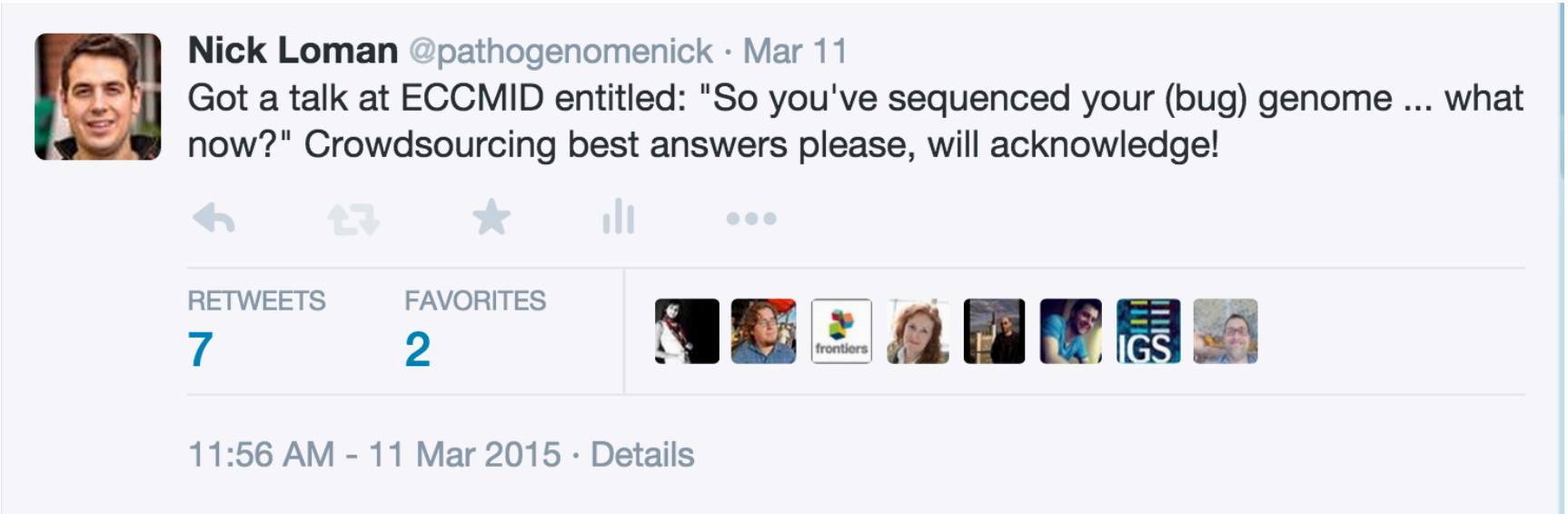
Mapping

Isheng Jason Tsai

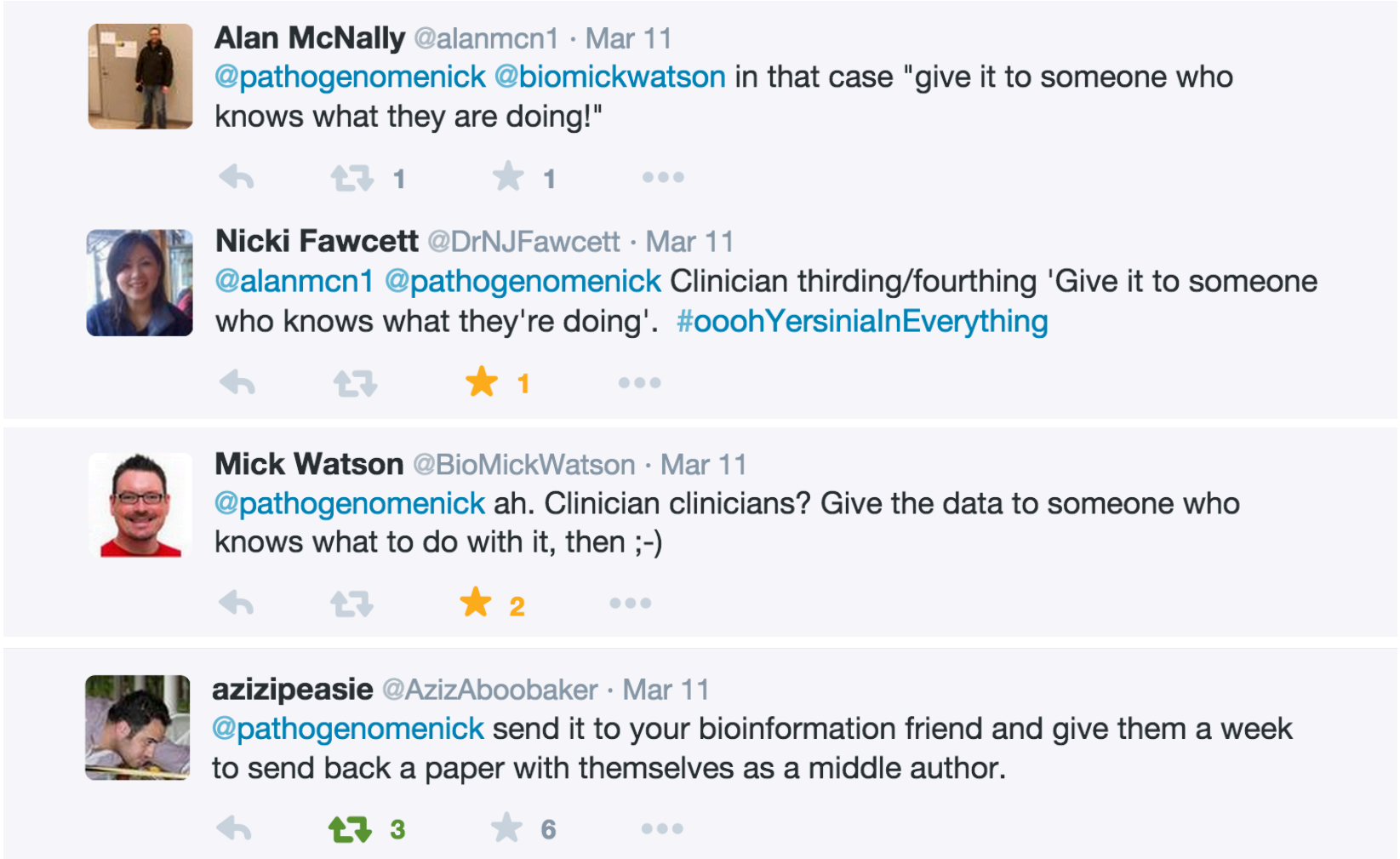
Introduction to NGS Data and Analysis
Lecture 4

Preface



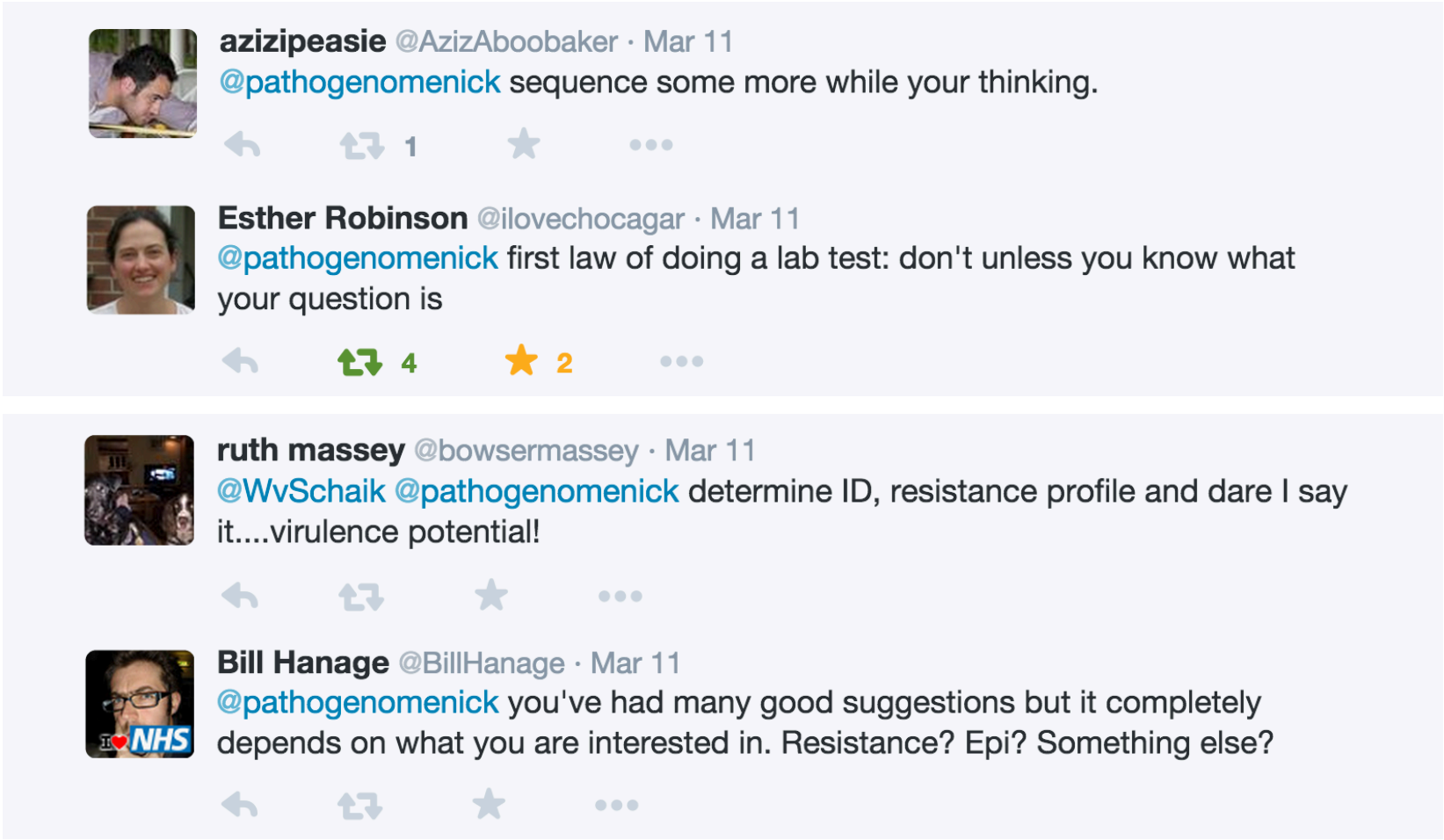
https://twitter.com/pathogenomenick/status/575626319616176128

We all know…



https://twitter.com/pathogenomenick/status/575626319616176128

Logical answer



azizipeasie @AzizAboobaker · Mar 11
@pathogenomenick sequence some more while your thinking.
1

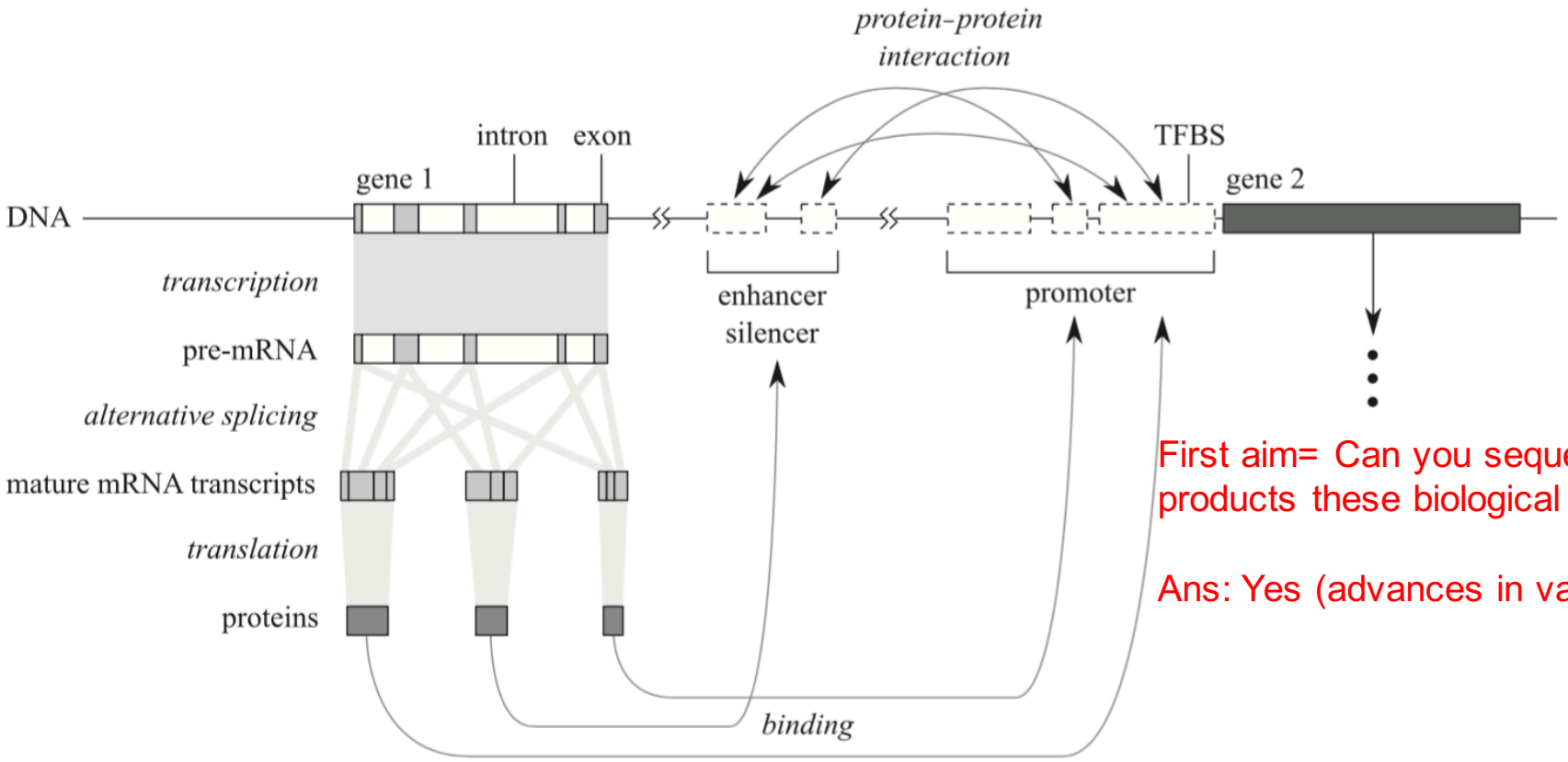
Esther Robinson @ilovechocagar · Mar 11
@pathogenomenick first law of doing a lab test: don't unless you know what your question is
4 2

ruth massey @bowsermassey · Mar 11
@WvSchaik @pathogenomenick determine ID, resistance profile and dare I say it....virulence potential!

Bill Hanage @BillHanage · Mar 11
@pathogenomenick you've had many good suggestions but it completely depends on what you are interested in. Resistance? Epi? Something else?

https://twitter.com/pathogenomenick/status/575626319616176128

High throughput sequencing applications

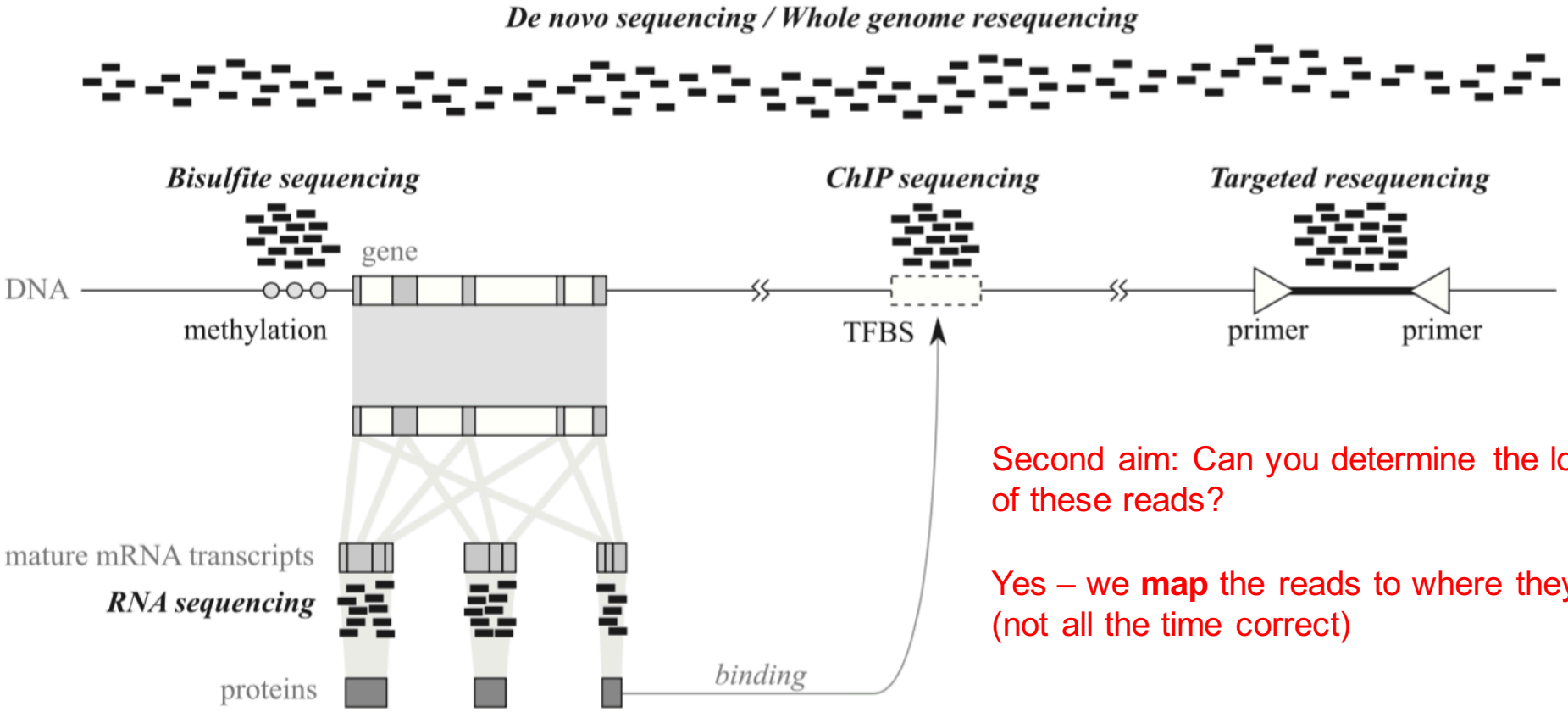


First aim= Can you sequence the products these biological processes?

Ans: Yes (advances in various experiments)

Figure 1.1 A schematic illustration of the central dogma. Gene 1 has three alternatively spliced transcripts. The relative expression of such transcripts affects the regulatory modules of gene 2, and eventually its expression. Definitions are given in Section 1.1.

High throughput sequencing applications



Second aim: Can you determine the location of these reads?

Yes – we **map** the reads to where they originate (not all the time correct)

Figure 1.2 A schematic summary of high-throughput sequencing applications. Details are described in Section 1.3.

Makinen *et al*

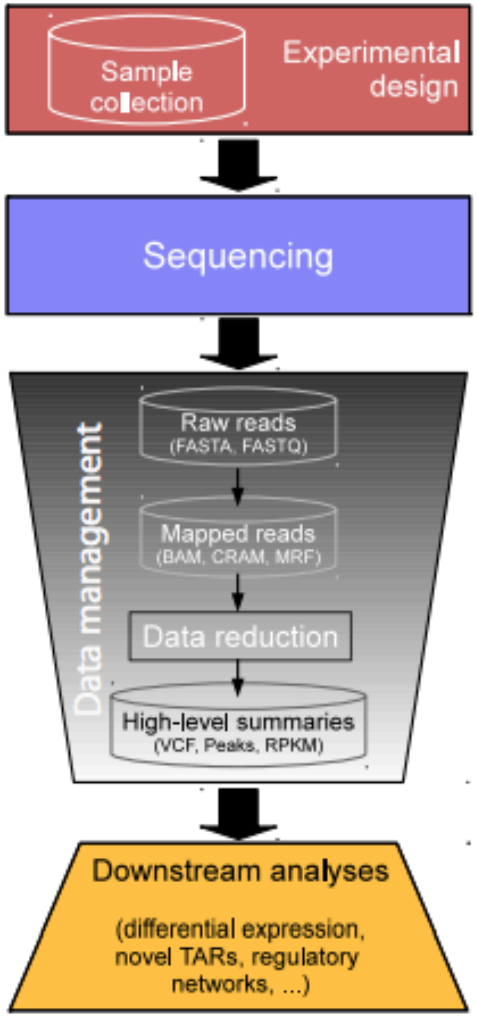
Types of sequencing

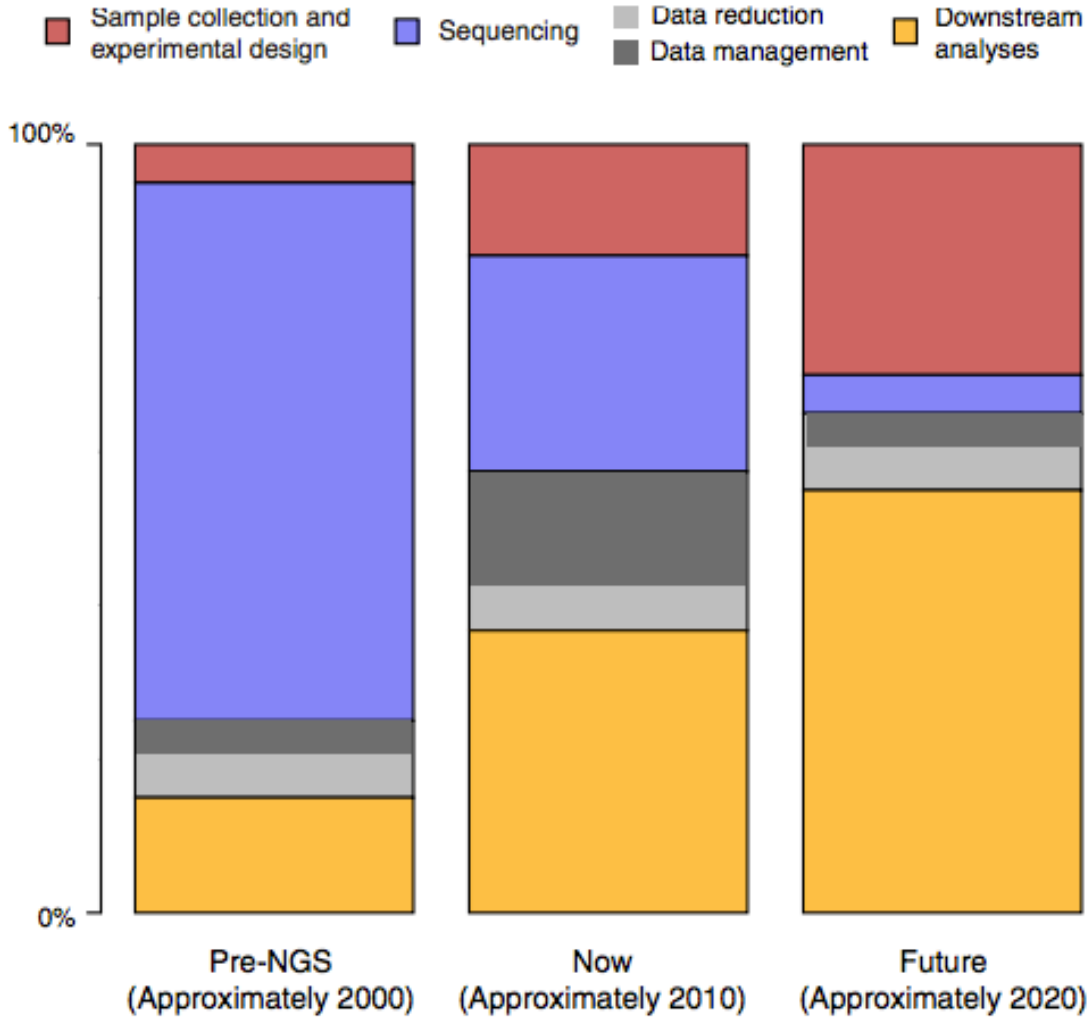
Paired-end Illumina (typically 150 – 400 bases)

~~Single-end Ion Torrent (typically 300-400 bases) (bad in my own experience)~~

Pacific Biosciences or Oxford Nanopore (long reads)

The real cost of sequencing



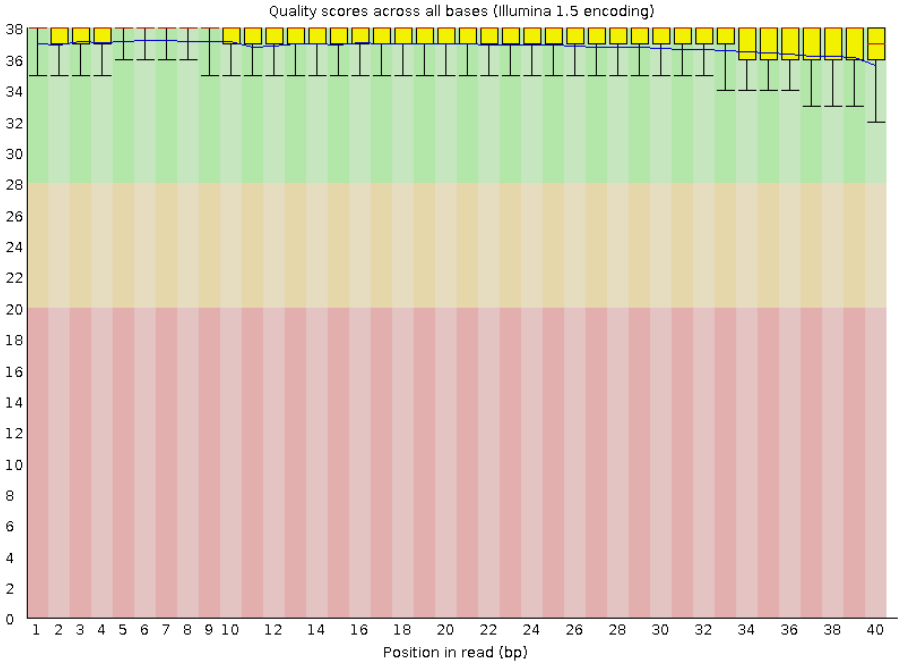


QC first - always always the first step

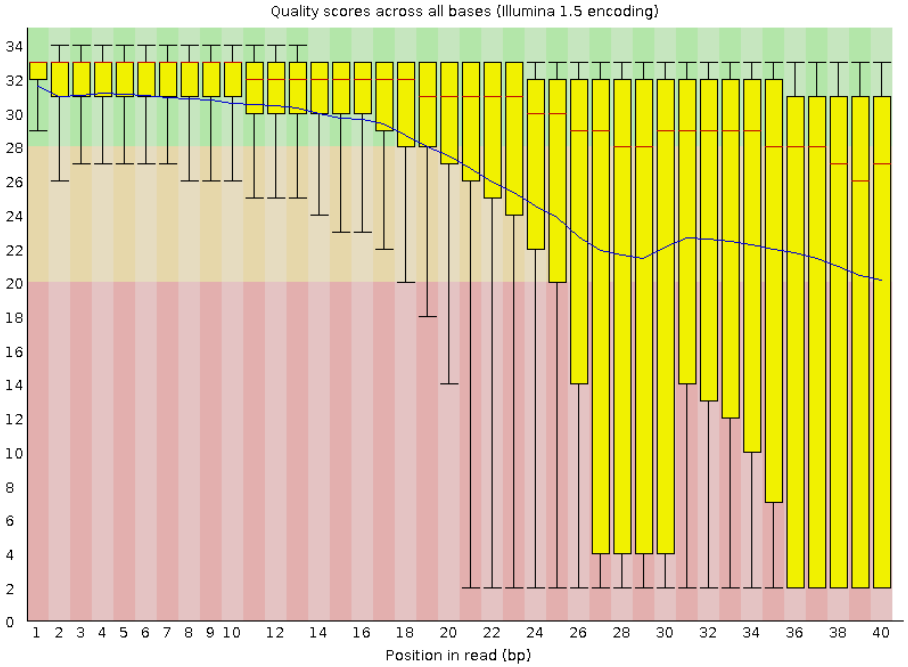
- Contamination! *
- Is it of good quality?
  - Read quality
  - Adaptor contamination
  - Insert size distribution
  - PCR duplicate rate

- Is it your species or someone else's (sample swap)?

Sequence quality - FastQC



Good (unlikely)



Bad

http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Type of adaptors (mate pair)

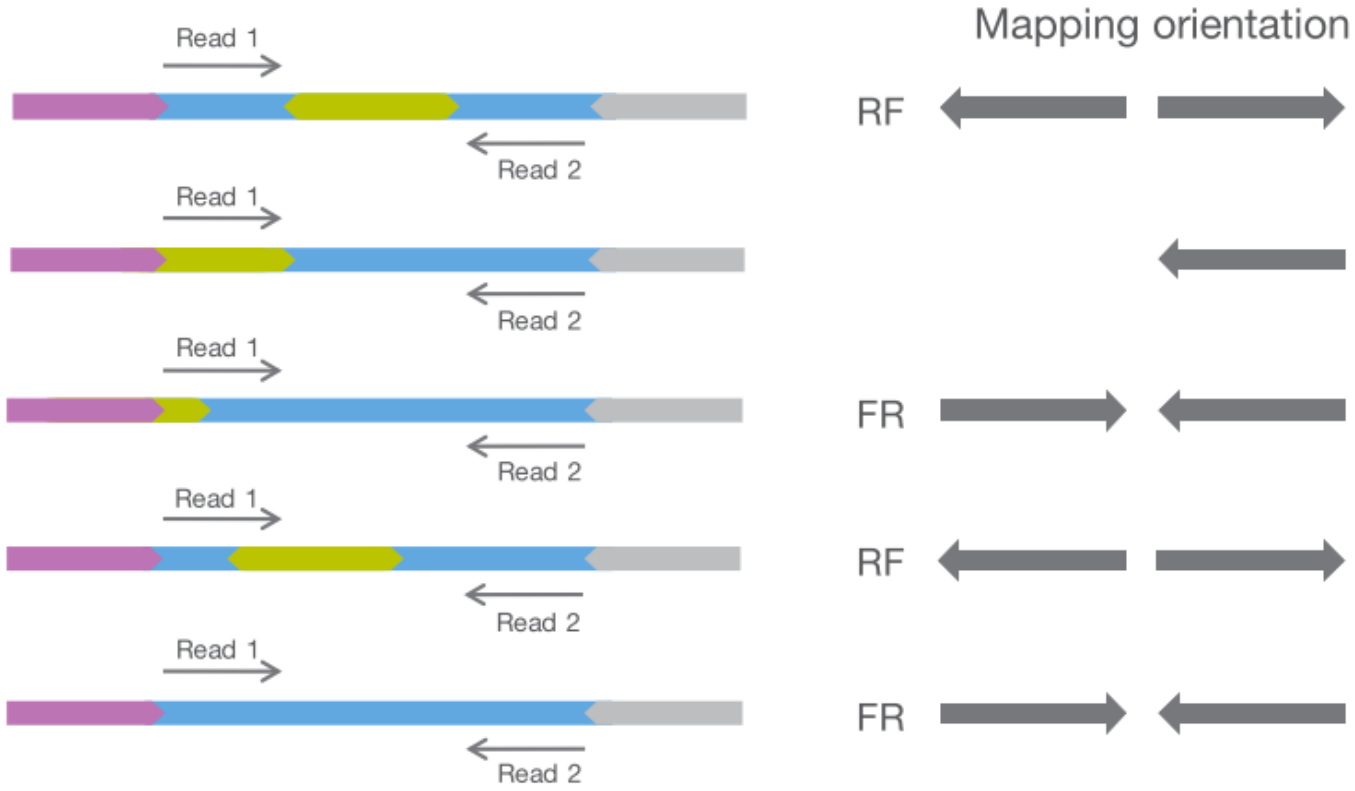
Best case
(a bit of adaptor at the end)

Bad

Okay

Okay

Totally fail to circula



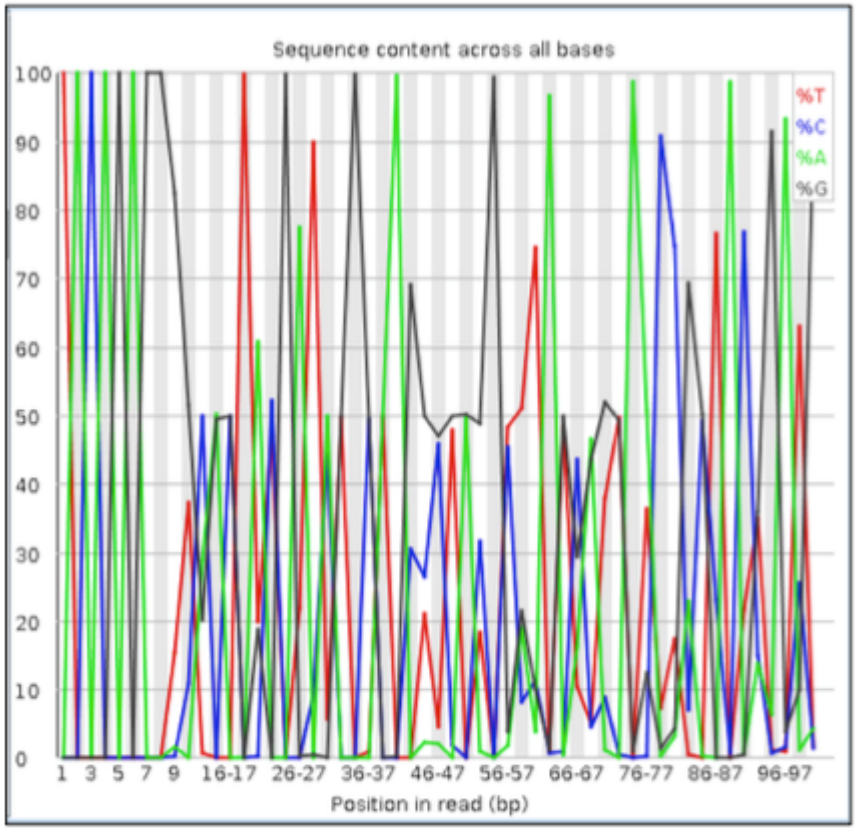
Composition of example library templates from a mate pair experiment. For each example (a–e), the position of the junction adapter sequence is shown in green and the mapping orientation (either FR or RF, 'forward-reverse' and 'reverse-forward', respectively) of the resulting read pairs is shown to the right. Sections of genomic DNA sequence are shown in blue and the TruSeq adapter sequences are shown in purple and grey.
Amplification/sequencing primer adapters are shown in grey and purple.

Illumina

Basically the adaptor sequence can appear everywhere (but in a logic way)

FastQC will offer some insights in adaptor

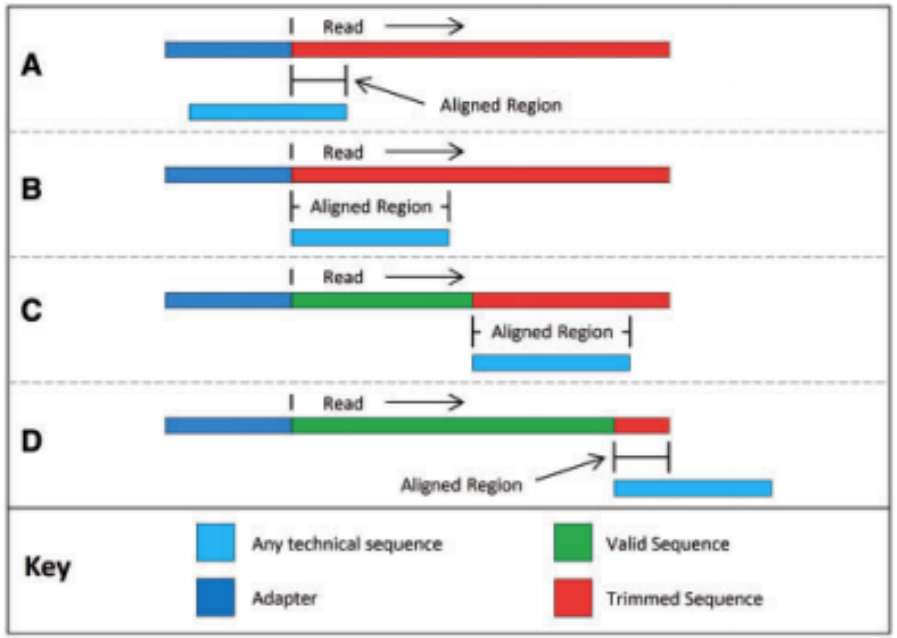
TACAGAGG overrepresented – what is it?

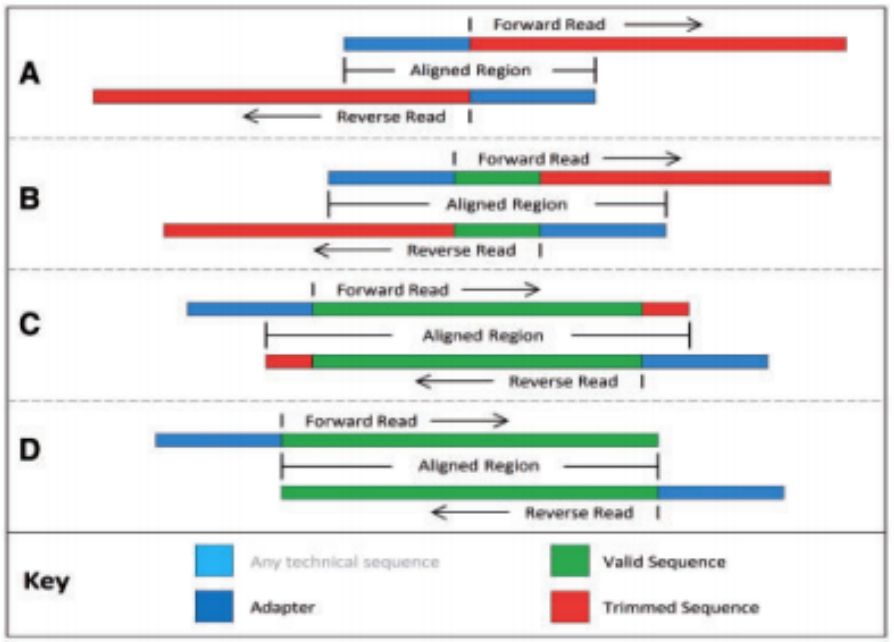


Overrepresented sequences

Sequence	Count	Percentage	Possible Source
AGCAGCATTGTACA...	3398	3.398	No Hit
TACAGTCCGACGAT...	1814	1.814	Illumina PCR Prime...
TCTACAGTCCGACG...	1570	1.57	RNA PCR Primer, In...
TATTGCACTTGTCCC...	1421	1.421	No Hit
TTCTACAGTCCGAC...	1181	1.181	RNA PCR Primer, In...
CTACAGTCCGACGA...	1168	1.168	Illumina PCR Prime...
CATTGCACTTGTCTC...	839	0.839	No Hit
ACAGTCCGACGATC...	835	0.835	RNA PCR Primer, In...
AGTTCTACAGTCCG...	648	0.648	Illumina PCR Prime...
AAAGTGCTGCGACA...	491	0.491	No Hit
TCGTATGCCGTCTT...	465	0.465	Illumina Single En...
CAGTCCGACGATCT...	436	0.436	Illumina PCR Prime...
TNNNNNNNNNNNN...	392	0.392	No Hit
TAGCTTATCAGACT...	388	0.388	No Hit
TATTGCACTCGTCC...	366	0.366	TruSeq Adapter, I...
ACCGGGCGGAAAC...	357	0.357	No Hit
ANNNNNNNNNNNN...	355	0.355	No Hit
GTTCTACAGTCCGA...	353	0.353	Illumina PCR Prime...
AAGTGCTGCGACAT...	341	0.341	No Hit

Trimmomatic for quality and adaptor trimming (many other tools also exist)

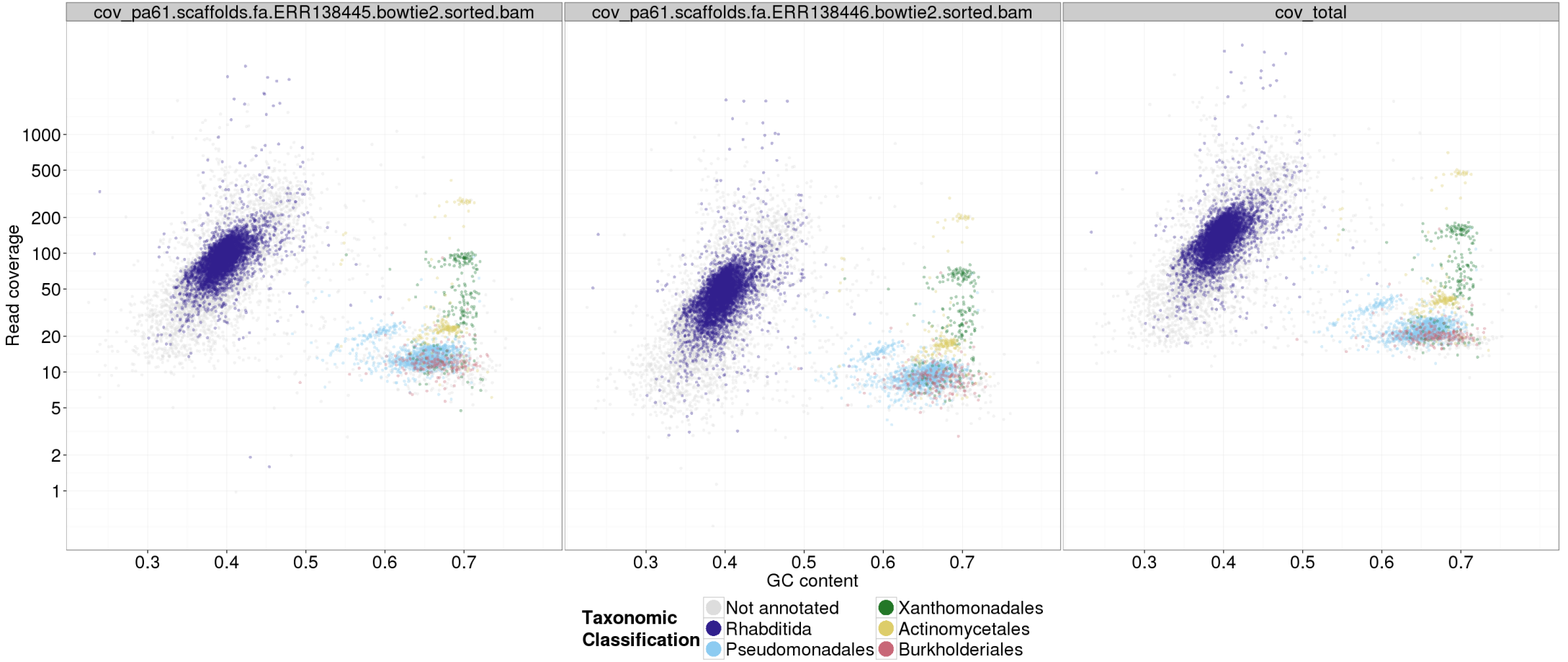




Trimmomatic: a flexible trimmer for Illumina sequence data
www.ncbi.nlm.nih.gov › ... › PubMed Central (PMC) ▾ 翻譯這個網頁
由 AM Bolger 著作 - 2014 - 被引用 850 次 - 相關文章
2014年4月1日 - However, short partial adapter sequences, which often occur at the ...
This scenario would result in the trimming of both reads as illustrated.

Bolger *et al.*, (2014)

Check what your samples contain - Blobology



https://github.com/blaxterlab/blobology

Source of contamination

- Difficult to remove (gut from microorganisms)
- Fail to remove
- Not careful
- Bad company
- Sequencer carry over (from previous run)
- Sample (barcode) mix up
- Or simply bad day (not your fault)

Salter *et al. BMC Biology* 2014, **12**:87
http://www.biomedcentral.com/1741-7007/12/87

RESEARCH ARTICLE Open Access

Reagent and laboratory contamination can critically impact sequence-based microbiome analyses

Sample storage matters (case of humans)

3 months storage resulted in less efficient DNA extraction

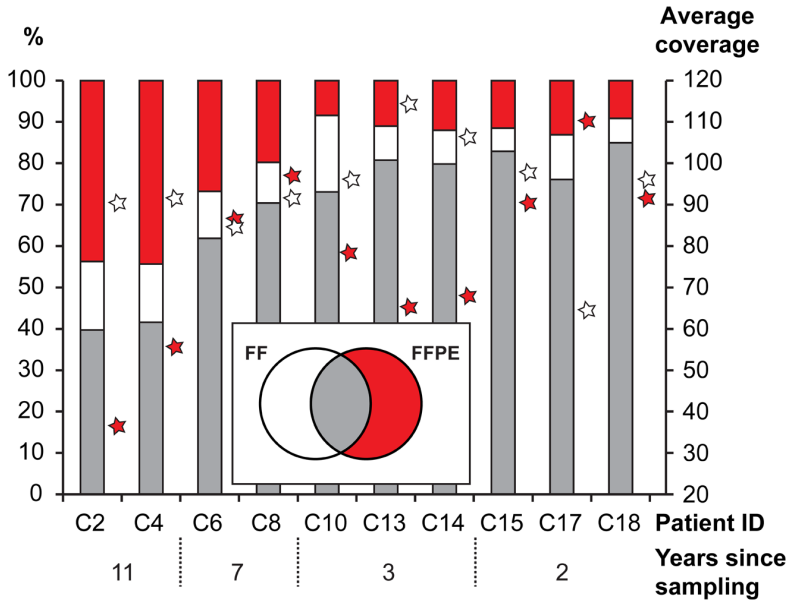
- High fragmentation: loss of material
- Decrease in library complexity
- High increase in PCR duplicates, 60-85% for FFPE vs. 30% for FF

C > U deamination is a common cause of artifacts

- U-tolerant polymerase didn't help
- Pattern, T <> C, A <> G transition

The fraction of mapped reads decreases with storage time

- Increase in partial mappings
- Increase in gapped mappings



Hedegaard et al. 2014

Mapping approach (the easier? way)

Mapping

Mapping is **aligning** the read to where **the most likely origin** within the reference/assembly

Sequence alignment has not changed and will remain a classic problem
Tradeoffs of speed, accuracy and sensitivity

Sequence data we want to map:
- Mostly nucleotide

Very short evolutionary distances (human to reference, isolate/strain to reference, 'slightly diverged' strain will map less)
Very short – needs faster processing per read (BLAST is too slow!)

There are some assumptions to make alignment process faster
(like allows most 2 mismatches)

How?

Brute force comparison
Smith-Waterman
Suffix Tree
Burrows-Wheeler Transform

Brute force

1.

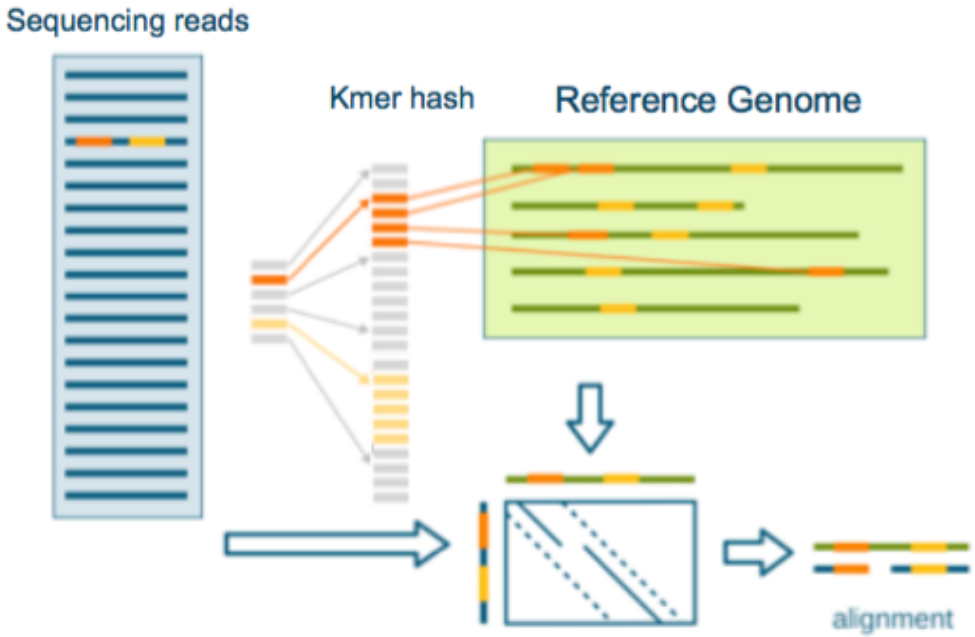
2.

3.

4.

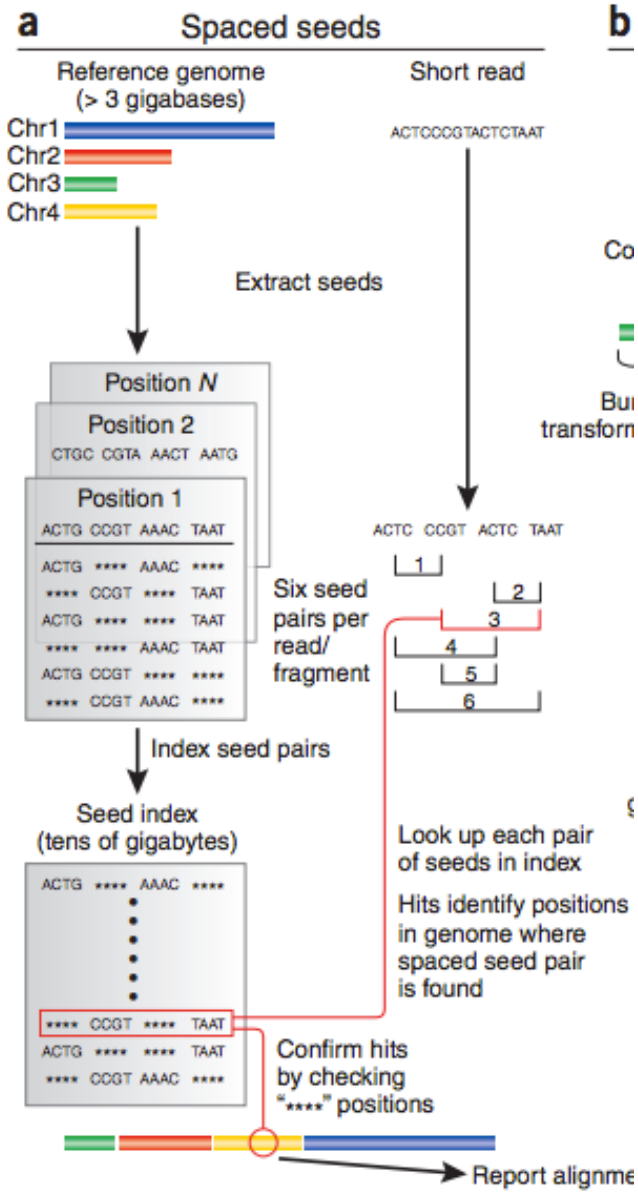
Credit: Mike Zody

Mapping (hash table)



Identify all the seeds in the index
Determine the most likely location
Perform Smith-Waterman alignment to fully align
Output (important)

Example: BLAST, MAQ (Heng Li 2008)

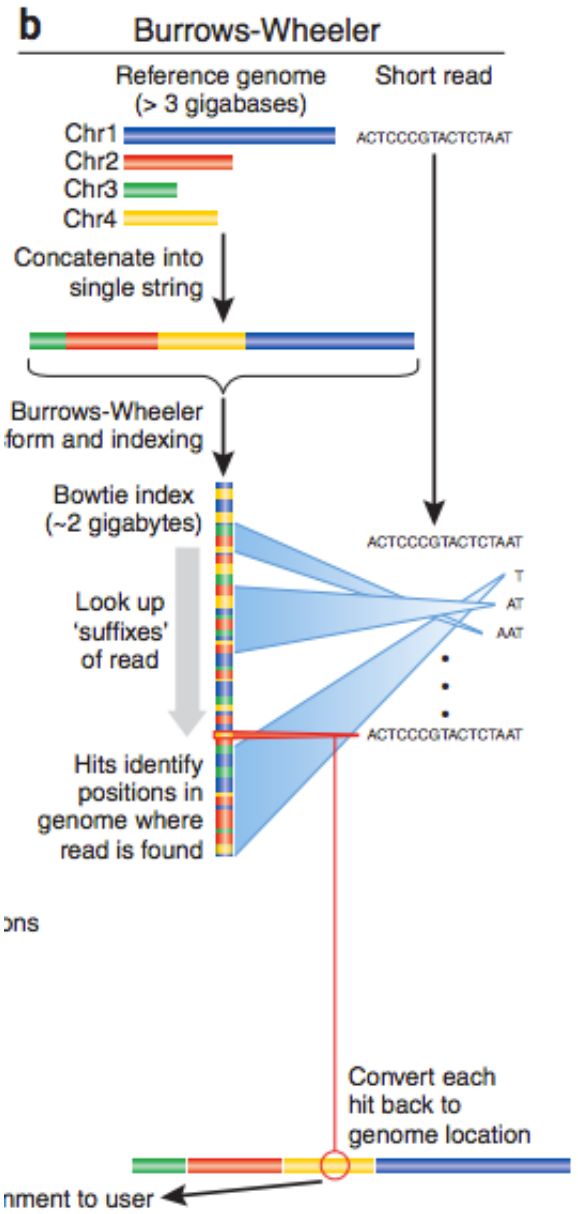


Suffix tree

GACCTCATCGATCCCACTG

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

Credit: Mike Zody ; Trapnell *et al* (2009)



Burrows-Wheeler Transform

A transformation that will result in many repeated characters

This means it's easy to compress

And an elegant way to search!

Transformation				
Input	All Rotations	Sorting All Rows into Lex Order	Taking Last Column	Output Last Column
^BANANA|	^BANANA| |^BANANA A|^BANAN NA|^BANA ANA|^BAN NANA|^BA ANANA|^B BANANA|^	**A**NANA|^B **AN**A|^BAN **A**|^BANAN **B**ANANA|^ **N**ANA|^BA **N**A|^BANA ^BANANA| |^BANANA	ANANA|^**B** ANA|^BA**N** A|^BANA**N** BANANA|^ NANA|^B**A** NA|^BAN**A** ^BANANA| |^BANAN**A**	BNN^AA|A

Wiki

F	...	L
ot look upon his like again.	...	n
ot look upon me; Lest with th	...	n
ot love on the wing,-- As I p	...	h
ot love your father; But that	...	n
ot made them well, they imita	...	n
ot madness That I have utter'	...	n
ot me'? Ros. To think, my lor	...	n
ot me; no, nor woman neither,	...	n
ot me? Ham. No, by the rood,	...	g
ot mend his pace with beating	...	n
ot mine own. Besides, to be d	...	n
ot mine. Ham. No, nor mine no	...	n
ot mock me, fellow-student. I	...	n
ot monstrous that this player	...	n
ot more like. Ham. But where	...	n
ot more native to the heart,	...	n
ot more ugly to the thing tha	...	n
ot more, my lord. Ham. Is not	...	j
ot move thus. Oph. You must s	...	n
ot much approve me.--Well, si	...	n

Fig. 1.5. Part of the BWT sorted list for Shakespeare's Hamlet

Adjeroh

banana

banana$		$banana
anana$b		a$banan
nana$ba	sort →	ana$ban
ana$ban		anana$b
na$bana		banana$
a$banan		nana$ba
$banana		na$bana

BWT(banana) =
annb$aa

Tends to put runs of the same character together.

Makes compression work well.

"bzip" is based on this.

Credit: Carl Kingsford

Credit: Mike Zody

BWT – a summary

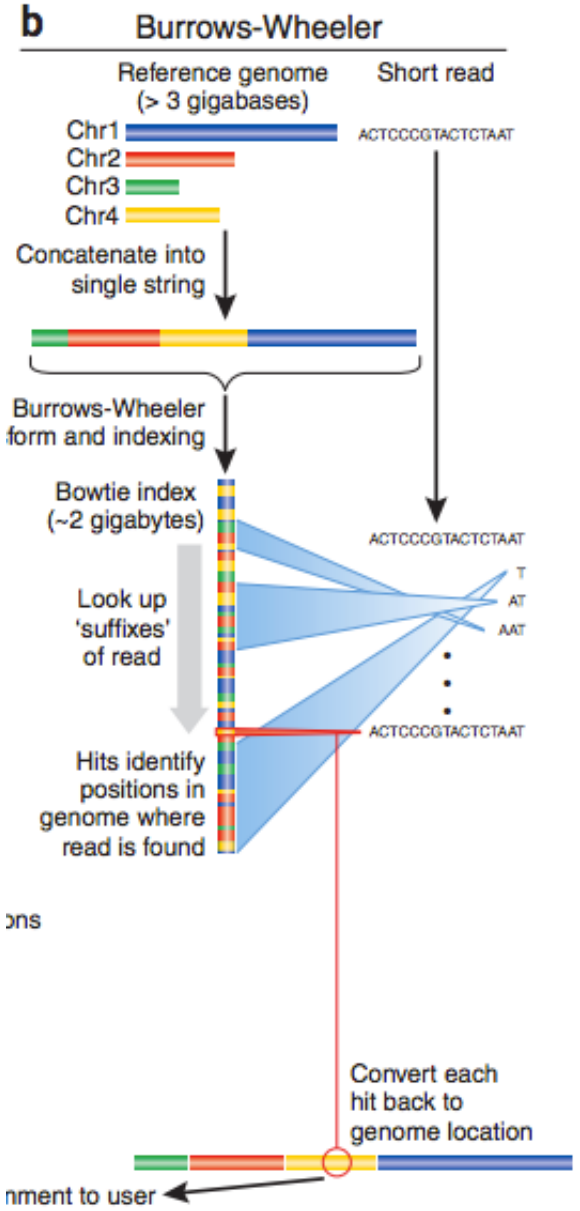
Stores all possible suffixes to enable fast string matching

Much smaller memory footprint than hash table
(hash table need to store all different kmers)

Examples:
MUMMER, bwa, bowtie2

Still need local alignment in final step

Trapnell *et al* (2009)



Hash table vs. BWT

BWT

Hash table

Table 1 A selection of short-read analysis software

Program	Website	Open source?	Handles ABI color space?	Maximum read length
Bowtie	http://bowtie.cbcb.umd.edu	Yes	No	None
BWA	http://maq.sourceforge.net/bwa-man.shtml	Yes	Yes	None
Maq	http://maq.sourceforge.net	Yes	Yes	127
Mosaik	http://bioinformatics.bc.edu/marthlab/Mosaik	No	Yes	None
Novoalign	http://www.novocraft.com	No	No	None
SOAP2	http://soap.genomics.org.cn	No	No	60
ZOOM	http://www.bioinfor.com	No	Yes	240

Trapnell *et al* (2009)

Hash table vs. BWT strengths and weaknesses

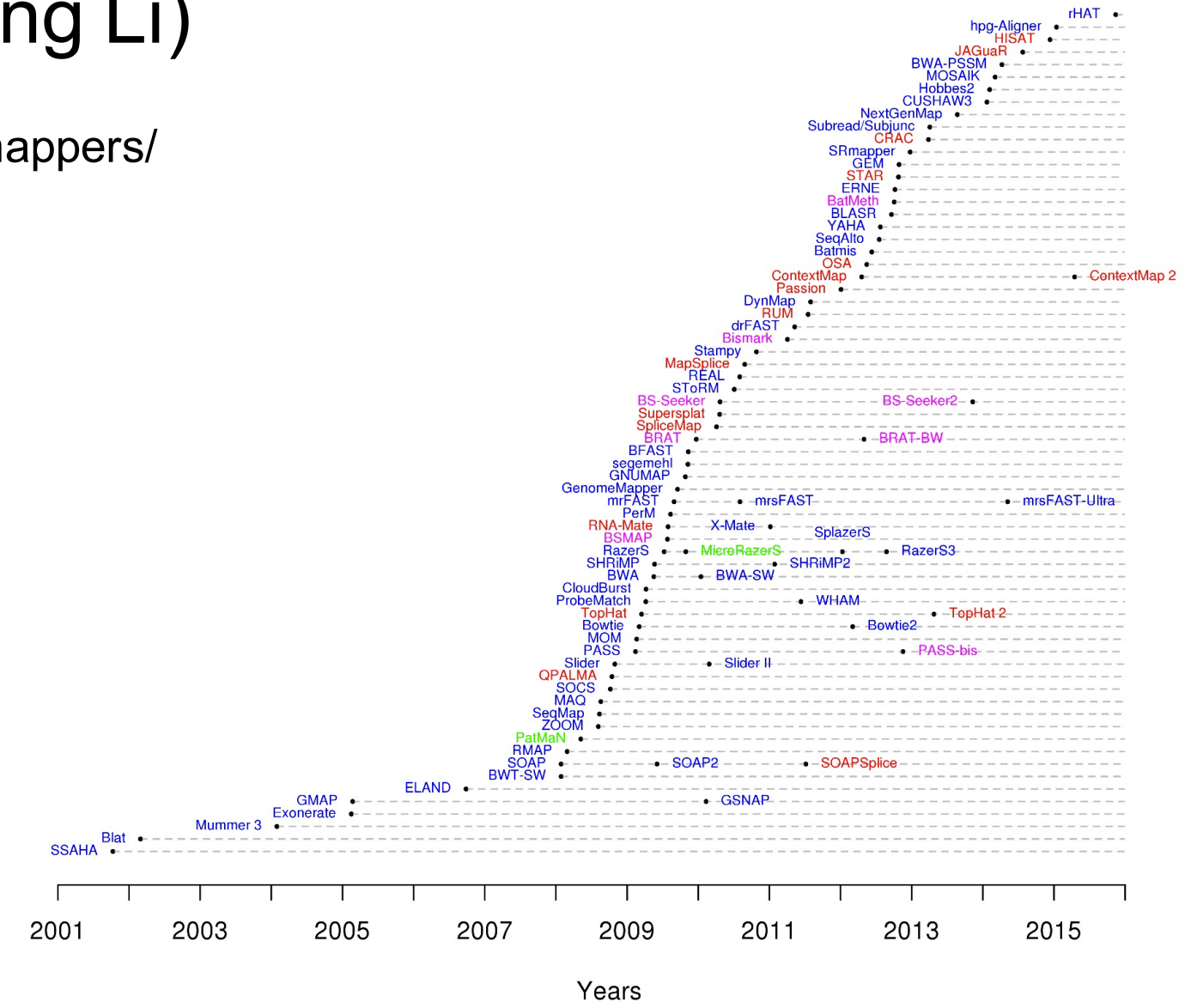
Burrows-Wheeler, e.g. bwa, bowtie
– Fast, esp. (multiple) exact matches
– High sensitivity at repetitive regions
– less robust at high genomic variation

Hashing (overlapping k-mer words, e.g SMALT, Stampy)
– Slower (more memory hungry)
– Less sensitivity at repetitive regions
– tolerate high genomic variation
– partial alignments (junction reads) easier
– Flexible (multiple sequencing platforms)

Ponstingl and Ning

Long live bwa (Heng Li)

http://www.ebi.ac.uk/~nf/hts_mappers/



Choose an aligner

	BWA	BWA-SW	BWA-MEM	Bowtie	Bowtie2	NovaAlign	MOSAIK	Isaac	Tmap
Affiliation	Heng Li			U of Maryland		Novacraft	Boston College	Illumina	Ion Torrent
First Published	2009	2010	2013	2009	2012	-	2014	2013	-
Read Length	<100	70bp-1Mbp		<100	>50				
Gapped Alignments				No					
Trimming				No					
Error Rates Allowed	Low	High	Med	Low	Med	Med	High	Low	Med
Chim Reads	No	Yes	**Yes**	No	Opt	Opt	**Yes**	No	No
Mem Usage	Med	Med	Med	Low	Low	Low	Med	**High**	Med
Speed	Med	Med	Fast	Fast	Fast	**Slow**	Fast	Fast	Fast

Credit: Golden Helix inc

Choose an aligner

Hash based approaches are more suitable for divergent alignments

General rule:

<2% divergence -> BWT

E.g. human samples

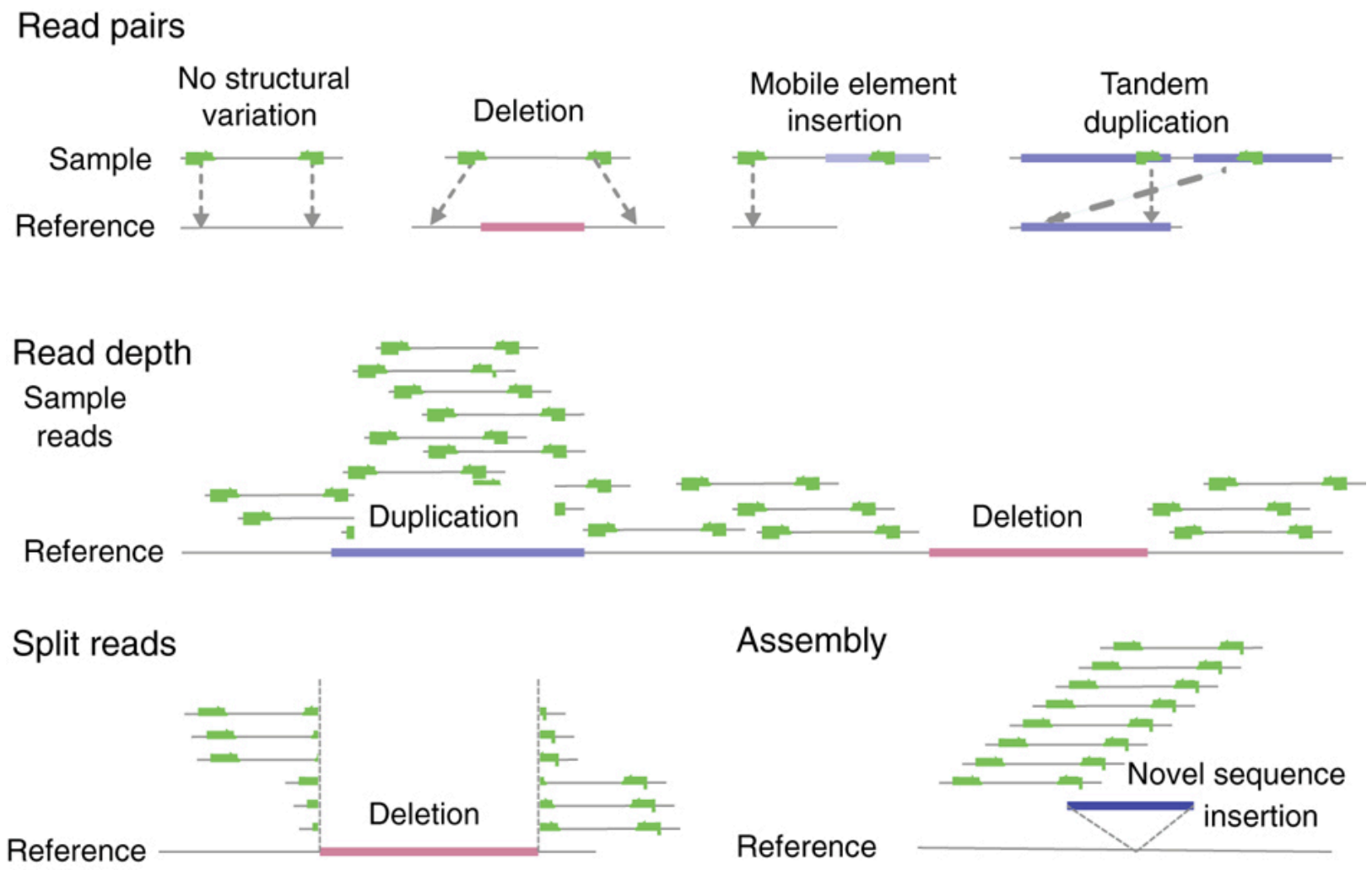
>2% divergence -> hash based approach

E.g. wild sample alignments ;

Watch out for latest advancement ; and don't stay at one for too long

Credit: Golden Helix inc

Detecting structural variations (ideally assembly is probably better)



Baker (2012)

What to do with repetitive (multi) reads?

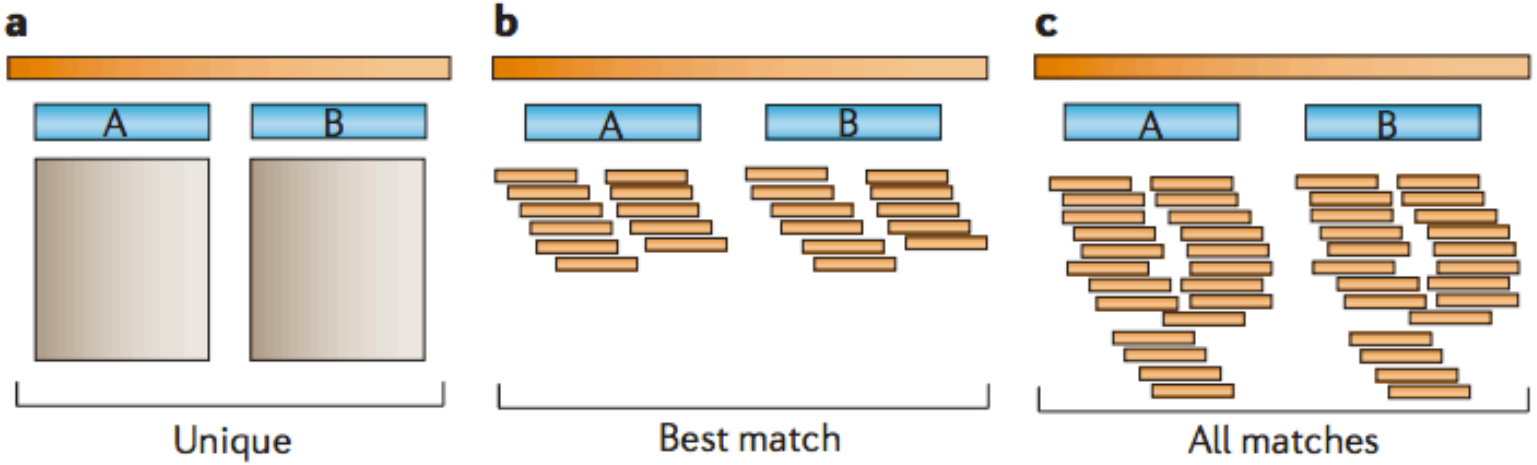
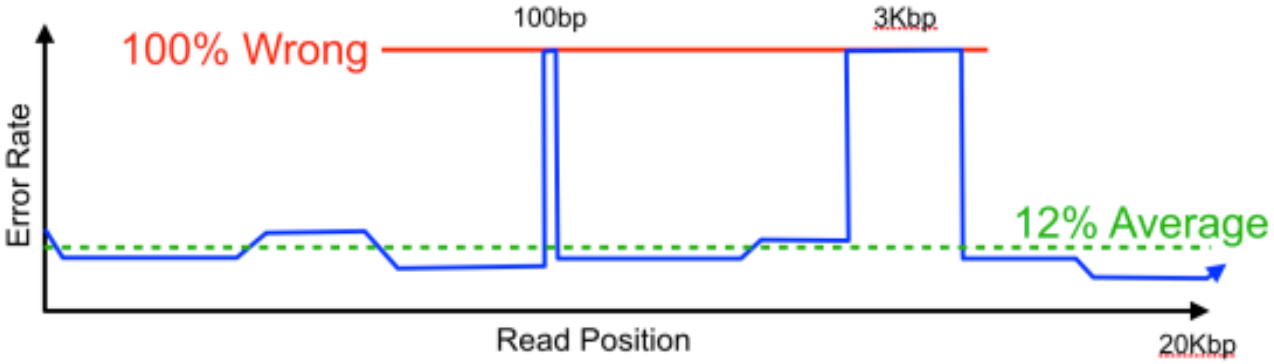
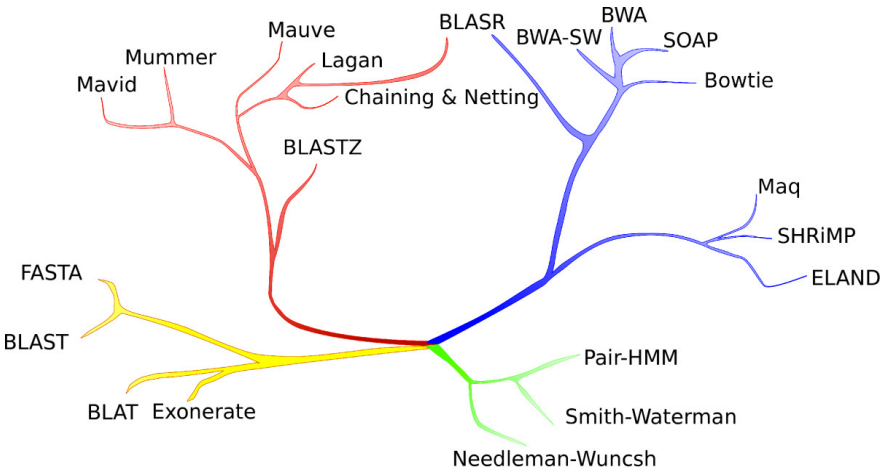


Figure 2 | **Three strategies for mapping multi-reads.** The shaded rectangles at the top represent intervals along a chromosome. The two blue rectangles below each region represent an identical two-copy repeat containing the paralogous genes A and B. The small orange bars represent reads aligned to specific positions. **a** | The 'unique' strategy reports only those reads that are uniquely mappable. Because A and B are identical, no alignments are reported. **b** | The 'best match' alignment strategy reports the best possible alignment for each read, which is determined by the scoring function of the alignment algorithm. In the case of ties, this strategy randomly distributes reads across equally good loci, as shown here. **c** | The 'all matches' strategy simply reports all alignments for each multi-read, including lower-scoring alignments.

Treangen *et al* (2012)

What about long read mapping?





- **BLASR** and **Daligner** designed for long error-prone (but random) reads (PacBio)
- Combines multiple methods
- Starts by finding short exact matches using suffix or B-W
- Next locally identifies a linear chain of shorter exact matches
- Performs banded Smith-Waterman constrained by the shorter exact matches

Chaisson et al (2012 ; https://dazzlerblog.wordpress.com/

Mapping algorithm – a summary

Feature	Hash table index tools	BWT tools
Speed	Slower	Faster
Memory	Higher	Lower
Sensitivity	Higher	Lower

Build an index of your reference

Align your reads to your index
- Choose an aligner!
  - Bowtie2, BWA-MEM
  - Isaac (much faster but less sensitivity ; good if you have huge data)
  - Blasr or Daligner (Pacbio)

Use the output to do subsequent analysis
- What's the output?
- How to use this output?

Back to the beginning: FASTQ

```
@HISEQ:409:HA7CJADXX:1:1101:1202:2113 1:N:0:GCNAAT
AAAAAAGTTTCCATACAATTACAAGCATCACACTGTGGGCATGCACTTGGGAAAGAA
+
==?DBD@<AA<ADAFHGGE<ECHHCG+:1::?D;G4::?BBGCFHI<BCCC;FCGC9
```

Read ID
Sequence

Quality score

Quality Score	Error Probability
Q40	0.0001 (1 in 10,000)
Q30	0.001 (1 in 1,000)
Q20	0.01 (1 in 100)
Q10	0.1 (1 in 10)

Q-Score Bins	Example of Empirically Mapped Q-Scores*
N (no call)	N (no call)
2–9	6
10–19	15
20–24	22
25–29	27
30–34	33
35–39	37
≥ 40	40

http://www.illumina.com/content/dam/illumina-marketing/documents/products/technotes/technote_understanding_quality_scores.pdf

Mapping output format: SAM/BAM

Spec defined by maq/bwa/samtools author Heng Li

SAM: text version

tab-delimited

Exome (GBs) ; Whole genome (TBs)

BAM: binary/compressed version

indexed so it's faster to look up using samtools

Exome (1-2GBs) ; Whole genome (GBs)

SAM file header

```
@HD     VN:1.4  SO:coordinate
@SQ     SN:PNOK.scaff0001.C     LN:7761079
@SQ     SN:PNOK.scaff0002.C     LN:4533150
@SQ     SN:PNOK.scaff0003.C     LN:3409659
@SQ     SN:PNOK.scaff0004.0     LN:3380754
@SQ     SN:PNOK.scaff0005.0     LN:2749859
@SQ     SN:PNOK.scaff0006.0     LN:2613677
@SQ     SN:PNOK.scaff0007.0     LN:1690816
@SQ     SN:PNOK.scaff0008       LN:1673160
@SQ     SN:PNOK.scaff0009.0     LN:1538597
@SQ     SN:PNOK.scaff0010       LN:1377172
@SQ     SN:PNOK.scaff0011       LN:633856
@SQ     SN:PNOK.scaff0012       LN:52253
@SQ     SN:PNOK.mito    LN:163443
@PG     ID:smalt        VN:0.7.4        CL:/h
```

Always start with @

Contains “background” information

@HD = Header

@SQ = Sequence dictionary

SAM file header

Very detailed in how one should specify the headers

Subsequent programs (like variant calling) will use these info

http://samtools.github.io/hts-specs/SAMv1.pdf

Tag	Description
@HD	The header line. The first line if present.
VN*	Format version. *Accepted format*: /^[0-9]+\.[0-9]+$/.
SO	Sorting order of alignments. *Valid values*: **unknown** (default), **unsorted**, **queryname** and **coordinate**. For coordinate sort, the major sort key is the RNAME field, with order defined by the order of @SQ lines in the header. The minor sort key is the POS field. For alignments with equal RNAME and POS, order is arbitrary. All alignments with '*' in RNAME field follow alignments with some other value but otherwise are in arbitrary order.
GO	Grouping of alignments, indicating that similar alignment records are grouped together but the file is not necessarily sorted overall. *Valid values*: **none** (default), **query** (alignments are grouped by QNAME), and **reference** (alignments are grouped by RNAME/POS).
@SQ	Reference sequence dictionary. The order of @SQ lines defines the alignment sorting order.
SN*	Reference sequence name. Each @SQ line must have a unique SN tag. The value of this field is used in the alignment records in RNAME and RNEXT fields. Regular expression: [!-)+-<>-~][!-~]*
LN*	Reference sequence length. *Range*: [1,2^{31}-1]
AS	Genome assembly identifier.
M5	MD5 checksum of the sequence in the uppercase, excluding spaces but including pads (as '*'s).
SP	Species.
UR	URI of the sequence. This value may start with one of the standard protocols, e.g http: or ftp:. If it does not start with one of these protocols, it is assumed to be a file-system path.
@RG	Read group. Unordered multiple @RG lines are allowed.
ID*	Read group identifier. Each @RG line must have a unique ID. The value of ID is used in the RG tags of alignment records. Must be unique among all read groups in header section. Read group IDs may be modified when merging SAM files in order to handle collisions.
CN	Name of sequencing center producing the read.
DS	Description.
DT	Date the run was produced (ISO8601 date or date/time).
FO	Flow order. The array of nucleotide bases that correspond to the nucleotides used for each flow of each read. Multi-base flows are encoded in IUPAC format, and non-nucleotide flows by various other characters. *Format*: /*|[ACMGRSVTWYHKDBN]+/
KS	The array of nucleotide bases that correspond to the key sequence of each read.
LB	Library.
PG	Programs used for processing the read group.
PI	Predicted median insert size.
PL	Platform/technology used to produce the reads. *Valid values*: CAPILLARY, LS454, ILLUMINA, SOLID, HELICOS, IONTORRENT, ONT, and PACBIO.
PM	Platform model. Free-form text providing further details of the platform/technology used.
PU	Platform unit (e.g. flowcell-barcode.lane for Illumina or slide for SOLiD). Unique identifier.
SM	Sample. Use pool name where a pool is being sequenced.
@PG	Program.
ID*	Program record identifier. Each @PG line must have a unique ID. The value of ID is used in the alignment PG tag and PP tags of other @PG lines. PG IDs may be modified when merging SAM files in order to handle collisions.
PN	Program name
CL	Command line

SAM file mapping

Read 1

Read 2

Sorted by chromosome position

SAM file spec

Col	Field	Type	Regexp/Range	Brief description
1	QNAME	String	`[!-?A-~]{1,254}`	Query template NAME
2	FLAG	Int	$[0,2^{16}-1]$	bitwise FLAG
3	RNAME	String	`\*|[!-()+-<>-~][!-~]*`	Reference sequence NAME
4	POS	Int	$[0,2^{31}-1]$	1-based leftmost mapping POSition
5	MAPQ	Int	$[0,2^{8}-1]$	MAPping Quality
6	CIGAR	String	`\*|([0-9]+[MIDNSHPX=])+`	CIGAR string
7	RNEXT	String	`\*|=|[!-()+-<>-~][!-~]*`	Ref. name of the mate/next read
8	PNEXT	Int	$[0,2^{31}-1]$	Position of the mate/next read
9	TLEN	Int	$[-2^{31}+1,2^{31}-1]$	observed Template LENgth
10	SEQ	String	`\*|[A-Za-z=.]+`	segment SEQuence
11	QUAL	String	`[!-~]+`	ASCII of Phred-scaled base QUALity+33

http://samtools.github.io/hts-specs/SAMv1.pdf

SAM file mapping

Col	Field	
1	QNAME	HWI-M01162:89:000000000-AC0DK:1:2103:24027:15590
2	FLAG	163
3	RNAME	PNOK.scaff0001.C
4	POS	539
5	MAPQ	60
6	CIGAR	300M
7	RNEXT	=
8	PNEXT	680
9	TLEN	441
10	SEQ	
11	QUAL	

Storing everything means fastq can be deleted (if you are not going to try different aligners and have same pipeline)

CGATAACAAATAATCACATGGTATGTTTCTATTGTCACTACAAGATTACTGATCATTCCATTTGCAACGATGG
CGTGATCCTCGCATCATACAATTACTAACTAAGGAGACCTGACGAGTATTTATACCAAAGAGGTCTACAGGA
GAGGGGAGTCAAATCCCCACCTCTCGTCTCTTTAGATCCTCTCATTACCTCGCTTCGCTGCGCTCAGCTTC
GAACCTAATAGTTAAGTGTCACGTGATTAGGATTAGTAAGCAAATTACTTAATCATATGGTCACTAATATGCT
TTGTCATAGAT

B>>FFFFFFFFFFFFEFFFFFFFFFFFFFFFFFGC4FGFFFGGFFCGGGGGGGGGFGGGGGGGGGGGGGGGGGGG
GGDDGGGGGGGGGF9GGGFGGGGGFFGGGGGGGGGGGGGGGGGGGGGFGGGGGGGGGGGGGGGGGG
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGEGGGGGGGGGGGGGG
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
GGGGGGGGGGGGGGGGGGGGGCCCCC

SAM flags

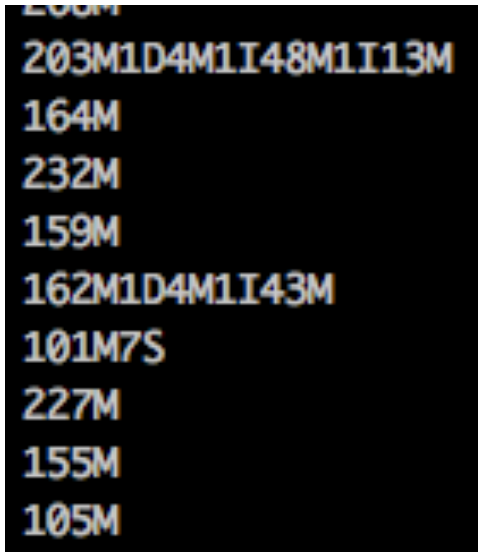
hexadecimal	decimal	binary bit; 0=no, 1=yes	position of bit	description
0x1	1	"0000 0000 0001"	1	paired-end (or multiple-segment) sequencing technology
0x2	2	"0000 0000 0010"	2	each segment properly aligned according to the aligner
0x4	4	"0000 0000 0100"	3	segment unmapped
0x8	8	"0000 0000 1000"	4	next segment in the template unmapped
0x10	16	"0000 0001 0000"	5	SEQ is reverse complemented
0x20	32	"0000 0010 0000"	6	SEQ of the next segment in the template is reverse complemented
0x40	64	"0000 0100 0000"	7	the first segment in the template
0x80	128	"0000 1000 0000"	8	the last segment in the template
0x100	256	"0001 0000 0000"	9	secondary alignment
0x200	512	"0010 0000 0000"	10	not passing quality controls
0x400	1024	"0100 0000 0000"	11	PCR or optical duplicate
0x800	2048	"1000 0000 0000"	12	supplementary alignment

http://gatkforums.broadinstitute.org/gatk/discussion/7019/sam-flags-down-a-boat

CIGAR String a few examples

ATCGATCGATCGATCG Reference

ATGGACGATTCGTGAA Read mapping = 5M1D3M1I3M4S



Soft clip usually the result of lower mapping quality

op	Description
M	Alignment match (can be a sequence match or mismatch
I	Insertion to the reference
D	Deletion from the reference
N	Skipped region from the reference
S	Soft clip on the read (clipped sequence present in <seq>)
H	Hard clip on the read (clipped sequence NOT present in <seq>)
P	Padding (silent deletion from the padded reference sequence)

Mapping quality

Probability that a read is mapped incorrectly

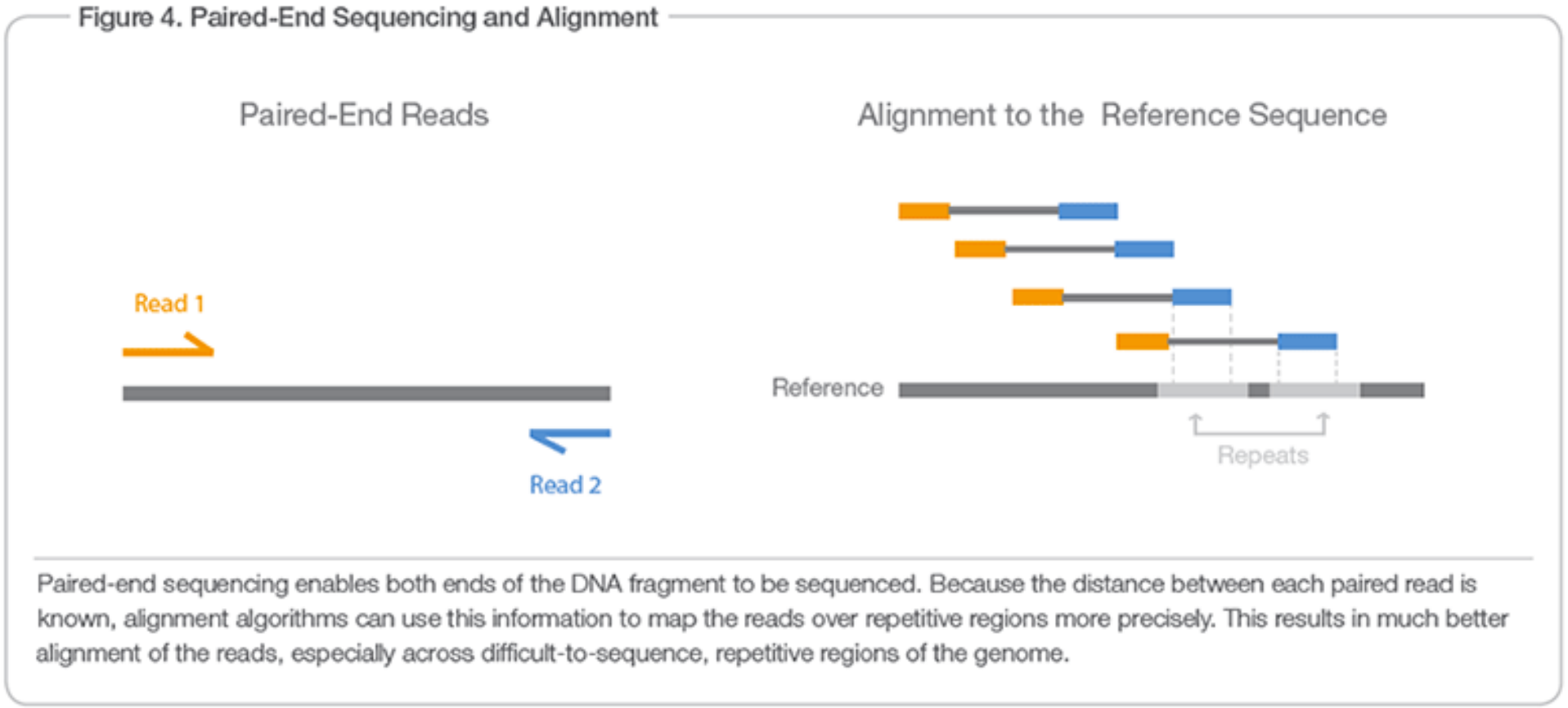
Useful for calling SNP later on

Function of

- Uniqueness
- Number of mismatches
- Number of indels
- Quality of bases in read

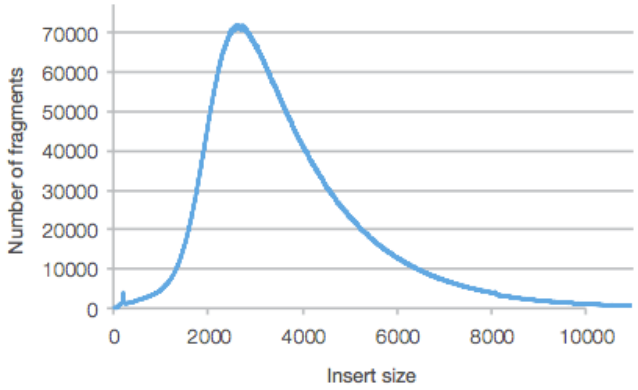
MQ30 = 1 in 1000 alignment is wrong
MQ40 = 1 in 10000 alignment is wrong

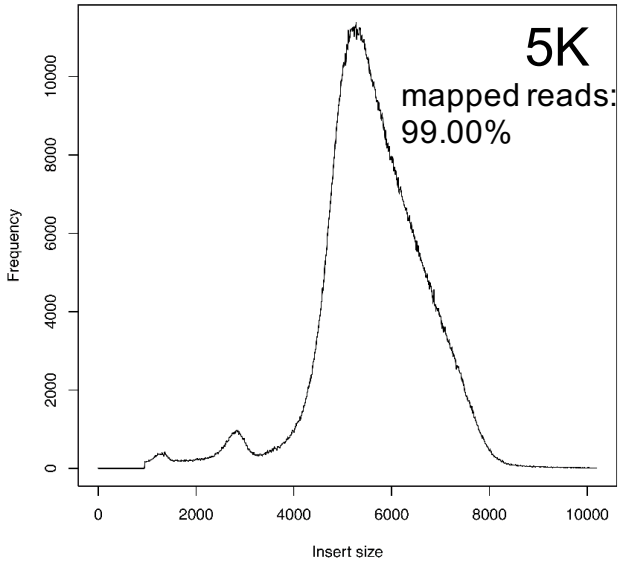
Post mapping QC: insert size in PE mapping?

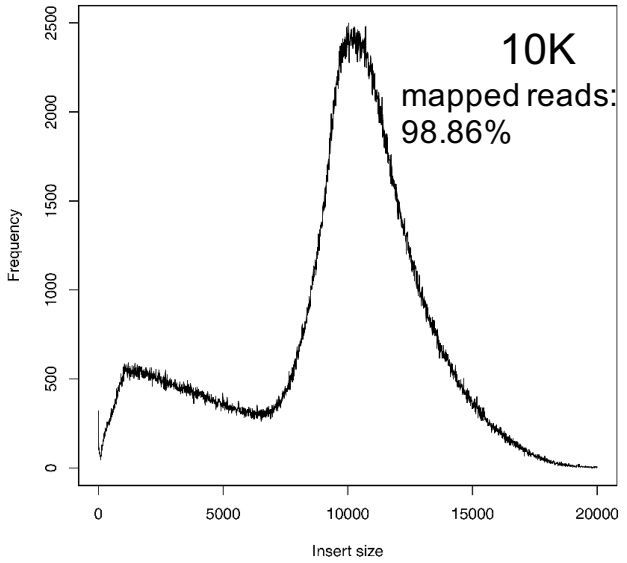


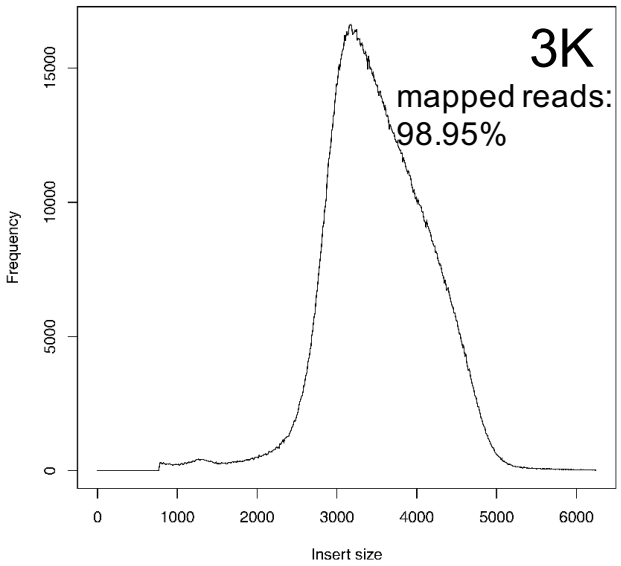
DNA fragment length should be longer than most repeat size in your genome
No point to boost up coverage if your fragment len < repeat lengtg

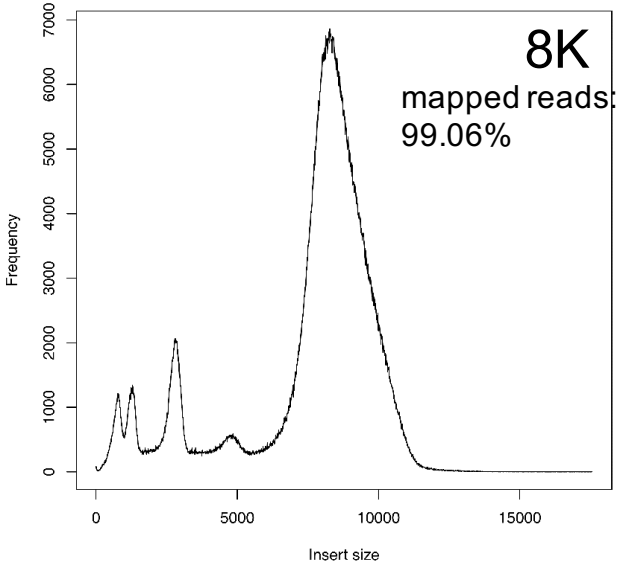
Insert size

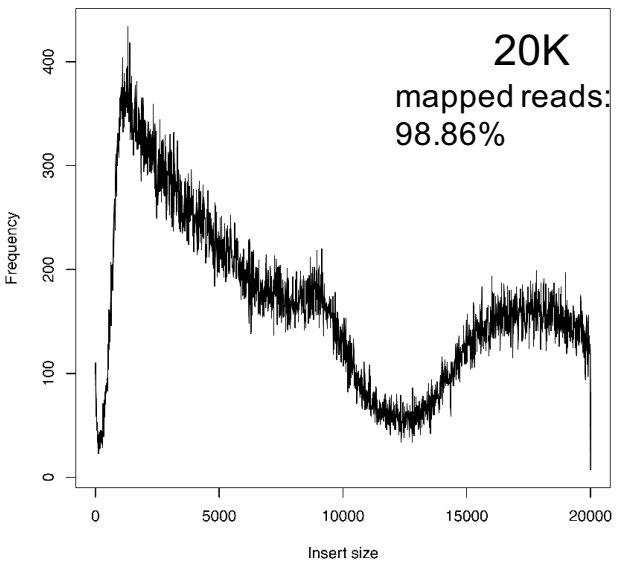

Number of fragments
70000
60000
50000
40000
30000
20000
10000
0
0
2000
4000
6000
8000
10000
Insert size


5K
mapped reads:
99.00%
Frequency
0
2000
4000
6000
8000
10000
0
2000
4000
6000
8000
10000
Insert size


10K
mapped reads:
98.86%
Frequency
0
500
1000
1500
2000
2500
0
5000
10000
15000
20000
Insert size


3K
mapped reads:
98.95%
Frequency
0
5000
10000
15000
0
1000
2000
3000
4000
5000
6000
Insert size


8K
mapped reads:
99.06%
Frequency
0
1000
2000
3000
4000
5000
6000
7000
0
5000
10000
15000
Insert size


20K
mapped reads:
98.86%
Frequency
0
100
200
300
400
0
5000
10000
15000
20000
Insert size

Post mapping QC – how much coverage?

```
**********************************************
Stats for BAM file(s):
**********************************************

Total reads:        2963812
Mapped reads:       2926492       (98.7408%)
Forward strand:     1463708       (49.386%)
Reverse strand:     1462784       (49.3548%)
Failed QC:          0     (0%)
Duplicates:         8469 (0.285747%)
Paired-end reads:   2963812       (100%)
'Proper-pairs':     2808018       (94.7435%)
Both pairs mapped:  2901342       (97.8922%)
Read 1:             1481906
Read 2:             1481906
Singletons:         25150         (0.848569%)
Average insert size (absolute value): 808.327
Median insert size (absolute value): 466
```

2963812 reads
x 300 bp per read
/ 32000000bp genome
= **27.8X**

This number is overestimated because
1. ~1.3% not mapped
2. Trimmed reads (not all reads have now 300bp)

Barnett *et al.*, (2011) **bamtools**

1 million dollar question: how much coverage is better

In mapping:

- ~15X for SNP calling in bacteria
- ~30X for SNP calling in diploid (to delineate heterozygous bases)
- >50X for exome (because you need to be sure)
- No point with >100X in the Illumina world

PCR duplicates

PCR duplicates during sample prep
= the same fragment is sequenced again and again and again

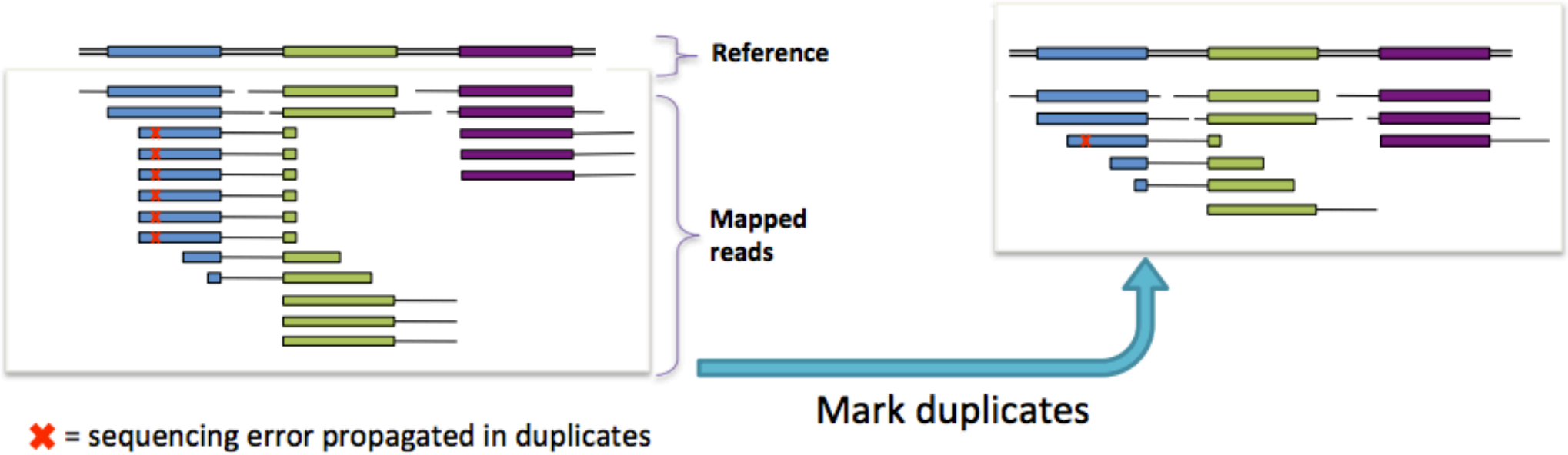
Some worse than others (because starting material is not good)
< 5% is good

High duplication rate will lead to problems in downstream analysis
Example: 30X ; 1 out of ~30 fragment get duplicated 15 times
= skew allele frequency
= false SNP discovery

Can be detected (and removed) by read pairs map at the complete position. **We usually keep one copy only**

PCR duplicates

Can be detected (and removed) by read pairs map at the complete position.
We usually keep one copy only

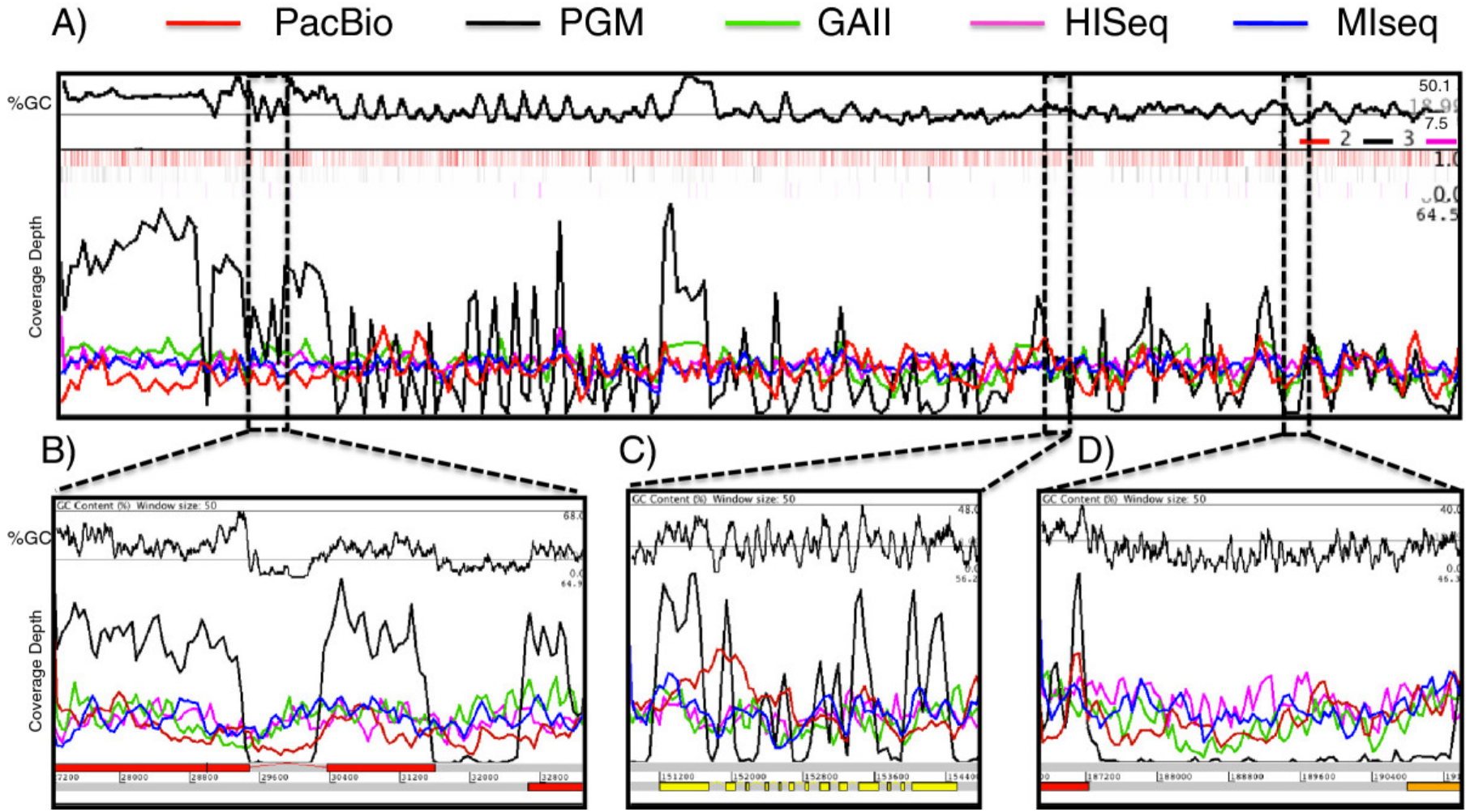


Broad - GATK

De- duplication

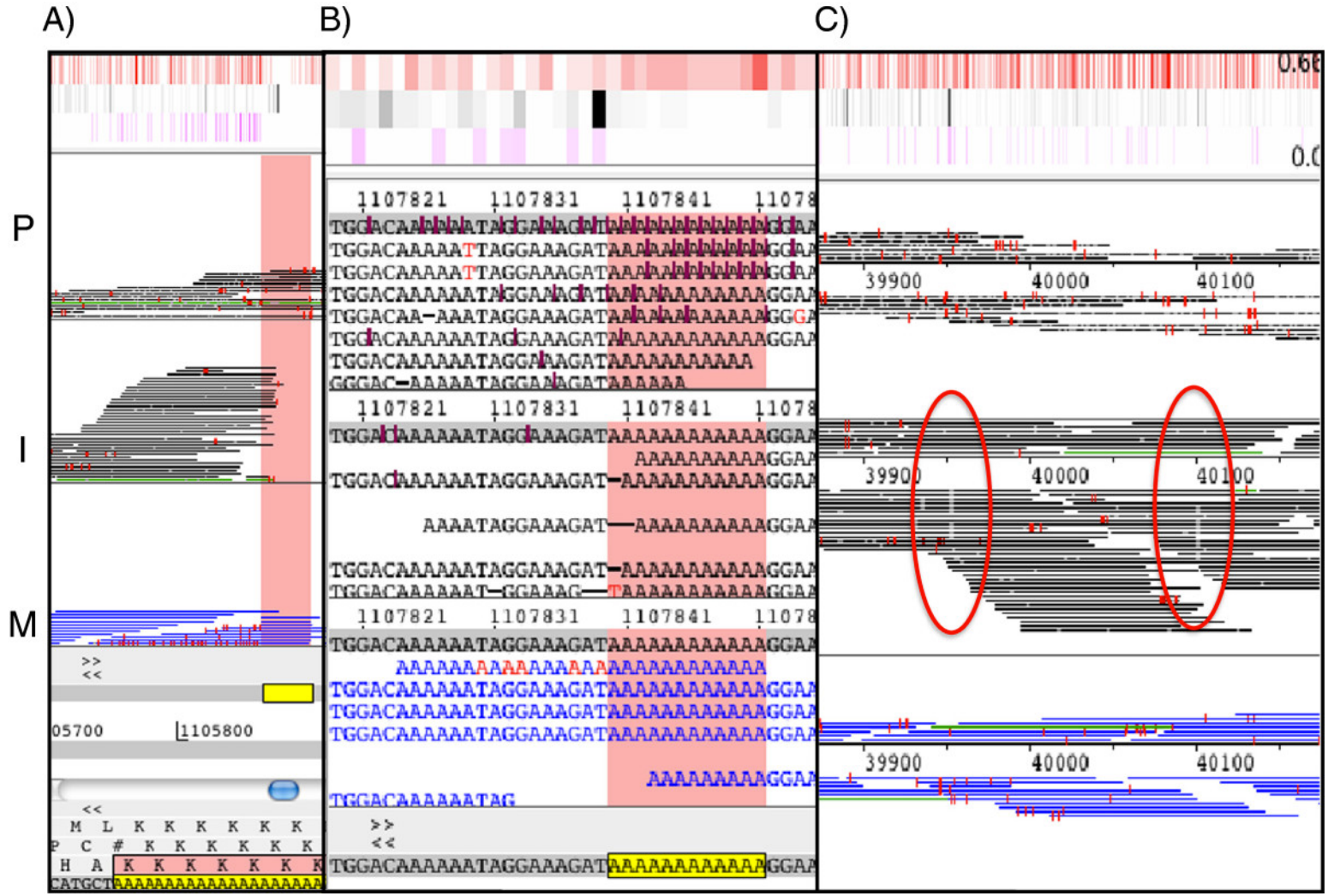
Broad - GATK

Sequencing biases



Quail *et al* 2012

Platform specific biases



Quail *et al* 2012

Experiment biases

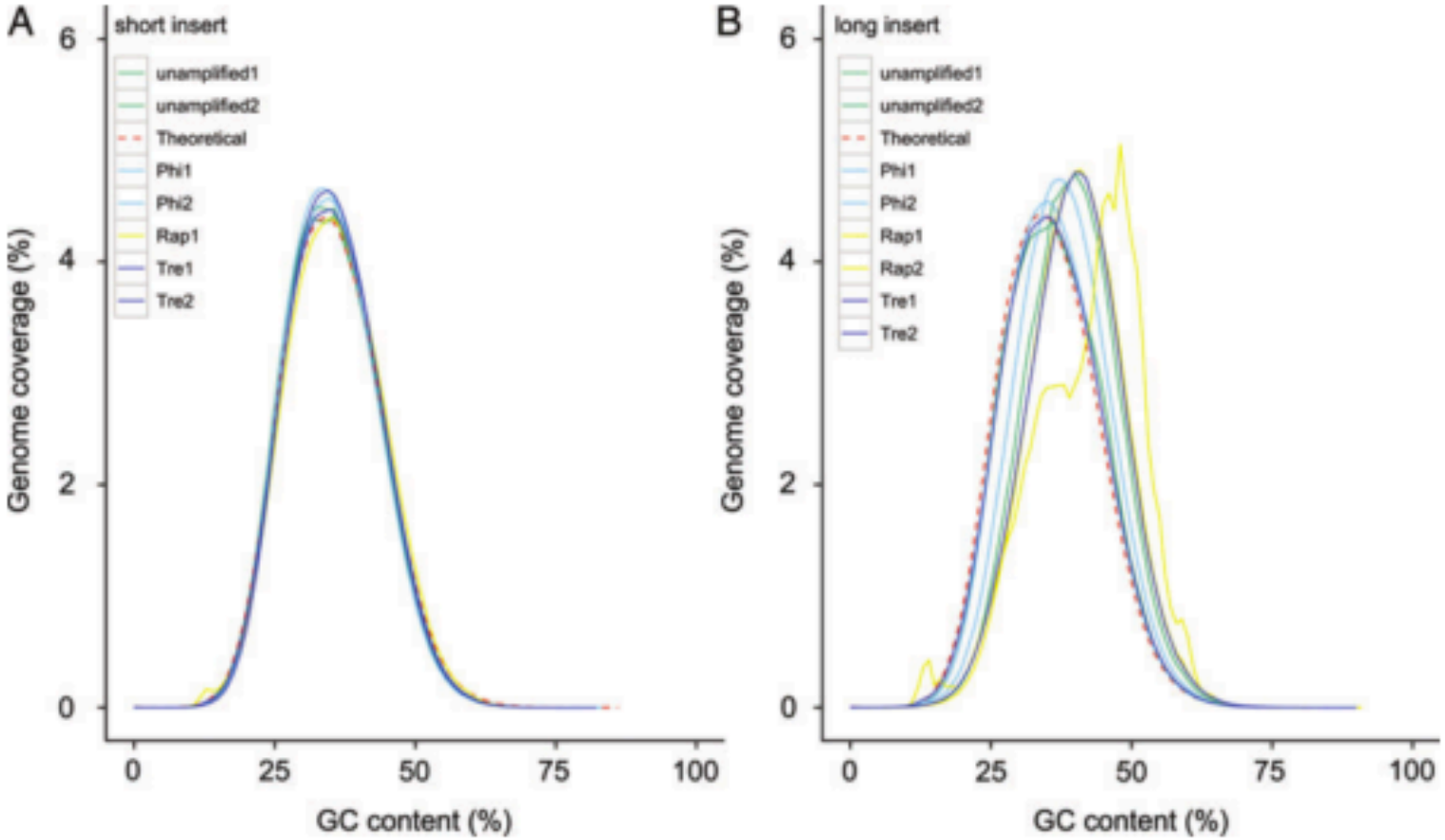


Figure 5. Distribution of GC content in sequenced reads of (A) short- and (B) long-insert libraries.

Tsai *et al* 2014

Mapping output: A summary

There is a lot you can do from the initial mapping output

- Post mapping QC
- Assembly QC

At this point you should decide whether

- it's a good run and you can go ahead to the next stage
- you need additional run
- you need to abandon the whole run

Variant calling

Variant calling

You have just:

- Mapped the reads to where they belong
- Provided accurate mapping quality scores

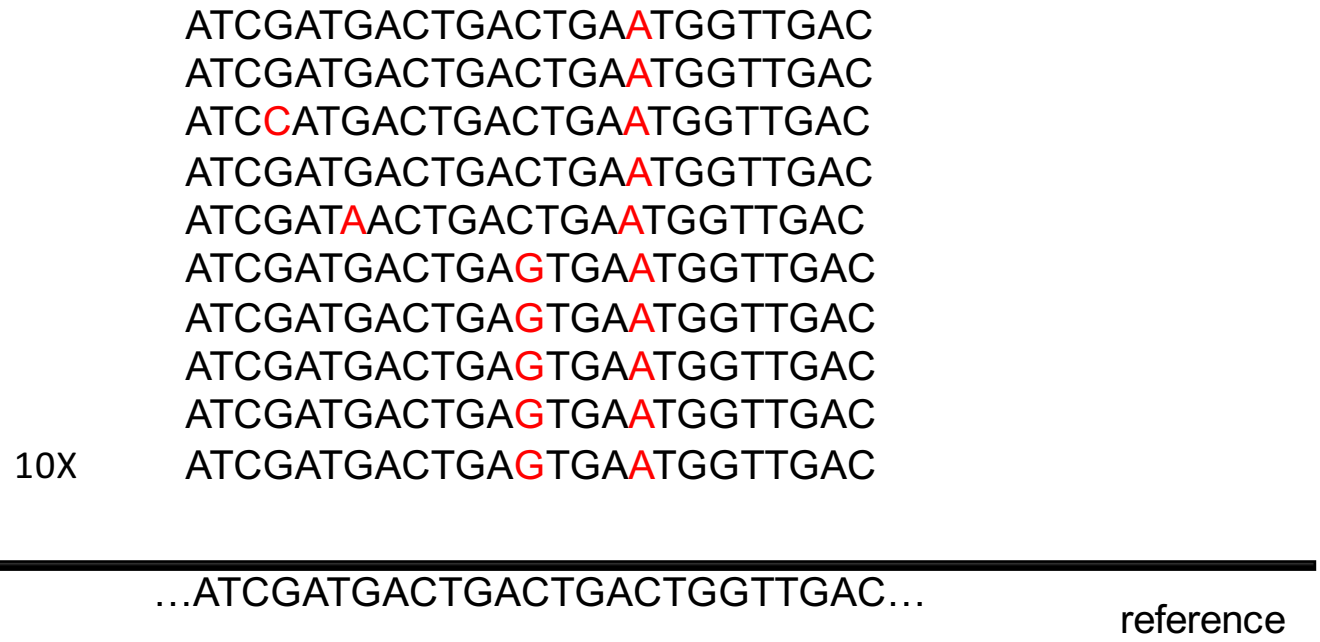
Next:

- Give the correct data (**BAM**) to variant callers

How to determine the above are correct?

SNP discovery

Heterozygous and homozygous SNP



INDELS (insertion deletions) and Structural variations

Indel examples

wild-type sequence

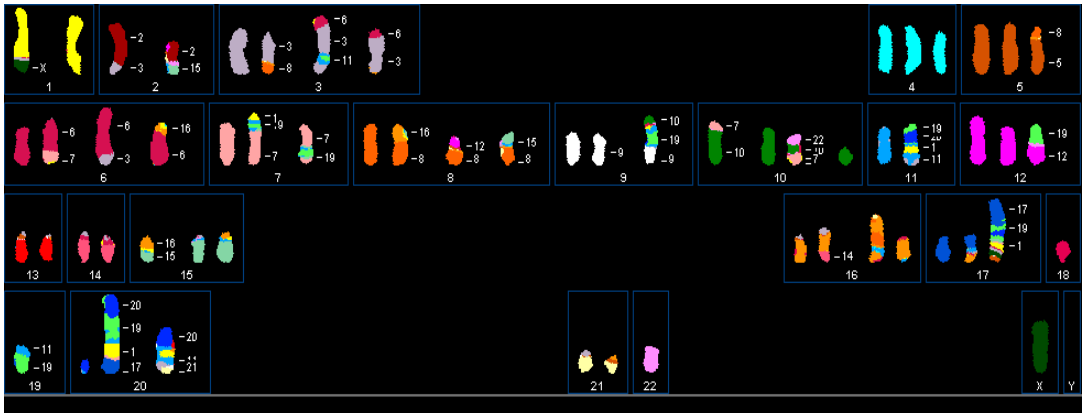
ATCTTCAGCCATAAAAGATGAAGTT

3 bp deletion

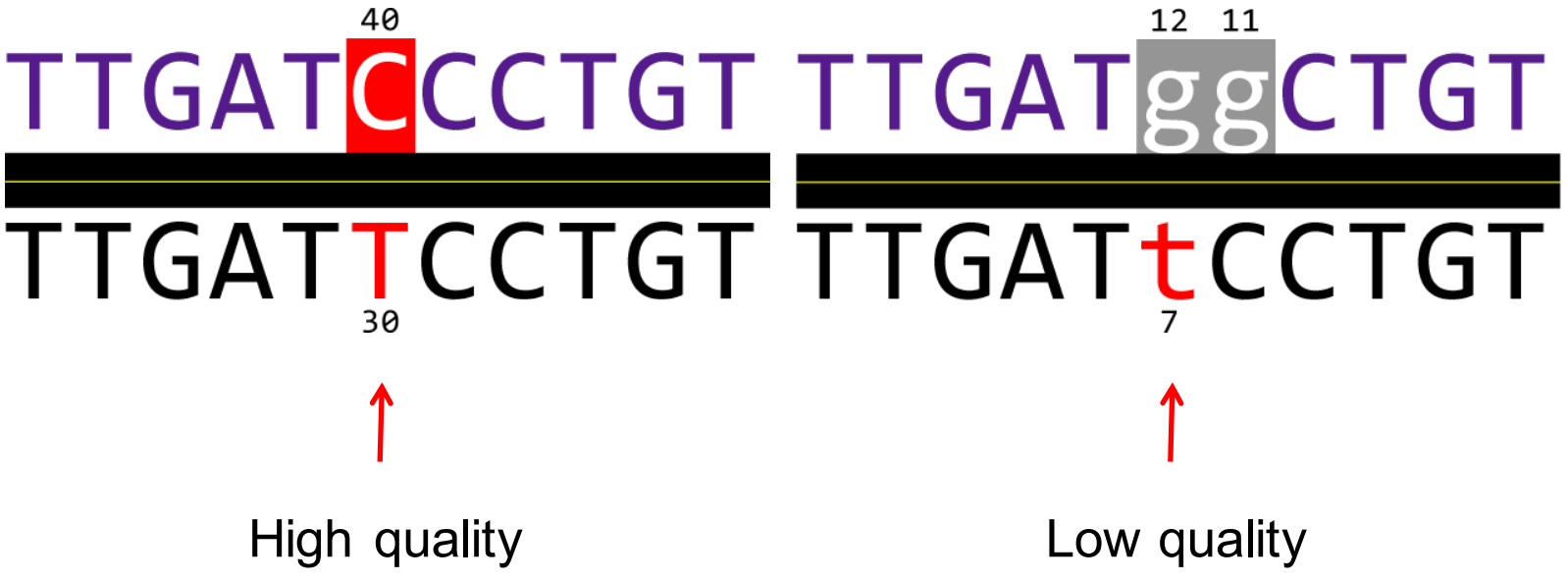
ATCTTCAGCCAAAGATGAAGTT

4 bp insertion (orange)

ATCTTCAGCCATATGTGAAAGATGAAGTT

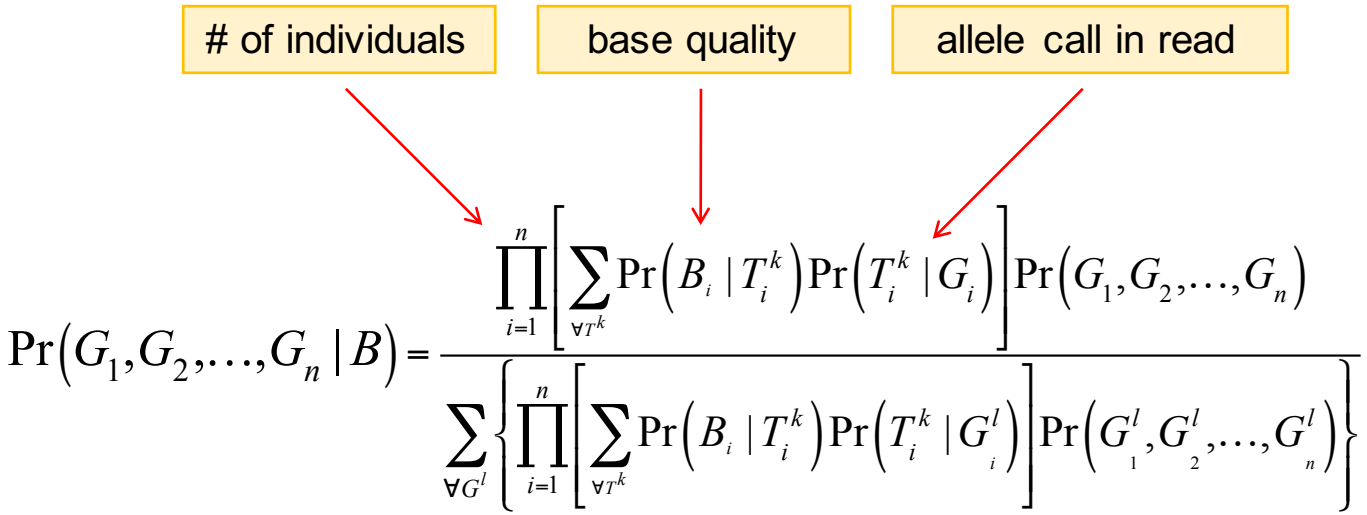


SNP Discovery: Base Qualities



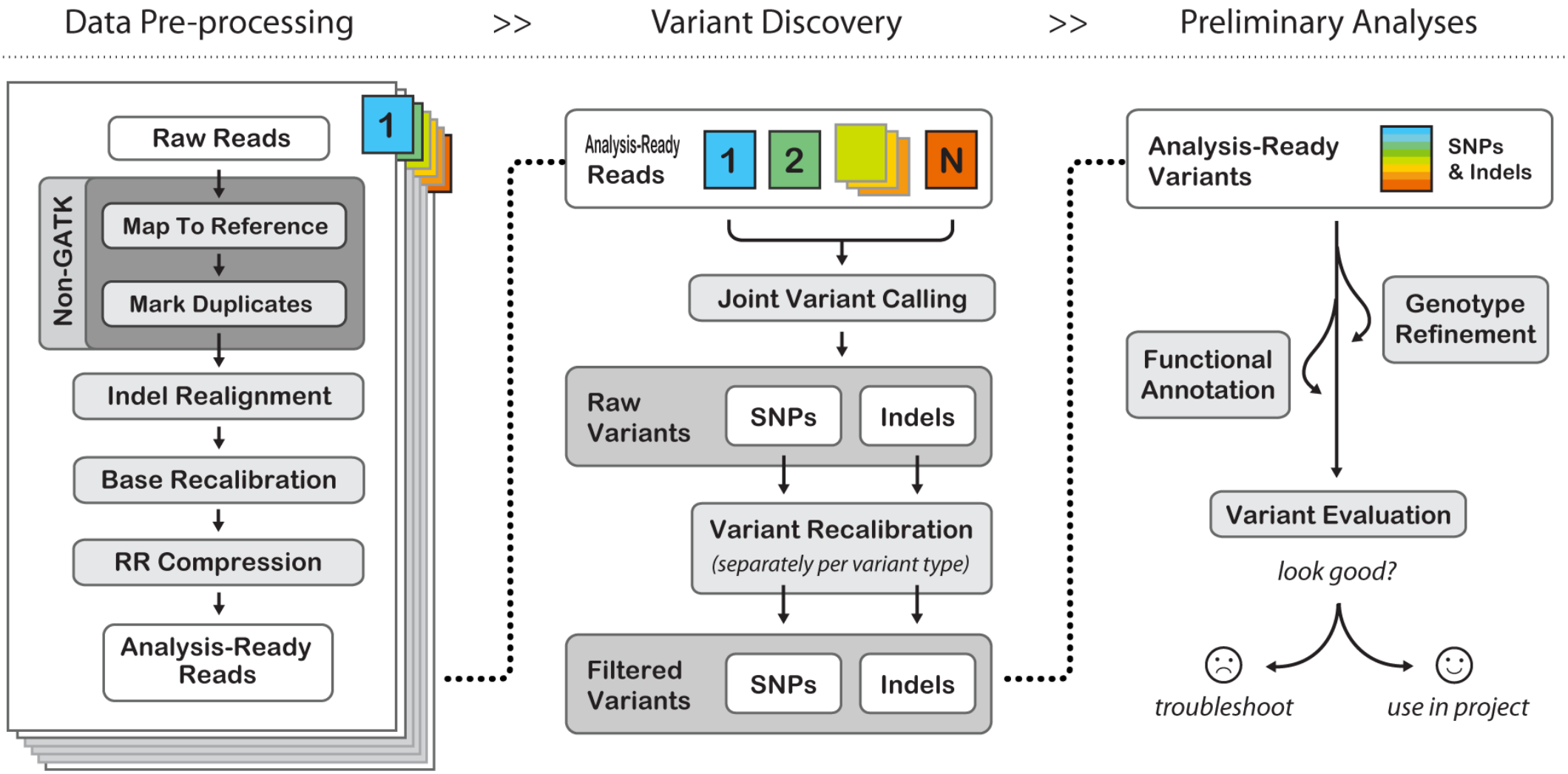
Michael Strömberg

SNPs & Bayesian Statistics

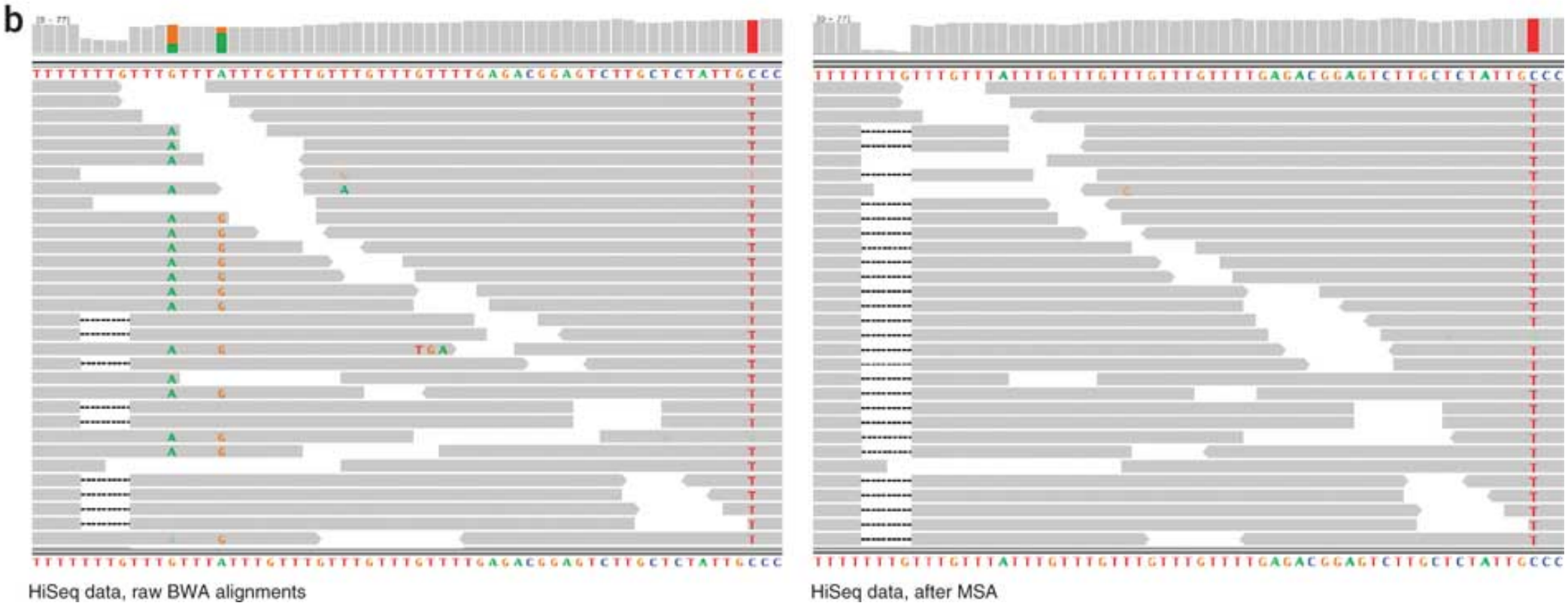


Michael Strömberg

Strategies that improve variant calling



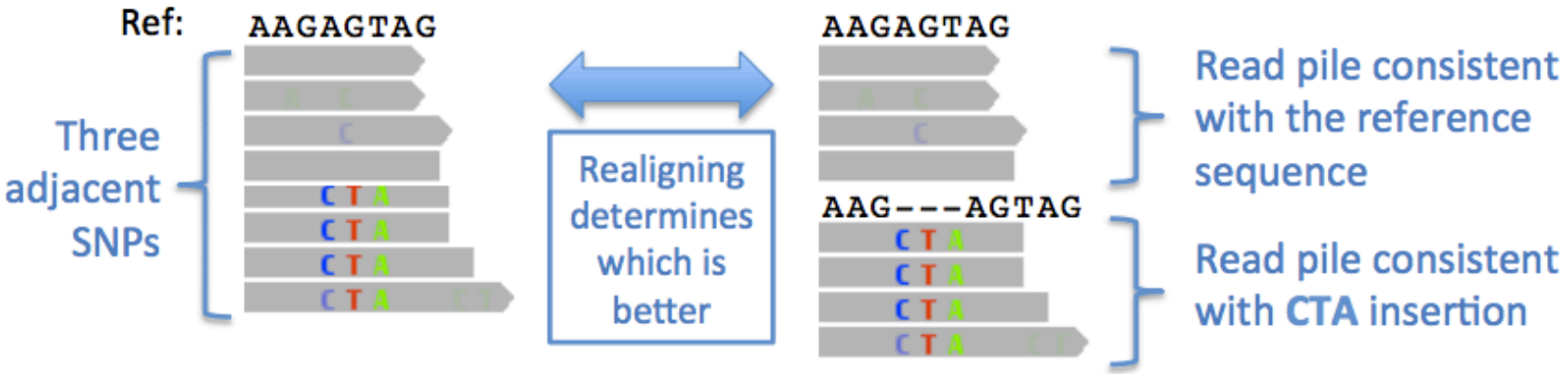
Local realignement



DePristo et al. *Nat Genet* 2011

Local realignement - principle

1. Find the best alternate consensus sequence that, together with the reference, best fits the reads in a pile (maximum of 1 indel)



2. The score for an alternate consensus is the total sum of the quality scores of mismatching bases

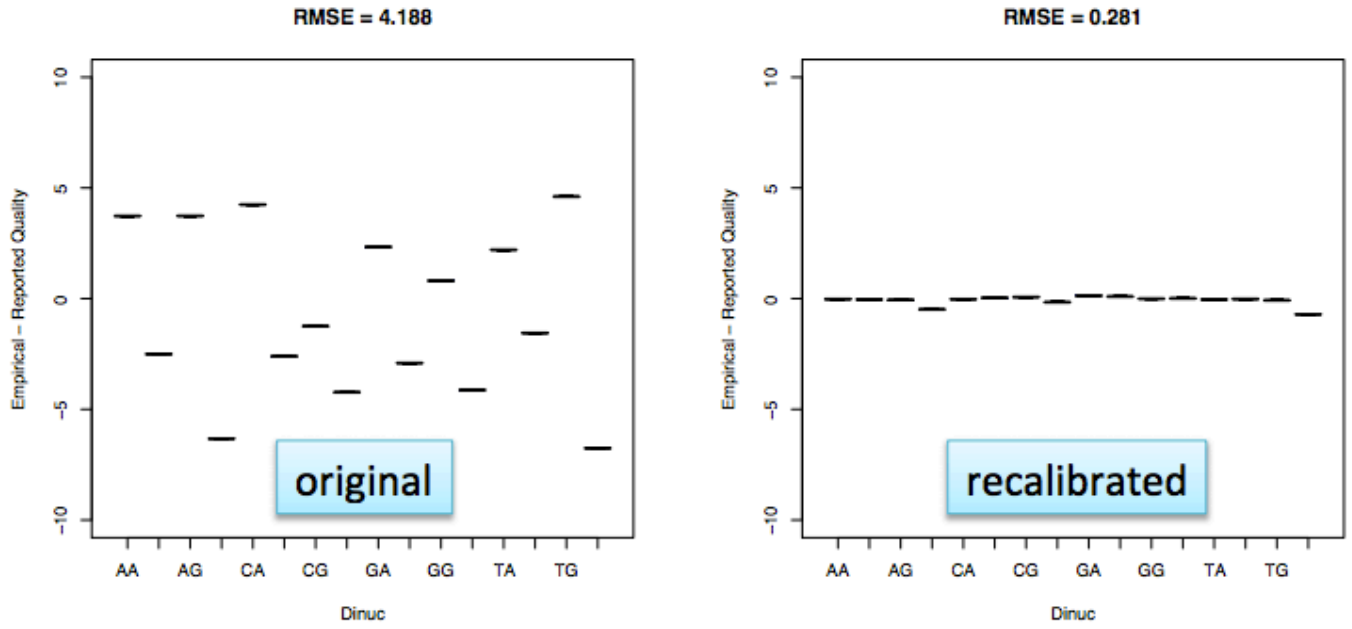
3. If the score of the best alternate consensus is sufficiently better than the original alignments (using a LOD score), then we accept the proposed realignment of the reads

GATK, Broad

Base quality recalibration

- Quality scores are critical for all downstream analysis
- Systematic biases are a major contributor to bad calls

Example of bias: qualities reported depending on nucleotide context

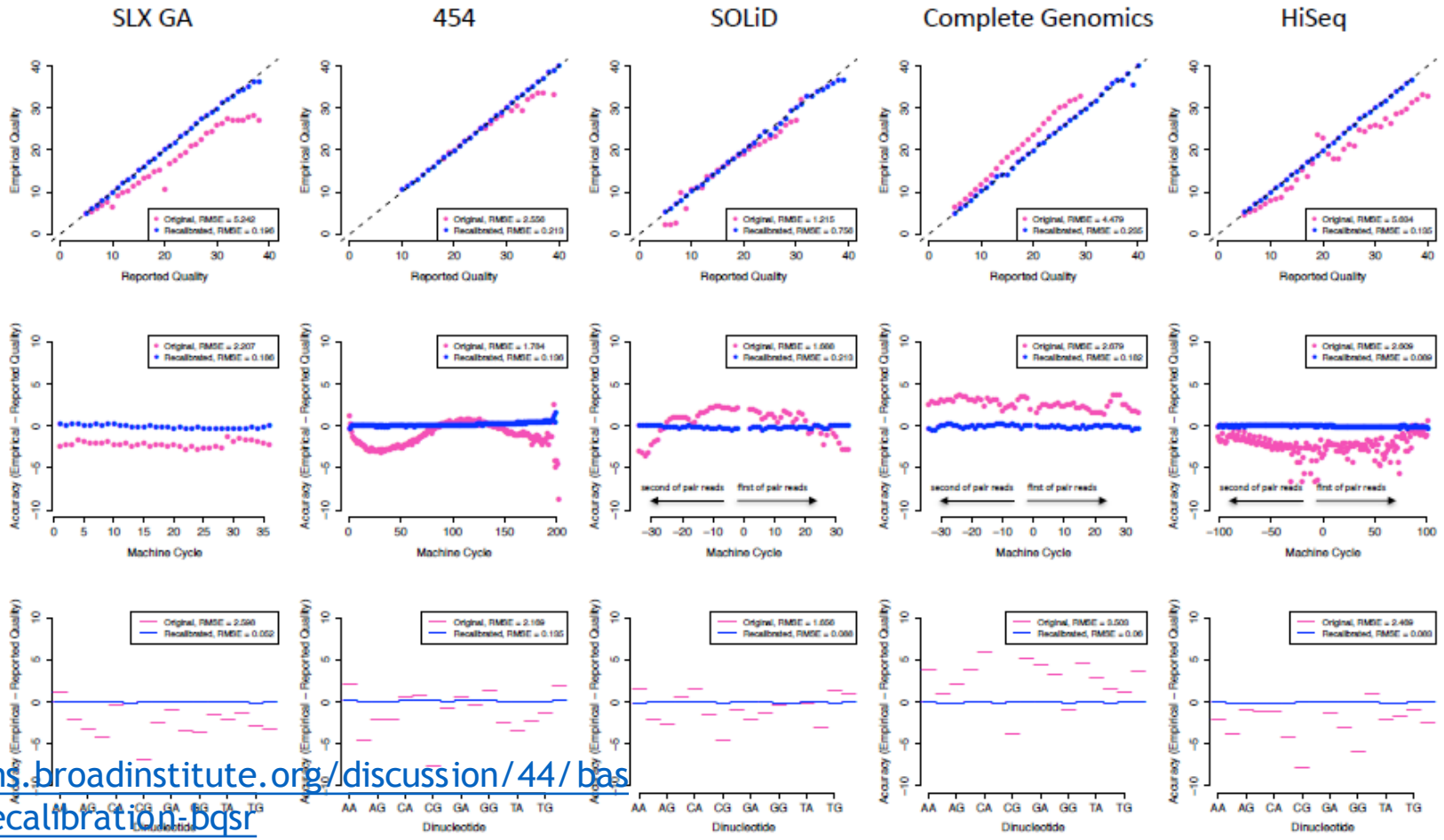


BQSR method identifies bias and applies correction

GATK, Broad

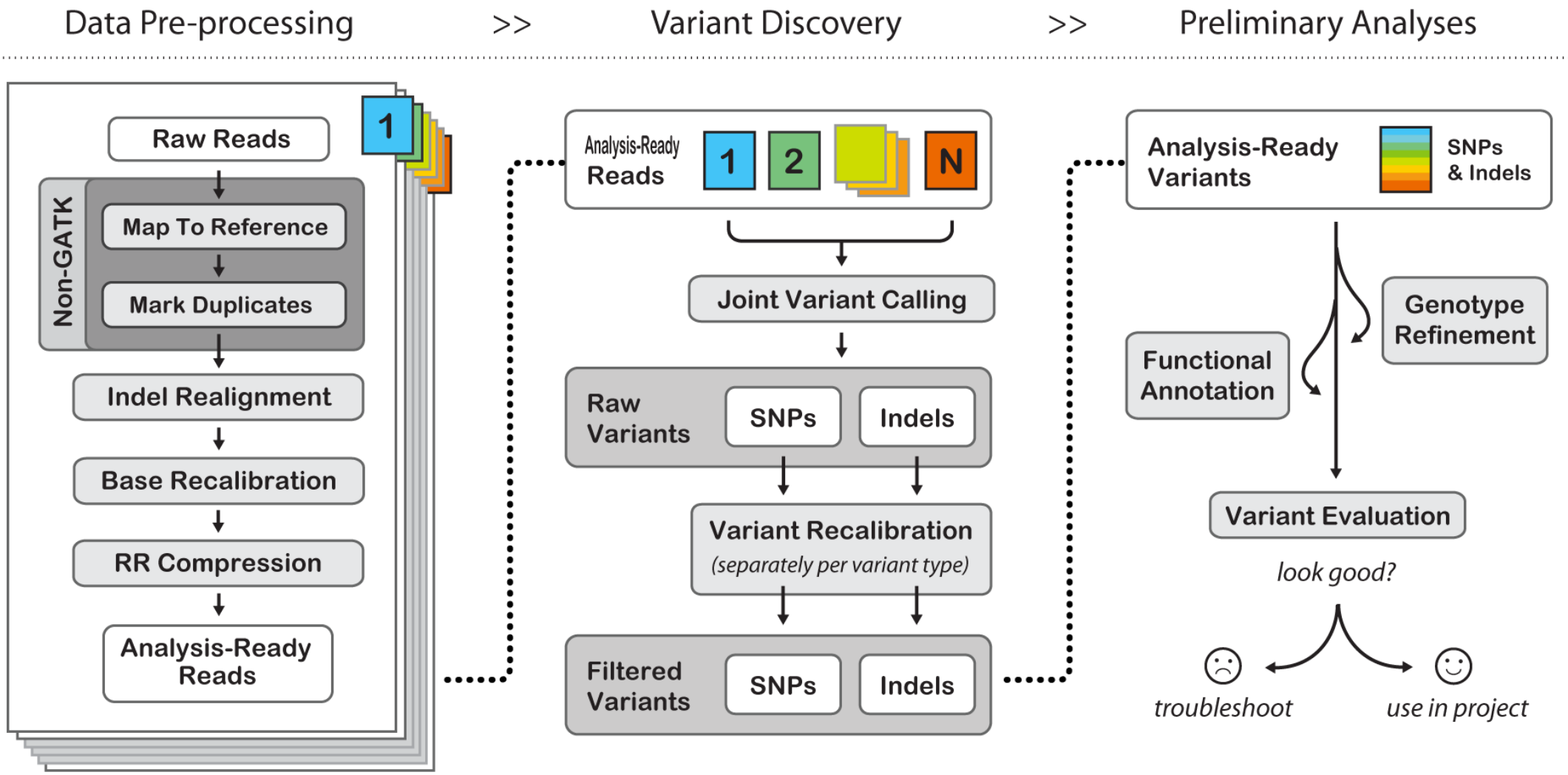
Highlighted as one of the major methodological advances of the 1000 Genomes Pilot Project!

Base Quality Score Recalibration provides a calibrated error model from which to make mutation calls



http://gatkforums.broadinstitute.org/discussion/44/base-quality-score-recalibration-bqsr

Improve beyond analysis-ready reads



Using haplotypes for base calling

- Suppose that only 2 haplotypes have been observed in a population:

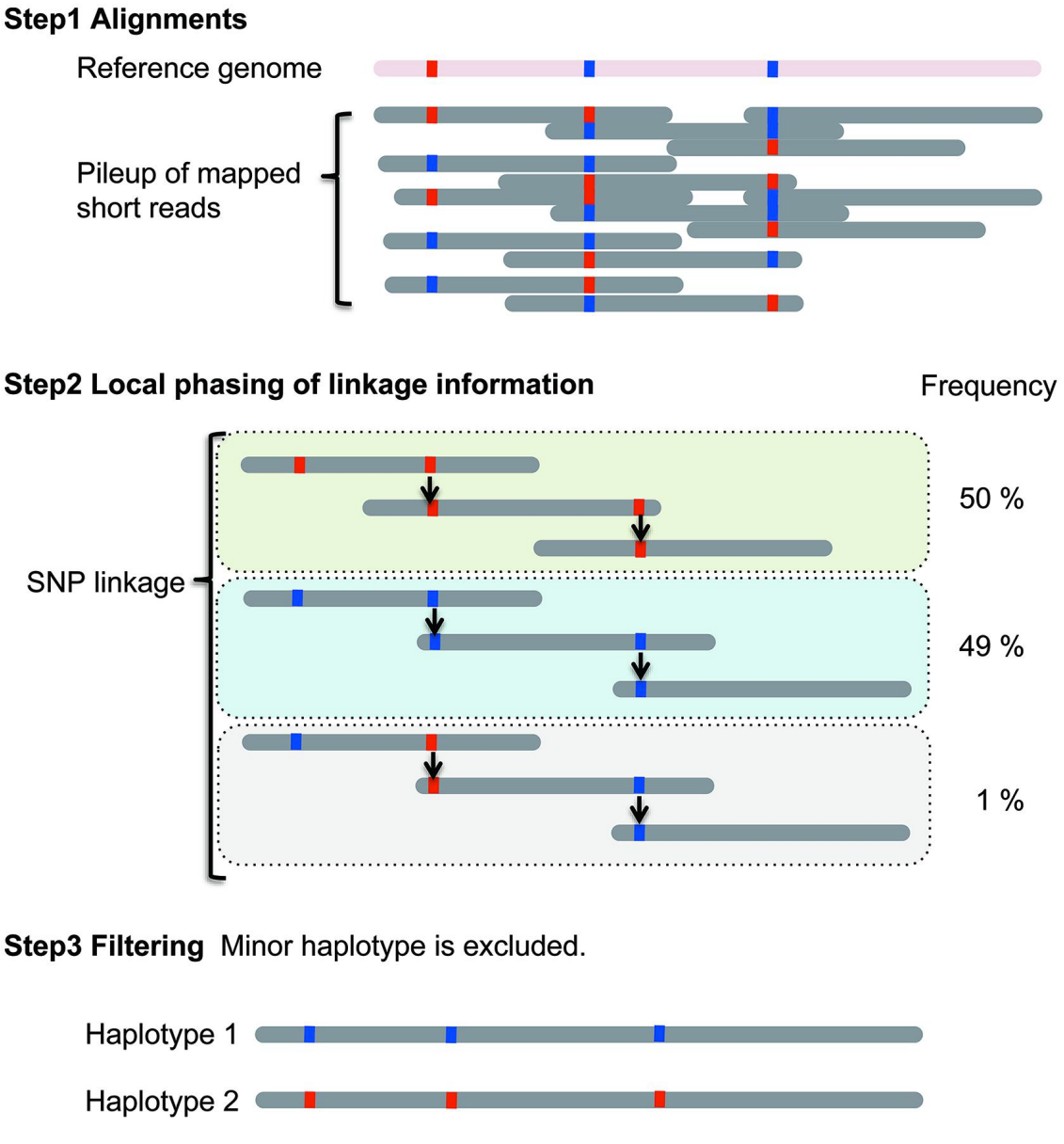
```
Chr1: ..........A....T.......G..........
Chr1: ..........C....G.......A..........
```

- And that you observe the following reads:

```
......A....N.......G..
   ..A....N.......G.....
  ...A....N.......G...
```

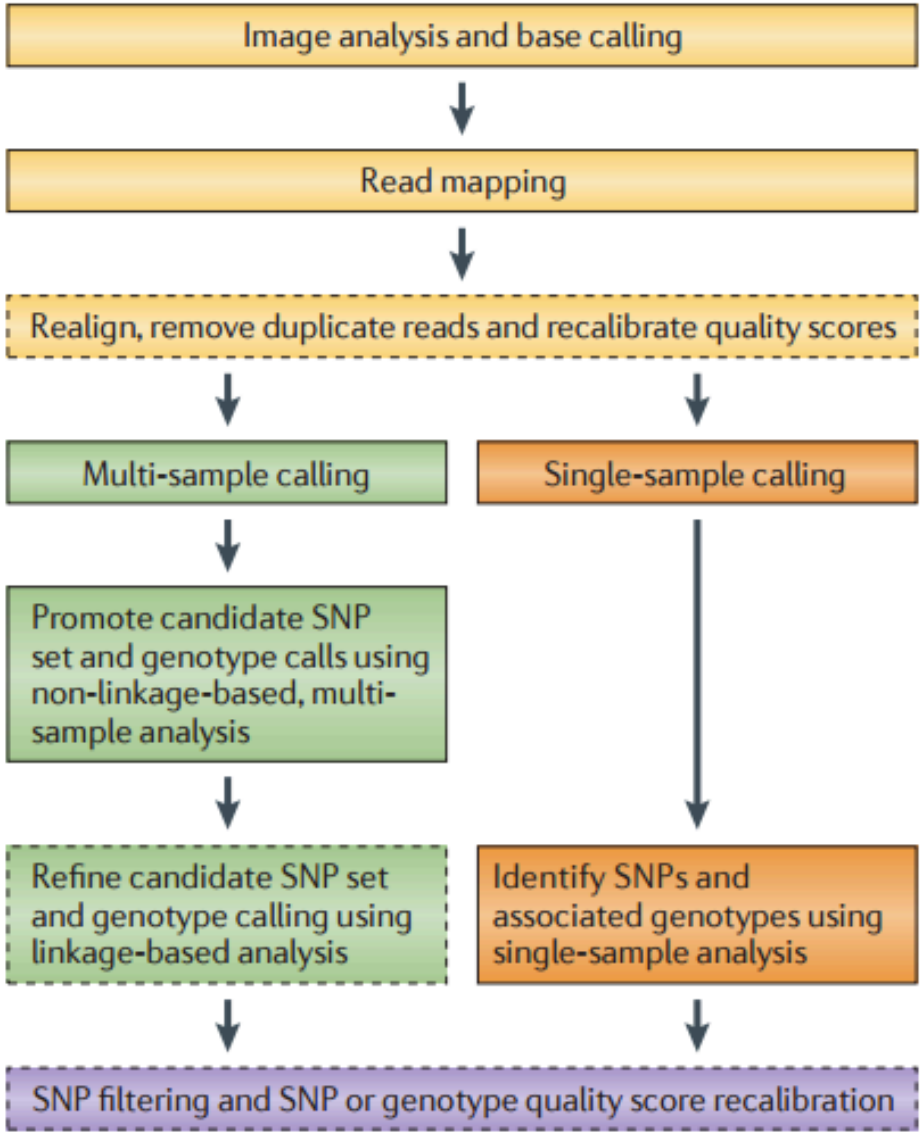
- Can you guess the value of `N` ?

Building haplotypes



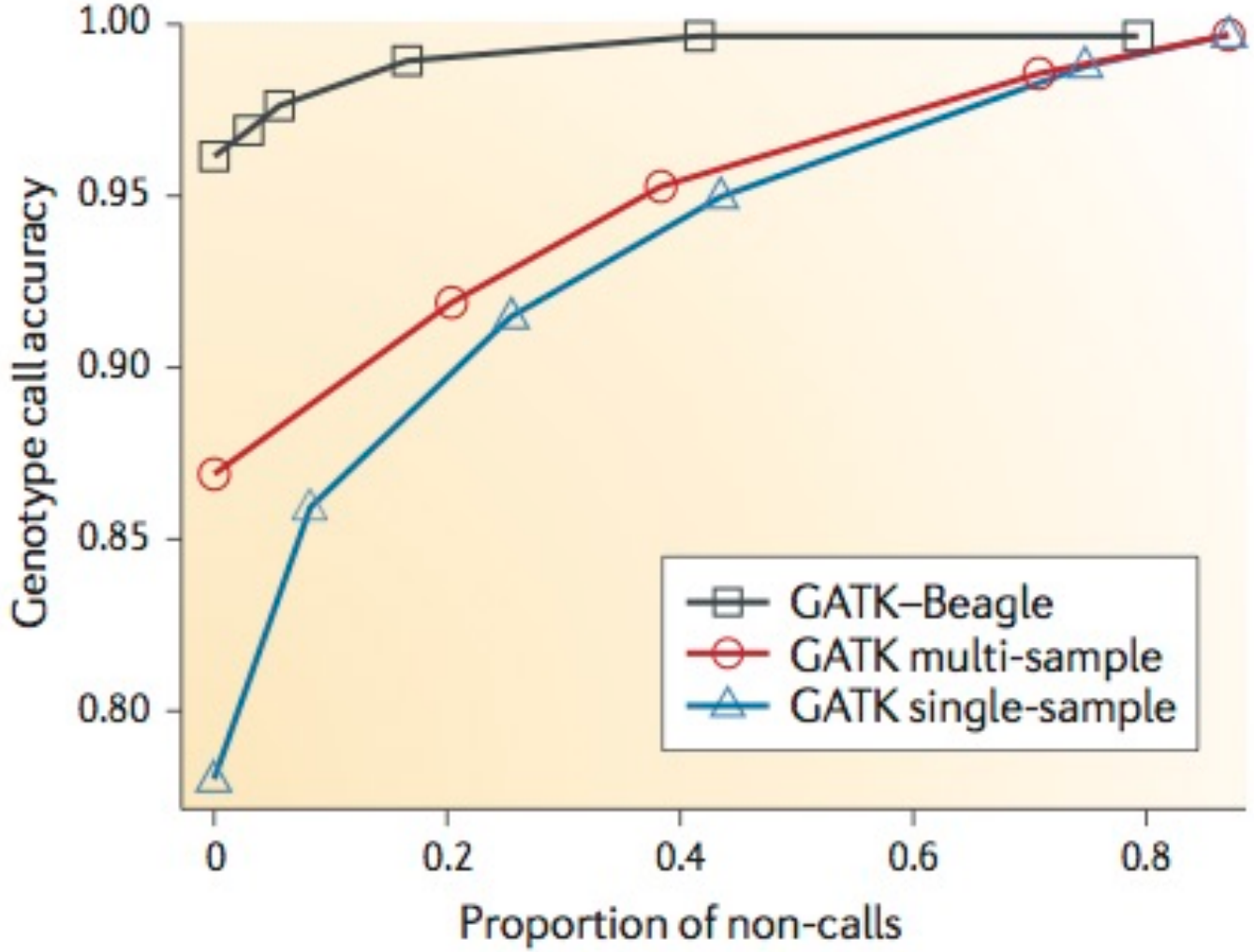
Yoshida et al (2015)

Use multiple samples

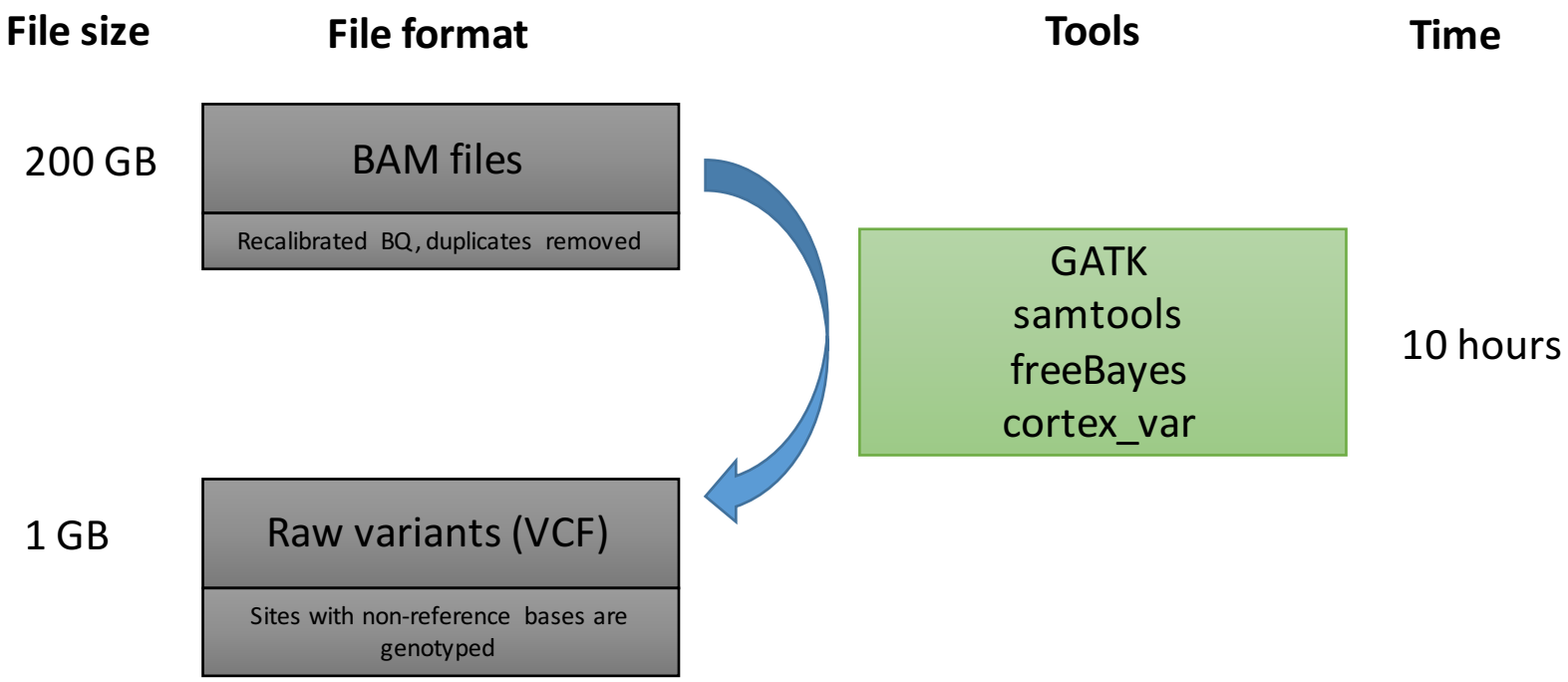


Nielsen et al (2012)

Haplotype imputation increase genotype accuracy



Nielsen et al (2012)



Adapted from Mark DePristo

VCF format

```
##fileformat=VCFv4.2
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS     ID        REF    ALT     QUAL FILTER INFO                              FORMAT      NA00001        NA00002        NA00003
20     14370   rs6054257 G      A       29   PASS   NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:48:1:51,51 1|0:48:8:51,51 1/1:43:5
20     17330   .         T      A       3    q10    NS=3;DP=11;AF=0.017               GT:GQ:DP:HQ 0|0:49:3:58,50 0|1:3:5:65,3   0/0:41:3
20     1110696 rs6040355 A      G,T     67   PASS   NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27 2|1:2:0:18,2  2/2:35:4
20     1230237 .         T      .       47   PASS   NS=3;DP=13;AA=T                   GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:51,51 0/0:61:2
20     1234567 microsat1 GTC    G,GTCT  50   PASS   NS=3;DP=9;AA=G                    GT:GQ:DP    0/1:35:4       0/2:17:2       1/1:40:3
```

Mandatory header line

Mandatory header line

Reference base

Alternative base

Quality score

Allele frequency, read depth, etc.

https://samtools.github.io/hts-specs/VCFv4.2.pdf

Variant filtering

Raw variant calls have a lot of false positives.

- How to filter?
- Which one do you look at first?

Manual filtering based on different parameters

- allele frequency, quality score, depth of coverge…
- Location (contig ends SNPs are usually inaccurate)
- Case by case

look at the strongest effect filter

Annotating variants

- Annotations using reference genomes

  Programs available: SNP-eff, annovar

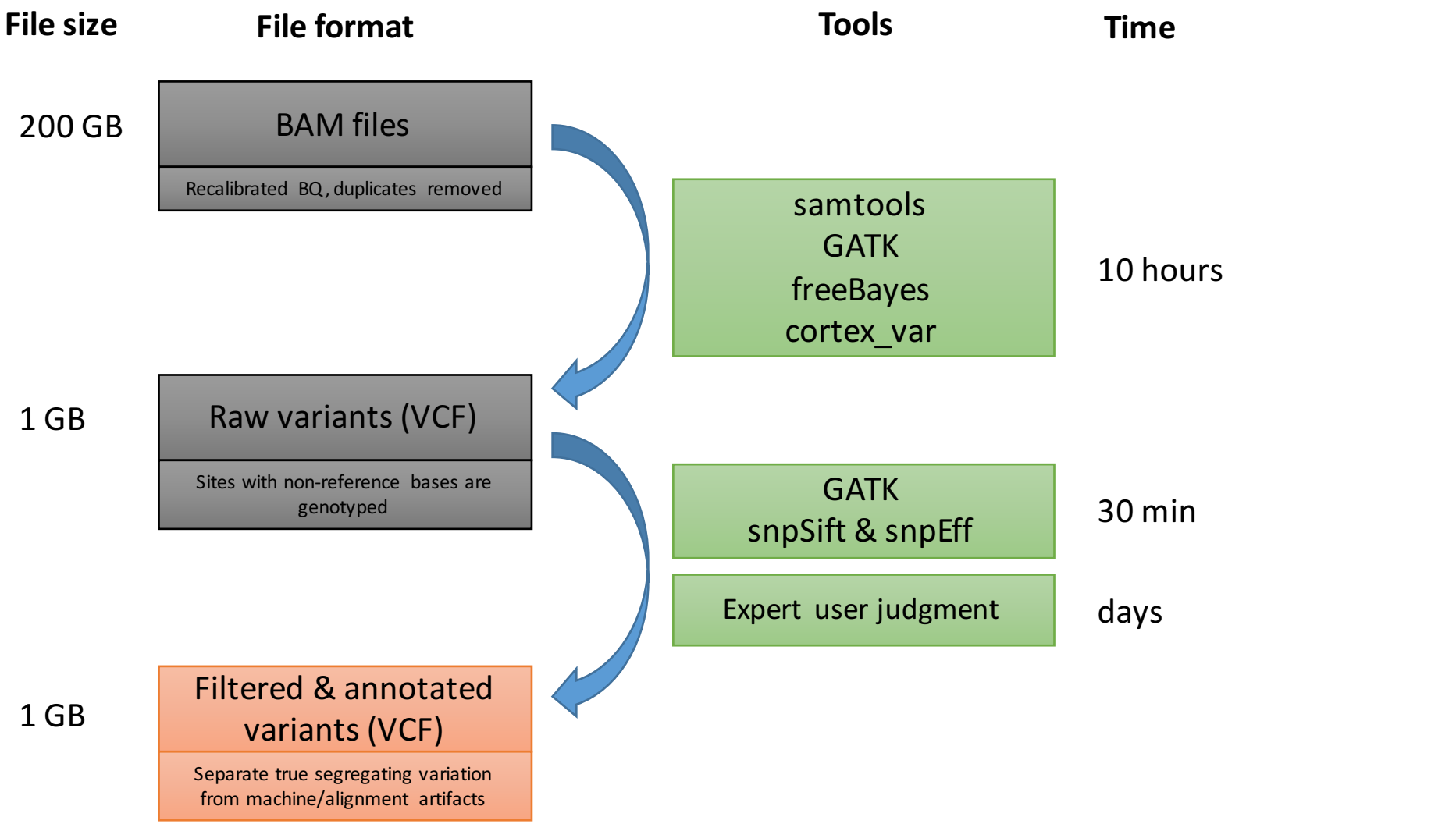
- Calculate effects:
  - Coding (e.g. Syn, Non-Syn, Stop gained, Splice)
  - Non-coding (e.g. TFBS)

One of the mostly intensively research areas:

Linking variation to function

Unfortunately, only applicable to humans

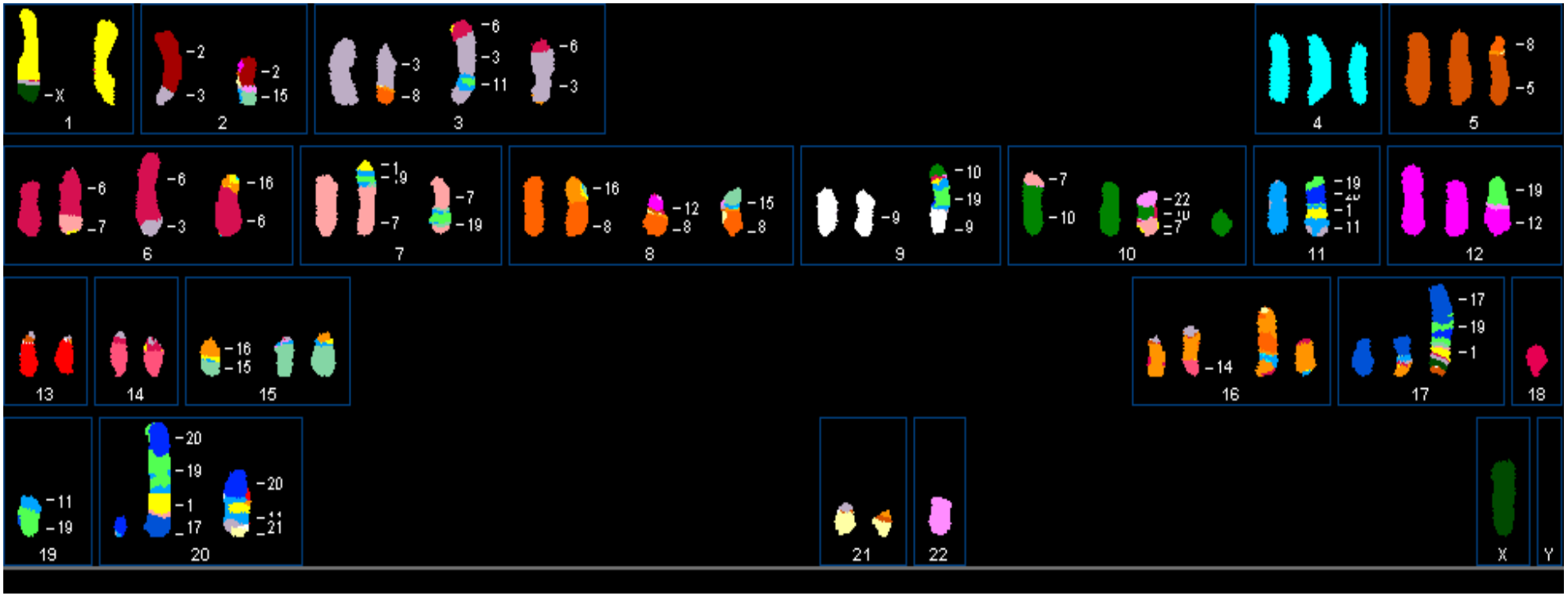
For a new species, you have to start from scratch



Adapted from Mark DePristo

Structural variation

More difficult structural variants



Traditional structural variations
Can we see them in higher resolution?

Alkan et al. 2011

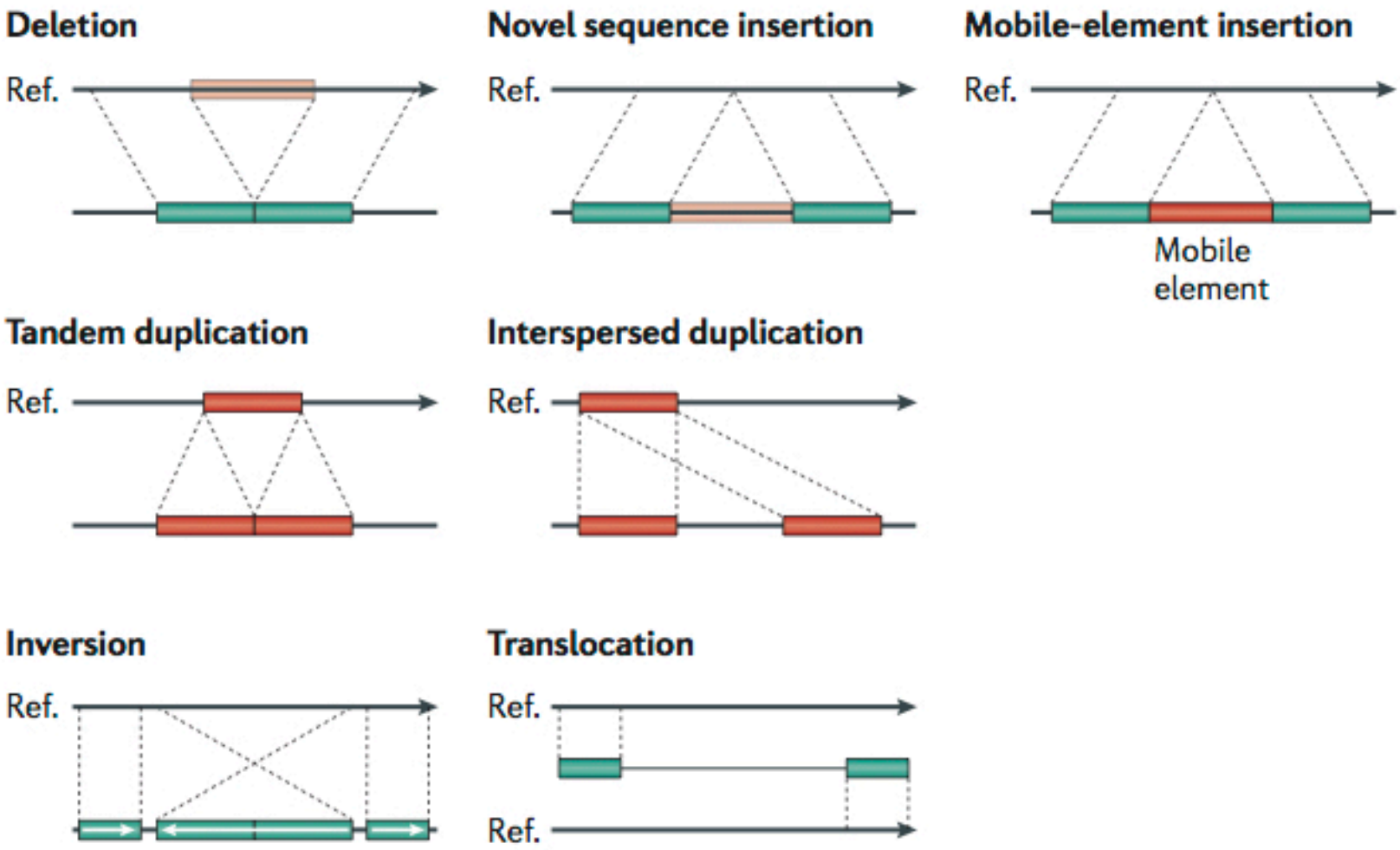
More difficult structural variants

Structural Variants (SVs): Genomic rearrangements that affect >50bp (or 100bp, or 1Kb) of sequence, including:

- deletions
- novel insertions
- inversions
- mobile-element transpositions
- duplications
- translocations

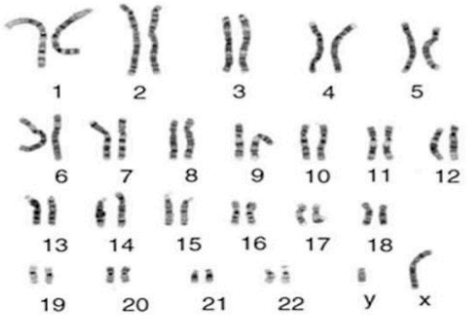
Alkan et al. 2011

SV classes

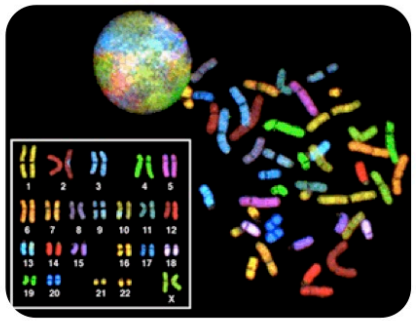


Alkan et al. 2011

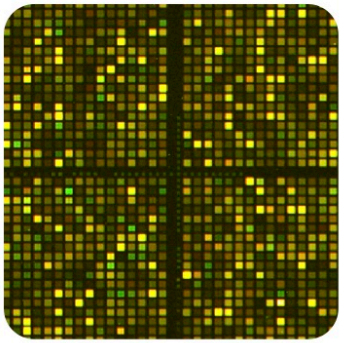
Again, our understanding is driven by technology



1940s - 1980s
Cytogenetics / Karyotyping



1990s
CGH / FISH /
SKY / COBRA

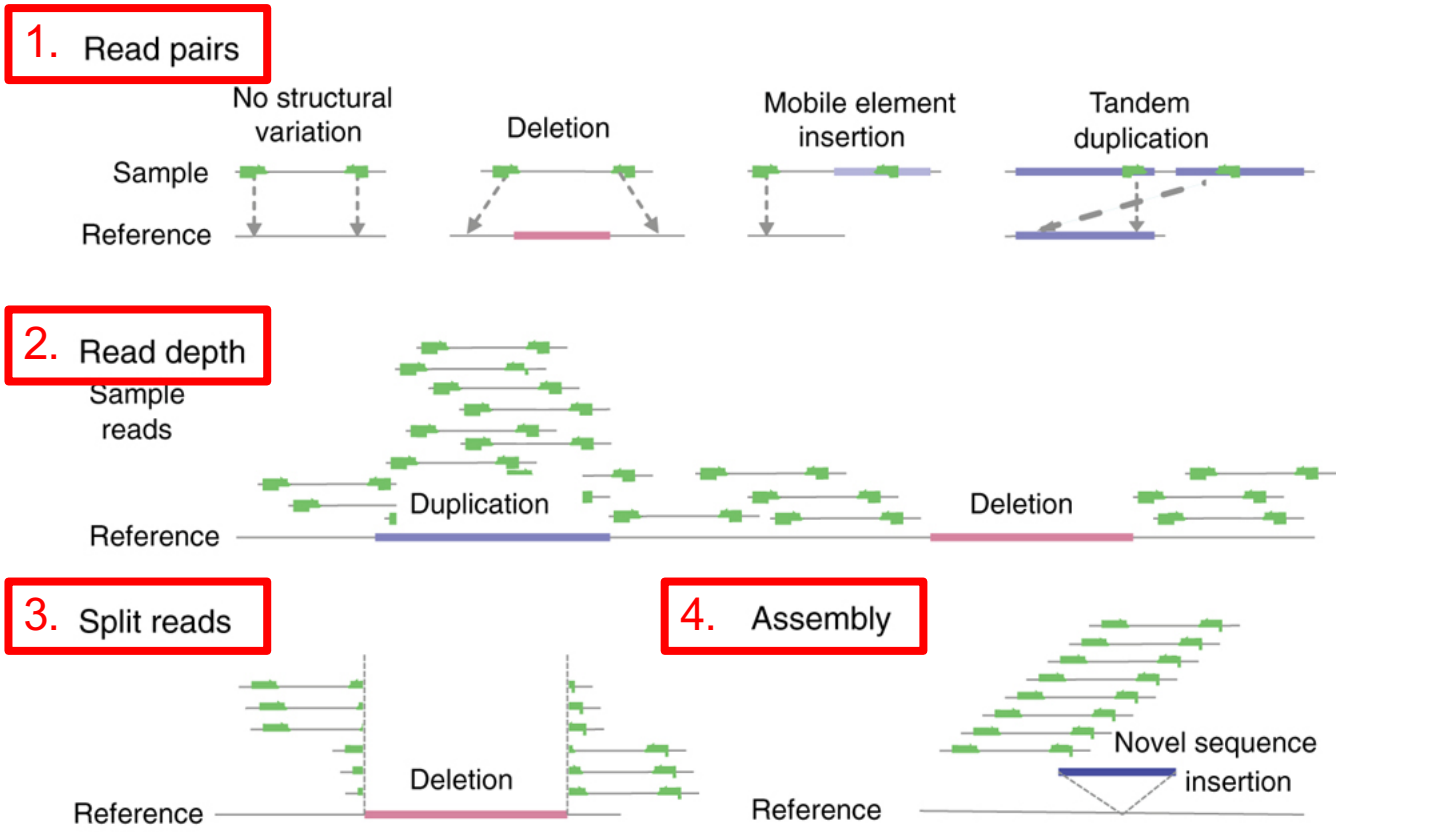


2000s
Genomic microarrays
BAC-aCGH / oligo-aCGH

Today
High throughput
DNA sequencing

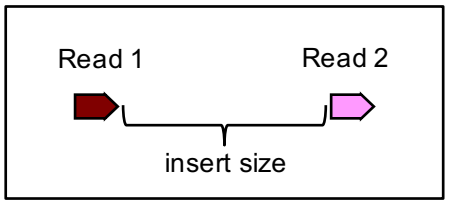
Aaron Quinlan

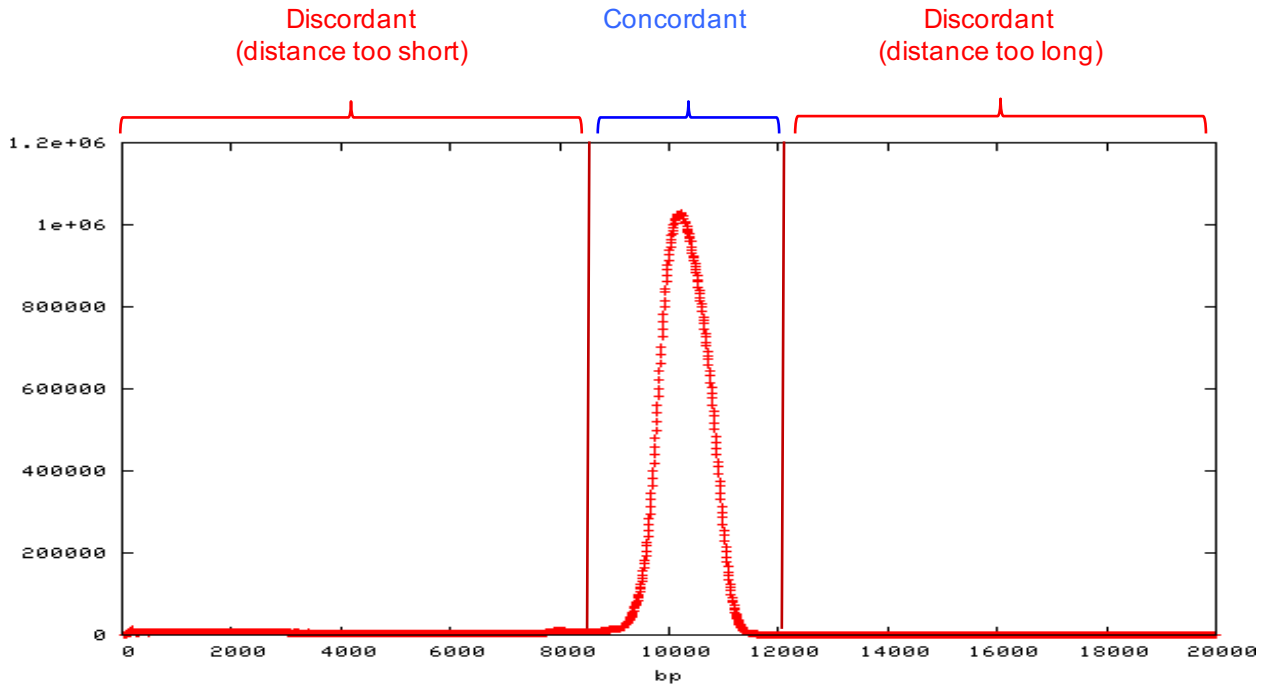
Strategies for calling SVs from NGS data



Baker *Nat Methods* 2012

Discordant read pairs

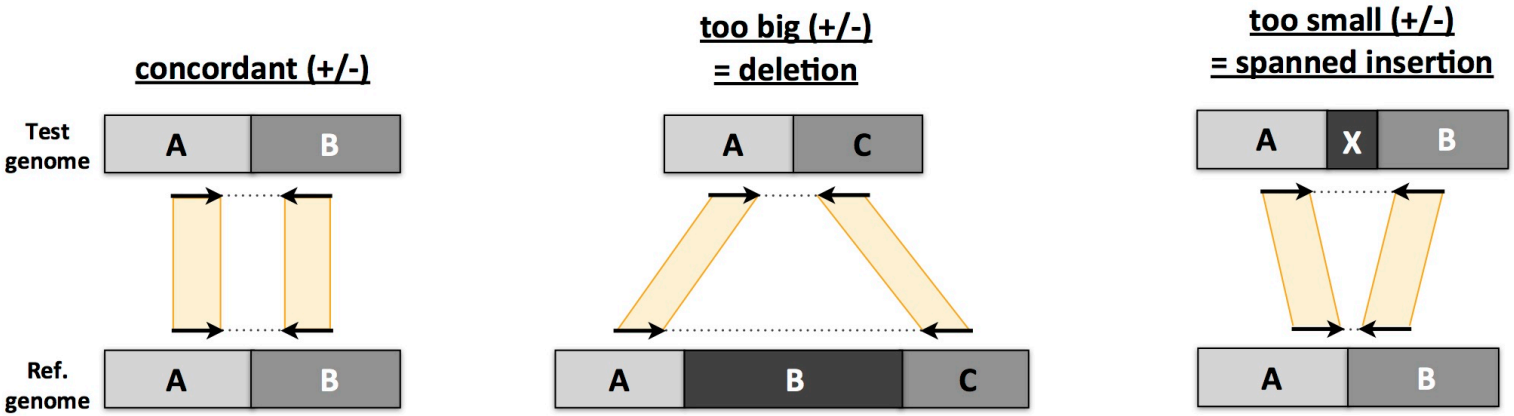


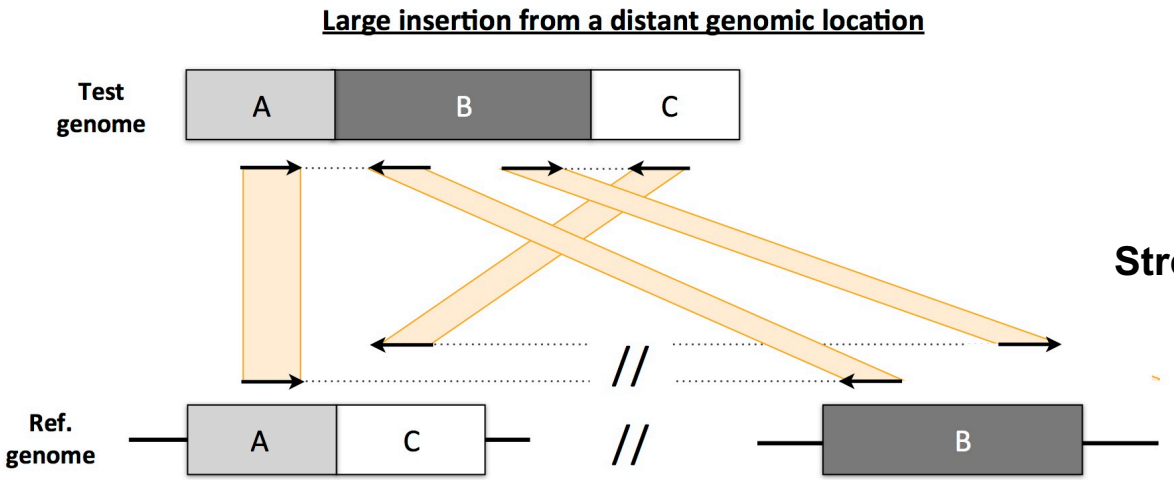


Genomic distance between mapped paired tags

Reads pairs are also Discordant when order or orientation isn't as expected.
Do they fall into particular region of the assembly?

Using discordant reads to detect SVs





Weaknesses

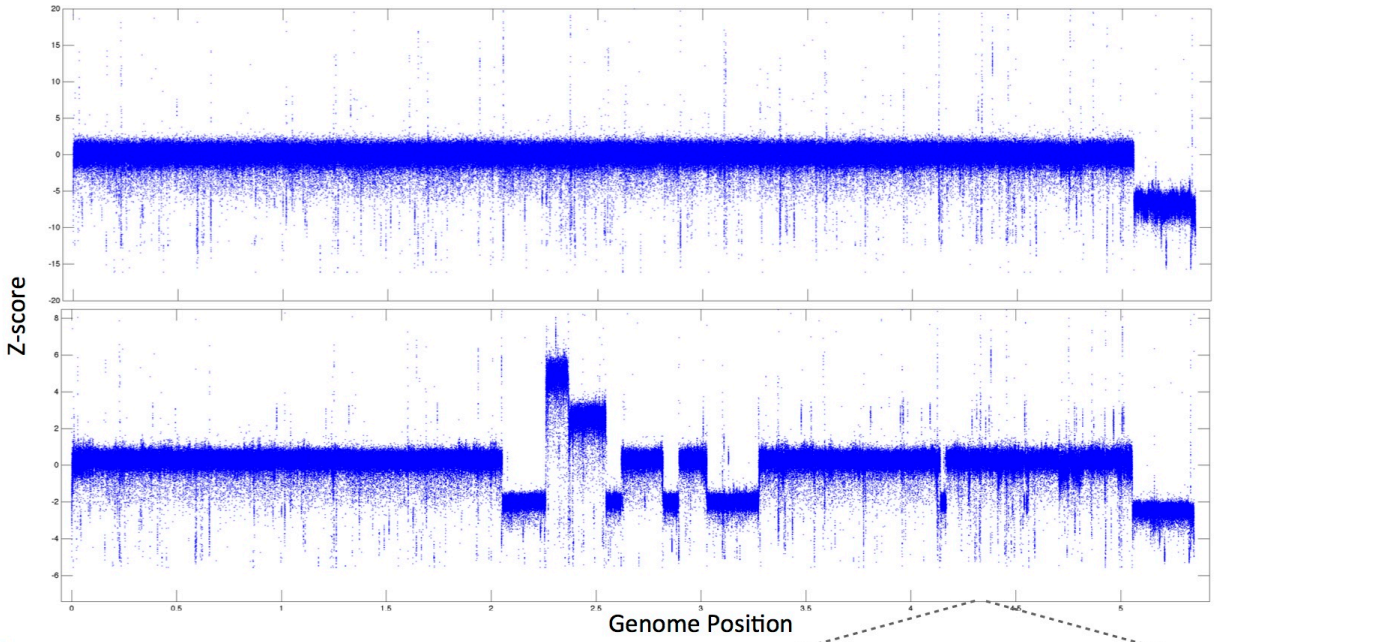
Difficult to interpret read-pairs in repetitive regions
Difficult to fully characterize highly rearranged regions
High rate of false positives

Strengths:

Most classes of variation can, in principle, be detected

Adapted from Aaron Quinlan

Read-depth



Strengths:

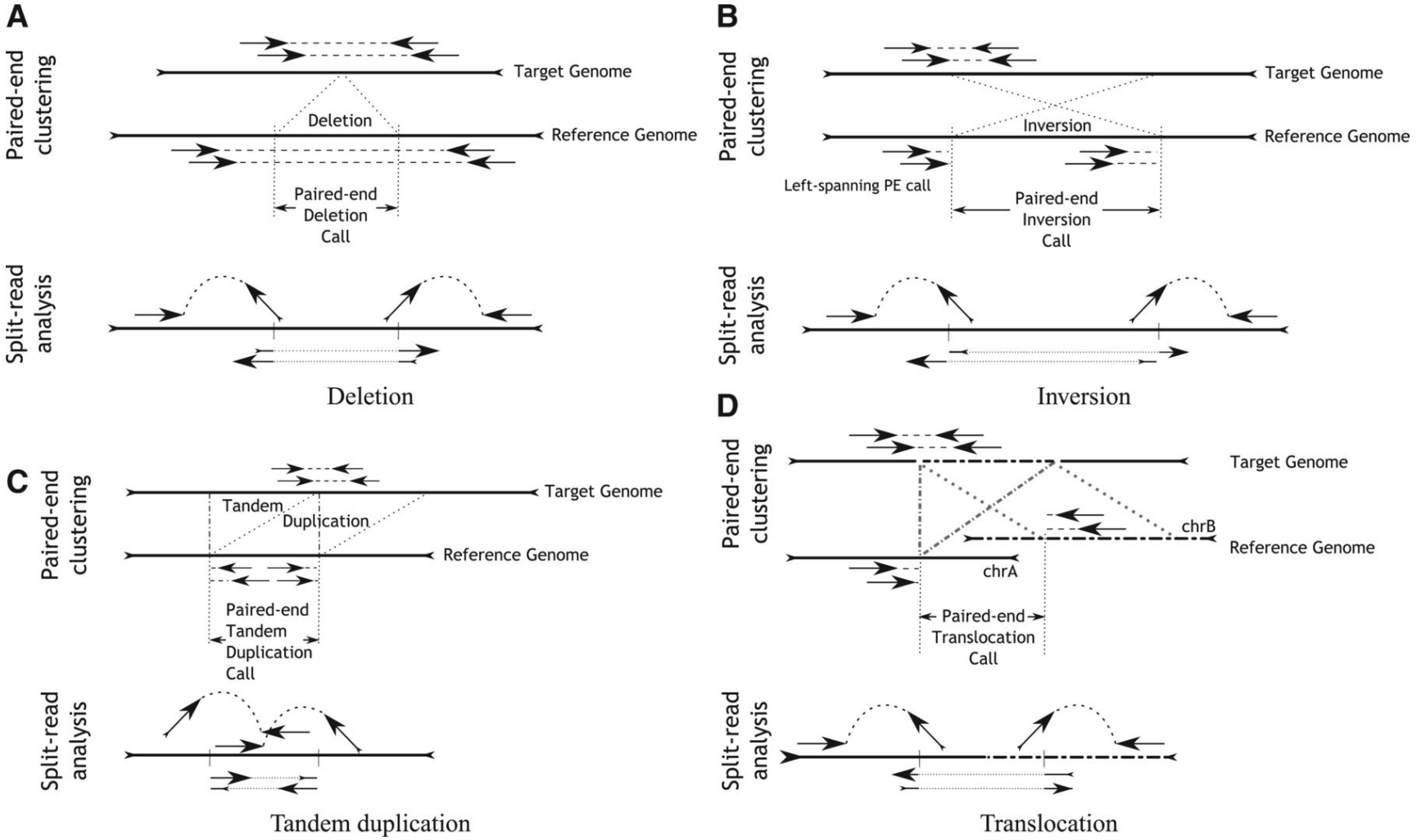
1) Fast and simple.
2) Easy to identify gene amplifications.
3) Relatively straightforward interpretation: is gene X amplified or deleted?

Weaknesses:

1) Limited resolution (5-10kb) = imprecise boundaries
2) Cannot detect balanced events or reveal variant architecture.

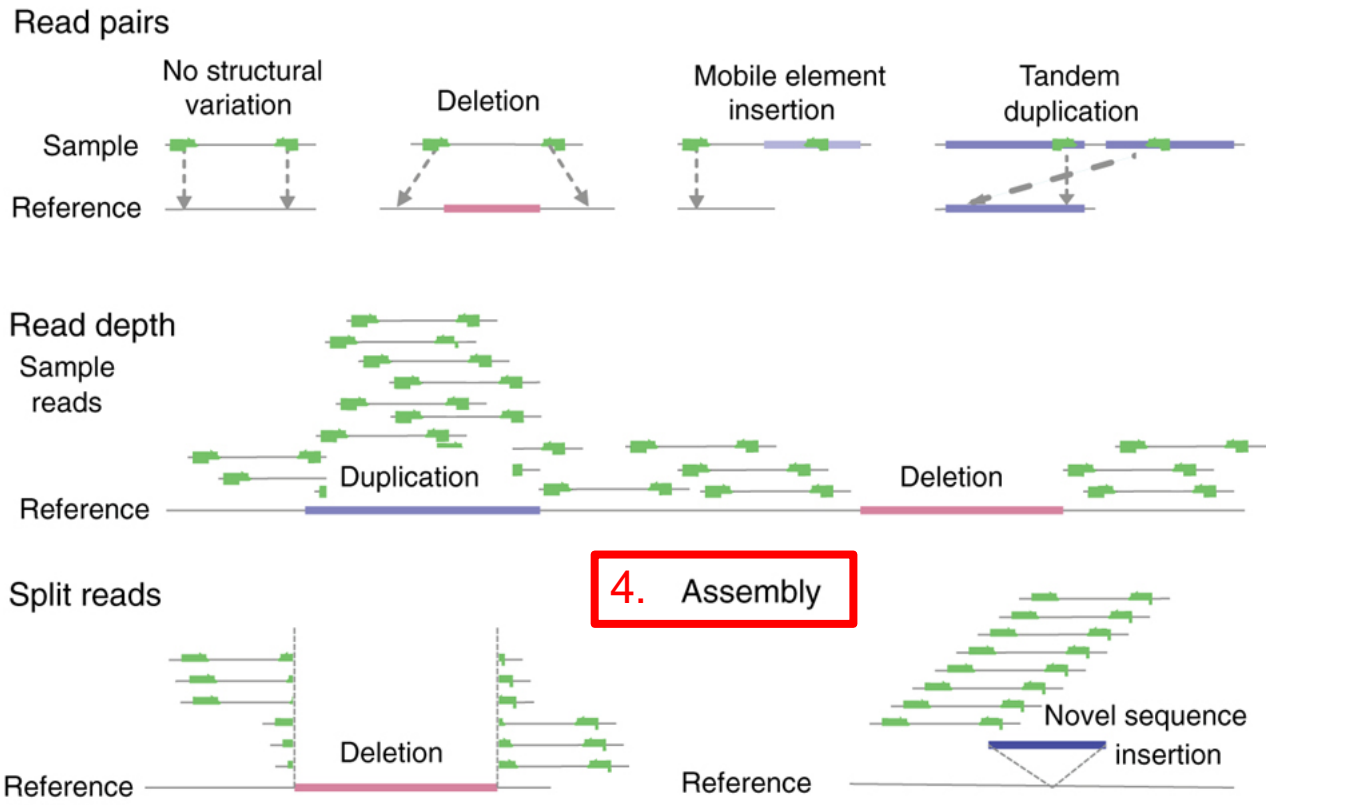
Aaron Quinlan

Split reads



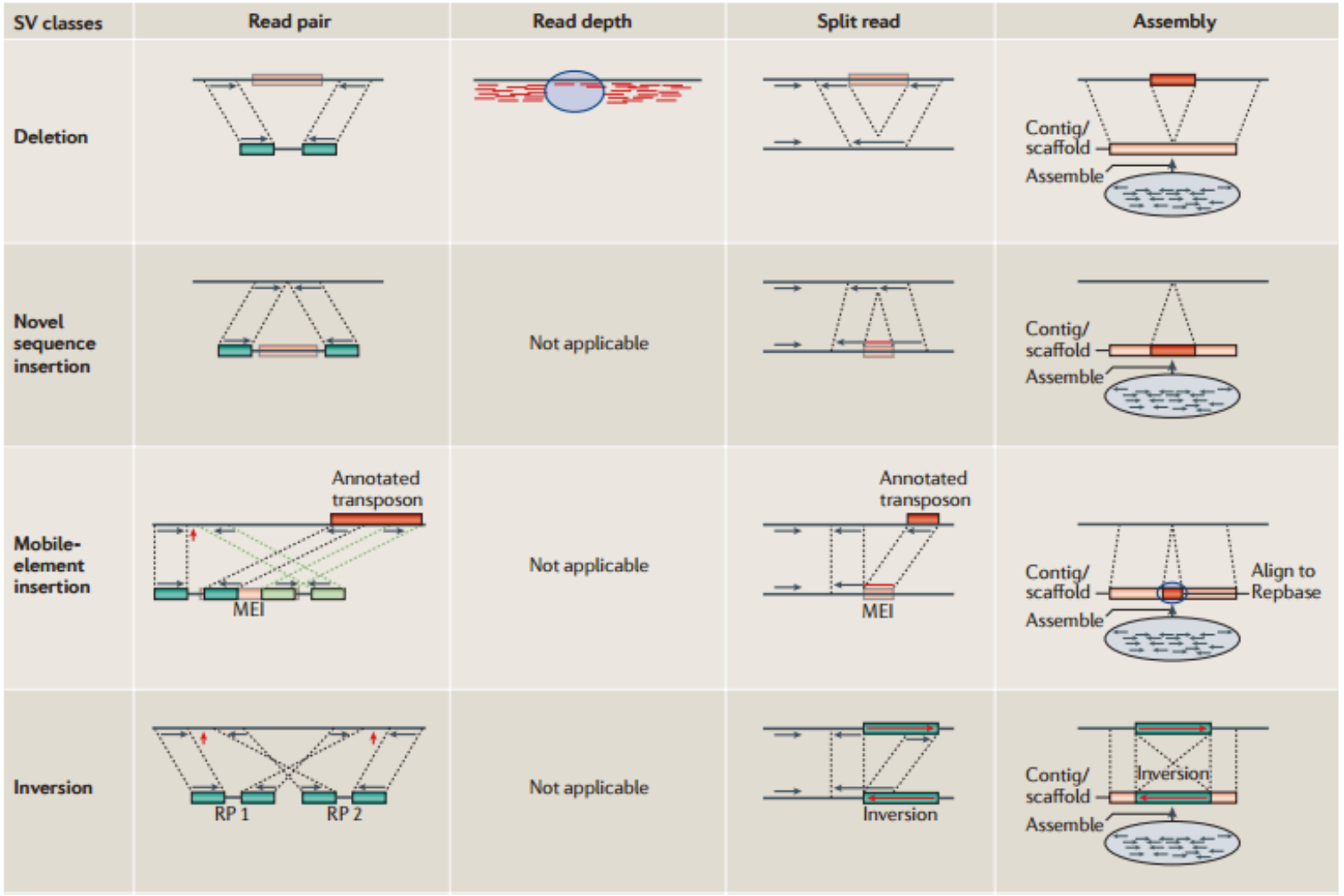
Rausch et al. *Bioinformatics* 2012

Strategies for calling SVs from NGS data



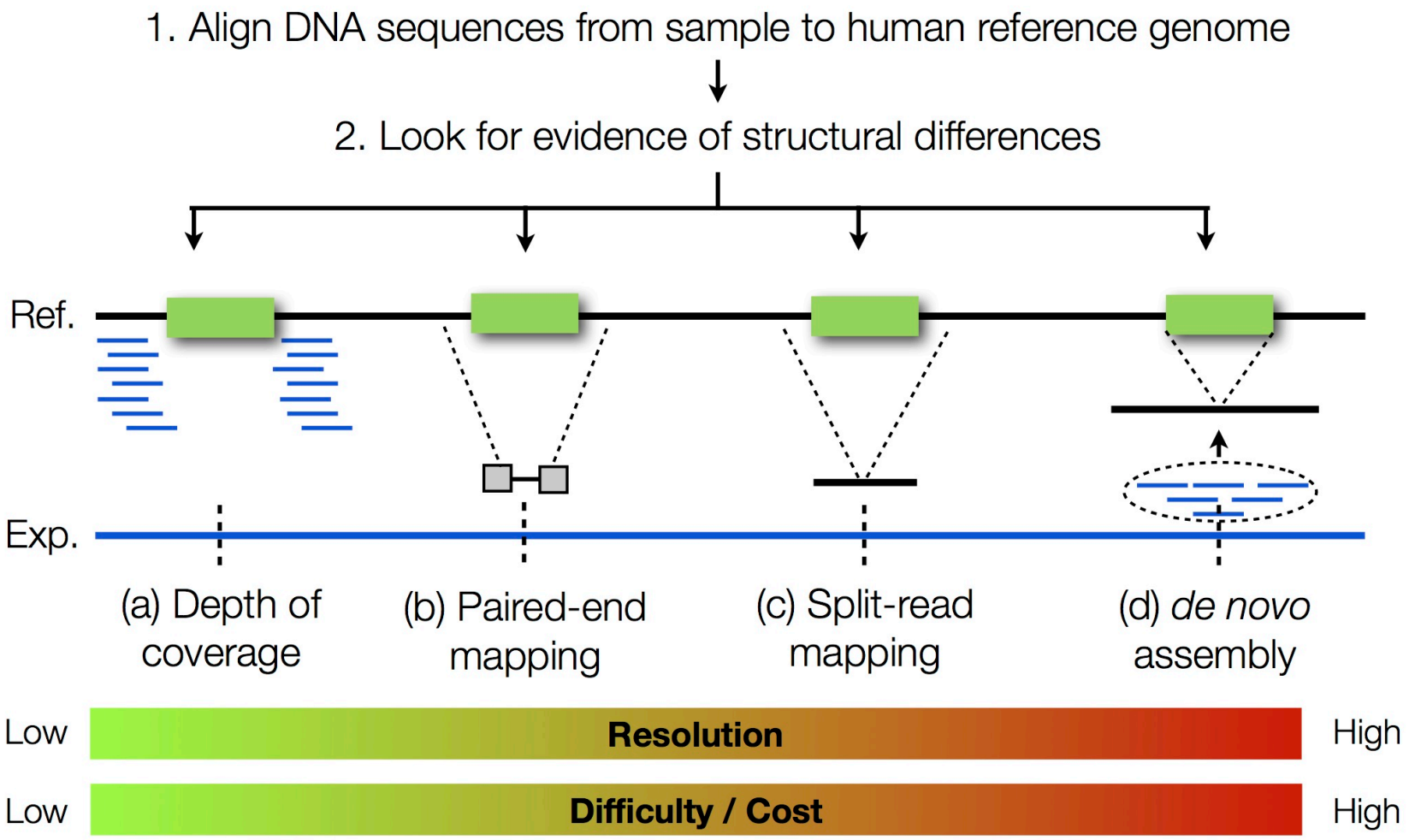
Baker *Nat Methods* 2012

De novo assembly for SVs



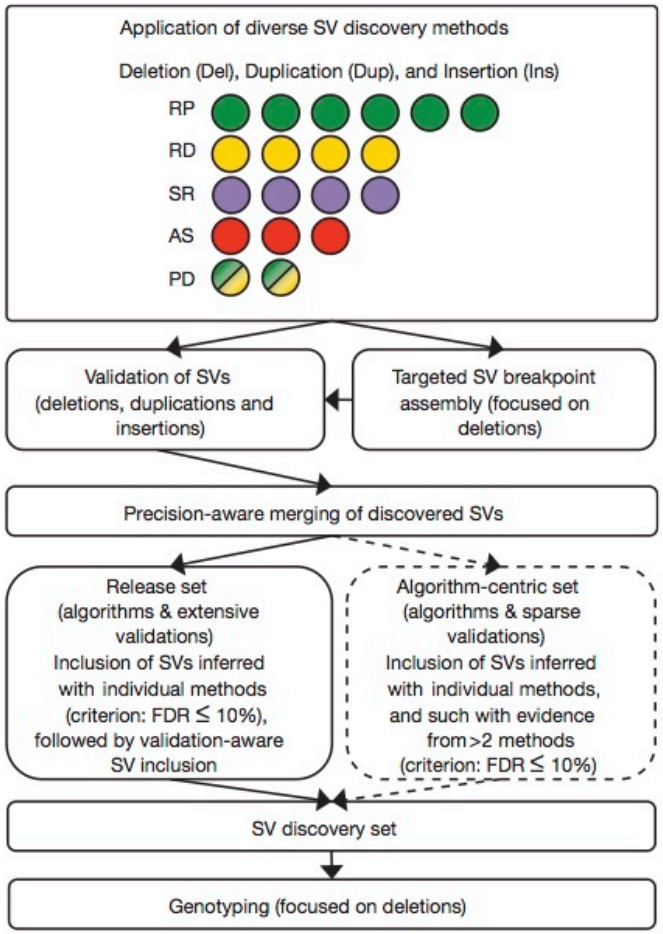
Alkan et al. 2011

Summary of strategies for calling SVs

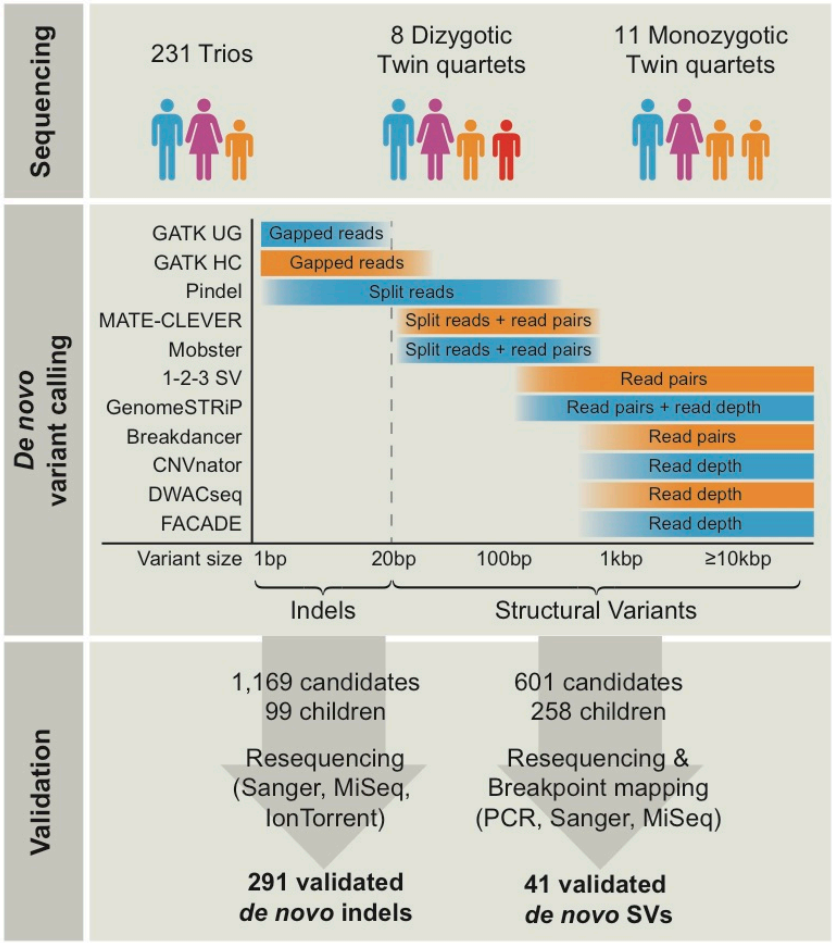


Aaron Quinlan

Bottom line: try many methods and validate

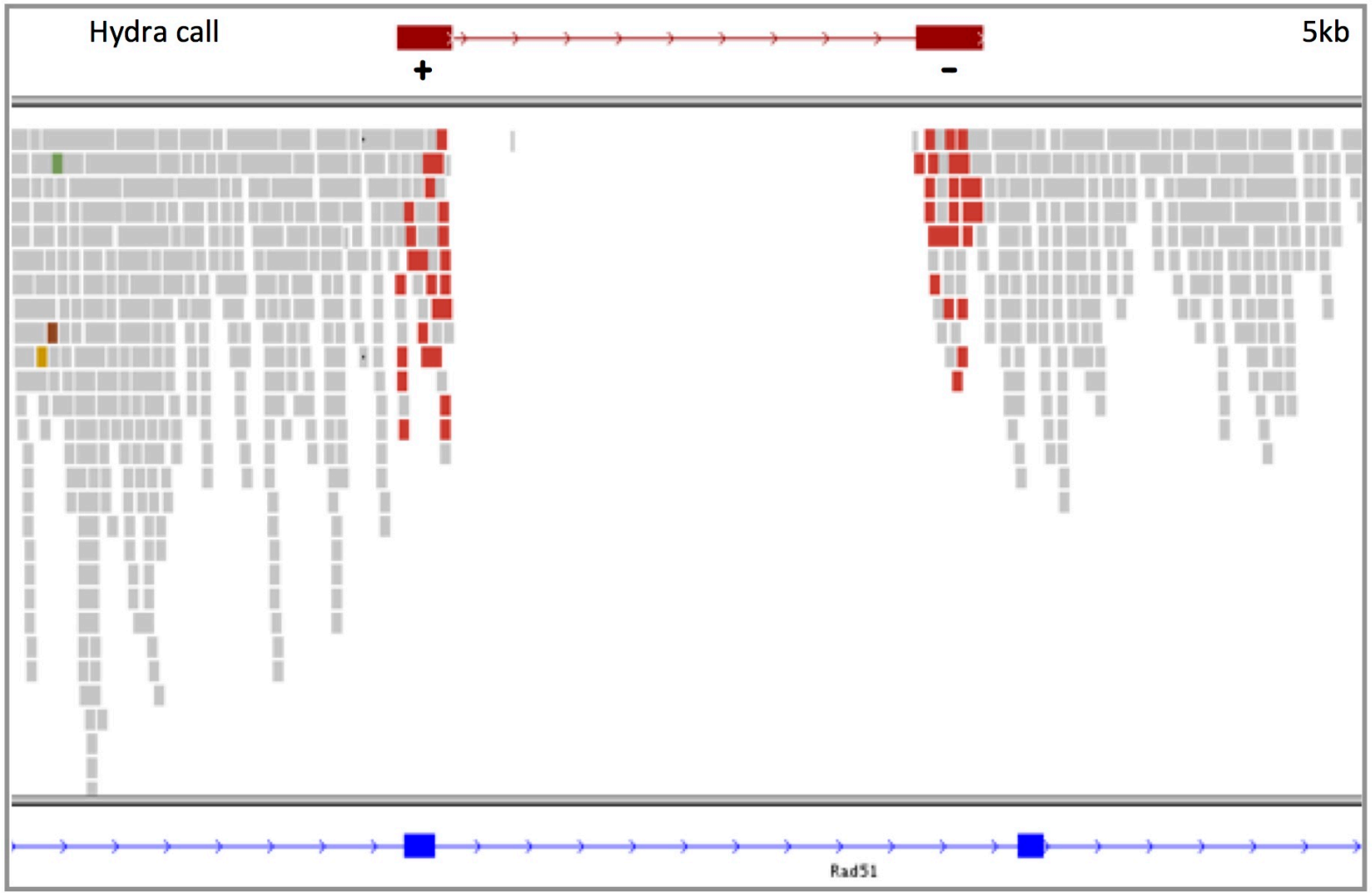


Mills et al. *Nature* 2011



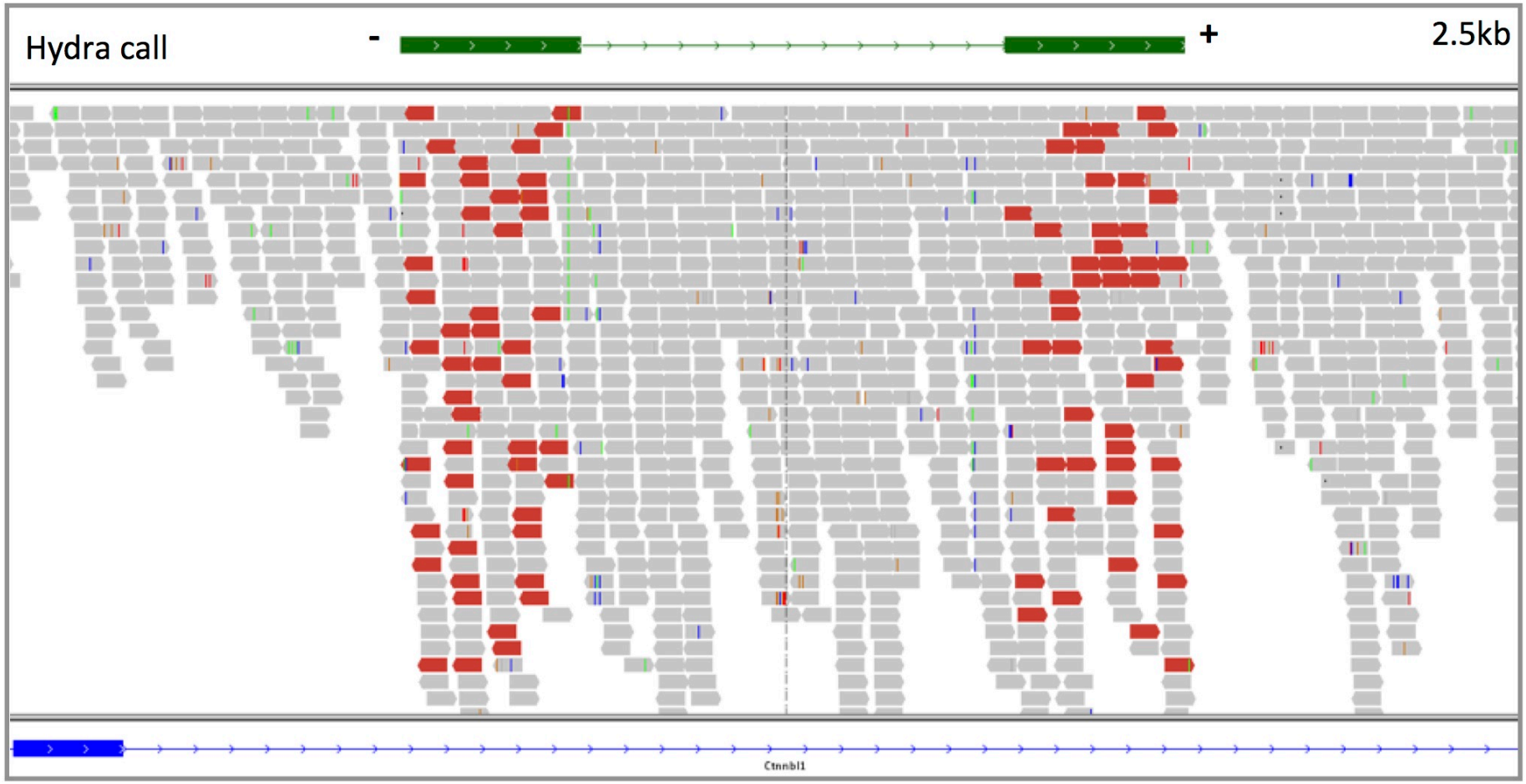
Kloosterman et al. 2015

Visual validation: a deletion



Aaron Quinlan

Visual validation: a duplication



Aaron Quinlan

Structural variants: A summary

Actually it’s all the same methods

Reference assembly -> check depth -> detect duplication

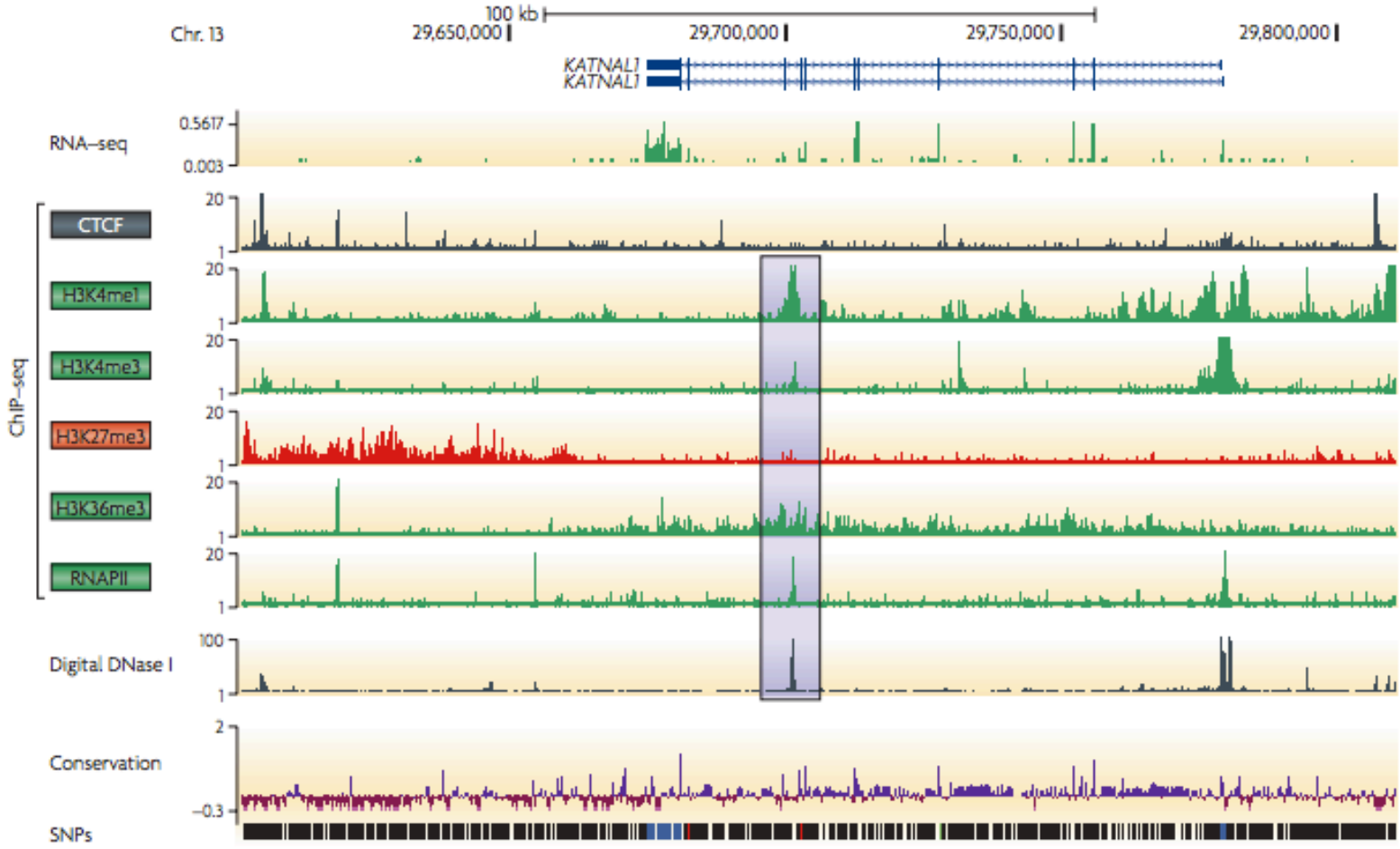
New assembly -> check depth -> detect ploidy chromosomes / mis-assemblies

Other experiment mapping approach

*-seq methodologies

Identify peaks!

How is peak different to coverage?

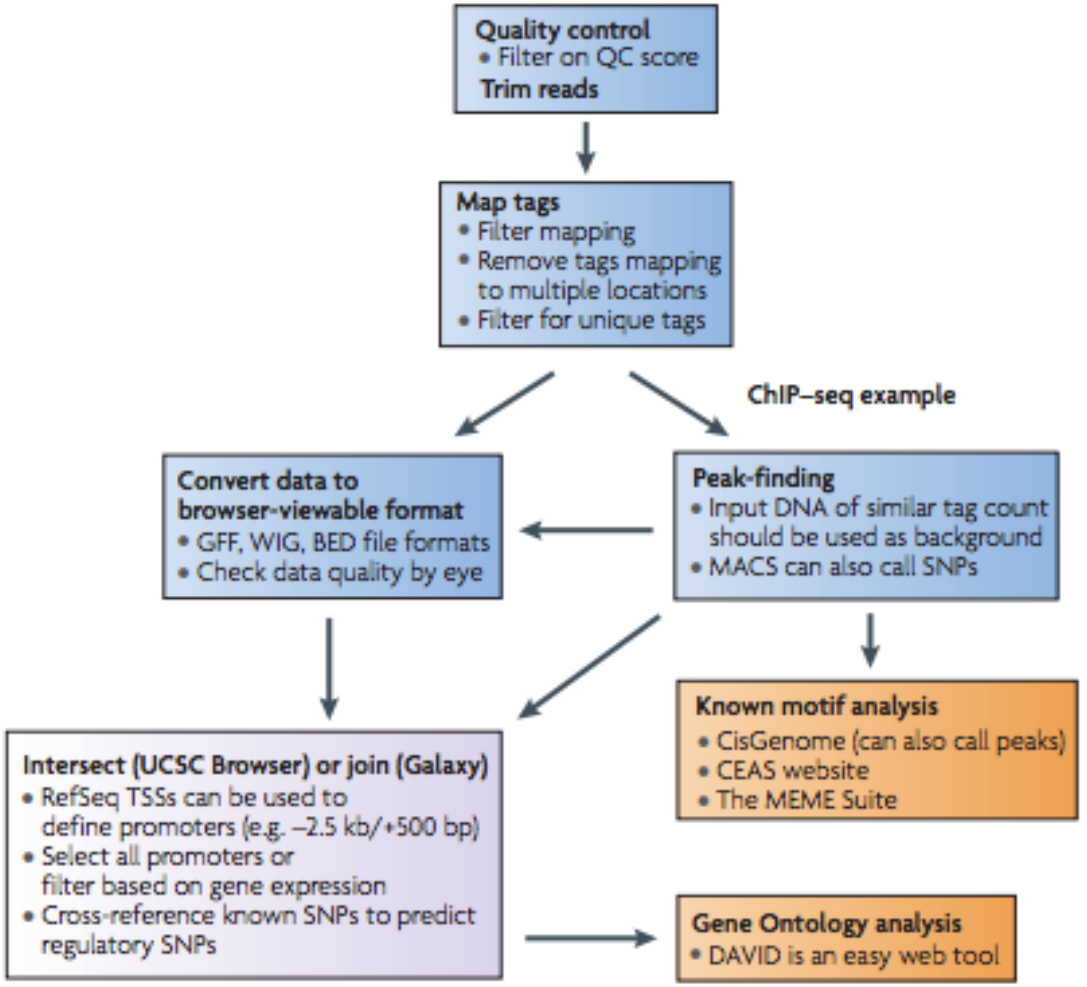


Hawkins et al 2010

Other experiment mapping approach

Similar methods

Different analysis



Hawkins et al 2010

Validation and standardisation

Genome in a Bottle Consortium

The Genome in a Bottle Consortium is a public-private-academic consortium hosted by NIST to develop the technical infrastructure (reference standards, reference methods, and reference data) to enable translation of whole human genome sequencing to clinical practice.

NA12878 cell line, sequenced many platforms, read lengths and sample preps ; A lot and lot of **Benchmarks**

https://sites.stanford.edu/abms/giab

Again, only in humans…

De novo vs mapping approach

Mapping is less complicated and more intuitive

Can gather lots of information from many individuals given a good ref

But, information on repeats/ gene families / *de novo* genes / large structural variants are more difficult to detect

Assembly is powerful but also computationally demanding

And is your question worth the trouble to assemble 100 strains?

In practice, people do a combination of both approaches

In humans, *de novo* genomes of references and cancer cells are being generated. In butterflies, many assemblies to reveal super gene

Case studies

Case study: Check lane quality and assembly

	Total reads	Mapped reads (%)	Duplicates	Proper-pairs	Both pairs mapped	Median
BRC PE	217,190,726	95.80%	1.47%	53.97%	92.85%	968
Old PE	249,742,439	**4.40%**	1.51%	2.81%	3.08%	59
Old PE	1,167,521,211	98.21%	**11.81%**	68.97%	97.18%	465
Old PE	917,638,787	97.97%	**5.12%**	75.54%	96.99%	261
Company hmm	38,508,236	94.15%	**7.51%**	48.22%	90.53%	1681
Company hmm	76,992,221	95.09%	**10.75%**	48.57%	92.43%	1675
Company hmm	26,348,302	93.54%	**6.23%**	47.58%	89.29%	1681
Company hmm	398,746,361	98.42%	**79.36%**	57.28%	97.23%	1500
Company hmm	396,241,991	98.42%	**79.03%**	57.31%	97.24%	1500
Company hmm	39,879,176	92.45%	**29.40%**	40.66%	88.55%	4623
Company hmm	43,010,934	92.10%	**31.27%**	40.40%	87.99%	4623
Company hmm	316,963,201	97.71%	**84.14%**	57.79%	96.11%	410
Company hmm	71,118,483	96.00%	**70.88%**	42.97%	93.67%	283
Company hmm	61,803,780	94.60%	**73.18%**	45.36%	91.55%	285

Case study: Check lane quality and assembly

PE from one of my projects

Directory	TotalReads	MappedReads	MappedReads(%)	ForwardStrand	ForwardStrand(%)	ReverseStrand	ReverseStrand(%)	FailedQC	FailedQC(%)	Duplicates	Duplicates(%)
map.718-S1	3171334	2920504	92.09%	1461181	46.07%	1459323	46.02%	0	0%	28099	0.89%
map.KPN91	2963812	2926683	98.75%	1463794	49.39%	1462889	49.36%	0	0%	8464	0.29%
map.KPN92	38811800	37864479	97.56%	18931099	48.78%	18933380	48.78%	0	0%	614696	1.58%
map.NTU	151505774	140827017	92.95%	70410162	46.47%	70416855	46.48%	0	0%	72381797	47.77%

MP from a company in Japan (Nextera)

3kb.map	29432914	28598764	97.17%	14280601	48.52%	14318163	48.65%	0	0%	2369419	8.05%
5kb.map	8887196	8436683	94.93%	4208676	47.36%	4228007	47.57%	0	0%	1455786	16.38%

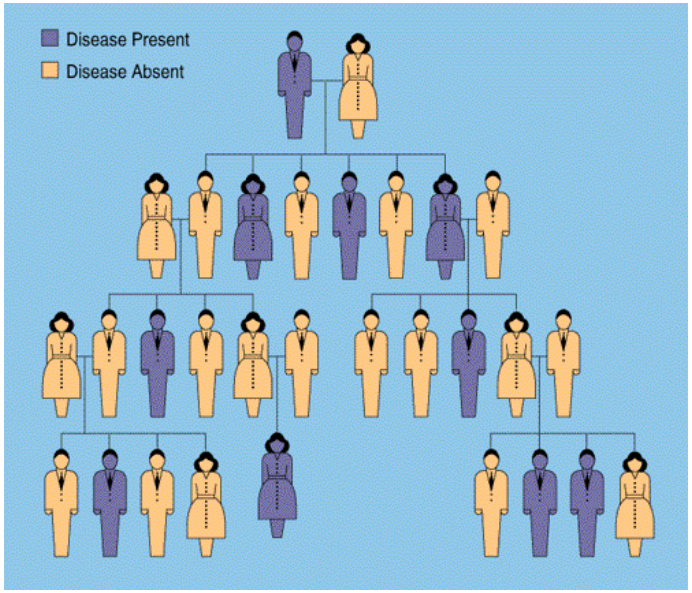
MP from another institute

Directory	TotalReads	MappedReads	MappedReads(%)	ForwardStrand	ForwardStrand(%)	ReverseStrand	ReverseStrand(%)	FailedQC	FailedQC(%)	Duplicates	Duplicates(%)
MP.2kb	87149392	61884157	71.01%	31033842	35.61%	30850315	35.40%	0	0%	6893810	7.91%
MP.4kb	92488172	60082343	64.96%	30124954	32.57%	29957389	32.39%	0	0%	6688542	7.23%
MP.6kb	79969510	50558184	63.22%	25273991	31.60%	25284193	31.62%	0	0%	3919754	4.90%
MP.9kb	63262972	44161175	69.81%	22132740	34.99%	22028435	34.82%	0	0%	6938278	10.97%

a project from BRC

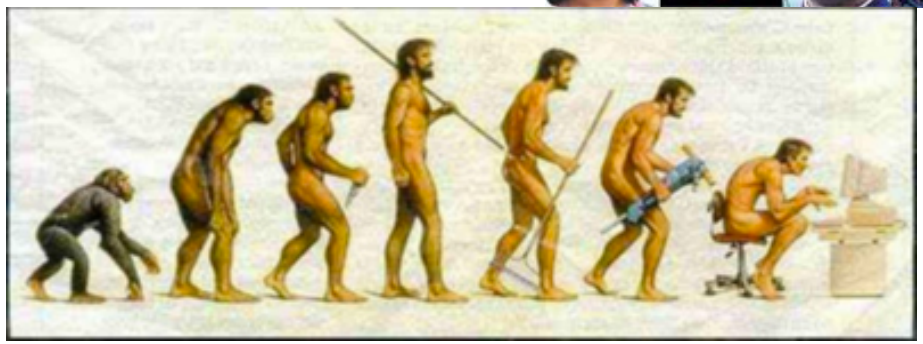
Directory	TotalReads	MappedReads	MappedReads(%)	ForwardStrand	ForwardStrand(%)	ReverseStrand	ReverseStrand(%)	FailedQC	FailedQC(%)	Duplicates	Duplicates(%)
MP10kb.map	4809196	4757184	98.92%	2380384	49.50%	2376800	49.42%	0	0%	31589	0.66%
MP15kb.map	4557418	4492023	98.57%	2247623	49.32%	2244400	49.25%	0	0%	101159	2.22%
MP4kb.map	5349212	5266083	98.45%	2633803	49.24%	2632280	49.21%	0	0%	26721	0.50%
MP6kb.map	5185824	5129611	98.92%	2566177	49.48%	2563434	49.43%	0	0%	30809	0.59%

Human genome **re**sequencing



Inherited diseases

Phenotypic differences



Ancestral history

Human genome resequencing

nature Vol 452|17 April 2008|doi:10.1038/nature06884

LETTERS

The complete genome of an individual by massively parallel DNA sequencing

David A. Wheeler[1]*, Maithreyan Srinivasan[2]*, Michael Egholm[2]*, Yufeng Shen[1]*, Lei Chen[1], Amy McGuire[3], Wen He[2], Yi-Ju Chen[2], Vinod Makhijani[2], G. Thomas Roth[2], Xavier Gomes[2], Karrie Tartaro[2]†, Faheem Niazi[2], Cynthia L. Turcotte[2], Gerard P. Irzyk[2], James R. Lupski[4,5,6], Craig Chinault[4], Xing-zhi Song[1], Yue Liu[1], Ye Yuan[1], Lynne Nazareth[1], Xiang Qin[1], Donna M. Muzny[1], Marcel Margulies[2], George M. Weinstock[1,4], Richard A. Gibbs[1,4] & Jonathan M. Rothberg[2]†

nature Vol 456|6 November 2008|doi:10.1038/nature07484

ARTICLES

The diploid genome sequence of an Asian individual

Jun Wang[1,2,3,4]*, Wei Wang[1,3]*, Ruiqiang Li[1,3,4]*, Yingrui Li[1,5,6]*, Geng Tian[1,7], Laurie Goodman[1], Wei Fan[1], Junqing Zhang[1], Jun Li[1], Juanbin Zhang[1], Yiran Guo[1,7], Binxiao Feng[1], Heng Li[1,8], Yao Lu[1], Xiaodong Fang[1], Huiqing Liang[1], Zhenglin Du[1], Dong Li[1], Yiqing Zhao[1,7], Yujie Hu[1,7], Zhenzhen Yang[1], Hancheng Zheng[1], Ines Hellmann[9], Michael Inouye[8], John Pool[9], Xin Yi[1,7], Jing Zhao[1], Jinjie Duan[1], Yan Zhou[1], Junjie Qin[1,7], Lijia Ma[1,7], Guoqing Li[1], Zhentao Yang[1], Guojie Zhang[1,7], Bin Yang[1], Chang Yu[1], Fang Liang[1,7], Wenjie Li[1], Shaochuan Li[1], Dawei Li[1], Peixiang Ni[1], Jue Ruan[1,7], Qibin Li[1,7], Hongmei Zhu[1], Dongyuan Liu[1], Zhike Lu[1], Ning Li[1,7], Guangwu Guo[1,7], Jianguo Zhang[1], Jia Ye[1], Lin Fang[1], Qin Hao[1,7], Quan Chen[1,5], Yu Liang[1,7], Yeyang Su[1,7], A. san[1,7], Cuo Ping[1,7], Shuang Yang[1], Fang Chen[1,7], Li Li[1], Ke Zhou[1], Hongkun Zheng[1,4], Yuanyuan Ren[1], Ling Yang[1], Yang Gao[1,6], Guohua Yang[1,2], Zhuo Li[1], Xiaoli Feng[1], Karsten Kristiansen[4], Gane Ka-Shu Wong[1,10], Rasmus Nielsen[9], Richard Durbin[8], Lars Bolund[1,11], Xiuqing Zhang[1,6], Songgang Li[1,2,5], Huanming Yang[1,2,3] & Jian Wang[1,2,3]

OPEN ACCESS Freely available online PLoS BIOLOGY

The Diploid Genome Sequence of an Individual Human

Samuel Levy[1*], Granger Sutton[1], Pauline C. Ng[1], Lars Feuk[2], Aaron L. Halpern[1], Brian P. Walenz[1], Nelson Axelrod[1], Jiaqi Huang[1], Ewen F. Kirkness[1], Gennady Denisov[1], Yuan Lin[1], Jeffrey R. MacDonald[2], Andy Wing Chun Pang[2], Mary Shago[2], Timothy B. Stockwell[1], Alexia Tsiamouri[1], Vineet Bafna[3], Vikas Bansal[3], Saul A. Kravitz[1], Dana A. Busam[1], Karen Y. Beeson[1], Tina C. McIntosh[1], Karin A. Remington[1], Josep F. Abril[4], John Gill[1], Jon Borman[1], Yu-Hui Rogers[1], Marvin E. Frazier[1], Stephen W. Scherer[2], Robert L. Strausberg[1], J. Craig Venter[1]

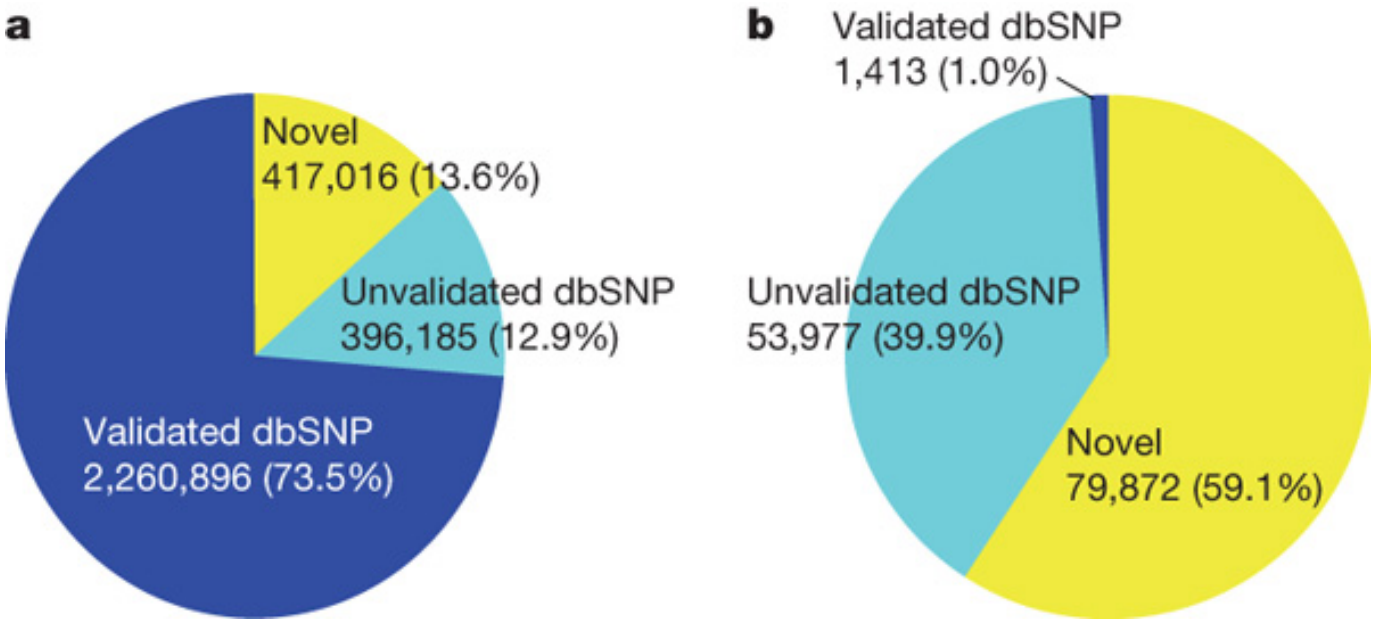
1 J. Craig Venter Institute, Rockville, Maryland, United States of America, 2 Program in Genetics and Genomic Biology, The Hospital for Sick Children, and Molecular and Medical Genetics, University of Toronto, Toronto, Ontario, Canada, 3 Department of Computer Science and Engineering, University of California San Diego, La Jolla, California, United States of America, 4 Genetics Department, Facultat de Biologia, Universitat de Barcelona, Barcelona, Catalonia, Spain

Sequencing of an Asian individual

- Illumina paired-end 35bp sequence reads
- 3.3billion reads -> 117.7 Gigabases of data
- Aligned to the NCBI genome using SOAP (87.4% of data);
- 36x fold coverage.

Wang et al., 2008 (Nature)

Sequencing of an Asian individual



Depth matters in sequencing

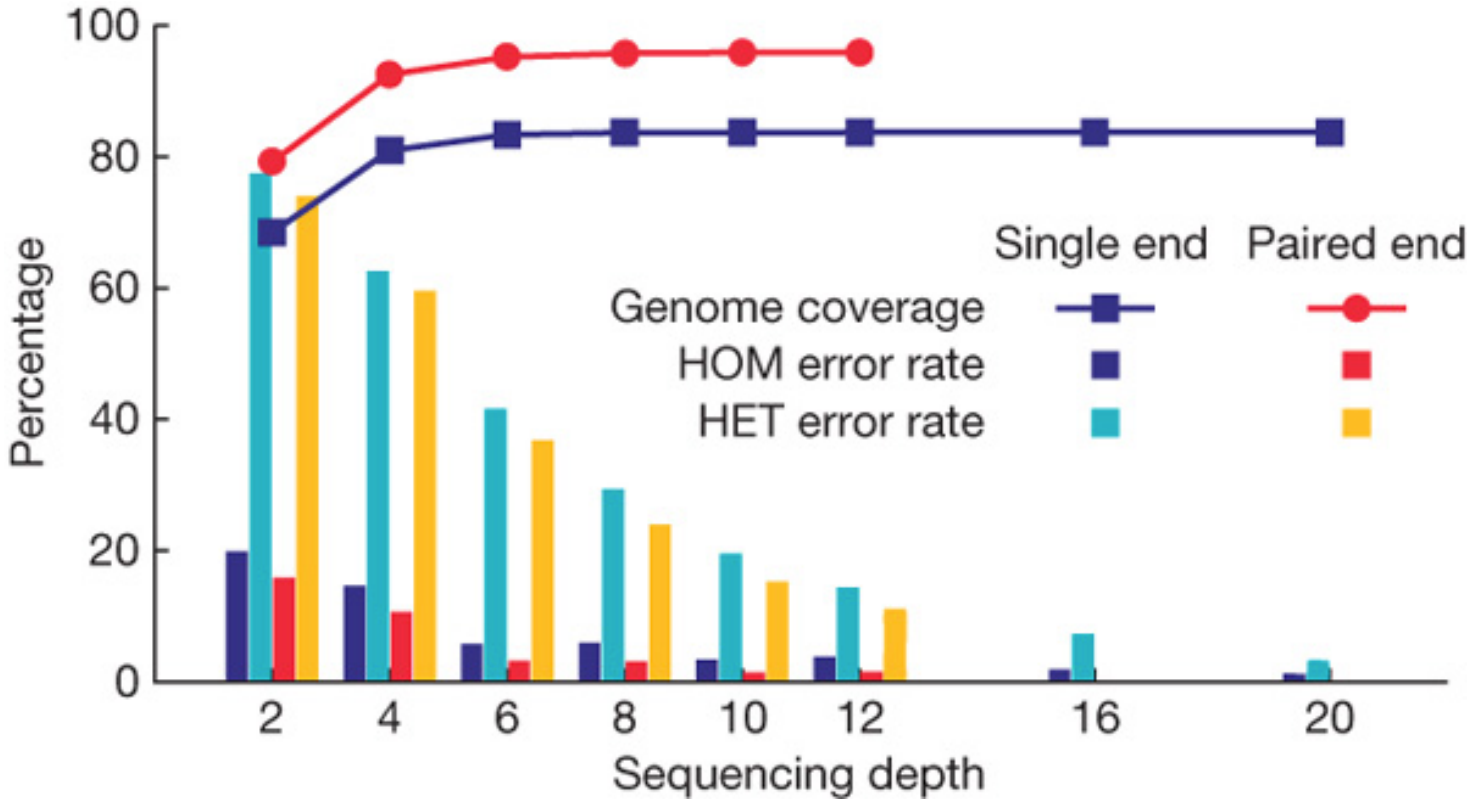


FIGURE 2. Genome coverage of the assembled consensus sequence and the accuracy of SNP detection as a function of sequencing depth.

Population Structure – tracing ancestries

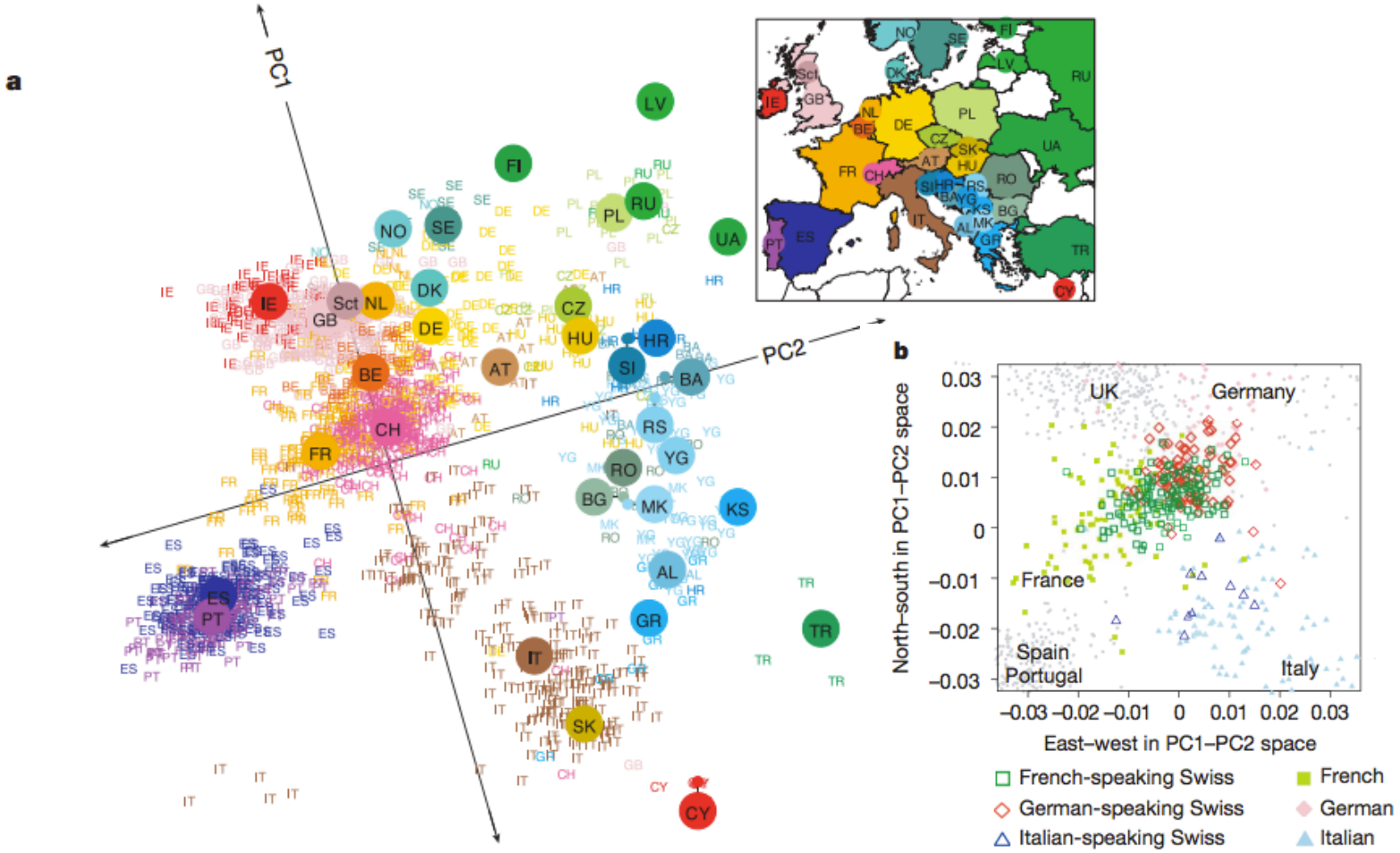
- Sequencing more individuas allow more SNP at different frequencies to be found). This allows high-throughput genotyping technologies with dense geographic samples can shed light on unanswered questions regarding human population structure

- Question: How precisely one can assign an individual to a geographic location based on genetic information alone?

- Initially genotyped 3,192 Europeans at 500,568 loci. After filtering,197,146 loci in 1,387 individuals were used to create 2d visual summary.

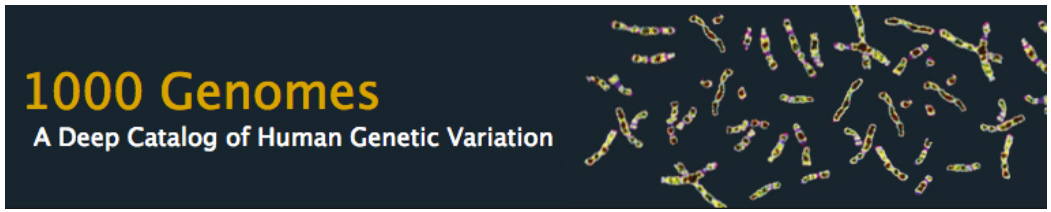
- The results are consistent with the theoretical expectation for model where genetic similarity decays with distance (not discrete well-differentiated populations). 50% of the individuals can be placed within 310 km of their reported origin. 90% within 700 km.

Population Structure – tracing ancestries



Novembre et al., (2008) Nature

Human genome **re**sequencing



Human genome **re**sequencing

Table 2 | Estimated numbers of potentially functional variants in genes

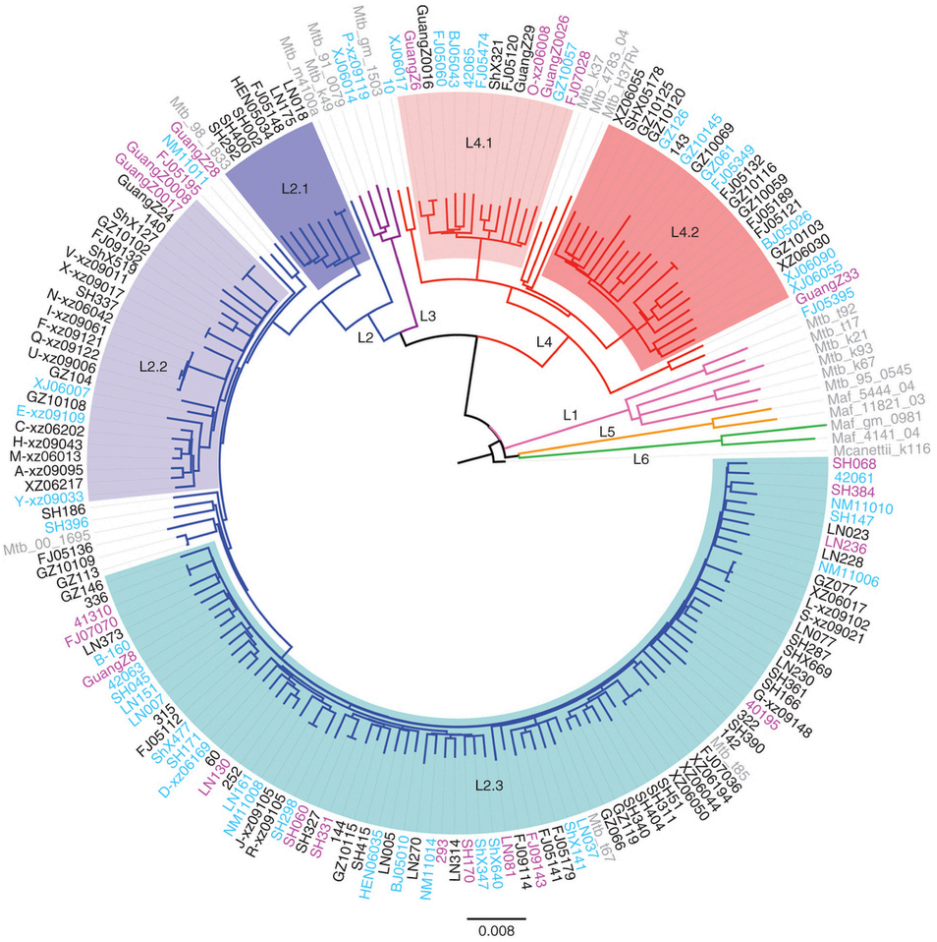
Class	Combined total	Combined novel	Low coverage		High-coverage trio		Exon capture		
			Total	Interquartile*	Total	Individual range	Total	Interquartile*	GENCODE extrapolation
Synonymous SNPs	60,157	23,498	55,217	10,572–12,126	21,410	9,193–12,500	5,708	461–532	11,553–13,333
Non-synonymous SNPs	68,300	34,161	61,284	9,966–10,819	19,824	8,299–10,866	7,063	396–441	9,924–11,052
Small in-frame indels	714	383	666	198–205	289	130–178	59	1–3	~25–75
Stop losses	77	40	71	9–11	22	4–14	6	0–0	~0–0
Stop-introducing SNPs	1,057	755	951	88–101	192	67–100	82	2–3	~50–75
Splice-site-disrupting SNPs	517	399	500	41–49	82	28–45	3	1–1	~50
Small frameshift indels	954	551	890	227–242	433	192–280	37	0–1	~0–25
Genes disrupted by large deletions	147	71	143	28–36	82	33–49	ND	ND	ND
Total genes containing LOF variants	2,304	NA	1,795	272–297	483	240–345	77	3–4	~75–100
HGMD 'damaging mutation' SNPs	671	NA	578	57–80	161	48–82	99	2–4	~50–100

NA, not applicable; ND, not determined.

* Interquartile range of the number of variants of specified type per individual.

doi:10.1038/nature09534

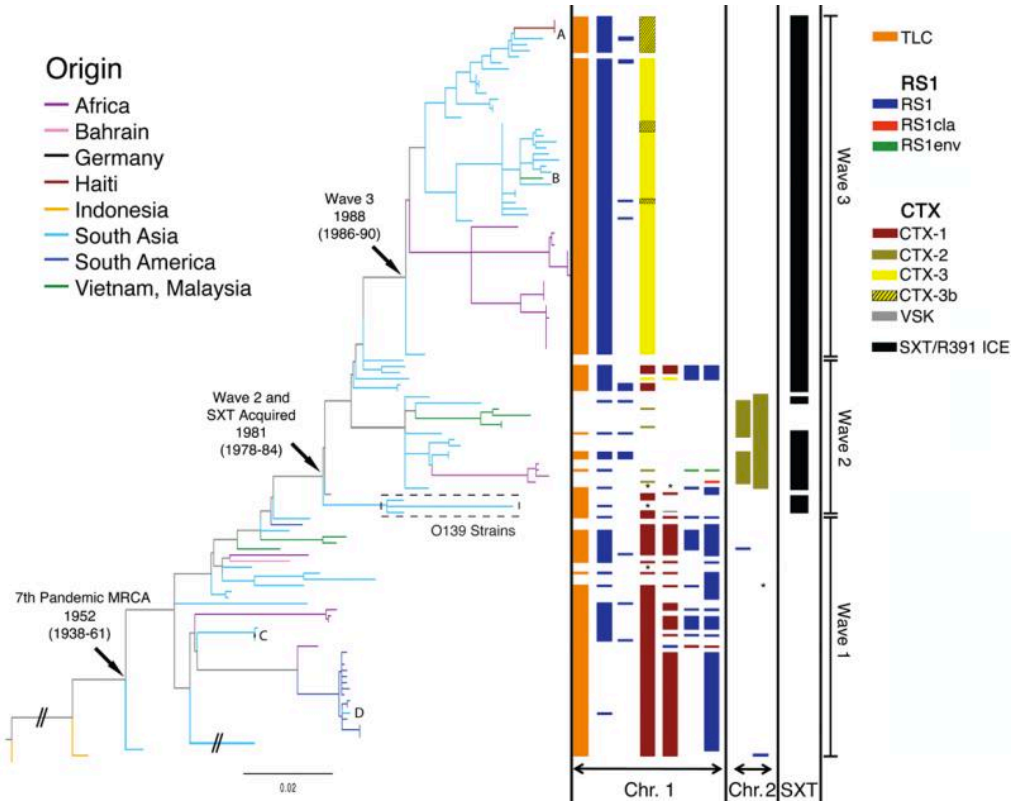
Phylogenetic analysis of *M. tuberculosis* isolates



- Sequencing of 161 multidrug-resistant (MDR) or extensively drug-resistant (XDR) TB isolates
- XDR likely arise as a non-synonymous mutations that arise as a result of second-line drug pressure

Zhang et al (2013)

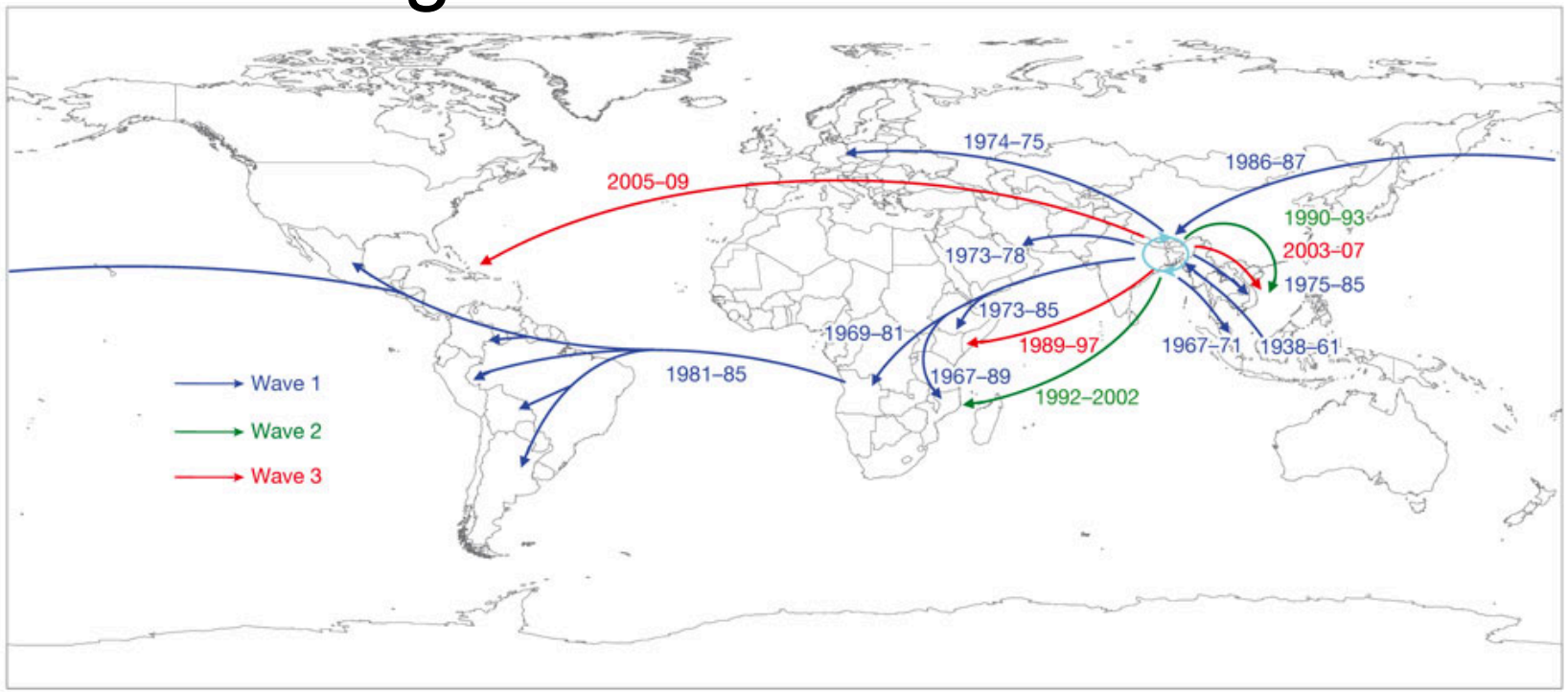
A maximum likelihood phylogenetic tree of the seventh pandemic lineage of *V. cholerae*



- Based on whole-genome analysis of 136 isolates of *V. cholerae*
- Tree provides clear evidence of clonal expansion of the lineage, with a strong temporal signature

Mutreja et al (2011)

Transmission events inferred from the 7th-pandemic tree; common origin



- Based on whole-genome analysis of 136 isolates of *V. cholerae*
- Tree provides clear evidence of clonal expansion of the lineage, with a strong temporal signature

Mutreja et al (2011)

霍亂逾8千死 海地患者告UN

REUTERS 路透社 – 2013年10月9日星期三下午4:04

（路透聯合國紐約總部8日電）代表海地霍亂患者的人權律師今天宣布，已在紐約法院對聯合國（UN）提起訴訟，要求聯合國賠償。

海地患者怪罪聯合國維和人員帶來霍亂，並要求聯合國支付數億美元賠償金。聯合國今年稍早表示，不會支付賠償金後，患者決定採取訴訟。

2010年10月以來，海地有超過8300人死於霍亂疫情，感染人數超過65萬。

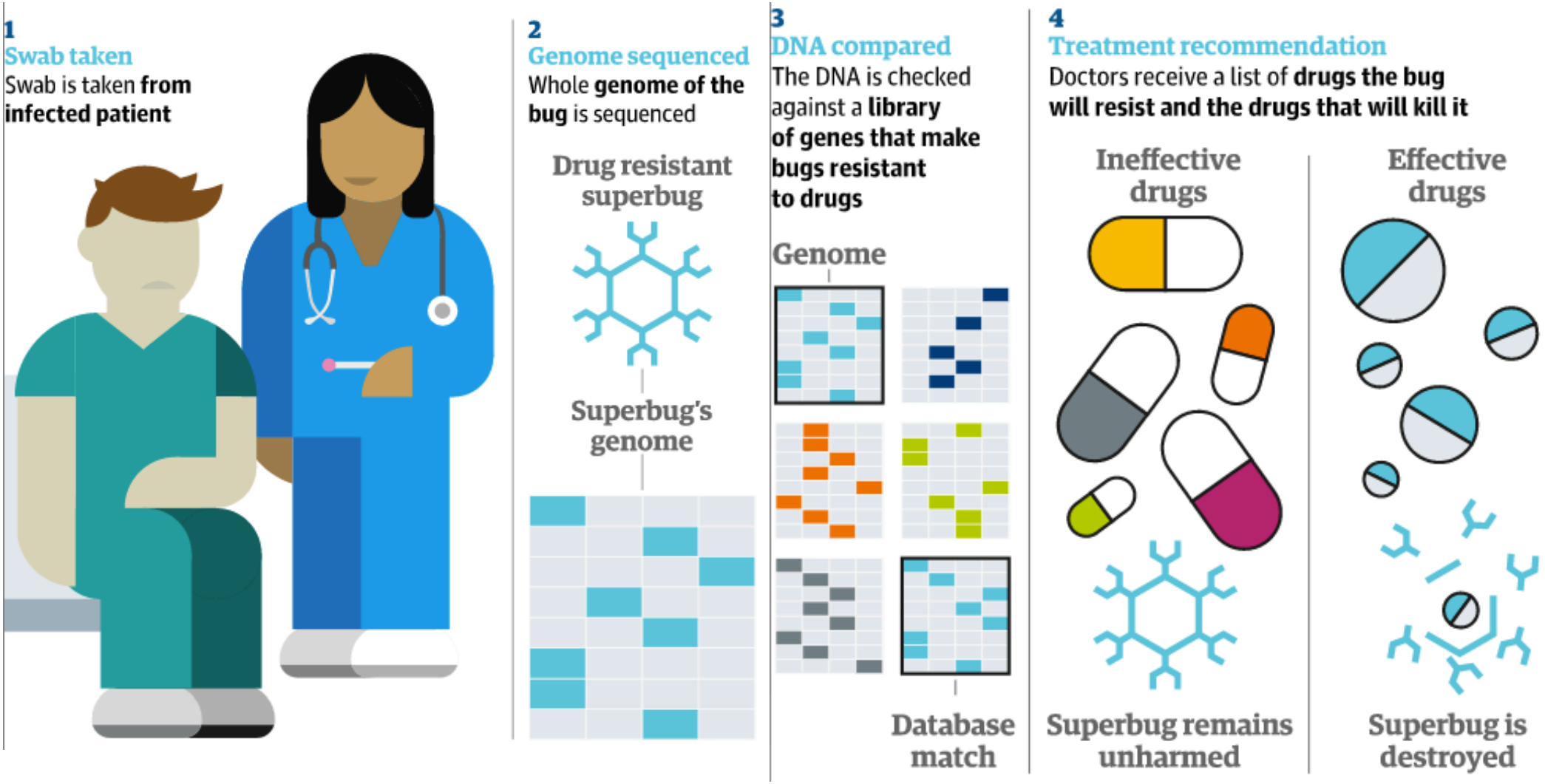
海地公理暨民主研究所（Institute for Justiceand Democracy）發布聲明表示：「提告者包括感染霍亂的海地民眾或海地裔美國人，及死亡病患家屬。」

根據聲明，律師在聯邦紐約南區地方法院（Southern District）提起訴訟，並沒有提到索賠金額的細節。

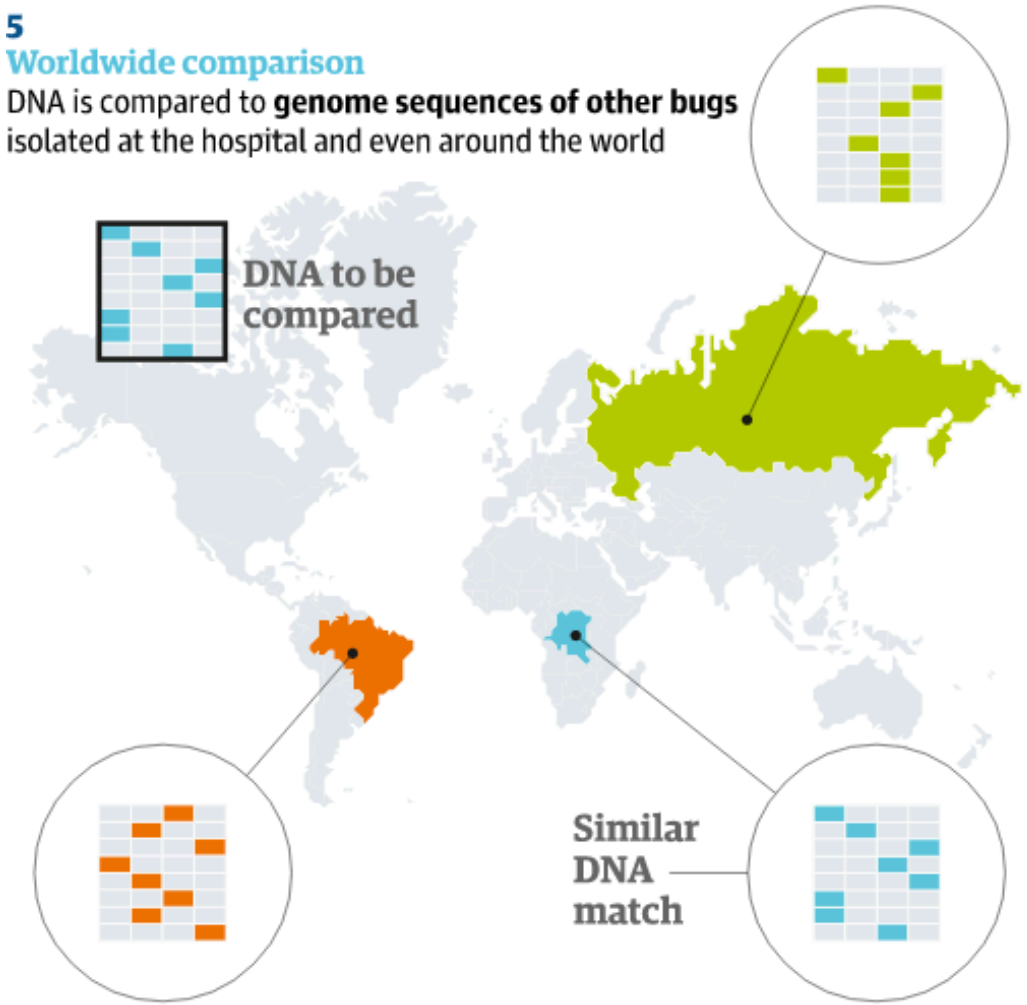
聯合國秘書長潘基文（Ban Ki-moon）指派的獨立專家小組2011年曾發布1份霍亂疫情報告，報告並未判定霍亂是如何傳到海地。

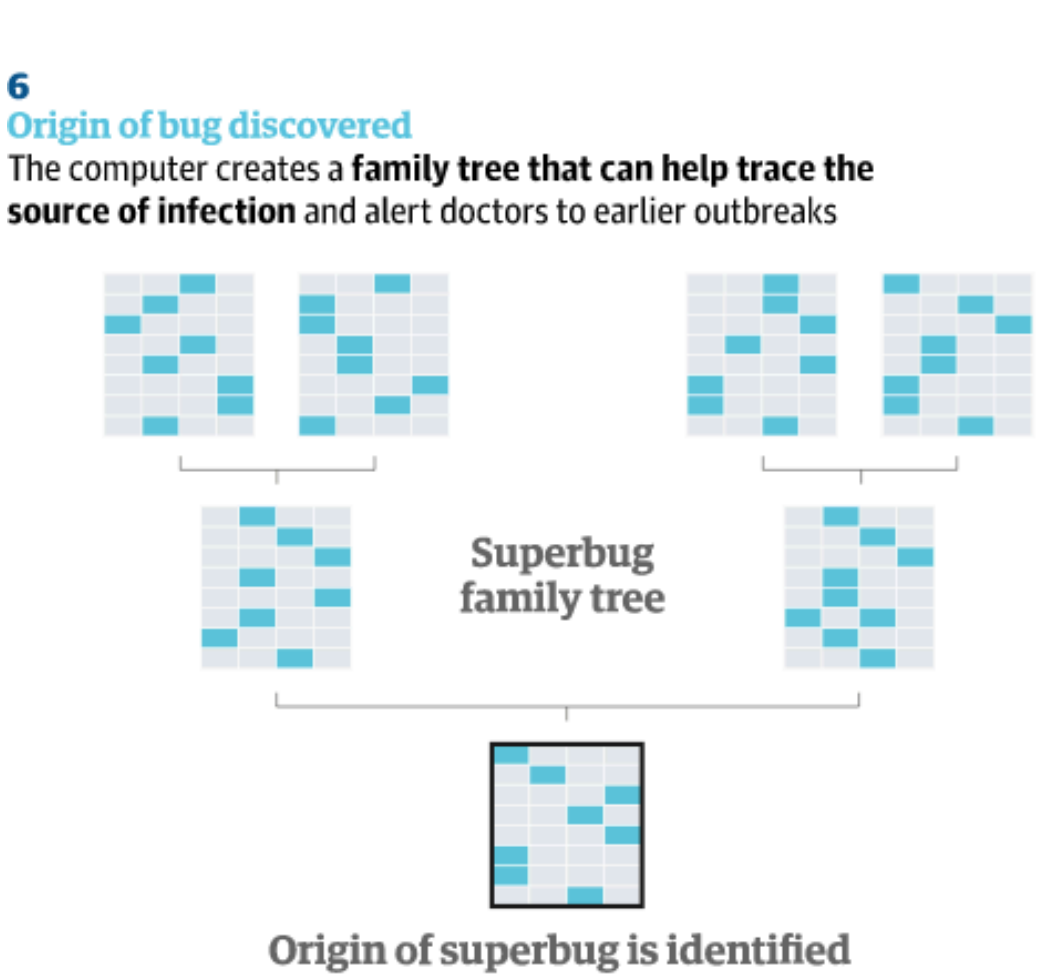
但美國疾病管制暨預防中心（U.S. Centers forDisease Control and Prevention）發現，證據強烈顯示，霍亂病源是來自尼泊爾的聯合國維和部隊。中央社（翻譯）

Current workflow of clinical labs

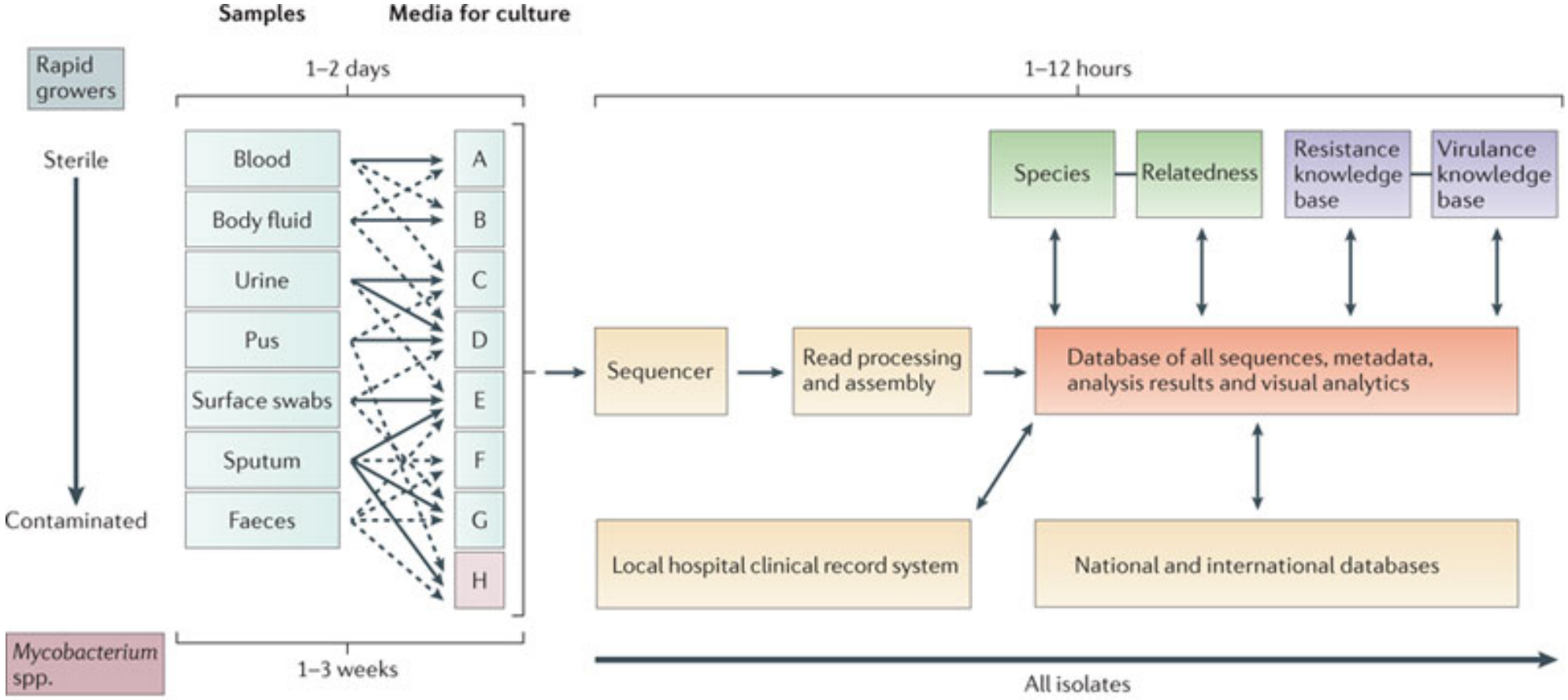


Current workflow of clinical labs





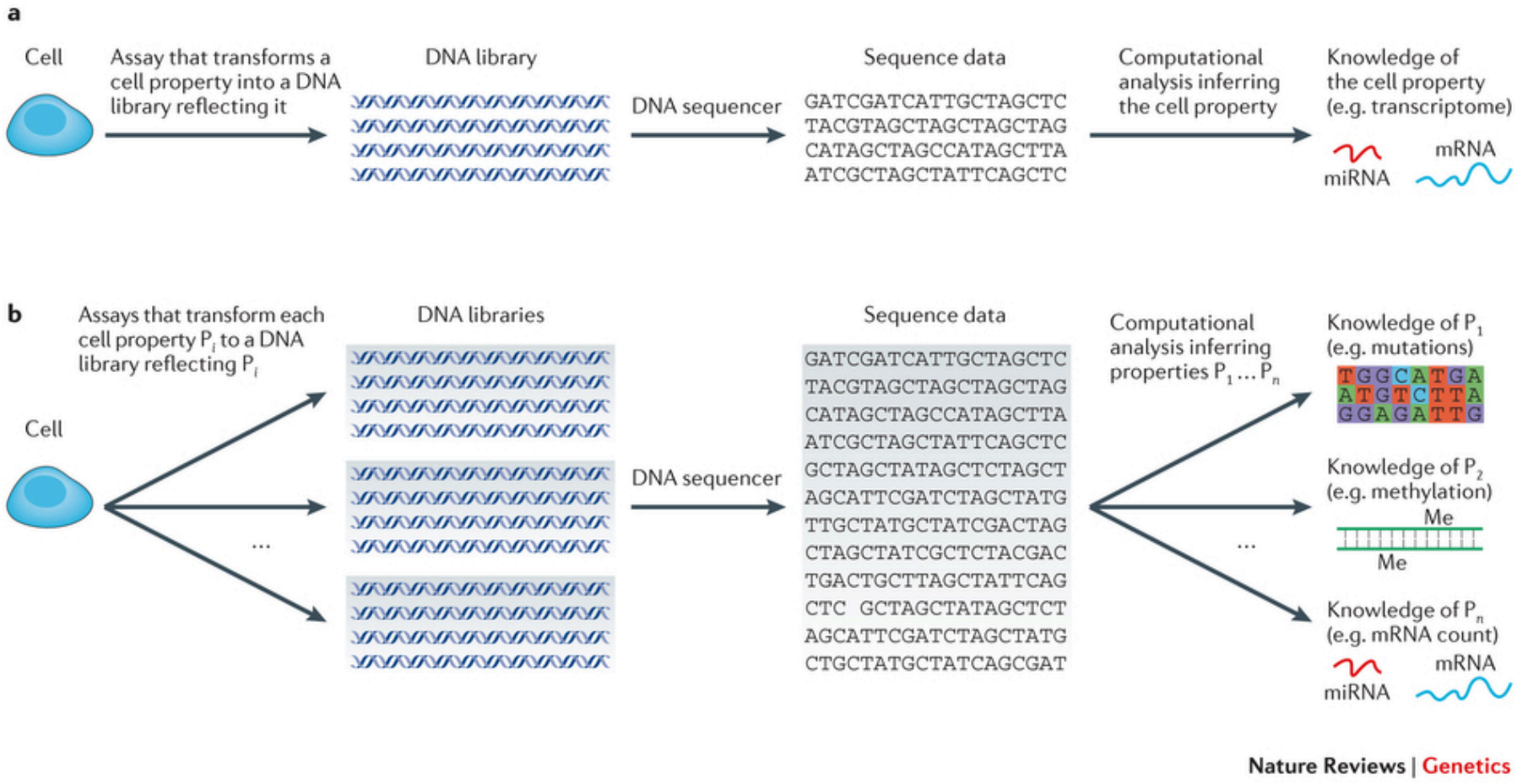
Future? – workflow of clinical labs / ecological samples



Nature Reviews | Genetics

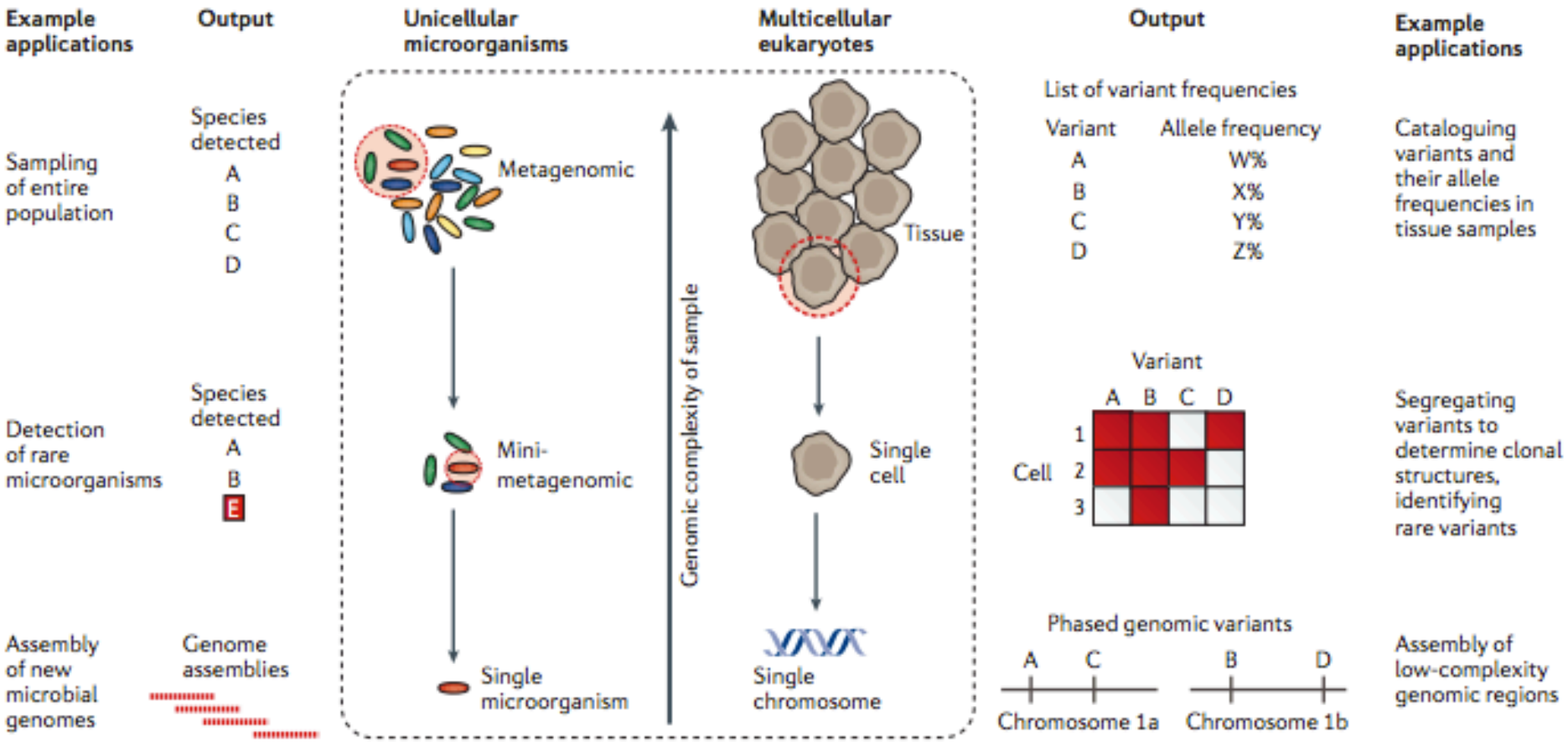
Didelot et al 2012

Future – single cell genomics will revolutionize whole-organism science



doi:10.1038/nrg3542

Current (2016) – single cell genomics has revolutionized whole-organism science

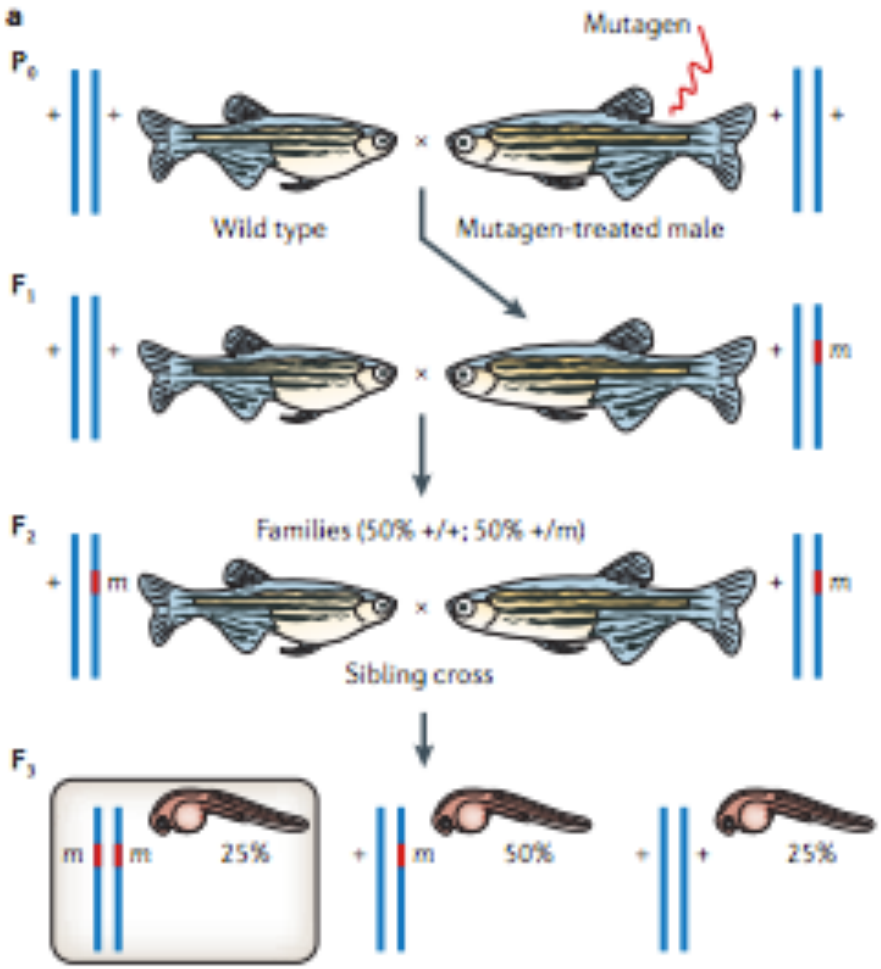


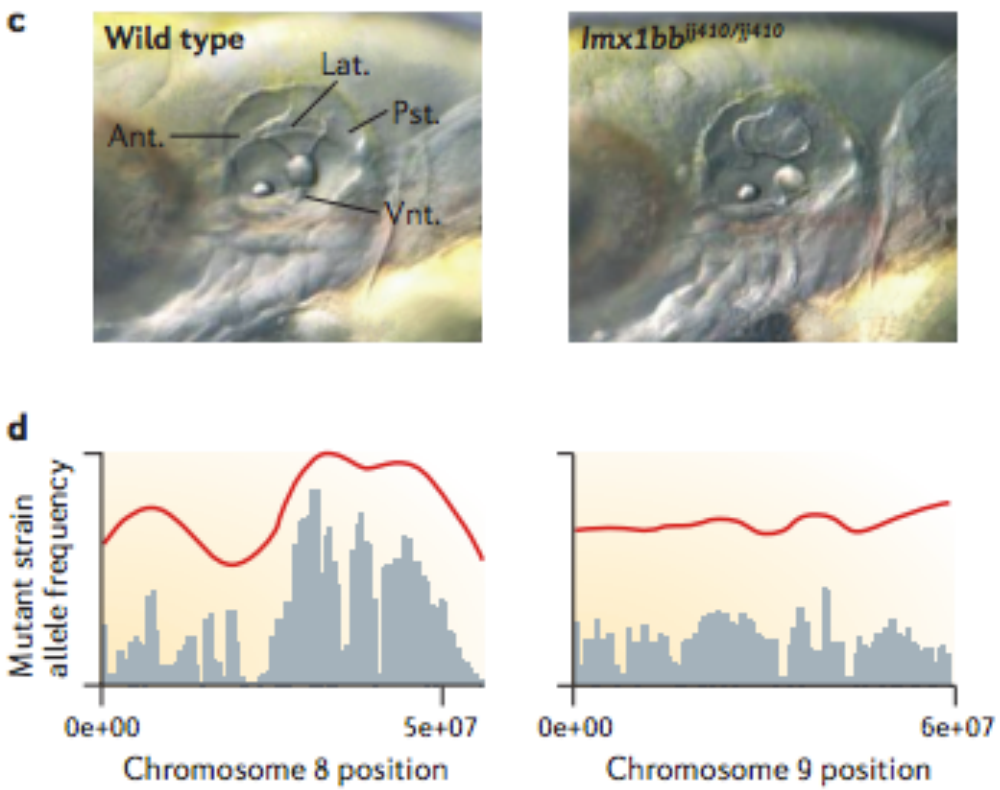
Gawad *et al* 2016

Using next-generation sequencing to isolate mutant genes from forward genetic screens

Korbinian Schneeberger

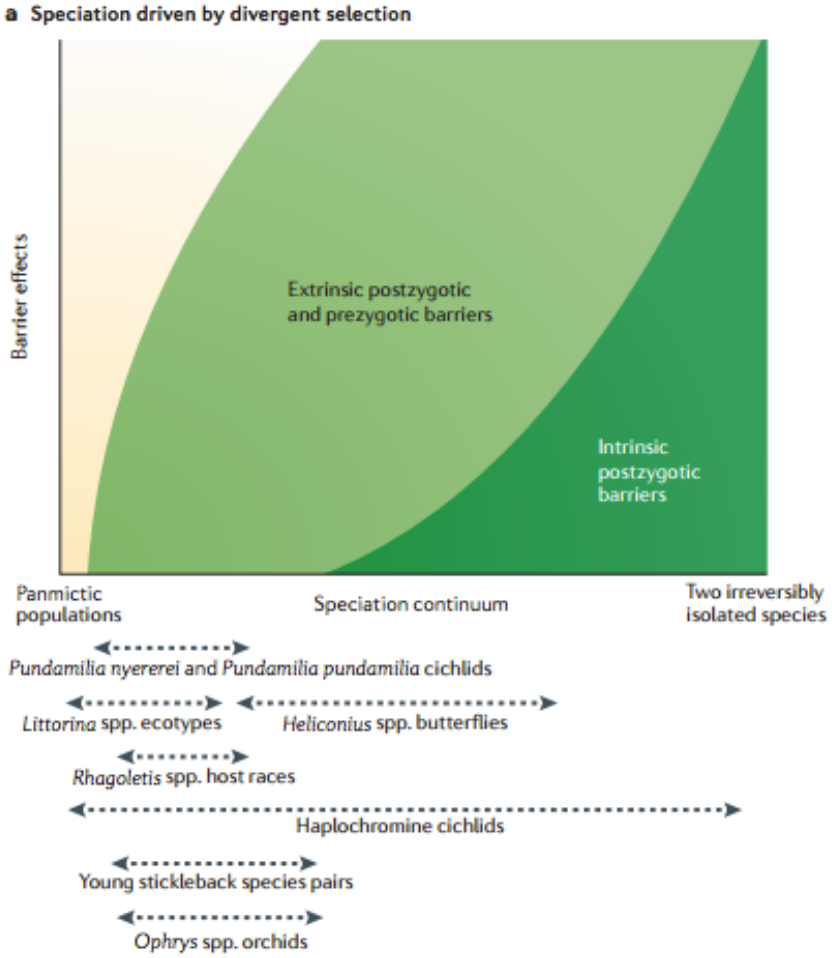
Abstract | The long-lasting success of forward genetic screens relies on the simple molecular basis of the characterized phenotypes, which are typically caused by mutations in single genes. Mapping the location of causal mutations using genetic crosses has traditionally been a complex, multistep procedure, but next-generation sequencing now allows the rapid identification of causal mutations at single-nucleotide resolution even in complex genetic backgrounds. Recent advances of this mapping-by-sequencing approach include methods that are independent of reference genome sequences, genetic crosses and any kind of linkage information, which make forward genetics amenable for species that have not been considered for forward genetic screens so far.

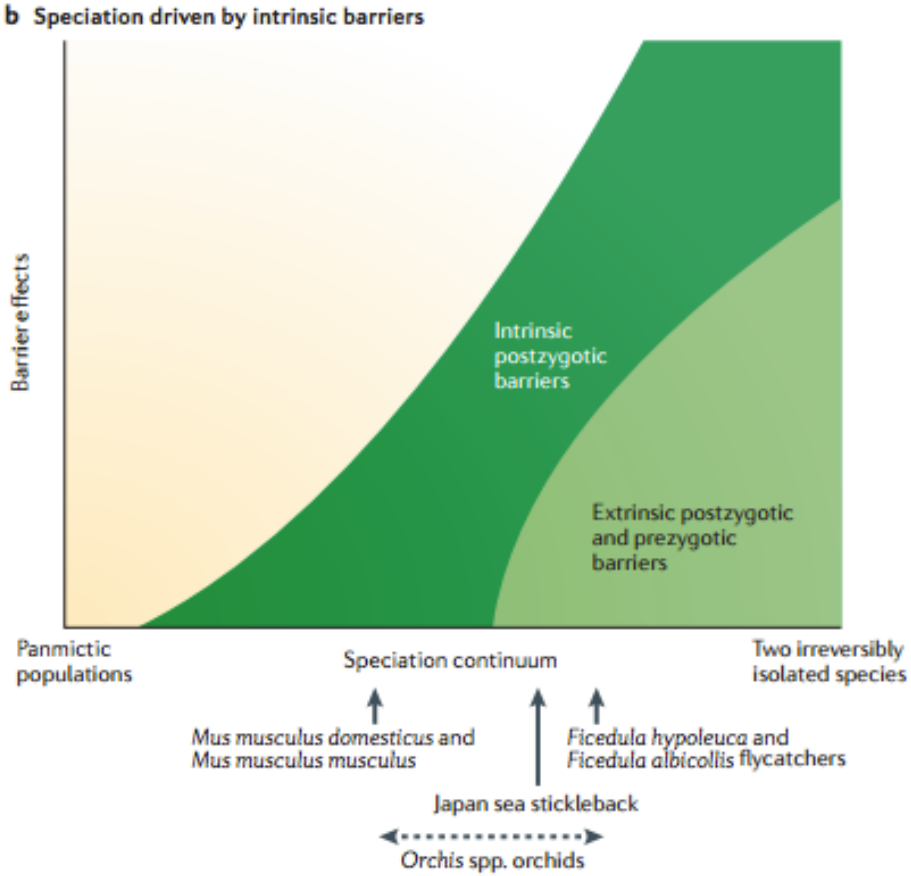




Schneeberger 2014

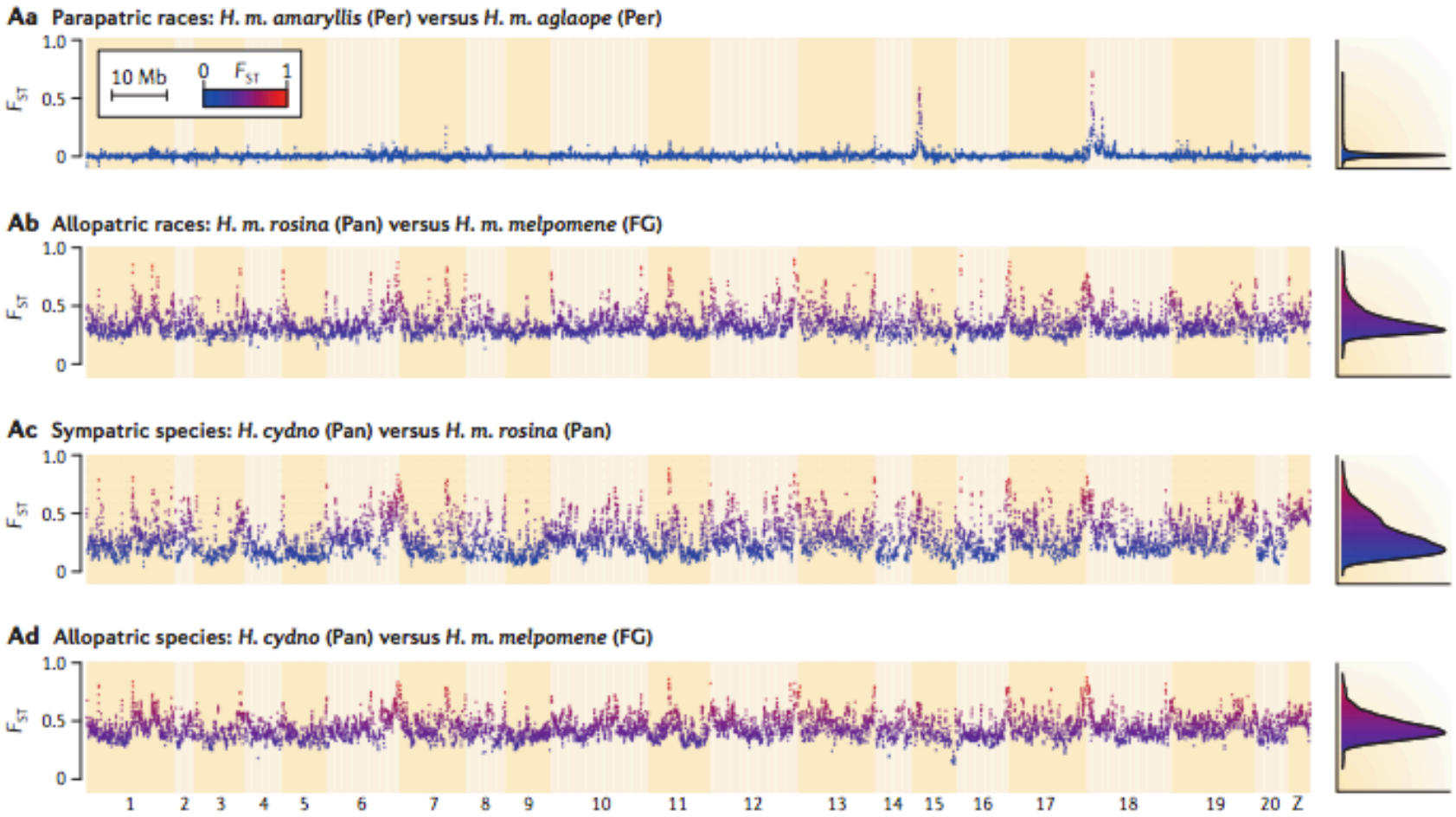
Need to consider different evolutionary scenarios





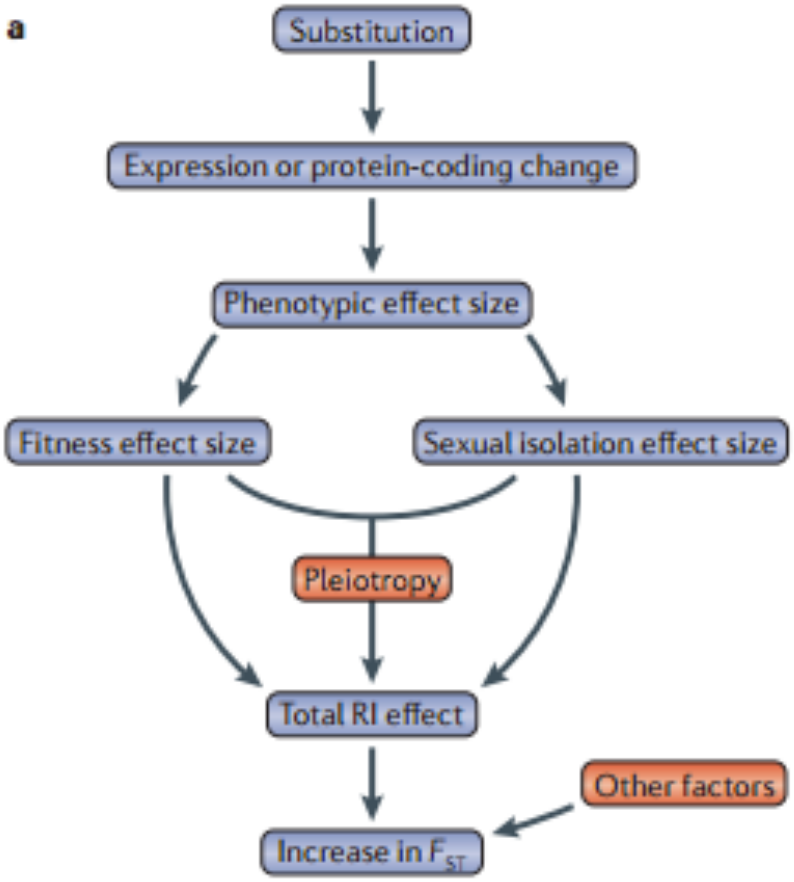
Seehausen et al 2016

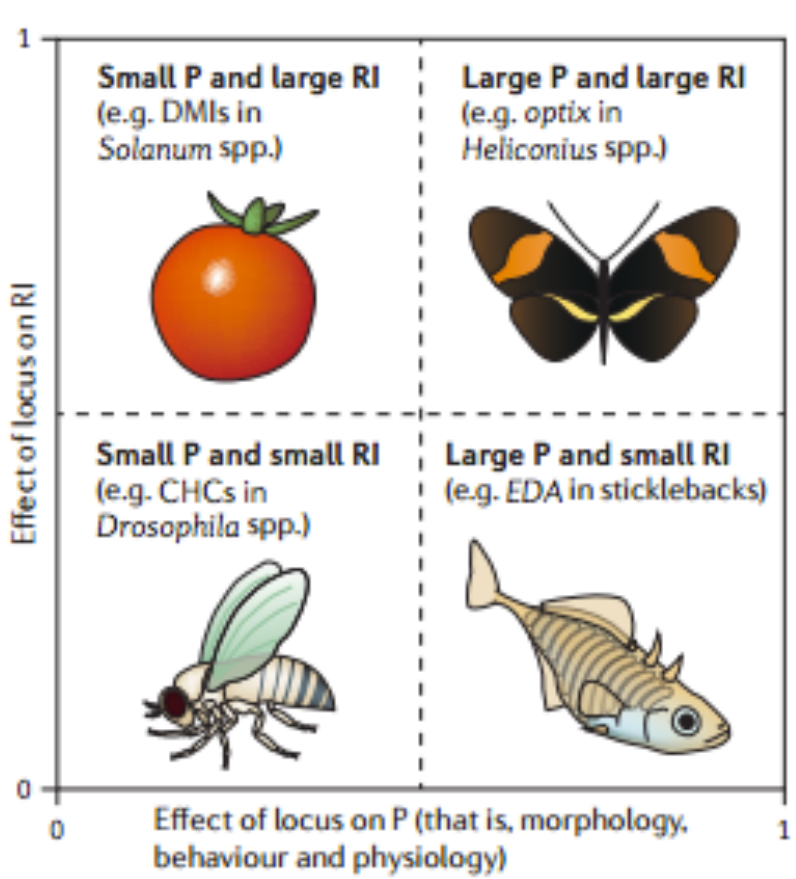
Hence different patterns of signals in genome scans



Seehausen et al 2016

Effect of P and RI (later lecture)





Seehausen et al 2016

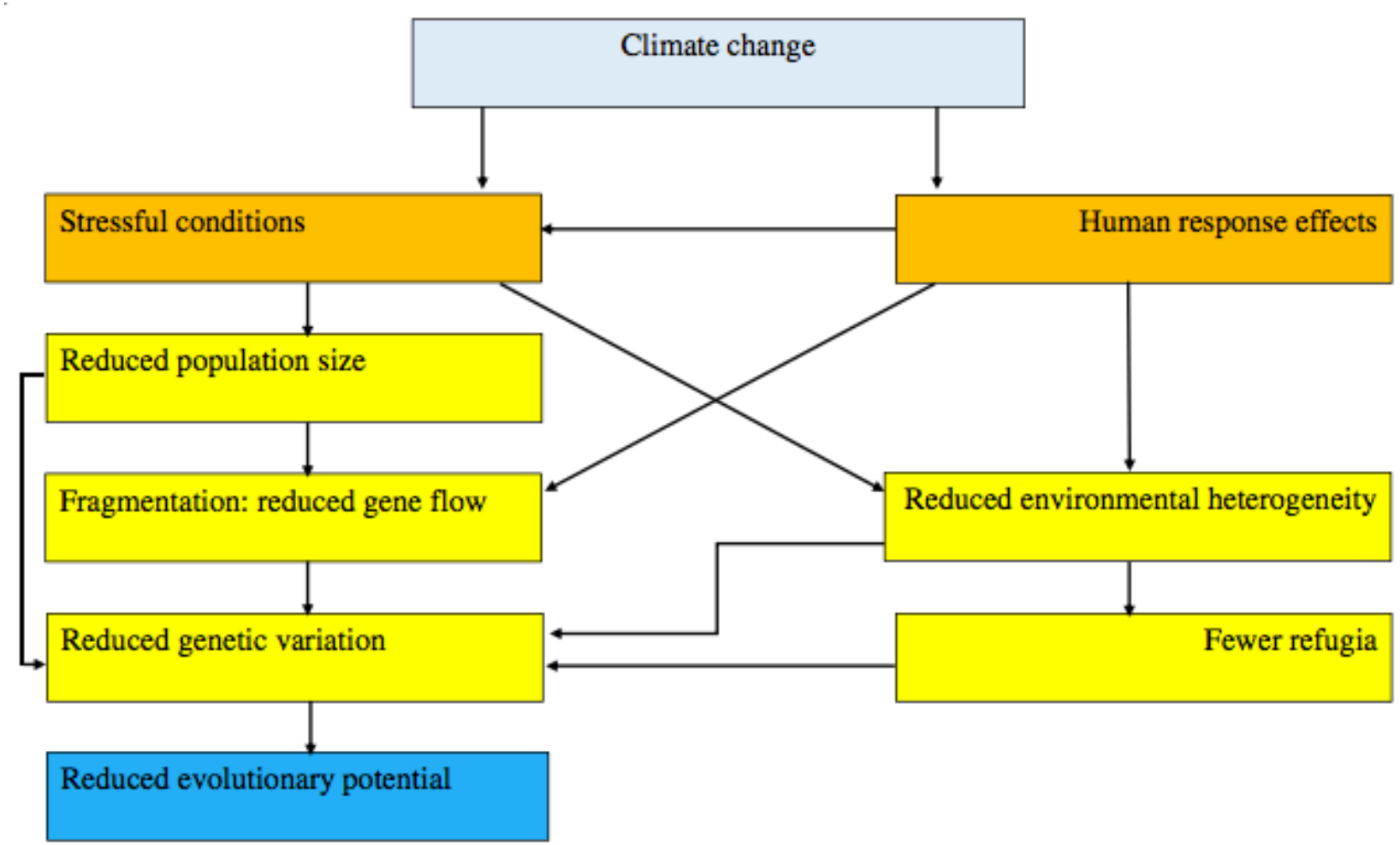
Hoffmann *et al. Climate Change Responses* (2015) 2:1
DOI 10.1186/s40665-014-0009-x

Climate Change Responses

REVIEW Open Access

A framework for incorporating evolutionary genomics into biodiversity conservation and management

Ary Hoffmann[1*], Philippa Griffin[1], Shannon Dillon[2], Renee Catullo[3], Rahul Rane[1], Margaret Byrne[4], Rebecca Jordan[1], John Oakeshott[5], Andrew Weeks[1], Leo Joseph[6], Peter Lockhart[7], Justin Borevitz[3] and Carla Sgrò[8]


Climate change
Stressful conditions
Human response effects
Reduced population size
Fragmentation: reduced gene flow
Reduced environmental heterogeneity
Reduced genetic variation
Fewer refugia
Reduced evolutionary potential

Ultimately, mapping is to quickly identify relationship between individuals once reference is known

Tradeoffs between $$$, sample size, sensitivity, speed

What kind of relationship do you want to see?

- To track disease transmission / species migration
  - Phylogenetics
  - May need to undertake assembly approach as well
- Selection on a particular locus between two populations?
  - Genomic scans
  - May need to undertake assembly approach as well
- Or, **everything!!!** (probably everyone's answer)

Reading materials

http://www.nature.com/nrg/series/nextgeneration/index.html

Surprising a good introduction
http://www.illumina.com/documents/products/technotes/technote_nextera_matepair_data_processing.pdf

Canadian Bioinformatics Workshops (all slides and video are available)
http://bioinformatics.ca/past-workshops