PaiRT: Accuarate Identification Pipeline helping for Real Time PCR

Doan Vu Thinh

Faculty of Information Technology, Nha Trang University

*thinhdv@ntu.edu.vn*

**TÓM TẮT**— Bài báo này trình bày những cách thức cần phải chuẩn bị cho khuôn dạng của một bài báo để được in ấn trong Kỷ yếu Hội nghị quốc gia về Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin (Fundamental and Applied Information Technology Research - FAIR). File được dùng như một khuôn mẫn (template) của Micrsosoft Word 6.0 hoặc mới hơn. Toàn bộ phần Tóm tắt được viết với font chữ Times New Roman, 9 pt. italic và bold cho cụm từ Tóm tắt

**Từ khóa**— Bao gồm khoảng 4 từ khóa hoặc cụm từ để qua đó người đọc có thể biết được hướng nghiên cứu của bài báo này. Các từ khóa hoặc cụm từ cách nhau bởi dấu phẩy.

# INTRODUCTION

Next-generation sequencing (NGS) has transformed the field of genetics into genomics by providing DNA sequence data at an ever increasing rate and reduced cost (Mardis, 2008). Three platforms for massively parallel DNA sequencing read production are in reasonably widespread use at present: the Roche/454 (http://www.454.com), the Illumina/Solexa (http://www.illumina.com), and SOLiDTM System (<http://marketing.appliedbiosystems.com>).

The Burkholderia genus contains over 60 species and occupies a large range of environments including soil, plants, rhizospheres, water, animals and humans (Ginther JL, 2015). The identification of novel species in new locations necessitates the need to identify the true global distribution of Burkholderia species, especially the members that are closely related to B. pseudomallei. Burkholderia ubonensis - whole genome shotgun sequence can be found at GenBank by ID ABBE01000101.1 (<https://www.ncbi.nlm.nih.gov/nuccore/ABBE01000101.1>)

Over 500 thousands reads for each forward and reverse sequence files, need to have pipeline to process huge data and to take full advantage of high performance PC. Here, the pipeline PairT is introduced as a tool for accurate identification helping for Real Time PCR genomic data; PairT is a wrapper script designed to take raw data, taking full advantage of both paired-end reads. As input, PairT takes paired FASTQ files for individuals and outputs show the percentage of reads which overlap through all pairs. The pipeline and a comprehensive online manual can be found at <https://github.com/thinhdoanvu/PairT>.

# METHODS

## Implementation and basic usage

#### Implementation and basic usage

The PairT pipeline is written in BASH and will run using most Unix-like operating systems through largely dependent on other bioinformatics software packages. Proper implementation depends on the correct installation of each third-party packages/tools. A full list of dependencies can be found at ([https://github.com/thinhdoanvu/PairT/requires.sh](https://github.com/thinhdoanvu/PairT)) and a sample script to automatically download and install the packages in a Linux environment can be found at the PairT repository (<https://github.com/thinhdoanvu/PairT>).

PairT is run by simply switching to a directory containing input data and starting the program. The program can be run with or no need configuration file, and PairT will proceed through a short series of command line prompts, allowing the user to establish analysis parameters.

#### Data input requirements

PairT requires both compressed forward and reverse FASTQ files for each individual in the analysis. A simple naming convention (a single-word locality code/name and a single-word sample identifier separated by an underscore) must be followed for every sample; examples are LOCAL\_IND01.F.fastq.gz and LOCAL\_IND01.R.fastq.gz. A sample script for naming convention, can be found at the github repository (https://github.com/thinhdoanvu/PairT/rename4PairT.sh).

#### Quality trimming

When working with high-throughput sequencing data, the quality of raw reads will need to assess using the Phred Quality Score by FASTQC tool (<http://www.bioinformatics.babraham.ac.uk/projects/fastqc>). PairT can use the program fastp (Chen, 2018) to simultaneously remove Illumina adapter sequences and trim ends of reads of low quality. By default, fastp trims bases with quality scores less than PHRED 10 (corresponding to a 10% chance of error in the base call).

#### Read maping

dDocent uses the MEM algorithm (Li, 2013) of BWA (Li & Durbin, 2009; Li & Durbin, 2010) to map quality-trimmed reads to the reference contigs. Users can deploy the default values of BWA or set an alternative value for each mapping parameter (match score, mismatch score, and gap-opening penalty). The default settings are meant for mapping reads to the human genome, so users are encouraged to experiment with mapping parameters. BWA output is ported to SAMtools (Li et al., 2009), saving disk space, and alignments are saved to the disk as binary alignment/Map (BAM). BAM files are then sorted and indexed.

## Kích thước khác

### Phần tiêu đề

Tên bài báo dùng chữ in hoa như Template file này (Font chữ Arial 14pt, in đậm)

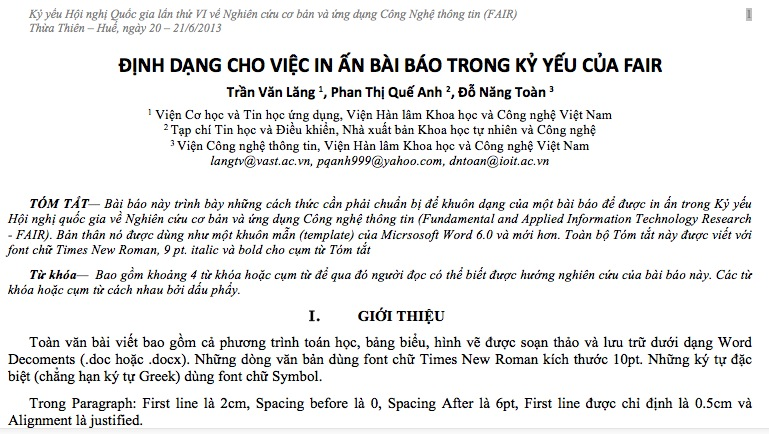
### Tác giả

Tên tác giả bao gồm cả cơ quan, địa chỉ email như Template file này

# THÔNG TIN KHÁC

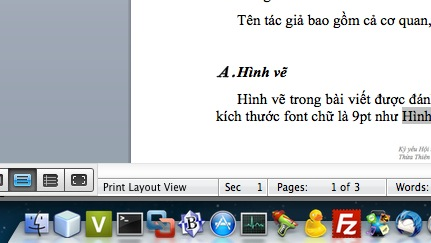
## Hình vẽ

Hình vẽ trong bài viết được đánh số bắt đầu từ 1, được canh lề Justified, lời chú thích được viết dưới hình vẽ với kích thước font chữ là 9pt như Hình 1.



1. Đây là hình về bài báo FAIR

Các hình vẽ tiếp theo được đánh số như Hình 2.



1. Đây là một hình khác

## Bảng biểu

Bảng biểu cũng tương tự như hình vẽ; tuy nhiên dòng chú thích được viết ở phía trên như Bảng 1.

1. Bảng biểu hướng dẫn

|  |  |
| --- | --- |
| Font chữ toàn văn | Times New Roman |
| Kích thước toàn văn | 10pt |
| Kích thước tiêu đề | 14pt |

# LỜI CẢM ƠN (NẾU CÓ)

# TÀI LIỆU THAM KHẢO

1. Mardis ER. 2008. Next-generation DNA sequencing methods. Annual Review of Genomics and Human Genetics 9:387–402 DOI 10.1146/annurev.genom.9.081307.164359.
2. Price EP, Sarovich DS, Webb JR, et al. Accurate and rapid identification of the Burkholderia pseudomallei near-neighbour, Burkholderia ubonensis, using real-time PCR. PLoS One. 2013;8(8):e71647. Published 2013 Aug 13. doi:10.1371/journal.pone.0071647
3. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018 Sep 1;34(17):i884-i890. doi: 10.1093/bioinformatics/bty560. Epub 2018 Sep 8. PubMed PMID: 30423086; PubMed Central PMCID: PMC6129281.
4. Author1\_Name, Author2\_Name, “Exploring the bounds of Web latency reduction from caching and prefetching”, In Proceeding(s) of the AICIT Symposium(Conference) on Internet Technologies, pp.13-22, 2019.

FORMAT FOR PREPARATION OF PAPER FOR PUBLICATION IN THE PROCEEDING’S FAIR

Tran Van Lang, Phan Thi Que Anh, Do Nang Toan

**ABSTRACT**— Inthis paper, we describe the submission guidelines for preparing papers for the Proceeding of national Conference on Fundamental and Applied Infromation Technology Research (FAIR). Use this document as a template with Microsoft Word 6.0 or later. Define all symbols used in the abstract. Do not cite references in the abstract. The abstract body is typed in Times New Roman, 9 pt. italic and in bold face.