II. IKERGAZTE
NAZIOARTEKO IKERKETA EUSKARAZ

2017ko maiatzaren 10, 11 eta 12
Iruñea, Euskal Herria

ANTOLATZAILEA:
Udako Euskal Unibertsitatea (UEU)

OSASUN ZIENTZIAK

Analisi proteogenomikoak minbiziaren ikerkuntzan

Alba Garin-Muga, Elizabeth Guruceaga, Fernando José Corrales eta *Victor Segura*

94-101 or.
https://dx.doi.org/10.26876/ikergazte.ii.04.12

Analisi proteogenomikoak minbiziaren ikerkuntzan

Garin-Muga Alba[1] **Guruceaga Elizabeth**[1,3] **Corrales Fernando**[1,2,3]

Segura Victor[1,3]

[1] Proteomika eta Bioinformatika saila. Medikuntza Aplikatuko Ikerketa Zentroa (CIMA). Iruñea (Nafarroa)
[2] Hepatologia eta Terapia Genikoko saila. Nafarroako Unibertsitatea (UNAV). Iruñea (Nafarroa)
[3] Nafarroako Osasun Ikerkuntzaren Institutua (IDISNA). Iruñea (Nafarroa)

Laburpena

Proteogenomika, proteomika eta genomika elkartzen dituen arloa, arrakastaz erabili ohi da minbizi-genometan gordetzen dituzten mutazioak identifikatzeko. Minbizi laginetan dauden mutazioak identifikatzeko gauzatu diren errendimendu handiko esperimentuek, *The Cancer Genome Atlas* (TCGA) gisako proiektuen garapena ahalbidetu dute. Proiektu hauetako datuak erabiliz, proteina sekuentzien datu-baseak sor ditzazkegu eta datu-base hauetako sarrera bakoitza minbizia mota jakin batekin zerikusia duen mutatutako peptido jakin bati dagokio. Artikulu honetan, bioinformatikako lan-fluxu bat deskribatuko dugu, zeina datu-base hauek sortzeko erabiltzeaz gain minbizi-laginetako mutatutako peptidoak proteomika esperimentuetan detektatzeko erabiliko dugun. Proposatutako metodoa lau minbizi desberdinen zeluletan atondutako datu publikoak erabilita balioztatuko da.

Hitz gakoak: Proteogenomika, minbiziaren ikerkuntza, peptido mutatuak

Abstract

Proteogenomics, the field that combines genomics and proteomics, has been successfully applied to the identification of mutations present in cancer genomes. High throughput experiments for genome-wide detection of mutations in cancer samples has allowed the development of projects such as the TCGA. Using these data, we can generate protein sequence databases in which each entry corresponds to a mutated peptide associated with certain cancer types. In this article, we describe a bioinformatics workflow for creating these databases and detecting mutated peptides in cancer samples from proteomic experiments. The performance of the proposed method has been evaluated using publicly available datasets from four cancer cell lines.

Keywords: *Proteogenomics, cancer research, mutated peptides*

1 Sarrera eta motibazioa

Minbizia, munduan gertatzen diren heriotzen eragile nagusiena da, heriotza guztien %15a delarik. Bestalde, datozen bi hamarkadetan, minbiziaren erruz heriotza-indizea bikoiztu egingo dela uste da. Gaur egun 200 minbizi mota baino gehiago ezagutzen dira eta hauek, euren ezaugarri molekular eta klinikoetan oinarrituta hainbat modutan sailka daitezke (Tomczak *et al.*, 2015). Minbizi mota hauetako bat baino gehiago, DNAren sekuentzian gertatzen diren mutazioengatik eraginda daude, mutazio hauek zelulen kontrolik gabeko hazkundea hasten bait dute gaixoetan, besteak beste. Beraz, minbiziaren prebentzio eta detekzio goiztiarraz gain, tratamendua ere hobetzeko, DNAn gertatzen diren mutazio guztiak identifikatzea behar beharrezkoa da.

Gizakiaren biologia konplexua ulertzeko, giza genomaren sekuentziazioa izan zen eman zen lehenengo pausoetako bat (Lander *et al.*, 2001; Venter *et al.*, 2001). Azken urteotan, errendimendu handiko teknologien garapenak giza proteomaren eta genomaren ikuspegi orokorra hobeto ezagutzeko aukera eskaini digu (Chin *et al.*, 2011). Sekuentziazio masiboak (*Next-generation sequencing*, NGS) mutazio genomikoen aurkikuntza bizkortu egin du (Meyerson *et al.*, 2010; Trapnell *et al.*, 2013) eta genoma- eta exoma-osoen sekuentziazioak dira gaur egun teknologia baliotsuenak giza gaixotasunen diagnosi eta tratamenduentzat, bereziki minbiziaren kasuan. Minbizi genomaren ikerkuntzan egin diren aurrerapenei esker, gaixotasun hauen inguruan karakterizazio proiektu handiak sortu dira:

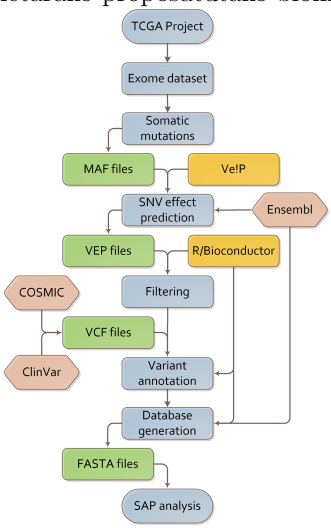
Cancer Genome Project (CGP[1]), *International Cancer Genome Consortium* (ICGC[2]) edota *The Cancer Genome Atlas* (TCGA[3]) izenekoak. 2006. urtean jaio zenetik, TCGA proiektuak, errendimendu handiko sekuentziazioa erabiliz, 34 minbizi motako 10000 tumore lagin baino gehiago karakterizatu ditu jada. Bioinformatika datuak aztertu ostean, proiektu honetan 10 milioi mutazio baino gehiago identifikatu dituzte.

2 Arloko egoera eta ikerketaren helburuak

Proteogenomika, proteomika, genomika eta transkriptomikako ikerketa elkartzen dituen arloa da. Proteogenomikaren helburu nagusia funtzio zelularrak guztiz ulertzea da (Faulkner *et al.*, 2015; Nagaraj *et al.*, 2015). Masa-espektrometriaren (MS) bidez identifikatutako peptidoak, genomikako sekuentziazio datuekin lerrokatzen dira ezagutzen diren geneak egiaztatu edo gene berriak identifikatzeko (Ansong *et al.*, 2008). Nahiz eta hainbat esparrutan aplikatu proteogenomikako ezagupena, minbiziaren arloan egin dira aurrerapenik esanguratsuenak, hain zuzen ere, tumoreek proteoman eragindako aldaketen gaineko ikerkuntzan. Hala nola, aminoazido bakarreko polimorfismoek (*Single amino-acid polymorphism*, SAP) edo mutatutako peptidoek gaixotasunaren hasiera edo tratamenduarekiko erantzunean eragin dezakete (Alfaro *et al.*, 2014; Zhang *et al.*, 2014). Guzti honetarako, proteomikan beharrezkoak diren neurrira egindako peptido datu-baseen sorkuntza izan da minbizi datuen analisiaren fase erabakigarriena. Datu-base hauek DNA edo RNA sekuentziazioak erabiliz sortu dira (Woo *et al.*, 2014; Nesvizhskii, 2014; Wang eta Zhang, 2013). Ondorioz, TCGA bezalako proiektuetan dauden datu kopuru handiak, baliabide botoretsua dira batez ere metodologia hau klinika onkologikoan ezartzeko. Ildo honetan, *Human Proteome Project* (HPP) deritzon ekimenak (Legrain *et al.*, 2011), transkriptomikako eta proteomikako datuen uztarketan jarduten du (Paik eta Hancock, 2012; Segura *et al.*, 2013) eta proteogenomikako metodoak eta erreminta bioinformatikoak garatzea da euren betebehar garrantzitsuenetako bat (Krasnov *et al.*, 2015; Tabas-Madrid *et al.*, 2015).

Artikulu honetan, proteogenomikarako datu-baseak sortzeko beharrekoa den lan-fluxu bioinformatikoaren garapena aurkezten dugu. Lan-fluxu honen helburua, TCGA proiektuan bilduak dauden DNAko mutazioak erabiliz, minbizi laginetan dauden proteinetan jasotzen diren aminoazidoen aldaketak identifikatzea da (1 irudia).

1 Irudia: Proiektu honetarako proposatutako bioinformatikako lan-fluxua



2.1 The Cancer Genome Atlas (TCGA)

The Cancer Genome Atlas (TCGA) proiektua, *National Cancer Institute* (NCI) eta *National Human Genome Research Institute* (NHGRI) izeneko erakundeen arteko elkarlana da. Proiektu honen xede nagusia, gaixotasunen oinarri molekularrak ulertzeko, hainbat minbizi motetan gertatzen diren aldaketa genomikoak deskribatzea da (Tomczak *et al.*, 2015). Helburu hau lortzeko, etekin altuko teknologiak erabili dira, mikroarraiak edo sekuentziazio masiboa esaterako.

[1] `http://www.sanger.ac.uk/genetics/CGP`
[2] `https://dcc.icgc.org`
[3] `http://cancergenome.nih.gov`

3 Ikerketaren muina

Minbizi laginetan mutatutako peptidoak aurkitzeko, errendimendu altuko proteomikako esperimentuen analisia egin genuen TCGAko datu multzoekin sortutako FASTA datu-baseak erabiliz. Atal honetan, 4 minbizi motatako zelulen proteogenomika azterketa azaltzen dugu. Kasu honetan, *shotgun* esperimentu publikoak erabili genituen sortutako datu-baseetan mutatutako peptidoak Mascot bilatzailearekin aurkitzeko. Izan ere, gure lan-fluxu proteogenomikoa egin daitekeela eta emaitzak zentzuzkoak direla ziurtatu nahi izan genuen.

3.1 Peptido mutatuen datu-baseen sorkuntza

Laginak aztertzen hasi baino lehen, peptido mutatuen datu-basea sortu genuen. Hortarako, TCGAn erabilgarri dauden datuetatik, 31 minbizi motari dagozkion 10183 exoma datu erabili genituen peptido mutatuen datu-basea sortzeko. Hiru minbizi motari dagokien exoma datuak ez zeuden eskuragarri. Gene-mutatuak dituzten artxiboak TCGAko web gunetik lortu ziren eta datu hauek gure departamentuan landutako R lengoaieko *script*-ak erabiliz prozesatu ziren. Guztira 3009480 aldaera identifikatu ziren eta hauetatik, gibeleko kartzinoma izan zen aldaera gehien (956761 mutazio) eta ubeako melanoma berriz aldaera gutxien azaldu zituena (3918 mutazio).

Mutazioen eragina aurreikusteko, prozesatutako mutazioen informazioa VEP erremintarekin osatu zen (McLaren *et al.*, 2010). VEP erremintak, geneetan gertatzen diren mutazioak aminoazidoen mailara bihurtzen ditu eta mutazio hauetako bakoitzak duen eragina bi puntuazio mota emanez deskribatzen du (Kumar *et al.*, 2009; Adzhubei *et al.*, 2010). Informazio hau erabiliz, mutazio mota guztietatik, sinonimoak ez diren mutazioak aukeratu genituen R/Bioconductor lengoaia erabiliz. Mutazio hauek, genetatik proteinara doan prozesuan aminoazido aldaketa bat eragiten dutenak dira. Honen ondorioz, bariante guztietatik sinonimoak ez diren mutazioak aukeratu ostean 1161751 aldaera identifikatu genituen.

Datu-basea sortzeko, jarraian azaldutako prozesua jarraitu genuen. Lehenik eta behin, proteina sekuentzia ezagunak erabiliz (Ensembl 75. bertsioa), 80 aminoazidoz osatutako sekuentziak sortu ziren mutatutako aminoazidoaren inguruan. Ondoren, erreferentziako proteoman agertzen ez diren sekuentziak bakarrik mantendu genituen sortutako datu-basean. Osotora, 92412 proteinetan aurkitzen diren 1525055 mutatutako aminoazidok osatzen dute gure datu-basea. Informazio hau, 19925 geneen sekuentzietatik ateratakoa zen. Gene bakoitzeko, batezbeste 77 mutatutako aminoazido identifikatu genituen. Azaleko melanoman aurkitu genituen SAP gehien (222665 mutatutako aminoazido) eta ubeako melanoman gutxien (2738 mutatutako aminoazido). Azkenik, datu hauek erabiliz, FASTA formatuko datu-basea sortu genuen, proteomikako bilaketak egiteko. Bilaketak egin ostean, guk egindako R lengoaiako *script*-ak erabiliz, beharrezko emaitzak lortu ziren.

3.2 *Shotgun* esperimentu publikoak

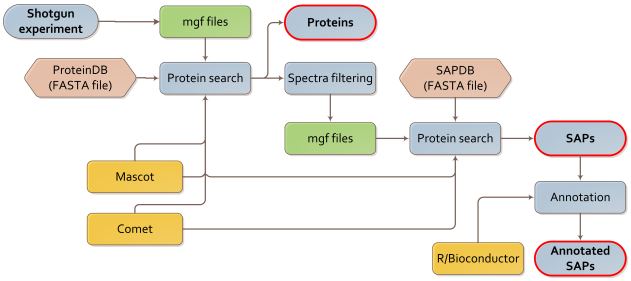
PRIDE datu-basean eskuragarri dauden lau giza minbiziren zelulen egindako esperimentuak analizatu genituen. Esperimentu hauek, Espainiako HPP ekimeneko lan taldeek ProteomeXchange biltegira PXD000039, PXD000442, PXD000443 eta PXD000449 kodeekin sartuak ziren. Minbizi mota desberdineko zelula bakoitzeko bi erreplika aukeratu genituen: Jurkat (T Linfoblasto motako giza zelulak), CCD18 (koloneko fibroblasto motako giza zelulak), MCF7 (bularreko adenokartzinoma motako giza zelulak) eta Huh7 (gibeleko kartzinoma motako giza zelulak). Adierazitako analisiak egiteko Mascot bilatzaileak erabiltzen dituen fitxategiak (*Mascot generic files*, MGF) jaitsi genituen. MGF fitxategi hauek, masa-espektrometria espektroak dituzte testu formatuan. MGF fitxategiak ProteinDB datu-basean bilatu genituen Mascot zerbitzaile bilatzailea (2.3 bertsioa, Matrix Science, London, UK) erabiliz. Bilaketa egiteko, ohiko parametroak erabili ziren. Protein mailako identifikazioek, espektro mailan positibo-faltsu tasa (*Peptide-spectrum match false discovery rate*, PSM FDR) $<\%$ 1 eta protein mailan positibo-faltsu tasa (proteina *False discovery rate* FDR) $<\%$ 1 irizpideak jarraitzen zituzten, C-HPP ekimenak gomendatzen duen moduan.

3.3 *Shotgun* datuen analisia

Lagin bakoitzarekin, 2. irudian ikusi daitekeen moduan, proteomikako 2 analisi egin ziren: lehenengoan, proteina ezagunen datu-base bat erabiliz, gure kasuan Ensembl proteina datu-basea (ProteinDB), eta bigarrenean, guk sortutako mutatutako peptido datu-basea erabiliz (SAPDB).

Proteinen inferentzia PAnalyzer algoritmoa erabiliz egin ziren eta eztabaidagarriak ziren proteinak alde batera baztertu ziren (Prieto *et al.*, 2012).1 taulan, zelula mota bakoitzaren proteomikako emaitzak erakusten ditugu.

2 Irudia: *Shotgun* esperimentu publikoak erabilita, mutatutako peptidoak detektatzeko bioinformatikako lan-fluxua. Analisi atazak (urdiñez), fitxategi formatuak (berdez), datu-baseak (gorriz) eta bioinformatika erremintak (laranjaz) ikus daitezke irudian. Gorriz markatutako ertzek, lan-fluxuaren irteera adierazten dute.

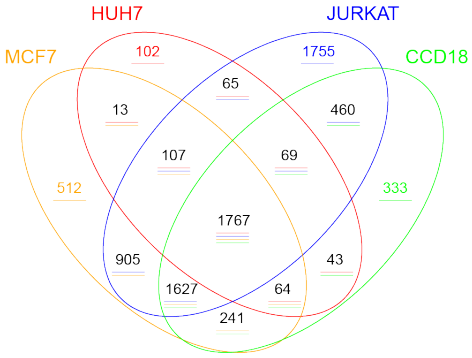


1 Taula: ProteinDB datu-basea erabiliz identifikatutako peptido, proteina eta gene kopurua (positibo-faltsu tasa (PSM FDR) <% 1, proteina mailan positibo-faltsu tasa (proteina FDR) <% 1).

	Peptidoak	Proteinak	Geneak
HUH7-1	1367	3596	1017
HUH7-2	3303	6159	1890
MCF7-1	28573	14000	4785
MCF7-2	12175	11324	3669
JURKAT-1	39433	19094	6409
JURKAT-2	36026	17888	6110
CCD18-1	17981	11581	3837
CCD18-2	17152	11005	3771

Emaitzen balioztatzea sinplifikatzeko, Ensembl protein kodeak gene kodera bihurtu genituen. Guztira 8063 gene detektatu ziren, % 21.92a 4 zelula motetan eta % 33.51a zelula mota bakoitzen berariazkoak izanda (3 irudia).

3 Irudia: ProteinDB datu-basea eta Mascot bilatzailea erabiliz identifikatutako geneak.



Ezagunak diren proteinak identifikatu ostean, analisi honetan esleitutako espektroak hasierako MGF fitxategietatik baztertu genituen. Horrela, mutatutako peptidoen bilaketa, proteina ezagunak ez direnen artean egin zen. Lagin iragaziak mutatutako peptidoen datu-basea (SAPDB) erabiliz analizatu genituen. Datu-base honek 1525055 mutatutako peptido dauzka. Honetaz gain, 3 datu-base berri sortu ziren:

- LIHC datu-basea (LIHCDB): Gibeleko kartzinomari dagokion 67694 mutatutako peptidoz osatua, Huh7 zelulak ikertzeko.
- COAD datu-basea (COADDB): Kolon-ondesteko adenokartzinomari dagokion 89771 mutatutako peptidoz osatua, CCD18 zelulak ikertzeko.
- BRCA datu-basea (BRCADB): Bularreko kartzinomari dagokion 72249 mutatutako peptidoz osatua, MCF7 zelulak ikertzeko.

2 Taula: SAPDB datu-baseak erabiliz identifikatutako peptido, protein eta gene kopurua (espektro mailan positibo-faltsuen tasa (PSM FDR) <% 1). SAPDB datu-base orokorra da eta LIHCDB, COADDB eta BRCADB zelula mota bakoitzaren berariazkoak.

	PSMak	Mutatutako peptidoak (SAP)	Mutatutako proteinak	Mutatutako geneak	datu-basea
HUH7-1	5	5	9	4	LIHCDB
HUH7-2	6	5	9	5	LIHCDB
MCF7-1	115	97	218	68	BRCADB
MCF7-2	22	8	24	5	BRCADB
CCD18-1	65	19	64	10	COADDB
CCD18-2	82	33	78	21	COADDB
HUH7-1	134	50	107	35	
HUH7-2	153	105	225	70	
MCF7-1	1972	1627	1093	343	
MCF7-2	27	22	63	12	SAPDB
CCD18-1	907	316	581	179	
CCD18-2	1238	466	841	256	
JURKAT-1	1111	690	1150	332	
JURKAT-2	554	364	795	239	

Jurkat zelulak ikertzeko, SAPDB datu-basea orokorra bakarrik erabili zen, mutazio informazio faltagatik, ez zegoelako berariazko datu-baserik egiteko aukerarik.

Mascot zerbitzariarekin bilaketak egin ziren, ohiko parametroak erabiliz, baina kasu honetan espektro mailan positibo-faltsu tasa (PSM FDR) <% 1 bakarrik erabili zen mutatutako peptidoak identifikatzeko. 2 taulan ikus daitezke lortutako emaitzak.

Iragazi ostean espektro kopurua handia izan arren, identifikatutako emaitzak, mutatutako peptido, proteina eta geneei dagokienez, eskasak izan ziren. Emaitz hauek espero diren mutatutako peptidoen ugaritasun bajuarekin eta masa-espektrometria esperimentuen zorizkotasunarekin bat egiten dute. SAPDB eta zelula mota bakoitzarekiko datu baseekin, emaitzak alderatu genituen, 4. irudian ikus daitekeen moduan. Bestalde, SAPDB orokorra erabiliz identifikazio gehiago lortu genituen zelula mota bakoitzaren berariazko datu-baseekin alderatuz. Gainera, aipagarria iruditzen zaigu mutatutako gene kopuruaren apaltasuna, mutatutako peptidoekin alderatuta, baina gertaera hau gene bakoitzeko mutazio asko daudelako izan daiteke.

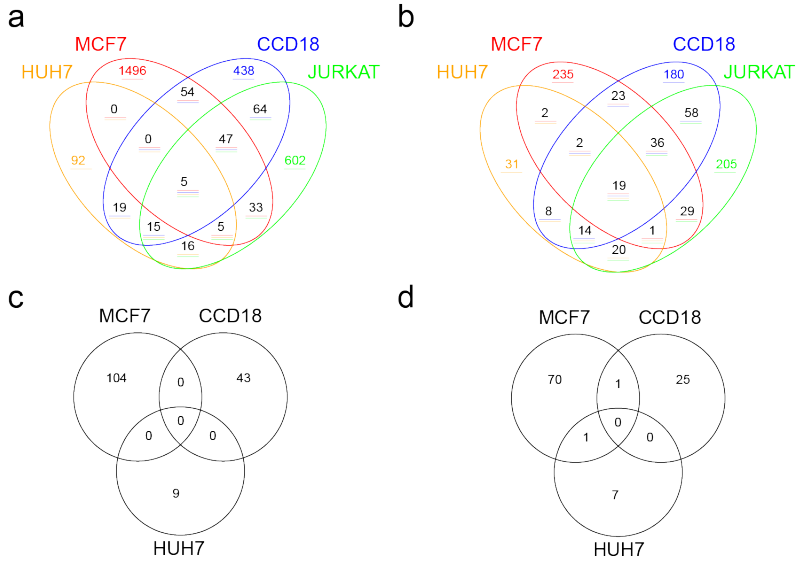
5. irudian ikus daitekeen moduan SAPDB orokorra eta zelula mota bakoitzaren berariazko datu-baseen arteko antzekotasuna nabaria da. Espero zen bezala, zelula mota bakoitzeko, identifikatutako mutatutako peptido gehienak bi datu-baseak erabilitakoan agertzen ziren. Halere, mutatutako peptido batzuk beste zelula motetan agertzen ziren eta ez ziren datu-base analisi bietan agertzen. Honen arrazoia datu-basearen tamainak positibo-faltsuen tasan duen eraginagatik edo hainbat minbizi motek mutazio berak konpartitzen dituztelako izan daiteke.

Laburbilduz, lan honetan 2916 mutatutako peptido (877 mutatutako gene) identifikatu ditugu 4 minbizia motatako zeluletan. Mutatutako gene multzoak ikertutako minbizia mota bakoitzean duen eragina aztertu genuen. Hortarako, mutatutako geneen analisi funtzionala egin genuen Ingenuity erreminta erabiliz. Mutatutako gene guztietatik, gehienak minbiziari dagokion kategoriarekin erlazionatuta zeuden. Gainera, aberastuta dauden gaixotasunen artean, "bularreko edo kolon minbizia", "kolon minbizia", "gibeleko minbizia", "hepatokartzinoma"eta "neoplasia hematologikoa"aurkitu genituen. Gure ustez emaitza hau oso garrantzitsua da, identifikatutako mutazioak ikertutako minbizi motekin erlazionatuta daudela egiaztatzen duelako.

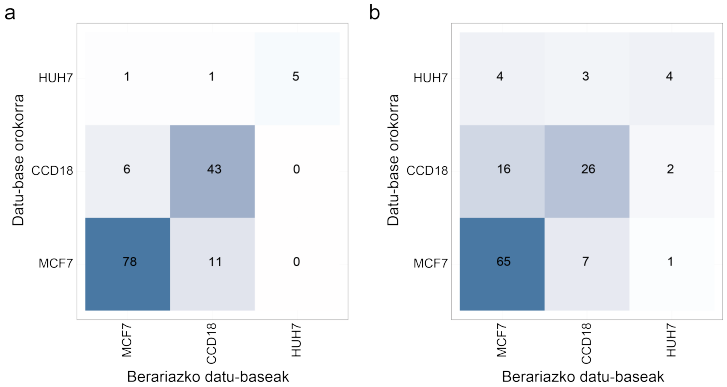
4 Ondorioak

Artikulu honetan, minbizi laginetan mutatutako peptidoak detektatzeko bioinformatikako lan-fluxua deskribatzen dugu. TCGA proiektuko mutazioez baliatuz, datu-base proteogenomikoak sortu genituen, 31 minbizi mota desberdinei dagokien datuak erabiliz. Datu-base hauek, 4 minbizi mota desberdinetako zelulak *shotgun* esperimentuen bilaketan erabili genituen, peptido mutatuak bilatzeko. Lortutako emaitzak funtzionalki ikertu ziren eta minbizian, eta batez ere guk ikertzen ditugun minbizi motetan, duten eragina egiaztatu genuen. Ikerketa honek, minbizi laginetan mutatutako peptidoak detektatzeko proteogenomikaren baliotasun garrantzitsua baieztatzen du.

4 Irudia: SAPDB orokorra (SAPDB) eta berariazko DBak erabiliz ikusitako mutatutako peptidoak eta geneak laburbiltzen dituen diagramak. a) Mutatutako peptido kopurua SAPDB datu-basea erabiliz. b) Mutatutako gen kopurua SAPDB datu-basea erabiliz. c) Mutatutako peptido kopurua zelula-lerro bakoitzeko berariazko datu-baseak erabiliz. d) Mutatutako gene kopurua zelula-lerro bakoitzeko berariazko datu-baseak erabiliz.



5 Irudia: Mutazioen arteko antzekotasunak erakusten dituen *heatmap*a. a) Zelula-lerroen artean konpartitzen dituzten mutatutako peptidoak. b) Zelula-lerroen artean konpartitzen dituzten mutatutako geneak.



5 Etorkizunerako planteatzen den norabidea

Minbizia ikertzeko genomikako sekuentziazioek minbizi genomen konplexutasuna eta heterogeneitatea erakutsi digute. Teknika honen bitartez, existitzen diren mutazio kopurua handia dela ikusi dugu azken urteotan. Honekin batera, proteinak zelula-funtzioan garrantzitsuak dira eta mutatuko geneak sortzen dituzten proteinek euren funtzioa aldatuta izaten dute. Hori dela eta, mutatutako proteinek tumoreen hasiera, progresioa eta tratamenduaren erantzuna gidatzen dute. Beraz, proteomikaren bidez, genomikan ikusten diren mutazioek duten eragina baieztatu daiteke. Alabaina, peptidoen bilaketak egiteko erabiltzen diren proteomikako datu-baseek, ezagunak diren proteinak barneratzen dituzte eta hau honela izanik, mutatutako peptidoak bilatzeko ez dute balio handirik. Horregatik, mutatutako geneetatik sortutako proteina datu-baseak sortzea ezin bestekoa da mutatutako peptidoak bilatzeko. Genomika bakarrik erabiliz, proteinan izango duen eragina aurresan dezakegu baina, ostera, proteogenomikaren bidez, mutatutako peptidoak ikus ditzakegu ere.

Aurrerantzean, artikulu honetan proposatutako lan-fluxua zehaztasunezko medikuntzan erabiltzea da gure

nahia. Horretarako, pazienteen datuak jasoko ditugu, hauen DNA genomikoa sekuentziatuko dugu eta pazienteek dituzten gene mutazioak kontuan hartuta, hauek sortzen dituzten peptido mutatuak masa-espektrometria bidez ikertuko ditugu. Honek, gure lan-fluxuari balioa emango dio zalantzarik gabe.

Erreferentziak

Adzhubei, Ivan A, Steffen Schmidt, Leonid Peshkin, Vasily E Ramensky, Anna Gerasimova, Peer Bork, Alexey S Kondrashov, eta Shamil R Sunyaev. 2010. A method and server for predicting damaging missense mutations. *Nature methods* 7.248–249.

Alfaro, Javier A, Ankit Sinha, Thomas Kislinger, eta Paul C Boutros. 2014. Onco-proteogenomics: cancer proteomics joins forces with genomics. *Nature methods* 11.1107–1113.

Ansong, Charles, Samuel O Purvine, JoshuaÑ Adkins, Mary S Lipton, eta Richard D Smith. 2008. Proteogenomics: needs and roles to be filled by proteomics in genome annotation. *Briefings in functional genomics & proteomics* 7.50–62.

Chin, Lynda, William C Hahn, Gad Getz, eta Matthew Meyerson. 2011. Making sense of cancer genomic data. *Genes & development* 25.534–555.

Faulkner, Sam, Matthew D Dun, eta Hubert Hondermarck. 2015. Proteogenomics: emergence and promise. *Cellular and Molecular Life Sciences* 72.953–957.

Krasnov, George Sergeevich, Alexey Alexandrovich Dmitriev, Anna Viktorovna Kudryavtseva, Alexander Valerievich Shargunov, Dmitry Sergeevich Karpov, Leonid Andreevich Uroshlev, Natalya Vladimirovna Melnikova, Vladimir Mikhailovich Blinov, Ekaterina Vladimirovna Poverennaya, Alexander Ivanovich Archakov, eta others. 2015. Ppline: An automated pipeline for snp, sap, and splice variant detection in the context of proteogenomics. *Journal of proteome research* 14.3729–3737.

Kumar, Prateek, Steven Henikoff, eta Pauline C Ng. 2009. Predicting the effects of coding non-synonymous variants on protein function using the sift algorithm. *Nature protocols* 4.1073–1081.

Lander, Eric S, Lauren M Linton, Bruce Birren, Chad Nusbaum, Michael C Zody, Jennifer Baldwin, Keri Devon, Ken Dewar, Michael Doyle, William FitzHugh, eta others. 2001. Initial sequencing and analysis of the human genome. *Nature* 409.

Legrain, Pierre, Ruedi Aebersold, Alexander Archakov, Amos Bairoch, Kumar Bala, Laura Beretta, John Bergeron, Christoph H Borchers, Garry L Corthals, Catherine E Costello, eta others. 2011. The human proteome project: current state and future direction. *Molecular & cellular proteomics* 10.M111–009993.

McLaren, William, Bethan Pritchard, Daniel Rios, Yuan Chen, Paul Flicek, eta Fiona Cunningham. 2010. Deriving the consequences of genomic variants with the ensembl api and snp effect predictor. *Bioinformatics* 26.2069–2070.

Meyerson, Matthew, Stacey Gabriel, eta Gad Getz. 2010. Advances in understanding cancer genomes through second-generation sequencing. *Nature Reviews Genetics* 11.685–696.

Nagaraj, Shivashankar H, Nicola Waddell, Anil K Madugundu, Scott Wood, Alun Jones, Ramya A Mandyam, Katia Nones, John V Pearson, eta Sean M Grimmond. 2015. Pgtools: a software suite for proteogenomic data analysis and visualization. *Journal of proteome research* 14.2255–2266.

Nesvizhskii, Alexey I. 2014. Proteogenomics: concepts, applications and computational strategies. *Nature methods* 11.1114–1125.

Paik, Young-Ki, eta William S Hancock. 2012. Uniting encode with genome-wide proteomics. *Nature biotechnology* 30.1065.

Prieto, Gorka, Kerman Aloria, Nerea Osinalde, Asier Fullaondo, Jesus M. Arizmendi, eta Rune Matthiesen. 2012. Panalyzer: A software tool for protein inference in shotgun proteomics. *BMC Bioinformatics* 13.288.

Segura, Víctor, Juan Alberto Medina-Aunon, Maria I Mora, Salvador Martínez-Bartolomé, Joaquín Abian, Kerman Aloria, Oreto Antúnez, Jesús M Arizmendi, Mikel Azkargorta, Silvia Barceló-Batllori, eta others. 2013. Surfing transcriptomic landscapes. a step beyond the annotation of chromosome 16 proteome. *Journal of proteome research* 13.158–172.

TABAS-MADRID, DANIEL, JOAO ALVES-CRUZEIRO, VICTOR SEGURA, ELIZABETH GURUCEAGA, VITAL VIALAS, GORKA PRIETO, CARLOS GARCÍA, FERNANDO J CORRALES, JUAN PABLO ALBAR, eta ALBERTO PASCUAL-MONTANO. 2015. Proteogenomics dashboard for the human proteome project. *Journal of proteome research* 14.3738–3749.

TOMCZAK, KATARZYNA, PATRYCJA CZERWIŃSKA, eta MACIEJ WIZNEROWICZ. 2015. The cancer genome atlas (TCGA): An immeasurable source of knowledge. *Contemporary Oncology (Pozn)* 19.A68–A77.

TRAPNELL, COLE, DAVID G HENDRICKSON, MARTIN SAUVAGEAU, LOYAL GOFF, JOHN L RINN, eta LIOR PACHTER. 2013. Differential analysis of gene regulation at transcript resolution with rna-seq. *Nature biotechnology* 31.46–53.

VENTER, J CRAIG, MARK D ADAMS, EUGENE W MYERS, PETER W LI, RICHARD J MURAL, GRANGER G SUTTON, HAMILTON O SMITH, MARK YANDELL, CHERYL A EVANS, ROBERT A HOLT, eta OTHERS. 2001. The sequence of the human genome. *science* 291.1304–1351.

WANG, XIAOJING, eta BING ZHANG. 2013. customprodb: an r package to generate customized protein databases from rna-seq data for proteomics search. *Bioinformatics* p. btt543.

WOO, SUNGHEE, SEONG WON CHA, SEUNGJIN NA, CLARK GUEST, TAO LIU, RICHARD D SMITH, KARIN D RODLAND, SAMUEL PAYNE, eta VINEET BAFNA. 2014. Proteogenomic strategies for identification of aberrant cancer peptides using large-scale next-generation sequencing data. *Proteomics* 14.2719–2730.

ZHANG, BING, JING WANG, XIAOJING WANG, JING ZHU, QI LIU, ZHIAO SHI, MATTHEW C CHAMBERS, LISA J ZIMMERMAN, KENT F SHADDOX, SANGTAE KIM, eta OTHERS. 2014. Proteogenomic characterization of human colon and rectal cancer. *Nature* 513.382–387.

6 Eskerrak eta oharrak

Ikerketa hau hurrengo erakundeek lagundu dute:

- Nafarroako Gobernuko Osasun Saileko 33/2015 proiektua (V Segura)
- Espainiako Zientzia eta Berrikuntza Ministerioko DPI2015-68982-R diru-laguntza (V Segura)
- PRBB eta Carlos III Nazioarteko Osasun Institutua, PRBB-ISCIII'
- Espainiako Zientzia eta Berrikuntza Ministerioko SAF2014-5478-R diru-laguntza (F Corrales)
- ISCIII-RETIC RD06/0020 (F Corrales)

Miren Josu Omaetxebarria (UPV-EHU) eta Xabier Aguirreri (CIMA-UNAV) euskera zuzendu eta gure artikulua irakurtzeko denbora hartzeagatik. Mila esker!