Paxata® | Precision Profile

CASE STUDY

Helping Empower Researchers to Save Lives

Precision Profile

PrecisionProfile is a bioinformatics technology company focused on enabling oncologists and research scientists to thoroughly analyze genomic profiles and to create personalized treatment plans for cancer patients. To advance cancer research and treatment, PrecisionProfile developed a genomic analytics platform that provides cohort analytics for researchers, universities and oncologists by using Paxata and other technologies.

Along with the ability to understand cancer-related data on the Paxata platform, PrecisionProfile offers the computational ability to correlate the attributes of a single patient to one or more relevant cohorts. Such correlation allows researchers to uncover shared disease drivers and enables physicians to develop treatment plans based on protocols that were effective for similarly-affected patients. The platform also has corresponding data from large cohorts of cancer patients.

EXECUTIVE SUMMARY

Empowered oncologists to leverage data in order to recommend personalized cancer treatment plans with the highest probability of success

Reduced the cycle time of a genome clinical study from **1-3 months to 2-8 hours**

Enabled completion of an advanced research project in **2 years instead of 3**

Opportunities and Challenges in Medical Science

With modern advances in medical science such as molecular profiling of patients, data – whether it is genomics, clinical, or patient data – is plentiful; it is also growing exponentially, with over 1.7 million new cancer diagnoses occurring annually in the United States alone.

While this proliferation of useful data is extremely beneficial, properly handling this data is becoming a major issue for many researchers. Because rather than spending their time analyzing the data and formulating how they can leverage it to save lives, many scientists are instead wasting precious hours preparing and massaging the data before analysis can even be conducted. With approximately 20,000 oncologists and a comparable number of pharmaceutical and academic researchers in the world, the problem is widespread.

In many cases, these studies require teams of research, clinical and informatics experts to manage results from a multitude of web-based and internal clinic information systems.

Oftentimes, however, even when various data sources are brought together and correlated, researchers are unable to perform what-if scenarios. Given the non-interactive nature of the data set, researchers must follow a trial-and-error approach: testing a hypothesis against one set of data, moving to another set of data, then testing the next hypothesis, and so on. This is a serial practice and, unfortunately, is not conducive to repeatability.

To provide a platform for scientific analysis and cancer research, PrecisionProfile aspired to design and develop a platform that significantly reduced the time, effort and data science expertise required for such research.

> “In building this platform we had one goal: empower researchers and oncologists to spend a fraction of their time restructuring the data, and the majority of their time on mapping patients to our research database to come up with a treatment plan.
>
> **– DAVE PARKHILL, CEO, PRECISIONPROFILE**

The Complexity and Variety of Research Data

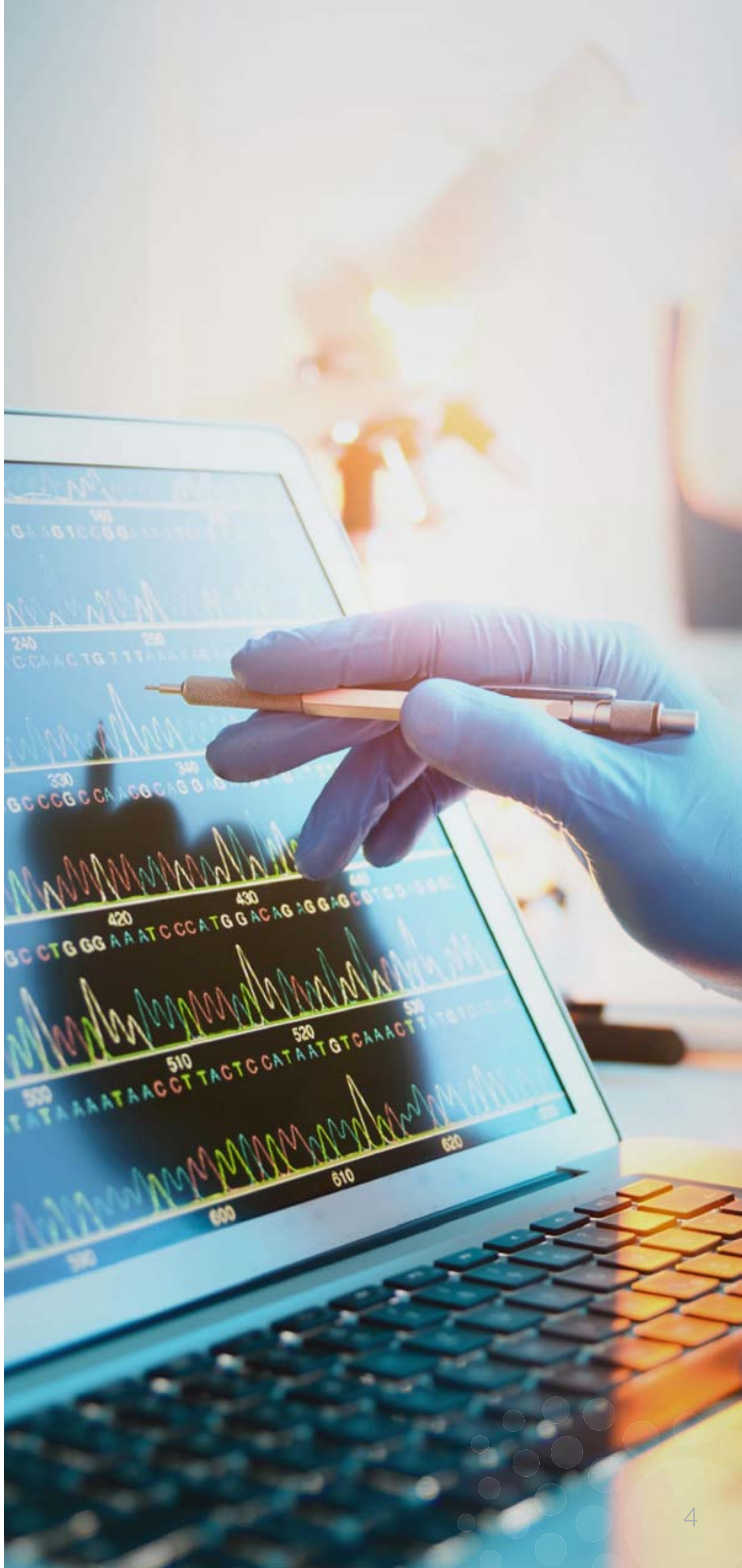
Data source complexity:

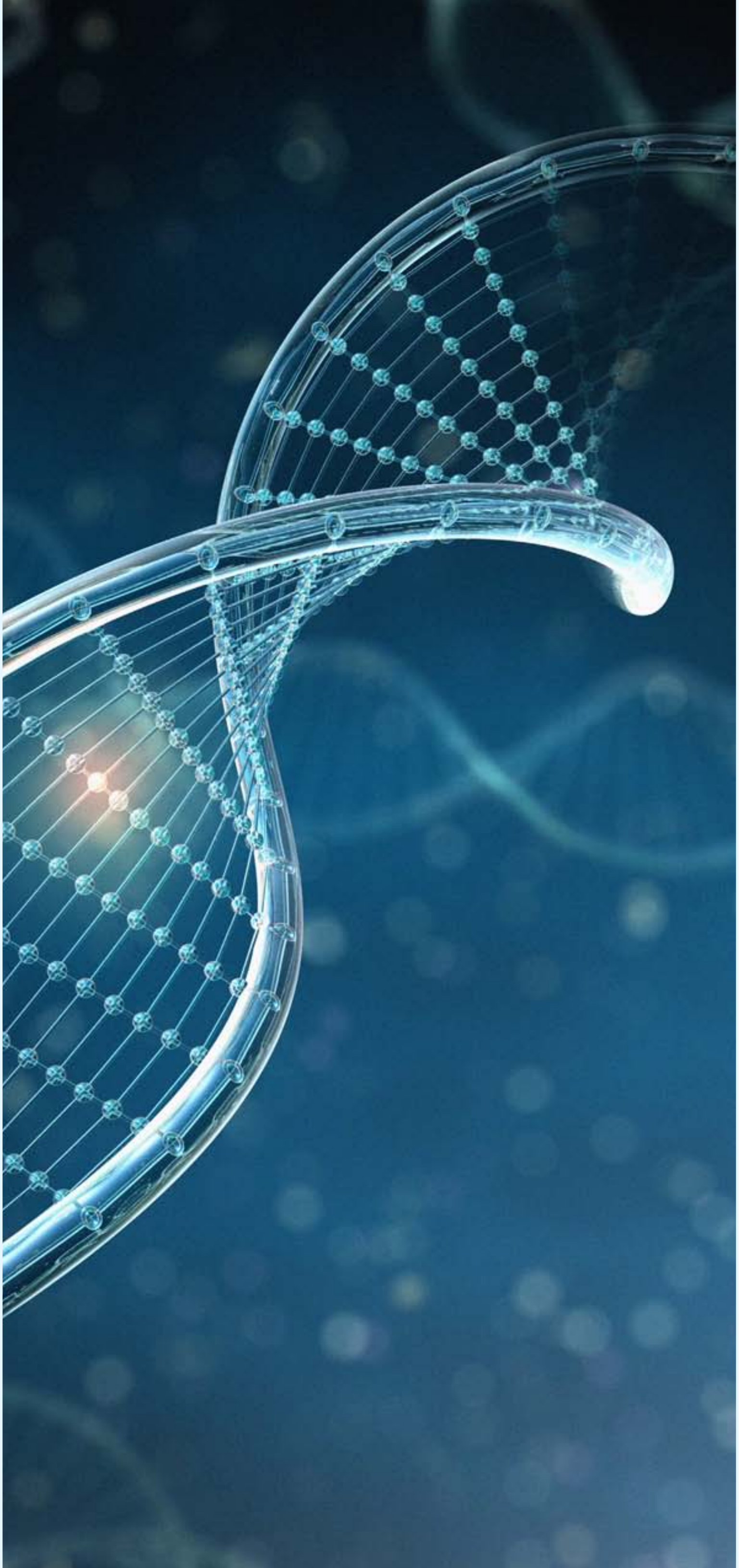
- Cancer genomic data, which is comprised of a molecular profile of patient cells and chromosomes. This data is in VCF files and includes a long list of tags and genomic variants for each patient that need to be parsed. This data is enormous in terms of file space, with one person's DNA taking up as much as 50GB.
- Demographics and clinical data. This data is in XML format.
- Genetics data. This is third party data which can be in diverse formats and may vary for different types of cancers. This data also changes every few months and expands as new discoveries are made.

Researchers traditionally use Excel or Perl scripts to bring these different data sources together, flatten XML structures, parse delimited files, or compare the profile of a given patient to a cohort of other patients – a manual and time-consuming process.

Additionally, while bioinformaticists (biologists with data science skills) can write programs, these hybrid skills are extremely rare. Most researchers and general oncologists don't possess this specific expertise; they are typically skilled instead at biology, medicine and genomic perspectives.

PrecisionProfile aimed to accelerate the research cycle for testing new drugs and combinations of therapies to save lives and revive cancer treatments.





Searching for the Right Solution

THE CRITERIA AND WHY PAXATA

PrecisionProfile initially considered other data prep and ETL solutions, such as those available in the open source community. However, the usability limitations of these tools (particularly for the research and scientist communities), the lack of interactivity with data, and limitations with large data volumes compelled PrecisionProfile to seek an alternative solution.

Why Paxata

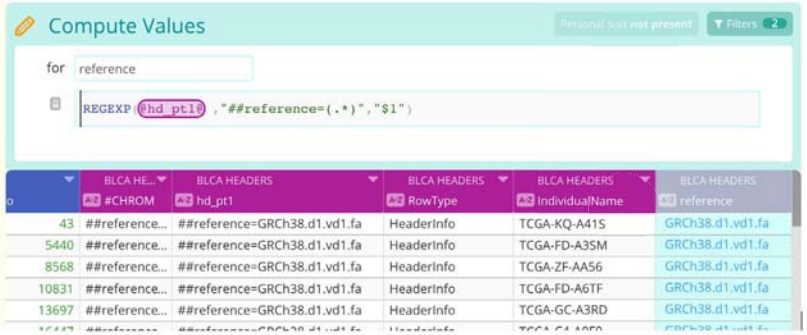
Ultimately, PrecisionProfile selected Paxata as their solution of choice, based on the following capabilities and features found in Paxata:

- Broad set of connectors
- Intuitive tabular view of data, especially turning XML and VCF nested data sets into simple columns and rows
- Scalability with Spark as the underlying architecture, and the ability to host data on thousands of patients, genomes, and clinical data sets
- Cloud-ready platform with elastic scale-out on AWS EC2 and native Amazon S3 connectivity
- Discovery of the data, statistics and interactivity with values
- Self-documenting solution provided a repeatable process to apply logic to new data ingestion
- Data lineage and the ability to show where a certain data set came from, and how it morphed and shaped into another set of data

The Solution

In order to match a new patient to hundreds or even thousands of previously treated patients, thereby quickly identifying similar patients along with their responses to treatments, PrecisionProfile is using Paxata's Self-Service Data Prep application.

With most of the incoming data stored in Amazon S3, local files and FTP sites, PrecisionProfile is using Paxata to consolidate these sources and create a discovery and preparation platform for researchers.

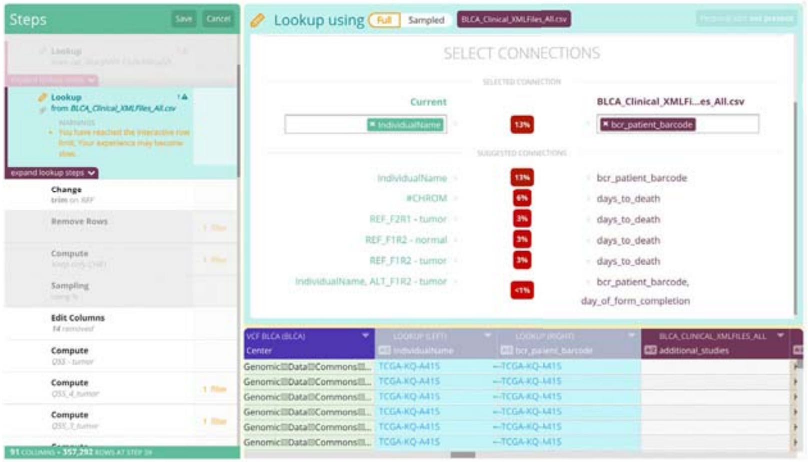


Compute Values

for: reference

```
REGEXP(@hd_pt1@ ,"##reference=(.*)","$1")
```

	BLCA HE... #CHROM	BLCA HEADERS hd_pt1	BLCA HEADERS RowType	BLCA HEADERS IndividualName	BLCA HEADERS reference
43	##reference...	##reference=GRCh38.d1.vd1.fa	HeaderInfo	TCGA-KQ-A41S	GRCh38.d1.vd1.fa
5440	##reference...	##reference=GRCh38.d1.vd1.fa	HeaderInfo	TCGA-FD-A3SM	GRCh38.d1.vd1.fa
8568	##reference...	##reference=GRCh38.d1.vd1.fa	HeaderInfo	TCGA-ZF-AA56	GRCh38.d1.vd1.fa
10831	##reference...	##reference=GRCh38.d1.vd1.fa	HeaderInfo	TCGA-FD-A6TF	GRCh38.d1.vd1.fa
13697	##reference...	##reference=GRCh38.d1.vd1.fa	HeaderInfo	TCGA-GC-A3RD	GRCh38.d1.vd1.fa



Steps

- Lookup from BLCA_Clinical_XMLFiles_All.csv
- Change: trim on REF
- Remove Rows
- Compute
- Sampling
- Edit Columns: 14 removed
- Compute: OSS - tumor
- Compute: OSS_4_tumor
- Compute: OSS_7_tumor

Lookup using Full / Sampled: BLCA_Clinical_XMLFiles_All.csv

SELECT CONNECTIONS

Selected connection: Current: IndividualName — 13% — BLCA_Clinical_XMLFi...es_All.csv: bcr_patient_barcode

Suggested connections:

Current		BLCA_Clinical_XMLFi...es_All.csv
IndividualName	13%	bcr_patient_barcode
#CHROM	6%	days_to_death
REF_F2R1 - tumor	3%	days_to_death
REF_F1R2 - normal	3%	days_to_death
REF_F1R2 - tumor	3%	days_to_death
IndividualName, ALT_F1R2 - tumor	<1%	bcr_patient_barcode, day_of_form_completion

VCF BLCA (BLCA) Center	LOOKUP (LEFT) IndividualName	LOOKUP (RIGHT) bcr_patient_barcode	BLCA_CLINICAL_XMLFILES_ALL additional_studies
Genomic Data Commons ...	TCGA-KQ-A41S	TCGA-KQ-A41S	
Genomic Data Commons ...	TCGA-KQ-A41S	TCGA-KQ-A41S	
Genomic Data Commons ...	TCGA-KQ-A41S	TCGA-KQ-A41S	
Genomic Data Commons ...	TCGA-KQ-A41S	TCGA-KQ-A41S	
Genomic Data Commons ...	TCGA-KQ-A41S	TCGA-KQ-A41S	
Genomic Data Commons ...	TCGA-KQ-A41S	TCGA-KQ-A41S	

IMPLEMENTATION STEPS

1

Import data into Paxata, including 24 chromosome datasets (from an FTP site), 24 VCF datasets (from S3), and 1 clinical dataset (from a local drive)

2

Concatenate large files into a single dataset

3

Reduce repetitive columns

4

Conduct shaping, parsing and other data prep steps on the consolidated datasets

5

Join the desired datasets together to create approximately 40 million rows

AnswerSets™ helps researchers to quickly and easily identify the specific locations of genes and pathways where a gene could be very detrimental to the patient, as well as clinical and drug outcomes that responded positively for other patients within the same cohorts and demographics

> “With Paxata, PrecisionProfile is now able to save valuable time for researchers and oncologists, so they can save lives.
>
> **– JOE KASPRZAK, VP OF SERVICES, PRECISIONPROFILE**

> “Being able to discover new information quicker has made a significant impact on many lives. For some patients, timely, targeted treatment may be the difference between life and death.
>
> **– DAVE PARKHILL, CEO, PRECISIONPROFILE**

The Results

INSPIRED BY INFORMATION, DRIVEN TO SAVE LIVES

Today, PrecisionProfile utilizes Paxata to empower researchers to understand and investigate the clinical and genomic profiles of patients. With Paxata, making the data usable, an oncologist can compare each patient with large cohorts of other patients and target third-party references and knowledge databases to devise the best treatment strategy with the greatest chance to succeed.

With Paxata, a genome clinical study that typically takes about 1-3 months has been reduced to 2-8 hours.

Additionally, with the new solution powered by Paxata, researchers, oncologists and academic professors now have a platform for collaboration and standardization. Because many sequences and processes in research are repeatable, scientists can now replicate a specific sequence for new sets of demographics or drugs as new cancer types, patients, and clinical data become available, thanks to Paxata.

In another example, a PrecisionProfile client foresaw a research project that was originally envisaged to be a 3-year project to instead be completed in only 2 years.

Paxata’s data prep, shaping, integration and collaboration capabilities, combined with PrecisionProfile’s proprietary “Patients Like This One” cohort analysis algorithm, have moved the company beyond competitor platforms.

Future Plans

Today, the solution is targeted primarily towards pharmaceutical, genomic, academic researchers, and bioinformaticists. In the future, PrecisionProfile plans to extend the solution to data stewards who prep the data for clinical studies and to oncologists who need decision support systems for their patient research and recommendations.

Additionally, PrecisionProfile envisions that these researchers and practitioners will want to bring and add new data sets to supplement what is provided out-of-the-box.

PrecisionProfile also intends to fully white-label Paxata as its own product. While today it currently offers only a hosted version in AWS, PrecisionProfile intends to offer customer Virtual Private Cloud (VPC) and on-premise versions in response to client requests.

> “
>
> The PrecisionProfile platform enables us to more quickly understand underlying genomic alterations. This will change the drug discovery process from one that is largely based on trial-and-error to one that is specific, targeted and more easily assessed in clinical trials.
>
> It will also enable me to be more informed in my role as an oncologist, by allowing me to fully understand and investigate the clinical and genomic profiles of my patients.
>
> **– DR. DAN THEODORESCU M.D., PHD DIRECTOR,**
> **UNIVERSITY OF COLORADO CANCER CENTER**

Companies around the globe rely on Paxata to get smart about information. Paxata is the pioneer that intelligently empowers all business consumers to transform raw data into ready information, instantly and automatically, with an enterprise-grade, self-service data preparation application and machine learning platform. Our Adaptive Information Platform weaves data into an information fabric from any source and any cloud to create trusted insights. Business consumers use clicks, not code to achieve results in minutes, not months. With Paxata, Be an Information Inspired Business.

Paxata is headquartered in Redwood City, California with offices in New York, Ohio, Washington D.C., and Singapore.

Paxata®

Paxata Headquarters 305 Walnut Street Redwood City, CA 94063 1-855-9-PAXATA paxata.com