Current new sequencing platforms

Peter Campbell,
Cancer Genome Project,
Sanger Institute.

454 GS FLX
Life Sciences (Roche)

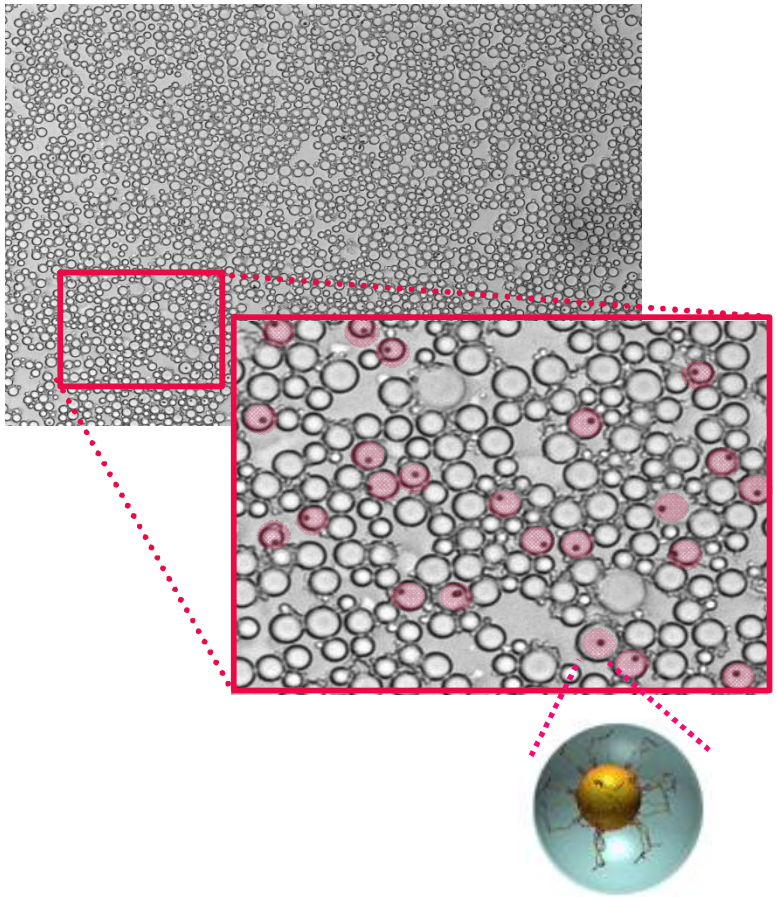
Genetic Analyzer
Solexa (Illumina)

SOLID sequencing
(Applied Biosystems)

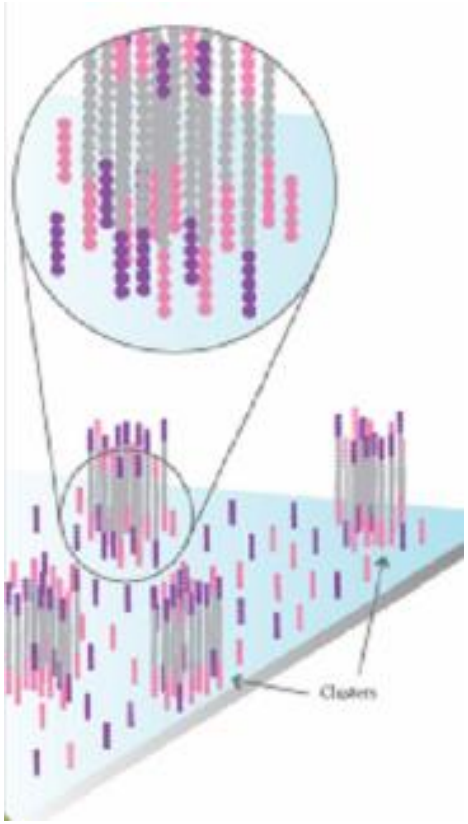
Common themes

- In situ amplification of single DNA molecule

Emulsion PCR (454 & SOLiD)

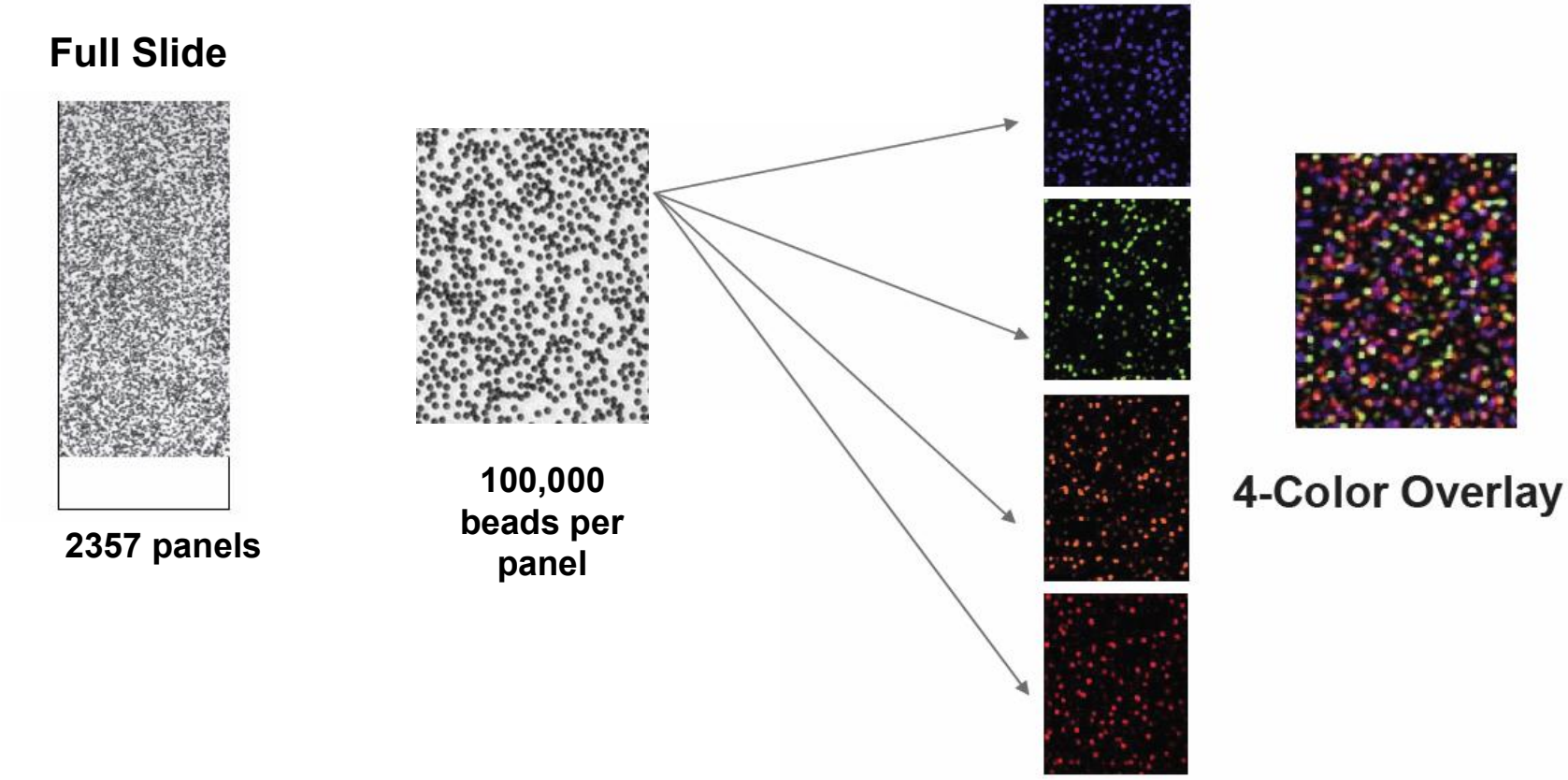


Bridge amplification (Solexa)



Common themes

- Massively parallel sequencing



Common themes

- Systematic and random sequencing errors
  - Phasing
  - Background signal
  - Low signal:noise ratio
  - PCR errors
  - Chemistry-specific issues

Common themes

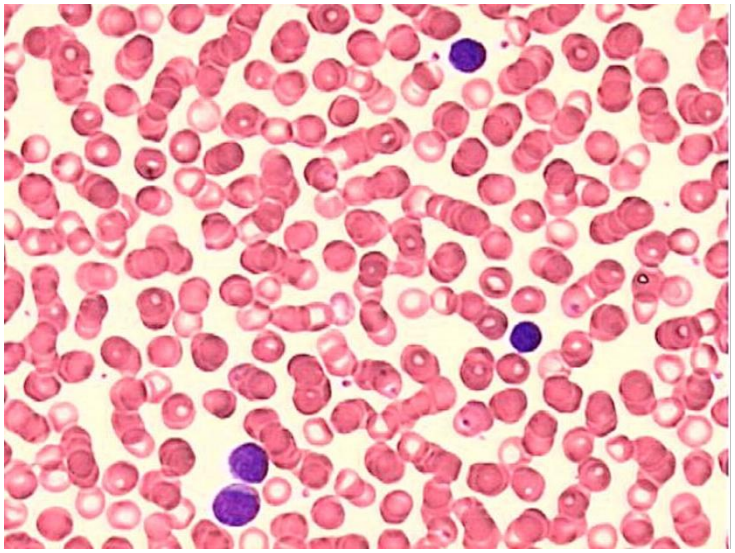
- Computational challenges
  - Short reads (Solexa & SOLiD)
  - Huge primary data image files (Tb)
  - On-machine image processing $\rightarrow$ FASTQ format
  - Currently, limited downstream analysis packages

454 GS FLX platform

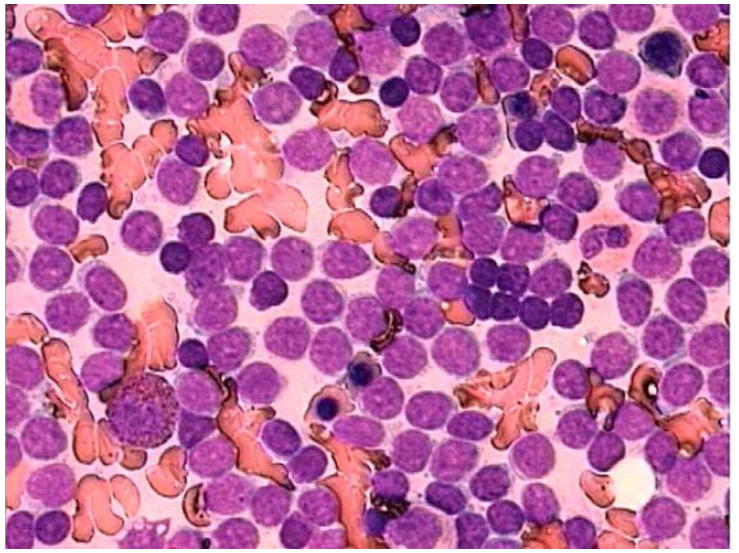
- Pyrosequencing chemistry
- Read lengths ~250bp
- ~ 400,000 sequences / run
- ~100 – 150Mb / run

Ultra-deep sequencing on the 454

Chronic lymphocytic leukaemia

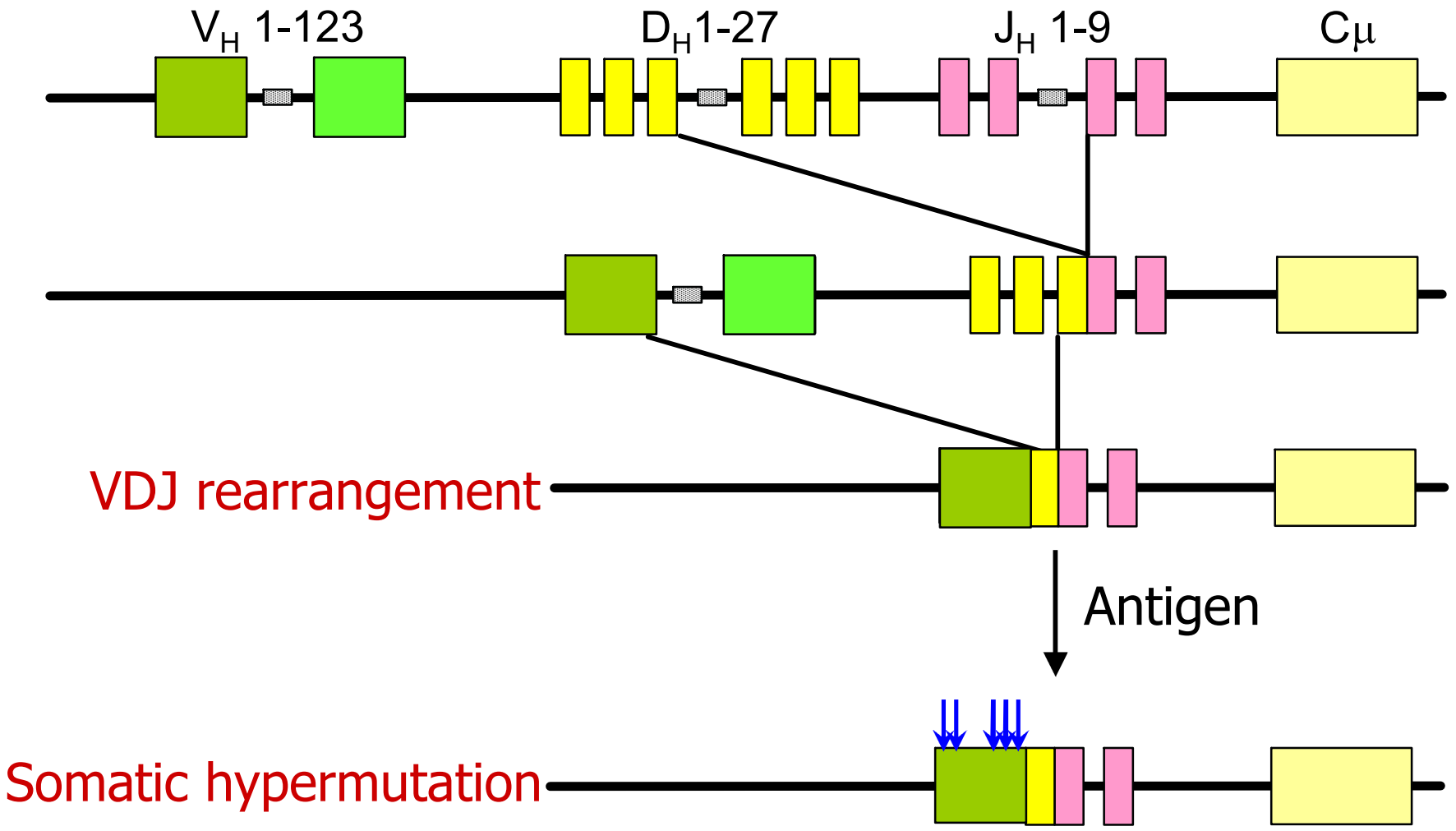


Blood

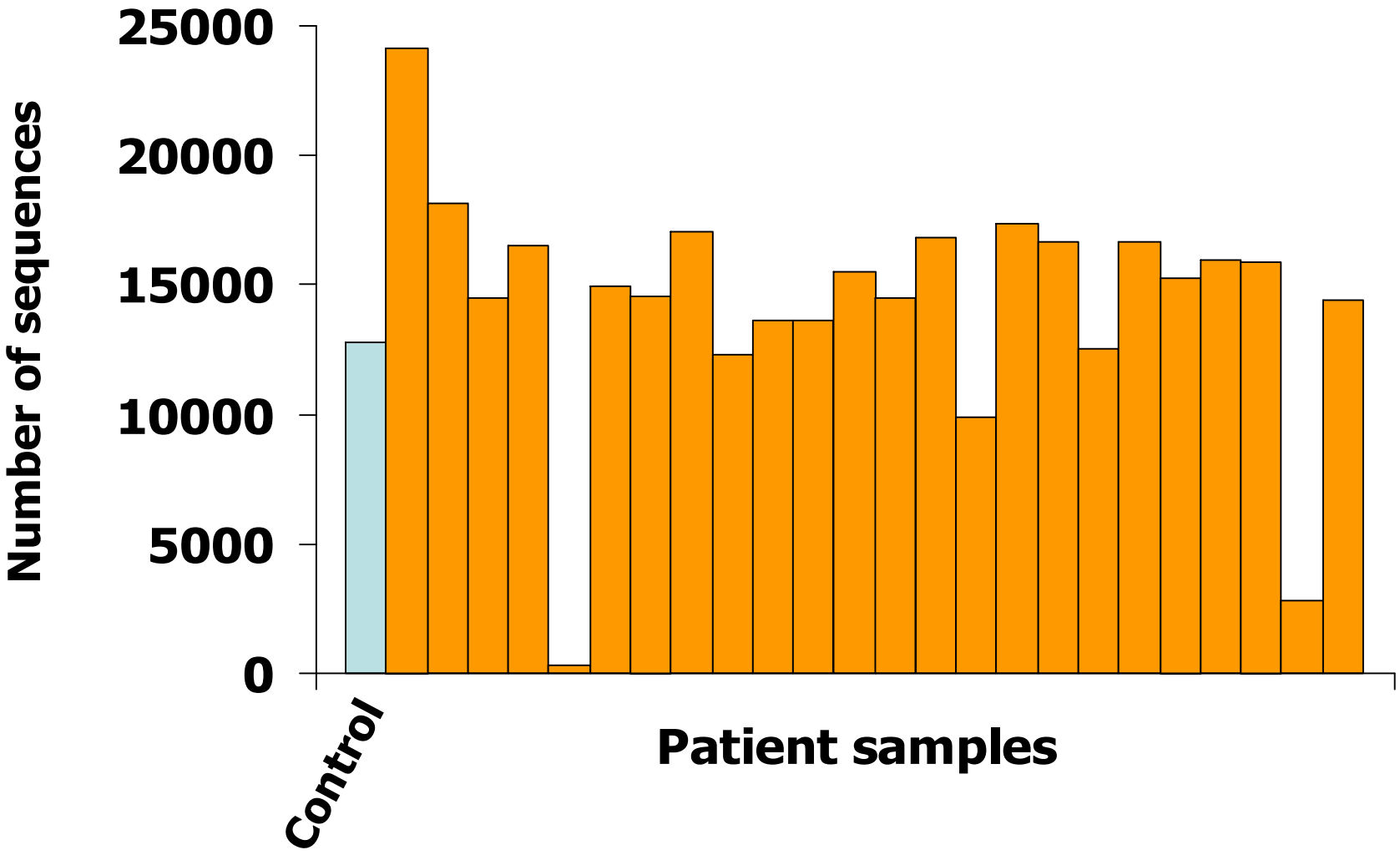


Bone marrow

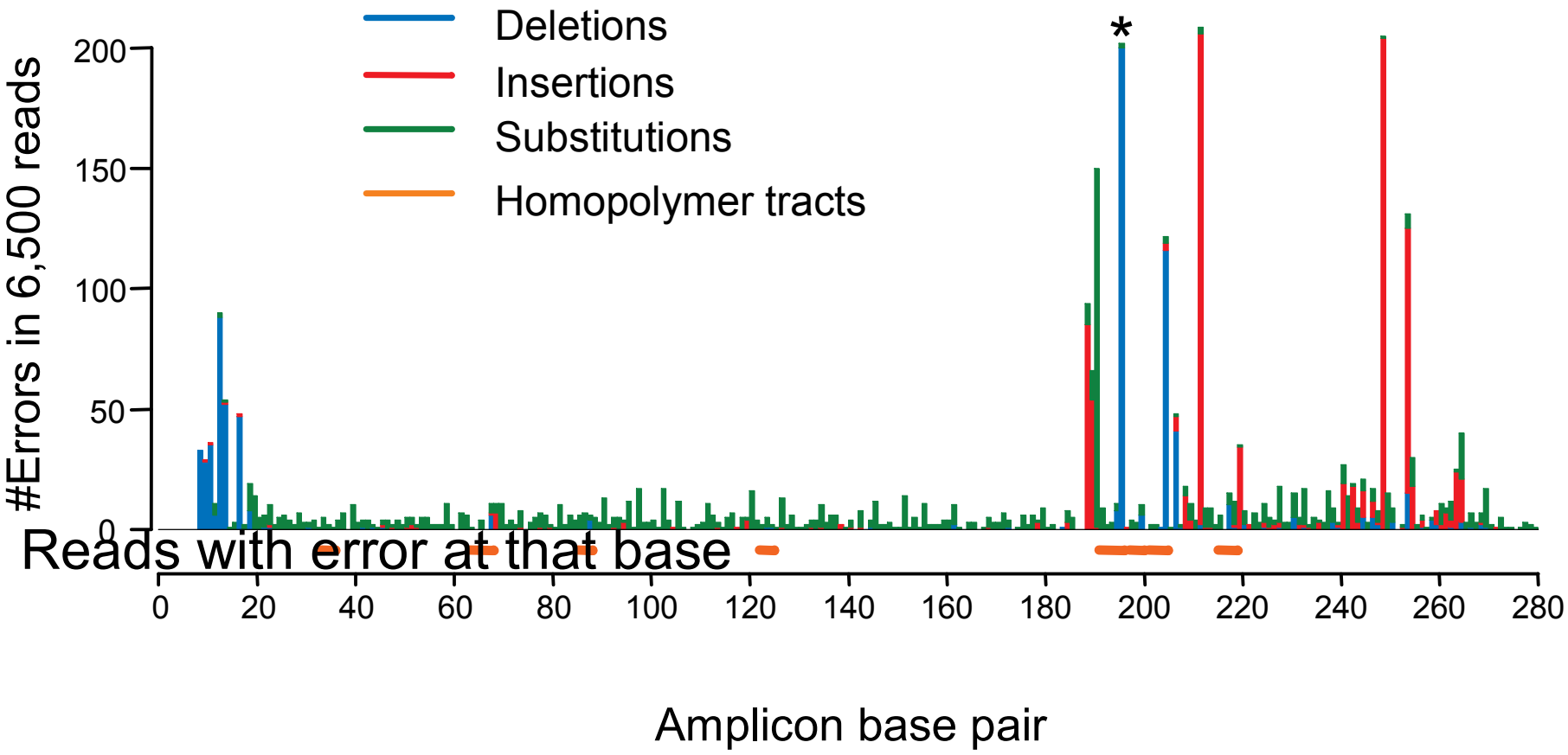
IGH rearrangement & somatic hypermutation



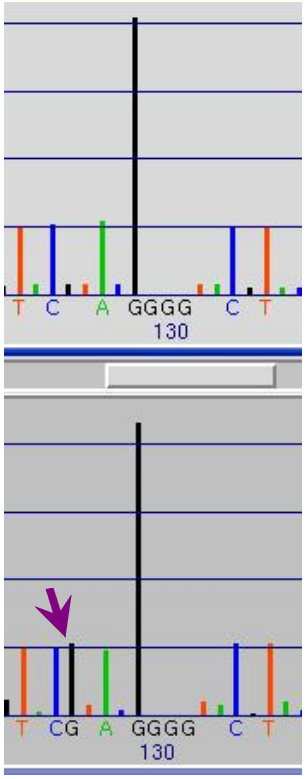
Distribution of sequences

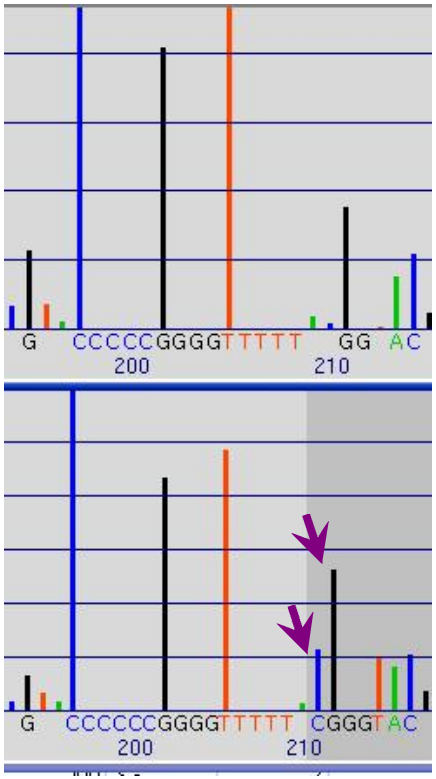


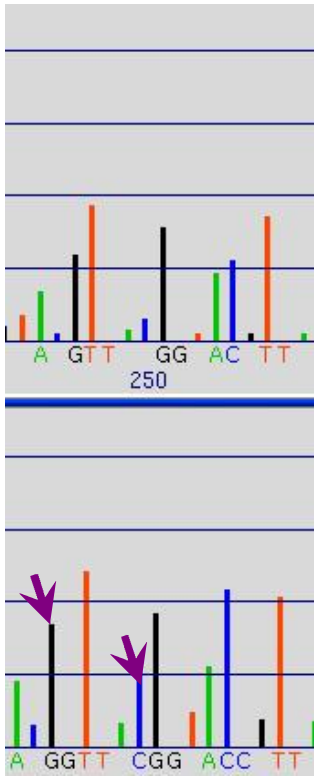
Sequencing errors – control locus



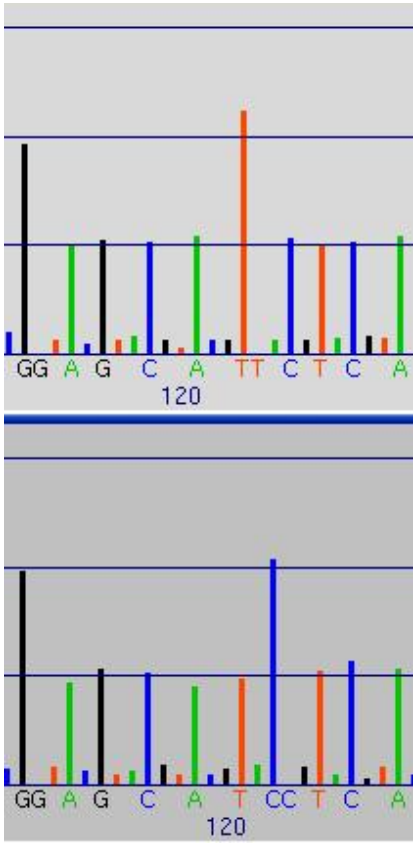
Examples of insertions





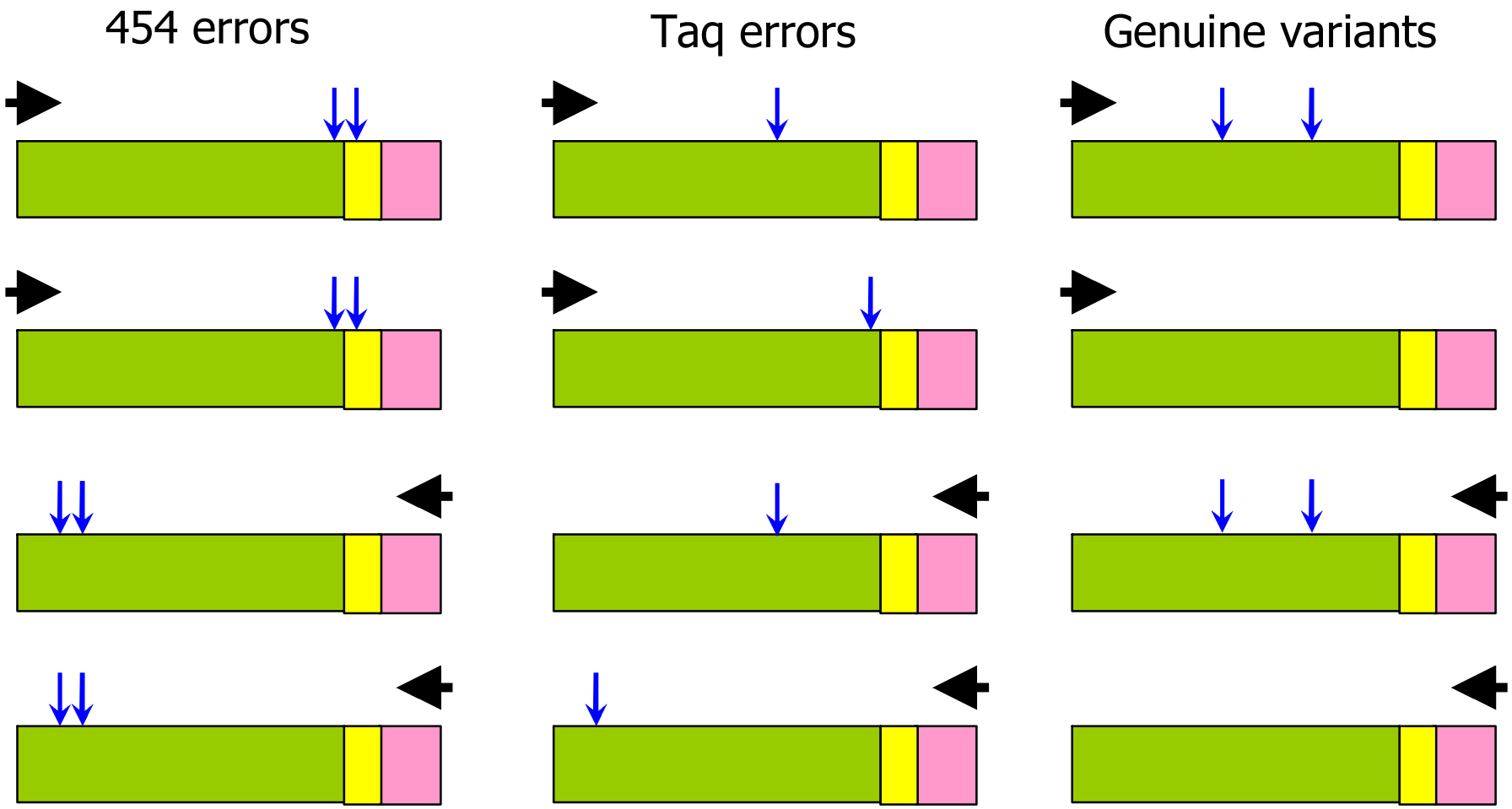


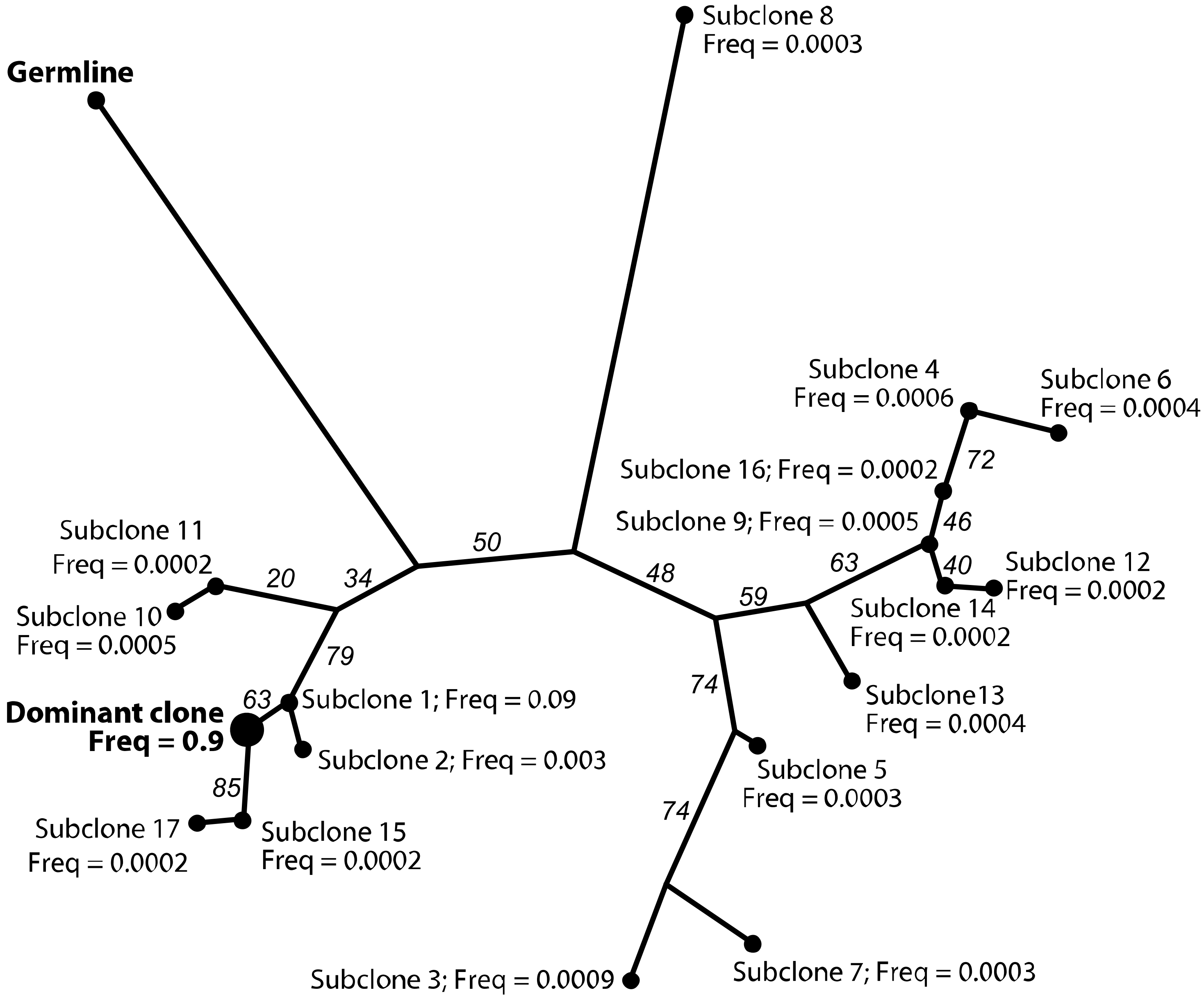
Substitutions



Post / Pre	A	C	G	T
A		22	132	15
C	5		53	161
G	118	9		34
T	19	300	29	

454 & Taq errors vs genuine variants

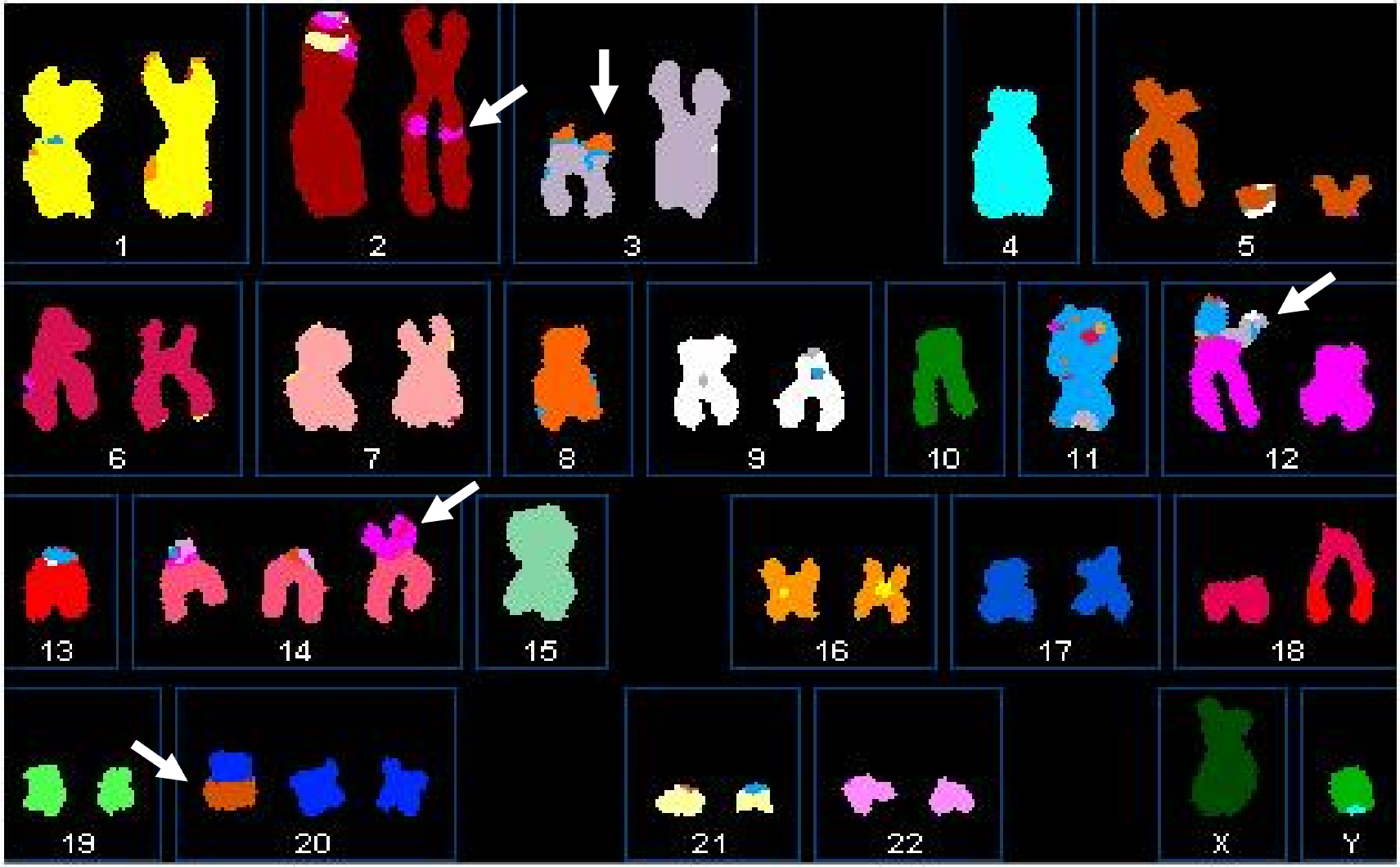



Subclone 8
Freq = 0.0003
Germline
Subclone 4
Freq = 0.0006
Subclone 6
Freq = 0.0004
72
Subclone 16; Freq = 0.0002
46
Subclone 9; Freq = 0.0005
Subclone 11
Freq = 0.0002
50
63
40
Subclone 12
Freq = 0.0002
20
34
48
59
Subclone 14
Freq = 0.0002
Subclone 10
Freq = 0.0005
79
74
Subclone13
Freq = 0.0004
63
Subclone 1; Freq = 0.09
Dominant clone
Freq = 0.9
Subclone 2; Freq = 0.003
Subclone 5
Freq = 0.0003
85
74
Subclone 17
Freq = 0.0002
Subclone 15
Freq = 0.0002
Subclone 3; Freq = 0.0009
Subclone 7; Freq = 0.0003

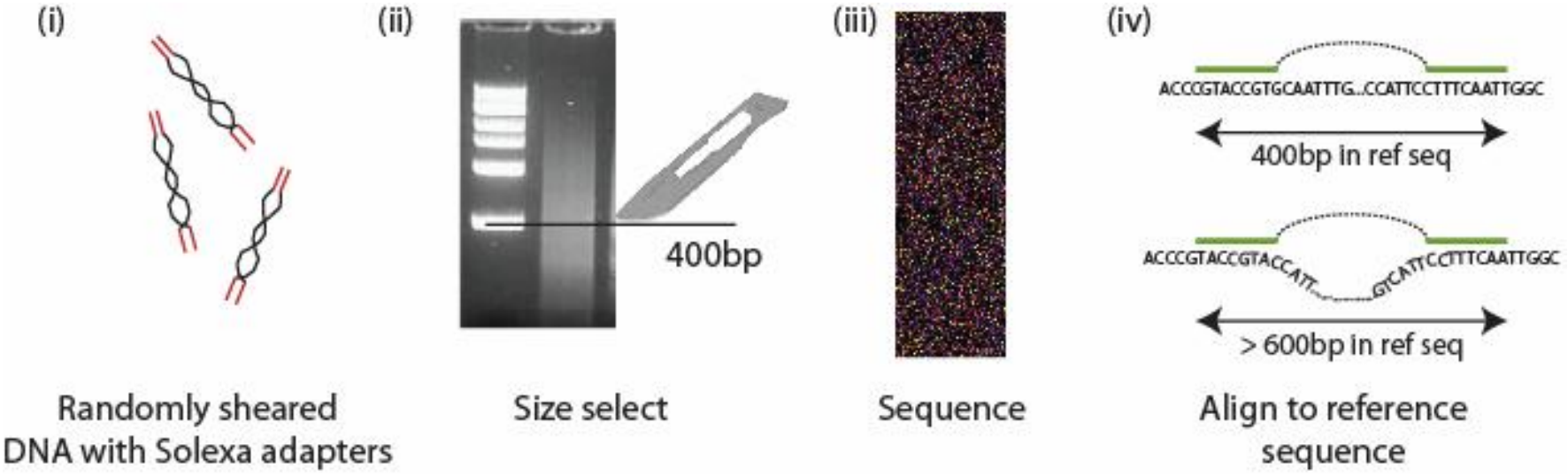
Solexa platform

- Chemistry: Uses reversible extension terminators
- ~30-50 million reads (paired end capability)
- ~35bp per read (now up to 70bp)
- 2-4 Gb per run

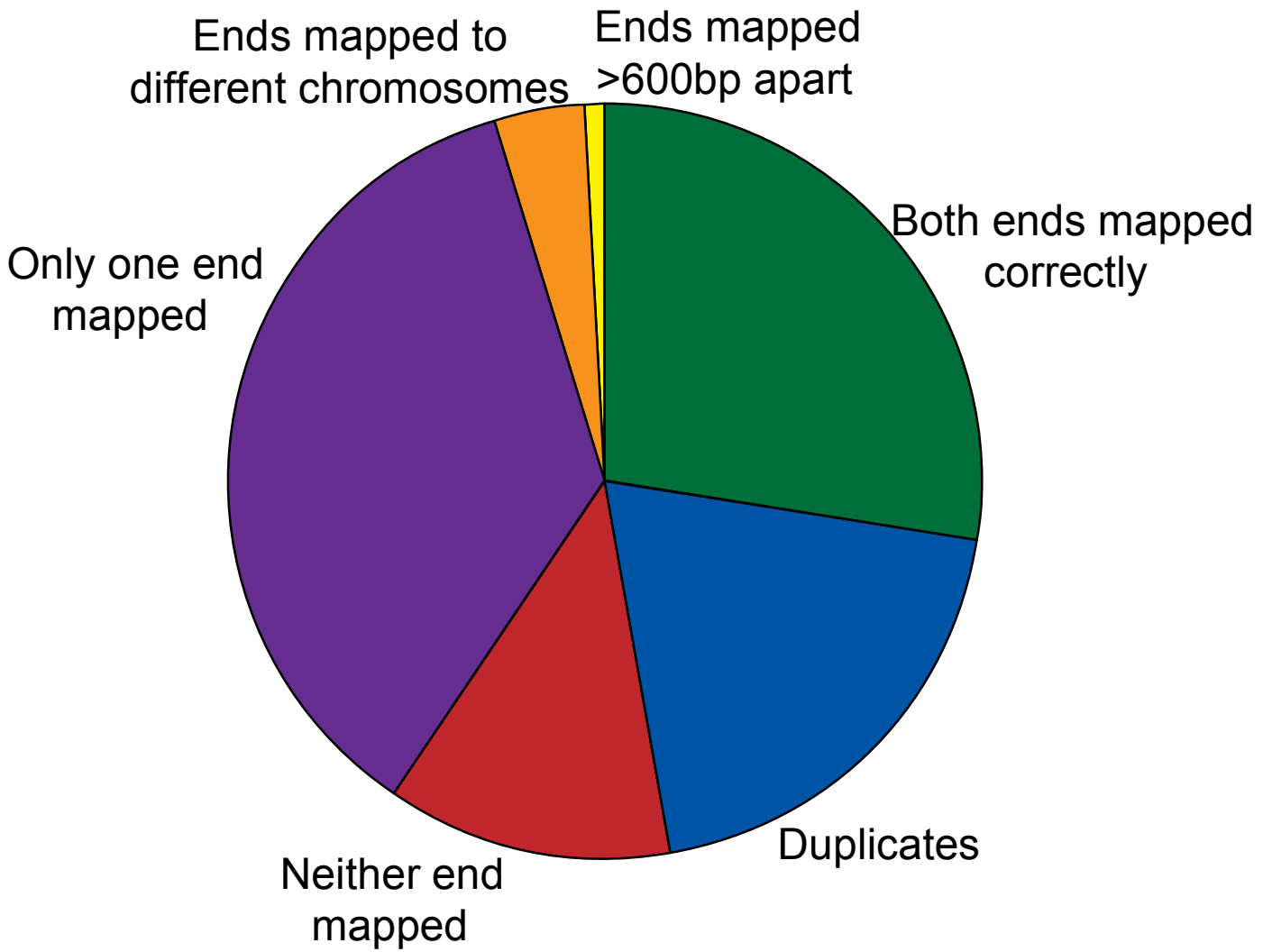
Genome-wide rearrangement screen



Protocol


(i)
(ii)
400bp
(iii)
(iv)
ACCCGTACCGTGCAATTTG..CCATTCCTTTCAATTGGC
400bp in ref seq
ACCCGTACCGTACCATT
GTCATTCCTTTCAATTGGC
> 600bp in ref seq
Randomly sheared DNA with Solexa adapters
Size select
Sequence
Align to reference sequence

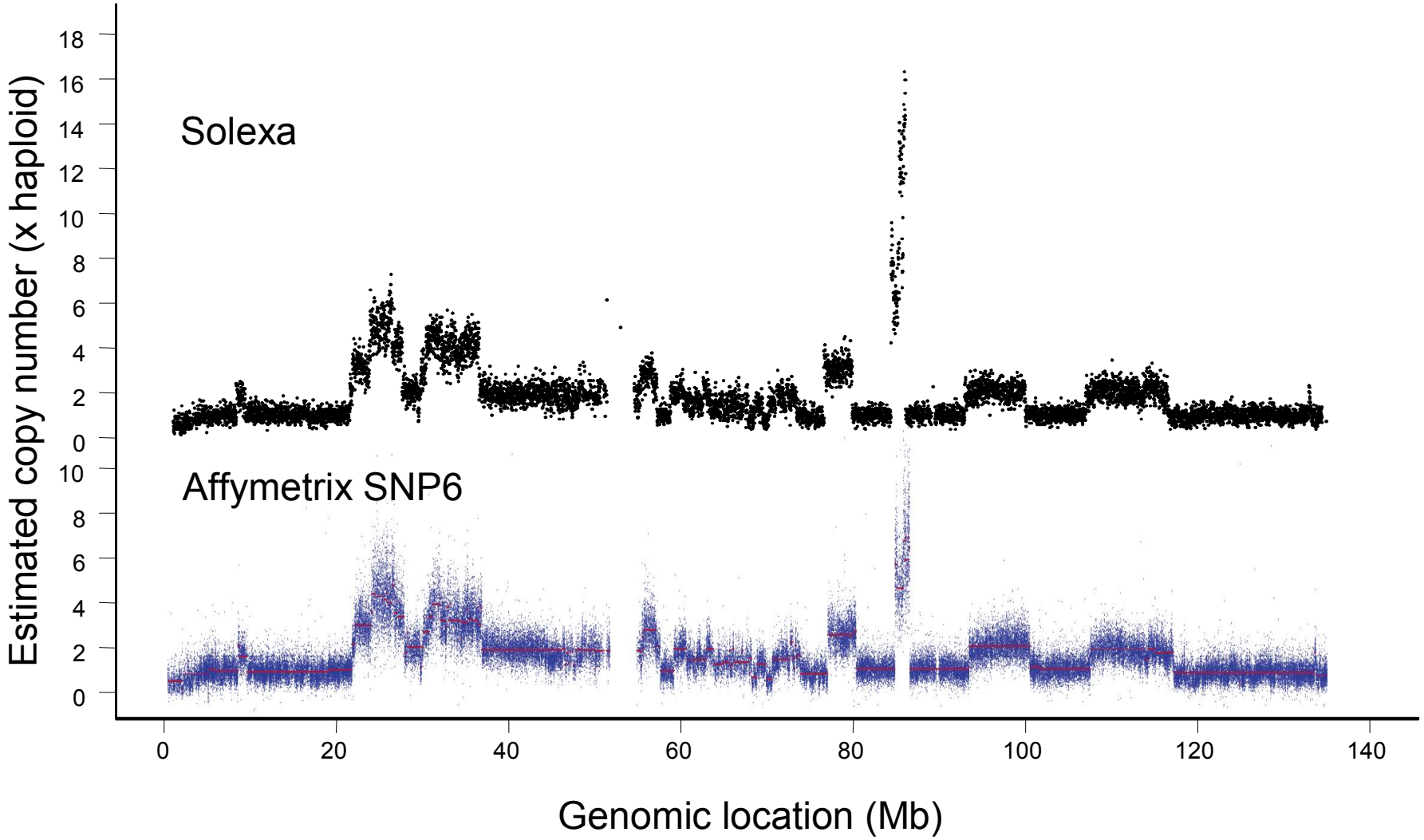
Mapping results



Artefacts

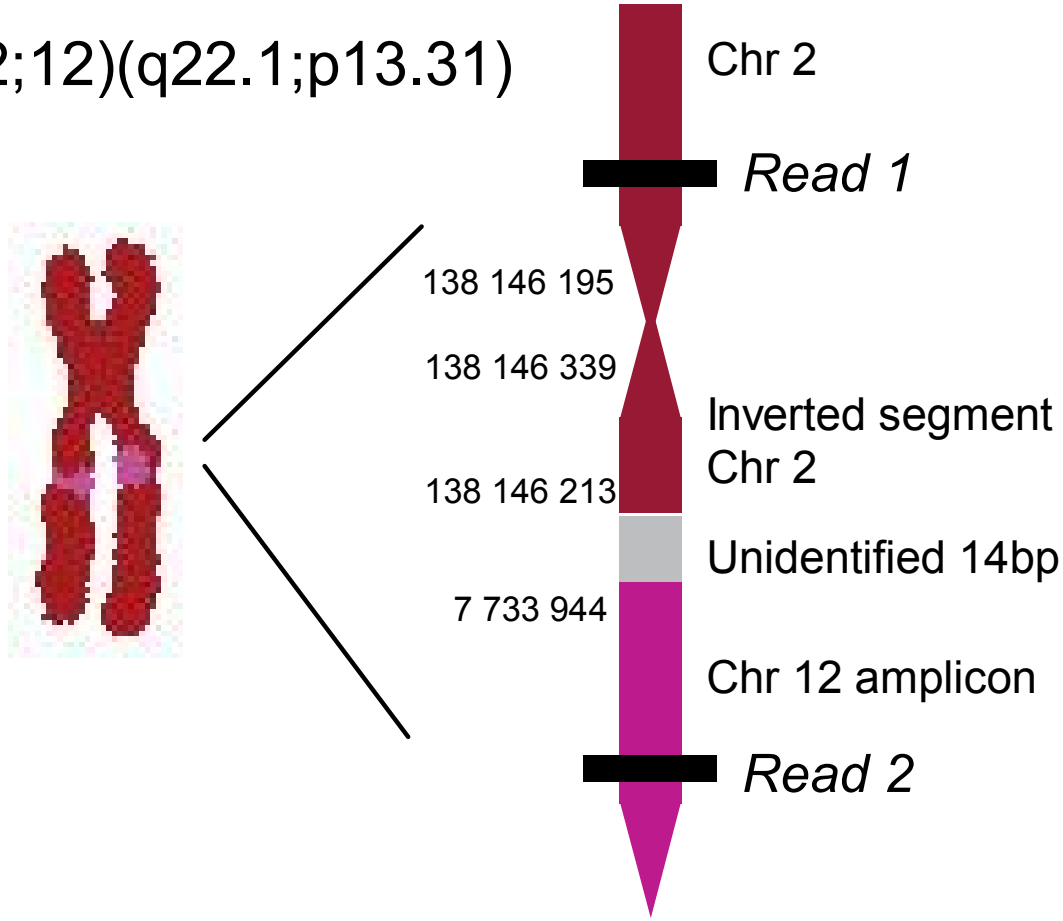
- Duplicate reads
- False priming
- Contaminating DNA
- Mapping errors
- Sequence gaps in reference genome
- False chimaeras

Copy number – Chr 11

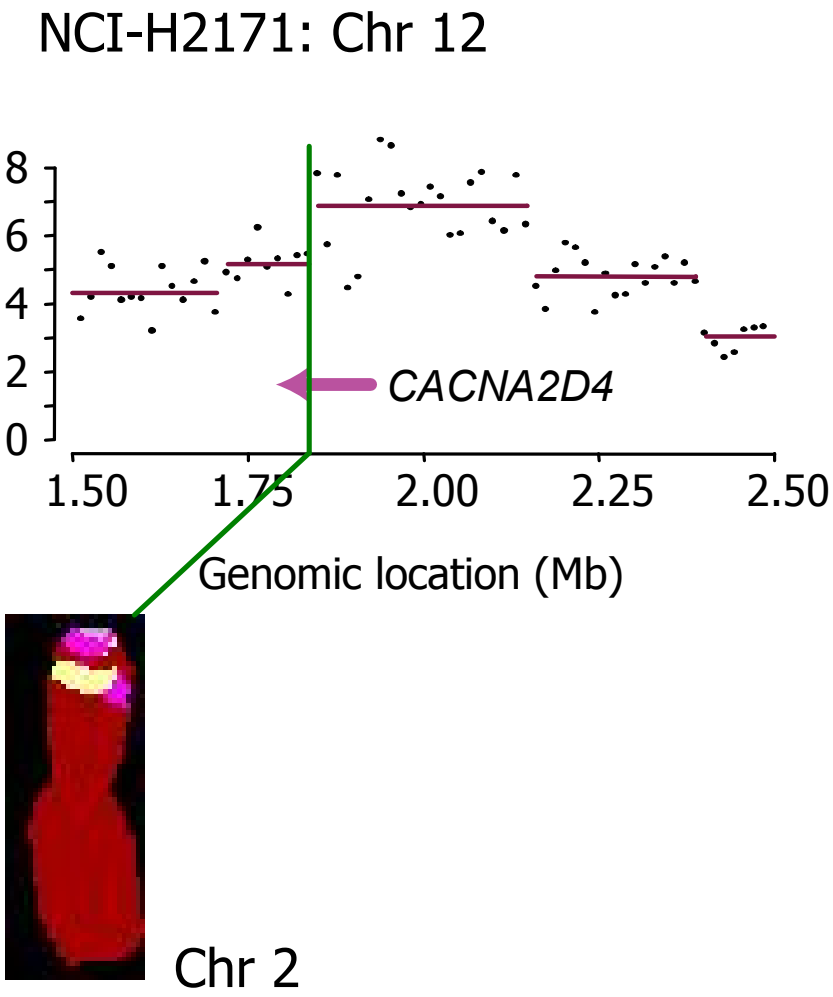


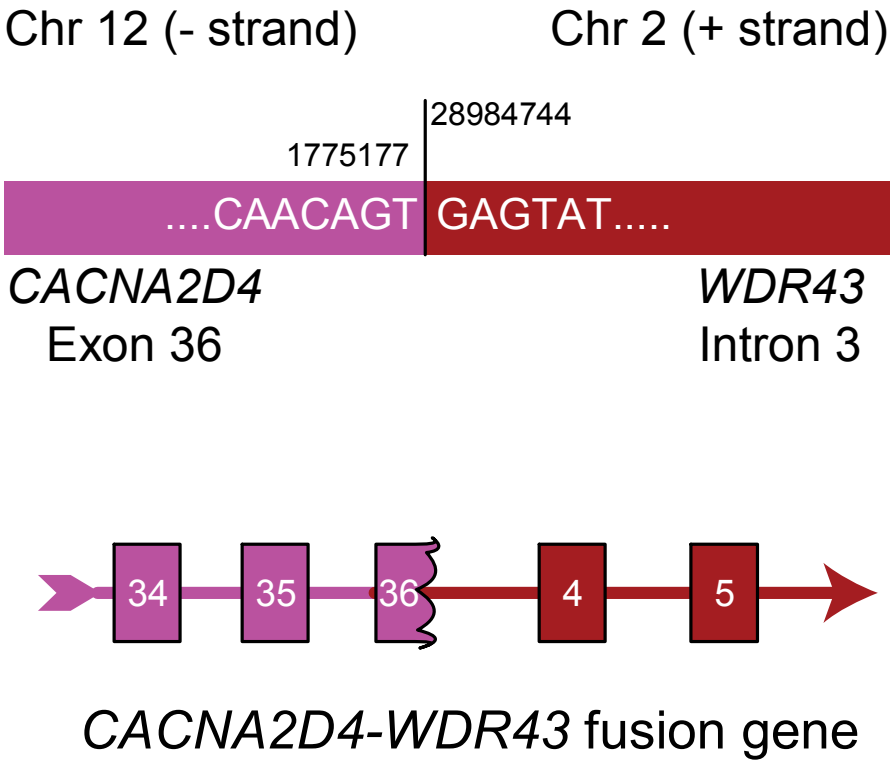
Rearrangement detection

der(2)ins(2;12)(q22.1;p13.31)

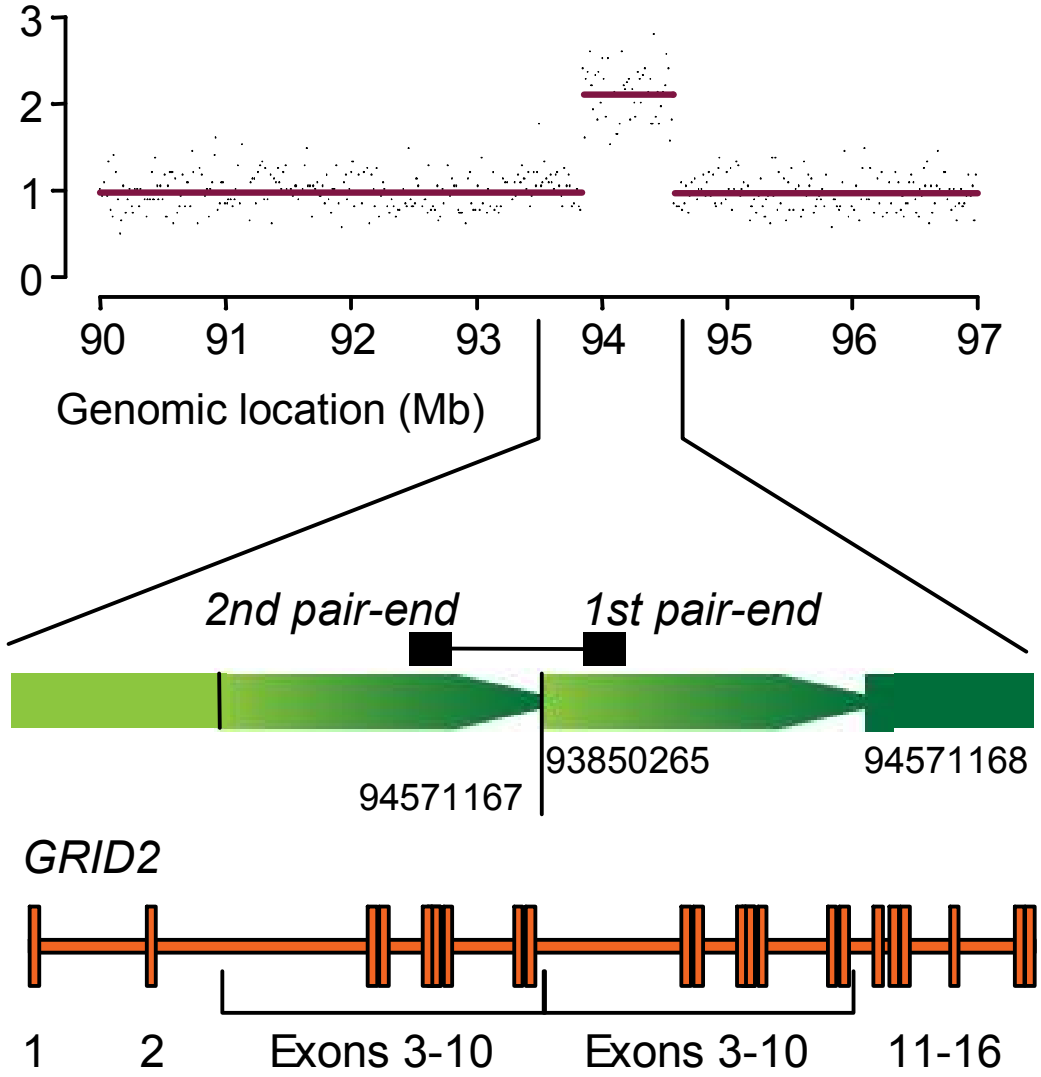


Fusion gene

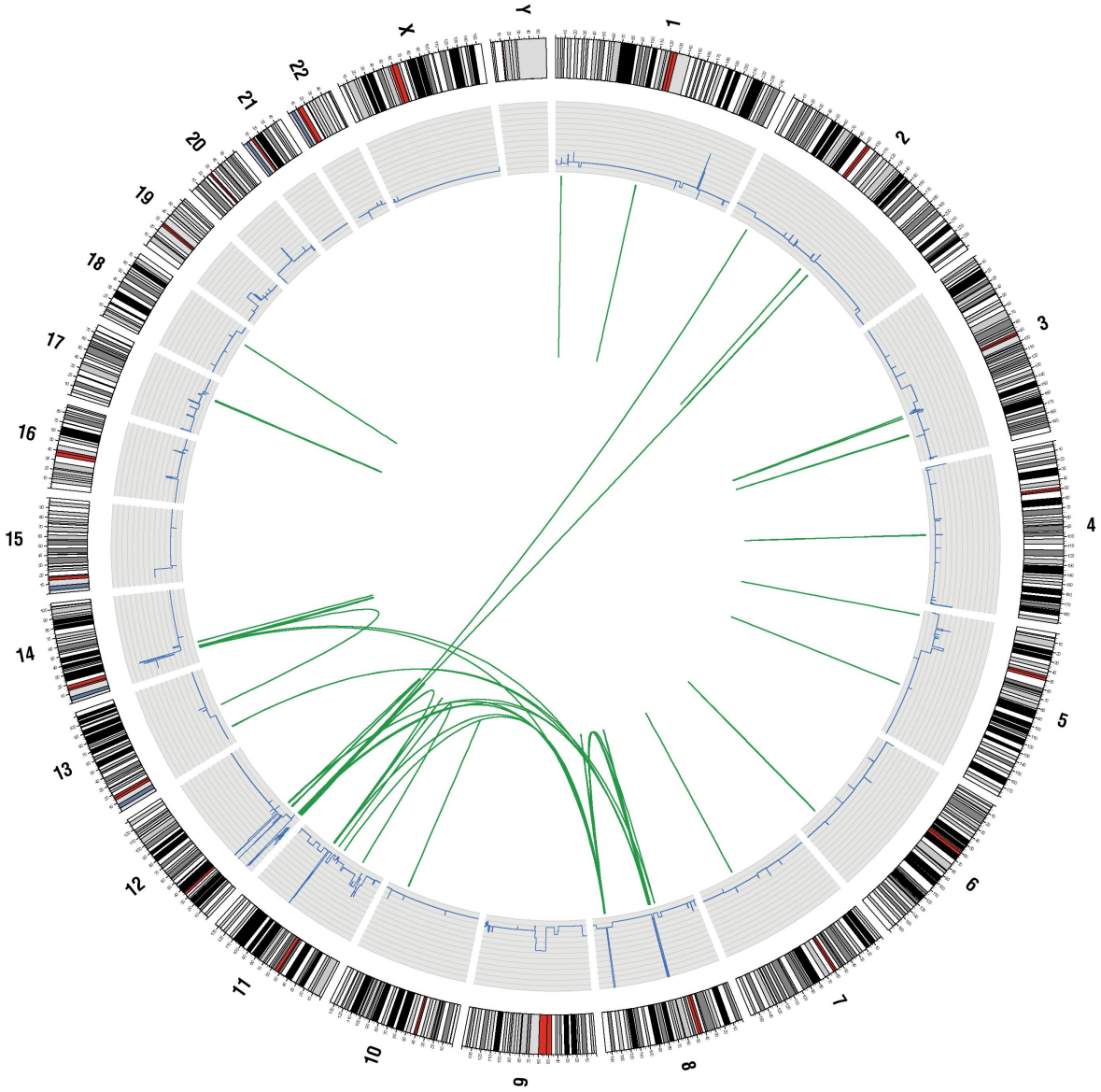

NCI-H2171: Chr 12
8
6
4
2
0
1.50
1.75
2.00
2.25
2.50
CACNA2D4
Genomic location (Mb)
Chr 2


Chr 12 (- strand)
Chr 2 (+ strand)
28984744
1775177
....CAACAGT
GAGTAT.....
CACNA2D4
Exon 36
WDR43
Intron 3
34
35
36
4
5
CACNA2D4-WDR43 fusion gene

Tandem duplication – Chr 4



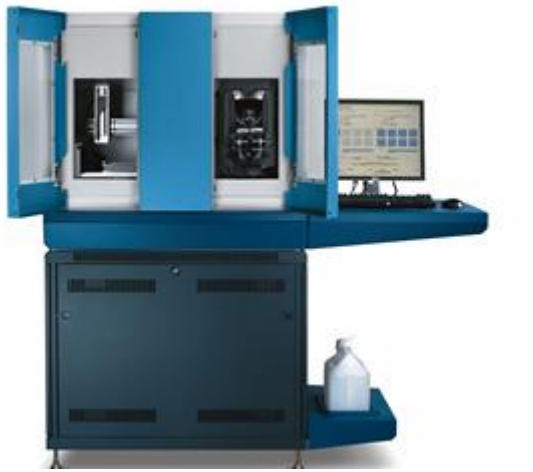
Acquired rearrangements



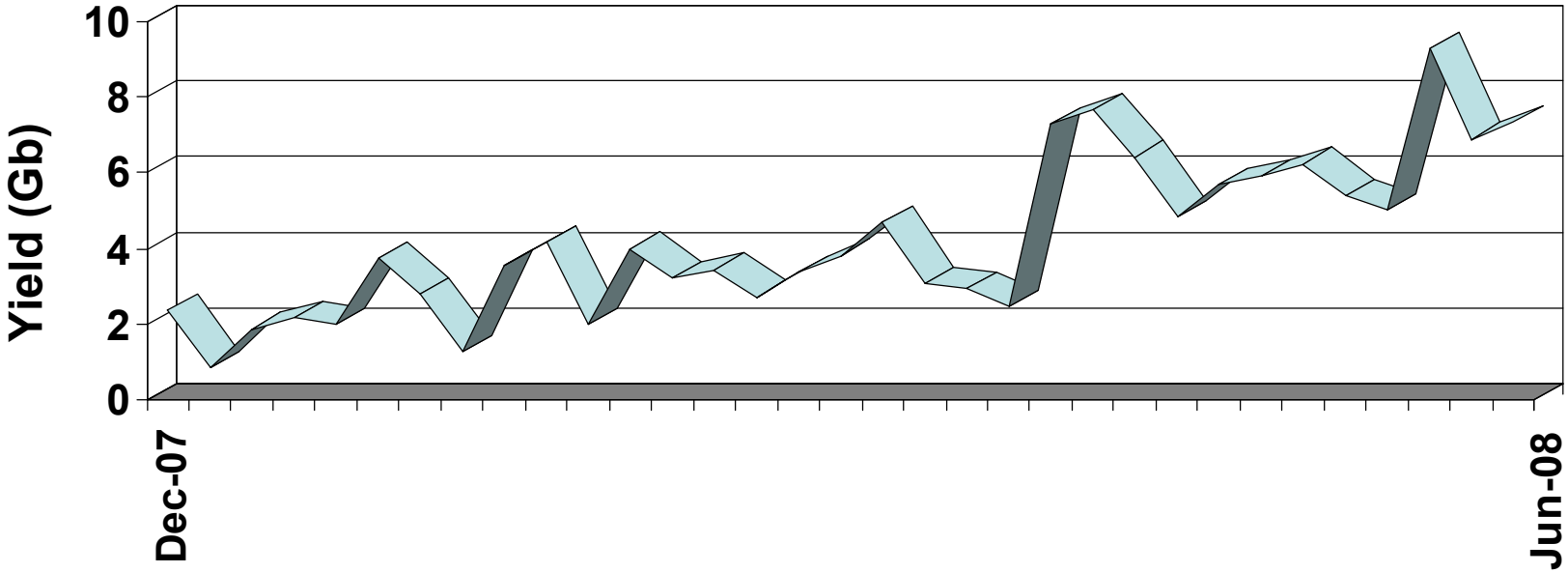
SOLiD platform

- Chemistry: Sequencing by ligation

- Sequencing & mapping in “colour space” (double base encoding)
- Paired end sequencing: 2 x 25bp
- Insert size: 600bp – 5kb



Sequencing a whole cancer genome



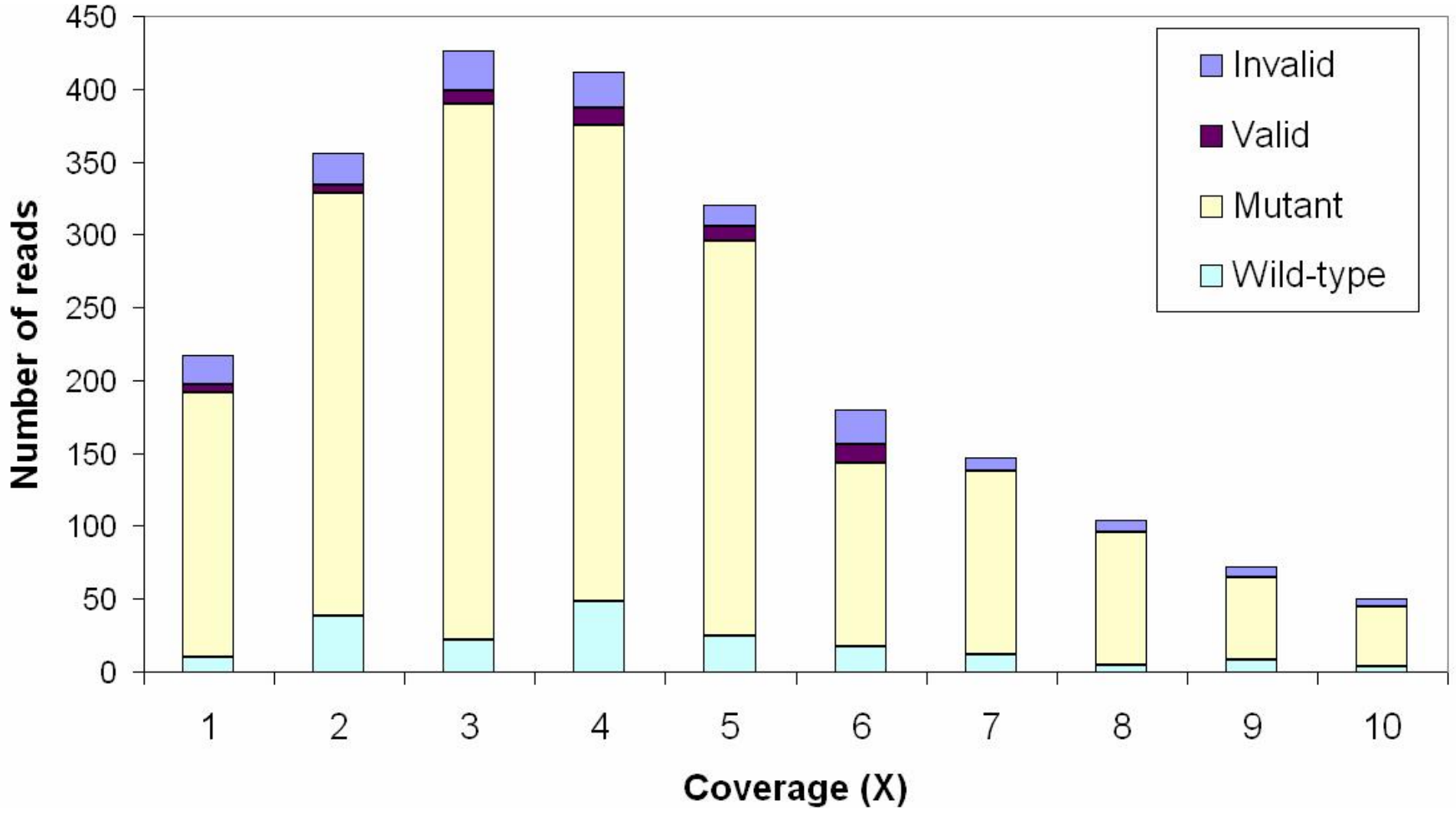
Error rates in known sequence

- 5 Mb of sequence analysed by capillary sequencing, known to have no variants

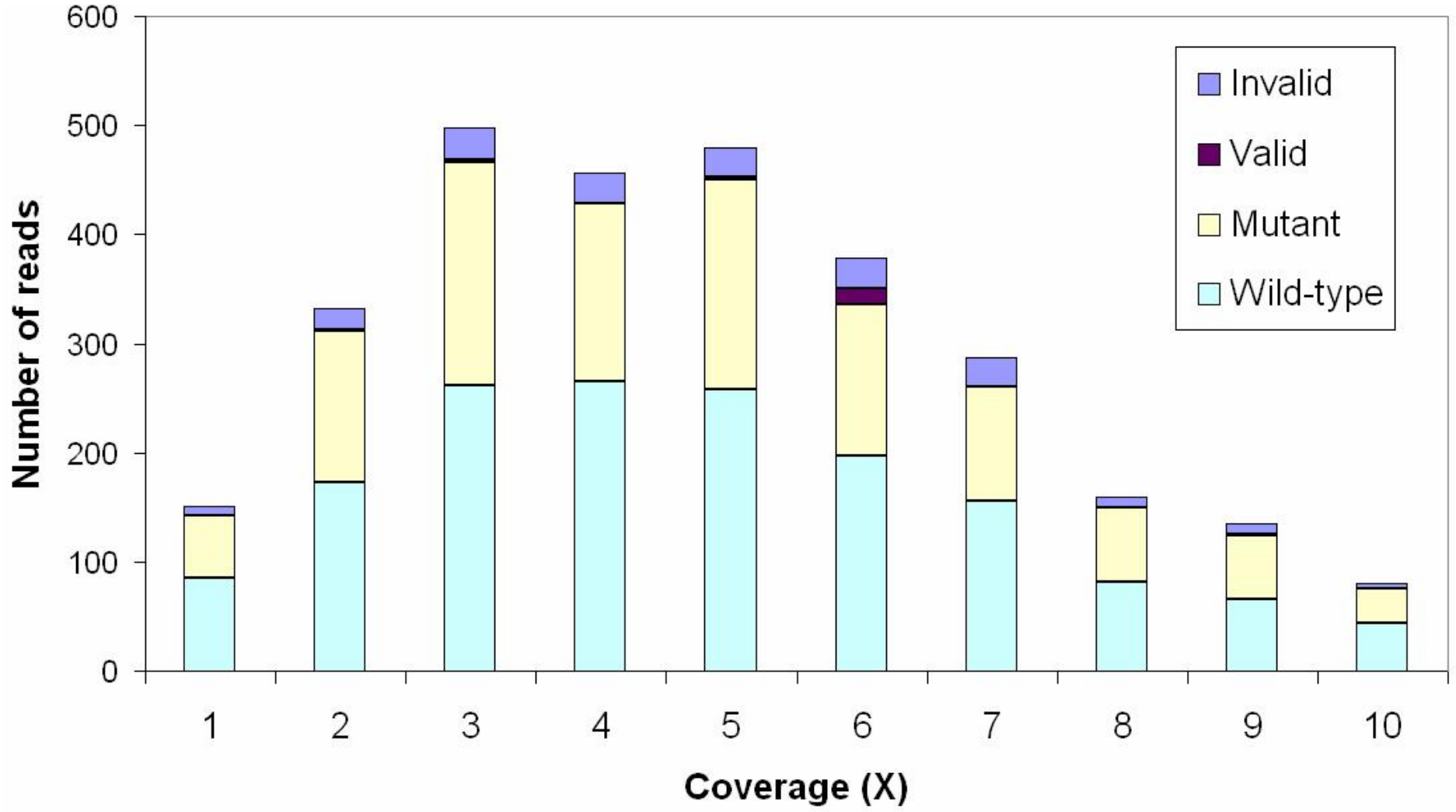
Total errors	'Invalid' errors	'Valid' errors
2.03%	0.13%	0.08%

- Only 'valid' errors could contribute to false positive variant calls
- Basic quality filtering reduces error rates by 50%

Homozygous variants



Heterozygous variants



	454	Solexa	SOLID
			
Chemistry:	Pyrosequencing	Fluorescent In-Situ	Ligation
Parallelisation:	400 K	30 million	50 million
Read length:	~250 bps	35 bps PE	25 bps PE
Sequence:	~150 Mb	~4 Gb PE	~14 Gb PE
Run time:	7.5 hours	6 days PE	10 days PE

Applications

- Ultra-deep sequencing
  - Minimal residual disease monitoring
  - Drug resistant subclones
  - Mutation screening in samples contaminated with normal cells

Applications

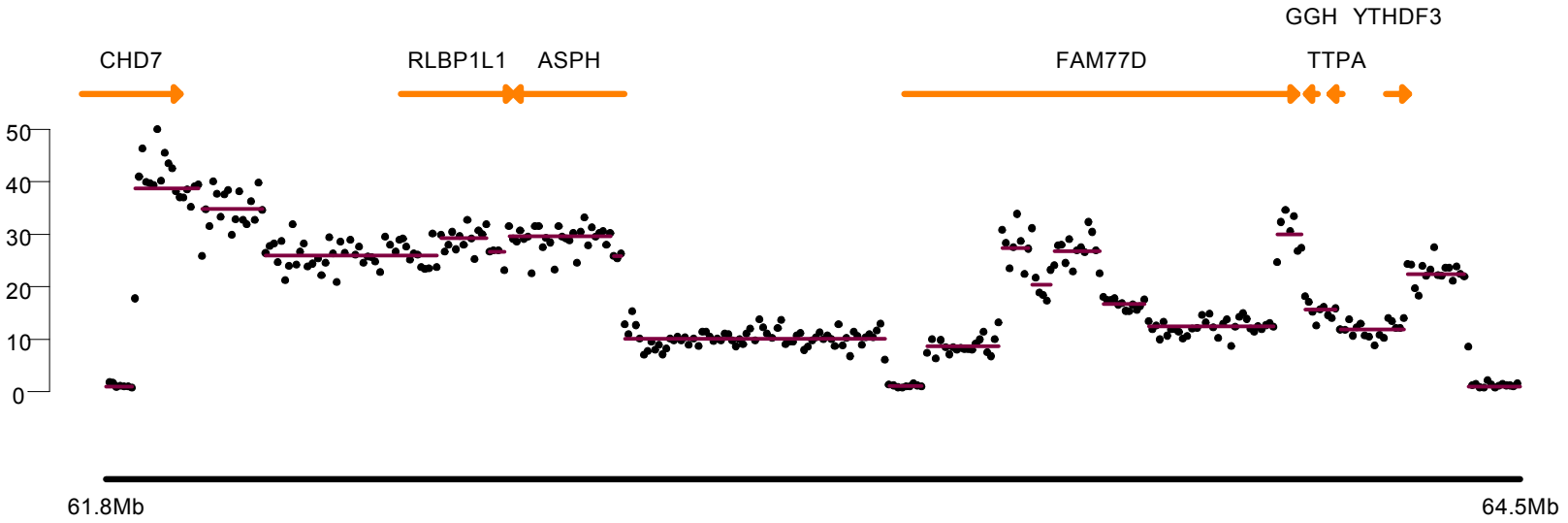
- Broad sequencing
  - Whole-genome germline and somatic rearrangements
  - Genetic screening in multiple samples
    - All 3 platforms have capacity to barcode samples
  - Whole genome resequencing in real time!

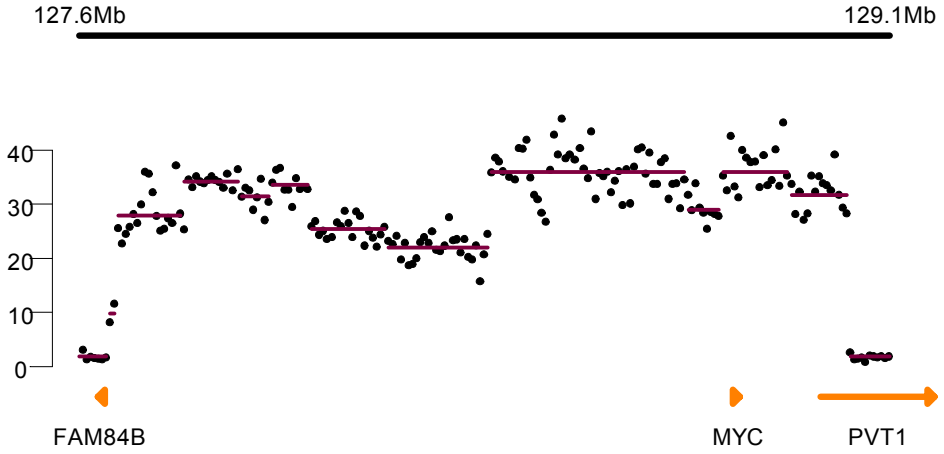
Acknowledgements

- Phil Stephens
- Erin Pleasance
- Sarah O’Meara and Claire Hardy
- Lucy Stebbings, Catherine Leroy & Andy Menzies
- Heng Li
- Ian Goodhead, Richard Rance and Dan Turner
- George Follows & Tony Green
- Mike Stratton & Andy Futreal

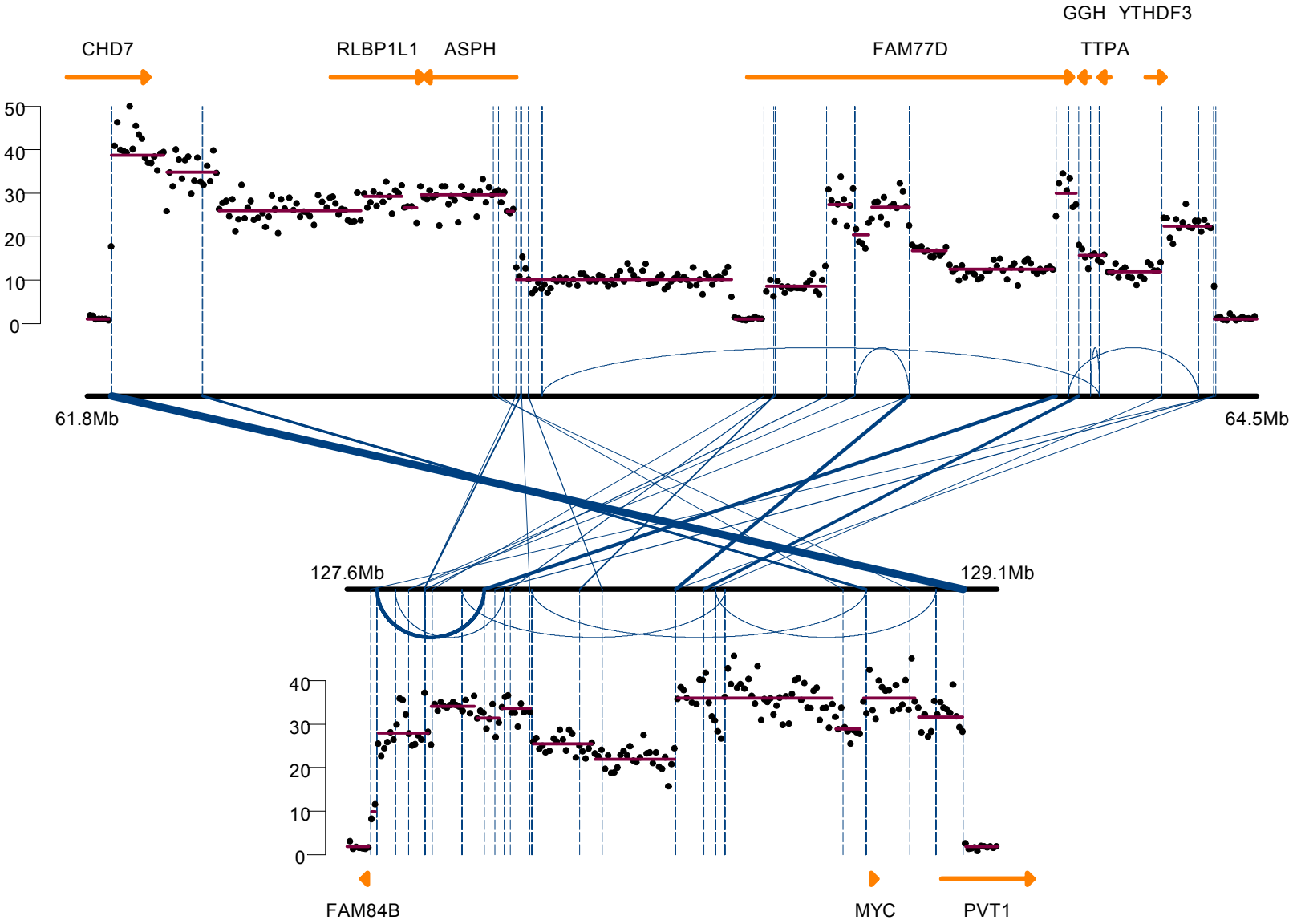
Amplicons on chr 8q

Amplicon breakpoint detection





Amplicon breakpoint detection



PVT1-CHD7 fusion gene