Population genomics

RNA-seq

Eyal Privman
Department of Ecology and Evolution, University of Lausanne

PhD Summer School
June 2012

UNIL | Université de Lausanne

SIB
Swiss Institute of Bioinformatics

Population genomics

Population ***genetics***:

The study of ***allele frequency distribution and change*** under the influence of evolutionary processes:

- genetic drift
- mutation
- natural selection
- recombination
- gene flow

(Wikipedia)

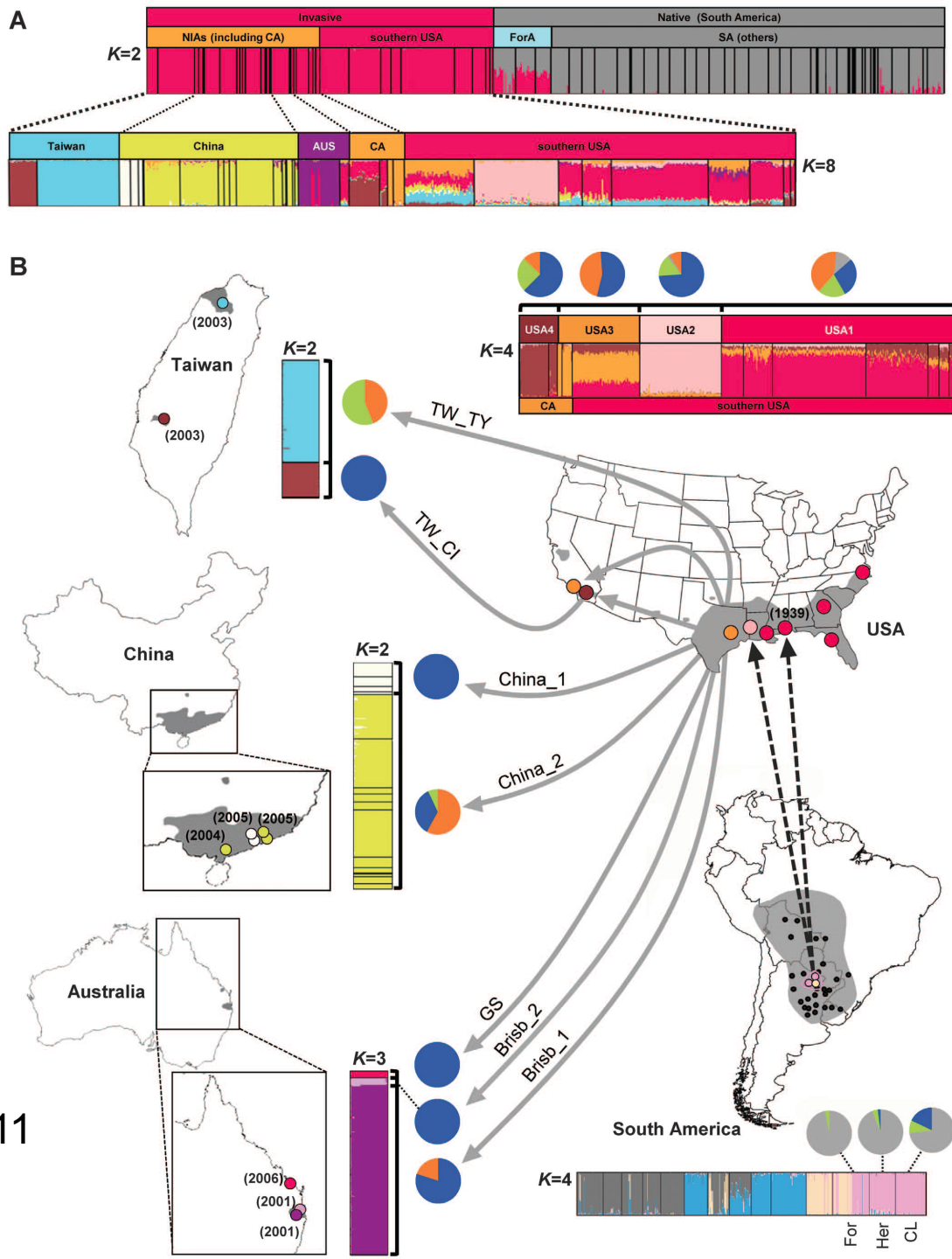
Population genomics: applications

- Reconstruct demographic history
  - Population expansion; bottlenecks; founder effect
  - Migration / invasion (fire ants, pharaoh ants, humans, flies…)

Ascunce *et al.* Science 2011

Population genomics:

- Reconstruct demographic h
  - Population expansion; b
  - Migration / invasion (fire



Ascunce *et al.* Science 2011

Population genomics: applications

- Reconstruct demographic history
  - Population expansion; bottlenecks; founder effect
  - Migration / invasion (fire ants, humans, flies…)
  - Admixture of populations (African Americans)

OPEN ACCESS Freely available online

PLoS GENETICS

Admixture Mapping of 15,280 African Americans Identifies Obesity Susceptibility Loci on Chromosomes 5 and X

Cheng *et al.* 2009

Population genomics: applications

- Reconstruct demographic history
  - Population expansion; bottlenecks; founder effect
  - Migration / invasion (fire ants, humans, flies…)
  - Admixture of populations (African Americans)
- Association studies – linking genotype & phenotype
  - Genome-Wide Association Studies (GWAS)
  - Admixture mapping
  - Quantitative Trait Loci (QTL) mapping; hybrid cross
- Detect selection
  - Local adaptation (F_{ST})
  - Selective sweep

Population genomic data

- SNP-chips
- RNA-seq
- RAD-seq
- Whole genome seq?

SNP discovery and genotyping from RNA-seq

- Short read RNA-seq assembled *de novo* or aligned to ref genome
- Each sample may contain one allele (haploid), two (diploid) or more, as in our case of pooled samples:
- **Fire ant RNA-seq data mapped to genome:**
  - 5 males = 5 alleles
  - 4 queens = 8 alleles
  - 200 workers = 400 alleles
- Note: this is not a “proper” population sample
  - It’s more useful to sequence individual separately
  - It’s more useful to detect SNPs using large cohorts of sequenced individuals

SAMtools

- Process SAM (Sequence Alignment/Map) and BAM format (compressed binary version of SAM)
- E.g. Process Bowtie or TopHat alignments of RNA-seq or RAD-seq to a reference genome
- Use the manual page to write samtools command lines: http://samtools.sourceforge.net/samtools.shtml

Convert BAM to SAM

Use **`samtools view`** to convert the binary format to readable text:

```
$ samtools view W422.accepted_hits.sort.bam | head
R2D2_0117:2:44:1714:17921#0 16  SIgn00001   79  50  75M *   0   0
ATTAAGTTCTAGTTCAAATAACTTAGGATTGTCTGTTGTATAGCTCACAAGCATGACGTAACCATTTGGTCCACG
HHHHHFGGFGEDHHHHFHFHHBHDHHHHHHHHHHGGGGGD3HHHHHFHHHGGG>GGHHHGGHHHHGHHHHEHHHH
XA:i:0  MD:Z:75 NM:i:0  NH:i:1
R2D2_0117:2:2:9366:2587#0   16  SIgn00001   96  50  75M *   0   0
ATAACTTAGGATTGTCTGTTGTATAGCTCACAAGCATGACGTAACCATTTGGTCCACGAACTTCCTGTATACCTG
IIIGIIIIIIIHIIHIIIIIIIIIIIIIIHIHIIHIIIFIIIIIIIIIIIIIIIIIIIIIIIIIIIIHHIII
XA:i:0  MD:Z:75 NM:i:0  NH:i:1
R2D2_0117:2:26:7962:5823#0  16  SIgn00001   101 50  75M *   0   0
TTAGGATTGTCTGTTGTATAGCTCACAAGCATGACGTAACCATTTGGTCCACGAACTTCCTGTATACCTGTCTTA
IIIIIIIIHHDGDDBIGGIIIIHIHIIIIIIIIIHIIIIFIIIIIIIIIIIGIIIIIIIIIIIIIIIIIIHIII
XA:i:0  MD:Z:75 NM:i:0  NH:i:1
R2D2_0117:2:75:3630:14196#0 0   SIgn00001   113 50  75M *   0   0
GTTGTATAGCTCACAAGCATGACGTAACCATTTGGTCCACGAACTTCCTGTATACCTGTCTTAGTCTTGTTCTTA
HIIHIIIIHIIIIIIIIIIIIIHIFIIIIHHIIIHIIHFIIEGIIIIIIIIIHIHIHIHHIIIIIIIIIIIHIG
XA:i:0  MD:Z:75 NM:i:0  NH:i:1
...
...
```

Convert BAM to SAM

`samtools view` can cut a desired region for us:

(Note: BAM file needs to be sorted)

```
$ samtools view W422.accepted_hits.sort.bam SIgn00002:100,000-110,000
R2D2_0117:2:68:6557:15142#0 0   SIgn00002   108809  50  75M *   0   0
CTTAGAACTCATCATGTCTCACAATATACGCATCGCAAAACAGAAATTATCATCTATGGATCGAGGGTAAGCGTC
IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIHIIIIIIIIIIIIIIIIIDHIIIIGI
XA:i:0  MD:Z:75 NM:i:0  NH:i:1
R2D2_0117:2:116:5195:5494#0 0   SIgn00002   109328  50  75M *   0   0
GTCGGCAATACTTTAGTGATCGCGGCTGTAATTACCACGAGGAGATTACGGTCTGTGACTAATTGTTACGTGTCT
IIIDIIIIIIIIIIIIIIIIHIIIHHEIIIIIIIIIHFIIEIGIIIIIIGIEIBG+GGEEIIIGIIIIGHFHI
XA:i:1  MD:Z:67T7   NM:i:1  NH:i:1
R2D2_0117:2:60:3752:12219#0 16  SIgn00002   109341  50  75M *   0   0
TAGTGATCGCGGCTGTAATTACCACGAGGAGATTACGGTCTGTGACTAATTGTTTCGTGTCTAGCTTGGCTGCTG
EHIIIIGIHIIHFHIIIIIIGIIGIIIIIIIIHIIIIIIIIIIIIIIHIIIIIIIIIIIIIIIHIIIIDIIIII
XA:i:0  MD:Z:75 NM:i:0  NH:i:1
R2D2_0117:2:26:8854:8712#0  16  SIgn00002   109374  50  75M *   0   0
TACGGTCTGTGACTAATTGTTACGTGTCTAGCTTGGCTGCTGCAGATTTACTGGTCGGTCTAGCGGTGATGCCAC
IDIHIIIIHIIGIHHHIIIIHGIIIIDIHIGIHHIIIIIIIIIIIIIIIIIIIIIIIHGIIIIIIIIIIIII
XA:i:1  MD:Z:21T53  NM:i:1  NH:i:1
```

Combine, sort, index BAM files

Concatenate: **`samtools cat`** combines several files into one:

```
$ samtools cat *.bam > all.bam
```

(Note: SAMtools commands were designed for combination by UNIX pipes)

Sort: Use **`samtools sort`** to sort the data in a BAM file:
(needed before indexing)

```
$ samtools sort W422.bam W422.sort
```

(Note: the ".bam" suffix will be added to the output parameter)

Index: **`samtools index`** creates an index file for a (sorted!) BAM file that allows instantaneous access to individual records:

```
$ samtools index W422.sort.bam
```

Merge: **`samtools merge`** combines (sorted!) BAM files:

```
$ samtools merge merge.sort.bam *.sort.bam
```

* **Note:** When merging files you may need to correct the header information and read group (RG) information. See exercise instructions.

GATK

- Package for population genomics from the BROAD Institute: http://www.broadinstitute.org/gsa/wiki/index.php/Main_Page (used in the 1000 genomes project)
- Disclaimer: GATK (and others) were designed for genotyping ***diploid individuals only***.
  - We will use them to genotype pools with >2 alleles, which does not fit with the probabilistic model used.
  - So our best are pools of 5 males (haploids)

GATK single sample genotype likelihoods

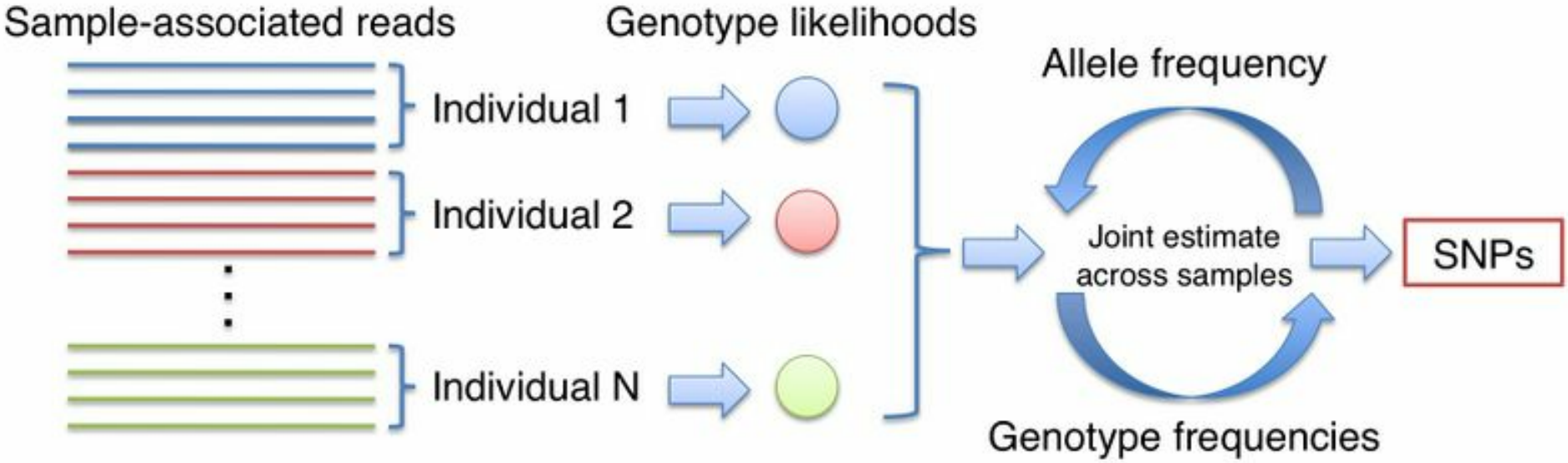
Bayesian model

Likelihood for the genotype | Prior for the genotype | Likelihood of the data given the genotype | Independent base model

$$L(G \mid D) = P(G)P(D \mid G) = \prod_{b \in \{good\_bases\}} P(b \mid G)$$

- Priors applied during multi-sample calculation; P(G) = 1
- Likelihood of data computed using pileup of bases and associated quality scores at given locus
- Only "good bases" are included: those satisfying minimum base quality, mapping read quality, pair mapping quality, NQS
- P(b | G) uses a platform-specific confusion matrix
- L(G|D) computed for all 10 genotypes

The Broad Unified Genotyper SNP caller multiple-sample allele frequency and genotype estimates



- This approach allows us to combine weak single sample calls to discover variation among samples with high confidence

See http://www.broadinstitute.org/gsa/wiki/index.php/Unified_genotyper for more information

Genotype: the Unified Genotyper

- The main program for genotyping SNPs: http://www.broadinstitute.org/gsa/wiki/index.php/Unified_genotyper
- Example usage: http://www.broadinstitute.org/gsa/gatkdocs/release/org_broadinstitute_sting_gatk_walkers_genotyper_UnifiedGenotyper.html
- Example usage: http://www.broadinstitute.org/gsa/gatkdocs/release/org_broadinstitute_sting_gatk_walkers_genotyper_UnifiedGenotyper.html

Genotype: the Unified Genotyper

The **UnifiedGenotyper** function in **GenomeAnalysisTK** detects SNPs and genotypes samples:

```
$ java -jar $GATK_HOME/GenomeAnalysisTK.jar \
-R SINV_subset_1.fa \
-I M350B.accepted_hits.sort.rg.bam \
-T UnifiedGenotyper \
-o M350B.accepted_hits.sort.rg.bam.snps.raw.vcf
```

(Note: Assumes the BAM file is sorted and indexed)

(Note: Requires read group information and corresponding header lines. See exercise instructions)

This will analyze one sample. The same command can be used for multiple samples.

The VCF output format

- GATK outputs a VCF (Variant Call Format) file:

```
[HEADER LINES]
#CHROM  POS     ID          REF     ALT     QUAL    FILTER  INFO          FORMAT          NA12878
chr1    873762  .           T       G       5231.78 PASS    [ANNOTATIONS] GT:AD:DP:GQ:PL  0/1:173,141:282:99:255,0,255
chr1    877664  rs3828047   A       G       3931.66 PASS    [ANNOTATIONS] GT:AD:DP:GQ:PL  1/1:0,105:94:99:255,255,0
chr1    899282  rs28548431  C       T       71.77   PASS    [ANNOTATIONS] GT:AD:DP:GQ:PL  0/1:1,3:4:25.92:103,0,26
chr1    974165  rs9442391   T       C       29.84   LowQual [ANNOTATIONS] GT:AD:DP:GQ:PL  0/1:14,4:14:60.91:61,0,255
```

- QUAL: The Phred scaled probability that REF/ALT polymorphism exists given sequencing data.
- Default threshold is 30.0
- Format field names describe the next list of values:
  - GT (genotype): 0/1 means heterozygote ref/alt
  - DP (depth): total number of reads mapped
  - AD (allele depth): count of ref/alt alleles