 09 May, 2018

Introduction

Users want to provide here background information about the design of their VAR-Seq project.

Background and objectives

This report describes the analysis of a VAR-Seq project studying the genetic differences among several strains ... from *organism*

Experimental design

Typically, users want to specify here all information relevant for the analysis of their NGS study. This includes detailed descriptions of FASTQ files, experimental design, reference genome, gene annotations, etc.

Generate workflow environment

Load workflow environment with sample data into your current working directory. The sample data are described here.

In the workflow environments generated by `genWorkenvir` all data inputs are stored in a `data/` directory and all analysis results will be written to a separate `results/` directory, while the `systemPipeVARseq.Rmd` script and the `targets` file are expected to be located in the parent directory. The R session is expected to run from this parent directory. Additional parameter files are stored under `param/`.

To work with real data, users want to organize their own data similarly and substitute all test data for their own data. To rerun an established workflow on new data, the initial `targets` file along with the corresponding FASTQ files are usually the only inputs the user needs to provide.

```
library(systemPipeRdata)
genWorkenvir(workflow="varseq")
setwd("varseq")
```

Alternatively, this can be done from the command-line as follows:

```
$ Rscript -e "systemPipeRdata::genWorkenvir(workflow='varseq')"
$ cd varseq
```

Now download the latest `systemPipeVARseq.Rmd` script for this course. From within R this can be done as follows.

```
download.file("https://raw.githubusercontent.com/tgirke/GEN242/gh-pages/_vignettes/13_VARseqWorkflow/sy
```

Or from the command-line one can do this with `wget`.

```
$ wget -O systemPipeVARseq.Rmd https://raw.githubusercontent.com/tgirke/GEN242/gh-pages/_vignettes/13_V
```

Now log in to a computer node on the HPC/biocluster. The following command sequence will connect the user from the command-line to a computer node on the cluster.

```
$ srun --x11 --partition=short --mem=2gb --cpus-per-task 1 --ntasks 1 --time 2:00:00 --pty bash -l
```

Load desired R version from module system (here R-3.4.2).

```
$ module load R/3.4.2
```

Now open the R markdown script `systemPipeVARseq.Rmd` in your R IDE (*e.g.* nvim-r or RStudio) and run the workflow as outlined below.

Note, Tmux sessions should always run on one of the headnodes and never on the computer nodes themselves. This is important since Tmux sessions are persistent meaning they don't close automatically when a computer job finishes. Thus, they are not controlled by the queueing system.

To check the environment of R session, one can execute the following commands from R. The first line returns the node name of the R session.

```
system("hostname") # should return name of a compute node starting with i or c
getwd() # checks current working directory of R session
dir() # returns content of current working directory
```

Required packages and resources

The `systemPipeR` package needs to be loaded to perform the analysis steps shown in this report (H Backman and Girke 2016).

```
library(systemPipeR)
```

If applicable users can load custom functions not provided by `systemPipeR`. Skip this step if this is not the case.

```
source("systemPipeVARseq_Fct.R")
```

Read preprocessing

Experiment definition provided by targets file

The `targets` file defines all FASTQ files and sample comparisons of the analysis workflow.

```
targetspath <- system.file("extdata", "targetsPE.txt", package="systemPipeR")
targets <- read.delim(targetspath, comment.char = "#")
targets[, -c(5,6)]
```

##	FileName1	FileName2	SampleName	Factor	Date
## 1	./data/SRR446027_1.fastq	./data/SRR446027_2.fastq	M1A	M1	23-Mar-2012
## 2	./data/SRR446028_1.fastq	./data/SRR446028_2.fastq	M1B	M1	23-Mar-2012
## 3	./data/SRR446029_1.fastq	./data/SRR446029_2.fastq	A1A	A1	23-Mar-2012
## 4	./data/SRR446030_1.fastq	./data/SRR446030_2.fastq	A1B	A1	23-Mar-2012
## 5	./data/SRR446031_1.fastq	./data/SRR446031_2.fastq	V1A	V1	23-Mar-2012
## 6	./data/SRR446032_1.fastq	./data/SRR446032_2.fastq	V1B	V1	23-Mar-2012
## 7	./data/SRR446033_1.fastq	./data/SRR446033_2.fastq	M6A	M6	23-Mar-2012
## 8	./data/SRR446034_1.fastq	./data/SRR446034_2.fastq	M6B	M6	23-Mar-2012
## 9	./data/SRR446035_1.fastq	./data/SRR446035_2.fastq	A6A	A6	23-Mar-2012
## 10	./data/SRR446036_1.fastq	./data/SRR446036_2.fastq	A6B	A6	23-Mar-2012

```
## 11 ./data/SRR446037_1.fastq ./data/SRR446037_2.fastq V6A V6 23-Mar-2012
## 12 ./data/SRR446038_1.fastq ./data/SRR446038_2.fastq V6B V6 23-Mar-2012
## 13 ./data/SRR446039_1.fastq ./data/SRR446039_2.fastq M12A M12 23-Mar-2012
## 14 ./data/SRR446040_1.fastq ./data/SRR446040_2.fastq M12B M12 23-Mar-2012
## 15 ./data/SRR446041_1.fastq ./data/SRR446041_2.fastq A12A A12 23-Mar-2012
## 16 ./data/SRR446042_1.fastq ./data/SRR446042_2.fastq A12B A12 23-Mar-2012
## 17 ./data/SRR446043_1.fastq ./data/SRR446043_2.fastq V12A V12 23-Mar-2012
## 18 ./data/SRR446044_1.fastq ./data/SRR446044_2.fastq V12B V12 23-Mar-2012
```

Read quality filtering and trimming

The following removes reads with low quality base calls (here Phred scores below 20) from all FASTQ files.

```
args <- systemArgs(sysma="param/trimPE.param", mytargets="targetsPE.txt")[1:4] # Note: subsetting!
filterFct <- function(fq, cutoff=20, Nexceptions=0) {
  qccount <- rowSums(as(quality(fq), "matrix") <= cutoff)
  fq[qccount <= Nexceptions] # Retains reads where Phred scores are >= cutoff with N exceptions
}
preprocessReads(args=args, Fct="filterFct(fq, cutoff=20, Nexceptions=0)", batchsize=100000)
writeTargetsout(x=args, file="targets_PTrim.txt", overwrite=TRUE)
```

FASTQ quality report

The following `seeFastq` and `seeFastqPlot` functions generate and plot a series of useful quality statistics for a set of FASTQ files including per cycle quality box plots, base proportions, base-level quality trends, relative k-mer diversity, length and occurrence distribution of reads, number of reads above quality cutoffs and mean quality distribution. The results are written to a PDF file named `fastqReport.pdf`.

```
args <- systemArgs(sysma="param/tophat.param", mytargets="targets.txt")
fqlist <- seeFastq(fastq=infile1(args), batchsize=100000, klength=8)
pdf("./results/fastqReport.pdf", height=18, width=4*length(fqlist))
seeFastqPlot(fqlist)
dev.off()
```

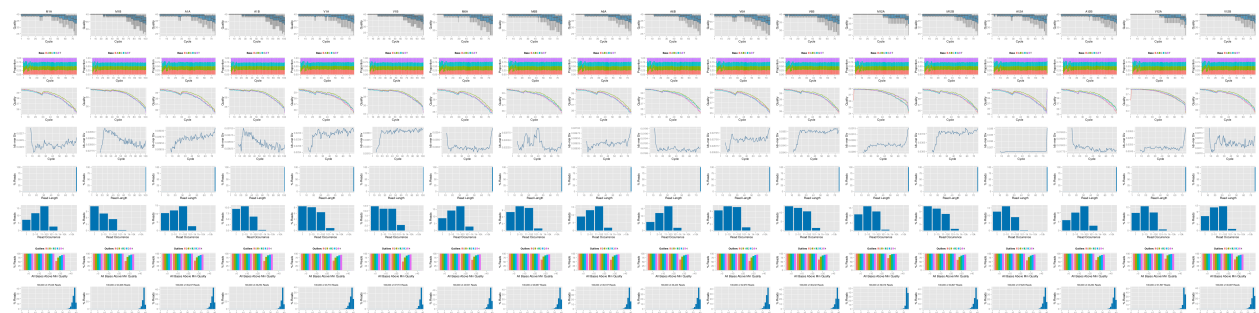


Figure 1: FASTQ quality report for 18 samples

Alignments

Read mapping with BWA-MEM

The NGS reads of this project are aligned against the reference genome sequence using the highly variant tolerant short read aligner BWA-MEM (Heng Li 2013; H Li and Durbin 2009). The parameter settings of the aligner are defined in the `bwa.param` file.

```
args <- systemArgs(sysma="param/bwa.param", mytargets="targets.txt")
sysargs(args)[1] # Command-line parameters for first FASTQ file
```

Runs the alignments sequentially (e.g. on a single machine)

```
moduleload(modules(args))
system("bwa index -a bwtsw ./data/tair10.fasta")
bampaths <- runCommandline(args=args)
writeTargetsout(x=args, file="targets_bam.txt", overwrite=TRUE)
```

Alternatively, the alignment jobs can be submitted to a compute cluster, here using 72 CPU cores (18 qsub processes each with 4 CPU cores).

```
moduleload(modules(args))
system("bwa index -a bwtsw ./data/tair10.fasta")
resources <- list(walltime="1:00:00", ntasks=1, ncpus=cores(args), memory="10G")
reg <- clusterRun(args, conffile=".BatchJobs.R", template="slurm.tmpl", Njobs=18, runid="01",
                  resourceList=resources)
waitForJobs(reg)
writeTargetsout(x=args, file="targets_bam.txt", overwrite=TRUE)
```

Check whether all BAM files have been created

```
file.exists(outpaths(args))
```

Read and alignment stats

The following generates a summary table of the number of reads in each sample and how many of them aligned to the reference.

```
read_statsDF <- alignStats(args=args)
write.table(read_statsDF, "results/alignStats.xls", row.names=FALSE, quote=FALSE, sep="\t")
```

Create symbolic links for viewing BAM files in IGV

The `symLink2bam` function creates symbolic links to view the BAM alignment files in a genome browser such as IGV. The corresponding URLs are written to a file with a path specified under `urlfile`, here `IGVurl.txt`.

```
symLink2bam(sysargs=args, htmdir=c("~/html/", "projects/gen242/"),
            urlbase="http://biocluster.ucr.edu/~tgirke/",
            urlfile="./results/IGVurl.txt")
```

Variant calling

The following performs variant calling with **GATK**, **BCFtools** and **VariantTools** in parallel mode on a compute cluster (McKenna et al. 2010; Heng Li 2011). If a cluster is not available, the `runCommandline` function can be used to run the variant calling with **GATK** and **BCFtools** for each sample sequentially on a single machine, or `callVariants` in case of **VariantTools**. Typically, the user would choose here only one variant caller rather than running several ones.

Variant calling with GATK

The following creates in the initial step a new `targets` file (`targets_bam.txt`). The first column of this file gives the paths to the BAM files created in the alignment step. The new targets file and the parameter file `gatk.param` are used to create a new `SYSargs` instance for running GATK. Since GATK involves many processing steps, it is executed by a bash script `gatk_run.sh` where the user can specify the detailed run parameters. All three files are expected to be located in the current working directory. Samples files for `gatk.param` and `gatk_run.sh` are available in the `param` subdirectory provided by `systemPipeRdata`.

```
moduleload("picard/1.130"); moduleload("samtools/1.3")
system("picard CreateSequenceDictionary R=./data/tair10.fasta O=./data/tair10.dict")
system("samtools faidx data/tair10.fasta")
args <- systemArgs(sysma="param/gatk.param", mytargets="targets_bam.txt")
resources <- list(walltime="00:20:00", ntasks=1, ncpus=1, memory="6G")
reg <- clusterRun(args, conffile=".BatchJobs.R", template="slurm.tmpl", Njobs=18, runid="01",
  resourceList=resources)
waitForJobs(reg)
# unlink(outfile1(args), recursive = TRUE, force = TRUE)
writeTargetsout(x=args, file="targets_gatk.txt", overwrite=TRUE)
```

Variant calling with BCFtools

The following runs the variant calling with **BCFtools**. This step requires in the current working directory the parameter file `sambcf.param` and the bash script `sambcf_run.sh`.

```
args <- systemArgs(sysma="param/sambcf.param", mytargets="targets_bam.txt")
resources <- list(walltime="00:20:00", ntasks=1, ncpus=1, memory="6G")
reg <- clusterRun(args, conffile=".BatchJobs.R", template="slurm.tmpl", Njobs=18, runid="01",
  resourceList=resources)
waitForJobs(reg)
# unlink(outfile1(args), recursive = TRUE, force = TRUE)
writeTargetsout(x=args, file="targets_sambcf.txt", overwrite=TRUE)
```

Inspect VCF file

VCF files can be imported into R with the `readVcf` function. Both `VCF` and `VRanges` objects provide convenient data structure for working with variant data (e.g. SNP quality filtering).

```
library(VariantAnnotation)
args <- systemArgs(sysma="param/filter_gatk.param", mytargets="targets_gatk.txt")
vcf <- readVcf(infile1(args)[1], "A. thaliana")
vcf
vr <- as(vcf, "VRanges")
vr
```

Filter variants

The function `filterVars` filters VCF files based on user definable quality parameters. It sequentially imports each VCF file into R, applies the filtering on an internally generated `VRanges` object and then writes the results to a new subsetted VCF file. The filter parameters are passed on to the corresponding argument as a character string. The function applies this filter to the internally generated `VRanges` object using the standard subsetting syntax for two dimensional objects such as: `vr[filter,]`. The parameter files (`filter_gatk.param`, `filter_sambcf.param` and `filter_vartools.param`), used in the filtering steps, define the paths to the input and output VCF files which are stored in new `SYSargs` instances.

Filter variants called by GATK

The below example filters for variants that are supported by $\geq x$ reads and $\geq 80\%$ of them support the called variants. In addition, all variants need to pass $\geq x$ of the soft filters recorded in the VCF files generated by GATK. Since the toy data used for this workflow is very small, the chosen settings are unreasonably relaxed. A more reasonable filter setting is given in the line below (here commented out).

```
library(VariantAnnotation)
library(BBmisc) # Defines suppressAll()
args <- systemArgs(sysma="param/filter_gatk.param", mytargets="targets_gatk.txt")[1:4]
filter <- "totalDepth(vr) >= 2 & (altDepth(vr) / totalDepth(vr) >= 0.8) & rowSums(softFilterMatrix(vr)) >= 2"
# filter <- "totalDepth(vr) >= 20 & (altDepth(vr) / totalDepth(vr) >= 0.8) & rowSums(softFilterMatrix(vr)) >= 2"
suppressAll(filterVars(args, filter, varcaller="gatk", organism="A. thaliana"))
writeTargetsout(x=args, file="targets_gatk_filtered.txt", overwrite=TRUE)
```

Filter variants called by BCFtools

The following shows how to filter the VCF files generated by `BCFtools` using similar parameter settings as in the previous filtering of the GATK results.

```
args <- systemArgs(sysma="param/filter_sambcf.param", mytargets="targets_sambcf.txt")[1:4]
filter <- "rowSums(vr) >= 2 & (rowSums(vr[,3:4])/rowSums(vr[,1:4]) >= 0.8)"
# filter <- "rowSums(vr) >= 20 & (rowSums(vr[,3:4])/rowSums(vr[,1:4]) >= 0.8)"
suppressAll(filterVars(args, filter, varcaller="bcftools", organism="A. thaliana"))
writeTargetsout(x=args, file="targets_sambcf_filtered.txt", overwrite=TRUE)
```

Check filtering outcome for one sample

```
length(as(readVcf(infile1(args)[1], genome="Ath"), "VRanges")[,1])
length(as(readVcf(outpaths(args)[1], genome="Ath"), "VRanges")[,1])
```

Annotate filtered variants

The function `variantReport` generates a variant report using utilities provided by the `VariantAnnotation` package. The report for each sample is written to a tabular file containing genomic context annotations (*e.g.* coding or non-coding SNPs, amino acid changes, IDs of affected genes, etc.) along with confidence statistics for each variant. The parameter file `annotate_vars.param` defines the paths to the input and output files which are stored in a new `SYSargs` instance.

Basics of annotating variants

Variants overlapping with common annotation features can be identified with `locateVariants`.

```
library("GenomicFeatures")
args <- systemArgs(sysma="param/annotate_vars.param", mytargets="targets_gatk_filtered.txt")
txdb <- loadDb("./data/tair10.sqlite")
vcf <- readVcf(infile1(args)[1], "A. thaliana")
locateVariants(vcf, txdb, CodingVariants())
```

Synonymous/non-synonymous variants of coding sequences are computed by the `predictCoding` function for variants overlapping with coding regions.

```
fa <- FaFile(systemPipeR::reference(args))
predictCoding(vcf, txdb, seqSource=fa)
```

Annotate filtered variants called by GATK

```
library("GenomicFeatures")
args <- systemArgs(sysma="param/annotate_vars.param", mytargets="targets_gatk_filtered.txt")
txdb <- loadDb("./data/tair10.sqlite")
fa <- FaFile(systemPipeR::reference(args))
suppressAll(variantReport(args=args, txdb=txdb, fa=fa, organism="A. thaliana"))
```

Annotate filtered variants called by BCFtools

```
args <- systemArgs(sysma="param/annotate_vars.param", mytargets="targets_sambcf_filtered.txt")
txdb <- loadDb("./data/tair10.sqlite")
fa <- FaFile(systemPipeR::reference(args))
suppressAll(variantReport(args=args, txdb=txdb, fa=fa, organism="A. thaliana"))
```

View annotation result for single sample

```
sampleDF <- read.delim(outpaths(args)[1])[38:40,]
sampleDF[sampleDF$LOCATION=="coding",]
```

Combine annotation results among samples

To simplify comparisons among samples, the `combineVarReports` function combines all variant annotation reports referenced in a `SYSargs` instance (here `args`). At the same time the function allows to consider only certain feature types of interest. For instance, the below setting `filtercol=c(Consequence="nonsynonymous")` will include only nonsynonymous variances listed in the `Consequence` column of the annotation reports. To omit filtering, one can use the setting `filtercol="All"`.

Combine results from GATK

```
args <- systemArgs(sysma="param/annotate_vars.param", mytargets="targets_gatk_filtered.txt")
combinedDF <- combineVarReports(args, filtercol=c(Consequence="nonsynonymous"))
write.table(combinedDF, "./results/combinedDF_nonsyn_gatk.xls", quote=FALSE, row.names=FALSE, sep="\t")
```


Combine results from BCFtools

```
args <- systemArgs(sysma="param/annotate_vars.param", mytargets="targets_sambcf_filtered.txt")
combinedDF <- combineVarReports(args, filtercol=c(Consequence="nonsynonymous"))
write.table(combinedDF, "./results/combinedDF_nonsyn_sambcf.xls", quote=FALSE, row.names=FALSE, sep="\t")
```

Summary statistics of variants

The `varSummary` function counts the number of variants for each feature type included in the annotation reports.

Summary for GATK

```
args <- systemArgs(sysma="param/annotate_vars.param", mytargets="targets_gatk_filtered.txt")
varSummary(args)
write.table(varSummary(args), "./results/variantStats_gatk.xls", quote=FALSE, col.names = NA, sep="\t")
```

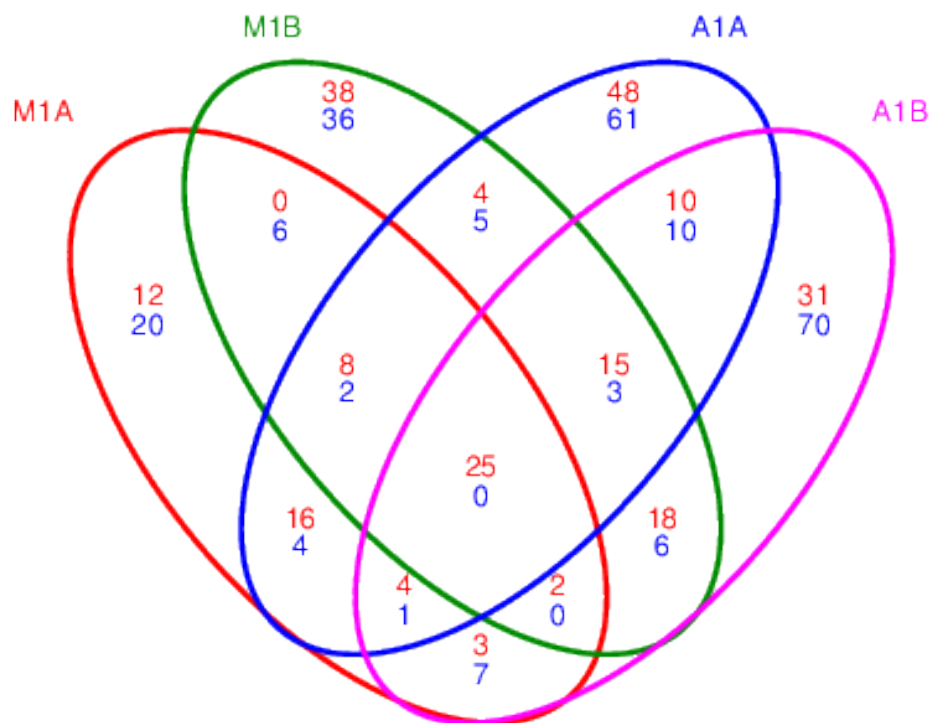
Summary for BCFtools

```
args <- systemArgs(sysma="param/annotate_vars.param", mytargets="targets_sambcf_filtered.txt")
varSummary(args)
write.table(varSummary(args), "./results/variantStats_sambcf.xls", quote=FALSE, col.names = NA, sep="\t")
```

Venn diagram of variants

The venn diagram utilities defined by the `systemPipeR` package can be used to identify common and unique variants reported for different samples and/or variant callers. The below generates a 4-way venn diagram comparing four samples for each of the two variant callers.

```
args <- systemArgs(sysma="param/annotate_vars.param", mytargets="targets_gatk_filtered.txt")
varlist <- sapply(names(outpaths(args))[1:4], function(x) as.character(read.delim(outpaths(args)[x])$VAR))
vennset_gatk <- overLapper(varlist, type="vennsets")
args <- systemArgs(sysma="param/annotate_vars.param", mytargets="targets_sambcf_filtered.txt")
varlist <- sapply(names(outpaths(args))[1:4], function(x) as.character(read.delim(outpaths(args)[x])$VAR))
vennset_bcf <- overLapper(varlist, type="vennsets")
pdf("./results/vennplot_var.pdf")
vennPlot(list(vennset_gatk, vennset_bcf), mymain="", mysub="GATK: red; BCFtools: blue", colmode=2, ccol=2,
dev.off()
```

GATK: red; BCFtools: blue

Figure 2: Venn Diagram for 4 samples from GATK and BCFtools

Plot variants programmatically

The following plots a selected variant with ggbio.

```
library(ggbio)
mychr <- "ChrC"; mystart <- 11000; myend <- 13000
args <- systemArgs(sysma="param/bwa.param", mytargets="targets.txt")
ga <- readGAlignments(outpaths(args)[1], use.names=TRUE, param=ScanBamParam(which=GRanges(mychr, IRanges(mystart, myend))))
p1 <- autoplot(ga, geom = "rect")
p2 <- autoplot(ga, geom = "line", stat = "coverage")
p3 <- autoplot(vcf[seqnames(vcf)==mychr], type = "fixed") + xlim(mystart, myend) + theme(legend.position = "bottom")
```

```
p4 <- autoplot(txdb, which=GRanges(mychr, IRanges(mystart, myend)), names.expr = "gene_id")
png("./results/plot_variant.png")
tracks(Reads=p1, Coverage=p2, Variant=p3, Transcripts=p4, heights = c(0.3, 0.2, 0.1, 0.35)) + ylab("")
dev.off()
```

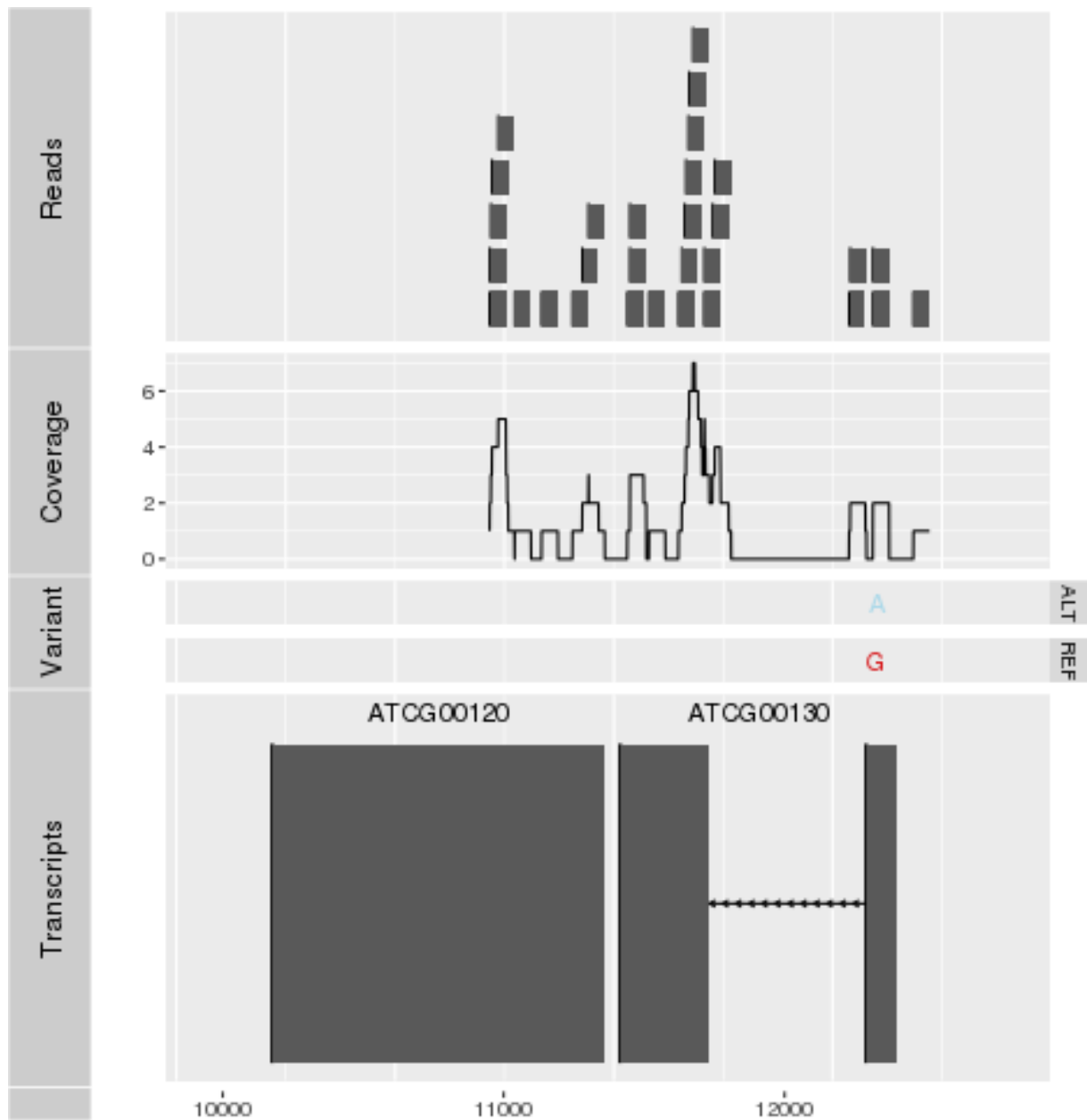


Figure 3: Plot variants with programmatically.

Version Information

```
sessionInfo()
```

```
## R version 3.4.0 (2017-04-21)
```

```

## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: Ubuntu 14.04.5 LTS
##
## Matrix products: default
## BLAS: /usr/lib/libblas/libblas.so.3.0
## LAPACK: /usr/lib/lapack/liblapack.so.3.0
##
## locale:
## [1] LC_CTYPE=en_US.UTF-8      LC_NUMERIC=C              LC_TIME=en_US.UTF-8
## [4] LC_COLLATE=en_US.UTF-8    LC_MONETARY=en_US.UTF-8   LC_MESSAGES=en_US.UTF-8
## [7] LC_PAPER=en_US.UTF-8      LC_NAME=C                 LC_ADDRESS=C
## [10] LC_TELEPHONE=C           LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C
##
## attached base packages:
## [1] stats4      parallel  methods    stats      graphics  utils      datasets  grDevices  base
##
## other attached packages:
## [1] ape_4.1                ggplot2_2.2.1            systemPipeR_1.10.0
## [4] ShortRead_1.34.0       GenomicAlignments_1.12.0 SummarizedExperiment_1.6.0
## [7] DelayedArray_0.2.0     matrixStats_0.52.2       Biobase_2.36.0
## [10] BiocParallel_1.10.0    Rsamtools_1.28.0         Biostrings_2.44.0
## [13] XVector_0.16.0         GenomicRanges_1.28.0     GenomeInfoDb_1.12.0
## [16] IRanges_2.10.0         S4Vectors_0.14.0         BiocGenerics_0.22.0
## [19] BiocStyle_2.4.0
##
## loaded via a namespace (and not attached):
## [1] edgeR_3.18.0           splines_3.4.0            latticeExtra_0.6-28      RBGL_1.52.0
## [5] GenomeInfoDbData_0.99.0 yaml_2.1.14              Category_2.42.0          RSQLite_1.1-2
## [9] backports_1.0.5        lattice_0.20-35          limma_3.32.0             digest_0.6.12
## [13] RColorBrewer_1.1-2     checkmate_1.8.2          colorspace_1.3-2         htmltools_0.3.5
## [17] Matrix_1.2-8           plyr_1.8.4               GSEABase_1.38.0          XML_3.98-1.6
## [21] pheatmap_1.0.8         biomaRt_2.32.0           genefilter_1.58.0        zlibbioc_1.22.0
## [25] xtable_1.8-2           GO.db_3.4.1              scales_0.4.1             brew_1.0-6
## [29] tibble_1.3.0           annotate_1.54.0           GenomicFeatures_1.28.0   lazyeval_0.2.0
## [33] survival_2.41-3        magrittr_1.5             memoise_1.1.0            evaluate_0.10
## [37] fail_1.3               nlme_3.1-131             hwriter_1.3.2            GOSTATS_2.42.0
## [41] graph_1.54.0           tools_3.4.0              BBmisc_1.11              stringr_1.2.0
## [45] sendmailR_1.2-1        munsell_0.4.3            locfit_1.5-9.1           AnnotationDbi_1.38.0
## [49] compiler_3.4.0         grid_3.4.0               RCurl_1.95-4.8           rjson_0.2.15
## [53] AnnotationForge_1.18.0 bitops_1.0-6             base64enc_0.1-3          rmarkdown_1.5
## [57] codetools_0.2-15       gtable_0.2.0             DBI_0.6-1                knitr_1.15.1
## [61] rtracklayer_1.36.0     rprojroot_1.2            stringi_1.1.5            BatchJobs_1.6
## [65] Rcpp_0.12.10

```

Funding

This project was supported by funds from the National Institutes of Health (NIH) and the National Science Foundation (NSF).

References

- H Backman, Tyler W, and Thomas Girke. 2016. “systemPipeR: NGS workflow and report generation environment.” *BMC Bioinformatics* 17 (1): 388. doi:10.1186/s12859-016-1241-0.
- Li, H, and R Durbin. 2009. “Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform.” *Bioinformatics* 25 (14): 1754–60. doi:10.1093/bioinformatics/btp324.
- Li, Heng. 2011. “A Statistical Framework for SNP Calling, Mutation Discovery, Association Mapping and Population Genetical Parameter Estimation from Sequencing Data.” *Bioinformatics* 27 (21): 2987–93. doi:10.1093/bioinformatics/btr509.
- . 2013. “Aligning Sequence Reads, Clone Sequences and Assembly Contigs with BWA-MEM.” *arXiv [Q-Bio.GN]*. <http://arxiv.org/abs/1303.3997>.
- McKenna, Aaron, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernysky, Kiran Garimella, et al. 2010. “The Genome Analysis Toolkit: A MapReduce Framework for Analyzing Next-Generation DNA Sequencing Data.” *Genome Res.* 20 (9): 1297–1303. doi:10.1101/gr.107524.110.