Tackling Genomic Data:

The Advantages and Applications of Bioinformatics Pipelines

Brad Reil

What's the plan…?

- What is a Bioinformatics Pipeline?
- Why do we need bioinformatics pipelines?
- How to we generate data for a pipeline?
- How does a pipeline operate (STACKS example)?
- What types of analyses do pipelines facilitate?
- What technical know-how is required?

What is a Bioinformatics Pipeline?

Bioinformatics – a Definition[1]

(Molecular) **bio** – **informatics:** bioinformatics is conceptualising biology in terms of molecules (in the sense of Physical chemistry) and applying **"informatics techniques"** (derived from disciplines such as applied maths, computer science and statistics) to **understand** and **organise** the **information** associated with these molecules, on a **large scale.** In short, bioinformatics is a management information system for molecular biology and has many **practical applications.**

[1]As submitted to the Oxford English Dictionary.

Pipeline

A set of **data processing elements** connected in **series**, where the output of one element is the input of the next one. - Wikapedia

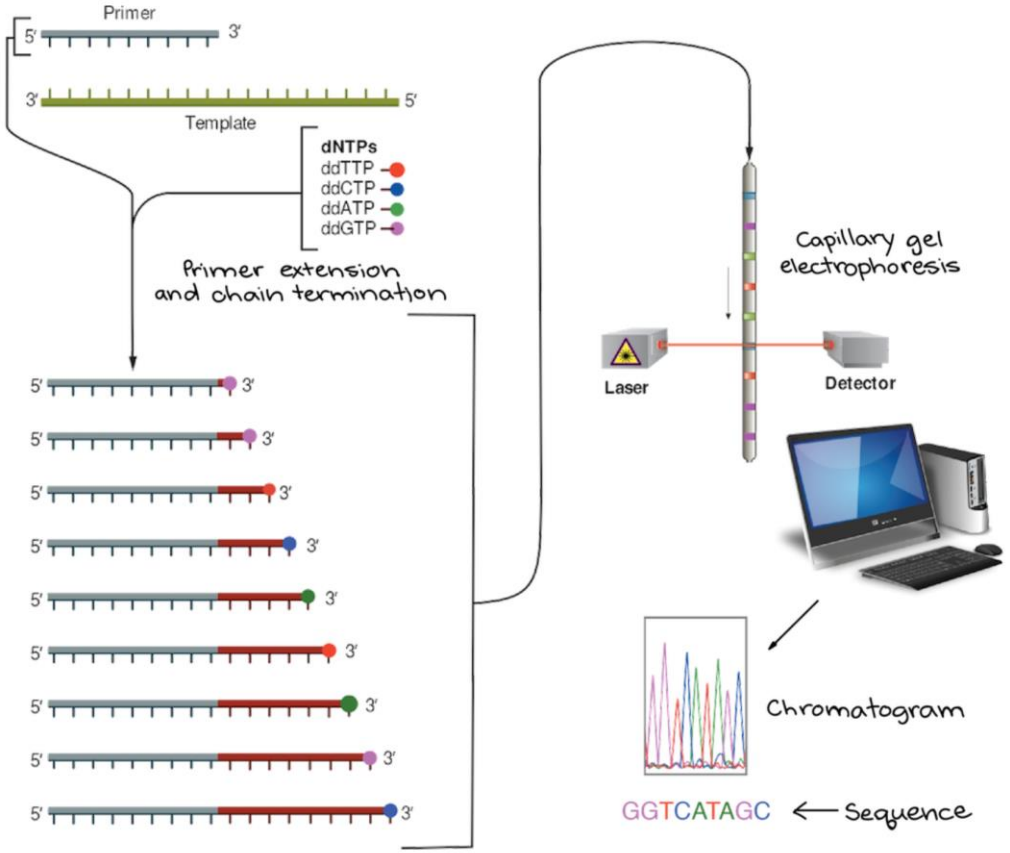
Bioinformatics Pipeline

A set of **data processing elements** connected in **series**, used to understand and organize the information associated with large scale data sets derived from biological molecules.

Luscombe, N. M., Greenbaum, D., & Gerstein, M. (2001). What is bioinformatics? A proposed definition and overview of the field. *Methods of information in medicine*, *40*(04), 346-358.

Why do we need these pipelines?

Well, we started with…



Sanger Sequencing

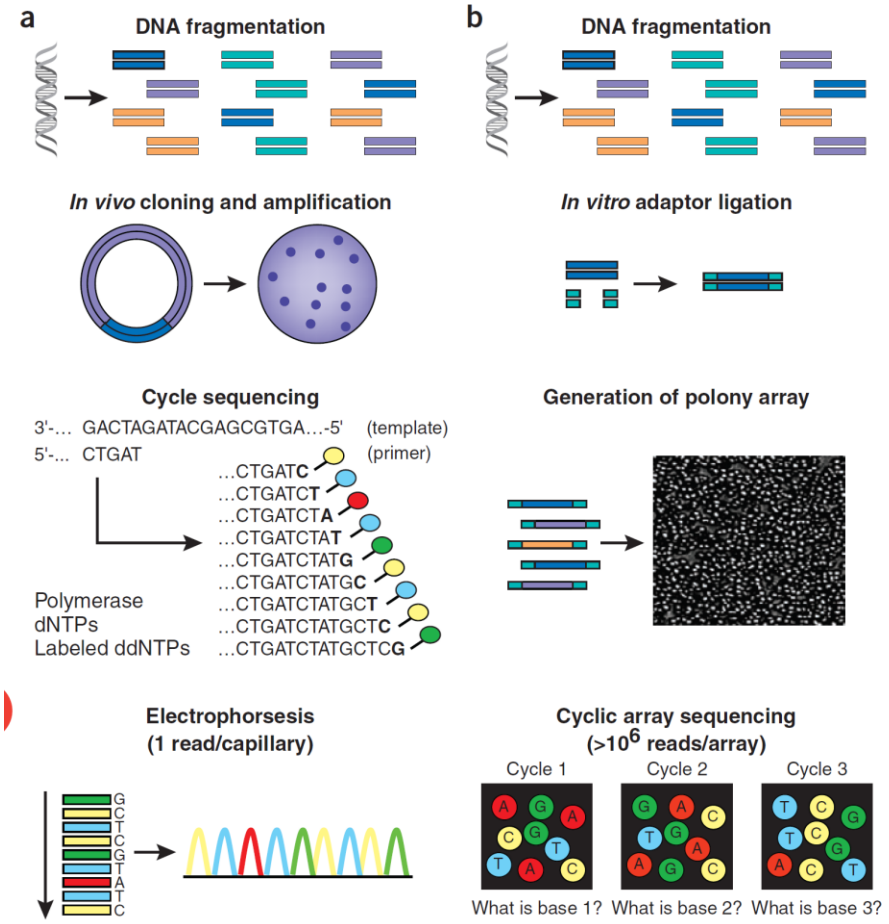
Sanger, F., Nicklen, S., & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the national academy of sciences*, *74*(12), 5463-5467.

Why do we need these pipelines?

Enter Next-Generation Sequencing Techniques…

Next-Generation Sequencing

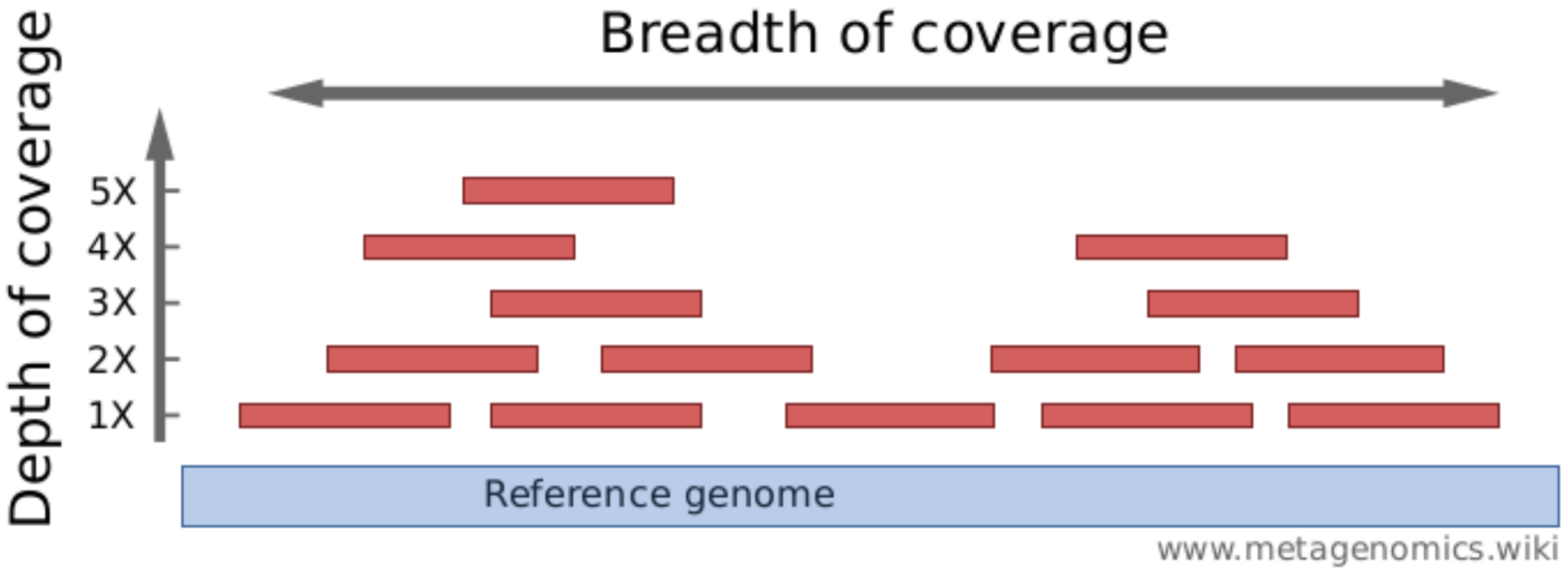
Term used to describe a collection of molecular sequencing techniques capable of producing reads for large arrays of different DNA fragments simultaneously.



Shendure, J., & Ji, H. (2008). Next-generation DNA sequencing. *Nature biotechnology*, *26*(10), 1135.

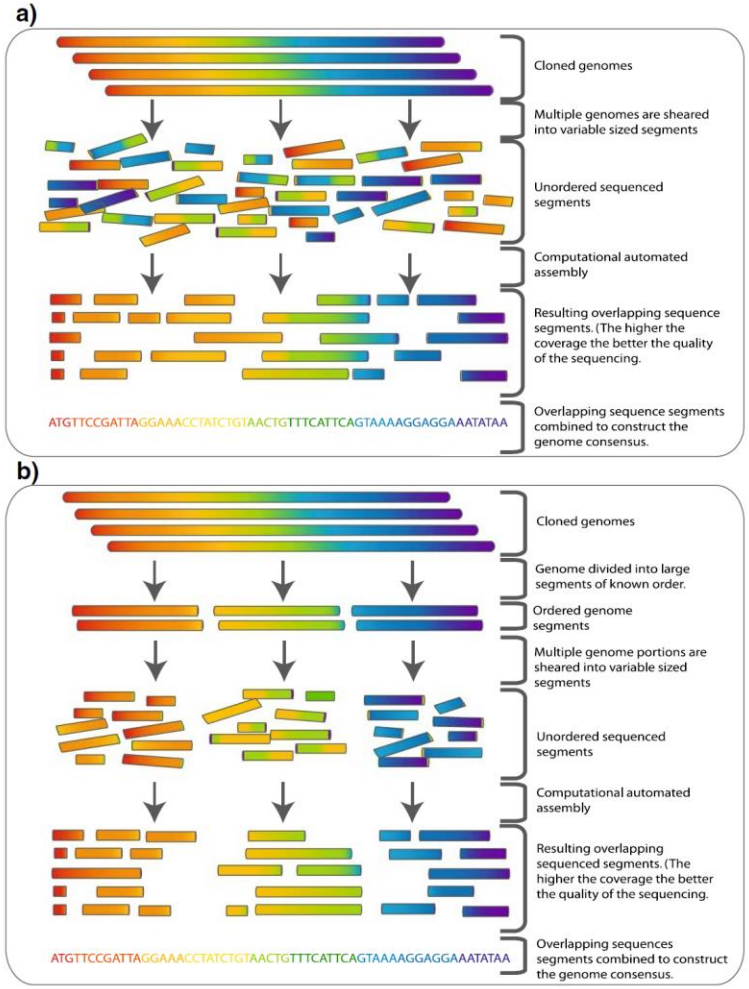
Sampling the Genome - DNA Fragmentation

Some quick terminology…



Sampling the Genome - DNA Fragmentation

So, how do we **fragment** and sample the genome…



Shotgun Sequencing

DNA Randomly Sheered By:

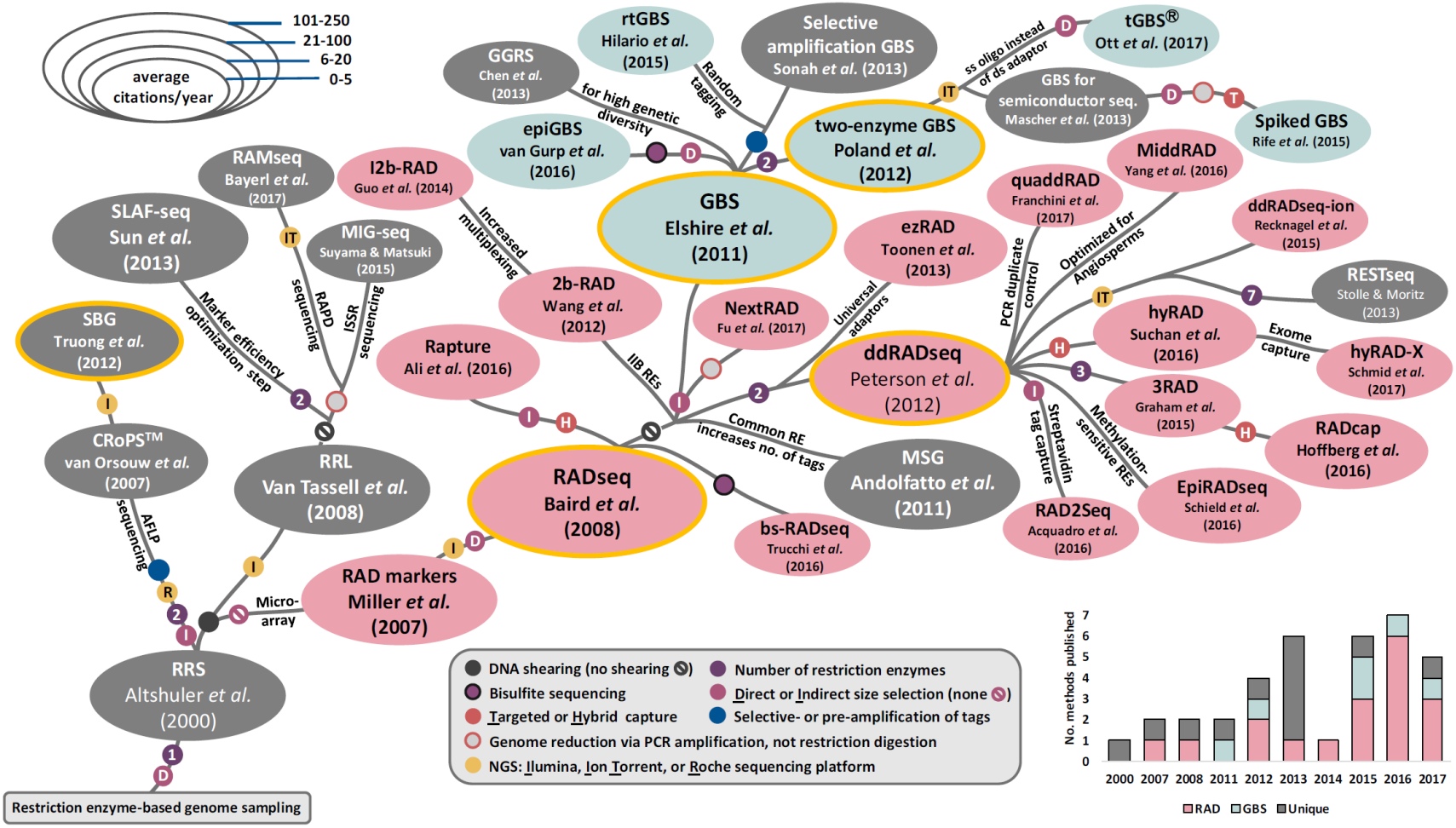
- Vortexing
- Glass Beads
- Sonication
- Radiation

The fragments in a specific size range are then isolated and sequenced

Commins, J., Toft, C., & Fares, M. A. (2009). Computational biology methods and their application to the comparative genomics of endocellular symbiotic bacteria of insects. *Biological procedures online*, *11*(1), 52.

DNA Fragmentation by Restriction Enzymes

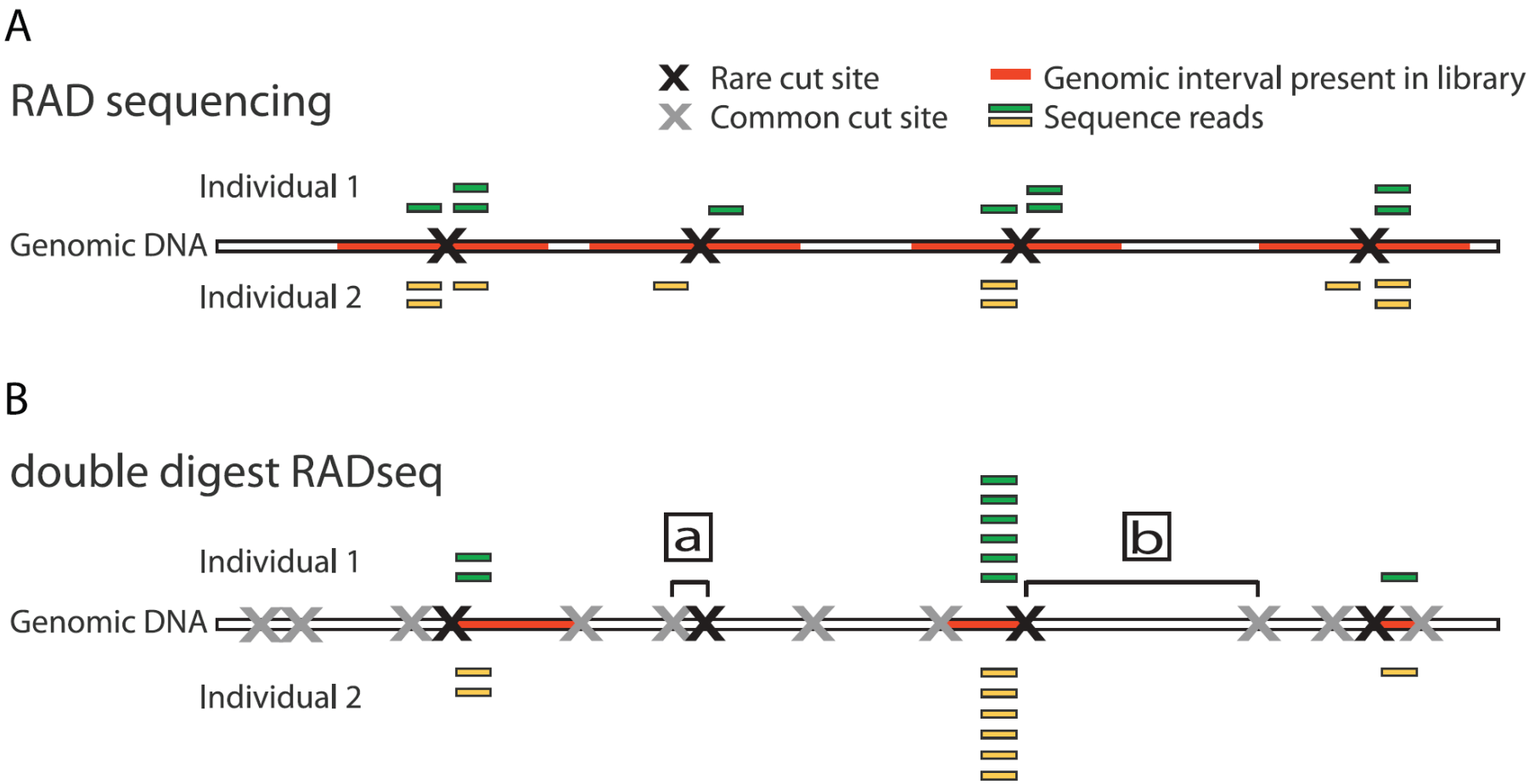
So, how do we **fragment** and sample the genome…



Campbell, E. O., Brunet, B. M., Dupuis, J. R., & Sperling, F. A. (2018). Would an RRS by any other name sound as RAD?. *Methods in Ecology and Evolution*, *9*(9), 1920-1927.

DNA Fragmentation by Restriction Enzymes

So, how do we **fragment** and sample the genome…



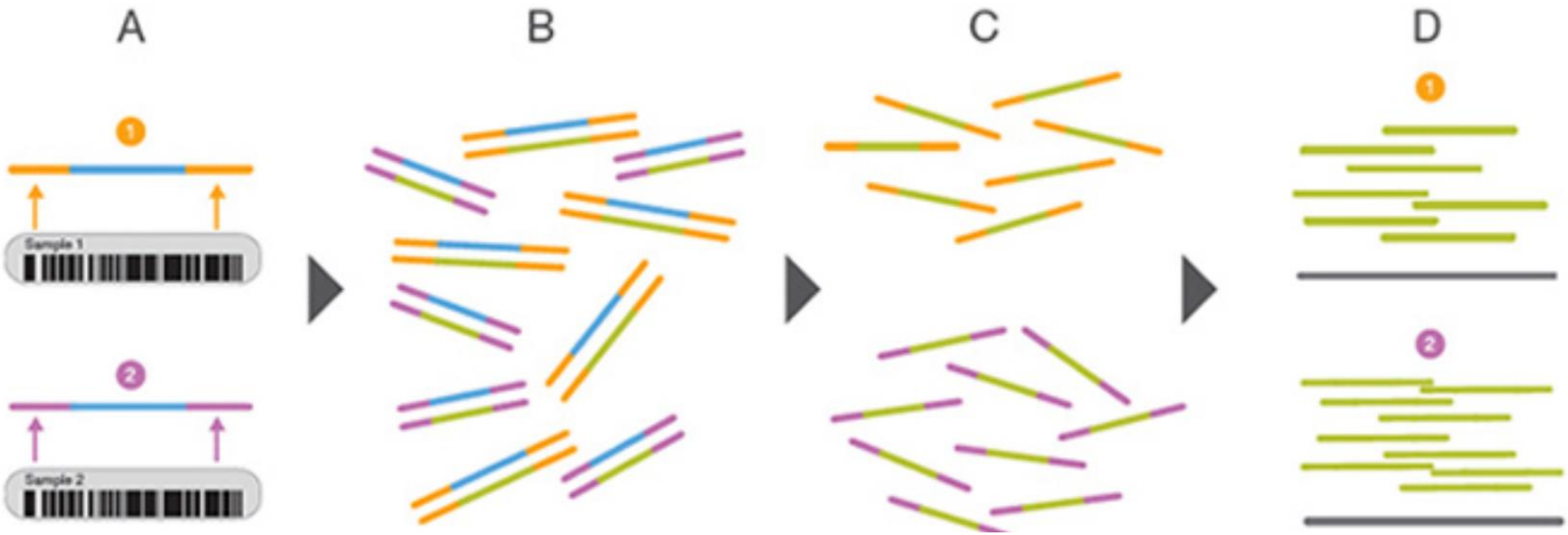
Peterson, B. K., et al. (2012). Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PloS one*, *7*(5), e37135.

DNA Multiplexing (Combinatorial Indexing)

And how do we **organize** large amounts of DNA…

DNA Multiplexing

The process of reducing repetitive sequencing steps by the mixing together of different DNA samples. This is achieved by tagging the samples isolated from a particular sample with a unique barcoding sequence via PCR.



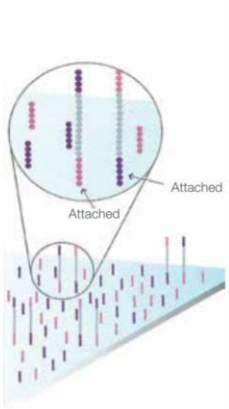
Shokralla, S., Porter, T. M., Gibson, J. F., Dobosz, R., Janzen, D. H., Hallwachs, W., ... & Hajibabaei, M. (2015). Massively parallel multiplex DNA sequencing for specimen identification using an Illumina MiSeq platform. *Scientific reports*, *5*, 9687.

Next-Gen Sequencing Platforms

Now the genome is sampled, organized, and ready to **sequence**

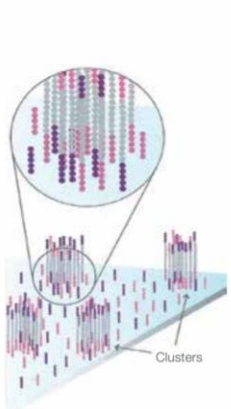
Illumina Sequencing

Figure 6: Denature the Double-Standed Molecules



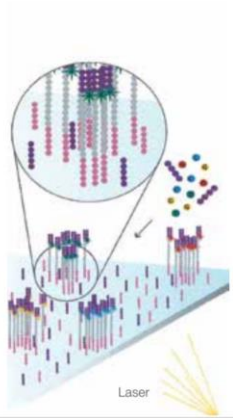
Denaturation leaves single-stranded templates anchored to the substrate.

Figure 7: Complete Amplification



Several million dense clusters of double-stranded DNA are generated in each channel of the flow cell.

Figure 8: Determine First Base



The first sequencing cycle begins by adding four labeled reversible terminators, primers, and DNA polymerase.

Figure 9: Image First Base

After laser excitation, the emitted fluorescence from each cluster is captured and the first base is identified.

Glenn, T. C. (2011). Field guide to next-generation DNA sequencers. *Molecular ecology resources*, *11*(5), 759-769.

Next-Gen Sequencing Platforms

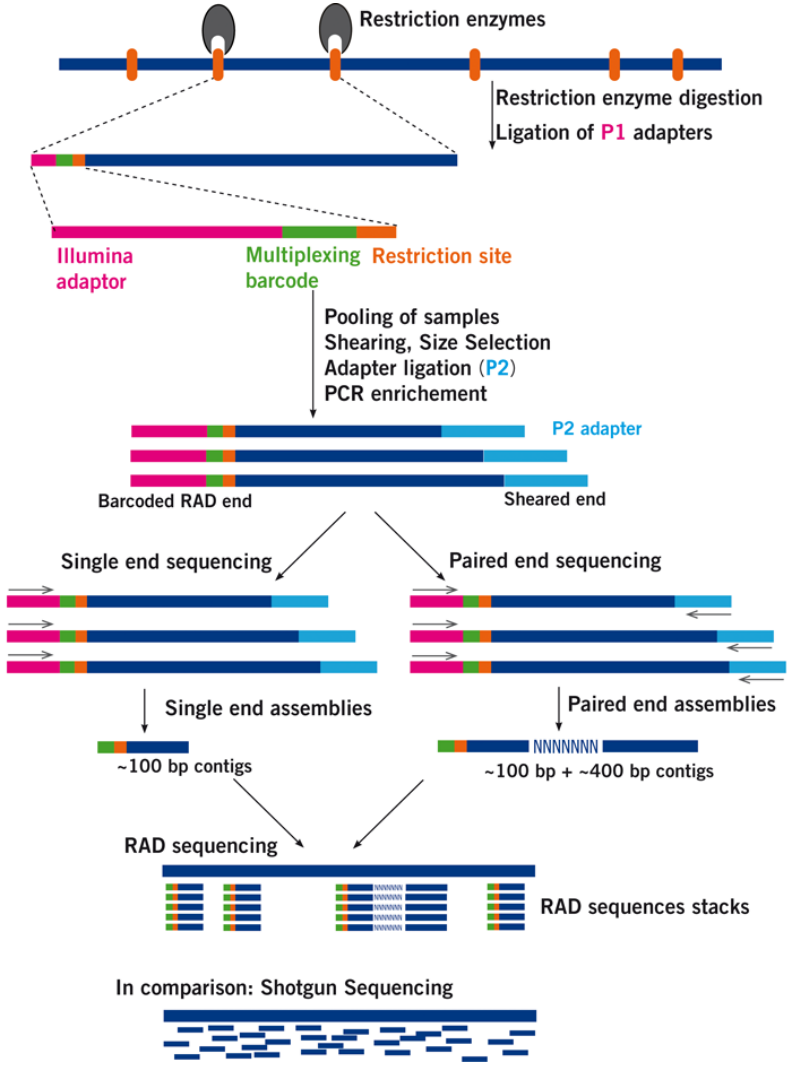
Now the genome is sampled, organized, and ready to **sequence**

Table 2 Comparison of sequencing instruments, sorted by cost/Mb, with expected performance by mid 2011

Instrument	Run time[a]	Millions of reads/run	Bases/read[b]	Yield Mb/run	Reagent cost/run[c]	Reagent cost/Mb	Minimum unit cost (% run)[d]
3730xl (capillary)	2 h	0.000096	650	0.06	$96	$1500	$6 (1%)
Ion Torrent – '314'chip	2 h	0.10	100	>10	$500	<$50	~$750 (100%)
454 GS Jr. Titanium	10 h	0.10	400	50	$1100	$22	$1500 (100%)
Starlight*	†	~0.01	>1000	†	†	†	†
PacBio RS	0.5–2 h	0.01	860–1100	5–10	$110–900	$11–180	†
454 FLX Titanium	10 h	1	400	500	$6200	$12.4	$2000 (10%)
454 FLX+[e]	18–20 h	1	700	900	$6200	$7	$2000 (10%)
Ion Torrent – '316'chip*	2 h	1	>100	>100	$750	<$7.5	~$1000 (100%)
Helicos[f]	N/A	800	35	28 000	N/A	NA	$1100 (2%)
Ion Torrent – '318'chip*	2 h	4–8	>100	>1000	~$925	~$0.93	~$1200 (100%)
Illumina MiSeq*	26 h	3.4	150 + 150	1020	$750	$0.74	~$1000 (100%)
Illumina iScanSQ	8 days	250	100 + 100	50 000	$10 220	$0.20	$3000 (14%)
Illumina GAIIx	14 days	320	150 + 150	96 000	$11 524	$0.12	$3200 (14%)
SOLiD – 4	12 days	>840[g]	50 + 35	71 400	$8128	<$0.11	$2500 (12%)
Illumina HiSeq 1000	8 days	500	100 + 100	100 000	$10 220	$0.10	$3000 (12%)
Illumina HiSeq 2000	8 days	1000	100 + 100	200 000	$20 120[h]	$0.10	$3000 (6%)
SOLiD – 5500 (PI)*	8 days	>700[g]	75 + 35	77 000	$6101	<$0.08	$2000 (12%)
SOLiD – 5500xl (4hq)*	8 days	>1410[g]	75 + 35	155 100	$10 503[h]	<$0.07	$2000 (12%)
Illumina HiSeq 2000 – v3[i]*	10 days	≤3000	100 + 100	≤600 000	$23 470[h]	≥$0.04	~$3500 (6%)

Glenn, T. C. (2011). Field guide to next-generation DNA sequencers. *Molecular ecology resources*, *11*(5), 759-769.

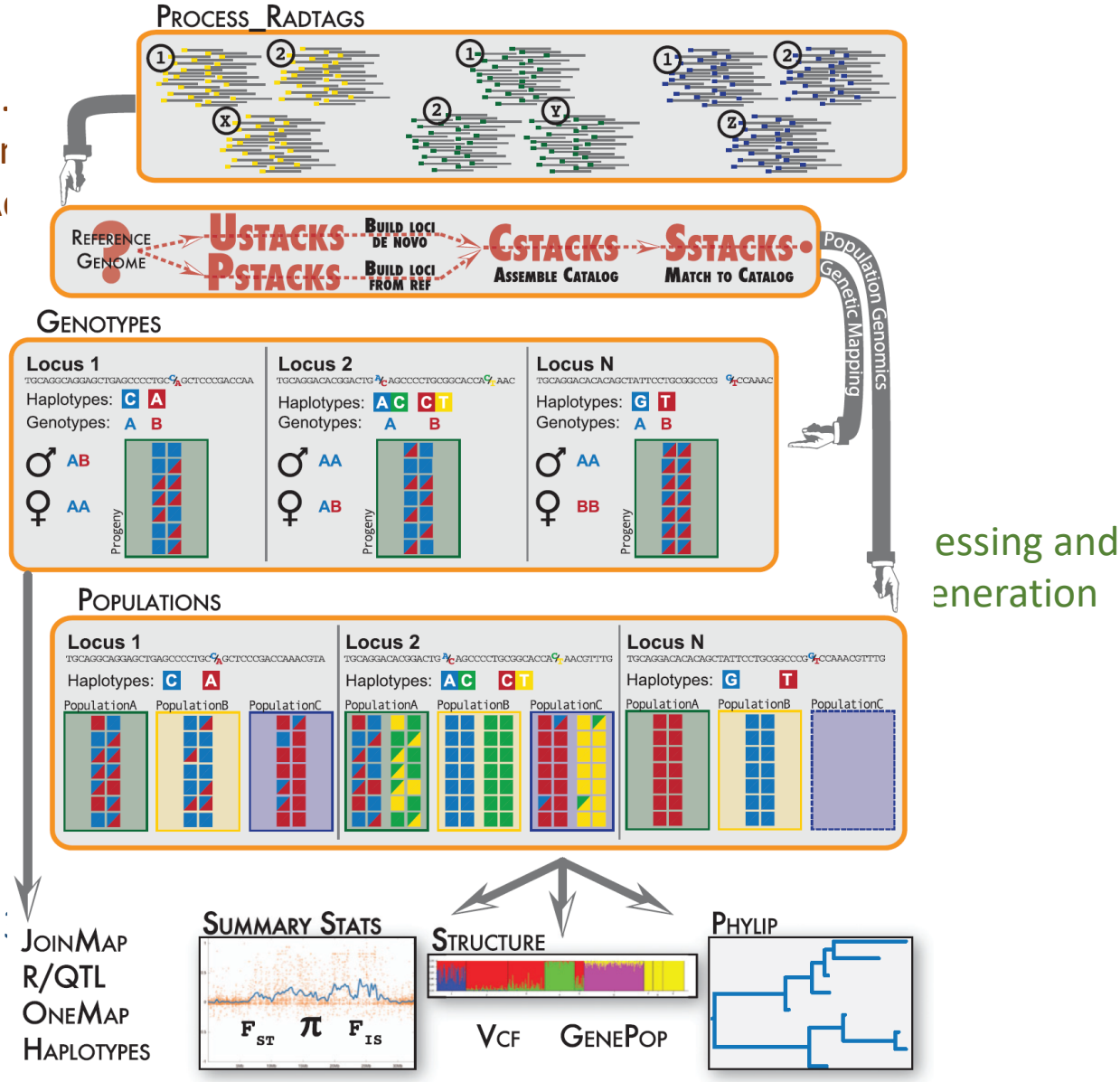
Alright so where are we...



http://www.floragenex.com/rad-seq/

And now we are ready for our pipeline…

1.
Sequer
A



essing and
eneration

The STACKS pipeline breakdown

Stacks is a software pipeline for building loci from short-read sequences, such as those generated on the Illumina platform. Stacks was developed to work with restriction enzyme-based data, such as RAD-seq, for the purpose of building genetic maps and conducting population genomics and phylogeography.

Raw Reads

- process_radtags **1**
- process_shortreads
- clone_filter
- kmer_filter

Core

- ustacks
- pstacks
- cstacks
- sstacks
- genotypes
- populations **3**
- rxstacks

Execution control

- denovo_map.pl **2**
- ref_map.pl
- load_radtags.pl

Utilities

- index_radtags.pl
- export_sql.pl
- sort_read_pairs.pl
- exec_velvet.pl

Catchen, J., Hohenlohe, P. A., Bassham, S., Amores, A., & Cresko, W. A. (2013). Stacks: an analysis tool set for population genomics. *Molecular ecology*, *22*(11), 3124-3140.

The STACKS pipeline breakdown

Some more quick terminology…

Stack

A series of identical reads from a given sample that have been anchored to a homologous location in the genome (reference genome or genome assembled de novo) and arranged thus to form a ‘stack’; a series of stacks is used to form a loci.

Loci

The location of a gene or mutation along the genome

Allele

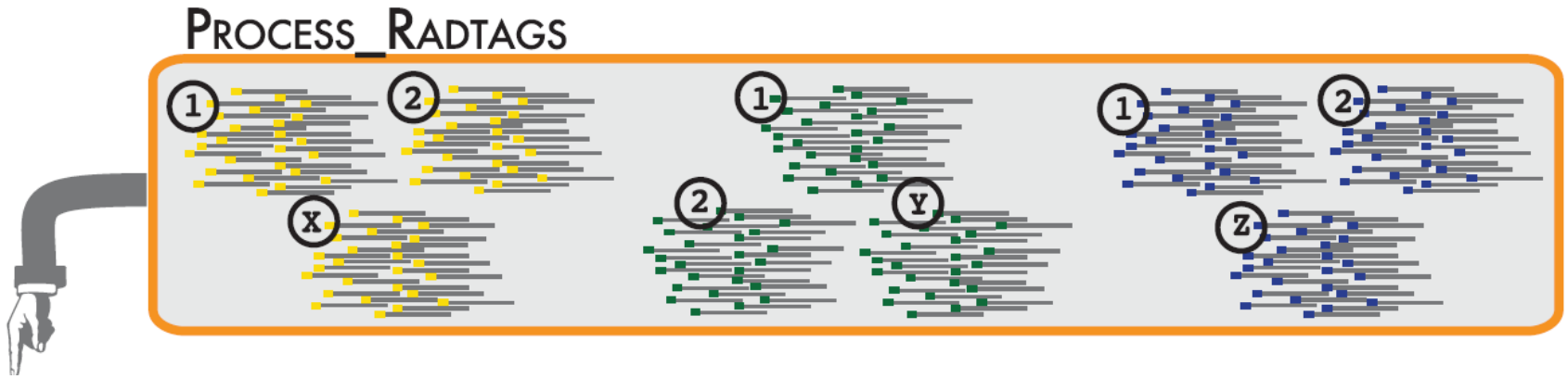
The particular form of a gene or mutation found at a given loci along the genome

STACKS pipeline 1: process_radtags

process_radtags

A function to de-multiplex and clean raw sequence ready

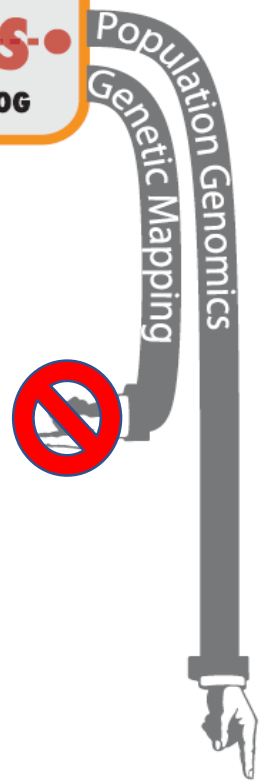
```
process_radtags -f XXXXX.fastq.gz -i gzfastq -o ./demultiplexed_fastqs/ -b
    XXXXX.txt -e nlaIII -c -q -r
```

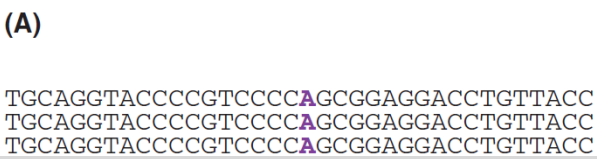


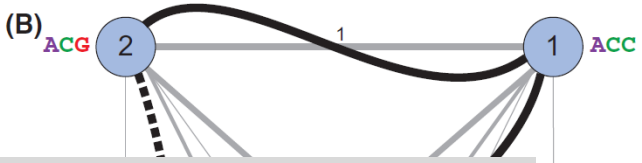
STACKS pipeline 2: denovo_map.pl

Denovo_map.pl

A wrapper function that runs all three main programs (ustacks, cstacks, and sstacks) comprised within the *de novo* loci assembly version of the stacks pipeline

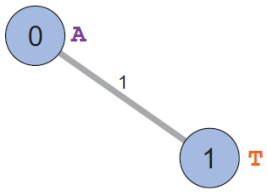


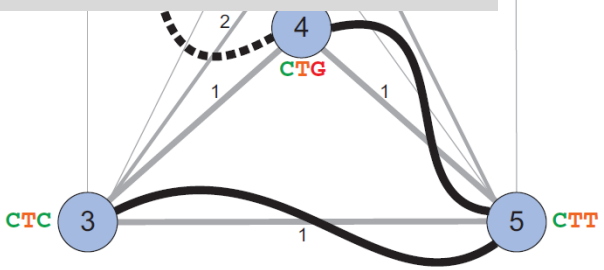




ustacks

Data from each individual are grouped into loci and polymorphic sites are identified

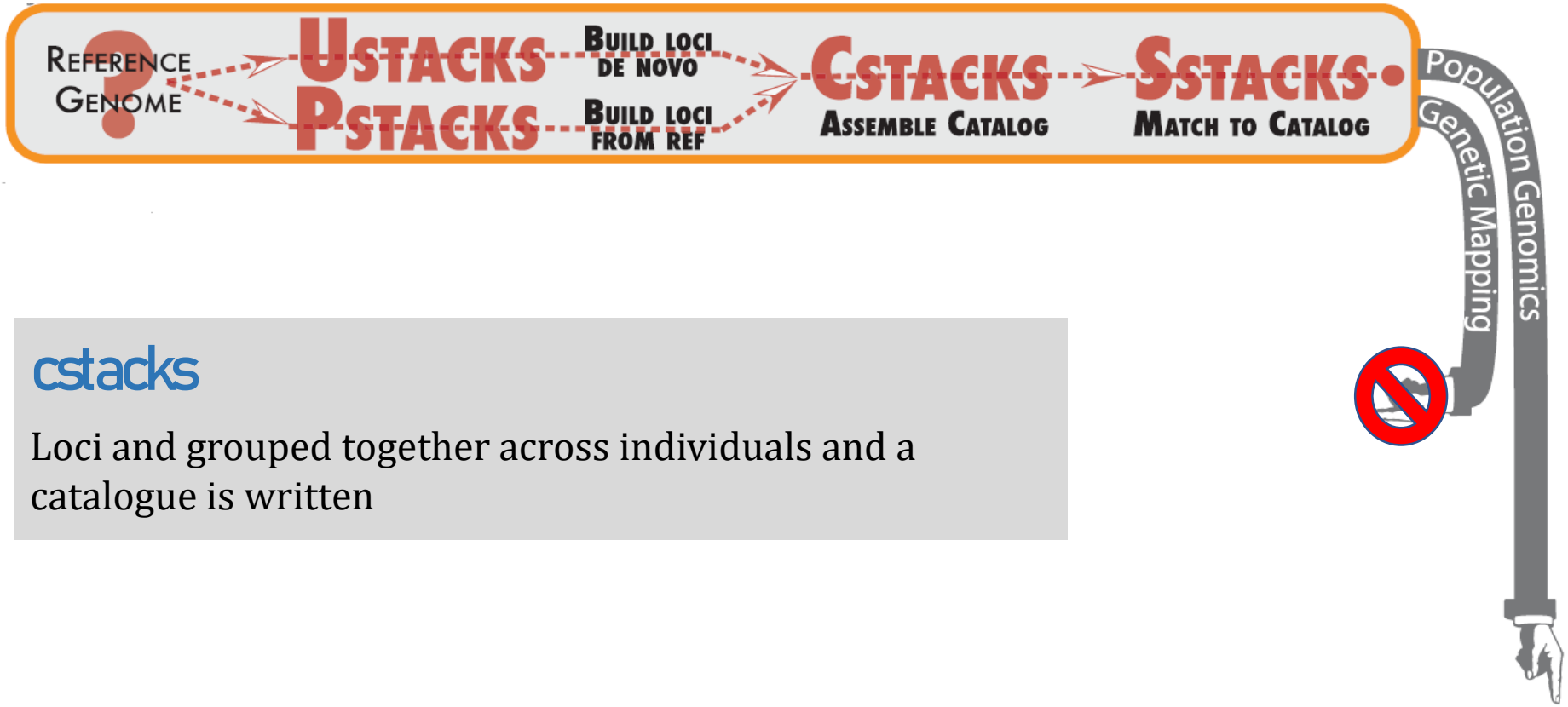




STACKS pipeline 2: denovo_map.pl

Denovo_map.pl

A wrapper function that runs all three main programs (ustacks, cstacks, and sstacks) comprised within the *de novo* loci assembly version of the stacks pipeline



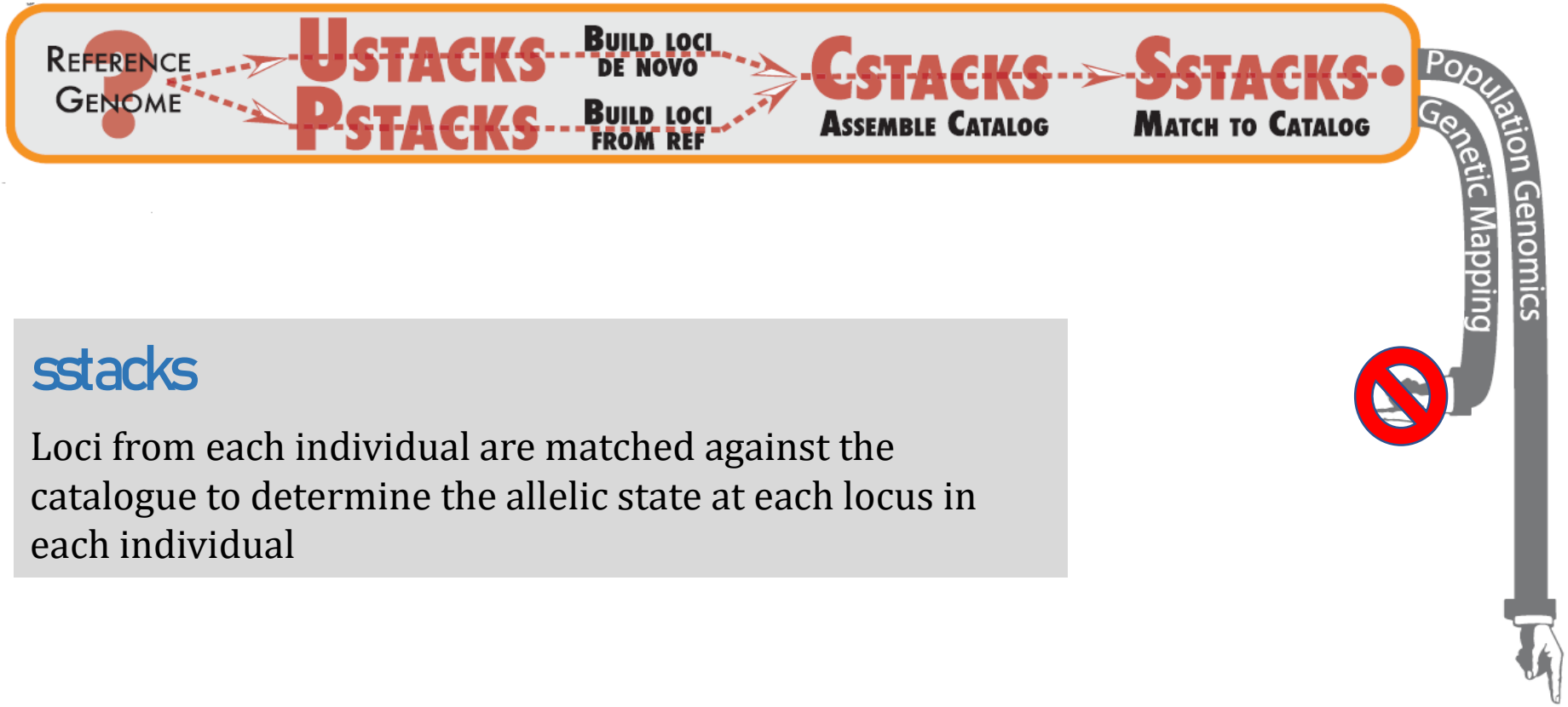
cstacks

Loci and grouped together across individuals and a catalogue is written

STACKS pipeline 2: denovo_map.pl

Denovo_map.pl

A wrapper function that runs all three main programs (ustacks, cstacks, and sstacks) comprised within the *de novo* loci assembly version of the stacks pipeline



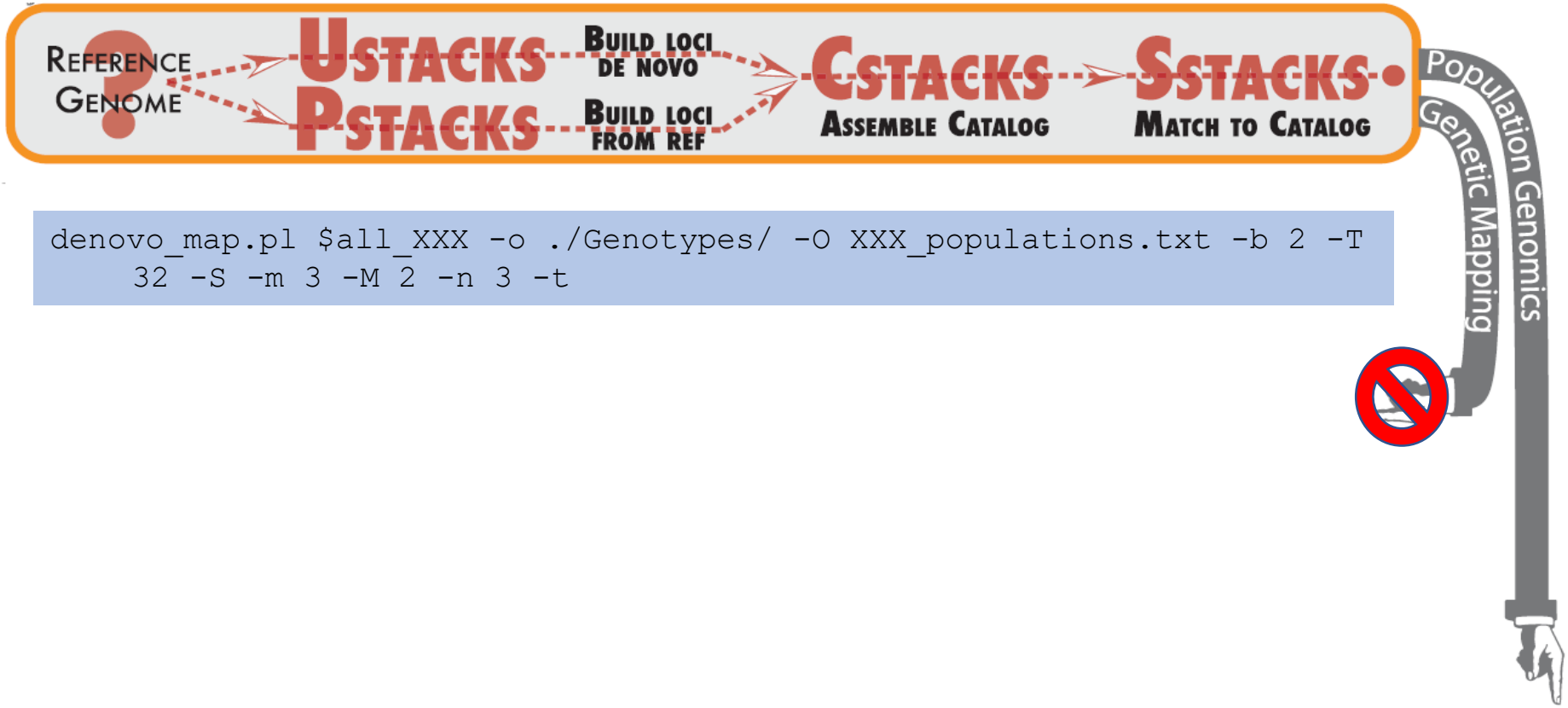
sstacks

Loci from each individual are matched against the catalogue to determine the allelic state at each locus in each individual

STACKS pipeline 2: denovo_map.pl

Denovo_map.pl

A wrapper function that runs all three main programs (ustacks, cstacks, and sstacks) comprised within the *de novo* loci assembly version of the stacks pipeline



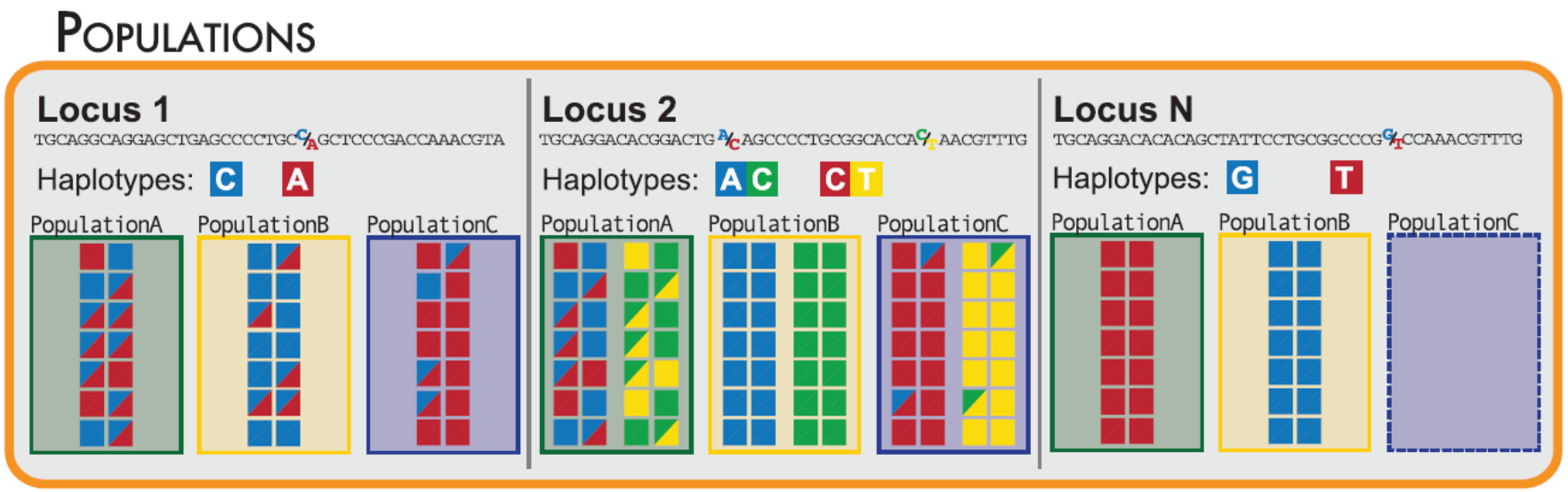

```
denovo_map.pl $all_XXX -o ./Genotypes/ -O XXX_populations.txt -b 2 -T
     32 -S -m 3 -M 2 -n 3 -t
```

STACKS pipeline 3: denovo_map.pl

populations

tabulates the state of loci within and among populations, calculates population genetics statistics and exports to a number of additional, useful formats

```
populations -b 2 -t 32 -P ./Genotypes/ -M ./XXX_populations.txt --renz nlaIII
    -m 10 -p 1 -r 0.5 -k --fstats -f p_value --genomic --ordered_export --
    write_random_snp --fasta --vcf --plink --genepop --structure --phylip
```



Ok so where are we now…

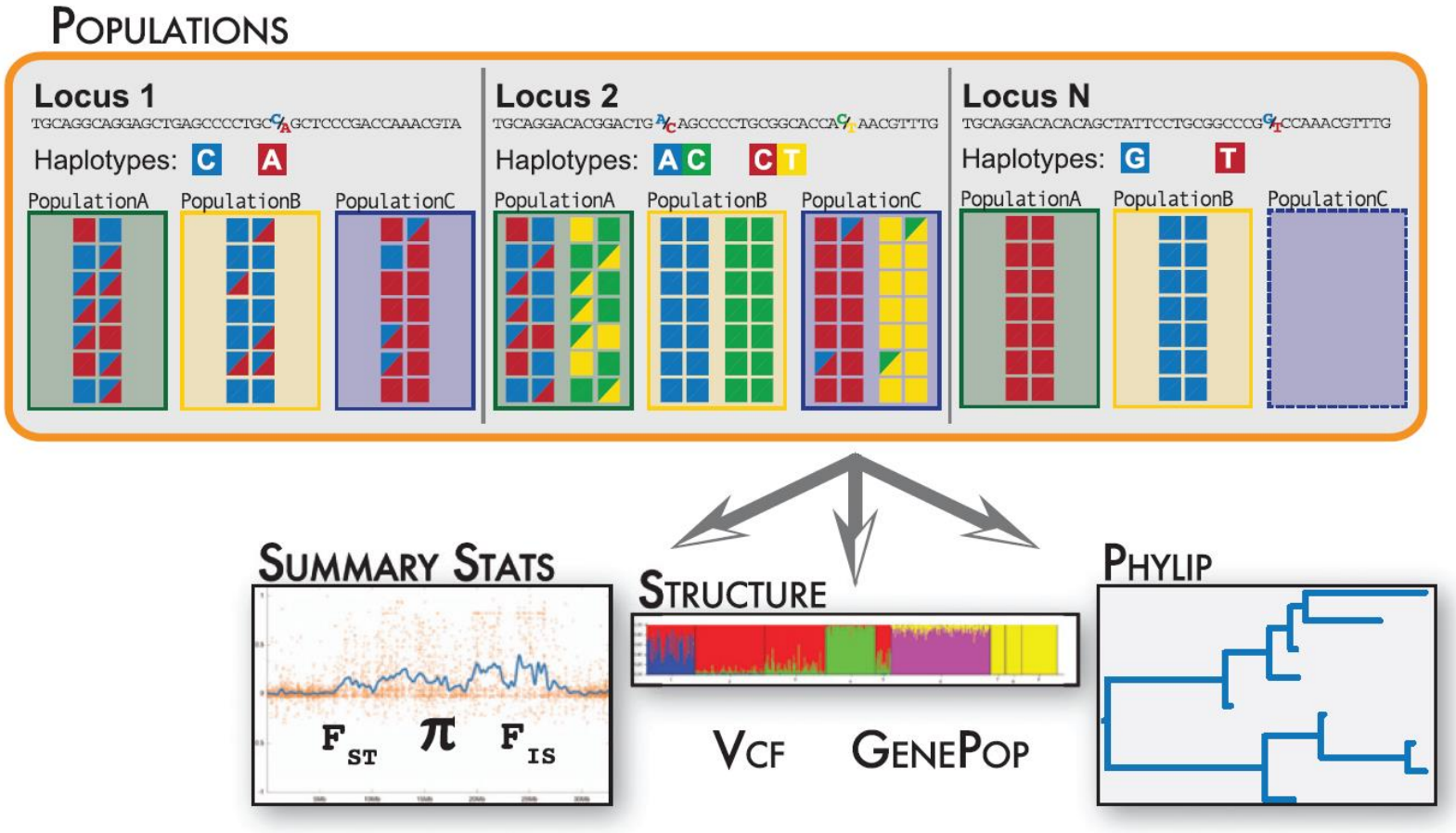
So we sent our raw sequence data through the pipeline and have…

Data Matrices With…

- **Series of SNP's From Across the Genome**
- **Unlinked**
- **Primarily Nuclear**
- **Coding and Noncoding**
- **Partially Filtered**
- **In VCF or other desired output format**

Ok so where are we now…

And now we can use this data to perform various analyses



NOTE: Methods are best for exploring recent evolutionary history (no more than ~10 mya)

Before that a few more tools

Some tools for additional filtering and formatting

Welcome to VCFtools

VCFtools is a program package designed for working with VCF files, such as those generated by the 1000 Genomes Project. The aim of VCFtools is to provide easily accessible methods for working with complex genetic variation data in the form of VCF files.

This toolset can be used to perform the following operations on VCF files:

- Filter out specific variants
- Compare files
- Summarize variants
- Convert to different file types
- Validate and merge files
- Create intersections and subsets of variants

PGDSpider version 2.1.1.2 (July 2017)

An automated data conversion tool for connecting population genetics and genomics programs

Ok now, on to the analyses...

And by way of example, here is a recent paper

ORIGINAL ARTICLE

WILEY MOLECULAR ECOLOGY

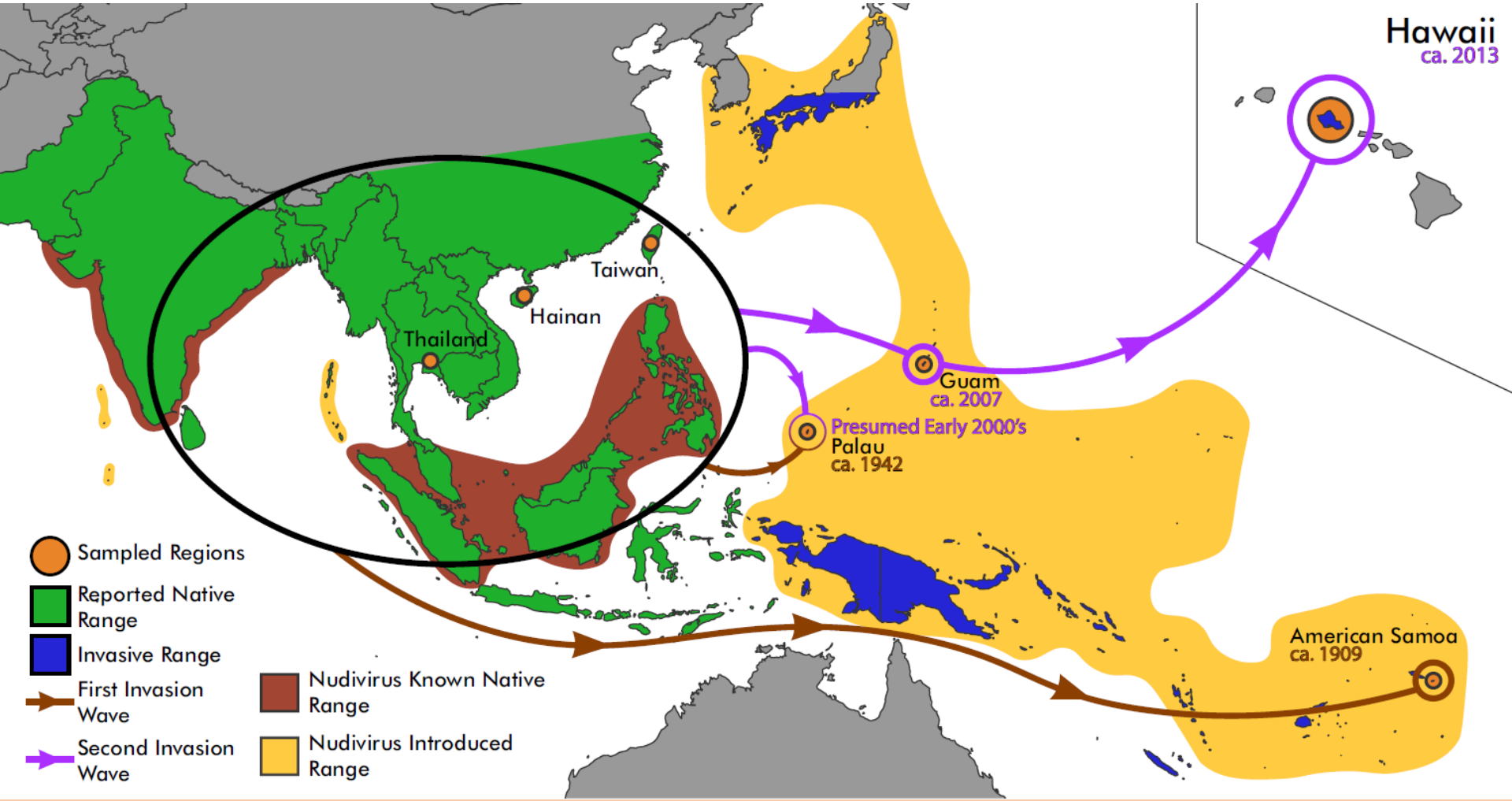
Transpacific coalescent pathways of coconut rhinoceros beetle biotypes: Resistance to biological control catalyses resurgence of an old pest

Jonathan Bradley Reil[1] | Camiel Doorenweerd[1] | Michael San Jose[1] | Sheina B. Sim[1,2] | Scott M. Geib[2] | Daniel Rubinoff[1]

Reil, J. B., Doorenweerd, C., San Jose, M., Sim, S. B., Geib, S. M., & Rubinoff, D. (2018). Transpacific coalescent pathways of coconut rhinoceros beetle biotypes: Resistance to biological control catalyses resurgence of an old pest. *Molecular ecology*, *27*(22), 4459-4474.

A working example: *Oryctes rhinoceros*

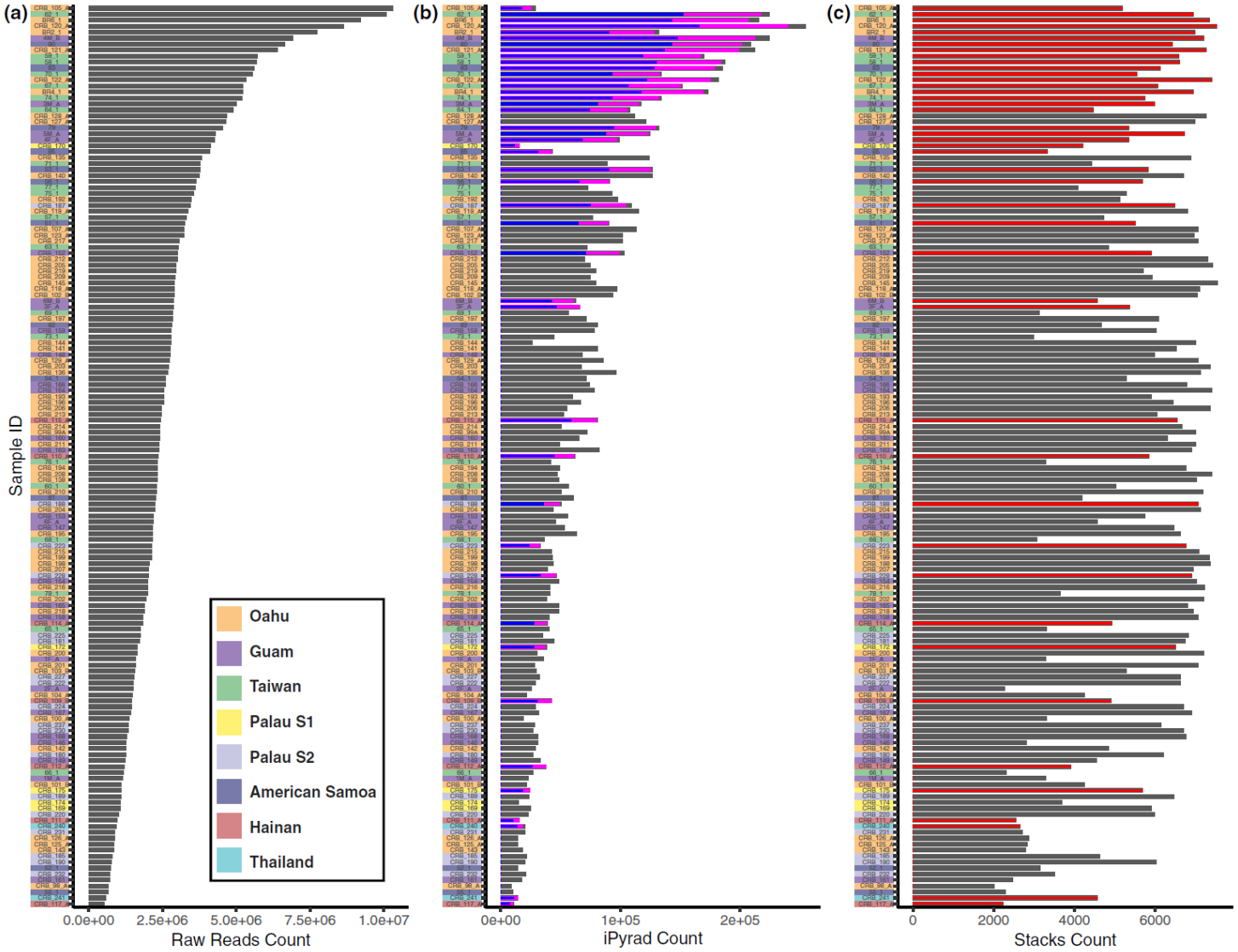
So what is the story…



Reil, J. B., Doorenweerd, C., San Jose, M., Sim, S. B., Geib, S. M., & Rubinoff, D. (2018). Transpacific coalescent pathways of coconut rhinoceros beetle biotypes: Resistance to biological control catalyses resurgence of an old pest. *Molecular ecology*, *27*(22), 4459-4474.

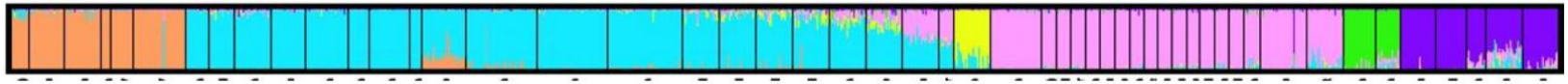
A working example: *Oryctes rhinoceros*

Data obtained from 172 individuals (151 post quality control) using ddRAD and Illumina Seq.



A working example: *Oryctes rhinoceros*

Began our analysis with some population genetics using STRUCTURE

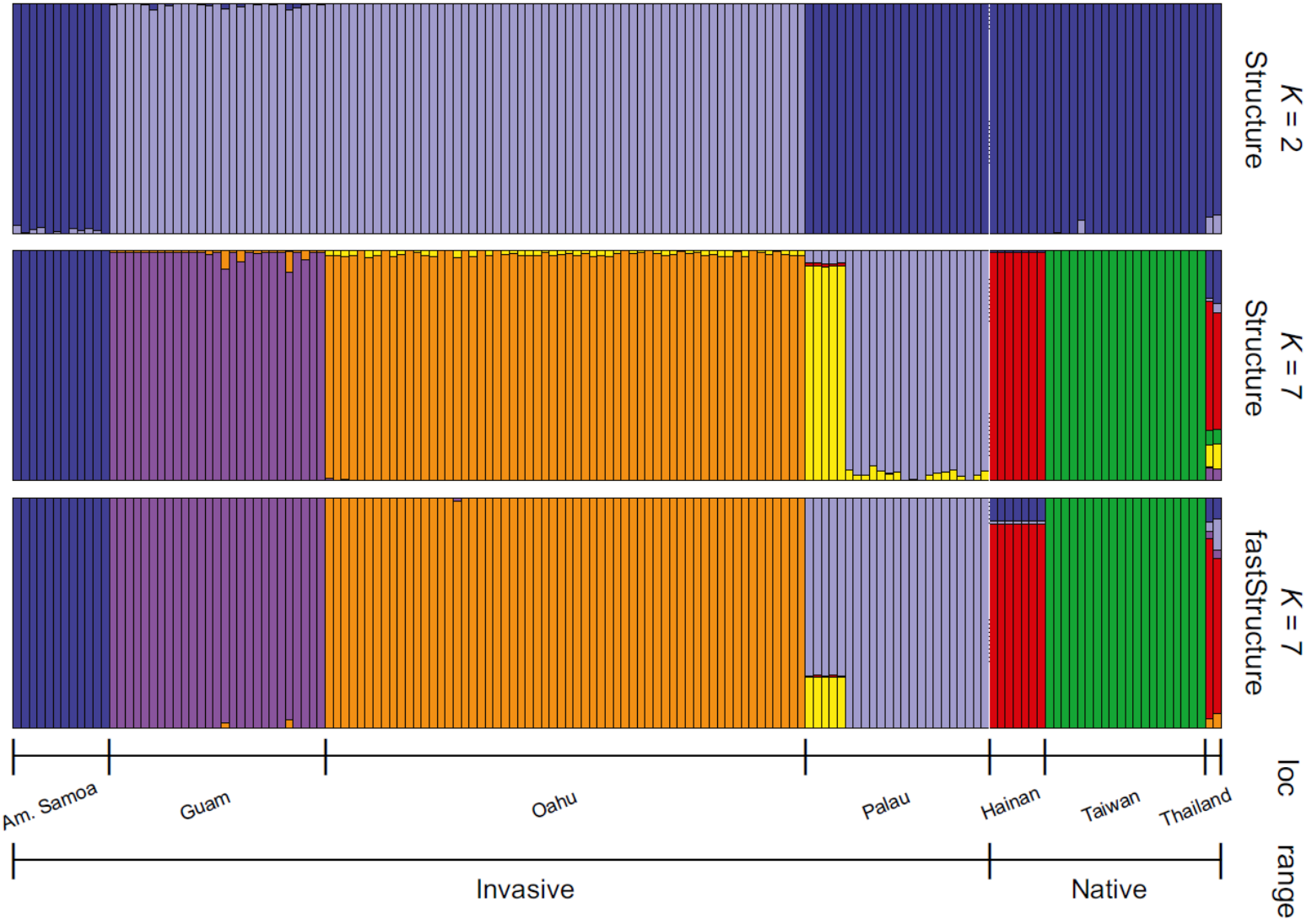


STRUCTURE analyses differences in the distribution of genetic variants amongst populations with a Bayesian iterative algorithm by placing samples into groups whose members share similar patterns of variation. *STRUCTURE* both identifies populations from the data and assigns individuals to that population representing the best fit for the variation patterns found. Typically *STRUCTURE* is the first step in examining population structures that emerge from the sample set to provide a preamble to further genetic analysis or to infer the origins of individuals with unknown population characteristics, especially when population admixture has occurred. As *STRUCTURE* uses the core Bayesian principle of comparing likelihoods, prior information about study samples can be supplied to further shape the analysis.

Porras-Hurtado, L., Ruiz, Y., Santos, C., Phillips, C., Carracedo, Á., & Lareu, M. (2013). An overview of STRUCTURE: applications, parameter settings, and supporting software. *Frontiers in genetics*, *4*, 98.

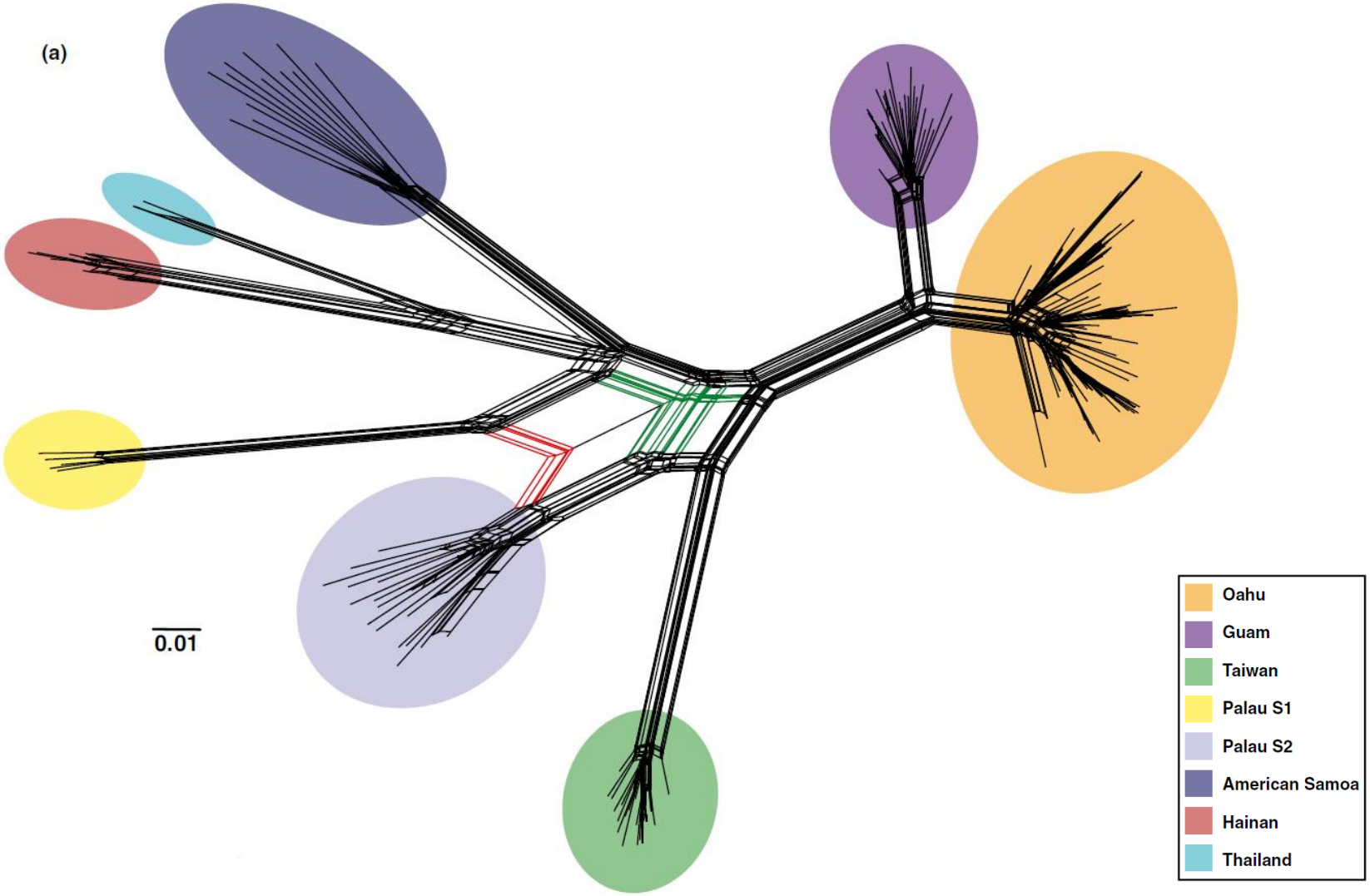
A working example: *Oryctes rhinoceros*

Began our analysis with some population genetics using STRUCTURE



A working example: *Oryctes rhinoceros*

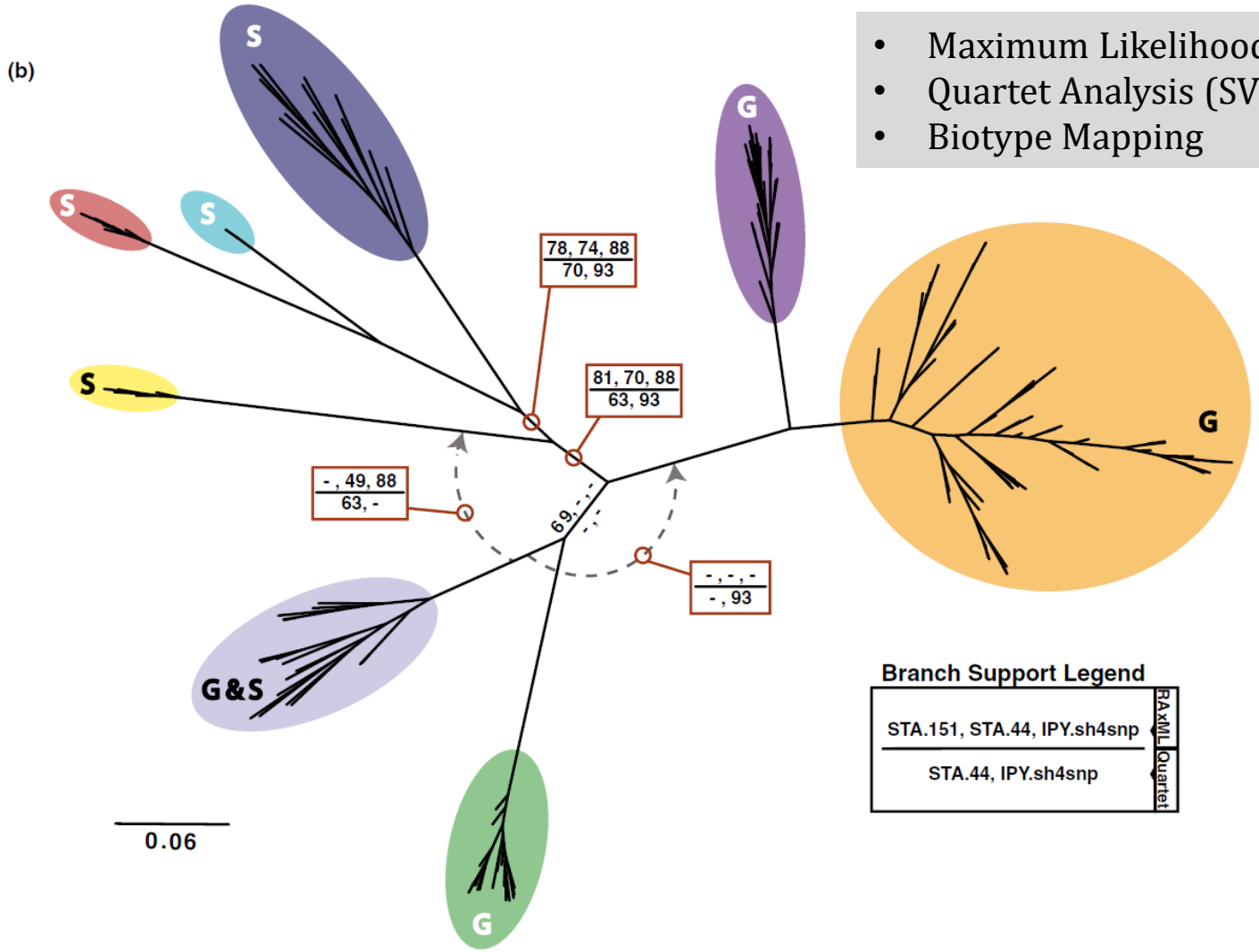
We then proceeded to begin working out Phylogenetic relationships



A working example: *Oryctes rhinoceros*

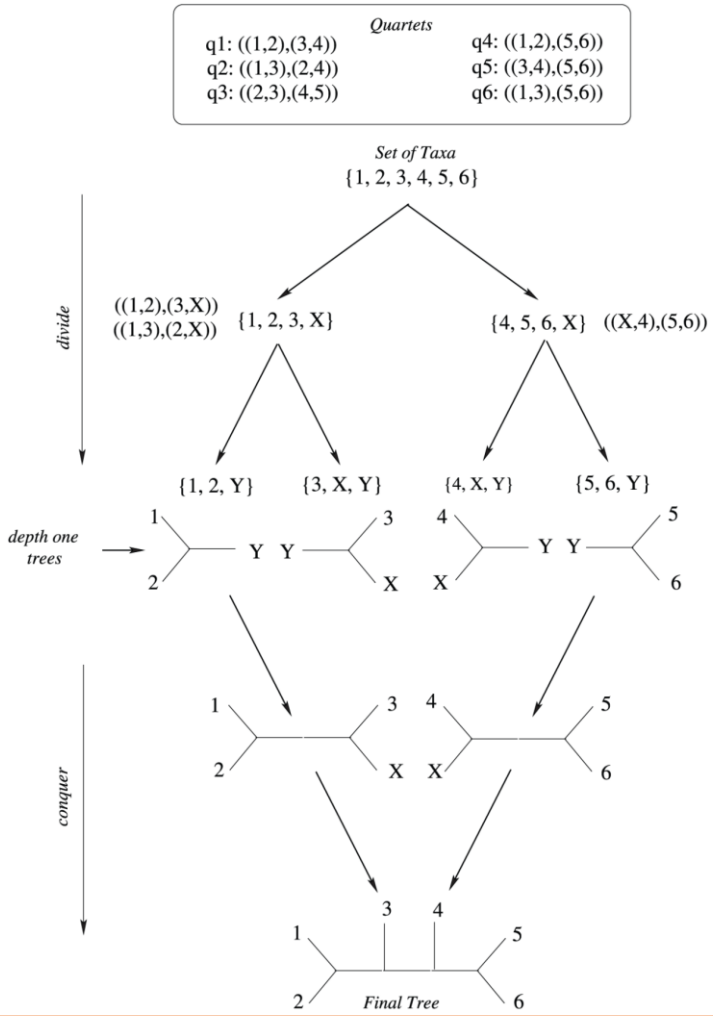
Ok something interesting is going on, lets dig deeper

- Maximum Likelihood (RAxML)
- Quartet Analysis (SVDQuartets)
- Biotype Mapping



A working example: *Oryctes rhinoceros*

A bit on SVDQuartets and quartet analysis…

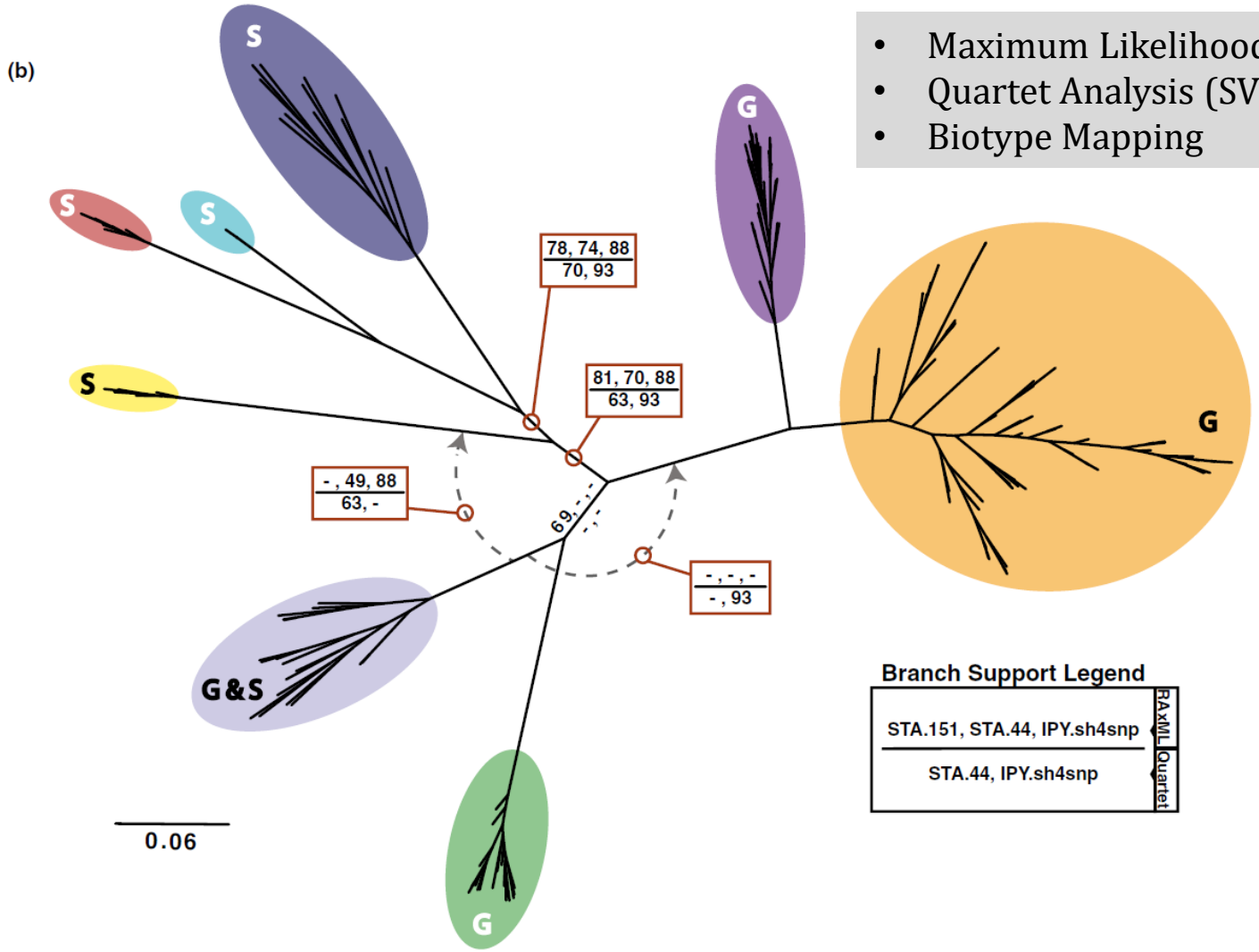


Eaton, D. A., Spriggs, E. L., Park, B., & Donoghue, M. J. (2017). Misconceptions on missing data in RAD-seq phylogenetics with a deep-scale example from flowering plants. *Systematic Biology*, *66*(3), 399-412.

A working example: *Oryctes rhinoceros*

Ok something interesting is going on, lets dig deeper

- Maximum Likelihood (RAxML)
- Quartet Analysis (SVDQuartets)
- Biotype Mapping

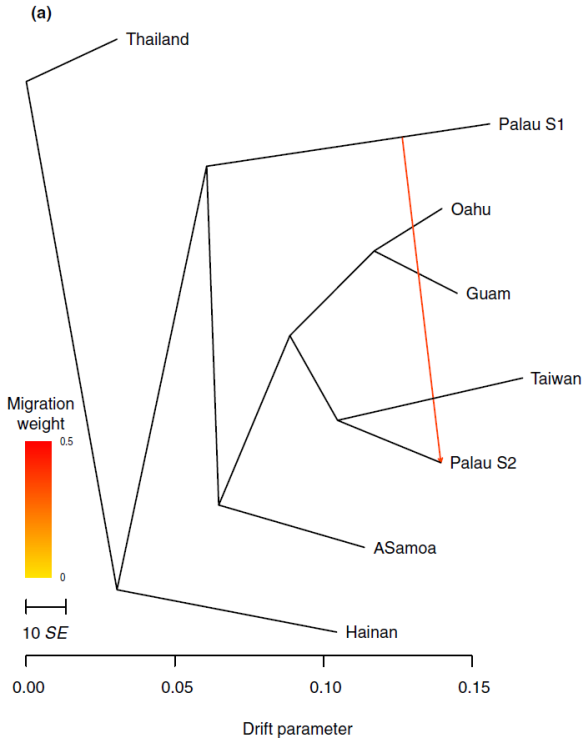


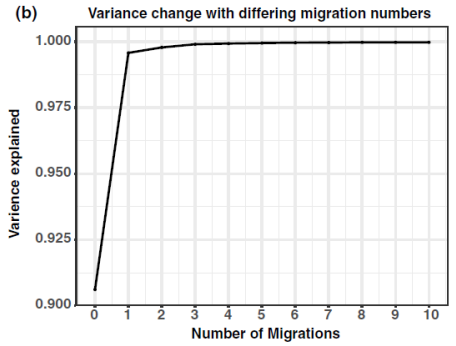
A working example: *Oryctes rhinoceros*

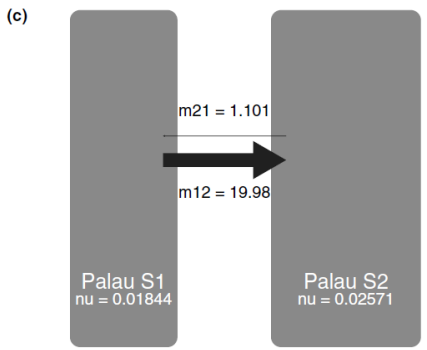
And so we concluded..

Conclusions

- Second Invasion catalyzed by the emergency of biotype G
- Both biotypes present in Palau, leading to gene tree variation
- Biotypes exchanging genetic material
- Beetles in Hawaii, likely originated in Guam
- Biotype G origin unknown, but Taiwan is suspect







So, what does it really take to do this

Who can reasonably take advantage of these techniques?

```
###STACKS PIPELINE SCRIPT (flags varied for different data matrices)####
##STEP 1: Process Radtags (De-multiplexing): Run this for each set of
##fastq.gz and combine the outputs into a single directory (Ex.
##demultiplexed_all)
#process_radtags -f XXXXX.fastq.gz -i gzfastq -o ./demultiplexed_fastqs/ -b
    XXXXX.txt -e nlaIII -c -q -r

##STEP 2: Pass this portion with denovo_map.pl: It parses your
##individuals.txt (file with sample ID's) and files in your combined directory
##into a list usable by denovo_map.pl
#all_XXX=""
#for x in `cat XXX_individuals.txt`;
#do
#all_XXX+="-s ./demultiplexed_all/$x.fq.gz "
#done

## Denovo assembly of loci, catalog generation, SNP flagging
#denovo_map.pl $all_XXX -o ./Genotypes/ -O XXX_populations.txt -b 2 -T 32 -S -
    m 3 -M 2 -n 3 -t

##STEP 3: Optional things like SNP/Loci filtering, basic pop. gen. stats, etc
#populations -b 2 -t 32 -P ./Genotypes/ -M ./XXX_populations.txt --renz nlaIII
    -m 10 -p 1 -r 0.5 -k --fstats -f p_value --genomic --ordered_export --
    write_random_snp --fasta --vcf --plink --genepop --structure --phylip
```