CS681: Advanced Topics in Computational Biology

Week 3, Lectures 2-3

Can Alkan
EA224
calkan@cs.bilkent.edu.tr

http://www.cs.bilkent.edu.tr/~calkan/teaching/cs681/

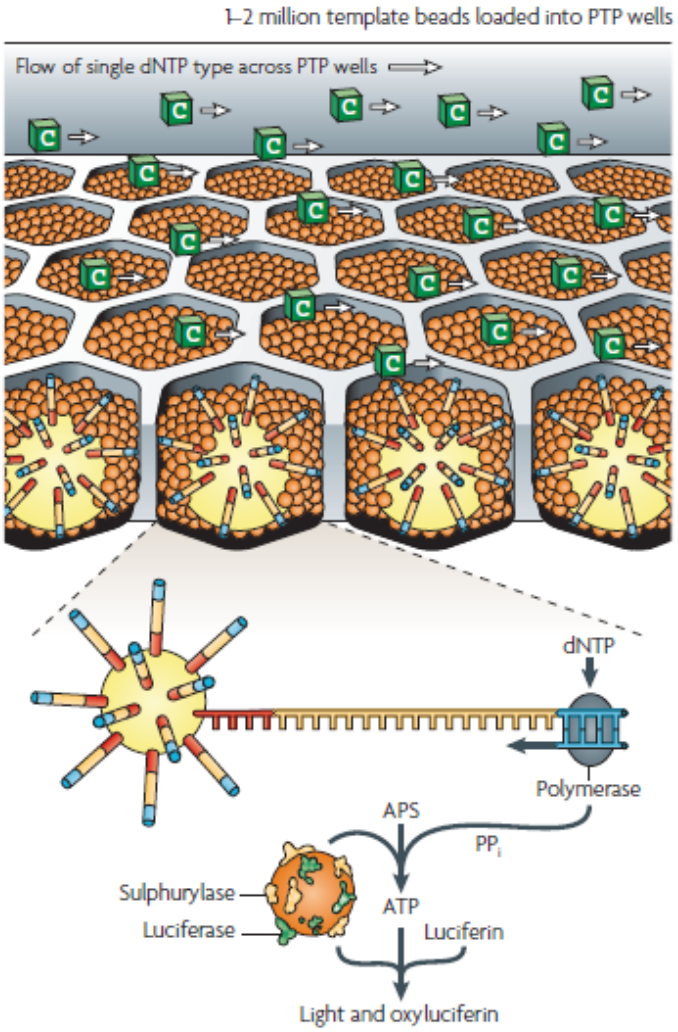
Features of NGS data

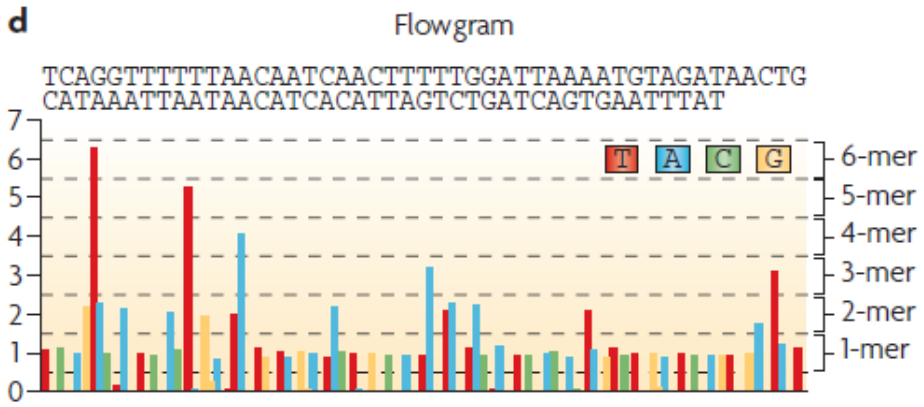
- Short sequence reads
  - ~500 bp: 454 (Roche)
  - 35 – 150 bp Solexa(Illumina), SOLiD(AB)
- Huge amount of sequence per run
  - Gigabases per run (> 600 Gbp for Illumina/HiSeq2000)
- Huge number of reads per run
  - Up to billions
- Bias against high and low GC content (most platforms)
  - GC% = (G + C) / (G + C + A + T)
- Higher error (compared with Sanger)
  - Different error profiles

454 Life Sciences (Roche)

- First “next-gen” sequencing technology
  - Genome of James Watson
- Based on *pyrosequencing*
- Current read length ~700bp
  - Matepair sequencing possible, but difficult
- Error ~1%
  - Indel errors dominate
  - Homopolymers (AAAAAA…, CCCCCC…, etc.)
- Cost: $7/Mb
- One each: Istanbul University, Sabanci University, Ankara University

454 Life Sciences (Roche)





454 Life Sciences (Roche)

Read:

>FL09RMR01D13PQ

TCAGGTTTATACACATGGCGATTTAAATATTTCCATATTTATAGAGATAGCCGTGTAGATATGTCCATGTT
CATGCAGATGGCGGTGAAGATATTTCCATGTTTATAGAGATGCGTAGTGAGAGTACTTTCGTACTAGGT
TAGTAGGAGAGGTATTCGGGTGTAGATAATTCCAGTGTTTATAGAGATGGCGATGTAGTATTTCCATGTT
ANAGAGATAGGTGTTGTANATATTCCAGTGTTATANAGATAGGTGGTGTAGTATTCCTATGTTTAGTAAC
CGAAGAAGTAGTAGGTTAGGTAGTAGTATATATAGTATAGTAGTAGTAGTAGTAGTAGTATATATAGTTAG
TAGTAGTAGTAGTAGTAGTAGTAGTATATAGTTAGTTAGTAGTAGTAGTAGTAGTAGTAGTAGTAGTA
TATAGTTAGTTAGTAGTAGTAGTAGTAGTAGTATATATATATAGTAGTAGTAGTAGTAGTAGTACGTTA
GTTANTAGTAGTANAG

Quality:

>FL09RMR01D13PQ

37 37 37 37 37 37 37 37 37 37 37 39 38 38 38 35 35 28 18 16 16 15 15 19 19 23 23 23 27 27 27 30 35
37 37 37 39 39 39 39 40 40 40 40 40 40 40 40 40 40 40 40 37 37 37 37 37 37 37 37
37 37 37 37 32 32 30 28 28 30 30 32 33 32 32 32 33 33 33 33 33 28 20 20 20 32 33 32 32 20 20 20 33
35 30 25 25 25 28 28 32 29 32 32 31 28 23 19 19

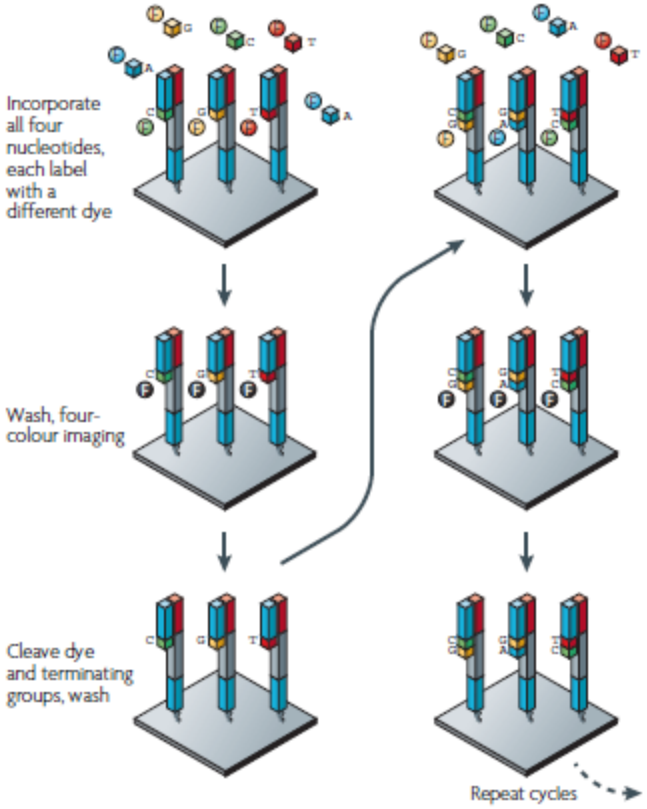
454 Life Sciences (Roche)

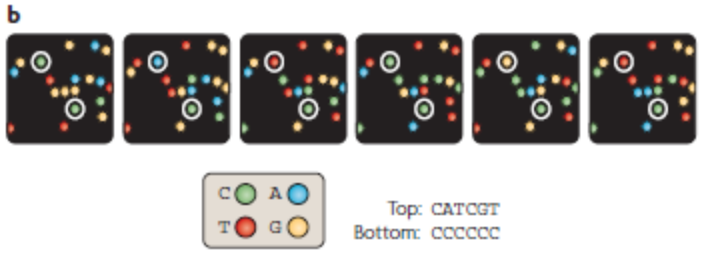
- Basecalling
  - PyroBayes
- Read mapping:
  - MegaBLAST, PASH, Newbler (company's own), SHRiMP, BFAST
- *De novo* assembly:
  - Newbler, Celera Assembler, EULER, Phusion

Illumina (Solexa)

- Current market leader
- Based on *sequencing by synthesis*
- Current read length 100-150bp
- Paired-end easy, longer matepairs harder
- Error ~0.1%
  - Mismatch errors dominate
- Throughput: >600 Gbp in one run (10 days)
- Cheapest sequencing technology
  - Cost: < 4 cents / Mb
- One: TUBITAK MAM
  - *Maybe one more: Ministry of Health*

Illumina (Solexa)





GA IIx

MiSeq

HiSeq 2000

Illumina (Solexa)

•Read length and quality string length are the same

Read and Quality (1)

@FC81ET1ABXX:3:1101:1215:2154/1

TTTTTCAAATGTTTGTTGCCTATTTTTATATCTTCTTTTGAGAATTGTCTGTTCATGTCNTNNGNNCNCNNTNTCANGGGATTGTTTGTT

+

HHGHHHHHGHHHHDHFHHHHHHHFHHHHHHEHHEHHHHEGGDEF2CGDCDFB0>DA####################################

Read and Quality (2)

@FC81ET1ABXX:3:1101:1215:2154/2

AAGCCANNTNNNNNNNNNNNNACTGGATCCTCATAGCTCACCTTATGCAAAAATCAACTCAAGATGGATGAAGGTCTTAAACCTAATAC

+

HHHBH?##;##############:83<9:;7FDFBFEFE;BEEBE8C>2D8@BBACDFG=E@=CDDHEGGDB;<,:19*23?=@#######

- Read length and quality string length are the same
- All read/1s are the same length in the same run
- All read/2s are the same length in the same run

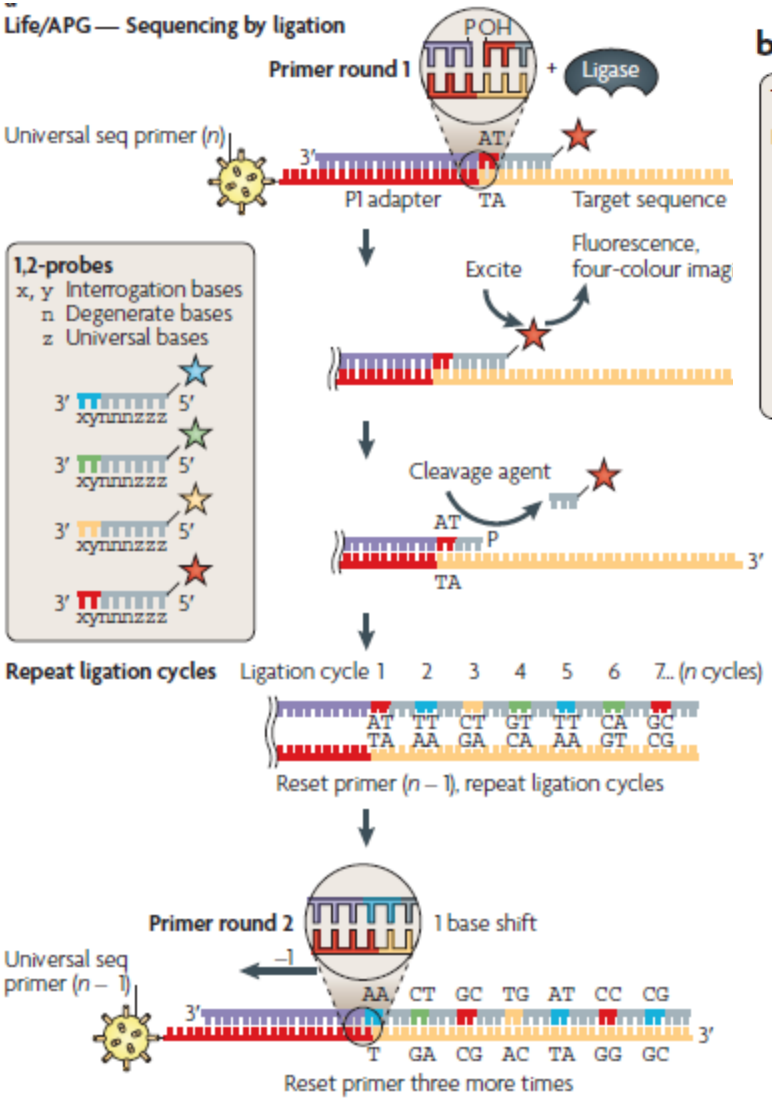
Illumina (Solexa)

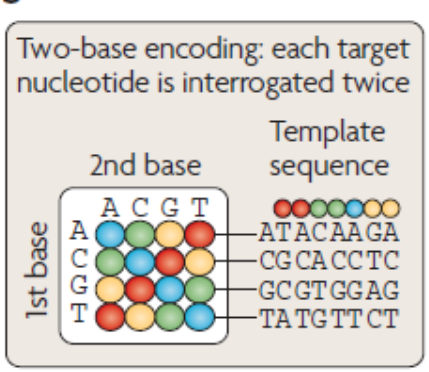
- Read mapping:
  - mrFAST, mrsFAST, BWA, MAQ, BFAST, MOSAIK, Bowtie, SOAP, SHRiMP, many more
- *De novo* assembly:
  - EULER, Velvet, ABySS, Hapsembler, SGA, ALLPATHS, ….

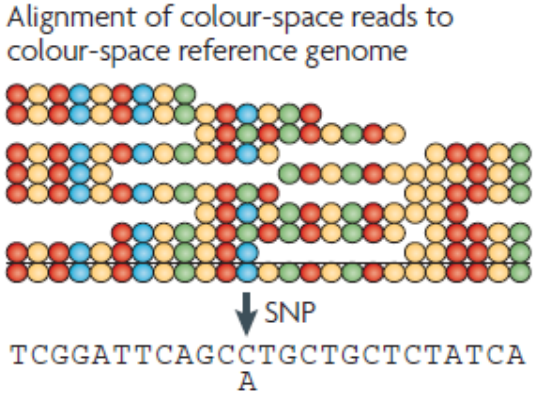
AB SOLiD

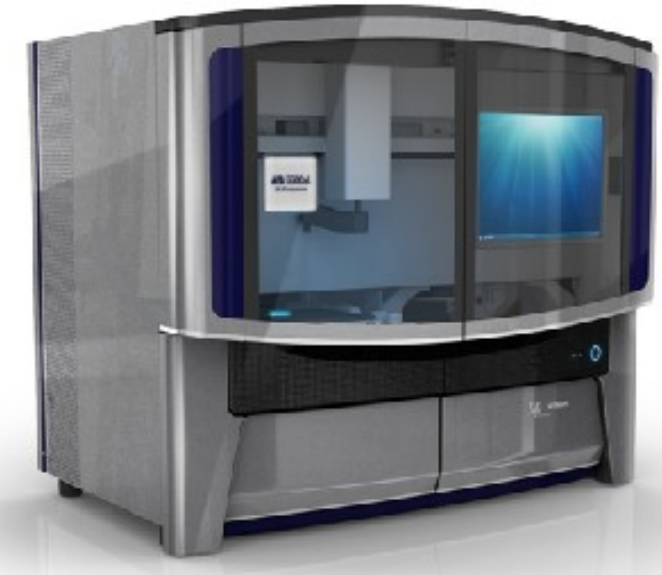
- Reads in “color-space” and not A/C/G/T
- Based on *sequencing by ligation*
- Read length ~75bp
- Paired end easy, longer matepair harder
- Error ~0.1%
- Cost
  - ~7 cents / Mb

AB SOLiD


Life/APG — Sequencing by ligation
Primer round 1
POH
Ligase
Universal seq primer (n)
3′
AT
P1 adapter
TA
Target sequence
1,2-probes
x, y Interrogation bases
n Degenerate bases
z Universal bases
3′ xynnnnzzz 5′
Excite
Fluorescence, four-colour imag
Cleavage agent
AT
P
TA
3′
Repeat ligation cycles
Ligation cycle 1 2 3 4 5 6 7... (n cycles)
AT TT CT GT TT CA GC
TA AA GA CA AA GT CG
Reset primer (n – 1), repeat ligation cycles
Primer round 2
1 base shift
Universal seq primer (n – 1)
–1
AA CT GC TG AT CC CG
T GA CG AC TA GG GC
Reset primer three more times

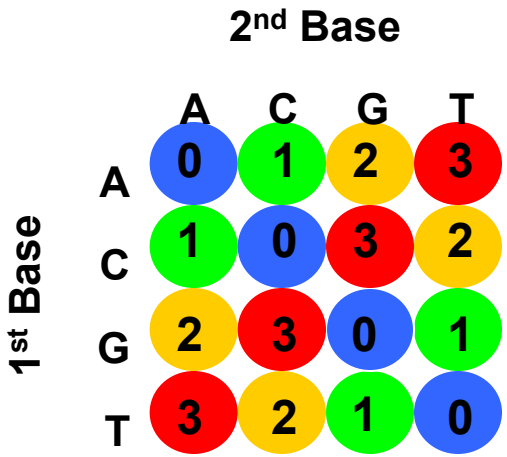

b
Two-base encoding: each target nucleotide is interrogated twice
2nd base
Template sequence
A C G T
1st base
A
C
G
T
ATACAAGA
CGCACCTC
GCGTGGAG
TATGTTCT

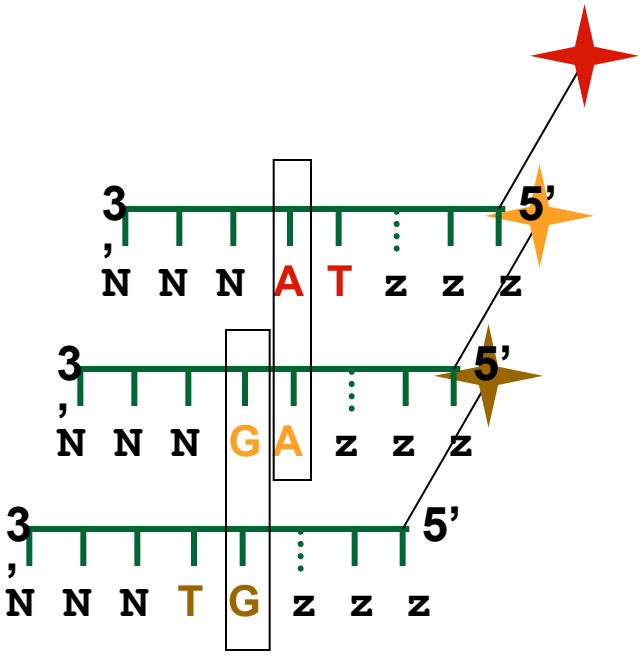

Alignment of colour-space reads to colour-space reference genome
SNP
TCGGATTCAGCCTGCTGCTCTATCA
A



AB SOLiD System dibase sequencing

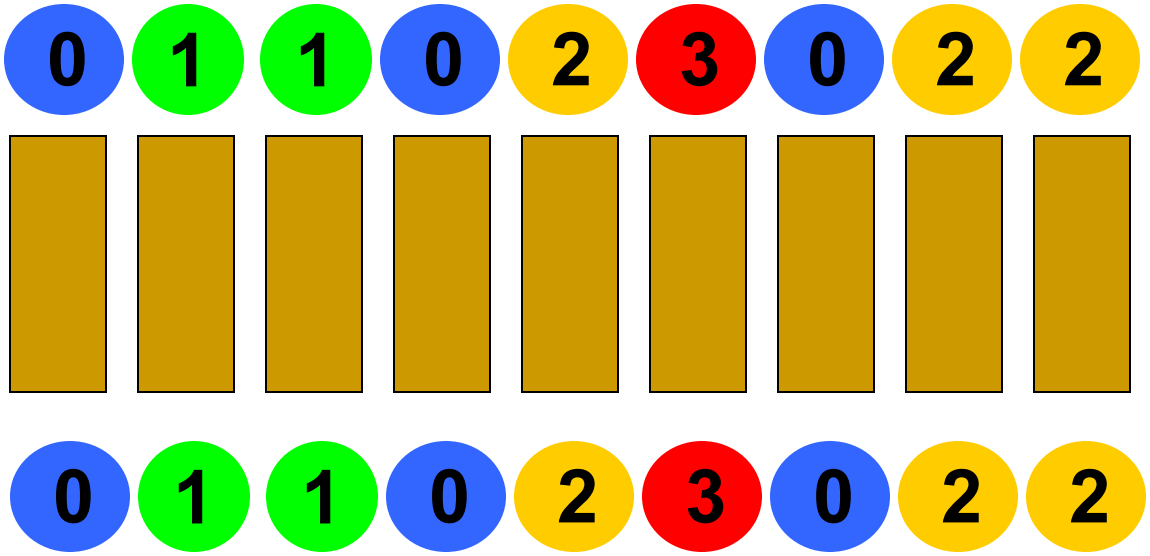
2-base, 4-color: 16 probe combinations

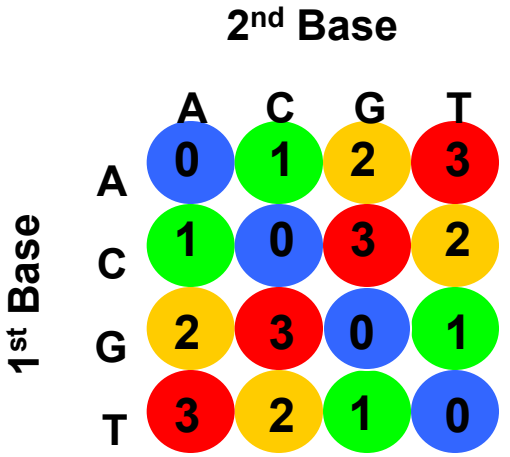


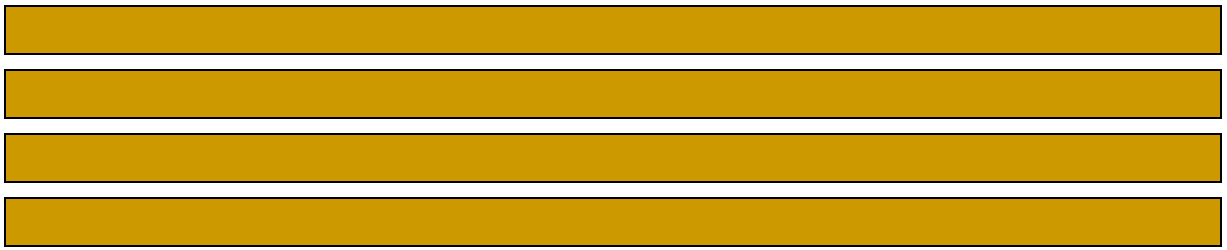


- 4 dyes to encode 16 2-base combinations
- Detect a single color indicates 4 combinations & eliminates 12
- Each color reflects position, not the base call
- Each base is interrogated by two probes
- Dual interrogation eases discrimination
  - errors (random or systematic) vs. SNPs (true polymorphisms)

Converting colors into letters



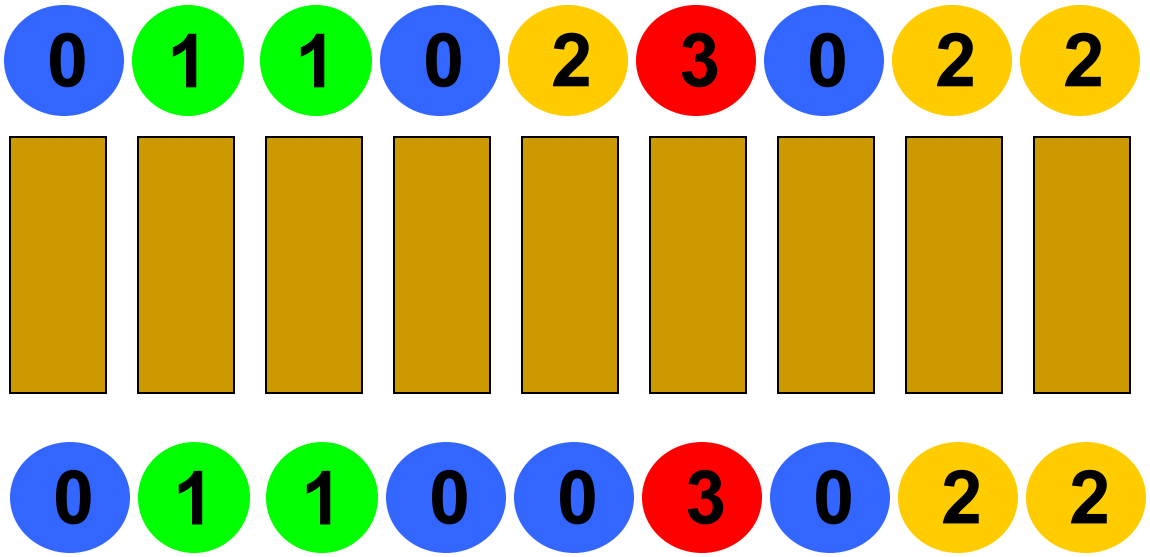


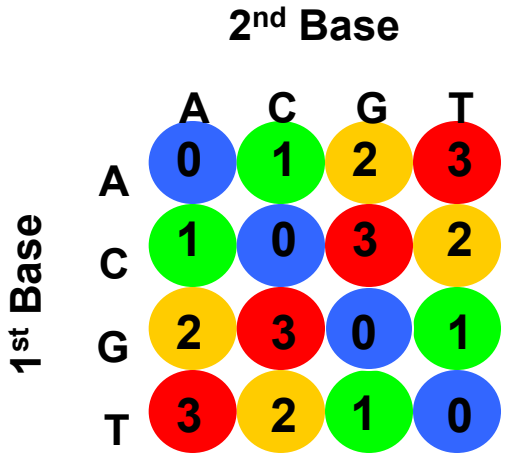


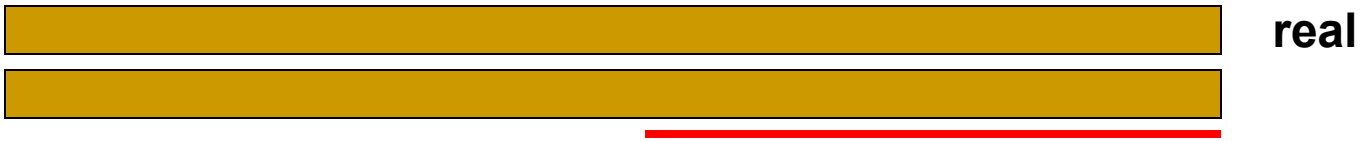
4 Possible Sequences

The decoding matrix allows a sequence of transitions to be converted to a base sequence, as long as one of two bases is known.

But….



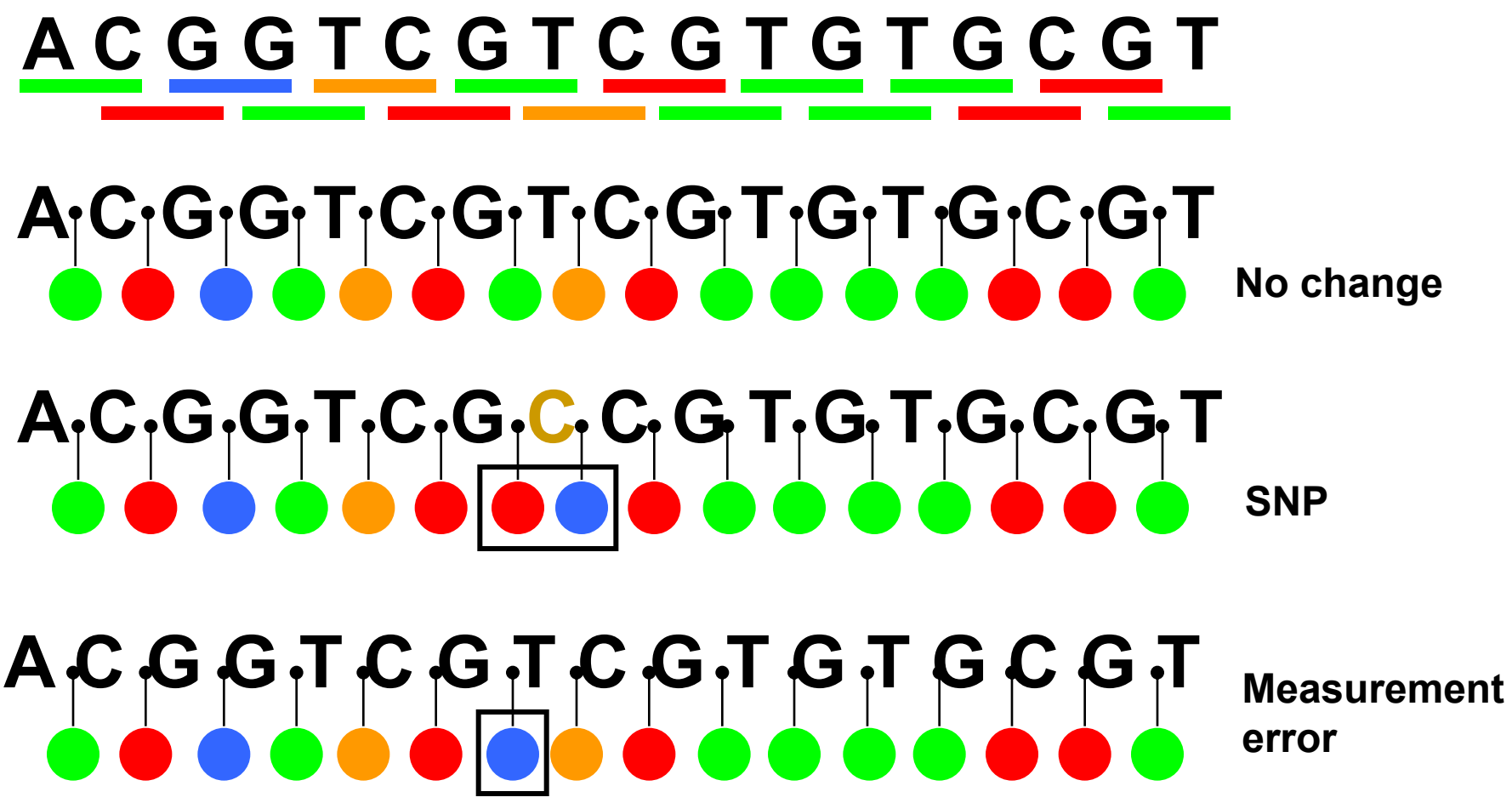




Conversion failure

The decoding matrix allows a sequence of transitions to be converted to a base sequence, as long as one of two bases is known.

SOLiD error checking code



AB SOLiD

Read:

>2_60_1020_F3
T113120222211222002211210221223001220203020032100 33

Quality

>2_60_1020_F3
4 33 29 26 4 27 25 28 29 28 13 22 30 9 27 5 32 4 13 26 16 14 29 5 26 7 4 9 19 14 14 30 16 5 11 7 17 30 8 7 17 20 26 5 26 28 22 4 8 25

- Read length and quality string length are the same
- All read/F3s are the same length in the same run
- All read/R3s are the same length in the same run

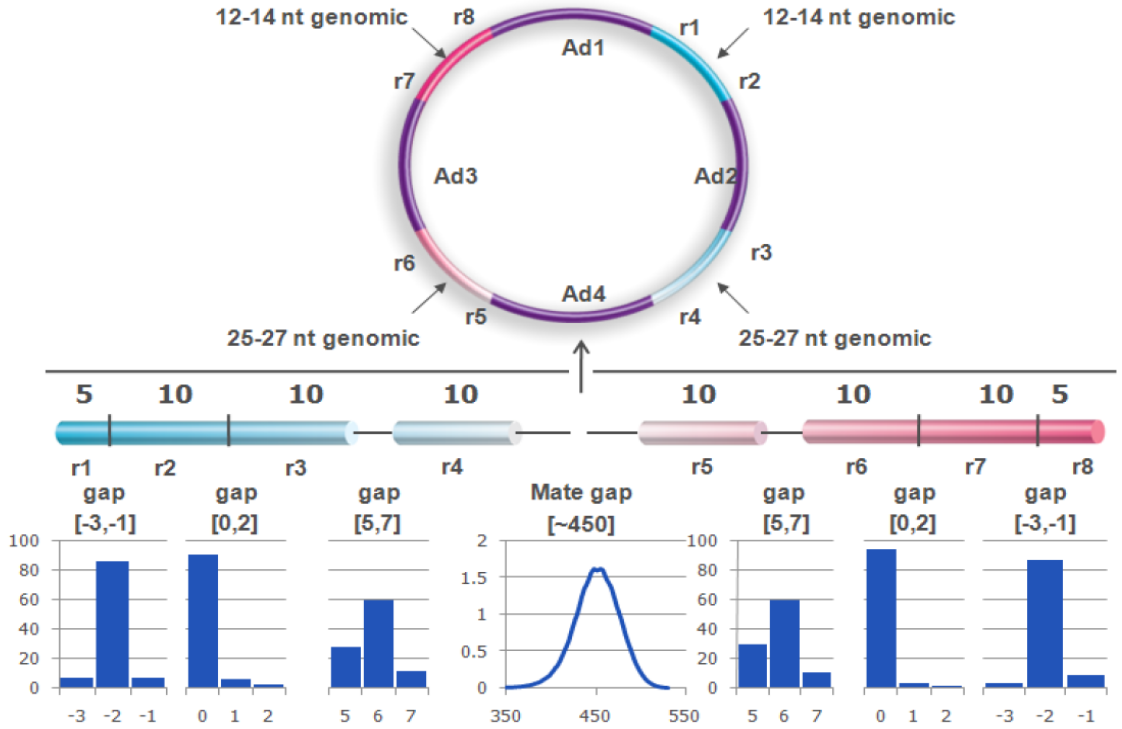
AB SOLiD

- Read mapping:
  - drFAST, BFAST, SHRiMP, BWA, Bowtie
- *De novo* assembly
  - ABySS, SHORTY, Velvet

Complete Genomics

- Provides service only – no instruments are sold
- Technology is propriety; no one really *knows* how it works
  - Based on ligation
- Generates very high coverage sequence per run (>80X)
- Very short reads (35bp paired-end)

Complete Genomics



TGNCNCCCCAATGAGTAACACAGTATTCAGAATGNTCCATAGCGTGCTACTCAGCAGTGCATTGGGGGAN

Read/1:

TGNCN

CCCCAATGAG

xxTAACACAGTA

xxxxxxxTTCAGAATGN

Read/2

TCCATAGCGT

xxxxxxxGCTACTCAGC

xxAGTGCATTGG

GGGAN

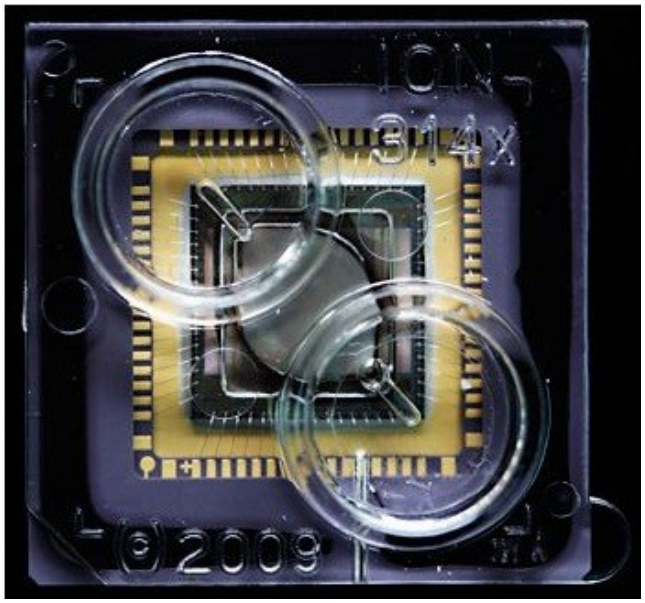
Complete Genomics

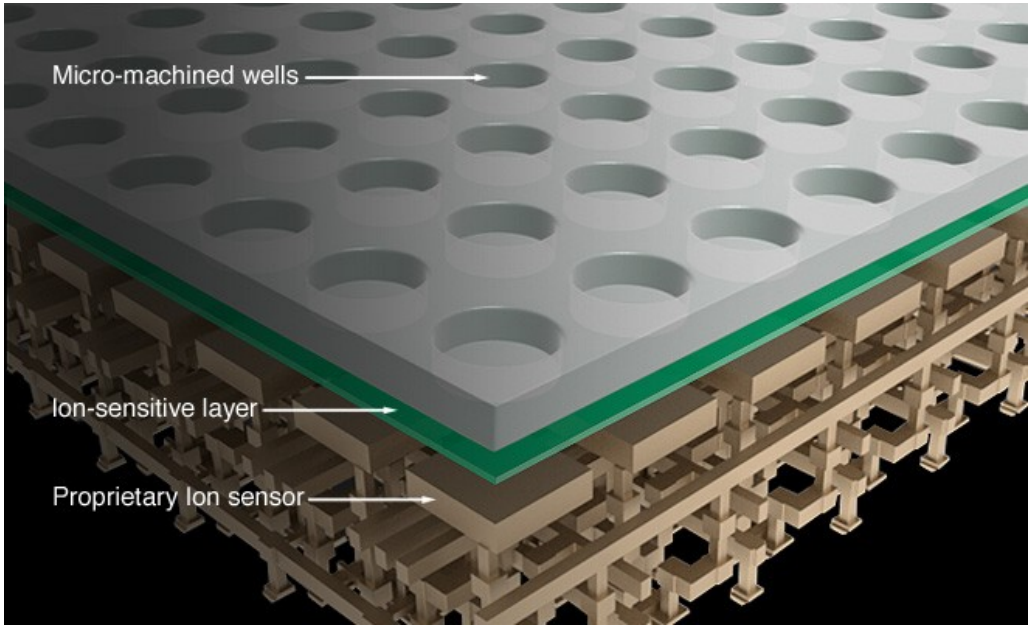
- All analysis tools are developed & used only inside the company; no rival algorithms
  - Public data available
  - Many research opportunities exist

Ion Torrent

- Newer technology, similar to pyrosequencing
- No laser, no image processing:
  - Sequencing is done on a microprocessor that measures pH level changes as bases incorporate
- Error ~1%
  - Indel dominated & homopolymers (454 Life Sci.)
- 95 cents / Mb
- Matepair sequencing possible, but difficult
- One: Istanbul University
- Analysis tools: same as 454 Life Sci.

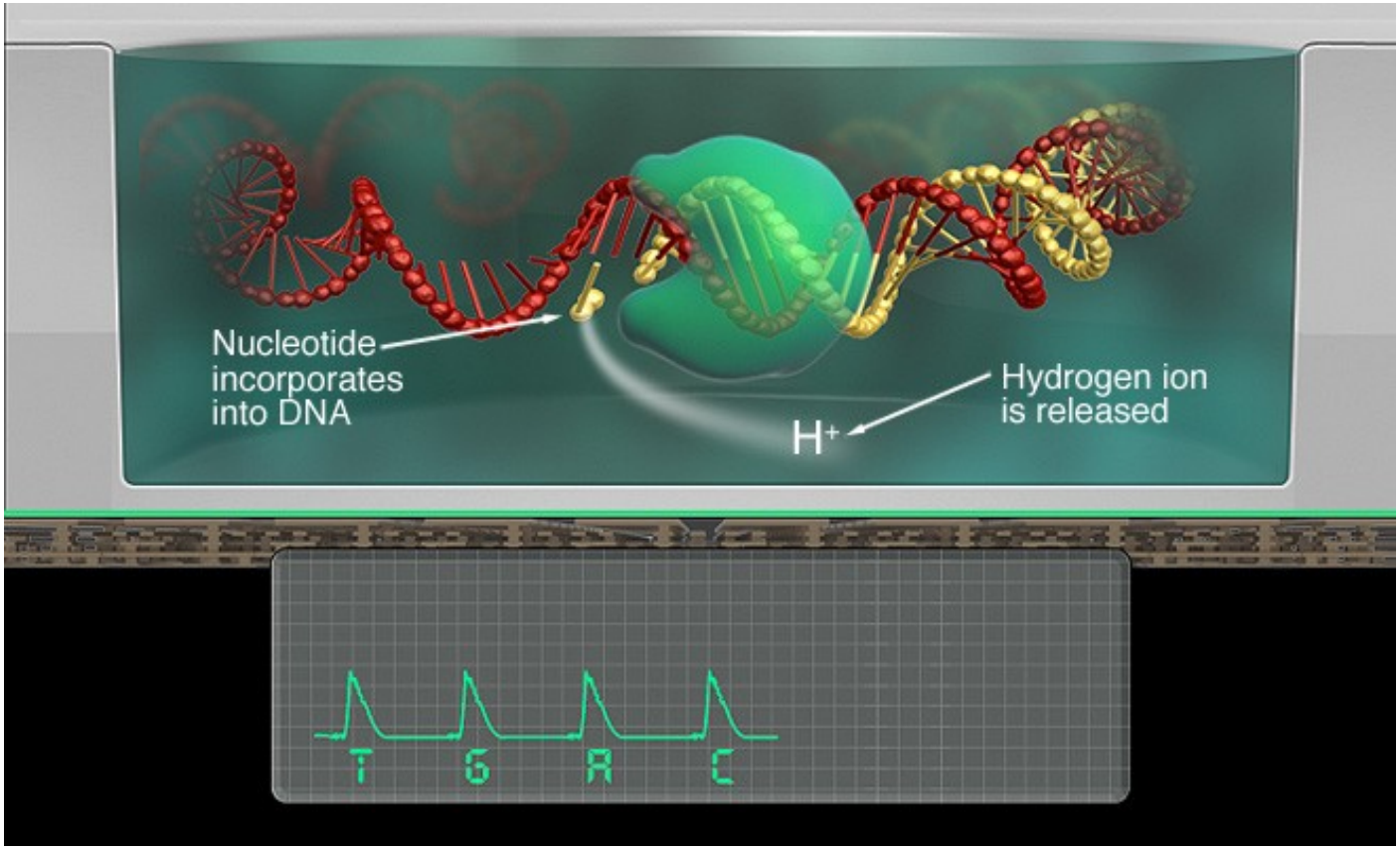
Ion Torrent


ION
314
2009

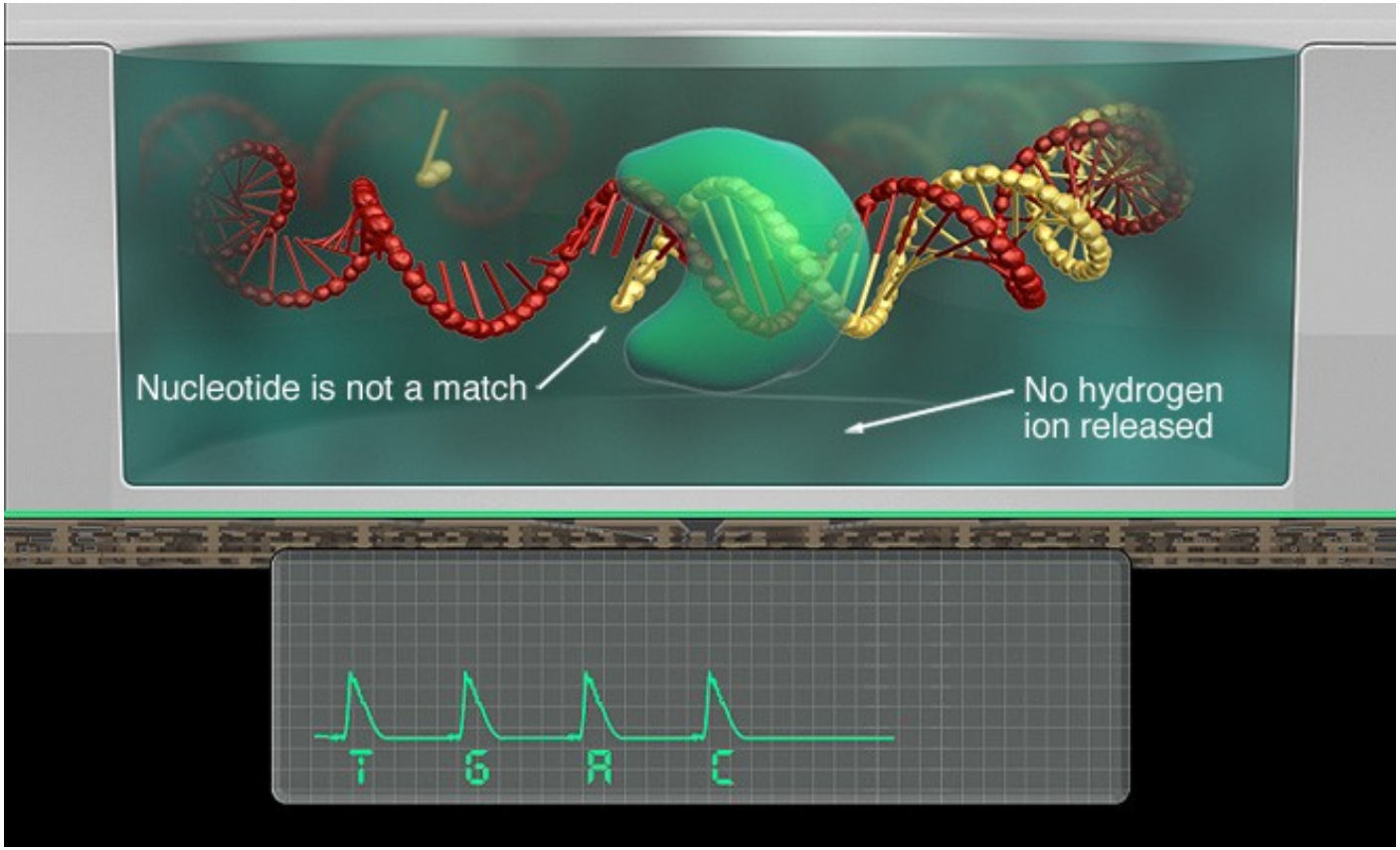

Micro-machined wells
Ion-sensitive layer
Proprietary Ion sensor


ion torrent
Proton

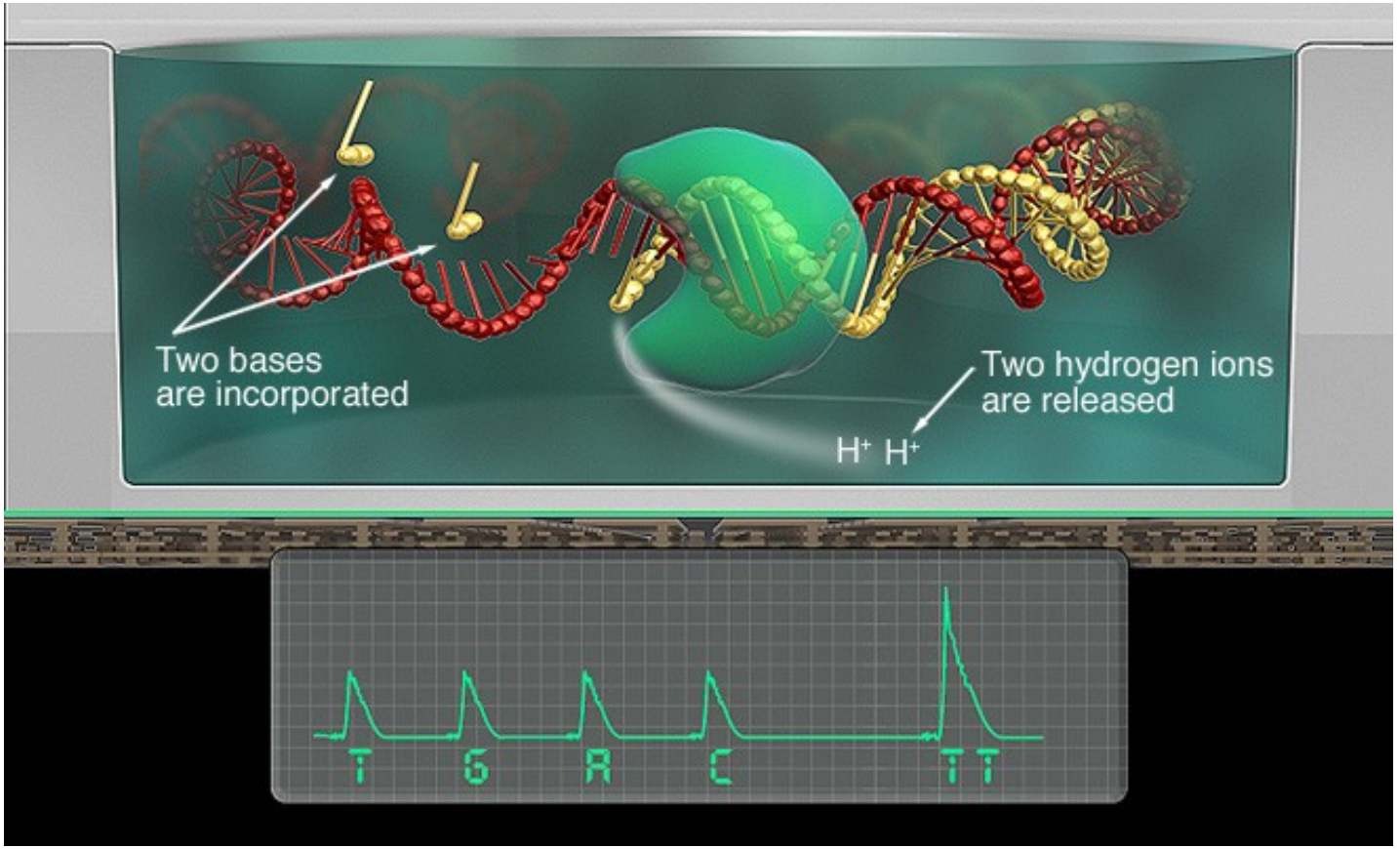
Ion Torrent


Nucleotide incorporates into DNA
H+
Hydrogen ion is released
T
G
A
C

Ion Torrent

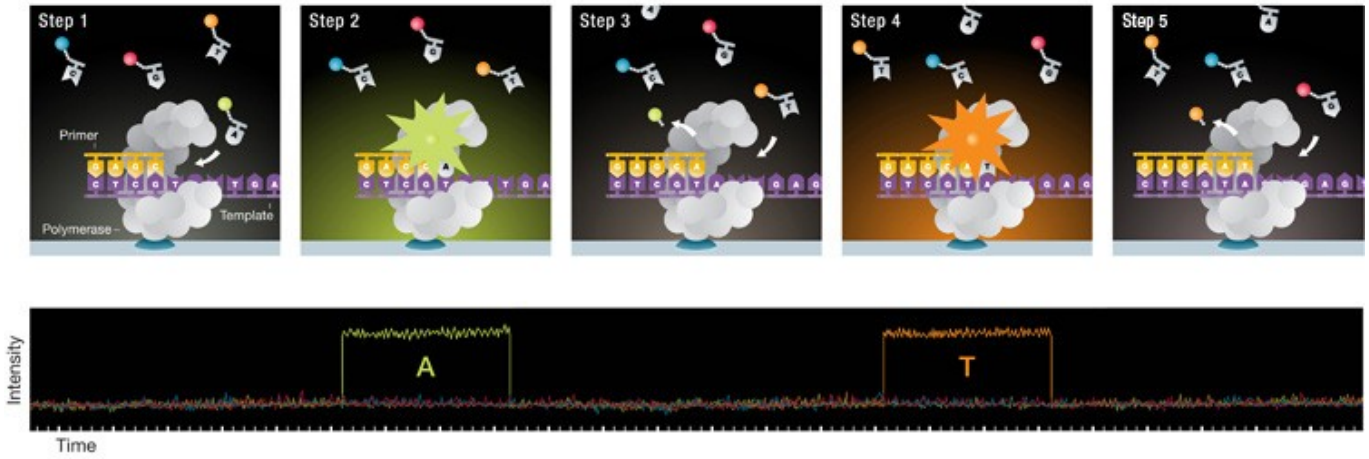


Ion Torrent


Two bases
are incorporated
Two hydrogen ions
are released
H+ H+
T G A C T T

Pacific Biosciences

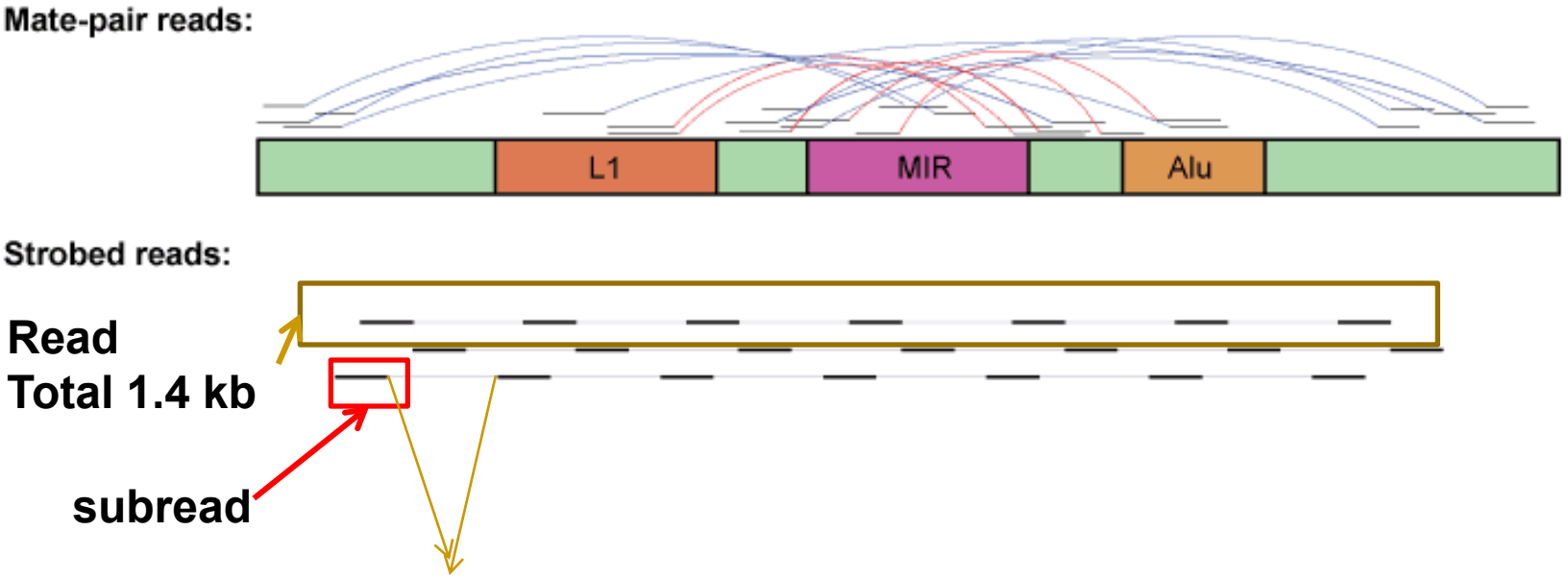
- “Third generation”; single molecule real time sequencing (SMRT)
- No replication with PCR
- Phosphates are labeled. Watches DNA polymerase in real-time while it copies single DNA molecules.
- Premise: long sequence reads in short time (median 1.4 kbp)
- Errors: ~15%; indel dominated
- $11 / Mb



Pacific Biosciences

- For any DNA polymerase you can read a total of ~1.4 kb (median) sequence
  - ~5% can generate > 3kb
- Three sequencing protocols:
  - Single: read one contiguous sequence
  - Circular consensus: Make a circle, re-read the same molecule 6-7 times
    - Multiple sequence alignment to correct errors
    - Median length = 1400 / 7 = 200 bp
  - Strobe sequencing

Pacific Biosciences: strobe sequencing



Distances between
subreads are approximately known

Upcoming

- Nanopore sequencing:
  - Oxford Biosciences
    - Protein based nanopore
    - 5.4 kb reads now; 100 kb reads “soon”
  - IBM
    - Silicon based nanopore
- Electron microscope:
  - Halcyon
- Many more..

NGS: Computational Challenges

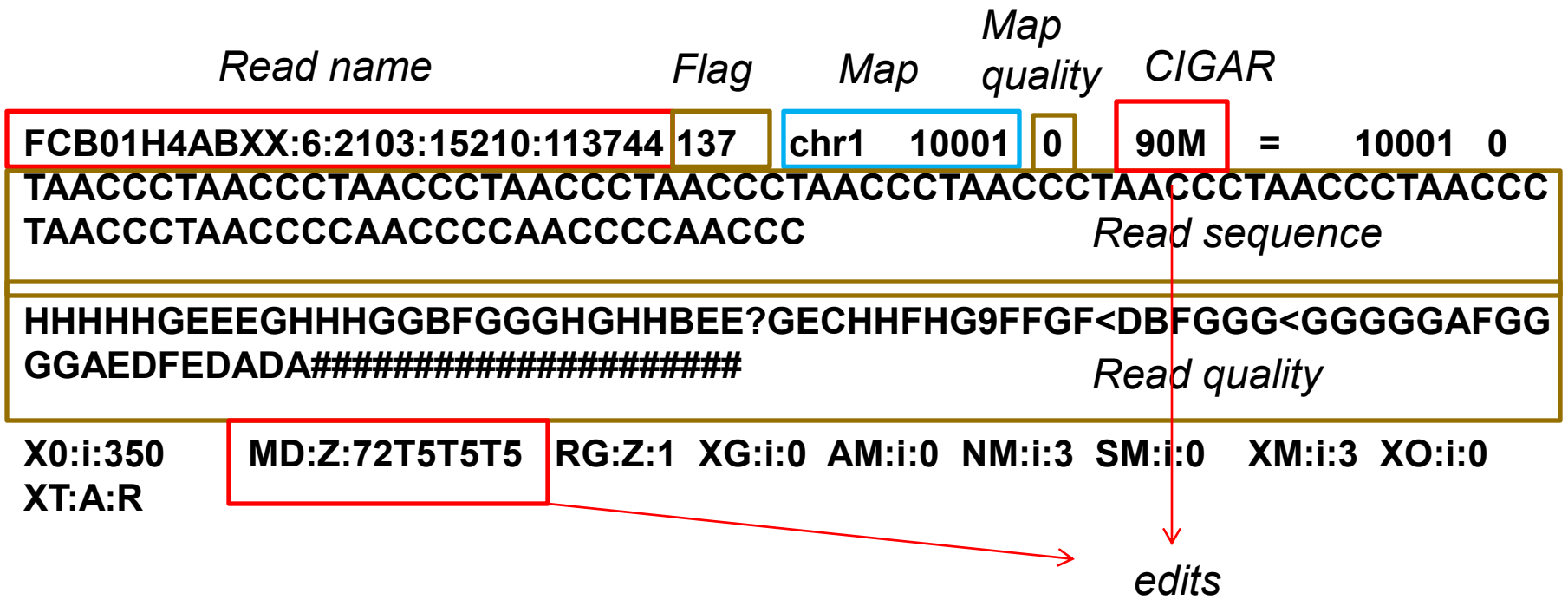
- Data management
  - Files are very large; compression algorithms needed
- Read mapping
  - Finding the location on the reference genome
  - All platforms have different data types and error models
  - Repeats!!!!
- Variation discovery
  - Depends on mapping
  - Again, all platforms has strengths and weaknesses
- *De novo* assembly
  - It's very difficult to assemble short sequences with high errors

Compression

- 1 – Reference based
  - Coding/decoding rather than real compression
  - Very high compression rate
  - Fast to encode
  - Slow to decode
  - Needs a reference genome
    - None, or poor quality for most species
    - Use same version of reference genome in decompression
  - Needs mapping (takes a long time)
    - Unmapped reads should be treated separately
  - CRAMtools, SlimGene, etc.
    - *Very* lossy

CRAMtools

Post mapping; SAM format:



- Read name is unnecessary
- Flag tells you whether /1 or /2
- Map location and edit fields (CIGAR & MD) can be used to regenerate reads
- Don't store quality if edit distance = 0; otherwise only keep the qualities of changed bases

Fritz et al. Genome Research, 2011

CRAMtools

Post mapping; SAM format:

Read name *Flag* *Map* *Map quality* *CIGAR*

FCB01H4ABXX:6:2103:15210:113744 137 chr1 10001 0 90M = 10001 0
TAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCC
TAACCCTAACCCCAACCCCAACCCCAACCC *Read sequence*

HHHHHGEEEGHHHGGBFGGGHGHHBEE?GECHHFHG9FFGF<DBFGGG<GGGGGAFGG
GGAEDFEDADA###################### *Read quality*

X0:i:350 MD:Z:72T5T5T5 RG:Z:1 XG:i:0 AM:i:0 NM:i:3 SM:i:0 XM:i:3 XO:i:0
XT:A:R

Keep: **137** ; **chr1:10001** ; **0** ; **90M**; **72T5T5T5** ; **(#,#,#)**
Add a layer of Huffman encoding

Fritz et al. Genome Research, 2011

CRAMtools: test case

- One human genome
  - 40X coverage
  - 134 GB gzipped = 479 GB raw text
  - Mapped with BWA; >1 day with 200 CPUs
  - SAM format converted to BAM file: 112 GB
  - BAM to CRAM: 7.5 GB
  - Decode CRAM to BAM: 33 GB (lossy!!!)

Compression

- 2 – Reference free
  - Less compression rate
  - No need for reference, applicable to any dataset from any species
  - Slower to compress, faster to decompress
  - Can be lossy or lossless
  - Multipurpose compressors:
    - gzip, bzip2, 7-zip, etc.
  - Specialized FASTQ compressors
    - SCALCE, ReCoil, G-SQZ, etc.

Reference-free compression

- Easy task (or gzip, etc.): Concatenate all sequences, then run Lempel-Ziv algorithm
- Problem: Locality

Lempel-Ziv Compression

a b b a a b b a a b a b b a a a a b a a b b a
0 1 1 0 2--- 4--- 2--- 6------- 5--- 5--- 7------- 3--- 0

Index	Entry	Index	Entry
0	a	7	baa
1	b	8	aba
2	ab	9	abba
3	bb	10	aaa
4	ba	11	aab
5	aa	12	baab
6	abb	13	bba

Reordering improves locality

File Size: 250MB, 5Mil 51bp Bacterial Genome

Pre-processing	Time (s)	Gzip time	Size (MB)	Comp. Factor	Boosting
-	-	70	65	4	-
Mapping	180	21	20	12.5	3.25
Lexo. Sorting	10	30	26	9.61	2.5
Cores*	10	21	21	11.9	3.1

*** Idea behind SCALCE**

Reordering example

Ref: AAAAAATGACGTCTCTCCTCCTTTTTTAAAACCT

Original	Mapping	Sorting	Cores
CTTTTT	AAAAAA	AAAAAA	AAAAAA
GATGAC	TAATGA	ATGACG	TAAAAC
CCCCCT	GATGAC	CCCCCT	CCCCCT
AAAAAA	ATGACG	CTTTTT	CTTTTT
ATGACG	CCCCCT	GATGAC	TAATGA
TAAAAC	CTTTTT	TAAAAC	GATGAC
TAATGA	TAAAAC	TAATGA	ATGACG

Cores: Locally Consistent Parsing

LCP (Sahinalp STOC 1994, Sahinalp FOCS 1996) is a combinatorial pattern matching technique that aims to identify building blocks of strings. For any user-specified integer *c* and with any alphabet, the LCP identies core substrings of length between *c* and *2c* such that:

- any string from the alphabet of length *3c* or more include at least one such core string
- there are no more than three such core strings in any string of length *4c* or less
- if two long substrings of a string are identical, then their core substrings must be identical

Increasing Locality

- Goal: Obtain a few core substrings for each read so that two highly overlapping reads will have common core substrings. We obtain a set of core strings such that
  - A long prefix of a core substring can not be a suffix of another core substring (this assures that two subsequent core substrings can not be too close to each other).
  - Each read includes at least one core substring.

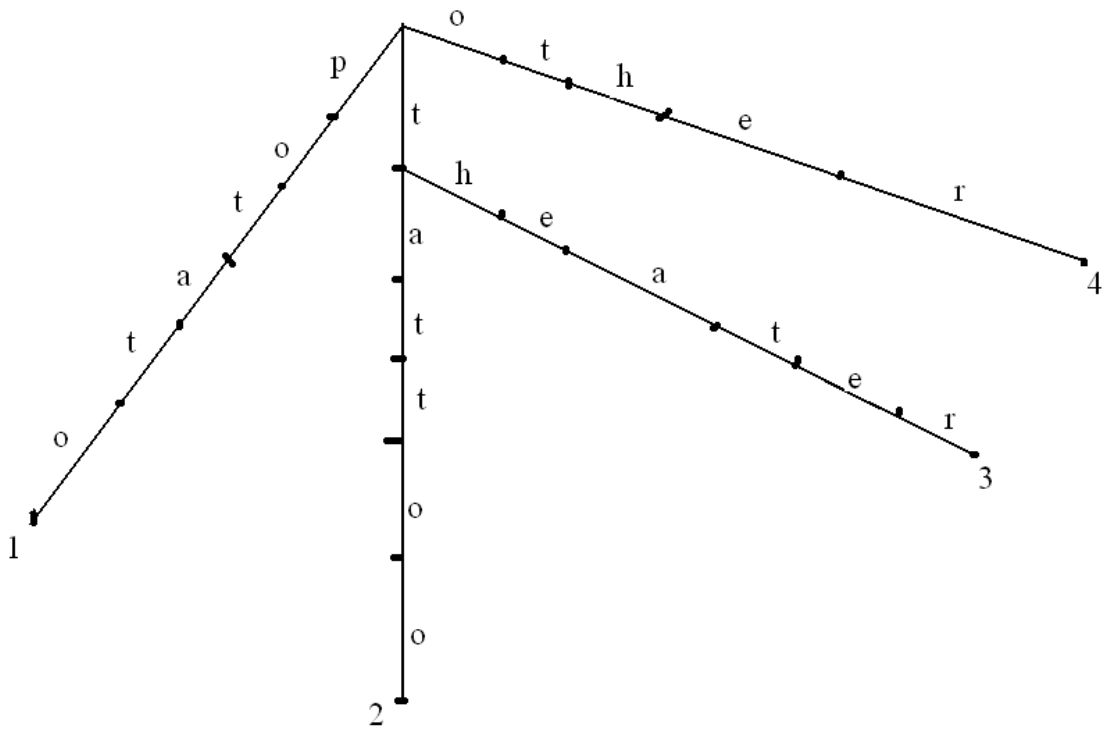
Finding cores

Find all “core substrings" in a given read and place it in a bucket which has the maximum number of reads.

- Trie data structure: finding all core substrings within a read would require O(cr) time (r: read length, c: length of all core substrings in that read).
- Improvement: Aho-Corasick dictionary matching algorithm using an automaton. O(r+k), where k is the number of core substring occurrences in each read.
- More improvement: Alphabet is small, and number of core substrings is fixed; pre-process automaton to calculate bucket in O(1) time, reduce total search time to O(r).

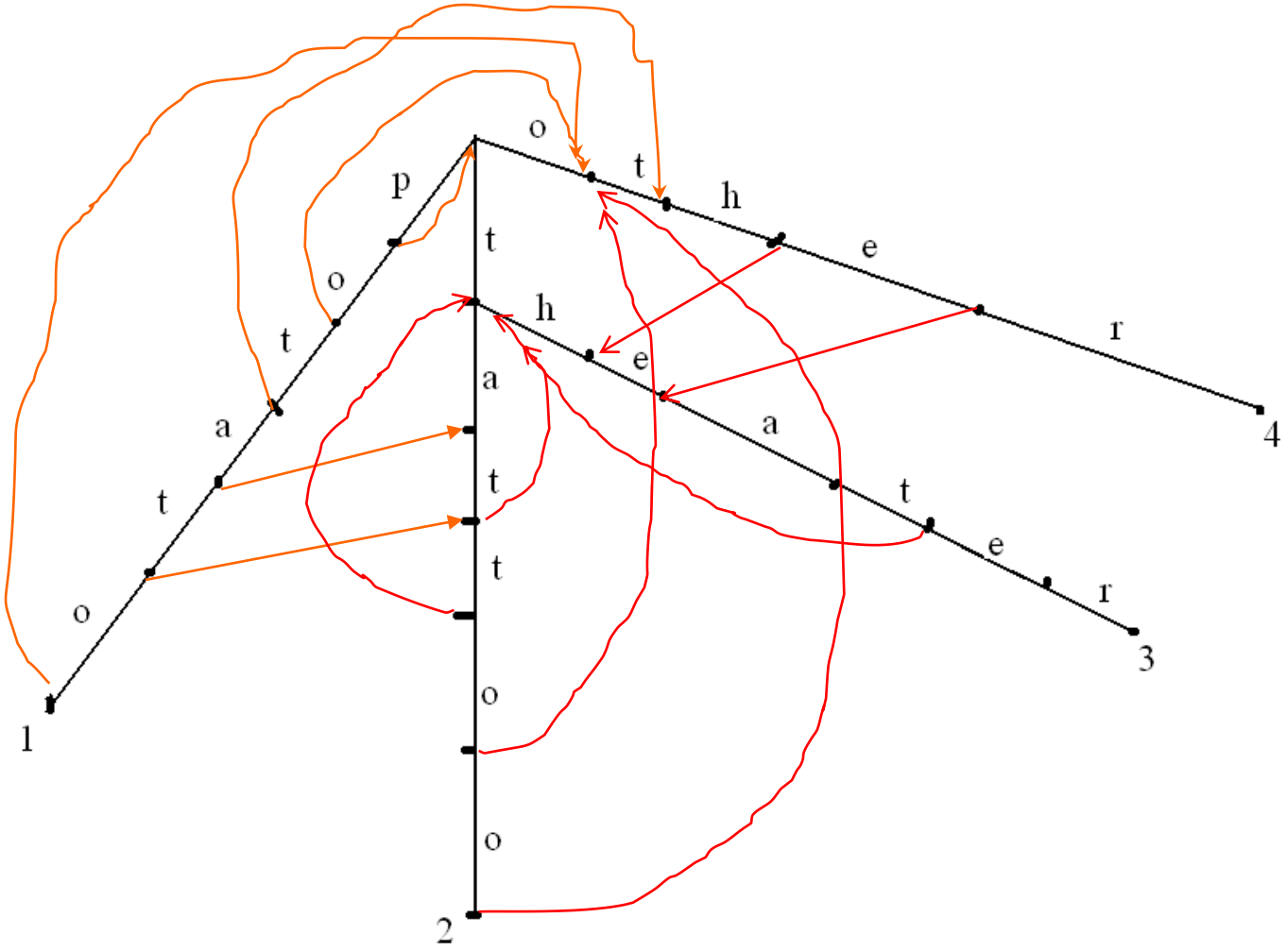
Trie data structure

P={potato, tattoo, theater, other}

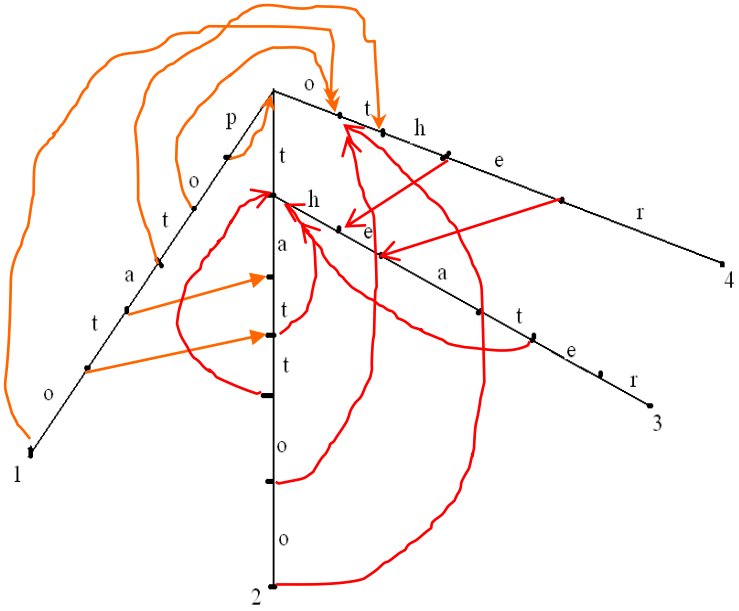


Failure links

P={potato, tattoo, theater, other}



Aho-Corasick Algorithm

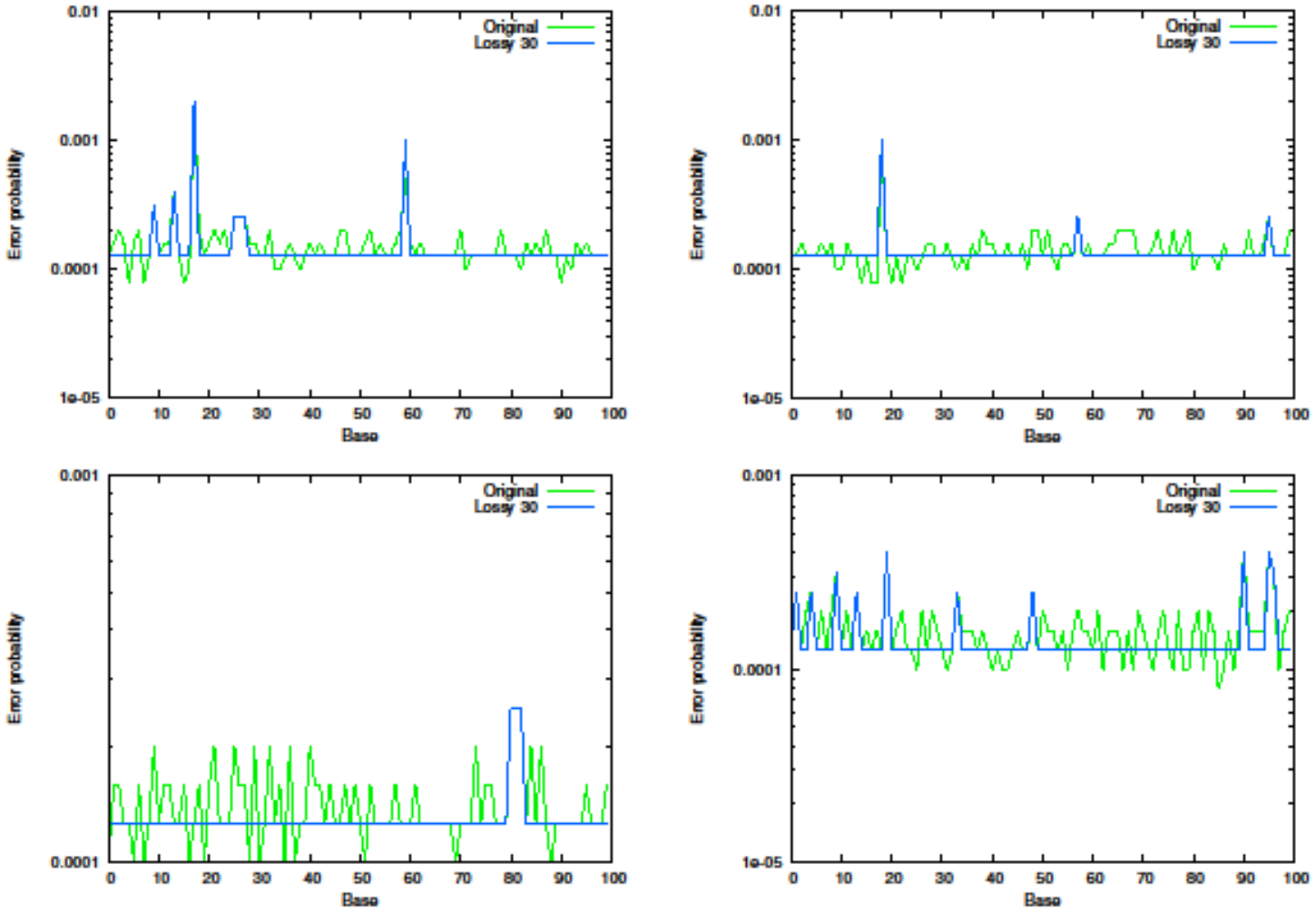


Input: Pattern set P and text T
Output: all occurrences in T any pattern from P
Algorithm AC
l=1;
c=1;
w=root
Repeat
    while there is an edge (w, w') labeled with T(c)
        if w' is numbered by pattern *i* then
            report that p_i occurs in T starting at *l*;
        w=w'; c++;
    w=n_w and *l*=c-lp(w);
Until c>m

(*optional*) Quality Score Transformation

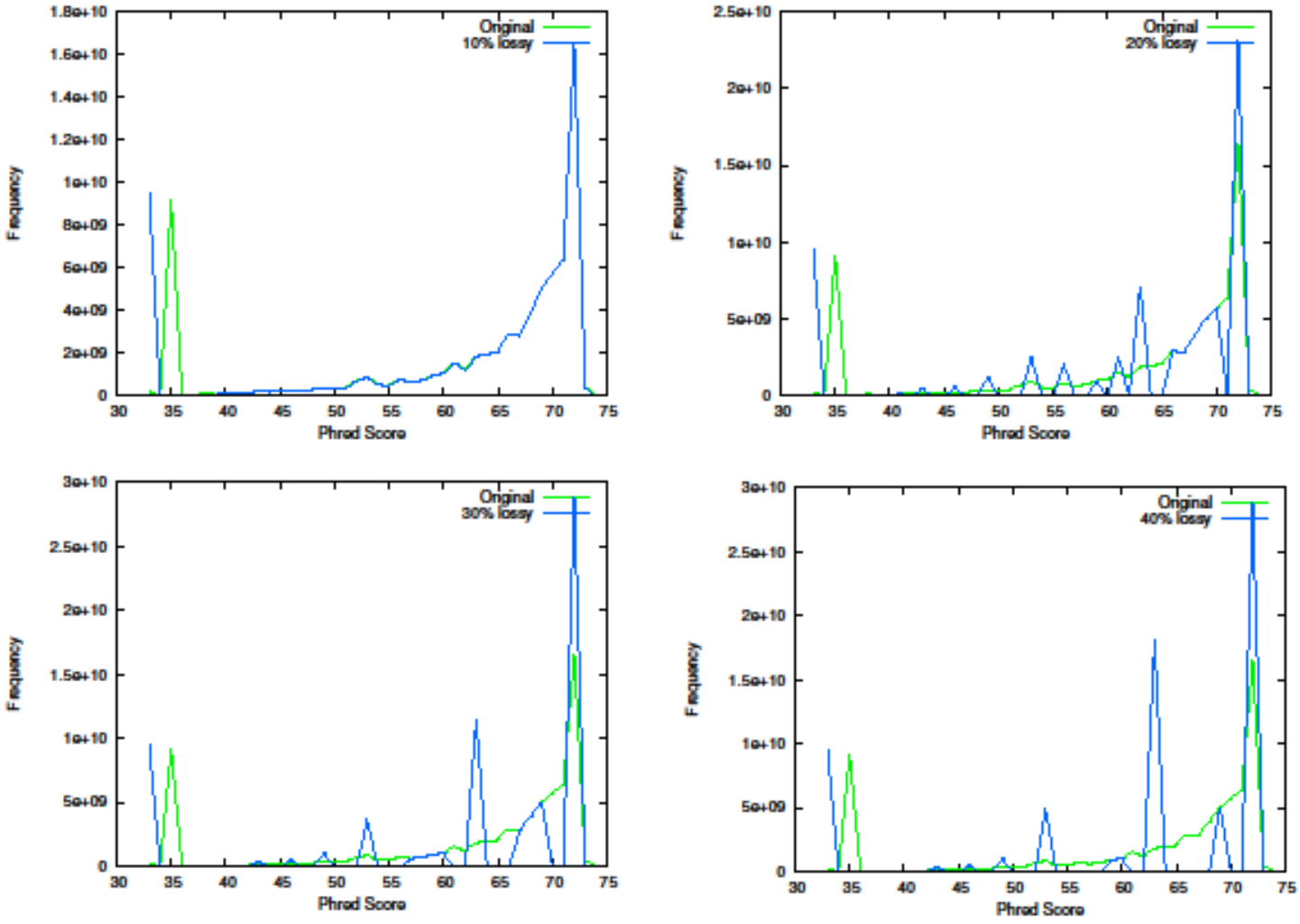
- Sequence alphabet has 5 characters (A,C,G,T,N); but quality string alphabet is larger, thus compresses less
  - Generate qualities with a smaller alphabet to improve compression
- Expect some small noise in a normal run of sequencing machine.
- Calculate the frequency of the alphabet and reduce the noise by merging the local maxima up to *e%* threshold.

(*optional*) Quality Score Transformation



Original and transformed quality scores for four random reads that are chosen from NA18507 individual.

(*optional*) Quality Score Transformation



Frequency plots after applying the greedy method with different error thresholds

Test case

Data Set			gzip		SCALCE +gzip		
Name	# of Reads	Size	Size	Rate	Size	Rate	Boosting factor
P. aeruginosa RNAseq	89M	11,812	3,473	3.40	1,599	7.38	2.17x
P. aeruginosa Genomic	81M	10,744	3,477	3.09	1,730	6.21	2x
NA18507 WGS	1.4B	479,106	133,984	3.58	76,011	6.30	1.76x
NA18507 Single Lane	36M	12,308	3,424	3.59	1,936	6.35	1.77x

	gzip	SCALCE -gzip		
Name	Time	Total	SCALCE Time	gzip Time
P. aeruginosa RNAseq	31m	22m	11m	11m
P. aeruginosa Genomic	30m	22m	11m	11m
NA18507 WGS	15h 25m	13h 19m	6h 32m	6h 47m
NA18507 Single Lane	25m	18m	8m	10m

SCALCE vs. CRAMtools

	# SNP Count	dbSNP v132	Novel	
			Total	in SD+CR
Original Qualities	4,296,152	4,092,923 (95.26%)	203,229	192,114 (94.53%)
Qualities using SCALCE	4,303,140	4,098,875(95.25%)	204,265	192,976 (94.47%)
Lost	7,931	4,596 (57.95%)	3,335	2,963 (88.84%)
New	14,919	10,548 (70.70%)	4,371	3,825 (87.51%)
Qualities using CRAM tools	4,202,298	4,013,401 (95.50%)	188,897	179,875 (95.22%)
Lost	101,957	84,607 (82.98%)	17,350	15,036 (86.66%)
New	8,103	5,085 (62.75%)	3,018	2,797 (92.67%)