CLAIMS

WHAT IS CLAIMED IS:

1. A method for training independent parallel cancer signal origin (CSO) classifiers, the method comprising:

obtaining training samples derived from subjects with a known cancer diagnosis, each training sample comprising methylation sequence reads corresponding to nucleic acid fragments in a biological sample collected from each subject and each known cancer diagnosis including a known organ or organ group of a plurality of organs or organ groups affected and a known tumor biology of a plurality of tumor biology classes;

generating, for each training sample, a feature vector based on the methylation sequence reads;

generating a first training data set comprising the feature vectors for the training samples and the known organ or organ group;

training an organ or organ group classifier with the first training data set to predict an organ or organ group from the plurality of organs or organ groups based on an input feature vector;

generating a second training data set comprising the feature vectors for the training samples and the known tumor biology classes; and

training a tumor biology classifier with the second training data set to predict tumor biology from the plurality of tumor biology classes based on input feature vector.

2. The method of any preceding claim, further comprising:

extracting, for each training sample, the known organ or organ group and the known tumor biology class from the known cancer diagnosis and clinical information.

3. The method of any preceding claim, wherein the feature vector is based, at least in part, on methylation features based on the methylation sequence reads.

4. The method of claim 3, wherein the methylation features include: methylation density at one or more loci, density of hypermethylated sequence reads at one or more loci, density of hypomethylated sequence reads at one or more loci, a count of methylation sequence reads determined to be anomalously methylated at one or more loci, or some combination thereof.

5. The method of any preceding claim, wherein generating the first training data set comprises excluding information regarding tumor biology.

6. The method of any preceding claim, wherein generating the second training data set comprises excluding information regarding the affected organ or organ group.

7. The method of any preceding claim, further comprising:

determining, for each feature, information gain in discriminating between the organs or organ groups;

identifying discriminatory features for the organ or organ group classifier based on the information gains; and

modifying the feature vectors of the first training set to consist of the discriminatory features, wherein the modified feature vectors are used in training of the organ or organ group classifier.

8. The method of any preceding claim, further comprising:

determining, for each feature, information gain in discriminating between the tumor biology classes;

identifying discriminatory features for the tumor biology classifier based on the information gains; and

modifying the feature vectors of the second training set to consist of the discriminatory features, wherein the modified feature vectors are used in training of the tumor biology classifier.

9. The method of any preceding claim, wherein the organ or organ group classifier or the tumor biology classifier are machine-learning models.

10. The method of any preceding claim, further comprising training the organ or organ group classifier and the tumor biology classifier in parallel training processes.

11. The method of any preceding claim, further comprising training the organ or organ group classifier prior to training the tumor biology classifier;

optionally wherein outputs of the organ or organ group classifier are appended to the feature vectors of the second training data set prior to training of the tumor biology classifier.

12. The method of any preceding claim, further comprising training the tumor biology classifier prior to training the organ or organ group classifier;

optionally wherein outputs of the tumor biology classifier are appended to the feature vectors of the first training data set prior to training of the organ or organ group classifier.

13. The method of any preceding claim, wherein the organs or organ groups include: breast; prostate; lung; head or neck; anus; cervix; ovary or fallopian tubes; uterus; bladder or urothelial; kidney; stomach or esophagus; liver or intrahepatic bile duct; pancreas, extrahepatic bile duct, or gall bladder; colon or rectum; bone or soft tissue; skin; blood, lymphatic system, or bone marrow; thyroid; ambiguous tissue; or some combination thereof; and/or

wherein the tumor biology classes include: lymphoid neoplasm, myeloid neoplasm, plasma cell neoplasm, neuroendocrine carcinoma or tumor, adenocarcinoma, squamous cell carcinoma and not human-papillomavirus-associated (HPV-associated), HPV-associated carcinoma, hepatocellular carcinoma, neoplasm of Mullerian origin, transitional cell carcinoma, mesenchymal tumor, melanocytic neoplasm, mesothelial neoplasm, other tumor biology, ambiguous tumor biology, or some combination thereof.

14. A non-transitory computer-readable storage medium storing instructions that, when executed by one or more computer processors, cause the one or more computer processors to perform the method of any of claims 1–13.

15. A system comprising:

one or more computer processors; and

the non-transitory computer-readable storage medium of claim 14.