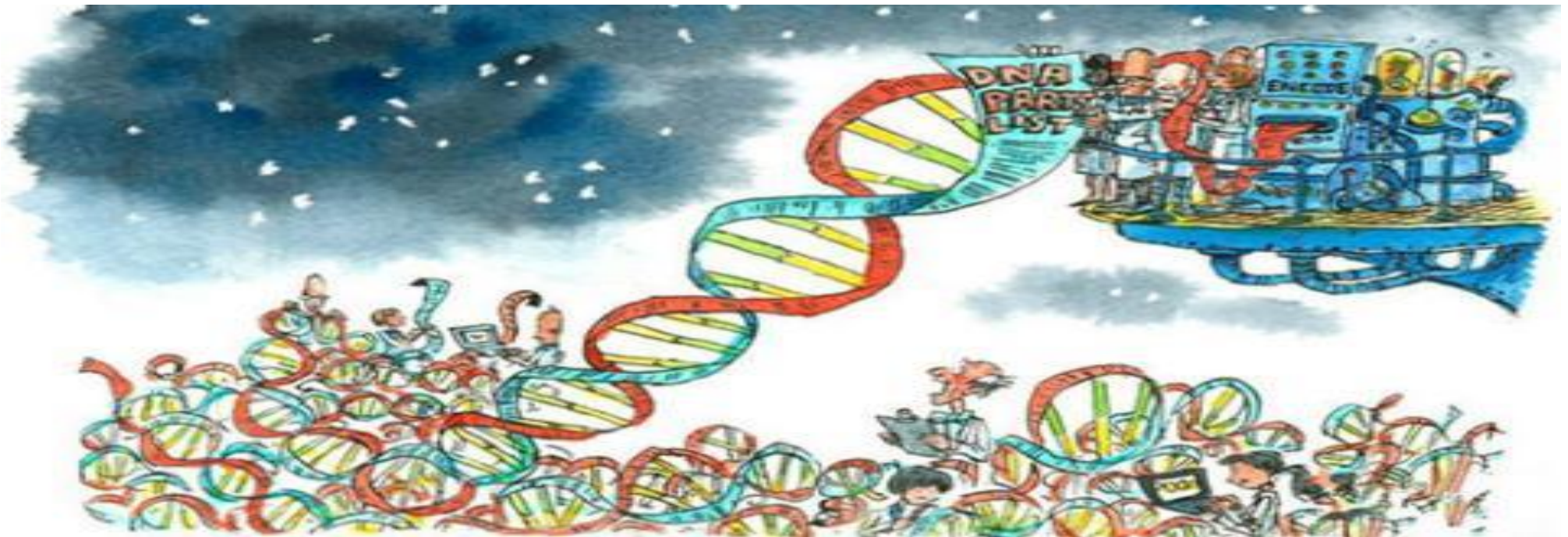


Bioinformatics: Introduction and Methods

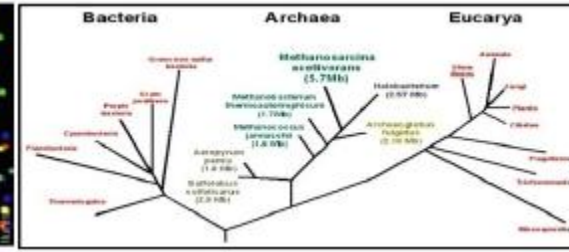
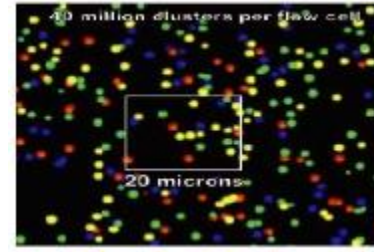
Le Zhang

Computer Science Department, Southwest University





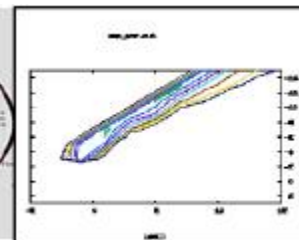
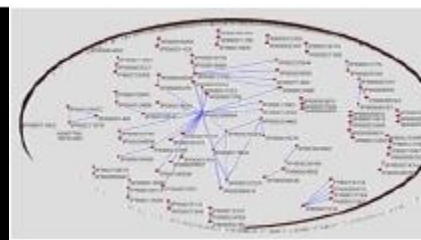
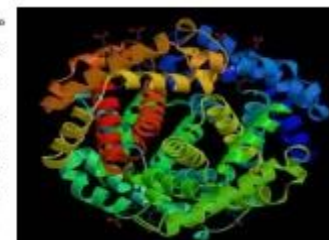
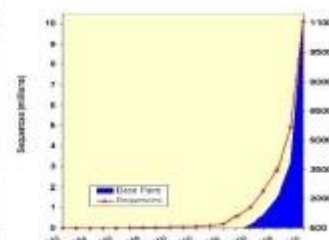
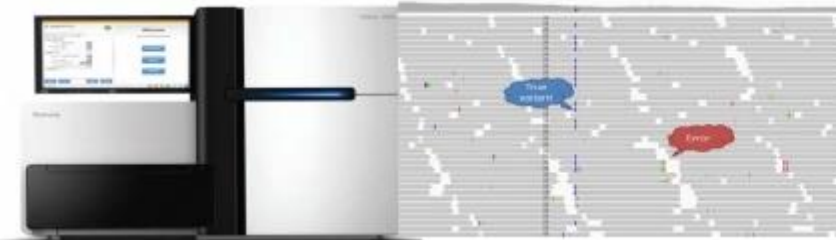
TAACCCTAACCCCTAACCCCTAACCCCTAACCCCTA
CCTAACCCCTAACCCCTAACCCCTAACCCCTAACCC
CCCTAACCCCTAACCCCTAACCCCTAACCCCTAAC
AACCCCTAACCCCTAACCCCTAACCCCTAACCCCTA
ACCCTAACCCCAACCCCAACCCCAACCCCAAC
CTACCCTAACCCCTAACCCCTAACCCCTAACCCCTA
ACCCTAACCCCTAACCCCTAACCCCTAACCCCTAA



Unit 1: From Sequencing to NGS

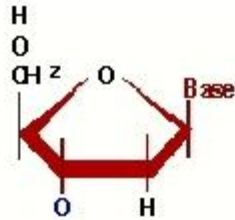
Le Zhang, Ph. D.

Computer Science Department
Southwest University

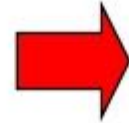
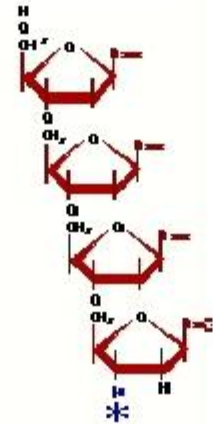
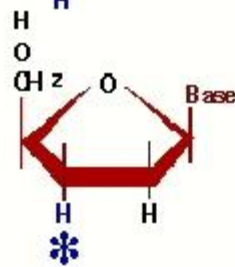


Chain Termination Sequencing

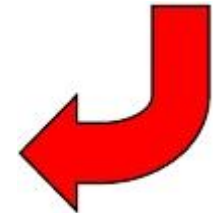
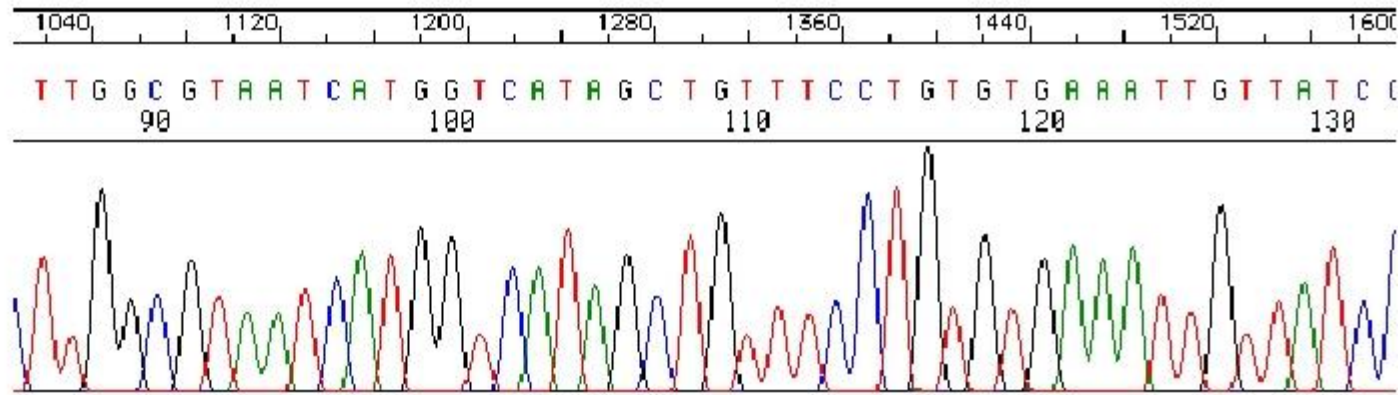
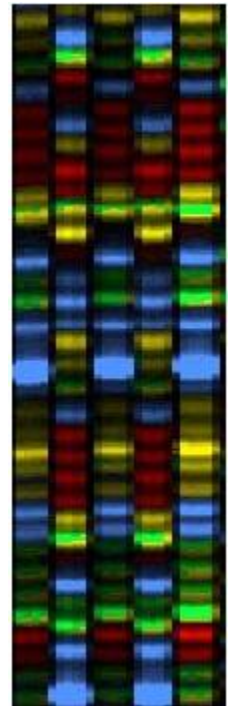
Normal nucleotides:



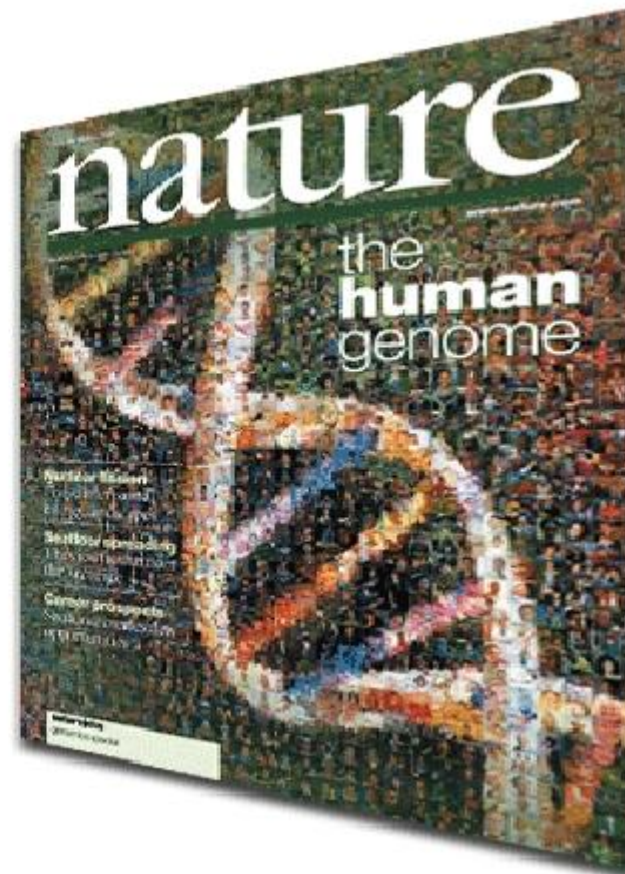
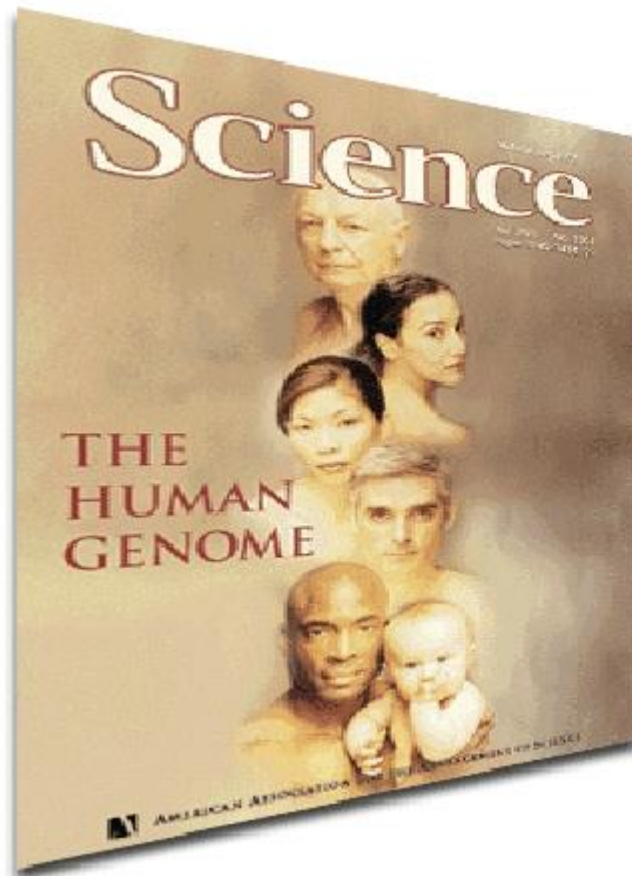
Dideoxy Chain Terminators:



█	G	GC	GAT	GC	GTCC	AC	AC	GC	TAC	AG	GTG
█	T	GC	GAT	GC	GTCC	AC	AC	GC	TAC	AG	T
█	G	GC	GAT	GC	GTCC	AC	AC	GC	TAC	AG	G
█	G	GC	GAT	GC	GTCC	AC	AC	GC	TAC	AG	G
█	A	GC	GAT	GC	GTCC	AC	AC	GC	TAC	A	
█	C	GC	GAT	GC	GTCC	AC	AC	GC	TAC	C	
█	A	GC	GAT	GC	GTCC	AC	AC	GC	TAC	A	
█	T	GC	GAT	GC	GTCC	AC	AC	GC	TAC	T	
█	C	GC	GAT	GC	GTCC	AC	AC	GC	TAC	C	
█	C	GC	GAT	GC	GTCC	AC	AC	GC	TAC	C	
█	G	GC	GAT	GC	GTCC	AC	AC	GC	TAC	G	
█	C	GC	GAT	GC	GTCC	AC	AC	GC	TAC	C	
█	A	GC	GAT	GC	GTCC	AC	AC	GC	TAC	A	
█	A	GC	GAT	GC	GTCC	AC	AC	GC	TAC	A	
█	C	GC	GAT	GC	GTCC	AC	AC	GC	TAC	C	
█	A	GC	GAT	GC	GTCC	AC	AC	GC	TAC	A	
█	C	GC	GAT	GC	GTCC	AC	AC	GC	TAC	C	
█	C	GC	GAT	GC	GTCC	AC	AC	GC	TAC	C	
█	T	GC	GAT	GC	GTCC	AC	AC	GC	TAC	T	
█	G	GC	GAT	GC	GTCC	AC	AC	GC	TAC	G	
█	C	GC	GAT	GC	GTCC	AC	AC	GC	TAC	C	
█	G	GC	GAT	GC	GTCC	AC	AC	GC	TAC	G	
█	T	GC	GAT	GC	GTCC	AC	AC	GC	TAC	T	





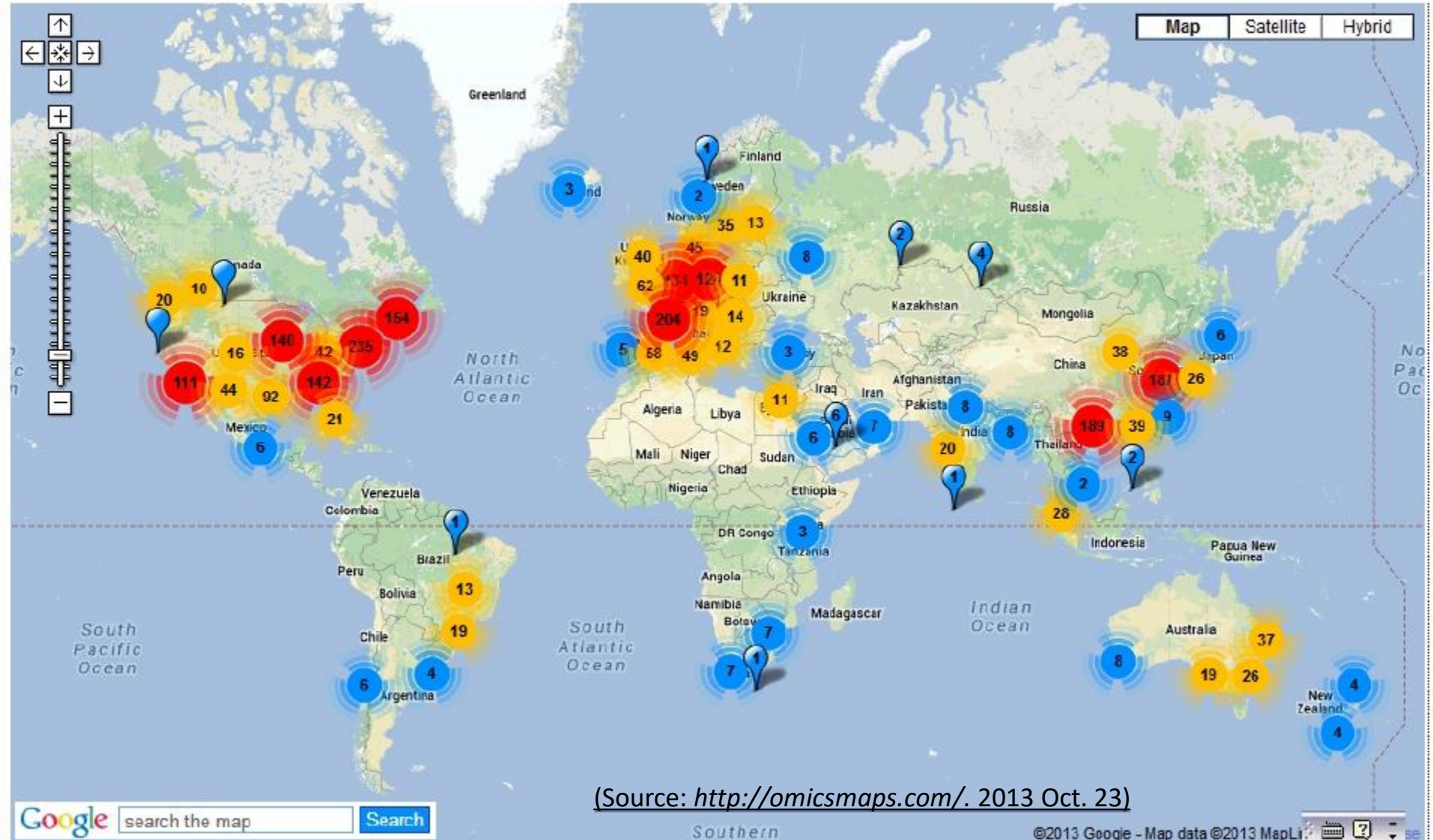


“90% of the three billion base pairs comprising the genome have been read and recorded. The completed work delivers surprises. Perhaps the biggest is that the human genome, estimated at the beginning of the project to contain 80,000 to 100,000 coding genes, appears to possess fewer than 25,000. ”



Next Generation Genomics: World Map of High-throughput Sequencers

Show all platforms
 454
 HiSeq
 Illumina GA2
 Ion Torrent
 MiSeq
 PacBio
 Polonator
 Proton
 SOLiD
 Service Provider



(Source: <http://omicsmaps.com/>. 2013 Oct. 23)

Next Generation Sequencing/Deep Sequencing Sanger Sequencing

Sequencer	454 GS FLX	HiSeq 2000	SOLiDv4	Sanger 3730xl
Sequencing mechanism	Pyrosequencing	Sequencing by synthesis	Ligation and two-base coding	Dideoxy chain termination
Read length	700 bp	50SE, 50PE, 101PE	50 + 35 bp or 50 + 50 bp	400~900 bp
Accuracy	99.9%*	98%, (100PE)	99.94% *raw data	99.999%
Reads	1 M	3 G	1200~1400 M	—
Output data/run	0.7 Gb	600 Gb	120 Gb	1.9~84 Kb
Time/run	24 Hours	3~10 Days	7 Days for SE 14 Days for PE	20 Mins~3 Hours
Advantage	Read length, fast	High throughput	Accuracy	High quality, long read length
Disadvantage	Error rate with polybase more than 6, high cost, low throughput	Short read assembly	Short read assembly	High cost low throughput



Read: A short DNA fragment which is *read out* by sequencer.

- DNA sequence (symbols)
- Quality information

In **FASTQ** format

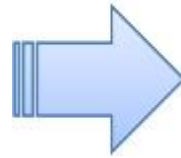
```
@test_fastq
GATTGGGGTTCAAAGCAGTATCGATCAAATAGTAA
+
!"*(((***+))%%%++)(%%%%).1***-+*"
```



```
Seq_ID: test_fastq
Sequence: GATTGGGGTTCAAAGCAGTATCGATCAAATAGTAA
Quality:  !"*(((***+))%%%++)(%%%%).1***-+*"
```


Quality: Given p = the probability of a base calling is *wrong*, its Quality Score can be written as

$$Q = -10 * \log_{10}(p)$$



p	Q
0.1	10
0.01	20
0.001	30
0.0001	40

0 10 20 30
 | | | |
 !"#\$%&'()*+,-./0123456789:;<=>?@ABCDEFGHI
 | | | |
 0 10 20 30

40
 |
 ←
 |
 40

```
@test_fastq
GATTGGGGTTCAAAGCAGTATCGATCAAATAGTAA
+
!"*((( (**+))%% % ++)(% % % %).1***_+*"
```

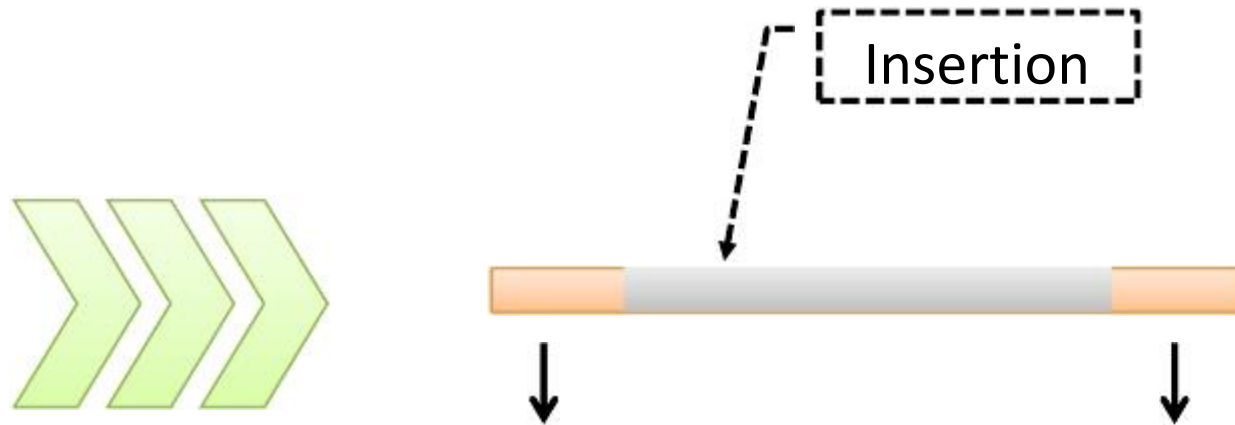
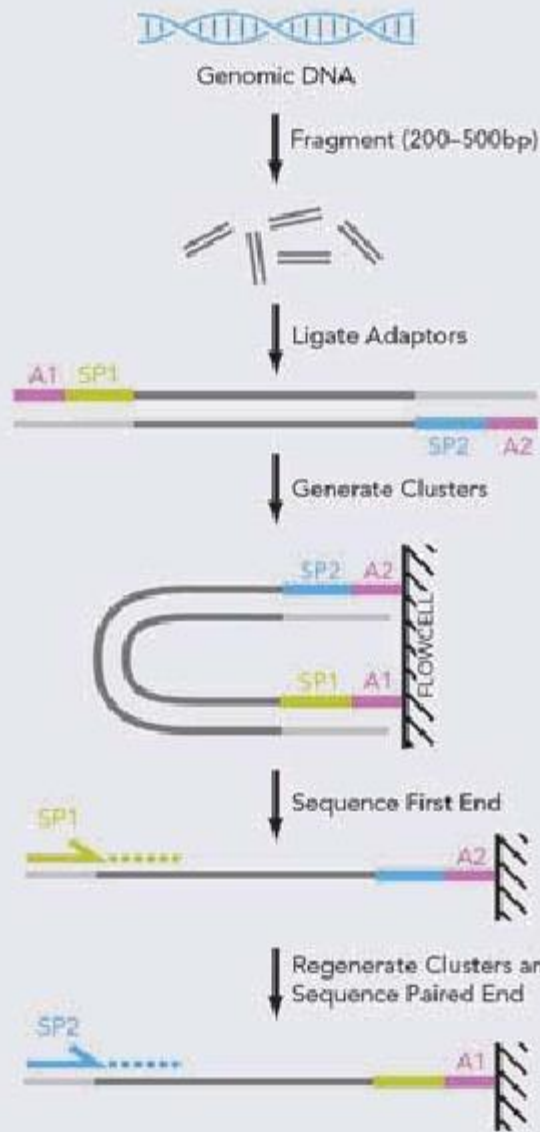


```

0           10           20           30           40
|           |           |           |           |
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHI
|           |           |           |           |
0           10           20           30           40
```

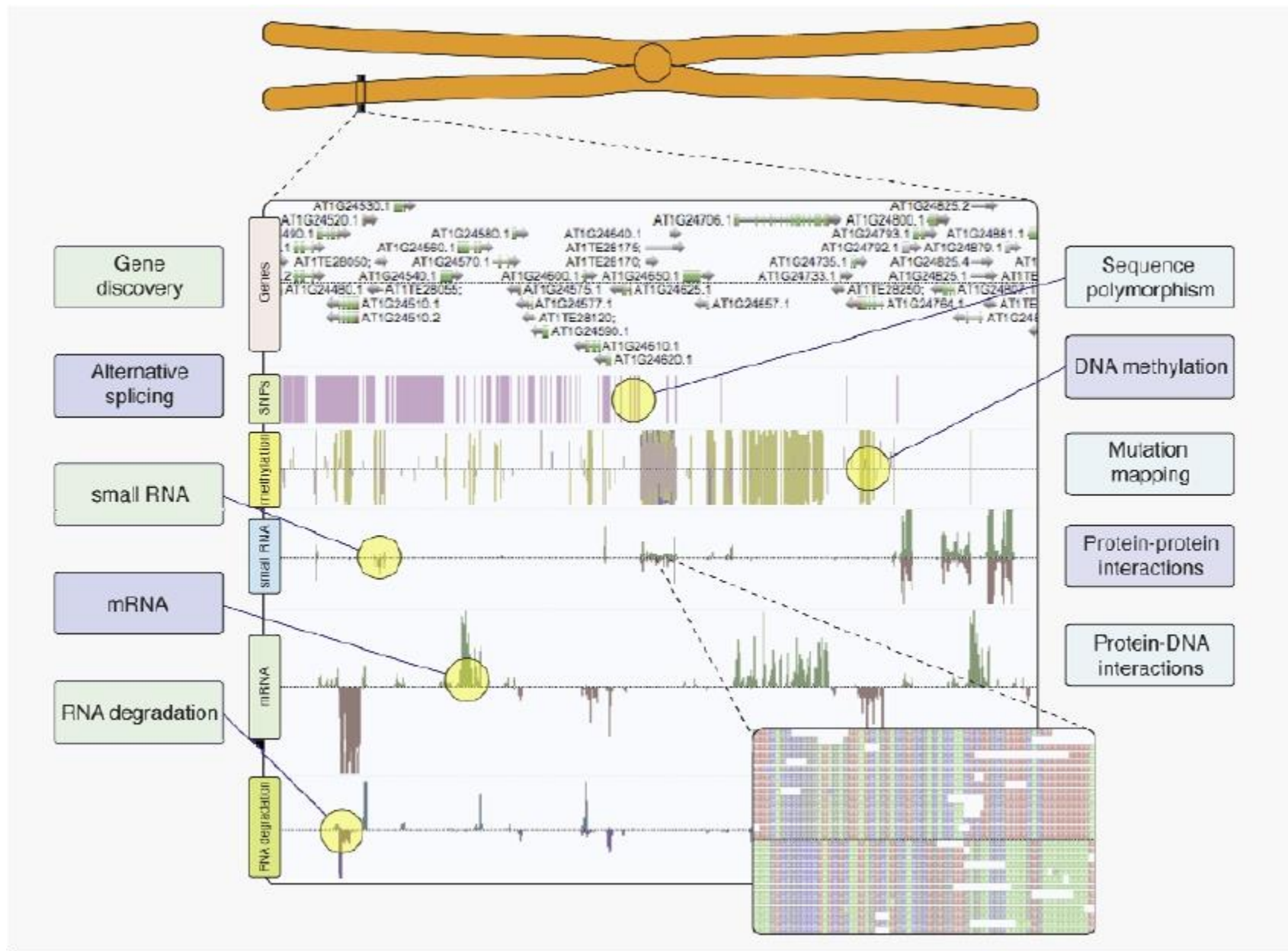
Seq	Quality Symbol	Quality Score	p
G	!	0	1.00
A'		6	0.25
T'		6	0.25
T	*	9	0.13
T(7	0.20
G(7	0.20
G(7	0.20
G(7	0.20
G	*	9	0.13
T	*	9	0.13
T	*	9	0.13
C	+	10	0.10
A)		8	0.16
A)		8	0.16
A	%	4	0.40
G	%	4	0.40
C	%	4	0.40
A	+	10	0.10
G	+	10	0.10
T)		8	0.16
A(7	0.20

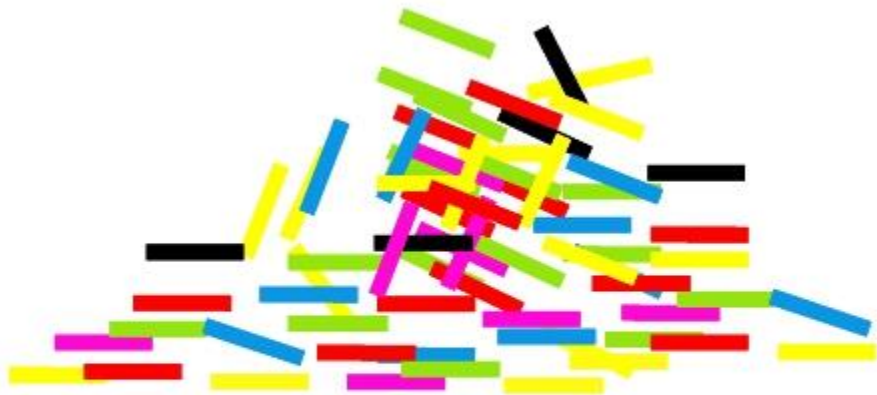
Paired-End Reads



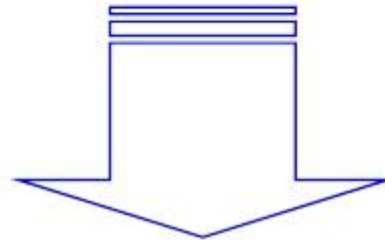
```
@test_fastq/1
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAA
+
!"*((( (**+))% % ++)(% % % %).1***-+*"
```

```
@test_fastq/2
ACATACTATTACTATTACTCCTCATANNNNTNCNN
+
BBB1',9,66<B>9<74<=BB@4=93'!!!!)!!'9
```



mapping



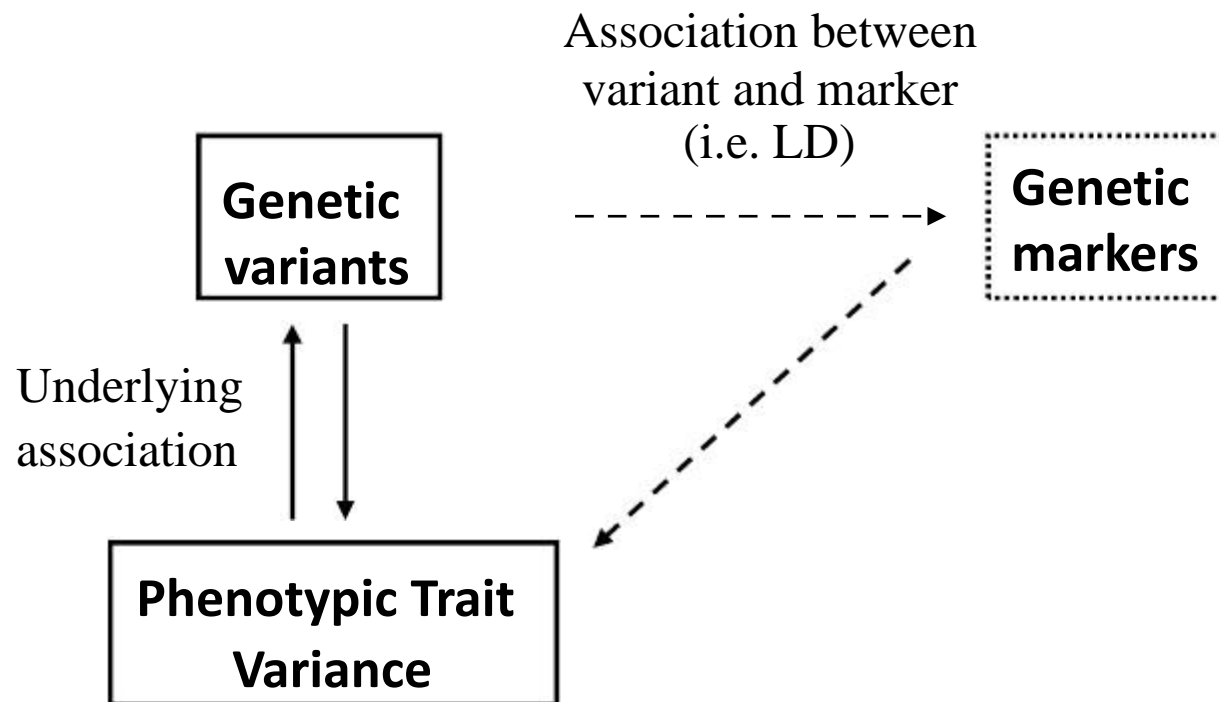
Reference Genome

```
...CCATAG      TAT CGCCC      CGGAATTTCGGTATAAC...
...CCAT      CTATAT CG      TCGGAATT  CGGTATAAC
...CCAT GGCTATAT CGC CTATCGGAAA  GCGGTATA
...CCA AGGCTATAT CGCCCTATCGGA  TTGCGGTA  C...
...CCA AGGCTATAT  GCCCTATCG  TTTGCGGT  C...
...CC  AGGCTATAT  GCCCTATCG  AAATTTGC  ATAC...
...CC  TAGGCTATA  CGCCCTA  AAATTTGC  GTATAAC...
...CCATAGGCTATATGGCGCCCTATCGGC  CAATTTGCGGTATAAC...
```



Genetic variants

Association Study: for the given phenotypic trait, “functional variants” could be identified by **comparing allele frequencies** at hundreds of thousands of polymorphic sites, *i.e.* allele A is associated with phenotypic trait P if (and only if) people who have P also have A more (or less) often than would be predicted from individual frequencies of A and P in the assessed population.



A

Reference	P L N I E V P K I S L H S L I L [*] D F S A V S F L D V S S V R G L K
GIT 264-1	P L N I E V P K I S L H S L I L N F S A V S F L D V S S V R G L K
Sense	5' - CCTCTCAACATTGAGGTCCCCAAAATCAGCCTCCACAGCCTCATTCTCGACTTTTCAGCAGTGTCTTTCTTGATGTTTCTTCACTGAGGGGCCTTAAA-3'
Antisense	3' - GGAGAGTTGTAACCTCCAGGGGTTTTAGTCGGAGGTGTCGGAGTAAGAGCTGAAAAGTCGTCACAGGAAAGAAGTACAAAGAAGTCACTCCCCGGAATTT-5'

3' - GGAGCGTTGTAACCTCCAGGGGTTTTAGTCGGAGGTGTCGGAGTAAGAGTT-5'

3' - GTTCTAACTCCAGGGTTTTTAGTCGGAGGTGTCGGAGTAAGAGTTGAAAA-5'

3' - AACTCCAGGGTTTTTCTGTCGGAGGGTTCGGAGTAAGAGTTGAAAAGTCGT-5'

5' - ctccaggggttttagtcggaggtgtcggagtaagagttgaaaagtcgtca-3'

3' - CCAGGGGTTTTAGTCGGAGGTGTCGGAGTAAGAGTTGAAAAGTCGTCACA-5'

5' - ggggttttagtcggaggtgtcggagtaagagttgaaaagtcgtcacagga-3'

3' - TTTTGGTGGGAGGTGTCGGAGTAAGAGTTGAAAAGTCGTCACAGGAAAG-5'

3' - TTTAGTCGGAGGTGTCGGAGTAAGAGTTGAAAAGTCGTCACAGGAAAGAA-5'

3' - GTCGGAGGCGTTCGGAGTAAGAGTTGAAAAGTCGTCACAGGAAAGAAGTAC-5'

5' - cggaggtgtcggagtaagagttgaaaagtcgtcacaggaagaactacaa-3'

3' - GGGGGGTCGGAGTAAGAGTTGAAAAGTCGTCACAGGAAAGAAGTACAAA-5'

5' - gaggtgtcggagtaagagatgaaaagtcgtcacaggaagaactacaaag-3'

3' - GGTCGGAGTAAGAGTTGAAAAGTCGTCACAGGAAAGAAGTACAAAGAAG-5'

5' - tcggagtaagagttgaaaagtcgtcacaggaagaactacaaagaagtc-3'

3' - GAGTAAGAGTAGAAAAGTCGTCACAGGAAAGAAGTACAAAGAAGTCACTC-5'

5' - agagttgaaaagtcgtcacaggaagaactacaaagaagtcactccccgg-3'

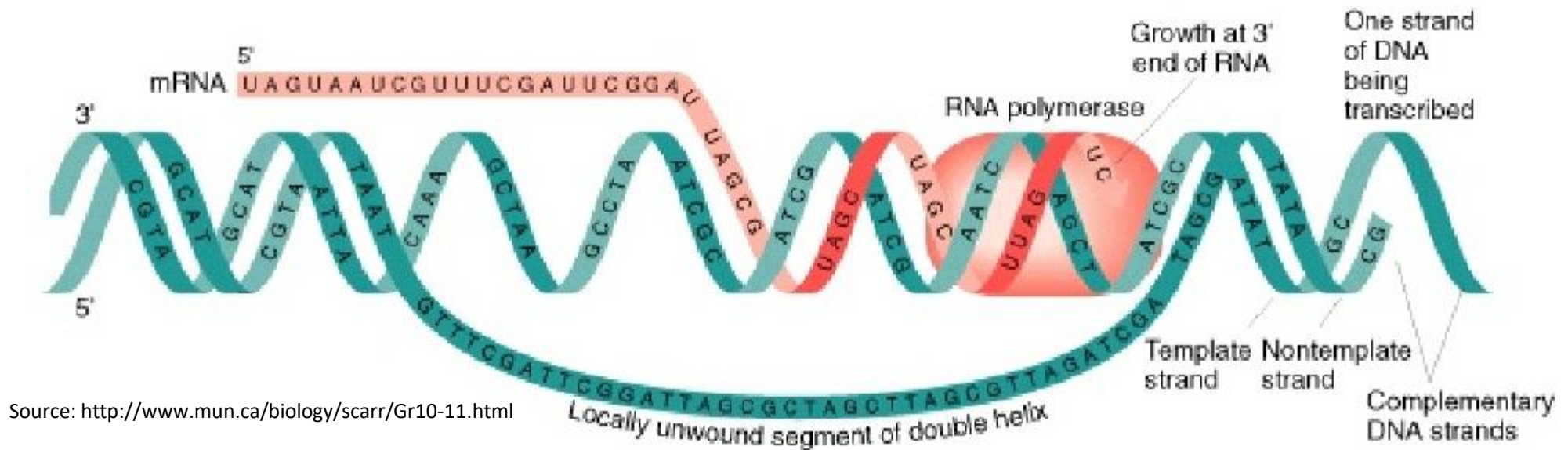
3' - GTTGAAGAAGTCGTCACAGGAAAGAAGTACAAAGAAGTCACTCCCCGGAAT-5'

(Source: *Proc Natl Acad Sci U S A.* 2009 Nov 10;106(45):19096)

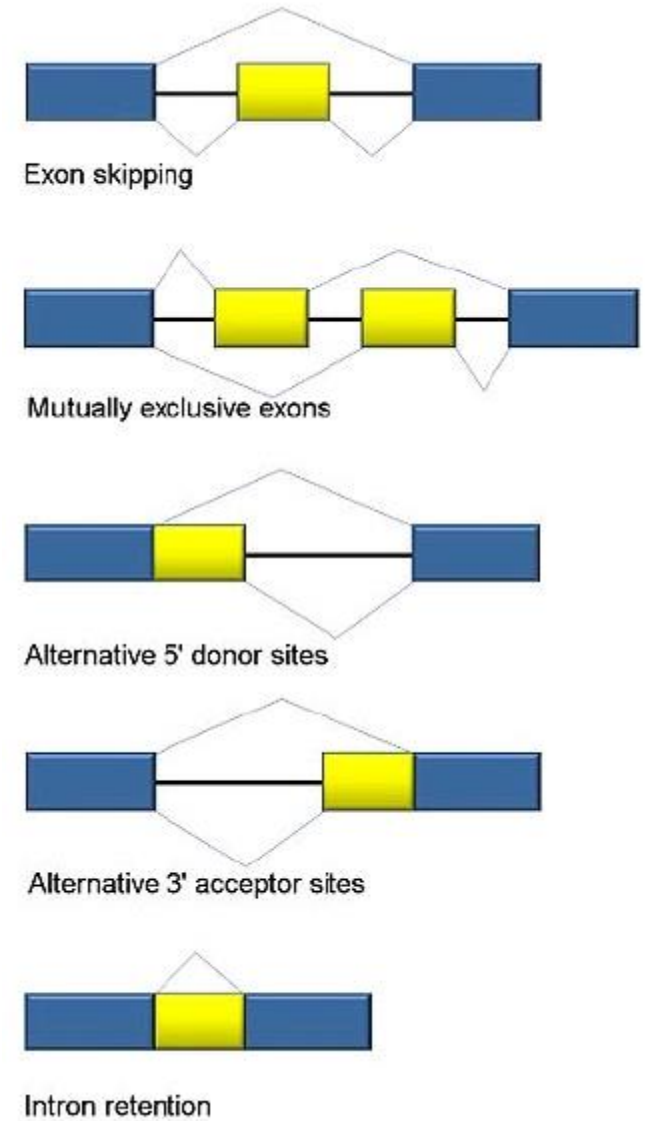
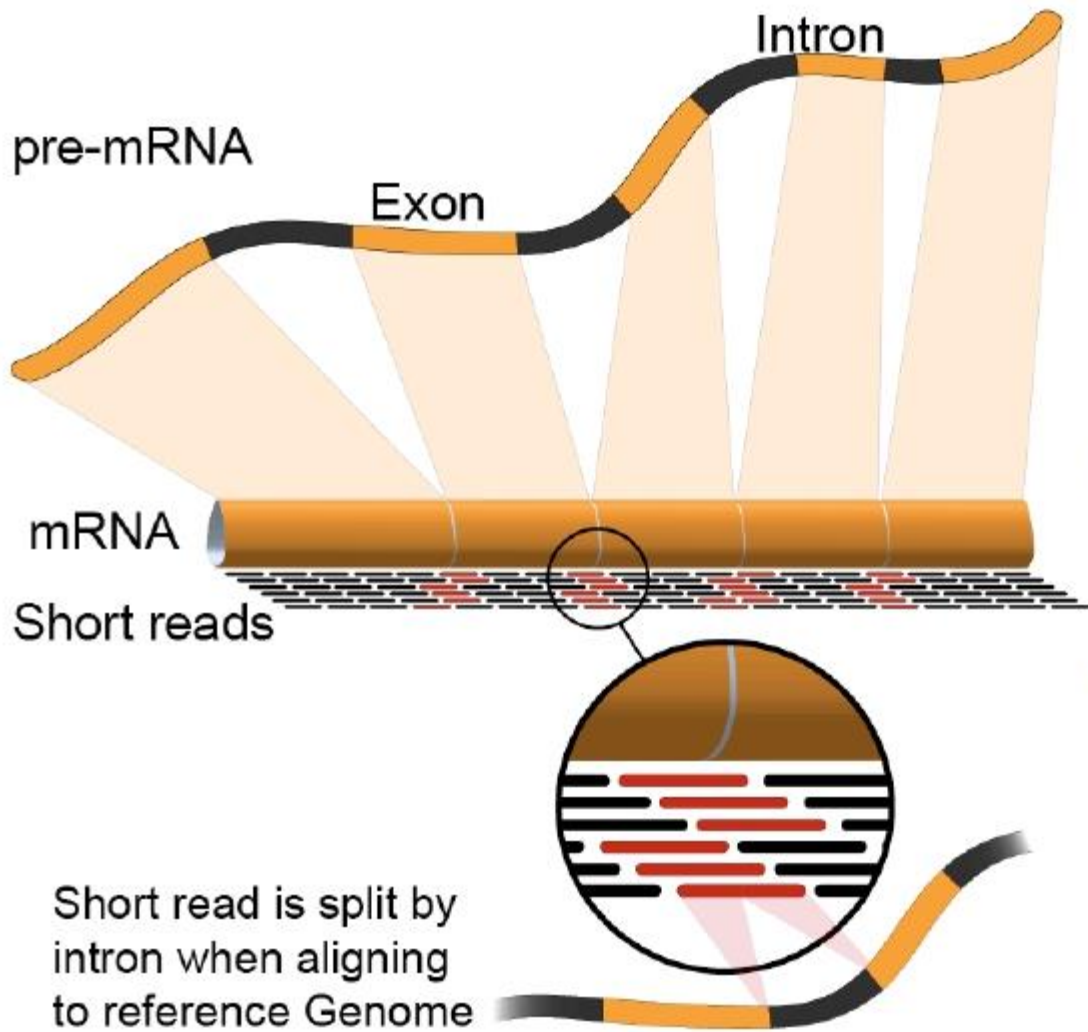


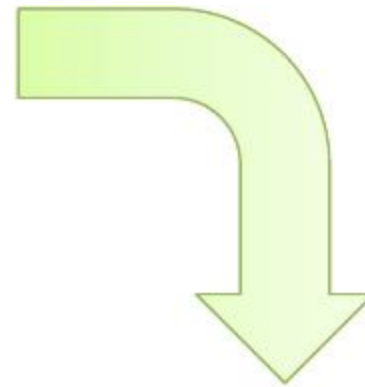
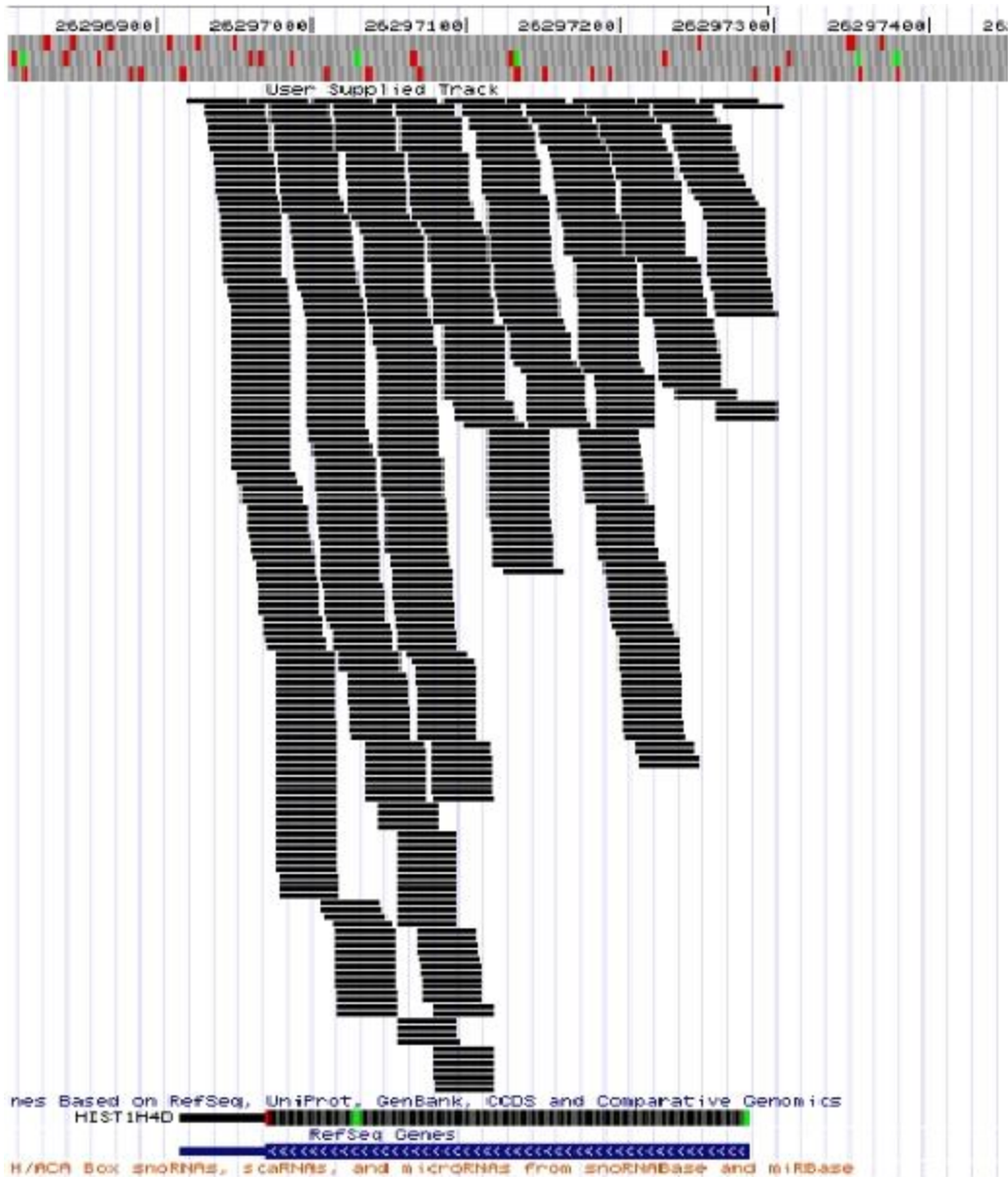
Choi *et al.* used whole-exome sequencing to discover the cause of disease in an individual with an unclear diagnosis. They identified a missense mutation in positions that were highly conserved from invertebrates to humans, in a gene known to cause congenital chloride-losing diarrhoea, consistent with the patient's symptoms.

RNA-Seq: Explore the transcriptome

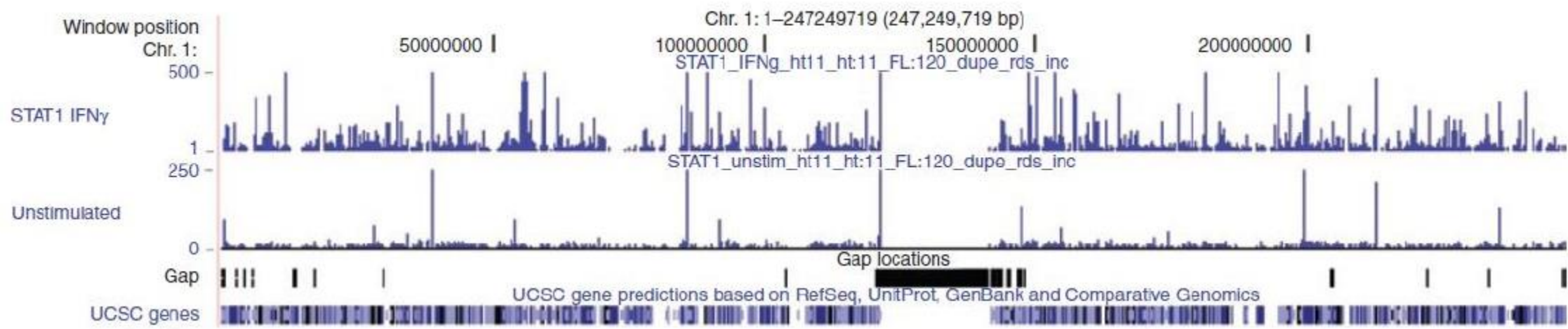


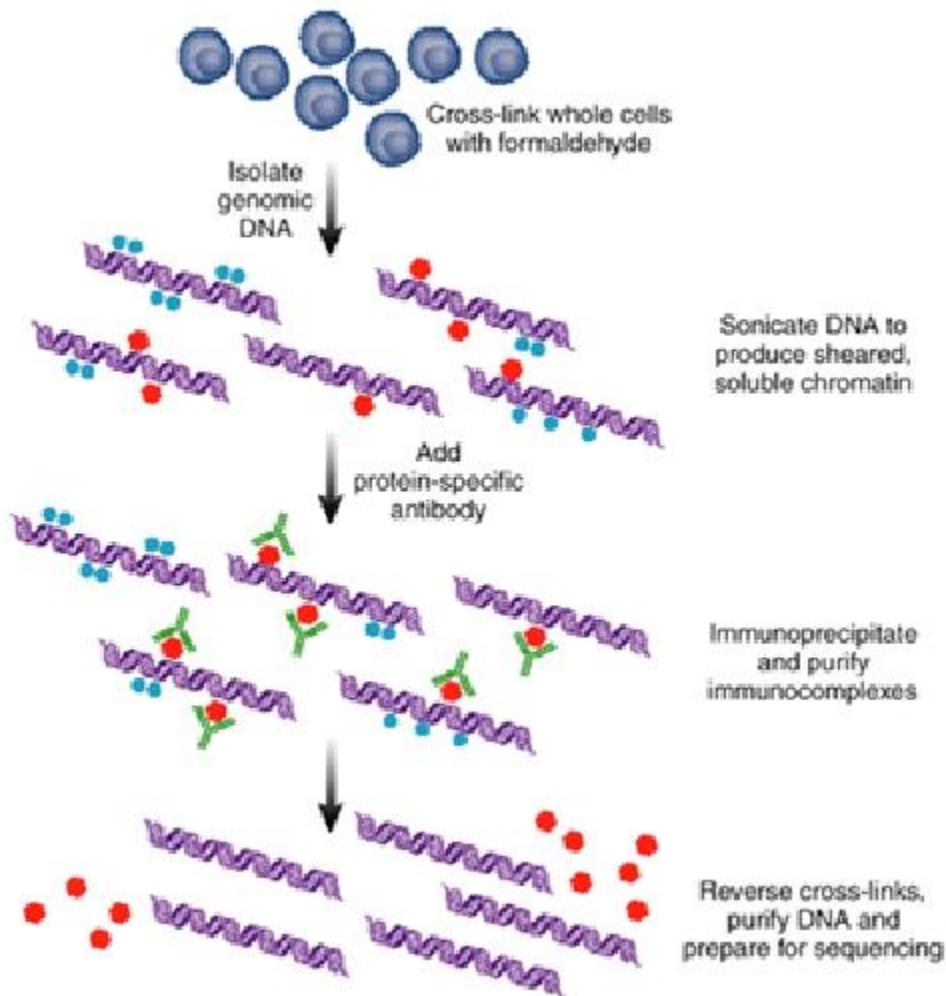
“A transcriptome is a collection of all the transcripts present in a given cell.” (NHGRI factsheet, NIH, US)





	B	C	D	E	F
1	gene	nsc1	nsc1 SE	nsc2	nsc2 SE
2	brain protein	18.9574	3.79952	21.5848	3.02241
3	Cluster Incl AW1	110.513	7.84625	114.894	7.95669
4	Cluster Incl AI8	235.873	35.6748	210.349	27.612
5	Cluster Incl AV3	47.4605	3.94976	29.6941	3.6586
6	Cluster Incl AV1	28.4527	3.74512	15.2986	3.62097
7	Cluster Incl AV1	80.302	6.45368	107.23	8.09591
8	Cluster Incl AV3	40.8113	5.13418	54.0835	3.18591
9	Cluster Incl AI1	53.1437	3.63392	58.635	5.50994

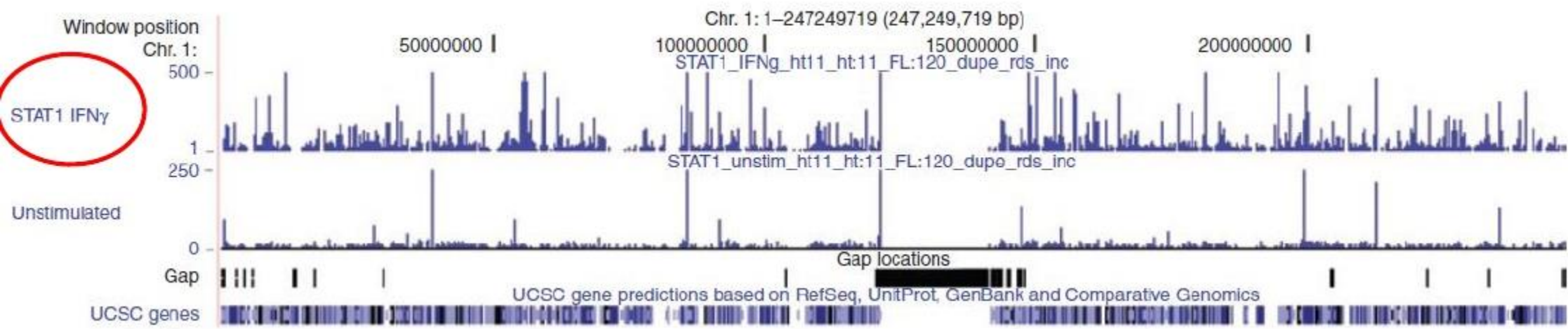




Katrin Rits

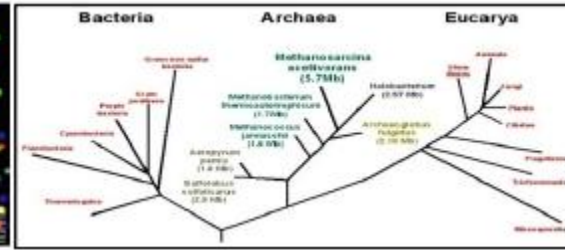
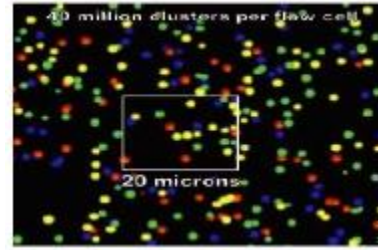
Chromatin ImmunoPrecipita- tion Sequencing (ChIP-Seq):

Profile Protein-DNA interaction





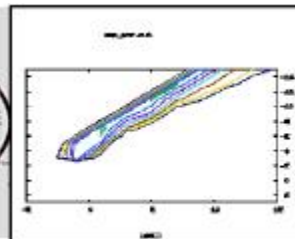
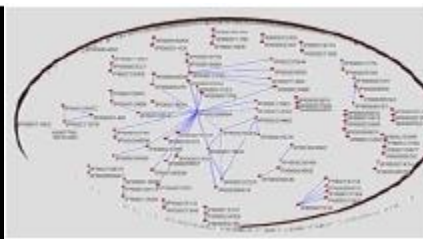
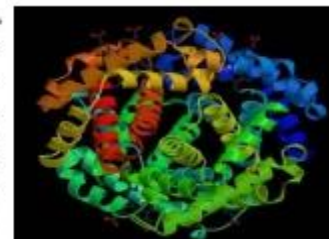
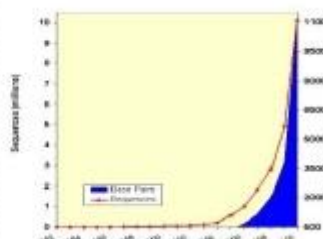
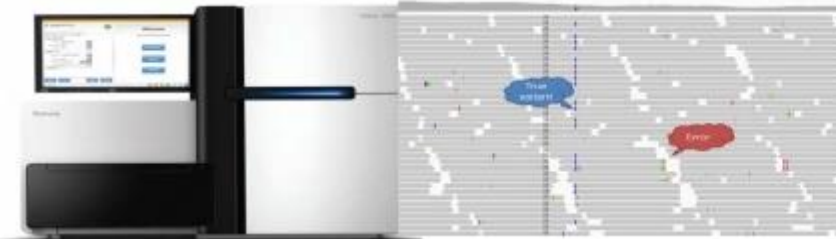
TAACCCTAACCCCTAACCCCTAACCCCTAACCCCTA
 CCTAACCCCTAACCCCTAACCCCTAACCCCTAACCC
 CCTAACCCCTAACCCCTAACCCCTAACCCCTAAC
 AACCCCTAACCCCTAACCCCTAACCCCTAACCCCTA
 ACCCTAACCCCAACCCCAACCCCAACCCCAAC
 CTACCCTAACCCCTAACCCCTAACCCCTAACCCCTA
 ACCCTAACCCCTAACCCCTAACCCCTAACCCCTAA



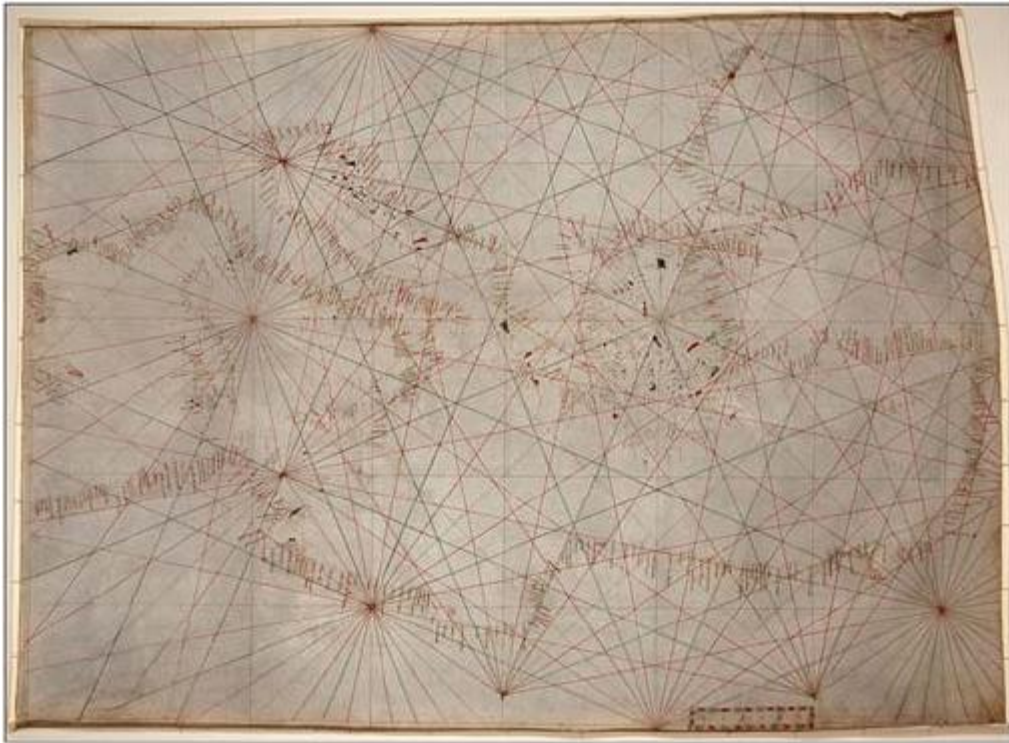
Unit 2: NGS: Reads Mapping

Le Zhang, Ph. D.

Computer Science Department
Southwest University



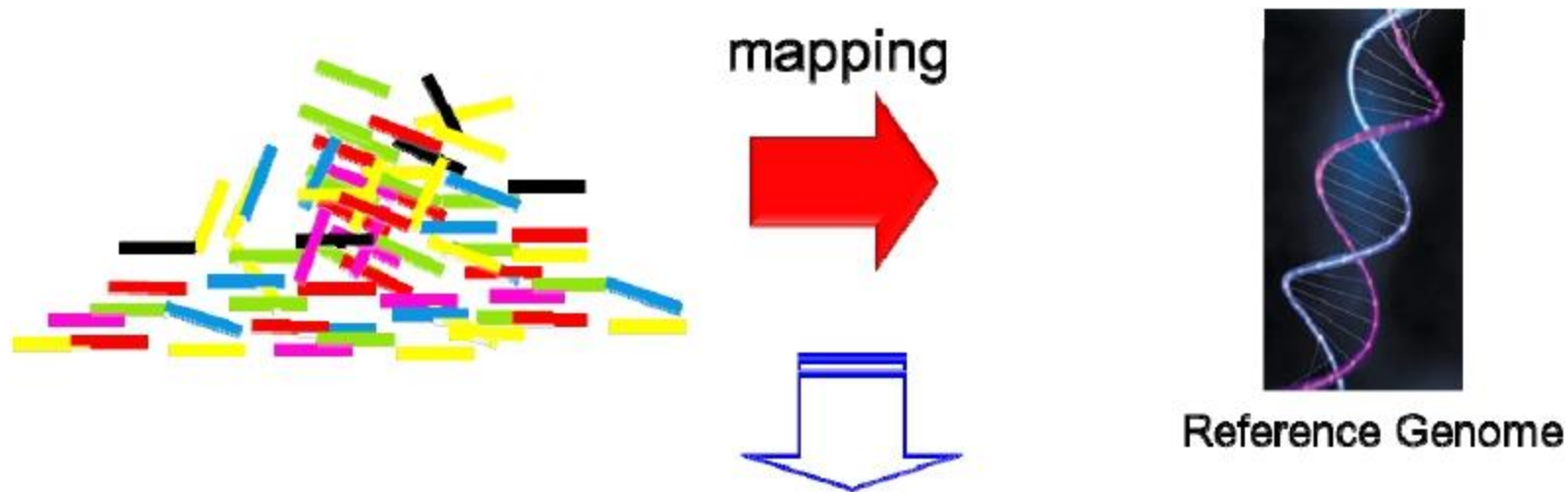
Reads Mapping



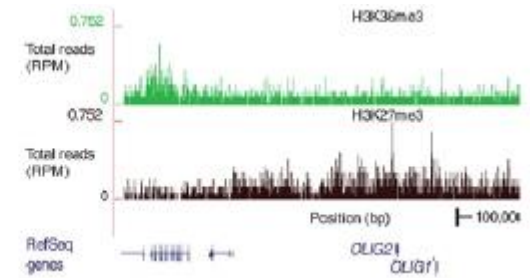
Map-Making / Cartography:
Establish relationship between locations

~~Technological~~: Reads is
usual too short to be
used/assembled *de novo*

~~Scientific~~: Taking full
usage of existing
annotation/knowledge



Calling Genetic Variants



Measuring Abundance: RNA-Seq, ChIP-Seq, etc.

Mapping: Input Data

- Reference Genome

- Nucleotide

- **Length**: Hundreds of Mb *per* chromosome.

- ~3 Gb in total (for human genome)



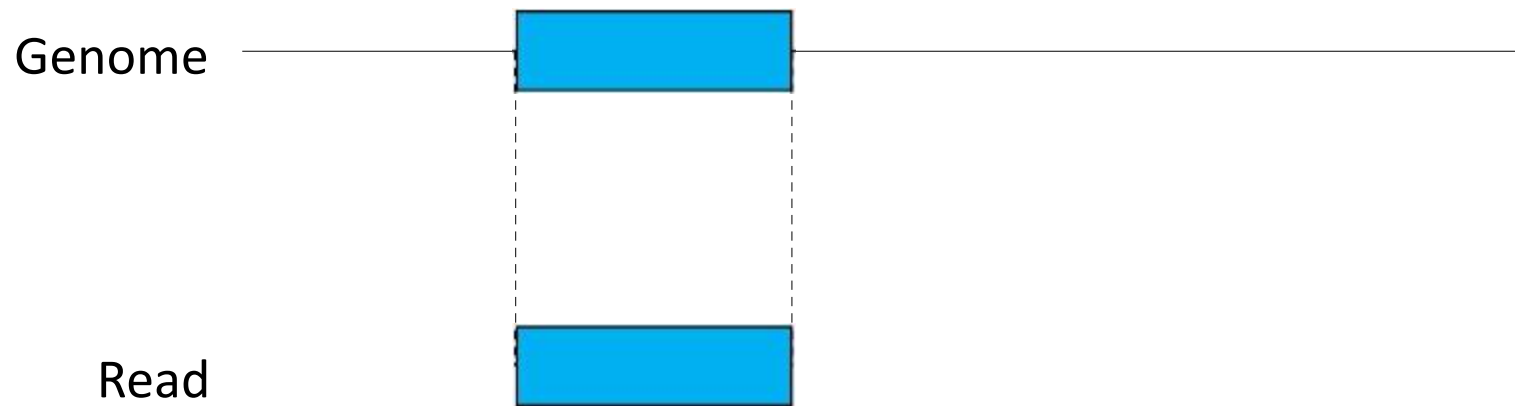
- Reads

- Nucleotide, with **various qualities** (relatively **high error rate**: $1e-2 \sim 1e-5$)

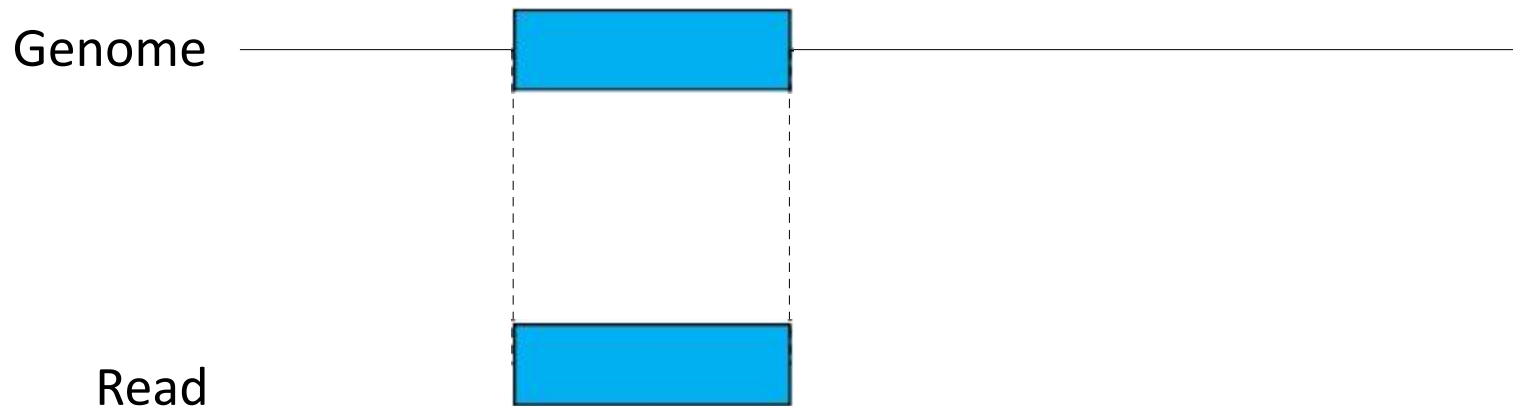
- **Length**: 36~80 bp *per* read

- Hundreds of Gbs *per* run

“Embedded” Alignment

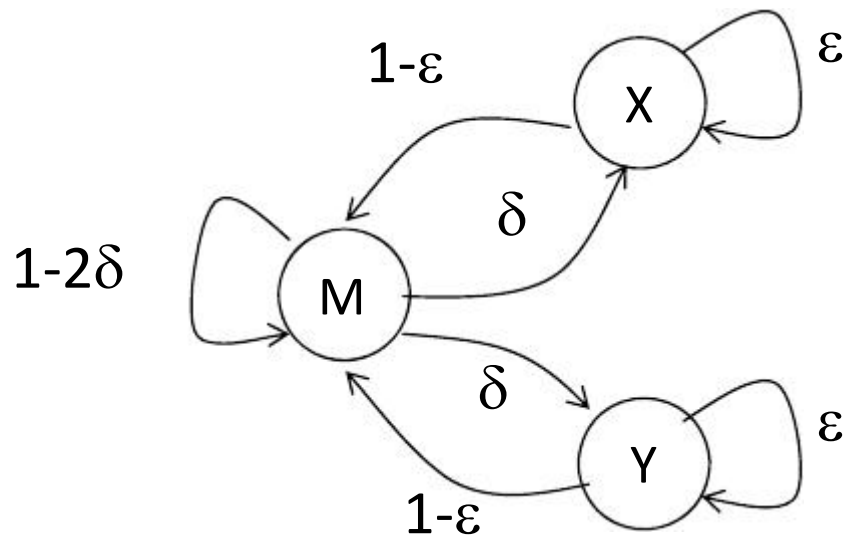


One sequence is “*embedded*” in the other sequence (NGS Reads, PCR primer, *etc.*)



What we need here is actually a hybrid “**global-local**” alignment

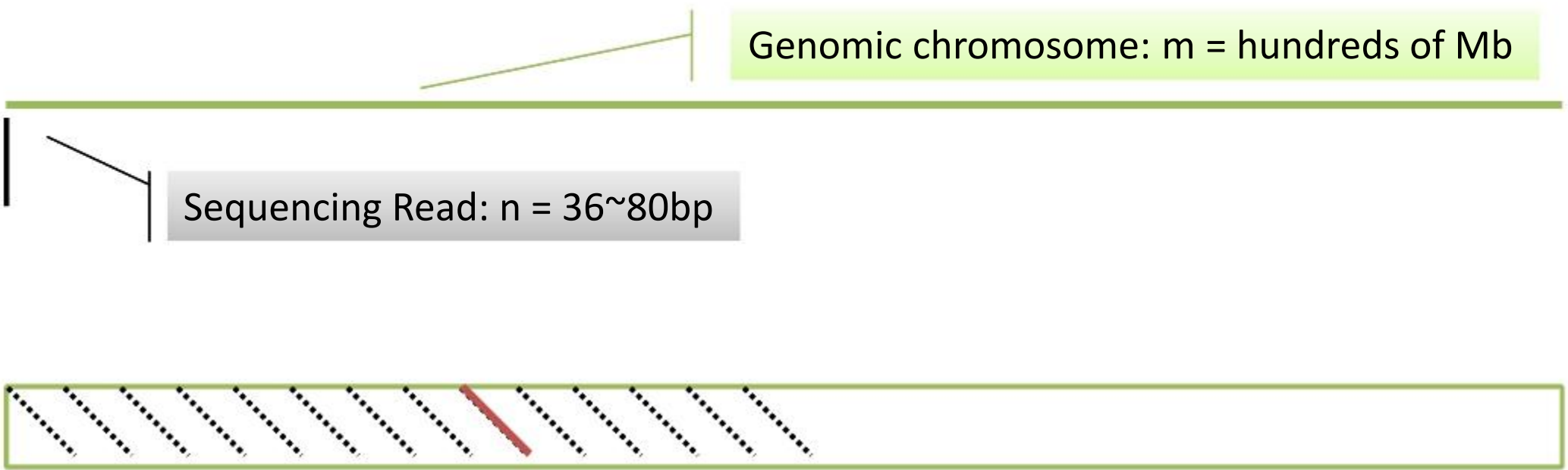
- ✓ “**Global**” for short sequence (i.e. NGS Read)
- ✓ But “**Local**” for long sequence (i.e. Reference Genome)
- ✓ In particular, the surrounding “overhang” gaps should be not penalized.



δ	Gap open
ε	Gap Extension

M	Match
X	Insert at sequence X (delete at sequence Y)
Y	Insert at sequence Y (delete at sequence X)

	M	X	Y
M	1-2δ	δ	δ
X	1-ε	ε	0
Y	1-ε	0	0



Genomic chromosome: $m = \text{hundreds of Mb}$

The diagram shows a long horizontal green bar representing a genomic chromosome. A thin vertical line is drawn across the bar, and a light green box highlights the text 'Genomic chromosome: m = hundreds of Mb' to its right. A thin vertical line is also drawn at the left end of the bar, and a grey box highlights the text 'Sequencing Read: n = 36~80bp' to its right. Below the bar, a series of diagonal lines represent sequencing paths. Most are black dotted lines, but one is a solid red line, indicating a specific path of interest.

Sequencing Read: $n = 36 \sim 80\text{bp}$

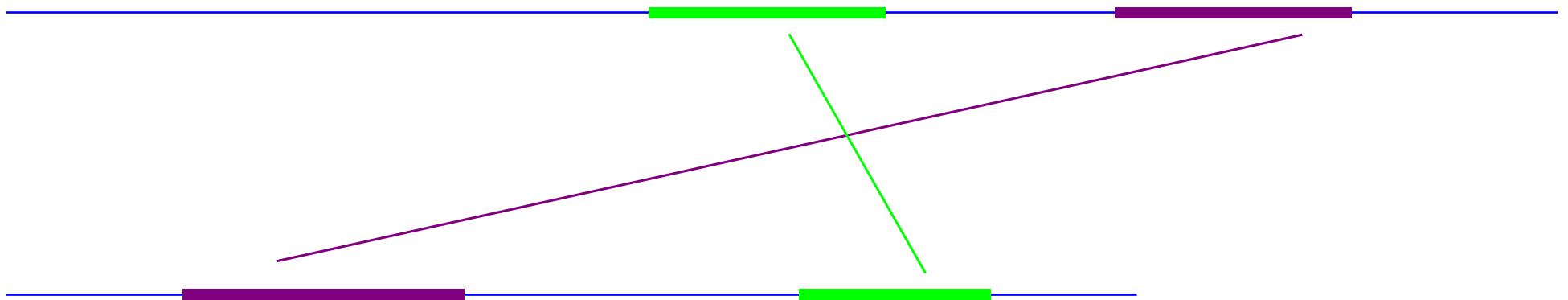
Most of paths will just fail eventually!



In real world, the speed will be a BIG problem!

BLAST Ideas: Seeding-and-extending

1. Find matches (**seed**) between the query and subject
2. Extend seed into High Scoring Segment Pairs (**HSPs**)
 - Run Smith-Waterman algorithm on the specified region only.
3. Assess the reliability of the alignment.



Alphabetical Index

A

Accelerator, 87
Accelerator Pin, 84
Aperture, 91
Arm, Follower, 40

B

Band, Locking, 148
Barrel, 148 -- 177
Barrel and Mount Base Assembly, 147
Barrel, Mount Base & Receiver Assembly, 6

Bolt, 34
Base and Pad Assembly, 21
Base, Altern, 171
Base, Rear Sight, 94
Blank, Rear Sight Screw, 108
Body, 207
Body, Follower Rod, 73
Bolt, 82 -- 83
Bolt Assembly, 79
Bottom Insert, 33
Bracket and Case Assembly Alternative Const., 200

Bracket and Case Assembly, M2, 191
Bracket Hinge Mount, 12
Bracket Knob Mount, 13
Bracket Mount Plunger, 15
Brackets, Flash Hider, 192
Bracket, Flash Hider (Alternative Const.), 201
Bracket, Mount, 11
Bracket, Mount, Assembly, 10
Buckle, 213
Buckle Slide Lock, 254 -- 216
Bullet Guide, 48
Butt Plate, 131

Butt Plate Assembly, 130
Butt Plate Cap, 127
Butt Plate Cap Pin, 129
Butt Plate Long Screw, 131
Butt Plate Plunger, 132
Butt Plate Plunger Spring, 136
Butt Swivel, 140

C

Cap, Butt Plate, 127
Cartridge Ejector, 84
Cartridge Ejector Assembly, 80
Cartridge Ejector Spring, 88
Catch, Operating Rod, 58
Catch, Operating Rod Assembly, 56
Cheek Pad Assembly, 72
Clip Ejector, 145
Clip Latch, 70
Clip Latch Pin, 69
Clip Latch Spring, 71
Clips, End, Strap-Type 1, 2 12
Cone, Flash Hider, 190, 199
Cover, Rear Sight, 95
Cylinder, Gas, 111

D

Detonator Spring, 18
Dog Release, 104
Dog, Indexing, 102
Dog, Indexing, Assm, 102

E

Early Exterior View M1D Sniper Rifle With M82 Telescope, 218
Early Longitudinal Section M1D Sniper Rifle With M82 Telescope, 239
Early Sectionalized Views M1D Sniper Rifle With M82 Telescope, 220
Ejector, Cartridge, 84
Ejector, Cartridge, Assembly, 83
Ejector, Clip, 145
Extractor, 85
Extractor Spring, 89
Extractor Spring Plunger, 87
Extractor Spring Plunger Assembly, 81
Eyelid - Lacking, 33
Eyelids, Metallic, and Eyelid Washers, Metallic, 24 -- 29

F

Finish, Front Hand Guard, 175
Finish, Stock, 128
Firing Pin, 86
Flash Hider, M2, 187
Flash Hider, M2 Alternate Construction, 197
Flash Hider, T 27, 114

Flash Hider Bracket (Alternative Const.), 200
Flash Hider Cone, 190, 199
Flat-External Tooth Lock Washer, 137
Follower, 61
Follower Arm, 60
Follower Arm Pin, 63
Follower Rod Assembly, 71
Follower Rod Body, 73
Follower Rod L.H. Plate, 75
Follower Rod R.H. Plate, 74
Follower Rod Rivet, 76
Follower Slide, 62
Front Hand Guard, 176
Front Hand Guard Assembly, 174
Front Hand Guard Female, 175
Front Hand Guard Spacer, 177
Front Sight, 113
Front Sight Cap Screw, 115

G

Gas Cylinder, 113
Gas Cylinder Lock Screw, 116
Gas Cylinder Lock Screw With Valve Assembly, 120

General Data, Heat Treatment and Finish, 193 -- 196

Geometric Symbols For Dimensioning and Tolerancing, 221
Guard, Hand, Front, Assembly, 174
Guard, Hand, Rear, 176
Guard, Trigger, 146
Guard, Trigger, Assembly, 147
Guide, Bullet, 68

H

Hammer, 130 -- 131
Hammer Pin, 129
Hammer Spring, 142
Hammer Spring Housing, 152
Hammer Spring Plunger, 159
Hammer Stop, 149
Handle & Tube Operating Rod Assembly, 51
Handle, Operating Rod, 31
Head, Valve, 117
Heat Treatments and Finish - General Data, 193 -- 196
Hider, Flash (T 27), 114
Hider, Flash, M2, 187

Hider, Flash, M2 List of Drawings & Specs, 186
Hinge Pin, 14
Hinge, Mount, Bracket, 12
Hook, 210
Hook Assembly, 288
Housing, Hammer Spring, 152
Housing, Trigger, 135 -- 134

I

Indexing Dog, 109
Indexing Dog Assembly, 102
Insert, Bottom, 31
Insert, Middle, 32
Insert, Top, 31

J

K

Keeper Assembly, 265
Knob Mount Screw, 16
Knob, Mount, Bracket, 11
Knob, Rear Sight Elevating, 106
Knobs, Windage, Rear Sight, 97
Knobs, Windage, Rear Sight, Assembly, 96

L

Lace, Leather, 16
Lacing Eyelets, 30
Latch, 198, 199
Latch, Clip, 70
Latches, 198, 38
Lather, Gun Hides, Lace, and Cut - Lace, 37 -- 41
Lack Washer, 137
Lock, Rear Sight Nut, 100
Loop, 209
Loops, Slide (For Equipment), 214 -- 216
Lower Band, 180
Lower Band Pin, 179
Lug Locking, Trigger Guard, 146

M

M1 Gun Ring Alternative Design, 204
M2 Bracket and Case Assembly, 191
M2 Flash Hider, 187
M2 Flash Hider Bracket, 191

M2 Flash Hider List of Drawings & Specs, 185 (Sheet 1 of 2), 197 (Sheet 2 of 2)
M2 Alternate Construction
M2 Flash Hider Spring Pin, 189
M84 Telescope, 181 -- 183
M84 Telescope Optical System, 184
Middle Insert, 32
Mount Base, 172
Mount Base & Receiver Barrel Assembly, 6
Mount Bracket, 11
Mount Bracket Assembly, 10
Mount Locking Pin, 173
Mount, Telescope, 8
Nut, Rear Sight, 99

O

Operating Rod Assembly, National Match (N.M.), 99
Operating Rod Assembly, 50
Operating Rod Catch, 53
Operating Rod Catch Assembly, 56
Operating Rod Handle, 52
Operating Rod Spring, 55
Operating Rod Tube, 53

P

Pad, 35
Pad, Check, Assembly, 32
Parts Identified By Part Number, M1, M1C, M1D 506, 222
Pin, Accelerator, 89
Pin, Butt Plate Cap, 129
Pin, Clip Latch, 69
Pin, Firing, 86
Pin, Follower Arm, 63
Pin, Hammer, 155
Pin, Release, 134
Pin, Spring, 179
Pin, Trigger, 157
Pin, Trigger (Superseded), 158
Pinion Assembly, Rear Sight Elevating, 107, 110
Pinion, Elevating, Rear Sight, 103
Pinion, Operating Rod, 54

Clayville, OH: 170
Clayville, Ohio: 177
Mt Pleasant: 153

Presbytery
Masking: 236

Socader Church
Clayville/New Concord, OH: 409

Crossroads: 277
Socaders: vii, viii, 141, 148, 149, 150, 153, 154, 156, 162, 167, 169, 170, 171, 172, 176, 184, 185, 186, 189, 190, 195, 196, 197, 198, 199, 201, 204, 210, 211, 212, 210, 231, 236, 237, 238, 239, 240, 241, 242, 243, 258, 256, 257, 258, 260, 261, 264, 265, 266, 268, 297, 323, 411, 416, 458, 506, 659, 799, 804, 806, 802, 814, 826, 843

and Tennessee: 179
Iowa: 237
North Carolina: 186
Ohio: 185, 186, 201, 231, 238
Peanut: 238
South: 148, 236

Washington: 627, 628, 630, 632, 634
Texas: 847

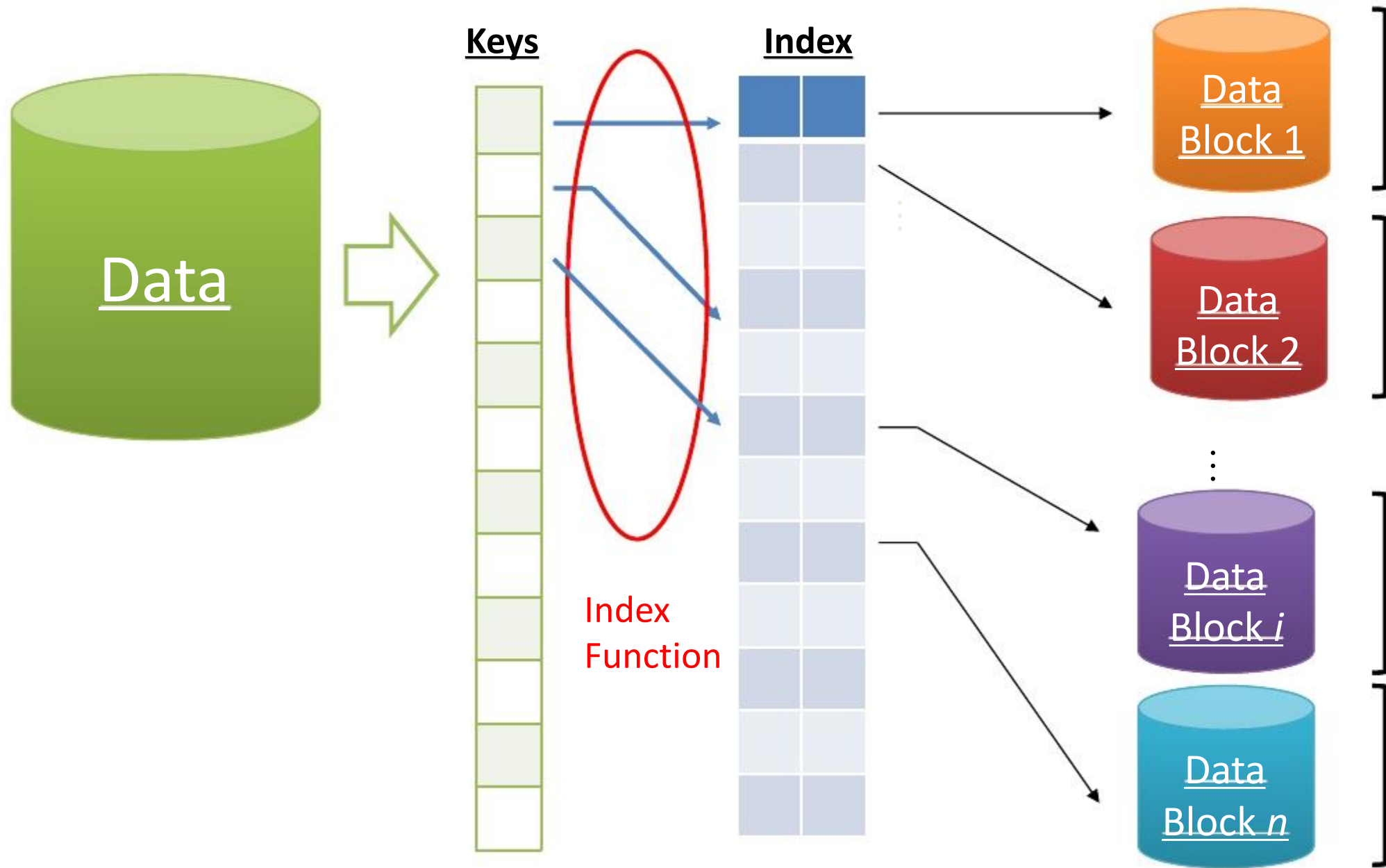
Sharon College: 512, 804
Shelbyville, KY: 399, 438, 520
Shenandoah Valley: 132, 314, 469, 471, 473, 629, 665, 668, 684, 693, 697, 698, 780, 847

Shenandoah Valley Campaign: 671, 684
Sheridan, Phil: 296, 319, 431, 523, 537, 565, 615, 684, 699, 796, 806, 807, 836
Washington: 400, 426, 430, 431, 432, 526, 542, 599, 620, 622, 645, 646, 665, 666, 667, 671, 683, 684, 693, 697, 698, 699, 700, 765, 769, 778, 789

Sherrin, William Tecumseh: 343, 348, 360, 361, 363, 375, 378, 379, 389, 400, 408, 499, 539, 543, 548, 565, 566, 567, 603, 616, 677, 687, 702, 734, 736, 747, 785, 787, 796, 798, 803, 836

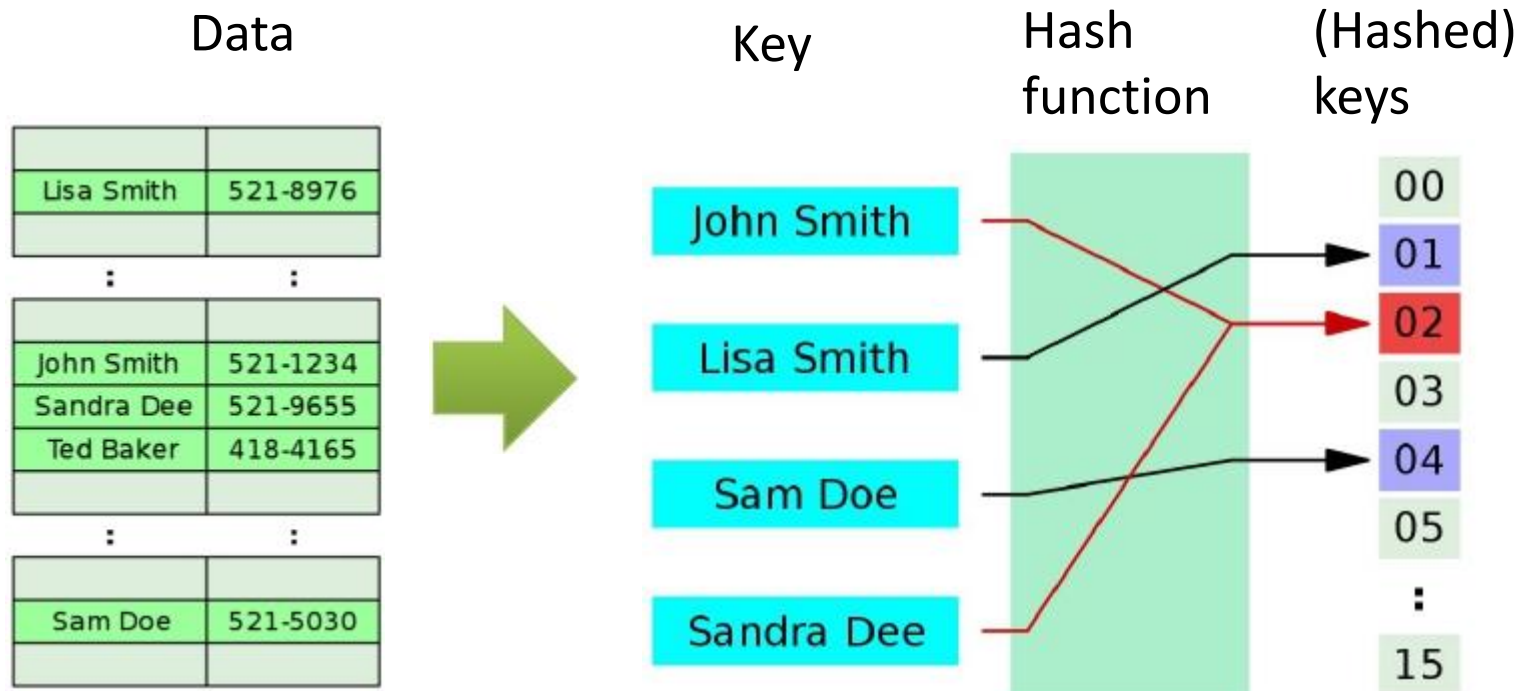
family: 675
March to the Sea: 697

Washington: 369, 370, 371, 432, 453, 460, 461, 463, 489, 537, 538, 539, 580, 581



Hash

Hash function maps (partial) data into (hashed) keys for following-up indexing



HBS: A naive hash function

Let's assume: A = 1, C = 2, G = 4, T = 8, then:

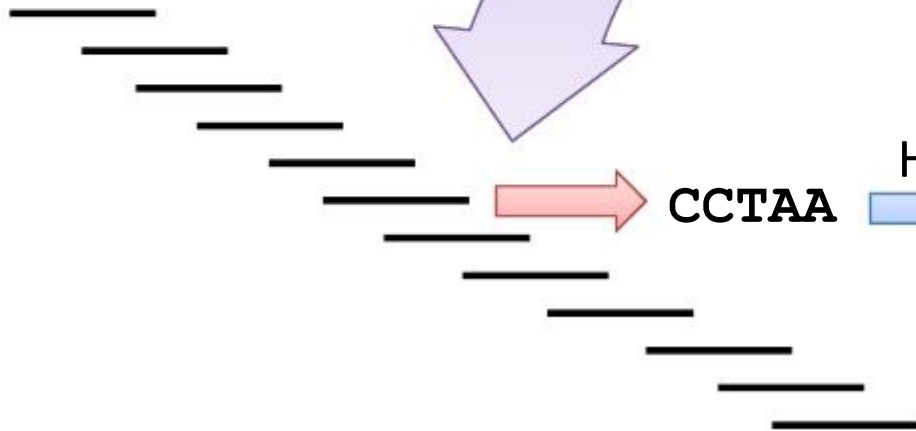
$$HBS(AAAAA) = 1 + 1 + 1 + 1 + 1 = 5$$

$$HBS(GTACG) = 4 + 8 + 1 + 2 + 4 = 19$$

...

() Σ , e.g:

123456789012345678901234567890
TAACCCTAACCCTAACCCTAACCCTAACC



HBS

$$2+2+8+4+4 \\ =20$$

Index Table	
...	
20	
...	

Address Table
(CCTAA, 11)
...

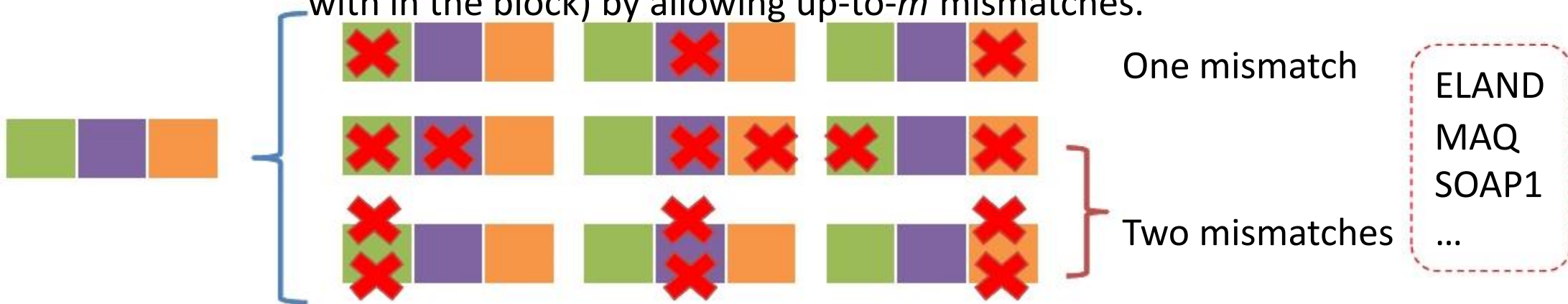
Pigeonhole principle (抽屜原理)

“In mathematics, the pigeonhole principle states that if n items are put into m pigeonholes with $n > m$, then at least one pigeonhole must contain more than one item.”

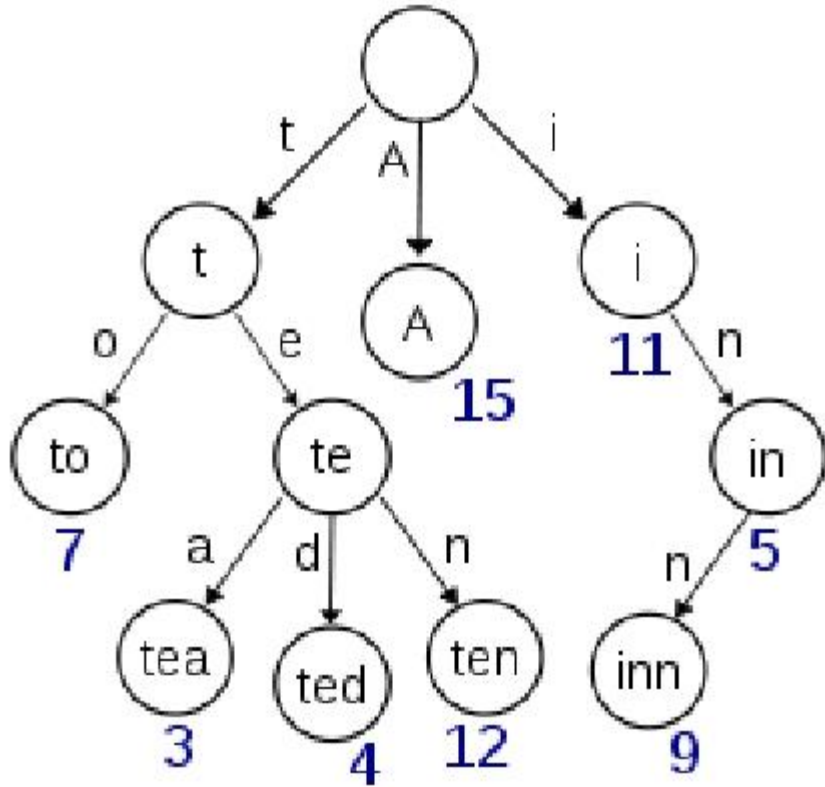


http://en.wikipedia.org/wiki/Pigeonhole_principle

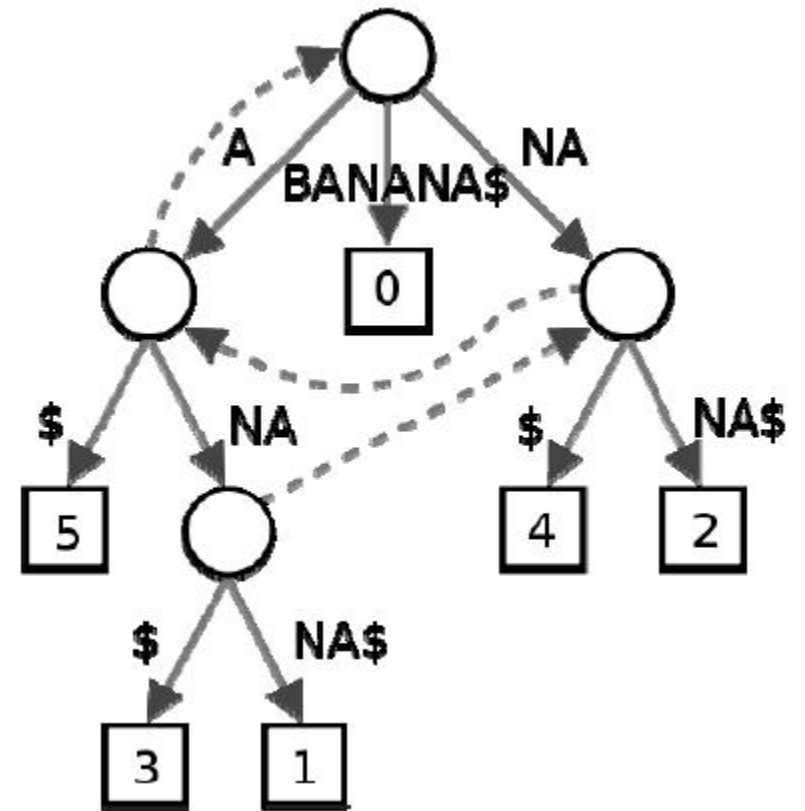
After splitting the read into n (non-overlapped) blocks, there will be **at least $n-m$ perfectly-matched blocks** (i.e. without any mismatch with in the block) by allowing up-to- m mismatches.



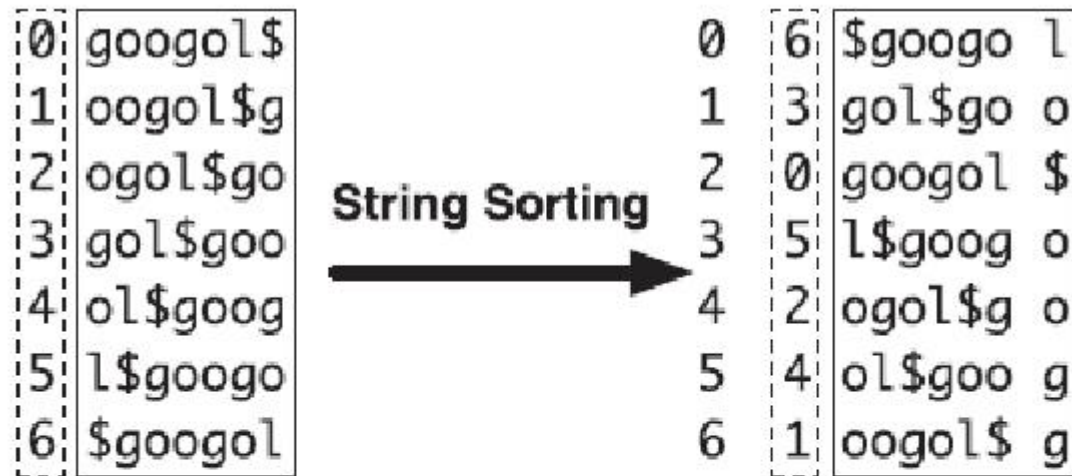
Prefix Tree



Suffix Tree



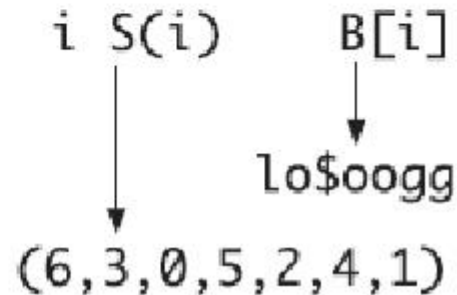
Burrows–Wheeler transform (BWT)



Pos

X = googol\$

(Li H, et. al, Bioinformatics, 2009)

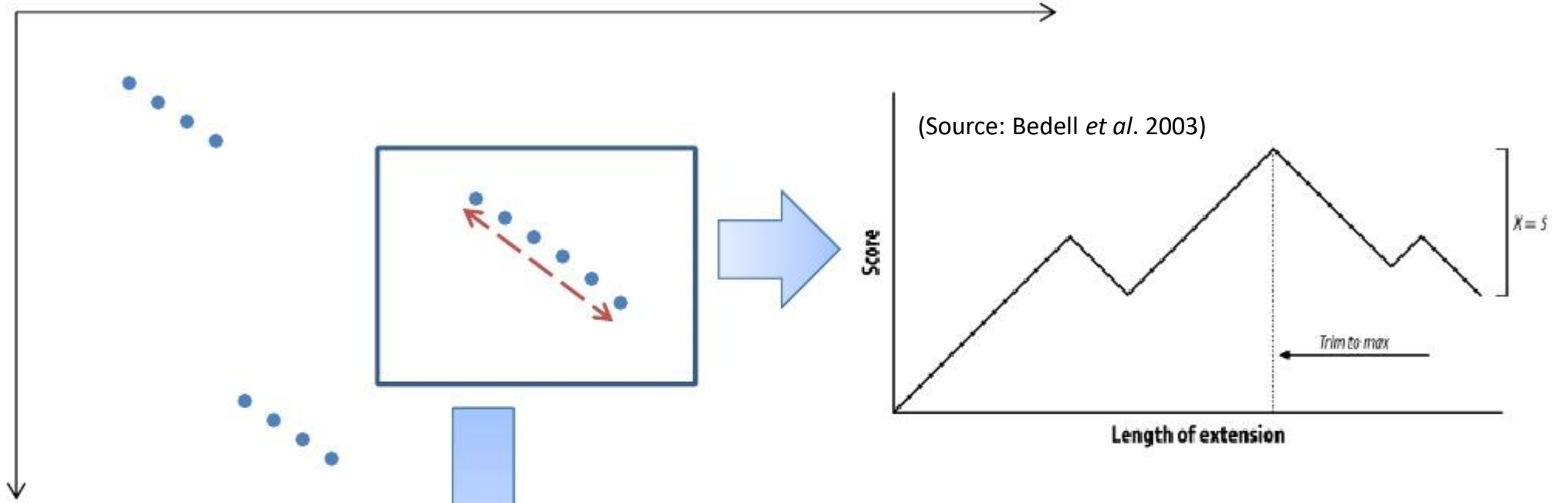


BOWTIE
BWA
SOAP3

...

One of candidate sequence

Query sequence

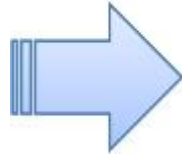


$$F(0,0) = 0$$

$$F(i,j) = \max \begin{cases} F(i-1, j-1) + s(x_i, y_j) \\ F(i-1, j) + d \\ F(i, j-1) + d \\ 0 \end{cases}$$

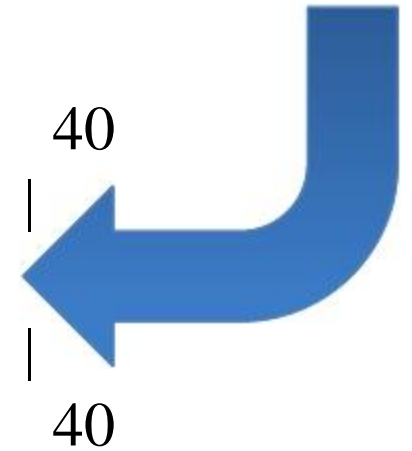
Quality: Given p = the probability of a base calling is *wrong*, its Quality Score can be written as

$$Q = -10 * \log_{10}(p)$$



p	Q
0.1	10
0.01	20
0.001	30
0.0001	40

0 10 20 30
 | | | |
 !"#\$%&'()*+,-./0123456789:;<=>?@ABCDEFGHI
 | | | |
 0 10 20 30



Mapping Quality

Given reference sequence z (length L), a read sequence x (length l), u is the alignment position of x on z , the probability that z actually coming from the position u is $p(z|x,u)$

(Genome Res. 2008 Nov;18(11):1851.)

$$p(z | x, u) = \prod_{\text{mismatch}} p(z_i) \quad SQ(u) = \log(p(z | x, u)) = \sum_{\text{mismatch}} p(z_i) = \sum_{\text{mismatch}} Q(z_i)$$

Read: **ACGT** (Quality: 30 30 25 20)

Ref: **ACGTACGGA**

ACGT	0+	0+	0+	0	SQ(0)
ACGT	30+	30+	25+	20	SQ(1)
ACGT	30+	30+	25+	20	SQ(2)
ACGT	30+	30+	25+	20	SQ(3)
ACGT	0+	0+	0+	20	SQ(4)
ACGT	30+	30+	0+	20	SQ(5)

Mapping Quality

If we assume that a **uniform NULL model**, i.e. the read can randomly come from all possible positions with equal probability, then the error of mapping to a specified position u could be written as

$$E(u) = \frac{SQ(u)}{\sum_i SQ(i)}$$

(Genome Res. 2008 Nov;18(11):1851.)

Read: ACGT (Quality: 30 30 25 20)

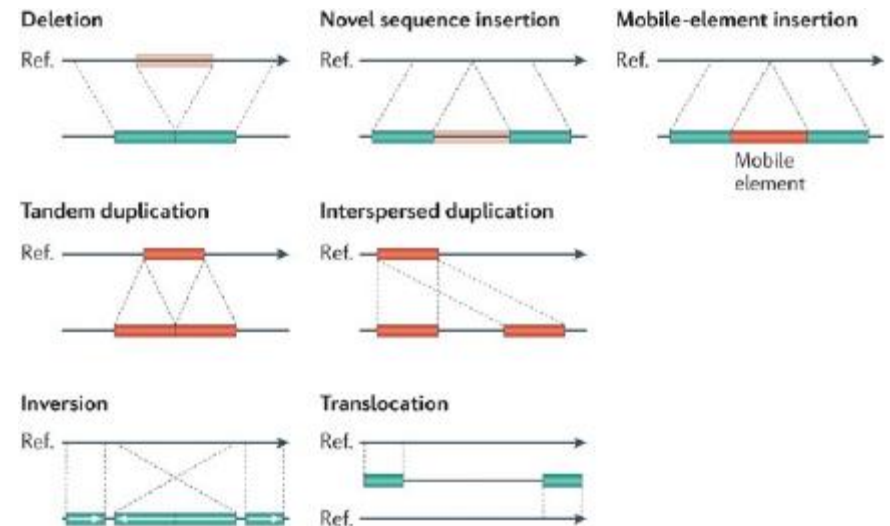
Ref: ACGTACGGA

		<u>SQ(u)</u>	<u>E(u)</u>
ACGT		0+ 0+ 0+ 0	0/415
ACGT		30+30+25+20	105/415
ACGT		30+30+25+20	105/415
ACGT		30+30+25+20	105/415
ACGT		0+ 0+ 0+20	20/415
ACGT		30+30+ 0+20	80/415

Genetic Variants

- SNV: Single Nucleotide Variant
 - Substitution (SNP)
 - Indel: insertion/deletion

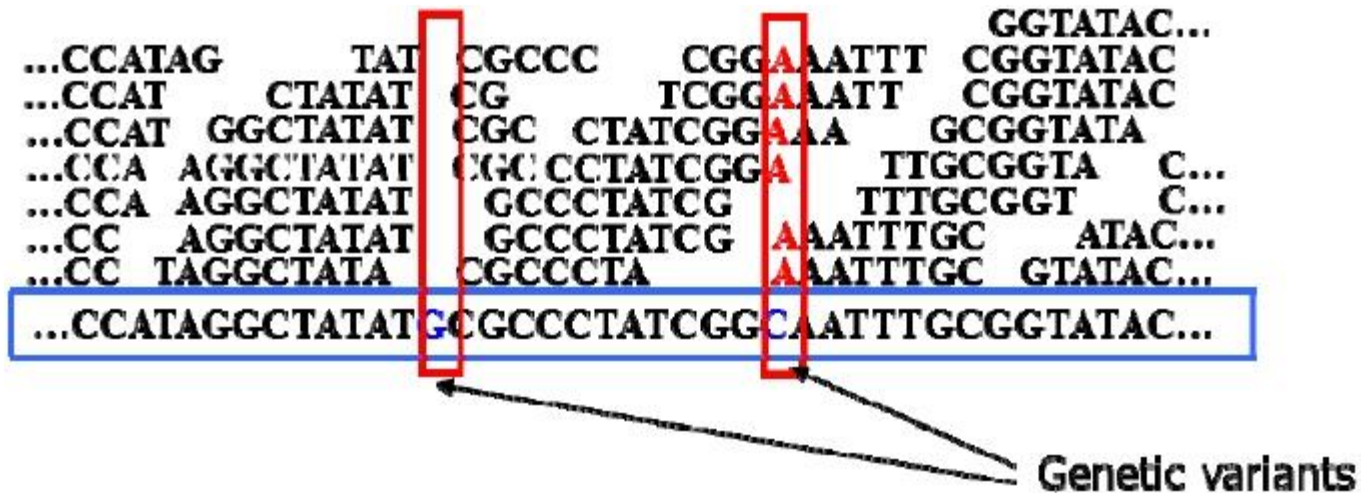
- Structural Variation (SV)
 - Large-scale insertion/deletion
 - Inversion
 - Translocation
 - Copy Number Variation (CNV)



SNP Calling is NOT Genotyping

- “**SNP calling** aims to determine in which **positions** there are polymorphisms or in which **positions** at least one of the bases differs from a reference sequence”
- “**Genotype calling** is the process of determining the **genotype** for each individual and **is typically only done for positions in which a SNP or a 'variant' has already been called.**”

Counting: an intuitive (and naïve) approach



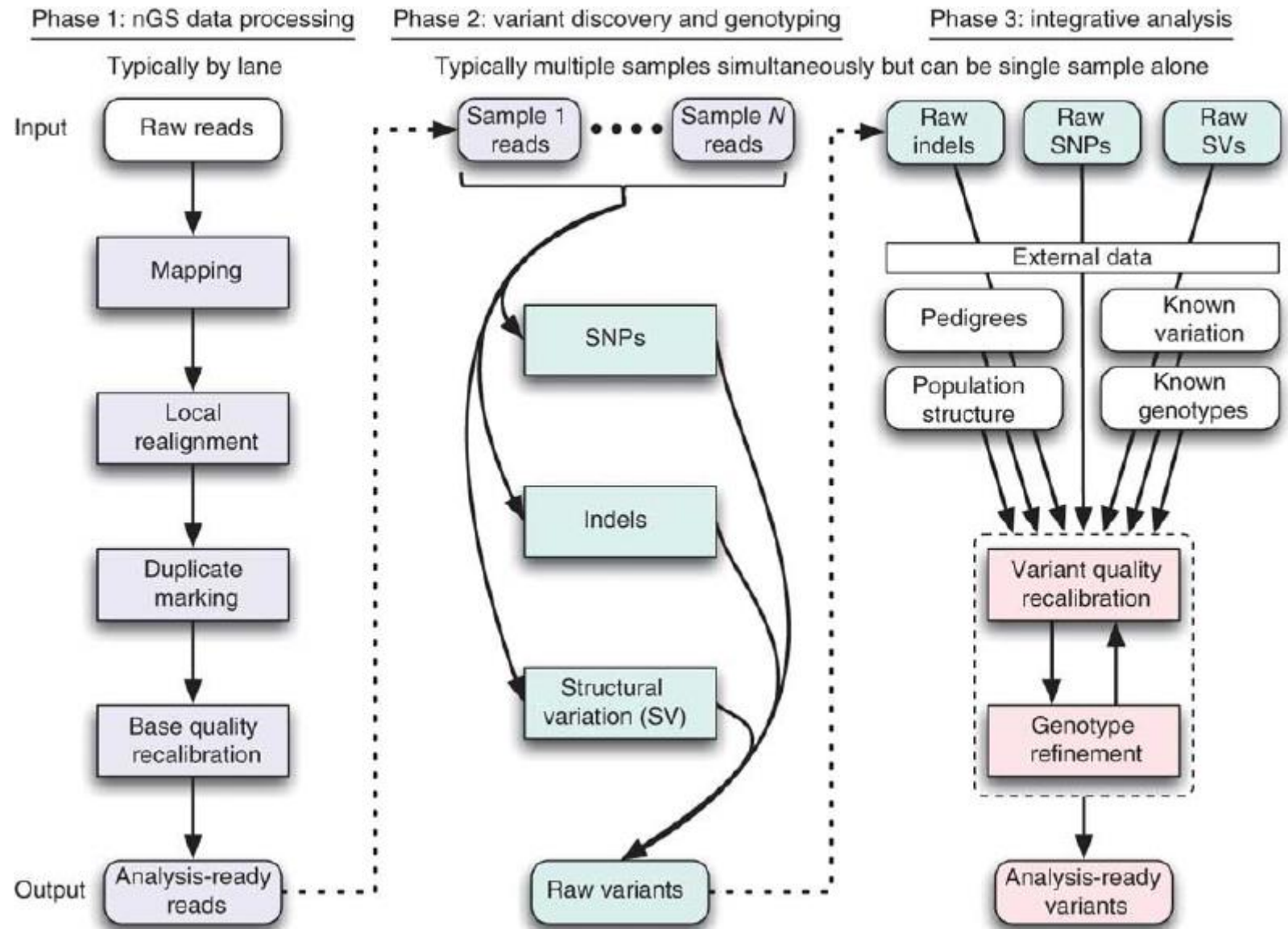
- Counting **high-confident** , **non-reference** allele (i.e. Quality ≥ 20)
 - Freq $< 20\%$ or $> 80\%$: **homozygous** genotype
 - Otherwise: **heterozygous**
- Works well for “**deeply sequenced regions**” (DSR), i.e. depth $> 25x$
 - But suffer from under-calling of heterozygous genotypes for low-coverage regions
 - And can't give an objective measurement for **reliability**

Reviews Genetics 12, 443-451)

A Simple Probabilistic Model for Genotyping

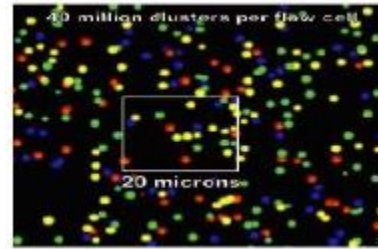
1. For a diploid genome, there will be at most two different alleles (A and a) observed at a given site:
 - 3 possible genotypes: $\langle A,A \rangle$, $\langle A,a \rangle$, $\langle a,a \rangle$
 - Number of A: k; Number of a: n-k
2. Then, the probability for each genotypes is
 - $P(D|\langle A,A \rangle)$ = the probability that we have (n-k) sequencing errors at this site
 \prod
 - Similarly, we can see the $P(D|\langle a,a \rangle) = \prod$
 - $P(D|\langle A,a \rangle) = 1 - (P(D|\langle A,A \rangle) + P(D|\langle a,a \rangle))$
3. Bayes Formula can be further employed to calculate posterior probabilities, i.e. $P(\langle A,A \rangle | D)$, $P(\langle a,a \rangle | D)$, and $P(\langle A,a \rangle | D)$ if we can estimate the prior probabilities $P(\langle A,A \rangle)$, $P(\langle a,a \rangle)$ and $P(\langle A,a \rangle)$

Genome Analysis ToolKit (GATK)





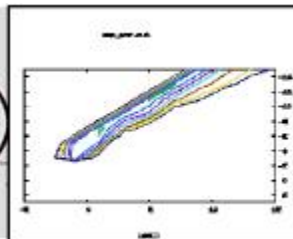
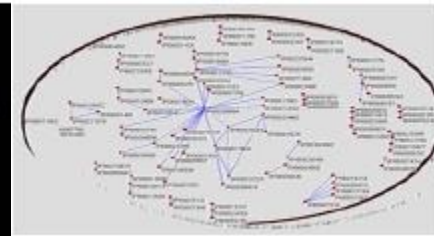
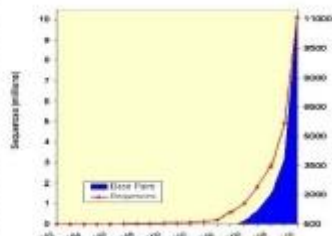
```
TAACCCTAACCCCTAACCCCTAACCCCTAACCCCTA
CCTAACCCCTAACCCCTAACCCCTAACCCCTAACCC
CCCTAACCCCTAACCCCTAACCCCTAACCCCTAAC
AACCCCTAACCCCTAACCCCTAACCCCTAACCCCTA
ACCCTAACCCCAACCCCAACCCCAACCCCAACCCCAAC
CTACCCTAACCCCTAACCCCTAACCCCTAACCCCTA
ACCCTAACCCCTAACCCCTAACCCCTAACCCCTAA
```



Unit 3: BWA & BWT algorithm

Le Zhang, Ph. D.

Computer Science Department
Southwest University



Outline

- BWA & BWT algorithm
- Variant caller
 - samtools
 - GATK

BWA / BWT algorithm

- The compression algorithm used in BWA
- Lossless compression
- Sort and transform the char matrix with string rotation
- Reverse-char method was utilized for match
- Cannot handle gap

ACTACGG

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

sort
→

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

TGAAGCC
I=2

TGAAGCC

l=2

L	F
T	A
G	A
A	C
A	C
G	G
C	G
C	T

sort →

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

F	L	
A	T	5
A	G	1
C	A	4
C	A	7
G	G	2
G	C	3
T	C	6

ACTACGG

→ TACG

F L

F L

F L

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

A	T	
A	G	1
C	A	
C	A	
G	G	2
G	C	
T	C	

A	T	4
A	G	
C	A	3
C	A	
G	G	1
G	C	2
T	C	

ACTACGG

||||

TACG

ACTACGG

||

AC

||

GG

→ GGAC?

ACTACGG → ACTACGG\$

F L

\$	G	
A	T	
A	\$	3
C	A	
C	A	2
G	G	
G	C	
T	C	1

→ GGAC

Variant caller

- samtools
 - mpileup + bcftools
- GATK
 - UnifiedGenotyper
 - HaplotypeCaller

GATK

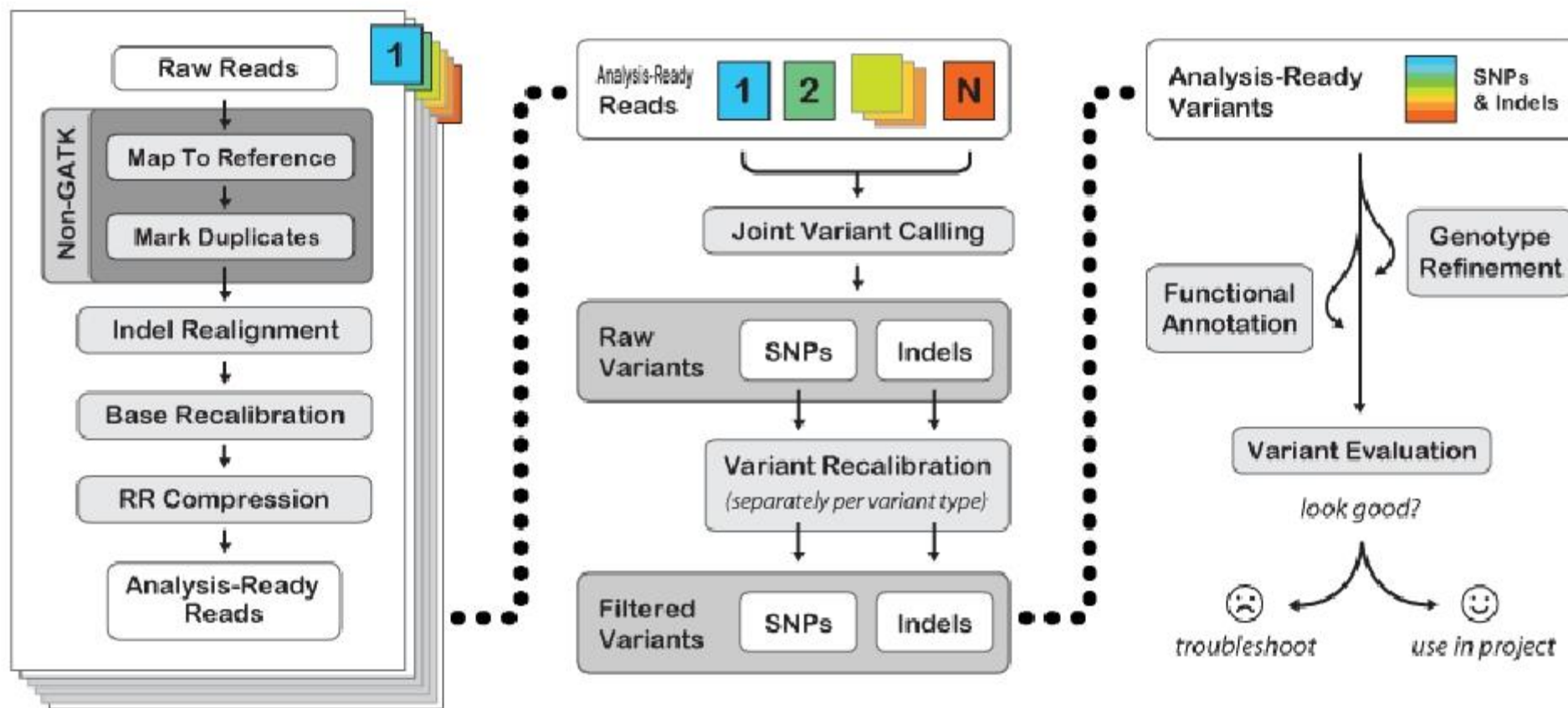
Data Pre-processing

>>

Variant Discovery

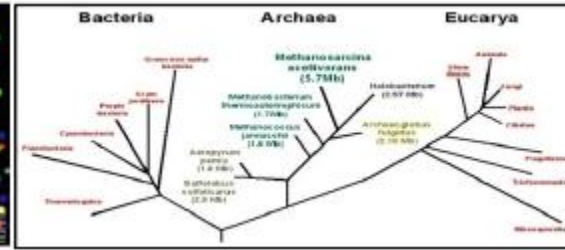
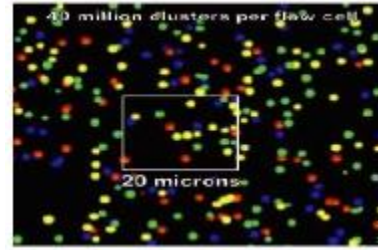
>>

Preliminary Analyses



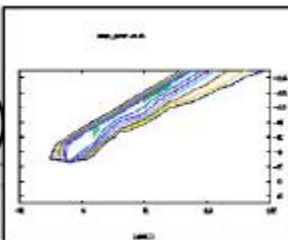
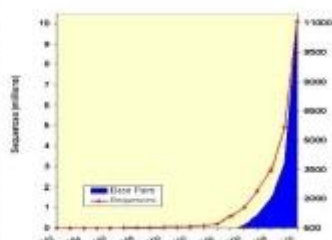


TAACCCTAACCCCTAACCCCTAACCCCTAACCCCTA
 CCTAACCCCTAACCCCTAACCCCTAACCCCTAACCC
 CCCTAACCCCTAACCCCTAACCCCTAACCCCTAAC
 AACCCCTAACCCCTAACCCCTAACCCCTAACCCCTA
 ACCCTAACCCCAACCCCAACCCCAACCCCAAC
 CTACCCTAACCCCTAACCCCTAACCCCTAACCCCTA
 ACCCTAACCCCTAACCCCTAACCCCTAACCCCTAA



Unit 4: Likelihood and Bayesian approach

Le Zhang, Ph. D.
 Computer Science Department
 Southwest University



Outline

- Introduction of Likelihood and Bayesian approach
- Genotyper of MAQ and SNVMix

Likelihood & Bayesian

- Likelihood function
 - a function of the parameters of a statistical model
 - $L(\theta) = P(\text{Data} | \theta)$
- Bayesian approach
 - $P(\theta | \text{Data}) \propto P(\theta) * P(\text{Data} | \theta)$
 - posterior \propto prior * likelihood

A Simple Demostration

- Toss a biased coin, let $\theta = P(\text{Head})$ in one trial
- Probability for seeing HTHH?

$$\begin{aligned}L(\theta) &= P(\text{Data}|\theta) = P(\text{HTHH}|\theta) \\ &= \theta \cdot (1 - \theta) \cdot \theta \cdot \theta = \theta^3(1 - \theta)\end{aligned}$$

Bernoulli distribution

- Probability for seeing 3 Heads in 4 trials?

$$\begin{aligned}L(\theta) &= P(\text{Data}|\theta) = P(3H \text{ in } 4|\theta) \\ &= \binom{4}{3} \theta^3(1 - \theta)\end{aligned}$$

binomial distribution

Models for SNP Calling and Genotyping

- MAQ
 - Li, H., Ruan, J., and Durbin, R. (2008). Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Research* 18, 1851–1858.
- samtools
 - Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* 27, 2987–2993.
- GATK
 - McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., et al. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* 20, 1297–1303.
 - DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., del Angel, G., Rivas, M.A., Hanna, M., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics* 43, 491–498.
- SNVMix
 - Goya, R., Sun, M.G.F., Morin, R.D., Leung, G., Ha, G., Wiegand, K.C., Senz, J., Crisan, A., Marra, M.A., Hirst, M., et al. (2010). SNVMix: predicting single nucleotide variants from next-generation sequencing of tumors. *Bioinformatics* 26, 730–736.
- ...

Genotyping Model used in MAQ

- Data: a pile of bases, with baseQ
 - k nucleotide b and (n-k) nucleotide b'
 - with error rate $\epsilon_1 \leq \dots \leq \epsilon_k \quad \epsilon_{k+1} \leq \dots \leq \epsilon_n$
- Goal: call genotype $\langle b, b \rangle, \langle b, b' \rangle, \langle b', b' \rangle$
- For $G = \langle b, b' \rangle$, $\Pr\{Data|G = \langle b, b' \rangle\} \approx \frac{1}{2^n} \binom{n}{k}$

Genotyping Model used in MAQ

- For $G = \langle b, b \rangle$,

$$\alpha_{nk} = \Pr\{\text{exactly } k \text{ errors in } n \text{ bases}\}$$

$$\bar{\alpha}_{nk}(\bar{e}) = \binom{n}{k} \bar{e}^k (1 - \bar{e})^{n-k}$$

Genotyping Model used in MAQ

$$\alpha_{nk} = \Pr\{\text{exactly } k \text{ errors in } n \text{ bases}\}$$

$$\beta_{nk} = \begin{cases} \Pr\{\text{more than } k \text{ errors} | \text{more than } k-1 \text{ errors in } n \text{ bases}\} & (k > 0) \\ \Pr\{\text{more than } 0 \text{ error in } n \text{ bases}\} & (k = 0) \end{cases}$$

$$\alpha_{nk} = (1 - \beta_{nk})\beta_{n(k-1)} \cdots \beta_{n2}\beta_{n1} = (1 - \beta_{nk}) \prod_{i=0}^{k-1} \beta_{ni} \quad \sum_{k=0}^n \alpha_{nk} = 1$$

$$\beta_{nk} = \frac{\sum_{i=k+1}^n \alpha_{ni}}{\sum_{i=k}^n \alpha_{ni}} = \frac{1 - \sum_{i=0}^k \alpha_{ni}}{1 - \sum_{i=0}^{k-1} \alpha_{ni}} \quad \beta_{nn} = 0$$

Genotyping Model used in MAQ

$$\bar{\alpha}_{nk}(\bar{\epsilon}) = \binom{n}{k} \bar{\epsilon}^k (1 - \bar{\epsilon})^{n-k} \quad \bar{\beta}_{nk}(\bar{\epsilon}) = \frac{1 - \sum_{i=0}^k \bar{\alpha}_{ni}}{1 - \sum_{i=0}^{k-1} \bar{\alpha}_{ni}}$$

$$\beta_{nk}(\bar{\epsilon}) = \bar{\beta}_{nk}^{f_k}(\bar{\epsilon}) \quad 0 < f_k \leq 1$$

$$\alpha_{nk}(\bar{\epsilon}) = (1 - \bar{\beta}_{nk}^{f_k}) \prod_{i=0}^{k-1} \bar{\beta}_{ni}^{f_i} = (1 - \bar{\beta}_{nk}^{f_k}) \prod_{i=0}^{k-1} \left(\frac{\bar{\beta}_{ni}}{\bar{\epsilon}} \right)^{f_i} \cdot \bar{\epsilon}^{f_i} = c_{nk}(\bar{\epsilon}) \cdot \prod_{i=0}^{k-1} \bar{\epsilon}^{f_i}$$

$$c_{nk}(\bar{\epsilon}) = (1 - \bar{\beta}_{nk}^{f_k}) \prod_{i=0}^{k-1} \left(\frac{\bar{\beta}_{ni}}{\bar{\epsilon}} \right)^{f_i}$$

Genotyping Model used in MAQ

$$\alpha_{nk}(\epsilon_1, \dots, \epsilon_k; \epsilon_{k+1}, \dots, \epsilon_n) \approx c_{nk}(\bar{\epsilon}) \cdot \prod_{i=0}^{k-1} \epsilon_{i+1}^{f_i}$$

$$\log \bar{\epsilon} = \frac{\sum_{i=0}^{k-1} f_i \log \epsilon_{i+1}}{\sum_{i=0}^{k-1} f_i} \quad \prod_{i=0}^{k-1} \bar{\epsilon}^{f_i} = \prod_{i=0}^{k-1} \epsilon_{i+1}^{f_i}$$

$$f_k = 0.85^k$$

$$\alpha_{nk}(\epsilon_1, \dots, \epsilon_k; \tilde{\epsilon}_1, \dots, \tilde{\epsilon}_k; \epsilon_{k+1}, \dots, \epsilon_n; \tilde{\epsilon}_{k+1}, \dots, \tilde{\epsilon}_n) \approx c_{nk}(\bar{\epsilon}) \prod_{i=0}^{k-1} \epsilon_{i+1}^{f_i} \cdot c_{\tilde{n}\tilde{k}}(\tilde{\bar{\epsilon}}) \prod_{\tilde{i}=0}^{\tilde{k}-1} \tilde{\epsilon}_{\tilde{i}+1}^{f_{\tilde{i}}}$$

Genotyping Model used in MAQ

- For $G = \langle b, b \rangle$,
$$\Pr\{Data|G = \langle b, b \rangle\} = \alpha_{n,k}(\epsilon_1, \dots, \epsilon_k; \epsilon_{k+1}, \dots, \epsilon_n)$$
- For $G = \langle b, b' \rangle$,
$$\Pr\{Data|G = \langle b, b' \rangle\} \approx \frac{1}{2^n} \binom{n}{k}$$
- For $G = \langle b', b' \rangle$,
$$\Pr\{Data|G = \langle b', b' \rangle\} = \alpha_{n,n-k}(\epsilon_{k+1}, \dots, \epsilon_n; \epsilon_1, \dots, \epsilon_k)$$

Genotyping Model used in MAQ

$$\Pr\{G|Data\} \propto \Pr\{G\} \cdot \Pr\{Data|G\}$$

- For $G=\langle b,b \rangle$,

$$\Pr\{G=\langle b,b \rangle|Data\} = \frac{\Pr\{G=\langle b,b \rangle\} \cdot \Pr\{Data|G=\langle b,b \rangle\}}{\Pr\{G=\langle b,b \rangle\} \cdot \Pr\{Data|G=\langle b,b \rangle\} + \Pr\{G=\langle b,b' \rangle\} \cdot \Pr\{Data|G=\langle b,b' \rangle\} + \Pr\{G=\langle b',b' \rangle\} \cdot \Pr\{Data|G=\langle b',b' \rangle\}}$$

- For $G=\langle b,b' \rangle$,

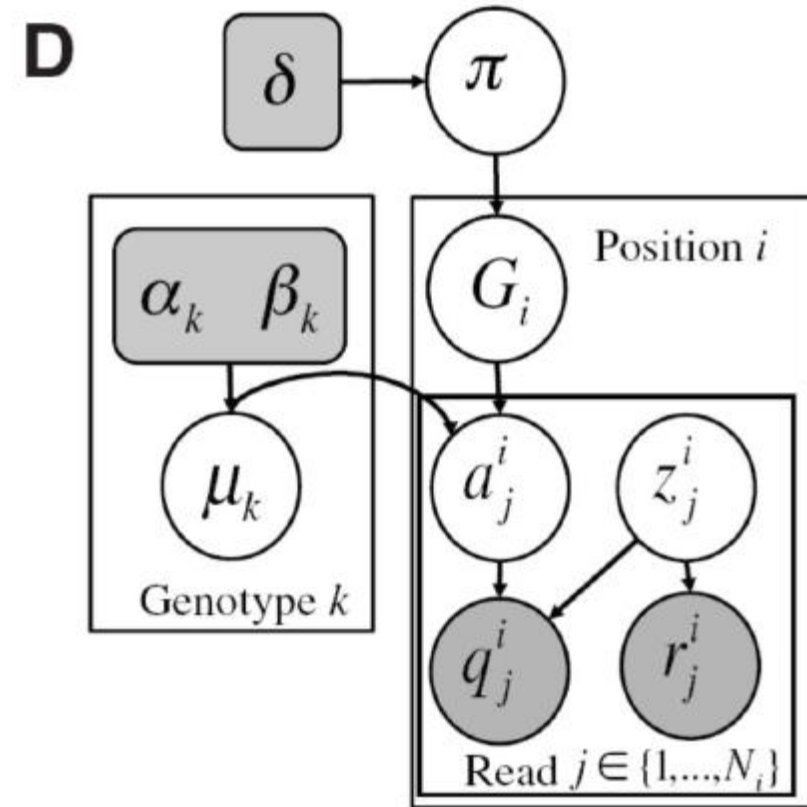
$$\Pr\{G=\langle b,b' \rangle|Data\} = \frac{\Pr\{G=\langle b,b' \rangle\} \cdot \Pr\{Data|G=\langle b,b' \rangle\}}{\Pr\{G=\langle b,b \rangle\} \cdot \Pr\{Data|G=\langle b,b \rangle\} + \Pr\{G=\langle b,b' \rangle\} \cdot \Pr\{Data|G=\langle b,b' \rangle\} + \Pr\{G=\langle b',b' \rangle\} \cdot \Pr\{Data|G=\langle b',b' \rangle\}}$$

- For $G=\langle b',b' \rangle$,

$$\Pr\{G=\langle b',b' \rangle|Data\} = \frac{\Pr\{G=\langle b',b' \rangle\} \cdot \Pr\{Data|G=\langle b',b' \rangle\}}{\Pr\{G=\langle b,b \rangle\} \cdot \Pr\{Data|G=\langle b,b \rangle\} + \Pr\{G=\langle b,b' \rangle\} \cdot \Pr\{Data|G=\langle b,b' \rangle\} + \Pr\{G=\langle b',b' \rangle\} \cdot \Pr\{Data|G=\langle b',b' \rangle\}}$$

Genotyping Model used in SNVMix

- Probabilistic Graphical Model
 - position i , read j , genotype k
 - G_i : genotype
 - a_{ji} : match reference allele or not?
 - q_{ji} : prob. of correct base calling
 - z_{ji} : alignment correct or not?
 - r_{ji} : prob. of correct mapping
 - μ_k : parameter of binomial for genotype k



SNVMix2 model

Goya, R., et al. (2010). SNVMix: predicting single nucleotide variants from

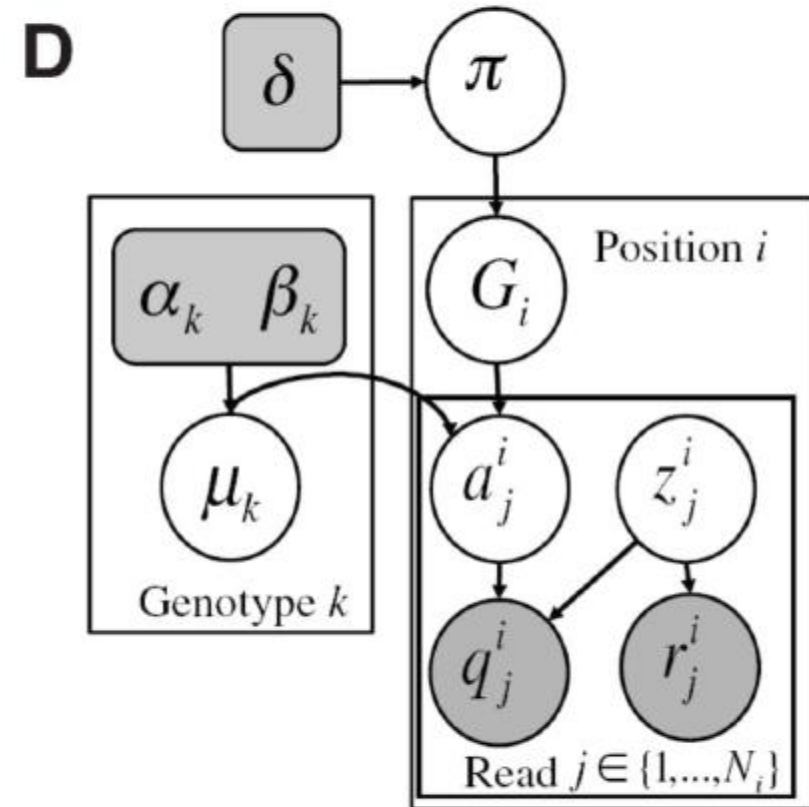
Genotyping Model used in SNVMix

$$p(\mathbf{G}_i | \boldsymbol{\pi}) = \text{Multinomial}(\mathbf{G}_i | \boldsymbol{\pi}, 1)$$

$$p(\boldsymbol{\pi} | \boldsymbol{\delta}) = \text{Dirichlet}(\boldsymbol{\pi} | \boldsymbol{\delta})$$

$$p(a_j^i | G_i = k, \mu_k) = \text{Bernoulli}(a_j^i | \mu_k)$$

$$p(\mu_k | \alpha_k, \beta_k) = \text{Gamma}(\mu_k | \alpha_k, \beta_k)$$



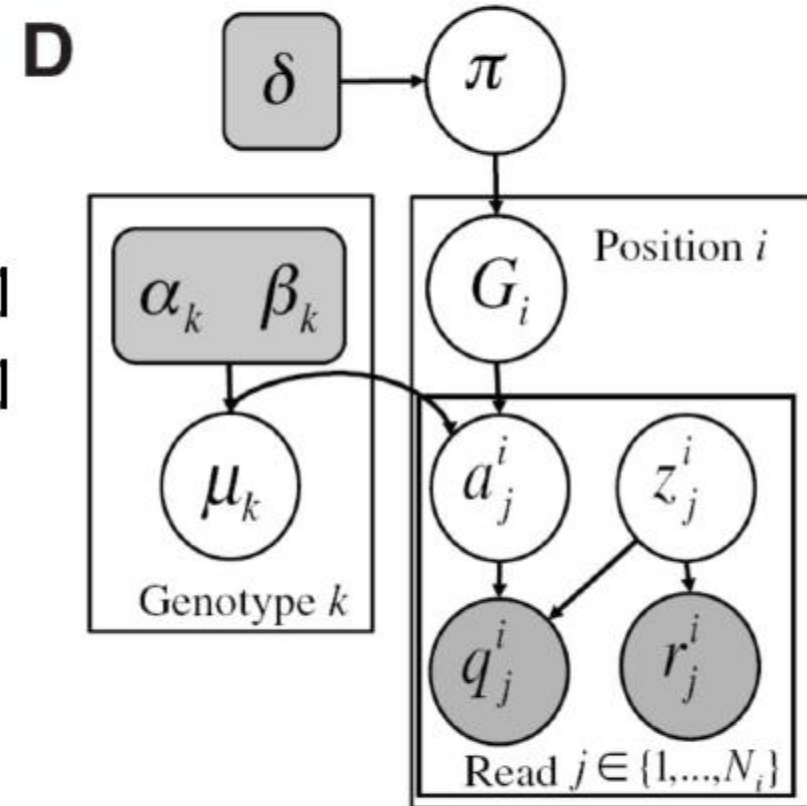
Goya, R., et al. (2010). SNVMix: predicting single nucleotide variants from

Genotyping Model used in SNVMix

$$p(z_j^i) = \text{Bernoulli}(z_j^i | 0.5)$$

$$p(q_j^i | a_j^i, z_j^i) = \begin{cases} q_j^i & \text{if } a_j^i = 1, z_j^i = 1 \\ 1 - q_j^i & \text{if } a_j^i = 0, z_j^i = 1 \\ 0.5 & \text{if } z_j^i = 0 \end{cases}$$

$$p(r_j^i | z_j^i) = \begin{cases} r_j^i & \text{if } z_j^i = 1 \\ 1 - r_j^i & \text{if } z_j^i = 0 \end{cases}$$



SNVMix2 model

Goya, R., et al. (2010). SNVMix: predicting single nucleotide variants from

Bioinformatics: Introduction and Methods

Computer Science Department, Southwest University

Thank you

