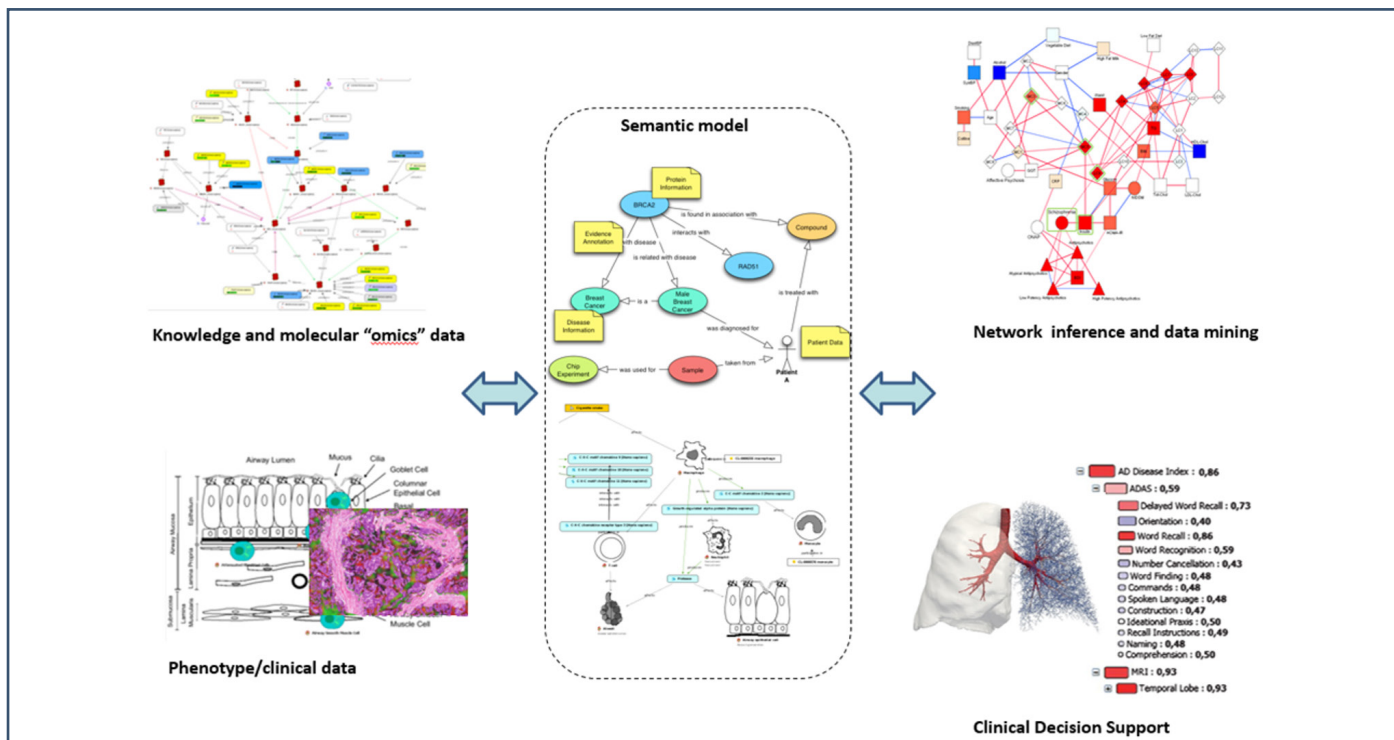


Biomax Knowledge Management - your scientific dashboard and connector to Systems Medicine



Get a deeper insight into your knowledge

Clinical Scenario: Stratification of risk patients

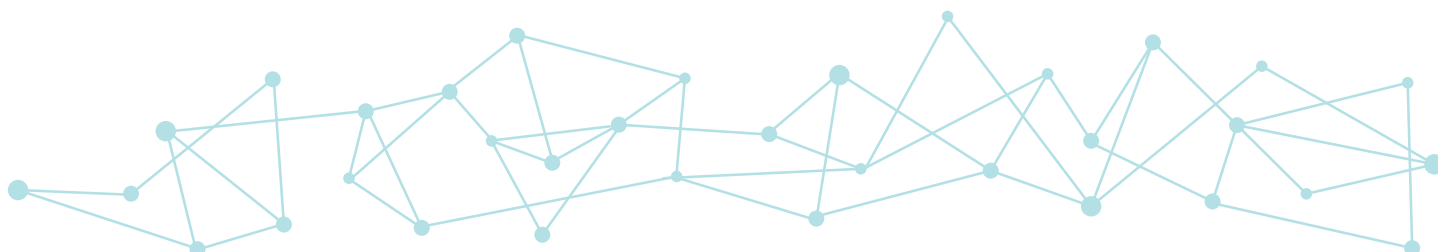
As a clinician I need to stratify my COVID-19 patients according to their risk for severe complications. Currently we lack information about most of the parameters determining risk except from age and co-morbidities.

If we collect relevant information (real-world data, observational trials) and integrate across the multitude of ongoing international initiatives such as the PREPARE clinical trials network, we can, within the first wave of a pandemic, produce evidence for informed decisions.

Clinical Scenario: Individualised patient treatment

As a clinician I would like to treat the right patient with the right treatment at the right time. Currently our understanding of the pathobiology of SARS-CoV-2 is not sufficient to provide specific therapy. Similarly, our understanding of the mechanisms causing symptoms is too limited to provide individualised treatment beyond generic approaches to fever, breathing support and infection handling.

Integrating all prior knowledge about SARS-CoV-2 as well as related human Coronaviruses with the huge omics datasets being rapidly generated and enabling



analytical data and knowledge exploration will improve our pathobiological understanding and assist informed decisions on individual treatment options.

Background

The use cases described above involve different capabilities and services that Biomax provides in the context of Knowledge Management for Systems Medicine:

- eCRF for clinical trial data capturing
- Clinical study data harmonisation
- Omics data management
- Disease maps
- Integrative data and knowledge exploration
- Interoperability
- Collaboration support
- Supporting clinical research privacy tasks (GDPR, DTAs, ISO 27001)

eCRF for clinical trial data capturing

Electronic case report forms (eCRF) enable automatic consistence and quality checks to ensure unprecedented completeness and quality of clinical data. In a recent clinical study on gastric cancer this approach helped to identify deficiencies in current standard testing and reporting procedure.

The secure Biomax technology platform facilitates collection, monitoring and management of data. A new eCRF can be developed very quickly on this platform. Libraries of eCRFs, evaluated in previously approved study protocols, and the wizard-based configuration of the specific forms allow fast, simple provision of new forms. Data standards are followed and, if necessary, mapping templates transform incoming data into the corresponding standard representation.

All newly configured eCRFs undergo extensive validation and acceptance testing before being deployed to production.

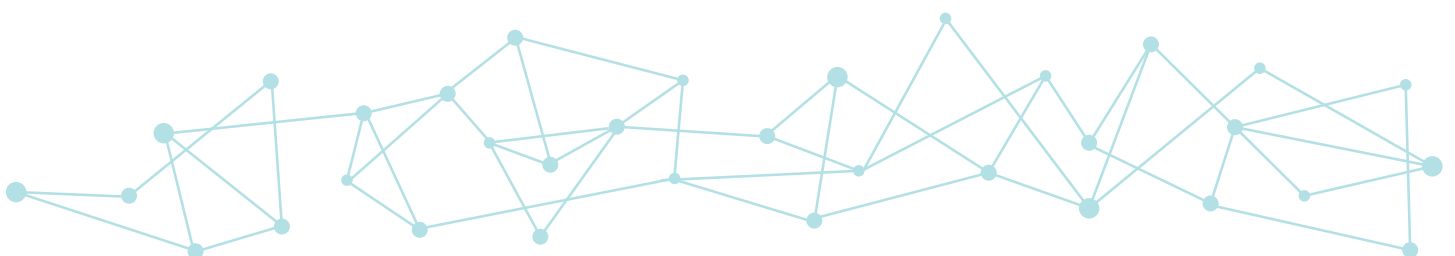
eCRFs and the knowledge management and analysis system support a number of data quality related tasks:

- Supporting data cleaning processes using online and offline ,query' procedures.
- Facilitating the source data verification process of collected study data.
- Creation of dedicated data sets for interim analysis purposes (e.g. for data and safety monitoring boards).
- Supporting of the collection and cleaning of non-eCRF data (e.g. lab data)
- Integrating all (raw) data sources to a federated data warehouse.
- Creating and sharing clean datasets on a secure data portal for final data analyses.

Clinical study data harmonisation

While the combination of data from multiple clinical studies would greatly improve statistical significance and analytical power for many research questions the differences in definitions and methods preclude direct pooling of original individual participant data. Data harmonisation can be defined as the process of adapting different variables from different sources in a way that makes them amendable for integrated analysis. In some cases this is a simple process, like harmonising variables by unit (e.g. blood pressure in mmHg or Pascal). In other situations, the differences in instruments, procedures or settings for data collection result in such differences that data cannot be combined in a straightforward manner. In such cases a successful harmonisation to achieve comparability is one of the main challenges.

Biomax has supported harmonisation processes for more than 50 clinical studies involving over 100,000 participants in a multitude of international projects.



Our technology provides automatic mapping between different formats, semantic concepts and standard units as well as mapping to Ontologies and rulebased data transformations.

For complex harmonisation processes we provide a collaborative, secure environment which enables structured management of variables, their definitions and mappings to each other, ensures transparency of discussions and decisions and traces changes over time. Following the DataSHaPEr harmonisation concept and Maelstrom research, the variables generated by each study are enhanced with rich metadata by the clinical experts and, as far as possible, mapped to ontologies by Biomax. In addition the clinical experts define “reference definitions” which represent an ideal of a variable. For example the reference for “education” might be defined as “categorical basic/secondary/high”. Each existing study variable is then mapped to the reference and mappings reviewed by clinical experts. The knowledge management infrastructure keeps track of references, variables, mappings and metadata to enable automatic data integration, algorithmic quality assessment, generation of summary statistics and application programming interface (API) based access of analysis methods.

Data formats provided by clinical partners, e.g. various Excel structures or statistical software formats are syntactically mapped by Biomax. When integrating data from multiple sites or follow-ups a number of data quality issues are frequently observed such as missing values or inconsistencies over time. Biomax provides automatic data quality assessments based on metadata including but not restricted to: completeness; availability of required data; data type check, e.g. float, thesaurus; data ranges based on realistic expected values, e.g. blood systolic between 60 and 250. Rule based inconsistency checks can be developed based on expert input, e.g. diagnosis “healthy” is in

conflict with PBRs score above 120. Units and codes are automatically converted and unified.

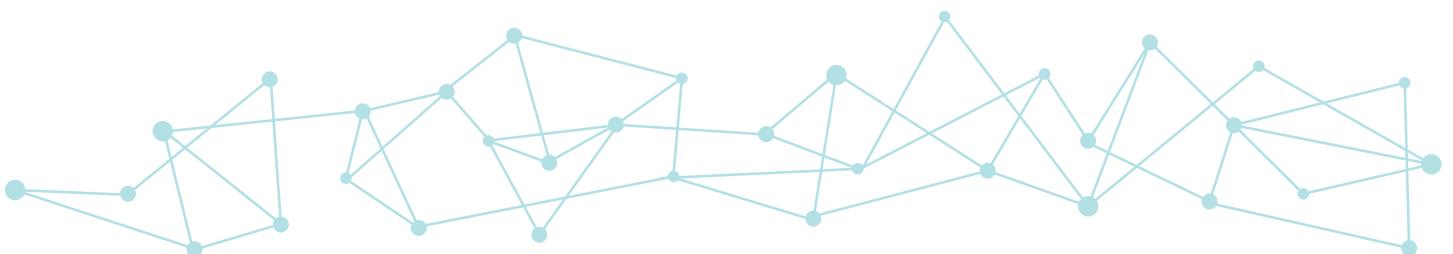
Omics data management

Since 2006 Biomax provides integrative management of clinical data, e.g. questionnaires, anthropometric and physiological measurements with standard clinical laboratory data as well as any type of omics data, from genetic variants, to DNase-seq, Meth-seq, ChIP-seq, gene expression (array and RNA-seq), proteomics and metabolomics with imaging derived information.

This integration is enabled by the semantic model which uses semantic concepts and mapped prior knowledge to establish a structured background network on which to map heterogeneous data.

Prior knowledge includes, among others, relevant molecular elements (genes, proteins, metabolites, etc.), functional information (GO, OMIM, etc.), functional interactions (e.g. protein-protein interaction, transcriptional regulation such as the mouse TF-regulatory network, miRNA network, etc.) and information about gene homologs (mouse, rat, human). These can be derived from structured databases, ontologies and mined from the literature.

In addition, genome features with coordinates for peak to gene associations of NGS data (e.g. mouse genome assembly mm10), metabolic and signal transduction pathways, cell types related to B-cell differentiation as well as ontologies such as the mouse anatomy ontology (MGI) have been incorporated. This integrated and dynamically organized knowledge serves as an information rich, structured background network. The semantic mapping of experimental data to this background network of prior knowledge enables complex integration and analysis approaches.



Disease maps

Biomax is involved in the EU funded PREPARE network for harmonized large-scale clinical research studies on infectious diseases, prepared to rapidly respond to any severe infectious disease outbreak, providing real-time evidence for clinical management of patients and for informing public health responses (<https://www.prepare-europe.eu/>).

Within PREPARE, Biomax has already generated and provided disease maps for a range of infectious diseases such as Zika or Influenza.

A disease map is a knowledge network, based on capturing, condensing and structuring of available information about disease facts, from symptoms to co-morbidities, affected or involved organs, tissues and cell types to molecular processes, pathways and single molecules. The disease map thus makes this information available for research management decisions, integrative data analysis or computational modelling.

In the light of the current global SARS-CoV-2 spread the EC FP7 PREPARE project has switched into outbreak mode level 3 to start clinical research and support of comparative human Coronavirus (h-CoV) pathobiological research including disease maps.

Building of a h-CoV disease map: Collection and assembly of h-CoV relevant information to a comprehensive knowledge network

A h-CoV disease map is produced in a stepwise process starting with automatic literature mining based on available ontologies, moving to optimization of ontologies for the h-CoV use case to manual curation and qualification of results.

Depending on the required use, completeness and quality of the disease map, different types of expertise and amounts of resources are necessary to achieve

significant impacts regarding SARS-CoV-2 pathobiology research. While PREPARE is focused on enabling pathobiology research, other stakeholders may be interested in supporting further specific uses such as vaccine development.

Integrative data and knowledge exploration

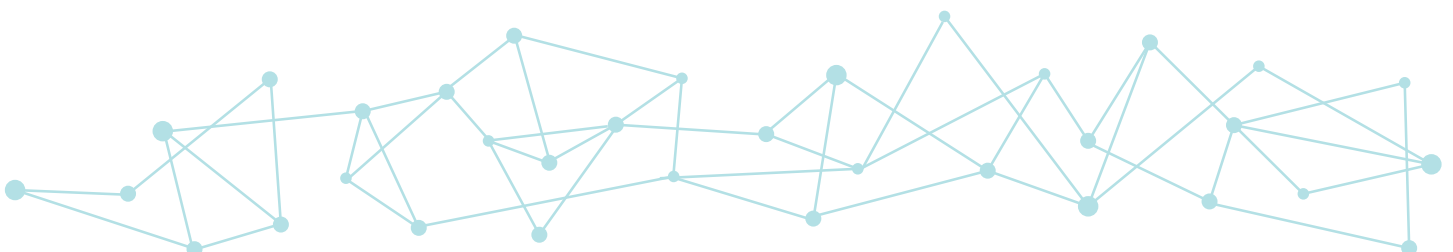
The purpose of the multiple knowledge extraction, data management, structuring and mapping functions described previously is to provide an integrated knowledge base which can be mined to find new associations, generate hypotheses or simply distribute knowledge within an organisation.

While the Biomax technology is not intended to provide specialised statistical or bioinformatics functionality and workflows, it delivers easy to use functions for clinical and biological researchers to quickly generate and pre-evaluate informed hypotheses which can then be rigorously tested by statistical or bioinformatics experts using the appropriate specialised tools.

To this end, multiple specific ways of exploring and mining the information are provided in addition to the AI-based language interface AILANI, offered by Biomax. Further information about AILANI can be downloaded at www.biomax.com/AILANI.

Integrative queries allow the seamless combination of heterogeneous information from multiple sources, e.g. “search for all patients aged above 35 with CRP value >30 and presenting at least one genetic variant in a gene in the EGFR-signalling pathway”. These queries are natural language based and do not require coding, SQL or SPARQL knowledge.

Summary statistics can be used to compare any two arbitrarily selected groups on any of the associated data.



Comparative analysis provides standard statistical methods such as generalised linear models to calculate multiple-testing corrected statistics across large datasets such as gene expression.

Functional significance of specific sets of genes, proteins, cell types etc. can be tested by methods such as gene-set-enrichment and overrepresentation. The test can be applied to any associated information such as molecular function, pathway participation, disease association etc.

Network search allows the exploration of the whole knowledge network and can reveal previously unknown associations. For example, combining signal transduction pathways with protein-protein interaction networks, disease association and gene-drug-interactions revealed potential drug candidates from Zika virus – host interaction data. The algorithms available for network search enable the discovery of connections within a group of objects of interest or between different groups.

Finally, machine learning brings a different dimension to complex data sets in which a single patient may be associated with several thousand attributes. Classical, uni- or multivariate statistics is quickly overwhelmed by the amount of possible tests that are needed to identify attributes of interest. The SOMiner (a neural network based similarity mapping and clustering method described elsewhere) of Viscovery, a subsidiary of Biomax, allows users to visually explore and reduce the set of attributes. Further information about the SOMiner can be downloaded at www.biomax.com/viscovery.

Interoperability

Biomax technology was the first commercial software to be awarded a certificate for system interoperability by the US National Cancer Institute. It has been designed to integrate federated data sources as well as interface with analysis methods and therefore provi-

des a multitude of APIs and plug-in mechanisms, for example SOAP and REST web services, plug-ins for R, KNIME, Cytoscape or connectors for Oracle and MySQL.

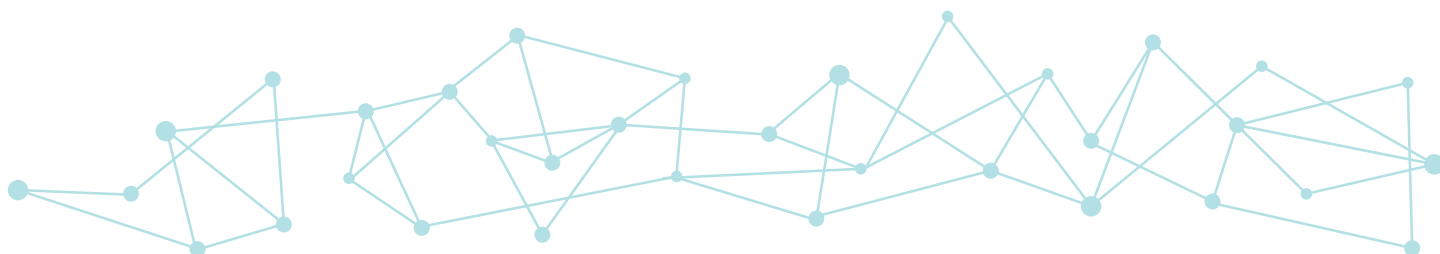
In addition, the semantic modelling concept used by Biomax ensures data FAIRness in terms of unique, stable IDs and rich metadata (mapped to ontologies whenever possible) as well as the ability to map to a large diversity of semantic and syntactic standards such as HL7, CDISC, SBML etc.

Collaboration support

In many international research projects, Biomax has generated collaborative environments to manage and discuss analysis plans, from the scientific question and hypothesis to description of analytical approach to the selection of harmonised variables. In this way data access requests can be made fully transparent to all partners and, based on template data analysis agreements (DTA), can be enacted quickly. In addition, this technical infrastructure opens up the possibility for secure collaboration and FAIR data sharing beyond institutional or project scopes by providing role-based access control and full audit of access and change. To this end Biomax can support data governance boards with knowledge and data management policies and SOPs for the granting of secure, legal and ethical access.

Supporting clinical research privacy tasks (GDPR, ISO 9001, ISO 27001)

Biomax is ISO 9001 and ISO 27001 certified and applies the corresponding procedures of quality management and IT-security. The BioXM™ Knowledge Management Environment provides encrypted data transfer, authentication, detailed access control based on resources and roles and full audit. All technological and organisational measures required by the GDPR are implemented. GDPR compatible data processing agreement templates are available and corresponding contracts have been approved by company and clinical data protection officers in major European countries.



The ISO 9001 based quality management ensures consistence between specification and reported result, including reviews of requirements and technical specifications, software implementation and documentation. Quality assessment according to ISO 29119 includes unit, integration, regression and acceptance testing.

Acknowledgement

The integration of COVID-19/SARS-CoV-2 specific sources was supported by the European Commission's Seventh Framework Research Programme project PREPARE (FP7-Health n°602525).

The development of the Systems Medicine workflows was supported by multiple European and German Research Programme projects, especially by BioBridge, Synergy-COPD, AirPROM, STATegra, MeDALL, SYS-Stomach, METSY and PREPARE.

References

Raimúndez E, Keller S, Ebert K, Hug S, Theis FJ, Maier D, et al. Model-based analysis of response and resistance factors of cetuximab treatment in gastric cancer cell lines. *PLoS Comput Biol*. 2020;16: e1007147. doi:10.1371/journal.pcbi.1007147

Gomez-Cabrero D, Tarazona S, Ferreirós-Vidal I, Ramirez RN, Company C, Schmidt A, et al. STATegra, a comprehensive multi-omics dataset of B-cell differentiation in mouse. *Sci Data*. 2019;6: 256. doi:10.1038/s41597-019-0202-7

Franssen FM, Alter P, Bar N, Benedikter BJ, Iurato S, Maier D, et al. Personalized medicine for patients with COPD: where are we? *International Journal of Chronic Obstructive Pulmonary Disease*. 2019 [cited 9 Jul 2019]. doi:10.2147/COPD.S175706

Benet M, Albang R, Pinart M, Hohmann C, Tischer CG, Annesi-Maesano I, et al. Integrating Clinical and Epidemiological Data on Allergic Diseases Across Birth Cohorts: a MeDALL Harmonization Study. *Am J Epidemiol*. 2018;188: 408–417. doi:10.1093/aje/kwy242

Keller S, Zwingerberger G, Ebert K, Hasenauer J, Wasmuth J, Maier D, et al. Effects of trastuzumab and afatinib on kinase activity in gastric cancer cell lines. *Mol Oncol*. 2018. doi:10.1002/1878-0261.12170

Frank E, Maier D, Pajula J, Suvitaival T, Borgan F, Butz-Ostendorf M, et al. Platform for systems medicine research and diagnostic applications in psychotic disorders-The METSY project. *Eur Psychiatry*. 2018. doi:10.1016/j.eurpsy.2017.12.001

Pinart M, Albang R, Maier D, Duran-Tauleria E, Mena G, Gimeno-Santos E, et al. Systematic Review on the Definition of Allergic Diseases in Children: The MeDALL Study. *Int Arch Allergy Immunol*. 2015;168: 110–121. doi:10.1159/000442414

Gomez-Cabrero D, Menche J, Vargas C, Cano I, Maier D, Barabási A-L, et al. From comorbidities of chronic obstructive pulmonary disease to identification of shared molecular mechanisms by data integration. *BMC Bioinformatics*. 2016;17: 1291. doi:10.1186/s12859-016-1291-3

Tényi Á, de Atauri P, Gomez-Cabrero D, Cano I, Clarke K, Falciani F, et al. ChainRank, a chain prioritisation method for contextualisation of biological networks. *BMC Bioinformatics*. 2016;17: 17. doi:10.1186/s12859-015-0864-x

Gehring U, Wijga AH, Hoek G, Bellander T, Berdel D, Brüske I, et al. Exposure to air pollution and development of asthma and rhinoconjunctivitis throughout childhood and adolescence: a population-based birth cohort study. *The Lancet Respiratory Medicine*. 2015 [cited 11 Nov 2015]. doi:10.1016/S2213-2600(15)00426-9

Cano I, Lluch-Ariet M, Gomez-Cabrero D, Maier D, Kalko S, Cascante M, et al. Biomedical research in a Digital Health Framework. *Journal of Translational Medicine*. 2014;12: S10. doi:10.1186/1479-5876-12-S2-S10

Cano I, Tényi Á, Schueller C, Wolff M, Migueláñez MMH, Gomez-Cabrero D, et al. The COPD Knowledge Base: enabling data analysis and computational simulation in translational COPD research. *Journal of Translational Medicine*. 2014;12: S6. doi:10.1186/1479-5876-12-S2-S6

Gomez-Cabrero D, Menche J, Cano I, Abugessaisa I, Huertas-Migueláñez M, Tenyi A, et al. Systems Medicine: from molecular features and models to the clinic in COPD. *Journal of Translational Medicine*. 2014;12: S4. doi:10.1186/1479-5876-12-S2-S4

Gomez-Cabrero D, Abugessaisa I, Maier D, Teschendorff A, Merckenschlager M, Gisel A, et al. Data integration in the era of omics: current and future challenges. *BMC Syst Biol*. 2014;8 Suppl 2: I1. doi:10.1186/1752-0509-8-S2-I1

Pison C, Tissot A, Magnan A, Botturi K, Stern M, Reynaud-Gaubert M, et al. The SysCLAD- Systems Prediction of Chronic Lung Allograft Dysfunction Study: Aims, Strategy and First Data. *The Journal of Heart and Lung Transplantation*. 2013;32: S220. doi:10.1016/j.healun.2013.01.555

Burrowes KS, De Backer J, Smallwood R, Sterk PJ, Gut I, Wirix-Speetjens R, et al. Multi-scale computational models of the airways to unravel the pathophysiological mechanisms in asthma and chronic obstructive pulmonary disease (AirPROM). *Interface Focus*. 2013;3. doi:10.1098/rsfs.2012.0057
Maier D, Kalus W, Wolff M, Kalko SG, Roca J, Marin de Mas I, et al. Knowledge management for systems biology a general and visually driven framework applied to translational medicine. *BMC Syst Biol*. 2011;5: 38. doi:10.1186/1752-0509-5-38



Biomax Informatics AG
Robert-Koch-Str. 2
82152 Planegg, Germany
info@biomax.com

Call:
+49 89 895574-0 (DE)
+44 7832 965 326 (UK)
+ 1 608 636 2222 (US)

www.biomax.com