Original Article: Artificial Intelligence Guided Discovery of Gastric Cancer Continuum

Authors: Daniella Vo[1], Pradipta Ghosh[2-4†] and Debashis Sahoo [1,2,5†]

Affiliations:

[1]Department of Pediatrics, University of California San Diego.
[2]Moores Cancer Center, University of California San Diego.
[3]Department of Cellular and Molecular Medicine, University of California San Diego.
[4]Department of Medicine, University of California San Diego.
[5]Department of Computer Science and Engineering, Jacob's School of Engineering, University of California San Diego.

[†]**Equal contribution:**

Debashis Sahoo, Ph.D.; Associate Professor, Department of Pediatrics, University of California San Diego; 9500 Gilman Drive, MC 0703, Leichtag Building 132; La Jolla, CA 92093-0703. **Phone:** 858-246-1803: **Fax:** 858-246-0019: **Email:** dsahoo@ucsd.edu

Running Head: AI Guided Discovery of GC Continuum

Word Count: 3163 words

ABSTRACT

Background: Detailed understanding of pre, early and late neoplastic states in gastric cancer helps develop better models of risk of progression to Gastric Cancers (GCs) and medical treatment to intercept such progression.

Methods: We built a Boolean Implication network of gastric cancer and deployed machine learning algorithms to develop predictive models of known pre-neoplastic states, e.g., atrophic gastritis, intestinal metaplasia (IM) and low- to high-grade intestinal neoplasia (L/HGIN), and GC. Our approach exploits the presence of asymmetric Boolean Implication relationships that are likely to be invariant across almost all gastric cancer datasets. Invariant asymmetric Boolean Implication relationships can decipher fundamental time series underlying the biological data. Pursuing this method, we developed a healthy mucosa→GC continuum model based on this approach.

Results: Our model performed better against publicly available models for distinguishing healthy versus GC samples. Although not trained on IM and L/HGIN datasets, the model could identify the risk of progression to GC via the metaplasia→dysplasia→neoplasia cascade in patient samples. The model could rank all publicly available mouse models for their ability to best recapitulate the gene expression patterns during human GC initiation and progression.

Conclusions: A Boolean implication network enabled the identification of hitherto undefined continuum states during GC initiation. The developed model could now serve as a starting point for rationalizing candidate therapeutic targets to intercept GC progression.

MINI-ABSTRACT

We developed predictive models of early and late neoplastic states in gastric cancer and identified gene clusters that are up/down-regulated at various points along the gastric cancer disease continuum.

KEYWORDS

Stomach neoplasms, computational biology, systems biology, transcriptome, machine learning

INTRODUCTION

Gastric cancer (GC) often presents as an advanced disease with patients either having inoperable conditions or surgery as the only potentially curative treatment [1]. There is evidence that 75% of all GCs are initiated by *Helicobacter pylori,* a known carcinogenic pathogen [2, 3]. Risk factors also include age, sex, smoking and family history [4]. This oncogenesis leads to Correa's cascade, a stepwise progression from normal, chronic active gastritis, atrophic gastritis, intestinal metaplasia, dysplasia then adenocarcinomas [3]. Intestinal metaplasia also has two subtypes, incomplete and complete intestinal metaplasia (IIM and CIM, respectively), with IIM having a higher probability of developing GC compared to CIM [5].

Research into GCs has used impactful approaches to investigate the genome [6], therapeutics [7] and survival [8], but these methods have not translated into actionable biomarkers of prognostication, targets, novel therapeutics, or changes in screening strategies. These genomic insights also have not provided insight into which genes are important in the progression of GC for preneoplastic detection and treatment.

Here we present a network-based approach for biomarker and target discovery that uses artificial intelligence (AI) to select genes and then perform rigorous validation in multiple independent GC datasets. Previously, we have successfully exploited this approach to identify biomarkers in IBD [9], COVID-19 [10] and macrophages [11]. We demonstrate how Boolean implications allow us to develop models that provide insight into the gastric cancer disease continuum.

METHODS

Detailed methods for computational modeling and AI-guided target identification are presented in Online Resource 1 and mentioned in brief here.

Construction of a Network of Boolean Implications

Modeling continuum states within the metaplasia → dysplasia → neoplasia cascade was performed using Boolean Network Explorer (BoNE) [9]. We created an asymmetric gene expression network, for the progression from normal to gastric cancer (GC), using a computational method based on Boolean logic [12]. To build the GC network, we analyzed a publicly available gastric cancer transcriptomic dataset, GSE66229[13] (n = 400; 300 GC tumor and 100 patient-matched normal tissue). A Boolean Network Explorer (BoNE; see **Online Resource 1** for more details) computational tool was introduced, which uses asymmetric properties of Boolean implication relationships (BIRs as in MiDReG algorithm [12]) to model natural progressive time-series changes in major cellular compartments that initiate, propagate, and perpetuate cellular state change and are likely to be important for GC progression. BoNE provides an integrated platform for the construction, visualization and querying of a network of progressive changes much like a disease map (in this case, GC map) in three steps: First, the expression levels of all genes in these datasets were converted to binary values (high or low) using the StepMiner algorithm [14]. Second, gene expression relationships between pairs of genes were classified into one-of-six possible BIRs and expressed as Boolean implication statements; two symmetric Boolean implications “equivalent” and “opposite” are discovered when two diagonally opposite sparse quadrants are identified and four asymmetric relationships, each corresponding to one sparse quadrant. While conventional symmetric analysis of transcriptomic datasets can recognize the latter 2 relationships, such an approach ignores the former. BooleanNet statistics are used to assess the significance of the Boolean implication relationships [12]. Prior work [9] has revealed how the Boolean approach offers a distinct advantage from currently used conventional computational methods that rely exclusively on symmetric linear relationships from gene expression data, e.g., differential, correlation-network, coexpression-network, mutual information-network, and the Bayesian approach. The other advantage of using BIRs is that they are robust

to the noise of sample heterogeneity (i.e., healthy, diseased, genotypic, phenotypic, ethnic, interventions, disease severity) and every sample follows the same mathematical equation, and hence is likely to be reproducible in independent validation datasets. Third, genes with similar expression architectures, determined by sharing at least half of the equivalences among gene pairs, were grouped into clusters, and organized into a network by determining the overwhelming Boolean relationships observed between any two clusters. In the resultant Boolean implication network, clusters of genes are the nodes, and the BIR between the clusters are the directed edges; BoNE enables their discovery in an unsupervised way while remaining agnostic to the sample type. All gene expression datasets were visualized using Hierarchical Exploration of Gene Expression Microarrays Online (HEGEMON) framework [9].

Ordering samples based on composite score of Boolean path

A Boolean path contains one or more clusters. A composite score is computed for each cluster and combined later. To compute the final score, first the genes present in each cluster were normalized and averaged. Gene expression values were normalized according to a modified Z-score approach centered around StepMiner threshold (formula = (expr - SThr)/3/stddev). A weighted linear combination of the averages from the clusters of a Boolean path was used to create a score for each sample. The weights along the path either monotonically increased or decreased to make the sample order consistent with the logical order based on BIR. The samples were ordered based on the final weighted and linearly combined score. A cluster highly expressed in a disease setting received a positive weight (ex: 1, 2, 3, etc.) and healthy setting received a negative weight (ex: -1, -2, -3, etc.).

Multivariate Analysis for Model Selection

Two microarray datasets (GSE37023 (only samples on GPL96 Affymetrix Human Genome U133A Array used for analysis), n = 65, non-malignant = 36, GC tumor = 29; GSE122401, n = 160, patient-matched normal = 80, GC tumor = 80) are used to train a network model to distinguish normal vs GC samples. Using Ordinary Least Squares (OLS) regression in Python statsmodels (version 0.12.2), we performed multivariate analysis to determine which models performed best in the two training datasets.

Statistical Analysis

Statistical significance between experimental groups was determined using Python scipy.stats.ttest_ind package (version 0.19.0) with Welch's Two Sample t-test (two-tailed, unpaired, unequal variance (equal_var=False), and unequal sample size). For all tests, a p-value of 0.05 was used as the cutoff to determine significance. Violin and bar plots are created using Python seaborn package version 0.10.1.

RESULTS

Machine learning identified two possible Boolean paths in the GC disease map

Using a publicly available GC dataset (GSE66229) with tumor (T) and adjacent normal (AN) samples, we built a Boolean implication network (See *Methods* and **Online Resource 1***;* **Fig. 1a**). Each cluster was evaluated to determine whether they fall on the healthy versus GC side of the disease map based on whether the average gene expression value of a cluster in healthy samples is up or down, yielding a GC map (**Fig. 1b**). We then used machine learning to identify Boolean paths (clusters connected by Boolean implication relationships) in the GC map that can distinguish tumor from AN samples in the training datasets (Figure 1C *top graphic*). Clusters #11-2-4-14 (C#11-2-4-14) performed the best with an ROC-AUC of 0.96 in training dataset #1 (GSE37023 AN versus T), while clusters #7-13-14 (C#7-13-14) performed best in training dataset #2

(GSE122401 AN vs T) with an ROC-AUC of 0.98 (**Fig. 1c**). Specific violin plots for both datasets and Boolean paths are presented in **Fig. 1d**. We performed Reactome pathway analysis on clusters in both paths to identify the top five biological processes associated with the clusters (**Fig. 1e**). Cluster 11 involves the downregulation of genes related to muscle contraction in GC. Cluster 2 represents genes relevant to cell cycle as many other studies pointed out their relevance in the context of GC [15, 16]. Cluster 4 had genes from the immune system including neutrophil degranulation as linked in other papers [17, 18]. Clusters 7 and 13 had genes involved in the downregulation of ion channel transport in GC [19, 20]. Cluster 14 represents genes increased in extracellular matrix processes [21, 22]. Since both Boolean paths C#11-2-4-14 and C#7-13-14 can distinguish AN versus GC samples, we identified a gene signature called GC-BoNE uses the path that best characterized the different samples (highest ROC-AUC score out of both paths) for classification of samples.

We tested how well the clusters identified by our Boolean approach would compare to previously established gene signatures (**Fig. 2a**). C#11-2-4-14 and C#7-13-14 individually (**Fig. 2b**) could classify the tumor and normal/adjacent normal samples in the 21 validation datasets (see **Online Resource 2** for a list of GSE IDs; ROC-AUC ranges from 0.57 - 1.00 in C#11-2-4-14, and 0.66 - 1.00 in C#7-13-14). We then compared GC-BoNE to other gene signatures (see **Online Resource 3** for list of genes in signatures; **Fig. 2c**) and found that our signature outperformed the others (average ROC-AUC for GC-BoNE is 0.933, and other signatures range from 0.690 - 0.921). There were minimal overlaps between clusters 11-2-4 (**Fig. 2d**), 7-13 (**Fig. 2e**) and the top three signatures (DEA (Li 2015), DEA+PPIN and Japanese GC). Cluster 14 and the Japanese GC signature had 8 overlapping genes (**Fig. 2f**). These findings suggest GC-BoNE provides a new list of potential biomarkers for GC that differ from previous signatures.

GC-BoNE identifies progressively increasing risk of GC along the metaplasia-dysplasia continuum

We next asked if the GC-BoNE signature is induced during the progression from normal to GC through the normal→inflammation (gastritis)→metaplasia→dysplasia→neoplasia cascade. In one dataset (E-MTAB-8889), we looked at the normal→inflammation (gastritis)→metaplasia cascade by comparing pairwise each sequential step, i.e., non-atrophic gastritis (NAG) vs chronic active gastritis (CG), CG vs chronic atrophic gastritis (CAG) and CAG vs intestinal metaplasia (IM) (**Fig. 3a**). We also looked at the first step in the cascade vs the other steps, i.e., NAG vs CAG and NAG vs IM (**Fig. 3a**). In another dataset (GSE55696), we studied the dysplasia→neoplasia cascade, which is typically scored by histopathological examination, as per the Vienna classification [23]; the latter comprises a continuum extending from low to high grade dysplasia to intramucosal carcinoma. Here, we looked at chronic gastritis (CG) vs low-grade intestinal neoplasia (LGIN), LGIN vs high-grade intestinal neoplasia (HGIN), HGIN vs early gastric cancer (EGC), CG vs HGIN and CG vs EGC (**Fig. 3b**). We compared GC-BoNE to the other signatures (**Fig. 3c**) and found that our signature again outperformed the others when looking at progression (see **Online Resource 2** for a list of GSE IDs; average ROC-AUC for GC-BoNE is 0.828, and other signatures range from 0.633 - 0.806). These findings suggest the genes identified in GC-BoNE may provide further insight into what initiates GC progression.

GC-BoNE can objectively assess the appropriateness of mouse models for studying human GC

Next, we wanted to identify mouse models that recapitulated human normal versus GC. We analyzed 38 mouse models [24-41] from 20 NCBI GEO datasets using C#11-2-4-14 and C#7-13-14 (see **Online Resource 2** for a list of GSE IDs; **Fig. 3d**). Many of the mouse models had a perfect ROC-AUC of 1.00 using C#11-2-4-14 and C#7-13-14 (see **Online Resource 4**). We then looked at which mouse models are significantly different using a t-test to determine the top ten models (**Fig. 3e**). It is noteworthy that the top two models represent the two common risk factors for GC in humans. The model that ranked #1 (GSE13873) is one in which the *H. pylori*

infection→GC cascade is modeled in C57Bl6 mouse model of experimental infection with the closely related *H. felis*. The authors showed that while most infected mice develop premalignant lesions such as gastric atrophy, compensatory epithelial hyperplasia and IM, a minority is completely protected from preneoplasia. The models that ranked #2-6 (GSE103639 (NGE vs pCP_GC), GSE45956, GSE103639 (NGE vs pChePS_GC), GSE16902, GSE93774) were all genetically engineered mouse models (GEMMs) in which targeted deletions were performed on genes (*CDH1, SMAD4, CLDN18* etc.) that are associated with risk of GC, by virtue of being either the most common germline mutation in GC (*CDH1 [42]*), or for harboring disease-associated SNPs (*SMAD4 [43]*) or being the target of the most frequent somatic genomic rearrangements [44] (*CLDN18*). These results suggest that *GC-BoNE* can objectively assess the degree of similarity between mouse models (both infection-induced and genetically-induced types) and human GC. In doing so, it can pinpoint which mouse models best recapitulate the patterns of gene expression that is observed during the transformation from healthy to GC in human samples.

GC-BoNE (C#11-2-4-14) can prognosticate the risk of IM→GC progression

Because we want to identify genes responsible for the progression of GC, we looked at a dataset that curated samples from a prospective study [45] with long-term follow-up (a mean of 12±3.4 years) to evaluate risk of progression to GC among patients with incomplete or complete intestinal metaplasia (IIM and CIM respectively) (**Fig. 4a**). It is known that among the types of intestinal metaplasia, IIM carries a greater risk for progression to GC compared to CIM [46]. A recent meta-analysis showed that compared with CIM, pooled RR of cancer/dysplasia in IIM patients was 4.48 (95% CI 2.50–8.03), and the RR was 4.96 (95% CI 2.72–9.04) for cancer, and 4.82 (95% CI 1.45–16.0) for dysplasia [47]. We found that C#11-2-4-14 best distinguished the healthy control patients (HC), patients with high risk-carrying IIM that progressed (IIM-GC) and those that did not progress (IIM-C) (ROC-AUC values: HC vs IIM-C: 0.86, HC vs IIM-GC: 0.94, IIM-C vs IIM-GC: 0.95; **Fig.**

4b). C#11-2-4-14 was not able to significantly distinguish (using Student's t-test) low risk-carrying CIM from HC. C#7-13-14 also could distinguish HC vs IIM-C (ROC-AUC = 0.80) and HC vs IIM-GC (ROC-AUC = 0.88), but not IIM-C vs IIM-GC (ROC-AUC = 0.71), however C#11-2-4-14 performed better (**Fig. 4c**). The DEA (Li 2015) gene signature similarly separates HC vs IIM-C (ROC-AUC = 0.90) and HC vs IIM-GC (ROC-AUC = 0.87) but is not able to distinguish IIM-C vs IIM-GC (ROC-AUC = 0.38) (**Fig. 4d**). The Japanese GC signature cannot significantly distinguish any of the samples (ROC-AUC values range from 0.42 - 0.74; **Fig. 4e**). These findings suggest genes in C#11-2-4-14 might be key to understanding why some IIM patients progress to GC.

GC-BoNE provides insights into the changes in cellular continuum states during healthy→IIM→GC progression

To understand which cellular processes change during cell transformation and which genes contribute to the progression of GC, we checked how clusters in C#11-2-4-14 and C#7-13-14 perform separately (**Fig. 4f**). When looking at HC vs IIM-C (**Fig. 4f row i**), cluster 14 is not able to distinguish the samples (ROC-AUC = 0.63), but both C#11-2-4 and C#7-13 are able to separate the samples (ROC-AUC = 0.87, 0.89 respectively). However, when you compare IIM-C vs IIM-GC (**Fig. 4f row ii**), cluster 14 is better able to distinguish the samples (ROC-AUC = 0.86), with C#11-2-4-14 best able to classify the samples (ROC-AUC = 0.95). These results show genes in C#11-2-4 might be responsible for the progression from HC to IIM, while C#14 is important for IIM to GC. Findings thereby suggest that the progression from HC to IIM may be impacted by genes related to muscle contraction, cell cycle and immune system, while the progression from IIM to GC is affected by extracellular matrix processes.

DISCUSSION

Although the incidence rates of GC have been decreasing around the world [4], there have not been any significant improvements in terms of new therapeutics, diagnostics and changes in screening designed for preneoplastic stages. In this study, we built a Boolean implication network using GSE66229 and used machine learning (on GSE37023 and GSE122401) to identify a gene signature (GC-BoNE) which could classify normal and gastric samples. Reactome pathway analysis of GC-BoNE revealed the following biological processes in the GC tumor samples: decrease in muscle contraction and ion transport, and increase in cell cycle, immune system and extracellular matrix functions (**Fig. 1e**). Although previous studies have identified most of these pathways [15-22], muscle contraction has not been widely identified, providing a new area to focus on researching. We then tested how GC-BoNE compares to gene signatures from past studies in both normal vs GC samples (**Fig. 2c**) and GC progression samples (**Fig. 3c and 4f**).

Our Boolean network-based approach improves upon past studies by *First*, identifying a gene signature (GC-BoNE) that is better able to classify samples along the GC disease continuum compared to previous signatures. When looking at normal vs GC samples, many of the signatures performed well (**Fig. 2c**). However, we are more interested in finding a gene signature that can distinguish samples earlier in the GC disease continuum. When looking at GC progression, our signature outperforms the other gene signatures (**Fig. 3c**). Since the genes in GC-BoNE do not overlap with many genes from the other gene signatures (**Fig. 1e**), this provides a list of new potential biomarkers for targeting therapeutics at different points along the GC disease continuum.

Second, we identified C#11-2-4 as important in the progression for HC to IIM-C, while C#14 is important for the progression from IIM-C to IIM-GC (**Fig. 4f**). Although the model was built and trained on N vs GC samples, using a Boolean network-based approach allows us to identify paths that can also determine the intermediate states of disease progression. The invariant asymmetric Boolean implications present in the GC-BoNE signature provide insight into the cellular changes occurring at various time points along the disease continuum. These findings

provide a list of gene targets that can be tested using the mouse models we identified (**Fig. 3e**) or other models. Genes in C#11-2-4 with cellular processes affecting muscle contraction, cell cycle and immune system can be targeted for drug development in patients with IIM before they advance to GC. Genes affecting extracellular matrix processes in C#14 can be targeted for patients with GC.

Overall, we demonstrate that the genes identified from our Boolean network-based approach were better able to classify samples along the GC disease continuum compared to the genes from previous work. The genes from GC-BoNE provide more opportunities to research the cellular processes behind GC progression. Results from this paper can be used to rationalize gene targets for diagnostics and therapeutics.

AUTHORS' DISCLOSURES

Authors declare that they have no competing interests.

AUTHORS' CONTRIBUTIONS

Conceptualization: D.S, P.G

Methodology: D.S, D.V.

Investigation: D.V, D.S, P.G

Visualization: D.V, D.S, P.G

Funding acquisition: D.S, P.G

Project administration: D.S, P.G

Supervision: D.S, P.G

Writing – original draft: D.V, D.S, P.G

Writing – review & editing: D.V, D.S, P.G

ACKNOWLEDGEMENTS

This work was supported by the National Institutes for Health (NIH) grant R01-AI155696 (to PG and DS). Other sources of support include: T32GM139790 (to DV), R01-GM138385 (to DS), R01-AI141630, CA100768 and CA160911 (to P.G), and UG3TR002968 (to D.S. and P.G). D.S was also supported by two Padres Pedal the Cause awards (Padres Pedal the Cause/RADY #PTC2017 and San Diego NCI Cancer Centers Council (C3) #PTC2017). D.S and P.G were also supported by the Leona M. and Harry B. Helmsley Charitable Trust.

We would also like to thank Saptarshi Sinha, Dharanidhar Dang and Sahar Taheri for providing feedback on the manuscript.

REFERENCES

1. Van Cutsem E, Sagaert X, Topal B, Haustermans K, Prenen H. Gastric cancer. Lancet. 2016;388(10060):2654-64. Epub 2016/05/10. doi: 10.1016/S0140-6736(16)30354-3. PubMed PMID: 27156933.

2. Amieva M, Peek RM, Jr. Pathobiology of Helicobacter pylori-Induced Gastric Cancer. Gastroenterology. 2016;150(1):64-78. Epub 2015/09/20. doi: 10.1053/j.gastro.2015.09.004. PubMed PMID: 26385073; PubMed Central PMCID: PMCPMC4691563.

3. Correa P, Piazuelo MB. The gastric precancerous cascade. J Dig Dis. 2012;13(1):2-9. Epub 2011/12/23. doi: 10.1111/j.1751-2980.2011.00550.x. PubMed PMID: 22188910; PubMed Central PMCID: PMCPMC3404600.

4. Karimi P, Islami F, Anandasabapathy S, Freedman ND, Kamangar F. Gastric cancer: descriptive epidemiology, risk factors, screening, and prevention. Cancer Epidemiol Biomarkers Prev. 2014;23(5):700-13. Epub 2014/03/13. doi: 10.1158/1055-9965.EPI-13-1057. PubMed PMID: 24618998; PubMed Central PMCID: PMCPMC4019373.

5. Gonzalez CA, Sanz-Anquela JM, Companioni O, Bonet C, Berdasco M, Lopez C, et al. Incomplete type of intestinal metaplasia has the highest risk to progress to gastric cancer: results

of the Spanish follow-up multicenter study. J Gastroenterol Hepatol. 2016;31(5):953-8. Epub 2015/12/03. doi: 10.1111/jgh.13249. PubMed PMID: 26630310.

6. Junnila S, Kokkola A, Mizuguchi T, Hirata K, Karjalainen-Lindsberg ML, Puolakkainen P, et al. Gene expression analysis identifies over-expression of CXCL1, SPARC, SPP1, and SULF1 in gastric cancer. Genes Chromosomes Cancer. 2010;49(1):28-39. Epub 2009/09/26. doi: 10.1002/gcc.20715. PubMed PMID: 19780053.

7. Park S, Nam CM, Kim SG, Mun JE, Rha SY, Chung HC. Comparative efficacy and tolerability of third-line treatments for advanced gastric cancer: A systematic review with Bayesian network meta-analysis. Eur J Cancer. 2021;144:49-60. Epub 2020/12/19. doi: 10.1016/j.ejca.2020.10.030. PubMed PMID: 33338727.

8. Korhani Kangi A, Bahrampour A. Predicting the Survival of Gastric Cancer Patients Using Artificial and Bayesian Neural Networks. Asian Pac J Cancer Prev. 2018;19(2):487-90. Epub 2018/02/27. doi: 10.22034/APJCP.2018.19.2.487. PubMed PMID: 29480983; PubMed Central PMCID: PMCPMC5980938.

9. Sahoo D, Swanson L, Sayed IM, Katkar GD, Ibeawuchi SR, Mittal Y, et al. Artificial intelligence guided discovery of a barrier-protective therapy in inflammatory bowel disease. Nat Commun. 2021;12(1):4246. Epub 2021/07/14. doi: 10.1038/s41467-021-24470-5. PubMed PMID: 34253728; PubMed Central PMCID: PMCPMC8275683.

10. Sahoo D, Katkar GD, Khandelwal S, Behroozikhah M, Claire A, Castillo V, et al. AI-guided discovery of the invariant host response to viral pandemics. EBioMedicine. 2021;68:103390. Epub 2021/06/16. doi: 10.1016/j.ebiom.2021.103390. PubMed PMID: 34127431; PubMed Central PMCID: PMCPMC8193764.

11. Ghosh P, Sinha S, Katkar GD, Vo DT, Taheri S, Dang D, et al. Machine Learning Identifies Signatures of Macrophage Reactivity and Tolerance that Predict Disease Outcomes. bioRxiv. 2022.

12. Sahoo D, Dill DL, Gentles AJ, Tibshirani R, Plevritis SK. Boolean implication networks derived from large scale, whole genome microarray datasets. Genome Biol. 2008;9(10):R157. Epub 2008/11/01. doi: 10.1186/gb-2008-9-10-r157. PubMed PMID: 18973690; PubMed Central PMCID: PMCPMC2760884.

13. Oh SC, Sohn BH, Cheong JH, Kim SB, Lee JE, Park KC, et al. Clinical and genomic landscape of gastric cancer with a mesenchymal phenotype. Nat Commun. 2018;9(1):1777. Epub 2018/05/05. doi: 10.1038/s41467-018-04179-8. PubMed PMID: 29725014; PubMed Central PMCID: PMCPMC5934392.

14. Sahoo D, Dill DL, Tibshirani R, Plevritis SK. Extracting binary signals from microarray time-course data. Nucleic Acids Res. 2007;35(11):3705-12. Epub 2007/05/23. doi: 10.1093/nar/gkm284. PubMed PMID: 17517782; PubMed Central PMCID: PMCPMC1920252.

15. Ma X, Huang M, Wang Z, Liu B, Zhu Z, Li C. ZHX1 Inhibits Gastric Cancer Cell Growth through Inducing Cell-Cycle Arrest and Apoptosis. J Cancer. 2016;7(1):60-8. Epub 2016/01/02. doi: 10.7150/jca.12973. PubMed PMID: 26722361; PubMed Central PMCID: PMCPMC4679382.

16. Zhang L, Kang W, Lu X, Ma S, Dong L, Zou B. LncRNA CASC11 promoted gastric cancer cell proliferation, migration and invasion in vitro by regulating cell cycle pathway. Cell Cycle. 2018;17(15):1886-900. Epub 2018/09/12. doi: 10.1080/15384101.2018.1502574. PubMed PMID: 30200804; PubMed Central PMCID: PMCPMC6152531.

17. Kono K, Nakajima S, Mimura K. Current status of immune checkpoint inhibitors for gastric cancer. Gastric Cancer. 2020;23(4):565-78. Epub 2020/05/30. doi: 10.1007/s10120-020-01090-4. PubMed PMID: 32468420.

18. Szor DJ, Dias AR, Pereira MA, Ramos M, Zilberstein B, Cecconello I, et al. Prognostic Role of Neutrophil/Lymphocyte Ratio in Resected Gastric Cancer: A Systematic Review and Meta-analysis. Clinics (Sao Paulo). 2018;73:e360. Epub 2018/06/21. doi: 10.6061/clinics/2018/e360. PubMed PMID: 29924187; PubMed Central PMCID: PMCPMC5996440.

19. Yuan D, Ma Z, Tuo B, Li T, Liu X. Physiological Significance of Ion Transporters and Channels in the Stomach and Pathophysiological Relevance in Gastric Cancer. Evid Based Complement Alternat Med. 2020;2020:2869138. Epub 2020/02/28. doi: 10.1155/2020/2869138. PubMed PMID: 32104192; PubMed Central PMCID: PMCPMC7040404.

20. Djamgoz MB, Coombes RC, Schwab A. Ion transport and cancer: from initiation to metastasis. Philos Trans R Soc Lond B Biol Sci. 2014;369(1638):20130092. Epub 2014/02/05. doi: 10.1098/rstb.2013.0092. PubMed PMID: 24493741; PubMed Central PMCID: PMCPMC3917347.

21. Moreira AM, Pereira J, Melo S, Fernandes MS, Carneiro P, Seruca R, et al. The Extracellular Matrix: An Accomplice in Gastric Cancer Development and Progression. Cells. 2020;9(2). Epub 2020/02/13. doi: 10.3390/cells9020394. PubMed PMID: 32046329; PubMed Central PMCID: PMCPMC7072625.

22. Jang M, Koh I, Lee JE, Lim JY, Cheong JH, Kim P. Increased extracellular matrix density disrupts E-cadherin/beta-catenin complex in gastric cancer cells. Biomater Sci. 2018;6(10):2704-13. Epub 2018/08/29. doi: 10.1039/c8bm00843d. PubMed PMID: 30151505.

23. Schlemper RJ, Riddell RH, Kato Y, Borchard F, Cooper HS, Dawsey SM, et al. The Vienna classification of gastrointestinal epithelial neoplasia. Gut. 2000;47(2):251-5. Epub 2000/07/18. doi: 10.1136/gut.47.2.251. PubMed PMID: 10896917; PubMed Central PMCID: PMCPMC1728018.

24. Sayi A, Kohler E, Hitzler I, Arnold I, Schwendener R, Rehrauer H, et al. The CD4+ T cell-mediated IFN-gamma response to Helicobacter infection is essential for clearance and determines gastric cancer risk. J Immunol. 2009;182(11):7085-101. Epub 2009/05/21. doi: 10.4049/jimmunol.0803293. PubMed PMID: 19454706.

25. Garay J, Piazuelo MB, Majumdar S, Li L, Trillo-Tinoco J, Del Valle L, et al. The homing receptor CD44 is involved in the progression of precancerous gastric lesions in patients infected with Helicobacter pylori and in development of mucous metaplasia in mice. Cancer Lett. 2016;371(1):90-8. Epub 2015/12/08. doi: 10.1016/j.canlet.2015.10.037. PubMed PMID: 26639196; PubMed Central PMCID: PMCPMC4714604.

26. Douchi D, Yamamura A, Matsuo J, Melissa Lim YH, Nuttonmanit N, Shimura M, et al. Induction of Gastric Cancer by Successive Oncogenic Activation in the Corpus. Gastroenterology. 2021;161(6):1907-23 e26. Epub 2021/08/16. doi: 10.1053/j.gastro.2021.08.013. PubMed PMID: 34391772.

27. Garay J, Piazuelo MB, Lopez-Carrillo L, Leal YA, Majumdar S, Li L, et al. Increased expression of deleted in malignant brain tumors (DMBT1) gene in precancerous gastric lesions: Findings from human and animal studies. Oncotarget. 2017;8(29):47076-89. Epub 2017/04/20. doi: 10.18632/oncotarget.16792. PubMed PMID: 28423364; PubMed Central PMCID: PMCPMC5564545.

28. An L, Nie P, Chen M, Tang Y, Zhang H, Guan J, et al. MST4 kinase suppresses gastric tumorigenesis by limiting YAP activation via a non-canonical pathway. J Exp Med. 2020;217(6). Epub 2020/04/10. doi: 10.1084/jem.20191817. PubMed PMID: 32271880; PubMed Central PMCID: PMCPMC7971137.

29. Choi W, Kim J, Park J, Lee DH, Hwang D, Kim JH, et al. YAP/TAZ Initiates Gastric Tumorigenesis via Upregulation of MYC. Cancer Res. 2018;78(12):3306-20. Epub 2018/04/20. doi: 10.1158/0008-5472.CAN-17-3487. PubMed PMID: 29669762.

30. Giannakis M, Backhed HK, Chen SL, Faith JJ, Wu M, Guruge JL, et al. Response of gastric epithelial progenitors to Helicobacter pylori Isolates obtained from Swedish patients with chronic atrophic gastritis. J Biol Chem. 2009;284(44):30383-94. Epub 2009/09/03. doi: 10.1074/jbc.M109.052738. PubMed PMID: 19723631; PubMed Central PMCID: PMCPMC2781593.

31. Shimada S, Akiyama Y, Mogushi K, Ishigami-Yuasa M, Kagechika H, Nagasaki H, et al. Identification of selective inhibitors for diffuse-type gastric cancer cells by screening of annotated compounds in preclinical models. Br J Cancer. 2018;118(7):972-84. Epub 2018/03/13. doi: 10.1038/s41416-018-0008-y. PubMed PMID: 29527007; PubMed Central PMCID: PMCPMC5931092.

32. Oshima H, Ishikawa T, Yoshida GJ, Naoi K, Maeda Y, Naka K, et al. TNF-alpha/TNFR1 signaling promotes gastric tumorigenesis through induction of Noxo1 and Gna14 in tumor cells. Oncogene. 2014;33(29):3820-9. Epub 2013/08/27. doi: 10.1038/onc.2013.356. PubMed PMID: 23975421.

33. Ihler F, Vetter EV, Pan J, Kammerer R, Debey-Pascher S, Schultze JL, et al. Expression of a neuroendocrine gene signature in gastric tumor cells from CEA 424-SV40 large T antigen-transgenic mice depends on SV40 large T antigen. PLoS One. 2012;7(1):e29846. Epub 2012/01/19. doi: 10.1371/journal.pone.0029846. PubMed PMID: 22253802; PubMed Central PMCID: PMCPMC3258231.

34. Liu J, Feng W, Liu M, Rao H, Li X, Teng Y, et al. Stomach-specific c-Myc overexpression drives gastric adenoma in mice through AKT/mammalian target of rapamycin signaling. Bosn J Basic Med Sci. 2021;21(4):434-46. Epub 2020/12/02. doi: 10.17305/bjbms.2020.4978. PubMed PMID: 33259779; PubMed Central PMCID: PMCPMC8292868.

35. Yu L, Wu D, Gao H, Balic JJ, Tsykin A, Han TS, et al. Clinical Utility of a STAT3-Regulated miRNA-200 Family Signature with Prognostic Potential in Early Gastric Cancer. Clin Cancer Res. 2018;24(6):1459-72. Epub 2018/01/14. doi: 10.1158/1078-0432.CCR-17-2485. PubMed PMID: 29330205.

36. Karasawa F, Shiota A, Goso Y, Kobayashi M, Sato Y, Masumoto J, et al. Essential role of gastric gland mucin in preventing gastric cancer in mice. J Clin Invest. 2012;122(3):923-34. Epub 2012/02/07. doi: 10.1172/JCI59087. PubMed PMID: 22307328; PubMed Central PMCID: PMCPMC3287219.

37. Loe AKH, Francis R, Seo J, Du L, Wang Y, Kim JE, et al. Uncovering the dosage-dependent roles of Arid1a in gastric tumorigenesis for combinatorial drug therapy. J Exp Med. 2021;218(6). Epub 2021/04/07. doi: 10.1084/jem.20200219. PubMed PMID: 33822841; PubMed Central PMCID: PMCPMC8034383.

38. Hagen SJ, Ang LH, Zheng Y, Karahan SN, Wu J, Wang YE, et al. Loss of Tight Junction Protein Claudin 18 Promotes Progressive Neoplasia Development in Mouse Stomach. Gastroenterology. 2018;155(6):1852-67. Epub 2018/09/10. doi: 10.1053/j.gastro.2018.08.041. PubMed PMID: 30195448; PubMed Central PMCID: PMCPMC6613545.

39. Park JW, Kim MS, Voon DC, Kim SJ, Bae J, Mun DG, et al. Multi-omics analysis identifies pathways and genes involved in diffuse-type gastric carcinogenesis induced by E-cadherin, p53, and Smad4 loss in mice. Mol Carcinog. 2018;57(7):947-54. Epub 2018/03/13. doi: 10.1002/mc.22803. PubMed PMID: 29528141.

40. Park JW, Jang SH, Park DM, Lim NJ, Deng C, Kim DY, et al. Cooperativity of E-cadherin and Smad4 loss to promote diffuse-type gastric adenocarcinoma and metastasis. Mol Cancer Res. 2014;12(8):1088-99. Epub 2014/05/03. doi: 10.1158/1541-7786.MCR-14-0192-T. PubMed PMID: 24784840; PubMed Central PMCID: PMCPMC4230498.

41. Itadani H, Oshima H, Oshima M, Kotani H. Mouse gastric tumor models with prostaglandin E2 pathway activation show similar gene expression profiles to intestinal-type human gastric cancer. BMC Genomics. 2009;10:615. Epub 2009/12/18. doi: 10.1186/1471-2164-10-615. PubMed PMID: 20015407; PubMed Central PMCID: PMCPMC2805698.

42. Luo W, Fedda F, Lynch P, Tan D. CDH1 Gene and Hereditary Diffuse Gastric Cancer Syndrome: Molecular and Histological Alterations and Implications for Diagnosis And Treatment. Front Pharmacol. 2018;9:1421. Epub 2018/12/21. doi: 10.3389/fphar.2018.01421. PubMed PMID: 30568591; PubMed Central PMCID: PMCPMC6290068.

43. Wu DM, Zhu HX, Zhao QH, Zhang ZZ, Wang SZ, Wang ML, et al. Genetic variations in the SMAD4 gene and gastric cancer susceptibility. World J Gastroenterol. 2010;16(44):5635-41. Epub 2010/11/26. doi: 10.3748/wjg.v16.i44.5635. PubMed PMID: 21105199; PubMed Central PMCID: PMCPMC2992684.

44. Zhang WH, Zhang SY, Hou QQ, Qin Y, Chen XZ, Zhou ZG, et al. The Significance of the CLDN18-ARHGAP Fusion Gene in Gastric Cancer: A Systematic Review and Meta-Analysis. Front Oncol. 2020;10:1214. Epub 2020/09/29. doi: 10.3389/fonc.2020.01214. PubMed PMID: 32983960; PubMed Central PMCID: PMCPMC7492548.

45. Companioni O, Sanz-Anquela JM, Pardo ML, Puigdecanet E, Nonell L, Garcia N, et al. Gene expression study and pathway analysis of histological subtypes of intestinal metaplasia that progress to gastric cancer. PLoS One. 2017;12(4):e0176043. Epub 2017/04/26. doi: 10.1371/journal.pone.0176043. PubMed PMID: 28441455; PubMed Central PMCID: PMCPMC5404762.

46. Du S, Yang Y, Fang S, Guo S, Xu C, Zhang P, et al. Gastric Cancer Risk of Intestinal Metaplasia Subtypes: A Systematic Review and Meta-Analysis of Cohort Studies. Clin Transl Gastroenterol. 2021;12(10):e00402. Epub 2021/10/02. doi: 10.14309/ctg.0000000000000402. PubMed PMID: 34597278; PubMed Central PMCID: PMCPMC8487777.

47. Wei N, Zhou M, Lei S, Zhong Z, Shi R. A meta-analysis and systematic review on subtypes of gastric intestinal metaplasia and neoplasia risk. Cancer Cell Int. 2021;21(1):173. Epub 2021/03/19. doi: 10.1186/s12935-021-01869-0. PubMed PMID: 33731114; PubMed Central PMCID: PMCPMC7968216.

48. Li H, Yu B, Li J, Su L, Yan M, Zhang J, et al. Characterization of differentially expressed genes involved in pathways associated with gastric cancer. PLoS One. 2015;10(4):e0125013. Epub 2015/05/01. doi: 10.1371/journal.pone.0125013. PubMed PMID: 25928635; PubMed Central PMCID: PMCPMC4415781.

49. Li L, Zhu Z, Zhao Y, Zhang Q, Wu X, Miao B, et al. FN1, SPARC, and SERPINE1 are highly expressed and significantly related to a poor prognosis of gastric adenocarcinoma revealed by microarray and bioinformatics. Sci Rep. 2019;9(1):7827. Epub 2019/05/28. doi: 10.1038/s41598-019-43924-x. PubMed PMID: 31127138; PubMed Central PMCID: PMCPMC6534579.

50. Takeno A, Takemasa I, Doki Y, Yamasaki M, Miyata H, Takiguchi S, et al. Integrative approach for differentially overexpressed genes in gastric cancer by combining large-scale gene expression profiling and network analysis. Br J Cancer. 2008;99(8):1307-15. Epub 2008/10/02. doi: 10.1038/sj.bjc.6604682. PubMed PMID: 18827816; PubMed Central PMCID: PMCPMC2570518.

51. Zang S, Guo R, Xing R, Zhang L, Li W, Zhao M, et al. Identification of differentially-expressed genes in intestinal gastric cancer by microarray analysis. Genomics Proteomics Bioinformatics. 2014;12(6):276-83. Epub 2014/12/17. doi: 10.1016/j.gpb.2014.09.004. PubMed PMID: 25500430; PubMed Central PMCID: PMCPMC4411479.

52. Wang JB, Li P, Liu XL, Zheng QL, Ma YB, Zhao YJ, et al. An immune checkpoint score system for prognostic evaluation and adjuvant chemotherapy selection in gastric cancer. Nat Commun. 2020;11(1):6352. Epub 2020/12/15. doi: 10.1038/s41467-020-20260-7. PubMed PMID: 33311518; PubMed Central PMCID: PMCPMC7732987.

53. Cho JY, Lim JY, Cheong JH, Park YY, Yoon SL, Kim SM, et al. Gene expression signature-based prognostic risk score in gastric cancer. Clin Cancer Res. 2011;17(7):1850-7. Epub 2011/03/31. doi: 10.1158/1078-0432.CCR-10-2180. PubMed PMID: 21447720; PubMed Central PMCID: PMCPMC3078023.

54. Wang H, Wu X, Chen Y. Stromal-Immune Score-Based Gene Signature: A Prognosis Stratification Tool in Gastric Cancer. Front Oncol. 2019;9:1212. Epub 2019/11/30. doi: 10.3389/fonc.2019.01212. PubMed PMID: 31781506; PubMed Central PMCID: PMCPMC6861210.

FIGURE LEGENDS

Fig. 1 Generation and validation of Boolean implication network-derived gastric cancer (GC) signature

a. Schematic summarizing the workflow to build a Boolean map using a gastric cancer microarray dataset containing tumor and adjacent normal samples (GSE66229)

b. Disease map representing the continuum from normal stomach to gastric cancer

c. Selection of Boolean path using machine learning on two training datasets (GSE37023 and GSE122401). Multivariate regression was used to determine which path best separated the tumor from the adjacent normal samples. Coefficient of each path score (at the center) with 95% confidence intervals (as error bars) and the p values were illustrated in the bar plot. The p value for each term tests the null hypothesis that the coefficient is equal to zero (no effect)

d. Violin plots showing the top Boolean paths in each of the training datasets

e. Reactome pathway analysis of the gene clusters in the GC-BoNE signature

Fig. 2 Comparison of classification accuracy using GC-BoNE signature versus gene signatures from previous literature for normal versus GC samples

a. Schematic summarizing the workflow to compare GC-BoNE to other gene signatures

b. Bar plots of GC datasets comparing normal (N)/AN vs T showing the ROC-AUC values for the Boolean paths in the GC-BoNE signature (11-2-4-14 and 7-13-14). Asterisks (*) after the ROC-AUC values represent the following: *p<=0.05, **p<=0.01, ***p<=0.001, no asterisk: p-value>0.05

c. Comparison of average ROC-AUC values for the datasets in B. using GC-BoNE and other gene signatures [6, 48-54] (DEA: Differential Expression Analysis, PPIN: Protein-Protein Interaction

Network, INGP: Ingenuity Pathway analysis, ISS: Immune Scoring System, SI: Stromal-Immune score; See Online Resource 3 for the complete list of genes in these signatures)

d-f. Venn diagrams showing the overlaps in genes in the top four gene signatures (GC-BoNE, DEA (Li 2015), DEA+PPIN and Japanese GC)

Fig. 3 GC-BoNE signature in GC progression and mouse models

a. Violin plots for GC progression of normal active gastritis (NAG) → chronic active gastritis (CG) → chronic atrophic gastritis (CAG) → intestinal metaplasia (IM) (E-MTAB-8889) using the GC-BoNE signature: 11-2-4-14 (*left*) and 7-13-14 (*right*)

b. Violin plots for chronic gastritis (CG) → low-grade intestinal neoplasia (LGIN) → high-grade intestinal neoplasia (HGIN) → early gastric cancer (EGC) (GSE55696) using the GC-BoNE signature: 11-2-4-14 (*left*) and 7-13-14 (*right*)

c. Comparison of average ROC-AUC values for GC progression datasets using GC-BoNE and other gene signatures

d. Schematic summarizing comparison of 38 mouse models from 20 GEO datasets using GC-BoNE

e. Top ten mouse models according to $-\log_{10}$(p-value) from Welch's Two Sample t-test separated by path (7-13-14: blue, 11-2-4-14: orange)

Fig. 4 GC-BoNE signature predicts outcome

a. Schematic summarizing GSE78523: samples collected from healthy patients (HC) and patients with incomplete IM (IIM) or complete IM (CIM). After a mean of 12 += 3.4 years, patients with IM were diagnosed as non-progressors (control: C) or progressors (GC)

b-e. Violin plots showing classification of samples using GC-BoNE, DEA (Li 2015), and Japanese GC signatures (**b**: 11-2-4-14, **c**: 7-13-14, **d**: DEA (Li 2015), **e**: Japanese GC)

f. GSE78523 is visualized as bubble plots of ROC-AUC values (radius of circles is based on the ROC-AUC) demonstrating the direction of gene regulation (Up: red, Down: blue) for the classification of samples (GC-BoNE clusters in columns; sample comparison in rows). P-values based on Welch's T-test (of composite score of gene expression values) are provided using the standard code (*p<=0.05, **p<=0.01, ***p<=0.001) next to the ROC-AUC

ONLINE RESOURCES

Online Resource 1 Supplementary methods

Online Resource 2 List of GSE IDs used in the analysis along with sample type (human vs mouse), use (network, training, validation) and figure panel

Online Resource 3 Complete list of genes used in all gene signatures (GC-BoNE and signatures from other sources)

Online Resource 4 Bubble plots of ROC-AUC values (radius of circles is based on the ROC-AUC) demonstrating the direction of gene regulation (Up: red, Down: blue) for the classification of samples in 38 mouse models (GC-BoNE clusters in columns; sample comparison in rows). P-values based on Welch's T-test (of composite score of gene expression values) are provided using the standard code (*p<=0.05, **p<=0.01, ***p<=0.001) next to the ROC-AUC

Figure 1

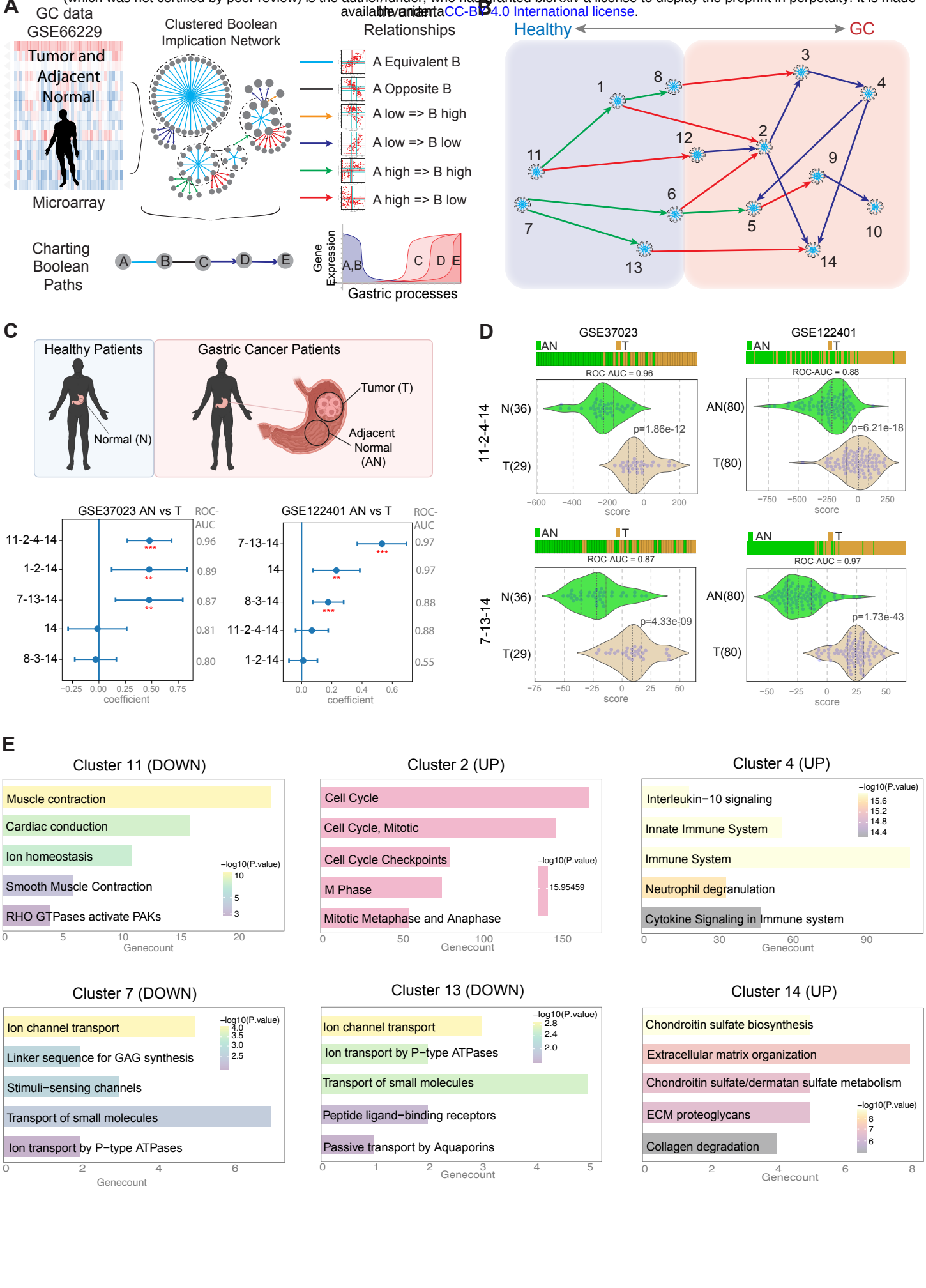
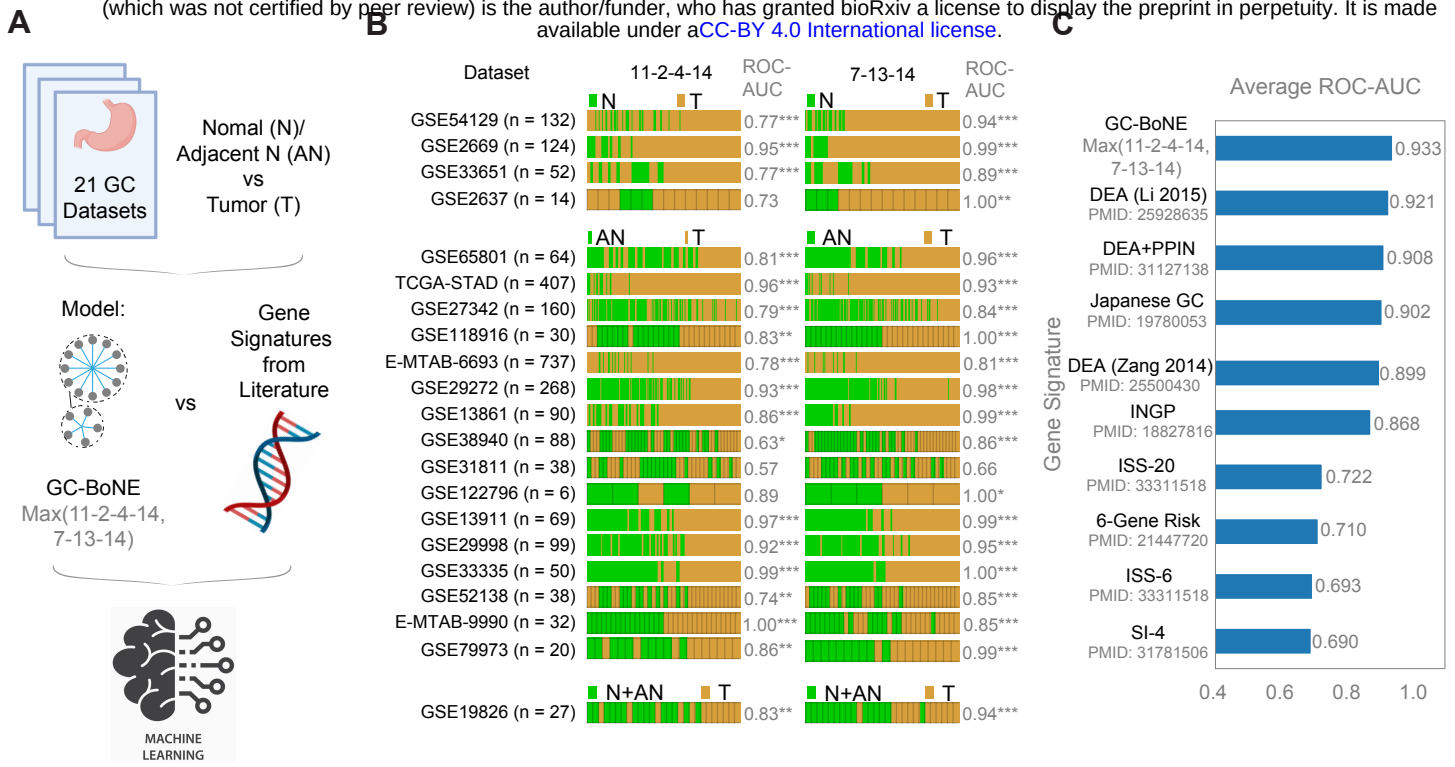
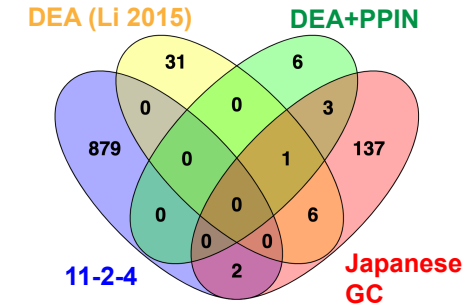
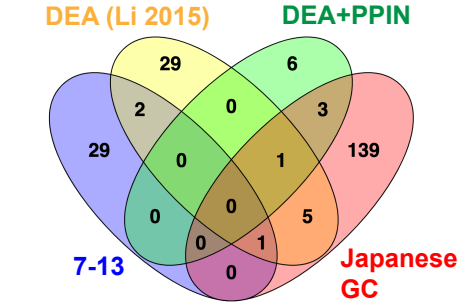


Figure 2

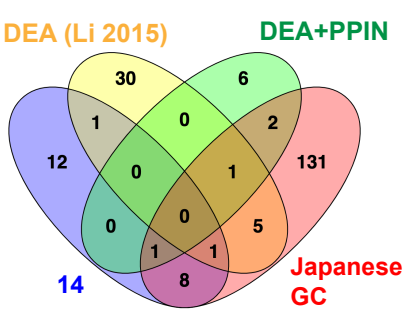




11-2-4, Japanese GC: *CXCL1, SLAMF8*



7-13, DEA (Li 2015), Japanese GC: *ATP4A*
7-13, DEA (Li 2015): *CCKBR, ATP4B*



14, DEA (Li 2015): *FAP*
14, PPIN (Li 2019), Japanese GC: *SPARC*
14, Japanese GC: *THY1, COL5A1, CTHRC1, OLFML2B, BGN, NID2, SULF1, COL5A2*

Figure 3

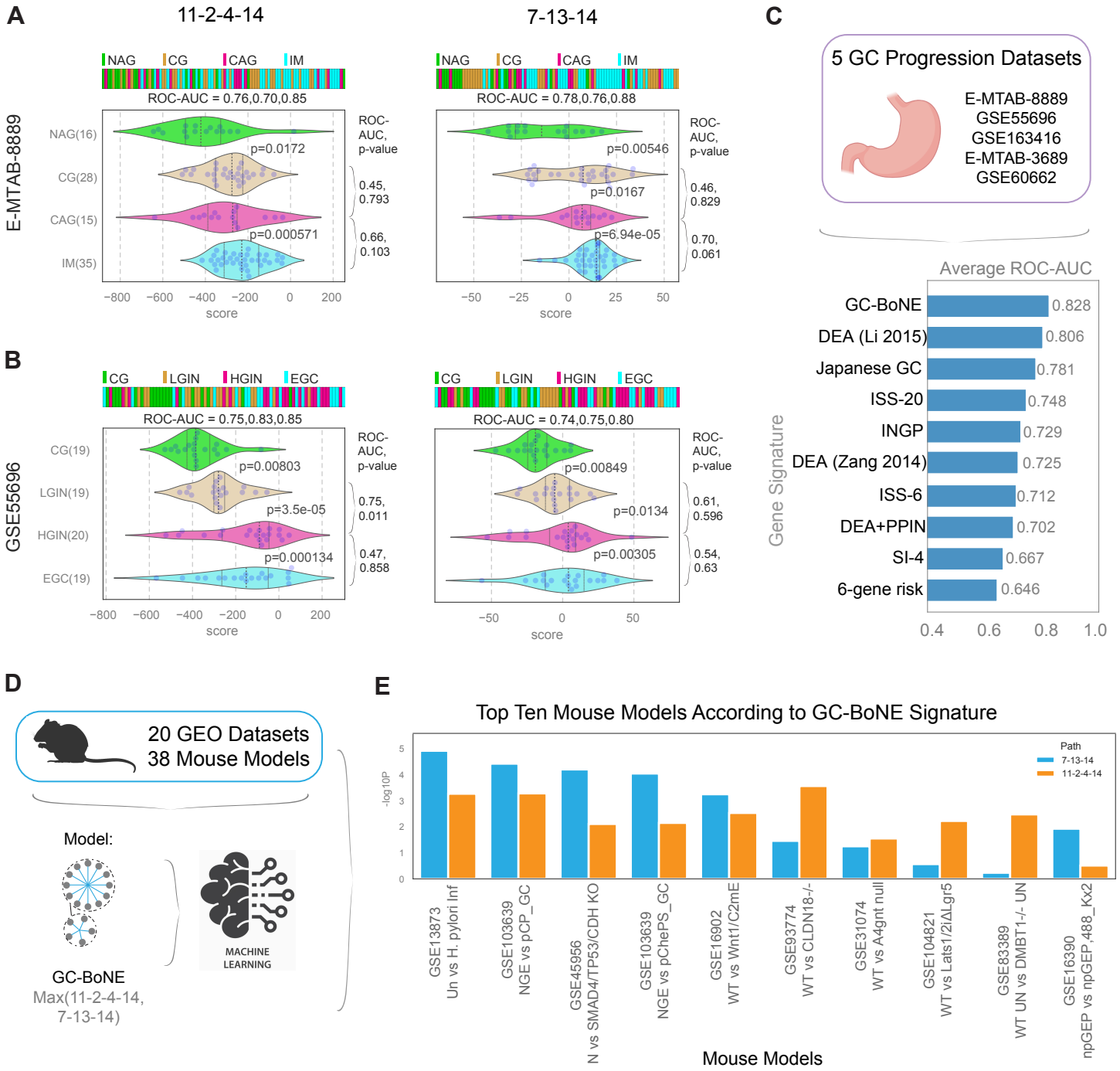


Figure 4