

Submission to Journal of Investigative Dermatology

Title

Research Techniques Made Simple: Bioinformatics for Genome-Scale Biology

Authors

Amy C. Foulkes PhD¹, David S. Watson², Christopher E.M. Griffiths¹, Richard B. Warren PhD¹,
Wolfgang Huber PhD³, Michael R. Barnes PhD²

Institutions

¹The Dermatology Centre, Salford Royal NHS Foundation Trust, The University of Manchester,
Manchester Academic Health Science Centre, M6 8HD, UK

²William Harvey Research Institute, Centre for Translational Bioinformatics, Barts and The
London School of Medicine and Dentistry, Charterhouse Square, London, UK

³European Molecular Biology Laboratory, Heidelberg, Germany

Corresponding author

Dr A. C. Foulkes, NIHR Academic Clinical Lecturer in Dermatology, The Dermatology Centre,
Salford Royal NHS Foundation Trust, The University of Manchester, Manchester Academic
Health Science Centre, M6 8HD. ORCID account 0000-0003-2680-750X.

Amy.foulkes@manchester.ac.uk

Disclosure

ACF has received educational support to attend conferences from or acted as a consultant or speaker for Abbvie, Almirall, Eli Lilly, Leo Pharma, Novartis, Pfizer, Janssen and UCB. CEMG has acted as a consultant and/or speaker for Abbvie, Janssen, Novartis, Sandoz, Rock Creek Pharma, Pfizer, Eli Lilly, UCB, Leo Pharma, Galderma and Celgene. RBW has acted as a consultant and/or speaker for Abbvie, Amgen, Almirall, Boehringer, Medac, Eli Lilly, Janssen, Leo Pharma, Pfizer, Novartis, Sun Pharma, Valeant, Schering-Plough (now MSD) and Xenoport.

Author roles and email addresses

Mr. D.S. Watson, Postgraduate researcher

D.Watson@qmul.ac.uk

Professor C.E.M. Griffiths, Foundation Professor of Dermatology

Christopher.griffiths@manchester.ac.uk

Professor. R.B. Warren, Reader in Dermatology and Honorary Consultant Dermatologist

Richard.warren@manchester.ac.uk

Dr. W. Huber, Group Leader and Senior Scientist

whuber@embl.de

Dr. M.R. Barnes, Reader in Bioinformatics, Director – Centre for Translational Bioinformatics

M.r.barnes@qmul.ac.uk

Abstract

High throughput biology presents unique opportunities and challenges for dermatological research. Drawing on a small handful of exemplary studies, we review some of the major lessons of these new technologies. We caution against several common errors and introduce helpful statistical concepts that may be unfamiliar to researchers without experience in bioinformatics. We recommend specific software tools that can aid dermatologists at varying levels of computational literacy, including platforms with command line and graphical user interfaces. The future of dermatology lies in integrative research, in which clinicians, laboratory scientists, and data analysts come together to plan, execute, and publish their work in open forums that promote critical discussion and reproducibility. In this article, we offer guidelines that we hope will steer researchers toward best practices for this new and dynamic era of data intensive dermatology.

Introduction

Modern dermatology has been revolutionized by the many so-called ‘omic’ profiling platforms enabled by high-throughput sequencing (HTS, also referred to as next generation sequencing, NGS). Plunging data generation costs have enabled dermatology researchers to generate genome scale data relating to genome sequence variation (Scott et al., 2013), epigenomes (Zhou et al., 2016), and transcriptomes (Li et al., 2014, Swindell et al., 2016), and these developments have increased the dermatology-relevant data openly available in repositories (Table 1).

Bioinformatics refers to the tools used to collect, classify, and analyze such datasets, collectively enabling the field of *computational biology*. Bioinformatics techniques have been developed to make sense of the output of omic platforms, including HTS, microarrays, liquid chromatography-mass spectrometry, and more (Kimball et al., 2012).

Physicians are key instigators of research data collection requiring computational biology. Structured and validated analysis ‘pipelines’ for most omic data have been implemented for researchers at various levels of complexity. Software has been designed for all ranges of computational ability, from simple “point and click” graphic user interfaces (GUIs) to highly customizable command line interfaces (CLIs), with the latter approach offering superior flexibility and analytical complexity. Although programming may seem like a daunting challenge for those without backgrounds in math, computer science, or statistics, with practice, computational methods for exploratory and inferential analytics can become a familiar part of the research toolkit. Of course, there is no substitute for expertise, and we advise all research teams working with omic data to consult a bioinformatician early and often. Here we highlight several points of special relevance to

the dermatologist and dermatology researcher, based on the first-hand experience of a junior clinician.

Considerations Prior to Data Collection

Experimental design

Researchers in dermatology employ a wide variety of HTS techniques, many of which have been discussed previously in the Research Techniques Made Simple (RTMS) series.. These include transcriptome analysis with RNA-Seq (Antonini et al., 2017, Whitley et al., 2016), immunosequencing (Matos et al., 2017), genome-wide epigenetics (Capell and Berger, 2013), proteomics, metabolomics, metagenomics and assessment of the microbiome (Jo et al., 2016). Additionally the Molecular Revolution in Cutaneous biology series provided an overview of HTS techniques (Anbunathan and Bowcock, 2017, Botchkareva, 2017, Johnston et al., 2017, Kong and Segre, 2017, Sarig et al., 2017) as do Grada et al. in an earlier RTMS publication (Grada and Weinbrecht, 2013). However, researchers often do not reach out to data analysts until a study is practically complete. At that point, they may look for a mathematically inclined colleague to fill in the blanks of a statistical model and provide a friendly p -value suitable for publication. This order of events is all wrong. As Ronald Fisher famously put it back in 1938, “To consult the statistician after an experiment is finished is often merely to ask him to conduct a post-mortem examination. He can perhaps say what the experiment died of.”

The data analysis strategy – including the choice of statistical approaches, should be integral to planning any research study. Hypothesis testing, regression and other statistical methods rely on rigorous collection and quality of the data, and any lapses here usually

cannot be fixed retrospectively. How many samples are required to adequately power your experiment? If samples cannot be processed all at once, does it matter how they are grouped into separate batches? If the data do not corroborate your hypothesis, can a modified research question generate interesting results? Failure to consider these questions before data collection may doom a study before it even begins. Statistical expertise is required to answer these questions, which is why we urge researchers to team up with a data analyst who can help guide them through these tricky issues. This will typically either be a statistician, with a background in math and statistics, or a bioinformatician, more likely with a background in computer science and machine learning. While there is considerable overlap in their respective areas of expertise, statisticians and bioinformaticians may offer differing (and sometimes complementary) perspectives on a given biological question.

One of the most fundamental tools in statistical analysis is hypothesis testing. The principles of hypothesis testing are illustrated in Table 2, illustrated with Li et al. (Li et al., 2014) as an exemplar study in the field (see also supplementary powerpoint). In this exploratory study, RNA-seq was utilised to evaluate the transcriptomes of lesional psoriatic and normal skin (from a large cohort of 174 individuals). A subset of these samples has been studied previously using microarrays, allowing for comparison of the methodologies; RNA-seq identified many more differentially expressed transcripts enriched in immune system processes.

Detailed discussion of requirements for testing a hypothesis will facilitate better downstream clinical data collection, ultimately maximising the opportunity to detect a clinically relevant association. Several key themes tend to dominate experimental design

considerations, including selection of appropriate numbers of biological replicates (Schurch et al., 2016), minimization of batch effects (Leek et al., 2010), and appropriate correction for multiple testing (Allison et al., 2006). For a general overview of issues related to HTS study design we recommend other excellent reviews (Allison et al., 2006, Conesa et al., 2016).

The steps outlined in Table 1 apply to most forms of omic data. Methods for computing test statistics vary depending on the data and underlying statistical assumptions. An overview of some common data types and test statistics in dermatological research are discussed elsewhere (Silverberg, 2015).

Batch Effects

Often a study's sample size exceeds the maximum number of samples that can be simultaneously processed by the available equipment. In such cases it is common to process the samples in multiple batches. This inevitably introduces batch effects, in which technical artifacts become significant, perhaps even dominant drivers of variation in a dataset. There are several methods for batch adjustment (Oytam et al., 2016).

Each method has its merits, but none can overcome poor study design. If batch is confounded with a clinical covariate—say, all disease samples were processed in Batch A, while all healthy samples were processed in Batch B—then there is no way to disentangle the technical from the biological variation. Ideally, each batch would represent a microcosm of the experiment itself, with proportionate numbers of samples from all relevant groups. While this cannot always be done in practice, the closer researchers come to attaining this goal, the more accurate their results will be.

Considerations after data collection

Software and workflows for omic analysis

As a rule of thumb, processing of raw HTS data, including genome alignment and assembly, is likely to require access to one or several devoted computers that can execute jobs in parallel. However, once the initial data processing is complete, in most cases the biological “downstream” analysis can be performed using a laptop. The analysis of omic data, including HTS, is supported by a range of widely used software packages that can be arranged into analysis workflows. Many packages have been made freely available by their authors with an open source license and in this field there is very little correlation between the price of software and its usefulness. A workflow is a software pipeline that takes raw data as input, transforms and summarises the data, conducts exploratory and/or inferential analytics, and exports results ready for biological interpretation. Command line genomic analysis tools can be scaled to use available computing resources, and are highly customizable to meet the requirement of an experiment. Many standard analysis tools can also be accessed remotely using the Galaxy workflow environment (<https://usegalaxy.org>). Galaxy offers users a simple but highly customisable GUI environment to perform many bioinformatics tasks. Galaxy is also well documented and serves as an excellent introduction to HTS analysis pipelines.

Processing HTS data

The short read is the common currency of HTS methods, but the way the read is processed is highly dependent on the analysis objective (Figure 1). In most cases processing commences with alignment to a reference genome using a tool, such as Burrows-Wheeler

Aligner (BWA) or bowtie2, producing binary alignment map (BAM) files. The alignment files can serve as the input to many other processes; in genetics they are used for variant calling, in epigenomics for peak calling, and in transcriptomics to estimate transcript abundance. A recent revolution in transcriptomics are alignment-free mapping methods, such as Kallisto (Bray et al., 2016) and Salmon (Patro et al., 2017). These tools circumvent the cumbersome alignment step and directly estimate transcript abundance; they are several orders of magnitude faster than alignment-based methods, and so computationally efficient that they can be run on a laptop computer. The workflow utilized by Li et al. (Li et al., 2014) is illustrated in Figure 2.

Programming environments

While many programming environments are used in bioinformatics, the most popular choices tend to be R(R Core Team, 2014) and Python(Python Software, 2013). Software packages for these languages are often released under open source licenses, which means the tools are free to use and the code is publicly accessible. Large user communities have developed around these languages, and R in particular has become a lingua franca for bioinformaticians. This has been aided in no small part by the Bioconductor project (Huber et al., 2015), a major repository for biostatistical software based primarily on R. The site also hosts discussion forums, encouraging active user engagement and collaborative learning.

Several programming environments are widely used in bioinformatics, including R, Matlab(MathWorks, 2012) and Java (see Table 3). These are open source and freely

available, enabling statistical and graphical data manipulation within large, active user communities.

Hypothesis testing in the age of big data

Hypothesis tests and p -values are a workhorse of medical research, but some additional complexities enter the scene when we do not only do one or a few tests, but thousands or millions. Interpreting p -values is quite different in omic contexts than in more traditional “low-throughput” research. Say you test 10,000 genes in search of biomarkers to distinguish between case and control samples. You find 500 with p -values below 0.05, not to mention 10 with p -values below 0.001. Not bad, right? Wrong! Since p -values are uniformly distributed under the null hypothesis, we should expect 5% of all tests to reach the nominal “significance” level of 0.05 by chance alone. That’s a manageable problem when testing one or two hypotheses, but in omic experiments we typically test something on the order of thousands to millions of hypotheses.

Some early papers attempted to mitigate the issue by controlling the family-wise error rate (FWER), defined as the probability of finding at least one false positive in a series of hypothesis tests. For example, the Bonferroni correction used by Li et al. (Li et al., 2014) strongly controls the FWER by setting the significance threshold as the quotient of the type I error α and the total number of hypothesis tests m , so that all and only tests with $p \leq \alpha/m$ are declared significant. While the Bonferroni correction is guaranteed to control the FWER, it is an overly conservative method that is likely to lead to many false negatives as m grows.

Current practice is to control not the false positive rate (i.e., the proportion of truly null features that are nominally significant) but the *false discovery rate* (i.e., the proportion of nominally significant features that are truly null). This latter value is typically estimated using the Benjamini-Hochberg algorithm (Hochberg and Benjamini, 1990) or some variant thereof. This method takes a list of p -values as input and returns a matched list of ‘adjusted’ p -values, also known as q -values. Applying a 5% false discover rate (FDR) threshold means that 1 in 20 genes in the hit list to be a false positive. Given 10,000 uniformly distributed p -values, as hypothesized above, minimum q -values are typically >0.5 .

Visualization

The communication of results is key for data exploration, summarization, and ultimately publication. Readers can more readily absorb a well-made graphic than any table of numbers. Visualizing HTS results can be challenging due to the data’s high dimensionality, but projection techniques like principal component analysis (PCA) (Pearson, 1901), multi-dimensional scaling (MDS) (Torgerson, 1952), and t-Distributed Stochastic Neighbor Embedding (t-SNE) (Van Der Maaten et al., 2008) can render large matrices as easily digestible 2D or 3D scatterplots. Matos et al. (Matos et al., 2017) demonstrate how these methods can reveal powerful insights for dermatological research. More recent interactive tools such as plotly (<https://plot.ly/>), shiny (<https://shiny.rstudio.com/>), and ggvis (<http://ggvis.rstudio.com/>) can also aid in data exploration, or even create widgets for HTML publication.

Code sharing and reproducibility

A number of studies have found an alarming lack of reproducibility in modern omic and clinical research (Open Science Collaboration, 2015). Many factors contribute to this problem, including the widespread failure to publish analysis code (Baker, 2016). Though some inroads have been made toward establishing best practices in molecular biology (Brazma et al., 2001), script sharing remains rare overall. Results may vary greatly depending on subtle, unstated analytic choices that are invisible without access to both raw data and the complete analysis script. Code sharing is a critical ingredient for open science; this will be apparent to researchers who have tried to re-use data in repositories, where code is absent and subject data often incomplete, making reproduction challenging at best. Excellent platforms exist for publishing code. Taking advantage of sites like GitHub can assist during peer review, enabling precise debate on the merits of particular methods. Set-up can be technically challenging, but user-friendly guides exist (<http://happygitwithr.com>). Researchers should ensure that they or their bioinformatician colleagues document and archive code, analogous to the use of a laboratory book as a record of research. This will ensure that bioinformatician turnover will not prevent ongoing analysis, since code will be clear, maintained, and transferable.

Summary and future directions

Embedding biostatisticians and computational biologists within clinical and academic research teams, as well as promoting better data and code sharing practices, will allow dermatologists to better document and communicate their research. The days of assembly line research—in which clinicians recruit patients, laboratory scientists process samples, and analysts crunch numbers—are coming to an end. The age of big data demands a

rigorous, integrated approach. Appropriate statistical design and analysis methods should be discussed and decided upon up front to meet most research objectives. By incorporating good experimental design and analytical work practice early, research quality and reproducibility will improve, and peer review by journals and grant awarding bodies is likely to be more favorable (Figure 3). Patients will be the ultimate beneficiaries of dermatology's drive to the forefront of life science research.

Acknowledgements

This forms part of the research themes contributing to the translational research portfolio of Barts and the London Cardiovascular Biomedical Research Centre, which is supported and funded by the National Institute of Health Research. (Huber et al., 2015)

References

- Allison DB, Cui X, Page GP, Sabripour M. Microarray data analysis: from disarray to consolidation and consensus. *Nat Rev Genet* 2006;7(1):55-65.
- Anbunathan H, Bowcock AM. The Molecular Revolution in Cutaneous Biology: The Era of Genome-Wide Association Studies and Statistical, Big Data, and Computational Topics. *The Journal of investigative dermatology* 2017;137(5):e113-e8.
- Antonini D, Mollo MR, Missero C. Research Techniques Made Simple: Identification and Characterization of Long Noncoding RNA in Dermatological Research. *The Journal of investigative dermatology* 2017;137(3):e21-e6.
- Baker M. 1,500 scientists lift the lid on reproducibility. *Nature* 2016;533(7604):452-4.
- Botchkareva NV. The Molecular Revolution in Cutaneous Biology: Noncoding RNAs: New Molecular Players in Dermatology and Cutaneous Biology. *The Journal of investigative dermatology* 2017;137(5):e105-e11.
- Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. *Nature biotechnology* 2016;34(5):525-7.
- Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, et al. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nature Genetics* 2001;29(4):365-71.
- Capell BC, Berger SL. Genome-wide epigenetics. *The Journal of investigative dermatology* 2013;133(6):e9.
- Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, et al. A survey of best practices for RNA-seq data analysis. *Genome Biol* 2016;17:13.
- Grada A, Weinbrecht K. Next-generation sequencing: methodology and application. *Journal of Investigative Dermatology* 2013;133(8):e11.
- Hochberg Y, Benjamini Y. More powerful procedures for multiple significance testing. *Stat Med* 1990;9(7):811-8.
- Huber W, Carey VJ, Gentleman R, Anders S, Carlson M, Carvalho BS, et al. Orchestrating high-throughput genomic analysis with Bioconductor. *Nature methods* 2015;12(2):115-21.
- Jo JH, Kennedy EA, Kong HH. Research Techniques Made Simple: Bacterial 16S Ribosomal RNA Gene Sequencing in Cutaneous Research. *The Journal of investigative dermatology* 2016;136(3):e23-7.
- Johnston A, Sarkar MK, Vrana A, Tsoi LC, Gudjonsson JE. The Molecular Revolution in Cutaneous Biology: The Era of Global Transcriptional Analysis. *The Journal of investigative dermatology* 2017;137(5):e87-e91.
- Kimball AB, Grant RA, Wang F, Osborne R, Tiesman JP. Beyond the blot: cutting edge tools for genomics, proteomics and metabolomics analyses and previous successes. *British Journal of Dermatology* 2012;166 Suppl 2:1-8.
- Kong HH, Segre JA. The Molecular Revolution in Cutaneous Biology: Investigating the Skin Microbiome. *The Journal of investigative dermatology* 2017;137(5):e119-e22.
- Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, Johnson WE, et al. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat Rev Genet* 2010;11(10):733-9.

- Li B, Tsoi LC, Swindell WR, Gudjonsson JE, Tejasvi T, Johnston A, et al. Transcriptome analysis of psoriasis in a large case-control sample: RNA-seq provides insights into disease mechanisms. *The Journal of investigative dermatology* 2014;134(7):1828-38.
- MathWorks. MATLAB and Statistics Toolbox Release. Boston, MA 2012.
- Matos TR, de Rie MA, Teunissen MBM. Research Techniques Made Simple: High-Throughput Sequencing of the T-Cell Receptor. 2017(1523-1747 (Electronic)).
- Open Science Collaboration. PSYCHOLOGY. Estimating the reproducibility of psychological science. *Science* 2015;349(6251):aac4716.
- Oytam Y, Sobhanmanesh F, Duesing K, Bowden JC, Osmond-McLeod M, Ross J. Risk-conscious correction of batch effects: maximising information extraction from high-throughput genomic datasets. *BMC Bioinformatics* 2016;17(1):332.
- Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. *Nature methods* 2017;14(4):417-9.
- Pearson K. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine* 1901;2(11):559-72.
- Python Software F. Python Language Reference. Python Software Foundation Wilmington, DE; 2013.
- R Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2014, <http://www.R-project.org/>. 2014 [accessed].
- Sarig O, Sprecher E, Kong HH, Segre JA, Anbunathan H, Bowcock AM, et al. The Molecular Revolution in Cutaneous Biology: Era of Next-Generation Sequencing
- The Molecular Revolution in Cutaneous Biology: Investigating the Skin Microbiome
- The Molecular Revolution in Cutaneous Biology: The Era of Genome-Wide Association Studies and Statistical, Big Data, and Computational Topics
- The Molecular Revolution in Cutaneous Biology: Noncoding RNAs: New Molecular Players in Dermatology and Cutaneous Biology. *The Journal of investigative dermatology* 2017;137(5):e79-e82.
- Schurch NJ, Schofield P, Gierlinski M, Cole C, Sherstnev A, Singh V, et al. How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? *RNA* 2016;22(6):839-51.
- Scott CA, Plagnol V, Nitoiu D, Bland PJ, Blaydon DC, Chronnell CM, et al. Targeted sequence capture and high-throughput sequencing in the molecular diagnosis of ichthyosis and other skin diseases. *The Journal of investigative dermatology* 2013;133(2):573-6.
- Silverberg JI. Study designs in dermatology: Practical applications of study designs and their statistics in dermatology. *J Am Acad Dermatol* 2015;73(5):733-40; quiz 41-2.
- Swindell WR, Sarkar MK, Liang Y, Xing X, Gudjonsson JE. Cross-Disease Transcriptomics: Unique IL-17A Signaling in Psoriasis Lesions and an Autoimmune PBMC Signature. *The Journal of investigative dermatology* 2016;136(9):1820-30.
- Torgerson WS. Multidimensional scaling: I. Theory and method. *Psychometrika* 1952;17(4):401-19.
- Van Der Maaten L, Hinton G, van der Maaten GH. Visualizing Data using t-SNE. *Journal of Machine Learning Research* 2008;9:2579-605.

Whitley SK, Horne WT, Kolls JK. Research Techniques Made Simple: Methodology and Clinical Applications of RNA Sequencing. The Journal of investigative dermatology 2016;136(8):e77-82.

Zhou F, Wang W, Shen C, Li H, Zuo X, Zheng X, et al. Epigenome-Wide Association Analysis Identified Nine Skin DNA Methylation Loci for Psoriasis. The Journal of investigative dermatology 2016;136(4):779-87.

Table 1. High throughput sequencing repositories

Repository	Website	Curator
Europe		
European Nucleotide Archive (ENA)	http://www.ebi.ac.uk/ena	European Bioinformatics Institute
ArrayExpress	http://www.ebi.ac.uk/arrayexpress	European Bioinformatics Institute
European Genome-phenome Archive (EGA)	https://www.ebi.ac.uk/ega/home	European Bioinformatics Institute
United States		
dbGAP	https://www.ncbi.nlm.nih.gov/gap	The National Center for Biotechnology Information
Gene Expression Omnibus (GEO)	https://www.ncbi.nlm.nih.gov/geo	The National Center for Biotechnology Information
Short Read Archive (SRA)	https://www.ncbi.nlm.nih.gov/sra	The National Center for Biotechnology Information

Table 2. Principles of hypothesis testing, from Li et al. (Li et al., 2014)

Step in hypothesis testing	Example
Ask a clinically relevant, testable question	Is there a significant difference between this set of genes expressed in subjects with psoriasis vs. those without?
Choose an experimental design and statistical framework	Gene expression is modeled as a linear function of disease condition
Set up a null hypothesis, i.e. a testable claim that becomes the target of statistical analysis	There is no significant difference between the average expression of gene g in subjects with and without psoriasis
Fix a rejection region, i.e. the degree of evidence against the null hypothesis at which it may be rejected	Genes whose t -statistics correspond to false discovery rates $\leq 5\%$ are declared differentially expressed
Conduct the experiment: collect data, compute the test statistics	Expression levels for each gene g_i are regressed onto one or several clinical predictors, generating a vector of t -statistics
Report results: all and only those genes that fall within the rejection region are declared differentially expressed	A number of genes were significantly differentially expressed in plaques of psoriasis when compared to controls

Table 3. Open source programming languages and resources for bioinformatics analysis of omic data

Open source resource	URL
Analysis code repositories	
Bioconductor	bioconductor.org
CRAN	www.cran.org
Bioperl	bioperl.org
Biopython	biopython.github.io
GitHub	github.com
BioJulia	github.com/BioJulia
Workflow tools	
Galaxy	usegalaxy.org
Visualization	
ShinyR	Shiny.rstudio.com
Plotly	plot.ly

Figure 1.

Common methodology for processing of short reads

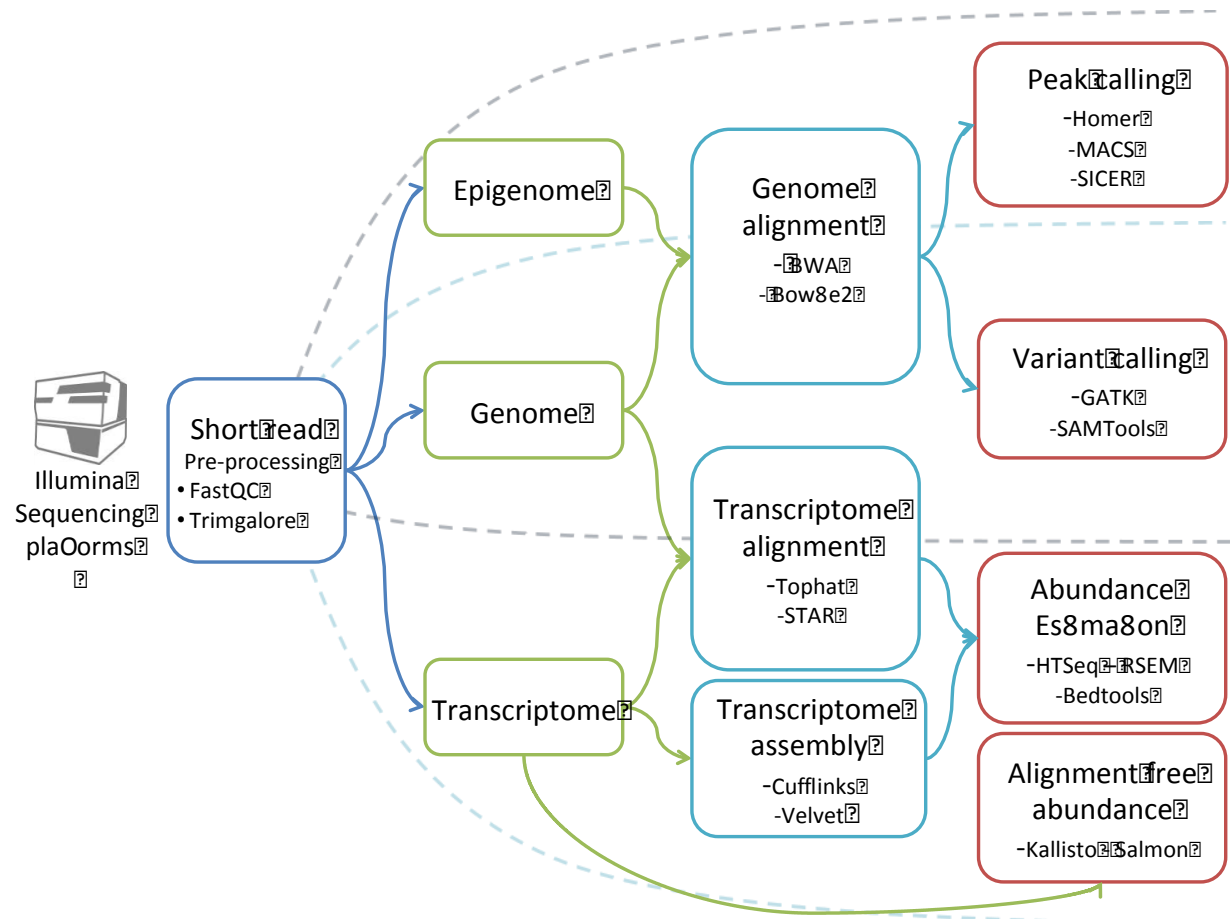


Figure 2. Example bioinformatic pipeline employed by Li et al.(Li et al., 2014)

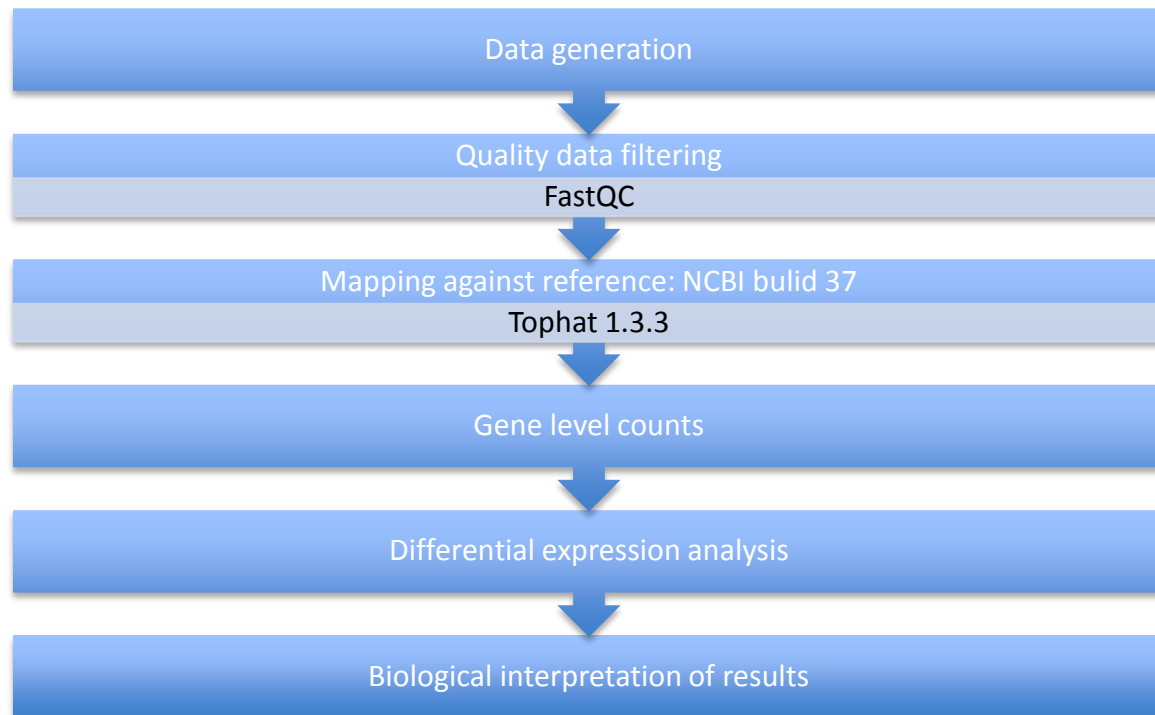
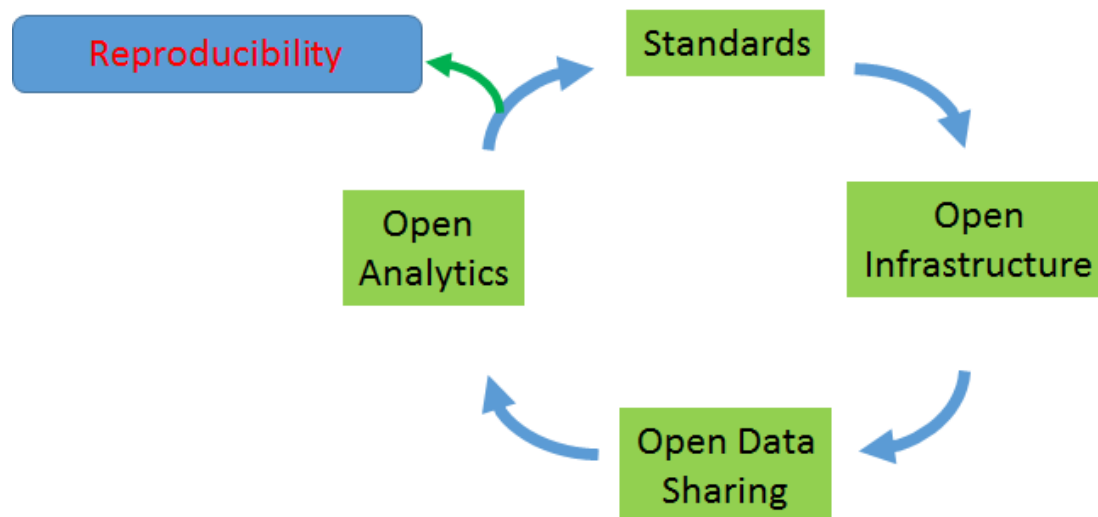


Figure 3. Reproducibility: Creating a virtuous circle



Summary

Advantages

- Bioinformatics methods allow efficient and powerful analysis of multi-omic data in a way that could not be achieved using simpler methods
- Bioinformatics software are customizable to all ranges of computational ability; however some informatics tasks are difficult and require experience
- Involving bioinformatician colleagues from project conception should improve project design, maximizing the opportunity to detect relevant association
- Sharing data, metadata, and code, and propagating the culture of bioinformaticians, will fuel best practices in dermatology research, promoting open research and reproducibility

Limitations

- Some statistical analysis methods require an understanding of underlying assumptions - erroneous assumptions can lead to false results
- The use of some analytical pipelines requires access to high performance computing facilities: this may be achieved by access to omic core facilities which provide researchers with compressed datasets that are amenable to PC-based analysis

Multiple Choice Questions

1. Which is an accurate description of batch effect?
 - a. Technical source of variation added to samples during handling
 - b. An uncommon problem in HTS experiments
 - c. Where proportionate samples are analyzed in each experiment
 - d. A problem which is not possible to adjust for using bioinformatic techniques

Answer A; Answer A describes batch effect, a common problem in HTS experiments. Answer c describes a way in which to avoid this problem and there are multiple techniques to adjust for batch effect, however no methodology will overcome poor study design.

2. The relevant significance measure in omic data is;
 - a. The p value
 - b. The false discovery rate
 - c. The false positive rate
 - d. The family-wise error rate

Answer B; The p value can be misleading in omic contexts. That is why the relevant significance measure in omic is not the false positive rate (i.e., the rate at which truly null features are declared significant) but the *false discovery rate* (i.e., the rate at which significant features are truly null) and this latter value is typically inferred from the former using the Benjamini-Hochberg algorithm or some variant thereof.

3. Which of the following is an analysis code repository;
 - a. GEO
 - b. R
 - c. Galaxy
 - d. GitHub

Answer D; GitHub is a web-based Git (version control) repository. Gene Expression Omnibus (GEO) is a public functional genomics data repository supporting MIAME-compliant data submissions. R is a statistical programming language. Galaxy is not a repository but is an open source, web-based platform for data intensive biomedical research.

4. Which of the following statements is true regarding sharing of analysis code;
 - a. This allows reproducibility of an analysis
 - b. Sharing of analysis code is technically challenging
 - c. Analysis code is required alongside submission of data and metadata for submission of original articles to major journals
 - d. There is no code sharing repository

Answer A; Sharing of analysis code is fundamental for reproducibility. Analysis scripts are required, alongside data and metadata. Sharing of code is easy and facilitated by excellent platforms, including GitHub (github.com). Major journals do not yet enforce submission of analysis code for peer review.

5. Which of the following is a major repository for biostatistical software?
- a. ShinyR
 - b. Plotly
 - c. Ggvis
 - d. Bioconductor

Answer D; Bioconductor is an open source, open development software project to provide tools for the analysis and comprehension of high-throughput genomic data. It is based primarily on the R programming language. ShinyR, plotly and ggvis are all interactive tools for data visualization.