medRxiv preprint doi: https://doi.org/10.1101/2024.05.03.24306738; this version posted May 5, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

A New Differential Gene Expression Based Simulated Annealing for Solving Gene Selection Problem: A Case Study on Eosinophilic Esophagus and Few Other Gastro-Intestinal Diseases

Koushiki Sinha, Sanchari Chakraborty, Arohit Bardhan, Riju Saha, Srijan Chakraborty, Surama Biswas*

Abstract:

Identification of the set of genes collectively causes a disease is an important problem, called gene selection problem. This study introduces two distinct approaches for gene selection in the context of biological diseases: the Ranked Variance (RV) method and Differential Gene Expression Based Simulated Annealing (DGESA). The RV method prioritizes genes based on their variance, offering an initial perspective on potential biomarkers. DGESA, on the other hand, employs simulated annealing, integrating differential gene expression data to refine gene selection further. Through a case study focused on Eosinophilic Esophagus (EoE) and other gastro-intestinal diseases, we compare and contrast the outcomes of both methods. Notably, we identify 10 common genes between RV and DGESA in EoE, highlighting their complementary nature. Validation analyses reveal that 13 out of 40 final genes identified by DGESA for EoE are corroborated by existing literature, indicating their biological relevance. Similarly, in Ulcerative Colitis (UC) and Crohn's Disease (CD), 8 and 7 genes, respectively, out of the final 40 genes identified exhibit confirmation in the literature. These findings underscore the efficacy of both RV and DGESA in elucidating molecular signatures associated with gastro-intestinal diseases, contributing to our understanding of their pathogenesis and potential therapeutic targets.

Introduction:

Computational genomics stands at the intersection of biology and computer science, harnessing advanced algorithms and computational tools to dissect genetic data with unprecedented precision. This interdisciplinary field endeavours to unravel the intricate mechanisms governing biological processes at the genomic level, ranging from DNA sequencing to gene expression analysis and variant identification. At the forefront of assessing the efficacy of computational methods in deciphering genetic variant impacts on phenotypes lies the Critical Assessment of Genome Interpretation (CAGI) [1]. Through a series of community experiments spanning five rounds and comprising 50 challenges, CAGI has emerged as a pivotal platform for evaluating the performance of computational models in predicting the effects of genetic variants, particularly within disease contexts. The global participation in CAGI, with 738 submissions from diverse backgrounds, underscores the significance and widespread interest in advancing computational genomics.

Despite the remarkable progress demonstrated by computational methods in predicting clinical pathogenic variants and enhancing the accuracy of estimating biochemical effects of missense variants, challenges persist in certain domains of genomic interpretation. Notably, assessing the impact of regulatory variants and predicting disease risks associated with complex traits remain formidable tasks that warrant further exploration and refinement. However, amidst these challenges, the utility of current methodologies in both research and potential clinical applications remains undeniable. The field continues to evolve rapidly, driven by the integration of emerging computational techniques and access to vast datasets for training and evaluation [2]. As CAGI's initiatives continue to steer the trajectory of genomic interpretation

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2024.05.03.24306738; this version posted May 5, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

research, there is an ever-growing opportunity to shape future research directions, refine clinical practices, and foster collaborative endeavors within the computational genomics community. Through these concerted efforts, computational genomics is poised to unlock deeper insights into the genetic underpinnings of diseases and pave the way for more effective therapeutic interventions and personalized healthcare approaches.

Gene expression data serves as a comprehensive repository of information that offers detailed insights into the dynamic utilization of genes within cells or organisms at specific moments. At its core, gene expression data unveils the intricate orchestration of biological processes by shedding light on which genes are actively engaged (expressed) and the extent to which they undergo transcription into RNA molecules. This transcriptional activity ultimately influences the production of proteins, thereby dictating the functionality and behavior of cells within an organism. As such, gene expression data plays a pivotal role in unravelling the molecular mechanisms underlying various physiological and pathological states.

One of the primary experimental techniques employed to generate gene expression data is microarray analysis. Microarrays empower researchers to concurrently measure the expression levels of thousands of genes by detecting the binding of RNA molecules to complementary DNA probes immobilized on a solid surface. This high-throughput approach enables comprehensive profiling of gene expression patterns across diverse biological samples, facilitating the identification of molecular signatures associated with specific conditions or experimental interventions [3]. Additionally, RNA sequencing (RNA-seq) has emerged as a powerful tool for transcriptome analysis, offering unparalleled sensitivity and resolution in quantifying gene expression levels. By sequencing RNA molecules extracted from biological samples, RNA-seq provides a comprehensive snapshot of the transcriptome, encompassing both protein-coding and non-coding RNA species [4]. Furthermore, quantitative polymerase chain reaction (qPCR) assays offer a precise and sensitive method for quantifying gene expression levels in a targeted manner, making them invaluable for validation studies and high-throughput screening assays.

The availability of gene expression data is bolstered by a plethora of public repositories, which serve as invaluable resources for the scientific community. Among these repositories, the Gene Expression Omnibus (GEO), ArrayExpress, and the National Center for Biotechnology Information (NCBI)[20] stand out as prominent platforms for sharing and disseminating gene expression datasets. These repositories host an extensive collection of datasets spanning various diseases, tissue types, experimental conditions, and organisms, thereby providing researchers with a rich and diverse pool of data for exploration and analysis. Leveraging these repositories, researchers can access curated gene expression datasets generated from a wide range of experimental techniques and biological systems, enabling comparative analyses, meta-analyses, and hypothesis-driven investigations. Additionally, these repositories often incorporate sophisticated data visualization and analysis tools, empowering researchers to extract meaningful insights from complex gene expression datasets and uncover novel biological findings.

Hence, gene expression data serves as a cornerstone of modern molecular biology, offering unprecedented insights into the dynamic regulation of gene activity and its impact on cellular function and phenotype. By harnessing advanced experimental techniques such as microarrays, RNA-seq, and qPCR, researchers can profile gene expression patterns with remarkable precision and scale. Furthermore, public repositories such as GEO, ArrayExpress, and NCBI [20] play a pivotal role in democratizing access to gene expression data, fostering collaboration, and accelerating scientific discoveries in diverse fields ranging

medRxiv preprint doi: https://doi.org/10.1101/2024.05.03.24306738; this version posted May 5, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

from basic biology to translational medicine. As our understanding of gene expression dynamics continues to evolve, gene expression data will remain a cornerstone of biological research, driving innovation and unlocking new avenues for therapeutic intervention and personalized medicine.

Gastrointestinal disorders represent a heterogeneous group of conditions that impact the functioning of the digestive system, encompassing a spectrum of ailments from functional disturbances like irritable bowel syndrome (IBS) to more inflammatory and allergic reactions such as inflammatory bowel disease (IBD) and eosinophilic esophagitis (EoE). These disorders manifest through a myriad of symptoms, including but not limited to abdominal pain, diarrhea, constipation, bloating, and alterations in bowel habits, significantly impairing patients' quality of life and posing substantial healthcare burdens [5].

A key feature of gastrointestinal disorders is their multifaceted etiology, which often involves a complex interplay of genetic predisposition, environmental triggers, immune dysregulation, and disruptions in the gut microbiota composition. Genetic factors play a significant role in predisposing individuals to certain gastrointestinal conditions, with studies implicating various genetic polymorphisms and susceptibility loci in the pathogenesis of disorders like IBD and IBS. Moreover, environmental factors such as diet, lifestyle, stress, and exposure to pathogens or toxins can profoundly influence disease development and progression. The intricate interplay between genetic and environmental factors underscores the heterogeneity and diverse clinical presentations observed among patients with gastrointestinal disorders.

Immune system dysregulation represents another critical component in the pathophysiology of gastrointestinal disorders, particularly those with an inflammatory component like IBD and EoE. Dysfunctional immune responses, characterized by aberrant activation of pro-inflammatory pathways or impaired regulatory mechanisms, contribute to chronic inflammation, tissue damage, and disease exacerbations. Furthermore, alterations in the gut microbiota composition, termed dysbiosis, have emerged as a significant factor in the pathogenesis of gastrointestinal disorders. Dysbiosis refers to disruptions in the balance of microbial communities residing in the gastrointestinal tract, leading to alterations in immune homeostasis, metabolic processes, and barrier integrity. Dysbiotic microbiota profiles have been implicated in the pathogenesis of IBD, IBS, and other gastrointestinal conditions, highlighting the intricate interplay between host genetics, environmental factors, and microbial dysbiosis in disease pathogenesis [6].

Gastrointestinal disorders, hence, encompass a diverse array of conditions characterized by disturbances in digestive system function and a broad spectrum of clinical manifestations. Understanding the multifactorial nature of these disorders, including the contributions of genetic susceptibility, environmental influences, immune dysregulation, and alterations in gut microbiota composition, is essential for elucidating their pathophysiology and developing targeted therapeutic strategies. By unravelling the complex interplay of these factors, researchers and clinicians can pave the way for personalized approaches to diagnosis, treatment, and management, ultimately improving outcomes and quality of life for individuals affected by gastrointestinal disorders.

Eosinophilic Esophagitis (EoE) stands as a chronic immune-mediated disorder characterized by inflammation within the esophagus, a result of eosinophil accumulation induced by allergic reactions. Its clinical presentation encompasses a range of symptoms, including dysphagia (difficulty swallowing), food impaction, chest pain, and heartburn, which significantly impact patients' daily lives. Diagnosis typically involves a combination of endoscopic evaluation and histological examination through biopsy, aimed at confirming the presence of eosinophilic infiltration within the esophageal tissue. Treatment strategies for

medRxiv preprint doi: https://doi.org/10.1101/2024.05.03.24306738; this version posted May 5, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

EoE are multifaceted, often starting with dietary modifications to identify and eliminate potential trigger foods, followed by pharmacological interventions such as proton pump inhibitors to reduce acid reflux and topical steroids to alleviate inflammation. In cases of severe esophageal strictures or stenosis, esophageal dilation procedures may be required to alleviate symptoms and improve swallowing function. Moreover, EoE frequently coexists with other allergic conditions, most notably asthma and eczema, underscoring the complex interplay between immune dysregulation and allergic predisposition in disease pathogenesis. This interconnectedness between allergic diseases highlights the importance of comprehensive patient evaluation and management to address potential comorbidities and optimize treatment outcomes. Additionally, ongoing research efforts aim to elucidate the underlying mechanisms driving EoE development and progression, with a particular focus on identifying novel therapeutic targets and precision medicine approaches tailored to individual patient profiles [7].

Expanding beyond EoE, eosinophilic gastrointestinal disorders (EGIDs) comprise a spectrum of conditions affecting the gastrointestinal (GI) tract, characterized by eosinophil-rich inflammation within various segments of the digestive system. EGIDs, including eosinophilic gastritis and eosinophilic colitis, have garnered increased recognition in recent years, with a growing prevalence linked to both allergic sensitization and genetic predisposition [4]. Immune dysregulation, particularly involving cytokines such as interleukin-5 (IL-5), plays a central role in driving eosinophilic infiltration and tissue inflammation within the GI tract. Diagnosis of EGIDs typically necessitates endoscopic evaluation with biopsy sampling to assess eosinophilic infiltration and exclude alternative etiologies of GI symptoms [5, 8].

Treatment strategies for EGIDs encompass a combination of dietary modifications, pharmacotherapy, and, in select cases, endoscopic interventions aimed at alleviating inflammation and improving GI function. However, significant gaps remain in our understanding of EGID pathogenesis and optimal treatment approaches, highlighting the need for continued research efforts to unravel the complex interplay between immune dysregulation, allergic sensitization, and genetic factors driving disease onset and progression [5, 7]. By advancing our understanding of EGIDs, clinicians and researchers can pave the way for more effective diagnostic strategies and personalized therapeutic interventions, ultimately enhancing patient care and quality of life for individuals affected by these challenging conditions.

Investigations into the prevalence of eosinophilic esophagitis (EoE) within patient populations exhibiting symptoms of gastroesophageal reflux have shed light on a noteworthy finding: a substantial proportion of individuals within this cohort are affected by EoE [5]. This discovery underscores the significance of incorporating EoE into the diagnostic assessment of patients presenting with symptoms suggestive of gastroesophageal reflux, particularly within the realm of tertiary care facilities in North India [9]. As gastroesophageal reflux symptoms can overlap with those of EoE, the inclusion of EoE in diagnostic considerations becomes paramount for ensuring comprehensive and accurate patient evaluations. By recognizing the potential coexistence of EoE alongside gastroesophageal reflux symptoms, healthcare providers can enhance their ability to identify and address underlying conditions contributing to patient morbidity and optimize therapeutic strategies accordingly. This finding serves as a crucial reminder of the intricate diagnostic challenges inherent in gastroenterological practice and highlights the importance of maintaining a high index of suspicion for EoE, particularly in regions where its prevalence may be underestimated or under-recognized. Ultimately, incorporating EoE into the diagnostic algorithm for patients presenting with gastroesophageal reflux symptoms can facilitate timely and appropriate management, ultimately improving patient outcomes and quality of life.

medRxiv preprint doi: https://doi.org/10.1101/2024.05.03.24306738; this version posted May 5, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Crohn's Disease (CD) stands as a chronic and debilitating inflammatory bowel disorder characterized by pervasive inflammation throughout the gastrointestinal tract, heralding a myriad of distressing symptoms including recurrent abdominal pain, persistent diarrhoea, fatigue, and malnutrition. This condition poses a considerable clinical burden, given its propensity to affect individuals across various age groups and significantly impair their quality of life. The pathogenesis of CD is complex and multifactorial, implicating a confluence of genetic predispositions, dysregulated immune responses, and environmental triggers. Such intricate interplay underscores the heterogeneous nature of CD and contributes to the variability observed in its clinical presentation and disease course [10].

Central to the management of CD is the pursuit of strategies aimed at mitigating inflammation and alleviating associated symptoms, thereby fostering disease remission and enhancing patient well-being. In this regard, therapeutic interventions encompass a diverse array of pharmacological agents, including but not limited to corticosteroids, immunosuppressants, biologics, and targeted immune modulators. These interventions are meticulously tailored to suit the individualized needs and preferences of patients, with a keen focus on optimizing therapeutic efficacy while minimizing adverse effects. Moreover, lifestyle modifications such as dietary adjustments, stress management techniques, and smoking cessation may complement pharmacotherapy in augmenting treatment outcomes and promoting long-term disease control.

Despite significant strides in the therapeutic armamentarium available for managing CD, the condition remains inherently complex and poses considerable challenges in clinical practice. Persistent inflammation and disease relapse are common occurrences, necessitating a nuanced and multidisciplinary approach to disease management that emphasizes regular monitoring, proactive symptom management, and patient education. Furthermore, ongoing research endeavors aimed at unraveling the intricate pathophysiological mechanisms underlying CD hold promise for uncovering novel therapeutic targets and refining treatment strategies [6, 10]. By harnessing insights gleaned from cutting-edge research and integrating them into clinical practice, healthcare providers can strive towards achieving more personalized and effective management approaches tailored to the unique needs of patients with CD, ultimately fostering improved clinical outcomes and enhancing overall quality of life.

Ulcerative Colitis (UC), a prevalent inflammatory bowel disease (IBD), represents a formidable challenge in the realm of gastroenterology, characterized by its predilection for the colon and rectum and its hallmark symptoms including abdominal pain, bloody diarrhea, urgency, and weight loss. The pathogenesis of UC is intricately woven, implicating a complex interplay between genetic susceptibilities, environmental triggers, and dysregulated immune responses. Genetic predispositions, encompassing polymorphisms in genes involved in mucosal integrity and immune regulation, confer susceptibility to UC, while environmental factors such as diet, smoking, and microbial dysbiosis serve as potent triggers in susceptible individuals [11]. The management of UC is predicated on a multifaceted approach aimed at controlling inflammation, alleviating symptoms, and ultimately inducing sustained remission to mitigate disease burden and optimize patient outcomes. Pharmacological interventions form the cornerstone of UC management, with a diverse array of medications at the disposal of healthcare providers. Aminosalicylates, encompassing compounds such as mesalamine and sulfasalazine, serve as first-line agents in inducing and maintaining remission in mild-to-moderate UC, exerting their anti-inflammatory effects through modulation of mucosal immune responses and suppression of pro-inflammatory cytokines [11]. Corticosteroids, while efficacious in inducing remission in acute flares, are often reserved for short-term use due to their adverse effect profile and potential for long-term complications. Immune modulators such

medRxiv preprint doi: https://doi.org/10.1101/2024.05.03.24306738; this version posted May 5, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

as azathioprine, 6-mercaptopurine, and methotrexate are employed as steroid-sparing agents in refractory cases or as maintenance therapy to prevent disease relapse. Biologic agents, including tumor necrosis factor-alpha (TNF-α) inhibitors (e.g., infliximab, adalimumab), integrin antagonists (e.g., vedolizumab), and interleukin inhibitors (e.g., ustekinumab), represent a revolutionary paradigm shift in UC management, targeting specific immune pathways implicated in disease pathogenesis to achieve profound and sustained remission. Surgical intervention, while reserved for severe and refractory cases or complications such as toxic mega colon or colorectal cancer, offers definitive management by removing the diseased colon and rectum, thereby alleviating symptoms and improving quality of life.

Beyond its immediate clinical ramifications, UC underscores broader implications in the landscape of immune-mediated inflammatory diseases (IMIDs), representing a paradigmatic example of dysregulated immune responses and aberrant inflammatory cascades [6]. Indeed, IMIDs share common immune pathogenic mechanisms, underpinned by disruptions in intestinal homeostasis, dysbiosis of the gut microbiome, and perturbations in mucosal immune responses, all of which converge to drive sustained and aberrant inflammatory processes within the gastrointestinal tract. This shared immune pathogenic framework underpins the therapeutic rationale for targeting specific immune pathways implicated in IMIDs, transcending the confines of individual diseases to offer overarching therapeutic strategies with potential applicability across diverse IMID spectra.

In navigating the complex and multifaceted landscape of UC and its broader implications in IMIDs, it becomes evident that a comprehensive and integrative approach is imperative to address the diverse needs and challenges posed by these conditions. By leveraging insights gleaned from cutting-edge research, embracing emerging therapeutic modalities, and fostering interdisciplinary collaboration among clinicians, researchers, and patients, we can aspire towards achieving a holistic understanding of UC and its immune pathogenic underpinnings, thereby paving the way for more efficacious and personalized management strategies that optimize outcomes and enhance the quality of life for individuals affected by UC and related IMIDs.

Artificial Intelligence (AI) and Machine Learning (ML) have emerged as transformative forces in the realm of disease genetics, reshaping the landscape of gene selection and disease association studies. These technologies have revolutionized the analysis of vast genomic datasets, offering powerful tools for uncovering disease-causing genes and elucidating the intricate genetic underpinnings of complex disorders [12]. By leveraging advanced algorithms and computational techniques, AI/ML platforms facilitate the prediction of functional impacts of genetic variants and prioritize candidate genes based on a myriad of criteria, including evolutionary conservation, protein structure, and known biological pathways. This holistic approach enables researchers to navigate the complexity of the genome and identify key genetic players implicated in disease pathogenesis with unprecedented accuracy and efficiency.

One of the hallmark capabilities of AI/ML in genomics lies in its capacity to integrate diverse omics data types, including genomic, transcriptomic, and epigenetic data, to unravel complex relationships between genetic variations and disease phenotypes. By leveraging multimodal datasets, AI/ML algorithms can decipher intricate regulatory networks, identify disease-specific expression patterns, and uncover novel genetic interactions that underpin disease susceptibility and progression. This integrative approach not only enhances our understanding of the molecular mechanisms driving disease but also paves the way for the development of personalized medicine approaches tailored to individual patients' genetic profiles [12, 13].

medRxiv preprint doi: https://doi.org/10.1101/2024.05.03.24306738; this version posted May 5, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Moreover, AI/ML technologies hold immense promise in accelerating the pace of discovery in disease genetics, particularly in the context of infectious diseases. By automating data processing and analysis, these tools streamline the identification of disease-associated genes and facilitate the translation of genomic insights into actionable clinical interventions [12, 13]. In the realm of infectious disease management, AI offers unprecedented opportunities for early detection, accurate diagnosis, and targeted treatment strategies. Through sophisticated image processing algorithms and predictive modeling techniques, AI can analyze medical images, decipher complex host-pathogen interactions, and predict disease outcomes with remarkable precision.

However, despite the remarkable strides made in harnessing AI/ML for disease genetics and infectious disease management, significant challenges and limitations persist [12]. The complex and multifaceted nature of genetic data poses inherent challenges in data quality, standardization, and interpretation, which can impact the reliability and reproducibility of AI-driven analyses. Moreover, ethical considerations surrounding data privacy, algorithm bias, and clinical interpretation present formidable obstacles that must be addressed to ensure the responsible and equitable deployment of AI technologies in healthcare settings.

Nevertheless, the burgeoning role of AI in computational biology holds immense promise for addressing pressing global health challenges and advancing the frontiers of precision medicine. As we continue to harness the power of AI/ML to unlock the mysteries of the genome and decipher the complexities of infectious diseases, interdisciplinary collaboration and concerted research efforts will be paramount in realizing the full potential of these transformative technologies to revolutionize healthcare delivery and improve patient outcomes on a global scale. Studies have revealed that functional gastrointestinal disorders present a complex challenge due to their symptom-based nature, lacking definitive biomarkers or understood pathophysiology. Enhancing our comprehension of the genetic underpinnings of these disorders holds promise for elucidating their intricate biology and explaining their frequent co-occurrence with persistent pain, mood disorders, and affective conditions. This understanding could potentially facilitate the identification of patient subgroups responsive to personalized therapeutic interventions. Unlike monogenic diseases, these disorders are polygenic, involving the influence of common variants across numerous genes, alongside environmental factors, in shaping individual susceptibility. While family and twin studies have underscored the genetic component in conditions like irritable bowel syndrome (IBS), efforts to link specific gene polymorphisms to the syndrome have been challenged by small sample sizes, lack of reproducibility in larger datasets, and variability in clinical phenotype reliability. Advancing our understanding in this field necessitates refining intermediate phenotypes with substantial effect sizes for the clinical phenotype, while also exploring gene-gene, environment-gene (epigenetics), and sex-gene interactions [14]. Utilizing genome-wide association studies and whole-genome sequencing in extensive datasets represents a promising avenue for future progress, offering insights into the genetic landscape of functional gastrointestinal disorders like eosinophilic esophagitis (EOE) [15].

Studies have illuminated the role of genetic mutations in severe atopy syndrome, particularly those related to DSG1, shedding light on the genetic basis of severe atopic disorders. Additionally, associations between genetic mutations linked to syndromes such as Loeys-Dietz and Ehlers-Danlos and the onset of EoE have been investigated, providing insights into the interplay between genetic abnormalities and EoE severity. Genome-wide association studies (GWAS) have identified EoE-risk loci, including CCL26, FLG,

medRxiv preprint doi: https://doi.org/10.1101/2024.05.03.24306738; this version posted May 5, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

CRLF2, and DSG1, harboring genes crucial for EoE pathogenesis. Furthermore, research has explored the correlation between DOCK8 mutations, Hyper-IgE syndrome, and EoE, underscoring the intricate genetic underpinnings of EoE and its potential overlap with other immune-related conditions.[16] The concordance of EoE is influenced not only by genetic factors but also by environmental variables, with studies demonstrating the impact of factors such as food and allergen exposure on EoE susceptibility and symptom severity.

The investigation into the prevalence of eosinophilic esophagitis (EoE) among patients exhibiting symptoms of gastroesophageal reflux has highlighted a notable incidence of EoE within this patient population. This underscores the imperative to consider EoE during the diagnostic assessment of individuals with gastroesophageal reflux symptoms, especially in tertiary care hospitals located in North India. Consecutive patients with suspected gastroesophageal reflux disease (GERD) underwent gastro duodenoscopy and subsequent esophageal biopsies were collected from specified regions, including the upper and lower esophagus, as well as any other visibly abnormal mucosal areas. Comprehensive analysis of demographic and clinical characteristics, endoscopic findings, peripheral blood eosinophil counts, and prior proton-pump inhibitor (PPI) usage was conducted. Additionally, stool examinations were performed to rule out parasitic infections. Diagnosis of EoE was established based on the presence of over 20 mucosal eosinophils per high-power field, with additional staining to exclude Helicobacter pylori. Among 190 consecutive patients screened for GERD symptoms, esophageal biopsies from 185 individuals were available for assessment. Among these cases, six were confirmed to have EoE, indicating a prevalence of 3.2% among GERD patients in North India [9]. Univariate analysis identified a history of allergy, lack of response to PPIs, and absolute eosinophil counts as significant predictors of EoE, with multivariable analysis corroborating a history of allergy and poor response to PPIs as significant indicators. Notably, the presence of EoE did not show a correlation with the severity of reflux symptoms.

Previous study delves into the utilization of artificial intelligence (AI) to enhance the diagnosis, treatment, and prevention of infectious diseases, with a focus on complex molecular data analysis. Notably, computer-aided detection (CAD) employing convolutional neural networks (CNN) is highlighted. Various machine learning models such as artificial neural networks (ANN), recurrent neural networks (RNN), support vector machines (SVM), and random forests (RF) are examined. AI shows promise in enhancing accuracy and efficiency in managing infectious diseases, although challenges and limitations are acknowledged, emphasizing the need for further research [12]. AI's application in computational biology facilitates early disease detection via image processing and prediction of host-pathogen interactions using genetic and molecular data.

In addition to traditional research methodologies, recent studies have embraced collective meta-heuristic approaches for identifying disease critical genes, as demonstrated in the application to preeclampsia. Furthermore, machine learning and bioinformatics techniques have been employed to predict diagnostic biomarkers associated with immune infiltration in Crohn's disease[10,13], offering promising avenues for understanding disease pathogenesis and improving diagnostic accuracy .In parallel, machine learning approaches have gained traction in EoE research, facilitating the analysis of complex datasets found in public repositories[20] and uncovering novel insights into disease mechanisms and treatment strategies[13]. By integrating genetic, environmental, and mechanistic findings, researchers aim to enhance our understanding of EoE and pave the way for more tailored and effective therapeutic interventions.

medRxiv preprint doi: https://doi.org/10.1101/2024.05.03.24306738; this version posted May 5, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Simulated Annealing (SA) stands as a stalwart in the realm of optimization algorithms, drawing inspiration from the metallurgical annealing process to navigate complex landscapes in search of global maxima. Its journey begins amidst the fervor of high "temperatures," allowing for a broad exploration of the solution space. As the metaphorical temperature gradually cools, SA refines its focus, honing in on the elusive global maximum. Its adaptability knows few bounds, deftly handling nonlinear models, navigating noisy data, and accommodating diverse constraints with aplomb. Yet, SA is not without its demands; achieving optimal results necessitates meticulous parameter tuning and judicious constraint handling. Simulated Annealing (SA) is a stochastic optimization technique inspired by metallurgical annealing. It gradually reduces a "temperature" parameter to explore the solution space of nonlinear problems, effectively escaping local minima and searching for global optima. SA's implementation involves defining solutions, introducing random alterations, evaluating problem functions, and setting an annealing schedule. This method offers a robust optimization approach, especially for complex problems [17].

In stark contrast, the Iterated Hill Climbing Search melds the randomness of a random search with the precision of a gradient search. Armed with an initial cohort of randomly selected solutions, it allocates increasing trials to regions exhibiting promising fitness. However, this method falters when faced with the challenge of locating a global maximum nestled within a diminutive oasis amidst vast stretches of low-fitness terrain. Though simplistic in nature, the Iterated Hill Climbing Search grapples with the intricacies of optimizing functions within convoluted landscapes. SA's reputation precedes it, lauded for its prowess in approaching global optimality and navigating the treacherous waters of nonlinear and stochastic systems. Its versatility and robustness outshine its peers in the domain of local search methods. Nonetheless, the efficacy of SA hinges upon the precision of numerical parameters within its implementation. Tuning SA to suit the idiosyncrasies of varied problems amplifies its efficacy, albeit at the cost of time and effort required to decipher the algorithm's nuances. Survey papers present SAGA, a hybrid approach combining Simulated Annealing (SA) and Genetic Algorithm (GA) for feature selection in high-dimensional microarray datasets. SAGA effectively explores and exploits the solution space, offering superior performance compared to existing algorithms. By integrating SA and GA, SAGA provides a balanced trade-off between exploration and exploitation, addressing challenges of high dimensionality and redundant features for improved classification ability [18].

In summation, SA emerges as a titan among optimization algorithms, wielding its prowess across a myriad of applications, particularly those embroiled in the complexities of nonlinear systems. Its deft balance between exploration and exploitation renders it an invaluable ally in the pursuit of global optima. Conversely, while the Iterated Hill Climbing Search offers simplicity, it struggles with complex optimization tasks, highlighting SA's superiority in addressing multifaceted challenges. Through the lens of SAGA, SA finds new heights, seamlessly integrating with Genetic Algorithms to tackle high-dimensional datasets with finesse and efficacy. In summary, the current state of art in EoE research encompasses a comprehensive exploration of genetic, environmental, and mechanistic aspects, alongside the application of advanced computational methods such as machine learning. By unravelling the complexities of EoE pathogenesis and treatment response, researchers strive to advance personalized medicine approaches and improve outcomes for individuals affected by this challenging condition.

In this paper, we embark on a comprehensive investigation aimed at unraveling the intricate molecular signatures underlying Eosinophilic Esophagus (EoE) and gastrointestinal diseases. Our primary objective is to identify potential biomarkers and elucidate key molecular pathways associated with disease

medRxiv preprint doi: https://doi.org/10.1101/2024.05.03.24306738; this version posted May 5, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

pathology. To achieve this, we employ two distinct yet complementary methodologies: the Ranked Variance (RV) method and Differential Gene Expression Based Simulated Annealing (DGESA). The RV method enables us to assess gene expression variability across samples, while DGESA facilitates the identification of gene sets optimized to maximize the discriminatory power between disease states. Through meticulous data analysis and validation procedures, we seek to provide novel insights into the molecular mechanisms driving these complex conditions. Our study aims to contribute to a deeper understanding of disease pathogenesis and to lay the groundwork for improved diagnostic and therapeutic strategies in gastroenterology. By leveraging innovative methodologies and conducting rigorous analyses, we aspire to identify robust biomarkers and elucidate molecular signatures with the potential to transform patient care in the field of gastroenterology.

Methodology:

The method section of this study details two distinct approaches employed for gene selection and analysis: the Ranked Variance (RV) method and Differential Gene Expression Based Simulated Annealing (DGESA). The RV method prioritizes genes based on their variance, providing an initial perspective on potential biomarkers. In contrast, DGESA utilizes simulated annealing to identify sets of genes exhibiting significant differences in expression between diseased and normal states, facilitating the discovery of disease-associated genetic signatures. Each method offers unique insights into gene selection and contributes to our understanding of molecular mechanisms underlying disease pathogenesis (see Figure 1).

Prior to analysis, several preprocessing steps were implemented to ensure data quality and compatibility. Firstly, rows lacking valid gene names were removed to maintain consistency across datasets. Subsequently, a normalization procedure was applied to each dataset, wherein gene expression values ($e_{i,j}$) were mapped to the range [0, 1]. This normalization step helped mitigate potential biases arising from variations in gene expression magnitude across samples. Finally, transposition of the datasets was performed to prepare the data matrix (denoted as X) for subsequent processing, facilitating the application of required analytical techniques. These preprocessing steps collectively ensured that the gene expression data were standardized and conducive to meaningful analysis and interpretation.

The Ranked Variance (RV) method employed in this study focused on leveraging gene expression variability as a means of discerning potential biomarkers associated with disease states. By computationally analyzing the variance of gene expression across samples, the RV method identified genes exhibiting significant variations in expression levels. This approach facilitated the separation of disease-associated genes from those with relatively stable expression patterns, thereby providing valuable insights into the molecular mechanisms underlying disease pathogenesis. Moreover, the identification of genes with pronounced expression variations enabled subsequent association studies, wherein these genes could be further investigated for their roles in disease development, progression, and potential therapeutic targeting. Overall, the RV method served as a powerful tool for elucidating the genetic signatures associated with various diseases, contributing to our understanding of their underlying biological processes and aiding in the discovery of novel biomarkers.

Differential Gene Expression Based Simulated Annealing (DGESA) is a methodology devised to address gene selection challenges in the context of biological diseases. At its core, DGESA operates on a transposed gene expression matrix (denoted as X), where each row represents a gene and each column corresponds to a sample. The method begins by defining a candidate solution, represented as a set of gene

medRxiv preprint doi: https://doi.org/10.1101/2024.05.03.24306738; this version posted May 5, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

indices (s), which is iteratively refined through a simulated annealing process. During each iteration, a perturbation is applied to the current solution by randomly altering a gene index from the gene expression matrix X that is not already present in the solution set s.

The fitness of each candidate solution is evaluated using a bespoke fitness function, represented by Equation (1):

$$\left|\sum_{i=1}^{g}\left(\overline{D(e_i)}\right)-\left(\overline{Norm(e_i)}\right)\right| \qquad (1)$$

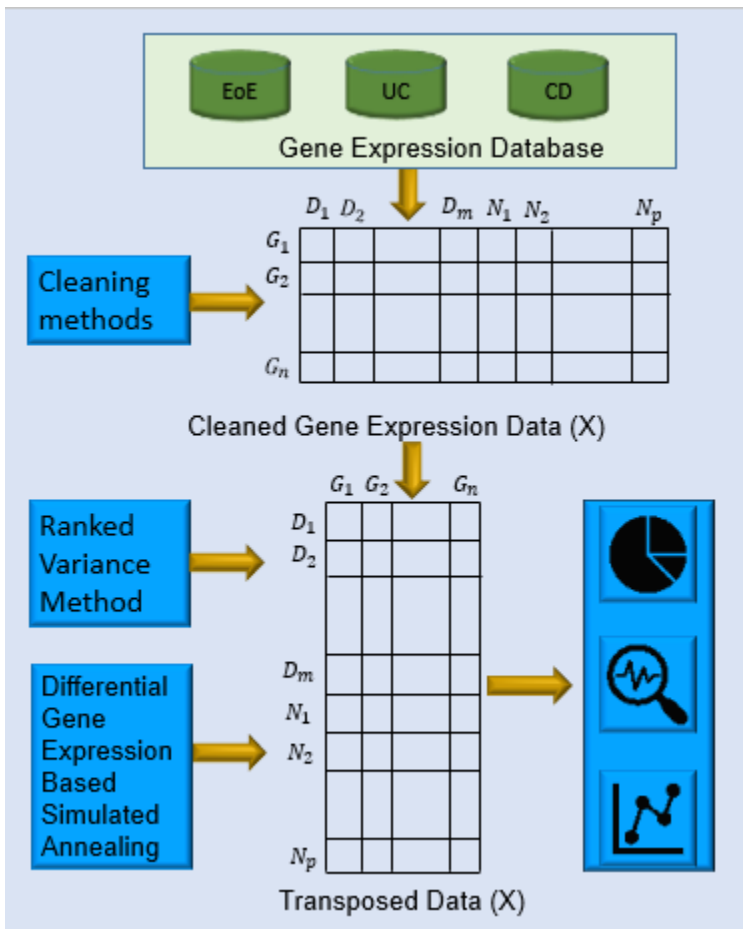


Figure 1. Solution Framework. This figure describes the overall structure of methodologies applied.

Here, g represents the number of genes in the candidate solution s. For each gene index i in s, the expression profile (e_i) is considered. The mean expression profile of the i-th gene in the candidate solution s for diseased samples is denoted by $\left(\overline{D(e_i)}\right)$, while the mean expression profile for normal samples is denoted by $\left(\overline{Norm(e_i)}\right)$. The fitness function computes the absolute difference between the mean expression profiles of diseased and normal samples across all genes in the candidate solution. This difference serves as a measure of the discriminative power of the selected genes in distinguishing between diseased and normal states.

medRxiv preprint doi: https://doi.org/10.1101/2024.05.03.24306738; this version posted May 5, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Through this iterative optimization process, DGESA aims to identify a set of genes that collectively exhibit significant differences in expression patterns between diseased and normal samples. The output of DGESA, denoted as $s*s*$, represents the final selection of genes optimized to maximize the discriminatory power between disease conditions, thereby facilitating the identification of potential biomarkers and elucidation of disease mechanisms (see Figure 2).

DGESA

```
1. Initialize temperature T,
   Iterations P and Cooling rate C;
2. Choose initial candidate solution
   s with random gene index;
3. Calculate fitness f(s)using
   Eq.(1);
4. Repeat
     A. for i = 1 to P do
        ○Randomly select s' ∈ N(s);
         // N(s): Neighbor of s
        ○if f(s') ≥ f(s) then
            s ← s';
        ○else
            s ← s' with
            probability e^(-(f(s')-f(s))/T);
        ○end if
     B. end for
     C. T = T × C;
5. Until stopping criteria not met
6. end
```

Figure 2: Algorithm DGESA

The methods applied in this study, including the Ranked Variance (RV) method and Differential Gene Expression Based Simulated Annealing (DGESA), have provided valuable insights into gene selection and analysis within the context of biological diseases. The RV method effectively identified genes with significant expression variations, aiding in disease gene separation and association studies. On the other hand, DGESA leveraged simulated annealing to pinpoint genes exhibiting differential expression patterns between diseased and normal samples, thereby contributing to the discovery of disease-associated genetic signatures. By employing these complementary methodologies, we have advanced our understanding of molecular mechanisms underlying disease pathogenesis and provided a robust framework for biomarker discovery and disease classification.

medRxiv preprint doi: https://doi.org/10.1101/2024.05.03.24306738; this version posted May 5, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Results:

The data collection for this study involved retrieving two gene expression datasets from the Gene Expression Omnibus (GEO) repository hosted by the National Center for Biotechnology Information (NCBI). The first dataset, GSE228083, comprised samples from patients with Eosinophilic Esophagus (EoE) compared to normal samples, facilitating the investigation of gene expression patterns specific to this condition. The second dataset, GSE24287, encompassed gene expression profiles from patients with Ulcerative Colitis (UC), Crohn's Disease (CD), and normal samples. From GSE24287, two distinct datasets were prepared by segregating samples into UC vs. Normal and CD vs. Normal categories.

Hyper-parameter tuning of Differential Gene Expression Based Simulated Annealing (DGESA) is a critical aspect of optimizing its performance in gene selection tasks. This iterative process involves systematically adjusting parameters that control the learning process, known as hyper-parameters, through experimentation with different configurations. In the case of DGESA, key hyper-parameters include the number of genes (g) in the candidate solution, the maximum number of iterations, the initial temperature (T), and the cooling rate. By systematically adjusting these hyper-parameters, the DGESA model can be fine-tuned to enhance its efficiency in identifying disease-associated genes. In this study, after thorough experimentation and analysis of resulting performance metrics, the final hyper-parameter configurations were determined as follows: g = 40 genes in the candidate solution, 200,000 iterations, T = 10^6, and a cooling rate of 0.9. These optimized hyper-parameters ensure the effectiveness of DGESA in identifying relevant genetic signatures associated with biological diseases, thereby advancing our understanding of disease mechanisms and aiding in biomarker discovery.

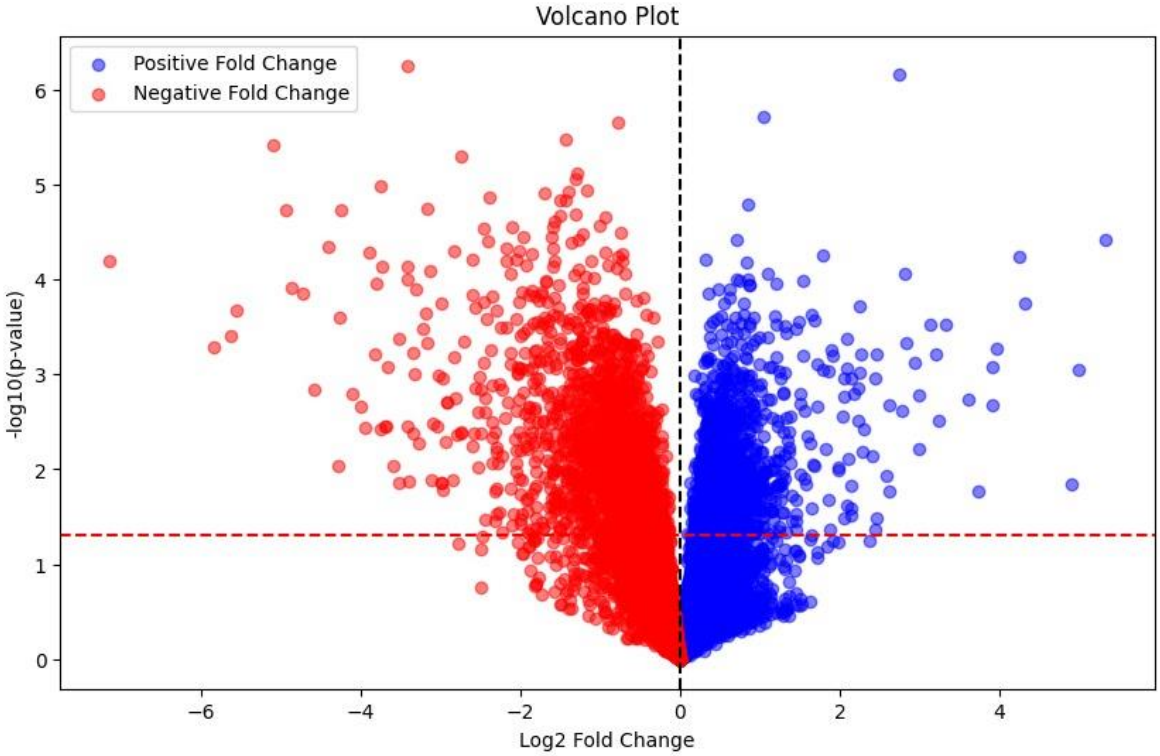


Figure 3: Volcano plot of EoE dataset

medRxiv preprint doi: https://doi.org/10.1101/2024.05.03.24306738; this version posted May 5, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

To gain insight into the differential expression patterns of genes in the EoE dataset, a volcano plot was generated, depicting the relationship between the log2 fold change and the log10 p-values of various genes. In this plot, the x-axis represents the log2 fold change, which quantifies the magnitude of gene expression differences between EoE samples and normal samples. Meanwhile, the y-axis displays the -log10 p-values, which serve as a measure of the statistical significance of these expression differences. The volcano plot (see Figure 3) revealed that the majority of genes exhibited negative fold changes, indicating underexpression in EoE compared to normal samples. This observation suggests a potential downregulation of gene expression associated with EoE pathology. However, it's essential to interpret these findings in conjunction with additional analyses to elucidate the specific genes and biological pathways underlying the disease's pathogenesis and progression.

The application of the Ranked Variance (RV) method to the EoE vs. normal dataset yielded insightful results regarding the variability of gene expression across samples. By plotting the curve (see Figure 4) where the x-axis represents the gene index and the y-axis denotes the corresponding variance, it was observed that approximately 40 genes exhibited decreasing variance. This observation suggests a notable reduction in the variability of expression levels for these genes in EoE samples compared to normal samples. Such a trend of decreasing variance may indicate a degree of regulatory homogeneity or consistent downregulation of gene expression within this subset of genes in the context of EoE pathology. These findings highlight the potential significance of these genes in contributing to the molecular mechanisms underlying EoE development and progression, warranting further investigation to elucidate their functional roles and potential implications as biomarkers or therapeutic targets. The variance graph of CD and UC are available on Supplement 1 and 2 respectively.

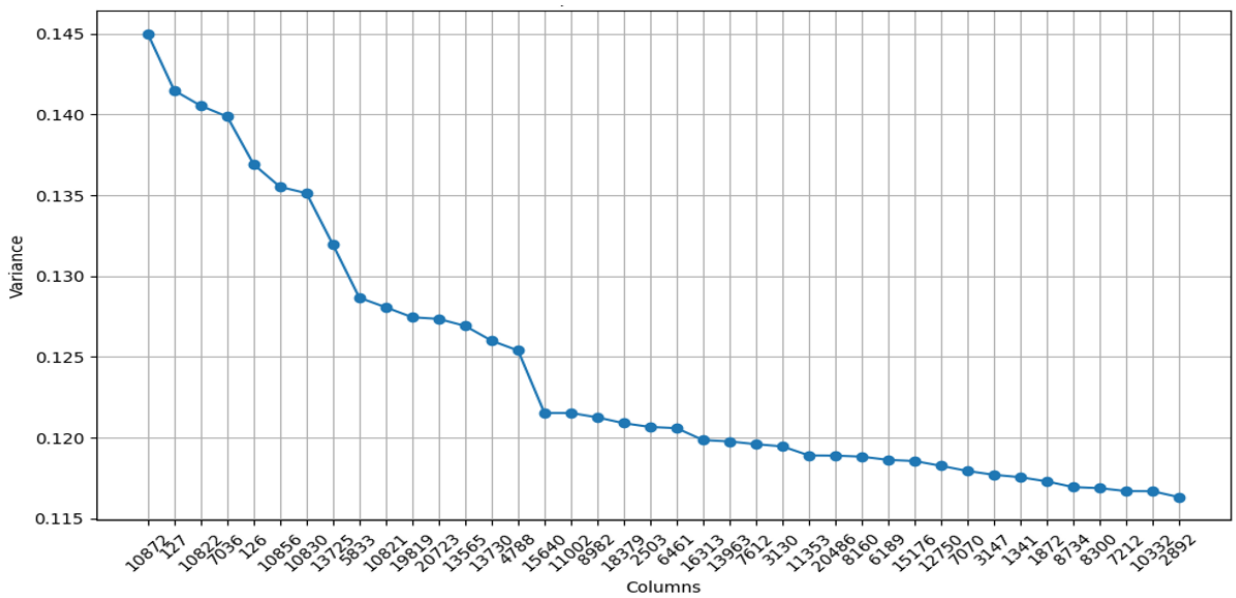


Figure 4: Variance of the first 40 genes of highest variability in EoE dataset

The application of the DGESA method to the EoE vs. normal dataset yielded a convergence curve that provides valuable insights into the optimization process. In this curve, the x-axis represents the iterations, reflecting the number of iterations or steps taken during the simulated annealing optimization procedure. Meanwhile, the y-axis denotes the corresponding fitness values, which quantify the effectiveness

medRxiv preprint doi: https://doi.org/10.1101/2024.05.03.24306738; this version posted May 5, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

of the candidate solutions at each iteration. The observed convergence of the curve indicates that as the optimization progresses through iterations, the fitness values gradually stabilize or improve, eventually reaching an optimal or near-optimal solution. This convergence phenomenon signifies the effectiveness of DGESA in iteratively refining the selection of genes to maximize their discriminative power between EoE and normal samples. The convergence curve underscores the robustness and efficiency of DGESA in (see Figure 5) identifying disease-associated genetic signatures and highlights its potential as a valuable tool for biomarker discovery and disease classification in biomedical research. The optimization curve of CD and UC are available on Supplement 3 and 4 respectively.

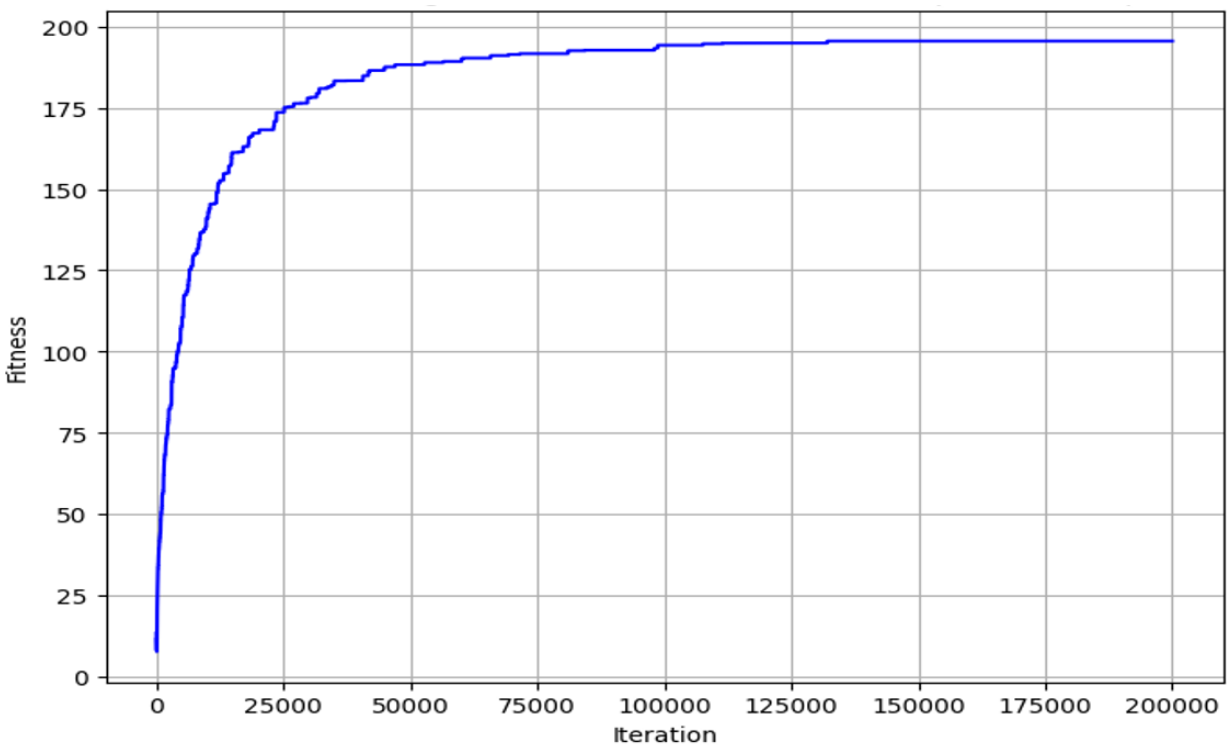


Figure 5: Optimization curve of DGESA applied on EoE vs. Normal dataset

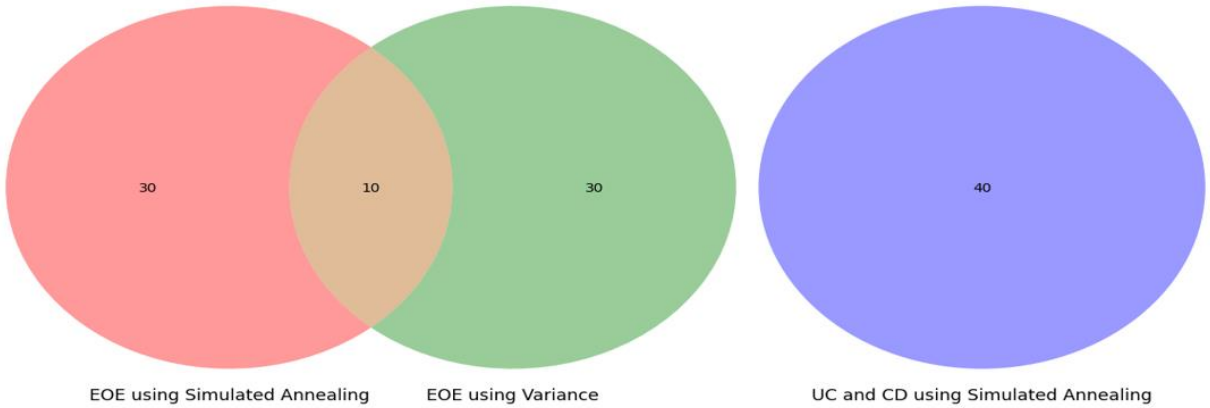


Figure 6: Venn diagram represent shared genes between methods (DGESA and RV) and Diseases (EoE and UC-CD)

medRxiv preprint doi: https://doi.org/10.1101/2024.05.03.24306738; this version posted May 5, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

The Venn diagram analysis comparing the gene sets identified by the DGESA and RV methods in the context of Eosinophilic Esophagus (EoE) versus normal samples revealed important findings. (see Figure 6). Specifically, the diagram indicated that 10 genes were shared between DGESA and RV, suggesting a degree of consistency or agreement between the two methods in identifying potential biomarkers associated with EoE. However, notably, no overlap was observed between the gene sets identified for EoE and the combined UC and CD versus normal dataset. This absence of connection between the gene sets for EoE and UC-CD may reflect distinct molecular mechanisms underlying these two gastrointestinal diseases. The lack of shared genes highlights the specificity of gene expression profiles associated with each disease entity and underscores the importance of tailored approaches for biomarker discovery and therapeutic targeting. Further investigation into the unique genetic signatures of EoE and UC-CD could offer deeper insights into their pathogenesis and facilitate the development of more precise diagnostic and therapeutic strategies.

The final gene set identified by the DGESA algorithm for Eosinophilic Esophagus EoE presents a compelling alignment with previous literature, underscoring its potential significance in the context of EoE pathology. Among the 13 genes listed, several, including DPCR1, SPRR2E, SPRR2B, SPRR2D, KRT79, RORC, CRISP2, IL36G, and CCND1, have been previously implicated in EoE and are highlighted in blue to denote their strong confirmation of association with the disease. These genes represent key players in various molecular pathways relevant to EoE, such as immune regulation, epithelial barrier function, and tissue remodeling. Additionally, the presence of other genes in the final gene set, although not explicitly highlighted, suggests potential connections to EoE based on their co-appearance with established EoE-associated genes. This comprehensive gene set derived from DGESA not only validates known associations but also offers new insights into the molecular mechanisms underlying EoE pathogenesis, paving the way for further research into diagnostic and therapeutic interventions for this complex disease. The unique genes of CD and UC by DGESA are available on Supplement 5 and 6 respectively.

Table 1: Final unique genes of EoE using DGESA

HSPA12A	DPCR1	FAM25G	FAM43B	IVD	PPP2R1B	MTHFD2L	SPRR2E	GGT6	KRT79
RORC	CRISP2	ZNF562	OAZ3	C18orf54	EXOC3	IL36G	TPPP2	ANXA8	CSN2
RECQL	RPAP3	SPINK13	TAF4B	LYPD6	COX6B1	CPB1	SPRR2B	SPRR2D	DMKN
FAM217B	HIP1	ARC	ZFAND4	CCND1	RMI1	LOC388780	CSNK1A1L	ADGRB1	STAT6

Conclusion

The discussion of results is a critical component of our study, as it provides an opportunity to interpret and contextualize the findings obtained from the application of the Ranked Variance (RV) method and Differential Gene Expression Based Simulated Annealing (DGESA) in the context of Eosinophilic Esophagus (EoE) and gastrointestinal diseases. Firstly, the observed convergence of genes identified by both DGESA and RV in EoE underscores the consistency and reliability of our approach. The 10 genes showing overlap between the two methods suggest a convergence of results and highlight their potential

medRxiv preprint doi: https://doi.org/10.1101/2024.05.03.24306738; this version posted May 5, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

relevance as biomarkers for EoE. This convergence provides confidence in the robustness of our methodologies and strengthens the validity of the identified gene sets. Additionally, the validation of 13 genes strongly associated with EoE in previous literature further corroborates the significance of our findings. These genes, implicated in various molecular pathways including immune regulation and tissue remodeling, underscore the complex nature of EoE pathogenesis and offer potential targets for future research and therapeutic interventions. However, the lack of overlap between EoE and the combined Ulcerative Colitis (UC) and Crohn's Disease (CD) datasets suggests distinct molecular signatures underlying these gastrointestinal diseases. This observation underscores the importance of tailored approaches to understanding the unique pathophysiology of each disease entity. Further investigation into the molecular mechanisms driving these diseases is warranted to identify disease-specific biomarkers and therapeutic targets. Moreover, the identification of genes with decreasing variance in EoE samples compared to normal controls through the RV method provides valuable insights into the regulation of gene expression in disease states. The reduction in variability suggests potential regulatory homogeneity or consistent downregulation of gene expression within this subset, highlighting their relevance to EoE pathology. Future studies could explore the functional roles of these genes and their implications for disease progression and therapeutic interventions. In conclusion, our study has contributed to a deeper understanding of gene expression patterns and their associations with EoE and gastrointestinal diseases. The convergence of results obtained from DGESA and RV, along with the validation of genes strongly associated with EoE in previous literature, underscores the significance of our findings. Moving forward, further research is needed to elucidate the molecular mechanisms underlying these diseases and identify novel biomarkers and therapeutic targets for improved patient care.

Reference

1. Cristianini, N., & Hahn, M. W. (2006). Introduction to computational genomics: a case studies approach. Cambridge University Press.

2. Derrien, T., Estellé, J., Marco Sola, S., Knowles, D. G., Raineri, E., Guigó, R., & Ribeca, P. (2012). Fast computation and applications of genome mappability. PloS one, 7(1), e30377.

3. Hackl, H., Charoentong, P., Finotello, F., & Trajanoski, Z. (2016). Computational genomics tools for dissecting tumour–immune cell interactions. Nature Reviews Genetics, 17(8), 441-458.

4. Emes, R. D., Pirooznia, M., Zou, Q., & Pellegrini, M. (2023). Insights in computational genomics: 2022. Frontiers in Genetics, 14, 1256011.

5. Clough, E., & Barrett, T. (2016). The gene expression omnibus database. Statistical Genomics: Methods and Protocols, 93-110.

6. https://www.ncbi.nlm.nih.gov/

7. Barmeyer, C., Schulzke, J. D., & Fromm, M. (2015, June). Claudin-related intestinal diseases. In Seminars in cell & developmental biology (Vol. 42, pp. 30-38). Academic Press.

8. Lucas López, R., Grande Burgos, M. J., Gálvez, A., & Pérez Pulido, R. (2017). The human gastrointestinal tract and oral microbiota in inflammatory bowel disease: a state of the science review. Apmis, 125(1), 3-10.

9. Cianferoni, A., & Spergel, J. (2016). Eosinophilic esophagitis: a comprehensive review. Clinical reviews in allergy & immunology, 50, 159-174.

medRxiv preprint doi: https://doi.org/10.1101/2024.05.03.24306738; this version posted May 5, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

10. Dellon, E. S., & Hirano, I. (2018). Epidemiology and natural history of eosinophilic esophagitis. Gastroenterology, 154(2), 319-332.

11. Muir, A., & Falk, G. W. (2021). Eosinophilic esophagitis: a review. Jama, 326(13), 1310-1318.

12. Saito, Y. A., Mitra, N., & Mayer, E. A. (2010). Genetic approaches to functional gastrointestinal disorders. Gastroenterology, 138(4), 1276-1285.

13. Rothenberg, M. E. (2015). Molecular, genetic, and cellular bases for treating eosinophilic esophagitis. Gastroenterology, 148(6), 1143-1157.

14. Kottyan, L. C., & Rothenberg, M. (2017). Genetics of eosinophilic esophagitis. Mucosal immunology, 10(3), 580-588.

15. Kottyan, L. C., Parameswaran, S., Weirauch, M. T., Rothenberg, M. E., & Martin, L. J. (2020). The genetic etiology of eosinophilic esophagitis. Journal of Allergy and Clinical Immunology, 145(1), 9-15.

16. Baruah, B., Kumar, T., Das, P., Thakur, B., Sreenivas, V., Ahuja, V., ... & Makharia, G. K. (2017). Prevalence of eosinophilic esophagitis in patients with gastroesophageal reflux symptoms: A cross-sectional study from a tertiary care hospital in North India. Indian journal of gastroenterology, 36, 353-360.

17. Nagarajan, K. V., Krishnamurthy, A. N., Yelsangikar, A., Mallappa, R. B., Bhat, V., Narasimhamurthy, V. M., & Bhat, N. (2023). Does eosinophilic esophagitis exist in India?. Indian Journal of Gastroenterology, 42(2), 286-291.

18. Cho JH, Lee D, Park JH, Lee IB (2003) New gene selection method for classification of cancer subtypes considering within class variation. FEBS Lett 551(1–3):3–7.

19. Díaz-Uriarte R, Andrés SA (2006) Gene selection and classification of microarray data using random forest. BMC Bioinform.

20. Yu L, Han Y, Berens ME (2012) Stable gene selection from microarray data via sample weighting. IEEE/ACM Trans Comput Biol Bioinform 9(1):262–272

21. Libbrecht, M. W., & Noble, W. S. (2015). Machine learning applications in genetics and genomics. Nature Reviews Genetics, 16(6), 321-332.

22. Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction. Computational and structural biotechnology journal, 13, 8-17.

23. Tabl, A. A., Alkhateeb, A., ElMaraghy, W., Rueda, L., & Ngom, A. (2019). A machine learning approach for identifying gene biomarkers guiding the treatment of breast cancer. Frontiers in genetics, 10, 256.

24. Biswas, S., Dutta, S., & Acharyya, S. (2019). Identification of disease critical genes using collective meta-heuristic approaches: an application to preeclampsia. Interdisciplinary Sciences: Computational Life Sciences, 11, 444-459.

25. Esengönül, M., Marta, A., Beirão, J., Pires, I. M., & Cunha, A. (2022). A systematic review of artificial intelligence applications used for inherited retinal disease management. Medicina, 58(4), 504.

26. Colak, C., Kucukakcali, Z., & Akbulut, S. (2023). Artificial intelligence-based prediction of molecular and genetic markers for hepatitis C–related hepatocellular carcinoma. Annals of Medicine and Surgery, 85(10), 4674-4682.

27. Bao, W., Wang, L., Liu, X., & Li, M. (2023). Predicting diagnostic biomarkers associated with immune infiltration in Crohn's disease based on machine learning and bioinformatics. European Journal of Medical Research, 28(1), 255.

medRxiv preprint doi: https://doi.org/10.1101/2024.05.03.24306738; this version posted May 5, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

28. Ridge, P. G., Hoyt, K. B., Boehme, K., Mukherjee, S., Crane, P. K., Haines, J. L., ... & Reitz, C. (2016). Assessment of the genetic variance of late-onset Alzheimer's disease. Neurobiology of aging, 41, 200-e13.

29. Chen, C. H., Kraemer, B. R., & Mochly-Rosen, D. (2022). ALDH2 variance in disease and populations. Disease Models & Mechanisms, 15(6), dmm049601.

30. Wolf, S., Melo, D., Garske, K. M., Pallares, L. F., Lea, A. J., & Ayroles, J. F. (2023). Characterizing the landscape of gene expression variance in humans. PLoS genetics, 19(7), e1010833.

31. Van Laarhoven, P. J., Aarts, E. H., van Laarhoven, P. J., & Aarts, E. H. (1987). Simulated annealing (pp. 7-15). Springer Netherlands.

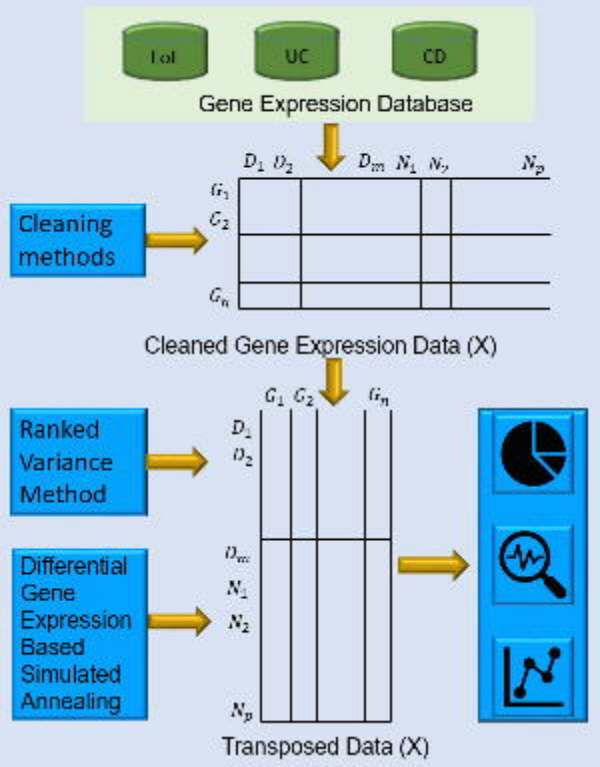
32. Bertsimas, D., & Tsitsiklis, J. (1993). Simulated annealing. Statistical science, 8(1), 10-15.

33. Aarts, E., Korst, J., & Michiels, W. (2005). Simulated annealing. Search methodologies: introductory tutorials in optimization and decision support techniques, 187-210.

34. Guilmeau, T., Chouzenoux, E., & Elvira, V. (2021, July). Simulated annealing: A review and a new scheme. In 2021 IEEE Statistical Signal Processing Workshop (SSP) (pp. 101-105). IEEE.

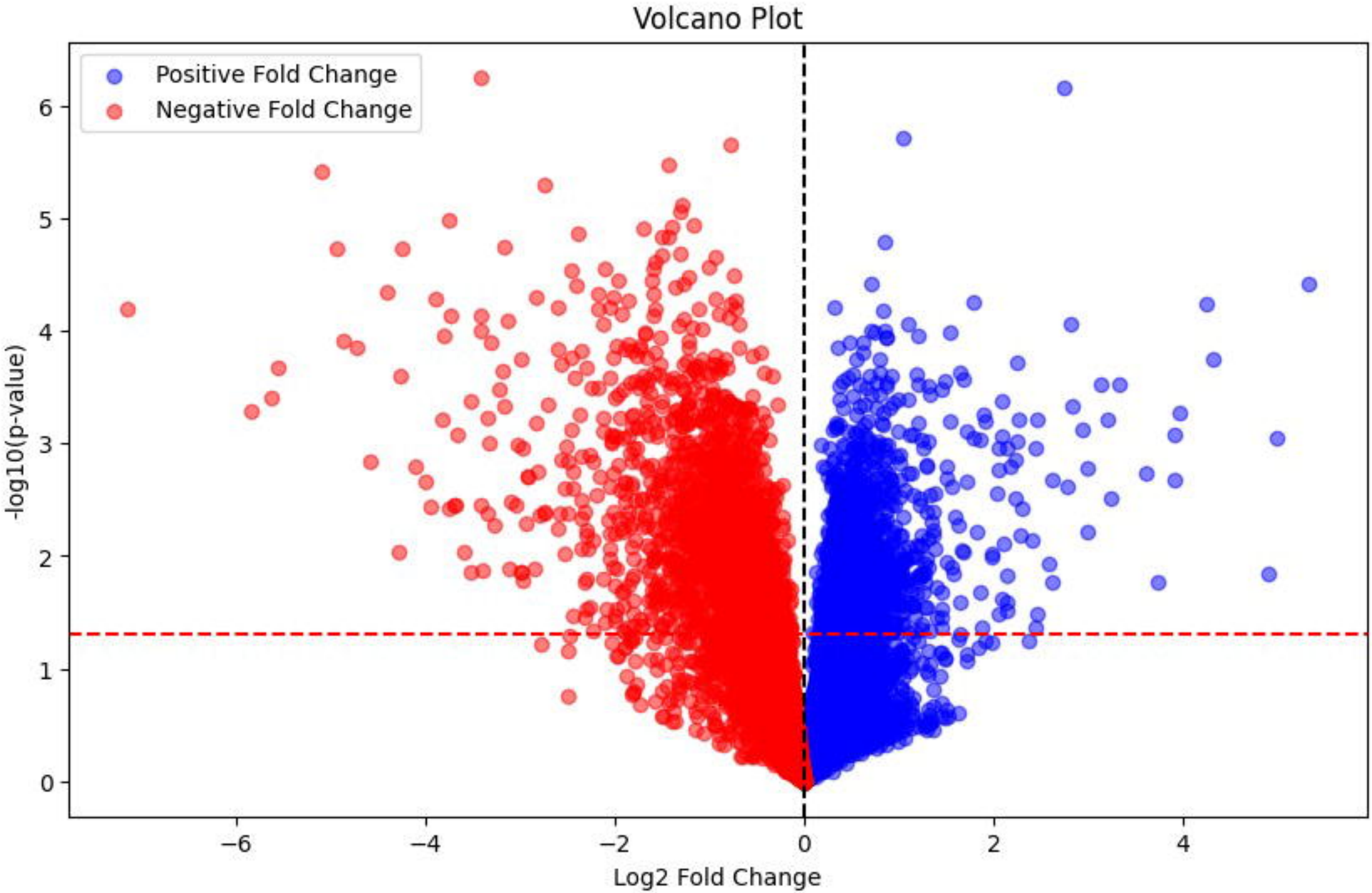
35. Koul, N., & Manvi, S. S. (2022). Feature selection from gene expression data using simulated annealing and partial least squares regression coefficients. Global Transitions Proceedings, 3(1), 251-256.

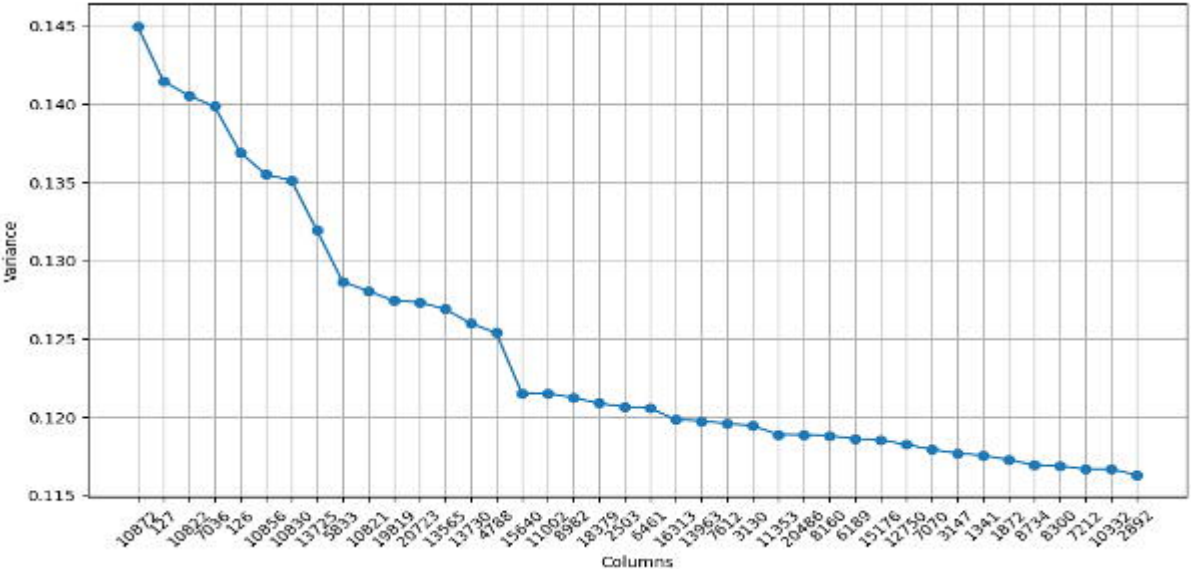
36. Marjit, S., Bhattacharyya, T., Chatterjee, B., & Sarkar, R. (2023). Simulated annealing aided genetic algorithm for gene selection from microarray data. Computers in Biology and Medicine, 158, 106854.

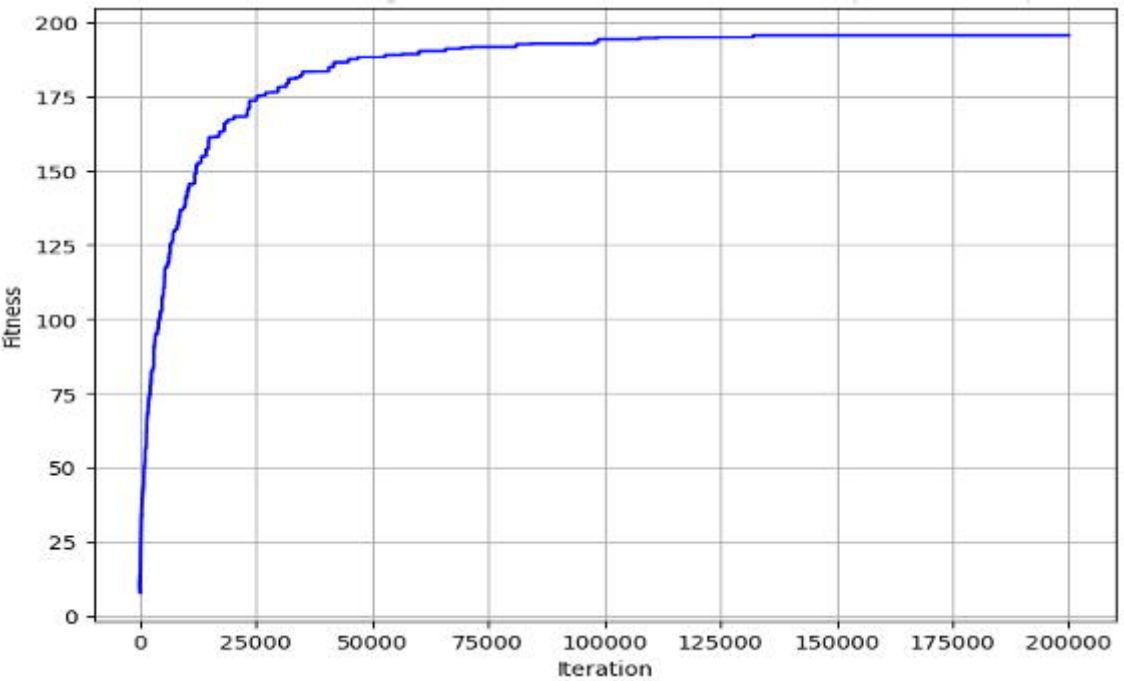

Lol
UC
CD
Gene Expression Database
D_1 D_2 D_m N_1 N_2 N_p
G_1
G_2
G_n
Cleaning methods
Cleaned Gene Expression Data (X)
G_1 G_2 G_n
D_1
D_2
D_m
N_1
N_2
N_p
Ranked Variance Method
Differential Gene Expression Based Simulated Annealing
Transposed Data (X)

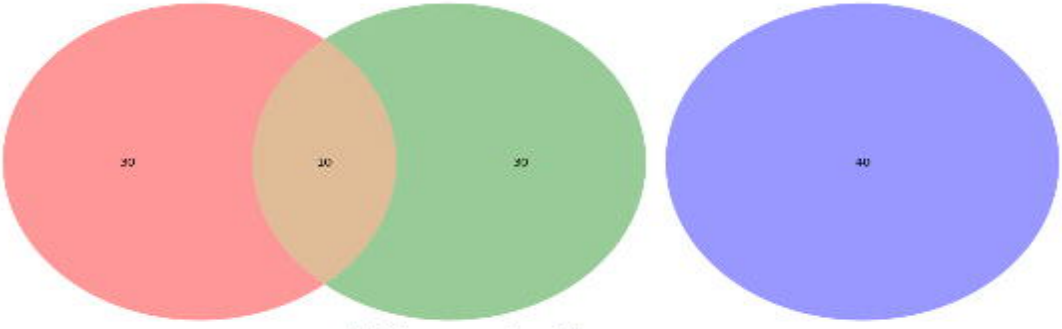
DGESA

```
1. Initialize temperature T,
   Iterations P and Cooling rate C;
2. Choose initial candidate solution
   s with random gene index;
3. Calculate fitness f(s)using
   Eq.(1);
4. Repeat
     A. for i = 1 to P do
        ○Randomly select s' ∈ N(s);
          // N(s): Neighbor of s
        ○if f(s') ≥ f(s) then
            s ← s';
        ○else
            s ← s' with
            probability e^(-(f(s')-f(s))/T);
        ○end if
     B. end for
     C. T = T × C;
5. Until stopping criteria not met
6. end
```


Volcano Plot
Positive Fold Change
Negative Fold Change
-log10(p-value)
Log2 Fold Change
6
5
4
3
2
1
0
−6
−4
−2
0
2
4


0.145
0.140
0.135
0.130
0.125
0.120
0.115
Variance
10872
127
10822
7036
126
10856
10830
13725
5833
10821
19819
20723
13565
13730
4788
15640
11002
8982
18379
2503
6461
16313
13963
7612
3130
11353
20486
8160
6189
15176
12759
7070
3147
1341
1872
8734
8300
7212
10332
2892
Columns


Fitness
Iteration
0
25
50
75
100
125
150
175
200
0
25000
50000
75000
100000
125000
150000
175000
200000


30
10
30
40
Final EoE genes using DGESA
EOE genes using RV
Final genes of UC-CD combined