Data Structures and Algorithms Underlying Genome Reconstruction from Short Reads

Bruce F. Cockburn

Department of Electrical and Computer Engineering
University of Alberta
Edmonton, AB

Outline

1. The Bioinformatics Revolution
2. DNA Sequence Reconstruction
3. Read Alignment Using the BWT Algorithm
4. An FPGA-based Hardware Accelerator
5. Conclusions

The Bioinformatics Revolution

- **Advances in DNA sequencing technology together with advances in the computer processing of biological data are the basis for a *Bioinformatics Revolution*.**
- **Many important potential applications:**
  - **Understanding the origins of genetically caused diseases**
  - **Development of more effective, targeted drugs**
  - **"Personalized medicine", where preventative care or the treatment of disease in a patient can be tailored to the genetics of that patient.**
- **Huge volumes of data must be processed. Computer Engineering can provide hardware accelerators to significantly speed up bioinformatics data processing.**

Next Generation DNA Sequencing Platforms

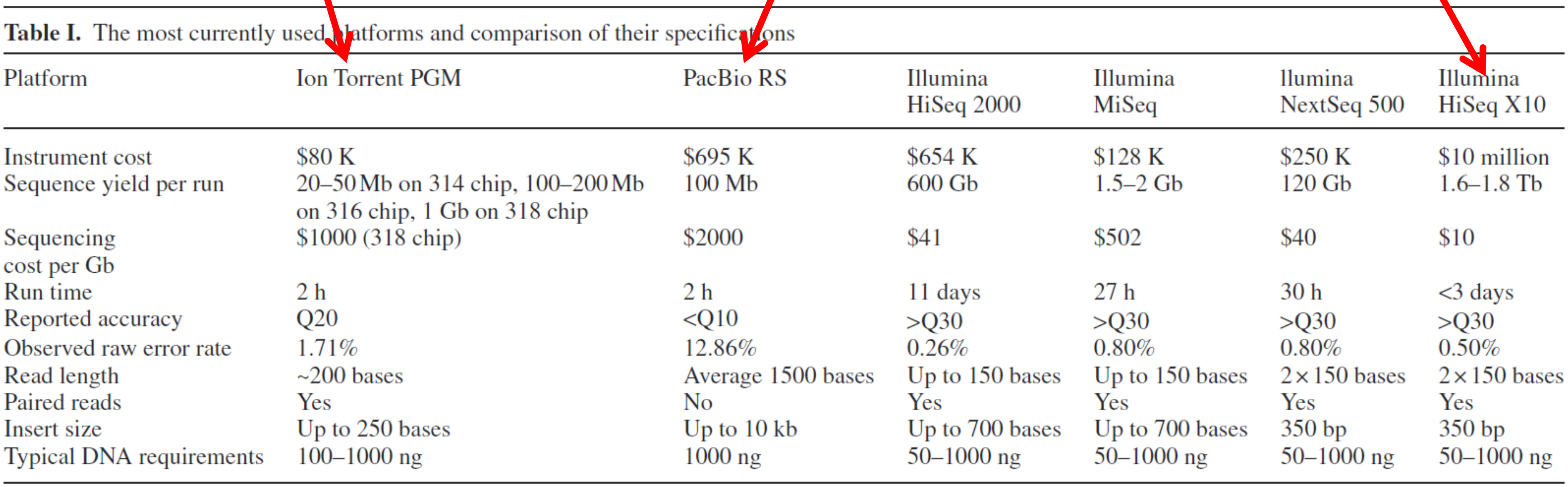


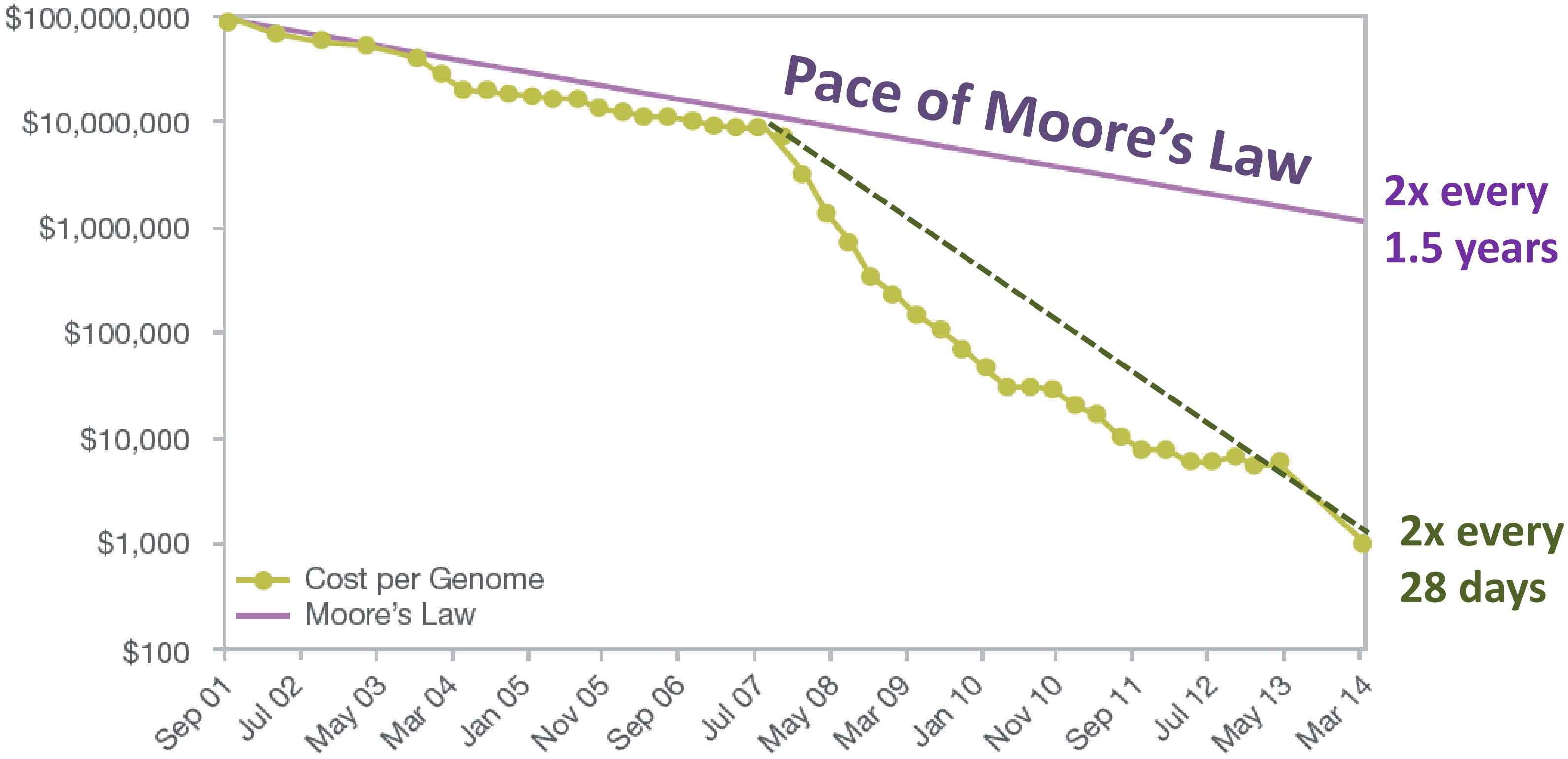
Table I. The most currently used platforms and comparison of their specifications

Platform	Ion Torrent PGM	PacBio RS	Illumina HiSeq 2000	Illumina MiSeq	llumina NextSeq 500	Illumina HiSeq X10
Instrument cost	$80 K	$695 K	$654 K	$128 K	$250 K	$10 million
Sequence yield per run	20–50 Mb on 314 chip, 100–200 Mb on 316 chip, 1 Gb on 318 chip	100 Mb	600 Gb	1.5–2 Gb	120 Gb	1.6–1.8 Tb
Sequencing cost per Gb	$1000 (318 chip)	$2000	$41	$502	$40	$10
Run time	2 h	2 h	11 days	27 h	30 h	<3 days
Reported accuracy	Q20	<Q10	>Q30	>Q30	>Q30	>Q30
Observed raw error rate	1.71%	12.86%	0.26%	0.80%	0.80%	0.50%
Read length	~200 bases	Average 1500 bases	Up to 150 bases	Up to 150 bases	2 × 150 bases	2 × 150 bases
Paired reads	Yes	No	Yes	Yes	Yes	Yes
Insert size	Up to 250 bases	Up to 10 kb	Up to 700 bases	Up to 700 bases	350 bp	350 bp
Typical DNA requirements	100–1000 ng	1000 ng	50–1000 ng	50–1000 ng	50–1000 ng	50–1000 ng

El Mustapha Bahassi and Peter J. Stambrook, "Next-generation sequencing technologies: breaking the sound barrier of human genetics," *Mutagenesis*, June 2014, vol. 29, no. 5, pp. 303-310.

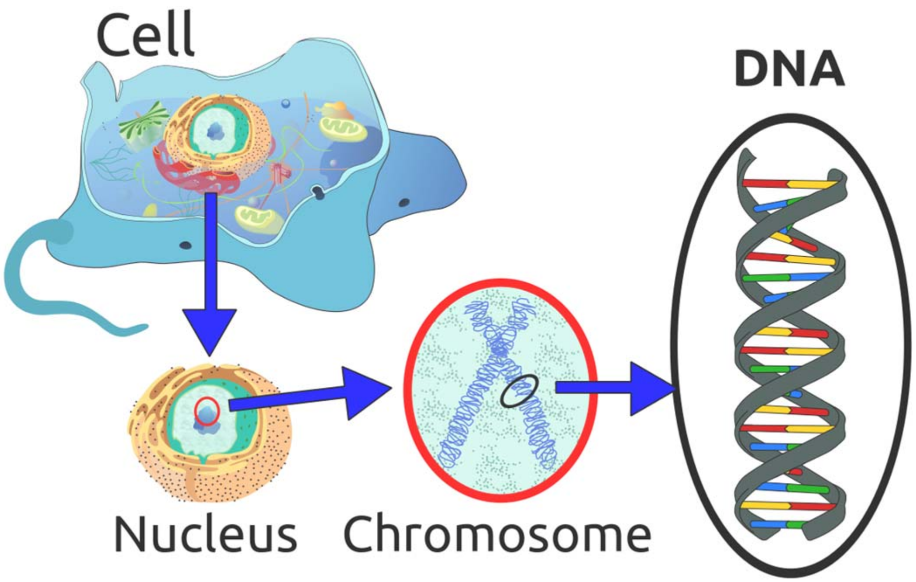
The $1,000 Human Genome

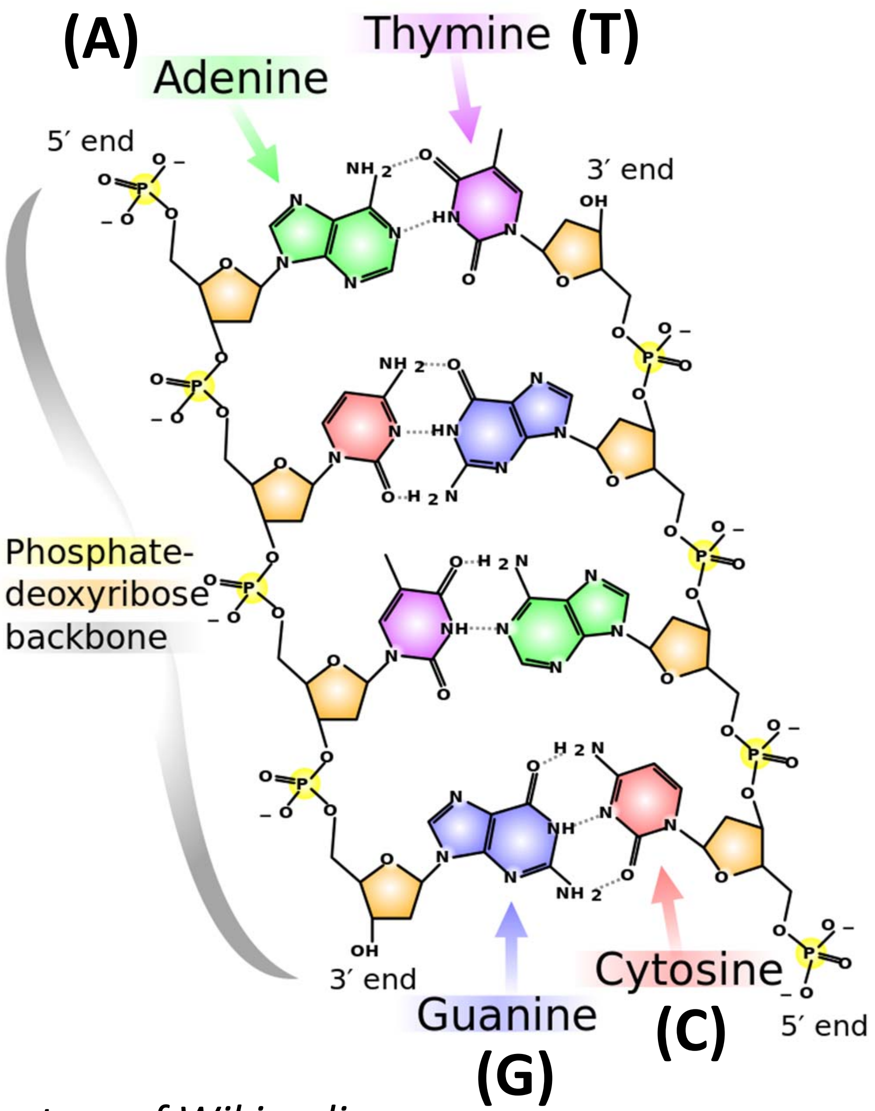
Figure 3: Illumina Sequencing Technology Outpaces Moore's Law for the Price of Whole Human Genome Sequencing



From Illumina's HiSeqX^{TM} Ten System Specification Sheet.

Deoxyribonucleic Acid (DNA)





- DNA has a ***double-helix*** structure, like a twisted ladder.
- The ladder rails are formed from ***deoxyribose*** sugar subunits.
- The ladder rungs are pairs of ***nucleotides*** of four kinds: (1) A-T, (2) T-A, (3) C-G, and (4) G-C.

Figures courtesy of Wikipedia

Nuclear DNA in 46 Chromosomes

- Each cell in the human body contains one copy of the human genome partitioned into 46 chromosomes.
- The chromosomes occur in pairs: there are 22 pairs of autosomal chromosomes plus two sex chromosomes (two X's in women; one X and one Y in men)..

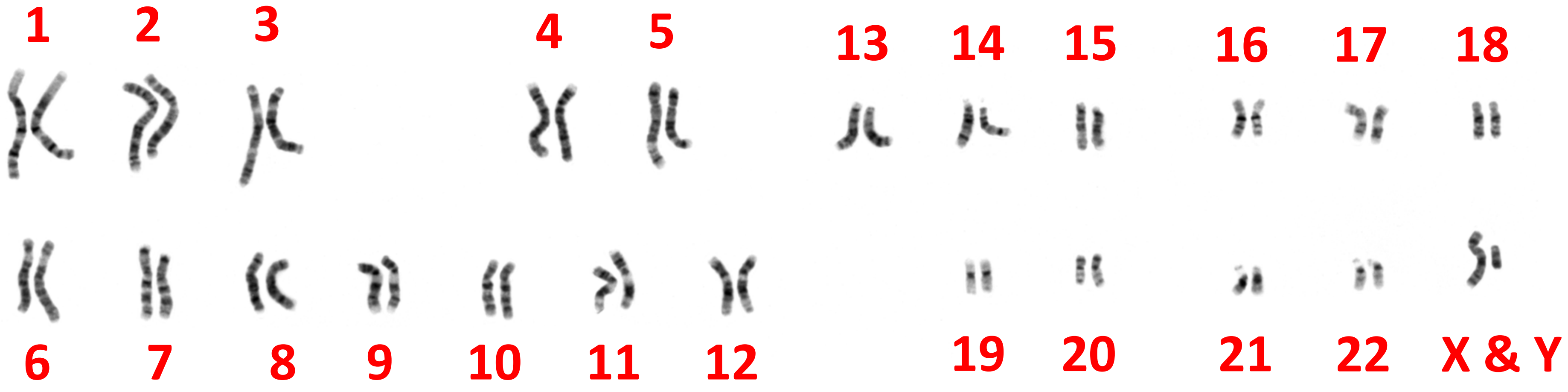
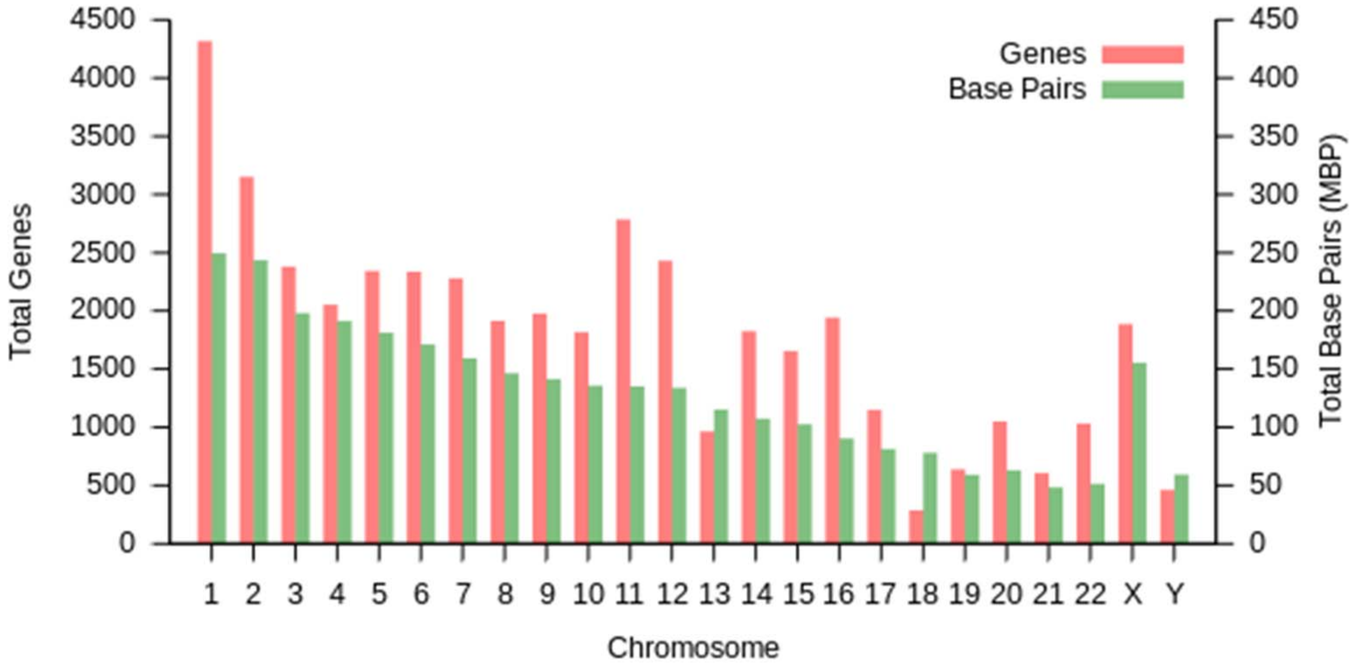


Figure modified from an original figure courtesy of Wikipedia

One genome is a lot of data!



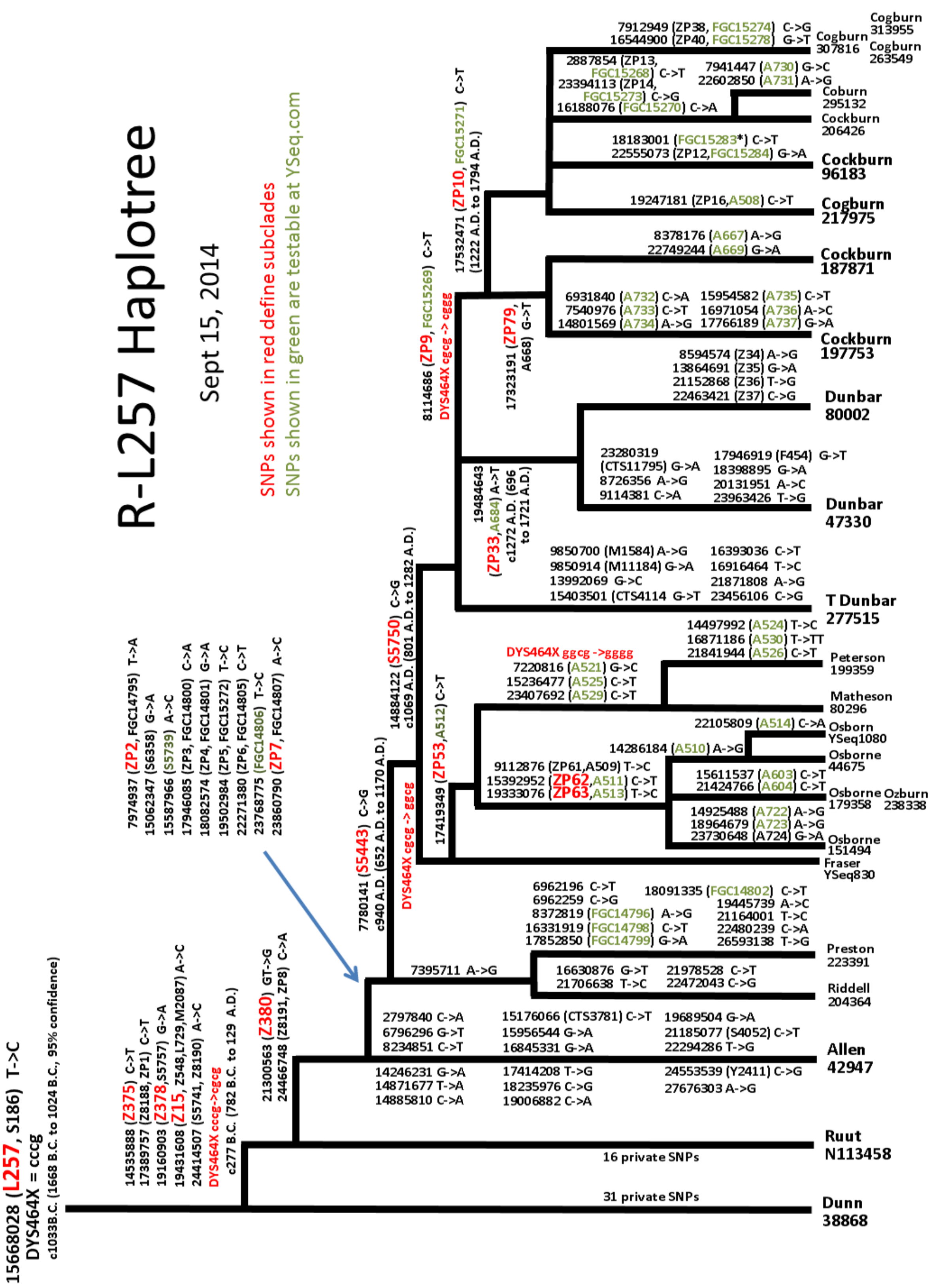
- The human genome contains 3.2 billion base pairs, which corresponds to 800 Mbytes of raw data.
- The genome specifies about 21,000 genes, which create 250,000 to 1,000,000 different proteins.

Figure courtesy of Wikipedia

Mutations in DNA

- **A mutation in DNA, RNA or a protein is a change in the sequence of subunits that comprise the polypeptide.**
  - ➢ ***substitution*: a subunit is replaced with one other**
  - ➢ ***deletion*: one or more subunits are removed**
  - ➢ ***insertion*: one or more subunits are inserted**
- **In DNA, mutations are created during meiosis, when gamete (sex) cells are created in the parents.**
- **Mutations can also occur as result of errors during DNA replication, RNA formation and protein synthesis. Biological processes are complex and occasionally go wrong. Environmental factors can trigger mutations.**

Ex: Substitution Mutations in the Y Chromosome



R-L257 Haplotree

Sept 15, 2014

SNPs shown in red define subclades

SNPs shown in green are testable at YSeq.com

15668028 (L257, S186) T->C
DYS464X = cccg
c1033B.C. (1668 B.C. to 1024 B.C., 95% confidence)

31 private SNPs

Dunn 38868

14535888 (Z375) C->T
17389757 (Z8188, ZP1) C->T
19160903 (Z378,S5757) G->A
19431608 (Z15, Z548,L729,M2087) A->C
24414507 (S5741, Z8190) A->C
DYS464X cccg->cgcg
c277 B.C. (782 B.C. to 129 A.D.)

16 private SNPs

Ruut N113458

21300563 (Z380) GT->G
24466748 (Z8191, ZP8) C->A

14246231 G->A
14871677 T->A
14885810 C->A
17414208 T->G
18235976 C->G
19006882 C->A
24553539 (Y2411) C->G
27676303 A->G

Allen 42947

2797840 C->A
6796296 G->T
8234851 C->T
15176066 (CTS3781) C->T
15956544 G->A
16845331 G->A
19689504 G->A
21185077 (S4052) C->T
22294286 T->G

7974937 (ZP2, FGC14795) T->A
15062347 (S6358) G->A
15587966 (S5739) A->C
17946085 (ZP3, FGC14800) C->A
18082574 (ZP4, FGC14801) G->A
19502984 (ZP5, FGC15272) T->C
22271380 (ZP6, FGC14805) C->T
23768775 (FGC14806) T->C
23860790 (ZP7, FGC14807) A->C

7395711 A->G

16630876 G->T
21706638 T->C
21978528 C->T
22472043 C->G

Riddell 204364

6962196 C->T
6962259 C->G
8372819 (FGC14796) A->G
16331919 (FGC14798) C->T
17852850 (FGC14799) G->A
18091335 (FGC14802) C->T
19445739 A->C
21164001 T->C
22480239 C->A
26593138 T->G

Preston 223391

7780141 (S5443) C->G
c940 A.D. (652 A.D. to 1170 A.D.)
DYS464X cgcg -> ggcg

Fraser YSeq830

17419349 (ZP53,A512) C->T

9112876 (ZP61,A509) T->C
15392952 (ZP62,A511) C->T
19333076 (ZP63,A513) T->C

14925488 (A722) A->G
18964679 (A723) A->G
23730648 (A724) G->A

Osborne 151494

15611537 (A603) C->T
21424766 (A604) C->T

Osborne 179358
Ozburn 238338

14286184 (A510) A->G

Osborne 44675

22105809 (A514) C->A

Osborn YSeq1080

DYS464X ggcg ->gggg
7220816 (A521) G->C
15236477 (A525) C->T
23407692 (A529) C->T

Matheson 80296

14497992 (A524) T->C
16871186 (A530) T->TT
21841944 (A526) C->T

Peterson 199359

14884122 (S5750) C->G
c1069 A.D. (801 A.D. to 1282 A.D.)

9850700 (M1584) A->G
9850914 (M11184) G->A
13992069 G->C
15403501 (CTS4114 G->T
16393036 C->T
16916464 T->C
21871808 A->G
23456106 C->G

T Dunbar 277515

8114686 (ZP9, FGC15269) C->T
DYS464X cgcg -> cggg

19484643
(ZP33,A684) A->T
c1272 A.D. (696
to 1721 A.D.)

23280319
(CTS11795) G->A
8726356 A->G
9114381 C->A
17946919 (F454) G->T
18398895 G->A
20131951 A->C
23963426 T->G

Dunbar 47330

8594574 (Z34) A->G
13864691 (Z35) G->A
21152868 (Z36) T->G
22463421 (Z37) C->G

Dunbar 80002

17323191 (ZP79,
A668) G->T

6931840 (A732) C->A
7540976 (A733) C->T
14801569 (A734) A->G
15954582 (A735) C->T
16971054 (A736) A->C
17766189 (A737) G->A

Cockburn 197753

8378176 (A667) A->G
22749244 (A669) G->A

Cockburn 187871

17532471 (ZP10, FGC15271) C->T
(1222 A.D. to 1794 A.D.)

19247181 (ZP16,A508) C->T

Cogburn 217975

18183001 (FGC15283*) C->T
22555073 (ZP12,FGC15284) G->A

Cockburn 96183

2887854 (ZP13,
FGC15268) C->T
23394113 (ZP14,
FGC15273) C->G
16188076 (FGC15270) C->A

Cockburn 206426

7941447 (A730) G->C
22602850 (A731) A->G

Coburn 295132

7912949 (ZP38, FGC15274) C->G
16544900 (ZP40, FGC15278) G->T

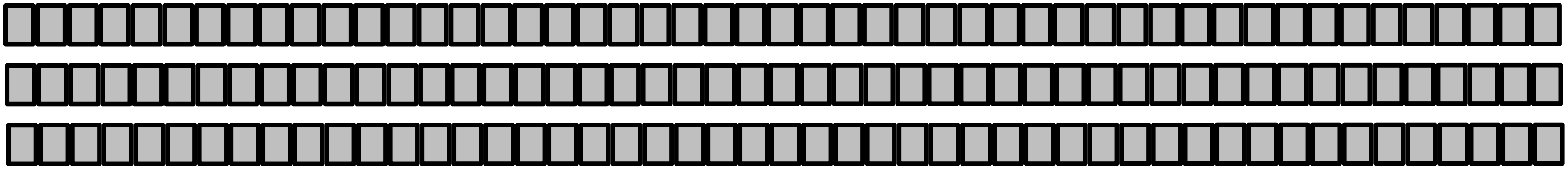
Cogburn 307816
Cogburn 313955
Cogburn 263549

Next Generation Sequencing of DNA

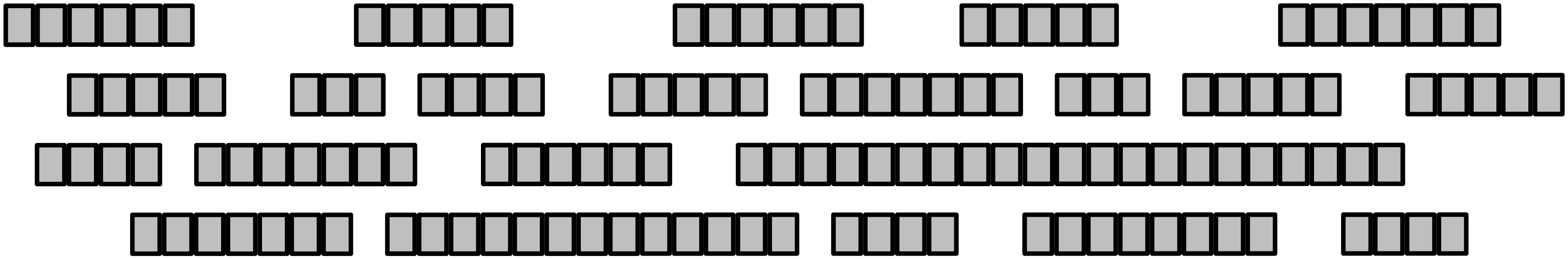
- In *Next Generation Sequencing* (NGS), the DNA in a large number of identical genomes is cleaved at random locations to create a large number of short segments called “short reads”.
- Each short read contains ~30 to ~200 symbols.
- Sequencing machines are used to sequence the millions of short reads in parallel.
- The resulting short read sequences must then be processed by computer to recreate the original genome.

Creation of Short Reads

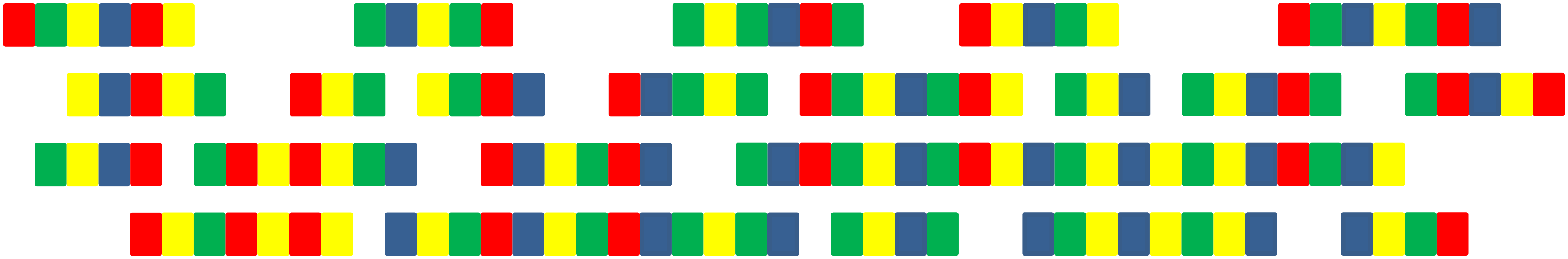
1) *Prepare Purified Sample of the Genome*



2) *Cleave the Sequences into Reads of Length ~30 to ~200*



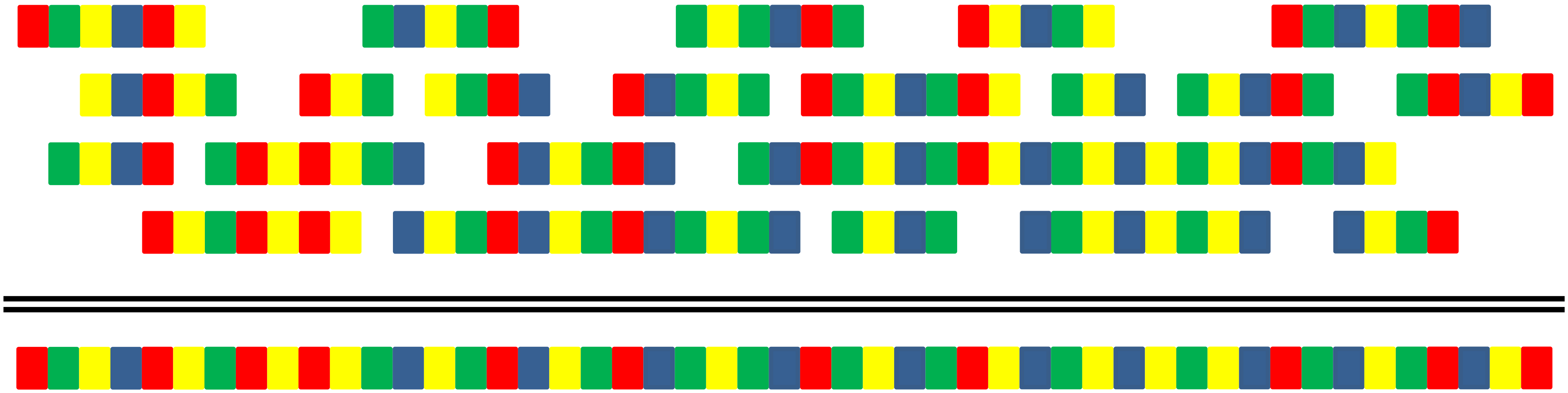
3) *Sequence the Short Reads using Parallel Techniques*



De Novo Reconstruction of Genomes

- When the genome of *new* organism needs to be sequenced, the short reads need to be assembled on their own into the most likely underlying genome.
- This is a very tough problem. DNA tends to have many regions where DNA subsequences are repeated many times. This leads to ambiguity.

Sequenced Short Reads

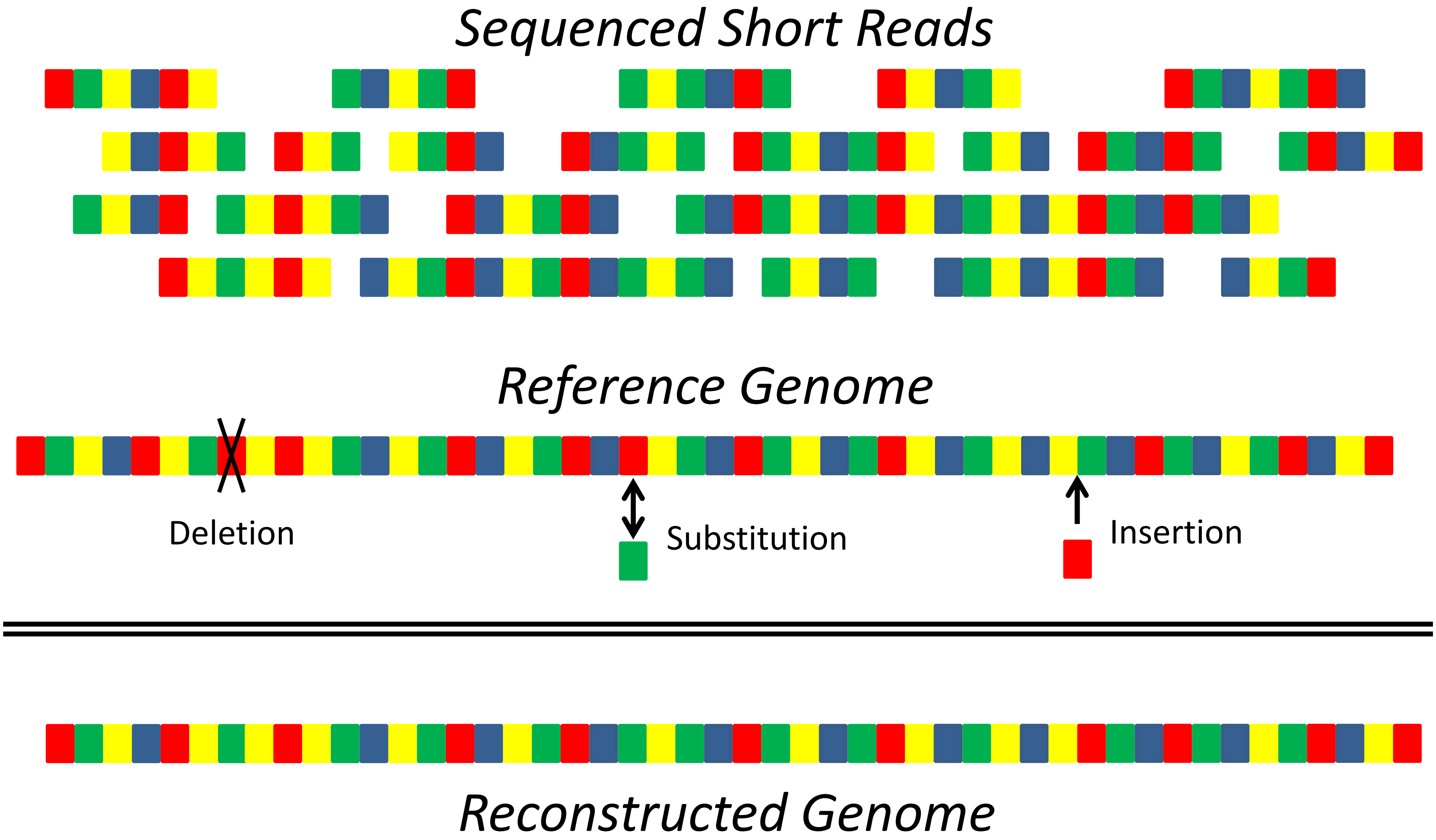


Reconstructed Genome

Mapping Short Reads to a Reference Genome

- *De novo* genome assembly is not required in many cases where there are already sequenced examples of the genome from the same species.
- For example, to assemble the genome of a particular human, we can exploit knowledge of the already sequenced “average” or ***reference human genome***.
- The sequenced short reads from the next generation sequencing machine can be “mapped” onto the reference genome, allowing for a reasonable number of substitutions, deletions and insertions.
- This is the ***short read mapping problem***.

Genome Reconstruction Using a Reference



Sequenced Short Reads

Reference Genome

Deletion

Substitution

Insertion

Reconstructed Genome

The Short Read Alignment Problem

Given:

- A reference human genome consisting of a sequence containing 3.2 billion symbols from the alphabet {A, C, G, T}.
- Over 500 million “short read” subsequences, each about 150 symbols long drawn from the same alphabet and taken from some unknown mutated genome.
- An assumed maximum number of mutations (substitutions, deletions, insertions) allowed in any read, usually 1, 2 or 3.

Objective:

- For each short read, find all local alignments against the reference genome using the smallest number of mutations, up to the given limit.
- In then last step, reconcile the local alignments to reconstruct the unknown mutated genome.

Examples of Short Read Alignments

Given:

Reference: CTAGATGATATACAGCTG

Short Read: AGAT

Alignments up to 1 mutation:

0 mutations: CTAGATGATATACAGCTG

1 substitution: CTAGATGATATACAGATG

1 deletion: CTAGA-GATATACAGCTG

1 insertion: CTAGATAGATATACAGCTG

Algorithms for Fast Short Read Alignment

- The short read alignment operation must be repeated 100s of millions of times in order to reconstruct one human genome.
- There is a need for a data structure that is:
  - ❖ sufficiently compact to store the necessary information from the genome reference
  - ❖ provides very fast to the reference to support the short read alignment operation
- Several data structures have been proposed.
- Many of the most advanced short read alignment algorithms use a data structure that is based on the Burrows-Wheeler Transform (BWT)

Burrows-Wheeler Transform (BWT)

1) Ref: CTAGATGATATA

2) Rotation matrix:

0	$CTAGATGATATA
1	CTAGATGATATA$
2	TAGATGATATA$C
3	AGATGATATA$CT
4	GATGATATA$CTA
5	ATGATATA$CTAG
6	TGATATA$CTAGA
7	GATATA$CTAGAT
8	ATATA$CTAGATG
9	TATA$CTAGATGA
10	ATA$CTAGATGAT
11	TA$CTAGATGATA
12	A$CTAGATGATAT

3) Sorted rotation matrix:

0	$CTAGATGATATA
12	A$CTAGATGATAT
3	AGATGATATA$CT
10	ATA$CTAGATGAT
8	ATATA$CTAGATG
5	ATGATATA$CTAG
1	CTAGATGATATA$
7	GATATA$CTAGAT
4	GATGATATA$CTA
11	TA$CTAGATGATA
2	TAGATGATATA$C
9	TATA$CTAGATGA
6	TGATATA$CTAGA

4) BWT: ATTTGG$TAACAA

Auxiliary Data Derived from the BWT

3) Sorted rotation matrix:

0	0		$CTAGATGATATA
12	1	→	A$CTAGATGATAT
3	2		AGATGATATA$CT
10	3		ATA$CTAGATGAT
8	4		ATATA$CTAGATG
5	5		ATGATATA$CTAG
1	6	→	CTAGATGATATA$
7	7	→	GATATA$CTAGAT
4	8		GATGATATA$CTA
11	9	→	TA$CTAGATGATA
2	10		TAGATGATATA$C
9	11		TATA$CTAGATGA
6	12		TGATATA$CTAGA

5) Occurrence matrix:

$	A	C	G	T
0	1	0	0	0
0	1	0	0	1
0	1	0	0	2
0	1	0	0	3
0	1	0	1	3
0	1	0	2	3
1	1	0	2	3
1	1	0	2	4
1	2	0	2	4
1	3	0	2	4
1	3	1	2	4
1	4	1	2	4
1	5	1	2	4

4) BWT: ATTTGG$TAACAA

6) C vector: (1,6,7,9)

Size of the BWT Occurrence Matrix

- Genome (reference) size: 3.2 billion x 2 bits = 800 MB
- Upper bound on occurrence matrix entry: 32 bits
- Occurrence matrix size: 3.2 billion x 4 x 4 bytes < 8 GB
- The 8-gigabyte size can be easily stored in a fast DRAM
- Further compression of the occurrence matrix would impose an undesirable access time penalty

Fast Suffix Search Algorithm

- Initialize two row indexes, *K* and *L*, to the smallest and largest possible row indexes for the occurrence matrix.
- Scan the short read $s = s_n \ldots s_1$ from right to left, and update *K* and *L* as follows for each symbol s_j:

  If $K_i = 0$, then $K_{i+1} = \text{c\_vec}[s_j]$

  If $K_i > 0$, then $K_{i+1} = \text{c\_vec}[s_j] + \text{occ\_mat}[K_i\text{-}1][\text{s}_j] – 1$

  $L_{i+1} = \text{c\_vec}[\text{s}_j] + \text{occ\_mat}[L_i][\text{s}_j] – 1$
- All occurrences of s appear in rows K_n to L_n inclusive of the rotation matrix.
- If $K_n > L_n$, then the read is not present in the reference

Example of Fast Suffix Search

- Find all occurrences of A, TA, ATA and CATA in the same reference sequence as before.

Ref = $\mathtt{\$}\underset{1}{\mathtt{C}}\mathtt{TAGATGATAT}\underset{12}{\mathtt{A}}$

			Positions in reference
	$K_0 = 0$	$L_0 = 12$	
s_0 = A	$K_1 = 1$	$L_1 = 1 + 5 – 1 = 5$	12, 3, 10, 8, 5
s_1 = T	$K_2 = 9 + 0 = 9$	$L_2 = 9 + 3 - 1 = 11$	11, 2, 9
s_2 = A	$K_3 = 1 + 2 = 3$	$L_3 = 1 + 4 - 1 = 4$	10, 8
s_3 = C	$K_4 = 7 + 0 = 7$	$L_4 = 7 + 0 - 1 = 6$	*Not present*

BWT-based Short Read Alignment *

```
Input:  BWT string $B$ for reference string $X$
        Array $C(.)$ and $O(.,.)$ from $B$
        BWT string $B'$ for the reverse of reference $X$
        Array $O'(.,.)$ from $B'$
Output: Suffix-array intervals
Procedures:
Calculated($W$)
begin
    Calculate $D(i)$ that gives a lower bound for the
    number of mismatches in $W$
end
InexRecur($W, i, z, k, l$)
begin
    /* On next slide */
end
main($W, z$)
begin
    Calculated(W)
    InexRecur($W, i, z, k, l$)
end
```

Reference sequence, X = (C,C,T,G,A,G,$)

Short read, W = (C,G,A)

W = short read to be mapped to reference X

i = index into short read W, initially 3

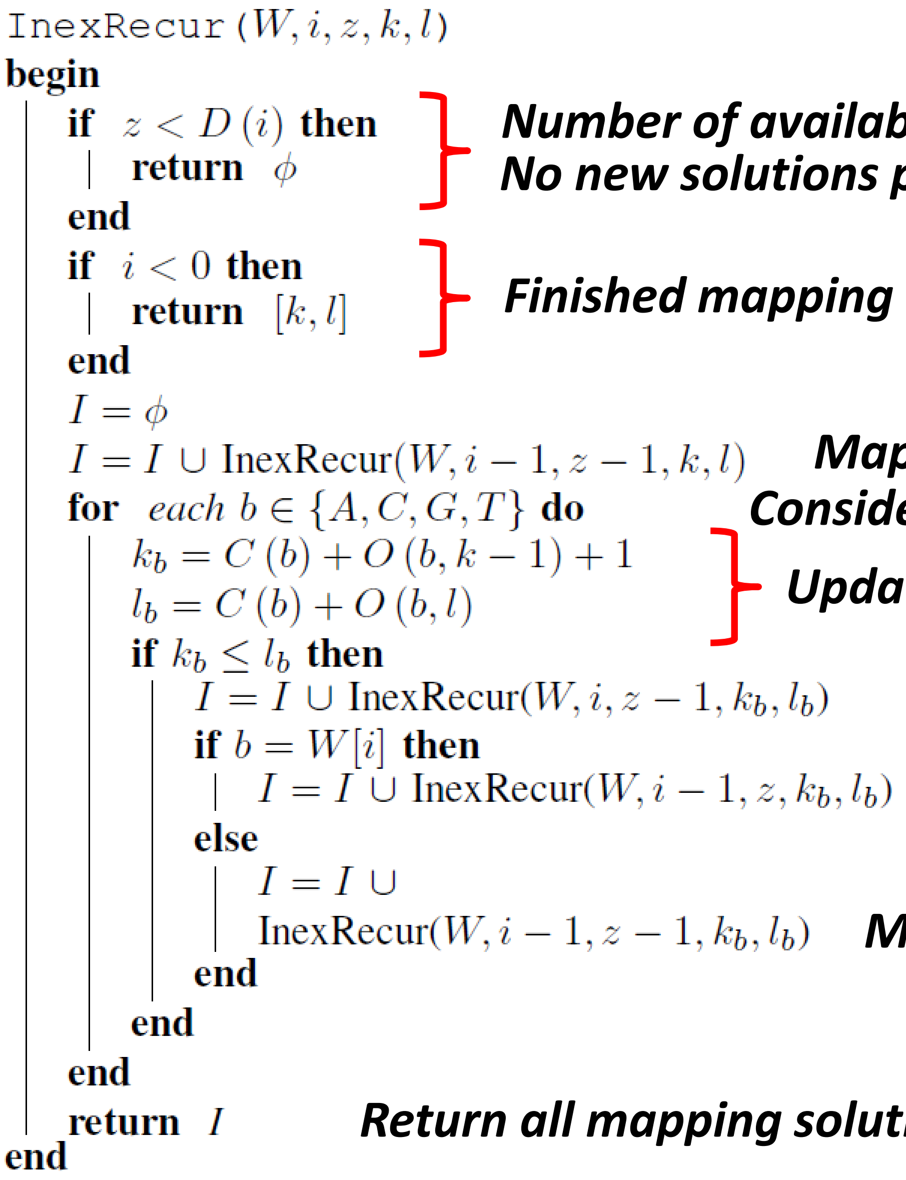
z = number of allowed mutations, initially 2

k = lower index into B(X), initially 0

l = upper index into B(X), initially 6

* Li & Durbin, *Bioinformatics*, vol. 25, no. 14, 2009, pp. 1754-1760.

BWT-based Short Read Alignment [Li & Durbin]



```
InexRecur($W, i, z, k, l$)
begin
    if $z < D(i)$ then
        return $\phi$
    end
    if $i < 0$ then
        return $[k, l]$
    end
    $I = \phi$
    $I = I \cup \text{InexRecur}(W, i-1, z-1, k, l)$
    for each $b \in \{A, C, G, T\}$ do
        $k_b = C(b) + O(b, k-1) + 1$
        $l_b = C(b) + O(b, l)$
        if $k_b \leq l_b$ then
            $I = I \cup \text{InexRecur}(W, i, z-1, k_b, l_b)$
            if $b = W[i]$ then
                $I = I \cup \text{InexRecur}(W, i-1, z, k_b, l_b)$
            else
                $I = I \cup$
                $\text{InexRecur}(W, i-1, z-1, k_b, l_b)$
            end
        end
    end
    return $I$
end
```

Number of available mutations fell below lower bound D(i). No new solutions possible in this scenario. Return null!

Finished mapping W to X. Return the new solution [k,l].

Map next symbol with one insertion

Consider all four possible next symbols b

Update upper and lower indexes into O(-,-)

Delete one symbol from X and continue

Map next symbol without mutations

Map next symbol with one substitution

Return all mapping solutions back up to the calling routine

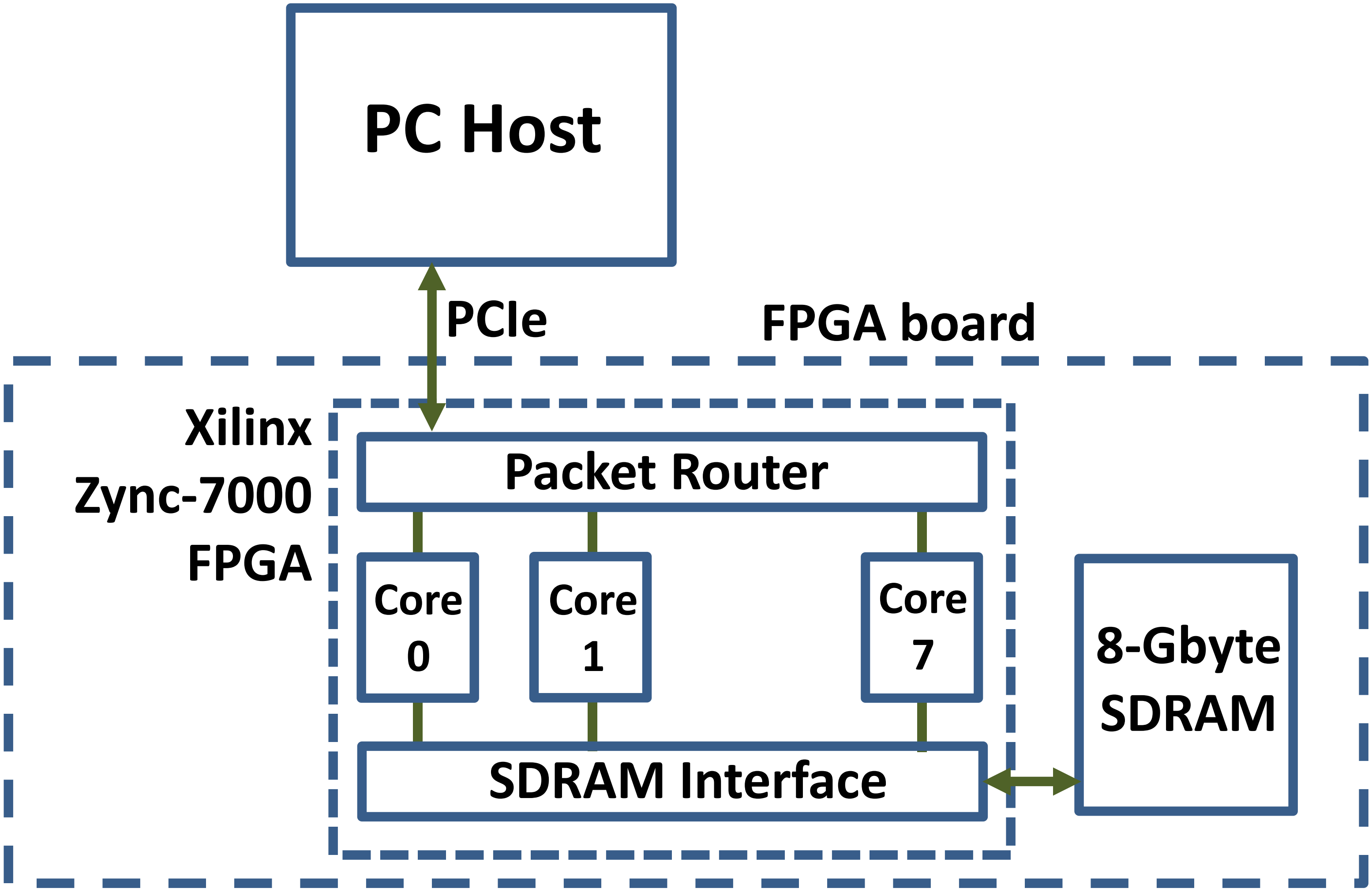
Some Recent Hardware Accelerators

- **J. Arram et al., "Hardware Acceleration of Genetic Sequence Alignment," *9th Int. Symp. on Applied Reconfigurable Computing*, Mar. 25-27, 2013, Los Angeles, pp. 13-24.**

- **J. Arram et al., "Reconfigurable Acceleration of Short Read Mapping," *2013 21st Ann. Int. IEEE Symp. on Field-Programmable Custom Computing Machines*, Apr. 28-30, Seattle, WA, pp. 210-217.**

- **J. Arram et al., "Reconfigurable Filtered Acceleration of Short Read Alignment," *2013 Int. Conf. on Field-Programmable Technology*, Dec. 9-11, Kyoto, pp. 441-438.**

New Accelerator Under Development



On-going and Future Work

- A prototype accelerator based on the Xilinx Zynq ZC702 Evaluation Board is nearing completion.
- This board is based around a Xilinx xc7z020clg484-1 field-programmable gate array (FPGA).
- One instance of the accelerator core requires 3133 flip-flops (~2% of 12,179 available), 6091 look-up tables (~11% of 53,200 LUTs) and 20 28-kbit block RAMs (~7% of 280).
- As many as 8 short read mapper cores can be fit onto this older FPGA.
- The recent Zynq xc7z100 FPGA is over 5 times bigger.

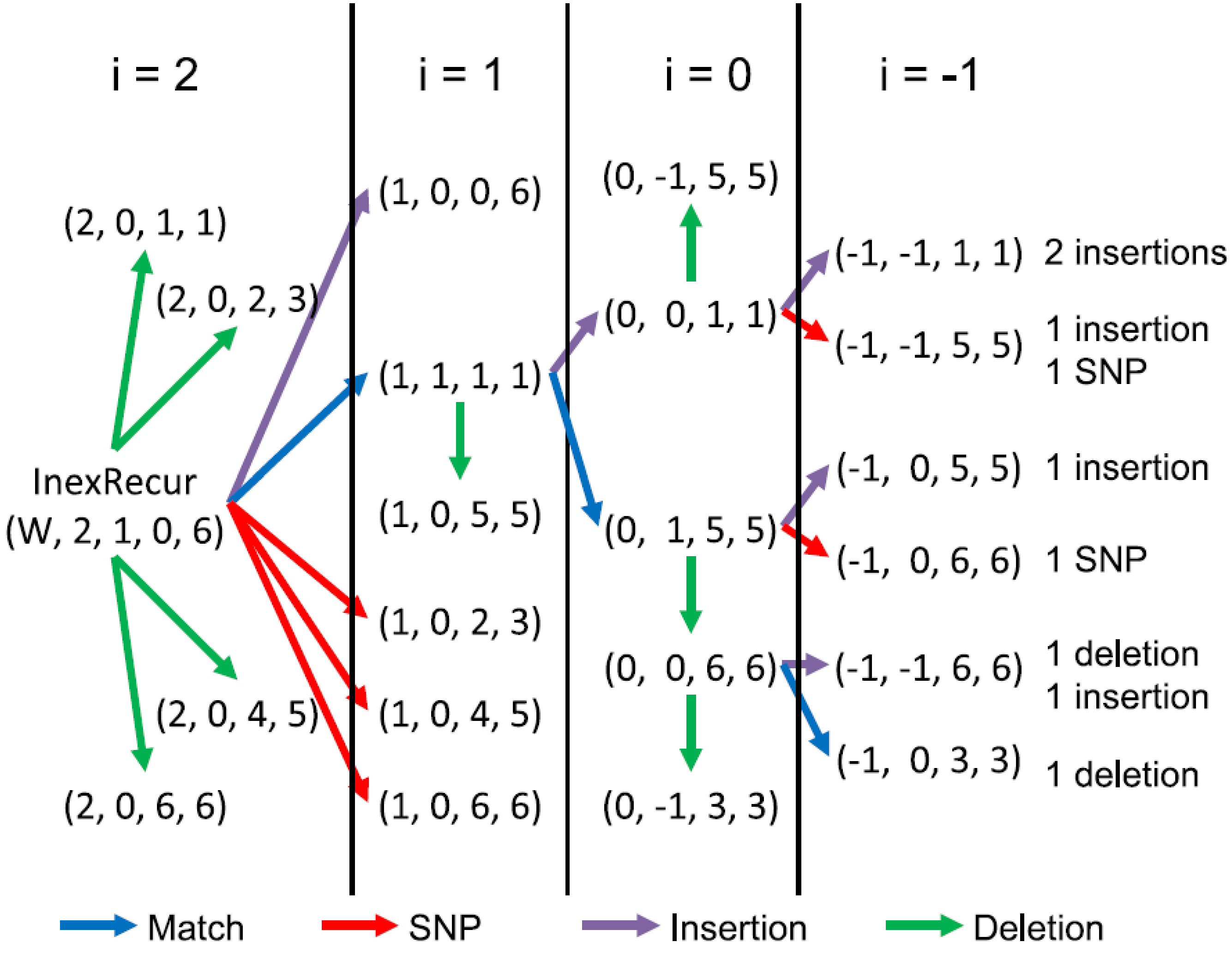
Conclusions

- The success of the bioinformatics revolution depends on advances that have been made in data structure and algorithm design.
- Genome reconstruction using short read mapping to a reference sequence is an increasingly serious data processing bottleneck.
- While workstation farms and cloud computing are the major computing platforms, steadily improving FPGA technology seems well suited to providing accelerators that can be optimized to the problem of genome reconstruction from short reads.

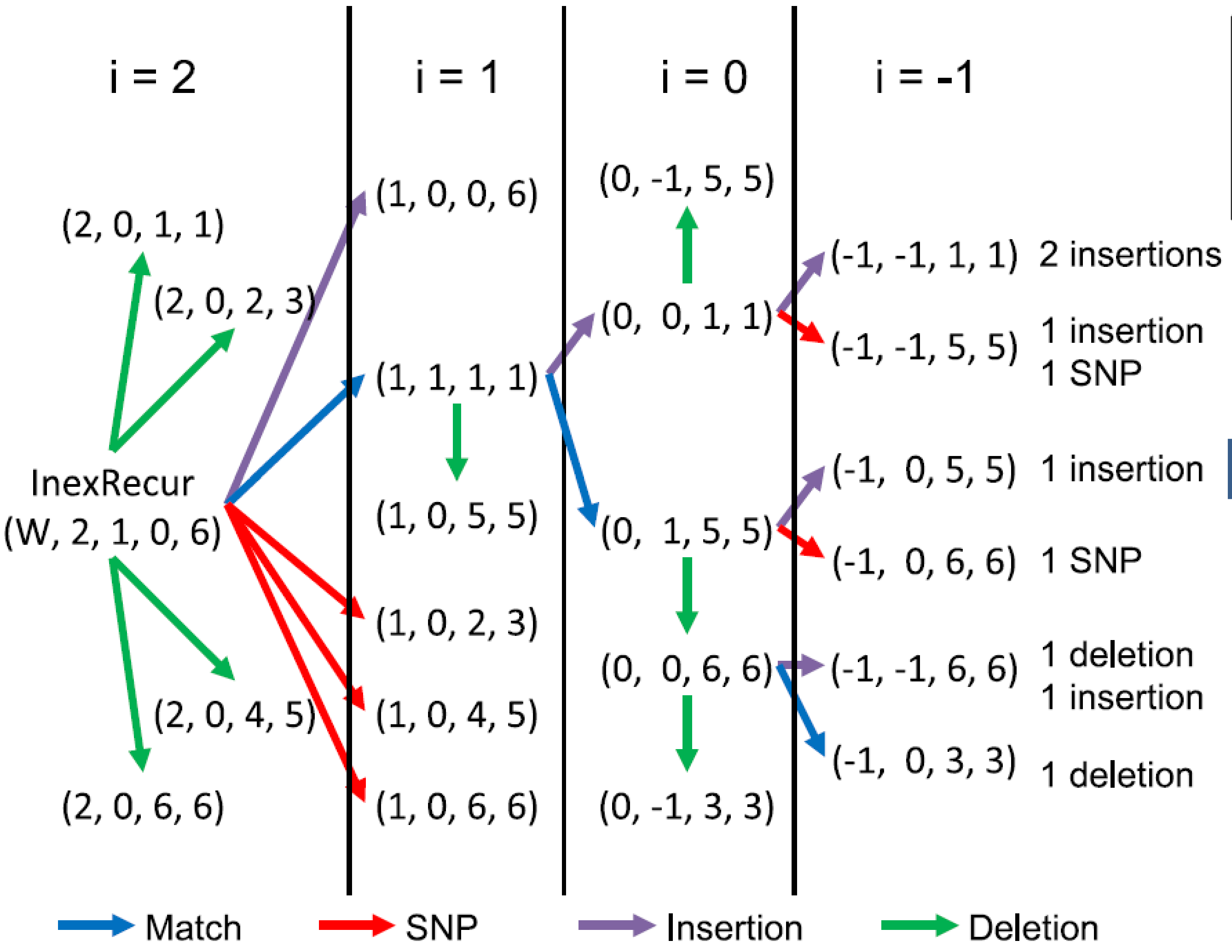
Extra Slides

Execution Pattern for the Mapping Algorithm

Reference sequence, X = (C,C,T,G,A,G,$) **Short read, W = (C,G,A)**



Read Mapping Using One Insertion into X



i = 2

(2, 0, 1, 1)

(2, 0, 2, 3)

InexRecur
(W, 2, 1, 0, 6)

(2, 0, 4, 5)

(2, 0, 6, 6)

i = 1

(1, 0, 0, 6)

(1, 1, 1, 1)

(1, 0, 5, 5)

(1, 0, 2, 3)

(1, 0, 4, 5)

(1, 0, 6, 6)

i = 0

(0, -1, 5, 5)

(0, 0, 1, 1)

(0, 1, 5, 5)

(0, 0, 6, 6)

(0, -1, 3, 3)

i = -1

(-1, -1, 1, 1) 2 insertions

(-1, -1, 5, 5) 1 insertion
1 SNP

(-1, 0, 5, 5) 1 insertion

(-1, 0, 6, 6) 1 SNP

(-1, -1, 6, 6) 1 deletion
1 insertion

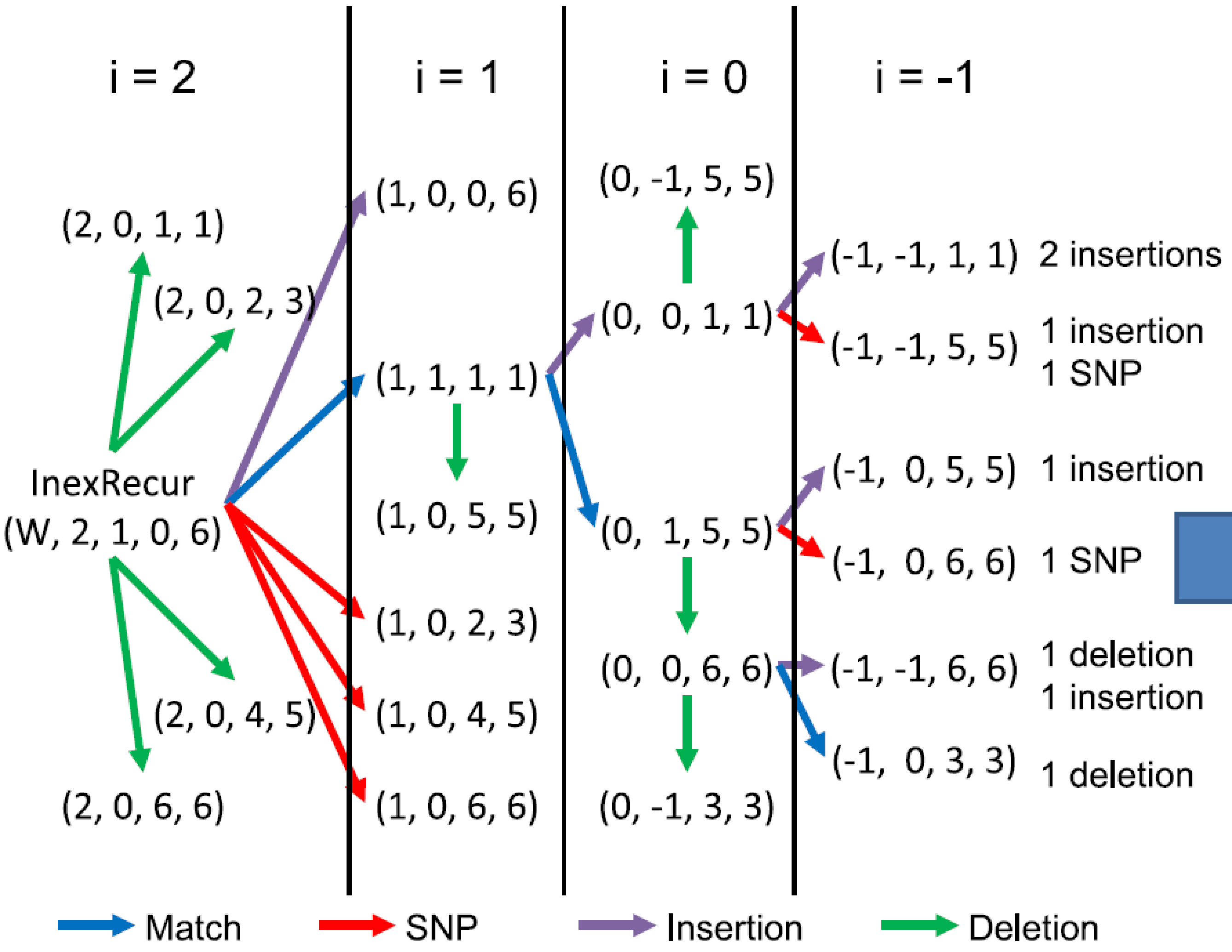
(-1, 0, 3, 3) 1 deletion

Match SNP Insertion Deletion

Position	0	1	2	3	4	5	6
X	C	C	T	G	A	G	
W				C	G	A	
Result	C	C	T	C	G	A	G

Reference sequence, X = (C,C,T,G,A,G,$)

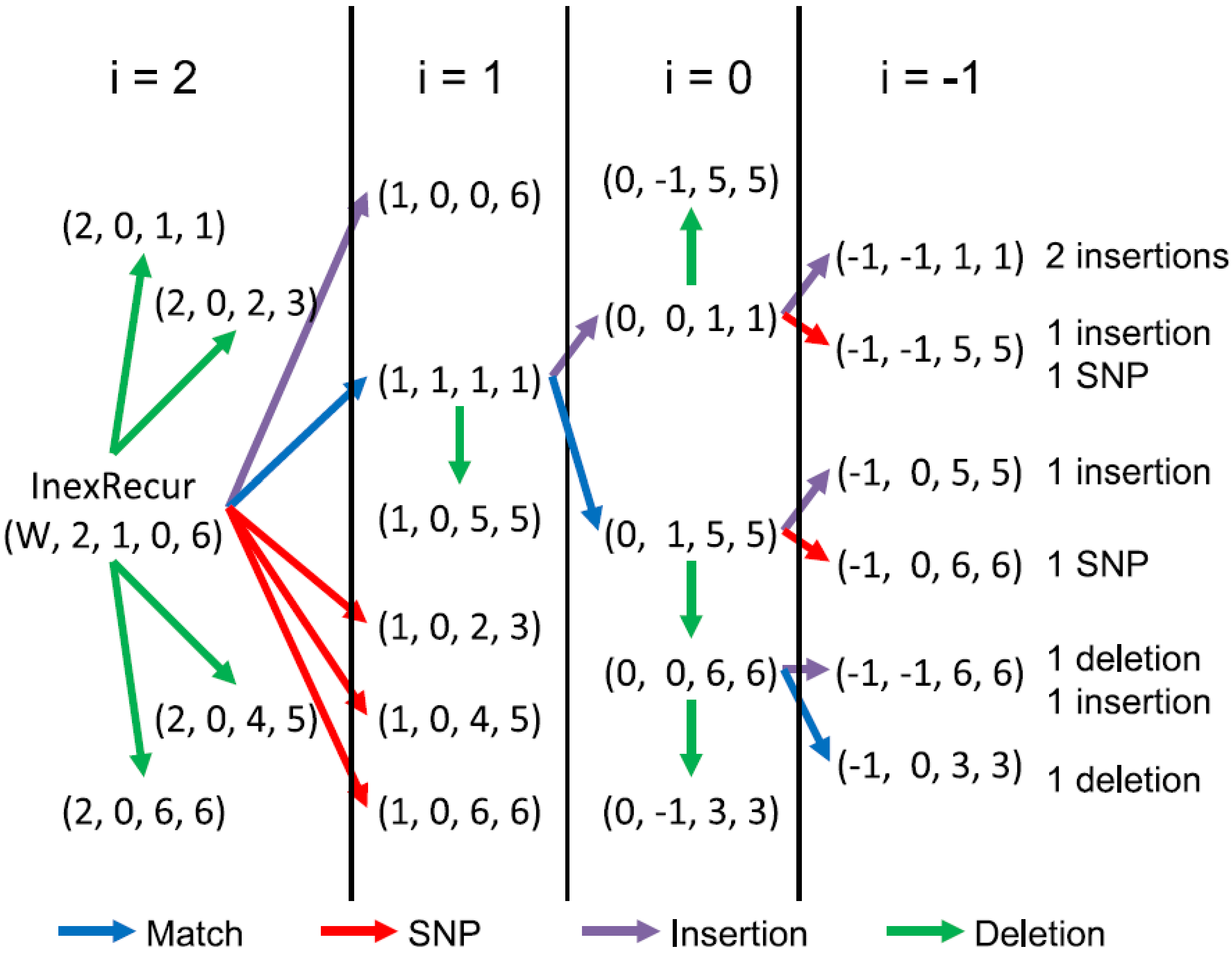
Read Mapping Using One SNP Mutation



Position	0	1	2	3	4	5	6
X	C	C	T	G	A	G	
W			C	G	A		
Result	C	C	C	G	A	G	

Reference sequence, X = (C,C,T,G,A,G,$)

Execution Pattern for the Mapping Algorithm



i = 2

i = 1

i = 0

i = -1

InexRecur (W, 2, 1, 0, 6)

(2, 0, 1, 1)

(2, 0, 2, 3)

(2, 0, 4, 5)

(2, 0, 6, 6)

(1, 0, 0, 6)

(1, 1, 1, 1)

(1, 0, 5, 5)

(1, 0, 2, 3)

(1, 0, 4, 5)

(1, 0, 6, 6)

(0, -1, 5, 5)

(0, 0, 1, 1)

(0, 1, 5, 5)

(0, 0, 6, 6)

(0, -1, 3, 3)

(-1, -1, 1, 1) 2 insertions

(-1, -1, 5, 5) 1 insertion 1 SNP

(-1, 0, 5, 5) 1 insertion

(-1, 0, 6, 6) 1 SNP

(-1, -1, 6, 6) 1 deletion 1 insertion

(-1, 0, 3, 3) 1 deletion

Match SNP Insertion Deletion

Position	0	1	2	3	4	5	6
X	C	C	T	G	A	G	
W		C	G	A			
Result	C	C	G	A	G		

Reference sequence, X = (C,C,T,G,A,G,$)

Three Key Molecules in Biology

- **Many of the key building blocks of organisms are long polymers whose structure is determined by a sequence of subunits (smaller molecules & residues).**
- ***Deoxyribonucleic Acid*** **(DNA) encodes genetic information that defines genes and hence proteins.**
- ***Ribonucleic Acid*** **(RNA) is used in the complex process of extracting genetic information from DNA, controlling gene expression, and producing proteins.**
- ***Proteins*** **are composed of sequences of amino acid residues. Humans have 20 kinds of amino acids.**