Gene Expression Imputation with Generative Adversarial Imputation Nets

Ramon Viñas [1] **Tiago Azevedo** [1] **Eric R. Gamazon** [2 3] **Pietro Lio** [1]

1. Introduction

High-throughput profiling of the transcriptome has revolutionised discovery methods in biological sciences. The resulting gene expression measurements can be used to uncover disease mechanisms (Emilsson et al., 2008; Gamazon et al., 2018), propose novel drug targets (Sirota et al., 2011; Evans & Relling, 2004), provide a basis for comparative genomics (King & Wilson, 1975; Colbran et al., 2019), and motivate a wide range of important biological problems.

One question of fundamental biological significance is to what extent the expression of a subset of genes can recover the full transcriptome. Genes involved in similar biological processes are likely to have similar expression profiles (Zhang & Horvath, 2005), raising the possibility of gene expression prediction from a minimal subset of genes. Moreover, gene expression measurements may suffer from unreliable or missing values because some genome regions are challenging to interrogate due to high genomic complexity or sequence homology (Conesa et al., 2016), highlighting the need for accurate imputation. Most gene expression studies continue to be performed with specimens derived from peripheral blood due to the difficulty of collecting other tissues, but gene expression may be highly tissue-specific, potentially limiting the utility of a proxy tissue.

To address these challenges, we develop an approach to gene expression imputation using Generative Adversarial Imputation Nets (GAIN) (Yoon et al., 2018). We present an architecture that recovers missing expression data for multiple tissue types under the *missing completely at random* assumption (Little & Rubin, 2019). To enlarge the possibility and scale of a study's expression data (e.g. by including samples from highly inaccessible tissues), we train our model on data from the Genotype-Tissue Expression (GTEx) project (Consortium et al., 2017), a reference resource (V8) that has generated a comprehensive collection of transcriptomes in a diverse set of tissues.

[1]Department of Computer Science and Technology, University of Cambridge, UK [2]Vanderbilt Genetics Institute and Data Science Institute, VUMC, Nashville, TN, USA. [3]Clare Hall, University of Cambridge, UK. Correspondence to: Ramon Viñas <rv340@cam.ac.uk>.

Under review for the Workshop on Computational Biology at ICML 2020, Vienna, Austria, 2020.

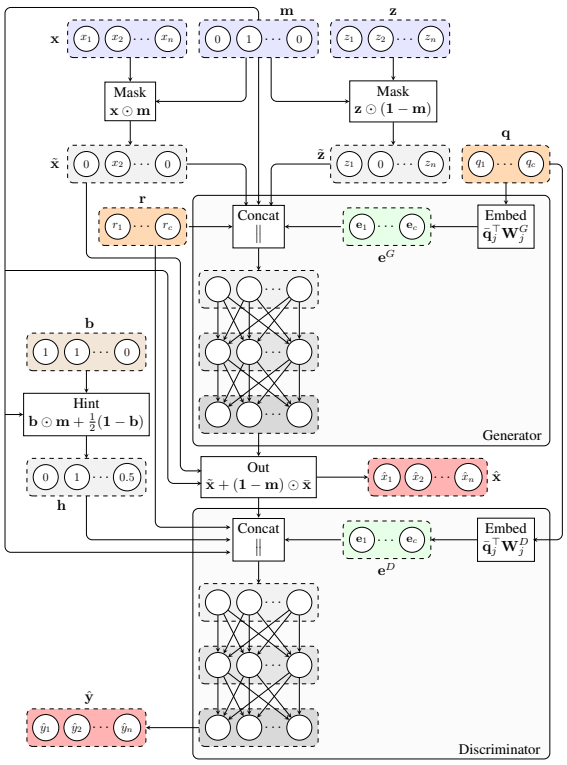

Figure 1. Architecture of the model. The generator takes gene expression values $\tilde{\mathbf{x}}$ with missing values as well as categorical (e.g. tissue type; $\mathbf{q}$) and numerical (e.g. age; $\mathbf{r}$) covariates, and outputs the imputed sample $\hat{\mathbf{x}}$. The discriminator receives $\hat{\mathbf{x}}$ and sample covariates, and produces the probabilities $\hat{\mathbf{y}}$ of each gene being observed as opposed to being imputed by the generator.

We show that our approach is superior to several standard and state-of-the-art imputation methods in terms of predictive performance and running time. Furthermore, we demonstrate that it is highly applicable across many different tissues and varying levels of missingness. To analyse the cross-study relevance, we evaluate our method on gene expression data from The Cancer Genome Atlas (TCGA) (Weinstein et al., 2013a). We show that our model generalises to RNA-Seq data from 3 cancer types, for which certain traditionally underrepresented populations, specifically in complex disease genomic studies (Wojcik et al., 2019), bear a disproportionate burden of poor health outcomes (Huo et al., 2017), with potentially important implications for ameliorating health disparities.

2. Methods

Let $\mathcal{D} = \{(\tilde{\mathbf{x}}, \mathbf{m}, \mathbf{r}, \mathbf{q})\}$ be a dataset where $\tilde{\mathbf{x}} \in \mathbb{R}^n$ represents a vector of gene expression values with missing components; $\mathbf{m} \in \{0,1\}^n$ is a mask indicating which components of the original vector of expression values $\mathbf{x}$ are missing or observed; n is the number of genes; and $\mathbf{q} \in \mathbb{N}^c$ and $\mathbf{r} \in \mathbb{R}^k$ are vectors of c categorical (e.g. tissue type or sex) and k quantitative covariates (e.g. age), respectively. Our goal is to recover the original gene expression vector $\mathbf{x} \in \mathbb{R}^n$ by modeling the conditional probability distribution $P(\mathbf{X} = \mathbf{x}|\tilde{\mathbf{X}} = \tilde{\mathbf{x}}, \mathbf{M} = \mathbf{m}, \mathbf{R} = \mathbf{r}, \mathbf{Q} = \mathbf{q})$. By modeling the distribution, we can also quantify the uncertainty of the imputed expression values.

Our model is based on a Generative Adversarial Imputation Network (GAIN, Yoon et al., 2018). The architecture of both players is shown in Figure 1.

Generator. The generator aims at recovering missing data from partial gene expression observations, producing samples from the conditional $P(\mathbf{X}|\tilde{\mathbf{X}}, \mathbf{M}, \mathbf{Q}, \mathbf{R})$. Formally, we define the generator as a function $G : \mathbb{R}^n \times \mathbb{R}^n \times \{0,1\}^n \times \mathbb{R}^k \times \mathbb{N}^c \to \mathbb{R}^n$ that imputes expression values as follows:

$$\bar{\mathbf{x}} = G(\mathbf{x} \odot \mathbf{m}, \mathbf{z} \odot (\mathbf{1} - \mathbf{m}), \mathbf{m}, \mathbf{r}, \mathbf{q}) \tag{1}$$

where $\mathbf{z} \in \mathbb{R}^n$ is a vector sampled from a fixed noise distribution. Here $\odot$ denotes element-wise multiplication. Similar to GAIN, we mask the n-dimensional noise vector as $\mathbf{z} \odot (\mathbf{1} - \mathbf{m})$ to encourage a bijective mapping between noise components and genes. Before passing the output $\bar{\mathbf{x}}$ to the discriminator, we replace the prediction for the non-missing components by the original, observed expression values:

$$\hat{\mathbf{x}} = \mathbf{m} \odot \tilde{\mathbf{x}} + (\mathbf{1} - \mathbf{m}) \odot \bar{\mathbf{x}} \tag{2}$$

Discriminator. The discriminator takes the imputed samples $\hat{\mathbf{x}}$ and attempts to distinguish whether the expression value of each gene has been observed or produced by the generator. This is in contrast to the original GAN discriminator, which receives information from two input streams (generator and data distribution) and attempts to distinguish the true input source.

Formally, the discriminator is a function $D : \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^k \times \mathbb{N}^c \to \mathbb{R}^n$ that outputs the probability $\hat{\mathbf{y}} \in \mathbb{R}^n$ of each gene being observed as opposed to being imputed by the generator:

$$\hat{\mathbf{y}} = D(\hat{\mathbf{x}}, \mathbf{h}, \mathbf{q}, \mathbf{r}) \tag{3}$$

Here, the vector $\mathbf{h} \in \mathbb{R}^n$ corresponds to the *hint* mechanism described in (Yoon et al., 2018), which provides theoretical guarantees on the uniqueness of the global minimum for the estimation of $P(\mathbf{X}|\tilde{\mathbf{X}}, \mathbf{M}, \mathbf{R}, \mathbf{Q})$. Concretely, the role of the hint vector $\mathbf{h}$ is to *leak* some information about the mask $\mathbf{m}$ to the discriminator. Similar to GAIN, we define the hint $\mathbf{h}$ as follows:

Table 1. Imputation performance with a missing rate of 50% across 10 runs. The running time of MICE and MissForest is > 7 days. When $R^2 < 0$, the mean of the data provides a better fit.

Method	R^2	Feasible?
MICE	–	×
MissForest	–	×
Blood surrogate	-0.924 ± 0.136	√
Median imputation	-0.024 ± 0.012	√
GAIN-MSE-GTEx	0.649 ± 0.021	√
GAIN-GTEx	0.659 ± 0.022	√

$$\mathbf{h} = \mathbf{b} \odot \mathbf{m} + \frac{1}{2}(\mathbf{1} - \mathbf{b}) \qquad \mathbf{b} \sim B(1, p) \qquad p \sim U(\alpha, \beta) \tag{4}$$

where $\mathbf{b} \in \{0,1\}^n$ is a binary vector that controls the amount of information from the mask $\mathbf{m}$ revealed to the discriminator. In contrast to GAIN, which discloses all but one components of the mask, we sample $\mathbf{b}$ from a Bernoulli distribution parametrised by a random probability $p \sim U(\alpha, \beta)$, where $\alpha \in [0,1]$ and $\beta \in [\alpha, 1]$ are hyperparameters. This accounts for a high number of genes n and allows to trade off the number of mask components that are revealed to the discriminator.

Optimisation. Similarly to GAN and GAIN, we optimise the generator and discriminator adversarially, interleaving gradient updates for the discriminator and generator.

For the discriminator, we penalise the errors for genes whose corresponding mask has not been revealed through the hint mechanism. We achieve this via the following loss function $\mathcal{L}_D : \{0,1\}^n \times \mathbb{R}^n \times \{0,1\}^n \to \mathbb{R}$:

$$\mathcal{L}_D(\mathbf{m}, \hat{\mathbf{y}}, \mathbf{b}) = \frac{-1}{Z}(\mathbf{1} - \mathbf{b})^\top \big(\mathbf{m} \odot \log \hat{\mathbf{y}} + (\mathbf{1} - \mathbf{m}) \odot (1 - \log \hat{\mathbf{y}})\big) \tag{5}$$

where $Z = 1 + (\mathbf{1} - \mathbf{b})^\top (\mathbf{1} - \mathbf{b})$ is a normalisation term.

For the generator, we penalise both reconstruction and imputation errors. Similar to GAIN, we use the following loss function $\mathcal{L}_G : \{0,1\}^n \times \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^n \times \{0,1\}^n \to \mathbb{R}$ for the generator:

$$\mathcal{L}_G(\mathbf{m}, \mathbf{x}, \bar{\mathbf{x}}, \hat{\mathbf{y}}, \mathbf{b}) = \frac{-1}{Z_1}\big((\mathbf{1} - \mathbf{b}) \odot (\mathbf{1} - \mathbf{m})\big)^\top \log \hat{\mathbf{y}} + \frac{\lambda}{Z_2}\mathbf{m}^\top (\mathbf{x} - \bar{\mathbf{x}})^2 \tag{6}$$

where $Z_1 = 1 + (\mathbf{1} - \mathbf{b})^\top (\mathbf{1} - \mathbf{b})$ and $Z_2 = \mathbf{m}^\top \mathbf{m}$ are normalisation terms, and $\lambda > 0$ is a hyperparameter.

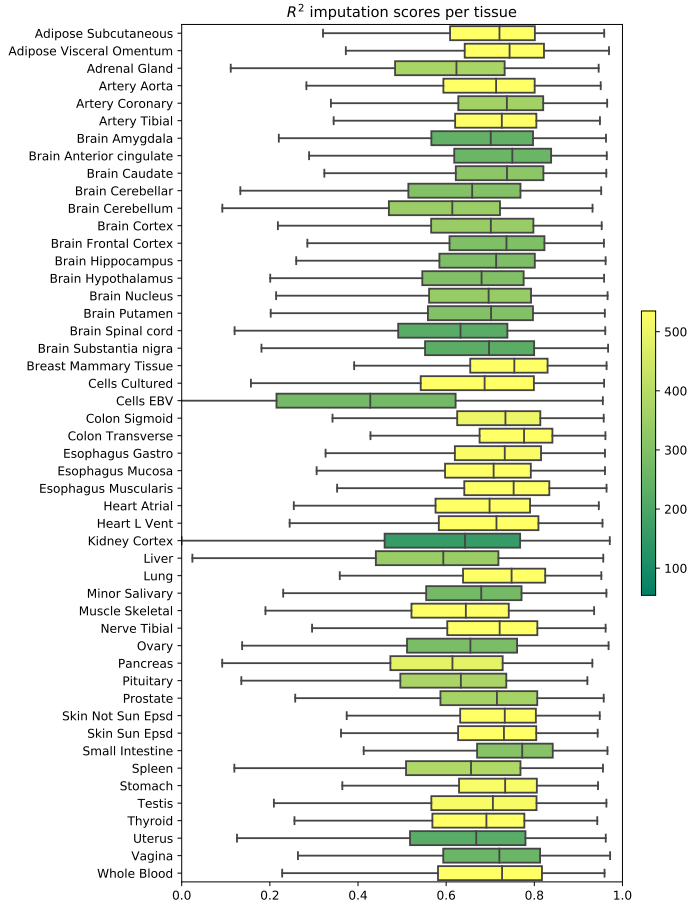

Figure 2. R^2 imputation scores per GTEx tissue with a missing rate of 50%. Each box shows the distribution of the per-gene R^2 scores in the test set. The color of each box represents the number of training samples per tissue.

3. Results

Comparison. We compare our approach, which we name GAIN-GTEx, to 2 standard approaches (blood surrogate[1] and median imputation), 2 state-of-the-art methods (MICE, Buuren & Groothuis-Oudshoorn, 2010; and MissForest, Stekhoven & Bühlmann, 2012), and a simplified version of our method trained exclusively on the mean squared error (GAIN-MSE-GTEx).

Table 1 shows a quantitative summary of the imputation performances across 12,557 unique human genes. In addition to the imputation scores, we include a *feasible* column that shows whether the gene expression imputation is computationally *feasible* under our experimental design. We label methods as *unfeasible* when they take longer than 7 days to run on our server[2], after which we halt the execution.

[1]We impute missing gene expression values in any given tissue with the corresponding expression values measured in whole blood.

[2]CPU: Intel(R) Xeon(R) Processor E5-2630 v4. RAM: 125GB

Table 2. Cross-study results for our model trained on GTEx. For a missing rate of 50%, we report the R^2 scores on data from 3 TCGA cancers and their *healthy* counterpart on GTEx (test set).

Tissue	R^2
TCGA LAML	0.386 ± 0.057
TCGA BRCA	0.408 ± 0.023
TCGA LUAD	0.439 ± 0.034
GTEx Whole blood	0.678 ± 0.031
GTEx Breast	0.724 ± 0.036
GTEx Lung	0.713 ± 0.033

Tissue-specific results. Figure 2 shows the R^2 scores achieved by GAIN-GTEx across 49 tissues. To obtain these results, we generate random masks with a missing rate of 50% for the extended test set where each tissue is equally represented, we perform imputation, and we plot the distribution of 12,557 gene R^2 scores for each tissue.

Cross-study results across missing rates. To evaluate the cross-study relevance of our method, we leverage the model trained on GTEx to perform imputation on The Cancer Genome Atlas (TCGA) gene expression data in acute myeloid leukemia (TCGA LAML; Network, 2013), breast cancer (TCGA BRCA; Network et al., 2012), and lung adenocarcinoma (TCGA LUAD; Network et al., 2014). For each TCGA tissue and its *non-diseased* test counterpart on GTEx, we show the imputation quality in Table 2 as well as the performance across varying missing rates in Figure 3.

4. Discussion

We develop an imputation approach to gene expression, facilitating the reconstruction of a high-dimensional molecular trait that is central to disease biology and drug target discovery. Our model builds on GAIN to learn complex probability distributions from incomplete gene expression data and relevant covariates.

To enlarge the possibility and scale of a study's expression data, we leverage the most comprehensive human transcriptome resource available (GTEx V8), allowing us to test the performance of our method in a broad collection of tissues (see Figure 2). The biospecimen repository includes model systems such as whole blood and Epstein Barr virus (EBV) transformed lymphocytes; central nervous system tissues from 13 brain regions; and a wide diversity of other primary tissues from *non-diseased* individuals. In particular, we observe that EBV transformed lymphocytes, an accessible and renewable resource for functional genomics, are a notable outlier in imputation performance. This is perhaps not surprising, consistent with studies about the transcriptional effect of EBV infection on the suitability of the cell lines as a model system for primary tissues (Carter et al., 2002).

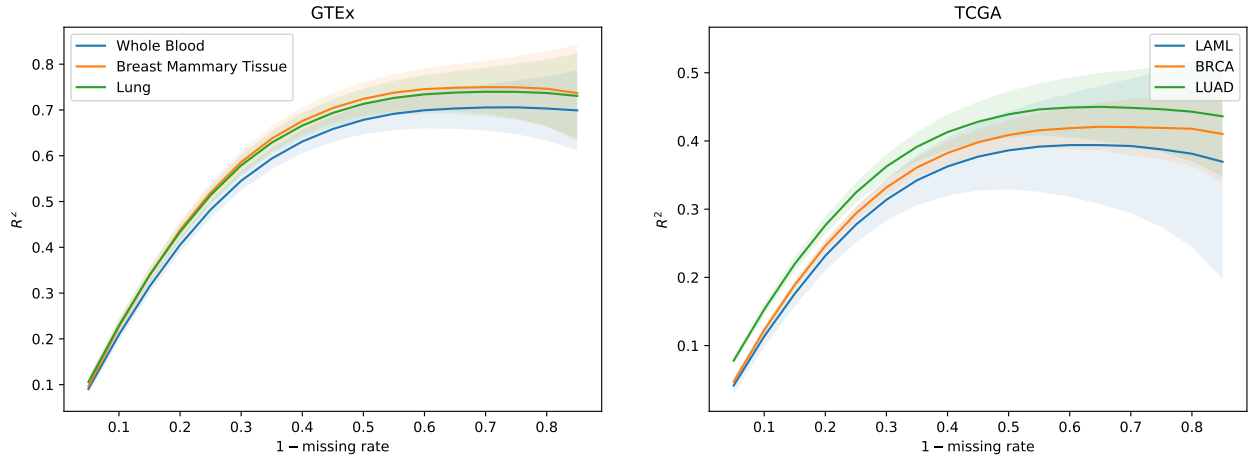

Figure 3. R^2 imputation scores per tissue across missing rate for 3 TCGA cancers and their healthy counterpart in GTEx (test set). The shaded area represents one standard deviation of the per-gene R^2 scores in the corresponding tissue. We note that the performance is stable and that the greater the proportion of missing values, the lower the prediction performance.

We compare our approach with several existing imputation methods and find that GAIN-GTEx outperforms them in terms of imputation performance and runtime (see Table 1). We observe that standard approaches such as leveraging the expression of missing genes from a surrogate blood tissue yield negative R^2 values and therefore do not perform well. Median imputation, although easy to implement, has a very limited predictive power. In terms of state-of-the-art-methods, we note that MICE and MissForest are computationally prohibitive given the high-dimensionality of the data and we halt the execution after running our experiments for 7 days. Finally, we observe that a simplification of our method, GAIN-MSE-GTEx, also performs well, suggesting that the mean squared error term of the generator's loss function has a major role in the learning process.

To evaluate the cross-study relevance of our method, we apply the trained model derived from GTEx to perform imputation on The Cancer Genome Atlas gene expression data in acute myeloid leukemia, lung adenocarcinoma, and breast cancer. In addition to technical artifacts (e.g. batch effects), generalising to this data is highly challenging because the expression is largely driven by features of the disease such as chromosomal abnormalities, genomic instabilities, large-scale mutations, and epigenetic changes (Weinstein et al., 2013b). Our results show that, despite these challenges, our method is robust to gene expression from multiple diseases in different tissues (see Table 2), lending itself to being used as a tool to extend independent transcriptomic studies. Finally, we evaluate the imputation performance of GAIN-GTEx for a range of values for the missing rate (see Figure 3). We note that the performance is stable and that the greater the proportion of missing values, the lower the prediction performance.

5. Conclusion

In this work, we develop a method for gene expression imputation that achieves state-of-the-art performance in terms of imputation quality and running time. Our analysis shows that the use of blood as a surrogate for inaccessible tissues, as widely practiced throughout biomedical research, has substantially degraded performance, with important implications for biomarker discovery and therapeutic development.

Our model can facilitate the straightforward integration and cost-effective repurposing of large-scale RNA biorepositories and resources into genomic studies of disease, with high applicability across diverse tissue types. Moreover, our approach generalises to gene expression in a disease class which has shown considerable health outcome disparities across population groups in terms of morbidity and mortality, with potential global health application to detection, diagnosis, and treatment (Hosny & Aerts, 2019). Finally, this study has the potential to catalyse research into the application of Generative Adversarial Networks (Goodfellow et al., 2014) for molecular reconstruction of cellular states and downstream gene mapping of complex traits (Gamazon et al., 2015).

Acknowledgements

The project leading to these results has received funding from "la Caixa" Foundation (ID 100010434), under agreement LCF/BQ/EU19/11710059. This research is supported by the National Institutes of Health / NHGRI R35 HG010718 (ERG).

References

Buuren, S. v. and Groothuis-Oudshoorn, K. mice: Multivariate imputation by chained equations in r. *Journal of statistical software*, pp. 1–68, 2010.

Carter, K. L., Cahir-McFarland, E., and Kieff, E. Epstein-barr virus-induced changes in b-lymphocyte gene expression. *Journal of virology*, 76(20):10427–10436, 2002.

Colbran, L. L., Gamazon, E. R., Zhou, D., Evans, P., Cox, N. J., and Capra, J. A. Inferred divergent gene regulation in archaic hominins reveals potential phenotypic differences. *Nature ecology & evolution*, 3(11):1598–1606, 2019.

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szcześniak, M. W., Gaffney, D. J., Elo, L. L., Zhang, X., et al. A survey of best practices for rna-seq data analysis. *Genome biology*, 17 (1):13, 2016.

Consortium, G. et al. Genetic effects on gene expression across human tissues. *Nature*, 550(7675):204–213, 2017.

Emilsson, V., Thorleifsson, G., Zhang, B., Leonardson, A. S., Zink, F., Zhu, J., Carlson, S., Helgason, A., Walters, G. B., Gunnarsdottir, S., et al. Genetics of gene expression and its effect on disease. *Nature*, 452(7186): 423–428, 2008.

Evans, W. E. and Relling, M. V. Moving towards individualized medicine with pharmacogenomics. *Nature*, 429 (6990):464–468, 2004.

Gamazon, E. R., Wheeler, H. E., Shah, K. P., Mozaffari, S. V., Aquino-Michaels, K., Carroll, R. J., Eyler, A. E., Denny, J. C., Nicolae, D. L., Cox, N. J., et al. A gene-based association method for mapping traits using reference transcriptome data. *Nature genetics*, 47(9):1091, 2015.

Gamazon, E. R., Segrè, A. V., van de Bunt, M., Wen, X., Xi, H. S., Hormozdiari, F., Ongen, H., Konkashbaev, A., Derks, E. M., Aguet, F., et al. Using an atlas of gene regulation across 44 human tissues to inform complex disease-and trait-associated variation. *Nature genetics*, 50(7):956–967, 2018.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial networks, 2014.

Hosny, A. and Aerts, H. J. Artificial intelligence for global health. *Science*, 366(6468):955–956, 2019.

Huo, D., Hu, H., Rhie, S. K., Gamazon, E. R., Cherniack, A. D., Liu, J., Yoshimatsu, T. F., Pitt, J. J., Hoadley, K. A., Troester, M., et al. Comparison of breast cancer molecular features and survival by african and european ancestry in the cancer genome atlas. *JAMA oncology*, 3 (12):1654–1662, 2017.

King, M.-C. and Wilson, A. C. Evolution at two levels in humans and chimpanzees. *Science*, 188(4184):107–116, 1975.

Little, R. J. and Rubin, D. B. *Statistical analysis with missing data*, volume 793. John Wiley & Sons, 2019.

Network, C. G. A. et al. Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418):61, 2012.

Network, C. G. A. R. Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *New England Journal of Medicine*, 368(22):2059–2074, 2013.

Network, C. G. A. R. et al. Comprehensive molecular profiling of lung adenocarcinoma. *Nature*, 511(7511): 543–550, 2014.

Sirota, M., Dudley, J. T., Kim, J., Chiang, A. P., Morgan, A. A., Sweet-Cordero, A., Sage, J., and Butte, A. J. Discovery and preclinical validation of drug indications using compendia of public gene expression data. *Science translational medicine*, 3(96):96ra77–96ra77, 2011.

Stekhoven, D. J. and Bühlmann, P. Missforest—nonparametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118, 2012.

Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R. M., Ozenberger, B. A., Ellrott, K., Shmulevich, I., Sander, C., Stuart, J. M., Network, C. G. A. R., et al. The cancer genome atlas pan-cancer analysis project. *Nature genetics*, 45(10):1113, 2013a.

Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R. M., Ozenberger, B. A., Ellrott, K., Shmulevich, I., Sander, C., Stuart, J. M., Network, C. G. A. R., et al. The cancer genome atlas pan-cancer analysis project. *Nature genetics*, 45(10):1113, 2013b.

Wojcik, G. L., Graff, M., Nishimura, K. K., Tao, R., Haessler, J., Gignoux, C. R., Highland, H. M., Patel, Y. M., Sorokin, E. P., Avery, C. L., et al. Genetic analyses of diverse populations improves discovery for complex traits. *Nature*, 570(7762):514–518, 2019.

Yoon, J., Jordon, J., and Van Der Schaar, M. Gain: Missing data imputation using generative adversarial nets. *arXiv preprint arXiv:1806.02920*, 2018.

Zhang, B. and Horvath, S. A general framework for weighted gene co-expression network analysis. *Statistical applications in genetics and molecular biology*, 4(1), 2005.