pdxBlacklist

Max Salm

Version 1.0: 10 March, 2017

Table of Contents

# pdxBlacklist

Patient-derived xenograft models are extremely valuable to immunology and oncology1. However, cross-contamination between host and graft tissue is inevitable2. Accordingly, sequencing data generated from these samples likely represent a mixture of mouse and human sequences.

*pdxBlacklist* catalogues genomic variation artefacts that arise by aligning mouse-derived short-read sequencing data to the human genome. To achieve this, mouse strain-specific short-reads generated by the [Mouse Genome Project](http://www.sanger.ac.uk/science/data/mouse-genomes-project)3 are aligned to the human genome, before identifying likely artefacts (e.g. SNVs and indels). The resulting variation catalogue enables annotation of existing variation datasets using tools such as *Vcfanno*4.

The *pdxBlacklist* pipeline is based on the bcbio-nextgen [pipeline](http://bcbio-nextgen.readthedocs.io/en/latest/), which affords great analytical flexibility in the choice of genome builds, aligners and variant callers. Currently, the defaults are set to:

* Genome: hg19
* Aligner: BWA
* Caller: VarDict5

To run the *pdxBlacklist* pipeline, exectute the following:

python pdxBlacklist.py --strain NOD\_ShiLtJ

where --strain specifies the mouse strain of interest.

## Features

* A choice of mouse strains (currently 40)
* Diverse variant analysis pipelines

## Installation

The scripts in this project have been developed and tested on an 8-core server with 30GB RAM running Ubuntu 14.04.5 LTS (GNU/Linux 3.19.0-74-generic x86\_64).

To install the *pdxBlacklist* pipeline, please run the following:

scripts/0\_setup\_bcbio.sh

## Contribute

* Issue Tracker: <https://github.com/MaxSalm/pdxBlacklist/issues>
* Source Code: <https://github.com/MaxSalm/pdxBlacklist>

## Support

If you are having issues, please let us know.

## License

The project is licensed under the GPL license.

## Citation

Paper in press

## References

1. Townsend, E. C. *et al.* The public repository of xenografts enables discovery and randomized phase iI-like trials in mice. *Cancer cell* **29,** 574–586 (2016).

2. Conway, T. *et al.* Xenome–a tool for classifying reads from xenograft samples. *Bioinformatics (Oxford, England)* **28,** i172–i178 (2012).

3. Keane, T. M. *et al.* Mouse genomic variation and its effect on phenotypes and gene regulation. *Nature* **477,** 289–294 (2011).

4. Pedersen, B. S., Layer, R. M. & Quinlan, A. R. Vcfanno: Fast, flexible annotation of genetic variants. *Genome biology* **17,** 118 (2016).

5. Lai, Z. *et al.* VarDict: A novel and versatile variant caller for next-generation sequencing in cancer research. *Nucleic acids research* **44,** e108 (2016).