

Genomic Dark Matter: The limitations of short read mapping illustrated by the Genome Mappability Score (GMS)

Hayan Lee
Advised by Prof. Michael Schatz

Sep. 28, 2011
Quantitative Biology Seminar

Outline

- Background on genome sequencing
- Challenges for accurately measuring genome variations
- Innovations for variations detection
 - Genome Mappability Score (GMS)
 - Genome Mappability Analyzer (GMA)
 - Applications of GMS
- Contributions and Future Work

Outline

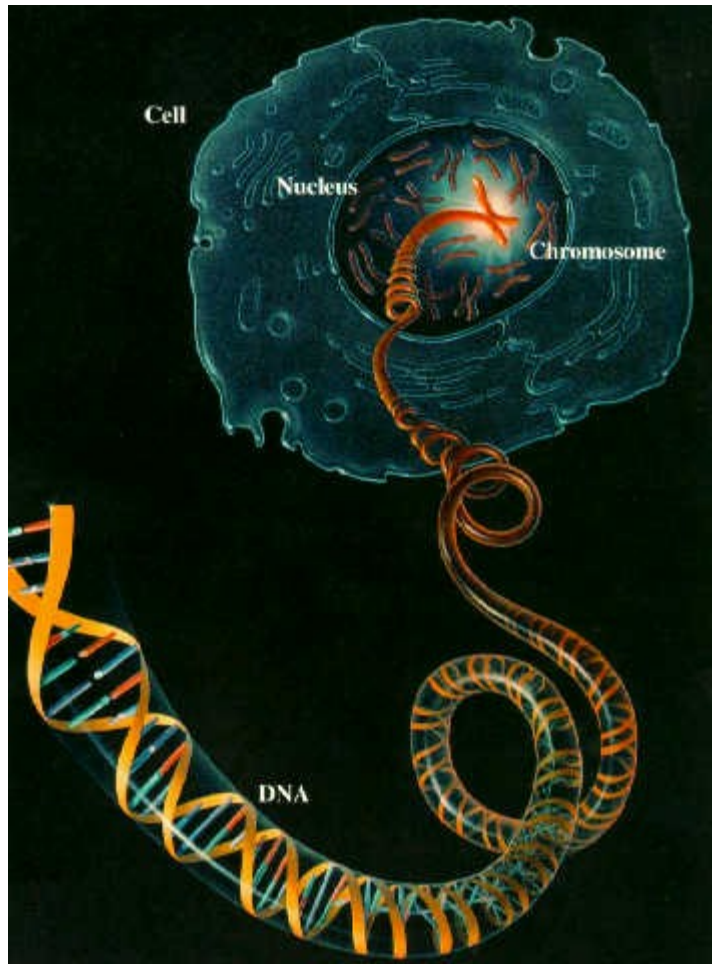
- Background on genome sequencing
 - Genome
 - Sequencing history
 - Applications of short read mapping - Resequencing
 - Alignment tools and algorithms
- Challenges for accurately measuring genome variations
- Innovations for variations detection
- Contributions and Future Work

Genome

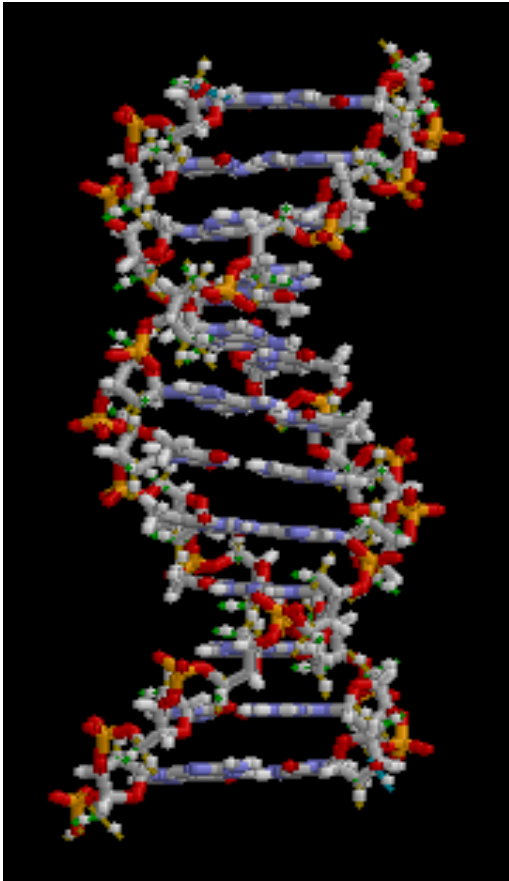
Gene + Chromosome

A container of an organism's hereditary information.

Encoded in long molecules of DNA.



Deoxyribonucleic acid(DNA)



- A cookbook with a lot of recipes
- The structure was discovered by James D. Watson and Francis Crick (Watson J.D. and Crick F.H.C. (1953). "A Structure for Deoxyribose Nucleic Acid" Nature 171 (4356): 737–738)
- Double Helix
- Nucleotides
 - A Adenine
 - C Cytosine
 - G Guanine
 - T Thymine

Back then...

- Sanger et al. sequenced the first complete genome in 1977

Nucleotide sequence of bacteriophage Φ X174 DNA

F. Sanger, G. M. Air*, B. G. Barrell, N. L. Brown†, A. R. Coulson, J. C. Fiddes, C. A. Hutchison III‡, P. M. Slocombe§ & M. Smith*

MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, UK

A DNA sequence for the genome of bacteriophage Φ X174 of approximately 5,375 nucleotides has been determined using the rapid and simple 'plus and minus' method. The sequence identifies many of the features responsible for the production of the proteins of the nine known genes of the organism, including initiation and termination sites for the proteins and RNAs. Two pairs of genes are coded by the same region of DNA using different reading frames.

THE genome of bacteriophage Φ X174 is a single-stranded, circular DNA of approximately 5,400 nucleotides coding for nine known proteins. The order of these genes, as determined by

strand DNA of Φ X has the same sequence certain conditions, will bind ribosomes fragment can be isolated and sequenced. was found. By comparison with the amino initiation of

At this sta with DNA synthesised a part of the ri the intercistron polymerase a tion techniqu labelled DNA produced. This decan



Radioactive Chain Termination
5000bp / week / person

<http://en.wikipedia.org/wiki/File:Sequencing.jpg>
<http://www.answers.com/topic/automated-sequencer>

Today

- Current DNA sequencing machines can sequence **billions of short (25-500bp) reads** from random positions
 - Per-base error rate estimated at 1-2% (Simpson et al, 2009)
 - *De novo* sequencing
 - 5375 b/week (1977) vs. 210 Gb /week (2008)

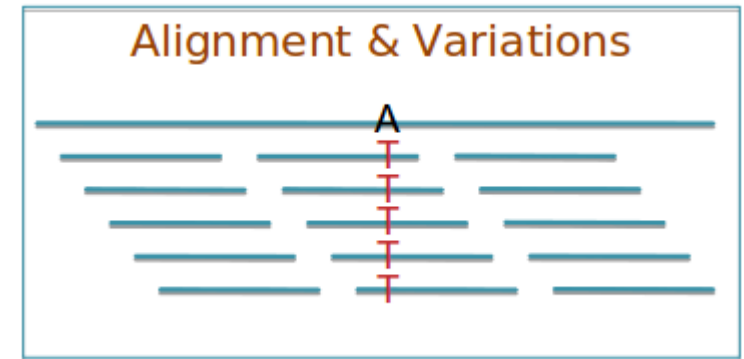


2008

ABI / Life Technologies
SOLiD Sequencing
Current Specs (5500xl):
5B 75bp reads / run
= 30Gbp / day
= 210Gbp / week
= 70 individuals /week

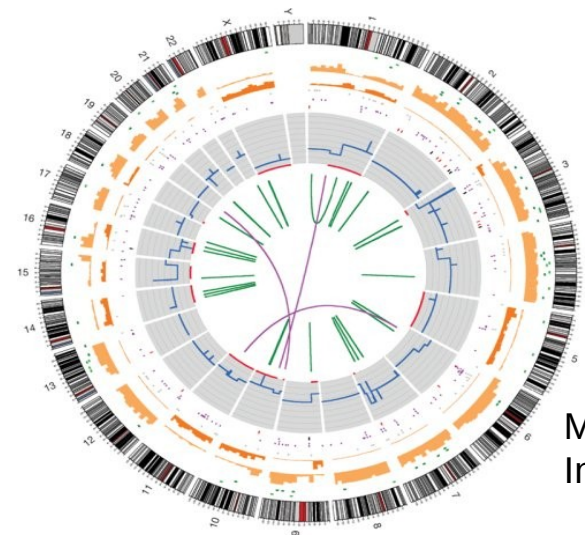
Short read mapping (Resequencing)

- Discovering genome variations
- Investigating the relationship between variations and phenotypes
- Profiling epigenetic activations and inactivations
- Measuring transcription rates

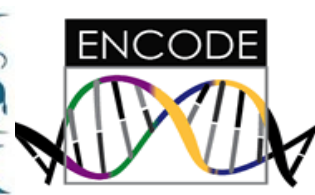


Projects using short read mapping

- Want to find all variations that may relate to disease and other phenotypes
 - 1000 Genome Project
 - Cancer Genome Atlas
 - ENCODE



Mutations
In Cancer

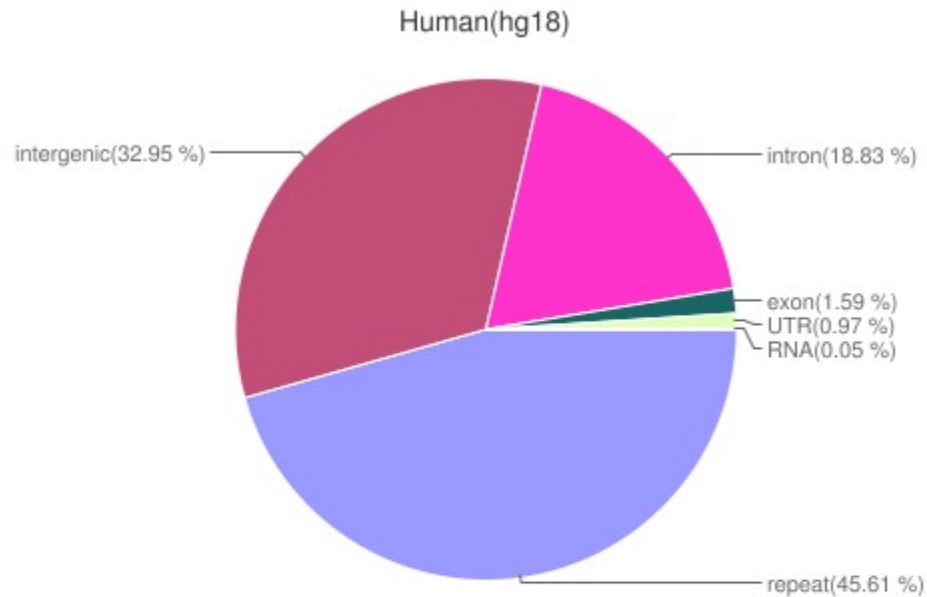


Challenges for variation discovery



- Low quality of reads/bases
- Coverage
- Repeat

Repeats



view ^	bases (bp)	%
repeat	1,405,061,157	45.61
intergenic	1,015,000,634	32.95
intron	580,090,705	18.83
exon	48,978,698	1.59
UTR	29,859,581	0.97
RNA	1,428,705	0.05
Whole Genome	3,080,419,480	100.0

- 46% of human genome is repetitive using standard repeat finding algorithms

(<http://www.ncrna.org/statgenome/index.html?view=class&gid=hg18>)

Tons of Mapping Tools

- BWA(Burrows-Wheeler Aligner)
 - Uses Burrows-Wheeler Transformation
 - BWA works very well for reads shorter than 200bp
 - Li, H., Ruan, J., and Durbin, R. (2008). Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome research*, 18(11), 1851–1858
- SOAP(Short Oligonucleotide Analysis Package)
 - "Li, R., Yu, C., Li, Y., Lam, T.-W. W., Yiu, S.-M. M., Kristiansen, K., and Wang, J. (2009b). SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics* (Oxford, England), 25(15), 1966–1967
- Bowtie
 - Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology* 2009, 10:R25
- Etc...

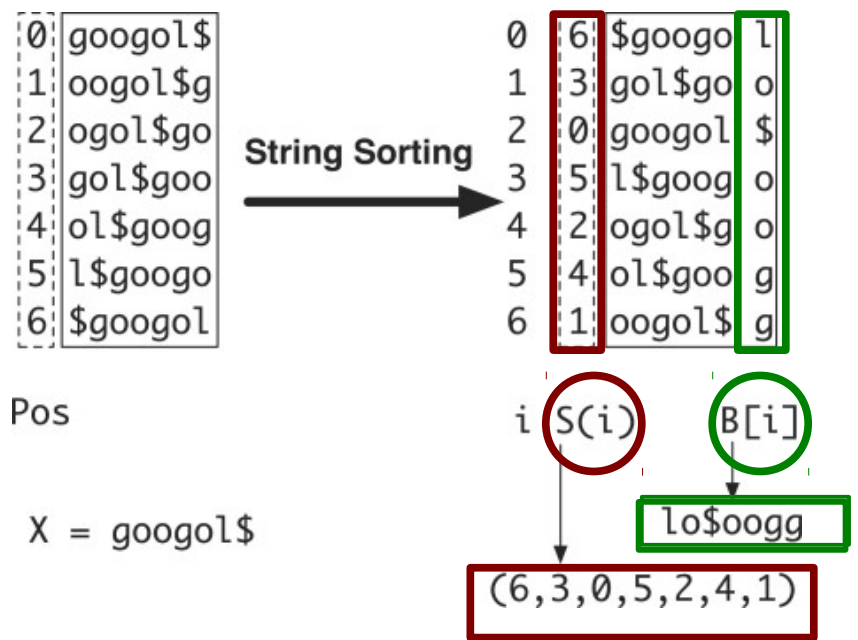
Burrows-Wheeler Aligner (BWA)

- Align relatively **short** sequences to a **long** reference genome
- Short reads mean less than 200bp with low error rate (<3%)
- Long reference genome
 - *Saccharomyces cerevisiae* – yeast (12M)
 - *Drosophila melanogaster* – fly (133M)
 - *Mus musculus* – mouse (2.7G)
 - *Homo sapien* – human (3 G)
 - *Pinus* - pine tree (24 Gbp)
 - *Protopterus aethiopicus* (130Gb)
 - Largest vertebrate genome known
 - *Pieris japonica* (150Gb)
 - Largest plant genome known



Burrows-Wheeler Transformation

- How to build BWT and suffix array
 - Suffix Array (Manber & Myers, 1991)
 - _ Lexicographically sorted list of suffixes
 - _ Fast binary search lookups: $O(\lg n) = 32$ probes / read
 - _ Relatively space efficient: $O(n \lg n) = 15\text{GB}$ / genome
 - BWT
 - _ BWT is a reversible permutation of the genome based on the suffix array
 - _ $< 1\text{GB}$ memory at peak time for constructing the BWT of human genome
 - _ implemented in BWT-SW (Lam et al., 2008)
 - _ Fast search and linear space requirements
- Given a string W, do binary search
 - Q: Find "go" in a given string "googol"
 - A: (1, 2)

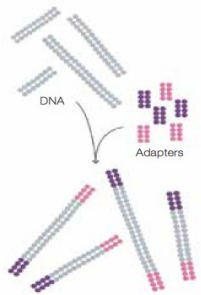


Outline

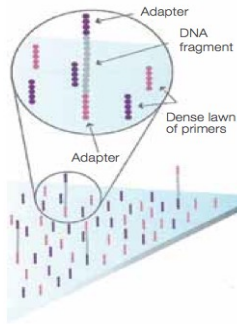
- Background on genome sequencing
- **Challenges for accurately measuring genome variations**
 - **Limitations of base quality score**
 - **Limitations of read quality score**
- Innovations for variations detection
 - Genome Mappability Score (GMS)
 - Genome Mappability Analyzer (GMA)
 - Applications of GMS
- Contributions and Future Work

Base Quality Score (1)

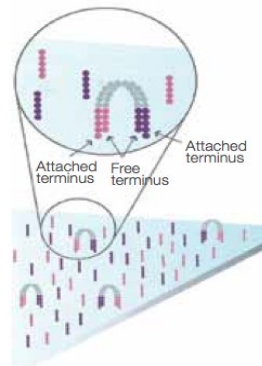
Illumina Sequencing by Synthesis



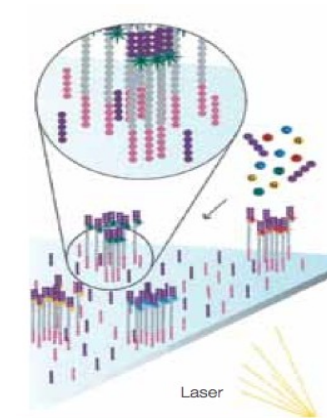
1. Prepare



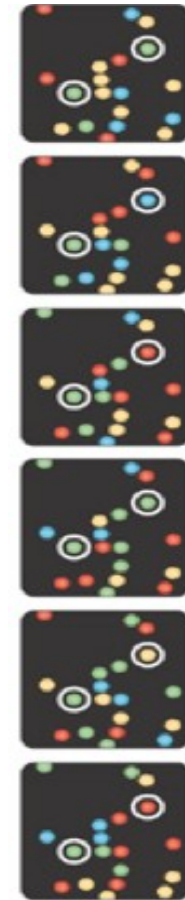
2. Attach



3. Amplify



4. Image



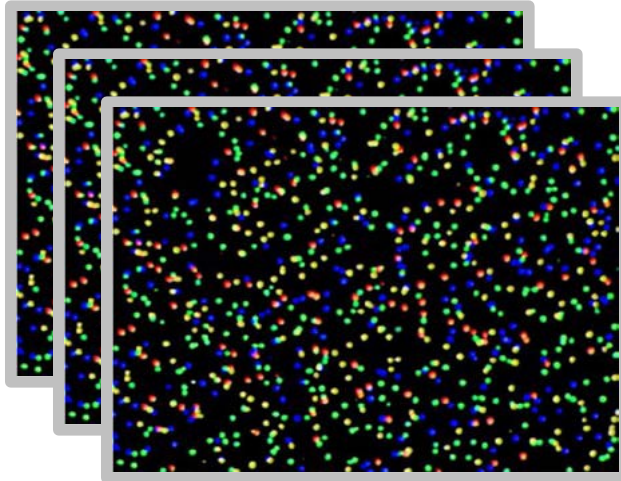
5. Basecall

Metzker (2010) Nature Reviews Genetics 11:31-46

http://www.illumina.com/documents/products/techspotlights/techspotlight_sequencing.pdf

Base Quality Score (2)

- Base-calling usually refers to the conversion of intensity data into sequences and quality scores.



Intensity
Analysis →

TTTACGATCGATCGTTGCATGCTGGGGTAGTGCTACTATA
 'GGGCTAGTTTCGATTTACGATCGATCGTTGCATGCTGGG
 CCG TTTACGATCGATCGTTGCATGCTGGGGTAGTGCTACTATA
 CGA 'GGGCTAGTTTCGATTTACGATCGATCGTTGCATGCTGGC
 ATG CCG
 CGA CGA TTTACGATCGATCGTTGCATGCTGGGGTAGTGCTACTATA
 GCC ATG 'GGGCTAGTTTCGATTTACGATCGATCGTTGCATGCTGGC
 ATT CCGATGGCCTGGGCTAGTTTCGATTTACGATCGATCGTT
 GCT CGATGGCCTGGGCTAGTTTCGATTTACGATCGATCGTTG
 :CTA GCC ATGCCGATGGCCTGGGCTAGTTTCGATTTACGATCGATC
 CCT ATT CGATGGCCTGGGCTAGTTTCGATTTACGATCGATCGTTG
 CTA/ GCT GCCTGGGCTAGTTTCGATTTACGATCGATCGTTGCATGC'
 TTC/ CTA/ ATTTACGATCGATCGTTGCATGCTGGGGTAGTGCTACTA
 TTT/ CTA/ GCTAGTTTCGATTTACGATCGATCGTTGCATGCTGGGGT
 :CTA CTA/ :CTAGTTTCGATTTACGATCGATCGTTGCATGCTGGGGT
 :TAG TTC/ CCTGGGCTAGTTTCGATTTACGATCGATCGTTGCATGCT
 CTG :CTA CTAGTTTCGATTTACGATCGATCGTTGCATGCTGGGGT
 TAG TTAGATTTACGATCGATCGTTGCATGCTGGGGTAGTGCTA
 TCG/ CTG :CTAGTTTCGATTTACGATCGATCGTTGCATGCTGGGGT
 TAG :TAGTTTCGATTTACGATCGATCGTTGCATGCTGGGGTAG
 TCG/ CTG/ CTGGGCTAGTTTCGATTTACGATCGATCGTTGCATGCTG
 TAGTTTCGATTTACGATCGATCGTTGCATGCTGGGGTAG'
 TCGATTACGATCGATCGTTGCATGCTGGGGTAGTGCTACT

Read

Base Quality Score(3)

- FASTQ format

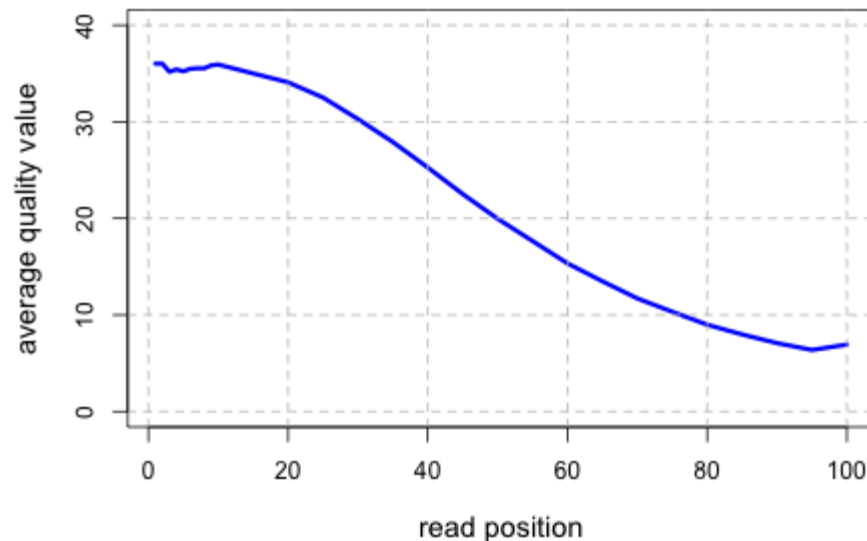
```
@SEQ_ID1
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!' '*((( (**+)) %%%++) (%%%) .1***-+*'') **55CCF>>>>>CCCCCCC65
...
...
...
```

- Phred-scaled base quality score

$$qv = -10 \log_{10} p_e$$

Base quality score	Error rate (%)
10	10%
20	1%
30	0.1%

Base Quality Score (4)



- Approximately the first 50 bp have qv 20, meaning the probability of error is less than 1%, while the latter half of the reads have considerably worse quality.

Quality value as a function of read position. The quality values at each read position were averaged from a sample of 100,000 100bp reads sequenced at the Broad Institute using an Illumina Genome Analyzer II (SRA study SRP001086).

Read Quality Score - MAQ(1)

$$Q_s = -10 \log_{10} [\Pr(\text{read is wrongly mapped})]$$

$$Q_s = -10 \log_{10} [1 - p_s(u|x, z)]$$

$$P_s(u|x, z) = \frac{P(z|x, u)}{\sum_{v=1}^{L-l+1} P(z|x, v)}$$

- The mapping quality score Q_s of a given alignment is typically written in Phred-scale
- Posterior probability P_s that the alignment is the correct alignment
 - $L = |x|$ the length of reference genome x ,
 - $l = |z|$ is a length of a read z
- $P(z|x, u)$, the probability of observing the particular read alignment
 - Defined as the product of the probability of errors recorded in the quality values.
 - The posterior error probability P_s is minimized when the alignment with the fewest mismatches is selected.
- Q_s will be lower for reads that could be mapped to multiple locations with nearly the same number of mismatches and Q_s will be zero if there are multiple positions with the same minimum number of mismatches weighted by quality value.

Read Quality Score – MAQ (2)

Reference ...GTCATCCTAATCGTATCTAGGCTCGATTCCGTA**T**GATTCCGGCCATGCAACGTCTCTGTTAGGTTCTC**G**TATCTAGGCTCGTATAGCTAGC...
CTCG**C**TTCCGTA**A**CTGTATAGATTCCGGCCA TACTGTATAGATTCCGGCCA

$$Q_s = -10 \log_{10} [1 - p_s(u|x, z)]$$

$$P_s(u|x, z) = \frac{P(z|x, u)}{\sum_{v=1}^{L-l+1} P(z|x, v)}$$

$$P(z|x, u)$$

$$\sum_{v=1}^{L-l+1} P(z|x, v)$$

- X is a reference
- Z is a read
- U is a position
- L = |x| the length of reference genome x,
- l = |z| is a length of a read z
- Position u has 2 mismatches
- Base quality scores are 20 for **C**, 10 for **A**
- Error probability of **C** is 1%, **A** is 10%
- Correctly mapped probability of position U is 0.1 %
- Q: If a read z is (almost) uniquely mapped?

Read Quality Score – MAQ (3)

Reference ...GTCATCCTAATCGTATCTAGGCTCGATTCCGTA**CT**GATTCCGGCCATGCAACGTCTCTGTTAGGTTCTC**G**TATCTAGGCTCGTATAGCTAGC...
 TCGTATCTAGGCTCGATTCCGTA TCGTATCTAGGCTCGATTCCGTA
 CTCG**C**TTCCGTA**C**TCTGTATAGATTCCGGCCA

$$Q_s = -10 \log_{10} [1 - p_s(u|x, z)]$$

- X is a reference
- Z is a read
- U is a position
- $L = |x|$ the length of reference genome x,
- $l = |z|$ is a length of a read z

$$P_s(u|x, z) = \frac{P(z|x, u)}{\sum_{v=1}^{L-l+1} P(z|x, v)}$$

$$\sum_{v=1}^{L-l+1} P(z|x, v)$$

- Q: If a read z is mapped to many positions?
- Q: What is the reliability of a specific position?
- Q: Do we have a metric to measure such reliability in a consistent view?

Challenges

- Base quality score is useful only inside a single read
- Read quality score provides only local view
- Read quality score is very sensitive to a minute change

Uniqueome

- Partially addressed by Uniqueome
 - Ryan Koehler et al., The uniqueome: a mappability resource for short-tag sequencing, BIOINFORMATICS, 27:2 (12 November 2010), pp. 272-274
- Measure if individual reads are mapped uniquely allowing a fixed number of mismatches
- Still sensitive because it does not consider all possible reads
- We need more stable “**GPS**” for a genome

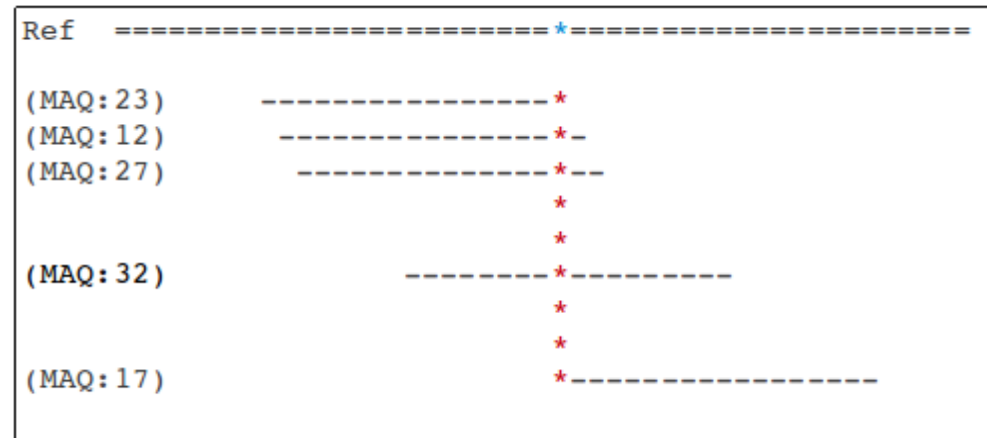
Outline

- Background on genome sequencing
- Challenges for accurately measuring genome variations
- **Innovations for variations detection**
 - **Genome Mappability Score (GMS)**
 - **Genome Mappability Analyzer (GMA)**
 - **Applications of GMS**
- Contributions and Discussion

The Global View (GPS for a genome)

- There is inherent uncertainty to mapping
- However, there is no tool to measure the reliability of mapped reads to the reference genome in a global perspective.

Genome Mappability Score (GMS)



$$GMS(u) = \frac{100}{|z|} \sum_{\forall z \ni u} p_s(u|x, z) = \frac{100}{l} \sum_{\forall z \ni u} \left(1 - 10^{-\frac{Q_s(u|x, z)}{10}}\right)$$

- u is a position
- x is a reference
- z is a read
- l is read length

Any Questions?

- At this moment, 2 questions can be raised naturally
- (1) Why are simulated reads used?
 - Some may want to argue that simulated reads are biased, thus cannot reflect what real reads in wet-lab experiments can.
- (2) Why is BWA used instead of other mapping tools?

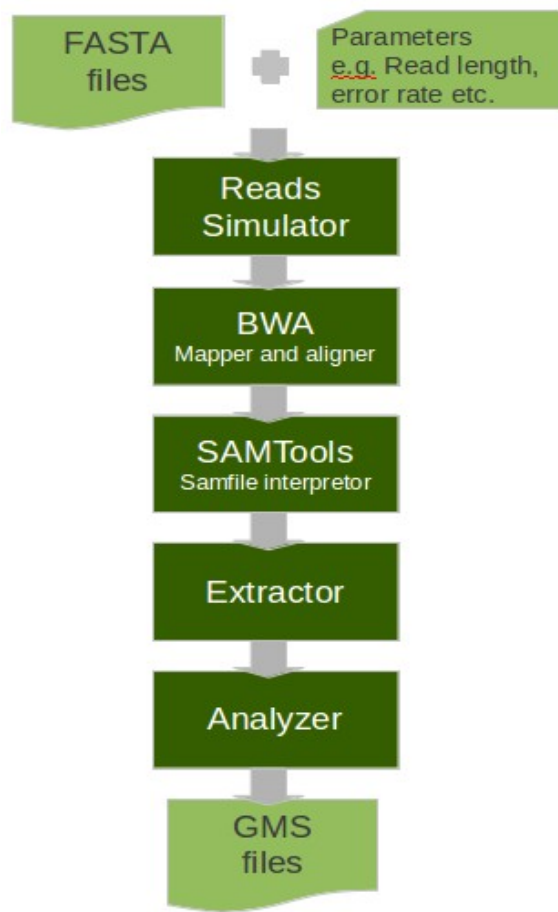
Why simulated reads are used

- To simulate ideal case
 - We need all possible reads related to specific position
 - Technically impossible to achieve
- To have a full control over all parameters
 - -l : Read length
 - -e : Error rate
 - -q : Quality value
 - -o : Expected distance for paired-end read
- Upper bound on what we can achieve

Why BWA is used instead of other mapping tools

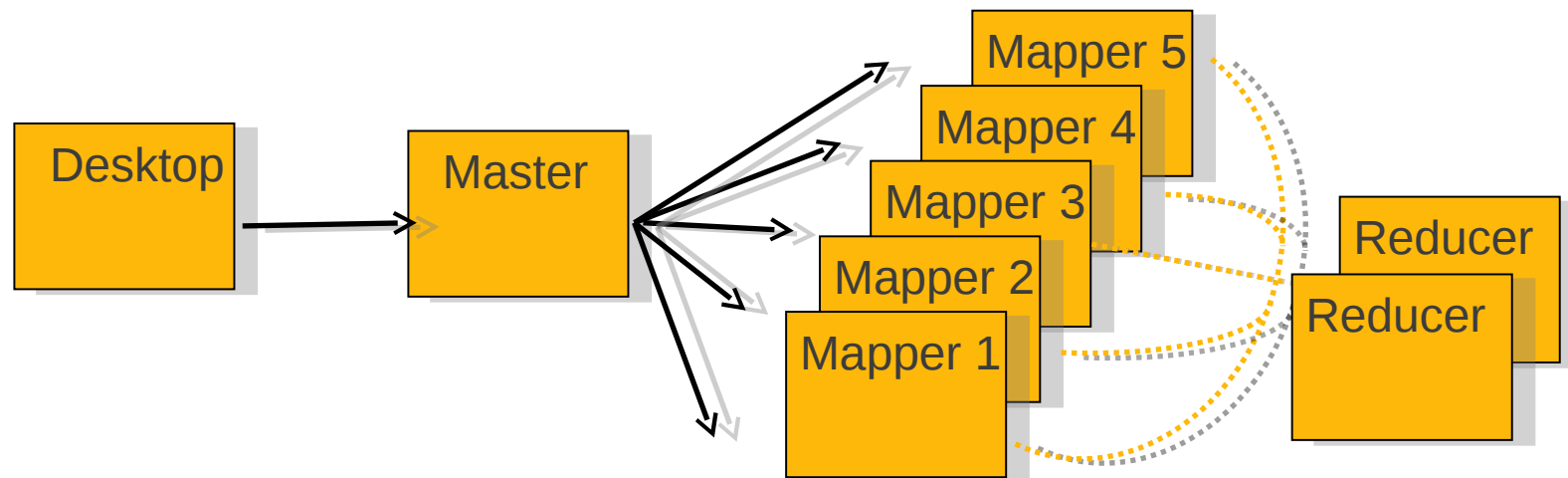
- Holtgrewe et al. (2011) benchmarked most reputable mapping tools and reported their performance.
 - + BWA and Shrimp2 outperforms and shows best and stable results
 - Bowtie and Soap2 is fluctuated by error rates and read length.
 - + BWA can tackle reads with indels in long reads
 - Bowtie and Soap2 cannot.
 - Shrimp2 is very sensitive to indels
- Remember: Our purpose is not finding all possible positions. We need one best probable result because we are looking for tool-independent and inherent tendencies of a genome.

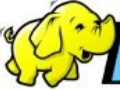

Genome Mappability Analyzer (GMA)



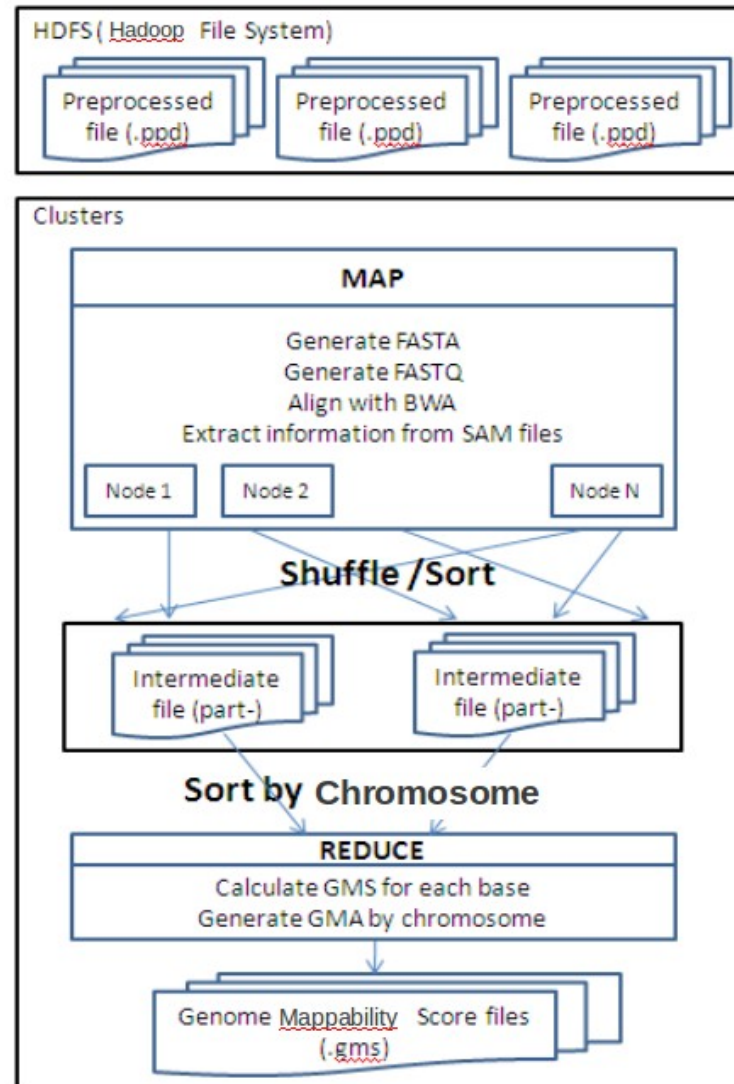
- GMA computes GMS for every position of a genome
- To compute GMS, all you need is FASTA file with parameter settings
 - -l : Read length
 - -e : Error rate
 - -q : Quality value
 - -o : Expected distance for paired-end read
- Local version for small-sized genome
 - Good enough for yeast(12M) or fly(133M)
 - Not efficient for large genome such as human(3G), which takes ~1 month

Cloud Computing using Hadoop



- We are in the age of data tsunami
- For large data, use parallel computing on cloud
-  **hadoop** is the leading open source implementation, developed by **YAHOO!**
 - Free version of MapReduce patented by 
 - Scalable, efficient, reliable, easy to write a program

Genome Mappability Analyzer (GMA)



- Hadoop version for large scale genomes
- ~ 1 month for entire human genome using 1 core
- 1 day on 48 cores with Hadoop
- 8 days for pine tree with Hadoop instead of 8 months

Outline

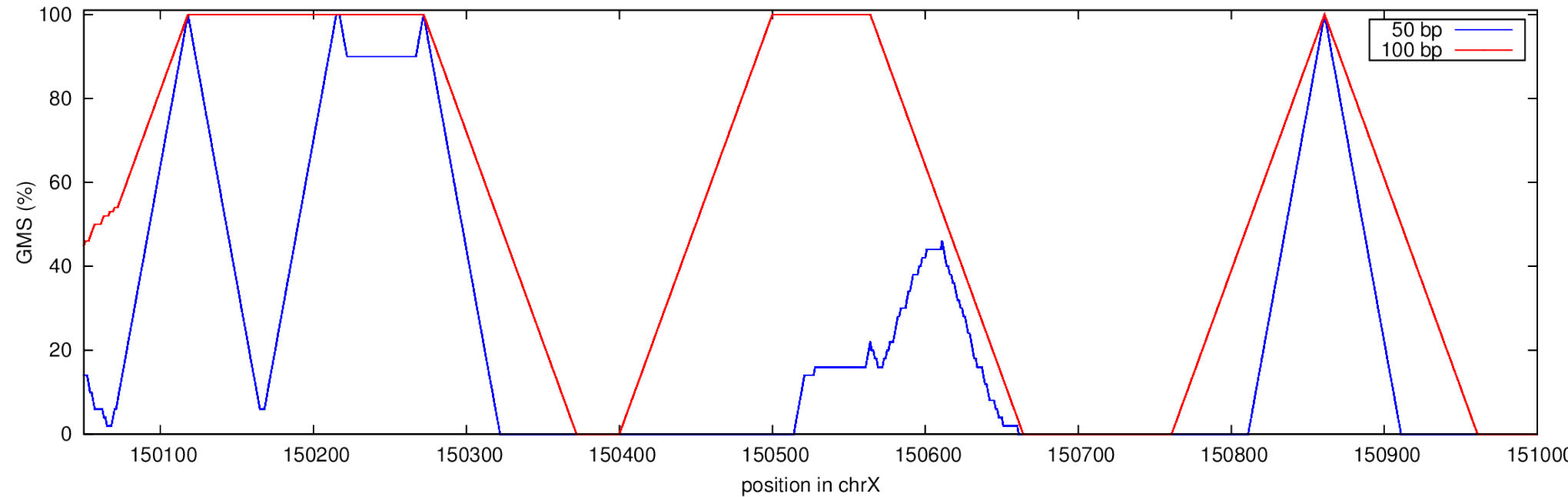
- Background on genome sequencing
- Challenges for accurately measuring genome variations
- Genome Mappability Score (GMS)
- **Innovations for variations detection**
 - Genome Mappability Score (GMS)
 - Genome Mappability Analyzer (GMA)
 - **Applications of GMS**
 - **Effect of parameters on GMS**
 - **GMS profiles**
 - Model organisms
 - Human pathogen *T. vaginalis*
 - **Variation Discovery and Dark Matter**
 - Variation Accuracy Simulator (VAS)
 - Dark Matter
- Contributions and Future Work

Effect of parameters on GMS

- Different parameters are used to trace the effect of different sequencing conditions
- -l : Read length
- -e : error rate
- -o :
 - expected distance for paired-end,
 - default is 0, which means single-end

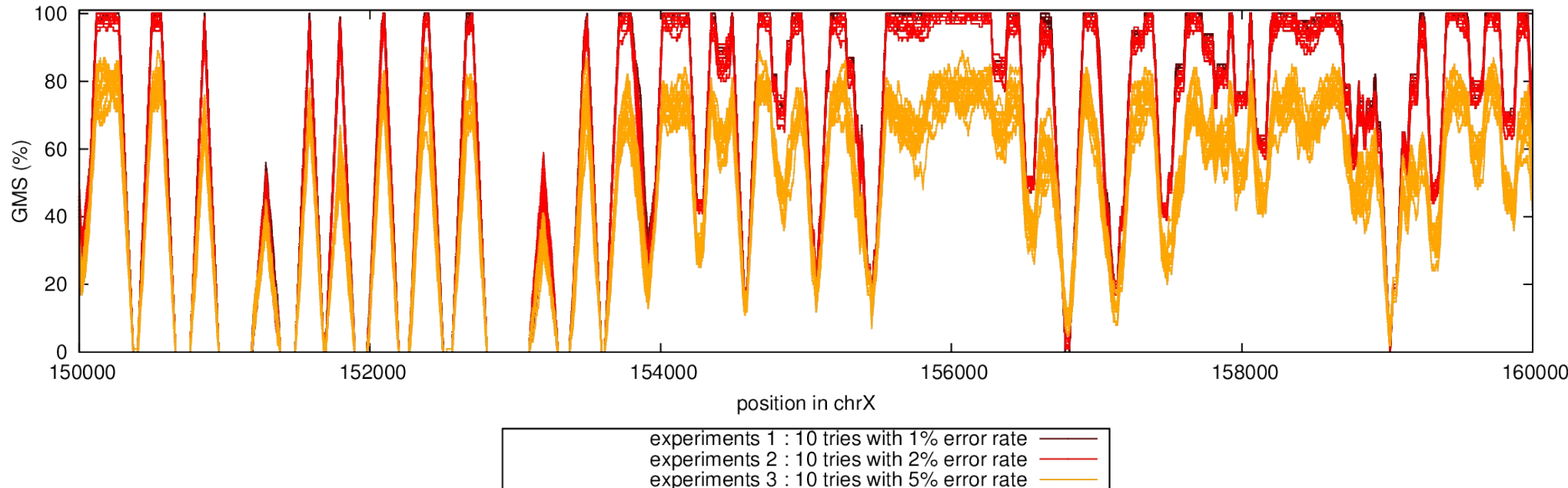
Read Length

Consistency of GMS (chrX of hg19, read length: 50bp vs. 100bp, error rate: 0%)



Error Rate

Consistency of GMS (chrX of hg19, read length: 100bp, error rate: 1% vs 2% vs 5%)



- Mutation rate(SNP) is 0.1%
- GMS shape will be consistent, independent from individual differences

GMS Profiles for Model Organisms

- Using common parameters
 - 100bp read length
 - 2% error rate
 - 300 expected distance for paired-end reads
- GMS $\geq 50\%$ is highly reliable region
- Percentage of highly mappable bases in the genomes of several model species. Approximately 90% of these genome can be mapped reliably.

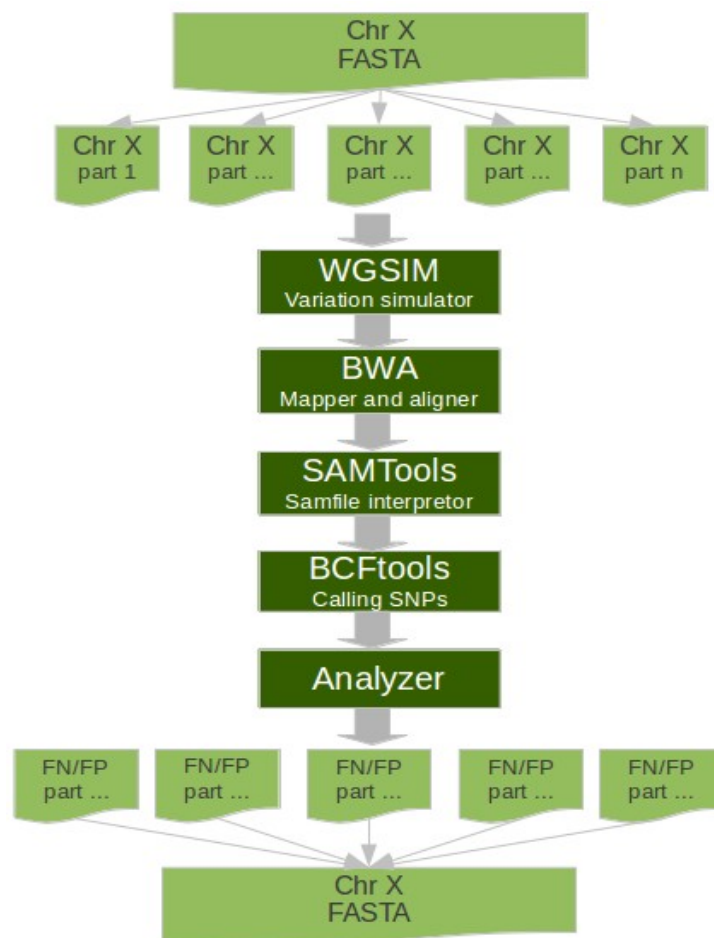
Species (build)	size	whole (%)	transcribed (%)	exon (%)
yeast (sc2)	12 Mbp	95.0	95.1	95.1
fly (dm3)	133 Mbp	88.9	91.7	92.8
mouse (mm9)	2.7 Gbp	86.5	91.1	91.2
human (hg19)	3.0 Gbp	86.1	94.2	94.4

GMS Profiles for *T. vaginalis*

GMS range	count (bp)	%
[0, 10]	66,295,633	42.3
(10, 20]	3,633,619	2.3
(20, 30]	3,198,130	2.0
(30, 40]	2,933,152	1.8
(40, 50]	2,738,858	1.7
(50, 60]	2,568,843	1.6
(60, 70]	2,434,898	1.5
(70, 80]	2,356,404	1.5
(80, 90]	2,332,921	1.4
(90, 100]	68,072,087	43.4

- Distribution of GMS values in the *T.vaginalis* genome. Since *T. vaginalis* has high proportion of repeats, over half of the genome cannot be reliably mapped.

Variation Accuracy Simulator (VAS)



- Simulation of resequencing experiments to measure the accuracy of variation detection
- SAMTools/BCFTools are leading programs to discover variations
- Local/Hadoop version

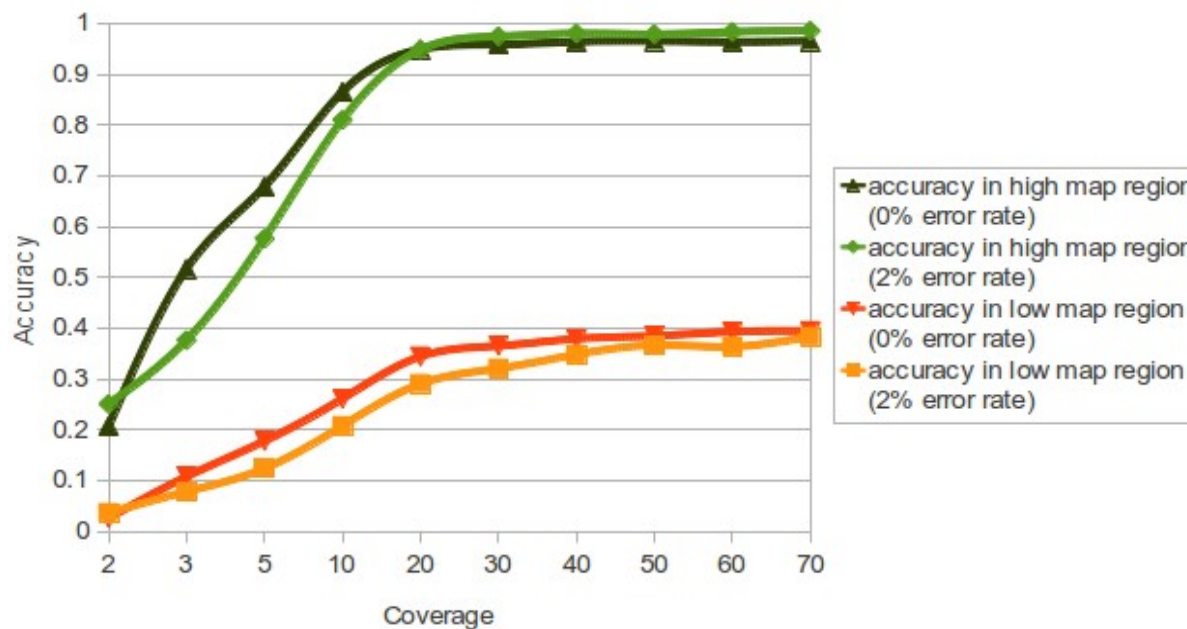
Accuracy Test

	Low GMS Region	High GMS Region
Total Simulated Mutations	5,636	145,094
Correct SNVs	2,381	144,845
False Positive	1	51
False Negative	3,255	249
Accuracy	0.4225	0.9983

- The overall variation detection accuracy is very high, and is twice as high (99.83%) in high GMS regions compared to low GMS regions (42.25%)
- Detection failure errors are dominated by false negatives
 - The SNP-calling algorithm will use the mapping quality score to filter out low confidence mapping.
- What is surprising is the extent of false negatives and the concentration of false negatives almost entirely within low GMS regions.
- Among all 3504 false negatives, 3255 (93%) are located in low GMS region, only 249 (7%) are in high GMS region.
 - Only 14% of human genome is low GMS region

Dark Matter

Variation Discovery Accuracy in High/Low GMS Region



- Unlike false negatives in high GMS region that can be discovered in high coverage (≥ 20 -fold), false negatives in low GMS regions cannot be discovered, because variation calling program will not use poorly mapped reads

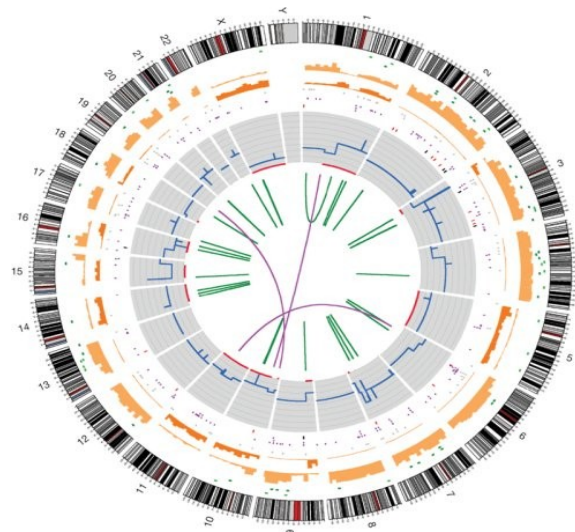
GMS ratio of Human (hg19)

	whole	transcription	coding	exon	SNP (dbSNP)	Clinical SNP
[0,10]	0.0918	0.0128	0.0109	0.0129	0.0060	0.0033
(10,20]	0.0056	0.0056	0.0054	0.0054	0.0052	0.0042
(20,30]	0.0051	0.0050	0.0048	0.0048	0.0046	0.0035
(30,40]	0.0052	0.0051	0.0049	0.0049	0.0048	0.0047
(40,50]	0.0053	0.0052	0.0050	0.0050	0.0049	0.0037
(50,60]	0.0055	0.0054	0.0051	0.0051	0.0052	0.0036
(60,70]	0.0058	0.0056	0.0054	0.0054	0.0056	0.0046
(70,80]	0.0063	0.0060	0.0058	0.0057	0.0062	0.0057
(80,90]	0.0073	0.0068	0.0065	0.0064	0.0073	0.0049
(90,100]	0.8620	0.9425	0.9462	0.9444	0.9503	0.9617
TOTAL	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Clinical SNPs in Low GMS Region

- There are very important clinical SNPs in low GMS Region
 - rs445114
 - PROSTATE CANCER, HEREDITARY, 10; HPC10
 - GMS : 3.597204
 - rs944289
 - thyroid carcinoma
 - GMS : 3.732166
 - rs1016732
 - AUTISM
 - GMS : 9.99999

Apply GMS to ongoing projects



- 1000 genome
 - Identify common variation in the human genome
- Cancer Genome Atlas
 - Identify variations related to various forms of cancer
- ENCODE
 - Use sequencing to identify all biologically active regions of the genome

Outline

- Background on genome sequencing
- Challenges for accurately measuring genome variations
- Innovations for variations detection
 - Genome Mappability Score (GMS)
 - Genome Mappability Analyzer (GMA)
 - Applications of GMS
- **Contributions and Discussion/Future Work**

Contributions (1)

- Short read mapping
- Mapping Algorithms are getting mature
- Open question how to interpret the mapping reliability
- Previous Works
 - Mapping quality score
 - Uniqueome
- Challenge
 - Narrowly focus on an individual read
 - Largely miss genomic context
 - Too sensitive to reflect genomic characteristics

Contributions (2)

- GMS(Genome Mappability Score)
 - A novel probabilistic metric
 - Measure how reliably reads are mapped
- GMA(Genome Mappability Analyzer)
 - Stand-alone version
 - Cloud version (Hadoop)
- GMS profiles for Model organisms
 - 14% of the human genome is in low mappability regions
 - 6% of exons are in low mappability regions

Contributions (3)

- Variations Accuracy Simulation
 - The overall variation detection accuracy is very high, and is twice as high (99.83%) in high GMS regions compared to low GMS regions (42.25%)
 - Detection failure errors are dominated by false negatives
 - Among all 3504 false negatives, 3255 (93%) are located in low GMS region, only 249 (7%) are in high GMS region.
 - Cannot be overcome by merely increasing coverage
- Hidden mutations are genomic dark matter.
 - Important in disease analysis
- GMS should be considered for all analysis of resequencing projects
 - False negative
 - False positive

Future Works

- Evaluate the GMS profile under the models with longer read length and higher error rates
 - e.g. Ion Torrent
 - e.g. Pacific Biosciences
 - Several thousand bases, 15% error rates.
- Find a way to call variations in low GMS region (if possible)

Acknowledgements

CSHL

Michael Schatz
Mitch Bekritsky

SBU

Fatma Betul
Matt Titmus
Steve Skiena
Yejin Choi

JHU

Ben Langmead



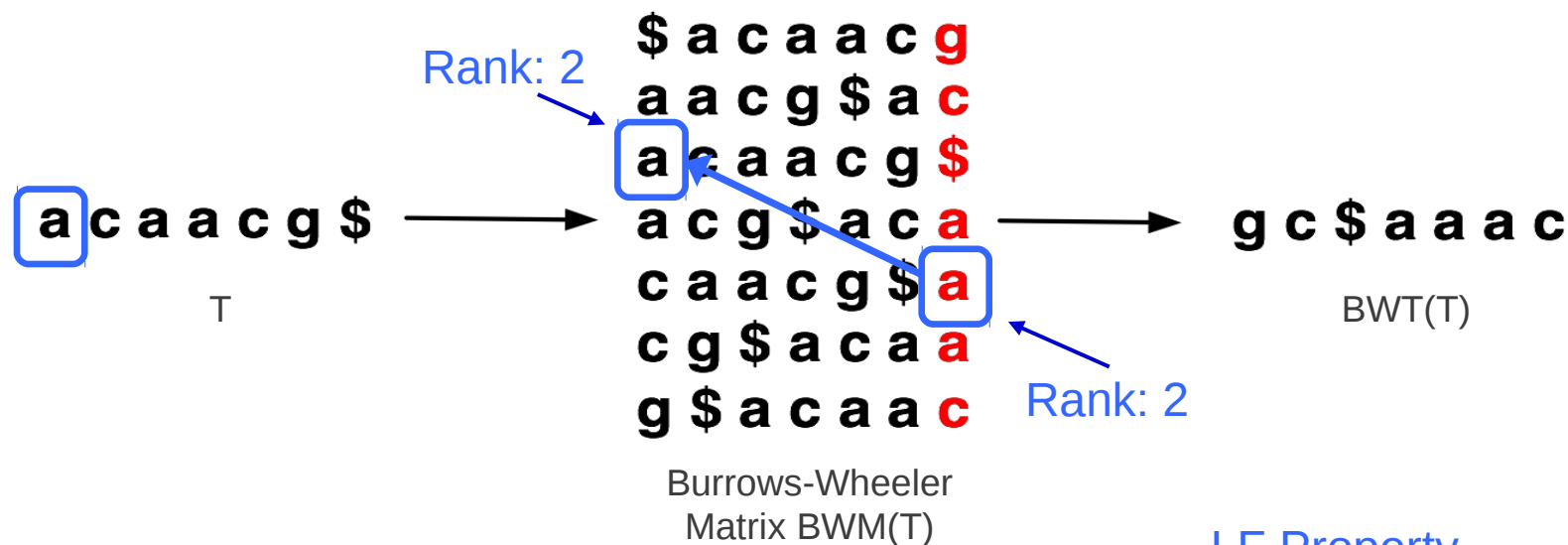
Thank You!

Appendix A

Burrows-Wheeler Transform

Burrows-Wheeler Transform

- Reversible permutation of the characters in a text



LF Property
 implicitly encodes
 Suffix Array

- $BWT(T)$ is the index for T

A block sorting lossless data compression algorithm.

Burrows M, Wheeler DJ (1994) *Digital Equipment Corporation*. Technical Report 124

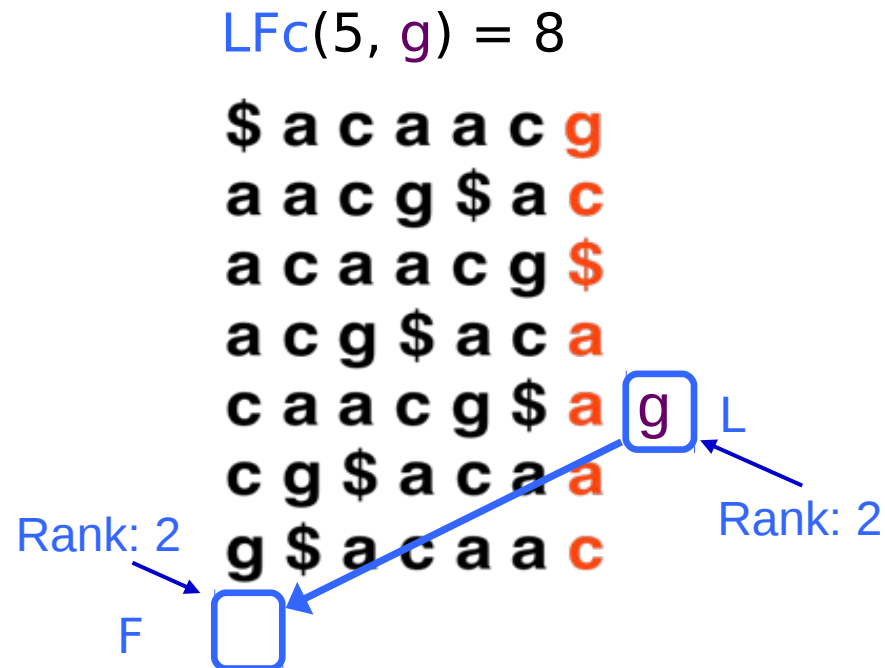
Burrows-Wheeler Transform

- Recreating T from BWT(T)
 - Start in the first row and apply **LF** repeatedly, accumulating predecessors along the way



Exact Matching

- **LFc**(r, c) does the same thing as **LF**(r) but it ignores r's actual final character and “pretends” it's c:

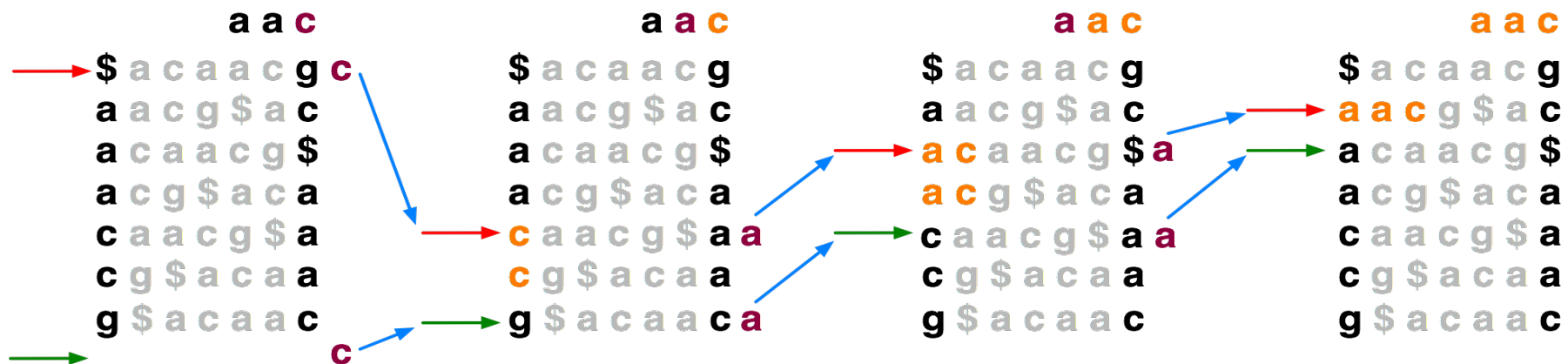


Exact Matching

- Start with a range, (**top**, **bot**) encompassing all rows and repeatedly apply **LFc**:

top = **LFc**(**top**, **qc**); **bot** = **LFc**(**bot**, **qc**)

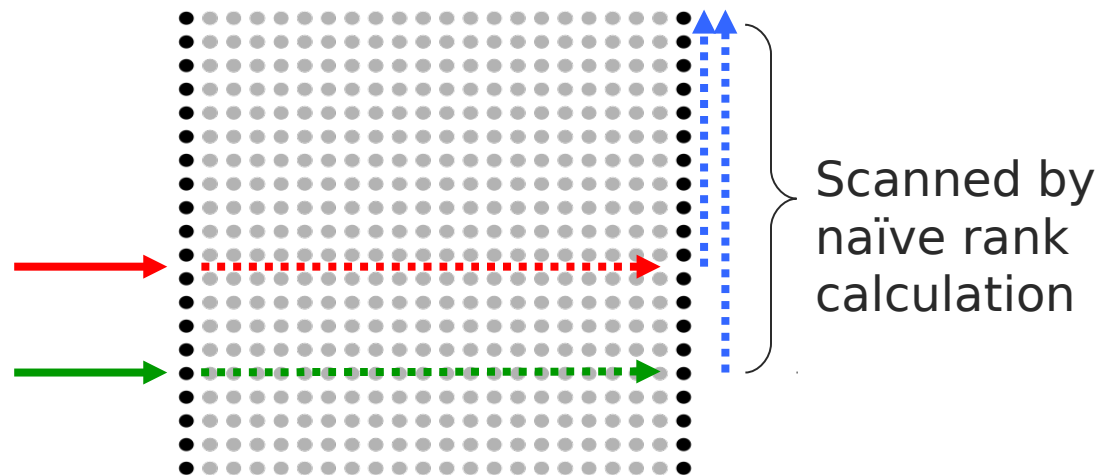
qc = the next character to the left in the query



Ferragina P, Manzini G: Opportunistic data structures with applications. *FOCS. IEEE Computer Society; 2000.*

Checkpointing in FM Index

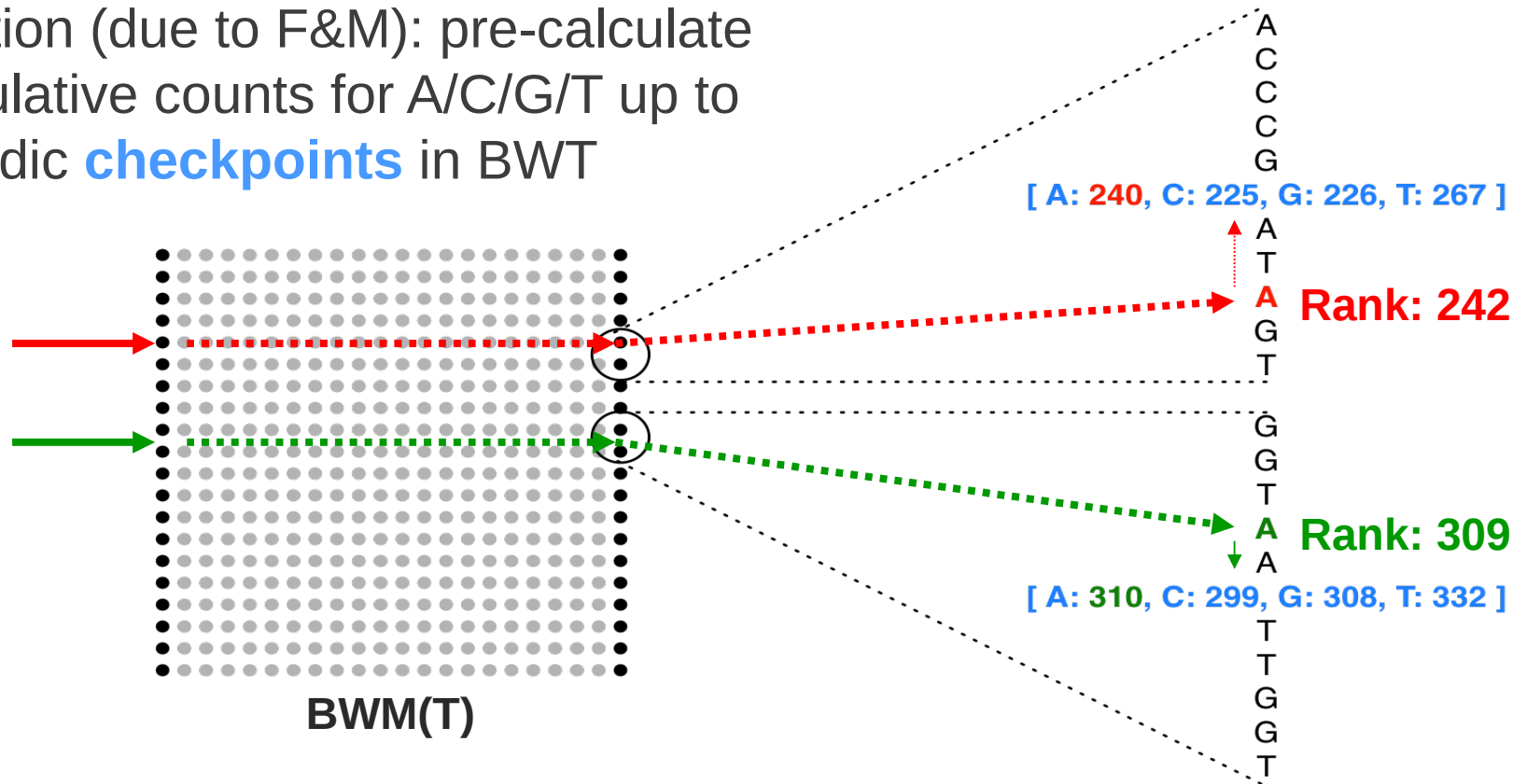
- $LF(i, qc)$ must determine the *rank* of qc in row i
- Naïve way: count occurrences of qc in all previous rows
 - Linear in length of text – too slow



BWM(T)

Checkpointing in FM Index

- Solution (due to F&M): pre-calculate cumulative counts for A/C/G/T up to periodic **checkpoints** in BWT

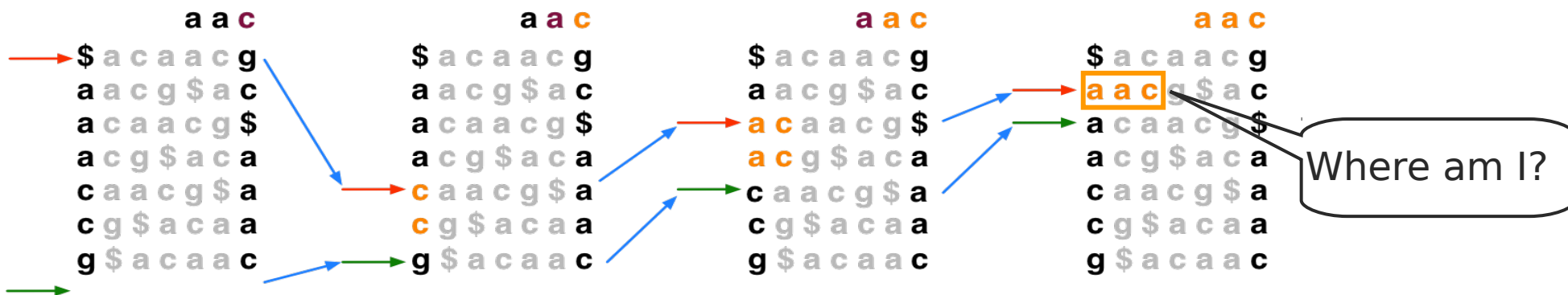


- $LF(i, qc)$ is now constant time

(if space between checkpoints is considered constant)

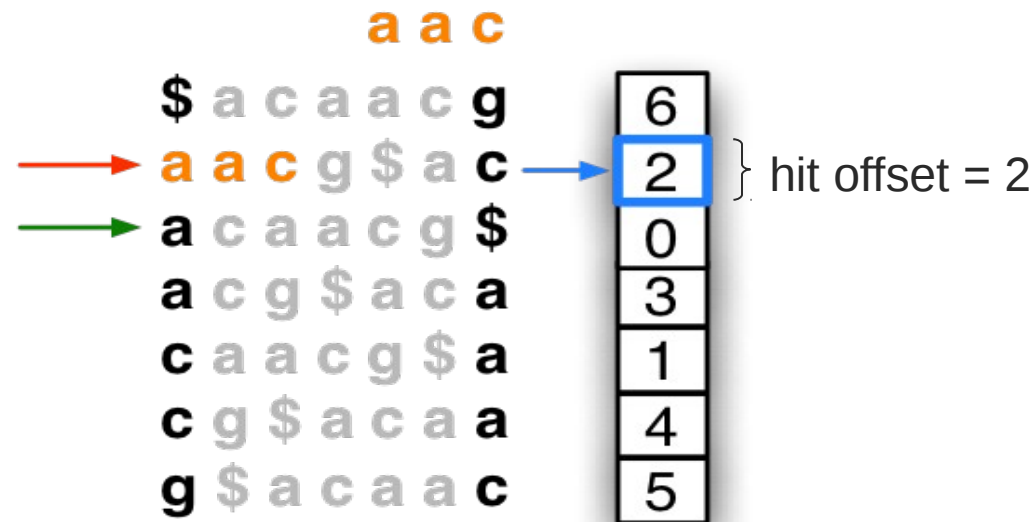
Rows to Reference Positions

- Once we know a row contains a legal alignment, how do we determine its position in the reference?



Rows to Reference Positions

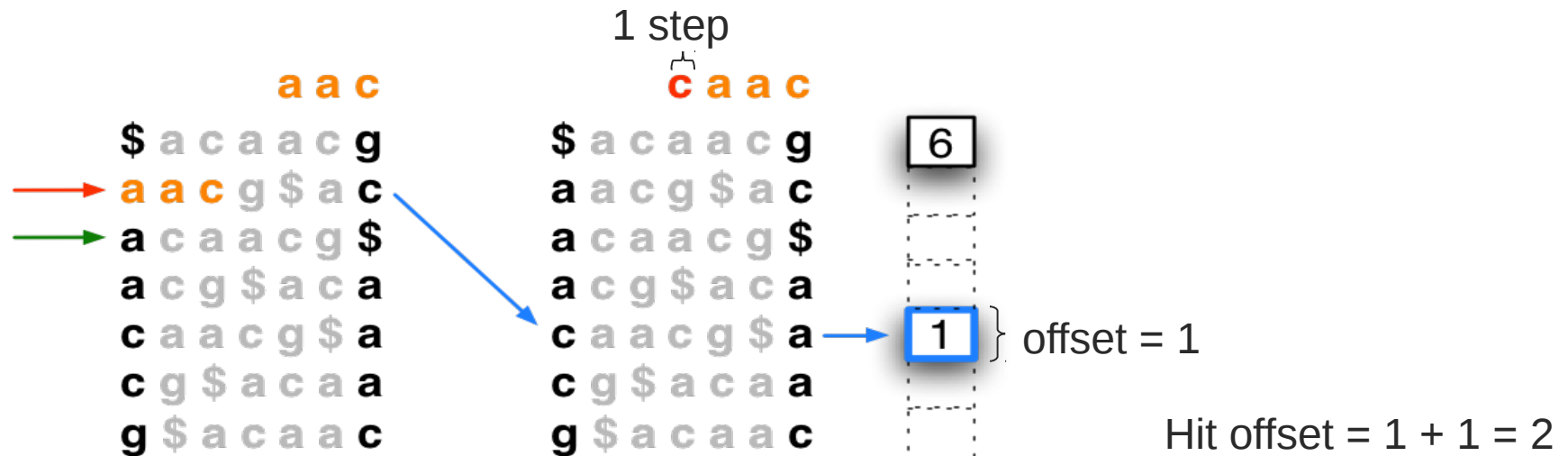
- Naïve solution 2: Keep pre-calculated offsets (the suffix array) in memory and do lookups



- Suffix array is ~12 GB for human – too big

Rows to Reference Positions

- Hybrid solution (due to F&M): Pre-calculate offsets for some “marked” rows; use UNPERMUTE to walk from the row of interest to next marked row to the left

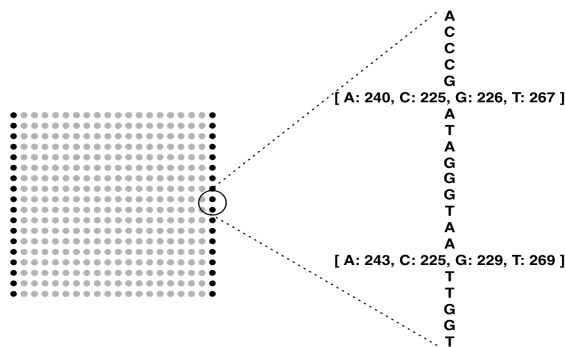


- Bowtie marks every 32nd row by default (configurable)

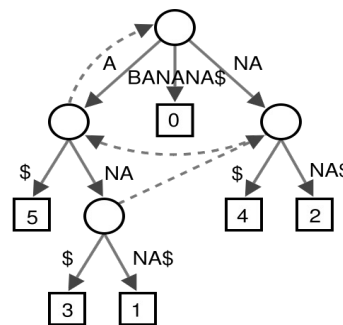
FM Index is Small

- Entire FM Index on DNA reference consists of:
 - BWT (same size as T)
 - Checkpoints (~15% size of T)
 - SA sample (~50% size of T)
- Total: ~1.65x the size of T

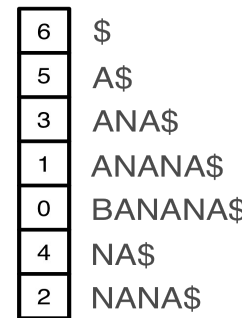
Assuming 2-bit-per-base encoding and no compression, as in Bowtie
 Assuming a 16-byte checkpoint every 448 characters, as in Bowtie
 Assuming Bowtie defaults for suffix-array sampling rate, etc



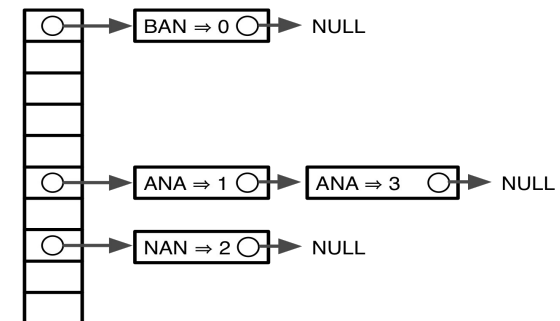
~1.65x



>45x



>15x



>15x