Human Background Variation DataBase - Standard Operating Procedure

The database does not store genotypes, only allele frequencies and tags. It is vital that no information that can be attributed to an identified single individual is retained. Obviously you are free to use the tools for such purposes yourself, for projects where you have legal reason to do so.

We primarily anticipate exome sequences for initial exchange, but the tools can handle other data.

STEP 1: Prepare **vcf-files** with variants.

- Ensure that no code-key mappable to a living or dead individual can be found anywhere in the files you intend to process. Inspect the filenames and vcf-headers once more to make sure these are **already de-identified**.
- Prepare cohorts of at least a few individuals if you wish to tag them for later optional exclusion. Please never tag cohorts of less than 4 individuals with a unique tag. Tagging ONE individual with a unique tag effectively renders that person identifiable.
- It is of importance to track the reference genome used for variant exchange. We currently use the 1000genomes reference hg19/GRCh37 build.
- **No strict quality cutoff** is given. Variants that are filtered to a level you would normally consider a candidate for calling a variant are appropriate; **perhaps PASS to Tranche99 for default GATK filters, or a samtools call quality of at least 20.** You do not need to manually inspect alignments or apply strict filters.
- **Normalize** to enhance small indel recall. We recommend using **GATK LeftAlign**.

STEP 2: Pick one or a small number of keywords (**tags**) describing your samples.

- Choose a tag that will allow you, or a like-minded researcher, to later exclude a group that might compromise filtering. We have so far used tags like HNPCC, HLH, bladder_exstrophy, and neutropenia.
- When you choose tags to identify disease (or healthy control) cohorts, be as **brief** and **general** as you feel is appropriate. The **tags are primarily intended for your own benefit,** and for researchers that work with a highly similar group of affected persons. So for example your osteoporosis cases can serve as a background population for myopia, but you can also use the database successfully for filtering even if another researcher has previously added osteoporosis cases.
- Giving **multiple tags with detailed information can compromise the deidentification**. E.g. giving both detailed geographic region and a disease type that occurs in 1/1'000'000 is not a good idea. It may be possible to identify the eight-fingered family from Grönköping without a code key.

STEP 3: Add your variants to your local database.

```
bvd-add.pl -T my_tag -d your_new_db_dir my_cohort.vcf
```

STEP 4: Share your variants.

Send a copy of your database directory to the other centers. If you send them to me (daniel.nilsson@scilifelab.se) I can include them in a common baseline db.

```
tar zcf collection_name_and_date.tgz your_new_db_dir
```

STEP 5: Obtain variants shared by others. Easily done when submitting your frequency data..

```
bvd-merge.pl -d common_db_dir large_common_database
```