High Throughput NGS Data Analysis

Bioinformatics Lab
Wen-Lian Hsu

中央研究院資訊科學研究所
INSTITUTE OF INFORMATION SCIENCE
ACADEMIA SINICA

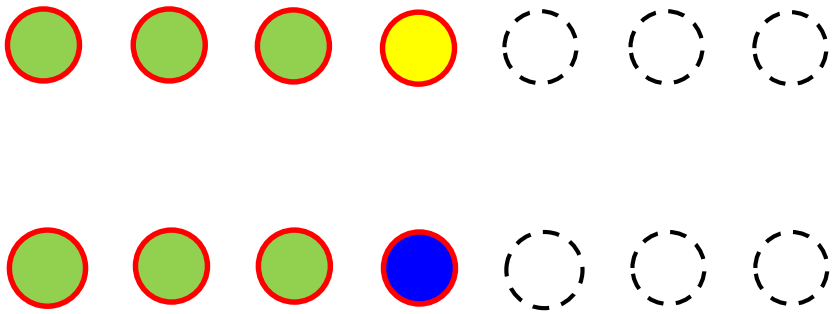
Kart -- An Ultra-fast NGS read mapping Algorithm

Bioinformatics Lab
Wen-Lian Hsu

Background

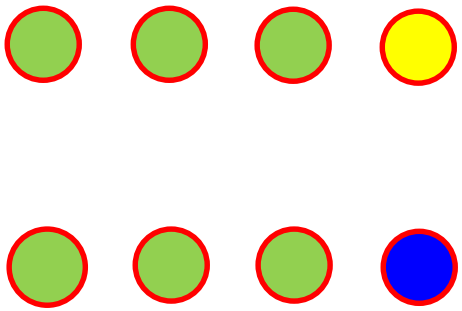
- Next-generation sequencing (NGS) allows biologists to investigate genome-wide variation at nucleotide resolution.
- NGS technologies can produce reads on the order of million/billion base-pairs in a single day.
- Many NGS applications require very fast alignment algorithms.

How to deal with a mismatch in an alignment



Keeping your options open
- Gap opening

1. Substitution



2. Deletion

3. Insertion

Normal Case

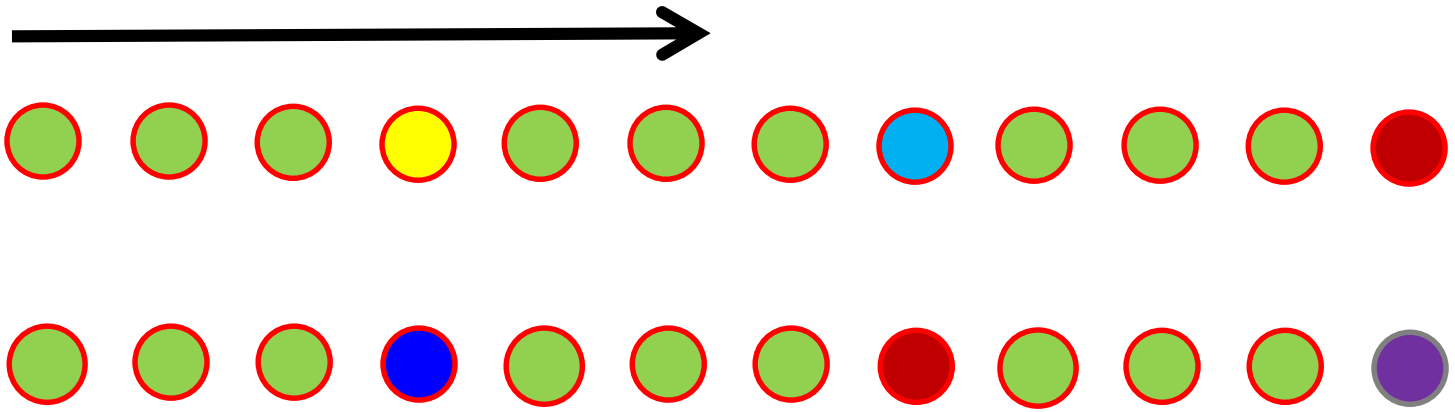
Gapped alignment -- expensive

- Need to consider a huge number of options
- Use Dynamic Programming to manage your options -- O(n^2) time.

Easy Case

Ungapped Alignment

- If we know that the best alignment only requires substitution (no gaps needed), then a linear scan will do -- O(n) time.



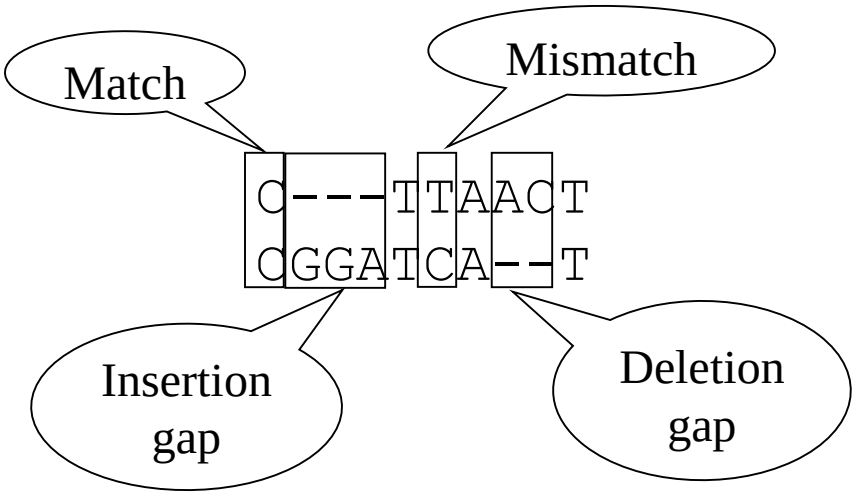
Traditional Pairwise Alignment

Dynamic programming (sequential, very slow)

Sequence a: CTTAACT
Sequence b: CGGATCAT

Time complexity: O(mn)

An alignment of a and b:



A simple scoring scheme

- Match: +8 ($w(x, y) = 8$, if $x = y$)
- Mismatch: -5 ($w(x, y) = -5$, if $x \neq y$)
- Each gap symbol: -3 ($w(-,x)=w(x,-)=-3$)

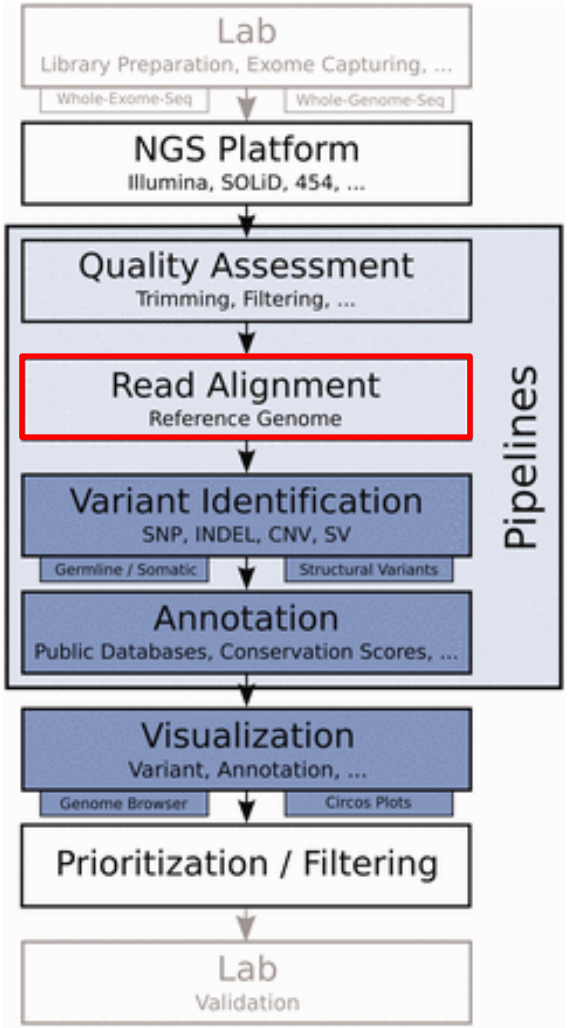
C	–	–	–	T	T	A	A	C	T	
C	G	G	A	T	C	A	–	–	T	
+8	-3	-3	-3	+8	-5	+8	-3	-3	+8	= +12

Alignment score

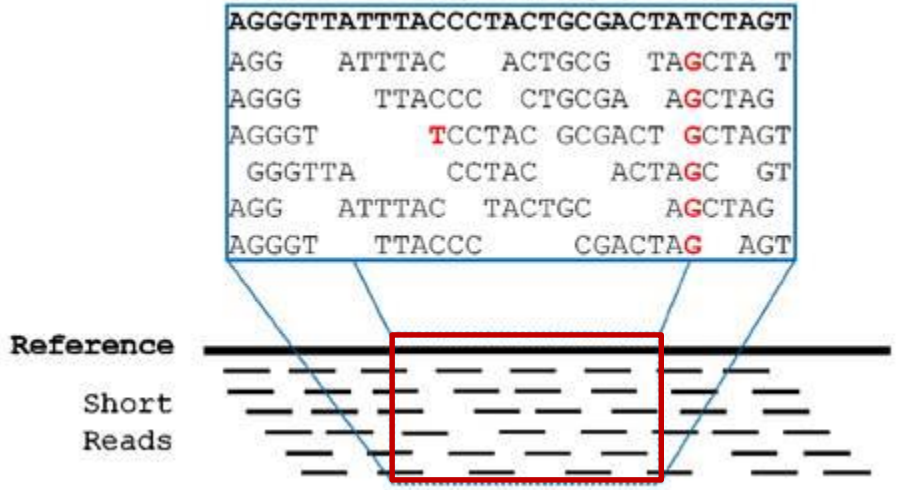
Different Types of Sequence Alignments

- Database Search
  - **BLAST**, FASTA, HMMER
- Pairwise/Multiple Sequence Alignment
  - **ClustalW**, T-Coffee, MAFFT
- Genomic Analysis
  - BLAT: to find regions in a target genomic database which are similar to a query sequence.
- Short Read Sequence Alignment
  - **BWA**, **Bowtie**, SOAP, MAQ,, GSNAP, SHRiMP

Basic workflow for NGS data analysis



Short read mapping



Short read mapping

- Input:
  - A reference genome
  - A collection of short reads
- Output:
  - One or more genomic coordinates for each read
- The mapping sensitivity depends on the read quality and the similarity between the sample genome and the reference genome.

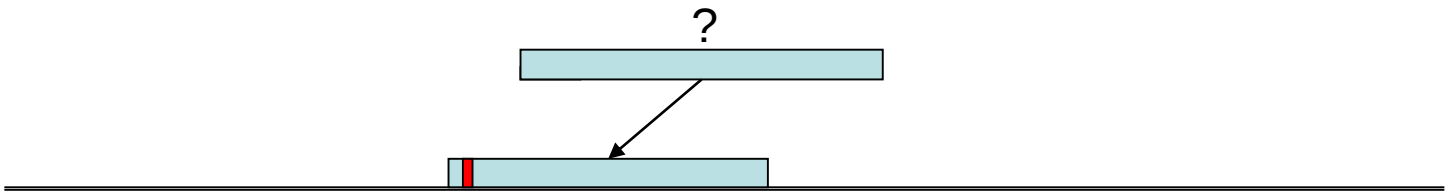
Existing methods

Based on indexing strategy

- BWT/suffix array based
  - Bowtie, BWA, BWA-SW, BWA-MEM, SOAPv2, CUSHAW, Subread, HISAT/HISAT2, HPG-aligner, segemehl
- Hash table
  - CloudBurst, Eland, MAQ, RMAP, SeqMap, SHRiMP, ZOOM, BFAST, NovoAlign, SSAHA, SOAPv1

Challenges of DNA read mapping (I)

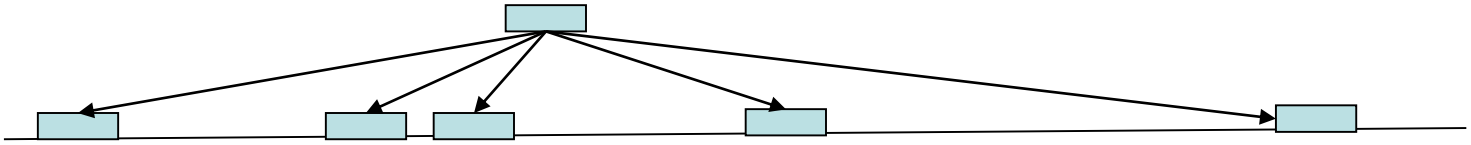
Inexact matching



- A read may not exactly match any position in the reference genome.
- Such mismatches may represent
  - a SNP (single-nucleotide polymorphism) or
  - a sequencing error.

Challenges of DNA read mapping (II)

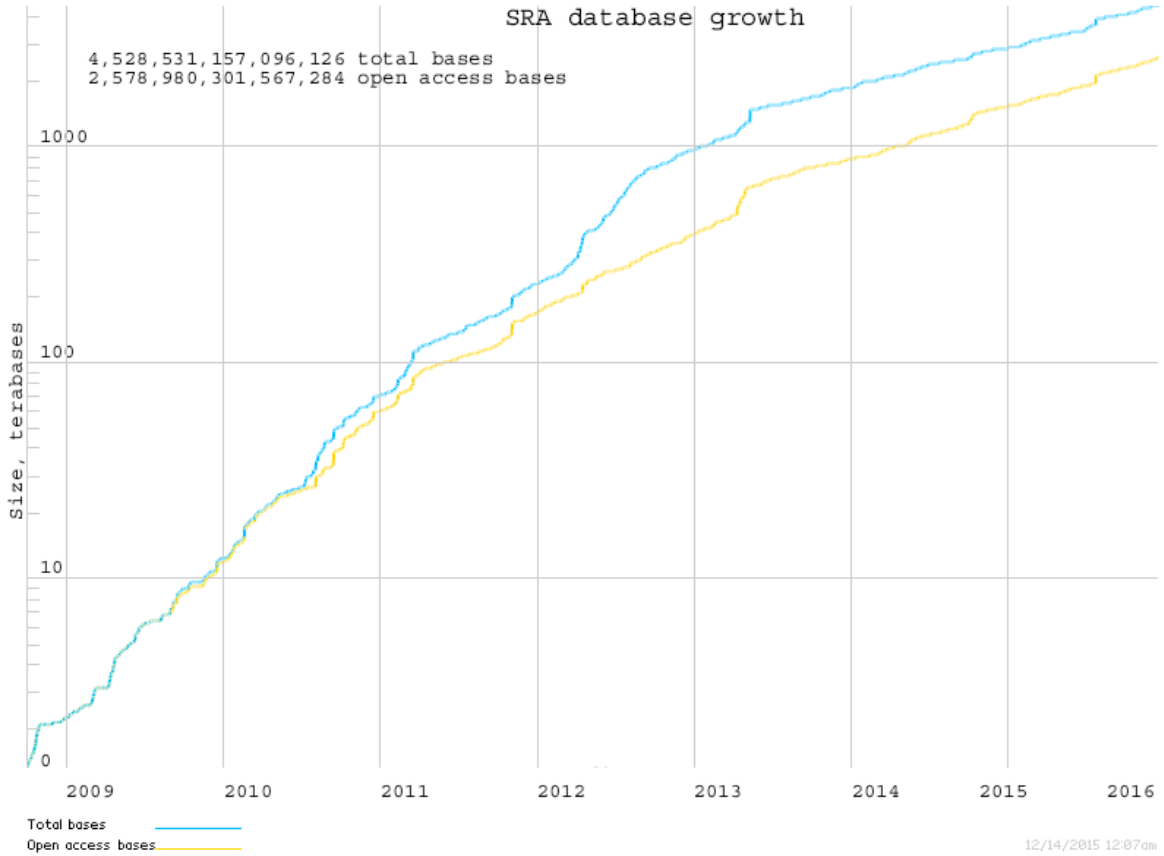
Multiple mapping



- A single read may occur more than once in the reference genome.
- The user may choose to ignore reads that appear more than *n* times.

Challenges of DNA read mapping (III)

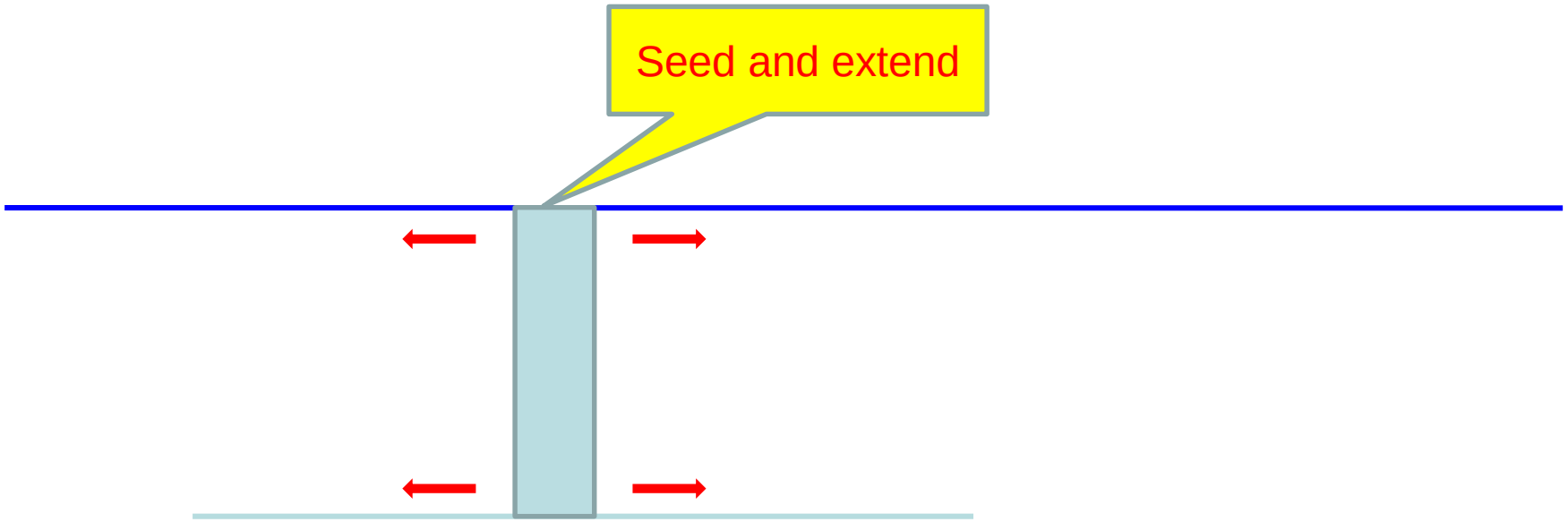
Huge amount of data to be processed



Algorithm Overview

- Seed-and-extend
  - Most aligners adopt seed-and-extend methodology (such as BLAST).
  - Initiate an alignment with a seed and extend the alignment with different dynamic programming strategies.

Seed-and-Extend



Our Strategy

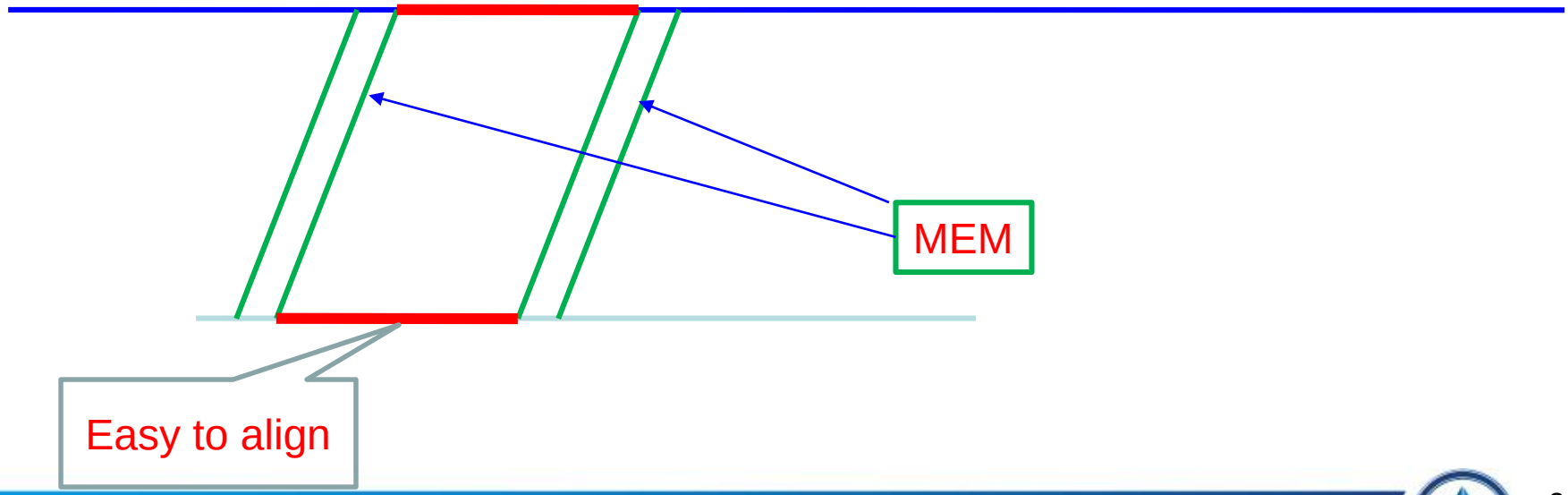
- Cluster close-by seeds together
- Eliminate overlapped seeds
- Map all remaining seeds simultaneously
- Extend parallel seeds to parallel segments
- Divide the read and align the remaining segments recursively

A Crucial Observation

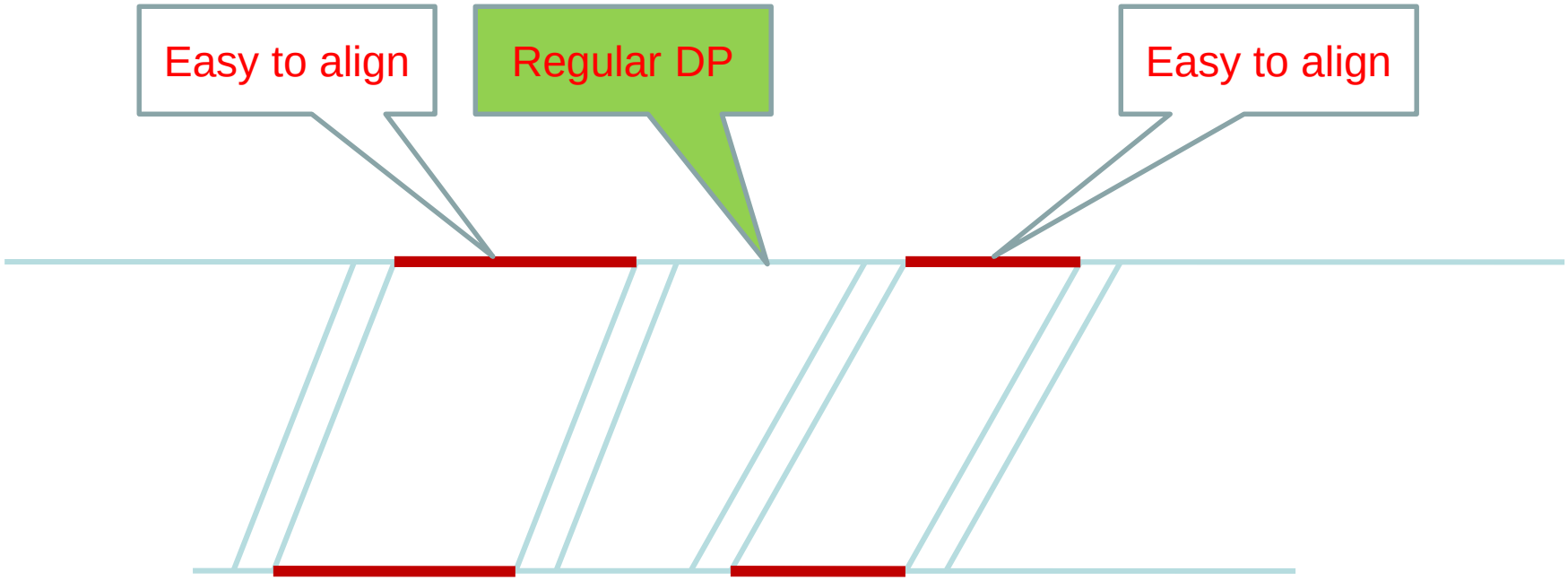
A MEM is a maximal exact match between them

Whenever you have two parallel MEMs, the region between them only has substitutions.

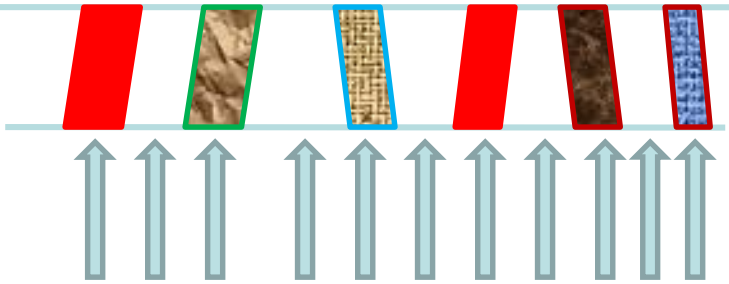
The probability of an exception is around 10^{-5}



Divide and Conquer



Divide and Conquer



Assume you have 10 segments. Original DP takes n^2 time. Now it takes 10 x $(n/10)^2$ = $n^2/10$ time.

The more segments (longer), the more you save.
Note, the colored segments are easy to align.

Performance on real data

Real dataset	Aligner	Sensitivity	Identical base pairs	MEM (Gb)	Runtime
SRR622458 Illumina-101bp (40 millions)	Kart	98.6	99	12	158
	Bowtie2	97.4	99	4.5	458
	BWA-MEM	98.8	97	8.5	1157
	HISAT2	86.0	99	5.5	298
SRR826460Illumina-150bp (40 millions)	Kart	99.3	149	12	186
	Bowtie2	98.4	149	4.5	769
	BWA-MEM	99.3	147	8.5	1374
	HISAT2	91.9	149	5.5	371

Performance on real data

Real dataset	Aligner	Sensitivity	Identical base pairs	MEM (Gb)	Runtime
SRR826471 Illumina-250bp (34 millions)	Kart	98.6	237	12	395
	Bowtie2	94.7	237	4.5	1729
	BWA-MEM	98.6	220	8.5	3027
M130929 PacBio-7118bp (1.2 millions)	Kart	100.0	5152	13	1811
	BWA-MEM	90.7	2953	9	7338
	LAST	97.2	5022	15	31295
	BLASR	97.8	5389	28.9	18682

The average size of segments requiring gapped alignment

Dataset	LMEM-seed	8-LMEM-seed	NP-gap free	NP-indels	NP-NW
SRR622458					17.5
SRR826460	112.7	13.7	4.5	1.9	19.5
SRR826471	104.2	12.4	7.5	1	22.8
M130929	21.3	12.4	10.8	1.4	21.3

Average length of segments requiring gapped alignment

中央研究院資訊科學研究所
INSTITUTE OF INFORMATION SCIENCE
ACADEMIA SINICA

DART -- A fast and robust alignment algorithm for RNA reads

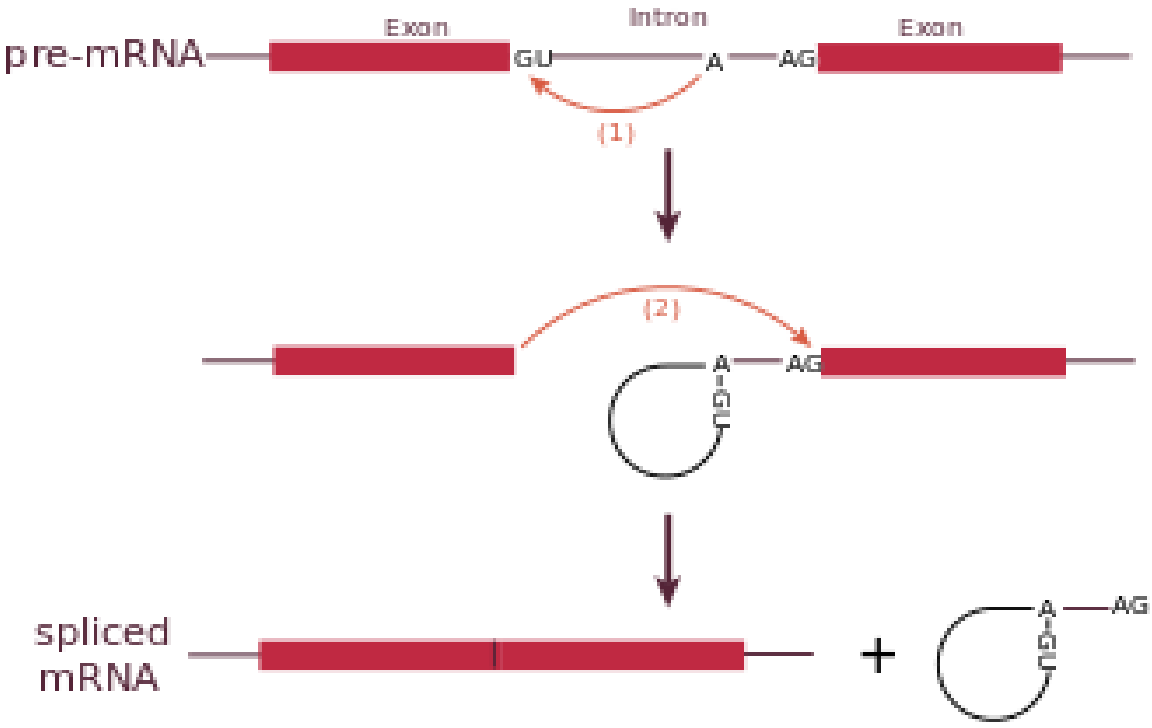
Bioinformatics Lab
Wen-Lian Hsu

DART

- Other DNA mappers only consider continuous alignment and cannot be used for RNA-seq.
- Kart can be easily adapted for RNA-seq
  - we consider fragmented alignment
- The same divide and conquer strategy can be extended to RNA-sequencing
  - Identify simple pairs and normal pairs (Divide)
  - Find the best alignment for each pair (Conquer)

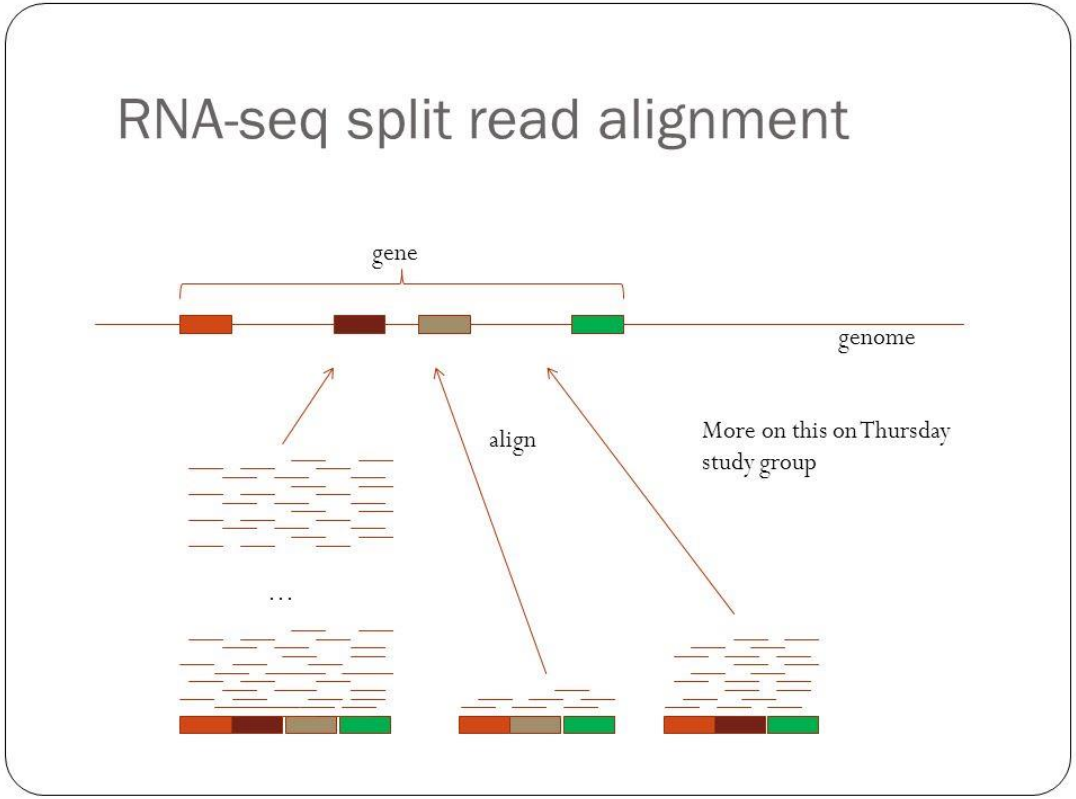
Background

- RNA-Seq technologies is a powerful tool to provide high resolution measurement of expression and high sensitivity in detecting low abundance transcripts.



Challenges of RNA-seq alignment

- The alignment of the corresponding RNA-seq read against the reference genome is not contiguous and it is separated by large gaps.



Existing methods

- QPALMA
- TopHat / TopHat2
- GSNAP
- PALMapper
- MapSplice
- RUM
- GEM
- STAR
- HISAT/HISAT2
- Subread

Algorithm Overview

Performance on simulation data

Synthetic datasets	Aligner	Sensitivity	Accuracy	Recall	SJ accuracy	Runtime
SimRead_76	DART	0.991	0.989	0.957	0.969	71
	STAR	0.978	0.981	0.958	0.935	129
	TopHat2	0.852	0.961	0.853	0.918	6172
	Subread	0.965	0.988	0.929	0.964	2610
	MapSplice2	0.962	0.976	0.940	0.967	3602
	HISAT2	0.911	0.977	0.889	0.964	353
SimRead_101	DART	0.992	0.988	0.965	0.968	95
	STAR	0.977	0.982	0.958	0.936	154
	TopHat2	0.809	0.967	0.809	0.912	10357
	Subread	0.955	0.987	0.925	0.961	2346
	MapSplice2	0.979	0.980	0.960	0.948	4736

STAR is the most read paper in Bioinformatics

Performance on real data

Real datasets	Aligner	Sensitivity	Seq Identity	SJ accuracy	Runtime
SRR3351428 (58.6 millions) 100 bp	DART	0.975	0.999	0.634	244
	STAR	0.922	0.996	0.562	270
	TopHat2	0.844	0.998	0.673	22464
	Subread	0.858	0.998	0.661	3312
	MapSplice2	0.966	0.996	0.620	67446
	HISAT2	0.883	0.998	0.865	404
ERR1518881 (66.6 millions) 100 bp	DART	0.874	0.997	0.636	369
	STAR	0.841	0.987	0.606	371
	TopHat2	0.640	0.995	0.680	21185
	Subread	0.759	0.992	0.660	4008
	MapSplice2	0.893	0.988	0.680	15021
	HISAT2	0.756	0.993	0.833	480

Performance on real data

Real datasets	Aligner	Sensitivity	Seq Identity	SJ accuracy	Runtime
SRR3439468 (88.5 millions) 150 bp	DART	0.930	0.996	0.655	481
	STAR	0.841	0.992	0.626	594
	TopHat2	NA	NA	NA	NA
	Subread	NA	NA	NA	NA
	MapSplice2	0.930	0.990	0.718	49320
	HISAT2	0.482	0.994	0.797	1306
SRR3439488 (64.5 millions) 250 bp	DART	0.899	0.995	0.790	427
	STAR	0.775	0.990	0.761	813
	TopHat2	NA	NA	NA	NA
	Subread	NA	NA	NA	NA
	MapSplice2	0.851	0.989	0.705	36240
	HISAT2	0.657	0.994	0.833	703

Application to whole genome alignment

Bioinformatics Lab
Wen-Lian Hsu

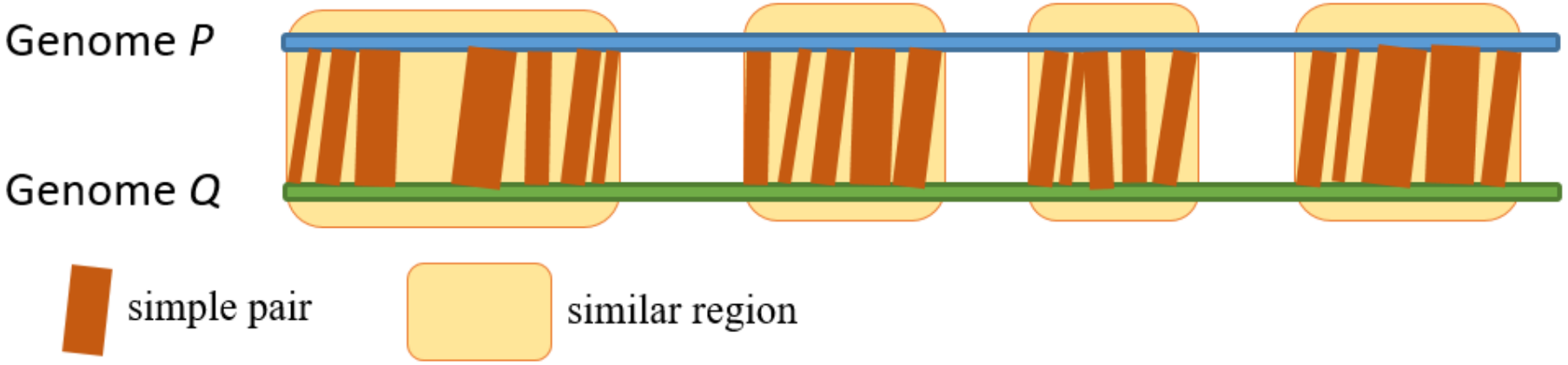
Genome Sequence Comparison

- Problem definition
  - Pairwise genome sequence alignment
- Challenges
  - Extremely long sequence length
  - Repetitive sequences
  - Sequence variations

WGAlign

- Input: Genome sequences G1 and G2
- Algorithm outlines
  - Index G1
  - Search simple pairs with G2 against G1 (parallel)
  - Cluster simple pairs
  - Fill gaps between simple pairs (parallel)
  - Generate sub-alignments of each normal pairs (parallel)
- Output: whole genome alignment, structural variants, dot plot.

WGAlign



Experiment result on real dataset

Dataset	Method	Precision		Recall		Memory (in MB)	Run Time
		Sub	Indel	Sub	Indel		
HG38 vs NA12878 (Diploid)	GSAlign	0.836	0.306	0.928	0.311	15,121	282
	MUMmer4	0.802	0.333	0.905	0.326	56,652	136,825

Whole Genome Alignment

```
#Identity = 23904 / 25846 (92.48%)

H_pyloriJ99_Eslice              52  CTCAAGAAATGCTCAATAGAGCT-AACGCTCAAGCAGAGATTTTGAGCTTAGCCCAACAAGTAGCGGACAATTTCCACAG
H_pylori26695_Eslice          9380  CTCAAGAAATGCTCAATAGAGCTGAA-GCTCAAGCAGAGATTTTAAATTTAGCTAAGCAAGTAGCGAACAATTTCCACAG

H_pyloriJ99_Eslice             131  CATTCAAGGGCCTATCCAACAAG---ATCTAGAAGAATGCACCGCAGGATCAGCTGGTGTGATTAACGACAACACTTATG
H_pylori26695_Eslice          9459  CATTCAAGGGCCTAT---TCAAGGGGATTTAGAAGAATGTAAAGCAGGATCGGCTGGCGTGATCACTAATAACACTTGGG

H_pyloriJ99_Eslice             208  GTTCAGGTTGCGCGTTTGTGAAAGAGACTCTCAATTCCTTAGAGCAACACACCGCTTATTATGGCAACCAGGTCAATCAG
H_pylori26695_Eslice          9536  GTTCAGGTTGCGCGTTTGTGAAAGAAACTTTAAACTCTTTAGAGCAACACACCGCTTATTACGGCAACCAGGTCAATCAG

H_pyloriJ99_Eslice             288  GATAGGGCTTTGTCTCAAACCATTTTGAATTTTAAAGAAGCCCTTAGCACTTTAGGGAACGACTCAAAAGCGATCAATAG
H_pylori26695_Eslice          9616  GATAGGGCTTTGGCTCAAACCATTTTGAATTTTAAAGAAGCCCTTAACACCCTGAATAAAGACTCAAAAGCGATCAATAG

H_pyloriJ99_Eslice             368  CGGTATCTCTAACTTGCCTAACGCTAAGTCCCTTCAAAACATGACGCATGCCACTCAAAACCCTAATTCCCCAGAAGGTT
H_pylori26695_Eslice          9696  CGGTATCTCCAACTTGCCTAACGCTAAATCTCTTCAAAACATGACGCATGCCACTCAAAACCCTAATTCCCCAGAAGGTC

H_pyloriJ99_Eslice             448  TGCTCACTTATTCTTTGGATACCAGCAAATACAACCAGCTCCAAACTGTTGCGCAAGAATTAGGCAAAAACCCCTTTAGG
H_pylori26695_Eslice          9776  TGCTCACTTATTCTTTGGATTCAAGCAAATACAACCAGCTCCAAACCATCGCGCAAGAATTGGGCAAAAACCCTTTCAGG

H_pyloriJ99_Eslice             528  CGCATCGGCGTGATTAACTATCAAAACAATAACGGGGCGATGAACGGCATCGGCGTGCAAGCGGGCTATAAGCAATTCTT
H_pylori26695_Eslice          9856  CGCTTTGGCGTGATTGACTTTCAAAACAACAACGGCGCAATGAACGGGATCGGCGTGCAAGTGGGTTATAAACAATTCTT

H_pyloriJ99_Eslice             608  TGGCAAAAAAAGGAATTGGGGGTTAAGGTATTATGGTTTCTTTGATTATAACCATGCTTATATCAAATCTAATTTTTTTA
H_pylori26695_Eslice          9936  TGGTAAAAAAAGGAATTGGGGGTTAAGGTATTATGGTTTCTTTGATTATAACCATGCTTATATCAAATCTAATTTTTTCA
```

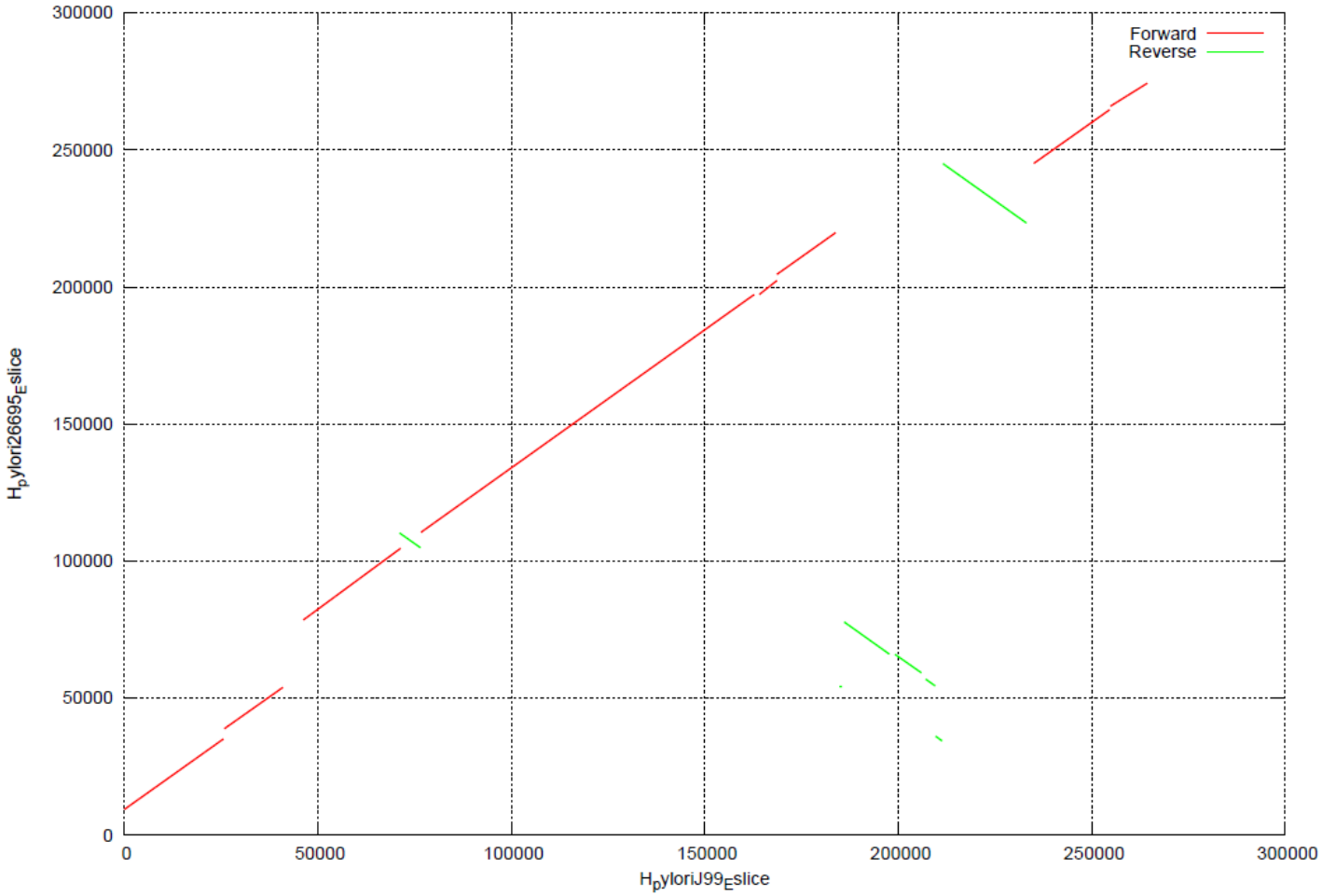
SNP Calling

```
SNP#1 Query=H_pyloriJ99_Eslice  :223  Ref=H_pylori26695_Eslice:9551
Q: TTGTGAAAGAGACTCTCAATTCCTTAGAGCAACACACCGCTTATTATGGCAACCAGG
R: TTGTGAAAGAAACTTTAAACTCTTTAGAGCAACACACCGCTTATTACGGCAACCAGG
             ^   ^ ^  ^  ^                        ^

SNP#2 Query=H_pyloriJ99_Eslice  :289  Ref=H_pylori26695_Eslice:9618
Q: TAGGGCTTTGTCTCAAACCAT
R: TAGGGCTTTGGCTCAAACCAT
             ^

SNP#3 Query=H_pyloriJ99_Eslice  :324  Ref=H_pylori26695_Eslice:9652
Q: GAAGCCCTTAGCACTTTAGGGAACGACTCAAAAGCGATCAATA
R: GAAGCCCTTAACACCCTGAATAAAGACTCAAAAGCGATCAATA
             ^   ^^ ^^^^  ^

SNP#4 Query=H_pyloriJ99_Eslice  :367  Ref=H_pylori26695_Eslice:9695
Q: GCGGTATCTCTAACTTGCCTAACGCTAAGTCCCTTCAAAACATGACGC
R: GCGGTATCTCCAACTTGCCTAACGCTAAATCTCTTCAAAACATGACGC
             ^                  ^  ^

SNP#5 Query=H_pyloriJ99_Eslice  :436  Ref=H_pylori26695_Eslice:9765
Q: CCCAGAAGGTTTGCTCACTTA
R: CCCAGAAGGTCTGCTCACTTA
             ^
```

INDELS Calling

```
Ind#209 Query=H_pyloriJ99_Eslice  :251129 Ref=H_pylori26695_Eslice:261176
Q: ATTCTTTTTGGCATCATATCCTAATAATTA-ATCTA----GCTTTTAAAATGGCCTTGATTATAACTAA
R: ATTCTTTTTGACATCGCATCCTAATAACTATAGCTATTCAGCTTTTAAAATAGCTTTGATTATAACTAA

Ind#210 Query=H_pyloriJ99_Eslice  :251210 Ref=H_pylori26695_Eslice:261262
Q: TAACACAGCCCTAATTTTAGGGGAAGTTAAAGAGCGTTTGAGCGTTATGCGTGCT
R: TAACACAGCC-TTATTTTAGGGGAAACTAAAGAGCATTTGAGCGTTATGCGTGCT

Ind#211 Query=H_pyloriJ99_Eslice  :252915 Ref=H_pylori26695_Eslice:262966
Q: TATGGCTCAACAGCGTGGAAAATAACGCGCAAATGTCCAAGTAAGTC--AA---AAAGTCAAATAAC-ATGGTAGTAGAAT
R: TATGGCTCAATAGCGTGGAAAATAACGTGCAAATGTCCAAGTGA-TCGGAATGTAAAGTTAAA--ACGATGGTAGTAGAAT

Ind#212 Query=H_pyloriJ99_Eslice  :255072 Ref=H_pylori26695_Eslice:266131
Q: GCCCGGTTTCAATACAGGTTTTAT-----TGAT----CGCAGTCAAAACCTCTTTGGCTTTCAAAAAAGCCTTGGAAAGTTCAGCGATGATTTCAT
R: GCCCGGTTTCTATGCAGGTTTTATAAGCTTGATGGATCGTAGTCAAAACTTCTTGGGCTTTCAAAAAAGCCTTTGAAAGCTCAACAATGATTTCAT

Ind#213 Query=H_pyloriJ99_Eslice  :256841 Ref=H_pylori26695_Eslice:267909
Q: ATTGGATTTAATTGGTATTTTGTTTTGGGTATTATAGCAAAAGA
R: ATTGGATTTA----GTATTTTCACT-----ATTATAGCAAAAGA
```

Dot Plotting



Q & A