Data Landscaping and Data Mining

Benefits, challenges & resolutions for medical research

fios GENOMICS

Contents

What is Data Landscaping?

There are vast amounts of data sets in existence for a wealth of different topics. Data landscaping refers to the process of researching and curating data sets for a specific purpose. At Fios Genomics, we perform data landscaping for biological data sets. Our clients tell us what type of biological data they require and for what purpose, then we conduct data landscaping to find the best available public data sets to meet their needs. Once we have conducted data landscaping, we also provide relevance scores for each data set to indicate which ones we believe are most relevant to the research requirement.

What is Data Mining?

Data mining is the process of analysing large data sets to generate new information. It is referred to as data mining as you 'mine' a data set for the valuable information it holds, similar to mining land for the valuable gold it contains. When we conduct data mining at Fios Genomics, we perform bioinformatics analyses of large biological data sets. We do this to reveal the meaningful information within the data that will help our clients to reach their research goals. We can even integrate multiple different data types for analysis.

Why are data landscaping and mining important for medical research?

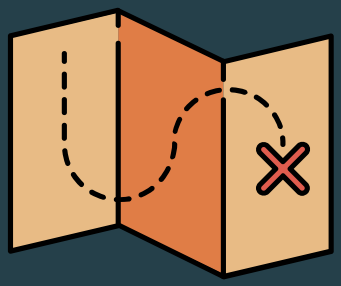
Data landscaping and mining play a significant role in medical research. Since public domain data can be used for hypothesis generation, it can dramatically **reduce the time and costs** associated with wet-lab experimentation and data generation. If suitable public data already exists for a research area of interest, it is much quicker and more economical to use this to support the research goal rather than conducting a new study to find the same information.

Also, in the early stages of medical research, data mining can be used to contextualise a research question. For example, when considering a certain line of enquiry, already available data sets can provide useful insights that will influence the direction of further research.

Researchers can also use data landscaping and mining to support their own medical research. This is possible because public domain data may contain information which can support and validate the results of the medical research they have conducted in-house.

In fact, a further benefit that data landscaping and mining bring to medical research is that they can be used to find suitable alternative data sets which researchers can use for comparison purposes. Medical researchers can draw comparisons between public data and in-house data to support their results or to uncover insights that will further their research or guide it in a new direction.

Challenges and Resolutions in Data Landscaping



One of the key challenges in data landscaping is knowing **where to find the right data sets.** With such a vast amount of biological data available in the public domain, finding and curating suitable data sets for a particular research area can be difficult and time-consuming, especially if you do not already have an idea of where to start your search. This is where a bioinformatics provider can be a great resource. At Fios Genomics, we have expert domain knowledge of available data repositories for 'omics data. We also have a flexible search tool which allows us to rapidly query multiple high quality data repositories for suitable data using filters and search terms related to things like data type, disease, organism and tissue type.

Another challenge in data landscaping is the practical challenge presented by the size of the data sets. Harbouring data generated from thousands or tens of thousands of samples then accessing this data and harmonising the various sources available is a technical challenge from both a computing and scientific perspective. Bioinformatics providers are uniquely equipped to overcome issues associated with analysing large data sets. This is due to their expertise in bioinformatic methods and thanks to their internal computer infrastructures and access to cloud computing.

Challenges and Resolutions in Data Mining

In data mining, sometimes the quality of a data set can present a challenge. This can be for a variety of reasons. For example, a lack of samples can lead to statistical limitations. If you are uncertain about the most suitable public data sets available to support your research, a bioinformatics provider can help. At Fios Genomics, we search a range of databases and return all relevant data sets including, where available, information on overall sample numbers and the number of replicates per condition. We then decide, in conjunction with our clients, which data sets would be most useful for further statistical analysis.

Another challenge in data mining can be difficulties **accessing the database** being mined. Due to the high volumes of traffic they receive, some databases will limit the amount of traffic to their website at any given time. This can make conducting a large number of queries time-consuming. However, database owners usually specify the quickest way to query their databases, as well as any limitations.

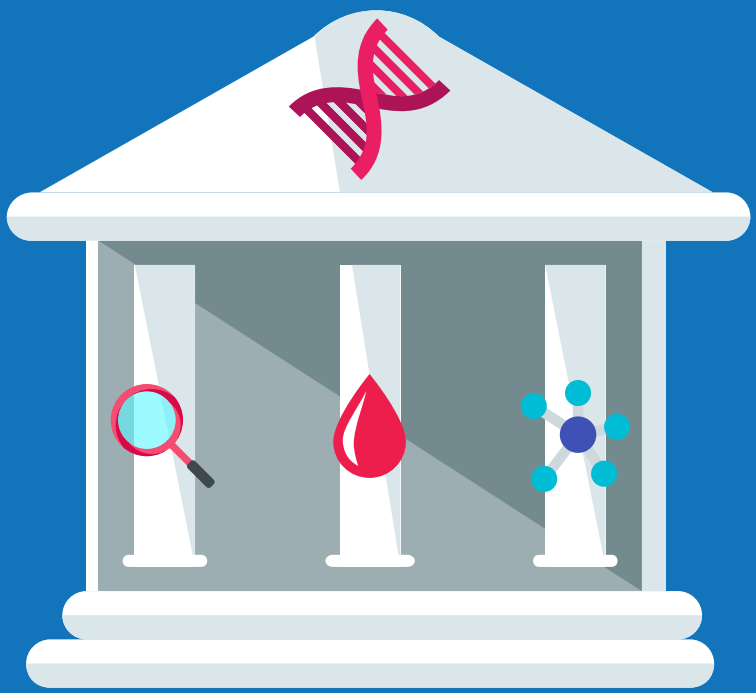
Public data repositories are regularly updated which is a great thing for advancing medical research. However, when updates occur after the data mining process has begun, downstream analyses may be affected.

Fortunately, some databases continue to provide the original data, while also making updated pre-processed data available for public use.

Medical Research and Data Mining: The Future

Data mining is likely to be relied upon for medical research much more heavily over the next few years. Public data sets were already being used by medical researchers prior to the COVID-19 pandemic. However, due to the disruption that COVID-19 caused for clinical trials and data generation, medical researchers have increasingly been turning to publicly available data to further their research. In 2022 and 2023, the reduced data generation caused by COVID-19's impact on wet lab work and the global supply chain will become even more apparent. In response, it is likely that medical researchers will increasing rely on public data sets, such as those available from biobanks and popular databases like the Cancer Cell Line Encyclopedia, to further their research.

What's more, given the vast amount of public domain data available; it is likely that as more researchers become aware of the available data sets and how to access them, public data mining will continue to rise in popularity, regardless of the impact of COVID-19 on data generation.



Our Expertise

At Fios, we can curate data sets from a range of publicly available sources and databases. We can also assist in identifying suitable sources as starting points for data set identification. In many instances, we can automate this as well, for example, quickly search for co-occurrences of specific search terms in journal abstracts, or relevant metadata from Gene Expression Omnibus (GEO) data sets to assist in identifying specific datasets of interest. We regularly use mined data in projects, combining multi-omics data sets. We can also combine your in-house generated data with public domain data.

Data Landscaping and Data Mining are core services at Fios Genomics that we conduct regularly. Some of the most popular data sets that we mine on behalf of clients include:

- The Cancer Genome Atlas (TCGA)
- Cancer Dependency Map (DepMap)
- Cancer Cell Line Encyclopaedia (CCLE)
- Gene Expression Omnibus (GEO)
- European Nucleotide Archive (ENA)
- Expression Atlas
- Database of Immune Cell EQTLs, Expression, Epigenomics (DICE)
- cBioPortal for Cancer Genomics

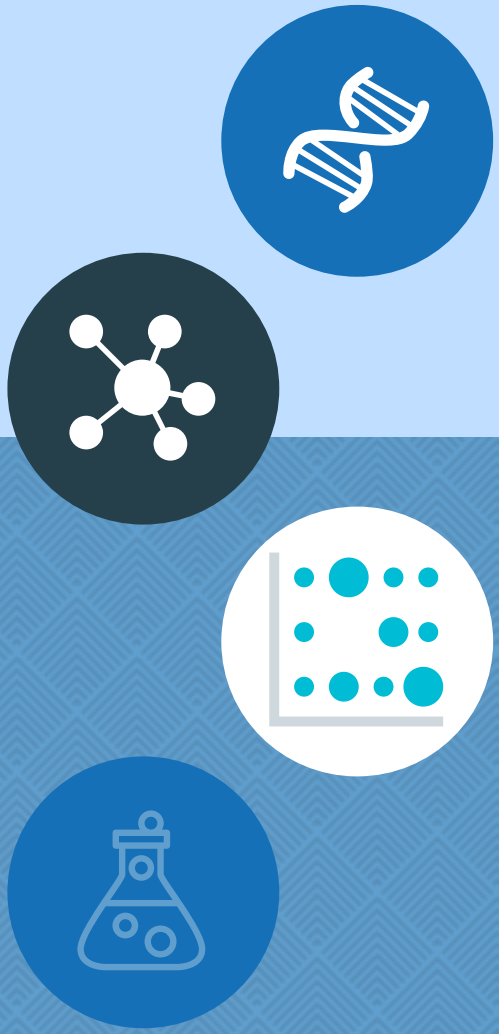
Projects We Have Helped With

Examples of client projects we have supported with our data landscaping/mining services include:

- Development of a gene expression–based prognostic signature for IDH wild-type glioblastoma
- Structural and functional annotation of the porcine immunome
- Coexpression analysis of large cancer data sets provides insight into the cellular phenotypes of the tumour microenvironment
- An expression atlas of human primary cells: inference of gene function from coexpression networks

Access a sample of our bioinformatics reports for data mining

Fios Genomics is a bioinformatic analysis provider helping clients to gain more insight from their research data.



OVERVIEW

With over 15 years of experience in supporting scientists, researchers and bioinformaticians in data analysis, Fios have extensive experience in handling all types of data sets for drug discovery & development, diagnostics, agricultural research, veterinary medicine and applied research across all species.

Our specialised team of bioinformaticians, statisticians, and biologists are able to analyse and interpret any genomic, transcriptomic, proteomic & metabolomic data, independent of the platform used.

learn more

"We have utilized the Bioinformatics team at Fios Genomics for many of our drug discovery projects, as they provide expertise in the analysis of complex bioinformatic datasets. We have been consistently impressed with the rigor of Fios Genomics' work, their communication throughout the projects, and the rapid speed at which they complete their analyses."

- Dr Scott Ribich, Vice President of Biology at Accent Therapeutics