

# Kokiri

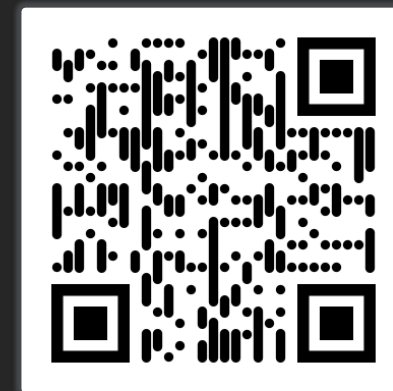
## Random Forest-Based Cohort Comparison and Characterization

Klaus Eckelt, Patrick Adelberger,  
Markus Bauer, Thomas Zichner, Marc Streit

✉ [klaus.eckelt@jku.at](mailto:klaus.eckelt@jku.at)

🐦 [@klaus\\_lml](https://twitter.com/klaus_lml)

<https://jku-vds-lab.at/kokiri>



### Cohort Comparison

Compare by Meta-Data

Compare by AA Mutated

☒ Exclude the cohorts' defining attributes

#### Attribute Importance

Aggr...	Rank	Sele...	# Importance	T Attribute	Distribution
...	...	↓↑	1.↑	↓↑ 🔍 ⚙️	↓↑ 🏷️ ...
		<input type="checkbox"/>	0.0%		100.0%
	1	<input type="checkbox"/>		EGFR	
	2	<input type="checkbox"/>		TP53	
	3	<input type="checkbox"/>		NF1	

#### Cohort Differentiation

Accuracy: 80.6%

Cohort	Gender	Predictions (approx. %)
#5	Female	80.0%
#6	Male	85.0%
#7	Female	85.0%
#8	Male	70.0%

### Cohort Characterization

#### Indistinguishable Items

Aggr...	Rank	Sele...	T Id	Cohort	Prob	# Max Prob...
...	...	↓↑	↓↑ 🔍 ⚙️	↓↑ 🏷️ ...	↓↑ ⚙️	1.↑
		<input type="checkbox"/>				
53		<input type="checkbox"/>	GENIE-JHU-00180-00433			
54		<input type="checkbox"/>	GENIE-JHU-00455-02583			
55		<input type="checkbox"/>	GENIE-JHU-00295-00493			
56		<input type="checkbox"/>	GENIE-JHU-00427-00641			
57		<input type="checkbox"/>	GENIE-JHU-00483-00038			
58		<input type="checkbox"/>	GENIE-JHU-00762-00874			
59		<input type="checkbox"/>	GENIE-JHU-00948-01097			
60		<input type="checkbox"/>	GENIE-JHU-00958-00670			
61		<input type="checkbox"/>	GENIE-JHU-01027-01169			
62		<input type="checkbox"/>	GENIE-JHU-01027-01171			
63		<input type="checkbox"/>	GENIE-JHU-01492-01868			
64		<input type="checkbox"/>	GENIE-JHU-01736-02137			

#### Cohort Association

Compare by *AA Mutated*

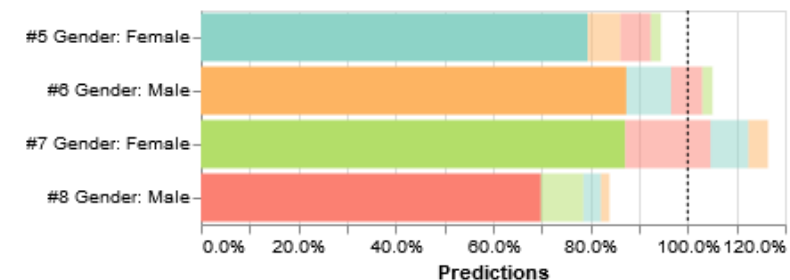
- ☒ Exclude the cohorts' defining attributes

Aggr...	Rank	Selec...	# Importance	T Attribute	Distribution
...	...	↓ ↑ ...	1. ↑ ↓ ...	↓ ↑ 🔍 ...	↓ ↑ ...
		<input type="checkbox"/>	0.0% 100.0%		
1	<input type="checkbox"/>		EGFR		
2	<input type="checkbox"/>		TP53		
3	<input type="checkbox"/>		NF1		

### Indistinguishable Items

Aggr...	Rank	Selec...	T Id	Cohort	Prob	# Max Prob...
...	...	↓ ↑ ...	↓ ↑ 🔍 🚫 ...	↓ ↑ 📅 🚫 ...	↓ ↑ 🚫 ...	1. ↓ 🚫 ...
		<input type="checkbox"/>				
	53	<input type="checkbox"/>	GENIE-JHU-00180-00433			
	54	<input type="checkbox"/>	GENIE-JHU-00455-02583			
	55	<input type="checkbox"/>	GENIE-JHU-00295-00493			
	56	<input type="checkbox"/>	GENIE-JHU-00427-00641			
	57	<input type="checkbox"/>	GENIE-JHU-00483-00038			
	58	<input type="checkbox"/>	GENIE-JHU-00762-00874			
	59	<input type="checkbox"/>	GENIE-JHU-00948-01097			
	60	<input type="checkbox"/>	GENIE-JHU-00958-00670			
	61	<input type="checkbox"/>	GENIE-JHU-01027-01169			
	62	<input type="checkbox"/>	GENIE-JHU-01027-01171			
	63	<input type="checkbox"/>	GENIE-JHU-01492-01868			
	64	<input type="checkbox"/>	GENIE-JHU-01736-02137			

Accuracy: 80.6%




## Cohort Association



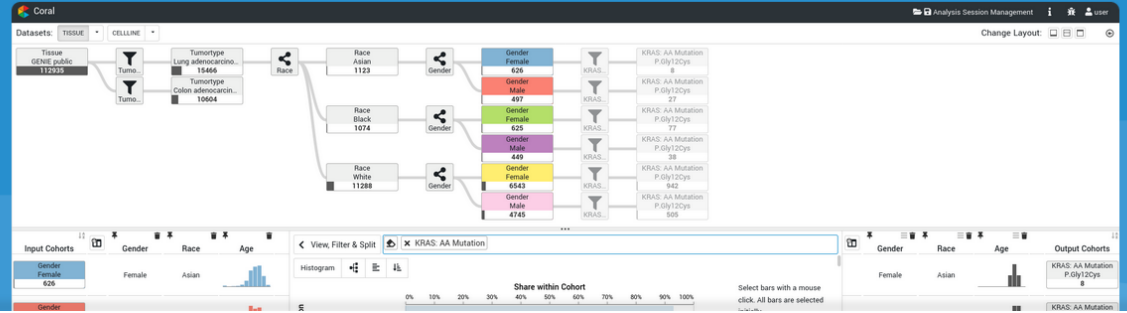
# Motivation



 WHAT'S NEW? FEATURES DATASETS PUBLICATIONS HELP [START ANALYSIS](#)

Coral is a cohort analysis tool to interactively create and refine patient cohorts, while visualizing their provenance in the Cohort Evolution Graph. The resulting cohorts can then be compared, characterized, and inspected down to the level of single entities.

[Watch intro video](#) [Learn more about Coral](#)




### Getting Started

The workflow of Coral consists of two steps: creating cohorts, and characterizing them. Operations from these two categories are carried out in an iterative workflow.

P. Adelberger et al., "Coral: a web-based visual analysis tool for creating and characterizing cohorts." in Bioinformatics 37.23, 2021

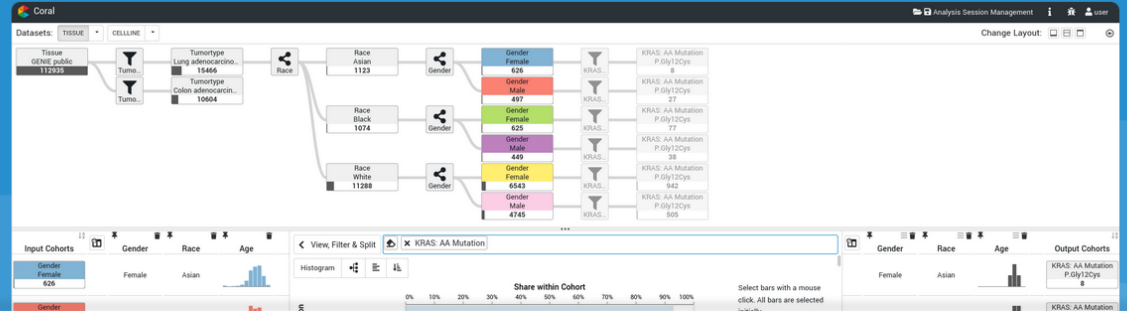
# Motivation



 WHAT'S NEW? FEATURES DATASETS PUBLICATIONS HELP [START ANALYSIS](#)

Coral is a cohort analysis tool to interactively create and refine patient cohorts, while visualizing their provenance in the Cohort Evolution Graph. The resulting cohorts can then be compared, characterized, and inspected down to the level of single entities.

[Watch intro video](#) [Learn more about Coral](#)

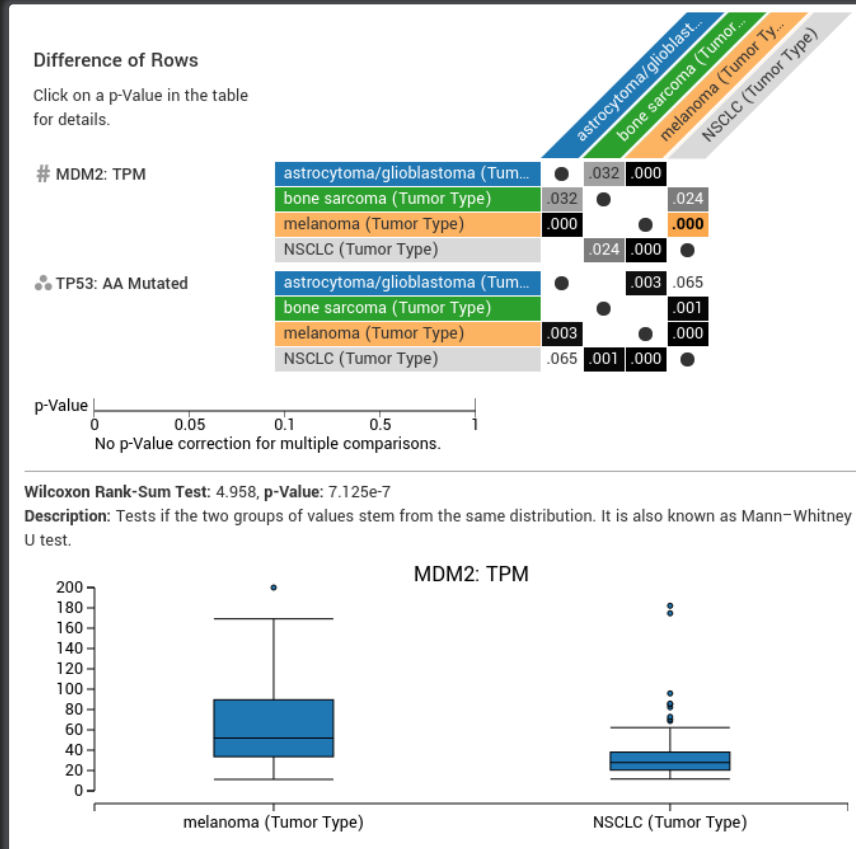


**Getting Started**

The workflow of Coral consists of two steps: creating cohorts, and characterizing them. Operations from these two categories are carried out in an iterative workflow.

P. Adelberger et al., "Coral: a web-based visual analysis tool for creating and characterizing cohorts." in Bioinformatics 37.23, 2021

# Motivation



**Coral** WHAT'S NEW? FEATURES DATASETS PUBLICATIONS HELP [START ANALYSIS](#)

Coral is a cohort analysis tool to interactively create and refine patient cohorts, while visualizing their provenance in the Cohort Evolution Graph. The resulting cohorts can then be compared, characterized, and inspected down to the level of single entities.

[Watch intro video](#) [Learn more about Coral](#)

**Getting Started**

The workflow of Coral consists of two steps: creating cohorts, and characterizing them. Operations from these two categories are carried out in an iterative workflow.

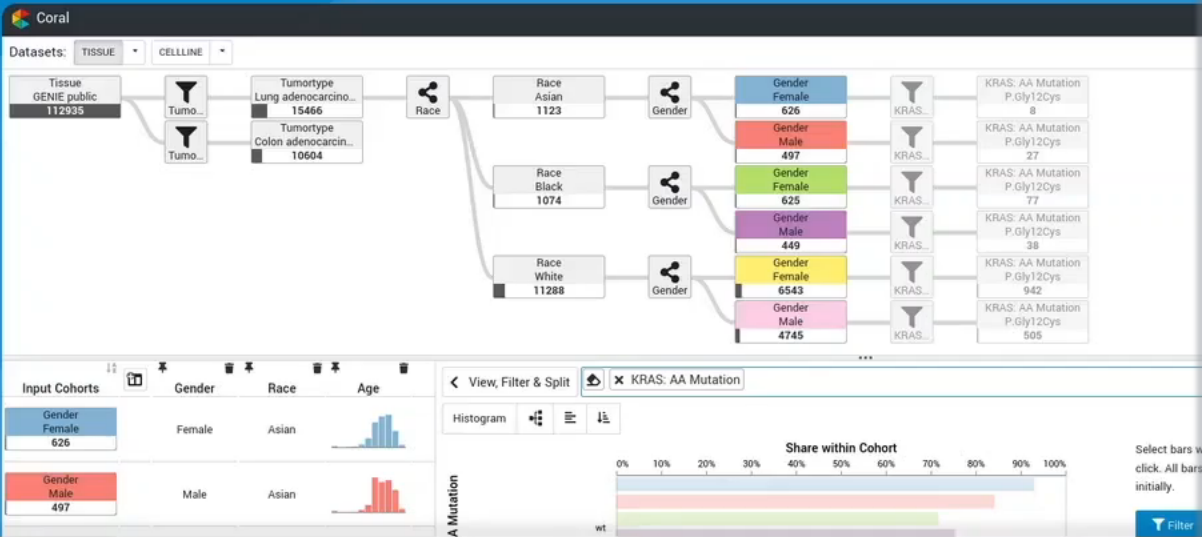
K. Eckelt et al., "TourDino: A Support View for Confirming Patterns in Tabular Data." EuroVis Workshop on Visual Analytics, 2019

P. Adelberger et al., "Coral: a web-based visual analysis tool for creating and characterizing cohorts." in Bioinformatics 37.23, 2021

Coral is a cohort analysis tool to interactively create and refine patient cohorts, while visualizing their provenance in the Cohort Evolution Graph. The resulting cohorts can be compared, characterized, and inspected down to the level of individual items.

Watch intro video

Learn more



## Getting Started

The workflow of Coral consists of two steps: creating cohorts, and characterizing them. Operations from these steps are grouped into two main workflows.

### Cohort Creation


An initial cohort that contains all items of the selected dataset is created automatically. Creation operations allow users to create new sub-cohorts based on different attributes and attribute combinations. Cohorts are refined with the *Filter* operation, or divided into multiple cohorts with the *Split* operation.

### Cohort Characterization

Characterization operations give insights into the cohorts. Similarities and differences between cohorts can be checked visually with the *View* operation, and statistically with the *Compare* operation. Additional operations give access to prevalence information and the data of individual items.

The NEW ENGLAND JOURNAL of MEDICINE

CORRESPONDENCE



**Distribution of KRAS<sup>G12C</sup> Somatic Mutations across Race, Sex, and Cancer Type**

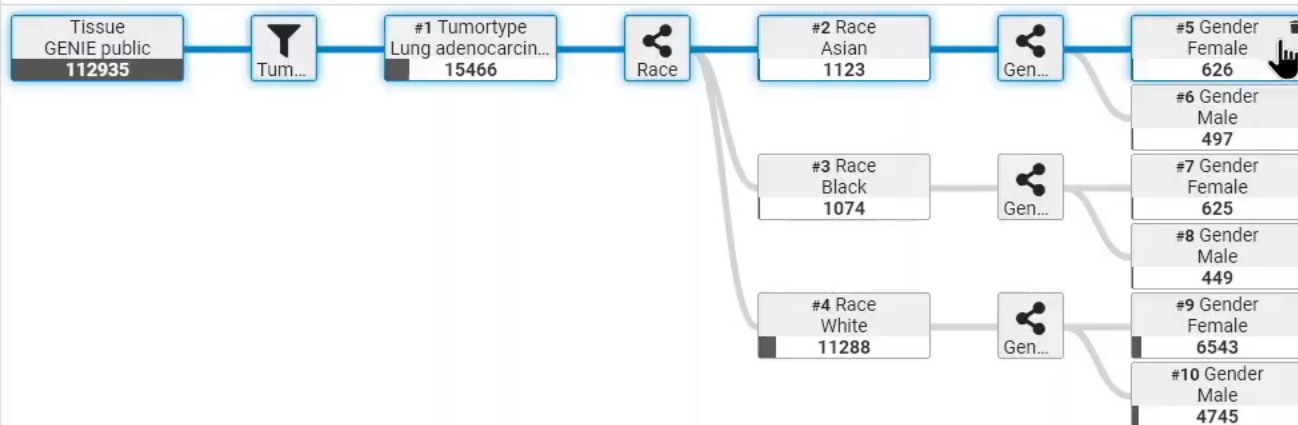

**TO THE EDITOR:** Hong et al. (Sept. 24 issue)<sup>1</sup> present results of an early-phase clinical trial of sotorasib, which showed promising clinical benefit. Patients had non–small-cell lung cancer (NSCLC; 46%) or colorectal cancer (33%), and White patients constituted 76% of the cohort. As drugs are being developed for the previously “undruggable” KRAS<sup>G12C</sup> mutation, it is imperative to study the distribution of this mutation across sex, race,<sup>2</sup> and all cancers.

We extracted data from the registry of the American Association for Cancer Research Project Genomics Evidence Neoplasia Information Exchange (GENIE), version 8.0<sup>3</sup> (see the Methods section in Supplementary Appendix 1). We studied the distribution of KRAS<sup>G12C</sup> mutations in 32,138 patients with cancer across race (Asian, Black, and White) and sex and in 10 cancer types (Table S1.1 in Supplementary Appendix 2). A total of 2045 patients (6.4%) were Asian, 2355 (7.3%) were Black, and 27,738 (86.3%) were White.

with colorectal cancer and those with NSCLC, female patients harbored significantly more KRAS<sup>G12C</sup> mutations than male patients (Table S1.3 in Supplementary Appendix 2).

Among patients with NSCLC, White and Black patient groups were enriched for KRAS<sup>G12C</sup> mutations more than Asians (White patients, 1153 of 8892 [13%]; Black patients, 94 of 862 [10.9%]; and Asian patients, 25 of 690 [3.6%]) (P<0.001). It is striking that there were differences by sex within the same ethnic groups of patients with NSCLC. KRAS<sup>G12C</sup> mutations occurred more often in White female patients than in White male patients with NSCLC (odds ratio, 1.4; 95% confidence interval [CI], 1.3 to 1.6; Q<0.001) and more often in Asian male patients than in Asian female patients (odds ratio, 5.2; 95% CI, 1.9 to 17.9; Q=0.01) (Fig. 1B, and Tables S1.4 and S1.5 in Supplementary Appendix 2). Among patients with colorectal cancer, White female patients were enriched for KRAS<sup>G12C</sup> mu-

A.H. Nassar et al., "Distribution of KRAS G12C somatic mutations across race, sex, and cancer type." in New England Journal of Medicine 384 2: 2021

Datasets: TISSUE CELLLINE [New Session](#)Change Layout: 



Datasets: TISSUE CELLLINE New Session

Change Layout: [Icons]

Input Cohorts

#5 Gender Female  
626

#6 Gender Male  
497

#7 Gender Female  
625

#8 Gender Male  
449

#9 Gender Female  
6543

#10 Gender Male  
4745

Gender

Female

Male

Female

Male

Female

Male

Race

Asian

Asian

Black

Black

White

White

< Characterize

Overlap of Cohorts Cohorts do not overlap.

Cohort Comparison

Compare by Meta-Data

Compare by Mutation Frequency

☒ Exclude the cohorts' defining attributes

Attribute Importance

Aggr... Rank Selec... # Importance T Attribute Distribution

1 TP53

2 EGFR

3 STK11

4 FGFR4

5 KRAS

6 PIK3CA

Cohort Differentiation

Accuracy: 74.5%

#5 Gender: Female

#8 Gender: Male

#7 Gender: Female

#8 Gender: Male

#9 Gender: Female

#10 Gender: Male

Predictions: 31.2%

predict: #8 Gender: Male

target: #9 Gender: Female

share: 0.311835577605

Cohort Characterization

Item Predictions

Aggr... Rank Selec... T Item Id Cohort Cohort Probability # Max Probabil...

1 GENIE-MSK-P-0009215-T01-IM5

2 GENIE-MSK-P-0004303-T01-IM5

3 GENIE-MSK-P-0006883-T01-IM5

4 GENIE-MSK-P-0012745-T01-IM5

5 GENIE-MSK-P-0004678-T01-IM5

6 GENIE-MSK-P-0005467-T01-IM5

7 GENIE-MSK-P-0012404-T01-IM5

8 GENIE-MSK-P-0012577-T01-IM5

9 GENIE-MSK-P-0006588-T01-IM5

10 GENIE-MSK-P-0010345-T02-IM5

11 GENIE-MSK-P-0000260-T01-IM5

Cohort Overview

[Loading Icon]

Clear



Datasets: TISSUE CELLLINE New Session

Change Layout: [Icons]

Input Cohorts

#5 Gender Female  
626

#6 Gender Male  
497

#7 Gender Female  
625

#8 Gender Male  
449

#9 Gender Female  
6543

#10 Gender Male  
4745

Gender

Female

Male

Female

Male

Female

Male

Race

Asian

Asian

Black

Black

White

White

Characterize

Overlap of Cohorts Cohorts do not overlap.

Cohort Comparison

Compare by Meta-Data Compare by Mutation Frequency ☒ Exclude the cohorts' defining attributes

Attribute Importance

Aggr... Rank Selec... # Importance T Attribute Distribution

0.0% 100.0%

1 TP53

2 EGFR

3 STK11

4 KRAS

5 FGFR4

6 PIK3CA

Cohort Differentiation

Accuracy: 74.3%

#5 Gender: Female

#6 Gender: Male

#7 Gender: Female

#8 Gender: Male

#9 Gender: Female

#10 Gender: Male

0.0% 40.0% 80.0% 120.0% 160.0% 200.0% 240.0%

Predictions

Cohort Characterization

Item Predictions

Aggr... Rank Selec... T Item Id Cohort Cohort Probability # Max Probabil...

0.0% 100.0% 0.0% 100.0%

1 GENIE-DFCI-001073-5215

2 GENIE-DFCI-008031-11008

3 GENIE-DFCI-009772-9552

4 GENIE-DFCI-000353-8791

5 GENIE-DFCI-001583-6568

6 GENIE-DFCI-001583-9841

7 GENIE-DFCI-008145-4847

8 GENIE-DFCI-008613-8293

9 GENIE-DFCI-009075-5945

10 GENIE-DFCI-009205-7970

11 GENIE-DFCI-009522-8607

12 GENIE-DFCI-010598-9170

Cohort Overview

Clear

Input Cohorts	Gender	Race
#5 Gender Female 626	Female	Asian
#6 Gender Male 497	Male	Asian
#7 Gender Female 625	Female	Black
#8 Gender Male 449	Male	Black
#9 Gender Female 6543	Female	White
#10 Gender Male 4745	Male	White

< Characterize

Overlap of Cohorts Cohorts do not overlap.

Cohort Comparison

Compare by Meta-Data

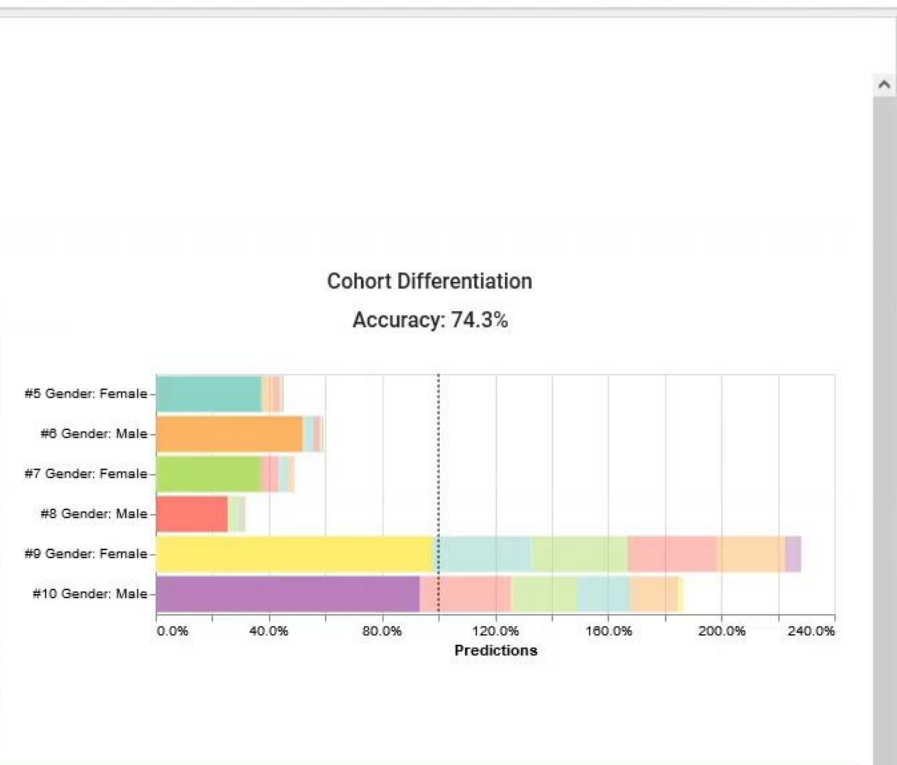
Compare by Mutation Frequency

☒ Exclude the cohorts' defining attributes

Attribute Importance

Aggr... Rank Selec... # Importance T Attribute Distribution

			0.0%	100.0%	
1	<input type="checkbox"/>			TP53	
2	<input type="checkbox"/>			EGFR	
3	<input type="checkbox"/>			STK11	
4	<input type="checkbox"/>			KRAS	
5	<input type="checkbox"/>			FGFR4	
6	<input type="checkbox"/>			PIK3CA	



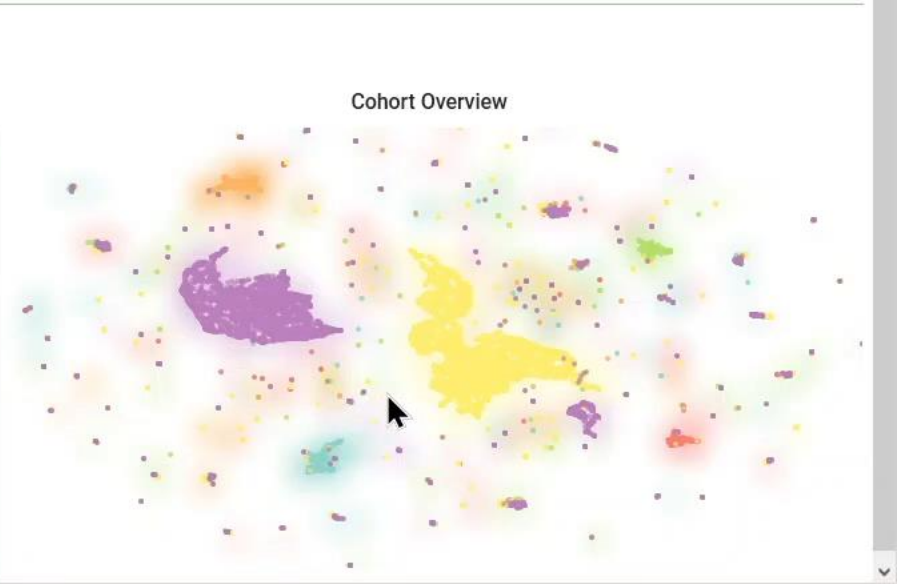
Clear

Cohort Characterization

Item Predictions

Aggr... Rank Selec... T Item Id Cohort Cohort Probability # Max Probabil...

						0.0%	100.0%	0.0%	100.0%
1	<input type="checkbox"/>		GENIE-MSK-P-0023910-T01						
2	<input type="checkbox"/>		GENIE-MSK-P-0023910-T04						
3	<input type="checkbox"/>		GENIE-MSK-P-0023910-T06						
4	<input type="checkbox"/>		GENIE-MSK-P-0023910-T08						
5	<input type="checkbox"/>		GENIE-MSK-P-0023910-T09						
6	<input type="checkbox"/>		GENIE-MSK-P-0023910-T11						
7	<input type="checkbox"/>		GENIE-MSK-P-0020717-T01						
8	<input type="checkbox"/>		GENIE-MSK-P-0039320-T01						
9	<input type="checkbox"/>		GENIE-PHS-0f1debd3-focus						
10	<input type="checkbox"/>		GENIE-PHS-2a4e01de-focus						
11	<input type="checkbox"/>		GENIE-PHS-6498c568-focus						
12	<input type="checkbox"/>		GENIE-PHS-891274e4-focus						



Input Cohorts

#3 Race  
Black  
1074

#1 Tumortype  
Lung adenocarcin...  
15466



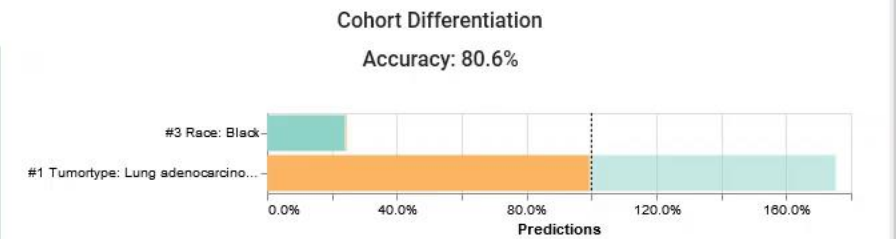
Cohort Comparison

Compare by *Meta-Data* Compare by *Mutation Frequency* ☒ Exclude the cohorts' defining attributes

75/500

Attribute Importance

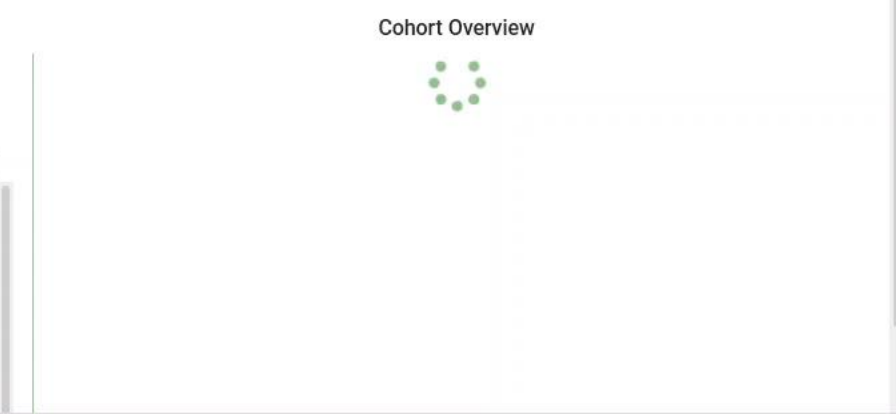
Aggr...	Rank	Selec...	# Importance	Attribute	Distribution
			1 1 1		
		<input type="checkbox"/>	0.0%	100.0%	
1	<input type="checkbox"/>			FGFR4	
2	<input type="checkbox"/>			TP53	<div><div></div><div></div></div>
3	<input type="checkbox"/>			EGFR	<div><div></div><div></div></div>
4	<input type="checkbox"/>			KRAS	<div><div></div><div></div></div>
5	<input type="checkbox"/>			BRCA1	<div><div></div><div></div></div>
6	<input type="checkbox"/>			STK11	<div><div></div><div></div></div>



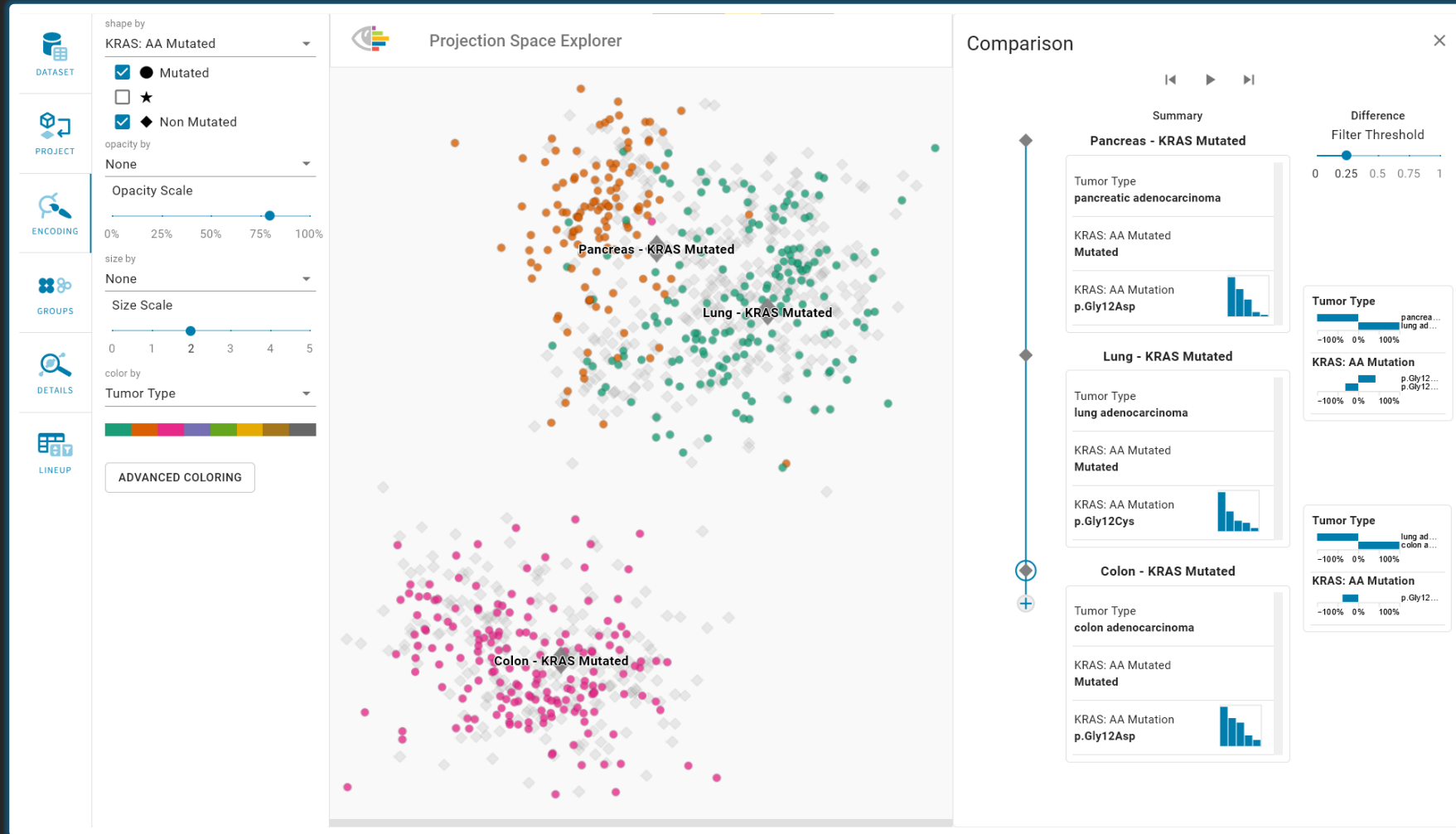
Cohort Characterization

Item Predictions

Aggr...	Rank	Selec...	Item Id	Cohort	Cohort Probability	# Max Probabil...
		<input type="checkbox"/>			0.0%	100.0%
1	<input type="checkbox"/>		GENIE-DUKE-P60-S60	<div><div></div><div></div></div>		
2	<input type="checkbox"/>		GENIE-MSK-P-0019570-T01-IM6	<div><div></div><div></div></div>		
3	<input type="checkbox"/>		GENIE-MSK-P-0021630-T02-IM6	<div><div></div><div></div></div>		
4	<input type="checkbox"/>		GENIE-MSK-P-0034970-T01-IM6	<div><div></div><div></div></div>		
5	<input type="checkbox"/>		GENIE-MSK-P-0044496-T01-IM6	<div><div></div><div></div></div>		
6	<input type="checkbox"/>		GENIE-MSK-P-0045887-T01-IM6	<div><div></div><div></div></div>		
7	<input type="checkbox"/>		GENIE-MSK-P-0046474-T02-IM6	<div><div></div><div></div></div>		
8	<input type="checkbox"/>		GENIE-DFCI-010651-11678	<div><div></div><div></div></div>		
9	<input type="checkbox"/>		GENIE-MSK-P-0026853-T01-IM6	<div><div></div><div></div></div>		



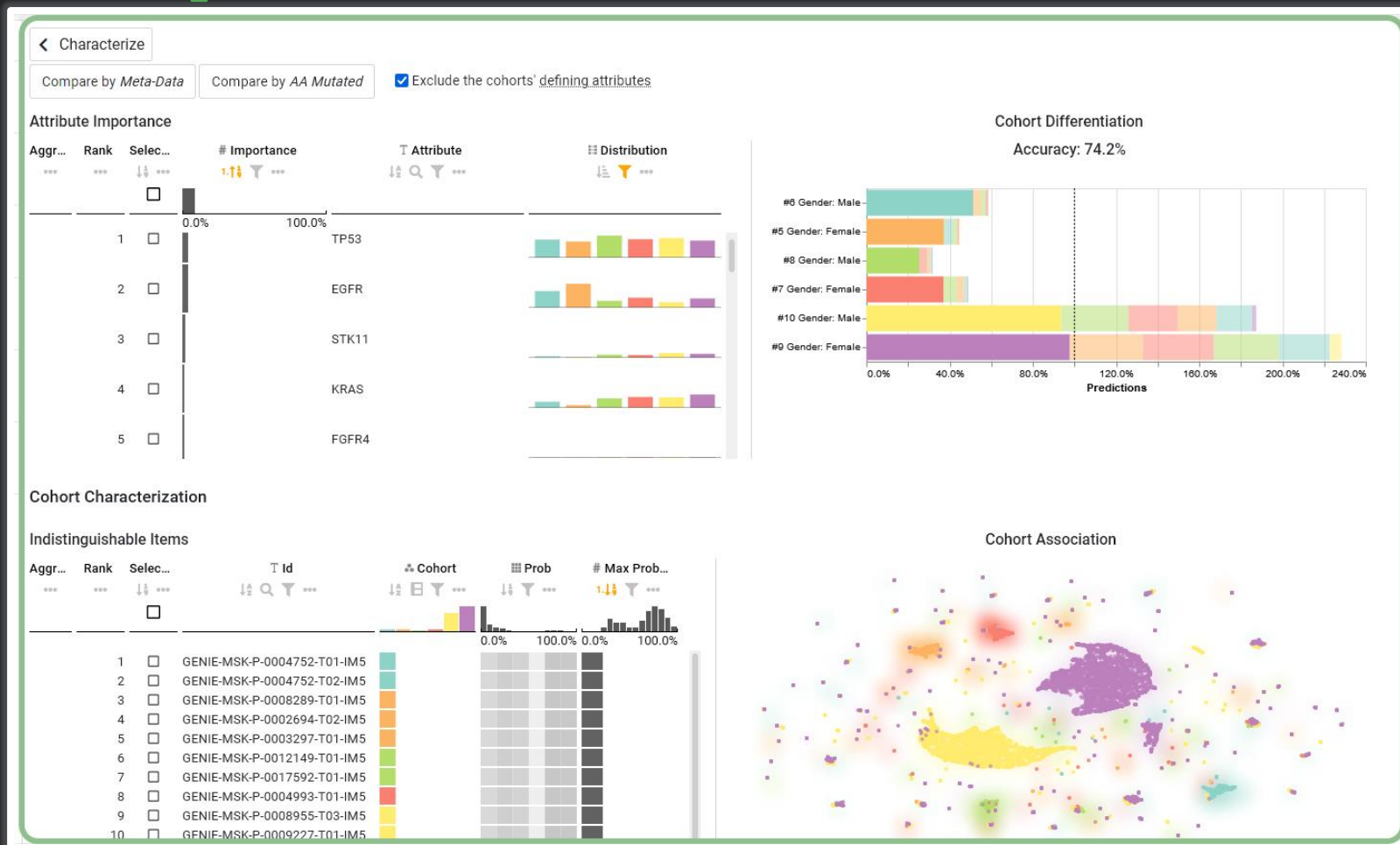
# Ongoing & Future Work



Friday, Oct 21<sup>st</sup>  
09:00 AM  
Oklahoma Station 1

K. Eckelt et al., "Visual Exploration of Relationships and Structure in Low-Dimensional Embeddings,"  
in IEEE Transactions on Visualization and Computer Graphics, 2022, doi: 10.1109/TVCG.2022.3156760.

# Kokiri Random Forest-Based Cohort Comparison and Characterization



Klaus Eckelt

 klaus.eckelt@jku.at

 @klaus\_lml

