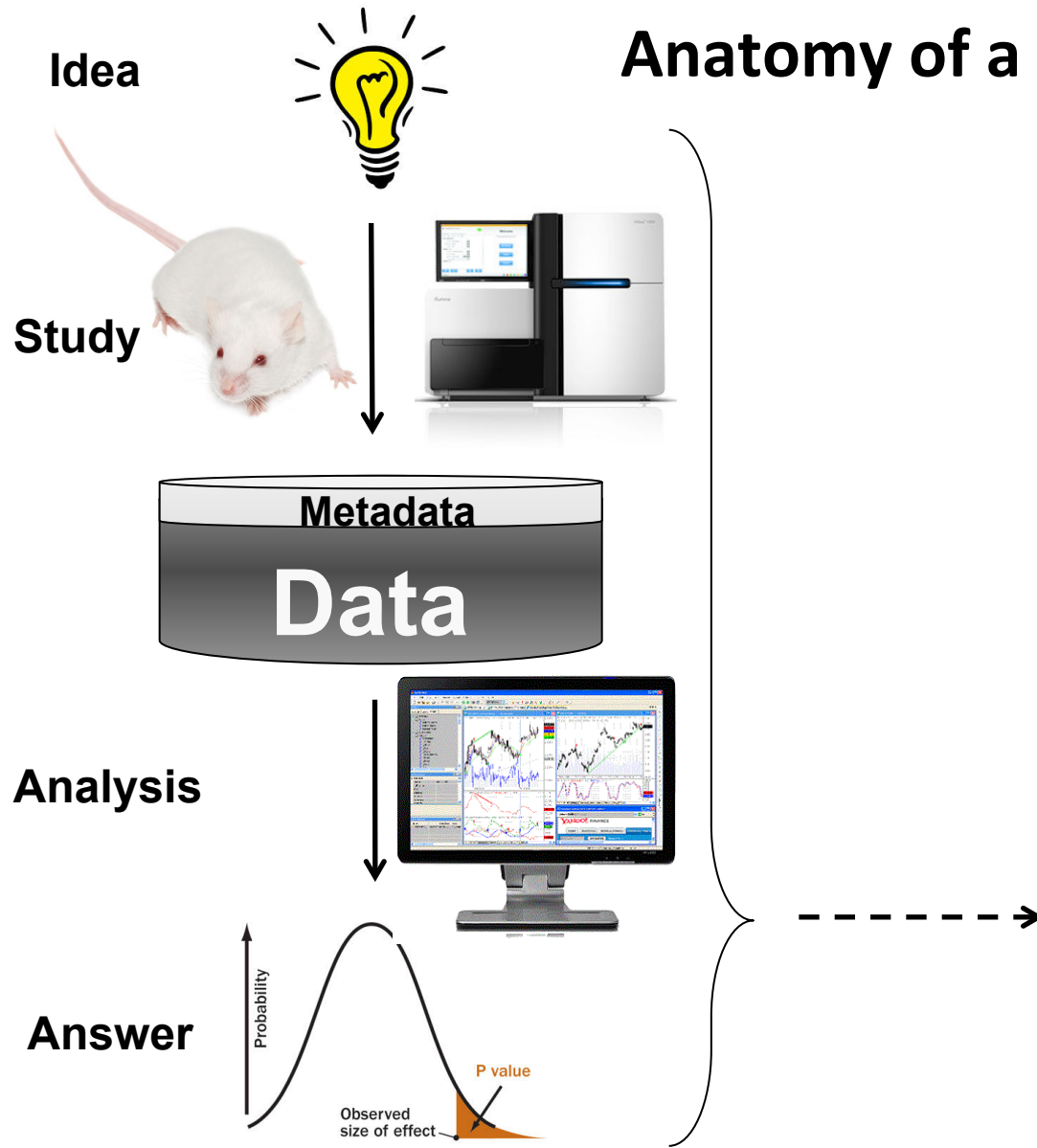


# Publish your tools For REAL!

Rob Davidson  
*GigaScience*  
@bobbledavidson   
GCC 2015

# Anatomy of a traditional Publication



Wilson et al. *GigaScience* 2012, 1:3  
<http://www.gigascejournal.com/content/1/1/3>

(GIGA)<sup>n</sup>  
SCIENCE

RESEARCH

Open Access

## Resources for methylome analysis suitable for gene knockout studies of potential epigenome modifiers

Gareth A Wilson<sup>1\*</sup>, Pawandeep Dhali<sup>1</sup>, Andrew Feber<sup>1</sup>, Daniel Cortázar<sup>2</sup>, Yuka Suzuki<sup>1</sup>, Reiner Schulz<sup>3</sup>, Primo Schär<sup>4</sup> and Stephan Beck<sup>5</sup>

### Abstract

**Background:** Methylated DNA immunoprecipitation (MeDIP) is a popular enrichment based method and can be combined with sequencing (termed MeDIP-seq) to interrogate the methylation status of cytosines across entire genomes. However, quality control and analysis of MeDIP-seq data have remained to be a challenge.

**Results:** We report genome-wide DNA methylation profiles of wild type (wt) and mutant mouse cells, comprising 3 biological replicates of Thymine DNA glycosylase (Tdg) knockout (KO) embryonic stem cells (ESCs), in vitro differentiated neural precursor cells (NPCs) and embryonic fibroblasts (MEFs). The resulting 18 methylomes were analysed with MeDUSA (Methylated DNA Utility for Sequence Analysis), a novel MeDIP-seq computational analysis pipeline for the identification of differentially methylated regions (DMRs). The observed increase of hypermethylation in MEF promoter-associated CpG islands supports a previously proposed role for Tdg in the protection of regulatory regions from epigenetic silencing. Further analysis of genes and regions associated with the DMRs by gene ontology, pathway, and ChIP analyses revealed further insights into Tdg function, including an association of Tdg with low-methylated distal regulatory regions.

**Conclusions:** We demonstrate that MeDUSA is able to detect both large-scale changes between cells from different stages of differentiation and also small but significant changes between the methylomes of cells that only differ in the KO of a single gene. These changes were validated utilising publicly available datasets and confirm Tdg's function in the protection of regulatory regions from epigenetic silencing.

**Keywords:** Methylome, MeDIP-seq, Epigenetics, Epigenomics, DNA methylation, Computational pipeline, MeDUSA

## Idea



## Study



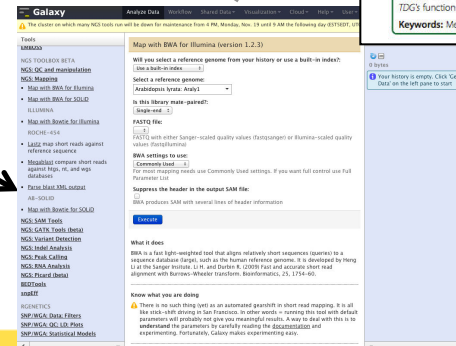
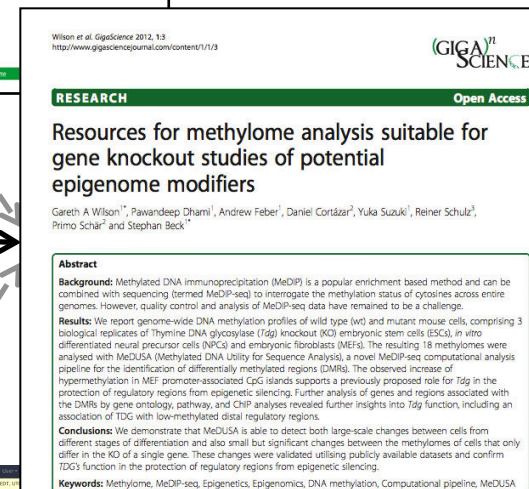
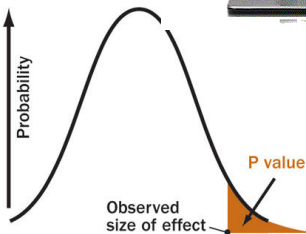
## Metadata

# Data

## Analysis



## Answer



# Reproducing results? SOAPdenovo2 *S. aureus* pipeline

Galaxy - Mozilla Firefox

Galaxy

192.168.171.50:8080/galaxy/root

Galaxy / CBIIT-Giga

Analyze Data Workflow Shared Data Visualization Admin Help User

Using 718.8 MB

Tools

search tools

BGI SOAP PACKAGE BETA

NGS: Mapping

NGS: De Novo Assembly

- SOAPdenovo1
- SOAPdenovo2

SOAPDENOV2 MODULES

- pregraph - construct Bruijn graph
- pregraph\_sparse
- contig identification from overlapping sequence reads

80 98.6 25 71.5 38 1086 2 1078

History

saureus 552.5 MB

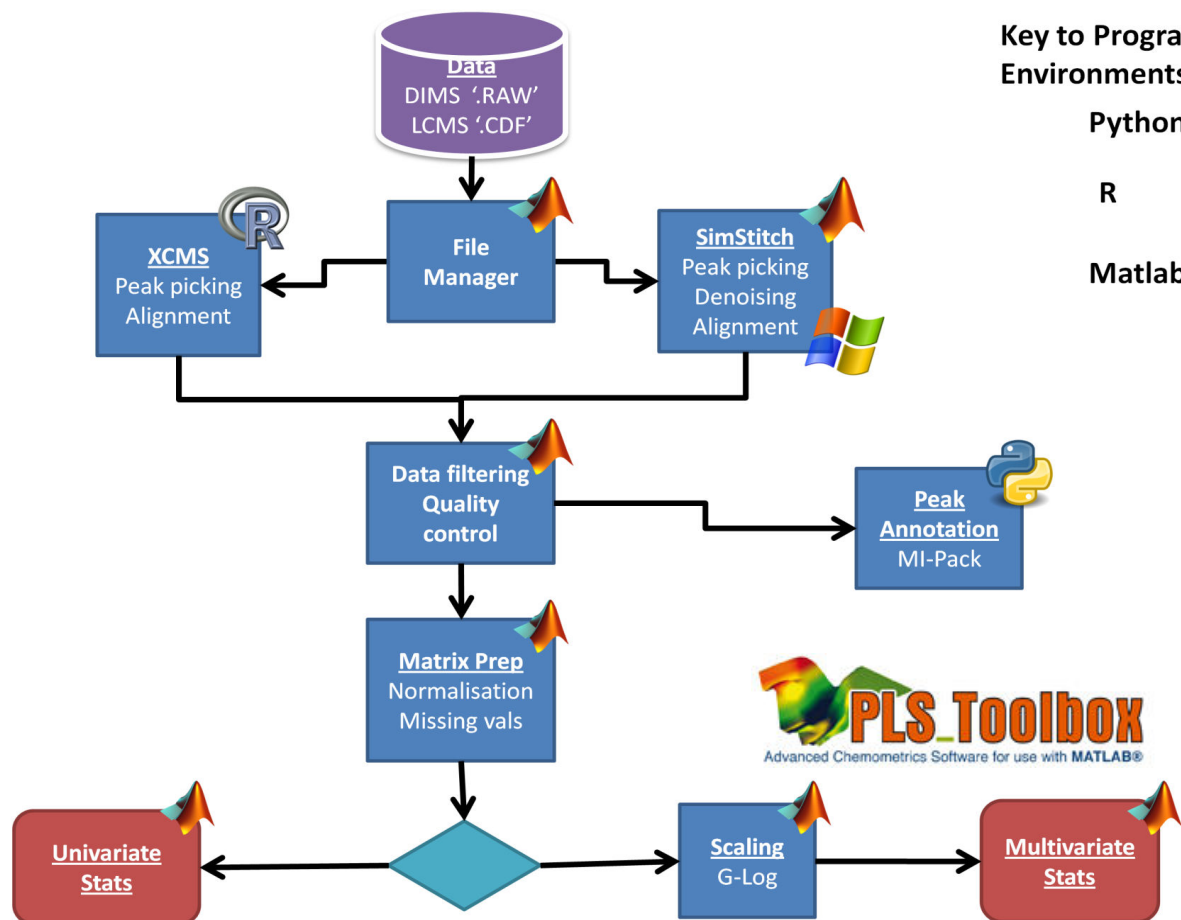
- 153: GAGE.output
- 146: scafseq.gc.ctg.fa
- 145: Information about closed gaps
- 144: gapclo.scaff
- 140: SOAPDvo2: Scaffolds
- 139: SOAPDvo2: Contigs

**Table 2 Assemblies of *S. aureus* and *R. sphaeroides***


Species	Version	Contigs				Scaffolds			
		Number	N50 (kb)	Errors	N50 corrected(kb)	Number	N50 (kb)	Errors	N50 corrected (kb)
<i>S. aureus</i>	SOAPdenovo1	79	148.6	156	23	49	342	0	342
	SOAPdenovo2	80	98.6	25	71.5	38	1,086	2	1,078
	ALLPATHS-LG*	37	149.7	13	117.6	10	1,477	1	1,093
<i>R. sphaeroides</i>	SOAPdenovo1	2,242	3.5	392	2.8	956	105	18	70
	SOAPdenovo2	721	18	106	14.1	333	2,549	4	2,540
	ALLPATHS-LG*	190	41.9	31	36.7	32	3,191	0	3,310

All datasets were downloaded from <http://gagc.chch.umd.edu/data/>

# github.com/Viant-Metabolomics/Galaxy-M



Key to Programming Environments:

Python 

R 

Matlab 

- Many tools
- Many languages
- Complex to learn
- Many parameters
- Complex to report

# Open Peer Review

- “I think important aspects of reproducibility are lost when building on closed source and non-free applications.”
- “To be frank, if this were a genomics article I would recommend not publishing a purely computational methods paper when large parts of the pipeline are non-free and closed source - limiting both the reproducibility and transparency of the pipeline. Realistically though my understanding is that this is quite common in metabolomics”
- “I would have indicated the paper was of more broad interest if there was at least one complete open source pipeline for data analysis”



# N.B.

- Big thanks to Dave Clements and Galaxy Team
  - Reviews, support, building galaxy
- Planemo for test-driven peer review??
- Containers for plug-and-play data-driven peer-review (bioboxes)

# Integrating workflows with papers

## Publish in the *GigaScience* Special Galaxy Series and benefit from:

- **Quick publication**– average time to first decision in 2013/14 less than **25 days**
- **15% Article Processing Charge discount** (£200) to all submissions from GCC2015
- **Free deposition and curation** of your data in **GigaDB database** with no size limit
- **All data and tools can be hosted** with the journal's **gigagalaxy.net** server
- **A home & citeable DOI for data & workflows**

(GIGA)<sup>n</sup>  
SCIENCE

(GIGA)<sup>n</sup> Galaxy

Editor-in-Chief: Laurie Goodman  
Executive Editor: Scott Edmunds

华大基因  
BGI