Program coveragePlot_region

Institute of Genetics, University of Bern
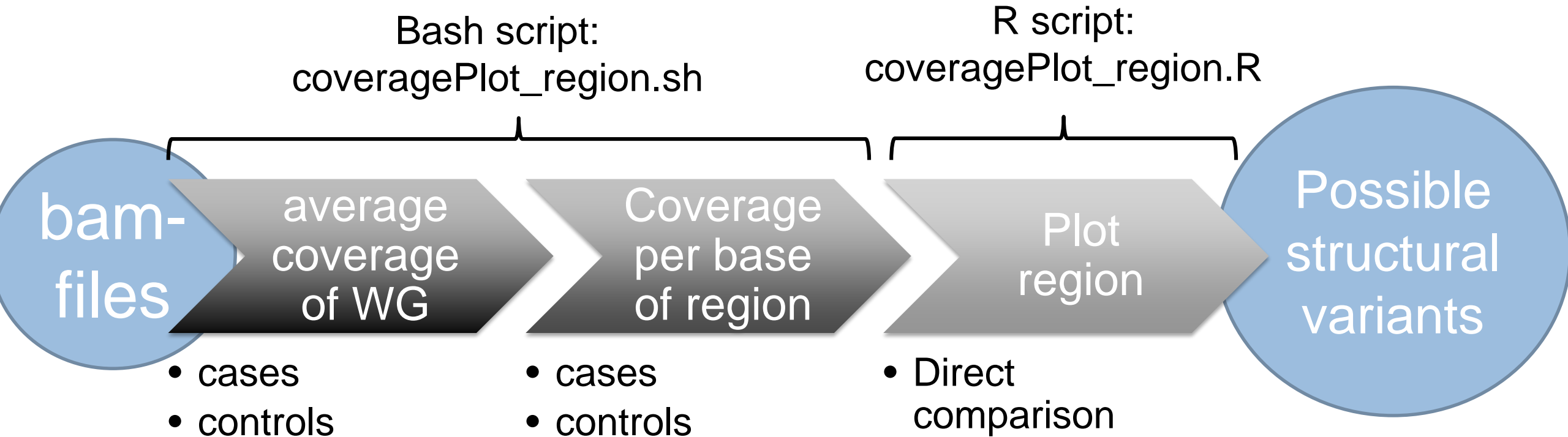
September 2018

Irene Häfliger

General

> Name: ***coveragePlot_region***

> Usage: to plot parts of a chromosome in more detail

> Available on vetsrv06

> programmed in bash and R scripting

> What is new?
— Can select regions in the genome
— Several cases and controls can be directly compared

Algorithm

Bash script:
coveragePlot_region.sh

R script:
coveragePlot_region.R

bam-files

average coverage of WG
- cases
- controls

Coverage per base of region
- cases
- controls

Plot region
- Direct comparison

Possible structural variants

Methods used

> Calculations:
  — **Goleft** is used to calculate the whole genome coverage
  — **Samtools depth –b region bamfiles.list** is used to calculate the coverage per base

> Plotting:
  — Up to 10'000 bases it will plot the coverage directly for each base with the usual function **plot()**
  — For windows > 10'000 the program calculates the average over several bases depending on the size of the window.
  it always plots 10'000 observations – so if the window has size 500kb it will calculate the average over 50 bases.
  The plot is made with the usual function **plot()**.

Prerequisites

> Access to vetsrv06

> Parameterfile
— Input for the program

> Bam files

Parameterfile

> Example file in folder on drive G:
> ..\..\..\Labor\computational protocols\current computational protocols\program_coveragePlot-region\controlparameters_example.ctr.sh

> **Important!!**
>   — Enter the whole path to the bamfiles
>   — Give it a unique job name
>   — Define the region in base
>   – 17 to 19 Mb would then be
>   start='17000000 '   end='19000000'

bamfiles on vetsrv06 - II

This is important if storage space becomes limited:

Be aware that the data stored per project will take up quite some space. So tidy up now and then when you are done with a project or when you see that some trials were not successfull anyways.

Instructions I: Preparations

1. Prepare the parameterfile

2. Login to vetsrv06 on a console (ssh, putty, etc)

3. Copy files to your working directory:
   1. Parameterfile
   2. Bash-script: coveragePlot_region.sh (can find it <u>here</u>)
   3. Rscript: coveragePlot_region.R (can find it <u>here too</u>)

Instructions II: Run the program

4. A) For small regions (< 1Mb) type:

  bash *coveragePlot_region.sh* parameterfile.ctr.sh *<ENTER>*

  Program will run on the console. You can follow the progress on the terminal.

 B) For big regions (≥ 1Mb) type:

  nohup bash *coveragePlot_region.sh* parameterfile.ctr.sh **&** *<ENTER>*

  Program will run in the background. You **cannot** follow the progress on the terminal. But you can close the terminal or do other things on the terminal while the program keeps running.

5. Enter the new folder of the project (**cd** foldername)

6. Run the Rscript by typing:
R CMD BATCH *../coveragPlot_region.R <ENTER>*
The plot will be produced. This may take some minutes but should not more than 20 minutes.

Output

> One folder will include the files
> - *output*.log   logfile summarizing what was done
> - parameterfile.ctr.sh   copy of the parameterfile
> - Av_cov_wg_taget_and_control.txt   file with the average coverage of all animals
> - cov_perBase_*chr_species*.txt   file with the coverage per base of all animals
> - cov_perBase_*chr_start_end_species*.pdf actual plots fo the coverage of that region

> The output folder will have the name of the job, which is defined in the parameterfile
> The folder will be placed in the current working directory

Error messages

> Some common mistakes
>
> — The input job-name is also the name of the output folder. Make sure that no folder with that name exists already
>
> — If ID's do not exist, check if you entered the right path to the bam files and the exact labID used in the name of the bamfiles
>
> — If the output file is just empty
>
>   → check if you defined sensible regions
>   → check if the region is defined correct (it is in **bases!!**)

Final remarks

> Do not run the program on vetpc1727
  — Files are mounted on vetpc1727 in read-only modus. So you would have to copy the whole file before you start working with it.
  — There is about 1TB of storage to use, what is not enough.
> Vetsrv06 has 32 cores
  — The program uses 1 core when it runs.
  — Other people are using the server too → do not run too many jobs at once!

Think before you start the program

> What are you trying to proof?

> Do you use all the knowledge you have about the cases?

> Did you set all parameters correct?

> How long will it run?

— Should it run in the background?

— Are there many jobs running on vetsrv06 already?
(type **top** on the console and have a look at the currently running processes)

Troubleshooting

> Nothing works the way it should?

Please contact me
— Office 012
— Phone 2328
— Mail irene.haefliger@vetsuisse.unibe.ch

Have fun using the program!!