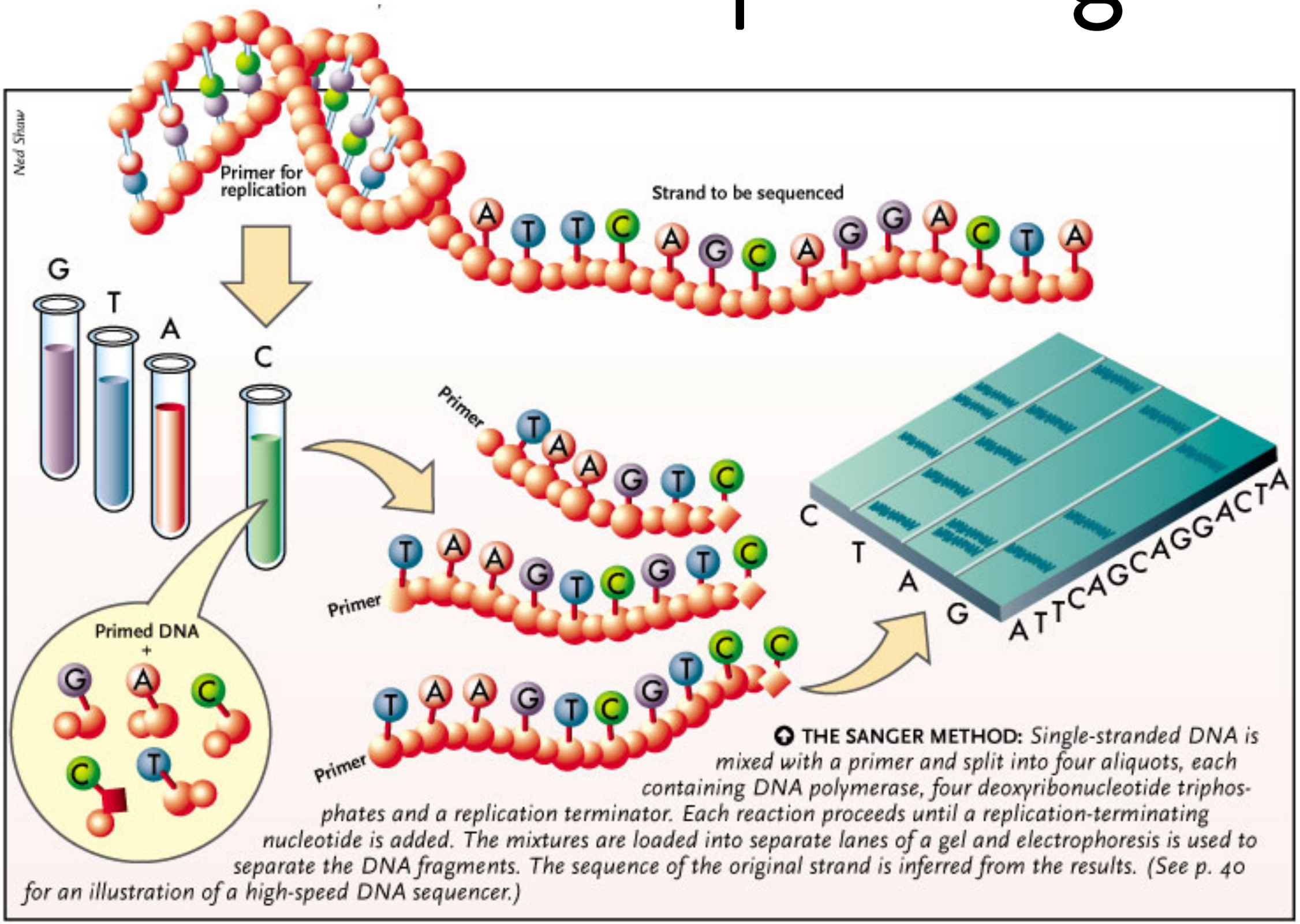
Detecting Gene Mutations in Cancer

androwis abumoussa | janice spence phd | john spence phd | richard burack md/phd

Background

sanger sequencing | next generation sequencing

DNA Sequencing



THE SANGER METHOD: *Single-stranded DNA is mixed with a primer and split into four aliquots, each containing DNA polymerase, four deoxyribonucleotide triphosphates and a replication terminator. Each reaction proceeds until a replication-terminating nucleotide is added. The mixtures are loaded into separate lanes of a gel and electrophoresis is used to separate the DNA fragments. The sequence of the original strand is inferred from the results. (See p. 40 for an illustration of a high-speed DNA sequencer.)*

DNA Sequencing

biology

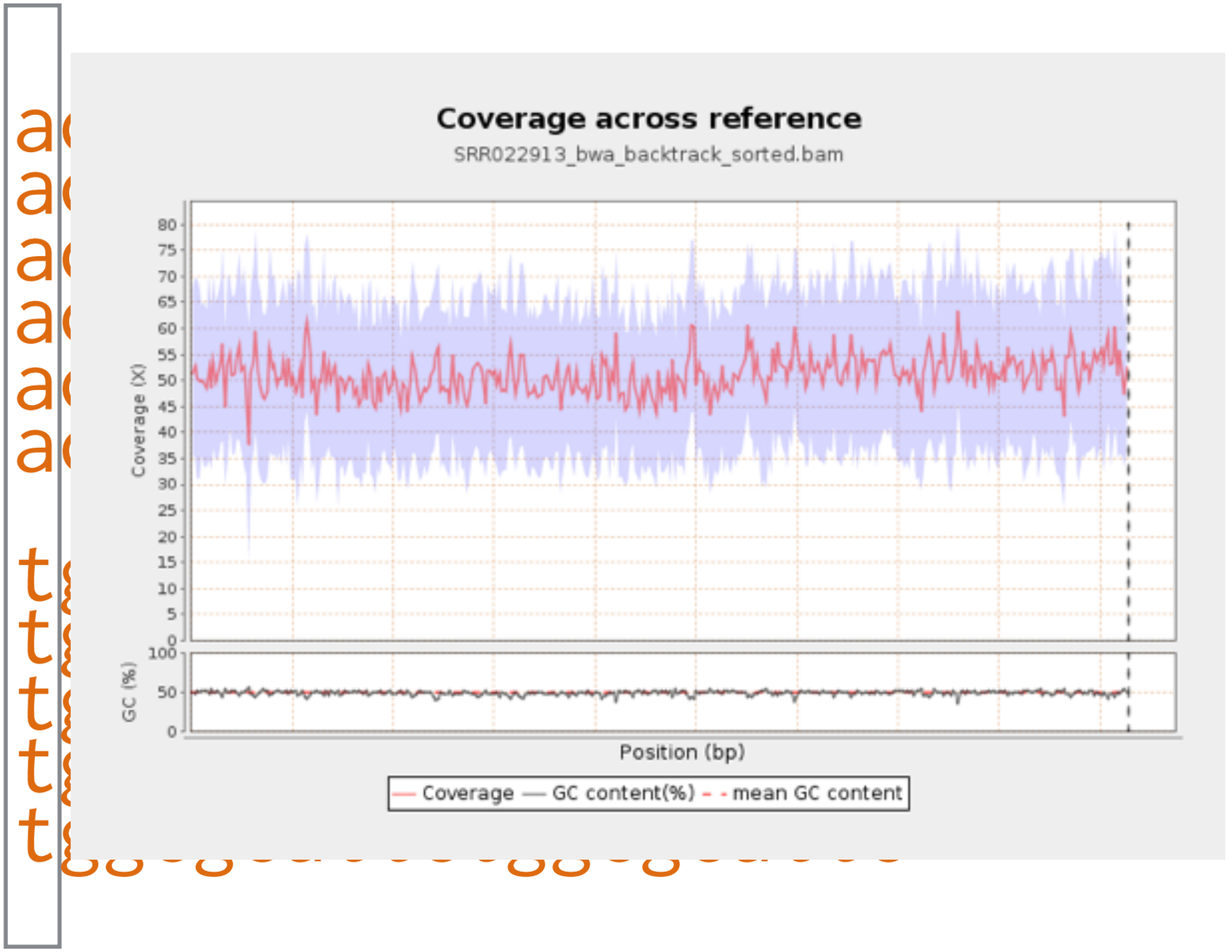
gcgtaagaccgcgtaagaccgcgt
cgcattctggcgcattctggcgca

DNA Sequencing

alignment

DNA Sequencing

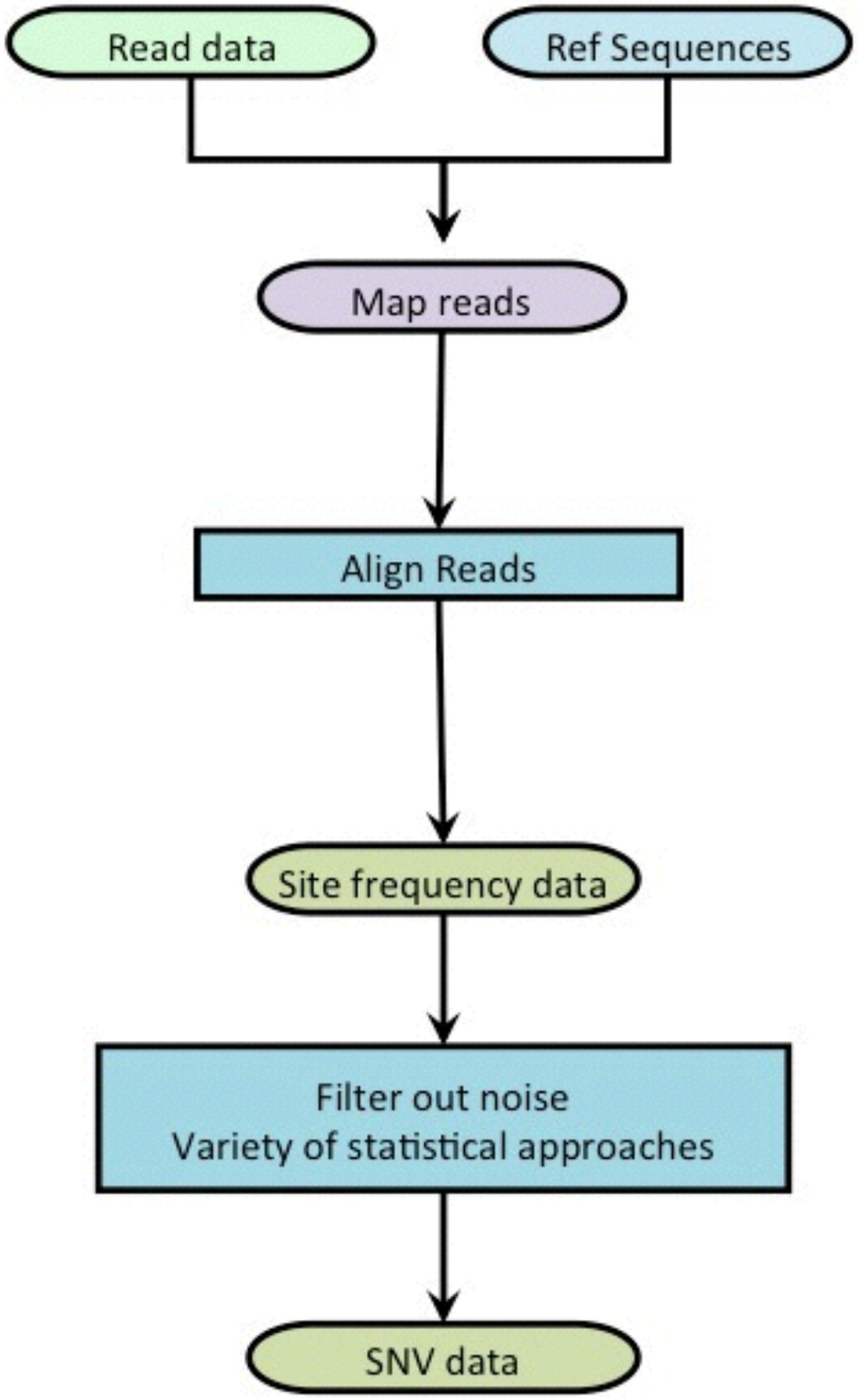
map reduce

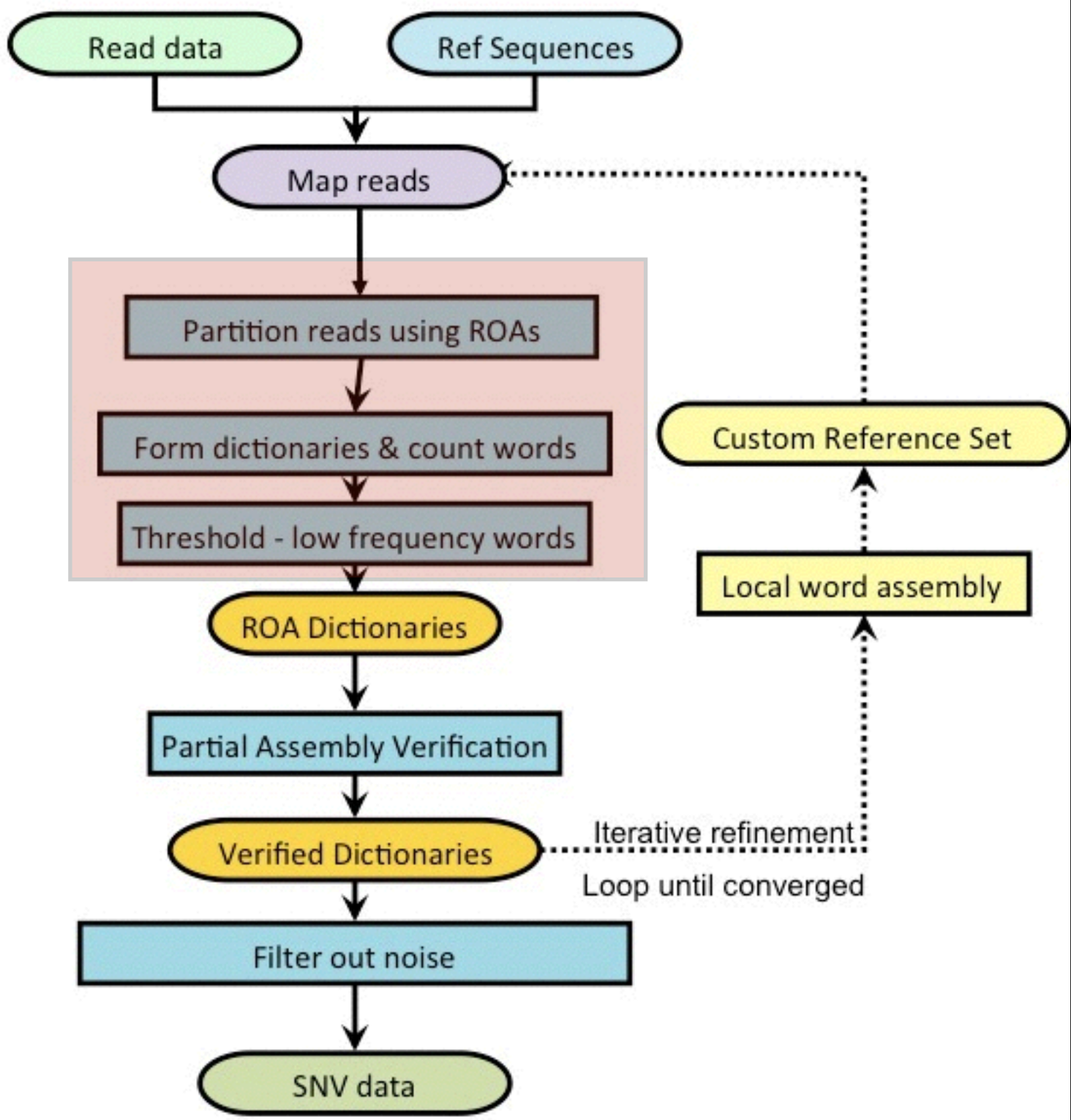


DNA Sequencing | Alignment

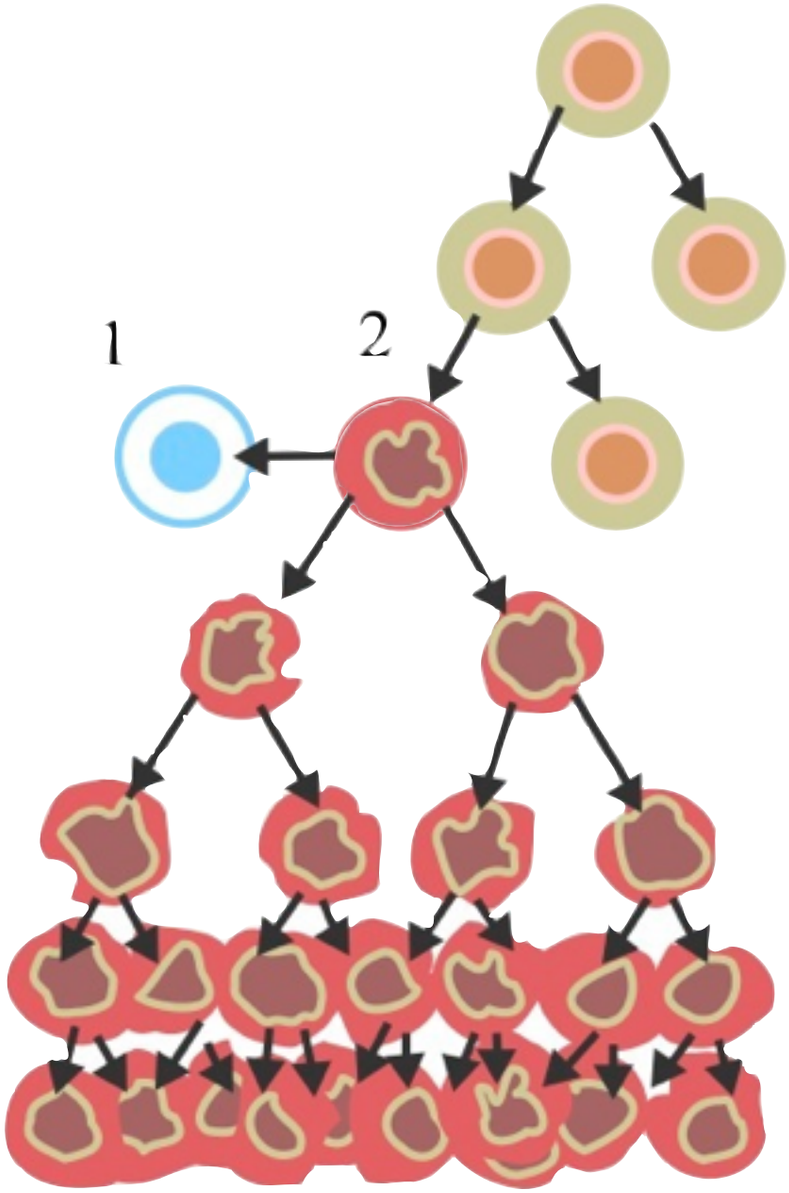
- Maximize Similarity
- Bound Search Space
- Make assumptions to do this

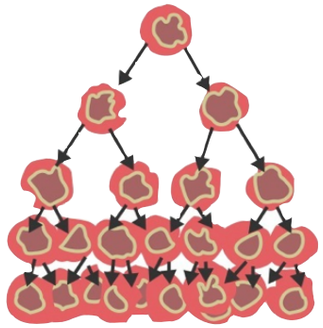
Targeted Resequencing





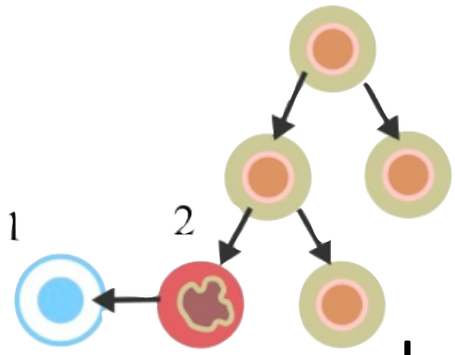
Cell Pathways



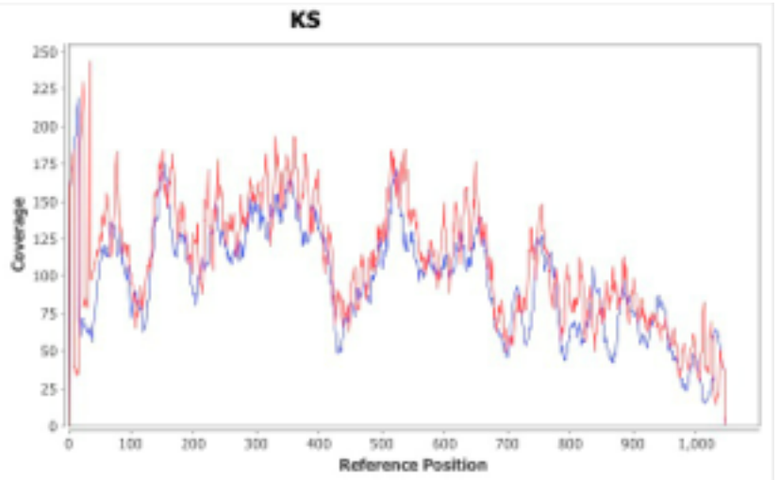


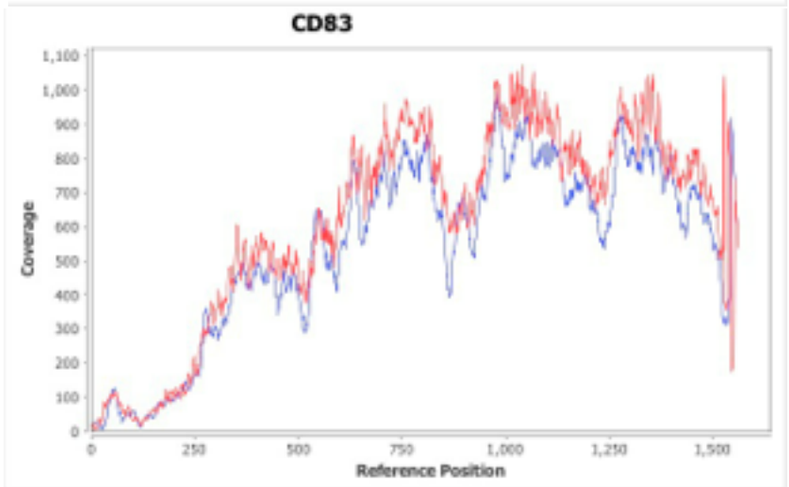
cancer

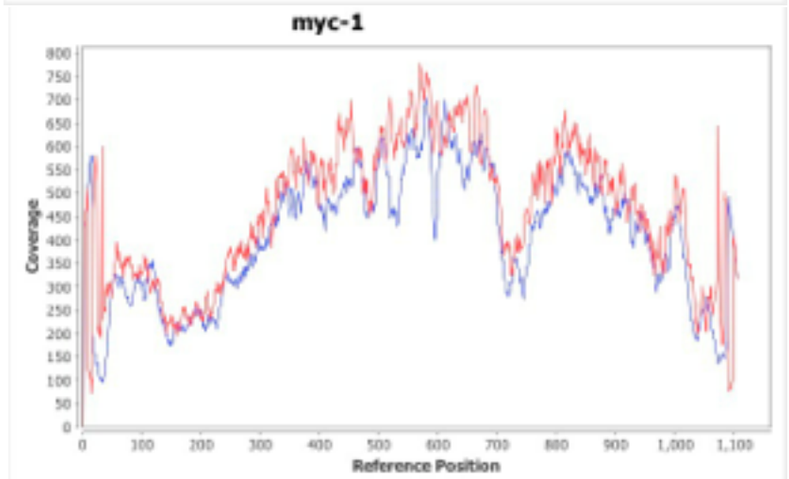
Cell Pathways

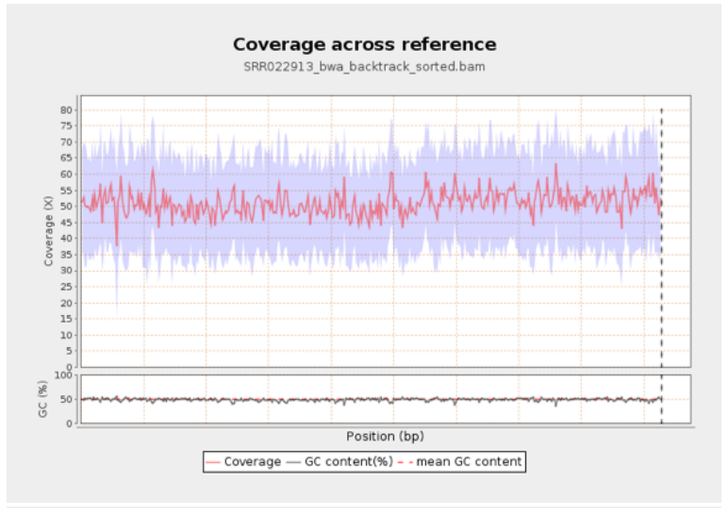


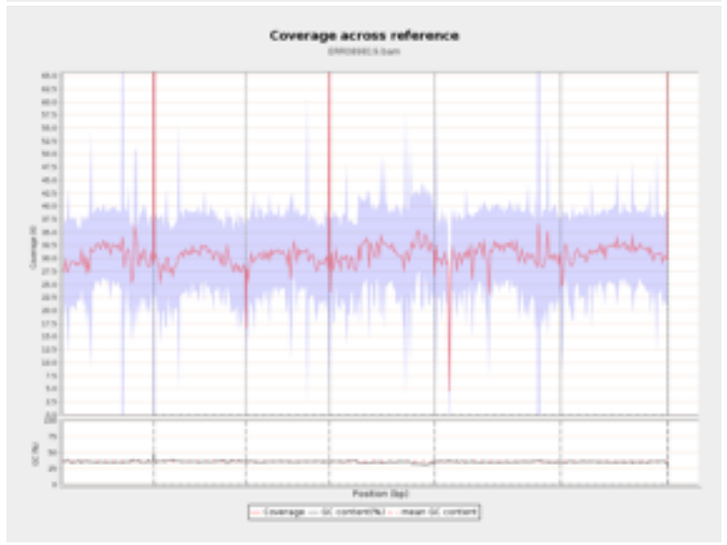
normal



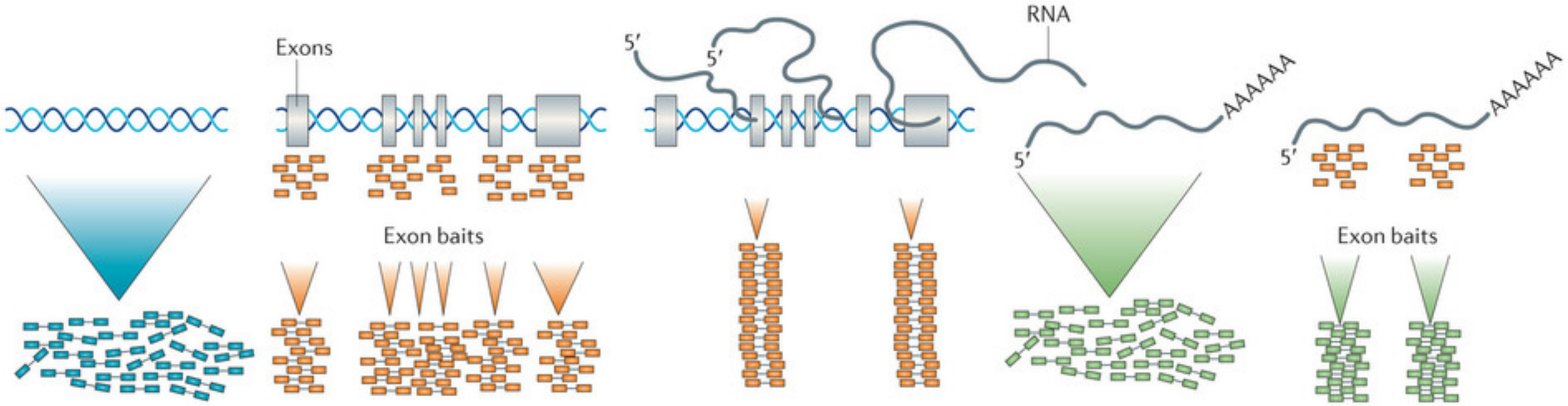








Next Gen Sequencing



Nature Reviews | Drug Discovery

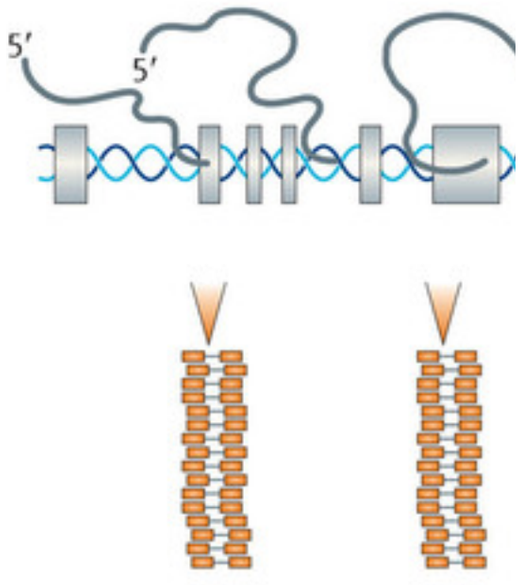
Big Data in Genetics

Parallel Data Structures | Map Reduce | Distributed Computing

Parallel Data Structures

trie | burrow-wheeler's algorithm

Trie Formation



PCR amplicon

Predominant applications:
- Point mutations
- Deletions

```
read : accgcgtaag.....
read : actgcgttag.....
read : actgcgtatg.....
read : accgcgtaag.....
read : actgcgttag.....
read : actgcgtatg.....
read : accgcgtaag.....
read : accgcgtaag.....
read : actgcgttag.....
read : actgcgtatg.....
read : accgcgtaag.....
read : actgcgttag.....
read : actgcgtatg.....
read : accgcgtaag.....
```

Trie Formation

read : accgcgtaag

Step 1 : create suffixes

```
read : accgcgtaag$
       ccgcgtaag$
       cgcgtaag$
       gcgtaag$
       cgtaag$
       gtaag$
       taag$
       aag$
        ag$
         g$
          $
```

Step 2 : alphabetize suffixes

```
read : aag$
       accgcgtaag$
       ag$
       ccgcgtaag$
       cgcgtaag$
       cgtaag$
       gcgtaag$
       gtaag$
       g$
       taag$
       $
```

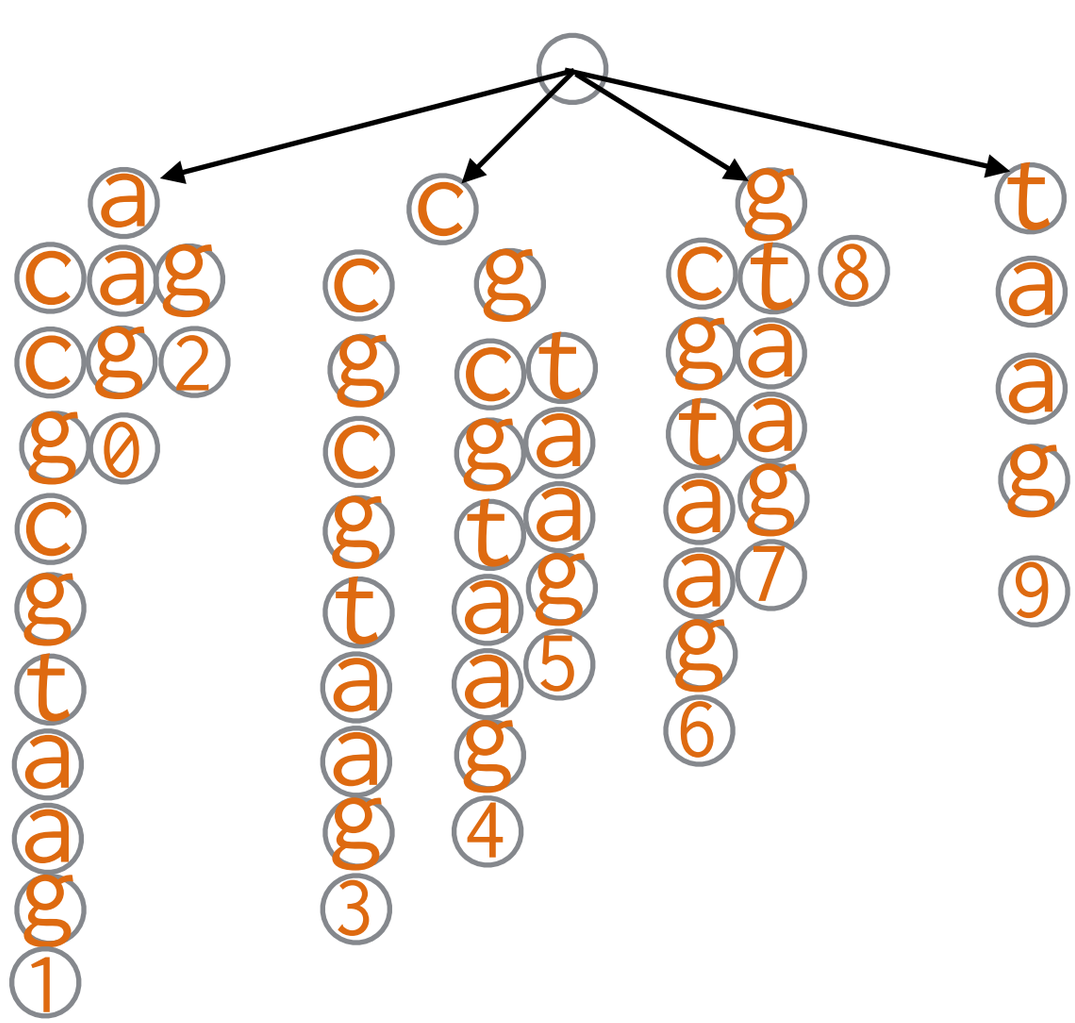
Trie Formation

```
read : accgcgtaag
hash : 1346579028
```

Step 2 : alphabetize suffixes

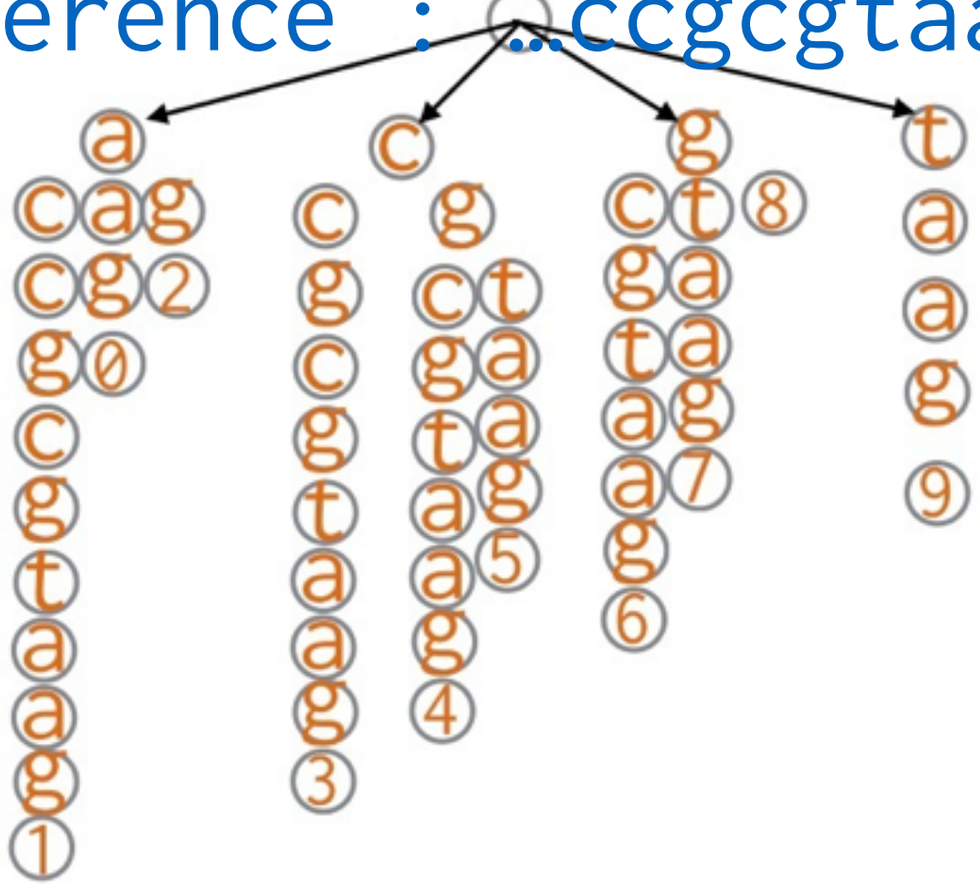
```
read : 0 aag$
1 accgcgtaag$
2 ag$
3 ccgcgtaag$
4 cgcgtaag$
5 cgtaag$
6 gcgtaag$
7 gtaag$
8 g$
9 taag$
$
```

Step 3 : build trie



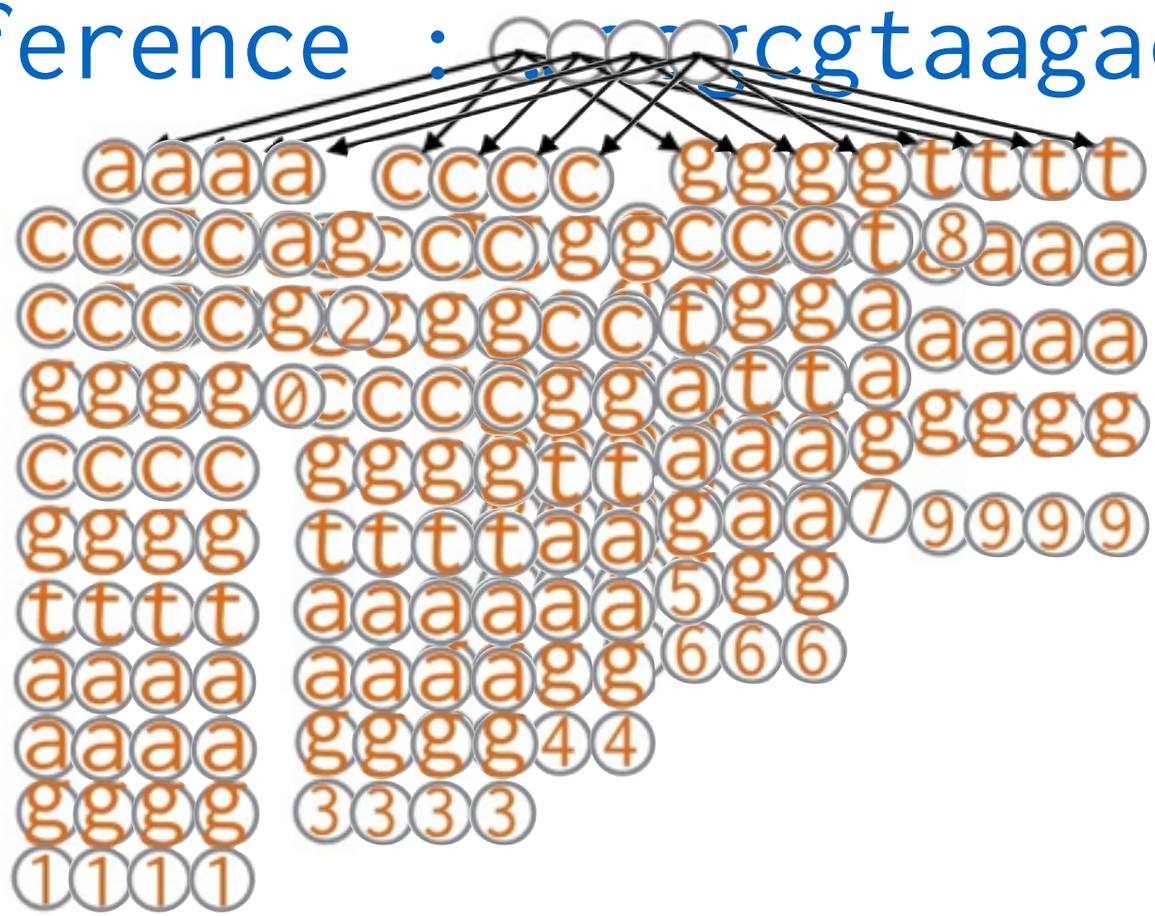
Burrows-Wheeler

Reference : …ccgcgtaagaccgcgtaagaccgcgtaaga…



Burrows-Wheeler (actual)

Reference : ...gcgtaagaccgcgtaagaccgcgtaaga...



Map-Reduce in Genetics

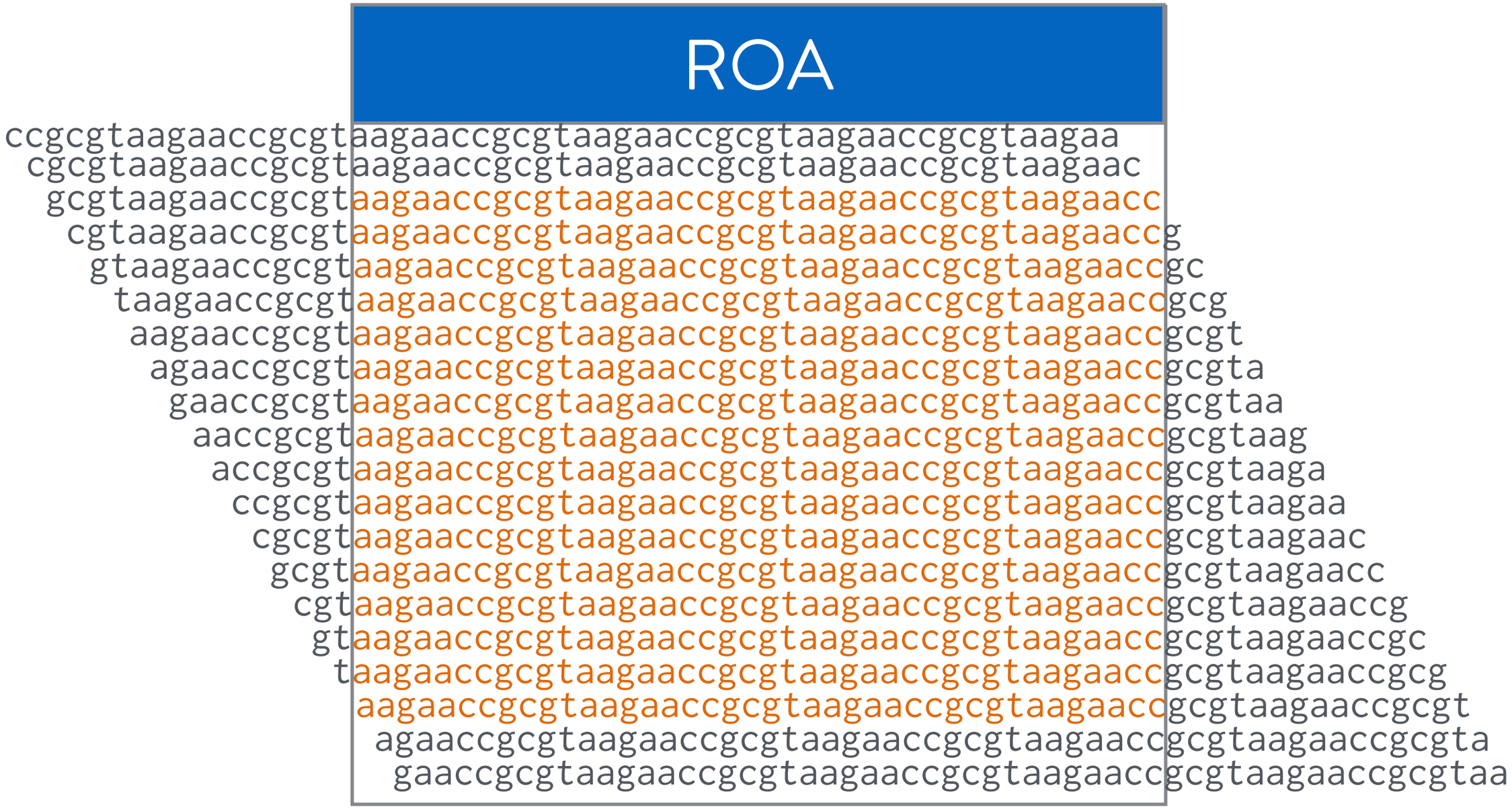
Regions of Analysis

Map Reduce


accgcgtaagaccgcgtaag
accgcgtaagaccgcgtaag
accgcgtaagaccgcgtaag
accgcgtaagaccgcgtaag
accgcgtaagaccgcgtaag
accgcgtaagaccgcgtaag
tggcgcattctggcgcattc
tggcgcattctggcgcattc
tggcgcattctggcgcattc
tggcgcattctggcgcattc
tggcgcattctggcgcattc

ROA Partitioning

Ref : …taagaaccgcgtaagaaccgcgtaagaaccgcgtaagaaccgcgtaagaaccgc…



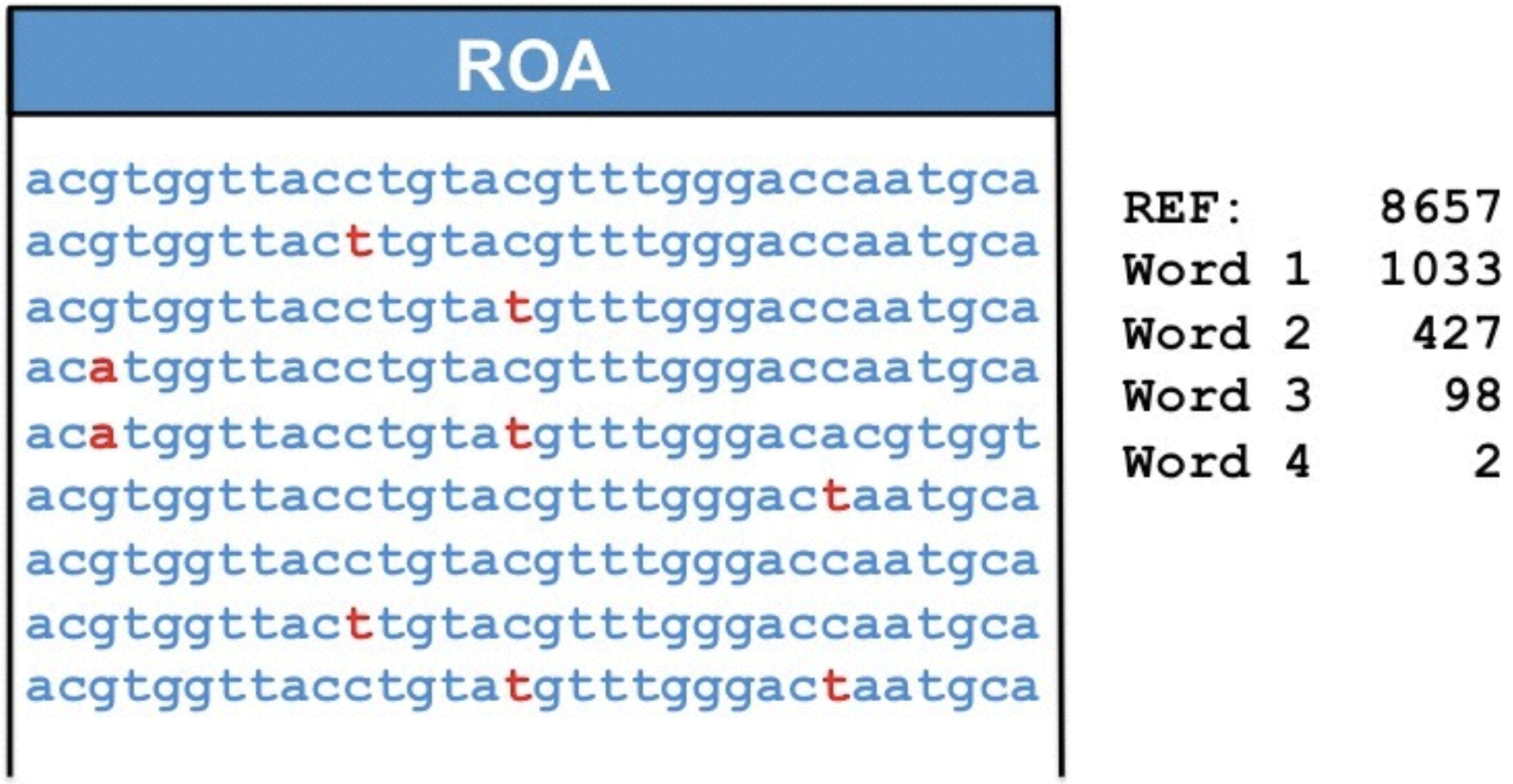
Dictionary Formation

Ref : …taagaaccgcgtaagaaccgcgtaagaaccgcgtaagaaccgcgtaagaaccgcgtaagaaccgc…

ROA
aagaaccgcgtaagaaccgcgtaagaaccgcgtaagaacc
aagaaccgcgtaagaaccgcgtaagaaccgcgtaagaacc
aagaaccgcgtaagaaccgcgtaagaacagcgtaagaacc
aagaaccgcgtaagaaccgcgtaagaaccgtgtaagaacc
aagaaccgcgtaagaaccgcgtaagatccgcgtaagaacc
aagaaccgcgtccgaaccgcgtaagaaccgcgtaagaacc
aagaaccgcgtaagaaccgcgtaagaaccgcgtaagaacc
aagaaccgcgtaagaaccgcgtaagaaccgcgtaagaacc
aagaaccgcgtaagaaccgcgtaagaaccgcgtaagaacc
aagaaccgcgtaagaaccgcgtaagaaccgcgtaagaacc
aagaaccgcgtaagaaccgcgtaagaaccgcgtaagaacc
aagaaccgcgtaagaaccgcgtaagaaccgcgtaagaacc
aagaaccgcgtaagaaccgcgtaagaaccgcgtaagaacc
aagaaccgcgtaagaaccgcgtaagaaccgcgtaagaacc
aagaaccgcgtaagaaccgcgtaagaaccgcgtaagaacc
aagaaccgcgtaagaaccgcgtaagaaccgcgtaagaacc

Length : 34 BP

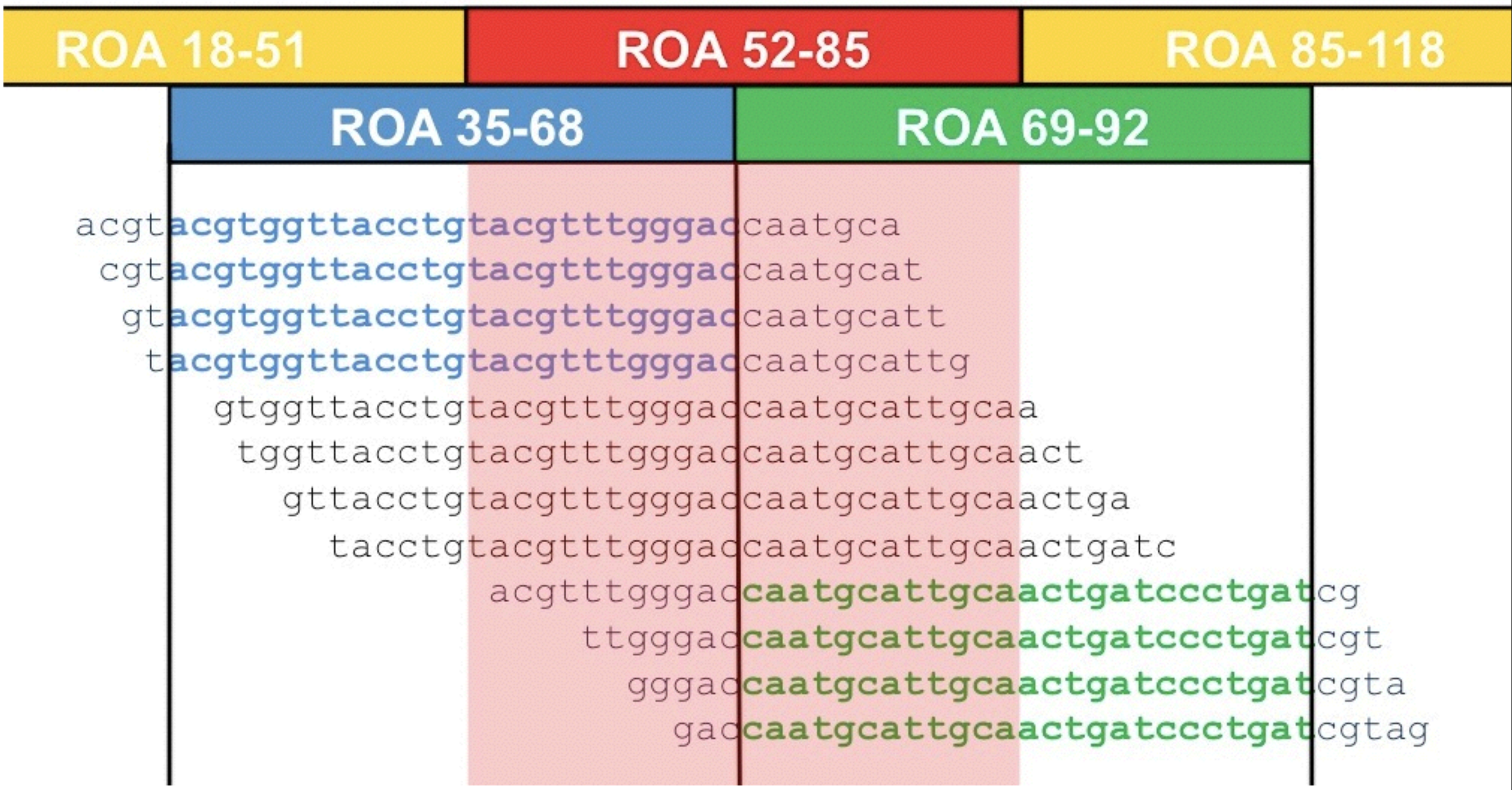
Dictionary Formation



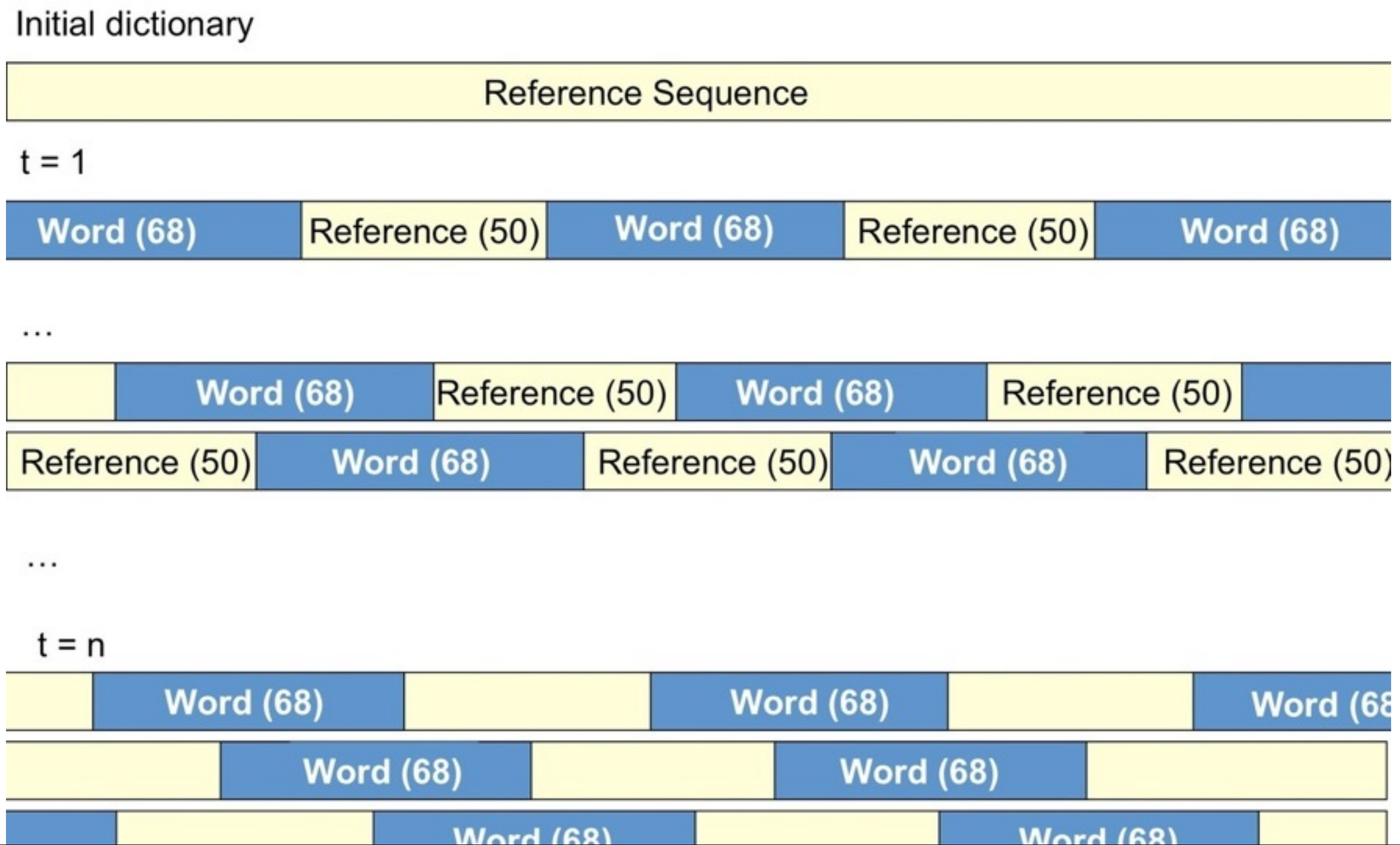
Dictionary Thresholding

REF:	8657
Word 1	1033
Word 2	427
Word 3	98

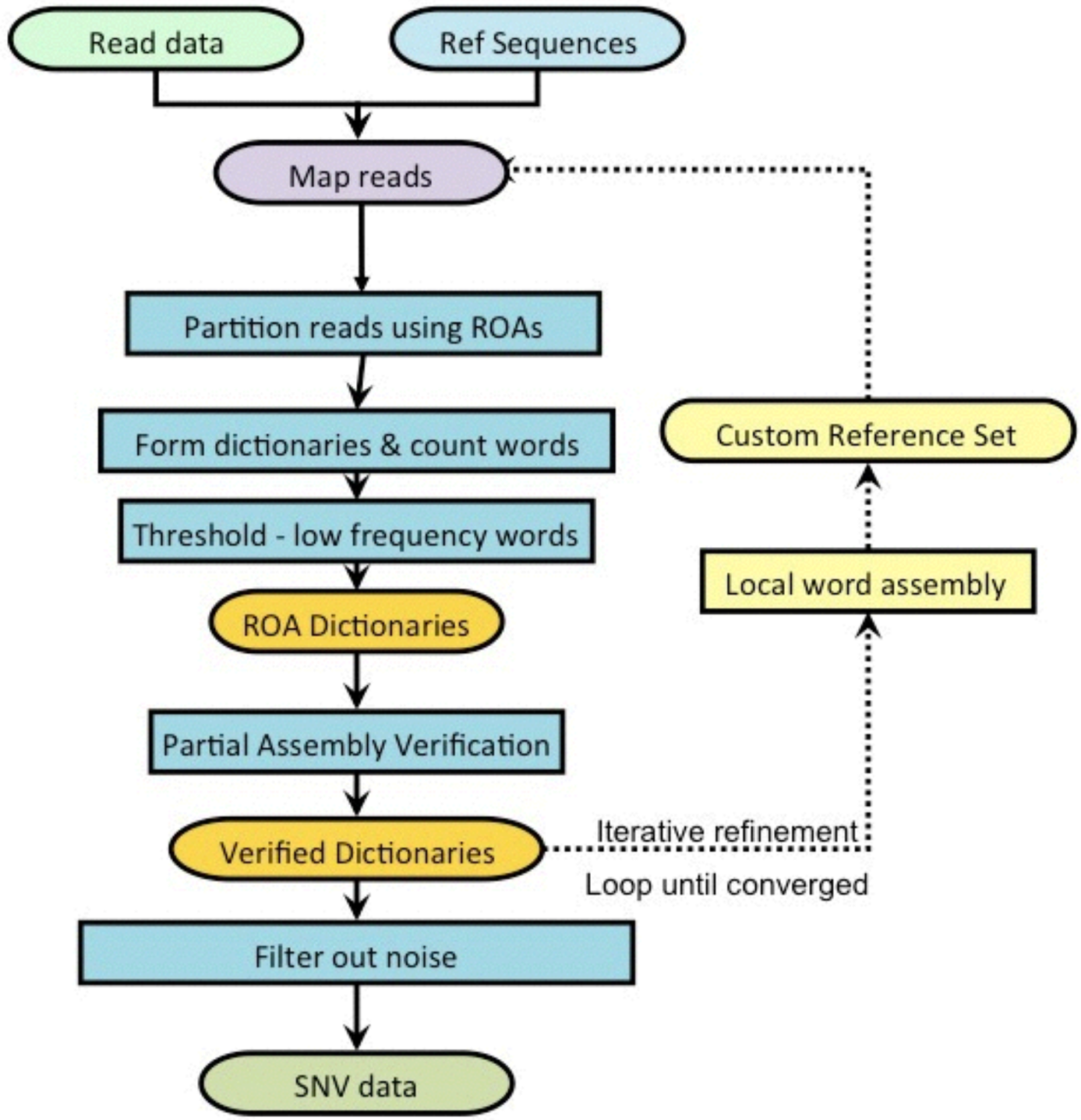
Baseline Reconstruction



Dictionary Reconstruction



Repeat


Read data
Ref Sequences
Map reads
Partition reads using ROAs
Form dictionaries & count words
Threshold - low frequency words
ROA Dictionaries
Partial Assembly Verification
Verified Dictionaries
Filter out noise
SNV data
Custom Reference Set
Local word assembly
Iterative refinement
Loop until converged

Distributed Systems in Genetics

Distributed Systems