Machine Learning Draft

D J Spaanderman *Broad Institute of MIT and Harvard*

This document provides an draft of the machine learning part of my Masters Major Internship, I will start by introducing the subject including some basics about machine learning, what my lab already has achieved and how I will go fourth in achieving our goal; predicting tumor growth media using genomics.

Keywords: Cancer cell line, Predicting growth media, Machine learning, AI

Introduction

The cancer cell line factory (CCLF) an initiative of the *Broad Institute* focussess on the production of cancer models (Boehm and Golub, 2015). These cancer models can be assessed for various aims. First, comprehensively characterizing the genomes of cancer models can infer information such as cancer dependencies, which is currently lacking in rarer types of cancers. Secondly, any lab can have access to these new cancer models enabeling faster and more tuned research. Thirdly, better representing cancer models to patients can improve treatment selection. Lastly, the scale at which these models are created will give us new insights into standardizing research protocals, which will be mainly the focus of this draft. Additionally, several papers have been recently published which use models derived from CCLF (Hong et al., 2016, Ben-David et al. (2017), Viswanathan et al. (2017), Joung et al. (2017)).

Due to the nature of different tumor and tissue types, effectively creating these cancer models harnasses many technical difficulties, one of which being the selection of growth media. Therefore, A major bottleneck in current procedures is the required use of media panels as large as 64 different types in order to effectively growth cancer models. By means of trial and error we have narrowed down the candidate growth media's for specific tumor and tissue types which we have grown multiple times over the years. However, rarer and less frequent grown cancer models still requires extensive media panels. In this draft I want to explain our pipeline, several machine learning approaches and hypothesis how we can go about using the genomic information gathered to train machine learning models to predict the best possible growth media.

Pipeline

CCLF receives patient samples from clinicians. In our pipeline we currently have a total of 1559 unique patient ID, ranging from the year 2001 till 2019 (based on the cohort dashboard from Tableau). In Figure 1, I have highlighted the top 10 most occuring cancer types in our pipeline out of a total of 223 cancer types. Total frequency for each tumor type in our database can be found in the supplementary table 1. Note, that due to some inconsistency in naming, multiple similar tumor types consists in the database. Additionally, I checked the current use of media types in the top 10 most occuring tumors (Figure 2) and present all the media types including brief explanation in supplementary table 2. Aside from the tumor type, the patient sample can have various biological origins (i.e. tumor site) and can originate from primary or metastasis. Furthermore, technical differences can also exists between samples, such as the way the sample is recieved (i.e. fresh tissue, cryopreserved tissue, frozen tissue etc.), types of biosp used (pleural

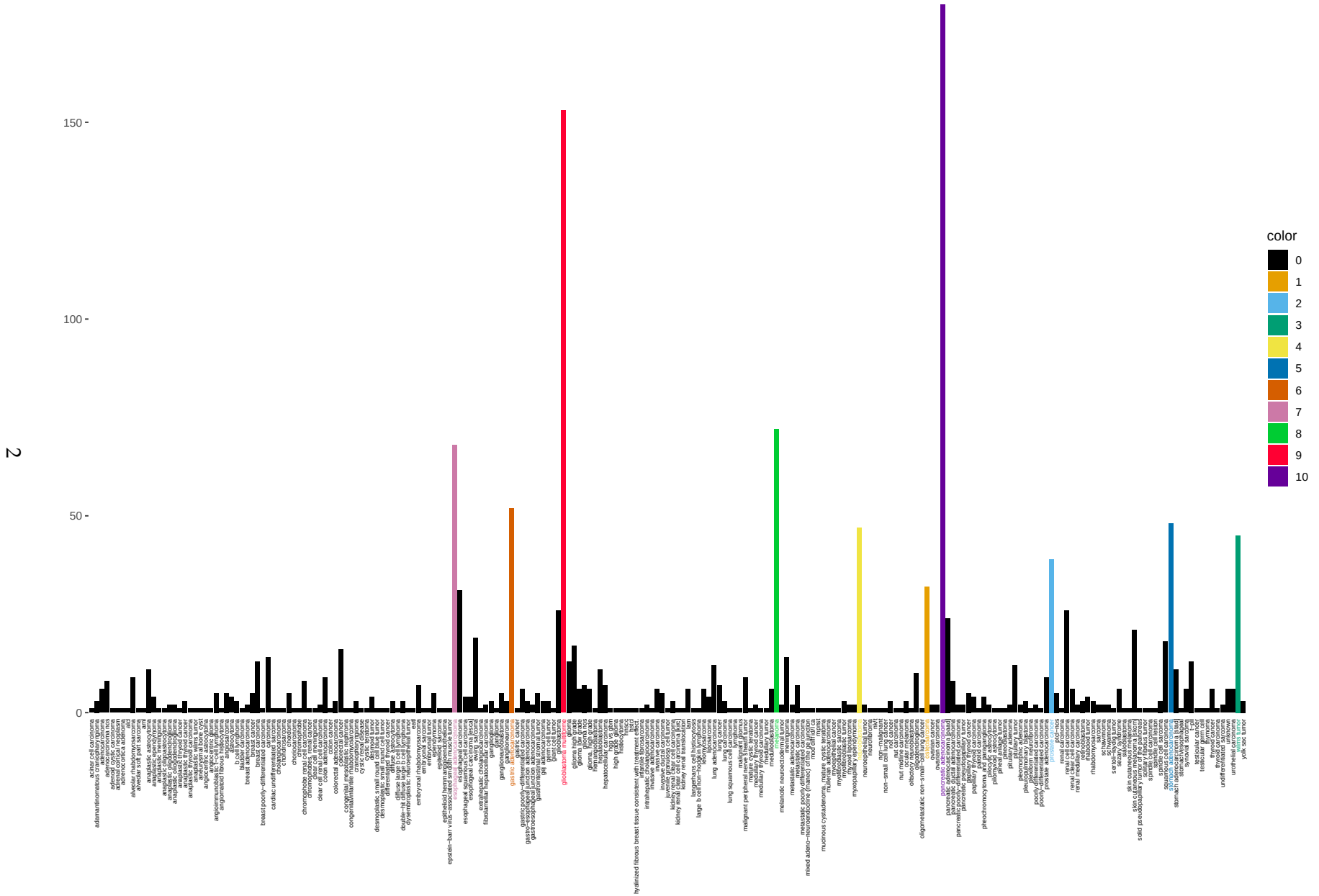

Figure 1: Number of media's tried for top 10 tumor types

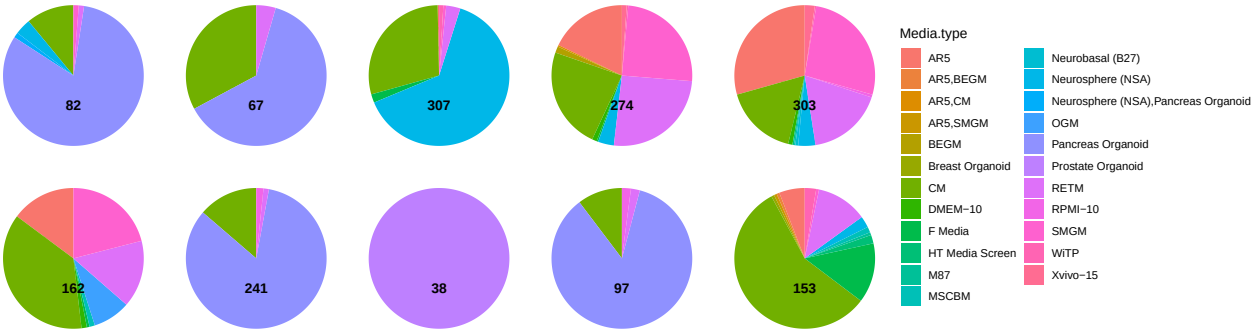

Figure 2: Number of unique patient IDs for each tumor type in the CCLF pipeline

effusion, needle, ascites, etc.) and from which clinician. Together these differences create noise between samples. For instance, it might be that tumor X growth's fine with media Y, however fails due to fact that the starting material was frozen down. It is important to take notes of these artifacts in the data as it can influence how further down the road, our model interpert the data, or might even learn features which are created due to these technical differences. Noteable, growing tumor models is a laborious and difficult task, so even when all conditions are right, the experiment might fail. Additionally, CCLF growths multiple types of cancer models, such as 2D (tradtional cell lines), 3D (organoid) and neurosphere (free floating cluster of cells), which should also be included for distinquishing tumors as these highly influence media condition.

Before we initialize the patient sample, first the DNA is sequenced using a large contig panel (a set of predefined regions). It is important to note that, during the time of CCLF this panel has been extended to include more regions of the genome, introducing either missing data or abundant data respectively for the old and new data if both datasets are aggegrated. Raw sequencing reads are mapped to reference genome GRCh37 and using GATK V4 including mutect1/2, copy number variations (CNVs), small nucleotide ploymorphisms (SNPs), insertions and deletions (Indels) are infered. Additionally, germline events are filtered by comparing matching mutations event in blood sample if applicable and filtering previously identified germline events in our pipeline.

When cancer models are finalized and have been through nurcery, genome and RNA is sequenced and compared to infer resemblance to its original patient sample. Currently, we neither have whole genome sequencing or RNA sequencing data for initial patient sample. Therefore, we could explore the possibility of using cancer model data as a manner of input data for our machine learning model as it is in theory more compleet. Note, that RNA data is timepoint dep endent and that culturing might have influenced the transcriptomics.

Arguable the most important descision for our media prediction model is the chosen input data. Some general notes to take into account when selecting input data, the higher dimensional the data, the more complex the model requires to be to infer features in this data. Contrary, less preprocessing such as protein pathway analysis or mutation calling has many advantages as an end-to-end model removes the introduction of additional technical noise, reduces analysis time and has better standarization capabilities. In Figure 3, I have depicted a simplified version of our current pipeline and possible input data for our model. This could also include biological data such as tumor type, site and metastasis or primary, however ideally this information is not included as it is often unknown or highly variable. Additionally, multiple data entries could be assessed as input data, however will require a more complex model. For all input possibility I have highlighted challenges and strengths.

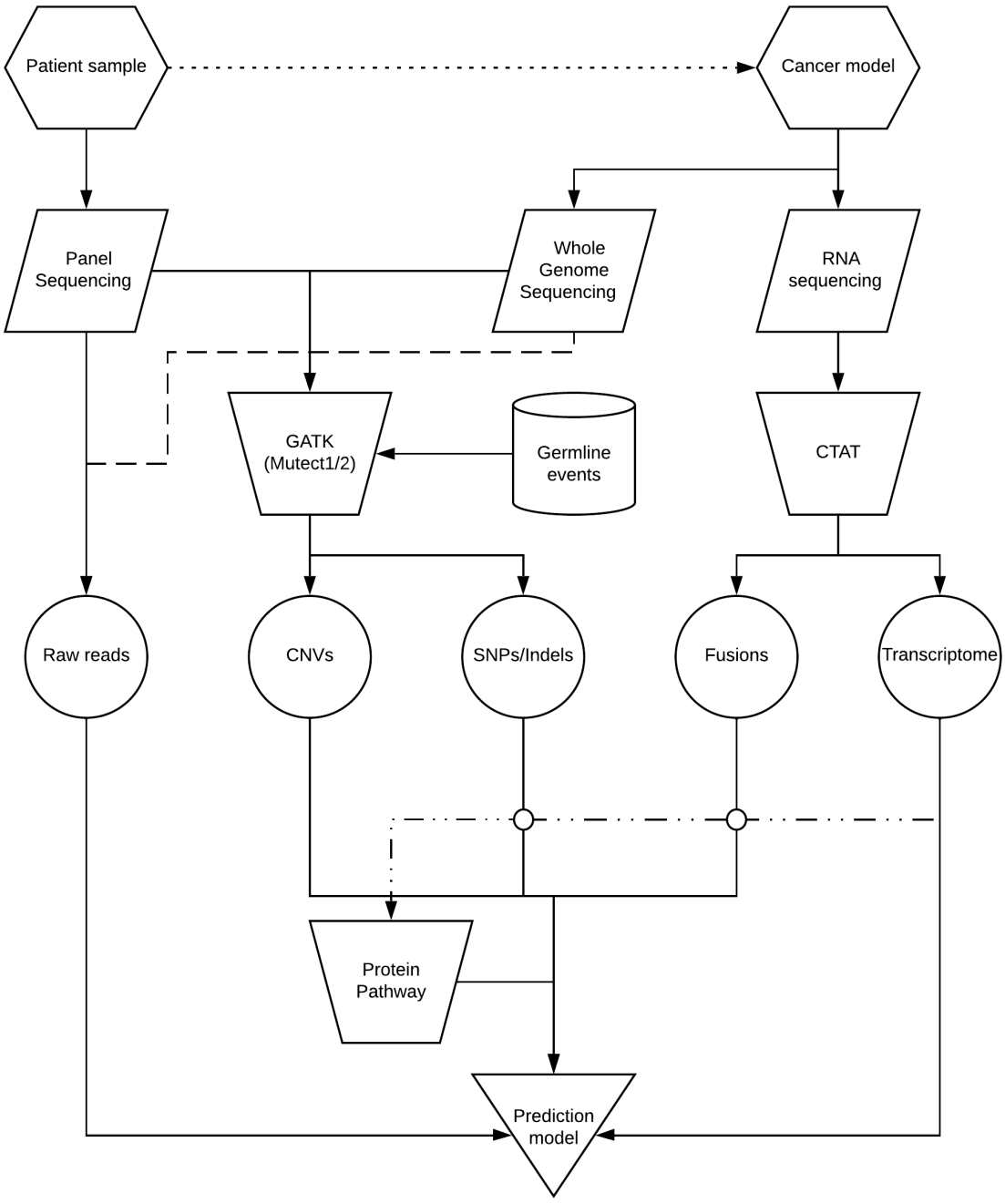

Figure 3: Pipeline of CCLF and possible data type entries for our machine learning application

- **Raw sequencing reads** would create an end-to-end model without any preprocessing however is far to high dimensional and sparce to use as an input data. Implementation of raw sequencing reads have to my knowlegde not been done in any genomics machine learning model, aside from raw nanopores sequencing signal. On the other hand, mapped sequencing reads have been used for more elaborate tasks such as SNP and Indel calling, but even here this would require a far to complex model (i.e. raw or mapped reads not an option as input data).
- **CNV** and/or **SNPs** and/or **Indels** (and/or **Fusions**): with the exception of CNVs, these data entries are very sparce. For example, it could well be that 40 similar tumor types only share 5 somatic variants/Fusions, while all the other data is unique for each sample. An higher overview, such as mutated gene map or protein pathway changes could reduce this limitation. However, would require a number of assumptions, such as identifying important fusion/variant.
- **RNA** have been reported in other relatable models, see the part literature models. Also, RNA data could better represent metabolic events in the data. However, RNA data is gathered from cancer models, which might have different transcriptomics than their patient sample counterpart.

Machine learning

Machine learning algorithms can be roughly divided between supervised and unsupervised methods. In supervised machine learning, labeled data is used to predict the classification or regression of data points. Examples of 'conventional' supervised machine learning algorithms are linear and logistic regression, the random forest classifier and support vector machines (SVM). On the other hand, in unsupervised machine learning, patterns in data are learned without relying on predefined labels. Two examples of unsupervised machine learning are clustering and principal component analysis. Additionally, semi-supervised models, which uses a combinaton of labeled and unlabeled data, in order to limit the required learning data can also be assessed. In our case, we have gathered a fair amount of labeled data in the form of a genomic profile for a patient sample with both succesful and unsuccesful experiments based on culture conditions (i.e. Media). Therefore either a supervised or an semi-supervised model should be assess to predict media condition. A major bottleneck for our model is the high amount of technical noise which consists in these experiments. *"A machine learning model is only as good as the data"* as described by IBM.

Aside from the input data, model selection is also important. Note that arguable the best way to approach this problem is to trial and error, initialy starting with simpeler models such as a random forest classifier, as most likely data is not linear seperable, ruling out linear regression models and support vector machines. Moving forward to more complex models such as feed-forward networks, convolutional neural networks and recurrent neural networks. These methods are well described in my literature study about Exploring Genomics using Deep Learning (confident but also present in my GitHub repository, please be mindfull when sharing). This review also describes shortly the bascis of deep learning such as back-propogation, activation functions, loss functions, hyperparamaterization, which I won't go into further here as it is an extensive subject. However, on the technical aspect there are several tools we could assess to create the best model for our needs. scikit-learn which is an extensive python library can be used for various machine learning algorithms such as random forest, dimensional reduction techniques and preprocessing (feature extraction and normalization). TensorFlow and Keras respectively low and higher level neural network API's to create deep learning models. Additionally, Hyperas and Talos are hyperparameter optimization algorithms and DeepReplay assist learning visualization.

Aside from these supervised learning models, we could also invest time in dimensional reduction techniques to infer relations in mutation data as a method of pre-processing such as principal component analysis (PCA), autoencoder or generative adversarial network (GAN).

Literature models

In my literature study, I have described many deep learning approaches in genomics. However, few models are described to achieve similar goals as predicting media conditions from genomic data. Nevertheless, Lyu and Haque (2018), et al. describes the use of gene expression data in order to classify tumor types. In order to achieve this goal, RNA data is log2 transformed and compared to Pan-Cancer Atlas in order to reduce the gene panel. Next these gene panels are embedded into a 2D image (102x102), which encodes for gene expression of 10404 important genes. These images are than analyzed using a convolutional neural network. Some things to note with this method, embedding is arguable the most dominant factor in model effectiveness. This is due to the nature of CNNs, as they take spatial dependencies in data into account. Therefore shifting around gene position would greatly impact model accurarcy. The main message we can take from this paper is the method they used to encode their genomic data by selecting important genes from Pan-Cancer Atlas and encoding into a 2D heatmap.

A more recent paper, uses a deep learning approach to classify primary and metastatic cancer using passenger mutation patterns (Jiao et al., 2020). somatic mutation were preprocessed to extract several features. For each sample, the mutational-type feature was based on counting the number of single nucleotide changes, nucleotide changes plus their 5′ and/or 3′ flanking nucleotides. Next, these mutational-type feautres were normalized for the total number of SNVs in the sample. The mutational distribution features are the number of SNVs, small indels, structural variation (SV) breakpoints and CNV in 1-megabase bins across the genome, normalized to the corresponding mutational event across the genome. Additional features are the total number of each type of mutational event per genome, number of each type of mutational event per chromosome (normalized for chromosome size), sample purity and sample ploidy. These features were than used in an feed-forward network. Noteable, adding information on driver mutations reduced model accuracy. This paper presents another method of encoding genomic data for our machine learning model.

Conclusion

In this paragraph I will discuss the best steps to follow in order to create a model which will be able to infer media condition from genomic data. Currently, the main bottleneck is selecting the type of input data and how to encode this information. Both models reported in the previous segment show promosing methods of encoding genomic data for deep learning approaches. In my opinion, RNA data seems more intuitive as it closer resemblence metabolic features, which are most important for media selection. However I am unsure if the transcriptomics data we have is applicable for predicting media condition as it originates from cancer models instead of patient sample. On the other hand, we have more abundant mutation data from patient sample. Preprocessing would be a more elaborate task when assessing mutation data. Similar to the paper described in the previous segment, we could encode mutation data in features. In contrast, we could encode mutation data as higher level structures such as gene level (i.e. gene x has y mutation data) or even in an protein pathway (i.e. pathway x has gene y with z mutation data). This would reduce the effect of sparseness in mutation data, however would require a fair amount of assumptions, such as which pathway/gene to feed the network, if and how to include functional

mutation annotation. In order to create a working machine learning model several steps have to be conducted:

1. Aggegrating compleet datasets, ideally a Terra workspace in which data is cleaned and displayed as Patient ID X - Biological/technical information (i.e. primary-biopsy-type etc.) - Raw/mapped reads - CNV - SNV/indels - Media types (csv file with all media types + succes rates for each type) - linked cancer model id - WGS (incl CNV/SNV) - RNA
2. Encoding either mutation information or transcriptomics based on feedback from Remi/Moony
3. Segregation of training, test and validation dataset
4. Building, training and validating machine learning algorithm
5. Blackbox features extraction. This is something very promising if the algorithm works. It would possibly give insight into why a certian mutational/transcriptomics landscape prefers a set media type.

Something to note is that ideally I would like to access all the created cancer models, by removing any predefined information such as tumor type and site. However, it also might be worthwhile to reduce noise by removing tumors types which are only few in our pipeline (<5).

Supplementary

Number of times reported	Tumor type
1	acinar cell carcinoma
3	adamantinomatous craniopharyngioma
6	adenocarcinoma
8	adenocarcinoma nos
1	adenoid cystic carcinoma
1	adrenal cortical neoplasm
1	adrenocortica adenoma
1	alcl
9	alveolar rhabdomyosarcoma
1	alveolar soft part sarcoma
1	aml
11	anaplastic astrocytoma
4	anaplastic ependymoma
1	anaplastic meningioma
1	anaplastic oligoastrocytoma
2	anaplastic oligodendroglioma
2	anaplastic oliogodendroglioma
1	anaplastic thryoid cancer
3	anaplastic thyroid cancer
1	anaplastic thyroid carcinoma
1	anaplastic wilms tumor
1	aneurysmal bone cyst
1	angiocentric astrocytoma
1	angiocentric glioma
5	angioimmunoblastic t-cell lymphoma
1	angiomatoid fibrous histiocytoma
5	angiosarcoma
4	astrocytoma
3	b cell lymphoma
1	bladder carcinoma
2	breast adenocarcinoma
5	breast cancer
13	breast carcinoma
1	breast poorly-differentiated carcinoma
14	carcinoma
1	cardiac undifferentiated sarcoma
1	cholangiocarcinoma
1	chondrosarcoma
5	chordoma
1	choriocarcinoma
1	chromophobe
8	chromophobe renal cell carcinoma
1	chromophobes, renal cancer

Number of times reported	Tumor type
1	clear cell meningioma
2	clear cell renal cell carcinoma
9	colon adenocarcinoma
1	colon cancer
3	colorectal adenocarcinoma
16	colorectal cancer
1	congenital mesoblastic nephroma
1	congenital/infantile rhabdomyosarcoma
3	craniopharyngioma
1	cystic renal disease
1	cystic teratoma
4	desmoid tumor
1	desmoplastic small round cell blue tumor
1	desmoplastic small round cell tumor
1	differentiated thyroid cancer
3	diffuse astrocytoma
1	diffuse large b cell lymphoma
3	double-hit diffuse large b cell lymphoma
1	dysembroplastic neuroepithelial tumor
1	eatl
7	embryonal rhabdomyosarcoma
1	embryonal sarcoma
1	embryonal tumor
5	ependymoma
1	epitelioid sarcoma
1	epithelioid hemmanggioendothelioma
1	epstein-barr virus-associated smooth muscle tumor
68	esophageal adenocarcinoma
31	esophageal carcinoma
4	esophageal squamous cell carcinoma
4	esophogeal carcinoma [esca]
19	ewing sarcoma
1	extrahepatic cholangiocarcinoma
2	fibrolamellar hepatocellular carcinoma
3	ganglioglioma
1	ganglioma
5	ganglioneuroblastoma
3	ganglioneuroma
52	gastric adenocarcinoma
1	gastric cancer
6	gastric poorly-differentiated carcinoma
3	gastro-esophageal junction adenocarcinoma
2	gastroesophageal junction adenocarcinoma
5	gastrointestinal stromal tumor
3	gej adenocarcinoma
3	germ cell tumor

Number of times reported	Tumor type
1	giant cell tumor
26	glioblastoma
153	glioblastoma multiforme
13	glioma
17	glioma high grade
6	glioma low grade
7	glioma nos
6	glioma, high grade
1	hemangioblastoma
11	hepatoblastoma
7	hepatocellular carcinoma
1	hgg vs gbm
1	high grade glioma
1	histiocytoma
1	hnscc
1	hstcl
1	hyalinized fibrous breast tissue consistent with treatment effect.
1	infantile fibrosarcoma
2	intrahepatic cholangiocarcinoma
1	invasive adenocarcinoma
6	invasive carcinoma
5	invasive ductal carcinoma
1	juvenile granulosa cell tumor
3	kidney renal clear cell carcinoma
1	kidney renal clear cell carcinoma [kirc]
1	kidney renal translocation
6	lam
1	langerhans cell histiocytosis
1	large b cell non-hodgkin lymphoma
6	leiomyosarcoma
4	liposarcoma
12	lung adenocarcinoma
7	lung cancer
3	lung carcinoma
1	lung squamous cell carcinoma
1	lymphoma
1	malignant glomus
9	malignant peripheral nerve sheath tumor
1	mature cystic teratoma
2	medullary thyroid cancer
1	medullary thyroid carcinoma
1	medullary tumor
6	medulloblastoma
72	melanoma
2	melanotic neuroectodermal tumor
14	meningioma

Number of times reported	Tumor type
2	metastatic adenocarcinoma
7	metastatic carcinoma
1	metastatic poorly differentiated carcinoma
1	mixed adeno-neuroendocrine (manec) of the ge junction
1	mod diff id/l ca
1	mpnst
1	mucinous cystadenoma, mature cystic teratoma
1	mullerian adenocarcinoma
1	myoepithelial cancer
1	myofibroblastic sarcoma
3	myofibroblastic tumor
2	myxoid liposarcoma
2	myxopapillary ependymoma
47	neuroblastoma
2	neuroepithelial tumor
1	neurofibroma
1	nk/t
1	non-malignant
1	non-small cell lung cancer
3	not cancer
1	nut carcinoma
1	nut midline carcinoma
3	ocular melanoma
1	olfactory neuroblastoma
10	oligodendroglioma
1	oligometastatic non-small-cell lung cancer
32	osteosarcoma
2	ovarian cancer
2	ovarian carcinoma
180	pancreatic adenocarcinoma
24	pancreatic adenocarcinoma [paad]
8	pancreatic ductal adenocarcinoma
2	pancreatic poorly-differentiated carcinoma
2	pancreatic pseudopapillary tumor
5	papillary thyroid cancer
4	papillary thyroid carcinoma
1	parotid carcinoma
4	pheochromocytoma and paraganglioma
2	pilocytic astrocytoma
1	pilomyxoid astrocytoma
1	pineal anlage tumor
1	pineoblastoma
2	pituitary adenoma
12	pituitary tumor
2	pleomorphic adenoma
3	pleuropulmonary blastoma

Number of times reported	Tumor type
1	plexiform neurofibroma
2	poorly differentiated carcinoma
1	poorly-differentiated carcinoma
9	prostate adenocarcinoma
39	prostate cancer
5	ptcl-nos
1	renal carcinoma
26	renal cell carcinoma
6	renal clear cell carcinoma
2	renal medullary carcinoma
3	retinoblastoma
4	rhabdoid tumor
3	rhabdomyosarcoma
2	sarcoma
2	schawnnoma
2	schwannoma
1	sertoli-leydig tumor
6	sezary syndrome
1	sialoblastoma
1	skin cutaneous melanoma
21	skin cutaneous melanoma [skcm]
1	solid pseudopapillary tumor of the pancreas
1	solitary fibrous tumor
1	spindle cell carcinoma
1	spindle cell lesion
3	spindle cell sarcoma
18	squamous cell carcinoma
48	stomach adenocarcinoma
11	stomach adenocarcinoma [stad]
1	stomach/esophageal
6	synovial sarcoma
13	t-pll
1	testicular cancer
1	testicular germ cell
1	thymoma
6	thyroid cancer
1	thyroid carcinoma
2	undifferentiated sarcoma
6	unknown
6	urothelial carcinoma
45	wilms tumor
3	yolk sac tumor

Table 1 : Frequency table of the occurence of tumor types in our pipeline, take note that some naming issues are currently present in our pipeline, such as stomach adenocarcinoma and stomach adenocarcinoma [stad], which should be considered the same, therefore cleaning of the data

should be conducted to aggegrate these results together, either by manually (cleaner) or artificially (faster) mapping these tumor types.

References

Ben-David, Uri, Gavin Ha, Yuen Yi Tseng, Noah F. Greenwald, Coyin Oh, Juliann Shih, James M. McFarland, Bang Wong, Jesse S. Boehm, Rameen Beroukhim and Todd R. Golub. 2017. "Patient-derived xenografts undergo mouse-specific tumor evolution." *Nature Genetics* 49(11):1567–1575.

Boehm, Jesse S. and Todd R. Golub. 2015. "An ecosystem of cancer cell line factories to support a cancer dependency map.".

Hong, Andrew L., Yuen Yi Tseng, Glenn S. Cowley, Oliver Jonas, Jaime H. Cheah, Bryan D. Kynnap, Mihir B. Doshi, Coyin Oh, Stephanie C. Meyer, Alanna J. Church, Shubhroz Gill, Craig M. Bielski, Paula Keskula, Alma Imamovic, Sara Howell, Gregory V. Kryukov, Paul A. Clemons, Aviad Tsherniak, Francisca Vazquez, Brian D. Crompton, Alykhan F. Shamji, Carlos Rodriguez-Galindo, Katherine A. Janeway, Charles W.M. Roberts, Kimberly Stegmaier, Paul Van Hummelen, Michael J. Cima, Robert S. Langer, Levi A. Garraway, Stuart L. Schreiber, David E. Root, William C. Hahn and Jesse S. Boehm. 2016. "Integrated genetic and pharmacologic interrogation of rare cancers." *Nature Communications* 7.

Jiao, Wei, Gurnit Atwal, Paz Polak, Rosa Karlic, Edwin Cuppen, Fatima Al-Shahrour, Gurnit Atwal, Peter J. Bailey, Andrew V. Biankin, Paul C. Boutros, Peter J. Campbell, David K. Chang, Susanna L. Cooke, Vikram Deshpande, Bishoy M. Faltas, William C. Faquin, Levi Garraway, Gad Getz, Sean M. Grimmond, Syed Haider, Katherine A. Hoadley, Wei Jiao, Vera B. Kaiser, Rosa Karlić, Mamoru Kato, Kirsten Kübler, Alexander J. Lazar, Constance H. Li, David N. Louis, Adam Margolin, Sancha Martin, Hardeep K. Nahal-Bose, G. Petur Nielsen, Serena Nik-Zainal, Larsson Omberg, Christine P'ng, Marc D. Perry, Paz Polak, Esther Rheinbay, Mark A. Rubin, Colin A. Semple, Dennis C. Sgroi, Tatsuhiro Shibata, Reiner Siebert, Jaclyn Smith, Lincoln D. Stein, Miranda D. Stobbe, Ren X. Sun, Kevin Thai, Derek W. Wright, Chin Lee Wu, Ke Yuan, Junjun Zhang, Alexandra Danyi, Jeroen de Ridder, Carla van Herpen, Martijn P. Lolkema, Neeltje Steeghs, Gad Getz, Quaid Morris and Lincoln D. Stein. 2020. "A deep learning system accurately classifies primary and metastatic cancers using passenger mutation patterns." *Nature Communications* 11(1):1–12.

Joung, Julia, Jesse M. Engreitz, Silvana Konermann, Omar O. Abudayyeh, Vanessa K. Verdine, Francois Aguet, Jonathan S. Gootenberg, Neville E. Sanjana, Jason B. Wright, Charles P. Fulco, Yuen Yi Tseng, Charles H. Yoon, Jesse S. Boehm, Eric S. Lander and Feng Zhang. 2017. "Genome-scale activation screen identifies a lncRNA locus regulating a gene neighbourhood." *Nature* 548(7667):343–346.

Lyu, Boyu and Anamul Haque. 2018. "Deep Learning Based Tumor Type Classification Using Gene Expression Data." *bioRxiv* p. 364323.

Viswanathan, Vasanthi S., Matthew J. Ryan, Harshil D. Dhruv, Shubhroz Gill, Ossia M. Eichhoff, Brinton Seashore-Ludlow, Samuel D. Kaffenberger, John K. Eaton, Kenichi Shimada, Andrew J. Aguirre, Srinivas R. Viswanathan, Shrikanta Chattopadhyay, Pablo Tamayo, Wan Seok Yang, Matthew G. Rees, Sixun Chen, Zarko V. Boskovic, Sarah Javaid, Cherrie Huang, Xiaoyun Wu, Yuen Yi Tseng, Elisabeth M. Roider, Dong Gao, James M. Cleary, Brian M. Wolpin, Jill P. Mesirov, Daniel A. Haber, Jeffrey A. Engelman, Jesse S. Boehm, Joanne D. Kotz, Cindy S. Hon, Yu Chen, William C. Hahn, Mitchell P. Levesque, John G. Doench, Michael E. Berens, Alykhan F. Shamji, Paul A. Clemons, Brent R. Stockwell and Stuart L. Schreiber. 2017. "Dependency of a therapy-resistant state of cancer cells on a lipid peroxidase pathway." *Nature* 547(7664):453–457.