Cortex for bacterial genomics: quickstart

1 Requirements

Download Cortex from github (https://github.com/iqbal-lab/cortex) - there are download instructions there.

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:
    /path/VCFTools_dir/perl:$PERL5LIB
export PATH = /path/cortex/scripts/analyse_variants/
   needleman_wunsch-0.3.0
```

Also, make an INDEX file, mapping sample-identifiers to sequence data.

2 Independent workflow: motivation

Do variant discovery independently for each sample (against a reference), then combine the callsets to make a single set of candidate sites (SNPs, indels, SVs), and then genotype all samples.

This new pipeline will parallelise and scale to hundreds of thousands of bacteria. This cheatsheet shows how to run it on a single machine with many CPUs/cores, using GNU parallels.

3 Independent workflow: how to run it

This script will compile Cortex for you, make reference genome binaries, Stampy indexes - all the things that used to have to be done manually. It will also choose appropriate memory-use parameters

```
perl cortex/scripts/calling/run_indep_wkflow_with_gnu_par.pl
--index INDEX  --ref_fa ref_genome.fa
--dir_for_ref_objects ref/
--vcftools_dir ~/installed_apps/vcftools_0.1.9/
--outdir results/
--stampy_bin ~/installed_apps/stampy-1.0.23/stampy.py
--kmer 31 --procs 20 --prefix salmonella
```

This starts with FASTQ (or BAM) files, and finishes with a single VCF with all samples genotyped at the same sites.

4 Troubleshooting

5 Segregating variants within our dataset (joint workflow)

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

6 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.

7 The End

For further information: