BMC Bioinformatics

POSTER PRESENTATION Open Access

Using large public data repositories to discover novel genetic mutations with prospective links to melanoma

Tamas S Gal[*], Sally R Ellingson, Chi Wang, Jinpeng Liu, Stuart G Jarrett, John A D'Orazio

From 14th Annual UT-KBRIN Bioinformatics Summit 2015
Buchanan, TN, USA. 20-22 March 2015

Background

Next generation sequencing (NGS) data analysis pipelines are frequently described in literature. NGS data is relatively easy to acquire from national data repositories and most software used in the pipelines are open source. This study extends research on the causal relation between changes in the ataxia telangiectasia and Rad3 related (ATR) pathways and melanoma [1].

Materials and methods

To study the effects of mutations in the ATR region on melanoma, we downloaded the Melanoma Genome Sequencing Project dataset (dbGaP Study Accession: phs000452.v1.p1) from the dbGaP repository [2]. The dataset contained full exome sequencing data of paired normal and tumor samples of 122 phenotyped subjects in the format of trimmed and aligned BAM files. The dataset also included basic demographic information, such as gender and age; as well as disease specific variables, such as the localization of the melanoma and stage. The total size of the dataset was over 4TB, so we only downloaded the region of interest (ATR gene region) with 50Kbp padding before and after the ATR gene region. We used an available pipeline (Figure 1) for analysis that was previously developed for a lung cancer project by the Biostatistics and Bioinformatics Shared Resource Facility of the University of Kentucky Markey Cancer Center. The details of the data analysis pipeline will be published elsewhere. We used Python to automate data submission to the pipeline in combination with a configuration file that allowed us to easily swap different versions of the tools used in the pipeline and to match normal and tumor samples for the same patient (Figures 2, 3). Our experiments were executed on the Lipscomb High Performance Computing Cluster at the University of Kentucky.

Results

Though analysis and validation of the results are still ongoing at the time of this report, we can share that we identified five previously unreported somatic missense or splice site SNP mutations in the ATR gene region in melanoma patients. Results will be further validated by analysis of NGS data from melanoma cell lines.

Conclusions

The main goal of this abstract was to describe a methodology that we used to identify novel genetic markers in publicly available data. This methodology offers a cost effective way to test hypotheses drawn from laboratory research on human genome data.

[*] Correspondence: tamas.gal@uky.edu
Markey Cancer Center, University of Kentucky, Lexington, KY 40536, USA

Acknowledgements
This research was supported by the Cancer Research Informatics and the Biostatistics and Bioinformatics Shared Resource Facilities of the University of Kentucky Markey Cancer Center (P30CA177558).
We would like to thank the University of Kentucky Information Technology department and Center for Computational Sciences for computing time on the Lipscomb High Performance Computing Cluster and for access to other supercomputing resources.

Published: 23 October 2015

References

1. Jarrett SG, Horrell EM, Christian PA, Vanover JC, Boulanger MC, Zou Y, D'Orazio JA: **PKA-mediated phosphorylation of ATR promotes recruitment of XPA to UV-induced DNA damage.** *Mol Cell* 2014, **54(6)**:999-1011.
2. Mailman MD, Feolo M, Jin Y, Kimura M, Tryka K, Bagoutdinov R, Hao L, Kiang A, Paschall J, Phan L, Popova N, Pretel S, Ziyabari L, Lee M, Shao Y, Wang ZY, Sirotkin K, Ward M, Kholodov M, Zbicz K, Beck J, Kimelman M,

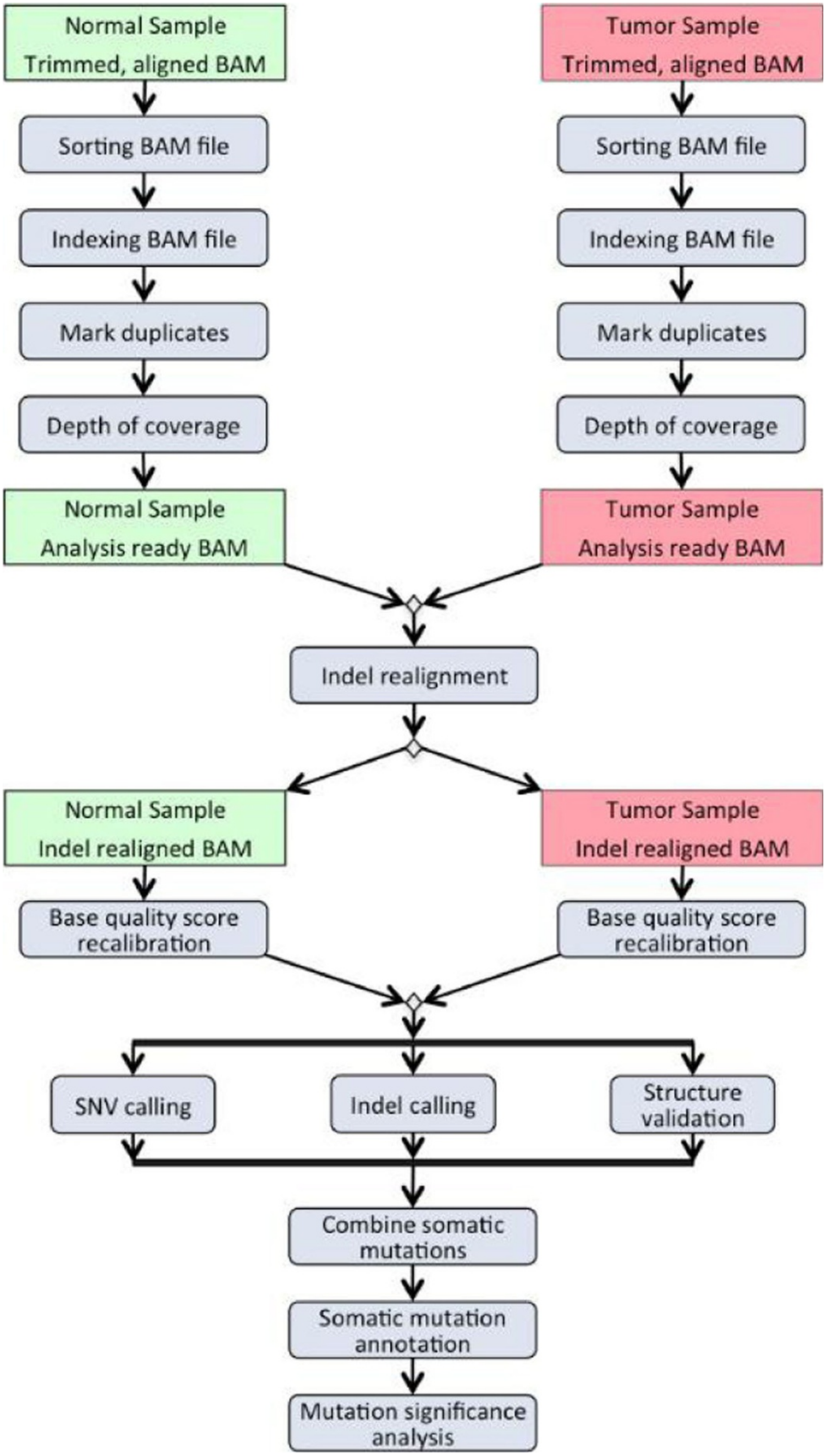
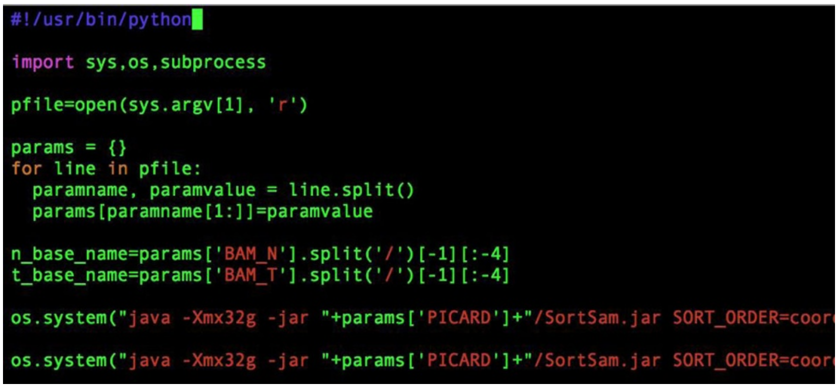
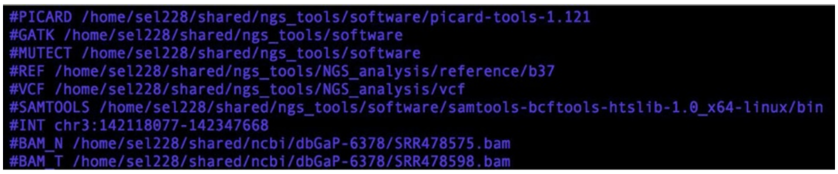


Figure 1 Pipeline for dbGaP analysis.



```
#!/usr/bin/python

import sys,os,subprocess

pfile=open(sys.argv[1], 'r')

params = {}
for line in pfile:
  paramname, paramvalue = line.split()
  params[paramname[1:]]=paramvalue

n_base_name=params['BAM_N'].split('/')[-1][:-4]
t_base_name=params['BAM_T'].split('/')[-1][:-4]

os.system("java -Xmx32g -jar "+params['PICARD']+"/SortSam.jar SORT_ORDER=coor

os.system("java -Xmx32g -jar "+params['PICARD']+"/SortSam.jar SORT_ORDER=coor
```

Figure 2 Python code for automated submission.



```
#PICARD /home/sel228/shared/ngs_tools/software/picard-tools-1.121
#GATK /home/sel228/shared/ngs_tools/software
#MUTECT /home/sel228/shared/ngs_tools/software
#REF /home/sel228/shared/ngs_tools/NGS_analysis/reference/b37
#VCF /home/sel228/shared/ngs_tools/NGS_analysis/vcf
#SAMTOOLS /home/sel228/shared/ngs_tools/software/samtools-bcftools-htslib-1.0_x64-linux/bin
#INT chr3:142118077-142347668
#BAM_N /home/sel228/shared/ncbi/dbGaP-6378/SRR478575.bam
#BAM_T /home/sel228/shared/ncbi/dbGaP-6378/SRR478598.bam
```

Figure 3 Configuration file.

Shevelev S, Preuss D, Yaschenko E, Graeff A, Ostell J, Sherry ST: **The NCBI dbGaP database of genotypes and phenotypes.** *Nat Genet* 2007, **39(10)**:1181-6.

doi:10.1186/1471-2105-16-S15-P3
Cite this article as: Gal *et al.*: Using large public data repositories to discover novel genetic mutations with prospective links to melanoma. *BMC Bioinformatics* 2015 **16**(Suppl 15):P3.