T5B: Developing a substitution calling algorithm to analyse breast cancer exomes by next generation sequencing

David Jones & Andy Menzies

Cancer Genome Project

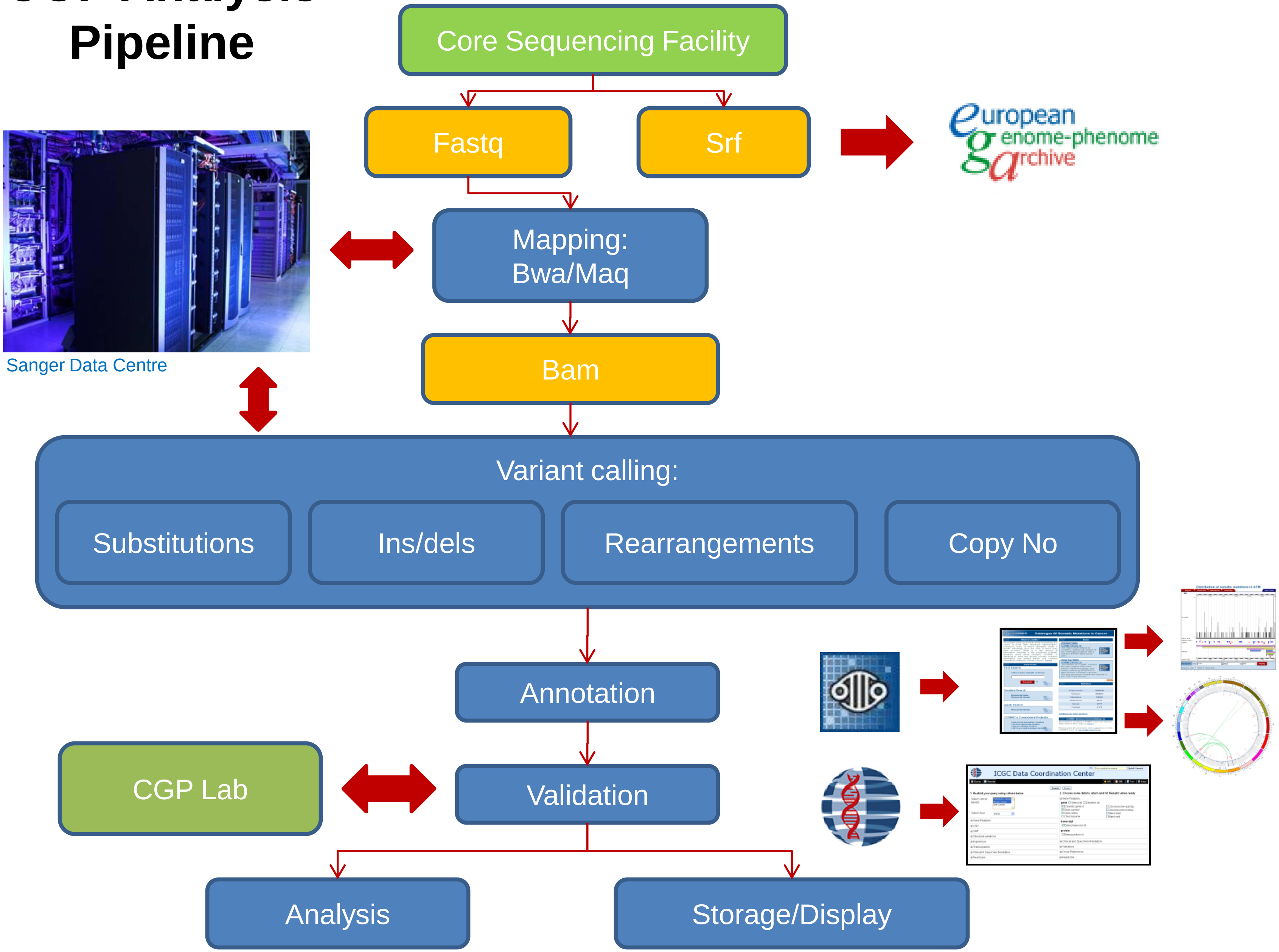
Wellcome Trust Sanger Institute

Introduction

- Overview of the analysis pipeline
  - Why do we need a pipeline?
  - Initial sequence analysis
  - Data release and presentation

- Variant analysis techniques
  - Rearrangement detection
  - Insertion/deletion detection
  - Substitution detection

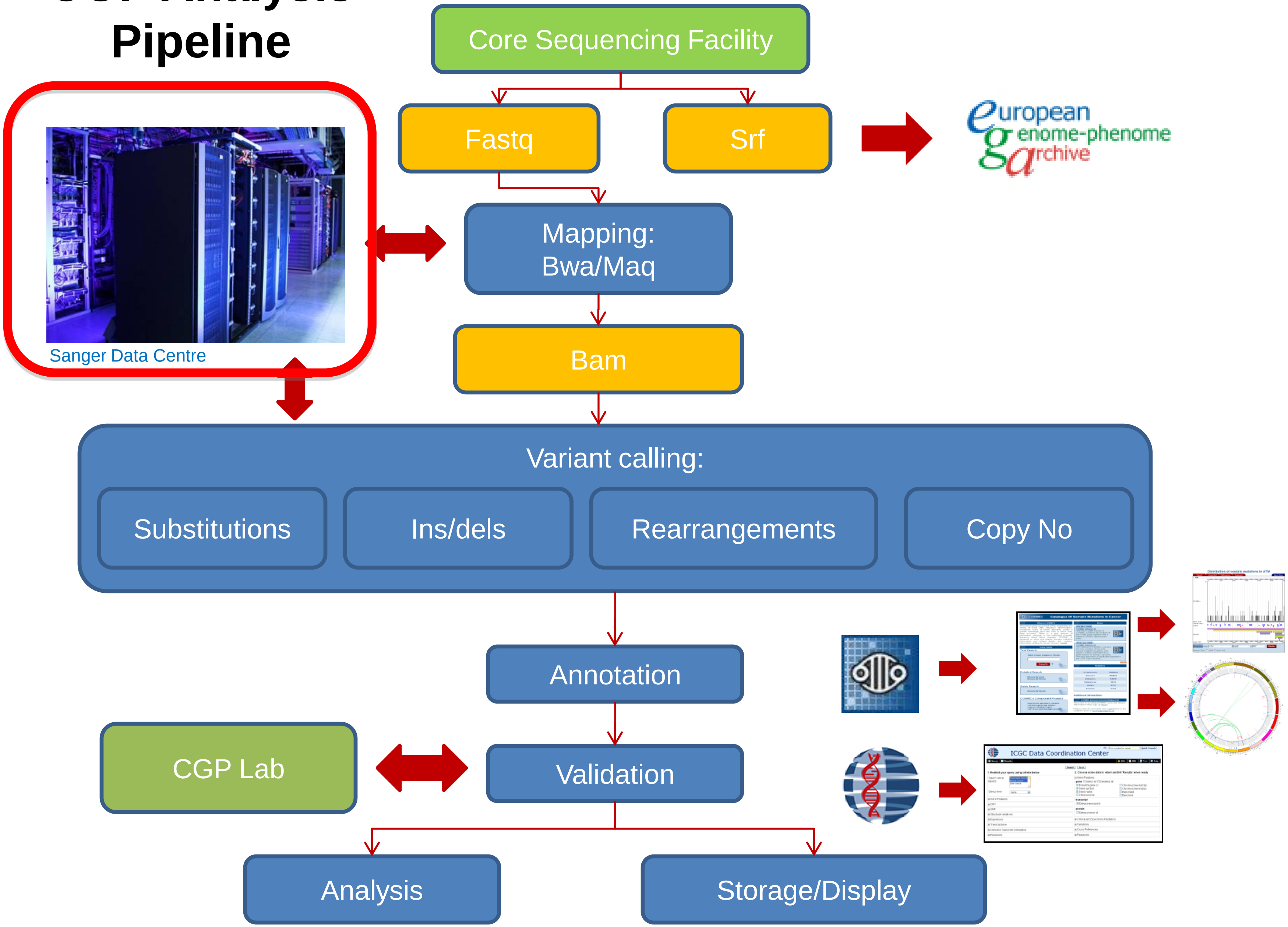
CGP Analysis Pipeline



Why: Data Volumes

- ICGC – 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.
  - Each ICGC project consists of 500 matched normal/tumour sample pairs
- EU BASIS
  - Breast ER+ve, HER2-ve – 500 norm/tum pairs
- Breast Cancer
  - Triple –ve, lobular & others – 500 norm/tum pairs
- Also assisting with a number of other ICGC projects
- And we have other non-ICGC projects too

CGP Analysis Pipeline



Core Sequencing Facility

Fastq

Srf

european genome-phenome archive

Sanger Data Centre

Mapping: Bwa/Maq

Bam

Variant calling:

Substitutions

Ins/dels

Rearrangements

Copy No

Annotation

CGP Lab

Validation

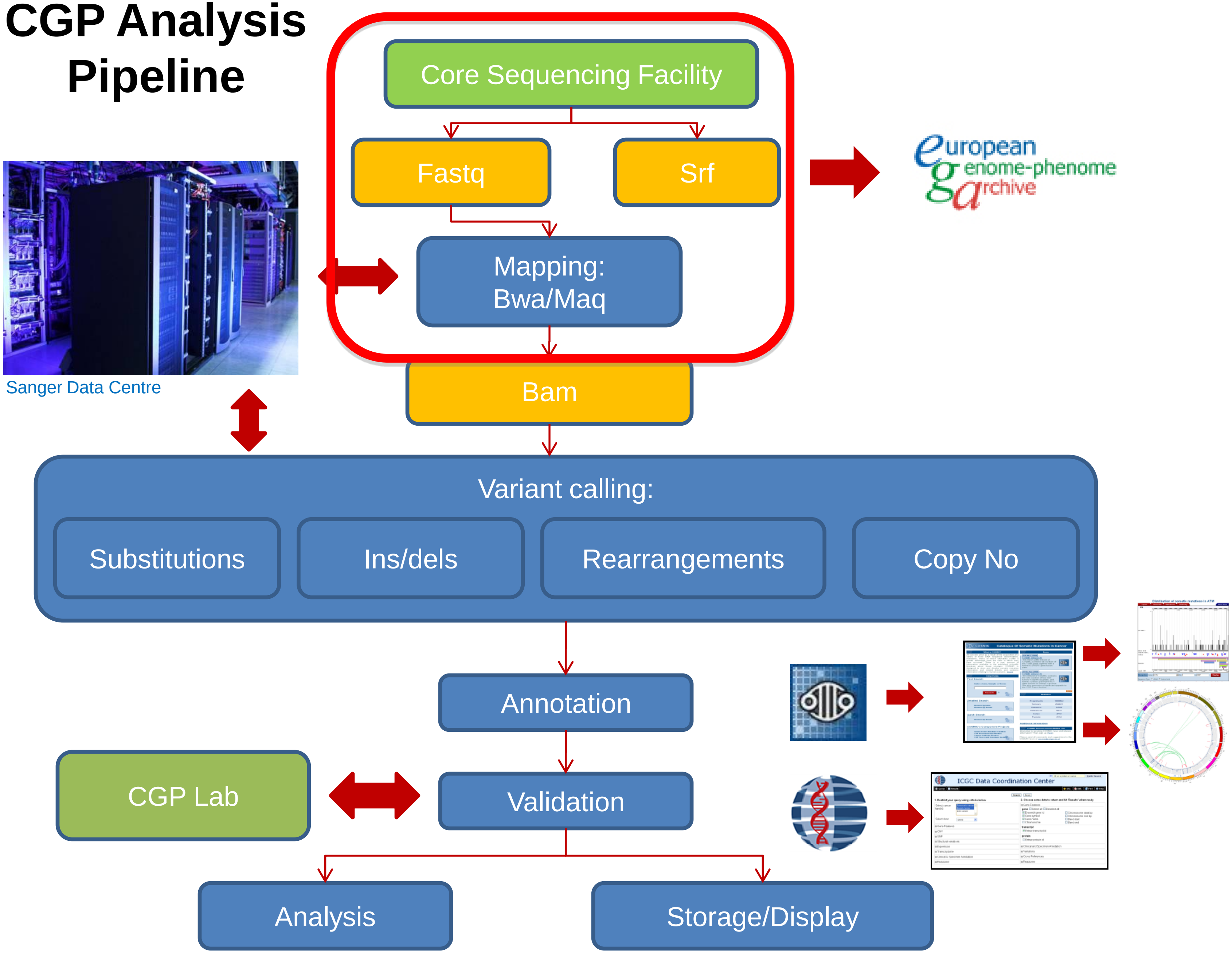
Analysis

Storage/Display

Hardware: The Farm

- CPUs
  - Mixture of dual and quad core multi processor machines
  - Total of 3920 cores
- Memory
  - Total of 8846Gb memory
  - Average of 2.2Gb per core
- Disk
  - 215TB Luster file system – working space
  - 350TB NFS file system – long term data storage
  - ~1TB to store a Normal/Tumor full genome pair & analysis

CGP Analysis Pipeline



Sanger Data Centre

Core Sequencing Facility

Fastq

Srf

Mapping: Bwa/Maq

Bam

european genome-phenome archive

Variant calling:

Substitutions

Ins/dels

Rearrangements

Copy No

Annotation

Validation

CGP Lab

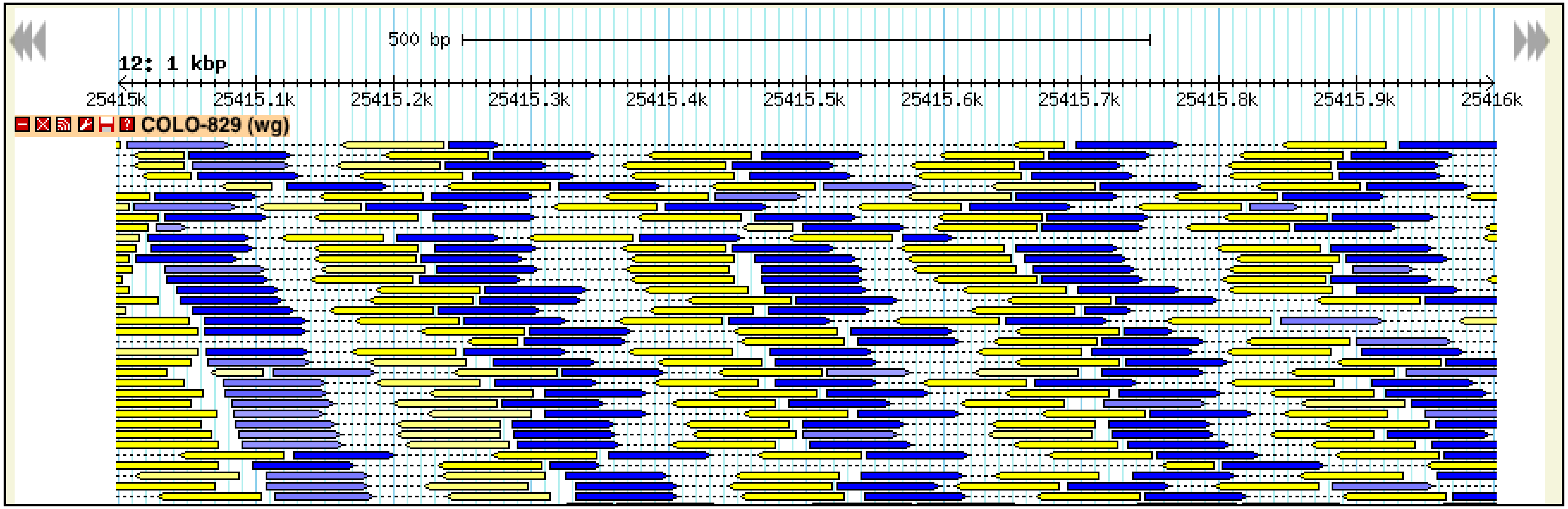
Analysis

Storage/Display

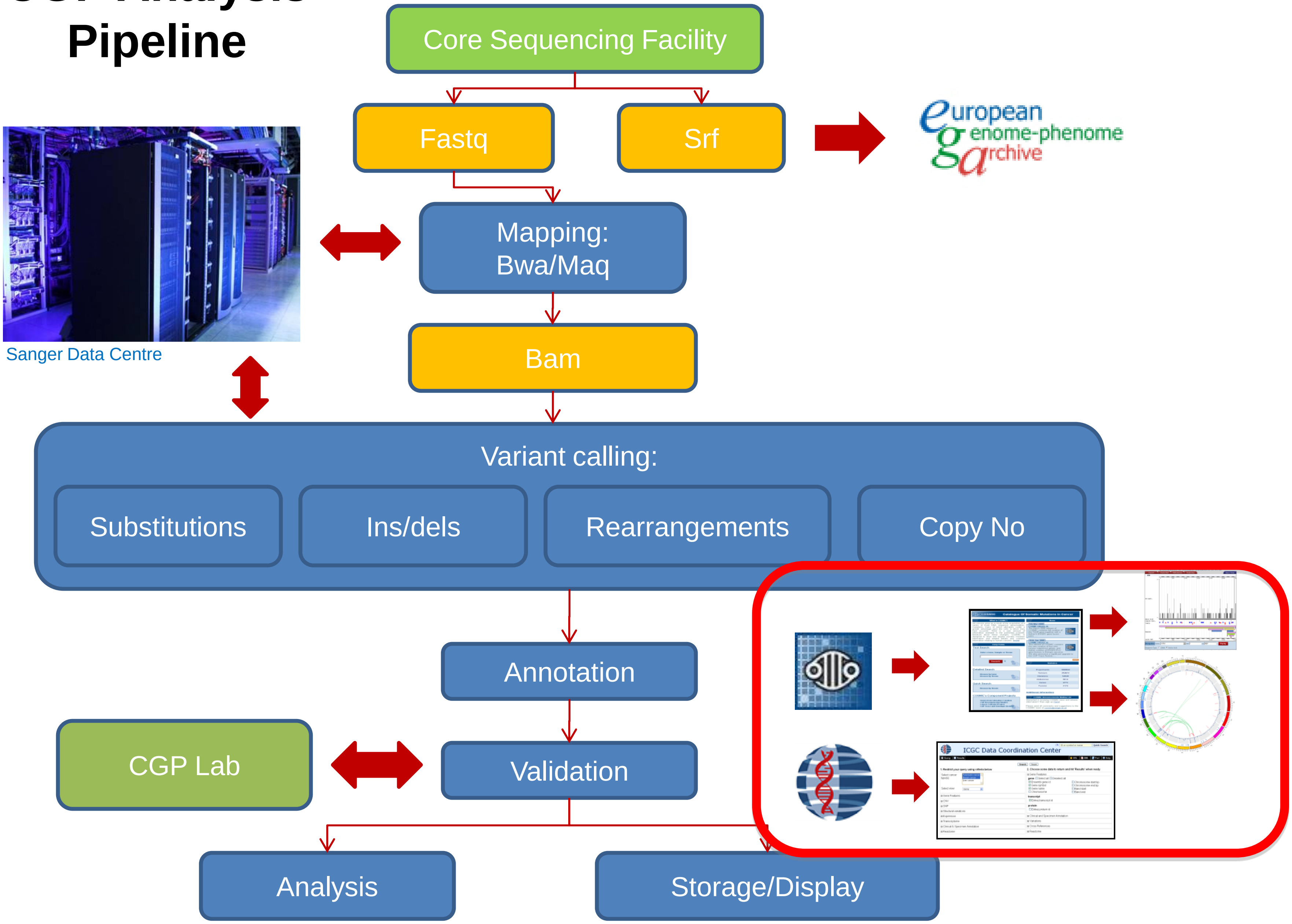
ICGC Data Coordination Center

Initial Data Processing

- Illumina GA2 – paired end sequencing
- Use BWA and MAQ
    - H Li et al Bioinformatics 2010
    - H Li et al Genome Research 2008
- Need to use fast aligners because of the vast number of individual reads generated by the sequencers



CGP Analysis Pipeline



Core Sequencing Facility

Fastq

Srf

european genome-phenome archive

Mapping: Bwa/Maq

Sanger Data Centre

Bam

Variant calling:

Substitutions

Ins/dels

Rearrangements

Copy No

Annotation

CGP Lab

Validation

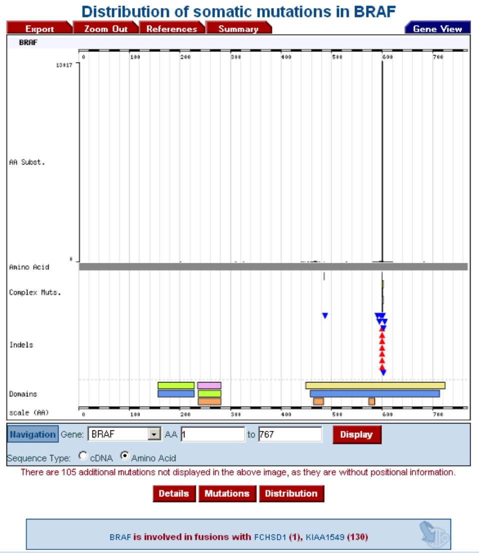
Analysis

Storage/Display

ICGC Data Coordination Center

COSMIC Interfaces and Web-based Tools

By Gene



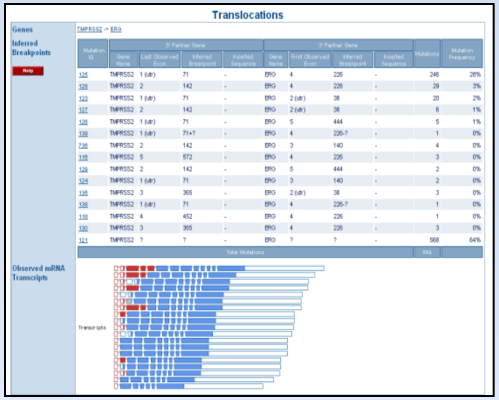
Distribution of somatic mutations in BRAF

Gene: BRAF AA 1 to 767

Sequence Type: ○ cDNA ● Amino Acid

There are 105 additional mutations not displayed in the above image, as they are without positional information.

BRAF is involved in fusions with FCHSD1 (1), KIAA1549 (130)

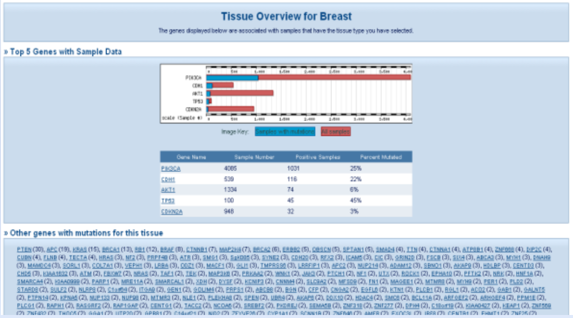


Translocations

TMPRSS2 -> ERG

Mutation ID	5' Partner Gene				3' Partner Gene				Mutations	Mutation Frequency
	Gene Name	Last Observed Exon	Inferred Breakpoint	Inserted Sequence	Gene Name	First Observed Exon	Inferred Breakpoint	Inserted Sequence		
125	TMPRSS2	1 (utr)	71	-	ERG	4	226	-	246	28%
128	TMPRSS2	2	142	-	ERG	4	226	-	29	3%
123	TMPRSS2	1 (utr)	71	-	ERG	2 (utr)	38	-	20	2%
127	TMPRSS2	2	142	-	ERG	2 (utr)	38	-	6	1%
126	TMPRSS2	1 (utr)	71	-	ERG	5	444	-	5	1%
139	TMPRSS2	1 (utr)	71+?	-	ERG	4	226-?	-	1	0%
736	TMPRSS2	2	142	-	ERG	3	140	-	4	0%
116	TMPRSS2	5	572	-	ERG	4	226	-	3	0%
129	TMPRSS2	2	142	-	ERG	5	444	-	2	0%
124	TMPRSS2	1 (utr)	71	-	ERG	3	140	-	2	0%
135	TMPRSS2	3	365	-	ERG	2 (utr)	38	-	3	0%
138	TMPRSS2	1 (utr)	71	-	ERG	4	226-?	-	1	0%
118	TMPRSS2	4	452	-	ERG	4	226	-	1	0%
130	TMPRSS2	3	365	-	ERG	4	226	-	3	0%
121	TMPRSS2	?	?	-	ERG	?	?	-	568	64%
Total Mutations									894	

By Tissue



Tissue Overview for Breast

The genes displayed below are associated with samples that have the tissue type you have selected.

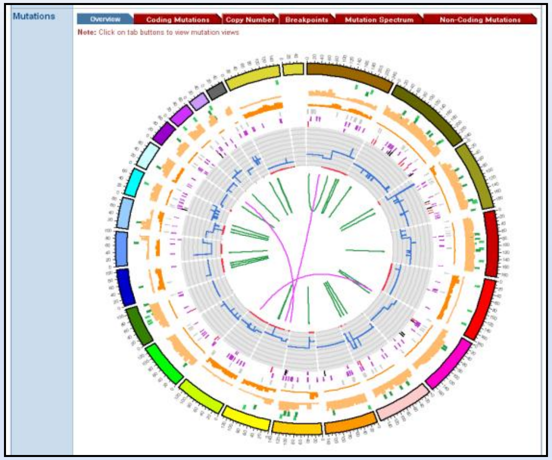
» Top 5 Genes with Sample Data

Gene Name	Sample Number	Positive Samples	Percent Mutated
PIK3CA	4085	1031	25%
CDH1	539	116	22%
AKT1	1334	74	6%
TP53	100	45	45%
CDKN2A	948	32	3%

» Other genes with mutations for this tissue

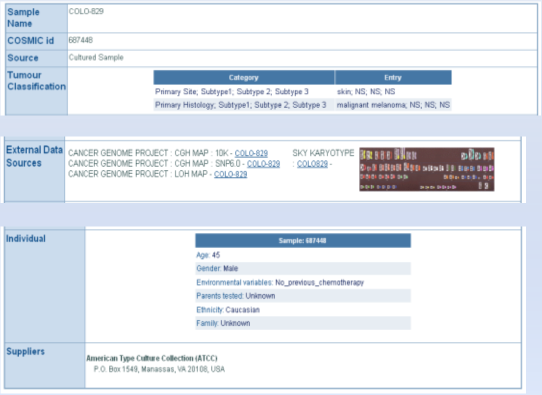
PTEN (30), APC (19), KRAS (15), BRCA1 (13), RB1 (12), BRAF (8), CTNNB1 (7), MAP2K4 (7), BRCA2 (6), ERBB2 (5), OBSCN (5), SPTAN1 (5), SMAD4 (4), TTN (4), CTNNA1 (4), ATP8B1 (4), ZNF668 (4), DIP2C (4), CUBN (4), FLNB (4), TECTA (4), HRAS (3), NF2 (3), PRPF4B (3), ATR (3), SMG1 (3), SgK085 (3), SYNE2 (3), CDH20 (3), RFX2 (3), ICAM5 (3), CIC (3), GRIN2D (3), FSCB (3), SIX4 (3), ABCA3 (3), MYH1 (3), DNAH9 (3), MAMDC4 (3), SORL1 (3), COL7A1 (3), VEPH1 (3), LRBA (3), ODZ1 (3), MACF1 (3), GLI1 (3), TMPRSS6 (3), LRRFIP1 (3), APC2 (3), NUP214 (3), ADAM12 (3), SBNO1 (3), AKAP9 (3), HDLBP (3), CENTD3 (3), CHD6 (3), KIAA1632 (3), ATM (2), FBXW7 (2), NRAS (2), TAF1 (2), TEK (2), MAP3K8 (2), PRKAA2 (2), WNK1 (2), JAK3 (2), PTCH1 (2), NF1 (2), UTX (2), ROCK1 (2), EPHA10 (2), PFTK2 (2), NRK (2), HNF1A (2), SMARCA4 (2), KIAA0999 (2), PARP1 (2), MRE11A (2), SMARCAL1 (2), XDH (2), DYSF (2), KCNIP3 (2), CNNM4 (2), SLC9A2 (2), MFSD9 (2), FN1 (2), MAGEE1 (2), MTMR8 (2), MYH9 (2), PER1 (2), PLD2 (2), STARD8 (2), SULF2 (2), NLRP8 (2), C1orf84 (2), ITGA9 (2), GEN1 (2), GOLIM4 (2), PRPS1 (2), ABCB8 (2), BGN (2), CFP (2), CNGA2 (2), EGFL6 (2), KTN1 (2), PLCB1 (2), RGL1 (2), ACO2 (2), GAB1 (2), GALNT5 (2), PTPN14 (2), KPNA5 (2), NUP133 (2), NUP98 (2), MTMR3 (2), NLE1 (2), PLEKHA8 (2), SPEN (2), UBR4 (2), AKAP6 (2), DDX10 (2), HDAC4 (2), SMC6 (2), BCL11A (2), ARFGEF2 (2), ARHGEF4 (2), PPM1E (2), PLCG1 (2), RAPH1 (2), RASGRF2 (2), RAP1GAP (2), CENTG1 (2), TACC2 (2), NCOA6 (2), SREBF2 (2), PKDREJ (2), SEMA5B (2), ZNF318 (2), ZNF277 (2), DPH4 (2), C10orf18 (2), KIAA0427 (2), KEAP1 (2), ZNF569 (2), ZNF432 (2), THOC5 (2), GGA1 (2), UTP20 (2), APBB1 (2), C14orf21 (2), NID2 (2), ZFYVE26 (2), CYP1A1 (2), SCNN1B (2), ZNF046 (2), AMFR (2), EXOC3L (2), IRF8 (2), CENTB1 (2), EHMT1 (2), ZNF25 (2),

By Genome



Note: Click on tab buttons to view mutation views

By Sample/Phenotype



Sample Name	COLO-829
COSMIC id	687448
Source	Cultured Sample
Tumour Classification	

Category	Entry
Primary Site; Subtype1; Subtype 2; Subtype 3	skin; NS; NS; NS
Primary Histology; Subtype1; Subtype 2; Subtype 3	malignant melanoma; NS; NS; NS

External Data Sources

CANCER GENOME PROJECT : CGH MAP : 10K - COLO-829
CANCER GENOME PROJECT : CGH MAP : SNP6.0 - COLO-829
CANCER GENOME PROJECT : LOH MAP - COLO-829

SKY KARYOTYPE : COLO829 -

Individual

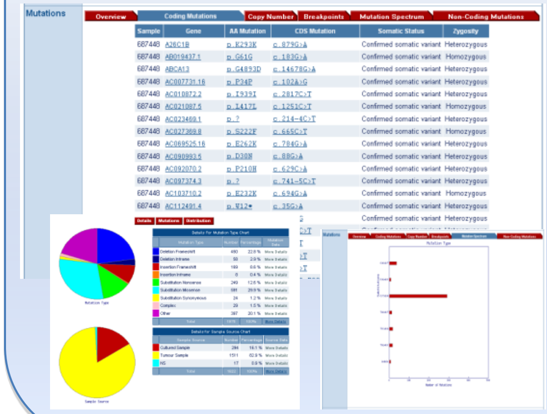
Sample: 687448

Age: 45
Gender: Male
Environmental variables: No_previous_chemotherapy
Parents tested: Unknown
Ethnicity: Caucasian
Family: Unknown

Suppliers

American Type Culture Collection (ATCC)
P.O. Box 1549, Manassas, VA 20108, USA

By Mutation



Sample	Gene	AA Mutation	CDS Mutation	Somatic Status	Zygosity
687448	A26C1B	p.K293K	c.879G>A	Confirmed somatic variant	Heterozygous
687448	AB019437.1	p.G61G	c.183G>A	Confirmed somatic variant	Homozygous
687448	ABCA13	p.G4893D	c.14678G>A	Confirmed somatic variant	Heterozygous
687448	AC007731.16	p.P34P	c.102A>G	Confirmed somatic variant	Heterozygous
687448	AC010872.2	p.I939I	c.2817C>T	Confirmed somatic variant	Heterozygous
687448	AC021087.5	p.L417L	c.1251C>T	Confirmed somatic variant	Homozygous
687448	AC023469.1	p.?	c.214-4C>T	Confirmed somatic variant	Heterozygous
687448	AC027369.8	p.S222F	c.665C>T	Confirmed somatic variant	Homozygous
687448	AC069525.16	p.E262K	c.784G>A	Confirmed somatic variant	Heterozygous
687448	AC090993.5	p.D30N	c.88G>A	Confirmed somatic variant	Heterozygous
687448	AC092070.2	p.P210H	c.629C>A	Confirmed somatic variant	Heterozygous
687448	AC097374.3	p.?	c.741-5C>T	Confirmed somatic variant	Heterozygous
687448	AC103710.2	p.E232K	c.694G>A	Confirmed somatic variant	Homozygous
687448	AC112491.4	p.W12*	c.35G>A	Confirmed somatic variant	Heterozygous
			G	Confirmed somatic variant	Heterozygous

Details For Mutation Type Chart

	Mutation Type	Number	Percentage	Mutation Data
🟦	Deletion Frameshift	450	22.8 %	More Details
🟦	Deletion Inframe	58	2.9 %	More Details
🟥	Insertion Frameshift	169	8.6 %	More Details
🟧	Insertion Inframe	8	0.4 %	More Details
🟩	Substitution Nonsense	249	12.6 %	More Details
🟦	Substitution Missense	591	29.9 %	More Details
🟨	Substitution Synonymous	24	1.2 %	More Details
🟪	Complex	29	1.5 %	More Details
🟪	Other	397	20.1 %	More Details
	Total	1975	100%	More Details

Details for Sample Source Chart

	Sample Source	Number	Percentage	Source Data
🟥	Cultured Sample	294	16.1 %	More Details
🟨	Tumour Sample	1511	82.9 %	More Details
🟦	NS	17	0.9 %	More Details
	Total	1822	100%	More Details

COSMIC Data Interoperability

Links to/from COSMIC & Resources that use COSMIC Data

e!Ensembl

Entrez Gene

intOGen

HG NC

OMIM Online Mendelian Inheritance in Man
Johns Hopkins University

Consensus CDS protein set
CCDS Database CCDS
EBI • NCBI • UCSC • WTSI

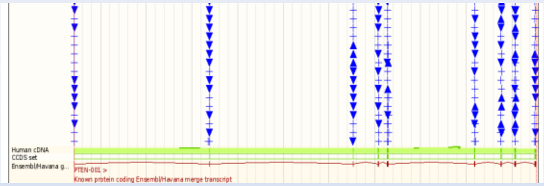
Atlas of Genetics and Cytogenetics in Oncology and Haematology

UniProt

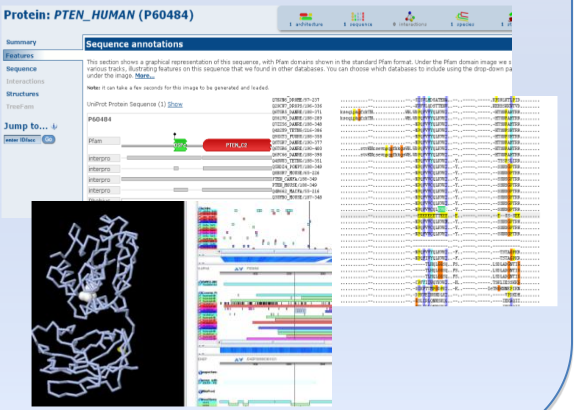
CGWB Home

Distribution Annotation System (DAS)

COSMIC/Ensembl DAS



COSMIC/Pfam/UniProt DAS



Protein: *PTEN_HUMAN* (P60484)

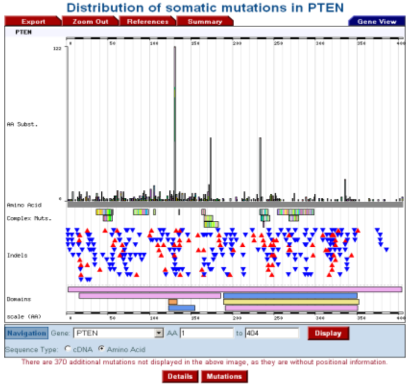
Sequence annotations

This section shows a graphical representation of this sequence, with Pfam domains shown in the standard Pfam format. Under the Pfam domain image we s various tracks, illustrating features on this sequence that we found in other databases. You can choose which databases to include using the drop-down pa under the image. More...

Note: it can take a few seconds for this image to be generated and loaded.

UniProt Protein Sequence (1) Show

COSMIC

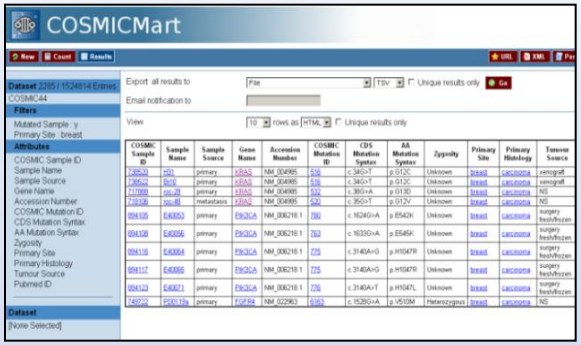


Distribution of somatic mutations in PTEN

There are 370 additional mutations not displayed in the above image, as they are without positional information.

Biomart

bio mart



Dataset 2265 / 1524814 Entries
COSMIC44

Filters
Mutated Sample : y
Primary Site : breast

Attributes
COSMIC Sample ID
Sample Name
Sample Source
Gene Name
Accession Number
COSMIC Mutation ID
CDS Mutation Syntax
AA Mutation Syntax
Zygosity
Primary Site
Primary Histology
Tumour Source
Pubmed ID

Dataset
[None Selected]

Export all results to File TSV □ Unique results only Go

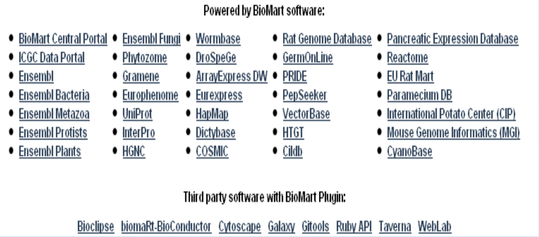
Email notification to

View 10 rows as HTML □ Unique results only

COSMIC Sample ID	Sample Name	Sample Source	Gene Name	Accession Number	COSMIC Mutation ID	CDS Mutation Syntax	AA Mutation Syntax	Zygosity	Primary Site	Primary Histology	Tumour Source
738520	H31	primary	KRAS	NM_004985	516	c.34G>T	p.G12C	Unknown	breast	carcinoma	xenograft
738622	Br10	primary	KRAS	NM_004985	516	c.34G>T	p.G12C	Unknown	breast	carcinoma	xenograft
717888	roc-28	primary	KRAS	NM_004985	532	c.38G>A	p.G13D	Unknown	breast	carcinoma	NS
719106	roc-48	metastasis	KRAS	NM_004985	520	c.35G>T	p.G12V	Unknown	breast	carcinoma	NS
894105	E40053	primary	PIK3CA	NM_006218.1	760	c.1624G>A	p.E542K	Unknown	breast	carcinoma	surgery fresh/frozen
894108	E40056	primary	PIK3CA	NM_006218.1	763	c.1633G>A	p.E545K	Unknown	breast	carcinoma	surgery fresh/frozen
894116	E40064	primary	PIK3CA	NM_006218.1	775	c.3140A>G	p.H1047R	Unknown	breast	carcinoma	surgery fresh/frozen
894117	E40065	primary	PIK3CA	NM_006218.1	775	c.3140A>G	p.H1047R	Unknown	breast	carcinoma	surgery fresh/frozen
894123	E40071	primary	PIK3CA	NM_006218.1	776	c.3140A>T	p.H1047L	Unknown	breast	carcinoma	surgery fresh/frozen
749722	PD0119a	primary	FGFR4	NM_022963	6163	c.1528G>A	p.V510M	Heterozygous	breast	carcinoma	NS

Biomart related resources

ICGC Franchise Database



Powered by BioMart software:

- BioMart Central Portal
- ICGC Data Portal
- Ensembl
- Ensembl Bacteria
- Ensembl Metazoa
- Ensembl Protists
- Ensembl Plants
- Ensembl Fungi
- Phytozome
- Gramene
- Europhenome
- UniProt
- InterPro
- HGNC
- Wormbase
- DroSpeGe
- ArrayExpress DW
- Eurexpress
- HapMap
- Dictybase
- COSMIC
- Rat Genome Database
- GermOnLine
- PRIDE
- PepSeeker
- VectorBase
- HTGT
- Cildb
- Pancreatic Expression Database
- Reactome
- EU Rat Mart
- Paramecium DB
- International Potato Center (CIP)
- Mouse Genome Informatics (MGI)
- CyanoBase

Third party software with BioMart Plugin:

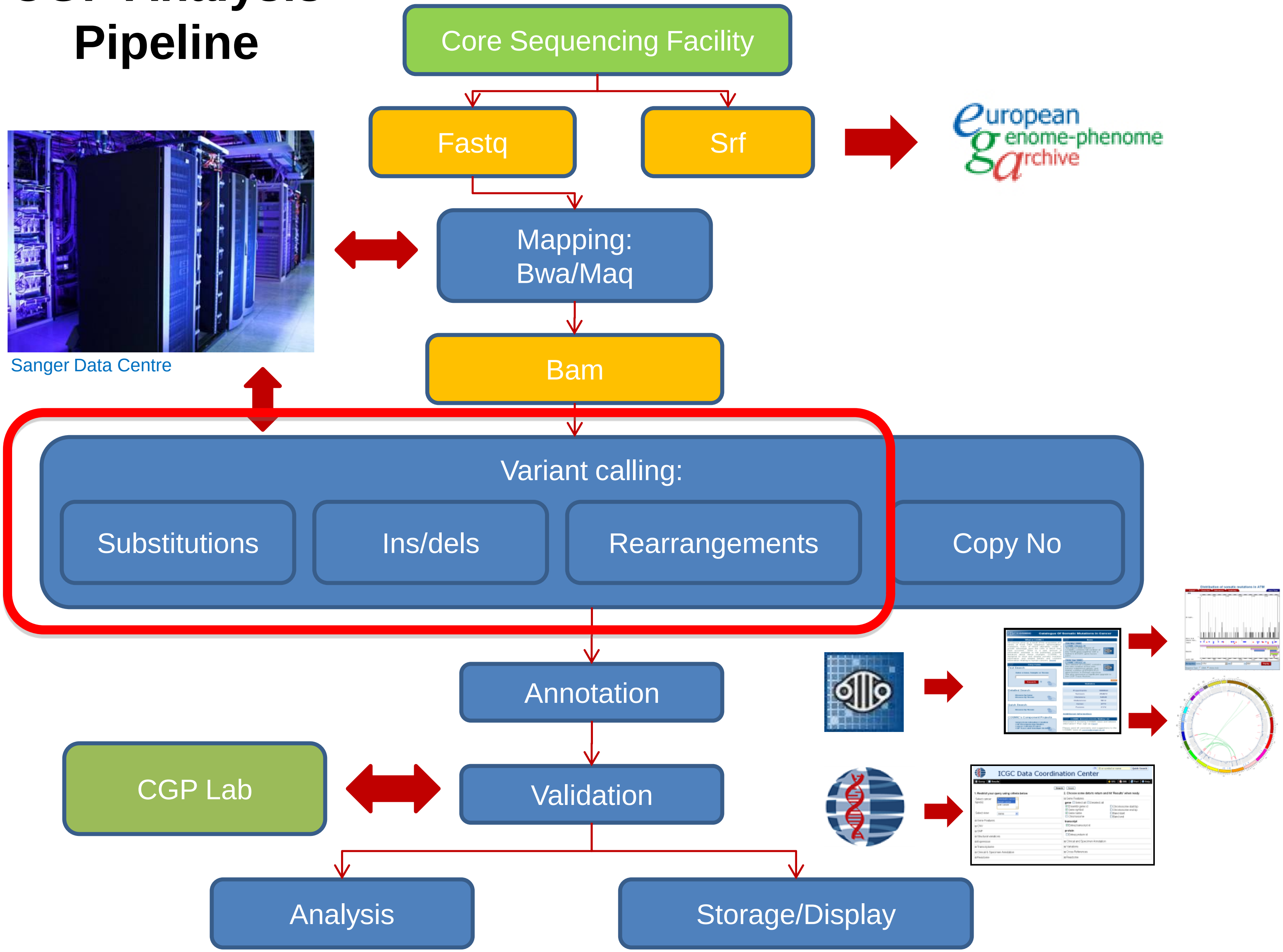
Bioclipse biomaRt-BioConductor Cytoscape Galaxy Gitools Ruby API Taverna WebLab

COSMIC Exports

ORACLE Export

Text

CGP Analysis Pipeline



Core Sequencing Facility

Fastq

Srf

european genome-phenome archive

Mapping:
Bwa/Maq

Sanger Data Centre

Bam

Variant calling:

Substitutions

Ins/dels

Rearrangements

Copy No

Annotation

CGP Lab

Validation

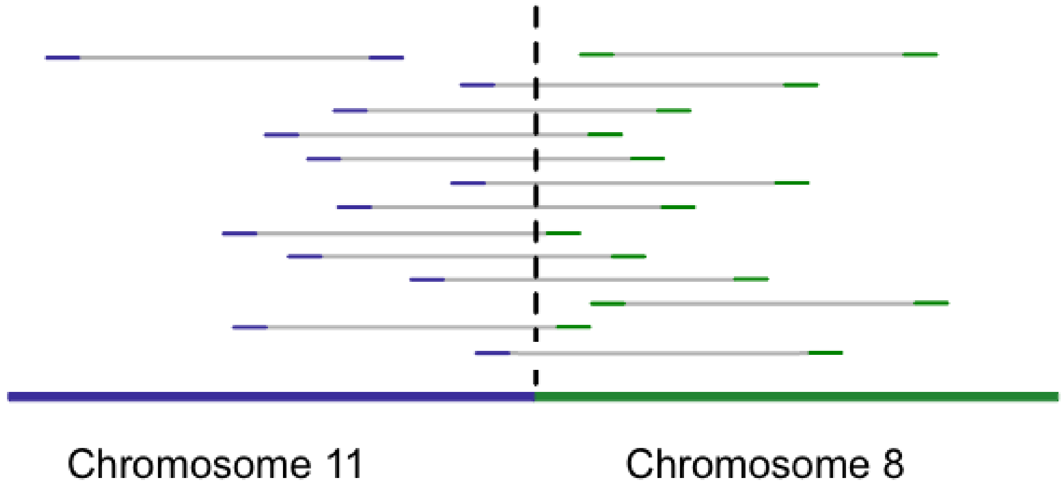
Analysis

Storage/Display

Catalogue Of Somatic Mutations In Cancer
Distribution of somatic mutations in ATM
ICGC Data Coordination Center

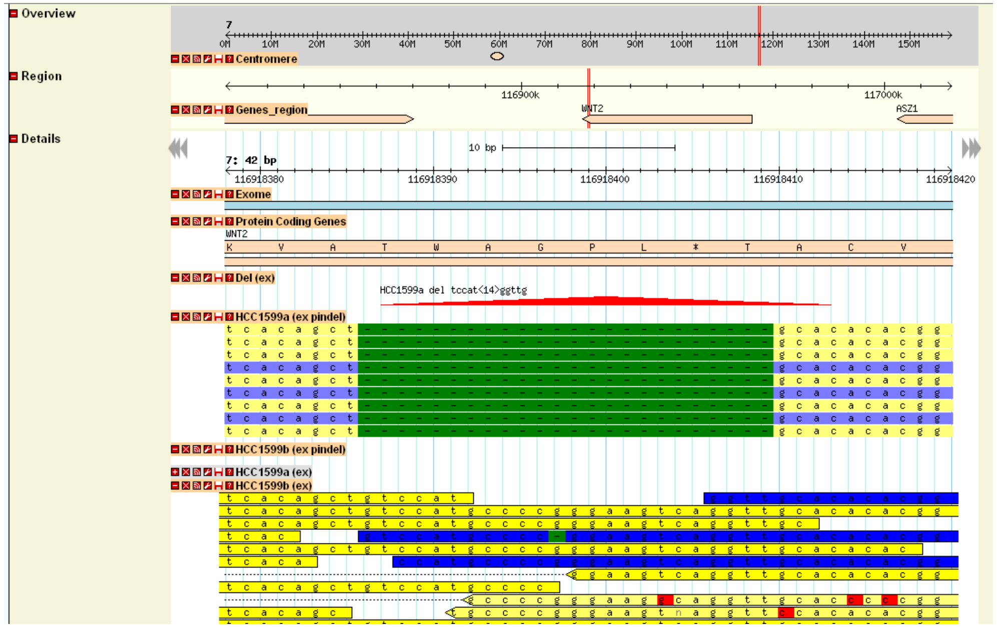
Variant Calling: Rearrangements

- See poster ‘Identifying Structural Rearrangements via Local Assembly of Next-Generation Sequence Data’ by John Marshall
- Identify informative read-pairs
  - Individual reads map accurately
  - Read-pair form an unexpected insert size
- Group read-pairs spanning the same putative break point
- Use the Velvet *de novo* assembler to reconstruct the sequence across the break



Variant Calling: Insertions/Deletions

- Pindel – K Ye *et al* Bioinformatics 2009
- Uses the read-pairs with:
    - One uniquely mapped read
    - One read is either unmapped or has gapped alignment
- Pindel then realigns the mis-mapped reads expecting greater divergence from the reference than BWA/MAQ



Variant Calling: Substitution

- Introduction to CaVEMan (**Ca**ncer **V**ariants through **E**xpectation **Ma**ximisatio**n**)
- Usage / Requirements
- Post processing
- Results

What is CaVEMan?

- Single base substitution calling algorithm
- Java
- Indexed bam as input
- Picard[1]
- 3 way comparison
  - reference
  - normal + tumour
- Expectation Maximisation algorithm[2]

[1] http://picard.sourceforge.net/index.shtml

[2] C.B. Do & S. Batzoglou (2008). ‘What is the expectation maximization algorithm?’. Nat Biotech 26(8):897-899

Expectation Maximisation Algorithm

- Two step iterative algorithm

- M(aximisation) step
  - Build profile of sequencing errors
  - reference base, called base, base quality, read position, mapping quality, lane as covariates…

- E(xpectation) step
  - Use profile to call substitutions
  - Naïve Bayesian classifier

What is CaVEMan?

- Modular - can update E-step (sub. calling) parameters without rerunning M-step.
- Flexible
  - More new sequencing technologies coming.
- Can be used on SOLiD (not yet tuned)
- Many optional parameters:

Normal contamination

Include Smith-Waterman reads

SNP probability cut-off

Expected mutation frequency

Reference bias

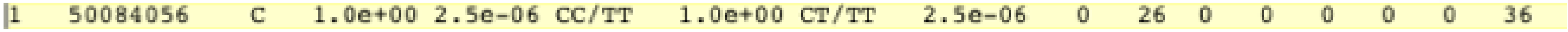
Mutation probability cut-off

Minimum base quality to include

Expected SNP frequency

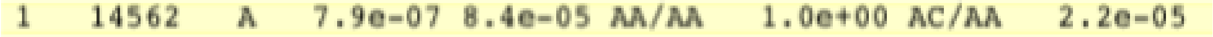
Output of Results

- Result for each genomic position
- Probability assigned for every possible genotype given the reference base
- If above defined cut off - called somatic/SNP, written to file



```
1    50084056    C    1.0e+00 2.5e-06 CC/TT    1.0e+00 CT/TT    2.5e-06    0    26   0    0    0    0    0    36
```

- ‘Normal’ bases - to file, info for every position.



```
1    14562    A    7.9e-07 8.4e-05 AA/AA    1.0e+00 AC/AA    2.2e-05
```

Usage / Requirements

Usage

- Designed for a compute farm environment.
- Initial split step - farm sized chunks.

	Exome	Genome
Jobs	~200	2000-3000
CPU sec/job	3372	6850
Mem (Mb)/job	3500	3500

Usage

- Somatic Calls (before post processing)

	avg.	min.	max.
Exome	8513.6	557	64824
Genome	84551.4	40178	163155

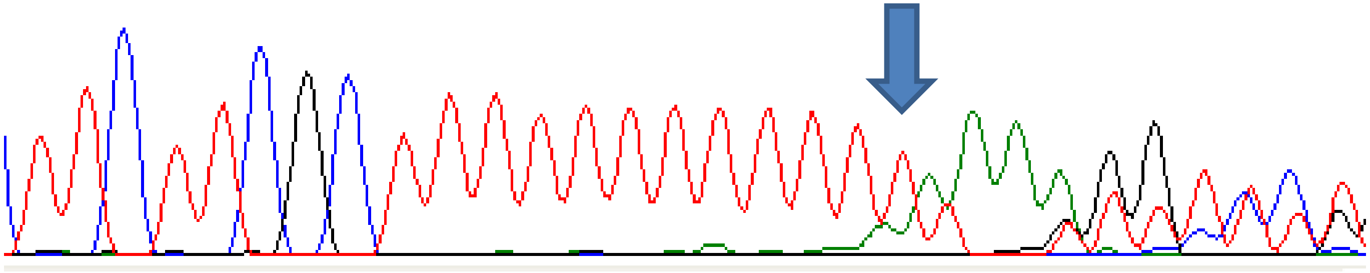
Post processing

Capillary confirmation before post processing

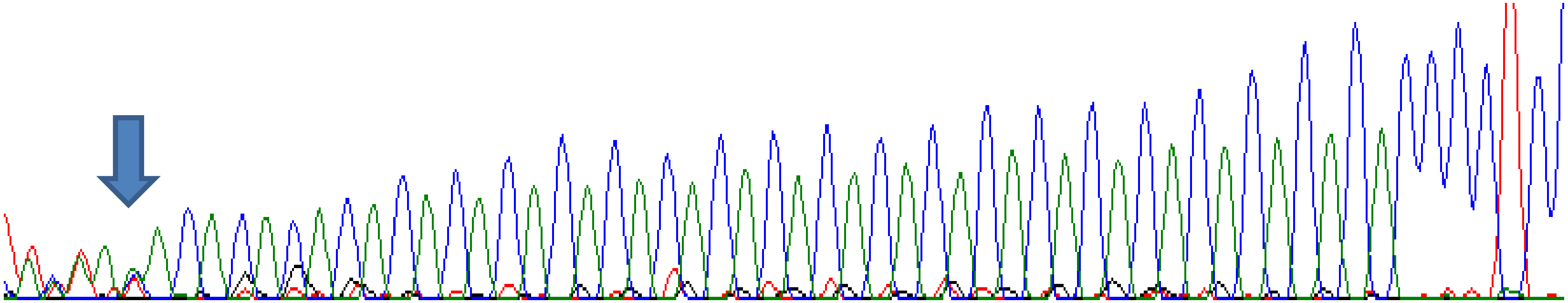
- Better to overcall than miss potentially interesting substitutions
    - Sensitivity over specificity
    - Computationally less intensive to filter after calling
- Example from exomic data:
    - 935 Putative somatic substitutions
    - 131 confirmed as somatic (14%)

Minor causes of false positives in exome data

Slippage at mononucleotide tracts

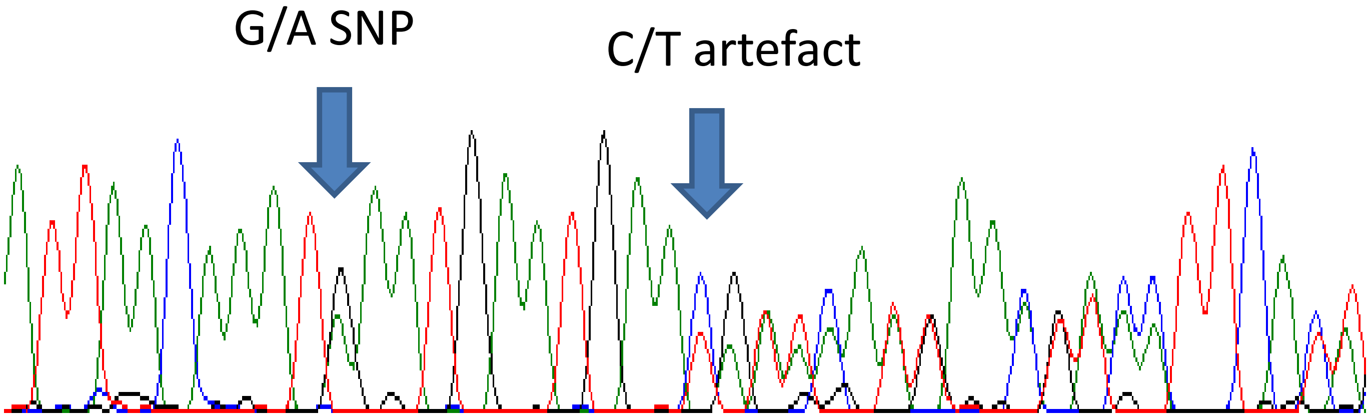


Slippage at poly (n) nucleotide tracts



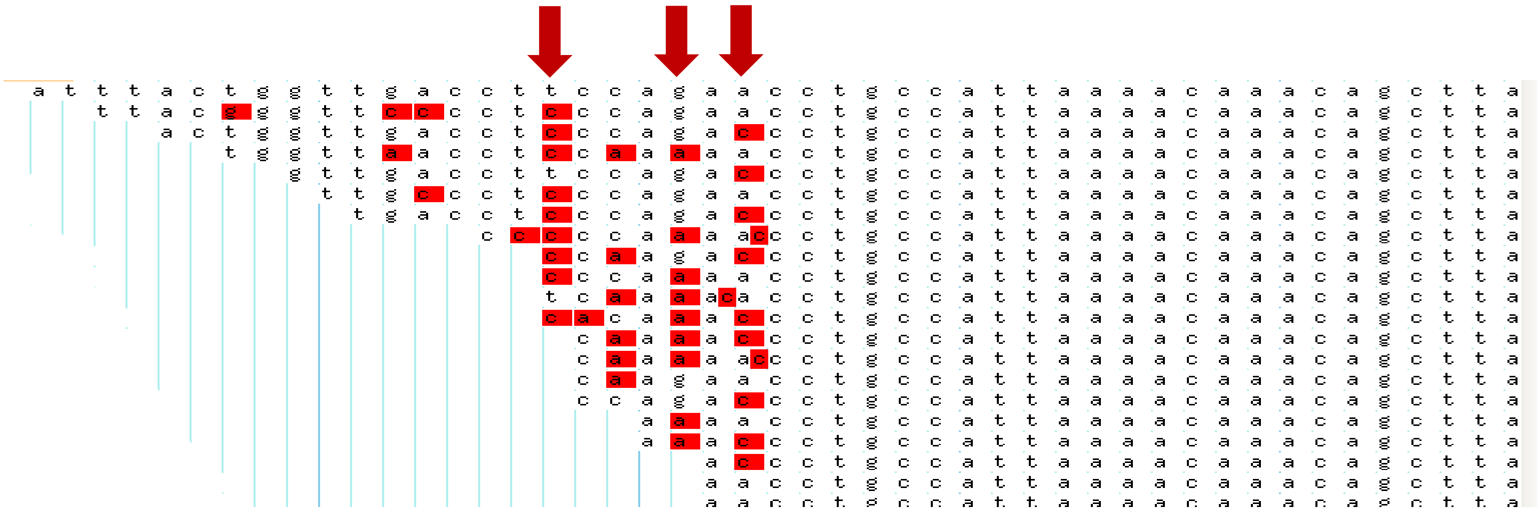
Minor causes of false positives in exome data

Germline INDELS (with & without SNPs)

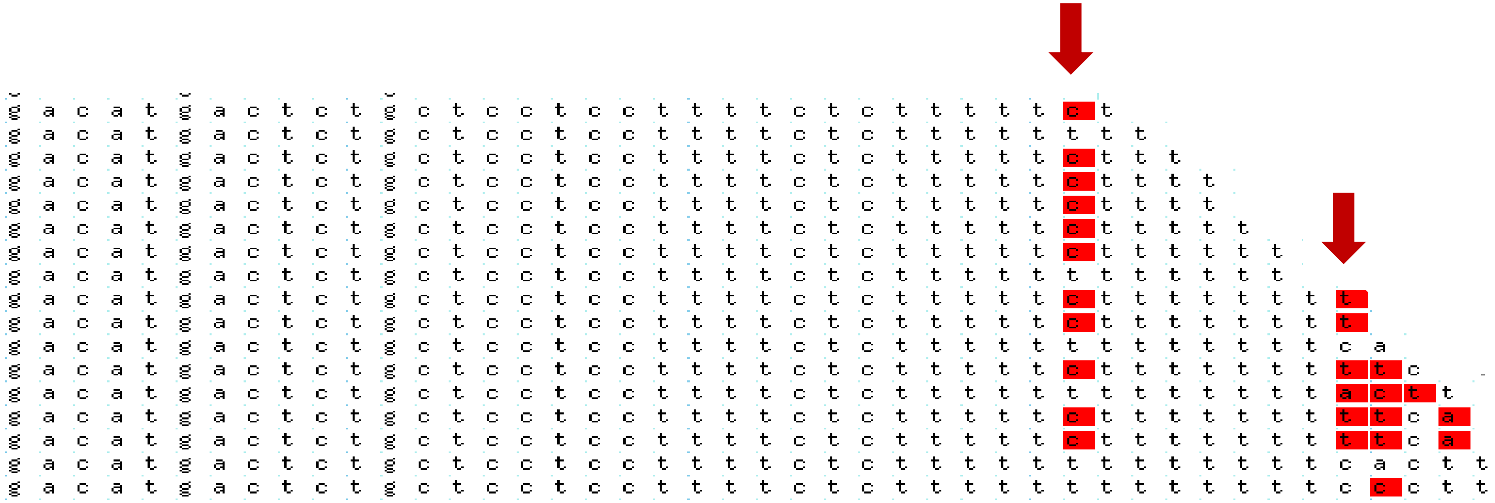


Major causes of false positives in exome data

i) Poor quality data at ends of reads



```
a t t t a c t g g t t g a c c t t c c a g a a c c t g c c a t t a a a a c a a a c a g c t t a
    t t a c g g g t t c c c c t c c c a g a a c c t g c c a t t a a a a c a a a c a g c t t a
        a c t g g t t g a c c t c c c a g a c c c t g c c a t t a a a a c a a a c a g c t t a
            t g g t t a a c c t c c a a a a a c c t g c c a t t a a a a c a a a c a g c t t a
                g t t g a c c t t c c a g a c c c t g c c a t t a a a a c a a a c a g c t t a
                  t t g c c c t c c c a g a a c c t g c c a t t a a a a c a a a c a g c t t a
                    t g a c c t c c c a g a c c c t g c c a t t a a a a c a a a c a g c t t a
                            c c c c c a a a acc c t g c c a t t a a a a c a a a c a g c t t a
                                c c a a g a c c c t g c c a t t a a a a c a a a c a g c t t a
                                c c c a a a a c c t g c c a t t a a a a c a a a c a g c t t a
                                t c a a aca c c t g c c a t t a a a a c a a a c a g c t t a
                                c a c a a a c c c t g c c a t t a a a a c a a a c a g c t t a
                                  c a a a a c c c t g c c a t t a a a a c a a a c a g c t t a
                                  c a a a acc c t g c c a t t a a a a c a a a c a g c t t a
                                  c a a g a a c c t g c c a t t a a a a c a a a c a g c t t a
                                  c c a g a c c c t g c c a t t a a a a c a a a c a g c t t a
                                      a a a a c c t g c c a t t a a a a c a a a c a g c t t a
                                      a a a c c c t g c c a t t a a a a c a a a c a g c t t a
                                          a c c c t g c c a t t a a a a c a a a c a g c t t a
                                          a a c c t g c c a t t a a a a c a a a c a g c t t a
                                          a a c c t g c c a t t a a a a c a a a c a g c t t a
```



```
g a c a t g a c t c t g c t c c t c c t t t t c t c t t t t t c t
g a c a t g a c t c t g c t c c t c c t t t t c t c t t t t t t t t
g a c a t g a c t c t g c t c c t c c t t t t c t c t t t t t c t t t
g a c a t g a c t c t g c t c c t c c t t t t c t c t t t t t c t t t t
g a c a t g a c t c t g c t c c t c c t t t t c t c t t t t t c t t t t
g a c a t g a c t c t g c t c c t c c t t t t c t c t t t t t c t t t t t
g a c a t g a c t c t g c t c c t c c t t t t c t c t t t t t c t t t t t t
g a c a t g a c t c t g c t c c t c c t t t t c t c t t t t t t t t t t t t
g a c a t g a c t c t g c t c c t c c t t t t c t c t t t t t c t t t t t t t t
g a c a t g a c t c t g c t c c t c c t t t t c t c t t t t t c t t t t t t t t
g a c a t g a c t c t g c t c c t c c t t t t c t c t t t t t t t t t t t t t c a
g a c a t g a c t c t g c t c c t c c t t t t c t c t t t t t c t t t t t t t t t c
g a c a t g a c t c t g c t c c t c c t t t t c t c t t t t t t t t t t t t t a c t t
g a c a t g a c t c t g c t c c t c c t t t t c t c t t t t t c t t t t t t t t t c a
g a c a t g a c t c t g c t c c t c c t t t t c t c t t t t t c t t t t t t t t t c a
g a c a t g a c t c t g c t c c t c c t t t t c t c t t t t t t t t t t t t t c a c t t
g a c a t g a c t c t g c t c c t c c t t t t c t c t t t t t t t t t t t t t c c c t t
```

Post processing substitutions

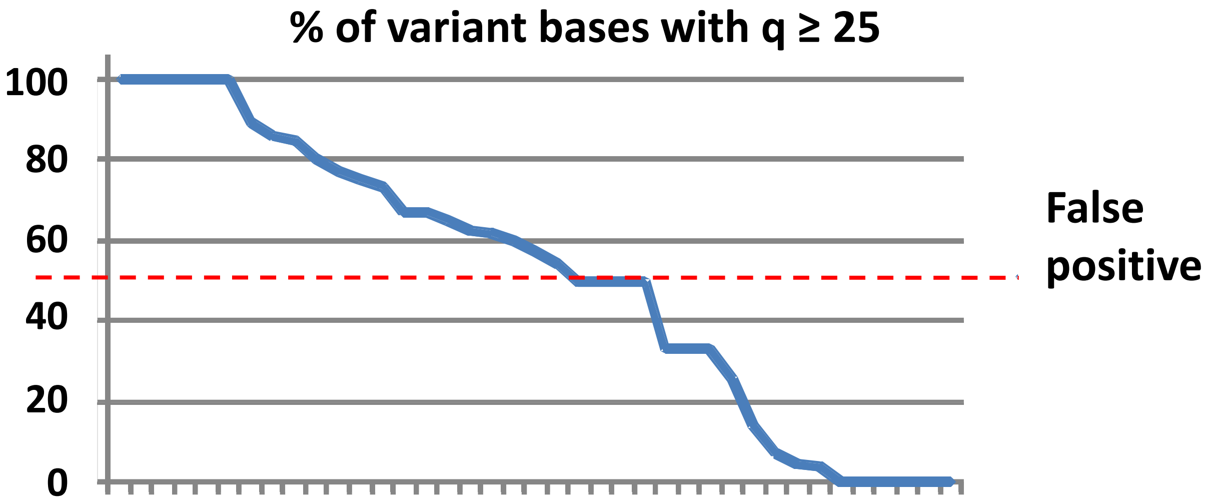
If coverage (in tumour) ≥10,

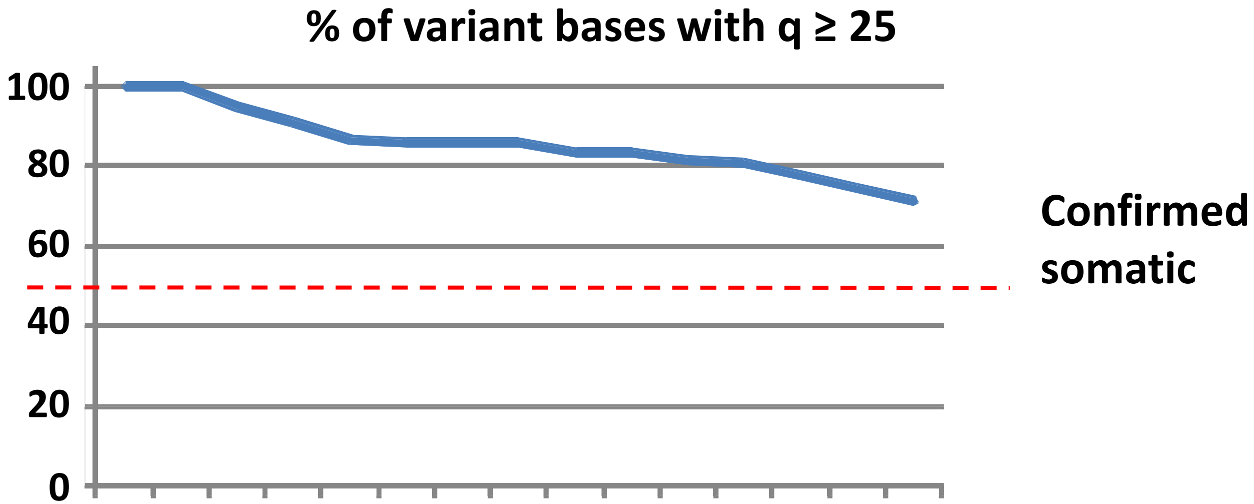
≥1 base call reporting a variant in the 2nd third of the read.

OR

If coverage (in tumour) ≤10,
≥ 1 base call reporting a variant in the 1st or 2nd third of the read.

ii) Low quality bases called as mutations





Post processing substitutions

If coverage (in tumour) ≥10,

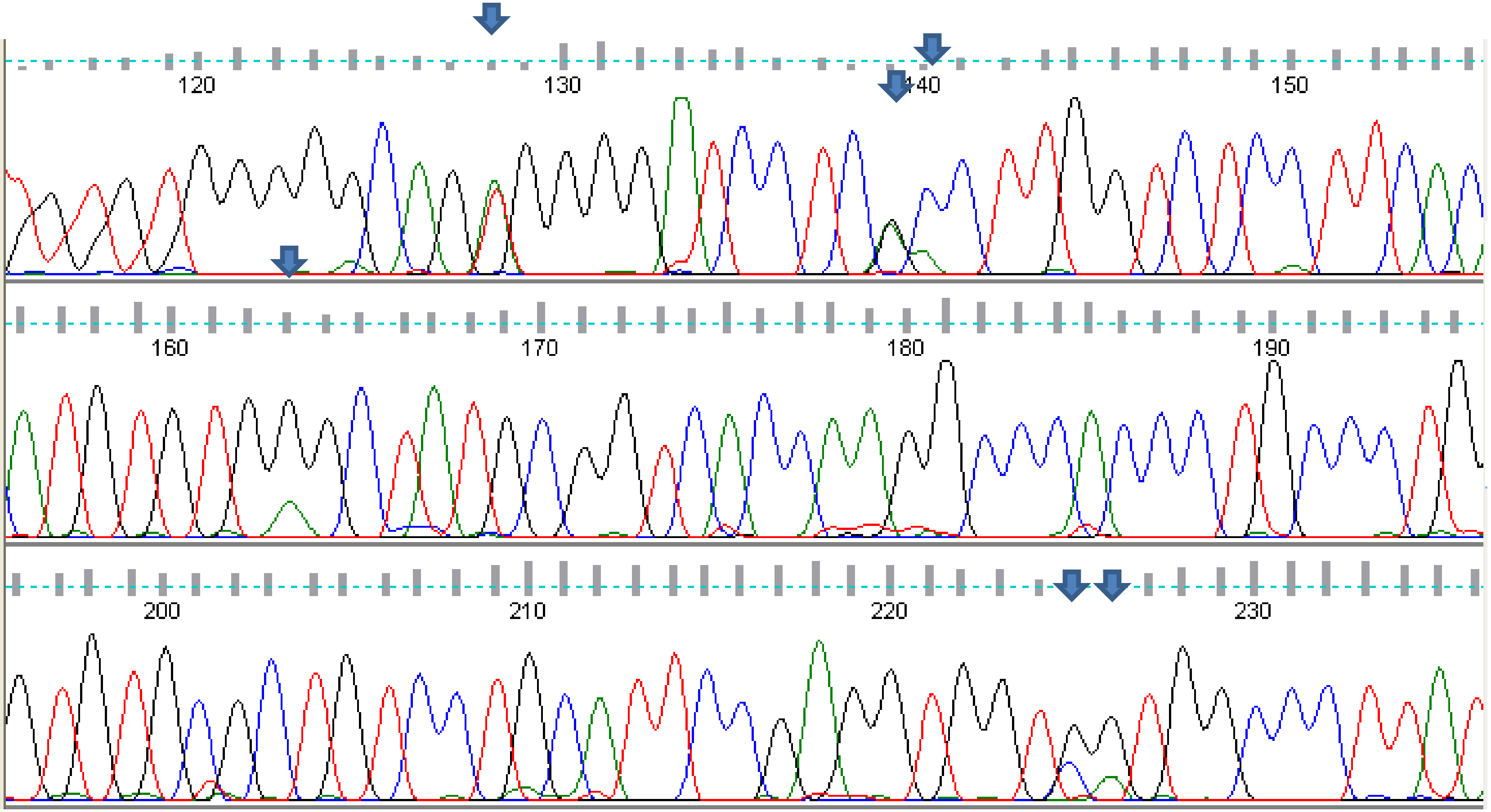
≥1 base call reporting a variant in the 2nd third of the read.

OR

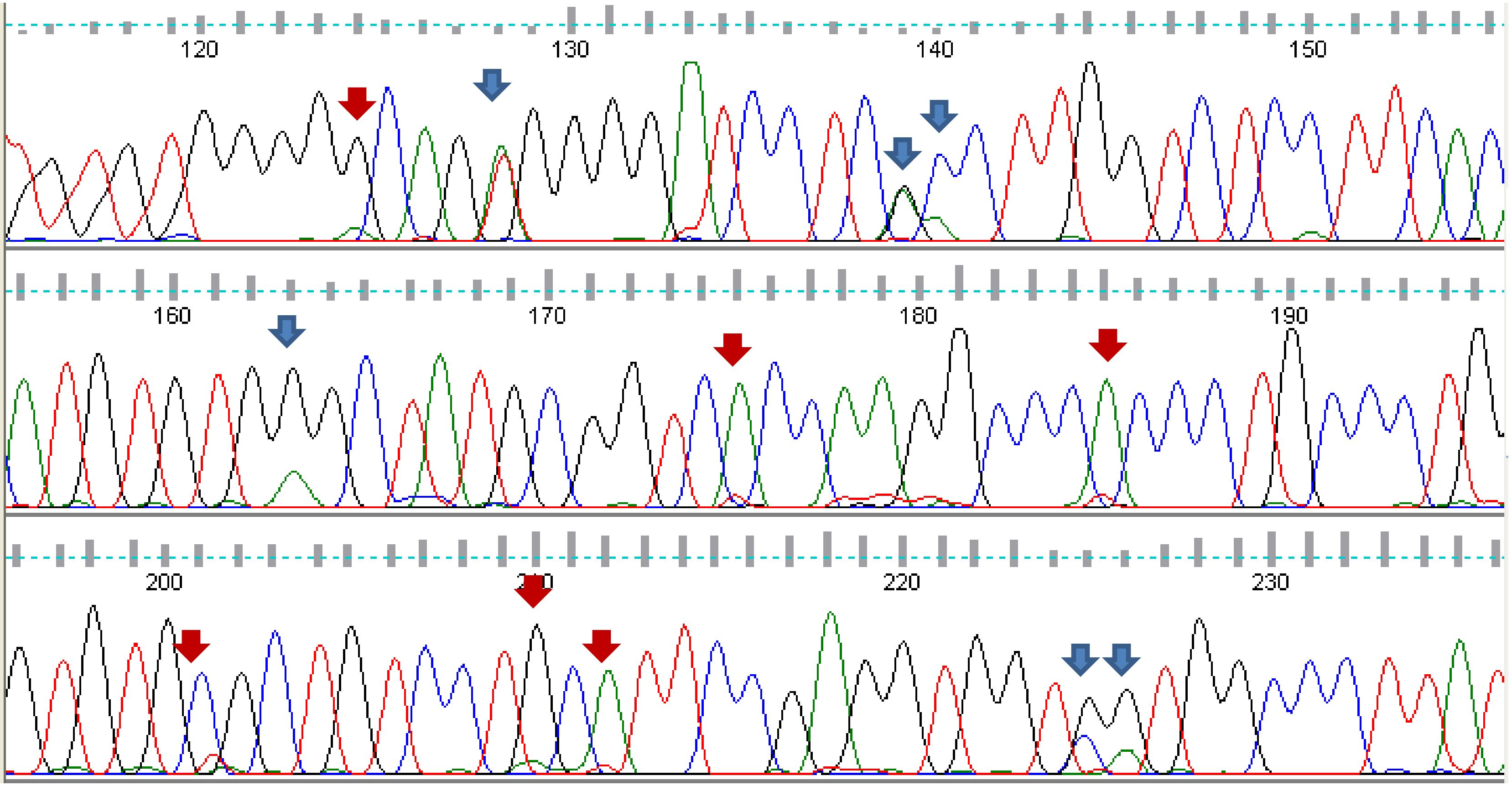
If coverage (in tumour) ≤10,
≥ 1 base call reporting a variant in the 1st or 2nd third of the read.

≥1/3 base calls reporting a variant (in tumour) high quality (≥ 25)

iii) Mapping errors called as mutations



iii) Mapping errors called as mutations



Match	Percent	Chr	Start	Finish
19-334	98.8%	**8**	11226692	11227007
19-334	98.5%	**18**	11634476	11634791
19-334	97.8%	**18**	11600261	11600576
19-334	97.5%	**17**	30544458	30544773
19-334	97.2%	**17**	7326694	7327009

Post processing substitutions

If coverage (in tumour) ≥10,

≥1 base call reporting a variant in the 2nd third of the read.

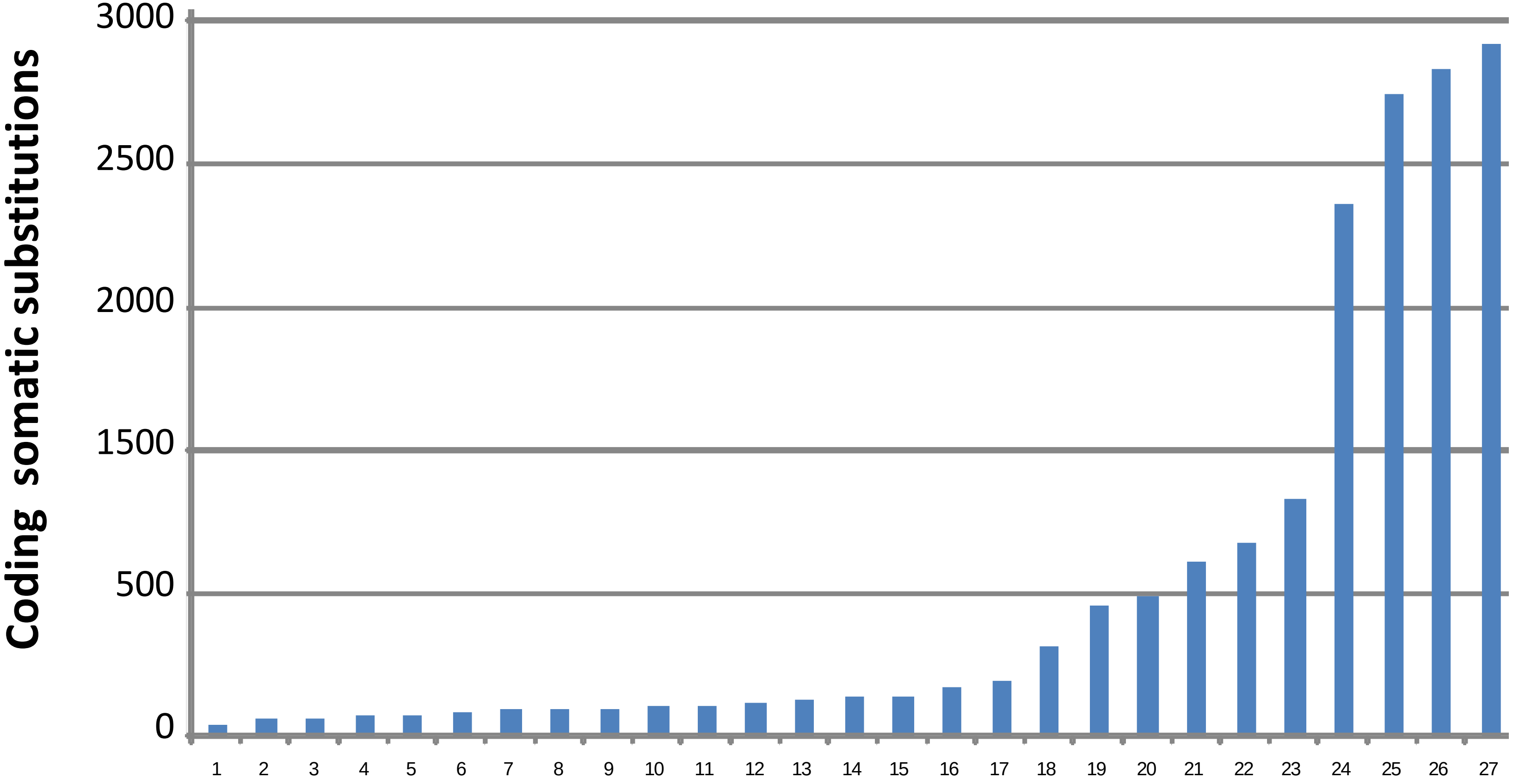
OR

If coverage (in tumour) ≤10,
≥ 1 base call reporting a variant in the 1st or 2nd third of the read.

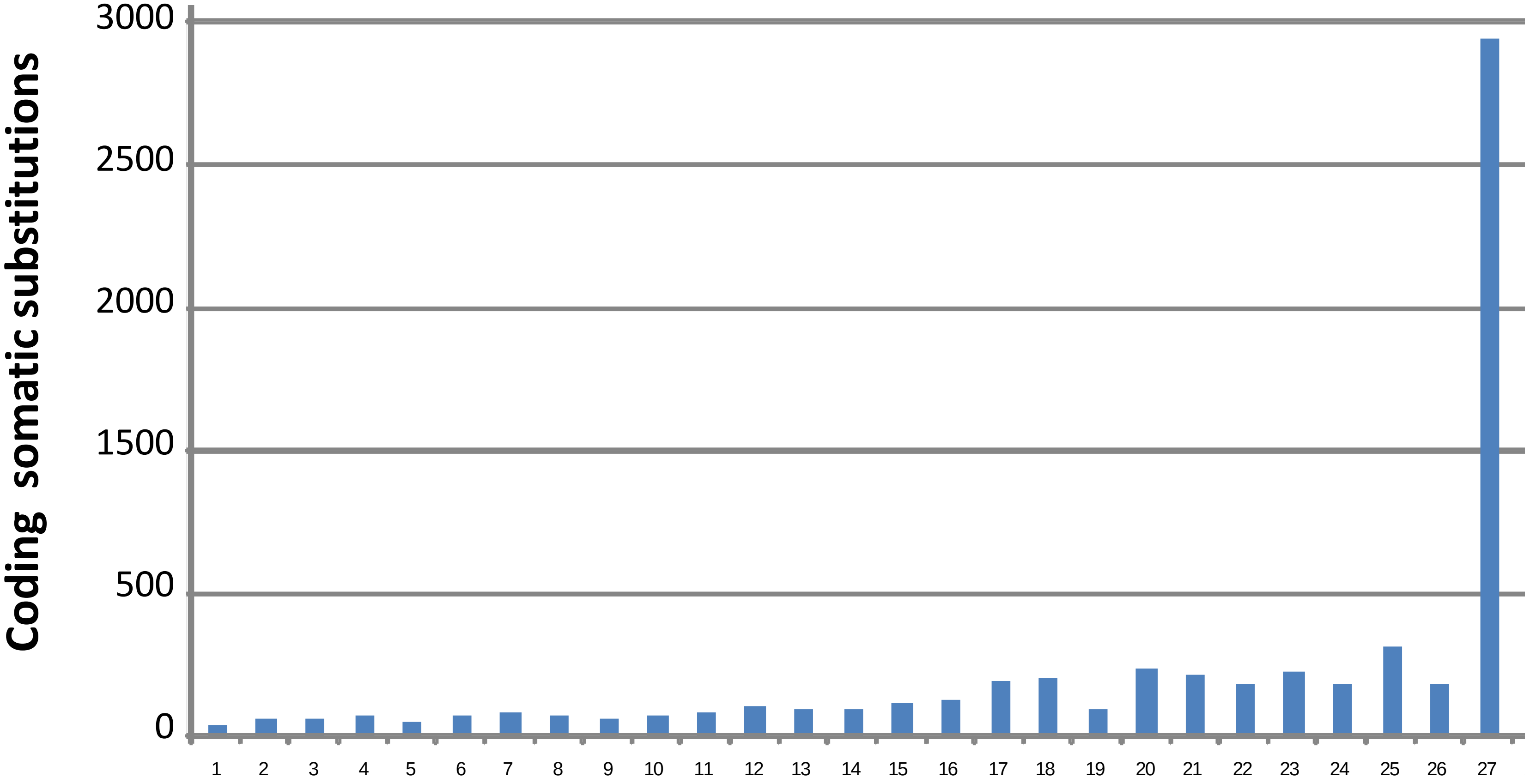
≥1/3 base calls reporting a variant (in tumour) high quality (≥ 25)

≤1 base calls reporting a variant in (ANY of 28 normals) high quality (≥20)

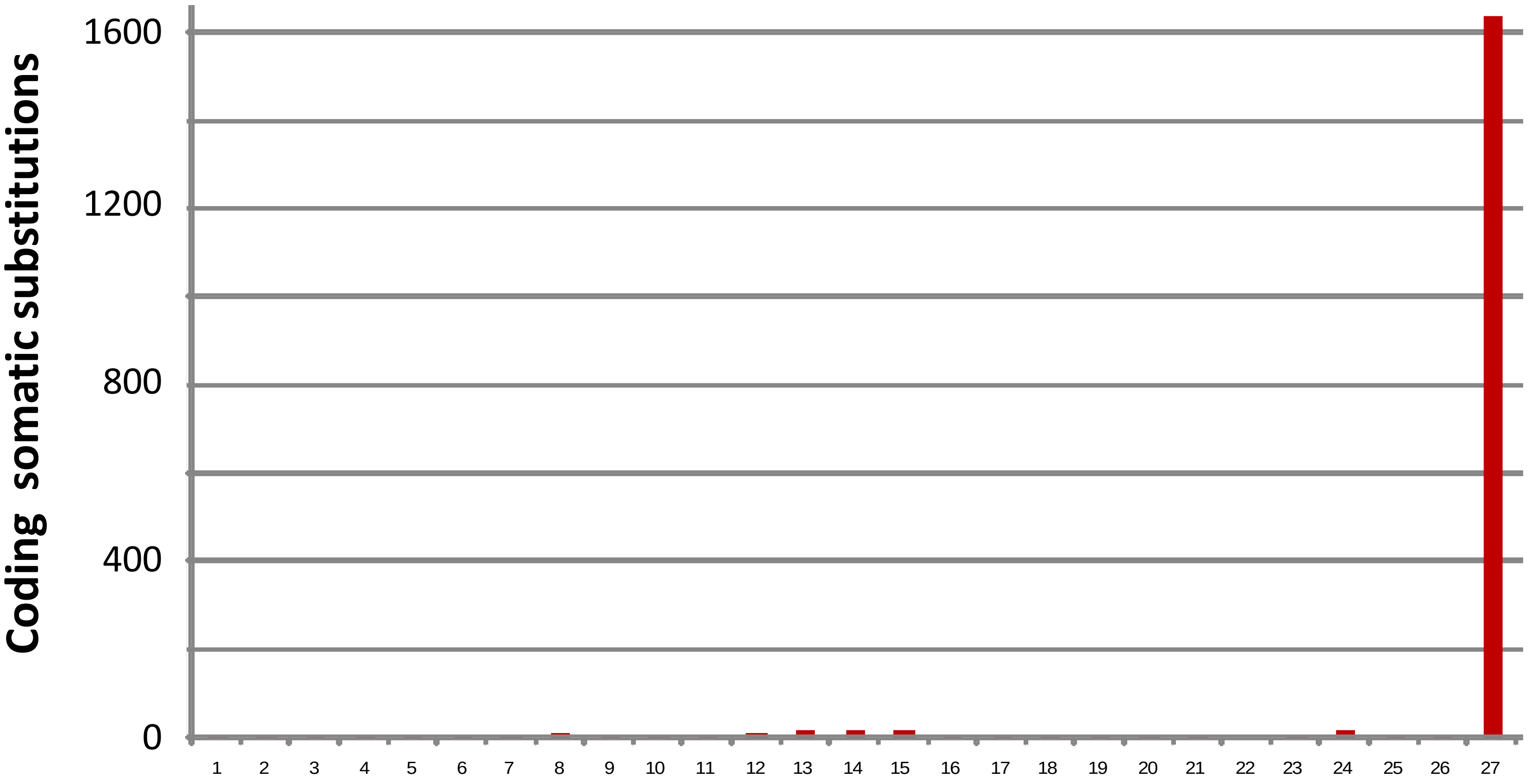
Variants called:- without post processing



Variants called:- with post processing



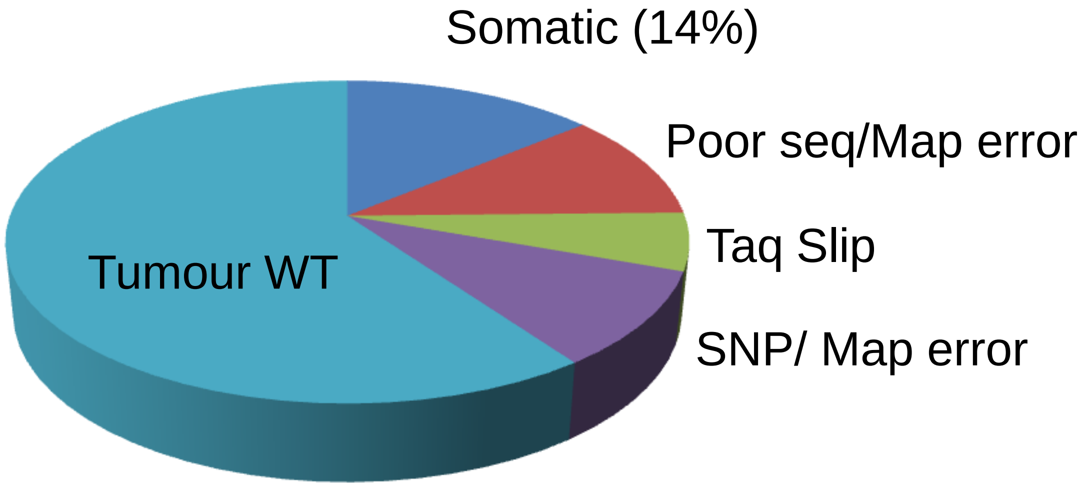
Variants called:- with known SNP I.D

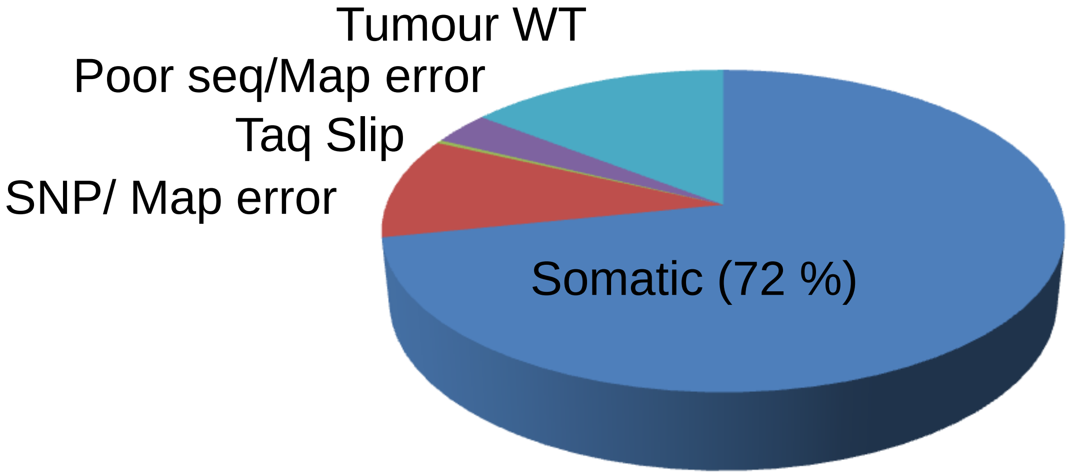


Variants called:-

Without

post processing





After

post processing

Results

COLO-829

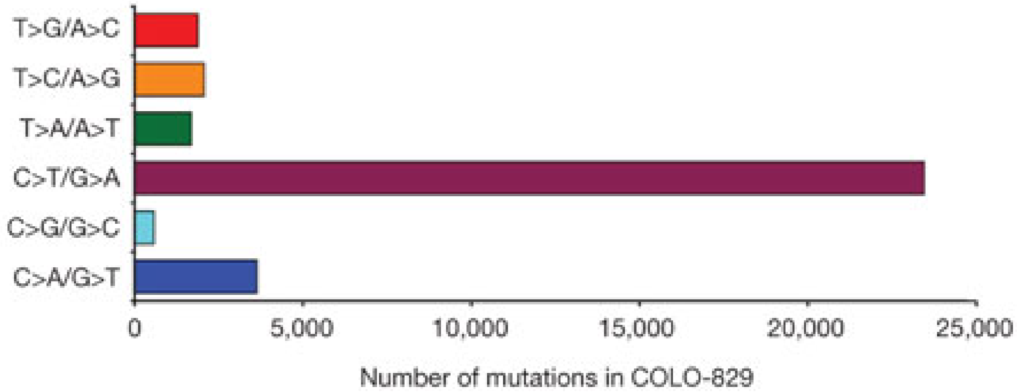
- E.D. Pleasance *et al*.
- 522 validated substitutions - NCBI36[1]
- CaVEMan - GRCh37 (liftover)

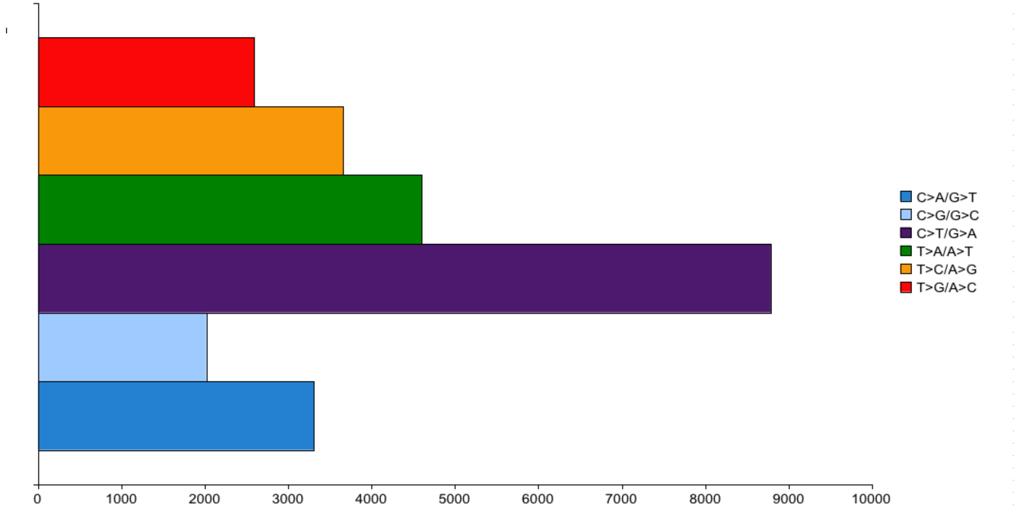
– Missed	3
– Called	519
– Novel (unvalidated)	24965

[1] E.D. Pleasance *et al* (2009). A comprehensive catalogue of somatic mutations from a human cancer genome. Nature 463(7278):191-196

COLO-829 - Novel Calls

- 24965 novel (unvalidated) subs
- Does the mutation spectrum match?





27 primary breast tumour and matching normal exomes

Coding Somatic Substitutions

27 Breast Exomes

- Known cancer genes at frequencies in concordance with literature
  - PIK3CA, TP53, AKT1, NF1, MAP2K4, GATA3, PTEN and CDH1

- ~500 subs in over 400 genes, many in >1 sample
  - Currently validating and evaluating with more breast exomes

Summary

- Pipeline
- CaVEMan
  - Substitution calling Expectation Maximisation algorithm attempting to deal with sequencing errors.
  - Performance / Useage
  - False positives + Post processing
  - Results - COLO-829 and 27 Breast exomes.

Thanks to.....

Peter Campbell

Phil Stephens

Keiran Raine

Serena Nik-Zainal

Ignacio Varela

Adam Butler

Jon Teague

Mike Stratton

Andy Futreal

Rest of the CGP

Wellcome Trust