Programa de Pós-graduação Interunidades em Bioinformática

Universidade de São Paulo

**Modelagem e implementação de banco de dados clínicos e moleculares de
pacientes com câncer e seu uso para identificação de marcadores em câncer de
pâncreas**

Documento apresentado ao Instituto de
Matemática e Estatística como requisito
para obtenção do título de mestre em
bioinformática

Aluna: Ester Riserio Matos Bertoldi

Orientador: Eduardo Moraes Rego Reis – IQ-USP

Co-orientador: João Eduardo Ferreira – IME-USP

Março de 2018

“Nada na vida deve ser temido, apenas compreendido.
Agora é a hora de entender mais, para que possamos temer menos.”
Maria Skłodowska-Curie

Resumo

O adenocarcinoma pancreático (PDAC) é uma neoplasia de difícil diagnóstico precoce e cujo tratamento não tem apresentado avanços expressivos desde a última década. As tecnologias de sequenciamento de nova geração (“*next generation sequencing*” - NGS) podem trazer importantes avanços para a busca de novos marcadores para diagnóstico de PDACs, podendo também contribuir para o desenvolvimento de terapias individualizadas. Bancos de dados são ferramentas poderosas para integração, padronização e armazenamento de grandes volumes de informação. O objetivo do presente estudo foi modelar e implementar um banco de dados relacional (CaRDIGAn - “*Cancer Relational Database for Integration and Genomic Analysis*”) que integra dados disponíveis publicamente, provenientes de experimentos de NGS de amostras de diferentes tipos histopatológicos de PDAC, com dados gerados por nosso grupo no IQ-USP, facilitando a comparação entre os mesmos. A funcionalidade do CaRDIGAn foi demonstrada através da recuperação de dados clínicos e dados de expressão gênica de pacientes a partir de listas de genes candidatos, associados com mutação no oncogene KRAS ou diferencialmente expressos em tumores identificados em dados de RNAseq gerados em nosso grupo. Os dados recuperados foram utilizados para a análise de curvas de sobrevida que resultou na identificação de 11 genes com potencial prognóstico no câncer de pâncreas, ilustrando o potencial da ferramenta para facilitar a análise, organização e priorização de novos alvos biomarcadores para o diagnóstico molecular do PDAC.

Abstract

Pancreatic Ductal Adenocarcinoma (PDAC) is a type of cancer difficult to diagnose early on and treatment has not improved over the last decade. Next Generation Sequencing (NGS) technology may contribute to discover new biomarkers, develop diagnose strategies and personalised therapy applications. Databases are powerfull tools for data integration, normalization and storage of large data volumes. The main objective of this study was the design and implementation of a relational database to integrate publicly available data of NGS experiments of PDAC patients with data generated in by our group at IQ-USP, allowing comparisson between both data sources. The database was called CaRDIGAn (“Cancer Relational Database for Integration and Genomic Analysis”), its functionalities were tested by retrieving clinical, and expression data of public data of genes differentially expressed genes in our samples or genes associated with KRAS mutation. The output of those queries were used to fit survival curves of patients, which led to the identification of 11 genes potentially useful for PDAC prognosis. Thus, CaRDIGAn is a tool for data storage and analysis, with promising applications to identification and prioritization of new biomarkers for molecular diagnosis in PDAC.

Sumário

Lista de Abreviaturas	01
Introdução	03
Objetivos	12
Material e Métodos	13
Resultados	18
Discussão	48
Conclusões	55
Referencias	56
Anexos	61

Lista de Abreviaturas

API	<i>Application Programming Interface</i> , interface de programação de aplicativos
CNV	<i>copy number variation</i> , alteração do número de cópias de uma dada sequência de DNA
CaRDIGAn	<i>Cancer Relational Database for Integration and Genomic Analysis</i> , Banco de dados relacional para integração e análises genômicas, banco de dados que foi modelado e implementado no presente estudo
COSMIC	<i>Catalogue of Somatic Mutations in Cancer</i> , banco de dados públicos curado sobre mutações em câncer
DDBJ	<i>DNA Data Bank of Japan</i> , banco de dados que armazena, entre outros, dados primários de sequências nucleotídicas
EMBL	<i>European Molecular Biology Laboratory</i> , banco de dados que armazena, entre outros, dados primários de sequências nucleotídicas
ETL	<i>Extraction, Formating and Loading</i> , ferramenta que permite extrair dados, transformá-los e, finalmente, carregá-los para um banco de dados estruturado, conforme o desenho experimental
GenBank	Banco de dados primários de sequências nucleotídicas
HGNC	<i>HUGO Gene Nomenclature Committee</i> , comitê que estabelece padrões para nomenclatura de genes
ICGC	<i>International Cancer Genome Consortium</i> , iniciativa internacional para gerar, analisar e disponibilizar clínicos e experimentais de diversos tipos de câncer
LGEGC	Laboratório de Genômica e Expressão Gênica em Câncer, laboratório do Instituto de Química da Universidade de São Paulo, onde o presente estudo foi desenvolvido
NCBI	<i>National Center for Biotechnology Information</i> , divisão do NIH que mantém o GenBank
NGS	<i>Next Generation Sequencing</i> , sequenciamento de nova geração, método para sequenciamento de ácidos nucleicos
NIH	<i>National Institute of Health</i> agência de saúde governamental americana
PDAC	<i>Pancreatic Ductal Adenocarcinoma</i> , Adenocarcinoma ductal de pâncreas, tumor mais comum do pâncreas
SGBD	Sistemas Gerenciadores de Banco de Dados, utilizados para implementação e gerenciamento de bancos de dados
SQL	<i>Structured Query Language</i> , Linguagem de Consulta Estruturada, linguagem de pesquisa declarativa padrão para banco de dados relacional.

SRA *Sequence Read Archive*, repositório público para o armazenamento de conjuntos de dados de sequências nucleotídicas

TCGA *The Cancer Genome Atlas*, programa do NIH que busca gerar, analisar e disponibilizar clínicos e experimentais de diversos tipos de câncer

Introdução

O Adenocarcinoma Ductal do Pâncreas

O câncer de pâncreas representa apenas 2% dos casos de câncer diagnosticados no país (“INCA - CÂNCER - Tipo - Pâncreas” 2017). Contudo, devido à dificuldade de diagnóstico e tratamento, esta neoplasia representa a quarta maior causa de morte por câncer no ocidente e apenas 6% dos pacientes sobrevivem pelo menos cinco anos após o diagnóstico (Schlitter et al. 2017). O tipo mais frequente de câncer de pâncreas é o adenocarcinoma ductal (PDAC, *pancreatic ductal adenocarcinoma*), que corresponde a 80% dos casos e acomete a região exócrina do órgão (Maitra e Hruban 2008). A única opção terapêutica curativa eficiente é a ressecção cirúrgica ainda em estágios iniciais, que é possível em menos de 20% dos casos (Hidalgo 2010).

Ainda que importantes progressos no conhecimento da biologia do câncer de pâncreas tenham sido feitos nos últimos anos, a prevenção, diagnóstico prematuro e tratamento do PDAC ainda representam grandes desafios. Sabe-se que lesões precursoras podem evoluir para PDAC, existindo, assim, a necessidade de biomarcadores que auxiliem no diagnóstico definitivo e prognóstico da doença em pacientes nos quais se detectam essas lesões (Matthaei et al. 2011). As proteínas CA19-9 (*carbohydrate antigen 19-9*) e CEA (*carcinoembryonic antigen*) são detectadas no sangue de indivíduos com câncer de pâncreas e foram os biomarcadores mais extensivamente avaliados para o diagnóstico da doença (Bai et al. 2004). Embora tenham se revelado úteis na clínica como indicadores de recidiva/progressão da doença, tanto após cirurgia quanto durante a quimioterapia adjuvante (Oettle et al. 2007), a baixa especificidade e sensibilidade desses marcadores prejudicam seu uso para a detecção precoce da doença (Duffy et al. 2010). Neste cenário, há grande interesse na identificação de novos marcadores moleculares para diagnóstico precoce que possam ser detectados através de métodos pouco invasivos. Nessa direção, um estudo preliminar identificou mRNAs presentes na saliva que possuem potencial como biomarcadores para identificar a presença do câncer em estágios iniciais e pré-invasivo (L. Zhang et al. 2010), mas o valor diagnóstico desse achado não foi confirmado até o momento.

Em relação à biologia molecular deste câncer, o acúmulo de mutações em genes específicos ao longo da progressão da doença já é bem conhecido, como por exemplo: a) ativação do oncogene *KRAS* em 95% dos casos (controlando uma ampla gama de funções celulares como proliferação, sobrevivência celular e remodelamento do citoesqueleto); b) inativação do gene supressor de tumor *CDKN2A* em 90% dos casos (fazendo com que não ocorra mais checagem do ciclo celular em G1-S); c) inativação do gene supressor de tumor *TP53* de 50 a 75% dos casos (permitindo com que a

célula sobreviva e se divida apesar de erros no DNA e do acúmulo de mutações e anormalidades); e d) inativação do gene *SMAD4*, também conhecido como *DPC4*, em 55% dos casos (com efeitos no crescimento e diferenciação celular) (Iacobuzio-Donahue 2012). A somatória destas e outras mutações foram associadas à progressão da doença de estágios baixos até terceiro e quarto grau, quando metástases são produzidas no duodeno, ducto biliar, tecidos peripancreáticos, estômago, baço, cólon e grandes vasos adjacentes (Aaltonen e Hamilton 2000). Por outro lado, estudos mais recentes utilizando abordagens genômicas têm revelado que os tumores pancreáticos se caracterizam por uma grande heterogeneidade e diversidade de alterações somáticas, com baixa ou média penetrância, que afetam um amplo número de oncogenes e genes supressores de tumor (Jones et al. 2008; Biankin et al. 2012). É plausível, portanto, que o desenvolvimento de novas estratégias terapêuticas envolva o mapeamento com alta resolução de mutações prevalentes em indivíduos ou subgrupos de pacientes.

As ômicas e sua contribuição para o conhecimento da biologia molecular dos PDAC

A primeira geração de sequenciadores de DNA foi baseada no método dideoxi de Sanger. A automatização do processo de sequenciamento de DNA/cDNA abriu a possibilidade de caracterização de genomas e transcritomas completos, inaugurando as áreas chamadas -ômicas (ou, em inglês, -omics), como genômica e transcriptômica, e motivando o desenvolvimento de tecnologias para análise em larga-escala de outras moléculas biológicas (ex. proteômica, metabolômica) (Choi e Pavelka 2012). Contudo, o método de Sanger de sequenciamento de DNA, desenvolvido em 1977, é hoje considerado lento, custoso e trabalhoso quando comparado às tecnologias mais recentes. As técnicas de sequenciamento de nova geração (NGS – *Next Generation Sequencing*) têm sido aplicadas principalmente para: a) resequenciamento do genoma para descoberta de mutações e polimorfismos; b) mapeamento de rearranjos estruturais, incluindo variação de número de cópias, pontos de quebra de translocações balanceadas e inversões cromossômicas; c) sequenciamento de RNA, ou *RNAseq*, que gera bibliotecas derivadas de RNA total ou frações selecionadas de RNAs (ex. RNA poliadenilado ou não, RNAs pequenos ou longos) que são sequenciadas em profundidade, permitindo uma análise quantitativa e qualitativa da expressão destes produtos (ex. expressão diferencial de genes e formas alternativas de *splicing*, uso alternativo de promotores e sítios de poliadenilação); d) análise em larga-escala de metilação de DNA, através do sequenciamento em profundidade de DNA tratado com bissulfito; e e) *ChIP-Seq*, ou mapeamento genômico completo de interações entre DNA e proteínas, realizado a partir do sequenciamento em profundidade de fragmentos de DNA obtidos a partir de imunoprecipitação de cromatina (Shendure e Ji 2008). A análise global de marcas

epigenéticas, como metilação do DNA e modificações químicas de histonas, e de ligação de proteínas regulatórias permite inferir sobre estados de ativação/repressão da cromatina.

A aplicação das técnicas de NGS tem revelado novas informações acerca da biologia dos tumores pancreáticos. O sequenciamento com alta-resolução da região codificadora do genoma (exoma) de amostras de PDAC revelou uma enorme heterogeneidade de mutações somáticas que afetam diversos genes e vias moleculares (Iacobuzio-Donahue 2012). No estudo de Jones et al, dos 20.661 genes codificantes analisados, 1.562 mutações somáticas foram detectadas, a maioria sendo substituição de bases (2008). Outro trabalho de sequenciamento do exoma de adenocarcinomas de pâncreas confirmou o alto grau de heterogeneidade de mutações somáticas entre tumores de diferentes indivíduos, e apontou novas vias moleculares ainda não descritas como afetadas nesta doença, como genes envolvidos na orientação de axônios (Biankin et al. 2012).

Trabalhos colaborativos entre instituições de diversos países vêm tentando classificar os PDACs em subtipos, baseando-se em análises de NGS, obtendo resultados interessantes. O estudo de Waddell e colaboradores (2015) analisaram 100 amostras de tumores primários de PDAC. Cada amostra passou por sequenciamento de genoma completo e análise de variação de número de cópias. Baseados nas alterações cromossômicas detectadas, os autores encontraram quatro padrões com potencial uso na clínica: estável, localmente rearranjado, difuso e instável. Foram também encontradas amplificações em oncogenes, embora pouco prevalentes nas amostras. São eles *ERBB2*, *MET*, *FGFR1*, *CDK6*, *PIK3R3* e *PIK3CA*. Curiosamente, a instabilidade genômica estava relacionada com a inativação de *BRCA1*, *BRCA2* ou *PALB2*, genes importantes para a manutenção celular.

Em um grande estudo recente, envolvendo 44 instituições de cinco países, foi avaliado o perfil de mutações somáticas e expressão gênica em tumores de pâncreas (Bailey et al. 2016). Foram incluídas 342 amostras de tumores primários e 41 de linhagens celulares primárias derivadas de pacientes. As amostras passaram por sequenciamento genômico com alta profundidade, com cobertura média de 75x para amostras com celularidade >40%. Amostras com baixa celularidade (12-40%) passaram por sequenciamento de exoma com cobertura média de 400x. Os resultados de RNA-seq passaram por análise bioinformática utilizando um método de classificação não supervisionada que permitiu a identificação de quatro subtipos de tumor, classificados como “Escamoso”, “Pancreático Progenitor”, “Imunogênico” e “ADEX” (Anormalmente Diferenciado Endócrino Exócrino), que tem dois subtipos, o “Endócrino” e o “Exócrino”. O enriquecimento gênico em cada um dos subtipos mostrou alteração em genes específicos para cada subtipo. A saber: Escamoso: *TP53*, *KDM6A* e *TP63*; Pancreático Progenitor: *FOXA2*, *PDX1* e *NMX1*; ADEX: *KRAS* e, finalmente,

ADEX Exócrino: *NR5A2* e *RBPJL* e ADEX Endócrino: *NEUROD1* e *NKX2-2*. Esses resultados levaram os pesquisadores a inferir que há diferença na evolução molecular dos subtipos de tumores de pâncreas e abre possibilidades para desenvolvimento de novas terapias. Este trabalho fez parte do ICGC, que será discutido em maior detalhe adiante.

Desafios e estratégias para a análise de dados de NGS

As técnicas NGS geram uma enorme quantidade de dados de sequência e evoluíram de forma que o principal fator limitante para o conhecimento deixou de ser a obtenção dos dados, mas sua organização e análise (Liu et al. 2012). A bioinformática tem sido crucial em todos os aspectos das ciências ômicas. As tecnologias atuais de NGS mais comumente utilizadas estão ainda limitadas ao sequenciamento de sequências com no máximo 500 nucleotídeos, tipicamente variando de 35 a 100 nucleotídeos. Considerando que um experimento de NGS pode gerar bilhões de pequenas sequências (*short reads*), um dos desafios na área é a reconstrução de genomas ou transcritomas por meio do alinhamento das sequências em sequências referência ou através de reconstrução baseada nas próprias sequências obtidas (montagem *de novo*). Os dados brutos e as saídas dos programas de alinhamento e montagem geram arquivos da ordem de *giga* ou mesmo *terabytes*, assim outro desafio é o desenvolvimento de estratégias de armazenamento e administração de grandes quantidades de dados, sendo crítico o desenvolvimento de ferramentas que facilitem a análise da informação gerada. Além do tamanho, outra dificuldade é a heterogeneidade observada nos formatos dos arquivos e das plataformas de sequenciamento utilizadas (Jun Zhang et al. 2011). Certamente, outro desafio da área é o que se chama de ruído. Observa-se grande variação nos resultados entre diferentes corridas, amostras, plataformas e laboratórios, implicando menor especificidade e sensibilidade, o que significa grande ocorrência de resultados falso-positivos e falso-negativos (Fortney e Jurisica 2011).

O campo da biologia computacional integrativa utiliza técnicas de ciência da computação, matemática, física e engenharia para analisar e interpretar dados biológicos. A criação de novos métodos de análise e visualização de dados, além do desenvolvimento de novas ferramentas e bancos de dados, possibilita abordar e diminuir as dificuldades derivadas dos desafios mencionados anteriormente. Essas aplicações podem ser classificadas em três categorias: a) padronização; b) integração de dados, e c) bancos de dados (Fortney e Jurisica 2011).

A padronização na maneira de se produzir, administrar, analisar e publicar os dados obtidos por técnicas de NGS facilita o compartilhamento dos mesmos, auxilia na redução de ruído e assegura

a reprodutibilidade tanto a nível experimental quanto computacional. A integração de dados auxilia a diminuição do ruído observado nas aplicações NGS, pois um resultado se torna mais confiável uma vez que tenha sido observado em outros experimentos similares. A integração de dados pode ser feita tanto entre dados obtidos em plataformas diferentes, como é comum no caso dos *microarrays*, quanto dados de tipos diferentes, como a combinação de análises de expressão, mutação e proteômica. Assim, a integração dos dados aumenta a confiança nos resultados que são consistentes entre diversos estudos. Bancos de dados integrados e atualizados, bem como os recursos a eles associados, auxiliam análises abrangentes e, simultaneamente, gerenciam o grande volume de dados existentes e gerados. Entre as vantagens do uso de bancos de dados, pode-se mencionar a organização de dados heterogêneos e a facilidade de reinterpretação de dados na eventualidade de novas descobertas ou métodos de análise (Fortney e Jurisica 2011).

Bancos de dados biológicos

Intuitivamente, poder-se-ia supor que uma boa forma de armazenar dados seria colocá-los em uma tabela única. Porém, esse tipo de organização traz diversos problemas, como alta redundância e muitas células em branco, além da dificuldade de agrupar os dados ou selecioná-los de maneira arbitrária. Dados experimentais e informações disponíveis em bancos de dados públicos podem ser armazenados em bancos de dados, o que facilita a busca e a análise, agilizando sua aplicação. O poder dos bancos de dados relacionais vem da sua capacidade de organizar fragmentos da informação de maneira não redundante, permitindo buscas que combinam os dados de diversas maneiras (Peitzsch 2003).

Os bancos de dados são implementados e geridos com auxílio de programas denominados genericamente de Sistemas Gerenciadores de Banco de Dados (SGBDs). Esses programas utilizam diferentes formas de representação de dados, também denominados modelos de bancos de dados. Os SGBDs existentes são implementados sob o modelo hierárquico, em rede, relacional, orientado a objetos ou objeto-relacional. Embora seja possível implementar bancos de dados biológicos baseados no modelo de dados orientado a objetos, esse tipo de banco segue, em sua maioria, o modelo relacional. O modelo relacional é baseado em relações que são interligadas entre si por meio de um ou mais relacionamentos, sendo especialmente úteis em situações nas quais se observa uma grande quantidade de elementos com atributos semelhantes, como é o caso, por exemplo, de dados de *microarray*. Nos bancos de dados orientados a objetos os registros são definidos com base em relações de herança de atributos, no qual uma hierarquia organiza, de maneira específica, o acesso a

um dado objeto. Embora seja mais intuitiva do ponto de vista de programação, esse tipo de banco de dado é mais frequentemente empregado em áreas comerciais específicas, na qual uma representação de dados mais complexa é necessária. Uma aplicação interessante de bancos orientados a objeto na área biológica é a anotação de informações de proteínas, incluindo sua estrutura, modificações pós-traducionais e interações proteína-proteína (Navarro et al. 2003).

Como já foi mencionado, a integração de dados entre os diversos bancos é de grande interesse para a comunidade científica. Nesse sentido, um projeto merece destaque. O BioMart (Kasprzyk 2011) foi desenhado para atender às necessidades da área biológica considerando o grande volume dos dados existente e o fato de alguns conceitos biológicos ainda não estarem bem definidos e, ainda, que podem mudar com frequência. De maneira geral, essas questões são abordadas a partir de duas diretrizes: modelo agnóstico de dados e federação de dados. O modelo agnóstico é um esquema relacional de dados pré-definido e otimizado para consultas que pode ser usado independente da natureza do dado. Já a federação de dados significa, de maneira sucinta, que dados de diferentes bancos de dados podem ser integrados e consultados por meio de uma interface única, aparentando ao usuário se tratar de uma fonte única (Kasprzyk 2011). O acesso pode ser feito via *website* ou APIs (*Application Programming Interface*, interface de programação de aplicativos). Em linhas gerais, um API é uma ferramenta para utilização de um *software* sem a necessidade de familiarização com a implementação do programa original. A plataforma BioMart tem APIs disponíveis para diversas linguagens de programação, incluindo Python, chamado BioServices (Cokelaer et al. 2013), e R, chamado Bioconductor (Huber et al. 2015). Assim, a plataforma BioMart é uma ferramenta de importância crescente para a comunidade científica.

O Chado é um esquema de banco de dados relacional utilizado para armazenar dados genômicos de uma variedade de organismos. Seu desenvolvimento começou com dados de *Drosophila* e hoje é aplicado para cerca de 20 organismos. Seu desenho prevê a representação de dados utilizando módulos, que armazenam dados relacionados a sequência, como gene, transcrito, exon, pseudogene etc, dados de filogenia, taxonomia, publicações, entre outros. O uso de ontologias é uma característica chave para a flexibilidade do modelo, que foi idealizado para permitir expansão (Mungall, Emmert, e FlyBase Consortium 2007). Alguns bancos de dados que utilizam o modelo Chado são o FlyBase (Marygold et al. 2016) e o ParameciumDB (Arnaiz e Sperling 2011).

Existem diversos bancos de dados biológicos disponíveis para consulta, que podem armazenar, de maneira geral, a) sequências, tanto de nucleotídeos quanto aminoácidos; b) estruturas proteicas, ou c) vias moleculares e redes gênicas (Toomula, et al. 2011).

Os bancos de dados biológicos podem ser divididos entre bancos primários e bancos secundários. Os bancos primários armazenam diretamente resultados obtidos experimentalmente. As maiores fontes de dados primários de sequências nucleotídicas são o DDBJ, EMBL e o GenBank, que compartilham dados diariamente. A troca de informação é facilitada pelo uso de formatos de arquivos similares (Baxeavanis e Ouellette, 2005). O GenBank é mantido pelo *National Center for Biotechnology Information* (NCBI), parte do *National Institute of Health* (NIH) uma agência governamental americana (Benson et al. 2015). De maneira similar, o DDBJ (*DNA Data Bank of Japan*) é mantido pelo governo japonês (Kodama et al. 2015). O EMBL (*European Molecular Biology Laboratory*) é mantido pelo consórcio europeu EBI (*European Bioinformatics Institute*) (Kulikova et al. 2007). O Ensembl é o banco de dados mantido pelo EBI em conjunto com o *Wellcome Trust Sanger Institute* e é uma referência em dados genômicos (Aken et al. 2016). Esses bancos armazenam genomas de diversos organismos. Atualmente, é possível acessar genomas completos de 123 animais, 50 protozoários, 102 plantas e 442 fungos através do NCBI (“Genome List” 2017). A importância desse tipo de repositório de informações pode ser comprovada pela sua própria expansão ao longo dos anos. Em dezembro de 1982, havia 606 sequências disponíveis no GenBank, enquanto atualmente há 201.663.568 (“GenBank and WGS Statistics” 2017). O advento e disseminação das tecnologias de NGS fomentaram o surgimento de repositórios públicos para o armazenamento de grandes conjuntos de dados de sequências, como o SRA (*Sequence Read Archive*) mantido pelo NCBI (Leinonen, Sugawara e Shumway 2011).

Os bancos secundários armazenam os produtos das análises realizadas com os dados provenientes de bancos primários. O *Expression Atlas*, também mantido pelo EBI, é um banco com informações sobre expressão genica e proteica de diversos organismos em condições biológicas variadas como diferentes tecidos, estágios do desenvolvimento e doença (“Expression Atlas – EMBL-EBI” 2017). Neste caso, o banco usa dados de outros projetos para gerar *heatmaps* mostrando a expressão de um dado gene em diversas condições biológicas, agrupamento hierárquico de genes mais variáveis em diferentes condições experimentais e, para um dado gene em uma dada condição, a distribuição da expressão basal entre diversas réplicas biológicas.

Ao longo dos anos, foram criadas iniciativas voltadas especificamente para o estudo do câncer. Essas iniciativas geraram e analisaram dados de diversos tipos de tumores, disponibilizando os resultados de suas análises. O *International Cancer Genome Consortium* (ICGC) (Zhang et al. 2011) é uma iniciativa para unir projetos com objetivo comum de elucidar mudanças nos genomas em diferentes tipos de câncer que desempenhem papéis relevantes. A estratégia adotada é, primeiramente, gerar catálogos com mutações, variação na expressão genica e modificações epigenéticas em tumores

de 50 tipos e/ou subtipos, entre eles o câncer de pâncreas. O *The Cancer Genome Atlas* (TCGA) (“The Cancer Genome Atlas Home Page” 2015) é um programa do NIH que tem objetivos semelhantes aos do ICGC, sendo estudados 20 tipos de câncer, incluindo câncer de pâncreas. Atualmente, seus dados encontram-se também no ICGC. Enquanto os ICGC e o TCGA são consórcios especializados em gerar e analisar dados, o projeto *Catalogue of Somatic Mutations in Cancer* (COSMIC) (Forbes et al. 2008) dedica-se a manter um banco de dados que extrai informações de trabalhos disponíveis na literatura, armazenando-as de maneira curada. Esse banco foi criado e é mantido pelo *Sanger Institute*, no Reino Unido, e seu objetivo é ser um depósito de dados de mutações, tanto estruturais quanto de poucos nucleotídeos, de diversos tipos de tumores, incluindo também dados de algumas síndromes nas quais se observa crescimento descontrolado da população de células de um determinado tecido.

Iniciativa para integração de dados públicos de câncer

O BioXpress (Wan et al. 2015) é um projeto que utiliza dados do TCGA, ICGC, *Expression Atlas* e de 536 publicações. Seu foco é a expressão gênica e sua relação com o câncer. Atualmente contém dados de 64 tipos de neoplasias, com 575 pacientes com amostras pareadas (tumor *versus* controle) para análise de miRNAs e 667 pacientes com amostras pareadas para análise de expressão gênica (“BioXpress Tools & Resources” 2017). Esse banco aceita consultas via *web* de genes pontuais ou do *dataset* inteiro. É possível usar identificadores de genes do *HUGO Gene Nomenclature Committee*, UniProtKB e RefSeq. A normalização da expressão a partir da contagem de *reads* é feita com o DEseq (Anders e Huber 2010). Quando há amostra normal e tumoral de um mesmo paciente, o dado é apresentado em *fold change*.

Ao se consultar a expressão diferencial de um gene, o resultado é uma tabela onde cada linha representa as informações desse gene em um dado tipo de câncer de uma certa fonte, que em maioria é o TCGA. As colunas têm dados de nomenclatura, um *fold change* com a estatística associada, o tipo do tumor e o número de pacientes utilizados para fazer essa comparação.

É possível também buscar genes diferencialmente expressos de acordo com o tumor desejado. Nesse caso, o resultado é uma tabela onde cada linha representa um gene. As colunas têm o número de amostras em que o gene foi detectado / total de amostras, a expressão mínima, média, mediana, máxima, além do primeiro e terceiro quartil.

O banco apresenta dados de expressão de genes em amostras não pareadas, ou seja, sem a

comparação com o tecido não tumoral. É possível consultar dados só de amostras tumorais, cuja principal fonte é o ICGC, ou só de amostras não tumorais, provenientes do *Expression Atlas*. Esses dados permitem comparação da expressão entre diferentes tumores e sua contraparte normal.

A seguir serão apresentados os objetivos do trabalho. Adiante, estará a metodologia desenvolvida para obtenção dos dados de interesse e modelagem do banco de dados local. Os resultados alcançados estão disponíveis na seção Resultados, onde se encontram maiores detalhes sobre os dados obtidos e o modelo conceitual do banco de dados local. No item Discussão são comparados os resultados obtidos com dados da literatura. Finalmente, são apresentadas as conclusões do presente estudo e as possíveis direções futuras do trabalho.

Justificativa

As tecnologias e NGS podem trazer importantes avanços para os estudos de PDACs. Considerando-se a heterogeneidade destes tumores e a riqueza de informações capturadas por meio de experimentos de RNAseq e DNAseq, estudos com essa abordagem metodológica podem ter relevância para ampliar a compreensão das bases moleculares da doença e para o desenvolvimento de terapias individualizadas. Bancos de dados são ferramentas poderosas tanto para armazenamento de dados já existentes quanto ponto de partida para sua análise, solucionando questões persistentes e recorrentes na área da bioinformática, como integração, padronização e armazenamento de grandes volumes de informação. Por um lado, existem diversos bancos de dados públicos que permitem o *download* livre de seu conteúdo, que contemplam desde expressão gênica até dados de proteômica. Porém, não existe uma plataforma unificada na qual se possa fazer o *upload* de dados próprios e comparação destes com os dados de bancos públicos de maneira eficiente e consistente. Outra dificuldade dos bancos existentes é a ausência de informações clínicas ou falta de flexibilidade para relacioná-las aos achados experimentais. O Laboratório de Genômica e Expressão Gênica em Câncer, grupo no IQ-USP no qual este estudo se insere, tem projetos em andamento que envolvem a geração de dados de NGS de pacientes de PDACs e lesões pré-malignas. A modelagem e a implementação de um banco de dados local, que incorpore dados disponíveis publicamente e facilite comparação com dados gerados pelo próprio laboratório resultarão na criação de uma ferramenta de grande importância para a análise, sumarização e priorização de novos alvos com potencial para o diagnóstico, prognóstico e tratamento de PDACs que serão selecionados para caracterização experimental detalhada.

Objetivos

Objetivo geral

Modelar e implementar um banco de dados integrando dados de câncer de pâncreas públicos aos gerados pelo Laboratório de Genômica e Expressão Gênica em Câncer. O banco local desenvolvido deve conter dados de experimentos de NGS e dados clínicos associados.

Objetivos específicos

- Geração de *scripts* para recuperação de dados públicos de exoma, transcrito e dados clínicos de diferentes bancos de dados públicos voltados para estudos em câncer de pâncreas.
- Modelagem do banco de dados local.
- Implementação do banco de dados local.
- Geração de *scripts* para extração, transformação e carregamento dos dados públicos para povoamento do banco de dados local.
- Povoamento do banco de dados local com dados gerados pelo laboratório.
- Desenvolvimento de consultas de interesse ao banco de dados local para priorização de genes candidatos para estudos em câncer de pâncreas.

Material e Métodos

Desenho experimental

O desenho experimental está representado na Figura 1. Foi modelado um banco de dados local chamado CaRDIGAn (*Cancer Relational Database for Integration and Genomic Analysis*), que conta com duas principais fontes de entrada de dados. A primeira consiste de dados provenientes da literatura, disponíveis nos bancos de dados públicos Ensembl, ICGC e TCGA, descritos separadamente adiante e resumidos na Tabela 1. A extração desses dados foi realizada por meio de *scripts* específicos para cada banco. A segunda tem como entrada os dados provenientes de três

projetos de doutorado em andamento no Laboratório de Genômica e Expressão Gênica em Câncer do Instituto de Química da Universidade de São Paulo (LGE GC), onde estão sendo gerados e processados dados de NGS de amostras de pâncreas de diferentes tipos histopatológicos.

Figura 1: Esquema da metodologia adotada. Foram recuperados dados de expressão gênica, mutações e expressão gênica de câncer de bancos públicos (em azul), que foram formatados para entrada no banco desenvolvido localmente (verde escuro), denominado CaRDIGAn (*Cancer Relational Database for Integration and Genomic Analysis*). Paralelamente, dados gerados por nosso grupo na qual este estudo está incluso também foram formatados e incluídos no CaRDIGAn (em verde claro). O banco, então, suporta consultas relevantes para a pesquisa de novos marcadores e genes de interesse.

Bancos de dados biológicos utilizados

Os bancos consultados foram selecionados de acordo com os dados disponíveis. A Tabela 1 detalha o tipo de dados que se obtém conforme o banco estudado. Embora também tenhamos acesso a dados de metilação, a princípio priorizamos nossos esforços em análises de expressão gênica e de mutação somática simples, ou seja, de até 200 nucleotídeos, o que exclui rearranjos cromossômicos.

Tabela 1: Bancos que foram utilizados para consulta e informações disponíveis, respectivamente

Banco	Dado disponibilizado	Observações	Referência
Ensembl	Nomes de genes e transcritos e respectivas coordenadas	Referência genômica	(Aken et al. 2016)
ICGC	CNV, rearranjos genômicos, expressão de RNA, proteômica e mutação	–	(Junjun Zhang et al. 2011)
TCGA	CNV, expressão de RNA e miRNA e mutação	Contido no ICGC	(“The Cancer Genome Atlas Home Page” 2015)

CNV: *copy number variation* – alteração do número de cópias de uma dada sequência de DNA

Estratégia para acesso de dados dos bancos públicos

Ensembl

Os dados recuperados nos diferentes bancos públicos usam diferentes nomenclaturas para genes e transcritos. Para o funcionamento correto do banco local é necessário ter os nomes dos genes com os identificadores e as respectivas coordenadas. A referência normal utilizada foi o Ensembl (Aken et al. 2016). Esse banco foi feito dentro da plataforma BioMart (Kasprzyk 2011) e tem sistema de acesso pelo API Bioconductor da linguagem R (Huber et al. 2015), que foi utilizado para obter os dados desejados.

ICGC

O ICGC (Zhang et al. 2011) é um esforço internacional para o estudo de diferentes tipos de câncer. Apesar de estar construído dentro da plataforma BioMart (Kasprzyk 2011), este banco ainda não é acessível via APIs. Assim, seus dados foram baixados com auxílio de um *script* em Python (van Rossum 2011) que usa recursos do Bash (“Bash Reference Manual” 2015) que foram incorporados ao CaRDIGAn.

O *The Cancer Genome Atlas* (TCGA) (“The Cancer Genome Atlas Home Page” 2015) é um programa do governo americano cujos dados foram incorporados ao ICGC. Assim, seus dados foram baixados e tratados da forma descrita para aquele banco.

Modelagem do CaRDIGAn

Uma vez recuperados os dados dos bancos públicos, verificou-se suas características e a natureza do dado representado. Com base nas observações realizadas, prosseguiu-se com a modelagem do banco de dados local. O modelo de abstrações de dados gerou o modelo conceitual que por sua vez foi mapeado para o modelo relacional como indicado para dados biológicos de NGS em estudo publicado anteriormente (Birney e Clamp 2004). Foram adotadas as abstrações de classificação e o modelo de entidade-relacionamento estendido. Por classificação entendem-se os construtores semânticos de restrição que definem uma classe ou tipo de entidade. Uma entidade tem um conjunto de atributos que a caracterizam. Cada instância dessa entidade é adicionada como uma tupla e cada tupla pode ser identificada univocamente através da chave primária (Elmasri e Navathe 2010). De maneira simplificada, pode-se comparar uma entidade a uma tabela onde os atributos são as colunas e as tuplas, as linhas, conforme a figura abaixo (Figura 2).

Paciente					
id_paciente	banco_de_origem	project_code	donor_sex	donor_vital_status	disease_status_last_followup
1	ICGC	PACA-AU	male	alive	progression
2	ICGC	PACA-AU		alive	stable
3	ICGC	PACA-AU			
4	ICGC	PACA-AU	female	alive	stable
5	ICGC	PACA-AU	male	deceased	progression
6	ICGC	PACA-AU	male	deceased	

Figura 2: Nomenclatura associada ao modelo entidade-relacionamento. A entidade Paciente está representada em uma relação. Cada atributo que a define é uma coluna, sendo o atributo id_paciente sua chave primária. Cada instância dessa entidade, ou seja, cada novo paciente, é uma nova tupla, representada nas linhas.

Um relacionamento ocorre entre n entidades e pode ter atributos associados ao próprio relacionamento. Tipicamente, os relacionamentos podem ter cardinalidade um para um (1:1), um para vários (1:N) e vários para vários (N:M). A notação de cardinalidade utilizada foi a descrita por

Elmasri e Navathe (2010), mais completa, onde os parênteses próximos a uma dada entidade representa a contribuição mínima e máxima dessa entidade no relacionamento, conforme a Figura 3.

Figura 3: Notação de cardinalidade adotada. Os números de mínimo e máximo indicam a participação da entidade E no relacionamento R.

A modelagem foi feita com auxílio do BrModelo (“brModelo - Ferramenta de Ensino: Modelagem de Dados (MER)” 2013), um *software* de acesso livre que permite a idealização do modelo conceitual e sua posterior representação no modelo lógico e físico. O programa draw.io (draw.io 2017) foi utilizado para geração da figura de modelo conceitual (Figura 5).

Implementação do CaRDIGAn

O SGBD utilizado foi o PostgreSQL, também de acesso livre (“PostgreSQL: The world’s most advanced open source database” 2015), associado a uma interface gráfica chamada pgAdmin (“pgAdmin III — pgAdmin III 1.20.0 documentation” 2015).

Geração de *scripts* para extração, transformação e carregamento dos dados para povoamento CaRDIGAn

A estratégia para manipulação dos dados está esquematizada na Figura 4. Os dados recuperados dos bancos públicos e os do laboratório foram importados, tais quais estavam em seu repositório, para um “banco” intermediário. Usa-se as aspas na palavra banco pois trata-se de tabelas com linhas duplicadas, não estruturadas e sem os construtores semânticos característicos de um banco de dados normalizado. Uma vez que os dados foram adicionados ao um ambiente de banco de dados, houve a possibilidade de manipulá-los utilizando-se SQL para formatação destes dados para então serem inseridos no CaRDIGAn. Esse processo normalmente é realizado por ferramentas de ETL (do inglês, *Extraction, Formating and Loading*). Essas ferramentas permitem extrair os dados, transformá-los e, finalmente, carregá-los para o banco de dados estruturado, conforme o desenho experimental. Tais funções permitem criar dados homogêneos e padronizados no CaRDIGAn

(Vassiliadis, Simitsis, e Skiadopoulos 2002). No presente estudo, as funções de ETL foram feitas diretamente através de *scripts* em SQL, detalhados na seção de Resultados.

Para melhorar o desempenho das buscas no banco intermediário, este foi indexado após a inserção dos dados.

Figura 4: Estratégia para manipulação dos dados públicos e dados gerados pelo laboratório (LGEGC) para entrada no banco de dados local (CaRDIGAn). Os dados públicos, recuperados via *scripts* em Python com auxílio de recursos nativos do Bash, bem como dados do laboratório, foram importados, tais quais estavam para um “banco” intermediário. Uma vez dentro da estrutura do banco de dados, os dados foram manipulados, via SQL, e adicionados ao CaRDIGAn de maneira padronizada.

Desenvolvimento de consultas de interesse

As consultas de interesse foram desenvolvidas em SQL com base no modelo desenvolvido. Mais detalhes sobre as consultas encontram-se na seção Resultados.

Análise de sobrevida

Foram geradas curvas de sobrevida de Kaplan Meier com dados clínicos de pacientes com PDAC avaliados pelo ICGC, agrupados em função da expressão de genes diferencialmente expressos identificados em nosso estudo, ou da presença de mutações no proto-oncogene *KRAS*. Diferenças estatísticas entre as curvas de sobrevida foram analisadas utilizando-se o teste *log-rank* com auxílio do pacote *Survival* do R (Therneau 2015). O tempo de sobrevida foi calculado da data do diagnóstico até a data do óbito. Os gráficos com as curvas de sobrevida foram gerados utilizando o pacote *ggfortify* (Tang, Horikoshi, e Li 2016).

Análise de genes diferencialmente expressos

Os dados de expressão gênica de amostras do LGEGC foram processados pelo aluno de doutorado Julio Sosa e inseridos no CaRDIGAn. Os dados foram previamente normalizados e quantificados (em nível de genes e transcritos) criando-se matrizes de contagem dos *reads* com o programa RSEM e considerando como referência o conjunto de transcritos reconstruídos na

montagem. A identificação dos genes diferencialmente expressos nas condições tumoral e normal foi realizada utilizando a ferramenta DESeq2 (Love et. al., 2014).

Resultados

Recuperação de dados a partir de repositórios públicos

Ensembl

Os *scripts* desenvolvidos recuperam dados de genes e transcritos nas versões das montagens GRCh37 e GRCh38 do Ensembl. Foram adquiridas tabelas distintas para cada versão. Os atributos selecionados foram os identificadores do próprio Ensembl, do RefSeq e o *Gene Name* do NCBI. Além disso, foram recuperadas informações acerca das coordenadas cromossômicas e orientação da fita do gene ou transcrito. Outra informação disponível que foi selecionada foi o biotipo do gene ou transcrito, que indica se este é um gene codificador de proteína, miRNA, etc. Abaixo encontra-se o pseudocódigo do programa desenvolvido para recuperação de dados referentes a genes (Algoritmo 1). A estrutura do programa para obtenção de dados de transcritos é semelhante a de genes, e portanto será omitida. Os programas desenvolvidos, bem como os atributos selecionados para gene ou transcrito, podem ser consultados no Anexo 2.

Algoritmo 1: Obtenção de dados sobre genes anotados no genoma humano a partir de diferentes montagens do Ensembl (GRCh37 e GRCh38)

1. Conecta-se ao Bioconductor
2. Ativa o pacote biomaRt
3. Função `pega_genes_selecionando_versao`
Recebe como entrada:
 base da Biomart, *default*: Ensembl
 o *dataset*, *default*: `hsapiens_gene_ensembl`
 o *host*, *default*: `www.ensembl.org`
 a versão do genoma desejada
Acessa o *mart* do Ensembl e obtém:
 Coordenadas (com cromossomo, início, fim e fita)
 Nomes no Ensembl, RefSeq, NCBI e MiRBase
 Biotipo (codificador, RNA longo intergênico, pseudogene, miRNA, RNA mitocondrial, antisense, etc)
4. Uso da função `pega_genes_selecionando_versao` para a montagem GRCh37 do genoma, salvando o resultado como arquivo tsv (*tab separated values*, formato de arquivo de tabela, onde as colunas são separadas por TABs)
5. Uso da função `pega_genes_selecionando_versao` para a versão GRCh38 do genoma, salvando o resultado como arquivo tsv

Foram recuperadas 63.677 coordenadas únicas para a montagem GRCh37 e 63.898 para a montagem GRCh38. Ressalta-se que tais coordenadas referem-se não apenas a genes que codificam proteínas, mas também pseudogenes, miRNAs e demais classes de genes, o que explica a quantidade obtida. No caso dos transcritos, tem-se 215.170 transcritos para a montagem GRCh37 e 218.207 para a montagem GRCh38.

As tabelas recuperadas dos bancos públicos, tais como estavam no banco de origem, foram importadas para um banco intermediário, servindo de origem para os *scripts* de ETL, detalhados a seguir. Os atributos de cada tabela original estão disponíveis no Anexo 1.

ICGC e TCGA

Os dados foram recuperados utilizando um *script* em Python (van Rossum 2011) descrito a seguir. O usuário escolhe as tabelas desejadas e qual versão do projeto será buscada, sendo o *default* todas as tabelas, com exceção dos dados de metilação, e a versão mais recente do banco. Assim, o programa acessa o site do projeto, processa o arquivo *readme* da versão desejada e armazena os nomes dos projetos. O *download* é feito com auxílio do *wget*, um recurso nativo do Bash (“Bash Reference Manual” 2015). Existem no ICGC 76 projetos de 16 países, que estudam 21 tipos de tumores. Os dados do TCGA estão contidos nesse conjunto. O programa desenvolvido obtém os dados de todos eles por *default*, mas há opção de seleção de projetos de interesse. O pseudocódigo do programa desenvolvido para recuperação de dados do ICGC está descrito abaixo (Algoritmo 2).

Após o *download*, checkou-se manualmente se os arquivos obtidos estavam íntegros e se o tamanho de cada um correspondia às informações disponíveis no *site* do ICGC (“DCC Data Releases | ICGC Data Portal” 2018).

Algoritmo 2: Obtenção de dados do ICGC e TCGA

Entradas do programa:
pasta onde serão salvos os arquivos baixados, <i>default</i> : pasta onde está o programa
versão do ICGC desejada, <i>default</i> : <i>current</i>
indicador para download de dados de metilação, <i>default</i> : Falso
projeto de interesse, <i>default</i> : “” (vazio)
1. Se Indicador de metilação for verdadeiro então:

Vetor de tabelas que serão recuperadas: *copy number somatic mutation, donor, donor exposure, donor family, donor therapy, exp array, exp seq, meth array, meth seq, mirna seq, protein expression, sample, simple somatic mutation.open, specimen, splice variant, structural somatic mutation*

Senão:

Vetor de tabelas que serão recuperadas: *copy number somatic mutation, donor, donor exposure, donor family, donor therapy, exp array, exp seq, mirna seq, protein expression, sample, simple somatic mutation open, specimen, splice variant, structural somatic mutation*

Fim se

2. Se versão desejada for especificada:

Se input da versão não é um inteiro então:

Fechar o programa e reportar erro

Fim se

Senão:

Versão que será recuperada será a especificada

Senão:

Versão que será recuperada será a *current*

Fim se

3. Se pasta de saída foi especificada:

Cria pasta com o nome especificado

Muda o diretório para a pasta especificada

Fim se

4. Download do arquivo *README* do banco ICGC na versão especificada usando o *wget* do *bash*

5. Se projeto_de_interesse foi determinado:

Vetor projetos terá uma instância, indicativa do projeto de interesse

Senão:

Leitura do arquivo *README* e obtenção dos nomes de projetos disponíveis através de expressões regulares. Projetos são salvos em um vetor

Fim se

6. Para cada projeto no vetor de projetos:

Cria uma pasta para o projeto

Muda o diretório para a pasta criada

Para cada tabela do vetor de tabelas que serão recuperadas:

Gera-se uma string combinada com o nome da tabela e do projeto, no padrão de nomeclaturas IGCG

Solicita-se ao servidor o download da tabela usando a string combinada, nome do projeto e a versão desejada do ICGC. O download é feito com o *wget* do *bash*

Fim para

Muda de diretório para a pasta original

Fim para

7. Apaga-se os arquivos vazios em todas as pastas criadas, que representam tabelas ainda não existentes dentro dos projetos.

Fim do programa

Cada projeto obrigatoriamente tem três tabelas disponíveis: “*donor*”, “*sample*” e “*specimen*”. Os projetos do ICGC/TCGA estão em diferentes pontos de andamento e portanto os dados experimentais disponíveis para cada projeto podem variar, assim as demais tabelas são opcionais. As tabelas existentes a respeito de câncer de pâncreas e a descrição de seu conteúdo estão listadas na Tabela 2.

Tabela 2: Tabelas disponíveis no ICGC (que inclui o TCGA) com dados de câncer de pâncreas e o tipo de dado disponibilizado

Nome	Descrição dos dados contidos
<i>donor</i>	Dados clínicos dos doadores como idade, idade quando diagnosticado o câncer, estado vital
<i>donor exposure</i>	Fatores de exposição do doador, como tabagismo e consumo de álcool
<i>donor family</i>	Histórico de câncer na família do doador
<i>donor therapy</i>	Terapias a que o doador foi submetido, intenção e resultado
<i>sample</i>	Identificação das amostras conforme o paciente e celularidade. As amostras estão relacionadas a um espécime cirúrgico
<i>specimen</i>	Detalhes da coleta de espécimes cirúrgicos conforme o doador. Contêm dados sobre o tipo do tumor e o armazenamento do espécime
<i>copy number somatic mutation</i>	Mutações somáticas de variação de cópias conforme a amostra
<i>simple somatic mutation open</i>	Mutações de até 200 nucleotídeos de acesso livre conforme amostra
<i>structural somatic mutation</i>	Alterações cromossômicas por amostra, incluindo translocações, deleções e inserções.
<i>exp array</i>	Dados de expressão gênica baseados em experimentos com <i>microarray</i>
<i>exp seq</i>	Dados de expressão gênica baseados em experimentos de <i>RNAseq</i>

<i>mirna seq</i>	Dados de expressão gênica de micro RNAs baseado em experimentos de <i>RNAseq</i>
------------------	--

Conforme mencionado anteriormente, os dados disponíveis do ICGC não estão em um formato normalizado, havendo redundância em diversas colunas. Como exemplo, pode-se observar na Tabela 3, os dados do ICGC de expressão por sequenciamento, os atributos referentes ao experimento realizado se repetem por todas as linhas, bem como dados do paciente e espécime coletado. Tal redundância está presente em todas as tabelas com dados experimentais, motivando a criação de um modelo otimizado para o armazenamento normalizado de tais dados.

Tabela 3: Recorte da tabela ICGC Exp Seq do projeto PACA-AU, com dados de quatro genes para um mesmo paciente. Nota-se o que algumas colunas repetem-se para todas as ocorrências de um mesmo paciente, particularmente as referente a dados experimentais e dados sobre o paciente, espécime e amostra

icgc donor id	project code	icgc specimen id	icgc sample id	submitted sample id	analysis id	gene model
DO35300	PACA-CA	SP78121	SA412587	MPCC_0016_Pa_C	release21	Ensembl
DO35300	PACA-CA	SP78121	SA412587	MPCC_0016_Pa_C	release21	Ensembl
DO35300	PACA-CA	SP78121	SA412587	MPCC_0016_Pa_C	release21	Ensembl
DO35300	PACA-CA	SP78121	SA412587	MPCC_0016_Pa_C	release21	Ensembl
DO35300	PACA-CA	SP78121	SA412587	MPCC_0016_Pa_C	release21	Ensembl

icgc donor id	gene id	normalized read count string	raw read count string	fold change string	assembly version	platform
DO35320	ENSG00000258429	603.166	11533		GRCh37	Illumina HiSeq
DO35164	ENSG00000151694	206.404	16499		GRCh37	Illumina HiSeq
DO35300	ENSG00000127124	256.542	8294		GRCh37	Illumina HiSeq
DO35300	ENSG00000127125	103.147	5632		GRCh37	Illumina HiSeq
DO35300	ENSG00000127184	554.19	66805		GRCh37	Illumina HiSeq

icgc donor id	total read count	experimental protocol	other analysis algorithm
DO35320		Nugen Encore mRNA Library http://www.nugeninc.com	
DO35164		Nugen Encore mRNA Library http://www.nugeninc.com	
DO35300		Nugen Encore mRNA Library http://www.nugeninc.com	
DO35300		Nugen Encore mRNA Library http://www.nugeninc.com	
DO35300		Nugen Encore mRNA Library http://www.nugeninc.com	

icgc donor id	alignment algorithm	normalization algorithm
DO35320	Tophat2.0.3 http://tophat.cbcb.umd.edu	Cufflinks2.0.2 http://tophat.cbcb.umd.edu

DO35164	Tophat2.0.3 http://tophat.cbcb.umd.edu	Cufflinks2.0.2 http://tophat.cbcb.umd.edu
DO35300	Tophat2.0.3 http://tophat.cbcb.umd.edu	Cufflinks2.0.2 http://tophat.cbcb.umd.edu
DO35300	Tophat2.0.3 http://tophat.cbcb.umd.edu	Cufflinks2.0.2 http://tophat.cbcb.umd.edu
DO35300	Tophat2.0.3 http://tophat.cbcb.umd.edu	Cufflinks2.0.2 http://tophat.cbcb.umd.edu

icgc donor id	sequencing strategy	raw data repository	raw data accession	reference sample type
DO35320	RNA-Seq	EGA	EGAS00001000395:EGAD00001001956:EGAN00001253	
DO35164	RNA-Seq	EGA	EGAS00001000395:EGAD00001001956:EGAN00001253	
DO35300	RNA-Seq	EGA	EGAS00001000395:EGAD00001001956:EGAN00001253	
DO35300	RNA-Seq	EGA	EGAS00001000395:EGAD00001001956:EGAN00001253	
DO35300	RNA-Seq	EGA	EGAS00001000395:EGAD00001001956:EGAN00001253	

LGEGC

O Laboratório de Genômica e Expressão Gênica em Câncer no IQ-USP gerou dados de RNAseq de 14 amostras de PDAC e de tecido pancreático sem evidência de transformação maligna (tecido pancreático sem alterações estruturais e morfológicas provenientes de bordas livres de tumor) obtidas do banco de tumores do AC Camargo Cancer Center, que colabora nesse projeto. A ausência de tratamento neoadjuvante pré-operatório foi um critério de inclusão no estudo. Foi sequenciado o transcrito poliadenilado e não-poliadenilado das amostras obtidas (RNA total depletado de RNA ribossomal). Os fragmentos de tecidos analisados foram congelados em nitrogênio líquido imediatamente após a ressecção cirúrgica, e fazem parte da coleção do banco de tumores mantido pelo Serviço de Anatomia Patológica do Hospital. O estudo foi aprovado pela Comissão de Ética da instituição e se encontra registrada na Plataforma Brasil (CAAE: 15059213.0.0000.5432).

Modelagem do CaRDIGAn

O CaRDIGAn foi modelado levando-se em conta as informações disponíveis nos bancos públicos e as informações obtidas por nosso grupo. Assim, organizou-se as entidades de uma forma representativa de nosso universo e utilizou-se os nomes dos atributos tais quais estão nos bancos consultados, quando possível. Novos atributos foram incluídos em português.

Tradicionalmente, nomes de entidades são substantivos no singular e relacionamentos são verbos. O nome de tabelas físicas criadas a partir de um relacionamento foi padronizado para estar

no formato *Entidade A Verbo Entidade B*. O Anexo 2 contém um dicionário de dados com todas as tabelas presentes no banco e uma breve descrição de seu conteúdo. Pode-se consultar também seus atributos e a descrição dos mesmos.

A Figura 5 mostra as entidades e relacionamentos do modelo conceitual do CaRDIGAn. Na região superior esquerda, em azul, estão os dados clínicos. A entidade ‘Paciente’ relaciona-se com ‘Exposição’ que detalha fatores de exposição de interesse para a biologia de muitos tumores, como tabagismo e consumo de álcool. Outro fator que pode ser relevante para certas neoplasias é o histórico familiar. Como os familiares não podem ser identificados de maneira unívoca com os dados disponíveis forma-se uma entidade fraca com ‘Paciente’ (visível no Anexo 2). Na entidade ‘Tratamento’ são detalhadas as terapias às quais o paciente foi submetido e os detalhes associados, como resposta e duração do tratamento.

A entidade ‘Espécime’ refere-se a um material biológico coletado de um paciente. Um paciente pode ter diversos tipos de espécimes coletados, como saliva ou peça de tumor proveniente de ressecção cirúrgica e pode, ainda, ter diversos espécimes coletados ao longo do tempo. A peça pode ser submetida a um diagnóstico, representado numa entidade de mesmo nome. ‘Diagnóstico’ contém informações especificações do tumor, como tipo, grau e sistema adotado para estadiamento. ‘Marcador’ é uma entidade criada para armazenar dados de marcadores sorológicos utilizados no diagnóstico. O resultado observado é armazenado na tabela ‘Marcador Auxilia Diagnóstico’, já que é um relacionamento N:M.

Um espécime pode ser dividido em várias amostras, que é o material efetivamente usado nos experimentos. Assim, a entidade ‘Amostra’ organiza os experimentos, ligando as amostras aos respectivos espécimes e, conseqüentemente, aos respectivos pacientes. A parte experimental está representada na cor lilás na Figura 5.

Os experimentos podem ser de mutação, expressão gênica ou expressão proteica. Considera-se um experimento uma aplicação pontual de uma metodologia, em um dia específico em um dado equipamento, seguindo um protocolo. Os relacionamentos do tipo ‘*Experimento “X” Utiliza Amostra*’, onde “X” pode ser Mutação, Expressão Gênica ou Expressão Proteica, indicam quais amostras foram usadas em certo experimento. Nos dados do ICGC não fica claro quais amostras fazem parte de um mesmo conjunto, então assume-se que todas as amostras que compartilham os mesmos atributos experimentais, dentro de um mesmo projeto, fizeram parte de um experimento. Assim, tomemos como exemplo amostras de um experimento de expressão gênica. Todas as amostras que por exemplo são do projeto PACA-AU e foram estudadas em um mesmo equipamento, com

mesmo protocolo e têm o mesmo total de *reads* são consideradas partes de um mesmo experimento e terão o mesmo identificador de experimento. Cada tupla conterá este identificador de experimento e o identificador de uma das amostras, havendo tantas tuplas quanto amostras utilizadas no experimento. o relacionamento N:M ‘Experimento de Expressão Gênica Utiliza Amostra’.

Os experimentos de mutação somática podem ser do tipo simples, estrutural ou variação de cópias. Os experimentos de expressão gênica podem ser derivados de *array* ou sequenciamento. Embora sejam muito diferentes, os experimentos de mutação simples, estrutural e variação de cópias compartilham diversos atributos, da mesma forma que os experimentos de *array* e sequenciamento o fazem entre si. Logo, foi descartada a necessidade de uma especialização para representar esses dados, mantendo-se apenas as entidades mais genéricas no modelo conceitual e físico: ‘Experimento de Mutação Somática’ e ‘Experimento de Expressão Gênica’. Para facilitar as consultas futuras, criou-se um atributo em ambas as entidades especificando o tipo do experimento.

Embora tenham atributos semelhantes, os resultados experimentais variam entre os tipos de experimento. Para evitar tabelas com grandes áreas em branco, cada tipo de experimento tem um relacionamento diferente com amostra e gene (ou transcrito). Os relacionamentos, bem como as tabelas no modelo físico, se chamam ‘*Nome do Tipo do Experimento Avalia Gene (ou Transcrito) em Amostra*’, por exemplo ‘Mutação Simples Avalia Gene em Amostra’. Tratam-se de relacionamentos quádruplos entre duas amostras, a experimental e a amostra controle, gene e experimento. Logo, cada tupla contém os dados experimentais de um gene ou transcrito medidos para uma dada amostra tumoral, uma amostra controle em um identificador de experimento. Na Figura 5 os relacionamentos desse tipo estão representados em negrito com letras em lilás.

A entidade ‘Experimento de Expressão Proteica’ não tem especializações, mas também tem um relacionamento ‘Experimento de Expressão Proteica Utiliza Amostra’ que define amostras usadas em um experimento, e um relacionamento ‘Expressão Proteica Avalia Gene em Amostra’. Embora o gene não esteja sendo avaliado de fato, as proteínas são codificadas a partir de genes e mantém-se esse relacionamento para nomeá-las.

A entidade ‘Gene’ e ‘Transcrito’ armazenam os diversos nomes que um gene ou transcrito pode ter em 4 bancos públicos usados como referência: Ensembl, RefSeq, HGNC e MiRBase. As tabelas de coordenada, ‘Coordenada Gene’ e ‘Coordenada Transcrito’ armazenam, em cada tupla, informação do cromossomo, fita, início e fim de um *locus*, bem como a montagem usada como referência. Tabelas de genes ou transcritos se relacionam com as tabelas de coordenadas de forma NxM. É simples pensar que uma coordenada pode ter diversos nomes a ela associados, já que ‘Gene’

e ‘Transcrito’ têm diversos bancos que usam identificadores diferentes, mas a ideia que uma coordenada gênica pode ter mais de um identificador de um banco associado a ela só é intuitiva quando se pensa em duas montagens ou mais para um mesmo *locus*. O banco modelado tem dados da montagem GRCh37 e GRCh38. Como exemplo, o nome *KRAS* da entidade ‘Gene’ está associado a duas coordenadas da tabela ‘Coordenada Gene’, ambas na fita reversa do cromossomo 12, mas com posições distintas de início e fim devido às diferentes montagens. Um caso especial é o dos miRNAs. Eles são transcritos a partir de um gene mas podem ser processados de maneiras diferentes, dando origem a vários nomes para um mesmo *locus* dentro de uma mesma montagem quando o banco para consulta selecionado é o MiRBase. Por essa razão, ao tratar RNAs nas buscas e conversões de nome, miRNAs devem ser tratados de forma distinta dos demais genes.

Desenvolvimento de *scripts* para extração, transformação e carregamento de dados

Após a recuperação dos dados, estes só devem ser incluídos em um banco local após verificado que obedecem às normas de formatação nele implementadas (Golfarelli e Rizzi 2009). Por exemplo, o atributo *donor_age_at_diagnosis*, que contém a idade em que o paciente foi diagnosticado com câncer, deve conter um número, nunca um texto. Considerando a diversidade das fontes dos dados integrados pelo banco local implementado, se fez necessário de um passo de tratamento dos dados antes de serem carregados no banco. Assim, foram desenvolvidos *scripts* capazes de realizar esta tarefa de forma automatizada. *Scripts* deste tipo são comumente chamados de ETL, de extração, transformação e carregamento (do inglês, *Extract, Transform and Load*).

Os *scripts* de ETL devem capturar a particularidade de cada banco, serem capazes de checar as restrições de cada atributo e incluí-lo ao banco de maneira correta. Os *scripts* escritos podem ser adaptados caso algum destes bancos mude seus atributos ou inclua novos atributos. Com auxílio dos *scripts* de ETL o banco foi povoado e foi testado com consultas de interesse. O processo de desenvolvimento desses *scripts* foi feito de forma conjunta ao de modelagem, pois ao se perceber particularidades dos dados e de como se relacionam entre si, torna-se necessário fazer ajustes no modelo conceitual e, conseqüentemente no lógico e físico. Atualmente o banco tem todas as tabelas do banco intermediário carregadas e as entidades do banco local mapeadas com *scripts* de ETL para todos os dados clínicos, representados em azul na Figura 5, dados de expressão gênica por RNASeq e mutações somáticas simples, ambos em lilás, bem como o mapeamento da entidade ‘Gene’, ‘Coordenada Gene’ e ‘Gene Nomeia Coordenada’, representados em verde na mesma figura.

Ao executar-se um comando de escrita em um banco de dados, tal comando é enviado ao SGBD, que se encarrega da tarefa. Se um erro é detectado, o comando é interrompido e tem-se uma mensagem de erro. Assim, os comandos de escrita só vão alterar efetivamente o banco de dados se executados por completo e sem erros (Elmasri e Navathe 2010). Tal propriedade do banco de dados garante que os *scripts* de ETL só alteraram o banco se estiverem de acordo com as restrições impostas

Figura 5: Modelo conceitual do CaRDIGAn mostrando as entidades e relacionamentos criados. Para facilitar a visualização, os atributos foram omitidos. As entidades ‘Paciente’, ‘Amostra’ e ‘Gene’ são centrais para o banco. ‘Paciente’ e as entidades associadas em azul compreendem os dados clínicos. As amostras utilizadas nos experimentos são rastreadas para ‘Paciente’ através da entidade ‘Espécime’. A entidade ‘Amostra’ organiza a parte experimental do banco, representada em lilás. Os relacionamentos com letras lilases se referem a dados experimentais. AGA: Avalia Gene em Amostra. Para experimentos de mutação somática: E, estrutural; MS, mutação simples; VC, variação de cópias. Para experimentos de expressão gênica: A, *array*; Seq, sequenciamento; SeqATA, sequenciamento avalia transcrito em amostra. ‘Gene’, ‘Transcrito’ e ‘Coordenada’ são entidades usadas como referência e estão representadas em verde. Para consultar os atributos associados a cada entidade ou relacionamento, favor referir-se ao Anexo 2.

pelo modelo elaborado, não existindo a possibilidade de alteração parcial do banco ou que algumas tuplas sejam incorporadas e outras não, e garantindo a integridade da informação armazenada.

Mapeamento de genes e coordenadas

A entidade ‘Gene’ armazena os identificadores de quatro bancos de dados diferentes. É esperado que muitos dos identificadores de genes de um mesmo banco se repitam entre as versões do genoma. Assim, para obtenção dos identificadores de cada banco é feita a união dos identificadores das duas montagens, garantindo que apenas um indicador seja inserido no CaRDIGAn. Para manter a compatibilidade com os nomes dos bancos utilizados pelo ICGC, o campo ‘banco alias’ é preenchido para os identificadores do Ensembl, *Gene Symbol* e miRBase'. O mapeamento está representado na Tabela 4.

Tabela 4: Mapeamento da relação Gene. São feitas diferentes consultas para cada identificador de interesse (Ensembl, HGCN, *Gene Symbol* e MiRBase). As consultas foram feitas a partir das tabelas Genes v37 e Genes v38 obtidas a partir da aplicação do Algoritmo 1. A união das tabelas garante que cada gene só seja inserido uma vez, mesmo quando presente nas duas versões do genoma

Gene	Genes v37 UNIÃO Genes v38
Ensembl	
nome_no_banco	← <i>ensembl gene id</i>
Banco	← ‘ENSEMBL’
banco alias	← ‘Ensembl’
HGCN	
nome_no_banco	← <i>hgnc id</i>
Banco	← ‘HGCN’
Gene Symbol	
nome_no_banco	← <i>external gene name</i>
Banco	← ‘Gene Symbol’
banco alias	← ‘GAF’
MiRBase	
nome_no_banco	← <i>mirbase id</i>
Banco	← ‘MiRBase’
banco alias	← ‘miRBase’

O mapeamento para inserção de dados na tabela ‘Coordenada Gene’ é feito duas vezes, uma para cada versão de montagem do genoma. Seleciona-se as coordenadas únicas para cada montagem. O mapeamento está representado na Tabela 5.

Tabela 5: Mapeamento da relação Coordenada Gene. As consultas foram feitas a partir das tabelas Genes v37 e Genes v38 obtidas a partir da aplicação do Algoritmo 1. O mapeamento é feito uma vez para a tabela com a versão 37 do genoma e outra, com os mesmos parâmetros, para a versão 38. Assim, armazena-se as coordenadas de ambas as montagens

Coordenada Gene		Genes v37 e Genes v38
Cromossomo	←	chromosome_name
Inicio	←	start_position
Fim	←	end_position
Fita	←	strand
Referencia	←	referencia
Montagem	←	montagem
Source	←	source
Status	←	status
gene_biotype	←	gene_biotype

O mapeamento da relação que representa o relacionamento ‘Gene Nomeia Coordenada’ é feito usando as tabelas recuperadas do Ensembl. Para cada montagem, cada coordenada é relacionada a todos os identificadores disponíveis na tabela de referência, Genes_v37 ou Genes_v38. O mapeamento está representado na Tabela 6.

Tabela 6: Mapeamento da relação Coordenada Nomeia Gene. Essa relação é usada para mapear o relacionamento NxM entre ‘Coordenada Gene’ e ‘Gene’, usando as tabelas Gene v37 e Gene v38 como referência. Os identificadores dos diferentes bancos são relacionados a uma coordenada da versão 37 e uma coordenada da versão 38. A notação Genes v3X representa a necessidade de executar os comandos de busca e inserção com os dados da tabela Genes v37 e, em seguida, utilizar os mesmos comandos para a tabela Genes v38

Coordenada Nomeia Gene		Gene
id gene	←	id gene
Inserção de Identificadores		
Gene.nome no banco	=	Genes v3X.ensembl gene id
Gene.nome no banco	ou	Genes v3X.hgnc id

Gene.nome no banco	ou	Genes v3X.external gene name
Gene.nome no banco	ou	Genes v3X.mirbase id
Coordenada Nomeia Gene		Coordenada Gene
id coordenada	←	id coordenada
Critérios de Consistência		
Coordenada Gene.cromossomo	=	Genes v3X.chromosome name
Coordenada Gene.Inicio	=	Genes v3X.start position
Coordenada Gene.Fim	=	Genes v3X.end position
Coordenada Gene.Fita	=	Genes v3X.strand
Coordenada Gene.Referencia	=	Genes v3X.referencia
Coordenada Gene.Montagem	=	Genes v3X.montagem

Mapeamento de dados do ICGC

O modelo do banco local usou como base os nomes dos atributos do ICGC. Assim, foram necessárias poucas transformações no dado para definir onde seriam adicionados ao banco local para povoá-lo.

A entidade ‘Paciente’ foi a primeira a do ICGC a ser povoada no banco de dados, já que é uma dependência para as entidades ‘Espécime’ e ‘Amostra’. Seu mapeamento está representado na Tabela 7. Para otimizar o mapeamento, as tabelas originais do ICGC foram indexadas. Outra alteração realizada antes do mapeamento foi a adição de uma coluna com o nome do banco e sua versão. Tais alterações estão indicadas com o sinal “#” na Tabela 7.

Tabela 7: Mapeamento da entidade ‘Paciente’ no banco local. A tabela *ICGC Donor* recuperada do repositório foi usada como fonte de dados

Paciente		ICGC Donor
<i>donor id banco de origem</i>	←	<i>icgc donor id # indexada</i>
<i>project code</i>	←	<i>project code</i>
<i>study donor involved in</i>	←	<i>study donor involved in</i>
<i>submitted donor id</i>	←	<i>submitted donor id</i>
<i>donor sex</i>	←	<i>donor sex</i>
<i>donor vital status</i>	←	<i>donor vital status</i>
<i>disease status last followup</i>	←	<i>disease status last followup</i>
<i>donor relapse type</i>	←	<i>donor relapse type</i>
<i>prior malignancy</i>	←	<i>prior malignancy</i>

<i>cancer type prior malignancy</i>	←	<i>cancer type prior malignancy</i>
<i>donor age at enrollment</i>	←	<i>donor age at enrollment</i>
<i>donor relapse interval</i>	←	<i>donor relapse interval</i>
<i>donor survival time</i>	←	<i>donor survival time</i>
<i>donor interval of last followup</i>	←	<i>donor interval of last followup</i>
<i>donor age at last followup</i>	←	<i>donor age at last followup</i>
banco de origem	←	‘ICGC’ # adicionado à tabela original
versão do banco de origem	←	‘v25’ # adicionado à tabela original
<i>donor age at diagnosis inicial</i>	←	<i>donor age at diagnosis</i>
<i>donor diagnosis icd10 inicial</i>	←	<i>donor diagnosis icd10</i>
<i>donor tumour staging system at diagnosis inicial</i>	←	<i>donor tumour staging system at diagnosis</i>
<i>donor tumour stage at diagnosis inicial</i>	←	<i>donor tumour stage at diagnosis</i>
<i>donor tumour stage at diagnosis supplemental inicial</i>	←	<i>donor tumour stage at diagnosis supplemental</i>
<i>cancer history first degree relative</i>	←	<i>cancer history first degree relative</i>

A seguir foi feito o mapeamento da entidade ‘Espécime’. Trata-se de um relacionamento 1:N, assim ‘Espécime’ tem chave estrangeira para a entidade ‘Paciente’. ‘Espécime’ e ‘Diagnóstico’ estão em um relacionamento de cardinalidade 1:1, assim formam uma tabela única no modelo físico. O mapeamento dessa relação está representado na Tabela 8.

Tabela 8: Mapeamento da relação ‘Especime’. A tabela utilizada como base é a ICGC *Specimen*. Há uma chave estrangeira para a relação ‘Paciente’, mapeada conforme indicado abaixo

Espécime		ICGC Specimen
id banco de origem	←	<i>icgc specimen id</i> # indexada
<i>study specimen involved in</i>	←	<i>study specimen involved in</i>
<i>submitted specimen id banco de origem</i>	←	<i>submitted specimen id</i>
<i>specimen type</i>	←	<i>specimen type</i>
<i>specimen type other</i>	←	<i>specimen type other</i>
<i>specimen interval</i>	←	<i>specimen interval</i>
<i>specimen donor treatment type</i>	←	<i>specimen donor treatment type</i>
<i>specimen donor treatment type other</i>	←	<i>specimen donor treatment type other</i>
<i>specimen processing</i>	←	<i>specimen processing</i>
<i>specimen processing other</i>	←	<i>specimen processing other</i>

<i>specimen storage</i>	←	<i>specimen storage</i>
<i>specimen storage other</i>	←	<i>specimen storage other</i>
<i>tumour confirmed</i>	←	<i>tumour confirmed</i>
<i>specimen biobank</i>	←	<i>specimen biobank</i>
<i>specimen biobank id</i>	←	<i>specimen biobank id</i>
<i>specimen available</i>	←	<i>specimen available</i>
<i>tumour histological type</i>	←	<i>tumour histological type</i>
<i>tumour grading system</i>	←	<i>tumour grading system</i>
<i>tumour grade</i>	←	<i>tumour grade</i>
<i>tumour grade supplemental</i>	←	<i>tumour grade supplemental</i>
<i>tumour stage system</i>	←	<i>tumour stage system</i>
<i>tumour stage</i>	←	<i>tumour stage</i>
<i>tumour stage supplemental</i>	←	<i>tumour stage supplemental</i>
<i>digital image of stained section</i>	←	<i>digital image of stained section</i>
<i>percentage cellularity string</i>	←	<i>percentage cellularity string</i>
<i>level of cellularity</i>	←	<i>level of cellularity</i>
Chave estrangeira		
Paciente		
Especime.id_paciente	←	Paciente.id_paciente
ICGC_Specimen.icgc_donor_id	=	Paciente.donor_id_banco_de_origem

A entidade ‘Amostra’ tem um relacionamento 1:N com a relação ‘Espécime’ e, por tanto, recebe a chave estrangeira dessa relação. Para otimizar consultas relacionando dados experimentais e dados clínicos, foi adicionada também a chave estrangeira para a relação ‘Paciente’. O mapeamento da relação ‘Amostra’ está representado na Tabela 9.

Tabela 9: Mapeamento da entidade ‘Amostra’. São importados os dados da tabela ICGC Sample. As relações ‘Paciente’ e ‘Espécime’, juntamente às tabelas correspondentes do ICGC são usadas para o mapeamento de chaves estrangeiras. O sinal “=” indica as condições para a correspondência

Amostra		ICGC Sample
<i>sample id banco de origem</i>	←	<i>icgc sample id # indexada</i>
<i>submitted sample id banco de origem</i>	←	<i>submitted sample id</i>
<i>specimen id banco de origem</i>	←	<i>icgc specimen id</i>
<i>submitted specimen id banco de origem</i>	←	<i>submitted specimen id</i>

<i>analyzed sample interval</i>	←	<i>analyzed sample interval</i>
<i>percentage cellularity</i>	←	<i>percentage cellularity</i>
<i>level of cellularity</i>	←	<i>level of cellularity</i>
Chaves estrangeiras		
Espécime		
Amostra.id especime	←	Especime.id especime
Amostra.specimen id banco de origem	=	ICGC Specimen.icgc specimen id
ICGC Specimen.icgc specimen id	=	Especime.id banco de origem
Paciente		
Amostra.id paciente	←	Paciente.id paciente
Amostra.sample id banco de origem	=	ICGC sample.icgc sample id
paciente.donor id banco de origem	=	ICGC sample.icgc donor id

Para diminuir a redundância nas tabelas de resultados experimentais, dados referentes aos experimentos foram mapeados para uma entidade chamada ‘Experimento de Mutação Somática’ ou ‘Experimento de Expressão Gênica’. O povoamento desse tipo de tabela depende de uma etapa de comparação de diversos campos da tabela original (Tabela 10). Cada conjunto único da combinação desses campos define um experimento, que tem um identificador que é usado como chave estrangeira na formação das tabelas *Experimento X Avalia Gene em Amostra*, que contém os resultados experimentais. A Tabela 10 mostra um exemplo de mapeamento de entidades do tipo “Experimento X”, no caso, da entidade ‘Experimento de Mutação Somática’. A mesma estratégia foi aplicada à entidade ‘Experimento de Expressão Gênica’.

Tabela 10: Mapeamento da relação Experimento de Mutação Somática. Esse mapeamento é feito com base na tabela ICGC *Simple Somatic Mutation* e busca ocorrências únicas dos atributos indicados. Assim, define-se uma instância da relação modelada que terá um identificador único, usado como chave estrangeira em diversos relacionamentos

Experimento de Mutação Somática	ICGC <i>Simple Somatic Mutation</i>
tipo experiment	← ‘simples’
projeto de origem	← <i>project code</i>
<i>assembly version</i>	← <i>assembly version</i>
<i>gene build version</i>	← <i>gene build version</i>
<i>Platform</i>	← <i>platform</i>
<i>experimental protocol</i>	← <i>experimental protocol</i>

<i>sequencing strategy</i>	← <i>sequencing strategy</i>
<i>base calling algorithm</i>	← <i>base calling algorithm</i>
<i>alignment algorithm</i>	← <i>alignment algorithm</i>
<i>variation calling algorithm</i>	← <i>variation calling algorithm</i>
<i>other analysis algorithm</i>	← <i>other analysis algorithm</i>
banco de origem	← 'ICGC'
versao do banco	← '25'

Um mesmo experimento pode utilizar diversas amostras e uma amostra pode ser utilizada em mais de um experimento. Assim, o relacionamento N:M gera uma nova tabela. O CaRDIGAn tem, para cada entidade experimental um relacionamento com ‘Amostra’ para representar essa informação. O mapeamento das relações *Experimento X Utiliza Amostra* está representado na Tabela 11. A princípio o mapeamento desse tipo de relação foi feita em uma única consulta ao banco, procurando as ocorrências únicas da combinação do identificador utilizado pelo ICGC com todos os atributos que descrevem o experimento, seguido da conversão do identificador do ICGC da amostra para o identificador gerado pelo CaRDIGAn. Um mesmo paciente pode ter uma amostra tumoral e uma amostra controle sendo utilizada para um mesmo experimento. Nesse caso, são inseridas duas linhas na tabela *Experimento X Utiliza Amostra* correspondente, cada uma com um identificador de amostra mas ambas com o mesmo identificador de experimento. O processo de povoamento é feito duas vezes, adicionando-se ‘tumor’ ou ‘controle’ no atributo ‘tipo amostra’, utilizando o campo *icgc_sample_id* ou *matched_icgc_sample_id* em cada uma das vezes, respectivamente. Ambas as relações ‘Experimento de Mutação Utiliza Amostra’ e ‘Experimento de Expressão Gênica Utiliza Amostra’ têm subtipos, acrescentados no atributo ‘tipo experimento’. O mapeamento feito para a relação ‘Experimento de Mutação Utiliza Amostra’ é mostrado na Tabela 11. A mesma estratégia é utilizada para a relação ‘Experimento de Expressão Gênica Utiliza Amostra’.

Tabela 11: Mapeamento da relação Experimento de Mutação Utiliza Amostra. Esse mapeamento é feito com base na tabela ICGC *Simple Somatic Mutation* e nas relações Amostra e Experimento de Mutação Somática. Utiliza-se as ocorrências únicas de uma consulta envolvendo as fontes citadas com os critérios de consistência especificados abaixo. A consulta é realizada uma vez para amostras tumorais e outra vez para amostras controle

Experimento de Mutação Utiliza Amostra	ICGC <i>Simple Somatic Mutation</i> (SSM)
<i>raw data repository</i>	← <i>raw data repository</i>

<i>raw data accession</i>	← <i>raw data accession</i>
seq coverage	← seq coverage
Amostra	
id amostra	← id amostra
Experimento de Mutação Somática (EMS)	
id exp mut somatica	← id exp mut somática
Texto livre	
tipo amostra	← ‘tumor’ ou ‘controle’
tipo experiment	← ‘simples’
Critérios de Consistência	
ICGC SSM.project code	= EMS.projeto de origem
ICGC SSM.assembly version	= EMS.assembly version
ICGC SSM.platform	= EMS.platform
ICGC SSM.gene build version	= EMS.gene build version
ICGC SSM.experimental protocol	= EMS.experimental protocol
ICGC SSM.alignment algorithm	= EMS.alignment algorithm
ICGC SSM.base calling algorithm	= EMS.base calling algorithm
ICGC SSM.variation calling algorithm	= EMS.variation calling algorithm
ICGC SSM.other analysis algorithm	= EMS.other analysis algorithm
ICGC SSM.sequencing strategy::TEXT	= EMS.sequencing strategy
Determinação do Tipo de Amostra	
Amostra.sample id banco de origem	= ICGC SSM.icgc sample id # amostra tumoral
Amostra.sample id banco de origem	= ICGC SSM.matched icgc sample id # amostra controle

O mapeamento das entidades com resultados experimentais do tipo *Experimento X Avalia Gene em Amostra* são um caso particular do mapeamento anterior, onde além da conversão dos identificadores de amostra e experimento do ICGC para o CaRDIGAn, é necessário obter o identificador do gene. Assim, buscou-se fazer uma consulta única com todos os parâmetros da consulta descrita anteriormente, mas acrescida da informação do gene e selecionando-se as colunas com as informações de resultados experimentais. Esse tipo de consulta foi muito custosa computacionalmente, assim foi necessário mudar a estratégia de mapeamento.

A estratégia adotada, então, foi a inserção de uma coluna na tabela original do ICGC e o identificador de amostra do ICGC foi convertido, linha a linha, para o correspondente CaRDIGAn e armazenado nessa coluna. Da mesma forma, os atributos com resultados experimentais foram comparados à ‘Experimento X’ e inseridos em uma coluna adicionada à tabela original. Os identificadores do CaRDIGAn para os genes foram inseridos na tabela original de forma semelhante. Como não há concorrência entre as colunas para inserção de dados, as tarefas puderam ser feitas simultaneamente.

O mapeamento da relação ‘Mutação Simples Avalia Gene em Amostra’ está representado na Tabela 12. O mapeamento da entidade ‘Sequenciamento Avalia Gene em Amostra’ é muito semelhante e será omitido.

Tabela 12: Após a inserção dos identificadores do CaRDIGAn na tabela original do ICGC, busca-se as ocorrências únicas da combinação dos atributos relacionados abaixo para inserção de dados do ICGC no CaRDIGAn

Mutação Simples Avalia Gene em Amostra		ICGC Simple Somatic Mutation (Ocorrências Únicas)
id Amostra experimental	←	id Amostra experimental # identificador CaRDIGAn
id Amostra controle	←	id Amostra controle # identificador CaRDIGAn
id gene	←	id gene # identificador CaRDIGAn
id exp mut somatica	←	id exp mut somatica
Chromossomo	←	chromosome
chromossomo inicio	←	chromosome start
chromossomo fim	←	chromosome end
Fita	←	chromosome strand
mutation type	←	mutation type
reference genome allele	←	reference genome allele
mutated from allele	←	mutated from allele
mutated to allele	←	mutated to allele
quality score	←	quality score
Probability	←	probability
total read count	←	total read count
mutant allele read count	←	mutant allele read count
verification status	←	verification status

verification platform	←	verification platform
biological validation status	←	biological validation status
biological validation platform	←	biological validation platform
consequence type	←	consequence type
aa mutation	←	aa mutation
cds mutation	←	cds mutation

Mapeamento de dados do LGEGC

Diferentemente do mapeamento feito para o ICGC, onde os nomes dos atributos pouco foram alterados, a correspondência entre os dados gerados por nosso grupo e o banco local modelado demandou maiores esforços. Em alguns casos, foi necessário que os dados fossem efetivamente processados, combinados ou formatados. Visando a aplicabilidade do modelo proposto com dados de outras fontes além de bancos públicos, abaixo é apresentada a estratégia usada para o mapeamento dos dados no LGEGC na entidade ‘Paciente’ a partir da tabela “Info Pacientes ACCamargo”. A tabela original, mostrada em recorte a seguir (Tabela 14) foi formatada substituindo notações em código por termos semelhantes aos utilizados pelo ICGC, que foi adotado no modelo desenvolvido. As informações foram complementadas com uma segunda tabela, chamada Espécimes ACCamargo. O mapeamento desenvolvido está esquematizado na Tabela 13.

Tabela 13: Mapeamento da entidade Paciente com dados gerados pelo LGEGC em colaboração com o Hospital ACCamargo. A tabela Info Pacientes ACCamargo foi usada como base para o mapeamento. A tabela Espécimes ACCamargo foi utilizada para obtenção da informação do identificador interno do Paciente, presente exclusivamente nessa tabela. Dados entre aspas simples (‘’) foram inseridos no campo de destino em forma de texto, tais quais indicados

Paciente		Info Pacientes ACCamargo
<i>project code</i>	←	'LGEGC'
<i>study donor involved in</i>	←	'ACCamargo2015'
<i>submitted donor id</i>	←	RGH
<i>donor sex</i>	←	sexo
<i>donor vital status</i>	←	status vida
<i>disease status last followup</i>	←	status apos terapia adjuvante
<i>prior malignancy</i>	←	outras neoplasias primarias
<i>cancer type prior malignancy</i>	←	tipo neop1 outro sitio primario

<i>donor age at enrollment</i>	←	data diagnostico - data nascimento (em anos)
<i>donor survival time</i>	←	data ultimo follow up - data diagnostico (em dias)
<i>donor age at last followup</i>	←	data ultimo follow up - data nascimento (em anos)
banco de origem	←	'LGEGC'
versao do banco de origem	←	'nov2017'
<i>donor age at diagnosis inicial</i>	←	data diagnostic - data nascimento (em anos)
<i>donor tumour staging system at diagnosis inicial</i>	←	'AJCC'
<i>donor tumour stage at diagnosis inicial</i>	←	estagio patologia peca operada
<i>cancer history first degree relative</i>	←	hist familiar
Paciente		Espécimes ACCamargo
donor id banco de origem	←	id paciente LGEGC
Paciente. <i>submitted donor id</i>	=	Especimes ACCamargo. id hospital

Dados disponíveis no CaRDIGAn no presente momento

Os dados clínicos do ICGC, TCGA e LGEGC acerca dos pacientes de câncer de pâncreas foram incorporados em sua totalidade ao banco de dados modelado. A Figura 6 representa alguns dos dados clínicos disponíveis no CaRDIGAn. O banco conta com dados clínicos de 954 pacientes.

Para que o banco local modelado fosse capaz de lidar com os diferentes identificadores de genes e diferentes montagens do genoma, os dados de coordenadas gênicas e seus respectivos identificadores foram povoados no banco. Os dados experimentais de mutação somática simples (de até 200 nucleotídeos) disponíveis no ICGC foram povoados com sucesso, totalizando 483 pacientes. A Figura 7 representa os dados experimentais disponíveis, conforme o projeto de origem. Dados experimentais de expressão gênica por sequenciamento do ICGC, TCGA e LGEGC também estão disponíveis, com um total de 496 pacientes investigados.

Tabela 14: Recorte da tabela de dados clínicos geradas por nosso grupo em colaboração com o AC Camargo Cancer Center, usada como exemplo de fonte de dados que pode ser incorporada ao modelo de banco de dados proposto. A tabela original, que contém 38 atributos, foi recortada para facilitar a visualização. Os atributos mapeados para a Entidade Paciente do banco local estão descritos na Tabela 13

RGH	dt_nasc	dt_diag	sexo	tipo_hist	outras_neop	tipo_neop1	estad_pat	dt_PD_adj	status_PD_adj	dt_ult_FUP	status_vida
9XXX9379	13-May-41	13-Mar-13	2	9	1	1		20-Sep-16	2	20-Sep-16	2
1XXX1160	15-Feb-51	06-Feb-14	2	3	2		1	17-Dec-14	2	17-Dec-14	2
1XXX9880	20-Jul-49	23-Jan-14	2	1	2		4	21-Aug-14	3	25-Jan-15	1
1XXX8920	08-Jun-52	09-May-13	1	1	1	2	4	30-Jan-15	3	16-Sep-15	1
1XXX5010	15-Apr-53	23-Mar-13	2	1	2		4	06-Jan-14	3	15-Apr-14	2
1XXX2730	30-Aug-33	11-Feb-13	1	1	2		3	01-Dec-13	3	15-Feb-14	1
1XXX4570	11-Nov-43	01-Mar-13	2	1	2		2	18-Apr-16	2	18-Apr-16	2
1XXX0430	18-Mar-50	11-Dec-12	2	1	2		3	11-Apr-13	3	26-May-14	1
1XXX0320	17-Feb-55	20-Aug-12	2	1	2		3	01-Jun-14	2	14-Mar-15	2
1XXX9880	20-Aug-48	10-Jul-12	2	1	2		4	23-Sep-15	3	23-Sep-15	2
1XXX3710	10-Mar-55	13-Jun-12	1	1	1	10	4	16-Jan-13	2	13-Jan-13	2
1XXX2990	12-Feb-28	27-Nov-12	1	3	1	6	2	09-Nov-15	2	09-Nov-15	2
1XXX2960	27-Mar-52	10-Jan-12	1	9	2			18-Feb-13	2	18-Feb-13	2
1XXX6790	17-Mar-70	05-Aug-11	2	1	2		1	30-Dec-15	2	30-Dec-15	2
1XXX7980	09-Nov-31	08-Sep-11	2	1	2		4	17-Aug-16	2	17-Aug-16	2
1XXX3450	12-Jan-60	24-Jun-11	1	1	1	23	4	12-Apr-13	2	12-Apr-13	2
1XXX0280	25-Jan-46	13-Oct-01	1	1	99		4	20-Aug-03	3	20-Aug-03	1

Figura 6: Dados dos pacientes presentes no banco de dados modelado. Foram recuperados dados de 954 pacientes com câncer de pâncreas do ICGC e TCGA e 24 pacientes do LGEGC. A- Projetos de Origem: O ICGC tem três projetos acerca de câncer de pâncreas, que foram incorporados ao banco de dados modelado, juntamente aos dados do LGEGC. B- Distribuição do sexo dos pacientes. C- Estado vital dos pacientes, sendo 384 vivos, 548 mortos e 47 sem informação. D- Idade ao Diagnóstico, mostrando uma faixa preferencial do câncer de pâncreas em pacientes de 65 a 80 anos. E- Sobrevida após o diagnóstico, em dias.

Figura 7: Dados dos experimentos presentes no banco de dados modelado. O ICGC tem 3 projetos acerca de câncer de pâncreas, que foram incorporados ao banco de dados modelado, juntamente aos dados do LGECC. Enquanto todos os projetos têm informações de expressão gênica (esquerda), apenas dois têm dados disponíveis de mutação somática simples, o PACA-AU e o PACA-CA. Há um total de 496 pacientes com dados de expressão gênica no CaRDIGAn e 483 pacientes com informações sobre mutação somática simples.

Figura 8: Sobrevida de pacientes com câncer de pâncreas do ICGC de acordo com a mutação em *KRAS*. Observou-se a presença da mutação em 89 pacientes e ausência em 9 casos, totalizando aproximadamente 91% de pacientes mutados. Não se observou significância estatística entre a presença da mutação de *KRAS* e a sobrevida dos pacientes analisados.

Desenvolvimento de consultas de interesse

O CaRDIGAn contém dados de experimentos de mutação somática simples e expressão gênica. Os dados podem ser combinados conforme desejado. Assim, se um espécime foi utilizado

para mais de um tipo de experimento, seus resultados podem ser combinados e relacionados aos dados clínicos do paciente.

Conforme citado anteriormente, a mutação de *KRAS* é muito prevalente em câncer PDAC (Iacobuzio-Donahue 2012). O oncogene *KRAS* regula diversos genes. Assim, pode-se perguntar se a presença de mutação em *KRAS* afeta o padrão de expressão gênica global dos tumores. Para fazer essa consulta no banco, deve-se inicialmente selecionar amostras vindas de um mesmo paciente que tenham feito parte tanto de experimentos de mutação somática simples como de expressão gênica. A Figura 9 representa essa consulta. A seguir, deve-se buscar quais amostras têm *KRAS* mutado e quais apresentam o gene sem mutação (selvagem). A partir daí pode-se criar uma tabela com os dados experimentais de expressão gênica, como *normalized read count*, *raw read count* e uma coluna com o tipo da mutação detectada em *KRAS* nessa amostra. A Figura 10 representa de forma esquematizada uma consulta no banco local para a recuperação da expressão de todos os genes em um subconjunto de pacientes que tenham mutação em um dado gene. A consulta para amostras sem a mutação é muito semelhante e, portanto, será omitida. O código utilizado para tal, em SQL, encontra-se no Anexo 3. Ressalta-se que as figuras de consultas mencionadas são esquemas meramente didáticos, pois o SGBD otimiza as consultas, realizando-as na ordem mais eficiente, que pode não ser a mais intuitiva.

Análise de sobrevida em pacientes do ICGC quanto à mutação de *KRAS*

Como exemplo de aplicação da consulta relacionando sobrevida e mutação em *KRAS*, foram geradas curvas de sobrevida de Kaplan Meier de pacientes com e sem mutação no gene, e analisado se estas são significativamente diferentes utilizando-se o teste *log-rank*. Inicialmente foram recuperados do CaRDIGAn os dados sobre a presença ou não de mutação em *KRAS* nos casos do ICGC. Observou-se a presença da mutação em 89 pacientes e ausência em 9 casos, totalizando aproximadamente 91% de pacientes mutados. Não se observou significância estatística entre a presença da mutação de *KRAS* e a sobrevida dos pacientes analisados. Os gráficos encontram-se na Figura 8.

Análise de genes diferencialmente expressos em pacientes do LGEGC

Os genes diferencialmente expressos no PDAC em relação ao tecido não tumoral adjacente dos pacientes analisados em nosso grupo foram selecionados de acordo com os seguintes critérios: *fold change* aumentado em pelo menos 10 vezes no tecido tumoral e *p*valor ajustado associado ao *fold change* $\leq 10^{-3}$. Os genes que atendiam a esses critérios foram então ordenados de acordo com o

valor de *base mean*, que representa a abundância média do gene em todas amostras avaliadas. Os 100 genes com maior *base mean* foram utilizados para a análise de sobrevida descrita a seguir.

Figura 9: Esquema de uma consulta ao banco local. Deseja-se identificar pacientes que tenham dados de experimentos de mutação simples e expressão gênica por RNASeq.

- 1- Seleciona-se amostras com dados de mutação simples na relação Experimento de Mutação Utiliza amostra
- 2- Busca-se essas amostras na Relação Amostra
- 3- Identifica-se os pacientes que apresentam dados de mutação simples (no exemplo, pacientes 1, 2 e 4)
- 4- Seleciona-se amostras com dados de expressão gênica por sequenciamento na relação Experimento de Expressão Gênica Utiliza Amostra
- 5- Busca-se essas amostras na relação Amostra
- 6- Identifica-se os pacientes que apresentam dados de expressão por sequenciamento (no exemplo, pacientes 1, 2 e 3)
- 7- Compara-se os dois conjuntos de pacientes obtidos, fazendo-se uma intersecção dos dois (no exemplo, os pacientes que atendem aos critérios da consulta são o paciente 1 e 2)

Figura 10: Esquema de uma consulta ao banco local. Deseja-se obter os dados de expressão gênica global de pacientes com mutação em um gene. Assume-se que já se sabe quais pacientes (1 e 2) tiveram amostras utilizadas nos experimentos de mutação simples e expressão gênica por sequenciamento (demonstrado na Figura 9).

- 1- Seleciona-se o identificador do gene desejado na relação 'Mutação Simples Avalia Gene em Amostra' e, opcionalmente, a mutação resultante na proteína
- 2- Identifica-se a amostra com a mutação
- 3- Busca-se na relação 'Amostra' os pacientes correspondentes às amostras selecionadas no passo anterior
- 4- Descarta-se, entre os pacientes com dados dos dois experimentos (1 e 2), aqueles sem a mutação (2)
- 5- Seleciona-se amostras com dados de RNASeq na relação 'Sequenciamento Avalia Gene em Amostra'
- 6- Busca-se essas amostras na relação 'Amostra', selecionando as amostras com pacientes com a mutação
- 7- Seleciona-se os dados de expressão gênica de todos os genes para as amostras de pacientes que obedecem aos critérios de seleção (dados nos dois experimentos e mutação no gene desejado)

Análise da sobrevivência de pacientes do ICGC a partir de genes diferencialmente expressos em amostras do LGEGC

Os 100 genes diferencialmente expressos identificados nas amostras do nosso grupo foram analisados em um conjunto de 91 pacientes do projeto PACA-AU do ICGC. As informações necessárias para a análise foram obtidas a partir de uma consulta ao CaRDIGAN onde se obteve as informações clínicas dos pacientes e a contagem de *reads* brutos de cada gene. A saída foi processada por um *script* em Python que transformou a saída da consulta em uma matriz com os pacientes nas

linhas e as informações clínicas, juntamente com a contagem de *reads* para cada gene, nas colunas. Foram obtidas as contagens de *reads* para 55.562 identificadores de genes do Ensembl. Após a eliminação de genes com contagem muito baixa (valores de contagem por milhão menores que 1), e considerando que as profundidades de cobertura não foram necessariamente iguais para cada amostra, procedeu-se a normalização dos dados para tornar as amostras comparáveis, utilizando a função *voon* do pacote *limma* do Bioconductor. A função assume uma distribuição normal levando em consideração a tendência da variância média. Após a etapa de filtragem e normalização prosseguiu-se a análise com 98 genes.

Os genes de interesse foram então analisados para avaliar seu impacto na sobrevida de pacientes com PDAC através de curvas de Kaplan-Meier seguido de análise estatística pelo teste *log-rank*. Para cada gene, os pacientes foram agrupados em relação ao valor da mediana da expressão do gene nas amostras, sendo os que se encontram acima da mediana considerado como o grupo com hiperexpressão e abaixo desta como o grupo com hipoexpressão. Onze dos genes selecionados apresentaram significância estatística de diferença nas curvas de sobrevida de pacientes com hiper ou hipoexpressão, conforme representado na Figura 11. Os genes identificados foram *B3GNT3*, *FOXQ1*, *GCNT3*, *HIST3H2BB*, *ISG15*, *LY6D*, *MUC13*, *PI3*, *S100A6*, *S100P* e *SMIM24*.

Figura 11: Sobrevida de pacientes com câncer de pâncreas do ICGC de acordo com a expressão de genes diferencialmente expressos em PDAC identificados em nosso laboratório. Noventa e oito genes diferencialmente expressos identificados nas amostras do LGEGC foram avaliados em pacientes do projeto PACA-AU do ICGC. A figura mostra os 11 genes obtidos que apresentaram diferença estatisticamente significativa nas curvas de sobrevida entre os pacientes com hiper (linhas vermelhas) e hipo (linhas azuis) expressão de tais genes. As sombras vermelhas e azuis representam o intervalo de confiança de seus respectivos grupos. Os grupos de pacientes com 47 hiper/hipoexpressão foram definidos com base no valor da expressão mediana do gene no conjunto de amostras.

Discussão

A integração de dados de diferentes instituições e experimentos é um desafio importante na biologia molecular (Fortney e Jurisica 2011). No presente estudo buscou-se construir uma ferramenta para auxiliar a exploração de dados públicos de câncer, bem como dados gerados pelo próprio laboratório. Abaixo são discutidas as principais etapas e contribuições do trabalho.

Recuperação de dados públicos

A recuperação de dados públicos contou com abordagens bem estabelecidas em aplicações computacionais de transmissão de dados entre servidores, como *sftp* e *wget* (“Bash Reference Manual” 2015), além de pacotes e APIs direcionados para a área biológica, como o Bioconductor (Huber et al. 2015). A combinação dessas metodologias com uma linguagem de programação de auto nível, no caso Python (van Rossum 2011), permitiu o desenvolvimento de *scripts* capazes de realizar a tarefa de recuperação dos dados públicos de maneira semi-automatizada a partir de dois bancos de interesse: Ensembl (Aken et al. 2016) e ICGC (Junjun Zhang et al. 2011), que contém o TCGA (“The Cancer Genome Atlas Home Page” 2015).

Modelagem do banco local

Embora existam modelos de bancos de dados genômicos disponíveis, como o Chado (Mungall, Emmert, e FlyBase Consortium 2007), observou-se que sua adaptação às necessidades do laboratório seria mais custosa do que o desenvolvimento de um novo modelo. Isso se dá por dois motivos principais. O primeiro é que o Chado é um modelo principalmente voltado para anotação genômica, taxonômica e filogenética, tendo diversos módulos que não seriam utilizados. Embora comporte anotação de resultados experimentais, certos módulos teriam que ser adaptados para capturar as informações referentes ao desenho de cada experimento que nosso grupo pretendia analisar. Em segundo lugar, a estrutura já disponível do Chado não prevê o armazenamento de dados clínicos, um dos pontos de maior interesse de nosso grupo. Assim, optou-se pela elaboração de um novo modelo de dados que atendesse às necessidades de nosso grupo.

O CaRDIGAn foi modelado com foco na integração dos dados dos bancos públicos mencionados anteriormente e de dados gerados em nosso laboratório. Porém, o modelo foi desenvolvido de maneira genérica o suficiente para acomodar dados de outras fontes, após escrita de novos *scripts* de ETL. Assim, o modelo desenvolvido é versátil, constituindo uma ferramenta interessante para integração de dados de fontes variadas, um dos benefícios do uso de banco de dados na área biológica mencionado por Fortney e Jurisica (2011).

Outra vantagem do modelo é que a representação dos genes, transcritos e suas respectivas coordenadas permite que os resultados experimentais possam ser remapeados quando novas montagens estiverem disponíveis ou mesmo quando um novo gene for identificado. Ao contrário de bancos como o BioXpress (Wan et al. 2015), onde se pesquisa ou se obtém resultados relacionados a genes, a abordagem desenvolvida no presente estudo permite que se busque coordenadas gênicas de interesse, levando a uma maior flexibilidade de consultas possíveis.

A modelagem do banco dados utilizando o modelo entidade relacionamento estendido permitiu captar as particularidades dos dados biológicos de interesse, seguindo-se as abstrações de dados (Elmasri e Navathe 2010). O modelo conseguiu representar os dados clínicos e experimentais de pacientes bem como as referências de modelos gênicos.

Povoamento do CaRDIGAn

Um dos fatores limitantes para o presente estudo foi a curva de aprendizado para a escrita dos *scripts* de ETL. Inicialmente as transformações de dados seriam feitas com auxílio da ferramenta *Kettle* do pacote Pentaho (*Pentaho Data Integration (aka Kettle)* (versão 7.1) 2017). A ferramenta, porém, não conseguiu lidar com as restrições do PostgreSQL, o SGBD utilizado, e não fazia atualizações de mais de 10.000 tuplas. Assim, foi abandonada e foi utilizado SQL puro para essa função.

Boas práticas no desenvolvimento de bancos de dados complexos como os que abrigam dados biológicos associados a dados de pacientes foram estudadas por Takecian e colaboradores (2013). Após a identificação das principais entidades, relacionamentos e atributos, deve-se criar um modelo conceitual e identificar as restrições semânticas dos dados. Se surgirem erros, deve-se corrigir o modelo conceitual e corrigir ou identificar novas restrições, repetindo esse ciclo. Assim, fica claro que o processo de criação de um banco eficiente é um processo iterativo. De fato, ao se elaborar os *scripts* de ETL foi necessário revisitar o modelo para melhor captar os dados disponíveis. O banco modelado sofreu diversas modificações desde a primeira vez que o modelo conceitual foi concebido.

O modelo proposto abordou a questão da grande redundância nas tabelas de dados experimentais do ICGC separando os dados referentes ao experimento em si de seus resultados. Para que o mapeamento das entidades do tipo *Experimento X Avalia Amostra* e *Experimento X Avalia Gene em Amostra* fosse bem-sucedido, inicialmente elaborou-se um *script* de ETL para cada uma dessas relações, contendo um passo em que se observava todos os dados experimentais de todas as tuplas da tabela de origem e os comparava com os presentes na entidade *Experimento X* para obter o identificador do experimento e na entidade ‘Amostra’ para obter o identificador da amostra. No caso das relações *Experimento X Avalia Gene em Amostra* era necessário, ainda, a busca do identificador

da tabela de origem na relação ‘Gene’. A abordagem inicial buscava, portanto, realizar os passos mencionados em uma consulta única para cada mapeamento, o que não foi computacionalmente eficiente, já que as tabelas originais têm cerca de dezenas de milhões de linhas. A abordagem que se provou mais eficaz foi a inserção dos identificadores de experimento, gene e mostra de forma independente nas tabelas experimentais originais. Ao invés de inserir os identificadores de todas as tuplas em uma única consulta, as inserções foram divididas em blocos de 1.000.000 de linhas. Como não havia concorrência do SGBD para escrita em colunas diferentes, a tarefa foi paralelizada, sendo inseridos os identificadores simultaneamente. Uma vez que as tabelas com dados experimentais originais foram acrescidas dos identificadores das tabelas de destino, os mapeamentos passaram a ser feitos em menos de 20min.

O desenvolvimento da abordagem descrita pôde, no tempo hábil do projeto, ser aplicado aos dados de experimentos de mutação somática simples e expressão gênica por sequenciamento. Novos *scripts* de ETL podem ser definidos e executados para os demais dados públicos disponíveis e, ainda, para dados de outros tumores.

Sabe-se que cada grupo organiza seus dados em tabelas de diferentes formatos e o mapeamento para o banco de dados é único para cada uma. Como exemplo de aplicabilidade do modelo proposto, realizamos o mapeamento com os dados gerados por nosso grupo em colaboração com o AC Camargo Cancer Center. O povoamento desses dados foi feito a partir de uma curagem manual dos dados anotados, seguida do carregamento com *scripts* de ETL. O processo demonstra que é o modelo elaborado atende às necessidades de um laboratório que gera e/ou analisa dados de NGS de câncer, bem como dados clínicos associados.

O modelo permite, ainda, que dados de outros tumores, também disponíveis no ICGC e TCGA sejam incorporados, permitindo comparação entre si.

Consultas de interesse

Foram feitas algumas consultas em SQL a fim de testar a funcionalidade do CaRDIGAn para a integração de dados gerados localmente e dados recuperados do ICGC. Conseguiu-se relacionar, com sucesso, dados clínicos e experimentais inseridos no CaRDIGAn. Como exemplo, foi possível consultar o *status* da mutação de um gene de interesse relacionando-a com os dados clínicos de pacientes. Também foi demonstrado o uso do banco para consultar a expressão global de genes de um subgrupo de pacientes de interesse, no caso, que apresentavam mutação no proto-oncogene *KRAS*. Os dados de expressão puderam ser relacionados à dados clínicos para análises de sobrevida. As consultas mencionadas foram concluídas em minutos, demonstrando que a implementação realizada

foi eficiente para a realização de consultas complexas, apesar do grande volume e dimensionalidade dos dados armazenados. Abaixo são discutidos os resultados obtidos nas consultas realizadas.

Análise de sobrevida em pacientes do ICGC quanto à mutação de *KRAS*

Os pacientes foram separados conforme a presença ou ausência de mutação nas posições G12A, G12C, G12D, G12R, G12S, G12V e Q61H da proteína codificada pelo gene *KRAS*. Obteve-se informações de 98 pacientes, dos quais 89 (91%) tinham ao menos uma das sete mutações de interesse. Tais dados estão de acordo com a observada na literatura (Iacobuzio-Donahue 2012). A mutação, *per se*, não teve impacto na sobrevida dos pacientes estudados.

Análise da sobrevida de pacientes do ICGC conforme a expressão de genes diferencialmente expressos em amostras do LGEGC

Partindo-se de um conjunto de 100 genes diferencialmente expressos selecionados no conjunto de genes do LGEGC, foi possível validar o impacto na sobrevida de pacientes com câncer de pâncreas de 11 genes.

Um dos genes validados, *MUC13* (*mucin 13, cell surface associated*) é um gene de particular interesse para o grupo. O CA19-9, um dos marcadores utilizados atualmente para exames laboratoriais de detecção de câncer de pâncreas é uma mucina. Por serem marcadores extracelulares que podem ser clivadas, as mucinas podem ter interesse para diagnóstico e prognóstico de câncer de pâncreas em exames não invasivos (Kufe, 2009). O potencial como marcador de prognóstico de *MUC13*, bem como de outros alvos, deverá ser estudado por nosso grupo através de um *tissue microarray*, como parte de um projeto de doutorado em andamento no laboratório desenvolvido pelo aluno Vinicius F Paixão. Trabalhos anteriores detectaram a proteína MUC13 em lesões pré-neoplásicas de pâncreas e em PDAC. A presença de MUC13 nuclear foi relacionada a menor sobrevida de pacientes, metástase em linfonodos e invasão tumoral a tecidos adjacentes (Khan et al. 2018). Em nosso trabalho, o aumento da expressão de *MUC13* parece estar relacionado a um melhor prognóstico (maior sobrevida).

Outro gene candidato que apresentou associação significativa com a sobrevida foi *B3GNT1*, em que a expressão aumentada se correlacionou a uma maior sobrevida dos pacientes. Este resultado está de acordo com um trabalho da literatura que descreveu a expressão diminuída de *B3GNT1* como associada a um pior prognóstico (Haider et al. 2014).

O *FOXQ1* (*forkhead box Q1*) é um gene que codifica um fator de transcrição de mesmo nome. Estudos anteriores em câncer de pâncreas mostraram a importância do aumento de expressão desse

gene para a manutenção do estado semelhante ao de células tronco, implicando em maior agressividade e capacidade de geração de tumor em modelos *in vivo* (Bao et al. 2014). Um trabalho posterior mostrou a associação do aumento da expressão gênica de *FOXQ1* com piora do prognóstico em pacientes com câncer de pâncreas, confirmando a presença aumentada da proteína por imuno-histoquímica (Zhan et al. 2015). Ambos os trabalhos divergem do que foi observado no presente estudo, onde a hiperexpressão de *FOXQ1* se mostrou associado com uma sobrevida maior dos pacientes do projeto PACA-AU do ICGC.

O gene *GCNT3* (*glucosaminyl (N-acetyl) transferase 3, mucin type*) codifica é uma proteína que catalisa a formação de dois compostos importantes para a formação de glicoproteínas, contribuindo para a formação de muco. Seu papel em câncer de pâncreas já foi estudado anteriormente. O grupo de Rao et al. (2016) avaliou a expressão da proteína GCNT3 por imuno-histoquímica em 88 pacientes, relacionando a baixa expressão da proteína a uma maior sobrevida. Em um modelo animal de ratos, o aumento da expressão de *GCNT3* se mostrou relacionando aumento de expressão de diversas mucinas (*Muc4*, *Muc5ac*, *Muc6*, *Muc1*, *Muc16* e *Muc20*), além de *EGFR*. Ao tratar esses ratos com um inibidor de *GCNT3*, observou-se diminuição da expressão de mucinas e maior resposta terapêutica com gefitinib, uma droga usada para o tratamento de tumores com aumento de expressão ou superatividade de *EGFR*. Em nosso estudo, a hiperexpressão de *GCNT3* se mostrou associada a uma maior sobrevida de pacientes com câncer de pâncreas, dado que diverge do estudo mencionado.

O gene *HIST3H2BB* (*histone cluster 3 H2B family member b*) codifica uma histona da família H2B. Seu papel no câncer de pâncreas ainda não foi estudado aprofundadamente. Em um trabalho onde realizou-se a análise transcritômica de amostras de câncer colorretal do TCGA foi mostrado o aumento da expressão do gene *HIST3H2BB* conforme o aumento do grau tumoral. Os autores ressaltam que as comparações foram feitas apenas entre amostras de diferentes graus tumorais, sem amostras de tecido não neoplásico como controle, o que é um fator limitante para o estudo (Huo et al. 2017). Em nosso estudo a expressão aumentada desse gene parece estar associada ao aumento da sobrevida dos pacientes analisados.

O *ISG15* (*ISG15 ubiquitin-like modifier*) codifica uma proteína *ubiquitina-like* que se liga a alvos intracelulares a partir de estímulos por interferon-alfa e interferon-beta. A proteína ISG15 tem diversos papéis associados a ela, como atividade quimiotática, sinalização célula-a-célula e atividade antiviral. Embora geralmente esteja ligada de forma não-covalente a filamentos intermediários, pode ser secretada ocasionalmente. Um trabalho mostrou a importância da ISG15 secretada por macrófagos

associados ao tecido tumoral (TAMs, *tumor-associated macrophages*) como fator necessário no microambiente para a manutenção de células tronco tumorais. Embora tenham observado expressão de *ISG15* em células de PDAC propriamente ditas, a maior parte da expressão desse gene provém de células imunológicas como os TAMs. A presença de *ISG15* foi associada a tumorigênese e aumento da migração em PDAC (Sainz et al. 2014). Em nosso estudo as amostras dos pacientes não foram separadas quanto a celularidade ou presença de células imunológicas. Observamos que o aumento da expressão de *ISG15* se relacionou com o aumento da sobrevida dos pacientes, o que parece divergir dos achados mencionados.

O *LY6D* (*lymphocyte antigen 6 family member D*) codifica uma proteína de mesmo nome expressa um antígeno que marca o início da diferenciação de linfócitos a células B. O trabalho de Luo et al. (2016) usou dados disponíveis publicamente do *gene expression omnibus* (GEO) e do *Georgetown Database of Cancer* (G-DOC) sobre diversos tipos de câncer, entre eles o câncer de pâncreas. O *LY6D* se mostrou com expressão aumentada em câncer de pâncreas, bexiga, encéfalo e sistema nervoso central, mama, ovário, pulmão, gástrico e cabeça e pescoço quando comparado ao tecido normal correspondente. A expressão desse gene teve impacto na sobrevida de pacientes com câncer de mama, encéfalo e sistema nervoso central, gástrico e ovário. Para todos as neoplasias onde a expressão do gene *LY6D* teve impacto na sobrevida dos pacientes a maior expressão do gene foi associada com menor sobrevida dos pacientes. Em nosso estudo a maior expressão do *LY6D* foi associada a uma maior sobrevida dos pacientes.

O *PI3* (*peptidase inhibitor 3*) codifica uma proteína de mesmo nome que inibe especificamente a elastase, agindo como um peptídio microbiano contra bactérias gram-positivas, gram-negativas e fungos. Sabe-se que responde a estímulo de lipopolisacarídeos e citocinas. Seu papel no câncer ainda não foi estudado até o presente momento, sendo um campo totalmente novo para investigação. Em nosso trabalho a expressão aumentada de *PI3* se mostrou relacionada com maior sobrevida dos pacientes estudados.

O *S100A6* (*S100 calcium binding protein A6*) codifica uma proteína de mesmo nome que faz parte da família S100 de proteínas, que possui pelo menos 13 membros com genes localizados no braço longo do cromossomo 1. Essa família de proteínas está envolvida com diversos processos celulares como progressão do ciclo celular e diferenciação. A *S100A6* proteína parece estar ligada a liberação de insulina dependente de cálcio, estímulo de secreção de prolactina e exocitose. Estudos funcionais desse gene mostraram que a *S100A6* é capaz de induzir transição epitélio-mesenquimal, promovendo migração celular e invasão de maneira dependente de β -catenina (Chen et al. 2015). Em

nosso estudo, o aumento da expressão de *S100A6* se mostrou relacionado a uma maior sobrevida dos pacientes.

O *S100P* (*S100 calcium binding protein P*) também pertence à família S100 de proteínas, embora seu locus esteja presente no cromossomo 4. É uma proteína que, além de cálcio, realiza ligação de zinco e magnésio. O gene *S100P* já foi descrito como diferencialmente expresso em câncer de pâncreas (Li et al. 2017) e supõe-se que pode ser um bom marcador para o diagnóstico de câncer de pâncreas a partir de material coletado de biópsias de agulha fina (Haider et al. 2014). Nosso estudo mostrou que o aumento da expressão de *S100P* está relacionado a uma maior sobrevida dos pacientes em relação a pacientes com hipoeexpressão desse gene.

O *SMIM24* (*small integral membrane protein 24*) é um gene pouco conhecido cujo papel no câncer ainda não foi estudado até o presente momento. No presente trabalho a expressão aumentada de *SMIM24* se mostrou relaciona ao aumento da sobrevida dos pacientes analisados.

Assim, o CaRDIGAn, se mostra uma ferramenta de interesse para laboratórios que buscam validar a expressão de alvos de interesse em um *dataset* com um grande número de pacientes e variedade de tumores, como é o caso do ICGC e TCGA.

Conclusões

O CaRDIGAn, *Cancer Relational Database for Integration and Genomic Analysis*, banco de dados modelado e implementado ao longo deste trabalho é uma ferramenta capaz de relacionar dados clínicos e dados experimentais de pacientes de câncer e integrar dados de fontes heterogêneas, com potencial para auxiliar na priorização de genes candidatos para caracterização detalhada.

Embora tenha sido modelado para armazenar dados genômicos e clínicos obtidos dos principais repositórios de dados públicos de câncer (ICGC/TCGA, Ensembl) e dados gerados em nosso laboratório a partir de amostras de pacientes com câncer de pâncreas, sua estrutura é genérica e pode suportar dados de qualquer fonte, necessitando apenas o desenvolvimento de *scripts* de ETL específicos, sendo potencialmente útil para outros grupos de pesquisa.

Foi concluído com sucesso o povoamento do banco local com os dados de expressão gênica e dados clínicos de pacientes com câncer de pâncreas analisados em nosso laboratório, além de dados de coordenadas e identificadores genômicos do Ensembl. O povoamento do banco local com dados recuperados do ICGC se mostrou desafiador devido ao grande volume e no tempo disponível para a conclusão do trabalho foram povoados apenas os dados de experimentos de mutação somática simples e expressão gênica por sequenciamento, além de dados clínicos de pacientes com câncer de pâncreas.

A funcionalidade do CaRDIGAn foi demonstrada através da realização de consultas complexas que integraram dados gerados localmente e dados recuperados do ICGC e retornaram o resultado em poucos minutos. Como exemplo, a partir de uma lista de genes diferencialmente expressos em tumores de pâncreas identificados em nosso laboratório, foram recuperados dados de expressão gênica e clínicos de pacientes do ICGC. Os dados recuperados foram utilizados em uma análise de sobrevida que resultou na identificação de 11 genes com potencial prognóstico no câncer de pâncreas, ilustrando o potencial da ferramenta para identificação de genes candidatos para futura validação experimental. A possibilidade de recuperação de dados genômicos e clínicos de pacientes com cancer de pâncreas de diferentes fontes primarias, de uma forma estruturada permite a automatização de análises de expressão diferencial e de sobrevida, que facilitará o desenvolvimento de diferentes estudos de busca de marcadores tumorais em andamento no laboratório.

Como perspectiva, o banco de dados modelado pode servir como base para o desenvolvimento de aplicações de aprendizado de máquina e outras técnicas operacionais para extração de informações relevantes para identificação de biomarcadores para diagnóstico molecular em câncer.

Referências

- Aaltonen, LA, e SR Hamilton. 2000. *Pathology and genetics of tumours of the digestive system*. Lyon: Oxford University Press.
- Aken, Bronwen L., Sarah Ayling, Daniel Barrell, Laura Clarke, Valery Curwen, Susan Fairley,

- Julio Fernandez Banet, et al. 2016. “The Ensembl gene annotation system”. *Database: The Journal of Biological Databases and Curation* 2016 (junho).
<https://doi.org/10.1093/database/baw093>.
- Anders, Simon, e Wolfgang Huber. 2010. “Differential Expression Analysis for Sequence Count Data”. *Genome Biology* 11 (10): R106. <https://doi.org/10.1186/gb-2010-11-10-r106>.
- Arnaiz, Olivier, e Linda Sperling. 2011. “ParameciumDB in 2011: New Tools and New Data for Functional and Comparative Genomics of the Model Ciliate Paramecium Tetraurelia”. *Nucleic Acids Research* 39 (Database issue): D632-636. <https://doi.org/10.1093/nar/gkq918>.
- Bai, Xiao-Feng, Xiao-Guang Ni, Ping Zhao, Shang-Mei Liu, Hui-Xin Wang, Bing Guo, Lan-Ping Zhou, et al. 2004. “Overexpression of Annexin 1 in Pancreatic Cancer and Its Clinical Significance”. *World Journal of Gastroenterology: WJG* 10 (10): 1466–70.
- Bailey, Peter, David K. Chang, Katia Nones, Amber L. Johns, Ann-Marie Patch, Marie-Claude Gingras, David K. Miller, et al. 2016. “Genomic Analyses Identify Molecular Subtypes of Pancreatic Cancer”. *Nature* 531 (7592): 47–52. <https://doi.org/10.1038/nature16965>.
- Bao, Bin, Asfar S. Azmi, Amro Aboukameel, Aamir Ahmad, Aliccia Bolling-Fischer, Seema Sethi, Shadan Ali, et al. 2014. “Pancreatic Cancer Stem-like Cells Display Aggressive Behavior Mediated via Activation of FoxQ1”. *The Journal of Biological Chemistry* 289 (21): 14520–33. <https://doi.org/10.1074/jbc.M113.532887>.
- “Bash Reference Manual”. [s.d.]. Acessado 10 de agosto de 2015.
<https://www.gnu.org/software/bash/manual/bashref.html>.
- Benson, Dennis A, Mark Cavanaugh, Karen Clark, Ilene Karsch-Mizrachi, David J Lipman, James Ostell, e Eric W Sayers. 2013. “GenBank”. *Nucleic acids research* 41 (Database issue): D36-42. <https://doi.org/10.1093/nar/gks1195>.
- Benson, Dennis A., Karen Clark, Ilene Karsch-Mizrachi, David J. Lipman, James Ostell, e Eric W. Sayers. 2015. “GenBank”. *Nucleic Acids Research* 43 (Database issue): D30-35.
<https://doi.org/10.1093/nar/gku1216>.
- Biankin, Andrew V, Nicola Waddell, Karin S Kassahn, Marie-Claude Gingras, Lakshmi B Muthuswamy, Amber L Johns, David K Miller, et al. 2012. “Pancreatic cancer genomes reveal aberrations in axon guidance pathway genes”. *Nature* 491 (7424): 399–405.
<https://doi.org/10.1038/nature11547>.
- “BioXpress Tools & Resources”. [s.d.]. Acessado 4 de agosto de 2017.
<https://hive.biochemistry.gwu.edu/bioexpress>.
- Birney, Ewan, e Michele Clamp. 2004. “Biological Database Design and Implementation”. *Briefings in Bioinformatics* 5 (1): 31–38.
- “brModelo - Ferramenta de Ensino: Modelagem de Dados (MER)”. [s.d.]. Acessado 19 de abril de 2013. <http://sis4.com/brModelo/Default.aspx>.
- Chen, Xue, Xinjuan Liu, Haibo Lang, Shiqi Zhang, Yanlin Luo, e Jie Zhang. 2015. “S100 Calcium-Binding Protein A6 Promotes Epithelial-Mesenchymal Transition through β -Catenin in Pancreatic Cancer Cell Line”. *PloS One* 10 (3): e0121319.
<https://doi.org/10.1371/journal.pone.0121319>.
- Choi, Hyungwon, e Norman Pavelka. 2012. “When One and One Gives More than Two: Challenges and Opportunities of Integrative Omics”. *Frontiers in Genetics* 2 (janeiro).
<https://doi.org/10.3389/fgene.2011.00105>.
- Cokelaer, Thomas, Dennis Pultz, Lea M. Harder, Jordi Serra-Musach, e Julio Saez-Rodriguez. 2013. “BioServices: A Common Python Package to Access Biological Web Services Programmatically”. *Bioinformatics (Oxford, England)* 29 (24): 3241–42.
<https://doi.org/10.1093/bioinformatics/btt547>.
- “DCC Data Releases | ICGC Data Portal”. [s.d.]. Acessado 6 de março de 2018.
<https://dcc.icgc.org/releases>.
- draw.io. 2017. 6.9.9. Londres, Inglaterra: JGraph Ltd. www.draw.io.

- Duffy, M J, C Sturgeon, R Lamerz, C Haglund, V L Holubec, R Klapdor, A Nicolini, O Topolcan, e V Heinemann. 2010. "Tumor Markers in Pancreatic Cancer: A European Group on Tumor Markers (EGTM) Status Report". *Annals of Oncology: Official Journal of the European Society for Medical Oncology / ESMO* 21 (3): 441–47. <https://doi.org/10.1093/annonc/mdp332>.
- Elmasri, Ramez, e Shamkant Navathe. 2010. *Fundamentals of Database Systems*. 6ª edição. Boston, USA: Pearson Education, Inc.
- "Expression Atlas – EMBL-EBI". [s.d.]. Acessado 4 de agosto de 2017. <http://www.ebi.ac.uk/gxa/download.html>.
- Forbes, S. A., G. Bhamra, S. Bamford, E. Dawson, C. Kok, J. Clements, A. Menzies, J. W. Teague, P. A. Futreal, e M. R. Stratton. 2008. "The Catalogue of Somatic Mutations in Cancer (COSMIC)". *Current Protocols in Human Genetics / Editorial Board, Jonathan L. Haines ... [et Al.]* Chapter 10 (abril): Unit 10.11. <https://doi.org/10.1002/0471142905.hg1011s57>.
- Fortney, Kristen, e Igor Jurisica. 2011. "Integrative computational biology for cancer research". *Human Genetics* 130 (4): 465–81. <https://doi.org/10.1007/s00439-011-0983-z>.
- "GenBank and WGS Statistics". [s.d.]. Acessado 1º de agosto de 2017. <https://www.ncbi.nlm.nih.gov/genbank/statistics/>.
- "Genome List". [s.d.]. Acessado 1º de agosto de 2017. <https://www.ncbi.nlm.nih.gov/genome/browse/#>.
- Golfarelli, Matteo, e Stefano Rizzi. 2009. *Data Warehouse Design: Modern Principles and Methodologies*. McGraw Hill Professional.
- Haider, Syed, Jun Wang, Ai Nagano, Ami Desai, Prabhu Arumugam, Laurent Dumartin, Jude Fitzgibbon, et al. 2014. "A multi-gene signature predicts outcome in patients with pancreatic ductal adenocarcinoma". *Genome Medicine* 6 (12). <https://doi.org/10.1186/s13073-014-0105-3>.
- Hidalgo, Manuel. 2010. "Pancreatic cancer". *The New England journal of medicine* 362 (17): 1605–17. <https://doi.org/10.1056/NEJMra0901557>.
- Huber, Wolfgang, Vincent J. Carey, Robert Gentleman, Simon Anders, Marc Carlson, Benilton S. Carvalho, Hector Corrada Bravo, et al. 2015. "Orchestrating High-Throughput Genomic Analysis with Bioconductor". *Nature Methods* 12 (2): 115–21. <https://doi.org/10.1038/nmeth.3252>.
- Huo, Tianyao, Ronald Canepa, Andrei Sura, François Modave, e Yan Gong. 2017. "Colorectal Cancer Stages Transcriptome Analysis". *PloS One* 12 (11): e0188697. <https://doi.org/10.1371/journal.pone.0188697>.
- Iacobuzio-Donahue, Christine A. 2012. "Genetic Evolution of Pancreatic Cancer: Lessons Learnt from the Pancreatic Cancer Genome Sequencing Project". *Gut* 61 (7): 1085–94. <https://doi.org/10.1136/gut.2010.236026>.
- "INCA - Câncer - Tipo - Pâncreas". [s.d.]. Acessado 1º de agosto de 2017. <http://www2.inca.gov.br/wps/wcm/connect/tiposdecancer/site/home/pancreas>.
- Jones, Siân, Xiaosong Zhang, D Williams Parsons, Jimmy Cheng-Ho Lin, Rebecca J Leary, Philipp Angenendt, Parminder Mankoo, et al. 2008. "Core Signaling Pathways in Human Pancreatic Cancers Revealed by Global Genomic Analyses". *Science (New York, N.Y.)* 321 (5897): 1801–6. <https://doi.org/10.1126/science.1164368>.
- Kasprzyk, Arek. 2011. "BioMart: Driving a Paradigm Change in Biological Data Management". *Database: The Journal of Biological Databases and Curation* 2011: bar049. <https://doi.org/10.1093/database/bar049>.
- Khan, Sheema, Nadeem Zafar, Shabia S. Khan, Saini Setua, Stephen W. Behrman, Zachary E. Stiles, Murali M. Yallapu, et al. 2018. "Clinical Significance of MUC13 in Pancreatic Ductal Adenocarcinoma". *HPB: The Official Journal of the International Hepato Pancreato*

- Biliary Association, janeiro. <https://doi.org/10.1016/j.hpb.2017.12.003>.
- Kodama, Yuichi, Jun Mashima, Takehide Kosuge, Toshiaki Katayama, Takatomo Fujisawa, Eli Kaminuma, Osamu Ogasawara, Kousaku Okubo, Toshihisa Takagi, e Yasukazu Nakamura. 2015. “The DDBJ Japanese Genotype-Phenotype Archive for Genetic and Phenotypic Human Data”. *Nucleic Acids Research* 43 (Database issue): D18-22. <https://doi.org/10.1093/nar/gku1120>.
- Kulikova, Tamara, Ruth Akhtar, Philippe Aldebert, Nicola Althorpe, Mikael Andersson, Alastair Baldwin, Kirsty Bates, et al. 2007. “EMBL Nucleotide Sequence Database in 2006”. *Nucleic Acids Research* 35 (Database issue): D16–20. <https://doi.org/10.1093/nar/gk1913>.
- Leinonen, Rasko, Hideaki Sugawara, e Martin Shumway. 2011. “The Sequence Read Archive”. *Nucleic Acids Research* 39 (Database issue): D19-21. <https://doi.org/10.1093/nar/gkq1019>.
- Li, Hongzhe, Xinjing Wang, Yuan Fang, Zhen Huo, Xiongxiang Lu, Xi Zhan, Xiaxin Deng, Chenghong Peng, e Baiyong Shen. 2017. “Integrated Expression Profiles Analysis Reveals Novel Predictive Biomarker in Pancreatic Ductal Adenocarcinoma”. *Oncotarget* 8 (32): 52571–83. <https://doi.org/10.18632/oncotarget.16732>.
- Liu, Lin, Yinhu Li, Siliang Li, Ni Hu, Yimin He, Ray Pong, Danni Lin, Lihua Lu, e Maggie Law. 2012. “Comparison of Next-Generation Sequencing Systems”. *Journal of Biomedicine and Biotechnology* 2012. <https://doi.org/10.1155/2012/251364>.
- Luo, Linlin, Peter McGarvey, Subha Madhavan, Rakesh Kumar, Yuriy Gusev, e Geeta Upadhyay. 2016. “Distinct Lymphocyte Antigens 6 (Ly6) Family Members Ly6D, Ly6E, Ly6K and Ly6H Drive Tumorigenesis and Clinical Outcome”. *Oncotarget* 7 (10): 11165–93. <https://doi.org/10.18632/oncotarget.7163>.
- Maitra, Anirban, e Ralph H. Hruban. 2008. “Pancreatic Cancer”. *Annual review of pathology* 3: 157–88. <https://doi.org/10.1146/annurev.pathmechdis.3.121806.154305>.
- Marygold, Steven J., Madeline A. Crosby, Joshua L. Goodman, e FlyBase Consortium. 2016. “Using FlyBase, a Database of Drosophila Genes and Genomes”. *Methods in Molecular Biology (Clifton, N.J.)* 1478: 1–31. https://doi.org/10.1007/978-1-4939-6371-3_1.
- Matthaei, Hanno, Richard D Schulick, Ralph H Hruban, e Anirban Maitra. 2011. “Cystic precursors to invasive pancreatic cancer”. *Nature reviews. Gastroenterology & hepatology* 8 (3): 141–50. <https://doi.org/10.1038/nrgastro.2011.2>.
- Mungall, Christopher J., David B. Emmert, e FlyBase Consortium. 2007. “A Chado Case Study: An Ontology-Based Modular Schema for Representing Genome-Associated Biological Information”. *Bioinformatics (Oxford, England)* 23 (13): i337-346. <https://doi.org/10.1093/bioinformatics/btm189>.
- Navarro, J Daniel, Vidya Niranjan, Suraj Peri, Chandra Kiran Jonnalagadda, e Akhilesh Pandey. 2003. “From Biological Databases to Platforms for Biomedical Discovery”. *Trends in Biotechnology* 21 (6): 263–68.
- Oettle, Helmut, Stefan Post, Peter Neuhaus, Klaus Gellert, Jan Langrehr, Karsten Ridwelski, Harald Schramm, et al. 2007. “Adjuvant Chemotherapy with Gemcitabine vs Observation in Patients Undergoing Curative-Intent Resection of Pancreatic Cancer: A Randomized Controlled Trial”. *JAMA: The Journal of the American Medical Association* 297 (3): 267–77. <https://doi.org/10.1001/jama.297.3.267>.
- Peitzsch, Robert M. 2003. “Modeling biology using relational databases”. *Current protocols in bioinformatics / editorial board, Andreas D. Baxevanis ... [et al.]* Chapter 9 (fevereiro): Unit9.3. <https://doi.org/10.1002/0471250953.bi0903s00>.
- Pentaho Data Integration (aka Kettle) (versão 7.1). [s.d.]. Hitachi Group Company. Acessado 5 de agosto de 2017. <http://wiki.pentaho.com/display/EAI/Latest+Pentaho+Data+Integration+%28aka+Kettle%29+Documentation>.
- “pgAdmin III — pgAdmin III 1.20.0 documentation”. [s.d.]. Acessado 28 de agosto de 2015.

- <http://pgadmin.org/docs/1.20/index.html>.
- “PostgreSQL: The world’s most advanced open source database”. [s.d.]. Acessado 4 de agosto de 2015. <http://www.postgresql.org/>.
- Rao, Chinthalapally V., Naveena B. Janakiram, Venkateshwar Madka, Gaurav Kumar, Edgar J. Scott, Gopal Pathuri, Taylor Bryant, et al. 2016. “Small-Molecule Inhibition of GCNT3 Disrupts Mucin Biosynthesis and Malignant Cellular Behaviors in Pancreatic Cancer”. *Cancer Research* 76 (7): 1965–74. <https://doi.org/10.1158/0008-5472.CAN-15-2820>.
- Rossum, Guido van. 2011. *The Python Language Reference Manual*. England: Network Theory Limited. Available at <http://www.python.org>.
- Sainz, Bruno, Beatriz Martín, Marianthi Tatari, Christopher Heeschen, e Susana Guerra. 2014. “ISG15 Is a Critical Microenvironmental Factor for Pancreatic Cancer Stem Cells”. *Cancer Research* 74 (24): 7309–20. <https://doi.org/10.1158/0008-5472.CAN-14-1354>.
- Schlitter, Anna Melissa, Angela Segler, Katja Steiger, Christoph W. Michalski, Carsten Jäger, Björn Konukiewitz, Nicole Pfarr, et al. 2017. “Molecular, Morphological and Survival Analysis of 177 Resected Pancreatic Ductal Adenocarcinomas (PDACs): Identification of Prognostic Subtypes”. *Scientific Reports* 7 (fevereiro): 41064. <https://doi.org/10.1038/srep41064>.
- Shendure, Jay, e Hanlee Ji. 2008. “Next-generation DNA sequencing”. *Nature biotechnology* 26 (10): 1135–45. <https://doi.org/10.1038/nbt1486>.
- Takecian, Pedro L., Marcio K. Oikawa, Kelly R. Braghetto, Paulo Rocha, Fred Lucena, Katherine Kavounis, Karen S. Schlumpf, et al. 2013. “Methodological Guidelines for Reducing the Complexity of Data Warehouse Development for Transactional Blood Bank Systems”. *Decision Support Systems* 55 (3): 728–39. <https://doi.org/10.1016/j.dss.2013.02.008>.
- Tang, Yuan, Masaaki Horikoshi, e Wenxuan Li. 2016. “ggfortify: Unified Interface to Visualize Statistical Result of Popular R Packages”. *The R Journal*. <https://journal.r-project.org/archive/accepted/tang-horikoshi-li.pdf>.
- “The Cancer Genome Atlas Home Page”. [s.d.]. The Cancer Genome Atlas - National Cancer Institute. Acessado 31 de março de 2015. <http://cancergenome.nih.gov/>.
- Therneau, Terry M. 2015. *A Package for Survival Analysis in S* (versão 2.38). <https://CRAN.R-project.org/package=survival>.
- Toomula, Nishant, Arun Kumar, Sathish Kumar D, e Vijaya Shanti Bheemidi. 2011. “Biological Databases- Integration of Life Science Data”. *Journal of Computer Science & Systems Biology* 04 (05). <https://doi.org/10.4172/jcsb.1000081>.
- Vassiliadis, Panos, Alk Simitsis, e Spiros Skiadopoulos. 2002. “Conceptual modeling for ETL processes”. In *Proceedings of the 5th ACM international workshop on data warehousing and OLAP*, 14–21. McLean, Virginia. USA.
- Waddell, Nicola, Marina Pajic, Ann-Marie Patch, David K. Chang, Karin S. Kassahn, Peter Bailey, Amber L. Johns, et al. 2015. “Whole Genomes Redefine the Mutational Landscape of Pancreatic Cancer”. *Nature* 518 (7540): 495–501. <https://doi.org/10.1038/nature14169>.
- Wan, Quan, Hayley Dingerdissen, Yu Fan, Naila Gulzar, Yang Pan, Tsung-Jung Wu, Cheng Yan, Haichen Zhang, e Raja Mazumder. 2015. “BioXpress: an integrated RNA-seq-derived gene expression database for pan-cancer analysis”. *Database: The Journal of Biological Databases and Curation* 2015 (março). <https://doi.org/10.1093/database/bav019>.
- Zhan, H.-X., J.-W. Xu, L. Wang, D. Wu, G.-Y. Zhang, e S.-Y. Hu. 2015. “FoxQ1 Is a Novel Molecular Target for Pancreatic Cancer and Is Associated with Poor Prognosis”. *Current Molecular Medicine* 15 (5): 469–77.
- Zhang, Jun, Rod Chiodini, Ahmed Badr, e Genfa Zhang. 2011. “The impact of next-generation sequencing on genomics”. *Journal of genetics and genomics = Yi chuan xue bao* 38 (3): 95–109. <https://doi.org/10.1016/j.jgg.2011.02.003>.
- Zhang, Junjun, Joachim Baran, A. Cros, Jonathan M. Guberman, Syed Haider, Jack Hsu, Yong

- Liang, et al. 2011. "International Cancer Genome Consortium Data Portal--a one-stop shop for cancer genomics data". *Database: The Journal of Biological Databases and Curation* 2011 (setembro). <https://doi.org/10.1093/database/bar026>.
- Zhang, Lei, James J Farrell, Hui Zhou, David Elashoff, David Akin, No-Hee Park, David Chia, e David T Wong. 2010. "Salivary Transcriptomic Biomarkers for Detection of Resectable Pancreatic Cancer". *Gastroenterology* 138 (3): 949-957.e1-7. <https://doi.org/10.1053/j.gastro.2009.11.010>.

Anexo 1

Tabelas geradas a partir dos bancos públicos, carregadas para o banco intermediário, que servem de base para o banco local desenvolvido.

ICGC

TABELA icgc_donor

icgc_donor_id
project_code
study_donor_involved_in
submitted_donor_id
donor_sex
donor_vital_status
disease_status_last_followup
donor_relapse_type
donor_age_at_diagnosis_string
donor_age_at_enrollment_string
donor_age_at_last_followup_string
donor_relapse_interval_string
donor_diagnosis_icd10
donor_tumour_staging_system_at_diagnosis
donor_tumour_stage_at_diagnosis
donor_tumour_stage_at_diagnosis_supplementa
l
donor_survival_time_string
donor_interval_of_last_followup_string
prior_malignancy
cancer_type_prior_malignancy
cancer_history_first_degree_relative

TABELA icgc_donor_exposure

icgc_donor_id
project_code
submitted_donor_id
exposure_type
exposure_intensity
tobacco_smoking_history_indicator
tobacco_smoking_intensity_string
alcohol_history
alcohol_history_intensity

TABELA icgc_donor_family

icgc_donor_id
project_code
submitted_donor_id
donor_has_relative_with_cancer_history
relationship_type
relationship_type_other

relationship_sex
relationship_age_string
relationship_disease_icd10
relationship_disease

TABELA icgc_donor_therapy

icgc_donor_id
project_code
submitted_donor_id
first_therapy_type
first_therapy_therapeutic_intent
first_therapy_start_interval_string
first_therapy_duration_string
first_therapy_response
second_therapy_type
second_therapy_therapeutic_intent
second_therapy_start_interval_string
second_therapy_duration_string
second_therapy_response
other_therapy
other_therapy_response

TABELA icgc_sample

icgc_sample_id
project_code
submitted_sample_id
icgc_specimen_id
submitted_specimen_id
icgc_donor_id
submitted_donor_id
analyzed_sample_interval
percentage_cellularity
level_of_cellularity
study

TABELA

icgc_copy_number_somatic_mutation
icgc_donor_id
project_code
icgc_specimen_id
icgc_sample_id
matched_icgc_sample_id
submitted_sample_id
submitted_matched_sample_id
mutation_type

copy_number
 segment_mean
 segment_median
 chromosome
 chromosome_start
 chromosome_end
 assembly_version
 chromosome_start_range
 chromosome_end_range
 start_probe_id
 end_probe_id
 sequencing_strategy
 quality_score
 probability
 is_annotated
 verification_status
 verification_platform
 gene_affected
 transcript_affected
 gene_build_version
 platform
 experimental_protocol
 base_calling_algorithm
 alignment_algorithm
 variation_calling_algorithm
 other_analysis_algorithm
 seq_coverage
 raw_data_repository
 raw_data_accession

 TABELA icgc_exp_array
 icgc_donor_id
 project_code
 icgc_specimen_id
 icgc_sample_id
 submitted_sample_id
 analysis_id
 gene_model
 gene_id
 normalized_expression_value
 fold_change
 platform
 experimental_protocol
 normalization_algorithm
 other_analysis_algorithm
 raw_data_repository
 raw_data_accession
 reference_sample_type reference_sample_type

 TABELA icgc_exp_seq

icgc_donor_id
 project_code
 icgc_specimen_id
 icgc_sample_id
 submitted_sample_id
 analysis_id
 gene_model
 gene_id
 normalized_read_count
 raw_read_count
 fold_change
 assembly_version
 platform
 total_read_count_string
 experimental_protocol
 alignment_algorithm
 normalization_algorithm
 other_analysis_algorithm
 sequencing_strategy
 raw_data_repository
 raw_data_accession
 reference_sample_type

 TABELA icgc_mirna_seq
 icgc_donor_id
 project_code
 icgc_specimen_id
 icgc_sample_id
 submitted_sample_id
 analysis_id
 mirna_db
 mirna_id
 normalized_read_count
 raw_read_count
 fold_change
 is_isomir
 chromosome
 chromosome_start_string
 chromosome_end_string
 chromosome_strand
 assembly_version
 verification_status
 verification_platform
 sequencing_platform
 total_read_count_string
 experimental_protocol
 reference_sample_type
 alignment_algorithm
 normalization_algorithm
 other_analysis_algorithm

sequencing_strategy
raw_data_repository
raw_data_accession

TABELA icgc_protein_expression

icgc_donor_id
project_code
icgc_specimen_id
icgc_sample_id
submitted_sample_id
analysis_id
antibody_id
gene_name
gene_stable_id
gene_build_version
normalized_expression_level
verification_status
verification_platform
platform
experimental_protocol
raw_data_repository
raw_data_accession

TABELA icgc_simple_somatic_mutation

icgc_mutation_id
icgc_donor_id
project_code
icgc_specimen_id
icgc_sample_id
matched_icgc_sample_id
submitted_sample_id
submitted_matched_sample_id
chromosome
chromosome_start
chromosome_end
chromosome_strand
assembly_version
mutation_type
reference_genome_allele
mutated_from_allele
mutated_to_allele
quality_score
probability
total_read_count
mutant_allele_read_count
verification_status
verification_platform
biological_validation_status
biological_validation_platform
consequence_type

aa_mutation
cds_mutation
gene_affected
transcript_affected
gene_build_version
platform
experimental_protocol
sequencing_strategy
base_calling_algorithm
alignment_algorithm
variation_calling_algorithm
other_analysis_algorithm
seq_coverage
raw_data_repository
raw_data_accession
initial_data_release_date

TABELA icgc_specimen

icgc_specimen_id
project_code
study_specimen_involved_in
submitted_specimen_id
icgc_donor_id
submitted_donor_id
specimen_type
specimen_type_other
specimen_interval_string
specimen_donor_treatment_type
specimen_donor_treatment_type_other
specimen_processing
specimen_processing_other
specimen_storage
specimen_storage_other
tumour_confirmed
specimen_biobank
specimen_biobank_id
specimen_available
tumour_histological_type
tumour_grading_system
tumour_grade
tumour_grade_supplemental
tumour_stage_system
tumour_stage
tumour_stage_supplemental
digital_image_of_stained_section
percentage_cellularity_string
level_of_cellularity

TABELA icgc_splice_variant

icgc_donor_id

project_code	sv_id
icgc_specimen_id	placement
icgc_sample_id	annotation
submitted_sample_id	interpreted_annotation
analysis_id	chr_from
junction_id	chr_from_bkpt
gene_stable_id	chr_from_strand
gene_chromosome	chr_from_range
gene_strand	chr_from_flanking_seq
gene_start	chr_to
gene_end	chr_to_bkpt
assembly_version	chr_to_strand
second_gene_stable_id	chr_to_range
exon1_chromosome	chr_to_flanking_seq
exon1_number_bases	assembly_version
exon1_end	sequencing_strategy
exon1_strand	microhomology_sequence
exon2_chromosome	non_templated_sequence
exon2_number_bases	evidence
exon2_start	quality_score
exon2_strand	probability
is_fusion_gene	zygosity
is_novel_splice_form	verification_status
junction_seq	verification_platform
junction_type	gene_affected_by_bkpt_from
junction_read_count	gene_affected_by_bkpt_to
quality_score	transcript_affected_by_bkpt_from
probability	transcript_affected_by_bkpt_to
verification_status	bkpt_from_context
verification_platform	bkpt_to_context
gene_build_version	gene_build_version
platform	platform
experimental_protocol	experimental_protocol
base_calling_algorithm	base_calling_algorithm
alignment_algorithm	alignment_algorithm
normalization_algorithm	variation_calling_algorithm
other_analysis_algorithm	other_analysis_algorithm
sequencing_strategy	seq_coverage
seq_coverage	raw_data_repository
raw_data_repository	raw_data_accession
raw_data_accession	

TABELA icgc_structural_somatic_mutation

icgc_donor_id
project_code
icgc_specimen_id
icgc_sample_id
submitted_sample_id
submitted_matched_sample_id
variant_type

Ensembl

TABELA genes_GRCh37 e GRCh38
ensembl_gene_id
hgnc_id

external_gene_name
mirbase_id
chromosome_name
start_position
end_position
strand
source
status
version
gene_biotype

Anexo 2

Dicionário de dados do banco modelado. Cada tabela do modelo físico está indicada em negrito, com uma breve descrição do seu papel no banco, seguida dos atributos que a definem e seus respectivos domínios e descrição.

Gene		
Contém os identificadores de genes em diferentes bancos de dados		
Atributo	Domínio	Descrição
id_Gene	SERIAL	Chave artificial da tabela Gene
Banco	TEXT	Banco de dados que contém a anotação
Banco_alias	TEXT	Por razões de compatibilidade, esse atributo contém o nome do banco com a grafia utilizada pelo ICGC
Nome_no_banco	TEXT	Identificador do gene no banco
Coordenada_gene		
Contém as coordenadas genômicas dos genes de origem e versão variadas		
Atributo	Domínio	Descrição
id_coordenada	SERIAL	Chave artificial da tabela Coordenada_gene
Cromossomo	TEXT	Cromossomo
Inicio	TEXT	Posição de inicio do gene, medido em nucleotídeos
Fim	TEXT	Posição de término do gene, medido em nucleotídeos
Fita	INTEGER	Fita do DNA em que se encontra o gene
Referencia	TEXT	Nome do banco ou projeto que definiu a coordenada
Montagem	TEXT	Versão utilizada da referência

Source	TEXT	Fonte da anotação. Utilizada caso a referencia especifique uma origem externa para a anotação
Status	TEXT	Um gene ou transcrito <i>known</i> (conhecido) corresponde a uma sequência de uma base de dados pública, como o NCBI RefSeq. Caso a sequência seja nova ou tenha sido descrita para uma espécie diferente, o gene é classificado como <i>novel</i> (novo) para essa espécie. As fusões Ensembl-Havana, chamadas <i>merged</i> , ocorrem quando um transcrito anotado manualmente do projeto Havana coincide exatamente com a sequência do Ensembl
gene_biotype	TEXT	Classificação do biotipo, especificando se trata-se de rRNA, lincRNA, codificador de proteína, etc.
Gene_Nomeia_Coordenada		
Mapeamento do relacionamento NxM entre Gene e Coordenada_gene		
Atributo	Domínio	Descrição
id_coordenada	INTEGER	Chave estrangeira
id_gene	INTEGER	Chave estrangeira
Transcrito		
Contém os identificadores de transcritos em diferentes bancos de dados		
Atributo	Domínio	Descrição
id_transcrito	INTEGER	Chave artificial da tabela Transcrito
id_gene	INTEGER	Chave estrangeira
Banco	TEXT	Banco de dados que contém a anotação
Nome_no_banco	TEXT	Identificador do gene no banco
Coordenada_transcrito		

Contém as coordenadas genômicas dos transcritos de origem e versão variadas		
Atributo	Domínio	Descrição
id_coordenada	SERIAL	Chave artificial da tabela Coordenada_transcrito
Cromossomo	TEXT	Cromossomo
Inicio	TEXT	Posição de inicio do gene, medido em nucleotídeos
Fim	TEXT	Posição de término do gene, medido em nucleotídeos
Fita	TEXT	Fita do DNA em que se encontra o gene
banda_cromossomo	TEXT	Banda cromossômica onde se encontra o transcrito
Referencia	TEXT	Nome do banco ou projeto que definiu a coordenada
Montagem	TEXT	Versão utilizada da referência
transcription_start_site	INTEGER	Fonte da anotação. Utilizada caso a referencia especifique uma origem externa para a anotação
comprimento	INTEGER	Extensão do transcrito maduro, em número de bases
Source	TEXT	Fonte da anotação. Utilizada caso a referencia especifique uma origem externa para a anotação
Status	TEXT	Um gene ou transcrito <i>known</i> (conhecido) corresponde a uma sequência de uma base de dados pública, como o NCBI RefSeq. Caso a sequencia seja nova ou tenha sido descrita para uma espécie diferente, o gene é classificado como <i>novel</i> (novo) para essa espécie. Um transcrito pode ser putativo (<i>putative</i>), que significa que não está descrito em outros bancos de dados mas tem um evidência de transcrição
transcript_biotype	TEXT	Classificação do biotipo, especificando se trata-se de rRNA, lincRNA, codificador de proteína, etc.
Transcrito_Nomeia_Coordenada		
Mapeamento do relacionamento NxM entre Transcrito e Coordenada_transcrito		

Atributo	Domínio	Descrição
id_coordenada	INTEGER	Chave estrangeira
id_transcrito	INTEGER	Chave estrangeira
Paciente		
Informações sobre os pacientes		
Atributo	Domínio	Descrição
id_paciente	SERIAL	Chave artificial da tabela Paciente
banco_de_origem	TEXT	Banco de dados do projeto da qual o paciente faz parte
versao_do_banco_de_origem	TEXT	Versão do banco de origem
donor_id_banco_de_origem	TEXT	Identificador do paciente no banco de origem
project_code	TEXT	Usado caso o banco de origem tenha mais de um projeto ligado a ele. No caso do ICGC, cada grupo que pesquisa um subtipo de câncer em cada país recebe um código de projeto
study_donor_involved_in	TEXT	Utilizado caso o paciente esteja envolvido em algum estudo além do projeto especificado em project_code
submitted_donor_id	TEXT	Identificador do paciente no projeto de origem
donor_sex	TEXT	Sexo do paciente
donor_vital_status	TEXT	Estado de vida do paciente. São aceitos os termos <i>alive</i> , <i>deceased</i> e campos em branco
disease_status_last_followup	TEXT	Estado da neoplasia no último acompanhamento. São aceitos os termos <i>complete remission</i> , <i>partial remission</i> , <i>stable</i> , <i>progression</i> , <i>no evidence of disease</i> , <i>deceased</i> , <i>relapse</i> e campos vazios
donor_relapse_type	TEXT	<i>Tipo de recaída. São aceitos os termos local recurrence, distant recurrence/metastasis, local recurrence and distant metastasis</i>

donor_age_at_diagnosis_inicial	TEXT	Idade do paciente no momento do diagnóstico, em anos
donor_age_at_enrollment	INTEGER	Idade do paciente no momento de coleta do primeiro espécime, em anos
donor_age_at_last_followup	INTEGER	Idade do paciente, em anos, no último <i>follow up</i>
donor_relapse_interval	INTEGER	Se o doador foi considerado livre de doença e a seguir teve uma recaída, esse campo representa o intervalo, em dias, em que o paciente esteve livre de doença
donor_diagnosis_icd10_inicial	TEXT	Identificador de diagnóstico do ICD-10
donor_tumour_staging_system_at_diagnosis_inicial	TEXT	Critério de estadiamento clínico do paciente no momento do diagnóstico, se houve
donor_tumour_stage_at_diagnosis_inicial	TEXT	Estadiamento clínico do paciente no momento do diagnóstico
donor_tumour_stage_at_diagnosis_supplemental_inicial	TEXT	Informações adicionais sobre o estadiamento clínico do paciente no momento do diagnóstico
donor_survival_time	INTEGER	Tempo de sobrevivência do paciente, em dias, desde o diagnóstico
donor_interval_of_last_followup	INTEGER	Intervalo entre o diagnóstico e a última visita do paciente ao serviço de saúde, em dias
prior_malignancy	TEXT	Indica se o paciente teve alguma neoplasia antes do diagnóstico
cancer_type_prior_malignancy	TEXT	Tipo de neoplasia que o paciente teve antes do diagnóstico
cancer_history_first_degree_relative	TEXT	Indica se o paciente tem algum parente de primeiro grau que teve uma neoplasia

Especime

Informações sobre os espécimes coletados cirúrgica ou clinicamente

Atributo	Domínio	Descrição
id_especime	SERIAL	Chave artificial da tabela Especime
id_paciente	INTEGER	Chave estrangeira
id_banco_de_origem	TEXT	identificador do espécime no banco de origem
study_specimen_involved_in	TEXT	Utilizado caso o espécime esteja envolvido em algum estudo além do projeto especificado em <code>paciente.project_code</code>
submitted_specimen_id_banco_de_origem	TEXT	Identificador do espécime no projeto de origem

specimen_type	TEXT	<p>Especifica o tecido coletado. Diferencia tecido normal e neoplásico, e local de coleta. São aceitos os termos:</p> <p><i>Normal - solid tissue, Normal - blood derived, Normal - bone marrow, Normal - tissue adjacent to primary, Normal - buccal cell, Normal - EBV immortalized, Normal - lymph node, Normal - other, Primary tumour - solid tissue, Primary tumour - blood derived (peripheral blood), Primary tumour - blood derived (bone marrow), Primary tumour - additional new primary, Primary tumour - other, Recurrent tumour - solid tissue, Recurrent tumour - blood derived (peripheral blood), Recurrent tumour - blood derived (bone marrow), Recurrent tumour - other, Metastatic tumour - NOS, Metastatic tumour - lymph node, Metastatic tumour - metastasis local to lymph node, Metastatic tumour - metastasis to distant location, Metastatic tumour - additional metastatic, Xenograft - derived from primary tumour, Xenograft - derived from tumour cell line, Cell line - derived from tumour, Primary tumour - lymph node, Metastatic tumour - other, Cell line - derived from xenograft tumour</i></p>
specimen_type_other	TEXT	<p>Descrição extra sobre o espécime, caso os termos <i>Normal - other, Primary tumour - other, Recurrent tumour - other</i> ou <i>Metastatic tumour - other</i> tenha sido utilizado no atributo anterior</p>
specimen_interval	TEXT	<p>Intervalo, em dias, entre o diagnóstico e a obtenção do espécime</p>
specimen_donor_treatment_type	TEXT	<p>Tratamento que o paciente estava fazendo no momento da obtenção do espécime. São aceitos os termos:</p> <p><i>No treatment, chemotherapy, radiation therapy, combined chemo+radiation therapy, immunotherapy, combined chemo+immunotherapy, surgery, other therapy, bone marrow transplant, stem cell transplant, monoclonal antibodies (for liquid tumours).</i> Campos vazios são aceitos</p>
specimen_donor_treatment_type_other	TEXT	<p>Caso o termo <i>other therapy</i> tenha sido especificado no atributo anterior, pode ser especificado nesse campo</p>

specimen_processing	TEXT	Processamento do espécime. São aceitos os termos: <i>cryopreservation in liquid nitrogen (dead tissue), cryopreservation in dry ice (dead tissue), cryopreservation of live cells in liquid nitrogen, cryopreservation, other, formalin fixed, unbuffered, formalin fixed, buffered, formalin fixed & paraffin embedded, fresh, other technique</i> . Campos vazios são aceitos
specimen_processing_other	TEXT	Caso o termo <i>other</i> tenha sido especificado no atributo anterior, pode ser especificado nesse campo
specimen_storage	TEXT	Método de armazenamento do espécime. São aceitos os termos: <i>frozen, liquid nitrogen, frozen, -70 freezer, frozen, vapor phase, RNA later frozen, paraffin block, cut slide, other</i> . Campos vazios são aceitos.
specimen_storage_other	TEXT	Caso o termo <i>other</i> tenha sido especificado no atributo anterior, pode ser especificado nesse campo
tumour_confirmed	TEXT	Confirmação do tumor por métodos histológicos
specimen_biobank	TEXT	Nome do biobanco, caso o espécime tenha sido obtido de um biobanco
specimen_biobank_id	TEXT	Identificador do espécime no biobanco, caso provenha de um
specimen_available	TEXT	Especifica se o tecido coletado está disponível para novos estudos
tumour_histological_type	TEXT	Classificação histológica do tumor de acordo com a Organização Mundial de Saúde
tumour_grading_system	TEXT	Nome do sistema de definição do grau do tumor
tumour_grade	TEXT	Grau tumoral de acordo com o sistema especificado anteriormente
tumour_grade_supplemental	TEXT	Informações adicionais sobre o grau do tumor
tumour_stage_system	TEXT	Nome do sistema de estadiamento do tumor
tumour_stage	TEXT	Estadiamento do tumor de acordo com o sistema especificado anteriormente
tumour_stage_supplemental	TEXT	Informações adicionais sobre o estadiamento do tumor
digital_image_of_stained_section	TEXT	Link para uma imagem representativa do tumor
percentage_cellularity_string	TEXT	Porcentagem da proporção de núcleos de células tumorais por células totais. Caso o valor não tenha sido avaliado de forma exata, utilizar o campo <i>level_of_cellularity</i>

level_of_cellularity	TEXT	Porcentagem da proporção de núcleos de células tumorais por células totais em valores aproximados. São aceitos os termos: 1-20%, 21-40%, 41-60%, 61-80%, >81%
responsavel_coleta	TEXT	Pessoa responsável pela coleta do espécime
Amostra		
Informações sobre as amostras derivadas de um espécime		
Atributo	Domínio	Descrição
id_amostra	SERIAL	Chave artificial da tabela Amostra
id_paciente	INTEGER	Chave estrangeira
id_especime	INTEGER	Chave estrangeira
sample_id_banco_de_origem	TEXT	Identificador da amostra no banco de origem
submitted_sample_id_banco_de_origem	TEXT	Identificador da amostra no projeto de origem
specimen_id_banco_de_origem	TEXT	Identificador do espécime de onde foi obtida a amostra no banco de origem
submitted_specimen_id_banco_de_origem	TEXT	Identificador do espécime de onde foi obtida a amostra no projeto de origem
analyzed_sample_interval	TEXT	Intervalo entre a obtenção do espécime e a utilização do mesmo em um experimento
percentage_cellularity	TEXT	Porcentagem da proporção de núcleos de células tumorais por células totais. Caso o valor não tenha sido avaliado de forma exata, utilizar o campo level_of_cellularity
level_of_cellularity	TEXT	Porcentagem da proporção de núcleos de células tumorais por células totais em valores aproximados. São aceitos os termos: 1-20%, 21-40%, 41-60%, 61-80%, >81%
Marcador		
Marcadores clínicos, sorológicos ou experimentais		

Atributo	Domínio	Descrição
id_marcador	SERIAL	Chave artificial da tabela Marcador
nome_marcador	TEXT	Nome do marcador
unidade_de_medida	TEXT	Unidade de medida do marcador
Marcador_Auxilia_Diagnóstico		
Mapeamento do relacionamento NxM entre Marcador e Especime		
Atributo	Domínio	Descrição
id_marcador	INTEGER	Chave estrangeira
id_especime	INTEGER	Chave estrangeira
valor_medido	TEXT	Valor medido para um dado marcador e uma amostra
pessoa_responsavel	TEXT	Pessoa responsável pela avaliação do marcador
data_da_analise	DATE	Dia da avaliação
protocolo	TEXT	Protocolo de avaliação adotado
equipamento	TEXT	Equipamentos utilizados para a análise do marcador
Familiar		
Informações sobre familiares dos pacientes que tiveram neoplasias		
Atributo	Domínio	Descrição
id_paciente	INTEGER	Chave estrangeira
id_paciente_banco_de_origem	TEXT	Identificador do paciente no banco de origem (mantido para facilitar o mapeamento)
relationship_type	TEXT	Grau de parentesco do paciente e o familiar em questão. São aceitos os termos: <i>Sibling, parent, grandparent, uncle/aunt, cousin, other, unknown.</i>

relationship_type_other	TEXT	Caso o termo <i>other</i> tenha sido especificado no atributo anterior, pode ser especificado nesse campo
relationship_sex	TEXT	Sexo do familiar
relationship_age_string	TEXT	Idade familiar, em anos, no momento do diagnóstico da neoplasia
relationship_disease_icd10	TEXT	Código do ICD-10 da neoplasia
relationship_disease	TEXT	Nome da neoplasia do familiar
Tratamento		
Tratamento realizado por paciente		
Atributo	Domínio	Descrição
id_tratamento	SERIAL	Chave artificial da tabela Tratamento
id_paciente	INTEGER	Chave estrangeira
donor_id_banco_de_origem	TEXT	Identificador do paciente no banco de origem (mantido para facilitar o mapeamento)
data_do_primeiro_tratamento	DATE	Data em que foi realizado o primeiro tratamento
first_therapy_type	TEXT	Tipo de tratamento dado após a primeira ressecção de espécime. Tratamentos fornecidos antes da primeira ressecção deve ser especificado em especime.specimen_donor_treatment_type
first_therapy_therapeutic_intent	TEXT	Objetivo do primeiro tratamento dado ao paciente após ressecção de espécime
first_therapy_start_interval	TEXT	Dias entre o diagnóstico e o primeiro tratamento
first_therapy_duration	TEXT	Duração, em dias, do primeiro tratamento
first_therapy_response	TEXT	Efeito da primeira terapia dada após ressecção de espécime
data_do_segundo_tratamento	DATE	Data em que foi realizado o segundo tratamento
second_therapy_type	TEXT	Tipo de tratamento dado após a primeira ressecção de espécime. Tratamentos fornecidos antes da primeira ressecção deve ser especificado em especime.specimen_donor_treatment_type
second_therapy_therapeutic_intent	TEXT	Tipo de tratamento dado após o primeiro tratamento pós ressecção de espécime

second_therapy_start_interval	TEXT	Intervalo, em dias, entre o diagnóstico e a segunda terapia fornecida ao paciente
second_therapy_duration	TEXT	Duração, em dias, do segundo tratamento pós ressecção de espécime
second_therapy_response	TEXT	Efeito da segunda terapia dada após ressecção de espécime
other_therapy	TEXT	Descrição extra caso a primeira ou a segunda terapia necessitem de mais informações
other_therapy_response	TEXT	Caso o atributo anterior tenha sido utilizado, resposta da terapia descrita no referido atributo
Exposicao		
Fatores ambientais a que o paciente pode ter sido exposto		
Atributo	Domínio	Descrição
id_exposicao	SERIAL	Chave artificial da Exposição
id_paciente	INTEGER	Chave estrangeira
donor_id_banco_de_origem	TEXT	Identificador do paciente no banco de origem (mantido para facilitar o mapeamento)
exposure_type	TEXT	Usado caso o paciente tenha sido exposto a algo além de tabaco e álcool
exposure_intensity	TEXT	Usar este campo para especificar a extensão da exposição especificada no atributo anterior
tobacco_smoking_history_indicator	TEXT	Histórico de consumo de tabaco do paciente
tobacco_smoking_intensity	TEXT	Intensidade da exposição ao tabaco
alcohol_history	TEXT	Histórico de consumo de álcool do paciente
alcohol_history_intensity	TEXT	Intensidade da exposição ao álcool
Experimento_de_Mutacao_Somatica		
Organiza os dados dos experimentos de mutação somática		

Atributo	Domínio	Descrição
id_exp_mut_somatica	SERIAL	Chave artificial de Experimento de Mutação Somática
tipo_experimento	TEXT	Especifica o tipo de experimento realizado. São aceitos os termos: <i>Estrutural</i> , <i>variacao_de_copias</i> e <i>simples</i>
projeto_de_origem	TEXT	Projeto que realizou o experimento
comparacao_realizada	TEXT	Tipo de comparação que o estudo fez. Normalmente faz-se comparações entre tecidos tumorais e não-neoplásicos, mas pode haver outros tipos
data_do_experimento	DATE	Data em que foi realizado o experimento
data_da_analise	DATE	Data em que foi feita a análise do experimento
pessoa_responsavel	TEXT	Pessoa responsável pelo experimento
analysis_id_experimento_banco_de_origem	TEXT	Identificador único para a análise realizada para todo o experimento
assembly_version	TEXT	Versão do genoma humano usado como referência para a análise
platform	TEXT	Identificador da plataforma utilizada para o experimento, incluindo sua versão
experimental_protocol	TEXT	Nome do protocolo utilizado e/ou link para o mesmo
sequencing_strategy	TEXT	Tipo de sequenciamento, quanto ao tipo de DNA capturado (exoma ou genoma total) e tipo de <i>reads</i>
base_calling_algorithm	TEXT	Nome do algoritmo utilizado para <i>base calling</i> e/ou link para o mesmo
alignment_algorithm	TEXT	Nome do algoritmo utilizado para alinhamento e/ou link para o mesmo
variation_calling_algorithm	TEXT	Nome do algoritmo utilizado para definir mutações e/ou link para o mesmo
other_analysis_algorithm	TEXT	Nomes de outros algoritmos que possam ter sido utilizados
raw_data_repository	TEXT	Repositório público onde estão armazenados os dados brutos do experimento
raw_data_accession	TEXT	Identificador dos dados brutos no repositório público especificado no atributo anterior
initial_data_release_date	DATE	Data em que os dados foram disponibilizados
estudo_de_origem	TEXT	Caso o dado tenha sido adicionado de um trabalho da literatura, especificar aqui a referência bibliográfica

banco_de_origem	TEXT	Banco de dados de onde o dado foi obtido
versao_do_banco_de_origem	TEXT	Versão do banco de origem
Experim_Mutacao_Utiliza_Amostra		
Entidade para mapear o uso de amostras pelos experimentos de mutação		
Atributo	Domínio	Descrição
id_exp_mut_somatica	INTEGER	Chave estrangeira
id_amostra	INTEGER	Chave estrangeira
tipo_experimento	TEXT	Tipo de experimento de mutação
analysis_id_amostra_banco_de_origem	TEXT	Identificador único para a análise realizada para a amostra em questão
raw_data_repository	TEXT	Repositório público onde estão armazenados os dados brutos do experimento para a amostra em questão
raw_data_accession	TEXT	Identificador dos dados brutos no repositório público especificado no atributo anterior
seq_coverage	DOUBLE	Cobertura da sequência, caso tenha sido utilizada plataforma de sequenciamento
tipo_amostra	TEXT	Especifica se a amostra é normal ou tumoral
Mutacao_Simples_Avalia_Gene_em_Amostra		
Relacionamento entre a amostra controle, amostra experimental, experimento e gene		
Atributo	Domínio	Descrição
id_Amostra_experimental	INTEGER	Chave estrangeira
id_Amostra_controle	INTEGER	Chave estrangeira
id_gene	INTEGER	Chave estrangeira
id_exp_mut_somatica	INTEGER	Chave estrangeira

chromossomo	TEXT	Cromossomo
chromossomo_inicio	INTEGER	Posição de início da mutação
chromossomo_fim	INTEGER	Posição de fim da mutação
fita	TEXT	Fita do gene em que foi detectada a mutação
mutation_type	TEXT	Tipo de mutação. São aceitos os termos: <i>single base substitution, insertion of <=200bp, deletion of <=200bp e multiple base substitution (>=2bp and <=200bp)</i>
reference_genome_allele	TEXT	Alelo de referência
mutated_from_allele	TEXT	Alelo de referência que se acredita ser o original na linhagem celular ancestral. Deve ser um dos alelos de referência
mutated_to_allele	TEXT	Novo alelo identificado no tumor que acredita-se ser resultado de uma mutação somática que ocorreu durante a tumorigênese
quality_score	TEXT	Qualidade atribuída a chamada da mutação
probability	TEXT	Probabilidade da chamada da mutação
total_read_count	INTEGER	Número total de <i>reads</i> brutos com cobertura da mutação
mutant_allele_read_count	TEXT	Número total de <i>reads</i> brutos que contém o alelo mutante
verification_status	TEXT	Indica se a mutação identificada foi validada por algum método diferente do que se utilizou para a referida identificação
verification_platform	TEXT	Plataforma utilizada para a validação da mutação, caso tenha sido indicada no atributo anterior
biological_validation_status	TEXT	Indica se a mutação identificada foi encontrada em estudos diferentes, com pacientes diferentes
biological_validation_platform	TEXT	Plataforma utilizada para a validação biológica da mutação, caso tenha sido indicada no atributo anterior
consequence_type	TEXT	Tipo de consequência da mutação.
aa_mutation	TEXT	Resultado da mutação descrita, quanto à mudança na proteína resultante
cds_mutation	TEXT	Resultado da mutação descrita, quanto à mudança no mRNA resultante

Experimento_de_Expressao_Genica		
Organiza os dados dos experimentos de expressão gênica		
Atributo	Domínio	Descrição
id_exp_genica	SERIAL	Chave artificial de Experimento de Expressão Gênica
projeto_de_origem	TEXT	Projeto que realizou o experimento
experimento_expressao_genica	TEXT	Tipo de experimento de expressão gênica
comparacao_realizada	TEXT	Tipo de comparação que o estudo fez. Normalmente faz-se comparações entre tecidos tumorais e não-neoplásicos, mas pode haver outros tipos
data_do_experimento	DATE	Data em que o experimento foi realizado
data_da_analise	DATE	Data em que foi feita a análise do experimento
peessoa_responsavel	TEXT	Pessoa responsável pela realização do experimento
tipo_de_fold_change	TEXT	Tipo de <i>fold change</i> adotado no experimento quanto a utilização de valores absolutos ou logarítmicos
analysis_id_experimento_banco_de_origem	TEXT	Identificador único para a análise realizada para todo o experimento
assembly_version	TEXT	Versão do genoma humano usado como referência para a análise
total_read_count	INTEGER	Número total de <i>reads</i> no RNASeq, indicador de profundidade do sequenciamento
platform	TEXT	Identificador da plataforma utilizada para o experimento, incluindo sua versão
experimental_protocol	TEXT	Nome do protocolo utilizado e/ou link para o mesmo
alignment_algorithm	TEXT	Nome do algoritmo utilizado para alinhamento e/ou link para o mesmo
normalization_algorithm	TEXT	Nome do algoritmo de normalização da expressão gênica
other_analysis_algorithm	TEXT	Nomes de outros algoritmos que possam ter sido utilizados
sequencing_strategy	TEXT	Tipo de sequenciamento, quanto ao tipo de RNA utilizado e o tipo de <i>reads</i> obtidos
raw_data_repository	TEXT	Repositório público onde estão armazenados os dados brutos do experimento

raw_data_accession	TEXT	Identificador dos dados brutos no repositório público especificado no atributo anterior
reference_sample_type	TEXT	Tipo de amostra usada para o cálculo de <i>fold change</i>
banco_de_origem	TEXT	Banco de dados de onde o dado foi obtido
versao_do_banco_de_origem	TEXT	Versão do banco de origem
Experim_Express_Genica_Utiliza_Amostra		
Entidade para mapear o uso de amostras pelos experimentos de expressão gênica		
Atributo	Domínio	Descrição
id_exp_genica	INTEGER	Chave estrangeira
id_amostra	INTEGER	Chave estrangeira
tipo_experimento	TEXT	Tipo de experimento de expressão gênica
analysis_id_amostra_banco_de_origem	TEXT	Identificador único para a análise realizada para a amostra em questão
raw_data_repository	TEXT	Repositório público onde estão armazenados os dados brutos do experimento para a amostra em questão
raw_data_accession	TEXT	Identificador dos dados brutos no repositório público especificado no atributo anterior
tipo_amostra	TEXT	Especifica se a amostra é normal ou tumoral
Sequenciamento_Avalia_Gene_em_Amostra		
Relacionamento entre a amostra experimental, experimento e gene		
Atributo	Domínio	Descrição
id_gene	INTEGER	Chave estrangeira
id_amostra	INTEGER	Chave estrangeira

id_exp_genica	INTEGER	Chave estrangeira
normalized_read_count	DOUBLE	Número de <i>reads</i> normalizados para o gene em questão
raw_read_count	INTEGER	Número de <i>reads</i> brutos para o gene em questão
fold_change	FLOAT	<i>Fold change</i> tumoral:normal do gene em questão
base_mean	FLOAT	Número médio de reads do gene em questão, normalizado
pvalue	DOUBLE	p-valor do <i>fold change</i>
p_adjusted	DOUBLE	p-valor ajustado do <i>fold change</i> para testes múltiplos, utilizando o algoritmo de Benjamini-Hochberg
lfcSE	FLOAT	Desvio padrão do <i>fold change</i>
stat	FLOAT	Teste de Wald, calculado pela divisão do log do <i>fold change</i> na base 2 pelo seu desvio padrão
is_isomir	TEXT	Usado para dados de miRNA, indicando se
log2FoldChange	DOUBLE	<i>Fold change</i> em logarítimo na base 2

Anexo 3

Consultas realizadas ao banco de dados

Obtenção de dados clínicos e informações sobre *status* da mutação de amostras com mutação em *KRAS*

```
SELECT
    s.id_amostra, s.id_exp_genica,
    s.id_gene, g.nome_no_banco,
    s.normalized_read_count, s.raw_read_count,
    a.id_amostra, a.id_paciente, a4.id_amostra,
    a4.id_paciente, m.id_exp_mut_somatica,
    m.id_Amostra_experimental, m.id_Amostra_controle,
    m.id_gene, m.cromossomo, m.cromossomo_inicio,
    m.cromossomo_fim, m.fita, m.mutation_type,
    m.mutated_from_allele, m.mutated_to_allele,
    m.quality_score, m.probability, m.total_read_count,
    m.mutant_allele_read_count, m.aa_mutation, m.cds_mutation,
    m.consequence_type
FROM
    Sequenciamento_Avalia_Gene_em_Amostra s, amostra a4, amostra a,
    Mutacao_Simples_Avalia_Gene_em_Amostra m, gene g
where
    s.id_amostra = a.id_amostra
    and s.id_amostra in (
        SELECT a3.id_amostra
            from
                amostra a3
            where
                id_paciente in (
                    select
                        a1.id_paciente
                    from
                        Experim_Mutacao_Utiliza_Amostra em, amostra a1
                    where
                        em.id_amostra = a1.id_amostra
                    INTERSECT
                    select
                        a2.id_paciente
                    from
                        Experim_Express_Genica_Utiliza_Amostra ee,
                        amostra a2
```



```

        where
            ee.id_amostra = a2.id_amostra
        )
    )
    and m.id_amostra_experimental = a4.id_amostra -- Selecionando apenas
as amostras de mutação que são de pacientes com dados dos dois experimentos
    and m.id_amostra_experimental in (
        select a4.id_amostra
        from amostra a4
        where
            id_paciente in (
                select a1.id_paciente
                from
                    Experim_Mutacao_Utiliza_Amostra em, amostra a1
                where
                    em.id_amostra = a1.id_amostra
            INTERSECT
                select
                    a2.id_paciente
                from
                    Experim_Express_Genica_Utiliza_Amostra ee,
                    amostra a2
                where
                    ee.id_amostra = a2.id_amostra
            )
    )
    and a.id_paciente = a4.id_paciente
    and m.id_gene = 72009
    and m.aa_mutation in ('G12A', 'G12C', 'G12D', 'G12R', 'G12S', 'G12V',
'Q61H')
    and g.id_gene = s.id_gene

```

Obtenção de dados de expressão de todos os genes e dados clínicos de pacientes do projeto PACA-AU do ICGC

```
SELECT
    g.nome_no_banco,
    s.normalized_read_count,
    s.raw_read_count,
    p.id_paciente, p.project_code,
    p.donor_id_banco_de_origem, p.banco_de_origem,
    p.donor_vital_status, p.donor_survival_time,
    p.donor_interval_of_last_followup

FROM paciente p, amostra a, gene g,
    sequenciamento_avalia_gene_em_amostra s,
    experimento_de_expressao_genica ee

WHERE p.id_paciente = a.id_paciente AND
    a.id_amostra = s.id_amostra AND
    g.id_gene = s.id_gene AND
    ee.id_exp_genica = s.id_exp_genica AND
    g.banco_alias = 'Ensembl'AND
    p.project_code = 'PACA-AU'
;
```