Analysis of Next Generation Sequence Data

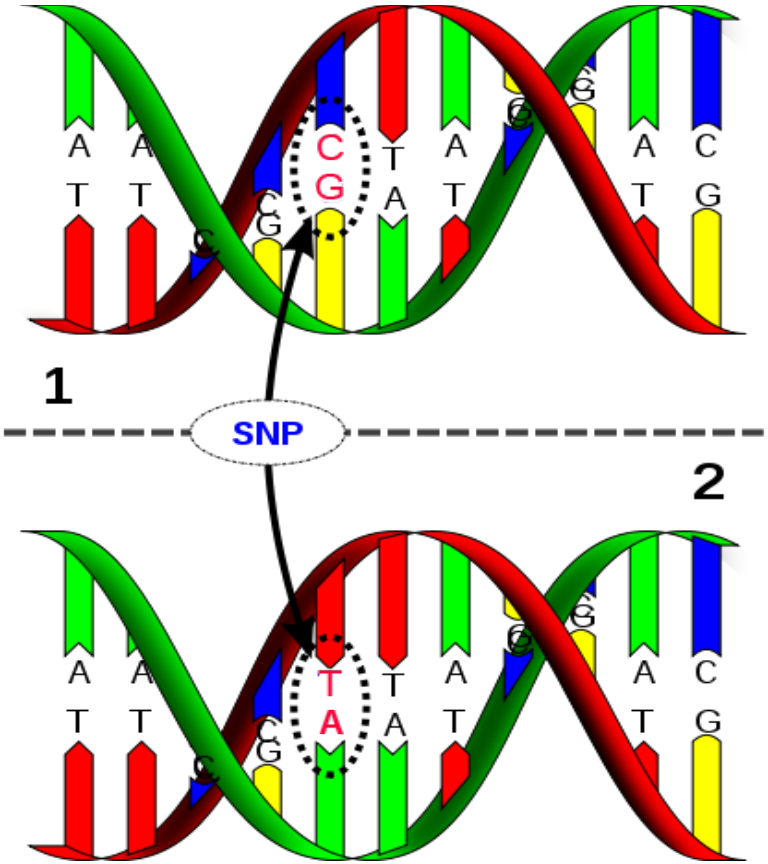
BIOST 2055

03/16/2012

Outline

- Background
- From sequence data to genotype
  - Alignment
  - SNP detection and genotype calling
  - Genotype refinement
- A walk-through example
  - Get familiar with various file formats
  - Get familiar with popular programs

Human Genome and Single Nucleotide Polymorphisms (SNPs)



- 23 chromosome pairs
- 3 billion bases

- A single nucleotide change between pairs of chromosomes

- E.g.

Haplotype1: AAGG**G**ATCCAC
Haplotype2: AAGG**A**ATCCAC

SNPs in Population

Association Study in Case Control Samples

SNP1 SNP2 SNP3 SNP4 SNP5 Disease

CAGATCGCTGGATGAATCGCATC
CGGATTGCTGCATGGATCGCATC

CAGATCGCTGGATGAATCGCATC
CAGATCGCTGGATGAATCCCATC

CGGATTGCTGCATGGATCCCATC
CGGATTGCTGCATGGATCCCATC

History of DNA Sequencing

1950
- DNA structure reported by Watson and Crick (1953)

1960
- DNA genetic code deduced for 20+ codons and their respective amino acids and reported by M. Niremberg, H. Mathaei and S. Ochoa (1966)

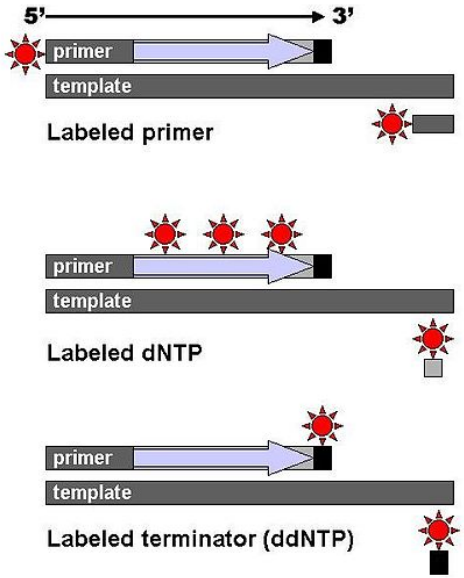
1970
- First recombinant DNA molecules were created (P. Berg 1972)
- Di-deoxy sequencing reported by F. Sanger (1977)

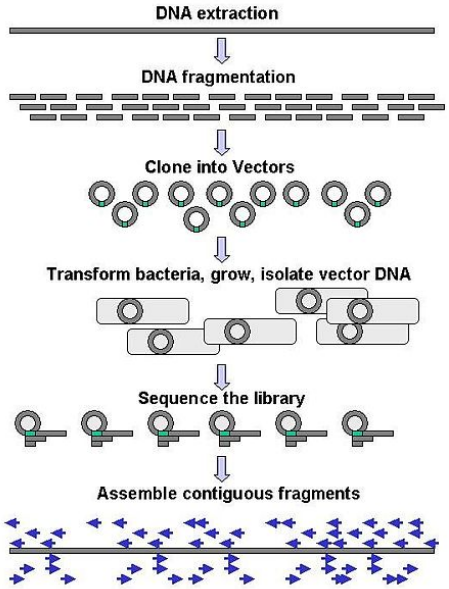
1980
- Polymerase chain reaction reported (Mullis 1983)

1990
- Human genome project initiated in public section (1990)
- First completed genome sequence in public sector for *M. genitalium* (Fraser et al. 1995)
- Human genome project initiated in private sector (Venter 1998)

2000
- Initial drafts of human genome reported simultaneously in private and public sectors (2001)
- Finished drafts of euchromatic sequence of human genome reported in public sector (2004)
- Rough draft (5x coverage) chimpanzee reported (2005)

2010
- Massively parallel sequencing technologies now being widely utilized about 2007 (technology made publicly available in 2004)

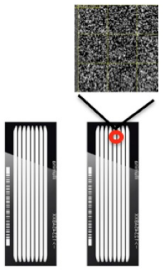




Illumina HiSeq2000

- 2 flow cells (can be run independently)
- Up to 320Gb mapped sequence per FC
- 64Gb sequence per day (2 flow cells)

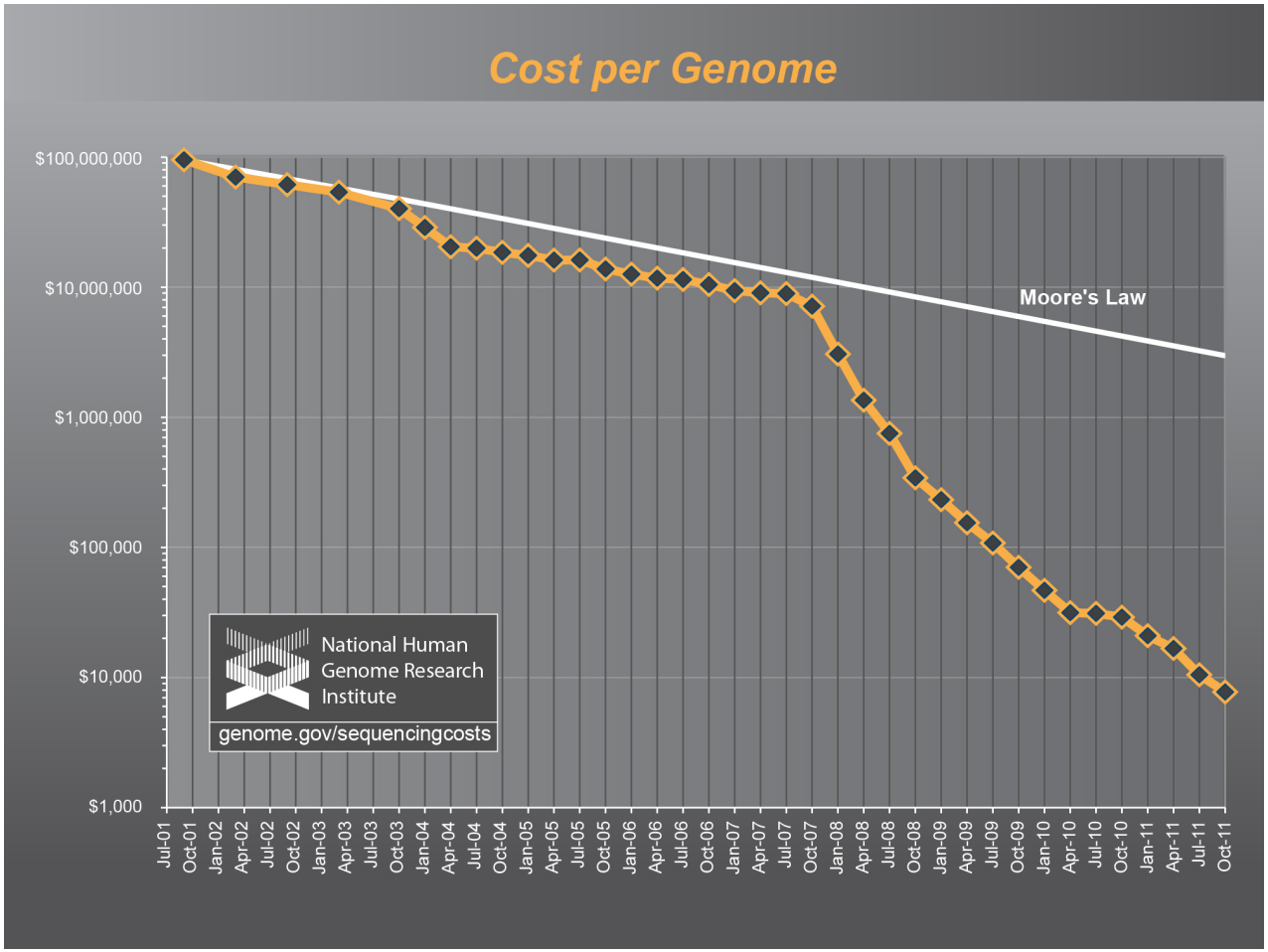
c-Bot
(automated cluster generation)



Cluster density
750-850/mm^2

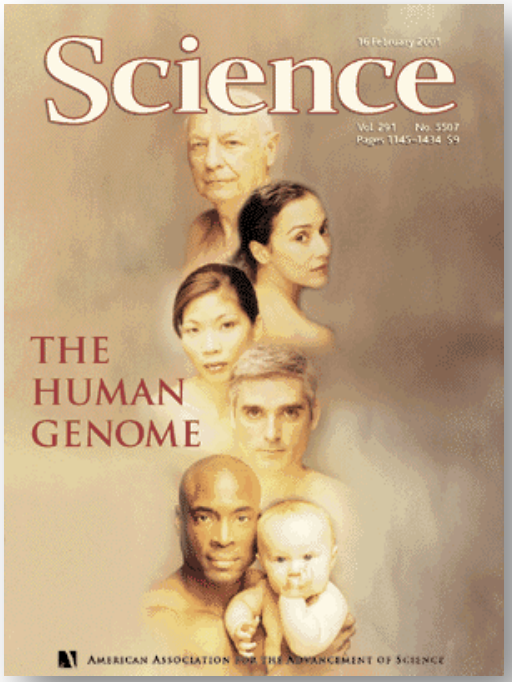
HiSeq Flow Cells

Sequencing Cost



http://www.genome.gov/

A Road to Discover Human Genome



1990-2003

2002 -

2008 -

Different Approaches

- Deep whole genome sequencing
  - Expensive, only can be applied to limited samples currently
  - Most complete ascertainment of all variations
- Low coverage whole genome sequencing
  - Modest cost, typically 100-1000 samples
  - Complete ascertainment of common variations
  - Less compete ascertainment of rare variants
- Exome capture and targeted region sequencing
  - Modest cost, high coverage
  - Most interesting part of the genome

With Complete Sequence Data

- What is the contribution of each identified locus to a trait?
  - Multiple variants, common and rare
  - Effect size

- What is the mechanism? What happens when we knockout a gene?
  - Most often, the causal variant will not have been examined directly
  - Rare coding variants will provide important insights into mechanisms

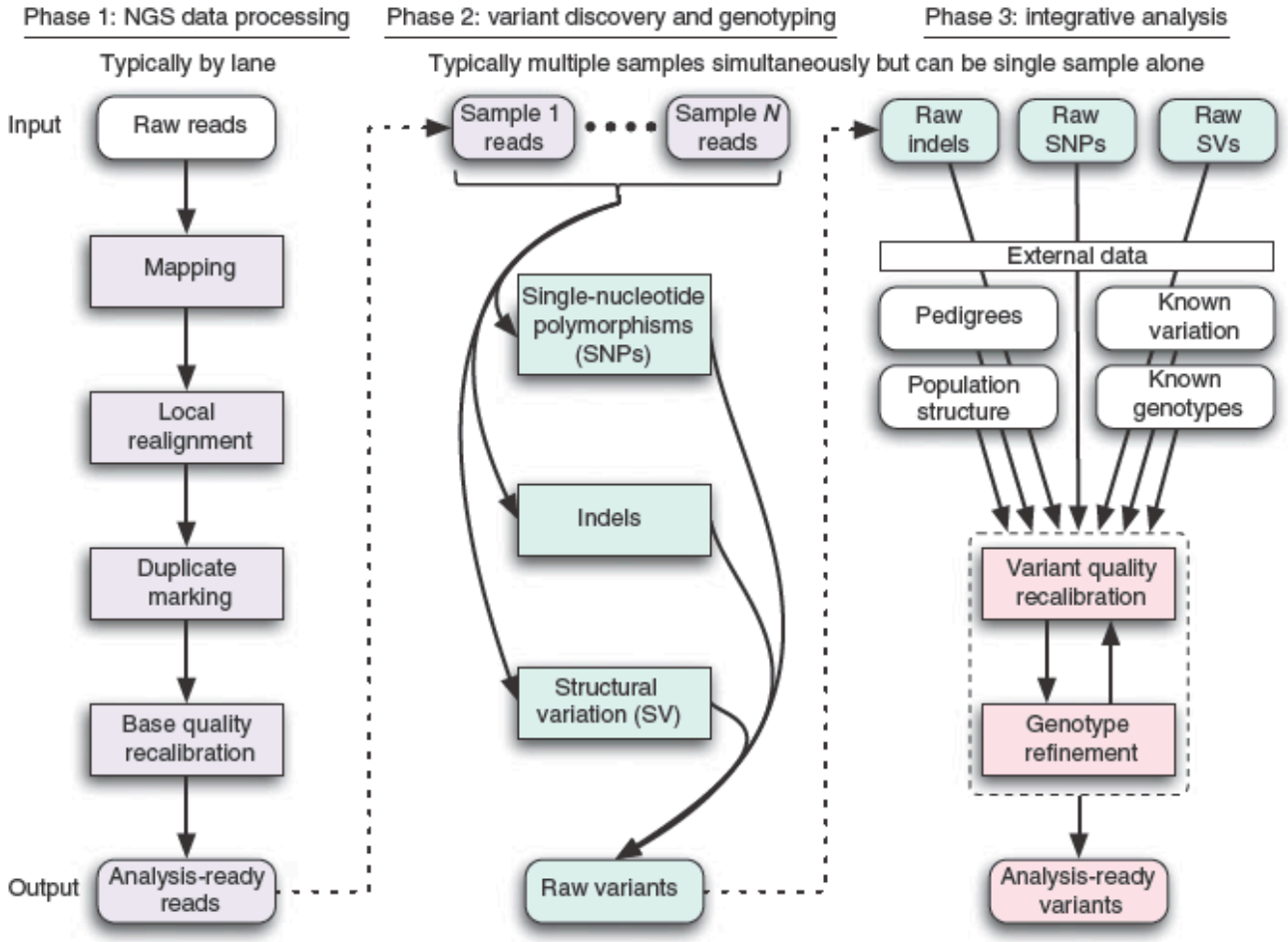
- What is the contribution of structural variation to disease?
  - These are hard to interrogate using current genotyping arrays

- Are there additional susceptibility loci to be found?
  - Only subset of functional elements include common variants
  - Rare variants are more numerous and thus will point to additional loci

Next Generation Sequencing

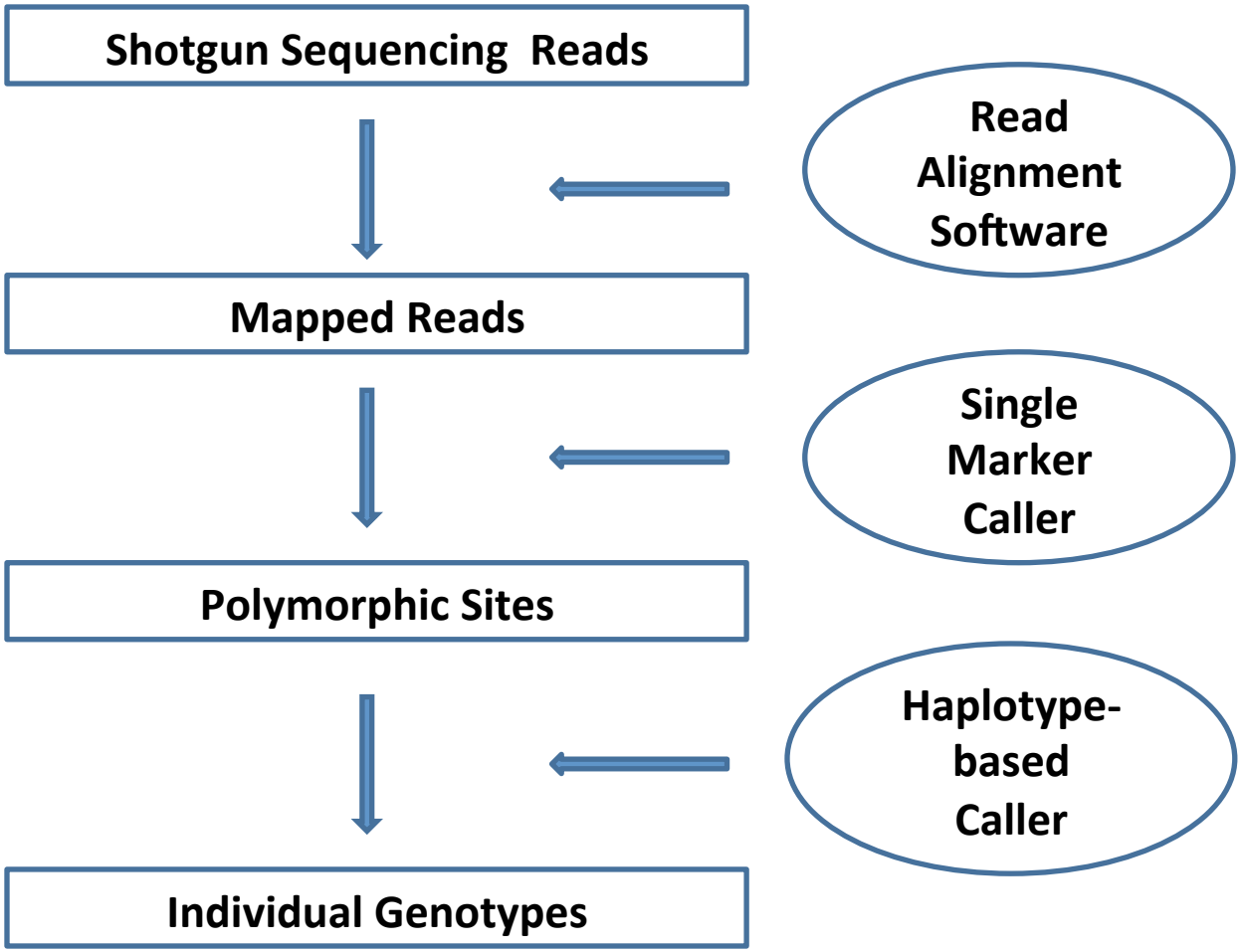
- Commercial platforms produce gigabases of sequence rapidly and inexpensively
  - ABI SOLiD, Illumina Solexa, Roche 454, Complete Genomics, and others…
- Sequence data consist of thousands or millions of short sequence reads with moderate accuracy 0.5 – 1.0% error rates per base may be typical
- High-throughput but hard to assemble

A Typical Pipeline



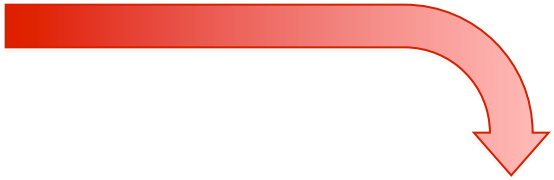
Monya Baker Nat Meth 2011

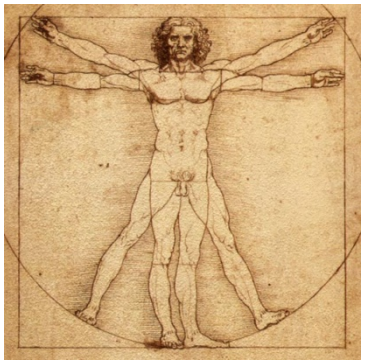
A Typical Pipeline



Short read alignment

Sequencer



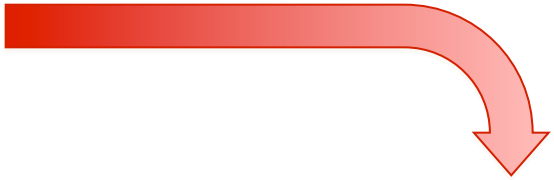


Human source

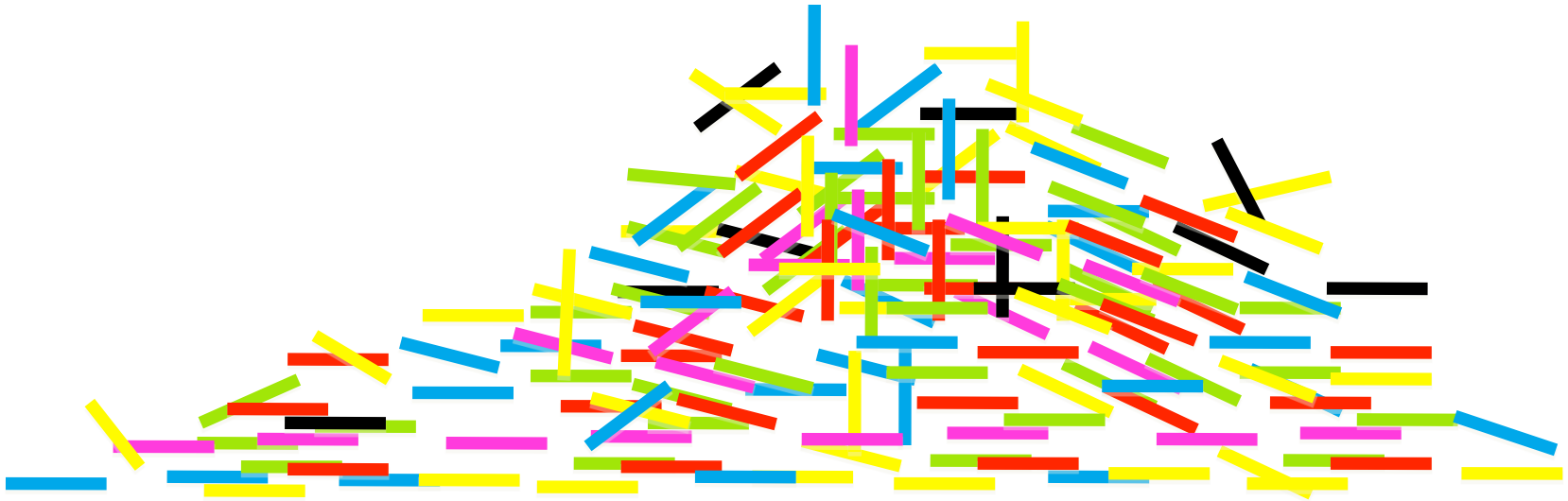
Reads from new sequencing machines are short: 30-400 bp

Short read alignment

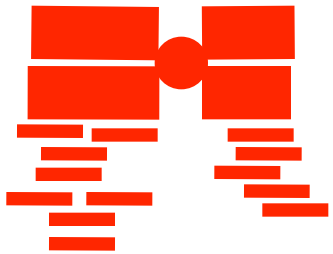
Sequencing machine

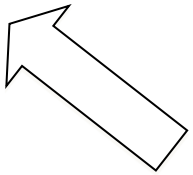


And you get
MILLIONS of them

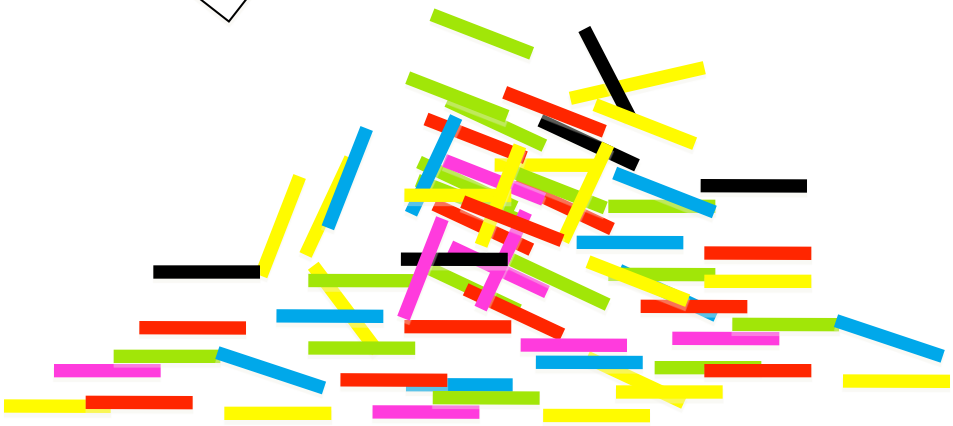


Short read alignment





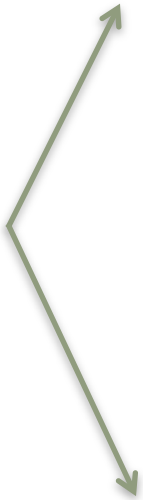
Need to map them back to human reference



Alignment

Reference sequence:

actgtagattag**ccgagtagctag**ctagtcgat



Find best match for each read in a reference sequence

ccgagaagctag

Which is better? Long words or Short words?

Existing Alignment by Category

- Hashing reference genome
  - SOAP1, MOSAIK, PASS, BFAST, …
- Hashing short reads
  - Eland, MAQ, SHRiMP, …
- Merge-sorting reference together with reads
  - Slider
- Based on Burrows-Wheeler Transform
  - BWA, SOAP2, Bowtie, …

Li and Durbin (2009), *Bioinformatics* 25 (14): 1754-60

A list

Sean Penn
Daniel Day Lewis
Forest Whitaker
Philip S. Hoffman
Jamie Foxx
Sean Penn
Adrien Brody
Denzel Washington
Russell Crowe
Kevin Spacey
Roberto Benigni
Jack Nicholson
Geoffrey Rush
Nicolas Cage
Tom Hanks
Tom Hanks
Al Pacino
Anthony Hopkins
Jeremy Irons
Daniel Day Lewis
Dustin Hoffman

Michael Douglas
Paul Newman
William Hurt
F. Murray Abraham
Robert Duvall
Ben Kingsley
Henry Fonda
Robert De Niro
Dustin Hoffman
Jon Voight
Richard Dreyfuss
Peter Finch
Jack Nicholson
Art Carney
Jack Lemmon
Marlon Brando
Gene Hackman
George C. Scott
John Wayne
Cliff Robertson
Rod Steiger

Paul Scofield
Lee Marvin
Rex Harrison
Sidney Poitier
Gregory Peck
Maximilian Schell
Burt Lancaster
Charlton Heston
David Niven
Alec Guinness
Yul Brynner
Ernest Borgnine
Marlon Brando
William Holden
Gary Cooper
Humphrey Bogart
José Ferrer
Broderick Crawford
Laurence Olivier
Ronald Colman
Fredric March

Ray Milland
Bing Crosby
Paul Lukas
James Cagney
Gary Cooper
James Stewart
Robert Donat
Spencer Tracy
Spencer Tracy
Paul Muni
Victor McLaglen
Clark Gable
Charles Laughton
Wallace Beery
Fredric March
Lionel Barrymore
George Arliss
Warner Baxter
Emil Jannings
Emil Jannings

An Ordered List

Adrien Brody
Al Pacino
Alec Guinness
Anthony Hopkins
Art Carney
Ben Kingsley
Bing Crosby
Broderick Crawford
Burt Lancaster
Charles Laughton
Charlton Heston
Clark Gable
Cliff Robertson
Daniel Day Lewis (2)
David Niven
Denzel Washington
Dustin Hoffman (2)
Emil Jannings (2)
Ernest Borgnine
F. Murray Abraham
Forest Whitaker
Fredric March (2)
Gary Cooper (2)
Gene Hackman
Geoffrey Rush
George Arliss
George C. Scott
Gregory Peck
Henry Fonda
Humphrey Bogart
Jack Lemmon
Jack Nicholson (2)
James Cagney
James Stewart
Jamie Foxx
Jeremy Irons
John Wayne
Jon Voight
José Ferrer
Kevin Spacey
Laurence Olivier
Lee Marvin
Lionel Barrymore
Marlon Brando (2)
Maximilian Schell
Michael Douglas
Nicolas Cage
Paul Lukas
Paul Muni
Paul Newman
Paul Scofield
Peter Finch
Philip S. Hoffman
Ray Milland
Rex Harrison
Richard Dreyfuss
Robert De Niro
Robert Donat
Robert Duvall
Roberto Benigni
Rod Steiger
Ronald Colman
Russell Crowe
Sean Penn (2)
Sidney Poitier
Spencer Tracy (2)
Tom Hanks (2)
Victor McLaglen
Wallace Beery
Warner Baxter
William Holden
William Hurt
Yul Brynner

Hash Table

- Function(key) = value

- F(name) =

  index(last letter of the first name)*26 +
  index(first letter of the last name)

- F(John Wayne) = F(n,w) = 13*26+22 = 360

A Hash Table

58: Alec Guinness
64: Fredric March (2)
80: Ronald Colman
81: Richard Dreyfuss
91: David Niven
96: Rod Steiger
104: George Arliss
105: Wallace Beery
109: José Ferrer
109: Jamie Foxx
111: Gene Hackman
116: Lee Marvin
118: Laurence Olivier
122: George C. Scott
130: F. Murray Abraham
147: Cliff Robertson
158: Bing Crosby
262: Broderick Crawford
266: Clark Gable
271: Jack Lemmon
273: Jack Nicholson (2)
287: Lionel Barrymore
287: Yul Brynner
288: Russell Crowe
289: Michael Douglas
295: Emil Jannings (2)
297: Daniel Day Lewis (2)
297: Paul Lukas
298: Paul Muni
299: Paul Newman
301: Al Pacino
304: Paul Scofield
308: Denzel Washington
319: William Holden
319: Tom Hanks (2)
319: William Hurt
339: Marlon Brando (2)
339: Adrien Brody
345: Dustin Hoffman (2)
345: Charlton Heston
348: Ben Kingsley
353: Sean Penn (2)
356: Maximilian Schell
356: Kevin Spacey
359: Jon Voight
360: John Wayne
365: Roberto Benigni
397: Philip S. Hoffman
443: Warner Baxter
447: Peter Finch
454: Victor McLaglen
461: Spencer Tracy (2)
470: Nicolas Cage
470: James Cagney
479: Charles Laughton
486: James Stewart
495: Ernest Borgnine
496: Art Carney
497: Robert Duvall
497: Robert De Niro
497: Robert Donat
505: Burt Lancaster
516: Forest Whitaker
605: Rex Harrison
625: Humphrey Bogart
626: Gary Cooper (2)
629: Henry Fonda
631: Anthony Hopkins
632: Jeremy Irons
636: Ray Milland
639: Gregory Peck
639: Sidney Poitier
641: Geoffrey Rush

Search In A Hash Table

58: Alec Guinness
64: Fredric March (2)
80: Ronald Colman
81: Richard Dreyfuss
91: David Niven
96: Rod Steiger
104: George Arliss
105: Wallace Beery
109: José Ferrer
109: Jamie Foxx
111: Gene Hackman
116: Lee Marvin
118: Laurence Olivier
122: George C. Scott
130: F. Murray Abraham
147: Cliff Robertson
158: Bing Crosby
262: Broderick Crawford
266: Clark Gable
271: Jack Lemmon
273: Jack Nicholson (2)
287: Lionel Barrymore
287: Yul Brynner
288: Russell Crowe
289: Michael Douglas
295: Emil Jannings (2)
297: Daniel Day Lewis (2)
297: Paul Lukas
298: Paul Muni
299: Paul Newman
301: Al Pacino
304: Paul Scofield
308: Denzel Washington
319: William Holden
319: Tom Hanks (2)
319: William Hurt
339: Marlon Brando (2)
339: Adrien Brody
345: Dustin Hoffman (2)
345: Charlton Heston
348: Ben Kingsley
353: Sean Penn (2)
356: Maximilian Schell
356: Kevin Spacey
359: Jon Voight
360: John Wayne
365: Roberto Benigni
397: Philip S. Hoffman
443: Warner Baxter
447: Peter Finch
454: Victor McLaglen
461: Spencer Tracy (2)
470: Nicolas Cage
470: James Cagney
479: Charles Laughton
486: James Stewart
495: Ernest Borgnine
496: Art Carney
497: Robert Duvall
497: Robert De Niro
497: Robert Donat
505: Burt Lancaster
516: Forest Whitaker
605: Rex Harrison
625: Humphrey Bogart
626: Gary Cooper (2)
629: Henry Fonda
631: Anthony Hopkins
632: Jeremy Irons
636: Ray Milland
639: Gregory Peck
639: Sidney Poitier
641: Geoffrey Rush

John Wayne = (n,w) = 13*26+22 = 360 = FOUND

Search In A Hash Table

58: Alec Guinness
64: Fredric March (2)
80: Ronald Colman
81: Richard Dreyfuss
91: David Niven
96: Rod Steiger
104: George Arliss
105: Wallace Beery
109: José Ferrer
109: Jamie Foxx
111: Gene Hackman
116: Lee Marvin
118: Laurence Olivier
122: George C. Scott
130: F. Murray Abraham
147: Cliff Robertson
158: Bing Crosby
262: Broderick Crawford
266: Clark Gable
271: Jack Lemmon
273: Jack Nicholson (2)
287: Lionel Barrymore
287: Yul Brynner
288: Russell Crowe
289: Michael Douglas
295: Emil Jannings (2)
297: Daniel Day Lewis (2)
297: Paul Lukas
298: Paul Muni
299: Paul Newman
301: Al Pacino
304: Paul Scofield
308: Denzel Washington
319: William Holden
319: Tom Hanks (2)
319: William Hurt
339: Marlon Brando (2)
339: Adrien Brody
345: Dustin Hoffman (2)
345: Charlton Heston
348: Ben Kingsley
353: Sean Penn (2)
356: Maximilian Schell
356: Kevin Spacey
359: Jon Voight
360: John Wayne
365: Roberto Benigni
397: Philip S. Hoffman
443: Warner Baxter
447: Peter Finch
454: Victor McLaglen
461: Spencer Tracy (2)
470: Nicolas Cage
470: James Cagney
479: Charles Laughton
486: James Stewart
495: Ernest Borgnine
496: Art Carney
497: Robert Duvall
497: Robert De Niro
497: Robert Donat
505: Burt Lancaster
516: Forest Whitaker
605: Rex Harrison
625: Humphrey Bogart
626: Gary Cooper (2)
629: Henry Fonda
631: Anthony Hopkins
632: Jeremy Irons
636: Ray Milland
639: Gregory Peck
639: Sidney Poitier
641: Geoffrey Rush

Adam Sandler= (m,s) = 12*26+18 = 320 = NOT FOUND

Back to Alignment

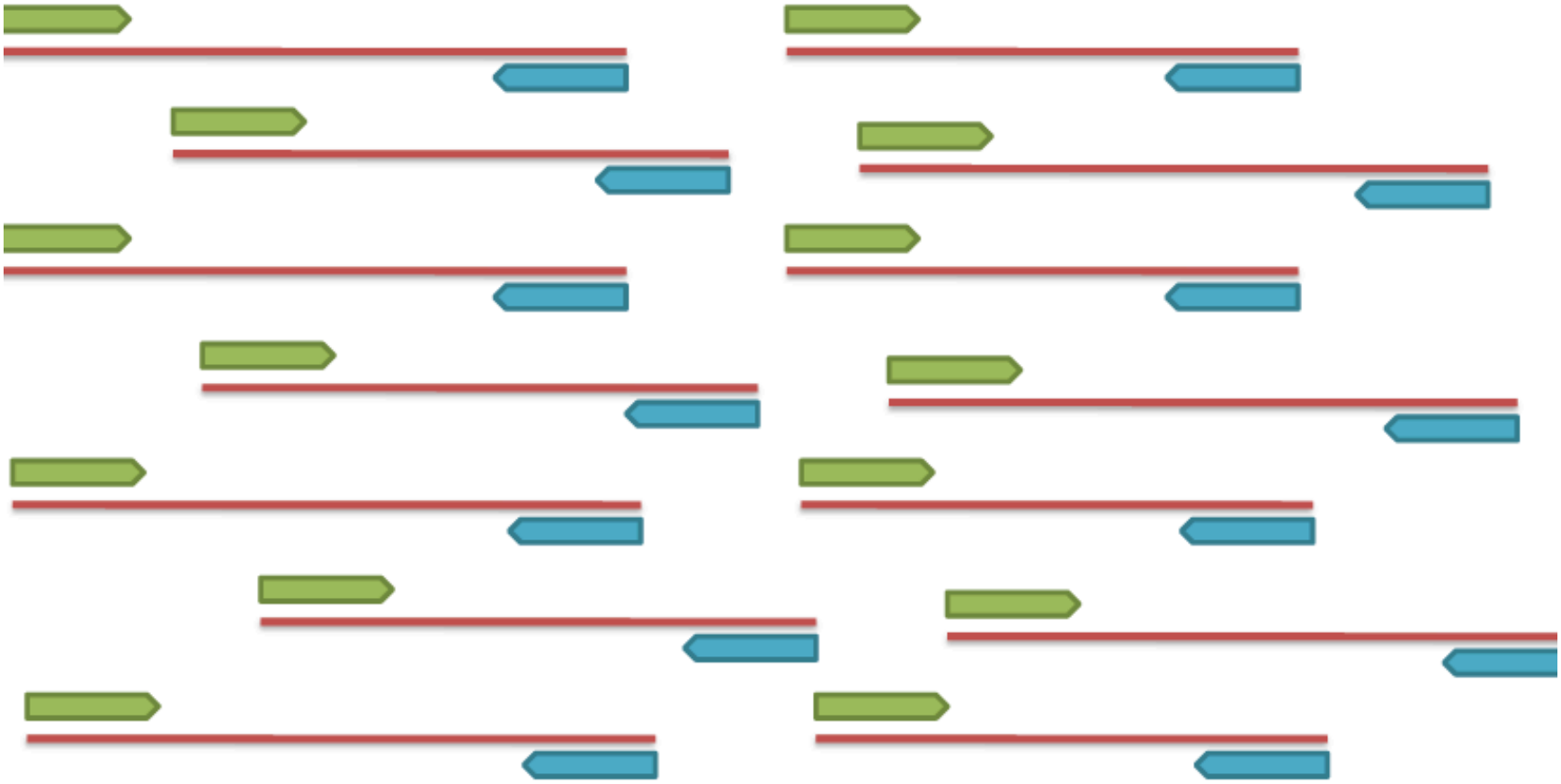
Reference sequence:

actgtagattag**ccgagtagctag**ctagtcgat

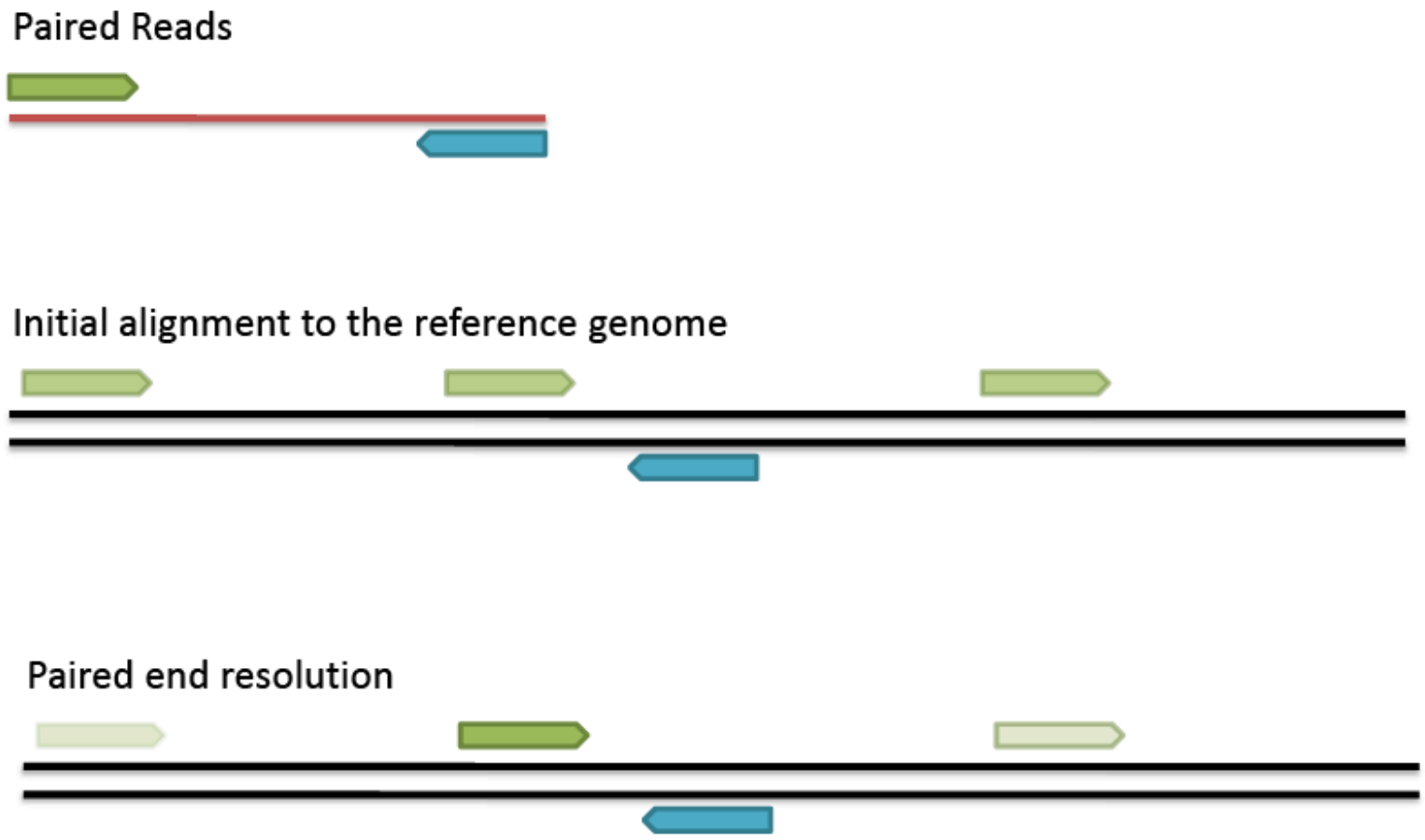
ccgagaagctag

- Hashing is time and memory consuming for millions of reads and billion-base long reference
- Each read may be mapped to multiple positions

Paired End Reads



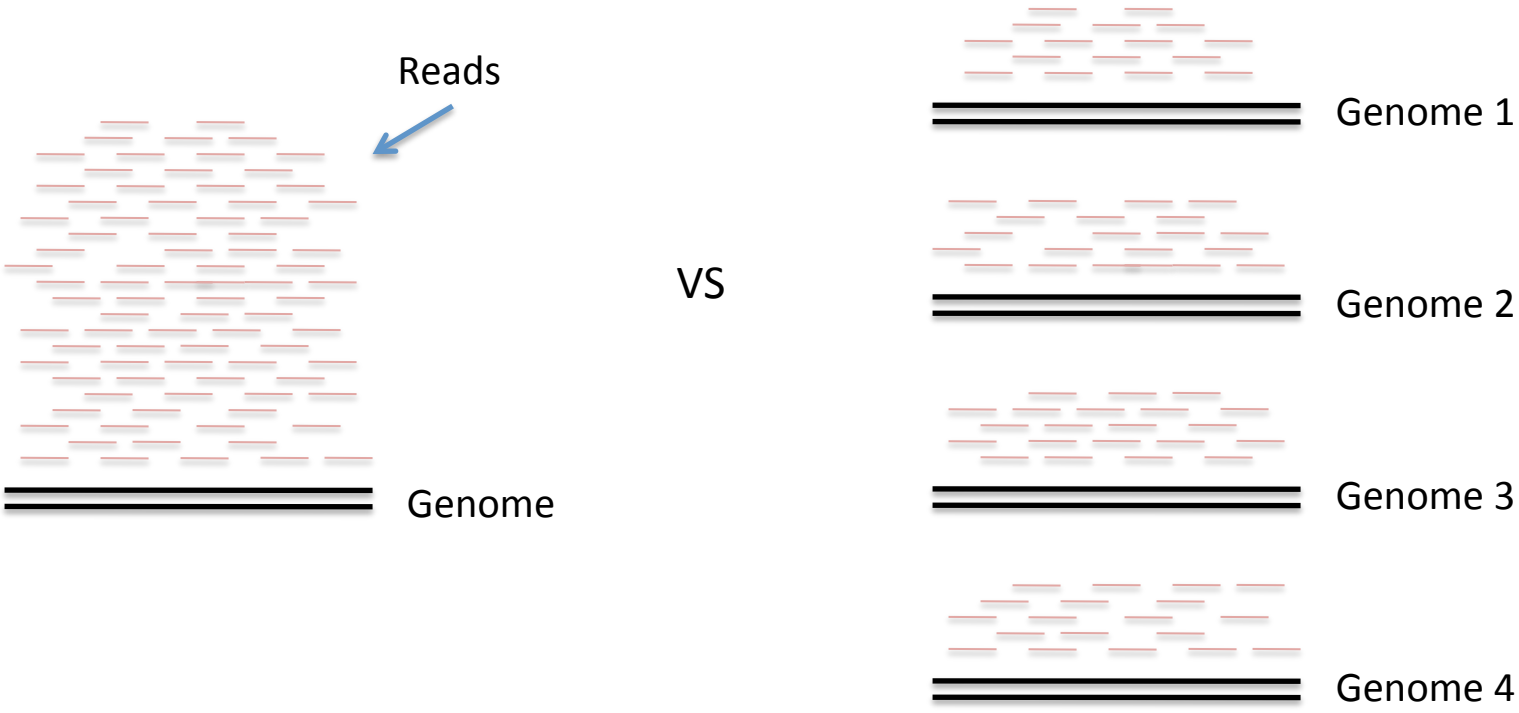
Paired End Reads



After Alignment

- Each read is mapped to reference genome with tolerated number of mismatches
  - Mismatches allow us to discover the individual variation
- Each site of reference genome is covered by multiple un-evenly distributed read
  - Some sites might not be covered

Coverage (High vs Low)



- Which one has more power to detect variations?

Genotype Calling from Sequence Data

TAGCTGATAGCTAG**A**TAGCTGATGAGCCCGAT

ATAGCTAG**A**TAGCTGATGAGCCCGATCGCTGCTAGCTC

ATGCTAGCTGATAGCTAG**C**TAGCTGATGAGCC

AGCTGATAGCTAG**C**TAGCTGATGAGCCCGATCGCTG

GCTAGCTGATAGCTAG**C**TAGCTGATGAGCCCGA

Sequence Reads

5'-ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3'

Reference Genome

A/C

Predicted Genotype

A Simple Model

At one site, Na reads carry A, Nb reads carry B

$$N_A \sim \begin{cases} Binomial(n_A + n_B, 1-\delta) & G = A/A \\ Binomial(n_A + n_B, 0.5) & G = A/B \\ Binomial(n_A + n_B, \delta) & G = B/B \end{cases}$$

Inference with no reads

Sequence Reads

5’-ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3’

Reference Genome

P(reads|A/A)= 1.0

P(reads|A/C)= 1.0

P(reads|C/C)= 1.0

Possible Genotypes

Inference with short read data

GCTAGCTGATAGCTAGCTAGCTGATGAGCCCGA

Sequence Reads

5’-ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3’

Reference Genome

P(reads|A/A)= P(C observed, read maps |A/A)

P(reads|A/C)= P(C observed, read maps |A/C)

P(reads|C/C)= P(C observed, read maps |C/C)

Possible Genotypes

Inference assuming error of 1%

GCTAGCTGATAGCTAGCTAGCTGATGAGCCCGA

Sequence Reads

5’-ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3’

Reference Genome

P(reads|A/A)= 0.01

P(reads|A/C)= 0.50

P(reads|C/C)= 0.99

Possible Genotypes

As data accumulate …

AGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTG

GCTAGCTGATAGCTAGCTAGCTGATGAGCCCGA

Sequence Reads

5’-ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3’

Reference Genome

P(reads|A/A)= 0.0001

P(reads|A/C)= 0.25

P(reads|C/C)= 0.98

Possible Genotypes

As data accumulate …

ATGCTAGCTGATAGCTAGCTAGCTGATGAGCC

AGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTG

GCTAGCTGATAGCTAGCTAGCTGATGAGCCCGA

Sequence Reads

5'-ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3'

Reference Genome

P(reads|A/A)= 0.000001

P(reads|A/C)= 0.125

P(reads|C/C)= 0.97

Possible Genotypes

As data accumulate …

TAGCTGATAGCTAGATAGCTGATGAGCCCGAT

ATGCTAGCTGATAGCTAGCTAGCTGATGAGCC

AGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTG

GCTAGCTGATAGCTAGCTAGCTGATGAGCCCGA

Sequence Reads

5’-ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3’

Reference Genome

P(reads|A/A)= 0.00000099

P(reads|A/C)= 0.0625

P(reads|C/C)= 0.0097

Possible Genotypes

In the "end"

TAGCTGATAGCTAGATAGCTGATGAGCCCGAT

ATAGCTAGATAGCTGATGAGCCCGATCGCTGCTAGCTC

ATGCTAGCTGATAGCTAGCTAGCTGATGAGCC

AGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTG

GCTAGCTGATAGCTAGCTAGCTGATGAGCCCGA

Sequence Reads

5'-ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3'

Reference Genome

P(reads|A/A)= 0.00000098

P(reads|A/C)= 0.03125

P(reads|C/C)= 0.000097

Not the "end" yet

Sequence Reads

5'-ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3'

Reference Genome

P(reads|A/A) = 0.00000098
P(reads|A/C) = 0.03125
P(reads|C/C) = 0.000097

Making a genotype call requires
combining sequence data with prior information

$$P(Genotype|reads) = \frac{P(reads|Genotype)Prior(Genotype)}{\sum_G P(reads|G)Prior(G)}$$

Not the "end" yet

Sequence Reads

5'-ACTGGTCGATGCTAGCTGATAGCTAGCTAGCTGATGAGCCCGATCGCTGCTAGCTCGACG-3'

Reference Genome

P(reads|A/A)= 0.00000098	Prior(A/A) = 0.00034	P(A/A|reads) < 0.01
P(reads|A/C)= 0.03125	Prior(A/C) = 0.00066	P(A/C|reads) = 0.175
P(reads|C/C)= 0.000097	Prior(C/C) = 0.99900	P(C/C|reads) = 0.825

Base Prior: every site has 1/1000 probability of varying

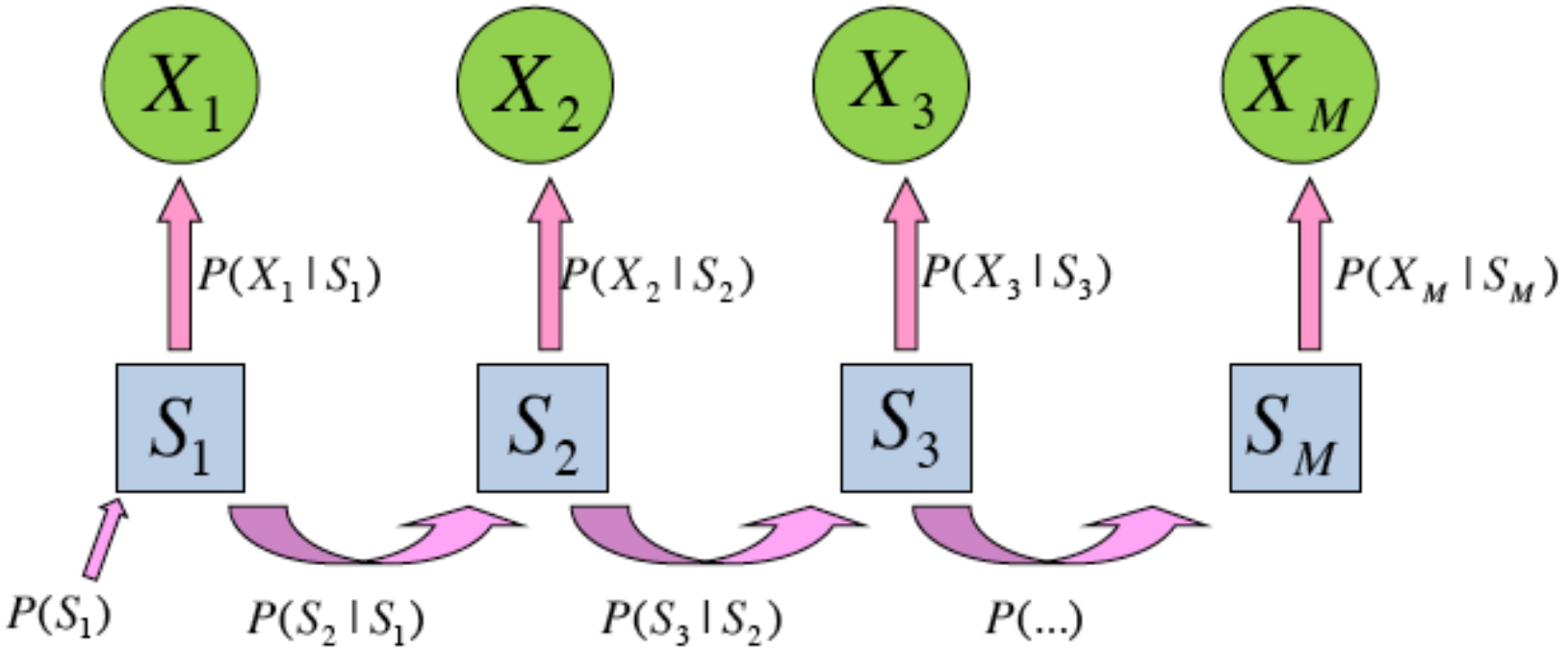
Prior Information

- Individual based prior
  - Equal probability of showing polymorphism
  - 1/1000 bases different from reference
  - Error Free and Poisson distribution
  - Single sample, single site
- Population based prior
  - Estimate frequency from many individuals
  - Multiple sample, single site
- Haplotype/Imputation based prior
  - Jointly model flanking SNPs, use haplotype information
  - Important for low coverage sequence data
  - Multiple samples, multiple sites

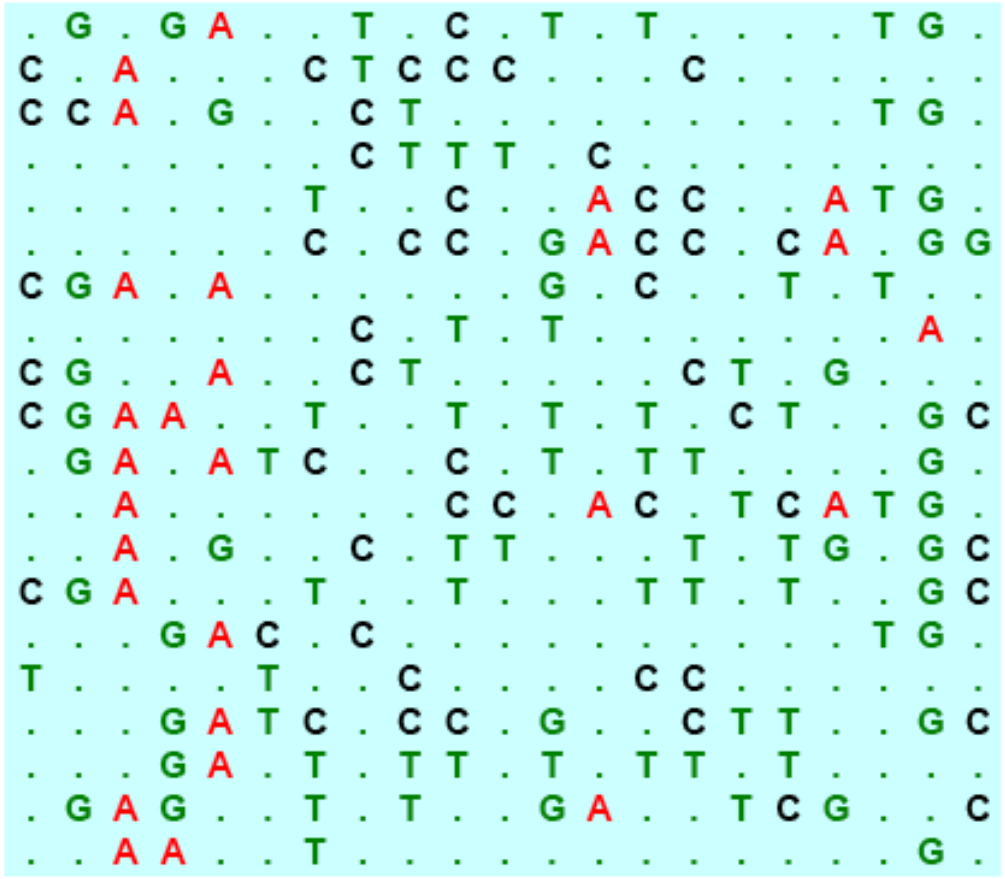
Haplotype-based Inference

- Start with some plausible configuration for each individual
- Use Markov model to update one individual condition on all others
- Repeat previous step many times
- Generate a consensus set of genotypes and haplotypes for each individual

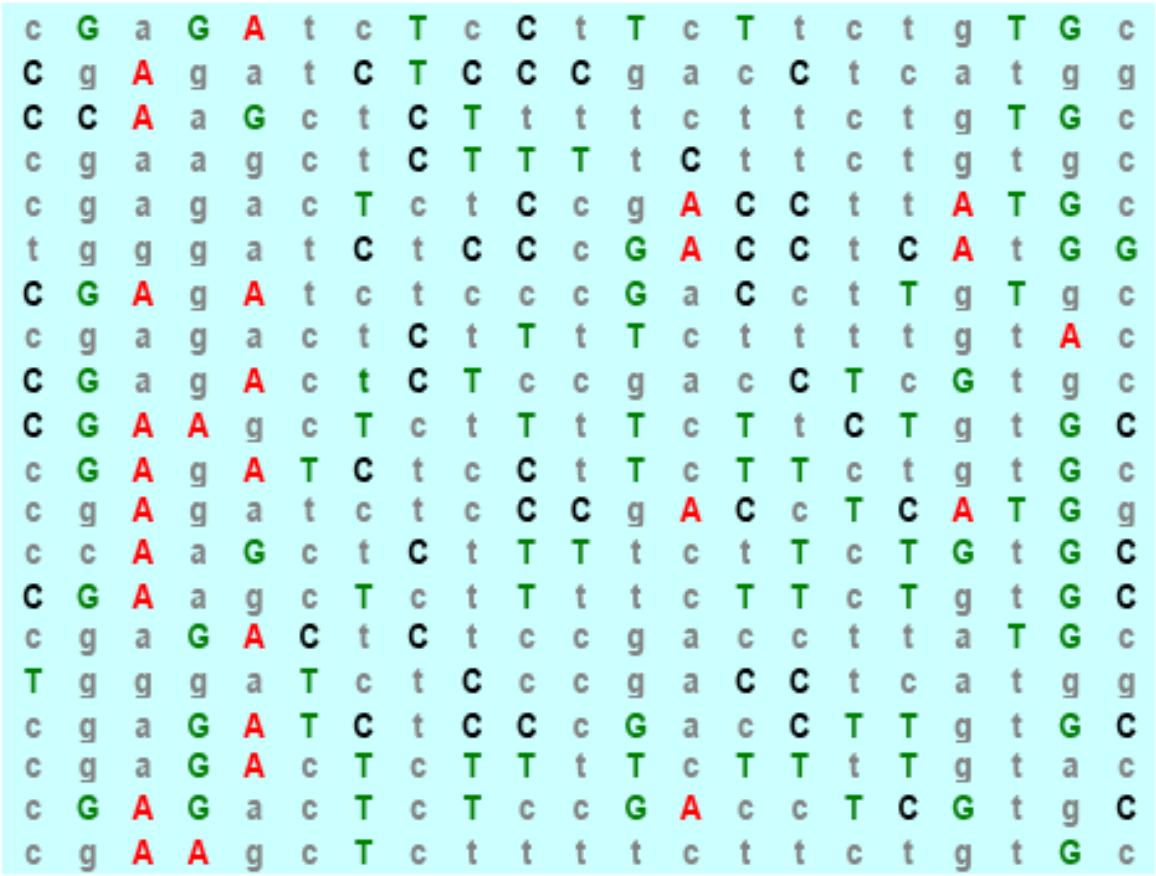
Hidden Markov Model



Cartoon View of Shotgun Data



Cartoon View of Shotgun Data



Cartoon View of the Method

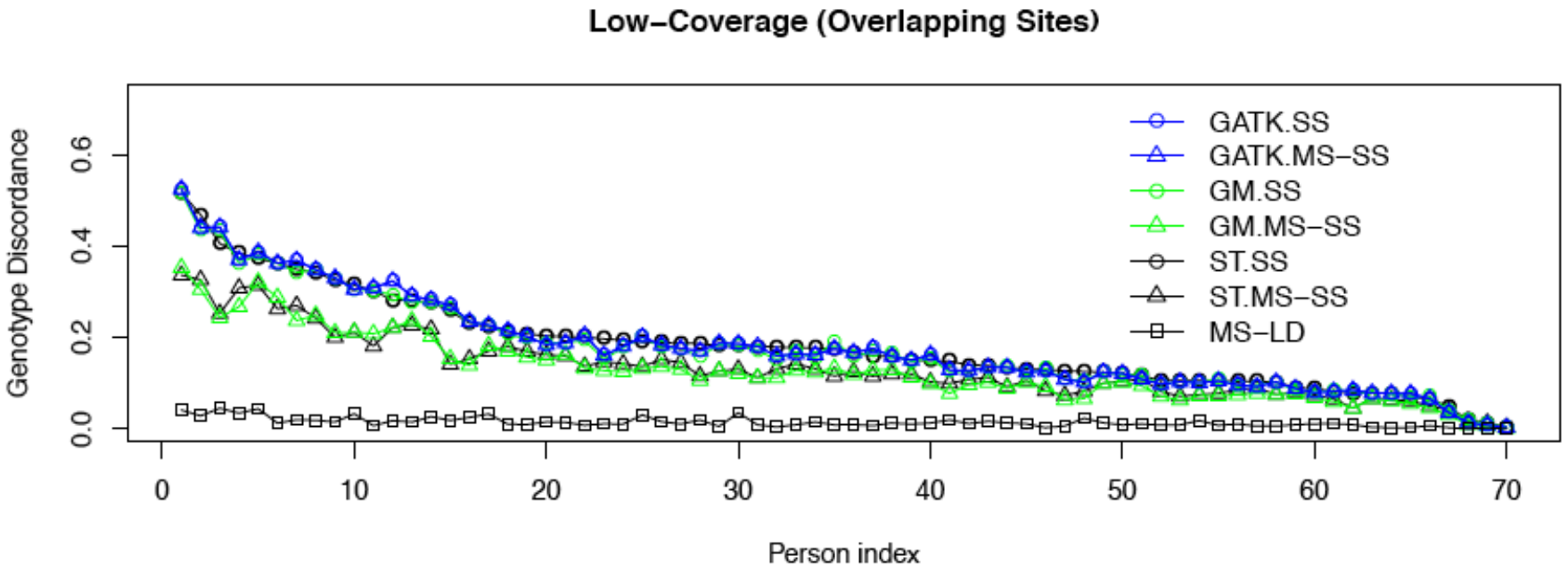
Current update Sample

C	T	C	C	C	G	A	C	C	T	T	G	T	G	C
T	C	T	T	T	T	A	C	C	T	C	A	T	G	G

Known Haplotype Panel

C	T	C	C	C	G	A	C	C	T	C	A	T	G	G
C	T	C	C	C	G	A	C	C	T	T	G	T	G	C
T	C	T	T	T	T	C	T	T	T	T	G	T	A	C
T	C	T	C	C	G	A	C	C	T	C	G	T	G	C
T	C	T	T	T	T	C	T	T	C	T	G	T	G	C
C	T	C	C	T	T	C	T	T	C	T	G	T	G	C
C	T	C	C	C	G	A	C	C	T	C	A	T	G	G
T	C	T	T	T	T	C	T	T	C	T	G	T	G	C
T	C	T	T	T	T	C	T	T	C	T	G	T	G	C
T	C	T	C	C	G	A	C	C	T	T	A	T	G	C

Comparisons of Different Genotype Calling Methods



Simulation Results: Common Sites

- Detection and genotyping of sites
  - MAF > 5%, 2116 simulated SNP sites/Mb

- Detected polymorphic sites:

Sample	Detected Sites	Accuracy (all)	Accuracy(Het)
100	2102	98.5%	90.6%
200	2115	99.6%	99.4%
400	2116	99.8%	99.7%

Simulation Results: Rarer Sites

- Detection and genotyping of sites
  - MAF > 1-2%, 425 simulated SNP sites/Mb

- Detected polymorphic sites:

Sample	Detected Sites	Accuracy (all)	Accuracy(Het)
100	139	98.6%	92.9%
200	213	99.4%	95.0%
400	343	99.6%	95.9%