The Gastric Cancer Registry: A Genomic Translational Resource for Multidisciplinary Research in Gastric Cancer

Alison F. Almeda[1], Susan M. Grimes[1], HoJoon Lee[1], Stephanie Greer[1], GiWon Shin[1], Madeline McNamara[1], Anna C. Hooker[1], Maya M. Arce[1], Matthew Kubit[1], Marie C. Schauer[1], Paul Van Hummelen[1], Cindy Ma[1], Meredith A. Mills[1], Robert J. Huang[2], Joo Ha Hwang[2], Manuel R. Amieva[3], Summer S. Han[4], James M. Ford[1], and Hanlee P. Ji[1]

ABSTRACT

Background: Gastric cancer is a leading cause of cancer morbidity and mortality. Developing information systems which integrate clinical and genomic data may accelerate discoveries to improve cancer prevention, detection, and treatment. To support translational research in gastric cancer, we developed the Gastric Cancer Registry (GCR), a North American repository of clinical and cancer genomics data.

Methods: Participants self-enrolled online. Entry criteria into the GCR included the following: (i) diagnosis of gastric cancer, (ii) history of gastric cancer in a first- or second-degree relative, or (iii) known germline mutation in the gene *CDH1*. Participants provided demographic and clinical information through a detailed survey. Some participants provided specimens of saliva and tumor samples. Tumor samples underwent exome sequencing, whole-genome sequencing, and transcriptome sequencing.

Results: From 2011 to 2021, 567 individuals registered and returned the clinical questionnaire. For this cohort 65% had a personal history of gastric cancer, 36% reported a family history of gastric cancer, and 14% had a germline *CDH1* mutation. 89 patients with gastric cancer provided tumor samples. For the initial study, 41 tumors were sequenced using next-generation sequencing. The data was analyzed for cancer mutations, copy-number variations, gene expression, microbiome, neoantigens, immune infiltrates, and other features. We developed a searchable, web-based interface (the GCR Genome Explorer) to enable researchers' access to these datasets.

Conclusions: The GCR is a unique, North American gastric cancer registry which integrates clinical and genomic annotation.

Impact: Available for researchers through an open access, web-based explorer, the GCR Genome Explorer will accelerate collaborative gastric cancer research across the United States and world.

Introduction

Gastric cancer is a leading cause of cancer morbidity and mortality worldwide (1). While incidence is lower in the United States, gastric cancer remains a major public health concern with an estimated 26,600 new cases in 2021 (2). Gastric cancer is diagnosed at generally advanced stages in the United States, where curative resection is often no longer possible (2, 3). These data underscore the need for additional translational research in gastric cancer etiology, prevention, early detection, and therapy.

Cancer registries are a valuable resource for collating clinical information. Some registries also contain biological specimen repositories of both cancerous and noncancerous tissue, allowing for somatic and germline genomic characterization through next-generation sequencing (NGS; 4, 5). Cancer registries that integrate clinical information with tumor genomic features are particularly useful in translational research (6, 7). Few registries focused on gastric cancer currently exist, particularly those with patient data and samples derived from the United States. The availability of medical records, epidemiologic data, and biospecimens of tumor tissue allow researchers to understand interactions between genetic, environmental, and other risk factors. This research is particularly relevant given the poor overall outcomes from gastric cancer in the United States, and the lack of established screening and surveillance programs for this deadly cancer.

To address this knowledge gap, we established the Gastric Cancer Registry (GCR) in 2011. The goal of this project is to integrate granular patient clinical data (collected through a detailed, 412-item online questionnaire) with a comprehensive genomic characterization of tumor samples. This includes gene expression, somatic mutations, copy-number variation (CNV), human leukocyte antigen genotypes, neoantigens, and intratumoral heterogeneity details. To facilitate public access to these data, we created the GCR Genome Explorer (https://gcregistry-explorer.stanford.edu/), a browser-based interactive tool which allows for querying of clinical and molecular annotation from the GCR. In this manuscript, we describe the overall study design, methods for sample collection and data generation, and characteristics of enrolled participants in the GCR over a 10-year period (2011–2021). We also review features of the GCR Genome Explorer and describe how this tool can be used by cancer researchers for translational research.

[1]Division of Oncology, Department of Medicine, Stanford University School of Medicine, Stanford, California. [2]Division of Gastroenterology and Hepatology, Department of Medicine, Stanford University School of Medicine, Stanford, California. [3]Division of Infectious Diseases, Department of Pediatrics, Stanford University School of Medicine, Stanford, California. [4]Department of Neurosurgery, Stanford University School of Medicine, Stanford, California.

Note: Supplementary data for this article are available at Cancer Epidemiology, Biomarkers & Prevention Online (http://cebp.aacrjournals.org/).

Corresponding Author: Hanlee P. Ji, Division of Oncology, Department of Medicine – Stanford University School of Medicine, CCSR 2245, 269 Campus Drive, Stanford, CA 94305-5151. Phone: 650-721-1503; Fax: 650-736-1454; E-mail: genomics_ji@stanford.edu

Cancer Epidemiol Biomarkers Prev 2022;XX:XX–XX

doi: 10.1158/1055-9965.EPI-22-0308

Materials and Methods

Study design

We gained approval from the Stanford University (Stanford, CA) Institutional Review Board (IRB) for the cross-sectional study design. The original letter of approval for IRB-20285 is found in Supplementary Materials. The study was performed in accordance with the ethical standards delineated in the 1964 Declaration of Helsinki and its later amendments. The final intended sample size of the GCR is 1,000 participants. Eligible participants were over 18 years of age and fulfilled one or more of the following criteria: personal history of histologically proven gastric cancer, a family history of gastric cancer in a first- or second-degree relative, and a known pathogenic or likely pathogenic mutation in the gene *CDH1*. From 2011 to 2014, only individuals with a gastric cancer diagnosis were included in the study. We added family history and a known germline *CDH1* mutation as eligibility criteria in 2014. Participants in the United States could donate biospecimens. Recruitment and enrollment for the GCR was conducted through referral by local clinicians, local and regional pamphlet distribution at conferences and interest/advocacy group events, social media advertisement, and a study website (https://gcregistry.stanford.edu/). Registrants used the website to access the questionnaire.

Study data were collected and managed using REDCap electronic data capture tools hosted at Stanford University. REDCap is a secure, web-based software platform designed to support data capture for research studies, providing (i) an intuitive interface for validated data capture; (ii) audit trails for tracking data manipulation and export procedures; (iii) automated export procedures for seamless data downloads to common statistical packages; and (iv) procedures for data integration and interoperability with external sources (8, 9). REDCap automatically assigns a numerical patient identifier that does not reveal personal information. Upon registration, participants reviewed a consent form within REDCap. After providing signed consent per an IRB-approved protocol and with a waiver of documentation, the participant proceeded to 412-item questionnaire (Supplementary Materials). The questionnaire focused on information relevant to genetic, lifestyle, environmental, and other risk factors related to gastric cancer. The questions regarded subjects' demographics, medical history, familial cancer history, and lifestyle behaviors. To provide updates in personal or family history, participants could submit additional information using a unique return code. Participants agreed to donating biospecimens in the consent form; therefore, the study team contacted individuals directly to obtain a signed release form for archival tissue specimens. The participants could also authorize the release of medical records related to the diagnosis and treatment of their cancer. This information included treatment summaries from any chemotherapy or radiotherapy, the operative notes from any surgeries and any pathology reports relating to the diagnosis of gastric cancer.

Biospecimen collection

We obtained tissue specimens from a subset of participants. These samples included archival tissue samples in the form of formalin fixed, paraffin embedded (FFPE) blocks, 5- to 20-μM–thick scrolls, and unstained slides. This type of archival tissue was a necessity given the logistic complexity of patient registrants that live across the United States and receive care at different medical centers. FFPE tissue is routinely made in pathology labs and easy to store, making it a more convenient source of genomic material compared with fresh-frozen (FF) tissue which requires complex logistic handling and refrigeration. While FFPE nucleic acids are more degraded than FF, our methods provided nucleic acids of sufficiently high quality for cancer genomic sequencing. Studies have shown that with specific methods of library preparation and bioinformatic adjustments accounting for the nature of FFPE DNA, CNV and gene mutations are highly concordant between FFPE and FF tissue (10–12).

This cohort included individuals with gastric cancer who had a surgical resection of the stomach or carriers of *CDH1* germline mutations who underwent a screening endoscopy with biopsy or preventative gastrectomy. From 2018 onward, participants had the option of donating a sample of saliva. We used the Oragene DNA collection kit (DNA Genotek Inc., catalog no. OGR-500, OGR-600) to collect these samples. All samples were assigned a study identification number which cannot be linked back to the participant or any protected health information.

Sample processing

Archival FFPE samples of gastric tumors were used for genomic studies. Genomic DNA and total RNA were extracted from the tumor samples using the Maxwell 16 system (Promega Corporation, catalog no. 4AS1130, PRAS1260). We assessed the FFPE nucleic acid quality at multiple stages through quantitation, genomic sizing, and fragmentation analysis. We used an ultrasonicator to shear genomic DNA to a desired length of 300 bp. The genomic DNA was transformed into sequencing libraries using the KAPA Hyper Prep Kit for Illumina (Roche, catalog no. KK8502) with 8-bp unique dual adapters. An amount of genomic DNA library was set aside for exome enrichment using xGen Lockdown Probes and reagents (Integrated DNA Technologies, catalog no. 1056114, 1075474). To prepare RNA sequencing (RNA-seq) libraries, the KAPA RNA HyperPrep Kit (Roche, catalog no. KK8540). The adapter ligation time was extended and followed by two-step bead clean-up to accommodate low quality samples. The genomic DNA, exome, and RNA libraries were pooled respectively and sequenced on the NovaSeq 6000 system (Illumina, catalog no. 20012850) for 150-base paired end sequencing at variable depths. Additional details about the methods are provided in Supplementary Methods.

NGS

The genomic DNA libraries underwent low-coverage 1–2X whole-genome sequencing (WGS). Exome libraries were sequenced at approximately 68X coverage to enable somatic mutation calling. The RNA libraries were sequenced at high depth for an average of 65 million reads per sample.

Bioinformatic analysis

Sequencing reads aligned to GRCh38. WGS reads were analyzed for CNV with CNVkit. To identify somatic copy-number changes for samples without a matched normal control, we used a normal reference genome data set as a comparison control (12). Whole-exome sequencing (WES) data was analyzed for SNVs and insertions/deletions of nucleotides in genomic DNA with Sentieon and TNhaplotyper2. We investigated RNA-seq data for multiple features, including: (i) gene expression level using HT-Seq, (ii) four-digit human leukocyte antigen (HLA) genotypes using OptiType, (iii) tumor immune infiltrate cell types with CIBERSORTx (13), and (iv) the microbiome using Kraken2 (14). Based on WES data and RNA-seq, candidate cancer neoantigens were identified. Additional detailed methods and bioinformatic pipeline information are provided in Supplementary Methods.

The GCR Genome Explorer

The GCR Genome Explorer contains clinical, genomic, and genetic data from the registry. Similar to practices of other cancer registries, the GCR only reports on somatic mutations; germline mutations are not reported due to privacy and confidentiality considerations. Germline *CDH1* mutations are stated as clinical data for GCR patients and the mutation details are not released. Within the GCR Genome Explorer, we also included other datasets derived from The Cancer Genome Atlas (TCGA) stomach adenocarcinoma (STAD) and esophageal carcinoma (ESCA) projects. The Variant Call Format files and CNVs for the TCGA cohorts were downloaded from the NCI Genomic Data Commons (15). Genomic and genetic TCGA data was processed using similar bioinformatics pipeline as GCR data (though with changes to accommodate for matched tumor-normal data), resulting in comparable bioinformatic analysis across all samples.

The portal is a two-tier client-server application written using Ruby on Rails (version 5.1.7, 2.4.9) with back-end database tables in MySQL 5.5.62 and deployed using Passenger and Apache2. The application server has 64GB RAM and 32 processors running Ubuntu 16.04. The database server has 32GB RAM and 16 processors running Ubuntu 16.04.

The user interface utilizes bootstrap version 3.4.1 for responsive sizing to different format clients and browsers. Standard formatting, search, and filtering capability for query tables is provided by the jQuery DataTables plugin. Highcharts is used for generation of all plots. All queries and plots are produced dynamically from the underlying database tables based on user query parameters.

Data accessibility

Primary data is available on the GCR Genome Explorer website (https://gcregistry-explorer.stanford.edu/). Other data generated in this study are available within the article and its supplementary data files.

Results

Participant population and demographics

From March 2011 to November 2021, 567 subjects enrolled in the study. For inclusion in the study, all participants were required to identify their eligibility status on the enrollment questionnaire. The majority reported only a personal history of gastric cancer ($N = 325$). Some participants met multiple eligibility criteria. Specifically, 10 patients with gastric cancer had a family history of gastric cancer, and another 10 patients with gastric cancer had a germline mutation in *CDH1*. 154 participants reported only a family history of gastric cancer, while 21 participants reported both a family history of gastric cancer and a germline *CDH1* mutation. 26 participants were only affected by a germline *CDH1* mutation, without family or personal gastric cancer history. Finally, 21 participants met all three eligibility criteria.

Eligibility status with respect to sex, age, race, and ethnicity is depicted in **Table 1**. Participants were predominantly female (63%), White (76%), and non-Hispanic (53%). The median age of all participants was 51 years (range: 18–92 years). The median age of participants with gastric cancer was 68 years. Some participants did not report their sex, race, ethnicity, or other demographic details as these questions were optional on the enrollment survey.

Commonly reported medications included nonsteroidal anti-inflammatories, proton pump inhibitors, and multivitamins. Common comorbidities included high blood pressure, high cholesterol, and other cancers. Nearly a quarter of participants with gastric cancer reported a history of gastric ulcers, gastroesophageal reflux disease, or gastric polyps. *Helicobacter pylori* (*H. pylori*) infection was reported by 16% of all participants, and 20% of all participants with gastric cancer. Notably, a high proportion of participants (65%) did not report or did not know their *H. pylori* status.

Biospecimens

For this study, 164 participants donated biological specimens. We collected 111 saliva samples and 89 tissue samples. 41 tumor GCR samples underwent sequencing with the results available on the GCR Genome Explorer (**Table 2**). Clinical and histologic characteristics of the sequenced tumors are depicted in **Table 2**. Of the specimens where Lauren classification was reported ($N = 24$), the majority were diffuse-type cancers ($N = 16$). The tumors were generally aggressive and poorly differentiated (59%). Many tumors were from patients with metastatic disease (54%). With respect to anatomic location, 10% of tumors arose from the cardia, 73% arose from the noncardia stomach, and 17% arose from an unreported location.

The GCR Genome Explorer and pilot data release

Public access is to the GCR Genome Explorer is available upon registration via the following URL: https://gcregistry-explorer.stanford.edu/users/sign_in. At time of publication, the Genome Explorer contains data from the initial 41 sequenced tumors through the GCR. In addition, genomic data from 443 TCGA gastric cancers and 185 TCGA esophageal cancers are available through the GCR Genome Explorer for cross-reference. Future releases will incorporate data from additional GCR tumor samples. A representative image of the GCR Genome Explorer home page showing available data sets is depicted in **Fig. 1**.

There are two tiers of results that are provided in the Genome Explorer. The first tier includes gene expression levels determined from RNA-seq, somatic copy number based on WGS, and somatic mutations derived from exome sequencing. The second tier uses the first-tier results. Different algorithms extrapolate characteristics of the clonal diversity (WGS, exome), cellular microenvironment (RNA-seq), microbiome content (RNA-seq), HLA genotypes (RNA-seq), and putative neoantigens (WGS, exome, RNA-seq).

Cellular and microbiome features reflect the content of the local tumor microenvironment. We extrapolated the cellular representation and microbial populations using each tumor's RNA-seq data. The cell results were based on processing with a deconvolution tool called CIBERSORTx (13). The analysis approximated the different cell types present in the tumor microenvironment (16). For those RNA-seq reads which did not align to the human genome, we used the Kraken2 program to determine if there were microbiome features that included bacterial genera (14).

Nonsynonymous mutations are a source of immunogenic peptides, called neoantigens, which are tumor-specific and not expressed in other normal cells. Tumor mutational load, and more specifically, neoantigen load, have been correlated with extent of T-cell reactivity, response to checkpoint therapy, and prognosis (17–19). Exome sequencing of these tumors allowed us to identify nonsynonymous mutations in the protein-coding portions of genes in patients with cancer and predict potential neoantigens. We identified candidate neoantigens from the exome and RNA-seq data (Supplementary Methods). A candidate neoantigen fulfilled the following criteria: (i) nonsynonymous somatic mutation, (ii) expressed in transcriptome data, and (iii) translated neopeptides with strong binding affinity

Table 1. Self-reported cohort characteristics by eligibility status from 2011 to 2021.

	All participants (*N* = 567)	Gastric cancer (*N* = 366)	Family history (*N* = 206)	*CDH1* mutation (*N* = 78)
	Frequency (%)			
Sex				
Female	357 (63%)	195 (53%)	160 (78%)	66 (85%)
Male	208 (37%)	170 (46%)	45 (22%)	12 (15%)
Unknown	2 (<1%)	1 (<1%)	1 (<1%)	0 (0%)
Age (years)				
<40	144 (25%)	61 (17%)	77 (37%)	28 (36%)
40–49	114 (20%)	76 (21%)	41 (20%)	17 (22%)
50–59	125 (22%)	92 (25%)	40 (19%)	17 (22%)
60–69	107 (19%)	77 (21%)	28 (14%)	13 (17%)
70–79	6 (1%)	6 (2%)	0 (0%)	0 (0%)
≥80	11 (2%)	8 (2%)	4 (2%)	0 (0%)
Unknown	60 (11%)	46 (13%)	16 (8%)	3 (4%)
Race/ethnicity				
White	432 (76%)	271 (74%)	158 (77%)	69 (88%)
Black	24 (4%)	20 (5%)	6 (3%)	2 (3%)
Native American	1 (<1%)	1 (<1%)	0 (0%)	0 (0%)
Asian	43 (8%)	34 (9%)	12 (6%)	4 (5%)
Pacific Islander	2 (<1%)	1 (<1%)	1 (<1%)	0 (0%)
Other	45 (8%)	25 (7%)	23 (11%)	3 (4%)
Missing	20 (4%)	14 (4%)	6 (3%)	0 (0%)
Ethnicity				
Hispanic	83 (15%)	52 (14%)	35 (17%)	6 (8%)
Non-Hispanic	443 (78%)	287 (78%)	154 (75%)	67 (86%)
Unknown	41 (7%)	27 (7%)	17 (8%)	5 (6%)
Comorbidities				
Diabetes	69 (12%)	57 (16%)	12 (6%)	8 (10%)
Hyperlipidemia	129 (23%)	85 (23%)	48 (23%)	18 (23%)
Hypertension	134 (24%)	99 (27%)	42 (20%)	19 (24%)
H. pylori				
Tested positive	91 (16%)	73 (20%)	22 (11%)	6 (8%)
Tested negative	108 (19%)	70 (19%)	37 (18%)	17 (22%)
Did not test	368 (65%)	223 (61%)	147 (71%)	55 (71%)
Risk factors				
Epstein-Barr virus	11 (2%)	5 (1%)	5 (2%)	2 (3%)
Gastric ulcer	112 (20%)	83 (23%)	34 (17%)	14 (18%)
Gastric polyps	40 (7%)	22 (6%)	22 (11%)	3 (4%)
Gastroesophageal reflux disease	131 (23%)	77 (21%)	56 (27%)	11 (14%)

As participants could fall into more than one eligibility category, some participants appear in multiple columns (i.e., a participant may have gastric cancer, a family history of gastric cancer, and a germline *CDH1* mutation, and therefore be represented in all columns).

to patient's own major histocompatibility molecules. Using a combination of the exome mutation calls, gene expression data and MHC genotypes, we generated a list of potential neoantigens for each tumor.

GCR Genome Explorer navigation

There are several ways of accessing results through the GCR Genome Explorer. Options include general summaries of the results as well as specific queries. All image files and tables are available for download. On the home page (**Fig. 1**), the sample sets are displayed. A user can query the GCR data set using the "Explore Study" feature, which directs to a landing page with multiple tabs. In the "Clinical Parameters" tab, the user will find visual and tabular representations of cohort characteristics. The cohort can be queried with regards to sex, race, ethnicity, cancer site, age at diagnosis, familial cancer history, smoking history, cancer diagnosis, Lauren classification, and histologic differentiation. Under "Gene Summary," two tables describe genes most frequently mutated or varied in copy number within the GCR cohort. All cancer-associated genes are listed, along with genes mutated in >10% of samples, or genes with copy number variation in more than 10% of samples. The percentage filter will change across cohorts due to rounding, and due to gene ranking ties (i.e., multiple genes being mutated across the same total number of samples). The tables include the gene name, cytoband, number of samples displaying the mutation or CNV and ranked order of genes based on the percentage of samples affected. In addition, we provide annotations for genes that are cancer-associated, oncogenes or tumor-suppressor genes. Copy number summaries of the pilot data set showed that many tumor samples had a high degree of genomic instability. WGS revealed extensive changes in gene copy number, with some amplified genes such as *ERBB2* (i.e., *HER2*) being clinically actionable. Amplifications of *ERBB2* are an indication for the use of trastuzumab, a therapeutic mAb. Exome sequencing allowed us to detect specific types of mutations. Each tumor sample contained unique combinations of missense mutations, frame shift deletions, and in-frame deletions across various genes.

Table 2. Overview of gastric tumor tissues in the GCR Genome Explorer ($N = 41$).

	Frequency (%)
Sex	
Male	24 (59%)
Female	16 (39%)
Not available	1 (2%)
Histology	
Gastric adenocarcinoma	37 (90%)
GIST	1 (2%)
Not available	2 (5%)
Histologic subtype	
Intestinal	4 (10%)
Diffuse	16 (39%)
Mixed	4 (10%)
Not available	17 (41%)
Histologic differentiation	
Well	1 (2%)
Moderate	3 (7%)
Moderately to poorly	3 (7%)
Poorly	24 (59%)
Not available	10 (24%)
AJCC tumor pathology	
T1	6 (15%)
T2	0 (0%)
T3	12 (29%)
T4	7 (17%)
Not available	16 (39%)
N0	9 (22%)
N1	3 (7%)
N2	4 (10%)
N3	7 (17%)
Not available	18 (44%)
M0	1 (2%)
MX	22 (54%)
Not available	18 (44%)
Lymph node involvement	
Positive	13 (32%)
Negative	12 (29%)
Not available	16 (39%)
Adjuvant radiation	
Yes	14 (34%)
No	20 (49%)
Not available	7 (17%)
Tumor anatomic site	
Gastroesophageal junction	5 (12%)
Cardia	4 (10%)
Pylorus	2 (5%)
Fundus	3 (7%)
Body	8 (20%)
Antrum	5 (12%)
Entire stomach	7 (17%)
Not available	7 (17%)

Abbreviation: AJCC, American Joint Committee on Cancer.

The "HLA Types" tab contains an overview and breakdown of specific HLA alleles for 20 patients in the GCR study. There are options to view the prevalence of specific HLA-A, HLA-B, and HLA-C types within the study and within individual patients. In the "Immune Cells" tab, users identify and quantify tumor-infiltrating immune cell types in a bar plot as a percentage of overall immune cells present within a patient's tumor. Alternatively, users can view the immune cell presence represented as a heatmap. The "Microbiome" tab contains an interactive bar chart displaying the major bacteria phylum found within each patient. Users can select and deselect phylum of interest for a more granular, sample-specific view of the microbiome composition. In addition, users can see the number of sequence reads within the genus *Helicobacter* for each patient.

For more discrete summaries, users can employ the "Gene" query function to view data for a specific gene. We cite an example using the *CDH1* tumor suppressor gene. When searching all studies for *CDH1* gene information, the user will find discrete percentages and counts on its mutations, CNVs, expression levels, and neoantigens within individual patients and across all cohorts (GCR and TCGA; **Fig. 2**). The *CDH1* gene was mutated in 7.7% of GCR samples, 11% of TCGA-STAD samples and 1.1% of TCGA-ESCA samples. Missense mutations were most common. Nearly all samples in the GCR, TCGA-STAD, and TCGA-ESCA cohorts showed high expression of *CDH1* (100%, 99.5%, and 99.4%). Mutation data can be linked to specific patients for independent querying.

The "Neoantigen" query gives the user the option to select an HLA allele type and view a list of the candidate neoantigens. For example, the most prevalent HLA-A type in the GCR study is A x 02:01 (18.9%). Selecting this HLA type in the Neoantigen query produces a breakdown of all neoantigen candidates. They are described by their gene, chromosomal location, amino acid change, and binding strength/rank based on their predicted properties in terms of MHC1 interaction (**Fig. 3**). Like the "Gene" query, the "Neoantigen" query function allows users to search data across multiple patient cohorts.

Finally, the "Patient" query function provides users a way to view all available data for an individual patient in the database (**Fig. 4**). The patient's summary page includes a count of mutations and copy number variations, their HLA-A, -B, and -C types, and their clinical characteristics. The "Mutation" tab details the patient's unique mutations with respect to chromosome position, reference sequence and alteration, variant type, and amino acid change. The "Copy Number" page lists genes with duplication and deletion events. In the "Gene Expression" tab, the user can view the expression level of a gene through fragments per kilobase of transcript per million mapped reads and qualitative values of high, medium, and low. The gene list can be filtered to include only cancer-associated genes, oncogenes, or tumor suppressor genes. The next tab contains putative neoantigens described by position, amino acid change, and binding strength to a specific HLA allele. The final tab displays the patient's microbiome, a taxonomy of microbial populations classified from kingdom to genus.

Discussion

We present the GCR: an integrated clinical and genomic registry of both individuals with gastric cancer and at heightened risk for gastric cancer drawn from the United States. The GCR contains comprehensive risk factor and treatment data through detailed questionnaire along with multi-level, multi-omic tumor profiling. It is a unique resource which will accelerate gastric cancer prevention, early detection, and personalized therapy. To facilitate collaboration, GCR is publicly accessible through a user-friendly, browser-based interface, the GCR Genome Explorer. This powerful tool will allow researchers from across the world to access comprehensive tumor data including patterns of gene expression, somatic mutations, CNV, human leukocyte antigen genotypes, neoantigens, and intratumoral heterogeneity.

The GCR differs from existing gastric cancer–focused registries in several important aspects. Most prior registries recruited from regions of the world with high *H. pylori* prevalence and high gastric cancer incidence. For instance, fewer than 10% of individuals who provided

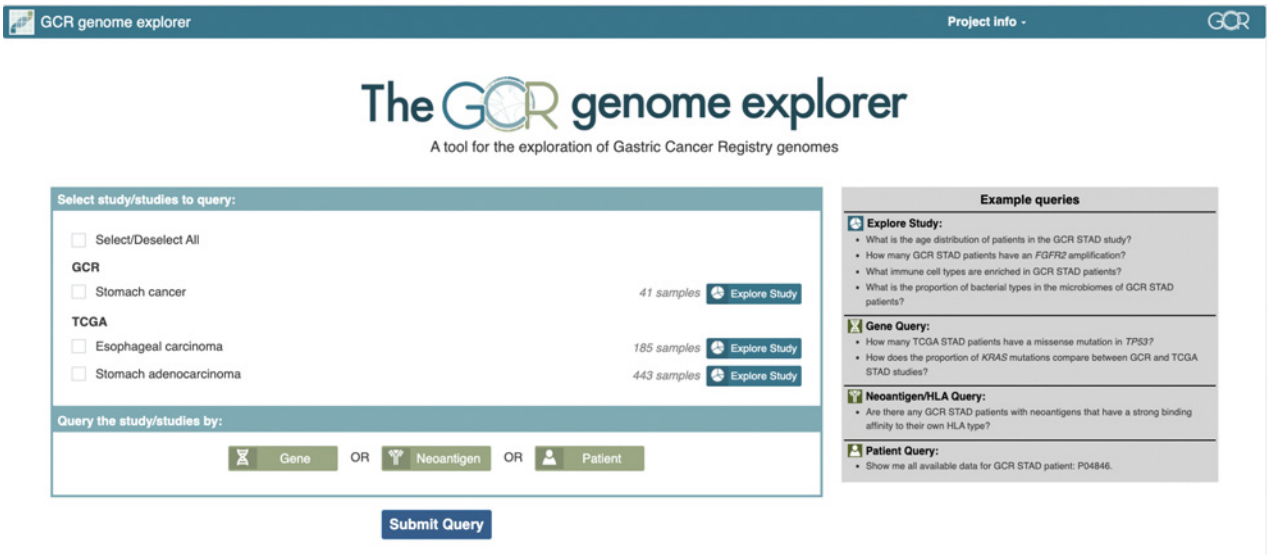


Figure 1.
The GCR Genome Explorer home page. The portal offers queries of gene, neoantigen, and patient datasets from one or multiple cohorts: the GCR, TCGA ESCA study, and the TCGA STAD study. The "Explore Study" feature enables the user to see the clinical and genomic characteristics of a single cohort.

tumor samples for TCGA were recruited from North America (20). Even larger international consortiums, such as the Stomach Cancer Pooling (StoP) Project, have fewer than 20% of participants from North America (21). There exists a need for genomic data specific to multiethnic, North American populations who have differing prevalence of *H. pylori* (22), and possibly different tumor anatomic distribution and histologic profiling. As one example, the GCR has a far higher frequency of diffuse-type cancers (per Lauren classification) compared with both TCGA and StoP, reflective of differences in disease burden between North America and other regions of the world. Also, while certain registries may have a wealth of granular clinical information (such as StoP), or genomic data (such as TCGA), very few integrate both into a single resource which can be parsed, queried, and categorized.

The Esophageal and Stomach Cancer Project is another North American registry that aims to create a clinical and genomic gastric cancer database (23). The GCR is distinguished by its more detailed questionnaire (412 fields compared with 16 fields) and sophisticated data portal. The GCR Genome Explorer enables queries across multiple study cohorts and down to the individual level.

The pilot release of the Genome Explorer contains genomic data for 41 gastric tumors. As we have biospecimens from over 160 individuals, we expect the number of sequenced tumors to rapidly increase. Moreover, though both promotion and through collaboration with research groups and healthcare centers across the globe, we are continuing to build out a robust repository of clinical datasets, biospecimens and genomic data. This influx of new participants brings forth additional clinical datasets such as pathology reports and other clinical metrics. On this expanded cohort we are conducting additional genomic studies which will greatly increase the overall number of tumors with genomic data. As the registry continues to accrue participants and tumor samples, we also anticipate that there will be

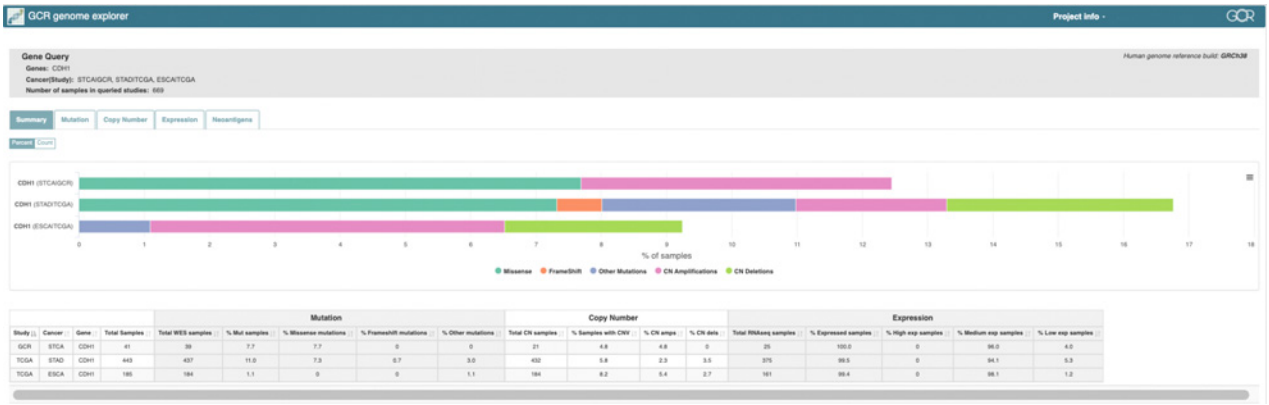
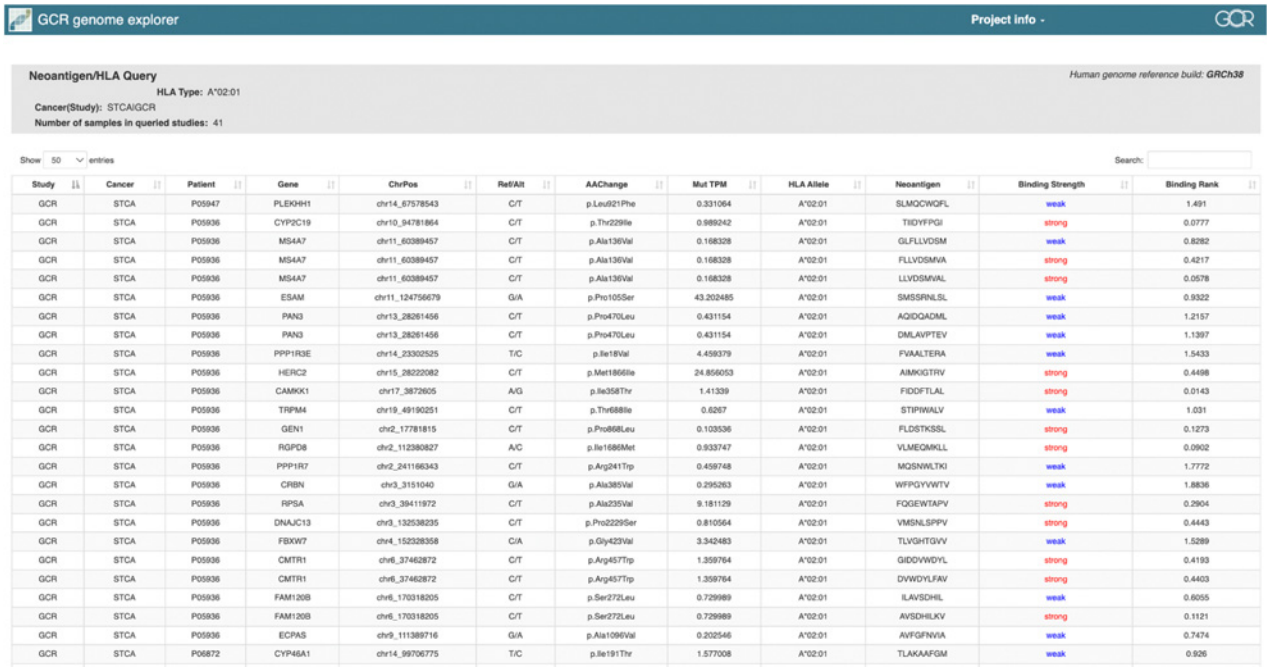


Figure 2.
The GCR Genome Explorer output of a "Gene" query across all study cohorts for *CDH1*. The "Summary" page displays counts and percentages of samples with *CDH1* mutations, CNVs, and expression levels. The "Mutation," "Copy Number," "Expression," and "Neoantigen" tabs contain more detailed gene alterations among the studies' patients.



Study	Cancer	Patient	Gene	ChrPos	Ref/Alt	AAChange	Mut TPM	HLA Allele	Neoantigen	Binding Strength	Binding Rank
GCR	STCA	P05947	PLEKHH1	chr14_67578543	C/T	p.Leu921Phe	0.331064	A*02:01	SLMQCWQFL	weak	1.491
GCR	STCA	P05936	CYP2C19	chr10_94781864	C/T	p.Thr229Ile	0.989242	A*02:01	TIDYFPGI	strong	0.0777
GCR	STCA	P05936	MS4A7	chr11_60389457	C/T	p.Ala136Val	0.168328	A*02:01	GLFLLVDSM	weak	0.8282
GCR	STCA	P05936	MS4A7	chr11_60389457	C/T	p.Ala136Val	0.168328	A*02:01	FLLVDSMVA	strong	0.4217
GCR	STCA	P05936	MS4A7	chr11_60389457	C/T	p.Ala136Val	0.168328	A*02:01	LLVDSMVAL	strong	0.0578
GCR	STCA	P05936	ESAM	chr11_124756679	G/A	p.Pro105Ser	43.202485	A*02:01	SMSSRNLSL	weak	0.9322
GCR	STCA	P05936	PAN3	chr13_28261456	C/T	p.Pro470Leu	0.431154	A*02:01	AQIDQADML	weak	1.2157
GCR	STCA	P05936	PAN3	chr13_28261456	C/T	p.Pro470Leu	0.431154	A*02:01	DMLAVPTEV	weak	1.1397
GCR	STCA	P05936	PPP1R3E	chr14_23302525	T/C	p.Ile18Val	4.459379	A*02:01	FVAALTERA	weak	1.5433
GCR	STCA	P05936	HERC2	chr15_28222082	C/T	p.Met1866Ile	24.856053	A*02:01	AIMKIGTRV	strong	0.4498
GCR	STCA	P05936	CAMKK1	chr17_3872605	A/G	p.Ile358Thr	1.41339	A*02:01	FIDDFTLAL	strong	0.0143
GCR	STCA	P05936	TRPM4	chr19_49190251	C/T	p.Thr688Ile	0.6267	A*02:01	STIPIWALV	weak	1.031
GCR	STCA	P05936	GEN1	chr2_17781815	C/T	p.Pro868Leu	0.103536	A*02:01	FLDSTKSSL	strong	0.1273
GCR	STCA	P05936	RGPD8	chr2_112380827	A/C	p.Ile1686Met	0.933747	A*02:01	VLMEQMKLL	strong	0.0902
GCR	STCA	P05936	PPP1R7	chr2_241166343	C/T	p.Arg241Trp	0.459748	A*02:01	MQSNWLTKI	weak	1.7772
GCR	STCA	P05936	CRBN	chr3_3151040	G/A	p.Ala385Val	0.295263	A*02:01	WFPGYVWTV	weak	1.8836
GCR	STCA	P05936	RPSA	chr3_39411972	C/T	p.Ala235Val	9.181129	A*02:01	FQGEWTAPV	strong	0.2904
GCR	STCA	P05936	DNAJC13	chr3_132538235	C/T	p.Pro2229Ser	0.810564	A*02:01	VMSNLSPPV	strong	0.4443
GCR	STCA	P05936	FBXW7	chr4_152328358	C/A	p.Gly423Val	3.342483	A*02:01	TLVGHTGVV	weak	1.5289
GCR	STCA	P05936	CMTR1	chr6_37462872	C/T	p.Arg457Trp	1.359764	A*02:01	GIDDVWDYL	strong	0.4193
GCR	STCA	P05936	CMTR1	chr6_37462872	C/T	p.Arg457Trp	1.359764	A*02:01	DVWDYLFAV	strong	0.4403
GCR	STCA	P05936	FAM120B	chr6_170318205	C/T	p.Ser272Leu	0.729989	A*02:01	ILAVSDHIL	weak	0.6055
GCR	STCA	P05936	FAM120B	chr6_170318205	C/T	p.Ser272Leu	0.729989	A*02:01	AVSDHILKV	strong	0.1121
GCR	STCA	P05936	ECPAS	chr9_111389716	G/A	p.Ala1096Val	0.202546	A*02:01	AVFGFNVIA	weak	0.7474
GCR	STCA	P06872	CYP46A1	chr14_99706775	T/C	p.Ile191Thr	1.577008	A*02:01	TLAKAAFGM	weak	0.926

Figure 3.
Example of the "Neoantigen" query for HLA type A x 02:01 within the GCR cohort. The table describes specific gene mutations that lead to the production of a neoantigen and the predicted binding strength of the neoantigen to HLA allele A x 02:01.

improved statistical power for future studies. The ongoing expansion of the GCR will provide an even more diverse cohort and broaden the utility of this data for research.

The novelty and significance of GCR must be weighed against several notable limitations. A significant limitation is missing data based on incomplete responses to the intake questionnaire. Participants may have chosen not to complete entries either due unwillingness to provide such information, or incomplete knowledge of the information (e.g., not knowing if they had previously been tested for *H. pylori* or not knowing details of their treatment history). We had biospecimens from only a subset of participants. The participants willing to provide biological samples may differ from the overall population enrolled in GCR. Furthermore, recruitment to GCR relied on a REDCap portal, limiting recruitment to individuals with means an ability to access the internet. These individuals may differ with respect to age, race and ethnicity, language ability, education, and socio-demographic characteristics compared with the overall population afflicted by gastric cancer (24).

Overall, the GCR and Genome Explorer represent a wealth of clinical and genomic information from a rapidly expanding cohort of patients either personally afflicted by gastric cancer or at high risk for gastric cancer. We sincerely hope the ongoing, open-access platform will both accelerate scientific discovery and foster collaborative research on this deadly disease.

Figure 4.
"Patient" query for P04906 from the GCR study. Multiple tabs present the entirety of clinical, genomic, and cellular data for this individual.

Authors' Disclosures

A.F. Almeda reports grants from Gastric Cancer Foundation during the conduct of the study. M. McNamara reports grants from Gastric Cancer Foundation during the conduct of the study. M.A. Mills reports other support from Gastric Cancer Fund during the conduct of the study. J.M. Ford reports grants from Gastric Cancer Foundation during the conduct of the study; grants from Genentech, AstraZeneca, Merus; and grants from PUMA outside the submitted work. H.P. Ji reports grants from Gastric Cancer Foundation and NIH during the conduct of the study. No disclosures were reported by the other authors.

Authors' Contributions

A.F. Almeda: Resources, investigation, writing–original draft, project administration. **S.M. Grimes:** Data curation, software, formal analysis, writing–original draft. **H. Lee:** Data curation, software, formal analysis, writing–review and editing. **S. Greer:** Software, formal analysis, writing–review and editing. **G. Shin:** Formal analysis, investigation. **M. McNamara:** Investigation, writing–original draft. **A.C. Hooker:** Investigation. **M.M. Arce:** Investigation. **M. Kubit:** Formal analysis, investigation. **M.C. Schauer:** Formal analysis, investigation. **P. Van Hummelen:** Conceptualization, formal analysis, supervision, funding acquisition, investigation, methodology, writing–original draft. **C. Ma:** Formal analysis, investigation, writing–review and editing. **M.A. Mills:** Conceptualization, formal analysis, supervision, funding acquisition, investigation, methodology. **R.J. Huang:** Conceptualization, supervision, funding acquisition, investigation, methodology, writing–original draft, writing–review and editing. **J.H. Hwang:** Conceptualization, supervision, funding acquisition, investigation, methodology, writing–original draft, writing–review and editing. **M.R. Amieva:** Conceptualization, supervision, funding acquisition, investigation, methodology, writing–original draft, writing–review and editing. **S.S. Han:** Conceptualization, supervision, funding acquisition, investigation, methodology, writing–original draft, writing–review and editing. **J.M. Ford:** Conceptualization, supervision, investigation, methodology. **H.P. Ji:** Conceptualization, supervision, funding acquisition, investigation, methodology, writing–original draft, writing–review and editing.

Acknowledgments

The work was supported by the Gastric Cancer Foundation. In addition support to H.P. Ji came from the Research Scholar Grant (grant no. RSG-13-297-01-TBG) from the American Cancer Society, and the Clayville Foundation. R.J. Huang is supported by the NCI of the NIH under Award Number K08CA252635. The REDCap platform services at Stanford are subsidized by Stanford School of Medicine Research Office, and the National Center for Research Resources and the National Center for Advancing Translational Sciences, NIH, (grant no. UL1 TR001085).

Received March 17, 2022; revised May 10, 2022; accepted June 23, 2022; published first June 30, 2022.

References

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin Nov 2018;68:394–424.
2. Howlader NNA, Krapcho M, Miller D, Brest A, Yu M, Ruhl J, editors [monograph on the internet]. SEER Cancer Statistics Review (CSR), 1975-2017. Bethesda, MD: National Cancer Institute; 2020. [cited 2020 May 1]. Available from: https://seer.cancer.gov/csr/1975_2017/.
3. Surveillance, Epidemiology, and End Results Program. SEER*Explorer: An interactive website for SEER cancer statistics. [cited 2020 Apr 15]. Available from: https://seer.cancer.gov/explorer/.
4. Huang RJ, Choi AY, Truong CD, Yeh MM, Hwang JH. Diagnosis and management of gastric intestinal metaplasia: current status and future directions. Gut Liver 2019;13:596–603.
5. Zhou L, Catchpoole D. Spanning the genomics era: the vital role of a single institution biorepository for childhood cancer research over a decade. Translational pediatrics 2015;4:93–106.
6. Liu A. Developing an institutional cancer biorepository for personalized medicine. Clin Biochem 2014;47:293–9.
7. Mccarty CA, Chisholm RL, Chute CG, Kullo IJ, Jarvik GP, Larson EB, et al. The eMERGE Network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies. BMC Med Genet 2011;4:1–11.
8. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)–a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform 2009;42:377–81.
9. Harris PA, Taylor R, Minor BL, Elliott V, Fernandez M, O'neal L, et al. The REDCap consortium: Building an international community of software platform partners. J Biomed Inform Jul 2019;95:103208.
10. Gao XH, Li J, Gong HF, Yu GY, Liu P, Hao LQ, et al. Comparison of fresh frozen tissue with formalin-fixed paraffin-embedded tissue for mutation analysis using a multi-gene panel in patients with colorectal cancer. Front Oncol 2020;10:310.
11. Greytak SR, Engel KB, Bass BP, Moore HM. Accuracy of molecular data generated with FFPE biospecimens: lessons from the literature. Cancer Res 2015;75:1541–7.
12. Xia LC, Van Hummelen P, Kubit M, Lee H, Bell JM, Grimes SM, et al. Whole genome analysis identifies the association of TP53 genomic deletions with lower survival in Stage III colorectal cancer. Sci Rep 2020;10:5009.
13. Newman AM, Steen CB, Liu CL, Gentles AJ, Chaudhuri AA, Scherer F, et al. Determining cell type abundance and expression from bulk tissues with digital cytometry. Nat Biotechnol 2019;37:773–82.
14. Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. Genome Biol 2019;20:257.
15. National Cancer Institute. Genomic data commons data portal. 2021. Available from: https://portal.gdc.cancer.gov/.
16. Aran D, Hu Z, Butte AJ. xCell: digitally portraying the tissue cellular heterogeneity landscape. Genome Biol 2017;18:220.
17. Brown SD, Warren RL, Gibb EA, Martin SD, Spinelli JJ, Nelson BH, et al. Neo-antigens predicted by tumor genome meta-analysis correlate with increased patient survival. Genome Res 2014;24:743–50.
18. Champiat S, Ferté C, Lebel-Binay S, Eggermont A, Soria JC. Exomics and immunogenics: Bridging mutational load and immune checkpoints efficacy. Oncoimmunology 2014;3:e27817.
19. Giannakis M, Mu XJ, Shukla SA, Qian ZR, Cohen O, Nishihara R, et al. Genomic correlates of immune-cell infiltrates in colorectal carcinoma. Cell Rep 2016;15:857–65.
20. Cancer Genome Atlas Research Network. Comprehensive molecular characterization of gastric adenocarcinoma. Nature 2014;513:202–9.
21. Pelucchi C, Lunet N, Boccia S, Zhang ZF, Praud D, Boffetta P, et al. The stomach cancer pooling (StoP) project: study design and presentation. Eur J Cancer Prev 2015;24:16–23.
22. Nguyen TH, Mallepally N, Hammad T, Liu Y, Thrift AP, El-Serag HB, et al. Prevalence of Helicobacter pylori positive non-cardia gastric adenocarcinoma is low and decreasing in a US population. Dig Dis Sci 2020;65:2403–11.
23. The esophageal and stomach cancer project patient data browser. Available from: https://escproject.org/data-release.
24. Buis LR, Janney AW, Hess ML, Culver SA, Richardson CR. Barriers encountered during enrollment in an internet-mediated randomized controlled trial. Trials 2009;10:76.