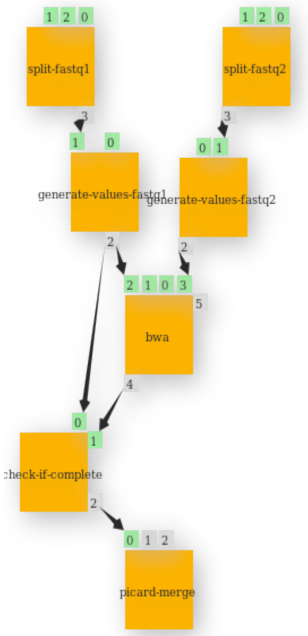
ER-flow Application Description Template

Application Name:
Sequence alignment

Application domain:
bioinformatics, life sciences, dna sequencing

Brief description of application

Aligns sequence fragments in fastq format to a reference database with BWA. The sequence reads have to be stored on the computing platform (grid/g-lite, pbs or local). The inputs supplied to this workflow are plain text files containing the path to the input files. The input files are split in smaller chunks, after which they are passed on to the "bwa" component. When all sequences were processed by BWA in parallel, a component checks if all results (alignment files in bam format) were produced. After this they are merged into one alignment file in bam format.



Description of the components

Split-fastq1 and split-fastq2

Inputs:

- FASTQ: text file containing the path to the sequence file in fastq format (fastq1 receives the forward reads, fastq2 the reverse reads)
- LINES: text file that contains a parameter that defines how large the splitted files should be, in number of lines per file. This should be a multiple of 4, since fastq entries are 4 lines long.
- OUTDIR: text file containing the directory path where the output should be stored. The directory for "split-fastq1" should be different from the directory for "split-fastq2". Outputs of the next components will be stored in a subdirectory of the "outdir" defined for "split-fastq1".

Output:

- FASTQLIST: a text file containing the file paths to the splitted fastq file

generate-values-fastq1 and generate-values-fastq2

This is a commonly used component in the DNA sequencing workflows and is explained in more detail in the paragraph "helper components". These are generator components.

Inputs:

- PREFIX: a text file containing the word "FASTQ1" or "FASTQ2", respectively
- VALUES: the list of fastq files that was generated by the previous component.

Outputs:

- FASTQ1_* and FASTQ2_*: text files containing the path to the fastq files. One path per file. The filenames follow the internal file naming scheme that WS-Pgrade/gUse expects for parameter sweeps.

Bwa

Performs an alignment of the sequence reads against a reference database, e.g. the human genome.
Inputs:

- REFERENCE: text file containing the path to the reference database. The file is a .tar.gz file with the reference in fasta format and with a BWA index for that reference genome. Please note that you need to make different indices for Solid and Illumina reads.
- PARAMETERS: text file with the command line parameters for BWA. Supply an empty file to use the default settings of BWA.
- FASTQ1: text file with the path to the forward sequence reads in fastq format
- FASTQ2: text file with the path to the reverse sequence reads in fastq format

Outputs:

- BAM: text file containing the path to the sequence alignment file in bam format
- BAI: text file containing the path to the index of the bam file

check-if-complete

Sometimes jobs fail. This component checks if all results are complete by a fairly simple comparison, i.e. the number of input files are compared with the number of output files. If these are equal, the output files are passed on to the next component, else this component will fail. This component makes sure that the last component is only fired when all results are complete. This is a collector and generator component.
Inputs:

- INPUT: the outputs that were generated by “generate-values-fastq1”
- OUTPUT: the outputs that were generated by “bwa”

Output:

- PASS: the same files as OUTPUT

picard-merge

Merges several alignment files (bam format) into one alignment file in bam format. This is a collector component.
Input:

- BAM: text files containing paths to the sequence alignment files in bam format

Output:

- BAMOUT: text file containing the path to the merged bam file
- BAIOUT: text file containing the path to the index on the merged bam file

Links and references

BWA - http://bio-bwa.sourceforge.net/
Li H. and Durbin R. (2009) Fast and accurate short read alignment with Burrows-Wheeler Transform. Bioinformatics, 25:1754-60. [PMID:19451168]

Picard tools - http://picard.sourceforge.net/

Sample data

Sequence data
http://www.1000genomes.org/
Store a fastq file for the forward and the reverse reads in fastq format on the compute resource and

provide the path via the FASTQ1 and FASTQ2 inputs.

Reference genome
ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/human_g1k_v37.fasta.gz
A BWA index can be build using these instructions:
http://bio-bwa.sourceforge.net/bwa.shtml
The reference genome (fasta) and bwa index are archived using tar zcvf.
Store this file on the compute resource and provide the path to this file via the REFERENCE input.

OLD TEXT:

data:
input data format:
Raw sequence fragments (fastq, sff or fasta format), reference database (fasta format), alignment parameters
output data format:
alignment files in bam format
sample data: (link) tbd
application (link)
BWA: http://bio-bwa.sourceforge.net/
BLAT: http://genome.ucsc.edu/FAQ/FAQblat.html#blat3
BFAST: http://sourceforge.net/apps/mediawiki/bfast/index.php?title=Main_Page
BLAST: http://blast.ncbi.nlm.nih.gov/Blast.cgi

documentation (link) see above
publication (link)
BWA short reads: Li H. and Durbin R. (2009) Fast and accurate short read alignment with Burrows-Wheeler Transform. Bioinformatics, 25:1754-60. [PMID:19451168]
BWA long reads: Li H. and Durbin R. (2010) Fast and accurate long-read alignment with Burrows-Wheeler Transform. Bioinformatics, Epub. [PMID: 20080505]
BLAT: WJ Kent (2002) BLAT--the BLAST-like alignment tool. Genome research, 12(4):656-64. [PMID: 11932250]
BFAST: Homer N, Merriman B, Nelson SF (2009) BFAST: An alignment tool for large scale genome resequencing. PLoS ONE. 2009 4(11): e7767. [PMID: 19907642] http://dx.doi.org/10.1371/journal.pone.0007767
BLAST: Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: Basic local alignment search tool. J Mol Biol 1990, 215(3):403-410. PubMed

Execution environment
DCI: grid (vlemed VO), cluster
middleware: glite, pbs workflow system: ws-pgrade/guse
PS: some of these workflows already exist for MOTEUR

Execution characteristics
data size (per unit, typical number of units):
input sequences: ranging from 1 to 100 GB (is growing)
input database: 5GB
output alignment file: same as input sequences
processing time (per unit):
Depending on input size. In the order of 5-24 hours for input up to 20GB
memory usage: disk usage:
Memory: 4.5GB (BWA, BLAT), tbd for BFAST

Target users
Community, projects: (link) number of users:
Bioinformatics, dna sequencing
Of general interest to researchers working with high throughput DNA sequence data.

user type: developer and end-user.

Usage scenario for workflow in the ER-FLOW

First the workflows will be developed and disseminated among the advanced users, then these will become available for end-users at the AMC ebioinfra gateway.
We do not foresee end-users running workflows from the SHIWA platform due to two reasons:
+ the user interface is too detailed
+ no data management is provided for such large files.

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