hbvdtools

Human Background Variation Database Tools

Jessada Thutkawkorapin
Daniel Nilsson

input

- variant files (vcf)
- tags (e.g. brief disease indication or perhaps platform, kit)
- PRIVACY CAVEAT: not to few individuals

output

- variant frequency table
  - (ANNOVAR avdb format by default)

data storage

- frequency table, tags

dependencies

- vcftools – Vcf.pm library interface

local git + github: hbvdb

Artistic license or GPL.

Perl implementation.

bvd-add

```
bvd-add.pl -T HNPCC cocaclingen.merged.vcf
bvd-add.pl -T 200danes 200danes.merged.vcf
```

bvd-get

```
bvd-get.pl -T HNPCC > my_background.avdb
```

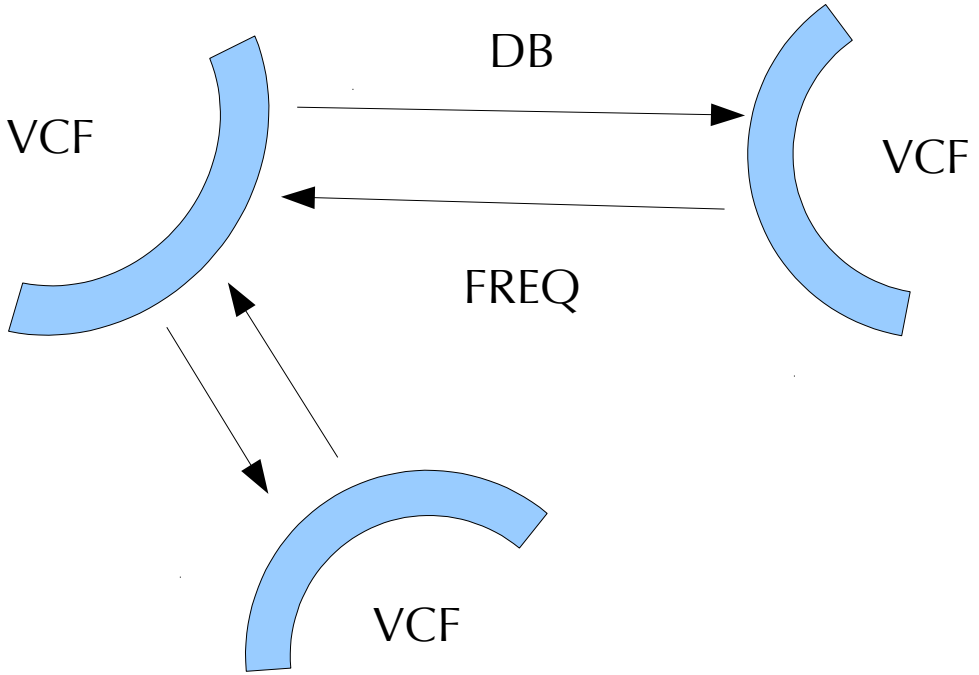
bvd-merge

merge two databases

```
bvd-merge.pl DB1/ DB2/
```

Did we both add that file?

Compares digests/checksums of all previous adds.



Assuming low intensity exchange.

Each of a few centers/larger groups adds own data
& exchanges dbs.

Total frequency tables provided for versioned download,
via standard http service,
with pre-computed filter options to exclude tags.

Outreach:
Clinical genetics, Karolinska -> SciLifeLab Sthlm
-> SciLifeLab U:a
& Clinical genetics, other (5 clinics, ~3 MPS+)

Performance rather IO-dependent.

Add VCF with 100 exomes ~ 20 mins.

Extract a frequency table with 6M variants ~ 334 exomes ~ 1 min.

Issues

Recognizing small indels – consistent nomenclature

`normalize-vcf.pl` catches a few more.

GATK LeftAlignVariants also.

```
1 GCCC GCC
2 CCC CC
2 CC C
2 C -
3 C -
4 C -
...
```

I want exact variants, not a region filter. Thoughts?