AN INTRODUCTION TO CANCER GENOMICS: TOOLS AND WORKFLOWS

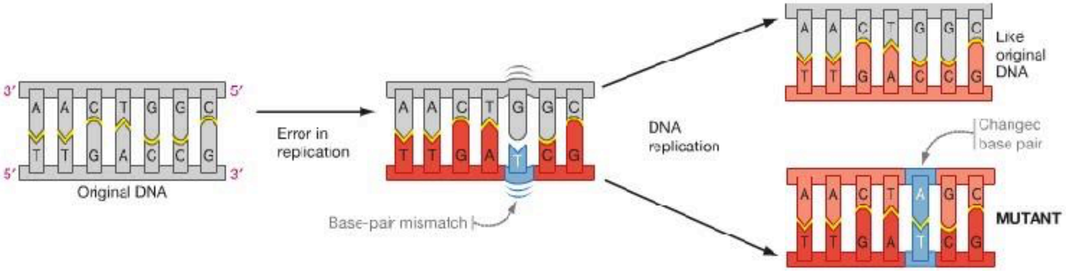
Anders Skanderup and Amanda Yu Guo

21 Sept 2016

Cancer is a genetic disease

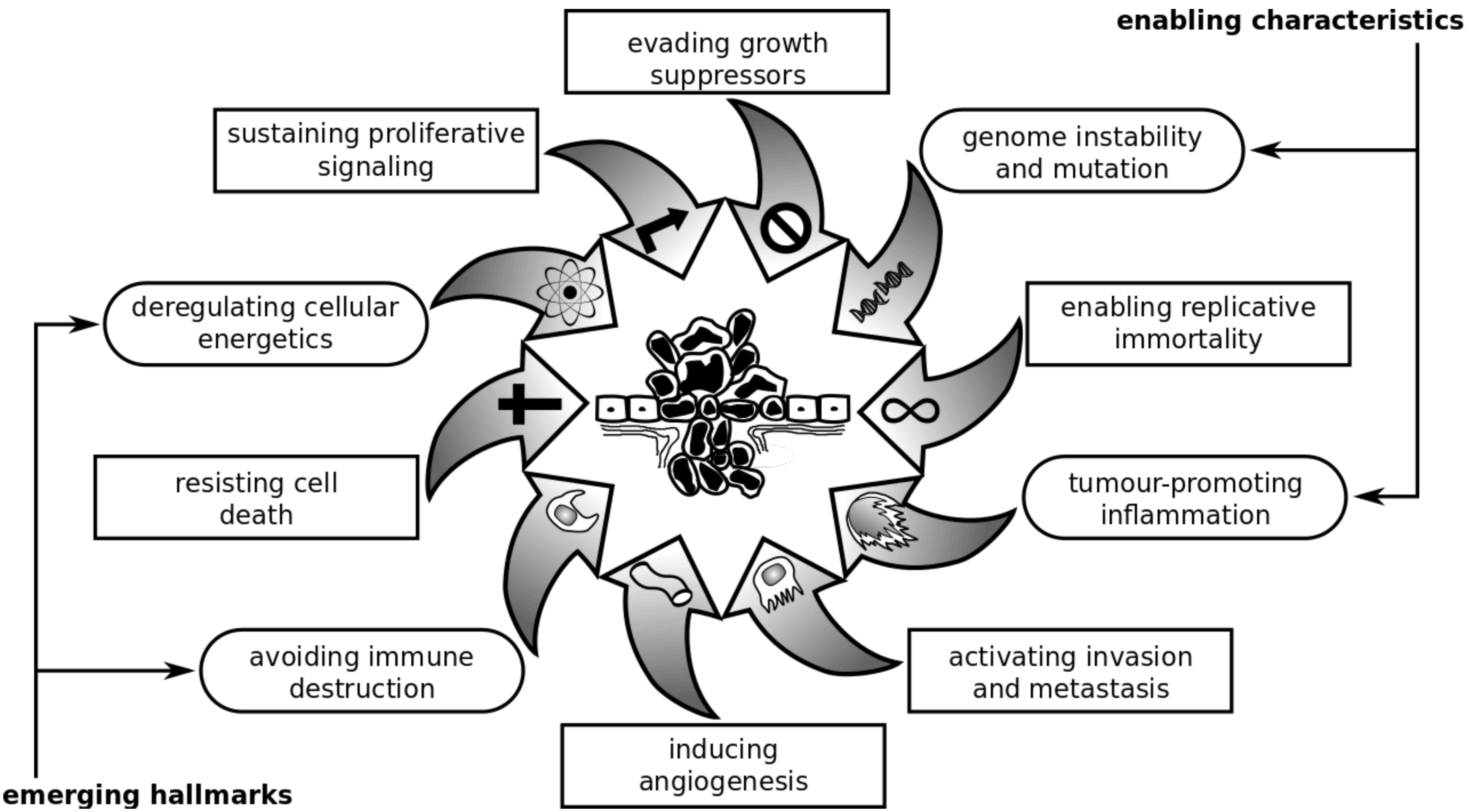
- Characterized by abnormal cell growth
- Caused by inherited or acquired genetic lesions (mutations), allowing affected cells to outcompete/ invade other cells/niches
- Cause and consequence varies by organ and cell type

Mutations



- Mutations may be
  - **germline** : inherited, all cells in an individual share these
  - **somatic** : spontaneous, exists *only* in a subset of (cancer) cells of given individual

Hallmarks of Cancer



evading growth suppressors

enabling characteristics

sustaining proliferative signaling

genome instability and mutation

deregulating cellular energetics

enabling replicative immortality

resisting cell death

tumour-promoting inflammation

avoiding immune destruction

activating invasion and metastasis

inducing angiogenesis

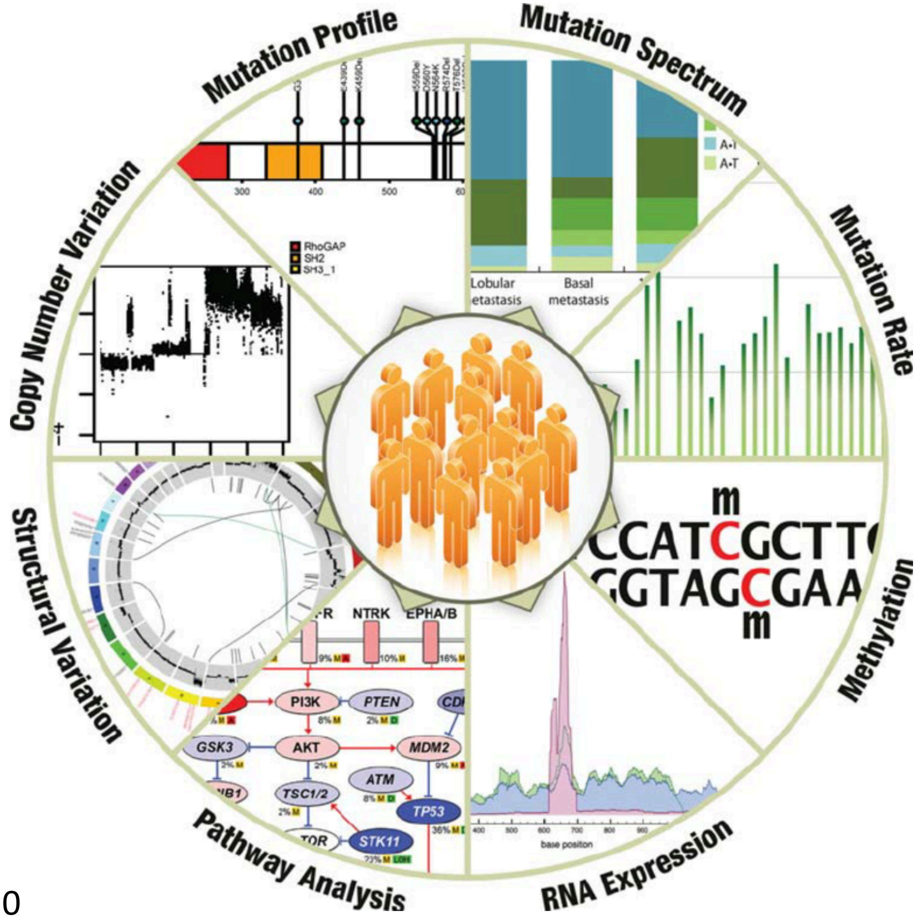
emerging hallmarks

Cancer is an umbrella of diseases: The problem of heterogeneity

- Number of cancer hallmarks acquired vary extensively between tumors
- The exact same capability may be acquired by mutating any of k genes
- Tumors of the same cancer type are genetically extremely heterogeneous
- Most individual cancer driver events occur in less than 5% of tumors and some likely at <1% frequency

Cancer Genomics

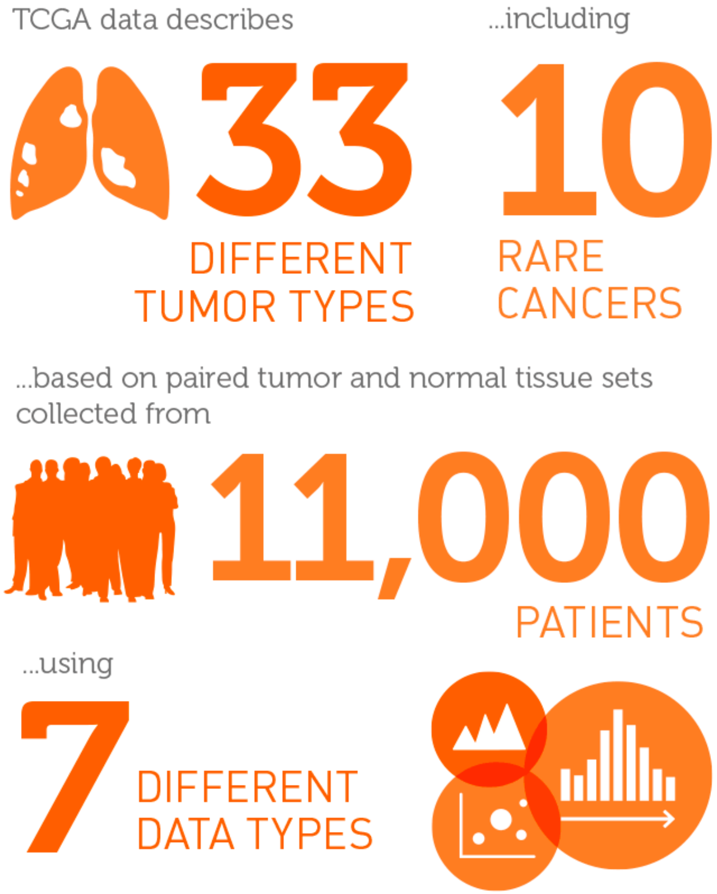
Cancer genomics is the study of the totality of DNA sequence and gene expression differences between tumour and normal cells



Ding et al, *Hum Mol Genet*, 2010

The Cancer Genome Atlas (TCGA)

Aims to catalogue and discover major cancer-causing genomic alterations to create a comprehensive ’atlas’ of cancer genomic profiles



TCGA data describes 33 DIFFERENT TUMOR TYPES

...including 10 RARE CANCERS

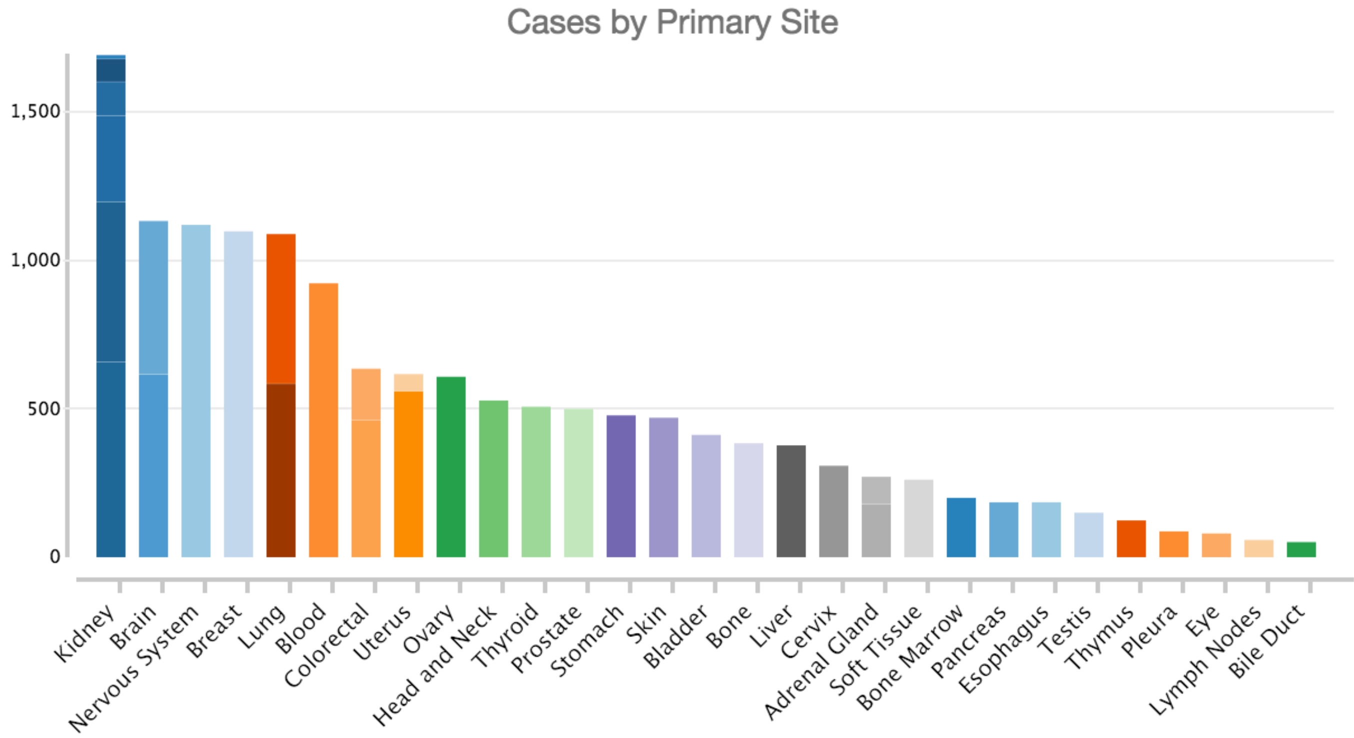
...based on paired tumor and normal tissue sets collected from 11,000 PATIENTS

...using 7 DIFFERENT DATA TYPES

- Clinical
- DNA sequencing
- RNA sequencing
- SNP arrays
- DNA methylation
- Protein array

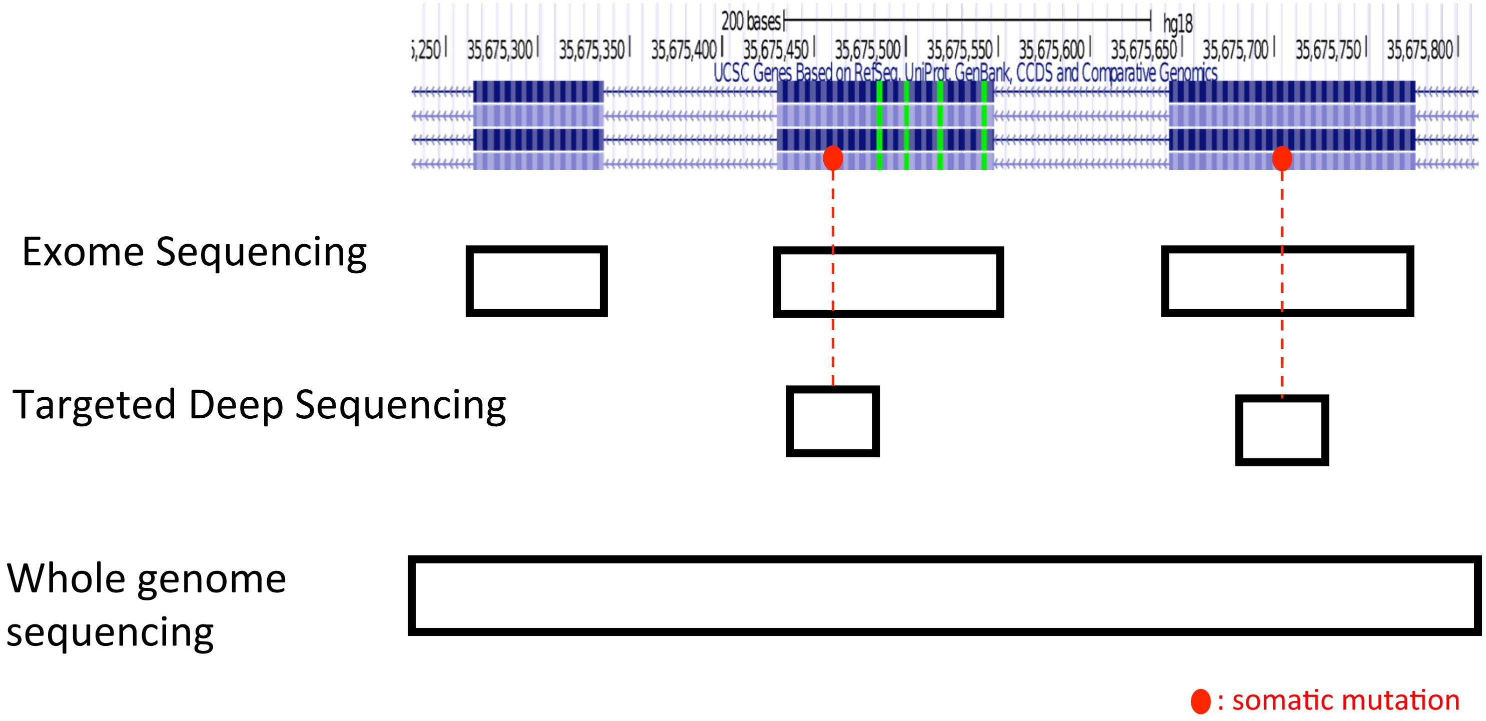
Genomic Data Commons Portal

Interactive data system for researchers to search, download, and analyze cancer genomic data sets



https://gdc-portal.nci.nih.gov/

Exome vs. whole genome sequencing



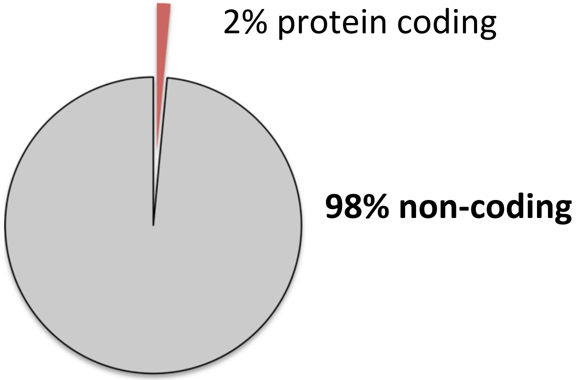
Exome Sequencing

Targeted Deep Sequencing

Whole genome sequencing

● : somatic mutation

Why whole genomes?



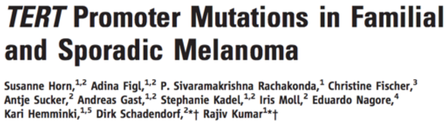
What are the advantages of WGS compared to WES

- Structural variation (!)
- Non-coding mutations (!)
- Base-point resolution DNA copy number profiles
- More data points (=better fit) for heterogeneity and mutation signature analysis

What are the disadvantages:

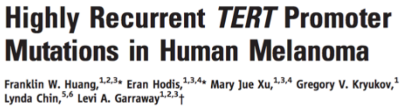
- Higher price (**6x**, at same depth, but only **4-5x** if you add a SNP-array)
- Data analysis (uses >10x resources)

Recent WGS studies



***TERT* Promoter Mutations in Familial and Sporadic Melanoma**

Susanne Horn,[1,2] Adina Figl,[1,2] P. Sivaramakrishna Rachakonda,[1] Christine Fischer,[3] Antje Sucker,[2] Andreas Gast,[1,2] Stephanie Kadel,[1,2] Iris Moll,[2] Eduardo Nagore,[4] Kari Hemminki,[1,5] Dirk Schadendorf,[2]*† Rajiv Kumar[1]*†



Highly Recurrent *TERT* Promoter Mutations in Human Melanoma

Franklin W. Huang,[1,2,3]* Eran Hodis,[1,3,4]* Mary Jue Xu,[1,3,4] Gregory V. Kryukov,[1] Lynda Chin,[5,6] Levi A. Garraway[1,2,3]†

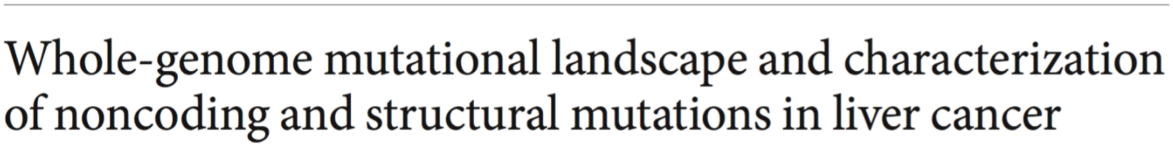
Science

AAAS

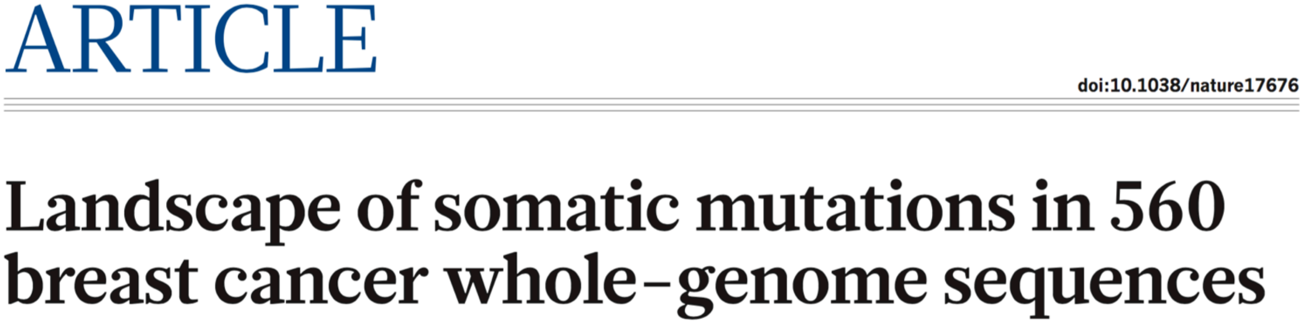
January 2013

nature genetics

(N=300)



Whole-genome mutational landscape and characterization of noncoding and structural mutations in liver cancer



ARTICLE

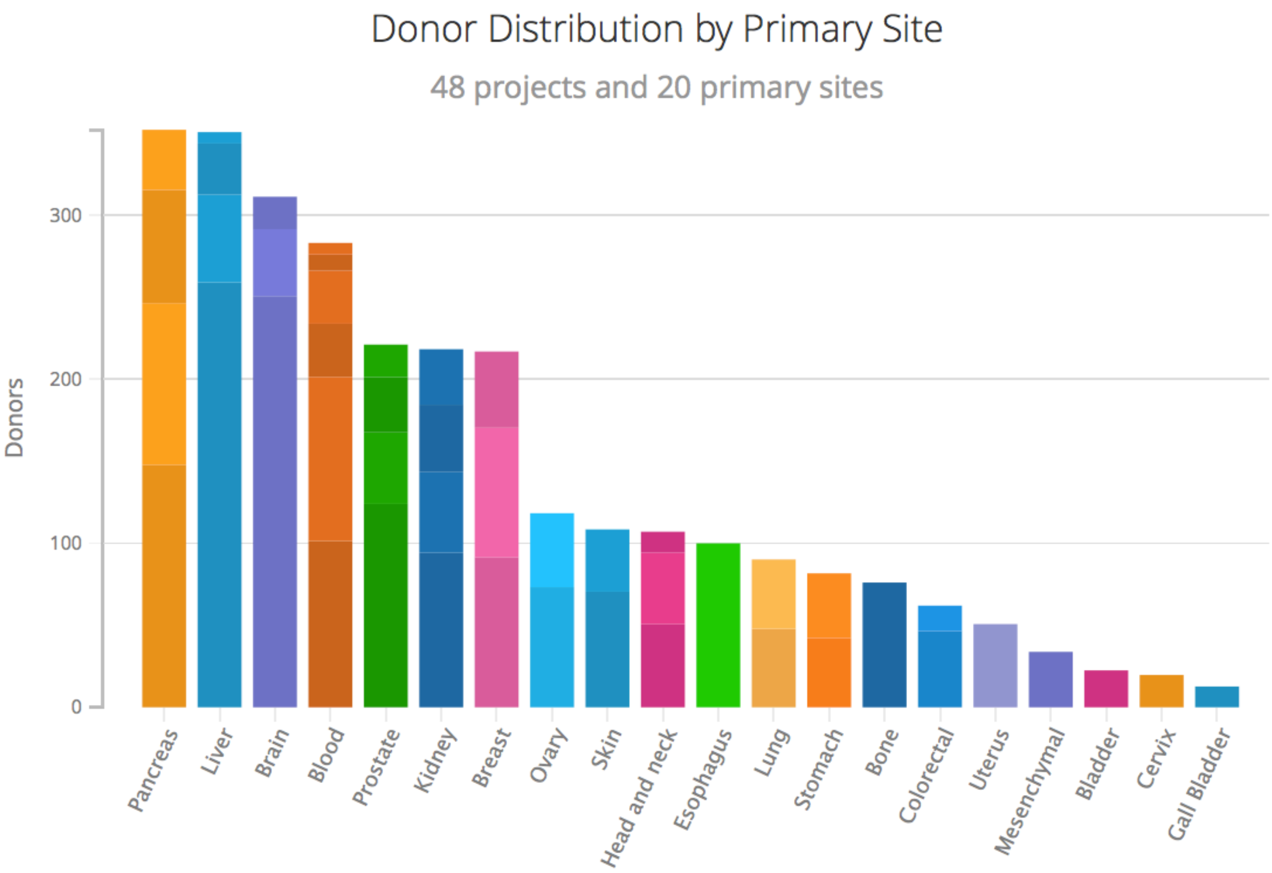
doi:10.1038/nature17676

Landscape of somatic mutations in 560 breast cancer whole-genome sequences

Pan-Cancer analysis of Whole Genomes (PCAWG)

- Co-coordinated by the International Cancer Genome Consortium (ICGC) and The Cancer Genome Atlas (TCGA)
- Analyzing more than 2,800 whole cancer genome
- Aims to explore somatic and germline variations in both coding and non-coding regions, with specific emphasis on cis-regulatory sites, non-coding RNAs, and large-scale structural alterations

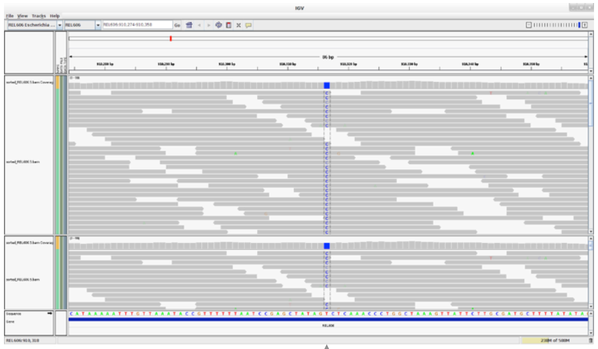
Resources: PCAWG



https://dcc.icgc.org

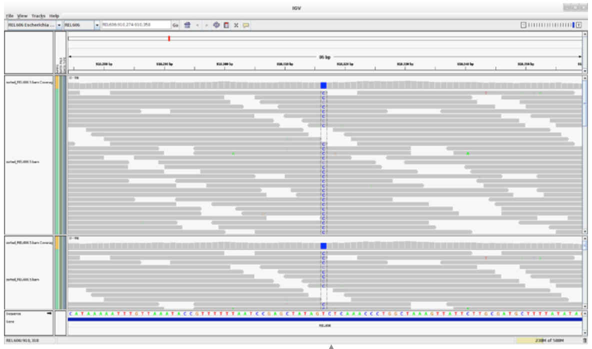
Compare genotypes in normal and tumor DNA

BAM file for sequenced normal tissue



Pileup

BAM file for sequenced tumor tissue



Pileup

What is the possible genotype
in the normal DNA?

A
A
A
A
C
A
A
A
-
C
C
C
G
C
A

Compute likelihood of a genotype difference between normal and tumor.

A
A
G
G
G
A
A
A
-
G
G
C
G
C
A

Somatic mutation calling – sources of error

- Cancer tissue is heterogeneous. Cell populations vary within tumor samples.
- Low-frequency mutations are hard to distinguish from sequencing errors.
- Sequencing bias: certain sequences are read with greater frequencies than others.
  - Amplification step in NGS

Mutation calling problems and heuristic filters

Problem: sequencing errors

Heuristic: only call mismatches represented by a threshold number of reads

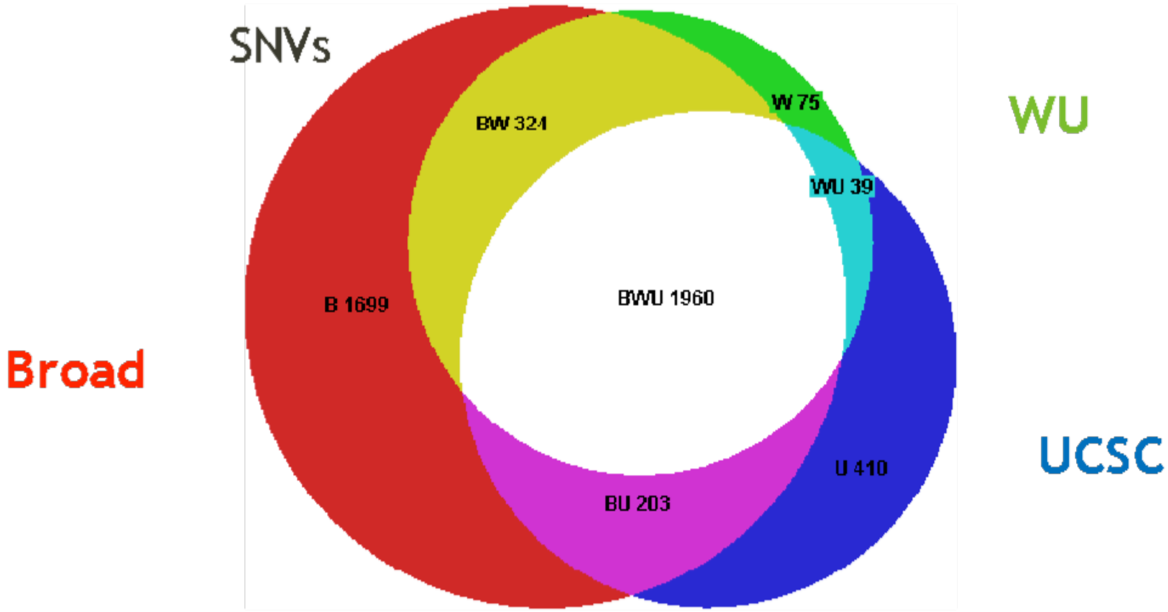
Problem: noisy, low-confidence reads

Heuristic: consult per-base read quality scores, and apply a quality score threshold.

Problem: low-confidence mapping, especially near indels

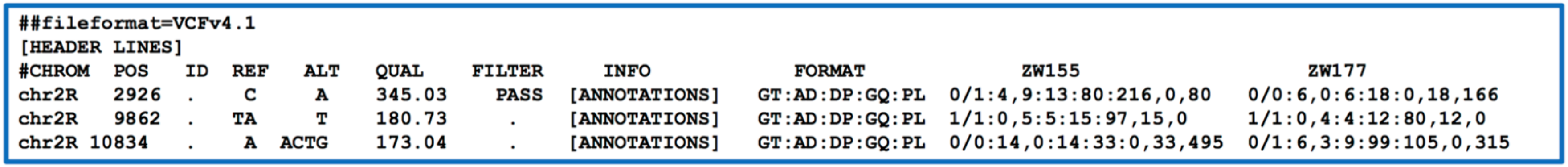
Heuristic: do not call mismatches near indels

Mutation calling is no solved problem. Different methods yield differing results



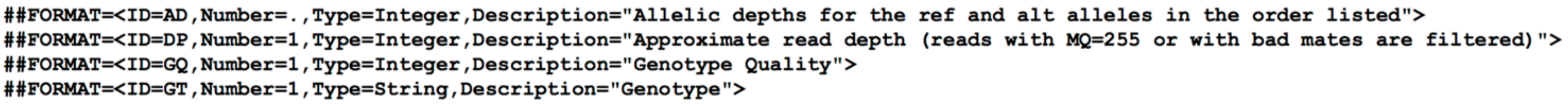
Overlap of mutation calls done for the same cancer samples at three different analysis centers

Standard format for variant calling: VCF files



```
##fileformat=VCFv4.1
[HEADER LINES]
#CHROM  POS    ID  REF   ALT   QUAL    FILTER     INFO           FORMAT            ZW155                   ZW177
chr2R   2926   .    C     A    345.03   PASS   [ANNOTATIONS]  GT:AD:DP:GQ:PL  0/1:4,9:13:80:216,0,80   0/0:6,0:6:18:0,18,166
chr2R   9862   .   TA     T    180.73    .     [ANNOTATIONS]  GT:AD:DP:GQ:PL  1/1:0,5:5:15:97,15,0     1/1:0,4:4:12:80,12,0
chr2R 10834    .    A  ACTG    173.04    .     [ANNOTATIONS]  GT:AD:DP:GQ:PL  0/0:14,0:14:33:0,33,495  0/1:6,3:9:99:105,0,315
```

[HEADER LINES]: start with “##”, describe all symbols found later on, e.g.,



```
##FORMAT=<ID=AD,Number=.,Type=Integer,Description="Allelic depths for the ref and alt alleles in the order listed">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Approximate read depth (reads with MQ=255 or with bad mates are filtered)">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
```

ID: some ID for the variant, if known (e.g., dbSNP)

REF, ALT: reference and alternative alleles (on forward strand of reference)

QUAL = -10*log(1-p), where p is the probability of variant being present given the read data

FILTER: whether the variant failed a filter (filters defined by the user or program processing the file)

Calling the somatic mutations in a tumor is only the first step

- Which genes are significantly mutated?
- Which mutations caused the cancer?
- Which mutations caused the cancer to progress?
- Which alterations are "actionable"?

Identification of driver mutations

- **Driver mutations**
  - Give a selective growth advantage to a cancer cell
  - Often occur in most cells in a tumor ("founders")
- **Passenger mutations**
  - Confer no selective growth advantage
  - May be present in founder cells or not

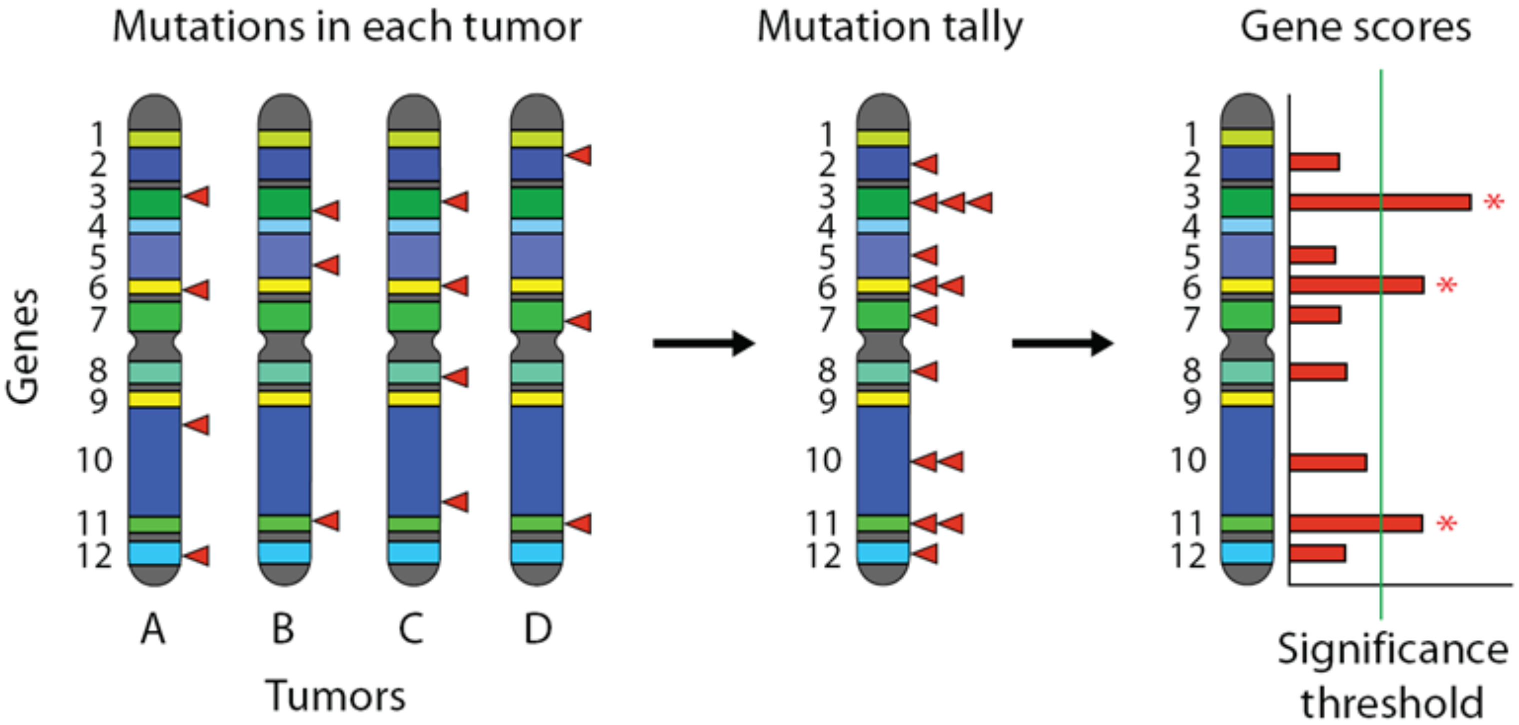
Problem, one tumor may have:

> 10 000 somatic mutations

> 100 mutations in protein-coding regions

The vast majority of these are ***passenger*** mutations

Identifying positive selection in tumors



- Model a null background mutation rate (BGM)
- Find genes/regions with more mutation than expected under the null

Parameters used to estimate the background mutation rate (BMR)

- Overall mutation frequency

- Relative frequencies of different categories of mutations
  - Transition vs transversion
  - CpG dinucleotides
  - Rest of C:G
  - A:T
  - small insertions and deletions

Kan et al., Nature, 2010
Seshagiri, Nature, 2012

Modeling background mutation rate in coding regions

$$f_i = \frac{s_i r_i}{n_i}$$

- f_i= background mutation rate for nucleotide category i
- n_i = # protein-coding nucleotide of category i
- s_i = # synonymous mutations of nucleotide category i
- r_i=NS/S ratio in nucleotide category i

Kan et al., Nature, 2010
Seshagiri, Nature, 2012

Problems with using an uniform BMR

- mutation rates vary across genomic loci
- mutation rates vary across samples
- longer lists of significant genes with more samples
- many false positive findings
  - olfactory receptors
  - list enriched for long genes: titin and mucin

Factors contributing to mutational heterogeneity in cancer genomes

- Cancer type
- Individual tumors

- Nucleotide context
- Replication timing
- Gene expression
- Chromatin organization
  - cell type specific epigenetic landscape

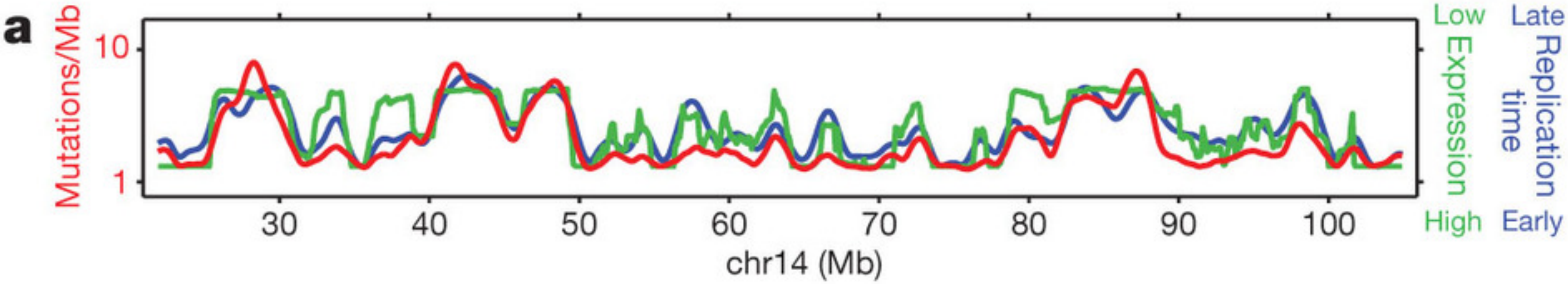
Mutation recurrence analysis in coding regions

MutSigCV –Mutsig with covariates

Builds a BMR model by pooling data from 'neighbor' genes in covariate space

Genomic covariates:

1. Gene expression level
2. Replication time during the cell cycle



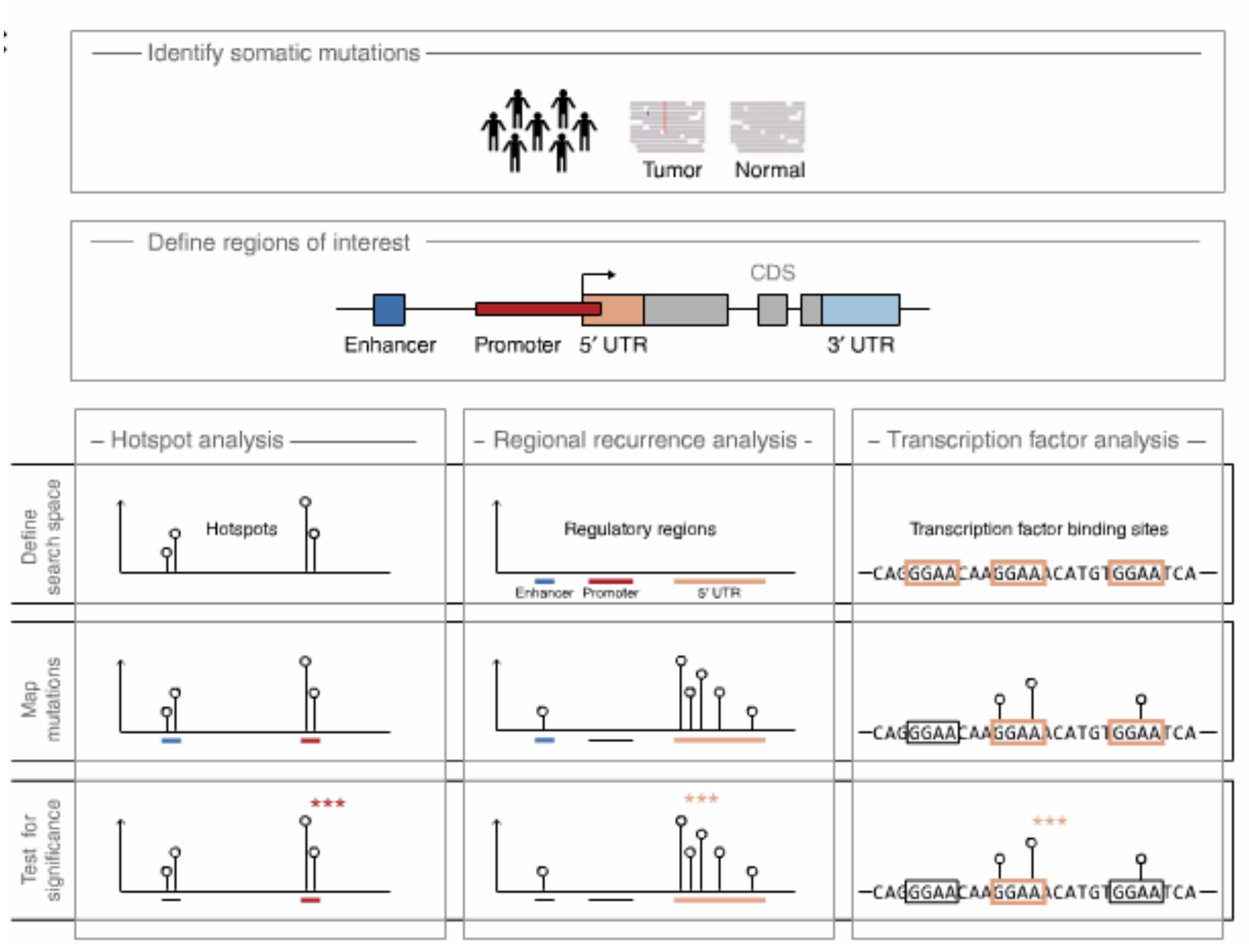
Lawrence at al., *Nature* 2013

MutSigCV: gene specific background rate

- directly estimate local BMR from:
  1. synonymous mutations
  2. non-coding mutations in the UTRs and introns

- bin genes according to gene expression levels and DNA replication time
  - find a set of nearest neighbors
  - pool data across the set of genes to estimate BMR

Mutation recurrence in non-coding regions


Identify somatic mutations
Tumor
Normal
Define regions of interest
CDS
Enhancer
Promoter
5′ UTR
3′ UTR
Hotspot analysis
Regional recurrence analysis
Transcription factor analysis
Define search space
Map mutations
Test for significance
Hotspots
Regulatory regions
Enhancer
Promoter
5′ UTR
Transcription factor binding sites
CAGGGAACAAGGAAACATGTGGAATCA
***

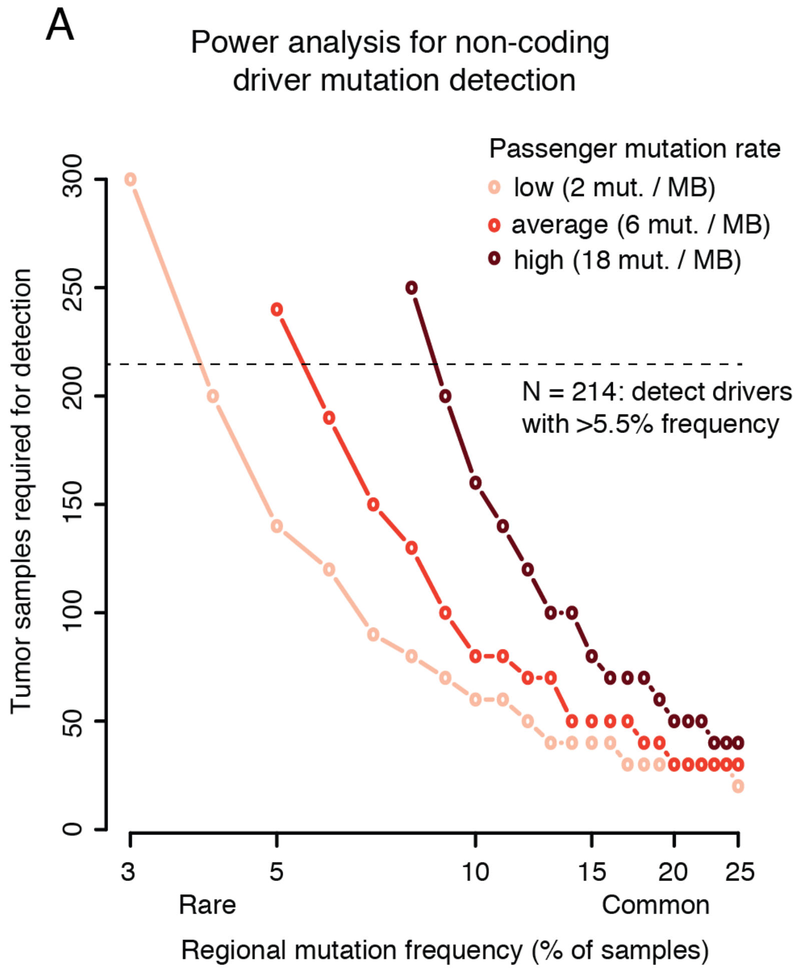
Weinhold et al., *Nature Genetics*, 2014

Mutation recurrence in non-coding regions

Some covariates to consider when modeling BMR

- patient ID
- replication timing bin
- nucleotide context
- transcription factor binding sites (ENCODE)
- histone modification profiles (Roadmap epigenomics)
- local mutation rate
- interactions among covariates

Power analysis



Factors affecting power of detection:

- passenger mutation rate
- mutation frequency among tumors

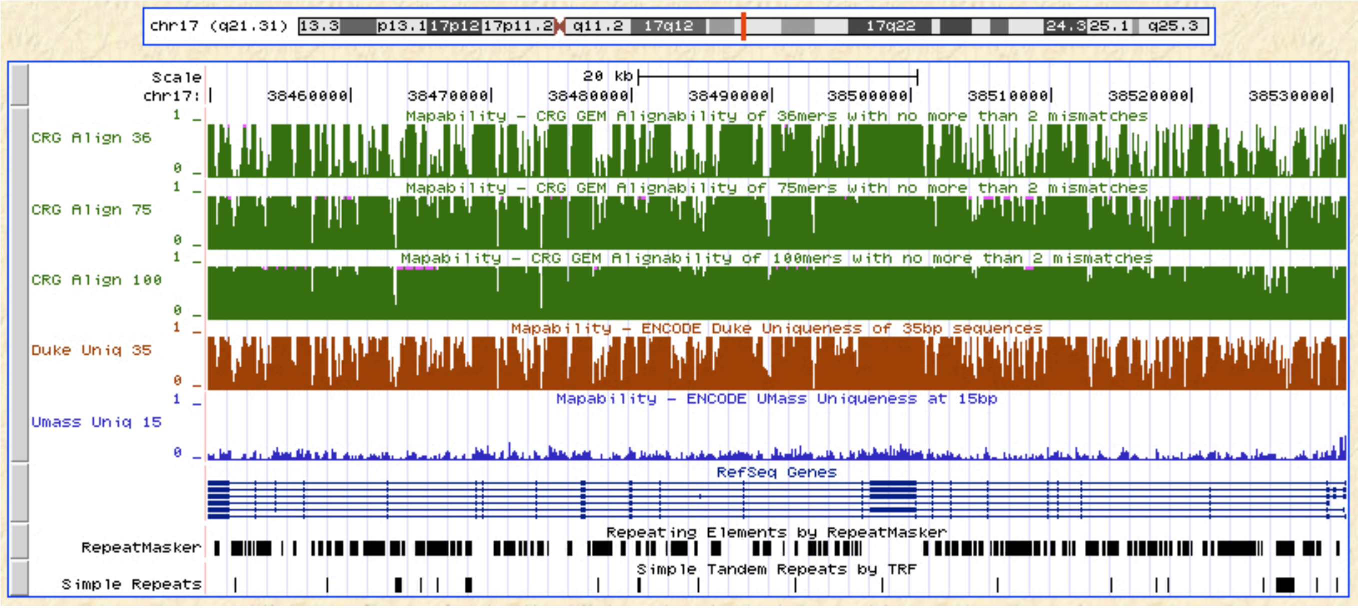
Common artifacts

genomic regions that tend to generate mapping errors

Reference TCGATCGATCGATCGATCGATCGA … TCGATCGAACGATCGATCGATCGA

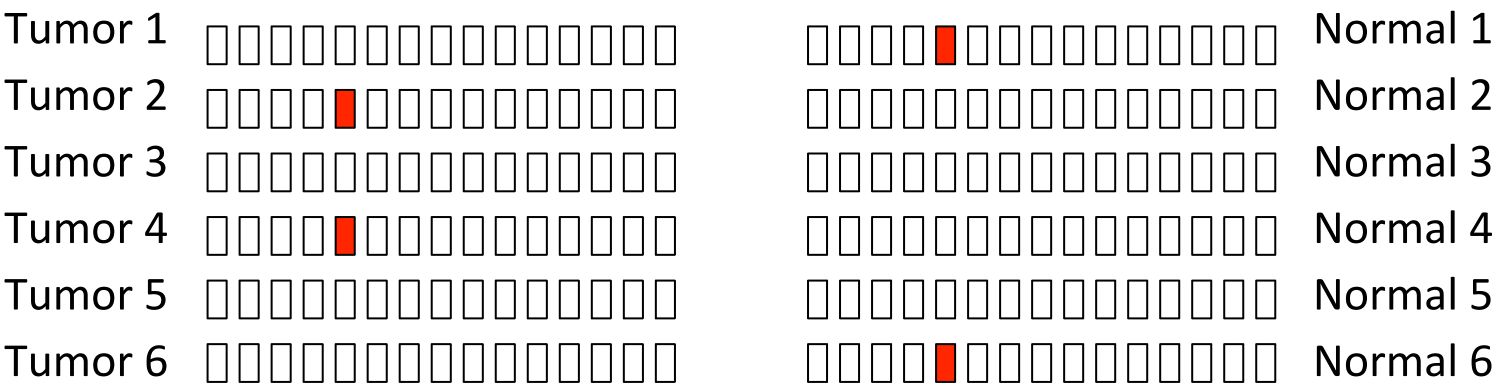
TCGATCGATCGATCGATC … TCGATCGATCGATCGAT

mask regions with low alignability/mappability


chr17 (q21.31) 13.3 p13.1 17p12 17p11.2 q11.2 17q12 17q22 24.3 25.1 q25.3
Scale
chr17:
20 kb
38460000
38470000
38480000
38490000
38500000
38510000
38520000
38530000
1
0
CRG Align 36
Mapability - CRG GEM Alignability of 36mers with no more than 2 mismatches
CRG Align 75
Mapability - CRG GEM Alignability of 75mers with no more than 2 mismatches
CRG Align 100
Mapability - CRG GEM Alignability of 100mers with no more than 2 mismatches
Duke Uniq 35
Mapability - ENCODE Duke Uniqueness of 35bp sequences
Umass Uniq 15
Mapability - ENCODE UMass Uniqueness at 15bp
RefSeq Genes
RepeatMasker
Repeating Elements by RepeatMasker
Simple Repeats
Simple Tandem Repeats by TRF

Common artifacts

systematic sequencing errors



flag/filter mutations that also appear in the panel of normal samples

Common artifacts

- Germline mutation wrongly called as somatic
  - Filter common SNPs in the general population

- Misalignment caused by germline insertions/deletions
  - Filter mutations close to common germline indels

Thank you!

contact: guoy1@gis-astar.edu.sg