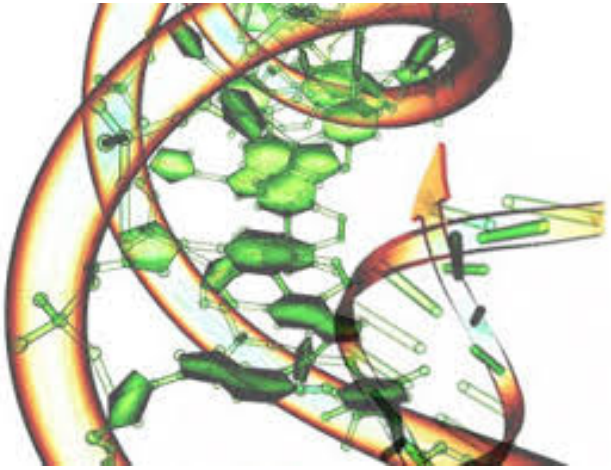
Detection of somatic mutations in cancer tumors



Malin Larsson

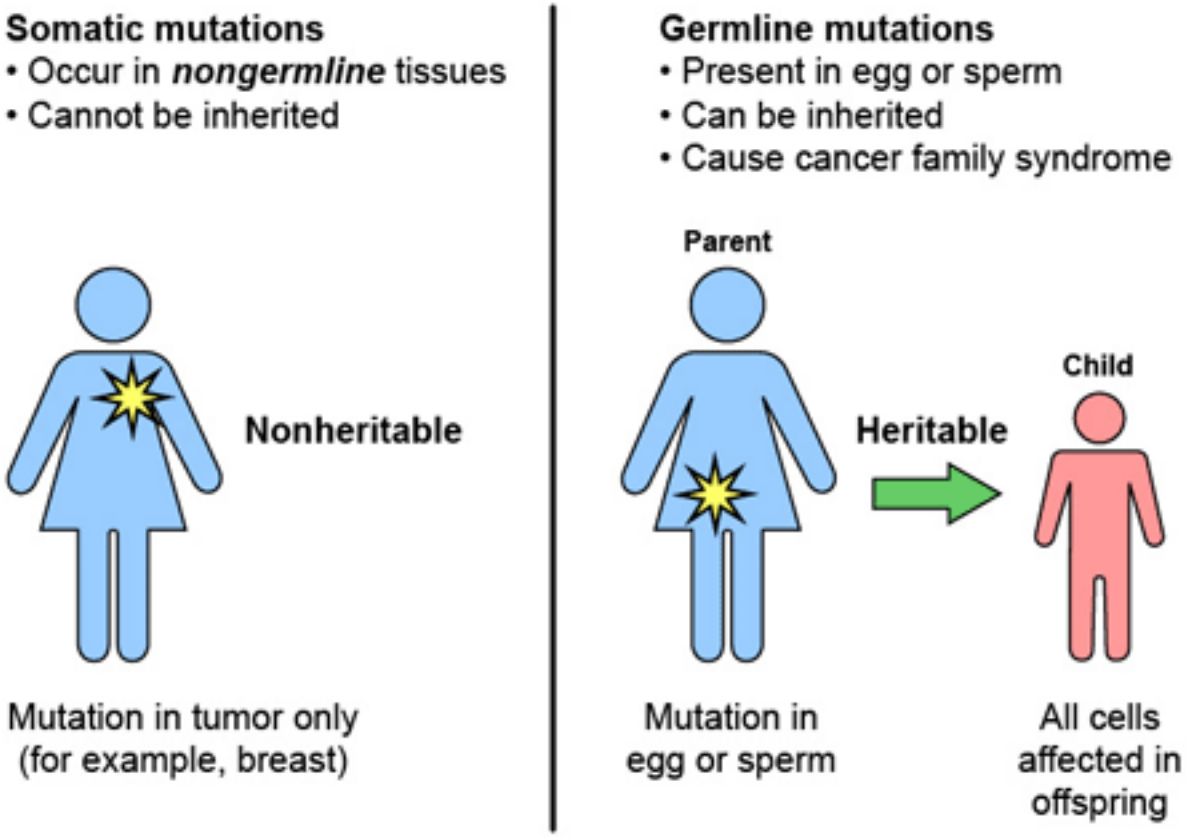
malin.larsson@scilifelab.se

Outline

- Introduction
- The mutational landscape of cancer
- Detection of cancer mutations
- recap of germline variant calling
- Somatic variant calling workflow
- Today's practical

Introduction

Somatic vs germline mutations

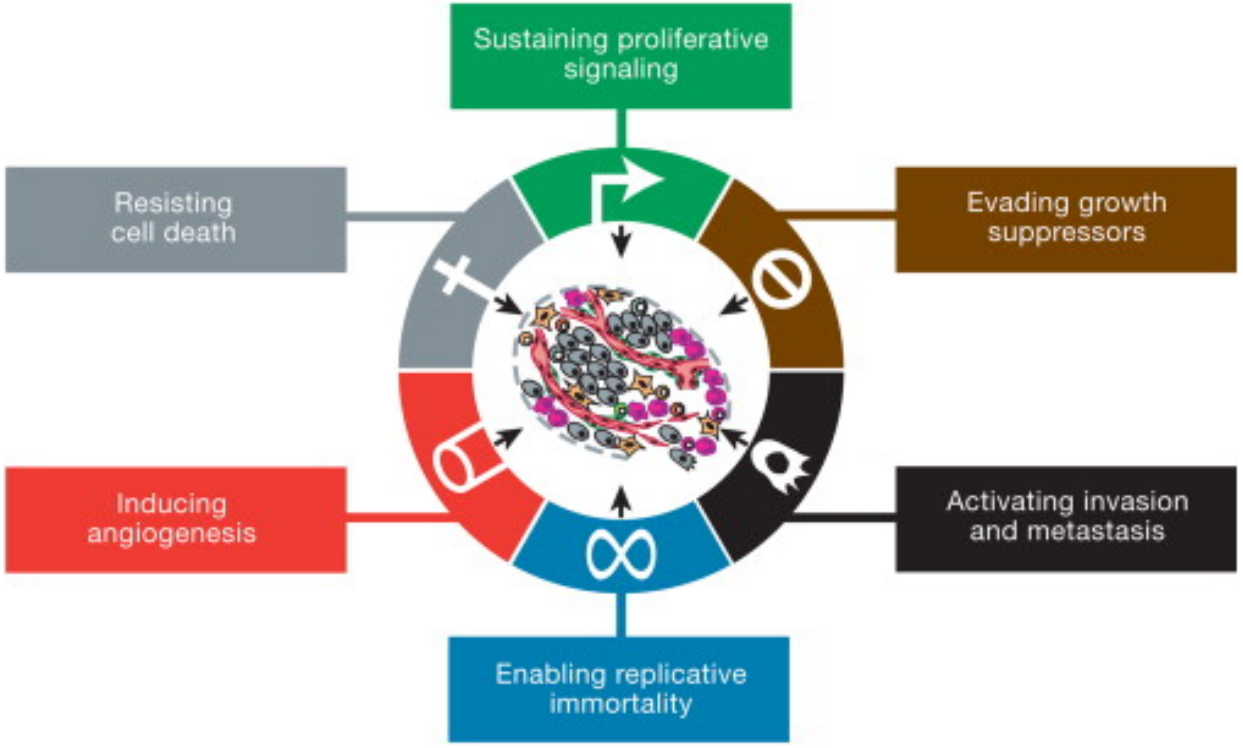


https://www.stjude.org/Images/hosp-misc-ipactr6-1101.gif

Cancer is an evolutionary process

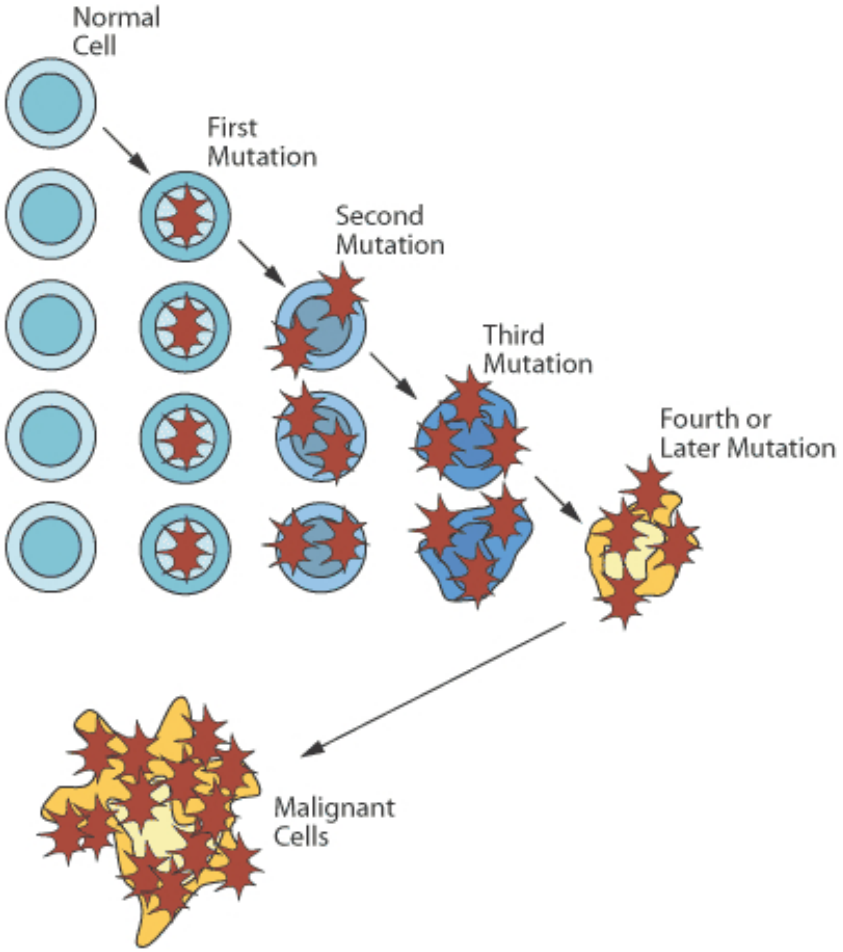
- Genetic variation introduced in individual cells
- more-or-less random mutations
- Clonal expansion - natural selection acting on the resultant phenotypic diversity

Hallmarks of Cancer



The Hallmarks of CancerThis illustration encompasses the six hallmark capabilities originally proposed by Hanahan et al 2000. The past decade has witnessed remarkable progress toward understanding the mechanistic underpinnings of each hallmark.
Hanahan and Weinberg, Hallmarks of Cancer: The Next Generation, Cell 2011

Development of cancer



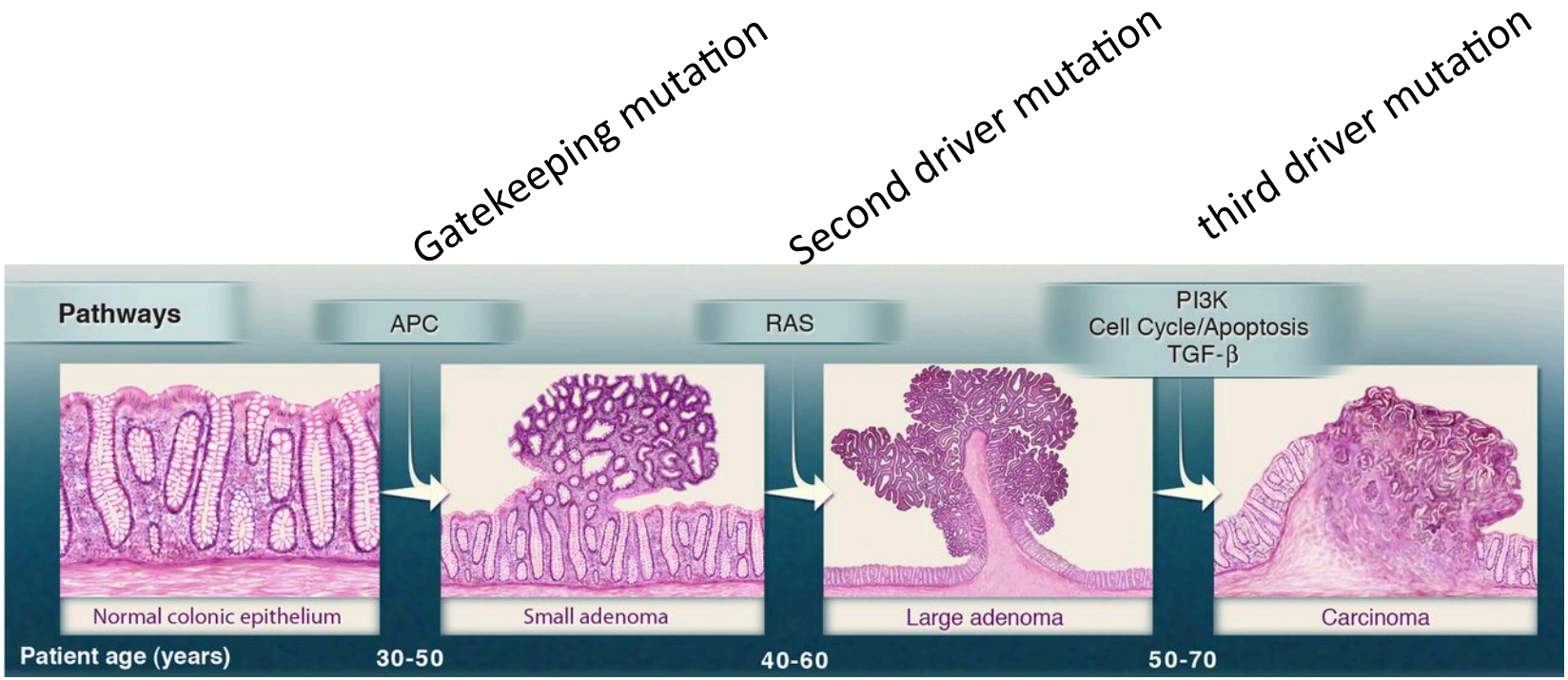
https://stanfordhealthcare.org/medical-clinics/cancer-genetics-program/understanding-genetics/how-genes-cause-cancer.html

Driver and passenger mutations

Driver' mutations confer a growth advantage of the cell. They are positively selected during the evolution of the cancer

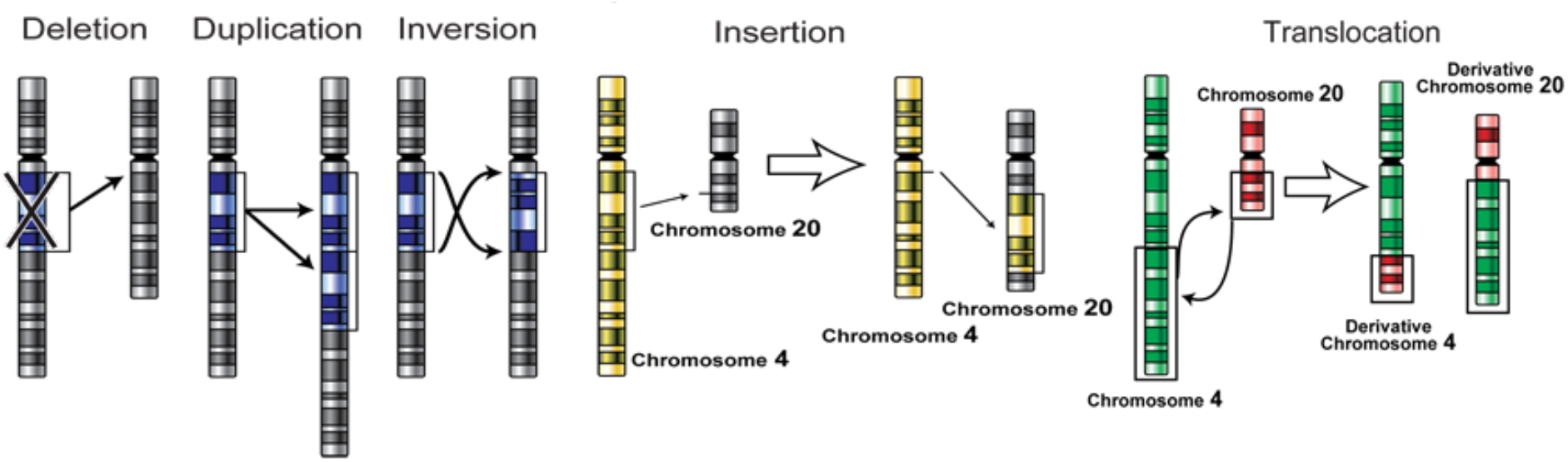
Passenger mutations are neutral, they just happened to be present in an ancestor of the cancer cell

Genetic alterations and the progression of colorectal cancer.



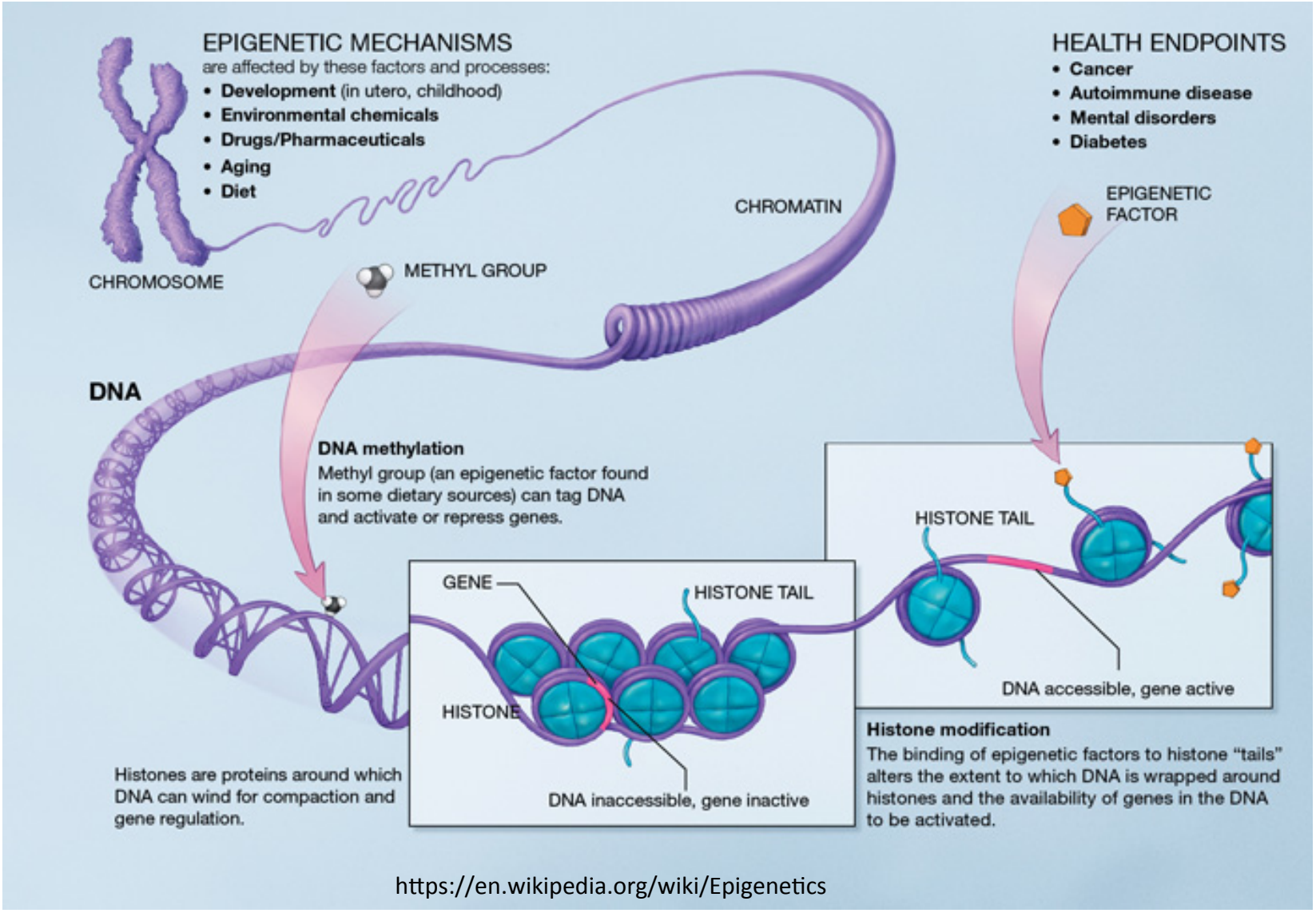
Bert Vogelstein et al. Science 2013;339:1546-1558

Types of mutations



http://socratic.org/questions/how-do-dna-mutations-occur

Epigenetic changes



https://en.wikipedia.org/wiki/Epigenetics

Mutational Landscape of Cancer

Some statistics…

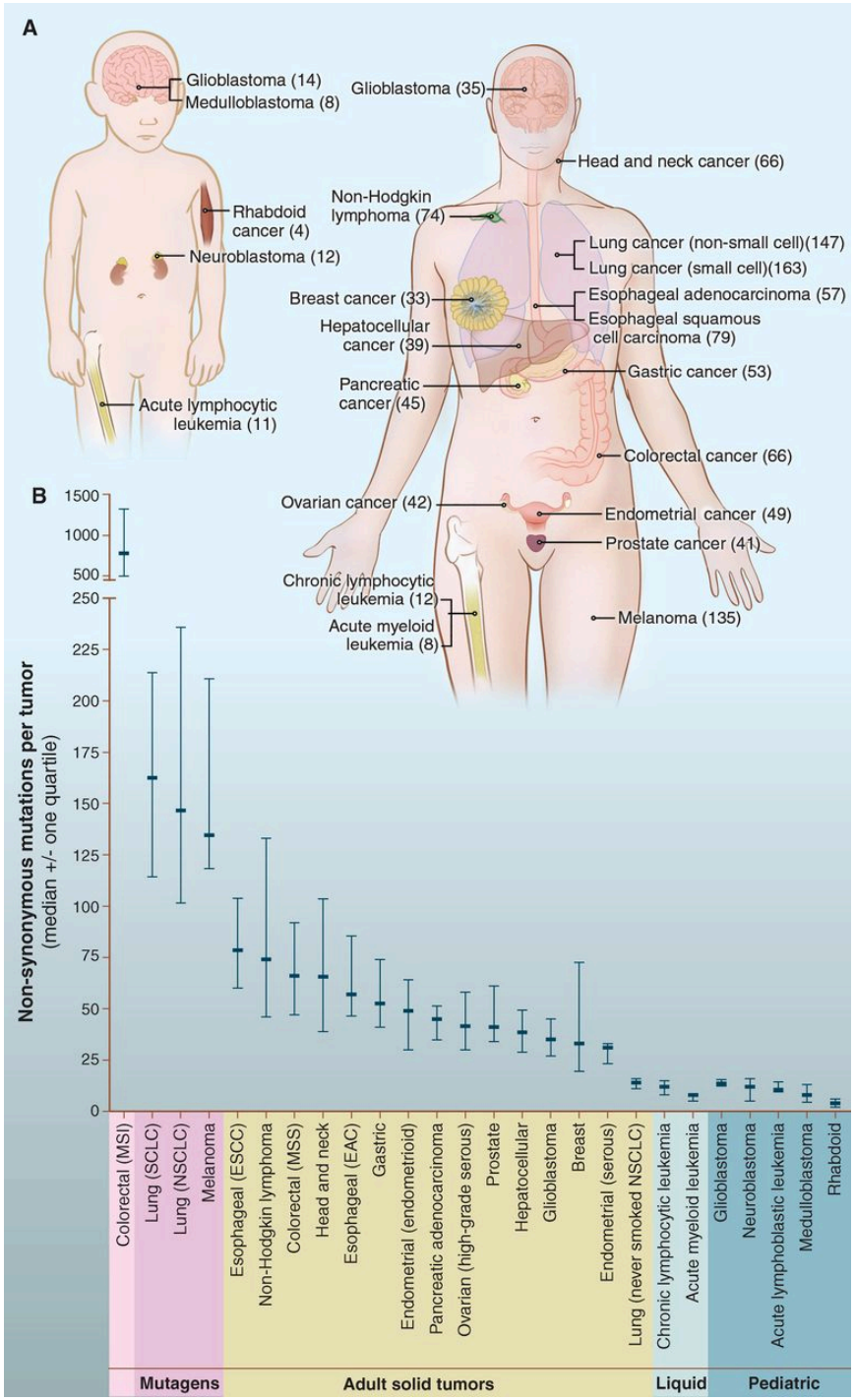
- From a review published 2013:
- ~350 cancer driver genes catalogued
- 5-7 driver mutations per tumor

  (Stratton et al, The Cancer Genome, Nature 2013)

- Exome seq/WGS studies suggest
  - higher number of driver genes
  - Up to 20 driver mutations per tumor

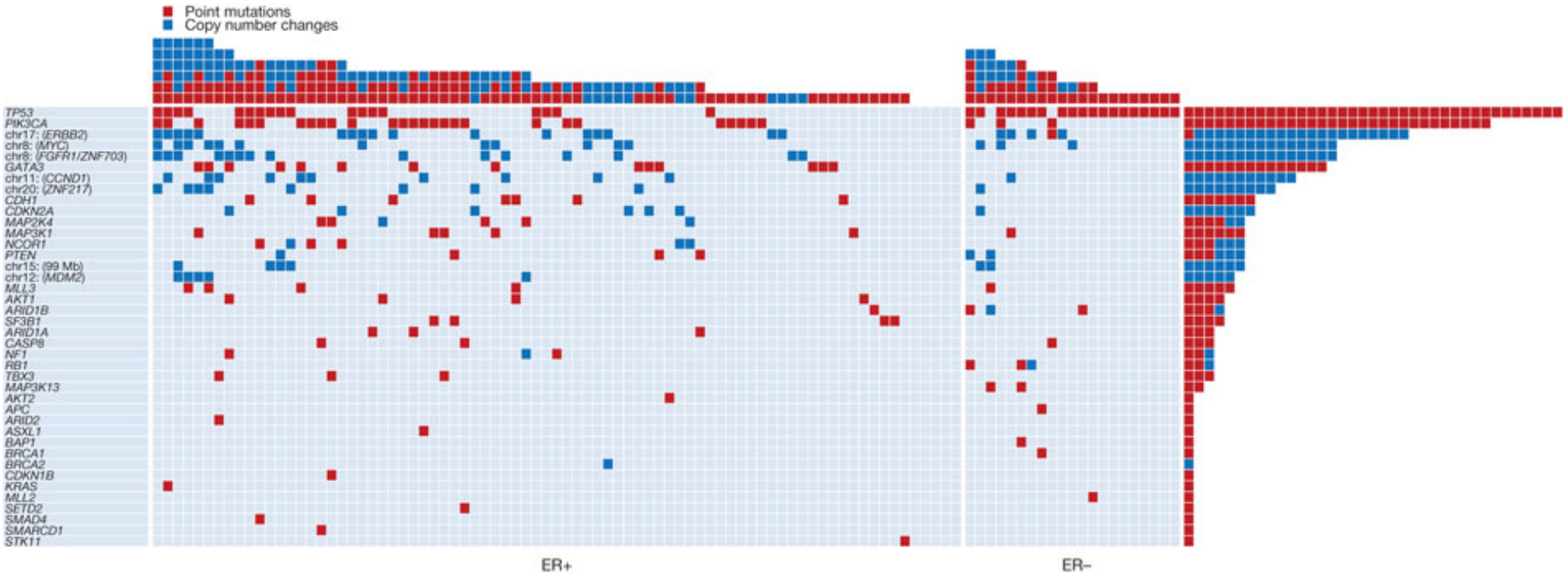
Number of somatic mutations in representative human cancers, detected by genome-wide sequencing studies.

Bert Vogelstein et al. Science 2013;339:1546-1558



Published by AAAS

The landscape of driver mutations in breast cancer



Rows: Cancer genes with driver mutations. In case of new
Columns: 100 primary breast cancer tumors (79 ER+, 21 ER-)

Coding exons of 21,416 protein coding genes and 1,664 microRNAs were sequenced

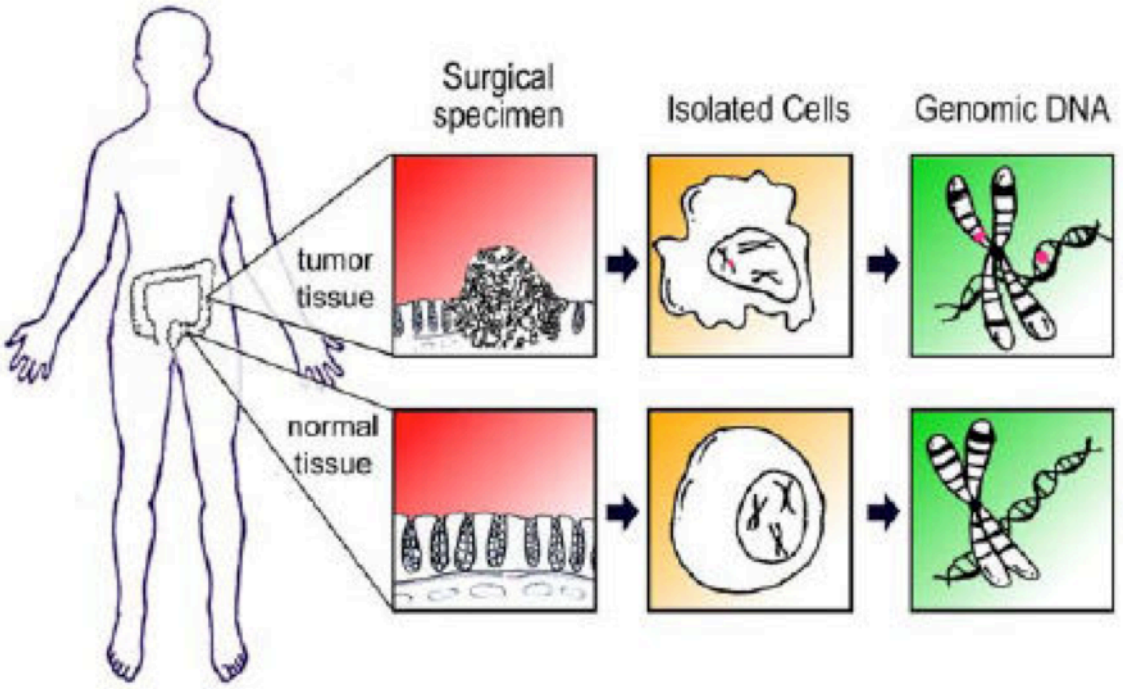
nature

PJ Stephens *et al*. *Nature* **000**, 1-5 (2012) doi:10.1038/nature11017

Detection of cancer mutations

We are interested in somatic events

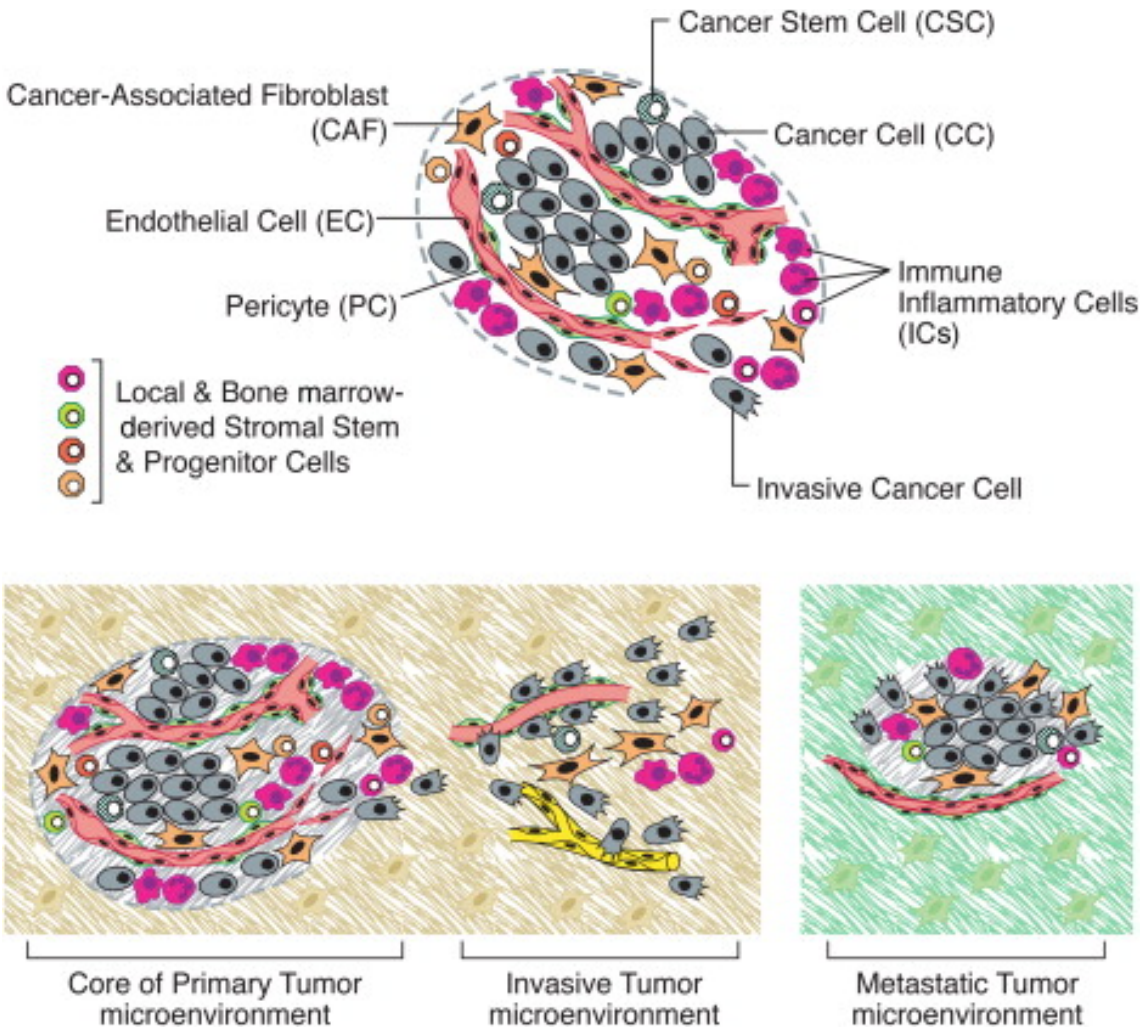
A matched “normal sample” needed to filter away germline variants



GGGATAC

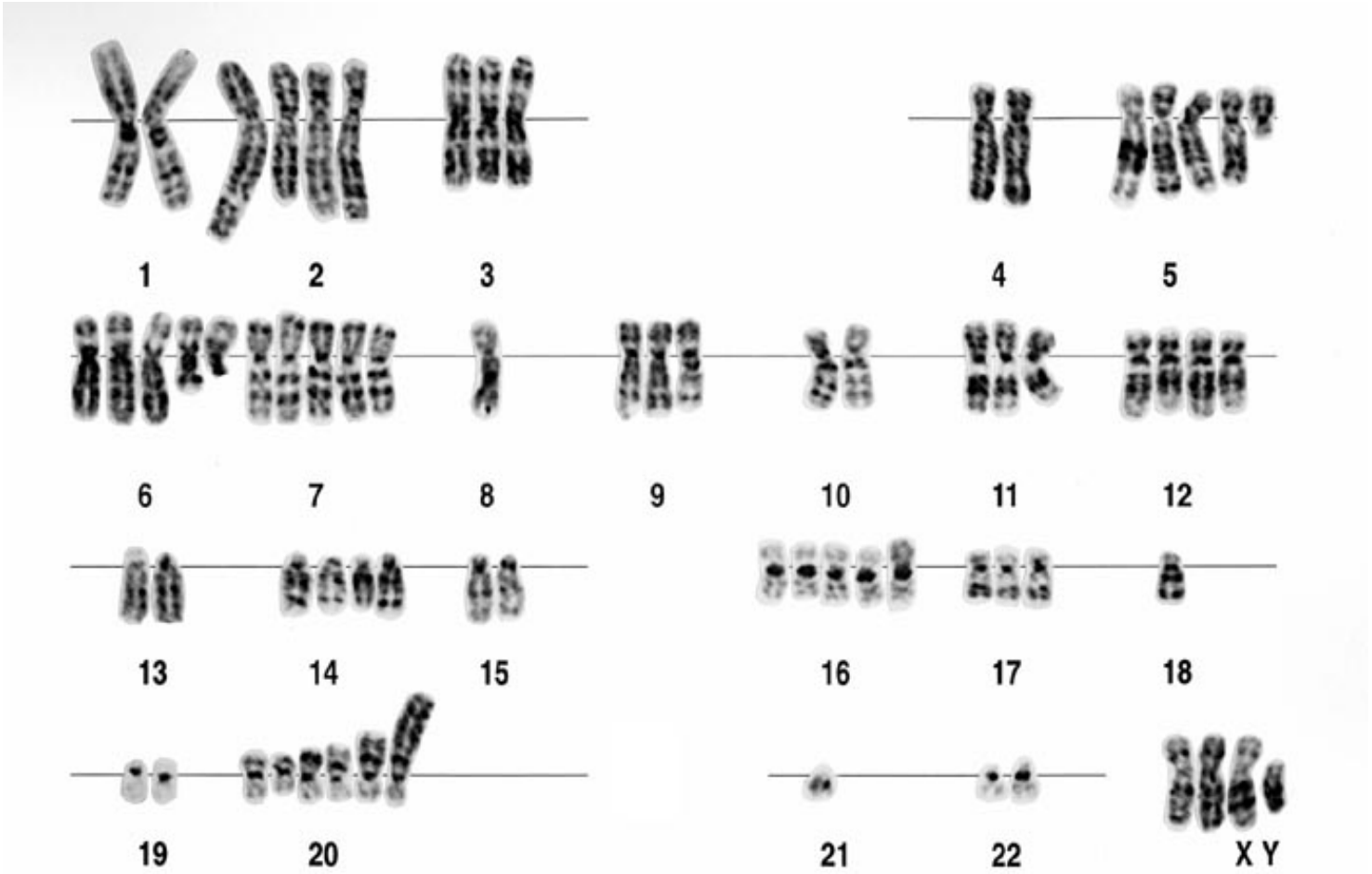
GGGAGAC

Tumor samples are often impure due to a mixture of tumor and normal cells

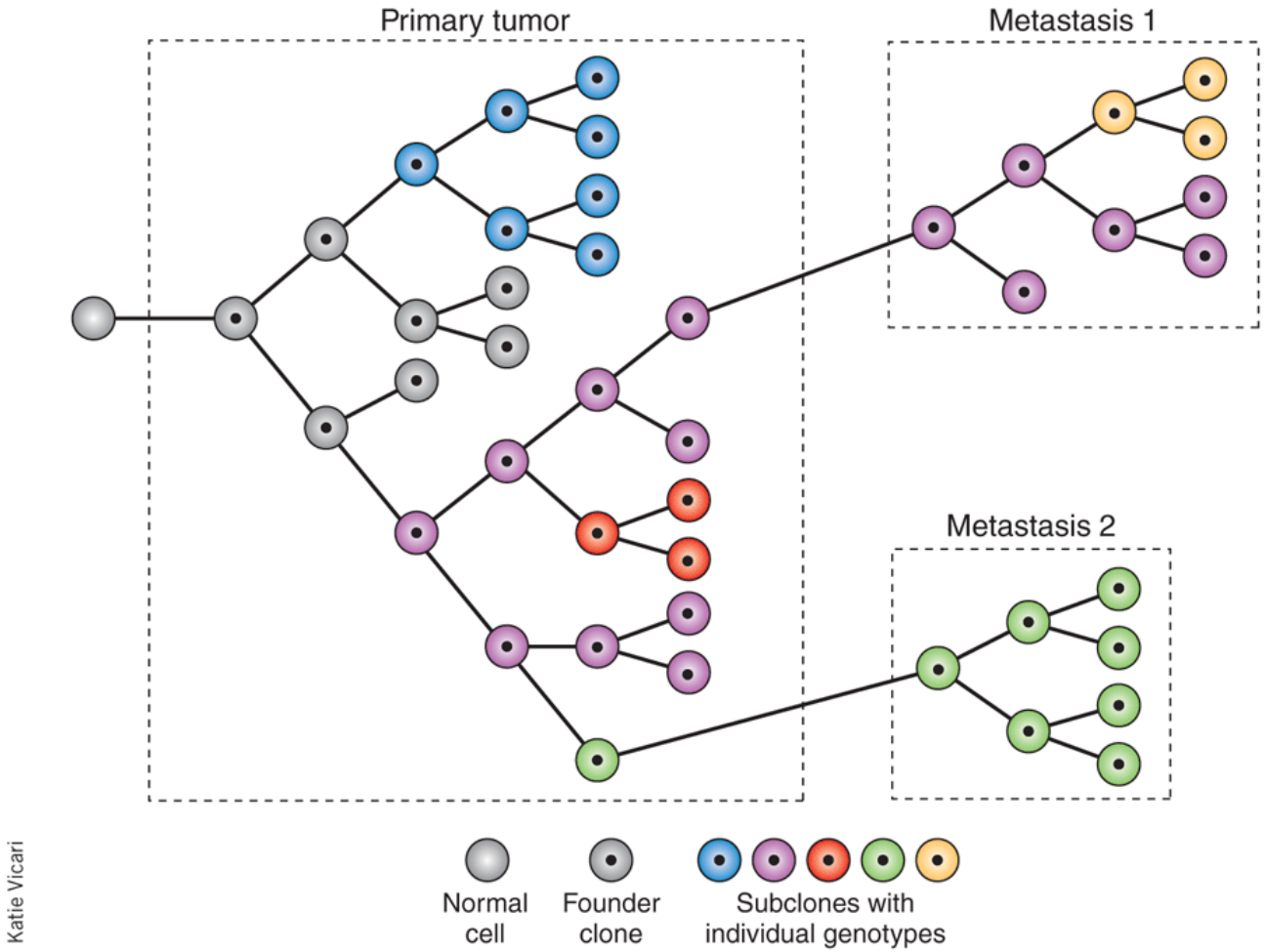


Hanahan and Weinberg, Hallmarks of Cancer: The Next Generation, Cell 2011

Aneuploidy



Tumors consists of subclones with different somatic mutations



Carlos Caldas, *Nature Biotechnology* **30**, 408–410 (2012) doi:10.1038/nbt.2213

So, detection algorithms must handle all of this!

Many tools available

- single nucleotide variants (SNVs)
  MuTect1, Strelka, MuTect2

- structural variants (SVs)
  Manta, Delly

- copy number variants (CNVs)
  Control-FREEC, ASCAT, Patchwork

Keep updated!

SciLifeLab WGS toolbox group:

https://wabi-wiki.scilifelab.se/display/SHGATG/SciLifeLab+human+genome+analysis+toolbox+group

Recommended tools and workflow for somatic variant calling (and other things)

Somatic variant calling Workflow

First…
recap of germline variant calling workflow

FastQ format

FASTQ format is a text-based format for storing both a nucleotide sequence and its corresponding quality scores.

```
@HWUSI-EAS100R:6:73:941:1973#0/1
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*(((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

1st row: sequence identifier (machine ID, x-y coordinates, additional info)

2nd row: The actual sequence

3rd row: starts with "+" and optionally the same identifier as in the 1st row

4th row: Quality score for each base in read

Quality score: ASCII representation of score for each base (i.e. the probability that the corresponding base call is incorrect.) Platform specific scaling!

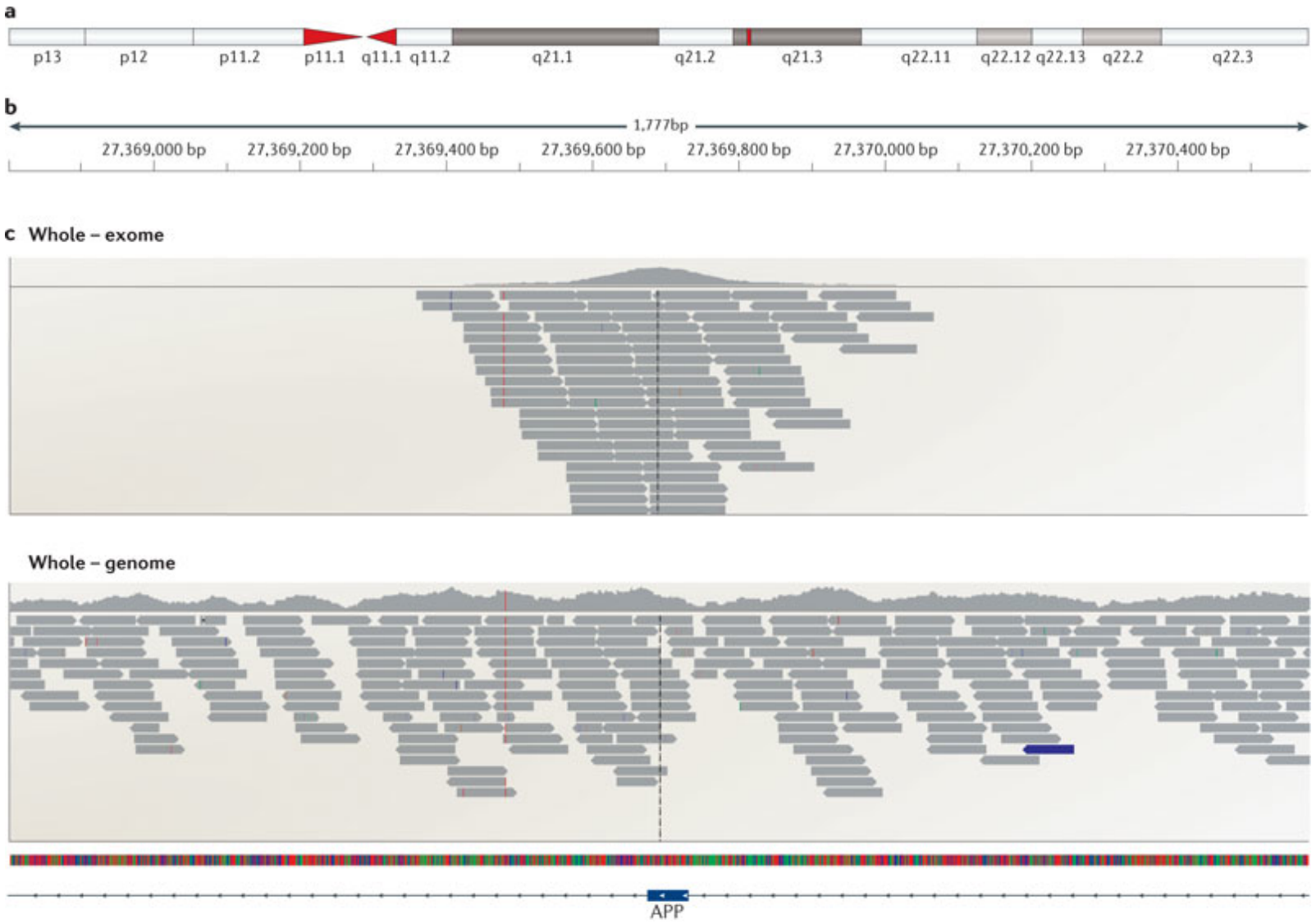
For more info: http://en.wikipedia.org/wiki/FASTQ_format

Output of experiment

```
:@M01674:9:000000000-A4148:1:1101:15048:1349 1:N:0:3
AGACGGTGACCGTGGTCCCTGTGCCCCAGACATCTCGGGTACTACCGTAGTAATCTTCTCTTGCACAG
TAATAGACTGCAGAGTCCTCTGATGTCAGGCTGCTGAGCTGCATGTAGGCTGTGTTGGA
+
AABCCCCCCFFCGGGGGGGGGGHHHHGGGHGHHHHHHGGHGHHHHHEFHGGHHHHHHHHHHHHHHHH
GHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHGHHEHHHHHHHHHHHHHHHHHHHH
@M01674:9:000000000-A4148:1:1101:15003:1351 1:N:0:3
CAGCCTTCATGCAGCTCAGCAGCCTTACATCTGAAGACTCTGCGGTCTATTTCTGCGCAAGAAAGGGG
AATTACTACGCCTAGGGGTACTTCGATGTCTGGGGCACAGGGACCACGGTCACCGTCTCCT
+
CCCCCFFFFFFFGGGGGGGGGGHHHGHHHHHHHHHHHHHHHHHGGGGGGHHHHHHHGGGGGGHGHHGG
GGHHHHHHHGGGHGHHHGGGHHHHGHHHHHHHHHGGGGHHHGGGGHHGGGGGHHGHGHHHH
@M01674:9:000000000-A4148:1:1101:14577:1352 1:N:0:3
CCTGCTTTTCGGGAAAACGGGATCACCACGATGGAACAGGTTAACGCAGGAATGCGCGTAGCCCGTCG
GCAGAATCGACCATTTCTGCCATCACCCGGGCAGTTTGTTGCATGGTGCCGGGAAGAAGCATCCGTTA
CCGCCGGACTGCCA
+
CCDDDFFFFFDDGGGGGGGGGGGGHHHHHGGHGGHHHHGHHHHHHGGGGGGHHHHGGGGGGGHHGGGG
GGGGGHHHHHGGGHHHHHHHHHHHHHHHGGFGGGHHHHHHHHHHHHHHHHHGGGGGGHHHHHHHHGGHG
GHGGGGGGGGGGGG
@M01674:9:000000000-A4148:1:1101:14770:1355 1:N:0:3
TCCAACACAGCCTTCATGCAACTCAGCAGCCTGACATCTGAGGACTCTGCAGTCTATTACTGTGCAAG
ATGGGGGTTACTAAGCGCTTACTGGGGCCAAGGGACTCTGGTCACTGTCTCTGCAGGT
+
CCDDDFCEEFDDGGGGGGGGGGHHHHHHHHHHHHHHHHHHHGHGGHHHHHHHHHHHHHIHHHHHHHHHH
FHHHHGGGGHHHHGHHGGGGGHHHHHGGGGHHGGHGHHHHHHHHHHHHHHHHHHHH
@M01674:9:000000000-A4148:1:1101:15309:1358 1:N:0:3
CCAACACAGCCTACATGCAGCTCAGCAGCCTGACATCTGAGGACTCTGCGGTCTATTACTGTGCAAGA
GGGGGGCTAATTACTACGGTAGTAGCCGACTACTGGGGCCAAGGCACCACTCTCACAGTCTCCTCAGG
TG
+
AACCCDCDFDCFGGGGGGGGGGHHHHHHHHHHHHHHHHHHHGHHHHHHHGGGGGHHHHHHHHHHHHH
HHGGGGGGHHHHHHHHHHGGHGGHHHHHHGGGGGHHHHHGGGGHHGHHHGGHHGGGGGGGGGGGGGGGG
FF
@M01674:9:000000000-A4148:1:1101:14985:1363 1:N:0:3
AGACGGTGACCGTGGTCCCTGTGCCCCAGACATCGAAGTCGGACCGTAGTAATAAGCCTCTTGCACAG
TAATAGACCGCAGAGTCCTCAGATGTCAGGCTGCTGAGTTGCATGAAGGCTGTGTTGGA
+
BCCCCCABBFFCGGGGGGGGGGHHHHGGGHGHHFGHFGEHEGEFGGGHGGHHHHHHHHHHHHHHHHH
GHHHHHHHHGGGGGHHHHHHHHGHHHHFHHHHHGHHFHHHHHHHHHHGHHFHHHHHHHHH
```

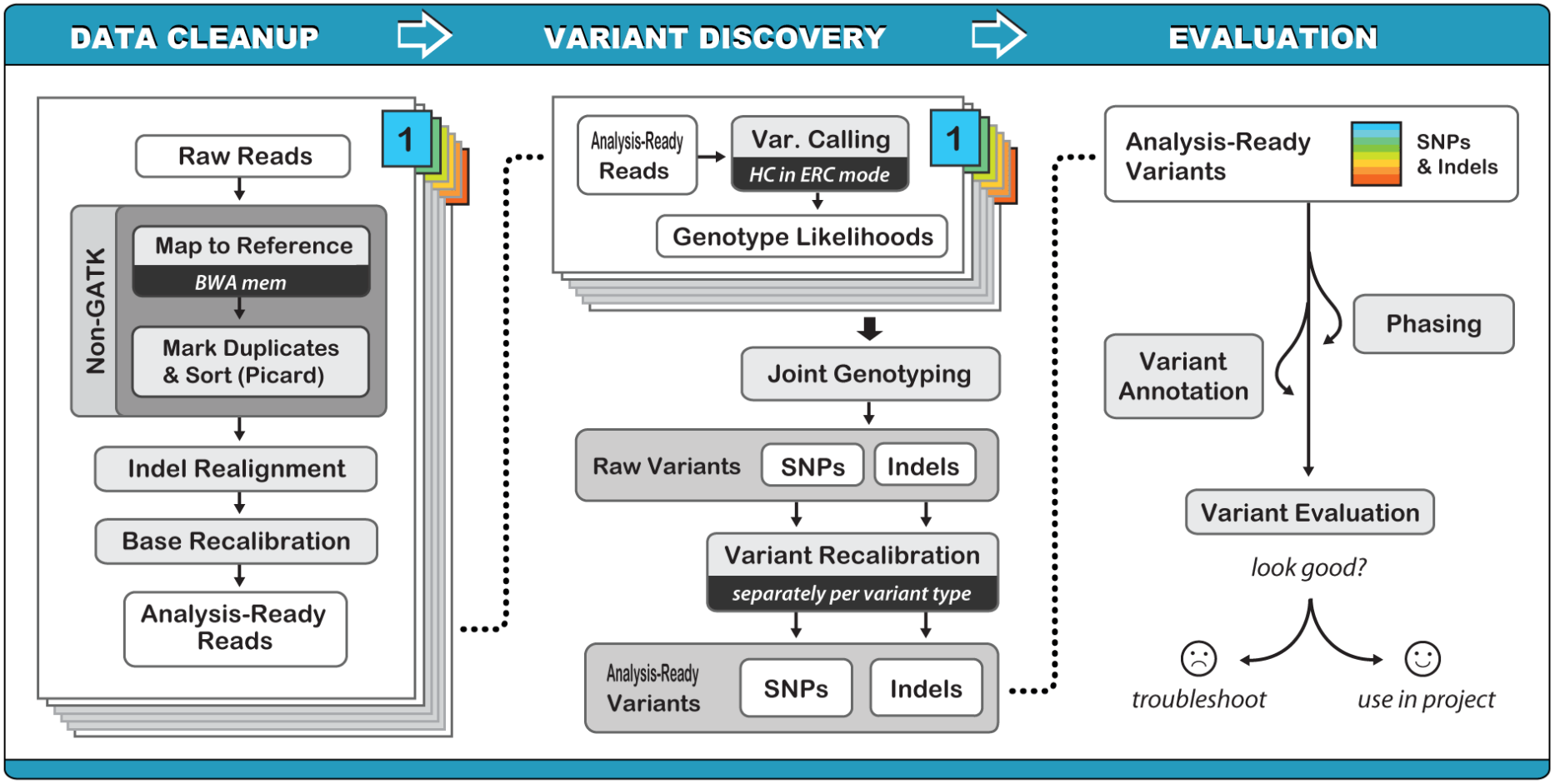
Fastq files
~7 Gb / exome

Goal:

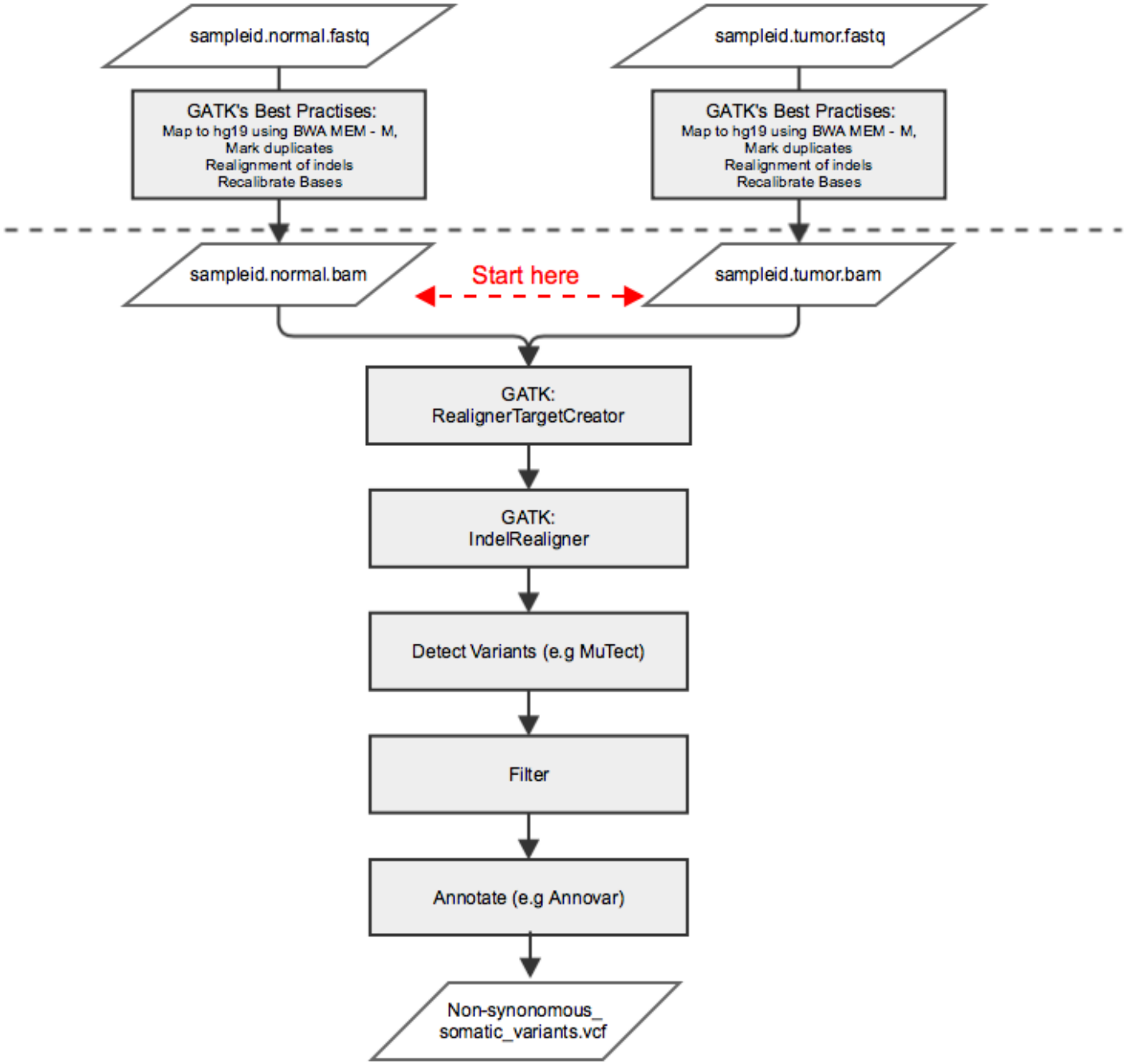


Nature Reviews | Neuroscience

Genome Analysis Tool Kit (GATK)



Somatic variant calling workflow


sampleid.normal.fastq
sampleid.tumor.fastq
GATK's Best Practises:
Map to hg19 using BWA MEM - M,
Mark duplicates
Realignment of indels
Recalibrate Bases
GATK's Best Practises:
Map to hg19 using BWA MEM - M,
Mark duplicates
Realignment of indels
Recalibrate Bases
sampleid.normal.bam
Start here
sampleid.tumor.bam
GATK:
RealignerTargetCreator
GATK:
IndelRealigner
Detect Variants (e.g MuTect)
Filter
Annotate (e.g Annovar)
Non-synonomous_
somatic_variants.vcf

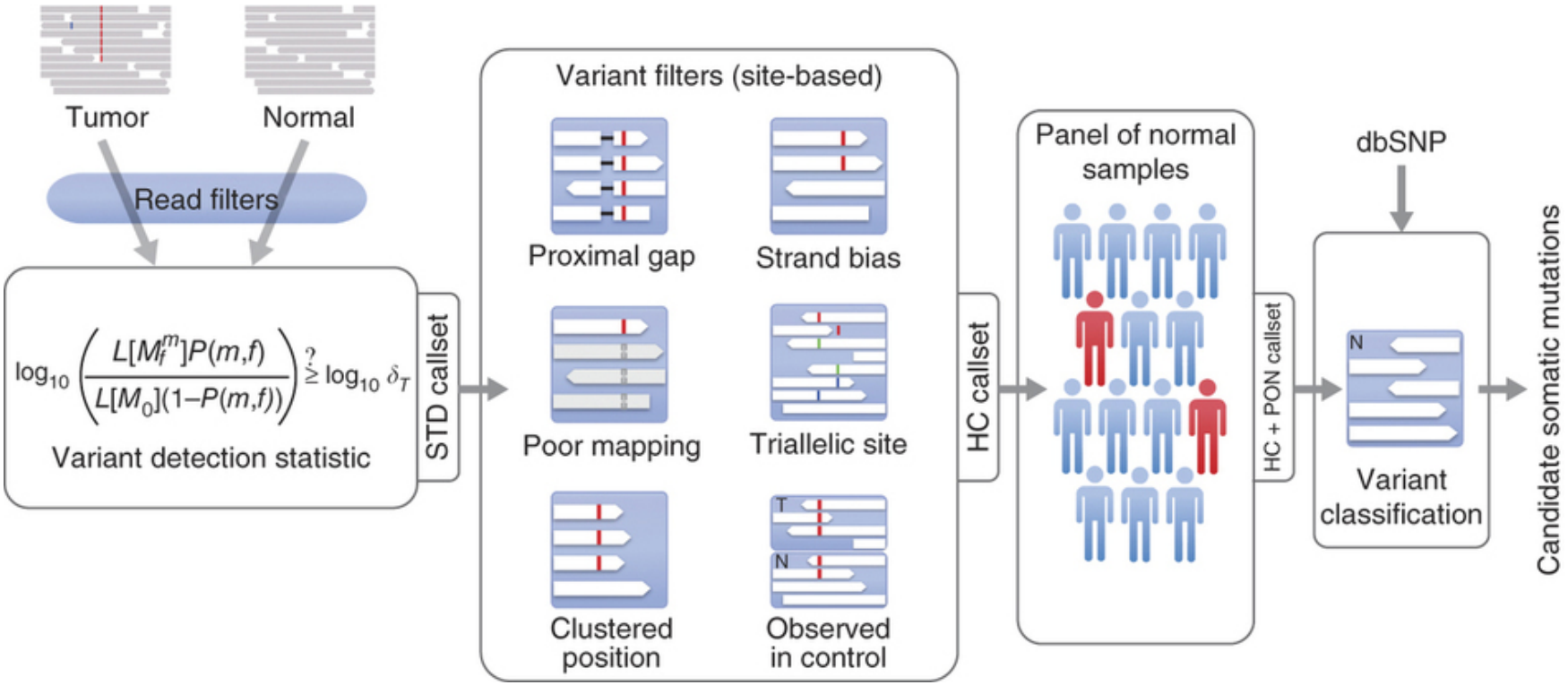
MuTect1

1. Identifies variants in tumor

   Differences between tumor DNA and human reference assembly (hg19)

2. Post detection filter to remove:
   - false positives due to non-independent sequencing errors
   - germ line variations (detected in normal)

MuTect1



Cibulskis, K. et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat Biotechnology (2013).doi:10.1038/nbt.2514

mutect.vcf

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	HCC1143.normal	HCC1143.tumor
17	1001315	.	C	T	.	REJECT	.	GT:AD:BQ:DP:FA	0:51,3:.:54:0.056	0/1:29,2:23:33:0.065
17	1001331	.	G	T	.	REJECT	.	GT:AD:BQ:DP:FA	0:30,3:.:33:0.091	0/1:15,2:34:17:0.118
17	1003390	.	G	A	.	REJECT	.	GT:AD:BQ:DP:FA	0:17,2:.:18:0.105	0/1:16,1:28:17:0.059
17	1004967	.	A	T	.	REJECT	.	GT:AD:BQ:DP:FA	0:28,1:.:29:0.034	0/1:16,4:15:20:0.200
17	1004974	.	C	T	.	REJECT	.	GT:AD:BQ:DP:FA	0:27,2:.:29:0.069	0/1:11,3:13:14:0.214
17	1024903	.	C	T	.	PASS	SOMATIC	GT:AD:BQ:DP:FA:SS	0:106,0:.:102:0.00:0	0/1:84,6:34:90:0.067:2
17	1277664	.	C	A	.	PASS	SOMATIC	GT:AD:BQ:DP:FA:SS	0:59,0:.:59:0.00:0	0/1:41,25:34:66:0.379:2
17	1527066	.	C	G	.	PASS	SOMATIC	GT:AD:BQ:DP:FA:SS	0:35,0:.:31:0.00:0	0/1:26,5:29:31:0.161:2

FORMAT (Each code is described in VCF header)

GT:AD:BQ:DP:FA

GT=Genotype

AD=Allelic depths for the ref and alt alleles in the order listed

BQ=Average base quality for reads supporting alleles

DP=Approximate read depth

FA=Allele fraction of the alternate allele with regard to reference

SS=Variant status

(0=wildtype,1=germline,2=somatic,3=LOH,4=post-transcriptional modification,5=unknown")

mutect.out file

All statistics used in post-detection filtering

Columns:

contig position context ref_allele alt_allele tumor_name normal_name score dbsnp_site covered power tumor_power normal_power normal_power_nsp normal_power_wsp total_reads map_Q0_reads init_t_lod t_lod_fstar t_lod_fstar_forward t_lod_fstar_reverse tumor_f contaminant_fraction contaminant_lod t_q20_count t_ref_count t_alt_count t_ref_sum t_alt_sum t_ref_max_mapq t_alt_max_mapq t_ins_c ount t_del_count normal_best_gt init_n_lod normal_f n_q20_count n_ref_count n_alt_count n_ref_s um n_alt_sum power_to_detect_positive_strand_artifact power_to_detect_negative_strand_artifact strand_ bias_counts tumor_alt_fpir_median tumor_alt_fpir_mad tumor_alt_rpir_median tumor_alt_rpir_mad observed_in_nor mals_count failure_reasons judgement

Example row:

17 1001315 TTTxTTT C T HCC1143.tumor HCC1143.normal 0 DBSNP COVERED 0.954491 0.954491 11 1 103 0 -3.640633 2.499583 0 3.065049 0.064516 0.02 -0.4105 76 41 29 2 893 47 70 70 0 6 CC 5.640677 0.055556 47 51 3 1476 91 0.560361 0.544179 (15,14,0,2) 2.5 0.5 83.5 8.5 0 fstar_t umor_lod,nearby_gap_events,possible_contamination,alt_allele_in_normal,clustered_read_position REJECT

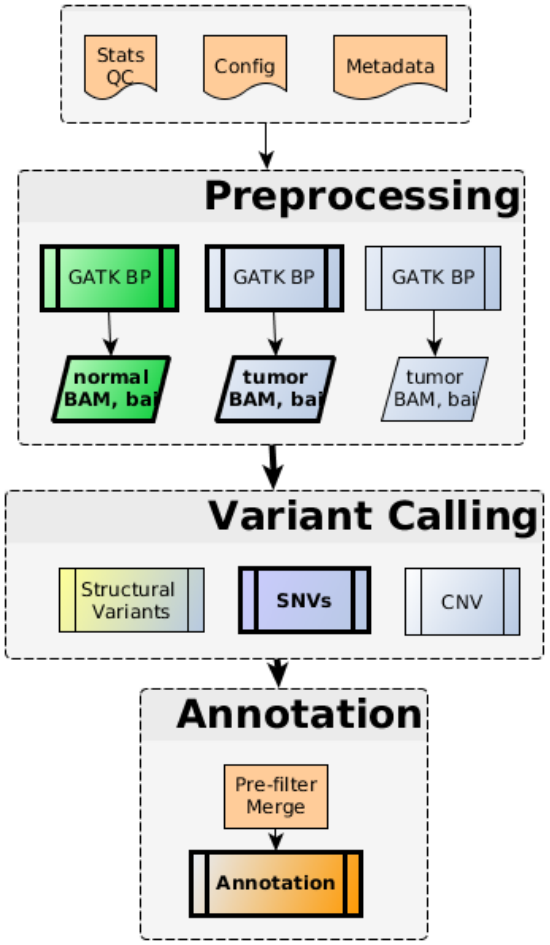
Annotation

Link detected variants to functional sites in the genome

- Protein coding exons
- UTR
- Regulatory regions
- Database of known variation
  - dbSNP / 1000 Genomes / ExAC for normal variants
  - Cosmic for cancer mutations

SciLifeLab Cancer Analysis Workflow

- https://github.com/SciLifeLab/CAW

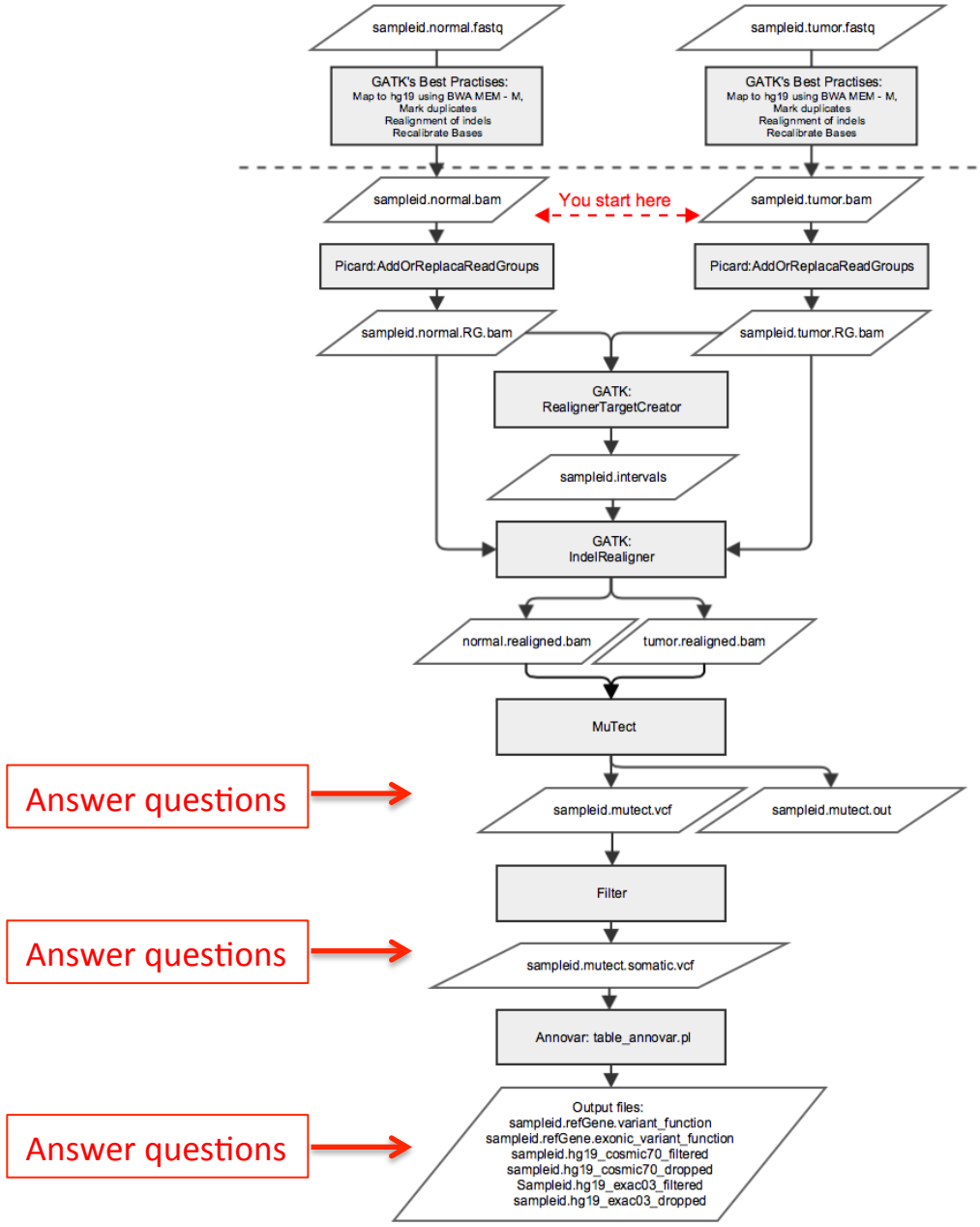


Todays practical
Part one

Analyze somatic mutations in WGS data from breast cancer cell lines and matched normal controls

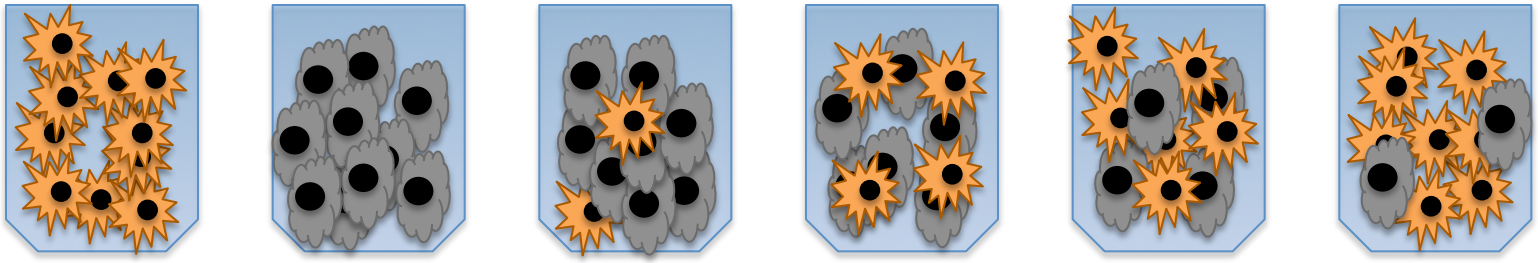
- Preprocess bam files
- Detect SNVs with MuTect
- Annotate variants with Annovar (RefGene, ExAC and Cosmic databases)
- Only for a small part of chromsome 17

Part One


sampleid.normal.fastq
sampleid.tumor.fastq
GATK's Best Practises:
Map to hg19 using BWA MEM - M,
Mark duplicates
Realignment of indels
Recalibrate Bases
GATK's Best Practises:
Map to hg19 using BWA MEM - M,
Mark duplicates
Realignment of indels
Recalibrate Bases
sampleid.normal.bam
You start here
sampleid.tumor.bam
Picard:AddOrReplacaReadGroups
Picard:AddOrReplacaReadGroups
sampleid.normal.RG.bam
sampleid.tumor.RG.bam
GATK:
RealignerTargetCreator
sampleid.intervals
GATK:
IndelRealigner
normal.realigned.bam
tumor.realigned.bam
MuTect
Answer questions
sampleid.mutect.vcf
sampleid.mutect.out
Filter
Answer questions
sampleid.mutect.somatic.vcf
Annovar: table_annovar.pl
Answer questions
Output files:
sampleid.refGene.variant_function
sampleid.refGene.exonic_variant_function
sampleid.hg19_cosmic70_filtered
sampleid.hg19_cosmic70_dropped
Sampleid.hg19_exac03_filtered
sampleid.hg19_exac03_dropped

Todays Practical
part two

- Same samples - data already generated for entire genome
- Check basic statistics (#detected mutations)
- Analyze how various degrees of normal contamination of the tumor sample affects allele frequencies



http://scilifelab.github.io/courses/ngsgu/cancergenomics/1610/