SIMPLEX
User Manual

Version 1.0
February 9, 2012

Andreas Dander
Maria Fischer
Stephan Pabinger
Rene Snajder
Gernot Stocker
Innsbruck Medical University
Division for Bioinformatics
Innrain 80-82/Level 4, 6020 Innsbruck, Austria

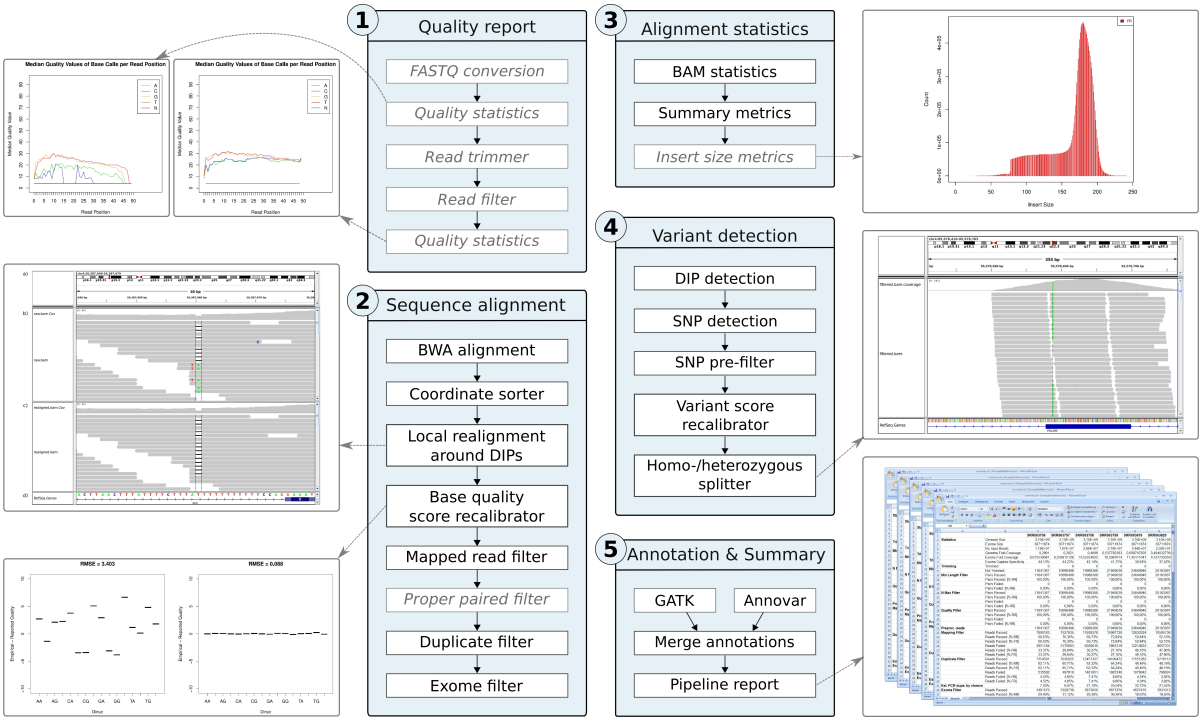
Contents

1 Introduction

This document is the user manual for the automatic analysis pipeline for exome data called SIMPLEX. The pipeline can be downloaded from http://www.icbi.at/simplex and is distributed under the GPL.

1.1 SIMPLEX

SIMPLEX is an autonomous analysis pipeline, which comprises the complete exome analysis workflow, including the following steps: (1) initial quality control, intelligent data filtering and pre-processing, (2) sequence alignment to a reference genome and refinement, (3) alignment statistics; (4) SNP and DIP detection, (5) functional annotation of variants using different approaches, and report generation during various stages of the workflow (see figure 1.1).



SIMPLEX connects the selected analysis steps, exposes all available parameters for customized usage, performs required data handling, and distributes computationally expensive tasks either on a dedicated high-performance computing infrastructure or on the Amazon cloud environment (EC2).

2 Download and installation

2.1 SIMPLEX Server

To run SIMPLEX you need both an instance of the server, which performs the analysis, and the client, which transfers data from and to the server and handles the flow of the pipeline. Since setting up the SIMPLEX server would be time consuming and complicated, we provide both a fully configured virtual machine image (in OVA format) as well as a Amazon EC2 cloud image.

2.1.1 Virtualization image (Virtual machine image)

The following steps should be performed to get a running instance of SIMPLEX.

1. Download the image at http://simplex.i-med.ac.at
2. Start your virtualization software (every software supporting Open Virtualization Format should work)
3. Load the downloaded image into the virtualization software
4. Start the image
5. Configure networking between host and guest system.

2.1.2 Cloud image

The Cloud image of SIMPLEX is based on the CloudMan image of Galaxy. For more information please visit http://wiki.g2.bx.psu.edu/Admin/Cloud. It can be instantiated from the Amazon AWS Console. The AMI name is "Simplex 1.02 Ubuntu 10.04 LTS based on CBL".

Detailed instructions on how to use SIMPLEX within the cloud are given in section 3.

2.1.3 Passwords

As explained in section 4.1.5 you need a username & password to access the SIMPLEX cluster (Whether it is the VirtualBox image or the cloud image..).

For both, the VirtualBox and the cloud image, the following default is used:

```
Username: simplex
Password: simplexdemo
```

2.2 SIMPLEX Client

The SIMPLEX Client can run on any office computer. Please perform the following steps to install SIMPLEX:

1. Download the SIMPLEX Client at http://simplex.i-med.ac.at
2. Install Oracle Java Runtime 7 (recommended) or 6 from www.oracle.com
3. See section 4 on how to use SIMPLEX

2.2.1 Java 6 Support

SIMPLEX works best with Java 7 installed. If you want to use it on Java 6, you need to perform a few more steps:

1. Download the Java 6 Support Libraries for Simplex from here
2. Create a directory in your JAVA_HOME directory with the following path: JAVA_HOMEjrelibend
3. Unpack the tgz file into the directory you just created
4. When calling Simplex, make sure you use the same Java version you just extended.

3 EC2 Cloud Usage

We created an AMI (Amazon Machine Image) containing SIMPLEX. This AMI is based on CloudBioLinux[1] and therefore uses CloudMan [1] for resource management.

3.1 Prerequisites

The only thing you need to start SIMPLEX in the Cloud is an EC2 Account. You can get one at http://aws.amazon.com/console/. Basic understanding of how to use the AWS Management Console is recommended. But as we will be giving a step-by-step instruction on how to use it, it is not entirely necessary.

3.2 Preparations

You will only have to perform these steps if you are using SIMPLEX for the first time. If you have used the CloudBioLinux or Galaxy in the Cloud before, you'll already have all required settings. Therefore, you could skip this step. What we have to do is to create a security group for our Instances and assign the proper rules. The following steps are the same as described in "Step 1" in the CloudMan Manual [2] with the addition of one rule (HTTPS):

- Create an EC2 account and log-in to the AWS management Console
- In the EC2 Tab, click "Security Groups".
- Click "Create Security Group".
- Enter a name and description (for example "SIMPLEX").
- After the Group has been created, select the Group in the list and click the "Inbound" tab in the lower panel.
- For each rule, enter the following values in the fields on the left side (Create a new rule, Port range, Source) and click "Add Rule":
  - Type: HTTP; Source: 0.0.0.0/0 (or you can enter your public IP address here if you want to limit access)
  - Type: HTTPS; Source: 0.0.0.0/0 (or you can enter your public IP address here if you want to limit access)

[1] http://cloudbiolinux.org/
[2] http://wiki.g2.bx.psu.edu/Admin/Cloud#Detailed_steps

  - Type: SSH; Source: 0.0.0.0/0 (or you can enter your public IP address here if you want to limit access)
  - Type: Custom TCP Rule; Port range: 42284; Source: 0.0.0.0/0
  - Type: Custom TCP Rule; Port range: 20-21; Source: 0.0.0.0/0
  - Type: Custom TCP Rule; Port range: 30000-30100; Source: 0.0.0.0/0
  - Type: All TCP; Source: (the name of the security group, e.g. "SIMPLEX")
- Click "Apply Rule Changes"

3.3 Starting the SIMPLEX AMI

- Login to the AWS Management Console and click the EC2 tab.
- Click "AMIs" in the Navigation panel on the left side
- Search for "SIMPLEX" in "All Images". If there is more than one result, please choose the one from Owner "372123314130", which is us.
- Click "Launch" to start the instance.
  - Select Instance Type "Extra Large" (8 ECUs) or higher (But NOT High-Memory Extra Large). Basically, the instance you start must have at least 4 CPU cores, otherwise some jobs will be queued forever. Click Continue.
  - On the next page, in the "User Data" field you must provide the following information:

    ```
    cluster_name: <a name for the \pipelinename{} cluster>
    password: <a password for the CloudMan webinterface>
    access_key: <your AWS access key>
    secret_key: <your AWS secret key>
    ```

    You got the AWS access key and AWS secret key when you created your Amazon Account. You can also retrieve it by clicking your username in the upper right corner and then "Security credentials" (must be logged in to a power user. For example it might look like this:

    ```
    cluster_name: mysimplexcluster
    password: supersecurepw
    access_key: AEIQEJ5RUZDHNWS69IJZ
    secret_key: DsopjdJSOdjqShkqhkw+DJQJKL+JD781Auo8D7Aj
    ```

  - Click Continue. On the next page you can select an existing key pair or create a new one. This step is only required if you plan to log on to the instance via SSH.
  - Click Continue. Now you are in the "Configure Firewall" page. Select the Security Group you configured in the Preparations chapter.
  - Click Continue and then Launch. Your instance should be launching.

3.4 Initializing the SIMPLEX Cluster

Now that the instance is starting, we have to open the CloudMan interface to initialize the Cluster in the Cloud.

- Click "Instances" on the left side, and select the instance from the list. In the bottom panel the parameters of this instance are displayed.
- See the "Public DNS:" field. It will show something like "ec2-50-19-171-174.compute-1.amazonaws.com". This is the public domain name of your instance.
- Open a new browser window, and enter the public domain name of your instance in the address bar. For example: http://ec2-50-19-171-174.compute-1.amazonaws.com
- If you cannot connect, try to give it a few minutes to start up. If after about 2 or 3 minutes you still cannot connect, check if the Security Group settings are set properly.
- You should now be asked for a username and password. The user name is "CM Administration" and the password is the one you set in the "User Data" field (in our example it was "supersecurepw".
- Now you will see the "Initial Cluster Configuration" dialog of CloudMan. Enter the number of GB you want to reserve on the data volume. This is important, as it depends on how big the data is you want to analyze with SIMPLEX. If you have 1GB input data, reserve at least 10GB in the Cloud as all the in-between results will also be stored there.
- Click "Start Cluster". Now the CloudMan and Galaxy services are starting up. Wait until both "Application" and "Data" have a green dot displayed (you may want to refresh the page from time to time). This may take a couple of minutes.
- Once the cluster initialization is done, you should be able to access https://–public instance domain name–/JClusterService. It should show an XML file. Don't bother trying to understand it. If it's there, initialization is completed.

Congratulations! The SIMPLEX cluster is running in your very own cloud instance!

You can now use this URL to in the cluster.properties file. See Section 4 on how to use Simplex.

4 Usage

SIMPLEX is designed as an intuitive, easy to use, and highly customizable command-line program to provide analyses for single end (SE) and paired end (PE) as well as Illumina (NS) and SOLiD (CS) experiments. All available parameters are listed in a document which can be downloaded at http://www.icbi.at/exome.

4.1 Input files

To analyze exome sequencing data with SIMPLEX, sequence read files, the exome (or target region) definition file, and server connection properties are needed. An example how to pass these files is given in section 4.2.

4.1.1 Illumina reads

Illumina raw reads (provided by parameter **-I**) are expected to be in FASTQ format. All three types, Sanger, Solexa (aka. Illumina 1.3-), and Illumina 1.3+ FASTQ, are supported by the pipeline. A detailed description of the differences between these format types is given in [2] and on http://en.wikipedia.org/wiki/FASTQ_format.

CAVEAT: to avoid interferance with downstream analyses, enable FASTQ conversion (parameters **-fqc** and **-cf** [**solexa** | **illumina**]) if Solexa or Illumina FASTQ format is used. This will trigger the conversion into Sanger FASTQ format before any other analysis steps are performed.

4.1.2 SOLiD reads

SOLiD data is provided to the pipeline by two (SE) or four (PE) files: csfasta for sequence information (parameter **-I**) and qual files (parameter **-Q**) containing each read's corresponding quality scores.

4.1.3 Naming convention for sequence read files

In order to generate reasonable output file names, SIMPLEX requires read input files to be named according to the following patterns:

NS/SE

runname.fq[.gz] - read data in FASTQ format

NS/PE

*runname*_R1.fq[.gz] - first reads of pairs in FASTQ format

*runname*_R2.fq[.gz] - second reads of pairs in FASTQ format

CS/SE

runname.csfasta[.gz] - read sequences in FASTA format

*runname*_QV.qual[.gz] - read quality values in FASTA format

CS/PE

*runname*_F3.csfasta[.gz] - sequence of first reads of pairs in FASTA format

*runname*_F3_QV.qual[.gz] - quality values of first reads of pairs in FASTA format

*runname*_F5.csfasta[.gz] - sequence of second reads of pairs in FASTA format

*runname*_F5_QV.qual[.gz] - quality values of second reads of pairs in FASTA format

4.1.4 Exome definition

This file (provided by parameter **-sfeb**) is needed to determin fold coverage measurments (e.g. overall fold coverage, identification of not/low captured regions) and to filter non-exonic/not targetted reads. It defines all exons/target regions by their positions within the genome (chromosome, start, and end position) in BED file format[1]. Idealy, the exon's name should include further description about corresponding gene name, exon number, and ccds id, e.g.:

```
track name=CCDS description="chrom start end name score strand"
chr1 69090 70008 OR4F5;Exon1;CCDS30547.1 850 +
chr1 367658 368597 OR4F29;Exon1;CCDS41220.1 850 +
chr1 621095 622034 OR4F16;Exon1;CCDS41221.1 850 -
chr1 861321 861393 SAMD11;Exon1;CCDS2.2 850 +
```

We do provide an exome definition files based on the released CCDS update for human build HsGRCh37.3 (September 7, 2011) which can be downloaded from the SIMPLEX webpage http://simplex.i-med.ac.at.

4.1.5 Server connection properties file

The server connection properties file (provided by parameter **-k** or **–clusterpropsfile**) is required to tell SIMPLEX where to contact the server instance. This could be either the VirtualBox image, or the AMI Cloud Image or - if you chose to set one up by yourself - your own server running JClusterService.

The format of the file looks like this:

```
url=https://URL_TO_YOUR_CLUSTER/JClusterService
user=username
password=password
```

[1] http://genome.ucsc.edu/FAQ/FAQformat.html#format1

For example: if you run the SIMPLEX Amazon Cloud Image, it would look like this (replace "ec2-50-19-171-174.compute-1.amazonaws.com" with your instance's public DNS):

```
url=https://ec2-50-19-171-174.compute-1.amazonaws.com/JClusterService
user=simplex
password=simplexdemo
```

4.2 Calling SIMPLEX

Depending on library preparation and sequencing technology used, different minimum parameter sets are required to start the pipeline. The following examples describe a minimal **testdata** pipeline call to test SIMPLEX with the provided testdata available on http://simplex.i-med.ac.at. We suggest to put the actual pipeline call into a script as it simplifies reproducing results.

4.2.1 Illumina Single End

```
./exomePipeline -c exomeSE -genP hg19.test -I testdata.fq.gz \
   -od path_to_output_directory -sfeb CCDS.20110907.test.bed \
   -dsP 80 -k connection.properties
```

Parameter description:

-c exomeSE ... use SE pipeline

-genP hg19.test ... prefix of the test reference genome

-I testdata.fq.gz ... input file in FASTQ format

-od path_to_outputdirectory ... output directory

-sfeb CCDS.20110907.test.bed ... bed file specifying the test exome

-dsP 80 ... percentage to distinguish between homo- and heterozygous DIPs

-k connection.properties ... server connection properties of the service

4.2.2 Illumina Paired End

```
./exomePipeline -c exomePE -I testdata_R1.fq.gz,testdata_R2.fq.gz \
   -od path_to_output_directory -genP hg19.test \
   -sfeb CCDS.20110907.test.bed -dsP 80 -k connection.properties
```

Parameter description:

-c exomePE ... use PE pipeline

-genP hg19.test ... prefix of the test reference genome

-I testdata_R1.fq.gz,testdata_R2.fq.gz ... first and second read in pair input files in FASTQ format

-od path_to_outputdirectory ... output directory

-sfeb CCDS.20110907.test.bed ... bed file specifying the test exome

-dsP 80 ... percentage to distinguish between homo- and heterozygous DIPs

-k connection.properties ... server connection properties of the service

4.2.3 SOLiD Single End

```
./exomePipeline -c exomeSE -CS -genP hg19.color.test \
   -I testdata_F3.csfasta.gz -Q testdata_F3_QV.qual.gz \
   -od path_to_output_directory -sfeb CCDS.20110907.test.bed \
   -dsP 80 -k connection.properties
```

Parameter description:

-c exomeSE ... use SE pipeline

-CS ... analyse color space data

-genP hg19.color.test ... prefix of the test reference genome

-I testdata_F3.csfasta.gz ... SOLiD read sequences in FASTA format

-Q testdata_F3_QV.qual.gz ... SOLiD read quality values in FASTA format

-od path_to_outputdirectory ... output directory

-sfeb CCDS.20110907.test.bed ... bed file specifying the test exome

-dsP 80 ... percentage to distinguish between homo- and heterozygous DIPs

-k connection.properties ... server connection properties of the service

4.2.4 SOLiD Paired End

```
./exomePipeline -c exomePE -CS -od path_to_output_directory\
   -I testdata_F3.csfasta.gz,testdata_F5.csfasta.gz \
   -Q testdata_F3_QV.qual.gz,testdata_F5_QV.qual.gz \
   -genP hg19.color.test -sfeb CCDS.20110907.test.bed \
   -dsP 80 -k connection.properties
```

Parameter description:

-c exomePE ... use PE pipeline

-CS ... analyse color space data

-genP hg19.color.test ... prefix of the test reference genome

-I testdata_F3.csfasta.gz,testdata_F5.csfasta.gz ... first and second read in pair SOLiD read sequences in FASTA format

-Q testdata_F3_QV.qual.gz,testdata_F5_QV.qual.gz ... first and second read in pair SOLiD read quality values in FASTA format

-od path_to_outputdirectory ... output directory

-sfeb CCDS.20110907.test.bed ... bed file specifying the test exome

-dsP 80 ... percentage to distinguish between homo- and heterozygous DIPs

-k connection.properties ... server connection properties of the service

4.2.5 Results

These minimal parameter sets would cause SIMPLEX to

- calculate raw coverage statistics,
- align the given reads to a subset of hg19 (chr1 and chr21, defined as SIMPLEX's test reference genome),
- local realign the reads around indels,
- recalibrate quality scores
- filter unmapped reads, duplicates, and not on target reads,
- calculate alignment metrics,
- identify and annotate DIPs (in vcf and tab delimited files),
- identify and annotate SNPs (in vcf and tab delimited files),
- generate indexes for vcf files to be viewed in genome browsers, and
- generate a detailed summary report (available as tab delimited and xls file)

Additional parameter settings (e.g. filtering parameters) are available to enable preprocessing of raw reads, including

- FASTQ conversion
- raw and filtered read statistics including
  - quality value metrics
  - nucleotide distribution
  - N content

- raw read trimming (currently available for Illumina reads only)
- raw read filtering based on (currently available for Illumina reads only)
  - N content
  - read length
  - quality values

Bibliography

[1] E Afgan, D Baker, N Coraor, B Chapman, A Nekrutenko, and J Taylor. Galaxy cloudman: delivering cloud compute clusters. *BMC Bioinformatics*, 11 Suppl 12:S4–S4, 2010.

[2] P J Cock, C J Fields, N Goto, M L Heuer, and P M Rice. The sanger fastq file format for sequences with quality scores, and the solexa/illumina fastq variants. *Nucleic Acids Res*, pages –, 2009.