Tutorial on genetic ancestry estimation: **How to use LASER**

Chaolong Wang

Sequence Analysis Workshop

June 2014 @ University of Michigan

LASER: **L**ocating **A**ncestry from **SE**quence **R**eads

- **Main functions of the software:**
  - Place samples in an ancestry space using sequence reads
    - Works well for very low-coverage data
    - Easy for parallel computation
  - Pace samples in an ancestry space using genotype data
    - Follows the same framework as for sequence data
    - Fast and robust to family relatedness
  - Perform standard PCA on genotype data
    - Fast
    - Can handle very large data sets

- **Download:** http://www.sph.umich.edu/csg/chaolong/LASER/
  wget http://www.sph.umich.edu/csg/chaolong/LASER/LASER-2.01.tar.gz
- **Wiki:** http://genome.sph.umich.edu/wiki/LASER

Outline of this tutorial

- What can you find in the package?
- Input files
- How to run LASER?
- Output files and interpretation
- Some advanced options
- Practice

What can be found in the software package?

- Two main programs:

  laser: estimate ancestry using sequence reads

  trace: estimate ancestry using genotype data

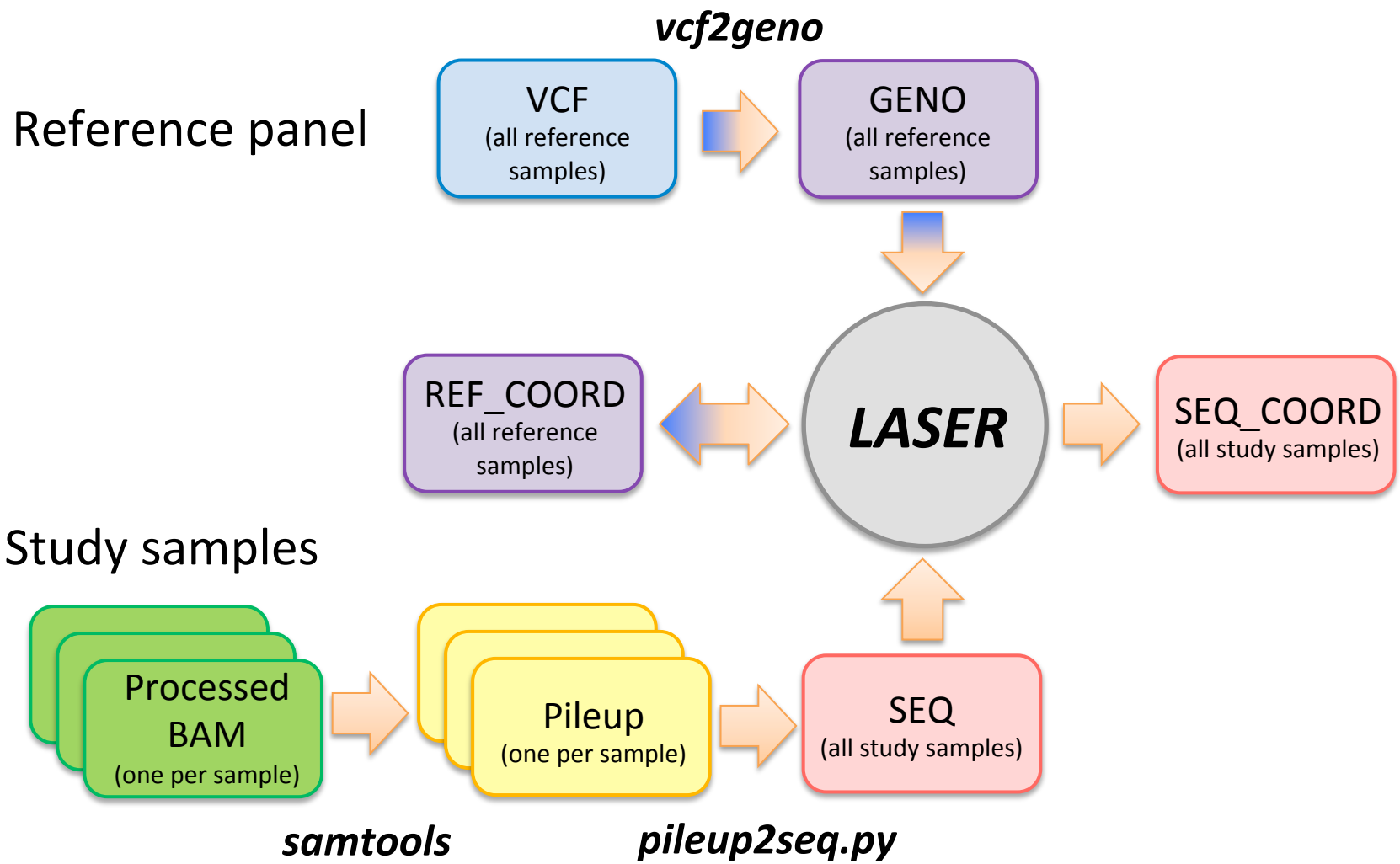
- Two tools for preparing input files:

  vcf2geno: convert VCF files to input genotype files

  pileup2seq: convert pileup files to input sequence files

- R codes for plotting results (in the "plot" folder)
- **User's manuals** (in the "doc" folder)
- Example data and commands
- Source codes (in the "src" folder)

Preparing input files for LASER



Reference genotype file and *vcf2geno*

Genotype data are provided in the geno file, with columns representing loci and rows representing samples.

Genotypes are coded as 0, 1, 2, representing counts of the reference alleles. Missing data are coded as -9.

Marker information are provided in the site file (one locus one line).

Important notes:

1. Reported on forward strand
2. Consistent genomic position (b37)
3. Tab-delimited

_.geno (tab-delimited)

POP_1	IND_1	2	0	1	...
POP_1	IND_2	2	-9	2	...
POP_2	IND_3	0	0	-9	...
POP_3	IND_4	1	2	1	...
...	...	...	...	...	...

_.site (tab-delimited, with a header)

CHR	POS	ID	REF	ALT
1	752566	rs3094315	G	A
1	768448	rs12562034	G	A
1	1005806	rs3934834	C	T
...	...	...	...	...

You can convert from vcf to geno format using our vcf2geno tool:

```
./vcf2geno --inVcf HGDP.vcf.gz --updateID newID.txt --out HGDP
```

The --updateID option to specify the popID and indivID in the output file.

Generate reference coordinate file (PCA)

Use LASER to perform PCA on the reference genotypes:

```
./laser -g ./example/HGDP_238_chr22.geno -k 20 -pca 1 -o HGDP
```

-g specifies the reference genotype file
-k specifies the number of PCs to compute
-pca turns on the PCA mode
-o specifies the prefix of the output files

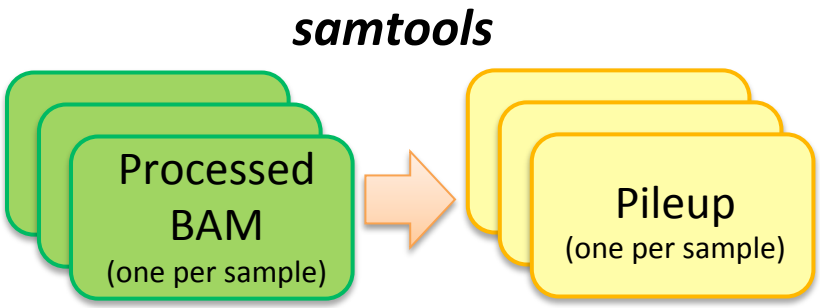
LASER will output the PC coordinates and proportion of variance explained by each PC, and has an option to output the SNP loadings (use -pca 3).

Computational speed:
LASER takes about 15 mins to perform PCA on the HGDP data set of 938 samples and 632,958 SNPs (on a 2.8GHz CPU).

_.RefPC.coord (tab-delimited)

popID	indivID	PC1	PC2	PC3	...
POP_1	IND_1	-3.5	0.2	0.7	...
POP_1	IND_2	-2.2	4.5	0.8	...
POP_2	IND_3	7.8	-0.8	-1.0	...
POP_3	IND_4	1.6	-3.8	-0.4	...
...	...	...	...	...	...

From BAM files to pileup files (*samtools*)



Format of a pileup file

CHR	POS	REF	DEPTH	BASES	QUAL
22	17094749	A	1	c	D
22	17202602	T	1	.	D
22	17411899	A	1	.	C
22	17450515	G	2	.,	9<
22	17452966	T	1	c	5
22	17470779	C	1	,	A
22	17492203	G	1	,	B
22	17504945	C	3	,..	BCA
22	17529814	T	3	..,	CCC

From BAM files to pileup files (*samtools*)

Step 1: make a BED file to specify a list of loci to extract from BAM files.

```
awk '{if (NR>1){print $1, $2-1, $2, $3;}}' HGDP.site > HGDP.bed
```

An example line in the BED file: `1 752565 752566 rs3094315`

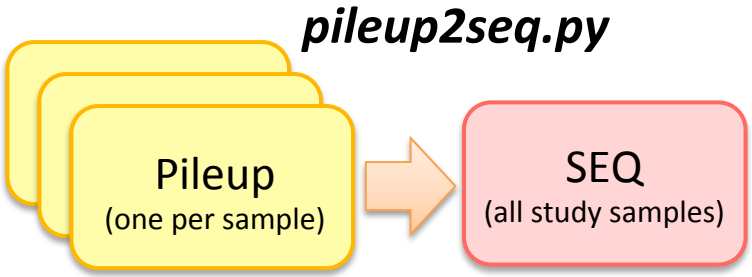
Step 2: use the *mpileup* option in *samtools* to generate pileup files.

```
samtools mpileup -q 30 -Q 20 -f hs37d5.fa.rz -l HGDP.bed A.bam > A.pileup
```

-q : minimum mapping quality score
-Q : minimum base quality score
-f : the reference file (human genome sequence in the FASTA format)
-l : the BED file (regions to pileup)

The above command processes sample A. *Repeat for other samples.*

From pileup files to a SEQ file (*pileup2seq*)



_.seq (tab-delimited)

```
POP_1   IND_1   2 1 20   0 0 0   15 5 22   ...
POP_1   IND_2   4 3 21   0 0 0   10 0 27   ...
POP_2   IND_3   2 2 30   0 0 0   0 0 0     ...
POP_3   IND_4   2 2 35   0 0 0   11 7 33   ...
...     ...     ...      ...     ...       ...
```

In the seq file, loci are in the same order as the reference panel.
Each locus has 3 space-delimited numbers:

1. sequencing depth (total number of reads);
2. number of reads that match the reference allele;
3. mean base quality score in *phred* scale.

From pileup files to a SEQ file (*pileup2seq*)

A typical command to generate a SEQ file named "test.seq":

```
python pileup2seq.py -m HGDP.site -o test A.pileup B.pileup C.pileup …
```

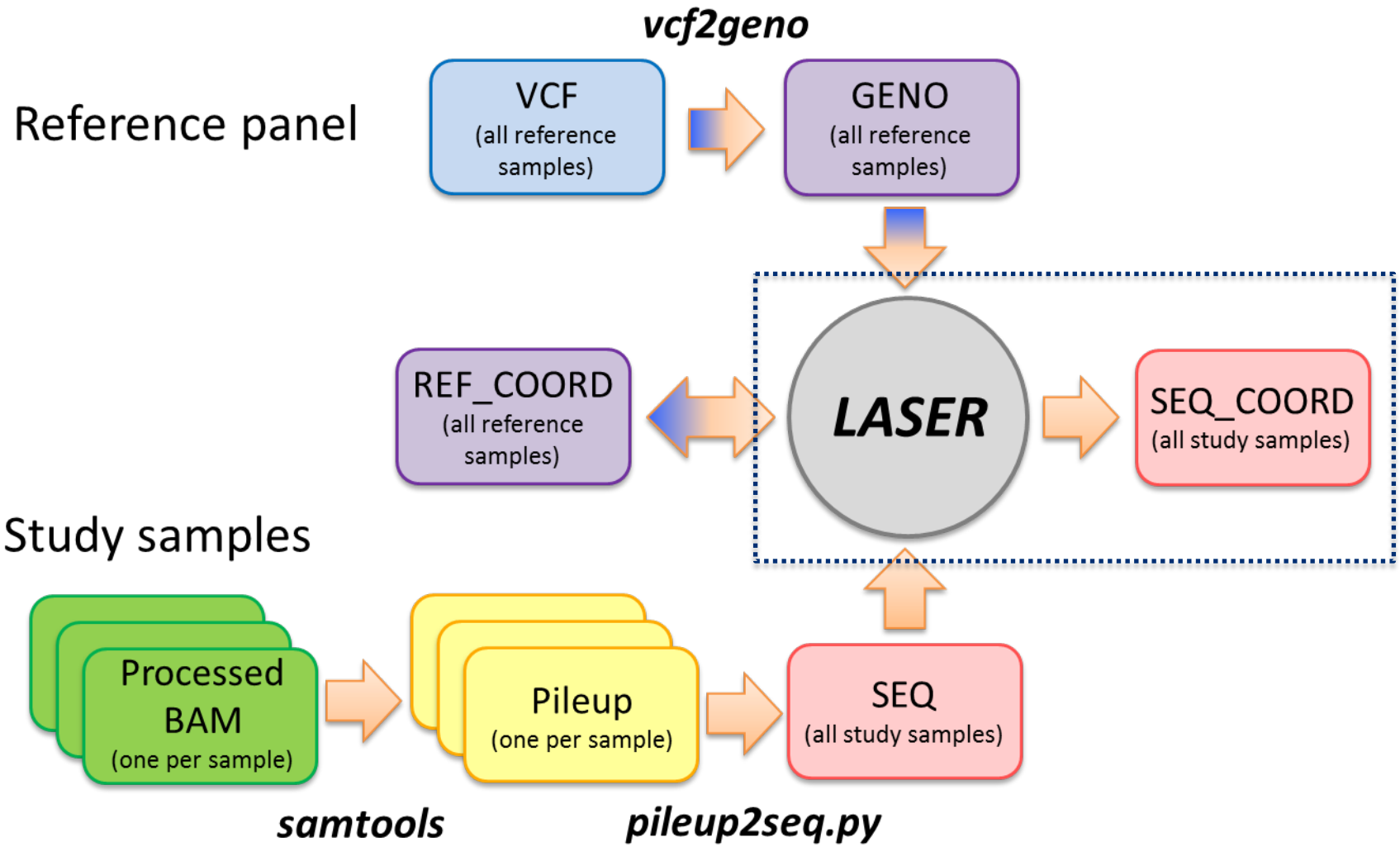
-m specifies the reference site file.
-o specifies the prefix of the output file.

Other options:
-b target.bed specifies target regions to exclude.
-i newID.txt specifies popID and indivID to output.

newID.txt

A	POP_1	IND_1
B	POP_1	IND_2
C	POP_2	IND_3
...	...	...

All input files are ready. Get started with LASER!



Parameters in LASER

There are 21 parameters in LASER, whose descriptions can be found in:
(1) LASER_Manual (section 5)
(2) The paramter file (laser.conf)

Ways to specify parameter values:
(1) Command line
(2) The parameter file
(3) If not (1) or (2), default values

If a parameter file is not specified (using -p), ***laser*** will search for a file named "laser.conf" under the working directory. If not found, ***laser*** will create a template "laser.conf" file under the working directory.

laser.conf

```
# This is a parameter file for LASER v2.01.
# The entire line after a '#' will be ignored.

###----Main Parameters----###

GENO_FILE          # File name of the reference genotype data (:

SEQ_FILE           # File name of the study sequence data (inclu

COORD_FILE         # File name of the reference coordinates (inc

OUT_PREFIX         # Prefix of output files (include path if out

DIM                # Number of PCs to compute (must be a positiv

DIM_HIGH           # Number of informative PCs for projection (r

MIN_LOCI           # Minimum number of covered loci in a sample


###----Advanced Parameters----###

###----Command line arguments----###

# -p     parameterfile (this file)
# -g     GENO_FILE
# -s     SEQ_FILE
# -c     COORD_FILE
# -o     OUT_PREFIX
# -k     DIM
# -K     DIM_HIGH
# -l     MIN_LOCI
# -e     SEQ_ERR
# -a     ALPHA
# -t     THRESHOLD
# -x     FIRST_IND
# -y     LAST_IND
# -r     REPS
# -R     OUTPUT_REPS
# -cov   CHECK_COVERAGE
# -fmt   CHECK_FORMAT
# -pca   PCA_MODE
# -ex    EXCLUDE_LIST
# -M     TRIM_PROP
# -minc  MAX_COVERAGE
# -maxc  MAX_COVERAGE

###----end of file----###
```

Basic usage

Use the parameter file:

```
./laser -p ./example/example.conf
```

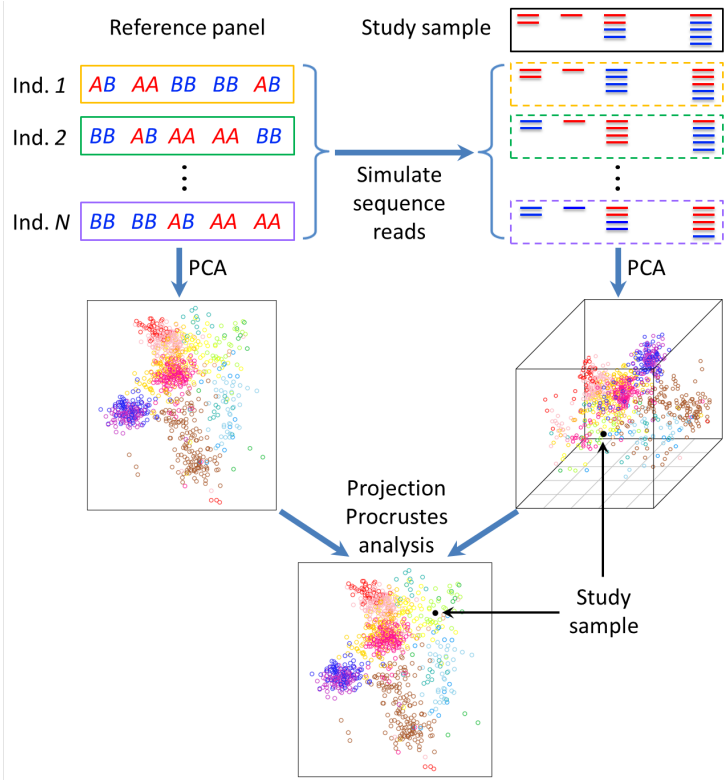
Or `./laser` (use laser.conf)

Use command lines:

```
./laser -g ./example/HGDP_238_chr22.geno \
    -c ./example/HGDP_238_chr22.RefPC.coord \
    -s ./example/HapMap_6_chr22.seq \
    -K 10 -k 5 \
    -o test
```

-p : the parameter file (default: laser.conf).
-g : the reference genotype file.
-c : the reference coordinate file.
-s : the sequence file for study samples.
-K : the dimension of the space to project from.
-k : the dimension of the reference space.
-o : the prefix of the output files.

The LASER alogrithm



Output results

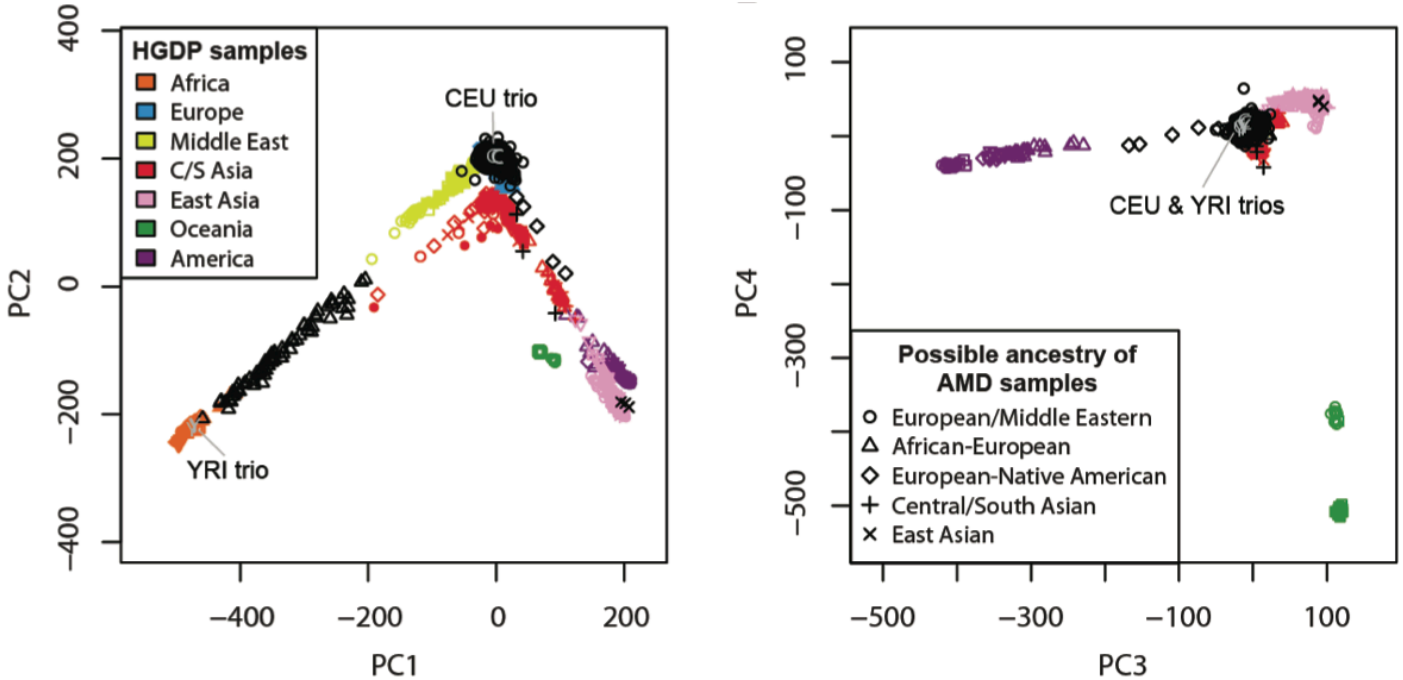
_.SeqPC.coord

popID	indivID	L1	Ci	K	t	PC1	PC2	PC3	PC4
YRI	NA19238	1411	0.3043	20	0.9818	59.116	30.930	3.358	-3.156
CEU	NA12892	1553	0.3301	20	0.9849	4.3483	-26.86	-0.155	3.2384
CEU	NA12891	1609	0.3621	20	0.9840	-7.538	-29.77	-10.71	11.523
CEU	NA12878	1581	0.3350	20	0.9845	4.1454	-32.19	-4.936	-4.433
YRI	NA19239	1558	0.3490	20	0.9852	61.863	24.512	-10.27	-15.23
YRI	NA19240	1735	0.4041	20	0.9869	70.584	35.145	2.871	18.385

L1 : number of loci being covered by at least one read
Ci : mean sequencing depth (averaged across loci)
K : number of PCs used for projection
t : sample-specific Procrustes similarity score
PCs : coordinates of sequenced samples in the reference ancestry space

Visualization and interpretation

Ancestry of the study samples can be identified by comparing with the reference individuals.



Example R scripts for plotting results can be found in the folder "plot" in the software package.

Running multiple LASER jobs in parallel

Because each sample is analyzed independently with the reference panel, we can run multiple jobs in parallel.

Submit two jobs to analyze samples 1 to 3, and 4 to 6:

```
./laser -p ./example/example.conf -x 1 -y 3  -o test.1-3 &
./laser -p ./example/example.conf -x 4 -y 6  -o test.4-6 &
```

-x : index of the first sample to analyze
-y : index of the last sample to analyze

Combine results

```
cp test.1-3.SeqPC.coord test.SeqPC.coord
more +2 test.4-6.SeqPC.coord >> test.SeqPC.coord
```

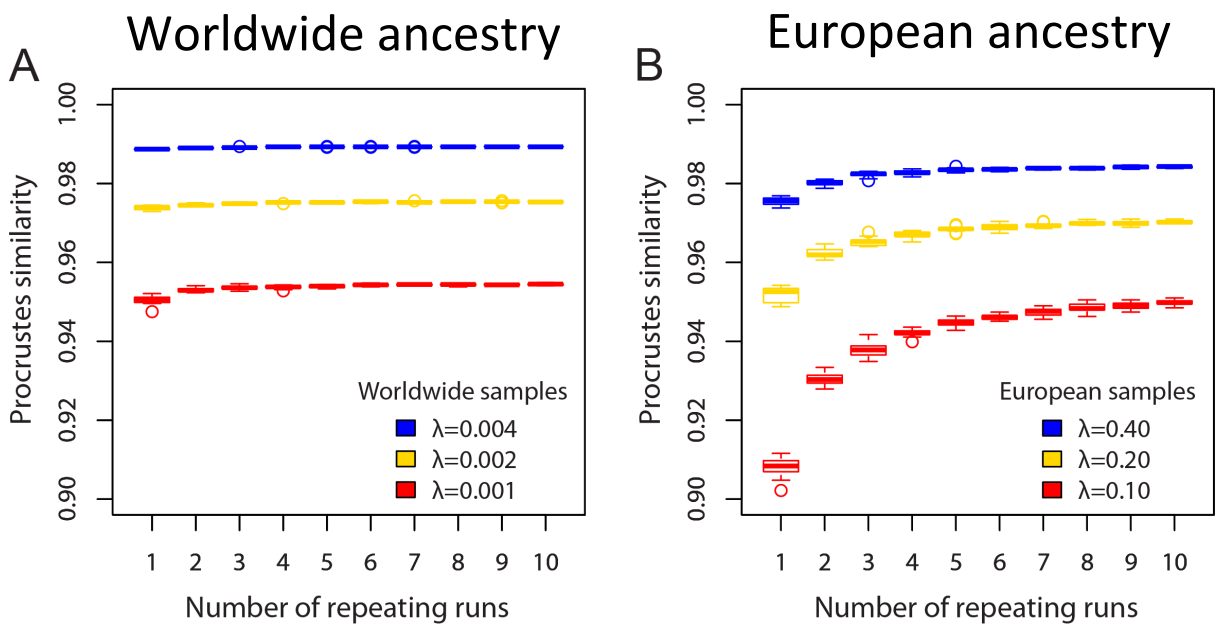
Using repeated runs to improve accuracy

Repeating LASER to analyze a sample multiple times and take the average coordinates can improve accuracy, especially for estimating fine-scale ancestry at very low coverage.

```
./laser -p ./example/example.conf -r 5 -R 1 -o test
```

-r : number of repeated runs

-R : if output detailed results of each run (0: No; 1: Yes)



Wang *et al.* 2014 *Nature Genetics*

Excluding loci from the analysis

Why we want to exclude some loci?

1. Targeted regions or special regions.
2. Memory limit (when the reference panel is large).
3. Computational speed.

Four options implemented in LASER to exclude loci:

-ex : exclude a given list of loci (also effective when running PCA mode)
-minc : exclude loci that have mean depth lower than a given value
-maxc : exclude loci that have mean depth higher than a given value
-M : randomly exclude a proportion of loci (from the reference panel)

The mean coverage per locus will be calculated based on samples being analyzed (results can be different when using together with -x and -y).

Example:

```
./laser -ex ./example/snps2exclude.txt -minc 0.001 -maxc 4 -M 0.5
```

Checking coverage distribution of your data

Check the mean sequencing depth across samples and loci:

```
./laser -p ./example/example.conf -cov 2 -o test
```

When -cov is set to 1 or 2, LASER will output the mean coverage for each sample and each locus to files named "test.ind.cov" and "test.loc.cov".

-cov :
1, check coverage and perform ancestry estimation;
2, check coverage and stop.
0, do not check coverage (default).

Practice

- Targeted sequencing data (BAM files) for 6 HapMap samples (two trios) are in the "$BAM" folder.
- Targeted regions are given in "$BAM/AMD_roi_1-based.bed".
- Sample IDs are given in "$BAM/AMD_hapmap_trios_id.txt".
- Reference panel: Human Genome Diversity Panel (HGDP).

Tasks:

1. Generate pileup files to the HGDP loci using ***samtools***
2. Prepare the SEQ file using ***pileup2seq.py***
3. Estimate ancestry in the HGDP ancestry space using ***laser***
4. Visualize results using R scripts

Commands and explanations are given on the wiki page:

http://genome.sph.umich.edu/wiki/SeqShop:_Estimates_of_Genetic_Ancestry_Practical

The best place to look for help: LASER_Manual.pdf

Have fun with LASER!

May the force be with you!