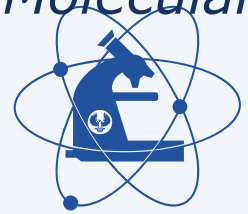
Introduction to next-generation sequencing and variant calling

BioInfoSummer 2013 Adelaide

5th December 2013

Dr Karin Kassahn
Head, Technology Advancement Unit,
Genetic & Molecular Pathology



SA PATHOLOGY

Environment/ Ecology

The Genome 10K project aims to **assemble a genomic zoo** — a collection of DNA sequences representing the genomes of 10,000 vertebrate species, approximately one for every vertebrate genus.

https://genome10k.soe.ucsc.edu/

SA PATHOLOGY

Environment/ Ecology

Re Fu Ge
2o 2o

Sea-quence

Generating core genetic data for corals from the Great Barrier Reef and Red Sea, helping to guide management responses in the face of climate change—

Sea-quence

The first first phase of this work is the Sea-quence Project, an initiative of the ReFuGe 2020 Consortium, convened by the Great Barrier Reef Foundation, and

http://www.barrierreef.org/our-projects/sea-quence

SA PATHOLOGY

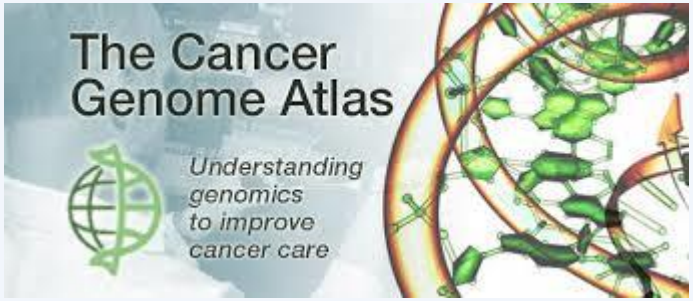
Human Health

£100 million to “sequence **100,000 whole genomes** of NHS patients at diagnostic quality over the next three to five years”.

SA PATHOLOGY

Human Health

International Cancer Genome Consortium



ICGC Goal: To obtain a **comprehensive**description of **genomic, transcriptomic and epigenomic changes** in **50 different tumor types and/or subtypes** which are of clinical and societal importance across the globe.

http://icgc.org/

TCGA: to chart the genomic changes involved in more than 20 types of cancer.

http://cancergenome.nih.gov/

SA PATHOLOGY

Clinical Diagnostics

NGS Clinical Test	Type	Applications
Illumina TruGenome Clinical Sequencing	whole-genome	Undiagnosed disease (single-gene etiology); predisposition screen
Foundation One	40 gene panel	Cancer stratified treatment and response
Maternity21	Low coverage whole-genome	Non-invasive prenatal (chromosomal abnormalities)

SA PATHOLOGY

16 May 2013

Media

NEWS

Just In | Australia | World | Business | Sport | Analysis & Opinion | More

Print | Email | Facebook | Twitter | More

DNA sequencing set to become routine medicine

By Melinda Howells and Kim Lyell

Updated Thu May 16, 2013 5:49pm AEST

PHOTO: Angelina Jolie underwent a double mastectomy after discovering a genetic variation dramatically increased her risk of breast cancer. (AFP: Ethan Miller/Getty Images - file image)

SA PATHOLOGY

Media

THE AUSTRALIAN LOGIN

NEWS

8 July 2013

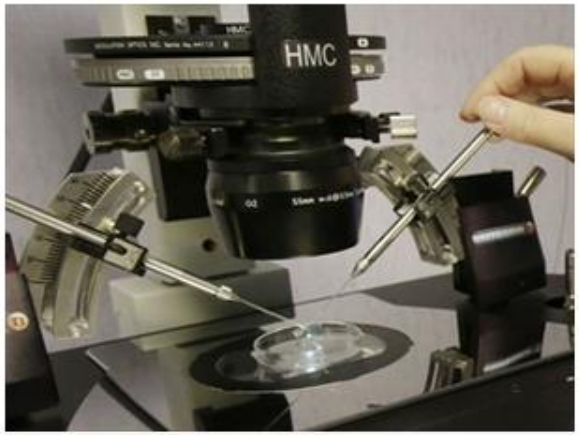
New gene sequencing yields healthy baby

AAP July 08, 2013 11:23AM

SCIENTISTS say they have used a new-generation gene sequencing technique to select a viable embryo for in-vitro fertilisation (IVF) that yielded a healthy baby boy.

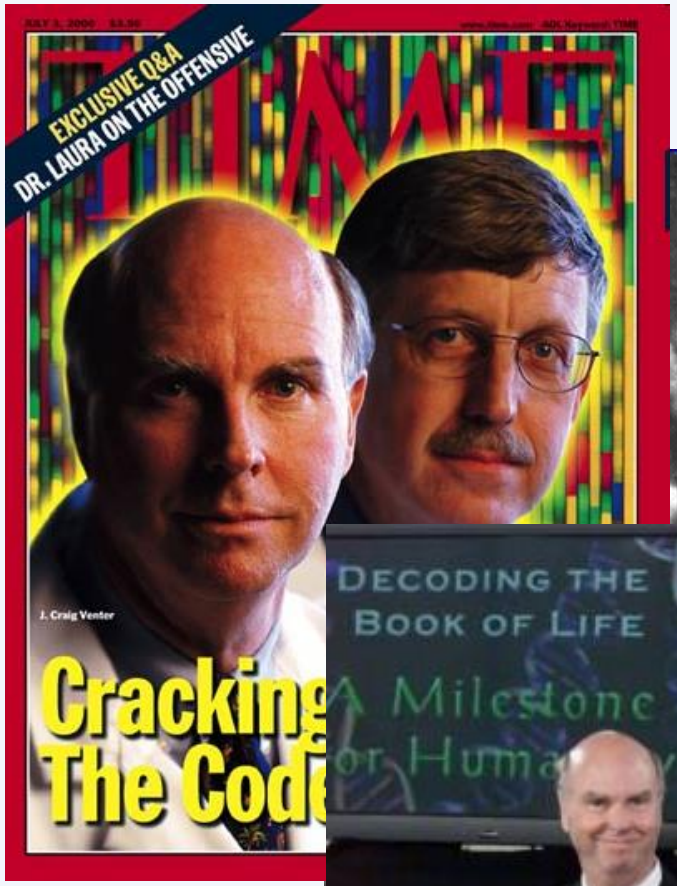
IVF, the process whereby a human egg is fertilised with sperm in the laboratory, is a hit-and-miss affair, with only about 30 per cent of fertilised embryos resulting in pregnancy after implantation.

The reason for the high failure rate is not clear but genetic defects are the prime suspects, according to the authors of the paper presented on Monday at a meeting in London of the European Society of Human Reproduction and Embryology.



New-generation gene sequencing helps identify the best embryos in IVF, a UK researcher says. *Source:* AAP

SA PATHOLOGY



June 25, 2000

SA PATHOLOGY

Sequencing of genomes is ubiquitous (evolutionary, ecological, medical studies)

It is becoming part of standard clinical practise

It is entering public life (media, …)

SA PATHOLOGY

Outline

The Technology Advance

Overview of NGS workflows

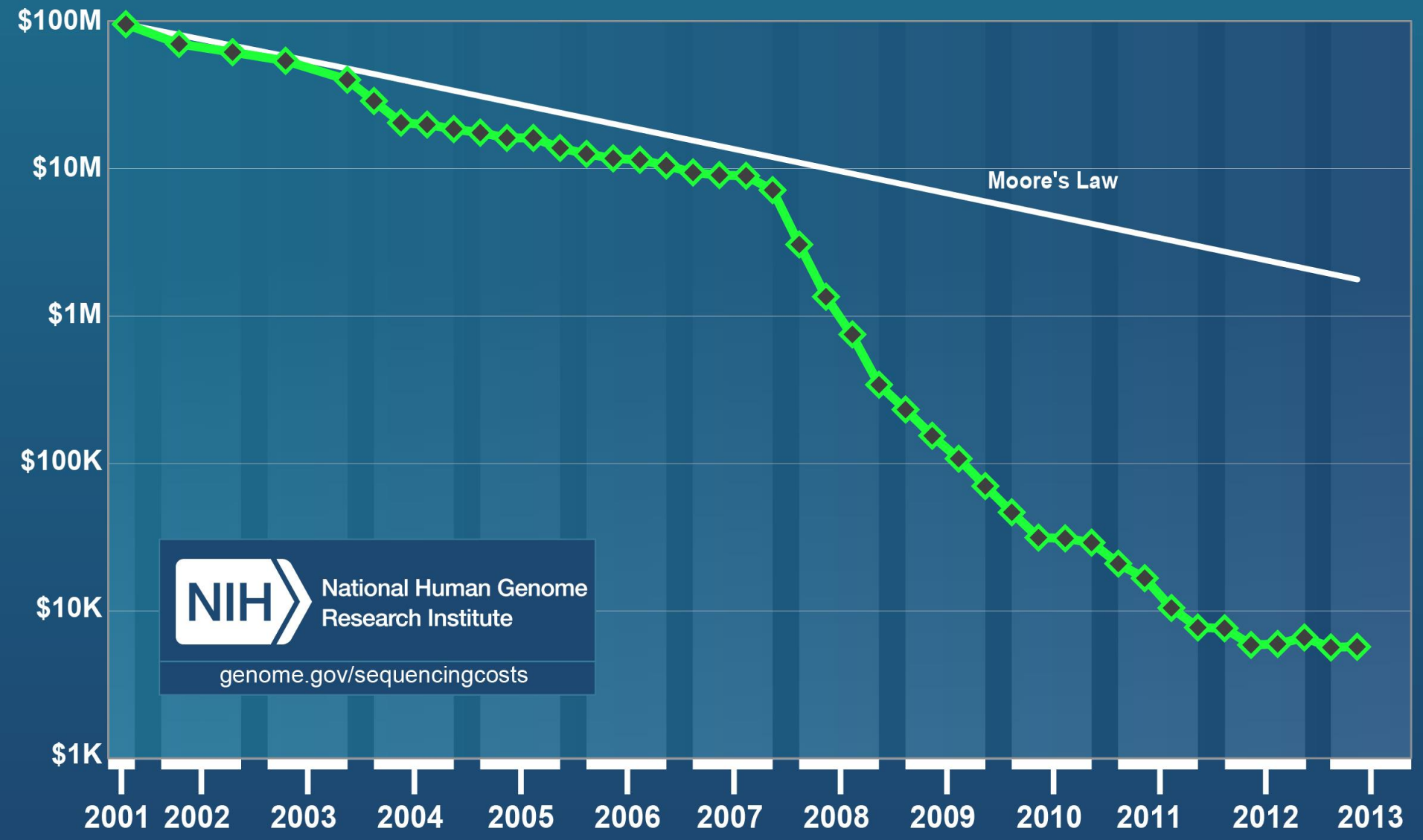
- Laboratory
- Bioinformatics analysis
  - Base calling
  - Alignment
  - Variant calling
  - Variant annotation

SA PATHOLOGY

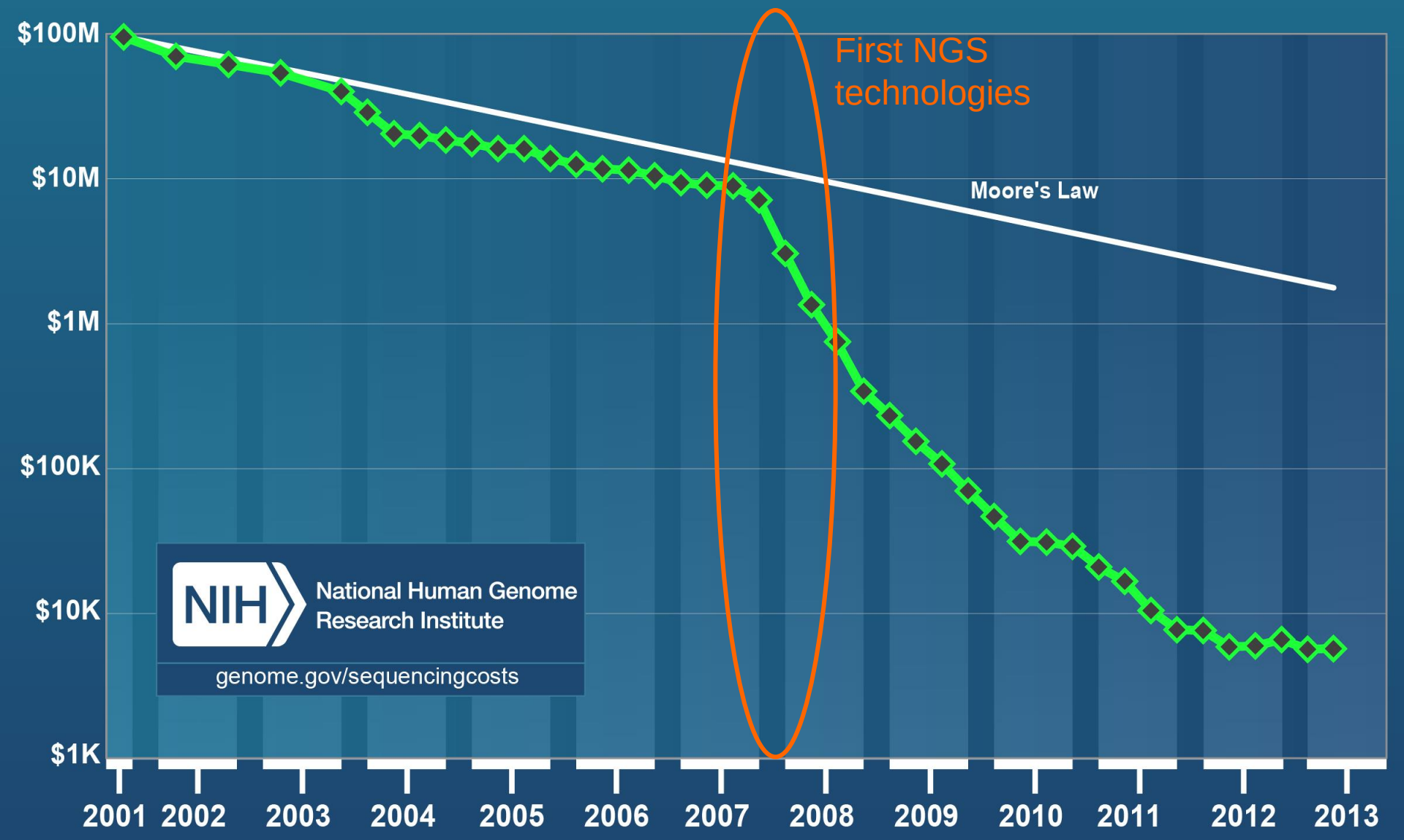
The Technology Advance

SA PATHOLOGY

Cost per Genome



Cost per Genome



Advances in Sequencing Technology

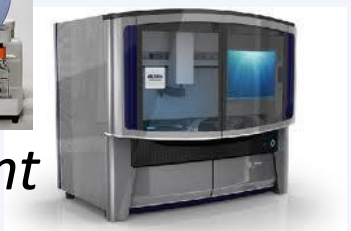
AB3700

1998

SOLiD

2008

IonTorrent



SOLiD 5500xl

2010

IonProton

2012

2005

454

2009

HiSeq

2011

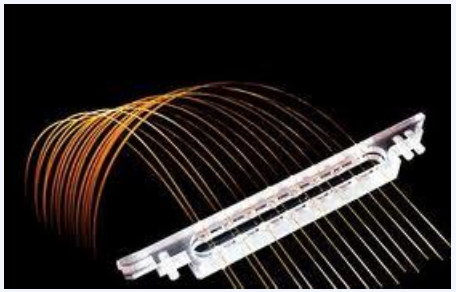
MiSeq

Complete Genomics

PacBio RS

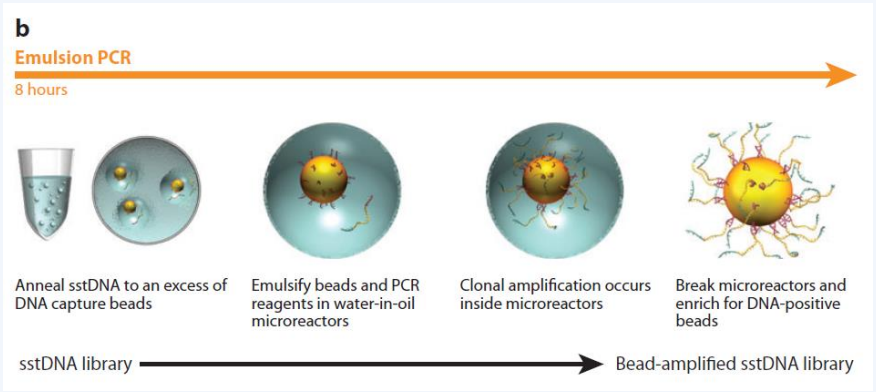
SA PATHOLOGY

Miniaturization and Parallelisation

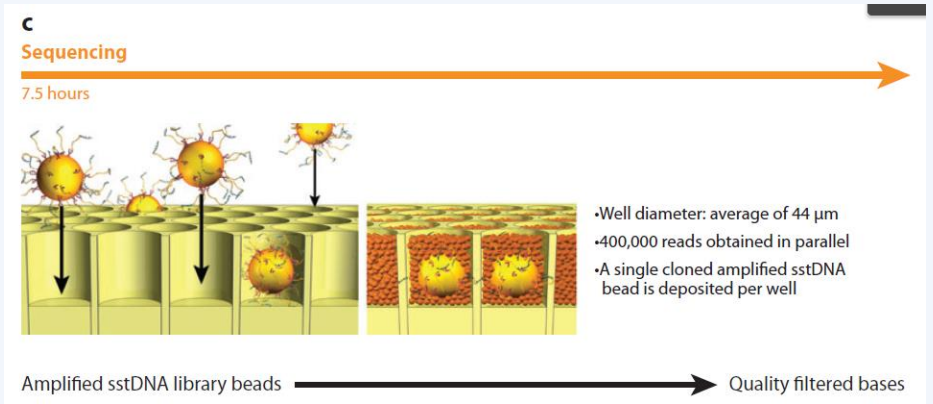


Capillary Sequencing (Sanger)

Emulsion PCR



Sequencing in wells



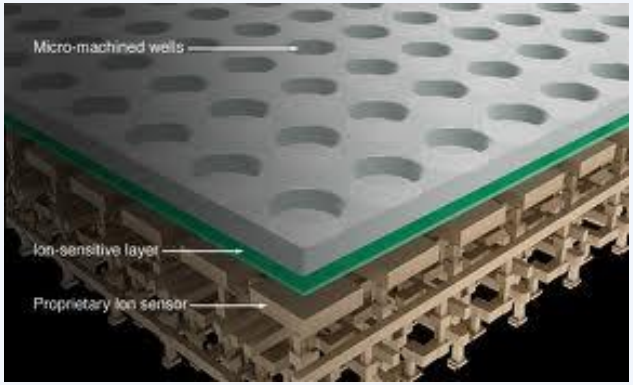
Images: Elaine Mardis, 2008

454 Pyrosequencing

SA PATHOLOGY

Miniaturization and Parallelisation *(cont)*

Sequencing in ever-smaller wells



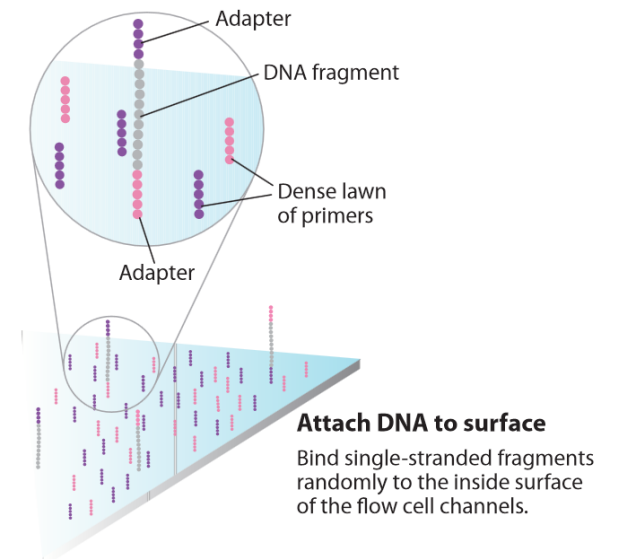
IonTorrent

IonProton

Chip	Wells	Reads
Torrent 314	1.2 million	400-500 thousand
Torrent 316	6.2 million	1.9 – 2.5 million
Torrent 318	11.1 million	3.3 – 4.4. million
Proton I	165 million	60 – 80 million

SA PATHOLOGY

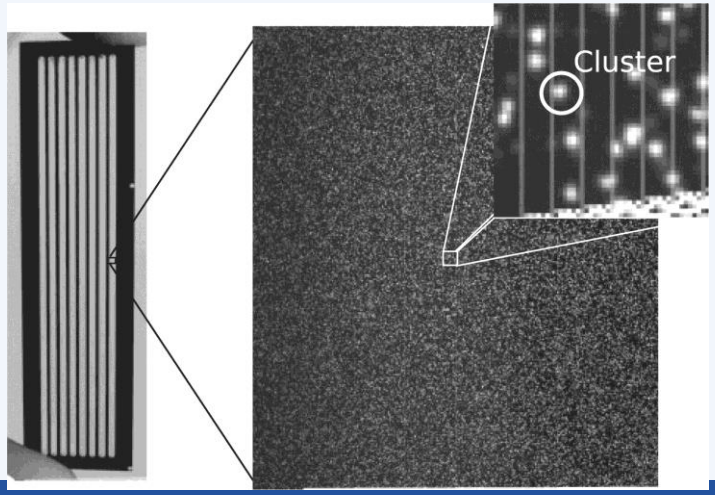
Miniaturization and Parallelisation *(cont)*

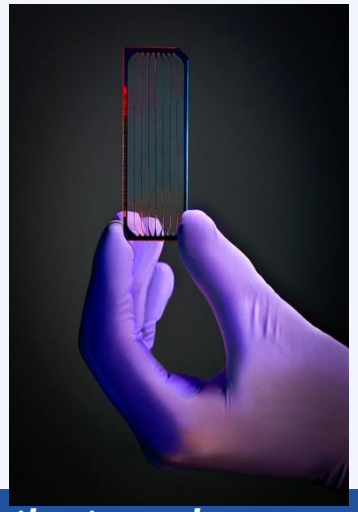


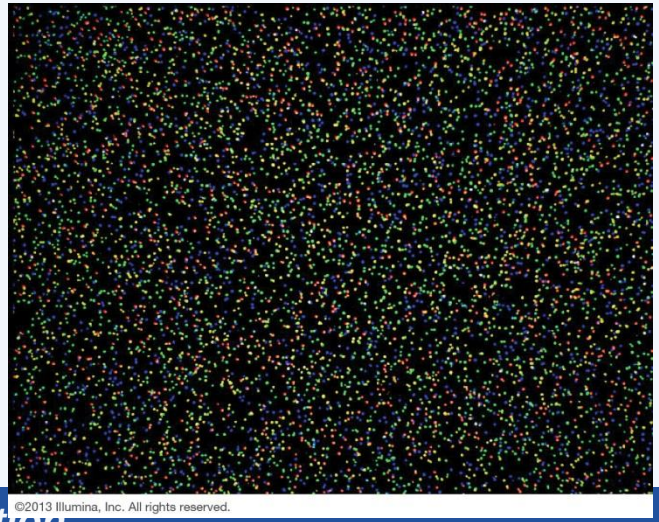
DNA on slides (Solexa)

Technology	Reads
Solexa	150-200 million
HiSeq 2000	3 billion

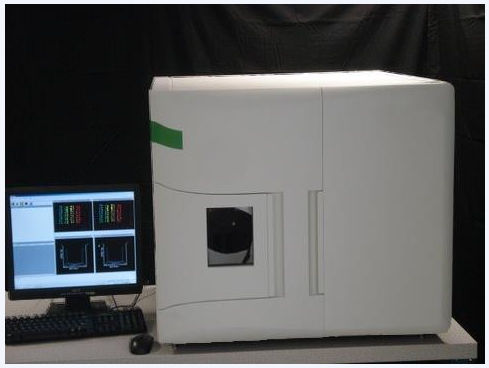
DNA flowcells (Illumina HiSeq)







New Sequencing Technologies on the Horizon



Intelligent Biosystems

MinIon Oxford Nanopore

gnubio

Qiagen GeneReader

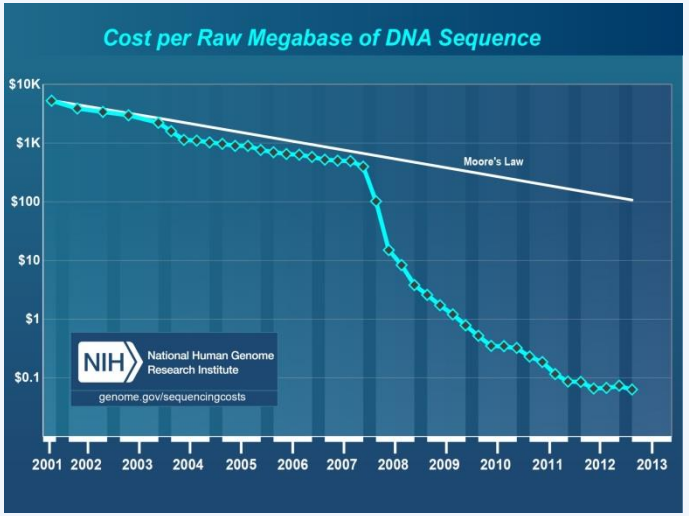
VISIGEN BIOTECHNOLOGIES, INC

LaserGen, Inc.

Helicos

SA PATHOLOGY

Technology Targets



reduce

- Cost
- DNA input
- Length of workflows

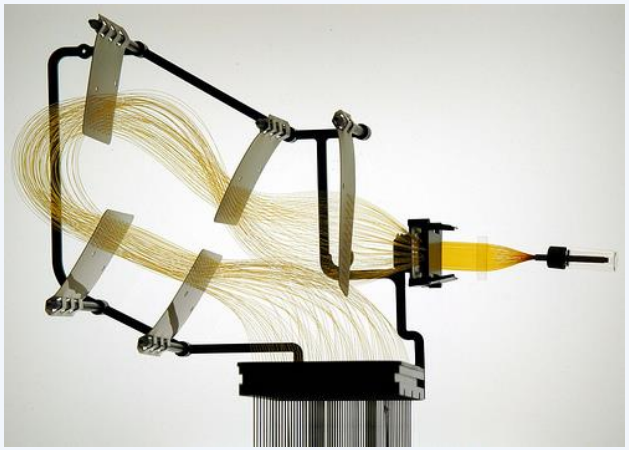
increase

- Sequencing accuracy
- Read length
- Detection of DNA base modification (methylation, …)
- Single-molecule sequencing (phasing, SVs …)

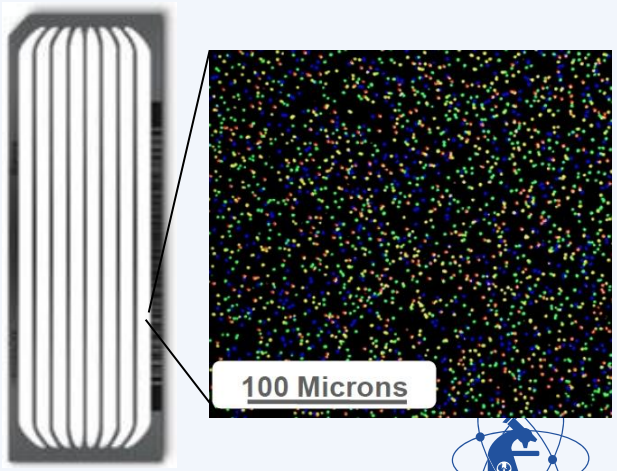
SA PATHOLOGY

First Gen vs Next Gen

Capillary DNA Analyzers	 3730xL
Number of Capillaries	96
Long Read Length	850b
Base Calls / Day	960,000



NGS Sequencers	 HiSeq 2500/2000	 MiSeq	 IonTorrent
Number of Reads	1.5x10^9	15x10^6	5x10^6
Long Read Length	2x100bp	2x300bp	400bp
Output (maximum)	600Gb	15Gb	2Gb
Run Time	11 Days	40 Hrs	7 Hrs



Joel Geoghegan

SA PATHOLOGY

What more data enables you to do…

- Sequence the human genome in few days at $1000
(compare to Sanger:
3.2 Billion bp @ 800bp = 4 Mio reactions or 42,000 96well plates at 16 plates per day = 2,625 days and $$$Mio)

- Clinical:
improved diagnostic pick-up rate
faster turn-around time
cheaper

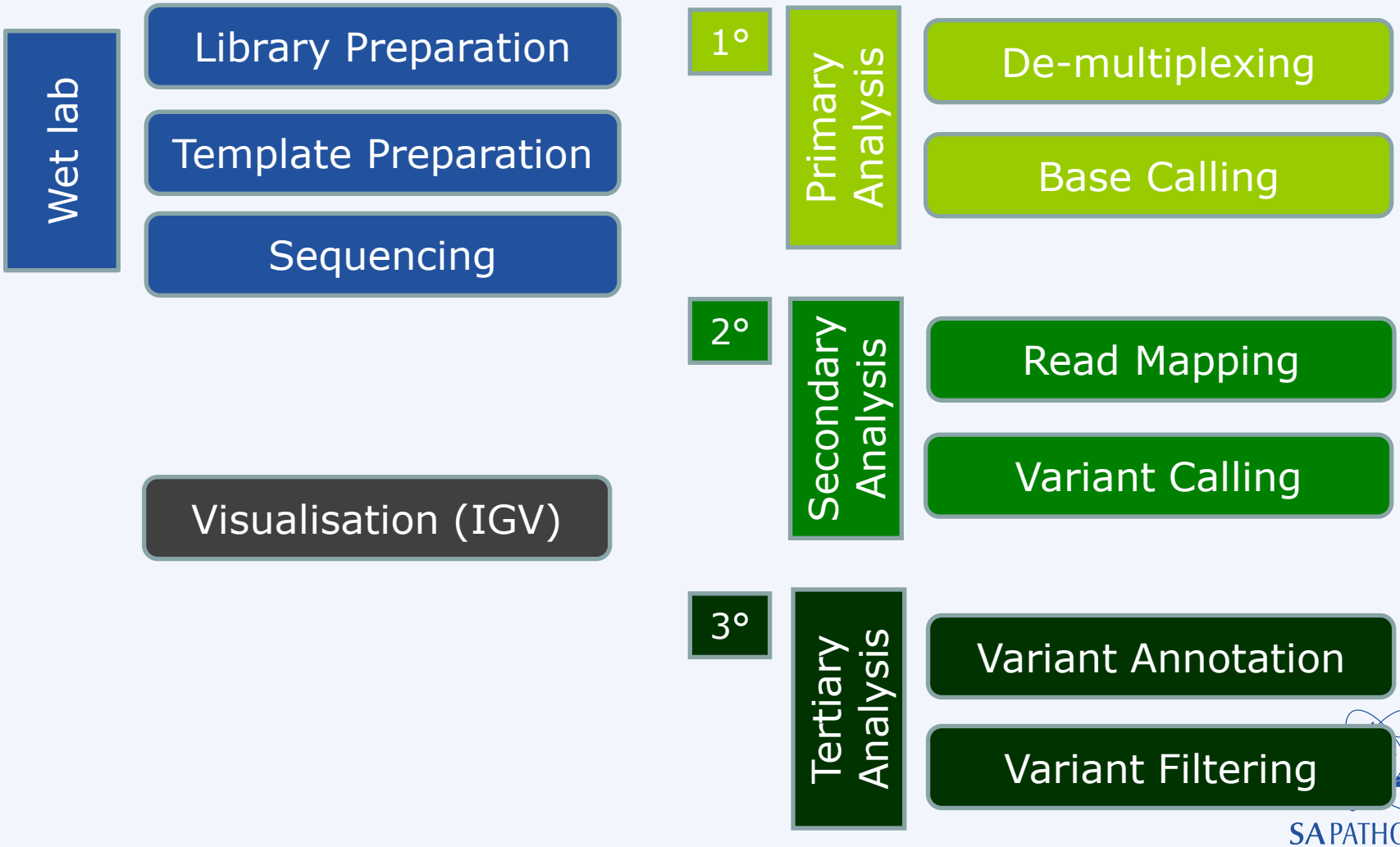
- Cancer:
improved sensitivity (sensitivity becomes a tunable parameter dependent on sequence depth)

SA PATHOLOGY

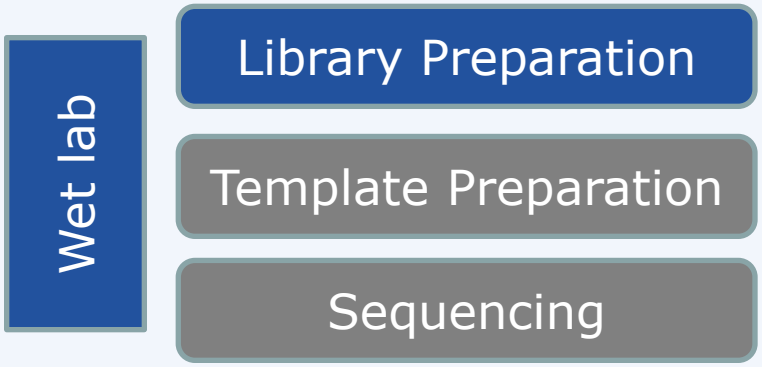
Overview of NGS workflows

SA PATHOLOGY

NGS workflow overview



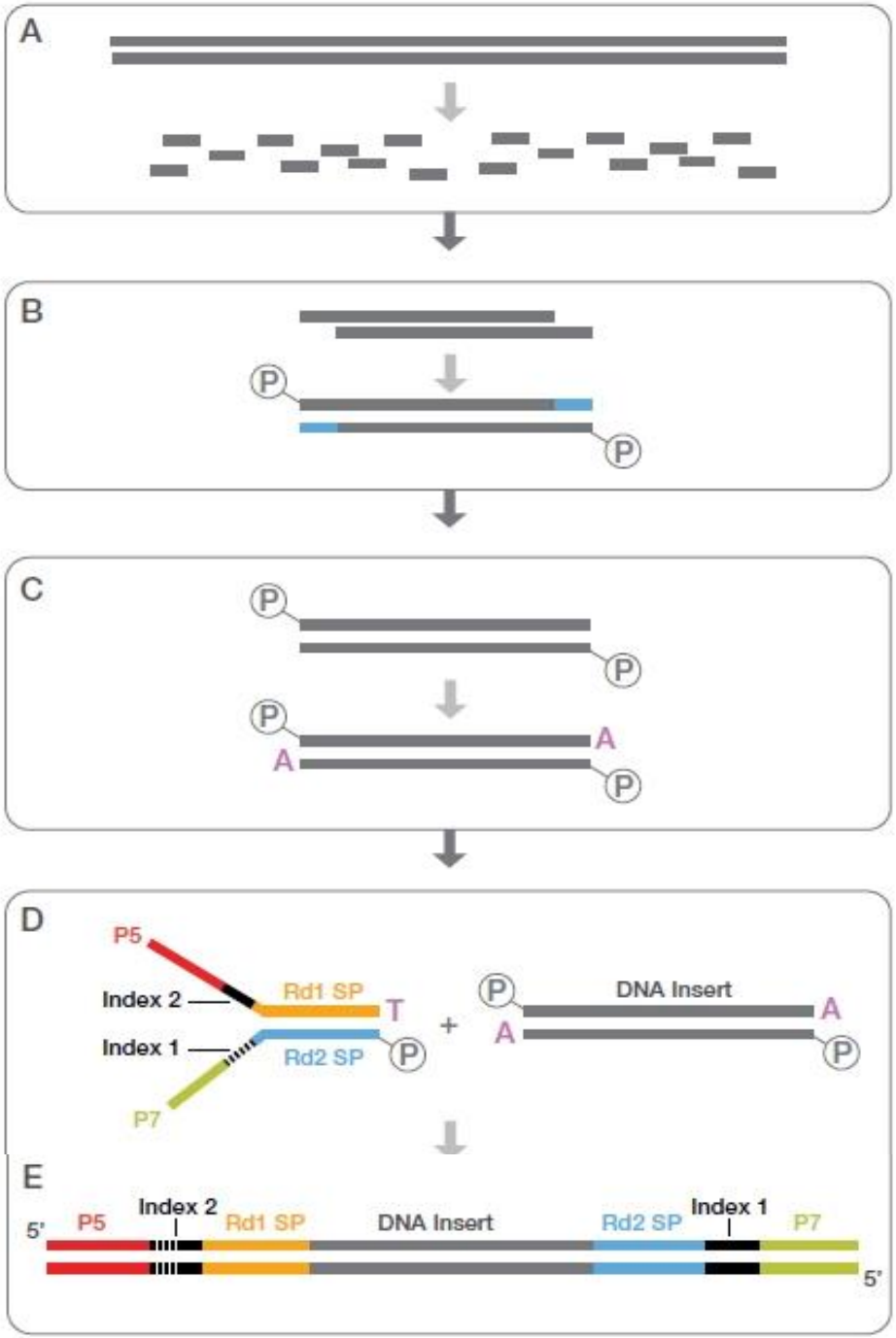
SA PATHOLOGY



- Select target
  - hybridization-based capture or PCR
- Add adapters
  - Contain binding sequences
  - Barcodes
  - Primer sequences
- Amplify material

SA PATHOLOGY

Illumina DNA library preparation



Fragment DNA

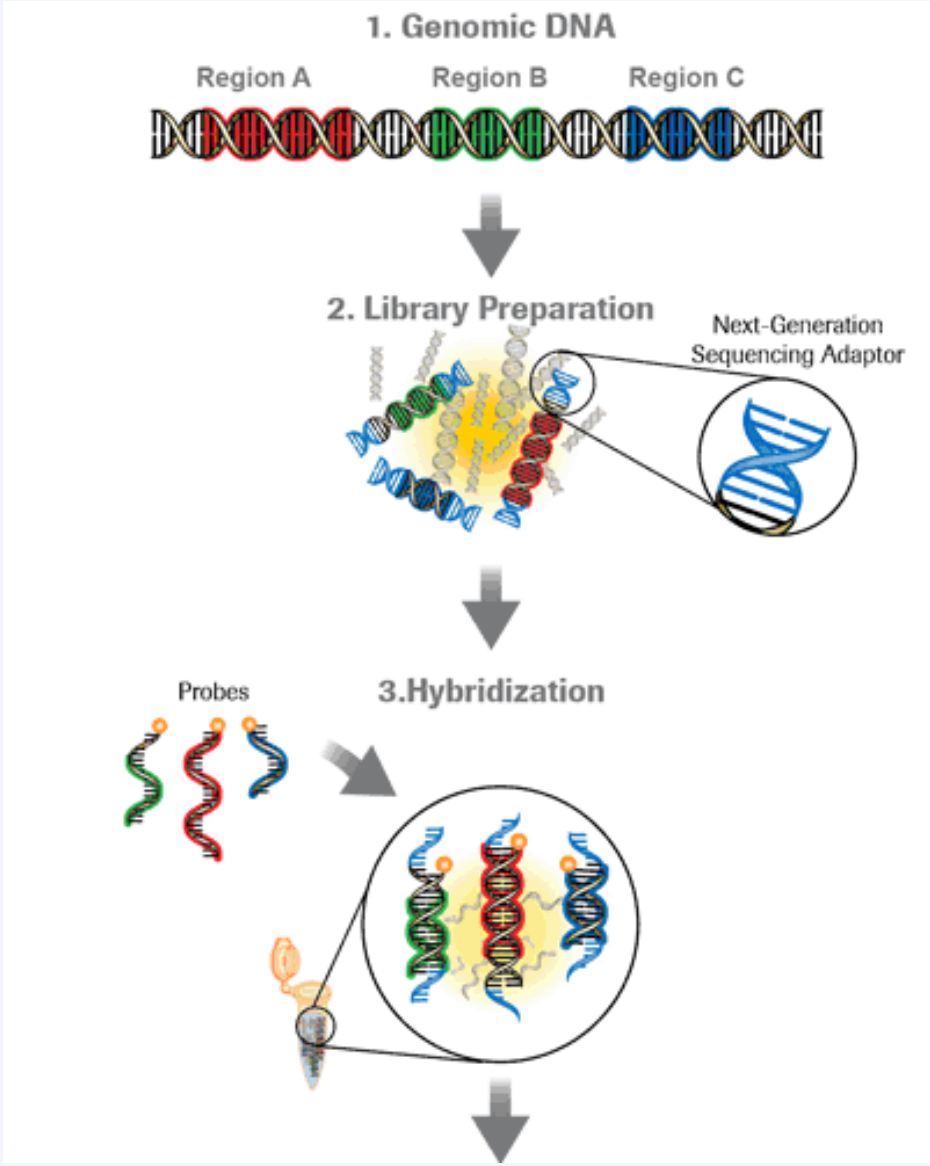
End-repair

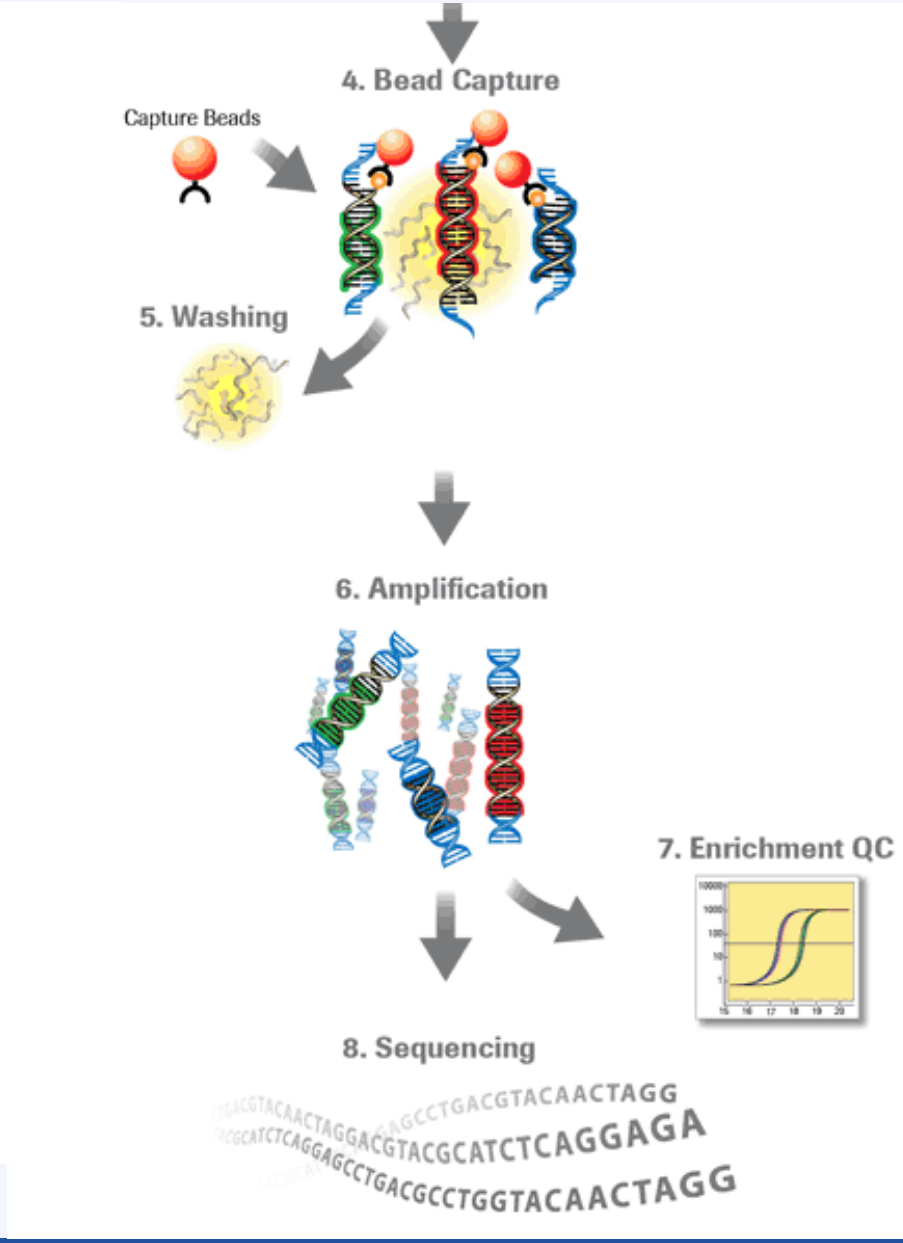
A-tailing, adapter ligation and PCR

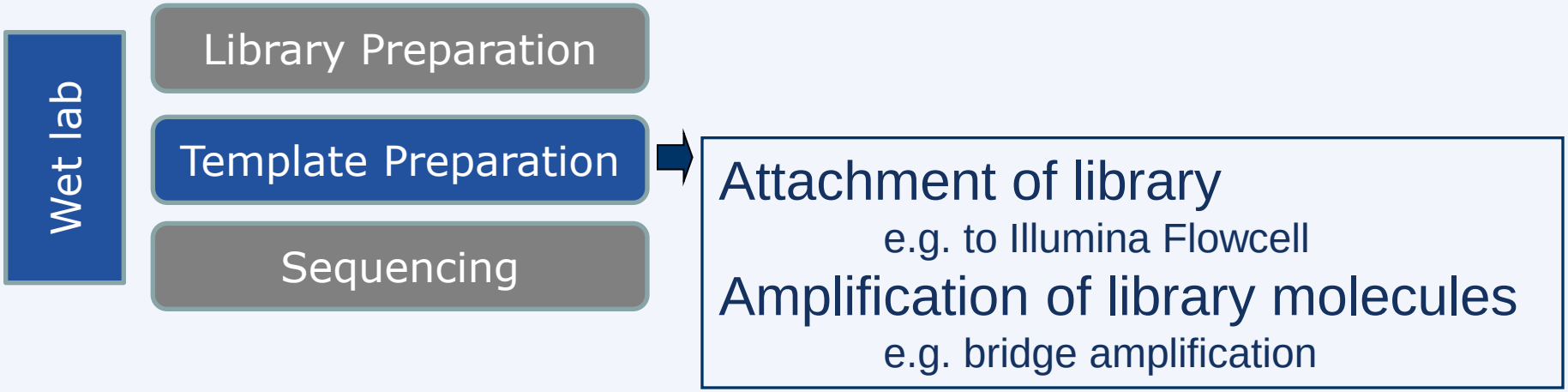
Final library contains

- sample insert
- indices (barcodes)
- flowcell binding sequences
- primer binding sequences

Hybridization capture/ Enrichment

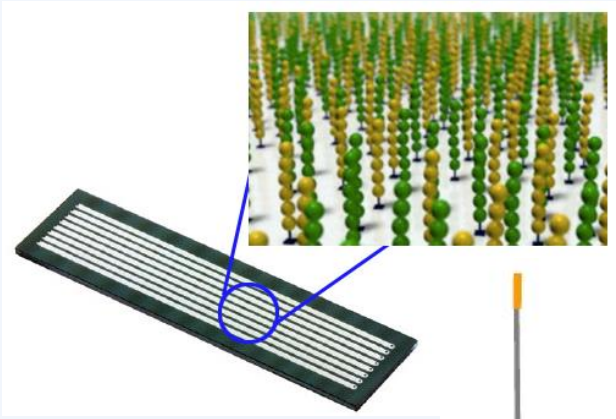


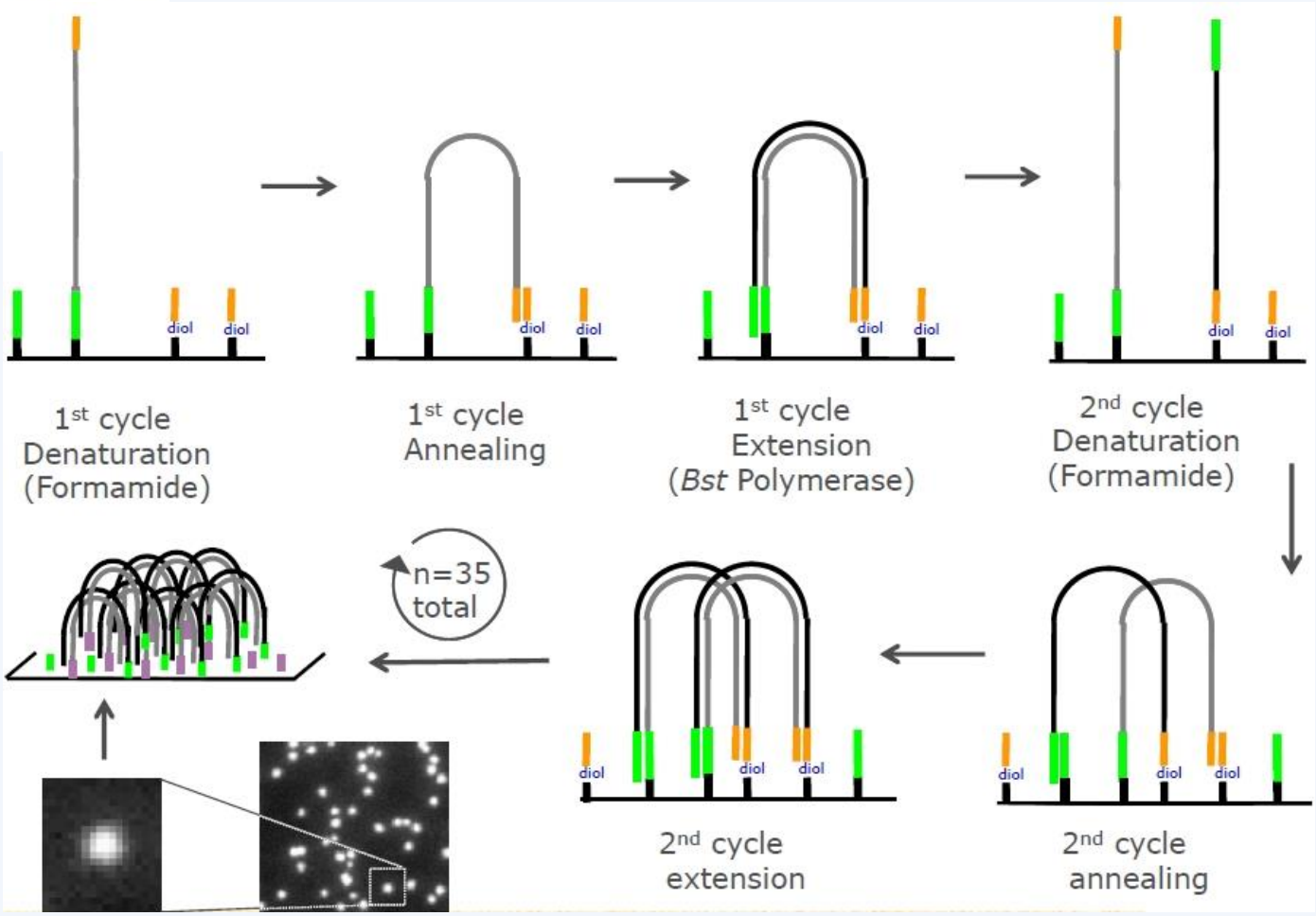




SA PATHOLOGY

Template Preparation





Wet lab

Library Preparation

Template Preparation

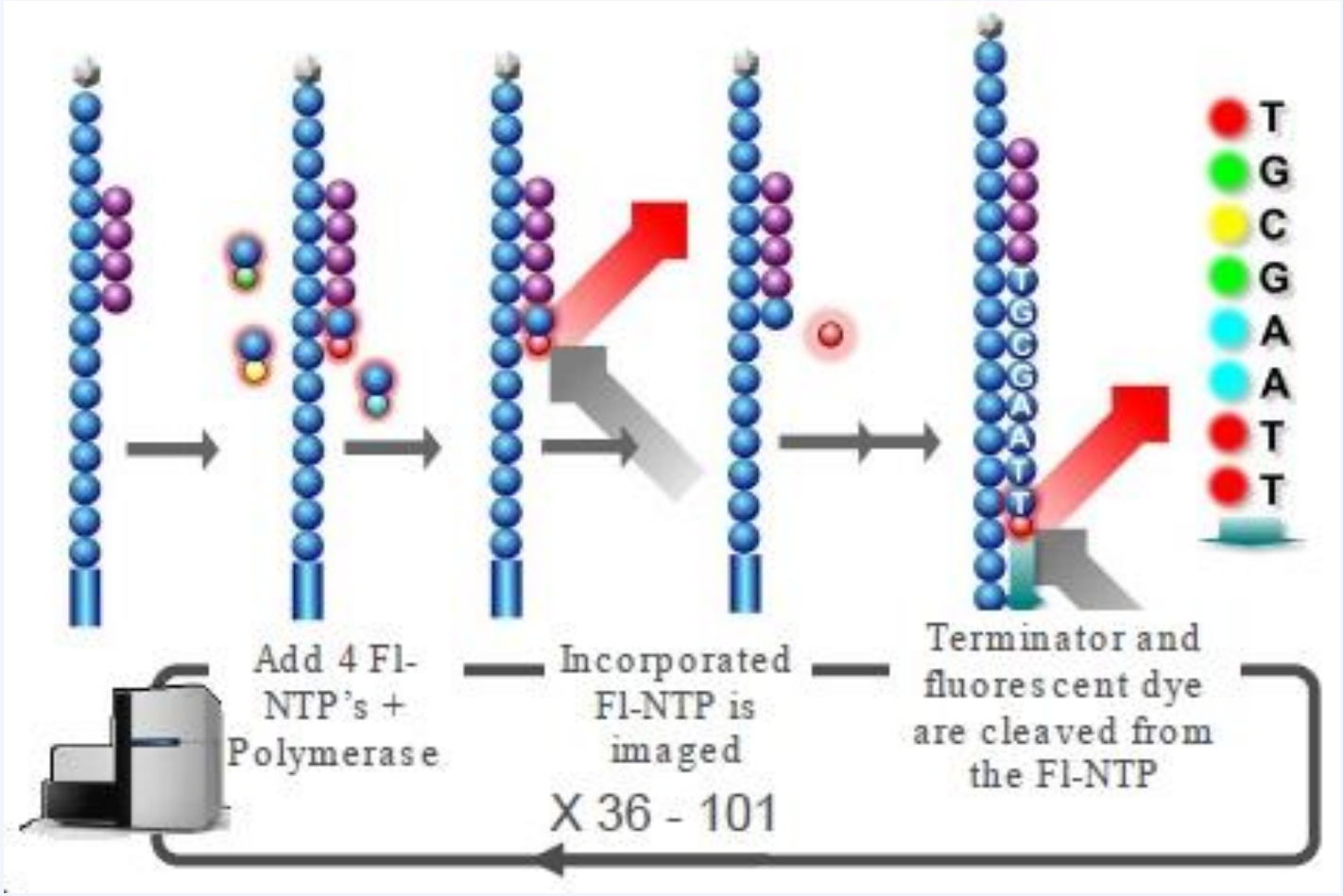
Sequencing

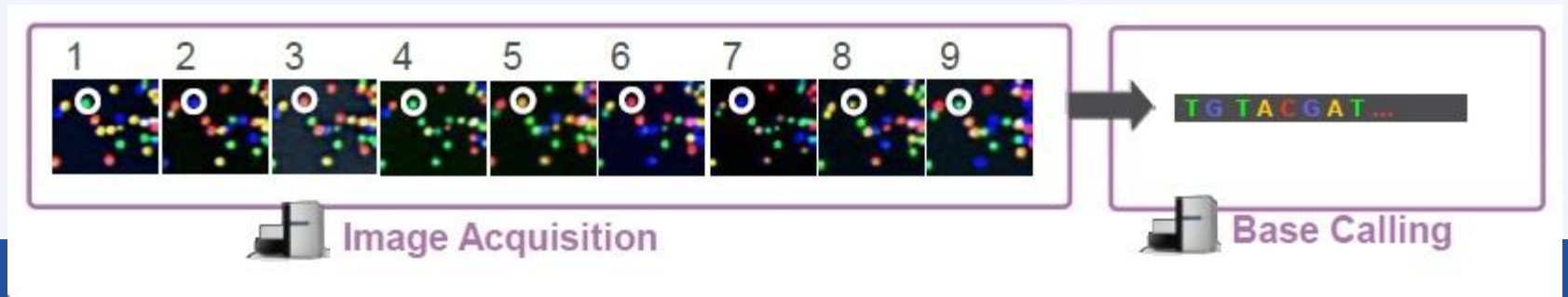
Sequencing-by-Synthesis
Detection by:

- Illumina – fluorescence
- Ion Torrent – pH
- Roche/454 – PO4 and light

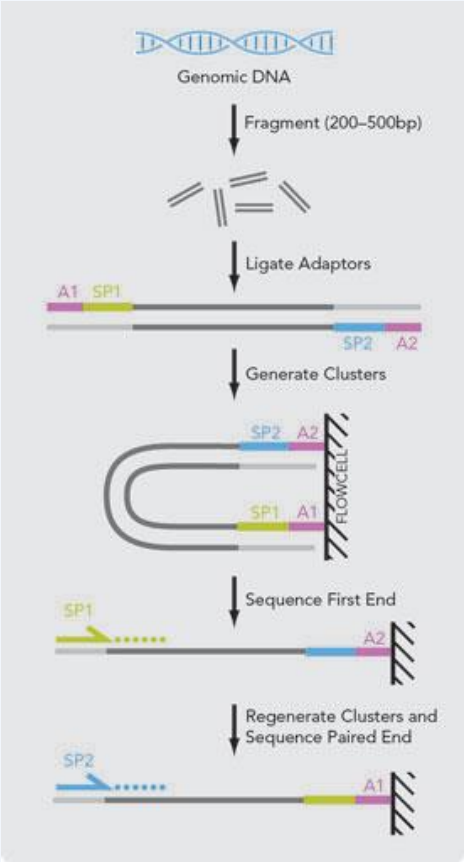
SA PATHOLOGY

Sequencing-by-Synthesis





Important sequencing concepts



- Barcoding/Indexing: allows multiplexing of different samples
- Single-end vs paired-end sequencing
- Coverage: avg. number reads per target
- Quality scores (Qscore): log-scales!

Quality Score	Probability of a wrong base call	Accuracy of a base call
Q 10	1 in 10	90%
Q 20	1 in 100	99%
Q 30	**1 in 1000**	**99.90%**
Q 40	1 in 10000	99.99%
Q 50	1 in 100000	100.00%

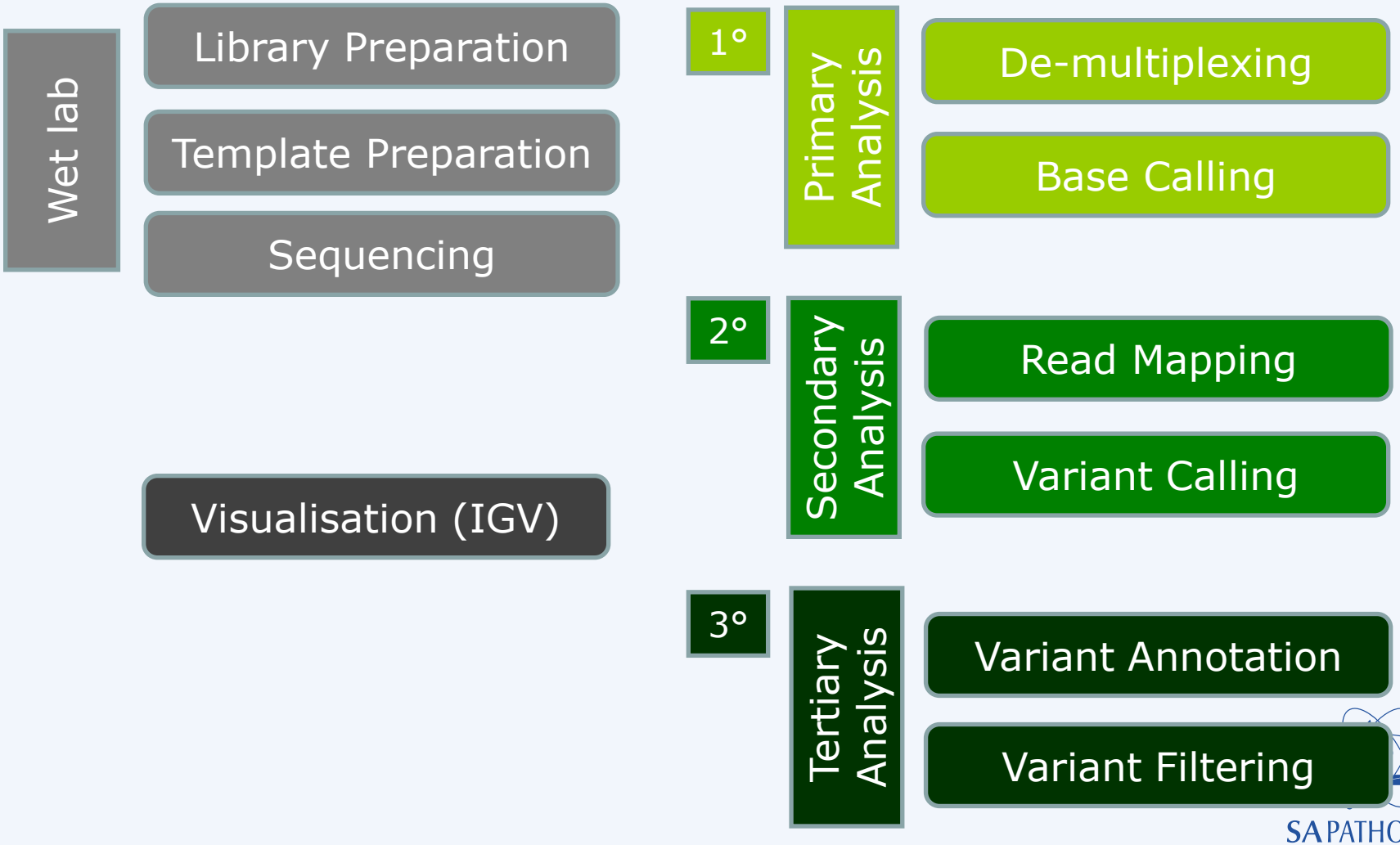
fragment ==========

Single read ------->

Paired-end reads R1-------> <------R2

SA PATHOLOGY

NGS workflow overview



SA PATHOLOGY

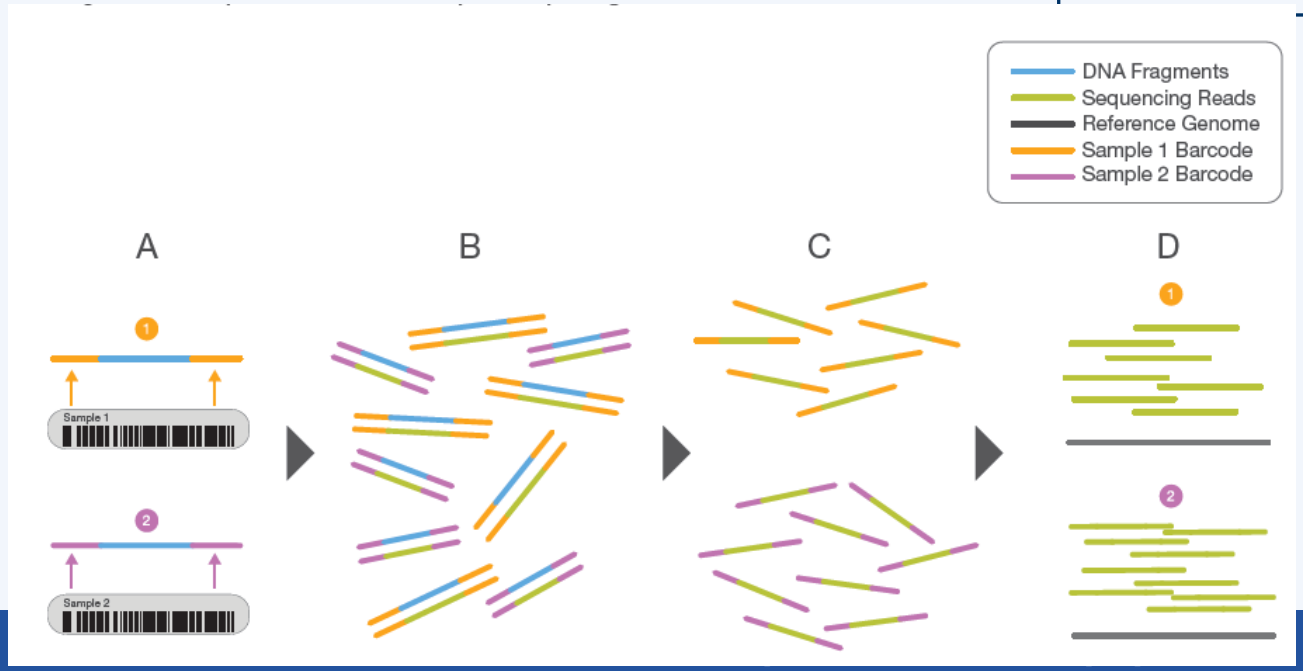
NGS workflow overview

1°

Primary Analysis

- De-multiplexing
- Base Calling

bcl2fastq software
Re-identifies samples
Error sources:
- Cross-talk (between bases)
- Phasing (incomplete removal of terminators)



SA PATHOLOGY

.FASTQ file format

Read Identifier	`@D3NZ4HQ1:111:D2DM2ACXX:1:1101:1243:2110 2:N:0:TGACCA`
Sequence	`GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTC`
+	`+`
Error probability (quality)	`!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CC`

!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJ

Phred score 0.....41

Probability 1.....0.0001

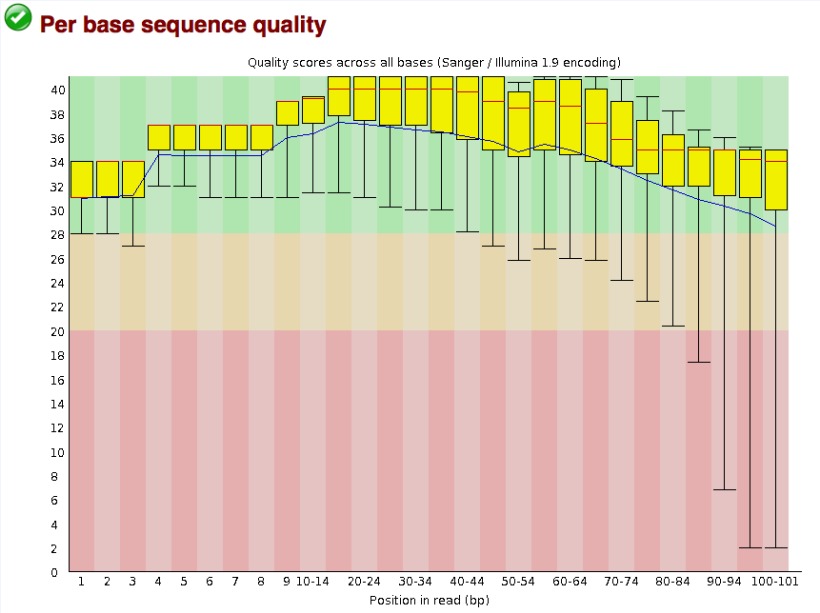
$$Phred\ score = -10\ log_{10}\ P$$

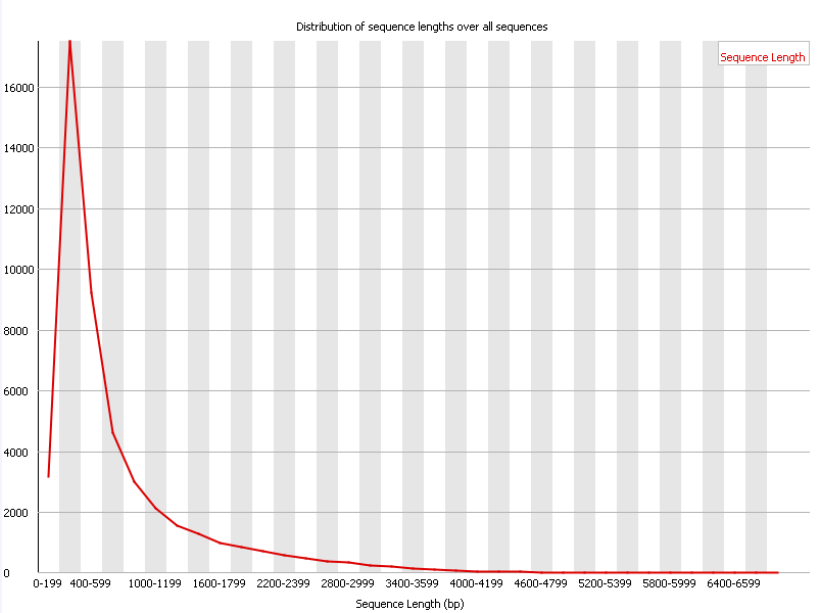
see en.wikipedia.org/wiki/FASTQ_format

Andreas Schreiber

SA PATHOLOGY

Sequence quality: fastQC





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

SA PATHOLOGY

2°

Secondary Analysis

Read Mapping

Variant Calling

Comparison against reference genome

(! not assembly !)

Many aligners

(short reads, longer reads, RNASeq…)

SAM/BAM files

SA PATHOLOGY

Burrows-Wheeler Alignment tool (BWA)

Popular tool for genomic sequence data (not RNASeq!)

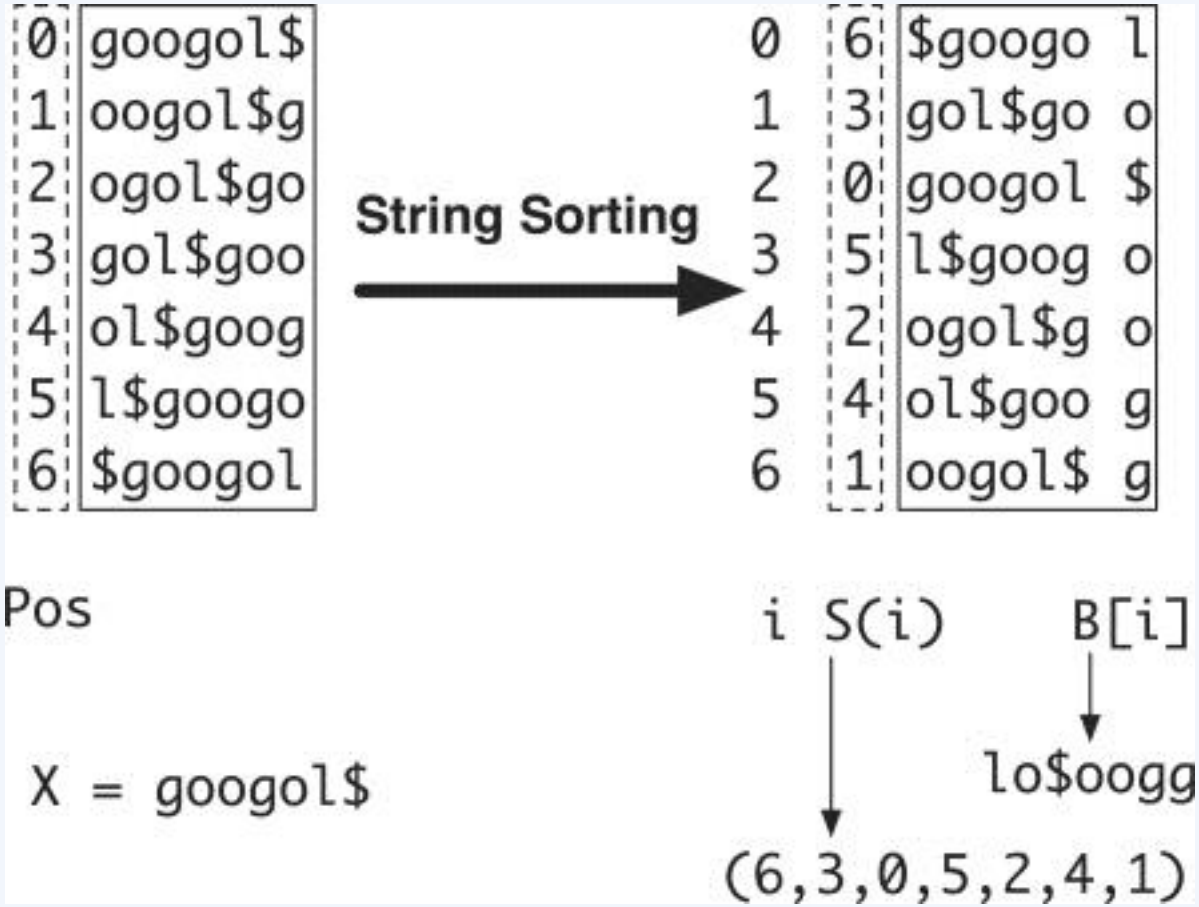
Li and Durbin 2009 Bioinformatics

Challenge:
compare billion of short sequence reads (.fastq file) against human genome (3Gb)

→ Uses Burrows-Wheeler Transform
"index" the human genome to allow ***memory-efficient*** and ***fast string matching*** between sequence read and reference genome

SA PATHOLOGY

Burrows-Wheeler Transform



SA PATHOLOGY

SAM/BAM files

Suppose we have the following alignment with bases in lower cases clipped from the alignment. Read `r001/1` and `r001/2` constitute a read pair; `r003` is a chimeric read; `r004` represents a split alignment.

```
Coor      12345678901234  5678901234567890123456789012345
ref       AGCATGTTAGATAA**GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT

+r001/1         TTAGATAAAGGATA*CTG
+r002         aaaAGATAA*GGATA
+r003       gcctaAGCTAA
+r004                      ATAGCT..............TCAGC
-r003                            ttagctTAGGC
-r001/2                                        CAGCGGCAT
```

The corresponding SAM format is:

```
@HD VN:1.5 SO:coordinate
@SQ SN:ref LN:45
r001   163 ref   7 30 8M2I4M1D3M = 37   39 TTAGATAAAGGATACTG *
r002     0 ref   9 30 3S6M1P1I4M *   0    0 AAAAGATAAGGATA     *
r003     0 ref   9 30 5S6M       *   0    0 GCCTAAGCTAA        * SA:Z:ref,29,-,6H5M,17,0;
r004     0 ref  16 30 6M14N5M    *   0    0 ATAGCTTCAGC        *
r003  2064 ref  29 17 6H5M       *   0    0 TAGGC              * SA:Z:ref,9,+,5S6M,30,1;
r001    83 ref  37 30 9M         =   7  -39 CAGCGGCAT          * NM:i:1
```

http://samtools.sourceforge.net/SAMv1.pdf

SAM/BAM files

@ Header (information regarding reference genome, alignment method…)

Read_ID **FLAG_field(first/second read in pair, both reads mapped…)**

ReferenceSequence Position(coordinate) MapQuality

CIGAR(describes alignment – matches, skipped regions, insertions..)

ReferenceSequence(Pair) Position(Pair) ReadSequence QUAL other…

```
@HD VN:1.5 SO:coordinate
@SQ SN:ref LN:45
r001  163 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG *
r002    0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA    *
r003    0 ref  9 30 5S6M       *  0   0 GCCTAAGCTAA       * SA:Z:ref,29,-,6H5M,17,0;
r004    0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC       *
r003 2064 ref 29 17 6H5M       *  0   0 TAGGC             * SA:Z:ref,9,+,5S6M,30,1;
r001   83 ref 37 30 9M         =  7 -39 CAGCGGCAT         * NM:i:1
```

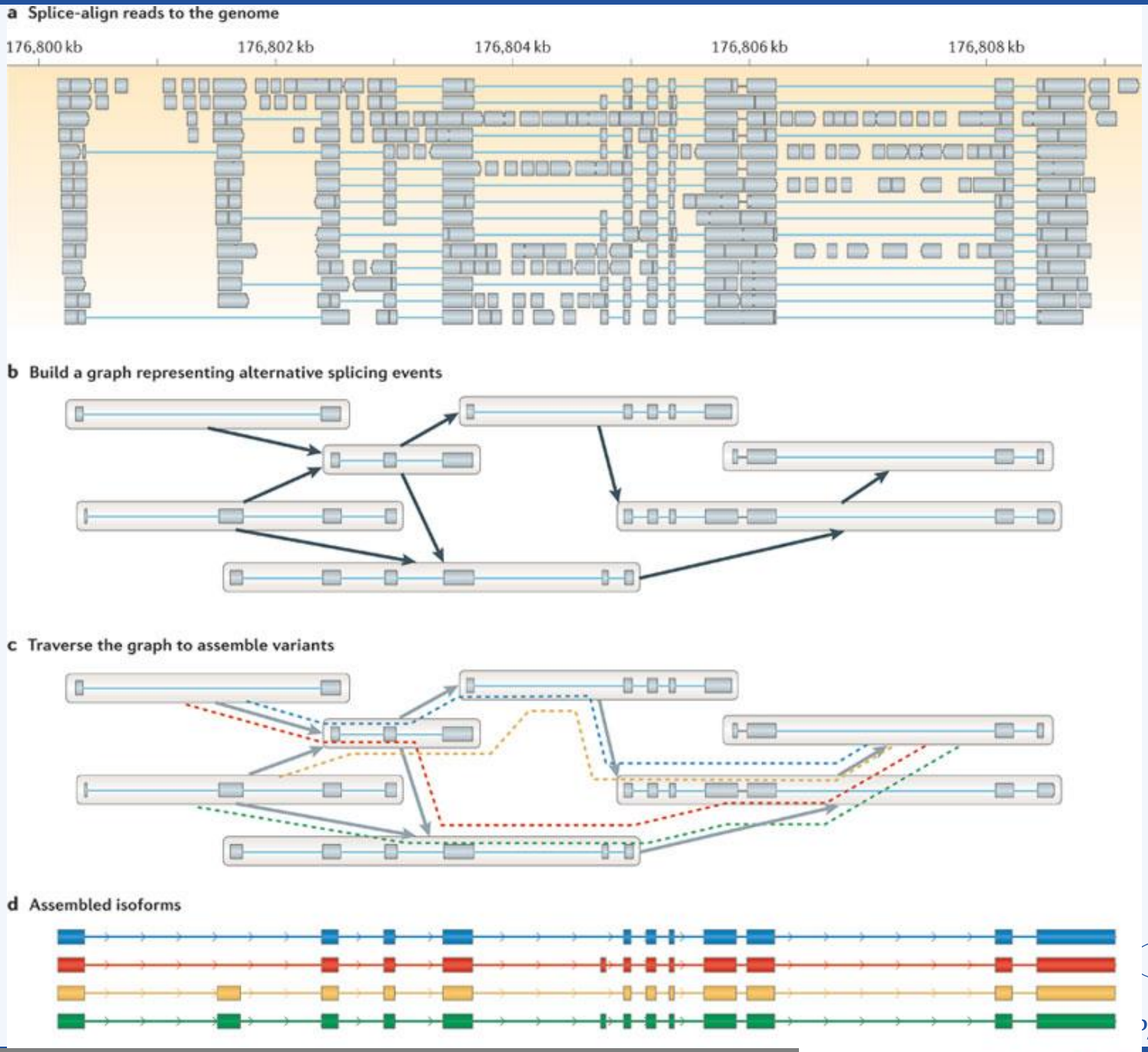
http://samtools.sourceforge.net/SAMv1.pdf

RNA-Seq mapping

Needs to account for splice-junctions and introns!
Needs to consider alternative splicing

- *de novo* transcript assembly (Abyss, Trinity…)
- reference-based approaches (TopHat, RNAMate…)
- Combined
- Feature-based approaches (Alexa-Seq)

SA PATHOLOGY



Martin and Wang 2011 Nature Reviews Genetics

PATHOLOGY

2°

Secondary Analysis

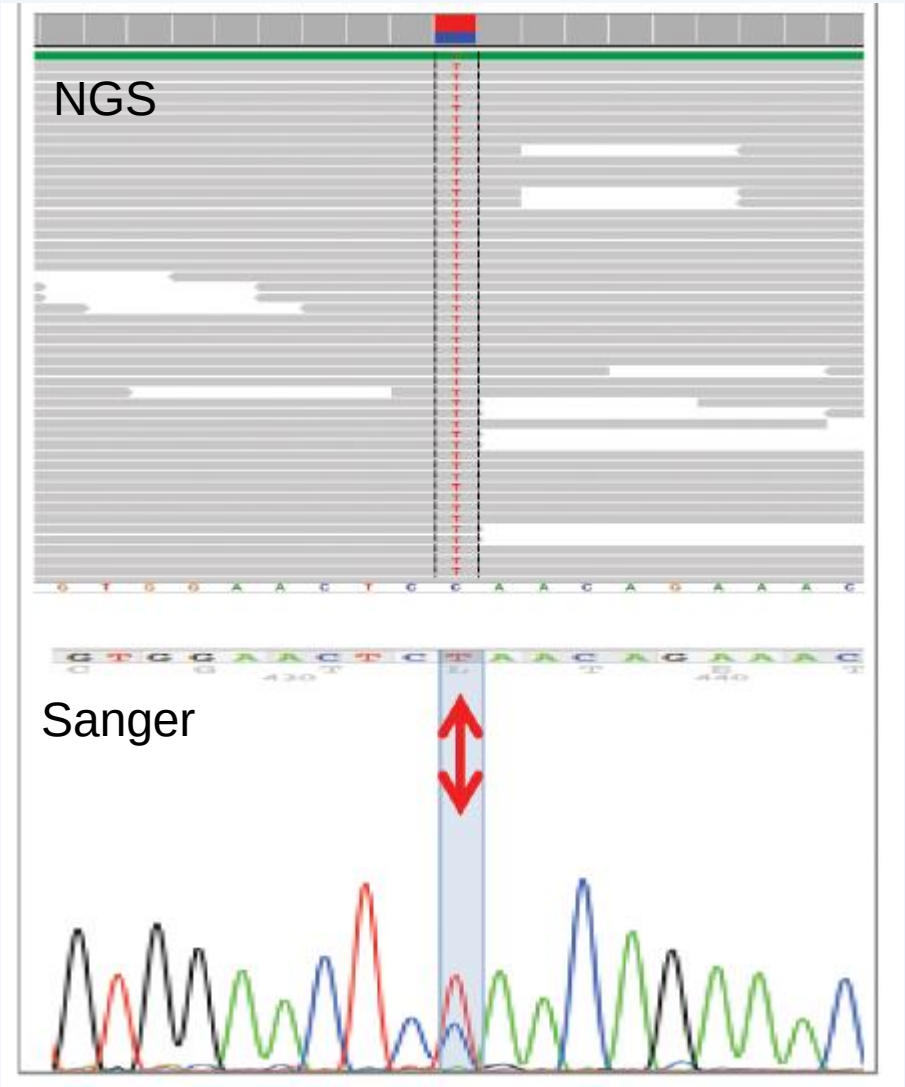
Read Mapping

Variant Calling

Identify sequence variants
Distinguish signal vs noise
VCF files

SA PATHOLOGY

Sequence variants



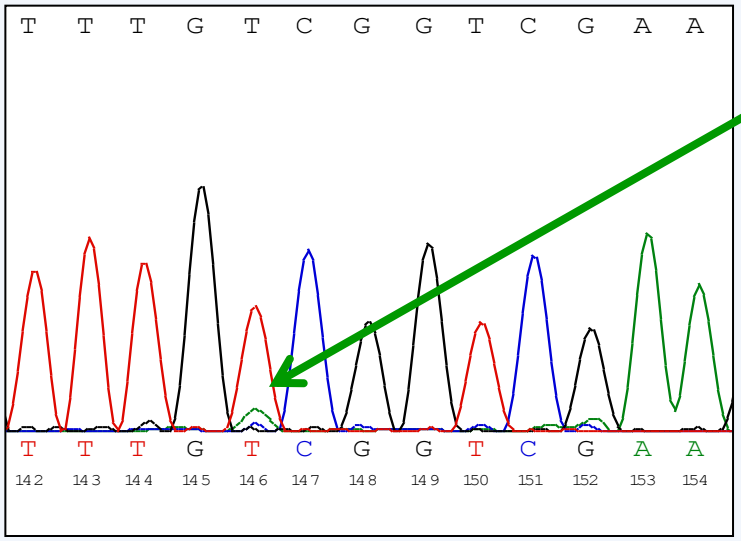
Differences to the reference

Reference: C

Sample: C/T

SA PATHOLOGY

Signal vs Noise



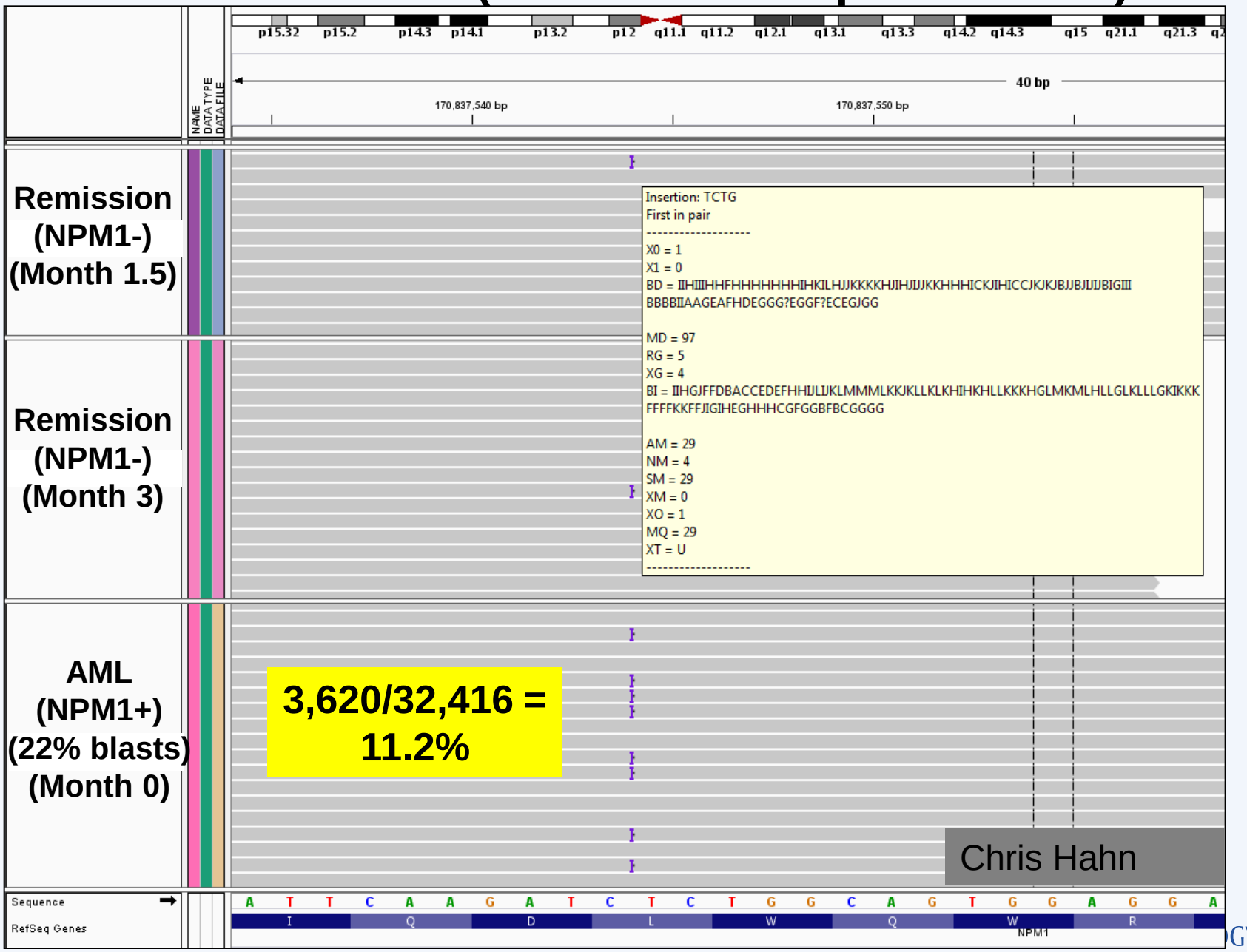
Sanger: is it real??

Total count: 204
A : 185 (91%, 92+, 93-)
C : 0
G : 1 (0%, 0+, 1-)
T : 18 (9%, 12+, 6-)
N : 0

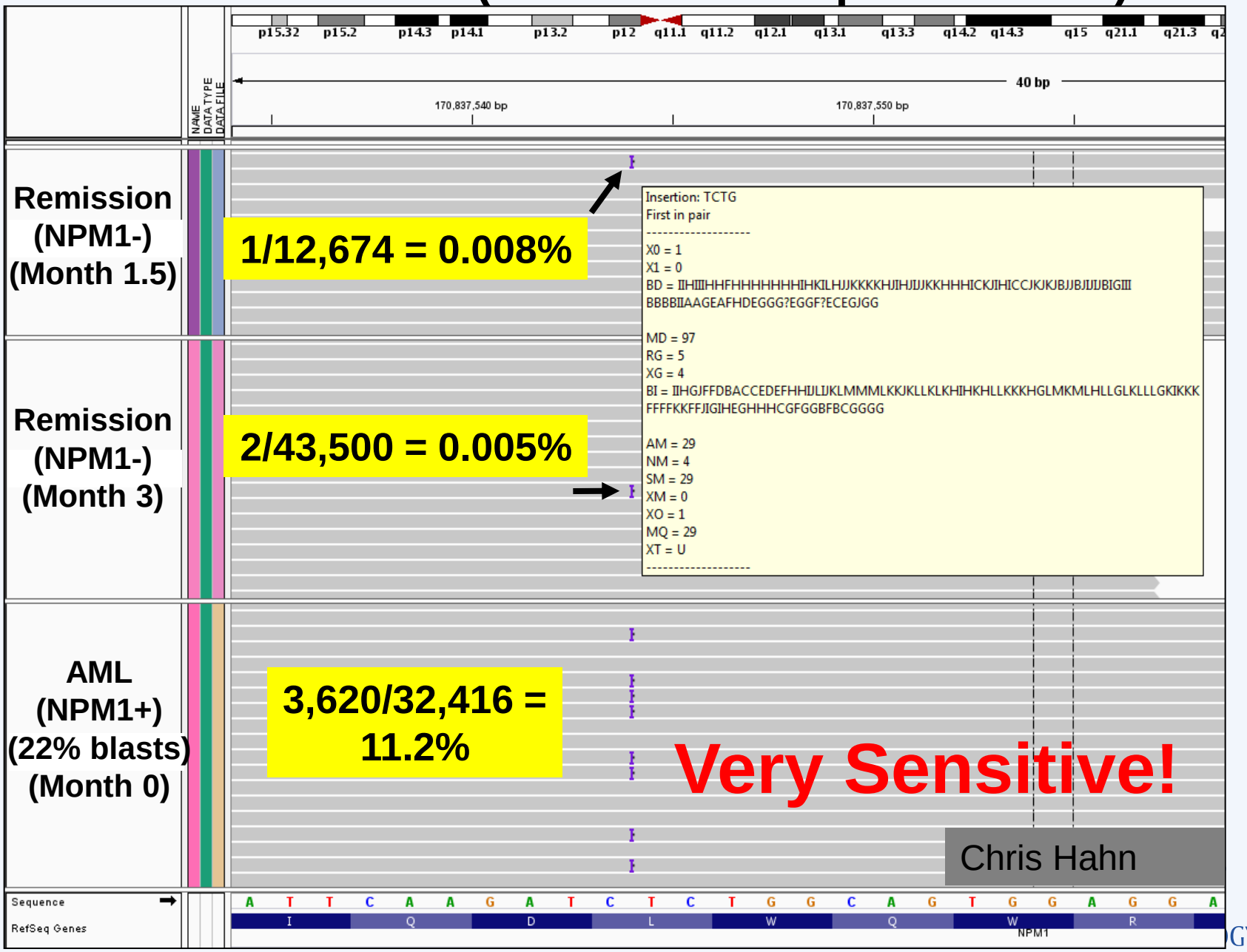
NGS: read count
Provides confidence (statistics!)
Sensitivity tune-able parameter (dependent on coverage)

SA PATHOLOGY

AML Patient NPM1 (TCTG insert → p.W288fs*12)



AML Patient NPM1 (TCTG insert → p.W288fs*12)



The GATK software

Genome Analysis Toolkit, BROAD Institute
http://www.broadinstitute.org/gatk/

- Initially developed for 1000 Genomes Project
- Single or multiple sample analysis (cohort)
- Popular tool for **germline variant calling**

SA PATHOLOGY

Unified Genotyper (GATK)

Bayesian genotype likelihood model

Evaluates probability of genotype given read data

McKenna *et al.* Genome Research 2010

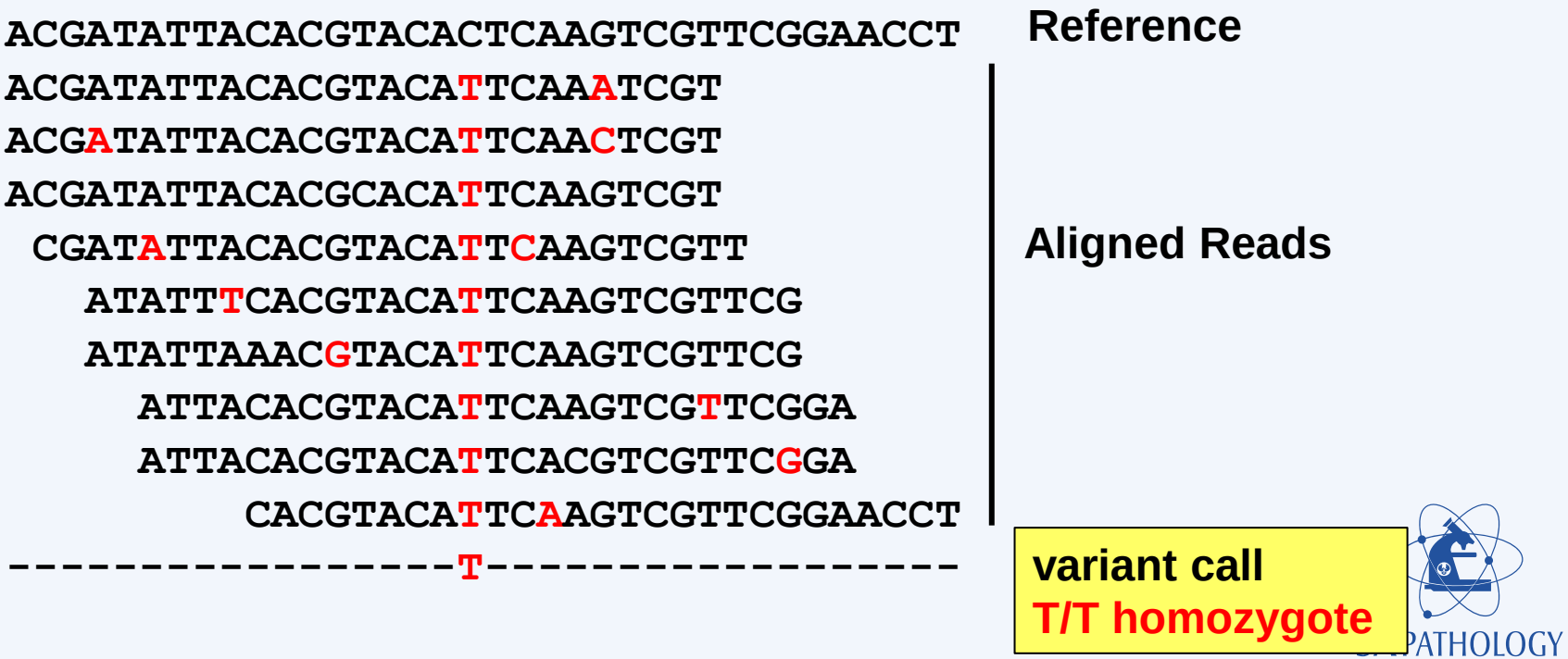
SA PATHOLOGY

Unified Genotyper (GATK)

Bayesian genotype likelihood model

Evaluates probability of genotype given read data

McKenna *et al.* Genome Research 2010



Somatic Variant Calling

- Somatic mutations can occur at **low freq. (<10%)** due to:
  - Tumor heterogeneity (multiple clones)
  - Low tumor purity (% normal cells in tumor sample)
- Requires different thresholds than germline variant calling when evaluating signal vs noise
- Trade-off between sensitivity (ability to detect mutation) and specificity (rate of false positives)

SA PATHOLOGY

Variant calling - indels

Small insertions/ deletions

The trouble with mapping approaches

coor	12345678901234	5678901234567890123456
ref	aggttttataaaac	----aattaagtctacagagcaacta
sample	aggttttataaaac	AAATaattaagtctacagagcaacta

Insertion AAAT in our sample!!!

modified from Heng Li (Broad Institute)

SA PATHOLOGY

Variant calling - indels

Small insertions/ deletions

The trouble with mapping approaches

```
coor      12345678901234     567890123456 7890123456
ref       aggttttataaaac----aattaagtctacagagcaacta
sample    aggttttataaaacAAATaattaagtctacagagcaacta
read1     aggttttataaaac     aaAtaa
read2      ggttttataaaac     aaAtaaTt
read3          ttataaaacAAATaattaagtctaca
read4              CaaaT     aattaagtctacagagc
read5                aaT     aattaagtctacagagc
read6                  T     aattaagtctacagagc
```

By default, aligners prefer placing reads w/ a mismatch than with an insertion, esp. at ends of read!!

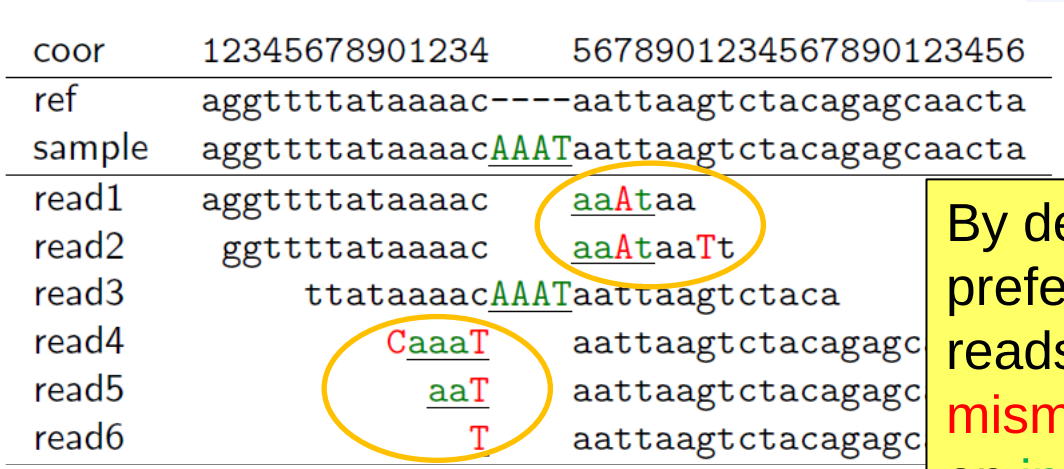
modified from Heng Li (Broad Institute)

SA PATHOLOGY

Variant calling - indels

Small insertions/ deletions

The trouble with mapping approaches



By default, aligners prefer placing reads w/ a mismatch than with an insertion, esp. at ends of read!!

modified from Heng Li (Broad Institute)

SA PATHOLOGY

Variant calling - indels

Small insertions/ deletions

The trouble with mapping approaches

```
coor      12345678901234     567890123456789012345678
ref       aggttttataaaac----aattaagtctacagagcaacta
sample    aggttttataaaacAAATaattaagtctacagagcaacta
read1     aggttttataaaac    aaAtaa
read2      ggttttataaaac    aaAtaaTt
read3          ttataaaacAAATaattaagtctaca
read4              CaaaT    aattaagtctacagagcaac
read5                aaT    aattaagtctacagagcaact
read6                 T     aattaagtctacagagcaacta
read1     aggttttataaaacaaataa
read2      ggttttataaaacaaataatt
read3          ttataaaacaaataattaagtctaca
read4                  caaataattaagtctacag
read5                    aataattaagtctacag
read6                     taattaagtctacag
```

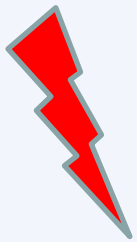
Information from other reads can be used to improve alignment; After local realignment the insertion has been correctly placed!!

modified from Heng Li (Bro

Variant calling - indels

Small insertions/ deletions

The trouble with mapping approaches



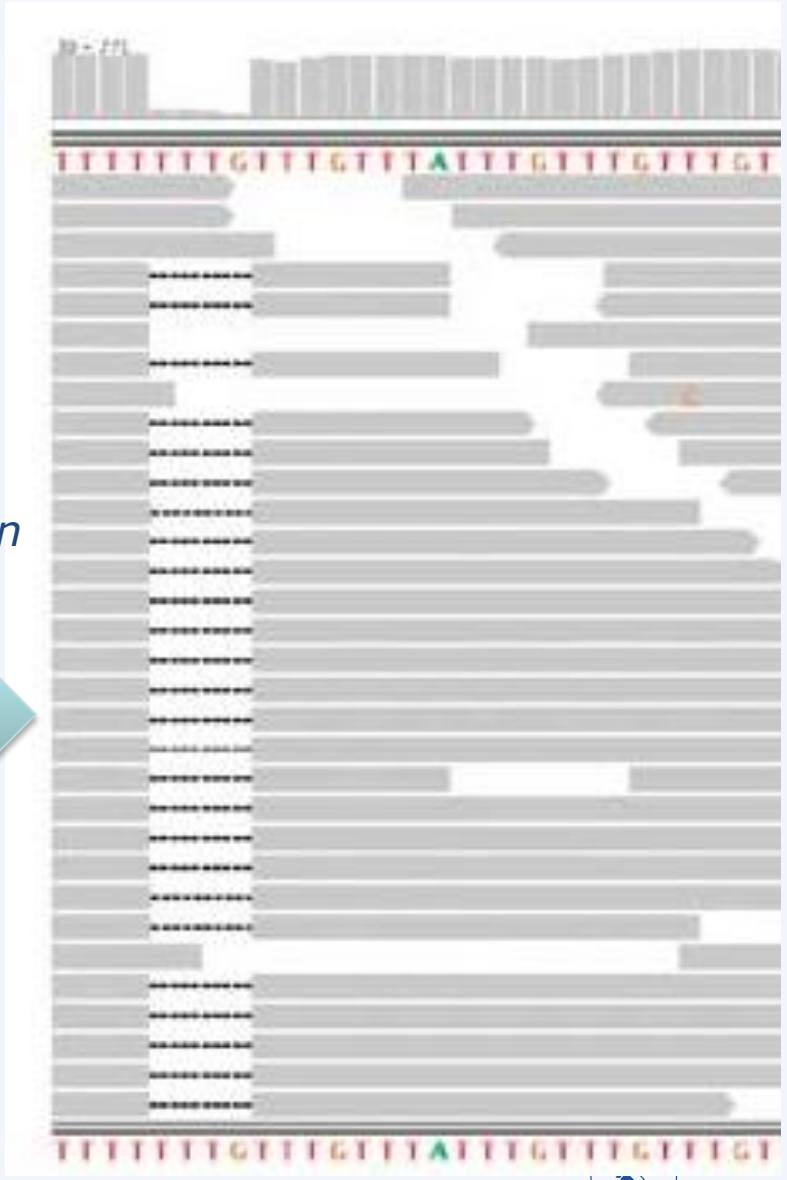
```
coor      12345678901234     567890123456789012345678
ref       aggttttataaaac----aattaagtctacagagcaacta
sample    aggttttataaaacAAATaattaagtctacagagcaacta
read1     aggttttataaaac    aaAtaa
read2      ggttttataaaac    aaAtaaTt
read3          ttataaaacAAATaattaagtctaca
read4              CaaaT    aattaagtctacagagcaac
read5                aaT    aattaagtctacagagcaact
read6                  T    aattaagtctacagagcaacta
read1     aggttttataaaacaaataa
read2      ggttttataaaacaaataatt
read3          ttataaaacaaataattaagtctaca
read4                caaataattaagtctacagagcaac
read5                  aataattaagtctacagagcaact
read6                    taattaagtctacagagcaacta
```

Improves indel calling

modified from Heng Li (Broad Institute)

SA PATHOLOGY

Re-align within multi-read context



Andreas Schreiber

SA PATHOLOGY

Local realignment in GATK

- Uses information from known SNPs/indels

  (dbSNP, 1000 Genomes)

- Uses information from other reads

- Smith-Waterman exhaustive alignment on select reads

SA PATHOLOGY

Evaluating Variant Quality

Consider:

- Coverage at position
- Number independent reads supporting variant
- Observed allele fraction vs expected (somatic,germline)
- Strand bias
- Base qualities at variant position
- Mapping qualities of reads supporting variant
- Variant position within reads (near ends or at centre)

SA PATHOLOGY

.VCF Files

Variant Call Format files

Standard for reporting variants from NGS

Describes metadata of analysis and variant calls

Text file format (open in Text Editor or Excel)

!!! Not a MS Office vCard !!!

http://www.1000genomes.org/wiki/Analysis/Variant%20Call%20Format/vcf-variant-call-format-version-41

SA PATHOLOGY

Example .VCF file

```
##fileformat=VCFv4.1
##fileDate=20090805
##source=myProgramV3
##reference=file:///seq/NCBI36.fasta
…
```

Header lines (marked by ##): Metadata of analysis

SA PATHOLOGY

Example .VCF file

Header lines
(marked by ##):
Metadata of analysis

```
##fileformat=VCFv4.1
##fileDate=20090805
##source=myProgramV3
##reference=file:///seq/NCBI36.fasta
…
```

#CHROM	POS	ID	REF	ALT	QUAL	FILTER
20	14370	rs6054257	G	A	29	PASS
20	17330	.	T	A	3	q10

INFO	FORMAT	SAMPLE1	…
NS=2;DP=14;AF=0.5;DB;H2	GT:GQ:DP	1|0:48:8	
NS=2;DP=11;AF=0.017	GT:GQ:DP	0|0:49:3	

Data lines:
Individual variant calls

SA PATHOLOGY

Example .VCF file

Header lines
(marked by ##):
Metadata of analysis

```
##fileformat=VCFv4.1
##fileDate=20090805
##source=myProgramV3
##reference=file:///seq/NCBI36.fasta
…
```

#CHROM	POS	ID	REF	ALT	QUAL	FILTER
20	14370	rs6054257	G	A	29	PASS
20	17330	.	T	A	3	q10

INFO	FORMAT	SAMPLE1	…
NS=2;DP=14;AF=0.5;DB;H2	GT:GQ:DP	1|0:48:8	
NS=2;DP=11;AF=0.017	GT:GQ:DP	0|0:49:3	

Data lines:
Individual variant calls

GT: genotype: 1|0 het, 0|0 hom
DP: read depth

3°

Tertiary Analysis

Variant Annotation

Variant Filtering

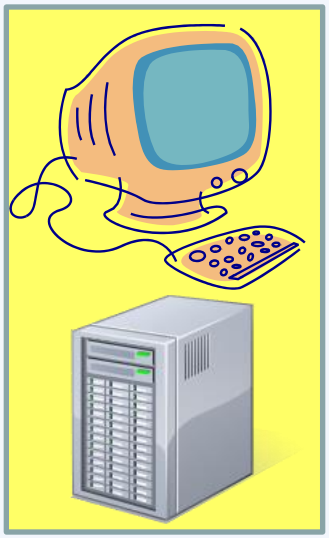
Provide biological/clinical context
Identify disease-causing mutation

(among thousands of variants)

SA PATHOLOGY

Annotation pipeline

Variant Calls
(.VCF file)
.....
.....
.....



Polymorphism DBs	Transcript Conseq.	Pathogenicity Pred.	Disease DBs
dbSNP	Ensembl VEP	PolyPhen	OMIM
1000 Genomes	snpEff	SIFT	HGMD
HapMap	ANNOVAR	Splice Site prediction	Gene Tests

Automated Annotation

PubMed

Annotated
Variant Calls
(.VCF file)
.....
.....
.....

Variant Filtering and Prioritization

Purpose:

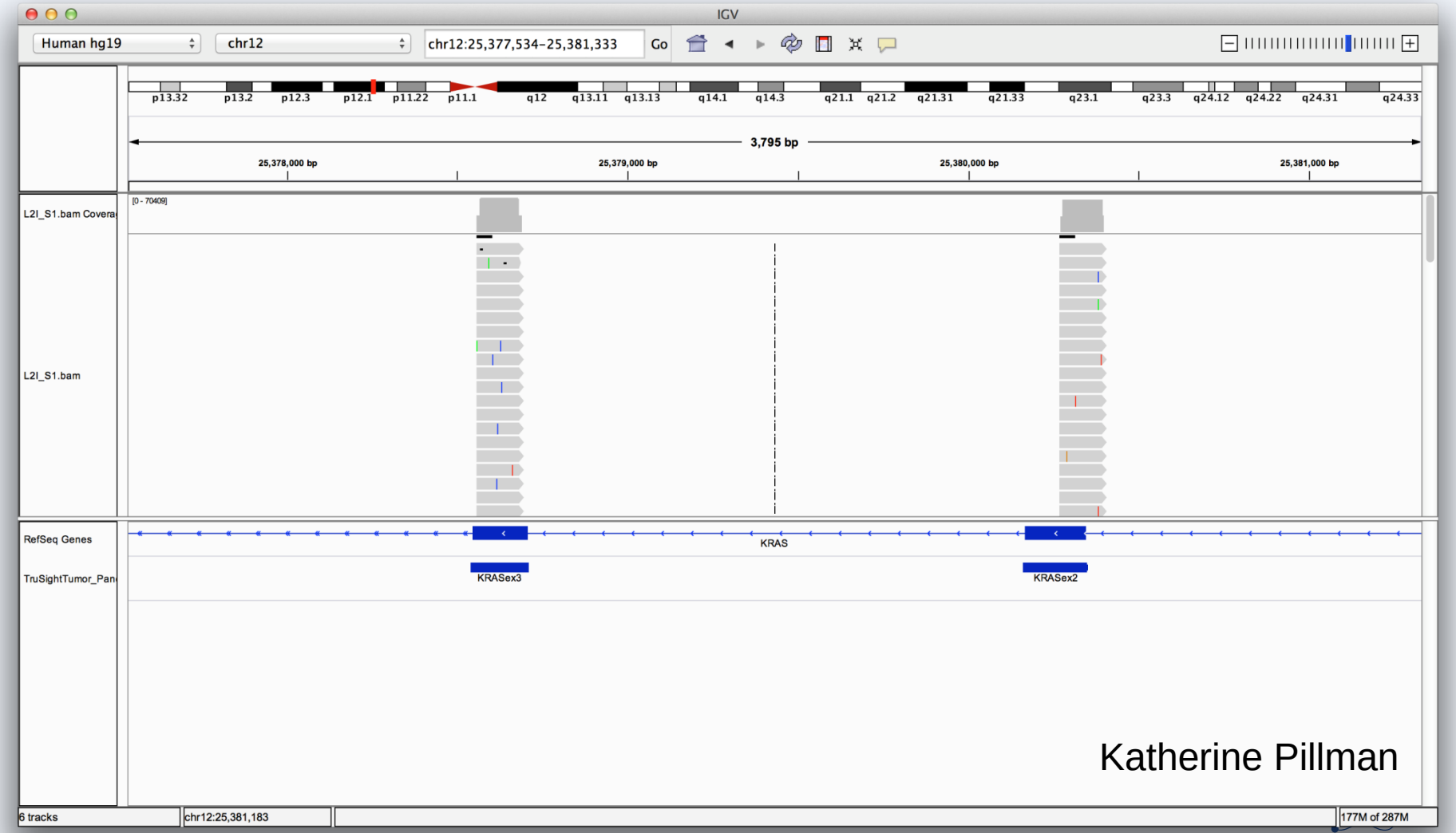
Identify pathogenic/disease-associated mutation(s)
Reduce candidate variants to reportable set

Common Steps:

1. Remove poor quality variant calls
2. Remove common polymorphisms
3. Prioritize variants with high functional impact
4. Compare against known disease genes
5. Consider mode of inheritance (autosomal recessive, X-linked…)
6. Segregation in family (where multiple samples avail.)

SA PATHOLOGY

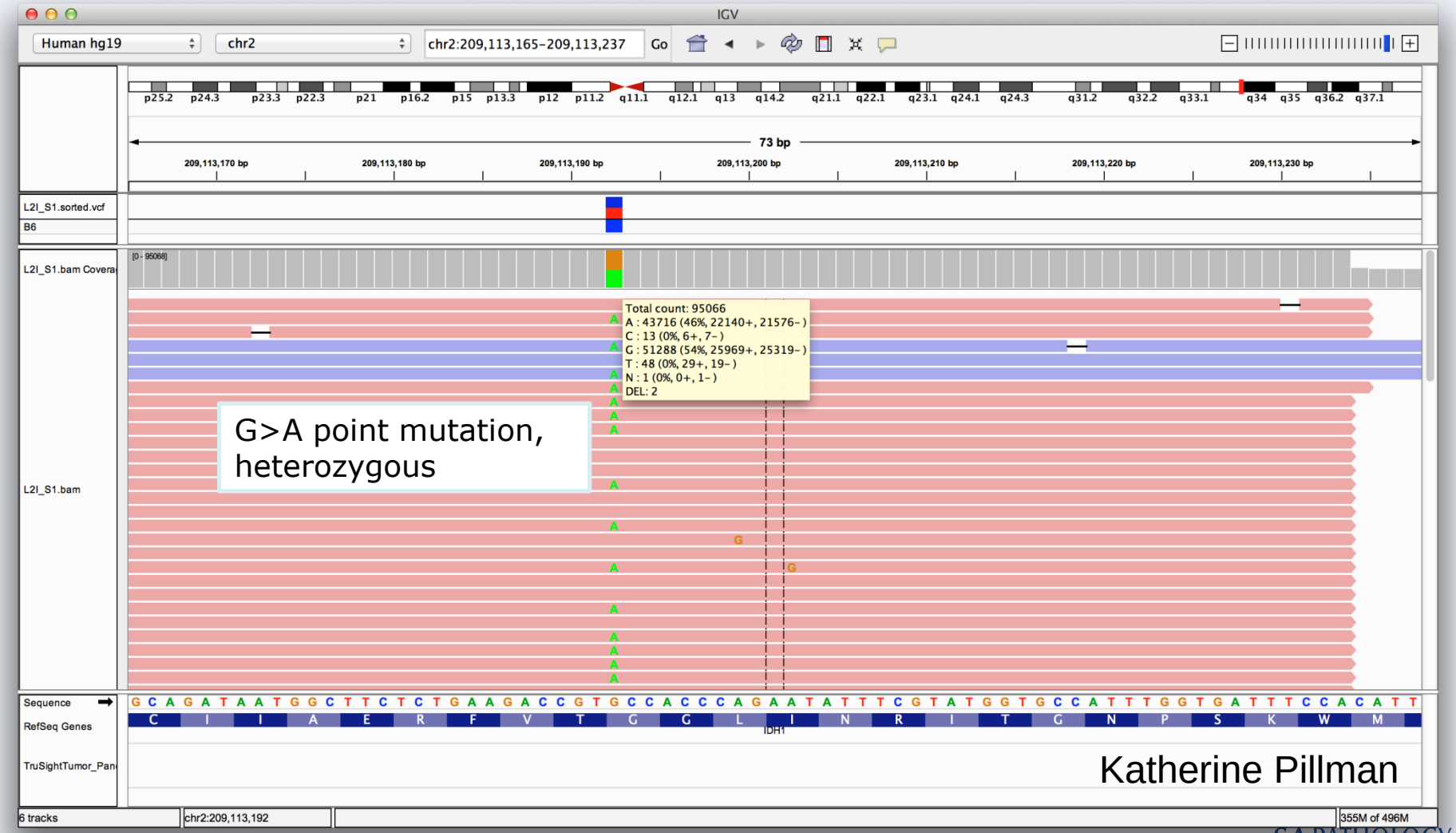
Visualisation (IGV)



SA PATHOLOGY

Integrative Genomics Viewer

http://www.broadinstitute.org/igv/



SA PATHOLOGY

“Common” pipeline

bcl2fastq (Illumina)
fastQC (open-source)

Exomes (HiSeq):
BWA (open-source), GATK (Broad)

Gene Panels (MiSeq, PGM)
MiSeq Reporter (Illumina)
Torrent Suite (Ion Torrent)

Custom scripts and third party tools
(Annovar, snpEff, PolyPhen, SIFT, ….)
Commercial annotation software
(Geneticist Assistant, VariantStudio…)

SA PATHOLOGY

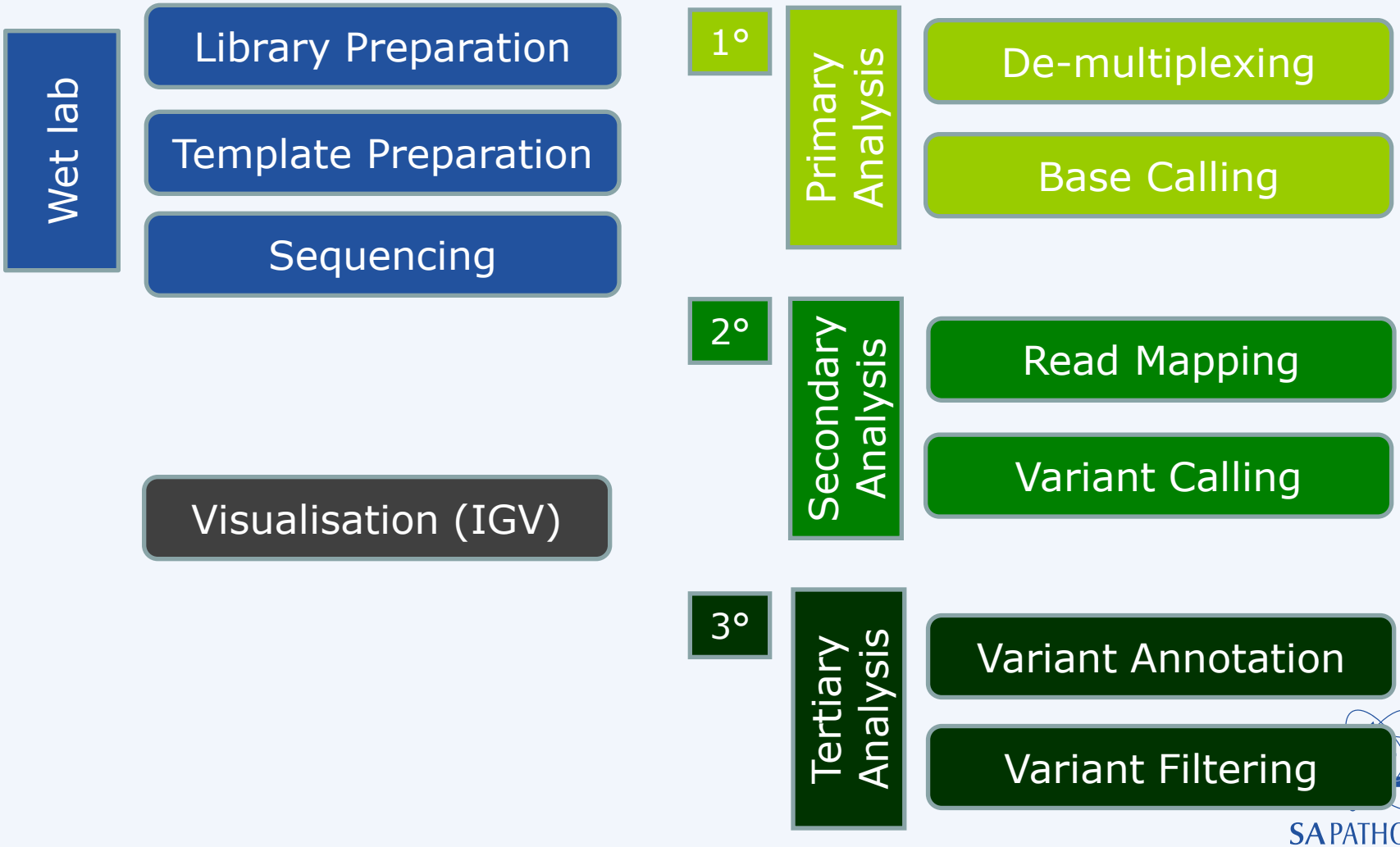
Common data formats

.bcl
.fastq

.BAM
.VCF

SA PATHOLOGY

NGS workflow overview



SA PATHOLOGY

Conclusions

- NGS data - the new currency of (molecular) biology
- **Broad applications** (ecology, evolution, ag sciences, medical research and clinical diagnostics…)
- **Rapidly evolving** (sequencing technologies, library preparation methods, analysis approaches, software)
- Bioinformatics pipelines typically combine vendor software, third-party tools and custom scripts
- Requires skills in **scripting, Linux/Unix, HPC**
- Understanding of data (SE, PE, RNA-Seq) important for successful analysis

SA PATHOLOGY

Thank you!

SA PATHOLOGY