CSE280A Class Projects

Please find included projects for the class. Some problems require knowledge of algorithms, but not as much biology. Others require a better understanding of the underlying biology but are methodologically easier. While these are ‘research projects’, the goal is not to necessarily do publication quality work, but rather to get started, and develop some insight into the problem. If taken in the right spirit, this should be the most fun part of the class. I will update this document in the first 3 weeks so please check the website often.

1. You can work on the projects in teams of 2 or 3.
2. You can use this program to satisfy MS exam requirements (CSE MS) students.
3. You can do some of these projects in conjunction with a rotation with the instructor (BISB).

1 Detecting Breakage Fusion Bridge Cycles from Genomic data

Introduction. *This project requires minimal biology, but a good understanding of graph algorithms.* The Breakage-Fusion-Bridge cycle was proposed by Barabara McClintock in 1940 as a mechanism for genomic variation, based on painstaking observations in maize. When first proposed,it was seen as an oddity, but it has become one of the central mechanisms for cancer pathogenicity. With a few BFB cycles, a cancer chromosome can dramatically increase the copy numbers of oncogenes which in turn can increase proliferative properties of the cell. We start by describing an abstract model of the BFB cycle.

1. Start with a string $x = 1, 2, \ldots, n$, representing chromosomal 'segments' from the centromere to just before the telomere.
2. In the following, strings $x^0 = x, x^1, \ldots$ will represent abstractions of the genome after BFB cycle. Let suf(y) describe a suffix of string y, and $-y$ describes a reversal of string y with the sign inverted on each symbol. Then, for all t,
$$x^t = x^{t-1} \cdot -\text{suf}(x^{t-1})$$
where the suffix is chosen arbitrarily (excluding the entire string), and '$\cdot$' implies concatenation. For example, the following is possible:
$$\begin{aligned} x^0 &= 1, 2, 3, 4, 5 \\ x^1 &= 1, 2, 3, 4, 5, -5, -4 \\ x^2 &= 1, 2, 3, 4, 5, -5, -4, 4, 5, -5, -4, -3, -2 \end{aligned}$$

Each of these numbers represents a large genomic segment containing entire genes, so for example, if a gene lay in segment 4, it would have 4 copies in 2 BFB cycles. Unfortunately, we cannot assemble chromosomes to get the architecture perfectly. Instead, we only have the number of copies of each segment. as well as 'fold-back' reads, for example, Illumina reads that have a $\langle +, + \rangle$, or $\langle -, - \rangle$ orientation.

Input: Chromosomal arm oriented from centromere to telomere with copy numbers of different segments. This input has the format: ID, begin-coordinate, end-coordinate, copy-number, where copy-number=2 is the default. A second input is a collection of fold-back reads with the format: coordinates, orientation, where the orientation $\in \{+, -\}$

Output. (a) a YES/NO describing if BFB explains the copy numbers and fold-back reads; (b) a sequence of BFB cycles that explain most of the (revised) copy number and the fold-back reads; and, (c) an error describing the weighted difference between the observed copy number and fold-backs versus the BFB suggested copy number and fold-backs.

Steps. The proposed steps can serve as a starting point.

1. Read some of the algorithmic descriptions for detection of BFB[10,24].
2. Download and run the bfb tool from bitbucket. What are the limitations of this tool?
3. Working with the instructor, download and understand examples of Amplicon Architect file output. AA is a tool for elucidation of the fine structure of an amplified region[4]. Extract the desired copy number and fold-back input starting with the AA output.
4. Working with the instructor, identify publications that claim to have found BFB in cancer genomes[5,7,8,14,20], and download some of the data-sets. Also collect AmpliconArchitect detected bfb data from any of the following: `https://www.dropbox.com/s/uyys2ocaj2v21g3/TumorAmpliconsFromMaster.csv?dl=0`
5. Given a proposed BFB string, describe and quantify an 'error' term that is 0 if the BFB is completely concordant with the observed copy numbers and increases with increasing discordance.
6. Develop a procedure that enumerates BFB strings based on and finds one with a low error. Report YES and the BFB string if the error term is below a threshold.

2 Multi-chromosomal Breakage Fusion Bridge

Introduction. *This project requires a deeper understanding of Biology and some knowledge of genomics.* The Breakage-Fusion-Bridge cycle was proposed by Barabara McClintock in 1940 as a mechanism for genomic variation, based on painstaking observations in maize. When first proposed,it was seen as an oddity, but it has become one of the central mechanisms for cancer pathogenicity. With a few BFB cycles, a cancer chromosome can dramatically increase the copy numbers of oncogenes which in turn can increase proliferative properties of the cell. A recent paper[22] provides a detailed investigation of the anaphase bridge. Specifically, it discusses the molecular signature of a multi-chromosomal breakage fusion bridge (BFB) structure (See Figure 2C). Further, the paper posits the occurrence of special rearrangements (TST jumps). To quote "`First, rather than being randomly distributed, breakpoints were tightly clustered into local 1- to 10-kb hotspots (fig. S10A). Second, tracking the connections between rearrangements revealed chains of tandemly arrayed short insertions [median insertion size, 183 base pairs (bp)] (fig. S10B), which we refer to as 'Tandem Short Template' (TST) jumps (Fig. 5).` " The goal of this project is to read the paper carefully to understand the molecular signatures for these events and to describe algorithms for detecting them in cancer genomes.

Input. Amplicon Architect[4] derived graphs of amplicon structures.
Output. Evidence for multi-chromosomal BFB event and/or a TST jump.

1. Read the Umbreit[22] paper and summarize the Figures with particular attention to Figures 2,5.
2. Devise a strategy for detecting multi-chromosomal BFB.
3. Devise a strategy for detecting TST jumps.
4. Discuss with the instructor and make a plan for analyzing cancer data with BFB events.

3 Genome skimming: Build a partially ordered variant matrix using genome skimming data.

Introduction. *This project requires a working (superficial) knowledge of string algorithms and genome assembly, as well as the ability to write scripts.* Anthropogenic pressure and other natural causes have resulted in severe disruption of global ecosystems in recent years, including climate change with extreme weather events, loss of biodiversity[3], and invasion of non-native flora and fauna. The deforestation of rain forests and the degradation of natural habitats is happening faster than efforts to study and understand the impact of these environmental changes. In North America alone, the bird population has declined by over a quarter since 1970[16]. Scientists would all benefit from an ability to quickly and inexpensively analyze the genomic ecology and biodiversity of a species.

The term ‘genome-skimming’ refers to a light (1-2X) light, whole genome shotgun sequencing of an organism. In this project, we exploit cheap genome-skimming data by building a variant matrix, which provides a foundation for future population genetics work.

Input. genome-skims for 2X data from n individuals of a species.

Output. Columns of a variant matrix: rows correspond to individuals, variants correspond to columns, and each entry is a genotype codes by $\{0, 1, 2\}$. In addition, output a graph on the variants that provides linking information.

Steps.

1. Read some background papers to understand the idea of genome-skimming[2,18]
2. Install and test *de novo* assembly software (e.g. Velvet, SPades[1]) that allows for identification of contigs and ‘bulges’. A bulge corresponds a path that diverges and comes back together in a contig and is suggestive of a genetic variant.
3. Start with a test genome, picking an insect of length < $500 Mpb. Simulate 2n copies of a single chromosome, using a coalescent simulator.
4. For each of the 2n chromosomes, sample 2X reads using a read simulator such as ART.
5. *Combine* reads from all genomes, assemble them, and identify bulge carrying contigs.
6. For each bulge, generate a column of the variant matrix.
7. Working with the instructor, provide statistics on the ‘coverage’ and power of this method.

4 Estimating the heterozygosity of a single, diploid individual using genome-skims.

Introduction. *This project requires knowledge of statistics and algorithms.*

Anthropogenic pressure and other natural causes have resulted in severe disruption of global ecosystems in recent years, including climate change with extreme weather events, loss of biodiversity[3], and invasion of non-native flora and fauna. The deforestation of rain forests and the degradation of natural habitats is happening faster than efforts to study and understand the impact of these environmental changes. In North America alone, the bird population has declined by over a quarter since 1970[16]. Scientists would all benefit from an ability to quickly and inexpensively analyze the genomic ecology and biodiversity of a species.

The term 'genome-skimming' refers to a light (1-2X) light, whole genome shotgun sequencing of an organism[2]. In a recent paper, we used genome-skims to estimate the length of a genome. In a diploid individual, there are two near-identical (homologous) copies of the genetic variation. Define *heterozygosity* as the fraction of sites that are variant (non-identical) in the two homologous copies. The goal of this project is to devise a method to estimate heterozygosity using genome-skim of an individual.

Input. Genome-skim of a diploid individual.

Output. Heterozygosity (θ).

Steps.

1. This is an open-ended research project, and the instructor is looking for new ideas.
2. Start by reading some Genome-skimming papers on Skmer[18] and the preprint of RESPECT[?] . Download and install Skmer (Contact: Shahabeddin Sarmashghi) to compute the distance $D(G, G')$ between two genome skims.
3. Generate a diploid genome by picking a haploid genome, and using a coalescent simulator to generate the homologous genome.
4. Sample reads from the two genomes using ART to get a genome skim G (Use 2-3X total coverage).
5. Algorithmically, partition reads from G into two skims G_1 and $G_2 = G - G_1$ s.t. $D(G_1, G_2)$ is maximized. Return $D(G_1, G_2)$.

5 Understanding the Genesis of extrachromosomal DNA (ecDNA) in cancer

Introduction. *This project requires some knowledge of genomics, but is not algorithmically challenging.* EcDNA refers to the formation of circular genomic structures (approximately 1M bp in length). See Verhaak[23] for a recent review. EcDNA are frequent in cancer occurring in nearly 20% of all cancers across multiple subtypes, and are possibly the dominant source of oncogene amplification. Understanding and targeting ecDNA is now recognized as one of the 'grand challenges' of cancer research (`https://cancergrandchallenges.org/challenges`) including mechanisms of genesis. Our early work suggests that ecDNA are formed from random breakages in the genome[9]. However, this assertion can be challenged because *specific oncogenes are preferentially amplified in specific cancer sub-types.* In this project, you should examine the breakpoints of cancer amplicons to identify if there are specific motifs that are more prone to breakage. For example, the genome is organized into a local structure marked by Topologically Associated Domains (TADs) and TAD boundaries are candidates for breakage. If the TAD boundaries are cell-type specific, that might explain why certain breakages are more common in certain cell types.

Input. Amplicon Architect constructed amplicon structure of ecDNA.

Output. Correlation of ecDNA breakpoints with other genomic features.

Possible Steps.

1. Skim through the amplicon architect publication[4] and study the amplicon graph output by AA. The instructor will provide AA graphs of ecDNA from cancer-cell lines or from pancancer primary tumor data[9].
2. Discard breakpoints that are smaller than 10Kbp in the reference. Specifically, discard 'everted' breakpoints less than 10kbp.
3. Partition the breakpoints into 'tight' (resolved to bp) and 'loose' (resolved to Kbp).
4. Is the location of breakpoints random? Specifically, if you partition the genomes into bins of size (N=10Kbp), is the distribution of number of breakpoints in a bin similar to a Poisson distribution?
5. Study Topologically Associated Domain boundaries (e.g. `http://3dgenome.fsm.northwestern.edu/publications.html`), and accessible genome regions in different subtypes. Are TAD boundaries and/or accessible regions fixed or variable in cancer subtypes?
6. Can we associate breakpoints with tissue specific TAD boundaries boundaries?

6 Understanding the maintenance of extrachromosomal DNA (ecDNA) in cancer

Introduction. *This project requires some knowledge of genomics, but is not algorithmically challenging. It is ideal for BISB/BMS students who are more interested in discovery and not tool building.* EcDNA refers to the formation of circular genomic structures (approximately 1M bp in length). See Verhaak[23] for a recent review. EcDNA are frequent in cancer occurring in nearly 20% of all cancers across multiple subtypes, and are possibly the dominant source of oncogene amplification. Understanding and targeting ecDNA is now recognized as one of the 'grand challenges' of cancer research (`https://cancergrandchallenges.org/challenges`). The challenge specifically mentions the role of proteins involved in maintaining ecDNA as those could be potential targets. The goal of this project is to identify the genes and pathways that are deferentially expressed in samples that carry ecDNA versus samples that do not carry ecDNA.

Input. A listing of samples from the TCGA project with their ecDNA positive or negative status. Normalized gene expression values can be downloaded from cBioportal.

Output. Identification of deferentially expressed genes and pathways and the role of various pathways in maintenance of ecDNA.

Steps:

1. Obtain a list of sample classifications from the instructor from a recent publication[9]. These samples are from the Cancer Genome atlas and each sample is classified as containing an ecDNA or not.
2. Download normalized gene expression, mutational, Copy number, methylation, and other OMICS data from cBioportal and other TCGA resources.
3. Test for significant differences of gene expression/methylation/CN against sample ecDNA status.
4. Test if gene fusions are correlated with ecDNA. You can use fusion data from Gao[6]. Although technically not connected to maintenance of ecDNA, the correlation will be of great interest.

7 Demography

The genome sequence of a person is like a fossil record, and can be read to understand the past history of the population. The goal of this project is to identify the historical changes in population sizes.

Input. The genome of an individual, or of a population given as a SNP matrix. Note that when there is one individual, we only look at heterozygous sites.

Output. Identify time epochs $\tau_1, \tau_2, \ldots, \tau_k$ measured in units of 2N generations, such that the population size *remains constant in each epoch.* Here, N is the current population size and is not known. Second, you must output population size ratios $\lambda_i = \frac{N_i}{N}$.

Goals. This is a difficult project. The expectation is not for you to come up with a novel algorithm. Instead:

1. Develop a simulation framework so you can generate a lot of simulation data.
2. Read and summarize some of the key papers, for example:Li and Durbin[11], Liu[13], Sheehan[19]. You must present a clear description of the algorithm/method used. When describing a learning method, provide an idea of the encoding, and the reasons for choosing a specific method.
3. For one of the methods, example PSMC, show the results of running the code on your simulated data.
4. Devise an implement an algorithm for the simplest case. For example, can you distinguish between 2 situations: one in which $k = 1, \lambda_1 = 1$, and the other in which $k = 2, \lambda_1 = 1, \lambda_2 = 0.5$? You will be scored based on the ideas presented, and negative results are OK as long as you tried different strategies.

8 Deep Dive into recent publications

Recent papers have started investigating DNA from ancient human individuals. If you choose this project, pick *one* of the following papers, explain the main conclusions. Specifically, for each of the main figures based upon computational work, describe in detail how the figure was generated. To understand the methods, you may need to consult the original reference for the tool. In addition, test the validity of some of the tools used by simulating appropriate data sets and applying the same tools.

1. Reconstructing Prehistoric African Population Structure[21]
2. The genomic landscape of Neanderthal ancestry in present-day humans[17]
3. Genomic analysis of Andamanese provides insights into ancient human migration into Asia and adaptation[15]
4. Ancient West African foragers in the context of African population history[12].
5. Mechanisms generating cancer genome complexity from a single cell division error[22].

9 Haplotype assembly from sequencing technologies with high error rate

Introduction. *This project requires good knowledge of algorithms and programming.*

The haplotype assembly problem - computational reconstruction of haplotypes in diploid genomes using sequence reads aligned to a reference - has been studied for almost two decades. HapCUT and other algorithms for haplotype assembly aim to reconstruct haplotypes that minimize or maximize an objective function that accounts for sequencing errors. DNA sequence datasets used for haplotype assembly have many additional types of errors such as false variants. This is particularly true for long read sequencing technologies such as Oxford Nanopore that have a high per-base error rate. HapCUT can tolerate a low frequency of false variants but its accuracy degrades as the fraction of false variants in the input data increases.

The goal of this project is to assess the impact of false variants on haplotype assembly and devise a method to enable accurate haplotype assembly in the presence of a high rate of false variants. For a pair of heterozygous variants, one can compute a statistic or score that captures the likelihood of observing the sequence data if the two variants are correlated (consistent with two haplotypes) versus un-correlated. This score should be positive for pairs of real variants and negative if one of the variants is false. Using this metric and graph algorithms, it should be feasible to identify the vast majority of false variants prior to haplotype assembly.

Steps:

1. Read the HapCUT and HapCUT2 papers (https://genome.cshlp.org/content/27/5/801)
2. Obtain a dataset for haplotype assembly for a human genome (the instructor can provide this)
3. Add false variants to the dataset and assess the impact on haplotyping accuracy
4. Develop a method to identify false variants before haplotype assembly using correlation metrics and graph algorithms

References

[1] http://cab.spbu.ru/software/spades/.

[2] K. Bohmann, S. Mirarab, V. Bafna, and M. T. P. Gilbert. Beyond DNA barcoding: The unrealized potential of genome skim data in sample identification. *Mol Ecol*, 29(14):2521–2534, 07 2020.

[3] ES Brondizio, J Settele, S Diaz, and HT Ngo. Global assessment report on biodiversity and ecosystem services of the intergovernmental science-policy platform on biodiversity and ecosystem services. *IPBES Secretariat, Bonn*, 2019.

[4] V. Deshpande, J. Luebeck, N. D. Nguyen, M. Bakhtiari, K. M. Turner, R. Schwab, H. Carter, P. S. Mischel, and V. Bafna. Exploring the landscape of focal amplifications in cancer using AmpliconArchitect. *Nat Commun*, 10(1):392, 01 2019.

[5] Anthony Ferrari, Anne Vincent-Salomon, Xavier Pivot, Anne Sophie Sertier, Emilie Thomas, Laurie Tonon, Sandrine Boyault, Eskeatnaf Mulugeta, Isabelle Treilleux, Gaëtan MacGrogan, Laurent Arnould, Janice Kielbassa, Vincent Le Texier, Hélène Blanché, Jean François Deleuze, Jocelyne Jacquemier, Marie Christine Mathieu, Frédérique Penault-Llorca, Frédéric Bibeau, Odette Mariani, Cécile Mannina, Jean Yves Pierga, Olivier Trédan, Thomas Bachelot, Hervé Bonnefoi, Gilles Romieu, Pierre Fumoleau, Suzette Delaloge, Maria Rios, Jean Marc Ferrero, Carole Tarpin, Catherine Bouteille, Fabien Calvo, Ivo Glynne Gut, Marta Gut, Sancha Martin, Serena Nik-Zainal, Michael R. Stratton, Iris Pauporté, Pierre Saintigny, Daniel Birnbaum, Alain Viari, and Gilles Thomas. A whole-genome sequence and transcriptome perspective on HER2-positive breast cancers. *Nature Communications*, 7, 2016.

[6] Q. Gao, W. W. Liang, S. M. Foltz, G. Mutharasu, R. G. Jayasinghe, et al. Driver Fusions and Their Implications in the Development and Treatment of Human Cancers. *Cell Rep*, 23(1):227–238, 04 2018.

[7] Dale W. Garsed, Owen J. Marshall, Vincent D.A. Corbin, Arthur Hsu, Leon DiStefano, Jan Schröder, Jason Li, Zhi Ping Feng, Bo W. Kim, Mark Kowarsky, Ben Lansdell, Ross Brookwell, Ola Myklebost, Leonardo Meza-Zepeda, Andrew J. Holloway, Florence Pedeutour, K. H.Andy Choo, Michael A. Damore, Andrew J. Deans, Anthony T. Papenfuss, and David M. Thomas. The Architecture and Evolution of Cancer Neochromosomes. *Cancer Cell*, 2014.

[8] David Gisselsson, Louise Pettersson, Mattias Höglund, Markus Heidenblad, Ludmila Gorunova, Joop Wiegant, Fredrik Mertens, Paola Dal Cin, Felix Mitelman, and Nils Mandahl. Chromosomal breakage-fusion-bridge events cause genetic intratumor heterogeneity. *Proceedings of the National Academy of Sciences of the United States of America*, 2000.

[9] H. Kim, N. P. Nguyen, K. Turner, S. Wu, A. D. Gujar, J. Luebeck, J. Liu, V. Deshpande, U. Rajkumar, S. Namburi, S. B. Amin, E. Yi, F. Menghi, J. H. Schulte, A. G. Henssen, H. Y. Chang, C. R. Beck, P. S. Mischel, V. Bafna, and R. G. W. Verhaak. Extrachromosomal DNA is associated with oncogene amplification and poor outcome across multiple cancers. *Nat Genet*, 52(9):891–897, 09 2020.

[10] Marcus Kinsella and Vineet Bafna. Combinatorics of the breakage-fusion-bridge mechanism. *Journal of Computational Biology*, 2012.

[11] H. Li and R. Durbin. Inference of human population history from individual whole-genome sequences. *Nature*, 475(7357):493–496, Jul 2011.

[12] Mark Lipson, Isabelle Ribot, Swapan Mallick, Nadin Rohland, Nicole Adamski, Nasreen Broomandkhoshbacht, Ann Marie Lawson, Jonas Oppenheimer, Kristin Stewardson, Neil Bradman, Brendan J Culleton, Els Cornelissen, Isabelle Crevecoeur, Pierre De Maret, Forka Leypey, Mathew Fomine, Philippe Lavachery, Christophe Mbida Mindzie, Rosine Orban, Elizabeth Sawchuk, Patrick Semal, Mark G Thomas, Wim Van Neer, Krishna R Veeramah, Douglas J Kennett, Nick Patterson, Garrett

Hellenthal, Carles Lalueza-fox, Scott Maceachern, Mary E Prendergast, and David Reich. Ancient West African foragers in the context of African population history. (November 2018), 2019.

[13] Xiaoming Liu and Yun-Xin Fu. Exploring population size changes using snp frequency spectra. *Nature genetics*, 47(5):555–559, 05 2015.

[14] Michael Marotta, Taku Onodera, Jeffrey Johnson, G. Thomas Budd, Takaaki Watanabe, Xiaojiang Cui, Armando E. Giuliano, Atsushi Niida, and Hisashi Tanaka. Palindromic amplification of the ERBB2 oncogene in primary HER2-positive breast tumors. *Scientific Reports*, 7(February):1–12, 2017.

[15] Mayukh Mondal, Ferran Casals, Tina Xu, Giovanni M Dall'Olio, Marc Pybus, Mihai G Netea, David Comas, Hafid Laayouni, Qibin Li, Partha P Majumder, and Jaume Bertranpetit. Genomic analysis of andamanese provides insights into ancient human migration into asia and adaptation. *Nature Genetics*, 48:1066 EP –, 07 2016.

[16] Kenneth V. Rosenberg, Adriaan M. Dokter, Peter J. Blancher, John R. Sauer, Adam C. Smith, Paul A. Smith, Jessica C. Stanton, Arvind Panjabi, Laura Helft, Michael Parr, and Peter P. Marra. Decline of the North American avifauna. *Science*, page eaaw1313, sep 2019.

[17] S. Sankararaman, S. Mallick, M. Dannemann, K. Prufer, J. Kelso, S. Paabo, N. Patterson, and D. Reich. The genomic landscape of Neanderthal ancestry in present-day humans. *Nature*, Jan 2014.

[18] S. Sarmashghi, K. Bohmann, M. T. P Gilbert, V. Bafna, and S. Mirarab. Skmer: assembly-free and alignment-free sample identification using genome skims. *Genome Biol*, 20(1):34, 02 2019.

[19] Sara Sheehan and Yun S. Song. Deep learning for population genetic inference. *PLOS Computational Biology*, 12(3):1–28, 03 2016.

[20] Noriaki Shimizu, Kenta Shingaki, Yukiko Kaneko-Sasaguri, Toshihiko Hashizume, and Teru Kanda. When, where and how the bridge breaks: Anaphase bridge breakage plays a crucial role in gene amplification and HSR generation. *Experimental Cell Research*, 2005.

[21] Pontus Skoglund, Jessica C. Thompson, Mary E. Prendergast, Alissa Mittnik, Kendra Sirak, Mateja Hajdinjak, Tasneem Salie, Nadin Rohland, Swapan Mallick, Alexander Peltzer, Anja Heinze, Iñigo Olalde, Matthew Ferry, Eadaoin Harney, Megan Michel, Kristin Stewardson, Jessica I. Cerezo-Román, Chrissy Chiumia, Alison Crowther, Elizabeth Gomani-Chindebvu, Agness O. Gidna, Katherine M. Grillo, I. Taneli Helenius, Garrett Hellenthal, Richard Helm, Mark Horton, Saioa López, Audax Z. P. Mabulla, John Parkington, Ceri Shipton, Mark G. Thomas, Ruth Tibesasa, Menno Welling, Vanessa M. Hayes, Douglas J. Kennett, Raj Ramesar, Matthias Meyer, Svante Pääbo, Nick Patterson, Alan G. Morris, Nicole Boivin, Ron Pinhasi, Johannes Krause, and David Reich. Reconstructing prehistoric african population structure. *Cell*, 171(1):59–71.e21, 2018/02/10.

[22] N. T. Umbreit, C. Z. Zhang, L. D. Lynch, L. J. Blaine, A. M. Cheng, R. Tourdot, L. Sun, H. F. Almubarak, K. Judge, T. J. Mitchell, A. Spektor, and D. Pellman. Mechanisms generating cancer genome complexity from a single cell division error. *Science*, 368(6488), 04 2020.

[23] R. G. W. Verhaak, V. Bafna, and P. S. Mischel. Extrachromosomal oncogene amplification in tumour pathogenesis and evolution. *Nat Rev Cancer*, 19(5):283–288, 05 2019.

[24] Shay Zakov, Marcus Kinsella, and Vineet Bafna. An algorithmic approach for breakage-fusion-bridge detection in tumor genomes. *Proceedings of the National Academy of Sciences of the United States of America*, 2013.