Research question:

Can a Random Forest model trained on array Comparative Genomic Hybridization (aCGH) data outperform a simple KNN in accurately predicting breast cancer subtype, specifically distinguishing between HER2 positive (HER2+), Hormone Receptor positive (HR+), and Triple Negative (TN) subtypes? Additionally, can this model identify important copy number variation (CNV) features associated with each subtype, potentially predicting biomarkers? Finally, what kinds of biomarkers are helpful in the early diagnosis of breast cancer: individual genes, corregulated gene networks, or chromosome locations?

Task assignment:

Date	Goals	Status	Assignee
Apr 2-Apr 4	1.1 Background reading on Canvas	Done	All
	1.2 Write-up a few possible research questions	Done	All
	1.3 Literature review - aGCH and BRCA subtype classification	Done	Jack
Apr 5-Apr 7	2.1 Read up on classification methods and choose two plus a baseline model to implement and compare	Done	Isabel
	2.2 Exploratory data analysis and visualization (Python)	Done	Xinyu
	2.3 Explore R library for QDNA Seq	Done	Shreya
Apr 8	3.1 Finalize the plan, task division, and assignment	Done	All
Apr 9	**CATS research questions & preliminary task list**		
Apr 10-Apr 14	4.1 Map the given data to genes using the library 'GenomicRanges' in R	To Do	Shreya
	4.2 Start implementing baseline models (log reg & kNN)	To Do	Xinyu
	4.3 Start implementing a random forest	To Do	Jack
	4.4 Explore dimensionality reduction/ feature selection for the raw data	To Do	Shreya, Isabel
Apr 15-Apr 16	5.1 Try nested cross-validation on the random forest	To Do	Jack, Isabel

	5.2 Feature importances from the random forest	To Do	Shreya, Xinyu
Apr 17-Apr 21	6.1 Validate all the above models and make predictions	To Do	All
	6.2 Give confidence estimates of predictions	To Do	Jack
	6.3 Calculate and Generate graphs for performance metrics	To Do	All
Apr 22-Apr 25	7.1 Write a draft report	To Do	All
	7.2 Model fine-tuning	To Do	All
	7.3 Look at the final features in detail and try to derive a biological explanation	To Do	Isabel
	7.4 Ensure github code is in the required format	To Do	Jack
Apr 26-May 5	Holidays!	To Do	
May 6	**CATS Draft**		
May 7-May 12	Biomarker prediction	To Do	All
May 13	**CATS peer review**		
May 14-May 20	Finish writing the paper and making the presentation	To Do	All
May 20	**CATS predictions**		
May 21	**CATS predictions & paper, presentations**		