Mapping short reads

But first…

- Friday: evening session (7:30pm? 9pm?)
- Saturday: no evening session; just BBQ on Windmill Island.
- Sunday: only lunch is served…

Mapping short reads

Note: long v short

- Mapping *long* reads is a different problem from mapping short reads.

- This is for two reasons, both of them pragmatic/practical:
  - The volume of data has traditionally been much less: 1m 454 reads vs 200m Illumina
  - Long reads are much more likely to have insertions or deletions in them

Long reads: BLAST vs 'blat'

- On Wed, I told you about the 11-mer heuristic that BLAST uses for DNA matches.
  - BLAST requires that a query sequence contains the same 11-mer as a database sequence before it attempts further alignment.
  - Any given 11-mer occurs only once in 2m sequences, so this filters out many database sequences quickly.
  - You can also store the list of all possible 11-mers in memory easily (~2mb), making it possible to keep track of everything quickly.
- 'blat' does the same thing as BLAST, but is faster because it uses longer k-mers.

How *alignment* works, and why indels are the devil

There are many alignment strategies, but most work like this:

```
GCGGAGatggac        GCGGAGatggac
||||||......  =>    ||||||x.....
GCGGAGgcggac        GCGGAGgcggac
```

At each base, try extending alignment; is total score still above threshold?

How *alignment* works, and why indels are the devil

There are many alignment strategies, but most work like this:

```
GCGGAGatggac        GCGGAGatggac
||||||......  =>    ||||||xx....
GCGGAGgcggac        GCGGAGgcggac
```

Each mismatch *costs*.

How *alignment* works, and why indels are the devil

Insertions/deletions introduce *lots* more ambiguity:

```
GCGGAGagaccaacc      GCGGAGag-accaacc
||||||          =>  ||||||
GCGGAGggaaccacc      GCGGAGggaacc-acc

GCGGAGagaccaacc      GCGGAGaga-ccaacc
||||||          =>  ||||||
GCGGAGggaaccacc      GCGGAGggaacca-cc
```

Mapping short reads, again

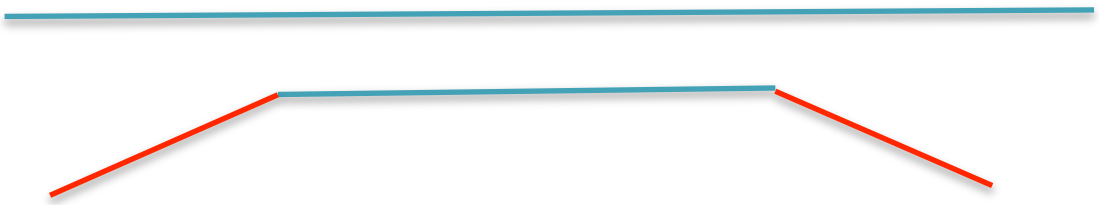
- What's hard about mapping
- Three mapping programs
- Decisions to be made
- Color space issues

Mapping, defined

- Exhibit A: 20m+ reads from genome/ transcriptome.
- Exhibit B: related genome/transcriptome, aka "the reference"
- Goal: assign all reads to location(s) within reference.
- Req'd for resequencing, ChIP-seq, and mRNAseq

Want *global*, not *local*, alignment

- You do not want matches *within* the read, like BLAST would produce.



- Do not use BLAST!

Mapping is "pleasantly parallel"

- Goal is to assign each individual read to location(s) within the genome.

- So, you can map each read *separately*.

What makes mapping challenging?

- Volume of data
- Garbage reads
- Errors in reads, and quality scores
- Repeat elements and multicopy sequence
- SNPs/SNVs
- Indels
- Splicing (transcriptome)

Volume of data

- Size of reference genome is not a problem: you can load essentially all genomes into memory (~3 gb).

- However, doing *any* complicated process 20m times is generally going to require optimization!

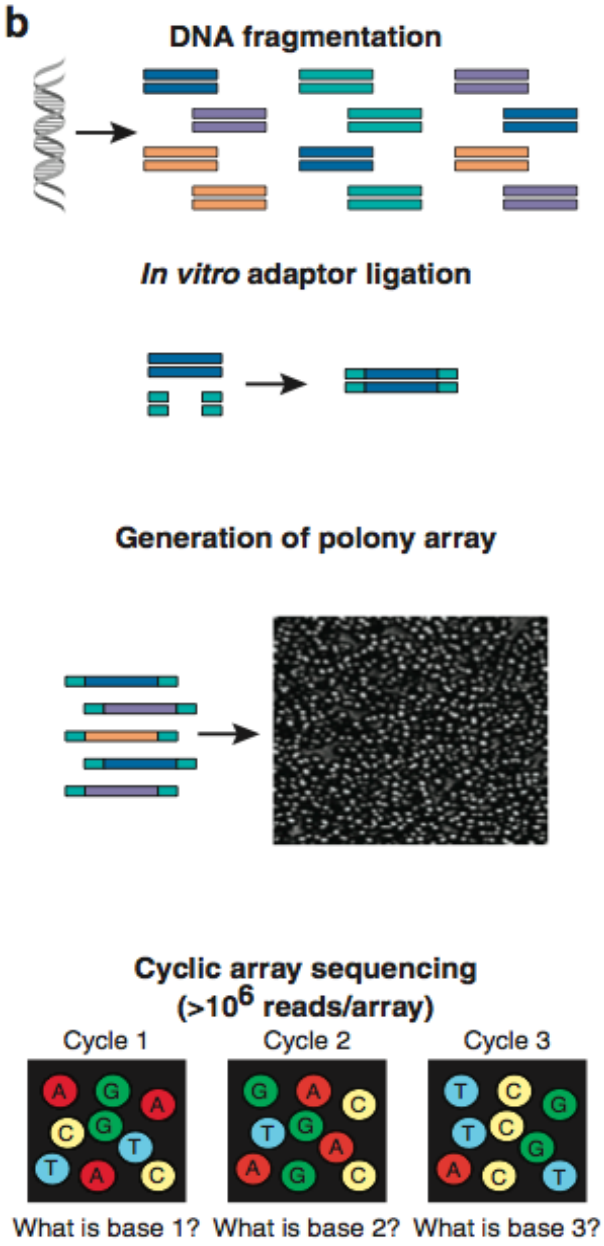
Garbage reads

Overlapping polonies result in mixed signals.

These reads will not map to anything!

Used to be ~40% of data.

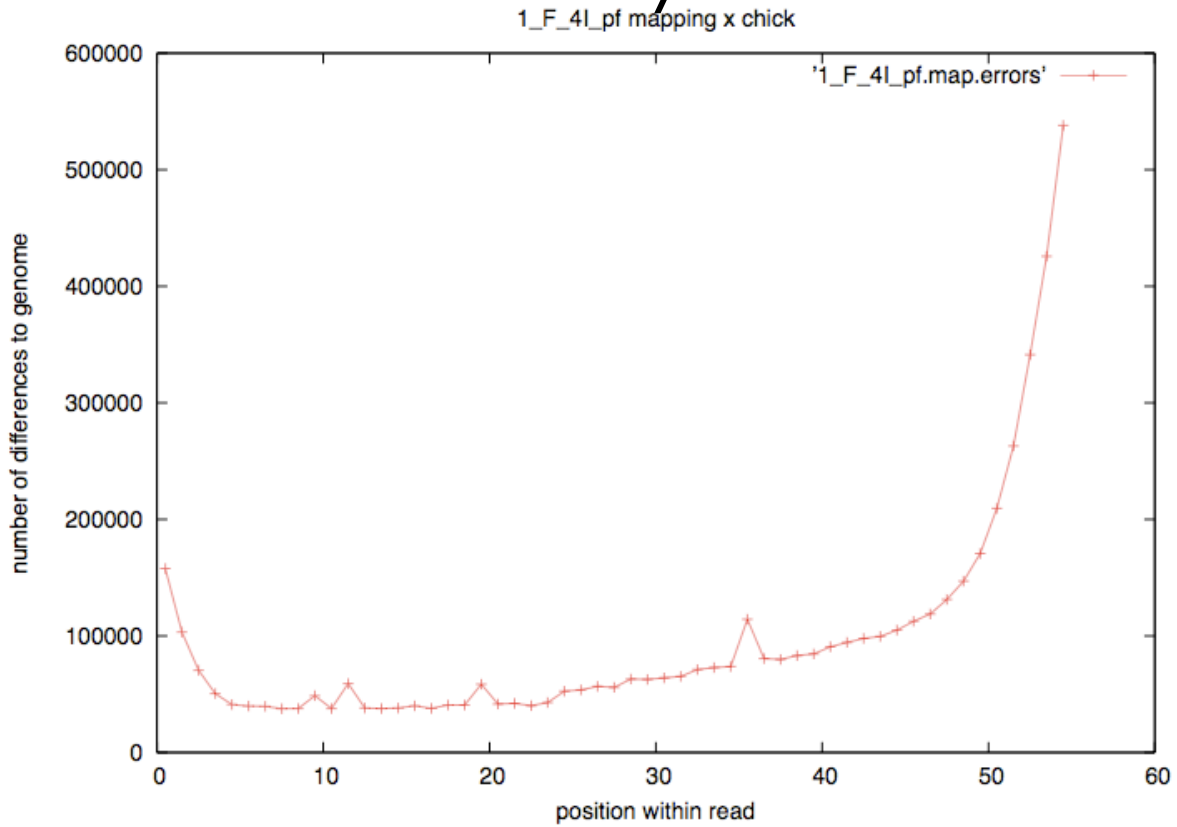
Increasingly, filtered out by sequencing software.



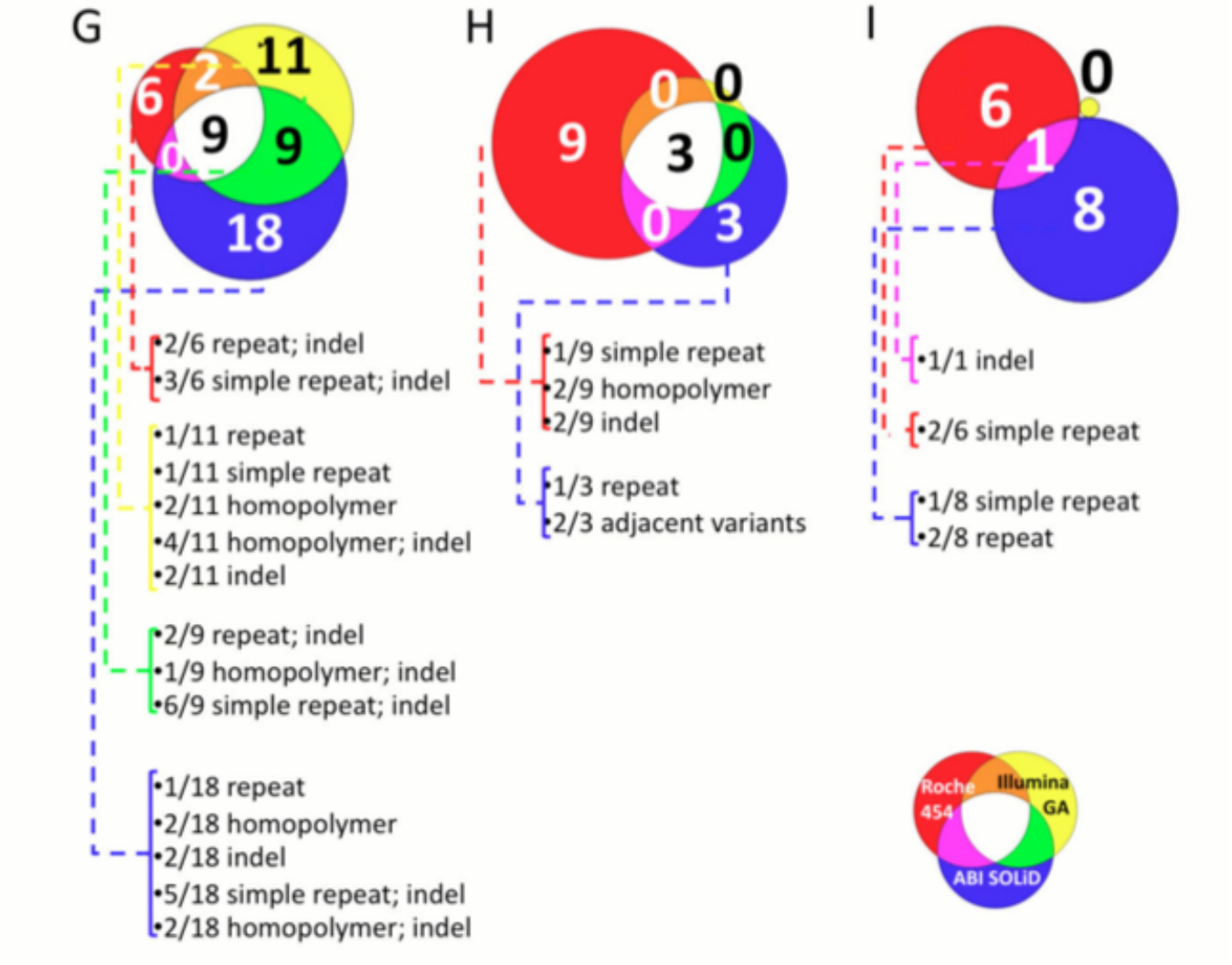
Shendure and Ji, Nat Biotech, 2008

Errors in reads

When mapping, a mismatch is not necessarily "real".



Technology-specific bias



Harismendy et al., Genome Biol. 2009, pmid: 19327155

Errors in reads

- Quality scores are based on Sanger sequencing-style quality scores: per base.
- But 454 is subject to different biases than Illumina, and the biases are not necessarily base-by-base (think: homopolymer runs)
- It's still not clear what quality scores on next-gen data mean.

Repeat/multi-copy elements

- Multi-copy sequence makes it impossible to map all reads uniquely.

- Repeats are particularly bad, because there are (a) lots of them and (b) they vary in sequence. They therefore may "attract" reads depending on what optimizations/heuristics you use.

SNP/SNVs

- Genuine mismatches between reference and sequence *do* exist, of course.
  - Polymorphism
  - Diploidy
  - Population studies
- You want to map these reads!
- Fast heuristic approaches exist, based on fuzzy matching.
- However, they are still biased towards mapping exact matches.
  - This can be a problem for allelotyping and population studies.
  - Likit will discuss next week.

Indels

- Remember, they are the devil…?
- Complicate mapping heuristics
- Complicate *statistics*
- At least at the moment, cause substantial increase in compute time: many more decisions to be made.

Indels: ambiguity & decisions…

```
TGACGATATGGCGATGGACTGGACG
|x|||||||||||x||x|x||||||
TcACGATATGGCGgTGaA-TGGACG

TGACGATATGGCGATGGACTGGACG
|x|||||||||||x|x||x||||||
TcACGATATGGCGgT-GAaTGGACG
```

Splice sites

- If you are mapping transcriptome reads to the genome, your reference sequence is different from your source sequence!
- This is a problem if you don't have a really good annotation.

- Main technique: try to map across splice sites, build new exon models.
- Another technique: assembly.

Three mapping programs

- maq
- Bowtie
- BWA

All open source; welcome to try all.

Bowtie is probably the fastest so we'll use that for examples. It may or may not be evil.

(There are many more, too.)

maq

- Ungapped alignment, up to two mismatches in first 28 bases.
- Aligns everything to the best of its ability, then estimates likelihood that it's a genuine alignment.
- Good for SNP calling (next week)

Bowtie

- Not indel-capable.

- Designed for:
  - Many reads have one good, valid alignment
  - Many reads are high quality
  - Small number of alignments/read

a.k.a. “sweet spot” :)

BWA

- Built by authors of maq
- Uses similar strategy to Bowtie, but does gapped alignment.
- 10x faster than maq.
- Newest, hottest tool.

Decisions to be made by you

- How many mismatches to allow?
  - Vary depending on biology & completeness of reference genome
- Report how many matches?
  - Are you interested in multiple matches?
- Require best match, or first/any that fit criteria?
  - It can be much faster to find *first* match that fits your criteria.

All of these decisions affect your results and how you treat your data.

Mapping best done on *entire reference*

- May be tempted to optimize by doing mapping to one chr, etc. "just to see what happens"
- Don't.
- Simple reason: if you allow mismatches, then many of your reads will match erroneously to what's present.

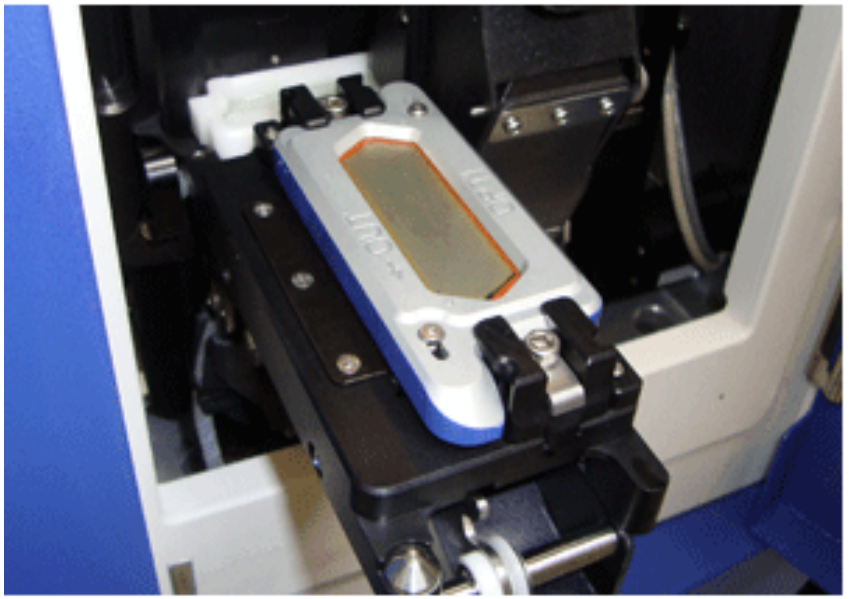
Look at your mapping

Just like statistics ☺

This is a lot of what we'll be doing today.

ABI *SOLiD™ platform & color space

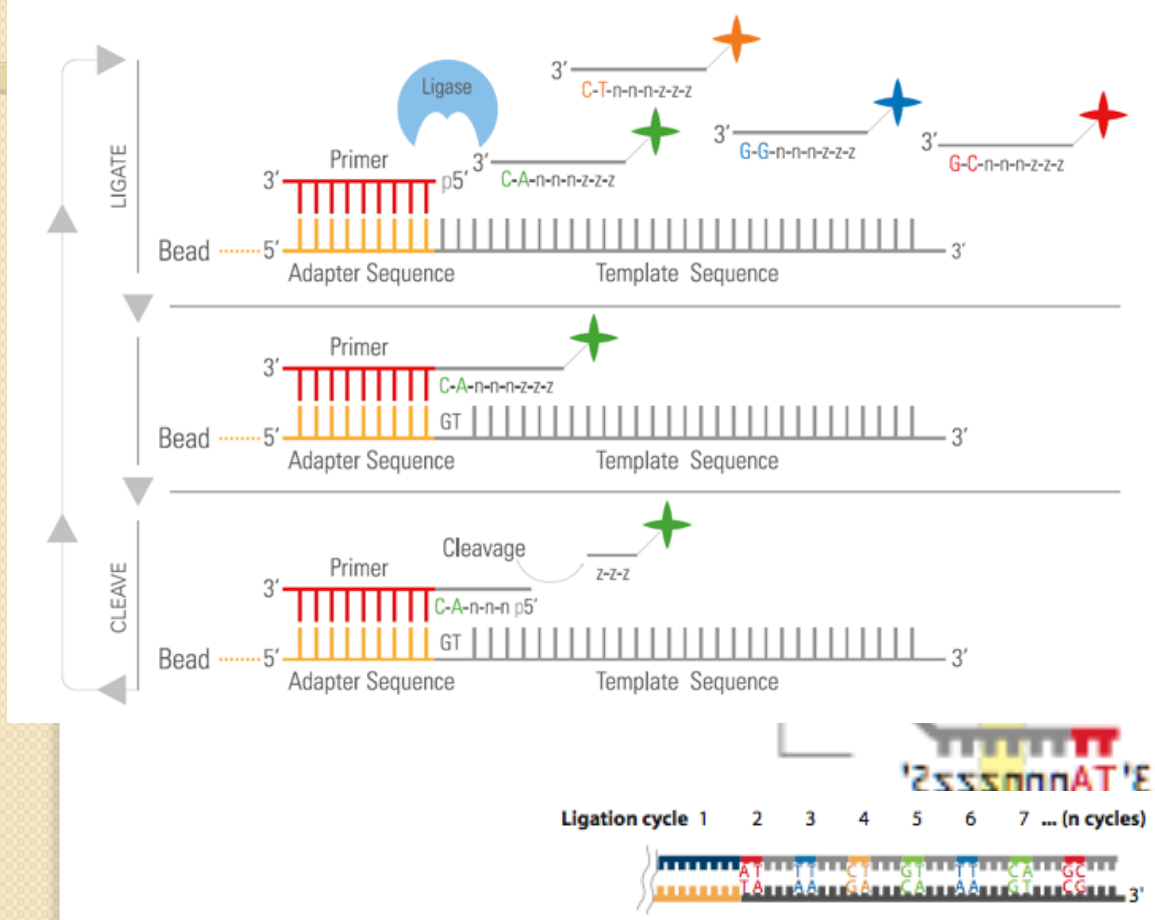
- High throughput, 60+ gigabases and >1 billion tags per run. (SOLiD ™ 3 plus)
- Sequencing by ligation.
- Di-base encoding
- Support single or paired-end tag.
- 99.94% accuracy.



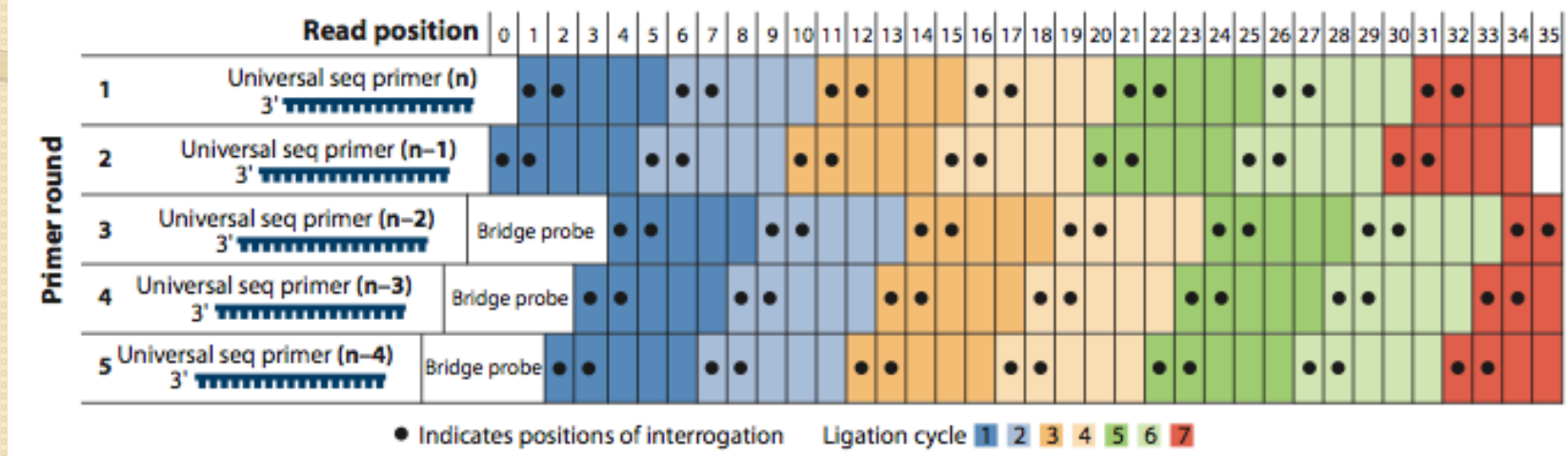
*SOLiD: Sequencing by Oligo Ligation and Detection.

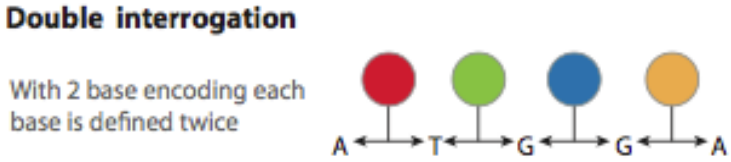
http://www3.appliedbiosystems.com/AB_Home/applicationstechnologies/SOLiDSystemSequencing/overviewofsolidsystem/index.htm

Di-base encoding



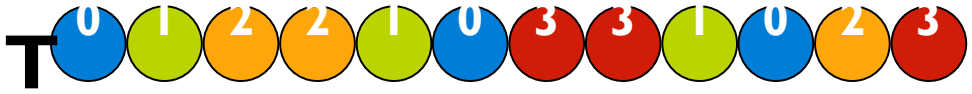
Double interrogation





With di-base encoding and double interrogation, 99.94% accuracy is claimed.

Sequencing Error and SNP Detection



Perfect alignment

```
G:  TGAGTTATGGAT
R:TTGAGTTATGGAT
   012210331023
```

Error

```
G:  TGAGTTATGGAT
R:TTGAGTCGCAAGC
   012212331023
```

SNP alignment

```
G:  TGAGTTATGGAT
R:TTGACTTATGGAT
   012120331023
```

**Mapping must be done in color space with color space aware mapper.

Color space: mapping v assembly

- Mapping is more complex, because of indels.
- New (or modified) software is required that is color-space aware.
- In contrast, assembly can operate in color space!

Perfect alignment

```
G: TGAGTTATGGAT
R:TTGAGTTATGGAT
   012210331023
```

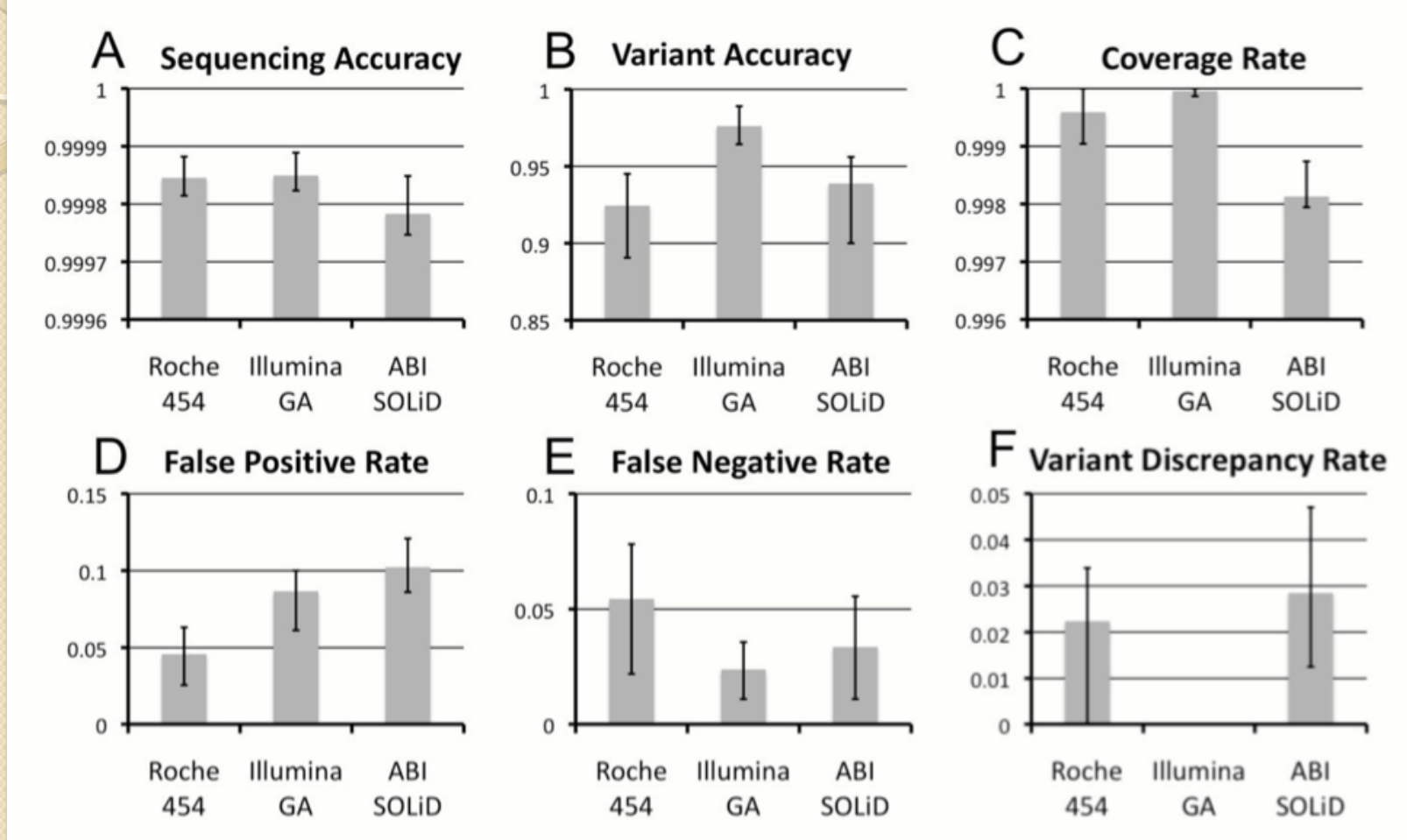
Error

```
G: TGAGTTATGGAT
R:TTGAGTCGCAAGC
   012212331023
```

SNP alignment

```
G: TGAGTTATGGAT
R:TTGACTTATGGAT
   012120331023
```

Is accuracy higher with SOLiD?



Harismendy et al., Genome Biol. 2009, pmid: 19327155

Today

- Morning & afternoon:
  - Installing Bowtie & running it on a small bit of data.
  - Visualizing mappings on the command line.
  - Running a full mapping
- Evening:
  - Reading bowtie's docs
  - Other mapping tools
  - Playing with real data
  - A Challenge