Title: ""

Pages budget:

Review:

Abstracts for Reviews should be no more than 100 words in length.

The word limit for the main text of Reviews (excluding the reference list and figure legends) is 3,000 words.

Up to 150 references can be cited.

Opinions can contain up to 3 display items (tables, boxes, and figures).

Abstract

- An argumentative piece on how Genomics is driving data science

https://www.dropbox.com/s/vfsed9mdm21ra0z/DSG%20mind%20map.mup?dl=0

1. Introduction

One of the major goals of data science is to improve decision making by leveraging insights obtained from large datasets through and the application of statistics and computer science to real-world domains. Data science has caught a lot of public's attention in the most recent years, mostly due its application and impact in the technology industry (Google, Facebook). Much less appreciated, however, is the fact that a lot of early genomics, in particular, bioinformatics, was often data science even before it became broadly adopted. Genomics has had an intertwined growth, perhaps earlier than the overall field of data science, providing key ideas that one can see today reflected in other areas of data science and also borrowing concepts from other data-intensive fields.

Many factors contributed to the premature data-driven nature of genomics. Most of the technical advancements and data collection growth we have seen in the life sciences was driven by the advancements in computing power, the massive reduction in memory costs, and DNA sequencing. DNA sequencing technologies have undoubtedly been one of the most transformative developments in the life sciences. The drastic drop in sequencing cost and growth in sequence throughput allowed the collection of so much genetic data that some estimates project that the DNA sequencing data will surpass the amount of data collected in astronomy, industry, and social media \cite{MSchatz}. Moreover, DNA sequencing data is monolithic, discrete and not subjective. Like weather data, the number of sensors (sequencers) around the globe has grown exponentially and allowed uniform and consistent collection of data.

The connection between Genomics and Data science is important to recognize as going forward. Genomics, to some degree, might provide a paradigm for Data science and other life sciences, and also should be recognized as one of the key early applications for data-intensive computing. Here we will show some aspects of the large scale in genomics currently and how it has both imported and exported a variety of different technical and cultural aspects from other disciplines in data science. We also explore the prospects for the future of genomics, as one increasingly sees a very data-rich ecosystem develop in the biological sciences.

- Here we explore how genomics is an early instantiation of data science
- We also explore the imports and exports and interplays between both disciplines
- We explore the cultural and methodological exchanges between genomics and data science to highlight genomics as one of the best models for newer data-driven disciplines.
- Furthermore, we use the big data framework (4V's - Volume, Velocity, Variety, and Veracity) to describe the recent developments observed in genomics.

- We propose a new pathway for genomics in which the discipline is moving from a descriptive culture (i.e. data mining) to a predictive culture - forecasting (predicting generative models) by integrating physical models and large datasets. Analogous to weather forecasting.

2. Contextualizing Genomics and Data Sciences

Life sciences and statistics have always long shown evidence of intertwinement. There are at least two anecdotes from early twentieth century that could be referenced as remarkable examples of life sciences, in particular generics, multidisciplinary nature.
Fisher, father of population genetics. Huge impact on statistics.
Shannon, from the information theory also defended a short Ph.D. in population genetics.

2.1 60's-90's: Origins: Tukey, Dayhoff – Struggling with computing/memory/storage

Genomics and Data Science are, compared to more established areas, new disciplines. The definition and foundation of genomics are clear. Genomics is defined as the branch of molecular biology that investigates the structure, function, evolution, and mapping of genomes. Thus, the origin of genomics is clearly associated with the first fully sequenced genomes in the 90's. The establishment of genomics as a mature discipline, however, was only achieved after the efforts to draft the human genome. Nevertheless, in the 70's Dayhoff published the first editions of a database, which can be considered one the first efforts to characterize organisms in a, far from complete, but genome-wide fashion \cite{30084940, Dayhoff books}.

In contrast, the data science definition and foundation are not as straightforward. In the 70's, Tukey called for a revisit of statistics and suggested a new focus. Tukey called statisticians to not only be concerned about the theory of modeling data generation but also use the growing amount of data being collected to model more applied problems. This new field, called Data Analysis, was suggested to be a multidisciplinary discipline combining statistics, mathematics, and rudiments of computer science. From its beginning, Tukey suggested that this discipline should not only be focused on modeling but also carefully study data management, processing, and data representation.

2.2 80's-90's: Early genomics era - computing commodity and internet

2.3 2000-today: Large-scale projects, deep learning, cheap storage/cloud

[[Figure 1 - Definition of biological data science (circos), Parallel timeline between DS and Genomics]]

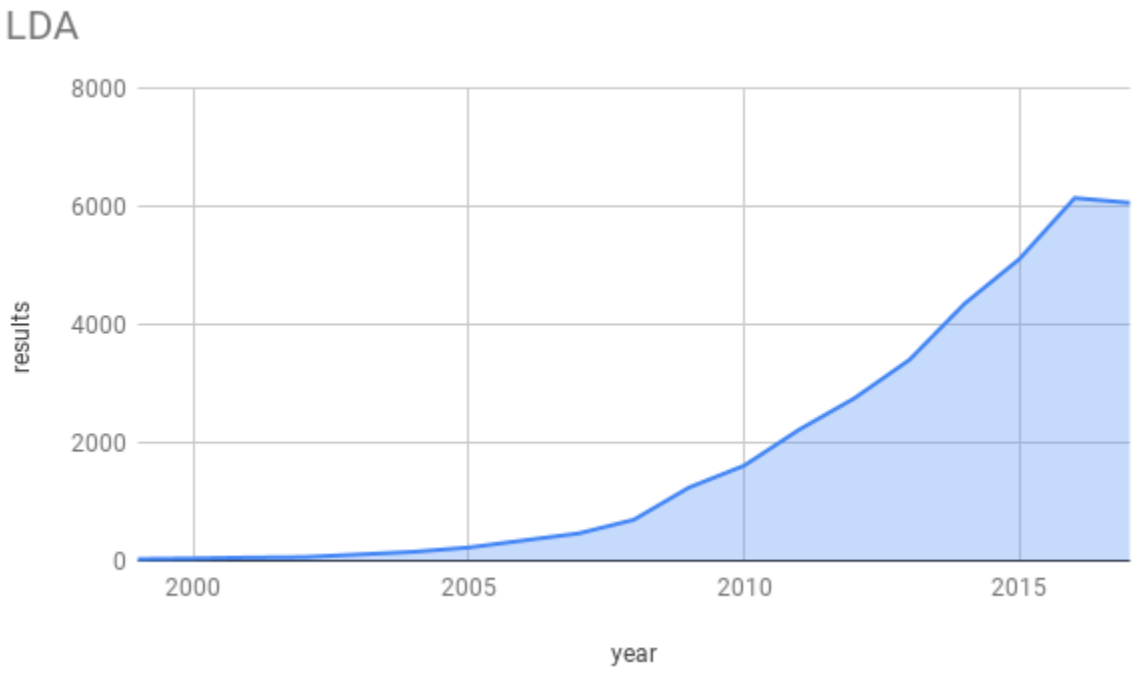
3. The exchanges between DS and Genomics

3.1 The exports

Over the years, the parallel progress of genomics and data science promoted numerous methodological exchanges between both disciplines. In general, the flow of ideas was mostly unilateral. Where genomics would import ideas developed by data scientists and apply to a

narrower set of problems involving molecular biology. However, that's not to say that genomics and bioinformatics have not contributed to many concepts adopted by data science.

One of the most notable examples of methods exported from genomics to data science is the Latent Dirichlet Allocation (LDA) model. LDA uses supervised sample classification to learn the most impactful factors that can predict future classifications. It has been used to, for example, classify the content of text excerpts based on the most informative words. LDA was first described in population genetics \cite{pritchard 2000}. Pritchard conceptualized LDA as a model to infer individuals' population structure based on most informative genetic variants. The concept was noticed by other and generalized using graphs.



Genomics has also contributed to concepts of data visualization in data science. One of the best examples is the Circos plot \cite{19541911}. Circos was initially conceptualized as a circular representation of the linear genomes. On its conception, the tool used to display chromosomal translocations or large synteny regions. As the visualization tool evolved to be more generic it was also used to display highly connected data sets. In particular, circos has been used by the media to display, for example, to track customer behavior, to track political citations, migration patterns.

Finally, genomics has also created statistical methods to detect the association between phenotypes and SNPs. Genome-Wide Association Studies (GWAS) is a model that testes the significance of the correlation between multiple variants and discrete or continuous phenotypes. Since XXX \cite{yale GWAS paper}, with the advent of microarrays, phenotypes could be tested to be associated with hundreds of thousands of single nucleotide variants in the human genome. GWAS is probably one the best representation of models that have an imbalance between the number of variables (millions of SNPs) and the number of samples (individuals). More recently, researchers have been moving from monogenic to polygenic models. The

models could certainly be used data science to models many multiple variables have small but significant effects on traits (i.e. hundreds of ratings/purchases could contribute to describing customer behavior) [[Expand]]

Prediction challenges are yet another example of pioneering concepts born and expanded in genomics that are widely adopted in data science. The Critical Assessment of protein Structure Prediction (CASP) is an organized effort to evaluate and assess the current state of the prediction of protein structures. Every two years, since 1994, a committee of researchers select a group of proteins that i) have their structure described by experiments – usually crystallography; and ii) are target of *in silico* prediction of structure for hundreds of groups around the world. CASP is a community-wide experiment to determine the state of the art in modeling protein structure from amino acid sequence. After predictions are submitted, independent board of assessors compare the prediction models to the experiment evidence and rank those methods. On the most recent instantiation of CASP more than 100 groups submitted more than 50,000 models for 82 targets. Similar initiatives have been seen in biomedicine and system biology, the DREAM challenges, for example, are annual challenges that are broadly distributed across many topics. In the same lines, today there are prediction challenges to define, for example, the state of are of image classification, to speech recognition.

3.2 The non-methodological exports of genomics

The evidence of interchange across genomics and DS is not only limited to methodological exchanges. Genomics has also exported and tested many principles of a successful data-driven discipline. These principles promote and embrace openness and reuse of data (data exhaustion) to a level rarely seen in academia. Here we organize and discuss the tenets of data-driven community.
\cite{Should neuroscience be like genomics?-24904347}

3.2.1. The Bermuda principles.

Initially conceptualized by the leaders of The Human Genome Project (HGP), the Bermuda Principles \cite{1996} is a document that defines how the data produced by the human genome project should be handled. In particular, the principles stated that data generated by the HGP should be publicly released, at most, 24 hours after it was generated. Many factors influenced the elaboration of these principles. First, the fierce competition from the private sector. At the same time, the public effort to sequence the human genome was being threatened by Celera. Second, the early discussions about the sequencing of the human genome were inspired by the Manhattan project \cite{UCSC}, thus, the public effort had a large consortium of research groups collaborating in order to sequence the human genome as soon as possible. Peer-to-peer exchange of data would only cause delays in the process of discovery and assembly. We argue that the early elaboration of the Bermuda principles and a large number of groups using the datasets produced by the project caused a paradigm shift in the genomics community. Today, many researchers and consortia projects adopt

Bermuda-like principles. For example, the 1000genomes \cite{}, ENCODE \cite{} and others release their datasets before publication to allow a broader number of researchers to use public datasets \cite{encodeauthors}

3.2.2. Large-scale repositories and universal file formats

As a result of the Bermuda-like principles, a large amount of sequencing data has been made available to the public. While a lot of datasets are being shared by the researchers themselves, genomics can also count on public data hosts, a huge contrast to other disciplines. The vast majority of sequencing data is hosted by public platforms that, essentially for free, host public and private datasets. The same is observed at the European Bioinformatics Institute (EBI). These central archives are a rich source of data for any genomics projects, contributing to a culture of data reuse and transparency.

The early adoption of data formats has also contributed to the standardization of genomics datasets. The majority of computations in genomics are based in a handful of file formats. For example, FASTA, BED, BAM, VCF, and bigwig, respectively represent sequences, coordinates, alignments, variants and coverage of DNA or amino acid sequences. Moreover, the early adoption and establishment of file formats is the result of the nature of the monolithic human genome. Both DNA and amino acids are discrete variables that are tied together by a reference coordinate system, the human reference genome. The monolithic nature of the genomics also contributes to the standardization of formats and allows the community to quickly test, adapt and switch for other methods that use the same input format.
\cite{3162770}

3.2.4. Community commitment to distribute data and methods openly

Probably the fact that data is open, pushed software code also to be open. Maybe there's a great interaction here, where the open source culture, particularly strong at UCSC, pushed for open data and both things retro-feed themselves. A huge amount of open source available. The recent push from journals to evaluate code quality.

The imports

On the flip side of the coin, over the years, genomics has been importing many concepts from data science. For example, most array and massively parallel sequencing platforms are heavily dependent on image processing algorithms. Improvements chemistry but also in image processing have been associated with better sequencing quality and the cost drop in many sequencing platforms \cite{}.

Another central aspect of genomics, the process of mapping reads to the human reference genome, also relies on a major technique on data science. Fast string processing algorithms. At its foundation protein, pairwise alignment predates DNA sequence alignment. The first implementations were based on Smith-Waterman \cite{1981} and Dynamic programming \cite{FASTA 85; BLAST 90}. These methods were highly reliant on computing power and more memory efficient. With the advancements of other string alignment techniques and the explosion on sequencing throughput genomics experienced had burst in sequence alignment

performance. Since most of the sequencing technologies produced short reads during, the first half of 2010 saw a growth in methods using index techniques. That advancement was accompanied by a drop in memory cost. In particular, we have seen many methods based on burrow-wheeler transformation (BWA, bowtie), De Bruijn graphs (Kallisto, Salmon), and Maximal Mappable Prefix (STAR).

- Multiple testing?
- LASSO regression?

Much more recently as the amount of DNA sequence, especially driven by the large accumulation of orthogonal methods such as functional genomics, and protein structure there has been an influx of deep learning solutions from data science. Very intriguing implementations of deep learning networks are being developed to, for example, predict protein structure \cite{}, classify tumor \cite{} or predict the chances of a patient to develop psychological diseases \cite{}. In particular, deep learning methods have been used to integrate large datasets to annotate and classify DNA sequences, to predict protein structures or even to model the interaction between genetic, transcriptomic and regulatory variations to predict pathways associated with diseases.

4. Framing key issues in Genomics in Data Science terms (The four big V's of Genomics)
4.1 Volume

The growth of genomics data has witnessed an exponential boom during the last 20 years. In Figure 1, we plot growth patterns of data generated from sequencing and microarray experiments in the European Nucleotide Archive (ENA) \ref{ena}. As a result of the swift decrease in sequencing costs \cite{NIHsequencingcostdrop},\cite{costofseq1 and 2}, the total size of the sequencing data surpassed that of microarray counterparts in 2007. The steep growth has remained consistent, a trend that was also observed in NCBI's Sequence Read Archive (SRA) as we show later in this paper. The gap between NGS and microarray data is expected to keep widening during the next decade should current trends be sustained. Interestingly, MS data has also witnessed an exponential growth rate since depositing the first MS dataset in 2009 \cite{the_ena_first_ms_dataset}, what indicates that the massive volume of generated data might transcend the borders of genomics and relates to other biomedical areas such as proteomics.

Current and Expected Growth Patterns across Fields

A consistent pattern of data growth in genomics, climate science, and social science can be observed in Figure 2B showing growth patterns in (logged) total size of deposited data over time in NASA's Earth Science Data Systems Program \ref{nasa}, NIH's Sequence Read Archive (SRA) \ref{sra}, and the Harvard Dataverse \cite{dataverse}, respectively. The rate of growth varies considerably among fields,

however, while climate science and astronomy (need refs) continue to dominate scientific fields *w.r.t* to data generation [[MG: AMT - Atmospheric Measurement Techniques]], genomics seems to have a faster rate during the last decade mainly because of the formation of mass scale consortia projects that leverage this kind of data and rapid advancements in sequencing technologies \refs{}.

Considering the patterns of progress in supercomputing power and towering growth of sequence data generation shown above, we predict that new challenges in data handling and processing will emerge during the next decade. The sum of top 500 supercomputers deployed for R&D is growing in a slower pace than that of genomics data, and further investments in infrastructure, a major part of which is supercomputing stations, are necessary. Furthermore, the growing interest in cloud computing \ref{cloudcomputingrefs} is expected to considerably increase web traffic, for which the infrastructure has been growing steadily and is expected to accommodate such emerging advancements as demonstrated above in IPTraffic plot \ref{IPdata}.

Nature of Different Fields

The total size of data generated by social scientific studies has been significantly smaller than that by studies in other scientific fields, especially natural sciences and medicine. One emblematic example is shown in Figure 1 with the Harvard Dataverse, predominantly comprising of social science studies' datasets. As this (social sciences) pattern seems to be consistent throughout a long stretch of time, i.e. decades in that particular example, we interject that certain fields intrinsically tend to not generate enormous amounts of data. A number of factors might be behind this issue:

- Types of experimental studies (surveys vs sequencing vs imaging)
- Nature of collected data (spreadsheets vs genomes vs images)
- Number of faculty positions and research centers (and consequently researchers) *[to look into this link https://www.humanitiesindicators.org/contentIndicatorDoc.aspx?i=71 + other articles on lack of funding in humanities might help]*
- Funding
- History of practice in a field (social sciences > astronomy > genomics yet order *w.r.t.* Data generation is different)

Data Eruption between Past and Present

Technological advancements have always escalated scientific data generation. Before genomics has been established as a major scientific field, Carlos Jaschek of the Center of Stellar Data and Observation is Strasbourg, France, conducted in 1978 \cite{jascheck1978} the first analysis of data growth in modern astronomy focusing on seven then-major subareas in the field. Not only is the title the author chooses to

describe the data growth phenomenon (“information explosion”) is different from the one commonly used today (“data eruption/deluge”) \ref{find refs for “data eruption/deluge”), but also his predictions underestimated actual values witnessed after the study. However, even when working on a much smaller scale compared to that of today, the study also relied on approximations because of inaccessibility and lack of annotation problems, two problems that considerably persist today. Different growth patterns were also observed among subareas, which defied the common belief back then of homogeneous exponential growth across astronomical subdisciplines. A similar conclusion can be drawn from the aforementioned analysis of data growth patterns among scientific fields we demonstrate in Figure 1.

4.2 Velocity

Increasing velocity of genomics data generation, as shown in Figures 1 and \fig{datagrowthfig} \cite{seq throughput growth} has been a driving factor behind the significant increase in volume. Nevertheless, most of the datasets generated by genomics is more of a static nature. For example, genomics change very little over time and there is little necessity of datasets being processed instantaneously.
New areas in genomics such as disease control, epidemiology are beginning to leverage cheap sequencing to track the spread of viruses and bacteria in the population \cite{ebola, zika}. Similar applications have been in ? microbiome? \cite{??}

4.3 Variety

Continuous emergence of new approaches to studying the genome has led to the availability of a wide variety of (i) data types, (ii) formats, and (iii) functional assays. Currently, the most widely used data types constitute raw Next Generation Sequencing (NGS) data, sequence alignments, annotation sets, and quantitative information derived from experiments and software pipelines (e.g. Chip-Seq files and variant files). Among a multitude of reasons, data variety has necessitated the need for standardization to facilitate the extraction, integration, and sharing of genomics data \cite{comprehensive paper here}. Additionally, the study of DNA-DNA interactions in three-dimensional space has grown further to append more data types to the genomics data variety equation. Primary growth of multi-dimensional, interaction genomics data took shape after the arrival of Hi-C technology in 2009 [V1], which measures interactions between all pair combinations of fragments under study. The technology arrived after a series of advancements that in 2002 with 3-C interaction studies focusing on single pairs of fragments [V2, V3]. (More on assays can be added if need be - list of references in Additional useful references below)

- 0.5 para explosion of *seq methods
- 1 para uniform data of DNA/AA

4.4 Veracity (cleanliness)

Numerous standards have been developed to scrutinize sequencing data accuracy \cite{here + this reference has many other details we can refer to should we need to

elaborate on Veracity}, most popular of which is Phred score in FASTQ files \cite{Cock et al. here}. In particular, the third generation of sequencers which produce longer sequencing reads at the expense of sequence veracity \cite{PBio+Nanopore}.
Algorithmic challenges to mapping sequence reads, calling variants and performing other core tasks in genomic pipelines has also led to discarding parts of datasets considered 'dirty' or of 'unknown' status. More recently, the field of single-cell genomics (SCG) has introduced a new set of challenges at the cellular level with datasets containing as little as 5-10% of the signal in many cases \cite{Nature paper here}. Data imputation is currently a central area in SCG research and is expected to maintain its position as the field evolves \cite{Nature Methods paper here}. More generally, data preprocessing, including data cleaning, has been an integral part of computational biology among other fields and has significant effects on downstream analyses.

5. The tripartite aspects (measurement, mining, and meaning) of genomics as a data science branch.

Genomics is vastly based on the collection of large amounts of data via sensors and the statistical and computational analytics to this data. Different from social sciences, economics, and even consumer datasets, most of the genomics data is strictly derived from sensors. A good analogy for the future of genomics is what has been achieved in weather forecasting, where the collection of large-scale sensor data around the globe and the fusion of this data to physical models has achieved great success. It seems reasonable that a way forward for genomics is to aim at using genetic data for forecast phenotype likelihoods, very similar to weather forecasters. Furthermore, the way that weather forecast is presented in terms of simple, probabilistic models suggests that the public also provides a model of, perhaps, how the underlying statistical predictions made by genomics may be useful in the future.

Genomics is unlike other branches of data science in that in addition to accumulating large amounts of data and mining it, there is also the notion in genomics of connecting the mining of the data to physically based models that describe molecules and biological processes. In that sense, there is a remarkable resemblance to the weather forecast.

Weather forecasting was one of the first applications of large-scale computing in the 1950s. This was done by XXX. At that time this was an abject flop. People tried to predict the weather solely based on physical models. As a result, they quickly found predictions were only correct in a very short degree into the future, mostly because of the great necessity to incorporate initial conditions.
This far from ideal attempt led to the development of the field of nonlinear dynamics, chaos and the coining of the term butterfly effect. However, the subsequent years' weather prediction was dramatically improved and is now weather prediction is a great

success story. Users routinely check their phones and TVs to get a probabilistic prediction about the weather. And these predictions are used in terms of what they are going to wear or how they are going to behave in the next short period of time. The improvement in weather forecasting had to do with synthesizing these physically based models with large-scale data gathering and data mining where the data is coming from satellites, weather balloons, and other sensors. This fusion of large-scale data collection, data mining and then connecting it to physically-based models is a model for the success that one might imagine in genomics. Today, as a community, genomics has displayed great aptitude to data collection and data mining. However, the predictive aspect of genomics, in particular when integrating physical models, is still lacking.

Pushbacks

[[Maybe? I like the idea of including a discussion of some of the pushbacks that genomics has had over the time because it is a data-driven discipline as opposed to an older view of genetics where whoever generated the data is the owner of the data.
The conflict and push backs from genomics being data-driven

-- Research parasitism and the major role played by genomics in the open data, open science scene.
]]
[[http://blogs.nature.com/naturejobs/2017/06/19/ask-not-what-you-can-do-for-open-data-ask-what-open-data-can-do-for-you/

http://blogs.sciencemag.org/pipeline/archives/2016/01/22/attack-of-the-research-parasites

https://www.forbes.com/sites/davidshaywitz/2016/01/21/data-scientists-research-parasites/2/]]
Concerns with privacy
[[GG? to add 1para]]