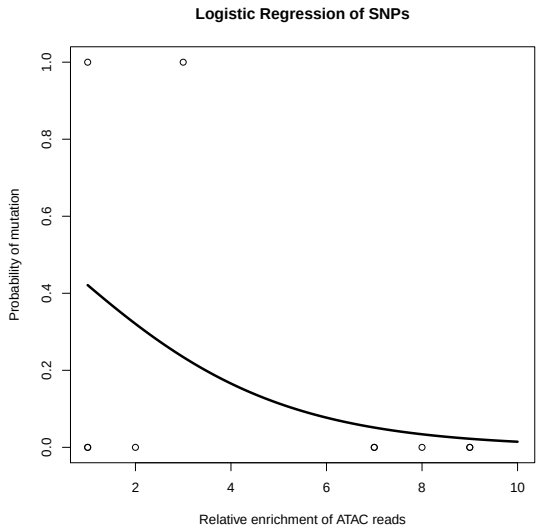
Evaluating the Relationship between Chromatin State and Mutation Rate in Cancer

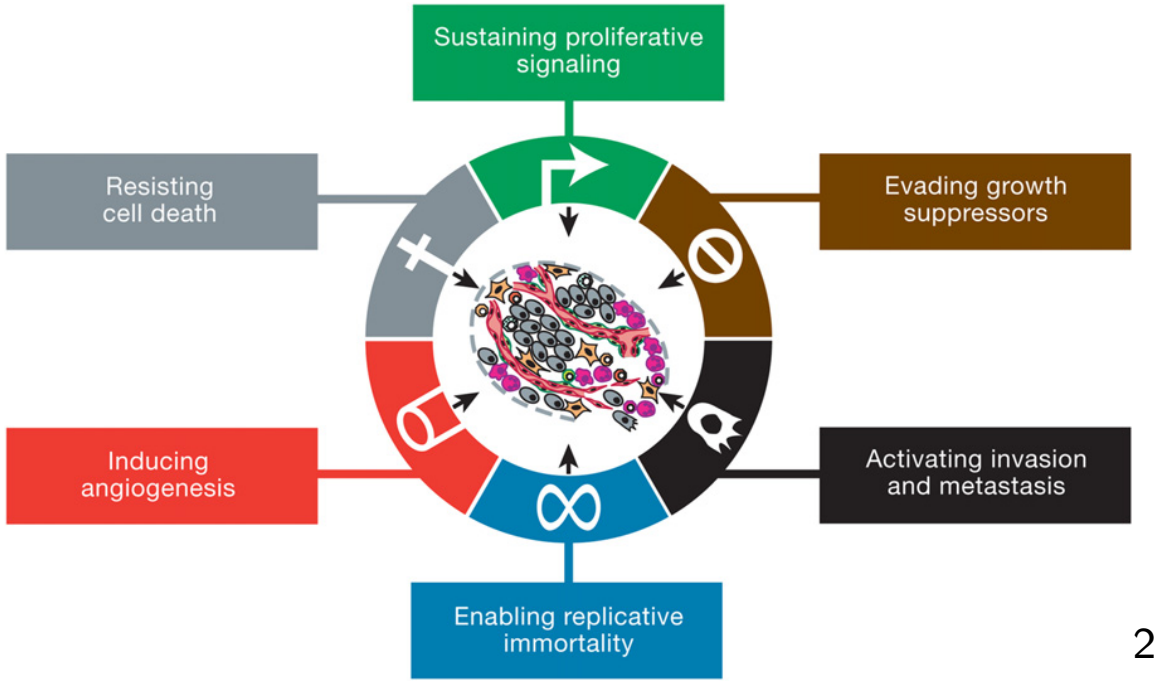
Adam Orr

11/9/17



Cancer is a major health problem

- Estimated 1.6 million new cancer cases and 600,000 deaths in 2016 [1]
- Cancer is difficult to treat
- Many cancers acquire drug resistance (Holohan *et al.* 2013)

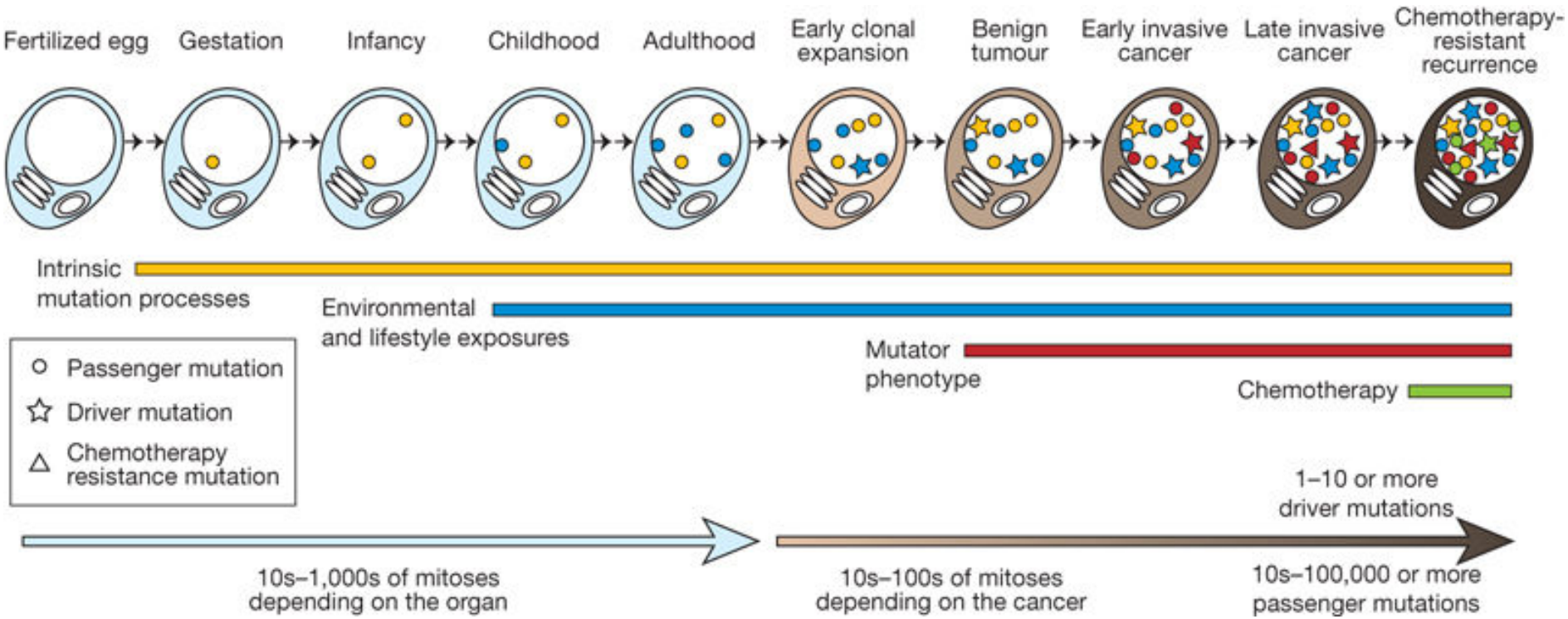


[2]

[1] https://www.cancer.gov/about-cancer/understanding/statistics

[2] Hanahan & Weinberg 2011

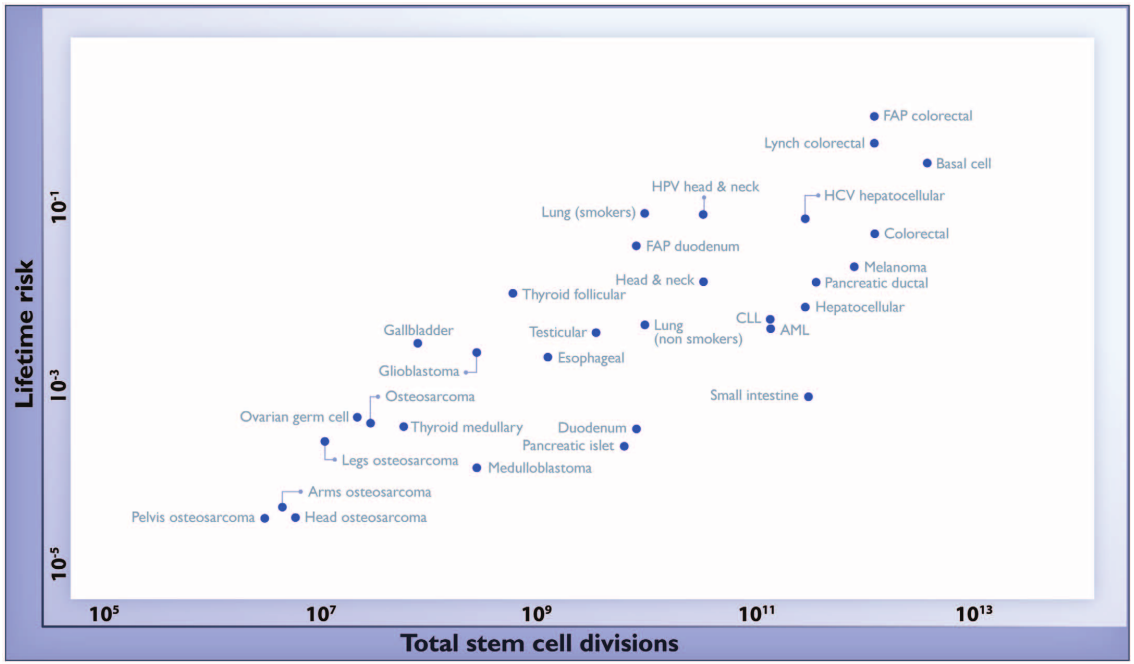
Cancer is believed to be caused by somatic mutations



[1] Stratton *et al.* 2009

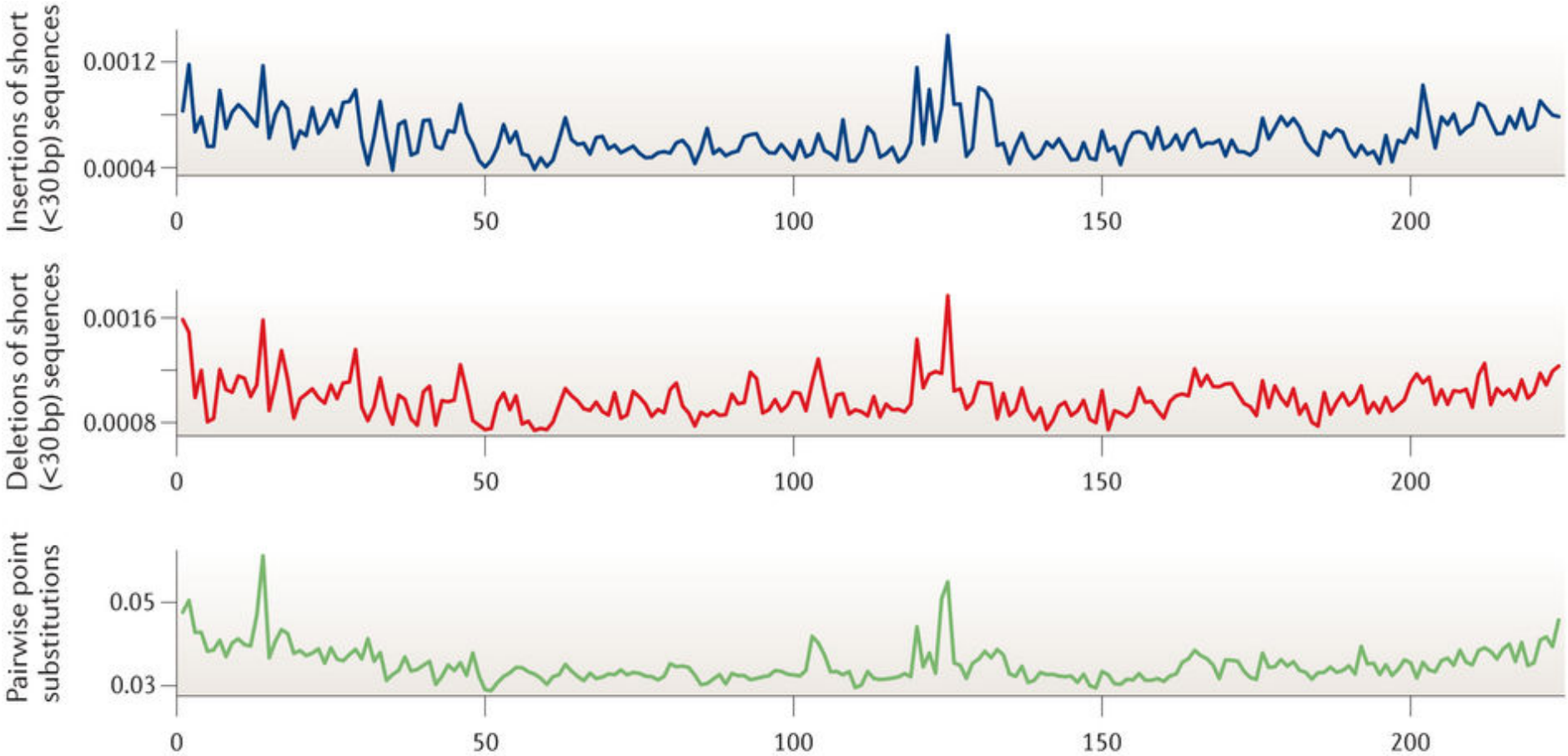
Understanding cancer initiation is critical for prevention

- Some claim Somatic Mutation Theory doesn't explain how sufficient mutations accumulate (Baker 2015)
- Disputed evidence that lifetime cancer incidence is linearly correlated with number of somatic cell divisions (Tomasetti & Vogelstein 2015)



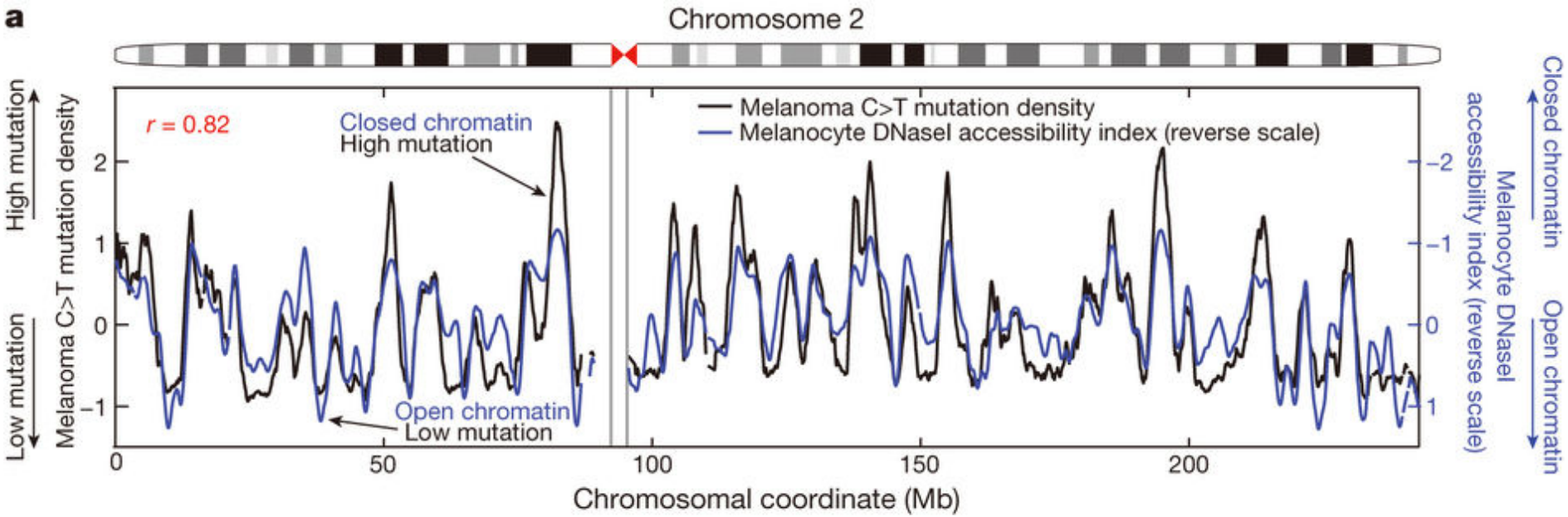
[1] Tomasetti & Vogelstein 2015

RViMR may explain how so many mutations occur in the correct places



Mutation rates in 1Mb windows across chromosome 1 for different types of mutations (Makova & Hardison 2015)

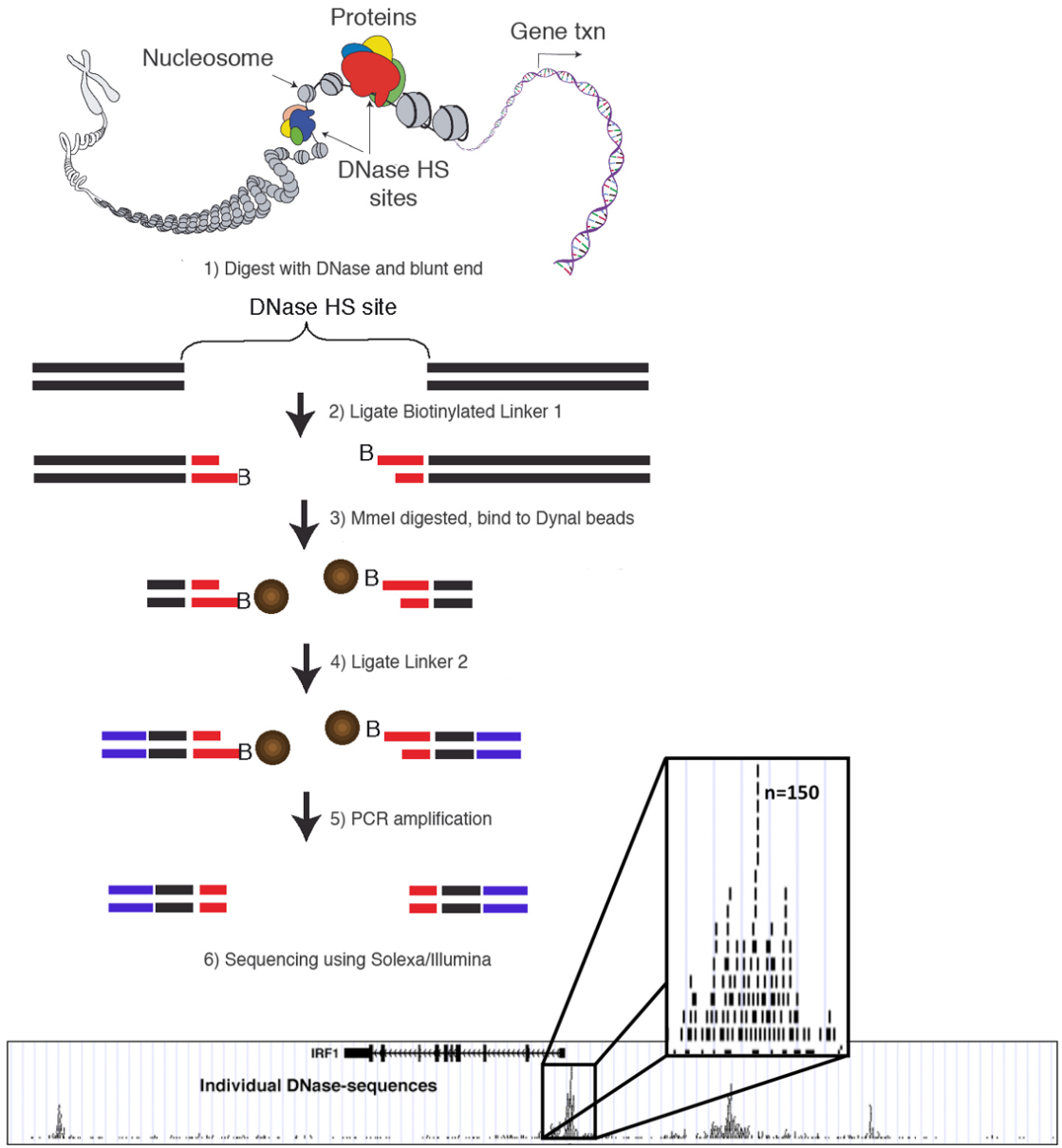
Epigenetic markers significantly correlate with local mutation rate



- Closed chromatin regions correlate with high single nucleotide mutation rates.
- Open chromatin regions correlate with high insertion/deletion rates (Makova & Hardison 2015)

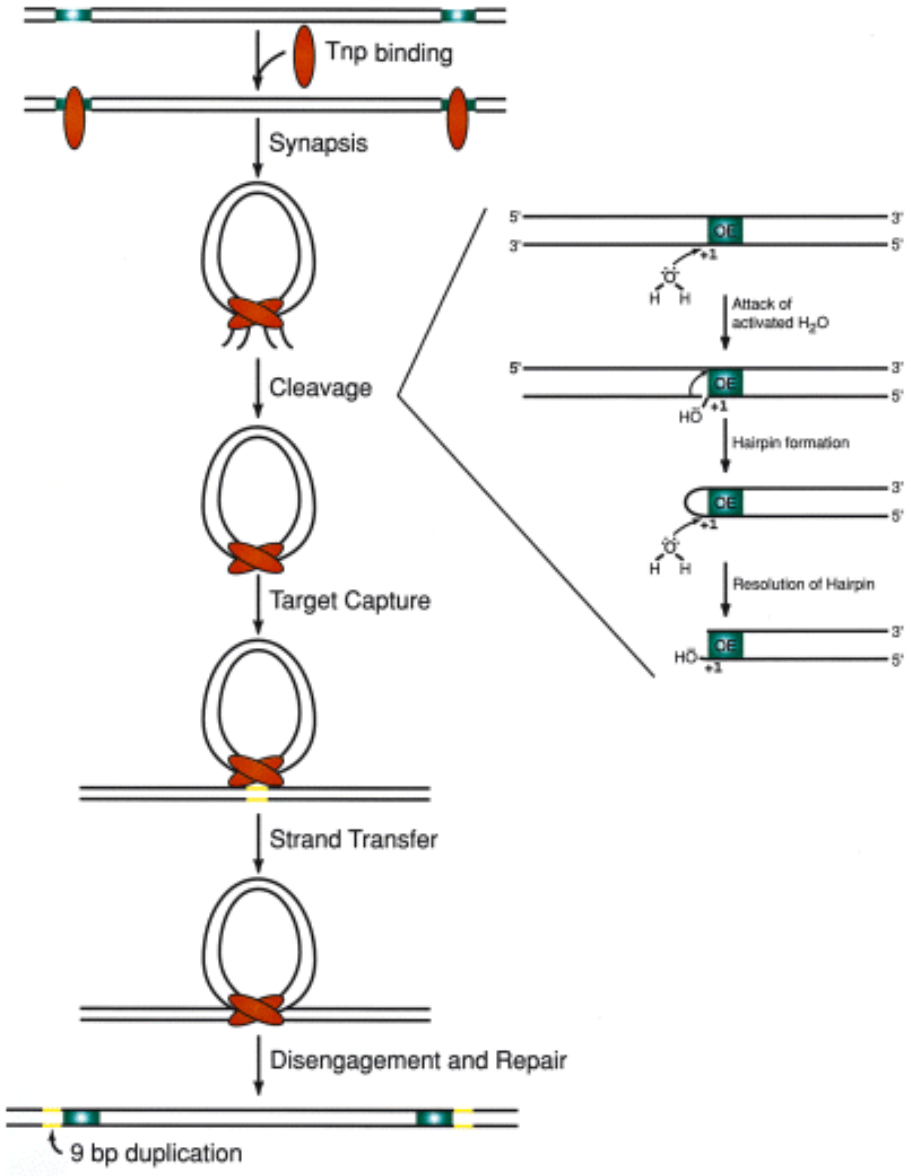
[1] Polak *et al.* 2015

DNAse-seq



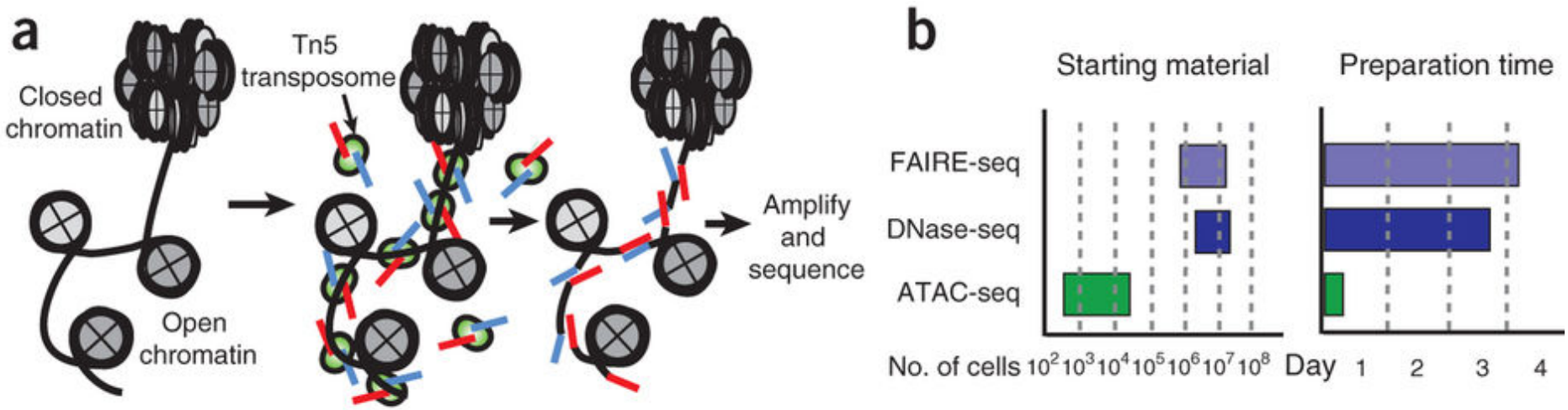
[1] Song & Crawford 2010

Transposase



[1] Reznikoff 2003

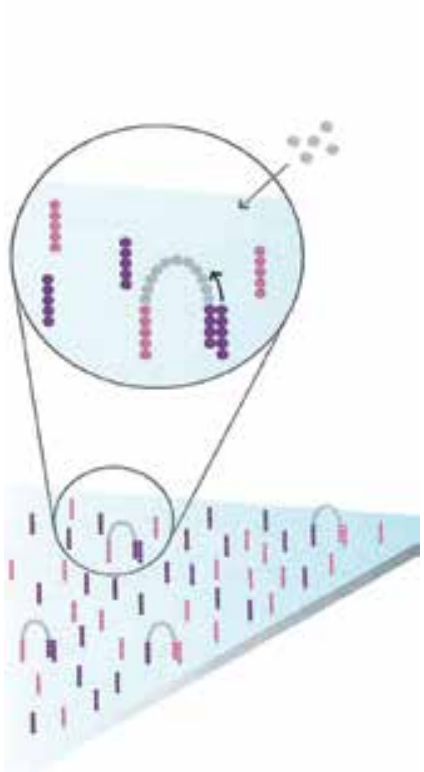
ATAC-seq



[1] Buenrostro *et al.* 2013

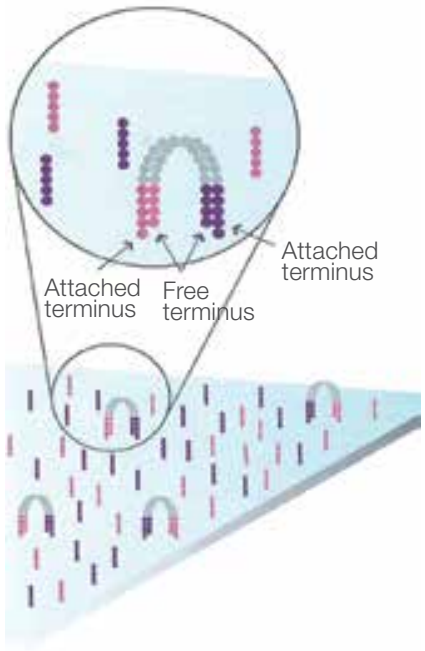
Illumina Sequencing is useful for detecting SNPs

Figure 4: Bridge Amplification



Add unlabeled nucleotides and enzyme to initiate solid-phase bridge amplification.

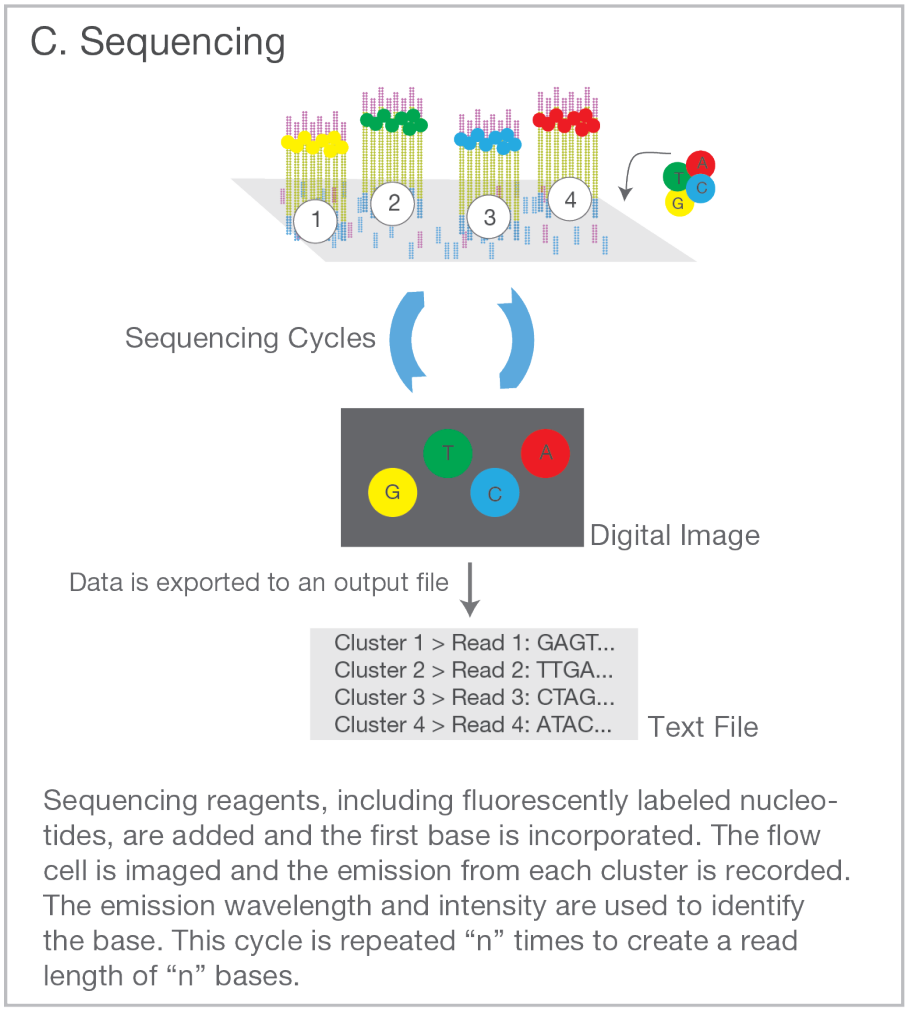
Figure 5: Fragments Become Double Stranded



The enzyme incorporates nucleotides to build double-stranded bridges on the solid-phase substrate.

[1] https://www.illumina.com/documents/products/techspotlights/techspotlight_sequencing.pdf

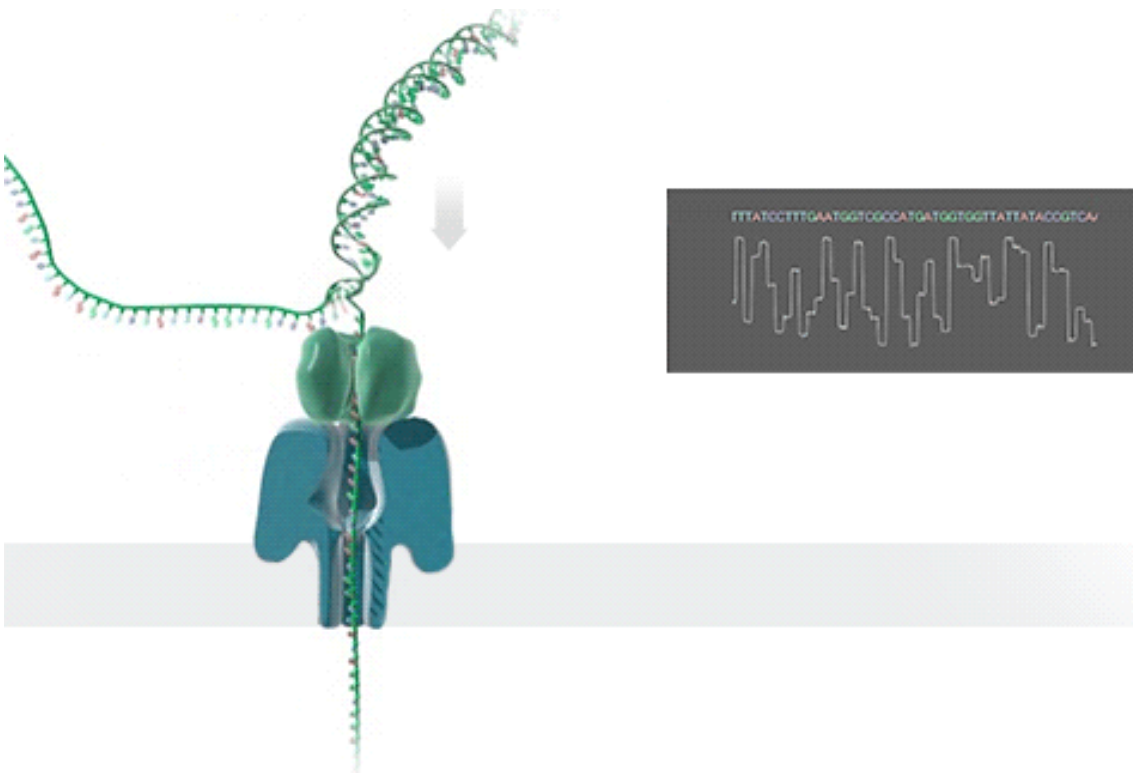
Illumina Sequencing is useful for detecting SNPs



[2] https://www.illumina.com/content/dam/illumina-marketing/documents/products/illumina_sequencing_introduction.pdf

Nanopore Sequencing is useful for detecting insertions and deletions

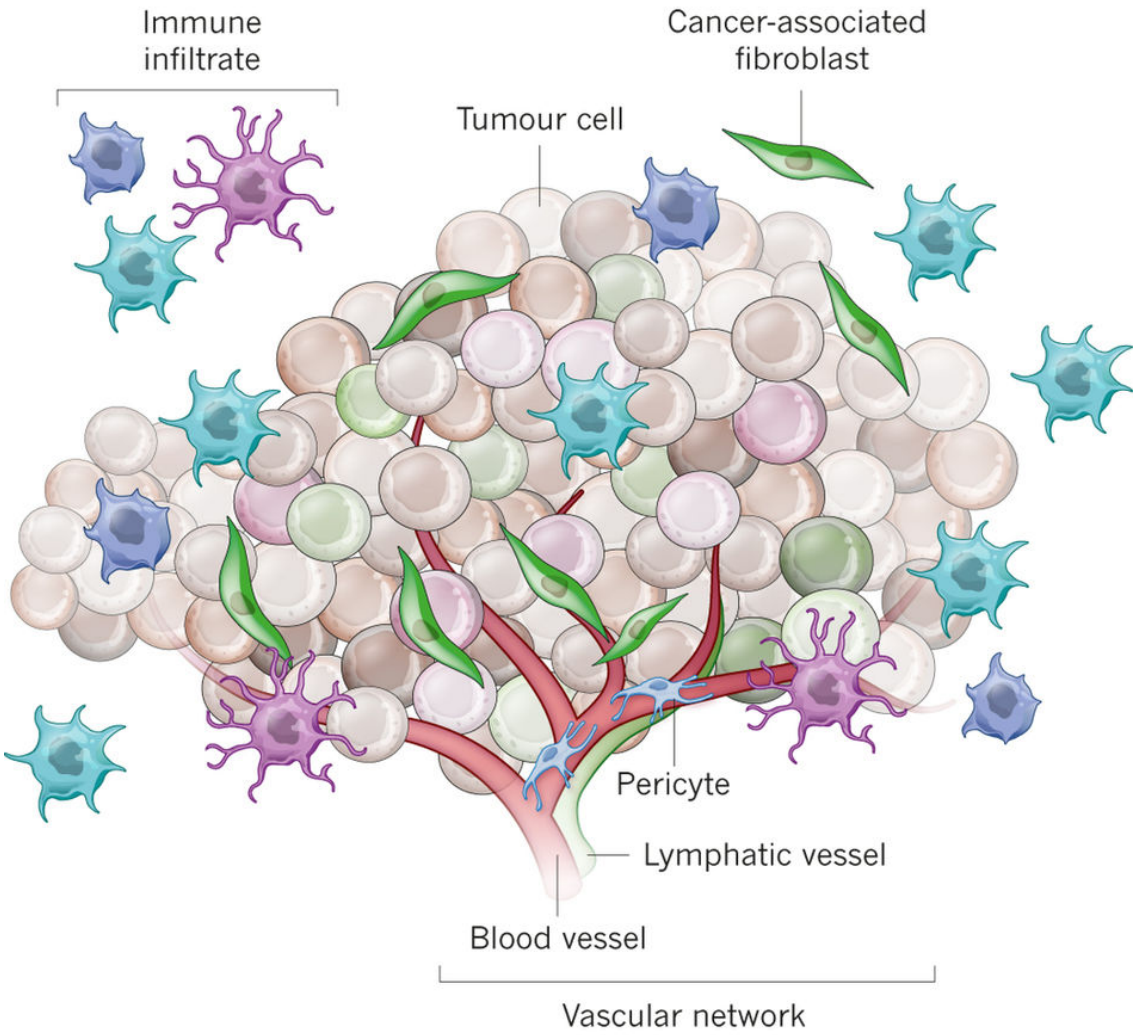
- A voltage is generated across the membrane containing the nanopore.
- Molecules that pass through the pore cause a detectable shift in current.
- α-hemolysin is used as a biological pore
- Helicase is used as a motor



[1] https://nanoporetech.com/how-it-works

Heterogeneity is a problem

This can be reduced by careful sampling of tissue



[1] Junttila & de Sauvage 2013

Sequencing errors are a problem

These errors are caused by

- DNA damage
- Mutations during library preparation
- Sequencer measurement error

The first two can be minimized by careful DNA extraction methods and minimizing the number of PCR steps required before sequencing.

Remaining errors can be corrected *in silico*

Read correction reduces errors

In short reads:

1. Create De Bruijn graph and K-mer count table
2. Determine a global threshold and local threshold
3. For every read containing a K-mer with a count lower than the threshold, make the minimum number of changes to the kmers to make them higher than the threshold.

In long reads:

- The reads are long enough that there is a significant level of overlap between reads. Use this overlap to find a consensus.
- **Or** align short reads to the long reads to "polish" the long reads and repair errors.

Aims

How does chromatin accessibility correlate with mutation rate in the same tissue? How does it change in cancer?

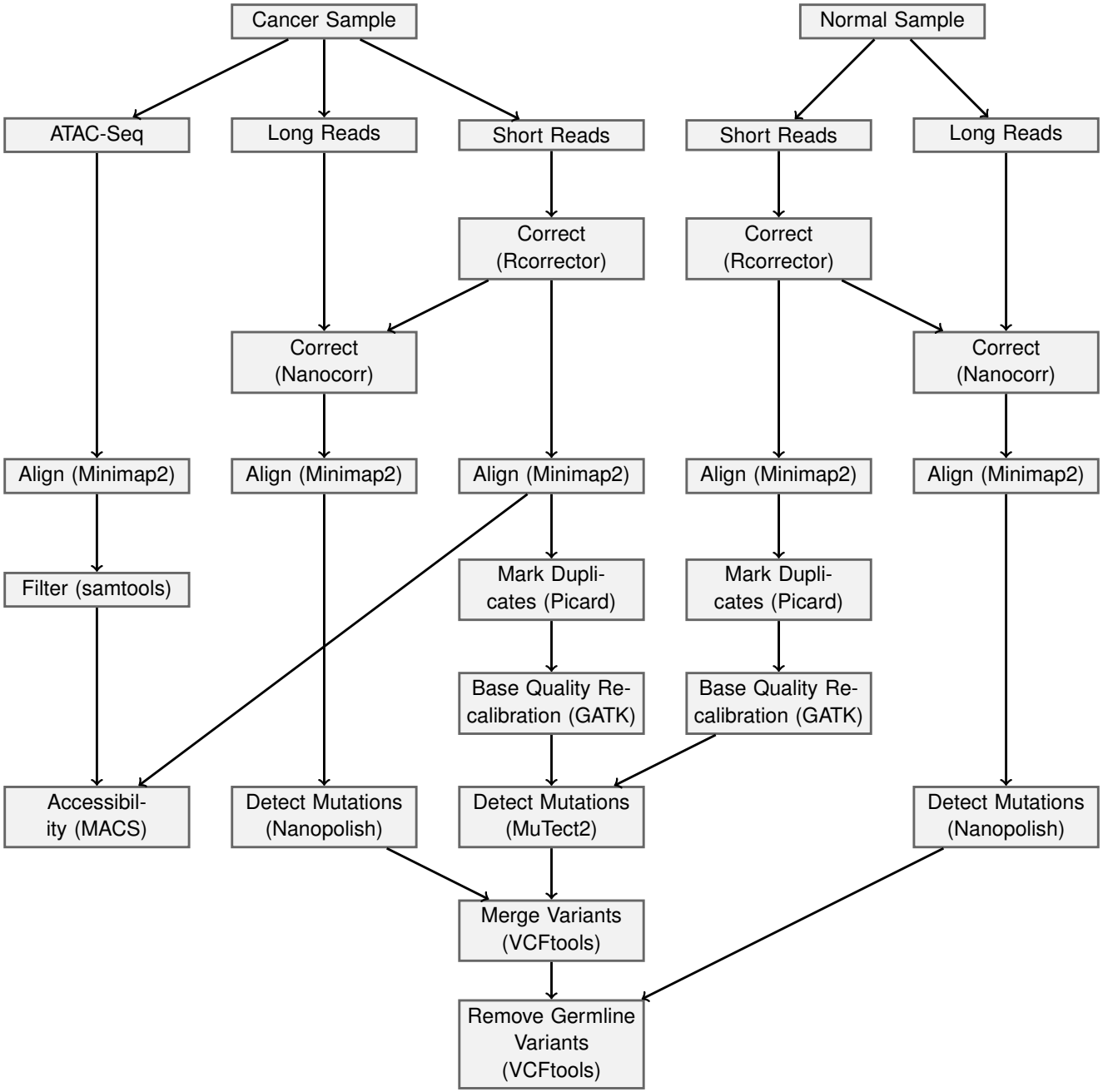
Aim 1: Develop a bioinformatic pipeline to detect somatic mutations and estimate chromatin accessibility across the genome in somatic samples.

- Use long reads for structural variation, short reads for single nucleotide changes
- Error correct reads

Aim 2: Test the hypothesis that chromatin accessibility significantly impacts mutation rate.

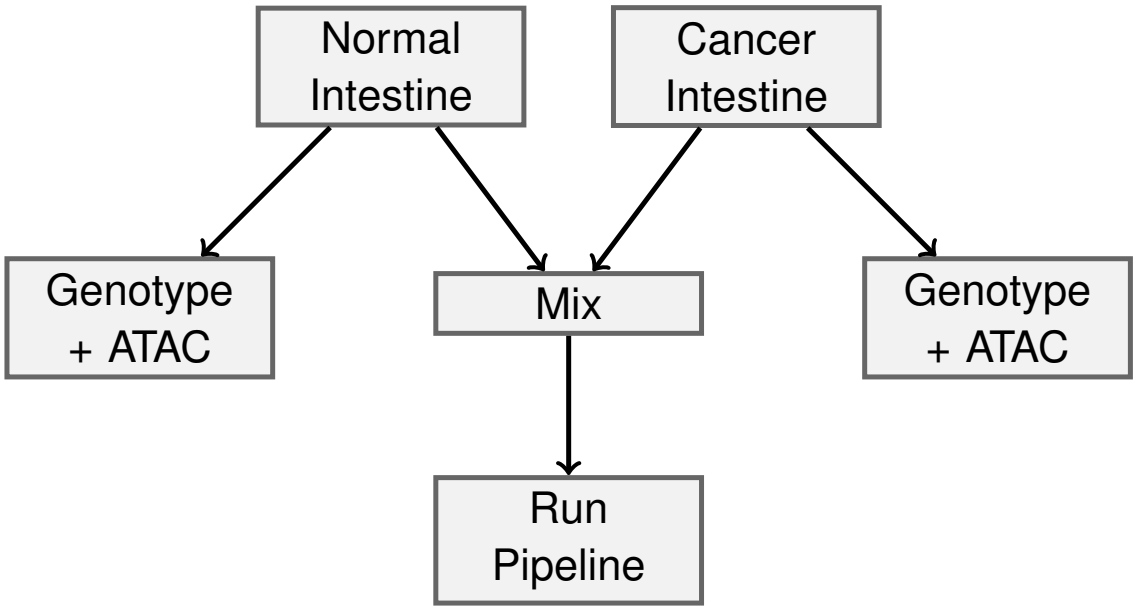
- Poisson test of mutation rate in "open" and "closed" regions
- Non-parametric test of ATAC-seq enrichment at mutated sites
- Logistic regression of ATAC-seq enrichment

Pipeline Overview



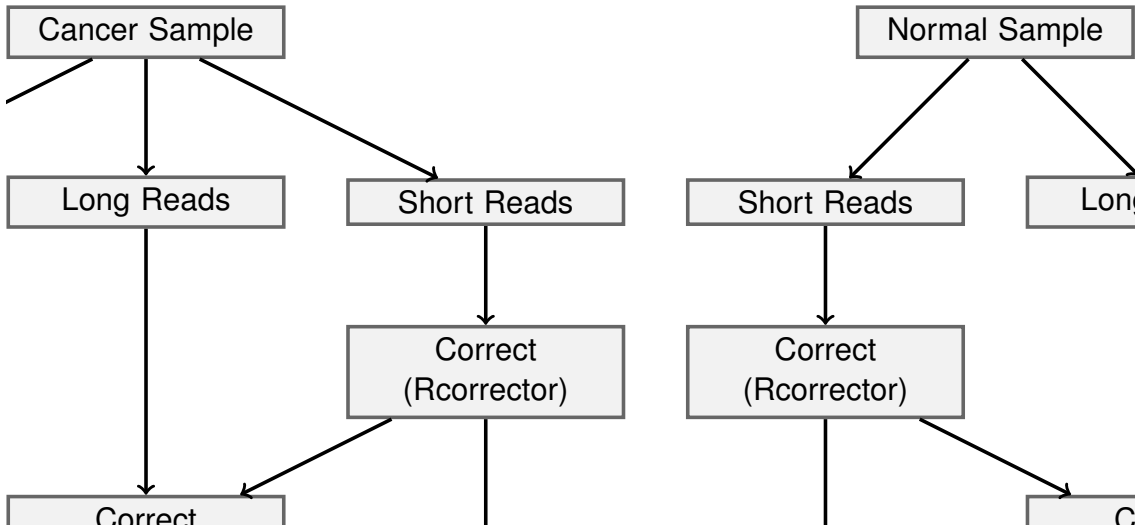
Pipeline Validation

- Genotype and ATAC-seq independent intestinal cell lines.
- Mix the cell lines, then run the pipeline to evaluate effectiveness.



Short Read Correction - Rcorrector

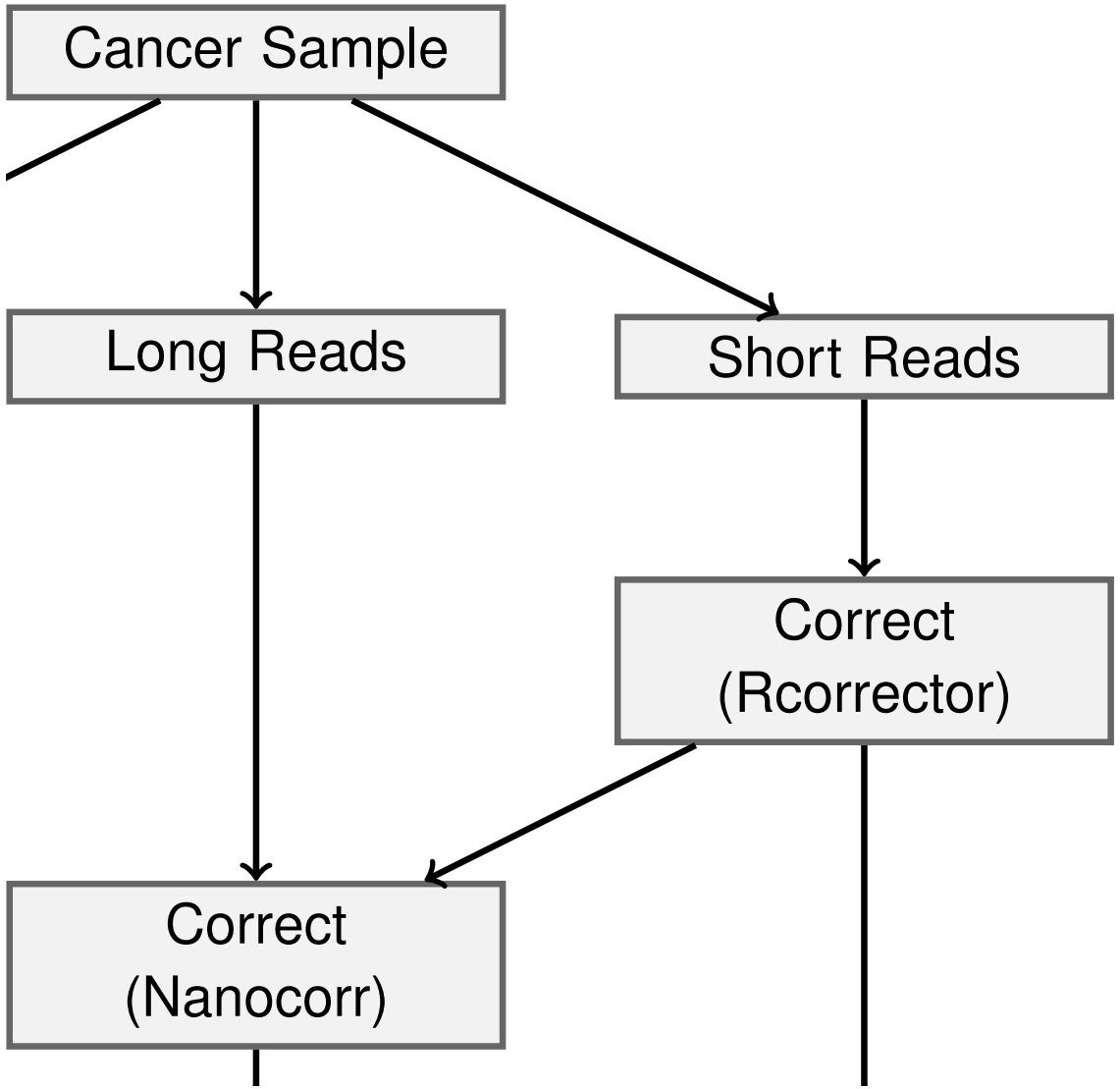
- Fix sequencing errors to reduce false positive mutation calls
- Do this by changing low-abundance kmers to high-abundance kmers



[1] Song & Florea 2015

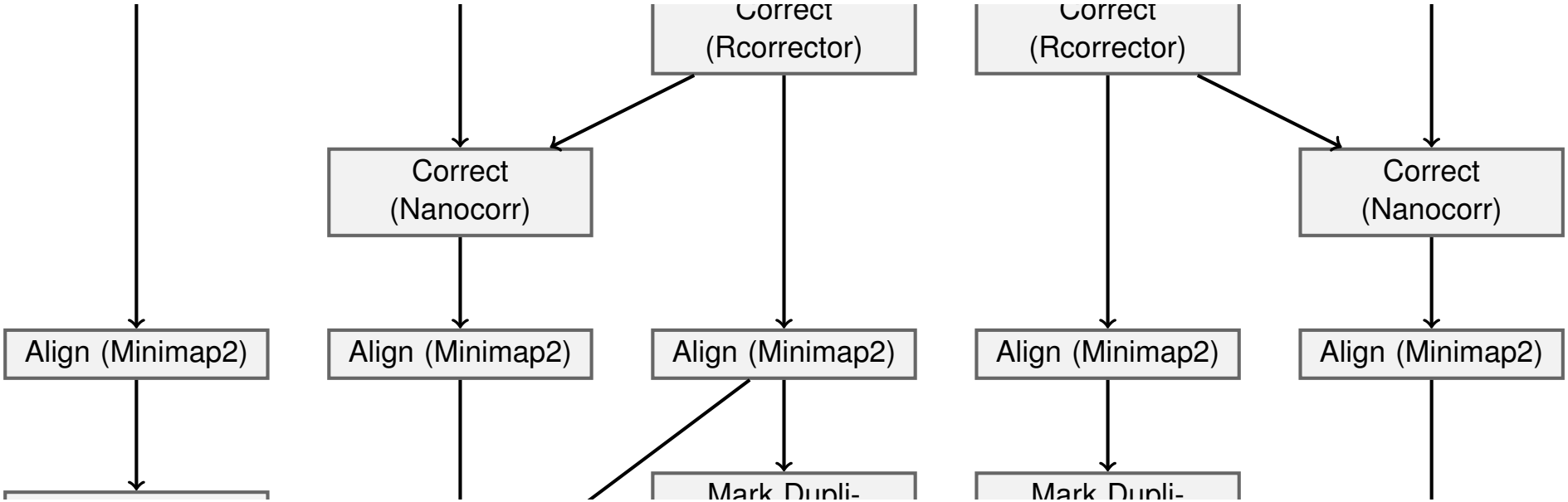
Long Read Correction - Nanocorr

- BLAST each short read to long reads
- Fix long reads using aligned short reads



[1] Goodwin *et al.* 2015

Alignment - Minimap2



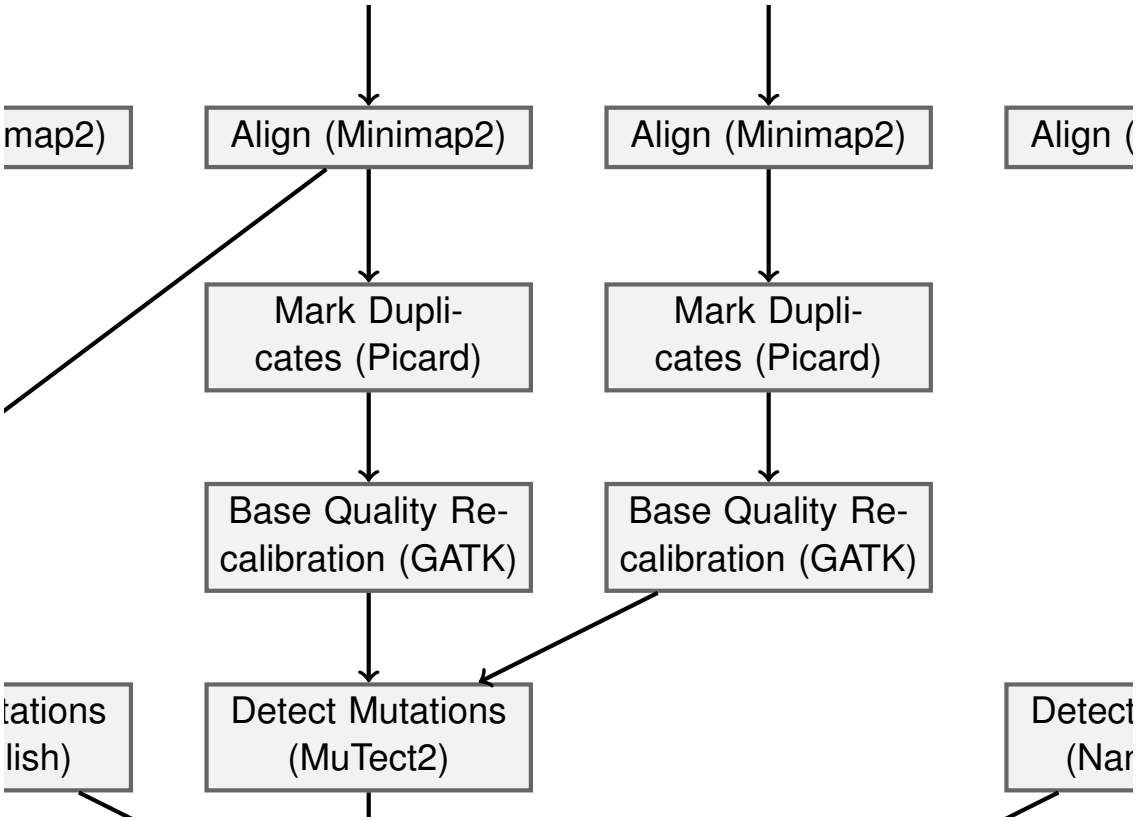
- Split the reference into windows, and for a fixed kmer size, rank the kmers and pick the smallest.
- This is the *minimizer*
- The *minimizer* of a query is used to find the a match in a large reference database, and extended

[1] Li 2017

Detect Mutations in Short Reads - MuTect2

- Ignore sites that are similar to control sample
- Calculate likelihood of a mutation for each non-reference base exists at frequency f, $L(M_f^m)$
- Determine whether $\log_{10} \frac{L(M_f^m)}{L(M_0)}$ exceeds a threshold

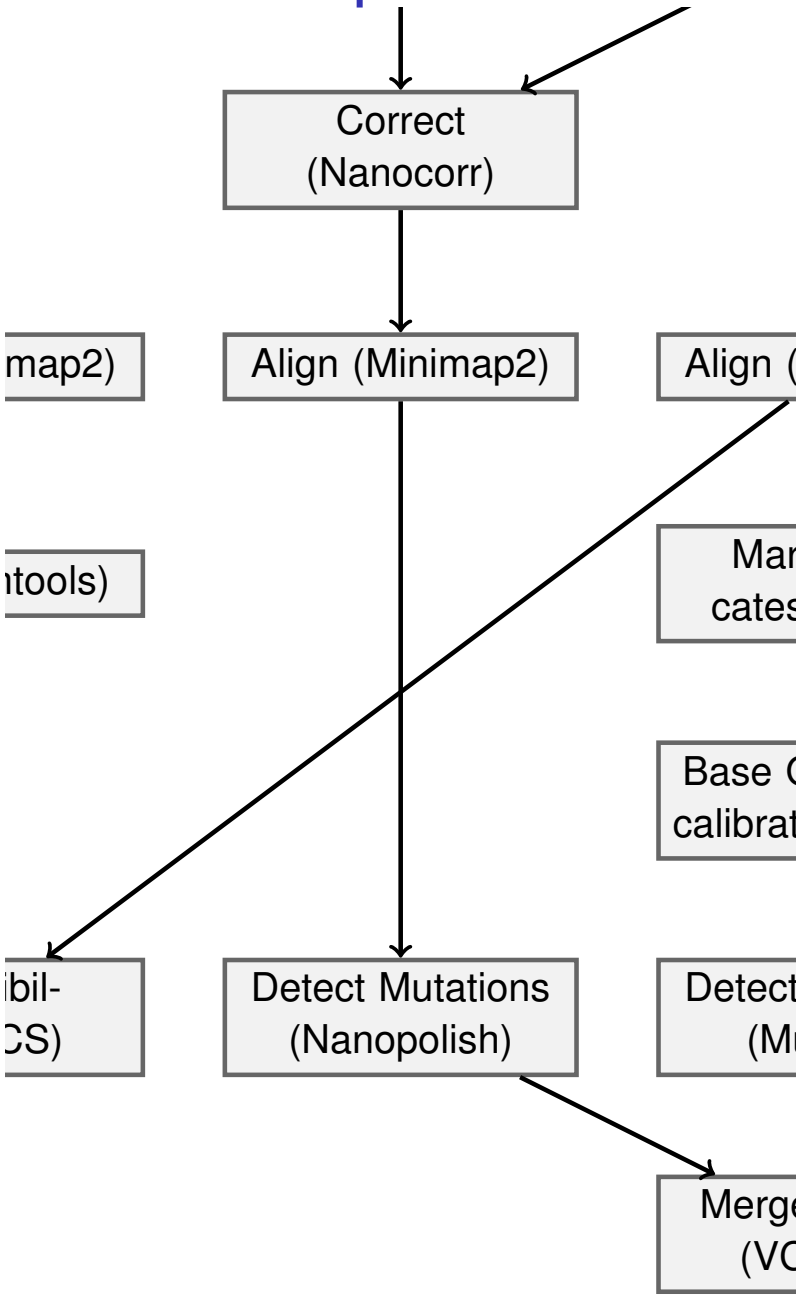
p



[1] Cibulskis *et al.* 2013

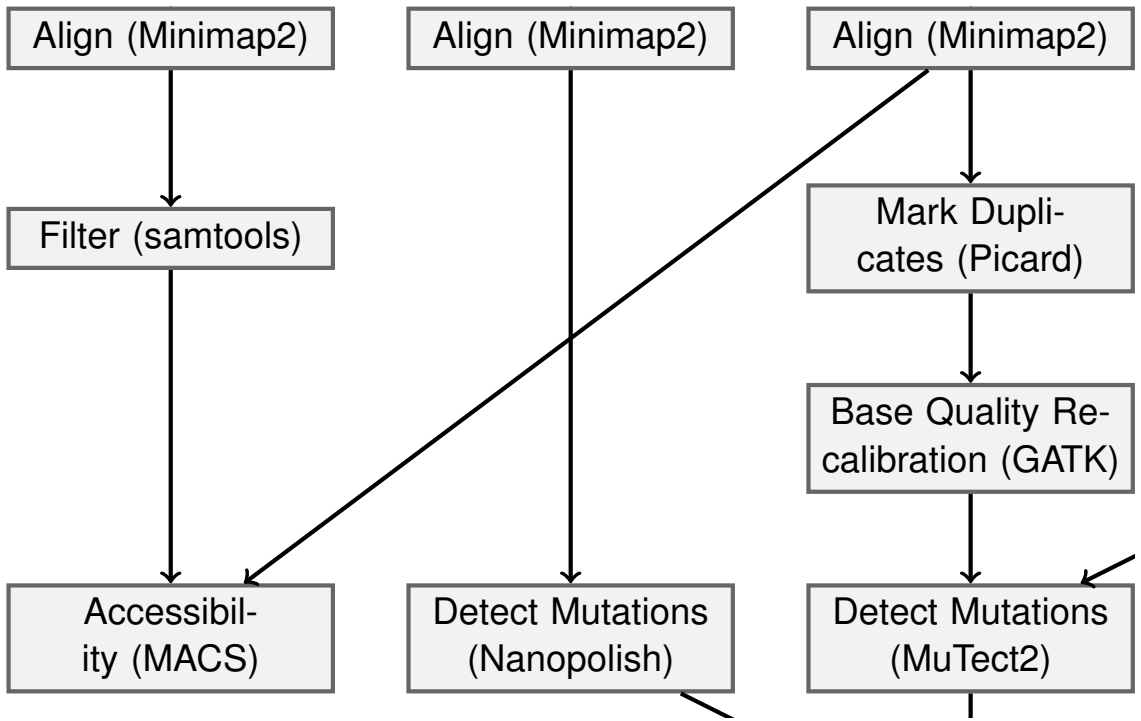
Detect Mutations in Long Reads - Nanopolish

- Use electrical signals and a Hidden Markov Model to detect variants



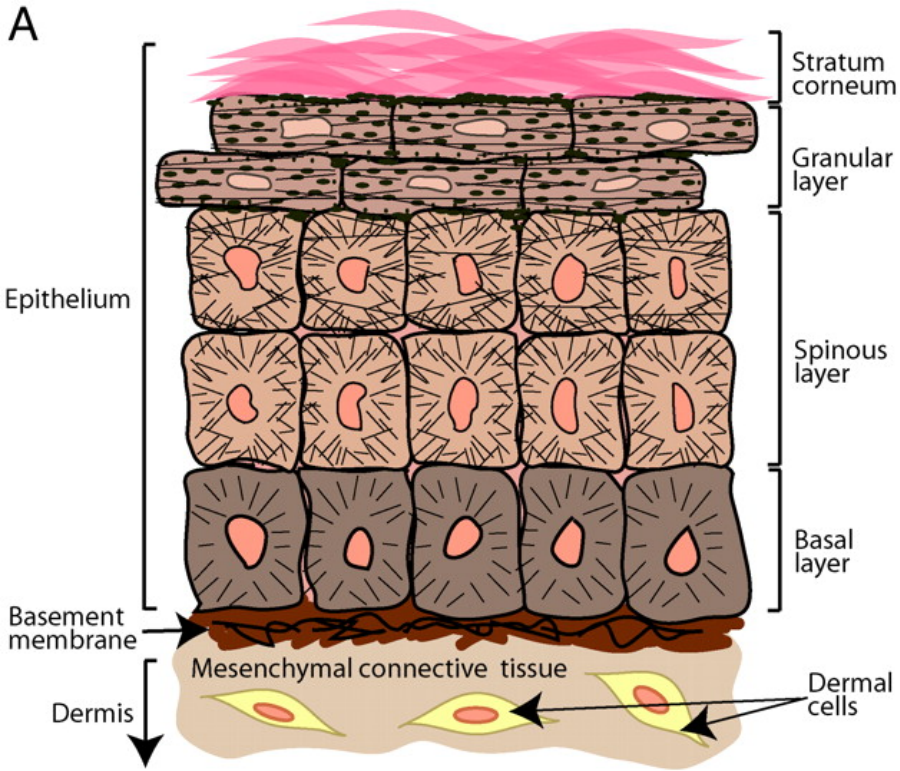
Calculate Chromatin Accessibility - MACS2

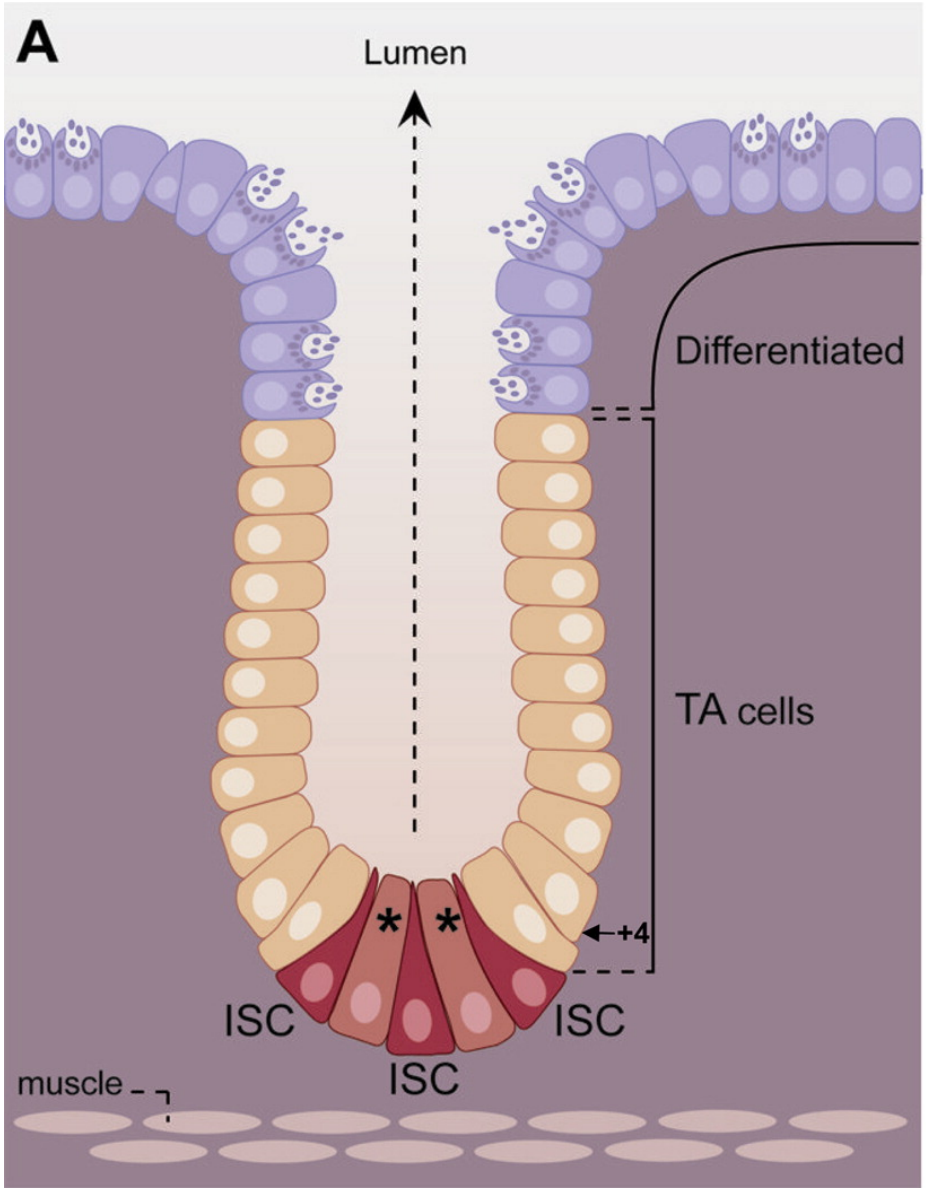
- In a sliding window count:
  - ATAC reads
  - control sequencing reads
- A Poisson rate ratio test is performed to see if the count of ATAC reads non-randomly exceeds the count of control reads.



[1] Zhang *et al.* 2008

Aim 2 - Tissue Collection





[1] Alonso & Fuchs 2003

[2] Casali & Batlle 2009

Statistical Analysis

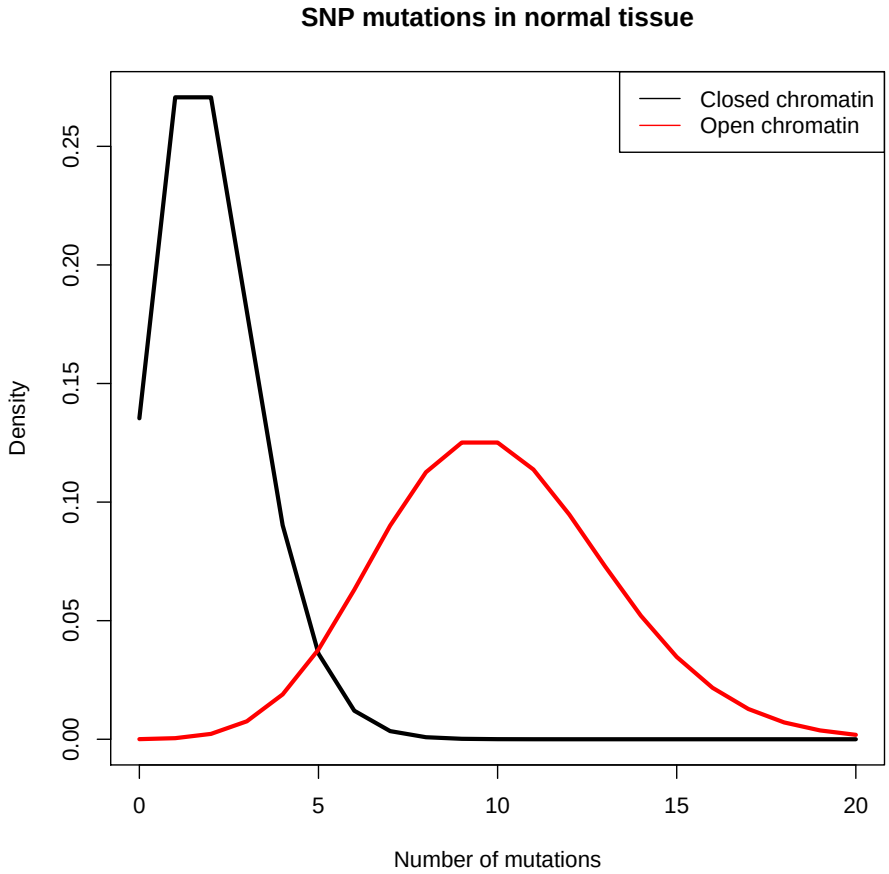
Interesting questions to test for an association and model it for SNPs, insertions, and deletions:

1. Is the mutation rate different in "open" and "closed" chromatin regions?
2. Is the distribution of accessibility enrichment by ATAC-seq different in a cancer-afflicted sample?
3. Does chromatin accessibility predict mutation probability in a logistic regression?

Expected Results - Mutation Rates in Open and Closed Chromatin

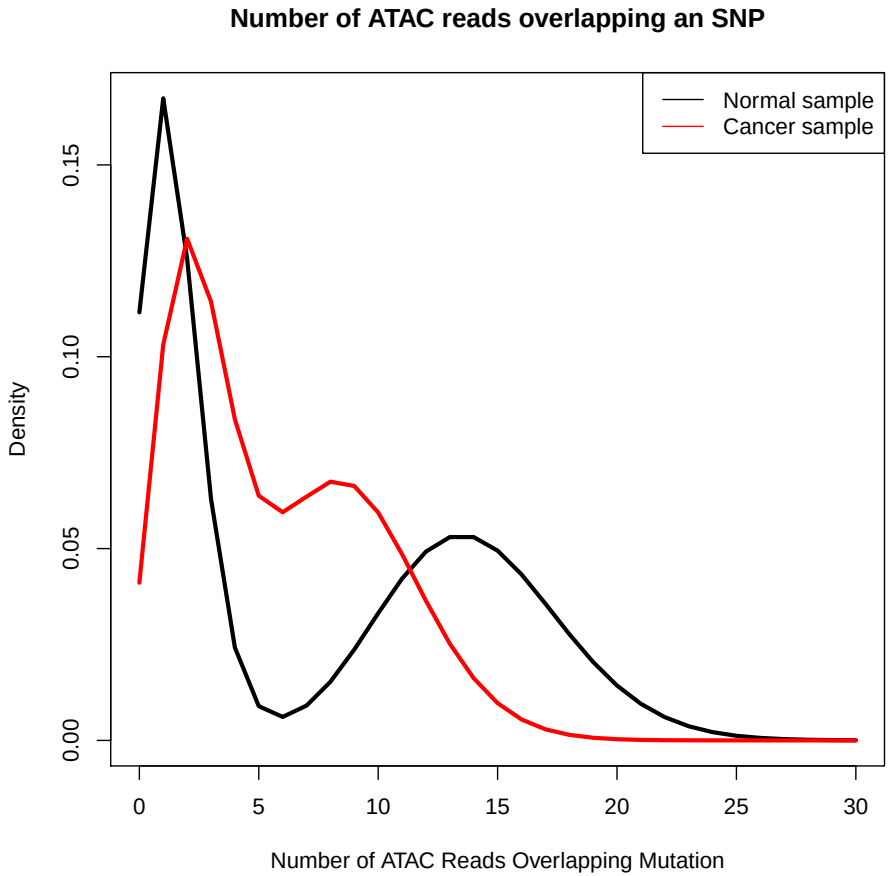
A Poisson rate test can be used to determine if the ratio of the rate parameters is 1.

The quantity $P(X_1|X_1 + X_2 = k)$ is Binomially distributed with k trials and p parameter $\frac{\frac{n_1}{n_2}}{1+\frac{n_1}{n_2}}$ when the rate ratio is 1.



Expected Results - Distribution of Chromatin State at Mutated Sites

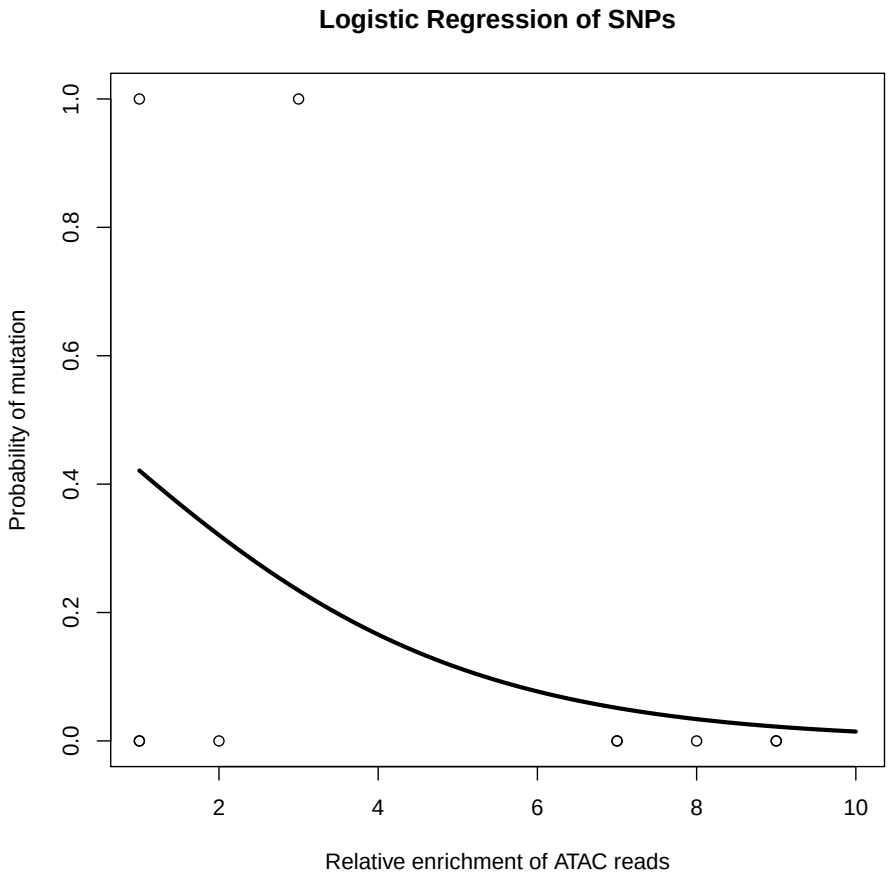
A 2-sample Anderson-Darling test can evaluate whether two samples were sampled from the same distribution.



Expected Results - Logistic Regression

A logistic regression regresses a binary response variable against a continuous one

An analysis of deviance can then be used to determine if the effect is significant.



Conclusion

- The pipeline presented here enables analysis of mutation and chromatin accessibility from a single somatic sample
- Genotyping and ATAC-seq of single somatic tissue will help elucidate the role of somatic mutation in cancer incidence

References I


Alonso, L. & Fuchs, E. Stem cells of the skin epithelium. *Proceedings of the National Academy of Sciences* **100,** 11830–11835 (Aug. 11, 2003).


Baker, S. G. A Cancer Theory Kerfuffle Can Lead to New Lines of Research. *JNCI: Journal of the National Cancer Institute* **107.** ISSN: 0027-8874. doi:10.1093/jnci/dju405. https://academic.oup.com/jnci/article/107/2/dju405/900483/A-Cancer-Theory-Kerfuffle-Can-Lead-to-New-Lines-of (2017) (Feb. 1, 2015).


Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature Methods* **10,** 1213–1218. ISSN: 1548-7091 (Dec. 2013).

References II


Casali, A. & Batlle, E. Intestinal Stem Cells in Mammals and Drosophila. *Cell Stem Cell* **4,** 124–127. ISSN: 1934-5909 (Feb. 6, 2009).


Cibulskis, K. *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature Biotechnology* **31,** 213–219. ISSN: 1087-0156 (Mar. 2013).


Goodwin, S. *et al.* Oxford Nanopore sequencing, hybrid error correction, and de novo assembly of a eukaryotic genome. *Genome Research* **25,** 1750–1756. ISSN: 1088-9051, 1549-5469 (Nov. 1, 2015).


Hanahan, D. & Weinberg, R. A. Hallmarks of Cancer: The Next Generation. *Cell* **144,** 646–674. ISSN: 0092-8674, 1097-4172 (Mar. 4, 2011).

References III


Holohan, C., Van Schaeybroeck, S., Longley, D. B. & Johnston, P. G. Cancer drug resistance: an evolving paradigm. *Nature Reviews Cancer* **13,** 714–726. ISSN: 1474-175X (Oct. 2013).


Junttila, M. R. & de Sauvage, F. J. Influence of tumour micro-environment heterogeneity on therapeutic response. *Nature* **501,** 346–354. ISSN: 0028-0836 (Sept. 19, 2013).


Li, H. Minimap2: fast pairwise alignment for long nucleotide sequences. *arXiv:1708.01492 [q-bio].* arXiv: 1708.01492. http://arxiv.org/abs/1708.01492 (2017) (Aug. 4, 2017).


Loman, N. J., Quick, J. & Simpson, J. T. A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nature Methods* **12,** 733–735. ISSN: 1548-7091 (Aug. 2015).


Makova, K. D. & Hardison, R. C. The effects of chromatin organization on variation in mutation rates in the genome. *Nature Reviews Genetics* **16,** 213–223. ISSN: 1471-0056 (Apr. 2015).

References IV


Polak, P. *et al.* Cell-of-origin chromatin organization shapes the mutational landscape of cancer. *Nature* **518,** 360–364. ISSN: 0028-0836 (Feb. 19, 2015).


Reznikoff, W. S. Tn5 as a model for understanding DNA transposition. *Molecular Microbiology* **47,** 1199–1206. ISSN: 1365-2958 (Mar. 1, 2003).


Song, L. & Florea, L. Rcorrector: efficient and accurate error correction for Illumina RNA-seq reads. *GigaScience* **4,** 48. ISSN: 2047-217X (Oct. 19, 2015).


Song, L. & Crawford, G. E. DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. *Cold Spring Harbor protocols* **2010,** pdb.prot5384. ISSN: 1940-3402 (Feb. 2010).


Stratton, M. R., Campbell, P. J. & Futreal, P. A. The cancer genome. *Nature* **458,** 719–724. ISSN: 0028-0836 (Apr. 9, 2009).

References V

Tomasetti, C. & Vogelstein, B. Variation in cancer risk among tissues can be explained by the number of stem cell divisions. *Science* **347,** 78–81. ISSN: 0036-8075, 1095-9203 (Jan. 2, 2015).

Zhang, Y. *et al.* Model-based Analysis of ChIP-Seq (MACS). *Genome Biology* **9,** R137. ISSN: 1474-760X (Sept. 17, 2008).