CSC 334 Project: Principle Component Analysis in Asia

CSC 334

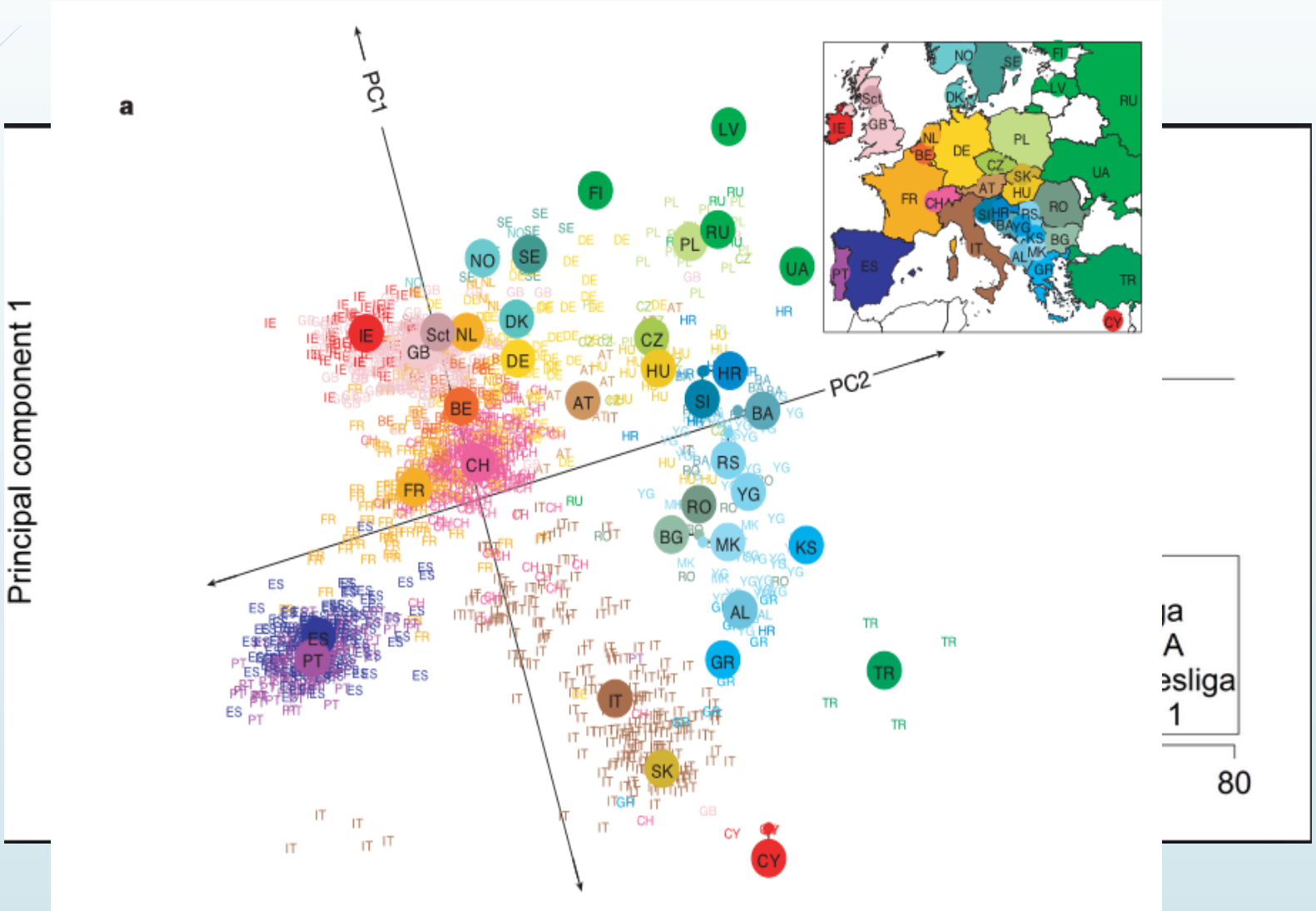
Hee Jin

Introduction

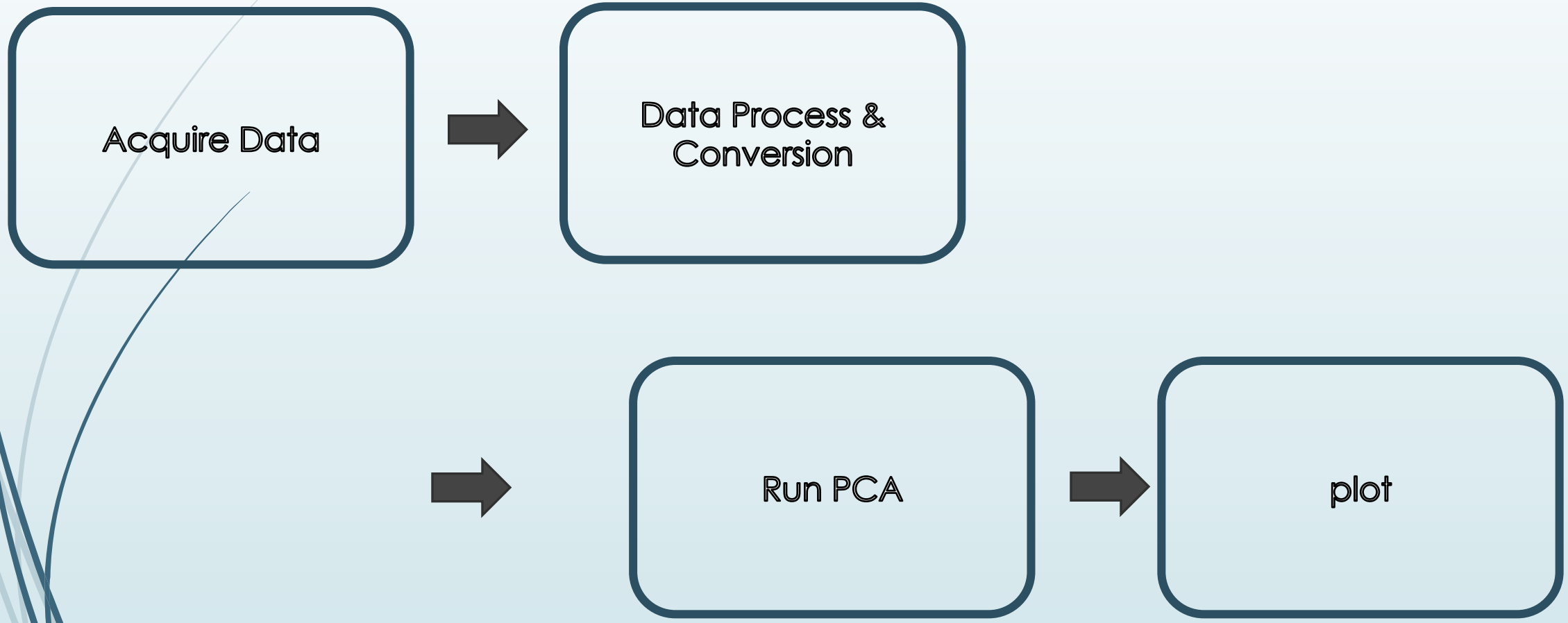
```
CSAKLECPQDWLSHRDKCFHVSQVSNTWEEGLVDCDGKGATLMLI-QDQEELRFLLDSIKEKYNSF
TTVNLECPQDWLLHRDKCFHVSQVSNTWEEGQADCGRKGATLLLI-QDQEELRFLLDSIKEKYNSF
CSVNLECPQDWLSHRDKCFRVFQVSNTWEEGQADCGRKGATLLLI-QDQEELRFLLDSIKEKYNSF
ARHCGHCPEEWITYSNSCYYIGKERRTWEESLLACTSKNSSLLSI-DNEEEMKFLSIISPSS----
DKVYWFC------YGMKCYYFVMDRKTWSGCKQACQSSSLS-LCL-KIDDEDELKFLQLVVPSDSC
VKVYWFC------YGMKCYYFVMDRKPWSRCKQSCQNSSLTLLKI-DDEDELKFLQLVVPSD--SC
FEKYWFC------YGIKCYYFNMDRKTWSGCKQTCQISSLSLLKI-DNEDELKFLQNLAPSD--IS
GVKHWFC------YGTKCFYFIMSKNTWSGCKQTCQHYSLPLVKI-EDEDELKFLQFQVISD--SY
GVKYWFC------YRTKCYYFIMNKNTWSGCKQNCQHYSLPLVKI-DDENELKFLQFQVIPD--SY
GVKYWFC------YGTKCYYFIMNKTTWSGCKANCQHYSVPIVKI-EDEDELKFLQRHVILE--SY
DSDCCSCQEKWVGYRCNCYFISSEQKTWNESRHLCASQKSSLLQI-QNTDELDFMSSSQQ----FY
ESYCGPCPKNWICYKNNCYQFFDESKNWYESQASCMSQNASLLKV-YSKEDQDLLKLVKS----YH
```

Principal Component Analysis

Introduction



Methods: General Pipeline



Methods: Details

Acquire Data

- 1000 genome project
- Focus on Han Population in East Asia

ALL.chr1.phase3_shapeit2_mvncall_integrated_v5a.20130502.genotypes.vcf.gz

ALL.chr1.phase3_shapeit2_mvncall_integrated_v5a.20130502.genotypes.vcf.gz.tbi

Methods: Details

Data Process & Conversion

- Tabix (filter data)

```
jhil914@ubuntu:~$ tabix -h ALL.chr1.phase3_shapeit2_mvnca
ll_integrated_v5a.20130502.genotypes.vcf.gz 1:1-50000 >my
File.vcf
```

- Vcftools / PLINK
- .vcf -> PLINK form

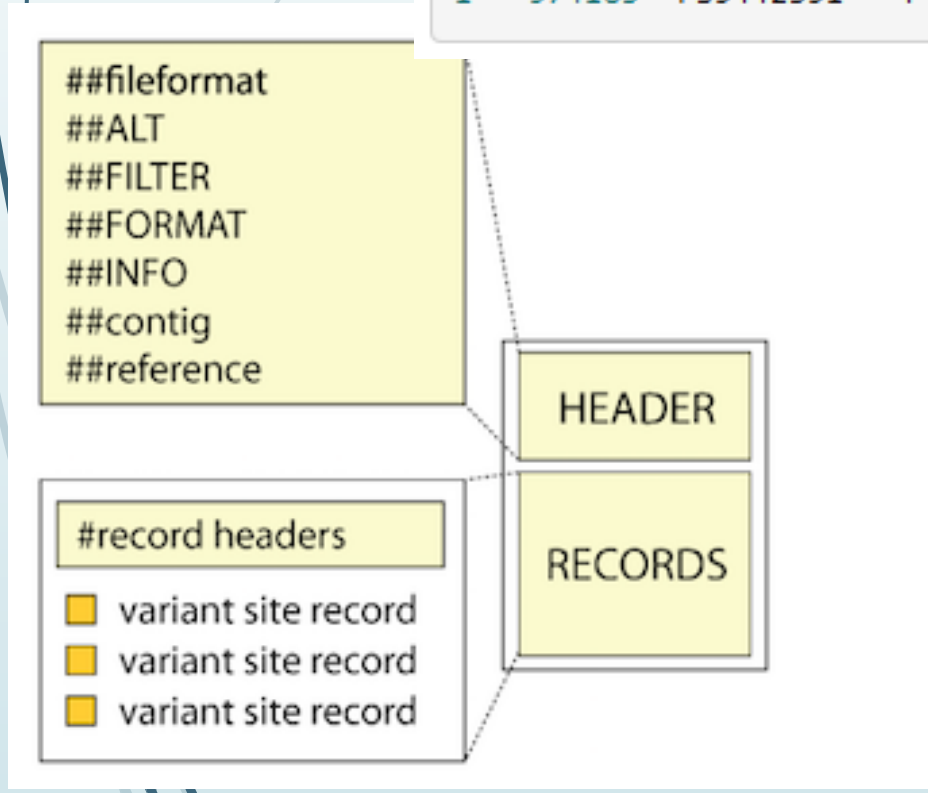
```
jhil914@ubuntu:~$ vcftools --vcf chr1.vcf -plink-tped -
chr 1 --out output_in_plink2
```

```
jhil914@ubuntu:~$ p-link --tfile output_in_plink --recode
```

- Output: .ped, .map

Methods: VCF structure

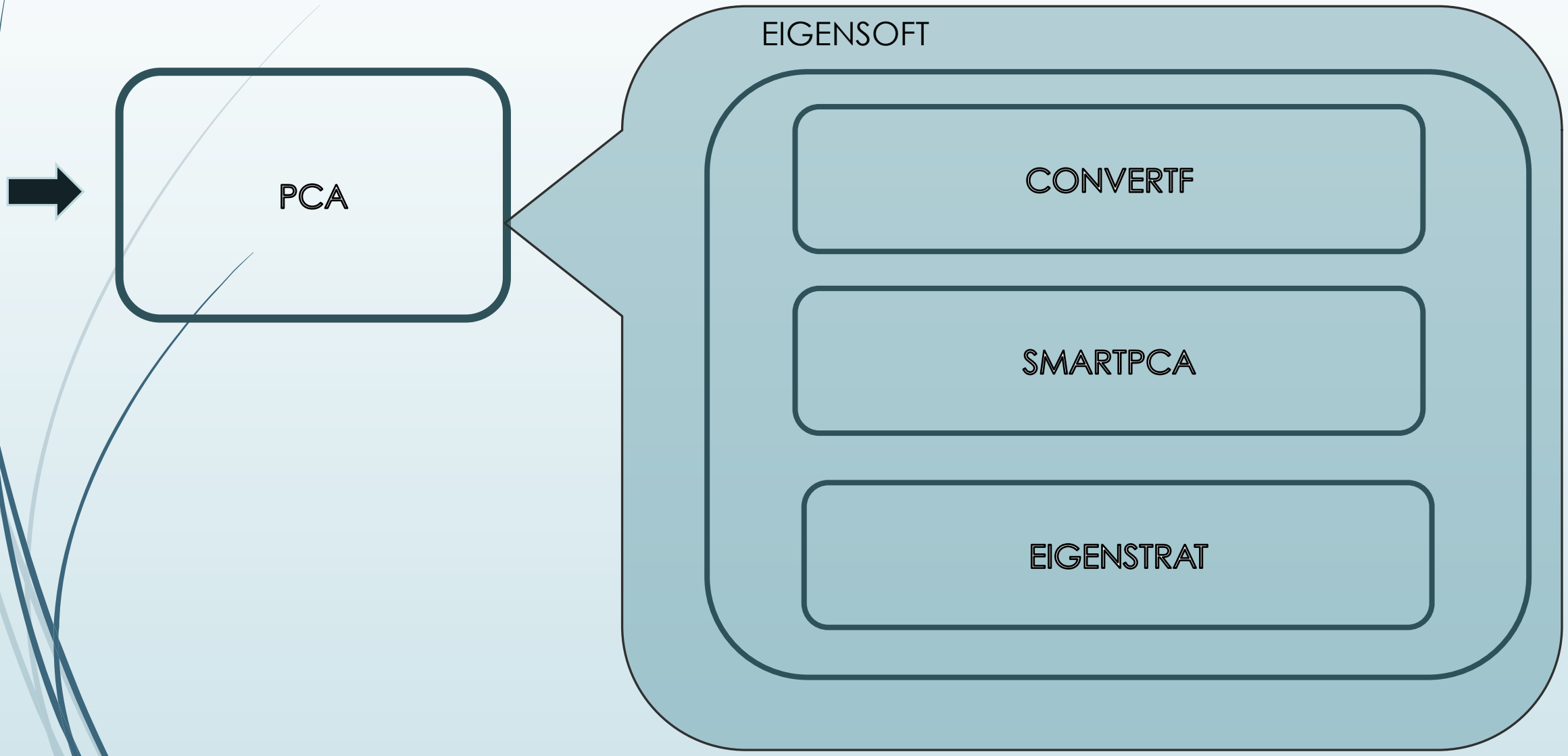
```
[HEADER LINES]
#CHROM  POS ID       REF ALT QUAL    FILTER  INFO          FORMAT          NA12878
1   873762  .        T   G   5231.78 PASS    [ANNOTATIONS] GT:AD:DP:GQ:PL  0/1:173,141:282:99:255,0,255
1   877664  rs3828047   A   G   3931.66 PASS    [ANNOTATIONS] GT:AD:DP:GQ:PL  1/1:0,105:94:99:255,255,0
1   899282  rs28548431  C   T   71.77   PASS    [ANNOTATIONS] GT:AD:DP:GQ:PL  0/1:1,3:4:26:103,0,26
1   974165  rs9442391   T   C   29.84   LowQual [ANNOTATIONS] GT:AD:DP:GQ:PL  0/1:14,4:14:61:61,0,255
```



Methods: PED structure

KHV_1	HG02023	0	0	0	0	A A	C C	G G	C C	C C	C C	G G	C C	C C	T T	T
T	G G	T T	A A	G G	G G	G G	G G	C C	T T	C C	C C	C C	C C	T T	G G	G
G	A A	T T	G G	A A	A A	G G	T T	A A	G G	C C	C C	C C	G G	G G	A G	G
G	C C	G G	C C	C C	G G	G T	G G	C C	G G	C C	G T	G G	G G	T T	G G	G
G	A A	A A	G G	C C	T T	T T	C C	G G	T T	T T	G G	G G	G G	C C	C C	G
G	C C	G G	G G	T T	T T	A A	C C	G G	C C	A A	G G	G G	C C	C C	G G	T
T	C C	G G	T T	A A	G G	G G	G G	G G	G G	C C	G G	G G	C C	G G	C C	C
C	T T	C C	T T	G G	C C	G G	T T	C C	T T	G G	A A	G G	G G	C C	A A	C
C	A A	G G	C C	T T	A A	T T	G G	A A	G G	C C	T T	A A				
KHV_2	HG01597	0	0	0	0	A A	C C	G G	C C	C C	C C	G G	C C	C C	T T	T
T	G G	T T	A A	G G	G G	G G	G G	C C	T T	C C	C C	C C	C C	T T	G G	G
G	A A	T T	G G	A A	A A	G G	T T	A A	G G	C C	C C	C C	G G	G G	A G	G
G	C C	G G	C C	C C	G G	T G	G G	C C	G G	C C	G T	G G	G G	T T	G G	G
G	A A	A A	G G	C C	T T	T T	C C	G G	T T	T T	G G	G G	G G	C C	C C	G
G	C C	G G	G G	T T	T T	A A	C C	G G	C C	A A	G G	G G	C C	C C	G G	T
T	C C	G G	T T	A A	G G	G G	G G	G G	G G	C C	G G	G G	C C	G G	C C	C
C	T T	C C	T T	G G	C C	G G	T T	C C	T T	G G	A A	G G	G G	C C	A A	C
C	A A	G G	C C	T T	A A	T T	G G	A A	G G	C C	T T	A A				

Methods: Details



Methods: Details

PCA

EIGENSOFT

CONVERTF

```
genotypename:    plink.ped
snpname:         plink.map # or example.map, either works
indivname:       plink.ped # or example.ped, either works
outputformat:    EIGENSTRAT
genotypeoutname: plink_output.eigenstratgeno
snpoutname:      plink_output.snp
indivoutname:    plink_output.ind
familynames:     YES
```

Methods: Details

PCA

EIGENSOFT

SMARTPCA

```
genotypename:     plink_output.eigenstratgeno
snpname:          plink_output.snp
indivname:        plink_output.ind
evecoutname:      example.evec
evaloutname:      example.eval
altnormstyle:     NO
numoutevec:       2
```

Methods: Details

PCA

EIGENSOFT

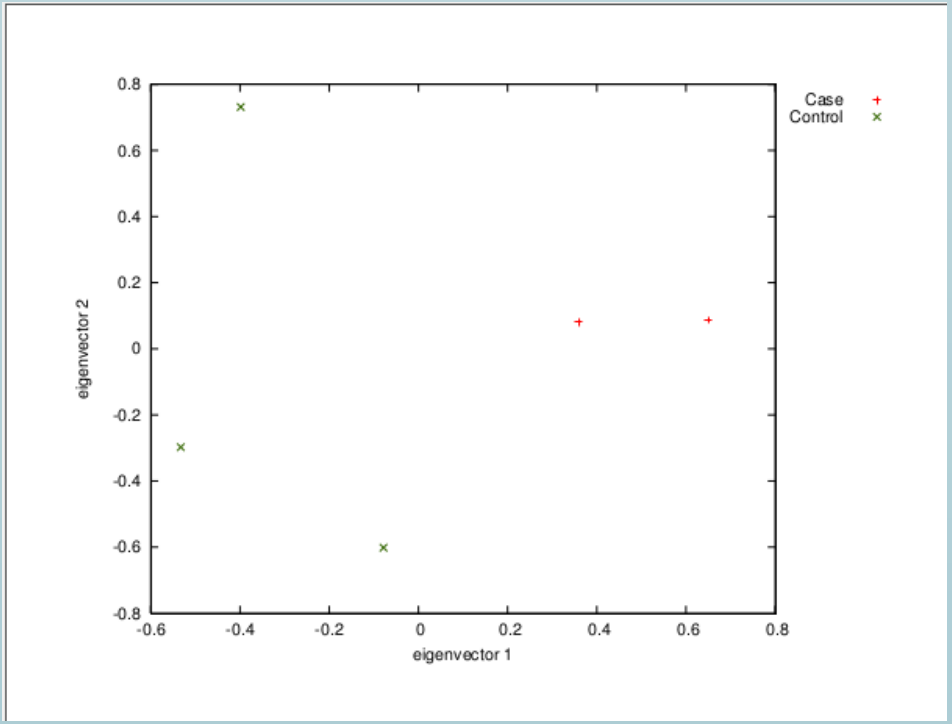
EIGENSTRAT

```
../bin/smarteigenstrat.perl
-i example.geno : genotype file in any format (see ../CONVERTF/README)
-a example.snp : snp file in any format (see ../CONVERTF/README)
-b example.ind : individual file in any format (see ../CONVERTF/README).
  We note that phenotype information will be contained in example.ind,
  either as Case/Control labels or quantitative phenotypes if -q set to YES.
-q YES/NO : If set to YES, use quantitative phenotypes in example.ind.
  If -q is set to YES, the third column of the input individual file
  in EIGENSTRAT format (or sixth column of input individual file in PED format)
  should be real numbers.  The value -100.0 signifies "missing data".
  If -q is set to NO, these values should be "Case" or "Control".
  The default value for the -q parameter is NO.
-p example.pca : input file of principal components (output of smartpca.perl)
-l l : (Default is 10) number of principal components along which to
  correct for stratification.  Note that l must be less than or equal to
  the number of principal components reported in the file example.pca.
-o example.chisq : chisq association statistics.  File contains log of
```

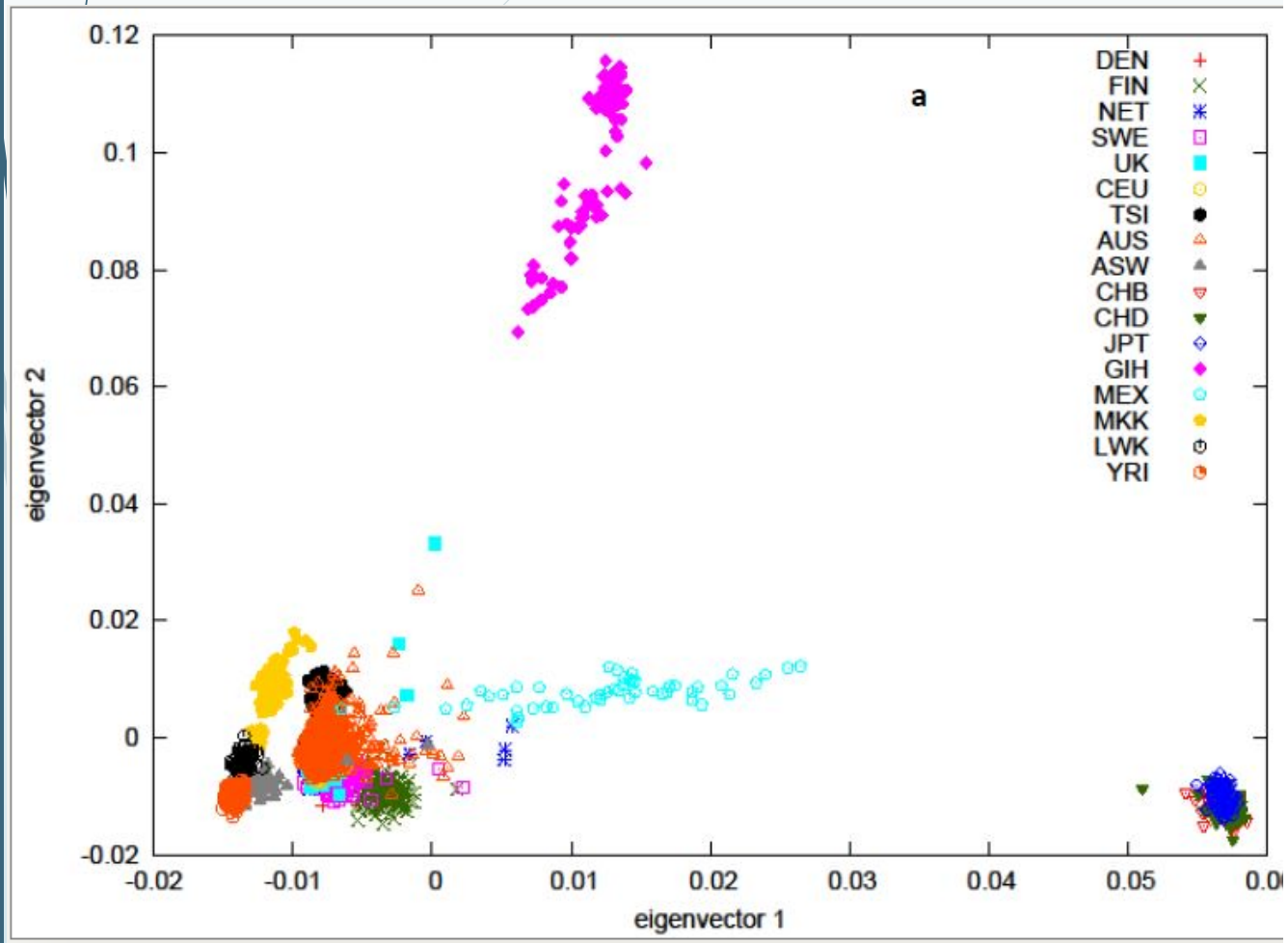
Methods: Details

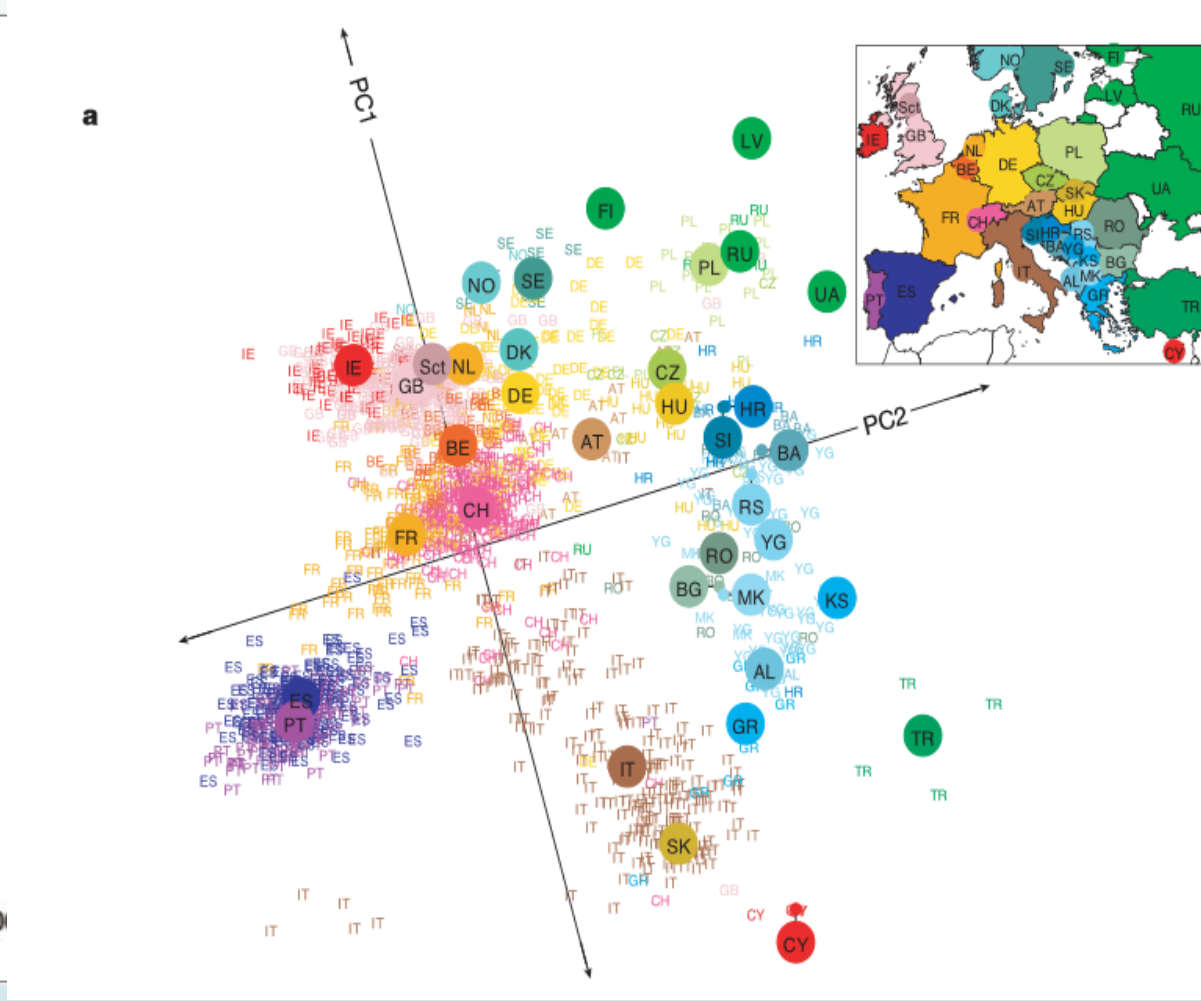
plot

- **Ploteig perl script**
  - Input: .evec
  - Output: plot graph



Expected Results





https://majorityrights.com/weblog/comments/some_initial_thoughts_on_yang_et_al