Synomics Studio

Accelerated discovery and validation of biomarker clusters including multiple genomic, phenotypic, and clinical factors

Highlights

Synomics Studio brings accelerated results by:

- Rapidly identifying biomarker clusters consisting of up to 30 genomic, phenotypic, and clinical factors acting in combination
- Delivering comprehensive, reproducible, and interpretable results in hours or days on a fully scalable compute platform
- Driving the discovery and validation of new, more-specific biomarker clusters that enable the delivery of precision medicine solutions
- Providing personalization at scale for applications from clinical trial designs, healthcare analytics, clinical decision support, and patient management tools
- Producing affordable multi-dimensional analysis of diseases and population-scale genomic study data

Next-generation sequencing, single molecule detection technologies, and rapid assay platforms are transforming the understanding of disease processes, and the integration of these new insights with electronic health records is beginning to fundamentally impact healthcare by enabling more personalized medicine.

One challenging consequence is that the closer the diseases are examined through a molecular lens, the more complex they appear to be. In oncology, scientists have turned what would have been thought of as a single disease 20 years ago into a whole series of rare diseases. While this provides more accurate diagnosis and can lead to better selection of treatments for patients, it requires a fundamentally new set of analytical tools to investigate, diagnose, advise, and prescribe therapies to individuals.

The metabolic processes underlying complex chronic diseases are highly interrelated and the factors driving disease risk and progression are usually polygenic and heterogeneous. This means that in several of the most prevalent and costly long-term conditions, multiple factors (including many that are not genomic) act in combination to lead to the disease risk.

Identifying clinically relevant biomarkers associated with specific outcomes, for example, disease risk or therapy response, in large patient populations is hugely complex and requires finding combinations of several features [for example single-nucleotide polymorphisms (SNPs)] that when found together are associated with the observed outcome.

As patients' disease risk or therapy response is often also heavily influenced by multiple phenotypic features such as their lifestyle, assay results, comorbidities, treatment history, and environment, these data must also be included in this multi-dimensional, hyper-combinatorial association analysis.

Until the development of Synomics Studio by Row Analytics the analysis of combinations of SNPs in large patient populations has not been a practical proposition. The current tools, for example, genome-wide association studies (GWAS), operate at the limits of current computational capacity, and yet have often proved inadequate to represent and unravel the full complexity of important diseases. Even GWAS studies looking for just one or two disease-associated SNPs have often failed to find reproducible and clinically useful biomarkers that can be used to accurately segment patient populations.

Single or two SNP association studies in large populations with tens of thousands of patients along with control samples and millions of SNPs often take months of computational time to identify and validate putative biomarkers, many of which have subsequently turned out not to be reproducible between different patient data sets, or to have limited clinical relevance as they have low penetrance in the targeted patient population.

When considering complex polygenic disorders, each additional associated factor in a biomarker cluster (that is, going from looking for two SNPs in combination to three SNPs in combination) can add 5 orders of magnitude to the number of possible combinations to be computed. This has imposed an unhelpfully low threshold on the complexity of associations that could be explored, even at a solely genomic level, and has precluded the use of phenotypic and clinical data alongside those genomic analysis.

When you know that complex diseases might involve interactions of 4, 8, 10, or even more factors, and only 50% of the disease prevalence can be explained by genetic factors, this has obviously been a big obstacle that will only get worse as the scale of disease population studies increases.

How Synomics Studio on IBM overcomes these challenges

Synomics Studio is a powerful new analytical approach to multi-dimensional, hyper-combinatorial association studies. Synomics Studio enables large-scale data sets including genomic, phenotypic, and clinical data to be fully analyzed for associations that meet stringent user-defined criteria for p-value, penetrance, and cluster size. The Synomics algorithms are massively parallelized, allowing them to be distributed across multiple graphics processing unit (GPU) compute devices, and ensuring that results can be returned in minutes, hours, or days, rather than weeks or months.

While massively quicker and more scalable than traditional GWA studies, Synomics is compute-intensive and requires large amounts of very dense compute power. A compute architecture based on the IBM® Reference Architecture for Genomics using IBM Power Systems™, IBM Spectrum Computing, and IBM Spectrum Storage™ software including IBM Spectrum LSF® and IBM Spectrum Scale™ (based on IBM GPFS™) delivers outstanding performance for such demanding requirements.

Synomics Studio

Synomics Studio makes it easy for genomics and precision medicine researchers to quickly identify, validate, and understand the metabolic relevance of disease-related features in novel biomarker clusters by combining multi-factor association analysis with integrated validation, biological annotation, and interpretation features. These include a full bioinformatics knowledge graph containing SNP, gene, and pathway annotations and systems pharmacology models.

An example of the various association studies that can be undertaken is shown by the study of 14,777 people, all of whom had BRCA1 or BRCA2 (or both) mutations. Mutations in these genes are associated with elevated breast, ovarian and other cancer risks due to abnormal DNA repair and recombination functionality. The cohort is split into those *affected* individuals who have or have had breast cancer and *unaffected'* who have not.

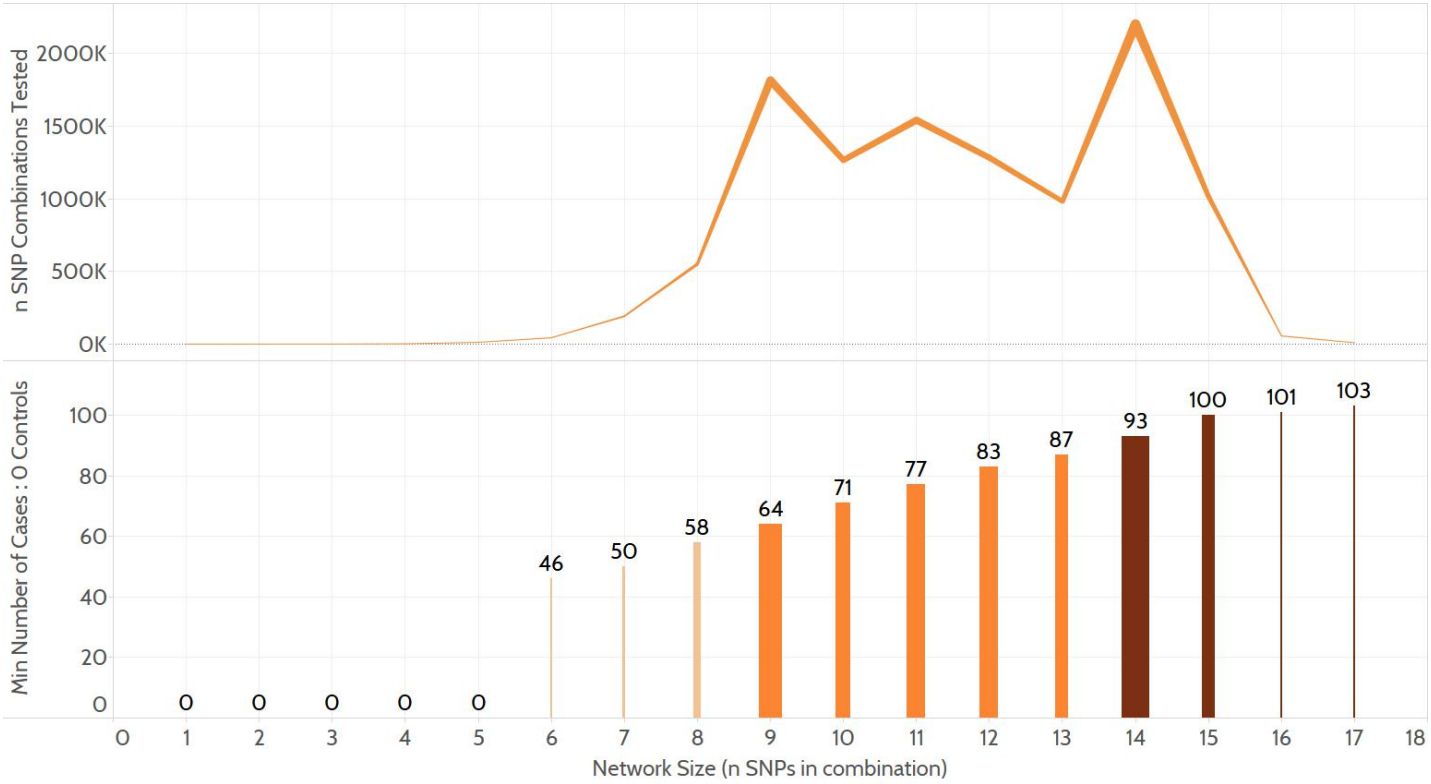

Figure 1. Number of computations (top graph) and minimum number of individuals in each n-combination cluster found only in the cases, not the controls (where n=1 to 17 SNPs – bottom graph)

Figure 1 shows that even for people with BRCA1 or BRCA2 (or both) mutants, there are no biomarker clusters containing less than six SNPs in combination that are represented only in the affected population and not in the unaffected. Clusters with combinations that do appear in the control set might well be random observations. The most populous biomarker clusters, present in 103 affected people and 0 controls, contain 17 SNPs acting in combination.

Because the combinations are automatically tested in series, running until no further significant clusters can be found, the analysis is both simple to perform and very robust. If there are no significant associations that meet the stringency of the parameters chosen, or if the parameters are returning too many low significance clusters, this will be apparent very quickly and the parameters can be adjusted easily.

Synomics Studio performs a configurable number (usually at least 1,000) of cycles of fully random permutations of the case:control population to eliminate the possibility of random observations and associations being reported from within the data. This gives highly reliable results which have been both reproduced with new data sets and clinically validated in previous studies.

The key features of the solution include:

Synomics Studio: This is the main association analysis tool performing initial analysis of genomics associations for up to 30* features (SNPs) in combination.
Phased analysis: Mining of the initial associations usually takes minutes or hours and gives an initial indication of the suitability of the parameters enabling early validation of the quality of the biomarker clusters that might be generated during the analysis phase.
Reporting and visualization: View reports on the progress of association study jobs as they are running. Full reports on the biological role, similarity, and significance of the features in the biomarker clusters can be generated.
Automated analysis: For smaller studies and for jobs on the Synomics Studio Community Server, the analysis can be fully automated.
Data security: Synomics Studio is designed to run securely across a range of research, production, and regulated environments, both in a hosted or dedicated mode. Depending on the level of access and sensitivity of information, the association analysis can be performed either with fully or partially identified, or fully deidentified data sets.
Data provenance: The data used to annotate and interpret the biological relevance and function of features in the biomarker clusters is derived from up-to-date high-quality sources.

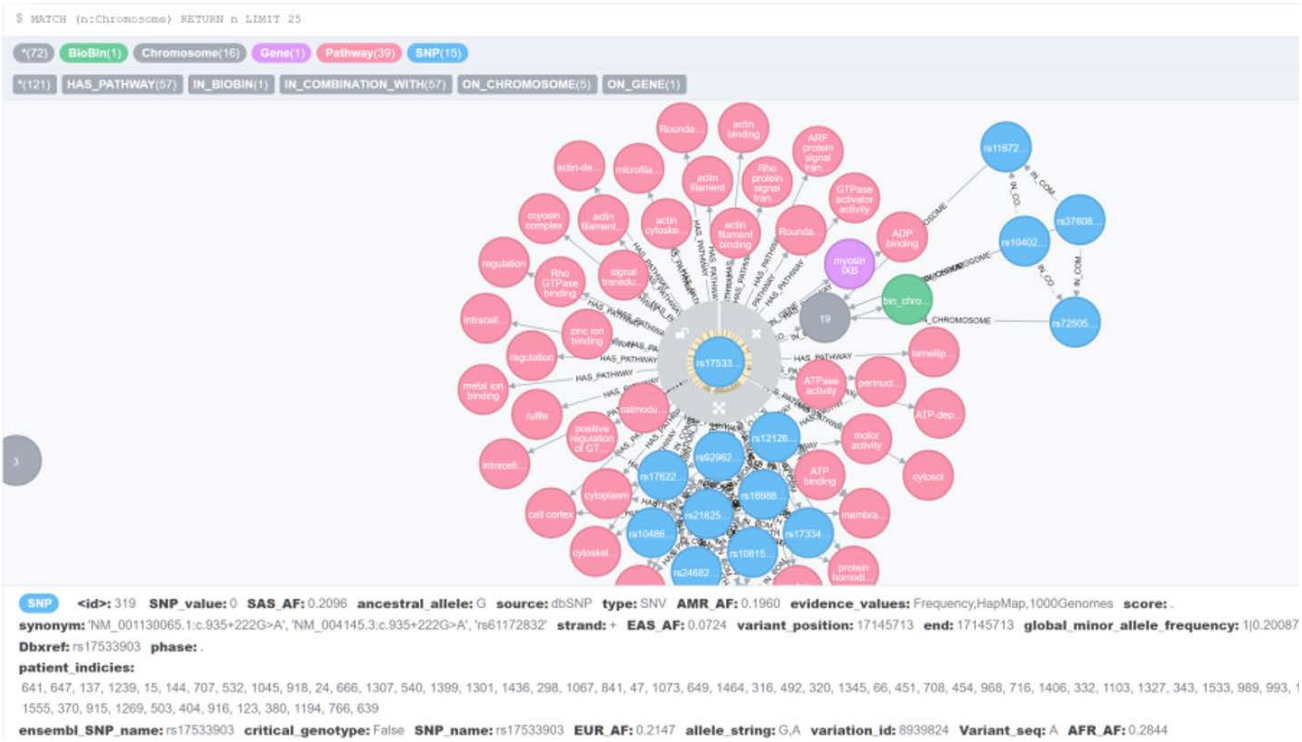

Figure 2. Synomics Studio knowledge graph exploration and visualization tools for biological annotation and interpretation of the abolic relevance and similarity of features within a single biomarker cluster

* Larger studies are possible but may require more compute resources.

IBM Reference Architecture for Genomics

IBM, in collaboration with key researchers and partners, created the IBM Reference Architecture for Genomics. It is an end-to-end reference architecture that defines the enterprise data management, workflow orchestration, and global access capabilities across key genomics, translational and personalized medicine platforms. It supports large-scale genomics sequencing and downstream data analytics, providing: data lifecycle management to support large scale data growth; software-based abstraction layers for compute, storage, big data and cloud; and workload and workflow orchestrator for applications.

IBM Power Systems

IBM Power Systems deliver trusted, state of the art technology at the small-, mid-, and enterprise-computing levels and are broadly deployed in production environments worldwide. These systems compare favorably with x86 servers on cost while delivering greater performance, higher use, and superior availability. With so much consolidated compute power, Power Systems can provide outstanding performance for next-generation sequencing workloads, especially for workloads that are highly parallelized or have large memory footprints. Power Systems achieve higher performance per core through:

- Massive parallelism (threads)
- Higher clock frequencies
- 8-way simultaneous multithreading (SMT) per core
- Larger IBM POWER® L3 on-chip cache
- IBM PowerVM® optimization

All IBM Power® server models include reliability, availability, and serviceability (RAS) features that help avoid unplanned downtime. IBM RAS engineers have optimized server design to help ensure that IBM Power servers support high levels of concurrent error detection, fault isolation, recovery and availability. IBM PowerSC™ enables security compliance automation and includes reporting for compliance measurement and audit. Automation capabilities include supplying prebuilt system profiles that enforce compliance to various industry regulations, such as the Health Insurance Portability and Accountability Act (HIPAA).

The IBM POWER8® processor-based servers used in these studies realize incredible advantage in application performance by using NVIDIA Tesla P100 accelerators in combination with NVLink technology which unlocks over 2.8 times CPU:GPU communication between POWER8 NVLink processors and Tesla P100 accelerators. This powerful pairing is a breakthrough for accelerated high-performance computing (HPC) delivering advanced performance, programming, and accessibility required for processing large volumes of data quickly.

IBM software-defined infrastructure

IBM software-defined infrastructure offerings transform a static IT infrastructure into an agile workload- , resource-, and data-aware environment. The offerings are part of a comprehensive portfolio of workload and resource management tools (IBM Spectrum Computing and IBM Spectrum Storage) that enables any organization to deliver IT services in the most efficient way possible, optimizing resource use based on workload and data volumes and on business priorities. It is the foundation for a fully integrated software-defined environment, enabling the compute, storage, and networking infrastructure to adapt dynamically to changing business requirements for faster time to results and lower costs.

IBM Spectrum LSF is a powerful workload-management platform for demanding distributed and mission-critical high-performance computing environments. It provides a comprehensive set of intelligent, policy-driven scheduling features to make full use of compute infrastructure resources and to ensure optimal application performance.

For I/O intensive workloads typical in next-generation sequencing, file systems and storage play a big role. IBM Spectrum Scale (based on IBM GPFS) is a scalable, high-performance data and file management solution that is proved to benefit various next-generation sequencing workloads. Spectrum Scale also provides seamless capacity expansion, improved enterprise-wide efficiency, commercial-grade reliability, business continuity, and the flexibility of supporting a wide variety of platforms.

The Synomics Studio on IBM difference

There are several compelling reasons why life science researchers should consider using Synomics Studio on IBM for their association studies, biomarker discovery, and patient stratification projects:

- A forward-looking, end-to-end collaborative solution for genomics research and medicine. IBM Reference Architecture for Genomics was developed with leading researchers and partners to address the compute, data, and workflow needs in genomics research, translational and personalized medicine.

- Massively scalable and flexible analysis platform: Synomics Studio is the only solution for multi-factor association studies, and the only solution that integrates genomic, phenotypic, and clinical data sets.
- Systems are based on powerful IBM POWER processors and NVLink technology paired with NVIDIA GPUs that deliver better performance per core compared to x86 and the latest NVIDIA GPU devices. Power Systems are widely used in supercomputing where scalability and throughput is critical. An IBM Power System solution, expertly integrated with a tuned IBM Spectrum Scale and IBM Spectrum LSF, can deliver superior performance and scalability and fully use all available resources to deliver maximum throughput. In addition, these systems possess outstanding enterprise-grade reliability and security.
- Platform flexibility: With the Synomics Studio solution on IBM Power Systems, users can distribute their workflow across compute platforms, if needed. This flexibility makes it easy to take advantage of the performance and reliability of Power Systems without sacrificing the familiarity and functionality of existing or legacy systems.
- Total cost of ownership: IBM Power Systems are affordable compared to like-sized x86 systems. While the hardware costs are comparable, there are significant gains on software and support costs.

Performance of Synomics Studio on IBM solutions

Current genome wide association study (GWAS) analysis takes weeks or months, even for a small number of features. Synomics Studio can do much larger studies, orders of magnitude quicker, with initial results in minutes or hours, helping researchers achieve more insightful results much faster.

The Synomics algorithm and architecture scales to meet the largest genomics challenges – the algorithm's compute and memory resources scale linearly with the number of factors (for example, SNPs), and current medium-scale studies with 15,000 patients and controls and 200,000 SNPs are using only a small fraction of the system's GPU memory resources. Larger jobs can be horizontally scaled across multiple GPU devices and servers to meet the demands of the largest genomics and precision medicine projects.

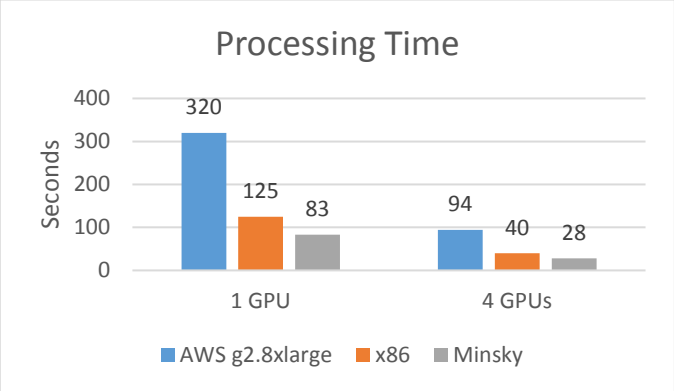

Table 1. Comparing the performance of Synomics Studio analysis on Power System S822LC to an AWS GPU instance and a generic x86 desktop (Intel Core i7 6850K @ 4.0 GHz with 4x GTX 1080).

Why IBM?

A compute architecture using IBM Reference Architecture for Genomics, IBM Power Systems with NVLink technology, and IBM software defined infrastructure offerings delivers outstanding performance for the large compute power and storage capacity demanded for the scale and scope of association studies enabled by Synomics Studio.

IBM Spectrum Computing offers a comprehensive portfolio of software-defined infrastructure solutions designed to help your organization deliver IT services in the most efficient way possible, optimizing resource utilization to speed up time to results and reduce costs. These offerings help maximize the potential of your infrastructure to accelerate your analytics, HPC, Apache Hadoop, Spark and cloud-native applications at any scale, extract insight from your data, and get higher-quality products to market faster.

Whether deployed in a data center or on the cloud, IBM Spectrum Computing solutions fuel product development, critical business decisions, and breakthrough insights in financial services, manufacturing, digital media, oil and gas, life sciences, government, research, and education. From designing Formula One race cars to credit risk analysis, organizations in a wide variety of industries are using IBM Spectrum Computing as a foundation for software-defined infrastructure solutions for big data, analytics, HPC, and cloud to improve business results.

For more information

To learn more about Synomics Studio, visit: https://rowanalytics.com/ and the synomics.studio product tab (at: https://rowanalytics.com/products/synomics-studio/)

To learn more about IBM Reference Architecture for Genomics, IBM Power Systems, and IBM software-defined infrastructure offerings including IBM Spectrum LSF and IBM Spectrum Scale, contact your IBM representative or IBM Business Partner, or visit:

- **ibm.com**/systems/spectrum-computing/industries/lifesciences.html
- **ibm.com**/systems/power
- **ibm.com**/spectrum-computing

Additionally, IBM Global Financing provides numerous payment options to help you acquire the technology you need to grow your business. The team provides full lifecycle management of IT products and services, from acquisition to disposition. For more information, visit: **ibm.com**/financing

IBM Systems
3039 Cornwallis Road
RTP, NC 27709

Produced in the United States of America