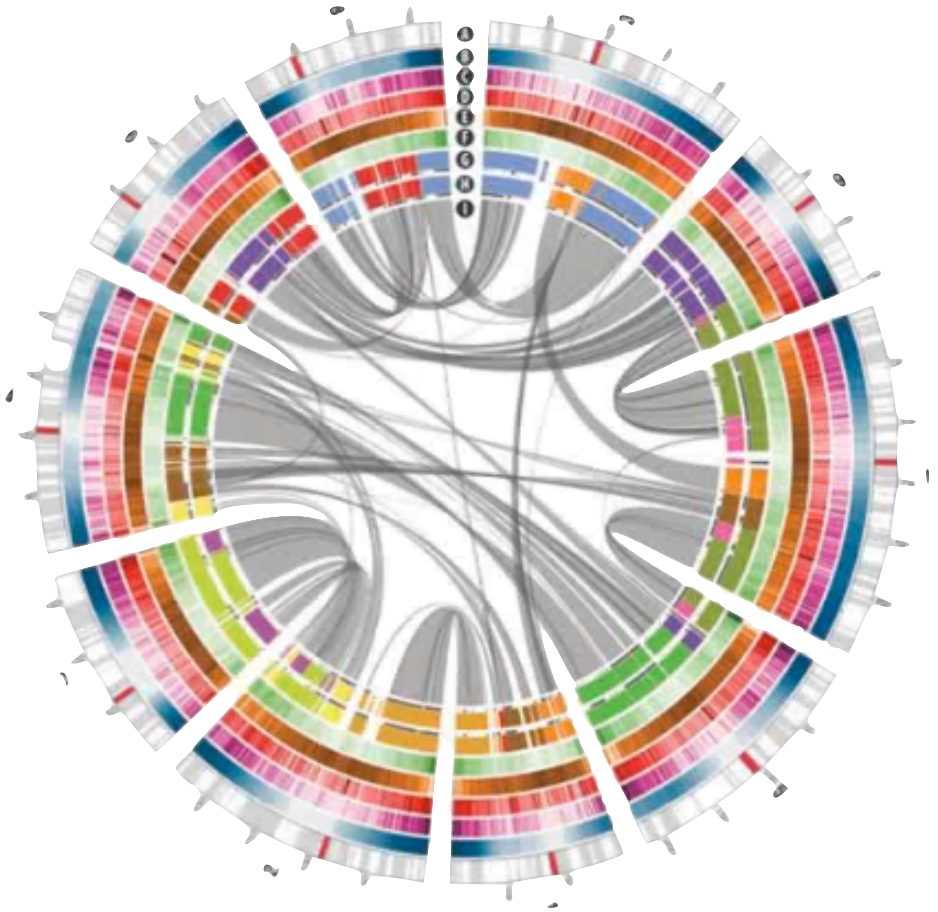
UNIVERSIDAD DE MÁLAGA scbi uma pab scbi PLATAFORMA ANDALUZA DE BIOINFORMÁTICA BioIn4NEXT ibima Instituto de Investigación Biomédica de Málaga LifeWatch ERIC

Bioinformatic analysis at SCBI



Scientific Community Meeting - LifeWatch-ERIC
25 to 27 Feb 2019
Porto (Portugal)

Rocío Bautista, PhD
Bioinformatic Unit
Supercomputing and Bioinnovation Center
University of Malaga

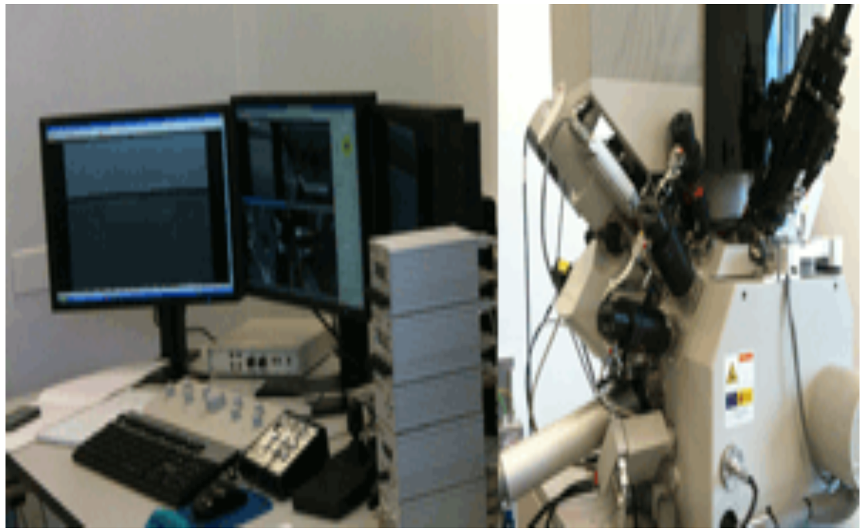
Supercomputing and Bioinnovation Center

Genomic Unit

Bioinformatics Unit

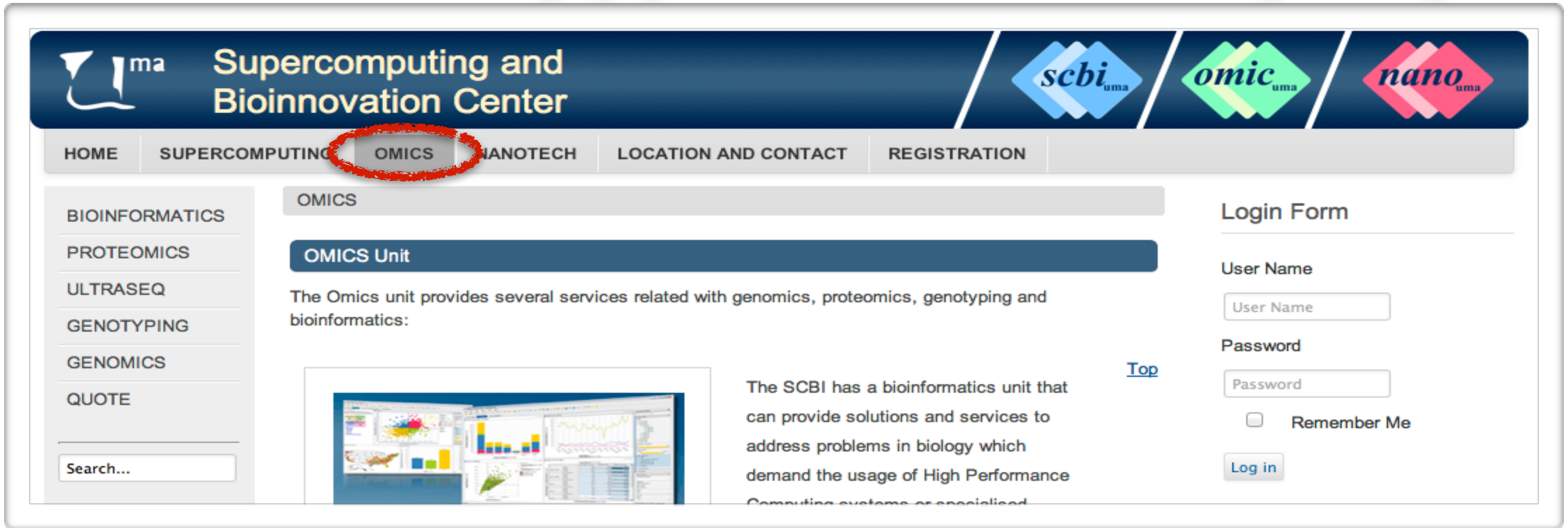
Nanotechnology Unit

Supercomputer (PICASSO)



rociobm@uma.es

Our web: http://www.scbi.uma.es/site/



OMICS

OMICS Unit

The Omics unit provides several services related with genomics, proteomics, genotyping and bioinformatics:

The SCBI has a bioinformatics unit that can provide solutions and services to address problems in biology which demand the usage of High Performance Computing systems or specialised

Login Form

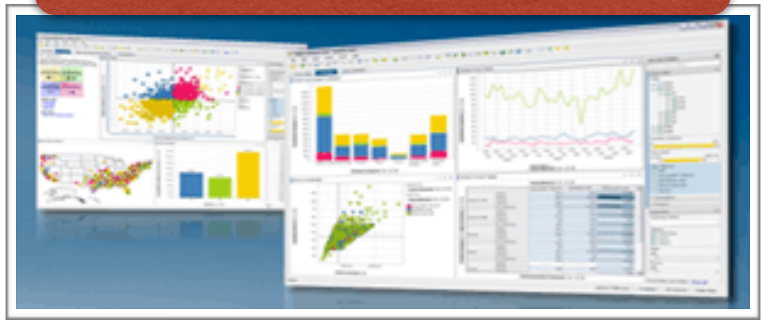
User Name: User Name

Password: Password

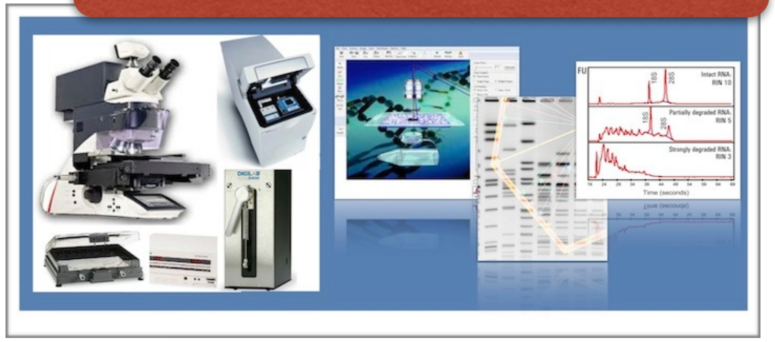
☐ Remember Me

NGS

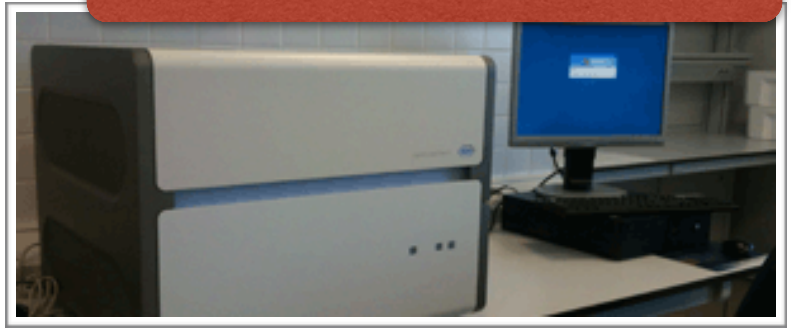
Bioinformatics



Genomics



Genotype



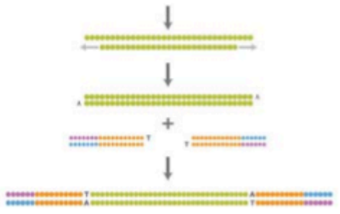
Integrating Next Generation Sequencing (NGS) Workflows

NextSeq500

MiSeq

1. **DNA/RNA**

2. **Library**



3. **Clustering**

4. **Sequencing**

5. **Data processing**

Report

rociobm@uma.es

Next Generation Sequencing Applications

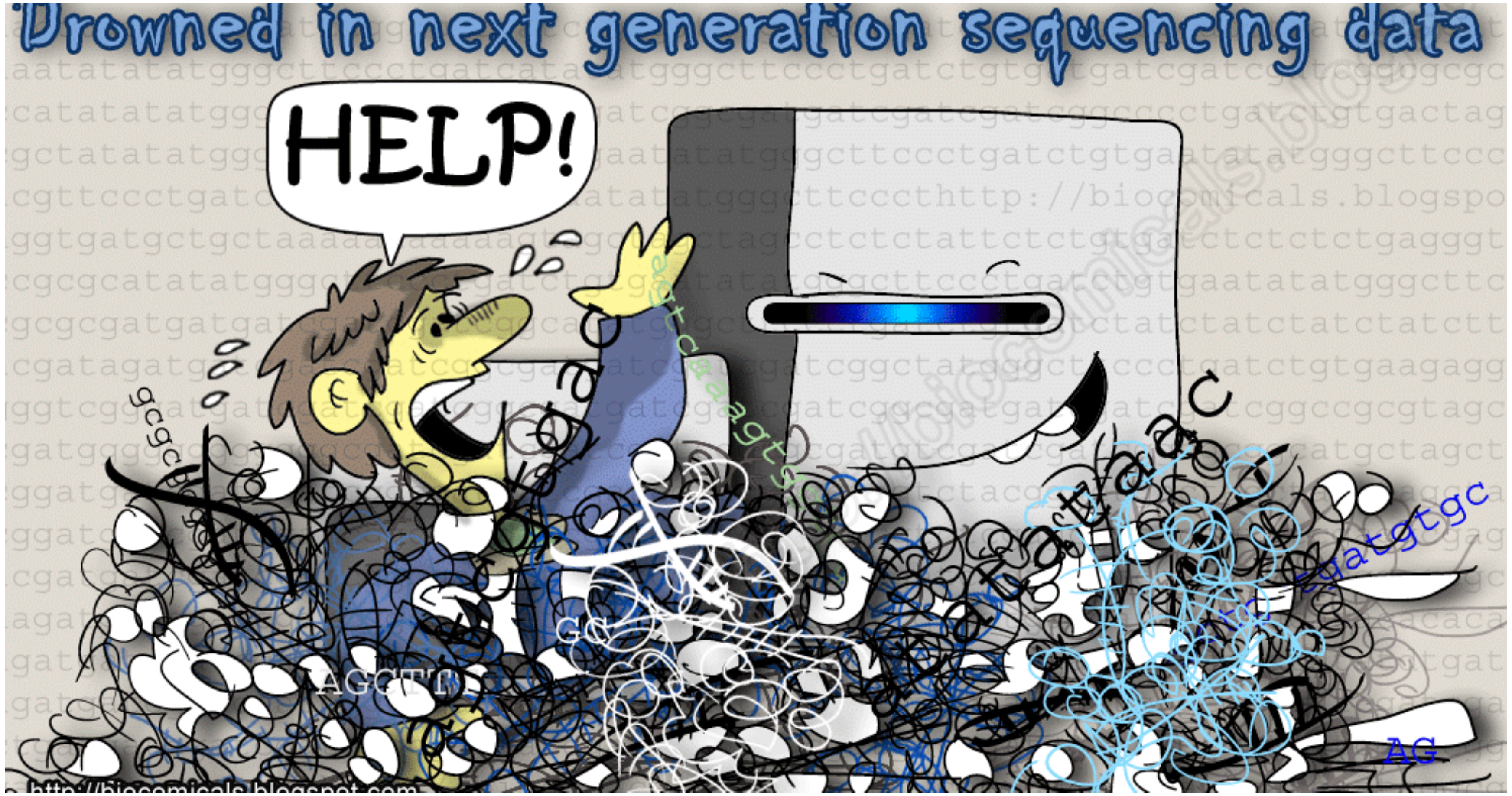
1. **Genome sequencing**
   - Prokaryotic genome
   - Eukaryotic genome
2. **Target sequencing**
   - Target genome
   - Target genes
3. **RNA-seq**
   - mRNA
   - Total RNA
   - Small RNA
4. **Metagenomics**
   - Shotgun
   - Amplicon 16S/18S
5. **Epigenetics**
   - Chip-Seq
   - Methyl-Seq

rociobm@uma.es

DATA, DATA, DATA... BIG DATA ERA

7.5 GB to 1.8 TB!!!

300.000 – 1.000.000 raw / sample !!!



Our users need supercomputers!!!

BIOINFORMATICS

Hardware / Software

http://www.scbi.uma.es/site/scbi/software

Hardware:

- ☑4016 cores
- ☑22 TB RAM
- ☑5 PB

Software:

- ☑ > 400 apps

Users:

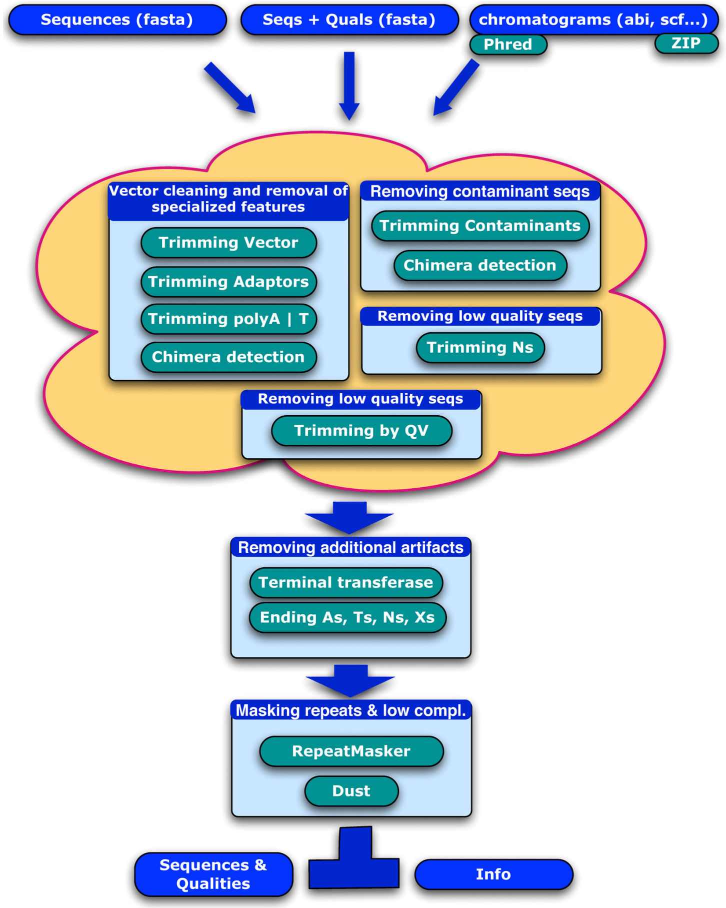
- ☑ > 230 research groups
- ☑ 47 biology groups

rociobm@uma.es

Our first workflows

SeqTrimNext

Preprocessing Reads



Sequences (fasta)
Seqs + Quals (fasta)
chromatograms (abi, scf...)
Phred
ZIP

Vector cleaning and removal of specialized features
Trimming Vector
Trimming Adaptors
Trimming polyA | T
Chimera detection

Removing contaminant seqs
Trimming Contaminants
Chimera detection

Removing low quality seqs
Trimming Ns

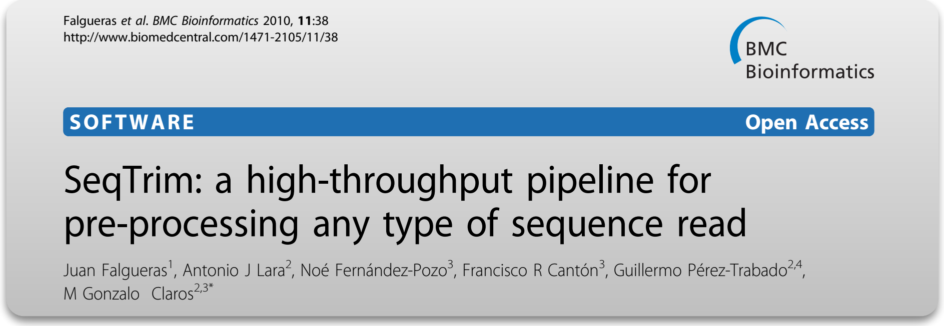
Removing low quality seqs
Trimming by QV

Removing additional artifacts
Terminal transferase
Ending As, Ts, Ns, Xs

Masking repeats & low compl.
RepeatMasker
Dust

Sequences & Qualities
Info

Useful NGS reads



Falgueras *et al. BMC Bioinformatics* 2010, **11**:38
http://www.biomedcentral.com/1471-2105/11/38

BMC Bioinformatics

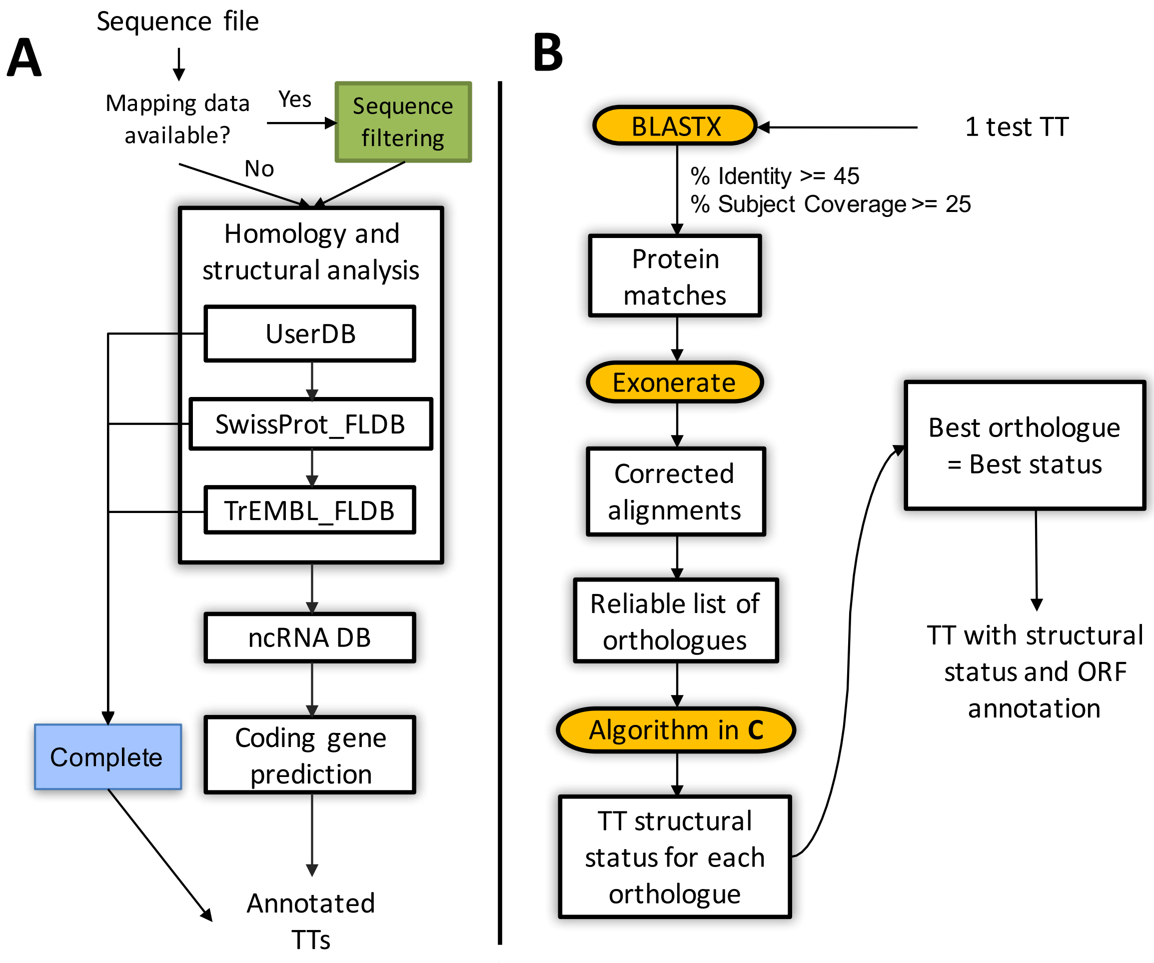
SOFTWARE **Open Access**

SeqTrim: a high-throughput pipeline for pre-processing any type of sequence read

Juan Falgueras[1], Antonio J Lara[2], Noé Fernández-Pozo[3], Francisco R Cantón[3], Guillermo Pérez-Trabado[2,4], M Gonzalo Claros[2,3*]

Full-LengtherNext

Gene Annotation



A

Sequence file

Mapping data available?

Yes

Sequence filtering

No

Homology and structural analysis

UserDB

SwissProt_FLDB

TrEMBL_FLDB

ncRNA DB

Coding gene prediction

Complete

Annotated TTs

B

BLASTX

1 test TT

% Identity >= 45
% Subject Coverage >= 25

Protein matches

Exonerate

Corrected alignments

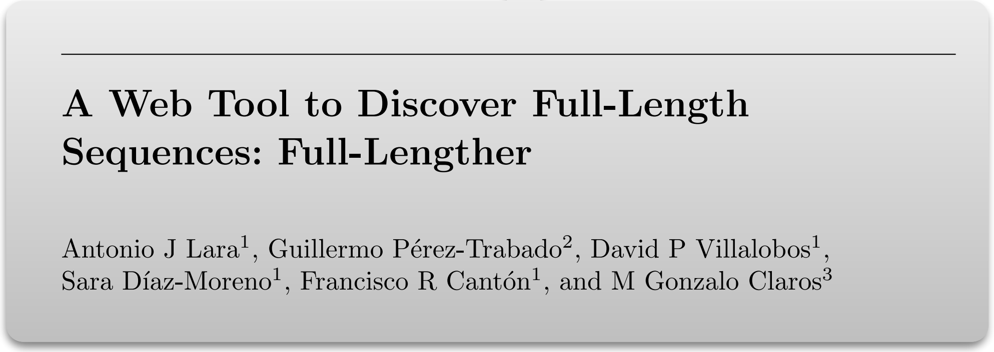
Reliable list of orthologues

Algorithm in C

TT structural status for each orthologue

Best orthologue = Best status

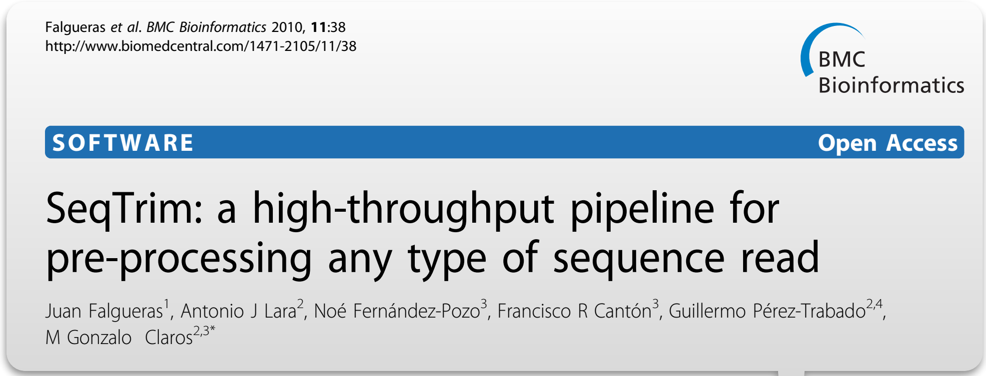
TT with structural status and ORF annotation



A Web Tool to Discover Full-Length Sequences: Full-Lengther

Antonio J Lara[1], Guillermo Pérez-Trabado[2], David P Villalobos[1], Sara Díaz-Moreno[1], Francisco R Cantón[1], and M Gonzalo Claros[3]

Transcriptomic workflow for non-model organisms



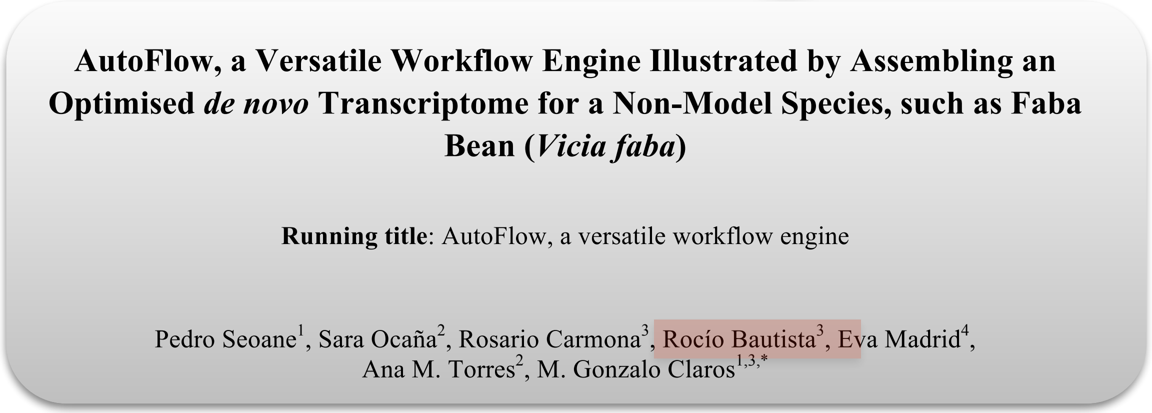
Falgueras *et al. BMC Bioinformatics* 2010, **11**:38
http://www.biomedcentral.com/1471-2105/11/38

BMC Bioinformatics

SOFTWARE **Open Access**

SeqTrim: a high-throughput pipeline for pre-processing any type of sequence read

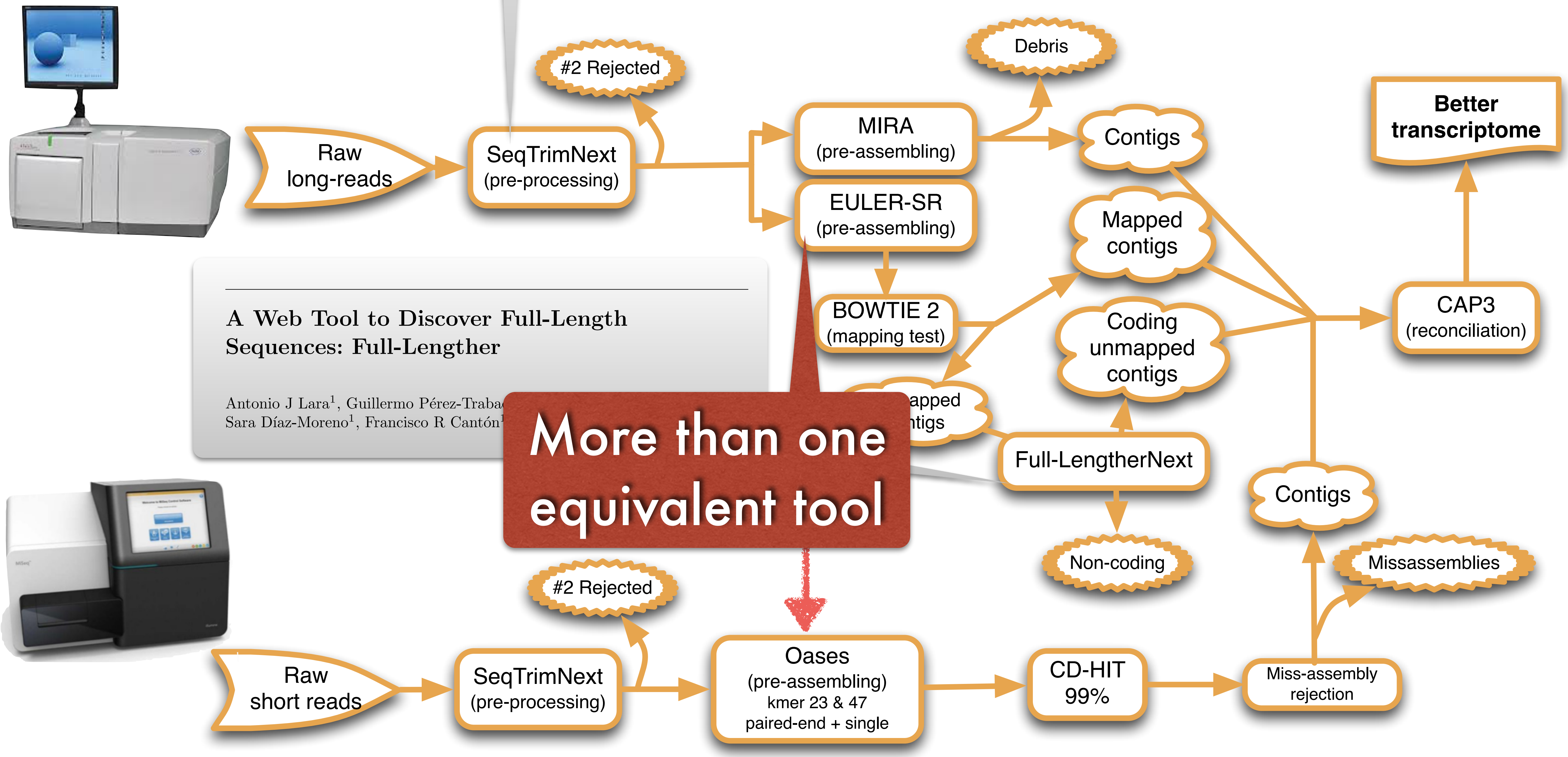
Juan Falgueras[1], Antonio J Lara[2], Noé Fernández-Pozo[3], Francisco R Cantón[3], Guillermo Pérez-Trabado[2,4], M Gonzalo Claros[2,3*]



AutoFlow, a Versatile Workflow Engine Illustrated by Assembling an Optimised *de novo* Transcriptome for a Non-Model Species, such as Faba Bean (*Vicia faba*)

Running title: AutoFlow, a versatile workflow engine

Pedro Seoane[1], Sara Ocaña[2], Rosario Carmona[3], Rocío Bautista[3], Eva Madrid[4], Ana M. Torres[2], M. Gonzalo Claros[1,3,*]



A Web Tool to Discover Full-Length Sequences: Full-Lengther

Antonio J Lara[1], Guillermo Pérez-Traba
Sara Díaz-Moreno[1], Francisco R Cantón[1]

Raw long-reads

SeqTrimNext (pre-processing)

#2 Rejected

MIRA (pre-assembling)

EULER-SR (pre-assembling)

Debris

Contigs

Mapped contigs

BOWTIE 2 (mapping test)

Coding unmapped contigs

apped ntigs

Full-LengtherNext

Non-coding

CAP3 (reconciliation)

Better transcriptome

More than one equivalent tool

Raw short reads

SeqTrimNext (pre-processing)

#2 Rejected

Oases (pre-assembling) kmer 23 & 47 paired-end + single

CD-HIT 99%

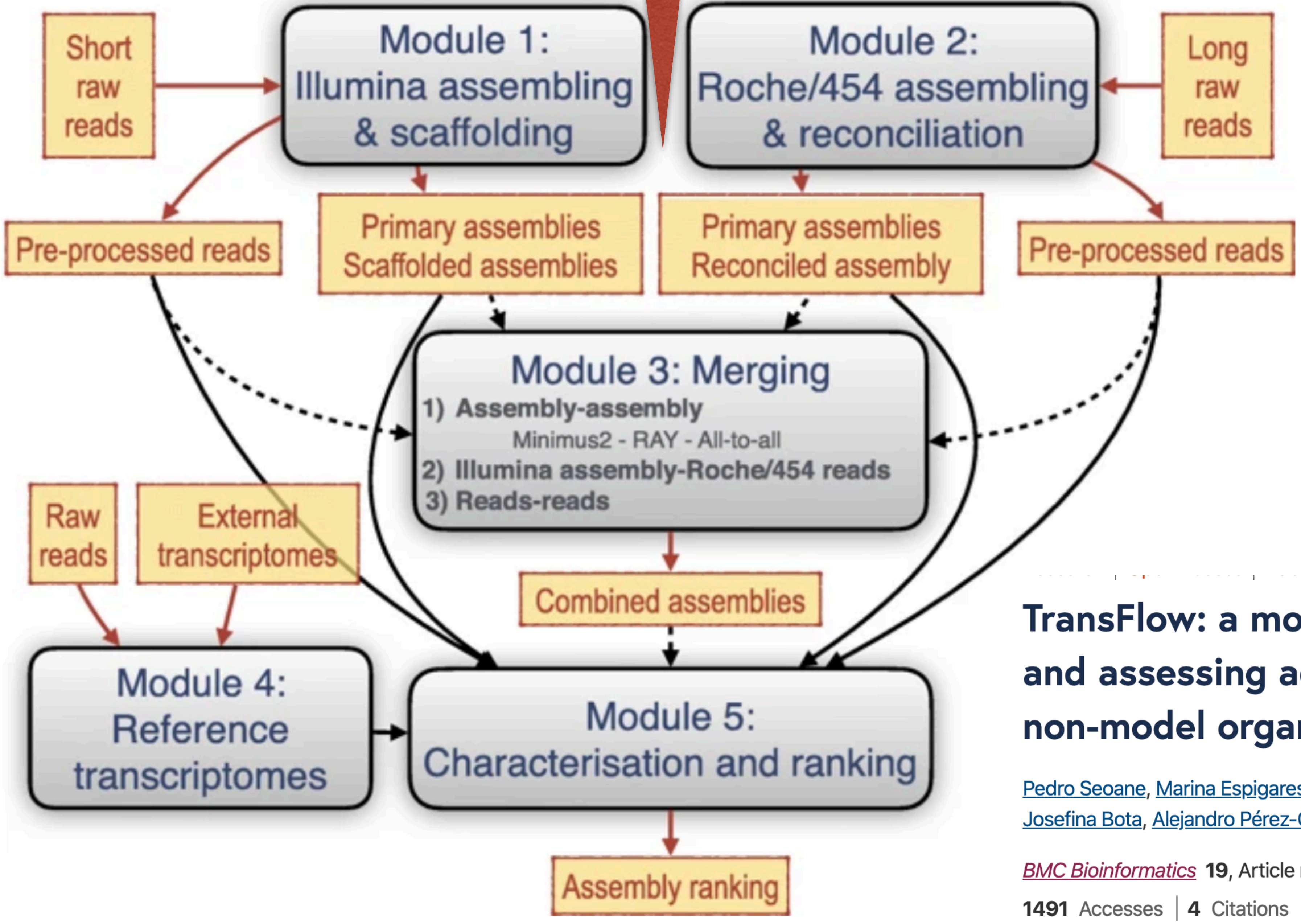
Miss-assembly rejection

Contigs

Missassemblies

The improvement...

More than one equivalent tool



Short raw reads

Module 1: Illumina assembling & scaffolding

Module 2: Roche/454 assembling & reconciliation

Long raw reads

Pre-processed reads

Primary assemblies
Scaffolded assemblies

Primary assemblies
Reconciled assembly

Pre-processed reads

Module 3: Merging
1) Assembly-assembly
Minimus2 - RAY - All-to-all
2) Illumina assembly-Roche/454 reads
3) Reads-reads

Raw reads

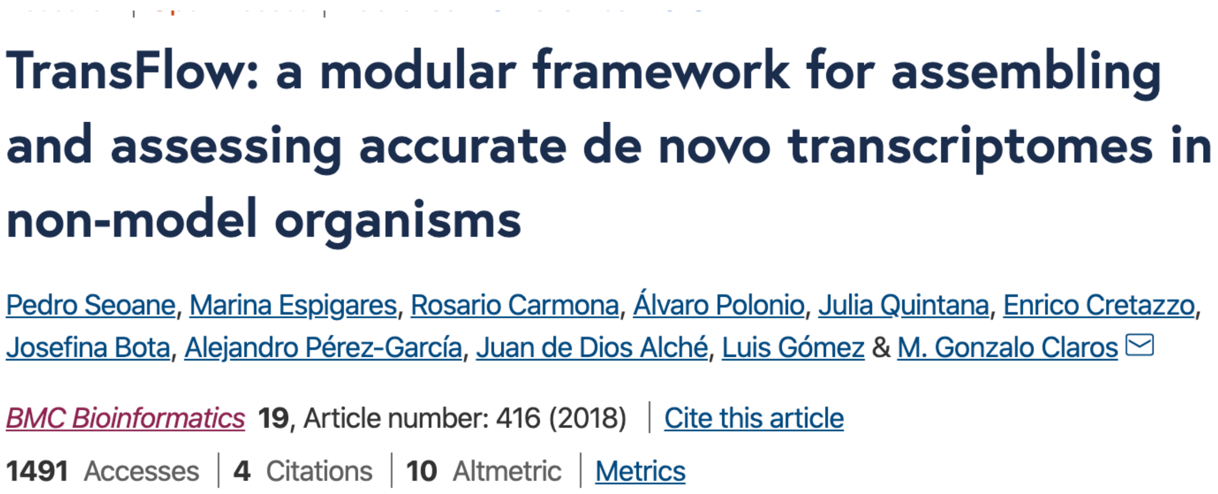
External transcriptomes

Combined assemblies

Module 4: Reference transcriptomes

Module 5: Characterisation and ranking

Assembly ranking



TransFlow: a modular framework for assembling and assessing accurate de novo transcriptomes in non-model organisms

Pedro Seoane, Marina Espigares, Rosario Carmona, Álvaro Polonio, Julia Quintana, Enrico Cretazzo, Josefina Bota, Alejandro Pérez-García, Juan de Dios Alché, Luis Gómez & M. Gonzalo Claros

BMC Bioinformatics **19**, Article number: 416 (2018) | Cite this article

1491 Accesses | **4** Citations | **10** Altmetric | Metrics

rociobm@uma.es

We have assembled several transcriptomes

Fernández-Pozo *et al. BMC Genomics* 2011, **12**:366
http://www.biomedcentral.com/1471-2164/12/366

BMC Genomics

DATABASE **Open Access**

EuroPineDB: a high-coverage web database for maritime pine transcriptome

Noé Fernández-Pozo[1], Javier Canales[1], Darío Guerrero-Fernández[2], David P Villalobos
Rocío Bautista[2], Arantxa Flores-Monterroso[1], M Ángeles Guevara[3], Pedro Perdiguero
M Teresa Cervera[3,4], Álvaro Soto[3,4], Ricardo Ordás[5], Francisco R Cantón[1], Concepción
and M Gonzalo Claros[1,2*]

Benzekri *et al. BMC Genomics* 2014, **15**:952
http://www.biomedcentral.com/1471-2164/15/952

RESEARCH ARTICLE **Open Access**

De novo assembly, characterization and functional annotation of Senegalese sole (*Solea senegalensis*) and common sole (*Solea solea*) transcriptomes: integration in a database and design of a microarray

Hicham Benzekri[1,2], Paula Armesto[3], Xavier Cou
David Mazurais[8], Rocío Bautista[2], Darío Guerrer
Jose Luis Zambonino[8], Sabine Nidelet[10], Marta
M Gonzalo Claros[1,2] and Manuel Manchado[3*]

PLOS | ONE

RESEARCH ARTICLE

Large-Scale Transcriptome Analysis in Faba Bean (*Vicia faba* L.) under *Ascochyta fabae* Infection

Sara Ocaña[1], Pedro Seoane[2], Rocio Bautista[3], Carmen Palomino[1], Gonzalo M. Claros[2,3], Ana M. Torres[1]*, Eva Madrid[4]

Plant Biotechnology Journal

Plant Biotechnology Journal (2014) **12**, pp. 286–299 doi: 10.1111/pbi.12136

De novo assembly of maritime pine transcriptome: implications for forest breeding and biotechnology

Javier Canales[1,†], Rocio Bautista[2,†], Philippe Label[3†], Josefa Gómez-Maldonado[1], Isabelle Lesur[4,5,6], Noe Fernández-Pozo[2], Marina Rueda-López[1], Dario Guerrero-Fernández[2], Vanessa Castro-Rodríguez[1], Hicham Benzekri[2], Rafael A. Cañas[1], María-Angeles Guevara[7], Andreia Rodrigues[8], Pedro Seoane[2], Caroline Teyssier[9], Alexandre Morel[9], François Ehrenmann[4,5], Grégoire Le Provost[4,5], Céline Lalanne[4,5], Céline Noirot[10], Christophe Klopp[10], Isabelle Reymond[11], Angel García-Gutiérrez[1], Jean-François Trontin[11], Marie-Anne Lelu-Walter[9], Celia Miguel[8], María Teresa Cervera[7], Francisco R. Cantón[1], Christophe Plomion[4,5], Luc Harvengt[11], Concepción Avila[1,2], M. Gonzalo Claros[1,2] and Francisco M. Cánovas[1,2,*]

---

frontiers in Plant Science

ORIGINAL RESEARCH
published: 11 August 2015
doi: 10.3389/fpls.2015.00625

ReprOlive: a database with linked data for the olive tree (*Olea europaea* L.) reproductive transcriptome

*Rosario Carmona[1,2†], Adoración Zafra[1†], Pedro Seoane[3], Antonio J. Castro[1], Darío Guerrero-Fernández[2], Trinidad Castillo-Castillo[4], Ana Medina-García[4], Francisco M. Cánovas[3], José F. Aldana-Montes[4], Ismael Navas-Delgado[4], Juan de Dios Alché[1] and M. Gonzalo Claros[2,3]**

Current Bioinformatics, 2016, *11*, 000-000 1

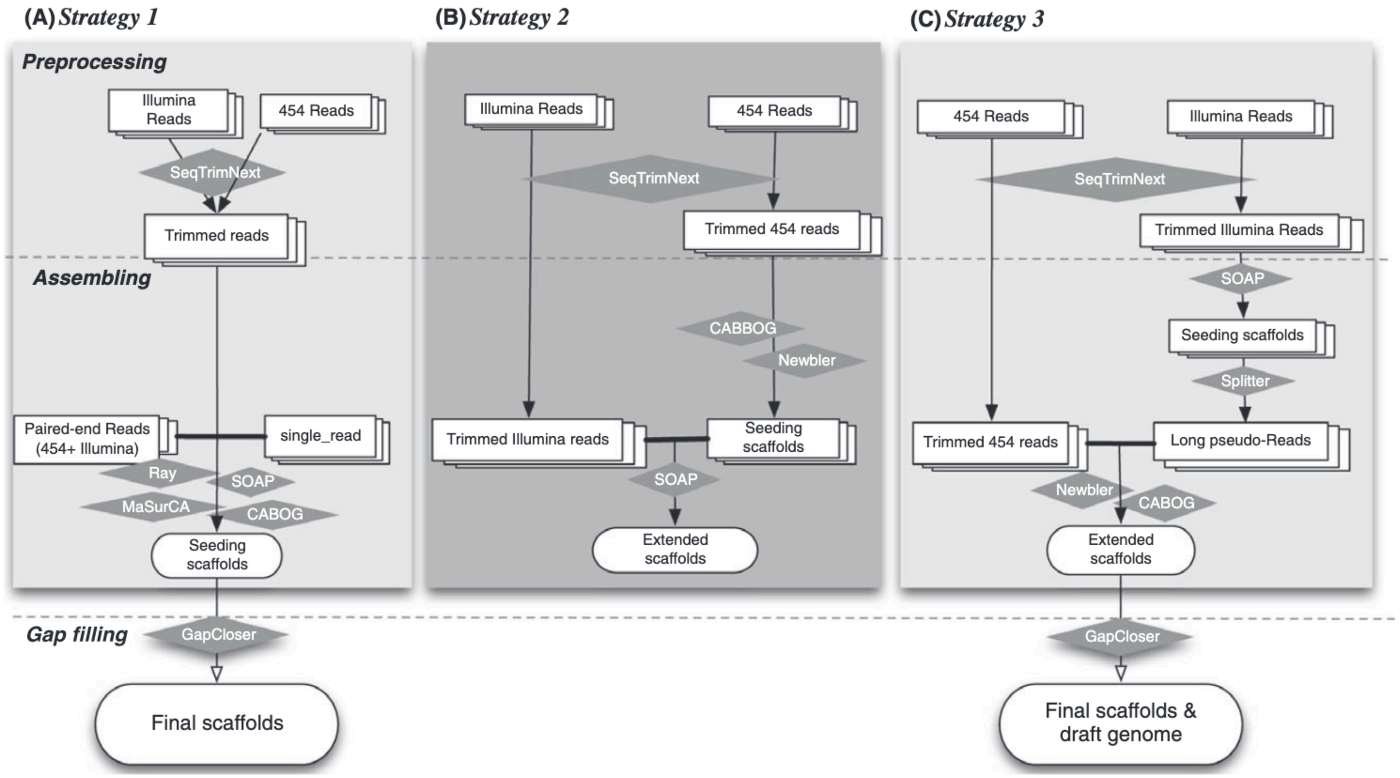
RESEARCH ARTICLE

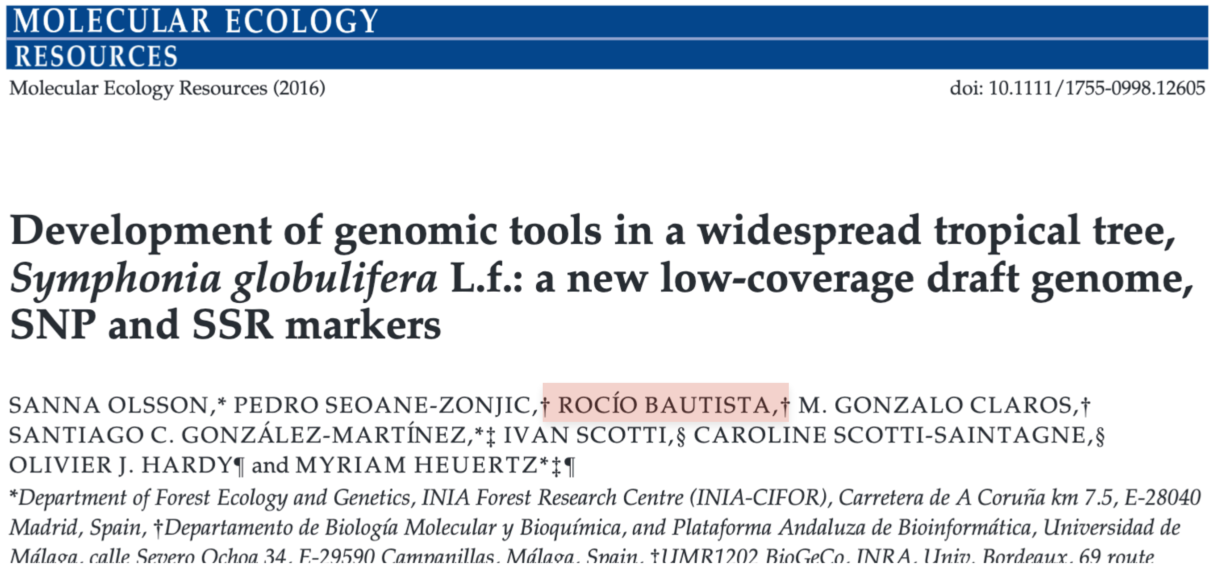
AutoFlow, a Versatile Workflow Engine Illustrated by Assembling an Optimised *de novo* Transcriptome for a Non-Model Species, such as Faba Bean (*Vicia faba*)

Pedro Seoane[1], Sara Ocaña[2], Rosario Carmona[3], Rocío Bautista[3], Eva Madrid[4], Ana M. Torres[2] and Manuel Gonzalo Claros[1,3,*]

De Novo Assembly too

Eukaryotic Genome





MOLECULAR ECOLOGY
RESOURCES

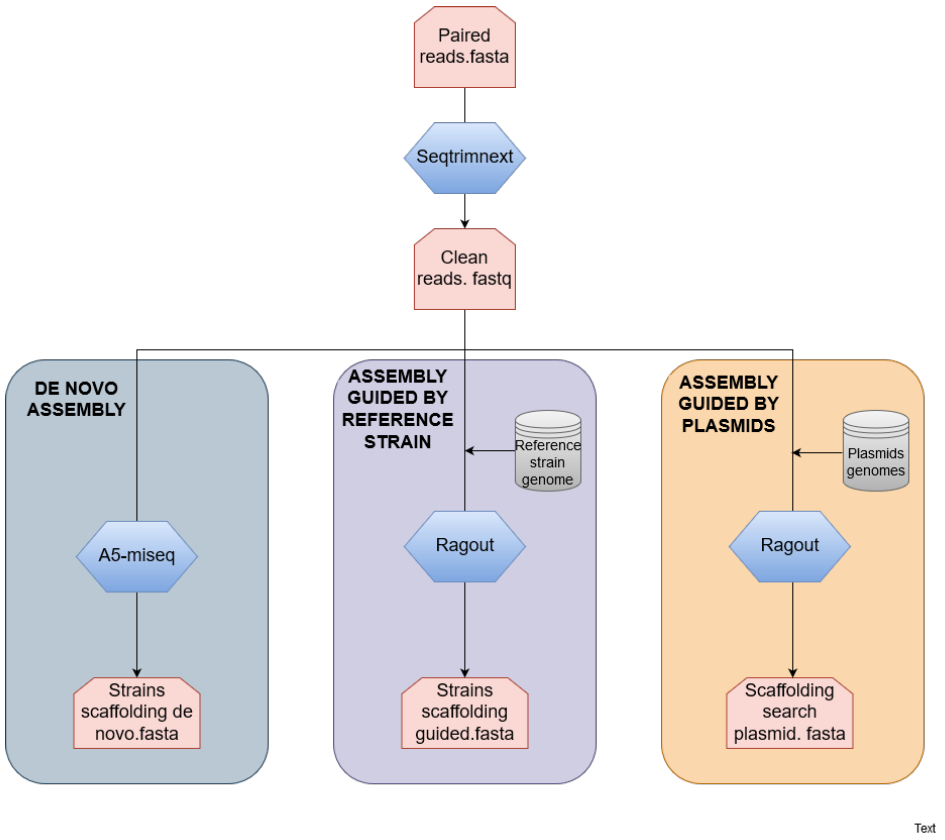
Molecular Ecology Resources (2016) doi: 10.1111/1755-0998.12605

Development of genomic tools in a widespread tropical tree, *Symphonia globulifera* L.f.: a new low-coverage draft genome, SNP and SSR markers

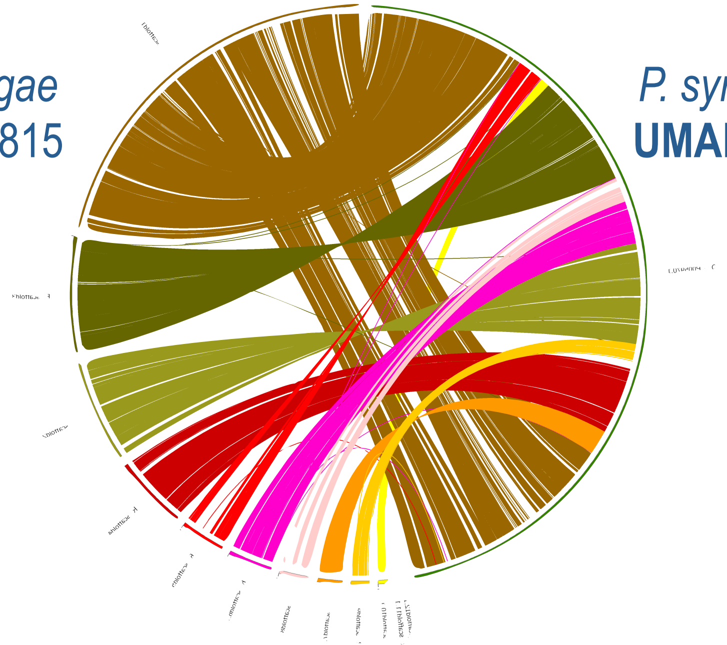
SANNA OLSSON,* PEDRO SEOANE-ZONJIC,† ROCÍO BAUTISTA,† M. GONZALO CLAROS,† SANTIAGO C. GONZÁLEZ-MARTÍNEZ,*‡ IVAN SCOTTI,§ CAROLINE SCOTTI-SAINTAGNE,§ OLIVIER J. HARDY¶ and MYRIAM HEUERTZ*‡¶

**Department of Forest Ecology and Genetics, INIA Forest Research Centre (INIA-CIFOR), Carretera de A Coruña km 7.5, E-28040 Madrid, Spain, †Departamento de Biología Molecular y Bioquímica, and Plataforma Andaluza de Bioinformática, Universidad de Málaga, calle Severo Ochoa 34, E-29590 Campanillas, Málaga, Spain, ‡UMR1202 BioGeCo, INRA, Univ. Bordeaux, 69 route*

Prokaryotic Genome

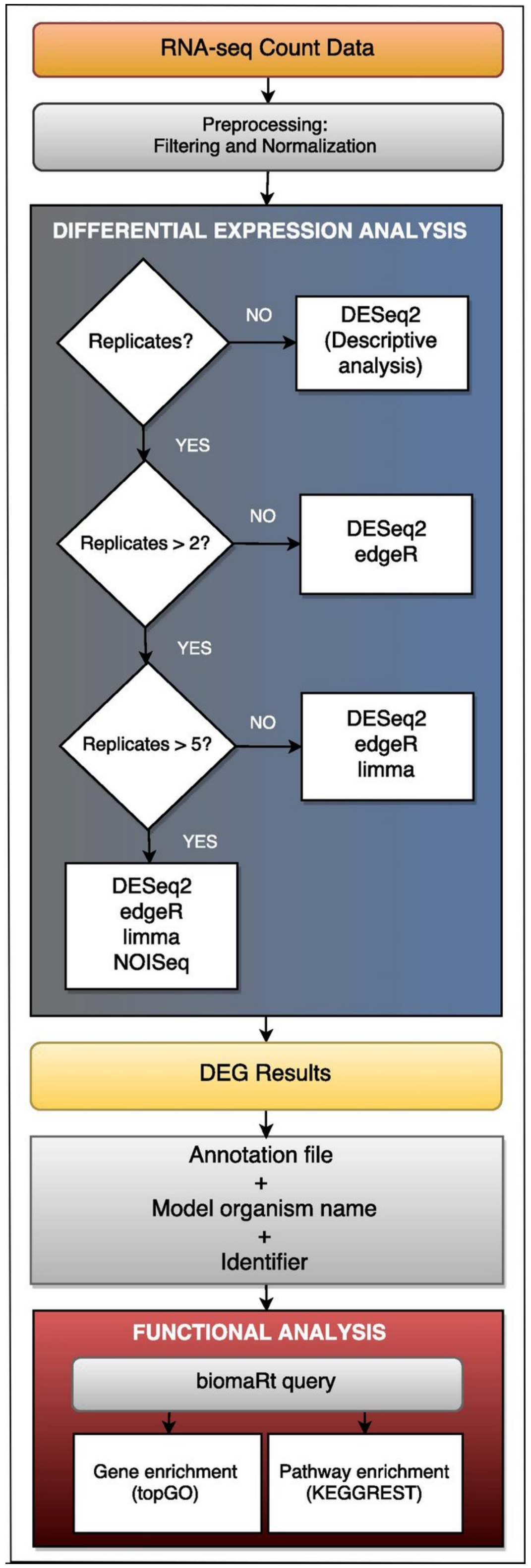


P. syringae
UMAF2815



P. syringae
UMAF0158

Differential expression analysis for RNA-Seq



RNA-seq Count Data

Preprocessing:
Filtering and Normalization

DIFFERENTIAL EXPRESSION ANALYSIS

Replicates?

NO

DESeq2
(Descriptive
analysis)

YES

Replicates > 2?

NO

DESeq2
edgeR

YES

Replicates > 5?

NO

DESeq2
edgeR
limma

YES

DESeq2
edgeR
limma
NOISeq

DEG Results

Annotation file
+
Model organism name
+
Identifier

FUNCTIONAL ANALYSIS

biomaRt query

Gene enrichment
(topGO)

Pathway enrichment
(KEGGREST)

GENOMICS AND COMPUTATIONAL BIOLOGY   Vol. 3 No. 3 (2017): e31

Analytical Workflow

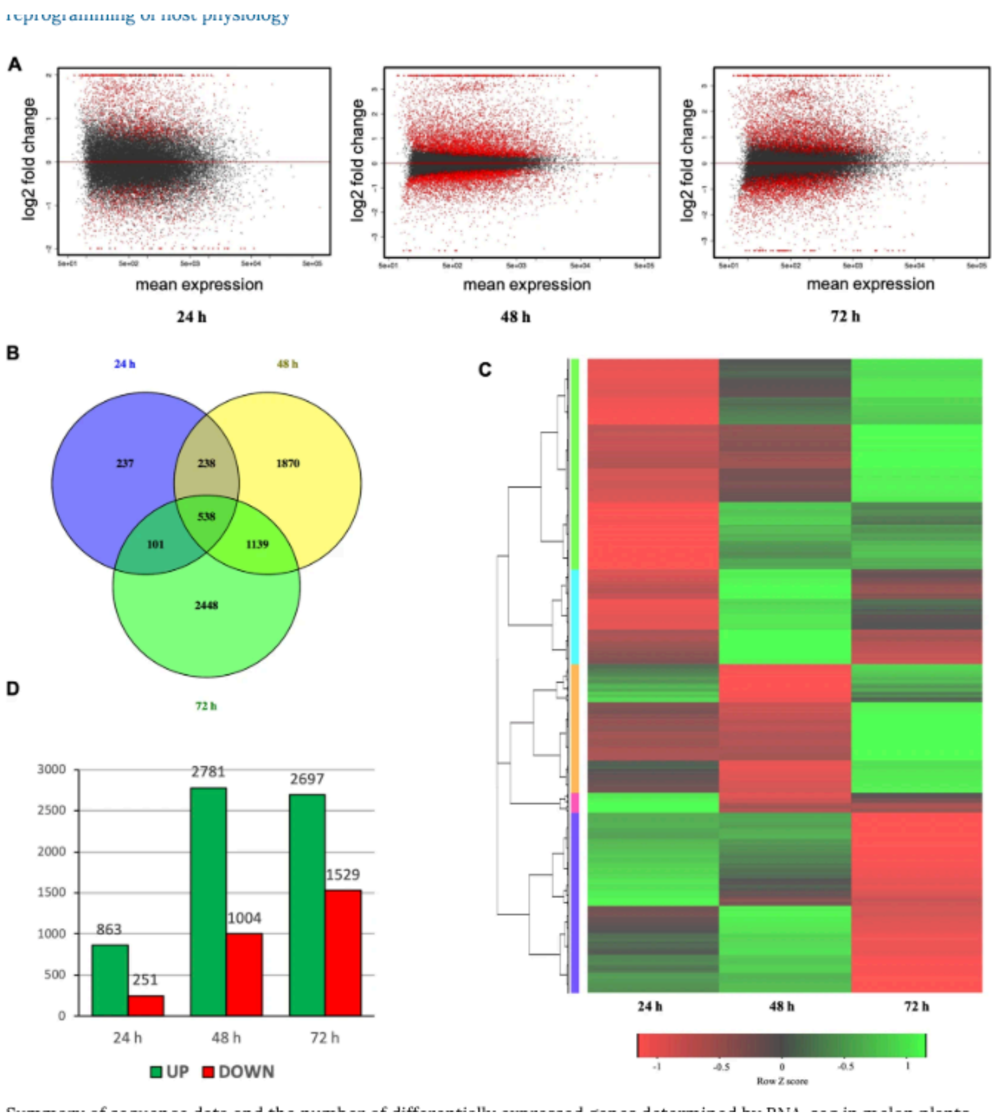
DEgenes Hunter - A Flexible R Pipeline for Automated RNA-seq Studies in Organisms without Reference Genome

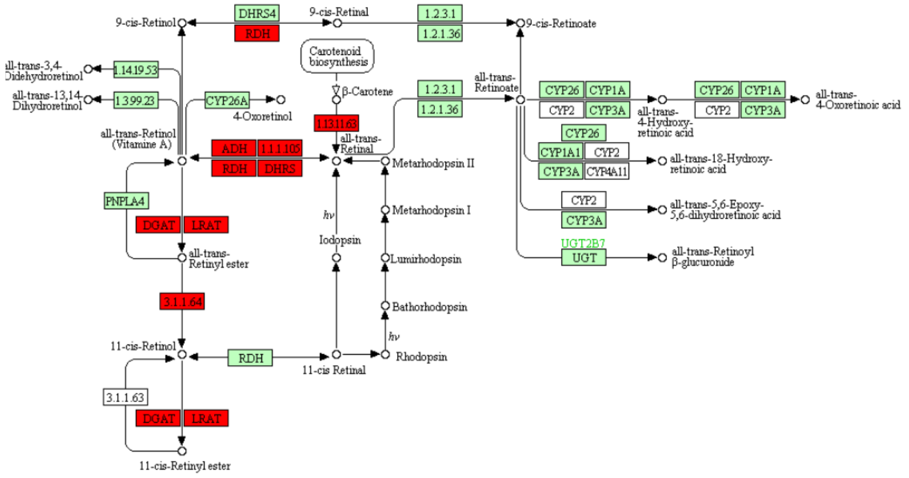
Isabel González Gayte[1], Rocío Bautista Moreno[2], Pedro Seoane Zonjic[1], M. Gonzalo Claros[1,2,*]

[1] Departmento de Biología Molecular y Bioquímica, Facultad de Ciencias, Universidad de Málaga, Malaga, Spain
[2] Plataforma Andaluza de Bioinformática, Universidad de Málaga, C/Severo Ochoa 34, Malaga, Spain

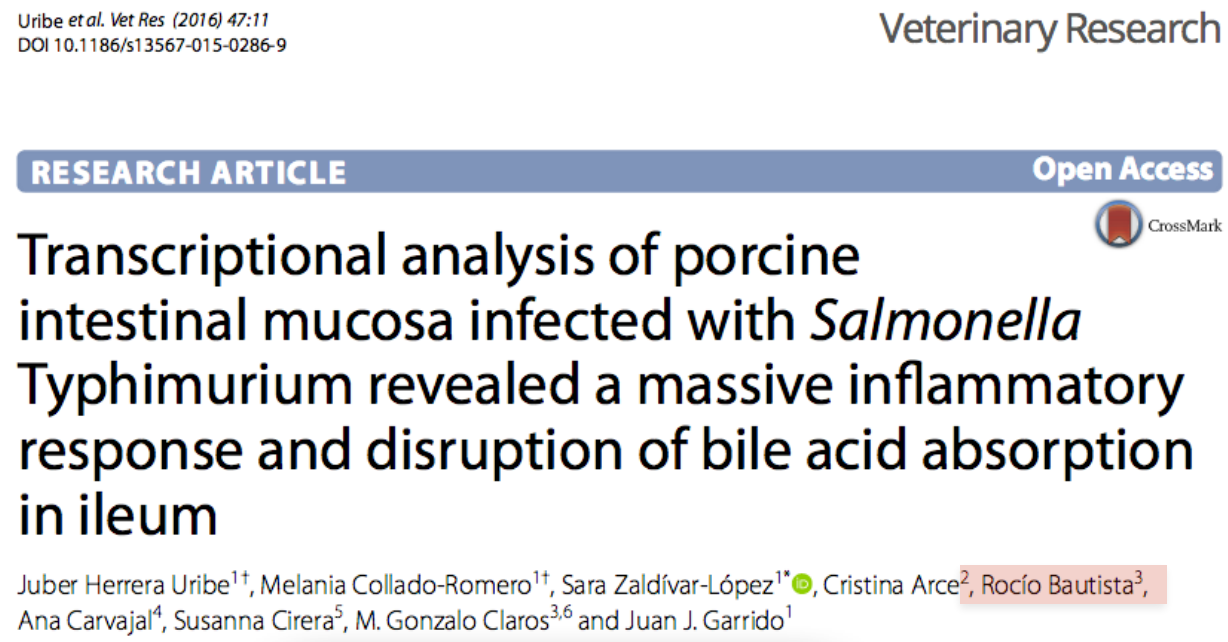
[*] Correspondence: Tel: +34 952 13 72 84; Fax: +34 952 13 40 21; Email: claros@uma.es

Received 2016-03-19; Accepted 2017-02-09





DEgenes Hunter on different studies



Uribe *et al. Vet Res* (2016) 47:11
DOI 10.1186/s13567-015-0286-9

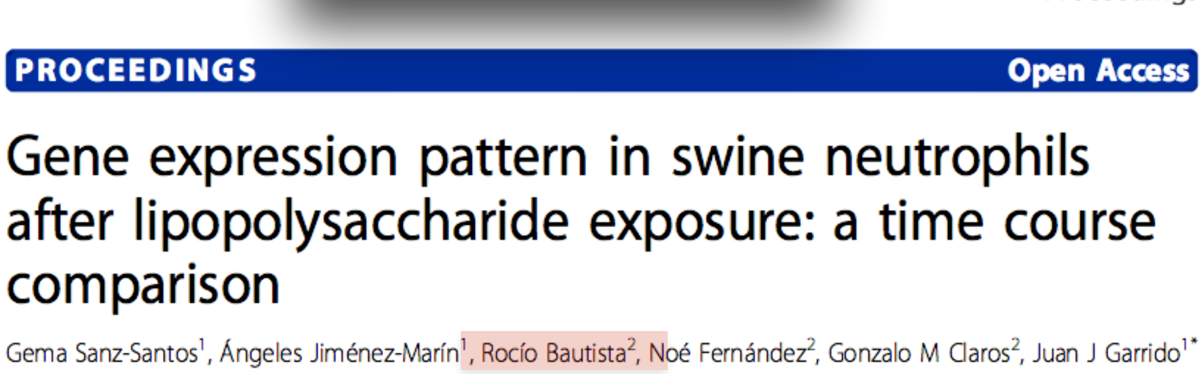
Veterinary Research

RESEARCH ARTICLE **Open Access**

CrossMark

Transcriptional analysis of porcine intestinal mucosa infected with *Salmonella* Typhimurium revealed a massive inflammatory response and disruption of bile acid absorption in ileum

Juber Herrera Uribe[1†], Melania Collado-Romero[1†], Sara Zaldívar-López[1*], Cristina Arce[2], Rocío Bautista[3], Ana Carvajal[4], Susanna Cirera[5], M. Gonzalo Claros[3,6] and Juan J. Garrido[1]



Sanz-Santos *et al. BMC Proceedings*
http://www.biomedcentral.com/17

BMC Proceedings

PROCEEDINGS **Open Access**

Gene expression pattern in swine neutrophils after lipopolysaccharide exposure: a time course comparison

Gema Sanz-Santos[1], Ángeles Jiménez-Marín[1], Rocío Bautista[2], Noé Fernández[2], Gonzalo M Claros[2], Juan J Garrido[1*]

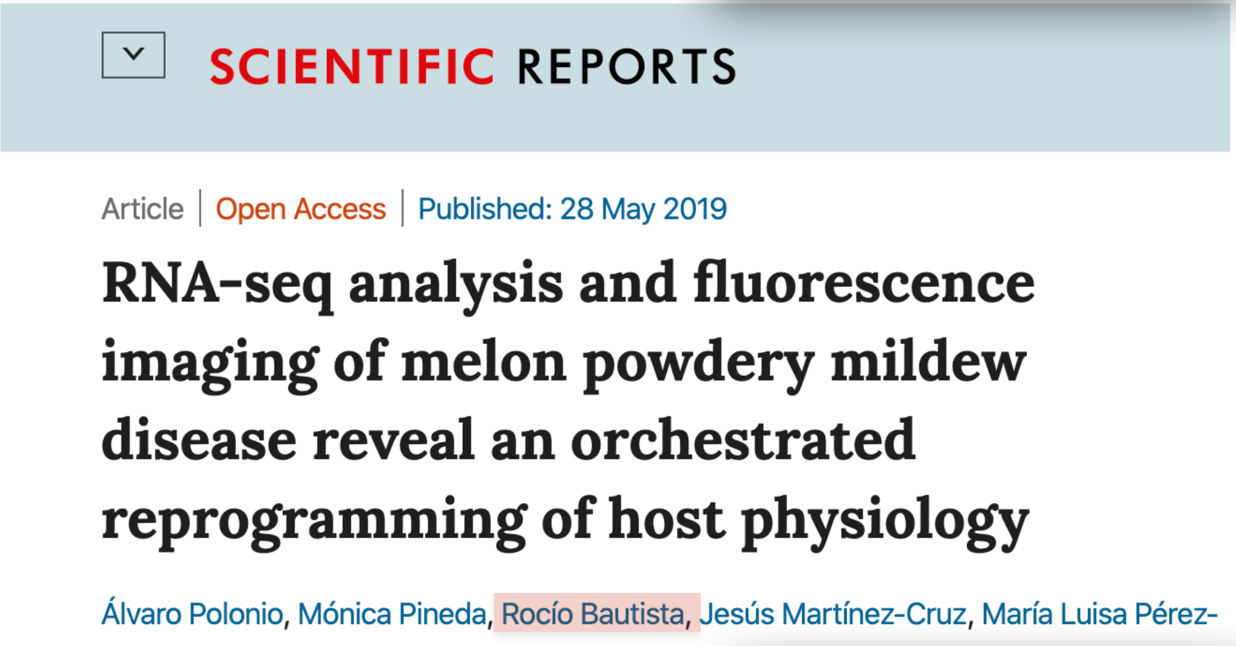
Journal of Experimental Botany
doi:10.1093/jxb/erz136 Advance Access Publication 29 March 2019
This paper is available online free of all access charges (see https://academic.oup.com/jxb/pages/openaccess for further details)

RESEARCH PAPER

Nitric oxide-dependent regulation of sweet pepper fruit ripening

Salvador González-Gordo[1], Rocío Bautista[2], M. Gonzalo Claros[2], Amanda Cañas[1], José M Palma[1], and Francisco J. Corpas[1,*]

[1] Antioxidant, Free Radical and Nitric Oxide in Biotechnology, Food and Agriculture
Superior de Investigaciones Científicas (CSIC), Granada, Spain
[2] Plataforma Andaluza de Bioinformática and Department of Molecular Biology and
Spain.



SCIENTIFIC REPORTS

Article | Open Access | Published: 28 May 2019

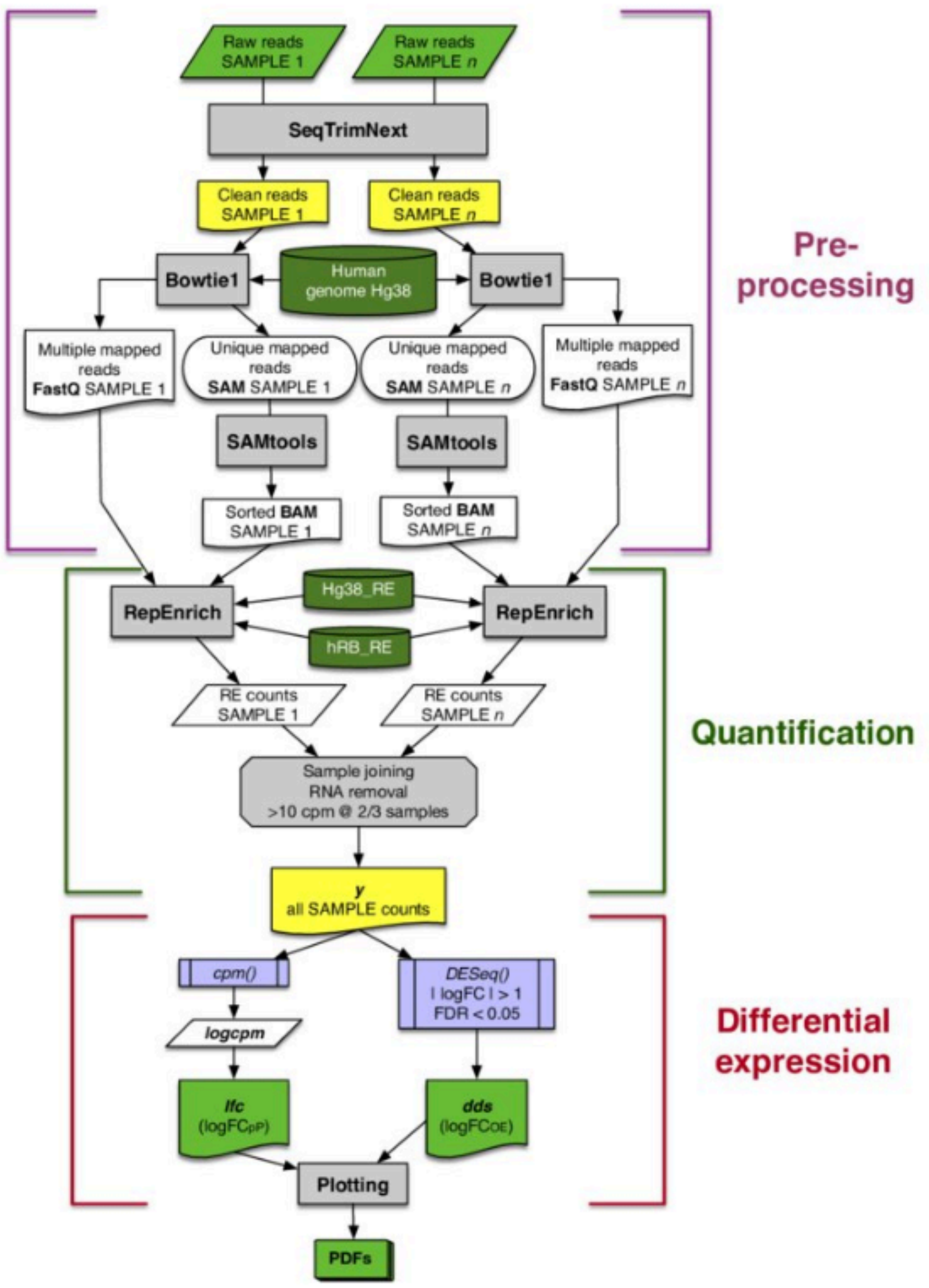
RNA-seq analysis and fluorescence imaging of melon powdery mildew disease reveal an orchestrated reprogramming of host physiology

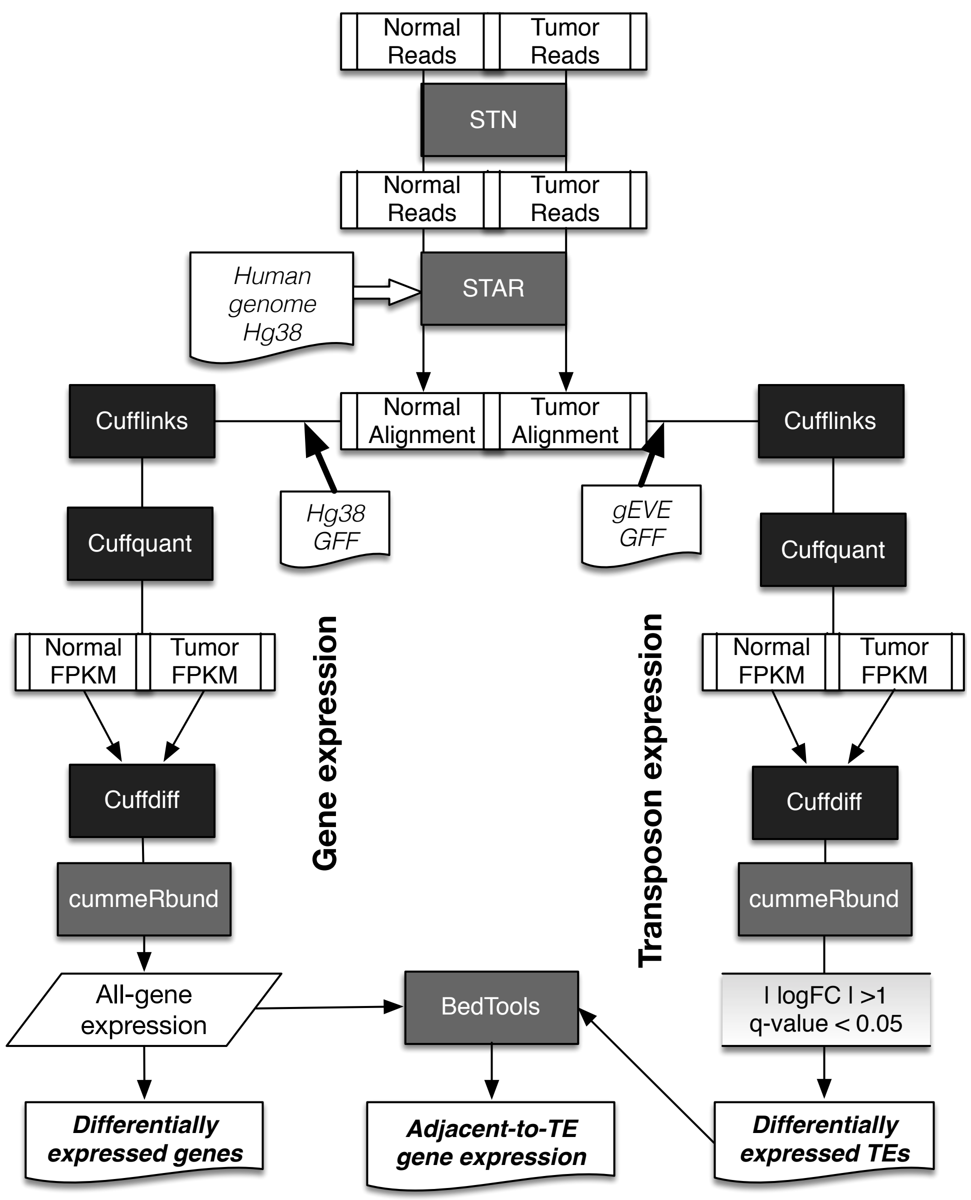
Álvaro Polonio, Mónica Pineda, Rocío Bautista, Jesús Martínez-Cruz, María Luisa Pérez-Bueno, Matilde Barón & Alejandro Pérez-García

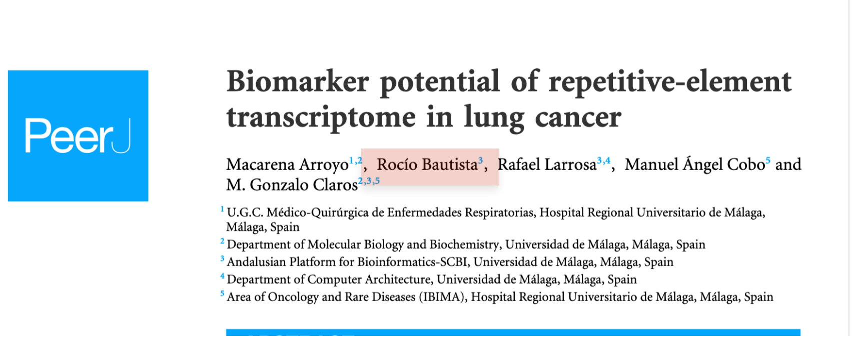
Scientific Reports **9**, Article number: 7978 (2019

1076 Accesses | **3** Citations | **4** Altmetric | M

Identification of potential tissue-specific cancer biomarkers





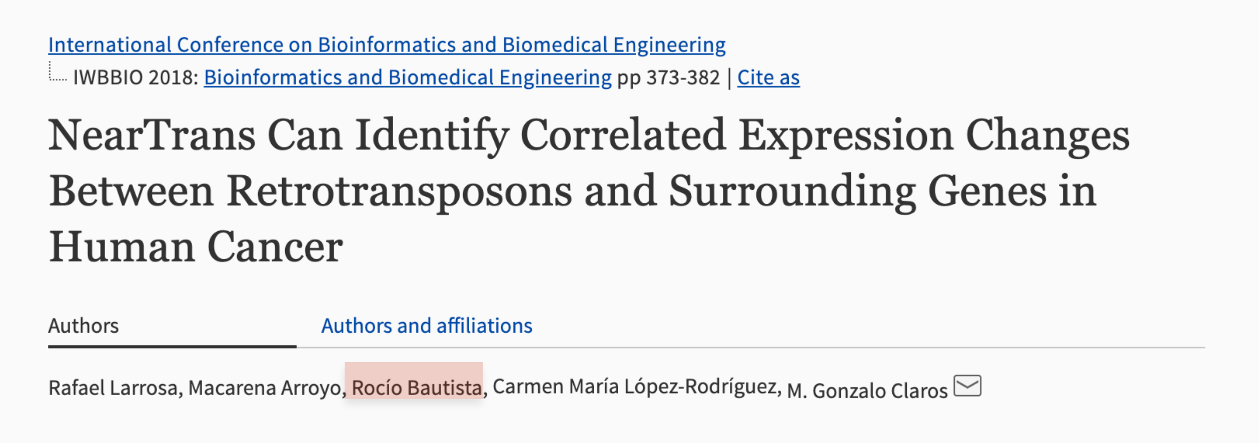


PeerJ

Biomarker potential of repetitive-element transcriptome in lung cancer

Macarena Arroyo[1,2], Rocío Bautista[3], Rafael Larrosa[3,4], Manuel Ángel Cobo[5] and M. Gonzalo Claros[2,3,5]

[1] U.G.C. Médico-Quirúrgica de Enfermedades Respiratorias, Hospital Regional Universitario de Málaga, Málaga, Spain
[2] Department of Molecular Biology and Biochemistry, Universidad de Málaga, Málaga, Spain
[3] Andalusian Platform for Bioinformatics-SCBI, Universidad de Málaga, Málaga, Spain
[4] Department of Computer Architecture, Universidad de Málaga, Málaga, Spain
[5] Area of Oncology and Rare Diseases (IBIMA), Hospital Regional Universitario de Málaga, Málaga, Spain



International Conference on Bioinformatics and Biomedical Engineering
IWBBIO 2018: Bioinformatics and Biomedical Engineering pp 373-382 | Cite as

NearTrans Can Identify Correlated Expression Changes Between Retrotransposons and Surrounding Genes in Human Cancer

Authors | Authors and affiliations

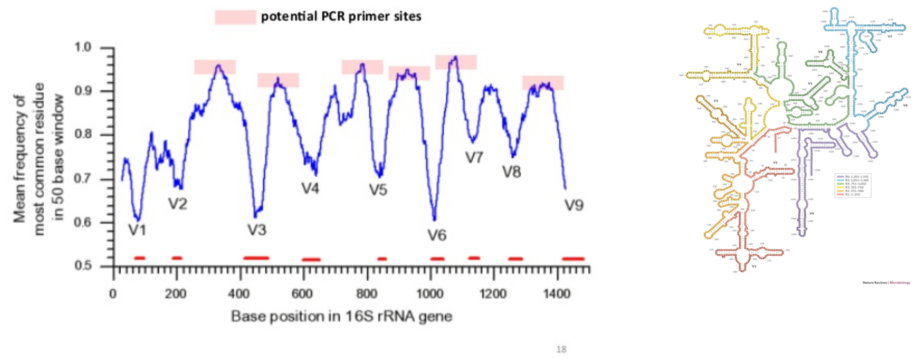
Rafael Larrosa, Macarena Arroyo, Rocío Bautista, Carmen María López-Rodríguez, M. Gonzalo Claros

rociobm@uma.es

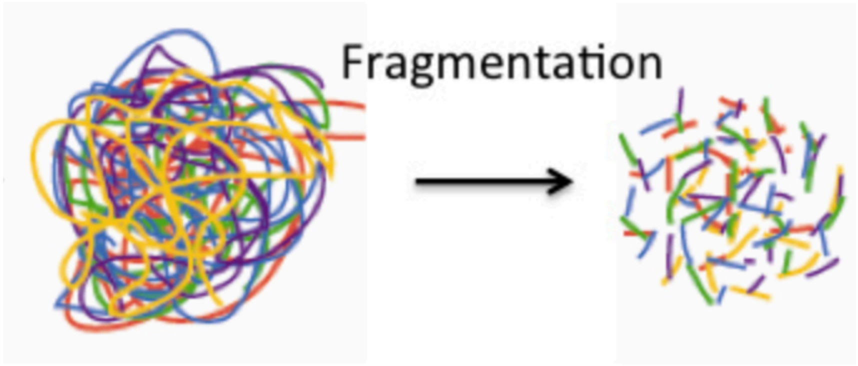
Workflow for analysis of metagenomic sequencing data

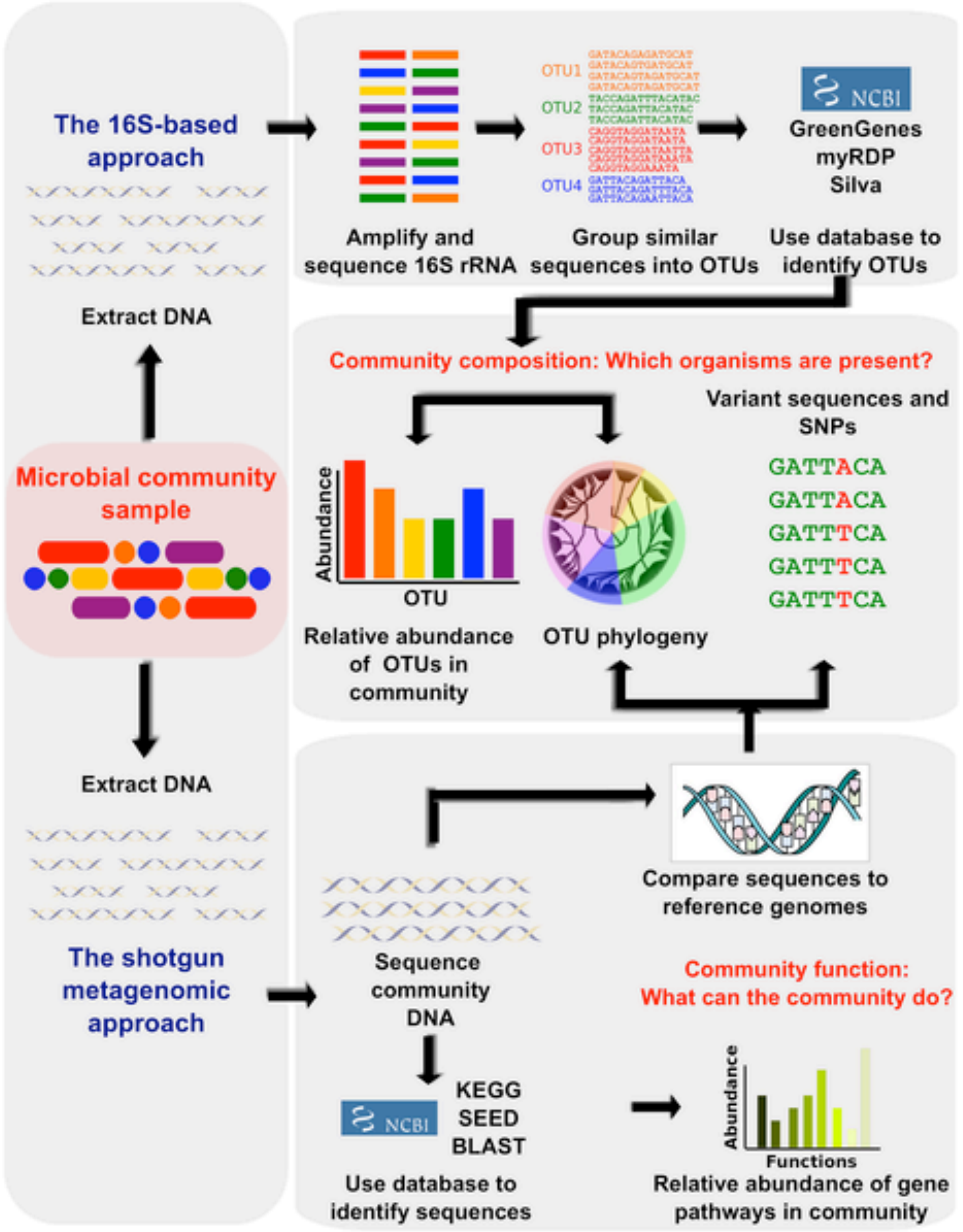
Target 16S

Variable Regions of the 16S rRNA:



Whole genome





The 16S-based approach

Extract DNA

Microbial community sample

Extract DNA

The shotgun metagenomic approach

Amplify and sequence 16S rRNA

Group similar sequences into OTUs

GreenGenes
myRDP
Silva

Use database to identify OTUs

Community composition: Which organisms are present?

Variant sequences and SNPs

Relative abundance of OTUs in community

OTU phylogeny

Compare sequences to reference genomes

Sequence community DNA

Community function: What can the community do?

KEGG
SEED
BLAST

Use database to identify sequences

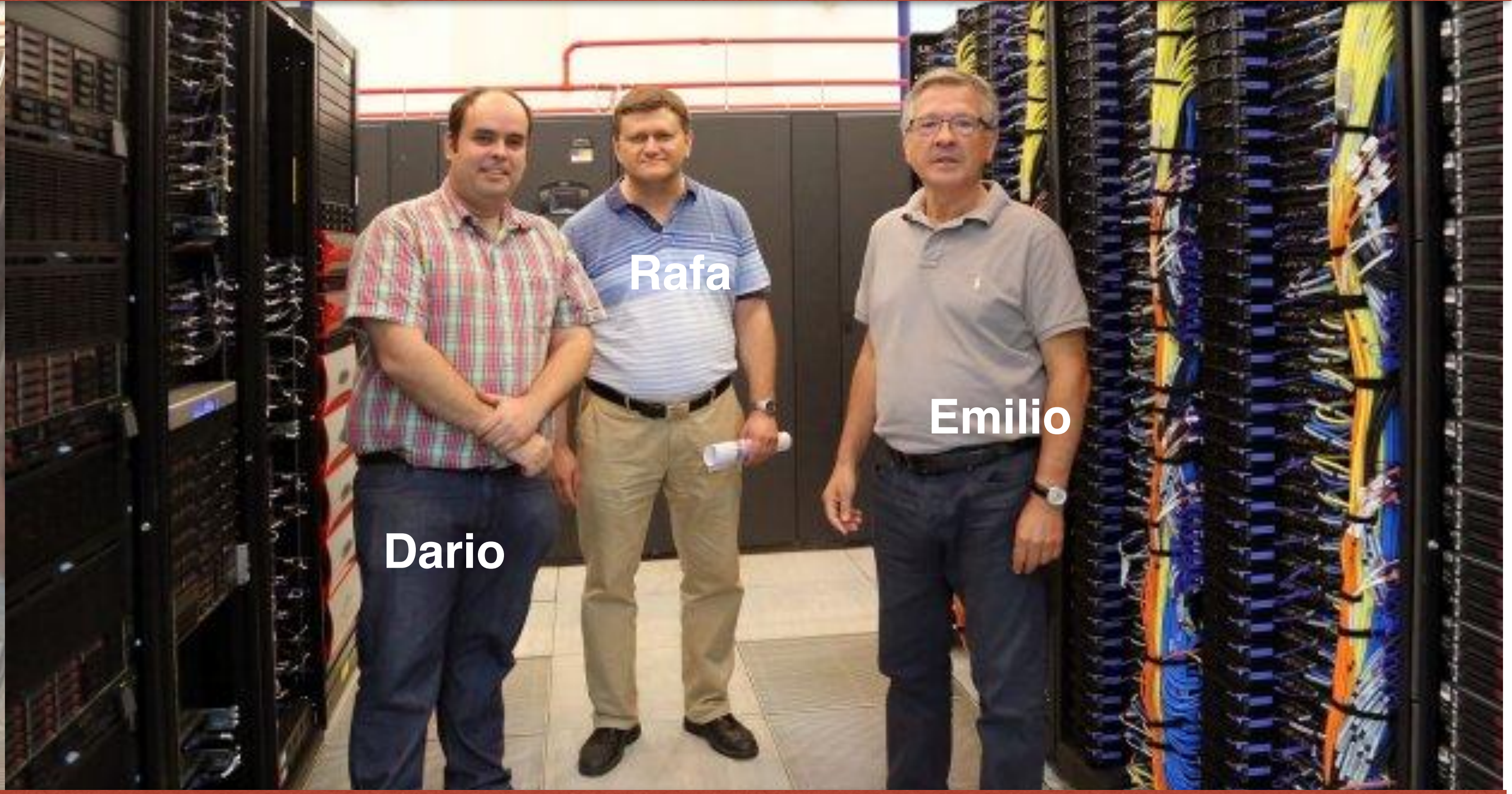
Relative abundance of gene pathways in community

Morgan XC, Huttenhower C (2012) Chapter 12: Human Microbiome Analysis. PLOS Computational Biology 8(12):

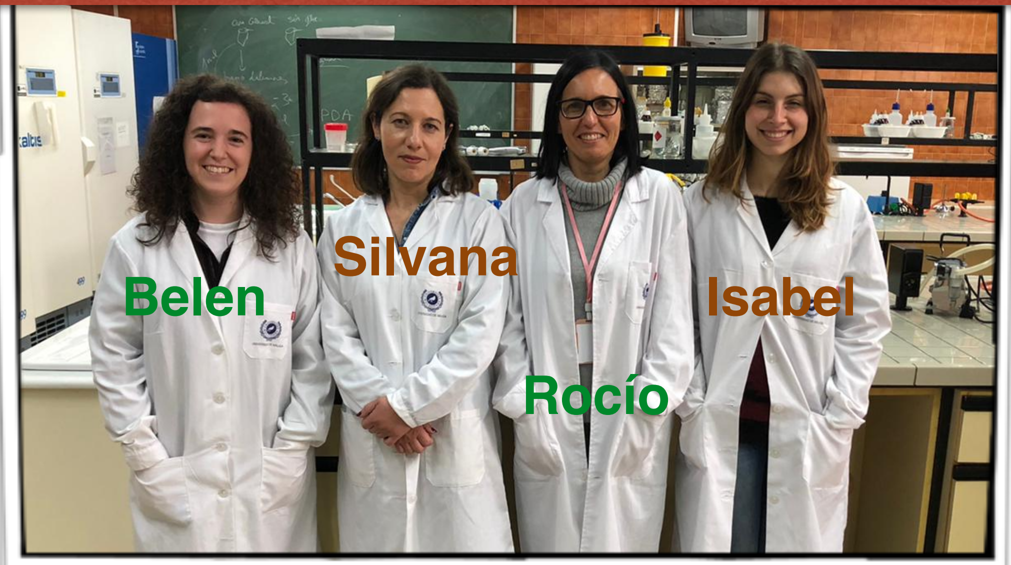
rociobm@uma.es

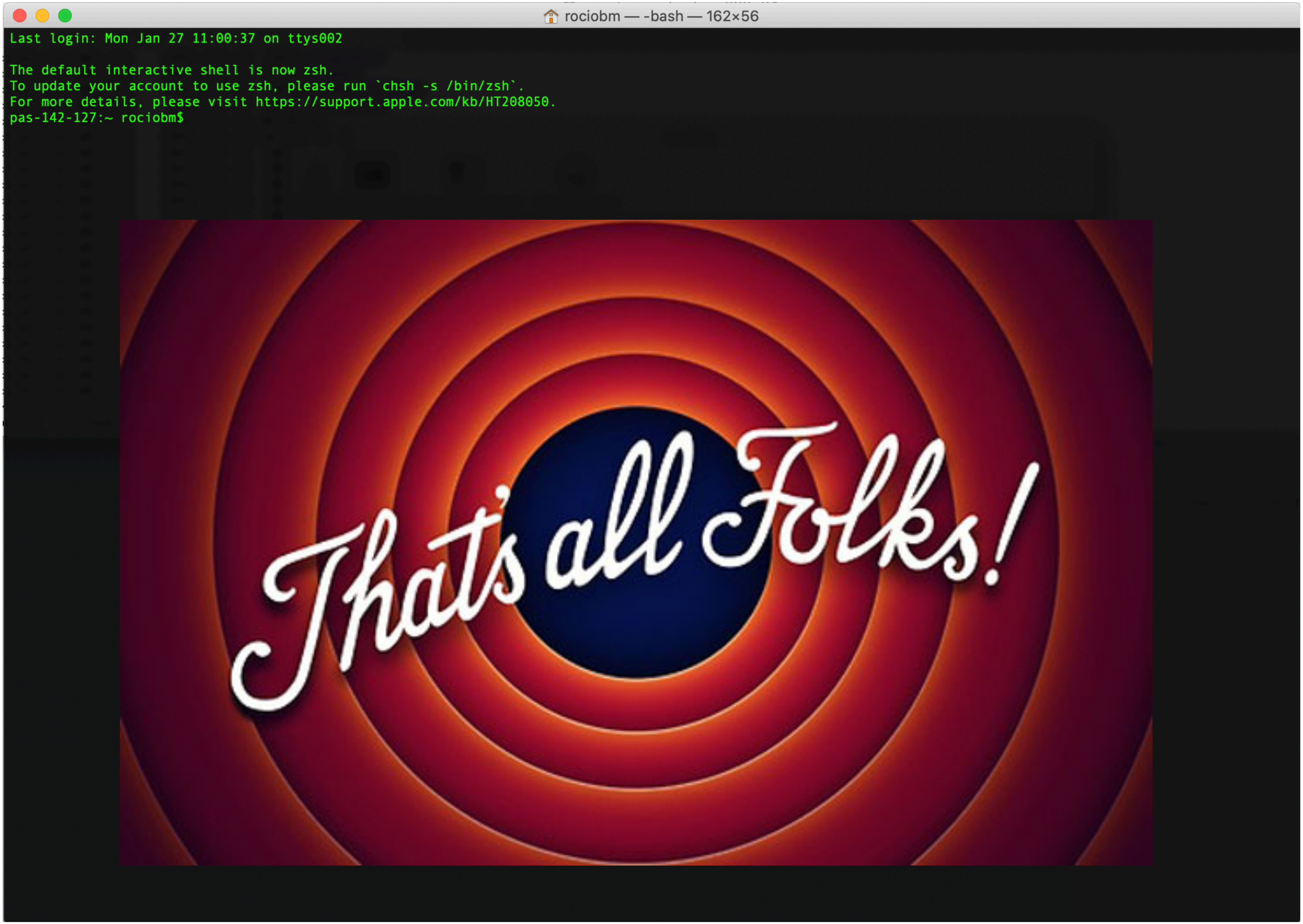
Co-workers!!!!!!!!!

Bioinformatics Team



Omics Team





```
Last login: Mon Jan 27 11:00:37 on ttys002

The default interactive shell is now zsh.
To update your account to use zsh, please run `chsh -s /bin/zsh`.
For more details, please visit https://support.apple.com/kb/HT208050.
pas-142-127:~ rociobm$
```