Population genomics

RAD-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

Fire ant population RAD-seq data

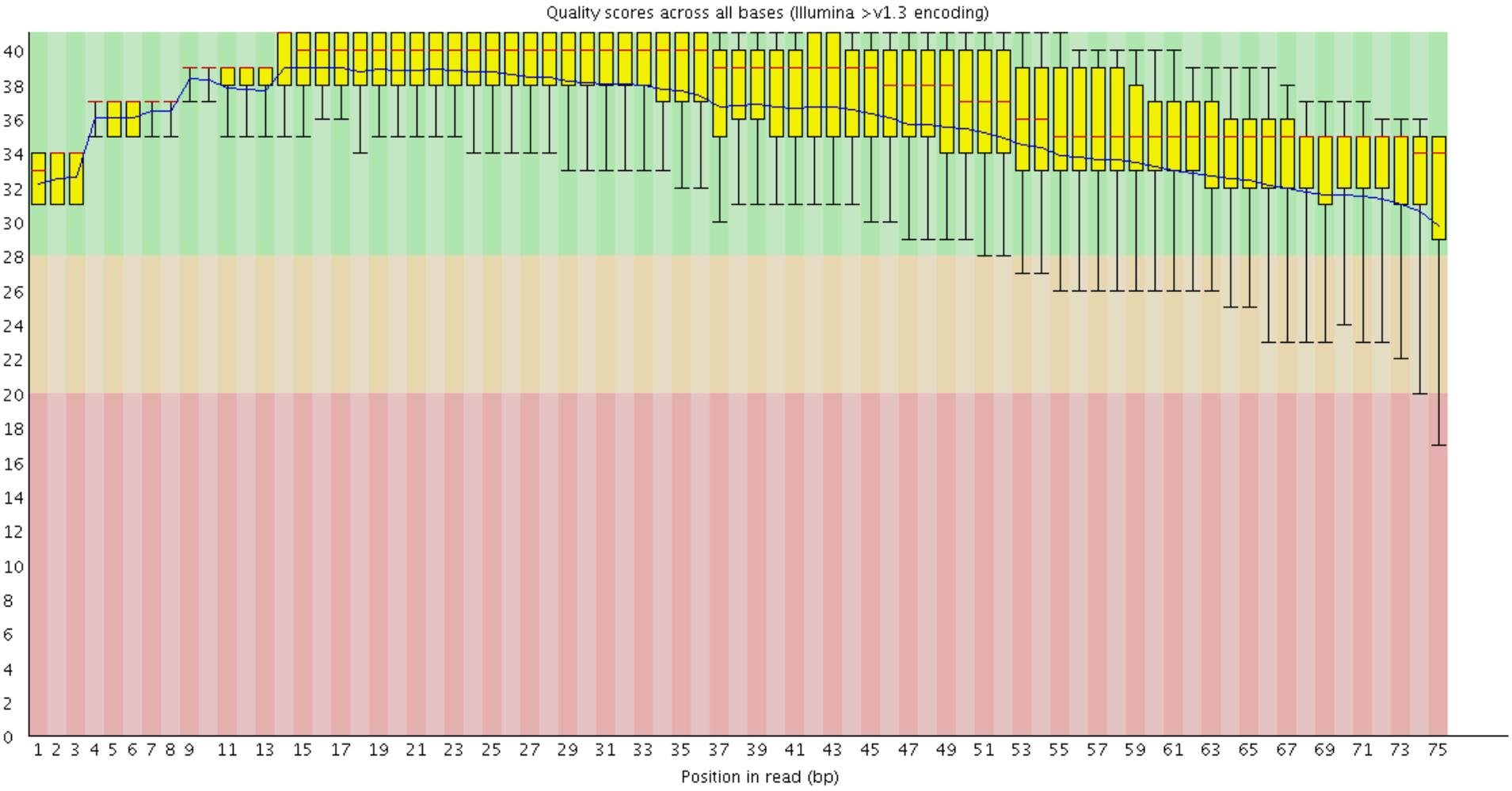
- 94 diploid workers (one per nest in sampled region)
- RAD protocol (Baird *et al.* Plos One 2008)
  PstI restriction enzyme:
  6-cutter, expect 1 site per 10Kbp

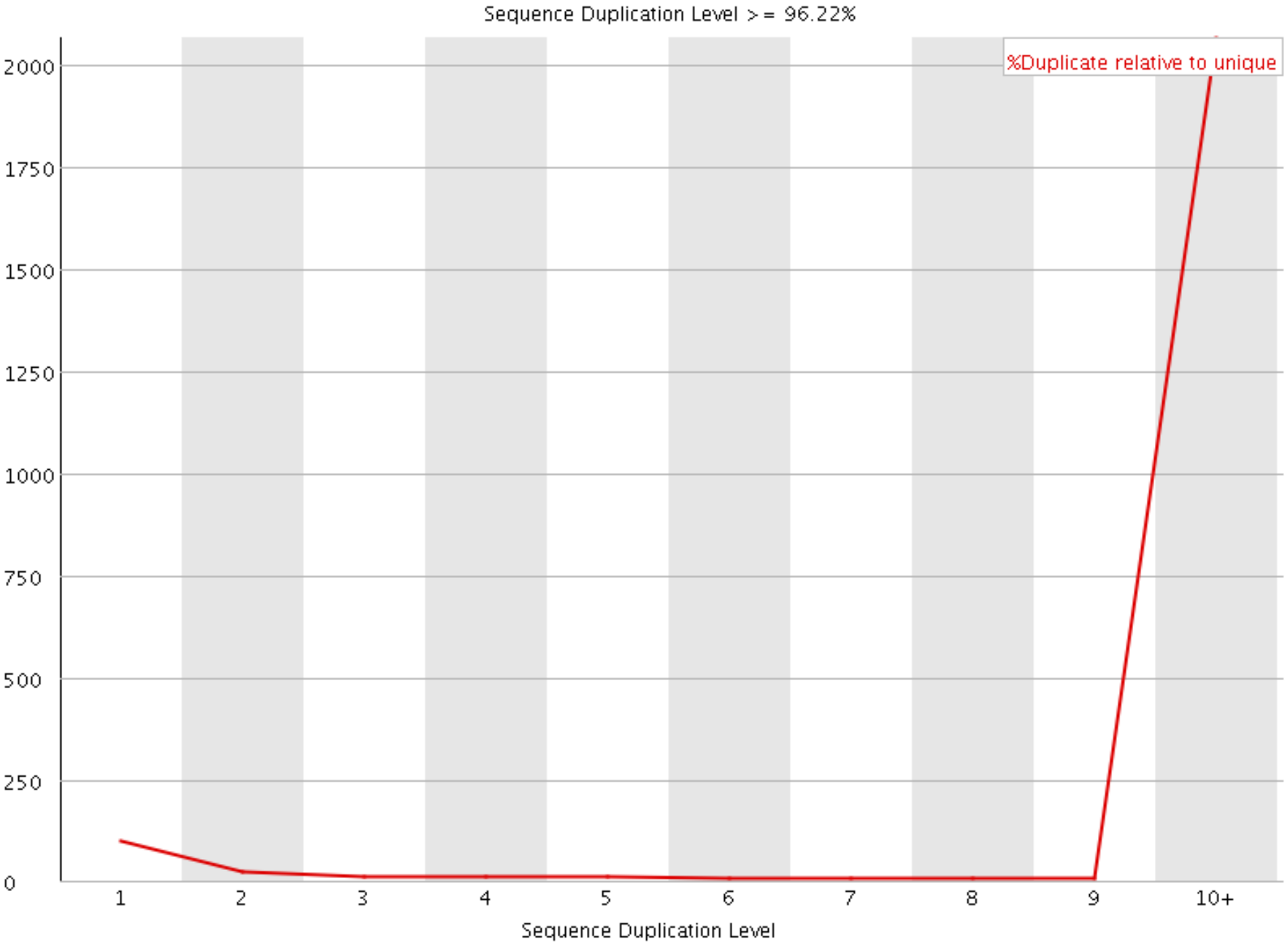
- Multiplexed Illumina 75bp single read sequencing
- ***Surprise:*** Not one simple population!

Stacks

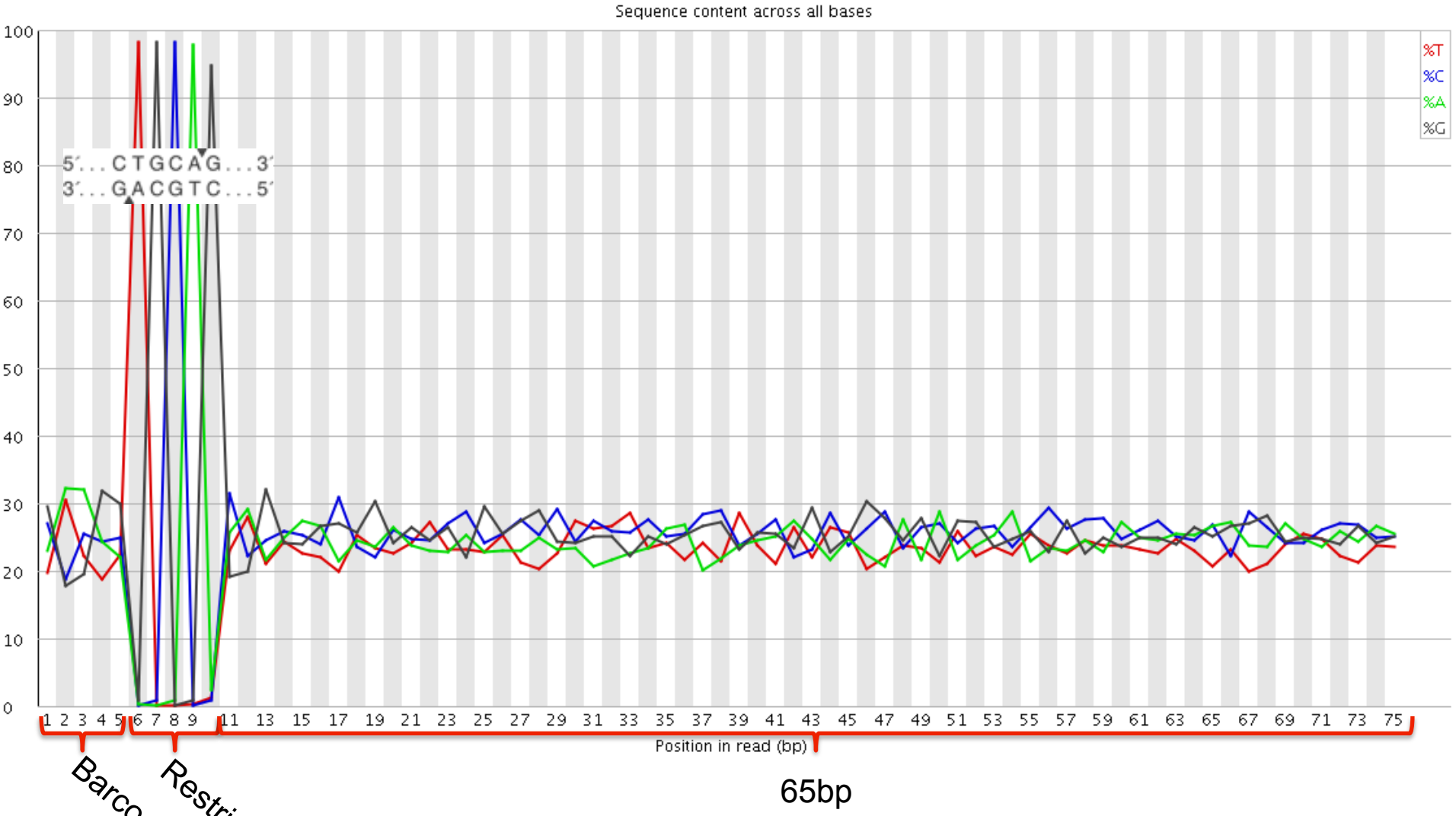
- http://creskolab.uoregon.edu/stacks/
- Process RAD-seq data from either
  - Parents and progeny samples → linkage map
  - Population samples → population genetics
- Can work with / without reference genome
- Can assemble mini-contigs of second-reads for paired-end sequencing

Quality


Quality scores across all bases (Illumina >v1.3 encoding)
40
38
36
34
32
30
28
26
24
22
20
18
16
14
12
10
8
6
4
2
0
1 2 3 4 5 6 7 8 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39 41 43 45 47 49 51 53 55 57 59 61 63 65 67 69 71 73 75
Position in read (bp)


Sequence Duplication Level >= 96.22%
%Duplicate relative to unique
2000
1750
1500
1250
1000
750
500
250
0
1
2
3
4
5
6
7
8
9
10+
Sequence Duplication Level

Base content

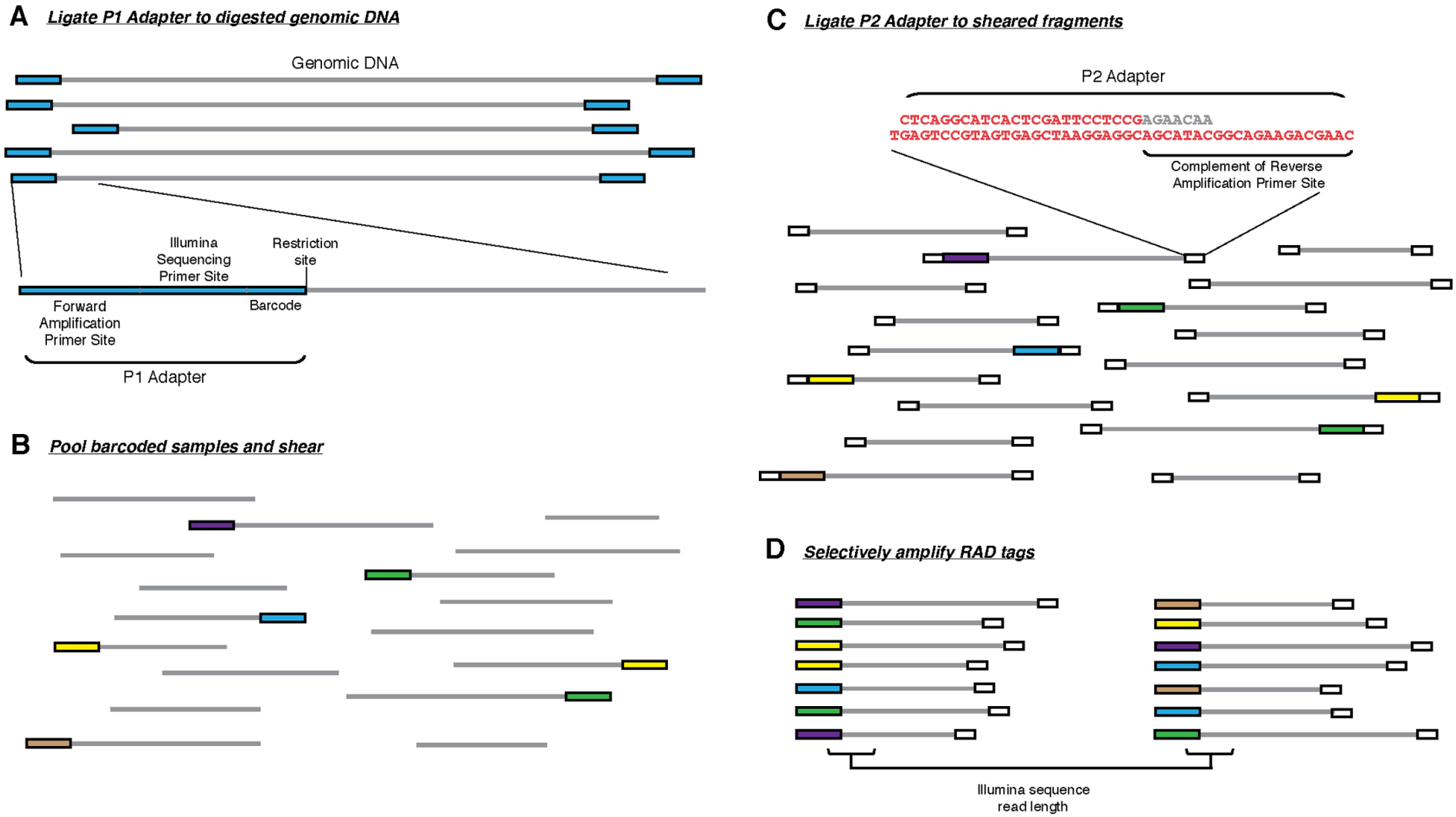


Barcode (5bp)

Restriction site (5bp)

65bp

RAD sequencing



Baird *et al.* Plos One 2008

Pre-processing

`process_radtags` processes raw Illumina input:

- De-multiplexing and removing barcodes
- Sliding-window quality filter
- Truncate reads

(alternatively – FASTX)

```
$ process_radtags -f in.fastq -o out_dir -i fastq -b
barcodes -e sbfI -c -q -r -D
```

```
f – path to the input file if processing single-end seqeunces.
i – input file type, either 'bustard' for the Illumina BUSTARD output
files, or 'fastq' (default 'fastq').
o – path to output the processed files.
b – a list of barcodes for this run.
e – specify the restriction enzyme to look for (either 'sbfI', 'pstI',
'ecoRI', or 'sgrAI').
c – clean data, remove any read with an uncalled base.
q – discard reads with low quality scores.
r – rescue barcodes and RAD-Tags.
D – capture discarded reads to a file.
```

Mapping

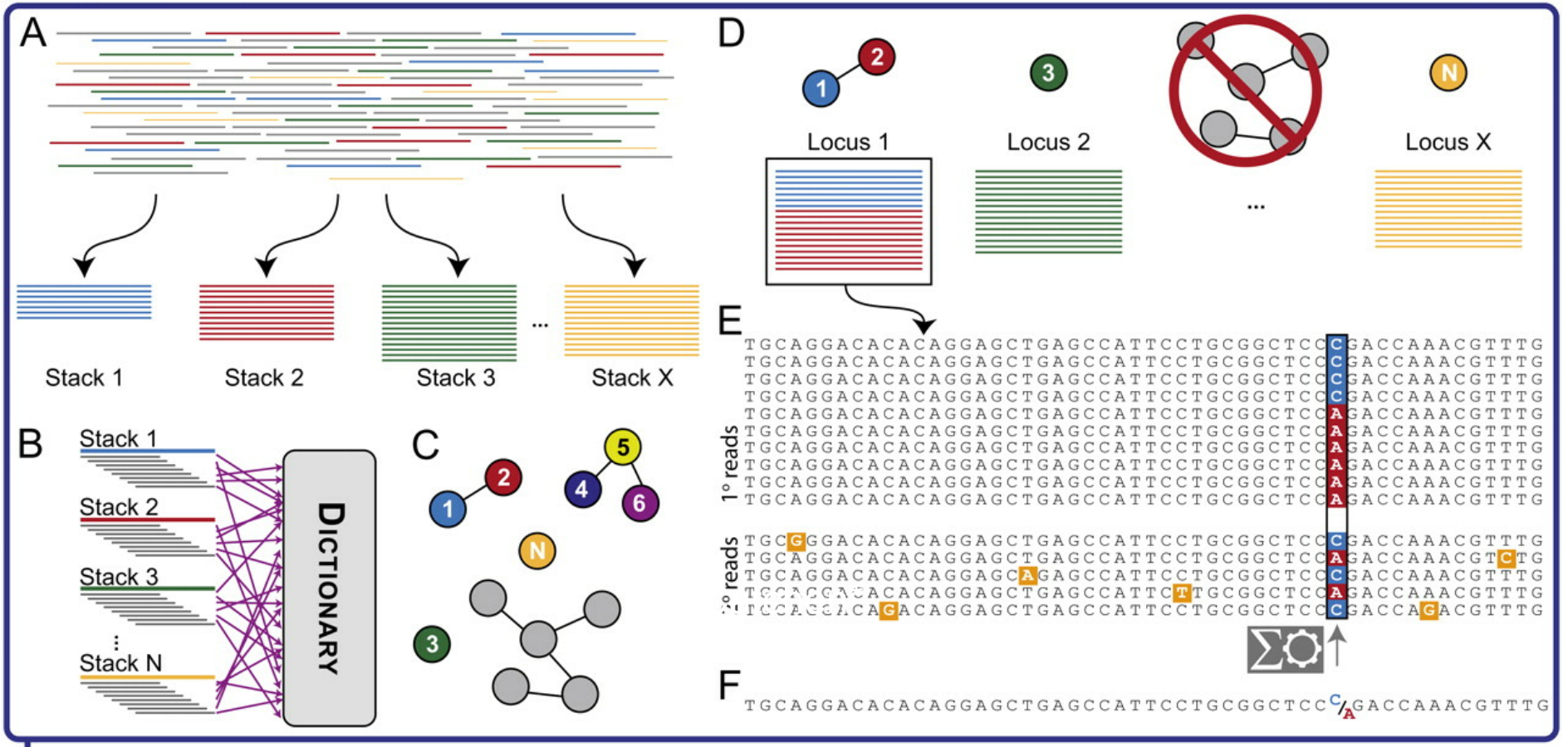
Use **`bowtie`** to map RAD tags to the genome

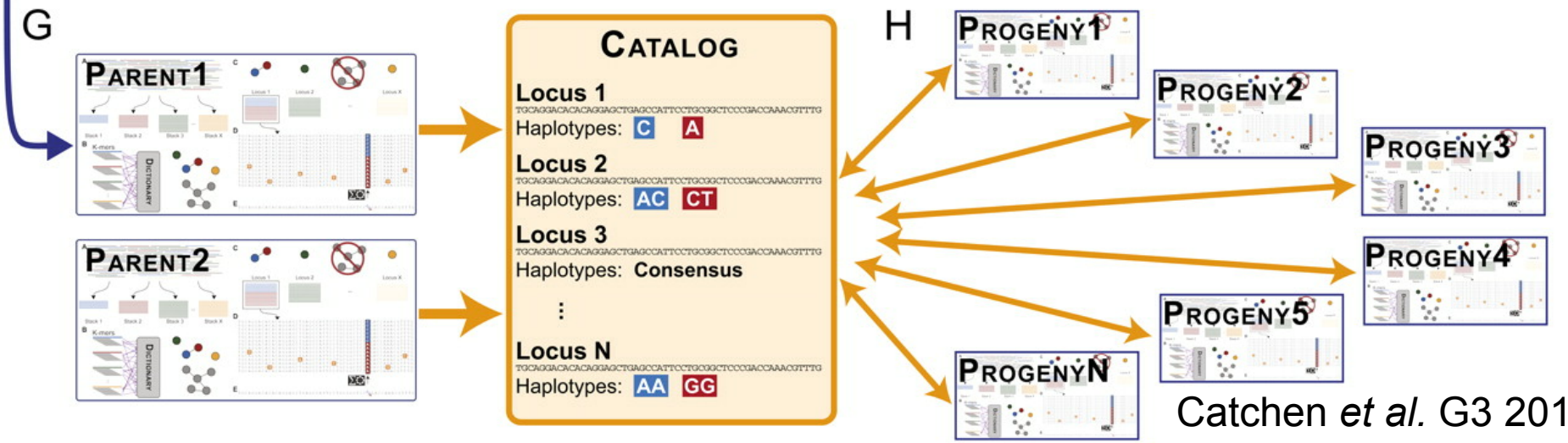
- Allow a few mismatches
  – *limits the number of SNPs per RAD tag!*
- Do not allow indels
  – *polymorphic indels will not be found!*

Stacks pipeline

Use RAD tags to the genome

- **`ustacks`** (unique stacks): Builds loci de novo and detects haplotypes in one individual
- **`cstacks`** (catalog stacks): Merges loci from multiple individuals to form a catalog
- **`sstacks`** (search stacks): Matches loci from an individual against a catalog
- **`pstacks`** (population stacks): Takes cleaned reads aligned to a reference genome, builds stacks based on the genomic locations of the reads, and detects haplotypes in one individual





Catchen *et al.* G3 2011