EGI CF 2015, BARI, Long tail of science: tools and services, 11 Nov. 2015

Opportunistic use of supercomputers: linking with Cloud and Grid platforms

Presented by Luis Cabellos

Co-authors: Fernando Aguilar, Jesús Marco, Aida Palacios

Instituto de Física de Cantabria (IFCA)

UC, University of Cantabria

CSIC, NATIONAL RESEARCH COUNCIL

SANTANDER, SPAIN

View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Digital.CSIC

Scope

- Introducing TRUFA: an HPC NGS application
- TRUFA and the long tail of science
- Impact on Supercomputer usage
- Linking to the Grid & Cloud worlds
- Opportunistic use of supercomputers: linking with Cloud and Grid

What is TRUFA?

- Research topic: Next Generation Sequencing (NGS) methods for transcriptome analysis (RNA-seq) TRUFA and the long tail research
- TRUFA is an open platform offering a web-based interface enabling de novo RNA-seq analysis and comparative transcriptomics
- Currently TRUFA uses for HPC a top500 Supercomputer (Altamira)

http://trufa.ifca.es

Published in 2015 (in Evolutionary Bioinformatics)

TRUFA: A User-Friendly Web Server for *de novo* RNA-seq Analysis Using Cluster Computing

Libertas Academica
FREEDOM TO RESEARCH

Etienne Kornobis[1], Luis Cabellos[2], Fernando Aguilar[2], Cristina Frías-López[3], Julio Rozas[3], Jesús Marco[2] and Rafael Zardoya[1]

[1]Departamento de biodiversidad y biología evolutiva, Museo Nacional de Ciencias Naturales MNCN (CSIC), Madrid, Spain. [2]Instituto de Física de Cantabria, IFCA (CSIC-UC), Edificio Juan Jordá, Santander, Spain. [3]Departament de Genètica and Institut de Recerca de la Biodiversitat (IRBio), Universitat de Barcelona, Barcelona, Spain.

ABSTRACT: Application of next-generation sequencing (NGS) methods for transcriptome analysis (RNA-seq) has become increasingly accessible in recent years and are of great interest to many biological disciplines including, eg, evolutionary biology, ecology, biomedicine, and computational biology. Although virtually any research group can now obtain RNA-seq data, only a few have the bioinformatics knowledge and computation facilities required for transcriptome analysis. Here, we present TRUFA (TRanscriptome User-Friendly Analysis), an open informatics platform offering a web-based interface that generates the outputs commonly used in *de novo* RNA-seq analysis and comparative transcriptomics. TRUFA provides a comprehensive service that allows performing dynamically raw read cleaning, transcript assembly, annotation, and expression quantification. Due to the computationally intensive nature of such analyses, TRUFA is highly parallelized and benefits from accessing high-performance computing resources. The complete TRUFA pipeline was validated using four previously published transcriptomic data sets. TRUFA's results for the example datasets showed globally similar results when comparing with the original studies, and performed particularly better when analyzing the green tea dataset. The platform permits analyzing RNA-seq data in a fast, robust, and user-friendly manner. Accounts on TRUFA are provided freely upon request at https://trufa.ifca.es.

KEYWORDS: transcriptomics, RNA-seq, *de novo* assembly, read cleaning, annotation, expression quantification

CITATION: Kornobis et al. TRUFA: A User-Friendly Web Server for *de novo* RNA-seq Analysis Using Cluster Computing. *Evolutionary Bioinformatics* 2015:11 97–104 doi: 10.4137/EBO.S23873.

CORRESPONDENCE: ekornobis@gmail.com

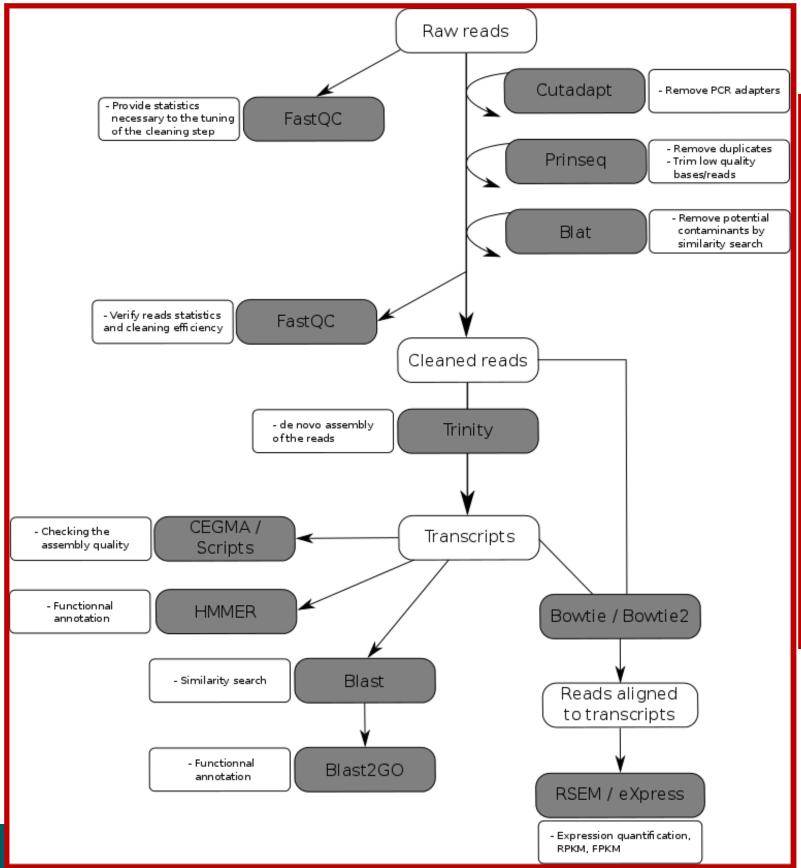
COPYRIGHT: © the authors, publisher and licensee Libertas Academica Limited. This is an open-access article distributed under the terms of the Creative Commons CC-BY-NC

UC UNIVERSIDAD DE CANTABRIA

CSIC CONSEJO SUPERIOR DE INVESTIGACIONES CIENTÍFICAS

Why TRUFA is useful?

- **Enables the "execution" of a simple but very useful pipeline**
- **Reserves the required HPC resources**



Software	Number of CPUs	Type
B2G	16	identify
BLASTP	16	assembly/mapping
BLAT	64	cleaning, identify
BOWTIE	16	assembly/mapping
CEGMA	16	assembly/mapping
CUFFLINKS	16	expression
CUTADAPT	16	cleaning
FASTQC	2	cleaning
HMMER	96	identify
INTERPROSCAN	64	identify
MPIBLAST	96	identify
PRINSEQ	16	cleaning
RSEM	1	expression
SAMTOOLS	1	assembly/mapping
TRINITY_RNA_SEQ	16	assembly/mapping

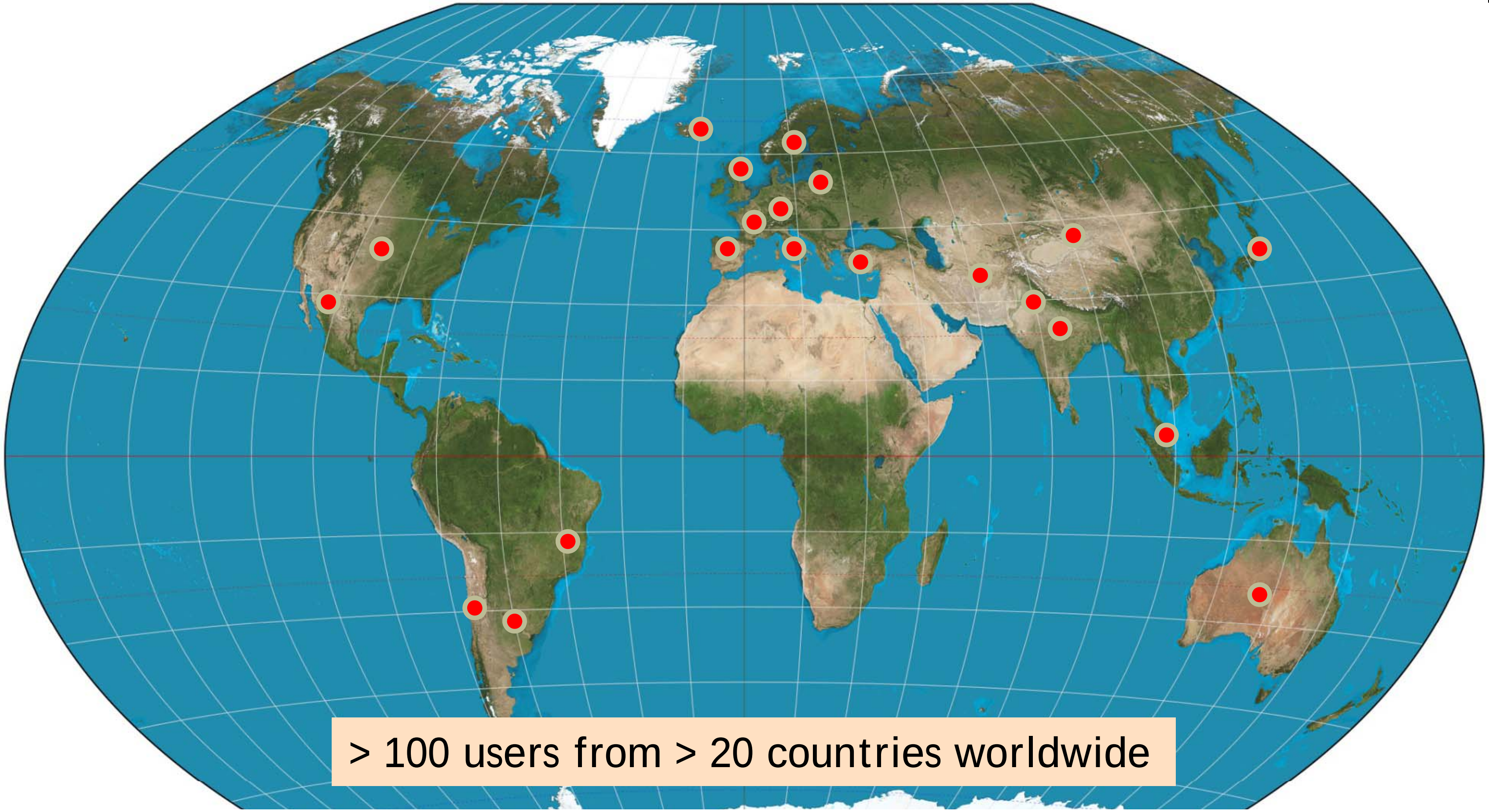
TRUFA and the long tail of science

- **When deploying TRUFA, the initial suggestion was**
  - **OPEN to any user worldwide**
  - **Even ANONYMOUS ACCESS (cf. reviewers of the paper)**
  - **FOR FREE**
- **Can we accept this "MODEL"?**
  - **FOR FREE: not completely, let's say NO ADDITIONAL COST**
  - **OPEN: ok (as long as there is COMPENSATION: citations in papers/projects)**
  - **ANONYMOUS: NO**
    - **security reasons, even if applications are constrained, storage is involved**
- **So when the paper was published, open/free access was granted to users that registered**
  - **But nothing happened...**
  - **... until a reference to TRUFA was published in a community BLOG**

UNIVERSIDAD DE CANTABRIA

CSIC

TRUFA registered users



> 100 users from > 20 countries worldwide

Spain	USA	India	Brazil	China	Australia	Chile	France	Iran	Iceland	Mexico
23	12	10	6	6	5	4	4	4	3	3

Poland	U K	Germany	Italy	Turkey	Argentina	Japan	Malaysia	Pakistan	Sweden
3	3	2	2	2	1	1	1	1	1

Impact on supercomputer usage

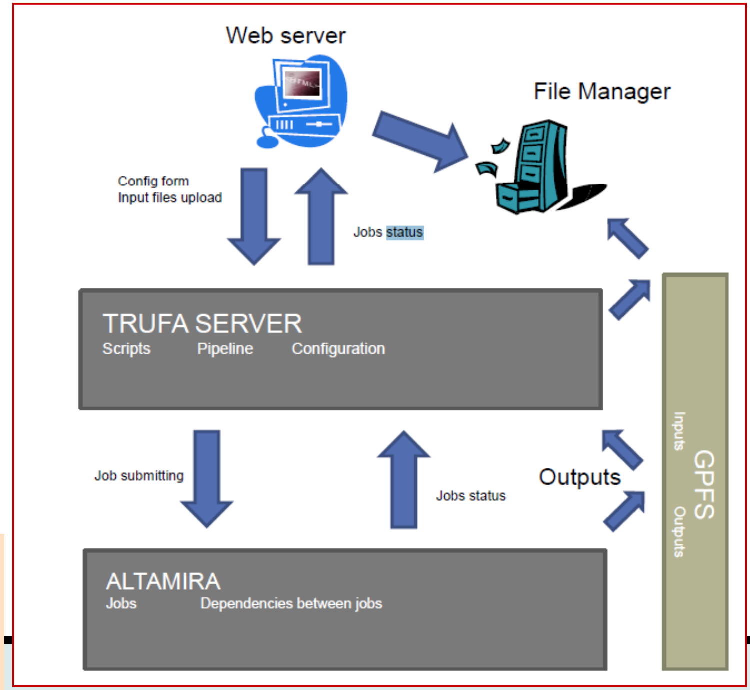
- **ALTAMIRA supercomputer uses SLURM as batch system**
  - **Open scheduler**
  - **Manages efficiently the requests for large number of cores**
    - **ALTAMIRA has >2500 cores linked by IB FDR, 512 cores jobs are usual**
  - **However it is difficult to get a global efficiency > 80%**
    - **Wall Time versus Time used in executions**
    - **Notice that ALTAMIRA uses a quite powerful GPFS over IB**
  - **Group Priorities are KEY to get access to the system**
    - **TRUFA users have minimum priority in the system**
- **TRUFA registered community has consumed more than 1.4 million hours in less than 6 months**
- **Despite this fact, the system has continued to provide adequate resources to all groups...**

Impact on supercomputer usage

- **Why?**
  - **We don’t have yet enough statistics**
  - **Notice the range of granularity of the processes in the pipeline**
    - **Different number of cores**
    - **Different execution time**
    - **Different dependencies**
  - **In particular, dependencies are managed by SLURM**
    - **not “inside” the job**
    - **i.e. a job requesting initially FASTQC (1 core) can start inmediatly even if there are not resources available for TRINITY, HMMER…**
- **Evolution**
  - **7K h in May, 100K h in June, >300K h/month July-October**
  - **November? Sudden drop up to now! WHY???**

Linking to the Grid/Cloud world

- TRUFA requires up to 96 cores for HPC
  - Practical limit is given by scalability ratio tests preformed (ALTAMIRA has extremely efficient FDR IB)
- 64 cores HPC-like resources are available in Grid/Cloud
  - Few IB enabled clusters plus MPI
  - 4 processor servers (Xeon E7 v3 4x12x2 (HT) = 96 cores)
- So, an obvious solution could be the "migration" of the existing scheme, to the Cloud
- BUT there are two approaches:
  - Deploy a cluster, install SLURM, reuse everything
  - Use Cloud solutions: OpenShift



see F.Aguilar talk on Thursday, in Community Clouds session @15h35

Linking to the Grid/Cloud world

- **The other way round from the Grid/Cloud:**

exploit Supercomputers available "backfilling" time

- **Needs a dedicated service to bridge towards the batch/queue system**
- **Would benefit of a common storage layer**
  - **GPFS is shared at IFCA site: ALTAMIRA, GRID, CLOUD (via NFS)**
- **This is becoming a very substantial resource**
  - **Race for the Exaflop: larger and larger systems**
  - **Supercomputers as very expensive resources per hour of processor for non HPC loads is no longer an argument**
  - **Traditional "isolation" is a barrier to be broken!**
- **Ongoing Project at University of Cantabria to use ALTAMIRA as opportunistic resource directly accessible from the Grid/Cloud world.**

UC UNIVERSIDAD DE CANTABRIA

CSIC CONSEJO SUPERIOR DE INVESTIGACIONES CIENTÍFICAS

Conclusions

- The long tail of research is really long…
- We believe in supporting Open Science ALSO offering e-infrastructure resources (third pillar of Open Science)
- We are impressed by the interest in TRUFA

  …but would like to see the citations!
- When the contact with the community is lost, it is not easy to know what is going on!
  - New OPEN/FREE resources?
  - Negative experience? Long waiting time? Storage?
- Is there a chance to connect the opportunistic use of supercomputers and an HPC Cloud-based vision?

A win-to-win strategy!