Sequencing & Analysing Cancer Genomes

Challenges of Variant Calling in Cancer

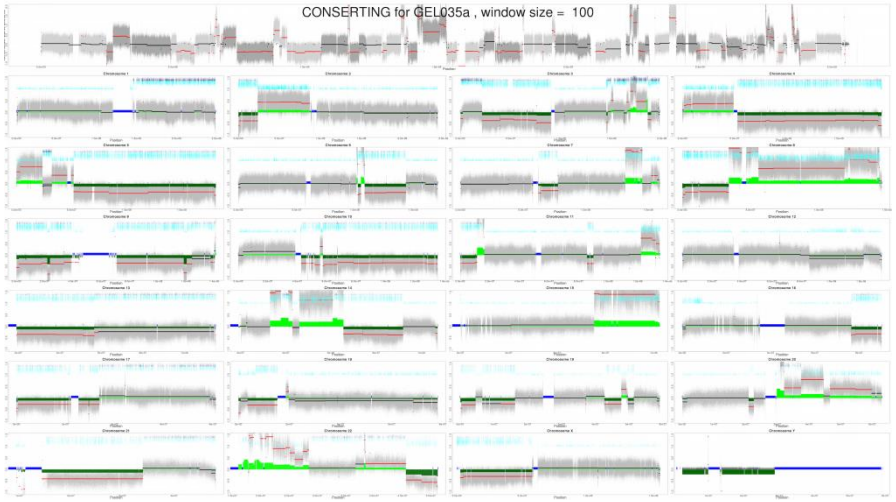
Matthew Parker, Ph. D
Lead Bioinformatician
Sheffield Diagnostic Genetics Service

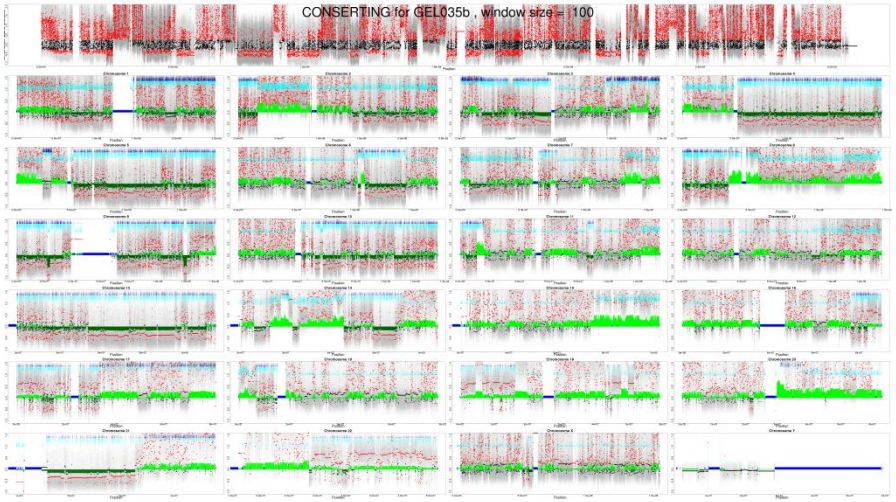
Sheffield Children's NHS
NHS Foundation Trust

The Problems with Cancer

Sheffield Children's NHS
NHS Foundation Trust

- Clinical
  - Turn around times
  - Appropriate samples
  - Variant prioritisation
  - "Action-ability"
- Analysis
  - Quality of DNA
  - Tumor Content
  - Clonality
  - BMT

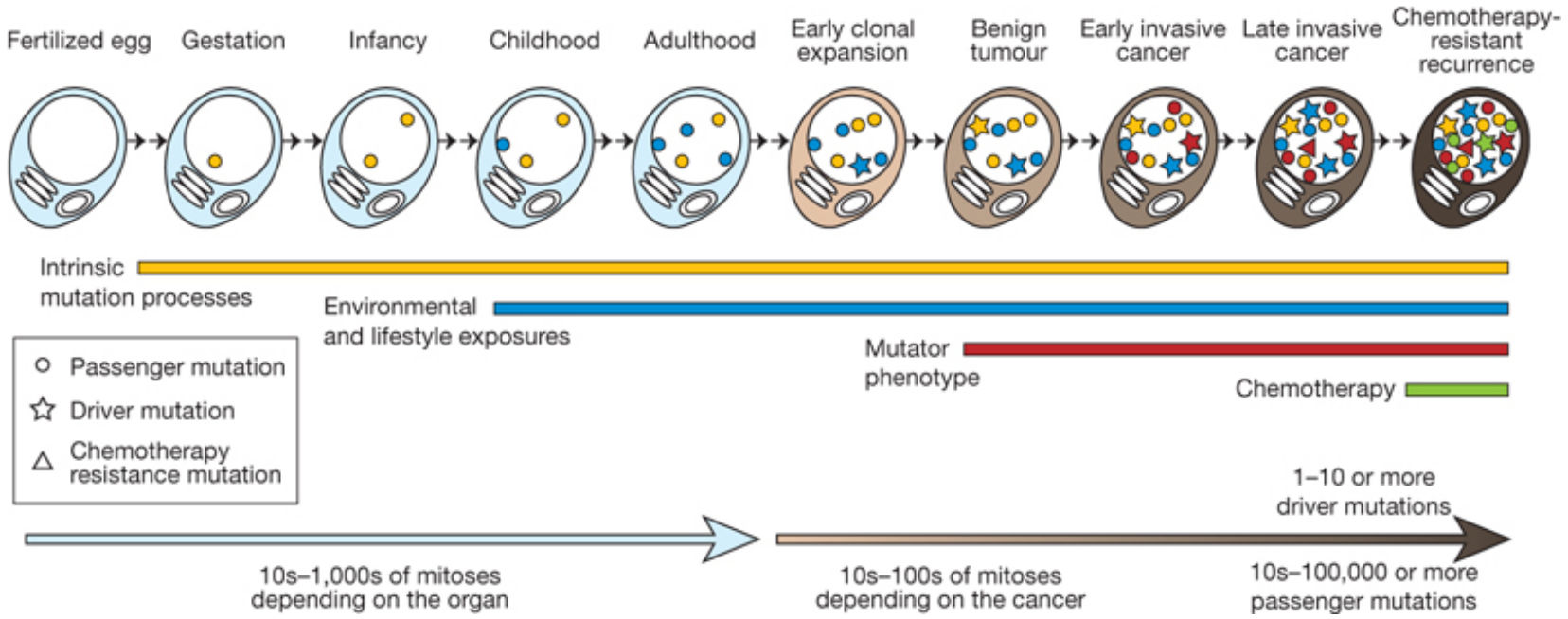




How do you think variants discovered in a cancer genome could be used?

Cancer 101

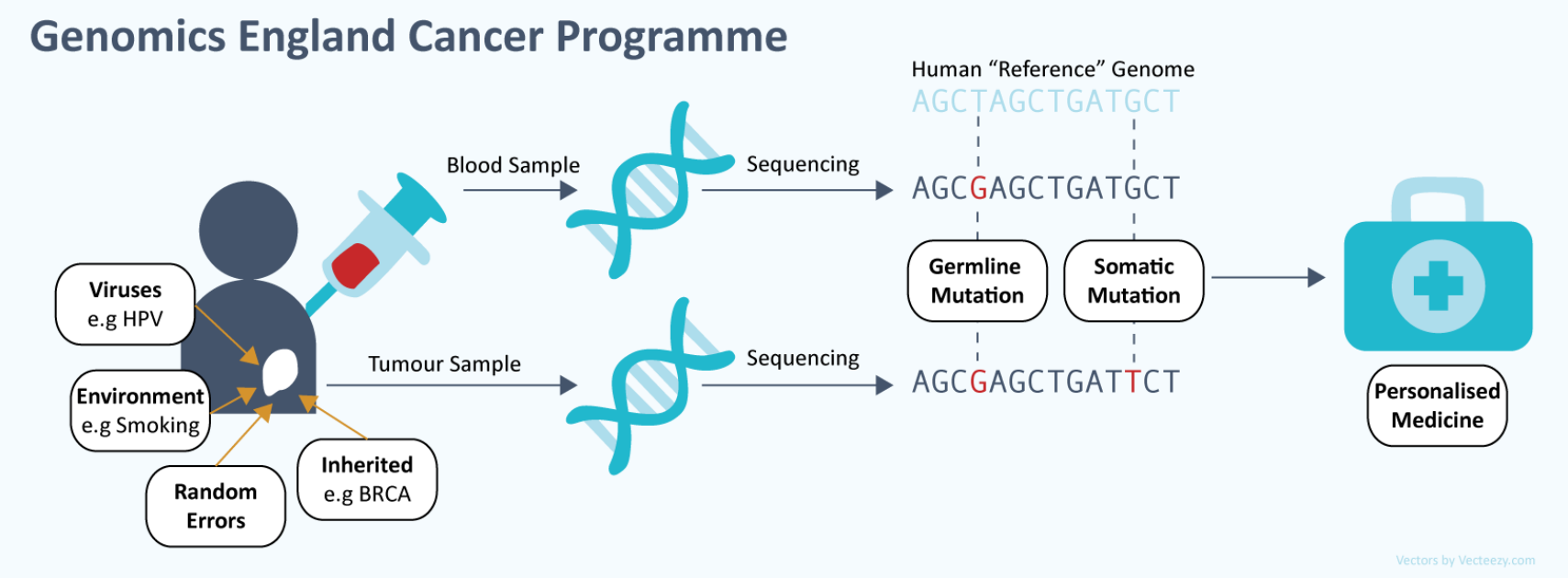
Sheffield Children's NHS
NHS Foundation Trust



Mike Stratton

Standard Cancer Sequencing

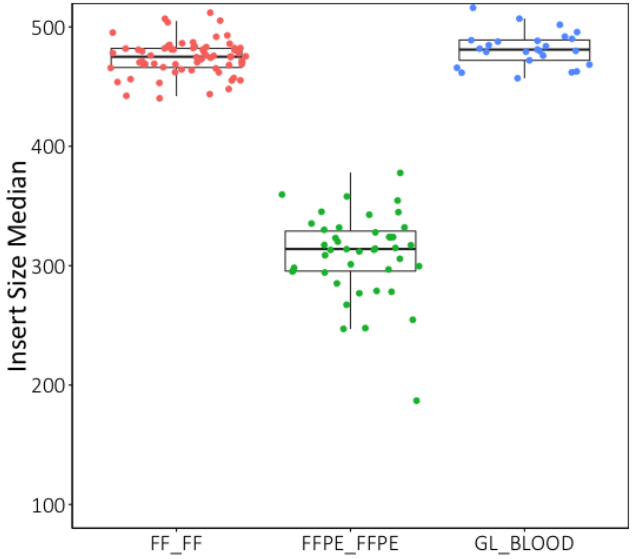
Sheffield Children's NHS
NHS Foundation Trust



Sample Quality

Sheffield Children's NHS
NHS Foundation Trust

- Sample quality affects downstream sequence quality
- Problems:
  - Difficult to get enough DNA sometimes
    - PCR May be required → Not ideal for WGS
  - Necrotic tissue – poor quality DNA
  - Storage method affects quality



Tumour Purity

Sheffield Children's NHS
NHS Foundation Trust

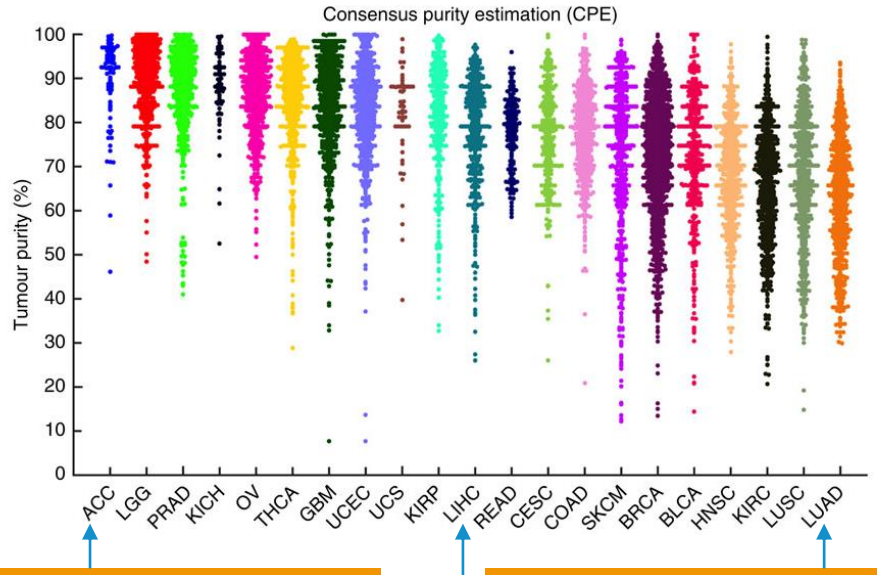
NATURE COMMUNICATIONS | ARTICLE OPEN

Systematic pan-cancer analysis of tumour purity

Dvir Aran, Marina Sirota & Atul J. Butte

Affiliations | Contributions | Corresponding author

Nature Communications **6**, Article number: 8971 | doi:10.1038/ncomms9971
Received 02 September 2015 | Accepted 19 October 2015 | Published 04 December 2015



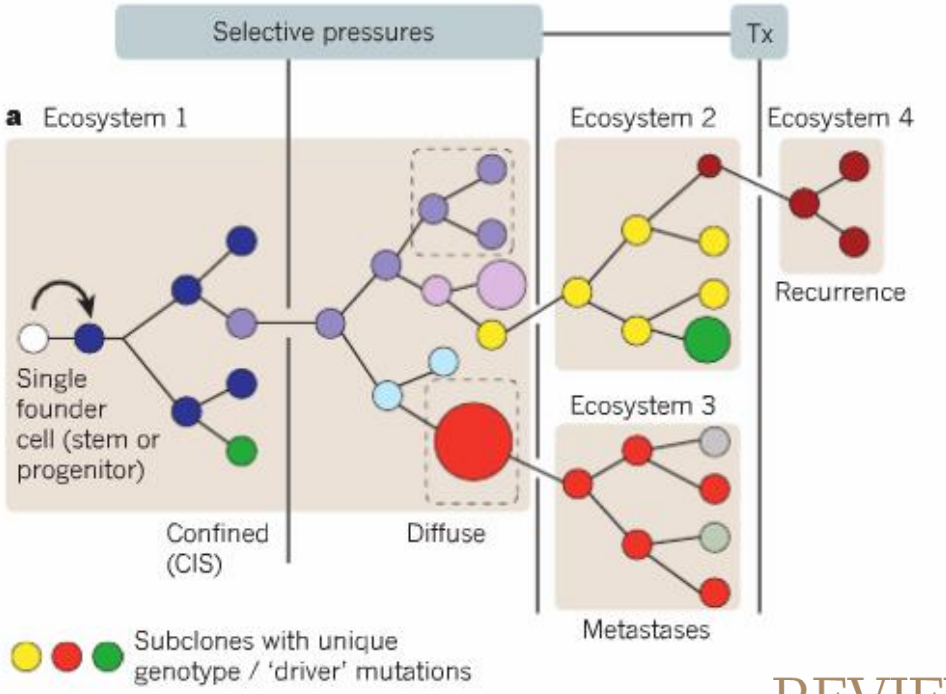
Adenoid cystic carcinoma

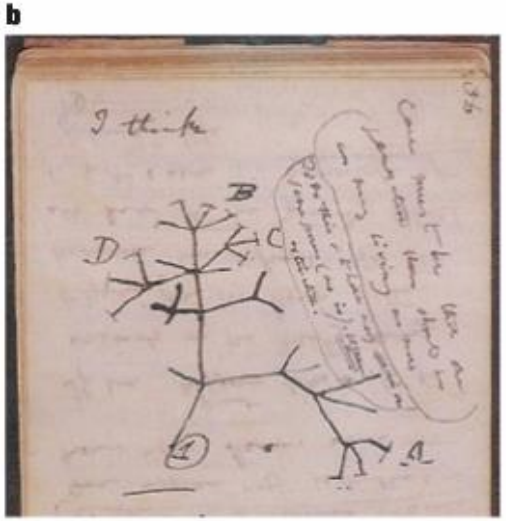
Lung Adenocarcinoma

Liver Hepatocellular Carcinoma

Tumour Heterogeneity (Clonality)

Sheffield Children's NHS
NHS Foundation Trust





REVIEW

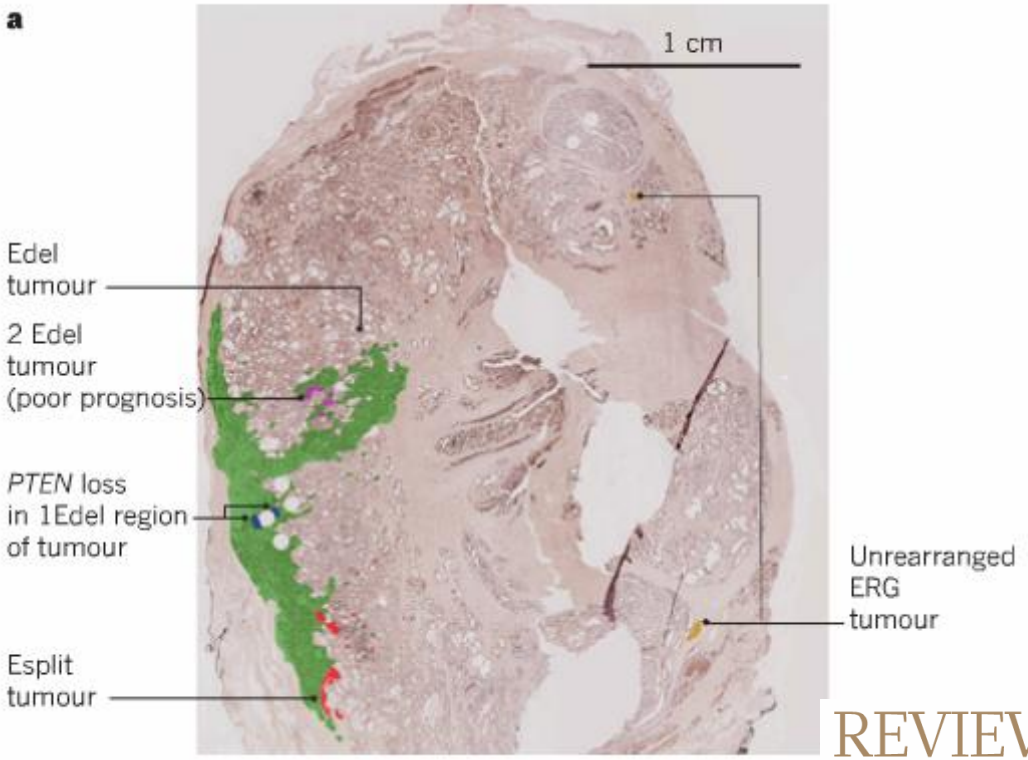
doi:10.1038/nature10762

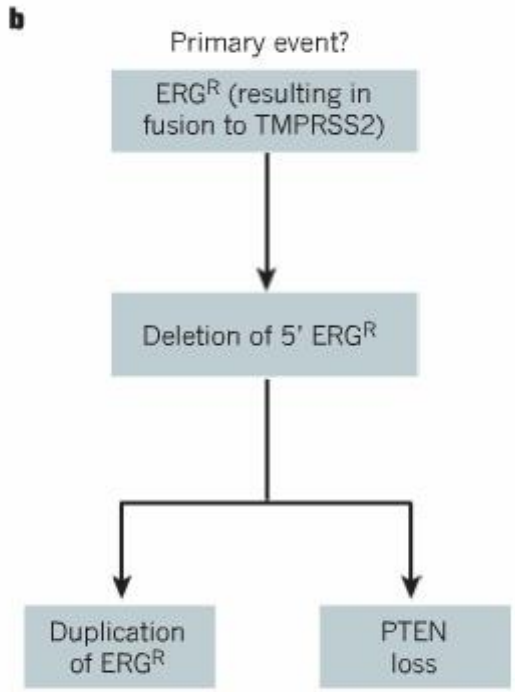
Clonal evolution in cancer

Mel Greaves[1] & Carlo C. Maley[2]

Tumour Heterogeneity

Sheffield Children's NHS
NHS Foundation Trust





REVIEW

doi:10.1038/nature10762

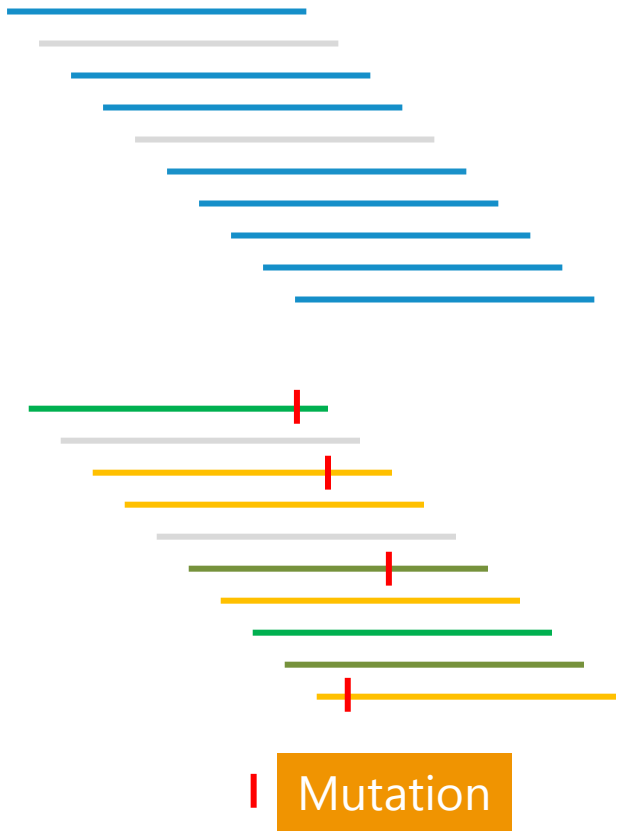
Clonal evolution in cancer

Mel Greaves[1] & Carlo C. Maley[2]

Worked Example

Sheffield Children's NHS
NHS Foundation Trust

- Imagine 80% tumor purity.
- Coverage of 10x
- 8/10 reads on average are from the tumor
- Imagine 3 clones, one at 50%, two at 25%
- Het Mutations – only one read in the two smaller clones
- Only 2 reads supporting variant in largest clone



Coverage is King!

Tumour Contamination of Normal

Sheffield Children's NHS
NHS Foundation Trust

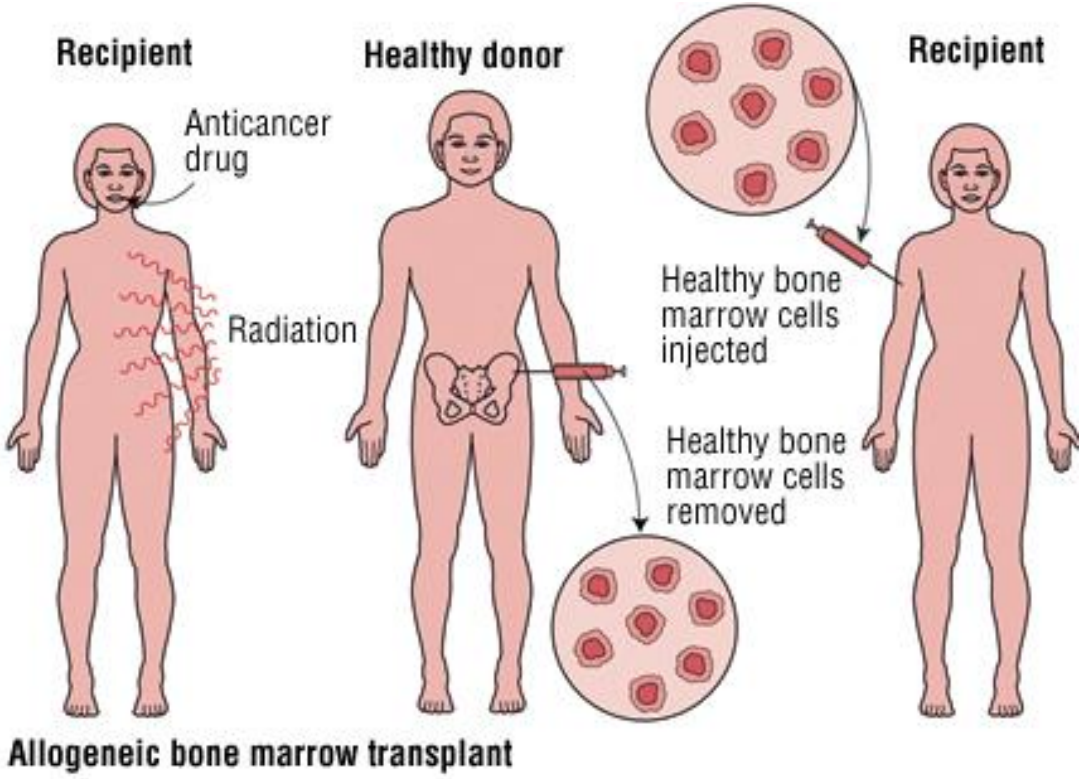
- Evidence of tumor material in the normal sample is sometimes evident
- Tumor cell infiltration of surrounding tissues/blood/saliva
- Some callers can't handle this and the SNV is excluded (for example Illumina!!)

GeneName	SJQuality	Chr	Pos	Tumor MAF	Normal MAF	Total Normal Coverage
HSPA12A	SJHQ	chr10	118464762	0.29	0.05	21
ZNF365	SJHQ	chr10	64415123	0.31	0.00	16
AKR1C1	SJHLQ	chr10	5005711	0.32	0.04	25
GRAMD1B	SJLQ	chr11	123448167	0.13	0.00	14
GIT2	SJHQ	chr12	110389058	0.29	0.06	35
KCNH5	SJHQ	chr14	63174997	0.34	0.07	29
KLC1	SJLQ	chr14	104139491	0.09	0.00	21
HMGN2P46	SJHQ	chr15	45808097	0.33	0.06	18
NR2E3	SJHQ	chr15	72109938	0.28	0.05	21
DSG1	SJHQ	chr18	28936281	0.30	0.04	25
NUDC	SJHQ	chr1	27250594	0.34	0.03	31
ISM1	SJHLQ	chr20	13251341	0.30	0.09	35
TRMU	SJHQ	chr22	46733794	0.29	0.04	24
CADPS	SJHQ	chr3	62459955	0.42	0.06	18
MTTP	SJHQ	chr4	100485294	0.32	0.09	33
ZNF718	SJHQ	chr4	154730	0.31	0.06	31
SETP8	SJHQ	chr5	132096615	0.33	0.05	20
LARP1	SJLQ	chr5	154195506	0.10	0.00	23
PCDHA8	SJLQ	chr5	140223016	0.20	0.00	8
PTPRK	SJHQ	chr6	128290476	0.31	0.09	34
ARHGEF10	SJHQ	chr8	1772199	0.40	0.00	18
OR5C1	SJHQ	chr9	125551552	0.23	0.04	28
GPR143	SJHQ	chrX	9733666	0.34	0.00	15
NHS	SJHQ	chrX	17743982	0.24	0.04	26

Evidence of Contamination?

Bone Marrow Transplant

Sheffield Children's NHS
NHS Foundation Trust



The patients blood is now the blood of the donor – using as a normal sample for somatic calls – not possible

Sequencing

- Because of Purity Heterogeneity – need a relatively high coverage Tumor sample > 30x
- Normal sample can be 20-30x
- Example Projects:
  - 1000G
  - PCGP
  - TCGA
- GEL: Tumor >75x, Normal 35x

Bioinformatically Challenging

- Mapping is the same – usually BWA
- Variant calling is done by algorithms specifically designed for cancer
  - Not A diploid genomes
  - Tumor admixture in the normal
  - Subcloncal variants
  - Copy number variants, copy-neutral LOH or major ploidy chnages
- Trade-off between sensitivity (detecting variants) and specificity (rate of false positives)
- Mutect is a good example of a cancer variant caller

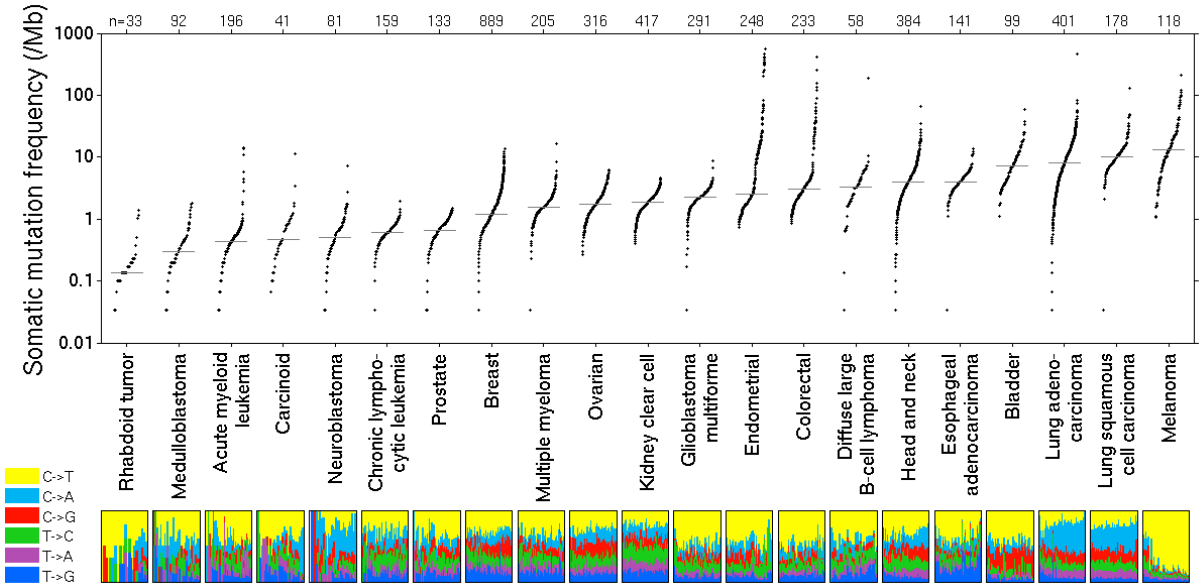
Somatic Variant Calling

- Variants for the cancer (somatic) are called with the germline sample in mind – at the same time.
- You cannot just do a simple subtraction
- Generally agorithms do some simple stats:
  - Probability of the variant being real given the reads from cancer sample and the reads from the germline sample
  - Usually involves post processing of data – filtering etc
  - Mutations found at high % in the population for example are usually filtered out

QC of Cancer Data

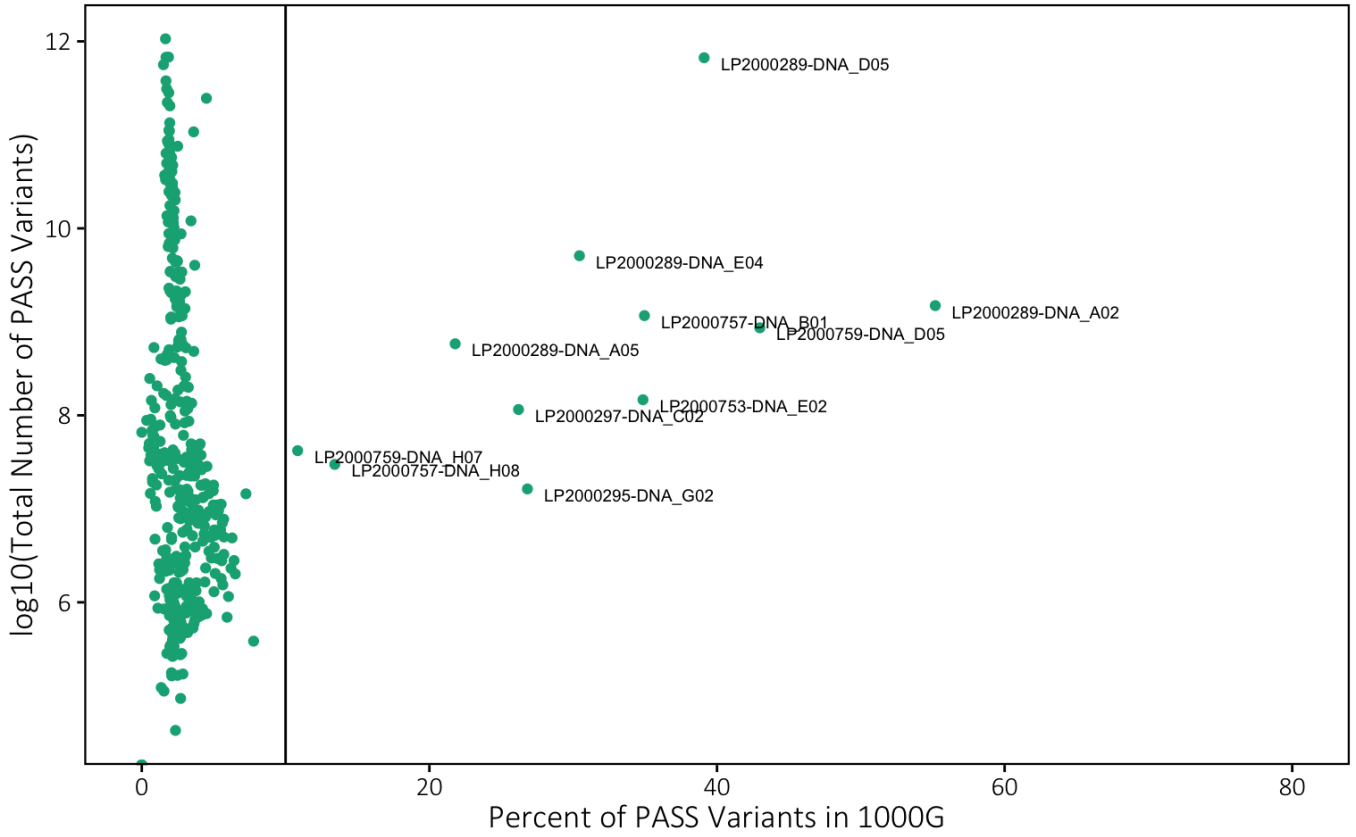
Sheffield Children's NHS
NHS Foundation Trust

- SNVs in cancer allow assessment of:
  - Sample-sample contamination
  - Tumour & Normal provenance – same patient?
  - Numbers/types of variants



Contamination Example

Sheffield Children's NHS
NHS Foundation Trust



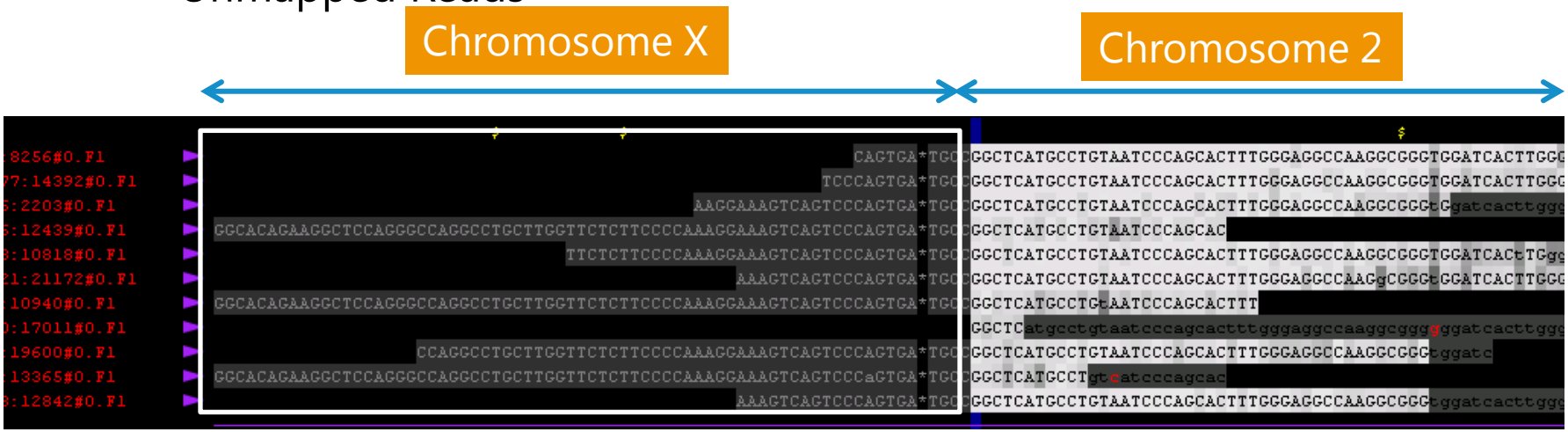
Variant Calling: Indels

- Aligners usually have a difficult time with indels
- Usually happen in more repetitive regions of DNA
- Left alignment

Variant Calling: Structural

Sheffield Children's NHS
NHS Foundation Trust

- 3 sources of information can be used to detect SVs:
  - Soft-clipped reads
  - Discordant read pairs
  - Unmapped Reads

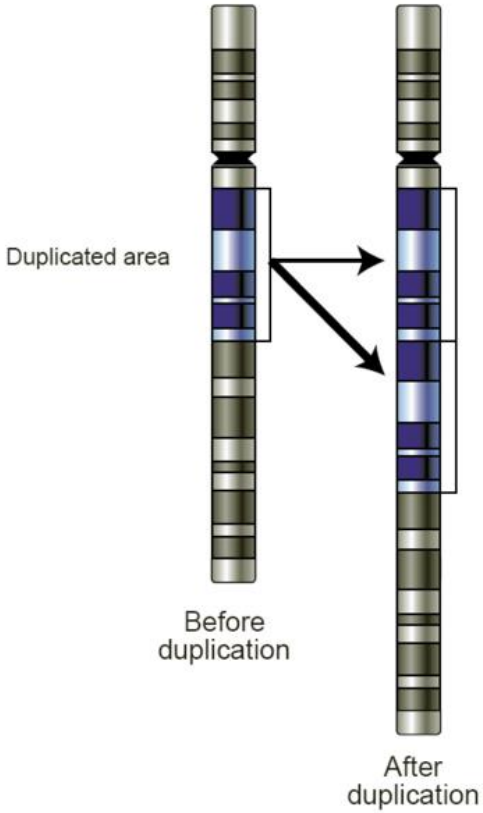


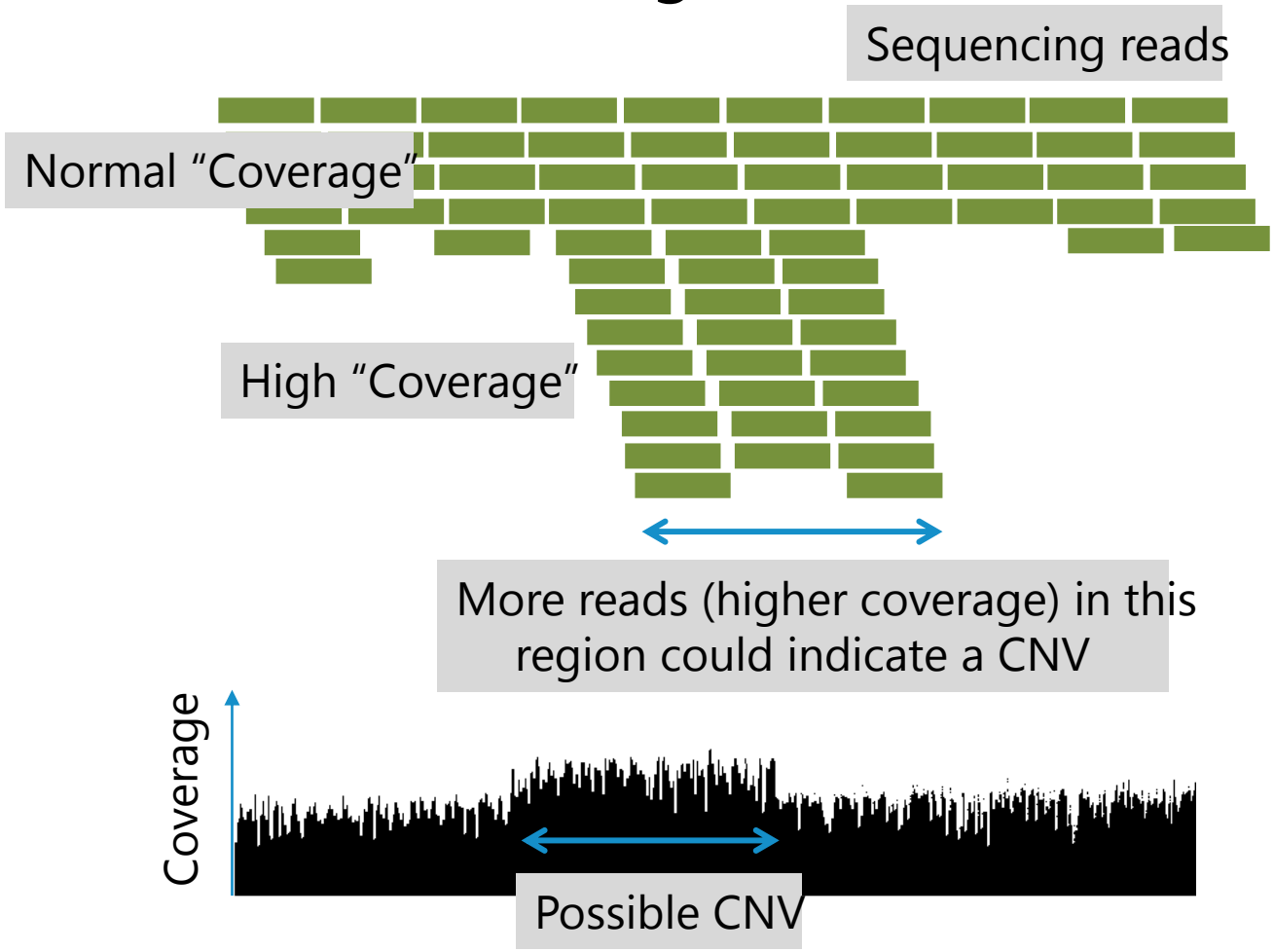
- Example: CREST: fhttp://www.stjuderesearch.org/site/lab/zhang

Variant Calling: Copy Number

Sheffield Children's NHS
NHS Foundation Trust

- Read depth information is leveraged



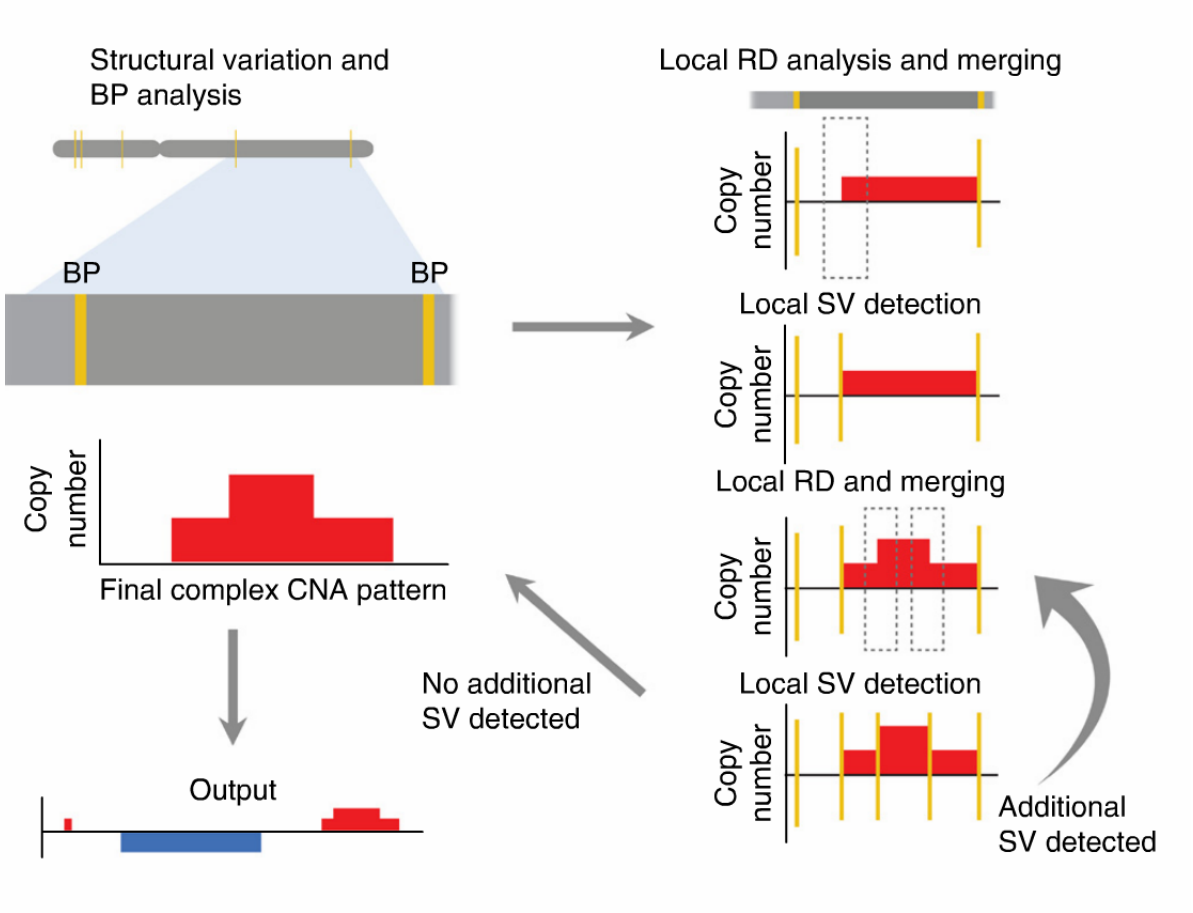


Combining Structural Variants & CNV

Sheffield Children's NHS
NHS Foundation Trust

CONSERTING: integrating copy-number analysis with structural-variation detection

Xiang Chen[1,2], Pankaj Gupta[1,2], Jianmin Wang[2–4], Joy Nakitandwe[2,5], Kathryn Roberts[5], James D Dalton[5], Matthew Parker[1,2], Samir Patel[5], Linda Holmfeldt[5], Debbie Payne[5], John Easton[2,6], Jing Ma[2,5], Michael Rusch[1,2], Gang Wu[1,2], Aman Patel[1,2], Suzanne J Baker[2,7], Michael A Dyer[2,7], Sheila Shurtleff[2,5], Stephen Espy[3], Stanley Pounds[8], James R Downing[2,5], David W Ellison[2,5], Charles G Mullighan[2,5] & Jinghui Zhang[1,2]



Telomere Length?

Sheffield Children's NHS
NHS Foundation Trust

Parker *et al. Genome Biology* 2012, **13**:R113
http://genomebiology.com/2012/13/12/R113

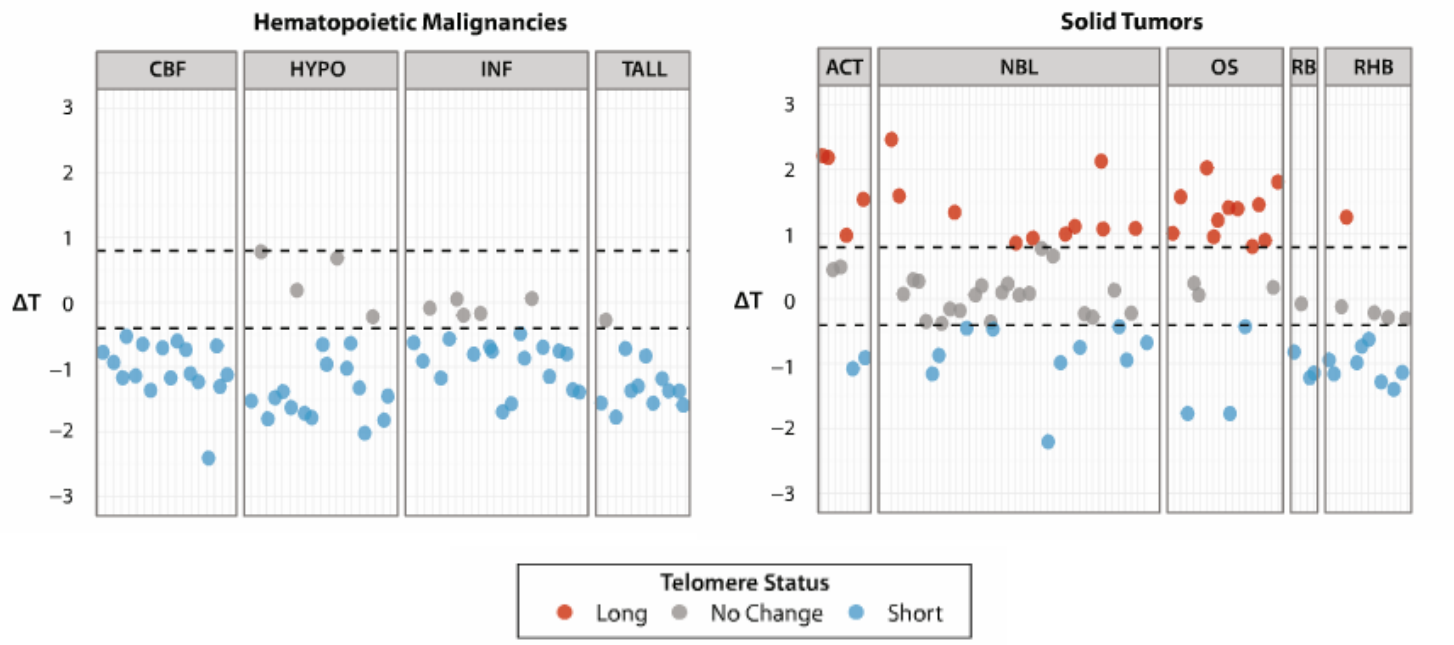
Genome Biology

RESEARCH Open Access

Assessing telomeric DNA content in pediatric cancers using whole-genome sequencing data

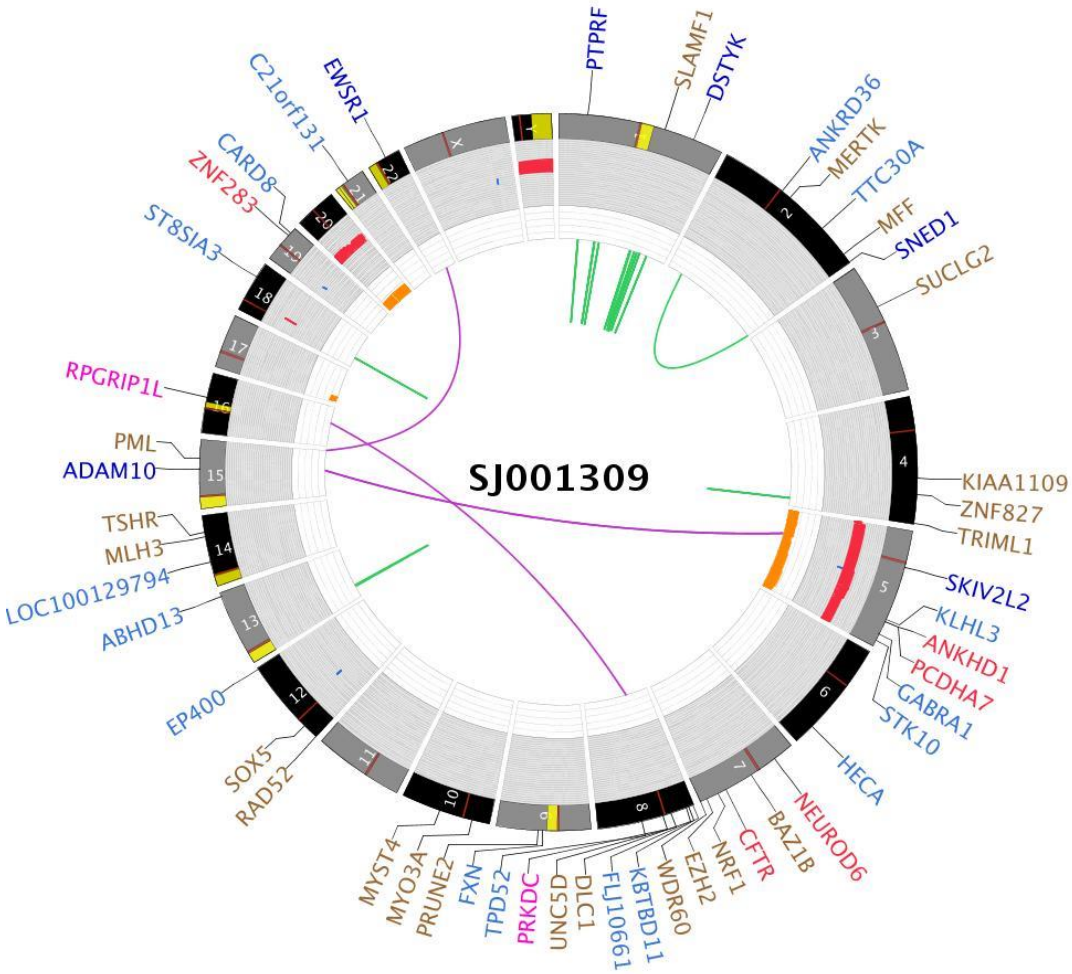
Matthew Parker[1], Xiang Chen[1], Armita Bahrami[2], James Dalton[2], Michael Rusch[1], Gang Wu[1], John Easton[3], Nai-Kong Cheung[4], Michael Dyer[5], Elaine R Mardis[6,7], Richard K Wilson[6,7], Charles Mullighan[2], Richard Gilbertson[5], Suzanne J Baker[5], Gerard Zambetti[8], David W Ellison[2], James R Downing[2] and Jinghui Zhang[1*], for the Pediatric Cancer Genome Project

- Deregulation of telomere maintenance is a hallmark of cancer.
- Most tumors find a way to maintain telomere length above a critical minimum
- Can be measured in WGS data



The Cancer Genome

Sheffield Children's NHS
NHS Foundation Trust



*silent mutations removed

Genomics england

GENOMONCOLOGY
BATTLING CANCER, ONE GENOME AT A TIME

Whole Genome Analysis from 100,000 Genomes Project Cancer Programme

Release notes for version 1.1 (composite WGA):

- WGAs are currently being returned from the Initiation Implementation Phase of the Cancer Main Programme on participants with solid tumours for whom paired fresh frozen and germline samples were submitted and passed quality assurance checks.

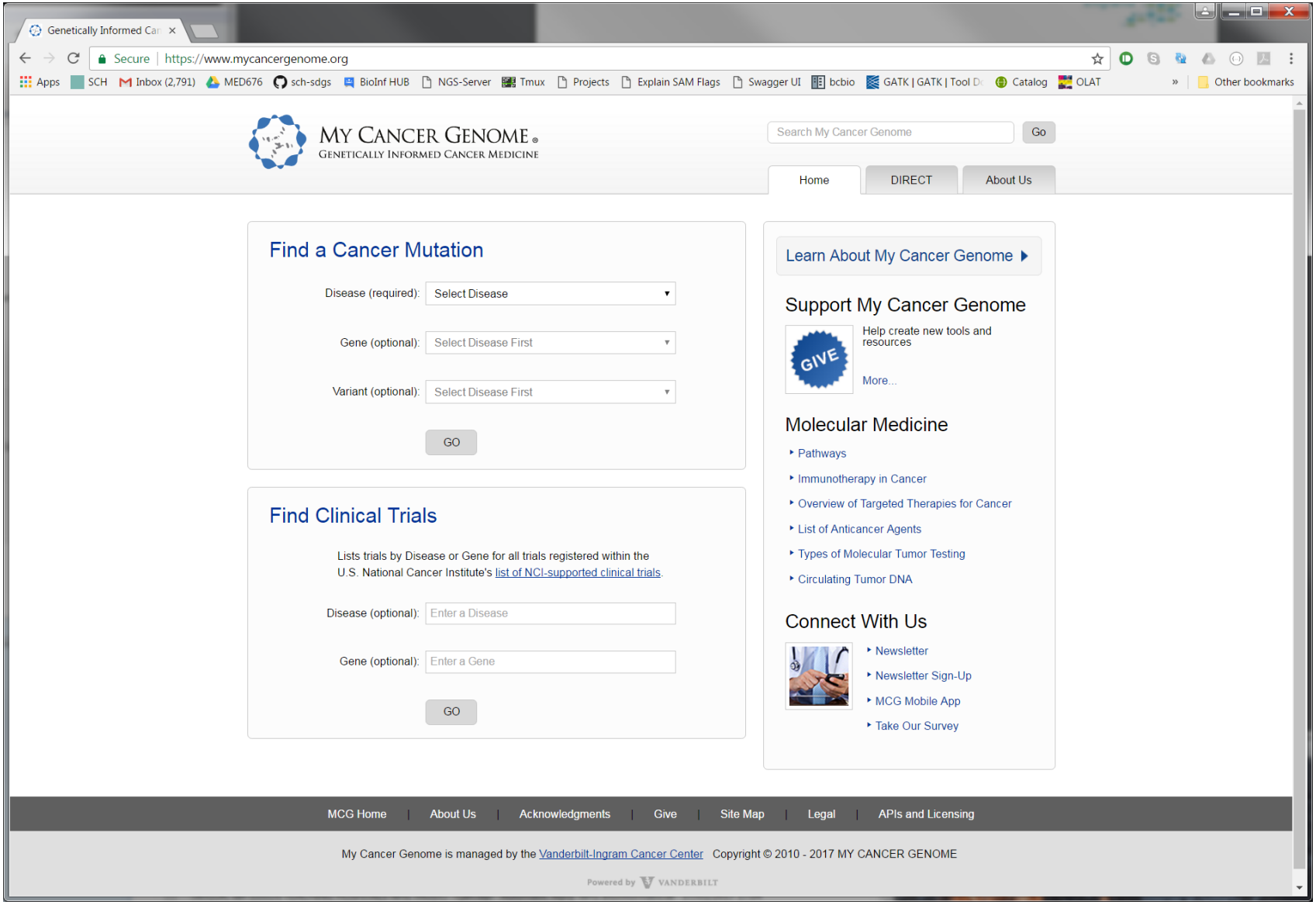
- It is envisaged that the initial review of these WGAs will be carried out within the GMC accredited Molecular Pathology/Genetics Laboratory therefore these analyses will be returned to GMC nominated individuals in these laboratories.

- WGAs are currently identified via two IDs; the Genomics England Participant ID and the Lab Sample ID for the fresh frozen tumour sample in line with NHS working practice. Provision of additional identifiers is under evaluation.

- The WGA is released as a paired set of formats:
    - a ***preliminary analysis*** containing small variants in Domain 1 (77 genes containing variants annotated as potentially actionable by Genome Oncology) and Domain 2 (590 genes listed in the Cancer Gene Census).
    - a ***supplementary analysis*** containing the content of the preliminary analysis **and** additional small variants (occurring outside Domains 1 & 2), copy number variants

Annotation of Cancer Variants

Sheffield Children's NHS
NHS Foundation Trust



Genomics England Cancer Report

Sheffield Children's NHS
NHS Foundation Trust

Whole Genome Analysis

100,000 Genomes Project Cancer Programme

Preliminary analysis of somatic small non-synonymous variants v1.1

Genomics england

Participant information

Participant name	D.O.B	Gender	NHS number	Laboratory sample ID	GeL participant ID	GMC	Sample date	Date analysis issued
XX								

Tumour information

Tumour type	Tumour subtype	ICD10 code	Sample type	Reported tumour content	Tumour sample cross-contamination
Colorectal	adenocarcinoma	N/A	FF	Medium 40-60%	PASS

Domain 1 variants

Variants in a virtual panel of potentially actionable genes*. Actionable genes are defined as genes in which small variants (SNVs and indels <50bp) have reported therapeutic, prognostic or clinical trial associations**, as defined by the GenomOncology Knowledge Management System. Where known, the "variant-level actionability" category and applicable tumour type are indicated. For other variants in these genes, their impact on gene function has not yet been characterised and therefore their actionability status is unclear. This means:
(i) local evaluation will be required for listed variants which are not yet characterised (i.e. "variant-level actionability" is denoted N/A)
(ii) even if well characterised as actionable for some tumour types, the listed variants may not be actionable in the participant's specific tumour type
*Current potentially actionable genes for solid tumours: 77 genes, listed at Actionable genes in solid tumour v1.1 document
**Links are provided to clinical trials within the United Kingdom which are both actively recruiting participants or closed to recruitment.

Right Tumour, Reported Actionability.

Genomics England Cancer Report

Sheffield Children's NHS
NHS Foundation Trust

Gene	Gene-level actionability	GRCh38 coordinates ref/alt allele	Transcript	cDNA and protein change	Predicted consequences	Population germline allele frequency (1KG)	VAF	Alt allele/ total read depth	COSMIC ID	Variant-level actionability	Gene mode of action
ALK	Therapeutic (NSC lung ca); Trial (NSC lung ca); Trial (NSC lung ca); Trial (NSC lung ca); Trial (NSC lung ca); Trial (NSC lung ca); Trial (solid neoplasm); Trial (solid neoplasm); Trial (solid neoplasm)	2:29220810 G>A	ENST00000389048	c.3541C>T p.(Arg1181Cys)	missense_variant	N/A	0.15	20/130	N/A	Trial (NSC lung ca); Trial (NSC lung ca); Trial (NSC lung ca); Trial (NSC lung ca); Trial (NSC lung ca); Trial (solid neoplasm); Trial (solid neoplasm); Trial (solid neoplasm)	oncogene
KRAS	Therapeutic (colorectal ca); Therapeutic (NSC lung ca); Trial (colorectal ca); Trial (colorectal ca); Trial (NSC lung ca); Trial (NSC lung ca); Trial (NSC lung ca); Trial (NSC lung ca); Trial (solid neoplasm); Trial (solid neoplasm); Trial (solid neoplasm); Trial (solid neoplasm); Trial (glioma); Trial (MPNST); Trial (melanoma); Trial (neuroblastoma); Trial (rhabdoid tu); Trial (rhabdomyosarcoma); Trial (schwannoma); Trial (sarcoma-ST)	12:25225628 C>T	ENST00000256078	c.436G>A p.(Ala146Thr)	missense_variant	N/A	0.3	35/118	COSM19404 COSM1165198	Therapeutic (colorectal ca); Therapeutic (NSC lung ca); Trial (colorectal ca); Trial (NSC lung ca); Trial (colorectal ca); Trial (glioma); Trial (MPNST); Trial (melanoma); Trial (neuroblastoma); Trial (NSC lung ca); Trial (rhabdoid tu); Trial (rhabdomyosarcoma); Trial (schwannoma); Trial (sarcoma-ST); Trial (solid neoplasm); Trial (solid neoplasm); Trial (solid neoplasm)	oncogene
TP53	Trial (ovarian ca)	17:7673796 C>A	ENST00000269305	c.824G>T p.(Cys275Phe)	missense_variant	N/A	0.49	40/82	COSM10701 COSM99932 COSM3723938 COSM1637959	Trial (ovarian ca)	oncogene, tumour suppressor

Domain 2 variants

Variants in a virtual panel of cancer-related genes***. Cancer-related genes are defined as genes in which any variants have been causally implicated in cancer, as defined by the Cancer Gene Census (Wellcome Trust Sanger Institute)
***Current cancer-related genes: 590 genes, listed at Cancer census genes v1.1 document

Genes causally implicated in cancer: Cancer Gene Census

http://cancer.sanger.ac.uk/census/

Genomics England Cancer Report

Sheffield Children's NHS
NHS Foundation Trust

Gene	GRCh38 coordinates ref/alt allele	Transcript	cDNA and protein change	Predicted consequences	Population germline allele frequency (1KG)	VAF	Alt allele/total read depth	COSMIC ID	Gene mode of action
APC	5:112839499 TGCAA>T	ENST00000508376	c.3906_3909delGCAA p.(Leu1302>fs)	frameshift_variant	N/A	0.13	14/110	N/A	tumour suppressor
CREBBP	16:3758051 T>TA	ENST00000262367	c.3370-4dupT	splice_region_variant	N/A	0.2	12/61	N/A	oncogene, tumour suppressor
FAT1	4:186620732 C>T	ENST00000441802	c.5854G>A p.(Val1952Ile)	missense_variant	N/A	0.28	28/99	COSM1054196 COSM1054194	tumour suppressor
FIP1L1	4:53453080 CAG>C	ENST00000337488	c.1459_1460delAG p.(Arg483>fs)	frameshift_variant	N/A	0.11	9/83	COSM249696 COSM4435275	N/A
IDH2	15:90088607 T>C	ENST00000330062	c.514A>G p.(Arg172Gly)	missense_variant	N/A	0.24	27/112	COSM33731	oncogene

Domain 3 variants

Small variants in genes not included in domains 1 & 2. These are not included in this document but are accessible via the Supplementary Analysis.

Sequencing quality information

See online Technical Information v1.1.main document and/or LabKey QC portal for details and expected ranges of QC metrics

Sample type	Mapped reads, %	Chimeric DNA fragments, %	Insert size median, bp	Genome-wide coverage mean, x	Unevenness of local genome coverage, x	COSMIC content with low coverage (<30x), %	Total somatic SNVs	Total somatic indels	Total somatic SVs
Germline	95.38	0.40	482.8	29.3	6.65	N/A	N/A	N/A	N/A
Tumour	95.68	0.37	447.4	84.6	13.45	1.15	31152	21184	262

Additional information

- **The pathways for sample processing and data analysis are not yet accredited end-to-end for diagnostic use. Accordingly, any result intended for use in informing clinical management should be confirmed using a test accredited for clinical use.**
- **Sensitivity: the depth of WGS used in this analysis will detect 99% of SNVs with an allele frequency of ≥0.3, 95% of SNVs with an allele frequency of ≥0.1 and 60% of indels with an allele frequency of ≥0.4 (estimate is based upon admixtures analysis of a highly accurate catalog of variants produced in the "platinum genomes" project). Consequently, variants with allelic frequencies below this level, or in areas of low coverage may not be detected. False negative results cannot be excluded.**
- Somatic calls are filtered according to the quality and quantity of reads. Full details of the filters used in this analysis can be found in the Technical Information v1.1.main.
- Variants present in the germline are subtracted to produce a list of somatic variants. Accordingly, variants detected in both the germline and the tumour will not be listed in this analysis.

Genomics England Cancer Report

Sheffield Children's NHS
NHS Foundation Trust

- In this analysis MNVs (multiple nucleotide variants) are reported as multiple consecutive SNVs and therefore the protein change may require correction.
- Only variants with specific consequences (transcript ablation, splice acceptor variant, splice donor variant, stop gained, frameshift variant, stop lost, initiator codon variant, transcript amplification, inframe insertion, inframe deletion, missense variant, splice region variant, incomplete terminal codon variant) in canonical transcripts are reported. The complete list of canonical transcripts can be accessed at List of canonical transcripts v1.1.
- A variant may have multiple entries in COSMIC database due to the use of different reference sequences. In these cases links to all COSMIC entries are provided.
- Structural variants (SVs) and copy number variants (CNVs) are not included in this analysis. These variants types are included in the Supplementary Analysis.
- For a full description of the methods used to produce these results and for further information regarding QC metrics please refer to the Technical Information v1.1.main. All related documentation is available at Genomics England Website.
- 'N/A' indicates that information is not available or not applicable.

Genomics England
Queen Mary University of London
Dawson Hall
Charterhouse Square
London
EC1M 6BQ

Sequencing Laboratory
Illumina Laboratory Services United Kingdom - Hinxton
The Ogilvie Building, Wellcome Trust Genome Campus
Hinxton Nr Saffron Walden
Essex
CB10 1DR

Sheffield Children's NHS
NHS Foundation Trust

GEL Supplementary Cancer Report

Take Home Messages

Sheffield Children's NHS
NHS Foundation Trust

- Utility of cancer genomes in the clinic not yet realised
- Need good samples with high tumour content
- Coverage is King!
- Variant calling is challenging
  - Heterogeneity
  - Tumour purity
  - Contamination
- Knowledgebases for prioritising variants increasingly important – "Actionability"