Millstone Documentation

Release 0.5

Dan Goodman, Gleb Kuznetsov, Kevin Chen, Changping Chen, Ma

Mar 07, 2018

User Guide

Millstone is a distributed bioinformatics software platform designed to facilitate genome engineering and evolutionary genomic analyses. With Millstone, you can automate and iterate genome analysis and debugging for sequencing projects involving tens to hundreds of microbial genomes.

CHAPTER 1

Introduction to Millstone

1.1 What can Millstone do for me?

Millstone 0.5 currently does:

- reference-based read alignment for multiple genomes
- single nucleotide variant calling & annotation
- structural variant calling & annotation
- visualization of variants via Jbrowse
- de-novo assembly and placement of unaligned reads into contigs
- genome versioning and creation & export of new reference genomes
- variant analysis among many genomes (i.e. searching, comparison, filtering)
- design of MAGE oligos to create or revert variants

Millstone is still in active development, and there are bound to be some bugs. Thanks for helping us find them! Please report them at our github repository.

1.2 How do I get a Millstone server of my own?

Currently the best way to use Millstone is through Amazon AWS. Using Amazon allows you to avoid the complexity of installing all the dependencies from scratch on your own server, so this should be the quickest and easiest way to deploy for most users. It requires registering an Amazon AWS account. For projects under 50 genomes, a suitable Amazon instance should cost less than 2 dollars per day. It is easy to stop and start an instance when not in use. Advanced users can also deploy their own Millstone instance locally.

We plan to write up a more complete AWS cost-guide in the future.

1.3 What do I need to get started?

You really just need two things to use Millstone:

One or more reference genomes (Genbank or FASTA format): If you are using a FASTA genome, you obviously cannot use SNPEff's variant annotation, so we recommend Genbank if it is available. If your genome is on Genbank, Millstone can pull the record straight from NCBI.

Note: Millstone is meant for smaller genomes (i.e. not H. sapiens). We use Millstone with E. coli genomes (4.6 MB) but Millstone should work well for most microbial genomes like Saccharomyces. Try larger genomes at your own risk.

Illumina HiSeq/MiSeq FASTQ Reads for one or more samples: We've thoroughly tested our pipeline with paired-end data, but single-end should work as well. You need two files per sample, one for read 1, and one for read 2. Millstone cannot (yet) split on multi-sample barcodes or on interleaved paired-end reads, so you'll have to do that yourself beforehand.

Note: Extremely high-coverage samples and short fragments with non-overlapping reads might cause difficulties. Try at your own risk. You might consider downsampling or cleaning the reads first

CHAPTER 2

Setting up Millstone

2.1 Using Millstone on Amazon AWS

Using Millstone via AWS is the preferred option for most users. We have pre-configured a Millstone installation into an Amazon Machine Image (AMI). This means you can sidestep all of the dependency installation, configuration, etc.

DISCLAIMER: The current Millstone Amazon setup leaves your application open to the web. Even though user accounts are password-protected, certain uploaded and/or processed data is downloadable without authentication if others "guess" the right urls. Realistically, this shouldn't be a problem for most projects, but we're letting you know just in case.

2.1.1 Create an Amazon AWS Account

You need to create to an Amazon Web Services (AWS) account. Brad Chapman's getting started guide for cloudbiolinux has a solid first chapter with instructions on getting everything set up.

2.1.2 Cloning the AMI

1. Login to https://console.aws.amazon.com/console/home and proceed to EC2. In the upper-right corner, be sure to select the N. Virgina region. We can't guarantee our AMI is visible outside of that region. From the EC2 dashboard, press `Launch Instance`, which will take you into a Wizard to have you configure your instance.
2. In the Choose AMI tab, select Community AMIs in the left panel, then search for "millstone". The Millstone AMI will have a name of the form 'millstone_combined_YYYY_mm_dd_hash'. Select the newest version.
3. On the 'Choose instance type' tab, select an instance according to your needs. We recommend m3.medium (select General Purpose on the left). The number of vCPUs will determine how many genomes can be simultaneously aligned.
4. In 'Configure instance', the only setting we recommend changing is explicitly setting the Availability Zone (we always use `us-east-1a`). You can only move EBS (Amazon hard drives) between instances in the same zone, so it'll make things easier to consistently make everything in the same zone.

5. In 'Add storage', increase the size of the root drive to the amount of space that you'll need. For bacterial genomes, about 2 GB per sample should be more than enough (i.e. 100 samples = 200 GB).
6. In 'Tag instance', fill in an informative value for the 'Name' key. We like the name to include the date it was created and a description of what the instance is running (e.g. `2014_04_01_mutate_all_the_things`).
7. For security group, configure a group appropriate to your needs. Most users will want to create a security group with all of the following open. (*This will make your instance publicly visible to someone trying random EC2 IPs, but login is still required.*):
    - All ICMP
    - All TCP
    - All UDP
    - SSH
8. Continue to the final tab where you'll press 'Launch the instance'. Select or create a public/private key pair. If you create the key, download and save the private key, and put it somewhere safe (we suggest `~/.ssh/`.) (*If you lose the private key there's no way to ssh back into your instance. You'll have to terminate it and create a new one.*)

It takes about 5-10 minutes for the instance to launch and all bootstrapping to finish, after which your Millstone is ready to grind!

2.1.3 Accessing your instance

Go back to the EC2 console Instances page and make sure you are in the correct region, using the dropdown in the top right. The instance you created should be visible in the list. When it is ready, its *Status Checks* column should say '2/2 checks passed'.

In the browser

Select the instance from the list, and the info pane should appear below the instance list. In the Description tab, the webpage URL to can be found under **Public DNS**. The url should look like: `ec2-xx-xx-xx-xx.compute-1.amazonaws.com`

It may take some time for your instance to initialize. Wait until all status checks are completed before attempting to log in. If the server doesn't come up, it might still be loading.

On the command line (just in case)

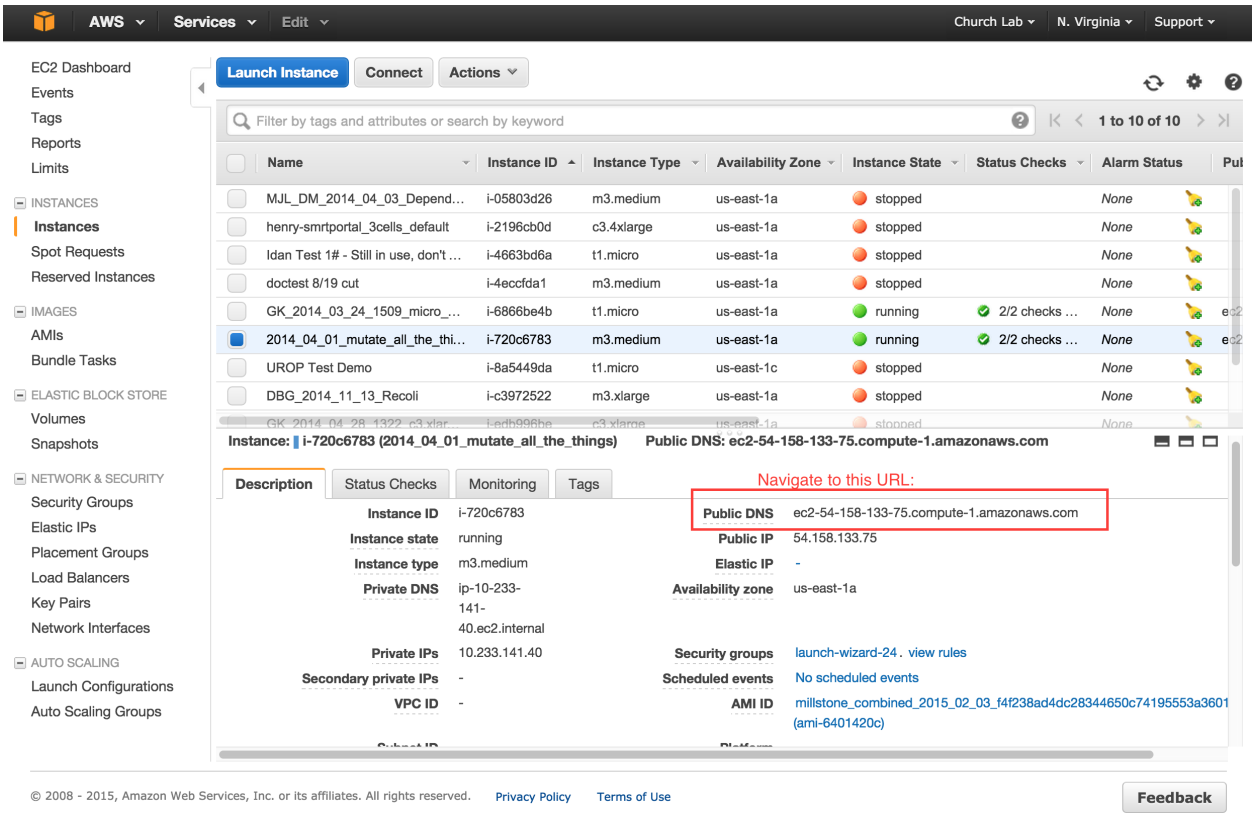
It should not be necessary at the moment, but if you need to SSH into the server, the command is:

```
ssh -i ~/.ssh/your-key.pem ubuntu@ec2-xx-xx-xx-xx.compute-1.amazonaws.com
```

(This assumes you put the private key you generated in `~/.ssh/`). If permissions fail on your key, `chmod` the key's permissions to 700.

2.2 Using Millstone locally

It is also possible to use Millstone locally on Mac OSX and Linux. **Local installation is meant for advanced users only**. It requires the manual installation and configuration of various dependencies, and requires root access. You can our find local installation guide in the Millstone github readme.


AWS
Services
Edit
Church Lab
N. Virginia
Support
EC2 Dashboard
Events
Tags
Reports
Limits
INSTANCES
Instances
Spot Requests
Reserved Instances
IMAGES
AMIs
Bundle Tasks
ELASTIC BLOCK STORE
Volumes
Snapshots
NETWORK & SECURITY
Security Groups
Elastic IPs
Placement Groups
Load Balancers
Key Pairs
Network Interfaces
AUTO SCALING
Launch Configurations
Auto Scaling Groups
Launch Instance
Connect
Actions
Filter by tags and attributes or search by keyword
1 to 10 of 10
Name
Instance ID
Instance Type
Availability Zone
Instance State
Status Checks
Alarm Status
MJL_DM_2014_04_03_Depend…
i-05803d26
m3.medium
us-east-1a
stopped
None
henry-smrtportal_3cells_default
i-2196cb0d
c3.4xlarge
us-east-1a
stopped
None
Idan Test 1# - Still in use, don't …
i-4663bd6a
t1.micro
us-east-1a
stopped
None
doctest 8/19 cut
i-4eccfda1
m3.medium
us-east-1a
stopped
None
GK_2014_03_24_1509_micro_…
i-6866be4b
t1.micro
us-east-1a
running
2/2 checks …
None
2014_04_01_mutate_all_the_thi…
i-720c6783
m3.medium
us-east-1a
running
2/2 checks …
None
UROP Test Demo
i-8a5449da
t1.micro
us-east-1c
stopped
None
DBG_2014_11_13_Recoli
i-c3972522
m3.xlarge
us-east-1a
stopped
None
Instance: i-720c6783 (2014_04_01_mutate_all_the_things)
Public DNS: ec2-54-158-133-75.compute-1.amazonaws.com
Description
Status Checks
Monitoring
Tags
Navigate to this URL:
Instance ID
i-720c6783
Instance state
running
Instance type
m3.medium
Private DNS
ip-10-233-141-40.ec2.internal
Private IPs
10.233.141.40
Secondary private IPs
-
VPC ID
-
Public DNS
ec2-54-158-133-75.compute-1.amazonaws.com
Public IP
54.158.133.75
Elastic IP
-
Availability zone
us-east-1a
Security groups
launch-wizard-24. view rules
Scheduled events
No scheduled events
AMI ID
millstone_combined_2015_02_03_f4f238ad4dc28344650c74195553a3601 (ami-6401420c)
© 2008 - 2015, Amazon Web Services, Inc. or its affiliates. All rights reserved.
Privacy Policy
Terms of Use
Feedback

CHAPTER 3

Projects and Alignments

3.1 Registering a new user

Once Millstone is installed, you should be greeted with the Millstone logo and a login/register page. Register a user with a login, email, and password. *Currently, we only allow one user per instance. After the first user is registered, registration is closed.* **Don't forget your username and password, as there is currently no 'reminder' functionality.** (The only way to change your password at present is to do so through the Django shell, using methods available on the Django auth model `User`.)

3.2 Creating a new project

Once you register, you can create a new project, and you will the prompted to give it a short name. Afterwards, you will be taken to the create alignment screen. There are 5 steps, each with a tab in the top bar. Choose a name for your first alignment, which will pair a reference genome with a set of samples to align. One project can have multiple alignments.

Note: If you have many/large samples, and would prefer to upload files via the command line instead of the browser, see this guide.

3.2.1 Reference Genome

Select the Reference Genome tab, and click the green 'New' button. You can select a reference genome from NCBI or upload a custom reference.

Note: If you use a FASTA there will be no variant annotation information, so Genbank is recommended if you have one.

Load file from NCBI: Simply fill in the accession number (for instance U00096.2 for E. coli) and give the reference genome a name. If you'd like to use a custom reference genome, you can upload a file from your desktop. You can check to make sure you've got the right accession number by comparing your genome's size to the number of nucleotides present in the reference genome.

Upload through browser: If you have a local file with your genome, you can upload it with this option. If you have a large cassette insertion or plasmid you would also like to align, you can edit the FASTA/Genbank file to insert it into the genome using a tool like Benchling or Geneious (in the case of a cassette insertion), or add it as a separate chromosome (an additional FASTA or GenBank record in the same file).

Finally, select the checkbox next to the uploaded genome to mark it as your reference.

3.2.2 Samples

Once that's done, move on to the samples tab. Each genome sample you upload must consist of a pair of forward and reverse FASTQ files. You can either upload samples through the browser, or you can upload them in batch to the server using a the command line via `scp`. The command line approach is better for large numbers of samples, but is more complicated. It is detailed in the *Manual Upload* section at the bottom of this guide.

Open the upload samples dialog via the green 'New' button, then choose 'Batch Upload through browser...'. In order to upload samples through the browser, you must first register samples to be uploaded by filling out a spreadsheet template with sample labels and corresponding data filenames (no path required). *Fields must be separated by tabs.* Here is an example:

```
Sample_Name Read_1_Filename Read_2_Filename
sample01    sample01_fwd.fq.gz  sample01_rev.fq.gz
sample02    sample02_fwd.fq.gz  sample02_rev.fq.gz
```

NOTE: Millstone can work with ''gzip''-ed FASTQ files, and they will be faster to upload.

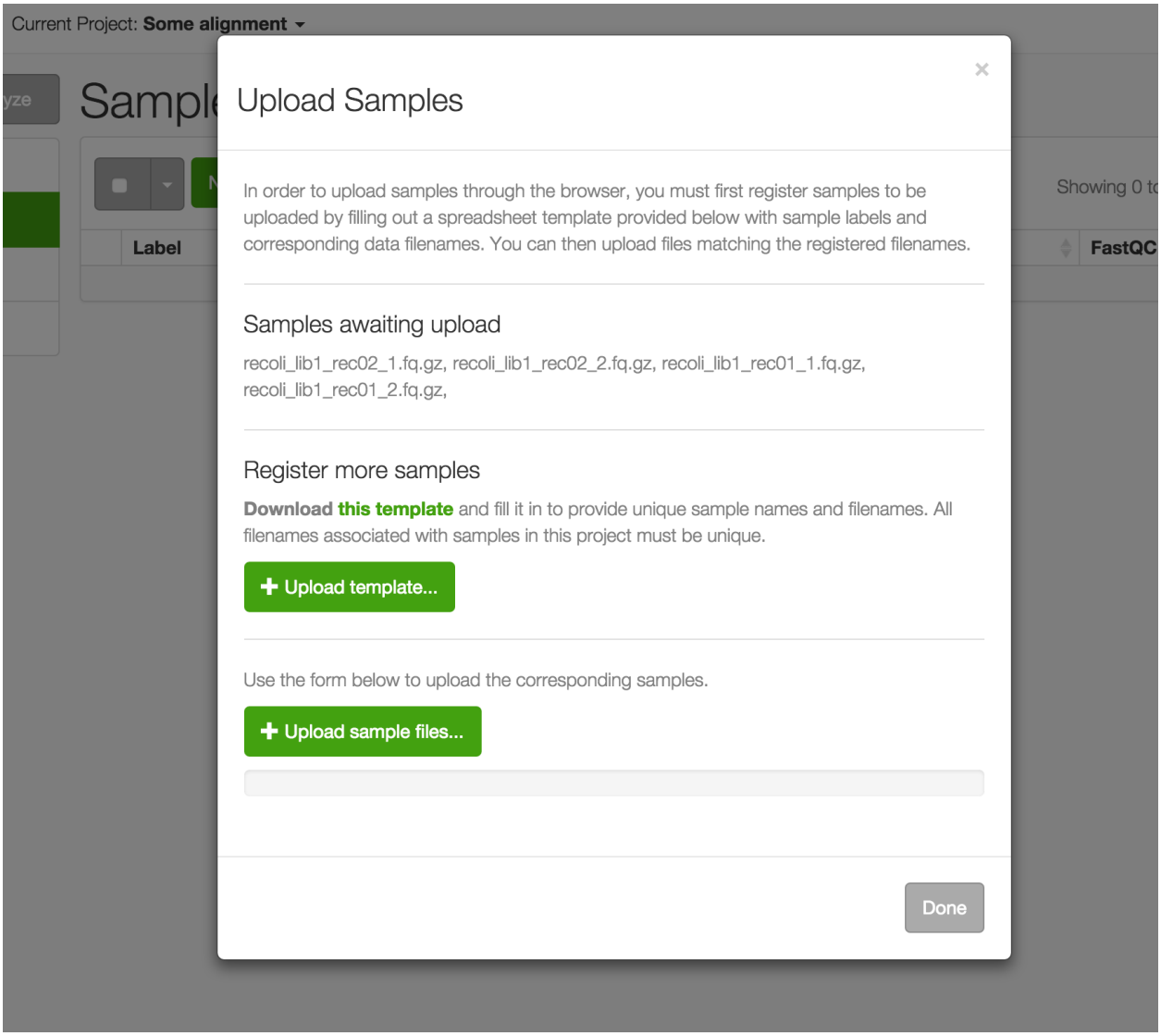
You can also include additional columns as per-sample metadata, like growth rates, plate and well, strain parentage, etc. Here is an example:

```
Sample_Name        Plate_or_Group        Well        Read_1_Path        Read_2_
↪Path        Parents        Growth_Rate
Test Sample 0        Group A        A01        /path/to/genome_0_read_1.fq        /
↪path/to/genome_0_read_2.fq                0.5
Test Sample 1        Group A        A02        /path/to/genome_1_read_1.fq        /
↪path/to/genome_1_read_2.fq        Test Sample 0        0.7
Test Sample 2        Group A        A03        /path/to/genome_2_read_1.fq        /
↪path/to/genome_2_read_2.fq        Test Sample 0        0.6
Test Sample 3        Group A        A04        /path/to/genome_3_read_1.fq        /
↪path/to/genome_3_read_2.fq        Test Sample 0        0.8
Test Sample 4        Group A        A05        /path/to/genome_4_read_1.fq        /
↪path/to/genome_4_read_2.fq        Test Sample 0        0.3
Test Sample 5        Group A        B01        /path/to/genome_5_read_1.fastq.
↪gz        /path/to/genome_5_read_2.fastq.gz        Test Sample 1        0.8
Test Sample 6        Group A        B02        /path/to/genome_6_read_1.fq        /
↪path/to/genome_6_read_2.fq        Test Sample 2        0.7
Test Sample 7        Group A        B03        /path/to/genome_7_read_1.fq        /
↪path/to/genome_7_read_2.fq        Test Sample 3,Test Sample 2        0.8
Test Sample 8        Group A        B04        /path/to/genome_8_read_1.fq.gz        /
↪path/to/genome_8_read_2.fq.gz        Test Sample 4        0.3
Test Sample 9        Group A        B05        /path/to/genome_9_read_1.fq        /
↪path/to/genome_9_read_2.fq                0.2
```

Once you upload the template, it will list the samples awaiting upload:

You can then upload the individual files matching the filenames in the template. Note that if you have many files, it may be necessary to select in batches of 10 or so files.

To confirm that files are uploaded, close the upload dialog and look at the table of samples. The status hsould change to RUNNING_QC, or when that's done, you should see a QC link. If the status still says AWAITING_UPLOAD, then


Current Project: Some alignment
Upload Samples
In order to upload samples through the browser, you must first register samples to be uploaded by filling out a spreadsheet template provided below with sample labels and corresponding data filenames. You can then upload files matching the registered filenames.
Samples awaiting upload
recoli_lib1_rec02_1.fq.gz, recoli_lib1_rec02_2.fq.gz, recoli_lib1_rec01_1.fq.gz, recoli_lib1_rec01_2.fq.gz,
Register more samples
Download this template and fill it in to provide unique sample names and filenames. All filenames associated with samples in this project must be unique.
Upload template...
Use the form below to upload the corresponding samples.
Upload sample files...
Done
Label
FastQC

the upload didn't go through successfully and you should try again.

3.2.3 Alignment Settings

By default, Millstone treats all samples as diploid. This allows ambiguous variants to be called as heterozygous. You can choose to keep all of these ambiguous variants, to keep only those where at least some samples are called as non-ambiguous, or throw away ambiguous variants all together. If you have many samples, we suggest the latter two options to keep the database size manageable.

3.2.4 Submit Alignment

Finally! Click the *Run Alignment* button in the last tab to start the alignment. Depending on your genome size, number of samples, and the size of the instance you chose, this could take time. You can see how individual sample alignments are progressing by clicking on the name of the alignment in the label column of the Alignments view. Every sample will have an *output log* link and a Job Status.

After the individual samples are done aligning, the Alignment status will change to `VARIANT_CALLING` as variants across all samples are called in aggregate. Once this step has completed, then the Alignment status will read `COMPLETED` and you can switch to the *Analyze* view to examine the called variants.

CHAPTER 4

Troubleshooting

- I can log in via SSH but the web interface doesn't load!

  You've probably forgotten to allow access to your instance through web interfaces. This can be fixed by adding the following connections to your security group: * All ICMP * All TCP * All UDP You can do this by going to the Network & Security -> Security Groups section of the EC2 dashboard and editing the security group that you created in your instance. If you've forgotten this can be found in the main instance dash on the far right under security groups. Click on that and you should be able to edit inbound rules by right clicking on the Group ID

- I've managed to load the webpage but get a 502 bad gateway error!

  Millstone is probably loading up, try again in a few minutes.

- Registration is closed.

  Only one user is allowed to register (as soon as the server boots up), and afterwards registration is closed.

- Millstone just sits there after importing a template file.

  This could be any number of things. If your template file is formatted correctly, it could be a completely out of space error, so check that you've got room on your drive containing Millstone. File formatting is often the biggest problem in this stage, so be careful that you've escaped spaces in file names.

- I want to make sure everything's going right, where can I find the logs?

  The logs are by default at /var/log/supervisor

CHAPTER 5

Variants

All sample, reference genome, and alignments are listed in the *Data* view (the toggle switch in the top left). Clicking over to the *Analyze* view will allow you to filter through multi-sample variants and view their aligned reads. Use the dropdowns on the left below the Data/Analyze toggle to select your alignment and your reference genome and choose *Variants*. Once the alignments are complete, you should see a list of all variants that have been identified across all samples.

5.1 Cast vs. Melted

There are two ways to view variants.

Cast: Cast displays a summary row for one variant across all samples. You can see how many samples the variant is present in, as well as the variant's effects.

Melted: 'Melting' the view shows one row for every combination of sample and and variant. It essentially multiplies the rows by the number of samples, so you can see data specific to individual samples. If a variant is not called in a sample, it's *Alt* column will be blank.

5.2 Links

There are three link icons next to every sample.

- *The magnifying glass* icon 'zooms in' to the melted view for that variant across all samples.
- *The read alignment* icon shows how individual fastq reads align around a variant. It is useful for doing visual QC on an alignment, to make sure your reads are properly aligned around your variant.
- *The bar graph* icon shows the coverage of your reads. Areas of high or low coverage might be of interest, and this view is more compact, which makes it easier to compare multiple samples.

Note: If an icon is gray in the Cast view it is disabled because it is too intensive to display many samples simultaneously. Zoom into the variant (with the magnifying glass) and inspect individual samples. You can manually add and remove tracks in Jbrowse via the track list on the left.

More information about using JBrowse and understanding its visualization can be found at its website.

5.3 Fields and Filtering

Millstone uses a simple language to understand query syntax for filtering variants.

Note: Currently some of the field names can be confusing. A list of all available fields can be found with the *Fields...* button. The default column names don't always correspond to the internal field names. There isn't currently a well-documented list of what each field means, but most of them are documented in the VCF specification. The `INFO_EFF_*` fields come from SnpEFF.

5.3.1 Examples

If you want to look at all variants in a certain gene:

```
INFO_EFF_GENE = tolC
```

If you want to look at all variants that have strong or moderate predicted phenotypic effects:

```
INFO_EFF_IMPACT = HIGH | INFO_EFF_IMPACT = LOW
```

If you want to look in a certain region:

```
CHROM = NC_000913 & POSITION > 500 & POSITION < 1000
```

5.4 Marginal Calls

We always run variant calling as diploid, even for haploid organisms like E. coli, so that some poorly-supported variants appear heterozygous. This allows marginal calls to be made in cases where only a portion of the reads show a SNV, in cases of regional duplications or if reads map to a non-unique region of the genome. Such marginal calls have an orange fraction icon in their ALT column, and can also be filtered on by using:

`IS_HET = TRUE` or `IS_HET = FALSE`

Additionally, the `GT_TYPE` field is another way to distinguish marginal from strongly called variants. `GT_TYPE` can take values between 0 or 1 for each sample/variant combination:

- 0 means the variant was called as reference in the sample
- 1 means the variant was called as heterozygous (i.e. marginal) in the sample
- 2 means the variant was called as homozygous (well-supported) in the sample

If you'd like to filter on only well-supported variants that have moderate to strong affects on genes, you can use the filter:

```
GT_TYPE = 2 & (INFO_EFF_IMPACT = HIGH | INFO_EFF_IMPACT = MODERATE)
```

5.5 Variant Sets

Variant sets are a way to group variants after filtering. The sets created by default correspond to regions where the alignment had problems; either there was insufficient coverage, no coverage, too much coverage, or poor mapping quality (corresponding perhaps to regions that are non-unique).

You can also create your own sets to group interesting variants, or those whose alignments you'd like to examine by eye.

5.5.1 Creating a blank set

You can create your own blank sets from the Sets tab in the *Data* view. After creating a set, you can add variants to it in the *Analyze* view using the checkboxes and the master checkbox dropdown on the left.

5.5.2 Uploading a set from a VCF file

You can also upload a variant set from a VCF file. Only the first 5 columns of the VCF will be used. The file must be tab delimited. Here is an example:

```
#CHROM          POS ID  REF ALT
NC_000913   2242    .   G   A
NC_000913   76  .   C   A
NC_000913   3170    .   T   C
NC_000913   1623    .   G   C
NC_000913   3879    .   A   G
NC_000913   3112    .   A   T
NC_000913   1577    .   C   T
NC_000913   5352    .   G   A
NC_000913   4386    .   A   T
NC_000913   1167    .   G   T
NC_000913   5425    .   T   A
NC_000913   951 .   C   A
NC_000913   3993    .   A   G
NC_000913   226 .   G   C
NC_000913   2939    .   T   G
NC_000913   92  .   C   A
NC_000913   5563    .   A   C
NC_000913   4446    .   A   C
NC_000913   607 .   A   G
NC_000913   5088    .   A   T
```

This way, you can identify variants you expected to be called in your samples, such as alleles targeted by MAGE oligonucleotides.

CHAPTER 6

Read Alignment Pipeline

CHAPTER 7

De-Novo Contig Assembly & Placement

CHAPTER 8

Django Models

CHAPTER 9

Postgres Database

CHAPTER 10

SNV Calling and Annotation

SNV calling is performed by Freebayes.

Annotation of variants is performed by SnpEff. The default arguments to SnpEff are specified in the code here and some can be overriden by modifying local_settings.py and restarting the Millstone web server and celery.

CHAPTER 11

Structural Variant Calling & Annotation

CHAPTER 12

Indices and tables

- genindex
- modindex
- search