ThermoFisher SCIENTIFIC

iontorrent

The Highs and Lows of Next Generation Sequencing

Dr Xosé M Fernández

Nottingham, 20th September 2016

The world leader in serving science

For Research Use Only. Not for use in diagnostic procedures

- Genetics
  - Genomics – Human Genome Project
- NGS Technologies
  - Different Techniques
  - Challenges of Interpretation
- Human Variation
- Precision Medicine and Tumour Analysis
  - Targeted Sequencing
  - Paradigm Shift – A Molecular Taxonomy
- Bioinformatics – A Helping Hand

Genetics Timeline

- **1866** G J Mendel discovered '*unit*' that conveyed information to offspring
- **1869** Johann Friedrich Miescher isolates *nuclein* (unknown function)
- **Chromosome Theory** (Sutton 1903) Genes lie on chromosomes.

- All DNA *identical* → Proteins were considered the genetic material due to their highly variable polymeric nature.

Genetics Timeline (2)

- **Griffith *et al.*** (1928) *Transforming principle (nuclein)* could be transferred to viable non-virulent bacteria
- **Oswald Avery** (1944) Substance from virulent strain could *transform* non-virulent bacteria
- **Hersey & Chase** (1952) The active component of the bacteriophage that transmits the infective characteristic is the **DNA**. There is a clear correlation between DNA and genetic information.
- **Erwin Chargaff** (1949) DNA composition species-specific (A=T; G=C)

- **Watson & Crick** (1953) Molecular structure of DNA

DNA – Genetic Blueprint

- Deoxyribonucleic acid (DNA)
- Located in the **nucleus**
- Wrapped up in structures called **chromosomes**
- Humans have **46** chromosomes (23 pairs in every cell)

What’s a Chromosome?

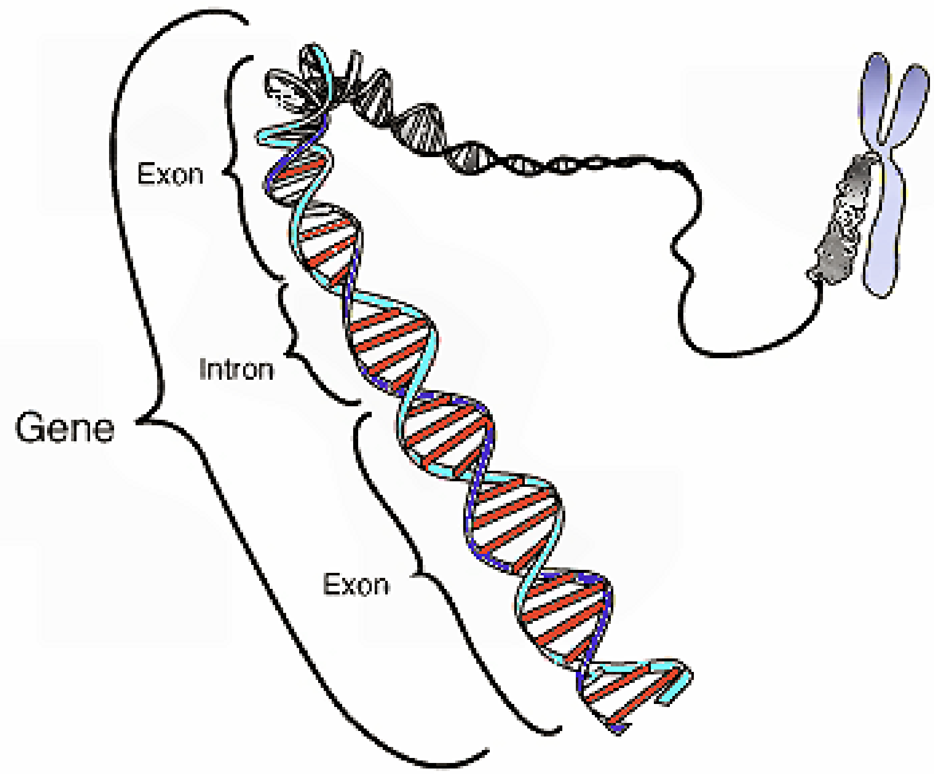
- A chromosome is DNA, which contains many genes, regulatory elements and other nucleotide sequences

- Cells may contain more than one type of chromosome (*e.g.* mitochondrial DNA, chloroplastin)

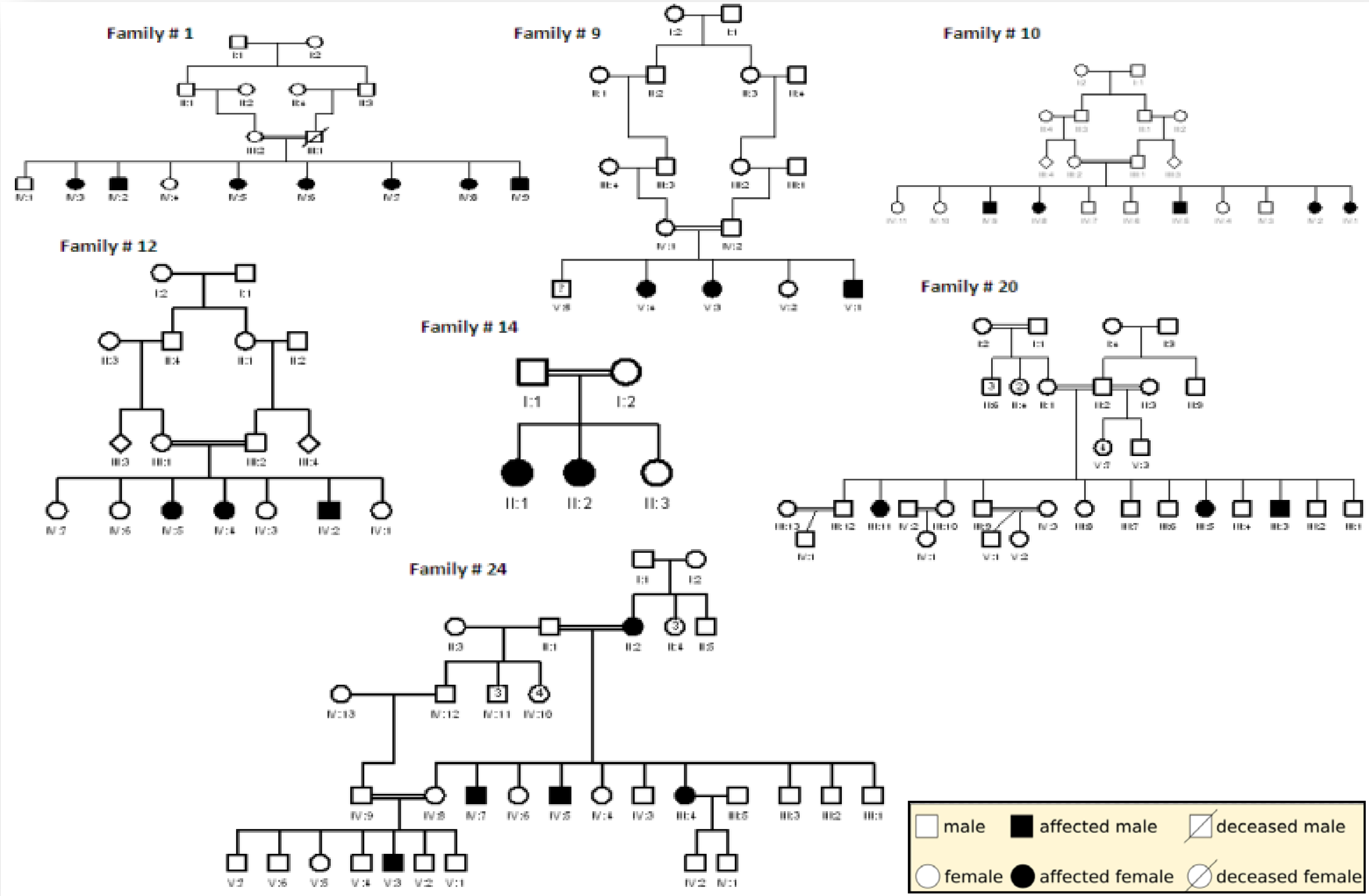
- In eukaryotes, nuclear chromosomes are packaged by proteins into a condensed structure called chromatin.

What’s a Gene?

- A gene is a stretch of DNA whose sequence determines the structure and function of a specific functional molecule (usually a protein)
- Not all the DNA codes for proteins
- 20,441 protein-coding genes in the human genome



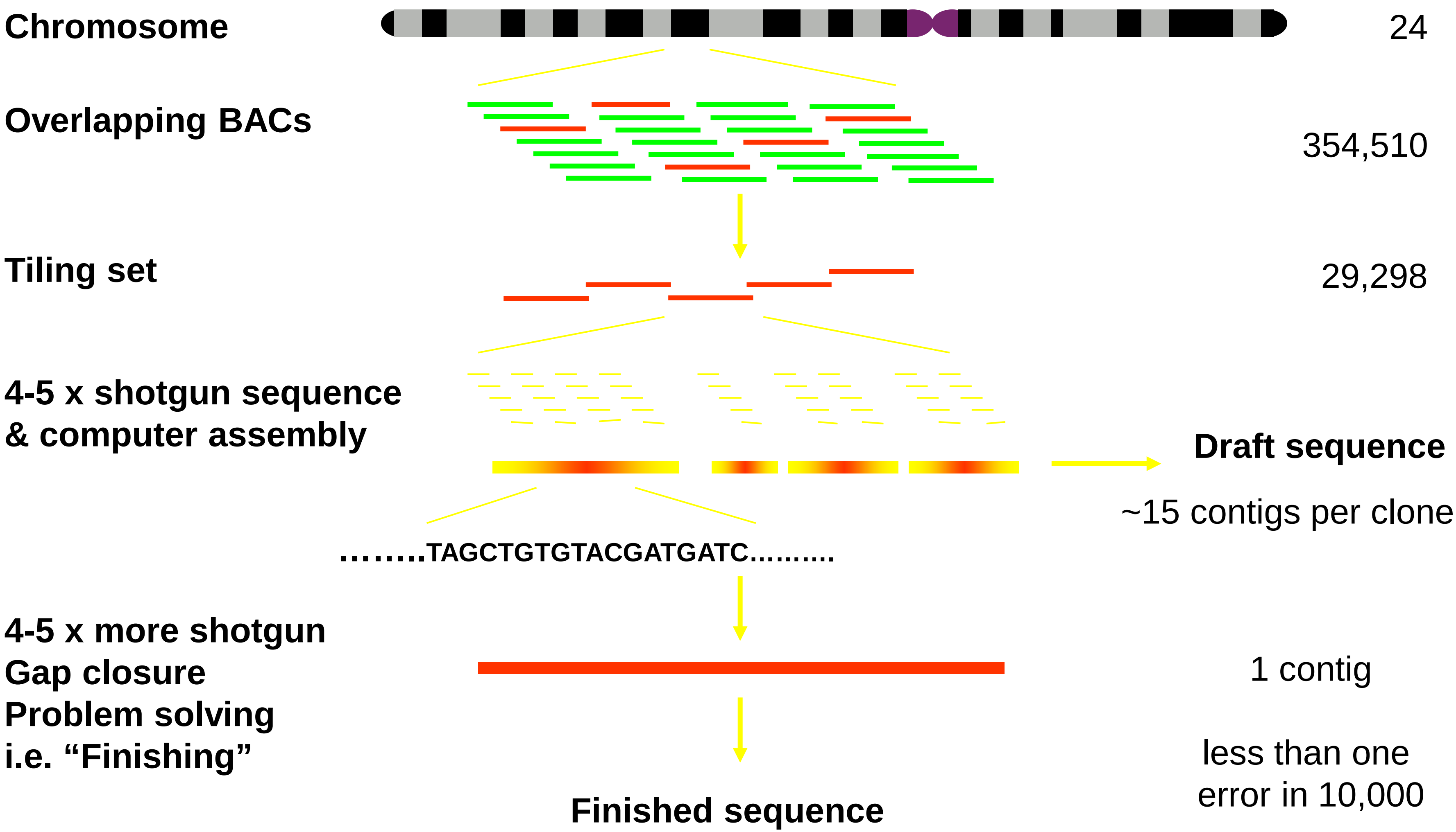
Pedigrees Reveal the Inheritance of Genes



What’s a Genome?

- **Genome** - All of the DNA for an organism
- Human Genome
  - Nucleus - 3.1 billion base pairs packaged into chromosomes
  - Mitochondria - 16,569 base pairs packaged in one circular chromosome
- Each cell:
  - 46 chromosomes
  - 2 m DNA
- Just 20,441 genes (just a bit more than *Drosophila*)
- Uneven gene density across chromosomes
- Hundreds of bacterial genes (horizontal transfer)
- Scores of genes acquired throughout transposable elements

Human Genome Project



Reading the Genome

TAGTCGAGGCTTTAGATCCGATGAGGCTTTAGAGACAG

AGTCGAG CTTTAGA CGATGAG CTTTAGA

GTCGGG TTAGATC ATGAGGC GAGACAG

TAGTCGA TCTAGAT GAGCCT GAGACAG

GAGGCTT GATCCGA GGCTTTAGA

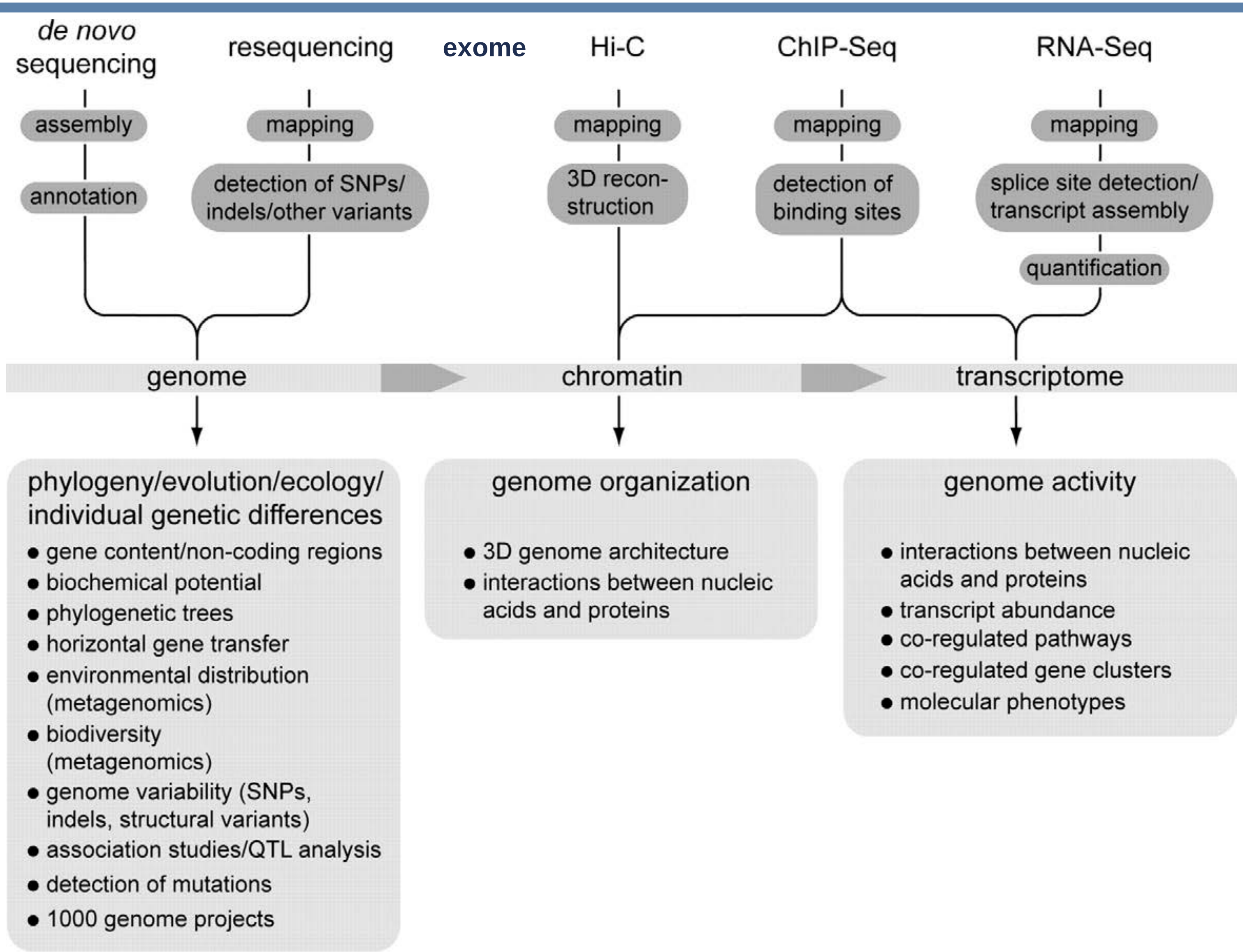
GGCTTTA ATCCGAT TTAGAG

Thermo Fisher
SCIENTIFIC

Key Metrics for an Assembly

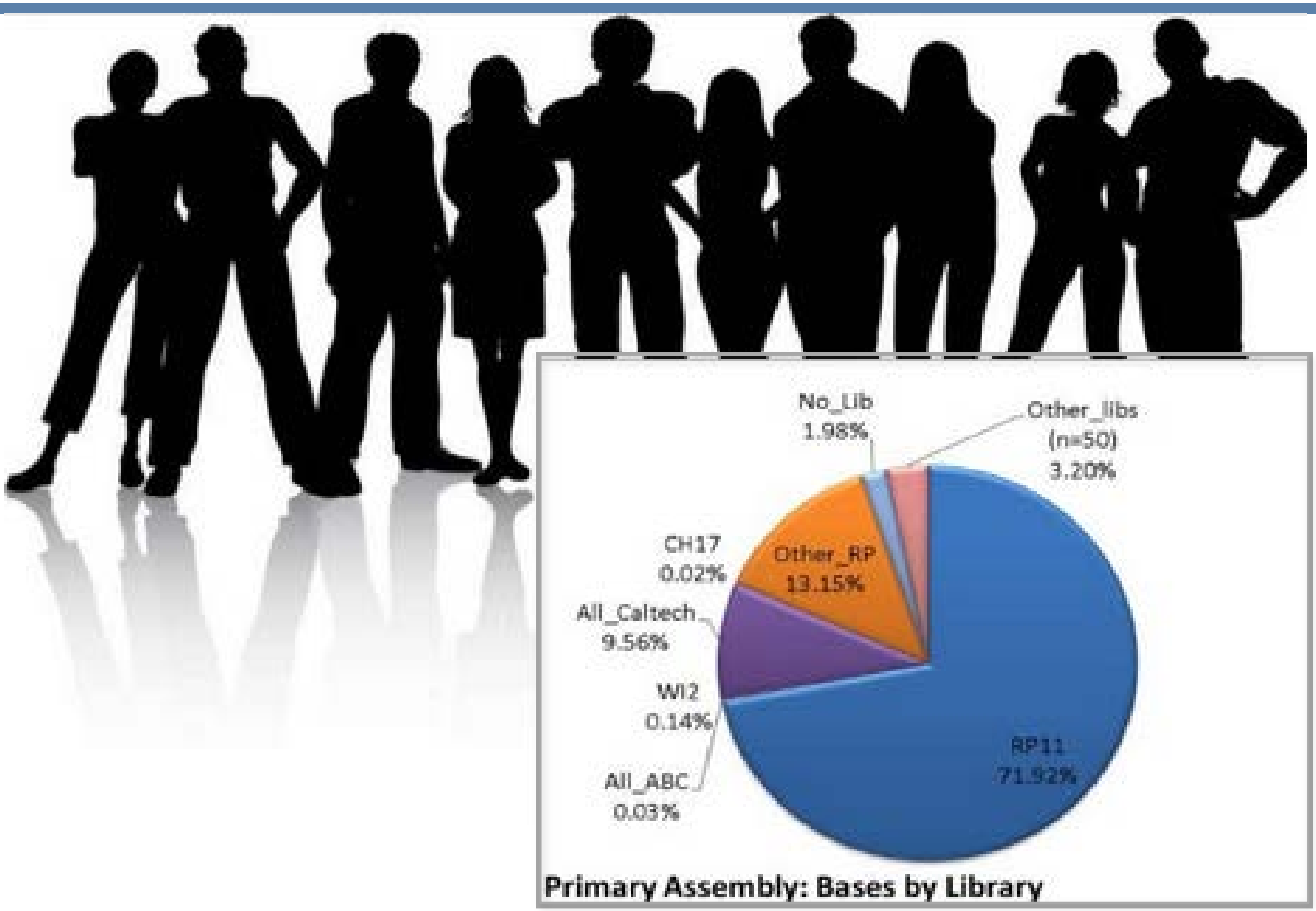
Metric	Description	Example
N50	"**Larger** is better" N50 is the *contig* length such that using equal or longer contigs produces 50% of the bases of the assembly.	N50 = 59.3 Mb
Number of contigs Contig: a set of overlapping DNA segments that together represent a consensus region of DNA	"**Lower** is better" As the number of contigs that can be placed into an assembly decreases this number gets lower. In the limit a completely assembled genome with zero gaps would have 1 contig	Nr of contigs = 1,449
% Reference Coverage	"Closer to **100%** is better" % Reference Coverage means what % of total bases in the genome were covered by at least one read	% Ref Coverage = 99.74%

Sequencing What?



Adapted from M Nowrousian (2010) *Cell* **9**:1300

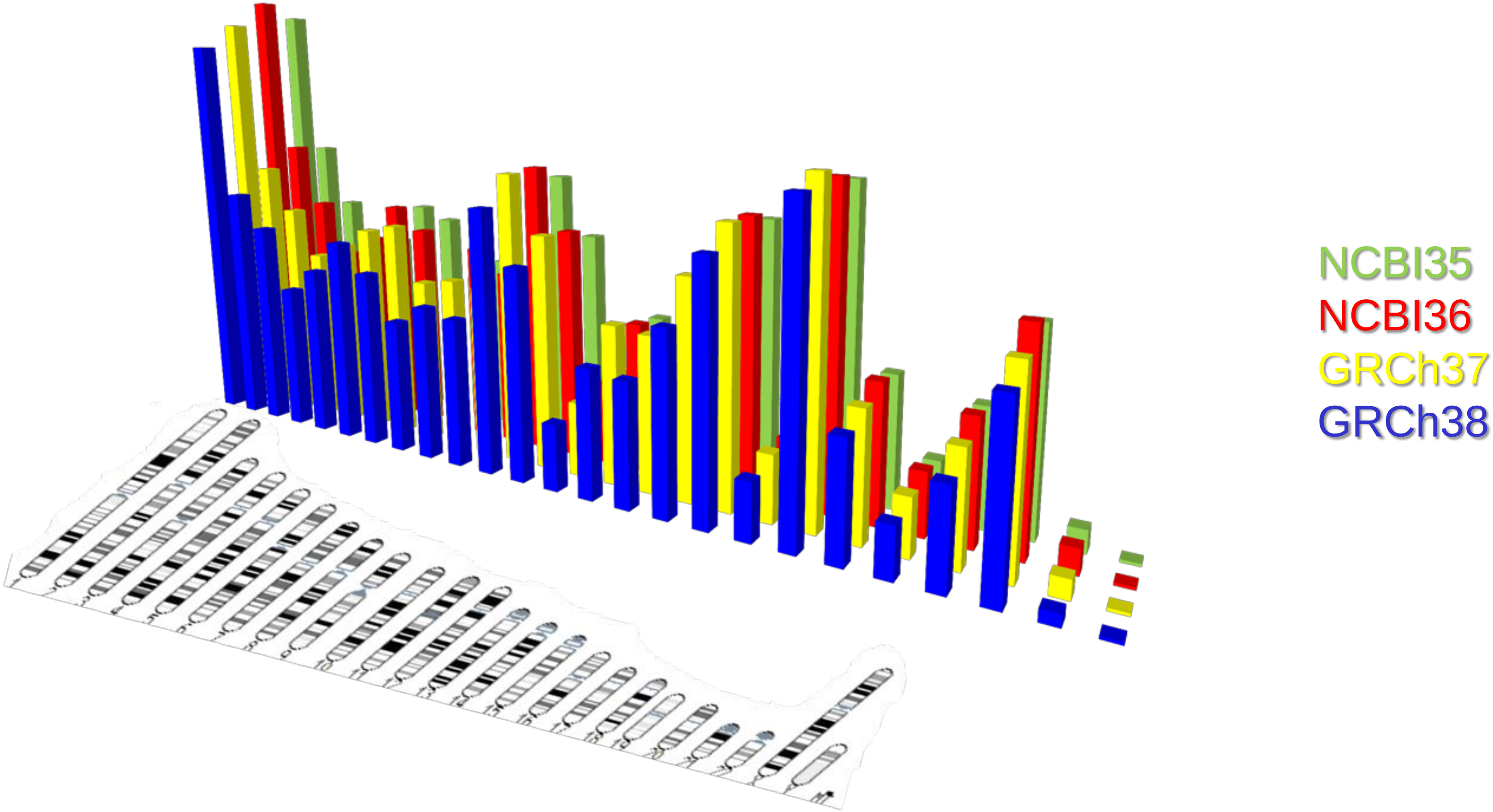
Whose Genome was Sequenced?



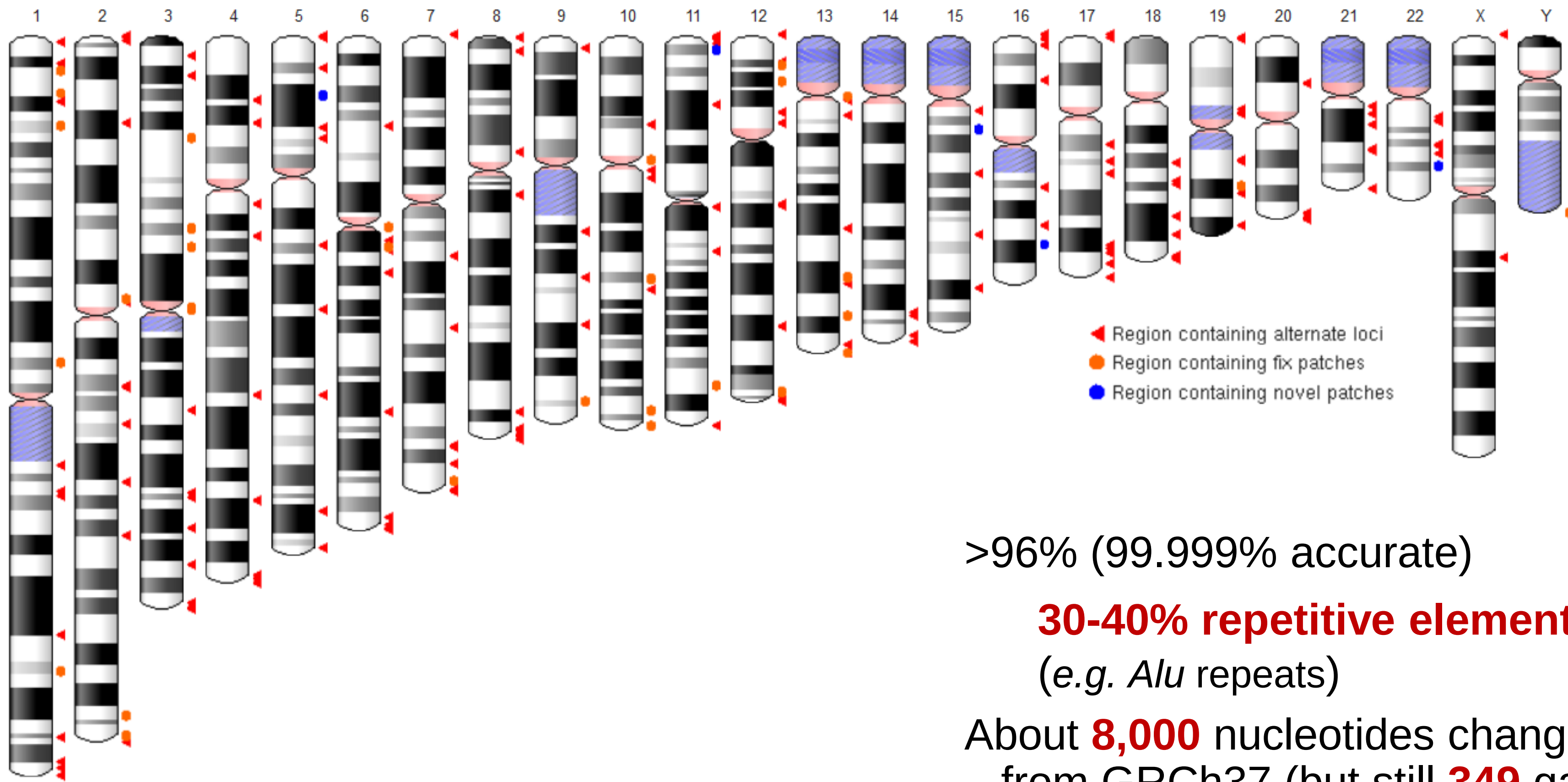
The human genome reference sequence does not represent an exact match for any one person's genome (composed of DNA from anonymous individuals across different racial and ethnic groups).

IHGSC. Nature (2001) **409** 860-921

Human Genome – Assemblies



Current Status – GRCh38



>96% (99.999% accurate)

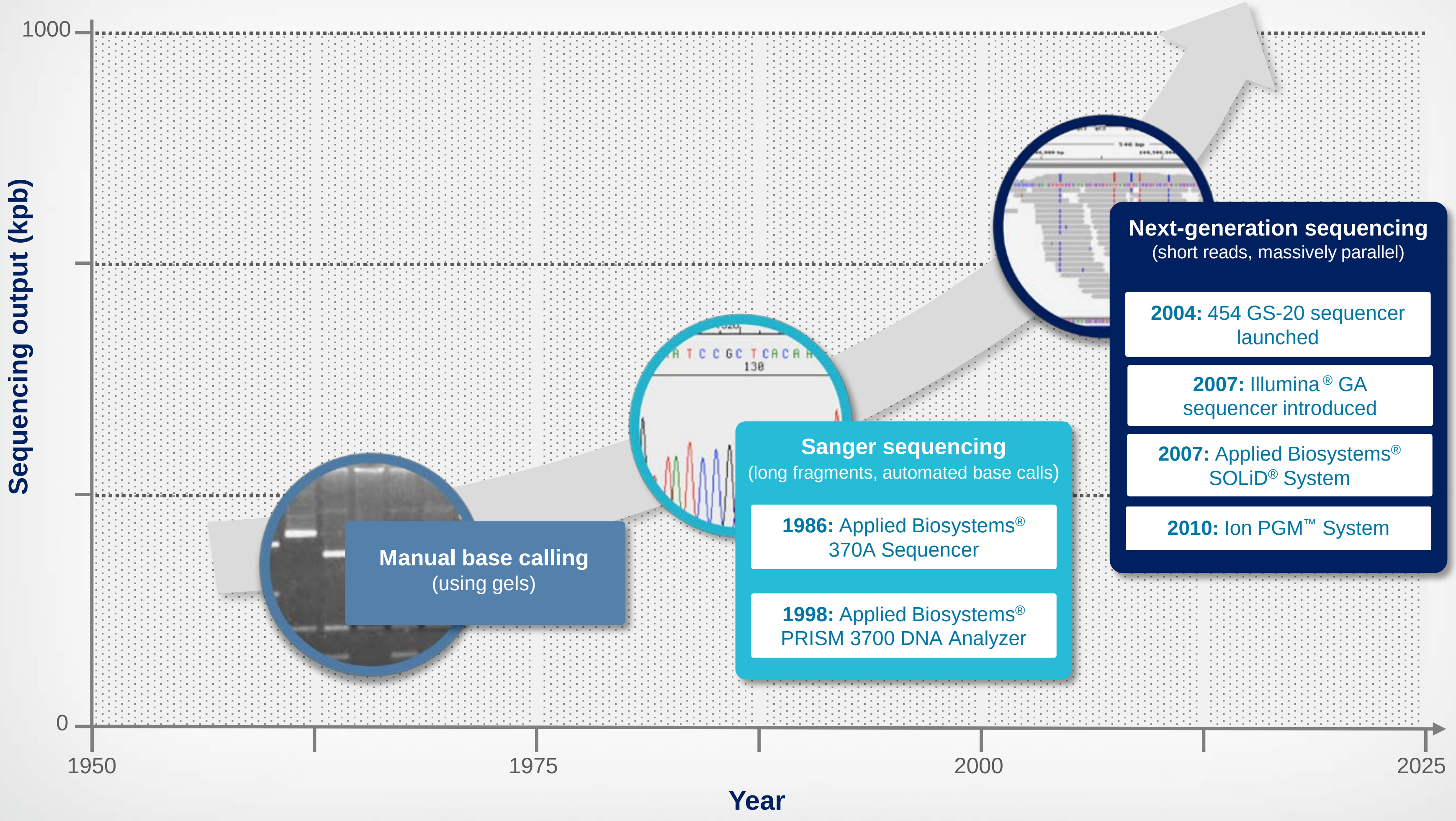
30-40% repetitive elements
(*e.g. Alu* repeats)

About **8,000** nucleotides changed from GRCh37 (but still **349** gaps)

All known genes, correctly identified (**99.74%**)

Assembled draft sequence totals 3.4GB

Exponential Advancements in Sequencing Technology

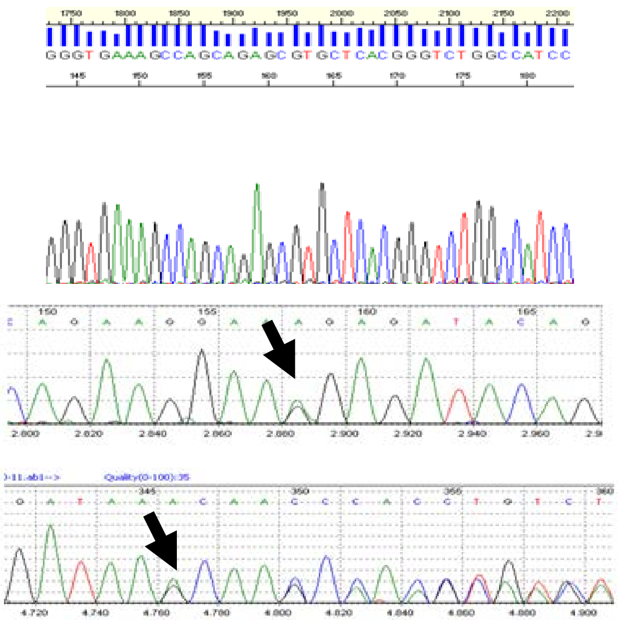


ThermoFisher SCIENTIFIC

Data Analysis – Basecalling

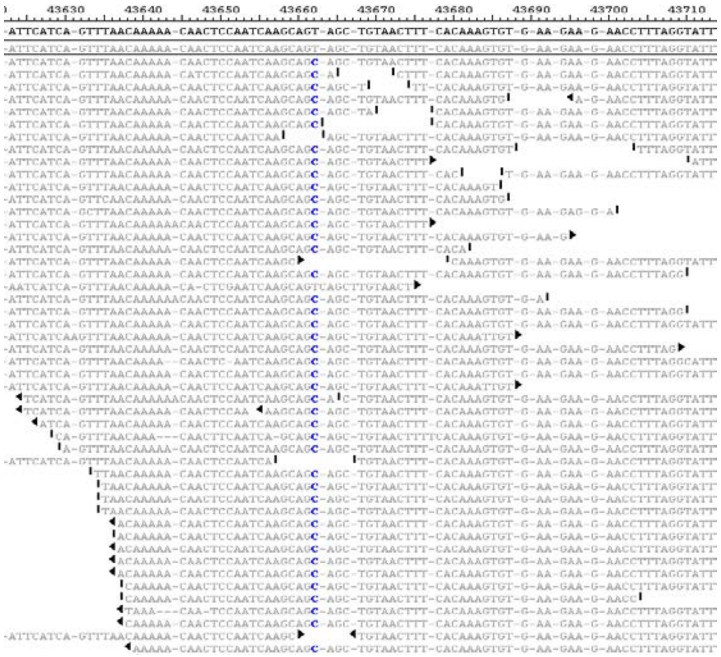
Sanger Sequencing

- The basecalling results from the signal from the **entire population** of molecules

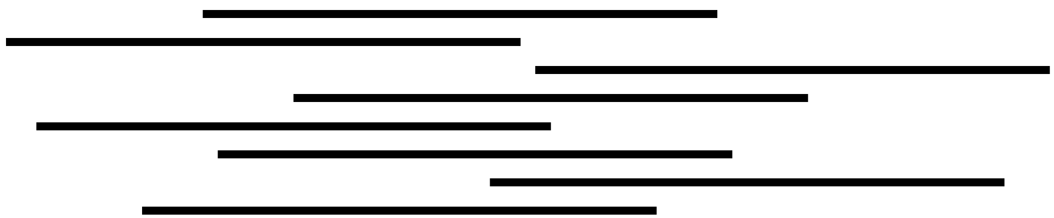


High Throughput Sequencing

- The basecalling is done for **each molecule**, clonally amplified

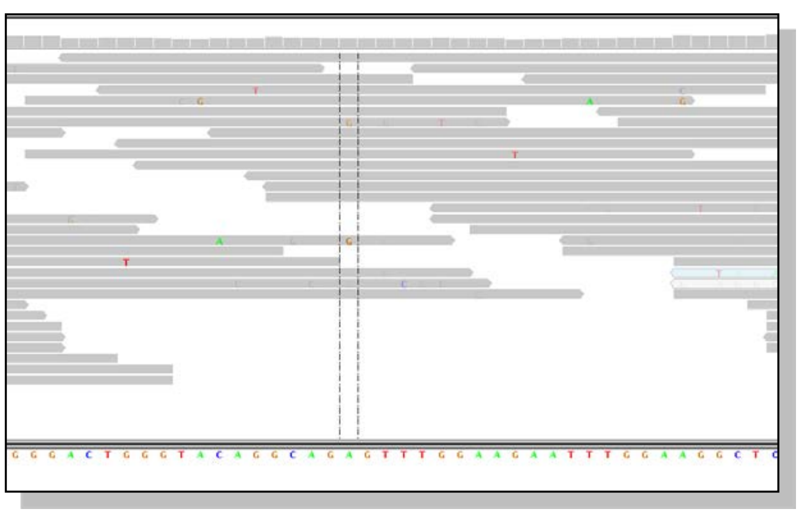


Building a *Digital* Consensus?



Reads do not share the same start or end point

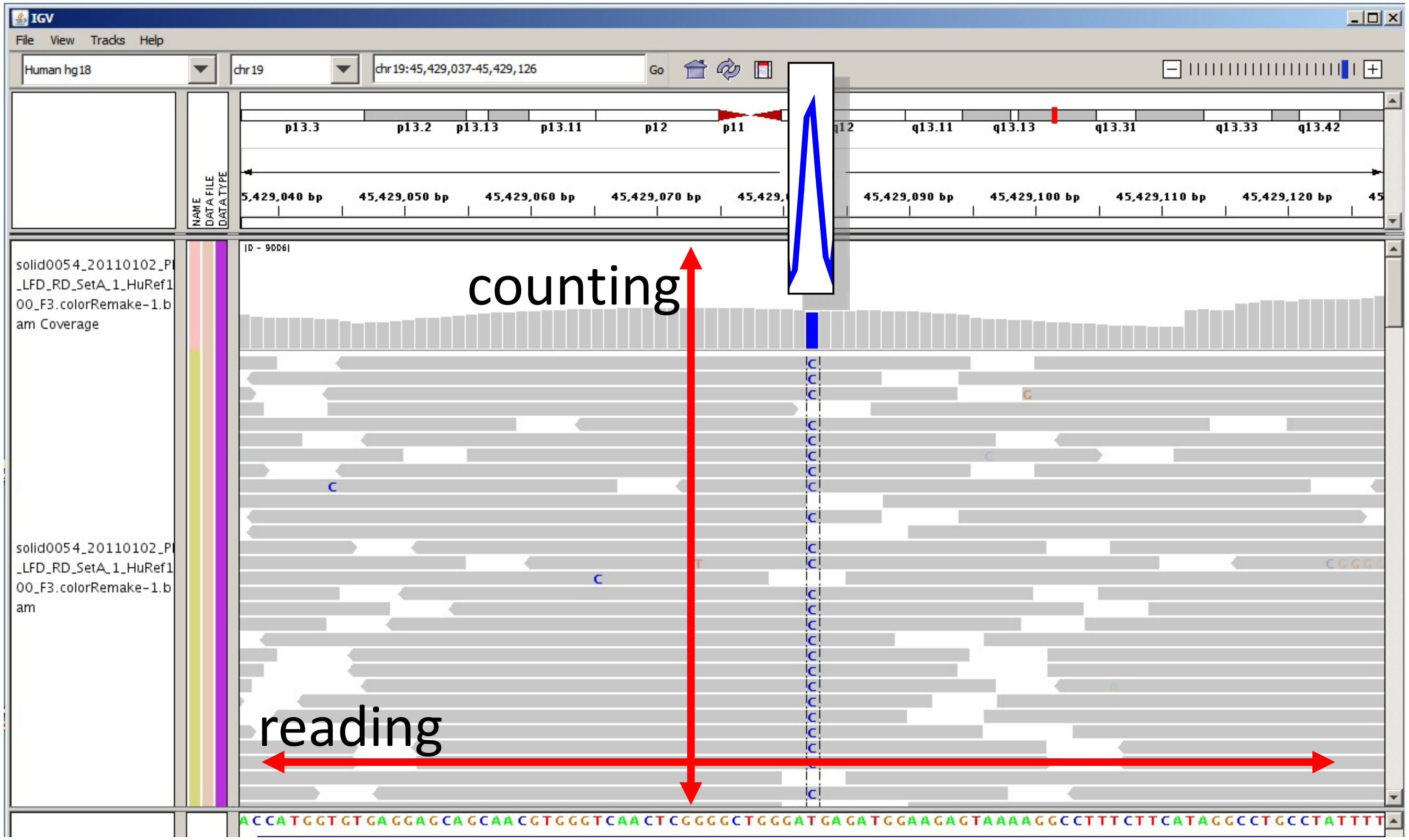
GGCGAGCGCCGGTCGTACCTCTGTGCGTCAC



...but here we know all bases of each read, not just the last one

...so we can align all reads to a reference

Digital Consensus Generation



An NGS consensus is generated bioinformatically aligning reads to the reference and counting all reads that cover each base position

Sanger *vs* NGS

Sanger or capillary sequencing

- Dominant for last ~40 years
- 1,000 bp longest read
- Based on primers so not good for repetitive or SNPs sites

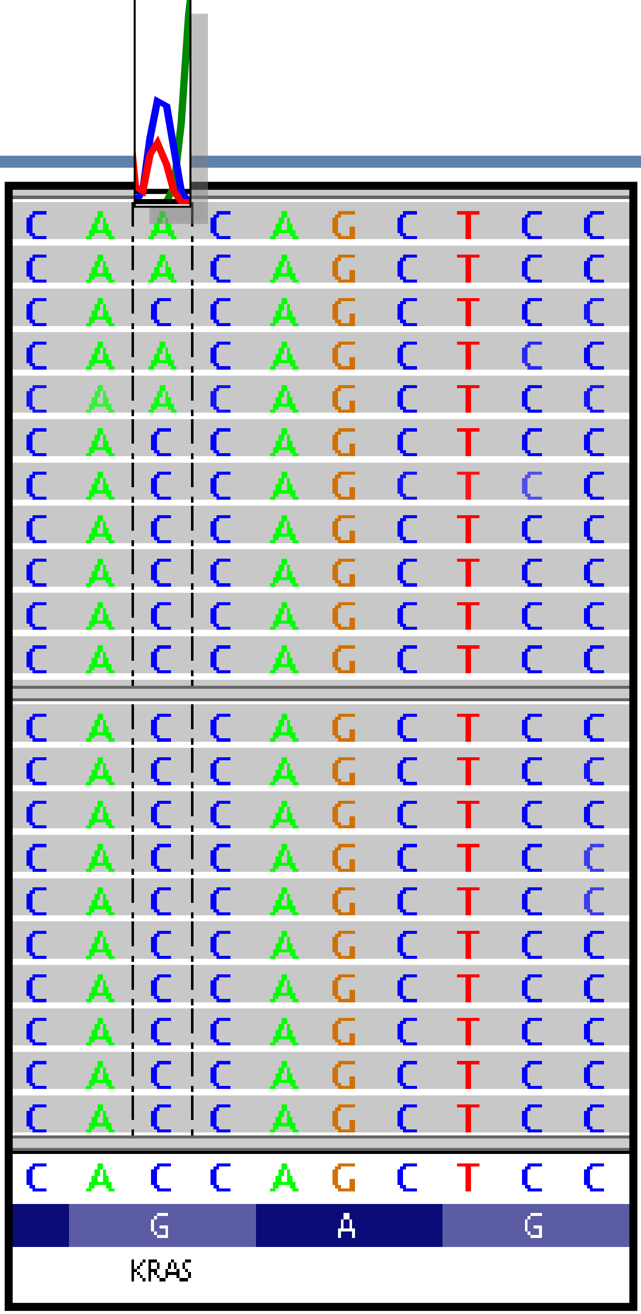
Next Generation Sequencing

- Shorter reads, 36 to 400-600* bp
- Higher throughput
- Cheaper cost per Mb
- Single molecule sequencing (no cloning step)
- More DNA sequenced since January 2008 than *all previous* years

PacBio® RS II reads can exceed 10,000 bp, with 5,000 average

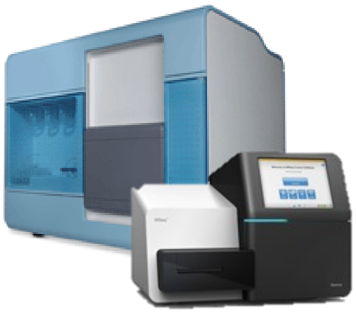
Deep Sequencing

- When minor allele representation is reduced, CE cannot call the SNP
- Here a higher number of measurement (deep sequencing) helps to build more confidence in the call
- With 1,000 reads on a single position we collect enough confidence to call the minor allele



Leaping into the new generation

454, Illumina, SOLiD, Ion Torrent, Pacific Biosciences…

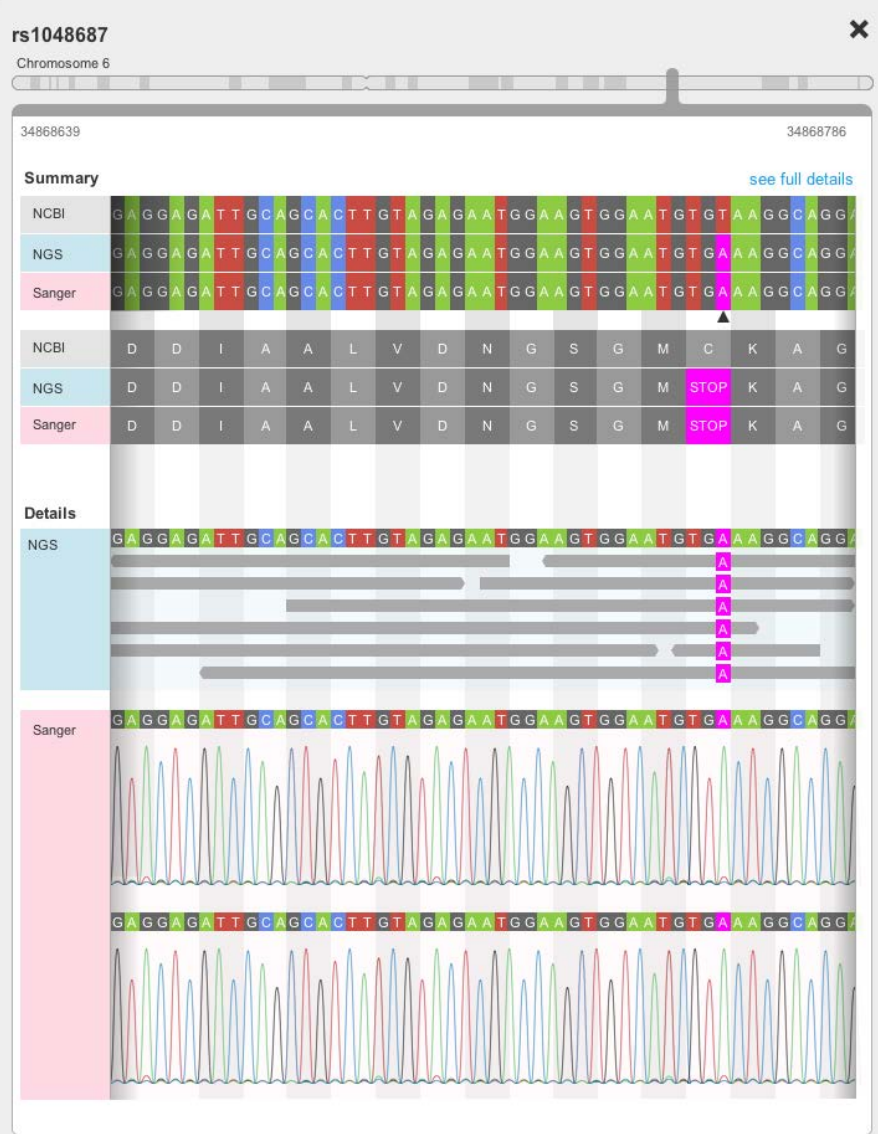


NGS - Sanger Confirmation

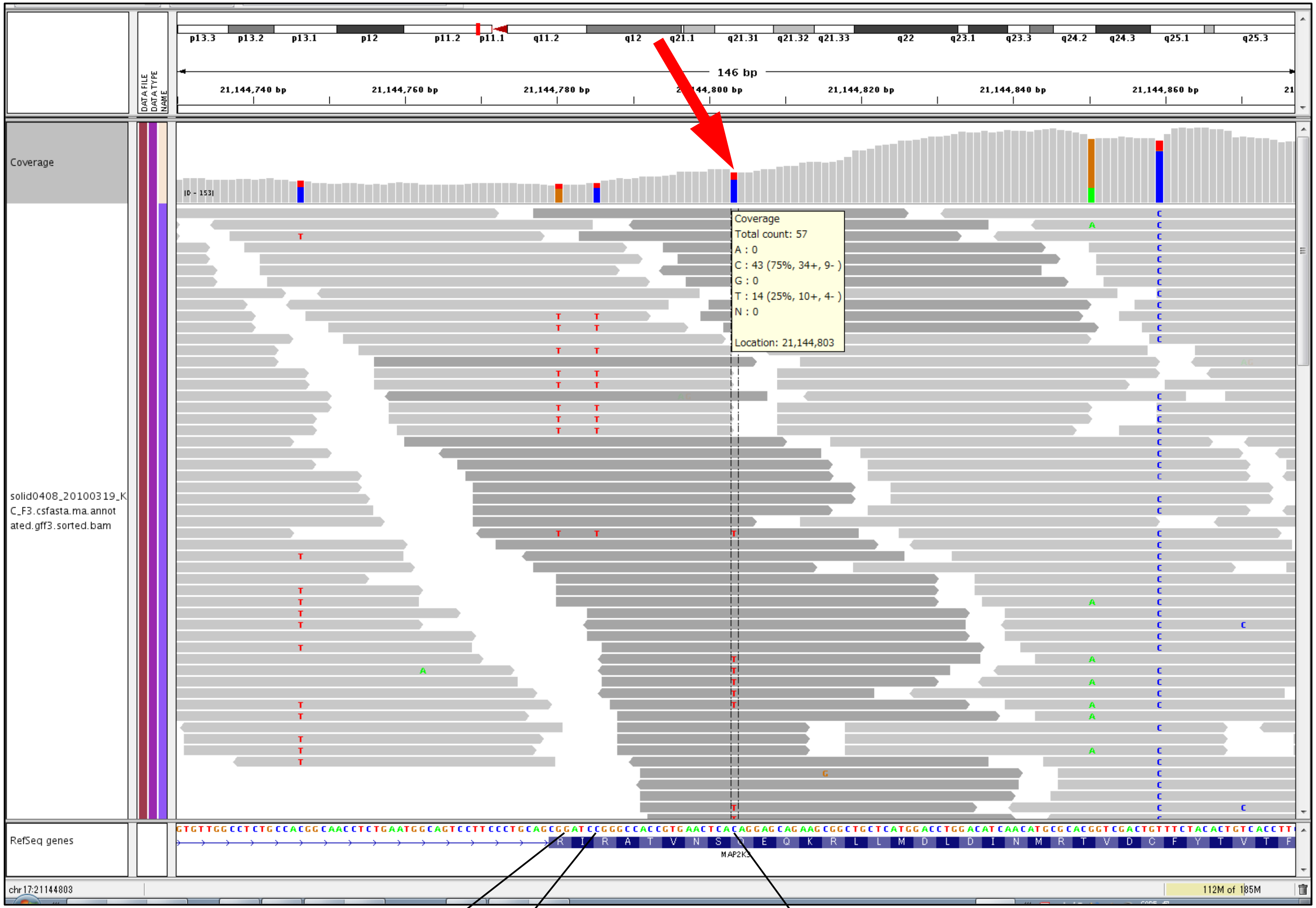
Summary

- Ion NGS
- Sanger CE

Variant	Affects	Sample	Ref	NGS	Sanger
rs1047979	UWY2	sample1b	T	TTGG	TTGG
rs1047801	GJAA6	sample1b	G	C	C
rs1047802	promoter for TYU2	sample1b	G	T	T
rs1048687	YULI5	sample1b	C	A	A
rs1049792	NLSZ11	sample1b	A	AC	AC
rs1057292	inhibitor for ACCP3	sample1b	T	T	G
rs1069202	HVLP	sample1b	A	A	G
rs1071029	no association	sample1b	G	G	C
rs1077922	ERTS2	sample1b	C	C	T
rs1081104	CLIB44	sample1b	C	C	C
rs1107923	PKK4	sample1b	A	G	A
rs1148922	DHWD	sample1b	T	A	T



SNP Detected by NGS not Detected by Sanger



HuRef (Craig Venter) CCTGCAGCTGATCTGGGCCACCGTGAACTCATAGGAGCAGAAGCGGCTG

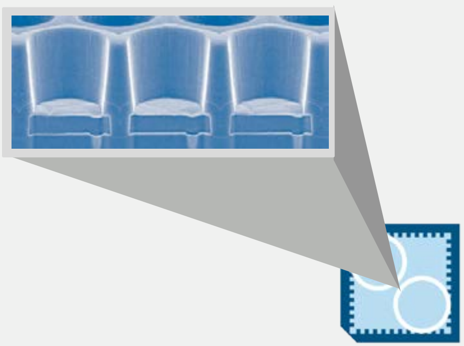
GRCh38 CCTGCAGCGGATCCGGGCCACCGTGAACTCACAGGAGCAGAAGCGGCTG

NGS – Some Initial Remarks

- Number of bases equals to “intensity”
  - More bases you sequence, the brighter the picture
  - But there is a limit once the picture is “bright enough”
- Readlength is “resolution” or zoom factor
  - Longer reads can see variations and haplotypes that cannot be seen with a shorter readlength
- Paired-end/Mate-pair reads are like a “digital zoom”
  - No change to fundamental zoom
  - But, software can synthesize additional zoom, if well implemented
- Read accuracy is “sharpness”
  - The more base errors, the more blurry the picture

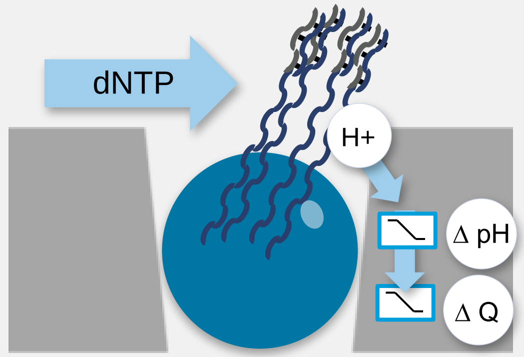
Sequencing Based on Natural Chemistry

1



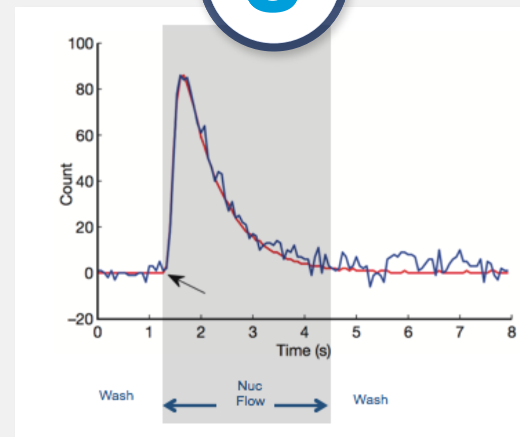
Add dNTPs that build the DNA sequence

2



Ions get released when DNA bases are matched

3



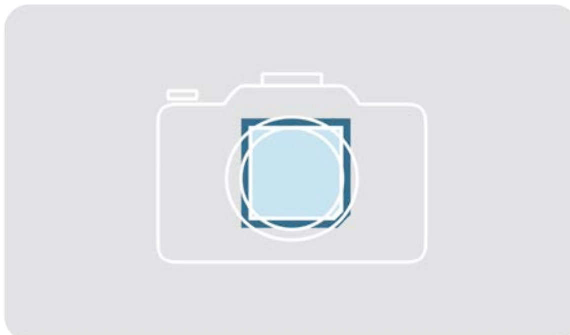
Sequence output is generated by the detection system

- Nucleotides flow sequentially over Ion semiconductor chip
- One sensor per well per sequencing reaction
- Direct detection of natural DNA extension
- Millions of sequencing reactions per chip

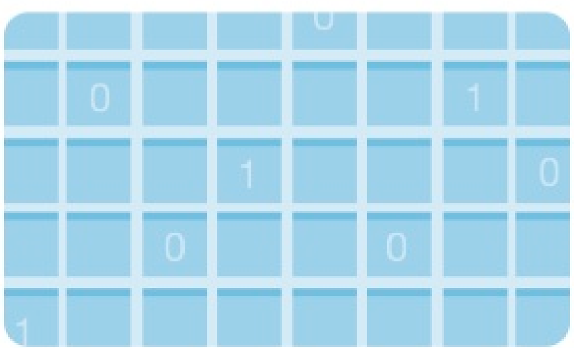
Ion Torrent™ Semiconductor Sequencing

Uses a process similar to that used in a digital camera

Digital camera chip



Covered in millions of pixels that convert light to digital information



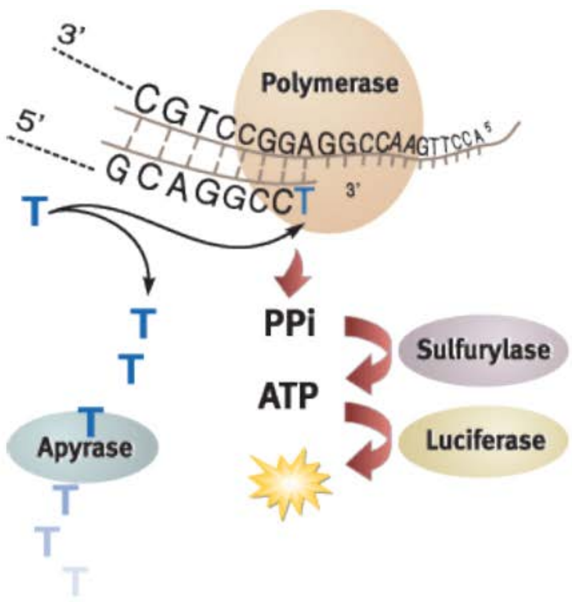
Ion Torrent™ sequencing chip

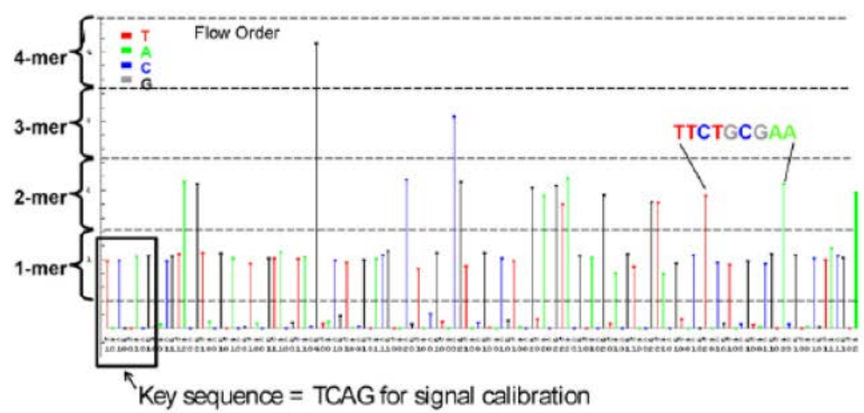
Millions of wells covering those pixels that convert chemical into digital information

ThermoFisher SCIENTIFIC

Pyrosequencing

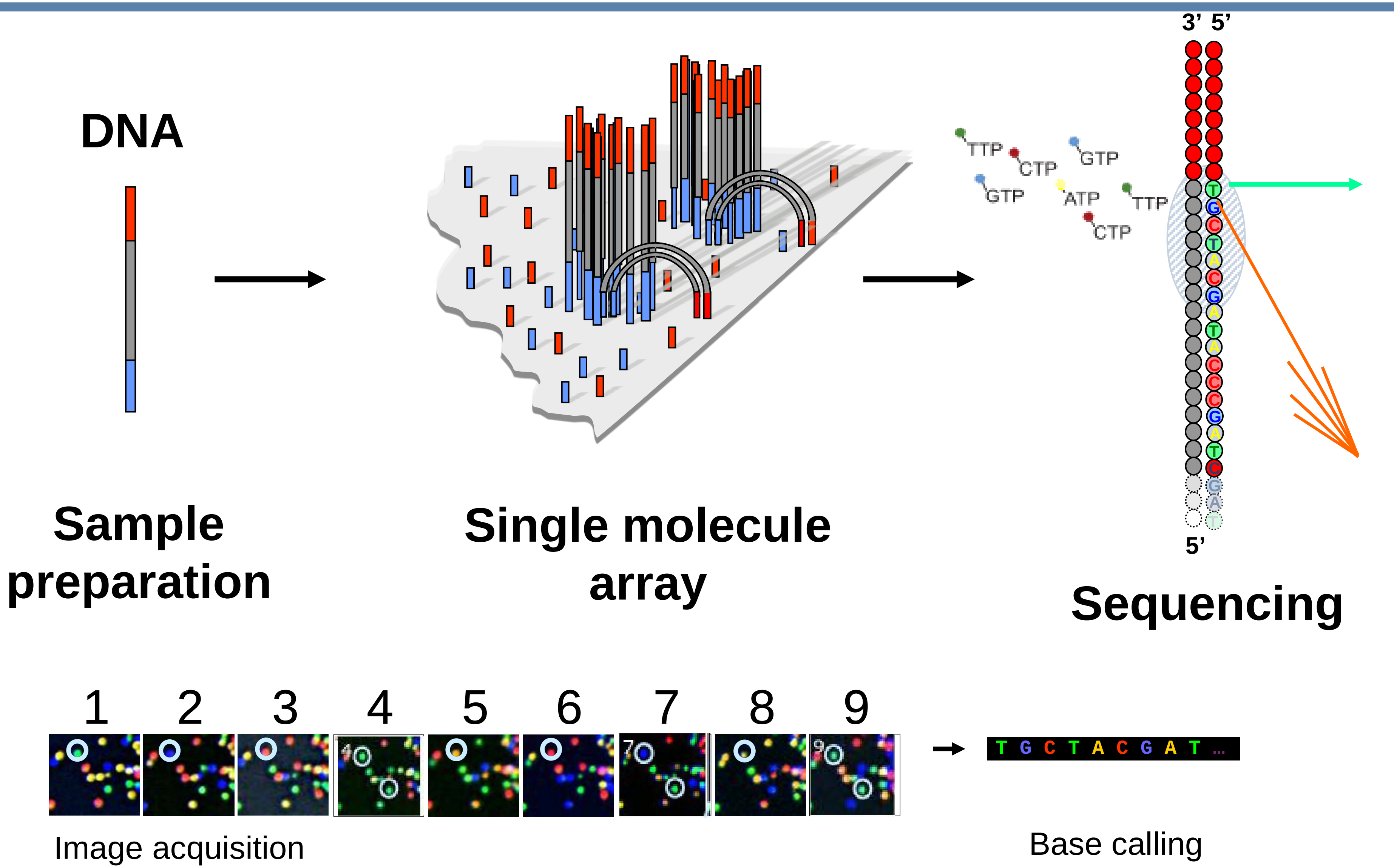
- Convergence of pyrosequencing and emulsion PCR
- Chemiluminescent detection of pyrophosphate released during polymerase-mediated deoxynucleoside triphosphate (dNTP) incorporation





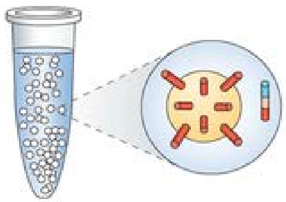
M Margulies *et al.* (2005) *doi:* 10.1038/nature03959

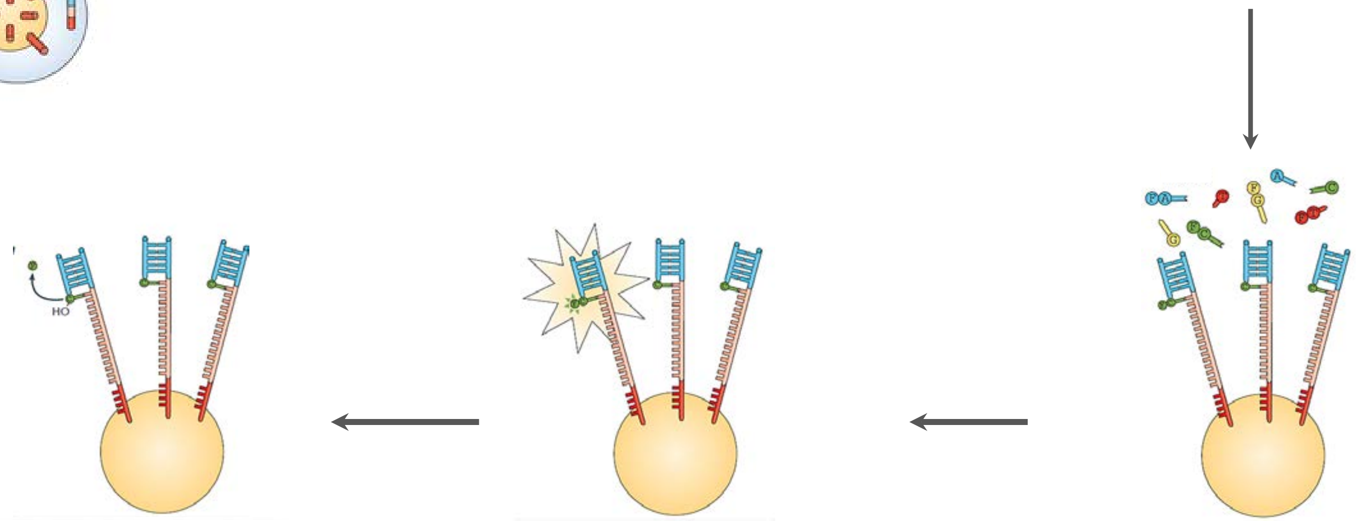
Sequencing based on Fluorescence

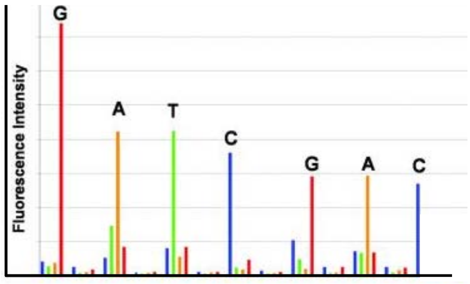


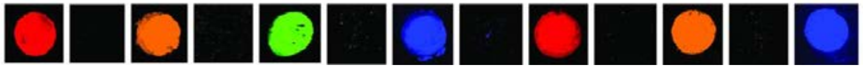
DR Bentley *et al.* (2008) *doi:* 10.1038/nature07517

SBS – Cyclic Reversible Termination





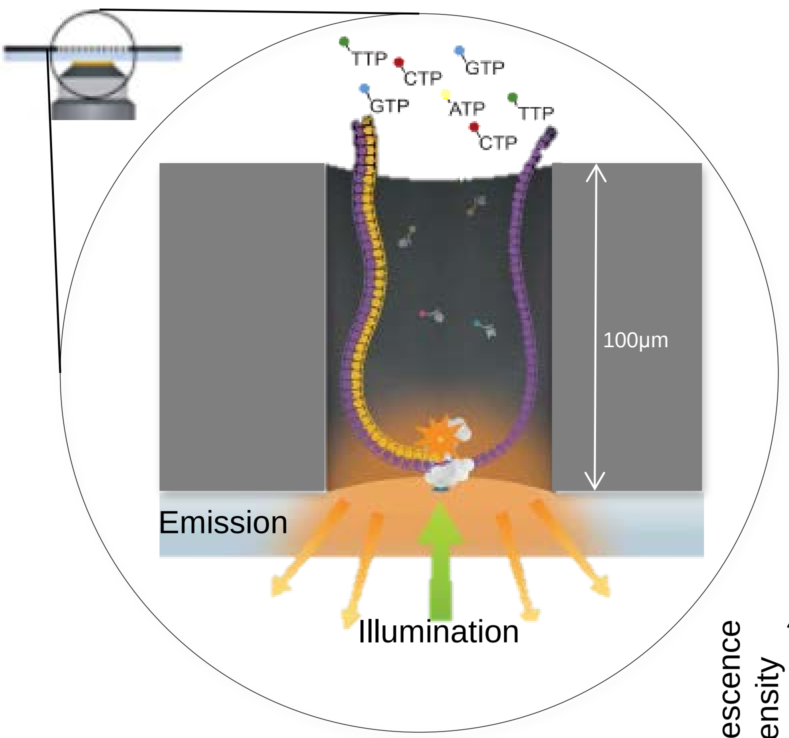


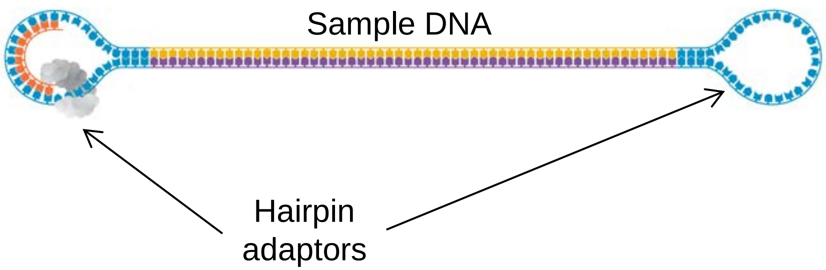


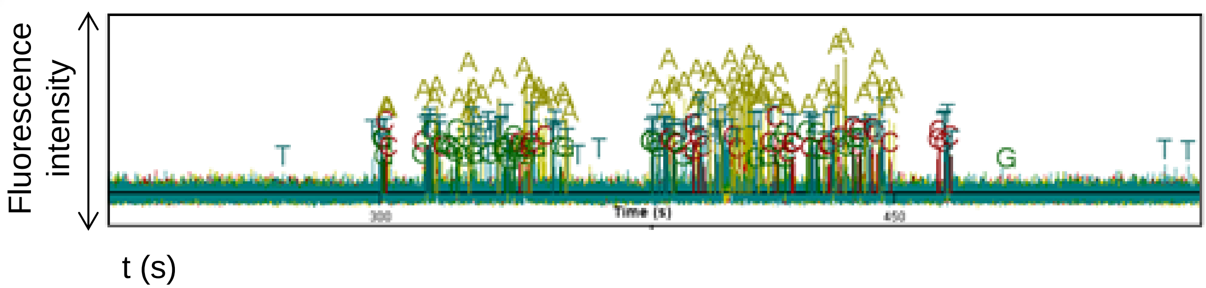
J Jingyue *et al.* (2006) *doi:* 10.1073/pnas.0609513103

ThermoFisher
SCIENTIFIC

Single Molecule Sequencing – Long Reads

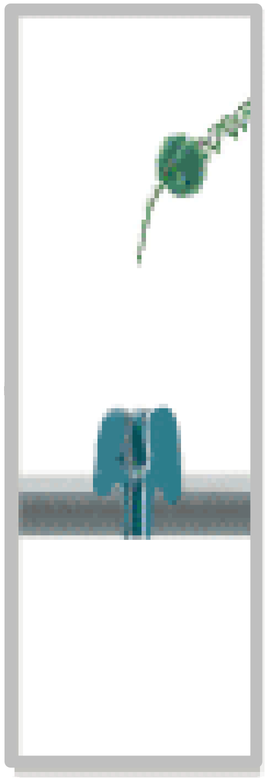


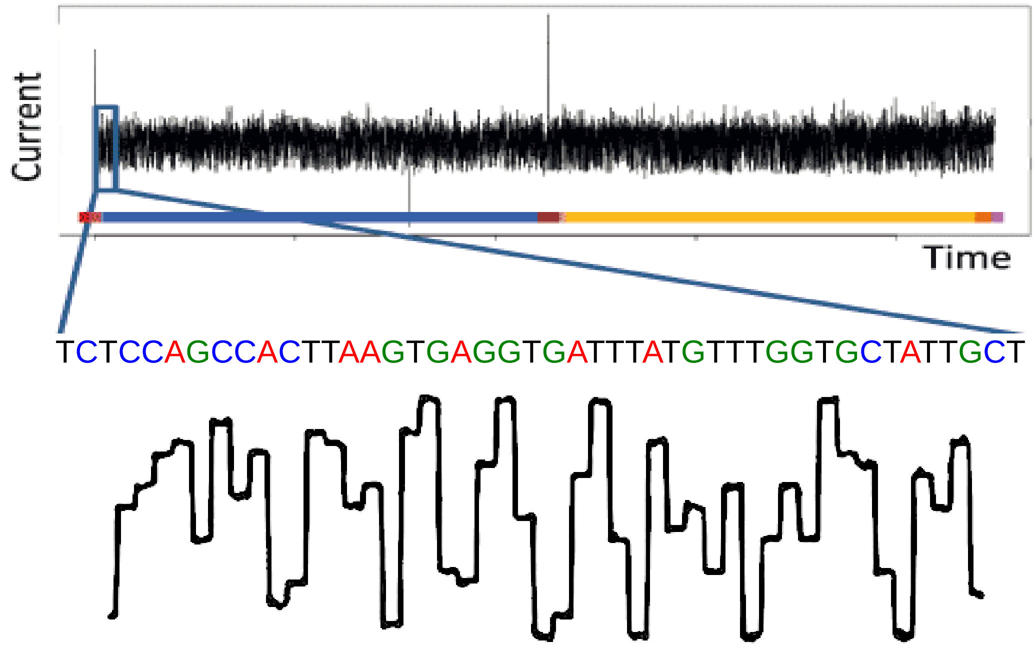




J Eid *et al.* (2009) *doi:10.1126/science.1162986*

Sequencing with Nanopores





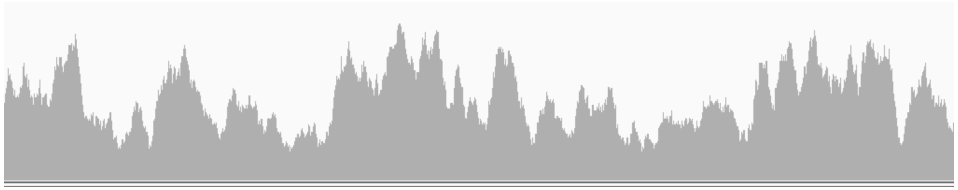
E Check Hayden *Nature doi:10.1038/nature.2012.10051*

Sequencing Data

GATGGGAAGA
GCGGTTCAGC
AGGAATGCCG
AGACCGATAT
CGTATGCCGT

Sequence data

- Precise
- Fairly unbiased
- Easy to QC

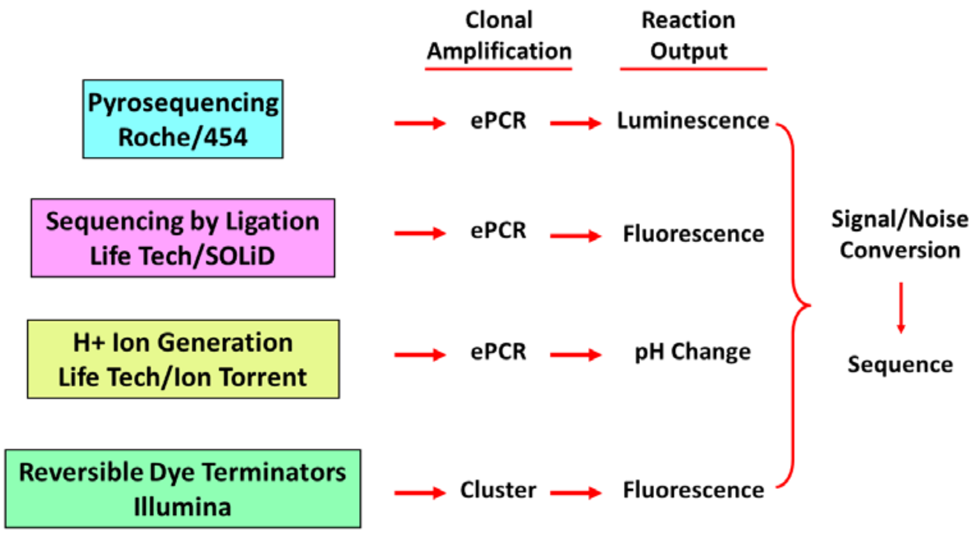


Coverage depth data

- Can be biased
- Hard to know what's true

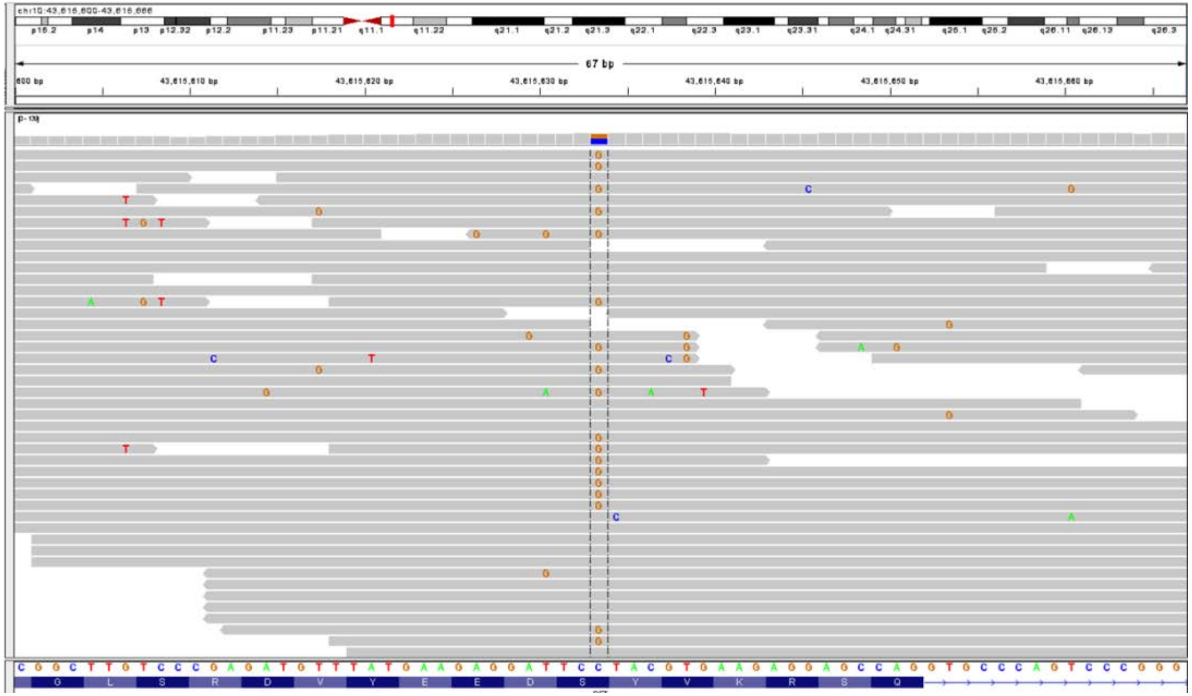
NGS Platforms

- Differ in design and chemistries
- Fundamentally related- sequencing of thousands to millions of clonally amplified molecules in a massively parallel manner
- Orders of magnitude more information-will continue to evolve
- Attractive for various applications – individual sequencing assays costly and laborious- serial “gene by gene” analysis.



NGS Application Examples

Inherited conditions



Discovery tool: Single gene disorders *i.e.* AD – Kabuki syndrome (MLL)

Causative mutations for multigenic diseases –superior to "one by one" approach of traditional sequencing

Diagnostic advancements for diseases with overlapping symptoms, multiple possible syndromes/genes

Quality Control

Questions you should ask (yourself or your sequencing provider):

- **Sequencing QC**
  - **How much** sequencing?
  - What's the sequencing **quality**?
- **Library QC**
  - What's the **base profile** across the reads?
  - Is there an unexpected **GC bias**?
  - Are there any library preparation **contaminants**?
- **Post mapping QC**
  - What is the **fragment length distribution**? (for paired end)
  - Is there an unexpected **duplicate** rate?

The Sequence Alignment / Map

- (SAM) format - SAM is text, BAM is binary
- Generic alignment format
- Supports short and long reads
- Supports different sequencing platforms (colour space)
- Flexible in style, compact in size,
- Efficient in random access

Sequencing Output

FASTQ format

Example FASTQ record:

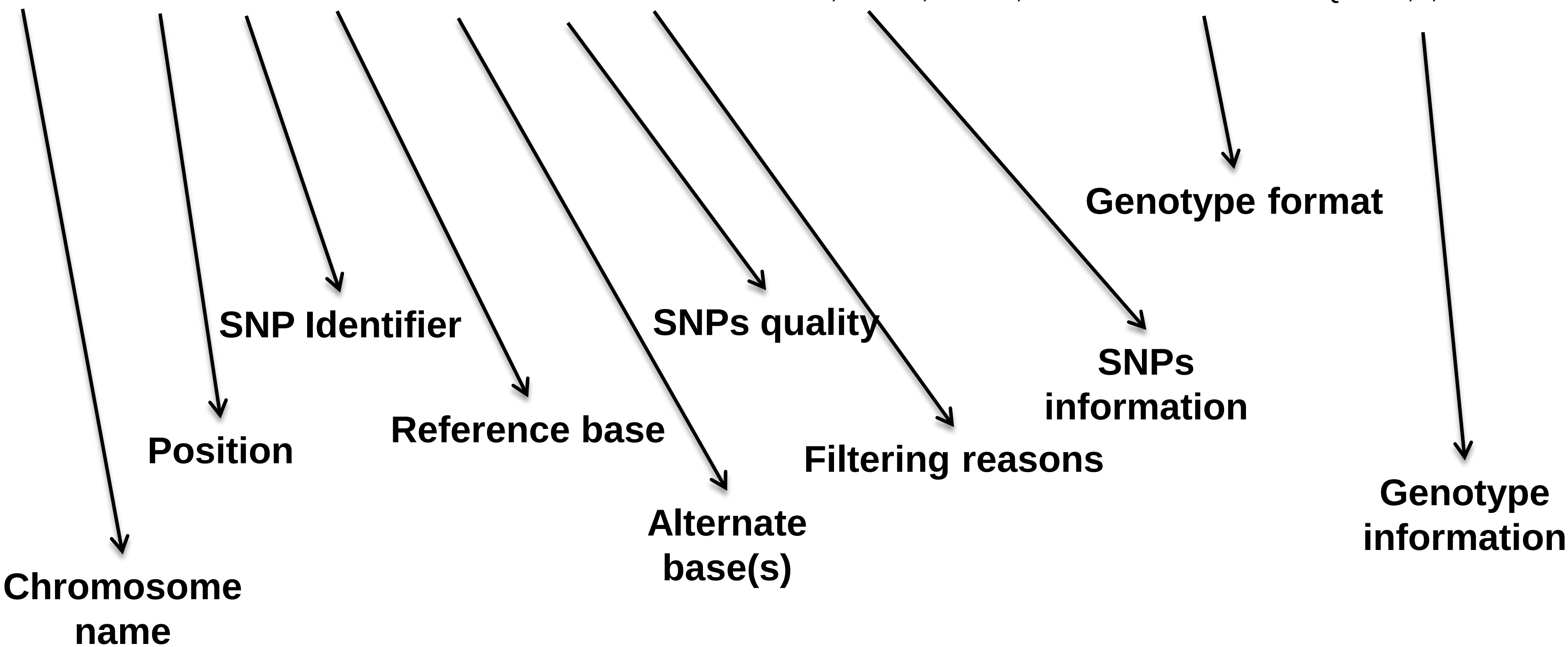
```
@06_0016:6:1:5388:12733#0
GATGGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATATCGTATGCCGTCTTCTGCTTGAAAAAAAAAAAAAAG
+
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCDDADACBCCCDADBDDCBCD;BBDBDBBB%%%%%%%%
```

```
 !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmn
 |                         |    |         |                               |
33                        59   64        73                              104
```

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90 %
20	1 in 100	99 %
30	1 in 1000	99.9 %
40	1 in 10000	99.99 %
50	1 in 100000	99.999 %

VCF Format

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	germline
chr4	27668	.	T	C	8.65	.	DP=2;AF1=1;AC1=4;…	GT:PL:DP:SP:GQ	0/1:0,0,0:0:0:3
chr4	27669	.	G	T	4.77	.	DP=2;AF1=1;AC1=4;…	GT:PL:DP:SP:GQ	0/1:0,0,0:0:0:3
chr4	27712	.	T	C	44	.	DP=2;AF1=1;AC1=4;…	GT:PL:DP:SP:GQ	1/1:40,3,0:1:0:8
chr4	27774	.	G	A	5.47	.	DP=2;AF1=0.5011; AC1=2; …	GT:PL:DP:SP:GQ	0/1:34,0,23:2:0:28
chr4	36523	.	A	T	10.4	.	DP=1;AF1=1;AC1=4;…	GT:PL:DP:SP:GQ	0/1:0,0,0:0:0:3



Header

```
@HD     VN:1.0   SO:coordinate
@PG     ID:MatesToBAM   VN:1.09
@RG     ID:BEV-1098767-F5-P2     SM:91550_F3_F5-P2       LB:Dh10b-50x25RR              PI:100   PU:CLARA_20091009_2     CN:BEV
@RG     ID:BEV-1099231-R3        SM:DH10B                LB:Dh10blibrary-50x50LMP      PI:2000  PU:JOAN_20090805_1      CN:BEV
@CO     history: ???
@SQ     SN:1     LN:247249719
@SQ     SN:2     LN:242951149
@SQ     SN:3     LN:199501827
<snip>
@SQ     SN:22    LN:49691432
@SQ     SN:X     LN:154913754
@SQ     SN:Y     LN:57772954
@SQ     SN:M     LN:16571
```

Type	Tag	Description
HD – header	VN*	File format version.
	SO	Sort order. Valid values are: *unsorted*, *queryname* or *coordinate*.
	GO	Group order (full sorting is not imposed in a group). Valid values are: *none*, *query* or *reference*.
SQ – Sequence dictionary	SN*	Sequence name. Unique among all sequence records in the file. The value of this field is used in alignment records.
	LN*	Sequence length.
	AS	Genome assembly identifier. Refers to the reference genome assembly in an unambiguous form. Example: HG18.
	M5	MD5 checksum of the sequence in the uppercase (gaps and space are removed)
	UR	URI of the sequence
	SP	Species.
RG – read group	ID*	Unique read group identifier. The value of the ID field is used in the RG tags of alignment records.
	SM*	Sample (use pool name where a pool is being sequenced)
	LB	Library
	DS	Description
	PU	Platform unit (e.g. lane for Illumina or slide for SOLiD); should be a full, unambiguous identifier
	PI	Predicted median insert size (maybe different from the actual median insert size)
	CN	Name of sequencing center producing the read.
	DT	Date the run was produced (ISO 8601 date or date/time).
	PL	Platform/technology used to produce the read.
PG – Program	ID*	Program name
	VN	Program version
	CL	Command line
CO – comment		One-line text comments

Alignment Records

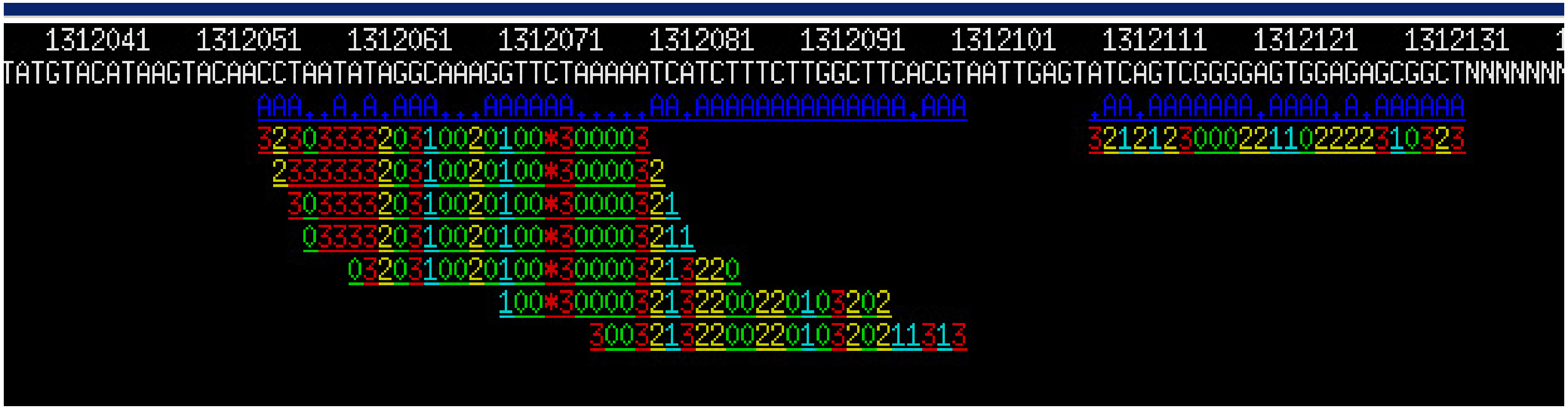
Part 1

```
1251_1005_1183  131   1   245   100   23M1I1M  =   4125   3905    TCTAAACCCTAAACCCTAACCCTTA   !NNNLNNNNNNNNNNNN=>FJN?9?!
737_1829_1497   131   1   600   50    25M      =   4009   3434    TGAGGAGAACGCAACTCCGCCGGCG   !NNNNNNKNNNN@1DN((7EN2&,!
1521_2026_1209  131   1   600   100   20M5H    =   4009   3434    TGAGGAGAACGCAACTCCGC        !NNNNNNNNNNND?KN%%+8
564_1311_1881   131   1   1670  50    25M      =   6095   4450    GTTCCTGCATGTAGTTTAAACGAGA   !NNNNNNNB7CNMFMJLN33GFNN!
737_1829_1497   67    1   4009  30    25M      =   600    -3434   TTAGGCTCTCAGCATGACTATTTTT   !GGJIHJJI55HIH36KJHFHKKJ!
1521_2026_1209  67    1   4009  50    25M      =   600    -3434   TTAGGCTCTCAGCATGACTATTTTT   *
1251_1005_1183  67    1   4125  100   25M      =   245    -3905   CCCTCTCATCCCAGAGAAACAGGTC   !II55I"""JGGKKKKKKJJJ55J!
```

Part 2

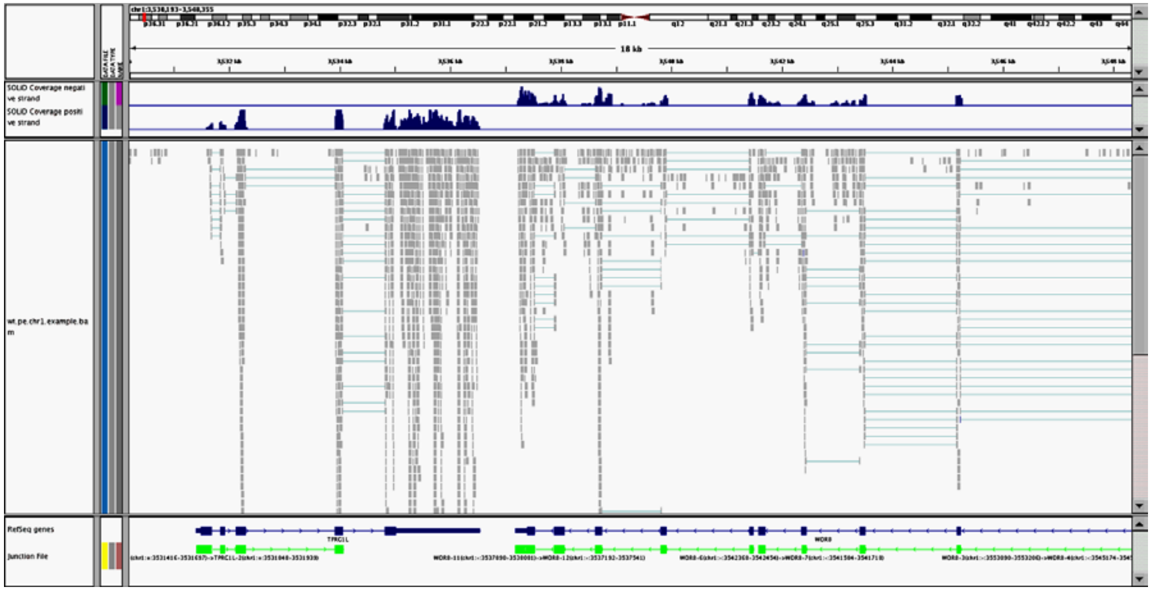
```
RG:Z:BEV-1099231-R3    CS:Z:G1223001002300100230100203 CQ:Z:<<<<3:;<<<<<3<;;7'8/<;%5+  MD:Z:25
RG:Z:BEV-1099231-R3    CS:Z:G1122022201331012223303233 CQ:Z:<<;;<<<48<<<9(*;<(,,:<2&'  MD:Z:25
RG:Z:BEV-1099231-R3    CS:Z:G1122022201331012223303233 CQ:Z:<<;<;<<<7<<<6/1;;%'%4<&&'  MD:Z:22GC1
RG:Z:BEV-1099231-R3    CS:Z:G1122022201331012223303233 CQ:Z:<<<<3:;<<<<<3<;;7'8/<;%5+  MD:Z:25
RG:Z:BEV-1099231-R3    CS:Z:T0032032222023130212330000 CQ:Z:6535645655654636665436665  MD:Z:25
RG:Z:BEV-1099231-R3    CS:Z:T0032032222123131212330003 CQ:Z:*                          MD:Z:25
RG:Z:BEV-1099231-R3    CS:Z:T2002022222001222200112212 CQ:Z:5555655655626666666565565  MD:Z:25 XW:z:23_27
```

SAMtools *tview* – Alignment Viewer

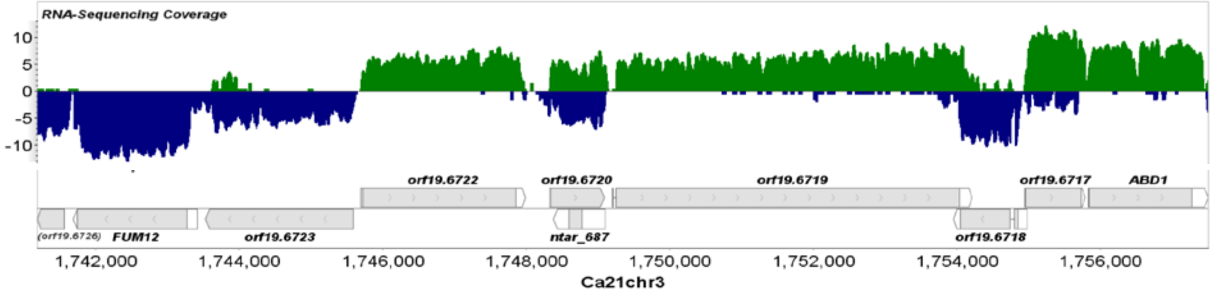


BAM and WIG Visualisation

Use .bam file to visualise coverage in IGV and load annotation tracks

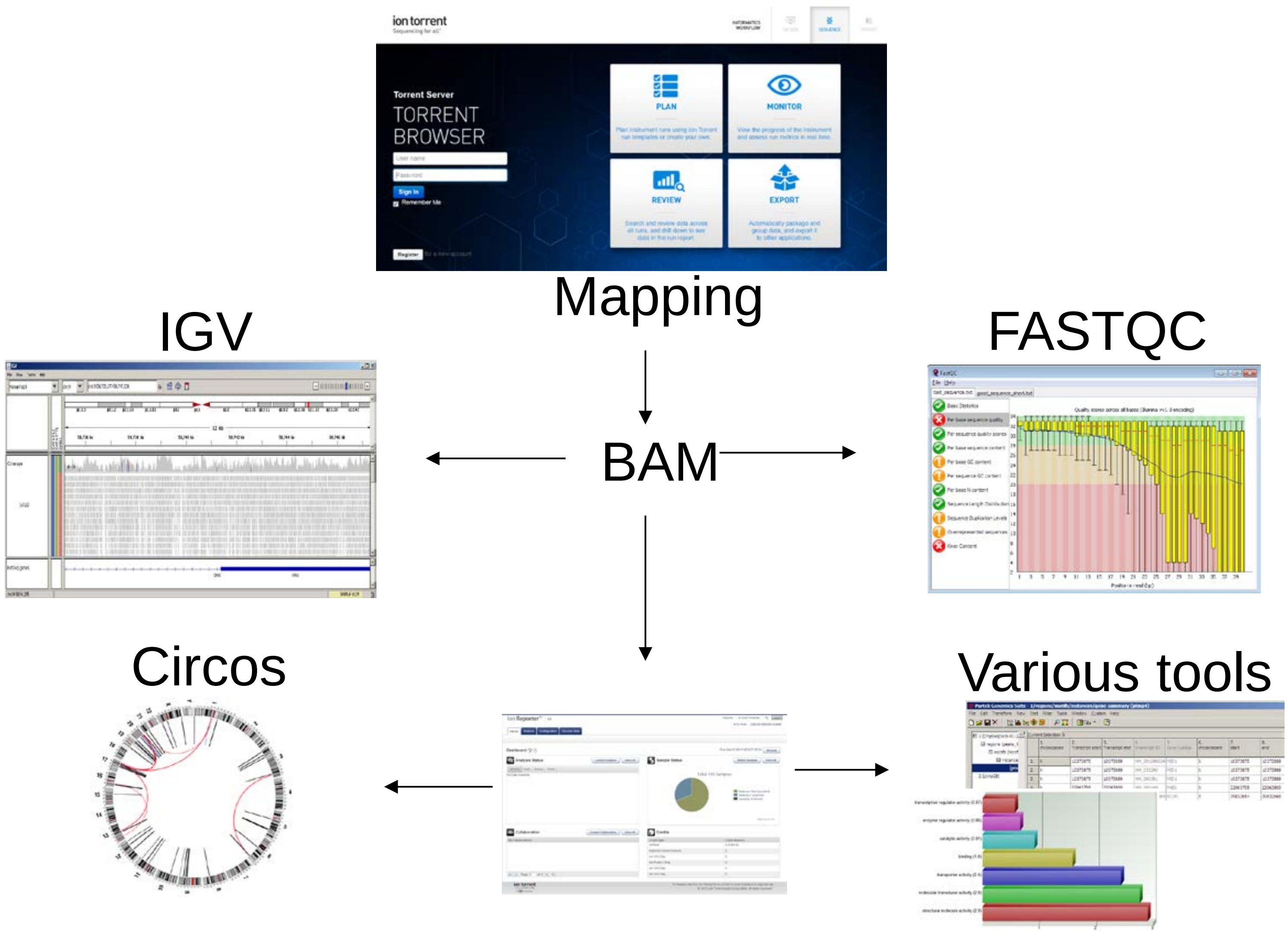


Use .wig files to visualise coverage in UCSC's Genome Browser



ThermoFisher SCIENTIFIC

Post-Mapping Workflows



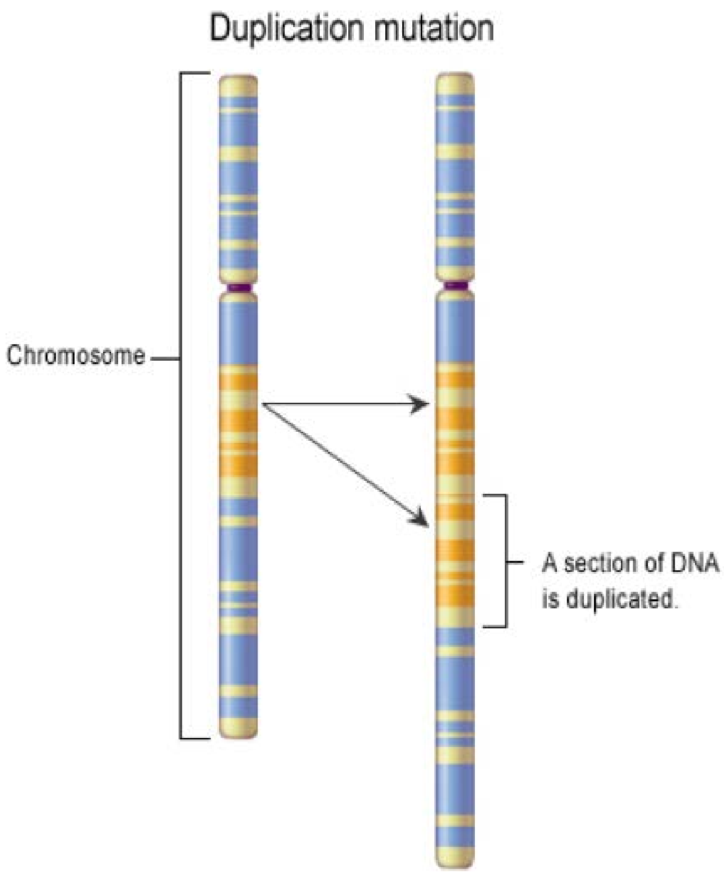
ThermoFisher SCIENTIFIC

Fernández-Suárez (2012), WIRE’s Data Mining Knowledge Discovery **2**:116

Human Variation

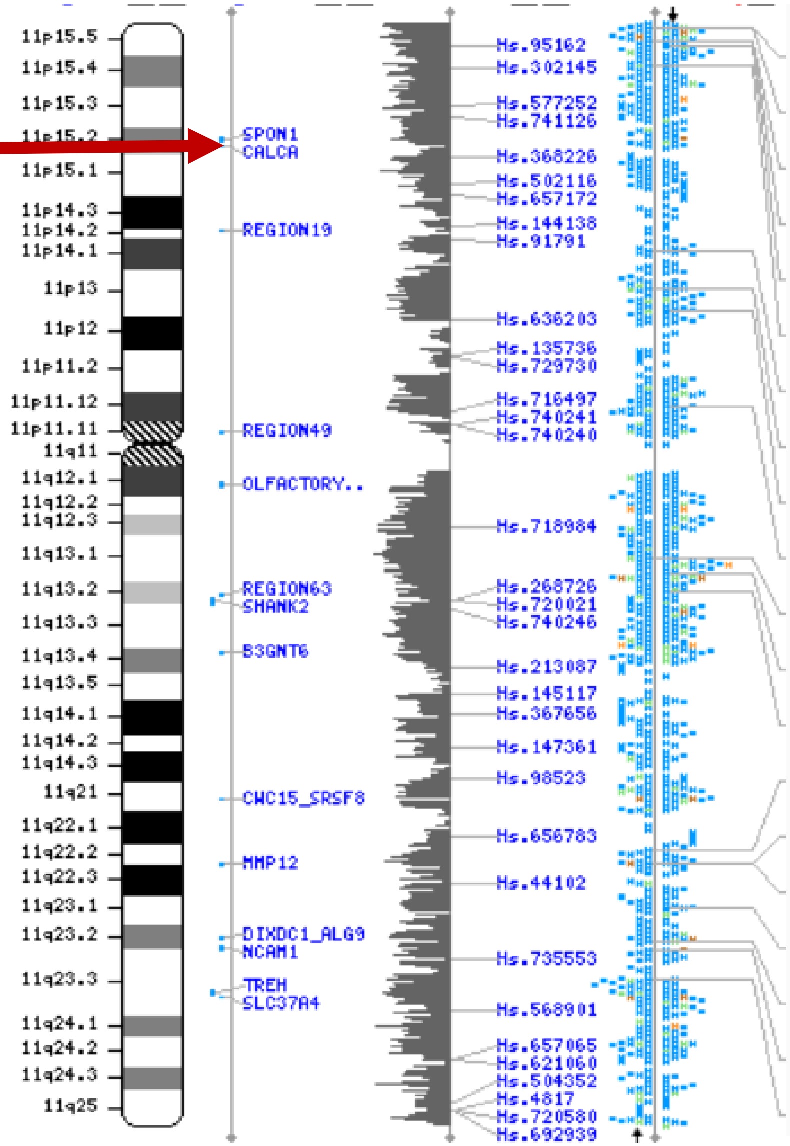
Changes in DNA

- **Deletion** − a section is missing
- **Translocation** − a section shifts from one chromosome onto another
- **Inversion** − a section gets snipped off and reinserted the wrong way around.
- **Single gene changes** − a small nucleotide change in a segment of the DNA that codes for a gene



Locus: A Region on the Chromosome

One locus

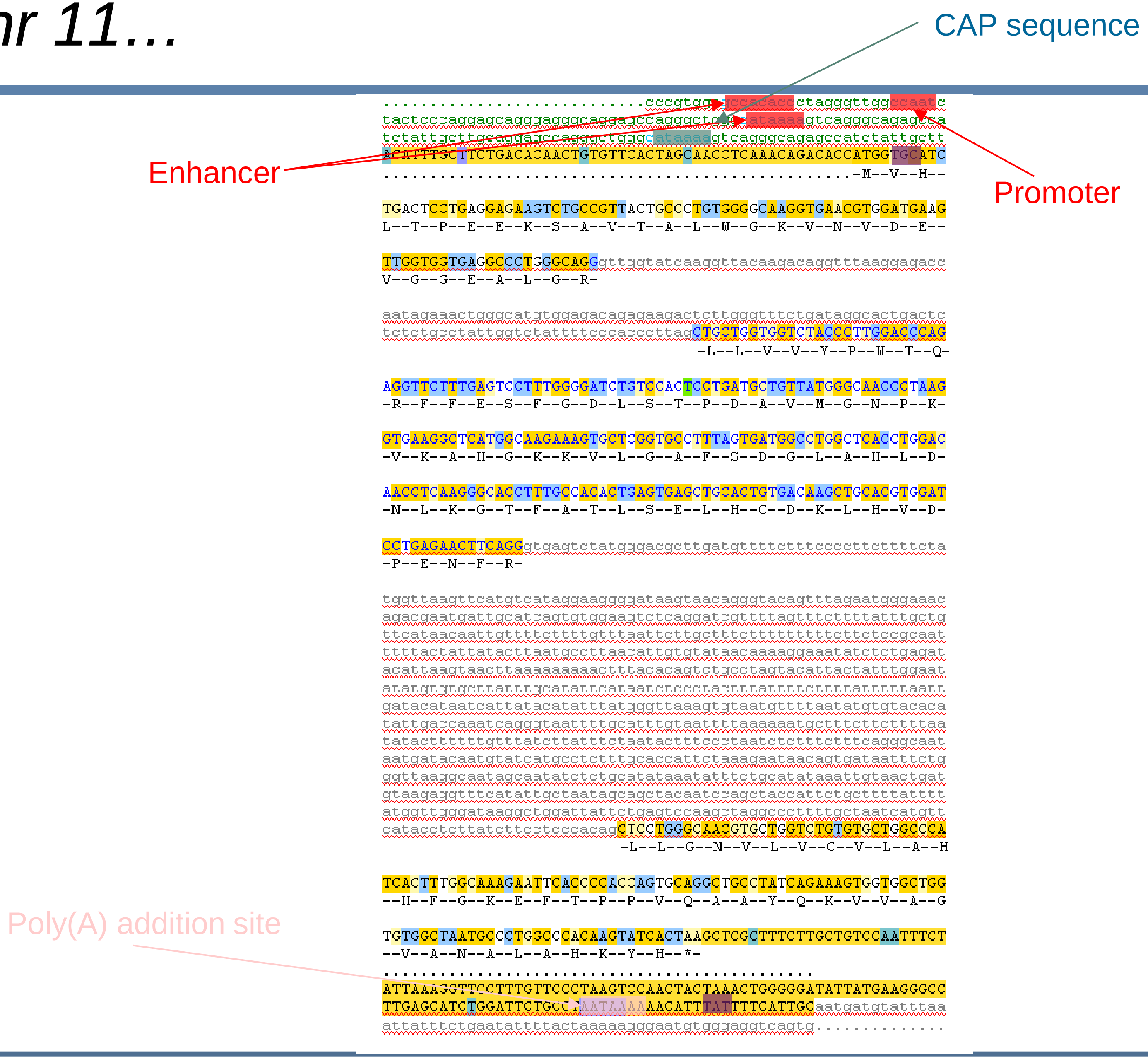


http://www.ncbi.nlm.nih.gov/mapview/

1800 bps in chr 11…

```
............................cccgtggagccacaccctagggttggccaatc
tactcccaggagcagggagggcaggagccagggctgggcataaaagtcagggcagagcca
tctattgcttgcaggagccagggctgggcataaaagtcagggcagagccatctattgctt
ACATTTGCTTCTGACACAACTGTGTTCACTAGCAACCTCAAACAGACACCATGGTGCATC
TGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGTGAACGTGGATGAAG
TTGGTGGTGAGGCCCTGGGCAGGGTTGGTATCAAGGTTACAAGACAGGTTTAAGGAGACC
AATAGAAACTGGGCATGTGGAGACAGAGAAGACTCTTGGGTTTCTGATAGGCACTGACTC
TCTCTGCCTATTGGTCTATTTTCCCACCCTTAGCTGCTGGTGGTCTACCCTTGGACCCAG
AGGTTCTTTGAGTCCTTTGGGGATCTGTCCACTCCTGATGCTGTTATGGGCAACCCTAAG
GTGAAGGCTCATGGCAAGAAAGTGCTCGGTGCCTTTAGTGATGGCCTGGCTCACCTGGAC
AACCTCAAGGGCACCTTTGCCACACTGAGTGAGCTGCACTGTGACAAGCTGCACGTGGAT
CCTGAGAACTTCAGGgtgagtctatgggacgcttgatgttttctttccccttcttttcta
tggttaagttcatgtcataggaaggggataagtaacagggtacagtttagaatgggaaac
agacgaatgattgcatcagtgtggaagtctcaggatcgttttagtttcttttatttgctg
ttcataacaattgttttcttttgtttaattcttgctttctttttttttcttctccgcaat
ttttactattatacttaatgccttaacattgtgtataacaaaaggaaatatctctgagat
acattaagtaacttaaaaaaaaactttacacagtctgcctagtacattactatttggaat
atatgtgtgcttatttgcatattcataatctccctactttattttcttttatttttaatt
gatacataatcattatacatatttatgggttaaagtgtaatgttttaatatgtgtacaca
tattgaccaaatcagggtaattttgcatttgtaattttaaaaaatgctttcttcttttaa
tatacttttttgtttatcttatttctaatactttccctaatctctttctttcagggcaat
aatgatacaatgtatcatgcctctttgcaccattctaaagaataacagtgataatttctg
ggttaaggcaatagcaatatctctgcatataaatatttctgcatataaattgtaactgat
gtaagaggtttcatattgctaatagcagctacaatccagctaccattctgcttttatttt
atggttgggataaggctggattattctgagtccaagctaggcccttttgctaatcatgtt
catacctcttatcttcctcccacagCTCCTGGGCAACGTGCTGGTCTGTGTGCTGGCCCA
TCACTTTGGCAAAGAATTCACCCCACCAGTGCAGGCTGCCTATCAGAAAGTGGTGGCTGG
TGTGGCTAATGCCCTGGCCCACAAGTATCACTAAGCTCGCTTTCTTGCTGTCCAATTTCT
ATTAAAGGTTCCTTTGTTCCCTAAGTCCAACTACTAAACTGGGGGATATTATGAAGGGCC
TTGAGCATCTGGATTCTGCCTAATAAAAAACATTTATTTTCATTGCaatgatgtatttaa
attatttctgaatattttactaaaaagggaatgtgggaggtcagtg..............
```

1800 bps in chr 11…



Allele − A Variation at a Locus

For example in the same locus we may have:

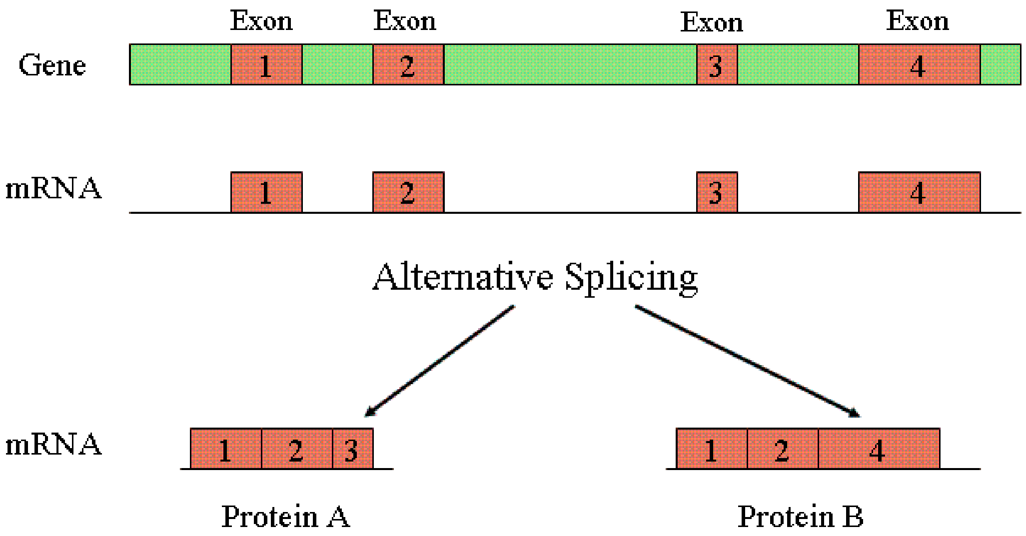
Allele 1: GTTTCTGATTTTTTGATGTCTTCATCCATCACTGTCCTTGTCAAATAGTTT…
Allele 2: GTTTCTGATTTTTTGATGTCTTCAGCCATCACTGTCCTTGTCAAATAGTTT…

allele frequency is the proportion
of a certain allele within a population

Population	Genotype Frequency (%)			Allele Frequency	
	MM	MN	NN	M	N
U.S. whites	29.16	49.38	21.26	0.540	0.460
U.S. blacks	28.42	49.64	21.94	0.532	0.468
U.S. Indians	60.00	35.12	4.88	0.776	0.224
Eskimos (Greenland)	83.48	15.64	0.88	0.913	0.087
Ainus (Japan)	17.86	50.20	31.94	0.430	0.570
Aborigines (Australia)	3.00	29.60	67.40	0.178	0.822

Alternative Splicing

- Alternative splicing is a process by which the exons of the RNA transcript (a primary gene transcript or pre-mRNA) are reconnected in alternative ways during RNA splicing.
- The resulting different mRNAs may be translated into different protein isoforms; thus, a single gene may code for multiple proteins



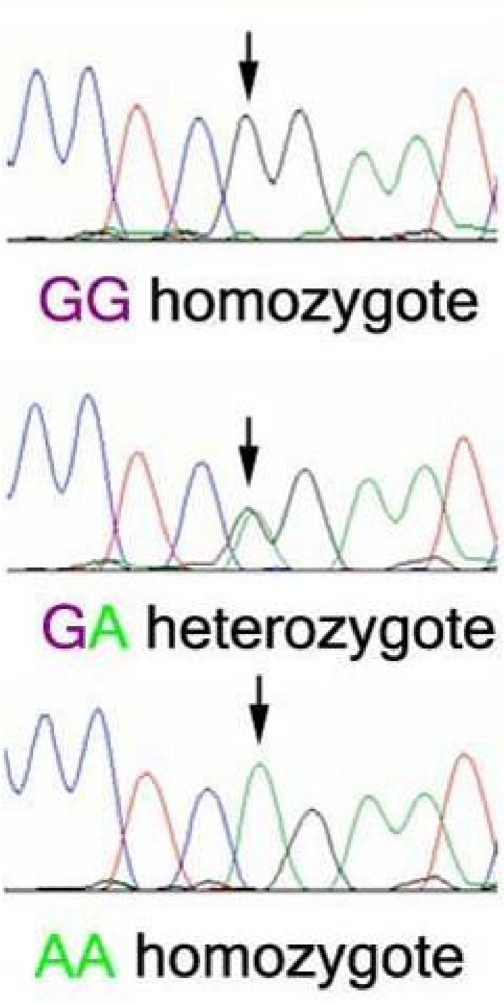
Genetic Markers – A Catalogue

Some commonly used types of genetic markers are:

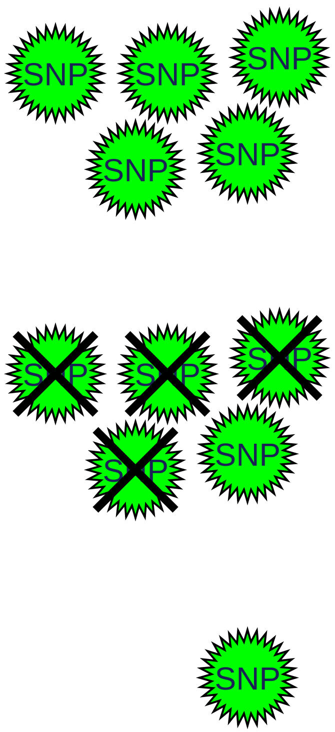
- **SNP (Single Nucleotide Polymorphism)**
- **VNTR (Variable Number of Tandem Repeat)**
- **Microsatellite or STR (Short Tandem Repeat)**
- **CNV (Copy number Variation)**

Single Nucleotide Polymorphism (SNP)

- A **single nucleotide polymorphism** (**SNP**)**,** is a single nucletide DNA sequence variation: **A**, **T**, **C**, or **G**
- In this case we say that there are two *alleles*: G and A. Almost all common SNPs have only two alleles
- For a variation to be considered a SNP, it must occur in at least **1%** of the population.



Origin of SNPs

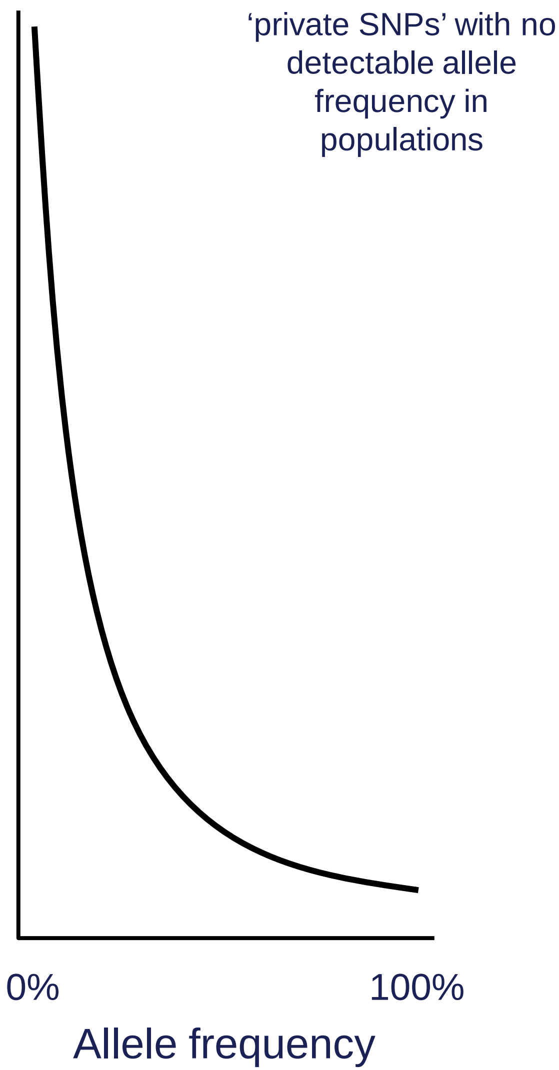


Appearance of new variants by mutation

Survival of alleles through early generations against the odds

Increase of the allele to a substantial population frequency

Fixation of the allele in populations



Adapted from Bioinformatics for Geneticists, Eds Barnes and Gray

Categories of SNPs

GAG >GAA
Glu > Glu

Synonymous
→ no change in amino acid

GAG >GGG
Glu > Gly

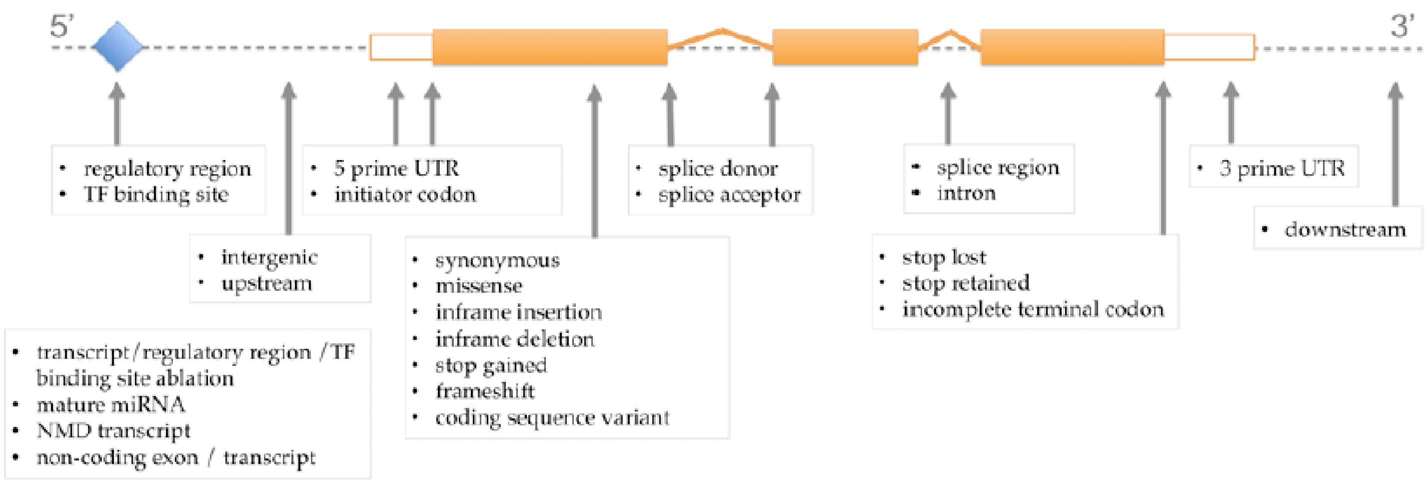
Missense/ Non-synonymous
→ change in amino acid

GAG >TAG
Glu > STOP

Nonsense
→ introduces a STOP codon

Consequence Types of Sequence Variants

Type	Change	Consequence
Non-synonymous or nonsense SNPs in coding areas	Alters the function and/or structure of the encoded protein	Cause of most monogenic disorders: Hemochromatosis (*HFE*), Cystic fibrosis (*CFTR*), Hemophilia (*F8*)
Synonymous SNPs in coding areas	No change in amino acid sequence of the protein	May alter splicing
Non-coding	Promoter or regulatory regions	May affect the level, location or timing of gene expression
Non-coding		No direct known impact on phenotype Useful as markers



Mutation or Polymorphism?

- **Mutation**: change in a DNA sequence.
  - Normal allele that is prevalent in the population
  - Mutation changes this to a rare and abnormal variant
- **Polymorphism**: change in a DNA sequence common in the population.
  - No single allele is the standard sequence
  - There are two or more equally acceptable alternatives
  - Arbitrary cut-off point between a mutation and a polymorphism is 1%

    $< 1\%$ = Mutation

    $> 1\%$ = Polymorphism

The Genetic Basis for Human Variation

Class of variation	Rules for assigning allele to class	Example	Frequency
Single Nucleotide Polymorphism (SNP)	Single base substitution involving A,T,C, or G	A/T	5,692,700 (~93%)
Deletion/Insertion Polymorphisms (DIPs)	Designated using the full sequence of the insertion as one allele, and either a fully defined string for the variant allele or a “-” character to specify the deleted allele.	T/-CCTA/G	431,319 (~7%)
Microsatellite or short tandem repeat (STR)	Alleles are designated by providing the repeat motif and the copy number for each allele.	(CAC)8/9/10/11	2,440 (0.04%)
Named variant	Applies to insertion/deletion polymorphisms of longer sequence features, such as retroposon .dimorphism for Alu or line elements	(alu) / -	1,859 (0.03%)

Data derived from dbSNP

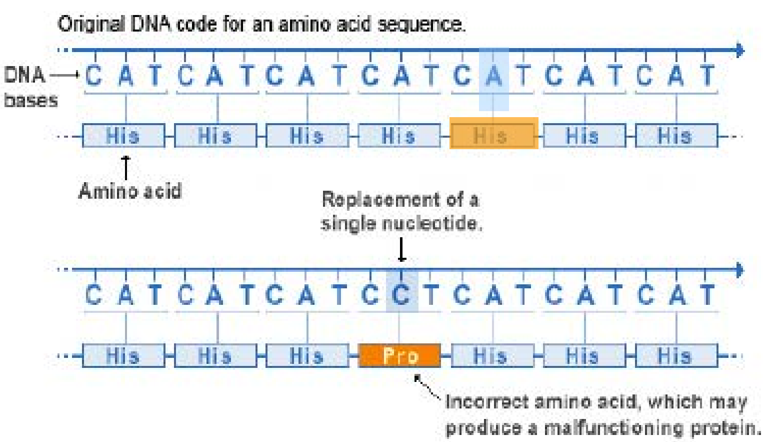
Genetic Markers - Introduction

- A **genetic marker** is a known DNA sequence. It can be described as a variation, which may arise due to mutation or alteration in the genomic loci
- Genetic markers can be used to study:
  - The relationship between an inherited disease and its genetic cause
  - How humans develop diseases and respond to pathogens, chemicals, drugs, vaccines, and other agents
- Genetic markers have to be associated with a specific locus, and highly polymorphic, because homozygotes do not provide any information on possible genetic differences.

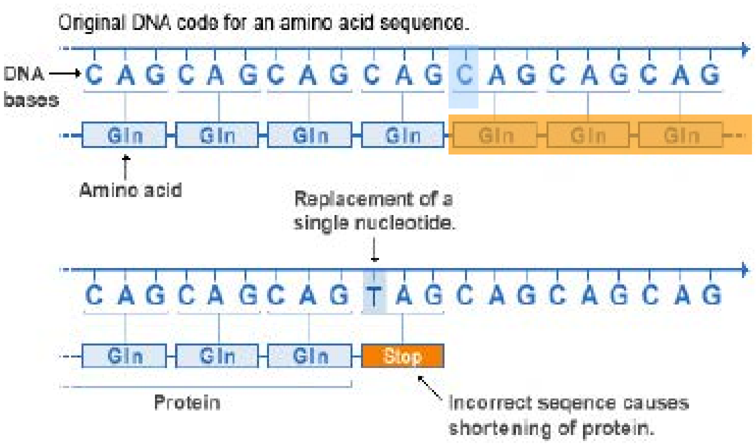
Point Mutations

- Most common forms of mutation
- Missense mutations can lead to changes in protein function (detrimental or beneficial)
- Nonsense mutation almost invariable lead to protein dysfunction)

Missense mutations

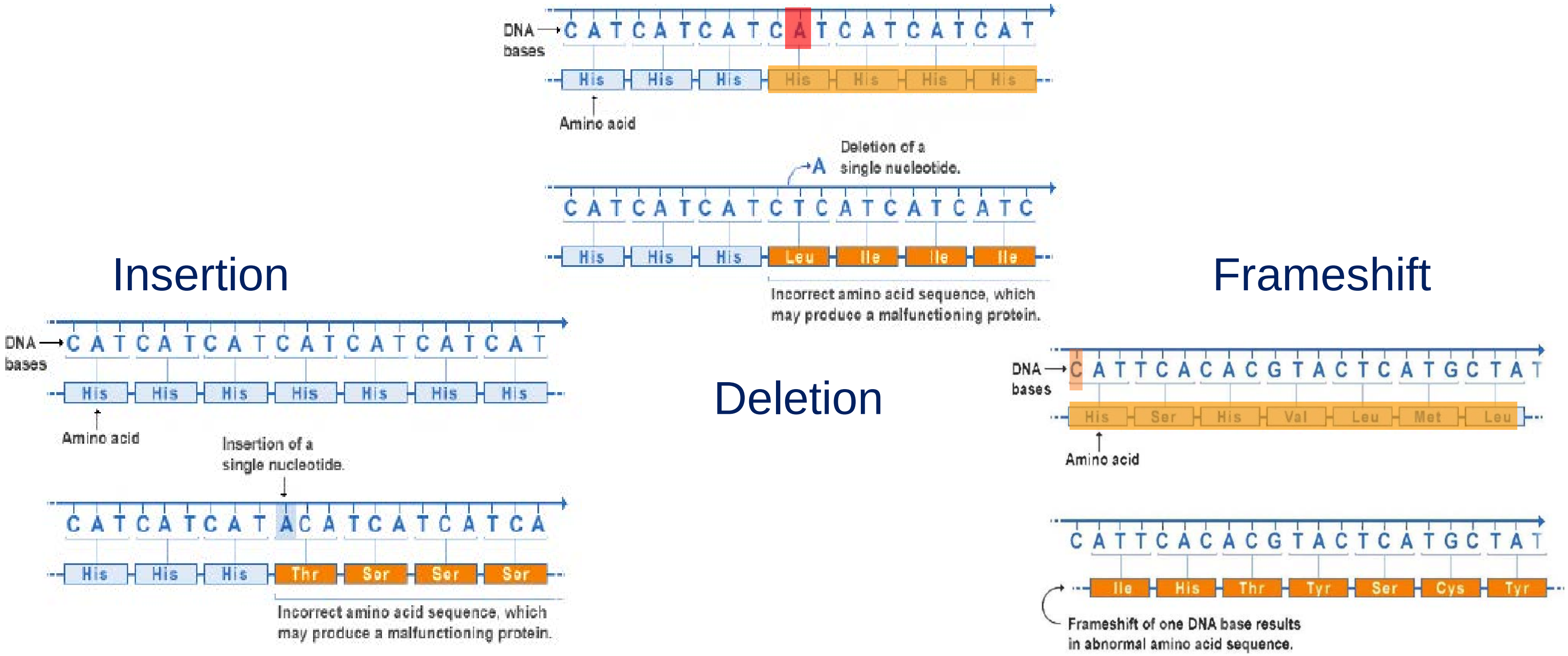


Nonsense mutation



Mutations: Insertion, Deletion and Frameshift

- Insertions and deletions are often more **deleterious** than missense mutations
- Insertion or deletion of 1 or 2 nucleotides will lead to a **frameshift** (change of ORF)



CNV are Ubiquitous in the Human Genome

Summary Statistics

Stat	Merged-level	Sample-level
CNVs:	353126	3024212
Inversions:	1645	2335

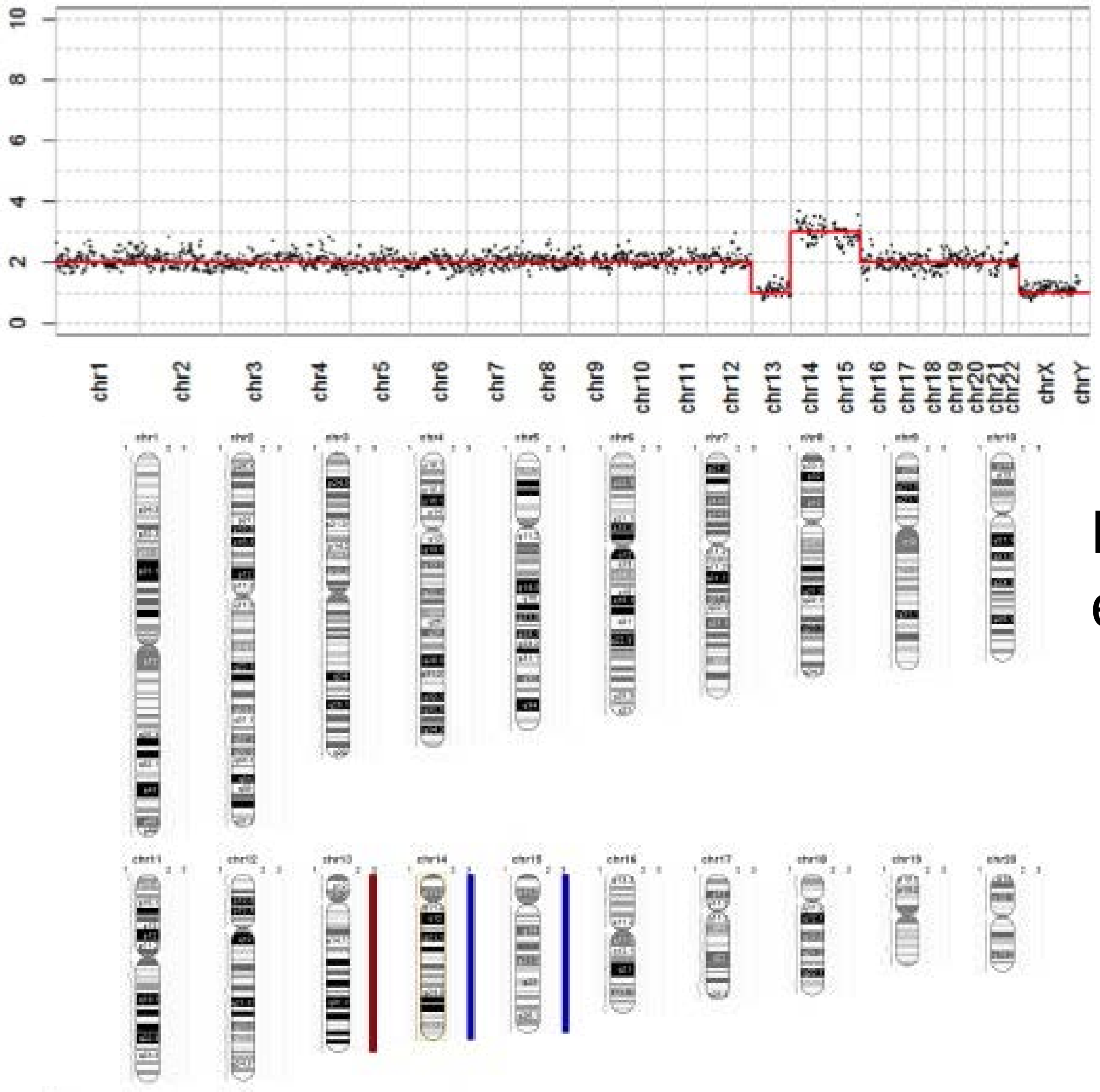
Number of Studies: 62

The number of genome structural variants (>1 kb) that distinguish genomes of different individuals is at least on the order of 600-900 per individual

J.O. Korbel *et al.* (2007) *Science* **318**(5894):420
R Redon *et al*. (2006) Nature **444**(7118):444

ThermoFisher SCIENTIFIC

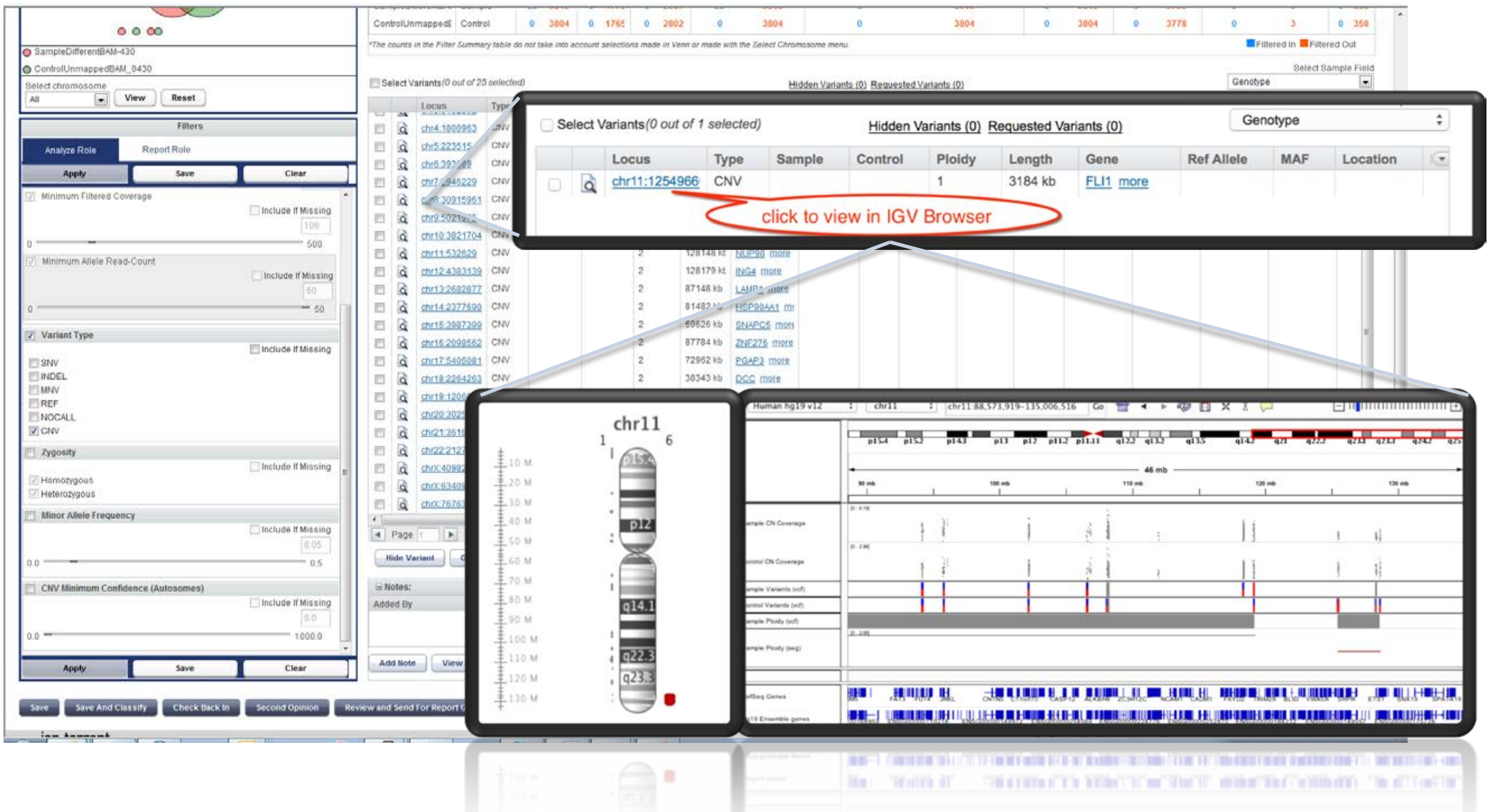
CNVs



Need internal control to establish baseline

- Paired sample analyses enable you to detect SNPs, indels, and CNVs in one analysis
- CNVs ≥ 100Kb

Visualising SNP, Indels and CNV



Genotyping: Practical Applications

Clinical

- Disease diagnosis
- Biomarker discovery
- Pharmacogenomics

Research

- Association studies
- Population genetics and evolutionary studies
- Marker-assisted breeding

Variant Frequency Sources

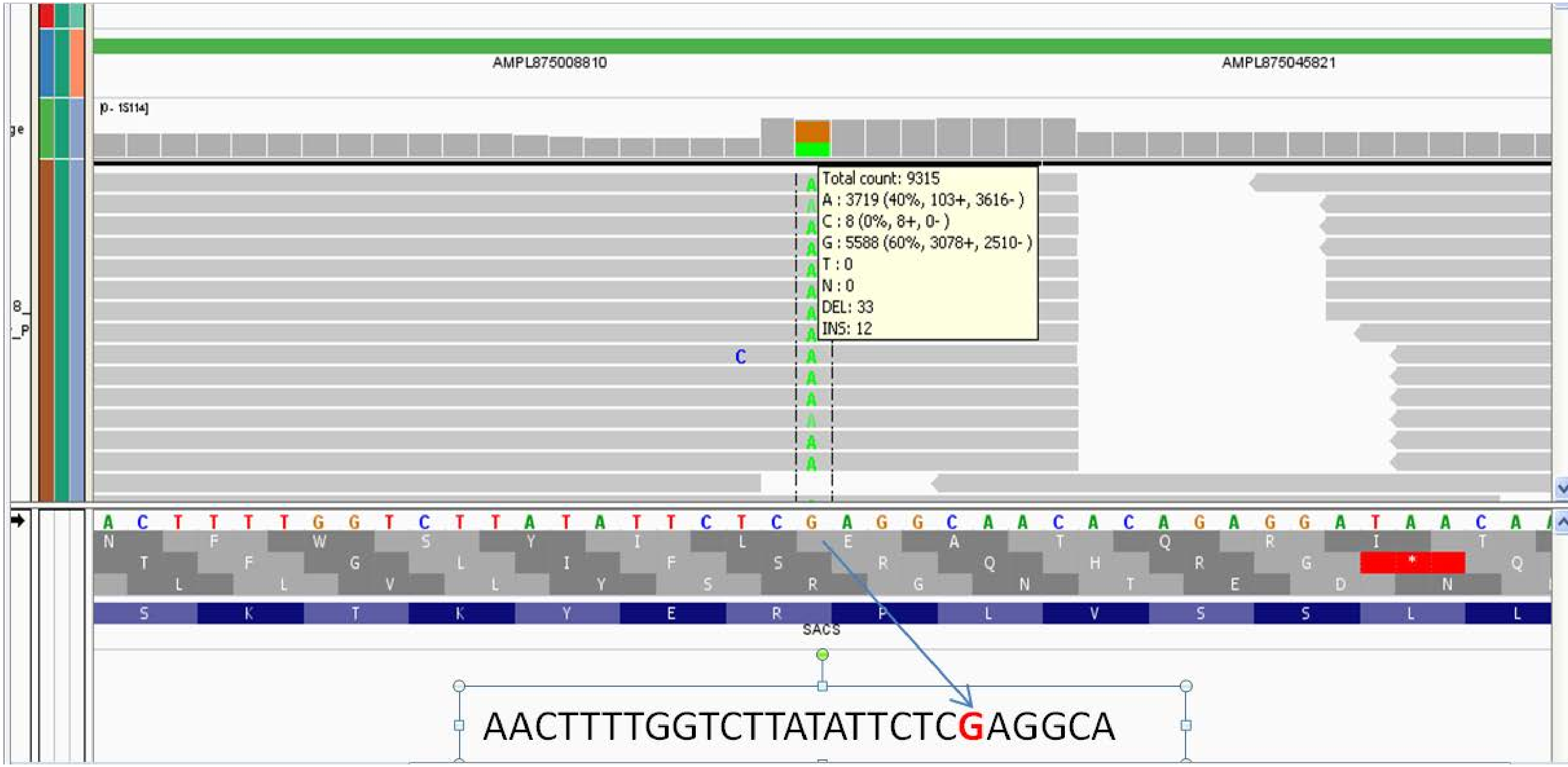
- **dbSNP –** Largest dataset, but “polluted”
- **1000 Genomes –** Frequencies available, but cell lines
- **Exome Sequencing Project –** No indels, patients, no validation
- **Published studies –** GoNL, Complete Genomics
- ***In house* databases –** Population/sequencing specific variants

Candidate Mutation

Affected individual

Control

Heterozygous

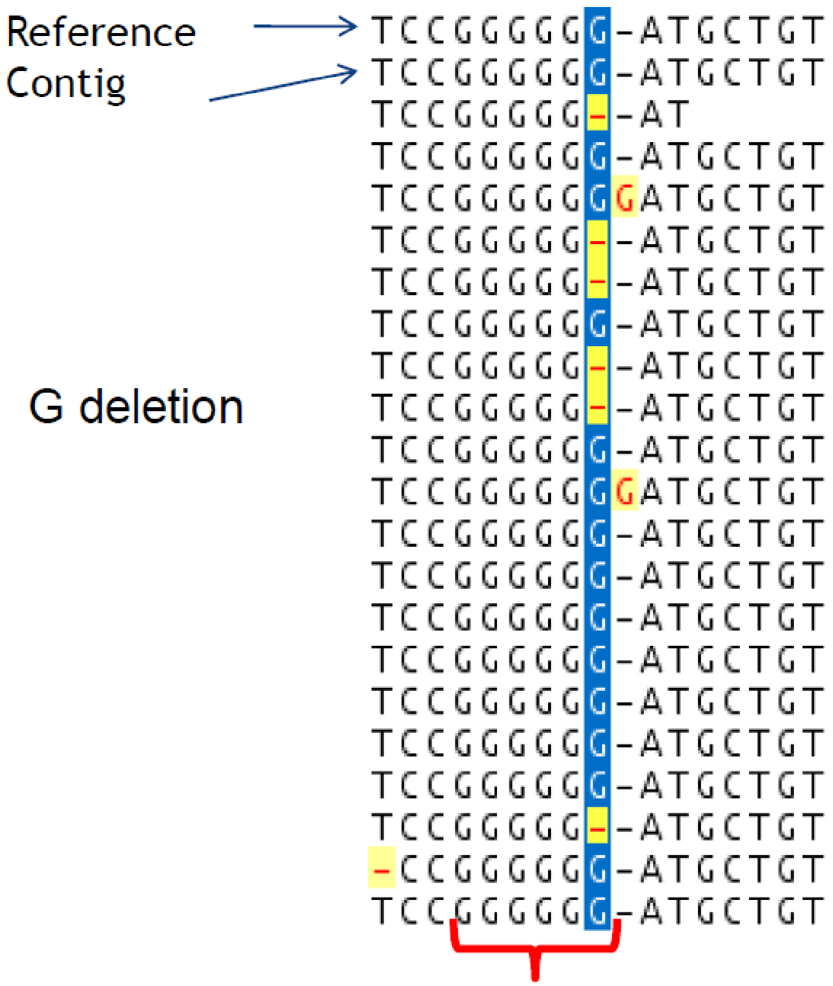


Variants Detection

T>A change

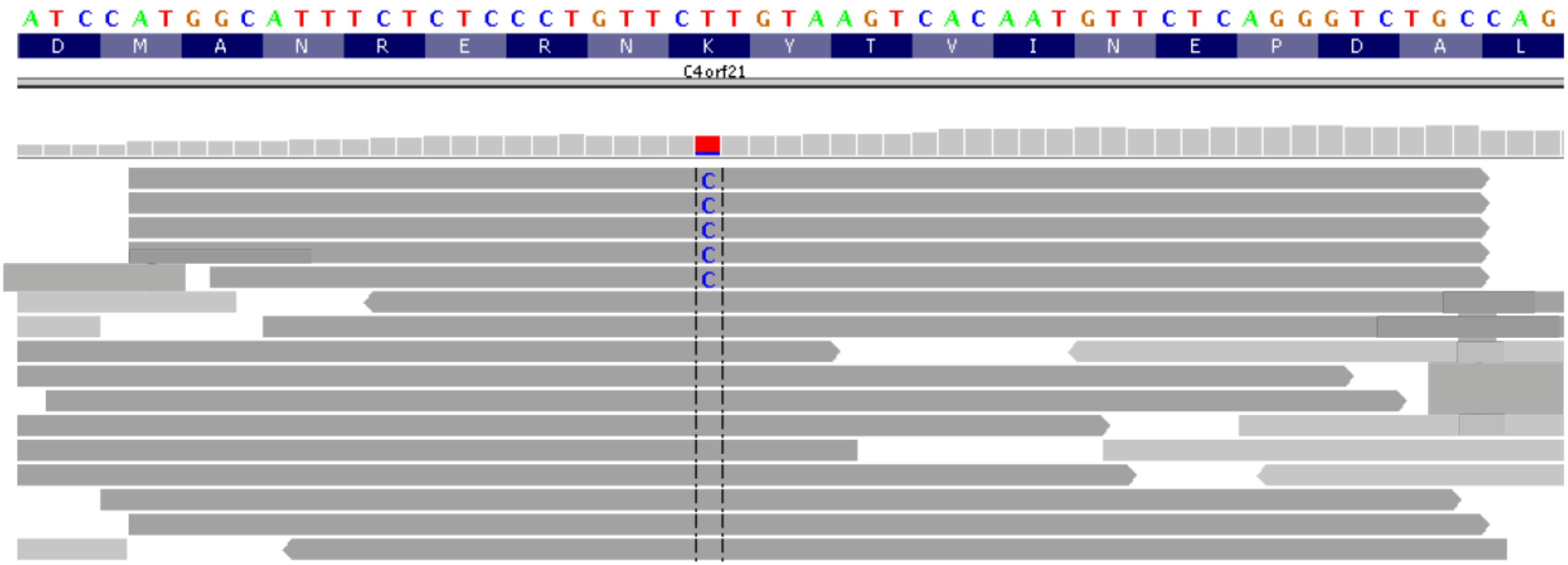
False positive due to an accumulation of errors

Variants Detection



False positive due to a homopolymer stretch

Variants Detection

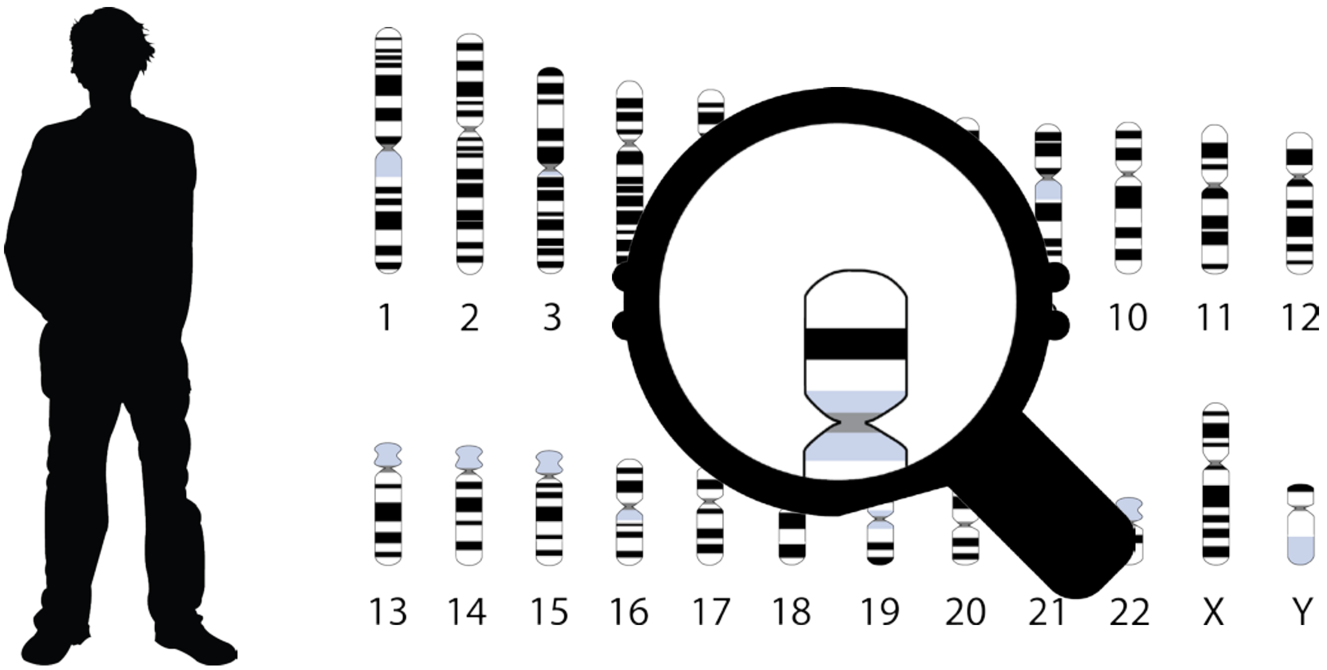


False positive due to a PCR amplification artifact

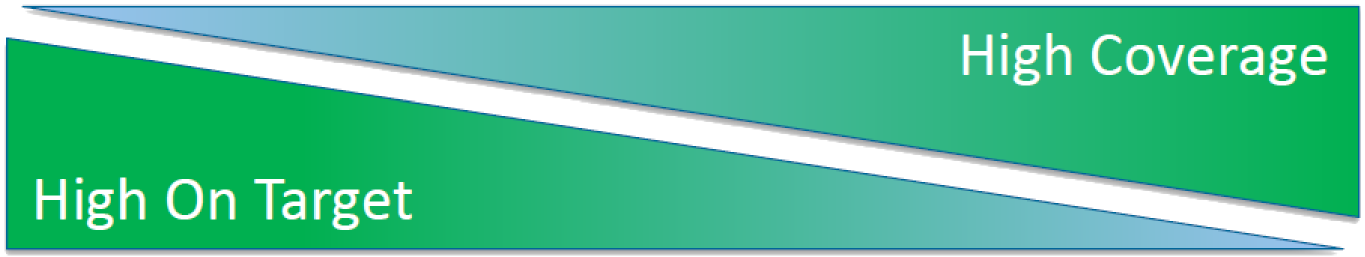
Panel Design

What is Targeted Sequencing?

- **Targeted Sequencing** isolates and focuses your sequencing on specific genes or genomic regions of interest rather than surveying the whole genome



Choosing the Correct Design

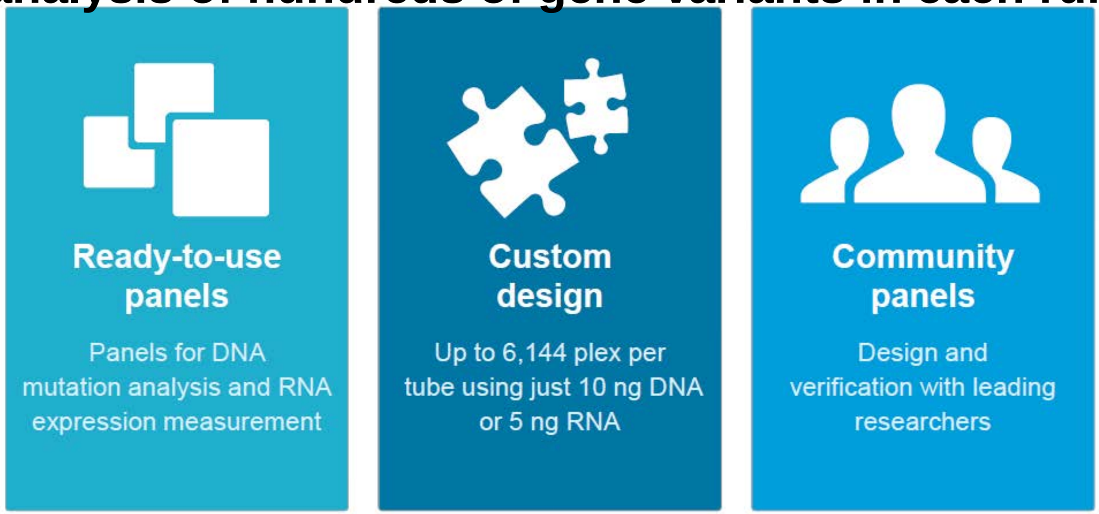


Primer Specificity	High Specificity (Recommended)	Medium Specificity	Relaxed Specificity
Design Criteria	Only high specificity primers	Slight reduction in primer specificity that will increase coverage and may increase off-target	More relaxed primer specificity that will increase coverage with greater off-target risk
How to choose?	If the *in silico* coverage is sufficient this design is the best option	For better coverage of GC-rich regions with slight increase in off-target	For more coverage of GC-rich regions and tolerance for higher off-target

Ion AmpliSeq™ Panels

Ultrahigh-multiplex PCR for targeted sequencing and expression

Highly multiplexed PCR for NGS library preparation enables the analysis of hundreds of gene variants in each run



"The chief advantage of these [hybridisation] methods is their ability to

capture large target regions in a single experiment,

more rapidly and conveniently than PCR. To capture the entire 30 Mb human exome, for example, would require at least

6,000 separate PCRs,

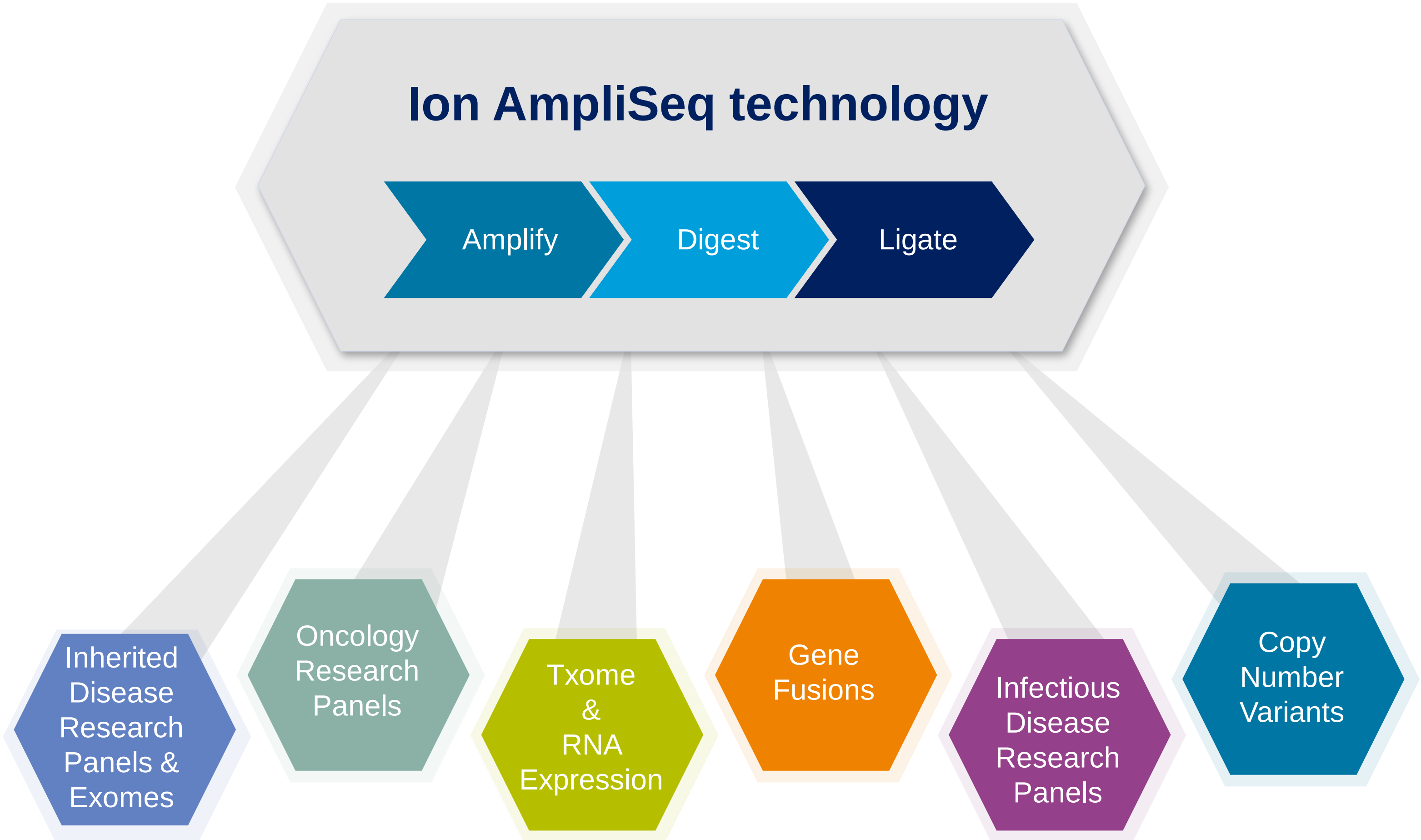
each of which would need to be optimised, the products would need to be normalised, and a total of around

120 µg of genomic DNA

would be required for the experiment."

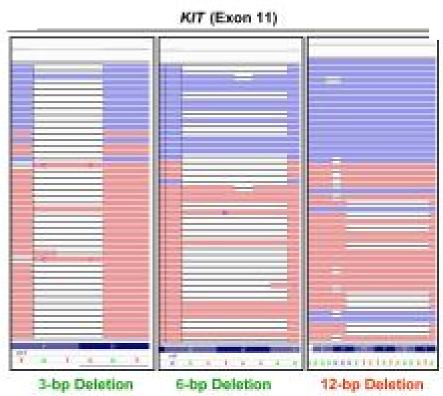
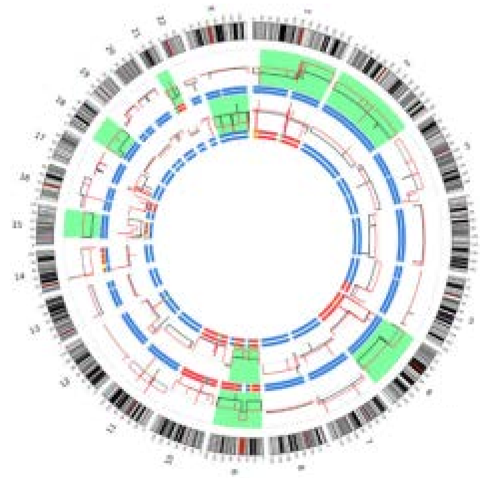
Nature Methods 2010

Broad Applications with Ion AmpliSeq Technology



Why Targeted Sequencing

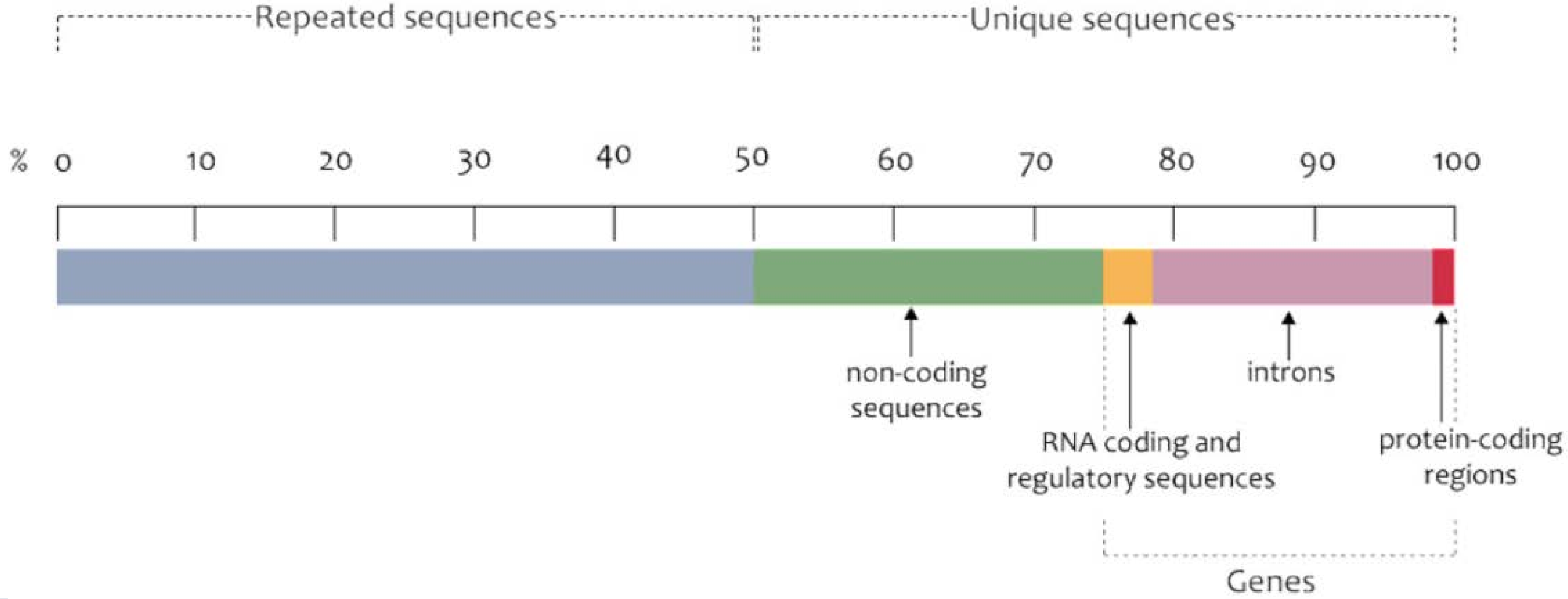
More cost effective, more time efficient and simpler to analyse

	Targeted Sequencing	Whole Exome	Whole Genome
			
Variants generated per run	**10s to 100s**	**~50,000**	**~3,000,000**
Likely number of variants for follow-up	**1-10s**	**1-10s**	**1-10s**
Time to analyze	**Hours to Days**	**Days to Weeks**	**Weeks to Months**
Total cost including analysis	**£**	**££**	**£££**

ThermoFisher SCIENTIFIC

What's the Exome?

- Exons are short sequences of DNA representing the protein coding regions in the genome
- It is estimated that there are about 180,000 exons arranged in 22,509 genes in the human genome
- The protein-coding exons comprise about 35-40Mb or ~1.3-1.5% of the human genome

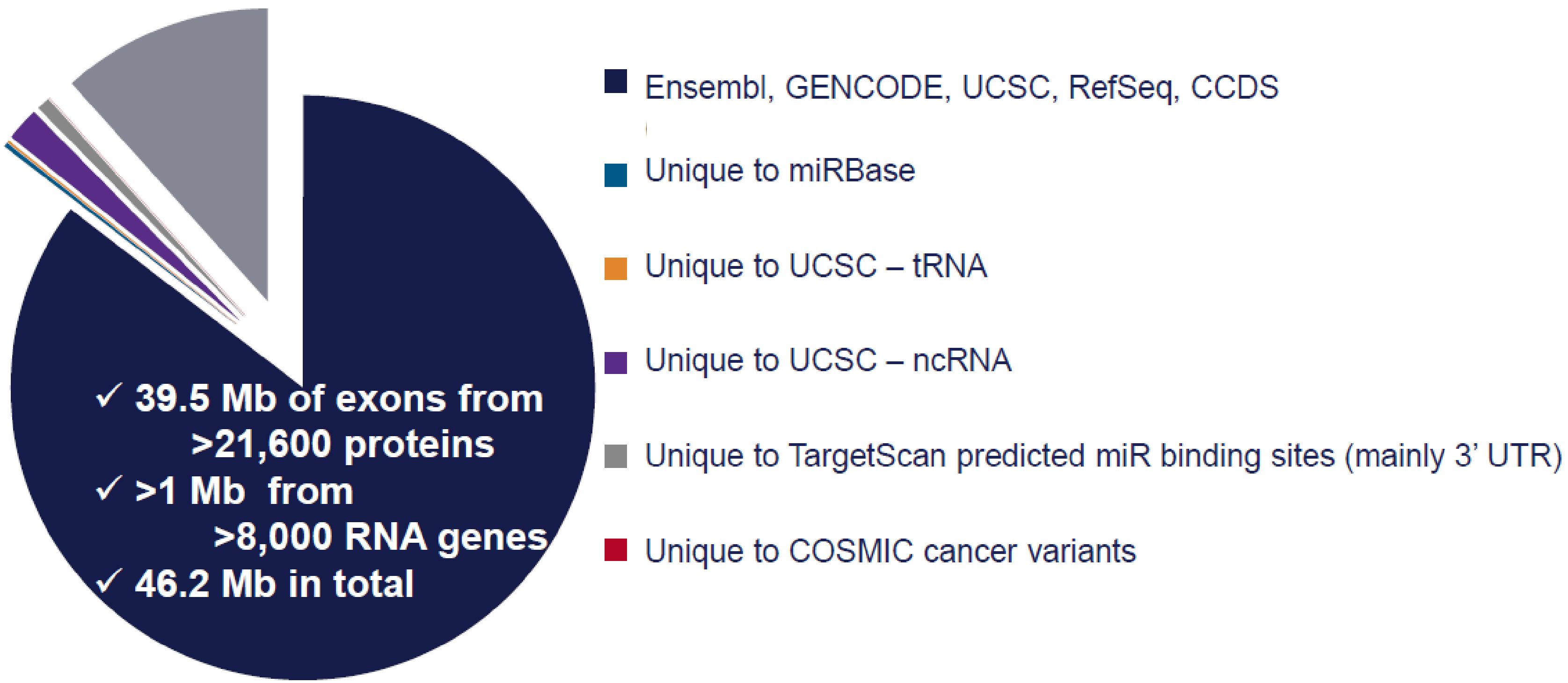


How is the Exome Defined in a Kit?

- The exome consists of all the exons of a genome that are transcribed into mature RNA.
- Multiple databases are used to derive exome content *e.g.* RefSeq, CCDS, Ensembl, GENCODE, etc.
- Content in the databases differ due to number of non-coding RNA's and start and end positions of transcripts
- An "exome kit" consists of a pool of oligos designed to hybridise or amplify the regions of interest

ThermoFisher SCIENTIFIC

Focus on Function

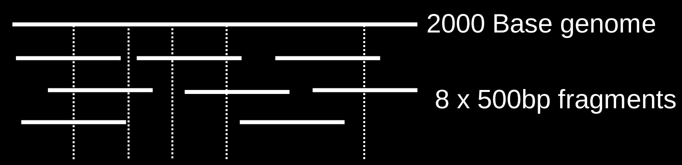


Key Metrics for Exome Sequencing

Metric	Description
% Reads On Target	"*More is better*" Fraction of total sequencing reads that uniquely align to the targeted regions as a percentage of all bases generated (*i.e.* 95% on-target)
Coverage Uniformity	"*More is better*" Fraction of total aligned bases within 0.2x of mean coverage; high uniformity comes from a lack of peaks and troughs of coverage, which can arise from bias
Sequencing Coverage	"*More is better*" The number of times an individual DNA base in the target region is covered via bases in mapped sequencing reads. This is usually expressed as "*X-fold*".
SNP Genotype Concordance	"*More is better*" Interrogation of a given sample and measuring % concordance of microarray SNP data with exome sequencing data

What is Coverage?

- The average number of reads representing a given nucleotide in the reconstructed sequence
- Enables you to estimate the % of the genome covered by reads
- High coverage overcomes errors in base-calling & assembly



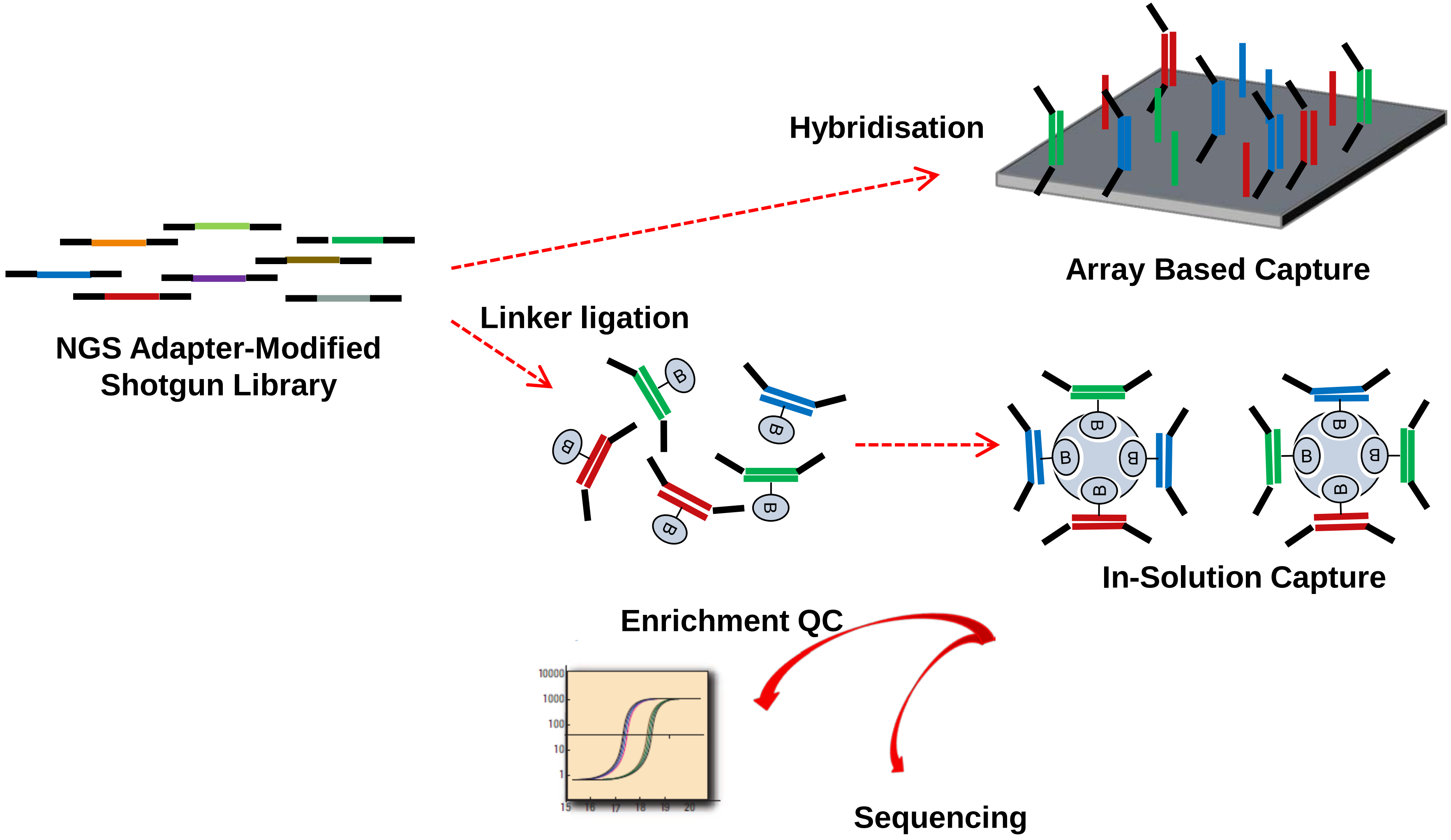
$$Coverage = N * \frac{L}{G}$$

$$Coverage = 8 * \frac{500}{2000} = \mathbf{2x}$$

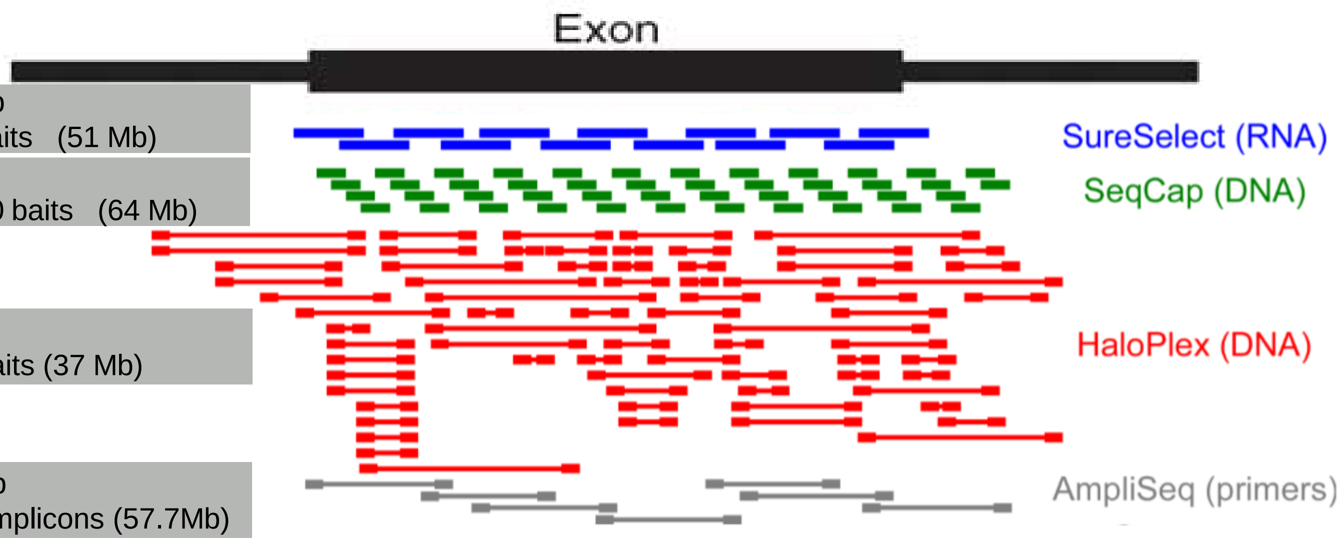
N=Number of reads
L=average read length
G=length of original genome

The typical desired coverage of a genome is 30x

Target Enrichment Strategies



Exome Enrichment



Length:	114-126 bp
Quantity:	655,872 baits (51 Mb)

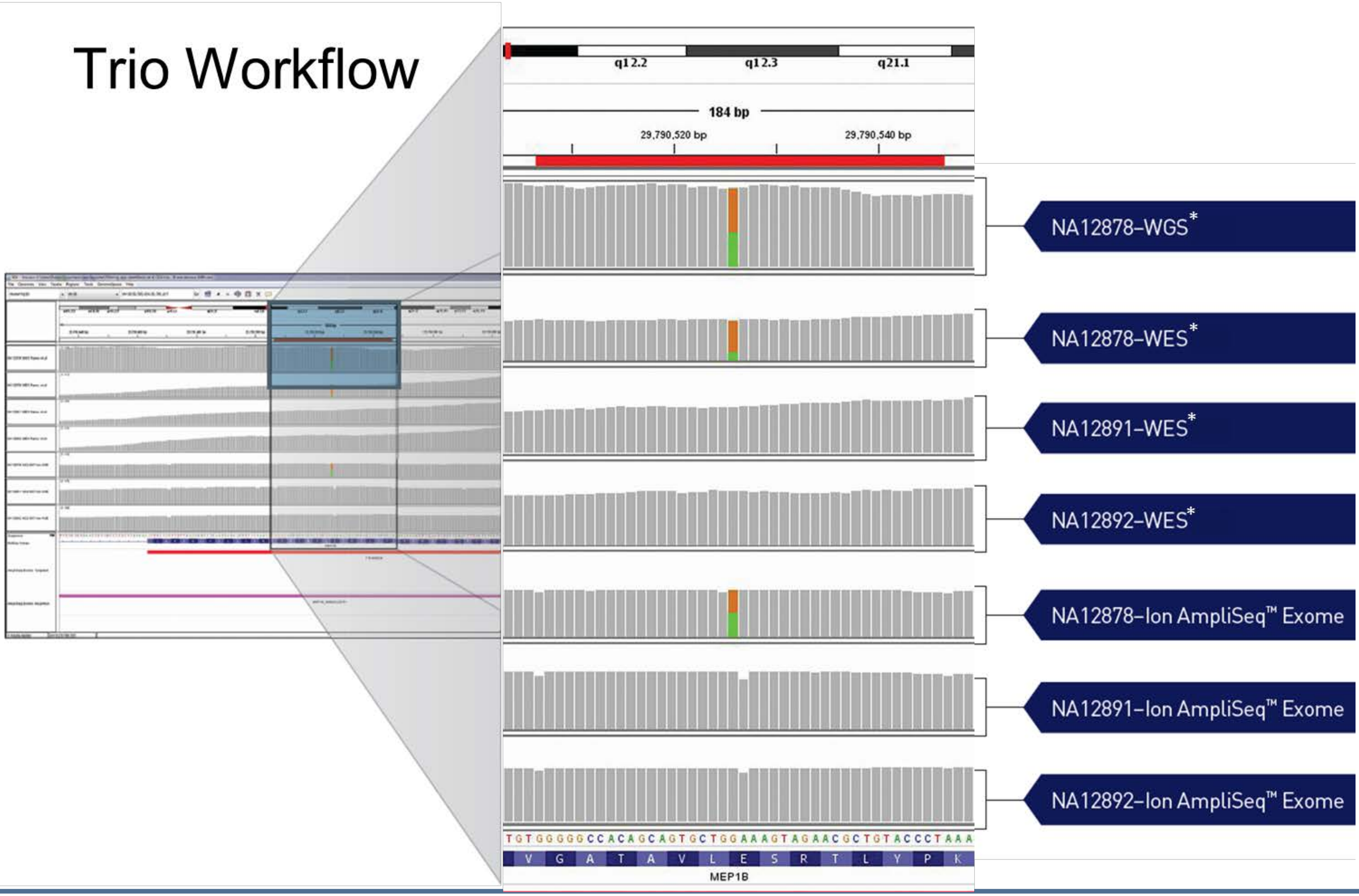
Length:	55-105 bp
Quantity:	>2,100,000 baits (64 Mb)

Lengths:	95 bp
Quantity:	340,427 baits (37 Mb)

Lengths:	225-275 bp
Quantity:	293,903 amplicons (57.7Mb)

Samorodnitsky *et al*. *Human Mutation* (2015)

Trio Workflow



*A. Ramu *et al*. *Nat Methods* (2013) **10**:985

Why Sequence the Exome?

Focus and Utility

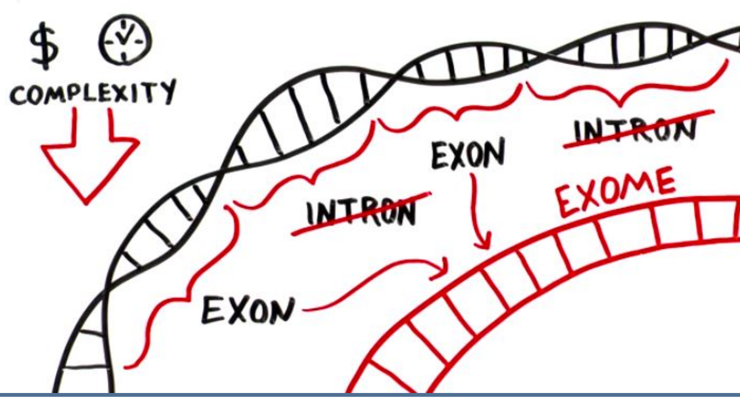
- So far, ~85% of human disease causing variants in exonic regions
- Practical to study rare Mendelian diseases with limited number of samples
- Simple sample preparation

Cost

- Generally 10% to 20% of Whole Genome Sequencing
- Typically can sequence 6-8 exomes for the price of one genome

Data:

- Fewer *Variants of Unknown Significance* to report (or not report)
- Faster sequencing times



Precision Medicine

Genomics-Driven Medicine

Personalised medicine is the most exciting change in cancer treatment since the invention of chemotherapy[1]

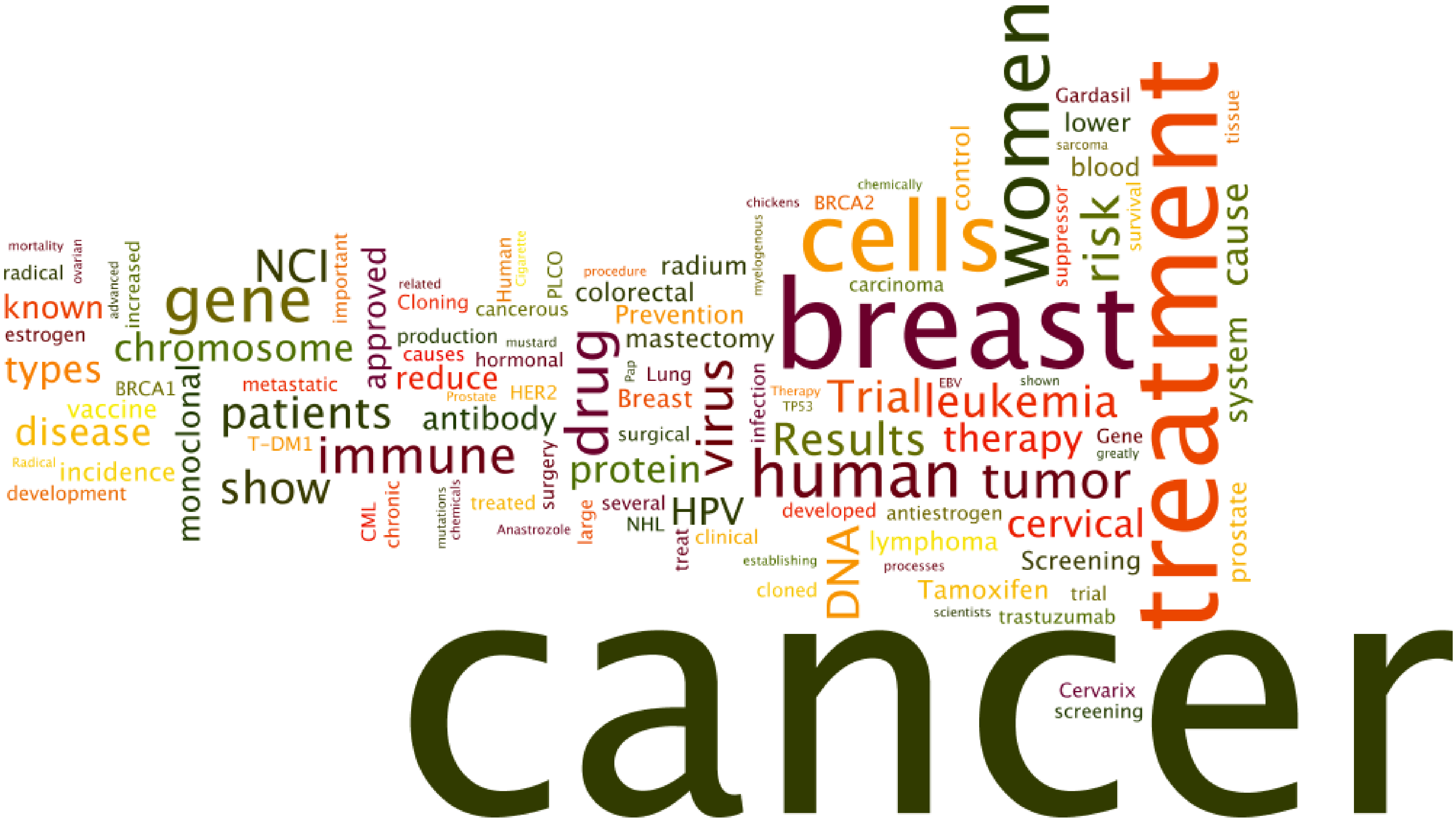
As a growing number of biomarkers become clinically actionable, the single-diagnostic/single-drug paradigm is becoming a challenge to manage[2]

Sources:

[1]Peter Johnson, CRUK, http://www.cancerresearchuk.org/support-us/donate/become-a-major-donor/how-you-can-give/the-catalyst-club/personalised-medicine

[2]Delivering precision medicine in oncology today and in future − the promise and challenges of personalised cancer medicine: a position paper by the European Society for Medical Oncology (2014) doi:10.1093/annonc/mdu217

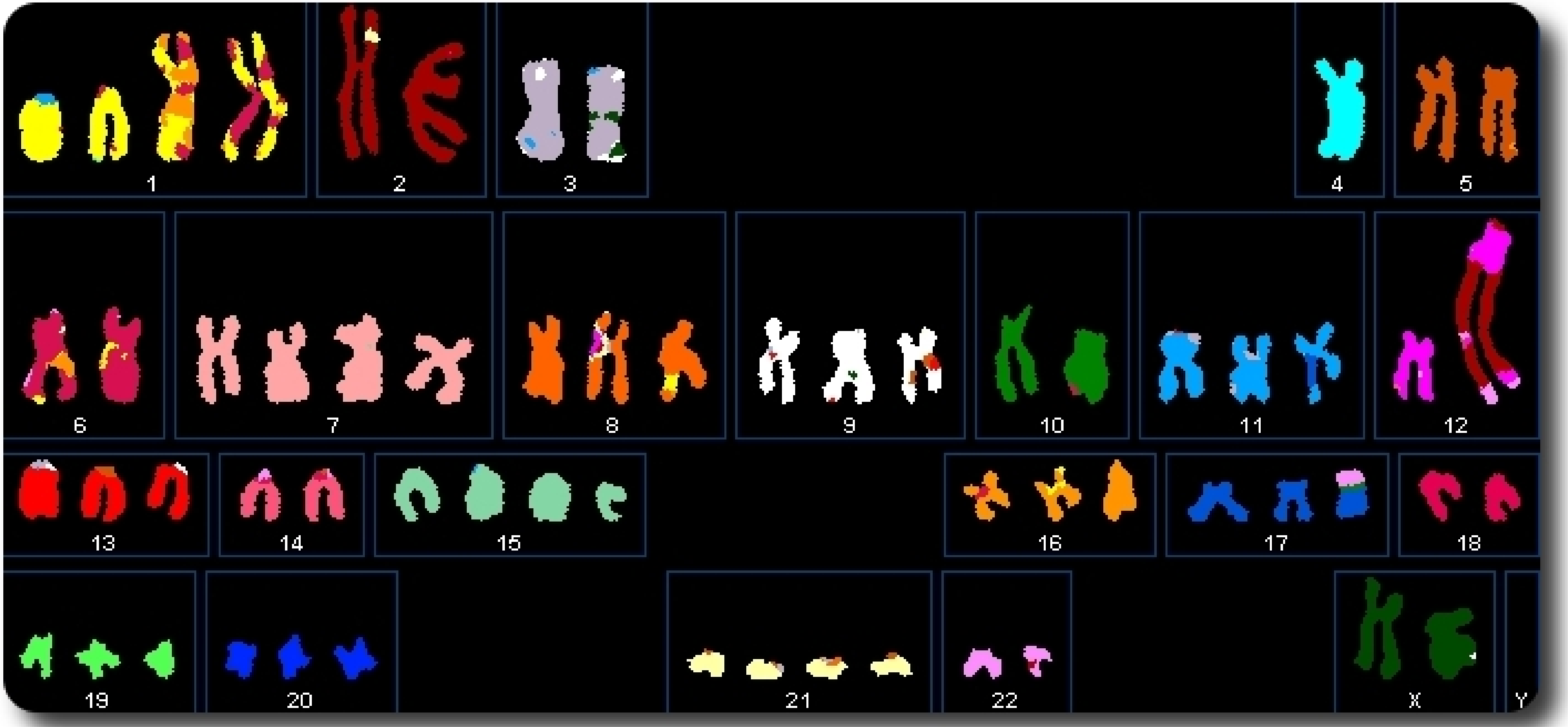
Advances Against Cancer – 250 years



Tag cloud generated from *250 Years of Advances Against Cancer*, NCI 2015

ThermoFisher SCIENTIFIC

Cancer



Cancer

Acute Lymphoblastic Leukaemia; Acute Myeloid Leukaemia; Adrenocortical Carcinoma; AIDS-Related Cancers; AIDS-Related Lymphoma; Anal Cancer; Appendix Cancer; Astrocytoma Cerebellar; Astrocytoma Cerebral; Atypical Teratoid; Atypical Teratoid/Rhabdoid Tumor; Basal Cell Carcinoma; Bile Duct Cancer; Bladder Cancer; Bone Cancer, Osteosarcoma and Malignant Fibrous Histiocytoma; Brain Stem Glioma; Brain Tumor; Breast Cancer; Bronchial Tumors; Burkitt Lymphoma; Carcinoid Tumor; Carcinoma of Unknown Primary; Central Nervous System Atypical Teratoid/Rhabdoid Tumor; Central Nervous System Lymphoma, Primary; Cerebellar Astrocytoma; Cerebral Astrocytoma/Malignant Glioma; Cervical Cancer; Childhood Cancers; Chordoma; Chronic Lymphocytic Leukaemia; Chronic Myelogenous Leukaemia; Chronic Myeloproliferative Disorders; Colon Cancer; Colorectal Cancer; Craniopharyngioma; Cutaneous T-Cell Lymphoma; Embryonal Tumors; Endometrial Cancer; Ependymoblastoma; Ependymoma; Esophageal Cancer; Ewing Family of Tumors; Extracranial Germ Cell Tumor; Extragonadal Germ Cell Tumor; Extrahepatic Bile Duct Cancer; Gallbladder Cancer

Gastric (Stomach) Cancer; Gastrointestinal Carcinoid Tumor; Gastrointestinal Stromal Cell Tumor; Gastrointestinal Stromal Tumor; Germ Cell Tumor, Extracranial; Germ Cell Tumor, Extragonadal; Germ Cell Tumor, Ovarian; Gestational Trophoblastic Tumor; Glioma; Glioma Brain Stem; Glioma Cerebral Astrocytoma; Glioma Visual Pathway and Hypothalamic; Hairy Cell Leukaemia; Head and Neck Cancer; Hepatocellular (Liver) Cancer (Primary); Histiocytosis; Hodgkin Lymphoma; Hypopharyngeal Cancer; Hypothalamic and Visual Pathway Glioma; Intraocular Melanoma; Islet Cell Tumors; Kaposi Sarcoma; Kidney (Renal Cell) Cancer; Kidney Cancer; Langerhans Cell Histiocytosis; Laryngeal Cancer; Leukaemia, Acute Lymphoblastic; Leukaemia, Acute Myeloid; Leukaemia, Chronic Lymphocytic; Leukaemia, Chronic Myelogenous; Leukaemia, Hairy Cell; Lip Cancer; Liver Cancer (Primary); Lung Cancer, Non-Small Cell; Lung Cancer, Small Cell; Lymphoma, AIDS-Related; Lymphoma, Burkitt; Lymphoma, Cutaneous T-Cell, see Mycosis Fungoides and Sézary Syndrome; Lymphoma, Hodgkin; Lymphoma, Non-Hodgkin; Lymphoma, Primary Central Nervous System; Macroglobulinemia, Waldenström; Malignant Fibrous Histiocytoma of Bone and Osteosarcoma; Medulloblastoma; Medulloepithelioma; Melanoma; Merkel Cell Carcinoma; Mesothelioma; Metastatic Squamous Neck Cancer with Occult Primary; Mouth Cancer; Multiple Endocrine Neoplasia Syndrome; Multiple Myeloma/Plasma Cell Neoplasm; Mycosis Fungoides; Myelodysplastic Syndromes; Myelogenous Leukaemia, Chronic; Myeloid Leukaemia Acute; Myeloma, Multiple; Myeloproliferative Disorders, Chronic; Nasal Cavity and Paranasal Sinus Cancer; Nasopharyngeal Cancer; Neuroblastoma; Non-Hodgkin Lymphoma; Non-Small Cell Lung Cancer; Oral Cancer; Oropharyngeal Cancer; Osteosarcoma and Malignant Fibrous Histiocytoma of Bone; Ovarian Cancer; Ovarian Epithelial Cancer; Ovarian Germ Cell Tumor; Ovarian Low Malignant Potential Tumor; Pancreatic Cancer, Pancreatic Cancer, Islet Cell Tumors; Papillomatosis, Paranasal Sinus and Nasal Cavity Cancer; Parathyroid Cancer, Penile Cancer; Pharyngeal Cancer, Pheochromocytoma; Pineal Parenchymal Tumors of Intermediate Differentiation; Pineoblastoma, Pituitary Tumor; Plasma Cell Neoplasm/Multiple Myeloma, Pleuropulmonary Blastoma; Pregnancy and Breast Cancer, Primary Central Nervous System Lymphoma; Prostate Cancer, Rectal Cancer; Renal Cell (Kidney) Cancer, Renal Pelvis and Ureter, Transitional Cell Cancer; Respiratory Tract Carcinoma Involving the NUT Gene on Chromosome 15; Retinoblastoma, Rhabdomyosarcoma, Salivary Gland Cancer; Sarcoma, Ewing Family of Tumors, Sarcoma, Kaposi; Sarcoma, Soft Tissue; Sarcoma, Uterine; Sézary Syndrome; Skin Cancer (Melanoma); Skin Cancer (Nonmelanoma); Skin Carcinoma, Merkel Cell; Small Cell Lung Cancer; Small Intestine Cancer; Soft Tissue Sarcoma; Spinal Cord Tumors; Squamous Cell Carcinoma, see Skin Cancer (Nonmelanoma); Squamous Neck Cancer with Occult Primary, Metastatic; Stomach (Gastric) Cancer; Supratentorial Primitive Neuroectodermal Tumors; T-Cell Lymphoma, Cutaneous, see Mycosis Fungoides and Sézary Syndrome; Testicular Cancer; Throat Cancer; Thymoma and Thymic Carcinoma; Thyroid Cancer; Transitional Cell Cancer of the Renal Pelvis and Ureter; Trophoblastic Tumor, Gestational; Unusual Cancers of Childhood; Ureter and Renal Pelvis, Transitional Cell Cancer; Uterine Cancer, Endometrial; Uterine Sarcoma; Vaginal Cancer; Visual Pathway and Hypothalamic Glioma; Vulvar Cancer; Waldenström Macroglobulinemia; Wilms Tumor

Where we are Today

Every year, 14M cancer cases diagnosed with 8.2M deaths worldwide[1]

13.1M projected cancer-related deaths by 2030[2]

$39.3B oncology therapeutics market for 2013[3]

$5.3B will be spent in next-gen cancer diagnostics by 2015[4]

Sources:
1 Cancer Fact Sheet. World Health Organization, Updated February 2014. Annals of Oncology 25
2 Cancer Fact Sheet. World Health Organization, Updated: January 2013.
3 Data extrapolated based on the following report: “Oncology Therapeutics Market to 2017." GBI Research, December 2011.
4 “Advanced Next Generation Cancer Diagnostic Devices Market 2012 – 2018.” Transparency Market Research, 2012.

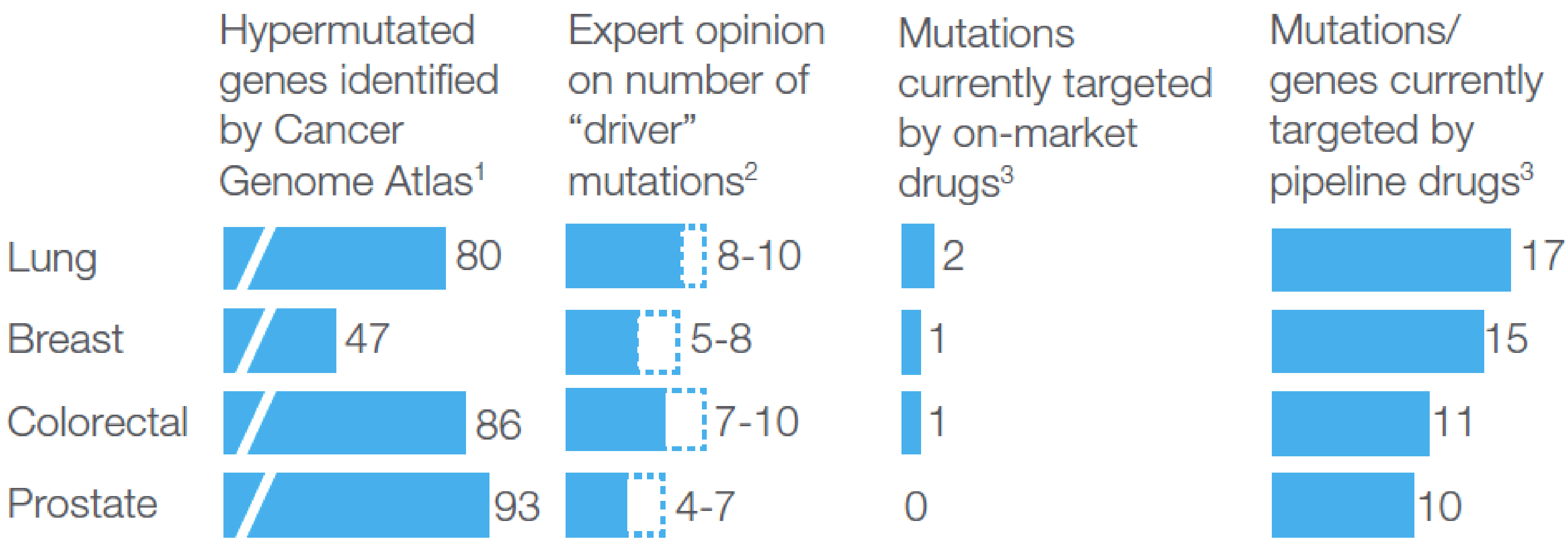
Routine Biomarker Analysis

Today's Challenges

- Limited sample material for analysis
- Multiple biomarkers for one disease indication
- Need to test multiple biomarkers simultaneously
- Growing number of biomarkers to address

Current methods to test samples against various biomarkers are slow and require more tumour sample than available

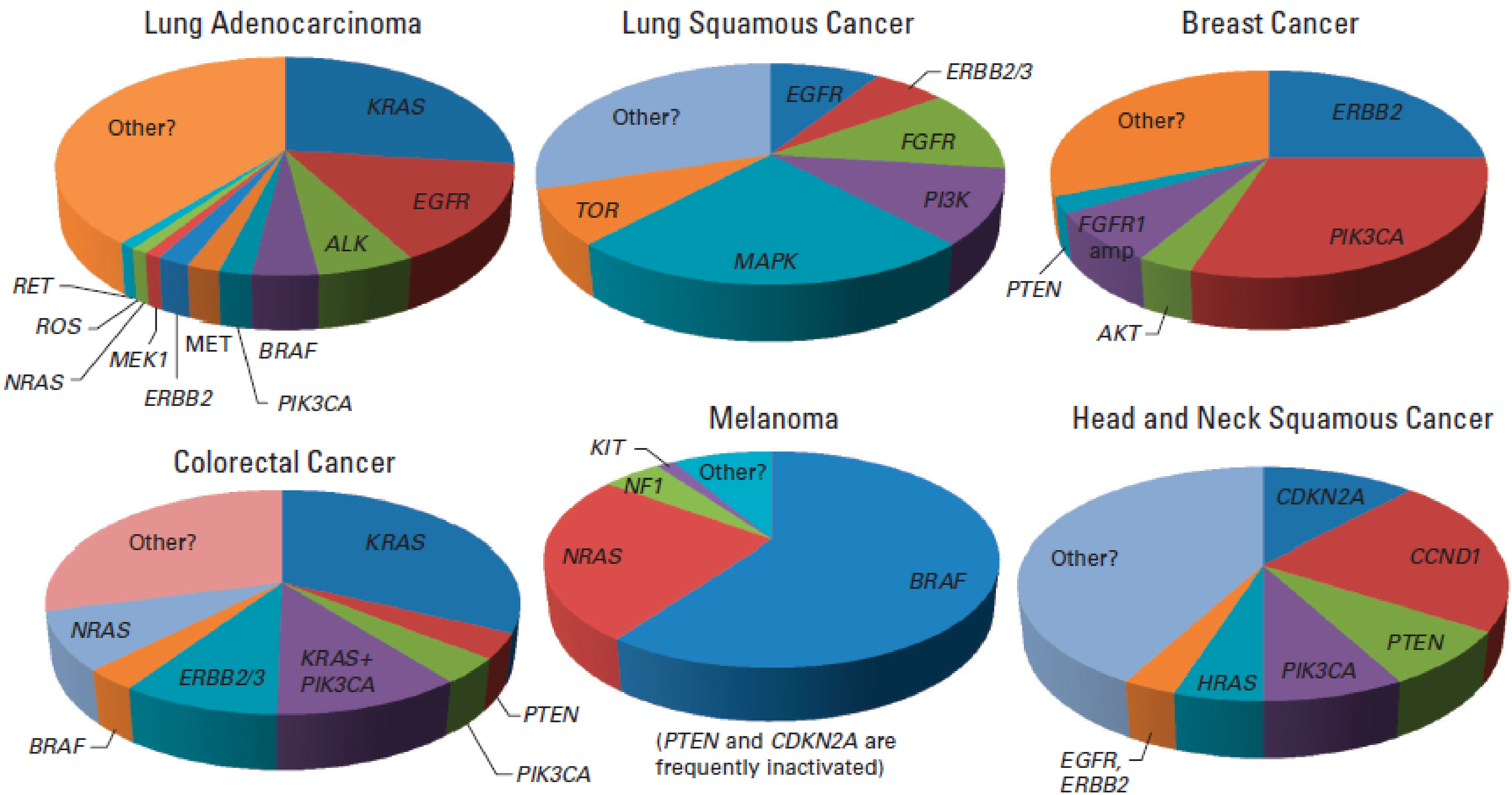
Ever-Expanding List of Genes and Variants



[1] Based on q-value analysis using MutSig software from the Broad Institute
[2] Based on expert interviews
[3] Based on Evaluate Pharmaceuticals database; for pipeline, includes Phase 1 and above only.

Personalized Medicine: The Path Forward, McKinsey & Company, 2013

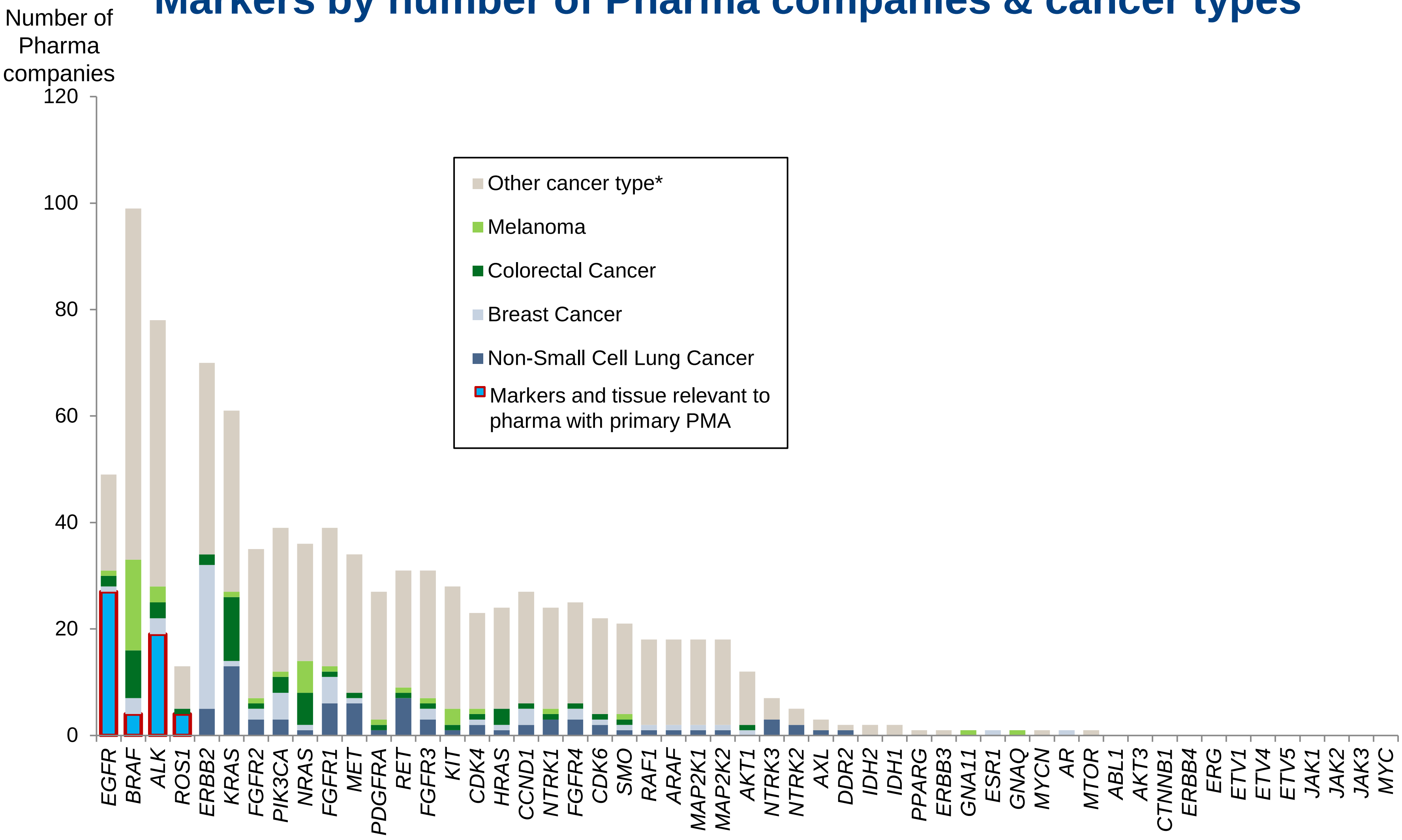
Actionable Signalling Pathways



Garraway *et al.* (2003) *J Clini Oncol* **31(15)**:1803

Ever-Expanding List of Genes and Variants

Markers by number of Pharma companies & cancer types



* Other cancer types include but not limited to gastric, oesophageal, thyroid, glioblastoma, etc.;

ThermoFisher SCIENTIFIC

Example: More Alterations Are Under Investigation

Alteration	Indication	Investigational drug(s)
AKT1 mutation	Multiple	MK-2206, MSC-2363318A
CCND1 amplification	Multiple	palbociclib
CDK4 amplification, mutation	Melanoma, NSCLC	palbociclib
CDK6 amplification	NSCLC	palbociclib
DDR2 mutation	Multiple	crizotinib + dasatinib
KRAS mutation	Multiple	various MEKi combinations
ERBB3 mutation	Multiple	neratinib
FGFR1-4 mutation, amplification, fusion	Multiple	BGJ-398, JNJ-42756493
GNA11 mutation	Melanoma	vorinostat
GNAQ mutation	Melanoma	vorinostat
HRAS mutation	Multiple	binimetinib + panitumumab, BVD-523
IDH1 mutation	Multiple	AG-120
KIT amplification	Melanoma	dasatinib
NRAS mutation	Multiple	various MEKi combinations
MET mutation	Multiple	AMG-337, crizotinib, INCB-028060
MTOR mutation	Multiple	MSC-2363318A
MYCN amplification	Multiple	GSK-525762
PDGFRA amplification	Glioblastoma	nilotinib, sorafenib
PIK3CA mutation	Multiple	various PI3K pathway combinations
PPARG fusion	Thyroid Cancer	pioglitazone
PTCH1 mutation	Multiple	vismodegib
RET mutation	NSCLC, Thyroid Cancer	ponatinib, sunitinib
SMO mutation	Multiple	vismodegib
STK11 mutation	Multiple	MSC-2363318A

Current State of Cancer Drug Development

Pharma

- Discovery of novel biomarkers and target validation
- To find appropriate patients in the right numbers to power clinical trials
- The ability to accelerate clinical trials
- Seamless transitions from biomarker development to CDx
- Top pharm expect to increase investments in analytic capabilities designed to turn data into knowledge[3]
- Top 15 pharm spend 4% of total R&D spend on biomarker research[2]

1,300 active drug projects for 45 cancer indications

321 compounds in Phase II trials

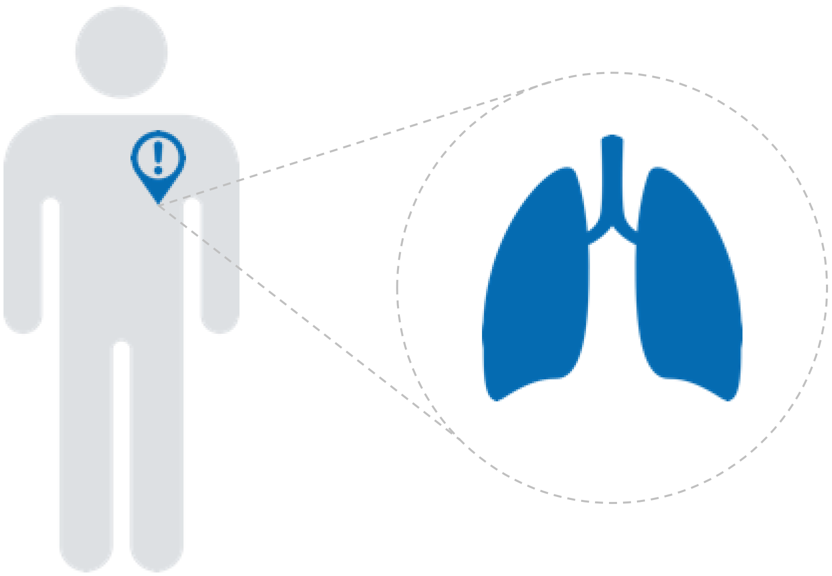
82 therapies in Phase III clinical trials

Sources:
[1] Companion Diagnostics in Personalized Medicine and Cancer Therapy – TriMark Publications, Feb. 2012
[2] Personalized Medicine: The Path Forward – McKinsey & Company (2013)
[3] FierceBiotechIT – Big Pharma OpX, Mar. 2014

Changing the Paradigm − Cancer as a Molecular Disease

From Anatomical to Molecular Approach

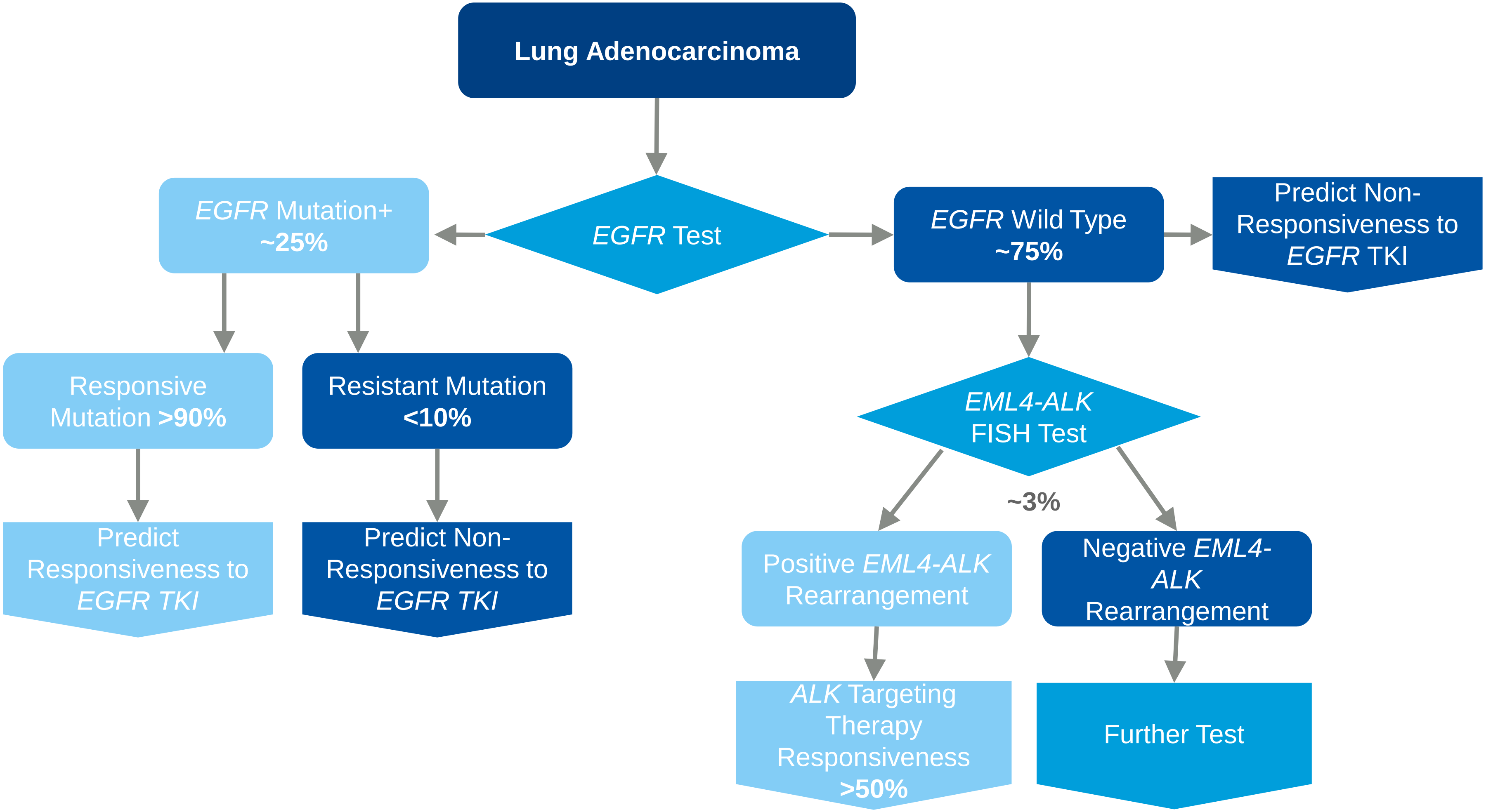
- Breast Cancer
- Cervical Cancer
- Colorectal Cancer
- Liver Cancer
- Lung Cancer
- Ovarian Cancer
- Pancreatic Cancer
- Prostate Cancer
- Skin Cancer
- Stomach Cancer
- Thyroid Cancer



- *ALK*
- *AKT1*
- *BRAF*
- *EGFR*
- *ERBB2*
- *KRAS*
- *NRAS*
- *MAP2K1*
- *PIK3CA*
- *RET*
- *ROS1*
- Undefined

Source: *Nature Medicine*, volume 18, number 3, March 2012

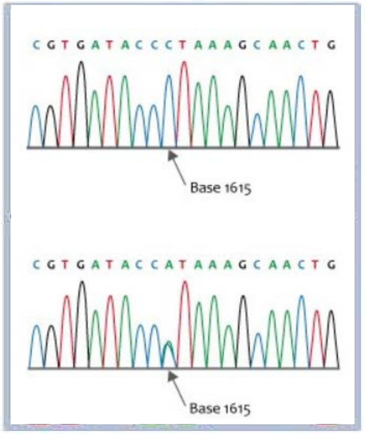
Sequential Testing – Lung Adenocarcinoma Example



1. Engstrom PF, Bloom MG, Demetri GD, *et al*. (2011) NCCN molecular testing white paper: effectiveness, efficiency, and reimbursement. *J Natl Compr Canc Netw*. 9 (6):S1-S16.
2. Dacic S. (2011) Molecular diagnostics of lung carcinomas. *Arch Pathol Lab Med*.135(5):622-629.
3. Wolff AC, Hammond ME, Schwartz JN, *et al*. (2007) American Society of Clinical Oncology/College of American Pathologists guideline recommendations for human epidermal growth factor receptor 2 testing in breast cancer. *Arch Pathol Lab Med*. 131(1):18-43.

ThermoFisher SCIENTIFIC

Genetic Variations and Cancer



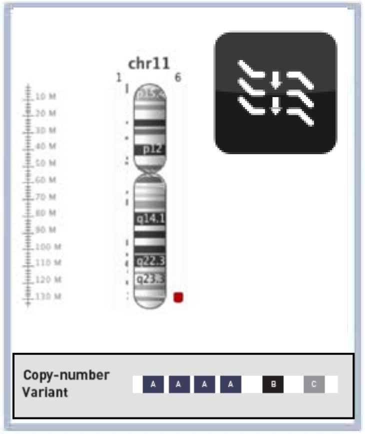
Mutations

AGCTCGTTGCTC
Reference genome

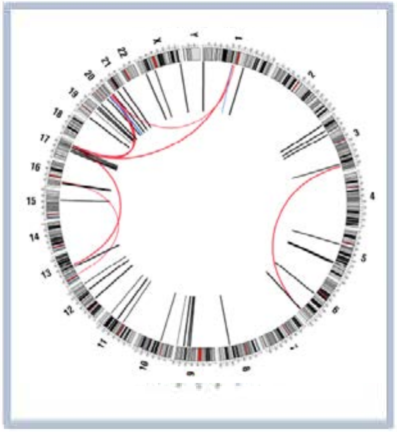
AGCTCGTGTTGCTC
Insertion

AGCTC---GCTC
Deletion

Indels



Copy number variation

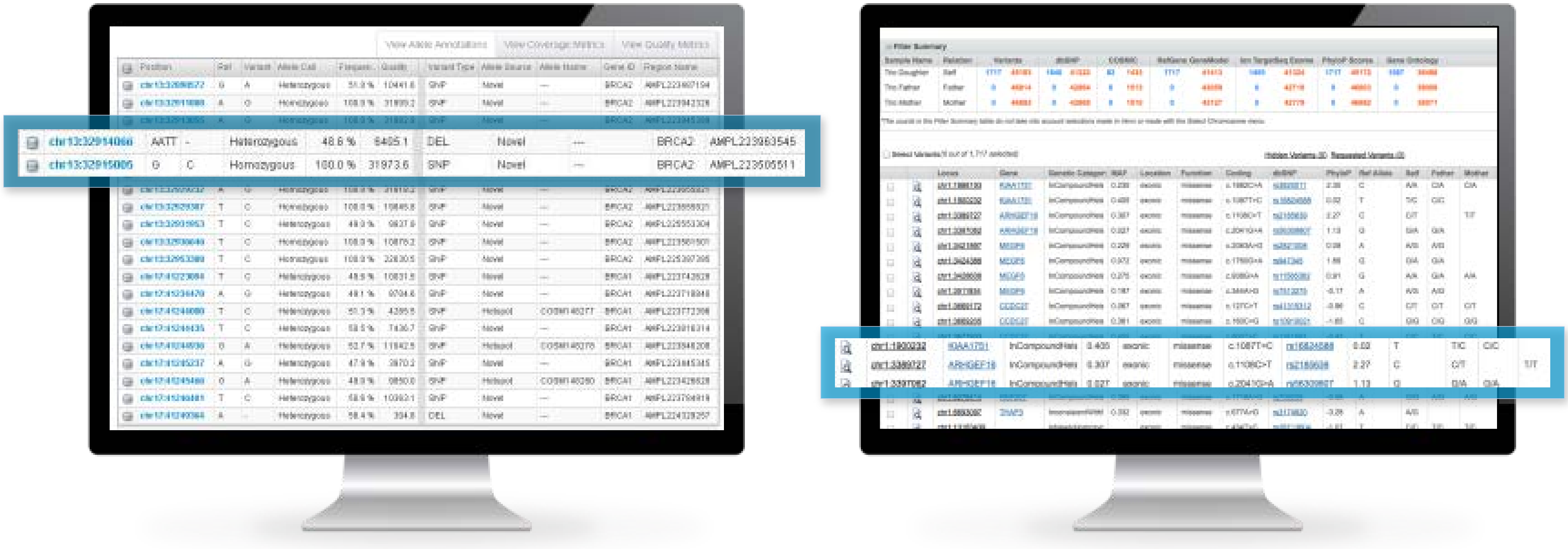


Fusion genes

Variant	*BRAF*^{V600E}	*EGFR* ΔE746-A750 *+ Kinase domain mutation*	*HER2 Overexpression*	*BCR-ABL*
Tumor	Melanoma	Lung adenocarcinomas	IDC-Breast cancer	Chronic myelogenous leukemia (CML)
Targeted drug	Vemurafenib (PLX4032)	Erlotinib/ Gefitinib	Trastuzumab	Imatinib

Translational Bioinformatics – New Found Knowledge

Adding Value to the Sequence

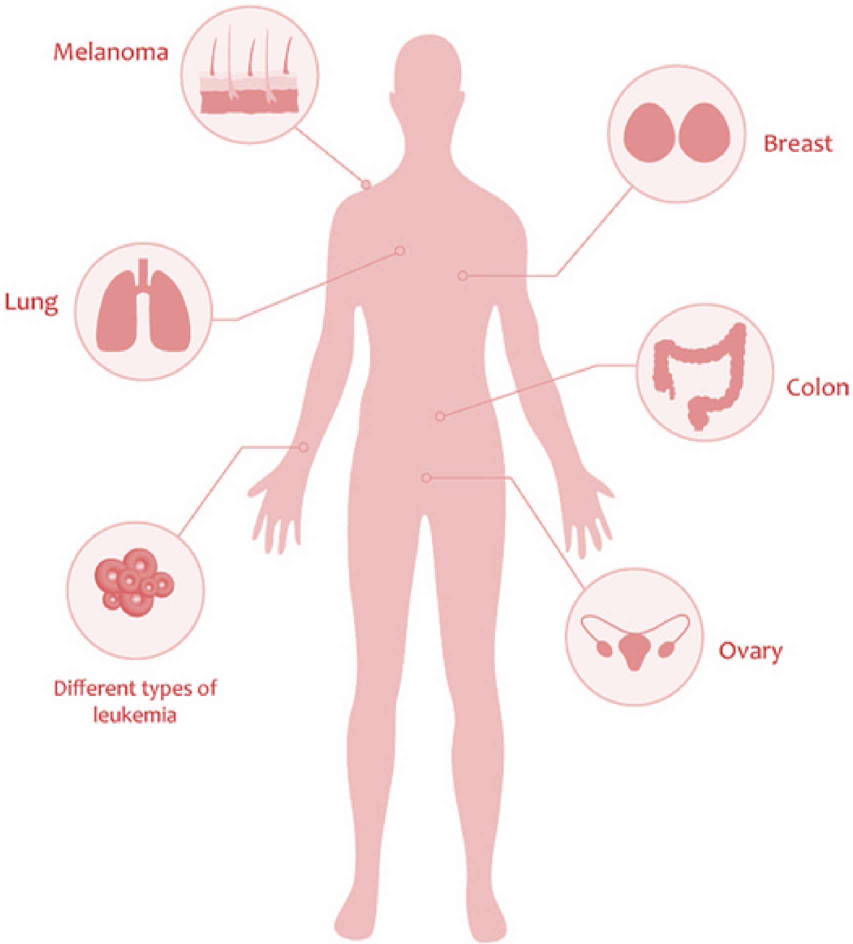


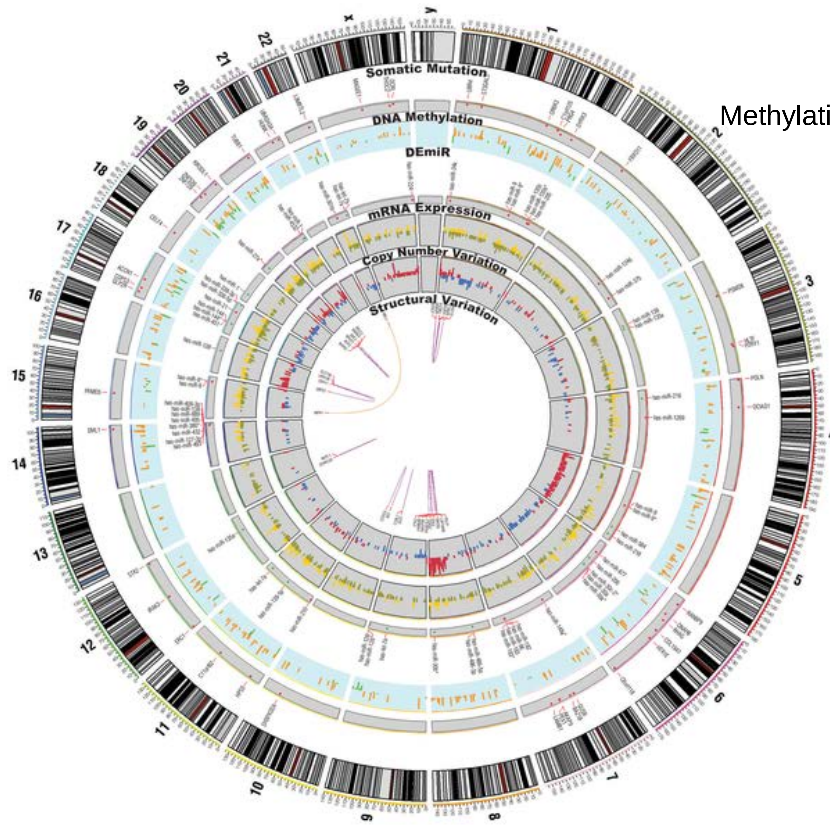
Torrent Suite™ Software → **Ion Reporter™ Software**

Variants → Biological meaning

Thermo Fisher SCIENTIFIC

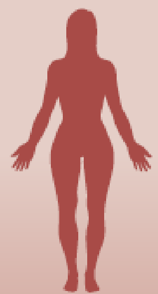
Multi-Dimensional Cancer Analysis





Gene expression
Mutations
CNV
miRNA
Methylation profile
EMRs
Clinical

Circos plot from SC Kim, *et al*. (2013) PLoS ONE 8(2): e55596



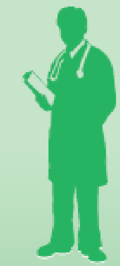
55 year old woman with carcinoma of unknown primary origin

Sequence analysis of 3 'actionable' cancer genes in fresh or archived biopsies

EGFR E746_A750del +
BRAF -
KRAS -
PIK3CA -
etc.

Comparison to cancer genome and clinical trials databases

- *EGFR* E746_A750del reported in lung (5.3%), breast (0.2%), ovary (0.4%), salivary glands (2.1%), etc.
- Functional consequence: Activating
- Clinical trials: Multiple



Clinical decision based on:

- Patient condition
- Prior therapies received by patient
- Mutation status
- Outcome information from prior clinical trials
- Availability of targeted therapy
- Patient preferences

TJ Hudson (2013) *J Int Med* **274**:440

The Oncomine Knowledgebase

The Oncomine Knowledgebase

Heterogeneous datasets

Data

- Identify and collect heterogeneous published cancer genomic data gathered from sources worldwide
- Data generated from strategic partnerships, including clinical sequencing collaborators and internal data at Thermo Fisher labs

Normalise Data

Expert Curation

- Dedicated team reviewing each sample, property, publication and mapping the data to the Oncomine® ontology
- Metadata curation and standardization
- Genomic data curation and re-annotation

Analysis Engine

Oncomine Ontology

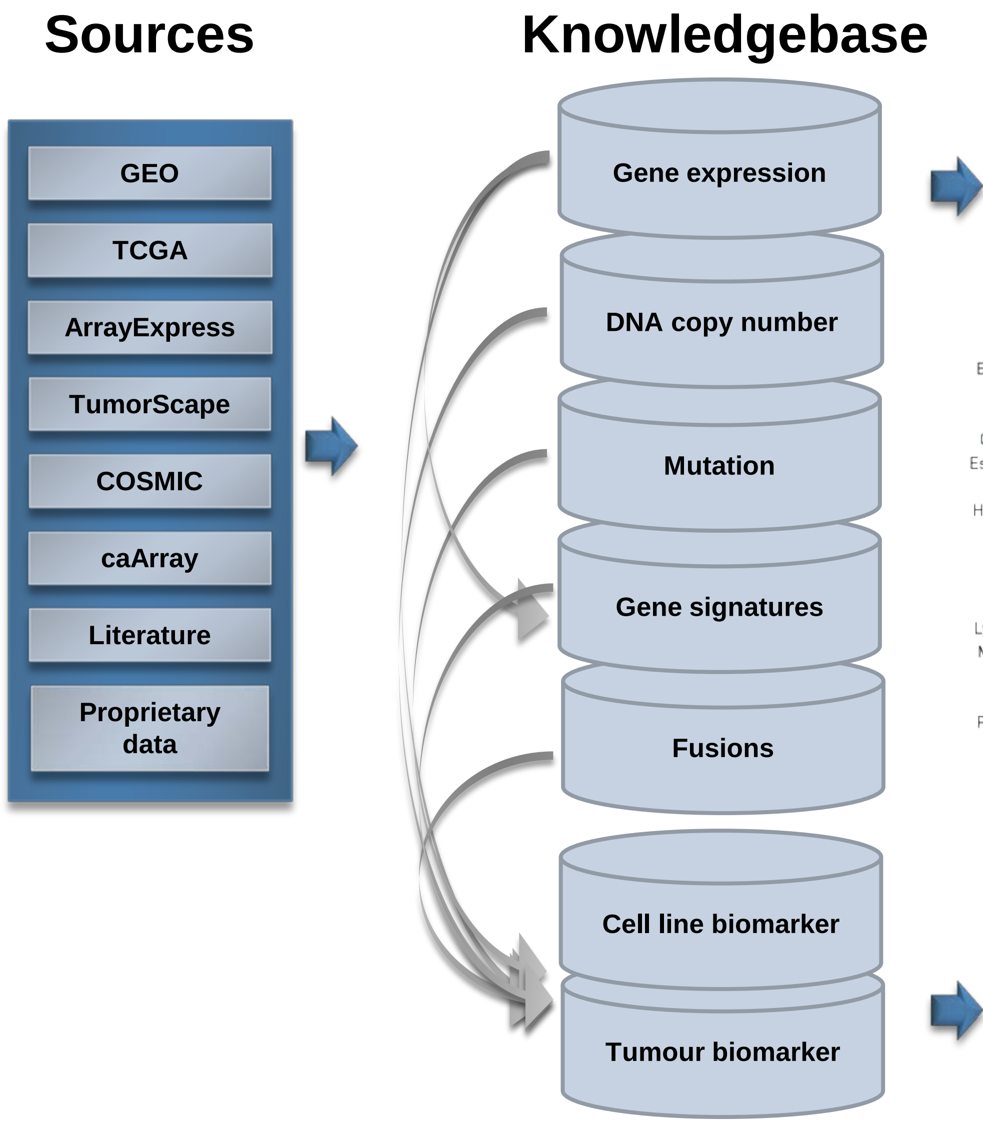
Multi-threaded hierarchy of terms and synonyms to describe the data

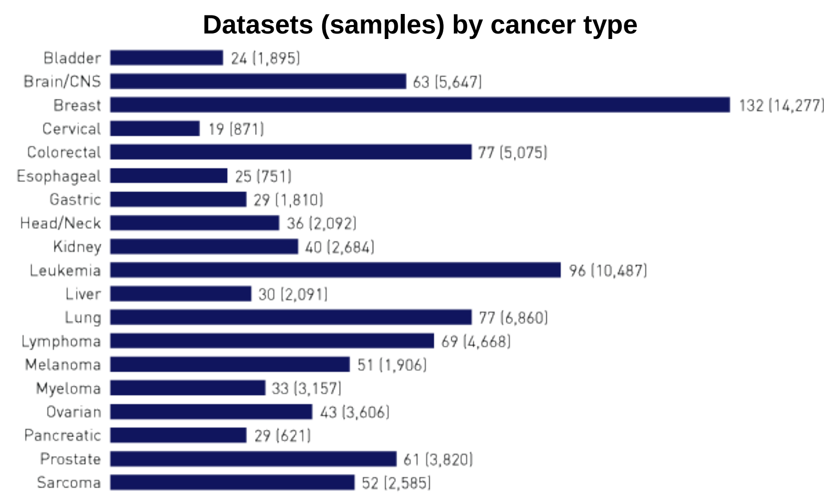
Findings Database

Standardised Analysis

Leveraging the Oncomine® Ontology, standardised analyses can be performed on every Oncomine® dataset

The Oncomine Knowledgebase





*This bar chart represents datasets. Some samples are represented in multiple cancer types. The Oncomine® knowledgebase also has 14,558 normal samples.

ICGC Data Portal

Cancer Projects | Advanced Search | Data Analysis | Data Repository

eg. BRAF, KRAS G12D, DO35100, MU7870, apoptosis, Cancer Gene Census, GO:0016049

About Us

The ICGC Data Portal provides tools for visualizing, querying and downloading the data released quarterly by the consortium's member projects.

To access ICGC controlled tier data, please read these instructions.

New features will be regularly added by the DCC development team. Feedback is welcome.

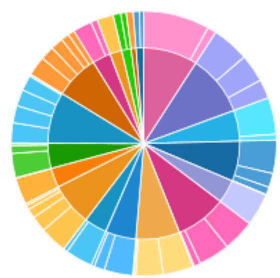
PCAWG
PanCancer Analysis of WHOLE GENOMES

The Pancancer Analysis of Whole Genomes (PCAWG) study is an international collaboration to identify common patterns of mutation in more than 2,600 cancer whole genomes from the International Cancer Genome Consortium.

Data Release 19

June 16th, 2015

Donor Distribution by Primary Site



Cancer projects	55
Cancer primary sites	21
Donors	12,979
Simple somatic mutations	16,459,160
Mutated genes	57,543

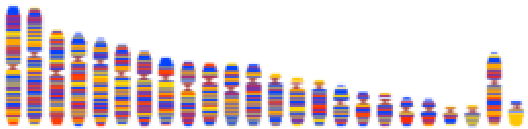
Information

Access Controlled Data
Methods
Submitter Tools

Tutorial

EXAMPLE QUERIES

1. BRAF missense mutations in colorectal cancer
2. Most frequently mutated genes by high impact mutations in stage III malignant lymphoma
3. Brain cancer donors with frameshift mutations and having methylation data available



International Cancer Genome Consortium

Challenges − Analysis & Interpretation

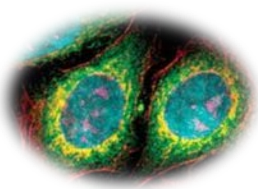
Role of Bioinformatics

The key to successful drug development and personalised medicine is understanding the data we are generating

Technology

NGS
Microarrays
High-content screening
Imaging

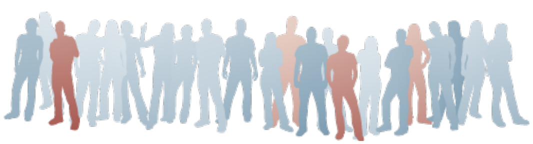
Cancer Biology



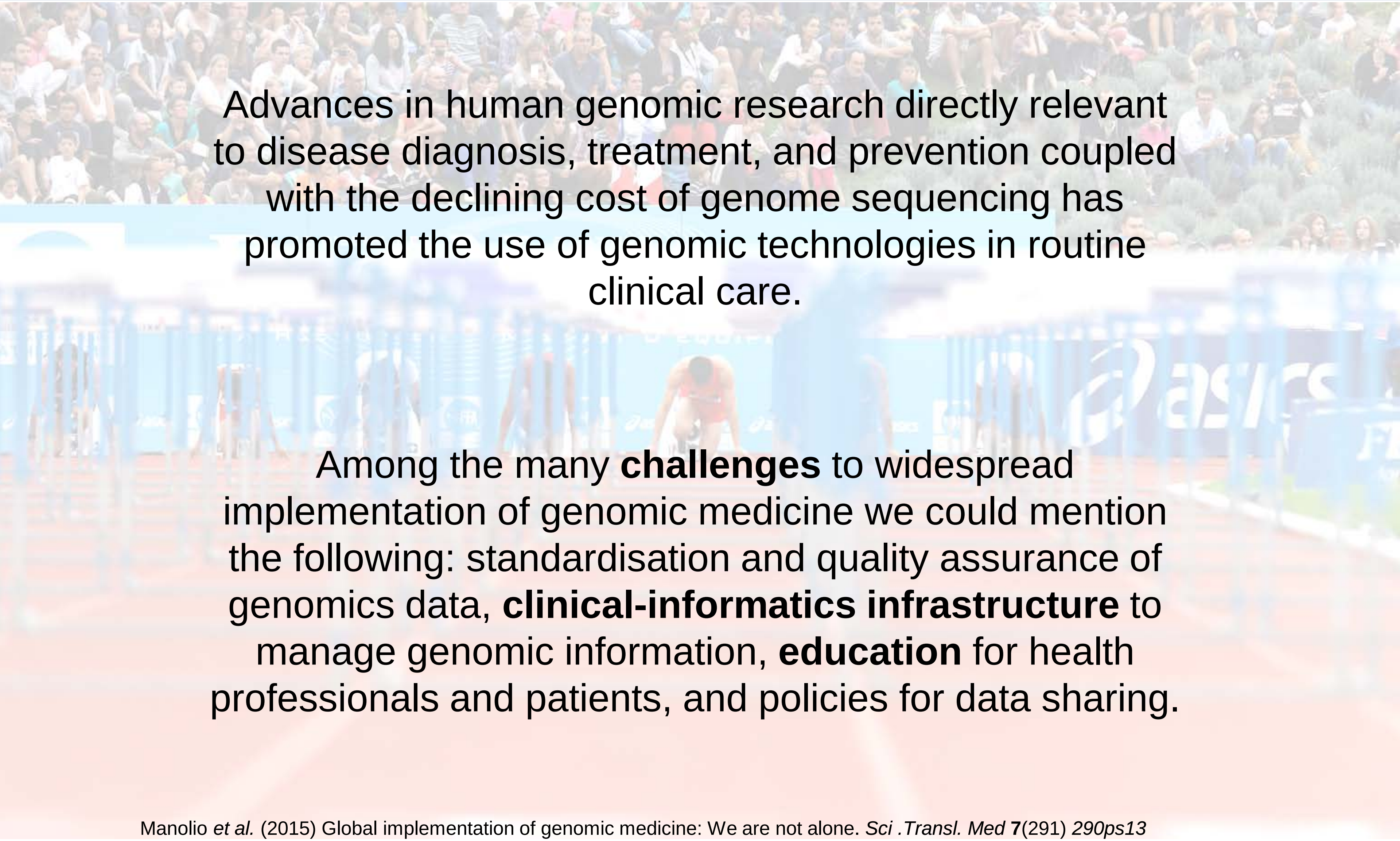
Bioinformatics

Data Management
Analytics
Visualisation
Interpretation

Improved Lives



Genomics-Driven Medicine − Clearing the Hurdles



Advances in human genomic research directly relevant to disease diagnosis, treatment, and prevention coupled with the declining cost of genome sequencing has promoted the use of genomic technologies in routine clinical care.

Among the many **challenges** to widespread implementation of genomic medicine we could mention the following: standardisation and quality assurance of genomics data, **clinical-informatics infrastructure** to manage genomic information, **education** for health professionals and patients, and policies for data sharing.

Manolio *et al.* (2015) Global implementation of genomic medicine: We are not alone. *Sci .Transl. Med* **7**(291) *290ps13*

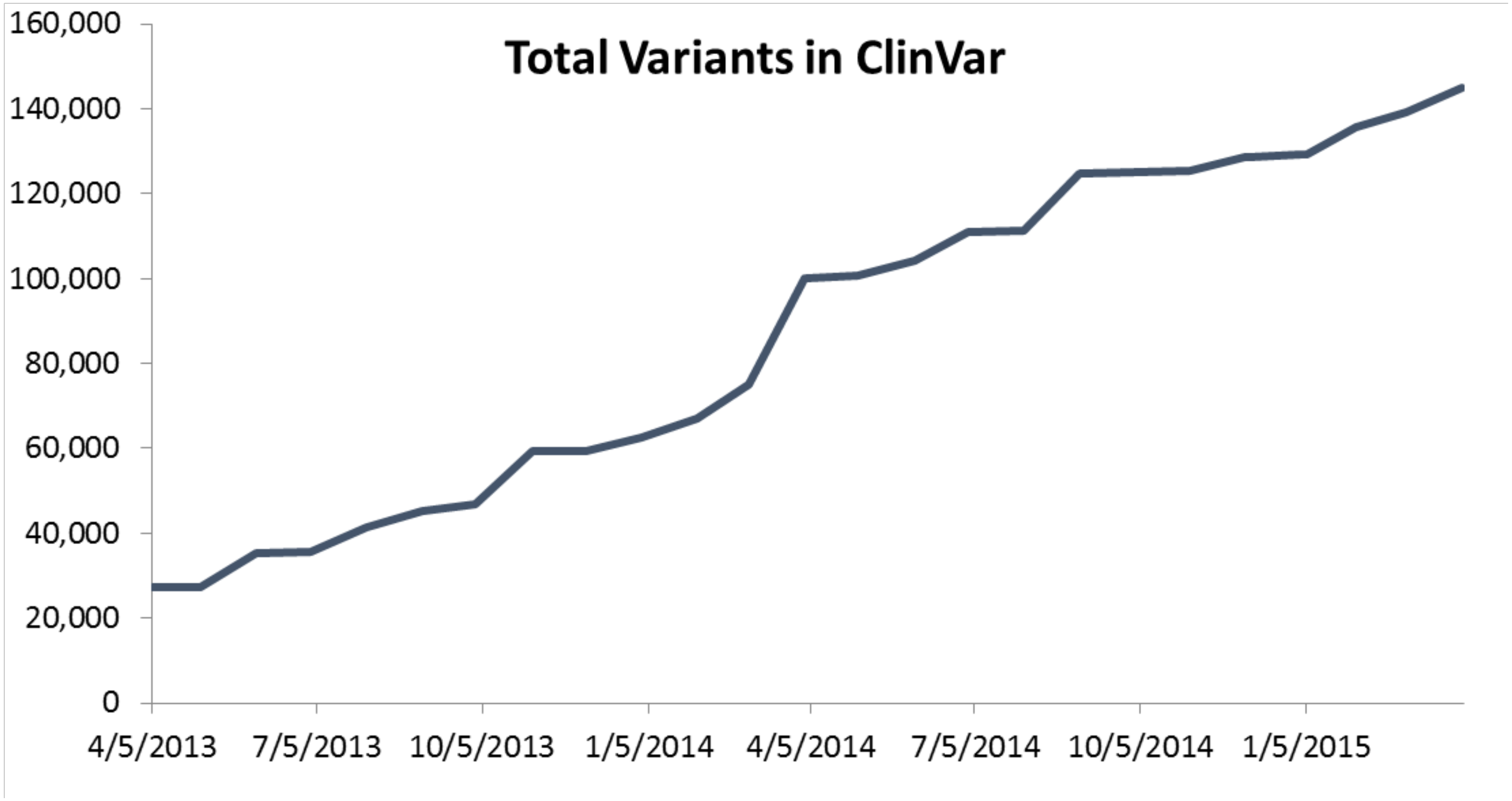
Cancer Genome Analysis

Databases

Database	Entities	Properties
Ensembl	Genes, proteins, transcripts, regulatory regions, variants	Genomic positions, relationships between them, identifiers in different formats, GO terms, PFAM domains
Entrez	Genes, articles	Articles for genes, abstracts of articles, links to full text
UniProt	Proteins	PDBs, known variants
KEGG, Reactome, Biocarta, Gene Ontology	Genes	Pathways, processes, function, cell location
TFacts	Genes	Transcription regulation
Barcode	Genes	Expression by tissue
PINA, HPRD, STRING	Proteins	Interactions
PharmaGKB	Drugs, proteins, variants	Drug targets, pharmacogenetics
STITCH, Matador	Drugs, proteins	Drug targets
Drug clinical trials	Investigational drugs	Diseases or conditions in they are being tested
GEO, ArrayExpress	Genes (microarray probes)	Expression values
ICGC, TCGA	Cancer Genomes	Point mutations, methylation, CNV, structural variants
dbSNP, 1000 genomes	Germline variations	Association with diseases or conditions
COSMIC	Somatic variations	Association with cancer types

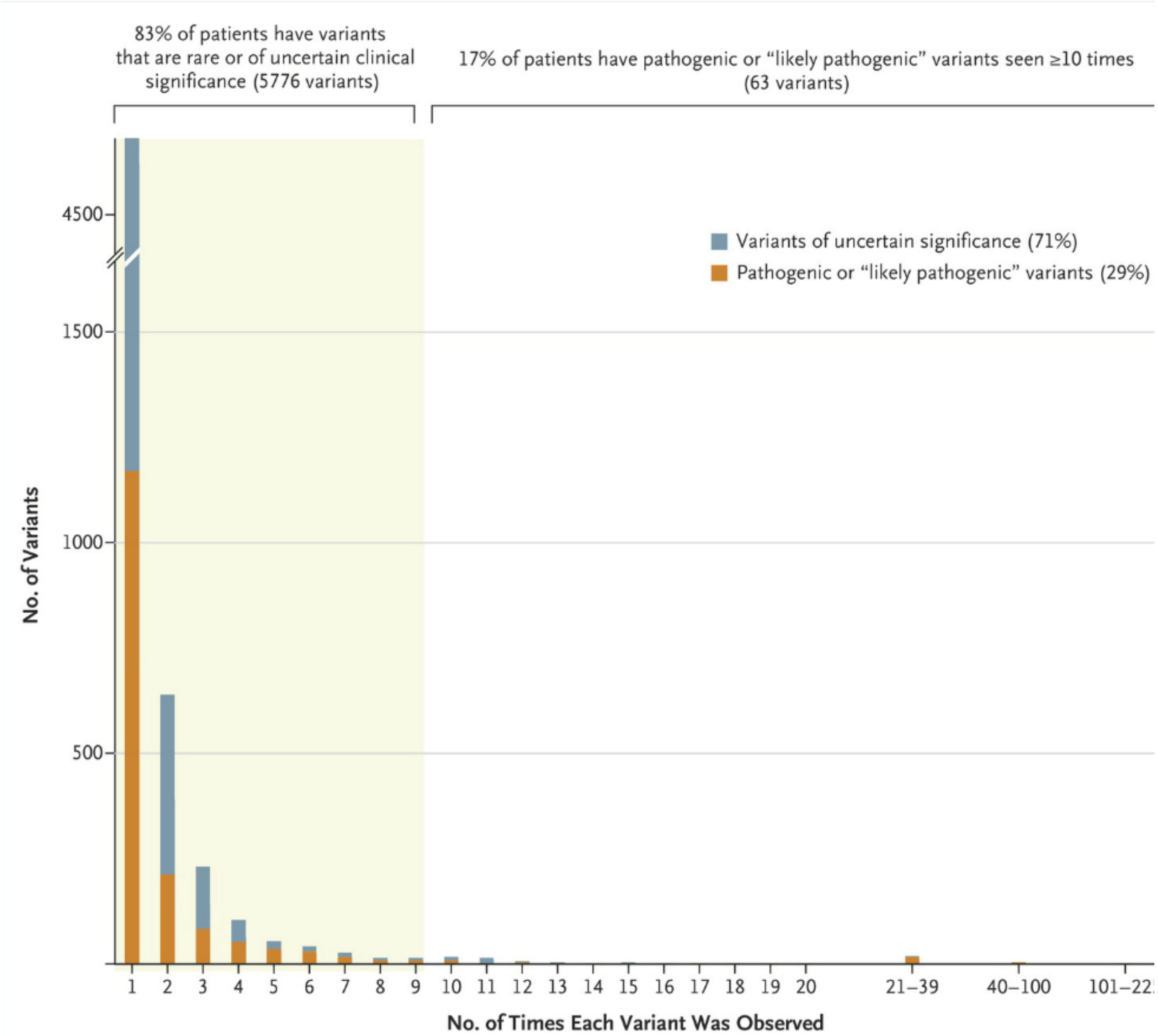
Vázquez *et al.* (2012) *PLoS Compt Biol*. **8**(12): e1002824.

ClinVar – Growing Content



ThermoFisher SCIENTIFIC

Variants in Mendelian Genes - ClinGen

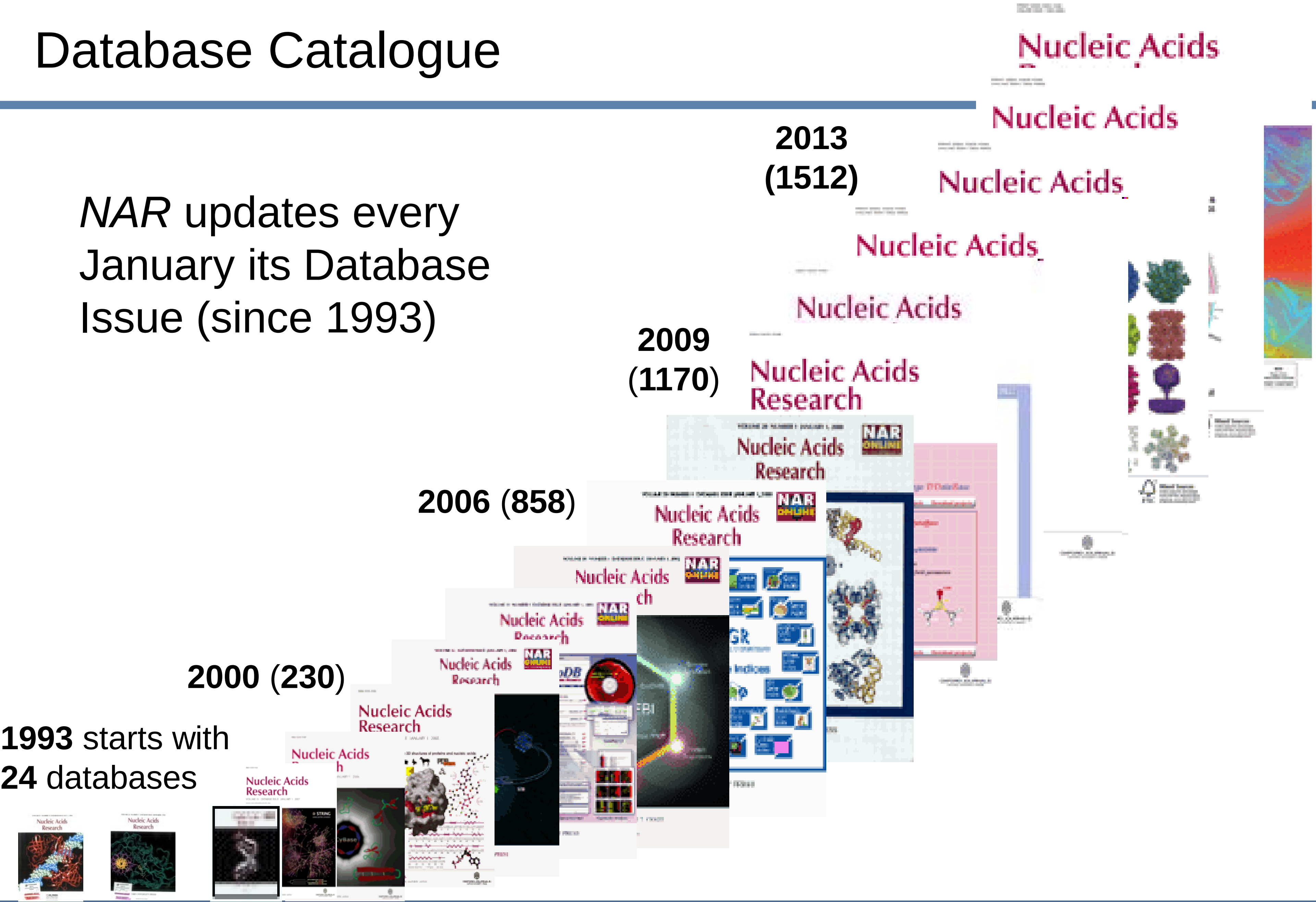


Rehm *et al.* (2015)

Thermo Fisher SCIENTIFIC

Database Catalogue

NAR updates every January its Database Issue (since 1993)



Nucleic Acids Research, 2015, Vol. 43, Database issue **D1–D5**
doi: 10.1093/nar/gku1241

The 2015 *Nucleic Acids Research* Database Issue and Molecular Biology Database Collection

Michael Y. Galperin[1,*], Daniel J. Rigden[2] and Xosé M. Fernández-Suárez[3]

[1]National Center for Biotechnology Information (NCBI), National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA, [2]Institute of Integrative Biology, University of Liverpool, Crown Street, Liverpool L69 7ZB, UK and [3]Thermo Fisher Scientific, Inchinnan Business Park, Paisley, Renfrew PA4 9RF, UK

Received November 10, 2014; Accepted November 11, 2014

Why this Kind of Training is Necessary?

Keeping up to date with professional developments is an integral part of good medical practice. The General Medical Council (GMC) has emphasised that continuing medical education should be tailored to the specific needs of the individual doctor, based on his or her personal practice.

- It is estimated that 50% of the knowledge acquired in medical school is obsolete after 7 years
- Continual Medical Education aims to update initial knowledge learnt in clinical school and add new knowledge to this
- Different pedagogical tools available (*e.g.* group discussions, surveys)