Genome Assembly

Sijia Huo Sean Kelly Gregory Raskind

University of Michigan

University of Illinois

14 July 2017

Background

Genomics BDSI

Sijia Huo, Sean Kelly, Gregory Raskind

In our cells, DNA carries genetic information:

- Each DNA strand contains a sequence of A, C, G, and Ts
- The human genome is around 3 billion base pairs long

Types of DNA Variation:

- Single base variations
- Copy number variations

```
>HSBGPG
GGCAGATTCCCCCTAGACCCGCCCGCACCATGGTCAG
TGGGCACAGCCCAGAGGGTATAAACAGTGCTGGAGGC
AGTCCTGAGCAGCAGCCCAGCGCAGCCACCGAGACAC
CTCGCCCTATTGGCCCTGGCCGCACTTTGCATCGCTG
CCACCTCCCCTCAGGCCGCATTGCAGTGGGGGCTGAG
CACCTCTTCTCACCCCTTTGGCTGGCAGTCCCTTTGC
AGGCTCAATCCATTTGCCCCAGCTCTGCCCTTGCAGA
AGCTGCCCGAGACGCAGGGGAAGGAGGATGAGGGCCC
ACCAGGCTCCCTTTCCTTTGCAGGTGCGAAGCCCAGC
GTGCAGGTATGAGGATGGACCTGATGGGTTCCTGGAC
CCCTCAGTCTCATTCCCCCACTCCTGCCACCTCCTGT
GCCTGCTCCCCACCTGATCCTCCCAAACCCAGAGCCA
CTCCACAGCCTTTGTGTCCAAGCAGGAGGGCAGCGAG
GCTACCTGTATCAATGGCTGGGGTGAGAGAAAAGGCA
```

(a) FASTA file stores reference genome

```
@H3GFVCCXX150415:8:2224:9627:35467/1
GGGAATTTTAACTGGCAAAACTCAGAACTCCATCCAAAC
+
AAAFF<<FFAFAFAAAAFFFAFAAAAAAFAAAFAAFAFA
@H3GFVCCXX150415:8:2224:8957:23407/1
CATACTTGATGGTCTCAGATATGTGTGGATTTTGGAATT
+
<FAFAFFAAAA7AAFAAAAFAAAAAAFAFAFFFAAFFFF
@H3GFVCCXX150415:8:2224:8907:25745/1
GTTAATTAAAAGCCCTTTACGAATGGACTAGATGTACCT
+
AAAFFAFAF/FFAAAA<FAA7FFAAAFAFAF7FAAFAAF
@H3GFVCCXX150415:8:2224:8825:55175/1
GCACCCTGTGTCAACAACCTGAGAGTGGCCTTGAGTTGC
+
AAFAAAFAAAAAFFAFFAAFAFAAAAFAAAFFFAFAFAA
```

(b) FASTQ file stores sequenced reads

Data Source: National Center for Biotechnology Information

Mapping vs Assembly

Genomics BDSI

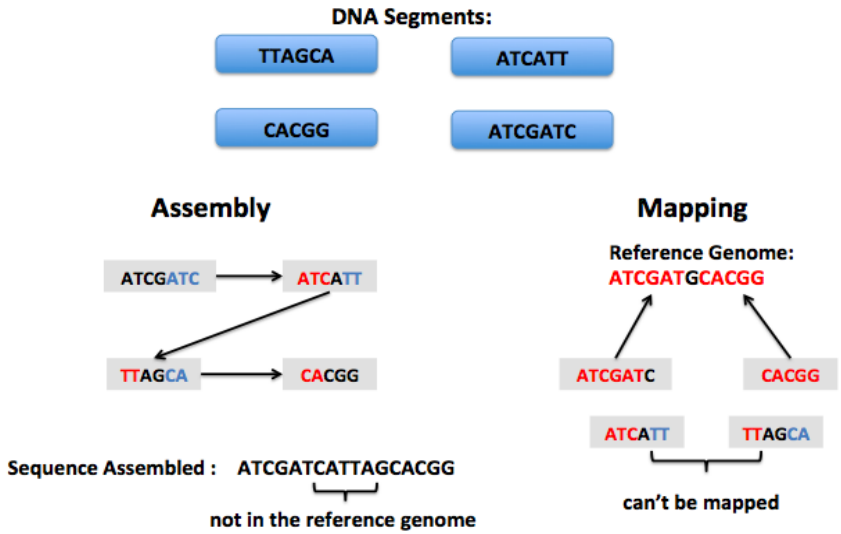
Sijia Huo, Sean Kelly, Gregory Raskind

Assembly

- Reconstructs genome from sequenced reads
- Particularly useful to detect large genetic variants. (i.e. free from bias towards reference genome)
- Slow, memory-intensive, and hard to implement

Mapping

- Match to reference genome to locate variation
- Fast, easy to understand
- Cannot handle sequenced reads containing complex variation

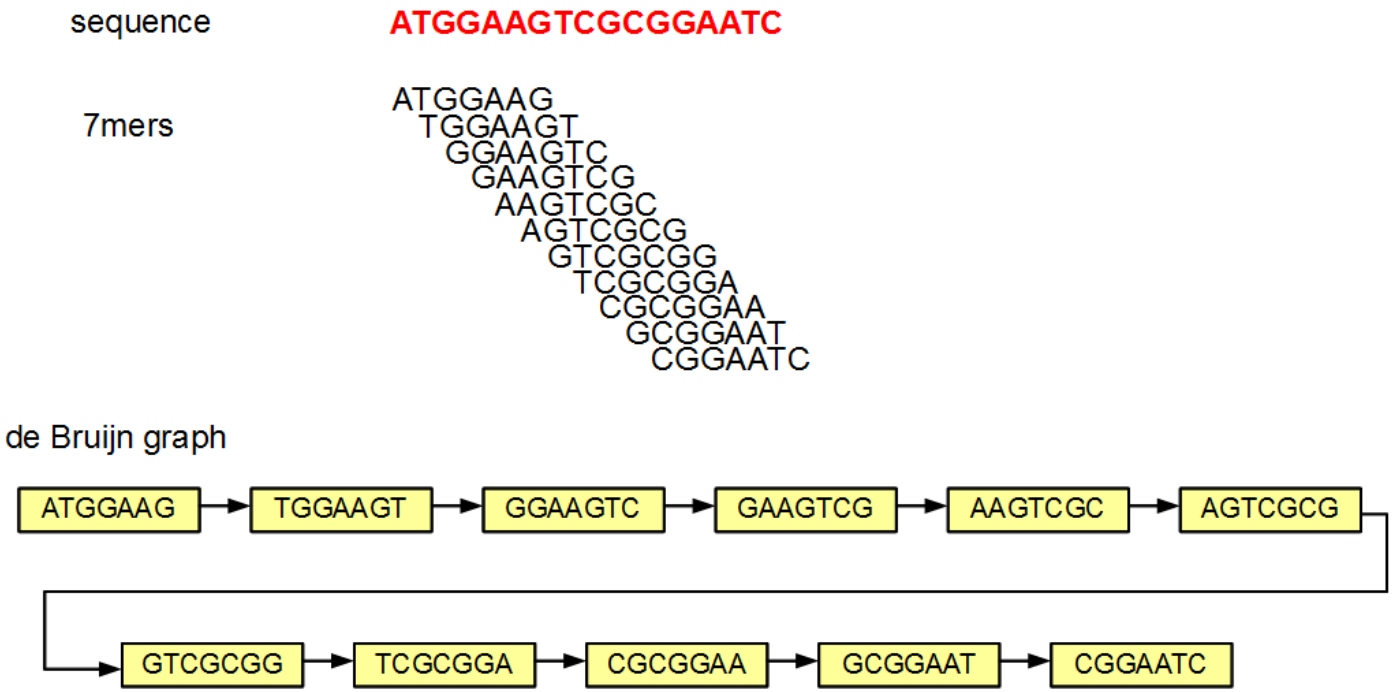


Constructing De Bruijn Graph of Reference Genome

Genomics BDSI

Sijia Huo, Sean Kelly, Gregory Raskind

Create process to store and reassemble a reference genome
Method: Break DNA string into 'kmers' and store in a De Bruijn Graph

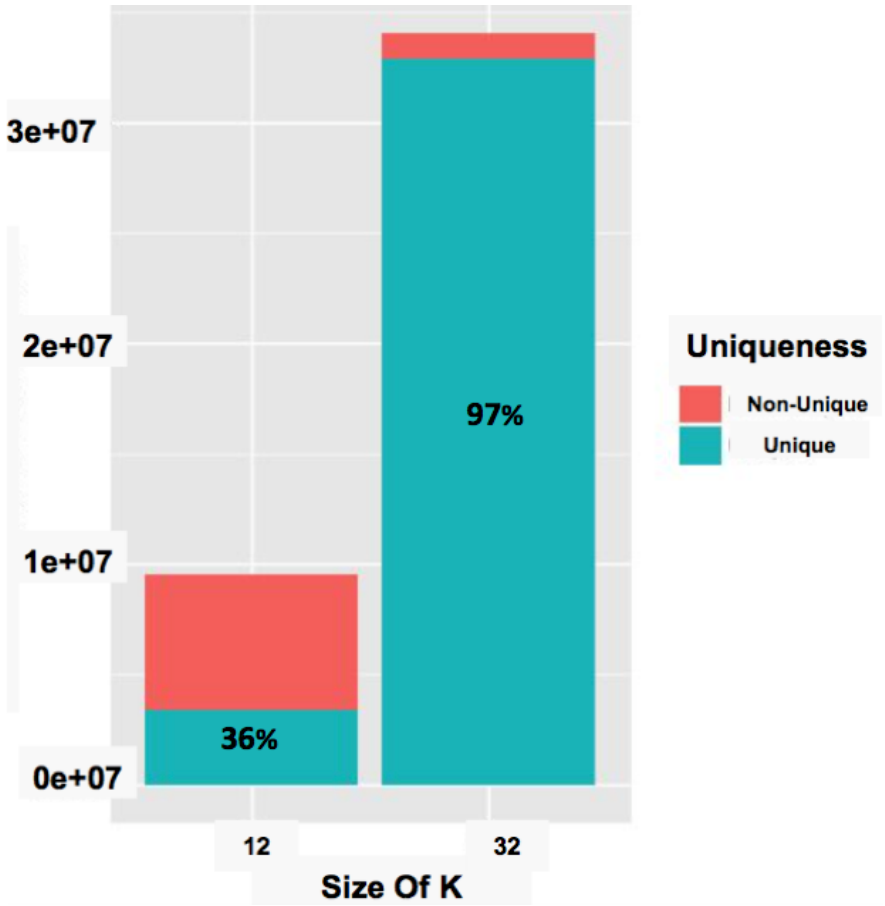


http://www.homolog.us/Tutorials/Tut-Img/Set1/fig5.png

Impact of K on complexity of De Bruijn Graph

Genomics BDSI

Sijia Huo, Sean Kelly, Gregory Raskind



Methodology: Functions

Genomics BDSI

Sijia Huo, Sean Kelly, Gregory Raskind

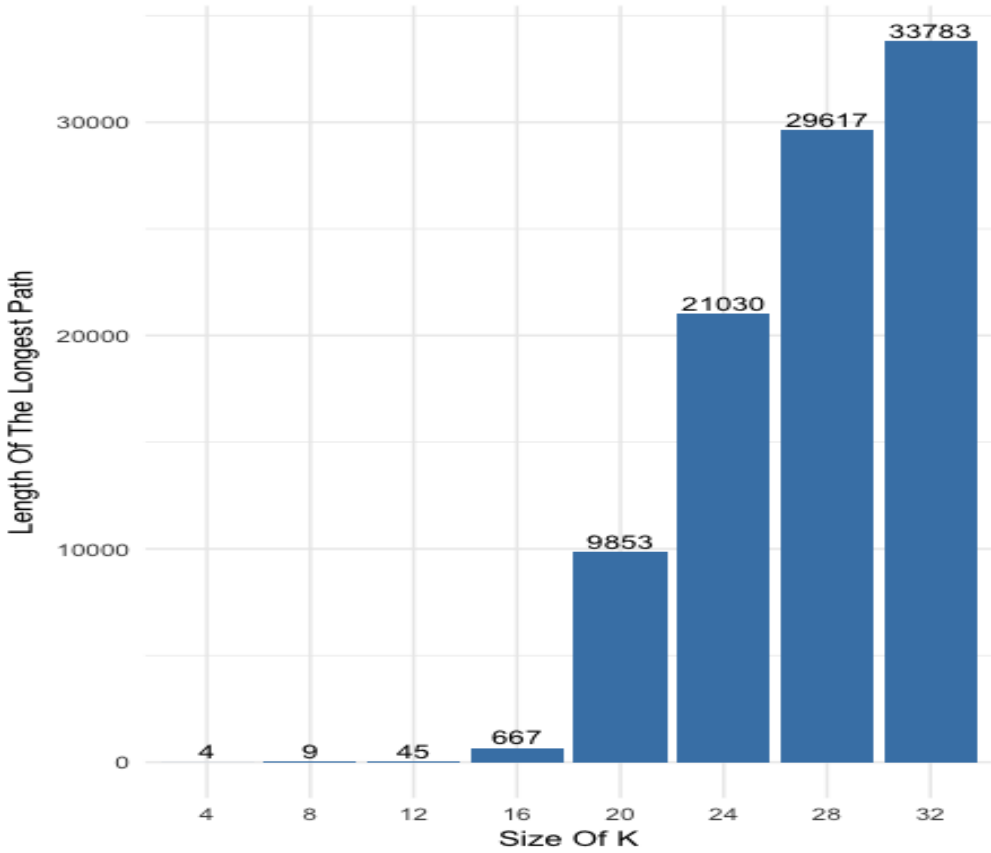
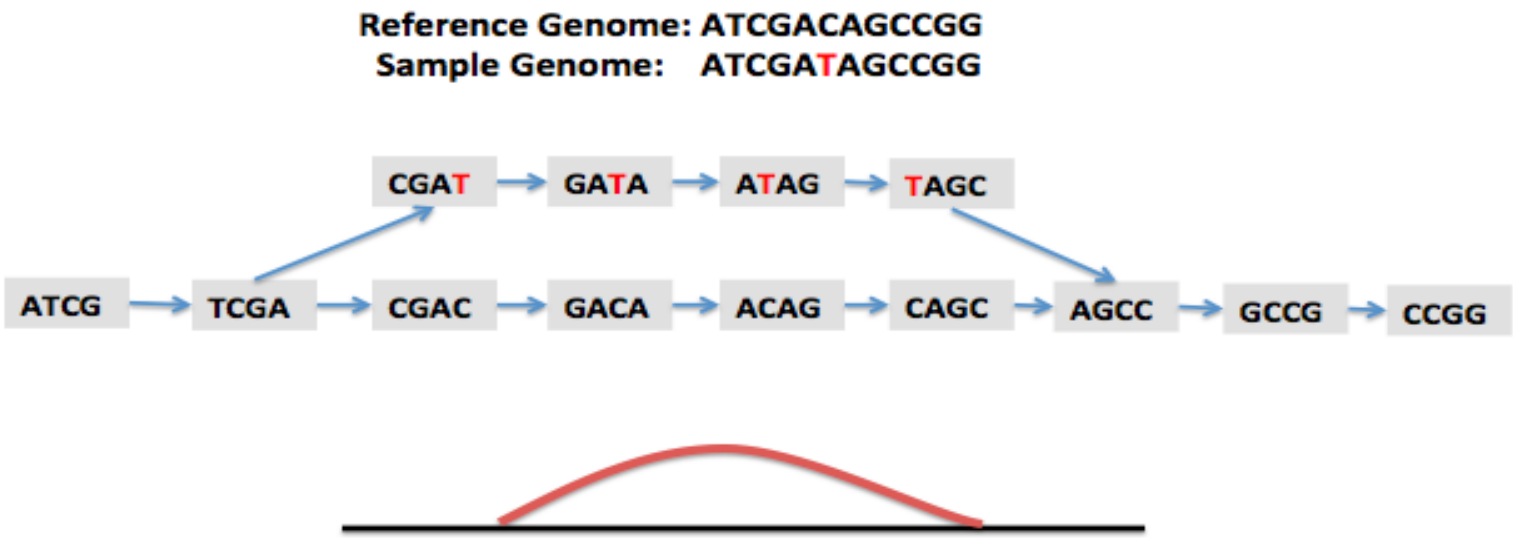


Figure: Comparison of The Lengths Of The Longest Paths

DNA Variation in a De Bruijn Graph

Genomics BDSI

Sijia Huo, Sean Kelly, Gregory Raskind



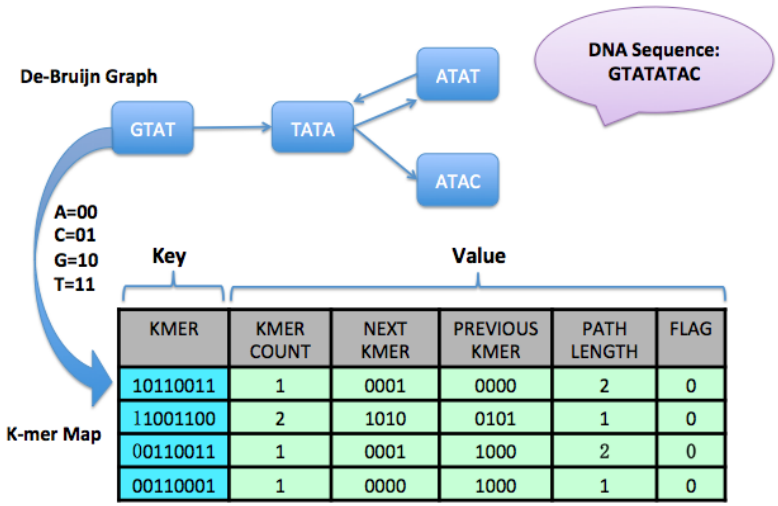
Given an individual's sequenced DNA, we want to locate where their genome varies from the reference

Method: Create De Bruijn graph for sequenced reads to store only variations in genome

Storage Structure

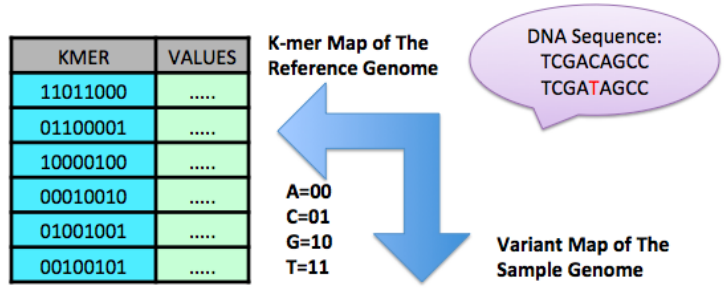
Genomics BDSI

Sijia Huo, Sean Kelly, Gregory Raskind



KMER	KMER COUNT	NEXT KMER	PREVIOUS KMER	PATH LENGTH	FLAG
10110011	1	0001	0000	2	0
11001100	2	1010	0101	1	0
00110011	1	0001	1000	2	0
00110001	1	0000	1000	1	0

(a) Kmer Graph Structure



KMER	VALUES
11011000	.....
01100001	.....
10000100	.....
00010010	.....
01001001	.....
00100101	.....

KMER	NEXT KMER COUNT OF A	NEXT KMER COUNT OF C	NEXT KMER COUNT OF G	NEXT KMER COUNT OF T
01100011	1	0	0	0
10001100	0	0	1	0
00110010	0	1	0	0
11001001	0	1	0	0

(b) Variant Graph Structure

Figure: Data Structures Of The Graphs

Implemented in C++

Future Plans

Cleaning/Pruning Variant Graph

- More accurate and usable
- Reduces storage size

FM indexing

- Allows compression of input text while still permitting fast substring queries

References

For Images:

- Iqbal, Zamin, Mario Caccamo, Isaac Turner, Paul Flicek, and McVean Gil. "De Novo Assembly and Genotyping of Variants Using Colored De Bruijn Graphs." Nature Genetics (2012): Nature Publishing Group. Web.

For information on sequencing and assembly methods

- Paolo Ferragina and Giovanni Manzini (2000). "Opportunistic Data Structures with Applications". Proceedings of the 41st Annual Symposium on Foundations of Computer Science. p.390.
- `http://genome.cshlp.org/content/18/5/821.long#sec-17`
- `http://cortexassembler.sourceforge.net/index_cortex_var.html`
- Umich Biostats lectures

Acknowledgments

University of Michigan and Bhramar Mukherjee
Hyun Min Kang, Goncalo Abecasis, Greg Zajac
NSF