# MinHash for overlapping and assembly

Sergey Koren

SMRT® Informatics Developers Conference
Gaithersburg, MD

August 26, 2015

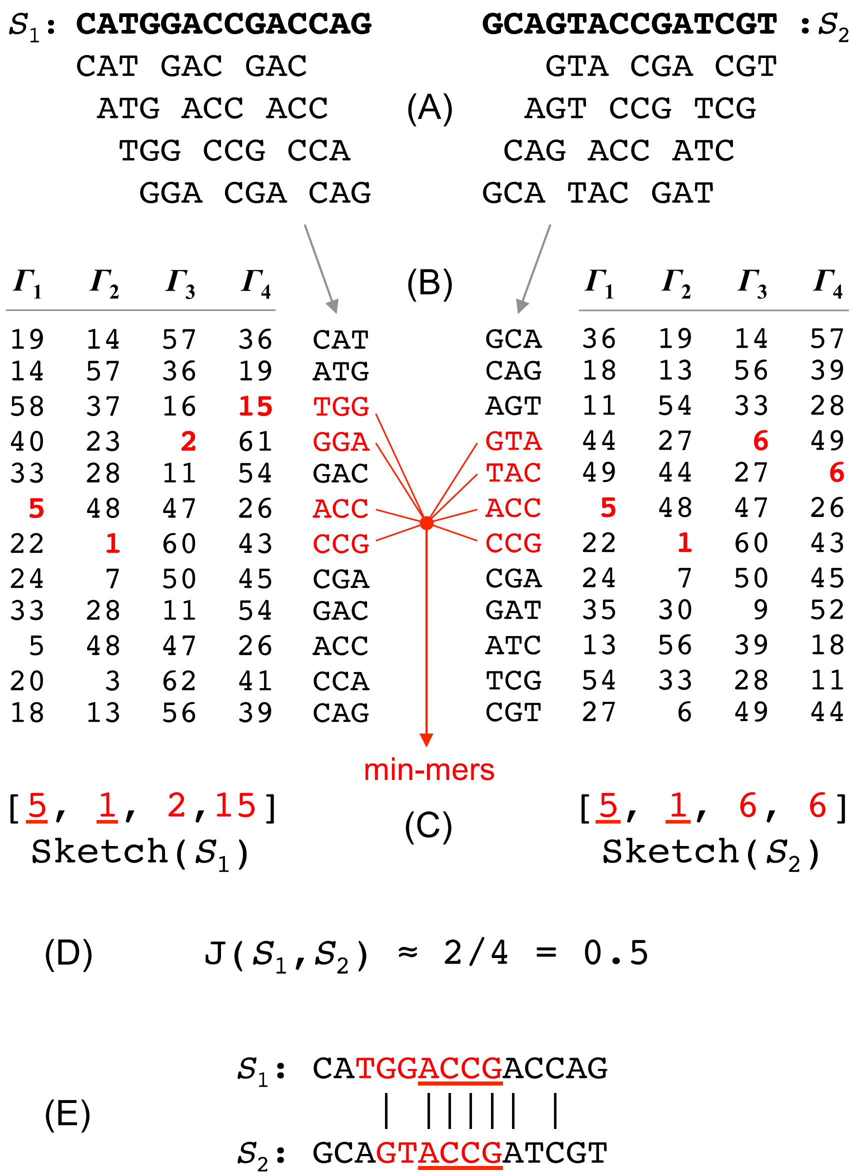
NBACC™
National Biodefense Analysis & Countermeasures Center

The NBACC logo is the intellectual property of DHS.

Homeland Security
Science and Technology

# Acknowledgement

This work was funded under Agreement No. HSHQDC-07-C-00020 awarded to Battelle National Biodefense Institute by the Department of Homeland Security (DHS) Science and Technology Directorate (S&T) for the management and operation of the National Biodefense Analysis and Countermeasures Center a Federally Funded Research and Development Center. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the U.S. Department of Homeland Security or the U.S. Government. The Department of Homeland Security does not endorse any products or commercial services mentioned in this presentation.



$S_1$: **CATGGACCGACCAG** GCAGTACCGATCGT : $S_2$

(A)

```
CAT GAC GAC          GTA CGA CGT
 ATG ACC ACC        AGT CCG TCG
  TGG CCG CCA      CAG ACC ATC
   GGA CGA CAG    GCA TAC GAT
```

(B)

| $\Gamma_1$ | $\Gamma_2$ | $\Gamma_3$ | $\Gamma_4$ | |
|---|---|---|---|---|
| 19 | 14 | 57 | 36 | CAT |
| 14 | 57 | 36 | 19 | ATG |
| 58 | 37 | 16 | **15** | TGG |
| 40 | 23 | **2** | 61 | GGA |
| 33 | 28 | 11 | 54 | GAC |
| **5** | 48 | 47 | 26 | ACC |
| 22 | **1** | 60 | 43 | CCG |
| 24 | 7 | 50 | 45 | CGA |
| 33 | 28 | 11 | 54 | GAC |
| 5 | 48 | 47 | 26 | ACC |
| 20 | 3 | 62 | 41 | CCA |
| 18 | 13 | 56 | 39 | CAG |

| | $\Gamma_1$ | $\Gamma_2$ | $\Gamma_3$ | $\Gamma_4$ |
|---|---|---|---|---|
| GCA | 36 | 19 | 14 | 57 |
| CAG | 18 | 13 | 56 | 39 |
| AGT | 11 | 54 | 33 | 28 |
| GTA | 44 | 27 | **6** | 49 |
| TAC | 49 | 44 | 27 | **6** |
| ACC | **5** | 48 | 47 | 26 |
| CCG | 22 | **1** | 60 | 43 |
| CGA | 24 | 7 | 50 | 45 |
| GAT | 35 | 30 | 9 | 52 |
| ATC | 13 | 56 | 39 | 18 |
| TCG | 54 | 33 | 28 | 11 |
| CGT | 27 | 6 | 49 | 44 |

min-mers

(C)

[5, 1, 2,15] Sketch($S_1$)

[5, 1, 6, 6] Sketch($S_2$)

(D) J($S_1$,$S_2$) ≈ 2/4 = 0.5

(E)

```
S1: CATGGACCGACCAG
       |  ||||| |
S2: GCAGTACCGATCGT
```

## ▪ The “AltaVista” algorithm

Invented in 1998 by Andrei Broder to detect duplicate web pages

Applied to DNA sequencing matching and alignment

– Quick estimator of Jaccard Similarity

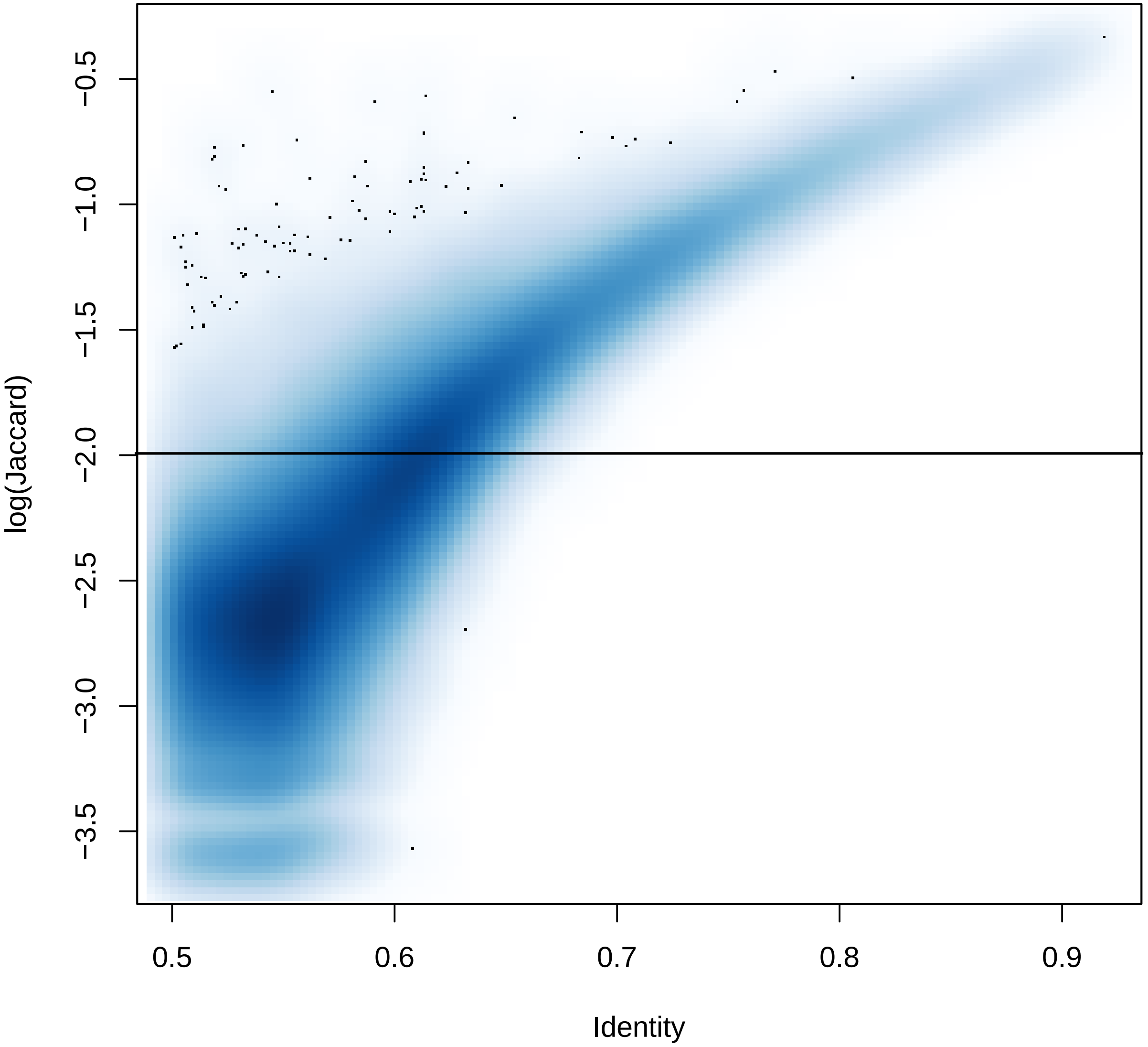
$$J(A,B)=\frac{|A \text{ Ç } B|}{|A \text{ È } B|}=\frac{4}{18}=0.22$$

- Position independent
- Length independent
- Correlated with identity

**Assembling Large Genomes with Single-Molecule Sequencing and Locality Sensitive Hashing**

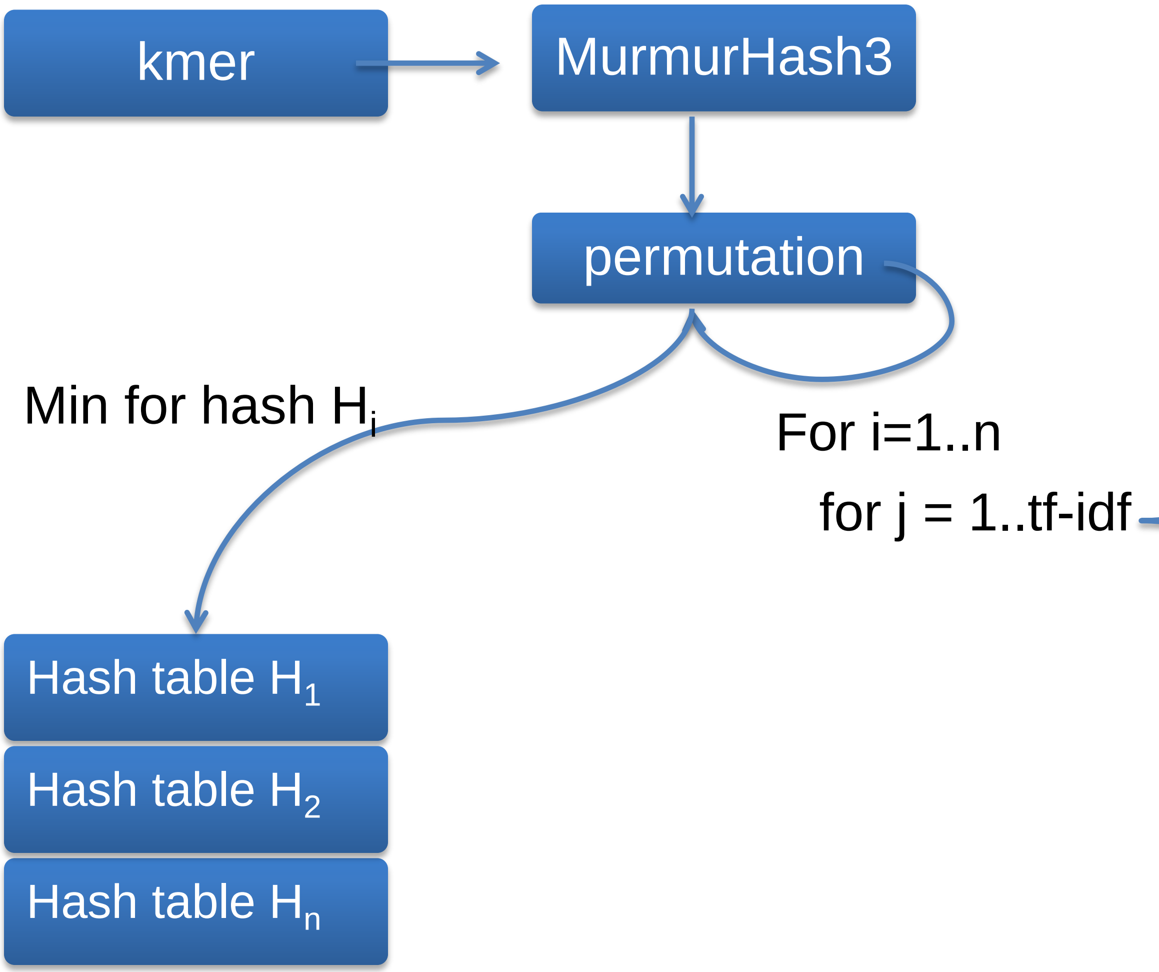
Berlin *et al.* (2015) *Nature Biotechnology*

# Jaccard score estimates identity



J=0.01

# In Practice, n hashes



kmer

MurmurHash3

permutation

Min for hash $H_i$

For i=1..n

for j = 1..tf-idf

Hash table $H_1$

Hash table $H_2$

Hash table $H_n$

## Tf-Idf Weight

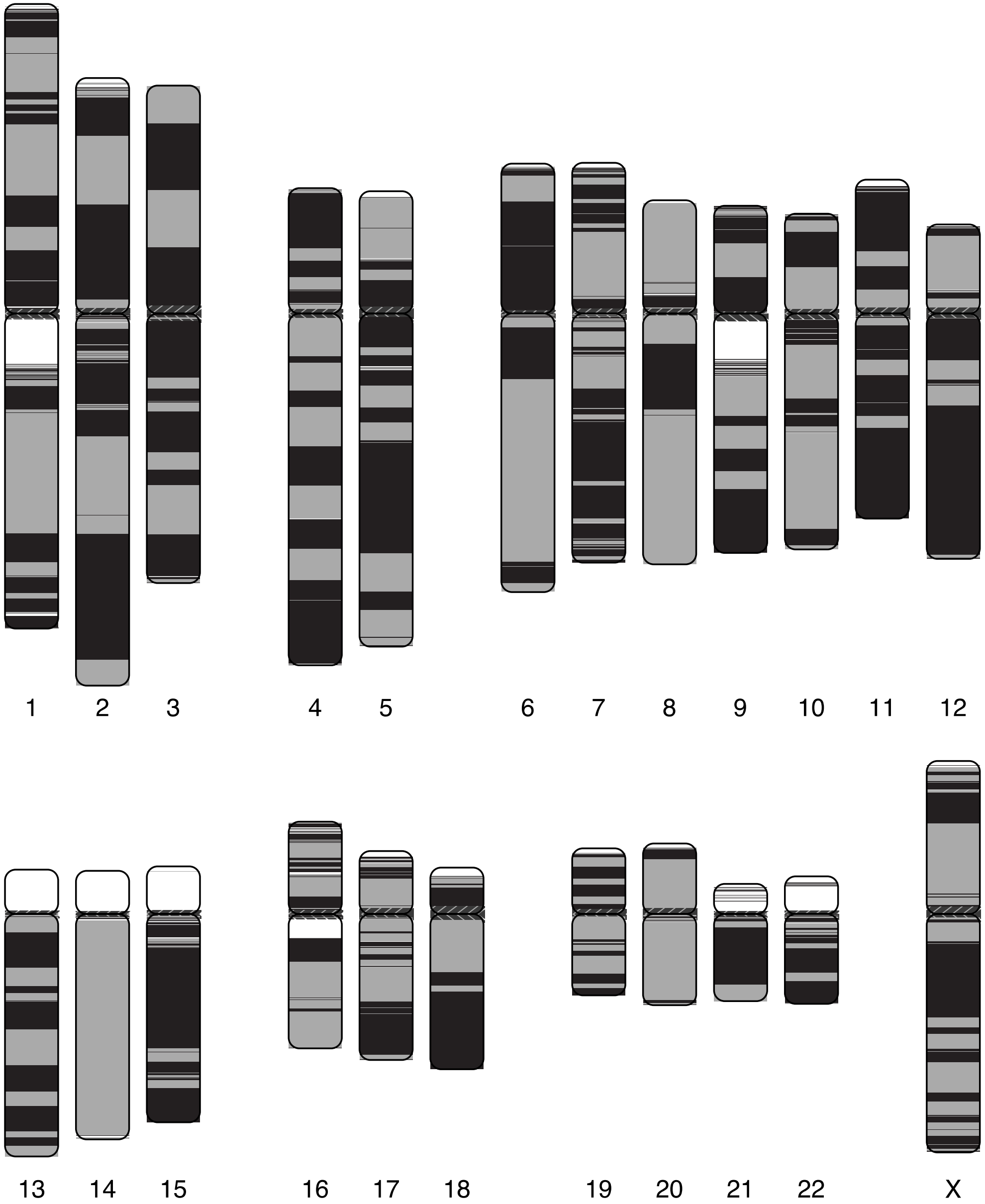
$$tf(kmer,r) = \#\ times\ kmer\ occurs\ in\ read\ r$$

$$f(kmer) = \sum_{r\ \in reads} tf(kmer,r)$$

$$idf(kmer) = \log\left(\frac{\max\,(f(kmer)\forall\ kmers}{f(kmer)}\right)$$

$$weight \propto tfidf = tf\ (kmer,read) *\ idf(kmer)$$

# Human Assembly, solved?



CHM13 CA 8.3

# Acknowledgements

- **MHAP & Canu**
  - Adam Phillippy
  - Konstantin Berlin
  - Brian Walenz
- **Parsnp & Gingr**
  - Todd Treangen
  - Brian Ondov
- **Mash**
  - Brian Ondov

GitHub
/MarBL

Or just Google “PBcR MHAP”

- **Join Phillippy Lab/MarBL at NIH**
  - Looking for two postdocs, talk to myself or Adam

NBACC™

National Biodefense Analysis & Countermeasures Center

*The NBACC logo is the intellectual property of DHS.*

This Document was prepared for the Department of Homeland Security (DHS) by the Battelle National Biodefense Institute, LLC (BNBI) as part of contract HSHQDC-07-C-00020 to manage and operate the National Biodefense Analysis and Countermeasures Center (NBACC), a Federally Funded Research and Development Center. In no event shall the DHS, BNBI or NBACC have any responsibility or liability for any use, misuse, inability to use, or reliance upon the information contained herein. **In addition, no warranty of fitness for a particular purpose, merchantability, accuracy or adequacy is provided regarding the contents of this document.**

Homeland Security

Science and Technology

# Human Assemblies with MHAP

| Genome | Chem | Cov | #Ctgs | Max (kb) | N50(kb) | Ovl CPU(h) |
|---|---|---:|---:|---:|---:|---:|
| CHM1 | P5 | 54X | 17,776 | 35,487 | 6,303 | 19,700 |
| CHM1 | P5+P6 | 120X | 8,011 | 143,469 | 23,254 | 26,305 |
| CHM13 | P5+P6 | 70X | 15,538 | 81,523 | 13,332 | 8,171 |
| | | | | | | |
| Trio HG002 | P5+P6 | 71X | 13,048 | 35,012 | 4,399 | 9,145 |
| Trio HG003 | P5+P6 | 33X | 23,493 | 9,815 | 912 | 30,625* |
| Trio HG004 | P5+P6 | 29X | 16,326 | 8,894 | 1,034 | 22,971* |

* Low coverage datasets run with sensitive parameters to improve assembly