10/06/22

WSI

* De momento he hecho la clasificación image-wise sin darme cuenta, 250 imágenes, 185 casos aprox. Es necesario pasar preprocesamiento y descargar todos los datos de GDC haciendo todo case-wise.
* Para ver más progreso a través de los epochs usar undersampling en vez de oversampling, así las epochs serán más cortas

RNASeq

* RNASeq, 178 casos, por lo que no hay datos de todas las modalidades para todos los casos.
* Descargo archivos STAR-COUNT .tsv 366 archivos, 183 no son autorizados

13/06/22

WSI

* Y si para entrenar el clasificador patch-wise uso todas las imágenes de TCGA pero luego para hacer la clasificación case-wise solo uso los diagnostic slide. Otra opción es expandir con la base de datos GTEX
* Son 244 GB, cinco veces más datos, solución, hacer un undersampling al crear las bases de datos para train, y tmb crear otras bases de datos para el test sin undersampling. De esta forma, no debería petar en el training al cargar todos los splits, ni al hacer el test.

RNASeq

* STAR-COUNTS, primary, solid tissue normal RNA SEQ
* Problema, todos son casos positivos de RNASeq, quizá necesito otras secuencias…
* Comprobar que efectivamente todos los datos son positivos.
* Ordenar los case\_ids y sacarle un sentido.

14/06/22

Sample type:

185 Primary tumor 42 Solid tissue normal

185 Casos en total

183 Casos con WSI -> 2 casos sin WSI

178 Casos con RNASeq -> 7 casos sin RNASeq

185 Casos con CNV

185 Casos con DNAmet

178 Casos con MIRNA -> 7 casos sin MIRNA

Si solo hay 42 muestras de solid tissue normal, tengo casos negativos??? EL POSITIVO O NEGATIVO NO ESTÁ EN CLINICAL, ESTÁ EN LA ETIQUETA

15/06/22

WSI

Hay tres IDs importantes. Patient ID, Case ID, y Sample ID. :

* El 10 CV se hace con el Patient ID
* La clasificación se hace con el Case ID
* El simple ID identifíca tu dato concreto

Patient ID = Case ID: TCGA-YB-A89D ¿?

Sample ID: TCGA-YB-A89D-01A

File ID: TCGA-YB-A89D-01A -01-TS1.53D8C086-00CC-4565-8C52-E81E8F8A8409

“In most cases for each sample, various modalities are available”

20/06/22

Ya entiendo los IDs, problema, viendo los sample types para RNASeq veo que hay 4 de negativos el resto positivos. Creo que hay que cambiar de base de datos.

Problema: desbalanceo de los datos, 10 cv no estratificado no es posible por como está hecho…

Solución: Hacer una estratificación teniendo en cuenta que casos tienen o no un sample negativo

Cambio a TGCA PRAD, Cáncer de próstata.

Duda: como junto los datos de WSI (DX) con el resto si vienen de samples distintas, es decir, tienen sample ID distinto.

Problema: Salen dobles ciertas etiquetas, eso es que para un mismo CASE ID puede haber TS o DX

21/06/22

Voy a usar los tissue slides, al menos de momento, quizá lo tenga que arreglar luego.

Voy a usar el set TCGA-COAD TCGA-READ

10 CV estratificado:

* Estratificar conforme al número de samples por clase, yo creo que con asignar una de las tres clases a la que más samples tenga para esa clase será suficiente para estratificar.
* Para eso debería tener todos los samples que se van a usar (Como)?? Quizá es suficiente con las WSI  
  USAR SAMPLE DENTRO DE BIOSPECIMEN AHÍ ESTAN TODOS LOS SAMPLES

23/06/22

K-fold estratificado:

* Por que tengo más samples usadas en WSI que samples totales?
* Contando usando sample\_sheet (wsi) obtengo más samples que usando sample (total) para todos los casos
* Solucionar problema np.argmax()

24/06/22

No se ha descomprimido entero o no se ha descargado entero

* LOS SAMPLE DE RECURRENT Y BLOOD LOS HE DESECHADO

29/06/22

Consigo entrenar con COAD-READ, mismo problema de las lables que tenia antes. Voy a buscar sol.

Resumen de como se organizan los datos:

De clinical.tsv se toman los pacientes y se hace el 10CV contando cuantas samples positivas y negativas tienen para estratificar. El número de samples se cuenta en sample.tsv, que es un .tsv obtenido de bioespecimen. Sample sheet contiene todos los datos de los archivos descargados. Los lables se obtienen mirando clinical.tsv y el nombre.

Al final si los lables están bien debería estar entrenando bien mas allá de la estratificación etc.

30/06/22

Cambio de BCEntropy loss a crossentropy loss

1/07/22

* Labels parecen estar bien
* No es overfitting
* El entrenamiento funciona con el training set
* Max 500 patches?

Tiene que ser el preproc o los labels

7/07/22

Reestructurar el proyecto,

GTEX: 864 casos

* RNASeq: 328 (328 casos)
* DX: 865 (864 casos)

Número de  samples con datos de WSI y RNASeq 328

GDC: 185 casos

* RNASeq: 182, 4 normal, 178 tumor (178 casos)
* DX: 189 samples, 6 normal tissue, 183 tumor (183 casos)

Número de  samples con datos de WSI y RNASeq 177

Preguntas:

* Cojo todos los samples o los que tienen datos de ambos?

De momento para el K-fold no voy a coger todas esas imágenes de GTEx y ya lo cambiaré luego si hace falta.

12/07/22

+ Approved Users further agree that the acknowledgment shall include the dbGaP accession number to the specific version of the dataset(s) analyzed.

13/07/22

Diferenciación de estadíos usando ómicas si tiene aplicabilidad clínica y sería un gran descubrimiento. El hecho de que no se usen imágenes histológicas me echa mucho para atrás. Pero siempre es verlo.

14/07/22

Knowseq, no se como usarlo con archivos star fusión. Los nombres de los genes me están rayando.

22/07/22

Consigo adaptar Knowseq, ahora solo queda hacer el feature selection con solo el training set

10 CV. Y cambiar el entrenamiento, luego probar con el estadío.

26/07/22

Intentando normalizar los datos RNASeq solo puedo aplicar batch effect correction, esto no parece ser suficiente:

“To determine whether uniform alignment and expression quantification was an essential step, or whether batch effect removal via ComBat by itself was sufficient, we also applied ComBat directly to the level 3 data from GTEx and TCGA (GTEx-quantified data was rescaled using quantile normalization). We found that batch effect removal by itself is not sufficient, and that the combination of uniform processing of sequencing reads followed by additional batch effect removal is required to make data from the TCGA and GTEx projects comparable”

Voy a empezar con el estadío de cáncer.

01/09/22

Resumen:

Estadío de cáncer <=IIA y >IIA:

* No se encuentran genes expresados diferencialmente, se prueba a variar LFC desde 0.5 a 2

Fusión de datos GTEx y GDC:

* Normalización de datos, de acuerdo con el artículo (art) es necesario empezar con los datos crudos, BAM y hacer el mismo preprocesamiento o alignment con esos datos, solo tengo acceso a los ficheros counts, según el artículo no es suficiente con el batch correction, usando PCA se comprueba que esto es verdad.
* Usando los datos de imágenes se obtiene un accuracy cercano al 98%-99%, existe la posibilidad de que usar RNASeq no ofrezca una ventaja.

05/09/22

Reunión con Luis, tres posibles rutas

1. Ver mejora gdc solo rnaseq e imágenes
2. Mirar si los datos del paper de Hao Fu et al. Son públicos
3. Mirar predicción estadio con imágenes
4. Pronóstico

12/09/22

Encuentro dos sets con buena pinta, uno de imágenes uno de RNASeq

13/09/22

En TCIA hay un specimen ID y un slide id, voy a ir por slide id puesto que un espécimen puede tener imágenes de cáncer y normales

23/03/22

PREOCUPACIÓN: PDAC no es lo mismo que PAAD. Quizás eso es lo que diferencia a las bases de GTEX de GDC. GDC tiene un 95% de ductal, pero tmb de otras. No creo que sea por eso.