evotec

DRUG DISCOVERY UPDATES #13

DDup

PanHunter

Evotec's next generation
multi-omics data analysis platform

Generating multi-omics insights

Interview with *John Szilagy*

Case study

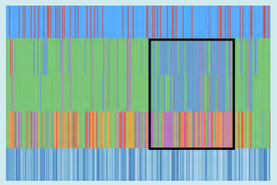
User interviews

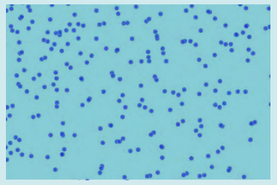
2022/10

evotec — DDup

What's inside this issue?

DDUP #13 CONTENTS





www.evotec.com/en/innovate/ddup

DEAR FRIENDS

of Evotec

A message from Evotec CEO
Dr Werner Lanthaler

Welcome to this thirteenth issue of DDup, an Evotec publication providing you with more insights into the company and its capabilities. This edition features our software as a service product PanHunter, which will drive enhanced disease understanding and accelerates to ultimately fight diseases. This multi-omics data analysis platform enables our partners and ourselves to achieve outstanding research insights and increase the probability of success throughout the whole drug discovery process.

With hundreds of dedicated researchers and PanHunter users we have a long and successful history of applying and evolving advanced systems to support our research throughout the whole value chain from target ID towards clinical studies. Our collaborations gave us vast experience in deriving disease understanding, molecular patient profiles and qualify compounds early. Such models cover challenges like tox prediction and reach out towards patient stratification and diagnostics. All these efforts converge into the upcoming release of our software-as-a-service platform that empowers our partners to run large scale multi-omics data analysis projects at ease and will give their researchers a highly interactive and powerful tool to excel further.

Thank you for reading this latest edition of DDup – we hope you found it of interest. We welcome your thoughts and input, and hopefully we will get the opportunity to collaborate in the exiting field of multi-omics, AI-driven medicine of the future.

Yours sincerely,
for the management of Evotec
Werner Lanthaler,
CEO of Evotec SE

CHAPTER 01

INTRODUCTION

PanHunter: *Generating multi-omics insights*

Evotec strongly believes in the power of omics-driven drug discovery. The unbiased nature and magnitude of insight enabled by multi-omics will allow us to tackle the challenges ahead. Healthcare and optimal medical treatment of patients will require more sophisticated, higher efficient and highly selective drugs.

If we want to uncover the full complexity of diseases, we must understand the underlying molecular processes driving them and decipher the reasons for their various phenotypes in a diverse range of patients. This will pave the way for the development of new and precise cures and how to better identify the correct patients to treat. Today, we can observe and explore the processes on a molecular level using unbiased genomics, transcriptomics, proteomics, and metabolomics to provide unprecedented insights. Each layer of understanding alone is already a milestone towards a more general understanding. Interconnecting these datasets in a multi-omics analysis, together with relevant clinical data, however, will reveal the whole spectrum of influences on the intracellular processes by diseases and treatments.

This drive for deep understanding motivates all steps we do, every step is therefore backed-up by extensive data collection and analysis, and we are in a great position for this approach as we can combine the various strengths we built-up with our partners over the last decade: from molecular patient databases, we derive disease profiles that we translate into - for example – iPSC-enabled disease models. Against those models, we can develop and screen very large amounts of compounds and really utilise the industry leading high-throughput platforms we have created in our EVO*panOmics* projects to generate the multi-omics data we need.

However, generating deep and high-quality omics data is just the starting point. Alongside advancing our data generation and interpretation capabilities, we have been working towards democratising access to data exploration for all scientists and give all researchers the bioinformatics skill set to handle and analyse high-dimensional omics data.

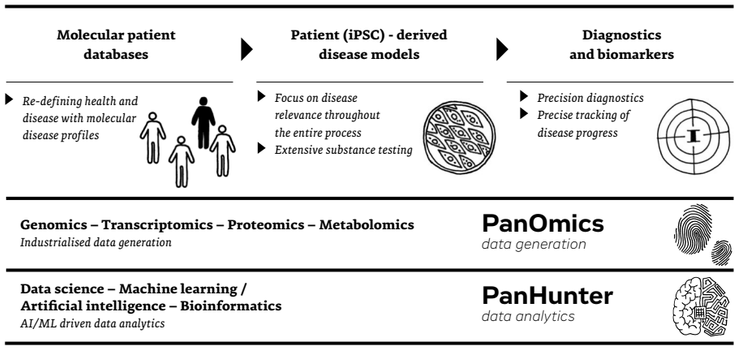


Figure 1: *Evotec's omics-driven and patient-based research platforms*

PanHunter is structured into apps, each with a specific focus, covering a broad spectrum of different analysis tasks. All this is built on top of a curated set of published and peer-reviewed, or in-house developed, robust algorithms.

Figure 2: *App-based web interface*

To empower our scientists to cope with the overwhelmingly increasing amount of data and to reduce the complexity of data analysis and interpretation, we built PanHunter, our versatile and interactive multi-omics data analysis platform. It provides a plethora of tools, visualisations, and reference information readily available. Via its easily accessible web interface, it allows our scientists to immerse with the data, analyse them faster and more focused, visualise them easier, put them into context, and derive deep and meaningful interpretations from them to make well-informed decisions.

PanHunter can be used through its graphical user interface in any modern web browser. This interactive interface provides immediate feedback to user interaction and reduces waiting times between setting up and interpreting analyses, especially compared to the classical pipelining approach.

This allows our users to focus on the results rather than the process of handling the data. Additionally, the modular structure of PanHunter allows quick and easy adaption to new tasks and necessity arising from our internal research projects or external partner requests.

The main motivation when designing PanHunter was to reduce existing obstacles for data access as much as possible and diminish the overhead and repetitive work usually associated with omics data analysis. The result is that both, our bioinformaticians as well as our lab scientists can focus more on data interpretation and less on data processing. However, using one general platform for omics data analysis comes with additional advantages: standardisation, reproducibility, and collaboration.

A standardised data processing workflow, which is an integral

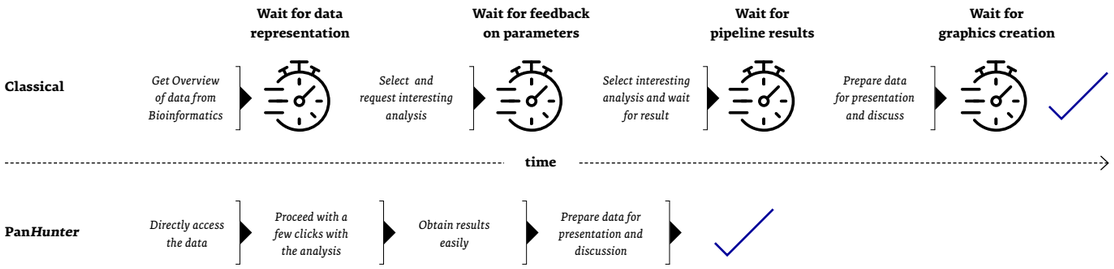


Figure3: *Comparison of a biologist's journey from data presentation towards result generation*

part of PanHunter, allows for easy re-purposing of already existing datasets in other contexts, which eventually increases the value of each individual experiment. It serves therefore as an exchange and access platform for the increasing number of linkable datasets at Evotec and its partners. Furthermore, our software generates reproducible and comparable results, since all intermediate steps, algorithms, and parameters are tracked automatically. And finally, all this is fully equipped for global collaboration of many scientists in larger projects, as simultaneous access to the same data base and immediate sharing of generated results is build right into the core of PanHunter.

After all, omics data alone only provide one part of the picture. To unfold their full potential, our software provides everything that is necessary to put the omics data into perspective by accessing additional meta information, reference data, chemical and structural information for compounds, and even clinical information for projects containing human samples.

In most situations, this vast amount of additional information is automatically associated with the experimental data. This allows the user to easily pull, e.g., gene or protein information from associated databases or run statistical tests to identify metabolic pathways or protein interaction networks that are significantly regulated within the tested data. Such networks can reveal the underlying molecular mechanisms for a disease or drug treatment.

Whenever performing such statistical methods, PanHunter applies well-selected default parameters, while being always transparent about them and providing the option to customise. This allows us to serve all scientists from a broad experience spectrum: new users to the field of omics data analysis will be guided and can rely on the default settings while experienced users can customise almost any small details of an analysis.

While the user gets empowered to perform a plethora of different analysis via the GUI of PanHunter

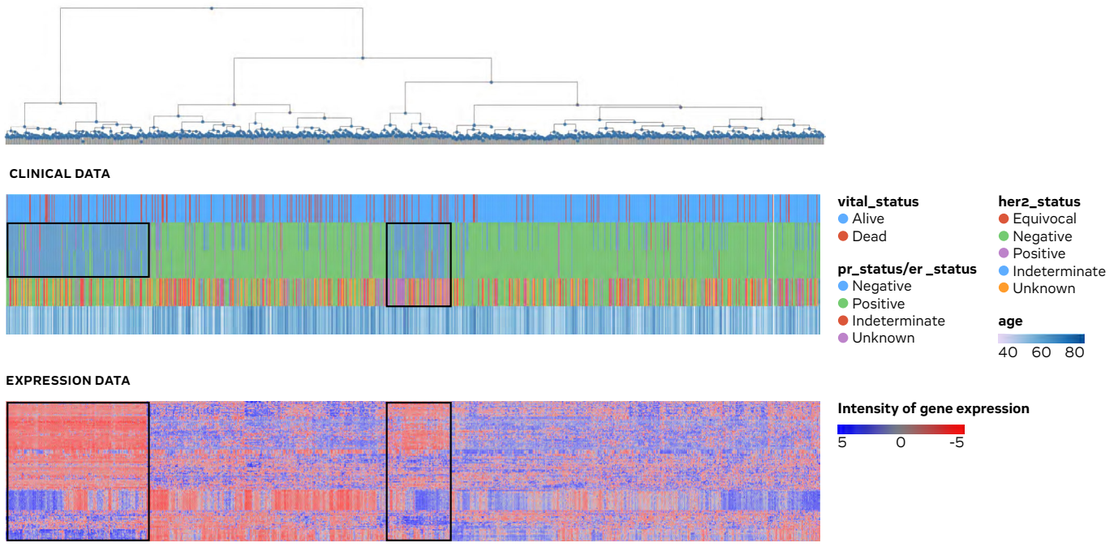


Figure 4: *Global view and correlation of clinical vs. transcriptomic data: tumour tissue samples from the public breast cancer cohort of The Cancer Genome Atlas (TCGA) are shown in PanHunter and allow to associate gene expression with clinical parameters. Hereby the user is able to identify potential connections in a global view on gene expression.*

that requires no coding or scripting skills, all other aspects, like managing the data and carrying out the analysis processes, are handled in the background. The omics data are stored in a fast yet secure data layer. In addition to the experimental and meta data provided by the user, this data layer handles and provides access to all additional information already mentioned above.

The processing, analysis, and interpretation of omics data involves a series of sequenctial steps, all of which are covered by PanHunter. As soon as new omics data are generated, e.g., via mRNA library sequencing using next-generation sequencing (NGS), the data need to be pre-processed to generate a so-called feature abundance table.

"Omics-driven drug discovery is the corner stone of disease understanding on the molecular level and the fast lane to cures!"
Cord Dohrmann, CSO of Evotec SE

In case of RNA-Seq, this table contains for all samples the abundance/activity information for all genes. For other omics types, e.g., the genotype for all single nucleotide polymorphisms (SNPs) of all individuals is determined (genomics), the relative abundance of proteins for all samples is quantified (proteomics), or the amount of certain metabolites is measured (metabolomics).

For all these quantification steps, PanHunter has its own, versatile, and robust pre-processing pipeline that can ingest the raw data and information from a variety of locations. Additionally, the software provides the necessary assistance tools to the user, to supplement the omics data with the experimental

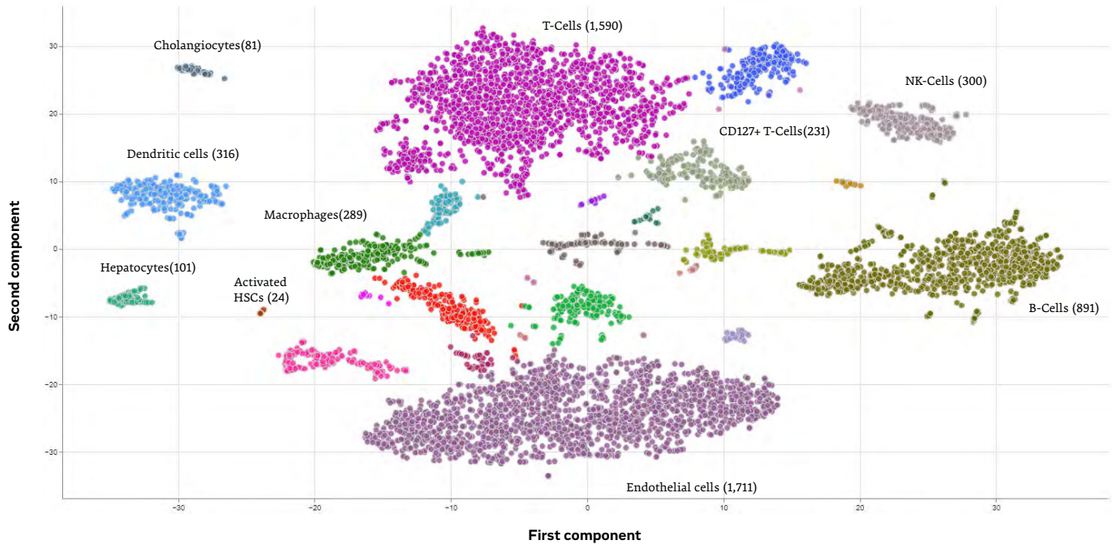


Figure 5: *Illustration of automated AI-driven cell type annotation for clusters of liver scRNAseq cells*

meta data (e.g., which tissue type, treatment, or time point belongs to which sample). Once all this information is integrated and available within PanHunter, the users can use them for differential analysis, to identify significantly regulated features (e.g., up- or downregulated genes, proteins, or metabolites), or statistical correlation testing, for example when trying to identify molecular marker genes based on the abundance of certain clinical parameters or survival rates. Moreover, PanHunter offers dedicated tools for specific omics data types, like single-cell or spatial transcriptomics. For example, it is easily possible for single-cell datasets to perform the key tasks of cell type clustering and annotation, with or without the assistance of machine-learning algorithms, or for spatial transcriptomics to associate a given cell's gene activity information with the exact location on its originating histological slice.

After all, PanHunter became the central platform for inter-disciplinary omics data analysis in a large variety of internal and external projects and keeps growing constantly.

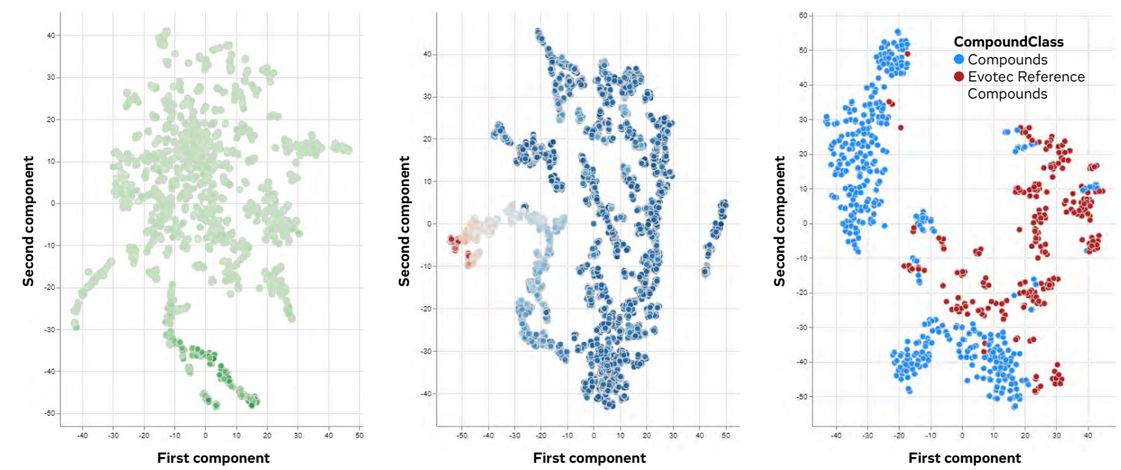


Figure 6: *A cross-omics analysis of samples linked by meta-data. The selection of a sample subset in one cluster, e.g., transcriptomics, is then highlighted in the other graphs as well and enables a deeper understanding of the underlying biological process.*

CHAPTER 02

5 MINUTES WITH

Dr John Szilagyi on PanHunter

Bristol Myers Squibb™

Dr John Szilagyi

Research scientist in the non-clinical investigative toxicology department at Bristol Meyers Squibb (BMS)

John is a research scientist in the non-clinical investigative toxicology department at Bristol Meyers Squibb (BMS). His area of focus is placental toxicology. He has a broad expertise in drug transport, enzyme and cell-based assays. John obtained his PhD in toxicology from Rutgers University in New Brunswick. He joined BMS in June 2020 after a postdoctoral research stay in the toxicology and environmental medicine department of the University of North Carolina at Chapel Hill.

In what project do you use PanHunter?

John: My active collaboration with Evotec involves high-content screening and bioinformatics investigations of differentiating pluripotent stem cells treated with protein-degrading teratogens, such as thalidomide. The overall goal of this work is to better understand the mechanisms behind stem cell differentiation in embryonic limb development and identify the sensitive pathways that can disrupt that process. In doing so, we hope to better inform drug discovery and development to improve the development strategy for safer and more effective medicines.

How does PanHunter leverage research outcomes at BMS?

John: PanHunter is a great platform to ergonomically see large bioinformatics datasets at multiple levels, allowing a researcher to easily build the context behind the observed biological changes. Notably in our current investigations, PanHunter allows the direct identification of high priority pathways during stem cell differentiation that can lead to

teratogenicity. These data can then be integrated into drug discovery to prioritise safer treatments for patients.

What is the impact of PanHunter in your daily business?

John: I spend 80% of my working time in the lab developing assays to assess underlying mechanisms of unexpected findings or problems encountered by other colleagues. For this purpose, we often use omics analyses. PanHunter speeds up and eases the analysis of these multi-omics datasets. More particular, PanHunter comes with a broad spectrum of tools to help me to really understand the mechanisms behind these unexpected findings.

What are the main points where PanHunter improved or accelerated your analysis?

John: The experimental setup of our multi-omics study is quite complex. Hence, there are many different angles to look at and analyse the data. I joked that I would need a post-doc to look at the data for three years to fully understand the findings. The interactivity of PanHunter makes it very easy for me dive into, sorting and pulling out data. My ultimate goal is to understand the mode-of-action of the compounds. It is very easy to set up differential analysis in PanHunter. To analyse the outcomes of these differential analyses I started with Euler and Venn diagrams to understand differences and similarities in the mechanisms of the compounds. Another very useful tool is the network visualisation that is based on BioGRID. This helped me a lot to identify the proteins that have physical interactions with the target protein.

What could you do faster/better because you had PanHunter at hand?

John: Definitively the identification of proteins that have physical interactions with the target protein. As well as the detection of pathways that are affected by the compounds. This helped me a lot to understand the mechanism behind the compounds. A thing that I could do much faster is the hierarchical clustering of abundances of proteins/transcripts based on predefined features lists. The predefined feature lists originate either from public databases like GO or from earlier research that I did.

Thank you for the interview.

CHAPTER 03

CASE STUDY

Rapid candidate biomarker discovery with PanHunter

According to the WHO Global Health estimates, colon and rectum cancers are among the top 10 causes of death in western countries, amounting to over 900,000 deaths every year, globally. As with many forms of cancer, early detection and clear diagnosis are key to apply the right therapy to achieving full recovery and the best possible life expectancy. Hence, identifying and linking molecular markers unambiguously to a specific type of cancer is a crucial achievement towards better and affordable point-of-care diagnostics in preventive medicine, as molecular markers can detect diseases long before the appearance of macroscopic symptoms and often also allow for a more targeted diagnosis.

"The Cancer Genome Atlas" (TCGA) collected molecular omics data from cancer patients in combination with clinical information to facilitate general cancer research including, e.g., candidate biomarker discovery. For this case study, the patient cohort for colon adenocarcinoma ('TCGA_COAD') will be used in PanHunter.

After an initial look at clinical information (e.g., the distribution of samples in the cohort), the next step is the exploration of the molecular data. PanHunter's New Comparisons app allows to explore the omics data, in this case transcriptomes, in a very straightforward way: a very common and easily accessible visualisation for this kind of exploration is a 2D plot of the dimensionality reduction (here, using the t-SNE algorithm) of the molecular expression data *(Figure 7)*.

In this plot, samples with similar gene activity are located close to each other, while those with larger differences are located further apart.

As a result, a strong correlation of the tissue definition with the two major clusters was immediately visible, indicating a significant difference in gene activity between both tissue types. To focus on tumorous samples only, non-tumours samples can be deselected and/or the dataset reduced by free selection.

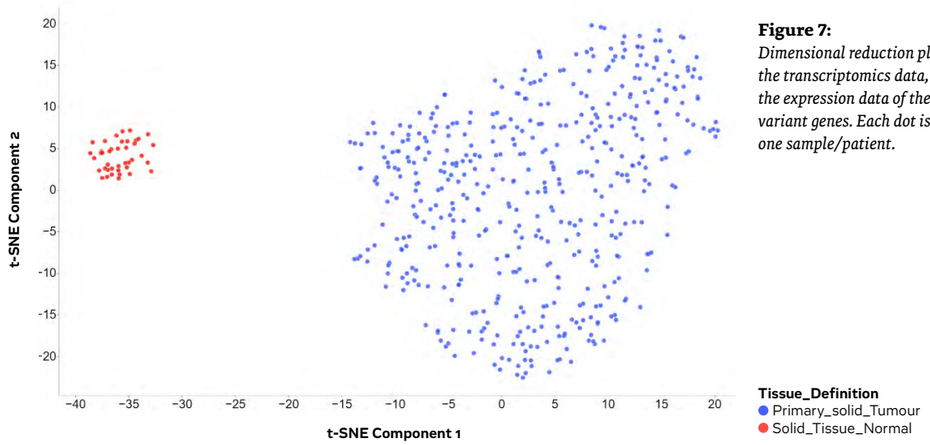


Figure 7:
Dimensional reduction plot of the transcriptomics data, based on the expression data of the 500 most variant genes. Each dot is representing one sample/patient.

A new 2D plot, only using the new sample subset, was generated immediately, showing an evenly distributed cloud of samples without any obvious clusters (not shown). To assist research, PanHunter allows to search for driving factors of sample distribution: Exploratory Analysis. It can be used to search for parameters in the metadata of the samples, or even in the expression values of the genes themselves, that are responsible for the distribution of the samples across the 2D plot, using statistical methods. As a result, a list of possible genes was generated, showing REG4 on top *(Figure 8)*.

Feature overview

ID	Symbol	StatScore	Correlation
text filter	text filter	number filter	number filter
ENSG00000134193	REG4	100	0.38
ENSG00000101470	TNNC2	100	0.35
ENSG00000211890	IGHA2	100	0.34
ENSG00000103485	QPRT	100	0.33
ENSG00000170835	CEL	100	0.32
ENSG00000179603	GRM8	100	0.32
ENSG00000196188	CTSE	100	0.31
ENSG00000211895	IGHA1	100	0.3

Figure 8: *Result from exploratory analysis, listing several genes which might be driving the distribution of samples within the 2D plot of Figure 7.*

To examine this result in a bit more detail, the coloration of the sample plot was changed to indicate the transcriptional intensity of that gene *(Figure 9)*.

The resulting plot showed a clear gradient within the large cloud of samples, from low (light green) to high (dark green) transcriptional activity of REG4. The correlation of REG4 expression with the overall clustering of the patients was a first hint that the gene could be a potential marker gene.

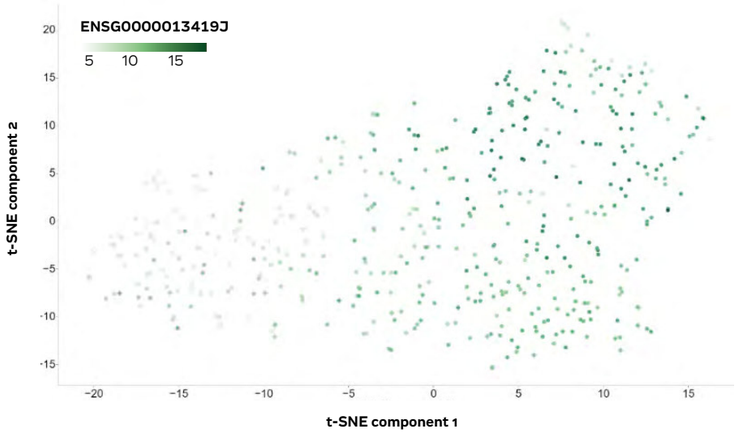


Figure 9: *t-SNE plot coloured according to transcription intensity of the REG4 gene.*

A comparison between the activity of REG4 and survival of patients was performed in the Patient Data app: First, the patient cohort was divided into two sub-cohorts via a respective selection in the graphical user interface (GUI). Subsequently, one sub-cohort containing the lower 50% (red) and another one containing the upper 50% quantile (cyan) with respect to their REG4 expression were generated. As a result, a Kaplan-Meier plot was generated, indicating patient survival for both sub-cohorts. In the plot, a slight difference between the sub-cohorts was already visible: patients with high REG4 expression showed a slightly better survival rate. However, PanHunter pointed out that the correlation observed here was not significant p-value = 0.137

As there are often gender-specific differences in cancer, both sub-cohorts were filtered subsequently, again via a quick GUI setting to leave only male patients selected. As a result, the survival plot *(Figure 10)* was updated on-the-fly and a much stronger difference for high vs. low REG4 expressing patients became visible, which was now also statistically significant (p-value = 0.0114). Hence, it was concluded that REG4 might indeed be a suitable biomarker to indicate the survival prognosis of colon adenocarcinoma patients and might help during diagnosis and assessing treatment responders and non-responder to drug treatment (which has also been reported in the literature).

This case-study is one of several case studies that are also available as demo videos for PanHunter on YouTube: youtu.be/qdz1lVdC9Io.

The entire procedure from start to finish would take an average user with the necessary disease-specific knowledge only a couple of hours when using PanHunter. All shown plots are taken directly from PanHunter, were generated on-the-fly, and no external/ additional tools were needed.

Pval log-rank: 0.0114; Pval CoxPH Wald: 0.0711

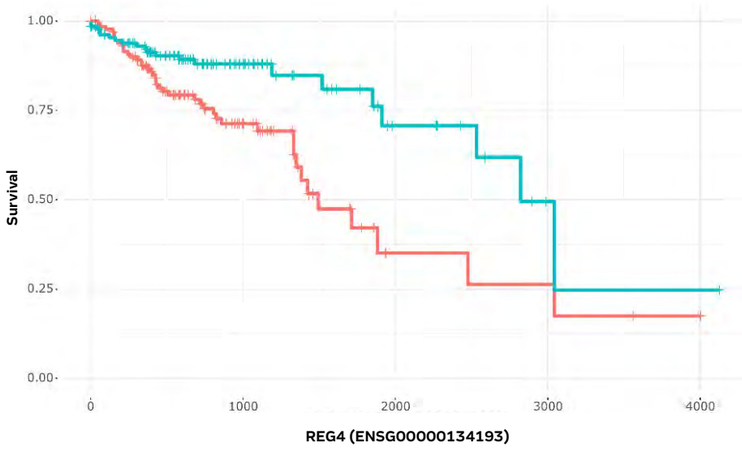


Figure 10: *Kaplan-Meier-Plot as Figure X for male patients. Here, we see a significant (p-value = 0.00143) correlation between low REG4 transcription (red curve) and poor survival.*

CHAPTER 04

USER PERSPECTIVES

How is PanHunter making an impact on your research here at Evotec?

... for wet lab biologists

PHILIPP SKROBLIN
Senior Scientist Metabolic Disease

My daily work revolves around data generation and analysis in NURTuRE: in this project we achieved to generate RNA-Seq data from more than 4,000 kidney disease patient blood samples with matching genotyping data, which is unparalleled in the kidney disease area. In addition, we have now sequenced a total of 600 patient kidney biopsies from NURTuRE and the Salford Kidney Study, providing an amazing resource for target identification for various strategic alliances. In the last 5 years of nearly daily use, PanHunter developed constantly to keep up with the increasing complexity and variety of data analysis tasks in NURTuRE.
It is not only the home to all these datasets, but it allows us to dive into the data facilitating research.

WINFRIED WUNDERLICH
Group and Project Leader Metabolic Disease

PanHunter has supported my team and me in several areas of drug discovery, but mostly on target ID and target validation, where using the various 'drill down' apps allow to generate a customised display of expression levels for individual genes as well as groups of related genes in just a few clicks.
This enables following-up directly with assays towards specific cell lines or tissues in which the target gene is expressed. A next step is the understanding of mechanisms-of-action. Here, PanHunter's linkage to the Pathway Mapping and Network Visualisation made it easy even for non-bioinformaticians to perform the automated comparison of transcriptomic datasets which enabled the identification of pathways that were affected, e.g., by compound treatments or genetic manipulations. Finally, translatability is established by the 'Compare Top Tables'-functionality. The app allows to effortlessly compare disease-associated signatures between a specific human disease and the respective animal model(s) to identify the suitability of a model to study a particular pathway in the context of a disease.

NURTURE
National Unified Renal Translational Research Enterprise

NURTuRE is just one of many large molecular patient cohort projects at Evotec. It contains clinical data and omics data such as transcriptomics, genomics, proteomics, and metabolomics from more than 10,000 chronic kidney disease (CKD) patients and 800 idiopathic nephrotic syndrome (INS) patients. All data is provided to the researchers via the PanHunter interface, all steps from data cleaning to deep insight generation are performed in the system.

PanHunter enables an integtrated analysis of omics and clinical data - Visualisation of NURTuRE data

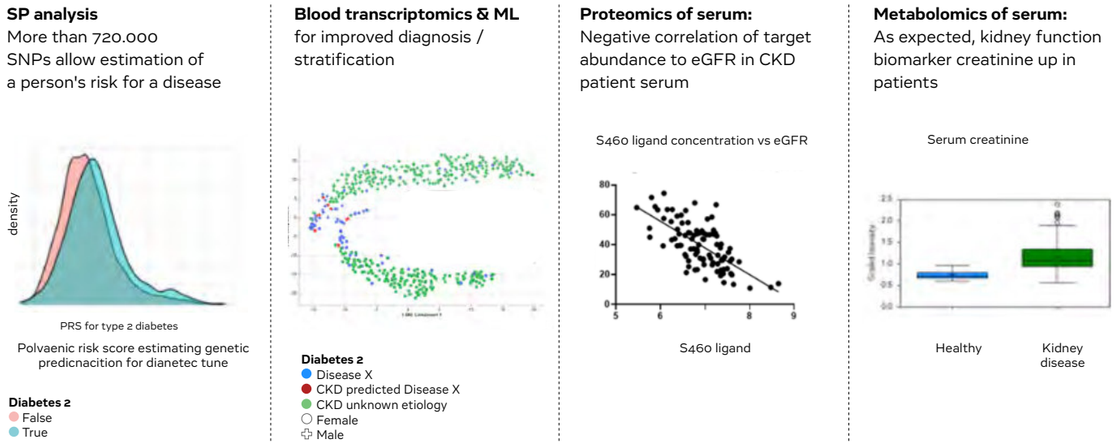
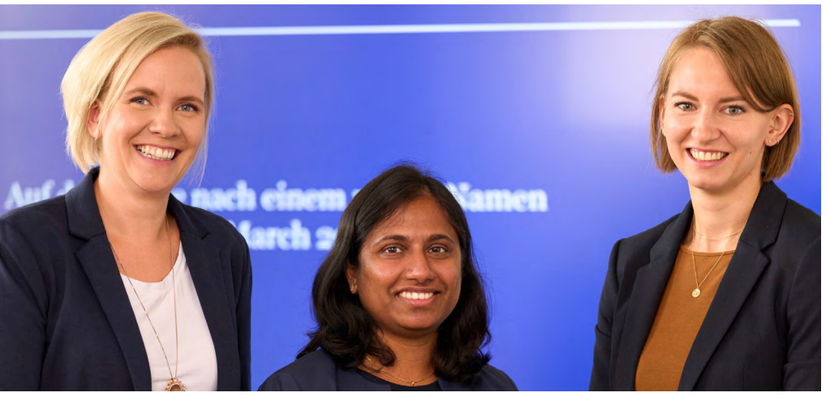


Figure 11: *Building the most comprehensive molecular profile of patients. Shown are SNPs, blood transcriptomics, proteomic and metabolomic data interconnected via the clinical parameters in the study.*

BRITTA SEIP
Research Scientist, Metabolic Disease

For me it is very easy to get started with data analysis without any prior coding knowledge by having PanHunter at hand. I especially appreciate that in PanHunter all analyses, starting from sample QC to the in-depths analysis of single-cell analysis are much more offered in one tool.

... for computational biologists



What does your professional life at Evotec look like? The fast development of omics technologies (e.g., next generation sequencing) in the recent years has been the starting point for truly data-driven drug discovery. We are working in a dedicated Computational Biology team of around 20 specially trained biologists with in-depth experience in analysing omics data and ensure efficient application of high-throughput omics technologies by analysis and biological interpretation of those resulting large and complex datasets.

COMPUTATIONAL BIOLOGY GROUP TEAM
representatives Maren Feist, Vijaja Kari and Elisa Buchberger

To integrate and analyse omics data, we can rely on PanHunter allowing us to work with different datatypes in an efficient, interactive, and reproducible manner. The strength of PanHunter is that it allows us to use one streamlined analysis workflow for most datasets, without adaptations and the need of developing new analysis scripts. The interactive user interface makes sample selection fast and intuitive. Results can easily be saved and shared via snapshot links – a strong advantage especially when multiple people working on the same project and ensuring reproducibility of the results.

A typical data analysis workflow starts with a quality check to spot technical biases or outliers that would interfere with biological interpretation. Subsequent dimension reduction gives a first idea of sample clustering and structure as well as provides the starting point for downstream analyses. Typically, we investigate differential gene expression, pathway, and gene ontology enrichment, network analysis, interpretation of patient data, and signature matching. An important role of a Computational Biologist is to present insights to projects and clients. With PanHunter we can produce high quality, interactive visualisations that are easy to share, understand, and can be integrated into our analysis reports. These visualisations include many different formats ranging from dimension reduction plots, heatmaps, networks and pathway visualisations, and each has multiple options to customise.

Together with our colleagues, we are constantly working on the implementation of new tools and the improvement of existing apps, ensuring state-of-the-art analysis of datasets. Thereby, PanHunter is organically growing with every new project, addresses real life questions, and incorporates the expertise of many, highly skilled scientists.

... for bioinformaticians

RAMON VIDAL
Research Scientist Bioinformatics

We have implemented great support for single-cell sequencing and its related technologies such as spatial transcriptomics in PanHunter: starting from a clean visualisation of dimension reductions, one can select individual sets of cells or entire cell clusters and annotate them manually. It is also possible to automatically annotate cell-types using our single-cell classifier, which is based on machine learning algorithms, which can be of great help for the analysis of complex or unknown tissue types.

MICHAELA BAYERLOVA
Research Scientist Bioinformatics

I am working on the NURTuRE dataset, which includes large patient cohorts with very rich and complex clinical data. The QC requirements regarding this patient data initiated the development of a PanHunter app to enable efficient analysis. It provides options for non-bioinformatic scientists to perform a detailed data QC, cleaning, and transformation of patient records which turned out to be a crucial step for subsequent statistical analysis. As an example, disease-relevant patient parameters can be associated with omics features to identify novel biomarker and target candidates.

MANUEL LANDESFEIND
Senior Scientist Bioinformatics

PanHunter allows our colleagues, often non-coding users, to execute “standard analyses” easily and efficiently. This reduces the workload for the bioinformatics team and allows to focus on the more complex and project specific analyses. The platform facilitates the communications over different departments or with customers from various fields and improves teamwork: if one colleague started an analysis in the software, somebody else can pick up the results and conduct further work easily. This also speeds up interaction between the researcher and the bioinformatics team significantly!

Cord Dohrmann
Chief Scientific Officer
cord.dohrmann@evotec.com

Matthias Evers
Chief Business Offiver
matthias.evers@evotec.com

CONTACT US

IMPRINT

EDITOR *Evotec SE* / **CHIEF EDITOR** *Michael Bayer* /
CONTENT *Michaela Bayerlova, Manuel Landesfeind, Timur Samatov, Sven Sauer, Erik Schliep, Britta Seip, Philipp Skroblin, John Szilagy, Florian Tegeler, Ramon Vidal, Winfried Wunderlich*
DESIGN *Alessandri, Design & Brand Manufactory*

evotec

DDup

EVOTEC SE
Manfred Eigen Campus, Essener Bogen 7
22419 Hamburg (Germany)
www.evotec.com/en/innovate/ddup