
Snoopy Documentation

Release 0.3.2

Daniel Rice

May 18, 2016

| | | |
|----------|--|-----------|
| 1 | Overview | 1 |
| 1.1 | Usage | 1 |
| 1.2 | Requirments | 1 |
| 1.3 | Installing | 1 |
| 2 | Loading Data | 3 |
| 2.1 | Modes | 3 |
| 2.1.1 | Manual | 3 |
| 2.1.2 | Batch | 3 |
| 2.2 | Data Access | 4 |
| 2.2.1 | Local Files | 4 |
| 2.2.2 | Remote Files | 4 |
| 2.2.3 | Summary | 5 |
| 2.3 | Input File Formats | 5 |
| 2.3.1 | Variant List File | 5 |
| 2.3.2 | Batch JSON File | 6 |
| 3 | Starting Snoopy | 9 |
| 3.1 | Examples | 9 |
| 3.1.1 | Start local server | 9 |
| 3.1.2 | Start local server at port 8888 | 9 |
| 3.1.3 | Start SSH-Bridge with username bob at big-bio-server | 9 |
| 4 | Perform Quality Control | 11 |
| 5 | Saving Results | 13 |
| 5.1 | JSON | 13 |
| 5.2 | Snapshots | 13 |
| 5.3 | HTML | 14 |
| 6 | User Interface | 15 |
| 6.1 | Dalliance | 15 |
| 6.2 | Snoopy | 16 |
| 6.2.1 | Variant Decisions | 16 |
| 6.2.2 | Navigation | 16 |
| 6.2.3 | Viewing | 17 |
| 6.2.4 | Admin | 17 |
| 6.2.5 | Settings | 17 |
| 6.2.6 | Help | 18 |

| | | |
|-------|--------|----|
| 6.2.7 | GitHub | 18 |
|-------|--------|----|

Overview

Snoopy is a tool designed to expedite the manual quality control of called variant sites in whole genome sequences. Given a set of sequence files (BAM/CRAM) and a list of questionable variant locations, Snoopy will present the reads at each variant location with the Dalliace genome browser. The user makes a judgement (variant or not variant) at each candidate variant. The results of these decisions are exported to a variety of human and machine friendly formats which can then be incorporated into your pipeline.

1.1 Usage

Snoopy is launched from the command line and runs in the browser. Once started on the command line, Snoopy facilitates different ways to access your data depending on where it resides. The easiest and most performant means of data access is with HTTP access but as this isn't always possible, Snoopy provides additional means of file access:

- a local http file server for local files
- an SSH-http bridge for remote files without http but can be accessed by SSH

(For more information on this please see Data Access section.)

In the browser, Snoopy provides an interface to record quality control decisions and to take snapshots for future reference. The sequence data is displayed using the Dalliace genome browser.

1.2 Requirments

- Python (2 or 3)
- Pip
- Modern web browser (Chrome is recommended)

1.3 Installing

```
pip install snoopy
```

Loading Data

There are a few choices to make when using snoopy: which mode to use and how to get to your data. The following usage matrix summarises the choices. Following this each use-case is discussed in more detail.

I want to review...

A single set of sequence files/candidate variants

My BAM files are local Use manual mode and load files within the browser

My BAM files are on a remote server with HTTP/S access Use manual mode and load files from remote HTTP

My BAM/CRAM files are on a remote server without HTTP/S access Use manual mode and load files with the SSH Bridge

Several sets of sequence files/candidate variants

My BAM files are local Use batch mode and load files from the local file server

My BAM files are on a remote server with HTTP/S access Use batch mode and load files from remote HTTP

My BAM/CRAM files are on a remote server without HTTP/S access Use batch mode and load files with the SSH Bridge

2.1 Modes

There are two ways to use Snoopy: **manual mode** and **batch mode**.

2.1.1 Manual

Quickly load a few files and start QC with little preparation.

In this mode, you select one-by-one a **single set** of sequence files (eg a single individual or a trio) and a list of candidate variant locations. Snoopy will then load the set of sequence files and visit each of the candidate variant locations.

2.1.2 Batch

Review several sets of sequence files.

In this mode, you can review **several** sets of sequence files (e.g. several sets of individuals or several different trios). As loading all of this information one-by-one with in the web interface would be cumbersome, the batch mode requires a JSON file which lists all of the sequence files and the variant locations. This file contains an array of **sessions**, where each session consists of:

- A set of sequence files (BAM, CRAM) which you want to view together at each of the locations in
- A list of variant locations (SNPs or CNVs) which you want to review in the sequence files.

For example, a session may consist of:

- a trio of sequence files `mother.bam`, `father.bam`, `offspring.bam`
- a set of de novo sites which need to be verified `chr5:96244805`, `chr11:28483998`, `chr12:106799289`, ...

2.2 Data Access

2.2.1 Local Files

Browser Load

It is possible to load local files within the browser but there are some limitations. Firstly, as Dalliace, the genome browser Snoopy wraps around, does not yet provide support for CRAM, it is only possible to upload BAM files. Secondly, as you can only load one file at a time in a web browser you are restricted to manual mode.

Modes Manual

File types Bam

Command line arguments None

Local File Server

If you have several local files you would like to view with batch mode Snoopy includes a local server. You will still be limited to BAM files however.

Modes Manual, Batch

File types Bam

Command line arguments `--local-server, -l` (see *Starting Snoopy* for more details)

2.2.2 Remote Files

HTTP/S

If your files exist on a remote HTTP/S server, you will be able to access these in either manual or batch mode. You will be limited to BAM files however.

Modes Manual, Batch

File types Bam

Command line arguments None

SSH Bridge

The SSH Bridge is useful if: your files exist on a remote server but cannot be accessed by HTTP/S; your files are in CRAM format. The SSH bridge works in the following way:

1. Establish SSH connection to remote machine.
2. Dalliace will make an HTTP/S request to Snoopy's local server for a specific sequence file (x.bam) at a specific location (chr:start-end).
3. HTTP/S request is parsed and turned into a samtools command: *samtools view x.bam chr:start-end*.
4. The samtools command is sent, via SSH, to remote machine.
5. The results of the command is sent back to the local machine over SSH.
6. The output of the samtools command is parsed to JSON.
7. JSON is served over local HTTP sever to the Dalliace browser.

This is the most flexible method, but as there quite a few steps involved in the process **this is also the slowest**.

Modes Manual, Batch

File types Bam, Cram

Command line arguments `--ssh SSH, -s` (see *Starting Snoopy* for more details)

2.2.3 Summary

| Access mode | File Types | Mode |
|-------------------|------------|---------------|
| Browser Load | BAM | Manual |
| Local File Server | BAM | Manual, Batch |
| HTTP/S | BAM | Manual, Batch |
| SSH Bridge | BAM, CRAM | Manual, Batch |

Note: When loading Local BAM files through the browser, you will also need to specify the accompanying BAI file. For the other access modes, as long as the BAI files exist in the same directory and have corresponding file names (ie `x.bam ==> x.bam.bai`) you do not need to explicitly load them.

2.3 Input File Formats

2.3.1 Variant List File

The variant text file can list both SNPs and CNVs. SNPs can be in any of the following formats:

```
chr:location
chr-location
chr,location
chr location
```

The format for CNVs is as SNPs except the location consists of two numbers. For example, a CNV location may be `16:start-end`.

2.3.2 Batch JSON File

A JavaScript Object Notation (JSON) file is used to specify a set of sessions when using the batch mode of Snoopy. The idea is to use construct the JSON file with some scripting language (e.g. Python or Perl) rather than having to manually loading files. The general structure of the batch file is an array of sessions as follows:

```
sessions:
  session 1:
    variants: array of variant locations | a file listing variant locations
    sequence_files: array of sequence files (BAM or CRAM)
  session 2:
    variants: array of variant locations | a file listing variant locations
    sequence_files: array of sequence files (BAM or CRAM)
  .
  .
  .
```

In the JSON format:

```
{
  "sessions": [
    {
      "variants": [
        "chr-loc",
        "chr-loc",
        "chr-loc"
      ],
      "sequence_files": [
        "path/to/sequence file 1",
        "path/to/sequence file 2",
        "path/to/sequence file 3"
      ]
    },
    .
    .
    .
  ]
}
```

The format of the variants takes the form as that in the [Variant List File](#) section.

Example

For an explicit example of a batch file:

```
{
  "sessions": [
    {
      "variants": [
        "16-48000491",
        "16-48001121",
        "16-48001200"
      ],
      "sequence_files": [
        "/users/joe/examples/mother1.bam",
        "/users/joe/examples/father1.bam",
        "/users/joe/examples/offspring1.bam"
      ]
    }
  ]
}
```

```
    },
    {
      "variants": [
        "12-18001",
        "12-1820491",
        "14-1803735",
        "15-1840848",
      ],
      "sequence_files": [
        "/users/joe/examples/individual.bam"
      ]
    }
  ]
}
```

Note: To test whether your JSON batch file is valid you can use the online tool [JSONLint](#)

Starting Snoopy

To start Snoopy goto the command line and enter:

```
$ snoopy <options>
```

Where options are summarised here:

- h, --help** show help message and exit
- local-server, -l** turn on local file server
DEFAULT: local-server not switched on
- port PORT, -p PORT** set the local HTTP server port number
DEFAULT: 4444, or next available port
- ssh SSH, -s SSH** *user@hostname* for SSH connection to sequence files on remote host
DEFAULT: SSH-Bridge not switched on

See the section *Loading Data* for more information about these options.

3.1 Examples

3.1.1 Start local server

```
$ snoopy -l
```

3.1.2 Start local server at port 8888

```
$ snoopy -l -p 8888
```

3.1.3 Start SSH-Bridge with username bob at big-bio-server

```
$ snoopy -s bob@big-bio-server
```

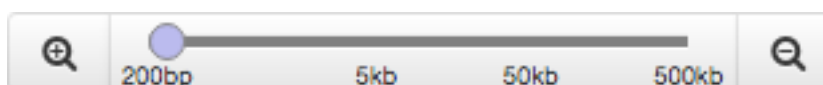

Perform Quality Control

After starting snoopy from the terminal, a new browser tab will open and present you with the first screen: mode selection. Using *Loading Data* as a guide, select the relevant mode and load your data. It's time to view each of the variant sites and record your decision. The following is a walk through guide, for a description of specific parts of the user interface refer to *User Interface*.

1. Upon starting, the first variant in the first session will be viewed.

2. **Explore the current variant location:**

- Drag the Dalliace track around.

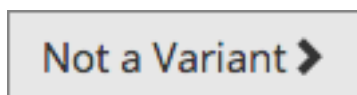


- Zoom in or out with
- There are several different view styles to present the sequence data which can be selected with the

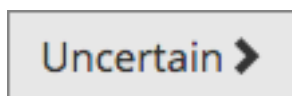


dropdown button.

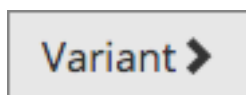
3. **Make a decision about the called variant site.**



- - You are certain that the called site is not actually a variant.



- - You are unsure if the called site is a variant.

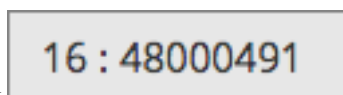


- - You are certain that the called variant is truly a variant.




4. Take a snapshot (PNG) of the current view with

5. After each decision, Snoopy will load the next variant in that session, or if you have reviewed all in that session, it will load the next session's sequence files and the first variant location.




6. If you wish to review your QC decisions made so far, click. From window you can also quickly navigate to a different variant too.



7. Save your results so far with . Refer to `/file_formats` for information about the file format in which the results are saved.

8. Once you have reviewed all variants in all of the sessions, you will be presented with a save dialoge.



9. If at any point you wish to stop reviewing the loaded sessions and start again click .

Saving Results

5.1 JSON

It is possible to export JSON files with the same as described in `batch-json-file` augmented with your QC decisions and any snapshot names (see next section) that you may have taken. The general structure is:

```
{
  "date": <date JSON file created>,
  "sessions": [
    {
      "sequence_files": [
        "path/to/sequence file 1",
        "path/to/sequence file 2",
        .
        .
        .
      ],
      "variants": [
        {
          "chr": <chr of variant>,
          "location": <base position of variant>,
          "qc_decision": <decision made when reviewing>,
          "snapshot": <PNG filename>
        },
        .
        .
        .
      ]
    },
    .
    .
    .
  ]
}
```

5.2 Snapshots

To accompany your QC decisions, a zip file of PNG snapshots can also be saved. These images are generated automatically if specified in [Settings](#) or when clicking the snapshot button [Snoopy](#). The filenames of the snapshots are formed by concatenating the sequence files and the variant location.

5.3 HTML

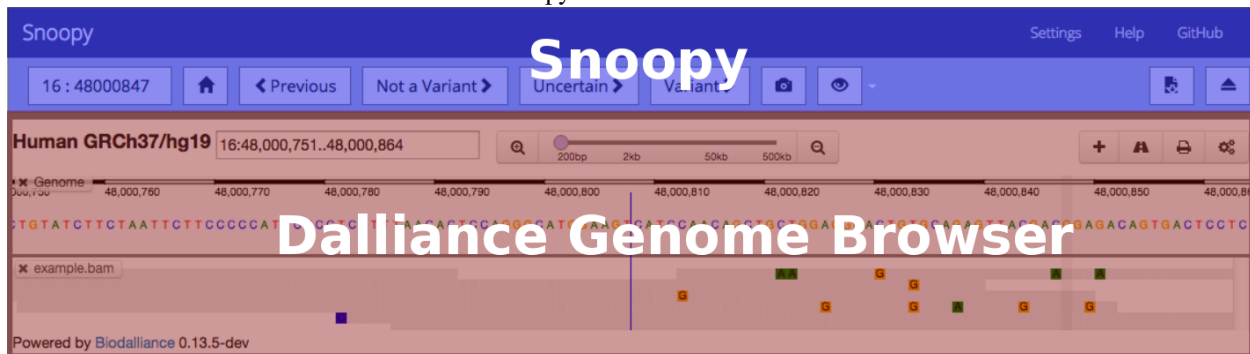
The HTML export is a convenient file which combines the JSON results and snapshots described in the sections above into a single standalone file. When opened in a web-browser (it is a static file so no web-server is needed) you can click through the sessions and variants to view recorded QC decisions and snapshots.

User Interface

When reviewing variants you'll have the following view:



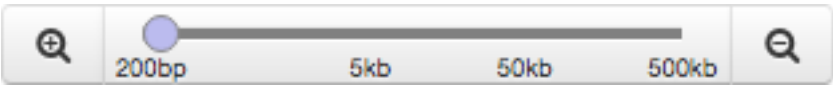
There are two main sections in this view: the Snoopy interface and the Dalliance interface.







6.1 Dalliance

Within Dalliance you have access to following:

- **Human GRCh37/hg19** 16:48,000,391..48,000,591
Lists the genome reference being used and the current chromosome and location. You can change the location by entering either a start and end position or a single base position.

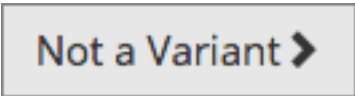
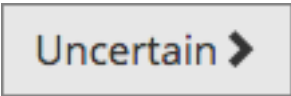

-  - Control the zoom depth.

-  - Adds a new track to the current view.
-  - Modify track options such as the max depth, and colors.
-  - Export current view as an SVG or PNG.
-  - Change some of Dalliance's options such as vertical guideline location, scrolling direction.



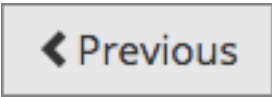
6.2 Snoopy

6.2.1 Variant Decisions

When either of the decision buttons are clicked, you will advance to the next variant.

-  - You are certain that the called site is not actually a variant.
-  - You are unsure if the called site is a variant.
-  - You are certain that the called variant is truly a variant.

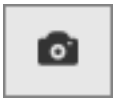
6.2.2 Navigation

-  - This displays the current variant and a QC decision, if available. If you click this button you will be presented with a window which summarizes all of the QC decisions made so far, as well as allowing you to quickly navigate to a different variant.
-  - Returns to the current variant of interest if you've dragged the Dalliance track away.
-  - Go to the previous variant. Clicking this button does not register any variant decisions.

6.2.3 Viewing



- **This dropdown button allows you to select four different track styles:**
 - Raw - Display the bases with color scheme as specified in the settings.
 - Condensed - Only display a base if it differs from the reference. If a read exists and is in agreement with the reference, use match color as set in your settings.
 - Mismatch - Color code plus strand / minus strand if reference agreement exists. If a base differs, display with the base color as given in settings for raw.
 - Coverage - Presents a histogram of coverage. If more than 20% of the bases differ from reference, the proportion of bases are displayed with their bases colors.



- - Clicking this will take a snapshot of whatever view is currently loaded into Dalliance.

6.2.4 Admin



- - Stop and restart Snoopy button, there will be a prompt asking if you wish to save your progress.



- - Save whatever progress you have made so far. The output format is described in [Saving Results](#).

6.2.5 Settings

View and change the following settings:

- **Connections**
 - Default Remote, Local and SSH-Bridge settings (URL, credential requirements)
- **Dalliance zoom level**
 - How deep should Dalliance be zoomed into after variant decision
- **Color settings**
 - nucleotide bases colors
- **Mismatch style**
 - Plus/minus strand colors
 - Show insertions
 - Reflect base quality with transparency
- **Condensed style**
 - Match color
 - Reflect base quality with transparency

- **Coverage histogram style**
 - Allele threshold (between 0 and 1): the minimum ratio of allele frequencies before a mismatch is displayed
 - Height
- **Snapshots**
 - Automatically take snapshots at each variant

6.2.6 Help

A link to the documentation hosted on Read The Docs.

6.2.7 GitHub

A link to the GitHub repository holding the snoopy's source code.