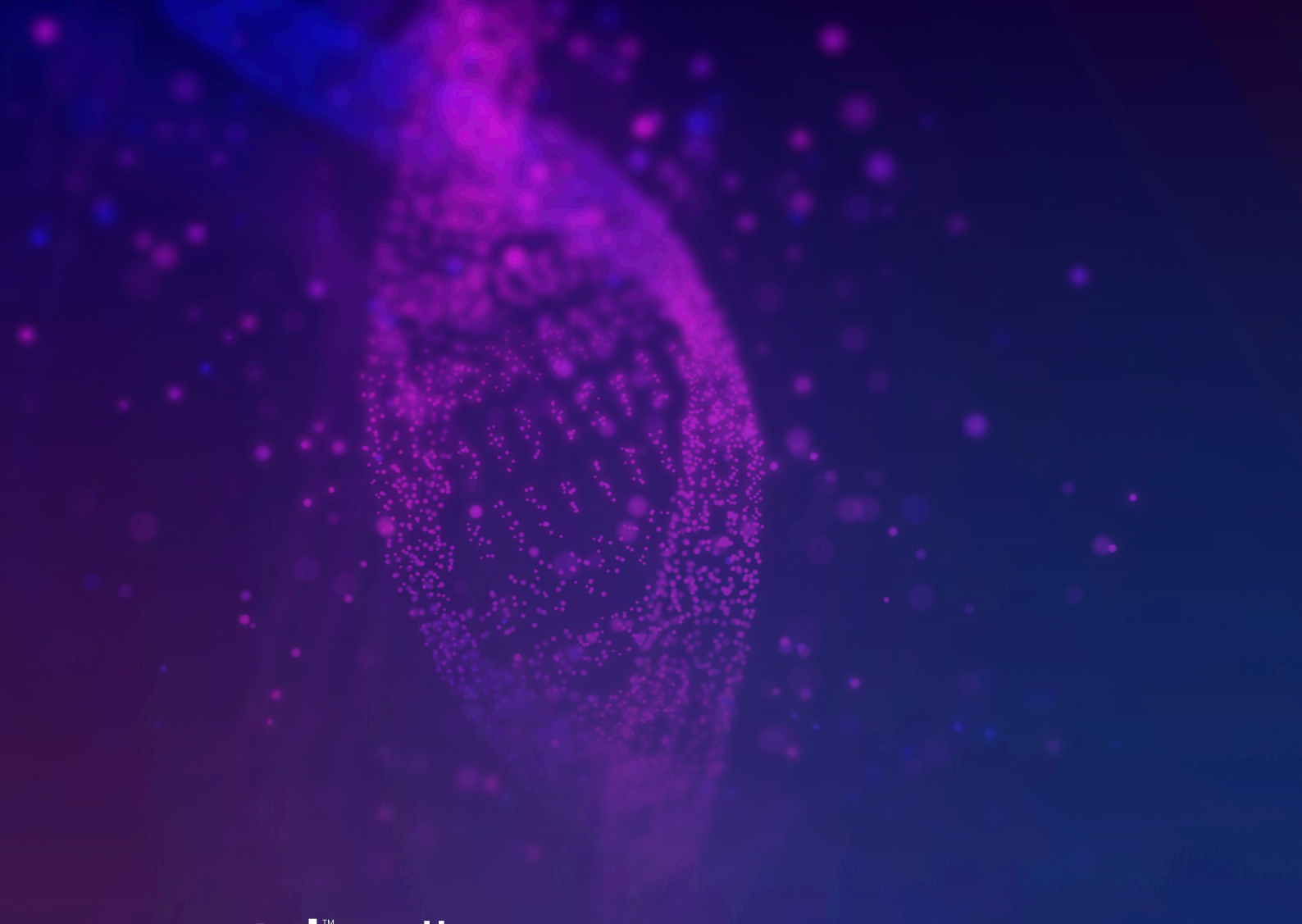

gretel™ + illumina®

Using AI to create safe synthetic datasets for genomics.

DECEMBER 14, 2021

GRETEL.AI

Amy Steier

Alex Watson

ILLUMINA

Pam Cheng

Vlad Sima

Geoff Nilsen

Thon de Boer

Table of contents

The Life Sciences Data Bottleneck

DNA microarrays and sequencing have revolutionized biology and are in the process of reinventing health care. Massive datasets of tens of thousands to millions of affected individuals (with matched controls) allow researchers to study genetic factors associated with disease with unprecedented resolution. Genetic test results combined with electronic health record data are helping us realize the potential of precision medicine. However, the inherent sensitivity of genetic data, strong privacy legislation, and terms of patient research consents limit the sharing of this data. The process for a researcher to request access to genomic data to test an idea or hypothesis must be approved on a case-by-case basis and often takes months. This can reduce the statistical power of scientific studies, and limit study enrollment to only those populations easily available to a given researcher.

Synthetic data, powered by recent advances in machine learning, is a promising technology to create artificial versions of sensitive datasets. Along with privacy enhancing technologies such as differential privacy, synthetic data has the potential to address the deep privacy concerns working with genomic data, enabling faster sharing of data and unlocking innovation.

Case Study Overview

Scientists explore the relationships between phenotypes (physical characteristics) and genotypes (measured genetic "spelling" at specific locations) in large groups of individuals in order to better understand the causes of disease, find genetic risk factors, and find targets for new drugs. One such widely-used study population is the UK Biobank, which contains more than 7,000 phenotypic fields and more than 800,000 SNP genotypes across more than 500,000 individuals, for a total of more than 400 billion data points.

Here we are using a more modestly-sized dataset from mice so that we can freely share the data and code with readers. We begin with the dataset and analysis of Parker et al[3], which describes genome-wide association analyses of 68 phenotypes with 92,734 single nucleotide polymorphisms (SNPs), in 1,200 mice. The authors used an outbred strain of mice to avoid linkage problems with common inbred strains.

Mice dataset characteristics *(used in this experiment)*	**UK Biobank** *(for comparison)*
1,220 samples, each containing a genotype record (as SNPs) and a matching phenotype record.	500,000+ samples
92,734 genotypes per mouse	800,000+ genotypes per sample
164 phenotypes	7,000+ phenotypes

Exploring the mouse genome dataset

There are two datasets that we will synthesize for this test: one for genotypes and another for phenotypes. The phenotype dataset represents both a series of measurements of various characteristics in mice as well as other external variables that can have an influence on the results. A comprehensive list is available in the original research paper.

	id	round	cageid	FCbox	PPIbox	methcage	methcycle	discard	mixup	earpunch	...	SW16	SW17	SW18	SW19	SW20	SW21	SW22	SW23	SW24	SW2
1	26305	SW18	1330002.0	1.0	1.0	1.0	1.0	no	no	R	...	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0
2	26306	SW18	1330002.0	2.0	3.0	2.0	1.0	no	no	R	...	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0
3	26307	SW18	1330002.0	3.0	4.0	3.0	1.0	no	no	L	...	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0
4	26308	SW18	1330002.0	4.0	5.0	4.0	1.0	no	no	L	...	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0
5	26309	SW18	1330003.0	1.0	1.0	5.0	1.0	no	no	R	...	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0

5 rows × 164 columns

Figure 1 - Example phenotype data

The genotype dataset contains the genomic information of the mice. Considering that most of the genome is the same across individuals in a species, only the differences with respect to a species-specific reference genome are reported. The simplest form of such a difference is called a SNP (single nucleotide polymorphism). In our dataset, the column names are SNP identifiers–either identifiers in NCBI's dbSNP or formed by joining the chromosome and its position on the chromosome.

	id	discard	cfw-1-3082859	cfw-1-3207478	cfw-1-3284999	cfw-1-4056451	rs241840178	cfw-1-4592184	rs214108183	rs31954814	...	rs239202862	cfw-19-60773695	rs212272420	rs5122300
0	26305	no	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.000	...	2.000	2.000	2.000	0.
1	26306	no	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000	...	1.929	1.980	1.999	0.
2	26307	no	0.0	0.0	0.0	0.0	0.0	2.0	0.0	0.000	...	1.933	1.603	1.993	0.
3	26308	no	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.776	...	2.000	2.000	2.000	0.
4	26309	no	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000	...	2.000	2.000	2.000	0.

5 rows × 92736 columns

Figure 2 - Example genotype data

In most animals, including mice and humans, there are two copies of each chromosome. So for a given SNP at a position you may have:

- A difference with respect to the reference genome in both chromosomes (which is represented as 2).
- A difference in only one of the copies (represented as 1).
- No differences at all (represented as 0).

Some genotypes have non-integer values, such as the final row in the figure above with a value of 1.439. There are multiple reasons why this can occur, but one of the most common is that the measuring process could not provide a definitive answer on the variant, and rather than make an approximation the data is provided “as-is”.

Finally, the mapping dataset maps each SNP identifier to its matching chromosome and position on the chromosome.

	id	chr	pos	ref	alt	quality
0	cfw-1-3082859	1	3082859	T	G	1.0000
1	cfw-1-3207478	1	3207478	G	A	0.9991
2	cfw-1-3284999	1	3284999	A	C	1.0000
3	cfw-1-4056451	1	4056451	A	C	1.0000
4	rs241840178	1	4289606	C	T	1.0000
...	...	...	...	...	...	...
92729	rs30654044	19	61003919	G	A	1.0000
92730	rs30990073	19	61004714	T	C	0.9914
92731	rs50978457	19	61012527	A	G	0.9991
92732	rs51755773	19	61067713	A	G	0.9931
92733	cfw-19-61107432	19	61107432	T	G	0.9914

92734 rows × 6 columns

Figure 3 - Example SNP to chromosome position mapping

Genome wide association study

In a genome-wide association study (GWAS), researchers seek to identify locations on the genome that are associated with a particular phenotype. This is done with a statistical test to determine whether any SNP displays a significantly different distribution of non-reference-to-reference alleles between the case and control populations. Because there are so many SNP locations on a genome, the distribution has to be quite different to be sure the observed differences aren't just due to chance. (If you flip a coin a thousand times, finding a run of ten heads is expected but a run of a hundred heads is not.) The likelihood that an association is due to chance is measured by the p-value; lower p-values are more significant.

Here we use the open source GEMMA software that is also used in the original research. The result of the GEMMA and linear regression algorithms for each phenotype comprise approximately 80,000 rows; each row containing a SNP and its calculated p value.

	index	snp	chr	pos	p
0	73583	cfw-1-3207478	1	3207478	0.921913
1	40919	cfw-1-4592184	1	4592184	0.496735
2	29303	rs31954814	1	5151352	0.353184
3	44285	rs31947195	1	5240999	0.540256
4	40335	rs30660852	1	5241015	0.489395
...	...	...	...	...	...
79640	72004	rs30654044	19	61003919	0.901308
79641	10500	rs30990073	19	61004714	0.116835
79642	18160	rs50978457	19	61012527	0.212267
79643	51684	rs51755773	19	61067713	0.637528
79644	44010	cfw-19-61107432	19	61107432	0.537170

79645 rows × 5 columns

Figure 4 - Example GWAS result

The p value threshold used for human GWAS studies is around 1e-8 magnitude, but multiple values can be used in practice. The authors of the original paper here computed the significance threshold at 2e-6, due to the lower allelic diversity in laboratory mice (even in this outbred population.) The abBMD bone-mineral density analysis from the experiment is shown below, where we can see two regions having multiple SNPs are above the threshold for statistical significance to this trait.

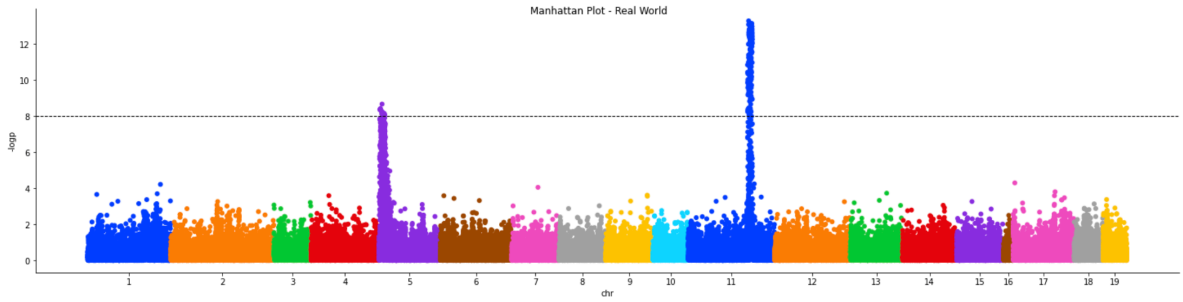

Figure 5 - Manhattan plot showing original GWAS results

Synthesizing the mouse genome dataset

To determine whether synthetic data models are capable of creating an artificial genomic dataset that captures the discoveries in the research paper, we will create synthetic versions of the real-world phenotypes and genotypes gathered from mice in the sample dataset using Gretel.ai's synthetic data APIs, perform the same GWAS analysis, and then compare the results of the GWAS analysis for real-world vs synthetic data.

https://github.com/gretelai/synthetic-data-genomics

Case study steps

Below are the general steps used to synthesize the genotype and phenotype datasets and to recreate the results of the original experiment with synthetic data. To recreate the results yourself, follow along with the Jupyter notebook for each step.

1 **Build phenotype training set**

First, create and format real-world training data for each phenotype batch that will be synthesized. This step can be recreated in 01_build_phenome_training_data.ipynb.

2 **Synthesize phenotypes**

Train a synthetic model for each batch of correlated phenotypes in the phenotype dataset, then generate synthetic phenotypes matching their size and shape. This step can be recreated in 02_create_synthetic_mouse_phenomes.ipynb.

3 **Build genotype training set**

Format and build a training set for genome data by batching the genome data by position on the chromosome along with synthesized phenotype data. 03_build_genome_training_data.ipynb.

4 **Synthesize genotypes**

Train a synthetic model on real world SNPs. To ensure that the synthetic phenotypes line up with our newly created synthetic genotypes, we use the synthetic phenotypes to prompt data generation, with a focus in this analysis on one phenotype (and its covariate), namely abBMD.

Initial analysis of the synthetic genomes

We can use Gretel's synthetic data report and Synthetic Quality Score (SQS) to compare the accuracy of the synthetic genotype and phenotype data to the real world training sets. Below is an example of a correlation matrix between an example batch of 19 genotype SNPs from the synthetic data report that demonstrates the model's ability to learn correlations in the data.

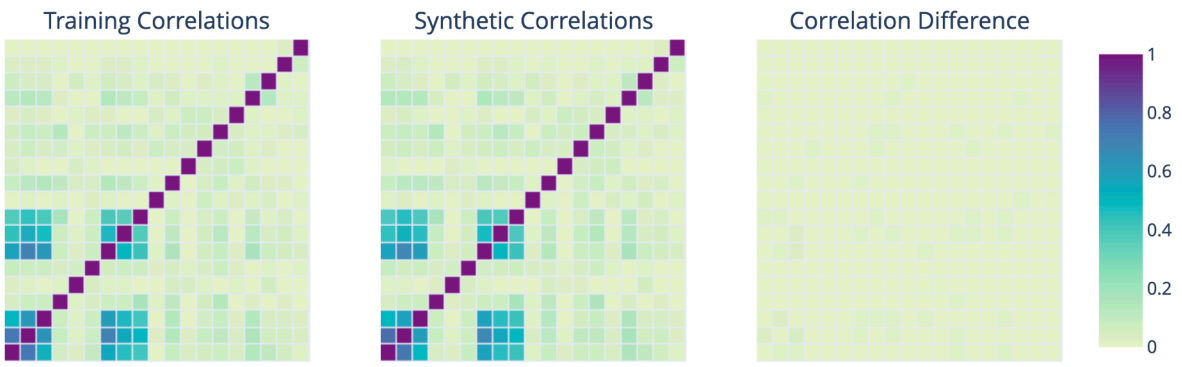

Figure 6 - Correlation matrices in synthetic data report

A principal component analysis (PCA) view of the real world vs synthetic data is another useful tool in the synthetic data quality report to examine how effectively the synthetic model learned the structure and distribution of the data. Once again, the results look quite promising on the sample batch of 19 SNPs and abBMD phenotype.

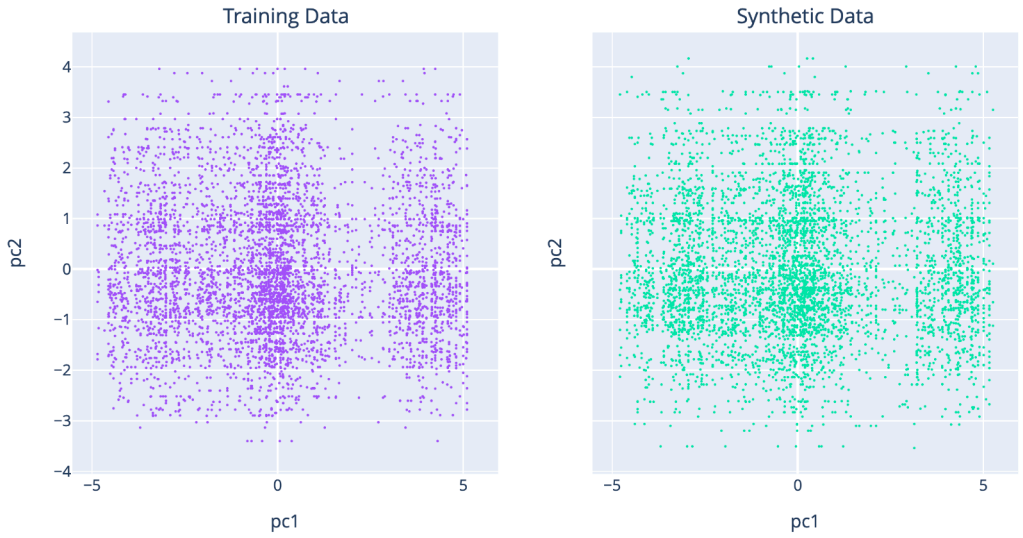

Figure 7 - PCA analysis in synthetic data report

Compute requirements

In this experiment, training time for Gretel.ai's synthetic data APIs, as described above for mice took approximately 36 hours using the default configuration in the attached notebooks, running on 80 Nvidia T4 GPUs in parallel. Nvidia T4 GPU instances in GCP, AWS, and Azure cost approximately $0.50/hour, bringing the total compute cost for model training and synthetic data generation of the mouse phenotypes and genotype dataset to approximately $1,440.

Comparing real world and synthetic datasets

We have now synthesized the abBMD phenotype data and all associated genotypes, and can compare the results of a genome-wide association study (GWAS) analysis using our real world and synthetic datasets. As the genome-wide significance (WGS) P value threshold of 1e-8 has become common for GWAS[7,8], we will use that as a cutoff to determine accuracy for comparisons between the synthetic and real-world results, although Parker et. al were able to use a lower threshold of 2e-6. The real world dataset analysis found 193 out of 71,315 SNPs with a p-value for the abBMD trait that was over the statistical significance threshold. In the synthetic dataset, 177 of the 193 SNPs were recreated by the GWAS analysis and false positives were low, with a SNP-wise precision (positive predictive value) of 93%.

We can use a Manhattan plot to represent the P values of the entire GWAS on a genomic scale and assess the performance of the model on a genomic locus-wise basis. In these plots, the -log 10 of the P values (y axis) are plotted in genomic order by chromosome and their position on the chromosome (x axis). SNPs associated with the studied trait will rise up high compared to the background, evoking skyscrapers in the Manhattan skyline. Generally, with a sufficient density of SNPs measured, clusters of SNPs (forming the "skyscrapers") are observed instead of just one or two high-flying causal SNPs as segments of chromosomes are what is inherited from our parents

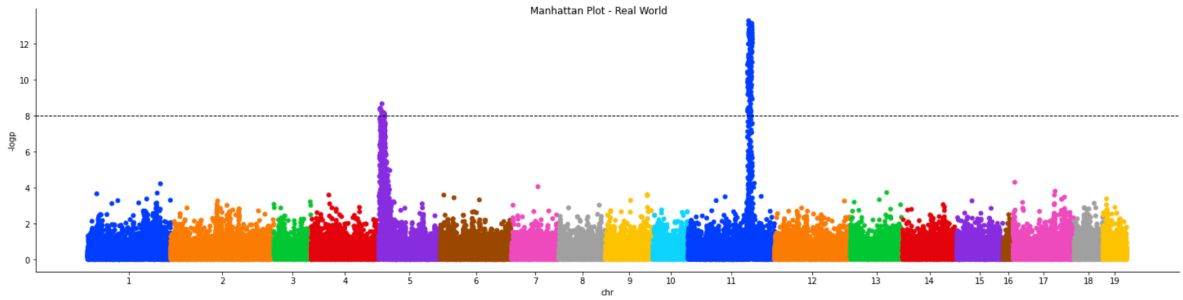

Figure 8 - GWAS results on real world data

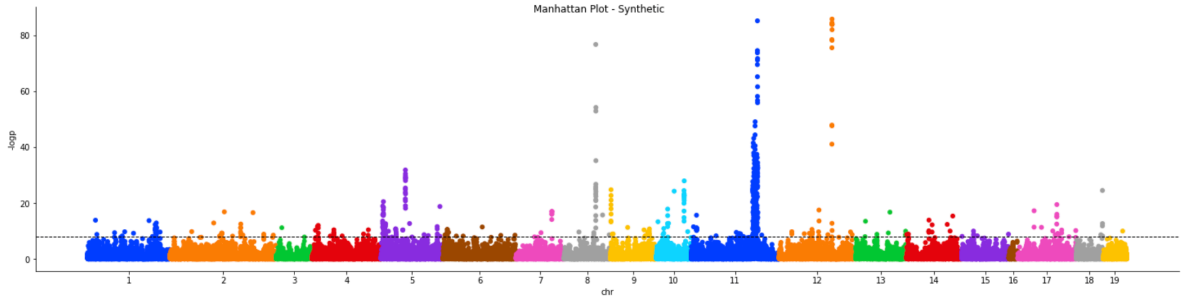

Figure 9 - GWAS results on synthetic data

These Manhattan plots demonstrate some interesting insights into our synthetic model and its performance. The synthetic model was able to capture and replay the strong associations in chromosomes 11 and 5. However, the synthetic model introduced notable false positive GWAS associations in chromosomes 8, 10, and 12. The false positives in the GWAS results above are most likely due to the small sample set size of 1,200 mice used to train the language model, where we generally recommend 10,000 or more examples for the network to sufficiently learn to recreate the data, especially with the complexity of the genome containing 92k SNPs per mouse. Similarly, the GWAS association noise floor and y scale is significantly higher in the synthetic data, indicating that the model might be amplifying characteristics in real world data that are being reflected in the GWAS analysis. These differences can also likely be minimized with additional examples and neural network parameter optimization, which we will explore in the next post.

Conclusion

While initial case study results are based on a relatively small sample set of 1,200 mice with limited testing and tuning, it demonstrates encouraging evidence that state of the art synthetic data models can produce artificial versions of even highly dimensional and complex genomic and phenotypic data. Our synthetic data model demonstrated the ability to recreate the key GWAS associations of the real world data, with total compute costs for training synthetic models on the genotype and phenotype data of only $1,440. With continued experiments in scale, accuracy, and privacy; synthetic data has the potential to enable sharing and collaboration on synthetic genomics datasets at a scale that is orders of magnitude larger than what is possible today.

What's next

We are working together to enable future genomics research and safe, private data sharing between researchers, health care providers, and industry. In our next posts, we will expand to human datasets, explore greater scale, train multiple phenotypes together with the genotypic data, and show the privacy guarantees that can be achieved working with synthetic data. If you have any questions or would like to discuss anything further we would love to talk to you. Feel free to reach out to us at hi@gretel.ai.

About Gretel

Gretel.ai was founded on a privacy-first mission to equip developers with the ability to unlock innovation through safe, efficient collaboration with sensitive data. Gretel pioneered Privacy Engineering as a Service and designed a synthetic data tool suite based on their open sourced AI-based core. These tools make it easier and faster to generate privacy-preserving data that can be safely shared.

Gretel created easy to use, accessible APIs for developers and data practitioners to generate high quality synthetic data, classify and label, and transform and anonymize data – tools that quickly remove privacy-related bottlenecks, and accelerate business innovation for organizations in financial services, life sciences, healthcare, technology, gaming and other industries.

About Illumina

Illumina is a leading developer, manufacturer, and marketer of life science tools and integrated systems for large-scale analysis of genetic variation and function. These systems are enabling studies that were not even imaginable just a few years ago, and moving us closer to the realization of personalized medicine. With rapid advances in technology taking place, it is mission-critical to offer solutions that are not only innovative, but flexible, and scalable, with industry-leading support and service.

We strive to meet this challenge by placing a high value on collaborative interactions, rapid delivery of solutions, and meeting the needs of our customers.

Our customers include a broad range of academic, government, pharmaceutical, biotechnology, and other leading institutions around the globe.

Concepts and Notation

Genomic data is the DNA of organisms. Genomic data often requires a large amount of storage, and is expected to generate exabytes of data over the next decade[1].

Synthetic data is artificial information generated by computer algorithms or simulations that can be used as an alternative to real world data[2]. Research has shown that synthetic data can be as good or even better than real world data for data analysis and training AI models, and that it can be engineered to reduce biases and increase privacy[2].

SNP - A single nucleotide polymorphism is a variation at a single position in a DNA sequence among individuals. Possible DNA bases are A, C, T, and G. For example, at position 19962213 on human chromosome 8 (Build 38), 90% of chromosomes have a C and 10% of chromosomes have a G.

P value - A P-value expresses the probability that a given result from a test is due to chance. “To account for multiple testing in genome-wide association studies (GWAS), a fixed P value threshold of $1x10^{-8}$ is widely used to identify association between a common genetic variant and a trait of interest.op

References

1. https://www.snowflake.com/trending/genomic-data

2. https://docs.gretel.ai/synthetics/synthetics-faqs#what-is-synthetic-data

3. Parker, C., Gopalakrishnan, S., Carbonetto, P. *et al.* Genome-wide association study of behavioral, physiological and gene expression traits in outbred CFW mice. Nat Genet 48, 919–926 (2016). https://doi.org/10.1038/ng.3609

4. https://www.genome.gov/genetics-glossary/Genome-Wide-Association-Studies

5. Xiang Zhou and Matthew Stephens (2012). Genome-wide efficient mixed-model analysis for association studies. Nature Genetics 44, 821–824.

6. Risch N , Merikangas K. 1996. The future of genetic studies of complex human diseases. Science. 273:1516–1517. doi:10.1126/science.273.5281.1516.

7. International HapMap Consortium: A haplotype map of the human genome. Nature 2005; 437: 1299–1320.

8. Pe'er I, Yelensky R, Altshuler D, Daly MJ : Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. Genet Epidemiol 2008; 32: 381–385.