Challenges and Solutions in the Analysis of Next Generation Sequence Data

Heng Li

Broad Institute

2nd CHOP/PENN NGS Symposium

About me

- One of the major contributors to the SAM specification
- Key developer of several popular software packages:
  - Short-read alignment: MAQ and BWA
  - Long-read alignment: BWA-SW
  - Variant calling and data processing: SAMtools and Tabix
- Involved in the early development of the 1000 Genomes Project
- Google (US/UK) "heng li" for the slides.

Outline

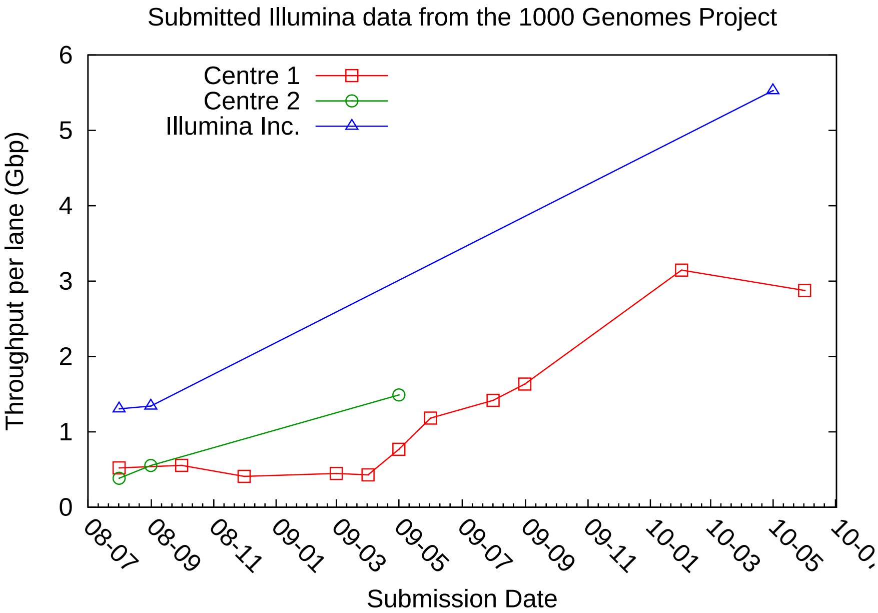
1. Overview of the next-generation sequencing
    - Messages from the 1000 Genomes Project
    - Sequencing machines vs. computers

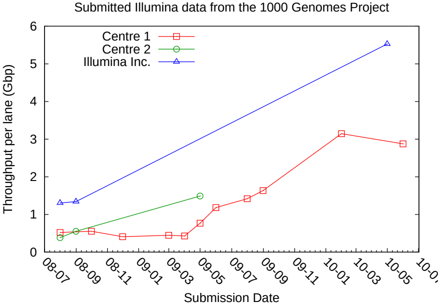
2. Quest for standards

3. The SAM-centric data processing
    - Making a choice: alignment
    - Making a choice: visualization
    - Making a choice: SNP/INDEL discovery
    - SAM-centric data processing

Outline

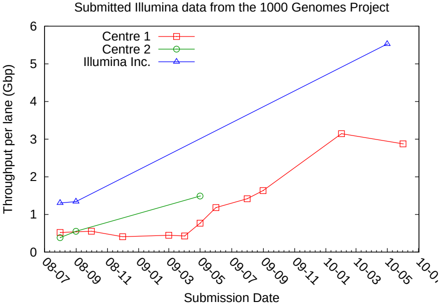
1. Overview of the next-generation sequencing
   - Messages from the 1000 Genomes Project
   - Sequencing machines vs. computers
2. Quest for standards
3. The SAM-centric data processing
   - Making a choice: alignment
   - Making a choice: visualization
   - Making a choice: SNP/INDEL discovery
   - SAM-centric data processing


Submitted Illumina data from the 1000 Genomes Project
Centre 1
Centre 2
Illumina Inc.
Throughput per lane (Gbp)
0
1
2
3
4
5
6
08-07
08-09
08-11
09-01
09-03
09-05
09-07
09-09
09-11
10-01
10-03
10-05
10-0
Submission Date



Illumina sequencing

- 5X increased throughput in $<$2 years
- 4–5Gbp raw sequences per machine day at present



Illumina sequencing

- 5X increased throughput in <2 years
- 4–5Gbp raw sequences per machine day at present
- HiSeq: 14Gbp per lane (not formally submitted yet); ∼30X mappable data to human per half run

Current sequencing technologies

	GA IIx	HiSeq	SOLiD4	454FLX
Read length (bp/color)	2x100	2x100	2x50	400
Run time (days)	9.5	8	14	0.4
Mappable per run (Gbp)	40	160	90	0.5
Throughput (Gbp/day)	4.2	20	6.4	1.1

- Machine yield obtained from vendors' website.
- Assuming 90% of raw seqences mappable for GA IIx and HiSeq.
- HiSeq and SOLiD4 sequence two flow cells/slides per run.

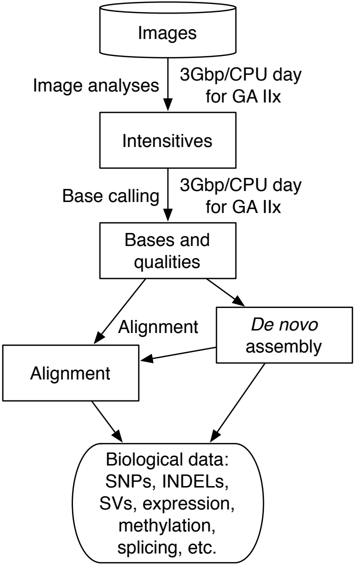
Current sequencing technologies

	GA IIx	HiSeq	SOLiD4	454FLX
Read length (bp/color)	2x100	2x100	2x50	400
Run time (days)	9.5	8	14	0.4
Mappable per run (Gbp)	40	160	90	0.5
Throughput (Gbp/day)	4.2	20	6.4	1.1

- Machine yield obtained from vendors' website.
- Assuming 90% of raw seqences mappable for GA IIx and HiSeq.
- HiSeq and SOLiD4 sequence two flow cells/slides per run.

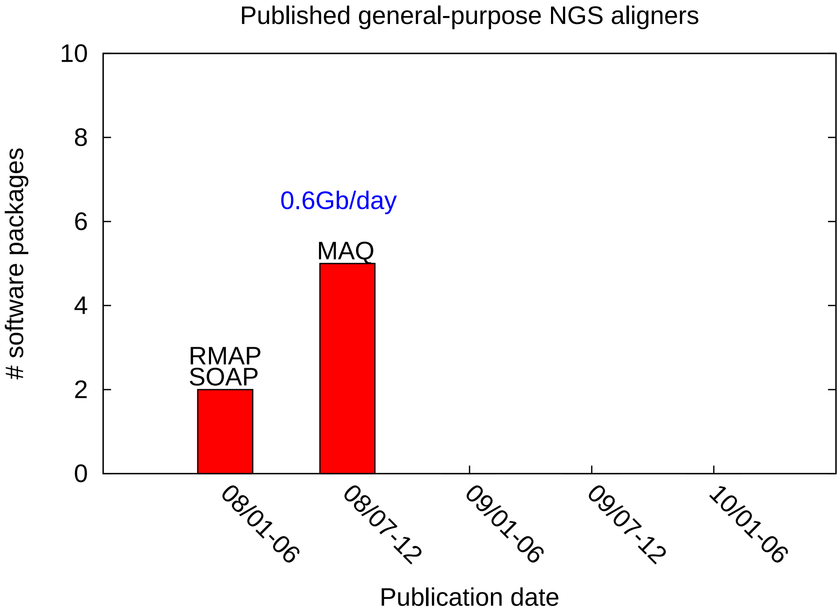
Do computers match sequencing machines in throughput?

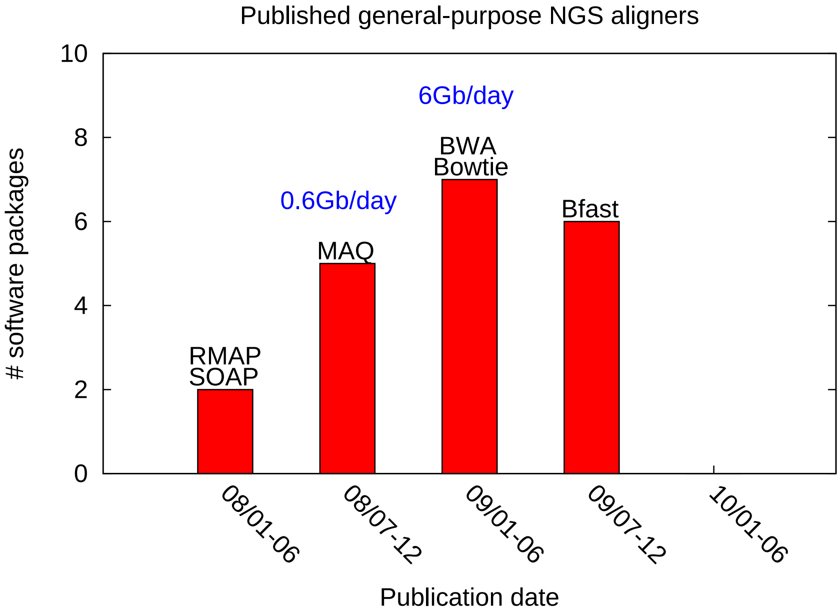
Typical NGS workflow

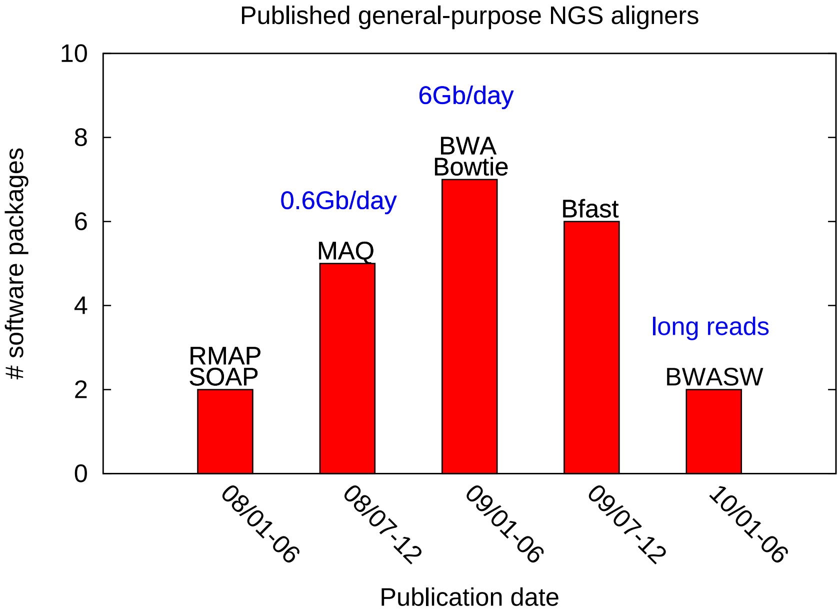


- Image analyses and base calling software are provided by vendors.
- Image are usually analyzed in real time.
- 454 alignment and assembly can be done on intesities.

For image analyses and base calling, an 8-core computer matches HiSeq in throughput.


Published general-purpose NGS aligners
# software packages
0
2
4
6
8
10
0.6Gb/day
RMAP
SOAP
MAQ
08/01-06
08/07-12
09/01-06
09/07-12
10/01-06
Publication date


Published general-purpose NGS aligners
10
8
6
4
2
0
# software packages
6Gb/day
0.6Gb/day
RMAP
SOAP
MAQ
BWA
Bowtie
Bfast
08/01-06
08/07-12
09/01-06
09/07-12
10/01-06
Publication date


Published general-purpose NGS aligners
# software packages
Publication date
0
2
4
6
8
10
08/01-06
08/07-12
09/01-06
09/07-12
10/01-06
RMAP
SOAP
MAQ
BWA
Bowtie
Bfast
BWASW
0.6Gb/day
6Gb/day
long reads

Current sequencing technologies

	GA IIx	HiSeq	SOLiD4	454FLX
Read length (bp/color)	2x100	2x100	2x50	400
Run time (days)	9.5	8	14	0.4
Mappable per run (Gbp)	40	160	90	0.5
Throughput (Gbp/day)	4.2	20	6.4	1.1
Image analysis (Gbp/CPU day)	3	3		
Base calling (Gbp/CPU day)	3	3		
Aln. to human (Gbp/CPU day)	6	6		3

- Machine yield obtained from vendors' website.
- Assuming 90% of raw seqences mappable for GA IIx and HiSeq.
- HiSeq and SOLiD4 sequence two flow cells/slides per run.
- Image analyses done in real-time (on sequencing machines).
- Illumina alignment done by Bowtie/BWA/SOAP2; 454 by BWASW.

A word about Pacific Biosciences (PacBio)

- Polymerase for sequencing
- Single-molecule sequencing
  - less amplification bias
  - DNA methylation (no bisulfite treatment)
- Long reads (In Sequence, 02/10/2009; Eid *et al.*, 2009)
  - $\sim$1000bp in average
  - exponentially distributed: a few very long, many short (Sanger and 454: normal distributed)
- Strobe reads – oriented fragments of a long DNA
- Relatively high error rate (a year ago, Eid *et al.*, 2009)

A word about Pacific Biosciences (PacBio)

- Polymerase for sequencing
- Single-molecule sequencing
  - less amplification bias
  - DNA methylation (no bisulfite treatment)
- Long reads (In Sequence, 02/10/2009; Eid *et al.*, 2009)
  - ∼1000bp in average
  - exponentially distributed: a few very long, many short (Sanger and 454: normal distributed)
- Strobe reads – oriented fragments of a long DNA
- Relatively high error rate (a year ago, Eid *et al.*, 2009)

- Potential for a variety of new applications
- Not an immediate replacement of current technologies

Sequencing machines vs. computers

- *At present*, a tie.
- *In future*, sequencing machines may pull ahead.
- Alignment used to be the bottleneck, but base calling is the slowest step now.
  - Base calling: platform dependent and mostly by vendors. (closed)
  - Alignment: community efforts. (open)

Outline

1. Overview of the next-generation sequencing
   - Messages from the 1000 Genomes Project
   - Sequencing machines vs. computers
2. Quest for standards
3. The SAM-centric data processing
   - Making a choice: alignment
   - Making a choice: visualization
   - Making a choice: SNP/INDEL discovery
   - SAM-centric data processing

The dilemma of openness

A year ago:

- 14 published aligners, 14 (primitive) alignment formats.
- No generic alignment viewers, no generic variant callers.
- Everyone writes their own pipeline from scratch.

Now (a year after the publication of SAM):

- Most popular aligners generate a single format: SAM.
- 10 SAM supported alignment viewers, 3 generic SNP/indel callers.
- Build pipeline upon high-performance tools as well as upon libraries in C, C++, Java, Perl, Python and Ruby.

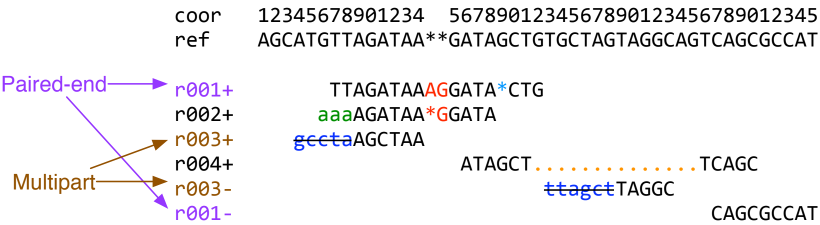
The dilemma of openness

A year ago:

- 14 published aligners, 14 (primitive) alignment formats.
- No generic alignment viewers, no generic variant callers.
- Everyone writes their own pipeline from scratch.

Now (a year after the publication of SAM):

- Most popular aligners generate a single format: SAM.
- 10 SAM supported alignment viewers, 3 generic SNP/indel callers.
- Build pipeline upon high-performance tools as well as upon libraries in C, C++, Java, Perl, Python and Ruby.



```
coor   12345678901234  5678901234567890123456789012345
ref    AGCATGTTAGATAA**GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT

r001+        TTAGATAAAGGATA*CTG
r002+       aaaAGATAA*GGATA
r003+     ~~gccta~~AGCTAA
r004+                   ATAGCT..............TCAGC
r003-                          ~~ttagct~~TAGGC
r001-                                        CAGCGCCAT
```

Ins & padding
Soft clipping

Splicing
Hard clipping

```
@SQ SN:ref LN:45
r001 163 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTA *
r002   0 ref  9 30 3S6M1P1I4M * 0    0 AAAAGATAAGGATA    *
r003   0 ref  9 30 5H6M       * 0    0 AGCTAA   *   NM:i:1
r004   0 ref 16 30 6M14N5M    * 0    0 ATAGCTTCAGC       *
r003  16 ref 29 30 6H5M       * 0    0 TAGGC    *   NM:i:0
r001  83 ref 37 30 9M         = 7  -39 CAGCGCCAT         *
```

```
ref  7 T 1 .    |ref 12 T 3 ...         |ref 17 T 3 ...
ref  8 T 1 .    |ref 13 A 3 ...         |ref 18 A 3 .-1G..
ref  9 A 3 ...  |ref 14 A 2 .+2AG.+1G.  |ref 19 G 2 *.
ref 10 G 3 ...  |ref 15 G 2 ..          |ref 20 C 2 ..
ref 11 A 3 ..C  |ref 16 A 3 ...         |...
```

Features

- Flexibility:
  - Variable read lengths, from short reads to BACs.
  - Indels, splicing, clipping and multi-part alignment.
  - User defined or aligner specific information.
- Efficiency:
  - Compact in file size (one byte per raw base).
  - Minimal memory requirement.
  - Random access.
- Open alignment files over FTP/HTTP.
  - Interested in a few genes, but cannot afford 40TB alignments.
- Matured and stable.

Features

- Flexibility:
  - Variable read lengths, from short reads to BACs.
  - Indels, splicing, clipping and multi-part alignment.
  - User defined or aligner specific information.
- Efficiency:
  - Compact in file size (one byte per raw base).
  - Minimal memory requirement.
  - Random access.
- Open alignment files over FTP/HTTP.
  - Interested in a few genes, but cannot afford 40TB alignments.
- Matured and stable.

The SAM/BAM format is the standard.

Outline

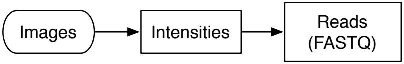
1 Overview of the next-generation sequencing

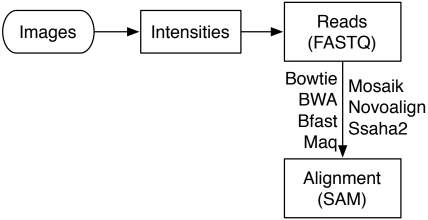
- Messages from the 1000 Genomes Project
- Sequencing machines vs. computers

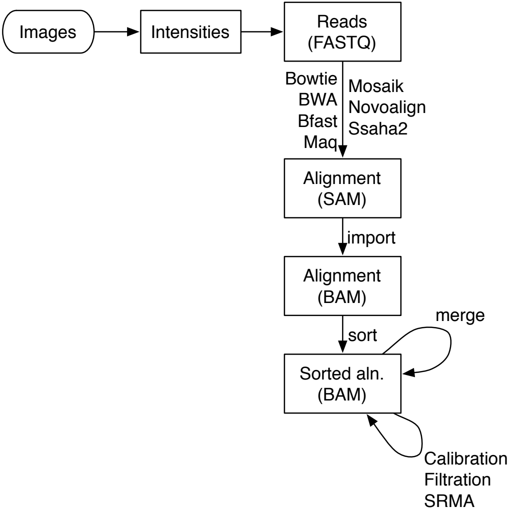
2 Quest for standards

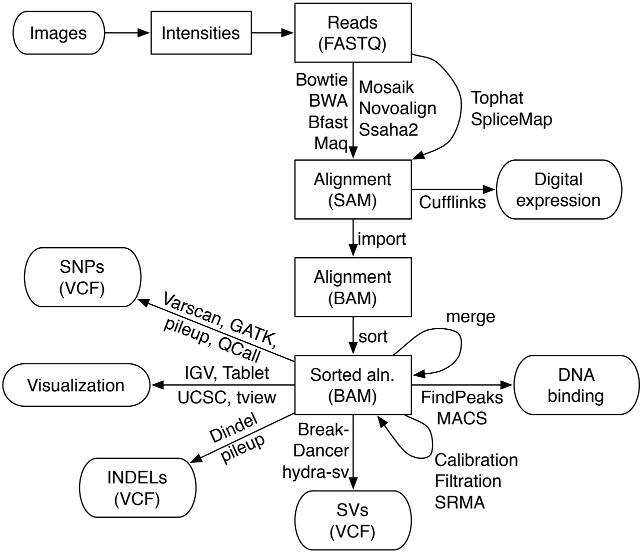
3 The SAM-centric data processing

- Making a choice: alignment
- Making a choice: visualization
- Making a choice: SNP/INDEL discovery
- SAM-centric data processing

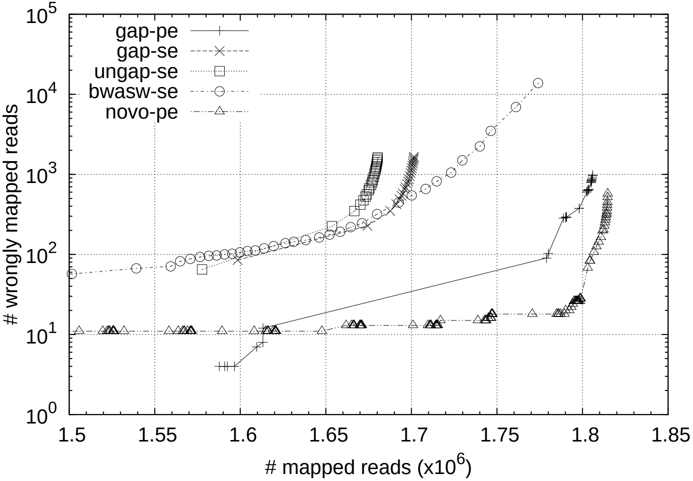

Images
Intensities
Reads (FASTQ)


Images
Intensities
Reads (FASTQ)
Bowtie
BWA
Bfast
Maq
Mosaik
Novoalign
Ssaha2
Alignment (SAM)

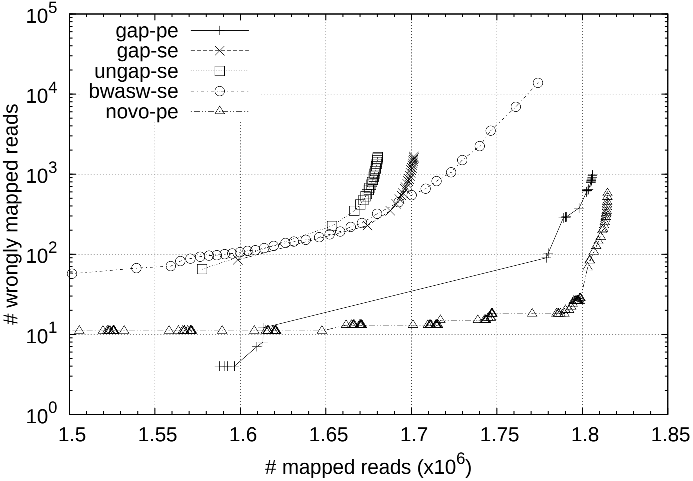

Images
Intensities
Reads (FASTQ)
Bowtie
BWA
Bfast
Maq
Mosaik
Novoalign
Ssaha2
Alignment (SAM)
import
Alignment (BAM)
sort
merge
Sorted aln. (BAM)
Calibration
Filtration
SRMA


Images
Intensities
Reads (FASTQ)
Bowtie
BWA
Bfast
Maq
Mosaik
Novoalign
Ssaha2
Tophat
SpliceMap
Alignment (SAM)
Cufflinks
Digital expression
import
Alignment (BAM)
sort
merge
Sorted aln. (BAM)
Varscan, GATK, pileup, QCall
SNPs (VCF)
IGV, Tablet
UCSC, tview
Visualization
Dindel
pileup
INDELs (VCF)
Break-Dancer
hydra-sv
SVs (VCF)
FindPeaks
MACS
DNA binding
Calibration
Filtration
SRMA

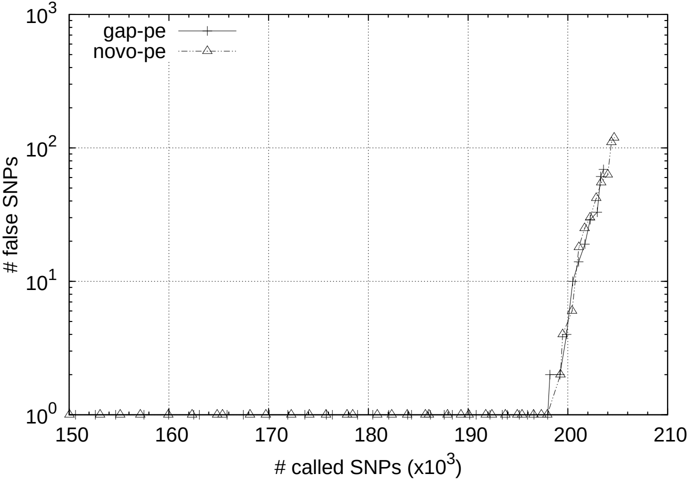
Not all aligners are equal.



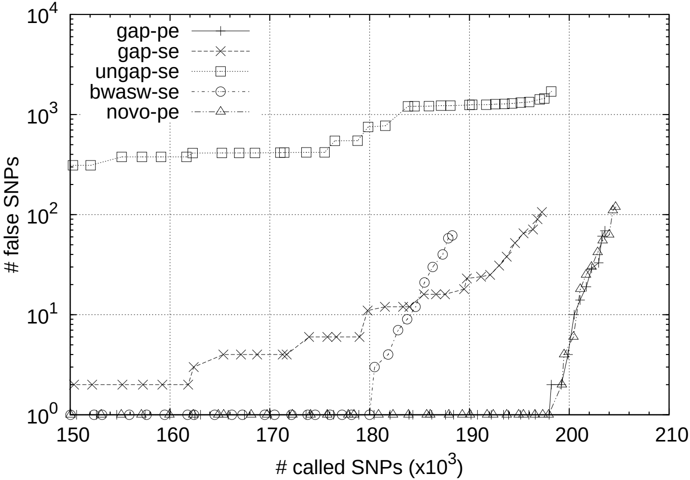
Not all aligners are equal.



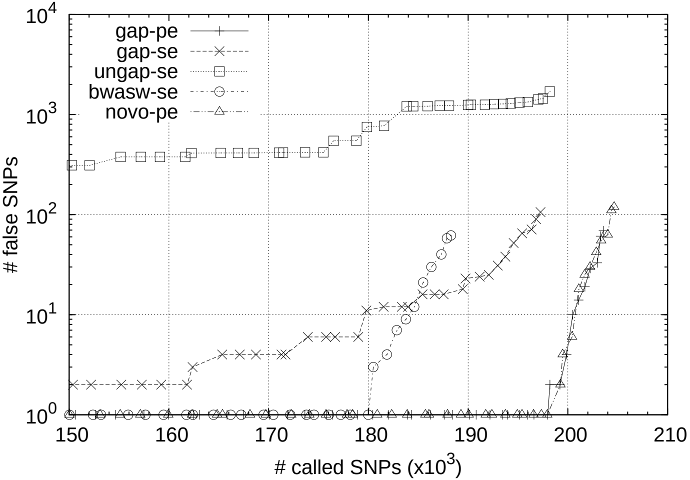
Systematic errors are more dangerous.



Systematic errors are more dangerous.



Systematic errors are more dangerous.



Choose an alignment algorithm

Sources of alignment errors:

- Repeats and *known* segmental duplications
- Approximations and heuristics
- Short indels
- Incompete reference genomes
- Typical alignment error rate: <1%.

Choose an alignment algorithm

Sources of alignment errors:

- Repeats and *known* segmental duplications – aligners know this
- Approximations and heuristics – getting better
- Short indels – getting better
- Incompete reference genomes – increasingly hurting
- Typical alignment error rate: $<1\%$.

Choose an alignment algorithm

Sources of alignment errors:

- Repeats and *known* segmental duplications – aligners know this
- Approximations and heuristics – getting better
- Short indels – getting better
- Incompete reference genomes – increasingly hurting
- Typical alignment error rate: $<1\%$.

The baseline:

Don't let the alignment errors be dominant.

Choose alignment algorithms based on needs

Depth based CNV discovery and study of highly expressed genes:

- Other sources of errors/variation dominate
- Use a fast aligner (e.g. Bowtie/SOAP2)

SNP/INDEL discovery and study of weakly expressed genes:

- 10–20% of short variants in human are indels.
- Use a gapped and accurate aligner (e.g. bwa/novoalign).

SV discovery; somatic mutations

- Tiny alignment errors are hurting.
- Combine distinct algorithms (e.g. bwa+mosaik/bwasw)

Choose alignment algorithms based on needs

Depth based CNV discovery and study of highly expressed genes:

- Other sources of errors/variation dominate
- Use a fast aligner (e.g. Bowtie/SOAP2)

SNP/INDEL discovery and study of weakly expressed genes:

- 10–20% of short variants in human are indels.
- Use a gapped and accurate aligner (e.g. bwa/novoalign).

SV discovery; somatic mutations

- Tiny alignment errors are hurting.
- Combine distinct algorithms (e.g. bwa+mosaik/bwasw)

Choose alignment algorithms based on needs

Depth based CNV discovery and study of highly expressed genes:

- Other sources of errors/variation dominate
- Use a fast aligner (e.g. Bowtie/SOAP2)

SNP/INDEL discovery and study of weakly expressed genes:

- 10–20% of short variants in human are indels.
- Use a gapped and accurate aligner (e.g. bwa/novoalign).
- All high-profile resequencing projects use gapped aligners.

SV discovery; somatic mutations

- Tiny alignment errors are hurting.
- Combine distinct algorithms (e.g. bwa+mosaik/bwasw)

Choose alignment algorithms based on needs

Depth based CNV discovery and study of highly expressed genes:

- Other sources of errors/variation dominate
- Use a fast aligner (e.g. Bowtie/SOAP2)

SNP/INDEL discovery and study of weakly expressed genes:

- 10–20% of short variants in human are indels.
- Use a gapped and accurate aligner (e.g. bwa/novoalign).
- All high-profile resequencing projects use gapped aligners.

SV discovery; somatic mutations

- Tiny alignment errors are hurting.
- Combine distinct algorithms (e.g. bwa+mosaik/bwasw)

Visualization

- Base-pair resolution:
  - IGV: highly tuned for SNP/SV discovery
  - SAMtools tview: fast over slow network
  - Tablet: Support many alignment/assembly formats; elegant interface
- Low resolution:
  - SeqMonk: highly tuned for ChIP/RNA-seq
- Web based:
  - UCSC: BAM support plus much more
  - GBrowse: BAM support

Types of SNP calling

- Single sample: samtools, GATK, QCall and VarScan
- Pooled sample: Syzygy
- Multi-sample, without LD: GATK and QCall
- Multi-sample, with LD: GATK+Beagle and QCall
- Contrast samples (e.g. tumor vs. normal): ?

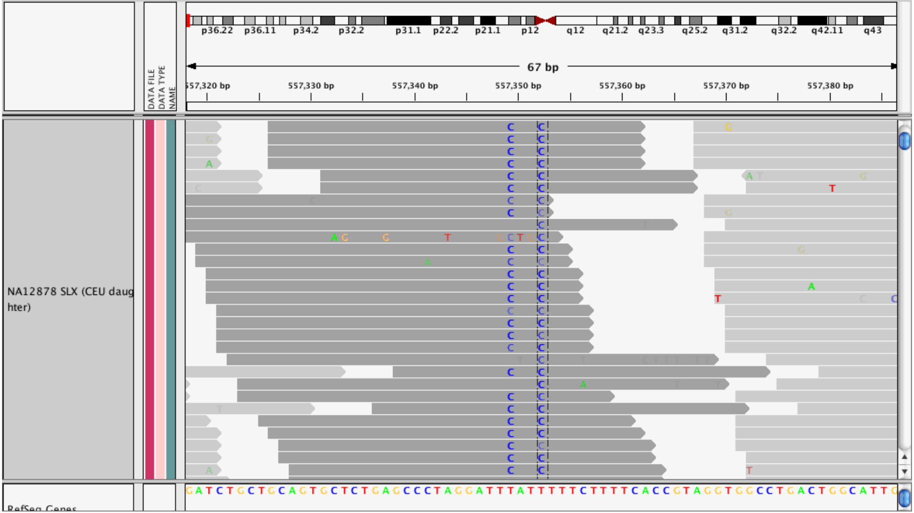
Use samtools' pileup to write your own variant caller.

Types of SNP calling

- Single sample: samtools, GATK, QCall and VarScan
- Pooled sample: Syzygy
- Multi-sample, without LD: GATK and QCall
- Multi-sample, with LD: GATK+Beagle and QCall
- Contrast samples (e.g. tumor vs. normal): ?

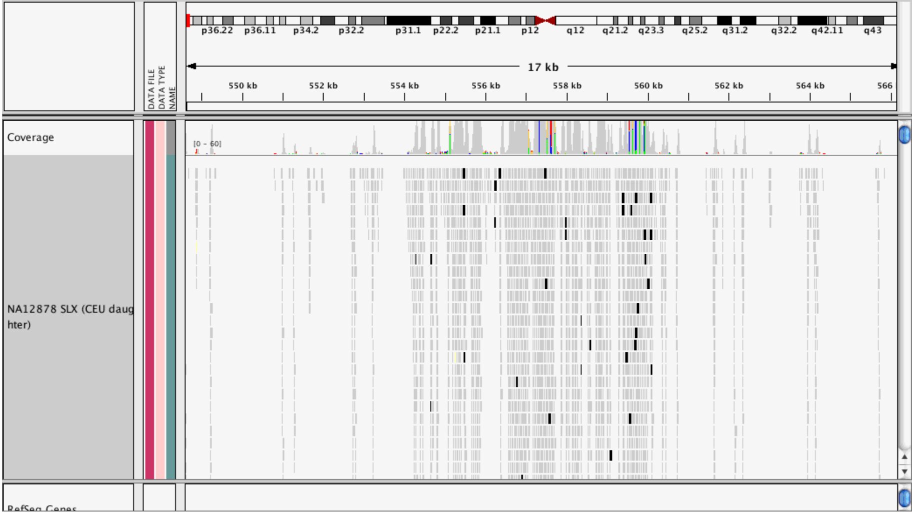
Use samtools' pileup to write your own variant caller.

Artefacts: strand bias



from the GATK website

Artefacts: excessive coverage



from the GATK website

Other factors considered in filtering

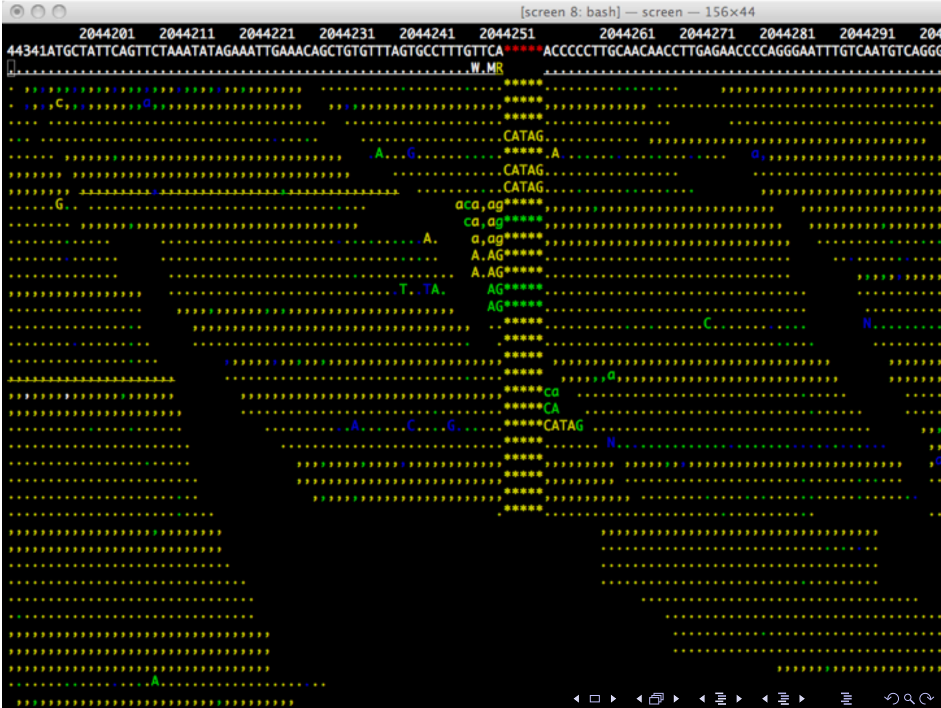
- Allele balance: #ref-alleles/#snp-alleles
- Mapping quality: the accessibility of the region
- Clustered SNPs: sign of paralogous mapping
- Length of the longest homopolymer run
- Break of diploidity: paralogous mapping

Other factors considered in filtering

- Allele balance: #ref-alleles/#snp-alleles
- Mapping quality: the accessibility of the region
- Clustered SNPs: sign of paralogous mapping
- Length of the longest homopolymer run
- Break of diploidity: paralogous mapping

Towards high accuracy SNP calling:

- Necessary for somatic mutations.
- A good alignment viewer (e.g. IGV and tview) is essential – SAM only!



[screen 8: bash] — screen — 156×44

```
         2044201   2044211   2044221   2044231   2044241   2044251       2044261   2044271   2044281   2044291   204
44341ATGCTATTCAGTTCTAAATATAGAAATTGAAACAGCTGTGTTTAGTGCCTTTGTTCA*****ACCCCCTTGCAACAACCTTGAGAACCCCAGGGAATTTGTCAATGTCAGGG
.............................................................W.MR     ..................................................
```

INDEL calling

Problem:

- Reads are indenpendent in alignment, but reads mapped to the same locus are correlated.
- Naive indel callers would not work well.

Solution:

- Sophisticated indel callers all do realignment implicitly (e.g. Dindel and samtools).
- Explicit realignment (e.g. SRMA and GATK).

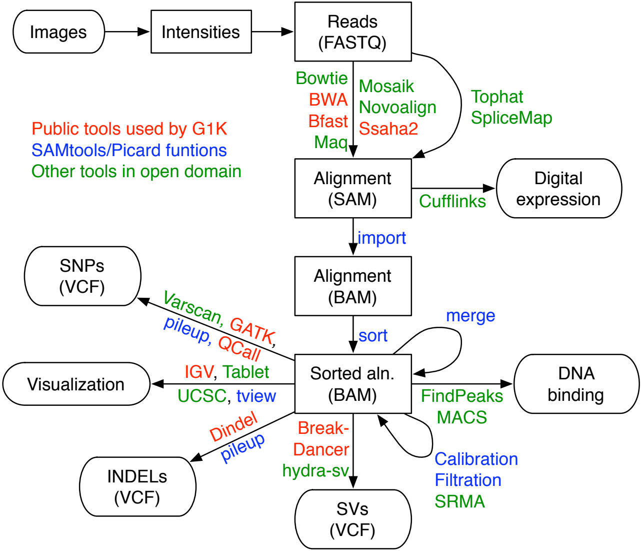
INDEL calling

Problem:

- Reads are indenpendent in alignment, but reads mapped to the same locus are correlated.
- Naive indel callers would not work well.

Solution:

- Sophisticated indel callers all do realignment implicitly (e.g. Dindel and samtools).
- Explicit realignment (e.g. SRMA and GATK).



Images

Intensities

Reads (FASTQ)

Bowtie

BWA

Bfast

Maq

Mosaik

Novoalign

Ssaha2

Tophat

SpliceMap

Public tools used by G1K

SAMtools/Picard funtions

Other tools in open domain

Alignment (SAM)

Cufflinks

Digital expression

import

Alignment (BAM)

sort

merge

SNPs (VCF)

Varscan, GATK,

pileup, QCall

Visualization

IGV, Tablet

UCSC, tview

Sorted aln. (BAM)

FindPeaks

MACS

DNA binding

Dindel

pileup

INDELs (VCF)

Break-Dancer

hydra-sv

SVs (VCF)

Calibration

Filtration

SRMA

The SAM-centric data processing

- Top tools for all applications.
- Vendor supports (Illumina, AB, Complete Genomics and potentially PacBio).
- Efficiency: processing 40X human resequencing data in a week (∼30 CPU days) with 500GB disk space.

The SAM-centric data processing

- Top tools for all applications.
- Vendor supports (Illumina, AB, Complete Genomics and potentially PacBio).
- Efficiency: processing 40X human resequencing data in a week (∼30 CPU days) with 500GB disk space.

- Following the best practices, small labs can afford analyzing deep human resequencing data.
- Analyzing data of smaller scale is even easier.

Summary

- A sequencing machine can produce up to 20G base pairs per day.
- Data processing is keeping the pace with the increasing throughput of sequence machines.
- The SAM/BAM format is the pivot of data processing. Adopted by most major sequencing centers.
- Following the best practices, small labs can handle data from the latest sequencing machines.

Acknowledgements

- 1000 Genomes Project analyses group
- Bob Handsaker and Richard Durbin
- Tim Fennel, Mark Depristo and the GSA group at Broad
- SAMtools/Picard users
- Altshuler/Daly lab and Reich lab
- Xiaowu Gai

Thank You