My research interests lie at the intersection of computing, machine learning, information theory and biology; and have developed from two seemingly divergent approaches:

1. Apply the tools and ideologies I honed during my academic training to solving mathematically-driven problems.
2. Discover **laws in biology** similar to what scientists were able to obtain in physics.

In this second direction, I am less focused on using familiar tools, and instead, use or develop whatever techniques are most-suited to not only answer the immediate question, but also to understand the underlying process. One of the key reasons biology has been difficult to codify is due to the technology bottleneck. For example, to obtain a data point which is as profound as "apple falling from a tree", in biology requires a huge amount of infrastructural and technological development. However, in the realm of DNA, we are now reaching a point, where we can obtain these "profound data points." Two key questions I targeted are:

- Can we infer the evolutionary history of DNA from its current snapshot?
- Can DNA evolution be characterized by a finite number of parameters?

To answer these questions, I have derived mathematical limits on fundamental aspects of evolutionary mutations like *diversity generation* [22, 19, 21], *evolutionary distance* [1] and *uniqueness of ancestory* [20, 19] using theory of computation and information theory tools. The insights gained from these theoretical contributions led me to develop computational techniques that find applications in *cancer classification from healthy DNA* [24, 23], characterizing *viral evolution* [26, 27] and *DNA storage* [20, 18].

While working with datasets, I came to appreciate just how easily sampling bias and noisy data can lead to false discoveries [25]. While these issues are particularly perilous in medical applications, they plague virtually every domain. As the world begins to rely more heavily on data science and ML models, we require a good understanding of their limitations since decisions made by them have important societal, ethical, political and health implications. Thus, we need to design robust techniques and be rigorous in our approach as the noise, biases, adversaries within data that can lead to false discoveries and unfair outcomes [2].

Moreover data can be limited and/or building a model from scratch can be computationally inefficient. In this situation, we can rely on using previously built models. More specifically, we can design "soft" integrated circuits where already built data-driven models can serve as "modules" and be used as components to solve a more complex prediction and decision problem. Here, a proper understanding of the limitations of combining these modules is necessary to make efficient use of them for different applications. These directions have motivated me to pursue the following questions:

- Can we define a rigorous metric to assess data quality? [41]
- Are there intrinsic ways to correct for sampling bias in machine learning? [37]
- Can new expertise be synthesized from known expertise? [36]

Below, I provide more details on my previous work and my future research goals.

1 Data Science/ML

Robust Correction of Sampling Bias: Cancer experts may wish to adapt their cancer classifier trained using the genetic data of American patients for patients in Europe. This will lead to distribution shift in the test domain, a problem popularly known as covariate shift [46]. We propose a new method using the recently proposed learning theory framework by Vapnik and Izmailov [54] to design robust classifiers that deal with the covariate shift problem. Unlike other state of the art methods [16, 46, 50, 57], our method doesn't require any parameter tuning. We also gave theoretical guarantees on our method's performance and experimentally showed much more stable and consistent performance over other methods on benchmark datasets [37]. In the future, I would like to work on understanding the limitations of the methods in higher dimensions, understand parameter-free covariate shift under the *PAC-learning* framework [53, 40] and develop ideas for the general problem of *domain adaptation* [3].

Expert Graphs: Given an expert physician that can distinguish between lung cancer and COVID-19 and another expert physician that can distinguish between COVID-19 and flu, can we design a new expert that can distinguish between lung cancer and flu? We recently proposed a new framework of expert graphs to analyse this problem [36]. In the context of machine learning, this framework can be used to derive new classifiers. For example, given pairwise classifiers between classes A and B and classes B and C, what can be said about the pairwise classifier between classes A and C? Analysing these questions in the context of experts and ML classifiers leads to an interesting phenomenon of *non-transitivity* finding connections to the Condorcet Paradox in voting theory and non-transitive dice in statistics [42]. In our current work, our framework assumes *perfect* experts, in future work, I am interested in analysing these questions for imperfect experts. For example, in the context of ML classifiers, these imperfections can arise due to sample complexity, model complexity, data quality and memory constraints.

2 Genome Evolution, Cancer and Viruses

The oldest information system was handed to us by nature in the form of life. One of the fundamental units of this biological information system is the genome or DNA. Today, due to the recent advances in DNA sequencing technology [29, 30, 4, 8], we have good quality DNA data and we find ourselves in a unique position to test and verify our hypotheses about evolution–hopefully leading us to find associations, causes, and cures for mutation based diseases like cancer, Alzheimer's, and autoimmune diseases.

Evolution from an information and computation perspective: A popular framework in computational biology views the genome as a *long computer program* where genes serve as *methods or functions* which are called at different times depending on an organism's state. Sometimes these methods can also have *errors* or *bugs* which can lead to dysfunction and can be contributing to a genetic disease. This has led us to identify genes for multiple diseases like cystic fibriosis, Huntington's disease, cancer [49, 13, 39]. However, this approach has its limitations [34] because

1. Evolution is a *dynamical system* and the current approach overlooks the statistical information about the *transitions* by only focusing on the *states.*
2. It ignores the non-functional part of the genome. It may be true that mutations in those areas are of no consequence for functional purposes, however the error pattern there may still have information about the evolutionary dynamics which can not only enhance our understanding of evolution but also help us identify patterns that can point to the risk affinity for catching mutation based diseases.

Why are evolutionary dynamics important?: The phenotypes or traits that we observe in living beings can be divided into two categories - *static* and *dynamic.* Static traits (ex. eye color) don't change for an individual, while diseases like cancer are primarily caused by an accumulation of mutations acquired during an individual's lifetime [6, 11, 5]. For dynamical traits (like cancer) having information about the transitions along with the states is crucial as they encode the *likelihood* of reaching an *absorbing* state (can be a lethal cancer). Hence, studying transitions could help predict the future risk or detect the disease early [6, 56, 23, 24].

Inferring evolutionary dynamics: One natural way to infer evolutionary dynamics is to collect genomic data for an individual and their family at *multiple* time points and use phylogenetic approaches [55] (as was recently seen with SARS-CoV-2 genomic data and other viruses in the past [9]), however we don't have an infrastructure yet to perform this experiment for the human genome as collecting a large dataset of the genomic sequences of individuals every day/month has economical and ethical constraints [44].
A question that we ask is, *Can we can gain information about these transitions* intrinsically *using a single genome snapshot without having access to temporal genomic data?* The genome evolves as cells multiply and the genome is copied to new cells. These copies may have mistakes or mutations. For many areas of the genome, it is hard to make predictions about the evolutionary path that has led to those mutations, as there are multiple equally likely possibilities that could have led to their existence. However, there are certain areas known as *tandem repeat regions* where predictions about the *evolution channel* (the term channel is motivated by the notion of communication channel introduced by Shannon [45] in 1948 which started the area of Information Theory) can be made. These regions have evolved by a sequence of tandem duplications (eg. $AG \rightarrow AGAGAGAG$) due to replication slippage events [33, 43] and point mutations (single changes like substitutions, insertions and deletions in the DNA, e.g. $AC\underline{T}G \rightarrow AC\underline{A}G$). Further, the rate of these mutation events is also higher in these regions [51]. When point mutations and duplication events are viewed together, one can learn the relative rates of these mutational events in these regions and infer information about their evolutionary history.

Cancer signal in healthy cells? We used this insight and worked with more than 5000 DNA samples (around 75 TB data) obtained using The Cancer Genome Atlas (TCGA) [38] on *DNA derived from blood ("healthy DNA")* for people with different cancer types. We quantified the evolutionary history of short tandem repeat regions [52] and then used gradient boosting [10] on these obtained histories. Our analysis found that signals for Glioblastoma (brain cancer) can be decoded by using the evolutionary history of tandem repeat regions in the healthy genome [24].

Theoretical Properties of the Evolution Channel: I also studied theoretical properties of tandem duplication mutations and focused on the analysis and characterization of the evolution channel using measures of *capacity*, *expressiveness* [22], *duplication distance* [1], and *uniqueness of ancestry* [20] and used these insights for the design of error correcting codes for DNA storage [20] and to explain inversion symmetry in the genome [21].

- *Diversity:* I found limits on the diversity of sequences that can be generated by tandem duplications by calculating *exact capacity* values and *fully* answering the *expressiveness* [22] question using regular languages [14] and constrained coding tools [35].
- *Duplication Distance:* I calculated *tight* bounds on the *duplication distance* which is used to measure the timing of generation by these duplications [1] which conveyed that a large duplication distance is needed to achieve

enough diversity *emphasizing* the role of short duplication lengths in evolution. More precisely, we found that almost all length n binary sequences (set of probability 1) required $\Theta(n)$ tandem duplication steps to evolve, even if *unbounded* duplication lengths were allowed which also means that short duplication lengths play a major role in generating capacity [1].

- *Uniqueness of Ancestry:* I also asked questions about the uniqueness of seed for a given sequence and completely characterized the duplication length sets where the seed is unique or non-unique [20, 19] which led me to design *capacity achieving* error correcting codes for any number of tandem duplication errors that are useful for DNA-storage based applications.

Parameterization of Genome generation: We hypothesize that genomic sequences have an underlying generator which can be characterized by a fixed number of parameters. Any genomic sequence that we observe in nature is an instance of this parametric generator. In [26], we made this characterization using state machines. More precisely, we parameterized each genomic sequence using a Markov Chain with a given number of states. From a computational perspective, these markov chains provided us with an efficient way of compressing genomic information that further allowed us to make *computationally efficient statistical comparisons* amongst large genomic datasets. They can also be used to characterize temporal and spatial evolution in a continuous manner, detect local sites with higher mutation activity, as well as to aid alignment techniques for fast and efficient discovery of mutations. As a case study, we demonstrated these advantages on different variants of the SARS-CoV-2 virus. *A variant is defined by deterministic stable mutations, however, evolution is characterized by the randomness of the transient mutations. We demonstrate that these markov chain based representations can be used to quantify the randomness associated with these transient mutations* hence giving us insights into the continuous evolution of a given species, for example SARS-CoV-2 virus [26]. Further, *we can envision a future where we can use these computational techniques to design vaccines as they give us predictive information about regions in viral sequences with high and low mutation activity.*

3 DNA Storage

DNA also provides us with an energy efficient and dense medium to *store information* both ex-vivo, i.e. outside a living organism in some chemical medium [7, 12] as well as in-vivo, i.e. inside a living organism [48, 47]. With the recent progress made in synthesis [31] & sequencing methods [8, 4] and gene editing technologies like CRISPR [15, 28], we can see DNA as a potential medium of data storage for engineering systems in the future.

1. *Live DNA Storage:* In live-DNA storage [47], information is embedded in the DNA of a cell and this cell replicates over time creating several copies of the same information. Due to substitution, indel and duplication mutations, these copies are not the same. One immediate application here is to design error correcting schemes so that the stored information can be uniquely recovered. In one of our works, we designed capacity achieving error correcting codes when errors are of tandem duplication type [20, 19]. Another application would be to use these erroneous copies and the initial information to estimate the evolution channel. Both these applications are also related, as a better characterization of the evolution channel will also result in a more realistic error correction scheme. The mathematical problems here have their fundamental roots in the reconstruction problem introduced by Levenshtein [32] and models used in phylogeny [55]. I would also like to look for potential collaboration where we can *design a wet lab experiment* to collect the data necessary to make the inference about the evolution channel, which can then be incorporated into the channel model for the design of error correcting codes for live DNA storage.
2. *ex-vivo DNA Storage:* DNA storage outside a living cell is a promising concept primarily due to its lower energy cost and high density [12, 7]. The roadblocks faced by this technology however are the *costs* and *error rates* incurred during synthesis (write) and sequencing (read) operations [17, 44]. Recently, an inexpensive but erroneous synthesis method was proposed [31]. We computed capacity and designed high rate coding schemes for the channel incurred by this writing mechanism [18]. As we make progress in synthesis and sequencing technologies, I am interested in designing efficient error correcting codes for channels that arise due to these new innovations.

Conclusion

The idea of decoding and quantifying evolutionary history from a genomic *snapshot* can serve as a *new computational microscope* to analyze the DNA. Moreover, due to advances in long read sequencing technology [4], in the future, we can analyze different *structural variations* in the genome such as long tandem repeats, interspersed repeats. I believe evolution has a structure remaining to be decoded. My long term goal is to discover those abstractions. Having abstract representations of the genome can also help us understand the interaction between genomes of different species. For example in current times, given the amount of research effort that has been put in solving and understanding COVID-19, we still don't have a good understanding of how the SARS-CoV-2 virus is *interacting* with the human genome!

References

[1] N. Alon, J. Bruck, F. Farnoud Hassanzadeh, and S. Jain. "Duplication Distance to the Root for Binary Sequences". In: *IEEE Transactions on Information Theory* 63.12 (Dec. 2017), pp. 7793–7803. ISSN: 0018-9448. DOI: 10.1109/TIT.2017.2730864.

[2] S. Barocas, M. Hardt, and A. Narayanan. *Fairness and Machine Learning*. http://www.fairmlbook.org. fairmlbook.org, 2018.

[3] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan. "A theory of learning from different domains". In: *Machine Learning* 79.1 (May 2010), pp. 151–175. ISSN: 1573-0565. DOI: 10.1007/s10994-009-5152-4. URL: https://doi.org/10.1007/s10994-009-5152-4.

[4] C. Bleidorn. "Third generation sequencing: technology and its potential impact on evolutionary biodiversity research". In: *Systematics and Biodiversity* 14.1 (2016), pp. 1–8. DOI: 10.1080/14772000.2015.1099575. eprint: https://doi.org/10.1080/14772000.2015.1099575. URL: https://doi.org/10.1080/14772000.2015.1099575.

[5] C. Tomasetti, L. Li, and B. Vogelstein. "Stem cell divisions, somatic mutations, cancer etiology, and cancer prevention". In: *Science* 6331.355 (2017), pp. 1330–1334.

[6] Cancer.Net. *The Genetics of Cancer*. URL: https://www.cancer.net/navigating-cancer-care/cancer-basics/genetics/genetics-cancer (visited on 05/11/2018).

[7] G. M. Church, Y. Gao, and S. Kosuri. "Next-Generation Digital Information Storage in DNA". en. In: *Science* 337.6102 (Sept. 2012), pp. 1628–1628. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.1226355.

[8] J. Eberwine, J.-Y. Sul, T. Bartfai, and J. Kim. "The promise of single-cell sequencing". In: *Nature Methods* 11.1 (2014), pp. 25–27. ISSN: 1548-7105. DOI: 10.1038/nmeth.2769. URL: https://doi.org/10.1038/nmeth.2769.

[9] S. Elbe and G. Buckland-Merrett. "Data, disease and diplomacy: GISAID's innovative contribution to global health". In: *Global Challenges* 1.1 (2017), pp. 33–46. DOI: https://doi.org/10.1002/gch2.1018. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/gch2.1018. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/gch2.1018.

[10] J. H. Friedman. "Greedy function approximation: A gradient boosting machine." In: *Ann. Statist.* 29.5 (Oct. 2001), pp. 1189–1232. DOI: 10.1214/aos/1013203451. URL: https://doi.org/10.1214/aos/1013203451.

[11] M. Gerstung et al. "The evolutionary history of 2,658 cancers". In: *Nature* 578.7793 (Feb. 2020), pp. 122–128. ISSN: 1476-4687. DOI: 10.1038/s41586-019-1907-7. URL: https://doi.org/10.1038/s41586-019-1907-7.

[12] R. N. Grass, R. Heckel, M. Puddu, D. Paunescu, and W. J. Stark. "Robust Chemical Preservation of Digital Information on DNA in Silica with Error-Correcting Codes". en. In: *Angewandte Chemie International Edition* 54.8 (Feb. 2015), pp. 2552–2555. ISSN: 1521-3773. DOI: 10.1002/anie.201411378.

[13] J. N. Hirschhorn and M. J. Daly. "Genome-wide association studies for common diseases and complex traits". In: *Nature Reviews Genetics* 6 (2005), pp. 95–108.

[14] J. E. Hopcroft, R. Motwani, and J. D. Ullman. *Introduction to Automata Theory, Languages, and Computation (3rd Edition)*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 2006. ISBN: 0321462254.

[15] P. D. Hsu, E. S. Lander, and F. Zhang. "Development and applications of CRISPR-Cas9 for genome engineering". eng. In: *Cell* 157.6 (June 2014). S0092-8674(14)00604-7[PII], pp. 1262–1278. ISSN: 1097-4172. DOI: 10.1016/j.cell.2014.05.010. URL: https://doi.org/10.1016/j.cell.2014.05.010.

[16] J. Huang, A. J. Smola, A. Gretton, K. M. Borgwardt, and B. Scholkopf. "Correcting Sample Selection Bias by Unlabeled Data". In: *Proceedings of the 19th International Conference on Neural Information Processing Systems*. NIPS'06. Canada: MIT Press, 2006, pp. 601–608. URL: http://dl.acm.org/citation.cfm?id=2976456.2976532.

[17] R. A. Hughes and A. D. Ellington. "Synthetic DNA Synthesis and Assembly: Putting the Synthetic in Synthetic Biology". eng. In: *Cold Spring Harbor perspectives in biology* 9.1 (Jan. 2017). 9/1/a023812[PII], a023812. ISSN: 1943-0264. DOI: 10.1101/cshperspect.a023812. URL: https://doi.org/10.1101/cshperspect.a023812.

[18] S. Jain, F. Farnoud, M. Schwartz, and J. Bruck. "Coding for Optimized Writing Rate in DNA Storage". In: *2020 IEEE International Symposium on Information Theory (ISIT)*. 2020, pp. 711–716. DOI: 10.1109/ISIT44484.2020.9174253.

[19] S. Jain, F. Farnoud, M. Schwartz, and J. Bruck. "Noise and Uncertainty in String-Duplication Systems". In: *IEEE Int. Symp. Information Theory (ISIT)*. Aachen, Germany, June 2017.

[20] S. Jain, F. Farnoud Hassanzadeh, M. Schwartz, and J. Bruck. "Duplication-Correcting Codes for Data Storage in the DNA of Living Organisms". In: *IEEE Transactions on Information Theory* 63.8 (Aug. 2017), pp. 4996–5010. ISSN: 0018-9448. DOI: 10.1109/TIT.2017.2688361.

[21] S. Jain, N. Raviv, and J. Bruck. "Attaining the 2nd Chargaff Rule by Tandem Duplications". In: *2018 IEEE International Symposium on Information Theory (ISIT)*. June 2018, pp. 2241–2245. DOI: 10.1109/ISIT.2018.8437526.

[22] S. Jain, F. Farnoud, and J. Bruck. "Capacity and Expressiveness of Genomic Tandem Duplication". In: *IEEE Trans. Information Theory* 63.10 (Oct. 2017). DOI: 10.1109/TIT.2017.2728079.

[23] S. Jain, B. Mazaheri, N. Raviv, and J. Bruck. "Cancer Classification from Healthy DNA". In: *bioRxiv* (2019). DOI: 10.1101/517839. eprint: https://www.biorxiv.org/content/early/2019/01/11/517839.full.pdf. URL: https://www.biorxiv.org/content/early/2019/01/11/517839.

[24] S. Jain, B. Mazaheri, N. Raviv, and J. Bruck. "Glioblastoma signature in the DNA of blood-derived cells". In: *PLOS ONE* 16.9 (Sept. 2021), pp. 1–12. DOI: 10.1371/journal.pone.0256831.

[25] S. Jain, B. Mazaheri, N. Raviv, and J. Bruck. "Short Tandem Repeats Information in TCGA is Statistically Biased by Amplification". In: *bioRxiv* (2019). DOI: 10.1101/518878. eprint: https://www.biorxiv.org/content/early/2019/01/11/518878.full.pdf. URL: https://www.biorxiv.org/content/early/2019/01/11/518878.

[26] S. Jain, X. Xiao, P. Bogdan, and J. Bruck. "Generator based approach to analyze mutations in genomic datasets". In: *Scientific Reports* 11.1 (Oct. 2021), p. 21084. ISSN: 2045-2322. DOI: 10.1038/s41598-021-00609-8. URL: https://doi.org/10.1038/s41598-021-00609-8.

[27] S. Jain, X. Xiao, P. Bogdan, and J. Bruck. "Predicting the Emergence of SARS-CoV-2 Clades". In: *bioRxiv* (2020). DOI: 10.1101/2020.07.26.222117. eprint: https://www.biorxiv.org/content/early/2020/07/27/2020.07.26.222117.full.pdf. URL: https://www.biorxiv.org/content/early/2020/07/27/2020.07.26.222117.

[28] M. Jinek, K. Chylinski, I. Fonfara, M. Hauer, J. A. Doudna, and E. Charpentier. "A Programmable Dual-RNA–Guided DNA Endonuclease in Adaptive Bacterial Immunity". In: *Science* 337.6096 (2012), pp. 816–821. ISSN: 0036-8075. DOI: 10.1126/science.1225829. eprint: https://science.sciencemag.org/content/337/6096/816.full.pdf. URL: https://science.sciencemag.org/content/337/6096/816.

[29] E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, et al. "Initial Sequencing and Analysis of the Human Genome". In: *Nature* 409.6822 (2001), pp. 860–921.

[30] E. S. Lander. "Initial impact of the sequencing of the human genome". In: *Nature* 470 (2011), pp. 187–197.

[31] H. H. Lee, R. Kalhor, N. Goela, J. Bolot, and G. M. Church. "Terminator-free template-independent enzymatic DNA synthesis for digital information storage". In: *Nature Communications* 10.1 (2019), p. 2383. ISSN: 2041-1723. DOI: 10.1038/s41467-019-10258-1. URL: https://doi.org/10.1038/s41467-019-10258-1.

[32] V. I. Levenshtein. "Efficient reconstruction of sequences". In: *IEEE Transactions on Information Theory* 47.1 (Jan. 2001), pp. 2–22. ISSN: 0018-9448. DOI: 10.1109/18.904499.

[33] G. Levinson and G. A. Gutman. "Slipped-Strand Mispairing: A Major Mechanism for DNA Sequence Evolution." In: *Molecular Biology and Evolution* 4.3 (1987), pp. 203–221.

[34] T. A. Manolio, F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff, D. J. Hunter, M. I. McCarthy, E. M. Ramos, L. R. Cardon, and A. Chakravarti. "Finding the missing heritability of complex diseases". In: *Nature* 461 (2009), pp. 747–753.

[35] B. H. Marcus, R. M. Roth, and P. H. Siegel. "An introduction to coding for constrained systems". In: *Lecture notes (* 2001 (). URL: http://www.math.ubc.ca/%E2%88%BCmarcus/Handbook/).

[36] B. Mazaheri, S. Jain, and J. Bruck. *Expert Graphs: Synthesizing New Expertise via Collaboration*. 2021. arXiv: 2107.07054 [cs.LG].

[37] B. Mazaheri, S. Jain, and J. Bruck. *Robust Correction of Sampling Bias Using Cumulative Distribution Functions*. 2020. arXiv: 2010.12687 [stat.ML].

[38] National Cancer Institute - TCGA, https://portal.gdc.cancer.gov/.

[39] P. M. Visscher, N. R. Wray, Q. Zhang, P. Sklar, M. I. McCarthy, M. A. Brown, and J. Yang. "10 years of GWAS Discovery: Biology, Function and Translation". In: *Am. J. Hum. Genet.* 101.1 (2017), pp. 5–22.

[40] A. Pagnoni, S. Gramatovici, and S. Liu. *PAC Learning Guarantees Under Covariate Shift*. 2018. arXiv: 1812.06393 [cs.LG].

[41] N. Raviv, S. Jain, and J. Bruck. "What is the Value of Data? on Mathematical Methods for Data Quality Estimation". In: *2020 IEEE International Symposium on Information Theory (ISIT)*. 2020, pp. 2825–2830. DOI: 10.1109/ISIT44484.2020.9174311.

[42] R. P. Savage. "The Paradox of Nontransitive Dice". In: *The American Mathematical Monthly* 101.5 (1994), pp. 429–436. ISSN: 00029890, 19300972. URL: http://www.jstor.org/stable/2974903.

[43] C. Schlötterer. "Evolutionary Dynamics of Microsatellite DNA". en. In: *Chromosoma* 109.6 (Sept. 2000), pp. 365–371. ISSN: 0009-5915, 1432-0886. DOI: 10.1007/s004120000089.

[44] K. Schwarze et al. "The complete costs of genome sequencing: a microcosting study in cancer and rare diseases from a single center in the United Kingdom". In: *Genetics in Medicine* 22.1 (Jan. 2020), pp. 85–94. ISSN: 1530-0366. DOI: 10.1038/s41436-019-0618-7. URL: https://doi.org/10.1038/s41436-019-0618-7.

[45] C. Shannon. "A Mathematical Theory of Communication". In: *The Bell System Technical Journal* 27.3 (July 1948), pp. 379–423. ISSN: 0005-8580. DOI: 10.1002/j.1538-7305.1948.tb01338.x.

[46] H. Shimodaira. "Improving predictive inference under covariate shift by weighting the log-likelihood function". In: *Journal of Statistical Planning and Inference* 90.2 (2000), pp. 227–244. ISSN: 0378-3758. DOI: https://doi.org/10.1016/S0378-3758(00)00115-4. URL: http://www.sciencedirect.com/science/article/pii/S0378375800001154.

[47] S. L. Shipman, J. Nivala, J. D. Macklis, and G. M. Church. "CRISPR–Cas Encoding of a Digital Movie into the Genomes of a Population of Living Bacteria". en. In: *Nature* 547.7663 (July 2017), pp. 345–349. ISSN: 0028-0836. DOI: 10.1038/nature23017.

[48] S. L. Shipman, J. Nivala, J. D. Macklis, and G. M. Church. "Molecular Recordings by Directed CRISPR Spacer Acquisition". en. In: *Science* (June 2016), aaf1175. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.aaf1175.

[49] A. Sud, B. Kinnersley, and R. S. Houlston. "Genome-wide association studies of cancer: current insights and future perspectives". In: *Nature Reviews* 17 (2017), pp. 692–704.

[50] M. Sugiyama, M. Krauledat, and K.-R. Müller. "Covariate Shift Adaptation by Importance Weighted Cross Validation". In: *J. Mach. Learn. Res.* 8 (Dec. 2007), pp. 985–1005. ISSN: 1532-4435. URL: http://dl.acm.org/citation.cfm?id=1314498.1390324.

[51] J. X. Sun et al. "A Direct Characterization of Human Mutation Based on Microsatellites". en. In: *Nature Genetics* 44.10 (Oct. 2012), pp. 1161–1165. ISSN: 1061-4036. DOI: 10.1038/ng.2398.

[52] M. Tang, M. Waterman, and S. Yooseph. "Zinc Finger Gene Clusters and Tandem Gene Duplication". In: *Journal of Computational Biology* 9.2 (2002), pp. 429–446.

[53] L. G. Valiant. "A theory of the Learnable". In: *Communications of the ACM* (1984), pp. 1134–1142.

[54] V. Vapnik and R. Izmailov. "Rethinking statistical learning theory: learning using statistical invariants". In: *Machine Learning* 108.3 (Mar. 2019), pp. 381–423. ISSN: 1573-0565. DOI: 10.1007/s10994-018-5742-0. URL: https://doi.org/10.1007/s10994-018-5742-0.

[55] T. Warnow. *Computational Phylogenetics: An Introduction to Designing Methods for Phylogeny Estimation*. 1st. USA: Cambridge University Press, 2017. ISBN: 1107184711.

[56] M. C. White, L. A. Peipins, M. Watson, K. F. Trivers, D. M. Holman, and J. L. Rodriguez. "Cancer prevention for the next generation". eng. In: *The Journal of adolescent health : official publication of the Society for Adolescent Medicine* 52.5 Suppl (May 2013). S1054-139X(13)00125-0[PII], S1–S7. ISSN: 1879-1972. DOI: 10.1016/j.jadohealth.2013.02.016. URL: https://doi.org/10.1016/j.jadohealth.2013.02.016.

[57] M. Yamada, T. Suzuki, T. Kanamori, H. Hachiya, and M. Sugiyama. "Relative Density-Ratio Estimation for Robust Distribution Comparison". In: *Advances in Neural Information Processing Systems 24*. Ed. by J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger. Curran Associates, Inc., 2011, pp. 594–602. URL: http://papers.nips.cc/paper/4254-relative-density-ratio-estimation-for-robust-distribution-comparison.pdf.