The Journal of Molecular Diagnostics, Vol. 26, No. 7, July 2024

ELSEVIER

the Journal of Molecular Diagnostics
jmdjournal.org

TECHNICAL ADVANCE

The Application of Knowledge Engineering via the Use of a Biomimetic Digital Twin Ecosystem, Phenotype-Driven Variant Analysis, and Exome Sequencing to Understand the Molecular Mechanisms of Disease

William G. Kearns,*† Georgios Stamoulis,‡ Joseph Glick,§ Lawrence Baisch,§ Andrew Benner,* Dalton Brough,* Luke Du,* Bradford Wilson,¶ Laura Kearns,*† Nicholas Ng,‖ Maya Seshan,‖ and Raymond Anchan‖

From Genzeva, Rockville, Maryland; LumaGene,† Rockville, Maryland; QIAGEN Digital Insights,‡ Redwood City, California; RYLTI BioPharma,§ Hauppauge, New York; IndyGeneUS AI,¶ Washington, District of Columbia; and the Department of Obstetrics and Gynecology,‖ Brigham and Women's Hospital, Harvard University, Boston, Massachusetts*

Accepted for publication March 19, 2024.

Address correspondence to William G. Kearns, Ph.D., Genzeva, 9420 Key West Ave., Ste. 100, Rockville, MD 20850.
E-mail: wgkearns@genzeva.com.

Applied artificial intelligence, particularly large language models, in biomedical research is accelerating, but effective discovery and validation requires a toolset without limitations or bias. On January 30, 2023, the National Academies of Sciences, Engineering, and Medicine (NAS) appointed an ad hoc committee to identify the needs and opportunities to advance the mathematical, statistical, and computational foundations of digital twins in applications across science, medicine, engineering, and society. On December 15, 2023, the NAS released a 164-page report, "Foundational Research Gaps and Future Directions for Digital Twins." This report described the importance of using digital twins in biomedical research. The current study was designed to develop an innovative method that incorporated phenotype-ranking algorithms with knowledge engineering via a biomimetic digital twin ecosystem. This ecosystem applied real-world reasoning principles to nonnormalized, raw data to identify hidden or "dark" data. Clinical exome sequencing study on patients with endometriosis indicated four variants of unknown clinical significance potentially associated with endometriosis-related disorders in nearly all patients analyzed. One variant of unknown clinical significance was identified in all patient samples and could be a biomarker for diagnostics. To the best of our knowledge, this is the first study to incorporate the recommendations of the NAS to biomedical research. This method can be used to understand the mechanisms of any disease, for virtual clinical trials, and to identify effective new therapies. *(J Mol Diagn 2024, 26: 543—551; https://doi.org/10.1016/j.jmoldx.2024.03.004)*

Artificial intelligence (AI), machine learning (ML), and large language models (LLMs) have and continue to transform biomedical research and health care. The comprehensive integration of these technologies into biomedical research holds the promise of enhancing operational efficiency and reducing costs, improving diagnostic ability, uncovering new therapeutic targets, and enabling increasingly personalized medical treatments. The future of health care will be defined by how we leverage massive amounts of data via AI/ML/LLM analysis.

Although AI, ML, and LLMs hold tremendous promise for driving advances in biomedical research, these technologies, like all technologies, have limitations. Traditional AI/ML/LLMs normalize data and remove outliers, thus hindering the identification of hidden or dark data. This removal of outliers is often performed to simplify data sets, and the degree of normalization may be adjustable. AI, ML, and LLMs also require a test training set to perform the analysis, which could unintentionally introduce bias into the process. However, in using AI/ML/LLMs, one can modify

Supported by Genzeva (W.G.K. and L.K.), LumaGene (W.G.K. and L.K.), RYLTI (L.B. and J.G.), and Brigham and Women's Hospital research programs (R.A.).

https://doi.org/10.1016/j.jmoldx.2024.03.004

and enhance training sets to more accurately identify the problem to be solved to reduce bias.

On July 5, 2023, former Google CEO Eric Schmidt wrote in MIT Technology Review, "we should be cognizant of the limitations—and even hallucinations—of current LLMs before we offload much of our paperwork, research, and analysis to them."[1,p.317] The huge data sets required by traditional AI/ML/LLMs and the associated scale of combinatorial math limit the ability of the algorithms to explore biological complexity,[1] relegating most key relationships and critical interactions into dark data—data that are unseen, unexplored, and as a result, unanalyzed (Gartner, Inc., Stamford, CT).

To address these issues and to provide guidance to the biomedical community, the National Academies of Sciences, Engineering, and Medicine (NAS), sponsored by the NIH, the National Science Foundation, and the Department of Energy, began advocating research into the use of biomimetic digital twins technology to more effectively model multidimensional and multiscale biological complexity.[2]

The NAS also published a *Physics of Life Report*,[3,sect.7,p.22] which concluded, "An important lesson from the long and complex history of neural networks and artificial intelligence is that revolutionary technology can be based on ideas and principles drawn from an understanding of life, rather than on direct harnessing of life's mechanisms or hardware."

As a follow-up to the NAS/NIH/National Science Foundation/Department of Energy workshop, the NAS appointed an ad hoc committee to identify needs and opportunities to advance the mathematical, statistical, and computational foundations of digital twins in applications across science, medicine, engineering, and society. On December 15, 2023, the NAS released a 164-page report, "Foundational Research Gaps and Future Directions for Digital Twins."[4]

Across multiple domains of science, engineering, and medicine, excitement is growing about the potential of digital twins to transform scientific research, industrial practices, and many aspects of daily life. A digital twin couples computational models with a physical counterpart to create a system that is dynamically updated through bidirectional data flows as conditions change. Going beyond traditional simulation and modeling, digital twins could enable improved medical decision-making at the individual patient level, predictions of future weather and climate conditions over longer timescales, and safer, more efficient engineering processes. However, many challenges remain before these applications can be realized.

This report identified the foundational research and resources needed to support the development of digital twin technologies. The report presents critical future research priorities and an interdisciplinary research agenda for the field, including how federal agencies and researchers across domains can best collaborate.[4]

The report's key conclusions include the following:

- An important theme that runs throughout this report is the notion that the digital twin virtual representation be fit for purpose, meaning that the virtual representation—model types, fidelity, resolution, parameterization, and quantities of interest—be chosen, and in many cases dynamically adapted, to fit the particular decision task and computational constraints at hand, as well as acceptable cost.
- A top priority, because of the heterogeneity, complexity, multimodality, and breadth of biomedical data, the harmonization, aggregation, and assimilation of data and models to effectively combine these data into biomedical digital twins require significant technical research.
- For many applications, the models that underlie the digital twin virtual representation must represent the behavior of the system across a wide range of spatial and temporal scales. For systems with a wide range of scales on which there are significant nonlinear scale interactions, it may be impossible to represent explicitly in a digital model the full richness of behavior at all scales and including all interactions.
- Technical challenges in modeling, computation, and data all pose current barriers to implementing digital twins for biomedical use. Because medical data are often sparse and collecting data can be invasive to patients, researchers need strategies to generate working models despite missing data.

The committee's guidance included the following.

- A combination of data-driven and mechanistic models can be useful to this end, but these approaches can remain limited because of the complexities and lack of understanding of the full biological processes even when sufficient data are available. In addition, data heterogeneity and the difficulty of integrating disparate multimodal data, collected across different time and size scales, also engender significant research questions.
- New techniques are necessary to harmonize, aggregate, and assimilate heterogeneous data for biomedical digital twins
- Furthermore, achieving interoperability and composability of models will be essential.

Taken collectively, the comprehensive report praises the use of AI/ML/LLMs in biomedical research but also identifies gaps and some limits of AI methods when it comes to modeling and exploring the biological complexity of the real world. Digital twins address many of the recommendations from the report that new theories and methods are required to address the multidimensional, multiscale characteristics of problems in modeling and advanced analytics in general, and in biomedicine in particular.

AI/ML/LLMs and digital twins could complement each other if all of the techniques are used most carefully and with enough knowledge in hands.

To address these issues, the current advanced genomics experimental protocol was updated by introducing an innovative biomimetic digital twin ecosystem. We believe this is the first report incorporating this method into research to understand the pathophysiology of disease.

A study using the current biomimetic knowledge engineering method was used to generate an ecosystem of digital twins that implemented real-world reasoning principles and analyzed data that were raw and in their original state. This meant that no cleansing or normalization was performed to remove outliers and hide relationships and impacts within data sets. The use of this method has both leveraged and used dark data and has enabled unexpected discovery.[3]

This study focused on the molecular mechanisms of endometriosis. Endometriosis is an inflammatory condition occurring in 5% to 10% of women of reproductive age and is associated with debilitating pelvic pain and infertility.[5,6] It is characterized by the presence of endometrial-like tissue outside the uterus, mainly on pelvic organs. Definitive diagnosis requires visualization of lesions during surgery, contributing to a delay in diagnosis that globally averages 7 years from symptom onset. Causes of endometriosis remain largely unknown, but the condition has an estimated heritability of approximately 50%,[7,8] with approximately 26% estimated to be due to common genetic variation in the populations studied.[8]

The molecular mechanisms involved in the development of endometriosis are still being actively researched, and our understanding of the exact processes is evolving.[9] Although the precise mechanisms are not fully elucidated, several key molecular factors and pathways have been implicated in the pathogenesis of endometriosis. These include the epithelial-mesenchymal transition, angiogenesis, and vascularization. Chronic inflammation and immune dysregulation are considered important contributors, and hormonal factors also play a significant role. Finally, the role that genetic and epigenetic factors play in the pathogenesis of endometriosis has yet to be comprehensively identified.[10]

Genetic and epigenetic alterations have been investigated for their role in endometriosis susceptibility and development. Various genetic polymorphisms and mutations have been associated with an increased risk of endometriosis. Epigenetic modifications, including DNA methylation, histone modifications, and miRNA expression changes, can influence gene expression patterns in endometrial cells, affecting processes such as hormone signaling, inflammation, and tissue remodeling. These molecular mechanisms are not mutually exclusive, and they likely interact and influence each other in a complex manner.

Previous multiomic studies, such as next-generation sequencing (NGS) to identify pathogenic variants, microarrays to identify polymorphic markers, RNA-sequencing analysis for gene expression, and epigenetic analysis, have provided limiting and confusing results in their attempts to identify genomic markers associated with the pathogenesis of endometriosis. A genome-wide association study meta-analysis, which included 60,674 cases and 701,926 controls of European and East Asian descent, identified 42 genome-wide significant loci comprising 49 distinct association signals.[11] Although this study is comprehensive and provides important information on the identification of genomic loci potentially associated with the pathogenesis of endometriosis, it did not involve a diverse cohort composed of a wide range of different ethnic populations. Additional research may yield information about the potential role of ethnicity in the mechanism of this disease.

In the current study, an innovative approach using exome sequencing, QIAGEN's Clinical Insight (QCI) phenotype-driven ranking analysis (QIAGEN, Redwood City, CA), and a biomimetic digital twin ecosystem was used to identify dark data associated with the molecular profile of endometriosis. Significantly, this study yielded evidence for a potential biomarker and a chromosomal hot spot associated with the pathogenesis of endometriosis.

Materials and Methods

Patient Population

The authors declare that all endometriotic and normal matched samples were biopsied in a clinical diagnostic setting. All research experimental protocols were approved by a Brigham and Women's Hospital, of Harvard University (Boston, MA), Human Institutional Review Board: number 2017P000184, Generation of Imaging Agents and Identification of Therapeutics Targeting Endometriosis. The patient ages ranged from 28 to 49 years, and they were all of White ethnicity.

Next-Generation Sequencing

In the experimental protocol, whole exome NGS was performed on each sample to determine the presence or absence of known pathogenic mutations, and variants of unknown clinical significance (VUSs) associated with endometriosis (Figure 1).

Whole-genome amplified DNA (50 ng) from each sample was used as input for library preparation (Thermo Fisher Scientific, Waltham, MA). The library preparation was done using xGen DNA Library Prep EZ UNI (Integrated DNA Technologies, Coralville, IA). The DNA sample underwent enzymatic preparation to produce fragment sizes of approximately 200 bp. This was followed by ligation using full-length adapters. The samples then underwent an AMPpure bead (Beckman Coulter, Sharon, IL) cleanup and were washed. A PCR amplification was then performed, followed by a second AMPure bead cleanup. The samples were then sized (4200 TapeStation; Agilent Technologies, Santa Clara, CA) and quantitated (Qubit 4 Florometer; Fisher Scientific, Waltham, MA). Samples were pooled with no more than 12 samples per pool, and a 16-hour

hybridization was preformed using xGen Exome Hyb Panel version 2 (Integrated DNA Technologies).

A bead capture (Bait-Capture) and a set of post-hybridization washes were performed using an xGen Hybridization and Wash Kit (Integrated DNA Technologies). The authors then did a post-hybridization amplification using xGen Library Amplification Primers (Integrated DNA Technologies), followed by an AMPure bead cleanup. The authors' pools were sized and quantitated once more. The pools were normalized and pooled into a single pool.

The pooled libraries were then denatured and loaded onto a NovaSeq 6000 (Illumina, San Diego, CA) and sequenced using a NovaSeq 6000 S1 Reagent Kit version 1.5 (Illumina). The libraries bind to grafted oligoes on the flow cell and then hybridize and bridge on their specific oligo and undergo multiple cycles of amplification. This forms clusters using an ExAmp technology. Then, the clusters undergo two-channel sequencing by synthesis chemistry.

Validation

The authors' experimental protocol included a NovaSeq 6000 for short-read NGS, the Illumina Dragen Germline pathway for secondary analysis, and QCI for tertiary analysis. First, the authors comprehensively validated their whole-exome sequencing against National Institute of Standards and Technology reference/validation samples. The authors performed accuracy, sensitivity, specificity, positive predictive value, negative predictive value, positive percentage agreement, and precision (inter-assay and intra-assay) assays to complete their validation process.

The authors also participate in the College of Pathologists surveys. The authors also perform blinded DNA sequencing to previously known samples to ensure the accuracy of their results.

Required passing quality control metrics for each sample sequenced include the following: total_input_reads: >49,000,000; number_of_duplicate_marked_reads_pct: <10%; uniformity_of_coverage_pct_gt_02mean_over_target_region: >95%; average_alignment_coverage_over_target_region: >85%; and Pct_of_target_region_with_coverage_20x_inf: >95%.

Short-read sequencing (approximately 350 bp) is the best modality to use for this study as long-range sequencing (approximately 3000 to 5000 bp) is more relevant to identify structural variants. Furthermore, the current standard of care for clinical NGS testing is using short-read sequencing.

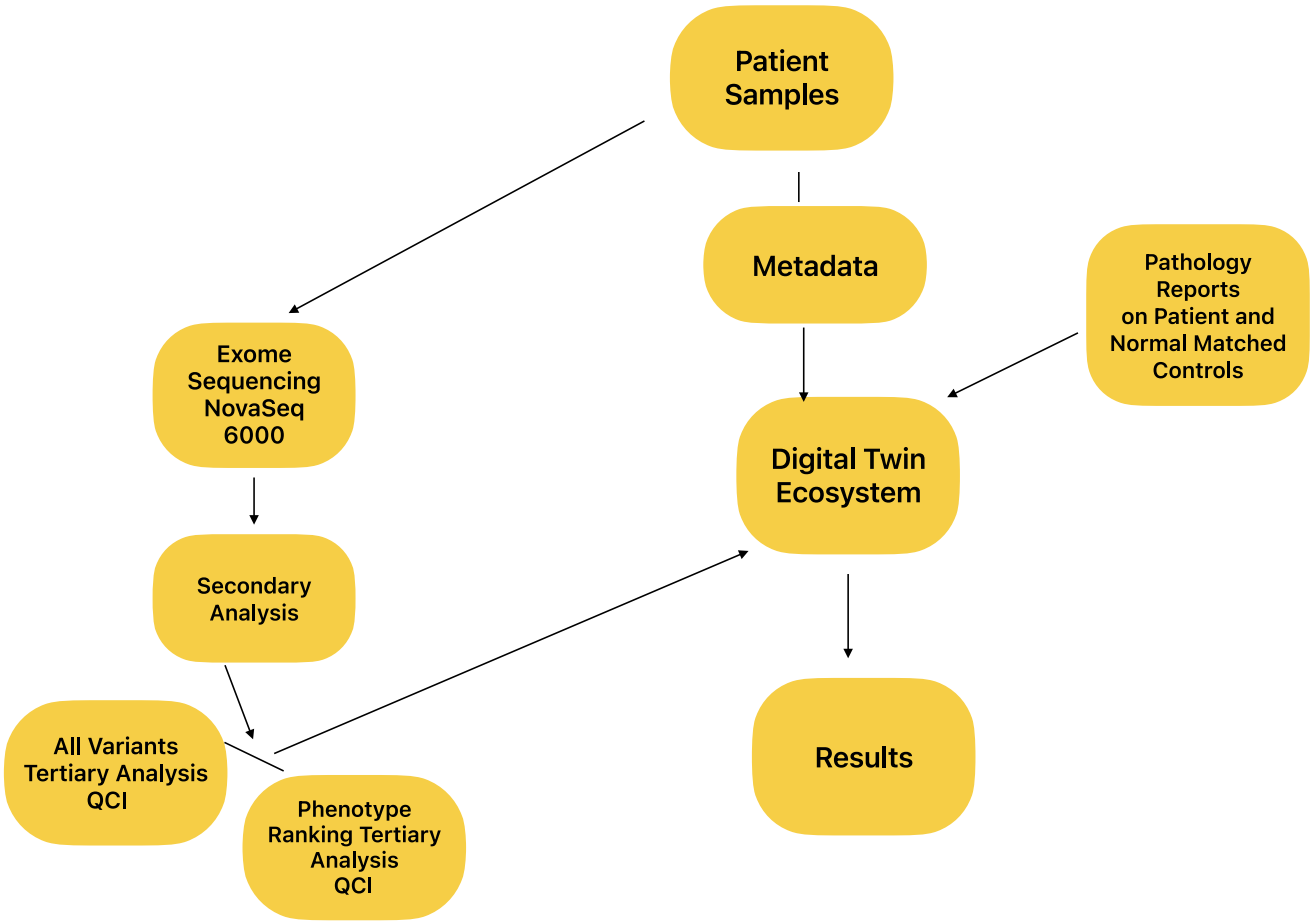

Figure 1 The experimental steps. QCI, QIAGEN's Clinical Insight.

Secondary Analysis

On sequencing, the authors used the Dragen Platform (Illumina) for secondary analysis. This enrichment is an accurate and efficient end-to-end [FASTQ-to-variant call format (VCF)] secondary analysis solution for whole-exome data. This application takes input files in FASTQ, BAM, and CRAM format. Files may be decompressed, go through map/align/sort, and go through variant calling using QCI.

Tertiary Analysis

For tertiary analysis, the authors downloaded all variants and all phenotype ranked variants for each sample using QCI (QIAGEN Digital Insights, Redwood City, CA). QCI Interpret is a clinical decision support software that accelerates variant interpretation and reporting of hereditary and oncology NGS tests at scale. QCI Interpret is powered by QIAGEN Knowledge Base, the biggest manually curated knowledge base, with insights about symptoms, phenotypes and gene-disease associations, biomedical databases, such as the Human Gene Mutation Database and the Catalogue of Somatic Mutations in Cancer, medical guidelines, and a wide variety of different bibliography content sources that are clinically relevant and are manually curated daily. QCI Interpret computes and combines all the relevant information related to the variant of interest, and distributes the relevant biological context. QCI Interpret over the past 20 years has been using >20 million curated findings and evidence and has analyzed >3 million clinical cases supported by AI and augmented molecular insights. Additionally, it offers the possibility of phenotype-driven analysis, where the user can submit phenotypes or symptoms of suspected disease or disease under investigation along with the .vcf file of the sample. On the basis of this information, the QCI Interpret phenotype-driven ranking algorithm estimates and ranks genomic variants based on the probability of being the causative one for the disease, symptoms, or the phenotypes under investigation by taking into account multiple variables, such as zygosity, predicted pathogenicity of variant, mode of inheritance, Combined Annotation Dependent Depletion (CADD) score, and more variant-centric variables as well as all the curated molecular insights from the QIAGEN Knowledge Base.

Knowledge Engineering Using a Biomimetic Engine[2–4,12–15]

- Each twin models a discrete component of the analytical scope of the ecosystem.
- Internal properties and behaviors must be modeled to a level of sufficient comprehensiveness to enable the reactions that are required for the ecosystem to reflect the real world to the scope of its design.
- Each twin can initiate an interaction with others or respond as prompted.
- Mitigation of bias is achieved by:
  - Independent design of each twin.
  - Abstract knowledge graphs populated without defining specific problems or events.
  - Autonomous interactions between the twins.

This real-world reasoning approach enables the construction of models that integrate highly diverse elements and information sources to enable exploration and discovery to a scope that traditional information architecture cannot accommodate.

Systems Thinking and Real-World Reasoning

The NAS recommends addressing complexity using systems thinking. Key observations are as follows:

- Bottom-up, mechanistic, linear approaches to understanding macro-level behavior are limited when considering complex systems.
- Bottom-up, reductionist hypotheses and approaches can lead to a proliferation of parameters; this challenge can potentially be addressed by applying top-down, system-level principles.
- Systems thinking can be used to predict macroscopic phenomena while bypassing the need to explicitly unmask all the quantitative dynamics operating at the microscopic level.

Although all knowledge engineering efforts seek to incorporate elements of cognitive science, a key aspect of the authors' innovation strategy is the driving role of a cognitive method, which is enabled by biomimetic information architectures. Brain processes are systemic and leverage what neuroscientists label plasticity and sparsity.

- Plasticity is the ability to engage diverse combinations of neurons and synapses by relevance to the purpose of the analysis, and to dynamically adapt internal functional architectures.
- Sparsity is the ability to identify the minimum data required. The brain can respond to situations that are simultaneously new on multiple dimensions and can even categorize one data point.

The neuronal and synaptic architecture of the brain is an ecosystem, which, according to the National Academies of Science, contains 100 trillion neurons. Systemic architecture, plasticity, and sparsity are core to biological learning, but are not similar to ML algorithms. The biomimetic technologies that enable elements of real-world reasoning are as follows.

- Expertise graphs.
- Neural system dynamics digital twins.

Researchers can imitate principles of plasticity and sparsity by implementing qualitative expertise graphs and leveraging them for contextual selection of data and methods from the in-memory model library. Unlike the deterministic methods to which traditional application

engineering is limited of necessity, systemic modeling requires the coexistence of chaotic and stochastic model elements, as well as their ability to dynamically interact with the deterministic elements.

For several years, AI has been looked to as the leading pathway to genetic understanding and drug development. However, deep learning and natural language processing have three key challenges that are addressed by the biomimetic digital twin ecosystem method presented in this article.

Biomimetic Digital Twin Ecosystem Process Tailored for the Analysis

i) Each patient endometrial DNA sample and matched control underwent exome sequencing, secondary analysis, and tertiary analysis. All DNA variants and phenotype ranked variants were exported to the digital twin ecosystem's data lake (Figure 2).
ii) Expert knowledge graphs were produced listing all previously reported DNA variants potentially associated with the pathophysiology of endometriosis and were exported to the digital twin ecosystem's data lake.
iii) Expert knowledge graphs were produced from pathology reports on each endometriosis sample and were exported to the digital twin ecosystem's data lake.
iv) Expert knowledge graphs were produced from each patient medical record and exported to the digital twin ecosystem's data lake.
v) The digital twin ecosystem's biomimetic engine then combined all data downloaded from Clinical Insight, including *in silico* calculations, phenotype ranked references, and multifactor correlations to the generated knowledge graphs, and produced a list of gene variants classified as VUSs potentially associated with the pathophysiology of endometriosis.
vi) The digital twin ecosystem's biomimetic engine ranked all VUSs according to the number of times that they were present in patient samples but absent from controls and provided output on genes that mapped to the same chromosome arm.
vii) The digital twin ecosystem's biomimetic engine's output pinpointed four genes, with DNA variants classified as VUSs, and phenotypes potentially associated with the pathophysiology of endometriosis, endometrial cancer, or an endometrial form of ovarian cancer.
viii) The digital twin ecosystem process output data were uploaded into GeneCards and VarElect to confirm results.
    a. GeneCards is a searchable, integrative database that provides comprehensive information on all annotated and predicted human genes. The knowledge base automatically integrates gene-centric data from approximately 150 web sources, including genomic, transcriptomic, proteomic, genetic, clinical, and functional information.[16,17]
    b. VarElect is a comprehensive phenotype-dependent DNA variant/gene prioritizer that can identify causal DNA variants with phenotypes. VarElect provides search and scoring capabilities, proficiently matching DNA variant-containing genes to submitted disease/symptom/phenotype keywords. The VarElect algorithm infers direct as well as indirect links between genes and phenotypes.[18]
ix) The digital twin ecosystem's biomimetic engine does not make recommendations or draw conclusions, but rather provides researchers with evidence for consideration that is not visible to AI or traditional bioinformatics platforms or approaches.

Statistical Analysis

Although statistical methods may ordinarily be applied to the data at this stage in the analysis, the authors' method enables researchers to discover real-world evidence that they cannot find using standard research software, including ML/AI tools. Assessing the statistical significance of the

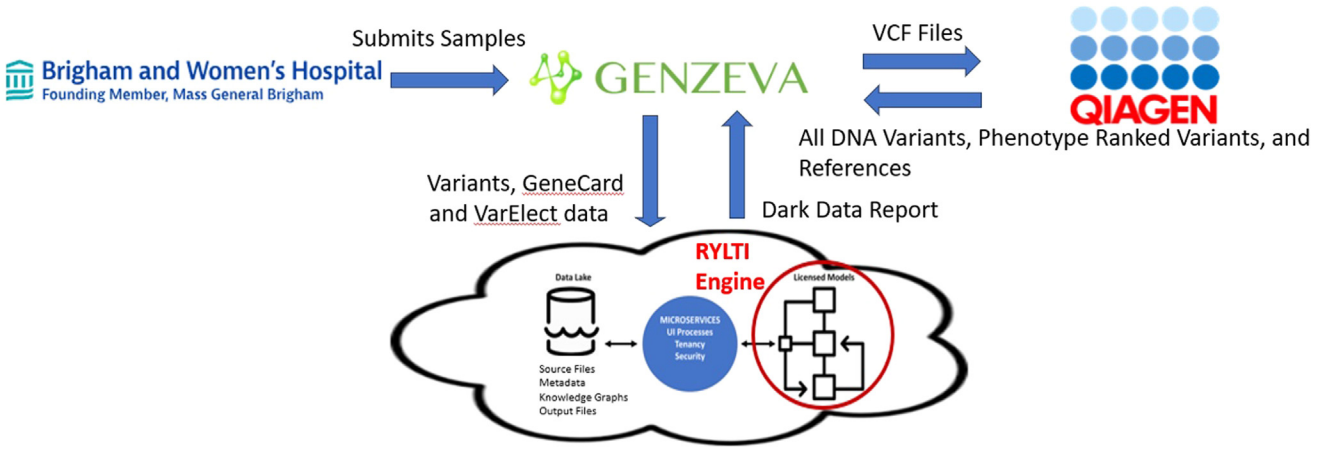

Figure 2 Biomimetic digital twin ecosystem process tailored for the analysis. UI, user interface.

evidence, if desired, can be performed, but the calculations depend on the researcher's hypotheses in combination with other available evidence. The use of *P* values and associated methods is not without controversy.[19] The authors' approach delivers the results and the supporting evidence, and adding a statistical component to the outcome could reduce the clarity of the results and possibly add bias.

Results

Result 1: Phenotype Ranking of DNA Variants

All DNA variants from QIAGEN's Clinical Insight from each patient's endometriotic and control tissue sample pairs were downloaded. Pathogenic mutations were identified in 8 of 12 patient samples in the following genes: *CPB1* (c.516C > A), *CD36* (c.447_450dupTCAA), *AXDND1* (c.125dupA), *PROKR2* (c.254G > A), *PKN1* (c.1663C > A), *DUOX2* (c.2654G > T), *CHST15* (c.1366dupC), and *ATP6V1B1* (c.988G > A).

All of these mutations are associated with endometriosis, endometrial cancer, or an endometrial form of ovarian cancer. No pathogenic mutations were identified in normal patient-matched control samples.

CPB1 (c.516C > A). This mutation is found in some patients with ovarian cancer in the analysis of germline and somatic variants in ovarian cancer.[1]

CD36 (c.447_450dupTCAA). Down-regulation of C*D36* results in reduced phagocytic ability of peritoneal macrophages of women with endometriosis.

AXDND1 (c.125dupA). Mutations in this gene are associated with ovarian cancer.

PROKR2 (c.254G > A). This mutation is associated with Kallmann syndrome. Kallmann syndrome combines an impaired sense of smell with a hormonal disorder that delays or prevents puberty. This hormonal disorder is due to the underdevelopment of specific neurons, or nerves, in the brain that signal the hypothalamus.

PKN1 (c.1663C > A). This mutation is associated in the development of cancer in populations of women studied with severe endometriosis.

DUOX2 (c.2654G > T). Hypoxia-inhibited dual-specificity phosphatase-2 expression in endometriotic cells regulates cyclooxygenase-2 expression and is thought to be associated with the development of endometriotic lesions.

CHST15 (c.1366dupC). *CHST15* expression in tissue is thought to be a prognostic factor of tumor cancer antigens in patients with endometrial cancer.

ATP6V1B1 (c.988G > A). This mutation is associated with epithelial ovarian cancer.

Result 2: Combining Phenotype Ranking and Digital Twin Analysis

All phenotype ranked variants were downloaded using specific key terms that described the phenotype of endometriosis, endometrial cancer, or an endometrial form of ovarian cancer using a phenotype-driven ranking filter (QIAGEN Clinical Insight Interpret) for each patient sample and matched controls. The data were then exported into their biomimetic digital twin ecosystem. Hidden or dark data for DNA variants were identified in four genes classified as VUSs in patient samples but not found in patient matched controls.

The VUSs identified were in genes *MUC 20* (12/12), *USP17L1* (8/12), FAM66B (8/12), and *DEFB109B* (12/12).

MUC20. *MUC20* polymorphisms, especially rs10794288 and rs10902088, are associated with endometriosis as well as endometriosis-related infertility.

USP17L1. This is predicted to enable cysteine-type endopeptidase activity and thiol-dependent deubiquitinase. It is predicted to be involved in protein deubiquitination and regulation of apoptotic process. Atypical regulation of apoptosis could be involved in the pathophysiology of endometriosis.

FAM66B. A long noncoding RNA, from the family of regulatory noncoding RNAs. Mechanistic studies indicate that long noncoding RNAs may regulate genes involved in endometriosis by acting as a molecular sponge for miRNAs, by directly targeting regulatory elements via interactions with chromatin or transcription factors, or by affecting signaling pathways.

DEFB109B. Understanding the biology of endometrial stem cell populations is important for defining normal and abnormal endometrial tissue regeneration and lineage cell commitment. This gene plays a role in the transmission of abnormalities across cell lineages and contributes to proliferative disorders, such as endometrial polyps, endometriosis, and endometrial hyperplasia/cancer.

Discussion

We believe this is the first study incorporating QIAGEN Clinical Insights Interpret phenotype-driven ranking filters with knowledge engineering via the use of a biomimetic digital twin ecosystem and genomic analysis, to provide greater understanding of the molecular mechanism of disease.

In this study, eight pathogenic mutations associated with the pathophysiology of endometriosis, endometrial cancer, or an endometrial form of ovarian cancer were identified in 8 of 12 patient samples analyzed. Additionally, QIAGEN's Clinical Insight and our biomimetic digital twin ecosystem identified four VUSs also associated with the development of endometriosis-related disorders.

One VUS, in the *MUC20* gene, maps to chromosome 3, and was identified in all 12 patient samples analyzed. *MUC20* polymorphisms, especially rs10794288 and rs10902088, are associated with endometriosis as well as endometriosis-related infertility.

The other identified VUSs in genes *USP17L1*, FAM66B, and *DEFB109B* all mapped to the short arm of chromosome 8.

USP17L1 is predicted to enable cysteine-type endopeptidase activity and thiol-dependent deubiquitinase. This gene is involved in protein deubiquitination and the regulation of the apoptotic process. Atypical regulation of apoptosis is hypothesized to be involved in the pathophysiology of endometriosis.

FAM66B is a long noncoding RNA, a type of regulatory noncoding RNA. Mechanistic studies indicate that long noncoding RNAs may regulate genes involved in endometriosis by acting as a molecular sponge for miRNAs, by directly targeting regulatory elements via interactions with chromatin or transcription factors or by affecting signaling pathways.

DEFB109B is involved in normal and abnormal endometrial tissue regeneration and lineage cell commitment. Transmission of abnormalities across cell lineages may contribute to proliferative disorders, such as endometrial polyps, endometriosis, and endometrial hyperplasia/cancer.

Significantly, a potential biomarker associated with endometriosis was identified and the presence of a VUS was demonstrated within the *MUC20* gene in all patient samples.

Furthermore, expression of the *DEFB109B* and FAM66B genes is regulated by the same enhancer sequence. These sequences are regulatory DNA sequences that, when bound by transcription factors, enhance the transcription of an associated gene(s). This suggests a possible hot spot on the short arm of chromosome 8 that could be associated with the molecular pathophysiology of endometriosis.

The identification of these VUSs does not confirm that they play a role in the development of endometrial-related disorders. The digital twin ecosystem was able to uncover variants classified as VUSs on genes potentially associated with endometriosis. GeneCards and VarElect were also incorporated into these analyses. GeneCards provides comprehensive information on all annotated and predicted human genes. VarElect is a comprehensive phenotype-dependent DNA variant/gene prioritizer that can identify causal DNA variants with phenotypes. VarElect provides search and scoring capabilities, proficiently matching DNA variant-containing genes to submitted disease/symptom/phenotype keywords. The VarElect algorithm infers direct as well as indirect links between genes and phenotypes.

These results provide evidence that the identified VUSs are highly likely to play some role in the pathophysiology of endometriosis, endometrial cancer, and an endometrial form of ovarian cancer. Furthermore, we describe an innovative way to potentially reclassify VUSs.

For several years, traditional AI has been looked upon as the leading pathway to genomic understanding and drug development. However, the three key challenges of deep learning and natural language processing are addressed by the biomimetic digital twin ecosystem method. These challenges include instability, blindness to dark data, and risks and biases that are being challenged by the US Food and Drug Administration (*https://www.fda.gov/news-events/fda-voices/fda-releases-two-discussion-papers-spur-conversation-about-artificial-intelligence-and-machine*, last accessed January 27, 2024).

Gartner, Inc., coined the phrase "dark data" to describe organizational information assets that are excluded from analytical processes, leaving the associated insights invisible and the value unharvested.

Addressing each of the above three limitations requires finding and connecting biological dark data. Standard AI is blind to the dark data because algorithms can only find what they have been engineered to find. Furthermore, AI can only work within the narrow limits of the algorithm's training data, which is large (many rows), narrow (limited attributes), and cleansed (outliers removed). These factors greatly diminish the ability to identify high-value insights.

Shedding light on the darkened insights requires small/wide data methods. Wide data allow analysts to examine and combine a variety of small and large attributes from diverse sources, whereas small data are focused on applying analytical techniques that look for useful information within limited sets of data.

The biomimetic digital twin ecosystem architecture enables finding and connecting dark data because:

- Each twin models a discrete component of the analytical scope of the ecosystem.
- Internal properties and behaviors must be modeled to a level of sufficient comprehensiveness to enable the reactions that are required for the ecosystem to reflect the real world to the scope of its design.
- Each twin can initiate an interaction with others or respond as prompted.
- Mitigation of bias is achieved by:
  - Independent design of each twin.
  - Abstract knowledge graphs populated without defining specific problems or events.[14]
  - Autonomous interactions between the twins.

All methods have limitations, including digital twins. One limitation is the requirement of expert knowledge graphs to drive the digital twin engine. Expert knowledge graphs are prepared by subject matter experts in their field of scientific discovery.

The outputs of our digital twin ecosystem are arrays of scenarios with associated evidence and predictions. Traditional AI/ML/LLM also produces output predictions and the outputs from the various methods should be, in the future, compared closely. The comparison results will be valuable in identifying additional gaps in biomedical research.

One potential limitation of these results is that only 12 patient samples and matched normal controls were analyzed in this study. However, a biomimetic digital twin ecosystem is powerful in its ability to analyze small, wide data sets and identify dark data. Another potential limitation is that all of the patients were of White descent. Additional studies are required to discover the role of ethnicity, if any, in the pathogenesis of endometrial-related disorders.

In conclusion, using our advanced genomic process, including exome sequencing, QIAGEN's Clinical Insight, and a biomimetic digital twin ecosystem, can help understand the molecular mechanism of disease. The current analysis identified a potential biomarker for a molecular test for endometriosis. These data also identified a potential hot spot for the molecular study of endometriosis on the short arm of chromosome 8. The combination of a knowledge engineering platform and comprehensive molecular analyses can be used for the identification of molecular mechanisms for any disease. It can include and clarify the role of factors such as ethnicity in the severity of disease, perform virtual clinical trials, and aid in the rapid identification of new therapies for the effective treatment of disease.

Disclosure Statement

W.G.K. and L.K. are owners of Genzeva and LumaGene, and they own stock in RYLTI, LLC, of which RYLTI BioPharma is a subsidiary. L.B. and J.G. own RYLTI, LLC, stock, of which RYLTI BioPharma is a subsidiary.

References

1. Kulkarni PA, Singh H: Artificial intelligence in clinical diagnosis: opportunities, challenges, and hype. JAMA 2023, 330:317−318
2. National Academies of Sciences, Engineering, and Medicine: Opportunities and Challenges for Digital Twins in Biomedical Research. Washington, DC: National Academies Press, 2022. doi: 10.17226/26922
3. National Academies of Sciences, Engineering, and Medicine: Physics of Life. Washington, DC: The National Academies Press, 2022
4. National Academies of Sciences, Engineering, and Medicine: Foundational Research Gaps and Future Directions for Digital Twins. Washington, DC: The National Academies Press, 2023. doi: 10.17226/26894
5. Zondervan KT, Becker CM, Missmer SA: Endometriosis. N Engl J Med 2020, 382:1244−1256
6. Nnoaham KE, Hummelshoj L, Webster P, d'Hooghe T, de Cicco Nardone F, de Cicco Nardone C, Jenkinson C, Kennedy SH, Zondervan KT: Impact of endometriosis on quality of life and work productivity: a multicenter study across ten countries. Fertil Steril 2011, 96:366−373
7. Saha R, Pettersson HJ, Svedberg P, Olovsson M, Bergqvist A, Marions L, Tornvall P, Kuja-Halkola R: Heritability of endometriosis. Fertil Steril 2015, 104:947−952
8. Treloar SA, O'Connor DT, O'Connor VM, Martin NG: Genetic influences on endometriosis in an Australian twin sample. Fertil Steril 1999, 71:701−710
9. Czyzyk A, Podfigurna A, Szeliga A, Meczekalski B: Update on endometriosis pathogenesis. Minerva Ginecol 2017, 69:447−461
10. Lee SH, Harold D, Nyholt DR; Gene Consortium; International Endogene Consortium; Genetic and Environmental Risk for Alzheimer's Disease Consortium, Goddard ME, Zondervan KT, Williams J, Montgomery GW, Wray NR, Visscher PM: Estimation and partitioning of polygenic variation captured by common SNPs for Alzheimer's disease, multiple sclerosis and endometriosis. Hum Mol Genet 2013, 22:832−841
11. Rahmioglu N, Mortlock S, Ghiasi M, Moller PR, Stefansdottir S, Galarneau G, et al: The genetic basis of endometriosis and comorbidity with other pain and inflammatory conditions. Nat Genet 2023, 55: 423−436
12. National Academies of Sciences, Engineering, and Medicine: Applying Systems Thinking to Regenerative Medicine: Proceedings of a Workshop. Washington, DC: The National Academies Press, 2021
13. National Academies of Sciences, Engineering, and Medicine: Closing Evidence Gaps in Clinical Prevention. Washington, DC: National Academies Press, 2022
14. Rottman BM, Genter D, Goldwater MB: Causal systems categories: differences in novice and expert categorizations of causal phenomena. Cogn Sci 2012, 36:919−932
15. Spivak DI, Kent RE: Ologs: a categorical framework for knowledge representation. PLoS One 2012, 7:e24274
16. Stelzer G, Rosen R, Plaschkes I, Zimmerman S, Twik M, Fishilevich S, Iny Stein T, Nudel R, Lieder I, Mazor Y, Kaplan S, Dahary D, Warshawsky D, Guan-Golan Y, Kohn A, Rappaport N, Safran M, Lancet D: The GeneCards suite: from gene data mining to disease genome sequence analyses. Curr Protoc Bioinformatics 2016, 54:1.30.1−1.30.33
17. Safran M, Rosen N, Twik M, BarShir R, Iny Stein T, Dahary D, Fishilevich S, Lancet D: The GeneCards Suite. Edited by Abugessaisa I, Kasukawa, T. Practical Guide to Life Science Databases. Singapore: Springer, 2021. pp. 27−56
18. Stelzer G, Plaschkes I, Oz-Levi D, Alkelai A, Olender T, Zimmerman S, Twik M, Belinky F, Fishilevich S, Nudel R, Guan-Golan Y, Warshawsky D, Dahary D, Kohn A, Mazor Y, Kaplan S, Iny Stein T, Baris H, Rappaport N, Safran M, Lancet D: VarElect: the phenotype-based variation prioritizer of the GeneCards Suite. BMC Genomics 2016, 17(Suppl 2):444
19. Wasserstein RL, Lazar NA, Lazar NA: The ASA statement on p-values: context, process, and purpose. Am Statistician 2016, 70: 129−133