Pset3_Exercises

your-name-here

Part 0: Warmup

We are interested to study the proportion of genes in the X chromosome.

Seven of the 100 human genes we sampled randomly from the human genome were found to occur on the X chromosome. The sample fraction of genes on chr. X was thus 7/100 = .07.

1. For each of the statements, indicate whether True or False.

.07 is the proportion of all human genes on the X chromosome.

.07 estimates p, the parameter of proportion of all human genes on the X chromosome.

The proportion of all human genes sampled that belong on the X chromosome has a sampling distribution.

Solution:

False, True, True.

Part 1: Discrete probability

We generate the plodiy vector from the first week again, with some modifications as shown below.

1. If I randomly sample a chromosome, what is the probability that I will get a diploid chromosome?

Solution: 13/23

2. If I randomly sample a chromosome, what is the probability that I will get a non-diploid chromosome?

Solution: 10/23

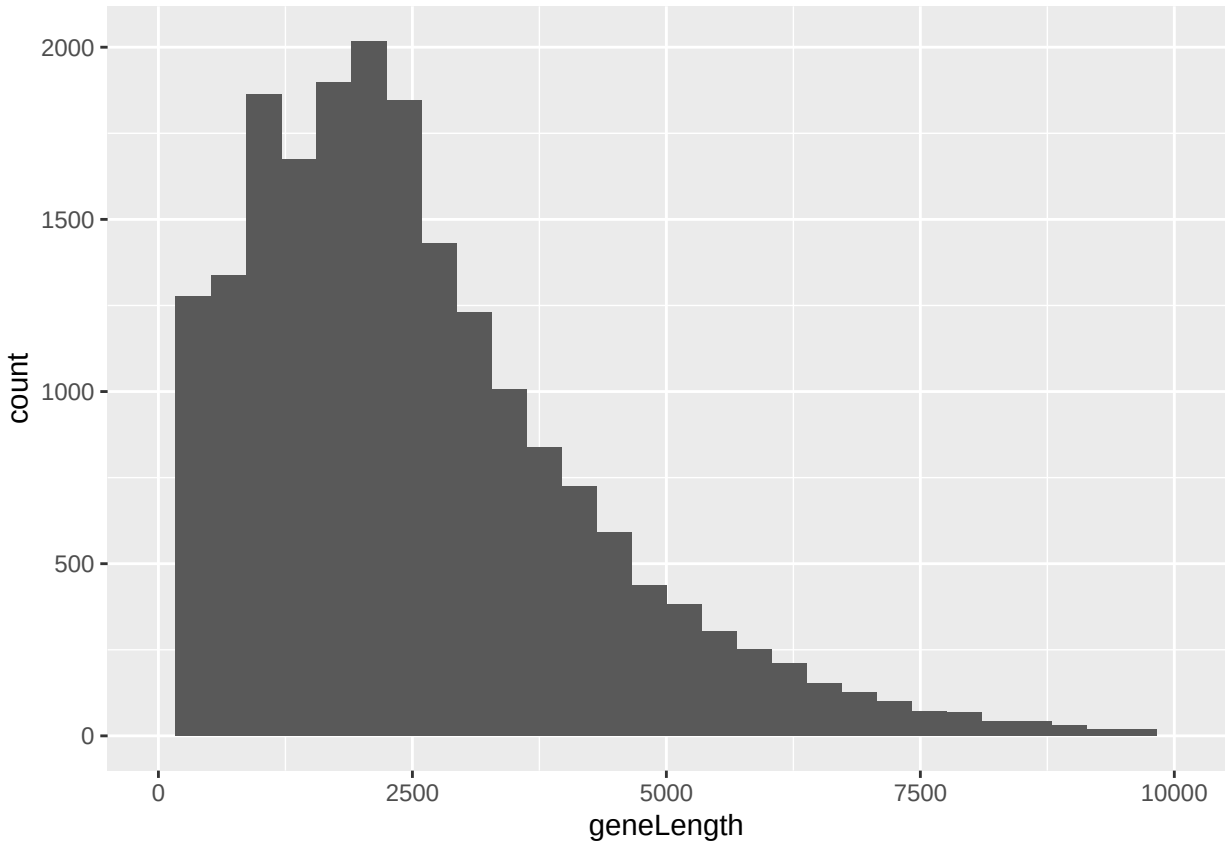
Part 3: Continuous probability

Load in the gene length dataset.

1. Make a histogram showing the distribution of the population using ggplot. Can you change the scale of the x-axis to remove some of the outliers?

```
## Warning: Removed 148 rows containing non-finite values (stat_bin).
```

```
## Warning: Removed 2 rows containing missing values (geom_bar).
```



2. What is the probability that a gene sampled is less than 5000bp? Use R to compute it.

```
## Probability:  0.9023657
```

3. What is the probability that a gene sampled between 4000bp and 5000bp?

```
## Probability:  0.08255298
```

Part 4: Probability in sequencing.

Turns out, there are multiple steps of sampling from patient to bioinformatics analysis that affects our final measurements.

Suppose that you are screening for early detection of lung cancer. For the population you are screening, you know that roughly 5% of the population has lung cancer.

1. You enroll 200 patients in your study via random sampling. How many patient do you expect to have lung cancer?

Solution: $5/200 = 2.5\%$

To test for lung cancer, you decided to study a specific somatic variant that occurs at 1% allelic fraction among those with lung cancer.

You collect blood draw samples, and sequence them. Each patient's sample has 5,000 genomic equivalence - that is, 5000 copies of this gene, whether it is wildtype or mutant.

The protocol you use to sequence the samples only picks up 50% of the 5,000 genomic equivalence.

2. Suppose a patient sample you are analyzing from the sequencer has the mutation. What is the number of mutant variants you expect to see?

Solution: 5000 * .5 * .01 = 25

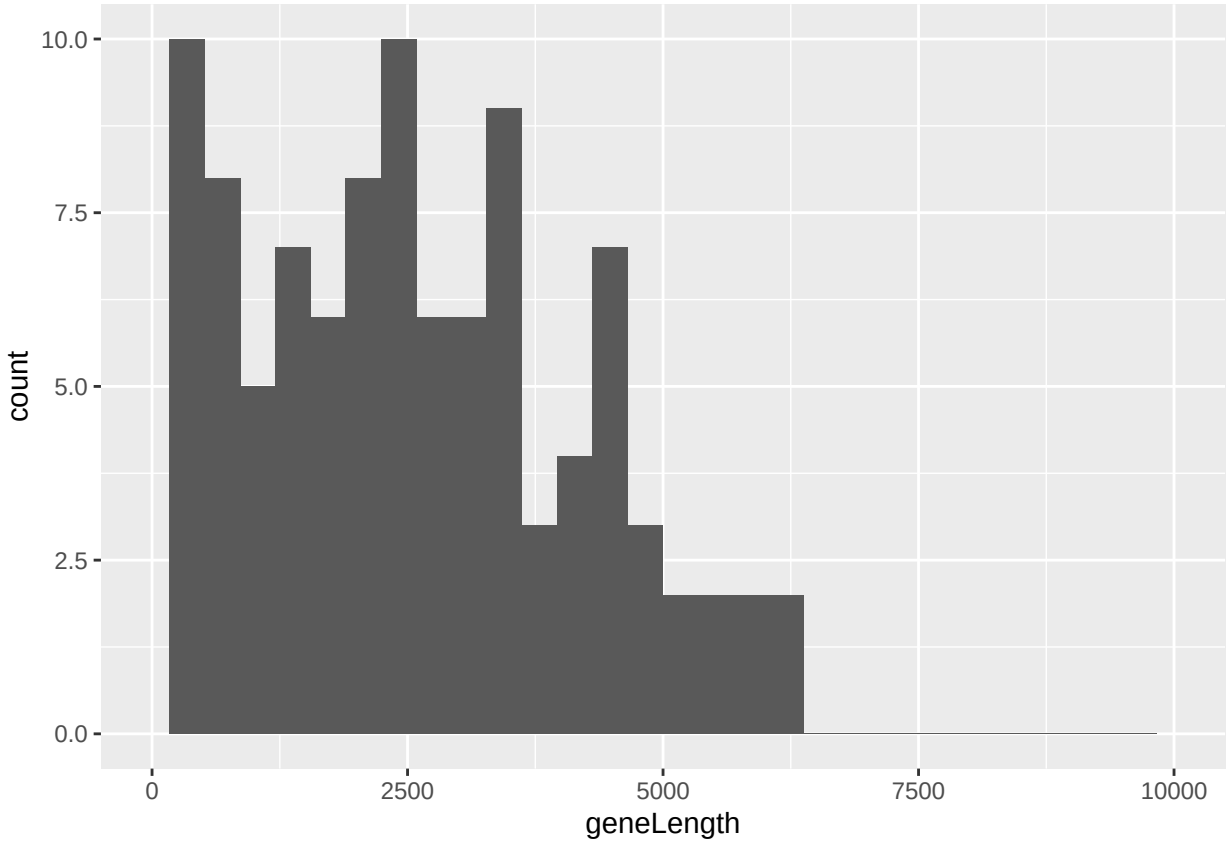
From the 200 patients in the study, what is total number of mutant variants aggregated you expect to see?

2.5 * 25 = 62.5

Part 5: One sample T-test

1. From the gene lengths dataframe, sample 100 genes using `sample_n()` function. We are going to treat the gene lengths dataframe as the true population, and our dataframe from sampling as our sample. Make a histogram showing the distribution of the sample.

```
## Warning: Removed 2 rows containing missing values (geom_bar).
```



2. Run a t-test on the mean of gene length, testing whether the sample mean is same as 2622, the true population mean. In the `t.test()` function, use the `mu` argument to indicate the null hypothesis.What is the null hypothesis? Do you reject the null hypothesis?

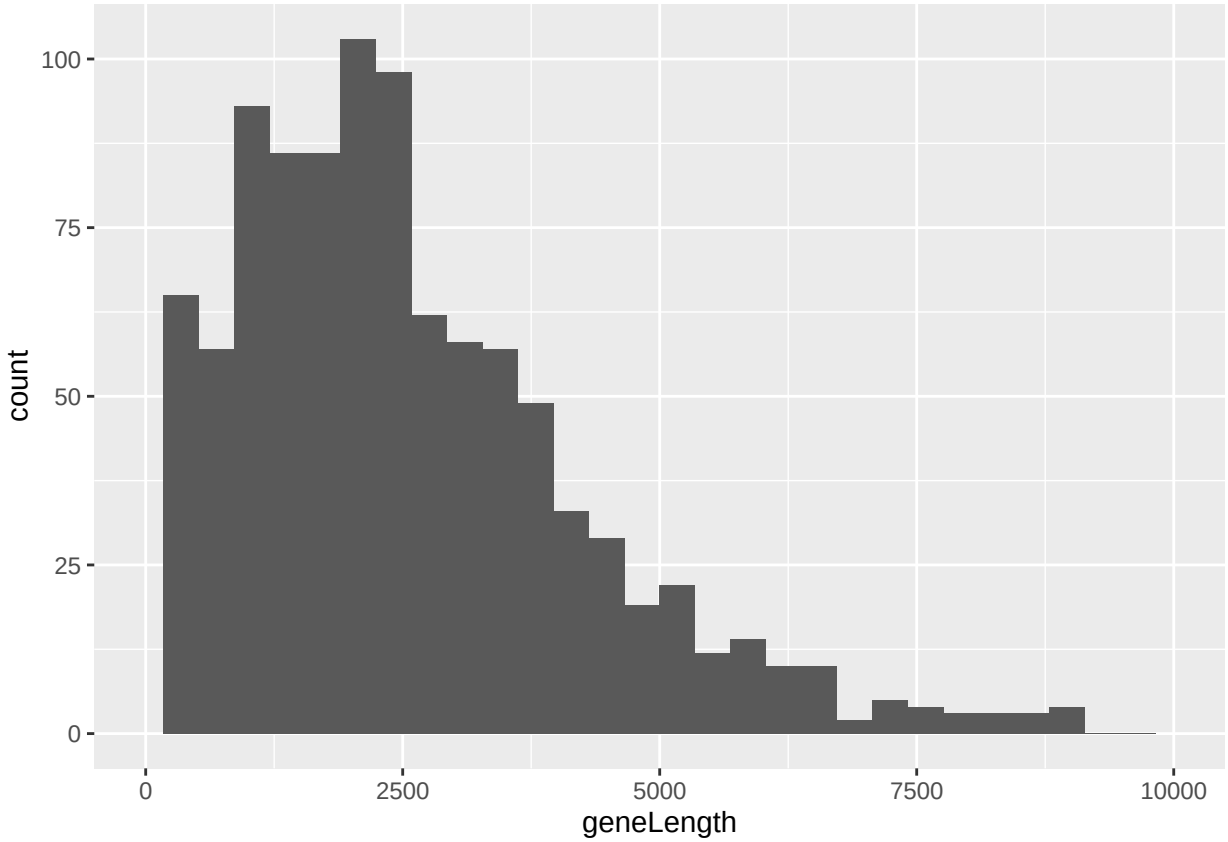
```
##
##  One Sample t-test
##
## data:  sample_genes$geneLength
```

```
## t = -0.11241, df = 99, p-value = 0.9107
## alternative hypothesis: true mean is not equal to 2622
## 95 percent confidence interval:
##  2289.451 2918.889
## sample estimates:
## mean of x
##   2604.17
```

3. Repeat the entire process, now with 1000 genes. Compare the 95% confidence interval between the two tests. What do you notice? What are some factors that explain the different confidence intervals?

```
## Warning: Removed 5 rows containing non-finite values (stat_bin).
```

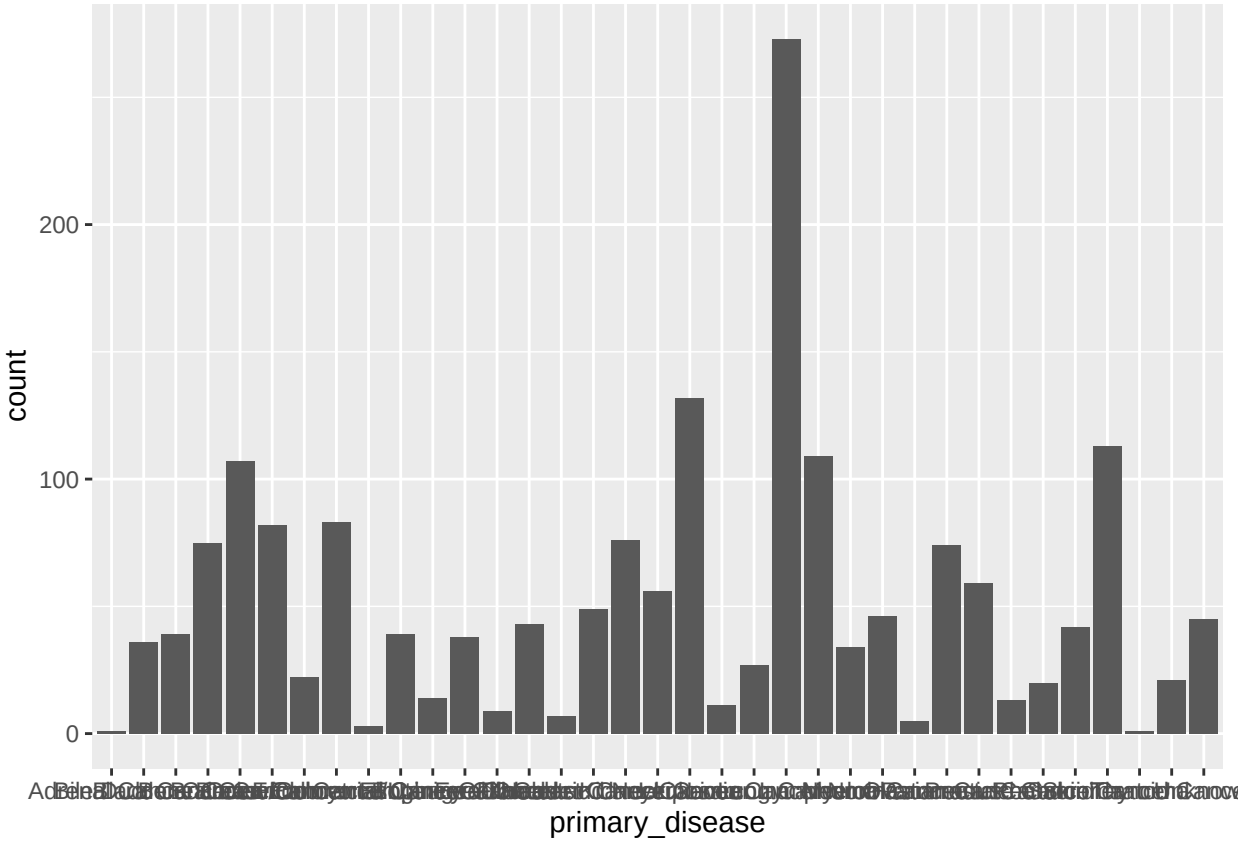
```
## Warning: Removed 2 rows containing missing values (geom_bar).
```



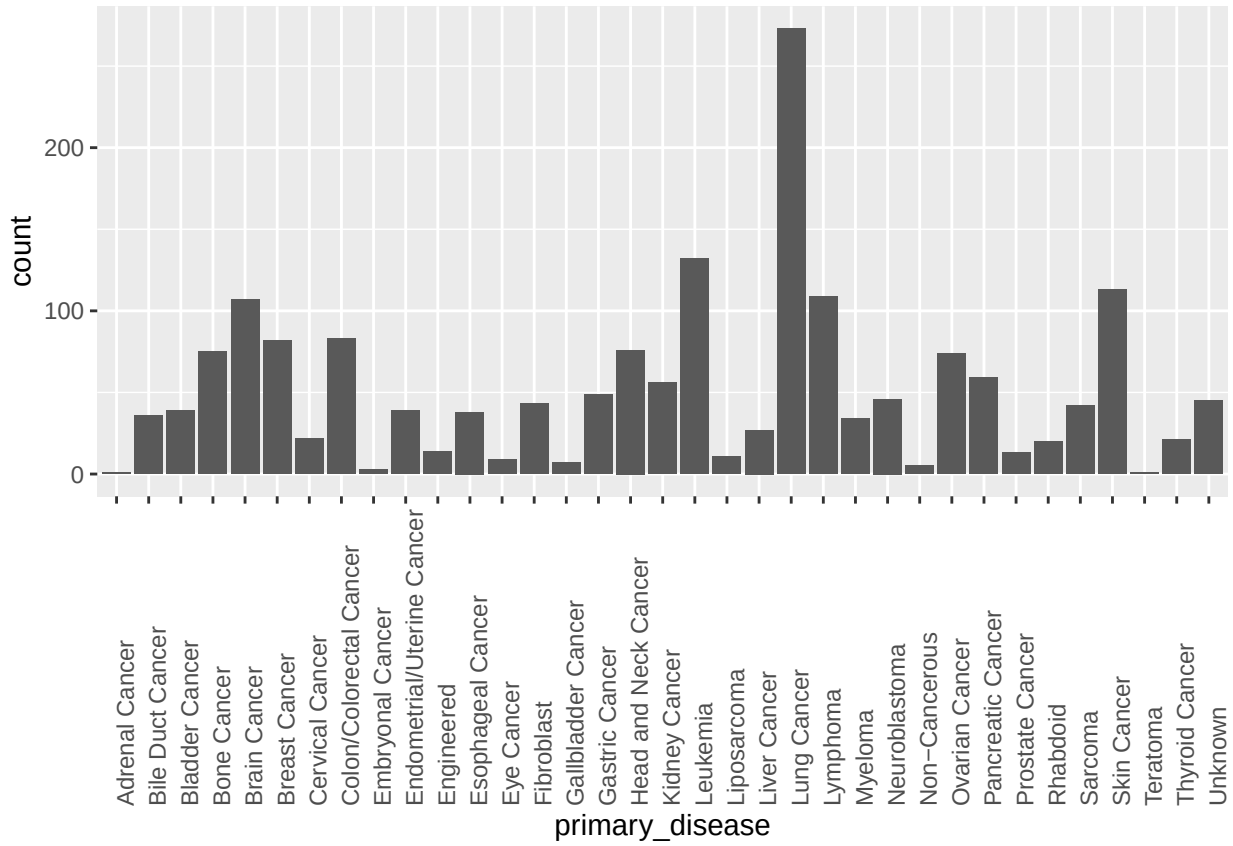

```
##
##  One Sample t-test
##
## data:  sample_genes$geneLength
## t = -0.09338, df = 999, p-value = 0.9256
## alternative hypothesis: true mean is not equal to 2622
## 95 percent confidence interval:
##  2503.583 2729.659
## sample estimates:
## mean of x
##  2616.621
```

Part 6: Plotting tricks

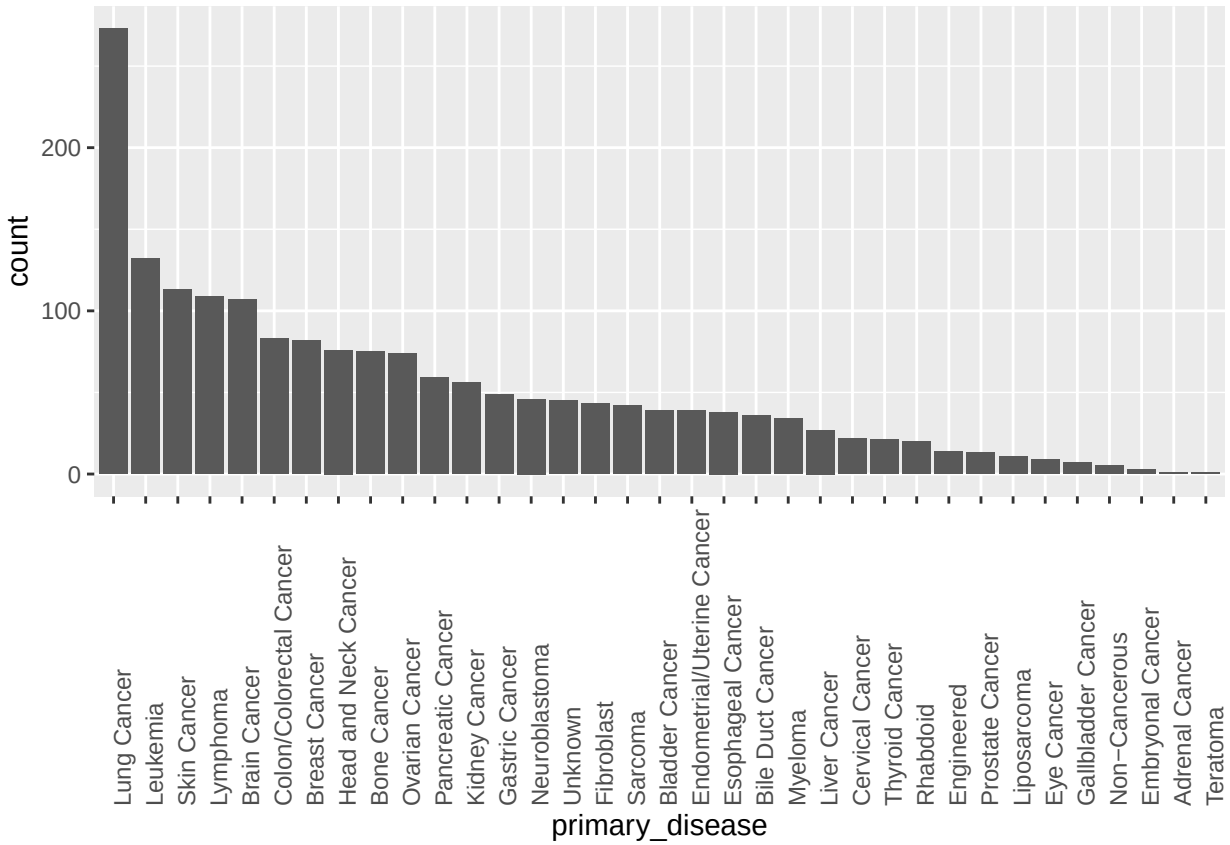
1. Load in metadata, and make a barplot of the `primary_disease` to understand the frequency of cancer types.



2. Let's try to make this more readable. Add the following code to your barplot to rotate the labels: `+ theme(axis.text.x=element_text(angle = 90, hjust = 0))`



3. Let's reorder the x-axis by the count. To do so, add an option called `+ scale_x_discrete(limits = new_axis)`, where `new_axis` is a vector of the x-axis labels in our desired order. Let `new_axis = names(sort(table(metadata$primary_disease), decreasing = T))`. Try it out.


count
200
100
0
Lung Cancer
Leukemia
Skin Cancer
Lymphoma
Brain Cancer
Colon/Colorectal Cancer
Breast Cancer
Head and Neck Cancer
Bone Cancer
Ovarian Cancer
Pancreatic Cancer
Kidney Cancer
Gastric Cancer
Neuroblastoma
Unknown
Fibroblast
Sarcoma
Bladder Cancer
Endometrial/Uterine Cancer
Esophageal Cancer
Bile Duct Cancer
Myeloma
Liver Cancer
Cervical Cancer
Thyroid Cancer
Rhabdoid
Engineered
Prostate Cancer
Liposarcoma
Eye Cancer
Gallbladder Cancer
Non−Cancerous
Embryonal Cancer
Adrenal Cancer
Teratoma
primary_disease