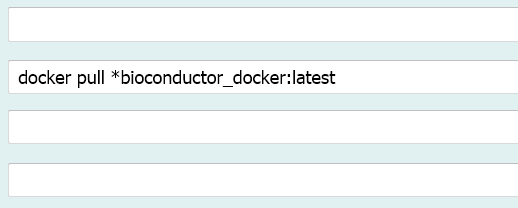
[GitHub Pages] (<https://github.com/william-p-kahley/williamkahley.github.io/William.P.Kahley.CovLet.121420.htm>)

|  |
| --- |
|  |
| |  | | --- | | **Running R on AWS**  **23 JUL 2015**  [**https://aws.amazon.com/blogs/big-data/running-r-on-aws/**](https://aws.amazon.com/blogs/big-data/running-r-on-aws/)  (October 2017 Update)  [**Aaron Friedman**](https://aws.amazon.com/blogs/big-data/author/ajfriedm/) is a Healthcare and Life Sciences Partner Solutions Architect with AWS.  [**https://aws.amazon.com/blogs/big-data/author/ajfriedm/**](https://aws.amazon.com/blogs/big-data/author/ajfriedm/) | | Instructions and best practices for running R on AWS.  Several months ago, I (Markus) wrote a post showing you how to connect R with Amazon EMR,  install RStudio on the Hadoop master node, and use R packages such as rmr2 or plyrmr to analyze a huge public weather dataset. | | Starting a server on AWS—called an EC2 instance—is easy with the [Getting Started instructions.](http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/EC2_GetStarted.html)  The first step is to [launch an Amazon EC2 instance](http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/ec2-launch-instance_linux.html).  We are going to focus on five of the launch steps that impact your R-based analysis environment on AWS:   * Choosing an Amazon Machine Image * Choosing an instance type * Configuring instance details: EC2 user data * Configuring instance details: IAM roles * Configuring a security group   After that, we show you how to load data into your R-based environment, analyze data located on [Amazon S3](http://aws.amazon.com/s3), and configure Shiny Server. We conclude by wrapping up these concepts in an [AWS CloudFormation](https://aws.amazon.com/cloudformation) template to simplify deployment. If you want to skip ahead to the CloudFormation template, [click here](https://aws.amazon.com/blogs/big-data/running-r-on-aws/#launch).  **Choosing an AMI for R** When launching an EC2 instance, you must choose an Amazon Machine Image (AMI), which contains all information required to start an instance. For example, an AMI defines which operating system is installed on your EC2 instance and which software is included.  You can choose the [Amazon Linux AMI](http://aws.amazon.com/amazon-linux-ami/), which is provided at no additional cost and has a stable version of R in the repository. This AMI is maintained by AWS and includes packages and configurations that provide native integration with AWS and other software.  **Choosing an Instance Type for R** Choose an EC2 instance type that matches the data size and processing that your analysis requires. By default, R runs only on one core node and, in many cases, requires a lot of memory.  For programming and development, the general-purpose T2 instance types are sufficient and cheap, and t2.micro is available through the [AWS Free Tier](https://aws.amazon.com/free/). If you don’t know what instance type to choose, start with t2.medium.  The M4 instance family is often a good choice for R workloads. If you use R packages such as foreach, parallel, or snow to parallelize, we recommend using the bigger M4 instance types. They provide a good mix of CPU power and memory.  To connect R to GPU hardware, you can choose the G2/3 or P2 instance families, to leverage packages like gputools. The following table, while a non-exhaustive list, is a good representation of how you might choose your instance types.    An advantage of using AWS is that you aren’t locked into the instance type that you originally choose. You can change your instance type in minutes: just stop your instance, change the instance type, and start the instance again.  RStudio Server lets you share your R-based analysis server with several other scientists. Provision a Linux user for each scientist, and several scientists can work on the same machine. Every user requires at least one CPU and some memory. For multiuser activities, use at least an m4.2xlarge instance type.  Configuring instance details: EC2 user data When you launch an EC2 instance, you can [pass in user data](http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/user-data.html) that can be used to perform common automated configuration tasks. The tasks can even run scripts for installation after the instance starts. In the EC2 launch wizard, you can add this at the **Configure Instance Details** step by expanding the **Advanced Details** pane:    Before running the following script to install R, RStudio Server, the Shiny package, and Shiny Server, visit <https://www.rstudio.com/products/rstudio/download-server>/ to check for the latest versions of RStudio Server. Modify the script to download and install the most recent version. This script also adds a user and password that you use for logging in later to RStudio.   |  | | --- | |  | | #!/bin/bash  #install R  yum install -y R  #install RStudio-Server 1.0.153 (2017-07-20)  wget https://download2.rstudio.org/rstudio-server-rhel-1.0.153-x86\_64.rpm  yum install -y --nogpgcheck rstudio-server-rhel-1.0.153-x86\_64.rpm  rm rstudio-server-rhel-1.0.153-x86\_64.rpm  #install shiny and shiny-server (2017-08-25)  R -e "install.packages('shiny', repos='http://cran.rstudio.com/')"  wget https://download3.rstudio.org/centos5.9/x86\_64/shiny-server-1.5.4.869-rh5-x86\_64.rpm  yum install -y --nogpgcheck shiny-server-1.5.4.869-rh5-x86\_64.rpm  rm shiny-server-1.5.4.869-rh5-x86\_64.rpm  #add user(s)  useradd username  echo username:password | chpasswd |  Configuring instance details: IAM roles On the same configuration page, you can add an AWS Identity and Access Management ([IAM) role](http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/iam-roles-for-amazon-ec2.html) to your EC2 instance.  -  -  - Configuring the security group In the EC2 launch wizard, you [define a security group](http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/using-network-security.htm), which acts as a virtual firewall that controls the traffic for one or more instances. For your R-based analysis environment, you have to open up port 8787 for RStudio Server and port 3838 for Shiny Server.  -  -  - Loading data into your R-based environment on AWS After your EC2 instance is running, you can connect using a web browser to RStudio Server and R. For login credentials, use the newly created user and password. The URL looks like the following:  http://ec2-YOUR-IP.compute-1.amazonaws.com:8787  You can find more details about your public DNS in the EC2 console. To change your Linux user password using RStudio, choose **Tools**, **Shell**, and type the Linux command passwd.  You can do most of your work using RStudio Server, but in some cases you might have to log in to your EC2 instance via SSH. For example, some R packages require installed Linux packages. For the next steps, install the curl-devel Linux package so that you can use the R package “RCurl”. [Connect to your EC2 instance](http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/ec2-connect-to-instance-linux.html) via SSH and execute the following command:  sudo yum install curl-devel Storing data in S3 [Amazon S3](http://aws.amazon.com/s3) is secure, durable, highly scalable object storage. It is easy to use, with a simple web service interface to store and retrieve any amount of data from anywhere on the web. It’s easy to [get started with S3](http://docs.aws.amazon.com/AmazonS3/latest/gsg/GetStartedWithS3.html).  Move your data to S3 for analysis, copy the data via the [AWS CLI](http://aws.amazon.com/cli/) to your EC2 instance, and read the data into R. If you make your S3 object permission “Everyone”, you can read the object directly into R using the RCurl package. You can also enable fine-grained permissions by specifying the appropriate read permissions in the previous IAM policy that you generated. In the following example, we read from the [CGIAR S3 bucket](https://aws.amazon.com/datasets/ccafs-climate-data/), which is publicly accessible as part of our [public data sets program](https://aws.amazon.com/datasets/). Configuring Shiny Server To use Shiny Server, you have to make some small configuration changes. Connect to your EC2 instance and run the following commands (you can also add this to the previous user data script): Automating deployment Now that you have gone through the preceding steps, here’s an [AWS CloudFormation](https://aws.amazon.com/cloudformation) template so that you can quickly and easily deploy this infrastructure in your own environment. |      |  | | --- | | **NGS Graph Generator on AWS Free Tier**  <https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&cad=rja&uact=8&ved=2ahUKEwjdn5_JvqrtAhUP2VkKHWdfAM0QFjABegQIBBAC&url=https%3A%2F%2Fwww.biorxiv.org%2Fhighwire%2Ffilestream%2F123738%2Ffield_highwire_adjunct_files%2F0%2F409573-1.docx&usg=AOvVaw1ctSQUOrm8NEQy_lD8FV1p> | | <https://github.com/systems-immunology-roslin-institute/ngs-graph-generator>  Choose an Instance Type page  This page lists all of the available instances type varying combinations of CPU, memory, storage and capacity.  In order to use this pipeline, it needs a small memory and low storage,  thus a user can select for t2.micro (free tier eligible) (1), then click ‘Review and Launch’ button (2).  BioLayout Express3D or Miru |      |  | | --- | | **Development of a cloud-based Bioinformatics Training Platform**  Briefings in Bioinformatics, Volume 18, Issue 3, May 2017, Pages 537–544  <https://academic.oup.com/bib/article/18/3/537/2453288> | | The 3 day NGS workshop has been organized into seven training modules:  introduction to the command line,  quality control of the NGS data and Alignment,  ChIPSeq, RNASeq, De novo Assembly and Post Workshop.  All the presentations, tools, data sets and tutorials for each training module are accessible and maintained on  GitHub (<https://github.com/BPA-CSIRO-Workshops>).  BTP has been given a sufficient allocation for running  50 instances with specifications of two CPUs and 8 GB of RAM each.  The data sets are stored on the cloud object storage and are organized into buckets (also known as containers), one per training module. The data sets used on the BTP are available on both the NeCTAR Research Cloud (Swift) and Amazon S3 object storage ([Table 1](javascript:;)). The data sets are downloaded on the training virtual machines using Puppet. Puppet reads a simple data set metadata file that describes what data set is required for each training module and its storage location. The data set metadata file is written in YAML (Yet Another Markup Language), a human readable format and easily parsable format. The use of cloud object storage is ideal, as the data sets are mostly static and are easily retrievable using standard Web protocols. An example of a data set metadata file based on the quality control module is described in the [supplementary document](https://oup.silverchair-cdn.com/oup/backfile/Content_public/Journal/bib/18/3/10.1093_bib_bbw032/3/bbw032_supp.zip?Expires=1608434484&Signature=iW6RrOlfjjUq0HHKEJ98jzY3aW4Yu8qqpdIv3hHqdivlrlVgpqCjg8oieyq3famhf3Ao18rXLcSG8hWHUkkxR3nsmPs2MwIYWf5sZV9R~AHFGYHtp9OcqHrLlDh07lyNo4gh~vpyh4P9uBy0OnhiCZJkCk~ZW-hjF8eiMhQVT3zsMVQ~fk1g2nAuMp2HyNWVaOMtrKyPCWttwK4KEZ5D-nM72HML~UPk9ohAnom5z2dCoHU3MKYC4xeKRU8Gu2e7ljBJJ4JjUWDBBTrPTLPl8YvWOqIeYVi2RTcb8uAq2ukUF~s1kUmcAqbJxajACk1JUVoBzOQPfwrsgeDqsx0Qxg__&Key-Pair-Id=APKAIE5G5CRDK6RD3PGA).    Specialist bioinformatics software tools for analysis of data are essential for any bioinformatics hands-on workshop. These software tools are packaged and installed into the BTP image on creation. To improve the software tool installation process onto the BTP virtual machine images, we have used a tool packaging system. All the software tools used during the workshop have been packaged into a Debian-based installer. A list of these tools can be found in [Table 2](javascript:;).  Table 2.  Current list of analysis tools included and maintained on the Bioinformatics Training Platform. These tools are automatically configured and installed on the BTP images and instances   | **Tool** | **Function** | **Link** | | --- | --- | --- | | AMOS Hawkeye | Genome data visualization | <http://sourceforge.net/projects/amos/> | | BEDTools | Genome data manipulation | <http://bedtools.readthedocs.org/en/latest/> | | BLAT | Sequence location lookup in the genome | <https://genome.ucsc.edu/FAQ/FAQblat.html> | | Bowtie | Read Alignment | <http://bowtie-bio.sourceforge.net/index.shtml> | | CummeRbund | RNA-Seq analysis using R | <http://bioconductor.org/packages/release/bioc/html/cummeRbund.html> | | Cufflinks | RNA-Seq analysis | <http://cole-trapnell-lab.github.io/cufflinks/> | | DESeq2 | Differential gene expression analysis using R | <https://bioconductor.org/packages/release/bioc/html/DESeq2.html> | | edgeR | Empirical gene expression analysis using R | <https://bioconductor.org/packages/release/bioc/html/edgeR.html> | | FastQC | FastQC | <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/> | | FASTX | Toolkit for short reads preprocessing | <http://hannonlab.cshl.edu/fastx_toolkit/> | | IGV | Interactive exploration of genomic data | <https://www.broadinstitute.org/igv/> | | igvtools | For preprocessing data before loading to IGV | <https://www.broadinstitute.org/igv/igvtools> | | MACS | ChIP-Seq analysis | <http://liulab.dfci.harvard.edu/MACS/> | | MUMmer | Rapid genome alignment, a dependency for AMOS | <http://mummer.sourceforge.net/> | | PeakAnalyzer | Multi-peak data analysis | <http://www.bioinformatics.org/peakanalyzer/wiki/> | | Picard | Sequence data analysis | <http://broadinstitute.github.io/picard/> | | SAMtools | For manipulating alignments in the SAM format | <http://samtools.sourceforge.net/> | | Skewer | Adapter trimmer for paired-end reads | <https://github.com/relipmoc/skewer> | |  |  | | --- | |  | |  |  |  | | --- | |  | |  | |

|  |
| --- |
|  |
| |  | | --- | | **Exploratory data analysis of genomic datasets using ADAM and Mango with Apache Spark on Amazon EMR**  13 JUL 2018  Amazon EMR, AWS Big Data  <https://aws.amazon.com/blogs/big-data/exploratory-data-analysis-of-genomic-datasets-using-adam-and-mango-with-apache-spark-on-amazon-emr/> | | As the cost of genomic sequencing has rapidly decreased, the amount of publicly available genomic data has soared over the past couple years. New cohorts and studies have produced massive datasets consisting of over 100,000 individuals. Simultaneously, these datasets have been processed to extract genetic variation across populations, producing mass amounts of variation data for each cohort. In this era of big data, tools like Apache Spark have provided a user-friendly platform for batch processing large datasets. However, in order to use such tools as a sufficient replacement to current bioinformatics pipelines, we need more accessible and comprehensive API’s for processing genomic data, as well as support for interactive exploration of these processed datasets.  [**ADAM**](https://github.com/bigdatagenomics/adam)  <https://github.com/bigdatagenomics/adam>  and  [**Mango**](https://github.com/bigdatagenomics/mango)  <https://github.com/bigdatagenomics/mango>  provide a unified environment for processing, filtering, and visualizing large genomic datasets on Apache Spark. ADAM allows users to programmatically load, process, and select raw genomic and variation data using SparkSQL, an SQL interface for aggregating and selecting data in Apache Spark. Mango supports visualization of both raw and aggregated genomic data in a Jupyter notebook environment, allowing users to draw conclusions from large datasets at multiple resolutions. This combined power of ADAM and Mango allows users to load, query and explore datasets in a unified environment, allowing users to interactively explore genomic data at a scale previously impossible using single node bioinformatics tools. **Configuring ADAM and Mango on Amazon EMR** First, we will launch and configure an EMR cluster. Mango uses Docker containers to easily run on Amazon EMR. Upon cluster startup, EMR will use the bootstrap action below to install Docker and the required startup scripts. The scripts will be available at /home/hadoop/mango-scripts  aws emr create-cluster  --release-label emr-5.19.0 \  --name 'emr-5.19.0 Mango example' \  --applications Name=Hadoop Name=Hive Name=Spark \  --ec2-attributes KeyName=<your-ec2-key>,InstanceProfile=EMR\_EC2\_DefaultRole \  --service-role EMR\_DefaultRole \  --instance-groups \ InstanceGroupType=MASTER,InstanceCount=1,InstanceType=c5.4xlarge \ InstanceGroupType=CORE,InstanceCount=4,InstanceType=c5.4xlarge \ --region <your-aws-region> \  --log-uri s3://<your-s3-bucket>/emr-logs/ \  --bootstrap-actions \  Name='Install Mango', Path="s3://aws-bigdata-blog/artifacts/mango-emr/install-bdg-mango-docker-emr5.sh"  To start the Mango notebook, run the following:  /home/hadoop/mango-scripts/run-notebook.sh  This file will set up all of the environment variables needed to run Mango in Docker on EMR. In your terminal, you will see the port and Jupyter notebook token for the Mango Notebook session. Navigate to this port on the public DNS URL of the master node for your EMR cluster. **Loading data from the 1000 Genomes Project** Now that we have a working environment, lets use ADAM and Mango to discover interesting variants in the child from the genome sequencing data of a trio (data from a mother, father, and child). These data are available from the  [1000 Genomes Project AWS Public Dataset](https://aws.amazon.com/1000genomes/).  <https://registry.opendata.aws/1000-genomes/>  **Resources on AWS**  Description  <http://www.internationalgenome.org/formats>  Resource type  S3 Bucket  Amazon Resource Name (ARN)  arn:aws:s3:::1000genomes  AWS Region  us-east-1  AWS CLI Access (No AWS account required)  aws s3  In this analysis, we will view a trio ([NA19685](http://www.internationalgenome.org/data-portal/sample/NA19685), [NA19661](http://www.internationalgenome.org/data-portal/sample/NA19661), and [NA19660](http://www.internationalgenome.org/data-portal/sample/NA19660)) and search for variants that are present in the child but not present in the parents.  In particular, we want to identify genetic variants that are found in the child but not in the parents, known as de novo variants. These are interesting regions, as they may indicate sights of de novo variation that may contribute to multiple disorders.  You can find the Jupyter notebook containing these examples in [Mango’s GitHub repository](https://github.com/bigdatagenomics/mango/blob/master/example-files/notebooks/aws-1000genomes.ipynb), or at /opt/cgl-docker-lib/mango/example-files/notebooks/aws-1000genomes.ipynb in the running Docker container for Mango.  First, import the ADAM and Mango modules and any Spark modules that you need:  # Import ADAM modules  from bdgenomics.adam.adamContext import ADAMContext  from bdgenomics.adam.rdd import AlignmentRecordRDD, CoverageRDD  from bdgenomics.adam.stringency import LENIENT, \_toJava  # Import Mango modules  from bdgenomics.mango.rdd import GenomicVizRDD  from bdgenomics.mango.QC import CoverageDistribution  # Import Spark modules  from pyspark.sql import functions as sf  Next, create a Spark session. You will use this session to run SQL queries on variants.  # Create ADAM Context  ac = ADAMContext(spark)  **Variant analysis with Spark SQL** Load in a subset of variant data from chromosome 17:  genotypesPath = 's3://1000genomes/phase1/analysis\_results/integrated\_call\_sets/ALL.chr17.integrated\_phase1\_v3.20101123.snps\_indels\_svs.genotypes.vcf.gz'  genotypes = ac.loadGenotypes(genotypesPath)  # repartition genotypes to balance the load across memory  genotypes\_df = genotypes.toDF()  You can take a look at the schema by printing the columns in the dataframe.  # cache genotypes and show the schema  genotypes\_df.columns  This genotypes dataset contains all samples from the 1000 Genomes Project. Therefore, you will next filter genotypes to only consider samples that are in the NA19685 trio, and cache the results in memory.  # trio IDs  IDs = ['NA19685', 'NA19661','NA19660']  # Filter by individuals in the trio  trio\_df = genotypes\_df.filter(genotypes\_df["sampleId"].isin(IDs))  trio\_df.cache()  trio\_df.count()  Next, add a new column to your dataframe that determines the genomic location of each variant. This is defined by the chromosome (contigName) and the start and end position of the variant.  # Add ReferenceRegion column and group by referenceRegion  trios\_with\_referenceRegion = trio\_df.withColumn('ReferenceRegion',  sf.concat(sf.col('contigName'),sf.lit(':'), sf.col('start'), sf.lit('-'), sf.col('end')))  Now, you can query your dataset to find de novo variants. But first, you must register your dataframe with Spark SQL.  # Register df with Spark SQL  trios\_with\_referenceRegion.createOrReplaceTempView("trios")  Now that your dataframe is registered, you can run SQL queries on it. For the first query, select the names of variants belonging to sample NA19685 that have at least one alternative (ALT) allele.  # filter by alleles. This is a list of variant names that have an alternate allele for the child  alternate\_variant\_sites = spark.sql("SELECT variant.names[0] AS snp FROM trios \  WHERE array\_contains(alleles, 'ALT') AND sampleId == 'NA19685'")  collected\_sites = list(map(lambda x: x.snp, alternate\_variant\_sites.collect())  For your next query, filter sites in which the parents have both reference alleles. Then filter these variants by the set produced previously from the child.  # get parent records and filter by only REF locations for variant names that were found in the child with an ALT  filtered1 = spark.sql("SELECT \* FROM trios WHERE sampleId == 'NA19661' or sampleId == 'NA19660' \  AND !array\_contains(alleles, 'ALT')")  filtered2 = filtered1.filter(filtered1["variant.names"][0].isin(collected\_sites))  snp\_counts = filtered2.groupBy("variant.names").count().collect()  # collect snp names as a list  snp\_names = map(lambda x: x.names, snp\_counts)  denovo\_snps = [item for sublist in snp\_names for item in sublist]  denovo\_snps[:10]    Now that you have found some interesting variants, you can unpersist your genotypes from memory.  trio\_df.unpersist()  **Working with alignment data** You have found a lot of potential de novo variant sites. Next, you can visually verify some of these sites to see if the raw alignments match up with these de novo hits.  First, load in the alignment data for the NA19685 trio:  # load in NA19685 exome from s3a  childReadsPath = 's3a://1000genomes/phase1/data/NA19685/exome\_alignment/NA19685.mapped.illumina.mosaik.MXL.exome.20110411.bam'  parent1ReadsPath = 's3a://1000genomes/phase1/data/NA19685/exome\_alignment/NA19660.mapped.illumina.mosaik.MXL.exome.20110411.bam'  parent2ReadsPath = 's3a://1000genomes/phase1/data/NA19685/exome\_alignment/NA19661.mapped.illumina.mosaik.MXL.exome.20110411.bam'  childReads = ac.loadAlignments(childReadsPath, stringency=LENIENT)  parent1Reads = ac.loadAlignments(parent1ReadsPath, stringency=LENIENT)  parent2Reads = ac.loadAlignments(parent2ReadsPath, stringency=LENIENT)  Note that this example uses s3a:// instead of s3:// style URLs. The reason for this is that the ADAM formats use Java NIO to access BAM files. To do this, we are using a [JSR 203](https://jcp.org/en/jsr/detail?id=203) implementation for the Hadoop Distributed File System to access these files. This itself requires the s3a:// protocol. You can view that implementation in [this GitHub repository](https://github.com/fnothaft/jsr203-s3a).  You now have data alignment data for three individuals in your trio. However, the data has not yet been loaded into memory. To cache these datasets for fast subsequent access to the data, run the cache() function:  # cache child RDD and count records  # takes about 2 minutes, on 4 c3.4xlarge worker nodes  childReads.cache()  # Count reads in the child  childReads.toDF().count()  # Output should be 95634679  **Quality control of alignment data** One popular analysis to visually re-affirm the quality of genomic alignment data is by viewing coverage distribution. Coverage distribution gives you an idea of the read coverage that you have across a sample.  Next, generate a sample coverage distribution plot for the child alignment data on chromosome 17:  # Calculate read coverage  # Takes 2-3 minutes  childCoverage = childReads.transform(lambda x: x.filter(x.contigName == "17")).toCoverage()  childCoverage.cache()  childCoverage.toDF().count()  # Output should be 51252612  Now that coverage data is calculated and cached, compute the coverage distribution of chromosome 17 and plot the coverage distribution:  # Calculate coverage distribution  # You can check the progress in the SparkUI by navigating to  # :8088 and clicking on the currently running Spark application.  cd = CoverageDistribution(spark, childCoverage, bin\_size = 1)  ax, results = cd.plotDistributions(normalize=True, cumulative=False)  ax.set\_title("Normalized Target Region Coverage")  ax.set\_ylabel("Fraction")  ax.set\_xlabel("Coverage Depth")  ax.set\_xscale("log")  plt.show()      This looks pretty standard because the data you are viewing is exome data. Therefore, you can see a high number of sights with low coverage and a smaller number of genomic positions with more than 100 reads. Now that you are done with coverage, you can unpersist these datasets to clear space in memory for the next analysis.  childCoverage.unpersist()  **Viewing sites with missense variants in the proband** After verifying alignment data and filtering variants, you have four genes with potential missense mutations in the proband, including YBX2, ZNF286B, KSR1, and GNA13. You can visually verify these sites by filtering and viewing the raw reads of the child and parents.  First, view the child reads. If you zoom in to the location of the GNA13 variant (63052580-63052581), you can see a heterozygous T to A call:  # missense variant at GNA13: 63052580-63052581 (SNP rs201316886)  # define alignment summary for child reads  childViz = AlignmentSummary(spark, ac, childReads)  # Takes about 2 minutes to collect data from workers  contig = "17"  start = 63052180  end = 63052981  childViz.viewPileup(contig, start, end)    It looks like there indeed is a variant at this position, possibly a heterozygous SNP with alternate allele A. Look at the parent data to verify that this variant does not appear in the parents:  # define alignment summary for parent reads  parent1Viz = AlignmentSummary(spark, ac, parent1Reads)  # view missense variant at GNA13: 63052580-63052581 in parent 1  contig = "17"  start = 63052180  end = 63052981  parent1Viz.viewPileup(contig, start, end)    This confirms the filter that this variant is indeed present only in the proband, but not the parents. **Summary** To summarize, this post demonstrated how to set up and run ADAM and Mango in Amazon EMR. We demonstrated how to use these tools in an interactive notebook environment to explore the 1000 Genomes dataset, a publicly available dataset on Amazon S3. We used these tools inspect 1000 Genomes data quality, query for interesting variants in the genome, and validate results through the visualization of raw data.  For more information about Mango, see the [Mango User Guide](http://bdg-mango.readthedocs.io/en/latest/). If you have questions or suggestions, please comment below. |  |  | | --- | | **Orchestrating analytics jobs by running Amazon EMR Notebooks programmatically**  23 NOV 2020 | Amazon EMR, AWS Big Data  <https://aws.amazon.com/blogs/big-data/orchestrating-analytics-jobs-by-running-amazon-emr-notebooks-programmatically/> | | [Amazon EMR](http://aws.amazon.com/emr) is a big data service offered by AWS to run Apache Spark and other open-source applications on AWS in a cost-effective manner. [Amazon EMR Notebooks](https://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-managed-notebooks.html) is a managed environment based on [Jupyter Notebook](https://en.wikipedia.org/wiki/Project_Jupyter#Jupyter_Notebook) that allows data scientists, analysts, and developers to prepare and visualize data, collaborate with peers, build applications, and perform interactive analysis using EMR clusters.  EMR notebook APIs are available on Amazon EMR release version 5.18.0 or later and can be used to run EMR notebooks via a script or command line. The ability to start, stop, list, and describe EMR notebook runs without the Amazon EMR console enables you to programmatically control running an EMR notebook. Using a parameterized notebook cell allows you to pass different parameter values to a notebook without having to create a copy of the notebook for each new set of parameter values. With this feature, you can schedule running EMR notebooks with cron scripts, chain multiple EMR notebooks, and use orchestration services such as [AWS Step Functions](https://aws.amazon.com/step-functions/) or Apache Airflow to build pipelines. If you want to use EMR notebooks in a non-interactive manner, this enables you to run ETL workloads, especially in production.  In this post, we show how to orchestrate analytics jobs by running EMR Notebooks programmatically with the following two use cases:   * Scheduling an EMR notebook run via crontab and the [AWS Command Line Interface](http://aws.amazon.com/cli) (AWS CLI) * Chaining your notebooks with Step Functions triggered by [Amazon CloudWatch Events](https://docs.aws.amazon.com/AmazonCloudWatch/latest/events/WhatIsCloudWatchEvents.html)   For our data source, we use the open-source, real-time COVID-19 US daily case reports provided by Johns Hopkins University CSSE in the following [GitHub repo](https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data/csse_covid_19_daily_reports_us). Prerequisites Before getting started, you must have the following prerequisites:   * An AWS account that provides access to the following AWS services at least:   + [AWS CloudFormation](http://aws.amazon.com/cloudformation)   + Amazon CloudWatch   + [Amazon Elastic Compute Cloud](http://aws.amazon.com/ec2) (Amazon EC2)   + Amazon EMR   + [Amazon EventBridge](https://aws.amazon.com/eventbridge/)   + [AWS Identity and Access Management](http://aws.amazon.com/iam) (IAM)   + [AWS Lambda](http://aws.amazon.com/lambda)   + [Amazon Simple Storage Service](http://aws.amazon.com/s3) (Amazon S3)   + AWS Step Functions * AWS CLI Version 1.18.128 or later installed on your work station. * [Jupyter](https://jupyter.org/install) installed on your work station (this is used for the output visualization part for this post only). * An EMR cluster running Amazon EMR release 5.18.0 or later, with Hadoop, Spark, and Livy installed. Record the value of the cluster ID (for example, <j-\*\*\*\*\*\*\*\*\*\*\*\*\*>); you use this for the examples later. * An EMR notebook created on the Amazon EMR console, using the following two input notebook files:   + [demo\_pyspark.pynb](https://aws-bigdata-blog.s3.amazonaws.com/artifacts/aws-blog-runnable-notebook/demo/notebook/demo_pyspark.ipynb) – Used for both use cases in this post.   + [trailing\_N day.ipynb](https://aws-bigdata-blog.s3.amazonaws.com/artifacts/aws-blog-runnable-notebook/demo/notebook/trailing_N_day.ipynb) – Used for the second use case.   Record the notebook ID (for example, <e-\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*>); you use this later for our examples later. Organize the notebook files in the Jupyter UI as follows:   * /demo\_pyspark.ipynb * /experiment/trailing\_N\_day.ipynb   See [Creating a Notebook](https://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-managed-notebooks-create.html) for more information on how to create an EMR notebook. |  |  | | --- | | **Interactive Analysis of Genomic Datasets Using Amazon Athena**  by Aaron Friedman, PhD | on 07 DEC 2016 | in Amazon Athena, AWS Big Data  [**https://aws.amazon.com/blogs/big-data/interactive-analysis-of-genomic-datasets-using-amazon-athena/**](https://aws.amazon.com/blogs/big-data/interactive-analysis-of-genomic-datasets-using-amazon-athena/) | | Aaron Friedman is a Healthcare and Life Sciences Solutions Architect with Amazon Web Services  The genomics industry is in the midst of a data explosion. Due to the rapid drop in the cost to sequence genomes, genomics is now central to many medical advances. When your genome is sequenced and analyzed, raw sequencing files are processed in a multi-step workflow to identify where your genome differs from a standard reference. Your variations are stored in a Variant Call Format (VCF) file, which is then combined with other individuals to enable population-scale analyses. Many of these datasets are publicly available, and an increasing number are hosted on AWS as part of our [Open Data](https://aws.amazon.com/public-data-sets/) project.  To mine genomic data for new discoveries, researchers in both industry and academia build complex models to analyze populations at scale. When building models, they first explore the datasets-of-interest to understand what questions the data might answer. In this step, interactivity is key, as it allows them to move easily from one question to the next.  Recently, we launched [Amazon Athena](https://aws.amazon.com/athena/) as an interactive query service to analyze data on [Amazon S3](https://aws.amazon.com/s3/). With Amazon Athena there are no clusters to manage and tune, no infrastructure to setup or manage, and customers pay only for the queries they run. Athena is able to query many file types straight from S3. This flexibility gives you the ability to interact easily with your datasets, whether they are in a raw text format (CSV/JSON) or specialized formats (e.g. Parquet). By being able to flexibly query different types of data sources, researchers can more rapidly progress through the data exploration phase for discovery. Additionally, researchers don’t have to know nuances of managing and running a big data system. This makes Athena an excellent complement to data warehousing on [Amazon Redshift](https://aws.amazon.com/redshift/) and big data analytics on [Amazon EMR](https://aws.amazon.com/emr/).   In this post, I discuss how to prepare genomic data for analysis with Amazon Athena as well as demonstrating how Athena is well-adapted to address common genomics query paradigms.  I use the [Thousand Genomes dataset](https://aws.amazon.com/1000genomes/) hosted on Amazon S3, a seminal genomics study, to demonstrate these approaches. All code that is used as part of this post is available in our GitHub [repository](https://github.com/awslabs/aws-big-data-blog/tree/master/aws-blog-athena-genomics/).  Although this post is focused on genomic analysis, similar approaches can be applied to any discipline where large-scale, interactive analysis is required.   Select, aggregate, annotate query pattern in genomics Genomics researchers may ask different questions of their dataset, such as:   * What variations in a genome may increase the risk of developing disease? * What positions in the genome have abnormal levels of variation, suggesting issues in quality of sequencing or errors in the genomic reference? * What variations in a genome influence how an individual may respond to a specific drug treatment? * Does a group of individuals contain a higher frequency of a genomic variant known to alter response to a drug relative to the general population?   All these questions, and more, can be generalized under a common query pattern I like to call “Select, Aggregate, Annotate”. Some of our genomics customers, such as Human Longevity, Inc., routinely use this query pattern [in their work](https://www.youtube.com/watch?v=CGbWEkszAlQ).  In each of the above queries, you execute the following steps:  **SELECT:** Specify the cohort of individuals meeting certain criteria (disease, drug response, age, BMI, entire population, etc.).  **AGGREGATE:** Generate summary statistics of genomic variants across the cohort that you selected.  **ANNOTATE:** Assign meaning to each of the variants by joining on known information about each variant. Dataset preparation Properly organizing your dataset is one of the most critical decisions for enabling fast, interactive analyses. Based on the query pattern I just described, the table representing your population needs to have the following information:   * A unique sample ID corresponding to each sample in your population * Information about each variant, specifically its location in the genome as well as the specific deviation from the reference * Information about how many times in a sample a variant occurs (0, 1, or 2 times) as well as if there are multiple variants in the same site. This is known as a genotype.   The extract, transform, load (ETL) process to generate the appropriate data representation has two main steps. First, you use ADAM, a genomics analysis platform built on top of Spark, to convert the variant information residing a VCF file to Parquet for easier downstream analytics, in a process similar to the one described in the [Will Spark Power the Data behind Precision Medicine?](https://blogs.aws.amazon.com/bigdata/post/Tx1GE3J0NATVJ39/Will-Spark-Power-the-Data-behind-Precision-Medicine) post. Then, you use custom Python code to massage the data and select only the appropriate fields that you need for analysis with Athena.  First, [spin up an EMR cluster](http://docs.aws.amazon.com/ElasticMapReduce/latest/ManagementGuide/emr-gs.html) (version 5.0.3) for the ETL process. I used a c4.xlarge for my master node and m4.4xlarges with 1 TB of scratch for my core nodes.  After you SSH into your master node, clone the [git repository](https://github.com/awslabs/aws-big-data-blog.git). You can also put this in as a bootstrap action when spinning up your cluster. |  |  | | --- | | **Building High-Throughput Genomics Batch Workflows on AWS: Introduction (Part 1 of 4)**  <https://aws.amazon.com/blogs/compute/building-high-throughput-genomics-batch-workflows-on-aws-introduction-part-1-of-4/>  by Andy Katz | on 30 MAY 2017 | | | **Genomics Research on AWS**  <https://github.com/aws-samples/aws-batch-genomics/tree/v1.0.0>  A tutorial on how to package and deploy a bioinformatics workflow on AWS using AWS Batch  This tutorial will cover the material presented within the "Genomics Workflows on AWS" blog post series  (Part 1, Part 2, Part 3, Part 4) that covers the basics of bootstrapping a bioinformatics analysis pipeline on AWS.  We break down the tutorial roughly as follows:  Setting up your AWS account (if you do not already have one)  Package a set of bioinformatics applications using Docker  Create a AWS Batch environment for analysis  Define and deploy AWS Step Functions to control the data processing steps  Initiate a workflow |  |  | | --- | | **Optimizing for cost, availability and throughput by selecting your AWS Batch allocation strategy**  by Bala Thekkedath | on 24 OCT 2019 | in Advanced (300), Amazon EC2, AWS Batch  <https://aws.amazon.com/blogs/compute/optimizing-for-cost-availability-and-throughput-by-selecting-your-aws-batch-allocation-strategy/> | | AWS offers a broad range of instances that are advantageous for batch workloads. The scale and provisioning speed of AWS’ compute instances allow you to get up and running at peak capacity in minutes without paying for downtime. Today, I’m pleased to introduce allocation strategies: a significant new capability in [AWS Batch](https://docs.aws.amazon.com/batch/latest/userguide/allocation-strategies.html) that  makes provisioning compute resources flexible and simple. In this blog post, I explain how the AWS Batch allocation strategies work, when you should use them for your workload, and provide an example CloudFormation script. This blog helps you get started on building your personalized Compute Environment (CE) most appropriate to your workloads.  **Overview**  AWS Batch is a fully managed, cloud-native batch scheduler. It manages the queuing and scheduling of your batch jobs, and the resources required to run your jobs. One of AWS Batch’s great strengths is the ability to manage instance provisioning as your workload requirements and budget needs change. AWS Batch takes advantage of AWS’s broad base of compute types. For example, you can launch compute based instances and memory instances that can handle different workload types, without having to worry about building a cluster to meet peak demand.  Previously, AWS Batch had a cost-controlling approach to manage compute instances for your workloads. The service chose an instance that was the best fit for your jobs based on vCPU, memory, and GPU requirements, at the lowest cost. Now, the newly added allocation strategies provide flexibility. They allow AWS Batch to consider capacity and throughput in addition to cost when provisioning your instances. This allows you to leverage different priorities when launching instances depending on your workloads’ needs, such as: controlling cost, maximizing throughput, or minimizing [Amazon EC2 Spot](https://aws.amazon.com/ec2/spot/) instances interruption rates.  There are now three instance allocation strategies from which to choose when creating an AWS Batch [Compute Environment](https://docs.aws.amazon.com/batch/latest/userguide/compute_environments.html) (CE). They are:  1.        Spot Capacity Optimized  2.        Best Fit Progressive  3.        Best Fit |  |  | | --- | | **BioContainers**  [**https://biocontainers-edu.readthedocs.io/en/latest/**](https://biocontainers-edu.readthedocs.io/en/latest/)  **Integration with BioConda**  [**https://biocontainers-edu.readthedocs.io/en/latest/conda\_integration.html**](https://biocontainers-edu.readthedocs.io/en/latest/conda_integration.html)  [**https://anaconda.org/bioconda/**](https://anaconda.org/bioconda/)  **bioconda / packages  View all (8203)**  [**https://anaconda.org/bioconda/repo**](https://anaconda.org/bioconda/repo) | | <https://bioconda.github.io/>  **Bioconda** is a channel for the [conda](https://conda.io/en/latest/index.html) package manager specializing in bioinformatics software. Bioconda consists of:   * a [repository of recipes](https://github.com/bioconda/bioconda-recipes) hosted on GitHub * a [build system](https://github.com/bioconda/bioconda-utils) turning these recipes into conda packages * a [repository of packages](https://anaconda.org/bioconda/) containing over 7000 bioinformatics packages ready to use with conda install * over 850 contributors and 570 members who add, modify, update and maintain the recipes   The conda package manager makes installing software a vastly more streamlined process. Conda is a combination of other package managers you may have encountered, such as pip, CPAN, CRAN, Bioconductor, apt-get, and homebrew. Conda is both language- and OS-agnostic, and can be used to install C/C++, Fortran, Go, R, Python, Java etc programs on Linux, Mac OSX, and Windows.  Conda allows separation of packages into repositories, or channels. The main defaults channel has a large number of common packages. Users can add additional channels from which to install software packages not available in the defaults channel. Bioconda is one such channel specializing in bioinformatics software.  **Browse packages in the Bioconda channel:** [Package Index](https://bioconda.github.io/conda-package_index.html)  Each package added to Bioconda also has a corresponding Docker [BioContainer](https://biocontainers.pro) automatically created and uploaded to [Quay.io](https://quay.io/organization/biocontainers). A list of these and other containers can be found at the [Biocontainers Registry](https://biocontainers.pro/#/registry). | | **bioconda / packages / samtools 1.11**  <https://anaconda.org/bioconda/samtools>  Tools for dealing with SAM, BAM and CRAM files  **bioconda / packages / igv 2.8.13**  <https://anaconda.org/bioconda/igv>  Integrative Genomics Viewer. Fast, efficient, scalable visualization tool for genomics data and annotations.  **bioconda / packages / bioconductor-deseq2 1.30.0**  <https://anaconda.org/bioconda/bioconductor-deseq2>  Differential gene expression analysis based on the negative binomial distribution  **bioconda / packages / bioconductor-pcaexplorer 2.16.0**  <https://anaconda.org/bioconda/bioconductor-pcaexplorer>  Interactive Visualization of RNA-seq Data Using a Principal Components Approach |   **Bioconductor is available as**  **1: AMI (Amazon Machine Image) images**  **Bioconductor in the cloud**  <https://www.bioconductor.org/help/bioconductor-cloud-ami/>    **2: Docker images**  **Docker containers for Bioconductor**  <https://www.bioconductor.org/help/docker/>  **DockerHub**  Bioconductor / bioconductor\_docker  <https://github.com/Bioconductor/bioconductor_docker>  **AWS Pricing Calculator**  <https://calculator.s3.amazonaws.com/index.html>   |  | | --- | | **DockerHub Bioconductor**  <https://hub.docker.com/u/bioconductor/>  37 repositories 2020/12 | |  |   **Working with Bioconductor's Docker Containers**  <https://www.youtube.com/watch?v=-Jr8k90JQFI>  Jun 18, 2016   |  | | --- | | **Bioconductor Docker Images**  bioconductor/bioconductor\_docker  By bioconductor  Dec 07 2020  <https://hub.docker.com/r/bioconductor/bioconductor_docker>  **bioconductor/bioconductor\_docker**  <https://hub.docker.com/r/bioconductor/bioconductor_docker> <<<<<<<  By bioconductor  Updated 12/04/2020  Bioconductor Docker Images | | **DockerHub** Excellent instructions Bioconductor / bioconductor\_docker  <https://github.com/Bioconductor/bioconductor_docker>  **Using the containers**  <https://github.com/Bioconductor/bioconductor_docker#using-the-containers>  A well organized guide to popular docker commands can be found here. For convenience, below are some commands to get you started.  The following examples use the **bioconductor/bioconductor\_docker:devel image**  **Running the container**  <https://github.com/Bioconductor/bioconductor_docker#running-the-container>  The above commands can be helpful but the real basics of running a Bioconductor Docker involves  pulling the public image and running the container.  **Mounting Additional Volume**  <https://github.com/Bioconductor/bioconductor_docker#mounting-additional-volume>  One such option for docker run is -v to mount an additional volume to the docker image.  This might be useful for say mounting a local R install directory for use on the docker.  The path on the docker image that should be mapped to a local R library directory is **/usr/local/lib/R/host-site-library**  **Modifying the images**  <https://github.com/Bioconductor/bioconductor_docker#modifying-the-images>  There are two ways to modify these images:  1: Making changes in a running container and then committing them using the docker commit command.  docker commit  2: Using a Dockerfile to declare the changes you want to make (recommended way).  The second way is the recommended way. | | **Docker containers for Bioconductor**  <https://www.bioconductor.org/help/docker/#current>  Current Containers  For each supported version of Bioconductor, we provide  bioconductor/bioconductor\_docker:RELEASE\_X\_Y  bioconductor/bioconductor\_docker:devel | | **DockerHub Bioconductor**  <https://hub.docker.com/u/bioconductor/>  37 repositories 2020/12 | | **Rocker Project**  Docker Containers for the R Environment  <https://www.rocker-project.org/>  Bioconductor’s Docker images are stored in Docker Hub;  the source Dockerfile(s) are in Github.  Ensure you have Docker installed and start R inside a container with:  **docker run --rm -ti rocker/r-base**  Or get started with an RStudio® instance:  **docker run -e PASSWORD=yourpassword --rm -p 8787:8787 rocker/rstudio**  **Browser URL localhost:8787**  Log in with user/password rstudio/yourpassword  (Please set your own password; it cannot be rstudio).    **GitHub Rocker**  rocker-org / rocker  <https://github.com/rocker-org/rocker/tree/master/rstudio> |   **Google Kubernetes Engine**  <https://cloud.google.com/kubernetes-engine/>  Secured and managed Kubernetes service with four-way auto scaling and multi-cluster support.  New customers get $300 in free credits to spend on Google Cloud during the first 90 days.  All customers get one zonal cluster per month for free, not charged against your credits. |



[**https://www.google.com/search?as\_q=&as\_epq=docker+pull+\*bioconductor\_docker%3Alatest&as\_oq=&as\_eq=&as\_nlo=&as\_nhi=&lr=&cr=&as\_qdr=all&as\_sitesearch=&as\_occt=any&safe=images&as\_filetype=&tbs=**](https://www.google.com/search?as_q=&as_epq=docker+pull+*bioconductor_docker%3Alatest&as_oq=&as_eq=&as_nlo=&as_nhi=&lr=&cr=&as_qdr=all&as_sitesearch=&as_occt=any&safe=images&as_filetype=&tbs=)

**Docker containers for Bioconductor**

<https://www.bioconductor.org/help/docker/>

Containers can run on any operating system including Windows and Mac (using modern Linux kernels) via the Docker engine.

Current Containers

<https://www.bioconductor.org/help/docker/#current>

For each supported version of Bioconductor, we provide

bioconductor/bioconductor\_docker:RELEASE\_X\_Y

bioconductor/bioconductor\_docker:devel

Bioconductor’s Docker images are stored in Docker Hub; the source Dockerfile(s) are in Github.

|  |
| --- |
| **GitHub Projects** |
| <https://github.com/hbctraining/Intro-to-R-with-DGE>  <https://github.com/hemberg-lab/scRNA.seq.course> |

<https://scrnaseq-course.cog.sanger.ac.uk/website/index.html>

**github.com jdidion biotools**

Excellent Outline of Omics Tools

<https://github.com/jdidion/biotools/blob/main/README.md#data-sets>

NGS RNA-Seq workflow

(quality control of reads,

alignment to reference genomes, <<== Alignment Free?

assembly, quantification, differential expression, visualization).

Study of standard workflow (bioinformatics tools R Bioconductor ).

Differential expression analysis (Pre-processing read count data, statistical principles and machine learning algorithms)

Computational Genomics, Statistical Genomics, Bioinformatics

[**RNA-seq Pre-analysis Tools**](https://bioinformaticshome.com/tools/rna-seq/pre-analysis.html)

**RNA-seq Core-analysis Tools**

**[5.1]** **-** Read open access full-text publications relevant to; use of tools  
**Knowledge of bioinformatics Tools, Workflows**

Standard Workflows

Best Practices

Standard Pipeline

**High-throughput Sequencing Analysis: StatQuest (Josh Starmer)**

**edgeR, part1: Library Normalization**

**DESeq2, part1: Library Normalization**

**edgeR and DESeq2, part2: Independent Filtering (removing genes with low read counts)**

NGS Genomics Workflows:

Whole Genome Sequencing Data Analysis Tools

Whole Exome Sequencing Data Analysis Tools

scRNA-Seq

RNA-Seq

**J Didion**

Mine this for URLs

<https://github.com/jdidion/biotools/blob/main/README.md>

Data Sets

<https://github.com/jdidion/biotools/blob/main/README.md#data-sets>

See this for Python related links – may want to add some of these links to Python table

<https://github.com/jdidion/biotools/blob/main/README.md#python-2>

|  |
| --- |
| **Knowledge of Statistics, Analytics, Machine Learning (Python)** |
| |  | | --- | | **Data Science (Python): Project Repositories** | | **GitHub Jupyter Notebook Topics**  <https://github.com/topics/jupyter-notebook>  **NoteBooks-Statistics-and-MachineLearning**  <https://github.com/leonvanbokhorst/NoteBooks-Statistics-and-MachineLearning/>  **Python-for-Probability-Statistics-and-Machine-Learning**  <https://github.com/unpingco/Python-for-Probability-Statistics-and-Machine-Learning>  **Data-Analysis-Science**  <https://github.com/Olow304/Data-Analysis-Science>  **Kaggle (public notebooks, public datasets); Python**  <https://www.kaggle.com/notebooks>  <https://www.kaggle.com/datasets>  **Python Data Science Handbook/notebooks/**  <https://github.com/jakevdp/PythonDataScienceHandbook/tree/master/notebooks>  **A-gallery-of-interesting-Jupyter-Notebooks**  <https://github.com/jupyter/jupyter/wiki/A-gallery-of-interesting-Jupyter-Notebooks#statistics-machine-learning-and-data-science> |  |  | | --- | | **Data Science (Python): Specific Expertise Tutorials** | | **Interesting Jupyter Notebooks (Statistics, Machine Learning, and Data Science)**  <https://github.com/jupyter/jupyter/wiki/A-gallery-of-interesting-Jupyter-Notebooks#statistics-machine-learning-and-data-science>  **Open Source data science projects**  <https://opensource.com/article/19/2/learn-data-science-ai>  **Pandas Tutorials**  <https://www.datacamp.com/community/tutorials/joining-dataframes-pandas>  <https://www.earthdatascience.org/courses/earth-analytics-bootcamp/data-wrangling/data-wrangling-pandas/> |  |  | | --- | | **Python Libraries:** | | **NumPy Reference**  <https://numpy.org/doc/stable>  **SciPy Reference** <https://docs.scipy.org/doc/scipy/reference/> <https://scipy-lectures.org/packages/statistics/index.html>  **scikit-learn**  <https://scikit-learn.org/stable/user_guide.html> <https://scikit-learn.org/stable/modules/classes.html> **scikit-learn-videos**  <https://github.com/justmarkham/scikit-learn-videos>  <https://www.youtube.com/playlist?list=PL5-da3qGB5ICeMbQuqbbCOQWcS6OYBr5A>  **StatsModels**  <https://www.statsmodels.org/stable/api.html>  **Matplotlib**  <https://matplotlib.org/> |  |  | | --- | | **JupyterLab:** | | **Documentation**  <https://jupyterlab.readthedocs.io/en/stable/>  **Notebook**  <https://jupyterlab.readthedocs.io/en/stable/user/notebook.html>  **Running Notebook**  <https://jupyter.readthedocs.io/en/latest/running.html>  **Exporting Notebooks**  <https://jupyterlab.readthedocs.io/en/stable/user/export.html>  **JupyterLab Features:**  **TOC**  <https://github.com/jupyterlab/jupyterlab-toc>  **Data Explorer**  <https://github.com/jupyterlab/jupyterlab-data-explorer>  **Git**  <https://github.com/jupyterlab/jupyterlab-git> |  |  | | --- | | **Dev Environment: Install, Setup, Configure** | | **How to Organize Your Project: Best Practices for Open Reproducible Science**  <https://www.earthdatascience.org/courses/intro-to-earth-data-science/open-reproducible-science/>  **Manage your Data Science project structure in early stage**  <https://towardsdatascience.com/manage-your-data-science-project-structure-in-early-stage-95f91d4d0600> | |

|  |
| --- |
|  |
| |  | | --- | | **Data Science (Python): Project Repositories** | |  |  |  | | --- | | Optimizing for cost, availability and throughput by selecting your AWS Batch allocation strategy  by Bala Thekkedath | on 24 OCT 2019 | in Advanced (300), Amazon EC2, AWS Batch  <https://aws.amazon.com/blogs/compute/optimizing-for-cost-availability-and-throughput-by-selecting-your-aws-batch-allocation-strategy/> | | AWS offers a broad range of instances that are advantageous for batch workloads. The scale and provisioning speed of AWS’ compute instances allow you to get up and running at peak capacity in minutes without paying for downtime. Today, I’m pleased to introduce allocation strategies: a significant new capability in [AWS Batch](https://docs.aws.amazon.com/batch/latest/userguide/allocation-strategies.html) that  makes provisioning compute resources flexible and simple. In this blog post, I explain how the AWS Batch allocation strategies work, when you should use them for your workload, and provide an example CloudFormation script. This blog helps you get started on building your personalized Compute Environment (CE) most appropriate to your workloads.  **Overview**  AWS Batch is a fully managed, cloud-native batch scheduler. It manages the queuing and scheduling of your batch jobs, and the resources required to run your jobs. One of AWS Batch’s great strengths is the ability to manage instance provisioning as your workload requirements and budget needs change. AWS Batch takes advantage of AWS’s broad base of compute types. For example, you can launch compute based instances and memory instances that can handle different workload types, without having to worry about building a cluster to meet peak demand.  Previously, AWS Batch had a cost-controlling approach to manage compute instances for your workloads. The service chose an instance that was the best fit for your jobs based on vCPU, memory, and GPU requirements, at the lowest cost. Now, the newly added allocation strategies provide flexibility. They allow AWS Batch to consider capacity and throughput in addition to cost when provisioning your instances. This allows you to leverage different priorities when launching instances depending on your workloads’ needs, such as: controlling cost, maximizing throughput, or minimizing [Amazon EC2 Spot](https://aws.amazon.com/ec2/spot/) instances interruption rates.  There are now three instance allocation strategies from which to choose when creating an AWS Batch [Compute Environment](https://docs.aws.amazon.com/batch/latest/userguide/compute_environments.html) (CE). They are:  1.        Spot Capacity Optimized  2.        Best Fit Progressive  3.        Best Fit |  |  | | --- | |  | |  |  |  | | --- | |  | |  |  |  | | --- | |  | |  | |

|  |
| --- |
| **Knowledge of Statistics, Analytics, Machine Learning (Python)** |
| |  | | --- | | **Data Science (Python): Project Repositories** | |  |  |  | | --- | |  | |  |  |  | | --- | |  | |  |  |  | | --- | |  | |  |  |  | | --- | |  | |  | |

**[K2]:** **Knowledge of Inferential Statistics, Analytics, Machine Learning (R, RStudio)**

|  |
| --- |
| **Knowledge of Statistics, Analytics, Machine Learning (R)** |
| **R / RStudio:**  Introduction to R and Rstudio 1:31:20 <https://www.youtube.com/watch?v=lL0s1coNtRk>  Introduction to R and RStudio part 2 1:27:23 <https://www.youtube.com/watch?v=ZA28sOmq7nU>  Introduction to ggplot in R 1:17:24 <https://www.youtube.com/watch?v=1GmQ5BdAhG4>  Cluster analysis <https://www.youtube.com/watch?v=PX5nSBGB5Tw>  Principal Components Analysis in R0**:**26:48 <https://www.youtube.com/watch?v=xKl4LJAXnEA>  **Tidyverse:** <https://www.tidyverse.org/>  - dplyr <https://dplyr.tidyverse.org/>  - tidyr <https://tidyr.tidyverse.org/>  - tibble <https://tibble.tidyverse.org/>  Introduction to dplyr  <https://dplyr.tidyverse.org/articles/dplyr.html>  Data Transformation with dplyr: Cheat Sheet  <https://github.com/rstudio/cheatsheets/blob/master/data-transformation.pdf> |

|  |
| --- |
| **geoquery**: a bridge between the gene expression omnibus (GEO) and bioconductor  Bioinformatics, **14**, 1846–1847. <https://www.bioconductor.org/packages/release/bioc/html/GEOquery.html>  <https://www.bioconductor.org/packages/release/bioc/vignettes/GEOquery/inst/doc/GEOquery.html#platforms> |

+++++++

RNA-Seq Workflows

Alignment Based

Alignment Free

**GOOD BIG PICTURE of Bioinformatics**

|  |
| --- |
| **Collection of Bioinformatics Tools**  <https://github.com/jdidion/biotools/blob/main/README.md>  Excellent |
| Click [Raw] for actual Markdown text which renders as a good hyperlinked page |

<div id="F1">

|  |
| --- |
| **[F1]: Concepts/terminology of Molecular Biology, Genetics** |
| |  | | --- | | **Scitable by Nature Education**  Excellent Hyperlinked Text  <https://www.nature.com/scitable/topics/>  <https://www.nature.com/scitable/topic/genetics-5/>  <https://www.nature.com/scitable/ebooks/>  <https://www.nature.com/scitable/index/> |  |  | | --- | | **Expanded encyclopaedias of DNA elements in the human and mouse genomes**  <https://www.nature.com/articles/s41586-020-2493-4> |  |  | | --- | | **NCBI Genome Assemblies and Resources**  Genome Assemblies and Annotation: Information concerning how assemblies are produced, maintained and annotated.  <https://www.ncbi.nlm.nih.gov/projects/genome/index.shtml> |  |  | | --- | | **NCBI Genome Glossary**  Commonly Used Genome Terms  <https://www.ncbi.nlm.nih.gov/projects/genome/glossary.shtml> |  |  | | --- | | **NIH / National Human Genome Research Institute (NHGRI) / Glossary of Terms**  <https://www.genome.gov/genetics-glossary> | |

|  |
| --- |
| **[F1]: Concepts/terminology of Bioinformatics (File Formats, )** |
| |  | | --- | | **Necessary for understanding software tools, algorithms** |  |  | | --- | | **1000 Genomes Project FAQ**  Reading the questions and answers is good for learning about file formats of the data. <https://www.internationalgenome.org/faq> |  |  | | --- | |  |  |  | | --- | |  |  |  | | --- | |  | |

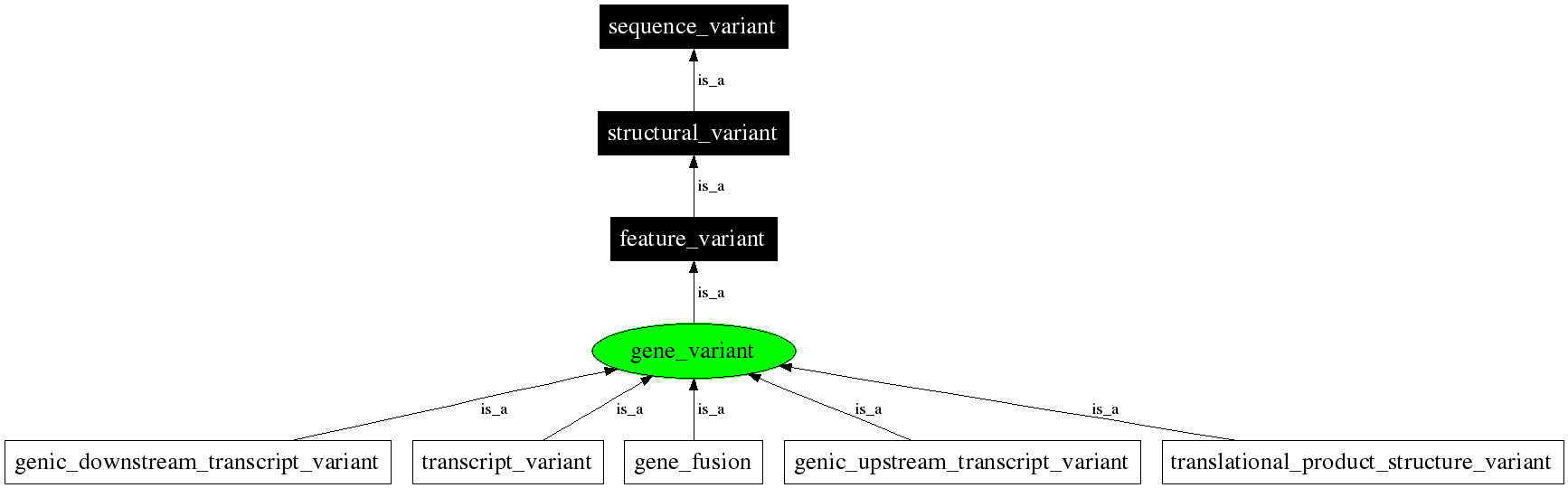
**Sequence Ontology (SO)**

Gene Ontology Consortium

<http://www.sequenceontology.org/>

Sequence Ontology is a set of terms and relationships used to describe the features and attributes of biological sequence. SO includes different kinds of features which can be located on the sequence. Biological features are those which are defined by their disposition to be involved in a biological process. Examples are binding\_site and exon. Biomaterial features are those which are intended for use in an experiment such as aptamer and PCR\_product. There are also experimental features which are the result of an experiment. SO also provides a rich set of attributes to describe these features such as “polycistronic” and “maternally imprinted”.

## gene\_variant



|  |
| --- |
| **[F2]: Genomic Sequencing Technologies** |
| |  | | --- | | **High Throughput**  <https://grcf.jhmi.edu/dna-services/sequencing/high-throughput-sequencing/>  **Long Read Sequencing**  <https://grcf.jhmi.edu/dna-services/sequencing/long-read-sequencing/>  **Medium Throughput**  <https://grcf.jhmi.edu/dna-services/sequencing/medium-throughput-sequencing/>  **PCR Support**  <https://grcf.jhmi.edu/dna-services/sequencing/pcr-support/>  **Pyrosequencing**  <https://grcf.jhmi.edu/dna-services/sequencing/pyrosequencing/>  **Sanger**  <https://grcf.jhmi.edu/dna-services/sequencing/sanger-sequencing/>  **Whole Exome/Targeted**  <https://grcf.jhmi.edu/dna-services/sequencing/whole-exome-targeted-sequencing/>  **Whole Genome**  <https://grcf.jhmi.edu/dna-services/sequencing/whole-genome-sequencing/> |  |  | | --- | | **Related data files formats**  Sanger FASTQ  alignment files  variant calls  Annotated variant lists  SNPS/indels in VCF format  BAM alignment files  QC report  BED files for regions targeted  Genotyping files |  |  | | --- | | **RNA-seqlopedia**  [**https://rnaseq.uoregon.edu/**](https://rnaseq.uoregon.edu/)  provides an overview of RNA-seq and of the choices necessary to carry out a successful RNA-seq experiment.  Experimental Design, RNA Preparation, Library Preparation, Sequencing, Analysis  Research Areas of RNA biology include:  RNA structure analysis  RNA alignment  RNA annotation  RNA-protein interaction  RNA-seq analysis  RNA target prediction  ribosome profiling | |

|  |
| --- |
| **[F3]: Open Access Journals (searching, reading): Identify current Best Practice bioinformatic tools, Standard Workflows** |
| **DOAJ (Directory of Open Access Journals)**  curated list of open access journals  <https://doaj.org/search?source=%7B%22query%22%3A%7B%22match_all%22%3A%7B%7D%7D%7D>  <https://doaj.org/subjects>  **Bioinformatics Organization / Journals**  <http://www.bioinformatics.org/wiki/Journals>     |  | | --- | | **BioMed Central (BMC) Springer Nature**  <https://www.biomedcentral.com/journals-a-z#jump-to-B> | | **BMC Bioinformatics**  <https://bmcbioinformatics.biomedcentral.com/>  An open access, peer-reviewed journal that considers articles on all aspects of the development, testing and novel application of computational and statistical methods for the modeling and analysis of all kinds of biological data, as well as other areas of computational biology.  **BMC Algorithms for Molecular Biology**  <https://almob.biomedcentral.com/>  **BMC BioData Mining**  <https://biodatamining.biomedcentral.com/> | | **BMC Genetics**  <https://bmcgenet.biomedcentral.com/>  **BMC Medical Genetics**  <https://bmcmedgenet.biomedcentral.com/>  is an open access journal publishing original peer-reviewed research articles in the effects of genetic variation in individuals, families and among populations in relation to human health and disease.  **BMC Genome Biology**  <https://genomebiology.biomedcentral.com/> |  |  | | --- | | **bioRxiv** Free online archive for unpublished preprints in the life sciences.  by Cold Spring Harbor Laboratory, a not-for-profit research and educational institution.  <https://www.biorxiv.org/>  <https://www.biorxiv.org/search> | | **Bioinformatics**  <https://www.biorxiv.org/collection/bioinformatics>  **Genomics**  <https://www.biorxiv.org/collection/genomics> |  |  | | --- | | **Public Library of Science (PLOS)**  <https://journals.plos.org/plosone/>  <https://journals.plos.org/plosone/search> | | **PLOS Computational Biology**  <https://journals.plos.org/ploscompbiol/>  <https://journals.plos.org/ploscompbiol/search>  **Subject Areas**  Eigenvalues  Linear algebra  Mathematical and statistical techniques  Mathematics  Multivariate analysis  Principal component analysis  Statistical methods  Statistics | | **PLOS Genetics**  <https://journals.plos.org/plosgenetics/search> | |

|  |
| --- |
| **[F4]: Publicly Available Genomic Data Resources (navigation, content, search)** |
|  |

|  |
| --- |
| **NCBI Gene Expression Omnibus (GEO)** <https://www.ncbi.nlm.nih.gov/geo/>  **NCBI GEO DataSets (GDS)**  <https://www.ncbi.nlm.nih.gov/gds/>  Functional genomics studies |
|  |
|  |

**NCI / Genetic Data Commons (GDC)**

<https://gdc.cancer.gov/>

<https://portal.gdc.cancer.gov/>

|  |
| --- |
| **Knowledge of Statistics, Machine Learning:** **[4]** |

**Knowledge of:**

**[K0]: Knowledge of Inferential Statistics, Analytics, Machine Learning**

**[K1]:** **Knowledge of Inferential Statistics, Analytics, Machine Learning (Python, JupyterLab, Jupyter Notebooks)**

**[K2]:** **Knowledge of Inferential Statistics, Analytics, Machine Learning (R, RStudio)**

**[K3]: Knowledge of Inferential Statistics, Analytics, Machine Learning (Bioinformatics Tools)**

**[K4]: Knowledge of Bioinformatics Tools (Bioconductor R Packages)**

**[K5]:** **Knowledge of Publicly Available Genomic Data Resources, Viewers**

(navigation, search, content interpretation, data accession)

|  |
| --- |
| **[K0]: Knowledge of Inferential Statistics, Analytics, Machine Learning** |
| |  | | --- | | **Inferential Statistics, Analytics, Machine Learning (R)** | | **P Values, clearly explained**  <https://www.youtube.com/watch?v=5Z9OIYA8He8&list=PLblh5JKOoLUJo2Q6xK4tZElbIvAACEykp&index=16>  **Linear Regression in R**  <https://www.youtube.com/watch?v=u1cc1r_Y7M0&list=PLblh5JKOoLUJo2Q6xK4tZElbIvAACEykp&index=24> **Linear Models Pt.1 - Linear Regression**  <https://www.youtube.com/watch?v=nk2CQITm_eo&list=PLblh5JKOoLUJo2Q6xK4tZElbIvAACEykp&index=23>  **Linear Models Pt.2 - t-tests and ANOVA**  <https://www.youtube.com/watch?v=NF5_btOaCig&list=PLblh5JKOoLUJo2Q6xK4tZElbIvAACEykp&index=25>  **Linear Models Pt.3 - Design Matrices (old version)**  <https://www.youtube.com/watch?v=2UYx-qjJGSs&list=PLblh5JKOoLUJo2Q6xK4tZElbIvAACEykp&index=26>  **Linear Models Pt.3 - Design Matrix Examples in R**  <https://www.youtube.com/watch?v=Hrr2anyK_5s&list=PLblh5JKOoLUJo2Q6xK4tZElbIvAACEykp&index=27>  **PCA main ideas**  <https://www.youtube.com/watch?v=HMOI_lkzW08&list=PLblh5JKOoLUJo2Q6xK4tZElbIvAACEykp&index=21>  **Principal Component Analysis (PCA) clearly explained**  <https://www.youtube.com/watch?v=_UVHneBUBW0&list=PLblh5JKOoLUJo2Q6xK4tZElbIvAACEykp&index=22>  **Principal Component Analysis (PCA), Step-by-Step**  <https://www.youtube.com/watch?v=FgakZw6K1QQ&list=PLblh5JKOoLUJo2Q6xK4tZElbIvAACEykp&index=3>  **PCA Practical Tips**  [**https://www.youtube.com/watch?v=oRvgq966yZg&feature=youtu.be**](https://www.youtube.com/watch?v=oRvgq966yZg&feature=youtu.be)  **PCA in R**  <https://www.youtube.com/watch?v=0Jp4gsfOLMs&list=PLblh5JKOoLUJo2Q6xK4tZElbIvAACEykp&index=4>  **PCA in Python**  <https://www.youtube.com/watch?v=Lsue2gEM9D0&list=PLblh5JKOoLUJo2Q6xK4tZElbIvAACEykp&index=5>  **RPKM, FPKM, TPM**  <https://www.youtube.com/watch?v=TTUrtCY2k-w&list=PLblh5JKOoLUJo2Q6xK4tZElbIvAACEykp&index=6>  **MDS and PCoA**  <https://www.youtube.com/watch?v=GEn-_dAyYME&list=PLblh5JKOoLUJo2Q6xK4tZElbIvAACEykp&index=7>  **MDS and PCoA in R**  <https://www.youtube.com/watch?v=pGAUHhLYp5Q&list=PLblh5JKOoLUJo2Q6xK4tZElbIvAACEykp&index=15>  **t-SNE, Clearly Explained**  <https://www.youtube.com/watch?v=NEaUSP4YerM&list=PLblh5JKOoLUJo2Q6xK4tZElbIvAACEykp&index=8>  **K-means clustering**  <https://www.youtube.com/watch?v=4b5d3muPQmA&list=PLblh5JKOoLUJo2Q6xK4tZElbIvAACEykp&index=28>  **Hierarchical Clustering**  <https://www.youtube.com/watch?v=7xHsRkOdVwo&list=PLblh5JKOoLUJo2Q6xK4tZElbIvAACEykp&index=10>  **Drawing and Interpreting Heatmaps**  <https://www.youtube.com/watch?v=oMtDyOn2TCc&list=PLblh5JKOoLUJo2Q6xK4tZElbIvAACEykp&index=11>  **False Discovery Rates, FDR, clearly explained**  <https://www.youtube.com/watch?v=K8LQSvtjcEo&list=PLblh5JKOoLUJo2Q6xK4tZElbIvAACEykp&index=17>  **Fisher's Exact Test and the Hypergeometric Distribution**  <https://www.youtube.com/watch?v=udyAvvaMjfM&list=PLblh5JKOoLUJo2Q6xK4tZElbIvAACEykp&index=18>  **Logs (logarithms), clearly explained**  <https://www.youtube.com/watch?v=VSi0Z04fWj0&list=PLblh5JKOoLUJo2Q6xK4tZElbIvAACEykp&index=20> |  |  | | --- | | **Knodge of Staics, Mae Lear** | | **5 Questions which can teach you Multiple Regression (with R and Python)** October 15, 2015  <https://www.analyticsvidhya.com/blog/2015/10/regression-python-beginners/>  **Going Deeper into Regression Analysis with Assumptions, Plots & Solutions** July 14, 2016  <https://www.analyticsvidhya.com/blog/2016/07/deeper-regression-analysis-assumptions-plots-solutions/>  **7 Regression Techniques you should know** August 14, 2015  <https://www.analyticsvidhya.com/blog/2015/08/comprehensive-guide-regression/>  **Statistics for Analytics and Data Science: Hypothesis Testing and Z-Test vs. T-Test** June 18, 2020  <https://www.analyticsvidhya.com/blog/2020/06/statistics-analytics-hypothesis-testing-z-test-t-test/>  **Commonly used Machine Learning Algorithms (with Python and R Codes)** September 9, 2017  <https://www.analyticsvidhya.com/blog/2017/09/common-machine-learning-algorithms/> |  |  | | --- | | **A Matrix Algebra Companion for Statistical Learning (matrix4sl)**  <https://www.gastonsanchez.com/matrix4sl/types-of-tables.html> |   **Biomedical Data Science**  <http://genomicsclass.github.io/book/> |

|  |
| --- |
| **Relevant Journal Publications:**  [**https://journals.plos.org/ploscompbiol/search**](https://journals.plos.org/ploscompbiol/search)  **All Fields: bioconductor**  **Publication Date: 2016-01-01 - 2020-10-09**  **Sort By: Most Bookmarked**  **Subject Area: Principal component analysis**  >> results  **Ten quick tips for effective dimensionality reduction**  20 Jun 2019 PLOS Computational Biology  <https://doi.org/10.1371/journal.pcbi.1006907>  <https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1006907>  Citations: 19  Context Specific and Differential Gene Co-expression Networks via Bayesian Biclustering  28 Jul 2016 PLOS Computational Biology  <https://doi.org/10.1371/journal.pcbi.1004791>  Citations: 22  Machine learning-based microarray analyses indicate low-expression genes might collectively influence PAH disease  12 Aug 2019 PLOS Computational Biology  <https://doi.org/10.1371/journal.pcbi.1007264>  Citations: 2 |

**[S1]:** **Special effort;**

NGS RNA-Seq bioinformatics tools, workflows

"differential expression"  
"exploratory data analysis"

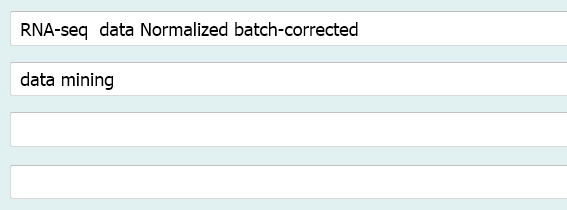
I am especially attracted to the aspects of workflows following the pre-processing, normalization (Quantification of Mapped Reads, Differential Gene Expression).

The Application of Inferential Statistics, Machine Learning (PCA, Feature/Dimensional Reduction, Clustering).

**Google Search**

[**enabling cross study analysis of rna-sequencing data**](https://www.google.com/searchlr=&as_qdr=all&sxsrf=ALeKk00QSjfjp_4g_LC4EyNExgXHpC1w4w:1606327451528&q=enabling+cross+study+analysis+of+rna-sequencing+data&sa=X&ved=2ahUKEwi-jZCJpJ7tAhVkw1kKHXaKCWMQ1QIoAHoECAUQAQ)

<https://www.google.com/searchlr=&as_qdr=all&sxsrf=ALeKk00QSjfjp_4g_LC4EyNExgXHpC1w4w:1606327451528&q=enabling+cross+study+analysis+of+rna-sequencing+data&sa=X&ved=2ahUKEwi-jZCJpJ7tAhVkw1kKHXaKCWMQ1QIoAHoECAUQAQ>

  
<https://www.google.com/search?as_q=RNA-seq++data+Normalized+batch-corrected+&as_epq=data+mining+&as_oq=&as_eq=&as_nlo=&as_nhi=&lr=&cr=&as_qdr=all&as_sitesearch=&as_occt=any&safe=images&as_filetype=&tbs=>

**Keywords:**

RNA-seq

data mining

cross study analysis

data normalization

batch-corrected RNA-seq

|  |
| --- |
| **[S1]:** **Special study of;   NGS RNA-Seq: Bioinformatics tools, Workflows** |
| **Relevant journal publications**   |  | | --- | | **RNA-seq workflow: gene-level exploratory analysis and differential expression**  2015 Oct 14  <http://master.bioconductor.org/packages/release/workflows/vignettes/rnaseqGene/inst/doc/rnaseqGene.html> | | Abstract  Here we walk through an end-to-end gene-level RNA-Seq differential expression workflow using Bioconductor packages. We will start from the FASTQ files, show how these were aligned to the reference genome, and prepare a count matrix which tallies the number of RNA-seq reads/fragments within each gene for each sample. We will perform exploratory data analysis (EDA) for quality assessment and to explore the relationship between samples, perform differential gene expression analysis, and visually explore the results.  Keywords  RNA-seq, differential expression, gene expression, Bioconductor, statistical analysis, high-throughput sequencing, visualization, genomics | |  |  |  | | --- | | **RNA-seq: Basic Bioinformatics Analysis**  Curr Protoc Mol Biol. 2018 Oct  <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6168365/> | | The workflow includes three parts:  (a) mapping sequencing reads to a reference genome or transcriptome;  (b) quantifying expression levels of individual genes and transcripts;  (c) identifying specific genes and transcripts that are differentially expressed between samples.  Alignment/Mapping:  the sequence of each read to a reference genome, annotation of genes  - **STAR** alignment tool  Quality Assessment:  After mapping reads to the genome, it is important to survey the quality of the RNA-seq data  - **Picard**  - **SAMTools**  Quantification of Mapped Reads:  - **HTseq** to quantify sequencing reads mapped to each gene  (a) identify genes that are differentially expressed between conditions (sample groups),  (b) derive gene expression values for each individual transcript  Approaches for normalization include  CPM (counts per million reads),  RPKM (reads per kilobase per million reads),  FPKM (fragments per kilobase per million reads),  TPM (transcripts per million reads).  Differential Gene Expression Analysis:  **EdgeR** (Robinson et al., 2010)  **DESeq2** (Love et al. 2014)  Download and install required tools:  STAR: <https://github.com/alexdobin/STAR>  Picard: <https://broadinstitute.github.io/picard/>  HTseq: <https://htseq.readthedocs.io/en/release_0.9.1/install.html>  R: <https://www.r-project.org> |  |  | | --- | | **ngs.plot: Quick mining and visualization of next-generation sequencing data by integrating genomic databases**  BMC Genomics 2014 volume 15, Article number: 284  415 Citations  <https://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-15-284> | |  |  |  | | --- | | **pcaExplorer: an R/Bioconductor package for interacting with RNA-seq principal components**  BMC Bioinformatics 2019 volume 20, Article number: 331  17 Citations  <https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-019-2879-1> | | **pcaExplorer**  <http://bioconductor.org/packages/release/bioc/html/pcaExplorer.html>  **pcaExplorer User Guide**  27 October 2020  <http://bioconductor.org/packages/release/bioc/vignettes/pcaExplorer/inst/doc/pcaExplorer.html>  **Up and running with pcaExplorer**  27 October 2020  <http://bioconductor.org/packages/release/bioc/vignettes/pcaExplorer/inst/doc/upandrunning.html>  data component (count matrix, experimental data, dds object, annotation) |      |  | | --- | | **A Beginner’s Guide to Analysis of RNA Sequencing Data**  <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6096346/>  Am J Respir Cell Mol Biol. 2018 Aug | | **Abstract:**  A general understanding of the principles underlying each step of RNA-seq data analysis allows investigators without a background in programming and bioinformatics to critically analyze their own datasets as well as published data. Our goals in the present review are to break down the steps of a typical RNA-seq analysis and to highlight the pitfalls and checkpoints along the way that are vital for bench scientists and biomedical researchers performing experiments that use RNA-seq.  **Keywords:**  RNA sequencing, transcriptomics, bioinformatics, data analysis  **Associated Data:**  The RNA-seq data reported in this article has been deposited in NCBI’s Gene Expression Omnibus (GEO) and are accessible through GEO Series accession number GSE116583.  **Experimental Design and Approach:**  A major goal of RNA-seq analysis is to identify differentially expressed and coregulated genes and to infer biological meaning for further studies. Source material can be cells cultured in vitro, whole-tissue homogenates, or sorted cells. The ability to interpret findings depends on appropriate experimental design, implementation of controls, and correct analysis. Every effort should be made to minimize batch effect, because small and uncontrolled changes in an environment can result in identification of differentially expressed genes (DEGs) unrelated to the designed experiment. Sources of batch effect can occur during the experiment, during the RNA library preparation, or during the sequencing run and include but are not limited to those listed in Table 1. Once a well-designed and controlled experiment is performed, a structured approach to the dataset allows for quality control followed by unbiased analysis of the data. In the present analysis, we use an approach that includes setting low count filtering, establishing a noise threshold, checking for potential outliers, running appropriate statistical tests to identify DEGs, clustering of genes by expression pattern, and testing for gene ontology (GO) enrichment. For each of these analysis components, we aim to highlight important checkpoints and quality controls that will streamline and strengthen data analysis, avoid bias, and allow investigators to maximally use their datasets.  **Clustering**  The two most common clustering methods used for RNA-seq data analysis are hierarchical and k-means clustering (see Clustering box). The most common form of hierarchical clustering is a bottom-up agglomerative approach that organizes the data into a tree structure without user input by starting with each data point as its own cluster and iteratively combining them into larger clusters or “clades.” In contrast, k-means clustering requires the investigator to define the number of clusters (k) a priori, and data are then sorted into the cluster with the nearest mean. It is possible to assess a range of k-values to decide how to best capture the trends. In addition, various tools such as Silhouette exist to help the investigator determine the ideal k-value, but some subjectivity remains (21). By adjusting the k, the investigator may set the degree of granularity they would like to achieve with the data. For either approach, the user must specify the distance metric by which data points are considered similar. Typically, Pearson’s correlation is used, and this is generally the default in software designed for RNA-seq analysis. Both approaches are widely used, and both aid the investigator in identifying groups of genes that display similar expression patterns, allowing for further downstream analyses. The clusters can then be used as input for an analysis of functional enrichment (see next section).  Why do we use clustering on RNA-seq data?  Clustering of RNA-seq data may be used to identify patterns of gene expression by grouping genes based on their distance in an unsupervised manner. Clustering RNA-seq data is used as an exploratory tool that allows the user to organize and visualize relationships between groups of genes, and to select certain genes for further consideration.  **Hierarchical clustering**  The most commonly used hierarchical clustering approach is a form of agglomerative, or bottom-up, clustering that iteratively merges clusters (originally consisting of individual data points) into larger clusters or “clades”.  **K-means clustering**  Data points are iteratively partitioned into clusters based on the minimum distance to the cluster mean. The number of clusters (k) is set by the investigator. |      |  | | --- | | **Review of RNA-Seq Data Analysis Tools**  February 17, 2016  <https://rna-seqblog.com/review-of-rna-seq-data-analysis-tools/> |      |  | | --- | | **Feature-based classification of human transcription factors into hypothetical sub-classes related to regulatory function**  BMC Bioinformatics volume 17, Article number: 459 (2016)  7 Citations  <http://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-016-1349-2> |      |  | | --- | | **GEO2RNAseq: An easy-to-use R pipeline for complete pre-processing of RNA-seq data**  September 16, 2019 bioRxiv Preprint  <https://www.biorxiv.org/content/10.1101/771063v1.full> |      |  | | --- | | **RNAseq data analysis in R - Notebook**  <http://monashbioinformaticsplatform.github.io/RNAseq-DE-analysis-with-R/RNAseq_DE_analysis_with_R.html> | | Install and load packages  Mapping reads to a reference genome  Count reads for each feature  QC and stats  Differential Expression  Gene Annotation  Gene Set Enrichment |      |  | | --- | | **Introduction to differential gene expression analysis using RNA-seq**  September 2015 updated November 14, 2019  <https://chagall.med.cornell.edu/RNASEQcourse/Intro2RNAseq.pdf> | |

|  |
| --- |
| **[S1]:** **Special effort;   NGS RNA-Seq: Statistical Methods, ML, Data Mining** |
| **Journal Publications:**   |  | | --- | |  | |  | |  |  |  | | --- | | **Google**  **enabling cross-study analysis of RNA-Sequencing data**  [https://www.google.com/search?lr=&as\_qdr=all&sxsrf=AleKk00QSjfjp\_4g\_LC4EyNExgXHpC1w4w:1606327451528&q=enabling+cross+study+analysis+of+rna-sequencing+data&sa=X&ved=2ahUKEwi-jZCJpJ7tAhVkw1kKHXaKCWMQ1QIoAHoECAUQAQ](https://www.google.com/search?lr=&as_qdr=all&sxsrf=ALeKk00QSjfjp_4g_LC4EyNExgXHpC1w4w:1606327451528&q=enabling+cross+study+analysis+of+rna-sequencing+data&sa=X&ved=2ahUKEwi-jZCJpJ7tAhVkw1kKHXaKCWMQ1QIoAHoECAUQAQ)  **Google**  **RNA-Seq data mining** | | **A survey of best practices for RNA-seq data analysis**  Genome Biology volume 17, Article number: 13 (2016)  769 Citations  <https://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-0881-8>  **Massive mining of publicly available RNA-seq data from human and mouse**  Nature Communications volume 9, Article number: 1366 (2018)  96 Citations  <https://www.nature.com/articles/s41467-018-03751-6>  **Omics Playground: Explore Omics Data Freely**  <https://omicsplayground.readthedocs.io/en/latest/index.html>  Omics Playground is a comprehensive self-service platform platform for visualization, analytics and exploration of Big Omics Data. It allows users to apply a multitude of state-of-the-art analysis tools to their own data to explore and discover underlying biology in a short time.  The platform offers a unique combination of features that distinguishes it from the other analytics platforms currently available. We believe that data preprocessing (primary analysis) and statistical testing (secondary analysis) are now well established, and the most challenging task is currently data interpretation (tertiary analysis) that often takes the longest time but where actual insights can be gained. Therefore, Omics Playground focuses strongly on tertiary analysis while providing good support for secondary analysis.  Reanalyzing Public Datasets  To illustrate the use case of the Omics Playground, we reanalyzed different types of publics datasets, including microarray, bulk RNA-seq, single-cell RNA-seq and proteomic datasets to recapitulate the results.  <https://omicsplayground.readthedocs.io/en/latest/examples/examples.html> |  |  | | --- | | **comparing rna-seq datasets** | | **Broad Institute Gene Set Enrichment Analysis (GSEA)**  <https://software.broadinstitute.org/cancer/software/gsea/wiki/index.php/Using_RNA-seq_Datasets_with_GSEA>  **BioStars**  **Question: RNA-seq data comparison across experiments**  <https://www.biostars.org/p/117451/>  **BioStars**  **Question: Comparing similarity of RNAseq datasets**  <https://www.biostars.org/p/334235/>  **Analysis of public RNA-sequencing data reveals biological consequences of genetic heterogeneity in cell line populations**  Scientific Reports volume 8, Article number: 11226 (2018) Cite this article  7 Citations  <https://www.nature.com/articles/s41598-018-29506-3> |   **Books:**  **NCBI Books Advanced Search**  <https://www.ncbi.nlm.nih.gov/books/advanced/>   |  | | --- | | **Computational Biology**  Brisbane (AU): Codon Publications; 2019 Nov 21.  ISBN-13: 978-0-9944381-9-5  <https://www.ncbi.nlm.nih.gov/books/NBK550339/>  Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0)  <https://creativecommons.org/licenses/by-nc/4.0/> | | Chapter 4 Biological Sequence Analysis  <https://www.ncbi.nlm.nih.gov/books/NBK550342/>  Chapter 5 Multivariate Statistical Methods for High-Dimensional Multiset Omics Data Analysis  <https://www.ncbi.nlm.nih.gov/books/NBK550343/>  Chapter 6 Statistical Methods for RNA Sequencing Data Analysis  <https://www.ncbi.nlm.nih.gov/books/NBK550334/> |      |  | | --- | |  | |  |      |  | | --- | |  |      |  | | --- | |  |      |  | | --- | |  |      |  | | --- | |  | |  |      |  | | --- | |  | |

|  |
| --- |
| **[K1]: Knowledge of; Inferential Statistics, Analytics, Machine Learning: Applied** |
| |  | | --- | | **StatQuest: (Josh Starmer)**  **U North Carolina Chapel Hill**  [**https://statquest.org/video-index/**](https://statquest.org/video-index/) | | **High-throughput Sequencing Analysis: StatQuest (Josh Starmer)**  **Introduction to RNA-seq**  <https://www.youtube.com/watch?v=tlf6wYJrwKY&list=PLblh5JKOoLUJo2Q6xK4tZElbIvAACEykp&index=1>  **Introduction to ChIP-Seq**  <https://www.youtube.com/watch?v=nkWGmaYRues&list=PLblh5JKOoLUJo2Q6xK4tZElbIvAACEykp&index=2> **RNA-seq**  <https://statquest.org/tag/rna-seq/>  **edgeR, part1: Library Normalization**  <https://www.youtube.com/watch?v=Wdt6jdi-NQo&feature=youtu.be>  **DESeq2, part1: Library Normalization**  <https://www.youtube.com/watch?v=UFB993xufUU&feature=youtu.be>  **edgeR and DESeq2, part2: Independent Filtering (removing genes with low read counts)**  <https://www.youtube.com/watch?v=Gi0JdrxRq5s&feature=youtu.be>  **FDR and the Benjamini-Hochberg Method**  <https://statquest.org/statquest-fdr-and-the-benjamini-hochberg-methoc-clearly-explained/>  **Linear Discriminant Analysis (LDA)**  <https://statquest.org/statquest-linear-discriminant-analysis-lda-clearly-explained/>  **Heatmaps how to draw and interpret them**  <https://statquest.org/heatmaps-how-to-draw-and-interpret-them/>  **RNA-seq: The Pitfalls of Technical Replicates**  <https://statquest.org/rna-seq-replicates-clearly-explained/>  **PCA**  <https://statquest.org/pca-clearly-explained/>  **RPKM, FPKM and TPM**  <https://statquest.org/rpkm-fpkm-and-tpm-clearly-explained/> | |

|  |
| --- |
| **[K5]:** **Knowledge of Publicly Available Genomic Data Resources, Viewers** |
| **Genome Data Viewer (GDV) <--> GEO <--> Integrative Genomics Viewer (IGV)**  NGS data deposited in the GEO database can be visualized through the genome data viewer function  To check the quality of raw sequence data in the FASTQ format,  NGS data (FASTQ files)   |  | | --- | |  | | **RNA-seq workflow: gene-level exploratory analysis and differential expression**  <http://master.bioconductor.org/packages/release/workflows/vignettes/rnaseqGene/inst/doc/rnaseqGene.html> |  |  | | --- | | **IGV** <https://igv.org> | | **IGV**  <https://igv.org/app>  variant visualization capabilities  is a Web application which runs in a web browser and requires no downloads.  Documentation  <https://igvteam.github.io/igv-webapp/>  <https://www.youtube.com/channel/UCb5W5WqauDOwubZHb-IA_rA>  Play All: <https://www.youtube.com/watch?v=sFeK25K5PE&list=PLSplvWwdPpSrhPn3V2iuPUzyxVIDYZ1xS> | | **Relevant Publications:**  **Variant Review with the Integrative Genomics Viewer**  <https://cancerres.aacrjournals.org/content/77/21/e31.long>  See video of IGV  **Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration**  <https://academic.oup.com/bib/article/14/2/178/208453>  **Integrative Genomics Viewer: Visualizing Big Data**  <https://ocg.cancer.gov/e-newsletter-issue/issue-9/integrative-genomics-viewer-visualizing-big-data> | |

|  |
| --- |
| **[K2]:** **Knowledge of Inferential Statistics, Analytics, Machine Learning (R, RStudio)** |
| |  | | --- | | **R** | | **r-crash-course**  **half-day introduction to the R language**  <https://bioinformatics-core-shared-training.github.io/r-crash-course/>  **Statistical Analysis: Introduction using R**  <https://en.wikibooks.org/wiki/Statistical_Analysis:_an_Introduction_using_R>  **Data Manipulation and Visualisation using R**  **Intermediate R Course**  <http://bioinformatics-core-shared-training.github.io/r-intermediate/> |  |  | | --- | | **R / RStudio** | | **Introduction to R and RStudio** 1:31:20  <https://www.youtube.com/watch?v=lL0s1coNtRk>  **Introduction to R and RStudio part 2** 1:27:23  <https://www.youtube.com/watch?v=ZA28sOmq7nU>  **Introduction to ggplot in R** 1:17:24  <https://www.youtube.com/watch?v=1GmQ5BdAhG4>  **Cluster analysis**  <https://www.youtube.com/watch?v=PX5nSBGB5Tw>  **Principal components analysis in R** 26:48  <https://www.youtube.com/watch?v=xKl4LJAXnEA> | |

**[K2]:** Bioinformatics Tools

**RNA-seq workflow: gene-level exploratory analysis and differential expression**

<http://master.bioconductor.org/packages/release/workflows/vignettes/rnaseqGene/inst/doc/rnaseqGene.html>

**[K3]: Knowledge of Inferential Statistics, Analytics, Machine Learning (Bioinformatics Tools)**

|  |
| --- |
| **Categorized Software Tools** <https://bioinformaticshome.com/tools/tools-main.html> |
| |  | | --- | | [**RNA-seq Pre-analysis Tools**](https://bioinformaticshome.com/tools/rna-seq/pre-analysis.html) <https://bioinformaticshome.com/tools/rna-seq/pre-analysis.html>  Pre-analysis quality control of raw reads includes assessment of tolerable GC and k-mer contents, removal of sequence adaptors, PCR artifacts, and contaminations. The assessment of duplicates and sequencing errors. In addition, sequencing quality tends to decrease towards the 3' end of the reads; Thus, the reads must be trimmed to remove the low-quality ends. [Data Quality Assessment](https://bioinformaticshome.com/tools/rna-seq/pre-analysis.html)   * [Filtering](https://bioinformaticshome.com/tools/rna-seq/pre-analysis.html#Filtering) * [Trimming](https://bioinformaticshome.com/tools/rna-seq/pre-analysis.html#Trimming) * [Filtering and Trimming](https://bioinformaticshome.com/tools/rna-seq/pre-analysis.html#Filtering-and-Trimming) * [Reporting/Visualization](https://bioinformaticshome.com/tools/rna-seq/pre-analysis.html#Reporting-Visualization) * [Other Pre-analysis RNA-seq Tools](https://bioinformaticshome.com/tools/rna-seq/pre-analysis.html#Other) | | **FASTX-Toolkit** <https://bioinformaticshome.com/tools/rna-seq/descriptions/FASTX-Toolkit.html> **FastQC** <https://bioinformaticshome.com/tools/rna-seq/descriptions/FastQC.html> |  |  | | --- | | **RNA-seq Core Analysis Tools**  <https://bioinformaticshome.com/tools/rna-seq/core-analysis.html>  **1. Transcriptome Profiling**  1.1 Read mapping or assembly  1.1.1 De novo (reference free) transcriptome assembly  1.1.1.1 Unstranded  1.1.1.2 Stranded  1.1.1.3 Quality Control  1.1.2 Mapping to a reference genome or transcriptome  1.1.2.1 Splice Aware  1.1.2.2 Splice unaware  1.1.2.3 Quality Control  1.2 Expression Quantification  1.2.1 Union-exon Based  1.2.2 Transcript Based  1.2.3 Bacterial genome  **2. Differential Expression Analysis**  2.1 Pre-processing DEA  2.2 Parametric  2.3 Non-parametric  2.4 Power analysis  **3. Functional Profiling**  3.1 Enrichment Analysis (GSEA), annotation, other  3.2 Comparison with Genome |  |  | | --- | | **Whole Genome Assembly (WGA) Analysis Tools** - Software and Resources <https://bioinformaticshome.com/tools/wga/wga.html> | | GAML  GAML is a tool for genome assembly based on maximum likelihood. It implements a probabilistic model to take into account sequencing error rates, insert lengths and other characteristics to produce a final genome assembly. This tool can work on sequenced data generated from multiple sequencing platforms (e.g. Illumina, 454, PacBio).  Operation: Genome assembly  Software interface: Command-line user interface  Language: -  Operating system: Linux  License: Not stated  Cost: Free  <https://bioinformaticshome.com/tools/wga/descriptions/GAML.html>  GAML: genome assembly by maximum likelihood  <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4454275/> | |

|  |
| --- |
| **[K3]: Tools: Python, BioPython** |
|  |

**Uniformed Services University / Learning Resource Center**

**Introduction to NCBI Bioinformatics Resources: NCBI Overview**

<https://usuhs.libguides.com/c.php?g=468091&p=3200594>

|  |
| --- |
| **[K5]:** **Knowledge of Publicly Available Genomic Data Resources, Viewers**  (navigation, search, content interpretation, data accession) |
| |  | | --- | | **NCBI Genome Data Viewer (GDV)**  <https://www.ncbi.nlm.nih.gov/genome/gdv/browser/genome/?id=GCF_000001405.38>  The NCBI Genome Data Viewer (GDV) is a genome browser supporting the exploration and analysis of eukaryotic RefSeq genome assemblies. Genome Data Viewer is also used by different NCBI resources, such as GEO and dbGaP, to display datasets associated with specified experiments or samples in a genome browser context. | | **Tutorials**  **NCBI Genome Data Viewer (Tutorial Page Functionality)**  <https://www.ncbi.nlm.nih.gov/genome/gdv/browser/help/#LAYOUT>  **NCBI Genome Data Viewer (Tutorial 11 videos Last updated Sep 23, 2020)**  <https://www.youtube.com/playlist?list=PLH-TjWpFfWruHgL0WRzZfQwp-MWzhIj16> |  |  | | --- | | **NCBI Gene Expression Omnibus (GEO)** <https://www.ncbi.nlm.nih.gov/geo/>  GEO is a public functional genomics data repository supporting MIAME-compliant data submissions. | | **GEO DataSets (GDS)**  <https://www.ncbi.nlm.nih.gov/gds/> This database stores curated gene expression DataSets, as well as original Series and Platform records in the Gene Expression Omnibus (GEO) repository. Enter search terms to locate experiments of interest. DataSet records contain additional resources including cluster tools and differential expression queries.  **About GEO DataSets** <https://www.ncbi.nlm.nih.gov/geo/info/datasets.html>  **GEO Profiles**  <https://www.ncbi.nlm.nih.gov/geoprofiles/> This database stores individual gene expression profiles from curated DataSets in the Gene Expression Omnibus (GEO) repository. Search for specific profiles of interest based on gene annotation or pre-computed profile characteristics.  **Querying GEO DataSets and GEO Profiles**  <https://www.ncbi.nlm.nih.gov/geo/info/qqtutorial.html> | | **Related Publications**  **The Gene Expression Omnibus database Methods Mol Biol. 2016; 1418: 93–110.**  <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4944384/>  **NCBI GEO: mining millions of expression profiles database and tools Nucleic Acids Res. 2005 Jan 1**  <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC539976/> |   **NCBI Sequence Read Archive (SRA)**  <https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi>?  **SRA Explorer**  tool aims to make datasets within the Sequence Read Archive more accessible.  <https://sra-explorer.info/> |

|  |
| --- |
| **1000 Genomes Project** |
| |  | | --- | | **1000 Genomes Project FAQ**  If you have any other questions you can’t find the answer to please email [info@1000genomes.org](mailto:info@1000genomes.org)  <https://www.internationalgenome.org/faq>  **IGSR: The International Genome Sample Resource**  Open human variation data  <https://www.internationalgenome.org/home>  **1000 Genomes 30x on GRCh38**  <https://www.internationalgenome.org/data-portal/data-collection/30x-grch38>  **1000 Genomes Browser (Phase 3)**  <https://www.ncbi.nlm.nih.gov/variation/tools/1000genomes/>  **Ensembl**  <http://uswest.ensembl.org/Homo_sapiens/Info/Index> |  |  | | --- | | **Data Portal (Search by Population, Technology, Data Collection)**  <https://www.internationalgenome.org/data-portal/sample>  Technology  Exome  Low coverage WGS  Integrated variant call sets  File Format  FASTQ  VCF | |

|  |
| --- |
| **Medical Data for Machine Learning**  <https://github.com/beamandrew/medical-data>  This is a curated list of medical data for machine learning.  This list is provided for informational purposes only, please make sure you respect any and all usage restrictions for any of the data listed. |
| **3. Data derived from Electronic Health Records (EHRs)**  Building the graph of medicine from millions of clinical narratives  Co-occurence statistics for medical terms extracted from 14 million clinical notes and 260,000 patients.  Paper: <http://www.nature.com/articles/sdata201432>  Data: <http://datadryad.org/resource/doi:10.5061/dryad.jp917>  **Learning Low-Dimensional Representations of Medical Concept**  Low-dimensional embeddings of medical concepts constructed using claims data. Note that this paper utilizes data from Building the graph of medicine from millions of clinical narratives  Paper: <http://cs.nyu.edu/~dsontag/papers/ChoiChiuSontag_AMIA_CRI16.pdf>  Data: <https://github.com/clinicalml/embeddings>  **MIMIC-III, a freely accessible critical care database**  Anonymized critical care EHR database on 38,597 patients and 53,423 ICU admissions. Requires registration.  Paper: <http://www.nature.com/articles/sdata201635>  Data: <http://physionet.org/physiobank/database/mimic3cdb/>  Clinical Concept Embeddings Learned from Massive Sources of Medical Data  Embeddings for 108,477 medical concepts learned from 60 million patients, 1.7 million journal articles, and clinical notes of 20 million patients  Paper: <https://arxiv.org/abs/1804.01486>  Embeddings: <https://figshare.com/s/00d69861786cd0156d81>  Interactive tool: <http://cui2vec.dbmi.hms.harvard.edu>  Evaluation of Embeddings of Laboratory Test Codes for Patients at a Cancer Center  200 dimensional Word2Vec embeddings of 1098 laboratory test codes (LOINCs) trained from 8,280,238 lab orders for 79,081 patients at City of Hope National Medical Center (Los Angeles, CA).  Paper: <https://arxiv.org/abs/1907.09600>  Embeddings and Code: <https://github.com/elleros/DSHealth2019_loinc_embeddings>  **4. National Healthcare Data**  Centers for Disease Control and Prevention (CDC)  Data from the CDC on many areas, including:  Biomonitoring  Child Vaccinations  Flu Vaccinations  Health Statistics  Injury & Violence  MMWR  Motor Vehicle  NCHS  NNDSS  Pregnancy & Vaccination  STDs  Smoking & Tobacco Use  Teen Vaccinations  Traumatic Brain Injury  Vaccinations  Web Metrics  Landing page: <https://data.cdc.gov>  Data Catalog: <https://data.cdc.gov/browse>  **Medicare Data**  Data from the Centers for Medicare & Medicaid Services (CMS) on hospitals, nursing homes, physicians, home healthcare, dialysis, and device providers.  Landing page: <https://data.medicare.gov>  Explorer: <https://data.medicare.gov/data>  **Texas Public Use Inpatient Data File Data on 11 Million inpatient visits with diagnosis, procedure codes and outcomes from Texas**  **between 2006 & 2009.**  Link: <https://www.dshs.texas.gov/thcic/hospitals/Inpatientpudf.shtm>  **Dollars for Doctors**  Propublica investigation of money paid by pharmaceutical companies to doctors.  Information: <https://www.propublica.org/series/dollars-for-docs>  Search tool: <https://projects.propublica.org/docdollars/>  Data request: <https://projects.propublica.org/data-store/sets/health-d4d-national-2>  **DocGraph Physician interaction network obtained through a freedom of information act request. Covers nearly 1 million entities.**  Main page: <http://www.docgraph.com>  Information: <http://thehealthcareblog.com/blog/2012/11/05/tracking-the-social-doctor-opening-up-physician-referral-data-and-much-more/>  Data: <http://linea.docgraph.org>  **5. UCI Datasets**  **Liver Disorders Data Set**  Data on 345 patients with and without liver disease. Features are 5 blood biomarkers thought to be involved with liver disease.  Data: <https://archive.ics.uci.edu/ml/datasets/Liver+Disorders>  **Thyroid Disease Data Set**  Data: <https://archive.ics.uci.edu/ml/datasets/Thyroid+Disease>  **Breast Cancer Data Set**  Data: <https://archive.ics.uci.edu/ml/datasets/Breast+Cancer>  **Heart Disease Data Set**  Data: <https://archive.ics.uci.edu/ml/datasets/Heart+Disease>  **Lymphography Data Set**  Data: <https://archive.ics.uci.edu/ml/datasets/Lymphography>  **Parkinsons Data Set**  Data: <https://archive.ics.uci.edu/ml/datasets/parkinsons>  **Parkinsons Telemonitoring Data Set**  Data: <https://archive.ics.uci.edu/ml/datasets/Parkinsons+Telemonitoring>  **Parkinson Speech Dataset with Multiple Types of Sound Recordings Data Set**  Data: <https://archive.ics.uci.edu/ml/datasets/Parkinson+Speech+Dataset+with++Multiple+Types+of+Sound+Recordings>  **Parkinson's Disease Classification Data Set**  Data: <https://archive.ics.uci.edu/ml/datasets/Parkinson%27s+Disease+Classification>  Primary Tumor Dataset Data: <https://archive.ics.uci.edu/ml/datasets/primary+tumor>  **6. Biomedical Literature**  **PMC Open Access Subset**  Collection of all the full-text, open access articles in Pubmed central.  Information: <http://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/>  Archived files: <http://www.ncbi.nlm.nih.gov/pmc/tools/ftp/#Data_Mining>  **PubMed 200k RCT**  Collection of pubmed abstracts from randomized control trials (RCTs). Annotations for each sentence in the abstract are available.  Paper: <https://arxiv.org/abs/1710.06071>  Data: <https://github.com/Franck-Dernoncourt/pubmed-rct>  **Web API of PubMed Articles**  NLM also provided Web API for accessing biomedical literatures in PubMed.  Instructions for getting PubMed articles: <https://www.ncbi.nlm.nih.gov/research/bionlp/APIs/BioC-PubMed/>  (not full text, just title, abstract, etc.)  For articles in PubMed Central, instructions for getting the whole articles: https://www.ncbi.nlm.nih.gov/research/bionlp/APIs/BioC-PMC/  **EBM NLP**  Collection of pubmed abstracts from randomized control trials (RCTs). Annotation of Population, Intervention, and Outcomes (PICO elements) are available.  Paper: <https://arxiv.org/abs/1806.04185>  Data: <https://ebm-nlp.herokuapp.com/annotations>  Website: <https://ebm-nlp.herokuapp.com/index>  **Evidence Inference**  A dataset for inferring the results of randomized control trials (RCTs). A collection of pubmed RCTs from the open access subset. Annotations of (intervention, comparison intervention, outcome, significance finding, evidence span) are available.  Paper: <https://arxiv.org/abs/1904.01606>  Data: <https://github.com/jayded/evidence-inference/tree/master/annotations>  Website: <http://evidence-inference.ebm-nlp.com/>  **PubMedQA**  A dataset for biomedical research question answering. Task is to use yes/no/maybe to answer naturally occuring questions in PubMed titles.  Paper: <https://arxiv.org/abs/1909.06146>  Data: <https://github.com/pubmedqa/pubmedqa>  Website: <https://pubmedqa.github.io/>  **7. TREC Precision Medicine / Clinical Decision Support Track**  Text REtrieval Conference (TREC) is running a track on Precision Medicine / Clinical Decision Support from 2014.  **2014 Clinical Decision Support Track**  Focus: Retrieval of biomedical articles relevant for answering generic clinical questions about medical records.  Information and Data: <http://www.trec-cds.org/2014.html>  **2015 Clinical Decision Support Track**  Focus: Retrieval of biomedical articles relevant for answering generic clinical questions about medical records.  Information and Data: <http://www.trec-cds.org/2015.html>  **2016 Clinical Decision Support Track**  Focus: Retrieval of biomedical articles relevant for answering generic clinical questions about medical records. Actual electronic health record (EHR) patient records are be used instead of synthetic cases.  Information and Data: <http://www.trec-cds.org/2016.html>  **2017 Clinical Decision Support Track**  Focus: Retrieve useful precision medicine-related information to clinicians treating cancer patients.  Information and Data: <http://www.trec-cds.org/2017.html> |

**RNA-seq workflow: gene-level exploratory analysis and differential expression**

<http://master.bioconductor.org/packages/release/workflows/vignettes/rnaseqGene/inst/doc/rnaseqGene.html>

**Genome Data Viewer**

<https://www.ncbi.nlm.nih.gov/genome/gdv/browser/genome/?id=GCF_000001405.39>

**Variation Viewer**

<https://www.ncbi.nlm.nih.gov/variation/view/>

**NCBI Sequence Viewer**

<https://www.ncbi.nlm.nih.gov/projects/sviewer/>

Example

**Human reference genomic region: NG\_000007**

<https://www.ncbi.nlm.nih.gov/projects/sviewer/?id=NG_000007>

|  |
| --- |
| **NCBI Reference Sequence Database (RefSeq)**  <https://www.ncbi.nlm.nih.gov/refseq/>  <https://ftp.ncbi.nlm.nih.gov/genomes/refseq/vertebrate_mammalian/Homo_sapiens/> |
|  |

|  |
| --- |
| **[5] Knowledge of Statistics, Machine Learning: Open Source Tools Applied to Bioinformatics**  **[5.1]** Browsing open access full-text publications on use of tools  **[5.2] Open Source Software Tools** |

|  |
| --- |
| **[K4]: Knowledge of Bioinformatics Tools (Bioconductor R Packages)** |
| **Most Common Tools Used for the Analysis of WGS Data**  <https://www.researchgate.net/figure/NGS-and-analysis-pipelines-Most-common-tools-used-for-the-analysis-of-WGS-data-QC_fig2_317413533> **Comprehensive Outline of Whole Exome Sequencing Data Analysis Tools** **Available in Clinical Oncology** <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6895801/>  **Comparative analysis of differential gene expression analysis tools for single-cell RNA sequencing data** 2019  <https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-019-2599-6>  **Most Common Tools Used for the Analysis of WGS Data**  <https://www.researchgate.net/figure/NGS-and-analysis-pipelines-Most-common-tools-used-for-the-analysis-of-WGS-data-QC_fig2_317413533> **Comprehensive Outline of Whole Exome Sequencing Data Analysis Tools Available in Clinical Oncology** <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6895801/>  **Comparitive Analysis of Differential Gene Expression Analysis Tools for Single-Cell Sequencing Data**  <https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-019-2599-6>  **FASTQC A quality control tool for high throughput sequence data**. 2014 September 29  <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>  **Free RNA-seq Analysis Tools – Software and Resources**  <https://bioinformaticshome.com/tools/rna-seq/rna-seq.html> |

|  |
| --- |
| **[K4]: Knowledge of Bioinformatics Tools (Bioconductor R Packages)** |
| **Most Common Tools Used for the Analysis of WGS Data**  <https://www.researchgate.net/figure/NGS-and-analysis-pipelines-Most-common-tools-used-for-the-analysis-of-WGS-data-QC_fig2_317413533> **Comprehensive Outline of Whole Exome Sequencing Data Analysis Tools** **Available in Clinical Oncology** <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6895801/>  **Comparative analysis of differential gene expression analysis tools for single-cell RNA sequencing data** 2019  <https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-019-2599-6>  **Most Common Tools Used for the Analysis of WGS Data**  <https://www.researchgate.net/figure/NGS-and-analysis-pipelines-Most-common-tools-used-for-the-analysis-of-WGS-data-QC_fig2_317413533> **Comprehensive Outline of Whole Exome Sequencing Data Analysis Tools Available in Clinical Oncology** <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6895801/>  **Comparitive Analysis of Differential Gene Expression Analysis Tools for Single-Cell Sequencing Data**  <https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-019-2599-6>  **FASTQC A quality control tool for high throughput sequence data**. 2014 September 29  <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>  **Free RNA-seq Analysis Tools – Software and Resources**  <https://bioinformaticshome.com/tools/rna-seq/rna-seq.html> |

|  |
| --- |
| **[K4]: Knowledge of Bioinformatics Tools (Bioconductor R Packages)**  **Analysis of single cell RNA-seq data (**University of Cambridge Bioinformatics)  <https://biocellgen-public.svi.edu.au/mig_2019_scrnaseq-workshop/public/index.html>  quality control, visualisation, data normalisation, exploratory data analysis, clustering, trajectory (pseudotime) inference, differential expression, batch correction, combining datasets, data integration, confounders, latent spaces, cell annotation, case studies |
| **2-day Course: (16 hours of video) RStudio**  Day 1: <https://www.youtube.com/watch?v=thHgPqQpkE4&feature=emb_err_woyt>  Processing Raw scRNA-Seq Data  Construction of Expression Matrix  Intro to R/Bioconductor  Seurat  Day 2: <https://www.youtube.com/watch?v=7dQ_pleDO2Y&feature=emb_err_woyt>  Clustering example  Feature Selection  Pseudotime Analysis  Differental Expression Analysis  DE Real Dataset  Comparing/Combining scRNA  Search scRNA-Seq Data  Seurat  scRNA-Seq Pipeline |

|  |
| --- |
| **[K4]: Knowledge of Bioinformatics Tools (Bioconductor R Packages)**  **Principal Components Analysis / Feature Selection**  **bioRxiv**  <https://www.biorxiv.org/>  <https://www.biorxiv.org/search> |
| **mixOmics: an R package for omics feature selection and multiple data integration Aug 2017**  <https://www.biorxiv.org/content/10.1101/108597v4.full>  **Differential Principal Components Reveal Patterns of Differentiation in Case/Control Studies Feb 2019**  <https://www.biorxiv.org/content/10.1101/545798v1.full>  **pathwayPCA: an R package for integrative pathway analysis with modern PCA methodology and gene selection April 2019**  <https://www.biorxiv.org/content/10.1101/615435v1.full>  **Accurate and Fast feature selection workflow for high-dimensional omics data June 2017**  <https://www.biorxiv.org/content/10.1101/144162v1.full> |

|  |
| --- |
| **[K4]: Knowledge of Bioinformatics Tools (Bioconductor R Packages)**  **SEURAT: R toolkit for single cell genomics**  <https://satijalab.org/seurat/v3.1/pbmc3k_tutorial.html> |
| Installation Instructions for Seurat  <https://satijalab.org/seurat/install.html>  Vignettes: Guided Analyses  <https://satijalab.org/seurat/vignettes.html>  **Seurat - Guided Clustering Tutorial** 2020 April  Setup the Seurat Object  Standard pre-processing workflow  Normalizing the data  Identification of highly variable features (feature selection)  Scaling the data  Perform linear dimensional reduction  Determine the ‘dimensionality’ of the dataset  Cluster the cells  Run non-linear dimensional reduction (UMAP/tSNE)  Finding differentially expressed features (cluster biomarkers)  Assigning cell type identity to clusters |

|  |
| --- |
| **[K3]: Knowledge of Inferential Statistics, Analytics, Machine Learning (Bioinformatics Tools)**  **Statistics for Genomics**  <https://www.youtube.com/playlist?list=PLdl4u5ZRDMQQpUcSDRKN3V2vvx3_SmMbr> |
| **17 Videos:** 2017 May  **Statistics for Genomcs: Distances and Clustering**  <https://www.youtube.com/watch?v=wQhVWUcXM0A&list=PLdl4u5ZRDMQQpUcSDRKN3V2vvx3_SmMbr&index=2>  **Statistics for Genomics Lab: Quick Introduction to R and Bioconductor**  <https://www.youtube.com/watch?v=J5h5WxOn3Gw&list=PLdl4u5ZRDMQQpUcSDRKN3V2vvx3_SmMbr&index=11>  **Statistics for Genomics Lab: Distances and Clustering RStudio**  <https://www.youtube.com/watch?v=PArRvqLUP6o&list=PLdl4u5ZRDMQQpUcSDRKN3V2vvx3_SmMbr&index=7>  **Statistics for Genomics: Introduction to RNAseq**  <https://www.youtube.com/watch?v=C8RNvWu7pAw&list=PLdl4u5ZRDMQQpUcSDRKN3V2vvx3_SmMbr&index=12>  **Statistics for Genomics: Advanced Differential Expression**  <https://www.youtube.com/watch?v=QINX3cI7qgk&list=PLdl4u5ZRDMQQpUcSDRKN3V2vvx3_SmMbr&index=15>  **Statistics for Genomics: Useful plots and bad plots**  <https://www.youtube.com/watch?v=46-t2jOYsyY&list=PLdl4u5ZRDMQQpUcSDRKN3V2vvx3_SmMbr&index=17> |

|  |
| --- |
| **[K4]: Knowledge of Bioinformatics Tools (Bioconductor R Packages)**  **Bioconductor** |
| **Bioconductor**  <http://bioconductor.org/>  **Bioconductor Courses & Conferences**  <http://bioconductor.org/help/course-materials/>  **BioC 2020 Conference** <http://bioc2020.bioconductor.org/schedule>  **BioC 2018 Workshops**  <https://bioconductor.github.io/BiocWorkshops/>  **BioC 2017 Workshops**  <http://bioconductor.org/help/course-materials/2017/BioC2017/>  **Bioconductor Support**  <https://support.bioconductor.org/>  <https://support.bioconductor.org/t/Tutorials/> |
| **Community Contributed Help Resources** <http://bioconductor.org/help/community/> **Videos:**  <https://www.youtube.com/user/bioconductor>  **R & Bioconductor Manual**  <http://manuals.bioinformatics.ucr.edu/home/R_BioCondManual>  Thomas Girke, UC Riverside  **Bioc-refcard**  <https://github.com/mikelove/bioc-refcard>  Mike Love |

|  |
| --- |
| **[K4]: Knowledge of Bioinformatics Tools (Bioconductor R Packages)**  **R/ Bioconductor  Principal Components Analysis / Feature Selection** |
| **mixOmics: an R package for omics feature selection and multiple data integration** 2017 Aug  <https://www.biorxiv.org/content/10.1101/108597v4.full>  **Differential Principal Components Reveal Patterns of Differentiation in Case/Control Studies** 2019 Feb  <https://www.biorxiv.org/content/10.1101/545798v1.full>  **pathwayPCA: an R package for integrative pathway analysis with modern PCA methodology and gene selection** 2019 April  <https://www.biorxiv.org/content/10.1101/615435v1.full>  **Accurate and Fast feature selection workflow for high-dimensional omics data** 2017 June  <https://www.biorxiv.org/content/10.1101/144162v1.full> |

|  |
| --- |
| **[K4]: Knowledge of Bioinformatics Tools (Bioconductor R Packages)**  **BioC 2020 Conference**  <http://bioc2020.bioconductor.org/schedule> |
| 100: **Annotating inter-sample DNA methylation and ATAC-seq variation with COCOA**  100: **Human RNA-seq data from recount2 and related packages**  100: **Introduction to Bioconductor annotation resources**  100: **A tidy transcriptomics introduction to RNA-Seq analyses**  200: **Functional enrichment analysis of high-throughput omics data**  200: **Best practices for ATAC-seq QC and data analysis**  200: **Copy number variation analysis with Bioconductor**  200: **Interactive visualization of SummarizedExperiment objects with iSEE**  200: **Integrated ChIP-seq data analysis workshop**  200: **An introduction to matrix factorization and principal component analysis in R**  500: **Bioconductor toolchain for usage and development of reproducible bioinformatics pipelines in CWL**  500: **Effectively Using the DelayedArray Framework to Support the Analysis of Large Datasets**  100: **Cloud-based genomics using Terra/AnVIL** |

|  |
| --- |
| **[K4]: Knowledge of Bioinformatics Tools (Bioconductor R Packages)** |
| |  | | --- | | **GenomicRanges**  <http://bioconductor.org/packages/release/bioc/html/GenomicRanges.html> | | **Software for Computing and Annotating Genomic Ranges**  <https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003118> | | **GenomicRanges HOWTOs**  April 27, 2020  <http://bioconductor.org/packages/release/bioc/vignettes/GenomicRanges/inst/doc/GenomicRangesHOWTOs.pdf> | | **A quick introduction to GRanges and GRangesList objects**  July 2015  <http://bioconductor.org/packages/release/bioc/vignettes/GenomicRanges/inst/doc/GRanges_and_GRangesList_slides.pdf> |  |  | | --- | | **GenomicAlignments**  <https://bioconductor.org/packages/release/bioc/html/GenomicAlignments.html> | | **Package GenomicAlignments** October 17, 2020  <https://bioconductor.org/packages/release/bioc/manuals/GenomicAlignments/man/GenomicAlignments.pdf>  Title: Representation and manipulation of short genomic alignments  **An Introduction to the GenomicAlignments Package** April 27, 2020  <https://bioconductor.org/packages/release/bioc/vignettes/GenomicAlignments/inst/doc/GenomicAlignmentsIntroduction.pdf> |  |  | | --- | | **IRanges**  <http://bioconductor.org/packages/release/bioc/html/IRanges.html> | | **An Overview of the IRanges package** May 21 2020  <http://bioconductor.org/packages/release/bioc/vignettes/IRanges/inst/doc/IRangesOverview.pdf>  **Package IRanges** October 17, 2020  <http://bioconductor.org/packages/release/bioc/manuals/IRanges/man/IRanges.pdf>  Foundation of integer range manipulation in Bioconductor | |

|  |
| --- |
| **[K4]: Knowledge of Bioinformatics Tools (Bioconductor R Packages)**  **Bioconductor: edgeR**  Empirical Analysis of Digital Gene Expression Data in R |
|  |
|  |

|  |
| --- |
| **[K4]: Knowledge of Bioinformatics Tools (Bioconductor R Packages)**  **Bioconductor: DESeq2** Differential gene expression analysis based on the negative binomial distribution |
| **Analyzing RNA-seq data with DESeq2** 2020 Oct  <https://bioconductor.org/packages/devel/bioc/vignettes/DESeq2/inst/doc/DESeq2.html>  **RNA-seq workflow: gene-level exploratory analysis and differential expression** 2019 Oct  <http://master.bioconductor.org/packages/release/workflows/vignettes/rnaseqGene/inst/doc/rnaseqGene.html>  **Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2** 2014 Dec  <https://genomebiology.biomedcentral.com/articles/10.1186/s13059-014-0550-8>  **The R package Rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads**  <https://academic.oup.com/nar/article/47/8/e47/5345150>  **Data preprocessing and creation of the data objects pasillaGenes and pasillaExons** 2020 May  <http://bioconductor.org/packages/release/data/experiment/vignettes/pasilla/inst/doc/create_objects.html> |
|  |

**Bioinformatics Workbook**

<https://bioinformaticsworkbook.org/list.html#gsc.tab=0>

|  |
| --- |
|  |

|  |
| --- |
| **GitHub Pages**  <https://pages.github.com/> |
| |  | | --- | | **Creating a GitHub Pages site**  <https://docs.github.com/en/free-pro-team@latest/github/working-with-github-pages/creating-a-github-pages-site> | | <https://docs.github.com/en/free-pro-team@latest/github/working-with-github-pages> |  |  | | --- | | **How to Host Your Website on GitHub Pages for Free**  Including custom domains, sub-domains, and https  <https://medium.com/swlh/how-to-host-your-website-on-github-pages-for-free-3302b0fe8956> | | <https://nbisweden.github.io/workshop-ngsintro/2001/slide_rnaseq.html> |  |  | | --- | | **William Kahley GitHub Pages Site** https://william-p-kahley.github.io/williamkahley.github.io/ <https://github.com/william-p-kahley/williamkahley.github.io> | | **GitHub Account**  william-p-kahley  **Repository**  Repository Settings Option to enable GitHub Pages  williamkahley.github.io  **Link to a Particular Web Page**  <https://william-p-kahley.github.io/williamkahley.github.io/William.P.Kahley.CovLet.111120.htm> | |  |  |  | | --- | | **How I Created GitHub Pages** | |  | |

|  |
| --- |
| **Tools: Genomics Browsers? Next Generation sequence data analysis using tools such as** |
| |  | | --- | | **IGV (Integrative Genomics Viewer)**  <http://broadinstitute.org/software/igv/>  **IGV-Web application**  <https://igv.org/app> | | **IGV User Guide**  <http://broadinstitute.org/software/igv/UserGuide>  **Integrative Genomics Viewer**  <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3346182/>  Video: <https://www.youtube.com/channel/UCb5W5WqauDOwubZHb-IA_rA>  Play All: <https://www.youtube.com/watch?v=sFeK25K5PE&list=PLSplvWwdPpSrhPn3V2iuPUzyxVIDYZ1xS> |  |  | | --- | | **Genome Analysis Toolkit (GATK)** Java framework  <https://gatk.broadinstitute.org/>  <https://gatk.broadinstitute.org/hc/en-us>  <https://www.broadinstitute.org/partnerships/education/broade/best-practices-variant-calling-gatk-1> A large Java library for variant analysis, discovery and genotyping,  powerful processing engine and high-performance computing features make it capable of taking on projects of any size. | | **GATK Best Practices Workflow for DNA-Seq**  <https://bioinformaticsworkbook.org/dataAnalysis/VariantCalling/gatk-dnaseq-best-practices-workflow.html#gsc.tab=0>  **GATK Getting Started**  <https://gatk.broadinstitute.org/hc/en-us/categories/360002302312>  Best Practices Workflows, Tutorials, Computing Platforms  **GATK Technical Documentation**  <https://gatk.broadinstitute.org/hc/en-us/categories/360002310591>  Troubleshooting, Glossary, Algorithms  **GATK Community Topics**  <https://gatk.broadinstitute.org/hc/en-us/community/topics>  Browse community discussions  **GATK / broadinstitute / gatk**  GitHub: <https://github.com/broadinstitute/gatk/releases>  Docker image: <https://hub.docker.com/r/broadinstitute/gatk/>  **Terra Support Quickstart New users overview**  Bioinformatics in the cloud on Terra  <https://support.terra.bio/hc/en-us/articles/360022714931-Bioinformatics-in-the-cloud-on-Terra> | |

**Human Genome Overview**

<https://www.ncbi.nlm.nih.gov/grc/human>

|  |
| --- |
| **[3]: Familiarity with Publicly Available Data Sources**  Explored Content, Search  **Open Genomic Data** |
| |  | | --- | | **UCSC Genome Bioinformatics Group**  <https://genome-euro.ucsc.edu/index.html> |  |  | | --- | | **NCBI Sequence Read Archive (SRA)**  <https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi>?  SRA is NIH primary archive of high-throughput sequencing data and is part of the international partnership of archives (INSDC). | | Stores raw sequence data from "next-generation" sequencing technologies including  Illumina, 454, IonTorrent, Complete Genomics, PacBio and OxfordNanopores.  In addition to raw sequence data, SRA now stores alignment information in the form of read placements on a reference sequence.  **SRA Toolkit**  <https://github.com/ncbi/sra-tools/wiki/02.-Installing-SRA-Toolkit>  The SRA Toolkit provides 64-bit binary installations for Linux distributions, Mac OS X, Windows.  **SRA Explorer**  tool aims to make datasets within the Sequence Read Archive more accessible.  <https://sra-explorer.info/> |  |  | | --- | | **ENCODE: Encyclopedia of DNA Elements** <https://www.encodeproject.org/> |  |  | | --- | | **1000 Genomes Project**  <http://www.1000genomes.org> |  |  | | --- | | **dbGaP**  **db Genotypes and Phenotypes**  <https://dbgap.ncbi.nlm.nih.gov/> |  |  | | --- | | **NCBI dbSNP database**  <https://www.ncbi.nlm.nih.gov/snp/> | | dbSNP contains human single nucleotide variations, microsatellites, and small-scale insertions and deletions along with publication, population frequency, molecular consequence, and genomic and RefSeq mapping information for both common variations and clinical mutations.  **dbSNP Data Access:**  <https://www.ncbi.nlm.nih.gov/snp/docs/RefSNP_about/#data-access>  RefSNP data, including genotype, frequency and associated metadata, are available without restrictions on the web, FTP, and API.  **dbSNP Overview:**  <https://www.ncbi.nlm.nih.gov/books/NBK21088/>  **dbSNP Tutorials on GitHub:**  <https://www.ncbi.nlm.nih.gov/snp/> |  |  | | --- | | **NIH Data Sharing Repositories/**  Open Domain-Specific Data Sharing Repositories  <https://www.nlm.nih.gov/NIHbmic/domain_specific_repositories.html> |  |  | | --- | | **GenBank**  <https://www.ncbi.nlm.nih.gov/genbank/>  GenBank is the NIH genetic sequence database, an annotated collection of all publicly available DNA sequences. | | GenBank is part of the **International Nucleotide Sequence Database Collaboration**, which comprises the DNA DataBank of  Japan (DDBJ), European Nucleotide Archive (ENA), GenBank at NCBI.  Ways to search and retrieve data from GenBank:  **-** Search GenBank for sequence identifiers and annotations with Entrez Nucleotide.  **-** Search and align GenBank sequences to a query sequence using BLAST (Basic Local Alignment Search Tool).  BLAST searches CoreNucleotide, dbEST, and dbGSS independently.  **-** Search, link, and download sequences programatically using NCBI e-utilities. |  |  | | --- | | **NCI / Genetic Data Commons (GDC)**  <https://gdc.cancer.gov/>  <https://portal.gdc.cancer.gov/> | | **GDC Data Transfer Tool: An Overview**  <https://docs.gdc.cancer.gov/Data_Transfer_Tool/Users_Guide/Getting_Started/>  Raw sequence data, stored as BAM files, make up the bulk of data stored at the NCI Genomic Data Commons (GDC).  **GDC Exploration**  <https://docs.gdc.cancer.gov/Data_Portal/Users_Guide/Exploration/>  The Exploration Page allows users to explore data in the GDC using advanced filters/facets, which includes those on a gene and mutation level. Users choose filters on specific Cases, Genes, and/or Mutations on the left of this page and then can visualize these results on the right.  **GDC Data Transfer Tool**  <https://docs.gdc.cancer.gov/Data_Transfer_Tool/Users_Guide/Preparing_for_Data_Download_and_Upload/>  is intended to be used in conjunction with the GDC Data Portal and the GDC Data Submission Portal to transfer data to or from the GDC. The GDC Data Portal's interface is used to generate a manifest file or obtain UUID(s) and (for Controlled-Access Data) an authentication token. The GDC Data Transfer Tool is then used to transfer the data files listed in the manifest file or identified by UUID(s). |  |  | | --- | | **GDC/TCGA - Genomic Data Commons (GDC) / The Cancer Genome Atlas (TCGA)**  <https://tcga-data.nci.nih.gov/tcga/>  Harmonized Cancer Datasets | | In order to download data from TCGA data portal:  1. Connect to <https://tcga-data.nci.nih.gov/tcga/>  2. Select the cancer subtype you are interested in (i.e breast invasive carcinoma)  3. Select mRNA  4. Now you can see a table where rows are representing different patients.  5. If present select the column (by clicking on header) that referse to RNASeq or RNASeqV2 if it is present for that  cancer subtype and then click BUILD archive.  6. Keep in mind that just below the header there is a number indicating the respective data level. Levels 1-4  <https://wiki.nci.nih.gov/display/TCGA/Data+level>  If you need RAW data such as FASTQ files you have find level 1 data, but often this kind of data is not publicly  available on TCGA and you might need to ask for permission in order to download it.  Marco has listed the steps to access open tier data on TCGA. In case you are interested in accessing lower level data,  such as raw bam files for rna seq samples, you can apply for the access here  <https://dbgap.ncbi.nlm.nih.gov/aa/wga.cgi?login=&page=login>  **GDC for TCGA Data Access Matrix Users**  <https://gdc.cancer.gov/gdc-tcga-data-access-matrix-users>  **Resources for TCGA Users**  <https://gdc.cancer.gov/resources-tcga-users>  **The Cancer Genome Atlas (TCGA)**  <https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga> |  |  | | --- | | **IGSR**: The International Genome Sample Resource  Supporting open human variation data  <https://www.internationalgenome.org/> |  |  | | --- | | **AnVIL**  supports the management, analysis and sharing of human disease data for the research community  <https://anvilproject.org/>  <https://gen3.theanvil.io/>  <https://gen3.theanvil.io/explorer> |  |  | | --- | | **REST APIs** <http://MyGene.info>  <http://MyVariant.info>  <http://BioThings.io>  <http://data.cvisb.org/home>  <http://Smart-API.info> | |

|  |
| --- |
| **Data Preparation and Transformation**  This deals with preprocessing raw data to convert it into a form that is  ready for analysis or model building and includes topics such as |
| a) handling missing data  b) data imputation  c) encoding categorical data  d) identification of predictor features and target features  e) data scaling (e.g., feature standardization, normalization)  f) feature selection, dimensionality reduction  g) advanced methods for data transformation (e.g., PCA, LDA)  Software that can be used for data preparation and transformation include  Python  Pandas package  R  Excel |
|  |
| **Data Exploration:**  Identifying patterns  - Inferential Statistical Analysis (T-Test, Kurtosis, Outliers, Linear Regression, …)  **Presenting Results:**  Independent, Related  Hypothesis  Ho = (T-value, P-value)  Alternate  Jupyter Notebook |
|  |
|  |
|  |
|  |
|  |
|  |
|  |

|  |
| --- |
| **Google Domain Name** |
| A domain name through a domain name registrar.  You can register a domain name through Google Domains or another domain registrar of your choice.  An IP address to point the A record of your zone. |

|  |
| --- |
| **Google Cloud Platform** |
| **Google Kubernetes Engine**  <https://cloud.google.com/kubernetes-engine/>  **Types of clusters**  <https://cloud.google.com/kubernetes-engine/docs/concepts/types-of-clusters> |

Start AWS Computational Genomics Image

1,686 views

Aug 25, 2018

<https://www.youtube.com/watch?v=AiSvnLamE5Y>

**AWS AMI bioinformatics**

Docker Hub

<https://hub.docker.com/_/microsoft-bioconductor-bioconductor-docker>

**Docker containers for Bioconductor**

<https://www.bioconductor.org/help/docker/>

Docker packages software into self-contained environments, called containers, that include necessary dependencies to run.

Containers can run on any operating system including Windows and Mac (using modern Linux kernels) via the Docker engine.

Current Containers

<https://www.bioconductor.org/help/docker/#current>

For each supported version of Bioconductor, we provide

bioconductor/bioconductor\_docker:RELEASE\_X\_Y

bioconductor/bioconductor\_docker:devel

Bioconductor’s Docker images are stored in Docker Hub; the source Dockerfile(s) are in Github.

**Bioconductor in the cloud**

<https://www.bioconductor.org/help/bioconductor-cloud-ami/>

Bioconductor software packages

<http://www.bioconductor.org/packages/stats/index.html>

Nov 2020

Bioconductor/AWSParallel

<https://rdrr.io/github/Bioconductor/AWSParallel/f/vignettes/AWSParallel-AWSBatchJobsParam-tutorial.Rmd>

Introduction to BiocParallel

1Martin.Morgan@RoswellPark.org

February 10, 2019

<https://bioconductor.statistik.tu-dortmund.de/packages/3.8/bioc/vignettes/BiocParallel/inst/doc/Introduction_To_BiocParallel.pdf>

**StarCluster**

<http://star.mit.edu/cluster/>

StarCluster is an open source cluster-computing toolkit for Amazon’s Elastic Compute Cloud (EC2) released under the LGPL license.

StarCluster has been designed to automate and simplify the process of building, configuring, and managing clusters of virtual machines on Amazon’s EC2 cloud.

StarCluster allows anyone to easily create a cluster computing environment in the cloud suited for distributed and parallel computing applications and systems.

Scroll down to Quick-Start Screencast

Quick-Start

<http://star.mit.edu/cluster/docs/latest/quickstart.html>

Installing StarCluster

<http://star.mit.edu/cluster/docs/latest/installation.html>

Starcluster on Ubuntu 14.04 - Quick Start - AMI Cookbook Recipe - Ubuntu 14

<https://www.youtube.com/watch?v=2RBupgpi_ec>

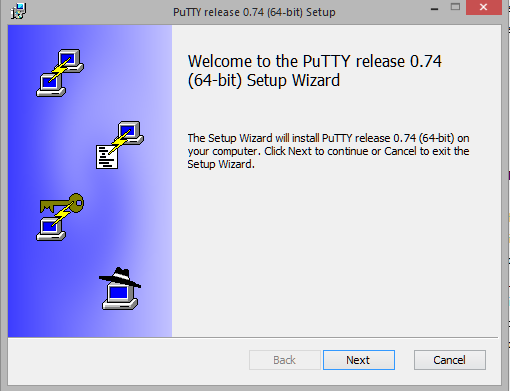
2014

|  |
| --- |
| **How to launch a preconfigured Bioconductor AMI session using the free tier with RStudio Server?**  <https://support.bioconductor.org/p/85061/> |
| Hello Bioconductor community,  I am learning how to a Bioconductor AMI using the steps in the link <https://www.bioconductor.org/help/bioconductor-cloud-ami/> under “Launching the AMI”.  I set up an Amazon Web Services account and launched an AMI using the recommended ID ([ami-64d43409](https://console.aws.amazon.com/ec2/home?region=us-east-1#launchAmi=ami-64d43409)).  However, I am encountering a problem in trying to configure the AMI to use the free tier.  Here are the details of the steps I followed to launch the AMI: From the AWS console, I first click "launch instance".  Then I click the "Community AMI" tab and paste in the AMI ID [ami-64d43409](https://console.aws.amazon.com/ec2/home?region=us-east-1#launchAmi=ami-64d43409) recommended on the Bioconductor AMI help site.  After clicking 'Select' on the pre-configured Bioconductor AMI, I accept the defaults for instance type (t2.micro, free tier) by clicking "Review and Launch".  However, I get a warning message pop up that says: "Your instance configuration is not eligible for the free usage tier".  I think this may be due to the fact that the storage is preconfigured for 100GB and the free tier is only eligible for 30GB.  When I try to change the storage amount by clicking "Edit Storage" on the "Review and Launch" page (and enter the maximum free tier amount of 30GB), I get the following error message after clicking "launch": "Launch Failed: Volume of size 30GB is smaller than snapshot 'snap-521fc0a9', expect size >= 100GB."  I am able to get the AMI up and running if I leave the storage at the default 100GB, but this will incur charges since it is not compatible with the free tier restrictions.  Is there a way to modify the preconfigured AMIs with a smaller storage so that the free tier can be utilized?  In addition, once the AMI is running, I would like to launch RStudio Server in a browser.  The Bioconductor AMI help site says this is possible, but I'm not sure where to find the username and password login info from my running instance details or how to login with those credentials.  Thanks in advance for your help.  Best,  Keegan Korthauer |
| Hi Keegan,  We pre-loaded the AMIs with a large number of Bioconductor software and annotation packages thinking this would be a convenience, i.e., users avoid spending time downloading packages themselves. Having a minimal configuration with a handful of Bioconductor packages that meets the free tier criteria sounds like a good idea and would be useful for those testing the waters. We'll look into adding a minimal AMI, it may be a week or so before this gets done.  Lori is looking into the RStudio credentials. We could send them to you off-line but would rather post them to the web site if this doesn't compromise security. Once we've sorted that out we'll let you know how to get the credentials.  Thanks for raising these issues.  Valerie |
| **Question: Best Amazon Machine Image (AMI) for bioinformatics**  <https://www.biostars.org/p/353590/>  AMI Bioconductor  Our AMIs have the following IDs.  **Bioconductor Version R Version AMI ID**  3.13 (devel) 4.1.0 ami-0e7efd11a6eab85a6  3.12 (release, recommended) 4.0.3 ami-04c69d122c1cf7e81  <http://www.bioconductor.org/help/bioconductor-cloud-ami/#connecting_rstudio> |

**PuTTy**

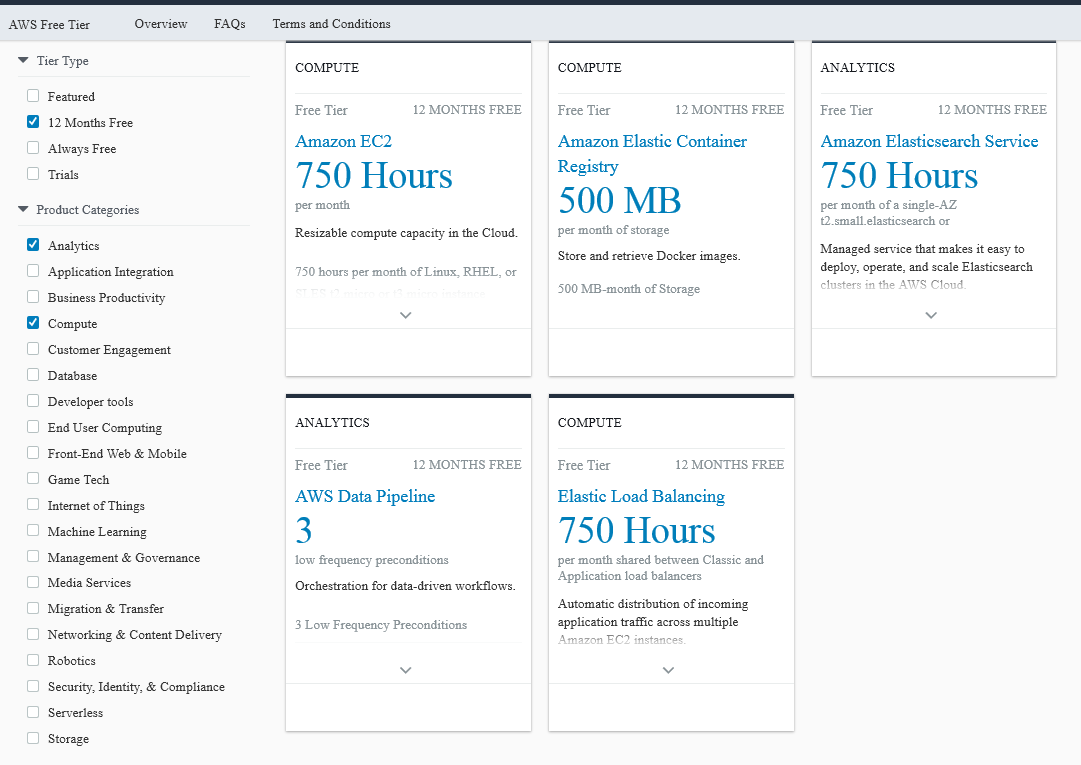
<https://www.chiark.greenend.org.uk/~sgtatham/putty/latest.html>

putty-64bit-0.74-installer.msi



**C:\Program Files\PuTTY\ putty.exe**

**puttygen.exe**



|  |
| --- |
| **Amazon EC2**  Amazon Elastic Compute Cloud  Secure and resizable compute capacity to support virtually any workload  <https://aws.amazon.com/ec2/>  Is a web service that provides secure, resizable compute capacity in the cloud. It is designed to make web-scale cloud computing easier for developers. Amazon EC2’s simple web service interface allows you to obtain and configure capacity with minimal friction. It provides you with complete control of your computing resources and lets you run on Amazon’s proven computing environment.  Amazon EC2 offers the broadest and deepest compute platform with choice of processor, storage, networking, operating system, and purchase model. We offer the fastest processors in the cloud and we are the only cloud with 400 Gbps ethernet networking. |
| **AWS Free Tier**  <https://aws.amazon.com/free/?all-free-tier.sort-by=item.additionalFields.SortRank&all-free-tier.sort-order=asc&awsf.Free%20Tier%20Types=tier%2312monthsfree&awsf.Free%20Tier%20Categories=categories%23compute%7Ccategories%23analytics>  Gain free, hands-on experience with the AWS platform, products, and services  Explore more than 85 products and start building on AWS using the free tier. Three different types of free offers are available depending on the product used. See below for details on each product. |
| **AWS Pricing Calculator**  <https://calculator.aws/#/addService> |
| **Amazon Elastic Block Store (EBS)**  <https://aws.amazon.com/ebs/>  allows you to create persistent block storage volumes and attach them to Amazon EC2 instances.  Amazon Elastic Block Store (EBS) is an easy to use, high performance block storage service designed for use with Amazon Elastic Compute Cloud (EC2) for both throughput and transaction intensive workloads at any scale. A broad range of workloads, such as relational  and non-relational databases, enterprise applications, containerized applications, big data analytics engines, file systems, and media workflows are widely deployed on Amazon EBS.  You can choose from six different volume types to balance optimal price and performance. You can achieve single digit-millisecond latency for high performance database workloads such as SAP HANA or gigabyte per second throughput for large, sequential workloads such  as Hadoop. You can change volume types, tune performance, or increase volume size without disrupting your critical applications, so you have cost-effective storage when you need it.  Designed for mission-critical systems, EBS volumes are replicated within an Availability Zone (AZ) and can easily scale to petabytes of data. Also, you can use EBS Snapshots with automated lifecycle policies to back up your volumes in Amazon S3, while ensuring  geographic protection of your data and business continuity. |
| **AWS Associate Solutions Architect Notes**<https://github.com/Apjo/AWSAssociateSolutionsArchitectNotes/blob/master/README.md>Notes prepared while doing the online course on AWS Associate Solutions Architect.These notes are a mixture of notes from a course taught by Ryan Kroonenburg, and one taught by Eissa**EC2 (Elastic Compute Cloud)**- Provides resizable compute capacity in the cloud.- You have root access to the EC2 instances, be able to restart, terminate, reboot.- You need to have a key & key pair to access the instance- Reduces the time required to obtain and boot new server instances to minutes,allowing you to quickly scale capacity, both up and down, as computing requirements change.- 2 types of block store devices namely Elastic Block store which are persistent and Network attached virtual drives,these are not directly connected to the host where the instance is but are attached to the network,where as Instance-store are not persistent(ephemeral), basically a virtual hard drive on the host allocated to this EC2 instanceandEBS-backed EC2 instance has a EBS root volume, and Instance-store backed EC2 instance has instance-store root volume**7: Analytics:****ElasticMapReduce(EMR):**A web service that makes it easy to quickly and cost-effectively process vast amounts of data.Uses Hadoop, an open source framework to distribute your data and process across a resizeable cluster of Amazon EC2 instances.It can also run other distributed framework such as Spark and Presto.EMR is used in a variety of applications including log analysis, web-indexing, data warehousing, machine learning, financial analysis, scientific simulation, and Bioinformatics, customers launch millions of Amazon’s EMR clusters each yearAllows you root access(i.e. login via SSH)**Cloud Search / Elastic Search:**Search engine for your website or your application,Cloud search is fully managed service provided by AWS, Elastic search uses open source framework**Data Pipeline:** Allows to move data from one place to another**Quick Sight:** used for creating visualizations, dashboards for BI/Analytics |
| **Introducing new Amazon EBS general purpose volumes, gp3** Posted On: Dec 1, 2020 Today AWS announced the availability of gp3, the next-generation general purpose SSD volumes for Amazon Elastic Block Store (Amazon EBS) that enable customers to  provision performance independent of storage capacity and provides up to 20% lower price-point per GB than existing gp2 volumes.  With gp3 volumes, customers can scale IOPS (input/output operations per second) and throughput without needing to provision additional block storage capacity, and pay only for the resources they need.  General purpose SSD volumes make it easy and cost effective for customers to meet the IOPS and throughput requirements for transaction-intensive workloads, such as virtual desktops, test and development environments, low-latency interactive applications, and boot volumes. With existing general-purpose SSD (gp2) volumes, performance is tied to storage capacity, enabling customers to get higher IOPS and throughput for their applications by provisioning a larger storage volume size. But customers want to scale performance and throughput without paying for storage that they don’t need.  Next generation gp3 volumes offer the ability to independently provision IOPS and throughput, separate from storage capacity. This enables customers to scale performance for transaction-intensive workloads without needing to provision more capacity, so they only pay for the resources they need. The new gp3 volumes also deliver a baseline performance of 3,000 IOPS and 125MB/s at any volume size. For use cases, where your application needs more performance than the baseline, customers can scale up to 16,000 IOPS and 1,000 MB/s for an additional fee. This makes the new gp3 volumes ideal for a wide variety of applications that require high performance at low cost, including MySQL, Cassandra, virtual desktops, and Hadoop analytics clusters.  Customers can easily migrate gp2 volumes to gp3 volumes using Elastic Volumes, which is an existing feature of Amazon EBS.  Elastic Volumes allow customers to modify the volume type, IOPS, or throughput of their existing EBS volumes without interrupting their Amazon EC2 instances. gp3 volumes are available in all AWS commercial and gov cloud regions. For more information, please see the gp3 announcement on the AWS News blog, and documentation. |
| **Volume Gateway**  Hybrid cloud block storage with local caching  <https://aws.amazon.com/storagegateway/volume/>  Volume Gateway presents cloud-backed iSCSI block storage volumes to your on-premises applications. Volume Gateway stores and manages on-premises data in Amazon S3 on your behalf and operates in either cache mode or stored mode. In the cached Volume Gateway mode, your primary data is stored in Amazon S3, while retaining your frequently accessed data locally in the cache for low latency access. In the stored Volume Gateway mode, your primary data is stored locally and your entire dataset is available for low latency access on  premises while also asynchronously getting backed up to Amazon S3. In either mode, you can take point-in-time copies of your volumes using AWS Backup, which are stored in AWS as Amazon EBS snapshots. Using Amazon EBS Snapshots enables you to make space-  efficient versioned copies of your volumes for data protection, recovery, migration, and various other copy data needs.  **Amazon Simple Storage Service (S3)**  Object storage built to store and retrieve any amount of data from anywhere  <https://aws.amazon.com/s3/>  (Amazon S3) is storage for the internet. You can use Amazon S3 to store and retrieve any amount of data at any time, from anywhere on the web.  **Amazon ElastiCache**  offers fully managed Redis and Memcached. Seamlessly deploy, run, and scale popular open source compatible in-memory data stores. Build data-intensive apps or improve the performance of your existing apps by retrieving data from high throughput and low latency in-memory data stores.  **AWS Storage Gateway**  <https://aws.amazon.com/storagegateway/>  AWS Storage Gateway is a hybrid storage service that enables your on-premises applications to seamlessly use AWS cloud storage. You can use the service for backup and archiving, disaster recovery, cloud data processing, storage tiering, and migration.  **Amazon Elastic Container Registry**  (ECR) is a fully-managed Docker container registry that makes it easy for developers to store, manage, and deploy Docker container images. Amazon ECR eliminates the need to operate your own container repositories or worry about scaling the underlying infrastructure.  With Amazon ECR, there are no upfront fees or commitments. You pay only for the amount of data you store in your repositories and data transferred to the Internet.  **Amazon RDS for PostgreSQL**  Amazon RDS makes it easy to set up, operate, and scale PostgreSQL deployments in the cloud. With Amazon RDS, you can deploy scalable PostgreSQL deployments in minutes with cost-efficient and resizable hardware capacity. |
|  |

|  |
| --- |
| **Using Amazon EC2 to run large data analyses cheaply Spot instances on Amazon Elastic Compute Cloud (EC2)**  Jul 5, 2018  <https://gex.netlify.app/post/using-amazon-ec2-to-run-large-data-analysis-cheaply/> |
| **Spot instances on Amazon Elastic Compute Cloud (EC2)** allow researchers to do high performance computing at very low cost.  For example, a 64-core workstation with 256 GB of memory can be rented at about $0.7 per hour  Main topics covered this tutorial  Set up Key Pairs and Security Groups to enable SHH login,  Request spot instance (virtual machine),  SHH access to the instance via Putty and Filezilla,  Install Docker software,  Build a Docker image based on the Bioconductor Docker definition files,  Start R and compute from within the container,  Create a “volume” (a virtual hard drive) and attach it to running instances,  Take a snapshot of a volume, copy it across regions, and use it to create a volume,  Google Compute Engine set-up. |
| Start the Spot Advisor and enter your request.  **Change the amount required to 1. Otherwise, 20 instances will be requested**.  A 64-core instance with 256GB memory costs $0.67 per hour or about $16 per day.  Note that Amazon has web servers all across the world. You can switch regions from the top right of the screen.  Prices and availability of resources vary greatly across the region. |

## **What is Packer?**

<https://www.packer.io/>

Packer is an open source tool for creating identical machine images for multiple platforms from a single source configuration.   
Packer is lightweight, runs on every major operating system, and is highly performant, creating machine images for multiple platforms in parallel.   
Packer does not replace configuration management like Chef or Puppet. In fact, when building images, Packer is able to use tools like Chef or Puppet to install software onto the image.

A machine image is a single static unit that contains a pre-configured operating system and installed software which is used to quickly create new running machines.   
Machine image formats change for each platform.   
Some examples include [**AMIs**](https://en.wikipedia.org/wiki/Amazon_Machine_Image) **for EC2**, VMDK/VMX files for VMware, OVF exports for VirtualBox, etc.

# **New AWS public datasets available from the National Cancer Institute, Massachusetts Institute of Technology, Amazon, the National Renewable Energy Laboratory, and others**

Posted On: Jul 15, 2020

<https://aws.amazon.com/about-aws/whats-new/2020/07/new-aws-public-datasets/>

Twenty-three new or updated Amazon Web Services (AWS) public datasets from the National Center for Bioinformatics, Johns Hopkins University, University of Texas at Southwestern, National Oceanic and Atmospheric Administration (NOAA), the National Cancer Institute, National Herbarium of New South Wales, and others are now available in the following categories:

COVID-19 response:

* [COVID-19 Molecular Structure and Therapeutics Hub](https://registry.opendata.aws/molssi-covid19-hub/) from the Molecular Sciences Software Institute
* [COVID-19 Genome Sequence Dataset](https://registry.opendata.aws/ncbi-covid-19) from the National Center for Biotechnology Information

Life sciences:

* [Cloud Genomic Indexes](https://registry.opendata.aws/jhu-indexes/) from Johns Hopkins University and the University of Texas at Southwestern
* [Refgenie Genomic Assets](https://registry.opendata.aws/refgenie/) from University of Virginia
* [Gabriella Miller Kids First Pediatric Research Program](https://registry.opendata.aws/kids-first/) from the National Cancer Institute
* [The Cancer Genome Atlas](https://registry.opendata.aws/tcga/) from the National Cancer Institute
* [Basic Local Alignment Sequence Tool (BLAST) Databases](https://registry.opendata.aws/ncbi-blast-databases/) from the National Library of Medicine
* [National Herbarium of New South Wales](https://registry.opendata.aws/nsw-herbarium/) from the Royal Botanic Gardens and Domain Trust

# **Cosine Similarity support in Amazon Elasticsearch Service**

Posted On: Jul 29, 2020

Amazon Elasticsearch Service now supports cosine similarity distance metric with k-Nearest Neighbor (k-NN) to power your similarity search engine. Cosine similarity is used to measure similarities between two vectors, irrespective of their sizes and is most commonly used in information retrieval, image recognition, text similarity, bioinformatics and recommendation systems.

### Reproducible and robust workshop materials

Workshops were authored using R Markdown, and compiled into a book (PDF and ePub) and website using Bookdown R package. Bookdown, in turn, uses the gitbook publishing system ( <https://www.gitbook.com/>) to produce a variety of formats from the same source material. R Markdown files intended to be part of a Bookdown project do not contain the required front matter of a typical stand-alone R Markdown document. To help authors use and test the correct format, we seeded each workshop document with the syllabus that had been submitted by that author, and successfully built the book of the submitted syllabi. Each workshop represented a chapter of a book compiled using the Bookdown software. This approach provided several advantages:

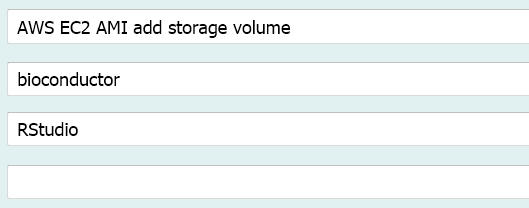
* R markdown syntax is already familiar to any developer of a Bioconductor package, since it is the standard approach to creating the package “vignette” or prose documentation.
* R markdown implements “literate programming” by including formatted text, runnable code, and output of the code
* Bookdown allows collating chapters as a clean, lightweight online book format, and pandoc additionally allows creation of PDF and ePub formats
* These formats can then be self-published with options to order paper copies through companies such as <https://leanpub.com>

This approach allowed automatic installation of required packages by listing them in the DESCRIPTION file required by R packages.

**bookdown: Authoring Books and Technical Documents with R Markdown**

<https://books.google.com/books?hl=en&lr=&id=_LrZDQAAQBAJ&oi=fnd&pg=PT12&dq=+Xie+Y+:+bookdown:+Authoring+Books+and+Technical+Documents+with+R+Markdown+%5BInternet%5D+.+Boca+Raton,+Florida:+Chapman+and+Hall/CRC%3B+2016++10.1201/9781315204963++&ots=tz1AhTAN96&sig=mmTuwDYYrl3sVm8SjtJD8tvnwd0#v=onepage&q&f=false>

AWS\_EC2\_AMI



Must include Bioconductor

32,500 results

<https://www.google.com/search?lr=&as_qdr=all&sxsrf=ALeKk01PJquBbFitcuwSWiEVFhAJCkGTzA:1606918695752&q=AWS+EC2+AMI+add+storage+volume+cost++RStudio+%22%22bioconductor+%22%22&sa=X&ved=2ahUKEwjXyavQvq_tAhXiUjUKHZMcB_4Q5t4CMAJ6BAgBEAo&biw=1146&bih=722>

<https://www.louisaslett.com/RStudio_AMI/>

|  |
| --- |
| **RStudio Server Amazon Machine Image (AMI)**  AMI specifically targeted at R and RStudio Server  <https://www.louisaslett.com/RStudio_AMI/> |
| To use the AMIs described on this page, you simply click your chosen AMI ID  which will take you through to the Amazon web interface and preselect the correct region and AMI.  Simply ensure that your ‘security group’ settings allow incoming HTTP (port 80) traffic and then  copy-and-paste the ‘Public DNS’ for your running instance to a web browser address bar to bring up the login page.  <https://www.louisaslett.com/RStudio_AMI/video_guide.html>  **US East, Virginia ami-0f2290fdad793f863**  <https://signin.aws.amazon.com/signin?redirect_uri=https%3A%2F%2Fconsole.aws.amazon.com%2Fec2%2Fhome%3Fregion%3Dus-east-1%26state%3DhashArgs%2523launchAmi%253Dami-0f2290fdad793f863%26isauthcode%3Dtrue&client_id=arn%3Aaws%3Aiam%3A%3A015428540659%3Auser%2Fec2&forceMobileApp=0&code_challenge=bEkiXYttLxSfqW-GZo7rCoYMLO0XpaFOtXitDHZkEwI&code_challenge_method=SHA-256> |
| **Features include:**   * 30GB EBS storage — compact, but enables storage of more sizeable datasets.   + Defaults to fast SSD storage (faster, zero IO costs, only $1 per month in most regions) * Easy to use Dropbox integration to up/down-load files and data.   + Setup can be completed entirely through RStudio in the web browser by running a single function.   + Selective syncing supported so that large Dropbox accounts don’t sync everything.   + Unlink and relink to new account supported. * Full LaTeX support enabling [R Markdown](http://rmarkdown.rstudio.com/), Sweave and regular document compiles within RStudio. * Java 8 JRE enabling full support for [H2O](https://www.h2o.ai/products/h2o/) and [Spark](https://spark.rstudio.com/). * GDAL dependencies for GIS packages. * [GSL](http://www.gnu.org/software/gsl/) and CURL libraries. * Database support:   + ODBC drivers installed.   + RMySQL package precompiled and installed. * Git and Subversion support out of the box. * MCMC samplers:   + [Stan](http://mc-stan.org/rstan.html) (RStan) installed and ready to use for Hamiltonian Monte Carlo sampling.   + [JAGS](http://mcmc-jags.sourceforge.net) (and rjags) installed and ready to use for Gibbs sampling.   + [Greta](https://greta-stats.org) supported for GPU MCMC sampling via Tensorflow support. * CUDA and cuDNN.   + Enables immediate use of GPU instances (e.g. p2.\* instances) without any setup.   + [Tensorflow](https://www.tensorflow.org/) and [Keras](https://keras.io/) deep learning libraries can be accelerated using nVidia GPUs on the GPU compute instances, just 3 lines of R code each to setup:   + # Tensorflow   + install.packages("tensorflow")   + library("tensorflow")   + install\_tensorflow(version = "gpu")   + # Keras   + install.packages("keras")   + library("keras")   + install\_keras(tensorflow = "gpu")   You will then be prompted to install Miniconda: say "Yes" to this option.  See the [RStudio Keras page](https://keras.rstudio.com/) and [RStudio Tensorflow page](https://tensorflow.rstudio.com/) for details on using from R.   * + Also preinstalled Magma GPU linear algebra libraries for accelerated matrix decompositions. * Swap space for compiling of large packages on constrained memory instances (such as rugarch). * Arbitrary precision arithmetic and number theory libraries supported out of the box:   + [GMP](http://gmplib.org/)   + [MPFR](http://www.mpfr.org/)   + [FLINT](http://flintlib.org/) * Optimised BLAS for automatically faster matrix operations than base R libraries ([OpenBLAS](http://www.openblas.net/)). |
|  |

|  |
| --- |
| **AWS** |
| What type of data transfer is free for Amazon s3?  Pick the right AWS region for your S3 bucket.  The main benefit of having S3 and EC2 in the same region is the performance and lower transfer cost.   Data transfer is free  between EC2 and S3 in the same region.  Downloading file  from another AWS region will cost $0.02/GB.  May 18, 2019 |
| Is t3 Micro free tier?  t3. micro is supported under AWS Free Tier.  For example, you can use  - 1 Linux instance continuously for a month, or  - 10 Linux instances for 75 hours a month.  In some cases, leaving your resources running maximizes your AWS Free Tier benefits. |
| **Access to Genomic Data for free**  In **S3**: <http://s3.amazonaws.com/1000genomes>  **IGSR: The International Genome Sample Resource**  Open human variation data  <https://www.internationalgenome.org/home>  **1000 Genomes Project FAQ**  If you have any other questions you can’t find the answer to please email [info@1000genomes.org](mailto:info@1000genomes.org)  <https://www.internationalgenome.org/faq> |
| Google:  AWS free  Docker for Bioconductor |
|  |

|  |
| --- |
| **DevOps Pipeline Experience** |
| NGC Virtual envir intended for DevOps Pipeline  GitLab, Docker Nexus, Jenkins --> Deployment Env  Git Local --> GitLab Remote -->  Docker Local --> Docker Hub --> Docker Hub Nexus  Dockerfile YAML  Jenkins  Jenkinsfile YAML |