ABSTRACT: Prognostic markers for cancer can assist in the evaluation of survival probability of patients and help clinicians to assess the available treatment modalities. Gallbladder cancer (GBC) is a rare tumor that causes 165,087 deaths in the world annually. It is the most common cancer of the biliary tract and has a particularly high incidence in Chile, Japan, and northern India. Currently, there is no accurate diagnosis test or effective molecular markers for GBC identification. Several studies have focused on the discovery of genetic alterations in important genes associated with GBC to propose novel diagnosis pathways and to create prognostic profiles. To achieve this, we performed data-mining of GBC in public repositories, harboring 133 samples of GBC, allowing us to describe relevant somatic mutations in important genes and to propose a genetic alteration atlas for GBC. In our results, we reported the 14 most altered genes in GBC: arid1a, arid2, atm, ctnnb1, erbb2, erbb3, kmt2c, kmt2d, kras, pik3ca, smad4, tert, tp53, and znf521 in samples from Japan, the United States, Chile, and China. Missense mutations are common among these genes. The annotations of many mutations revealed their importance in cancer development. The observed annotations mentioned that several mutations found in this repository are probably oncogenic, with a putative loss-of-function. In addition, they are hotspot mutations and are probably linked to poor prognosis in other cancers. We identified another 11 genes, which presented a copy number alteration associated with GBC to propose novel diagnosis pathways and to create prognostic profiles. To achieve this, we performed data-mining of GBC cancer of the biliary tract and has a particularly high incidence in Chile, Japan, and northern India. Currently, there is no accurate diagnosis test or effective molecular markers for GBC identification. Several studies have focused on the discovery of genetic alterations in important genes associated with GBC to propose novel diagnosis pathways and to create prognostic profiles. To achieve this, we performed data-mining of GBC in public repositories, harboring 133 samples of GBC, allowing us to describe relevant somatic mutations in important genes and to propose a genetic alteration atlas for GBC. In our results, we reported the 14 most altered genes in GBC: arid1a, arid2, atm, ctnnb1, erbb2, erbb3, kmt2c, kmt2d, kras, pik3ca, smad4, tert, tp53, and znf521 in samples from Japan, the United States, Chile, and China. Missense mutations are common among these genes. The annotations of many mutations revealed their importance in cancer development. The observed annotations mentioned that several mutations found in this repository are probably oncogenic, with a putative loss-of-function. In addition, they are hotspot mutations and are probably linked to poor prognosis in other cancers. We identified another 11 genes, which presented a copy number alteration in gallbladder database samples, which are ccnd1, ccnd3, ccone, cdk12, cdkn2a, cdkn2b, erbb2, erbb3, kras, mdk2, and myc. The findings reported here can help to detect GBC cancer through the development of systems based on genetic alterations, for example, the development of a mutation panel specifically for GBC diagnosis, as well as the creation of prognostic profiles to accomplish the development of GBC and its prevalence.

KEYWORDS: Altered genes, gallbladder cancer, public databases, cancer diagnostic, cancer prognostic, mutation, computational approach

Introduction

Gallbladder cancer (GBC) is one of the most common malignant tumors of the extrahepatic bile ducts, with high incidence in Japan, Chile, and northern India.1 Gallbladder cancer is a rare but highly aggressive neoplasm, with a dismal prognosis and a median survival of less than 1 year in the locally advanced or metastatic setting.2 Symptoms only appear in advanced stages (stage III or IV) with 5-year survival rates <10%.3,4 Overall survival of early GBC (stage I) is close to 90%, however, most GBCs are diagnosed at advanced stages. Concerning its prevalence, annually this tumor occurs in 219,420 cases, and resulting in 165,087 deaths in the world.5 Previous works demonstrated significant racial/ethnic disparities in survival improvements for GBC.6,7 In Chile, the incidence of GBC is rising; it is one of the most frequent causes of cancer mortality in Chilean women (about 1620 cases), according to data from the World Health Organization (2014).8 Mapuche Indians from Valdivia, Chile, and South America exhibit the highest rate of GBC: 12.3/100,000 for men and 27.3/100,000 for women.9 Similarly, northern India (21.5/100,000) and south Karachi Pakistan (13.8/100,000) have been reported as 2 of the most affected regions regarding women.4 Gallbladder cancer is also found in high frequency in Asia and Eastern Europe, including Poland (14/100,000 in Poland), The Czech Republic, and Slovakia. Whereas, South Americans of Indian descent (3.7-9.1 per 100,000), Israel (5/100,000), and Japan (7/100,000) have shown an intermediate prevalence of GBC.10,11 China is 1 of the 5 Asian countries with the highest rates of GBC. These 5 countries have the highest number of GBC deaths which include China, Japan, India,
Bangladesh, and the Republic of Korea, and represent 88% of all GBCs observed in Asia.13 Residents of the Andean-area, North American Indians, and Mexican Americans are especially predisposed to GBC.14 The incidence in the United States is lower than that around the world, with a rate of 1.4 per 100,000 among women and 0.8 among men.15

Explanations for this geographic distribution are related to genetic susceptibility (more frequent in the Amerindian population), hormonal factors (mainly relating to estrogen), and environmental factors (lifestyle, infections, insufficient access to health services, and diet).5,11,16

Concerning the infections role, Iyer et al7 performed computational analyses to search for DNA sequences of Salmonella sp. and human papillomavirus (HPV) in GBC samples and their normal counterparts. Furthermore, they found a high incidence of Salmonella typhoidal and non-typhoidal sequences among the GBC samples data, which suggested a possible role of Salmonella infection in GBC. They suggested the possible linking of Salmonella sp. infections to gallbladder carcinogenesis, which could stimulate an inflammatory initiation process.17

Another work conducted by Scanu et al18 showed that Salmonella enterica could induce cellular transformation in GBC samples. For this, specific mutations (inactivation of arf/tp53 genes and amplification of c-myc genes) are required, which can pre-transform the host cell. The damage induced by S enterica during its infection cycle is translate in a cell transformation.18

Another important factor is age, with women above 65 years old who presented a history of gallstones.11,19,20 Although GBC is more common in female patients, in some countries like Korea, Iceland, and Costa Rica, higher mortality rates have been reported for male patients.21

When it comes to the genetic basis of GBC, like others neoplasia, this tumor is a multifactorial disorder involving multiple genetic alterations seen in several ethnicities.7,8 Many studies were performed to understand how certain types of genetic alterations act in GBC. For example, it is known that in samples of GBC, the genes kras,14 tp53,7,8,16,22 pik3ca,14 and c-erbB-222 are more highly altered. In addition to these genes, others presented several alterations in gallbladder tumors, like loss of heterozygosity, changes in the methylation pattern of DNA, as well as in their expressions.12 Despite all genetic knowledge already published, it is necessary to describe the genetic basis of the gallbladder carcinogenesis process in more depth to understand which genetic factors lead to its initiation, development, and progression. One way to achieve this is through the screening of predicted mutations for GBC.

Concerning the scenario of GBC in Chile and in other countries, we were motivated to perform a study to identify the most mutated genes in GBC through data-mining of public repositories. The goal of this work is to generate data to propose the identification of new molecular markers based on the more common mutations and specifics of GBC. In this approach, we performed our analysis on samples from United States, Japan, China, and Chile, which were already available on a cancer database. All the information generated could be applied to propose the development of diagnostics able to detect GBC in early stages, and the establishment of new tools of prognosis. We made available a descriptive atlas of genetic alterations, which are present in specific genes of GBC patients. The future should therefore be engaged in good quality research focused on early diagnosis and refinement of prognostic profile information.

Methods
Public repository

We performed genetic data-mining from 133 GBC samples. The repositories used were Gallbladder Cancer (MSK, Cancer 2018) and Gallbladder Carcinoma (Shanghai, Nat Genet 2014). The first dataset has targeted sequencing samples from 101 GBC patients. This repository includes samples from the United States (n = 49), Chile (n = 21), Japan (n = 11), and others (n = 20), and this database contains information regarding mutation and copy number alterations (CNAs) (heterozygous deletion, homozygous deletion, low-level gain, and high-level amplification) in gallbladder samples. The second repository available includes 32 samples from 32 GBC patients from China. All data were obtained from CBioPortal (www.cbioportal.org).

Based on these repositories, information about the somatic mutation of the most mutated and altered genes in GBC was obtained. The calculation to obtain the frequencies used the ratio of the number of somatic mutation occurrence in the studied gene and the total number of samples (n = 103 or n = 32), in GBC databases. The program to generate the mutation figures for 3 genes was Mutation Mapper (www.cbioportal.org). Heatmaps were constructed using the software “Java Treeview” to show and compare the number of cases in a graphical way.24

Results
Overview of gallbladder cancer data

In the public repositories analyzed, there were 135 samples from 133 patients. Within this dataset, the frequency of the occurrence of this type of cancer is higher in women (n = 80 cases, 60.2%) than in men (n = 53 cases, 39.8%). According to the literature, more than half of patients have a history of gallbladder stones (51.1%) before developing GBC. In the samples that had information about primary tumor sites, the gallbladder was the most common primary site (68.6%) followed by the liver (10.8%), for Gallbladder Cancer (MSK, Cancer 2018) data only. Furthermore, the common metastatic sites for patients with GBC were the liver (11.1%), pelvis (1.5%), and peritoneum (1.5%). The other data from GBC patients are summarized in Table 1.

In the studied dataset, the most prevalent GBC stages were stages III and IV. Gallbladder cancer is asymptomatic in the early stages. Patients only look for health attention when the symptoms appear, causing its prevalence to be higher in more
advanced phases. At this level, generally, the prognosis is poor. The scientific literature appointed similar findings; about 90% of patients are detected at advanced stages, and systemic chemotherapy is the mainstay of their treatment.

Most altered genes in gallbladder cancer

We found—in the gallbladder samples from the United States, Japan, Chile, and China—14 genes that presented significant values of somatic mutation, which are arid1a, arid2, atm, cttnb1, erbb2, erbb3, kmt2c, kmt2d, kras, pik3ca, smad4, tert, tp53, and znf521. Comparisons between the studied countries for frequencies of mutations are represented as a heatmap in Figure 1A. The genes tp53, smad4, and arid1a were the most mutated genes within the mentioned ethnicities, as described by Narayan et al with similar results. Mutations in genes atm, tert, kmt2d from the Gallbladder Cancer (MSK) repository and znf521 from the Gallbladder Carcinoma (Shanghai) database were not observed in previous researches for GBC; however, they were recurrent in the present GBC databases used. The US group was the only ethnicity that presented mutations in

Table 1. Patients' outcome with gallbladder cancer.

|                  | CHILE  | CHINA  | THE UNITED STATES | JAPAN  |
|------------------|--------|--------|-------------------|--------|
| Age Mean         | 59.0   | 60.0   | 66.4              | 72.0   |
| Sex Female       | 81.0%  | 66.7%  | 36.4%             | 61.2%  |
| Sample type      |        |        |                   |        |
| Primary          | 0      | 75%    | 87.8%             | 100%   |
| Metastasis       | 0      | 25%    | 12.2%             | 0      |
| No information   | 100%   | 0      | 0                 | 0      |
| TNM stage        |        |        |                   |        |
| I                | 0      | 0      | 2.0%              | 9.1%   |
| II               | 0      | 16.70% | 4.1%              | 27.3%  |
| III              | 23.8%  | 83.4%  | 44.9%             | 54.5%  |
| IV               | 76.2%  | 0      | 44.9%             | 9.1%   |
| Unknown          | 0      | 0      | 4.1%              | 0      |
| Tissue site      |        |        |                   |        |
| Gallbladder      | 76.2%  | 0      | 72.5%             | 81.8%  |
| Liver            | 23.8%  | 0      | 27.5%             | 18.2%  |
| No information   | 0      | 100%   | 0                 | 0      |

Data obtained from CBioPortal (www.cbioportal.org), data accessed on November 5, 2019.

TNM: tumour, node, metastasis.

Figure 1. Most mutated genes in gallbladder cancer in samples from China, Japan, Chile, the United States, and others centers. (A) Most mutated genes in gallbladder cancer. Chile (n = 21), China (n = 32), Japan (n = 11), the United States (n = 49), and others (n = 20). (B) Most altered genes in gallbladder cancer. Chile (n = 21), China (n = 32), Japan (n = 11), the United States (n = 49), and others (n = 20). Data obtained from CBioPortal (www.cbioportal.org, accessed on November 5, 2019). The numbers are in percentage of alterations.
almost all genes, except for the *znf521* gene. In the China and Japan datasets, mutations in the *kras* gene were not found. Chile and China presented the highest rate of mutation for the *kmt2d* gene and Chile still presented several mutations in the *kmt2c* gene, in contrast to the other countries. Despite, lacking information regarding the country of origin for the data of 20 samples, the tendency of mutation seems to be the same, with the most mutated genes being *tp53*, *smad4*, and *arid1a*.

Regarding CNAs, the most altered genes found in these databases were *ccnd1*, *ccnd3*, *ccne1*, *cdk12*, *cdkn2a*, *cdkn2b*, *erb2b*, *erb3b*, *kras*, *mdm2*, and *myc*. Comparisons between the studied countries for frequencies of genetic alterations are represented as a heatmap in Figure 1B. In samples from Japan, the United States, and Chile, *ccnd1*, *ccne1*, *cdk12*, *erb2b*, *erb3b*, *kras*, *mdm2*, and *myc* are amplified. However, the most frequent alteration for the *cdkn2a* and *cdkn2b* genes was deletion, which was observed in all ethnicities. The genes *mdm2* (13.2%), *erb3b* (7.8%), and *ccne1* (7.8%) are the most altered, mainly in samples from the United States. The repository containing the data from China does not present data of CNA alterations.

**Annotation of gallbladder cancer mutations**

Most mutations described with GBC presented some kind of annotation. In the database studied, 6 types of annotations were identified in accordance with previous studies of other cancers. Regarding clinical implications, some mutations received the annotation as probably oncogenic. When the mutations already have biological effects, they present a loss-of-function as an annotation. The others types are CIViC variants, my cancer genome, recurrent hotspot mutation, and clinical evidence levels. Many mutations do not present any annotation. All data about annotations of GBC mutations are summarized in Table 2.

**Tumor protein p53**

Figure 1A shows that the *tp53* gene is one of the most mutated genes in GBC, as seen in other cancers.26 Here, we described 75 mutations in the *tp53* gene for GBC samples (50.4%) (Table 2). We removed redundant mutations from the table. More common alterations in *tp53* gene were missense mutations (n = 46), followed by nonsense (n = 15) and frameshift deletions (n = 6). All 6 frameshift deletion and 3 frameshift insertion mutations received the annotation as probably oncogenic mutation, which can cause a probable loss-of-function of the *tp53* product. The 3 in-frame deletion mutations observed in the *tp53* gene were annotated as likely oncogenic and a hotspot of mutation in other cancers. The changed amino acids were identified as a recurrent hotspot (statistically significant) in a population-scale cohort of tumor samples of various cancer types using methodology based in part on Chang et al.27

Forty-five missense mutations have information about their probability to present oncogenic function and probable loss of biological function. Forty-three missense alterations were annotated as recurrent hotspot sites in other cancers and 26 missense mutations present the CIViC annotation.28 This annotation informs us of the relevance of clinical features linking these genetic alterations. Among these, 10 mutations (R273C, R248Q, R273H, R280T, R248W, R175H, V157F, S241F, G275S, and R213V) presented information regarding prognosis and function.

**SMAD family member 4**

Figure 1A shows that *smad4* is the second most mutated gene studied in the gallbladder repository. In this study, *smad4* presented 33 mutations (about 20.4%). The most common mutation was missense type (n = 18), with 14 of them annotated with oncogenic function and as recurrent hotspot sites in other cancers; 9 missense mutations were annotated with all mentioned information and also with genomic information (Table 2). Among these 9 missense alterations are R361C, R361H, D355V, D351N, and E330K. According to data from CBioPortal, these mutations occur inside of the MH2 domain of the SMAD4 protein. This MH2 domain plays a role in transcriptional activation and formation of the Smads homo- and heteromeric complex.29 According to literature, such mutations could reduce the DNA binding capacity of *smad4* and were linked to several cancer stages and their progressions.29 Other types of mutations like nonsense (n = 7), frameshift deletions (n = 2), frameshift insertions (n = 4), and in-frame deletions (n = 2) were annotated as probable oncogenic mutations with loss-of-function. Other in-frame deletions were annotated as recurrent hotspot site of mutations (see Table 2). In this present study, we also described the annotation of 5 *smad4* mutations, including 1 in-frame deletion and 4 missense mutations (F329del, H317Y, H528D, G359A, and Q83H), which did not present any annotation.

**AT-rich interactive domain-containing protein 1A**

The *arid1a* gene encodes a subunit of the barrier to autointegration factor (BAF) chromatin-remodeling complex and plays roles in transcriptional regulation and DNA damage response. Mutations in the *arid1a* gene that lead to inactivation or loss of expression are frequent and widespread across many cancer types.27 This is the third most mutated gene in GBC samples in almost all studied ethnicities and presented almost all genetic alterations as truncating mutations (nonsense [14 cases], frameshift insertion [4 cases], and deletion [7 cases]), revealing an altered frequency of around 20.7% (31 mutations, see Figure 1A and Table 2). All mutations of *arid1a* were annotated as oncogenic alterations with a loss-of-function, with 2 splice mutations included in this set of annotation and 4 missense mutations which did not present any annotation. These likely correspond to new mutations described in this dataset.
Table 2. Mutations and their annotations in GBC public data.

| GENE | MUTATIONS | ANNOTATION | PROBABLY ONCOGENIC | LOSS OF FUNCTION | HOTSPOT | CIVIC | MY CANCER GENOME | LEVEL PROGNOSTIC | WITHOUT ANNOTATION |
|------|-----------|------------|--------------------|------------------|---------|-------|------------------|------------------|-------------------|
| ARID1A | Missense | – | – | – | – | – | – | – | G856E; S2113A; K2146M; P1873R |
|        | FSDel A1978Sfs*36; L657Pfs*81; N2109Ffs*37; R727Afs*11 | – | – | – | – | – | – | – |
|        | FSIns D1850Gfs*4; Y551Lfs*72; Q1519Pfs*13; E38Rfs*73; Q1306Tfs*17; S2559Tfs*112 | – | – | – | – | – | – | – |
|        | Nonsense E1032*; Q1473*; Q1172*; Q708*; Q581*; Q1358*; Q529*; Q1573*; E2255*; Q487*; K2146*; L2147*; Q449*; Q1614* | – | – | – | – | – | – | – |
|        | Splice X960_splice; X1289_splice | X960_splice; X1289_splice | – | – | – | – | – | – |
| ARID2 | Missense S297P | S297P | – | – | – | – | S319F; G1425S; G1281W; |
|        | FSDel L1423Ws*6; N67tfs*23 | – | – | – | – | – | – | – |
|        | Nonsense Q819*; E267*; L57* | Q819*; E267*; L57* | – | – | – | – | – | – |
|        | Splice X572_splice | X572_splice | – | – | – | – | – | – |
| ATM | Missense D2870H | D2870H | – | – | – | – | D2870H | S2394L; F2410L; D2889H; R35Q; D2725H; G3051R; L739R; V2256L |
|        | FSDel V566tfs*6 | V566tfs*6 | – | – | – | – | V566tfs*6 | – |
|        | FSIns R2138Kfs*8; F2571Yfs*4; L2953Tfs*3 | – | – | – | – | – | R2138Kfs*8; F2571Yfs*4; L2953Tfs*3 | – |
|        | IFDel | – | – | – | – | – | – | I149del |
|        | Nonsense R1446*; E343* | R1446*; E343* | – | – | – | – | – | – |
|        | Splice X221_splice; X301_splice; X2951_splice | X221_splice; X301_splice; X2951_splice | – | – | – | – | X221_splice; X301_splice; X2951_splice | – |
| CTNNB1 | Missense D32V; S45P | D32V; S45P | D32V; S45P | S45P | S45P | – | R717C | (Continued) |
| GENE   | MUTATIONS | ANNOTATION |
|--------|-----------|------------|
|        |           | PROBABLY ONCOCGENIC | LOSS OF FUNCTION | HOTSPOT | CIVIC | MY CANCER GENOME | LEVEL PROGNOSTIC | WITHOUT ANNOTATION |
| IFDel  | S45_V57del | S45_V57del | S45_V57del | S45_V57del | S45_V57del | S45_V57del | – | – |
| ERBB2  | Missense  | V842I; S310Y; S310F; G222R; E265K | V842I; S310Y; S310F; G222R; E265K | V842I; S310Y; S310F | V842I; S310Y | V842I; S310Y; G222R; E265K | L1098M |
| ERBB3  | Missense  | D297Y; T355I; G284R; V104L | D297Y; T355I; G284R; V104L | D297Y; T355I; G284R; V104L | V1035D; G284R; V104L | – | – | R444 Q; V586M; G994D |
| KMT2C  | Missense  | – | – | – | – | – | – | N1477D; R284Q; R909K |
|        | FS Del    | S4517Vfs*109 | S4517Vfs*109 | – | – | – | – |
|        | Nonsense  | Q2539*; E84*; E2496*; Q75* | Q2539*; E84*; E2496*; Q75* | – | – | – | – |
|        | Splice    | X923_splice | X923_splice | – | – | – | – |
| KMT2D  | Missense  | – | – | – | – | – | – | K5493E; E1159Q; P565L; L3470R; D5462H |
|        | FS Del    | Q809Rfs*121; Q791Rfs*139 | Q809Rfs*121; Q791Rfs*139 | – | – | – | – |
| IFDel  | – | – | – | – | – | – | R755_P763del |
|        | Nonsense  | Q3634*; E1171* | Q3634*; E1171* | – | – | – | – |
| KRAS   | Missense  | G12D; G13D; G12A; Q61H; L19F | G12D; G13D; G12A; Q61H; L19F | G12D; G13D; G12A; Q61H; L19F | G12D; G13D; G12A; Q61H | G12D; G13D; G12A; Q61H | G12D; G13D; G12A; Q61H; L19F | – |
|        | FS Del    | E545K; H1047R; R88Q; E81K; N345T | E545K; H1047R; R88Q; E81K; N345T | E545K; H1047R; R88Q; E81K | E545K; H1047R; R88Q; E81K | E545K; H1047R; R88Q; E81K | E545K; H1047R; R88Q; E81K | – |
| PIK3CA | Missense  | E330K; R361C; R361H; G386V; D355V; D351N; A118V; Y95H; E526Q; D493N | E330K; R361C; R361H; G386V; D355V; D351N; A118V; Y95H; E526Q; D493N | E330K; R361C; R361H; G386V; D355V; D351N; A118V; Y95H; E526Q; D493N | E330K; R361C; R361H; G386V; D355V; D351N | E330K; R361C; R361H; G386V; D355V; D351N | E330K; R361C; R361H; G386V; D355V; D351N | – |
|        | FS Del    | S144Wfs*7; L536Kfs*14; S144Wfs*7; L536Kfs*14; – | – | – | – | – | – |
| SMAD4  | Missense  | E388Cfs*28; L535Sfs*42; T73Nfs*31; H111Kfs*4 | E388Cfs*28; L535Sfs*42; T73Nfs*31; H111Kfs*4 | – | – | – | – |

(Continued)
### Table 2. (Continued)

| GENE | MUTATIONS | ANNOTATION | PROBABLY ONCOGENIC | LOSS OF FUNCTION | HOTSPOT | CIVIC | MY CANCER GENOME | LEVEL PROGNOSTIC | WITHOUT ANNOTATION |
|------|-----------|------------|--------------------|-----------------|---------|-------|------------------|------------------|--------------------|
| IFDel | L536_E538delinsQ | L536_E538delinsQ | L536_E538delinsQ | – | – | – | F329del | |
| Nonsense | R135*; S242*; Q116*; S154*; Q448*; K428* | R135*; S242*; Q116*; S154*; Q448*; K428* | – | – | – | – | – | – |
| TERT | Missense | – | – | – | – | – | D1126N | |
| Promoter | Promoter | Promoter | Promoter | – | – | – | – | – |
| TP53 | Missense | R273C; R248Q; R273H; R280T; E238K; Y163C; R248W; R175H; C275F; V157F; G245D; R156P; S241F; E271K; G245V; G266E; C242R; G245S; M133K; K132E; C141Y; P190L; N239D; C141W; I232T; H179L; V172F; C242S; Y234N; R282Q; V274D; E287K; V216G; M246K; S241C; R175H; P250L; H179R; S241Y | R273C; R248Q; R273H; R280T; E238K; Y163C; R248W; R175H; C275F; V157F; G245D; R156P; S241F; E271K; G245V; G266E; C242R; G245S; M133K; K132E; C141Y; P190L; N239D; C141W; I232T; H179L; V172F; C242S; Y234N; R282Q; V274D; E287K; V216G; M246K; S241C; R175H; P250L; H179R; S241Y | – | – | – | – |
| FSDel | R342Efs*3; Q52Lfs*68; K24Nfs*20; E343Gfs*2; P87ifs*54 | R342Efs*3; Q52Lfs*68; K24Nfs*20; E343Gfs*2; P87ifs*54 | – | – | – | – | – | – |
| FSIns | R267Tfs*5; Q167Hfs*4; L323Nfs*24 | R267Tfs*5; Q167Hfs*4; L323Nfs*24 | – | – | – | – | – | – |
| IFSdel | I255del; C242_248delinsW; T140_141del | I255del; C242_248delinsW; T140_141del | I255del; C242_248delinsW; T140_141del | – | – | – | – |
| Nonsense | E285*; R213*; R306*; Q192*; R196*; G266*; S183*; Y163*; S94*; C275*; W53*; W91* | E285*; R213*; R306*; Q192*; R196*; G266*; S183*; Y163*; S94*; C275*; W53*; W91* | – | – | R213* | – | – | – |
| Splice | X125_splice; X367_splice | X125_splice; X367_splice | X125_splice; X367_splice | – | – | – | – | – |
| ZNF521 | Missense | – | – | – | – | – | E942K; V574F; P465T; K599T | – | – |

Abbreviations: GBC, gallbladder cancer; FSDel, frameshift deletion; FSIns, frameshift insertion; IFSdel, in-frame deletion.

Data from CBioPortal (www.cbioportal.org, accessed on November 5, 2019).
Histone methyltransferases

KMT2D (Lysine methyltransferase 2D), a histone H3 lysine 4 (H3K4) methyltransferase, is mutated (missense, nonsense, silent, frameshift deletions and insertions, and in-frame deletions) in several cancers such as intestine, skin, and stomach tumors. In our analysis, this gene presented 10 mutations (7.4%), with truncating alterations the most frequent (missense [5 cases], nonsense [2 cases], frameshift deletion [2 cases], and in-frame deletion [1 case]). All the truncating (frameshift deletion and nonsense) mutations presented annotations suggesting clinical implications as probably oncogenic mutations which can lead to a loss of biological function. None of the missense and in-frame deletion mutations presented any associated annotation (see Table 2). This gene appeared most mutated in Chilean patient samples. In the other ethnicities, the values of mutations in kmt2d were lower. In addition, the kmt2d gene presented 2 missense mutations (K5493E and D5462H) inside its catalytic site, the SET (Su(var)3-9, E(z) and Trithorax) domain (see Table 2).

In other cancers and according to literature, the major rate of mutation seen in kmt2d in cancer is in lymphoma (30.44%), followed by bladder carcinoma (29.03%), non-small cell lung carcinoma (9.98%), colorectal carcinoma (5.83%), and pancreatic carcinoma (5.01%). The most common alterations in KMT2D present in other cancers are P2354Lfs*30 (0.14%), P647Hfs*283 (0.09%), KMT2D amplification (0.08%), and G1235Vfs*95 (0.08%).

Zinc finger protein 521

The znf521 gene was mutated only in Chinese samples of GBC, with 4 missense mutations (3%) (E942K, V574F, P465T, and K599T), without any annotations (Figure 1A and Table 2).

Discussion

The generation of a genetic alterations atlas is a useful tool to study the main cancer-causing alterations. Currently, it is feasible through repository public data-mining to detect the most mutated genes, in several types of cancers, which include tumor suppressors, proto-oncogenes and genes involved in DNA repair. These genes encode several kinds of proteins that help control cell growth and proliferation. It is widely known that mutations in these genes can contribute to the development of tumors. Here, we described in samples of GBC from the United States, Japan, China and Chile, that 14 genes presented significant values of somatic mutations (arid1a, arid2, atm, ctnnb1, erbb2, erbb3, kmt2d, kras, pik3ca, smad4, tert, tp53, and znf521). These genes are altered in 109 samples out of 133 cases (82%) in these databases. The literature reinforces the data-mining performed in our work, mainly for mutations in tp53, smad4, and arid1a. Several previous studies reported, through whole exome sequencing and whole genome sequencing approaches, that the most recurrently mutated genes in hepatobiliary pathways were tp53, cdkn2a/b, kras, arid1a, and idh1 in intrahepatic cholangiocarcinoma (ICC); kras, tp53, cdkn2a/b, and smad4 in extrahepatic cholangiocarcinoma (ECC); and tp53, cdkn2a/b, arid1a, and erbb2 in GBC. In this sense, Iyer et al reported 5060 somatic mutations in 17 tumors, 3239 missenses, 1449 silent, 131 nonsense, 135 indels, and 160 splice site mutations. The most altered genes in this study were tp53 (35.2%) followed for erbb2, tp53, atm, akap11, and ctnnb1. In our work, we described in addition of most common mutated genes of GBC, other mutated genes, such as arid2, kmt2c, and znf521, which also presented high importance in other cancers like hepatocellular, leukemia, urothelial, and oral squamous cell tumors. Furthermore, one of the most common features in several cancers are genetic amplifications and/or deletions, that is, CNAs that can activate oncogenes and inactivates tumor suppressors. In the work of Iyer et al, they identified CNA in GBC, where the loci cdk4, mdm4, cond1, cond3, cmyc, stk11, and brd13 harbor amplifications. Similarly, Lucio-Eterovic et al found that the histone methyltransferase nsd3 gene is over-amplified in about 15% of cases of breast cancer. In our work, several CNAs were found in 11 genes (cond1, cond3, cond5, cdk12, cdkn2a, cdkn2b, erbb2, erbb3, kras, mdm2, and mbc) in GBC samples. In our study, we found that in samples from Japan, the United States, and Chile, cond1, cond3, cdk12, erbb2, erbb3, kras, mdm2, and myc are amplified, with the genes mdk2 (13.2%), erbb3 (7.8%), and cond1 (7.8%) most altered in the database studied. The high amplification in mdk2, erbb3, and cond1 can be an indicative issue that these genes are involved in the process of carcinogenesis of the gallbladder, as this feature is common for other genes involved with the tumorigenesis process in many types of cancer. In addition, these sets of genes could be useful as possible genetic targets to study GBC in patients from the ethnicities studied.

Conversely, the most frequent alteration for cdkn2a and cdkn2b genes was deletion, observed in all ethnicities. In this respect, Iyer et al also identified that the loci fhit, smad3, trim33, and apc present deletions for GBC samples, reinforcing that the events of amplifications and deletions are present in several genes at GBC samples. The suppressor tumor cdkn2a encodes the p16 protein, a kinase-dependent cyclin inhibitor. Some studies suggest that the methylation and/or deletion in this gene might lead to gallbladder carcinogenesis. Roa et al found about 35% of inactivation of the cdkn2a gene in samples from Chile. This study found that, in male patients of Mapuche ethnicity, the inactivation of p16 was more frequent and associated with the worst prognosis. Besides this, previous studies reported mutations in the cdkn2a gene in ~50% of GBC cases suggesting the importance of cdkn2a in GBC, after evaluating its expression, deletion, and methylation events. However, when we evaluated some public databases of cancer, we found few studies have explored the mutations in these mentioned genes in samples of GBC.
Additional in-depth studies are needed to explore genetic targets like cdkn2a, which are involved in GBC cancer and in the carcinogenesis process, as it shares a common feature (dele-tion) with others genes drive-cancers.

Are tp53 and smad4 mutations prognosis markers?

In relation to the tp53 gene, more than half of GBC samples harbor 1 or more mutations. As we know, the function of the tp53 product is related to the control of cell cycle progression, cell death, and DNA repair; thus, it could participate in controlling the tumorigenesis process in GBC.\cite{41,42} Comparison of the frequency of tp53 mutations among Chilean, Peruvian,\cite{43} Bolivian,\cite{44} Hungarian,\cite{45} and Japanese\cite{46} cohorts resulted in no significant differences (\(P = .41\)).\cite{43} No mutations were found in exon 8 for Hungarian patients.\cite{45} When comparing data from Japanese, Hungarian, and Chilean samples, significant differences in frequency of mutations were found between exons 5 (35\%, 7/20) and 8 (5\%, 1/20) in the Chilean cohort.\cite{46} Based on literature, the rate of mutations in the tp53 gene from Chilean cases of GBC was around 55\%, and according to the database studied, the rate was 76.2\%. This high incidence is also seen in Japan and Hungary.\cite{45} The genetic changes of tp53 in gallbladder carcinoma seem to drive the carcinogenesis process in this type of tumor, based on its frequency.

Mutations R273C, R273H, R248Q, R248W, and R175H found in the tp53 gene in our data-mining may result in a gain-of-function, because, they are able to promote carcinogenesis in the murine model, according to the CBioPortal annotation for these genes. Patients harboring these mutations are more responsive to doxorubicin treatment. R248Q mutation is related to an increased invasive tendency in cell lines (CBioPortal, 2019).\cite{47} It is important to mention that, most mutations in tp53 occur in these amino acid positions (arginine positions 175, 248, and 273).\cite{48} They are well-known mutations in other cancers as well as GBC described here. In this respect, in the literature, missense mutations are the most common alterations in the tp53 gene. These are hotspot mutations for other types of tumors and are present in about 60\% of all samples analyzed. Mutations in the tp53 gene are an indicator of poor prognosis, metastasis, and poor conditions regarding patient survival.\cite{49,50}

Using the Kato et al\cite{51} data, it was possible to confirm the correlation between a loss of activity in TP53 and a high frequency of tp53 mutations in cancers. In most cases, tp53 is mutually exclusive with known driver genes.\cite{52} It is interesting to highlight that mutations in tp53 genes are required in several moments within process of GBC carcinogenesis, for example, the inactivation of arf/tp53 pathway along with c-myc overexpression could be initial requirements that lead Salmonella sp. to induce host cell transformation.\cite{19}

On the contrary, for GBC, the tumor progression results in dysplasia, in situ carcinoma, invasive adenocarcinoma, and finally, metastatic disease. Each phase, having specific mapped mutations. For example, normal tissue with hyperplasia (GBC initial stage) commonly has identified mutations in the tp53 gene. To identify the moment that specific mutations occur, during cancer progression, can bring new strategies of tumor management. In relation to that, Barreto et al\cite{53} provide a genetic model for GBC carcinogenesis, based on specific genetic alteration that occur during stages of tumor formation. This model can help to predict earlier patients with risk to develop GBC, making the decision to extract gallbladder organ before the disease appearance, as a prophylactic treatment.\cite{53} The first observed pathological alteration in the Barreto’s model is chronic inflammation, generating a hyperplasia. Our results corroborate this model’s findings, as one the most altered gene for GBC in a database studied was tp53. It can be indicative that mutations in this gene could play a role in initiation of the GBC process.\cite{53} Besides, in accordance to this, Iyer et al\cite{54} analyzed the effects of somatic alterations on survival of GBC patients. They found that poor survival rate was observed in patients with mutated tp53. These works reinforce the importance of describing the most common mutations in GBC and put them in a context of pathologic progression of tumors to propose their use as genetic markers for prognosis.

Another highly mutated gene in GBC was the smad4 gene, which encodes a protein subject to complex regulation by post-translational modifications.\cite{54} Mutations or deletions in this gene were described as linked to the GBC process. This gene plays an important role in the transforming growth factor-β pathway (TGF-β), which regulates diverse cellular processes, including proliferation, differentiation, apoptosis, and migration.\cite{54} In colorectal cancer, the presence of mutations in this gene can lead to a poor prognosis.\cite{55} In addition, smad4 is inactivated in approximately 55\% of pancreatic tumors and this is associated with poor prognosis and widespread disease.\cite{54} Based on literature, the SMAD4 protein has been suggested as a tumor suppressor in cholangiocarcinoma.\cite{57} Loss of smad4 activity was noted in 19 of 42 (45.2\%) cases with ICC, and had a positive correlation with clinical stages and prognosis.\cite{58,59}

Although 39\% of the cancers had genetic alterations in at least one of the TGF-β pathway genes, gastrointestinal (GI) cancers were particularly enriched with them. Gastrointestinal cancers were most influenced by recurrent hotspot mutations in 6 genes that encode TGF-β ligands (bmp5), receptors (tgfb2, avcr2a, bmpr2), and smad4 genes (smad2, smad4). In studies conducted by Ohshiro et al,\cite{60} they identified hotspot mutations in 6 genes, with increased expressions of tert, bmgα2, il6, mmp9, col1a1/1a2/3a1, myc, and foxp3. Alterations in these core genes are correlated positively with the expression of metastasis-associated genes, and poor patient survival.\cite{64} These data suggest that when combined with other specific genes, such as smad4 and tert, the TGF-β superfamily genes may represent strong prognostic markers and targets in some cancer types, such as GBC.

Therefore, the importance is clear for studying the mutations in the genes indicated above and their implications in GBC, mainly for tp53 and smad4 genes, which are described as
being linked to poor prognosis and to the carcinogenesis process. Many mentioned mutations present several annotations because they are recurrent in other types of tumors, now including GBC. This information is valuable and can provide a good tool to propose new tests for prognosis and evaluation of GBC progress in patients.

The arid1a: a new tumor-suppressor gene in gallbladder cancer?

One face of the genetic alterations that occur in cancer cells is disordered chromatin organization and truncating mutations, which were observed in the arid1a (AT-rich interaction domain 1A) gene. In vitro studies demonstrated that the depletion of arid1a can increase colony formation and cell proliferation, as well as decrease the apoptosis. These studies showed that that heterozygosity of arid1a could lead to embryonic lethality in the murine model.\textsuperscript{61-63} We found in GBC that the arid1a gene is mutated in 20.5% of the GBC samples and almost all mutations described were truncating (around 90%), of the type nonsense and splicing mutations. The truncating mutation in arid1a in GBC patients seems to be a common event in other cancers too, such as ovarian tumors.

In addition, this gene is the most frequently mutated gene, also in human colorectal cancer, where it is mutated in 10.9% of cases (TCGA PanCancer Atlas dataset, CBioPortal for Cancer Genomics).\textsuperscript{64} In colorectal cancer, it was shown that the activity of BAF-occupied enhancers is reduced in arid1a-deficient cells and it is accompanied by a loss of the active H3K27ac mark (acetylation of lysine 27 on histone 3).\textsuperscript{65} This is a gene involved in several biological processes including replication, DNA repair, and controlling cell division. The arid1a is the most frequently mutated member of the SWItch/Sucrose Non Fermentable (SWI/SNF) family and it has a high incidence of inactivating mutations in several cancer types; thus emerging functional studies consider arid1a to be a novel tumor suppressor.\textsuperscript{66}

Chilean cohort: a study of case

As GBC is the most frequent cause of cancer death in Chilean women and the third-highest cause of tumor death among Chilean men, in contrast to other countries worldwide, we believe it is very important to study the cause and genes related to the prevalence, which is still unclear; however, the high rates of obesity and presence of previous gallstones as well genetic susceptibility could explain, at least in part, the high risk in this population.\textsuperscript{67,68}

Our computational analysis showed that the frequency with which this type of cancer occurs was higher in women (60.2%) than in men (39.8%), in all places analyzed. This result is consistent with the worldwide literature, which states that the prevalence rate of GBC in women can reach 5:1 regarding men (10; 16; 17). A rate of 4:1 in Chilean women, which had already been described in previous studies,\textsuperscript{7} was also observed in the repository we used.\textsuperscript{7}

Regarding genetic alteration in Chilean patients with GBC, previous studies performed the comparison among Chilean, Chinese, and US cohorts and showed that genetic aberrations in DNA repair pathway were the most frequent alterations, in particular in the atm gene (25.2% vs 8.3% and 1.9%, \textit{P} = 0.3). A low frequency of variation was observed in the Chilean cohort (10.7%), compared with the Chinese (37%) and US (33%) cohorts.\textsuperscript{7}

In our analysis, we found that all mutations in genes \textit{tp53} (R280T, G245S, R273H, R175H, R248W, and R273C), \textit{pik3ca} (E545K, E542K, and H1047R), \textit{atm} (L2953TFS*3, E343*, X301_SPLICE), and \textit{kras} (G12A, G12D) in Chilean patients are annotated as putative cancer drivers, according to CBioPortal information. It is interesting to mention that these mutations described above for \textit{tp53}, \textit{pik3ca}, \textit{atm}, and \textit{kras} genes present a complete annotation in databases: function, biological importance, whether are hotspot mutation, and whether they are associated with prognosis (Figure 2). For \textit{tp53}, the mutations found in the Chilean cohort are also common in other cancers, resulting in a gain-of-function. They are able to promote tumorigenesis in murine models and probably lead the patient to be more responsive to treatment with doxorubicin. Generally, these mutations are associated with a worse prognosis (CBioPortal, 2019).

The mutations found in Chilean GBC patients for \textit{pik3ca} are recurrent mutations found in many cancers. E542K and E545K mutations gene can confer resistance to epidermal growth factor receptor (EGFR) inhibitors, like cetuximab. It is already known that the common mutation H1047R is generally associated to a poor prognosis in other cancers, according to the CBioPortal annotation. \textit{ATM} mutation described in the Chilean GBC cohort, in lymphoma for example, can lead to a predisposition to develop this tumor and acute leukemia, when the \textit{atm} gene is mutated in germ-lines.\textsuperscript{70} Finally, for \textit{kras}, we identified important mutations in Chilean patients, which can lead the patients to be resistant to tyrosine kinase inhibitors (CBioPortal, 2019).

In addition, in our analysis with the Chilean GBC cohort, it is important to highlight mutations in the \textit{kmt2d} gene. The frequency of mutations in this gene have been seen only in the Chilean cohort (Figure 2). In our analysis, 4 mutations (2 frameshift deletion and 2 nonsense) have annotations as probable oncogenic and could lead to a loss of biological function (Q809RFS*121, Q291RFS*139, Q363*, and E1171*) in the KMT2D protein.\textsuperscript{71} The mutations D5462H and K5493E are inside the SET domain, the active domain of the codified protein. These mentioned mutations could have a role in the loss of function of this gene as the active domain is an important site to normal function of the codified protein. In addition, the silencing of \textit{kmt2d} in bladder cancer can lead to a significant increase of the cell viability, migration, and invasion acting as an anti-tumor factor, suggesting its role in the carcinogenesis process through the enhanced H3K4me1 activity.\textsuperscript{69}
Figure 2. (Continued)
In spite of our results, it is important to highlight that our study has some limitations. The study was not presented with a big cohort to analyze; we used only 133 samples from 4 countries: Chile, China, Japan, and the United States. In addition, 20 samples did not have information regarding the country of origin. Finally, some clinical data such as the primary site of tumors, perineural invasion, and metastatic sites were not known for some countries. Therefore, it was not possible to analyze the relationship between mutations and the clinical outcomes. However, our work brings a new descriptive approach regarding the annotation of the GBC mutations and their putative importance in the type of cancer studied.

**Conclusions**

The genetic basis of the development of GBC is still scarce. Thus, it is necessary to generate more knowledge regarding the most important mutations and genetic alterations in this tumor, to propose more effective diagnosis and new molecular markers of predisposition and prognosis, especially in Chile, where there is a particularly high prevalence of this disease.

Here, we performed data-mining on a dataset of GBC to encourage the scientific community to propose specific analysis to control GBC with the development of new diagnostic tests and the use of new prognosis makers. We were able to describe the most mutated genes in GBC, which are *arid1a,*
Finally, we found 11 genes with CNAs in the public repository used, which are cond1, cond3, cond1e, dkk12, dkk2a, dkk2b, erb2b, erb3b, kras, mdm2, and myc. In samples from Japan, the United States, and Chile, cond1, cond1e, dkk12, erb2b, erb3b, kras, mdm2, and myc are amplified. Conversely, the more frequent genetic event for dkk2a and dkk2b was deletion, in all ethnicities, suggesting that the lack of function of these genes could influence the GBC carcinogenesis process.

According to the above indicated, we are able to describe an atlas containing several genetic alterations with their respective annotations for GBC samples, which can be useful to establish new molecular markers to control GBC.

Acknowledgements
The authors want to thank Universidad Católica del Maule through the Centro de Biotecnología de los Recursos Naturales (CENBIO) for the support to perform this work and the Sociedad Latinoamericana de Farmacogenómica y Medicina Personalizada (SOLFAGEM).

Author Contributions
VD’A was responsible for the conception of the research, organization of the database, analysis of data, and writing the manuscript. ADA, AE-V, LC, and JPC organized the database. NMV contributed to the analysis of data and writing the manuscript. MS and LAQ contributed to the conception of the research, analysis of data, and writing the manuscript.

ORCID iDs
Alex Echeverría-Vega https://orcid.org/0000-0002-1110-1079
Nelson M Varela https://orcid.org/0000-0002-5229-3007

REFERENCES
1. Ferlay JI, Soerjomataram I, Siegel RL, et al. Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN. Int J Cancer. 2009;126:3597-3607. doi:10.1002/ijc.24695.
2. Hemant KB. Unravelling molecular genetics of gallbladder cancer: implications for targeted therapy. Curr Opin Oncol. 2017;10:6-11. doi:10.1093/cco/cco214.
3. Hussain K, Miki K, Kagawa Y. Lecithinosomes are associated with gallbladder carcinoma: results of a Japanese survey of 498 patients. J Hepatobiliary Pancreat Surg. 2002;9:256-260. doi:10.1005/jhps.2002.203102.
