SMRT® Analysis System Requirements

Introduction	The SMRT Analysis software suite performs assembly and variant detection analysis of sequencing data generated by the PacBio® RS II instrument.
Installation	For installation instructions, see the document **SMRT Analysis Software Installation**, available on the Pacific Biosciences web site.

Hardware Guidelines

Submit Hosts

- Minimum 16 cores, with 4 GB RAM per core.
- Minimum 250 GB of disk space.

Execution Hosts

- 5 nodes.
- 16 cores per node with 4 GB RAM per core.
- Minimum of 250 GB of disk space per node.

Software Prerequisites

Operating Systems

- SMRT Analysis is supported on:
  - English-language **Ubuntu: versions 12.04**, **10.04**, **8.04**
  - English-language **RedHat/CentOS: versions 6.3**, **5.6, 5.3**
- SMRT Analysis **cannot** be installed on Mac OS® or Windows® systems.

Client OS

To run SMRT Portal and SMRT View, we recommend:

- Microsoft Windows XP or later.
- Mac OS X 10.6 or later.

Client Web Browser

- We recommend using the Google Chrome® 21 web browser to run SMRT Portal for consistent functionality. We also support Apple's Safari® and Internet Explorer® web browsers; however some features may not be optimized on these browsers.

Client Java

To run SMRT View, we recommend installing the **latest** version of Java. The minimum recommended versions are:

- **Oracle Java**: Java Version 7 Update 67 or later for Linux, Windows, and Mac OS X.
- **Apple Java**: Java for OS X 2013-004 (1.6.0_51-b11-457-10M4509) or later.

Compatible Computing Infrastructure

SMRT Analysis can generally be deployed on three different types of computing infrastructures. We recommend and support deployment on a multi-node cluster computing environment, though it is possible, but **unsupported**, to install SMRT Analysis on other computing environments as well. Considerations for typical deployment scenarios are described below:

1. Multi-node clusters

For production-level processing of multiple SMRT Cells and analyses per day, a multi-node cluster is **necessary**. SMRT Analysis can be configured to use the SGE, PBS, or LSF job-management systems.

- SGE is **preferred** and most extensively tested.
- UGE is similar enough to SGE such that it can be configured exactly the same way as SGE, though we do not do any testing on UGE.
- For large genome (> 200 Mb) assemblies, you **must** use SGE and provide a custom `.spec` file to use with Celera Assembler.

Example Applications:

- All SMRT Analysis protocols for genomes up to 100 Mb including:
  - `RS_HGAP_Assembly.2`, `RS_HGAP_Assembly.3` - *De novo* Assembly with PacBio data only.
  - `RS_AHA_Scaffolding` - Hybrid Scaffolding and gap filling of an existing assembly using PacBio long reads.
- Multiple SMRT Analysis jobs running concurrently.
- Experimental large genome assemblies > 200 Mb.

2. High-powered single-node computer

It is possible, but **not advisable**, to install SMRT Analysis on a single high-powered computer if it meets or exceeds the sum of the minimum CPU and memory requirements for a multi-node cluster. There is a risk in having multiple jobs fail at the same time when resources are exhausted. This is because jobs are **not** managed by any queuing system such as SGE, PBS, or LSF. Therefore, if you initiate many jobs at once, the computer will try to allocate all CPU and memory resources at once, causing failures on all jobs when these resources are exhausted.

Example Applications:

- All SMRT Analysis protocols for genomes up to 100 Mb including:
  - `RS_HGAP_Assembly.2`, `RS_HGAP_Assembly.3` - *De novo* Assembly with PacBio data only.
  - `RS_AHA_Scaffolding` - Hybrid Scaffolding and gap filling of an existing assembly using PacBio long reads.
- Only a few SMRT Analysis jobs running concurrently.

3. Amazon Machine Instance (AMI)

If you do not have **any** computing resources available, consider running SMRT Analysis on the Amazon Elastic Compute Cloud (EC2) infrastructure. We provide a publically-accessible and SMRT-Analysis-specific Amazon Machine Image (AMI) for every SMRT Analysis release. For details, see the document **Running SMRT Analysis on Amazon**, available on the Pacific Biosciences web site.

Example Applications:

- Certain SMRT Analysis protocols for genomes up to 10 Mb including:
  - `RS_HGAP_Assembly.2`, `RS_HGAP_Assembly.3` - *De novo* Assembly with PacBio data only. (E. coli)
  - `RS_AHA_Scaffolding` - Hybrid Scaffolding and gap filling of an existing assembly using PacBio long reads.
- Only **one** SMRT Analysis job running at a time.

Network Configuration

- Please refer to the **IT Site Prep guide** provided with your instrument purchase for more details.
- For network connectivity considerations, see the network diagram in the **Computer Requirements** section of the **IT Site Prep guide**. The PacBio RS II requires four IP addresses and data can be transferred to the SMRT Analysis Server via three methods. Please open the port required for the data transfer method of your choice.

Data Storage

- 10 TB (Actual storage depends on usage.)
  - Each SMRT Cell generates roughly 5 to 8 Gb of data, and the PacBio RS II can sequence roughly 8 SMRT Cells per day. Disk space occupied by raw data alone can expand to ~7.3 Tb if continuously sequencing for 6 months.
  - **Each** SMRT Portal job may also add ~1 Gb of data, and one SMRT Cell can be analyzed multiple times.
- By default, **all** PacBio Data will only occupy **four** directories on your file system: The **SMRT Analysis software directory,** the **SMRT Cell input directory**, the **SMRT Analysis output directory,** and the **SMRT Analysis temporary directory.** These are described below.
- The **SMRT Analysis software directory** (we recommend `$SMRT_ROOT=/opt/smrtanalysis`) **must** have the same path and be **readable** by the `smrtanalysis` user across **all** compute nodes via **NFS**.
- The **SMRT Cell input directory** (we recommend `$SMRT_ROOT/pacbio_instrument_data/`) **must** have the same path and be **readable** by the `smrtanalysis` user across **all** compute nodes via **NFS**. This directory contains data from the instrument, and can either be:

– A directory configured by RS Remote during instrument installation, or

– A directory you created when you received data from a core lab.

- The **SMRT Analysis output directory** (we recommend `$SMRT_ROOT/userdata)` **must** have the same path and be **writable** by the `smrtanalysis` user across **all** compute nodes via **NFS**. This directory is usually softlinked to a large storage volume.

- The **SMRT Analysis temporary directory** is used for fast I/O operations during runtime. The software accesses this directory from `$SMRT_ROOT/tmpdir` and you can softlink this directory manually or using the install script. This directory should be a **local** directory (**not** NFS-mounted) and be writable by the `smrtanalysis` user and exist as independent directories on **all** compute nodes.

File System Considerations

- The **only** supported file system is NFS; we have **only** tested SMRT Analysis on NFS. There are several known issues associated with alternate or distributed file systems such as GlusterFS. One workaround for these problems is to install the software on a local directory first. Then move to the `smrtanalysis` install directory manually to a mounted directory on the file system.

Cluster Configuration

- Pacific Biosciences has explicitly validated **Sun Grid Engine (SGE)**, and provides job submission templates for **LSF** and **PBS**. You only need to configure the software **once** during initial install.

P/N 100-593-300-01