Disseny d'una aplicació web per a la integració, visualització i anàlisi de les dades genòmiques obtingudes del projecte Pan-Cancer Whole Genome Sequencing (PCAWGS)

Nom Estudiant: Anna Pedrola Gómez
Pla d'estudis: Màster en Bioinformàtica i Bioestadística
Àrea de treball final: Estadística i Bioinformàtica 3

Nom Professor/a Consultor/a: Laia Bassaganyas Bars
Nom Professor/a responsable de l'assignatura: Ferran Prados Carrasco

Data de lliurament: 5 de juny del 2019

FITXA DEL TREBALL FINAL

Títol del treball:	*Disseny d'una aplicació web per a la integració, visualització i anàlisi de les dades genòmiques obtingudes del projecte Pan-Cancer Whole Genome Sequencing (PCAWGS) .*
Nom de l'autor:	*Anna Pedrola Gómez*
Nom del consultor/a:	*Laia Bassaganyas Bars*
Nom del PRA:	*Ferran Prados Carrasco*
Data de lliurament (mm/aaaa):	*06/2019*
Titulació o programa:	*Máster en bioinformàtica i bioestadística*
Àrea del Treball Final:	*Estadística i bioinformàtica*
Idioma del treball:	*Català*
Paraules clau	*Whole Genome Sequence, càncer, immunitat, expressió gènica, aneuploïdia*

Resum del Treball (màxim 250 paraules): *Amb la finalitat, context d'aplicació, metodologia, resultats i conclusions del treball*

Aquest treball de final de màster, situat en l'àmbit de la genòmica del càncer esta basat en les dades del projecte internacional, Pan-Cancer Analyses of Whole Genome Sequences (PCAWGS). Dades fetes publiques recentment que inclouen informació sobre varis anàlisis duts a terme on hi podem trobar més de 2.800 mostres de seqüenciació completa del genoma (Whole Genome Sequencing o WGS), per a diferents tipus de càncer, fins a 40. Entre aquests anàlisis en trobem per exemple: alteracions en nombre de copia (CNA), variants estructurals (SV), dades d'expressió i dades clíniques.

En aquest treball s'han utilitzat aquestes dades per a partir d'elles, calcular variables de gran interès en la genòmica del càncer actual, com són l'immunophenoscore (IPS) i les puntuacions de CNA (CNA scores), amb la finalitat d'integrar aquestes noves dades al conjunt de PCAWGS i poder estudiar el paper de les CNAs i SVs en la determinació de les característiques immunitàries del tumor. Com a resultat, s'ha creat una base de dades que relaciona tota aquesta informació i a partir de la qual s'ha desenvolupat una aplicació web amb el framework de Python Django, per tal de proporcionar a grups de recerca clínics una eina de visualització senzilla i intuïtiva, d'on extreure informació sobre les relacions d'aquestes variables en molts tipus de càncer diferents, per poder interpretar i complementar els seus estudis.

Abstract (in English, 250 words or less):

This master's final project, located in the field of cancer genomics is based on the data of the recently published international project Pan-Cancer Analyzes of Whole Genome Sequences (PCAWGS). This project contains more than 2800 samples of the complete sequencing of the genome (WGS), for up to 40 different types of cancer. This data includes: copy number alterations (CNA), structural variants (SV), expression data and clinical data.

In this project we have used PCAWGS data for the calculation of some variables of great interest in the genetics of the current cancer, which are immunophenoscore (IPS) and scores of CNA (BSC and FCS), with the purpose of integrating these new data to the set of PCAWGS and study the impact of CNAs and SVs in the immune characteristics of tumours. As a result, a database was created to contain such calculations and from which a web application has been developed with the Python Django framework, in order to provide clinical research groups a tool of simple and intuitive visualization, to extract information about the relationships of these variables in many different cancer types to interpret and complement their studies.

Índex

Llista de figures

Llista de taules

Llista d'acrònims

PCAWGS	Pan-Cancer Analyses of Whole Genome Sequences
SVs	Structural Variants
CNV-s	Copy Number Variants
CNAs	Copy Number Alterations
SNVs	Single nucleotide variant
Kb	Kilo base
WGS	Whole genome sequence
WES	Whole exome sequencing
NGS	Next-generation sequencing
ADN	Àcid desoxiribonucleic
ARN	Àcid ribonucleic
ICGC	International Cancer Genome Consortium
TCGA	The Cancer Genome Atlas
SQL	Structured Query Language
UCSC	University of California, Santa Clara
TMB	Tumor mutational burden
BCS	Broad CNA score
FCS	Focal CNA score
GCS	Global CNA score

1. Introducció

1.1 Context i justificació del Treball

Actualment, s'emmagatzemen quantitats molt grans de dades. En l'àmbit científic, això esdevé tant pel fet de que les tècniques d'investigació són capaces de recollir i/o generar moltes dades, com perquè moltes investigacions es basen en la cerca de repeticions o patrons, que requereixen conjunts d'informació molt grans.

Un context on s'exemplifica clarament la generació i ús de dades massives en el camp científic és en el desenvolupament de la medicina personalitzada i, molt específicament, en l'estudi del càncer, on el tractament de la malaltia passa cada vegada més pel correcte i estricte anàlisi de les característiques especifiques del tumor tant a nivell genòmic com transcriptòmic (expressió gènica) i epigenòmic. Això implica l'obtenció de grans quantitats de dades biològiques (bàsicament d'ADN i ARN) per cada pacient, majoritàriament a partir de tecnologies de seqüenciació de nova generació (*next-generation sequencing* o NGS, en anglès) i, en conseqüència, la capacitat de saber-les processar, analitzar i interpretar. A més, en els darrers anys s'han anat implementant iniciatives a gran escala per poder estudiar el màxim nombre de tumors possibles, amb l'objectiu d'explorar la complexitat del càncer i poder desenvolupar estratègies terapèutiques pels diferents tipus de tumor. Si bé aquest fet ha fet possible un salt qualitatiu en la nostra comprensió i tractament de la malaltia, encara estem lluny de poder-la combatre amb total efectivitat, i és per això que el re-anàlisi i la re-interpretació de les dades massives ja generades proporciona una bona oportunitat per avançar.

Així doncs, aquest treball està basat en la integració d'un gran conjunt de dades genòmiques, clíniques i d'expressió gènica de diferents tipus de tumors, generades a partir de diferents tipus d'anàlisis, que es troben dipositades públicament per a la comunitat científica.

Concretament, es tracta de dades obtingudes pel projecte internacional, Pan-Cancer Analyses of Whole Genome Sequences (PCAWGS) [1], una col·laboració internacional establerta per tal d'identificar patrons comuns d'alteracions genòmiques i transcriptòmiques (d'expressió gènica) a partir de l'estudi massiu de diferents tipus de càncer. En aquest projecte convergeixen dades de seqüenciació completa del genoma (Whole Genome Sequencing o WGS) i d'expressió gènica (RNA sequencing o RNAseq), amb la intenció d'explorar la naturalesa i les conseqüències de les variacions somàtiques i germinals, tant a regions codificants com no codificants, i amb especial èmfasis als llocs cis-regulador, els ARN no codificants i les modificacions estructurals. Actualment hi ha dades de més de 2.800 mostres corresponents a més de 40 tumors diferents, procedents del International Cancer Genome Consortium (ICGC) i del projecte The Cancer Genome Atlas (TCGA).

Entre elles hi podem trobar, per exemple, variants en nombre de copia (copy number alterations o CNA, en angles), variants estructurals (structural variants o SVs, en anglès), firmes mutacionals (mutational signatures), variants de nucleòtids únics (single nucleotide

variants o SNVs, en anglès) i dades d'expressió. A més a més, també hi ha recollides dades clíniques dels pacients, com l'edat, l'estadi del tumor o variables de supervivència.

Aquest conjunt de dades, malgrat estar localitzades en un navegador específic, el Xena Browser de la UCSC [2], fet amb l'objectiu de permetre'n la consulta i exploració, no estan presentades en una forma fàcil d'analitzar. Per altra banda, l'opció de baixar-les per estudiar-ne el contingut, i degut a que estan en forma de diferents taules molt grans, requereix coneixements computacionals que moltes vegades no estan a l'abast, almenys en aquests moments, per la majoria de grups clínics. Això dificulta el seu aprofitament per part d'equips de recerca que se'n podrien beneficiar. El més important, però, és que es tracta d'un conjunt de dades actualment molt valuós ja que són resultats molt complets extrets d'un nombre important d'anàlisis de WGS, representant així un salt qualitatiu important al que s'havia acumulat fins ara en els estudis de genòmica del càncer. La gran majoria de dades públiques emmagatzemades fins ara eren extretes de whole – exome sequencing (WES), pel que tenir aquestes noves dades on hi trobem milers de mostres analitzades amb WGS, permet explorar més enllà, incloent les zones no codificants del genoma o tenir més oportunitats d'analitzar amb més detall les variacions estructurals, les quals majoritàriament presenten els punts de trencament en zones no codificants.

Així doncs, amb aquest treball es pretén optimitzar i simplificar la manipulació i visualització de les dades del PCAWGS, crear relacions entre els resultats generats fins ara, per tal de a continuació, poder respondre un conjunt de preguntes biològiques interessants per a l'estudi del càncer d'una manera fàcil i intuïtiva mitjançant la creació d'una aplicació.

Això es durà a terme a partir de la inicial creació d'una base de dades en SQL Server que reculli tota aquesta informació publicada i on es pugui consultar fàcilment, per a després poder procedir amb al desenvolupament d'una aplicació, la qual, recollint les dades de la base de dades, permeti visualitzar d'una forma fàcil i optima els resultats als anàlisis plantejats.
En resum, la finalitat no és altra que crear una eina que reculli unes dades molt útils i que pugui ser utilitzada per investigadors i investigadores en l'àmbit de la genòmica del càncer per tal de recolzar-se en els seus anàlisis i estudis.

La temàtica d'aquest Treball Final de Màster s'ha escollit donat que en ell convergeixen dos dels grans temes de l'actualitat que em provoquen gran interès. Per una part, el càncer (més concretament la genòmica del càncer), i per l'altra, l'anàlisi de grans conjunts de dades. Ambos, presents en el meu dia a dia, i amb moltes possibilitats per dur a terme un projecte d'aquesta envergadura.
També un motiu pel qual s'ha escollit aquest tema, és pel fet de que les dades genòmiques de tots els tipus de càncers en el que es basa, són una mina de dades molt interessant i amb moltes oportunitats per a realitzar diferents anàlisis, així com que crec que és molt necessària la creació d'eines que apropin aquestes dades als científics i les científiques i que les presentin d'una manera intuïtiva i senzilla d'interpretar per a que, tot hi no tenir coneixements profunds d'anàlisis de dades d'aquest tipus, puguin extreure'n informació útil.

1.2 Qüestions actualment obertes en la genòmica del càncer

El càncer sorgeix per l'acumulació de mutacions al llarg del temps en una sèrie de gens. És per això que amb els avenços en el coneixement del genoma humà, s'ha potenciat molt la investigació de la anomenada genòmica del càncer [3].
La genòmica del càncer és l'estudi de la totalitat de la seqüenciació de l'ADN i les diferencies d'expressió gènica entre les cèl·lules tumorals i les cèl·lules normals. El seu principal objectiu és comprendre la base genètica de la proliferació de les cèl·lules tumorals i l'evolució del genoma del càncer segons les mutacions, per entendre com tot això afecta en la progressió de la malaltia i en la resposta al tractament [4].

Malgrat el progrés dels últims anys en aquest camp, tant a nivell de coneixement bàsic de la biologia dels tumors com a nivell de desenvolupament de noves estratègies terapèutiques, encara queden qüestions importants a resoldre. Un dels aspectes importants és el que fa referència a la resposta intrínseca del sistema immunitari dels pacients contra el tumor i els mecanismes que té aquest per adaptar-se i escapar-se de l'atac de les cèl·lules immunitàries. Aquesta relació entre els dos sistemes és el que s'ha utilitzat com a fonament per al desenvolupament de les immunoteràpies (ITs), que no són més que l'intent d'estimular/regular el sistema immunitari del pacient perquè ataqui i destrueixi les cèl·lules tumorals. Les ITs han significat un dels avanços més important en càncer a nivell terapèutic, però malauradament només hi responen bé un subgrup de pacients. Així, un dels actuals punts més importants en la recerca oncològica és aconseguir entendre els mecanismes que regulen la immunitat a nivell tumoral, així com la cerca de biomarcadors de resposta que permetin identificar ràpidament els pacients amb possibilitats de respondre a les ITs [5] [6].

Tenint en compte tot això, s'integraran a l'aplicació una sèrie d'anàlisis que permetran visualitzar les característiques immunitàries de cada tipus de tumor analitzat, i la correlació entre aquestes i algunes dades genòmiques particulars, com el nivell d'aneuploïdia i el nombre de SVs. Estudis recents suggereixen que la composició immunitària dels tumors ve condicionada per les alteracions en nombre de còpia (aneuploïdia) que carreguen les cèl·lules tumorals [7]. Més específicament, alts nivells de CNAs correlacionarien amb una disminució de l'expressió de marcadors de cèl·lules immunitàries citotòxiques, fet que seria indicatiu de l'evasió del tumor a la resposta immunitària contra ell. Així doncs, s'intentarà obtenir resultats que ens permetin avaluar aquesta correlació en les mostres recollides del PCAWGS.

1.3 Objectius del Treball

- **Objectius generals:**

1. Simplificar l'anàlisi d'un gran conjunt de dades genòmiques, PCAWGS.
2. Utilitzar la base de dades per a respondre diverses preguntes biològiques prèviament establertes.

- **Objectius específics:**

1. Simplificar l'anàlisi d'un gran conjunt de dades genòmiques, PCAWGS.
   1.1 Crear una base de dades relacional que reculli les dades genòmiques de PCAWGS.
   1.2 Proporcionar una visualització senzilla i intuïtiva de les dades.
   1.3 Programar una aplicació que reculli les diferents visualitzacions i anàlisis.

2. Utilitzar la base de dades per a respondre diverses preguntes biològiques prèviament establertes.
   2.1 Estudiar el detall de totes les dades i estudis d'on s'extreuen.
   2.2 Definir una sèrie de preguntes biològiques que es puguin respondre amb els dades estudiades, essent aquestes: l'estudi del paper de les CNAs i SVs en la determinació de les característiques immunitàries dels tumors.
   2.3 Realitzar els anàlisis pertinents per a respondre les preguntes establertes.

1.4 Enfocament i mètode seguit

L'enfocament general del projecte és basa en la recollida inicial de dades, la creació de la base relacional amb el programari SQL Server, i el previ estudi d'aquestes, per tal de poder entendre bé amb quines variables es tracta i que signifiquen aquestes en l'àmbit de la genòmica del càncer. El pas següent va ser la formulació de varies preguntes biològiques que es podien respondre amb aquestes dades, per tal de poder programar una aplicació que dugui a terme els anàlisis necessaris per a poder presentar-ne els resultats d'aquestes, posant especial èmfasis en la visualització senzilla dels resultats e utilització intuïtiva de l'aplicació.
En quant a estratègies més apropiades, podem avaluar-les en varis àmbits. Pel que fa al programari utilitzat s'escull SQL Server per a la gestió de la base de dades ja que és un gestor de base de dades intuïtiu i fàcil de fer servir, i amb molts bons resultats. El plantejament de les preguntes extres sobre les quals es basarà l'anàlisi de les dades recollides, és va dur a terme un cop analitzades correctament les dades, i amb ple coneixement de les variables de que es disposen, ja que plantejar aquestes preguntes des de un inici, sense saber exactament massa bé les dades de les quals es disposa, no era un mètode gaire òptim i amb moltes possibilitats d'acabar canviat a mesura que es va realitzant el treball. Per últim, sobre la programació que s'utilitzarà per l'anàlisi, s'ha dut a terme mitjançant el llenguatge de programació Python, amb l'ajuda puntual del programari R. L'elecció de Python és simplement pel fet de que té més flexibilitat pel que fa a la manipulació de grans conjunts de

dades, però alhora és pot combinar amb R en el cas de ser necessari. Així com que permet opcions més amplies a l'hora de crear l'aplicació final.

1.5 Planificació del Treball

- **Tasques:**

Tasca	Descripció	Durada prevista (setmanes)
1	PAC0 - Definició dels continguts del treball	1
2	PAC1 - Pla de treball.	2
3	Planificació TFM - Diagrama de Gantt.	1
4	Definició dels objectius.	1
5	Redacció del pla de treball.	2
6	PAC2 - Desenvolupament del treball - Fase 1	5
7	Obtenció de dades PCAWGS.	5
8	Creació BBDD SQL amb les dades PCAWGS.	4
9	Estudi de les dades recopilades i de les tècniques d'obtenció.	3
10	Definició de les preguntes biològiques a resoldre.	2
11	Recopilar documentació teòrica per al background del TFM.	3
12	PAC3 - Desenvolupament del treball - Fase 2	4
13	Anàlisi de les dades segons les preguntes definides.	4
14	Visualització general de les dades.	2
15	Programació de l'aplicació.	6
16	Redacció de la memòria.	4
17	PAC4 - Tancament de la memòria.	2
18	Acabar la redacció general de la memòria.	2
19	Redacció de resultats i conclusions.	1
20	PAC5a - Elaboració de la presentació.	1
21	Realització de la presentació.	1
22	Pràctica oral.	1
23	PAC5b - Defensa pública	2

Taula 1 - *Detall de les tasques del treball.*

- **Calendari:** Annexat al final del document (Annex 9.1 - Calendari del projecte)

- **Fites:**

Les dates que s'han utilitzat com a fites durant la realització del treball són les marcades per les PAC's que s'han ant entregant al llarg del semestre. Aquestes dates clau han servit per a revisar com anava la feina, que s'havia fet, que faltava per fer, i si s'estava seguint correctament el calendari de l'apartat anterior. Tot i això, internament les tasques petites han anat tenint un seguiment setmanal o bisetmanal per tal d'avançar progressivament el en projecte i no patir una acumulació de feina els dies abans de les dates clau.

Fita	Descripció	Data clau
1	PAC0 - Definició dels continguts del treball	04/03/19
2	PAC1 - Pla de treball	18/03/19
3	PAC2 - Desenvolupament del projecte - Fase 1	24/04/19
4	PAC3 - Desenvolupament del projecte - Fase 2	20/05/19
5	PAC4 - Tancament de la memòria	05/06/19
6	PAC5a - Elaboració de la presentació	13/06/19
7	PAC5b - Defensa pública	26/06/19

Taula 2 - *Detall de les fites del treball.*

1.6 Breu sumari de productes obtinguts

El producte final d'aquest projecte és una aplicació web que recull i visualitza de forma senzilla les dades del projecte PCAWGS i les variables extres calculades a partir d'aquestes dades. Pretén ser una eina útil per a la comunitat científica on poder consultar les relacions entre varies variables relacionades amb el càncer, per a un gran nombre de càncers diferents.

1.7 Breu descripció dels altres capítols de la memòria

Pel que fa a la resta de la memòria, s'hi troba un apartat de materials i mètodes, on s'expliquen detalladament les dades utilitzades de PCAWGS, així com les calculades a partir d'aquestes i les eines amb els que s'han calculat. En ell també s'hi explica l'eina SQL Server, i l'estructura amb la que s'hi ha creat la base de dades del projecte. I finalment, el programari utilitzat, on s'entra en detall en les eines i mètodes seguint per a la programació de l'aplicació, així com el detall tècnic d'aquesta.

Després hi ha un capítol de resultats i discussió, on es descriu detalladament el resultat de l'aplicació web, mostrant totes les pantalles i opcions d'interacció possibles entre l'usuari i aquesta. I també es defineixen els mètodes utilitzats per a interpretar els resultats a nivell biològic, com les diferents correlacions i tests estadístics realitzats per a les variables estudiades.

A continuació, es descriuen en un breu apartat les problemàtiques i els inconvenients que han anat sorgint a mesura que s'ha anat avançant en la realització del treball i com s'han anat solucionant.

Després trobem un apartat de conclusions, on s'hi resumeixen els resultats i es valora el projecte, tant a nivell de resultat final com de procediments seguit al llarg de la seva realització.

Seguit d'un apartat de treball futur, on es comenten més idees mitjançant les quals es pot seguir explotant les dades que s'han integrat i acabar desenvolupant una aplicació més completa, innovadora i útil per a la comunitat científica.

Finalment hi trobem el glossari amb les definicions necessàries i la bibliografia, que es pot trobar citada al llarg de tot el document. I els annexos, que recullen els fitxers de codi utilitzats per al desenvolupament del projecte.

2. Materials i mètodes

2.1. Dades

2.1.1. Dades PCAWGS

Les dades públiques generades pel PCAWGS i utilitzades en aquest treball estan dipositades al XenaBrowser [2] de la University of California, Santa Clara (UCSC). Bàsicament consisteixen en diferents tipus de resultats generats gràcies a l'anàlisi exhaustiva dels genomes complets (WGS) de les més de 2800 mostres tumorals, així com l'anàlisi de les seves dades de variacions estructurals, dades d'expressió gènica, obtingudes a partir de la seqüenciació de l'ARN (RNAseq), més la informació clínica de cada pacient.

Per a la primera versió de l'aplicació web pensada per aquest treball, s'ha seleccionat el conjunt de dades que permetés explorar, visualitzar i abordar l'aspecte biològic que finalment s'ha definit com a prioritari: l'estudi de la correlació entre el nivell d'aneuploïdia dels tumors (definit com a nivell de CNAs) i la seva composició immunitària. L'interès per l'abordatge d'aquesta associació rau en dades recents obtingudes a nivell de pan-cancer que suggereixen que les variacions estructurals i/o l'aneuploïdia, sobretot la definida per CNAs grans, jugaria un paper essencial en el mecanisme d'evasió immunitària que desenvolupen alguns tumors [7], fet que es creu que està també relacionat amb el grau d'eficàcia en la resposta a la immunoteràpia. Malgrat que s'apunta cap a aquesta direcció, encara no s'arriba a comprendre el motiu pel qual l'aneuploïdia influencia al sistema immunitari, i tampoc se sap del cert que l'associació entre els dos factors és igual en tots els tipus de tumors. Addicionalment, també s'ha decidit incorporar alguns gràfics d'estadística bàsica. Queda pendent per un futur a curt termini la incorporació dels resultats sobre la càrrega mutacional dels tumors (*tumor mutational burden o TMB*).

Variacions en nombre de còpia - Copy Number Alterations (CNAs)

Les alteracions en nombre de còpia (a partir d'ara, CNAs) són alteracions de guanys i pèrdues de segments genòmics més grans d'1-kb.

S'ha demostrat que aproximadament entre el 4.8% i el 9.5% del genoma humà es pot classificar com variacions d'aquest tipus [8] i, de fet, s'ha vist que presenten un paper molt important en la generació de la variació necessària de la població i en el fenotip d'algunes malalties. En els tumors però, les CNAs són molt més abundants i es creu que poden tenir un paper clau tant en el desenvolupament del càncer com en la progressió de la malaltia i la resistència al tractament.

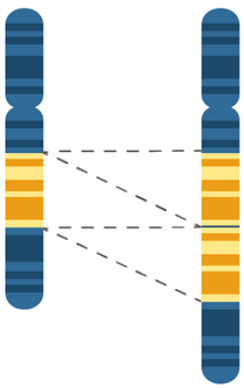


***Figura 1** - Exemple duplicació*

Originàriament, les CNAs s'obtenien mitjançant tècniques citogenètiques, les quals permetien observar l'estructura física del cromosoma. N'és un exemple, la hibridació in situ fluorescent, comunament anomenada FISH, mitjançant les quals s'insereixen sondes fluorescents que requereixen un alt grau de complementarietat en el genoma per unir-se [9].

Posteriorment, amb els avenços tecnològics adquirits, un dels mètodes més utilitzats van passar a ser els microarrays d'alta resolució [10]. De fet, les CNAs més recurrents en càncer s'han descrit gràcies a l'anàlisi massiva de dades de microarrays, generades dins el consorci del TCGA, però la seva capacitat resolutiva és encara limitada per poder localitzar els punts reals de trencament. Més recentment, amb l'arribada de les NGS, també s'han utilitzat dades de whole-exome sequecing (WES) per inferir CNAs en gens, tot i que es tracta d'una tècnica sub-òptima ja que no avarca tot el genoma. El WGS representa la forma més completa d'anàlisis de CNAs, ja que en ella s'explora tot el genoma i amb una capacitat de resolució dels punts de trencament molt més gran que els microarrays.

Les dades de CNAs del PCAWGS són el resultat d'un procediment que combina sis algoritmes de detecció d'aberracions en nombre de còpies: ABSOLUTE [11], ACEseq [12], Battenberg [13], CloneHD [14], JaBbA [33] i Sclust [15]. Tenint en compte les dades dels segments de ploidia, ajustades amb un algoritme per tal de maximitzar l'acord, es consideren individualment per tal d'assignar tres estats diferents de número de còpia: acord clonal (3 estrelles, acord de vot majoritari i acord després d'arrodonir el número de còpia subclonal (2 estrelles) i la relació amb el millor mètode d'aquesta mostra (1 estrella).

sampleID	chr	start	end	value	total_cn	major_cn	minor_cn	star
DO51110	1	1	8764299	2	2	1	1	3
DO51110	1	8764300	8764330	2	2	1	1	3
DO51110	1	8764331	47205247	2	2	1	1	2
DO51110	1	47205248	48396134	1	1	1	0	2
DO51110	1	48396135	49966602	2	2	1	1	2
DO51110	1	49966603	69887288	2	2	1	1	3
DO51110	1	69887289	71938491	2	2	1	1	3
DO51110	1	71938492	72059993	0	0	0	0	1
DO51110	1	72059994	72509357	2	2	1	1	2

Taula 3 – *Mostra de les dades de CNAs de PCAWGS.*

Variacions estructurals - Structural Variants (SVs)

Les variacions estructurals (a partir d'ara SVs) poden ser de dos tipus: balancejades o no balancejades. Pel que fa a les balancejades, no presenten canvi de dosi de material genòmic i el seu impacte es dona només en el punt de trencament del material genètic. Aquestes poden ser inversions o translocacions. En canvi, en les no balancejades, aquelles en els quals si que hi ha canvi de dosi de material genòmic, l'impacte inclou tot el segment alterat i poden ser guanys o pèrdues, anomenades també, CNAs [16] [17].

Una translocació és un re-ordenament cromosòmic, a nivell intra o inter-cromosòmic, on una secció del cromosoma canvia de posició sense afectar al contingut total de l'ADN.

Una inversió és una secció d'ADN en un cromosoma la qual inverteix la seva orientació respecte el genoma de referencia.

Pel que fa a les dades de SV concretament, el projecte PCAWGS representa una gran oportunitat ja que les SVs només es poden analitzar de forma amplia i detallada a partir de WGS, ja que aquest, permet anar més enllà dels CNAs, i estudiar a fons les inversions i translocacions que no es possible obtenir per arrays.

A continuació podem observar una petita mostra de les dades processades de SV que s'han extret de PCAWGS.

sample	chr	start	end	reference	alt	gene	altGene	effect	sample
DO46416	chr1	17378698	17378698	A	A]6:21872484]	SDHB	intergenic region	TRA	DO46416
DO46416	chr1	27256756	27256756	G	]1:28028201]G	NUDC	intergenic region	DUP	DO46416
DO46416	chr1	32976096	32976096	A	]1:33461699]A	RP1-2705.3	RP1-11703.2	DUP	DO46416
DO46416	chr1	33354395	33354395	G	]1:33554784]G	HPCA	ADC	DUP	DO46416
DO46416	chr1	33461699	33461699	T	T[1:32976096[	RP1-11703.2	RP1-2705.3	DUP	DO46416
DO46416	chr1	33554784	33554784	C	C[1:33354395[	ADC	HPCA	DUP	DO46416

***Taula 4** – Mostra de les dades de SV de PCAWGS.*

Sobre aquestes, s'ha hagut de fer una reestructuració de les variables per tal d'obtenir una taula amb la informació ordenada de la manera desitjada. Aquest procediment s'ha dut a terme mitjançant un codi de Python que s'adjunta a l'annex (Annex 9.2 - Codi Python per adaptar les dades SV de PCAWGS) obtenint el resultat que es mostra en la taula 5.

sample	chr1	start	chr2	end	size	reference	gene	alt_gene	effect
DO46416	chr1	17378698	6	21872484	4493786	A	SDHB	intergenic region	TRA
DO46416	chr1	27256756	1	28028201	771445	G	NUDC	intergenic region	DUP
DO46416	chr1	32976096	1	33461699	485603	A	RP1-2705.3	RP1-11703.2	DUP
DO46416	chr1	33354395	1	33554784	200389	G	HPCA	ADC	DUP
DO46416	chr1	32976096	1	33461699	485603	T	RP1-11703.2	RP1-2705.3	DUP
DO46416	chr1	33354395	1	33554784	200389	C	ADC	HPCA	DUP

***Taula 5** - Mostra de les dades SV després de ser processades.*

Informació fenotípica

La informació fenotípica dels pacients inclosos en el projecte PCAWG la trobem dividida en varies taules diferents:

- **Dades fenotípiques / informació clínica - Phenotype / Donor Clinical**

icgc_donor_id	donor_sex	donor_vital_status	donor_diagnosis_icd10	first_therapy_type	first_therapy_response	donor_age_at_diagnosis	donor_survival_time	donor_interval_of_last_followup	tobacco_smoking_history_indicator	tobacco_smoking_intensity	alcohol_history	alcohol_history_intensity	icgc_donor_id
DO1000	female	alive	NULL	other therapy	NULL	61	NULL	NULL	Smoking history not documented	NULL	Don't know/Not sure	Not Documented	DO1000
DO1001	female	NULL	NULL	other therapy	NULL	41	NULL	NULL	Smoking history not documented	NULL	Don't know/Not sure	Not Documented	DO1001
DO1002	female	alive	NULL	other therapy	unknown	39	NULL	NULL	Smoking history not documented	NULL	Don't know/Not sure	Not Documented	DO1002
DO1003	female	alive	C50.4	chemotherapy	unknown	34	NULL	NULL	Smoking history not documented	NULL	Don't know/Not sure	Not Documented	DO1003

***Taula 6** - Mostra de les dades Donor Clinical de PCAWGS.*

Aquí hi trobem les dades clíniques dels individus. D'aquesta taula bàsicament s'han utilitzat dades molt bàsiques com poden ser el consum de tabac o d'alcohol, el gènere de l'individu o l'edat, amb la finalitat de correlacionar-les amb els diferents tipus de càncers que tenim.

- **Dades fenotípiques / Codi de projecte - Phenotype / Project Code**

En aquesta taula trobem la informació que ens relaciona el codi del pacient amb el codi del projecte. Si bé, és la taula per la qual es regeix tot el treball ja que és on trobem tots els pacients inclosos en el projecte, així com la informació del tipus de càncer que tenen.

icgc_donor_id	dcc_project_code
DO1000	BRCA-UK
DO1001	BRCA-UK
DO1002	BRCA-UK
DO1003	BRCA-UK
DO1004	BRCA-UK
DO1005	BRCA-UK

***Taula 7** - Mostra de les dades Project Code de PCAWGS.*

Podem observar 40 tipus de càncers diferents, els quals també estan detallats segons el país d'on s'ha reclutat el pacient.

- **Dades fenotípiques / Histologia del tumor - Phenotype / Tumour Histology**

Finalment, aquí trobem tota la informació que descriu la histologia del tumor, des de la localització (òrgan), fins dades de l'estadi i la cel·lularitat.

Aquesta informació, i la relació amb la taula 7 ens permet classificar els tipus de càncer segons l'òrgan on es troben. D'aquesta manera obtenim la classificació que s'ha utilitzat per a mostrar la informació a l'aplicació web final, i que podem veure a la taula 9.

icgc_specimen_id	organ_system	histology_abbreviation	histology_tier1	histology_tier2	histology_tier3	histology_tier4	tumour_histological_code	tumour_histological_type	tumour_stage	tumour_grade	percentage_cellularity	level_of_cellularity
DO496	URINARY BLADDER	Bladder-TCC	ENDODERM	Bladder	Transitional cell carcinoma	Transitional cell carcinoma, papillary	'8130/3	Papillary trans. cell carcinoma	NULL	NULL	NULL	NULL
DO4766	BREAST	Breast-AdenoCA	ECTODERM	Breast	Adenocarcinoma	Infiltrating duct carcinoma	'8500/3	Infiltrating duct carcinoma, NOS	NULL	NULL	90	90-90
DO498	URINARY BLADDER	Bladder-TCC	ENDODERM	Bladder	Transitional cell carcinoma	Transitional cell carcinoma	'8120/3	Transitional cell carcinoma, NOS	NULL	NULL	80	80

***Taula 8** - Mostra de dades Tumour Histology de PCAWGS.*

Òrgan	Tipus de càncer
Sang, os, sistema hematopoètic	Leucèmia limfocítica crònica (CLLE) Trastorns mieloides crònics (CMDI) Leucèmia Mieloide (LAML)
Os, teixit suau	Càncer d'os (BOCA) Sarcoma (SARC)
Cervell	Glioblastoma multiforme (GBM) Glioma cerebral de grau baix (LGG) Càncer cerebral pediàtric (PBCA)
Pit	Càncer de mama (BRCA)
Cèrvix uterí	Carcinoma cervical de cèl·lules escamoses (CESC)
Esòfag	Adenocarcinoma d'esòfag (ESAD)
Vesícula biliar i conductes extrahepàtics	Càncer del tracte biliar (BTCA)
Boca	Carcinoma de cèl·lules escamoses de cap i coll (HNSC) Càncer oral (ORCA)
Ronyó	Cromòfob Renal (KICH) Carcinoma renal de cèl·lules clares (KIRC) Carcinoma renal de cèl·lules papil·lars renals (KIRP) Càncer de cèl·lules renals (RECA)
Intestí gruixut	Adenocarcinoma de colon (COAD)
Fetge	Càncer de fetge (LICA) Carcinoma hepatocel·lular hepàtic (LIHC) Càncer de fetge (LINC) Càncer de fetge (LIRI)
Pulmó i bronquis	Adenocarcinoma de pulmó (LUAD) Carcinoma de cèl·lules escamoses de pulmó (LUSC)
Nodes limfàtics	Limfoma limfàtic difús de cèl·lules B grans neoplasma limfàtic (DLBC) Limfoma maligno (MALY)
Ovari	Càncer d'ovari (OV)
Pàncreas	Càncer de pàncreas (PACA) Càncer de pàncreas neoplàsies endocrines (PAEN)
Glàndula prostàtica	Càncer de pròstata (PRAD) Càncer de pròstata de aparició temprana (EOPC)
Recte	Adenocarcinoma de recte (READ)
Pell	Càncer de pell (MELA) Melanoma cutani (SKCM)
Estomac	Càncer gàstric (GACA) Adenocarcinoma gàstric (STAD)
Glàndula tiroides	Carcinoma de tiroides (THCA)
Bufeta urinària	Càncer urotelial de bufeta (BLCA)
Úter	Carcinoma d'endometri del cos de l'úter (UCEC)

***Taula 9** - Tipus de càncers del projecte PCAWGS segons òrgan.*

Puresa tumoral i ploidia / Purity & Ploidy

La puresa dels tumors (purity) que es troba a les dades PCAWGS, s'ha aconseguit mitjançant 10 mètodes diferents amb la finalitat de trobar un consens, utilitzant tant les CNAs com les SVs. És per aquesta raó que cada valor te assignat un valor de confiança, per tal d'identificar que un valor d'aquesta confiança baix, representa un alt acord entre els diferents mètodes.
Pel que fa a la plodia (ploidy), és el nombre de jocs complets d'un cromosoma en una cèl·lula. També te una anotació addicional que denota l'estat de duplicació del genoma complet del tumor.

samplename	purity	ploidy	purity_conf_mad	wgd_status	wgd_uncertain
DO46416	0.885	3.355	0.039	wgd	FALSE
DO36062	0.774	2.001	0.022	no_wgd	FALSE
DO45049	0.8	2.428	0.011	no_wgd	FALSE
DO22145	0.837	1.831	0.03	no_wgd	FALSE

Taula 10 *- Mostra dades puresa i ploidia de PCAWGS.*

Dades d'expressió gènica - Gene expression

En aquest conjunt de dades hi ha recollida la quantificació de l'expressió gènica de les mostres obtinguda a partir de l'RNAseq i calculada mitjançant la mitjana de les estimacions d'expressió gènica derivades de l'alineació de TopHat2 i STAR en valors de nivell de gen fpkm del quartil superior normalitzat [18].

sample	D045145	D026512	D06350
ENSG00000000003	6.023	4.2452	-7.3673
ENSG00000000005	-9.9658	-6.5955	-9.9658
ENSG00000000419	4.5004	5.0261	4.3377

Taula 11 - *Mostra dades expressió PCAWGS.*

Inicialment les dades mostraven els noms dels gens en nomenclatura Ensembl (taula RNAseq_1 i RNAseq_2 de la base de dades) i per tal d'harmonitzar les dades, poder-les relacionar amb més facilitat i poder calcular altres variables a partir d'aquesta, es van traduir a nomenclatura Approved Gene symbol (taula EXPR_IPS_1 i EXPR_IPS_2 de la base de dades).

2.1.2. Dades Calculades

Després d'estudiar detalladament les dades proporcionades pel projecte PCAWGS, i d'escollir-ne les que més ens interessaven per proporcionar-nos la informació biològica rellevant per als nostres objectius, s'ha detallat el càlcul de varies variables més, a partir de les dades, amb l'objectiu de poder, inicialment, afegir aquestes noves variables genòmiques a la base de dades, per a posteriorment correlacionar-les, entre elles, i amb altres variables, amb la finalitat de respondre biològicament a l'efecte d'unes en les altres.

Les dues variables que s'han calculat són la composició immunitària del tumor, calculada a partir d'una adaptació pròpia de l'Immunophenoscore (IPS) [19] i el seu nivell d'aneuploïdia, calculada a partir dels CNA scores obtinguts a través de l'aplicació CNApp [20].

IPS

L'IPS és una puntuació agregada basada en l'expressió de gens o conjunts de gens representatius que comprenen quatre categories [19] la qual reflecteix l'activitat immunitària dels tumors, i esta calculada a partir de les dades d'expressió RNAseq. Les quatre categories són els següents:

- Molècules presentadores d'antígens (MHC)
- Inmunomoduladors i checkpoints immunitaris (CP)
- Cèl·lules immunitàries efectores (CE)
- Cèl·lules immunitàries supressores (SC)

Per tal de realitzar el càlcul d'aquest nou paràmetre amb les dades de PCAWGS, inicialment s'han transformat les taules d'expressió RNAseq de PCAWGS, per tal d'adaptar la nomenclatura dels gens. A continuació s'ha utilitzat un script d'R (adjuntat a l'Annex 9.3 - Codi R per al càlcul de l'IPS) proporcionat per l'equip de Recerca Traslacional en Càncer Hepàtic de l'IDIBAPS, prèviament adaptat i simplificat per tal de poder ser executat mitjançant el programari R. Amb aquest codi, i un fitxer en format text on es detallen diferents tipus de cèl·lules immunes, amb funcions diferents (adjunt a l'Annex 9.4 - Fitxer IPS_genes.txt) al introduir les nostres dades d'expressió extrèiem una taula resultant amb el càlcul d'un grup de scores i les categories nombrades anteriorment, per a cada mostra d'expressió.

CNA scores

Pel que fa a l'altre conjunt de variables calculades, es tracta de tres CNA scores, calculant el nivell de CNAs grans (broad CNA score o BCS), focals (focal CNA score o FCS) i global (global CNA score o GCS).
Aquests han estat calculats mitjançant les dades de CNA extretes del projecte PCAWGS i amb l'ajuda de l'aplicació CNApp [21]. Aquesta eina, realitza una anàlisi completa i integradora dels CNA mitjançant l'avaluació dels perfils i nivells de CNA, obtenint les tres puntuacions comentades anteriorment.

Per tal de realitzar el càlcul amb el CNApp, inicialment es fa una adaptació de les dades de CNAs mitjançant un petit codi de R (adjuntat a l'Annex 9.5 - Codi R d'obtenció del input pel CNApp). En aquesta adaptació, inicialment s'eliminen els *outliers*. Es defineix com a punt de tall, la desviació estàndard (mesura que quantifica la dispersió del conjunt de dades) multiplicada per dos, i per tant, s'eliminen totes aquelles mostres que presenten un nombre d'alteracions major a aquest resultat. També eliminem els registres que no corresponen a dades de copy number. Finalment, transformem el valor nombre de copia a valor log ratio, i donem format a les columnes que ens interessen, amb els noms de columna determinants per l'aplicació CNApp.
El resultat d'aquesta fase prèvia es pot veure en la imatge 12, on les dades estan preparades per ser introduïdes a continuació a l'aplicació CNApp.

ID	chr	loc start	loc end	seg mean
DO1000	16	7765935	8099989	1
DO1000	1	128900000	142602938	0.5849625
DO1000	16	27809800	27811540	0.5849625
DO1000	20	47794973	49713515	1.5849625
DO1000	22	49687501	51304565	-1
DO1000	16	26709827	27809786	1
DO1000	3	8459980	22349999	0.5849625
DO1000	16	26113874	26304247	1
DO1000	16	8099990	8249678	1.32192809

***Taula 12** - Dades processades preparades per introduir a CNApp.*

Amb motiu de la gran quantitat de dades, executem des del programa R, l'aplicació en local a l'ordinador, ja que des de la pàgina web té limitacions de mida de les mostres.
Quan accedim a la pantalla de l'aplicació, introduïm les nostres dades de CNA del projecte PCAWGS formatades, i executem l'aplicació.

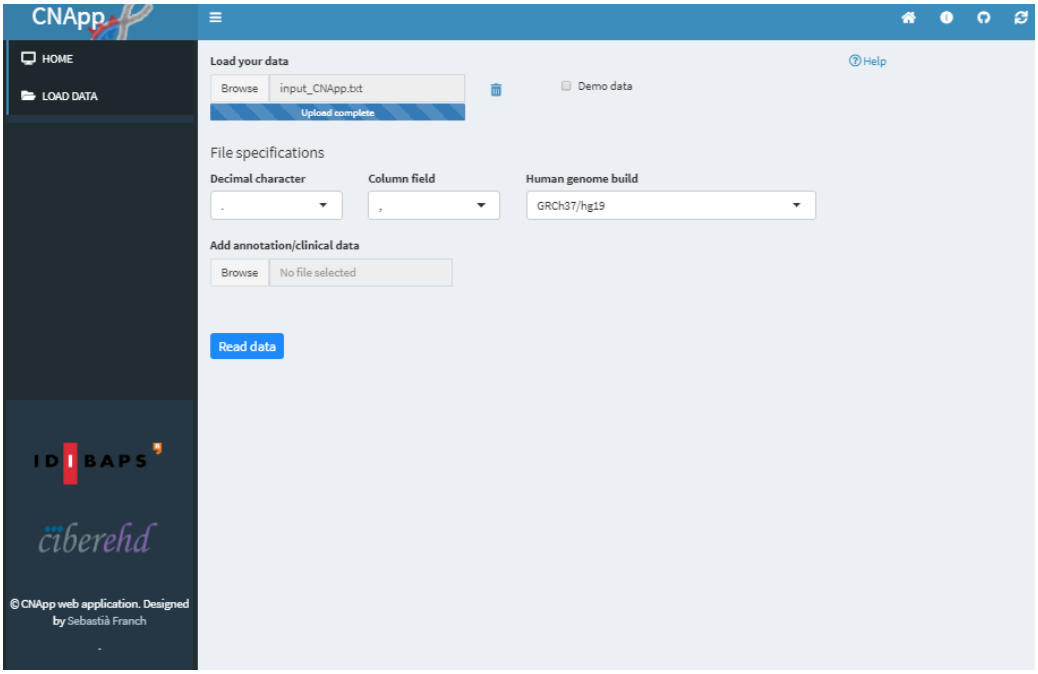


Figura 2 *- Aplicació CNApp.*

Després dels càlculs de l'aplicació, podem entre d'altres coses, veure diferents gràfics dels scores, i descarregar els scores calculats per a les nostres dades. A la taula 13 podem veure un exemple de les dades que hem calculat amb CNApp mitjançant les dades de copy number introduïdes.

sampleID	FCS	BCS	GCS
DO1000	299	71	2.86757576
DO1001	373	24	1.19330567
DO1002	328	19	0.75151068
DO1003	735	14	2.56389775
DO1004	551	56	3.47220949
DO1005	311	33	1.27537729
DO1007	337	98	4.23090739
DO1008	361	45	2.04643182
DO1009	419	18	1.16186453
DO1010	730	35	3.4519338

Taula 13 - *Resultat de dades calculades amb el CNApp.*

Totes les dades explicades anteriorment són les que s'han utilitzat per al desenvolupament de l'aplicació resultant d'aquest treball. Tot i això, a la base de dades creada a SQL Server s'han introduït altres dades d'interès de PCAWGS, les quals són:

- Phenotype – Overall survival
- Driver Mutations
- Gene Fusion RNAseq
- Mutational Signatures: Mutagenesis Analysis i Scores

2.2. Base de dades SQL Server

Per tal de integrar totes les dades i poder treballar amb elles amb facilitat, s'ha utilitzat SQL Server. Aquest és un sistema de gestió de bases de dades relacional de Microsoft, basat en el llenguatge de consulta Transact – SQL. SQL Server ens permet, alhora de poder gestionar grans quantitats de dades d'una forma senzilla, la connectivitat amb altres aplicacions per tal de poder mostrar i analitzar les nostres dades.

Per tal d'introduir les dades de PCAWGS a SQL Server, inicialment s'ha creat una base de dades amb el nom "PCAWGS" (veure Annex 9.6 – Backup base de dades PCAWGS). En aquesta base de dades, s'hi han creat les diferents taules amb tota la informació pertinent extreta de la pàgina web del projecte [2].

Depenent del tipus de dades, s'han utilitzats diferents procediments per a introduir-les a SQL. Generalment, aquest_procediment s'ha dut a terme mitjançant la opció "*Import data*" que proporciona SQL Server per introduir dades a una base de dades des de fitxers plans en format text o fitxers Excel.

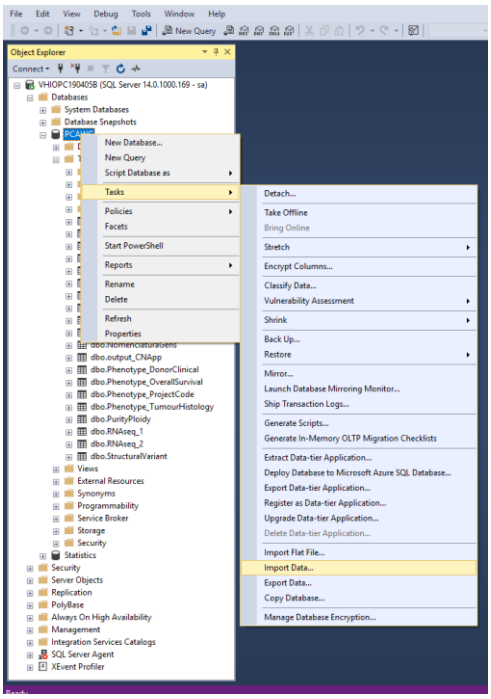


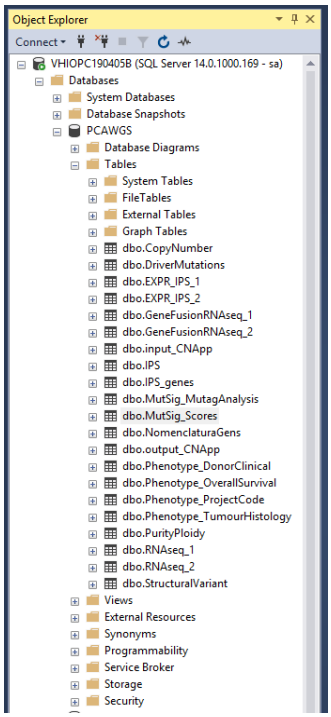
Figura 3 *- Procés d'importació de dades a SQL Server.*

Així doncs, introduint el fitxer de dades desitjat en cada cas i establint uns paràmetres determinats, les dades són introduïdes en una taula on a l'hora de pujar-la o posteriorment podem editar-ne el codi de creació per tal d'establir correctament els tipus de variables, noms, claus primàries, etc.

```
CREATE TABLE [dbo].[Phenotype_TumourHistology](
    [icgc_specimen_id] [nvarchar](255) NULL,
    [organ_system] [nvarchar](255) NULL,
    [histology_abbreviation] [nvarchar](255) NULL,
    [histology_tier1] [nvarchar](255) NULL,
    [histology_tier2] [nvarchar](255) NULL,
    [histology_tier3] [nvarchar](255) NULL,
    [histology_tier4] [nvarchar](255) NULL,
    [tumour_histological_code] [nvarchar](255) NULL,
    [tumour_histological_type] [nvarchar](255) NULL,
    [tumour_stage] [nvarchar](255) NULL,
    [tumour_grade] [float] NULL,
    [percentage_cellularity] [float] NULL,
    [level_of_cellularity] [nvarchar](255) NULL
) ON [PRIMARY]
GO
```

Figura 4 *- Exemple de codi per a la creació d'una taula a SQL Server.*

En altres casos, on el volum de dades era molt gran i aquesta opció de SQL Server no permetia manipular els fitxers, s'ha hagut d'utilitzar un script programat amb Python (veure Annex 9.7 - Codi Python per a carregar els conjunts de dades molt grans de PCAWGS a SQL Server), connectat a la base de dades y al fitxer de text contenidor de les dades, per tal de pujar les dades a SQL Server.



Finalment, a SQL Server aconseguim una base de dades com la que es pot veure en la figura 5, amb tota la informació recollida del projecte PCAWGS.

Figura 5 *- Taules a la base de dades final.*

En la figura 6 podem veure un diagrama de la base de dades ER (entitat – relació) que resumeix la informació de cada taula, així com les relacions entre les set taules que finalment s'han utilitzat per al desenvolupament de l'aplicació web. Centrades en la seva relació amb la taula del "Project Code" la qual hem utilitzat com a base per a recollir les identificacions de tots els pacients així com els tipus de càncer.

Figura 6 *- Diagrama ER base de dades.*

A sota podem veure també altres taules que s'han utilitzat per als càlculs de les taules finals, però que les seves dades no intervenen activament en la aplicació.

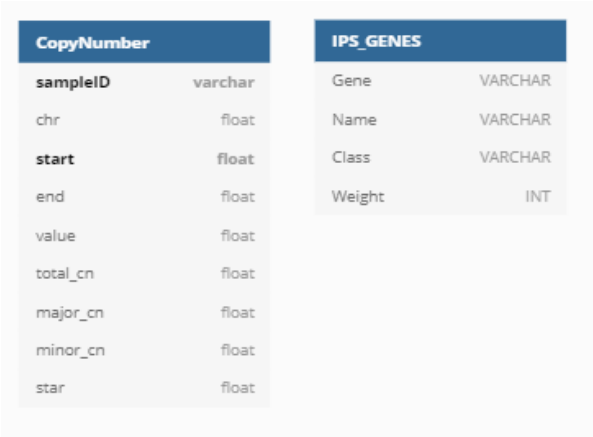


Figura 7 *- Taules inicials de PCAWGS utilitzades per a calcular les noves variables.*

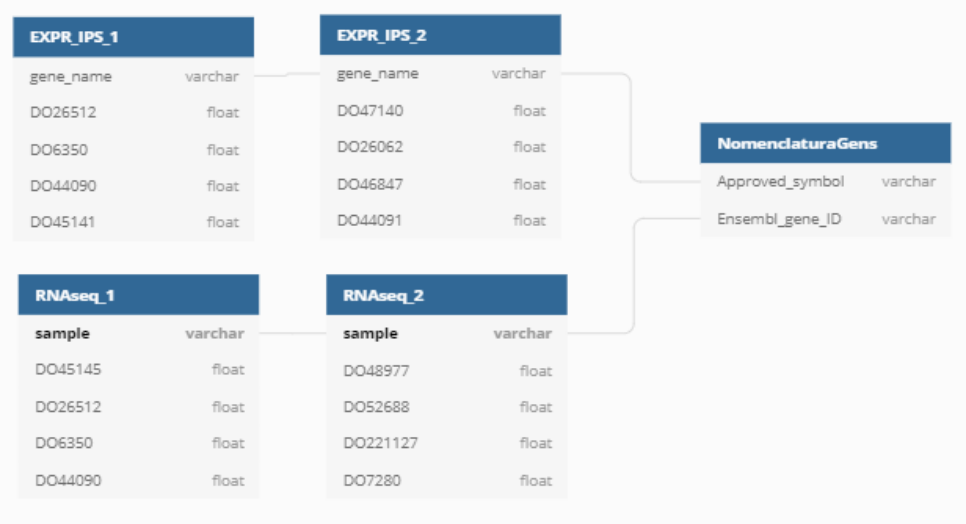


Figura 8 - *Taules d'expressió utilitzades per a calcular l'IPS.*

Cal tenir en compte que en els diagrames anteriors algunes de els taules simplement mostren una idea de l'estructura però degut a la gran quantitat de variables que tenen no hi estan totes especificades.

2.3. Programari

2.3.1. Django

L'aplicació web resultant d'aquest treball s'ha realitzat amb Django. Aquest és un framework de desenvolupament web de codi obert, escrit amb Python.

Django treballa amb l'estructura tradicional d'un lloc web, l'aplicació web espera peticions HTTP de l'explorador web, i quan n'arriba una, l'aplicació elabora el que es necessita segons la URL. L'aplicació retorna una resposta a l'explorador web, creant dinàmicament una pàgina HTML per tal de que l'explorador presenti les dades recuperades, des de la base de dades en el nostre cas.

L'aplicació creada agrupa el codi que gestiona aquests passos en diversos fitxers separats, que es relacionen entre ells tal i com es pot veure en el diagrama inferior.

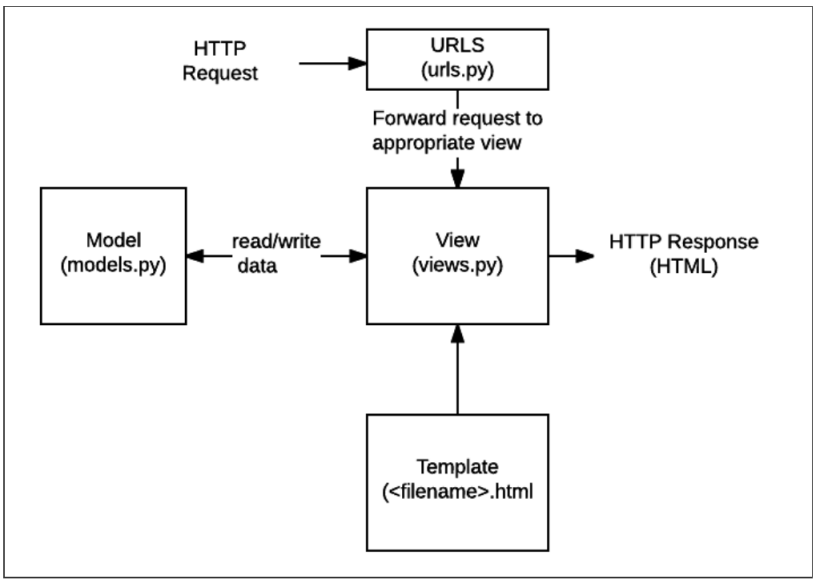


Figura 9 - *Esquema bàsic d'estructura d'una aplicació Django* [22] .

A continuació es defineixen aquests elements bàsics:

- **URLS:**

Aquest fitxer s'utilitza com a enllaç per tal de redirigir les peticions HTTP a la vista apropiada segons la URL de la petició.

- **Vista (view):**

Una vista és una funció que gestiona les peticions que arriben i com es retornen les respostes HTTP. Aquestes accedeixen a les dades necessàries i deleguen el format de la resposta a les plantilles (templates).

- **Models (models):**

Són objectes de Python que defineixen l'estructura de les dades de l'aplicació i proporcionen mecanismes per a gestionar-los, com per exemple fer consultes a una base de dades.

- **Plantilles (templates):**

Són plantilles HTML a partir de les quals es construeixen la resta de fitxers HTML mitjançant marcadors de posició que s'utilitzen per a representar el contingut real.

2.3.2. Visual Studio

Visual Studio és un entorn integrat de desenvolupament, que serveix per desenvolupar aplicacions.

S'ha utilitzat aquest entorn ja que disposa de les extensions necessàries per a poder desenvolupar una aplicació web amb Django.

2.3.3. GitHub

Durant el desenvolupament del projecte també s'ha fet ús de la plataforma GitHub, la qual permet allotjar projectes utilitzant el sistema de versions Git. Bàsicament, s'ha utilitzat per tal de mantenir el codi segur, i poder treballar d'una forma ordenada. Així com poder tenir documentats tots els canvis que s'han anat fent a mesura que s'ha anat desenvolupant l'aplicació web, així com les diferents versions que s'han anat creant.

2.3.4. R

Paral·lelament, al llarg del projecte, també s'ha fet us del programari R en diverses ocasions amb diferents finalitats.

D'una banda, en el cas dels CNA, per a pre - processar les dades extretes de PCAWGS abans i preparar el fitxer amb la informació necessària per a executar l'aplicació amb la que s'han calculat els CNA scores (CNApp).

S'ha utilitzat també per al càlcul de l'IPS, ja que l'script proporcionat, i posteriorment adaptat, es trobava en aquest llenguatge.

I finalment, durant tot el procés, s'ha utilitzat R puntualment degut a la major agilitat de treball amb aquest programari, per tal d'avaluar els resultats esperats. És a dir, per exemple, per fer proves amb les dades i representar alguns gràfics i així poder decidir quins tipus de gràfics eren els més interessants per a reproduir després a l'aplicació web.

2.3.5. Llibreries Javascript

- **Plotly**

És una llibreria molt complerta per a la representació de gràfics en aplicacions web [23].

- **Bootstrap**

Es tracta d'un conjunt d'eines de codi obert pel disseny d'aplicacions web. Conté plantilles de disseny, tipografies, formularis, botons, etc. basades en HTML i CSS [24].

2.4 Detall tècnic de l'aplicació

En aquest apartat es mostra gràficament l'esquema de l'aplicació i se'n detallen els component.

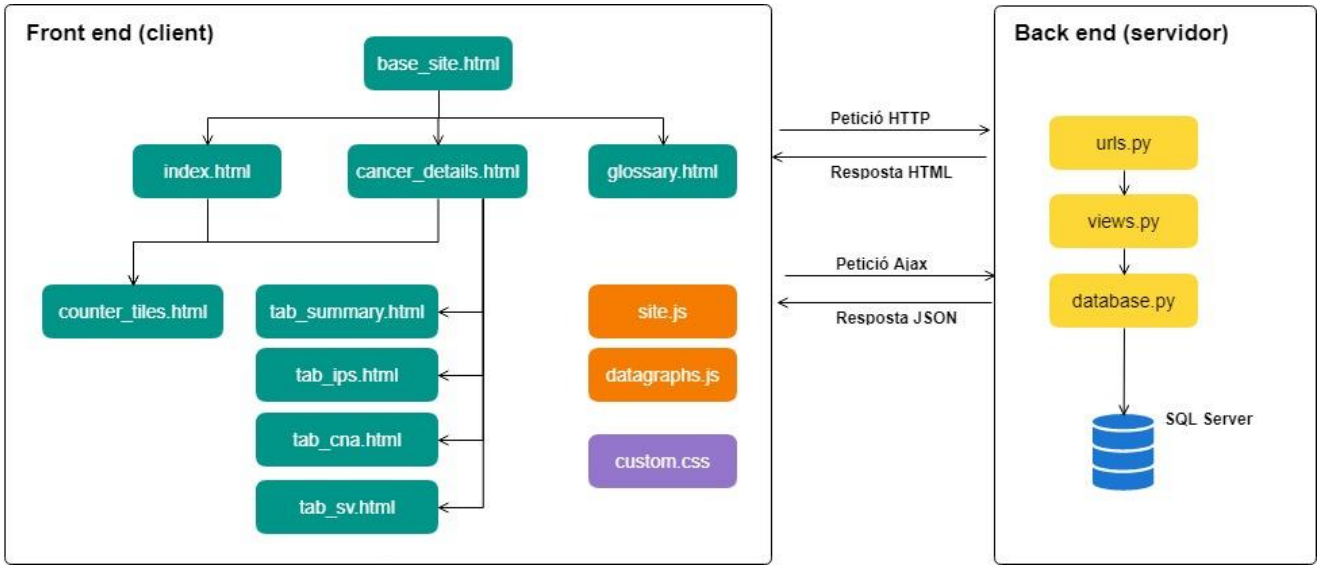


Figura 10 - *Esquema de l'aplicació web (site map).*

A continuació es detalla cada d'una forma molt general cadascun dels fitxers de l'aplicació creada. Per a més informació, es poden consultar a [l'Annex 9.8 - Codi aplicació web Django] on es troben complets i degudament comentats.

Els fitxers HTML de l'aplicació són:

base_site.html

Aquest fitxer HTML és al plantilla que forma l'estructura bàsica de l'aplicació web. Conté enllaços a la resta de pàgines. A més a més, conté les referencies a les diferents llibreries externes de JavaScript i els fulls d'estil.

cancer_details.html

Aquesta és la pàgina principal. S'hi construeixen les pestanyes (tab) dels diferents anàlisis: Summary, IPS, CNA i SV. Des d'ella es gestiona el títol de cada pàgina segons el tipus de càncer escollit, així com les dades que hi surten passant els paràmetres segons la URL de la pàgina. També es criden els gràfics per a que al carregar aquesta es carreguin. Aquest HTML fa servir els pertinents HTML per omplir el contingut de la pàgina.

counter_tiles.html

Es defineixen els comptadors de la capçalera de cada pàgina.

glossary.html

És la pàgina on es troba el glossari i les descripcions d'aquest.

index.html

Aquesta és la pàgina de benvinguda (Home) de l'aplicació web. En ella es crida el counter_tiles.html per a que carregui els dades generals de la capçalera, s'hi descriu l'aplicació al paràgraf "About", i s'hi introdueix en un "block" de JavaScript el gràfic amb el nombre de mostres per cada càncer.

tab_summary.html / tab_ips.html / tab_cna.html / tab_sv.html

Aquestes pàgines creen el contingut de les quatre seccions de la pàgina cancer_details.html.

Pel que fa als fitxers Python són els següents:

database.py

Aquest fitxer conté tots els accessos a la base de dades, i la crida a les llibreries pertinents per dur a terme aquests processos. Conté diferents funcions que fan les consultes a la base de dades, per a continuació processar-les i utilitzar-les a les representacions gràfiques. També és aquí on es duen a terme els càlculs de les línies de regressió i els test de Spearman. Les dades es retornen cap a l'aplicació web en una estructura JSON [25].

urls.py

Aquest fitxer conté la llista de URLs que connecten el client amb el servidor.

views.py

S'hi troben les funcions que gestionen les peticions que arriben del client i retornen respostes, algunes HTTP i d'altres en forma de JSON.

Pel que fa als fitxer JavaScript:

datagraphs.js

Realitza les crides al servidor mitjançant Ajax [26] (tècnica que permet actualitzar els continguts d'una pàgina sense recarregar-la) per tal de recuperar les dades en format JSON i mostrant-les en els gràfics. Aquí es construeixen els gràfics amb totes les seves característiques.

site.js

Conté utilitats per al comportament general de la pàgina: menús, desplegables, botons, etc.

I finalment hi ha un fitxer CSS:

custom.css

Es tracta d'una fulla d'estils (style sheet) que defineix l'aparença de l'aplicació (colors, tipografia, etc).

Un cop finalitzat el desenvolupament de la web, s'ha hagut de dur a terme la configuració per tal de poder veure la pàgina web públicament. Aquest passos, estan detallats en el manual d'instal·lació que es pot consultar a l'Annex 9.9 - Manual d'instal·lació d'una aplicació web Django a un servidor Windows.

3. Resultats i discussió

En aquets apartat es descriuen i mostren els resultats del projecte, centrant-nos en l'aplicació web dissenyada i en les conclusions biològiques que en podem extreure gràcies a aquesta.

El producte final d'aquest treball és una aplicació web que recull les correlacions entre l'IPS, les CNAs, la ploidia, la puresa, i les SV per a quaranta tipus de càncers diferents.

En aquesta hi trobem inicialment una pantalla principal, on podem veure a la capçalera un conjunt de nombres que ens resumeixen la quantitat de dades que estem analitzant en l'aplicació.

Exactament podem veure com hi ha un total de 2.809 individus dels quals s'han recollit mostres. Veiem també que la mitjana d'edat de tota la cohort és de 55.95, i que el rang d'edat és molt ampli ja que hi trobem pacients amb edats compreses entre 1 any i 90 anys.
Veiem també la distribució de gènere en la cohort general, la qual és molt paritària, ja que disposem de dades de 1.263 dones (44.96%) i 1.546 homes (55.04%). A nivell de resultats genòmics, hem incorporat les mostres amb dades de SV, 2.480 pacients tenen mostres d'aquest tipus, el que representa un 88.29% del total de pacients de la cohort. I les mostres amb resultats de CNAs, que són un total de 2.434, un 86.65% del total. D'altra banda veiem també quantificades les mostres amb dades d'expressió de les quals disposem, que són poc menys de la meitat de les mostres totals. Aquestes mostres d'expressió es refereixen a les directament extretes de PCAWGS, i amb les quals s'ha calculat l'IPS.

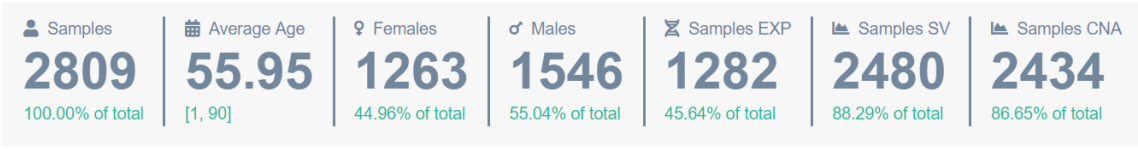


Figura 11 *– Comptadors capçalera general.*

Aquest conjunt de dades les podem trobar, com veurem més endavant, per a cada cohort analitzada individualment. D'aquesta manera es pot obtenir una visió general de la cohort d'interès, així com tenir clares les característiques bàsiques de les nostres anàlisis.

En aquesta pàgina inicial (Home) també hi podem veure una breu descripció de l'aplicació, i un gràfic de barres on es mostren tots els tipus de càncers analitzats i, el nombre de casos que en trobem en cadascun.

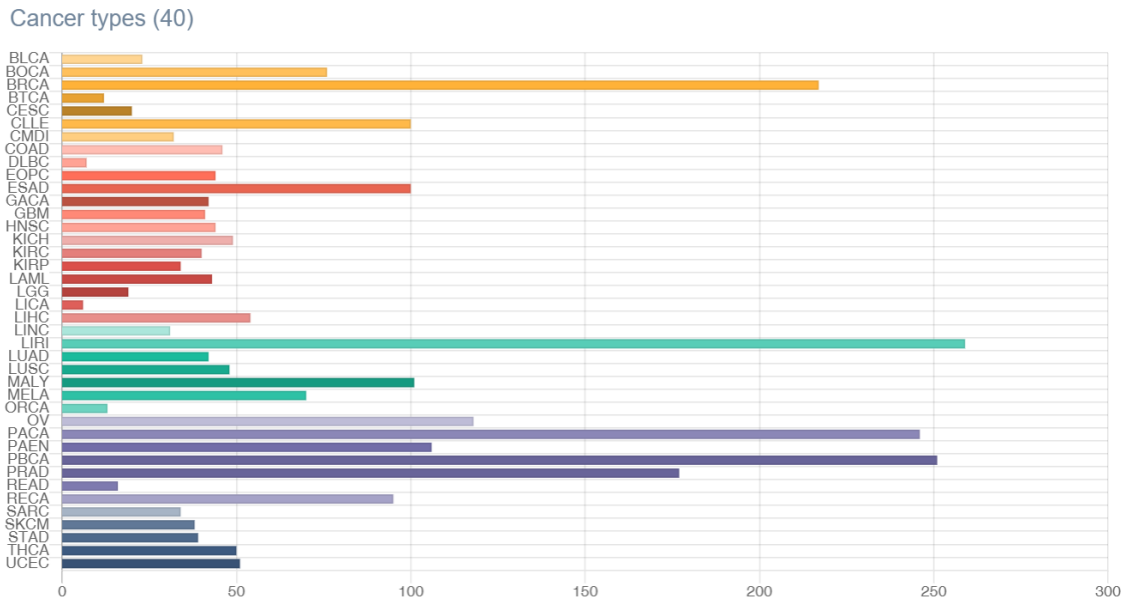


Figura 12 *- Gràfic pàgina benvinguda, nombre de casos per cada tipus de càncer del projecte.*

Al menú lateral, a més de la pàgina inicial "Home", visualitzem una pàgina anomenada "Glossary" on s'hi recullen petites descripcions dels termes més claus utilitzats en tota la resta de l'aplicació web. En aquesta pàgina, s'hi pot accedit directament, clicant l'acrònim o paraula que es desitja consultar des de la icona d'interrogant de l'esquerra de cada pestanya dintre dels tipus de càncers.

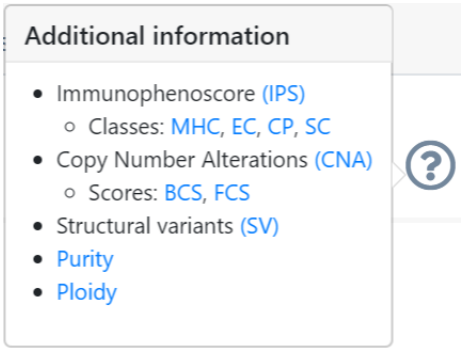


Figura 13 *- Informació addicional i links al glossari.*

Aquesta informació addicional és diferent segons la pestanya on ens trobem, ja que depenen de la informació que s'hi troba en cada pestanya.

Seguint amb el detall del menú lateral, es pot veure un desplegable anomenat "Primary site" el qual esta dividit en 22 submenús, corresponents als òrgans, on estan distribuïts els 40 tipus de càncers diferents estudiats. Desplegant cada òrgan, podem veure els tipus de càncer que s'hi ha classificat en ell, amb el nom i acrònim designat The Cancer Genome Atlas (TCGA).

Aquests submenús tenen dues variants, ja que, en cas de ser un òrgan que té tan sols un tipus de càncer associat en les dades PCAWGS, simplement surt aquest càncer i al clicar-hi veiem l'anàlisi per aquell tipus determinat de càncer. En canvi, si es tracta d'un òrgan que té associats més d'un tipus de càncer, es mostren els diferents tipus amb els seus respectius anàlisis individualment i a més a més, es mostra la opció "All" on podem veure l'anàlisi de totes les dades referents a aquell òrgan, és a dir, dels varis tipus de càncer conjuntament.

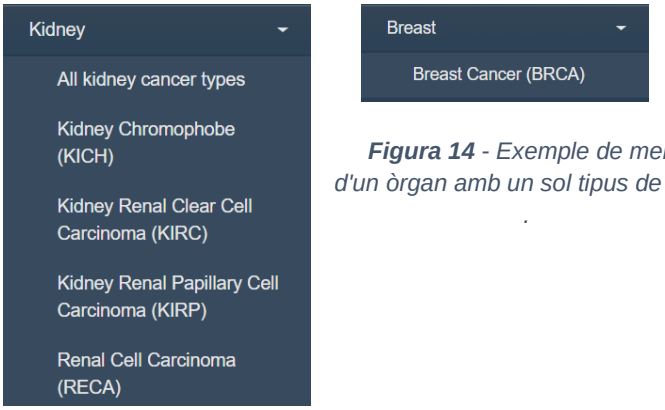


***Figura 14** - Exemple de menú d'un òrgan amb un sol tipus de càncer.*

***Figura 15** - Exemple de menú d'un òrgan amb varis tipus de càncers.*

A continuació, quan es vol inspeccionar la informació d'algun càncer en concret, clicant a sobre visualitzem una pàgina com la inferior.

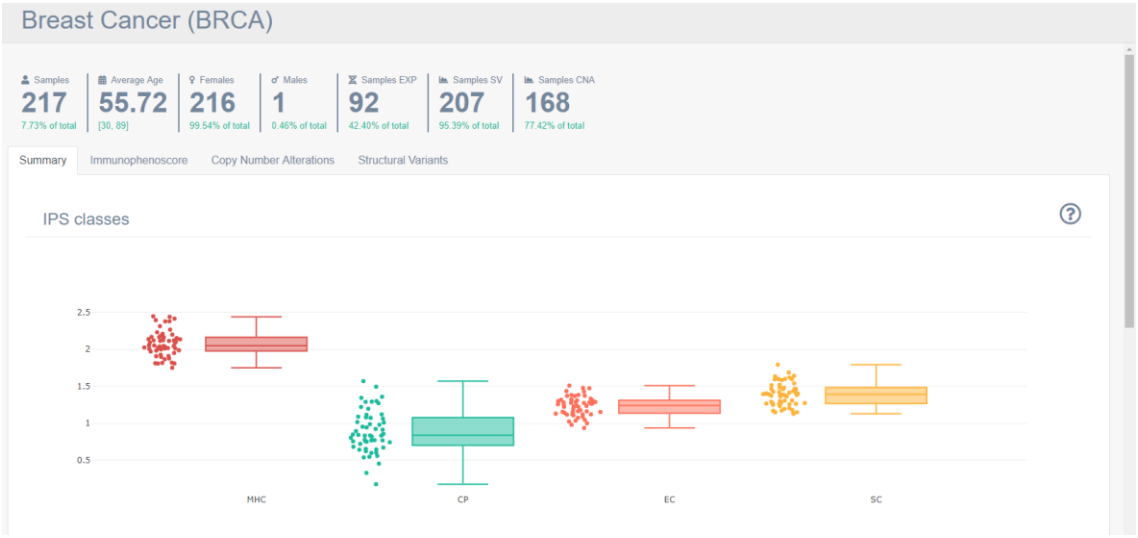


***Figura 16** - Exemple pàgina amb anàlisis d'un determinat tipus de càncer, BRCA.*

Primer de tot, veiem el títol superior on es detalla el tipus de càncer. Seguit a sota per els comptadors que ja s'han comentat de la pàgina inicial. Aquests comptadors però, ens donen informació de les mostres que tenim en aquest tipus de càncer determinat. Per exemple, observant la pàgina de *Breast Cancer* (BRCA) veiem que hi ha 217 mostres, les quals suposen un 7.73% del total de les mostres del projecte PCAWGS analitzades a l'aplicació web.
A continuació tenim la mitjana d'edat, que en aquest cas és de 55.72 anys, en un interval entre els 30 i els 80 anys.

Pel que fa a la distribució de gènere veiem com el 99.54% de les mostres de BRCA, és a dir, 216 són dones, i tan sols hi ha 1 home, és a dir, un 0.46%.
Dels valors que tenim a continuació, nombre de mostres d'expressió, de SV i de CNA, els percentatges que veiem a la part inferior corresponen al tant per cent sobre el nombre de mostres totals d'aquest tipus de càncer determinat.

Finalment, trobem les diferents pestanyes amb els anàlisis gràfics i estadístics correlacionant les diferents variables.
La primera pestanya, "Summary" esdevé un espai de descripció de les diferents variables que posteriorment es correlacionaran i conté quatre espais diferenciats. En el primer, veiem un boxplot amb les quatre categories d'IPS més importants, les quals són: MHC, CP, EC i SC.

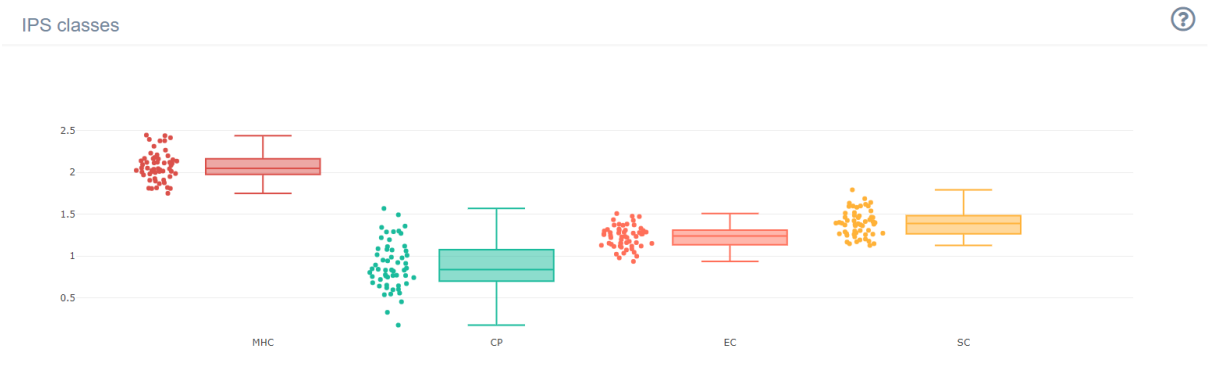


Figura 17 *- Boxplot categories IPS per BRCA.*

En el segon espai, observem les distribucions dels CNA scores, amb un histograma per al BCS i un altre per al FCS.

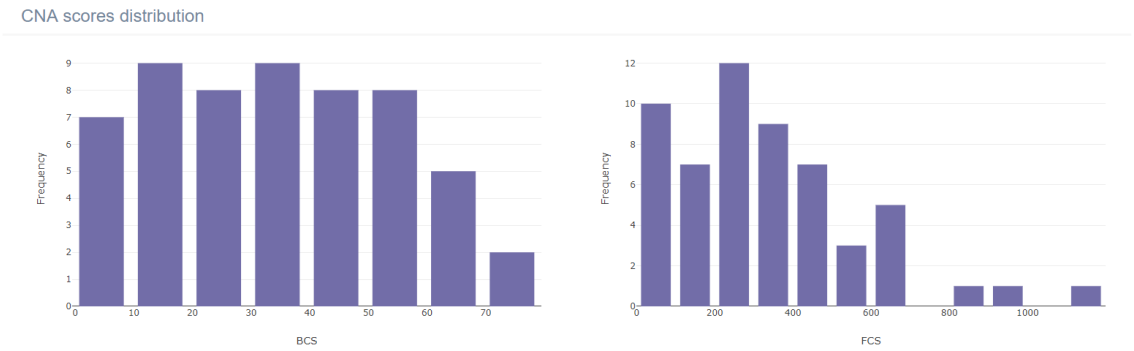


Figura 18 - *Histogrames CNA scores per BRCA.*

En el tercer, veiem un histograma també, aquest cop amb la distribució del nombre de SV en cada pacient.

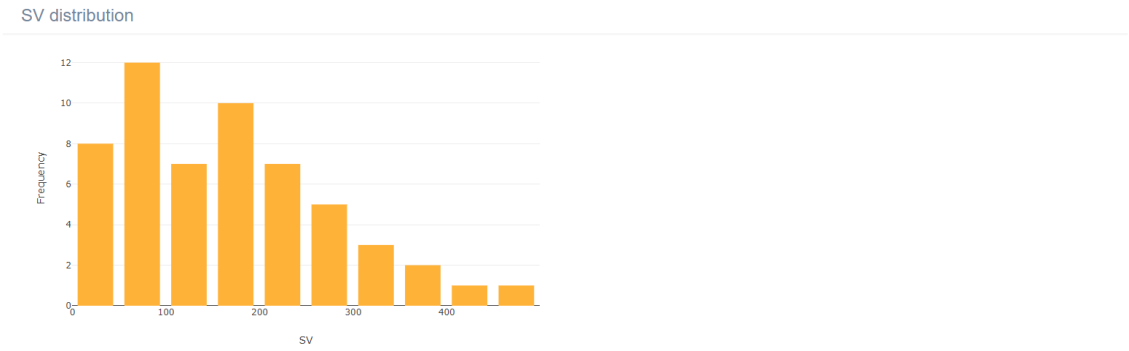


Figura 19 *- Histograma per distribució de nombre de SV, per BRCA.*

I finalment, les distribucions de la puresa i la ploidia.

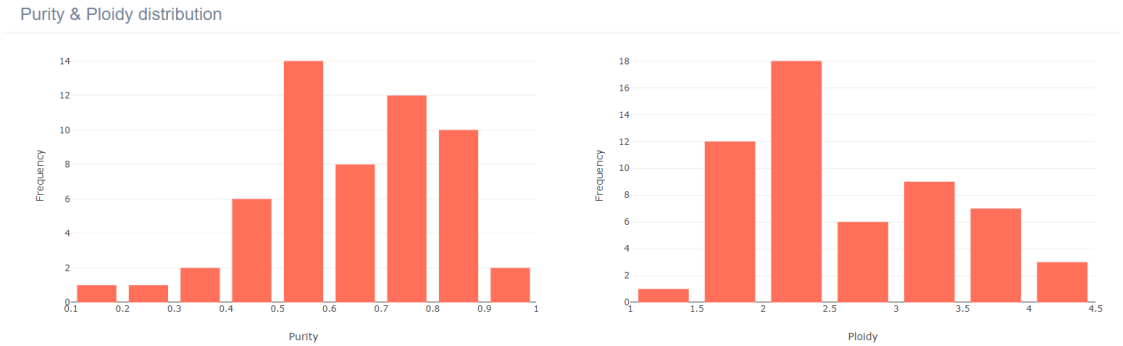


Figura 20 *- Histogrames de distribució de ploidia i puresa, per BRCA.*

Passant el ratolí per sobre del gràfic, a la part superior de cadascun d'ells trobem diferents botons que ens permeten interactuar amb el gràfic. Les opcions són tant com per descarregar el gràfic en format png, fer zoom, seleccionar una part del gràfic i visualitzar-la en detall, etc.

Figura 21 - *Botons d'interacció amb els gràfics.*

La següent pestanya és "Immunophenoscore". En ella, trobem dues seccions: les correlacions, per cada mostra tumoral, de les quatre categories més destacades de l'IPS amb els CNA scores i un resum en forma de taula dels resultats de les correlacions entre un conjunt més gran de puntuacions de l'IPS amb cadascun dels CNA scores. L'objectiu d'aquest apartat es investigar si existeix una associació (positiva o negativa) entre el nivell d'aneuploïdia dels tumors i el seu nivell d'infiltrat immunitari. El fet que hi hagi dos CNA scores, un per les alteracions grans i l'altre per les alteracions focals, ens permet explorar si, tal com s'ha descrit recentment, aquestes tenen un impacte funcional diferent [7].

Figura 22 *- Pestanya d'IPS amb gràfics de correlació entre la categoria d'IPS, MHC i els CNA scores, per BRCA.*

Els gràfics estan ordenats segons les categories de l'IPS, i es poden observar quatre parelles de gràfics correlacionant MHC, CP, EC i SC amb el BCS a l'esquerra i el FCS a la dreta.

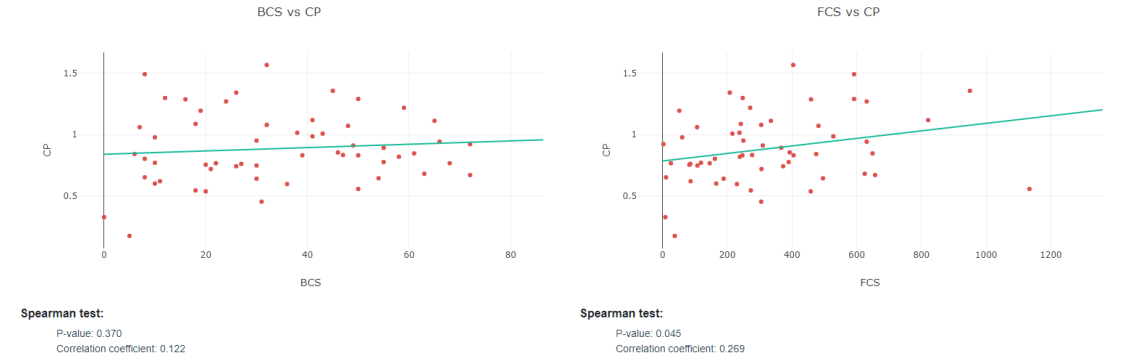


Figura 23 *- Gràfics de correlació entre la categoria d'IPS, CP i els CNA scores, per BRCA. Resultats dels Tests d'Spearman, p-valor i coeficient de correlació.*

A sota de cada representació de correlació hi trobem els resultats del test d'Spearman que s'ha realitzar per tal d'avaluar-ne el nivell de correlació [27]. Es mostren en cada cas el p-valor i el coeficient de correlació resultants. Aquests valors s'han obtingut mitjançant la llibreria **scipy.stats.spearmanr()** [28] de Python, en la qual s'han introduint els dos vectors de les variables que es volen correlacionar per a realitzar el càlcul.

La correlació de Spearman [29] és una mesura no paramètrica de la monotonicitat de la relació entre dos conjunts de dades. En aquest tipus de correlació, no s'assumeix inicialment que ambdós conjunts de dades segueixen una distribució normal, a diferencia d'altres tipus de correlació.

Pel que fa al coeficient de correlació, pot prendre valors dintre del rang -1 i +1, on un valor de 0 implica que no hi ha correlació entre els dues variables, i valors iguals a -1 i 1 indiquen una relació monotonica exacta. Així com una correlació positiva, implica que a mesura que x augmenta, y també augmenta, i en canvi una correlació negativa implica que si x augmenta, y disminueix.

Pel que fa al p-valor, ens indica la significança estadística dels resultats, avaluant una correlació com a significativa si el aquest és menor de 0.05.

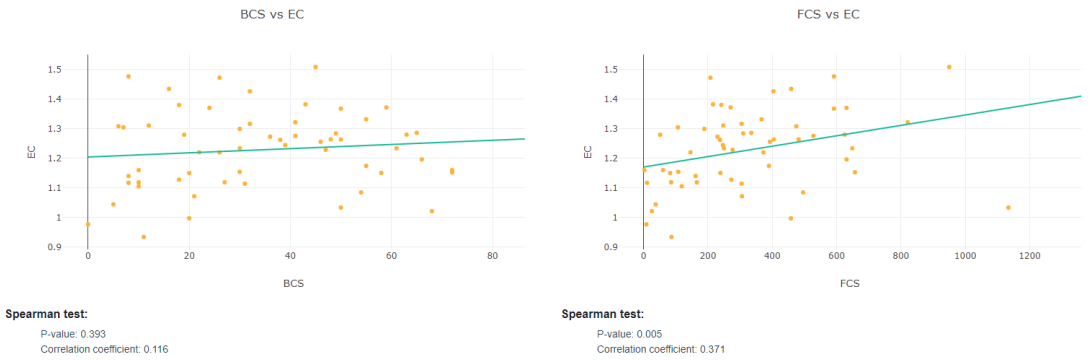


Figura 24 *- Gràfics de correlació entre la categoria d'IPS, EC i els CNA scores, per BRCA.*

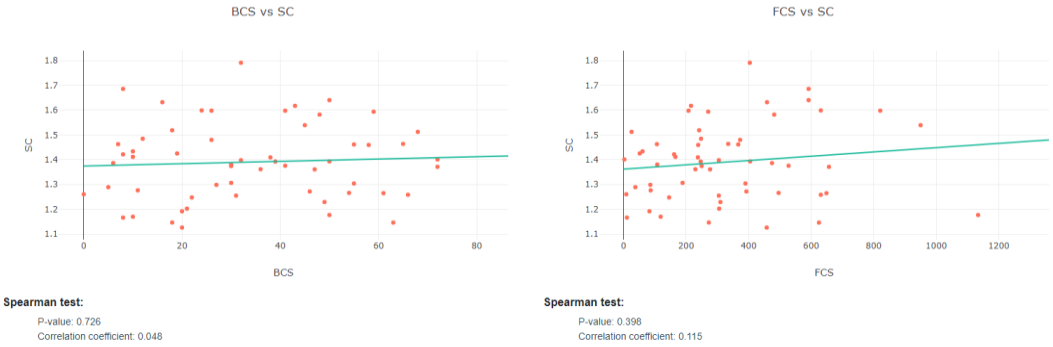


Figura 25 - *Gràfics de correlació entre la categoria d'IPS, SC i els CNA scores, per BRCA.*

Finalment, es completa la informació de l'IPS amb les correlacions de diferents scores d'aquest amb el BCS I FCS. Aquests resultats es mostren en una taula la qual permet el filtratge per a facilitar la cerca d'un score determinat, així com la descàrrega en diversos formats.

Correlations summary

IPS	BCS		FCS	
	p-value	cor. coef.	p-value	cor. coef.
b2m	0.654	-0.061	0.899	0.017
tap1	0.926	-0.013	0.175	0.184
tap2	0.933	0.011	0.031	0.289
hla_a	0.293	-0.143	0.643	0.063
hla_b	0.574	-0.077	0.346	0.128
hla_c	0.292	-0.143	0.782	-0.038
hla_dpa1	0.527	-0.086	0.706	-0.051
hla_dpb1	0.766	-0.041	0.910	-0.015
hla_e	0.884	-0.020	0.128	0.206
hla_f	0.528	-0.086	0.226	0.164

Showing 1 to 10 of 27 rows 10 rows per page ‹ 1 2 3 ›

Figura 26 *- Taula resum de correlacions entre IPS i CNA scores.*

A nivell d'interpretació biològica, els resultats de les correlacions entre els valors de l'IPS i els CNA scores ens serveixen per saber si el nivell d'aneuploïdia de les mostra tumorals d'una cohort te alguna influencia amb l'expressió de gens relacionats amb la presencia de cèl·lules immunitàries. És a dir, si per exemple observem en un subtipus tumoral que els valors obtinguts per la categoria d'expressió de cèl·lules MHC correlacionen positivament amb els valors de BCS (nivell de CNAs grans), podem afirmar que les alteracions en nombre de còpia grans podrien influenciar positivament la presencia de cèl·lules presentadores d'antígens i, per tant, l'efecte del sistema immunitari en el tumor.

La següent pestanya recull la informació referent al càlculs dels nivells de CNAs amb altres variables recollides. En aquesta hi podem veure els gràfics de correlació de BCS i FCS amb la ploidia i la puresa, amb els seus respectius resultats del test de Spearman.

Figura 27 *- Pestanya de CNA amb gràfics de correlació purity & ploidy i el BCS, per BRCA.*

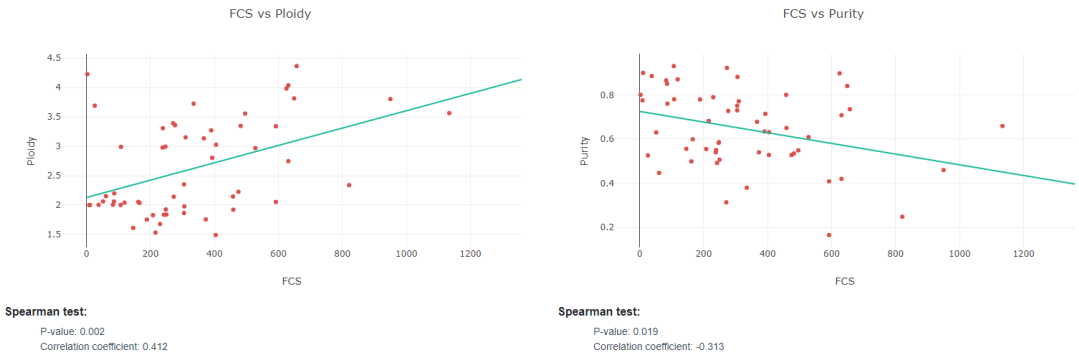


Figura 28 - *Gràfics de correlació purity & ploidy i el FCS, per BRCA.*

També en aquesta pestanya podem observar un gràfic que relaciona el nombre de CNAs per pacient, amb la ploidia. Aquest, s'ha dividit en quatre seccions (representades en diferents colors) segons el valor que pren la ploidia. Aquestes són: ploidia menor de 2, entre 2 i 3, entre 3 i 4 o major de 4.

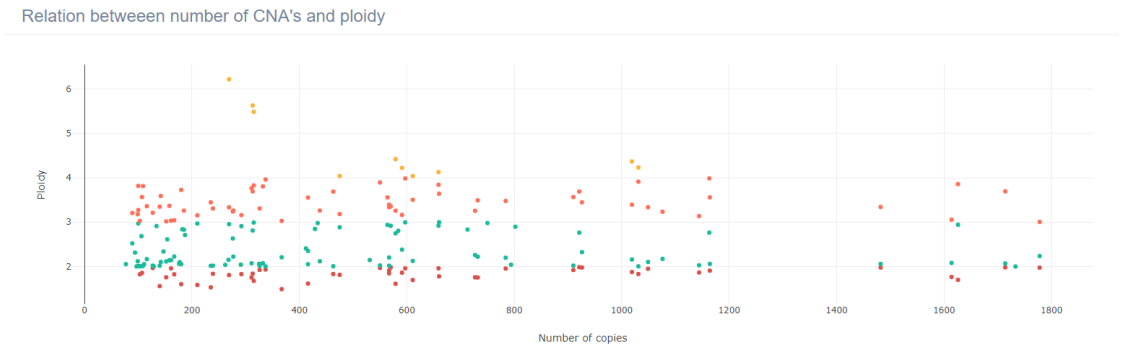


Figura 29 - *Relació entre el nombre de CNA i la ploidia, per BRCA.*

En la imatge inferior es mostra com gràcies als botons de la part superior dreta dels gràfics, podem seleccionar una part del gràfic per a veure-la més ampliada, resultat que veiem a la figura 30. Per tal de tornar al gràfic inicial tan sols es necessari fer doble click.

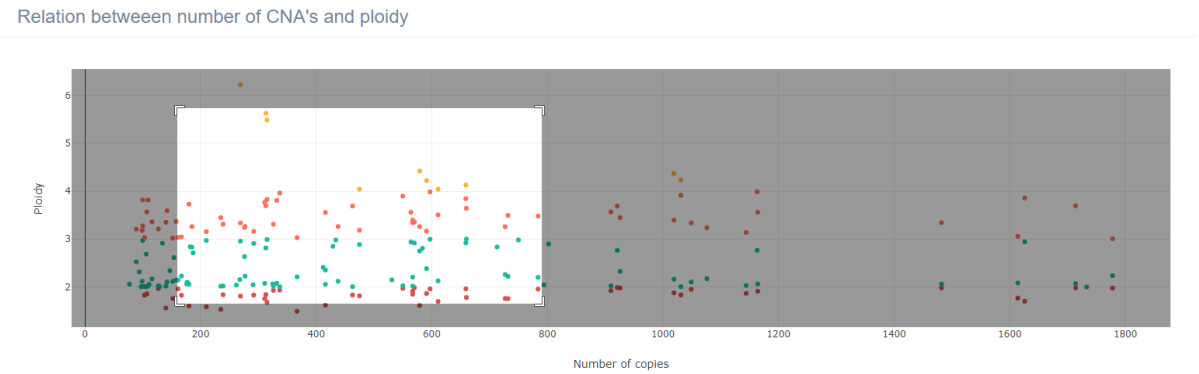


Figura 30 *- Mostra de les opcions per ampliar i interactuar amb els gràfics.*

Figura 31 *- Mostres d'interacció amb els gràfics.*

A la figura 31 també podem veure com, situant-nos sobre els punts del gràfics podem obtenir més informació sobre aquell valor.

Finalment, a la pestanya de “Structural Variants” veiem dues seccions. La primera presenta un gràfic circular amb les proporcions en els que es presenten els efectes (delecions, translocacions, duplicacions, etc.) de les variants estructurals en cada cohort.

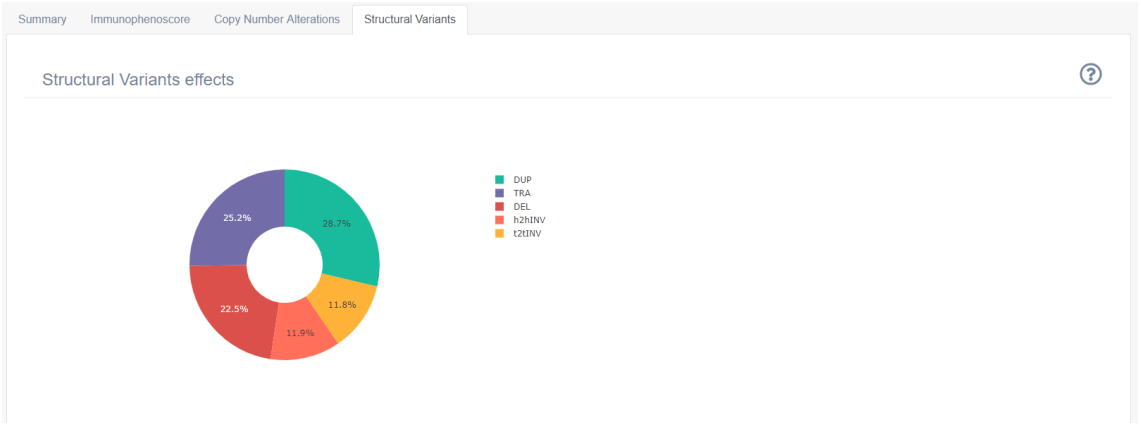


Figura 32 *- Pestanya SV amb gràfic circular dels diferents efectes.*

I finalment la correlació del nombre de SV per pacient amb la ploidia i la puresa, amb els seus respectius tests de Spearman.

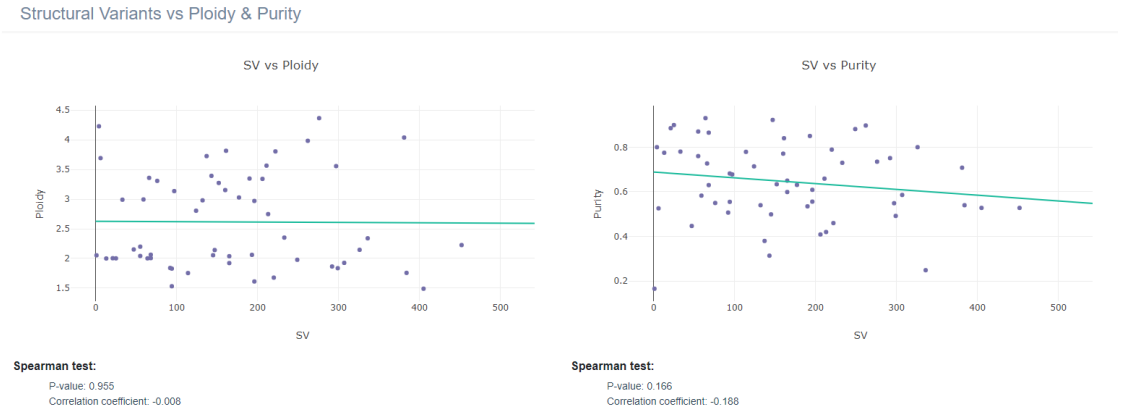


Figura 33 *- Gràfic de correlació entre nombre de SV i ploidy & purity.*

Com es pot comprovar amb els comptadors superiors, el nombre de mostres d'expressió, de SV i de CNAs, no sempre coincideix amb el nombre total de mostres per cohort, ja que alguns pacients no tenen dades informades per aquestes variables. Això no suposa cap problema pel que fa a CNAs i SV ja que no hi ha cap cohort amb zero mostres, tot i que per l'estudi de l'IPS si que hi ha alguns casos en que no hi ha cap mostra d'expressió en aquell cohort.

En aquests casos, s'ha ressaltat aquest comptador amb el zero en vermell, així com un missatge alertant de que no hi ha dades disponibles. Conseqüentment també s'ha eliminat la pestanya d'anàlisis de L'IPS.

En relació a aquest aspecte, i com es pot veure a la imatge inferior, també hi ha algun cas on les edats dels pacients no estan informades, pel que també apareix el comptador en vermell.

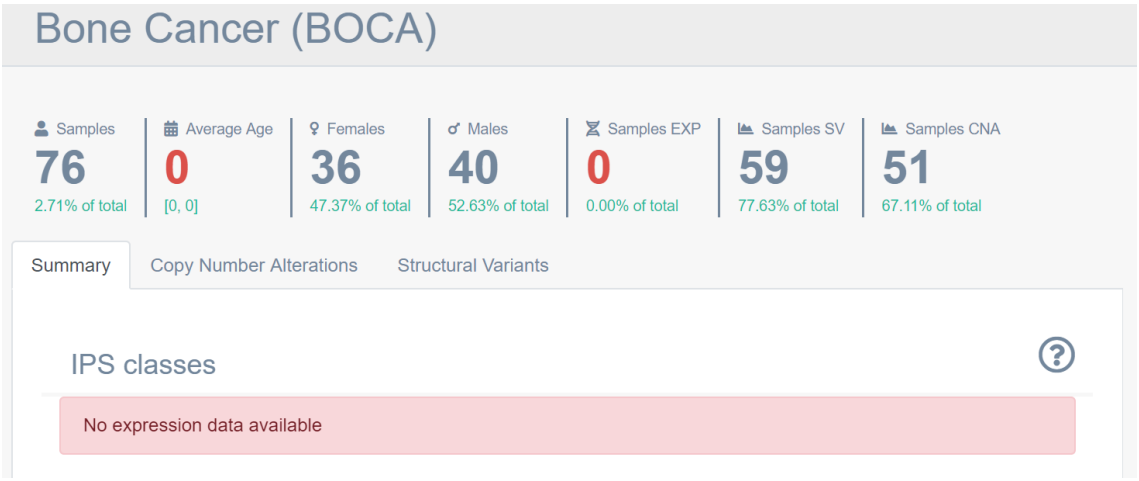


Figura 34 - *Mostra de la gestió de l'aplicació web a la falta de dades.*

4. Problemàtiques i solucions implementades

Al llarg del treball s'han trobat diferents dificultats que s'han hagut d'anar sortejant per tal de tirar endavant amb el projecte. La més repetida d'aquestes ha estat la dificultat del maneig de les dades degut a la gran quantitat d'aquestes.
Relacionat amb les dades també , i el qual és un problema molt comú en l'àmbit biomèdic, és que sovint aquestes són incomplertes. És molt complex obtenir el mateix tipus de dades, provinents de diferents plataformes d'anàlisis, per a tots els pacients que s'inclouen en un projecte. Així, en el projecte PCAWGS, i tot i ser un conjunt de dades molt processat i complet, hi ha molts camps buits per alguns individus. Aquest problema s'ha vist molt clar en les dades d'expressió, ja que com s'ha explicat anteriorment, en alguns tipus de càncer no hi havia pacients amb aquestes informades i per tant a l'aplicació web s'han hagut de gestionar aquestes situacions per tal d'evitar errors de visualització. Un altre cas per exemple ha estat el de les dades de l'edat dels individus, ja que no tots la tenien informada i per tant, en el càlcul de la mitjana d'edat que es duu a terme per cada tipus de càncer, en alguns casos aquest valor surt zero ja que no hi ha dades, o bé, s'han hagut de tractar les dades per tal de que els valors "null" que es rebien des de la base de dades no influenciessin en el càlcul de la mitjana de la resta.

Una altra problemàtica sorgida ha estat la tria de la llibreria més adient per a la representació dels gràfics. Inicialment, es va començar a treballar amb Chart.js [30] ja que eren uns gràfics amb moltes possibilitats, dinàmics i estèticament correctes. Més endavant, quan es van definir exactament quins tipus de gràfics s'havien de mostrar a l'aplicació web, va sorgir la necessitat de realitzar histogrames, i Chart.js no presentava aquesta possibilitat. És per això que es va decidir optar per Plotly (llibreria explicada anteriorment) ja que és una llibreria coneguda d'haver-la utilitzat amb R, i permet ser utilitzada en JavaScript i Python, també. A més a més, aquesta llibreria recull totes les opcions gràfiques que es necessitaven per al desenvolupament de la web, i un conjunt d'eines per una major interacció amb l'usuari. Això va produir haver de refer el codi dels gràfics en un moment determinat per adaptar-los a la nova llibreria.

També, i ja cap al final del projecte s'han tingut bastants problemes per fer visible la web des d'internet, sense disposar de la infraestructura necessària habitualment ni coneixements profunds en aquest àmbit. Tot i això, i gràcies a tenir el treball ja bastant avançat, s'ha pogut dedicar temps a això i fer que la web sigui visible.

5. Conclusions

Com a inici de les conclusions d'aquest treball, em sembla important destacar el que significa aquets treball a nivell de tancament del màster de Bioinformàtica i Bioestadística. Si més no, ja que crec que recull i consolida d'una forma homogènia gran part dels continguts del màster. Des de l'estudi i anàlisi de dades biològiques, tant genòmiques com clíniques, passant per l'anàlisi estadístic, la utilització de R, així com de plataformes de la comunitat científica per al compartiment de dades. I finalment la part de programació, l'ús de Python, de les diverses llibreries que s'han necessitat tant per a realitzar els gràfics com per a els càlculs estadístics, i la creació i gestió de bases de dades. Tot això va lligat directament a l'aprenentatge i sobretot, consolidació de coneixements que m'ha aportat aquest treball.

Comentar també la importància de les PACS, que m'han ajudat molt a mantenir-me constant, a organitzar-me i a tenir sempre una visió clara de com anava el treball. Així doncs, crec que la planificació s'ha seguit bastant fidelment, tot i que en alguns moments, les prioritats de les tasques han anat canviat i algunes tasques han acabat o començat abans o després del previst. En resum, a nivell global, la metodologia emprada ha permès que el volum de feina hagi estat constant al llarg de tots els mesos de duració del projecte i que aquest s'hagi pogut realitzar amb satisfacció.

Entrant en més detall en els objectius, crec que s'han assolit correctament, si més no, s'ha aconseguit crear una aplicació web on s'analitzen diverses variables d'interès del gran conjunt de dades que conté el projecte PCAWGS amb l'afegit de que s'han calculat i introduït altres variables d'interès biològic per aquest conjunt de dades, i se n'han estudiat les relacions entre elles. Tot això d'una manera simplificada i amigable.

Finalment, pel que fa a les línies de treball futur en parlaré en el següent punt més detalladament ja que la riquesa de les dades PCAWGS permet encara avançar més en el desenvolupament de l'aplicació web, introduint altres variables d'interès i explotant encara més la gran quantitat de dades de les que es disposa.

6. Treball futur

Donada la riquesa de les dades PCAWGS, com ja s'ha comentat abans, el nombre de possibles anàlisis és molt elevat i dona peu a l'estudi de moltes correlacions entre variables.
Al llarg del projecte han anat sorgint idees de diferents anàlisis que es podien realitzar amb les eines de les que disposàvem, tot i això, s'ha intentat centrar el focus del treball i concretar els estudis per tal de donar-li un sentit comú a l'aplicació.
Donat el temps limitat per realitzar el projecte, s'han prioritzat aquells anàlisis amb un interès biològic més actual i centrat els esforços en el desenvolupament de l'aplicació. Però hi ha varies línies obertes per tal d'acabar aconseguint una aplicació web més complerta i interessant.

Un d'aquests punts és el càlcul de la càrrega mutacional del tumor (TMB, de l'anglès Tumor Mutational Burden). Aquesta càrrega es defineix com la proporció de genoma alterat per SNVs i es considera, almenys per alguns subtipus de càncer, que està correlacionat amb la presència de neoantigens en el tumor i que serveix per a la predicció de la resposta als tractaments amb immunoteràpia. Aquesta variable normalment s'ha calculat a partir de mutacions no sinònimes, és a dir, mutacions que provoquen canvis d'aminoàcids en les seqüencies dels gens. Així, les dades de seqüenciació de l'exoma han estat àmpliament utilitzades amb aquesta finalitat. Ara bé, com que l'associació TMB-resposta a la immunoteràpia no acaba de ser del tot directe ni clara per tots els tipus de càncer, es reconeix que o bé hi ha altres mecanismes influenciant fortament la resposta al tractament, o bé el càlcul de la TMB també hauria d'incloure altres tipus de mutacions, com les que afecten a les regions no codificants. En aquest sentit doncs, la informació que es pugui extreure de les dades de WGS pot ser molt més rellevant [31][32].

De totes maneres, el càlcul del TMB és bastant complex i necessita d'un coneixement elevat de les seves característiques i de les dades amb les que es calcula. Però una de les possibles línies de futur seria calcular-lo i integrar-lo a la base de dades, per tal d'afegir-ne una pestanya amb correlacions amb les altres variables a l'aplicació web.
Així, s'obtindria una aplicació que correlacionaria tres variables molt importants actualment per a la genòmica del càncer: la presència d'infiltrat immunitari (mitjançant l'IPS), els nivells d'aneuploïdia dels tumors (a partir dels scores de CNAs), junt amb les dades ploidia i puresa, i la TMB, donant la possibilitat de veure com afecten les unes en les altres, en cada tipus de càncer.

Una altre possible afegit seria la creació d'un altre nivell d'anàlisi per sota dels tipus de càncer, afegint la divisió segons la histologia del tumor, i així poder visualitzar les correlacions de les variables en les diferents histologies de cada tumor. Això ens permetria fer un pas endavant en el tractament personalitzat dels tumors, ja que ens podria permetre sub-classificar pacients amb tumors en el mateix òrgan, per intentar discriminar els que podrien respondre a un tipus de teràpia o altra.

I finalment, la major optimització de l'aplicació web, tant a nivell d'estètica, com de fer-la més intuïtiva i amigable per a l'usuari, ja que aquests aspectes sempre és poden millorar.

7. Glossari

- **Alteracions en nombre de còpia (CNA)**

Les alteracions en nombre de còpia (CNAs) són alteracions de guanys i pèrdues de segments genòmics més grans d'1-kb.

- **Puntuació amplificada CNAs (BCS)**

Puntuació més especifica dels CNA que representa el nivell de CNAs amplificats.

- **Puntuació focal CNAs (FCS)**

Puntuació més especifica dels CNA que representa el nivell de CNAs focals.

- **Immunophenoscore (IPS)**

L'IPS és una puntuació agregada basada en l'expressió de gens o conjunts de gens representatius que comprenen quatre categories la qual reflecteix l'activitat immunitària dels tumors, i esta calculada a partir de les dades d'expressió.

- **Molècules presentadores d'antígens (MHC)**

Complex d'histocompatibilitat major que inclou un gran grup de glicoproteïnes de membrana altament polimòrfiques, implicades en el reconeixement de l'antigen i més específicament en la presentació de l'antigen.

- **Inmunomoduladors i checkpoints immunitaris (CP)**

Substància que estimula o deprimeix el sistema immunitari.

- **Cèl·lules immunitàries efectores (CE)**

Cèl·lules que realitzen una funció específica per respondre a un estímul. En general, descriu les cèl·lules del sistema immunitari.

- **Cèl·lules immunitàries supressores (SC)**

Tipus de cèl·lules immunitàries que impedeixen l'acció d'alguns altres tipus de limfòcits amb l'objectiu del sistema immunitari no es torna hiperactiu.

- **Variacions estructurals (SV)**

Les variacions estructurals (SVs) són aquelles variacions (inversions o translocacions) que no presenten canvi de dosi de material genòmic i el seu impacte es dona només en el punt de trencament del material genètic.

8. Bibliografia

[1] "About - ICGC DCC Docs." [Online]. Available: https://docs.icgc.org/pcawg/. [Accessed: 31-May-2019].

[2] "UCSC Xena." [Online]. Available: https://xenabrowser.net/datapages/?cohort=PCAWG (donor centric)&addHub=https%3A%2F%2Fpcawg.xenahubs.net&removeHub=https%3A%2F%2Fxena.treehouse.gi.ucsc.edu%3A443. [Accessed: 31-May-2019].

[3] J. R. Pollack, "Cancer genomics," in *The Molecular Basis of Human Cancer*, 2016.

[4] L. Chin, J. N. Andersen, and P. A. Futreal, "Cancer genomics: From discovery science to personalized medicine," *Nature Medicine*. 2011.

[5] S. Ferro, V. Huber, and L. Rivoltini, "Mechanisms of tumor immunotherapy, with a focus on thoracic cancers," *Journal of Thoracic Disease*. 2018.

[6] I. Nisbet, "Cancer immunotherapy comes of age (Finally!)," *Australas. Biotechnol.*, 2016.

[7] T. Davoli, H. Uno, E. C. Wooten, and S. J. Elledge, "Tumor aneuploidy correlates with markers of immune evasion and with reduced response to immunotherapy," *Science (80-.).*, 2017.

[8] M. Zarrei, J. R. MacDonald, D. Merico, and S. W. Scherer, "A copy number variation map of the human genome.," *Nat. Rev. Genet.*, 2015.

[9] J. L. Freeman *et al.*, "Copy number variation: New insights in genome diversity," *Genome Research*. 2006.

[10] N. P. Carter, "Methods and strategies for analyzing copy number variation using DNA microarrays," *Nat. Genet.*, 2007.

[11] "GenePattern." [Online]. Available: http://software.broadinstitute.org/cancer/software/genepattern/modules/docs/ABSOLUTE. [Accessed: 31-May-2019].

[12] K. Kleinheinz *et al.*, "ACEseq – allele specific copy number estimation from whole genome sequencing," *bioRxiv*, 2017.

[13] E. Battenberg and I. Bischofs-pfeifer, "A System for Automatic Cell Segmentation of Bacterial Microscopy Images," *Cell*, 2006.

[14] A. Fischer, I. Vázquez-García, C. J. R. Illingworth, and V. Mustonen, "High-definition reconstruction of clonal composition in cancer," *Cell Rep.*, 2014.

[15] Y. Cun, T. P. Yang, V. Achter, U. Lang, and M. Peifer, "Copy-number analysis and inference of subclonal populations in cancer genomes using Sclust," *Nat. Protoc.*, 2018.

[16] L. Feuk, A. R. Carson, and S. W. Scherer, "Structural variation in the human genome," *Nature Reviews Genetics*. 2006.

[17] G. Escaramís, E. Docampo, and R. Rabionet, "A decade of structural variants: Description, history and methods to detect structural variation," *Brief. Funct. Genomics*, 2015.

[18] P. G. Engström *et al.*, "Systematic evaluation of spliced alignment programs for RNA-seq data.," *Nat. Methods*, 2013.

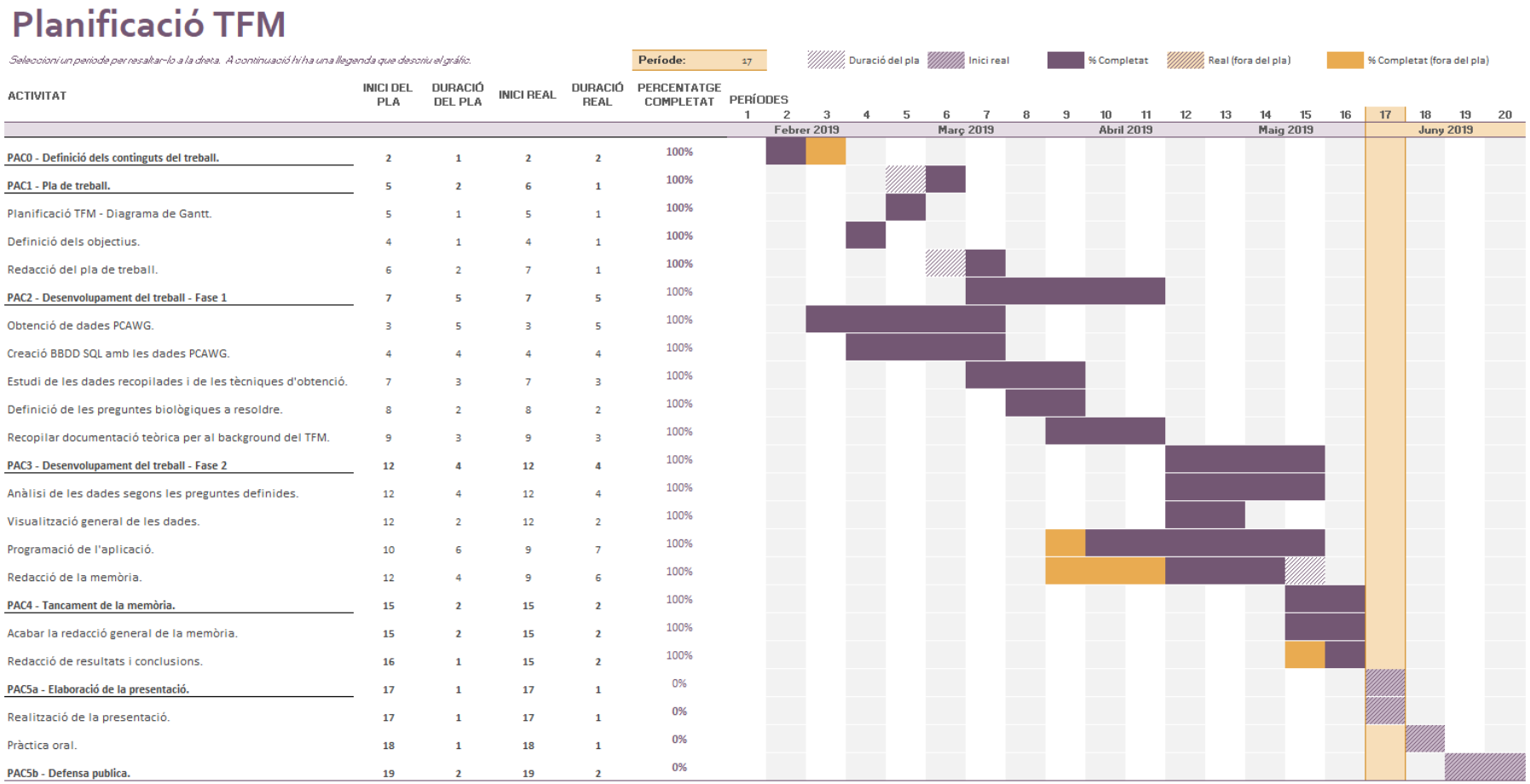
[19] P. Charoentong *et al.*, "Pan-cancer Immunogenomic Analyses Reveal Genotype-Immunophenotype Relationships and Predictors of Response to Checkpoint

Blockade," *Cell Rep.*, 2017.
[20] S. Franch-Exposito *et al.*, "CNApp: a web-based tool for integrative analysis of genomic copy number alterations in cancer," *bioRxiv*, 2018.
[21] "CNApp: copy number alterations integrative analysis." [Online]. Available: http://bioinfo.ciberehd.org/CNApp/. [Accessed: 30-May-2019].
[22] "Introducción a Django - Aprende sobre desarrollo web | MDN." [Online]. Available: https://developer.mozilla.org/es/docs/Learn/Server-side/Django/Introducción. [Accessed: 31-May-2019].
[23] "plotly.js | JavaScript Graphing Library." [Online]. Available: https://plot.ly/javascript/. [Accessed: 31-May-2019].
[24] "Bootstrap · The most popular HTML, CSS, and JS library in the world." [Online]. Available: https://getbootstrap.com/. [Accessed: 31-May-2019].
[25] "Implementations | JSON Schema." [Online]. Available: https://json-schema.org/implementations.html. [Accessed: 31-May-2019].
[26] "AJAX - Guía de Desarrollo Web | MDN." [Online]. Available: https://developer.mozilla.org/es/docs/Web/Guide/AJAX. [Accessed: 31-May-2019].
[27] W. Brown, G. H. Thomson, W. Brown, and G. H. Thomson, "Introduction to Correlation.," in *The essentials of mental measurement.*, 2006.
[28] "scipy.stats.spearmanr — SciPy v0.14.0 Reference Guide." [Online]. Available: https://docs.scipy.org/doc/scipy-0.14.0/reference/generated/scipy.stats.spearmanr.html. [Accessed: 31-May-2019].
[29] "Spearman's Rank-Order Correlation - A guide to when to use it, what it does and what the assumptions are." [Online]. Available: https://statistics.laerd.com/statistical-guides/spearmans-rank-order-correlation-statistical-guide.php. [Accessed: 31-May-2019].
[30] "Chart.js | Open source HTML5 Charts for your website." [Online]. Available: https://www.chartjs.org/. [Accessed: 31-May-2019].
[31] M. Allgäuer *et al.*, "Implementing tumor mutational burden (TMB) analysis in routine diagnostics—a primer for molecular pathologists and clinicians," *Transl. Lung Cancer Res.*, 2018.
[32] A. Stenzinger *et al.*, "Tumor mutational burden standardization initiatives: Recommendations for consistent tumor mutational burden assessment in clinical samples to guide immunotherapy treatment decisions," *Genes Chromosom. Cancer*, 2019.
[33] "mskilab/JaBbA", GitHub, 2019. [Online]. Available: https://github.com/mskilab/JaBbA. [Accessed: 30- May- 2019]

9. Annexos

9.1 Calendari del projecte

Aquest diagrama de Gantt es pot veure amb més detall a la imatge en format jpeg amb el nom “Annex_9.1” de la carpeta comprimida “Annexos_ PCAWGS” lliurada conjuntament amb aquest document.



ACTIVITAT	INICI DEL PLA	DURACIÓ DEL PLA	INICI REAL	DURACIÓ REAL	PERCENTATGE COMPLETAT
PAC0 - Definició dels continguts del treball.	2	1	2	2	100%
PAC1 - Pla de treball.	5	2	6	1	100%
Planificació TFM - Diagrama de Gantt.	5	1	5	1	100%
Definició dels objectius.	4	1	4	1	100%
Redacció del pla de treball.	6	2	7	1	100%
PAC2 - Desenvolupament del treball - Fase 1	7	5	7	5	100%
Obtenció de dades PCAWG.	3	5	3	5	100%
Creació BBDD SQL amb les dades PCAWG.	4	4	4	4	100%
Estudi de les dades recopilades i de les tècniques d'obtenció.	7	3	7	3	100%
Definició de les preguntes biològiques a resoldre.	8	2	8	2	100%
Recopilar documentació teòrica per al background del TFM.	9	3	9	3	100%
PAC3 - Desenvolupament del treball - Fase 2	12	4	12	4	100%
Anàlisi de les dades segons les preguntes definides.	12	4	12	4	100%
Visualització general de les dades.	12	2	12	2	100%
Programació de l'aplicació.	10	6	9	7	100%
Redacció de la memòria.	12	4	9	6	100%
PAC4 - Tancament de la memòria.	15	2	15	2	100%
Acabar la redacció general de la memòria.	15	2	15	2	100%
Redacció de resultats i conclusions.	16	1	15	2	100%
PAC5a - Elaboració de la presentació.	17	1	17	1	0%
Realització de la presentació.	17	1	17	1	0%
Pràctica oral.	18	1	18	1	0%
PAC5b - Defensa publica.	19	2	19	2	0%

9.2 Codi Python per adaptar les dades SV de PCAWGS

Aquest codi es pot trobar en un fitxer format text amb el nom “Annex_9.2” a la carpeta comprimida “Annexos_ PCAWGS” lliurada conjuntament amb aquest document.

9.3 Codi R per al càlcul de l’IPS

Aquest codi es pot trobar en un fitxer format text amb el nom “Annex_9.3” a la carpeta comprimida “Annexos_ PCAWGS” lliurada conjuntament amb aquest document.

9.4 Fitxer IPS_genes.txt

Aquest fitxer es pot trobar amb el nom “Annex_9.4” a la carpeta comprimida “Annexos_ PCAWGS” lliurada conjuntament amb aquest document.

9.5 Codi R d’obtenció del input pel CNApp

Aquest codi es pot trobar en un fitxer format text amb el nom “Annex_9.5” a la carpeta comprimida “Annexos_PCAWGS” lliurada conjuntament amb aquest document.

9.6 Backup base de dades PCAWGS

El backup de la base de dades es pot trobar en un fitxer format bak amb el nom “Annex_9.6” a la carpeta comprimida “Annexos_PCAWGS” lliurada conjuntament amb aquest document.

Per tal de carregar la base de dades a SQL Sever, simplement cal crear una base de dades, i amb el botó dret, seleccionant Tasks\Restore\Database, per finalment carregar el fitxer adjuntat a l’annex a l’opció Source – Device.

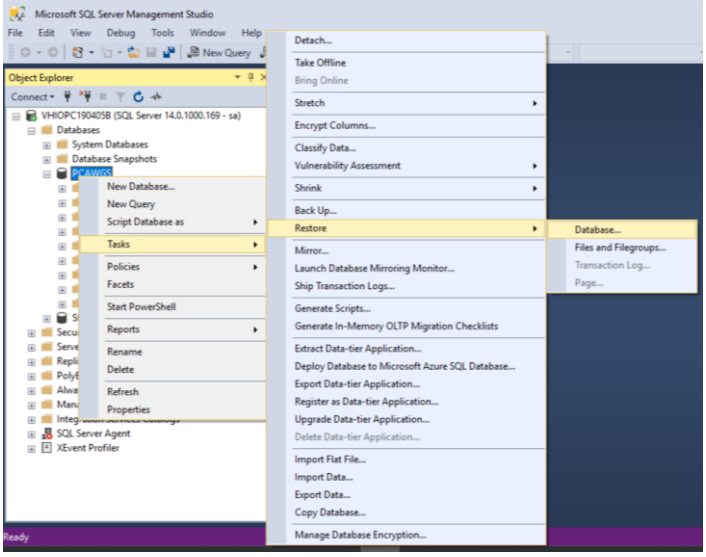


Figura 35 *– Càrrega de la BBDD a SQL Server.*

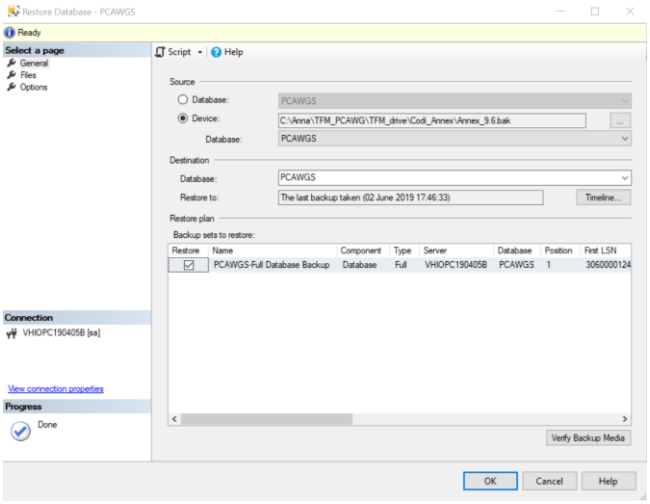


Figura 36 *- Càrrega del fitxer PCAWGS.bak*

9.7 Codi Python per a carregar els conjunts de dades molt grans de PCAWGS a SQL Server.

Aquest codi es pot trobar en un fitxer format text amb el nom “Annex_9.7” a la carpeta comprimida “Annexos_PCAWGS” lliurada conjuntament amb aquest document.

9.8 Codi aplicació web Django

Aquest annex consta de tots els fitxers descrits a l’apartat de descripció tècnica a més a més de la configuració necessària per a poder instal·lar l’aplicació, per tant s’hi troben fitxers en formats: Python, JavaScript, HTML i CSS entre d’altres.

Aquests fitxers es poden consultar a la carpeta comprimida “Annex_9.8” dins de la carpeta “Annexos_PCAWGS” lliurada conjuntament amb aquest document.

9.9 Manual d’instal·lació d’una aplicació web Django a un servidor Windows.

Aquest codi es pot trobar en un fitxer PDF amb el nom “Annex_9.9” a la carpeta comprimida “Annexos_PCAWGS” lliurada conjuntament amb aquest document.