Managing data and code in the chemical biology laboratory

Sergio Martínez Cuesta

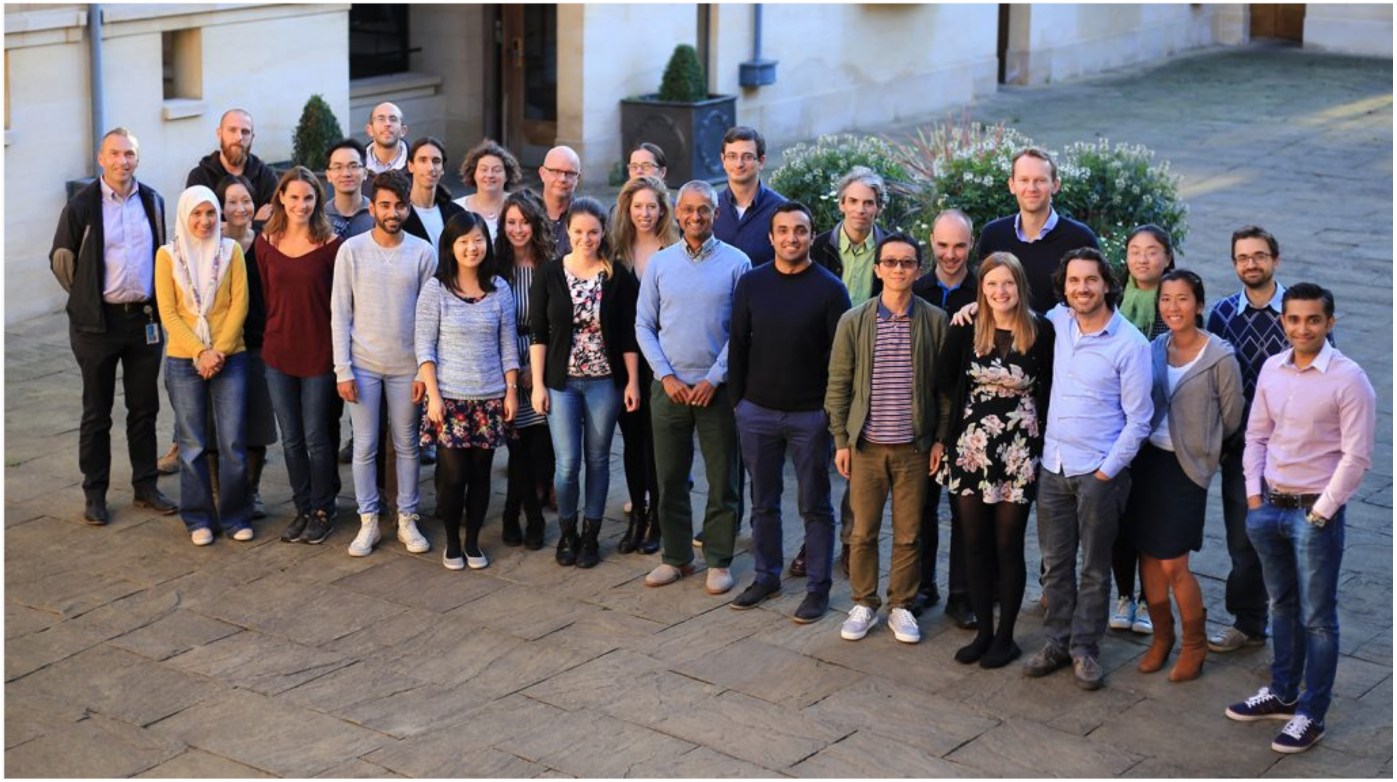
WOLFSON COLLEGE
CAMBRIDGE

CANCER
RESEARCH
UK

CAMBRIDGE
INSTITUTE

The Balasubramanian laboratories

The chemistry and biology of nucleotide modifications and G-quadruplexes in DNA and RNA



90% Experimental   10% Computational

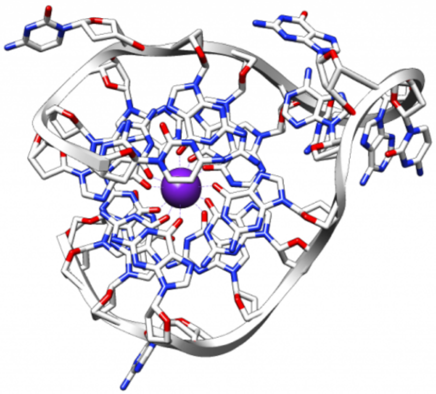
Projects and methods

Nucleotide modifications

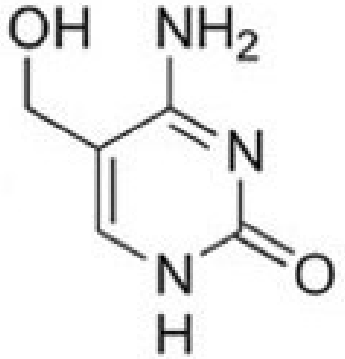
- Mapping modified bases (e.g. 5hmC and 5fC) in genomes and transcriptomes
- Quantifying abundances using mass spectrometry
- Chemical synthesis

G-quadruplexes

- Mapping in DNA and RNA
- Biophysical characterization
- Drug discovery



G4



5hmC

Data

Illumina MiSeq and NextSeq

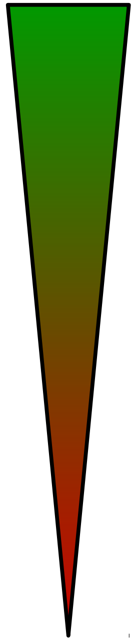
Primary: `.fastq` files

Containing raw sequencing reads

Secondary: `.bam` and `.bed` files

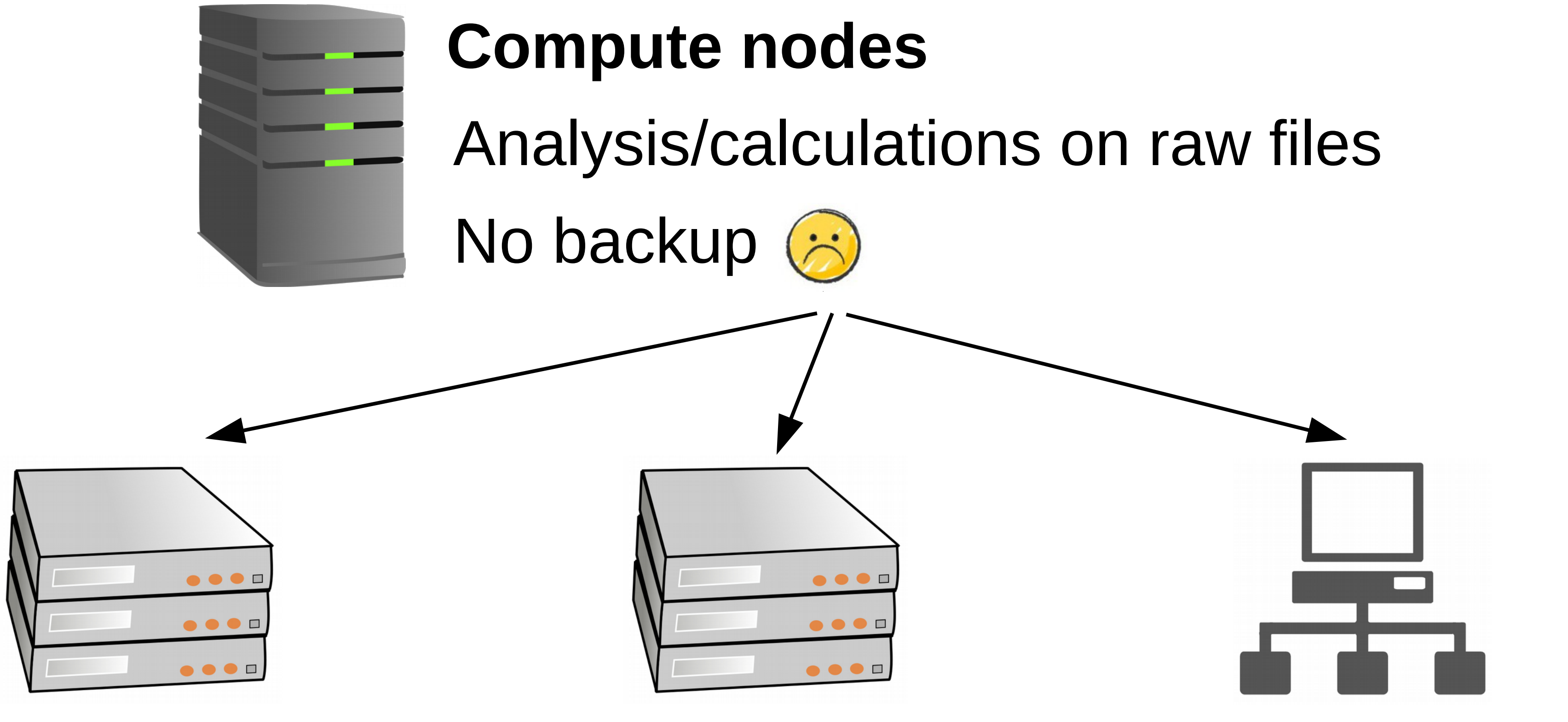
Containing processed (aligned) reads

Tertiary: tables and figures



Long-term storage importance

Data management



Compute nodes

Analysis/calculations on raw files

No backup ☹

Archive

Primary

Backup 😀

Non-deletable
(after 48h)

Server

Secondary

Backup 😀

Limited (12TB)

Public folders

Tertiary

Backup 😀

To share with
collaborators

Data management

But what if archive (and backup) fails?

Genomics core

LIMS

BaseSpace

SEQUENCE HUB

Does it work? Ideally ...

Primary ← **Secondary** ← **Tertiary**

Primary files naming convention

e.g. fk468_PCF-b2-oxhyd-1.fastq.gz

Collaborator: fk (Fumiko Kawasaki)

Experiment number: 468

Life cycle stage: PCF

Batch: b2

Treatment: oxhyd

Replicate: 1

Project timeline

Close data

Open data

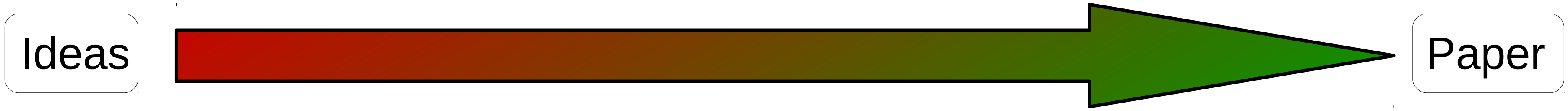
Initial experiments

Method optimisation

Answering biological questions

Write up story

ArrayExpress



Ideas

Paper

Tests

First scripts

Robust analysis pipeline

Figures Tables

Document computational method

ENA

Private code

Public code

GitHub

Code management

From tests to computational method ...

Python R C
Java Bash

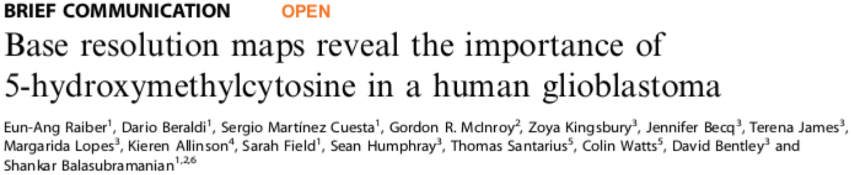
ATOM textmate

Markdown
(.md)

https://github.com/
sblab-bioinformatics

/projects 🔒 early stages of development

/epigenetics-of-glioblastoma 🔓

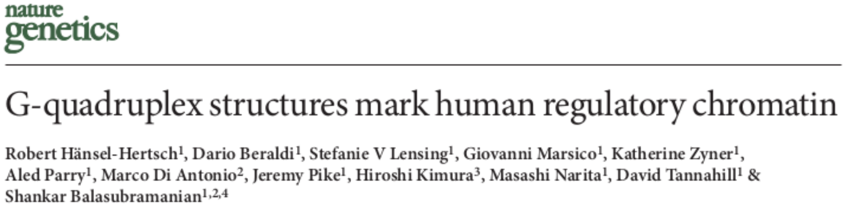


BRIEF COMMUNICATION OPEN

Base resolution maps reveal the importance of 5-hydroxymethylcytosine in a human glioblastoma

Eun-Ang Raiber[1], Dario Beraldi[1], Sergio Martínez Cuesta[1], Gordon R. McInroy[2], Zoya Kingsbury[3], Jennifer Becq[3], Terena James[3], Margarida Lopes[3], Kieren Allinson[4], Sarah Field[1], Sean Humphray[3], Thomas Santarius[5], Colin Watts[5], David Bentley[3] and Shankar Balasubramanian[1,2,6]

/dna-secondary-struct-chrom-lands 🔓



nature genetics

G-quadruplex structures mark human regulatory chromatin

Robert Hänsel-Hertsch[1], Dario Beraldi[1], Stefanie V Lensing[1], Giovanni Marsico[1], Katherine Zyner[1], Aled Parry[1], Marco Di Antonio[2], Jeremy Pike[1], Hiroshi Kimura[3], Masashi Narita[1], David Tannahill[1] & Shankar Balasubramanian[1,2,4]

...

Challenges

What data/code to share? When?

What do we do with the private code when the project is finished?

Is internal peer review of code useful before public release?

Questions? Thanks!