Dear Hiring Managers, Others,

I am very much interested in being considered for the position; and other similar roles, with special interest in   
Genome sequencing analysis [Computational Genomics, Statistical Genomics (Statistics, Machine Learning), Bioinformatics],

and

Genome sequencing pre-processing

Also,

When reviewing my resume, you may find that I do not have all of the experience that you are looking for in a candidate. However, for the areas of experience not found satisfied in the content of my resume, it may perhaps be reasonable to allow consideration of the following.   
  
Online courses:

Amazon Web Services

• AWS Technical Professional 2020/04

<https://www.aws.training/Details/Curriculum?id=45423>

Skillport

• Python for Data Science: Introduction to Pandas 2020/05

• Data Science Statistics: Applied Inferential Statistics 2020/05

Udemy

• Machine Learning A-Z: Hands-On Python & R in Data Science 44 hour In-Progress

<https://www.udemy.com/course/machinelearning/>

Software Development Env: install, setup, configure, use of

Python  
 • Python scripting

• Python Libraries: NumPy, SciPy, Pandas, SciKit-Learn, StatsModels, MatPlotLib

• IDE: Spyder, PyCharm

• JupyterLab, Jupyter Notebook

• Anaconda Navigator

R

• R scripting

• R Packages: Bioconductor

• IDE: RStudio

• BiocManager  
  
Applied Maths:   
 Inferential Statistics, Analytics, Machine Learning

Gained the following by diciplined self-study, exploration; 30 hrs / week X 20 weeks (as of 10/04/2020)

**Familiarity with**

**[F1]:** Concepts/terminology of Molecular Biology, Genetics

**[F2]:** Genome sequencing technologies

**[F3]:** Open Access Journals: (BioMed Central (BMC), PubMed Central (PMC), Public Library of Science (PLOS), bioRxiv)   
 Identify current Best Practice bioinformatic tools, Standard Workflows

**[F3]:** Data Engineering (Docker images, AWS AMI )

- Amazon Web Services: Free-Tier EC2 account/configuration (Linux Amazon Machine Image (AMI), T2/T3 Instances, SSH)

Genomic Data Commons which AWS has made freely available

- Google Cloud Platform: Nginx server, Static Web content (pages of notes, outlines, hyperlinks, MP4 video files)

- GitData: Genomic data (Ancestry DNA, )

- GitHub: Bioinformatics projects

**Knowledge of**

By browsing, reading open access journal publications in order to identify; current Best Practice bioinformatic tools, Standard Workflows, - - -  
(BioMed Central (BMC), PubMed Central (PMC), Public Library of Science (PLOS), bioRxiv)

**[K1]:** Inferential Statistics, Machine Learning (Applied to Computational Biology)

**[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>

Tools: R/RStudio, Bioconductor Packages, (Use Cases: Bioconductor Package Vignettes, )

Tools: Python, BioPython

**[K3]:** Publicly Available Genomic Data Resources, Viewers (navigation, content interpretation, search, data accession, )

**NCBI databases** <https://www.ncbi.nlm.nih.gov/search/>

**NCBI Genome Data Viewer (GDV)**

<https://www.ncbi.nlm.nih.gov/genome/gdv/>

**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

**NCBI Sequence Read Archive (SRA)**

<https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi>?

**Special effort**

Bioconductor Packages associated with following

**[S1]**: RNA-Seq (related data files formats, pre-processing) FASTQ BAM VCF

**[S2]**: Differential Expression analysis [ Expression Analysis, Differential Expression Analysis ]

**[S3]**:

https://github.com/william-p-kahley/williamkahley.github.io.git

**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>

**william-p-kahley / williamkahley.github.io**  
<https://github.com/william-p-kahley/williamkahley.github.io>

<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>

**GitHub Projects**

<https://github.com/hbctraining/Intro-to-R-with-DGE>

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

<https://github.com/hemberg-lab/scRNA.seq.course>

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)**

Gained overview/survey understanding of bioinformatics concepts, terminology [n

Greater effort given to statistical principles and machine learning algorithms [n], and in particular the implementation of principles [n](R / RStudio, Bioconductor)

With particular attention to identifying standard workflow tools [1: Publications], [2 \_\_\_\_]

and using R scripting in RStudio

NGS Genomics Workflows:

Whole Genome Sequencing Data Analysis Tools

Whole Exome Sequencing Data Analysis Tools

scRNA-Seq

RNA-Seq

|  |
| --- |
| **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>  **scikit-learn-videos**  <https://github.com/justmarkham/scikit-learn-videos>  <https://www.youtube.com/playlist?list=PL5-da3qGB5ICeMbQuqbbCOQWcS6OYBr5A>  **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**  <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>  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> |

|  |
| --- |
| **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 R **0:**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> |

|  |
| --- |
| **Google:**  **NCBI RStudio Bioconductor** |
| **geoquery**: a bridge between the gene expression omnibus (GEO) and bioconductor <https://www.bioconductor.org/packages/release/bioc/html/GEOquery.html>  <https://www.bioconductor.org/packages/release/bioc/vignettes/GEOquery/inst/doc/GEOquery.html#platforms>  Davis S, Meltzer P (2007). “GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor.” Bioinformatics, **14**, 1846–1847. |
|  |
|  |

+++++++

RNA-Seq Workflows

Alignment Based

Alignment Free

|  |
| --- |
| - Knowledge of pipeline frameworks ()  - Knowledge of HPC/Parallel/Grid-Computing ()  **Public accessable projects as examples demonstrating my applying the above:**  Knowledge of Omics concepts/terminology, Bioinformatics,  **GitHub**    **GitData**  AncestryDNA Data upload to GitData  - Knowledge of Statistics, Machine Learning:  Longitudinal data analysis,  correlation analysis,  linear and logistics regression,  robust linear regression,  linear mixed model  clustering  Principle Component Analysis (PCA)  Eigen something  **Familiarity with:**  RNASeq  Data formats (FASTQ, BAM->SAM,  **a**CGH  Use of **r**CGH (R coded tool) **[2]**    **Knowledge of:**  **Public accessable projects as examples demonstrating my applying the above:**  Knowledge of Omics concepts/terminology, Bioinformatics,  **GitHub**    **GitData**  AncestryDNA Data upload to GitData |

Bioinformatics workflows, tools by study of;

browsing recent publications in Open Access Journal Publications

open access academic publications (BMC Bioinformatics, PubMed Central, ),

**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 |

|  |
| --- |
| **[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> | |

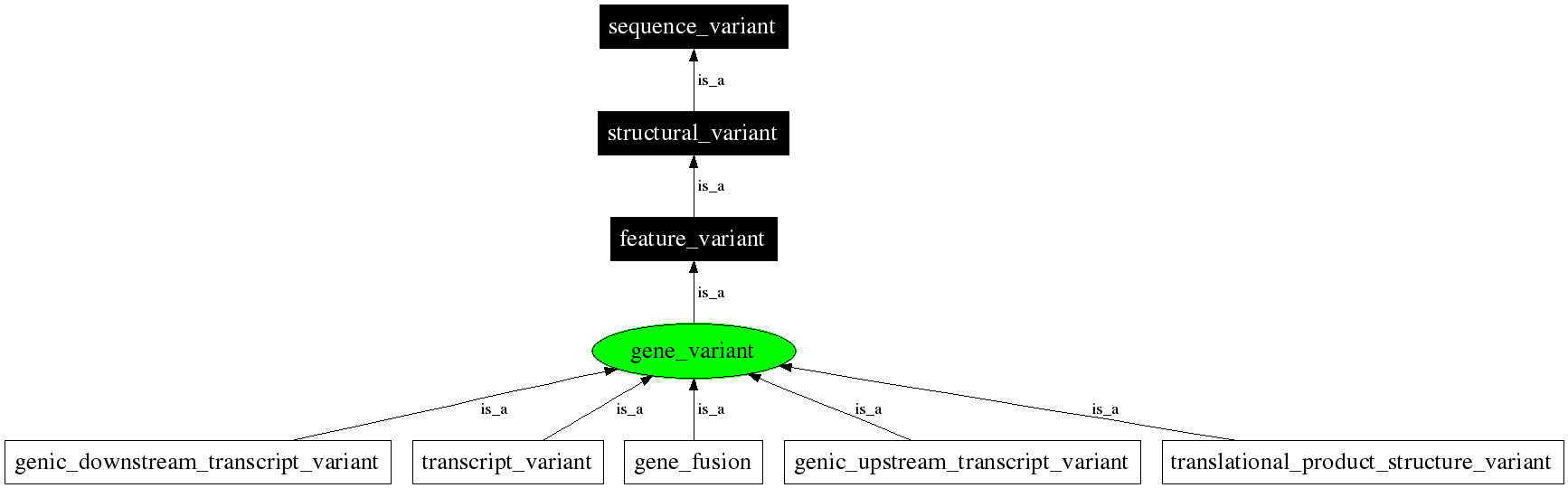
**Sequence Ontology (SO)**

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**

**[K1]:** Inferential Statistics, Analytics, Machine Learning

**[K2]:** Tools: R/RStudio, Bioconductor Packages, (Use Cases: Bioconductor Package Vignettes, )

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

**[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>

**[K4]:** Publicly Available Genomic Data Resources (navigation, content, search)

|  |
| --- |
| **[K1]: Inferential Statistics, Analytics, Machine Learning** |
| |  | | --- | | **Knowledge of Inferential Statistics, Analytics, Machine Learning (R)** | | **15:55 P Values, clearly explained**  <https://www.youtube.com/watch?v=5Z9OIYA8He8&list=PLblh5JKOoLUJo2Q6xK4tZElbIvAACEykp&index=16>  **5:01 Linear Regression in R**  <https://www.youtube.com/watch?v=u1cc1r_Y7M0&list=PLblh5JKOoLUJo2Q6xK4tZElbIvAACEykp&index=24> **27:27 Linear Models Pt.1 - Linear Regression**  <https://www.youtube.com/watch?v=nk2CQITm_eo&list=PLblh5JKOoLUJo2Q6xK4tZElbIvAACEykp&index=23>  **11:38 Linear Models Pt.2 - t-tests and ANOVA**  <https://www.youtube.com/watch?v=NF5_btOaCig&list=PLblh5JKOoLUJo2Q6xK4tZElbIvAACEykp&index=25>  **14:40 Linear Models Pt.3 - Design Matrices (old version)**  <https://www.youtube.com/watch?v=2UYx-qjJGSs&list=PLblh5JKOoLUJo2Q6xK4tZElbIvAACEykp&index=26>  **8:20 Linear Models Pt.3 - Design Matrix Examples in R**  <https://www.youtube.com/watch?v=Hrr2anyK_5s&list=PLblh5JKOoLUJo2Q6xK4tZElbIvAACEykp&index=27>  **6:05 PCA main ideas**  <https://www.youtube.com/watch?v=HMOI_lkzW08&list=PLblh5JKOoLUJo2Q6xK4tZElbIvAACEykp&index=21>  **20:16 Principal Component Analysis (PCA) clearly explained**  <https://www.youtube.com/watch?v=_UVHneBUBW0&list=PLblh5JKOoLUJo2Q6xK4tZElbIvAACEykp&index=22>  **21:58 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)  **8:57 PCA in R**  <https://www.youtube.com/watch?v=0Jp4gsfOLMs&list=PLblh5JKOoLUJo2Q6xK4tZElbIvAACEykp&index=4>  **11:37 PCA in Python**  <https://www.youtube.com/watch?v=Lsue2gEM9D0&list=PLblh5JKOoLUJo2Q6xK4tZElbIvAACEykp&index=5>  **10:14 RPKM, FPKM, TPM**  <https://www.youtube.com/watch?v=TTUrtCY2k-w&list=PLblh5JKOoLUJo2Q6xK4tZElbIvAACEykp&index=6>  **8:18 MDS and PCoA**  <https://www.youtube.com/watch?v=GEn-_dAyYME&list=PLblh5JKOoLUJo2Q6xK4tZElbIvAACEykp&index=7>  **7:45 MDS and PCoA in R**  <https://www.youtube.com/watch?v=pGAUHhLYp5Q&list=PLblh5JKOoLUJo2Q6xK4tZElbIvAACEykp&index=15>  **11:48 t-SNE, Clearly Explained**  <https://www.youtube.com/watch?v=NEaUSP4YerM&list=PLblh5JKOoLUJo2Q6xK4tZElbIvAACEykp&index=8>  **8:31 K-means clustering**  <https://www.youtube.com/watch?v=4b5d3muPQmA&list=PLblh5JKOoLUJo2Q6xK4tZElbIvAACEykp&index=28>  **11:19 Hierarchical Clustering**  <https://www.youtube.com/watch?v=7xHsRkOdVwo&list=PLblh5JKOoLUJo2Q6xK4tZElbIvAACEykp&index=10>    **16:49 Drawing and Interpreting Heatmaps**  <https://www.youtube.com/watch?v=oMtDyOn2TCc&list=PLblh5JKOoLUJo2Q6xK4tZElbIvAACEykp&index=11>  **18:27 False Discovery Rates, FDR, clearly explained**  <https://www.youtube.com/watch?v=K8LQSvtjcEo&list=PLblh5JKOoLUJo2Q6xK4tZElbIvAACEykp&index=17>  **5:15 Fisher's Exact Test and the Hypergeometric Distribution**  <https://www.youtube.com/watch?v=udyAvvaMjfM&list=PLblh5JKOoLUJo2Q6xK4tZElbIvAACEykp&index=18>    **15:37 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/> |      |  | | --- | | **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  A complete statistical model for calibration of RNA-seq counts using external spike-ins and maximum likelihood theory  11 Mar 2019 PLOS Computational Biology  <https://doi.org/10.1371/journal.pcbi.1006794>  Citations: 1  A Bayesian mixture modelling approach for spatial proteomics  Research Article | published 27 Nov 2018 PLOS Computational Biology  <https://doi.org/10.1371/journal.pcbi.1006516>  Citations: 5 |      |  | | --- | | **A Matrix Algebra Companion for Statistical Learning (matrix4sl)**  <https://www.gastonsanchez.com/matrix4sl/types-of-tables.html> | |
|  |
|  |

|  |
| --- |
| **[K1]: Inferential Statistics, Machine Learning (Applied to Computational Biology)** |
| |  | | --- | | **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  A complete statistical model for calibration of RNA-seq counts using external spike-ins and maximum likelihood theory  11 Mar 2019 PLOS Computational Biology  <https://doi.org/10.1371/journal.pcbi.1006794>  Citations: 1  A Bayesian mixture modelling approach for spatial proteomics  Research Article | published 27 Nov 2018 PLOS Computational Biology  <https://doi.org/10.1371/journal.pcbi.1006516>  Citations: 5 |      |  | | --- | |  |      |  | | --- | |  |      |  | | --- | |  |      |  | | --- | |  |      |  | | --- | |  | |
|  |
|  |

|  |
| --- |
| **[K1]: 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/> | |
|  |

|  |
| --- |
| **[K2]:** Bioinformatics Tools  Tools: R/RStudio, Bioconductor Packages, (Use Cases: Bioconductor Package Vignettes, )  Tools: Python, BioPython |
| **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]: Tools: R/RStudio, Bioconductor Packages, (Use Cases: Bioconductor Package Vignettes, )** |
| |  | | --- | | **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>

**[K2]: Tools: R/RStudio, Bioconductor Packages, (Use Cases: Bioconductor Package Vignettes, )**

|  |
| --- |
| **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>

|  |
| --- |
| **[K4]: Publicly Available Genomic Data Resources, Viewers (navigation, content, search)** |
| |  | | --- | | **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.  **NCBI » GEO » Info » 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>? |

**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** |

|  |
| --- |
| **[5.1]** **-** Browsing open access full-text publications on use of tools **Knowledge of bioinformatics Tools, Workflows: Relevant Publications** |
| **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> |
|  |

**[5.2] – Open Source Software Tools, Browsers**

|  |
| --- |
| **Knowledge of DNA/RNA Bioinformatics** |
|  |
|  |
|  |

|  |
| --- |
| **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 videos)** Day 1: <https://www.youtube.com/watch?v=thHgPqQpkE4&feature=emb_err_woyt> **RStudio**  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> **RStudio**  Clustering example  Feature Selection  Pseudotime Analysis  Differental Expression Analysis  DE Real Dataset  Comparing/Combining scRNA  Search scRNA-Seq Data  Seurat  scRNA-Seq Pipeline |

|  |
| --- |
| **bioRxiv**  <https://www.biorxiv.org/>  <https://www.biorxiv.org/search> |
| **Principal Components Analysis / Feature Selection**  **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> |

|  |
| --- |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |

|  |
| --- |
| **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> |

|  |
| --- |
| **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  What does data in a count matrix look like?  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 |

**BIOCONDUCTOR**

|  |
| --- |
| **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 Hefin Rhys** |
| **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> |

|  |
| --- |
| **Bioconductor** |
| **Bioconductor**  <http://bioconductor.org/>  **Bioconductor Courses & Conferences**  <http://bioconductor.org/help/course-materials/>  **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 |

|  |
| --- |
| **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> |

|  |
| --- |
| **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** |

|  |
| --- |
|  |
| |  | | --- | | **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**  <https://bioconductor.org/packages/release/bioc/manuals/GenomicAlignments/man/GenomicAlignments.pdf>  October 17, 2020  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 IRangespackage** 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 | |

|  |
| --- |
| **Tidyverse** |
| **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>  **Bioinformatics in the Tidyverse**  <https://chapmandu2.github.io/post/2017/02/21/bioinformatics-in-the-tidyverse/> 2017/02/21  **The tidyverse: dplyr, ggplot2, and friends**  <https://monashbioinformaticsplatform.github.io/r-more/topics/tidyverse.html#an-rna-seq-example>  ggplot2 revisited  dplyr  The pipe %>%  tidyr  An RNA-Seq example  <https://www.tidyverse.org/>  <https://dplyr.tidyverse.org/>  <https://tidyr.tidyverse.org/>  <https://tibble.tidyverse.org/> |
|  |
|  |

|  |
| --- |
| **Bioconductor: edgeR**  Empirical Analysis of Digital Gene Expression Data in R 24 |
|  |
|  |

|  |
| --- |
| **Bioconductor: DESeq2** Differential gene expression analysis based on the negative binomial distribution 27 |
| **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>

|  |
| --- |
|  |

|  |
| --- |
| **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> |
|  |
| **Sequence Read Archive (SRA)**  <https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi>?  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 is NIH’s primary archive of high-throughput sequencing data and is part of the international partnership of archives (INSDC).  **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. |
| **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/> |
| **dbSNP**  **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), the European Nucleotide Archive (ENA), and 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; see BLAST info for more information about the numerous BLAST databases.  **-** 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> |

|  |
| --- |
| **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. |

|  |
| --- |
| **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. |
| 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 |

|  |
| --- |
| **[rCGH]**  **Pipeline for Analyzing and Visualizing Array-Based CGH Data** |
| **Pipeline for Analyzing and Visualizing Array-Based CGH Data 042720.pdf**  **rCGH**,  a comprehensive array-based CGH analysis workflow, integrating functionalities specifically designed for precision medicine.  rCGH ensures a full traceability by saving all the process parameters, and facilitates genomic profiles interpretation and decision-making through interactive visualizations.  rCGH supports commercial arrays:  Agilent (from 44K to 400K arrays),  Affymetrix SNP6.0,  cytoScanHD  Notice that rCGH is a superclass designed for calling common methods.  Depending on the type of array and there adfunctions used, the resulting objects will be assigned to classes  rCGH-Agilent,  rCGH-SNP6,  rCGH-cytoScan,  rCGH-generic.  These classes inherit from the superclass, and allow array-specific pre-parametrizations. |

|  |
| --- |
|  |
|  |
| **data analytics features modeling**  **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  Pandas package  Excel  R  Python |
|  |
| **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 |
|  |
|  |
|  |
|  |
|  |
|  |