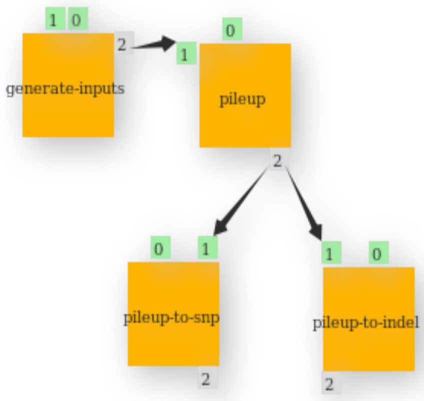
ER-flow Application Description Template

Application Name:
SNP calling with samtools and varscan

Application domain:
bioinformatics, life-sciences, dna sequencing
Brief description of application
One of the goals of next generation sequencing is to find variants in individuals compared to a reference genome. Therefore raw sequence data is aligned (mapped) to the reference genome (see the “Sequence alignment” application). After the alignment variant with respect to the reference can be determined. This application calls variants from (human) genome re-sequencing data with the programs 'samtools' and 'varscan'. The samtools program calls raw variants from a dataset, after which the varscan program determines which SNP calls have more evidence to be true positive calls.



Description of
the
components

Generate-inputs

Takes a text file with paths to alignment files and creates separate text files with one path per file. These files are passed on to the next component. This is a generator component. More information about this component can be found in the paragraph “Helper components”.

Input:

- PREFIX: a text file with the word “BAM”
- VALUES: a text file with a list of paths to alignment files in bam format

Output:

- BAM: text files containing one path per file

pileup

Calls raw variants with “samtools pileup”.

Input:

- REFERENCE: path to the reference genome. The reference genome is in fasta format and compressed with gzip.
- BAM: path to the alignment file in bam format

Output:

- PILEUP: path to the variant file in pileup format pileup-to-snp and pileup-to-indel

Implements the varscan programs “pileup-to-snp” and “pileup-to-indel”.
Input:

- PARAMETERS: a text file containing the parameters for Varscan. Supply an empty file to use the default settings of Varscan
- PILEUP: text file with the path to the raw variant file in pileup format

Output:

- SNP and INDEL: path to the results of Varscan. The file where SNP points to contains the single nucleotide polymorphisms, the INDEL file to the file with short insertions and deletions.

References and links

Samtools - http://samtools.sourceforge.net/

Li H.*, Handsaker B.*, Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G., Durbin R. and 1000 Genome Project Data Processing Subgroup (2009) The Sequence alignment/map (SAM) format and SAMtools. Bioinformatics, 25, 2078-9. [PMID: 19505943]

Varscan - http://varscan.sourceforge.net/
Koboldt DC, Chen K, Wylie T, Larson DE, McLellan MD, Mardis ER, Weinstock GM, Wilson RK, & Ding L (2009). VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics (Oxford, England), 25* (17), 2283-5 PMID:19542151
Koboldt, D., Zhang, Q., Larson, D., Shen, D., McLellan, M., Lin, L., Miller, C., Mardis, E., Ding, L., & Wilson, R. (2012). VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing *Genome Research* DOI: 10.1101/gr.129684.111

Sample data

Alignment files in bam format can be downloaded from
http://www.1000genomes.org/

The reference genome can be downloaded from
ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/human_g1k_v37.fasta.gz

OLD TEXT

Input: reference genome (fasta format), aligned sequence fragments (bam format), parameters for SNP calling programs

data:
input data format: fasta (reference database), bam (alignment file)
output data format: text output data value range
sample data: tba
application

- http://varscan.sourceforge.net/
- http://samtools.sourceforge.net/

documentation (link)

- see links above

publication (link)

- **VarScan 1:** Koboldt DC, Chen K, Wylie T, Larson DE, McLellan MD, Mardis ER, Weinstock GM, Wilson RK, & Ding L (2009). VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics (Oxford, England), 25* (17), 2283-5 PMID:19542151
  VarScan 2: Koboldt, D., Zhang, Q., Larson, D., Shen, D., McLellan, M., Lin, L., Miller, C., Mardis, E., Ding, L., & Wilson, R. (2012). VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing *Genome Research* DOI: 10.1101/gr.129684.111
- **Samtools:** Li H.*, Handsaker B.*, Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G., Durbin R. and 1000 Genome Project Data Processing Subgroup (2009) The Sequence alignment/map (SAM) format and SAMtools. Bioinformatics, 25, 2078-9. [PMID: 19505943]

Execution environment
DCI: (computing, data, VO, etc): grid, cluster
middleware: glite, pbs workflow system: moteur, ws-pgrade/guse

Execution characteristics
data size (per unit, typical number of units):
typical data size: 5GB (reference genome) + 10-500GB (per alignment file)
processing time: depends on the size of the alignment file
memory and disk usage: tbd

input temporary output
processing time (per unit):
memory usage: disk usage:

Target users
Community, projects: (link) number of users:
Bioinformatics, DNA sequencing
Number: of general interest to biomedical and bioinformatics researchers working with DNA re-sequencing data world-wide

user type: developer and end-user

Usage scenario for workflow in the ER-FLOW (how workflow will be reused, metaworkflow, how expected to contribute to project indicators, etc.).

The development of SNP calling programs is ongoing. Several methods exist. The described methods are two commonly used programs.

Usage scenario: the sequencing facility or the bioinformatician stores the sequence alignment files on grid or cluster storage. The end-user (bioinformatician or biomedical researcher) defines paths to the files as input for the workflow. Results stay on grid or cluster, because the files are too large to download via the browser.

Contact information (author)
name: Barbera van Schaik, Bioinformatics Laboratory, AMC, Amsterdam, the Netherlands
e-mail: b.d.vanschaik@amc.uva.nl, support@ebioscience.amc.nl