Mapping, Alignment and SNP Calling

Heng Li

Broad Institute

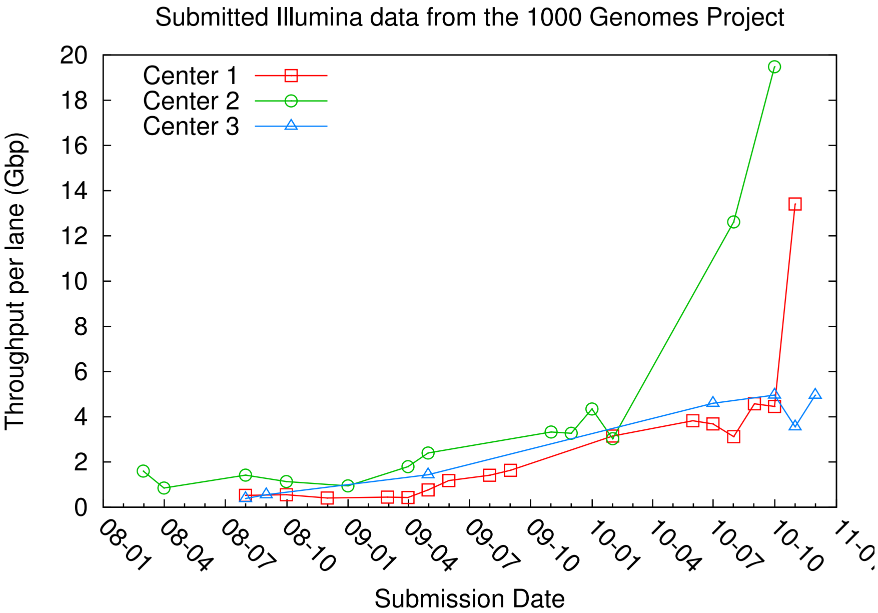
MPG Next Gen Workshop 2011

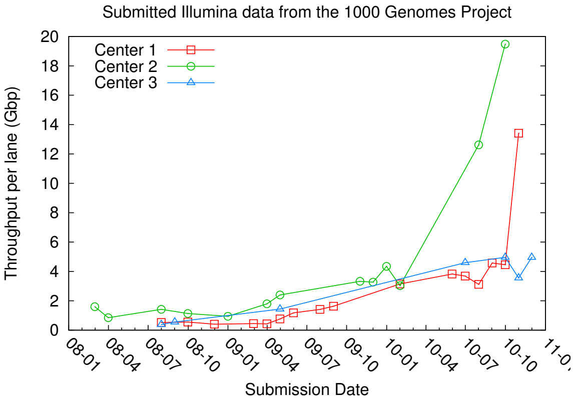
Outline

1. Mapping
   - Messages from the 1000 Genomes Project
   - A race in throughput
2. Alignment
   - Mapping vs. alignment
   - Fixing wrong alignments
3. SNP calling
   - Single-sample SNP calling
   - Multi-sample SNP calling

Outline

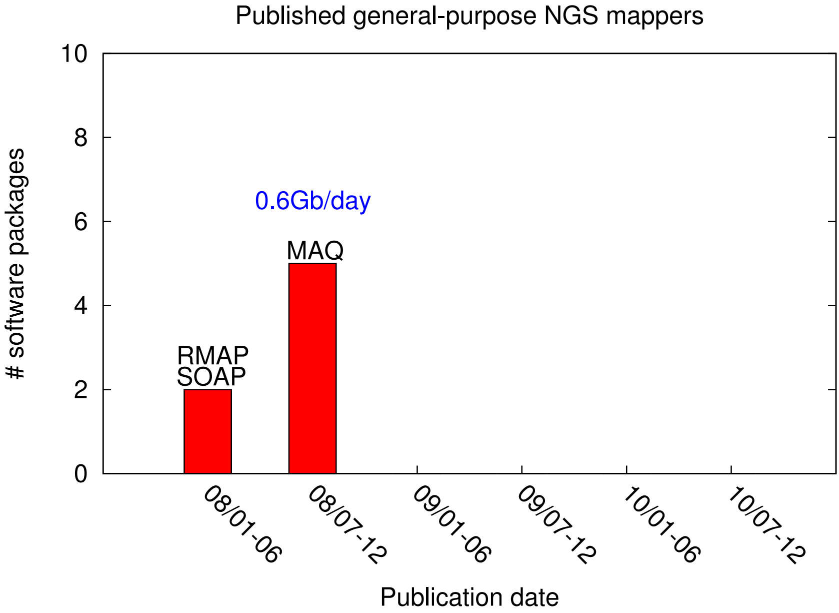
1. Mapping
   - Messages from the 1000 Genomes Project
   - A race in throughput
2. Alignment
   - Mapping vs. alignment
   - Fixing wrong alignments
3. SNP calling
   - Single-sample SNP calling
   - Multi-sample SNP calling

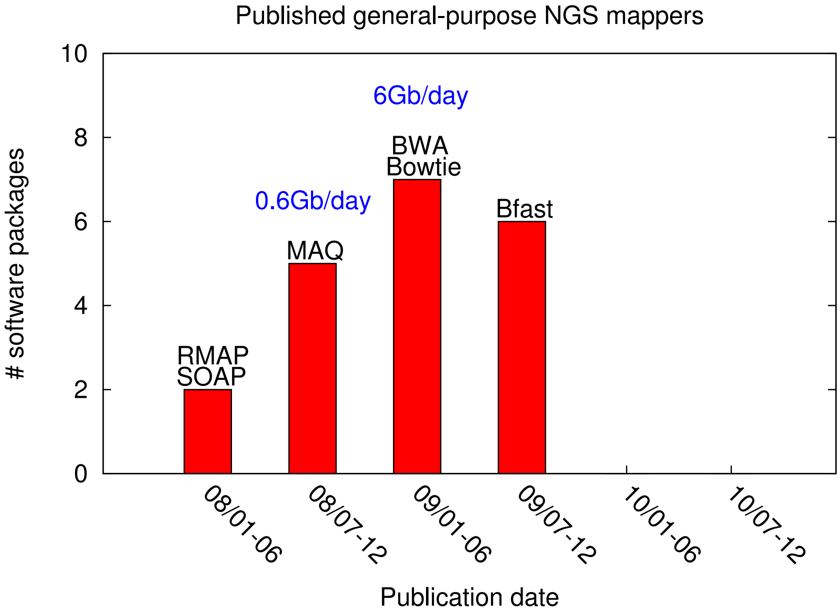

Submitted Illumina data from the 1000 Genomes Project
Throughput per lane (Gbp)
0
2
4
6
8
10
12
14
16
18
20
Center 1
Center 2
Center 3
08-01
08-04
08-07
08-10
09-01
09-04
09-07
09-10
10-01
10-04
10-07
10-10
11-0
Submission Date

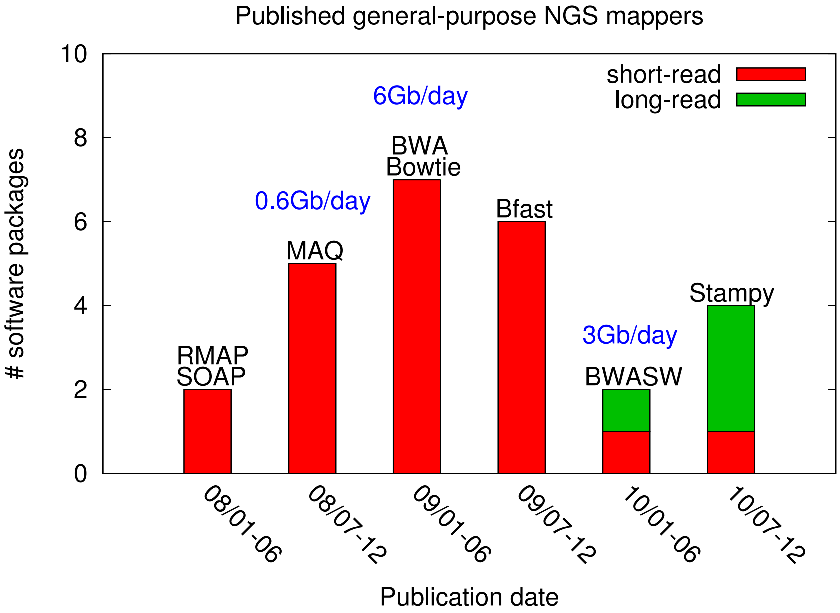


Illumina sequencing

- >20X increased throughput in 3 years
- ∼20Gbp raw sequences per machine day at present


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


Published general-purpose NGS mappers
# software packages
0
2
4
6
8
10
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
10/07-12
Publication date


Published general-purpose NGS mappers
short-read
long-read
6Gb/day
0.6Gb/day
3Gb/day
RMAP
SOAP
MAQ
BWA
Bowtie
Bfast
BWASW
Stampy
# software packages
0
2
4
6
8
10
08/01-06
08/07-12
09/01-06
09/07-12
10/01-06
10/07-12
Publication date

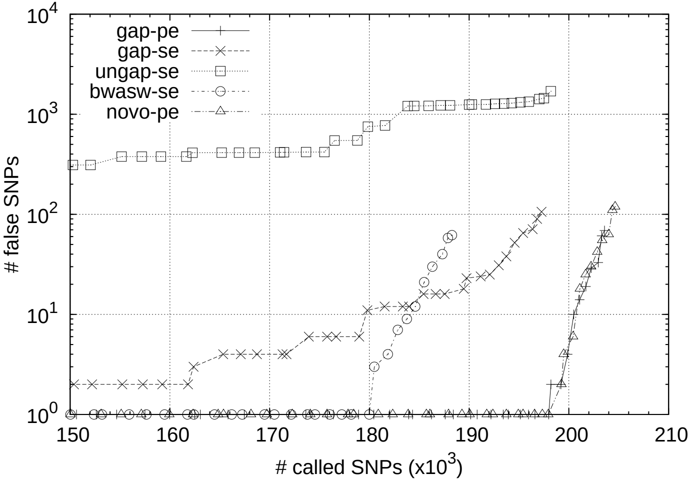
Convergence in mapping algorithms

- Recommended mappers for *variant calling*:
  - Illumina: BWA, Eland2, Novoalign, Stampy
  - SOLiD: Bfast, BioScope
  - 454: SSAHA2, gsMapper, BWA-SW
- Modern short-read mappers are faster than image analysis and base calling.
  - No need for further speed improvements for short reads
  - Long-read and multi-reference alignments still pose challenges

Outline

1. Mapping
   - Messages from the 1000 Genomes Project
   - A race in throughput
2. Alignment
   - Mapping vs. alignment
   - Fixing wrong alignments
3. SNP calling
   - Single-sample SNP calling
   - Multi-sample SNP calling

Ungapped mappers perform badly for SNP calling



A typical SNP caller sees...

```
9    t   ttt
10   a   aaaC
11   a   aaaaa
12   a   aaaaaa
13   a   aaaaaa
14   c   cccTTT
15   a   aaaaaa
16   a   aaaaaa
17   t   AAtttt
18   t   tttttt
19   a   aaaaaa
20   a   aaaaaa
21   g   Tgggg
```

The alignment looks like...

9	t	ttt
10	a	aaaC
11	a	aaaaa
12	a	aaaaaa
13	a	aaaaaa
14	c	cccTTT
15	a	aaaaaa
16	a	aaaaaa
17	t	AAtttt
18	t	tttttt
19	a	aaaaaa
20	a	aaaaaa
21	g	Tgggg

```
coor     12345678901234    5678901234567890123456
ref      aggttttataaaac----aattaagtctacagagcaacta
sample   aggttttataaaacAAATaattaagtctacagagcaacta
read1    aggttttataaaac    aaAtaa
read2     ggttttataaaac    aaAtaaTt
read3         ttataaaacAAATaattaagtctaca
read4             CaaaT    aattaagtctacagagcaac
read5               aaT    aattaagtctacagagcaact
read6                 T    aattaagtctacagagcaacta
```

But what is really happening is...

9	t	ttt
10	a	aaaC
11	a	aaaaa
12	a	aaaaaa
13	a	aaaaaa
14	c	cccTTT
15	a	aaaaaa
16	a	aaaaaa
17	t	AAtttt
18	t	tttttt
19	a	aaaaaa
20	a	aaaaaa
21	g	Tgggg

```
coor     12345678901234    5678901234567890123456
ref      aggttttataaaac----aattaagtctacagagcaacta
sample   aggttttataaaacAAATaattaagtctacagagcaacta
read1    aggttttataaaac    aaAtaa
read2     ggttttataaaac    aaAtaaTt
read3         ttataaaacAAATaattaagtctaca
read4             CaaaT    aattaagtctacagagcaac
read5               aaT    aattaagtctacagagcaact
read6                 T    aattaagtctacagagcaacta
read1    aggttttataaaacaaataa
read2     ggttttataaaacaaataatt
read3         ttataaaacaaataattaagtctaca
read4                 caaataattaagtctacagagcaac
read5                   aataattaagtctacagagcaact
read6                     taattaagtctacagagcaacta
```

Mapping vs. alignment

Mapping

- A mapping is the region where a read sequence is placed.
- A mapping is regarded to be correct if it overlaps the true region.

Alignment

- An alignment is the detailed placement of each base in a read.
- An alignment is regarded to be correct only if each base is placed correctly.

The problem

- A read mapper is fairly good at mapping, may not be good at alignment.
- This is because the true alignment minimizes differences between reads, but the read mapper only sees the reference.

Mapping vs. alignment

Mapping

- A mapping is the region where a read sequence is placed.
- A mapping is regarded to be correct if it overlaps the true region.

Alignment

- An alignment is the detailed placement of each base in a read.
- An alignment is regarded to be correct only if each base is placed correctly.

The problem

- A read mapper is fairly good at mapping, may not be good at alignment.
- This is because the true alignment minimizes differences between reads, but the read mapper only sees the reference.

Fixing wrong alignments

Multi-sequence realignment

- Perform multi-seq alignment to minimize differences between reads.
- Effective for long gaps.

Base Alignment Quality (BAQ)

- Measure the probability of a read base being misaligned
  - BAQ is low if the read base is aligned to a different reference base in a suboptimal alignment.
  - Bases with low BAQ ignored or downweighted in SNP calling.
- Effective even if no reads are mapped with gaps.
- Work by traversing all the possible alignment between the read and the reference.
- Computed efficiently with an HMM.

Fixing wrong alignments

Multi-sequence realignment

- Perform multi-seq alignment to minimize differences between reads.
- Effective for long gaps.

Base Alignment Quality (BAQ)

- Measure the probability of a read base being misaligned
  - BAQ is low if the read base is aligned to a different reference base in a suboptimal alignment.
  - Bases with low BAQ ignored or downweighted in SNP calling.
- Effective even if no reads are mapped with gaps.
- Work by traversing all the possible alignment between the read and the reference.
- Computed efficiently with an HMM.

Evaluation on simulated data

GATKrealn	BAQ	FNR	# false SNPs
No	No	7.3%	116
Yes	No	7.6%	4
No	Yes	8.3%	2
Yes	Yes	8.3%	0

- No filtering applied except a quality cutoff
- BAQ and multi-sequence realignment complement each other:
  - BAQ is less effective given long gaps.
  - The current realignment algorithm is less effective if no reads are mapped with gaps.

Outline

1. Mapping
   - Messages from the 1000 Genomes Project
   - A race in throughput
2. Alignment
   - Mapping vs. alignment
   - Fixing wrong alignments
3. SNP calling
   - Single-sample SNP calling
   - Multi-sample SNP calling

Single-sample Bayesian caller: a toy example

Input

Reference is C, observing 4C and 2T, all with base quality 30.

Likelihood of data

- $P(D|CC) = \Pr\{\text{two Q30 errors}\} = 10^{-(30+30)/10} = 10^{-6}$
- $P(D|TT) = \Pr\{\text{four Q30 errors}\} = 10^{-(30*4)/10} = 10^{-12}$
- $P(D|CT) = \Pr\{\text{sample 6 reads from 2 chr}\} = 1/2^6 = 1.56 \times 10^{-2}$

Posterior

- Prior: $P(CC) = 0.9985$, $P(CT) = 0.001$ and $P(TT) = 0.0005$

$$P(CC|D) = \frac{P(D|CC)P(CC)}{P(D|CC)P(CC) + P(D|CT)P(CT) + P(D|TT)P(TT)}$$

- Get: $P(CC|D) = 0.06$, $P(CT|D) = 0.94$ and $P(TT|D) = 3 \times 10^{-11}$

Single-sample Bayesian caller: a toy example

Input

Reference is C, observing 4C and 2T, all with base quality 30.

Likelihood of data

- $P(D|CC) = \Pr\{\text{two Q30 errors}\} = 10^{-(30+30)/10} = 10^{-6}$
- $P(D|TT) = \Pr\{\text{four Q30 errors}\} = 10^{-(30*4)/10} = 10^{-12}$
- $P(D|CT) = \Pr\{\text{sample 6 reads from 2 chr}\} = 1/2^6 = 1.56 \times 10^{-2}$

Posterior

- Prior: $P(CC) = 0.9985$, $P(CT) = 0.001$ and $P(TT) = 0.0005$

$$P(CC|D) = \frac{P(D|CC)P(CC)}{P(D|CC)P(CC) + P(D|CT)P(CT) + P(D|TT)P(TT)}$$

- Get: $P(CC|D) = 0.06$, $P(CT|D) = 0.94$ and $P(TT|D) = 3 \times 10^{-11}$

Single-sample Bayesian caller: a toy example

Input

Reference is C, observing 4C and 2T, all with base quality 30.

Likelihood of data

- $P(D|CC) = \Pr\{\text{two Q30 errors}\} = 10^{-(30+30)/10} = 10^{-6}$
- $P(D|TT) = \Pr\{\text{four Q30 errors}\} = 10^{-(30*4)/10} = 10^{-12}$
- $P(D|CT) = \Pr\{\text{sample 6 reads from 2 chr}\} = 1/2^6 = 1.56 \times 10^{-2}$

Posterior

- Prior: $P(CC) = 0.9985$, $P(CT) = 0.001$ and $P(TT) = 0.0005$

$$P(CC|D) = \frac{P(D|CC)P(CC)}{P(D|CC)P(CC) + P(D|CT)P(CT) + P(D|TT)P(TT)}$$

- Get: $P(CC|D) = 0.06$, $P(CT|D) = 0.94$ and $P(TT|D) = 3 \times 10^{-11}$

Multi-sample Bayesian caller: an overview

- Similar to single-sample calling except replacing the individual genotype with the genotype configuration of multiple samples.
- Math magic to accelerate computation.
- Allele frequency estimated at the same time.

Multi-sample vs. pooled SNP calling

An example

- 1 sample covered by 3 Q20 C bases (1% error rate); 99 samples covered by 297 Q20 T bases.
- Very unlikely for 3 errors appear in one sample.
- Without sample information, the 3 C look like perfect sequencing errors.

Combining pooling and barcoding

- Pool less than 100 samples together, barcode each pool and then sequence.

Acknowledgements

- 1000 Genomes Project analyses group
- Mark Depristo and the GSA group at Broad
- SAMtools/Picard users
- Altshuler/Daly lab and Reich lab

Thank You