Rare Variant Analysis Pipeline

https://github.com/vforget/rare-variant-pipeline

Nov 15, 2012

Objective

Use multiple rare variant association (RVA) tests to analyze a set of targets within a parallel computing environment

RVA tests: SKAT, RR, VT

Targets: exons, introns, whole genes, regulatory regions, conserved regions, and any combination thereof.

PE: Grid Engine

Overview

- GRINUX
- Grid Engine
- Rare variant analysis
- Pipeline
- Future work
- Discussion

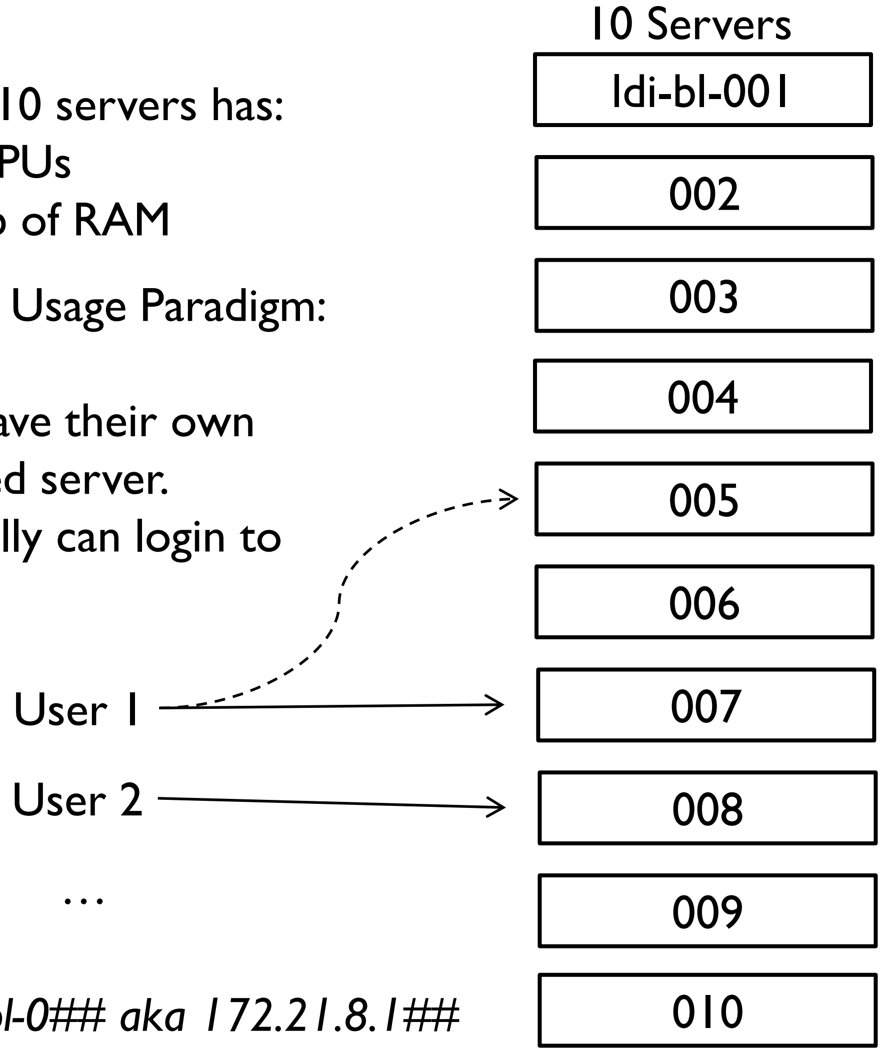
GRINUX

Each of 10 servers has:

- 24 CPUs
- 64Gb of RAM

Current Usage Paradigm:

Users have their own dedicated server. Optionally can login to others



ldi-bl-0## aka 172.21.8.1##

Limitations:

- 24 CPUs per user, or
- manual starting of jobs across multiple machines
- what happens when we pass 10 users? …
- What sharing policy to employ?

Parallel Computing w/ Grid Engine

- A software layer that schedules and executes programs (jobs) across multiple servers (nodes). Resources (cores, memory) are automatically allocated.

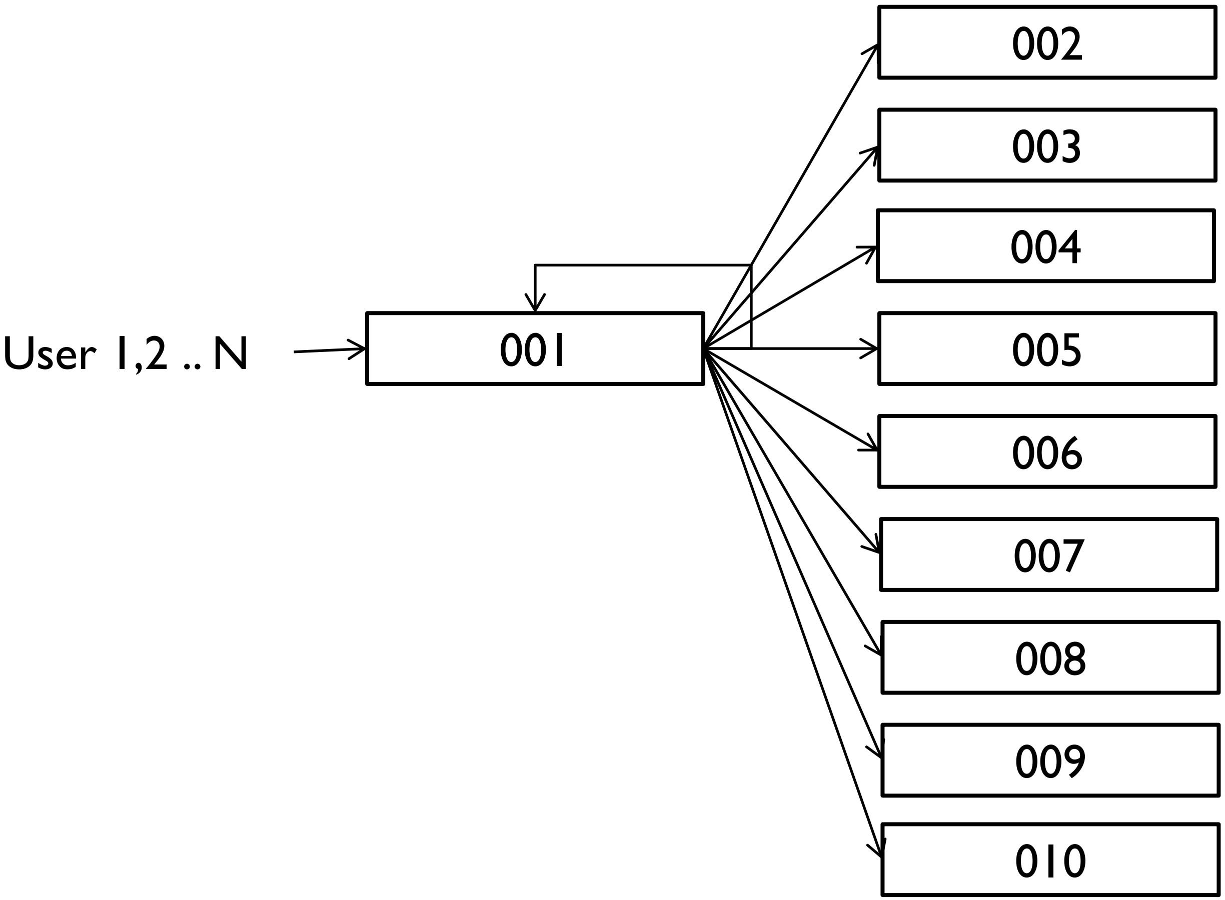
  - qsub: the program that submits jobs.
    - Ways to submit a job with qsub:
      - □ echo “command” | qsub [easy, not powerful]
      - □ qsub script.bash [requires more work, more powerful]
      - □ qsub –t 1-100 script.bash [launch numerically ordered jobs]

  - qdel: delete jobs

  - qstat: job status (still running?, on what server?)

  - qrsh: shell prompt
    - request cores interactive use
    - Dedicated resource similar to current setup

GRINUX ON Grid Engine



User has access to

- 240 cores
- Automatic queuing of jobs (>10k jobs).
- Share policy is built-in w/ Grid Engine
- Shell access via qrsh can simulate current "dedicated" setup

Limitations:

- With queuing no dedicated resources (share policy*)
- Small learning curve (but worth it ☺)

Rare Variant Analysis (RVA)

- SNPs are collapsed (or grouped) within a target (e.g., gene).

- A target may consist of multiple separate regions, e.g, exons of a gene.

- Generalization – RVA is conducted using SNPs from a set of regions. This group of SNPs is identified by a unique target name.
  - Example: SNPs from all coding exons (regions) of WNT16 (identifier).
  - Example: the entire gene space + regulatory regions of WNT 16.

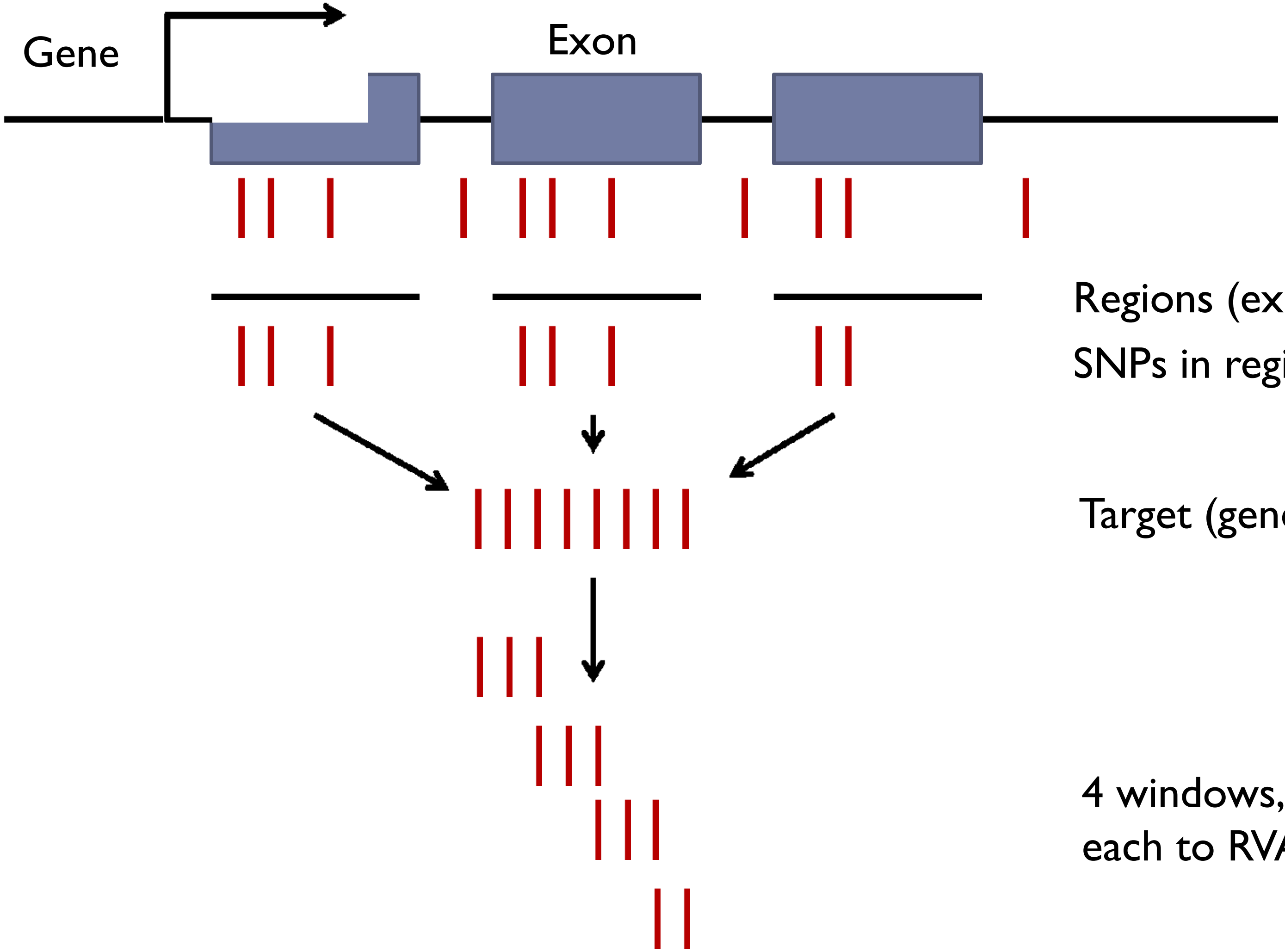
RVA (cont'd)

- SNP count will vary across different targets due to:
  - number and size of regions per target,
  - SNP density at locus,
  - etc.

- To normalize SNP count across different targets we split targets into windows containing an equal number of SNPS, with optional and user-defined window overlap.

- Generalize: Windows are just new smaller targets built from larger targets.
  - E.g. WNT16 target split into 3 sub-targets (windows) would be identified as WNT16.1, WNT16.2, WNT16.3.

Windows Visualized



Regions (exons)

SNPs in regions

Target (gene) >> RVA

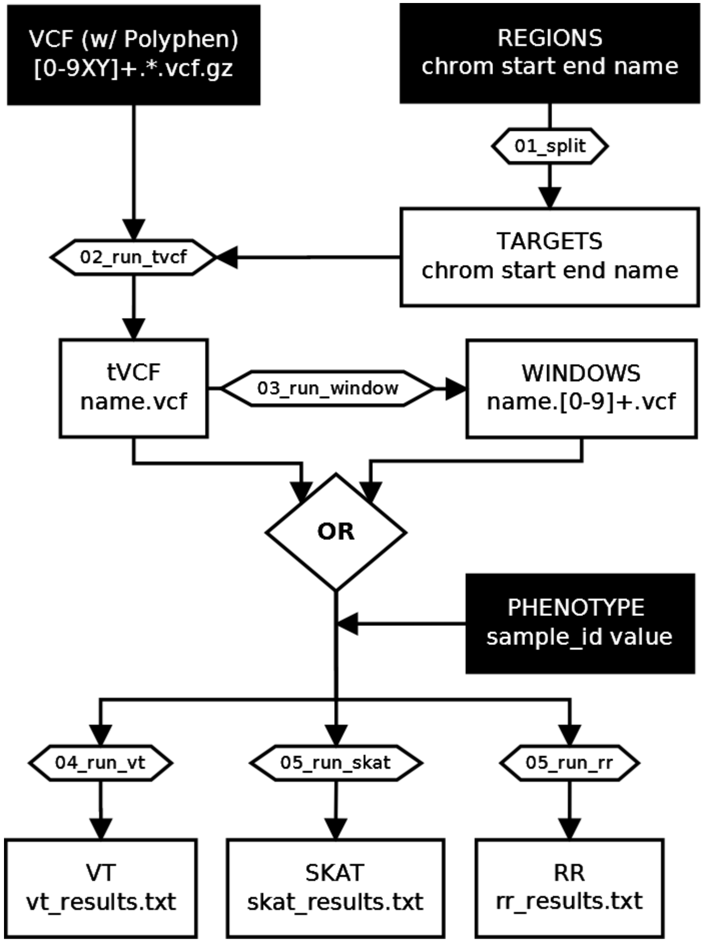
4 windows,
each to RVA

RVA (cont'd)

- Summary:
  - SNPs are collected from multiple regions per target.
  - Targets can be further split into windows into smaller targets.

** Targets are fed into RVA (SKAT, RR, VT) **

Pipeline



Input:

Chromosome VCF files.
Regions: genomic coordinates
Phenotype file.

Grid Engine is used for:

Fetching SNPs from targets (list of regions)

Splitting SNPs into windows.

Running rare variant tests (SKAT, VT, RR).

Pipeline Steps (Step 0)

- Input files:
  - VCF files, one per chromosome (polyphen scores optional)
  - Region file:

Target

Regions

```
10 100003847 100004653 C10orf28
10 100007442 100008748 LOXL4
10 100010821 100010933 LOXL4
10 100011322 100011459 LOXL4
10 100012109 100012225 LOXL4
10 100013309 100013553 LOXL4
10 100015333 100015496 LOXL4
10 100016536 100016704 LOXL4
```

  - Phenotype file:

```
UK10K_124350     -2.769557
UK10K_88736      -2.529971
QTL210350        -2.521555
QTL218819        -2.39639
QTL211631        -2.383679
```

Pipeline (Step 1)

- Split region file into targets:

```
$ mkdir ~/my_project && cd ~/my_project
$ mkdir targets/
$  cd targets/
$  01_split.py exome_ranges.txt
```

- This will create one file for each target name (e.g. gene name). Each file is named <target_name>.txt, e.g., WNT16.txt will contain all exon coordinate for that gene.

Pipeline (Step 2)

- Fetch SNPs by target:

```
$ 02_tvcf.bash targets.txt \
        targets/
        tvcf/
        ~/share/UK10K_COHORT/REL-2011-12-01/v4
        0.01
```

- This step fetches the SNPs within the set of regions for each target. Results for each target are saved to a file names <target_name>.vcf, e.g., WNT16.vcf.

Pipeline (Step 3, optional)

- Split targets into windows:

```
$ 03_window_tvcf.bash targets.txt \ # target list
      tvcf/output/ \ # whole targets
      windows/ \ # output directory
      20 \ # window size
      15 \ # window step
      10   # min window size
$ ls windows/output | perl -p -e "s/\.vcf//g;" >
     windows.txt
```

- This step further splits each target VCF file into overlapping windows of SNPs. Results of each window are saved to a file named <target_name>.<window_num>.vcf, e.g., the first window for WNT16 will be named WNT16.1.vcf.

Pipeline (Step 4)

- Run RVA (e.g. SKAT):

```
$ 05_skat.bash windows.txt
        windows/output/
        skat.window/
        pheno/pheno_FA_uk10k.txt
        TRUE
        C
```

- This step will convert the VCF files to a format compatible for SKAT and perform the rare variant analysis.
- Results from SKAT for each target or window are saved to files named either WNT16.skline or WNT16.1.skline, respectively.
- These are then merged into one file named skat_results.txt.

Output

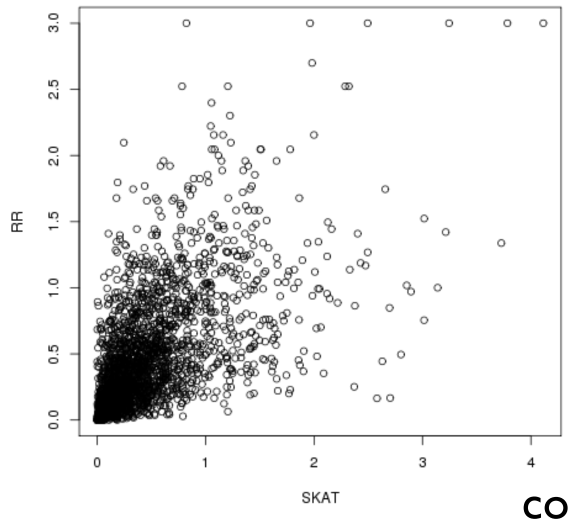
So far, output is p-values from SKAT, VT and RR.

SKAT Output:

```
TARGET,NMARKER,NMARKER.TEST,P.VALUE,PVALUE.LIU,PVALUE.RESAMP
A1BG,15,9,0.326773278859148,0.326315495949641,0.342657342657343
A1CF,16,9,0.150561608661568,0.154691372295823,0.141858141858142
A2BP1,60,35,0.665793975893963,0.646130589787688,0.67032967032967
A2LD1,8,5,0.0566838959172191,0.0579290181811024,0.0589410589410589
A2ML1,64,41,0.288998114631238,0.284245649833552,0.291708291708292
A2M,71,47,0.119405845320115,0.122793520008053,0.110889110889111
A4GALT,21,14,0.122589290322937,0.124604582536437,0.136863136863137
A4GNT,17,13,0.892642862724579,0.892244306075528,0.888111888111888
AAA1,11,7,0.902766800639803,0.959459030763467,0.888111888111888
```

Compare p-values for SKAT w/ RR
** preliminary results **

- Run RVA for all 20,846 human genes (no window for now), using FA_adj_std phenotype from UK10K_exomes
- Get results back for 19,252 genes (1580 have no SNPs)
  - ** 13 remaining genes need to be investigated **
- Comparing p-values for SKAT and RR:



Genes with –log10(pv) > 2 in both SKAT and RR:
TBC1D10C: neg. feedback inhibitor calcineurin path
TLE2: Enhancer of split groucho-like protein 2
TRIM47: tripartite motif-containing 47
TTC33: Osmosis-responsive factor
VPS4A: vacuolar protein sorting
ZNF479: zinc finger protein 479

cor = 0.6154802

Future Work

- Support IMPUTE2 format
  - Fetch Polyphen scores from most recent database.
- Wrap entire pipeline in a single script:
  - Execute each step as a "task".
  - Save parameters to a file (logging, ease of use).
- Improve logging of errors
- Provide summary statistics of run (failures, etc).

Discussion