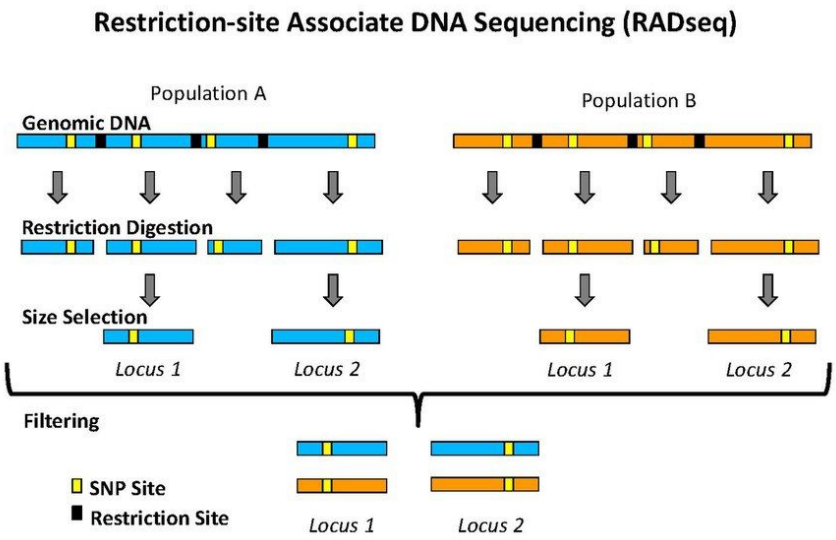
RadSeq vs. Genome Skimming: Variant Matrix Construction

Jennifer Chien, Sigal Shaul, and Akshat Singhal

What is RAD-seq?

- Restriction site associated DNA sequencing
- Restriction enzymes fragment DNA, creating reads
- Targets a subset of a genome; provides better depth per locus



Motivation

RadSeq

- Pros:
  - High amplification for sections of the genome that contain restriction site (more depth)
  - Alignment by **restriction site**
- Cons:
  - Less coverage - Only captures region around restriction sites

Genome Skimming

- Pros:
  - More **coverage** across the genome
  - More variant loci (more likely to capture mutations)
- Cons:
  - Brute force approach is computationally heavy

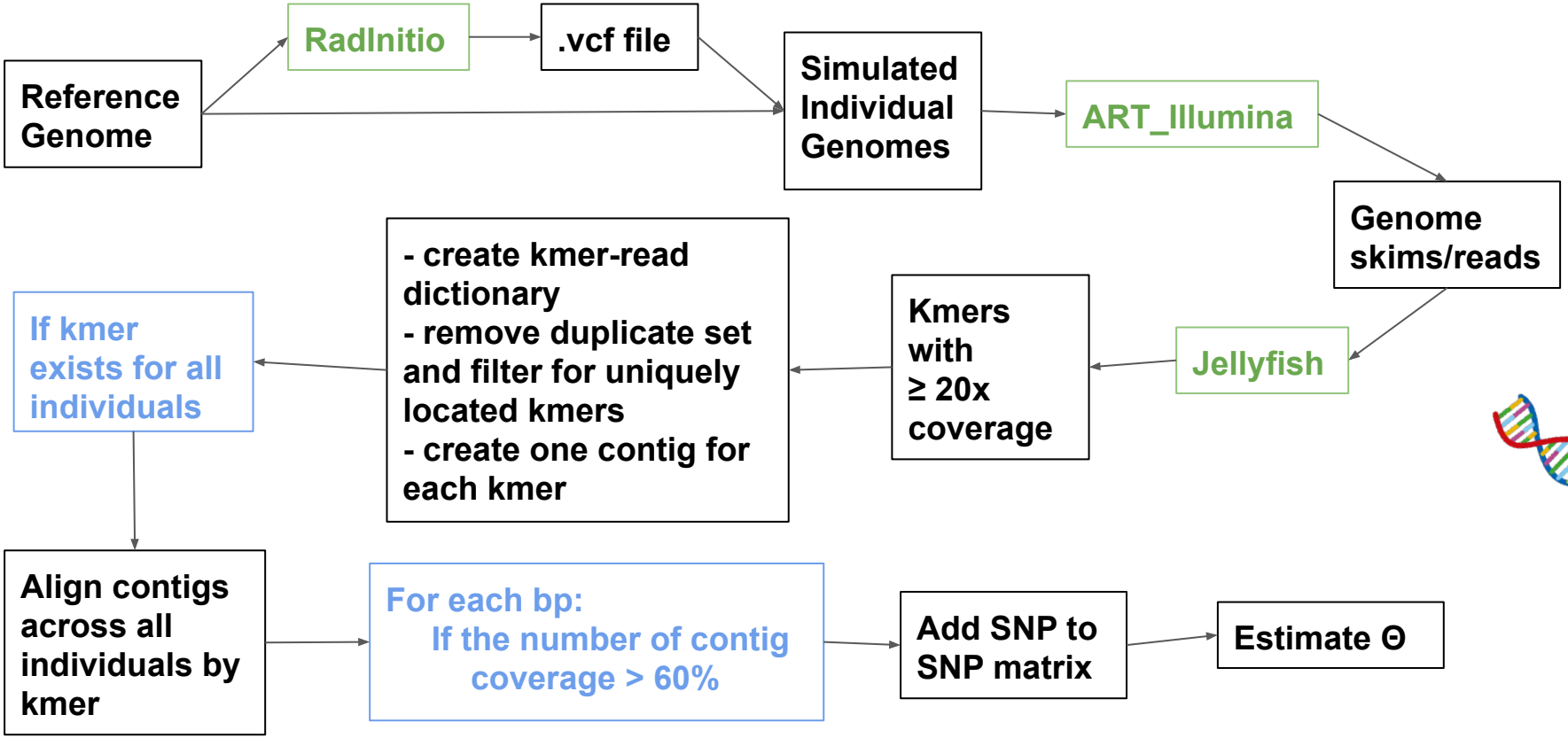
Simulated RAD-Seq Experiment

- C. Elegans Genome (100Mbp)
- RADInitio - Using reference genome and msprime, creates simulated population of individuals with mutations;
  returns a vcf file, restriction site loci files, and read files
- Directly use restriction site files to estimate θ - The files contain restriction site loci; each loci is 1000bp long and there are 634 such loci

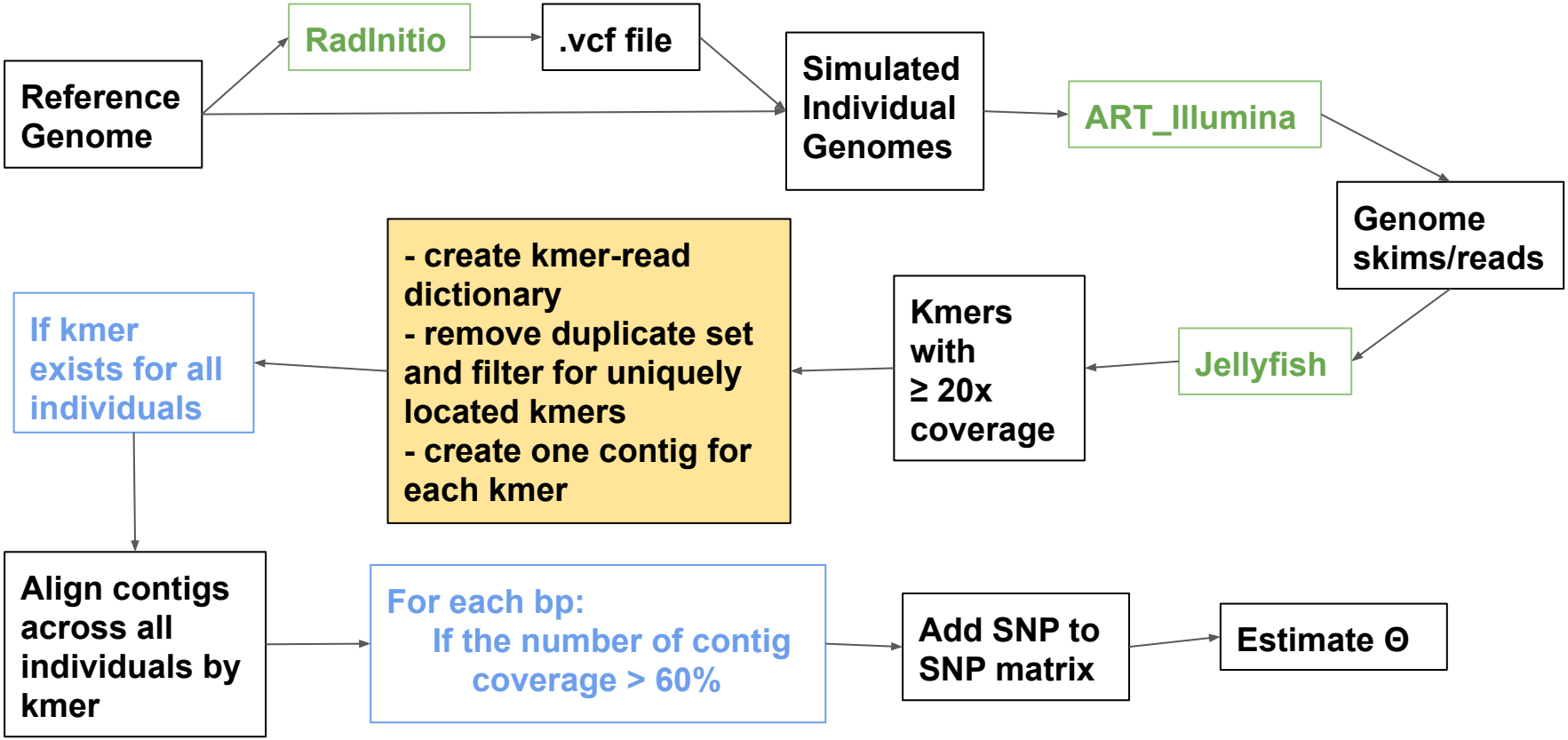
Proposed Method Using Genome Skimming

1. **Take a reference data** - Chromosome 1 of C. elegans
2. **Simulate a population** - 20 sequences (10 diploid individuals) created using RADinitio variant simulation
3. **Generate skims** - Reads generated using art-illumina with 5x coverage
4. **Get the k-mers for each skim** - Used jellyfish to get the common k-mers with at least 20x frequency (n)
5. ***Compare the skims** - Align the reads for each k-mer to form a SNP matrix
6. ***Provide an estimate of θ** - Watterson's estimate or Tajima's estimate

Our Pipeline



Our Pipeline



5. Comparing the Skims

kmers
AAAA
...

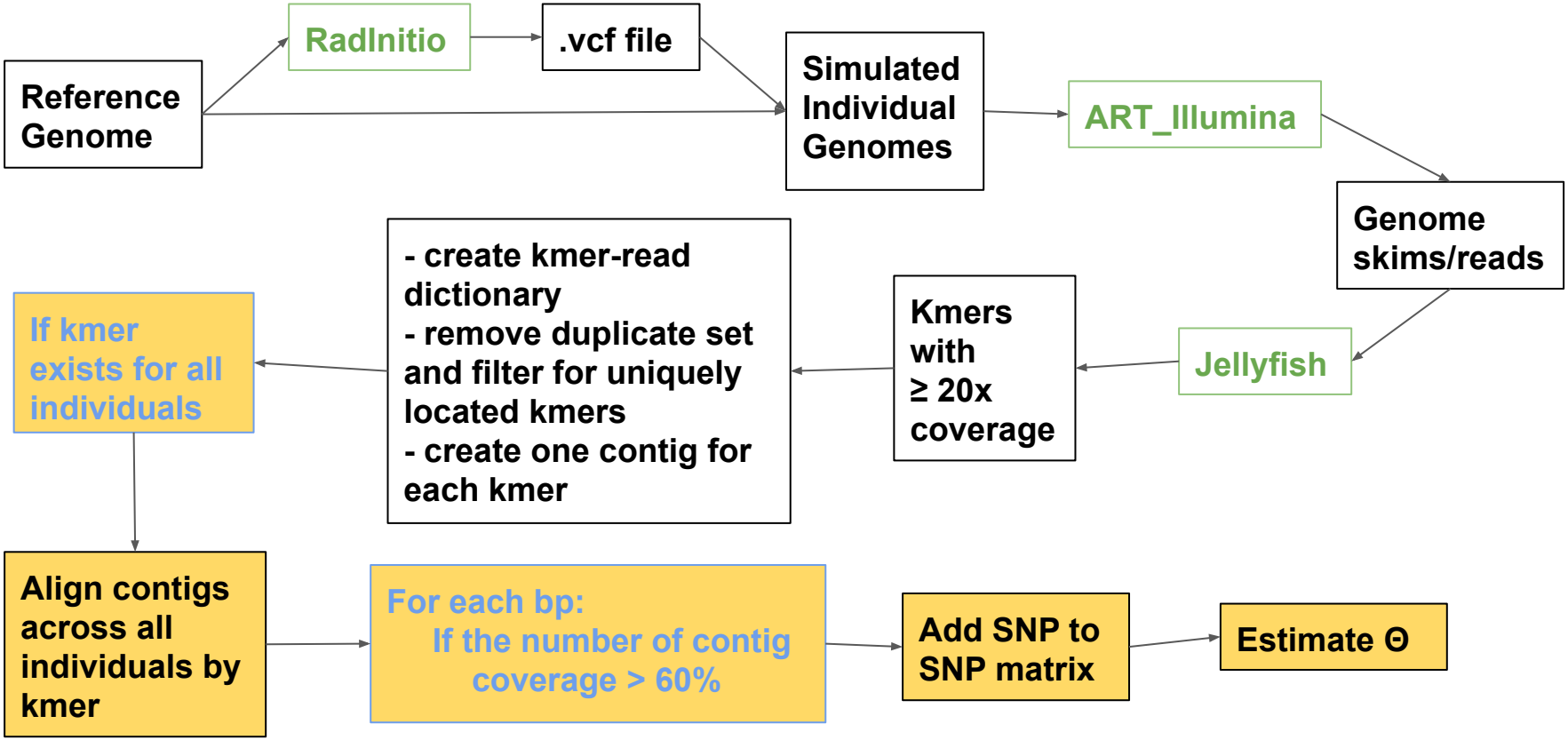
Reads for Indiv_0
ACGGAAAATCCA
...

kmersDict (kmer → read index)
AAAA [0, 2, 10]
…
GCTA [0, 1, 3, 4, 5, ...]

0 ACGAAAATCCA
2 GTTAACGAAAA
10 CGAAAATCCACT

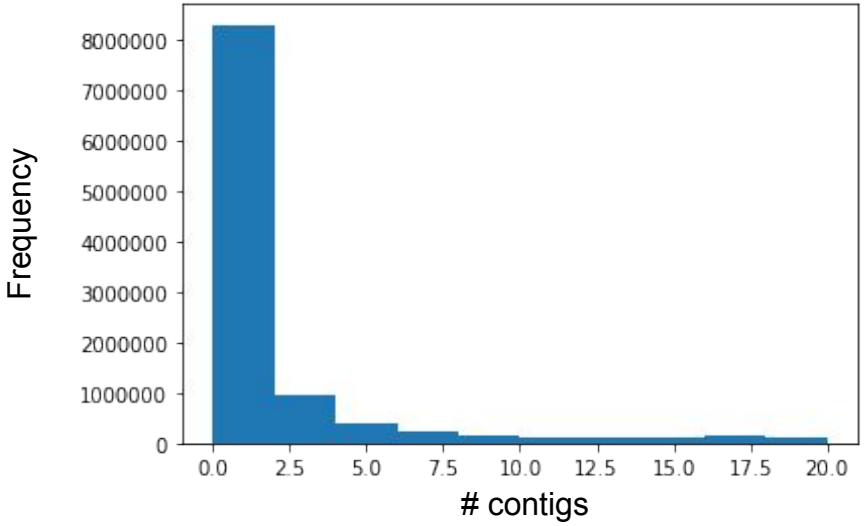
kmerDict (kmer → contig)
AAAA GTTAACGAAAATCCACT
...

Our Pipeline



Consider kmers in all Individuals

1. Total number of kmers reported: 10599778
2. Limited to kmers that are present in all individuals



Creating SNP Matrix and Calculating Theta

1. Align contigs for each kmer

kmer: AAAAAAAA

TAAAAAAAAAATTGAAAAA
GCCTTAAAAAAAATCGTTT
GGGAAAAAAAATAGATTG
CCCCGGAAAAAAAATCGT
TTAAAAAAAATC

-	-	-	-	-	T	A	A	A	A	A	A	A	A	A	A	T	T	G	A	A	A	A	A
-	G	C	C	T	T	A	A	A	A	A	A	A	A	T	C	G	T	T	T	-	-	-	-
-	-	-	G	G	G	A	A	A	A	A	A	A	A	T	A	G	A	T	T	G	-	-	-
C	C	C	C	G	G	A	A	A	A	A	A	A	A	T	C	G	T	-	-	-	-	-	-
-	-	-	-	T	T	A	A	A	A	A	A	A	A	T	C	-	-	-	-	-	-	-	-

Creating SNP Matrix and Calculating Theta

1. Align contigs for each kmer
2. Convert each site to SNP column
   a. Set initial allele to 0, mutated alleles to 1+

kmer: AAAAAAAA

-	-	-	-	-	T	A	A	T	T	G	A	A	A	A	A
-	G	C	C	T	T	T	C	G	T	T	T	-	-	-	-
-	-	-	G	G	G	T	A	G	A	T	T	G	-	-	-
C	C	C	C	G	G	T	C	G	T	-	-	-	-	-	-
-	-	-	-	T	T	T	C	-	-	-	-	-	-	-	-

```
[[-1 -1 -1 -1 -1  0  0  0  0  0  0  0  0  0  0  0]
 [-1  0  0  0  0  0  1  1  1  1  1  1 -1 -1 -1 -1]
 [-1 -1 -1  1  1  1  1  0  1  0  1  1  1 -1 -1 -1]
 [ 0  1  0  0  1  1  1  1  1  1 -1 -1 -1 -1 -1 -1]
 [-1 -1 -1 -1  0  0  1  1 -1 -1 -1 -1 -1 -1 -1 -1]]
```

Creating SNP Matrix and Calculating Theta

Theta Estimate:

1. Count number of columns with mutations
   a. Only if number alleles present is greater than threshold
2. Estimate with m / ln(n)

$$
\begin{bmatrix}
-1 & -1 & -1 & -1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
-1 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \\
-1 & -1 & -1 & 1 & 1 & 1 & 1 & 0 & 1 & 0 & 1 & 1 & 1 & -1 & -1 & -1 \\
0 & 1 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 & -1 & -1 \\
-1 & -1 & -1 & -1 & 0 & 0 & 1 & 1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1
\end{bmatrix}
$$

Results

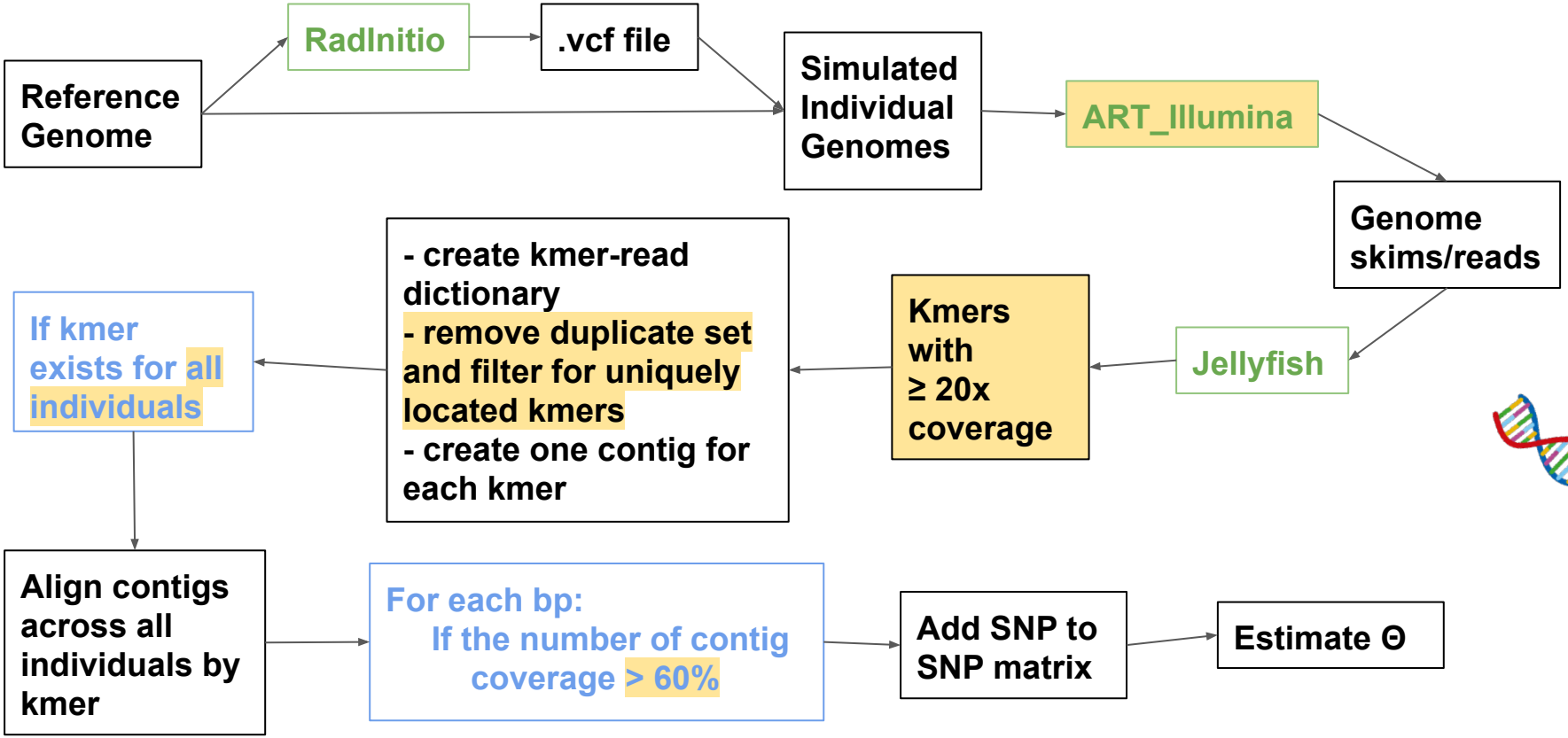
	M	θ
True SNPs	180,568	60275.1
RADSeq	100,000	63,485.7*
Our Implementation	177,696	59316.4

* Effective θ: RadInitio covers about 52.59% of the genome. We projected this coverage linearly to Chromosome 1, i.e., effective θ = M/(ln(n) * effective coverage)

Assumptions

- Determining genome skimming coverage (5x)
- Determining unique loci for kmers (how much coverage is too much?)
  - Used 8x for the genome skimming experiment
- Determining kmer coverage threshold (for both if statements)
  - For kmer to be considered (all individuals)
  - For SNP to be considered (60%)
    - Should use binomial probability

Our Pipeline



Special Thanks: Shahabeddin Sarmashghi and Vineet Bafna

And to Jennifer for making this amazing theme that Vineet should use in all future lectures (no thanks to Akshat who does not wish to be associated with such a zest for life)

Clipart creds: https://www.kissclipart.com/clipart-genetic-engineering-nucleic-acid-double-he-n0sgl0/

Here is a question for you to answer using Lander Waterman like statistics as discussed in class.

Suppose you have a sample of n haploid individuals, each sampled to a coverage of c. For example, n=10, c=5. What is the probability that a random nucleotide on the genome is sampled with at least k reads (e.g. k>=30) from all sampled individuals. With high values of k we should be able to get high polymorphic sites up to a certain minor allele frequency.

Please give your answer in terms of n,c,k.

Pr(coverage of a random nucleotide with ≥ k reads) =

$$\sum_{k=30}^{nc} (1-e^{-nc})^k \,/\, \sum_{k=1}^{nc} (1-e^{-nc})^k$$