VirAmp Documentation

Release 1.0

Yinan Wan

January 18, 2017

Contents

VirAmp is a galaxy-based system for fast virus genome assembly and variation discovery.

Quick Start Guide:

1. Launch the latest version of the “Szpara_Viramp” AMI from Amazon Web Services

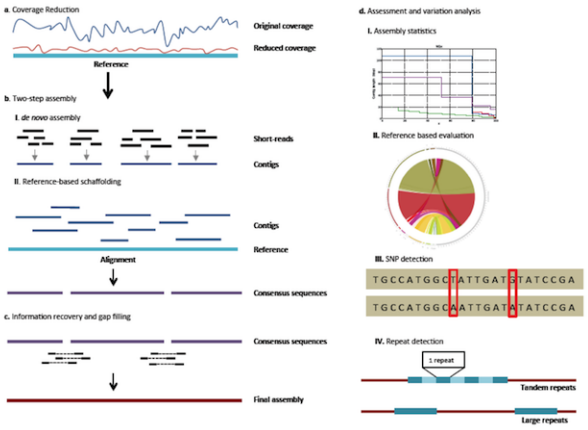
2) SSH into the server and start the run.sh script using screen `./run.sh`

Contents:

CHAPTER 1

Introduction

The following graphic is an overview of how the VirAmp platform works:



Advances in next generation sequencing make it possible to obtain high-coverage sequence data for large numbers of viral strains in a short time. However, since most bioinformatics tools are developed for command line use, the selection and accessibility of computational tools for genome assembly and variation analysis limits the ability of individual labs to perform further bioinformatics analysis.We have developed a multi-step viral genome assembly pipeline named VirAmp, that combines existing tools and techniques and presents them to end users via a web-enabled Galaxy interface. Our pipeline allows users to assemble, analyze and interpret high coverage viral sequencing data with an ease and efficiency that was not possible previously. Our software makes a large number of genome assembly and related tools available to life scientists and automates the currently recommended best practices into a single, easy to use interface. We tested our pipeline with three different datasets from human herpes simplex virus (HSV).VirAmp provides a user-friendly interface and a complete pipeline for viral genome analysis. We make our software available via an Amazon Elastic Cloud disk image that can be easily launched by anyone with an Amazon web service account. A demonstration version of our system can be found at http://www.viramp.com. We also maintain detailed documentation on each tool and methodology at http://docs.viramp.com.

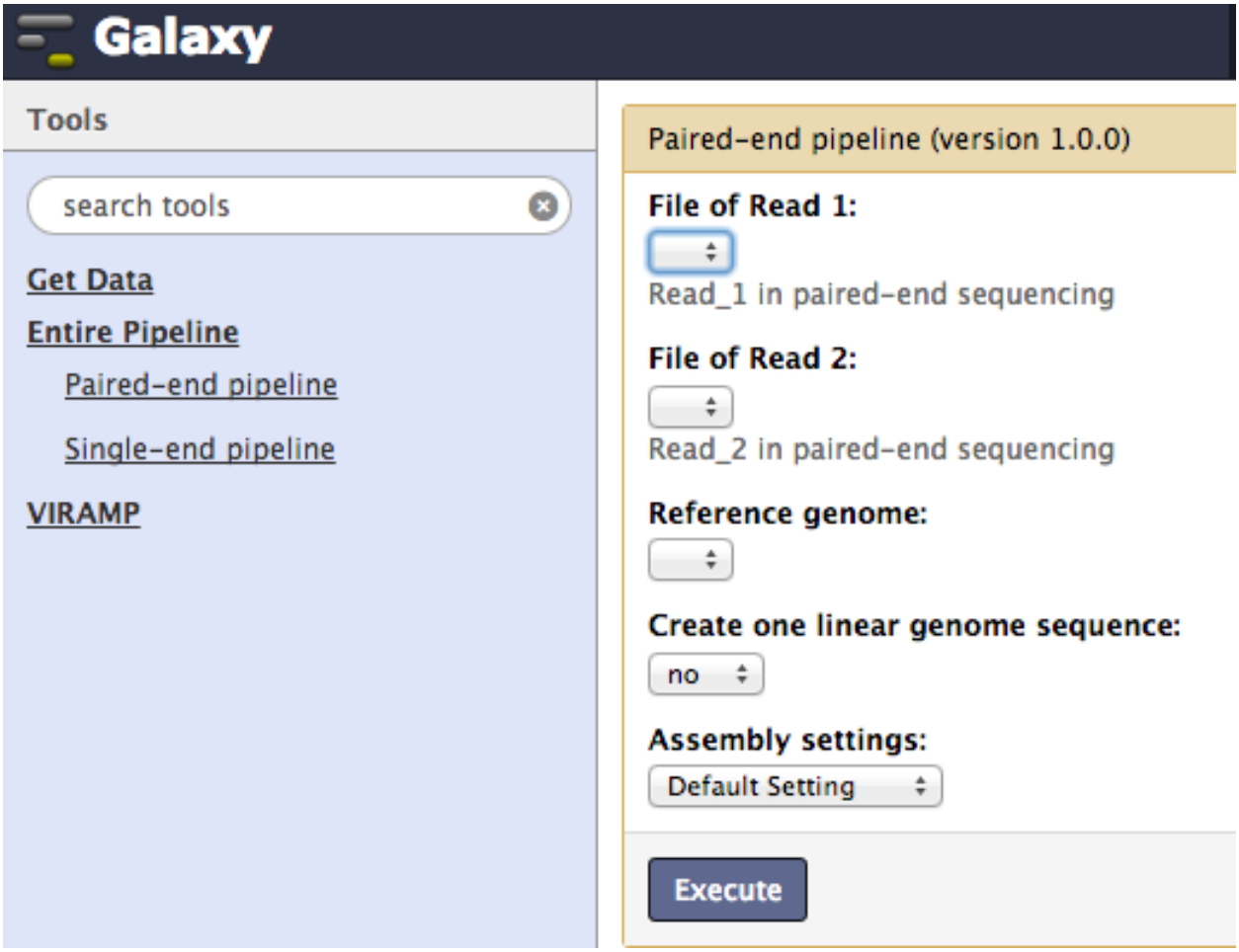
CHAPTER 2

Usage

This is a general description of the usage and function of each tool found in the VirAmp pipeline. A more detailed description can be found at the webpage of each tool.

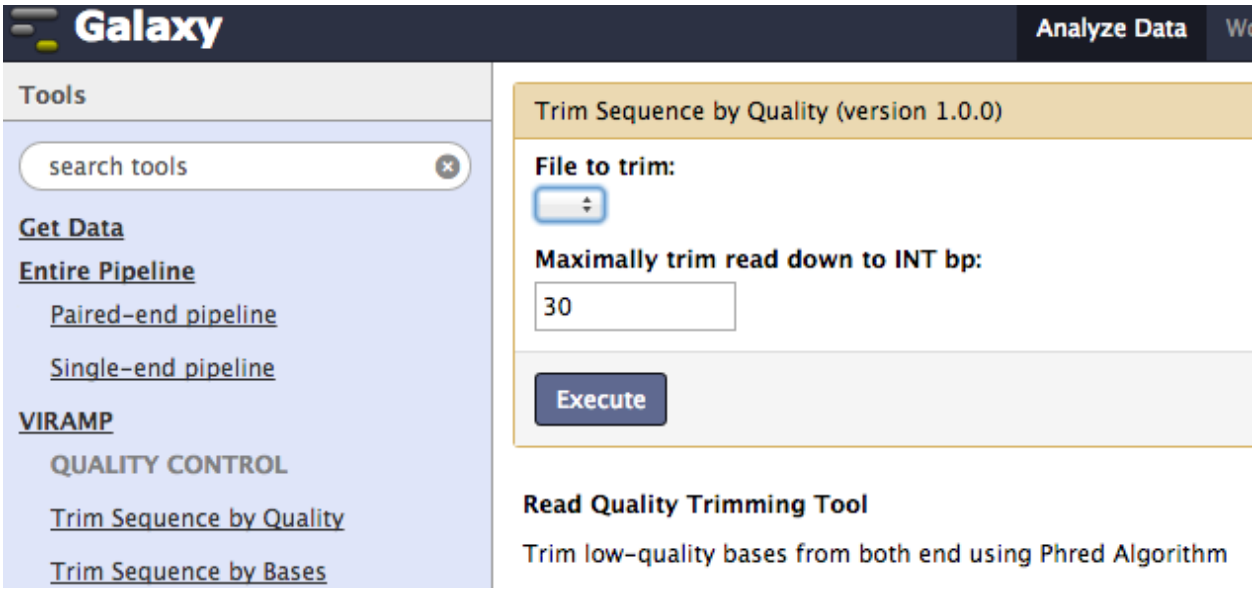
2.1 One-click pipeline

Two general pipelines are provided with a one-click option, one for paired-end data and the other for single-end data. Users are only required to submit read files and a reference file corresponding to their data. Alongside the default settings, users may use the "advanced setting" option to custom configure the pipeline with alternative parameters.



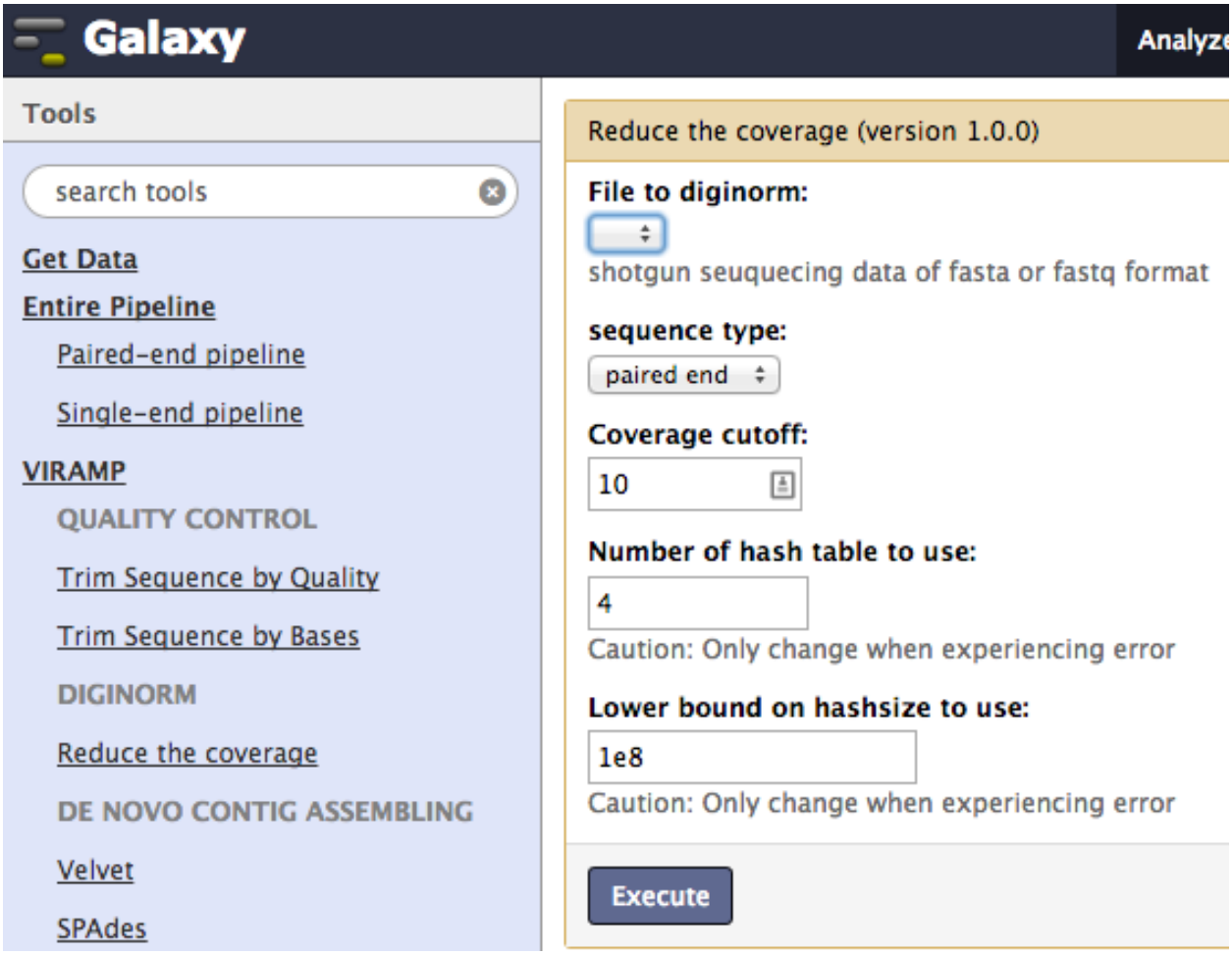
2.2 Quality Control

First, trim out the low quality bases of the input fastq files. This can be achieved by either removing low quality bases or trimming a certain length from each end.



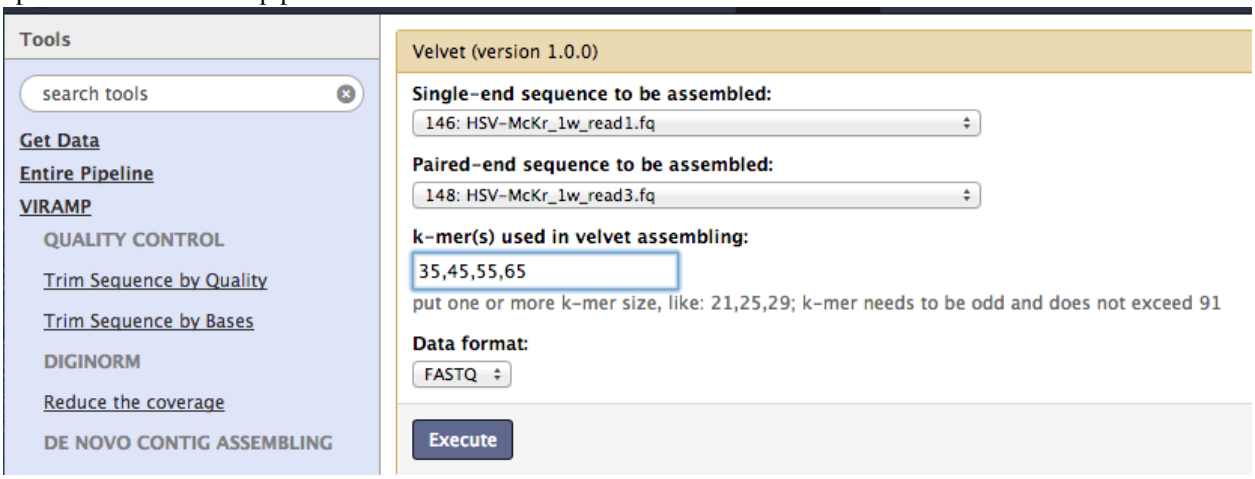
2.3 Diginorm

Next, reduce coverage and bias using Digital normalization. This step reduces the sample variation as well as sample bias.



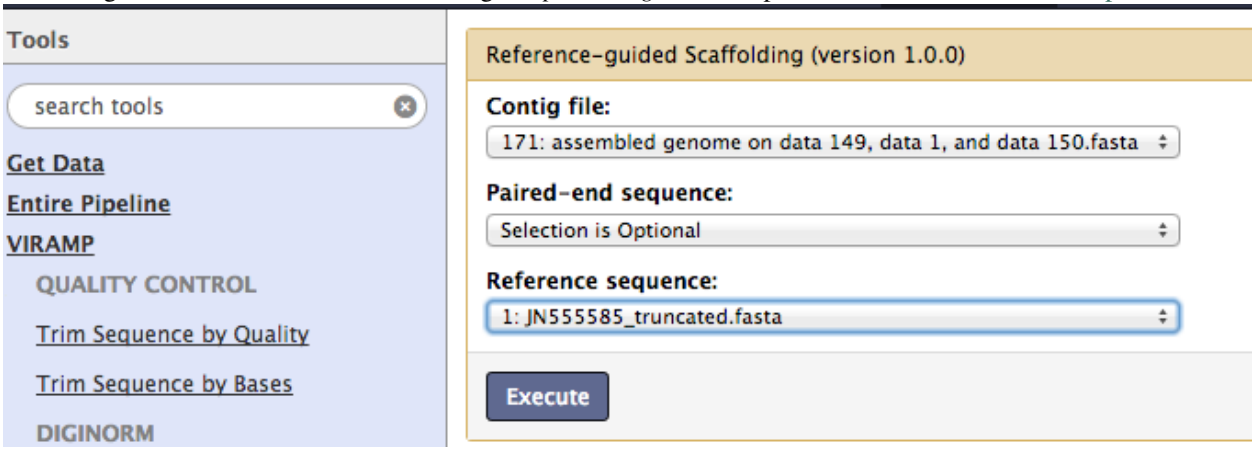
2.4 *de novo* Contig assembly

Now, the pipeline assembles the short reads into longer contigs. By default the **One-click pipeline** uses velvet. Two alternatives, SPAdes and VICUNA , are provided and can be selected as either individual tools or through the advanced options in the one-click pipeline.



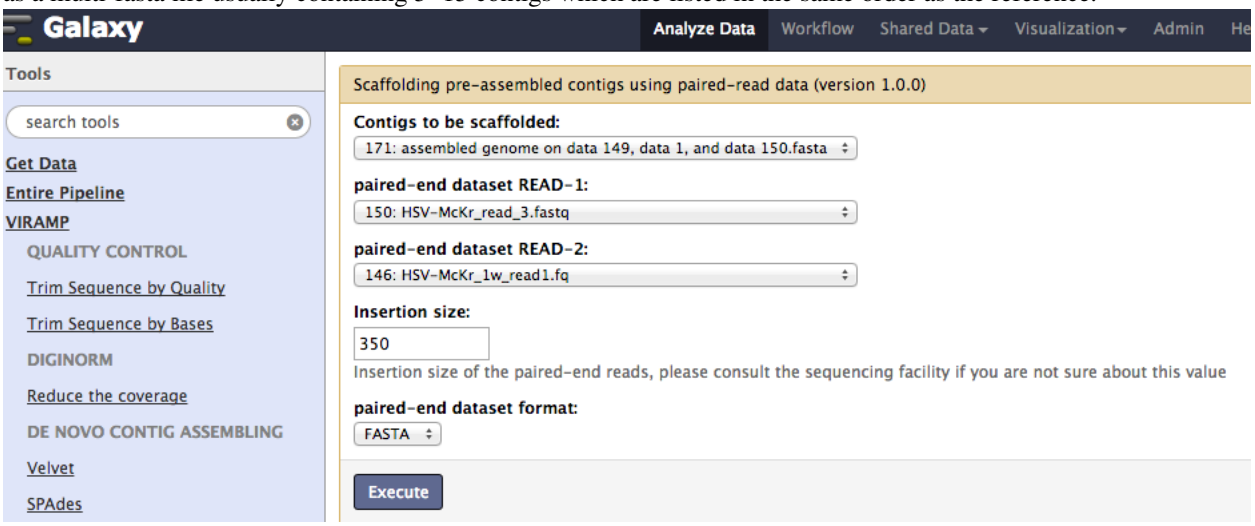
2.5 Reference-based scaffolding

The contigs are then assembled into even longer *super-contigs*. This step is a modification of AMOScmp



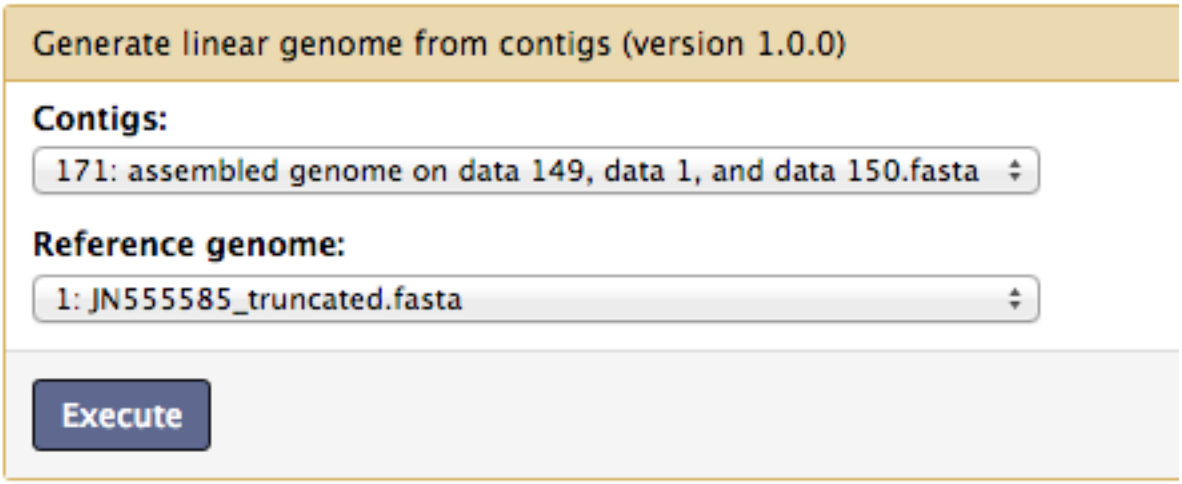
2.6 Reference-independent scaffolding

The next step extends the super-contigs and connects them using SSPACE. The pipeline will produce a draft genome as a multi-fasta file usually containing 5~15 contigs which are listed in the same order as the reference.



2.7 Gap closing

This step connects all the contigs in the multi-fasta from the previous step into one linear genome for the convenience of downstream functional analysis. However, this is **optional** and highly recommended to be done only after assessing the draft genome, as the gaps between the contigs could be from misassembly, sequencing, genome feature, etc.


Generate linear genome from contigs (version 1.0.0)
Contigs:
171: assembled genome on data 149, data 1, and data 150.fasta
Reference genome:
1: JN555585_truncated.fasta
Execute

CHAPTER 3

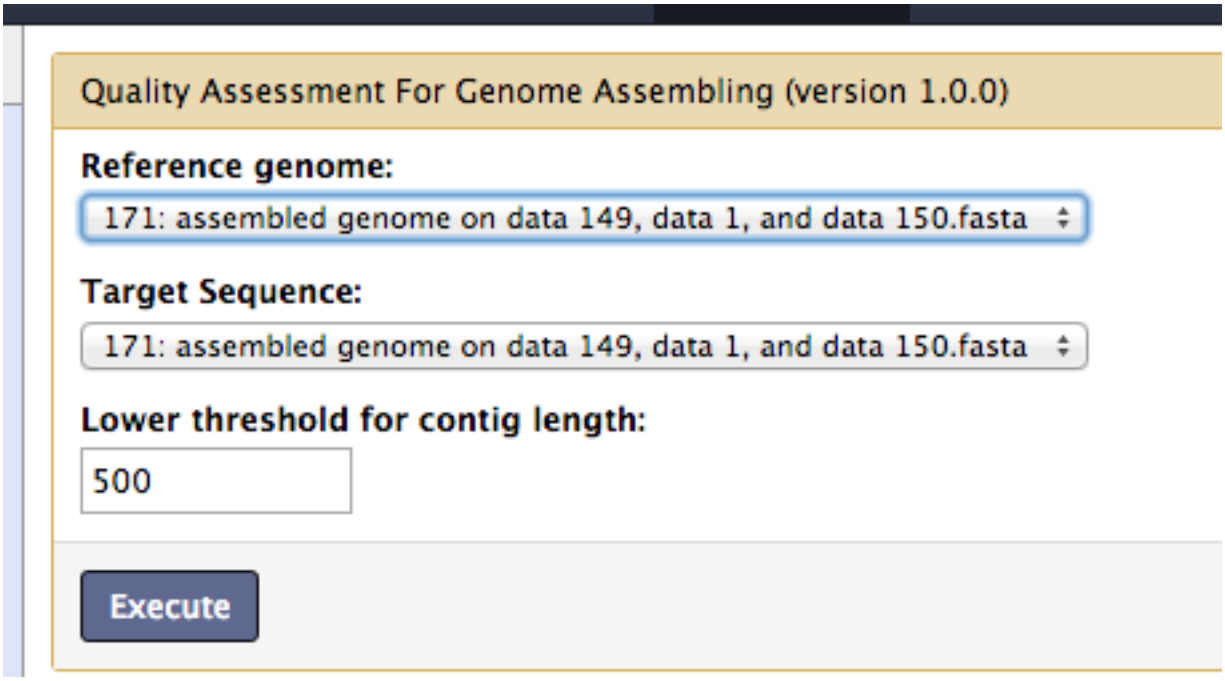
Post-Assembly Analysis

VirAmp not only provides all the processes related to assembly, but also integrates multiple tools for post-assembly processing including quality assessment and variation analysis.

3.1 QUAST REPORT

It is important to evaluate how robust the new assembly is before it is fed into the downstream functional analysis. VirAmp constructs a report of common assembly evaluation metrics based on comparisons with the reference. A detailed QUAST report can be downloaded for further evaluation.

The inputs required are the reference genome and the newly created assembly.



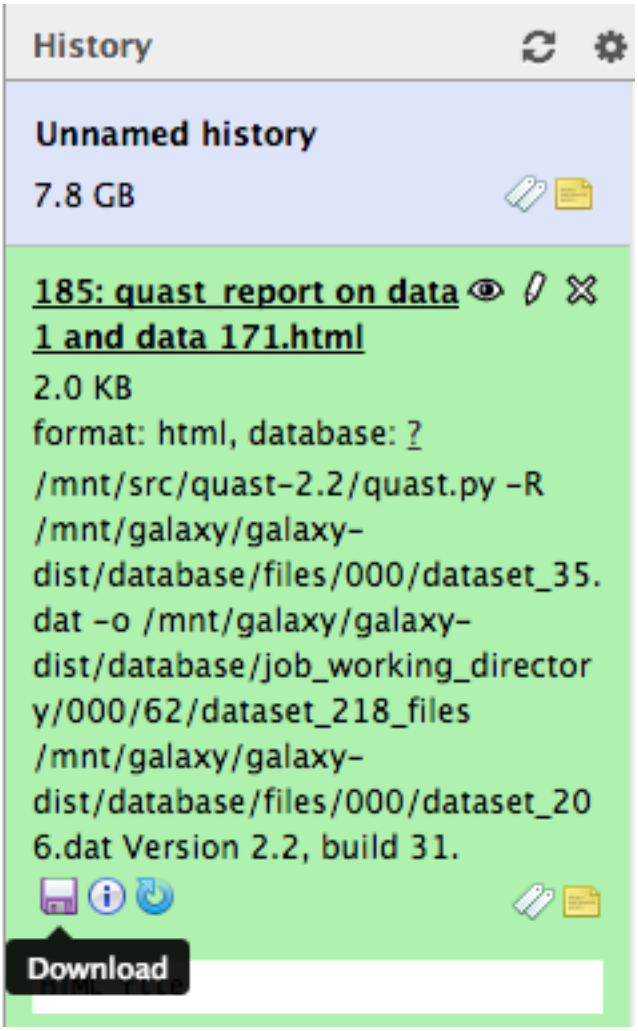
The primary output of QUAST is a summary of common assembly evaluation metrics.

Assembly Statistics

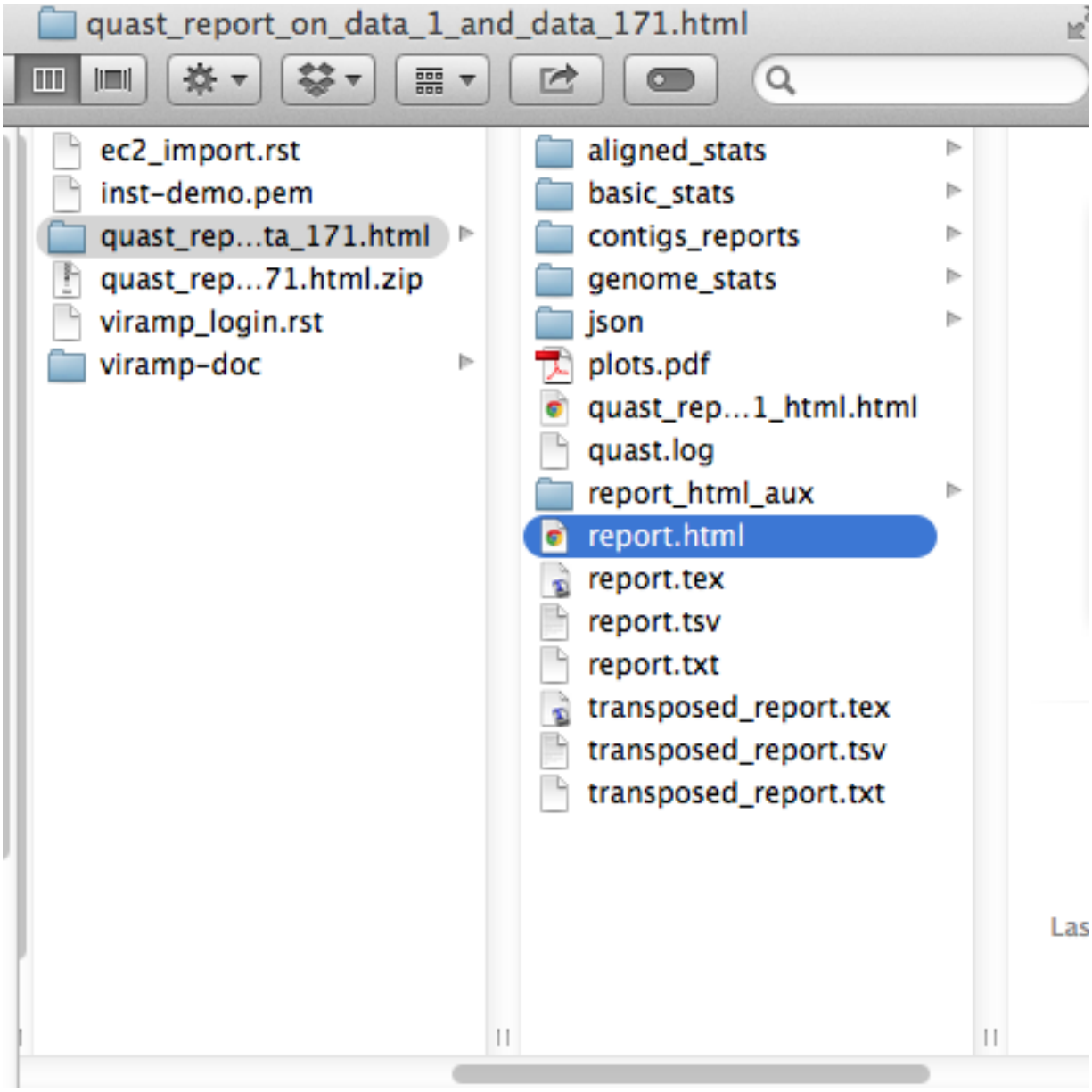
All statistics are based on contigs of size >= 500 bp, unless otherwise noted (e.g., "# contigs (>= 0 bp)" and "Total length (>= 0 bp)" include all contigs).

Assembly	dataset_206.dat
# contigs (>= 0 bp)	7
# contigs (>= 1000 bp)	5
Total length (>= 0 bp)	133874
Total length (>= 1000 bp)	133166
# contigs	6
Total length	133678
Largest contig	71396
Reference length	136376
GC (%)	67.20
Reference GC (%)	67.53
N50	71396
NG50	71396
N75	36562
NG75	36562
L50	1
LG50	1
L75	2
LG75	2
# misassemblies	1
Misassembled contigs length	16097
# local misassemblies	3
# unaligned contigs	1 + 0 part
Unaligned length	512
Genome fraction (%)	97.485
Duplication ratio	1.002
# N's per 100 kbp	120.44
# mismatches per 100 kbp	483.66
# indels per 100 kbp	64.69
Largest alignment	71309
NA50	71309
NGA50	71309
NA75	36560
NGA75	36560
LA50	1
LGA50	1
LA75	2
LGA75	2

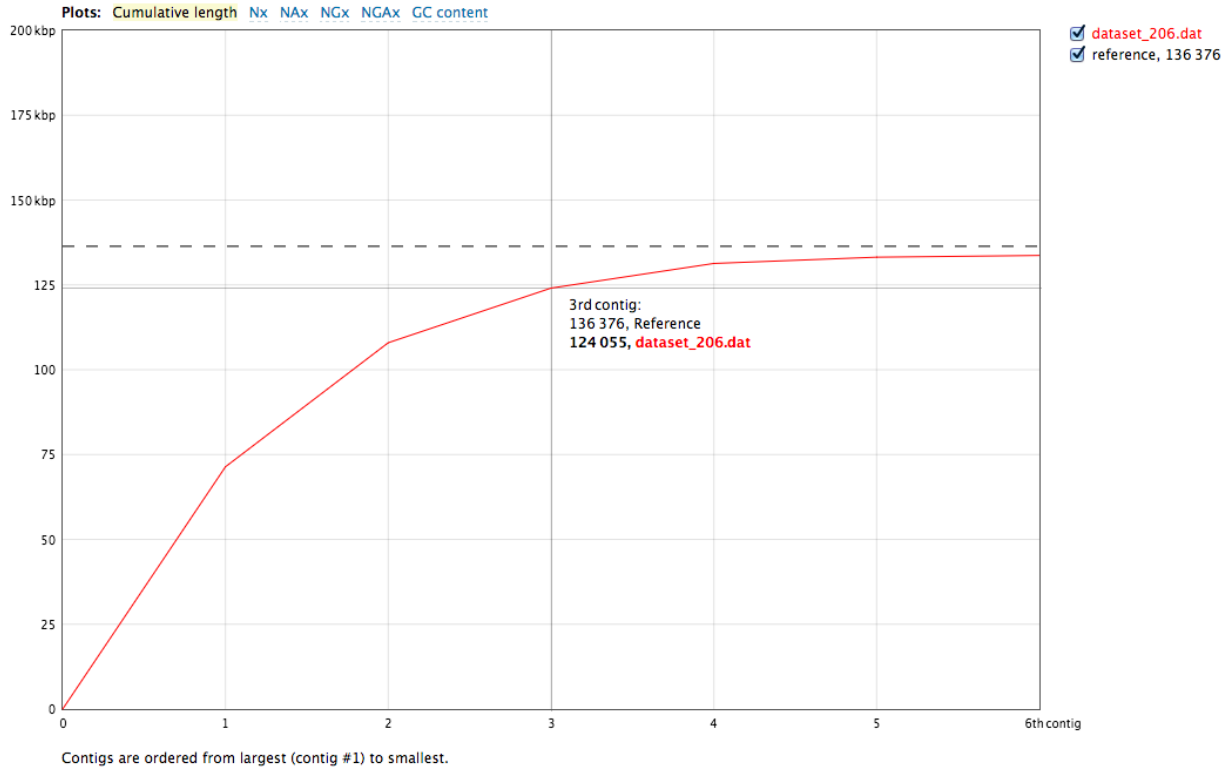
Alternatively, a more detailed QUAST report can also be downloaded.



Unzip and open the report.



A demonstration of a QUAST plot:



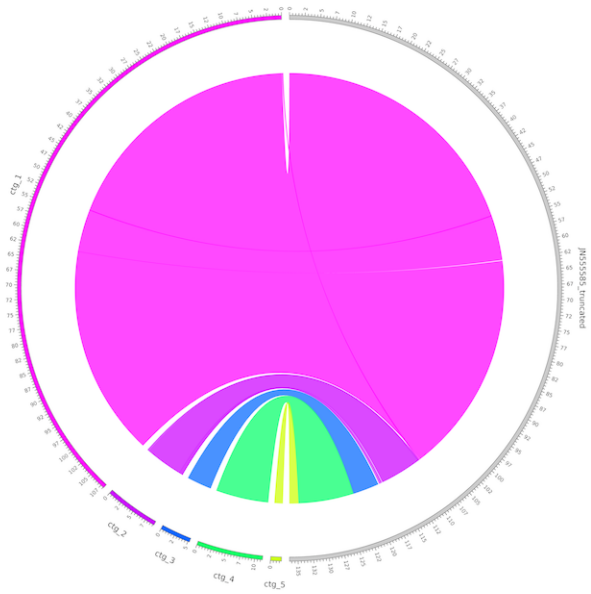
3.2 Assembly-Reference Alignment

VirAmp provides information about the difference between the reference and the new assembly based on a MUMmer alignment. Coordinates and percentage identities are displayed for each aligned region between these two sequences. This is useful in identifying large INDELs as well as other complex structural variations. Table 1 demonstrates an example of the comparison report generated by this tool.

[R_St]	[R_Ed]	[T_St]	[T_Ed]	[% IDY]	[LEN_R]	[LEN_T]	[COV_R]	[COV_T]	[REF_ID]	[CGT_ID]
1	62450	47	62569	99.37	136376	71396	45.79	87.57	JN555585_truncated	scaffold2|size71396
53191	53334	53367	53224	100.00	136376	71396	0.11	0.20	JN555585_truncated	scaffold2|size71396
53191	53273	53451	53369	100.00	136376	71396	0.06	0.12	JN555585_truncated	scaffold2|size71396
62633	71434	62611	71396	99.20	136376	71396	6.45	12.31	JN555585_truncated	scaffold2|size71396
71435	108009	3	36562	99.53	136376	36562	26.82	99.99	JN555585_truncated	scaffold1|size36562
108299	108496	5	193	92.96	136376	196	0.15	96.43	JN555585_truncated	scaffold7|size196
108879	116167	1	7268	96.69	136376	7268	5.34	100.00	JN555585_truncated	scaffold4|size7268
116305	116864	1	483	85.71	136376	512	0.41	94.34	JN555585_truncated	scaffold6|size512
117906	119748	1	1843	99.57	136376	1843	1.35	100.00	JN555585_truncated	scaffold5|size1843
120363	123237	16097	13249	97.30	136376	16097	2.11	17.70	JN555585_truncated	scaffold3|size16097
123245	123394	13307	13158	100.00	136376	16097	0.11	0.93	JN555585_truncated	scaffold3|size16097
123383	134543	1	11179	98.86	136376	16097	8.18	69.45	JN555585_truncated	scaffold3|size16097
134332	134565	11181	11413	99.57	136376	16097	0.17	1.45	JN555585_truncated	scaffold3|size16097
134593	136376	11351	13104	97.42	136376	16097	1.31	10.90	JN555585_truncated	scaffold3|size16097

3.3 Circos graph visualization

Circos projects the assembled draft genome to the aligned part of the reference genome, creating a straightforward visualization for the above alignment and providing insight into large structural variations.



3.4 SNP analysis

Using the alignment between the assembly and the reference, SNP information is displayed in VCF format.

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	
Reference	1700	nucmer	T	C	0	PASS	.	
Reference	1704	nucmer	G	T	0	PASS	.	
Reference	1867	nucmer	G	A	0	PASS	.	
Reference	1909	nucmer	T	G	0	PASS	.	
Reference	2239	nucmer	C	T	0	PASS	.	
Reference	2463	nucmer	C	CC	0	PASS	.	
Reference	2466	nucmer	A	G	0	PASS	.	
Reference	2467	nucmer	T	C	0	PASS	.	
Reference	2558	nucmer	AAG	A	0	PASS	.	
Reference	2568	nucmer	G	T	0	PASS	.	
Reference	2731	nucmer	T	C	0	PASS	.	

3.5 Repeat and Tandem repeat analysis

By aligning the assembly against itself, VirAmp additionally provides repeat and tandem repeat information. The starting coordinates and lengths of the repeats are dervied from this alignment.

Start	Extent	UnitLen	Copies	Ctg#
18534	20	3	6.7	scaffold1|size36562
36515	48	7	6.9	scaffold1|size36562
1	46	7	6.6	scaffold2|size71396
20931	19	4	4.8	scaffold2|size71396
62510	60	12	5.0	scaffold2|size71396
62570	41	1	41.0	scaffold2|size71396
7819	63	21	3.0	scaffold3|size16097
8474	19	1	19.0	scaffold3|size16097
11139	41	15	2.7	scaffold3|size16097

CHAPTER 4

Custom installation of the VirAmp AMI

Access http://aws.amazon.com/, in a Web browser.

Select 'My Account/Console' on the top right if you already have an account; otherwise sign up with a new account.

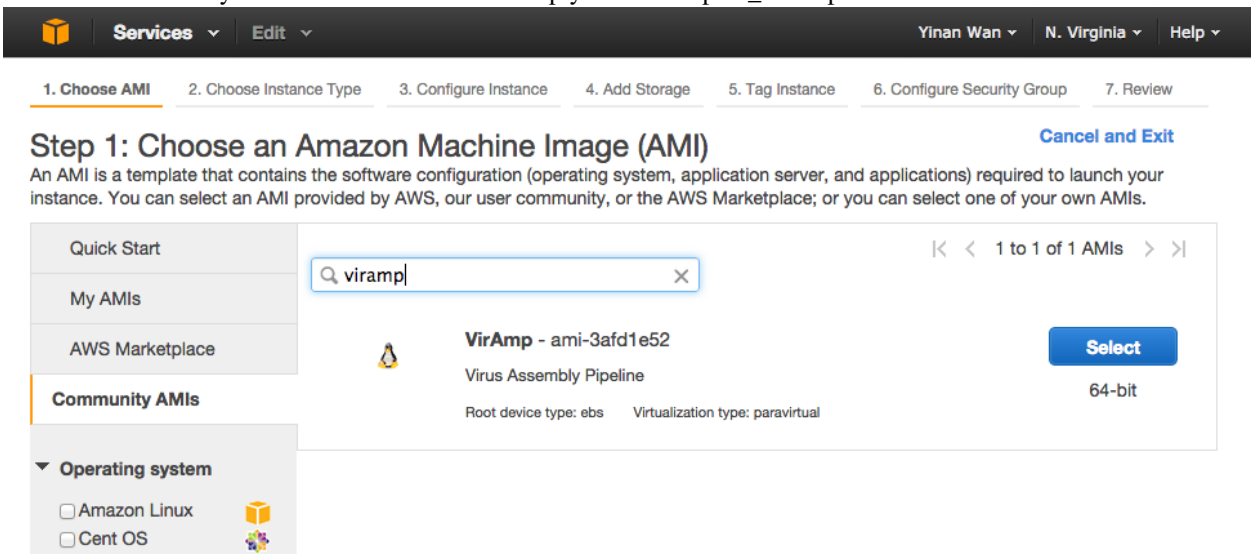
Go to the 'AWS Management Console' option, click the 'EC2' at upper left.

Before importing the AMI, make sure you are in the correct Availability zone. Amazon EC2 is hosted in multiple locations world-wide with multiple Availability zones, and resources cannot be replicated across regions until specified. Our AMI is stored in region "US East(N. Virginia)". Check the upper right corner next to your account name, and make sure it's set at the correct region. If not, just click and select the correct one from the dropdown menu.

Next, click the blue 'Launch Instance' button.

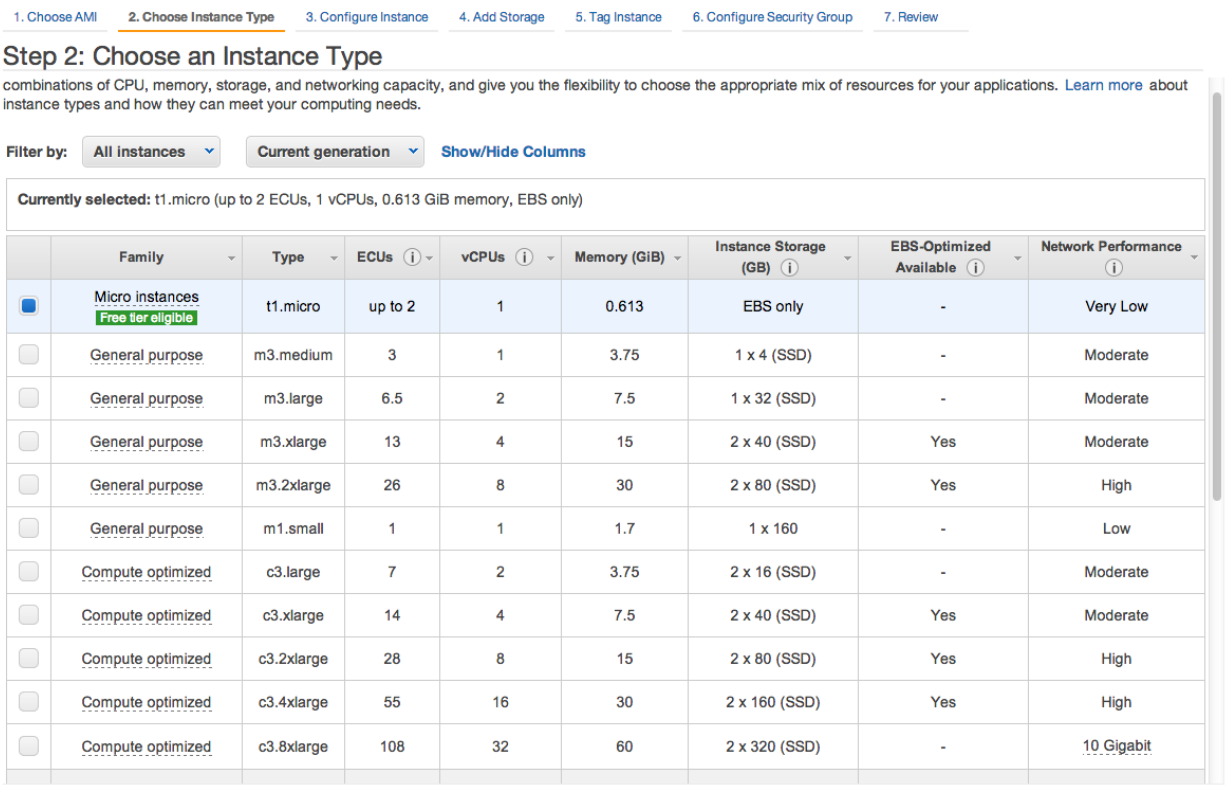
4.1 Step-1: Choosing the instance

Click the Community AMIs tab at mid-left and simply search "Szpara_Viramp"



4.2 Step-2: Review Instance type

Due to storage and computational requirements, free tier instances are not usable with our AMI. For trial runs it is possible to choose smaller instance types, but for serious usage it is advised to select at least the m3.large (third option)



1. Choose AMI | 2. Choose Instance Type | 3. Configure Instance | 4. Add Storage | 5. Tag Instance | 6. Configure Security Group | 7. Review

Step 2: Choose an Instance Type

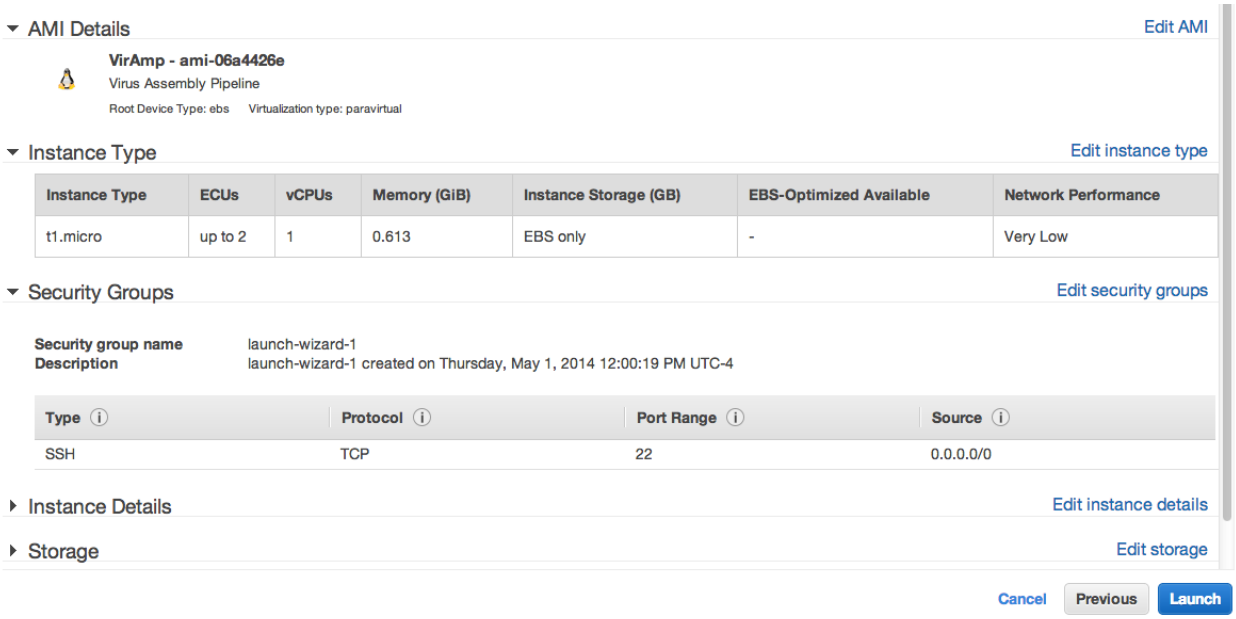
combinations of CPU, memory, storage, and networking capacity, and give you the flexibility to choose the appropriate mix of resources for your applications. Learn more about instance types and how they can meet your computing needs.

Filter by: All instances | Current generation | Show/Hide Columns

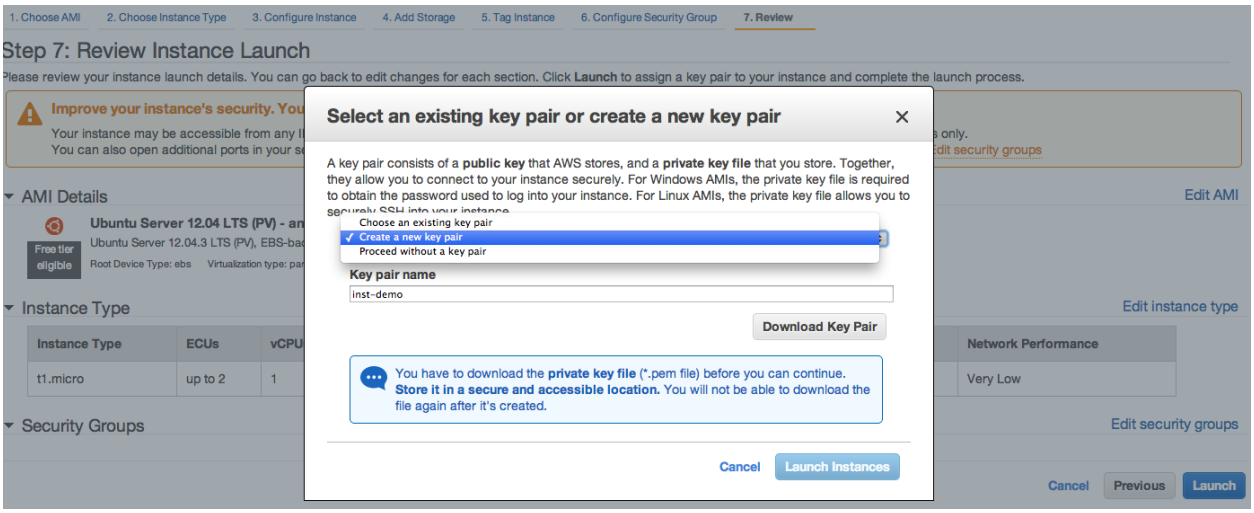
Currently selected: t1.micro (up to 2 ECUs, 1 vCPUs, 0.613 GiB memory, EBS only)

	Family	Type	ECUs	vCPUs	Memory (GiB)	Instance Storage (GB)	EBS-Optimized Available	Network Performance
☑	Micro instances Free tier eligible	t1.micro	up to 2	1	0.613	EBS only	-	Very Low
☐	General purpose	m3.medium	3	1	3.75	1 x 4 (SSD)	-	Moderate
☐	General purpose	m3.large	6.5	2	7.5	1 x 32 (SSD)	-	Moderate
☐	General purpose	m3.xlarge	13	4	15	2 x 40 (SSD)	Yes	Moderate
☐	General purpose	m3.2xlarge	26	8	30	2 x 80 (SSD)	Yes	High
☐	General purpose	m1.small	1	1	1.7	1 x 160	-	Low
☐	Compute optimized	c3.large	7	2	3.75	2 x 16 (SSD)	-	Moderate
☐	Compute optimized	c3.xlarge	14	4	7.5	2 x 40 (SSD)	Yes	Moderate
☐	Compute optimized	c3.2xlarge	28	8	15	2 x 80 (SSD)	Yes	High
☐	Compute optimized	c3.4xlarge	55	16	30	2 x 160 (SSD)	Yes	High
☐	Compute optimized	c3.8xlarge	108	32	60	2 x 320 (SSD)	-	10 Gigabit

4.3 Step-3: Launch the Instance



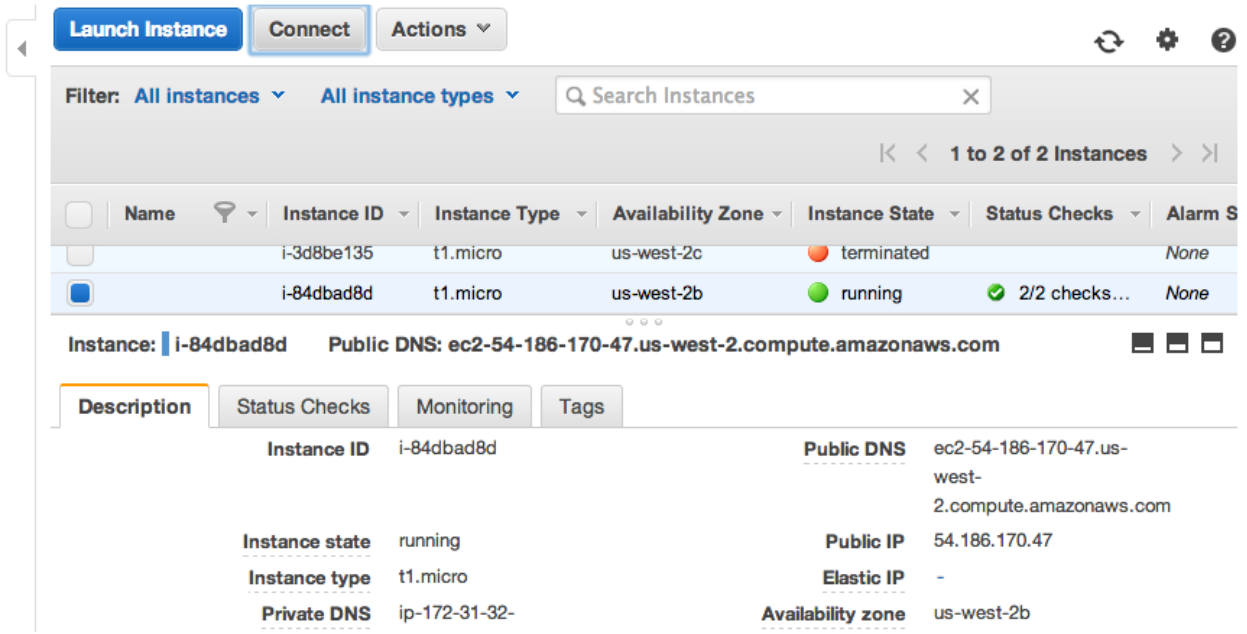
4.4 Step-4: Create Key-pairs



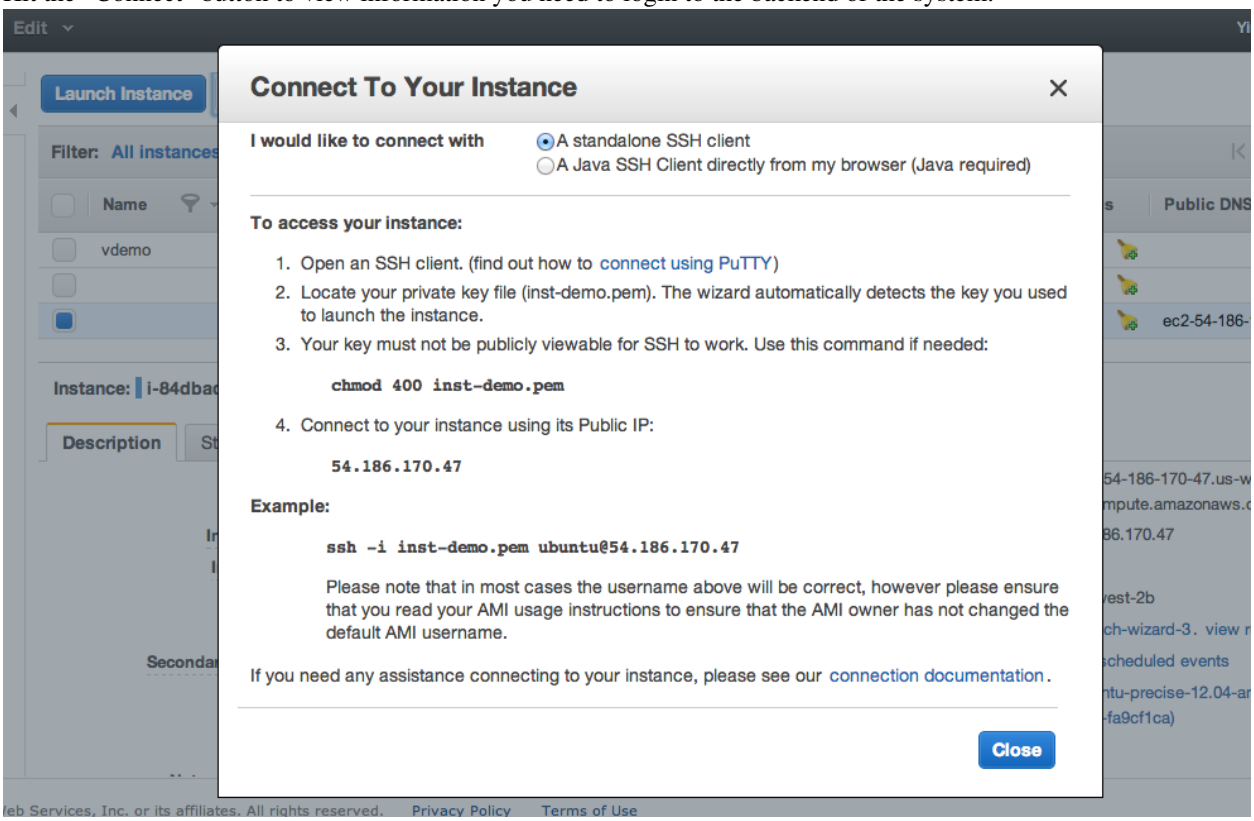
You have now successfully launched your own version of the instance. For information on logging in and starting your instance, please go to VirAmp instance login

4.5 Log in to the new instance

Instructions and an overview of the basic steps and parameters you need to login to the instance are provided at the console.



Hit the "Connect" button to view information you need to login to the backend of the system.



Start your terminal and type the following command:

```
chmod 400 myPemName.pem
```

Connect to your instance using your public IP:

```
ssh -i myPemName.pem ubuntu@public_IP
```

Change to the galaxy directory:

```
cd /mnt/galaxy/galaxy-dist/
```

Change viramp settings:

```
vi universe_wsgi.ini
```

Line 596: admin_users = dwr19@psu.edu should be changed to reflect the current administrators email address Line 662: ftp_upload_site = viramp.com should be changed from viramp.com to your public ip address

Start the viramp server:

```
screen ./run.sh CTRL-a-d
```

- For further information on the individual tools VirAmp utilizes please see the following websites:
  - seqtk
  - diginorm
  - velvet
  - AMOS
  - Quast
  - MUMmer
  - Circos