Whole genome Next Generation DNA Sequencing (NGS) and Third Generation DNA Sequencing (TGS) data analysis

Fredrick M. Mobegi, PhD

BHKi hybrid seminar series

linkedin.com/fmobegi

twitter.com/mobeginomics

github.com/fmobegi

rpubs.com/fmobegi

Once upon a time in a small village…

2004-2008

B.Sc. Biochemistry

Graduated with a BSc degree from **Kenyatta University, Kenya,** and continued to work as a clinical laboratory technician at NPDL

M.Sc. Bioinformatics

Graduated with a Masters degree from **Wageningen University,**The Netherlands, and continued to work as a **graduate assistant** at **ILRI/BecA**

2009-2011

2012-2016

Ph.D. Bioinformatics & Computational Biology

Successfully defended my Thesis and was awarded a PhD at **Radboud University,** The Netherlands

2016

Postdoctoral Fellow

1st Postdoc: collaboration between **Netherlands Cancer Institute** and **GlaxoSmithKline**

2017-2020

Postdoctoral Fellow

2nd Postdoc: Microbiome Research Laboratory at **SAHMRI (Flinders Medical Centre)**

2020-2022

Research Fellow

3rd Postdoc: **The Centre for Crop & Disease Management, Curtin University**

2017-2020

Senior Bioinformatician

HLA/KIR immunology at PathWEST Department of Health **Government of Western Australia**

2011. UCSC, USA
2014. WTSI, UK
2014. Family

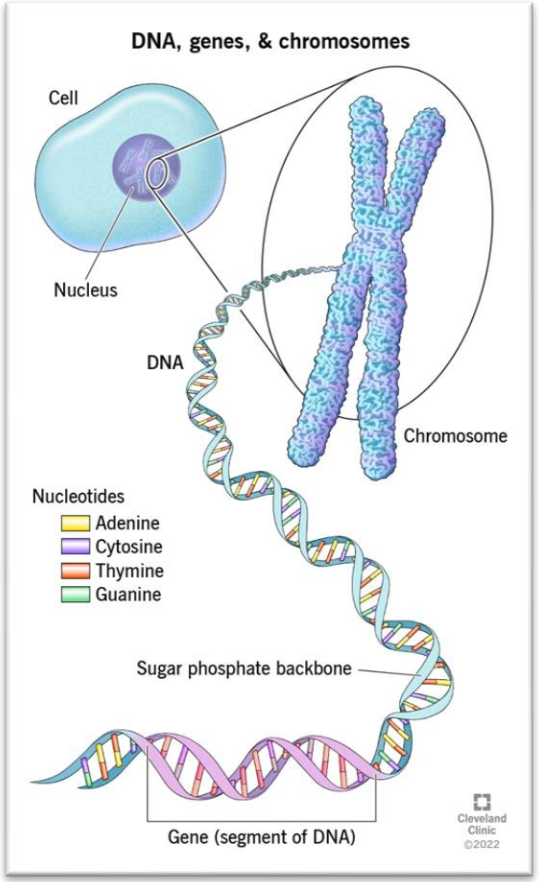
Let's get some terminology out of the way…

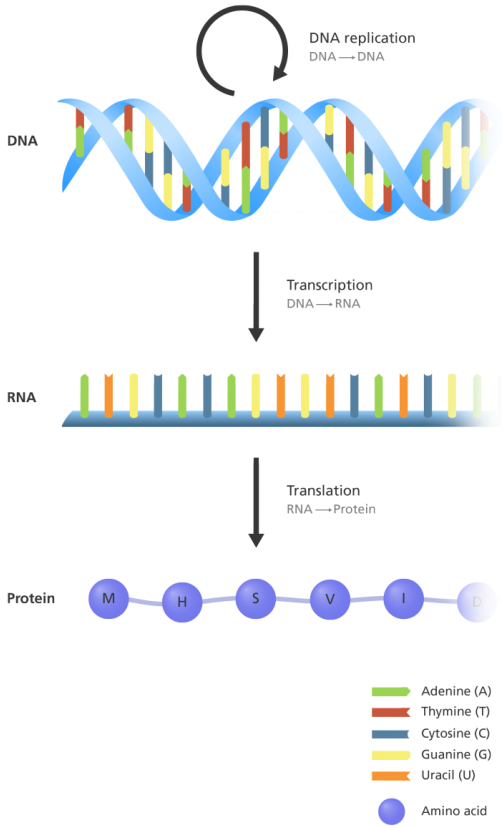
Genetics is the study of how genes and traits are passed down from one generation to the next: **Heredity**

Genomics is the study of the total or part of the genetic or epigenetic sequence information of organisms, and attempts to characterize the structure and function of these sequences and of downstream biological products: **Architecture**

A genome is an organism's complete set of deoxyribonucleic acid (DNA), a chemical compound that contains the genetic instructions needed to develop and direct the activities of every organism.

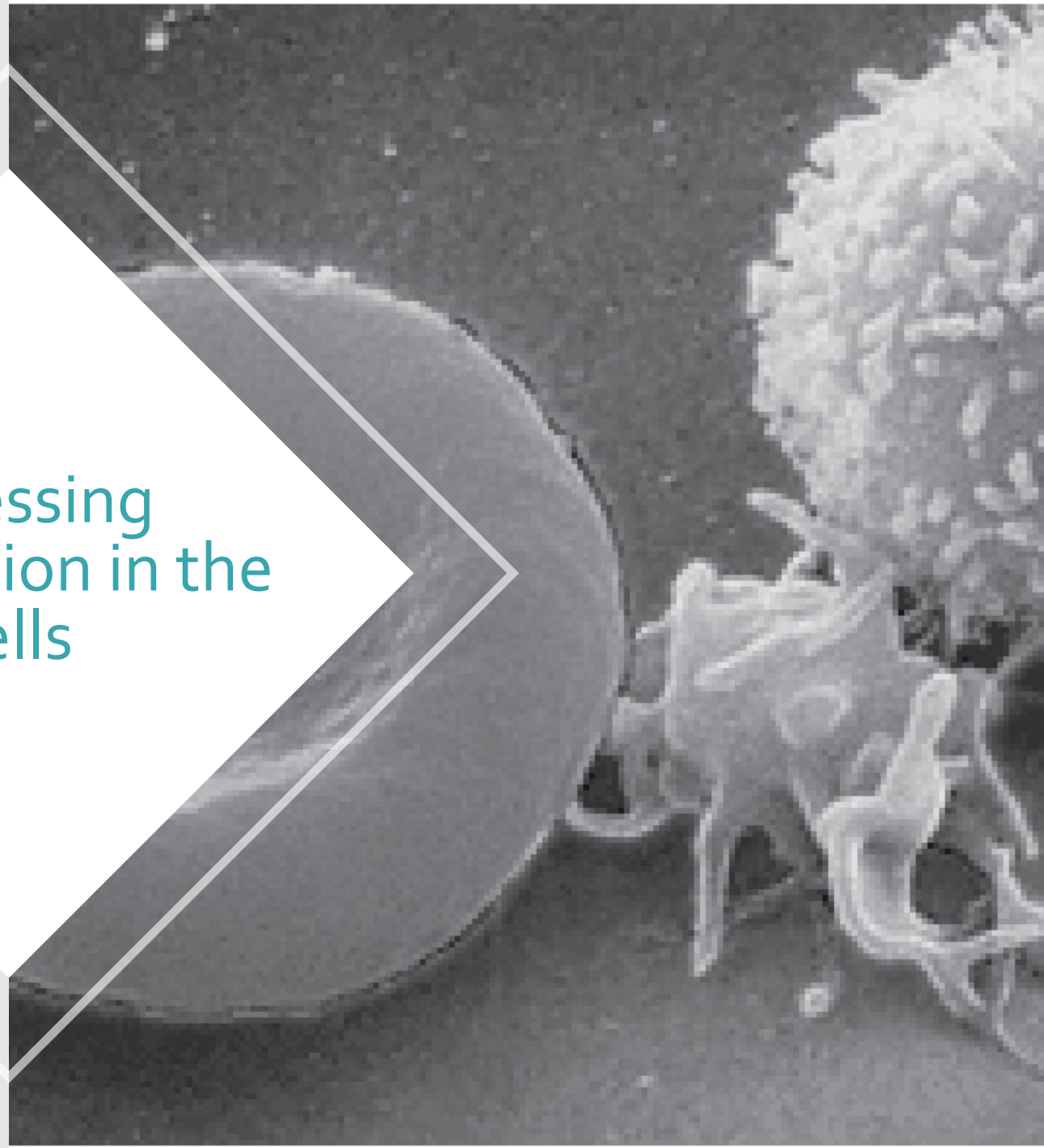
A gene is the basic physical and functional unit of heredity.



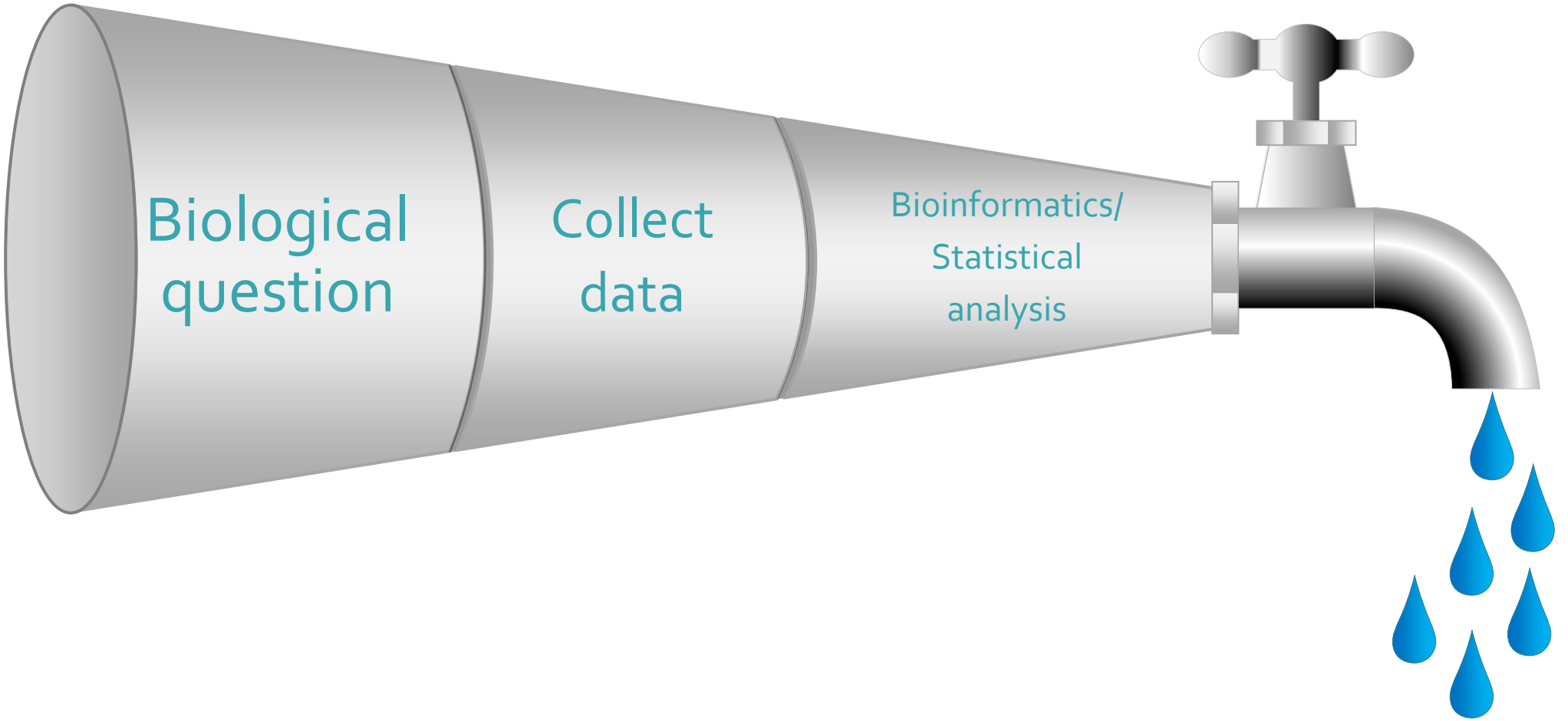


- https://www.who.int/news-room/questions-and-answers/item/genomics
- https://www.amnh.org/explore/ology/genetics/what-is-genetics

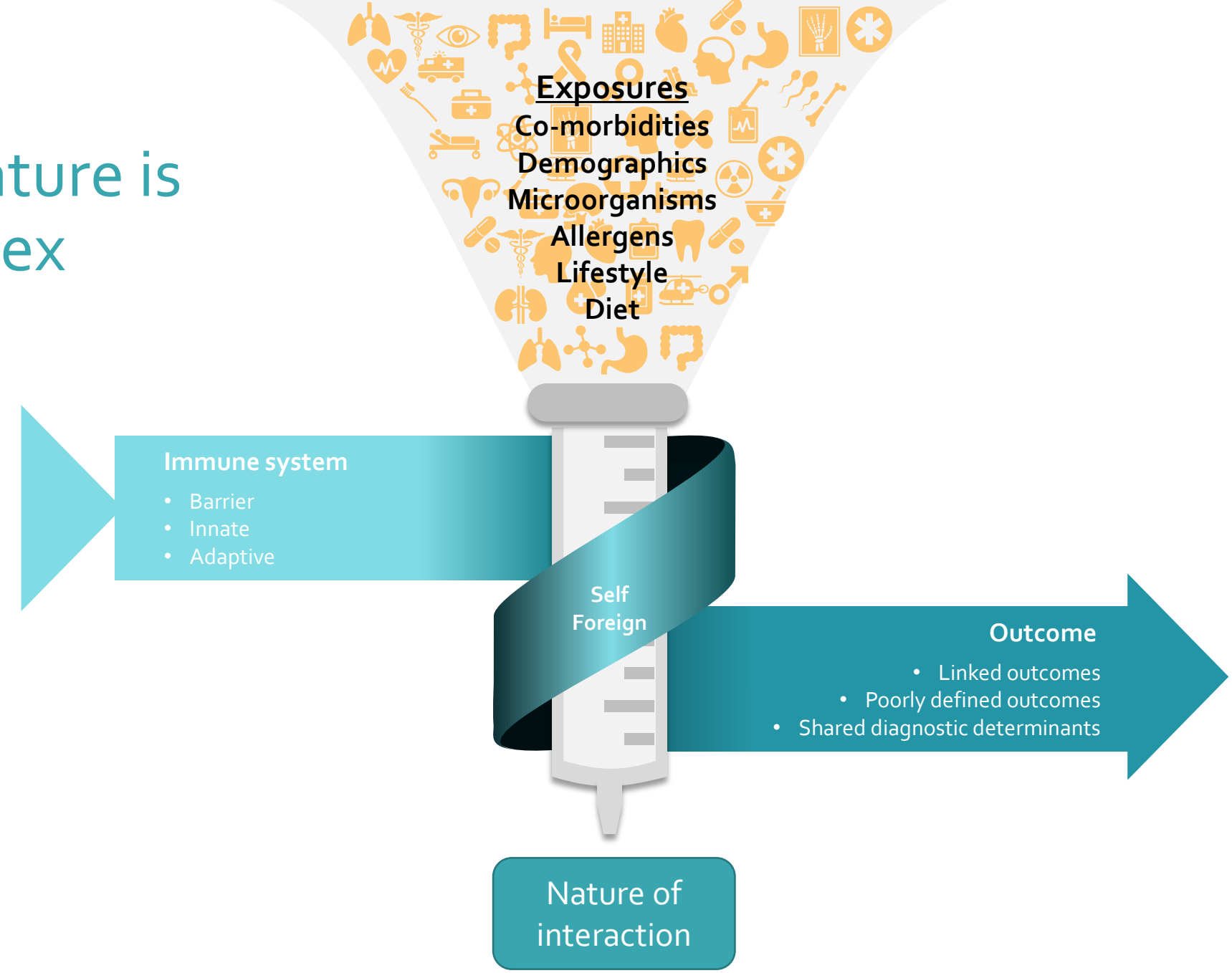
Accessing information in the cells

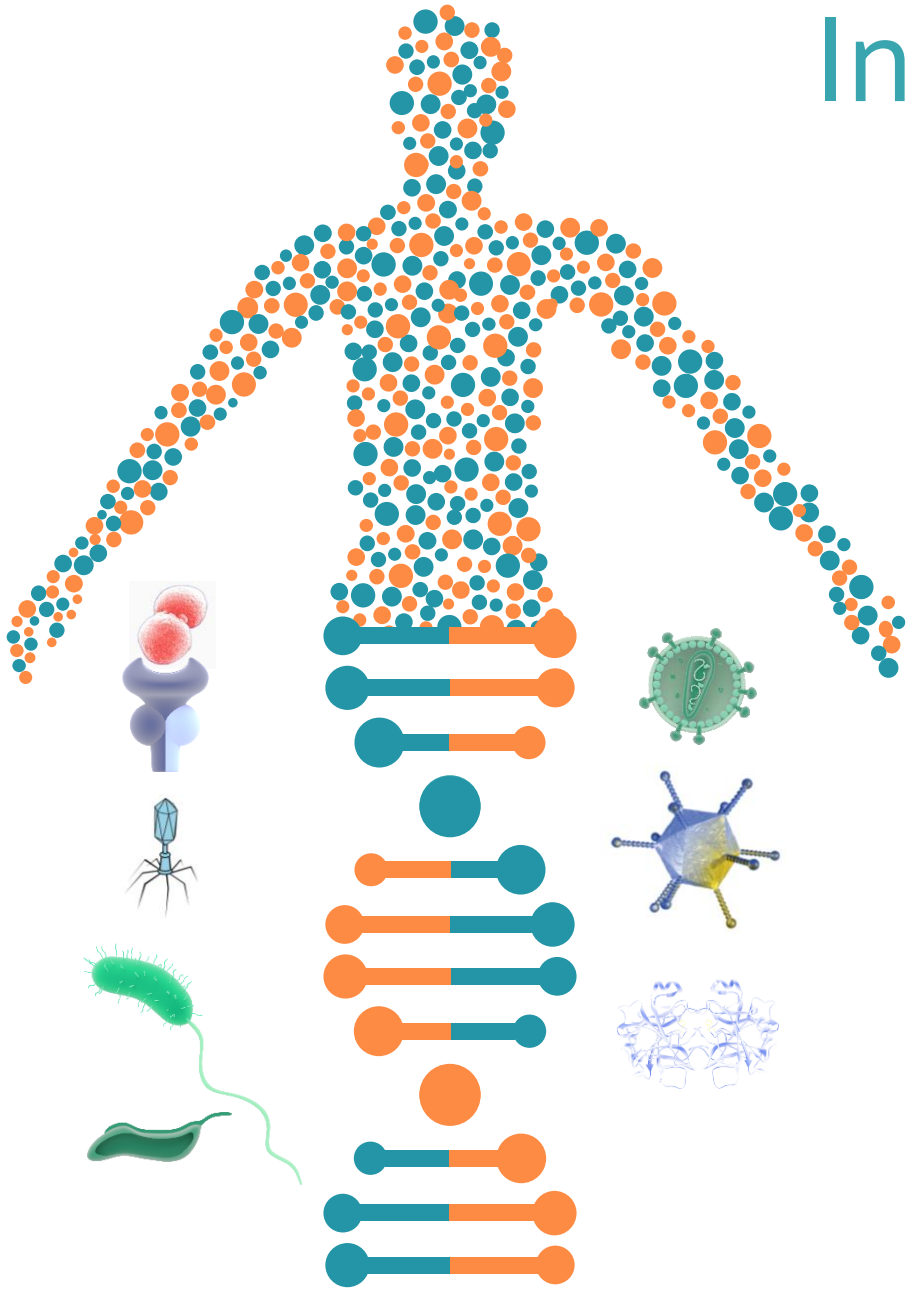


In an ideal situation, we like our science simple…



But nature is complex





Integrated genomics concept

- The immune system's ability to distinguish self from non-self is a critical process for our health.
- Breakdown of checkpoints for the "self-nonself" discrimination results in the hundred different autoimmune diseases
- Technological advances in computing and genomics, and the reducing cost of sequencing have opened-up possibilities to "cut through" the immunological complexity of clinical disease to identify causal genes, molecules, and cellular mechanisms.

Malfunctioning immune system has been associated with the development or progression of several human diseases.

Cardiometabolic disease

- Obesity
- Diabetes
- CVD /CKD
- Asthma
- coronary artery disease
- Heart disease
- stroke

Inflammation

- IBD
- Ulcerative colitis
- Crohn's disease
- Multiple sclerosis
- Systemic lupus erythematosus
- Arthritis

Others

- Cancer
- Alzheimer's
- Autoimmune diseases
- Parkinson's
- Allergies
- Graves' disease
- Coeliac disease

Genomics workflow projects: planning

Incentive
Identify the user needs
Establish objectives
Identify stakeholders
Discuss expected outcomes
Sketch workflows

Interviews
Best-performance often highly contextual
Identify technical requirements
Face to face meetings

Plan individual needs before-hand
Source for different test datasets

Research
Summarize available tools and licensing
Map features to technical requirements
Follow-up to clear ambiguities

Prototype
Implement coding aspects
Collaborate with engineers
Iterate and self-test
Modify- with version control
Self-test again

External testing
Collaboration is key at this stage
Open channels for communication
Set up a project issues-reporting

Use real-life data to test functionality
Improve and scale accordingly

Deployment
Deliver final product in live environment
Plan for maintenance
Plan for features update/addition

Genomics workflow projects: IT infrastructure

Developer

Determine resources needed
Profiling and Timing Code
Tracing & visualisation

Scalable strategies
Shared-Memory Multicore Architecture
Special hardware

Portability and platform scalability
Multi-Node HPC scalability
Cloud scalability
Container Scalability

Support

Kind of support
Full-time support
occasional support

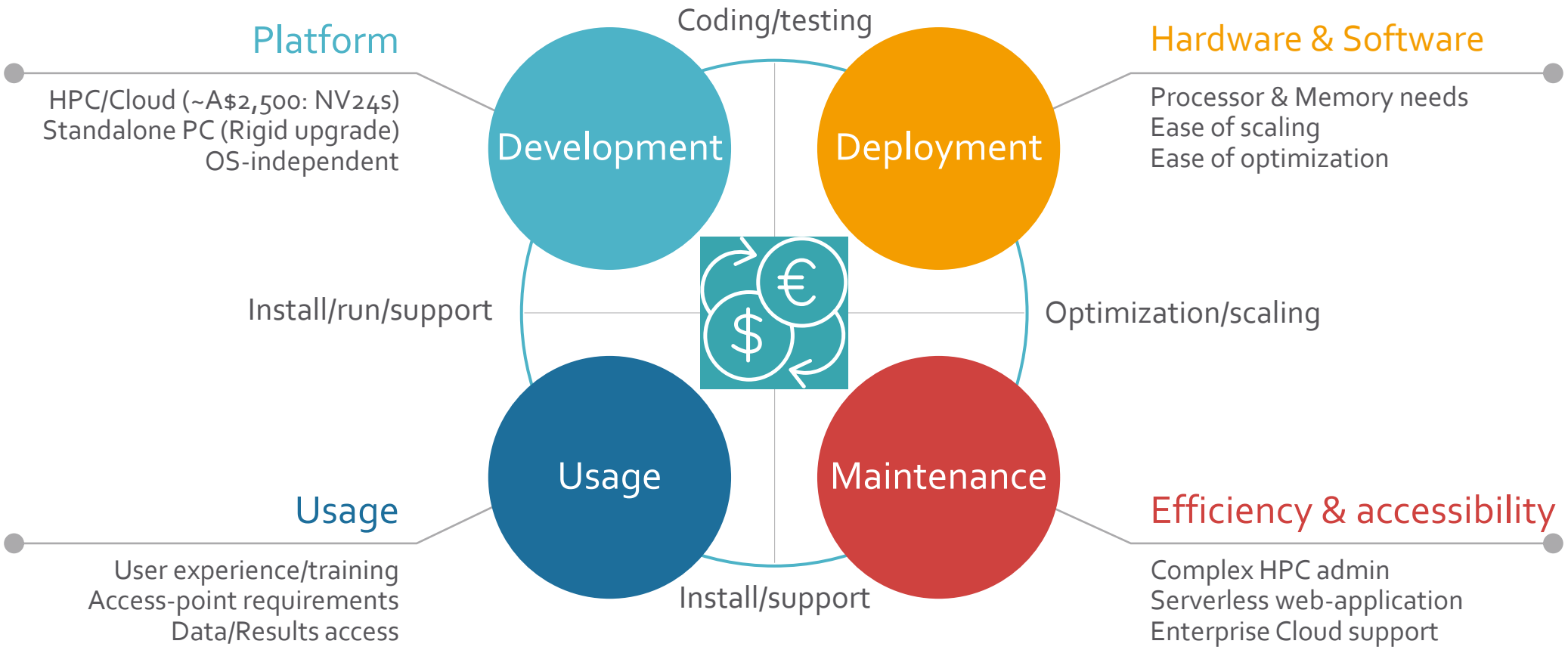
Efficiency
Serverless web services
HPC/Cloud services

Level of support
MPI tools require seasoned engineers to handle memory control, data communication, and task synchronization

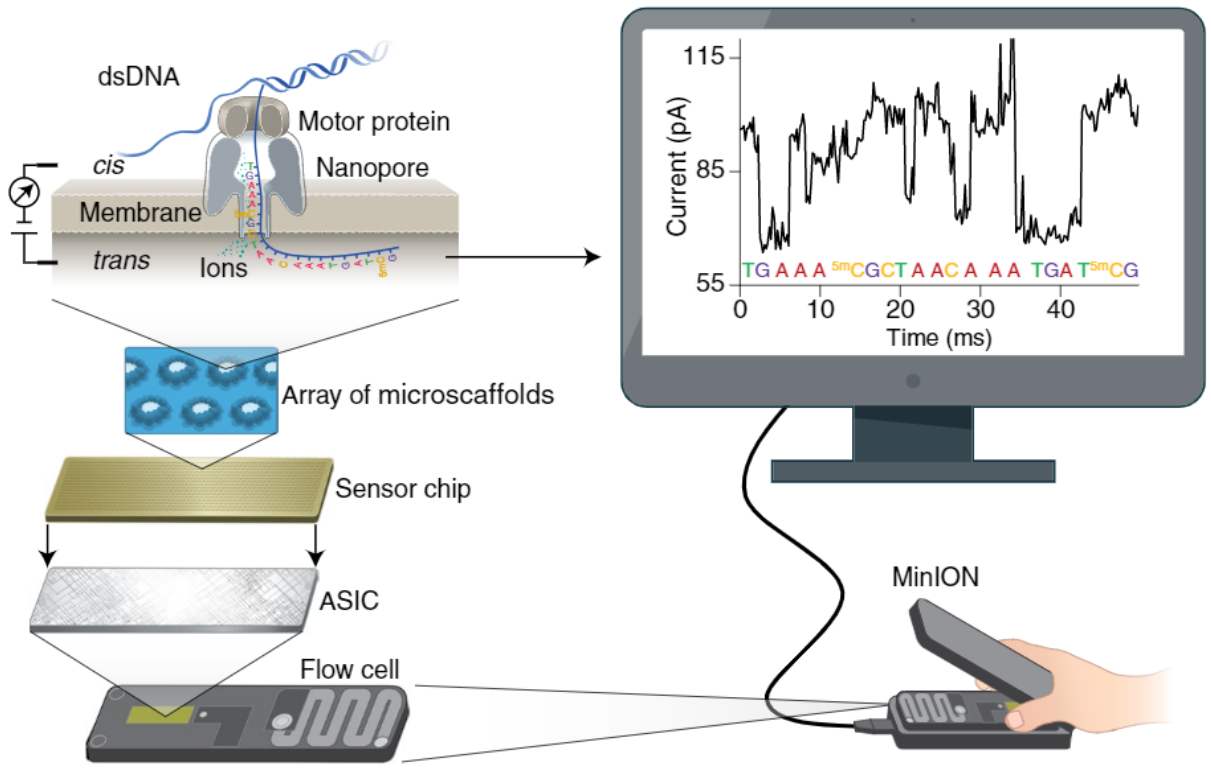
Parallelization	Scalability and optimization	Integration & portability
How much compute power is needed?	Can we extend beyond primary state?	Can we use existing resources?

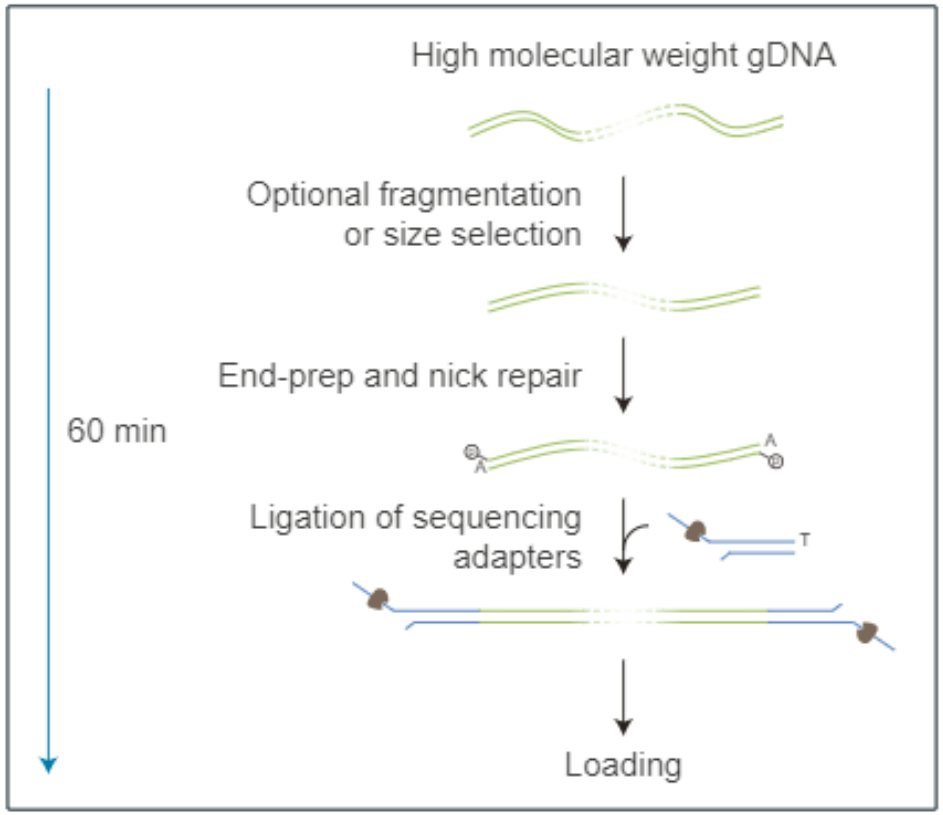
Shi L, Wang Z. Computational Strategies for Scalable Genomics Analysis. Genes. 2019; 10(12):1017. https://doi.org/10.3390/genes10121017

Genomics workflow projects: Budget



Genome sequencing: Oxford Nanopore Technology



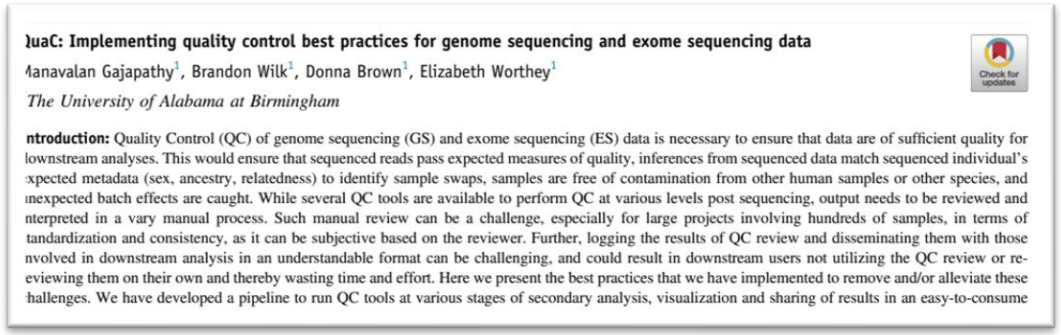


Genome sequencing: Oxford Nanopore Technology

SINGLE MOLECULE SEQUENCING

ENABLES REAL-TIME ANALYSIS

DIRECT RNA SEQUENCING

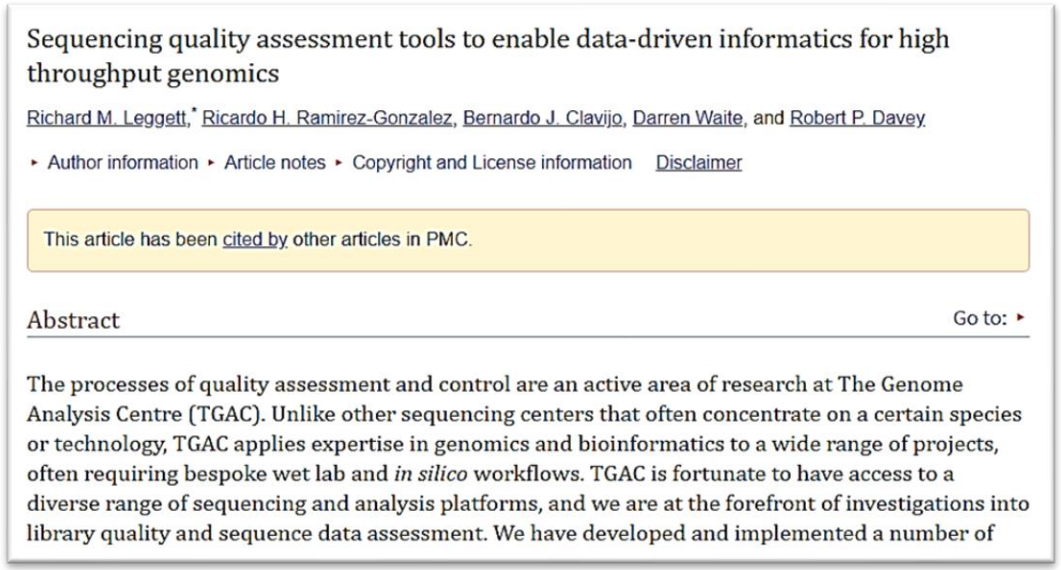


luaC: Implementing quality control best practices for genome sequencing and exome sequencing data

lanavalan Gajapathy[1], Brandon Wilk[1], Donna Brown[1], Elizabeth Worthey[1]

The University of Alabama at Birmingham

ntroduction: Quality Control (QC) of genome sequencing (GS) and exome sequencing (ES) data is necessary to ensure that data are of sufficient quality for lownstream analyses. This would ensure that sequenced reads pass expected measures of quality, inferences from sequenced data match sequenced individual's xpected metadata (sex, ancestry, relatedness) to identify sample swaps, samples are free of contamination from other human samples or other species, and nexpected batch effects are caught. While several QC tools are available to perform QC at various levels post sequencing, output needs to be reviewed and nterpreted in a vary manual process. Such manual review can be a challenge, especially for large projects involving hundreds of samples, in terms of tandardization and consistency, as it can be subjective based on the reviewer. Further, logging the results of QC review and disseminating them with those nvolved in downstream analysis in an understandable format can be challenging, and could result in downstream users not utilizing the QC review or re-eviewing them on their own and thereby wasting time and effort. Here we present the best practices that we have implemented to remove and/or alleviate these hallenges. We have developed a pipeline to run QC tools at various stages of secondary analysis, visualization and sharing of results in an easy-to-consume



Sequencing quality assessment tools to enable data-driven informatics for high throughput genomics

Richard M. Leggett,* Ricardo H. Ramirez-Gonzalez, Bernardo J. Clavijo, Darren Waite, and Robert P. Davey

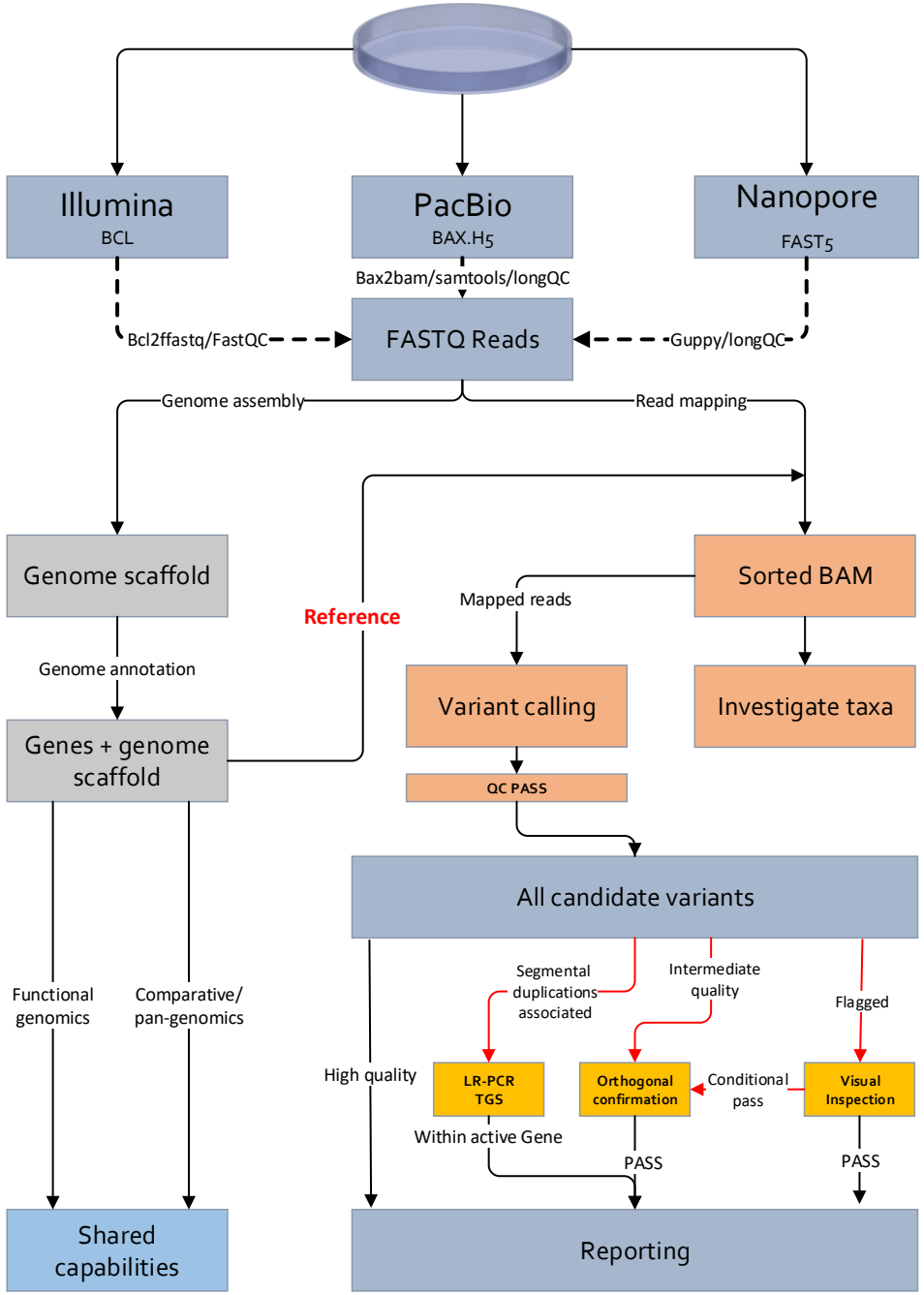
▸ Author information ▸ Article notes ▸ Copyright and License information Disclaimer

This article has been cited by other articles in PMC.

Abstract Go to: ▸

The processes of quality assessment and control are an active area of research at The Genome Analysis Centre (TGAC). Unlike other sequencing centers that often concentrate on a certain species or technology, TGAC applies expertise in genomics and bioinformatics to a wide range of projects, often requiring bespoke wet lab and *in silico* workflows. TGAC is fortunate to have access to a diverse range of sequencing and analysis platforms, and we are at the forefront of investigations into library quality and sequence data assessment. We have developed and implemented a number of

- Virtually all downstream analysis and interpretation processes rely on accurate data QC and validation.
- Most genomics workflows are interconnected on the critical reads pre-processing steps.



Genome assembly

Quality trim and filter

Assess quality
- FastQC (NGS)
- LongQC/Falco (TGS)

Trimming and filtering
- Trimmomatic
- Canu2

Assess quality
- FastQC (NGS)
- LongQC/Falco (TGS)

FASTQ

Assembly

De novo assembly
- Canu2
- SPAdes
- Unicycler
- SMRT
- SKESA

Reference-guided
- SPAdes
- Ragoo

FASTA

Scaffolding

Sort by reference
- RaGoo/RagTag
- Optical mapping
- Nullarbor (Pipeline for public health microbiology involving bacterial data)

Sorted-FASTA

Annotation

De novo
- Augustus
- GenMark-ES
- GlimmerHMM
- SNAP
- Prokka
- Funannotate
- RAST

Evidence-based
- Apollo web
- Augustus
- GlimmerHMM

GFF3 + FATSA

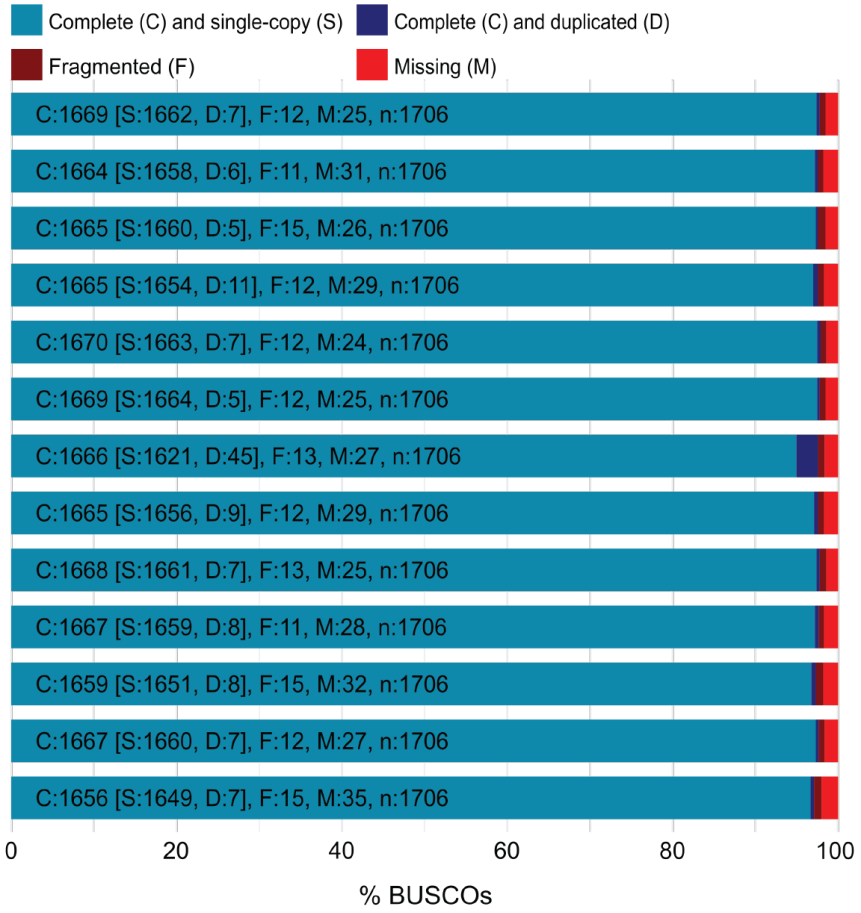
Publishing

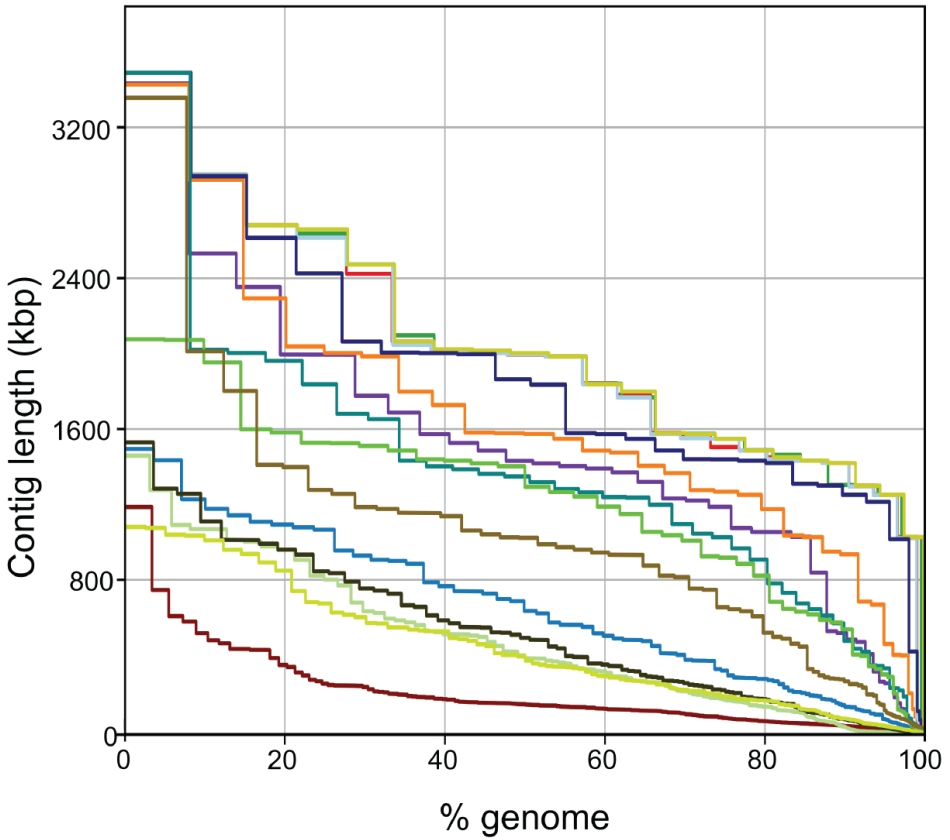
Convert to ncbi/ena/jgi
- NCBI tbl2asn
- Geneious ™

SQN/GBFF/GBK

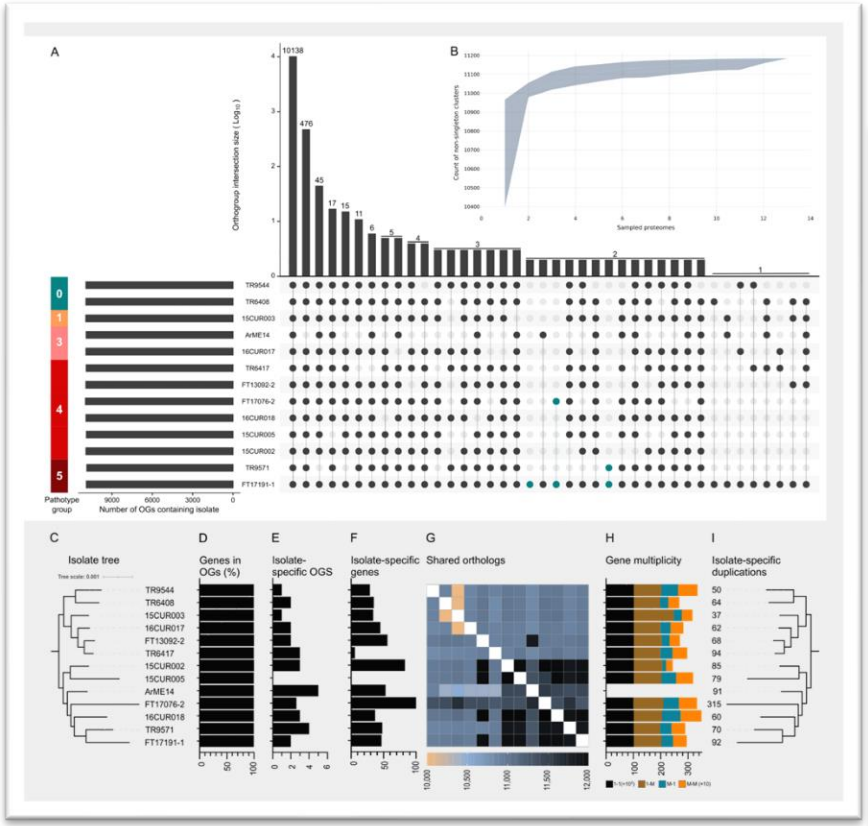
https://bioinformaticshome.com/tools/wga/wga.html

Genome assembly: quality assessment

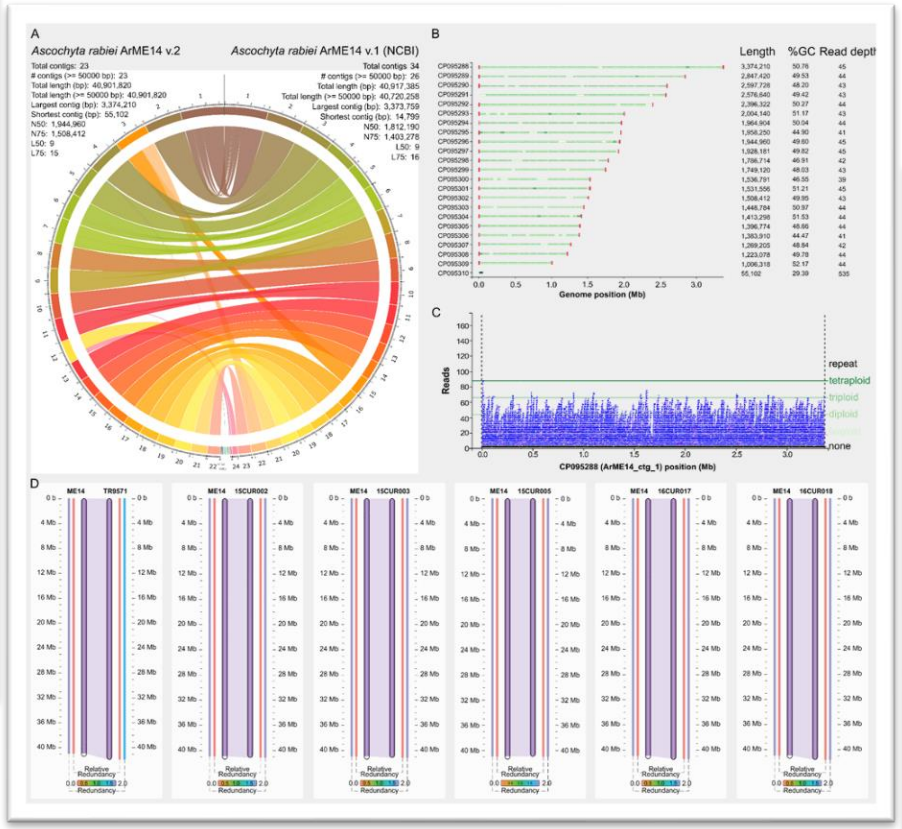




Comparative analysis

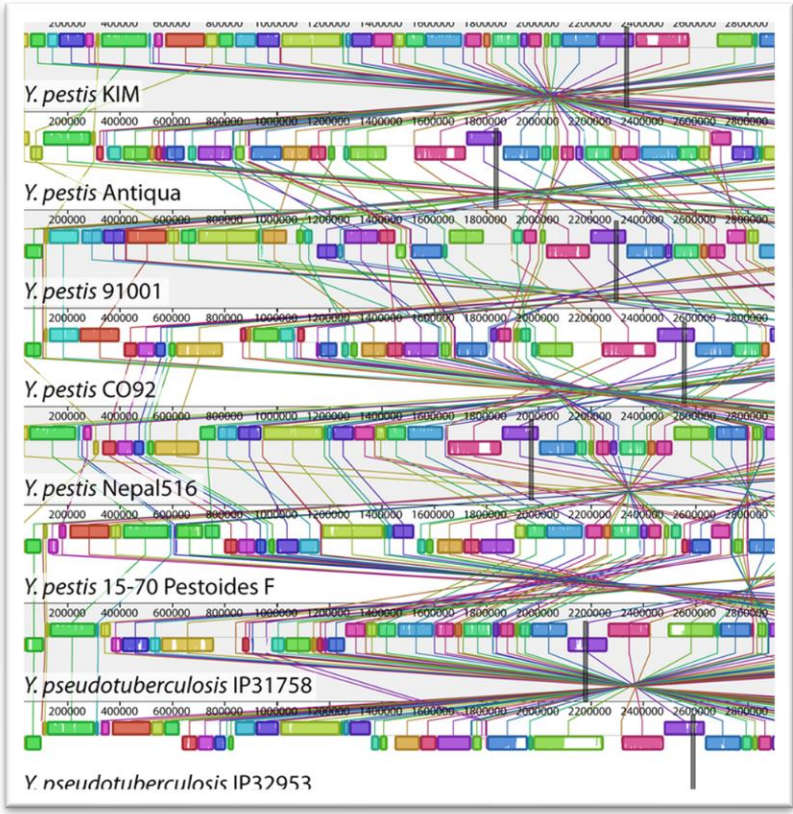


Orthologs



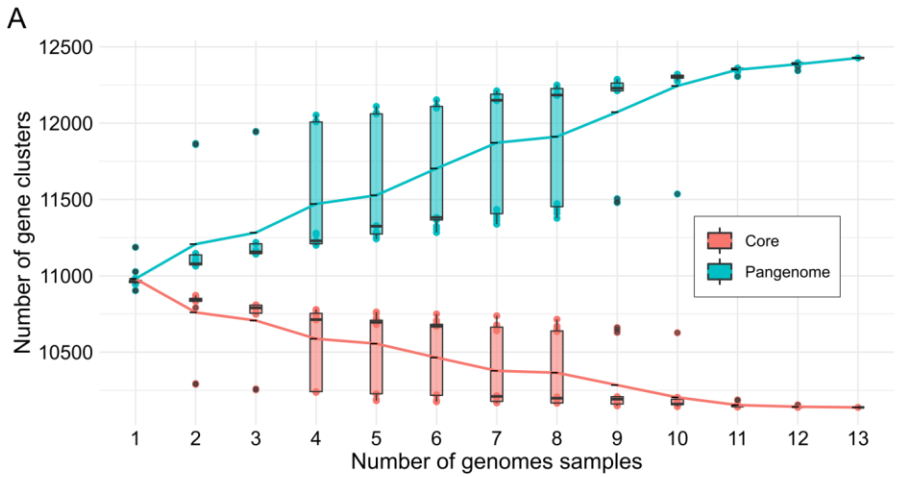
Genome structure

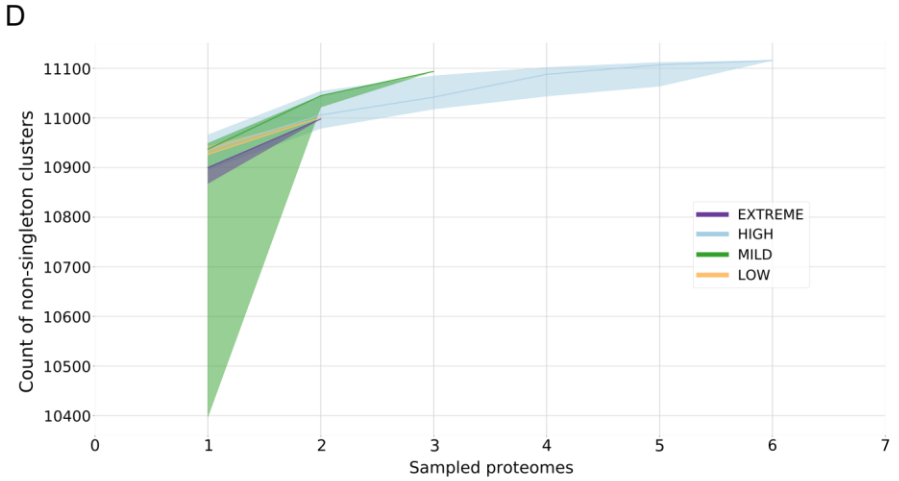
Whole genome alignment

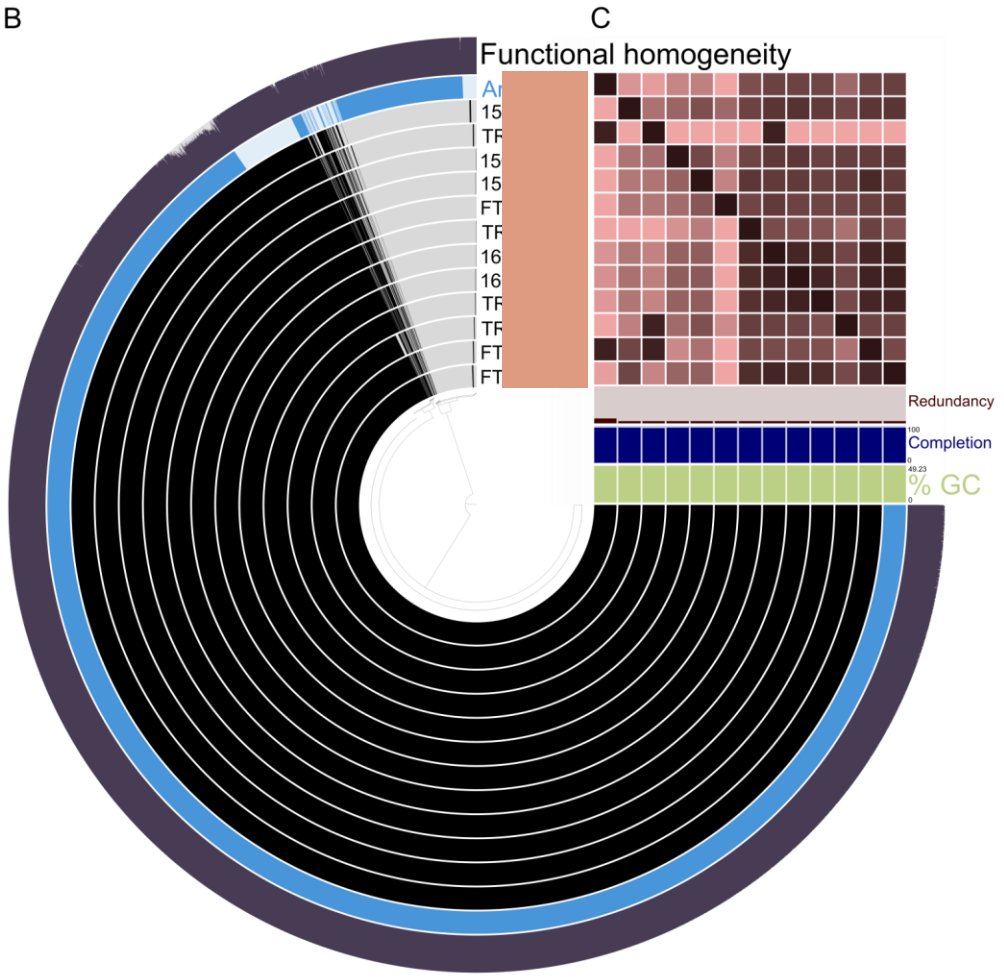


PLOS Genetics. DOI:10.1371/journal.pgen.1000128

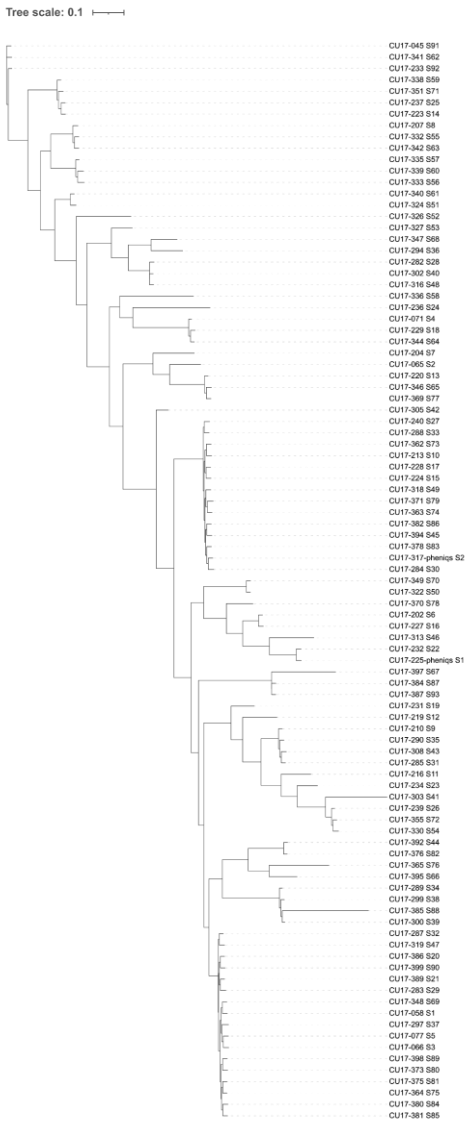
Pan-genomics

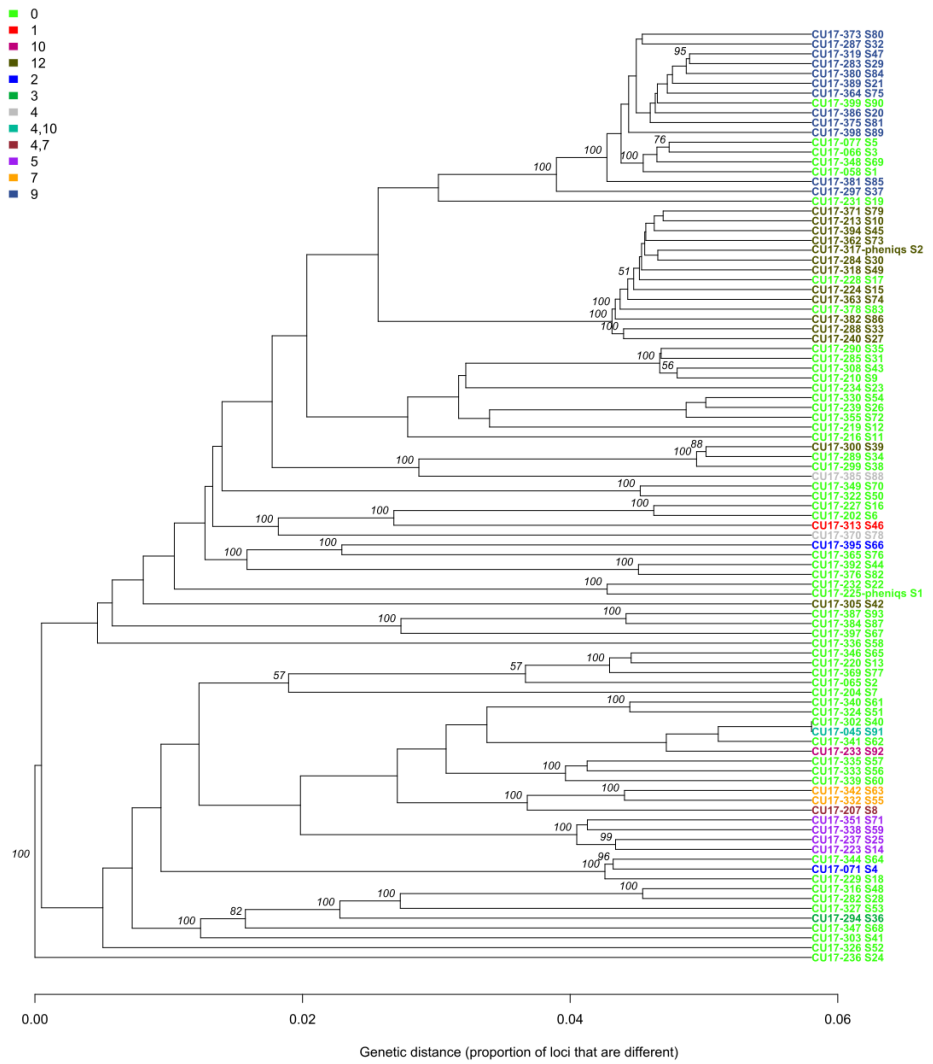






Phylogenetic analysis





Variant calling

Quality trim and filter

Assess quality
- FastQC (NGS)
- LongQC/Falco (TGS)

Trimming and filtering
- Trimmomatic
- Canu2

Assess quality
- FastQC (NGS)
- LongQC/Falco (TGS)

FASTQ

Align to reference

Index reference
- Aligner-specific

Align QC reads
- BWA (splice-unaware)
- STAR; Hisat2

Mark duplicates
- Picard; Sambamba
- GATK

Compress SAM output
- Samtools

BAM

Alignment cleanup

Sort by coordinates
- Samtools sort

Index BAM (accessible)
- Samtools index

Assess BAM Quality

Coverage depth; %mapped; %unmapped; MapQ score
- Samtools flagstat
- QualiMap; BEDtools

Sample quality
- KING; VerifyBamID

Sorted-BAM

Variant calling

Generate VCF
- GATK; FreeBayes; BCFtools (SNV/indels)
- VarScan2; DeepSNV (Somatic)
- CoNVEX; ExomeCNV (CNV)
- InDelible; Sniffles; SVMerge (SV)

Filter VCF & annotate

#PASS; #QUAL; #DP; SNV; INDELS
CDS; Non-CDS(Intron; Silent, UTR)
- GATK; BCFtools
- VCFannotate; AWK; ALFA

VCF

Variant reporting

Functional implications
- AnnoVAR; ENSEMBL VEP; SnpEff
- Clinical data, family history, signaling pathways

Disease/drug-associated
- ClinVar; ClinPred; Clinotator >> JIGV report

- Variant Annotation Integrator - UCSC Genome Browser
- Overall pipeline run summaries: MultiQC

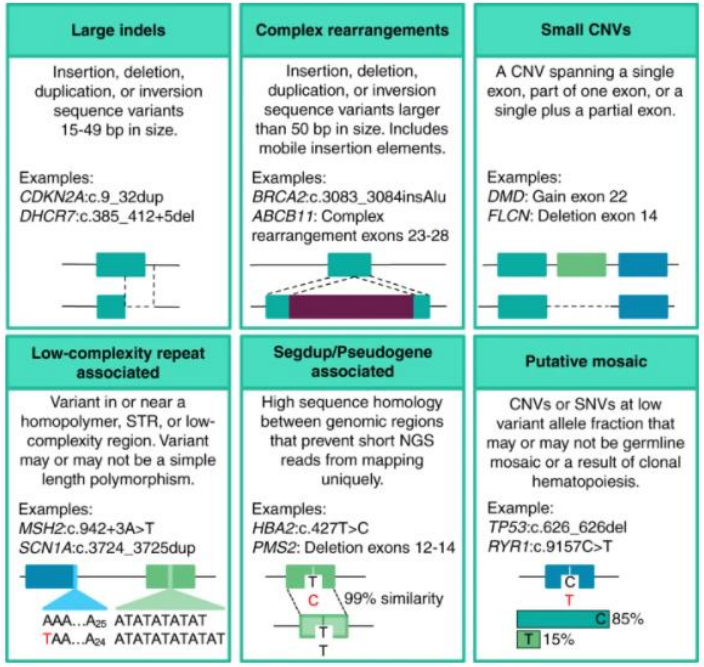
VCF;HTML;SVG;TXT

Pipelining in NetFlow:
- Fast to prototype and has scalability without extra code.
- Resumable tasks with continuous checkpoints.
- Portable (AZURE/AWS/HPC), reproducible, multi-language.
- Supports software containers and use of existing code "as is".
- Internal parallelization eliminates resources wastage.
- Focused resources-allocation to compute-intensive processes

Di Tommaso et al. Nat Biotechnol. 35(4):316-319. doi: 10.1038/nbt.3820. PMID: 28398311.

Expecting the expected…

Fig. 1: Technically challenging variant types.



Variants were categorized as being technically challenging, or not, based on these six criteria. Note that some variants could be considered challenging for multiple reasons (e.g., a single-exon deletion within a segmentally duplicated region). Examples provided are variants observed in the prevalence analysis. Detailed criteria are provided in the Supplemental Methods. CNV copy-number variant, indels, insertions or deletions, NGS next-generation sequencing, Segdup, segmental duplication, SNVs single-nucleotide variants, STR short tandem repeat.

Despite advances in technology, some variants are still complex to predict reliably

ELSEVIER

The Journal of Molecular Diagnostics

Volume 21, Issue 2, March 2019, Pages 318-329

Regular article

A Rigorous Interlaboratory Examination of the Need to Confirm Next-Generation Sequencing–Detected Variants with an Orthogonal Method in Clinical Genetic Testing

Stephen E. Lincoln *, Rebecca Truty *, Chiao-Feng Lin †, ‡, Justin M. Zook §, Joshua Paul *, Vincent H. Ramey *, Marc Salit §, ¶, Heidi L. Rehm †, ‡, ‖, **, ††, Robert L. Nussbaum *, ‡‡, Matthew S. Lebo †, ‡, **, ††

Show more

+ Add to Mendeley Share Cite

Standardized confirmation protocols are still in their infancy stages

BJHaem
BRITISH JOURNAL OF HAEMATOLOGY
British Society for Haematology

Review | Full Access

Integrating germline variant assessment into routine clinical practice for myelodysplastic syndrome and acute myeloid leukaemia: current strategies and challenges

Kiran Tawana, Anna L. Brown, Jane E. Churpek

First published: 18 October 2021 | https://doi.org/10.1111/bjh.17855 | Citations: 1

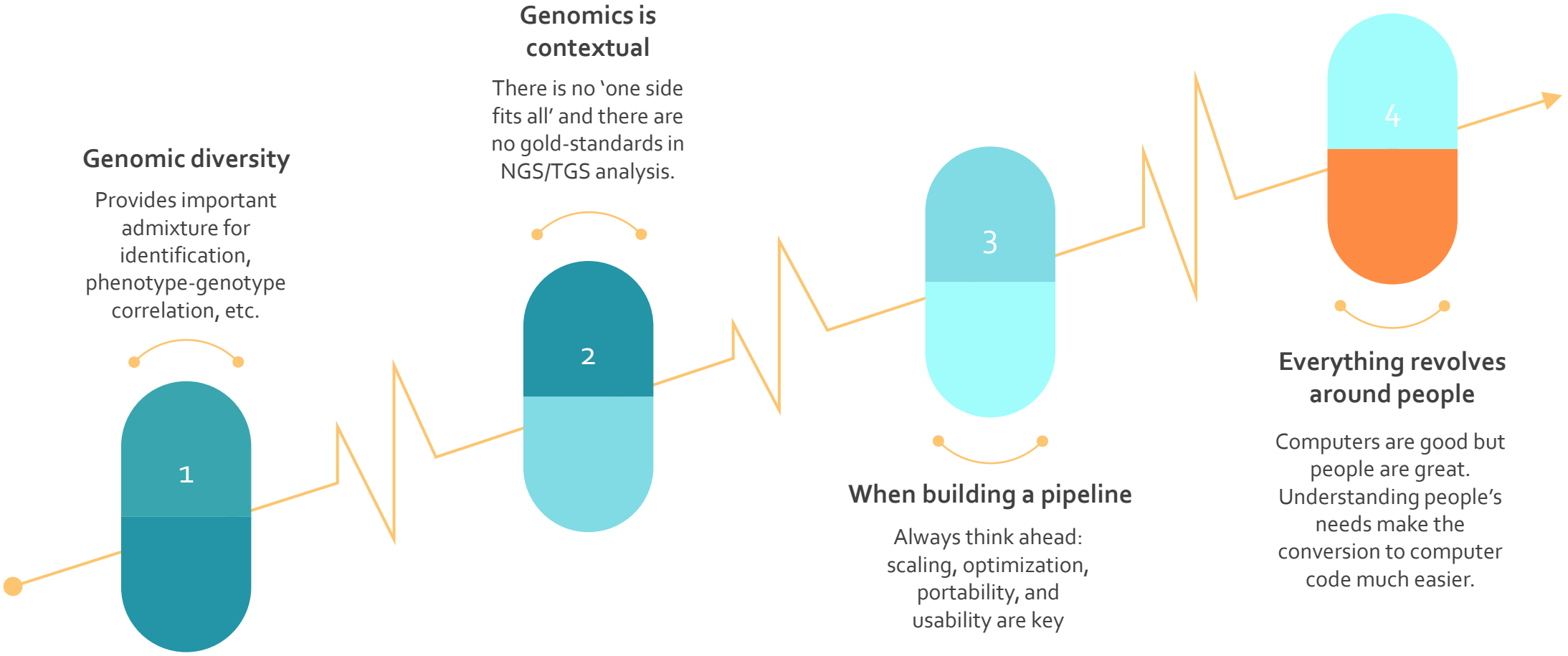
SECTIONS PDF TOOLS SHARE

Summary

Over the last decade, the field of hereditary haematological malignancy syndromes (HHMSs) has gained increasing recognition among clinicians and scientists worldwide. Germline mutations now account for almost 10% of adult and paediatric myelodysplasia/acute myeloid leukaemia (MDS/AML). As our ability to diagnose HHMSs has improved, we are now faced with the challenges of integrating these advances into routine clinical practice for patients with MDS/AML and how to optimise management and surveillance of patients and asymptomatic carriers. Discoveries of novel syndromes combined with clinical, genetic and epigenetic profiling of tumour samples, have highlighted unique patterns of disease evolution across HHMSs. Despite these advances, causative lesions are detected in less than half of familial cases and evidence-based guidelines are often lacking, suggesting there is much still to learn. Future research efforts are needed to sustain current momentum within the field, led not only by advancing genetic technology but essential collaboration between clinical and academic communities.

Cross-platform variant curation, universal specialist knowledge & metadata limited

Summarising everything



Genomic diversity

Provides important admixture for identification, phenotype-genotype correlation, etc.

Genomics is contextual

There is no 'one side fits all' and there are no gold-standards in NGS/TGS analysis.

When building a pipeline

Always think ahead: scaling, optimization, portability, and usability are key

Everything revolves around people

Computers are good but people are great. Understanding people's needs make the conversion to computer code much easier.

“It’s one small step for man, one giant leap for mankind.”

- Neil Armstrong

Skills (Proteomics & transcriptomics)

ILRI/BecA (CGIAR)

- East Coast fever is a lymphoproliferative disease caused by the tick-borne protozoan parasite *Theileria parva*.
- Sporozoite proteome defined by LC–MS/MS analysis.
- Raw MS/MS data files were analysed with Peaks software (Bioinformatics solutions) using a database containing all Uniprot database entries for R. appendiculatus and the re-annotated proteome of *T. parva*
- Proteins were classified according to their putative localization in the sporozoite using TargetP. Trans-membrane domains were predicted by TMHMM Server, signal peptides by SignalP and glycosylphosphatidylinositol (GPI)-anchor signal by GPI-SOM.
- We clustered all sporozoite protein coding sequences of *T. parva* with the complete predicted proteome of *Plasmodium falciparum* into putative orthologous .
- We identified proteins predicted to be orthologs of *Plasmodium falciparum* sporozoite surface molecules and invasion organelle proteins, and proteins that may contribute to the phenomenon of bovine lymphocyte transformation.

James Nyagwange [a, b], Edwin Tijhaar [b], Nicola Ternette [c], Fredrick Mobegi [f], Kyle Tretina [d], Joana C. Silva [d, e], Roger Pelle [a], Vishvanath Nene [a]

Show more

+ Add to Mendeley Share Cite

https://doi.org/10.1016/j.ijpara.2017.09.007

Get rights and content

Under a Creative Commons license

Open access

Highlights

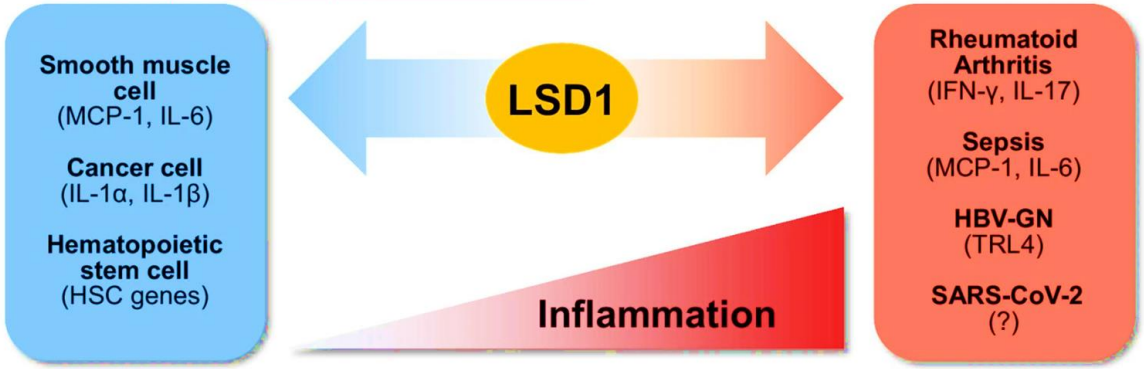
- 2007 *Theileria parva* proteins expressed in the sporozoite were identified.
- Proteins include known *T. parva* antigens targeted by antibodies and cytotoxic T cells.
- Proteins predicted to be orthologs of *Plasmodium falciparum* sporozoite surface molecules were identified.
- Proteins predicted to be orthologs of *P. falciparum* invasion organelle proteins were identified.
- Proteins that may contribute to the phenomenon of bovine lymphocyte transformation were identified.

Skills (Epigenetics & Transcriptomics)

Netherlands Cancer Institute & GSK: The dichotomy of LSD1

- Lysine specific demethylase 1 (LSD1) suppresses gene expression through demethylation of H3K4.
- LSD1 acts as a transcriptional coactivator when bound to the androgen receptor (AR) via demethylation of repressive histone marks at H3K9, resulting in the de-repression of AR target genes .

From: Roles of lysine-specific demethylase 1 (LSD1) in homeostasis and diseases



LSD1 functions as a key epigenetic regulator in inflammatory disease. LSD1 controls gene expression in two ways during inflammation. LSD1 increases the expression levels of inflammatory response genes by acting as a positive regulator in inflammatory diseases such as rheumatoid arthritis (RA), sepsis, hepatitis B virus-associated glomerulonephritis (HBV-GN), and SARS-CoV-2 infection. By contrast, LSD1 functions as a negative regulator of the inflammatory response that decreases the expression of cytokine genes in smooth muscle cells (SMCs), cancer cells, and hematopoietic stem cells (HSCs)

Kim et al (2021). J Biomed Sci 28(41). https://doi.org/10.1186/s12929-021-00737-3

Haematologica. 2019 Jun; 104(6): 1156–1167. PMCID: PMC6545850
Prepublished online 2018 Dec 4. doi: 10.3324/haematol.2018.199190 PMID: 30514804

Lysine specific demethylase 1 inactivation enhances differentiation and promotes cytotoxic response when combined with all-*trans* retinoic acid in acute myeloid leukemia across subtypes

Kimberly N. Smitheman,[1] Tesa M. Severson,[2] Satyajit R. Rajapurkar,[1] Michael T. McCabe,[1] Natalie Karpinich,[1] James Foley,[1] Melissa B. Pappalardi,[1] Ashley Hughes,[3] Wendy Halsey,[3] Elizabeth Thomas,[3] Christopher Traini,[3] Kelly E. Federowicz,[1] Jenny Laraio,[1] Fredrick Mobegi,[2] Geraldine Ferron-Brady,[4] Rabinder K. Prinjha,[1] Christopher L. Carpenter,[1] Ryan G. Kruger,[1] Lodewyk Wessels,[2,5] and Helai P. Mohammad[1]

▸ Author information ▸ Article notes ▸ Copyright and License information Disclaimer

- LSD1 is a known regulator of normal haematopoiesis.
- LSD1 inhibition promotes differentiation of leukemic cells.
- Therapeutic potential of a selective, potent inhibitor of LSD1, GSK2879552,in combination with all-trans retinoic acid in relapsed refractory AML patients was explored.
  - RNA-seq (Active treatments vs DMSO) :: DESeq2
  - ChIP-seq (H3K4me1, H3K4me2, H3K4me3, H3K9ac, H3K27ac vs DMSO) :: Peaks call (MACS2); differential analysis (DESeq2); nearest-gene + peak intersection (BEDTools).

(clinicaltrials.gov identifiers: 02273102, 02717884, 02842827)

Skills (WGS; Metagenomics; Comparative)

ELSEVIER

Journal of Cystic Fibrosis

Volume 20, Issue 3, May 2021, Pages 413-420

The cystic fibrosis gut as a potential source of multidrug resistant pathogens

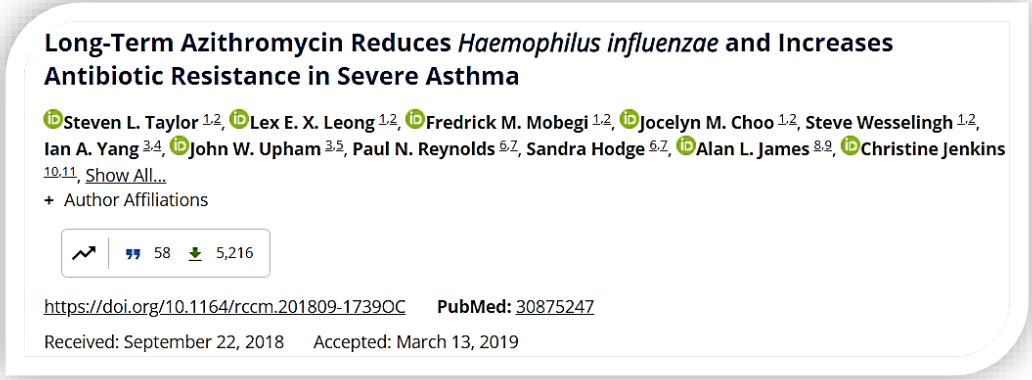
Steven L. Taylor [a, b, 1], Lex E.X. Leong [c, 1], Sarah K. Sims [a, b], Rebecca L. Keating [d], Lito E. Papanicolas [a, b], Alyson Richard [a, b], Fredrick M. Mobegi [a, b], Steve Wesselingh [b], Lucy D. Burr [d, e, 2], Geraint B. Rogers [a, b, 2]

Mechanisms linking low-calorie sweeteners to impaired glycaemic control

D Kreuch, K Ivey, F M Mobegi, L Leong, N J Isaacs, N Pezos, M Horowitz, C K Rayner, G B Rogers, R L Young

Microbiome And Host Health, Intestinal Nutrient Sensing

Research output: Contribution to journal › Article › peer-review



Long-Term Azithromycin Reduces *Haemophilus influenzae* and Increases Antibiotic Resistance in Severe Asthma

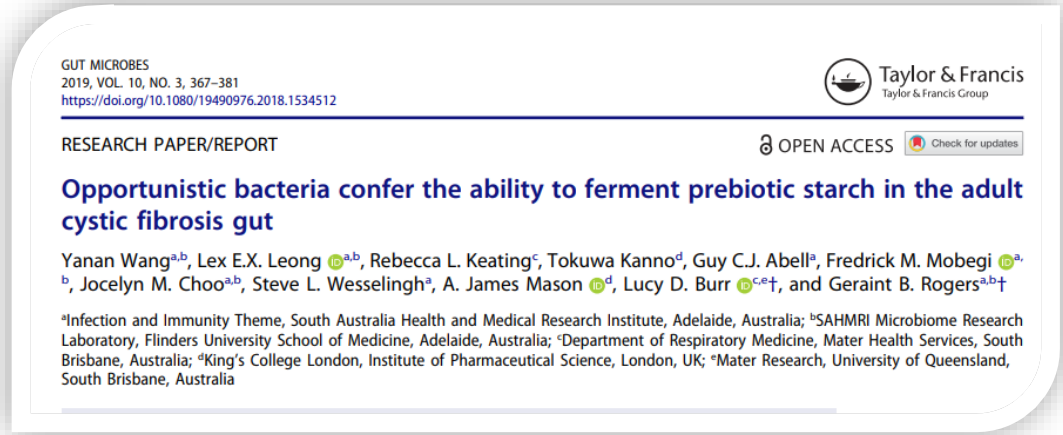
Steven L. Taylor [1,2], Lex E. X. Leong [1,2], Fredrick M. Mobegi [1,2], Jocelyn M. Choo [1,2], Steve Wesselingh [1,2], Ian A. Yang [3,4], John W. Upham [3,5], Paul N. Reynolds [6,7], Sandra Hodge [6,7], Alan L. James [8,9], Christine Jenkins [10,11], Show All...

+ Author Affiliations

58 5,216

https://doi.org/10.1164/rccm.201809-1739OC **PubMed:** 30875247

Received: September 22, 2018 Accepted: March 13, 2019



GUT MICROBES
2019, VOL. 10, NO. 3, 367–381
https://doi.org/10.1080/19490976.2018.1534512

Taylor & Francis
Taylor & Francis Group

RESEARCH PAPER/REPORT

OPEN ACCESS Check for updates

Opportunistic bacteria confer the ability to ferment prebiotic starch in the adult cystic fibrosis gut

Yanan Wang[a,b], Lex E.X. Leong [a,b], Rebecca L. Keating[c], Tokuwa Kanno[d], Guy C.J. Abell[a], Fredrick M. Mobegi [a,b], Jocelyn M. Choo[a,b], Steve L. Wesselingh[a], A. James Mason [d], Lucy D. Burr [c,e†], and Geraint B. Rogers[a,b†]

[a]Infection and Immunity Theme, South Australia Health and Medical Research Institute, Adelaide, Australia; [b]SAHMRI Microbiome Research Laboratory, Flinders University School of Medicine, Adelaide, Australia; [c]Department of Respiratory Medicine, Mater Health Services, South Brisbane, Australia; [d]King's College London, Institute of Pharmaceutical Science, London, UK; [e]Mater Research, University of Queensland, South Brisbane, Australia

eLife

RESEARCH ARTICLE

Intestinal microbiology shapes population health impacts of diet and lifestyle risk exposures in Torres Strait Islander communities

Fredrick M Mobegi[1,2], Lex EX Leong[1], Fintan Thompson[1,3], Sean M Taylor[3], Linton R Harriss[3], Jocelyn M Choo[1,2], Steven L Taylor[1,2], Steve L Wesselingh[4], Robyn McDermott[3,5], Kerry L Ivey[1,6,7†*], Geraint B Rogers[1,2†*]

Skills (WGS &Transcriptomics, NGS, TGS)

Curtin University: Characterization fungal isolates

- Comprehensive characterization of the biology and survival of fungal pathogens is a prerequisite to developing more effective disease management strategies.
- At the CCDM, I lead the bioinformatics analyses (in two projects) focusing on Ascochyta blight of pulses.
- We employ whole-genome sequencing using nanopore, PacBio, and Illumina technologies, RNA-seq, and targeted genotyping to study Ascochyta interactions with the hosts and identify novel mechanisms of resistance and pathogenicity.
- These datasets provide unmatched resolution at a molecular level into various mechanisms of pathogenicity and resistance, allowing us to identify genetic strengths and vulnerabilities that can be exploited in breeding programs.

Research | Open Access | Published: 08 May 2021

Analysis of differentially expressed *Sclerotinia sclerotiorum* genes during the interaction with moderately resistant and highly susceptible chickpea lines

Virginia W. Mwape, Fredrick M. Mobegi, Roshan Regmi, Toby E. Newman, Lars G. Kamphuis & Mark C. Derbyshire

BMC Genomics **22**, Article number: 333 (2021) | Cite this article

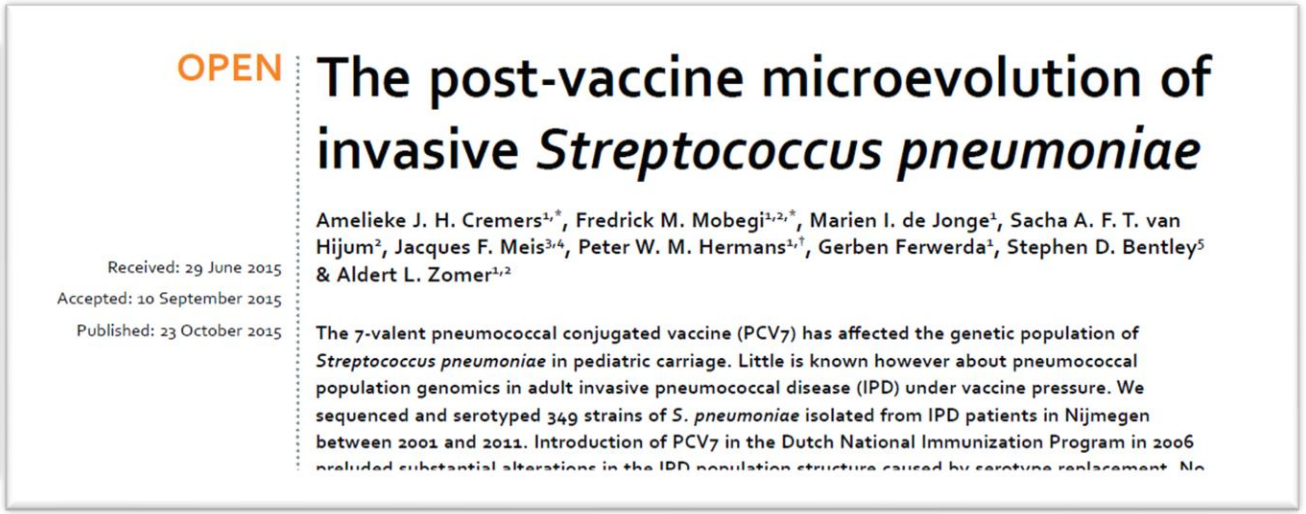
1063 Accesses | **1** Citations | **5** Altmetric | Metrics

Skills (Molecular epidemiology)

RESEARCH Open Access

Investigating potential transmission of antimicrobial resistance in an open-plan hospital ward: a cross-sectional metagenomic study of resistome dispersion in a lower middle-income setting

Anushia Ashokan[1,2,3], Josh Hanson[4,5], Ne Myo Aung[5,6], Mar Mar Kyi[5,6], Steven L. Taylor[1,2], Jocelyn M. Choo[1,2], Erin Flynn[1,2], Fredrick Mobegi[1,2], Morgyn S. Warner[3,7], Steve L. Wesselingh[8], Mark A. Boyd[3] and Geraint B. Rogers[1,2*]



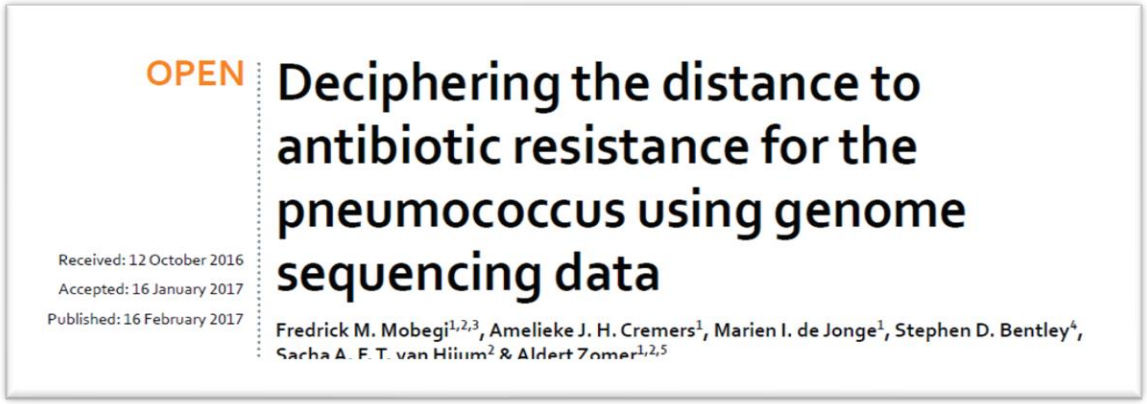
OPEN

The post-vaccine microevolution of invasive *Streptococcus pneumoniae*

Received: 29 June 2015
Accepted: 10 September 2015
Published: 23 October 2015

Amelieke J. H. Cremers[1,*], Fredrick M. Mobegi[1,2,*], Marien I. de Jonge[1], Sacha A. F. T. van Hijum[2], Jacques F. Meis[3,4], Peter W. M. Hermans[1,†], Gerben Ferwerda[1], Stephen D. Bentley[5] & Aldert L. Zomer[1,2]

The 7-valent pneumococcal conjugated vaccine (PCV7) has affected the genetic population of *Streptococcus pneumoniae* in pediatric carriage. Little is known however about pneumococcal population genomics in adult invasive pneumococcal disease (IPD) under vaccine pressure. We sequenced and serotyped 349 strains of *S. pneumoniae* isolated from IPD patients in Nijmegen between 2001 and 2011. Introduction of PCV7 in the Dutch National Immunization Program in 2006 preluded substantial alterations in the IPD population structure caused by serotype replacement. No



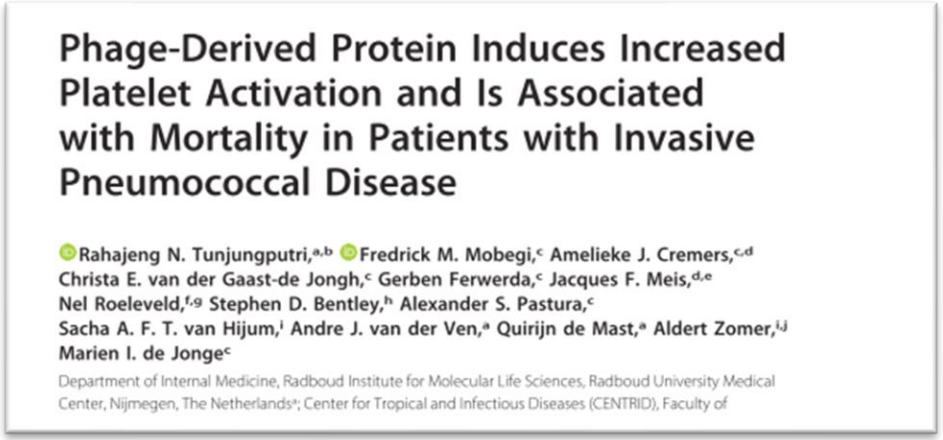
OPEN

Deciphering the distance to antibiotic resistance for the pneumococcus using genome sequencing data

Received: 12 October 2016
Accepted: 16 January 2017
Published: 16 February 2017

Fredrick M. Mobegi[1,2,3], Amelieke J. H. Cremers[1], Marien I. de Jonge[1], Stephen D. Bentley[4], Sacha A. F. T. van Hijum[2] & Aldert Zomer[1,2,5]

Skills (Genotype-phenotype association)



Phage-Derived Protein Induces Increased Platelet Activation and Is Associated with Mortality in Patients with Invasive Pneumococcal Disease

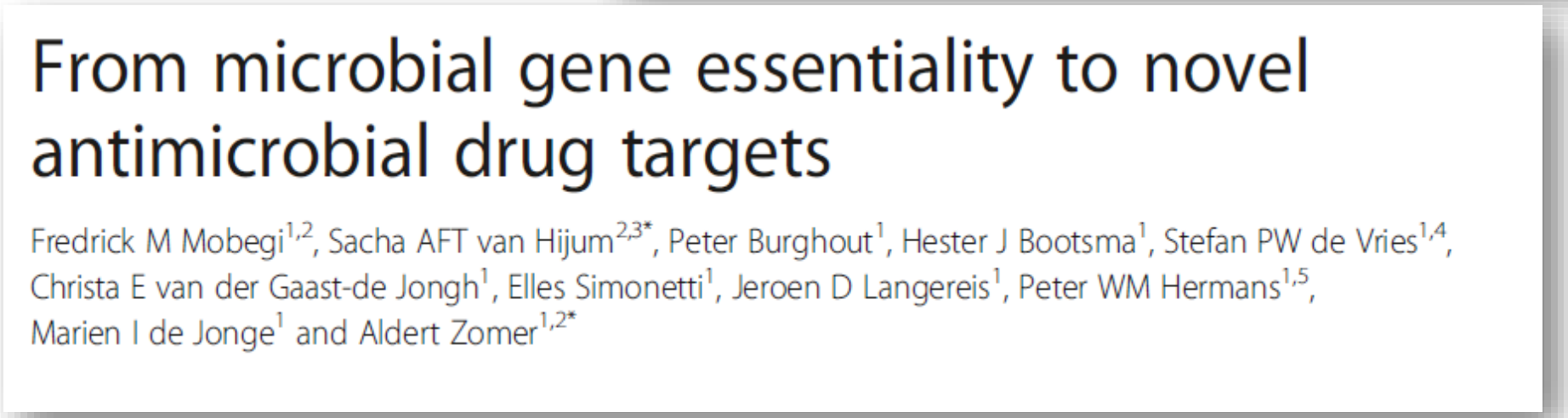
Rahajeng N. Tunjungputri,[a,b] Fredrick M. Mobegi,[c] Amelieke J. Cremers,[c,d] Christa E. van der Gaast-de Jongh,[c] Gerben Ferwerda,[c] Jacques F. Meis,[d,e] Nel Roeleveld,[f,g] Stephen D. Bentley,[h] Alexander S. Pastura,[c] Sacha A. F. T. van Hijum,[i] Andre J. van der Ven,[a] Quirijn de Mast,[a] Aldert Zomer,[i,j] Marien I. de Jonge[c]

Department of Internal Medicine, Radboud Institute for Molecular Life Sciences, Radboud University Medical Center, Nijmegen, The Netherlands[a]; Center for Tropical and Infectious Diseases (CENTRID), Faculty of

The Contribution of Genetic Variation of *Streptococcus pneumoniae* to the Clinical Manifestation of Invasive Pneumococcal Disease

Amelieke J. H. Cremers,[1,2,3] Fredrick M. Mobegi,[1,2,4] Christa van der Gaast–de Jongh,[1,2] Michelle van Weert,[1,2] Fred J. van Opzeeland,[1,2] Minna Vehkala,[5] Mirjam J. Knol,[6] Hester J. Bootsma,[6] Niko Välimäki,[5] Nicholas J. Croucher,[7] Jacques F. Meis,[8] Stephen Bentley,[9] Sacha A. F. T. van Hijum,[2,4,10] Jukka Corander,[5,9,11] Aldert L. Zomer,[12] Gerben Ferwerda,[1,2] and Marien I. de Jonge[1,2]

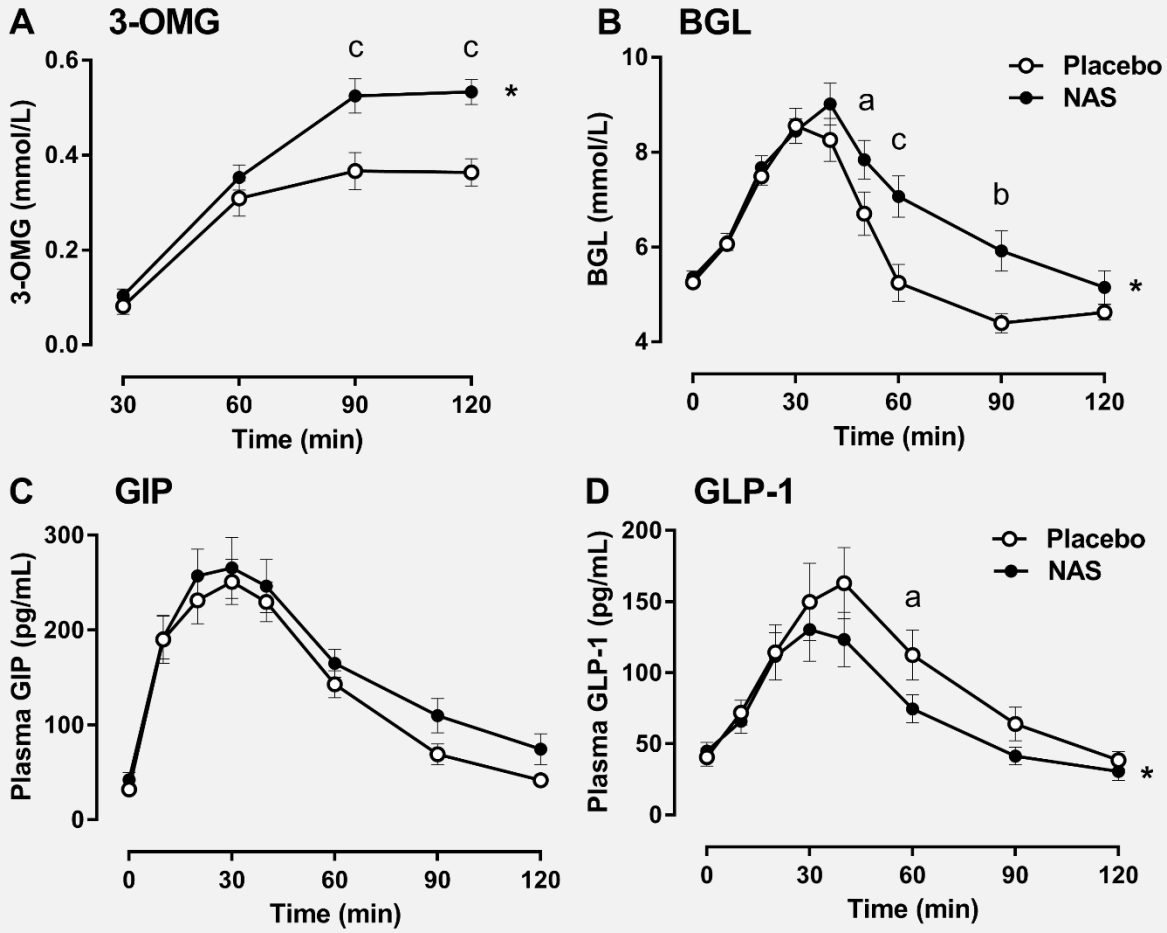
[1]Section of Pediatric Infectious Diseases, Laboratory of Medical Immunology, Radboud Institute for Molecular Life Sciences, [2]Radboud Center for Infectious Diseases, [3]Department of Medical Microbiology, and [4]Bacterial Genomics Group, Center for Molecular and Biomolecular Informatics, Radboudumc, Nijmegen, The Netherlands; [5]Department of Mathematics and Statistics, University of Helsinki, Finland; [6]Centre for Infectious Disease Control, National Institute for Public Health and the Environment (RIVM), Bilthoven, The Netherlands; [7]Medical Research Council Centre for Outbreak Analysis and Modelling, Department of Infectious Disease Epidemiology, Imperial College London, United Kingdom; [8]Department of Medical Microbiology and Infectious Diseases, Canisius-Wilhelmina Hospital, Nijmegen, The Netherlands; [9]Wellcome Trust Sanger Institute, Pathogen Genomics Group, Hinxton, Cambridge, United Kingdom; [10]NIZO, Ede, The Netherlands; [11]Department of Biostatistics, University of Oslo, Norway; and [12]Department of Infectious Diseases and Immunology, Faculty of Veterinary Medicine, Utrecht University, The Netherlands

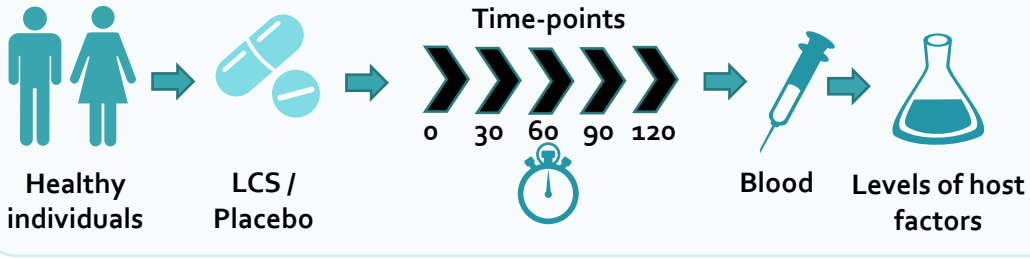


From microbial gene essentiality to novel antimicrobial drug targets

Fredrick M Mobegi[1,2], Sacha AFT van Hijum[2,3*], Peter Burghout[1], Hester J Bootsma[1], Stefan PW de Vries[1,4], Christa E van der Gaast-de Jongh[1], Elles Simonetti[1], Jeroen D Langereis[1], Peter WM Hermans[1,5], Marien I de Jonge[1] and Aldert Zomer[1,2*]

Artificial sweeteners: the microbiome perspective





Contribution of LCS to dysglycaemia.

- Regular high intake of beverages sweetened with **low-calorie sweeteners (LCS)** increase the risk of developing type 2 diabetes mellitus (T2DM)
- Underlying mechanisms remain unknown.
- 2-week intervention (randomised double-blind experiment) supplementation with a LCS combination (92 mg sucralose + 52 mg acesulfame-K, equivalent to ~1.5L of diet beverage consumption/day) or placebo

RBG and Insulin

Intake of LCS caused increased blood sugar and insulin levels in healthy non-diabetic subjects.

3-OMG

Intake of LCS caused augmented glucose absorption (3-OMG) in healthy non-diabetic subjects.

GLP-1 and GLP-2

Intake pf LCS caused attenuated release of glucagon-like peptide-1 (GLP-1) levels in healthy non-diabetic subjects.

Does the gut microbiota play a role for LCS effects in humans just like in rodents?

RNA sequencing data analysis

Fredrick M. Mobegi, PhD

BHKi hybrid seminar series

linkedin.com/fmobegi

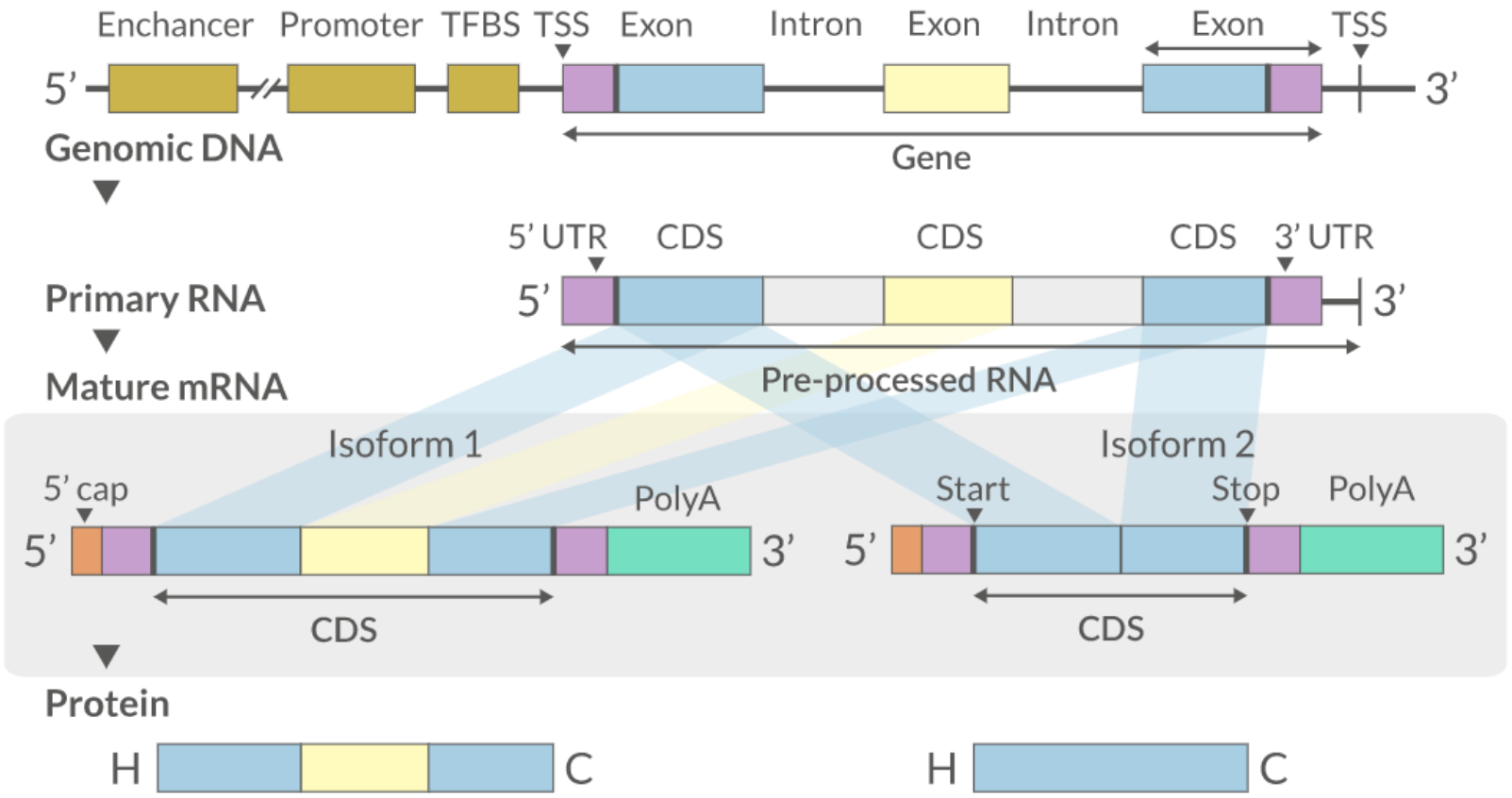
twitter.com/mobeginomics

github.com/fmobegi

rpubs.com/fmobegi

Contents

RNA sequencing



- The transcriptome is spatially and temporally dynamic
- Data comes from functional units (coding regions)
- Only a tiny fraction of the genome

Image credits: scilifelab

RNA-seq experiment design

Scotty: a web tool for designing RNA-Seq experiments to measure differential gene expression FREE

Michele A. Busby, Chip Stewart, Chase A. Miller, Krzysztof R. Grzeda, Gabor T. Marth
Author Notes

Bioinformatics, Volume 29, Issue 5, 1 March 2013, Pages 656–657,
https://doi.org/10.1093/bioinformatics/btt015
Published: 12 January 2013 **Article history**

PDF Split View Cite Permissions Share

Abstract

Motivation: A common question arises at the beginning of every experiment where RNA-Seq is used to detect differential gene expression between two conditions: How many reads should we sequence?

- Balanced experiment
- Biological replicates
- Statistical power
- Technical replicates

Scotty performs Power analysis with cost estimates

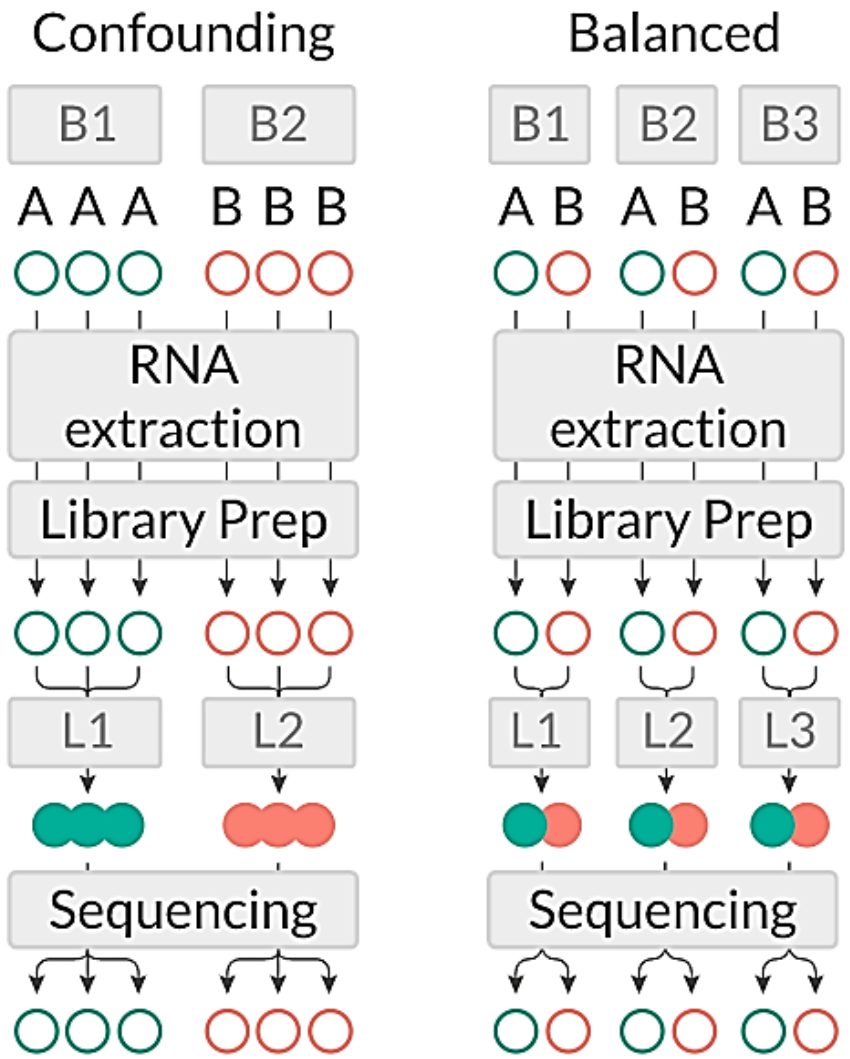
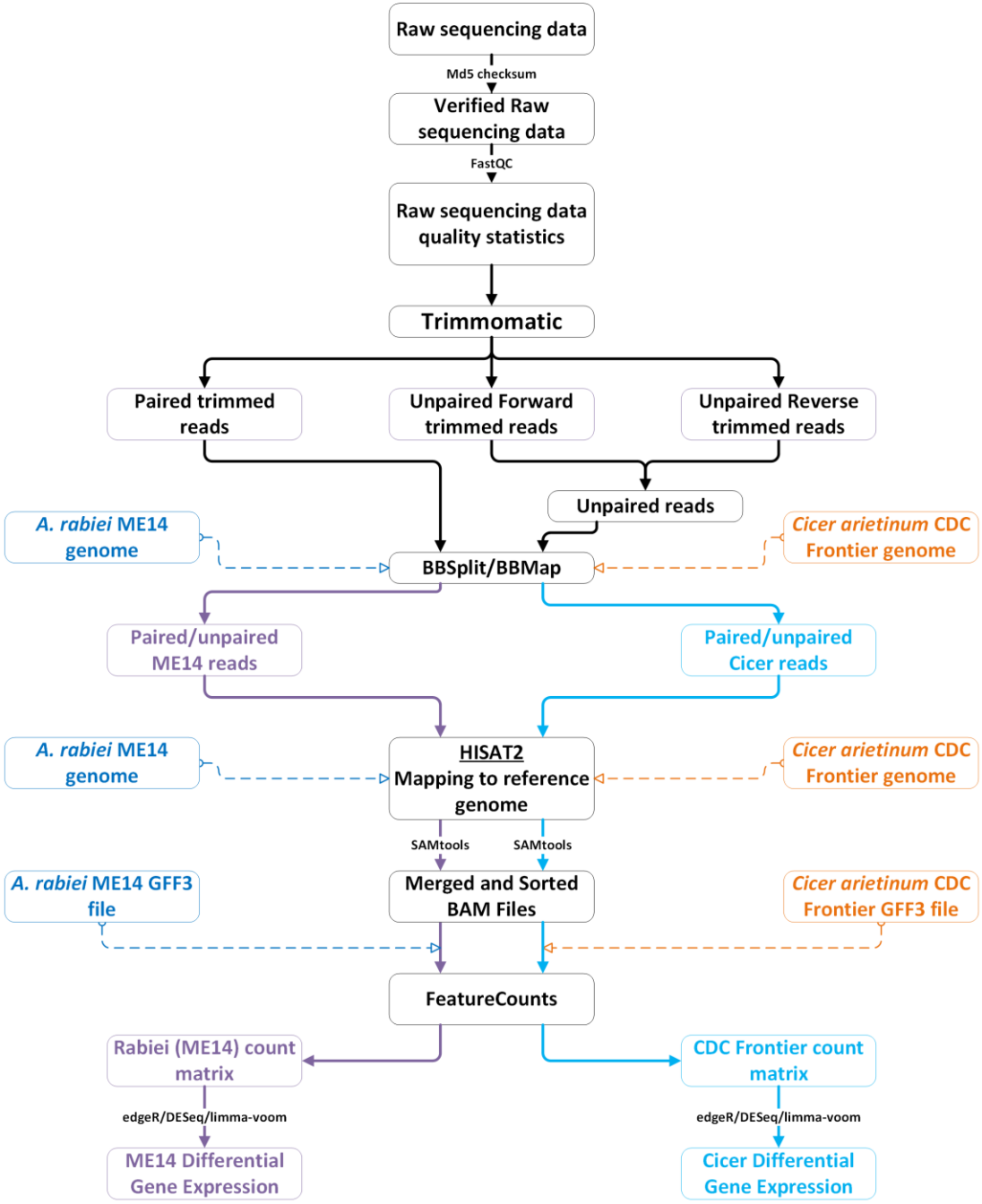


Image credits: scilifelab

RNA-seq data analysis pipeline



RNA-seq data analysis pipeline

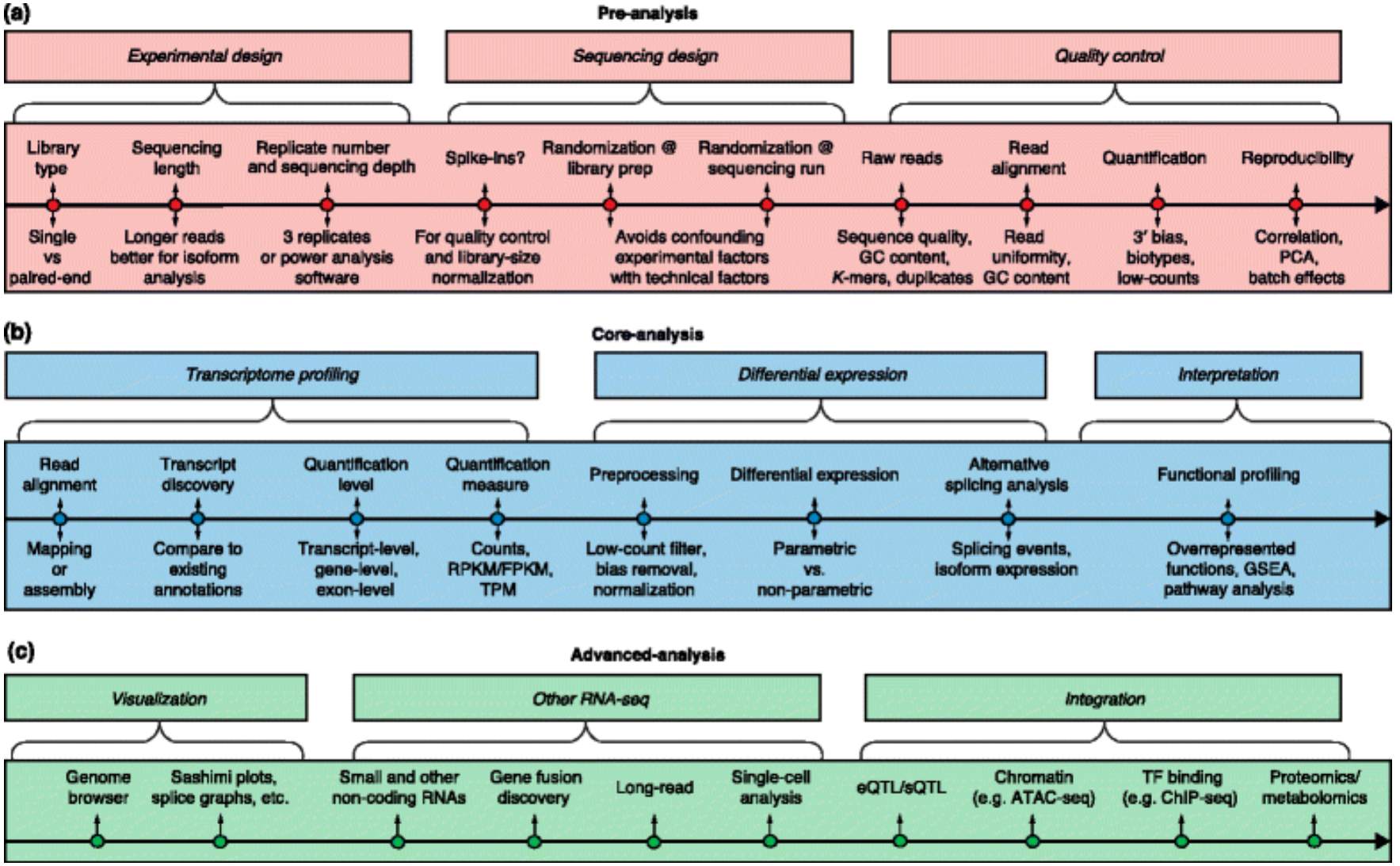
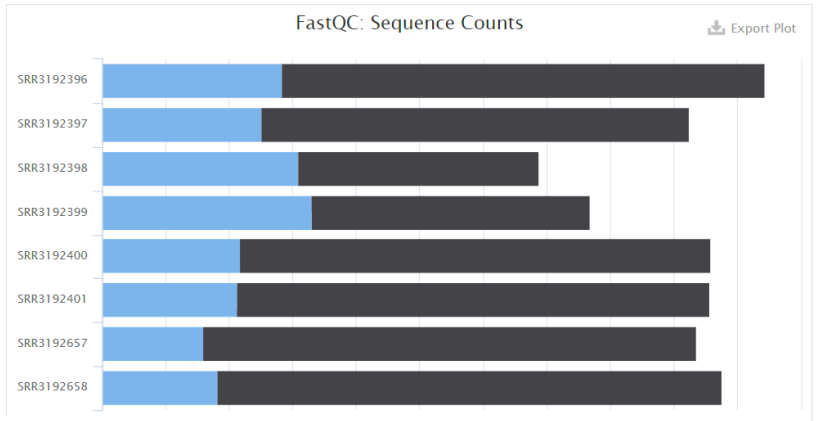
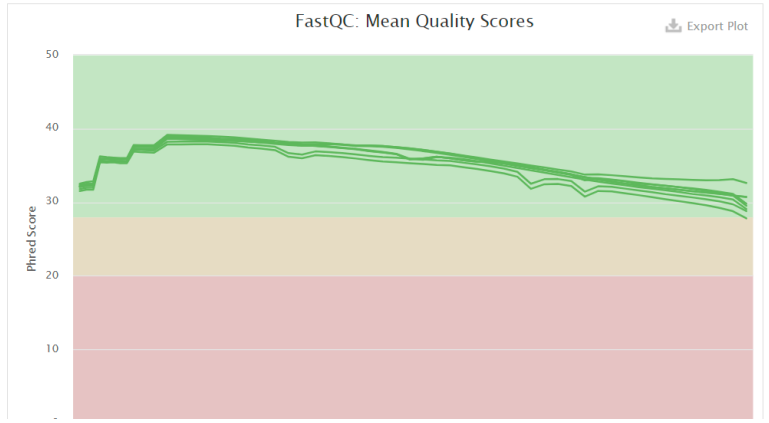


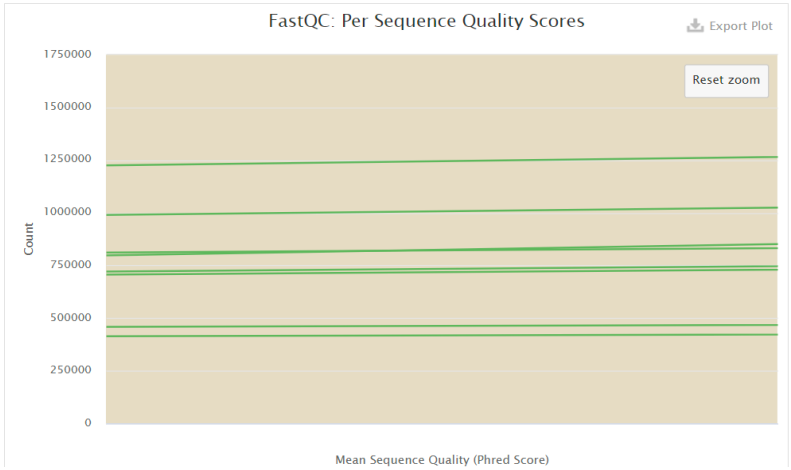
Figure 1: Conesa *et al*. A survey of best practices for RNA-seq data analysis. Genome Biology volume 17, Article number: 13 (2016)

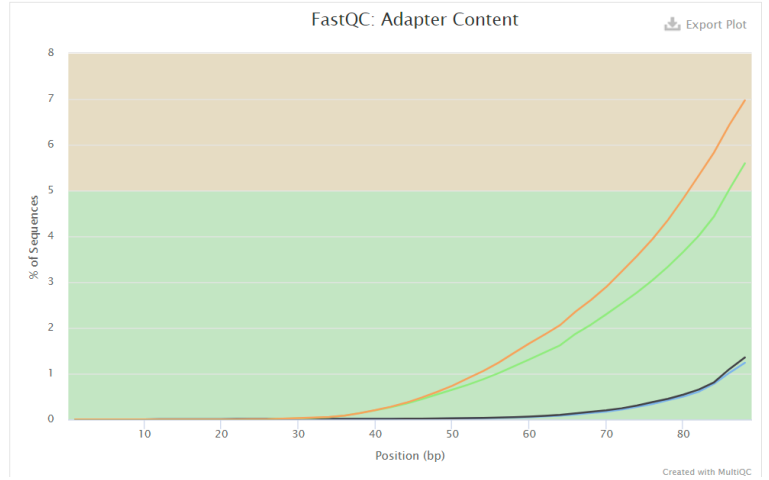
Quality assessment

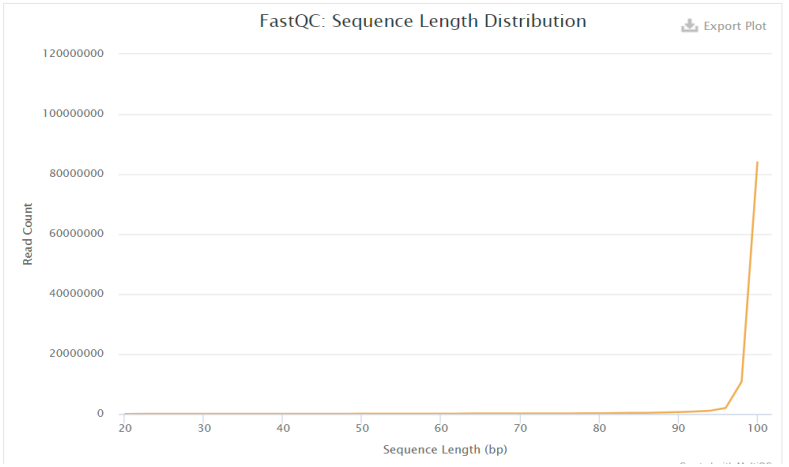
- MultiQC
- FastQC
- Trimmomatic
- CutAdapt

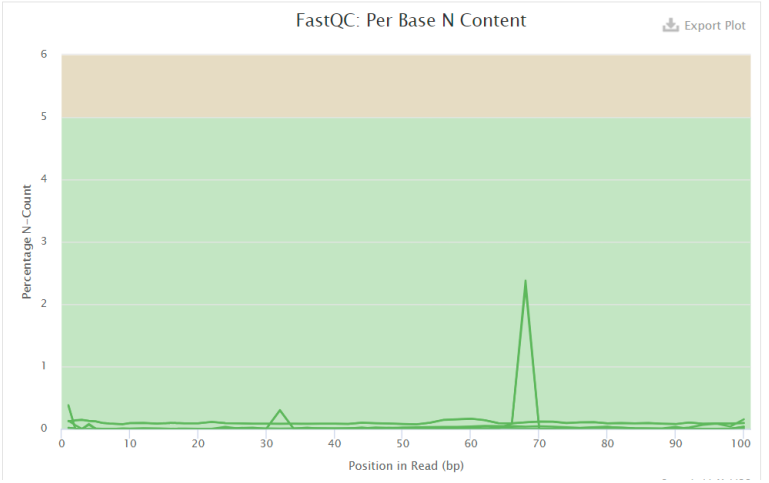




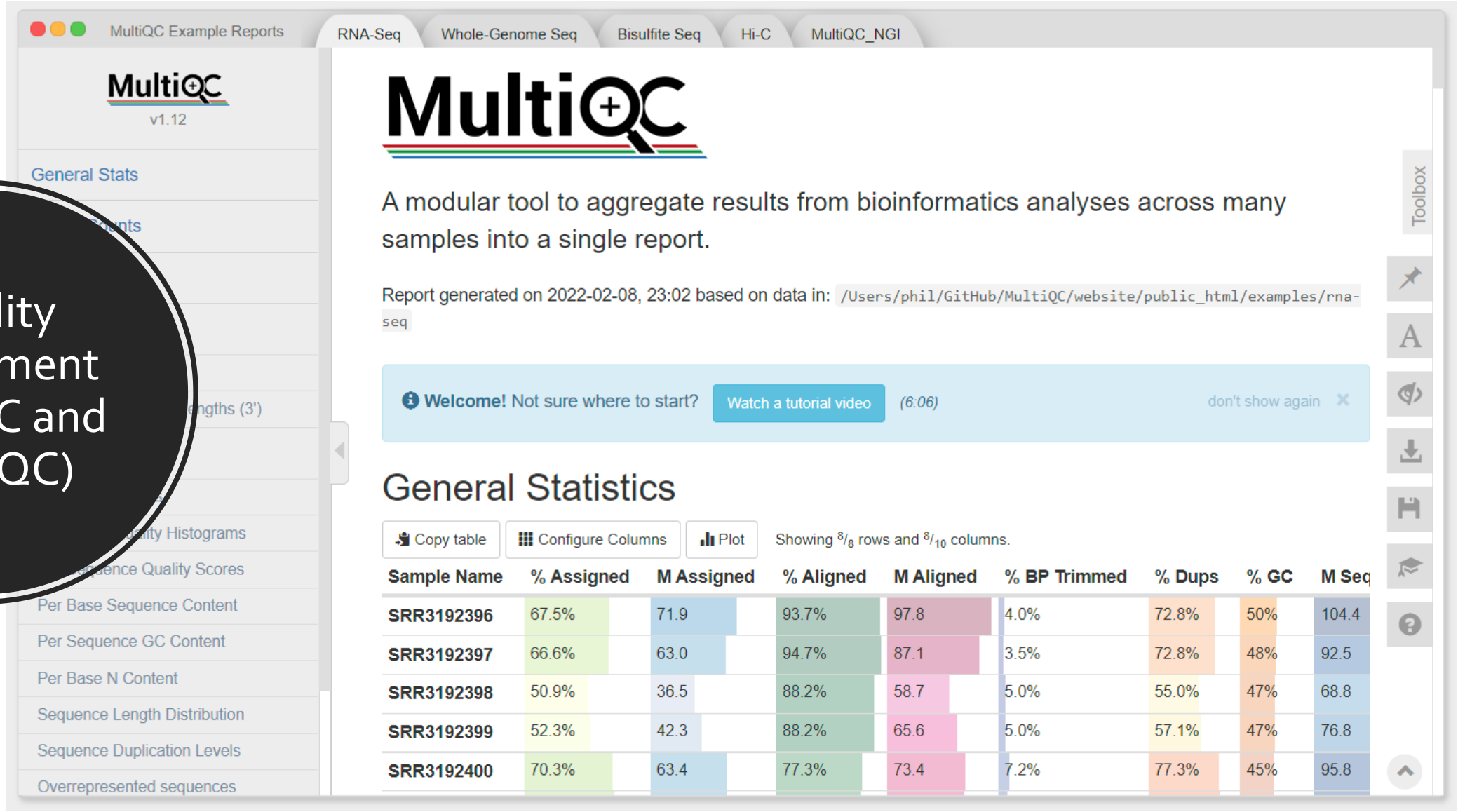






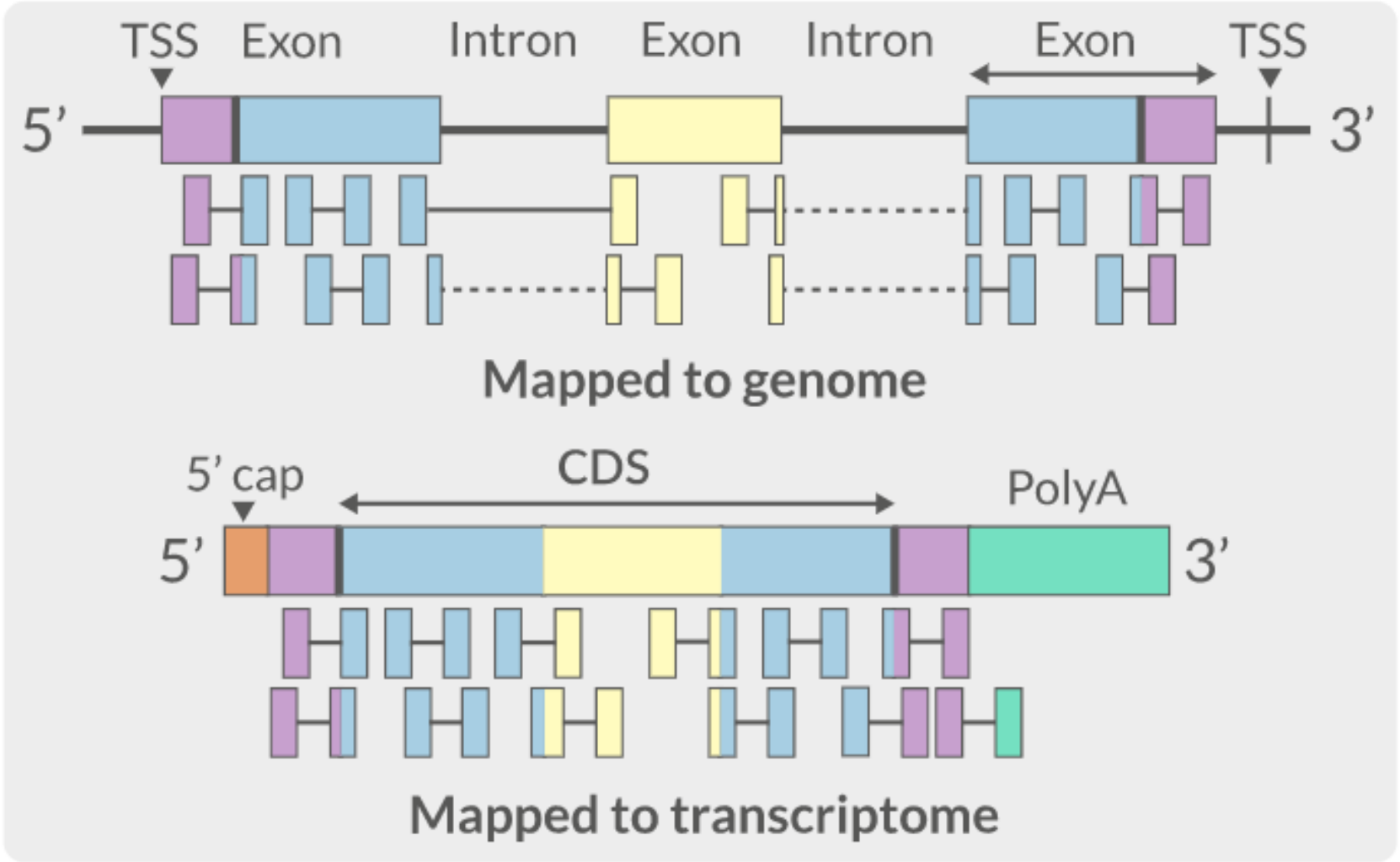


Quality assessment (FastQC and MultiQC)



Mapping and splice-aware alignment

- STAR
- HiSat2
- GSNAP
- SOAPsplice
- TopHat2
- Novoalign™
- CLC ™



- Some tools like Kallisto can quantify reads directly without a reference genome
- Viable aligners must (i) align reads across splice junctions, (ii) handle paired-end reads, (iii) handle strand-specific data, and (iv) run efficiently.
- Ability to align reads across unannotated splice junctions is also a plus

Baruzzo, G., *et al.* 2017. Simulation-based comprehensive benchmarking of RNA-seq aligners. Nature Methods 14(2): 135-139.

Sequence Alignment Map (SAM) QC

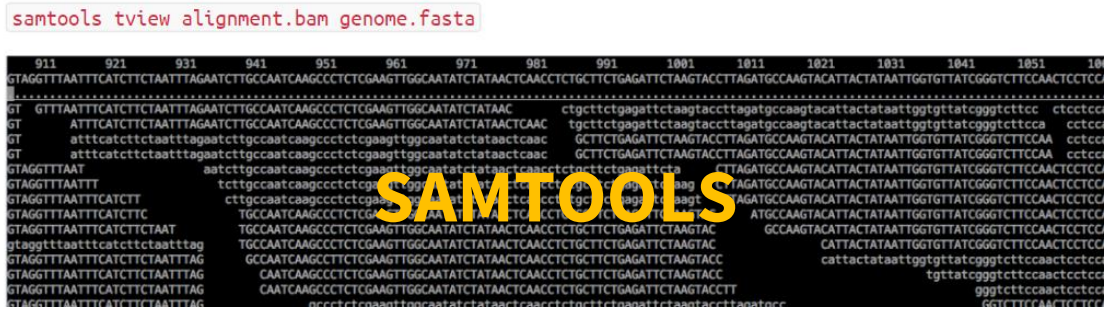
- samtools stats
- bamtools stats
- QoRTs
- RSeQC
- Qualimap

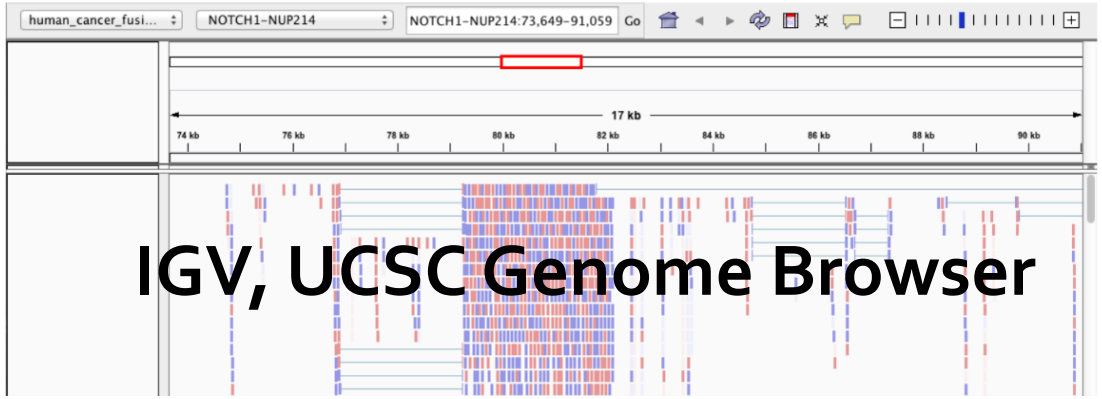
Size matters

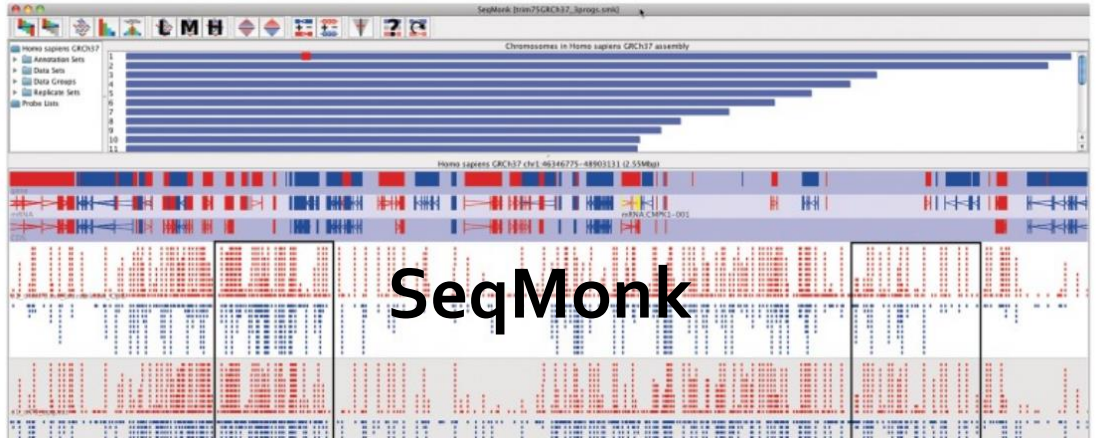
Format	Size_GB
SAM	7.4
BAM	1.9
CRAM lossless Q	1.4
CRAM 8 bins Q	0.8
CRAM no Q	0.26

- # of reads mapped/unmapped
- Uniquely mapped
- Insert size distribution
- Coverage +Gene body coverage
- Biotype counts / Chromosome counts
- Counts by region: gene/intron
- Sequencing saturation
- Strand specificity

Looks matter too..

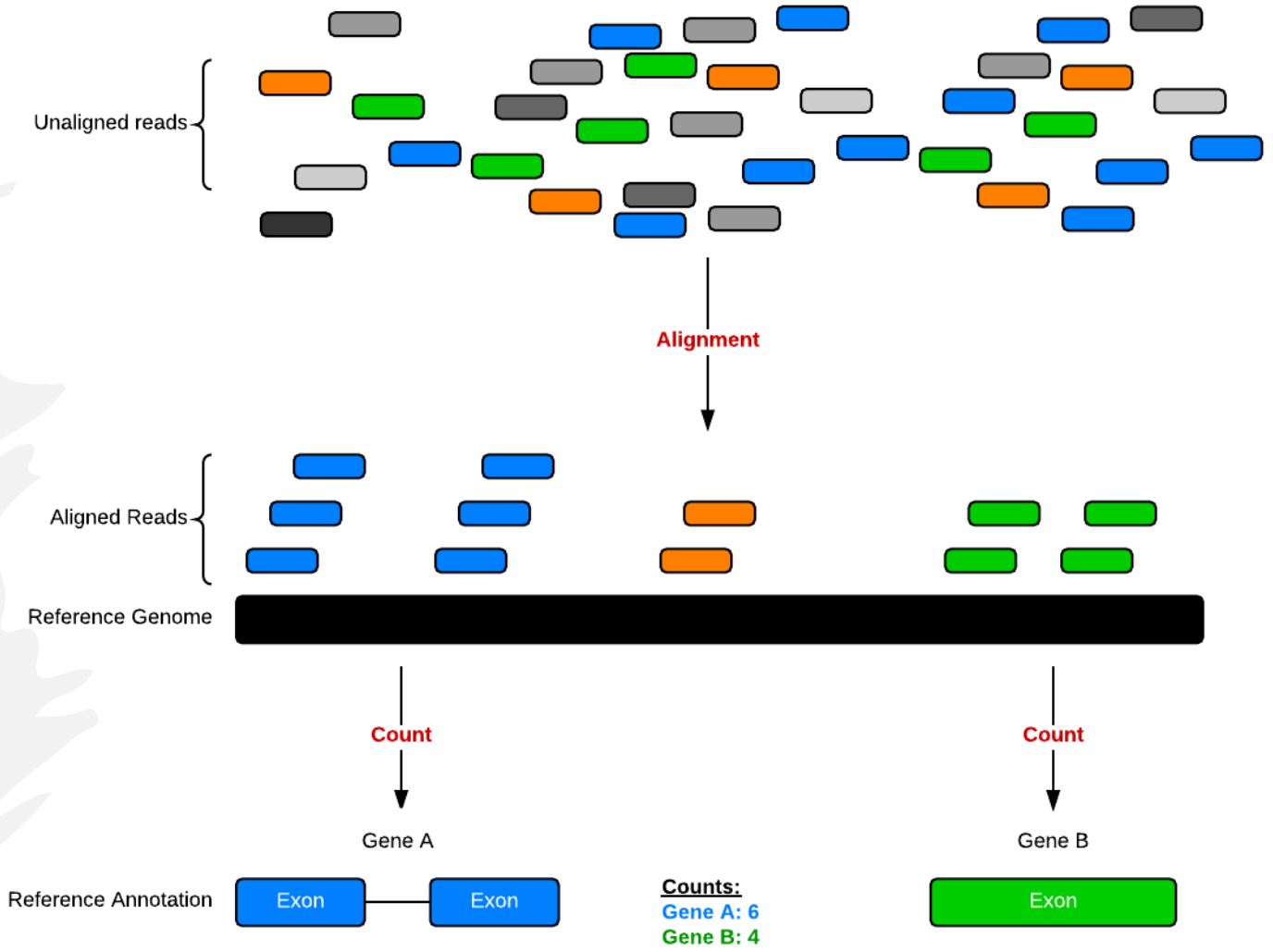






Quantifying counts

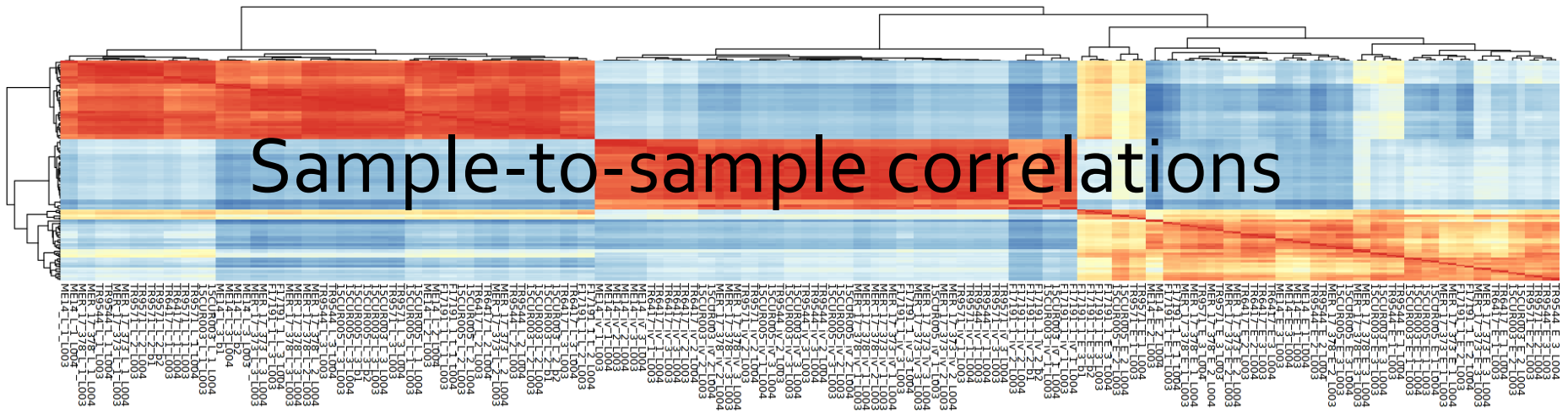
- Htseq, StringTie, RSEM, ~~Cufflinks~~, and Trinity are some of the popular tools.
- Reference annotation tells counting software the assembly position/location of features to quantify.
- Reads can be quantified on any feature in the reference annotation (gene, transcript, exon).
- Gene and transcript level are most common.



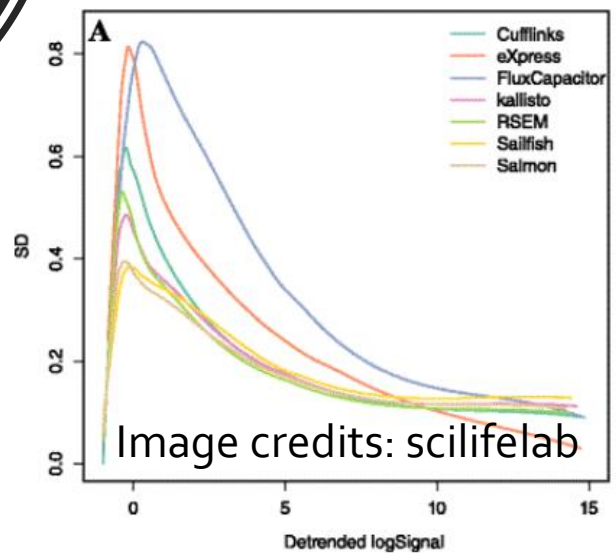
http://bioconnector.github.io/bims8382/r-rnaseq-airway.html

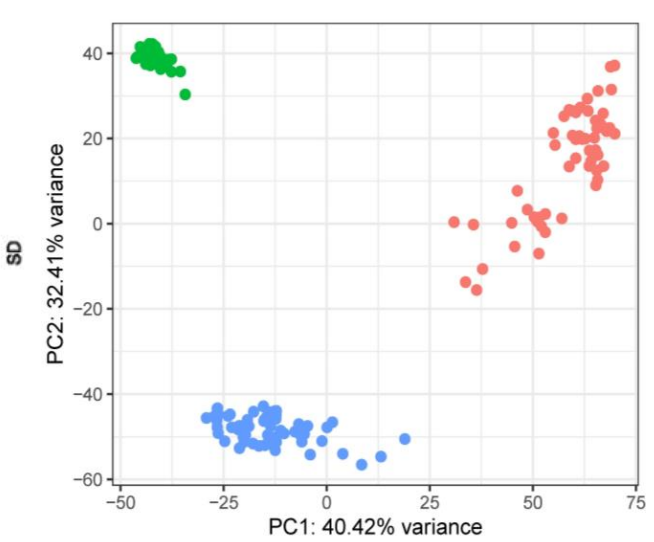
QC on Counts

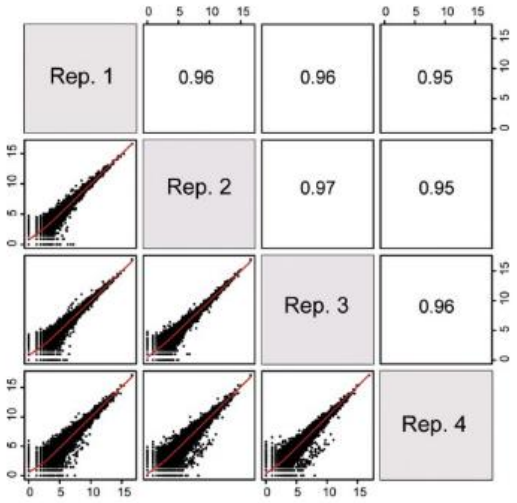
- RNASeqComp
- MultiQC (reports)
- R data wrangling



ENSG00000000003	140	242	188	143	287	344	438	280	253
ENSG00000000005	0	0	0	0	0	0	0	0	0
ENSG00000000419	69	98	77	55	52	94	116	79	69
ENSG00000000457	56	75	104	79	157	205	183	178	153

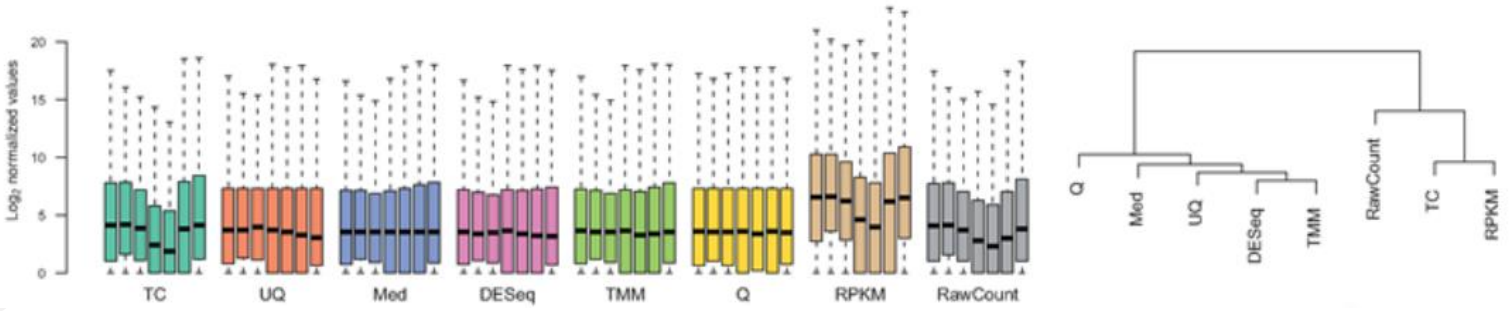


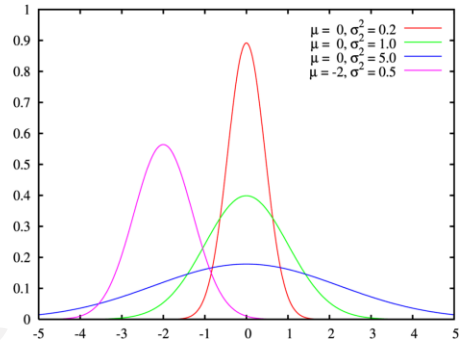


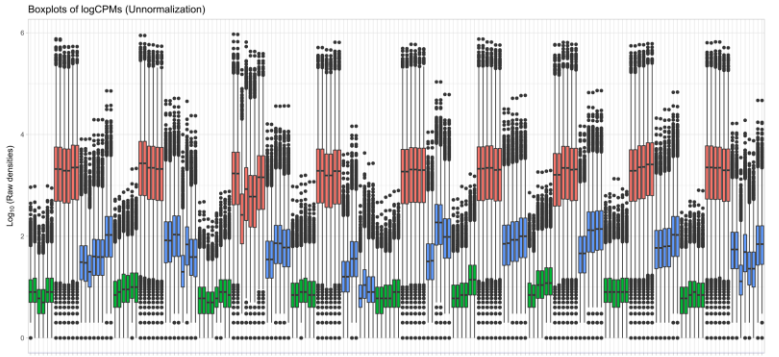


Normalizing counts

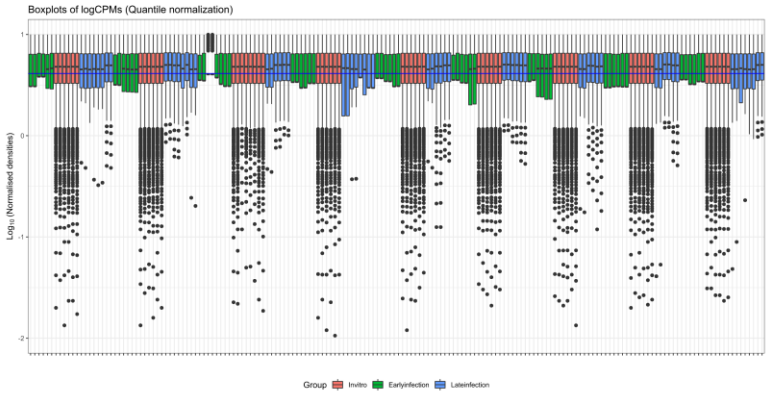
- Normalization controls for compositional bias and sequencing/library depth.
- Own analysis and plots: use normalized counts like TPM/FPKM.
- Heatmaps and clustering: use VST, R-log, Voom.
- Differential Gene Expression analysis: use raw counts.





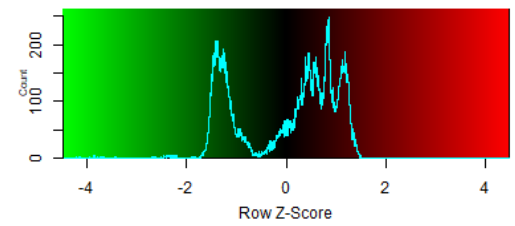


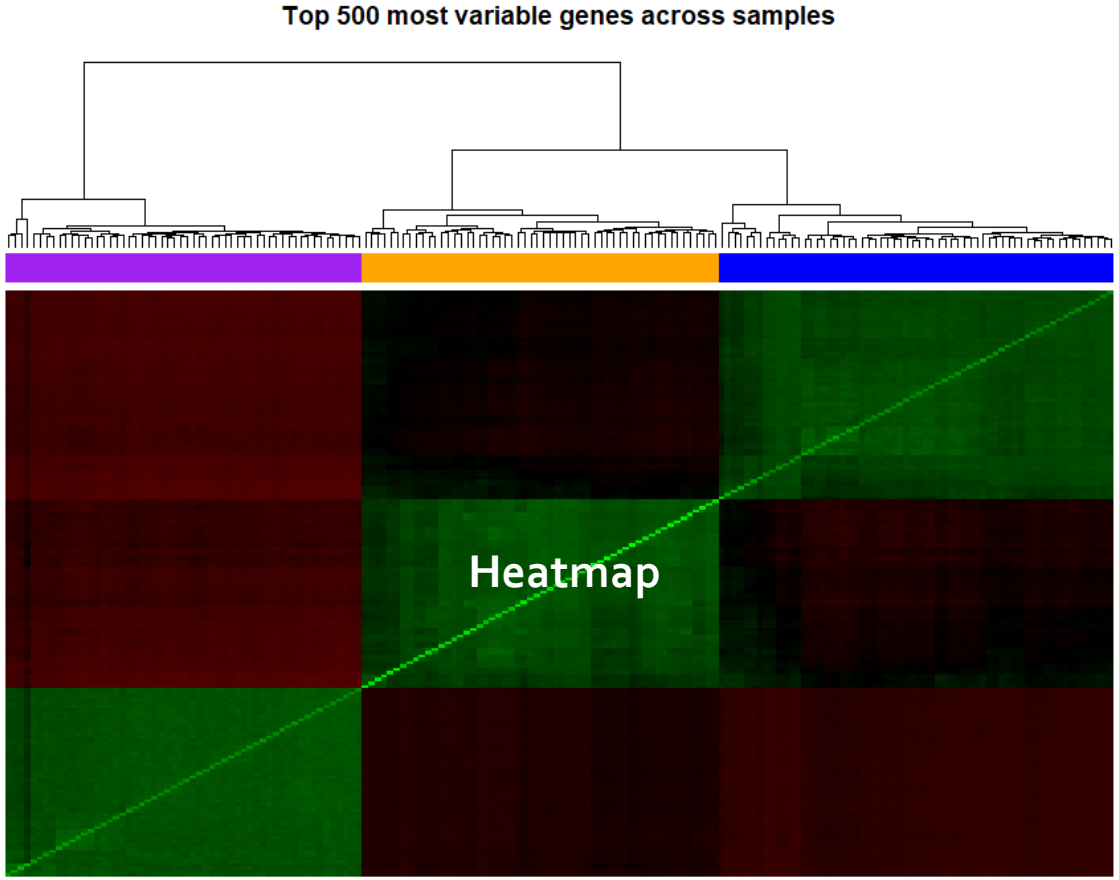
- TMM (edgeR)
- Limma-Voom
- TPM/FPKM/RPM
- R-Log
- Median of Ratios (DESeq2)

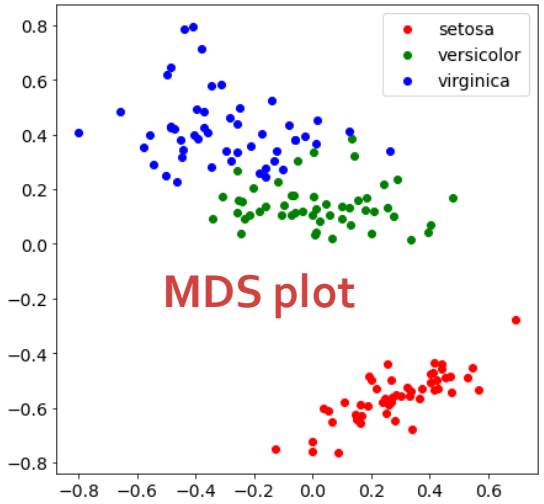


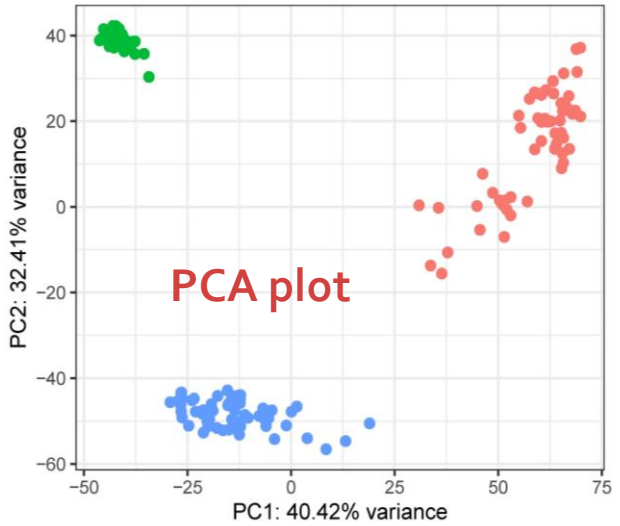
Exploration

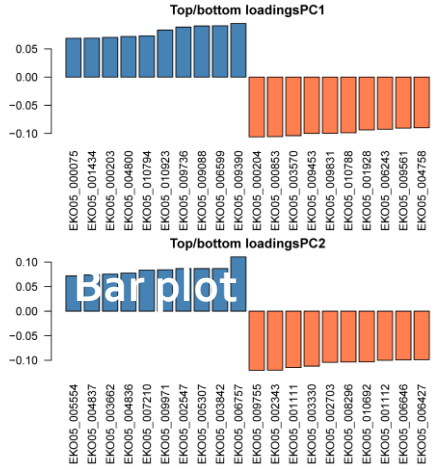
- SVA
- PVCA
- BatchQC
- ComBat
- R











DGE analysis

countData

	ctrl_1	ctrl_2	exp_1	exp_1
geneA	10	11	56	45
geneB	0	0	128	54
geneC	42	41	59	41
geneD	103	122	1	23
geneE	10	23	14	56
geneF	0	1	2	0
…	…	…	…	…
…	…	…	…	…
…	…	…	…	…

colData

	treatment	sex
ctrl_1	control	male
ctrl_2	control	female
exp_1	treatment	male
exp_2	treatment	female

Sample names:
ctrl_1, ctrl_2, exp_1, exp_2

countData is the count matrix
(number of reads mapping to each gene for each sample)

colData describes metadata about the *columns* of countData

colnames(countData) == rownames(colData)

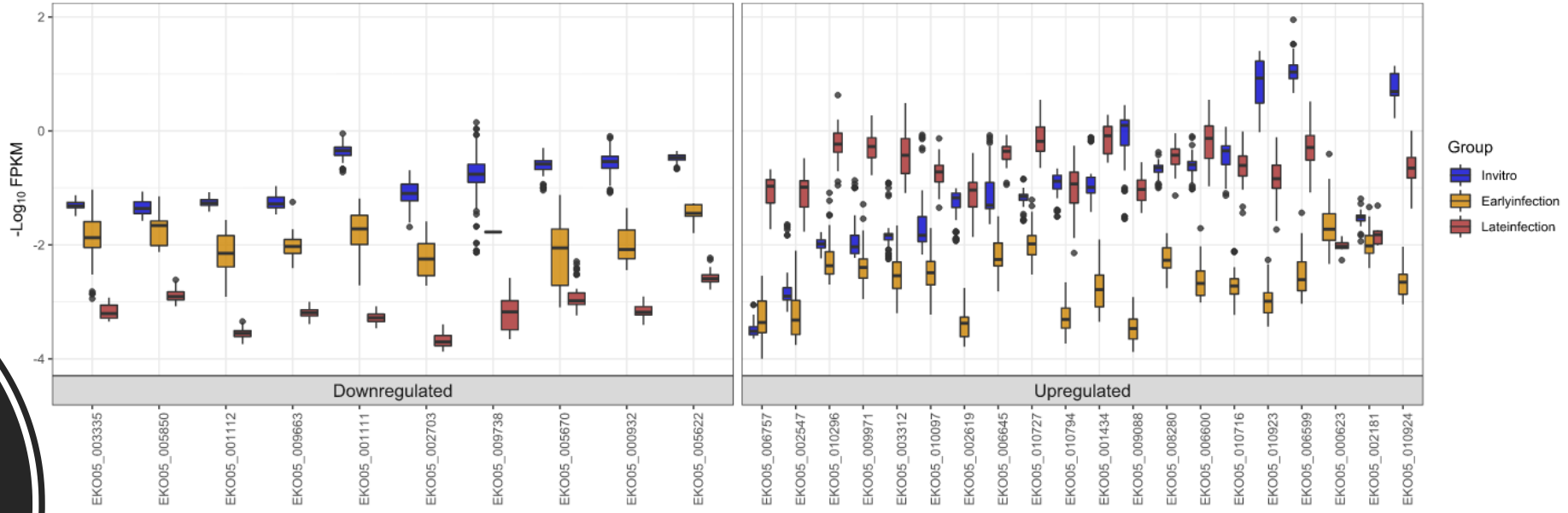
Experimental design

- Factors always ordered alphabetically by default
- The first variable is used as the base (intercept) upon which all other variables are compared
- Set variable levels manually to ensure the model is consistent with the hypothesis.

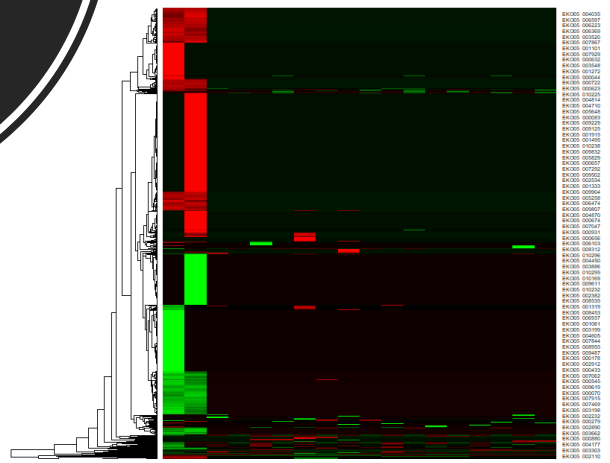
relevel(colData$treatment, 'control')
x$col <- factor(x$col, levels = c(a,b..))

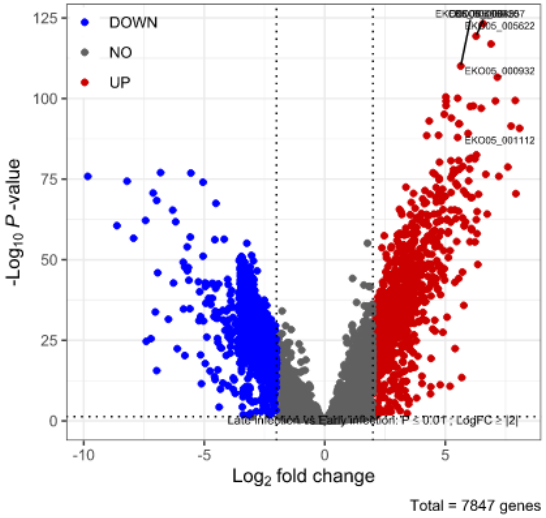
design = ~treatment ::: compare treated-vs-control using control as reference.
design = ~treatment+sex ::: a + M vs F with female as reference.
design = ~o+sex+treatment ::: a and b but with sex not contrasted.
design = ~treatment+sex+treatment:sex ::: a+b+interaction between a & b.
design = ~treatment+sex+batch+treatment:sex ::: d + model for batch effects.

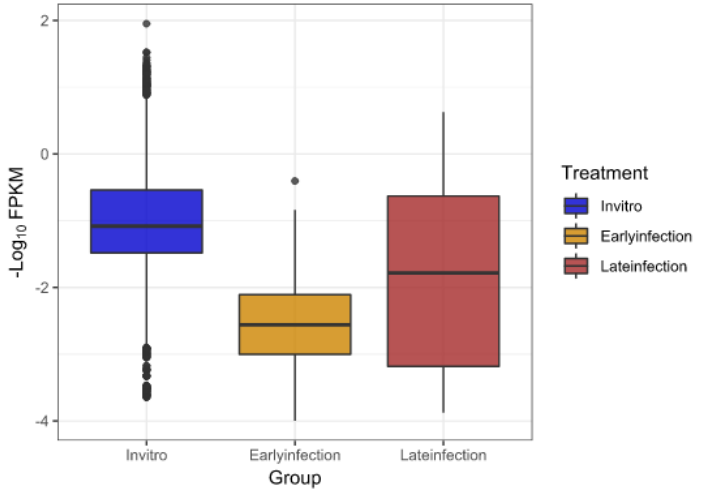
Visualize



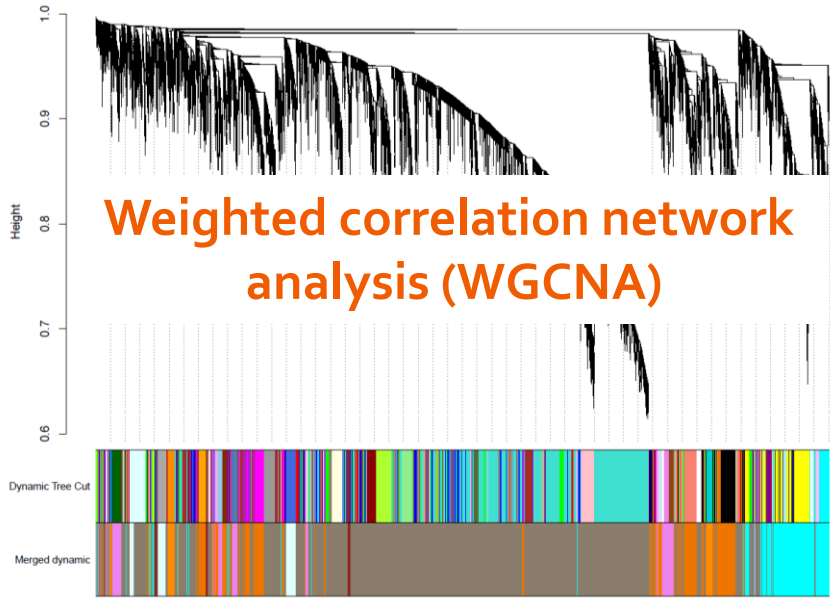
Gene differentialy expresed under different conditions



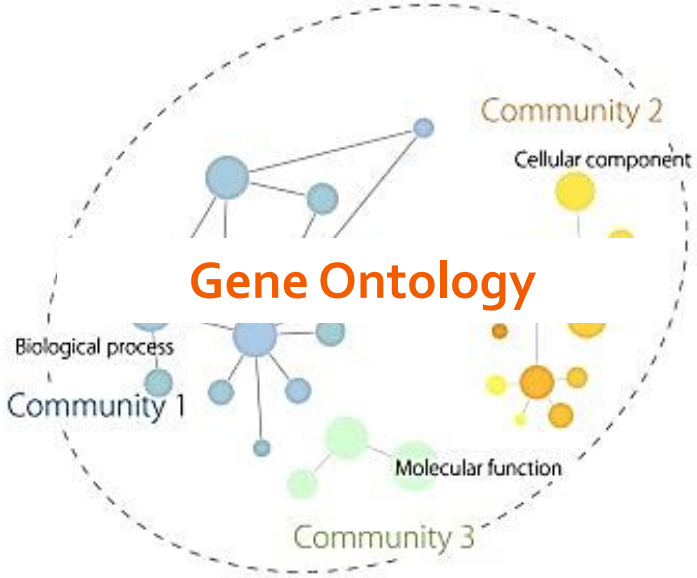




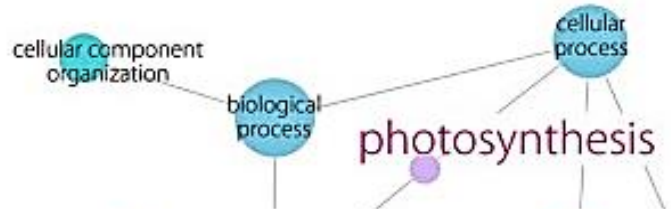
Functional analysis



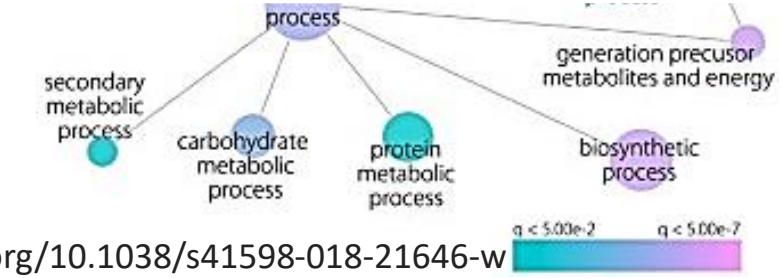
Weighted correlation network analysis (WGCNA)

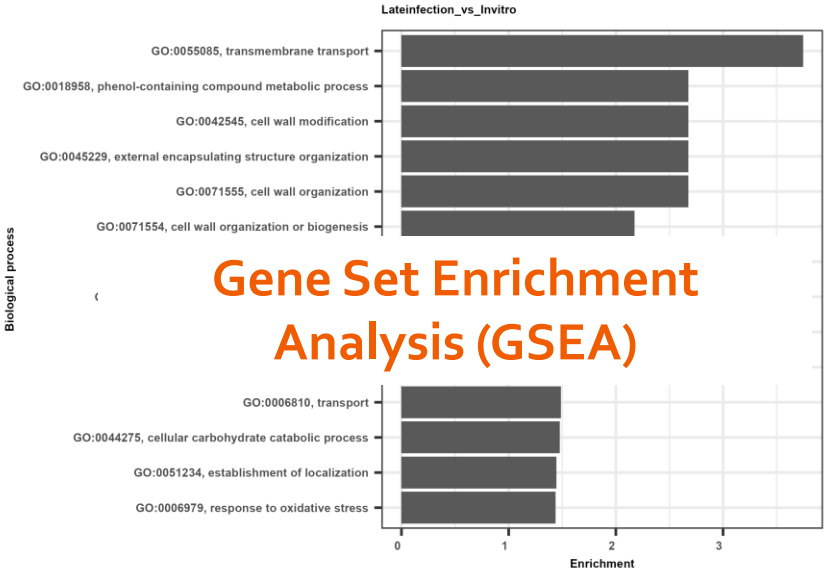


Gene Ontology



Self Organising Maps (SOM)





Gene Set Enrichment Analysis (GSEA)

doi.org/10.1038/s41598-018-21646-w

The aim is to identify important pathways or functional modules represented in your expression data..

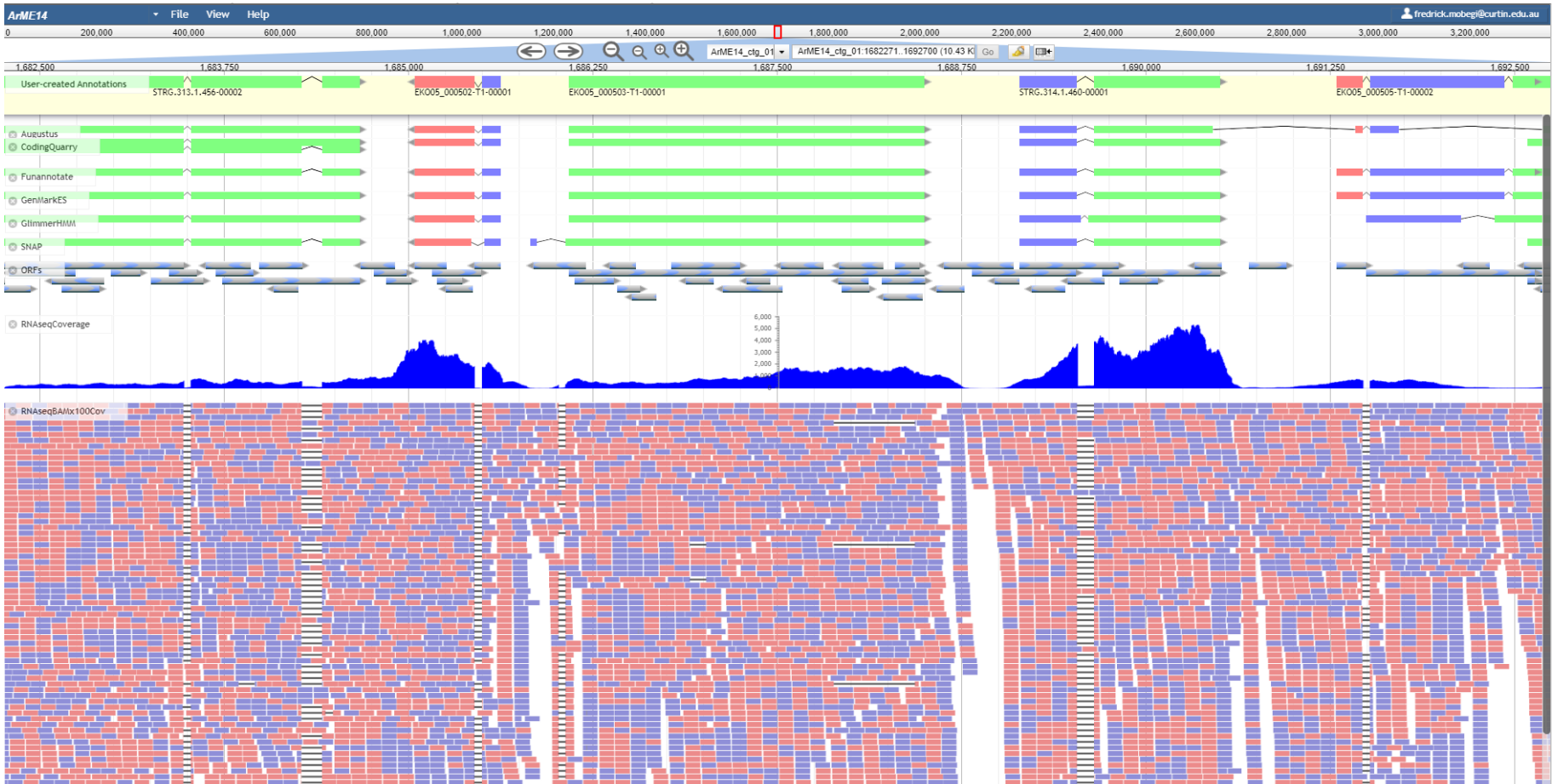
Genome annotation

Software | Open Access | Published: 30 August 2013

Web Apollo: a web-based genomic annotation editing platform

Eduardo Lee, Gregg A Helt, Justin T Reese, Monica C Munoz-Torres, Chris P Childers, Robert M Buels, Lincoln Stein, Ian H Holmes, Christine G Elsik & Suzanna E Lewis

Genome Biology **14**, Article number: R93 (2013) | Cite this article



Further reading

- Fu, Yu *et al.*, "Elimination of PCR duplicates in RNA-seq and small RNA-seq using unique molecular identifiers." BMC genomics 19.1 (2018): 531.
- Parekh *et al.*, "The impact of amplification on differential expression analyses by RNA-seq." Scientific reports 6 (2016): 25533.
- Klepikova *et al.*, "Effect of method of deduplication on estimation of differential gene expression using RNA-seq." PeerJ 5 (2017): e3091.
- Soneson *et al.*, "Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences." F1000Research 4 (2015).
- Zhang *et al.*, "Evaluation and comparison of computational tools for RNA-seq isoform quantification." BMC genomics 18.1 (2017): 583.
- Teng *et al.*, "A benchmark for RNA-seq quantification pipelines." Genome biology 17.1 (2016): 74.
- Conesa *et al.*, "A survey of best practices for RNA-seq data analysis." Genome biology 17.1 (2016): 13.
- Hsieh *et al.*, "Effect of de novo transcriptome assembly on transcript quantification" 2018 bioRxiv 380998.
- Wang, Sufang, and Michael Gribskov. "Comprehensive evaluation of de novo transcriptome assembly programs and their effects on differential gene expression analysis." Bioinformatics 33.3 (2017): 327-333.

Summary

- Quality Control is KING in everything transcriptomics.
- A good experimental design helps avoid/minimize confounding effects.
- Biological replicates rule above all other replicates and/or sequencing methods.
- Carefully planning sample prep and sequencing ensures you capture the intended hypothesis in your experiment.
- Always discard low quality bases, reads, genes and samples.
- Experiment with different pipelines and software to identify biases.
- Corroborate assumptions about the data with methods and tools used for analysis.
- No two tools will give you identical results. Choose wisely
- You can never have too many illustrations in RNA-seq.

Questions