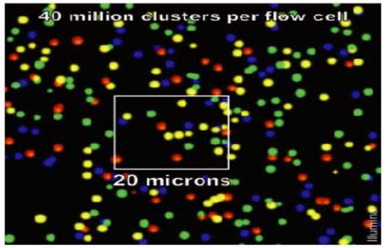
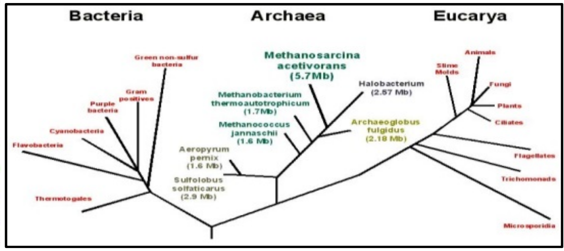
TAACCCTAACCCTAACCCTAACCCTAACCCTA
CCTAACCCTAACCCTAACCCTAACCCTAACCC
CCCTAACCCCTAACCCTAACCCTAACCCTAAC
AACCCTAACCCTAACCCTAACCCCTAACCCTA
ACCCTAACCCCAACCCCAACCCCAACCCCAAC
CTACCCTAACCCTAACCCTAACCCTAACCCTA
ACCCTAACCCTAACCCTAACCCCTAACCCTAA



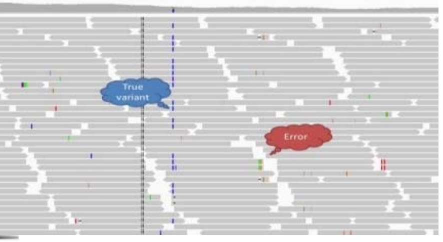

Bacteria
Archaea
Eucarya
Green non-sulfur bacteria
Gram positives
Purple bacteria
Cyanobacteria
Flavobacteria
Thermotogales
Methanosarcina acetivorans (5.7Mb)
Methanobacterium thermoautotrophicum (1.7Mb)
Methanococcus jannaschii (1.6 Mb)
Halobacterium (2.57 Mb)
Archaeoglobus fulgidus (2.18 Mb)
Aeropyrum pernix (1.6 Mb)
Sulfolobus solfataricus (2.9 Mb)
Animals
Slime Molds
Fungi
Plants
Ciliates
Flagellates
Trichomonads
Microsporidia

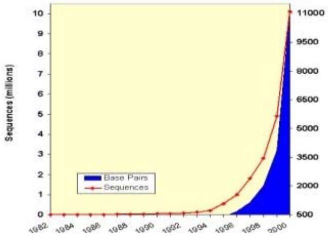
Next Generation Sequencing (NGS): Reads Mapping

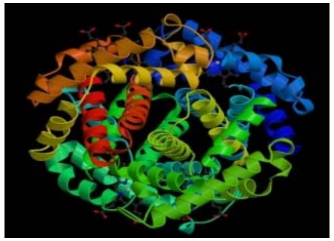
北京大学生物信息学中心 高歌

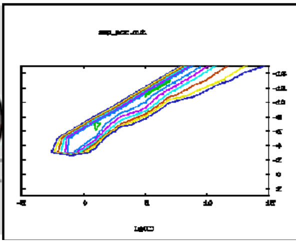
Ge Gao, Ph.D.

Center for Bioinformatics, Peking University

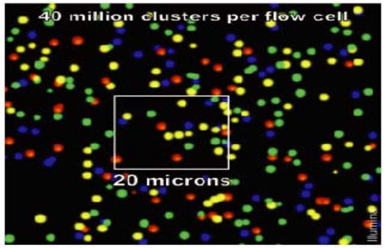


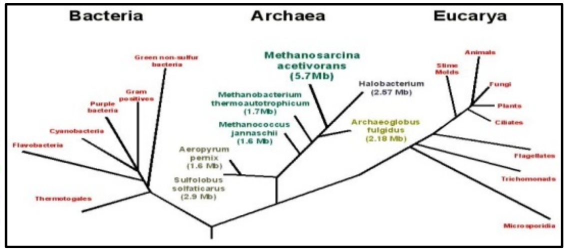






```
TAACCCTAACCCTAACCCTAACCCTAACCCTA
CCTAACCCTAACCCTAACCCTAACCCTAACCC
CCCTAACCCCTAACCCTAACCCTAACCCTAAC
AACCCTAACCCTAACCCTAACCCCTAACCCTA
ACCCTAACCCCAACCCCAACCCCAACCCCAAC
CTACCCTAACCCTAACCCTAACCCTAACCCTA
ACCCTAACCCTAACCCTAACCCCTAACCCTAA
```



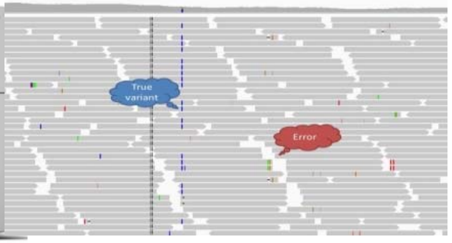

Bacteria
Archaea
Eucarya
Green non-sulfur bacteria
Gram positives
Purple bacteria
Cyanobacteria
Flavobacteria
Thermotogales
Methanosarcina acetivorans (5.7Mb)
Methanobacterium thermoautotrophicum (1.7Mb)
Methanococcus jannaschii (1.6 Mb)
Aeropyrum pernix (1.6 Mb)
Sulfolobus solfataricus (2.9 Mb)
Halobacterium (2.57 Mb)
Archaeoglobus fulgidus (2.18 Mb)
Animals
Slime Molds
Fungi
Plants
Ciliates
Flagellates
Trichomonads
Microsporidia

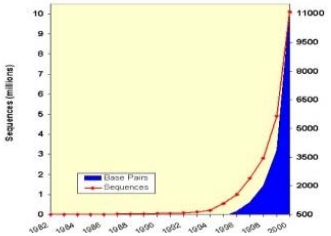
Unit 2:
NGS: Reads Mapping

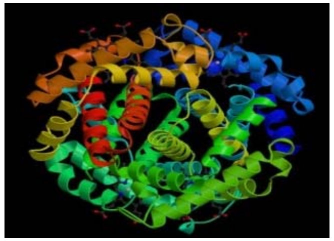
北京大学生物信息学中心 高歌

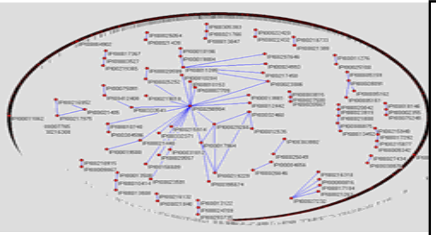
Ge Gao, Ph.D.

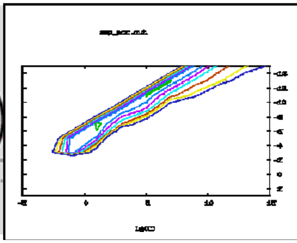
Center for Bioinformatics, Peking University



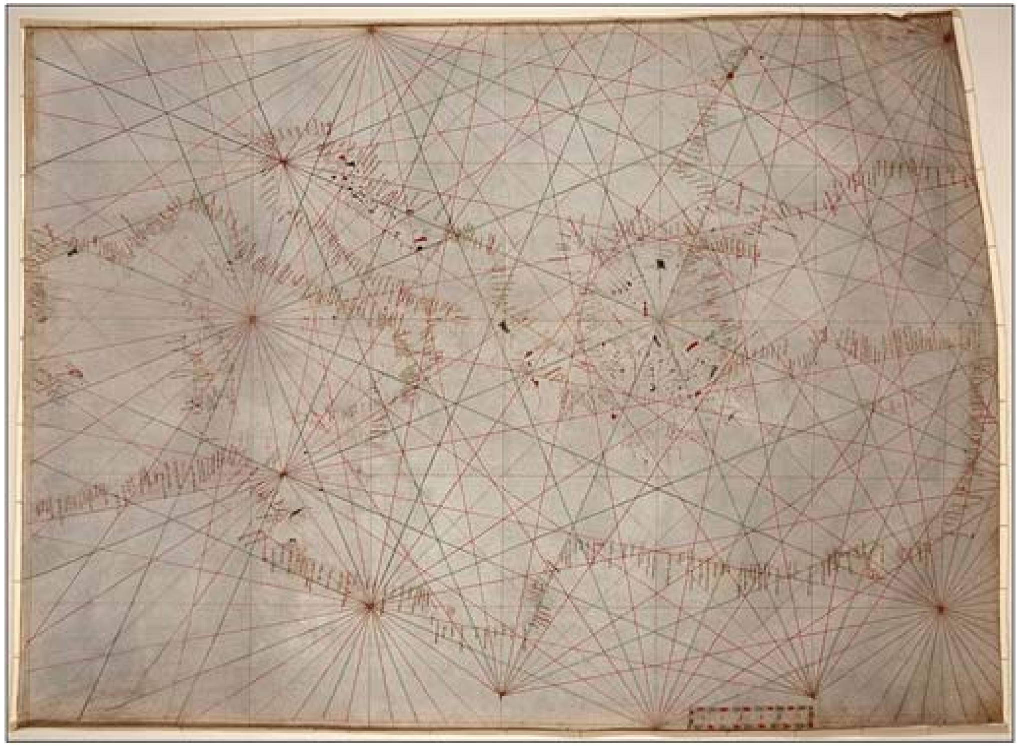








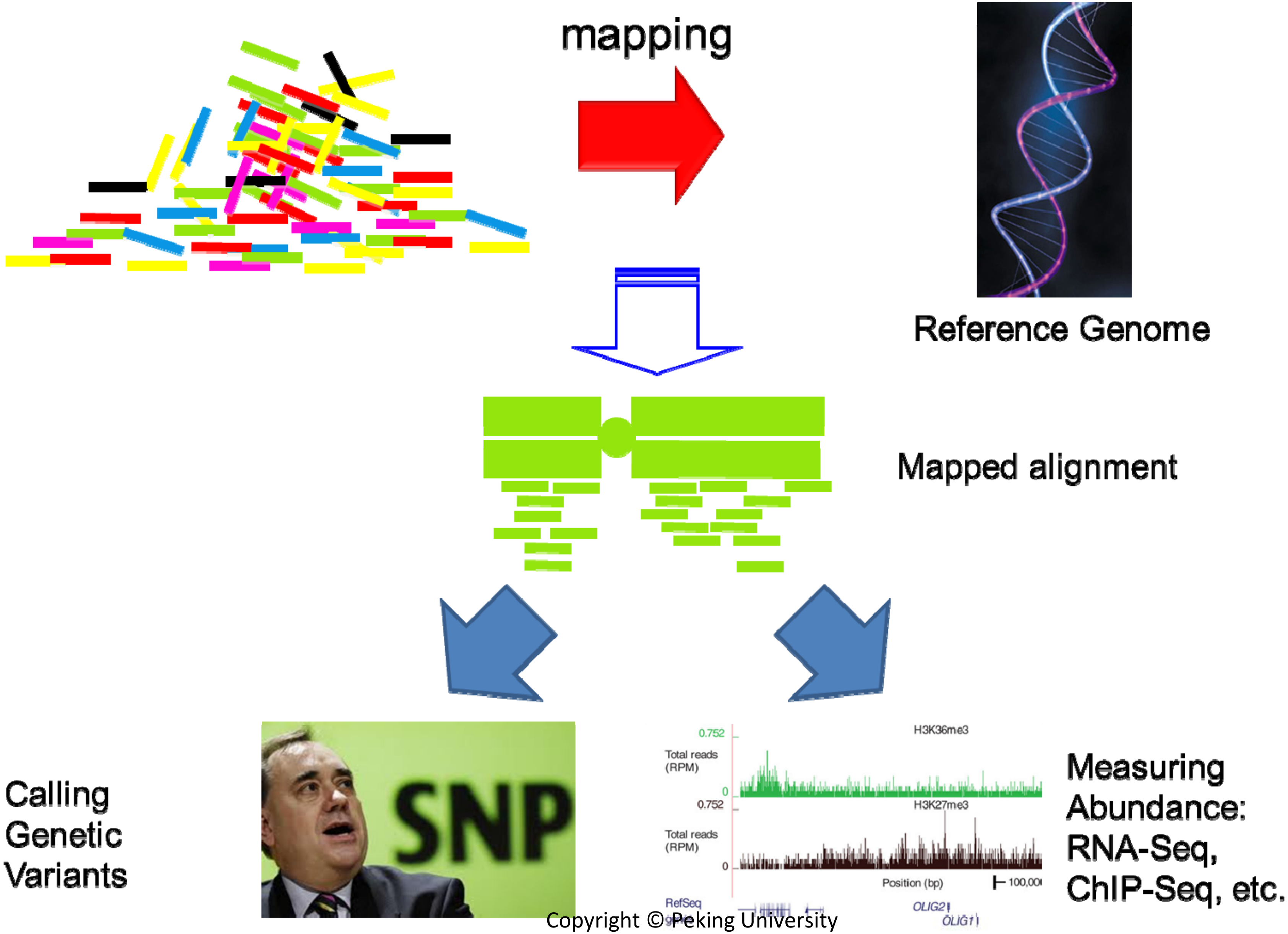
Reads Mapping



Map-Making / Cartography:
Establish relationship between locations
(http://en.wikipedia.org/wiki/Cartography)

<ins>Technological</ins>: Reads is usual too short to be used/assembled *de novo*

<ins>Scientific</ins>: Taking full usage of existing annotation/knowledge



mapping

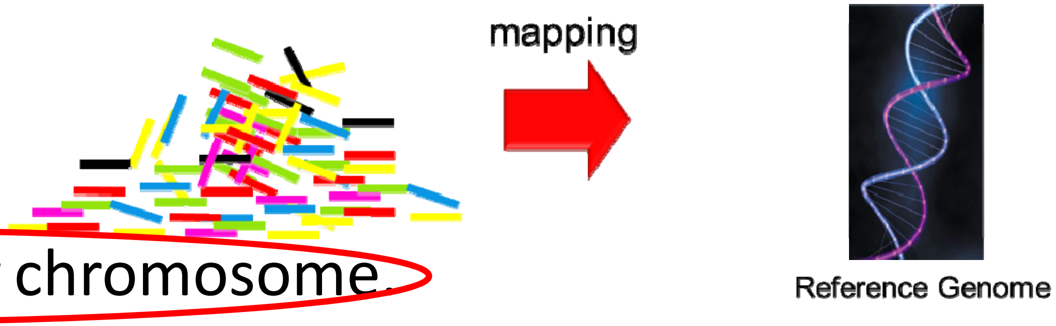
Reference Genome

Mapped alignment

Calling Genetic Variants

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

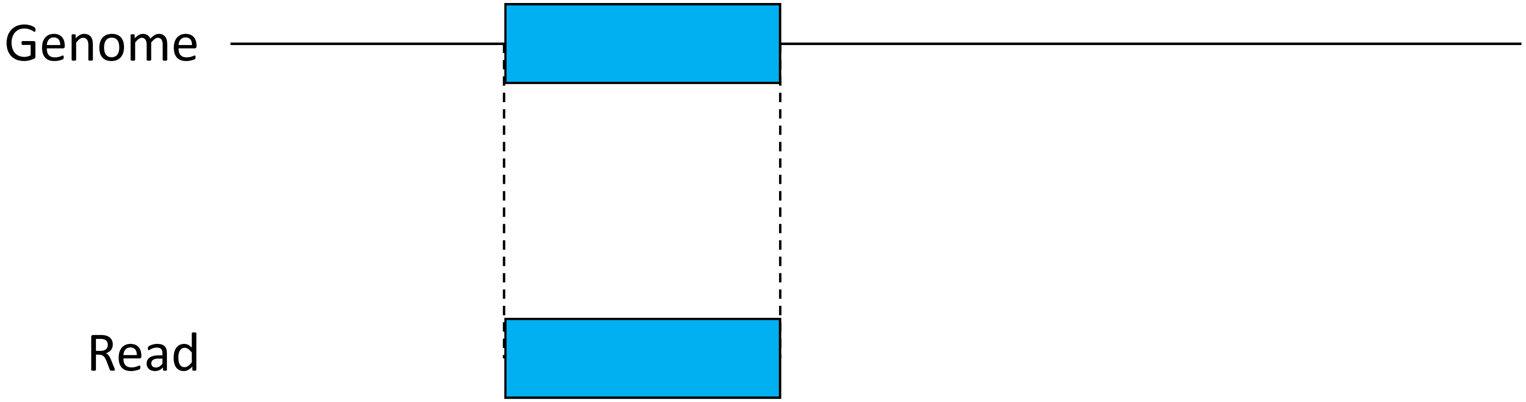
Mapping: Input Data



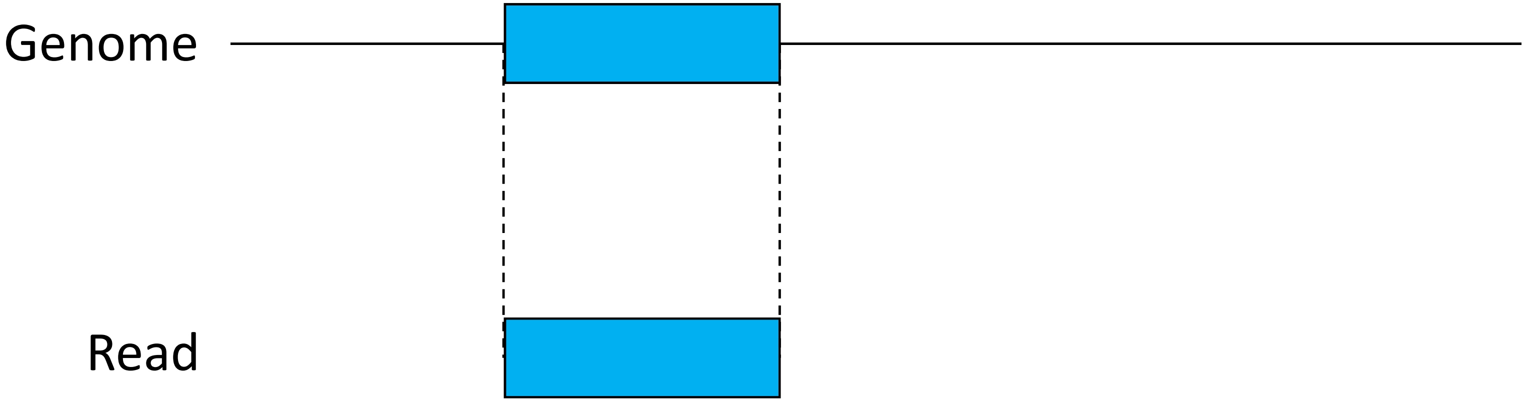
- Reference Genome
  - Nucleotide
  - **Length**: Hundreds of Mb *per* chromosome.
  - ~3 Gb in total (for human genome)

- Reads
  - Nucleotide, with **various qualities** (relatively **high error rate**: 1e-2 ~ 1e-5)
  - **Length**: 36~80 bp *per* read
  - Hundreds of Gbs *per* run

"Embedded" Alignment

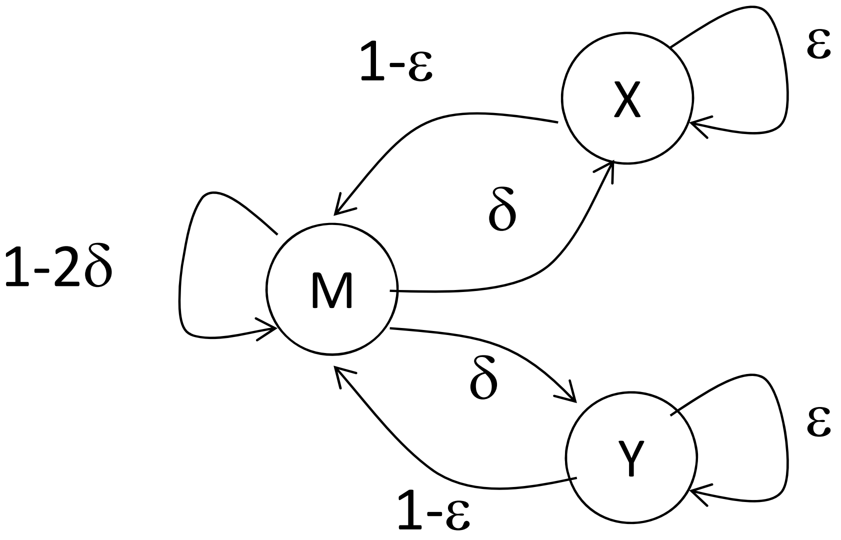


One sequence is "*embedded*" in the other sequence (NGS Reads, PCR primer, *etc.*)



What we need here is actually a hybrid "global-local" alignment

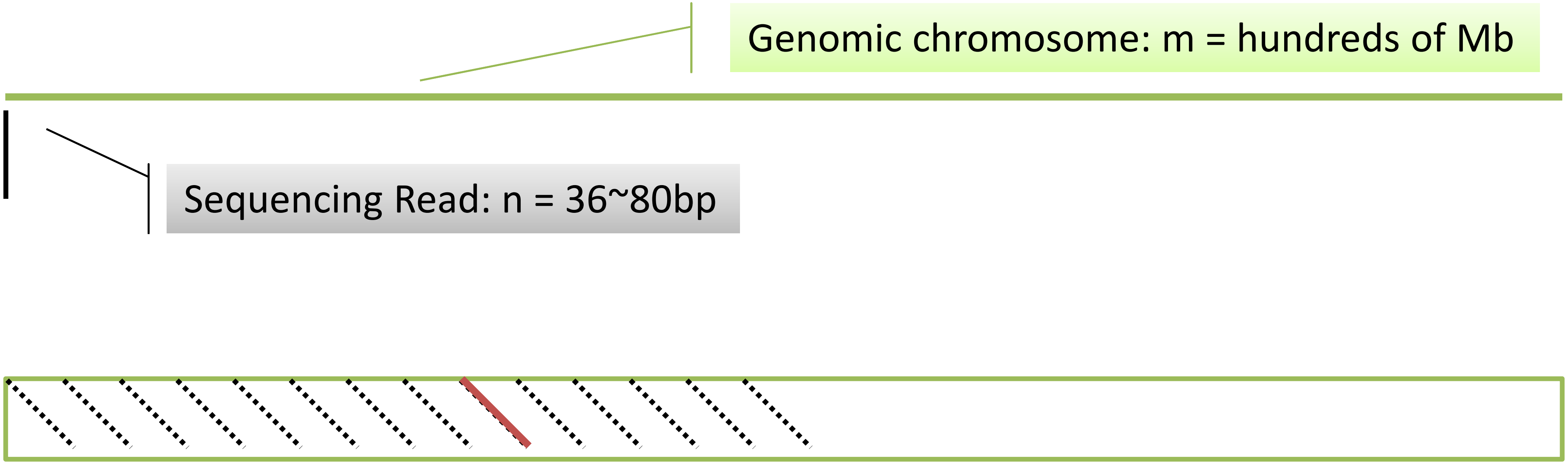
- ✓ "Global" for short sequence (i.e. NGS Read)
- ✓ But "Local" for long sequence (i.e. Reference Genome)
- ✓ In particular, the surrounding "overhang" gaps should be not penalized.



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 = hundreds of Mb

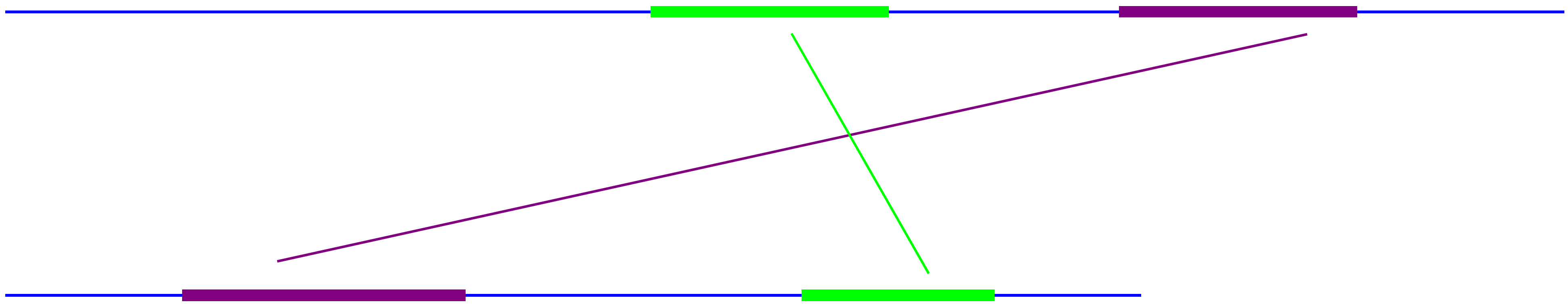
Sequencing Read: n = 36~80bp

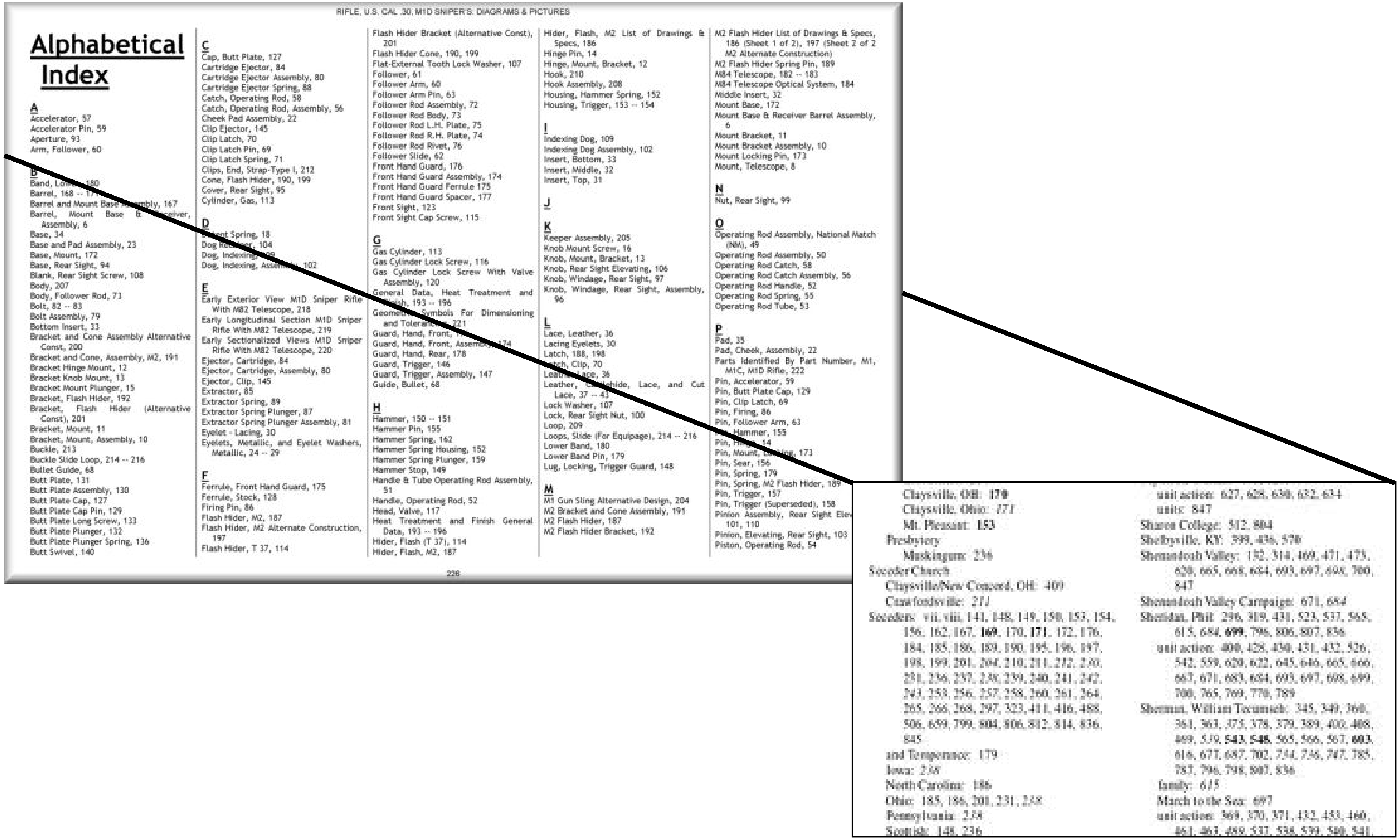
Most of paths will just fail eventually!

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.





RIFLE, U.S. CAL .30, M1D SNIPER'S: DIAGRAMS & PICTURES

Alphabetical Index

A
Accelerator, 57
Accelerator Pin, 59
Aperture, 93
Arm, Follower, 60

B
Band, Lower, 180
Barrel, 168 -- 171
Barrel and Mount Base Assembly, 167
Barrel, Mount Base & Receiver, Assembly, 6
Base, 34
Base and Pad Assembly, 23
Base, Mount, 172
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 Cone Assembly Alternative Const, 200
Bracket and Cone, Assembly, M2, 191
Bracket Hinge Mount, 12
Bracket Knob Mount, 13
Bracket Mount Plunger, 15
Bracket, Flash Hider, 192
Bracket, Flash Hider (Alternative Const), 201
Bracket, Mount, 11
Bracket, Mount, Assembly, 10
Buckle, 213
Buckle Slide Loop, 214 -- 216
Bullet Guide, 68
Butt Plate, 131
Butt Plate Assembly, 130
Butt Plate Cap, 127
Butt Plate Cap Pin, 129
Butt Plate Long Screw, 133
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, 22
Clip Ejector, 145
Clip Latch, 70
Clip Latch Pin, 69
Clip Latch Spring, 71
Clips, End, Strap-Type I, 212
Cone, Flash Hider, 190, 199
Cover, Rear Sight, 95
Cylinder, Gas, 113

D
Detent Spring, 18
Dog Retainer, 104
Dog, Indexing, 109
Dog, Indexing, Assembly, 102

E
Early Exterior View M1D Sniper Rifle With M82 Telescope, 218
Early Longitudinal Section M1D Sniper Rifle With M82 Telescope, 219
Early Sectionalized Views M1D Sniper Rifle With M82 Telescope, 220
Ejector, Cartridge, 84
Ejector, Cartridge, Assembly, 80
Ejector, Clip, 145
Extractor, 85
Extractor Spring, 89
Extractor Spring Plunger, 87
Extractor Spring Plunger Assembly, 81
Eyelet - Lacing, 30
Eyelets, Metallic, and Eyelet Washers, Metallic, 24 -- 29

F
Ferrule, Front Hand Guard, 175
Ferrule, Stock, 128
Firing Pin, 86
Flash Hider, M2, 187
Flash Hider, M2 Alternate Construction, 197
Flash Hider, T 37, 114
Flash Hider Bracket (Alternative Const), 201
Flash Hider Cone, 190, 199
Flat-External Tooth Lock Washer, 107
Follower, 61
Follower Arm, 60
Follower Arm Pin, 63
Follower Rod Assembly, 72
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 Ferrule 175
Front Hand Guard Spacer, 177
Front Sight, 123
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, 176
Guard, Hand, Front, Assembly, 174
Guard, Hand, Rear, 178
Guard, Trigger, 146
Guard, Trigger, Assembly, 147
Guide, Bullet, 68

H
Hammer, 150 -- 151
Hammer Pin, 155
Hammer Spring, 162
Hammer Spring Housing, 152
Hammer Spring Plunger, 159
Hammer Stop, 149
Handle & Tube Operating Rod Assembly, 51
Handle, Operating Rod, 52
Head, Valve, 117
Heat Treatment and Finish General Data, 193 -- 196
Hider, Flash (T 37), 114
Hider, Flash, M2, 187
Hider, Flash, M2 List of Drawings & Specs, 186
Hinge Pin, 14
Hinge, Mount, Bracket, 12
Hook, 210
Hook Assembly, 208
Housing, Hammer Spring, 152
Housing, Trigger, 153 -- 154

I
Indexing Dog, 109
Indexing Dog Assembly, 102
Insert, Bottom, 33
Insert, Middle, 32
Insert, Top, 31

J

K
Keeper Assembly, 205
Knob Mount Screw, 16
Knob, Mount, Bracket, 13
Knob, Rear Sight Elevating, 106
Knob, Windage, Rear Sight, 97
Knob, Windage, Rear Sight, Assembly, 96

L
Lace, Leather, 36
Lacing Eyelets, 30
Latch, 188, 198
Latch, Clip, 70
Leather Lace, 36
Leather, Cowhide, Lace, and Cut Lace, 37 -- 43
Lock Washer, 107
Lock, Rear Sight Nut, 100
Loop, 209
Loops, Slide (For Equipage), 214 -- 216
Lower Band, 180
Lower Band Pin, 179
Lug, Locking, Trigger Guard, 148

M
M1 Gun Sling Alternative Design, 204
M2 Bracket and Cone Assembly, 191
M2 Flash Hider, 187
M2 Flash Hider Bracket, 192
M2 Flash Hider List of Drawings & Specs, 186 (Sheet 1 of 2), 197 (Sheet 2 of 2 M2 Alternate Construction)
M2 Flash Hider Spring Pin, 189
M84 Telescope, 182 -- 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

N
Nut, Rear Sight, 99

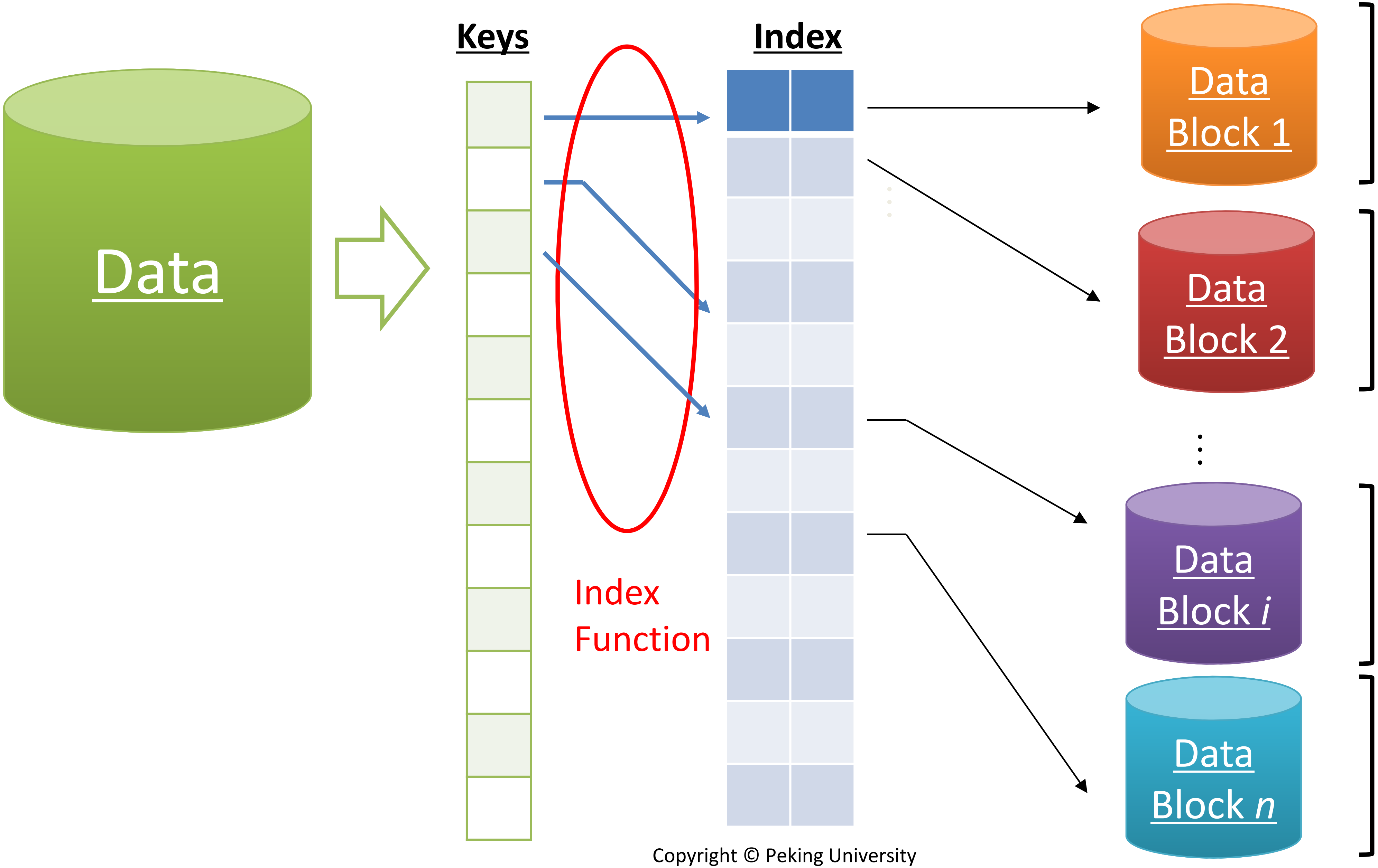
O
Operating Rod Assembly, National Match (NM), 49
Operating Rod Assembly, 50
Operating Rod Catch, 58
Operating Rod Catch Assembly, 56
Operating Rod Handle, 52
Operating Rod Spring, 55
Operating Rod Tube, 53

P
Pad, 35
Pad, Cheek, Assembly, 22
Parts Identified By Part Number, M1, M1C, M1D Rifle, 222
Pin, Accelerator, 59
Pin, Butt Plate Cap, 129
Pin, Clip Latch, 69
Pin, Firing, 86
Pin, Follower Arm, 63
Pin, Hammer, 155
Pin, Hinge, 14
Pin, Mount, Locking, 173
Pin, Sear, 156
Pin, Spring, 179
Pin, Spring, M2 Flash Hider, 189
Pin, Trigger, 157
Pin, Trigger (Superseded), 158
Pinion Assembly, Rear Sight Elev 101, 110
Pinion, Elevating, Rear Sight, 103
Piston, Operating Rod, 54

228

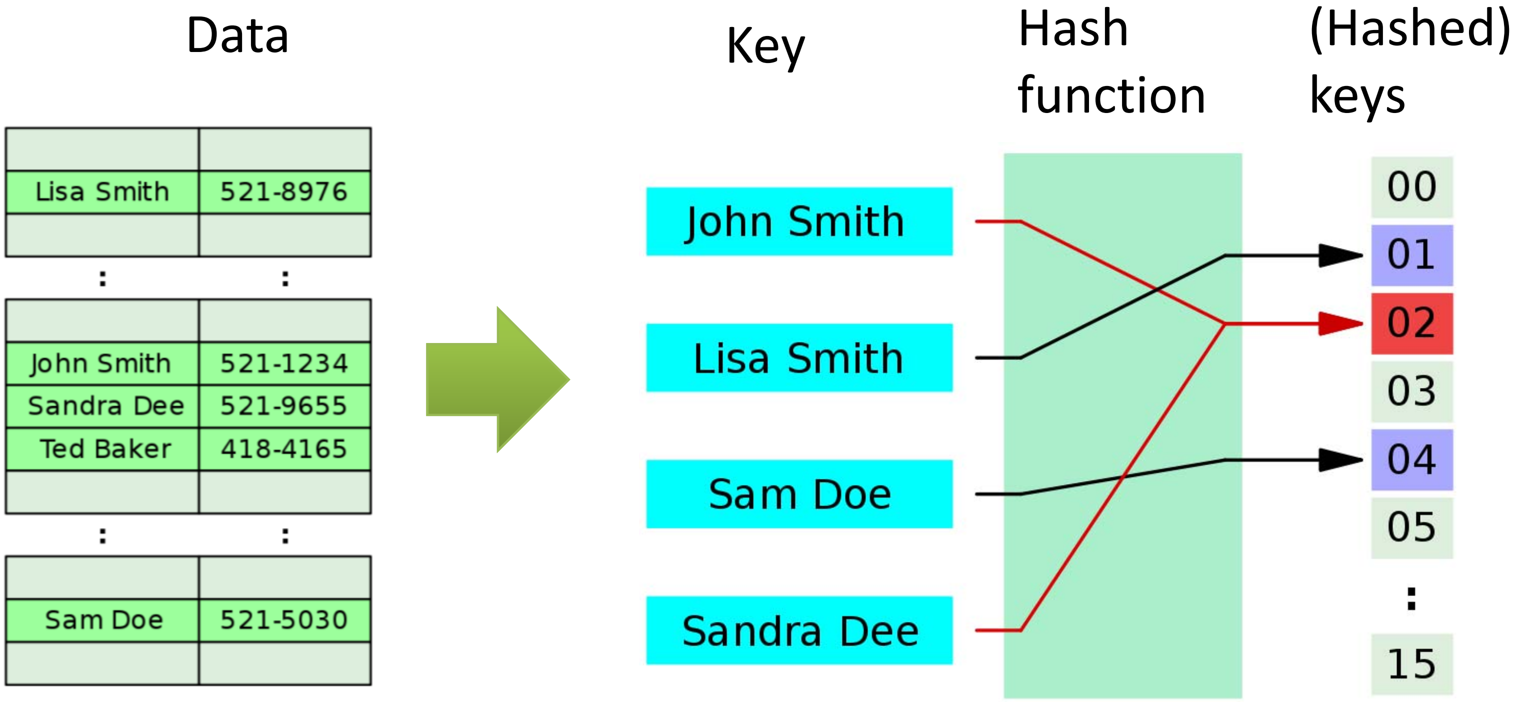
Claysville, OH: **170**
Claysville, Ohio: *171*
Mt. Pleasant: **153**
Presbytery
Muskingum: 236
Seceder Church
Claysville/New Concord, OH: 409
Crawfordsville: *211*
Seceders: vii, viii, 141, 148, 149, 150, 153, 154, 156, 162, 167, **169**, 170, **171**, 172, 176, 184, 185, 186, 189, 190, 195, 196, 197, 198, 199, 201, 204, 210, 211, *212*, *230*, 231, 236, 237, *238*, 239, 240, 241, *242*, *243*, 253, 256, *257*, 258, 260, 261, 264, 265, 266, 268, 297, 323, 411, 416, 488, 506, 659, 799, 804, 806, 812, 814, 836, 845
and Temperance: 179
Iowa: *238*
North Carolina: 186
Ohio: 185, 186, 201, 231, *238*
Pennsylvania: *238*
Scottish: 148, 236

unit action: 627, 628, 630, 632, 634
units: 847
Sharon College: 512, 804
Shelbyville, KY: 399, 436, 570
Shenandoah Valley: 132, 314, 469, 471, 473, 620, 665, 668, 684, 693, 697, *698*, 700, 847
Shenandoah Valley Campaign: 671, *684*
Sheridan, Phil: 296, 319, 431, 523, 537, 565, 615, *684*, **699**, 796, 806, 807, 836
unit action: 400, 428, 430, 431, 432, 526, 542, 559, 620, 622, 645, 646, 665, 666, 667, 671, 683, 684, 693, 697, 698, 699, 700, 765, 769, 770, 789
Sherman, William Tecumseh: 345, 349, 360, 361, 363, 375, 378, 379, 389, 400, 408, 469, *539*, **543**, **548**, 565, 566, 567, **603**, 616, 677, 687, 702, *734*, *736*, *747*, 785, 787, 796, 798, 807, 836
family: *615*
March to the Sea: 697
unit action: 369, 370, 371, 432, 453, 460, 461, 463, 469, 537, 538, 539, 540, 541,


Data
Keys
Index
Index Function
Data Block 1
Data Block 2
Data Block *i*
Data Block *n*

Hash

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



Data	
Lisa Smith	521-8976
:	:
John Smith	521-1234
Sandra Dee	521-9655
Ted Baker	418-4165
:	:
Sam Doe	521-5030

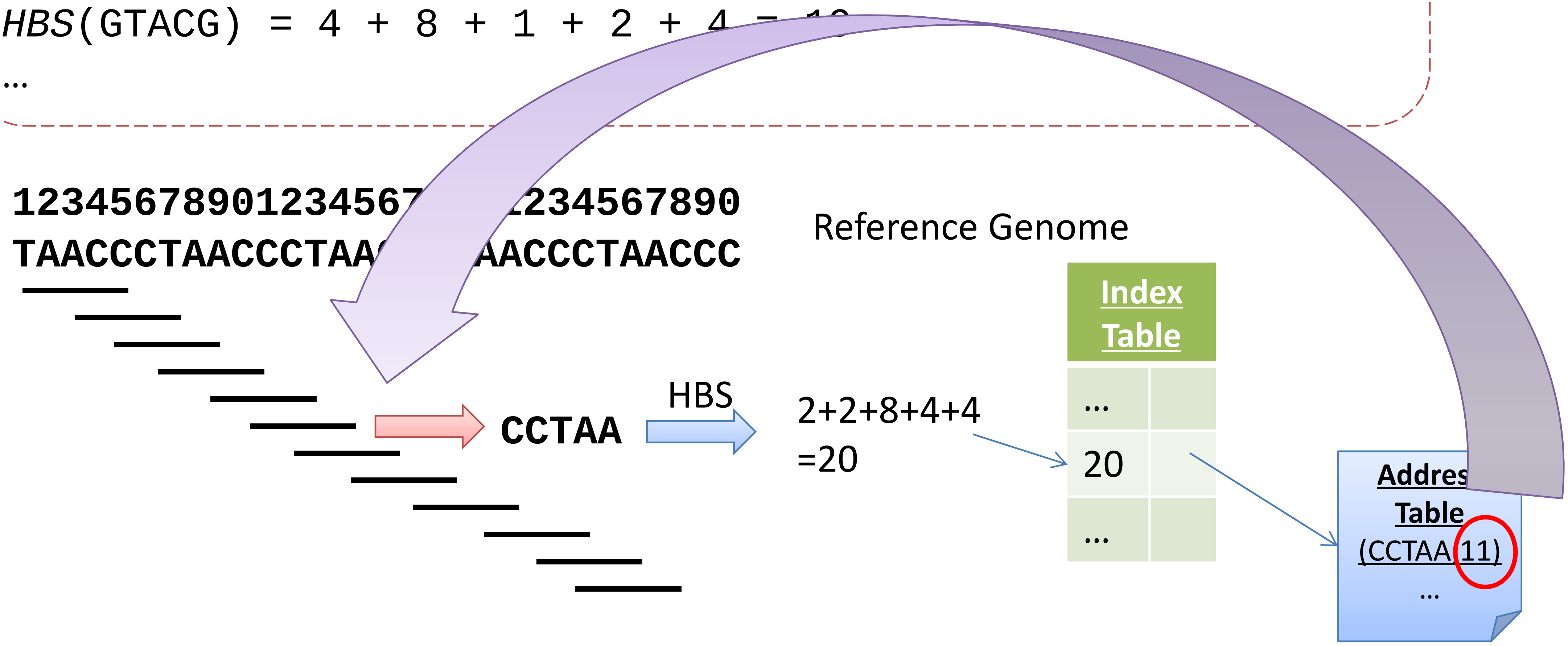
HBS: A naive hash function

Let’s assume: A = 1, C = 2, G = 4, T = 8, then: $HBS(S) = \sum_i HBS(S_i)$, e.g:

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

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

…

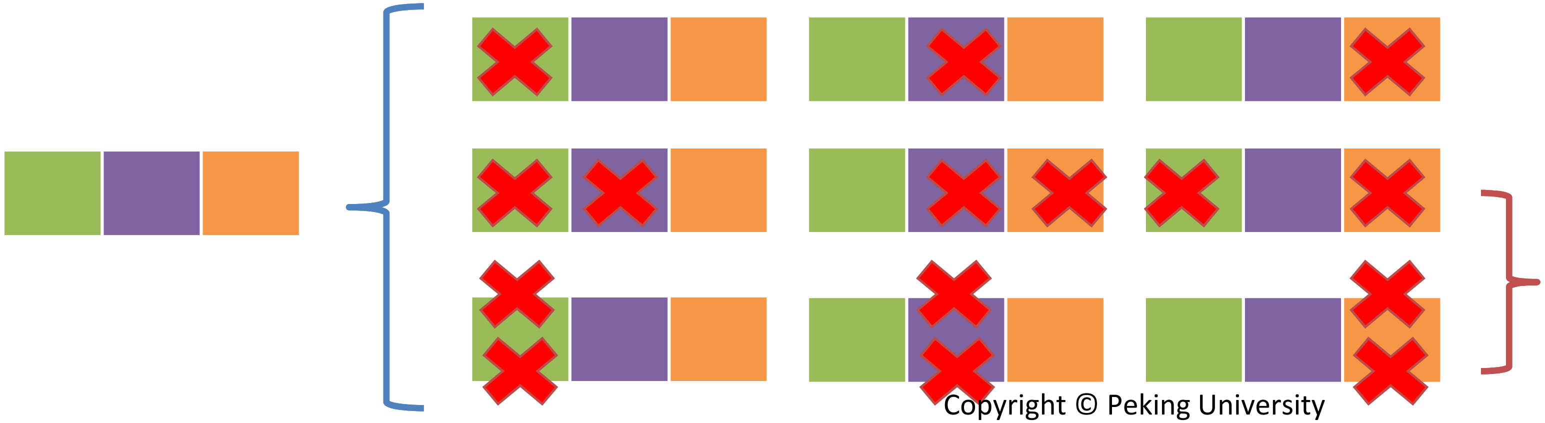


Pigeonhole principle (抽屉原理)

"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. "

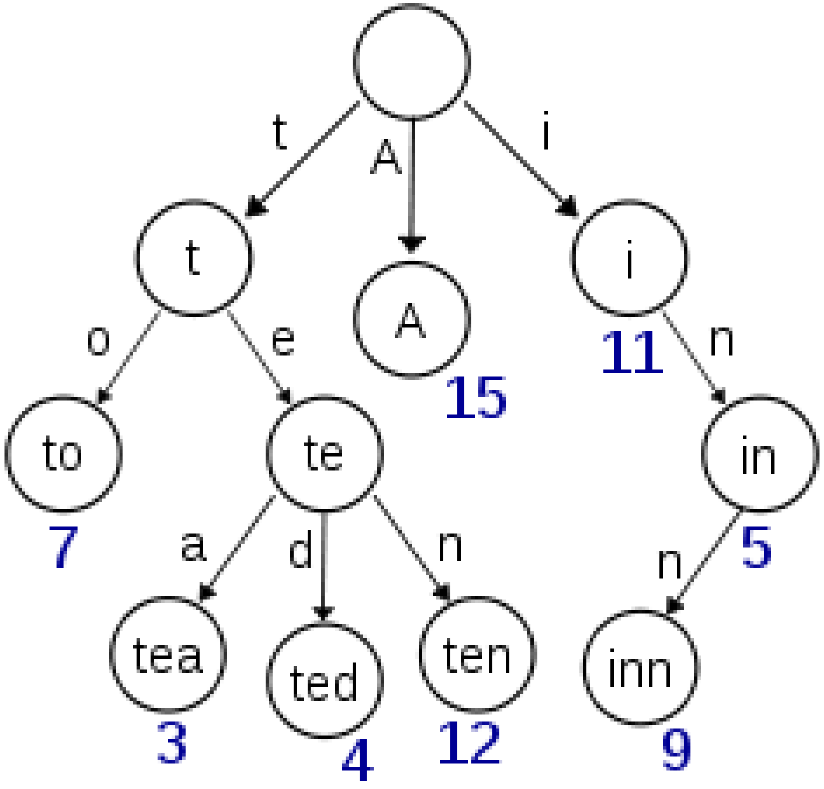
http://en.wikipedia.org/wiki/Pigeonhole_principle

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.



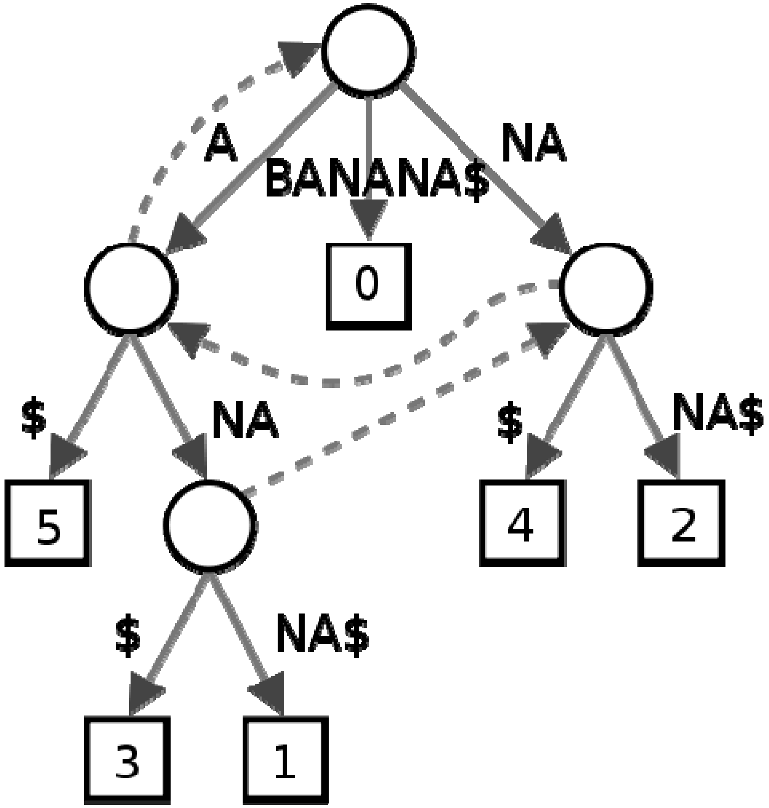
ELAND
MAQ
SOAP1
…

Prefix Tree



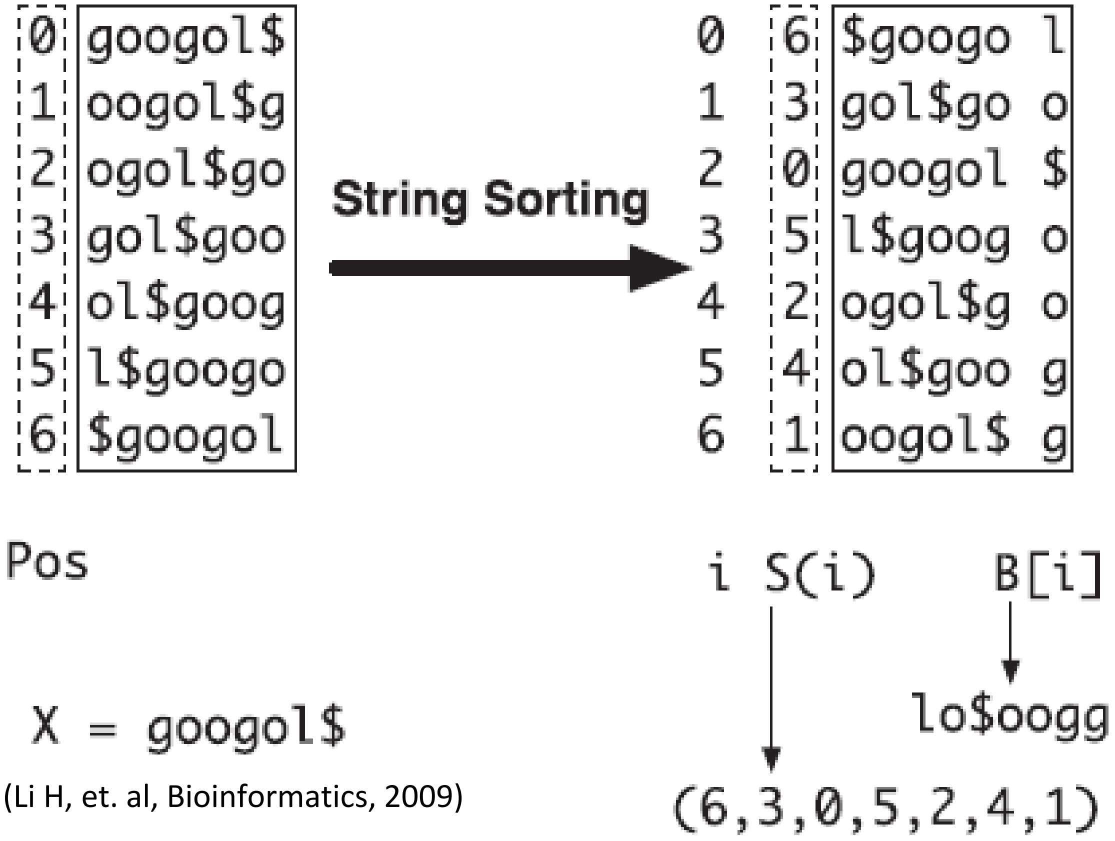
http://en.wikipedia.org/wiki/Trie

Suffix Tree



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

Burrows–Wheeler transform (BWT)



Pos	
0	googol$
1	oogol$g
2	ogol$go
3	gol$goo
4	ol$goog
5	l$googo
6	$googol

String Sorting

i	S(i)		B[i]
0	6	$googo	l
1	3	gol$go	o
2	0	googol	$
3	5	l$goog	o
4	2	ogol$g	o
5	4	ol$goo	g
6	1	oogol$	g

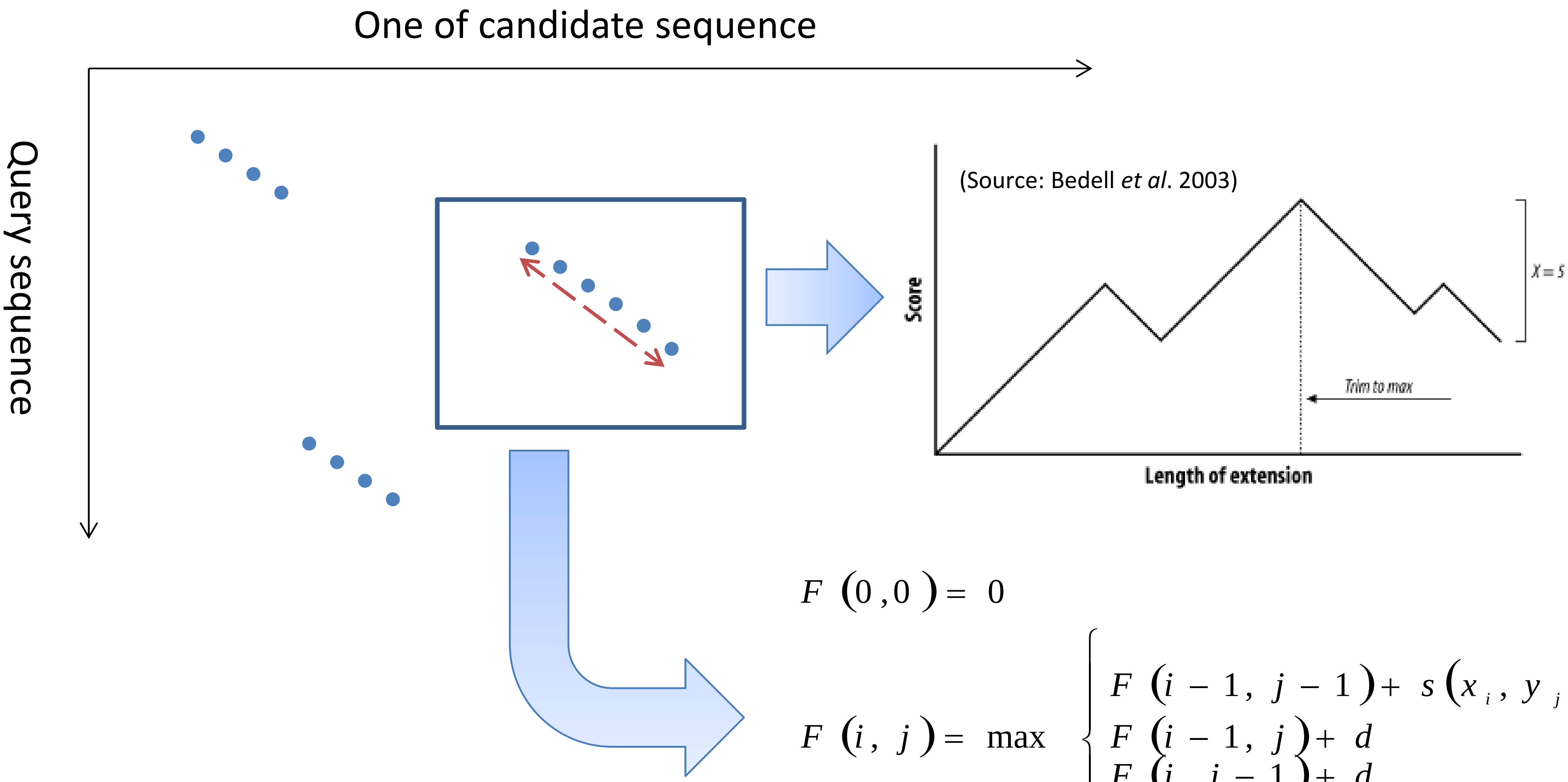
X = googol$

lo$oogg

(6,3,0,5,2,4,1)

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

BOWTIE
BWA
SOAP3
…



(Source: Bedell *et al*. 2003)

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

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

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

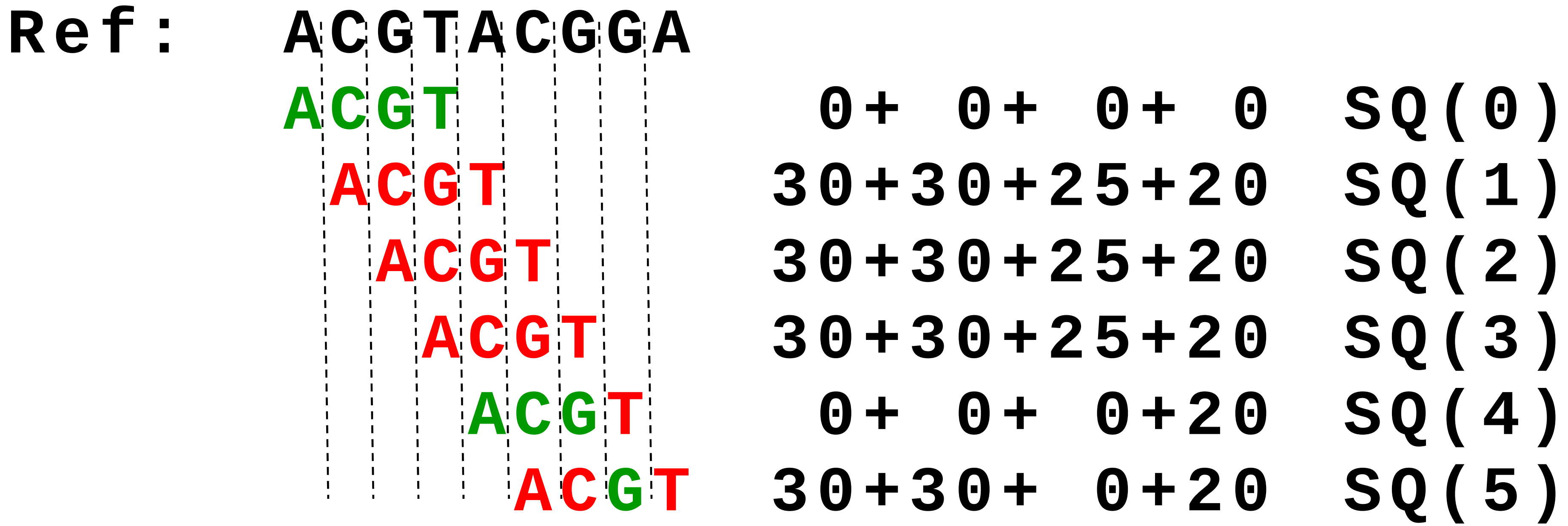
Mapping Quality

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.)

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

```
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)
```

Copyright © Peking University

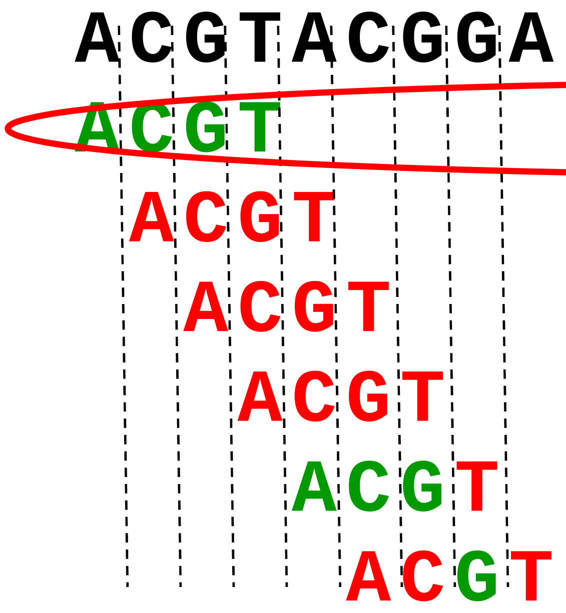
Mapping Quality

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

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

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

Read: ACGT (Quality: 30 30 25 20)

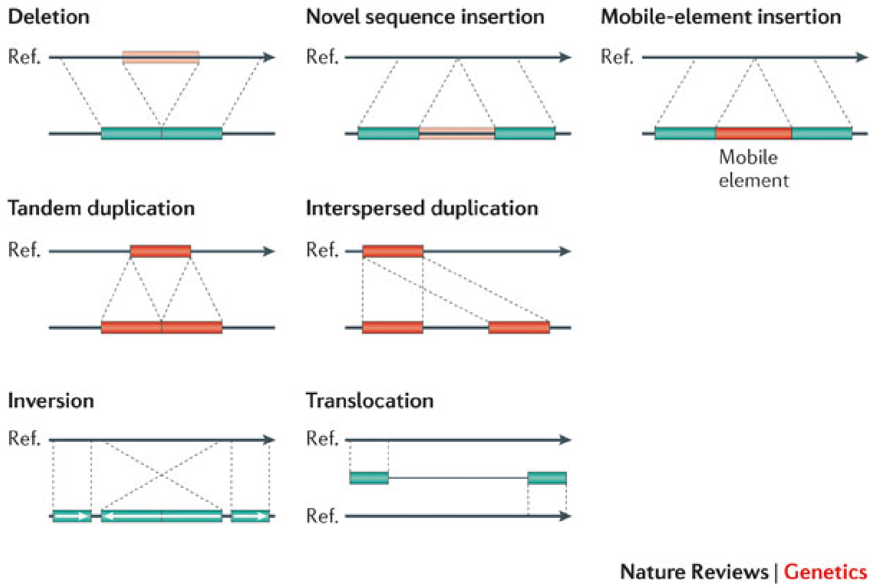


Ref:	ACGTACGGA	SQ(*u*)	E(*u*)
	ACGT	0+ 0+ 0+ 0	0/415
	ACGT	30+30+25+20	105/415
	ACGT	30+30+25+20	105/415
	ACGT	30+30+25+20	105/415
	ACGT	0+ 0+ 0+20	20/415
	ACGT	30+30+ 0+20	80/415

Genetic Variants

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

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



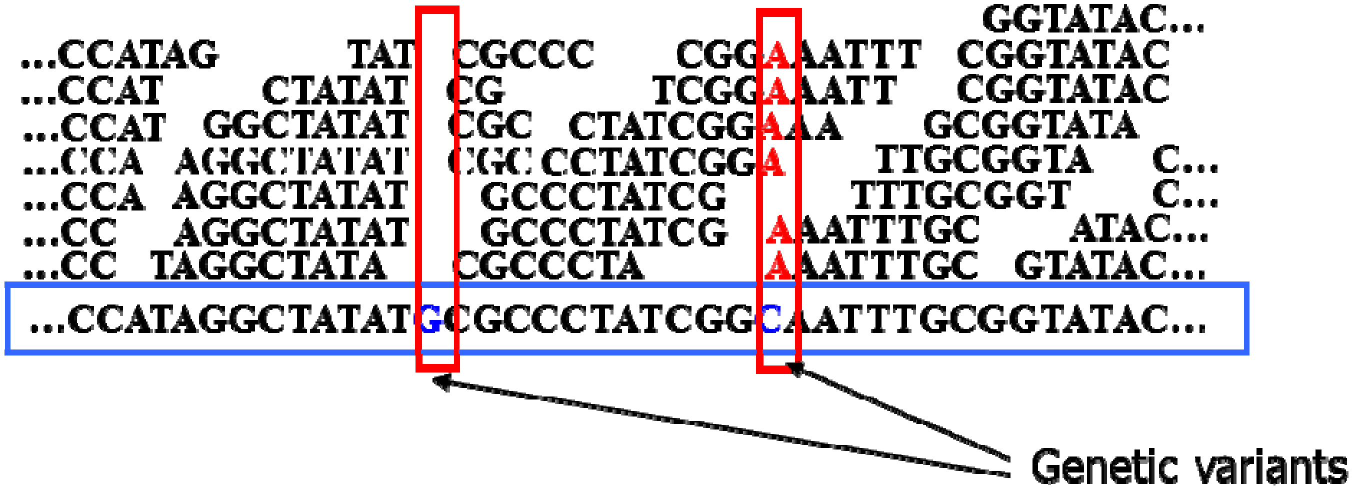
Nature Reviews | Genetics

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.”

(Source: *Nature Reviews Genetics* 12, 443-451)

Counting: an intuitive (and naïve) approach



GGTATAC...
...CCATAG TAT CGCCC CGGAAATTT CGGTATAC
...CCAT CTATAT CG TCGGAAATT CGGTATAC
...CCAT GGCTATAT CGC CTATCGGAAA GCGGTATA
...CCA AGGCTATAT CGC CCTATCGGA TTGCGGTA C...
...CCA AGGCTATAT GCCCTATCG TTTGCGGT C...
...CC AGGCTATAT GCCCTATCG AAATTTGC ATAC...
...CC TAGGCTATA CGCCCTA AAATTTGC GTATAC...
...CCATAGGCTATATGCGCCCTATCGGCAATTTGCGGTATAC...
Genetic variants

- 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

(Source: *Nature Reviews Genetics* 12, 443-451)

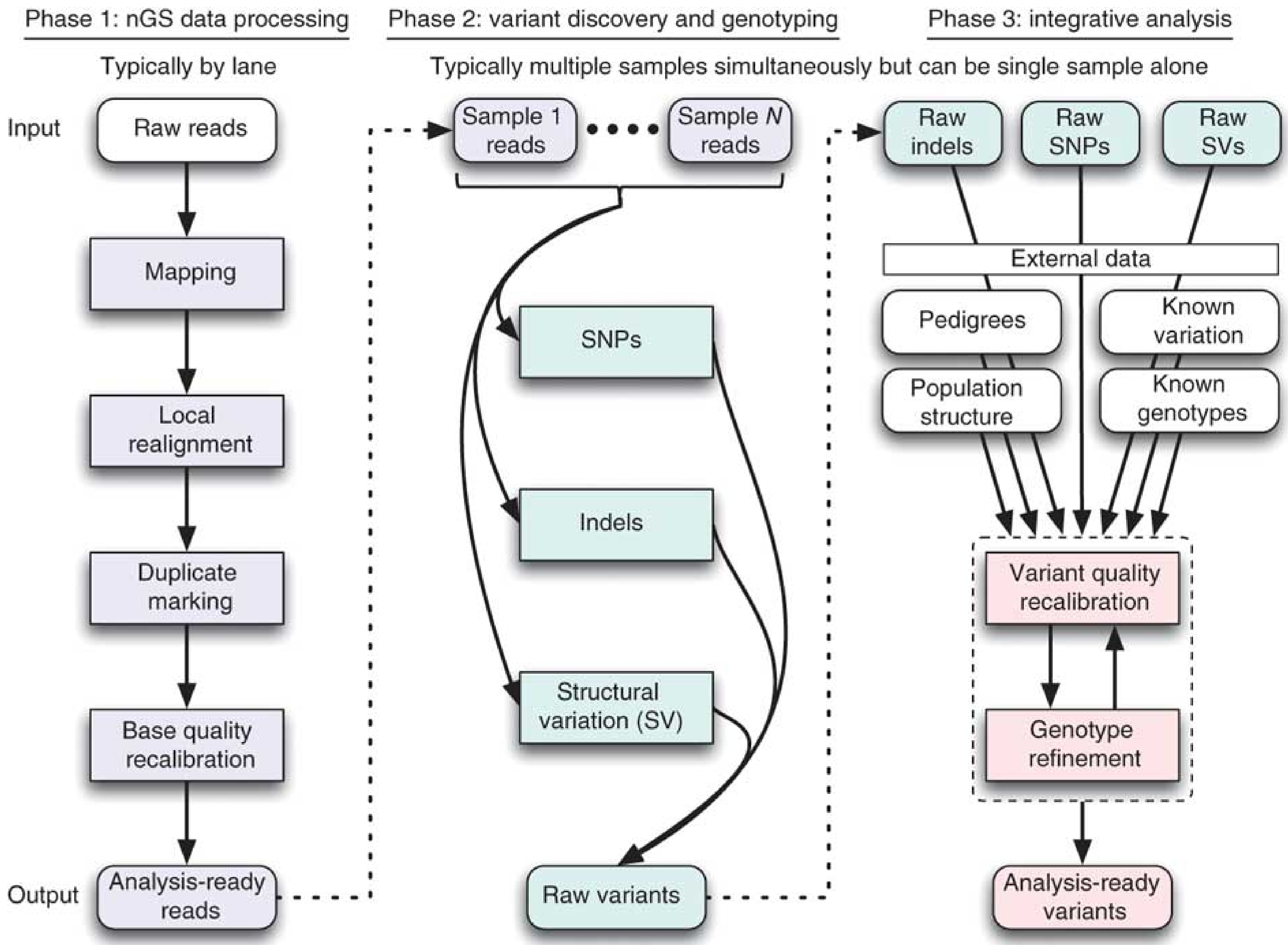
A Simple Probabilistic Model for Genotyping

1. For a diploid genome, there will be at most two different alleles (A and a) observed at a given site:
   - 3 possible genotypes: <A,A>, <A,a>, <a,a>
   - Number of A: k; Number of a: n-k

2. Then, the probability for each genotypes is
   - P(D|<A,A>) = the probability that we have (n-k) sequencing errors at this site $\prod_{n-k} P(x_i)$
   - Similarly, we can see the P(D|<a,a>) =$\prod_{k} P(x_i)$
   - P(D|<A,a>) = 1 – (P(D|<A,A>) + P(D|<a,a>))

3. Bayes Formula can be further employed to calculate posterior probabilities, i.e. P(<A,A>|D), P(<a,a>|D), and P(<A,a>|D) if we can estimate the prior probabilities P(<A,A>), P(<a,a>) and P(<A,a>)

<u>G</u>enome <u>A</u>nalysis <u>T</u>ool<u>K</u>it (GATK)



Nature Genetics 43, 491–498 (2011)

生物信息学：导论与方法

Bioinformatics: Introduction and Methods

https://www.coursera.org/course/pkubioinfo