Cortex cheat sheet for bacterial genomics

1 Requirements

Download Cortex from github (https://github.com/iqbal-lab/cortex) as follows:

```
git clone --recursive
                https://github.com/iqbal-lab/cortex.git
```

You must have installed VCFtools (and have the entire directory, not just the binary executable), R and Stampy. R must be in your path. Add the following directories to PERL5LIB and PATH

```
export PERL5LIB= /path/cortex/scripts/analyse_variants/
    bioinf-perl/lib;/path/cortex/scripts/calling:$PERL5LIB
export PATH = /path/cortex/scripts/analyse_variants/
   needleman_wunsch-0.3.0
```

2 Preparation (once per species)

There are a set of files Cortex needs to use, which you should not need to worry about, so we wrap it all up:

```
perl scripts/calling/prepare.pl --index INDEX
     --ref_fa species.fasta
     --dir_for_ref_objects /path/refdir/
     --vcftools_dir /path2/vcftools_1.0.9/
     --outdir /path3/results
     --stampy_bin /path4/stampy.py  --kmer 31
```

This script will create a config script in /path3/results containing information about your parameter choices, so you do not need to enter them again in subsequent scripts.

3 Compare samples against a reference; combine results; genotype

Step1 First we build per-sample graphs in parallel using commands such as this GNU parallels command. This will create a directory for each sample within directory /path3/results/. A typical command would be (for 1700 samples):

```
parallel --gnu -j 20  perl scripts/calling/par.pl
  --num {}  --index INDEX
  --out_dir /path3/results  ::: {1..1700}
```

Step2 Combine all the per-sample VCFs to get one combined set of sites (SNPs, indels, structural variants).

```
perl scripts/analyse_variants/combine/combine_vcfs.pl
 --prefix XYZ   --outdir  /path3/results
```

Step 3 Finally, independently genotype each sample at all of these sites. Here using GNU parallels to spread across 20 cores of a server. Memory use now drops as we only use the graph of poymorphisms.

```
cat /path3/results/combine/list_args_for_final_step | parallel
     perl scripts/calling/genotype_1sample_against_sites.pl
--config /path3/results/combine/config.txt
 --invcf /path3/results/combine/XYZ.sites_vcf
--sample {1} --outdir {2}   --sample_graph {3}
```

This will give you one VCF file per sample in /path3/results/{sample_id}/union_calls/

4 Segregating variants within our dataset (joint workflow)

First build sample graphs as Step1 in the previous example. Then UNDETERMINED

5 Pan-genome analysis

To detect presence of a set of predefined genes (genes.fasta) among your samples

To look at pan-genome graph of all samples and see which samples have which contigs, allowing you to stratify them by frequency or look for differentiating/segregating contigs.

6 The End

For further information: