Bioinformatics for translational benefit

“visualisation and web based tools bring sequence data to the clinic”

NHMRC Emerging Researcher Grant - Pitch

Dr Miles Benton

2017/07/28

Motivation

identification of known or novel pathogenic variants

- found among a host of common and rare polymorphisms
- identification of clinically relevant variants is time consuming
    - large potential for analysis induced false negatives
- larger datasets (WES/WGS) contain **thousands** to **millions** of variants
- aggravated by complex conditions/disorders
    - multitudes of genes / mutations responsible for symptoms or important for treatment

Motivation

identification of known or novel pathogenic variants

- found among a host of common and rare polymorphisms
- identification of clinically relevant variants is time consuming
  - large potential for analysis induced false negatives
- larger datasets (WES/WGS) contain **thousands** to **millions** of variants
- aggravated by complex conditions/disorders
  - multitudes of genes / mutations responsible for symptoms or important for treatment

current tools are limited

Motivation

identification of known or novel pathogenic variants

- found among a host of common and rare polymorphisms
- identification of clinically relevant variants is time consuming
    - large potential for analysis induced false negatives
- larger datasets (WES/WGS) contain **thousands** to **millions** of variants
- aggravated by complex conditions/disorders
    - multitudes of genes / mutations responsible for symptoms or important for treatment

current tools are limited

bringing data back to the end-users

Motivation

identification of known or novel pathogenic variants

- found among a host of common and rare polymorphisms
- identification of clinically relevant variants is time consuming
    - large potential for analysis induced false negatives
- larger datasets (WES/WGS) contain **thousands** to **millions** of variants
- aggravated by complex conditions/disorders
    - multitudes of genes / mutations responsible for symptoms or important for treatment

current tools are limited

bringing data back to the end-users

visualisation and interactivity incredibly powerful tools

Motivation

identification of known or novel pathogenic variants

- found among a host of common and rare polymorphisms
- identification of clinically relevant variants is time consuming
  - large potential for analysis induced false negatives
- larger datasets (WES/WGS) contain **thousands** to **millions** of variants
- aggravated by complex conditions/disorders
  - multitudes of genes / mutations responsible for symptoms or important for treatment

current tools are limited

bringing data back to the end-users

visualisation and interactivity incredibly powerful tools

I'm an emerging researcher (3 years out from PhD)

Major drivers for change

paywalls

- nearly all software suites are licensed == '$$$'

Major drivers for change

paywalls

- nearly all software suites are licensed == '$$$'

'blackbox' software

- mysterious to users
- difficult/impossible to modified (closed-source)
    - those not closed source are tailored to specific lab/job (i.e. CPIPE)
- can be hard to use

Major drivers for change

paywalls

- nearly all software suites are licensed == '$$$'

'blackbox' software

- mysterious to users
- difficult/impossible to modified (closed-source)
  - those not closed source are tailored to specific lab/job (i.e. CPIPE)
- can be hard to use

NO software readily accessible and user-friendly for end users 'on the ground'

Significance

current methods fall into either:

pay to use

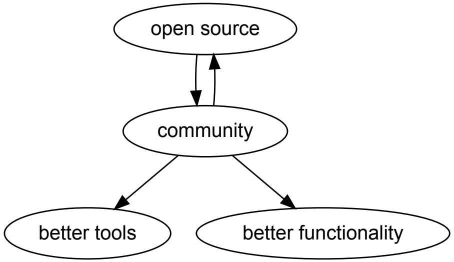
free but tricky

Significance

current methods fall into either:

pay to use **free but tricky**

Accessible, flexible, powerful, and ... free



Aims

address shortfall in current methods

- free and simple

Aims

address shortfall in current methods

- free and simple

Shiny integration

- the power of R/RStudio
- visualisation/interactivity front and center

Aims

address shortfall in current methods

- free and simple

Shiny integration

- the power of R/RStudio
- visualisation/interactivity front and center

ease of use

- want it to work in the hands of all end users
  - clinic
  - research
  - public?

Aims

address shortfall in current methods

- free and simple

Shiny integration

- the power of R/RStudio
- visualisation/interactivity front and center

ease of use

- want it to work in the hands of all end users
  - clinic
  - research
  - public?

modularity (full extensible and upgradable)

Aims

address shortfall in current methods

- free and simple

Shiny integration

- the power of R/RStudio
- visualisation/interactivity front and center

ease of use

- want it to work in the hands of all end users
    - clinic
    - research
    - public?

modularity (full extensible and upgradable)

cloud integration and deployment options

Team members on this submission

Dr Miles Benton - Principal Investigator (emerging investigator)[1]

Prof Lyn Griffiths[1]	AI (*Human Genetics*)
A/Prof Rod Lea[1]	AI (*Genome Informatics*)
Dr Robert Smith[1]	AI (*Diagnostics*)
Prof Greg Gibson[2]	AI/Mentor (*Integrative Genomics*)

[1] CDA, MM, IHBI, QUT
[2] Georgia Tech, USA

VCF files are powerful but clumsy*

[*] *if you are not familiar with the commandline*

To 95% of people these are just really large complex text files

- no easy interaction
- no ready visualisation

VCF files are powerful but clumsy*

[*] *if you are not familiar with the commandline*

To 95% of people these are just really large complex text files

- no easy interaction
- no ready visualisation

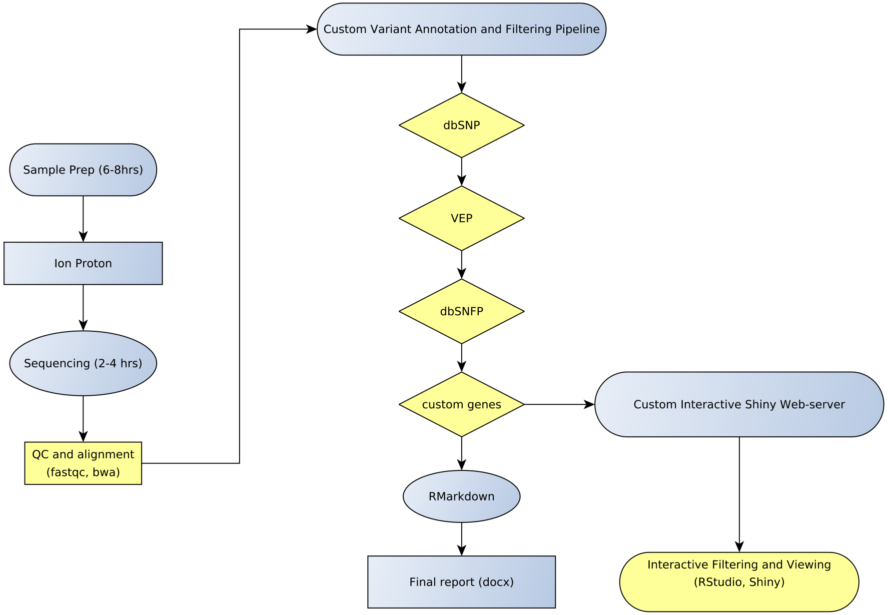
variant filtering process is complicated**

[**] *even if you **are** familiar with the commandline*

A large number of people still do this manually...
...this is what computers are for!

Progress to date

WESTARC - live demonstration


Custom Variant Annotation and Filtering Pipeline
dbSNP
VEP
dbSNFP
custom genes
Custom Interactive Shiny Web-server
RMarkdown
Final report (docx)
Interactive Filtering and Viewing (RStudio, Shiny)
Sample Prep (6-8hrs)
Ion Proton
Sequencing (2-4 hrs)
QC and alignment (fastqc, bwa)

Research Design

Develop a series of modules, each achieving a specific task:

- initial QC and sequence alignment

  + including functionality for structural variation (routinely overlooked)

- VCF annotation and manipulation

  + currently only accessible to 'advanced' users

- simple interactive 'base' fontend (i.e. WESTARC)

  + include database interfacing

- Additional analysis interface

  + phenotype, case/control, sample comparison ...

Research Design

Develop a series of modules, each achieving a specific task:

- initial QC and sequence alignment

  + including functionality for structural variation (routinely overlooked)

- VCF annotation and manipulation

  + currently only accessible to 'advanced' users

- simple interactive 'base' fontend (i.e. WESTARC)

  + include database interfacing

- Additional analysis interface

  + phenotype, case/control, sample comparison ...

Provide to clinical end users:

- have access to several engaged clinical geneticists (testing group?)

Research Design

Develop a series of modules, each achieving a specific task:

- initial QC and sequence alignment
  + including functionality for structural variation (routinely overlooked)
- VCF annotation and manipulation
  + currently only accessible to 'advanced' users
- simple interactive 'base' fontend (i.e. WESTARC)
  + include database interfacing
- Additional analysis interface
  + phenotype, case/control, sample comparison ...

Provide to clinical end users:

- have access to several engaged clinical geneticists (testing group?)

Distribute:

- GitHub, docker, & online cloud server (Amazon S3)

Expected outcomes

> *simple, scalable and robust method for the annotating and categorising genetic variants enabling more rapid and effective analysis of potentially pathogenic variants*

Expected outcomes

> *simple, scalable and robust method for the annotating and categorising genetic variants enabling more rapid and effective analysis of potentially pathogenic variants*

A fully operational suite of software 'modules':

- integrate into an easy-to-use workflow
- can handle all forms of sequence data
- open-source / free to use and develop

Deployment of an user friendly app version / suite of apps:

- integration with existing databases
- cloud deployment
- docker integration

Direct to consumer:

- putting the 'power' back in the hands of those that matter