Personalized Medicine:

Redefining Cancer Treatment

Member

Piyatida
Meesatean
6205008

Rujikorn
Sangsumrit
6205024

Ornrakorn
Mekchaiporn
6205025

Steps in Data Science

Outline

01
DATA UNDERSTANDING

02
DATA PREPARATION

03
MODELING

04
RESULT

05
CONCLUSION & DISCUSSION

01 Data Understanding

- เป็นข้อมูล genetic mutations ที่ผลมาจากผลตรวจทาง clinical ในรูปแบบ text
- ข้อมูล genetic mutation นี้ แบ่งเป็น **9 class***
- ข้อมูลที่ใช้มีทั้งหมด 2 ไฟล์ ได้แก่
  - data_variants (ID, Gene, Variations, Class)
  - data_text (ID, Text)

	ID	Gene	Variation	Class
0	0	FAM58A	Truncating Mutations	1
1	1	CBL	W802*	2
2	2	CBL	Q249E	2
3	3	CBL	N454D	3
4	4	CBL	L399V	4

ตัวอย่างข้อมูลจากไฟล์ data_variants

	ID	TEXT
0	0	Cyclin-dependent kinases (CDKs) regulate a var...
1	1	Abstract Background Non-small cell lung canc...
2	2	Abstract Background Non-small cell lung canc...
3	3	Recent evidence has demonstrated that acquired...
4	4	Oncogenic mutations in the monomeric Casitas B...

ตัวอย่างข้อมูลจากไฟล์ data_text

01 Data Understanding

ตัวอย่างข้อมูลจากไฟล์ data_variants

	ID	Gene	Variation	Class
0	0	FAM58A	Truncating Mutations	1
1	1	CBL	W802*	2
2	2	CBL	Q249E	2
3	3	CBL	N454D	3
4	4	CBL	L399V	4

1. **ID** = (the id of the row used to link the mutation to the clinical evidence),
2. **Gene** (the gene where this genetic mutation is located),
3. **Variation** (the aminoacid change for this mutations),
4. **Class** (1-9 the class this genetic mutation has been classified on)

ตัวอย่างข้อมูลจากไฟล์ data_text

	ID	TEXT
0	0	Cyclin-dependent kinases (CDKs) regulate a var...
1	1	Abstract Background Non-small cell lung canc...
2	2	Abstract Background Non-small cell lung canc...
3	3	Recent evidence has demonstrated that acquired...
4	4	Oncogenic mutations in the monomeric Casitas B...

1. **ID** (the id of the row used to link the clinical evidence to the genetic mutation),
2. **Text** (the clinical evidence used to classify the genetic mutation)

01 Data Understanding: Exploratory Data Analysis (EDA)

ไฟล์ data_variants

```
train_variants.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3321 entries, 0 to 3320
Data columns (total 4 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   ID         3321 non-null   int64
 1   Gene       3321 non-null   object
 2   Variation  3321 non-null   object
 3   Class      3321 non-null   int64
dtypes: int64(2), object(2)
memory usage: 103.9+ KB
```

```
#ตรวจสอบ na ในข้อมูล
train_variants.isna().sum()
```

```
ID           0
Gene         0
Variation    0
Class        0
dtype: int64
```

ไฟล์ data_text

```
train_text.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3321 entries, 0 to 3320
Data columns (total 2 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   ID      3321 non-null   int64
 1   TEXT    3316 non-null   object
dtypes: int64(1), object(1)
memory usage: 52.0+ KB
```

```
#ตรวจสอบ na ในข้อมูล
train_text.isna().sum()
```

```
ID      0
TEXT    5
dtype: int64
```

มี na ในข้อมูลอยู่ 5

01 Data Understanding: Exploratory Data Analysis (EDA)

- **จัดการกับ na ในไฟล์ data_text**

```
#จัดการกับ na ในข้อมูล โดยเติม unknown เข้าไป
train_text['TEXT'].fillna('unknown',inplace=True)
```

```
#ตรวจสอบ na ในข้อมูล
train_text.isna().sum()
```

```
ID      0
TEXT    0
dtype: int64
```

ไม่มี na ในข้อมูลแล้ว

01 Data Understanding: Exploratory Data Analysis (EDA)

data_variants + data_text = data_merge

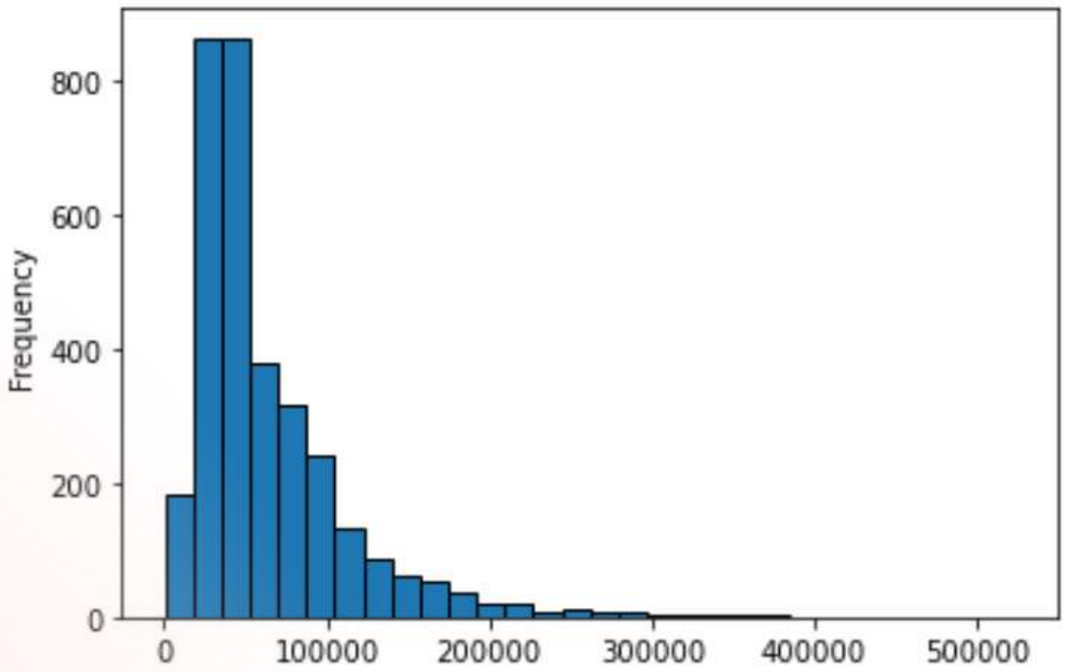
```
[9]  # merge ไฟล์ทั้งสอง
     train_merge = pd.merge(train_variants, train_text, on='ID')
     train_merge.head()
```

	ID	Gene	Variation	Class	TEXT
0	0	FAM58A	Truncating Mutations	1	Cyclin-dependent kinases (CDKs) regulate a var...
1	1	CBL	W802*	2	Abstract Background Non-small cell lung canc...
2	2	CBL	Q249E	2	Abstract Background Non-small cell lung canc...
3	3	CBL	N454D	3	Recent evidence has demonstrated that acquired...
4	4	CBL	L399V	4	Oncogenic mutations in the monomeric Casitas B...

01 Data Understanding: Exploratory Data Analysis (EDA)

- **ความยาวของ Text ในข้อมูล**

```
# เพิ่ม col length
train_merge['length'] = train_merge['TEXT'].str.len()
```



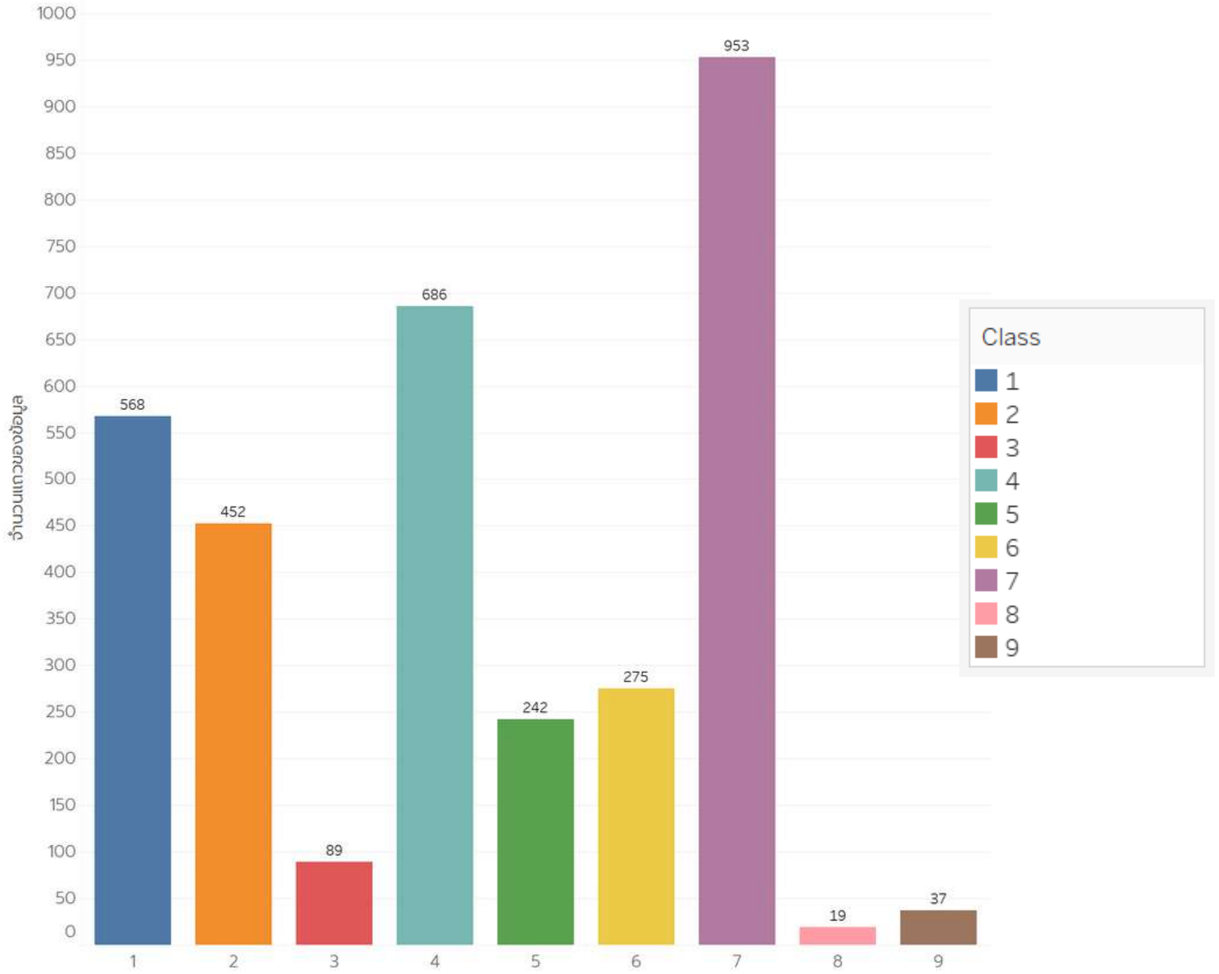
01 Data Understanding: Exploratory Data Analysis (EDA)

ตัวอย่างข้อมูลจาก column TEXT เพียง 1 row โดยมีความยาว (str.len()) อยู่ที่ 39672

"Cyclin-dependent kinases (CDKs) regulate a variety of fundamental cellular processes. CDK10 stands out as one of the last orphan CDKs for which no activating cyclin has been identified and no kinase activity revealed. Previous work has shown that CDK10 silencing increases ETS2 (v-ets erythroblastosis virus E26 oncogene homolog 2)-driven activation of the MAPK pathway, which confers tamoxifen resistance to breast cancer cells. The precise mechanisms by which CDK10 modulates ETS2 activity, and more generally the functions of CDK10, remain elusive. Here we demonstrate that CDK10 is a cyclin-dependent kinase by identifying cyclin M as an activating cyclin. Cyclin M, an orphan cyclin, is the product of FAM58A, whose mutations cause STAR syndrome, a human developmental anomaly whose features include toe syndactyly, telecanthus, and anogenital and renal malformations. We show that STAR syndrome-associated cyclin M mutants are unable to interact with CDK10. Cyclin M silencing phenocopies CDK10 silencing in increasing c-Raf and in conferring tamoxifen resistance to breast cancer cells. CDK10/cyclin M phosphorylates ETS2 in vitro, and in cells it positively controls ETS2 degradation by the proteasome. ETS2 protein levels are increased in cells derived from a STAR patient, and this increase is attributable to decreased cyclin M levels. Altogether, our results reveal an additional regulatory mechanism for ETS2, which plays key roles in cancer and development. They also shed light on the molecular mechanisms underlying STAR syndrome.Cyclin-dependent kinases (CDKs) play a pivotal role in the control of a number of fundamental cellular processes (1). The human genome contains 21 genes encoding proteins that can be considered as members of the CDK family owing to their sequence similarity with bona fide CDKs, those known to be activated by cyclins (2). Although discovered almost 20 y ago (3, 4), CDK10 remains one of the two CDKs without an identified cyclin partner. This knowledge gap has largely impeded the exploration of its biological functions. CDK10 can act as a positive cell cycle regulator in some cells (5, 6) or as a tumor suppressor in others (7, 8). CDK10 interacts with the ETS2 (v-ets erythroblastosis virus E26 oncogene homolog 2) transcription factor and inhibits its transcriptional activity through an unknown mechanism (9). CDK10 knockdown derepresses ETS2, which increases the expression of the c-Raf protein kinase, activates the MAPK pathway, and induces resistance of MCF7 cells to tamoxifen (6).Here, we deorphanize CDK10 by identifying cyclin M, the product of FAM58A, as a binding partner. Mutations in this gene that predict absence or truncation of cyclin M are associated with STAR syndrome, whose features include toe syndactyly, telecanthus, and anogenital and renal malformations in heterozygous females (10). However, both the functions of cyclin M and the pathogenesis of STAR syndrome remain unknown. We show that a recombinant CDK10/cyclin M heterodimer is an active protein kinase that phosphorylates ETS2 in vitro. Cyclin M silencing phenocopies CDK10 silencing in increasing c-Raf and phospho-ERK expression levels and in inducing tamoxifen resistance in estrogen receptor (ER)+ breast cancer cells. We show that CDK10/cyclin M positively controls ETS2 degradation by the proteasome, through the phosphorylation of two neighboring serines. Finally, we detect an increased ETS2 expression level in cells derived from a STAR patient, and we demonstrate that it is attributable to the decreased cyclin M expression level observed in these cells.Previous SectionNext SectionResultsA yeast two-hybrid (Y2H) screen unveiled an interaction signal between CDK10 and a mouse protein whose C-terminal half presents a strong sequence homology with the human FAM58A gene product [whose proposed name is cyclin M (11)]. We thus performed Y2H mating assays to determine whether human CDK10 interacts with human cyclin M (Fig. 1 A–C). The longest CDK10 isoform (P1) expressed as a bait protein produced a strong interaction phenotype with full-length cyclin M (expressed as a prey protein) but no detectable phenotype with cyclin D1, p21 (CIP1), and Cdi1 (KAP), which are known binding partners of other CDKs (Fig. 1B). CDK1 and CDK3 also produced Y2H signals with cyclin M, albeit notably weaker than that observed with CDK10 (Fig. 1B). An interaction phenotype was also observed between full-length cyclin M and CDK10 proteins expressed as bait and prey, respectively (Fig. S1A). We then tested different isoforms of CDK10 and cyclin M originating from alternative gene splicing, and two truncated cyclin M proteins corresponding to the hypothetical products of two mutated FAM58A genes found in STAR syndrome patients (10). None of these shorter isoforms produced interaction phenotypes (Fig. 1 A and C and Fig. S1A).Fig. 1.In a new window Download PPTFig. 1.CDK10 and cyclin M form an interaction complex. (A) Schematic representation of the different protein isoforms analyzed by Y2H assays. Amino acid numbers are indicated. Black boxes indicate internal deletions. The red box indicates a differing amino acid sequence compared with CDK10 P1. (B) Y2H assay between a set of CDK proteins expressed as baits (in fusion to the LexA DNA binding domain) and CDK interacting proteins expressed as preys (in fusion to the B42 transcriptional activator). pEG202 and pJG4-5 are the empty bait and prey plasmids expressing LexA and B42, respectively. lacZ was used as a reporter gene, and blue yeast are indicative of a Y2H interaction phenotype. (C) Y2H assay between the different CDK10 and cyclin M isoforms. The amino-terminal region of ETS2, known to interact with CDK10 (9), was also assayed. (D) Western blot analysis of Myc-CDK10 (wt or kd) and CycM-V5-6His expression levels in transfected HEK293 cells. (E) Western blot analysis of Myc-CDK10 (wt or kd) immunoprecipitates obtained using the anti-Myc antibody. "Inputs" correspond to 10 μg total lysates obtained from HEK293 cells coexpressing Myc-CDK10 (wt or kd) and CycM-V5-6His. (F) Western blot analysis of immunoprecipitates obtained using the anti-CDK10 antibody or a control goat antibody, from human breast cancer MCF7 cells. "Input" corresponds to 30 μg MCF7 total cell lysates. The lower band of the doublet observed on the upper panel comigrates with the exogenously expressed untagged CDK10 and thus corresponds to endogenous CDK10. The upper band of the doublet corresponds to a nonspecific signal, as demonstrated by it insensitivity to either overexpression of CDK10 (as seen on the left lane) or silencing of CDK10 (Fig. S2B). Another experiment with a longer gel migration is shown in Fig. S1D.Next we examined the ability of CDK10 and cyclin M to interact when expressed in human cells (Fig. 1 D and E). We tested wild-type CDK10 (wt) and a kinase dead (kd) mutant bearing a D181A amino acid substitution that abolishes ATP binding (12). We expressed cyclin M-V5-6His and/or Myc-CDK10 (wt or kd) in a human embryonic kidney cell line (HEK293). The expression level of cyclin M-V5-6His was significantly increased upon coexpression with Myc-CDK10 (wt or kd), and to a lesser extent, that of Myc-CDK10 (wt or kd) was increased upon coexpression with cyclin M-V5-6His (Fig. 1D). We then immunoprecipitated Myc-CDK10 proteins and detected the presence of cyclin M in the CDK10 (wt) and (kd) immunoprecipitates only when these proteins were coexpressed pair-wise (Fig. 1E). We confirmed these observations by detecting the presence of Myc-CDK10 in cyclin M-V5-6His immunoprecipitates (Fig. S1B). These experiments confirmed the lack of robust interaction between the CDK10.P2 isoform and cyclin M (Fig. S1C). To detect the interaction between endogenous proteins, we performed immunoprecipitations on nontransfected MCF7 cells derived from a human breast cancer. CDK10 and cyclin M antibodies detected their cognate endogenous proteins by Western blotting. We readily detected cyclin M in immunoprecipitates obtained with the CDK10 antibody but not with a control antibody (Fig. 1F). These results confirm the physical interaction between CDK10 and cyclin M in human cells.To unveil a hypothesized CDK10/cyclin M protein kinase activity, we produced GST-CDK10 and Strep‖-cyclin M fusion proteins in insect cells, either individually or in combination. We observed that GST-CDK10 and Strep‖-cyclin M copurified, thus confirming their interaction in yet another cellular model (Fig. 2A). We then performed in vitro kinase assays with purified proteins, using histone H1 as a generic substrate. Histone H1 phosphorylation was detected only from lysates of cells coexpressing GST-CDK10 and Strep‖-cyclin M. No phosphorylation was detected when GST-CDK10 or Strep‖-cyclin M were expressed alone, or when Strep‖-cyclin M was coexpressed with GST-CDK10(kd) (Fig. 2A). Next we investigated whether ETS2, which is known to interact with CDK10 (9) (Fig. 1C), is a phosphorylation substrate of CDK10/cyclin M. We detected strong phosphorylation of ETS2 by the GST-CDK10/Strep‖-cyclin M purified heterodimer, whereas no phosphorylation was detected using GST-CDK10 alone or GST-CDK10(kd)/Strep‖-cyclin M heterodimer (Fig. 2B).Fig. 2.In a new window Download PPTFig. 2.CDK10 is a cyclin M-dependent protein kinase. (A) In vitro protein kinase assay on histone H1. Lysates from insect cells expressing different proteins were purified on a glutathione Sepharose matrix to capture GST-CDK10(wt or kd) fusion proteins alone, or in complex with STR-CycM fusion protein. Purified protein expression levels were analyzed by Western blots (Top and Upper Middle). The kinase activity was determined by autoradiography of histone H1, whose added amounts were visualized by Coomassie staining (Lower Middle and Bottom). (B) Same as in A, using purified recombinant 6His-ETS2 as a substrate.CDK10 silencing has been shown to increase ETS2-driven c-RAF transcription and to activate the MAPK pathway (6). We investigated whether cyclin M is also involved in this regulatory pathway. To aim at a highly specific silencing, we used siRNA pools (mix of four different siRNAs) at low final concentration (10 nM). Both CDK10 and cyclin M siRNA pools silenced the expression of their cognate targets (Fig. 3 A and C and Fig. S2) and, interestingly, the cyclin M siRNA pool also caused a marked decrease in CDK10 protein level (Fig. 3A and Fig. S2B). These results, and those shown in Fig. 1D, suggest that cyclin M binding stabilizes CDK10. Cyclin M silencing induced an increase in c-Raf protein and mRNA levels (Fig. 3 B and C) and in phosphorylated ERK1 and ERK2 protein levels (Fig. S3B), similarly to CDK10 silencing. As expected from these effects (6), CDK10 and cyclin M silencing both decreased the sensitivity of ER+ MCF7 cells to tamoxifen, to a similar extent. The combined silencing of both genes did not result in a higher resistance to the drug (Fig. S3C). Altogether, these observations demonstrate a functional interaction between cyclin M and CDK10, which negatively controls ETS2.Fig. 3.In a new window Download PPTFig. 3.Cyclin M silencing up-regulates c-Raf expression. (A) Western blot analysis of endogenous CDK10 and cyclin M expression levels in MCF7 cells, in response to siRNA-mediated gene silencing. (B) Western blot analysis of endogenous c-Raf expression levels in MCF7 cells, in response to CDK10 or cyclin M silencing. A quantification is shown in Fig. S3A. (C) Quantitative RT-PCR analysis of CDK10, cyclin M, and c-Raf mRNA levels, in response to CDK10 (Upper) or cyclin M (Lower) silencing. ** ≤ 0.01; *** ≤ 0.001.We then wished to explore the mechanism by which CDK10/cyclin M controls ETS2. ETS2 is a short-lived protein degraded by the proteasome (13). A straightforward hypothesis is that CDK10/cyclin M positively controls ETS2 degradation. We thus examined the impact of CDK10 or cyclin M silencing on ETS2 expression levels. The silencing of CDK10 and that of cyclin M caused an increase in the expression levels of an exogenously expressed Flag-ETS2 protein (Fig. 4A). This increase is not attributable to increased ETS2 mRNA levels, which marginally fluctuated in response to CDK10 or cyclin M silencing (Fig. S4B). We then examined the expression levels of the Flag-tagged ETS2 protein when expressed alone or in combination with Myc-CDK10 or -CDK10(kd), with or without cyclin M-V5-6His. Flag-ETS2 was readily detected when expressed alone or, to a lesser extent, when coexpressed with CDK10(kd). However, its expression level was dramatically decreased when coexpressed with CDK10 alone, or with CDK10 and cyclin M (Fig. 4B). These observations suggest that endogenous cyclin M levels are in excess compared with those of CDK10 in MCF7 cells, and they show that the major decrease in ETS2 levels observed upon CDK10 coexpression involves CDK10 kinase activity. Treatment of cells coexpressing Flag-ETS2, CDK10, and cyclin M with the proteasome inhibitor MG132 largely rescued Flag-ETS2 expression levels (Fig. 4B).Fig. 4.In a new window Download PPTFig. 4.CDK10/cyclin M controls ETS2 stability in human cancer derived cells. (A) Western blot analysis of endogenous ETS2 expression levels in MCF7 cells, in response to siRNA-mediated CDK10 and/or cyclin M silencing. A quantification is shown in Fig. S4B. (B) Western blot analysis of exogenously expressed Flag-ETS2 protein levels in MCF7 cells cotransfected with empty vectors or coexpressing Myc-CDK10 (wt or kd), or Myc-CDK10/CycM-V5-6His. The latter cells were treated for 16 h with the MG132 proteasome inhibitor. Proper expression of CDK10 and cyclin M tagged proteins was verified by Western blot analysis. (C and D) Western blot analysis of expressions levels of exogenously expressed Flag-ETS2 wild-type or mutant proteins in MCF7 cells, in the absence of (C) or in response to (D) Myc-CDK10/CycM-V5-6His expression. Quantifications are shown in Fig. S4 C and D.A mass spectrometry analysis of recombinant ETS2 phosphorylated by CDK10/cyclin M in vitro revealed the presence of multiple phosphorylated residues, among which are two neighboring phospho-serines (at positions 220 and 225) that may form a phosphodegron (14) (Figs. S5–S8). To confirm this finding, we compared the phosphorylation level of recombinant ETS2wt with that of ETS2SASA protein, a mutant bearing alanine substitutions of these two serines. As expected from the existence of multiple phosphorylation sites, we detected a small but reproducible, significant decrease of phosphorylation level of ETS2SASA compared with ETS2wt (Fig. S9), thus confirming that Ser220/Ser225 are phosphorylated by CDK10/cyclin M. To establish a direct link between ETS2 phosphorylation by CDK10/cyclin M and degradation, we examined the expression levels of Flag-ETS2SASA. In the absence of CDK10/cyclin M coexpression, it did not differ significantly from that of Flag-ETS2. This is contrary to that of Flag-ETS2DBM, bearing a deletion of the N-terminal destruction (D-) box that was previously shown to be involved in APC-Cdh1–mediated degradation of ETS2 (13) (Fig. 4C). However, contrary to Flag-ETS2 wild type, the expression level of Flag-ETS2SASA remained insensitive to CDK10/cyclin M coexpression (Fig. 4D). Altogether, these results suggest that CDK10/cyclin M directly controls ETS2 degradation through the phosphorylation of these two serines.Finally, we studied a lymphoblastoid cell line derived from a patient with STAR syndrome, bearing FAM58A mutation c.555+1G>A, predicted to result in aberrant splicing (10). In accordance with incomplete skewing of X chromosome inactivation previously found in this patient, we detected a decreased expression level of cyclin M protein in the STAR cell line, compared with a control lymphoblastoid cell line. In line with our preceding observations, we detected an increased expression level of ETS2 protein in the STAR cell line compared with the control (Fig. 5A and Fig. S10A). We then examined by quantitative RT-PCR the mRNA expression levels of the corresponding genes. The STAR cell line showed a decreased expression level of cyclin M mRNA but an expression level of ETS2 mRNA similar to that of the control cell line (Fig. 5B). To demonstrate that the increase in ETS2 protein expression is indeed a result of the decreased cyclin M expression observed in the STAR patient-derived cell line, we expressed cyclin M-V5-6His in this cell line. This expression caused a decrease in ETS2 protein levels (Fig. 5C).Fig. 5.In a new window Download PPTFig. 5.Decreased cyclin M expression in STAR patient-derived cells results in increased ETS2 protein level. (A) Western blot analysis of cyclin M and ETS2 protein levels in a STAR patient-derived lymphoblastoid cell line and in a control lymphoblastoid cell line, derived from a healthy individual. A quantification is shown in Fig. S10A. (B) Quantitative RT-PCR analysis of cyclin M and ETS2 mRNA levels in the same cells. ***P ≤ 0.001. (C) Western blot analysis of ETS2 protein levels in the STAR patient-derived lymphoblastoid cell line transfected with an empty vector or a vector directing the expression of cyclin M-V5-6His. Another Western blot revealing endogenously and exogenously expressed cyclin M levels is shown in Fig. S10B. A quantification of ETS2 protein levels is shown in Fig. S10C.Previous SectionNext SectionDiscussionIn this work, we unveil the interaction between CDK10, the last orphan CDK discovered in the pregenomic era (2), and cyclin M, the only cyclin associated with a human genetic disease so far, and whose functions remain unknown (10). The closest paralogs of CDK10 within the CDK family are the CDK11 proteins, which interact with L-type cyclins (15). Interestingly, the closest paralog of these cyclins within the cyclin family is cyclin M (Fig. S11). The fact that none of the shorter CDK10 isoforms interact robustly with cyclin M suggests that alternative splicing of the CDK10 gene (16, 17) plays an important role in regulating CDK10 functions.The functional relevance of the interaction between CDK10 and cyclin M is supported by different observations. Both proteins seem to enhance each other's stability, as judged from their increased expression levels when their partner is exogenously coexpressed (Fig. 1D) and from the much reduced endogenous CDK10 expression level observed in response to cyclin M silencing (Fig. 3A and Fig. S2B). CDK10 is subject to ubiquitin-mediated degradation (18). Our observations suggest that cyclin M protects CDK10 from such degradation and that it is the only cyclin partner of CDK10, at least in MCF7 cells. They also suggest that cyclin M stability is enhanced upon binding to CDK10, independently from its kinase activity, as seen for cyclin C and CDK8 (19). We uncover a cyclin M-dependent CDK10 protein kinase activity in vitro, thus demonstrating that this protein, which was named a CDK on the sole basis of its amino acid sequence, is indeed a genuine cyclin-dependent kinase. Our Y2H assays reveal that truncated cyclin M proteins corresponding to the hypothetical products of two STAR syndrome-associated FAM58A mutations do not produce an interaction phenotype with CDK10. Hence, regardless of whether these mutated mRNAs undergo nonsense-mediated decay (as suggested from the decreased cyclin M mRNA levels in STAR cells, shown in Fig. 5B) or give rise to truncated cyclin M proteins, females affected by the STAR syndrome must exhibit compromised CDK10/cyclin M kinase activity at least in some tissues and during specific developmental stages.We show that ETS2, a known interactor of CDK10, is a phosphorylation substrate of CDK10/cyclin M in vitro and that CDK10/cyclin M kinase activity positively controls ETS2 degradation by the proteasome. This control seems to be exerted through a very fine mechanism, as judged from the sensitivity of ETS2 levels to partially decreased CDK10 and cyclin M levels, achieved in MCF7 cells and observed in STAR cells, respectively. These findings offer a straightforward explanation for the already reported up-regulation of ETS2-driven transcription of c-RAF in response to CDK10 silencing (6). We bring evidence that CDK10/cyclin M directly controls ETS2 degradation through the phosphorylation of two neighboring serines, which may form a noncanonical β-TRCP phosphodegron (DSMCPAS) (14). Because none of these two serines precede a proline, they do not conform to usual CDK phosphorylation sites. However, multiple so-called transcriptional CDKs (CDK7, -8, -9, and -11) (to which CDK10 may belong; Fig. S11) have been shown to phosphorylate a variety of motifs in a non–proline-directed fashion, especially in the context of molecular docking with the substrate (20). Here, it can be hypothesized that the high-affinity interaction between CDK10 and the Pointed domain of ETS2 (6, 9) (Fig. 1C) would allow docking-mediated phosphorylation of atypical sites. The control of ETS2 degradation involves a number of players, including APC-Cdh1 (13) and the cullin-RING ligase CRL4 (21). The formal identification of the ubiquitin ligase involved in the CDK10/cyclin M pathway and the elucidation of its concerted action with the other ubiquitin ligases to regulate ETS2 degradation will require further studies.Our results present a number of significant biological and medical implications. First, they shed light on the regulation of ETS2, which plays an important role in development (22) and is frequently deregulated in many cancers (23). Second, our results contribute to the understanding of the molecular mechanisms causing tamoxifen resistance associated with reduced CDK10 expression levels, and they suggest that, like CDK10 (6), cyclin M could also be a predictive clinical marker of hormone therapy response of ERα-positive breast cancer patients. Third, our findings offer an interesting hypothesis on the molecular mechanisms underlying STAR syndrome. Ets2 transgenic mice showing a less than twofold overexpression of Ets2 present severe cranial abnormalities (24), and those observed in STAR patients could thus be caused at least in part by increased ETS2 protein levels. Another expected consequence of enhanced ETS2 expression levels would be a decreased risk to develop certain types of cancers and an increased risk to develop others. Studies on various mouse models (including models of Down syndrome, in which three copies of ETS2 exist) have revealed that ETS2 dosage can repress or promote tumor growth and, hence, that ETS2 exerts noncell autonomous functions in cancer (25). Intriguingly, one of the very few STAR patients identified so far has been diagnosed with a nephroblastoma (26). Finally, our findings will facilitate the general exploration of the biological functions of CDK10 and, in particular, its role in the control of cell division. Previous studies have suggested either a positive role in cell cycle control (5, 6) or a tumor-suppressive activity in some cancers (7, 8). The severe growth retardation exhibited by STAR patients strongly suggests that CDK10/cyclin M plays an important role in the control of cell proliferation.Previous SectionNext SectionMaterials and MethodsCloning of CDK10 and cyclin M cDNAs, plasmid constructions, tamoxifen response analysis, quantitative RT-PCR, mass spectrometry experiments, and antibody production are detailed in SI Materials and Methods.Yeast Two-Hybrid Interaction Assays. We performed yeast interaction mating assays as previously described (27).Mammalian Cell Cultures and Transfections. We grew human HEK293 and MCF7 cells in DMEM supplemented with 10% (vol/vol) FBS (Invitrogen), and we grew lymphoblastoid cells in RPMI 1640 GlutaMAX supplemented with 15% (vol/vol) FBS. We transfected HEK293 and MCF7 cells using Lipofectamine 2000 (Invitrogen) for siRNAs, and Jetprime (Polyplus) for plasmids/siRNAs combinations according to the manufacturers' instructions. We transfected lymphoblastoid cells by electroporation (Neon, Invitrogen). For ETS2 stability studies we treated MCF7 cells 32 h after transfection with 10 μM MG132 (Fisher Scientific) for 16 h.Coimmunoprecipitation and Western Blot Experiments. We collected cells by scraping in PBS (or centrifugation for lymphoblastoid cells) and lysed them by sonication in a lysis buffer containing 60 mM β-glycerophosphate, 15 mM p-nitrophenylphosphate, 25 mM 3-(N-morpholino)propanesulfonic acid (Mops) (pH 7.2), 15 mM EGTA, 15 mM MgCl2, 1 mM Na vanadate, 1 mM NaF, 1mM phenylphosphate, 0.1% Nonidet P-40, and a protease inhibitor mixture (Roche). We spun the lysates 15 min at 20,000 × g at 4 °C, collected the supernatants, and determined the protein content using a Bradford assay. We performed the immunoprecipitation experiments on 500 μg of total proteins, in lysis buffer. We precleared the lysates with 20 μL of protein A or G-agarose beads, incubated 1 h 4 °C on a rotating wheel. We added 5 μg of antibody to the supernatants, incubated 1 h 4 °C on a rotating wheel, added 20 μL of protein A or G-agarose beads, and incubated 1 h 4 °C on a rotating wheel. We collected the beads by centrifugation 30 s at 18,000 × g at 4 °C and washed three times in a bead buffer containing 50 mM Tris (pH 7.4), 5 mM NaF, 250 mM NaCl, 5 mM EDTA, 5 mM EGTA, 0.1% Nonidet P-40, and a protease inhibitor coktail (Roche). We directly added sample buffer to the washed pellets, heat-denatured the proteins, and ran the samples on 10% Bis-Tris SDS/PAGE. We transferred the proteins onto Hybond nitrocellulose membranes and processed the blots according to standard procedures. For Western blot experiments, we used the following primary antibodies: anti-Myc (Abcam ab9106, 1:2,000), anti-V5 (Invitrogen R960, 1:5,000), anti-tubulin (Santa Cruz Biotechnology B-7, 1:500), anti-CDK10 (Covalab pab0847p, 1:500 or Santa Cruz Biotechnology C-19, 1:500), anti-CycM (home-made, dilution 1:500 or Covalab pab0882-P, dilution 1:500), anti-Raf1 (Santa Cruz Biotechnology C-20, 1:1,000), anti-ETS2 (Santa Cruz Biotechnology C-20, 1:1,000), anti-Flag (Sigma F7425, 1:1,000), and anti-actin (Sigma A5060, 1:5,000). We used HRP-coupled anti-goat (Santa Cruz Biotechnology SC-2033, dilution 1:2,000), anti-mouse (Bio-Rad 170–6516, dilution 1:3,000) or anti-rabbit (Bio-Rad 172–1019, 1:5,000) as secondary antibodies. We revealed the blots by enhanced chemiluminescence (SuperSignal West Femto, Thermo Scientific).Production and Purification of Recombinant Proteins.GST-CDK10(kd)/Strep‖-CycM. We generated recombinant bacmids in DH10Bac Escherichia coli and baculoviruses in Sf9 cells using the Bac-to-Bac system, as described by the provider (Invitrogen). We infected Sf9 cells with GST-CDK10- (or GST-CDK10kd)-producing viruses, or coinfected the cells with Strep‖-CycM–producing viruses, and we collected the cells 72 h after infection. To purify GST-fusion proteins, we spun 250 mL cells and resuspended the pellet in 40 mL lysis buffer (PBS, 250 mM NaCl, 0.5% Nonidet P-40, 50 mM NaF, 10 mM β-glycerophosphate, and 0.3 mM Na-vanadate) containing a protease inhibitor mixture (Roche). We lysed the cells by sonication, spun the lysate 30 min at 15,000 × g, collected the soluble fraction, and added it to a 1-mL glutathione-Sepharose matrix. We incubated 1 h at 4 °C, washed four times with lysis buffer, one time with kinase buffer A (see below), and finally resuspended the beads in 100 μL kinase buffer A containing 10% (vol/vol) glycerol for storage.6His-ETS2. We transformed Origami2 DE3 (Novagen) with the 6His-ETS2 expression vector. We induced expression with 0.2 mM isopropyl-β-d-1-thiogalactopyranoside for 3 h at 22 °C. To purify 6His-ETS2, we spun 50 mL cells and resuspended the pellet in 2 mL lysis buffer (PBS, 300 mM NaCl, 10 mM Imidazole, 1 mM DTT, and 0.1% Nonidet P-40) containing a protease inhibitor mixture without EDTA (Roche). We lysed the cells at 1.6 bar using a cell disruptor and spun the lysate 10 min at 20,000 × g. We collected the soluble fraction and added 0.2 mL Cobalt beads (Interchim Scientific). After 1 h incubation at 4 °C on a rotating wheel, we washed four times with lysis buffer. To elute, we incubated 10 min at 4 °C with elution buffer (PBS, 250 mM imidazole, pH 7.6) containing the protease inhibitor mixture, spun 30 s at 10,000 × g, collected the eluted protein, and stored it at −80 °C.Protein Kinase Assays. We mixed glutathione-Sepharose beads harboring GST-CDK10 or GST-CDK10(kd) complexed with Strep‖-CycM, 22.7 μM BSA, 15 mM DTT, 100 μM ATP, 5 μCi ATP[γ-32P], 7.75 μM histone H1, or 1 μM 6His-ETS2 and added kinase buffer A (25 mM Tris-HCl, 10 mM MgCl2, 1 mM EGTA, 1 mM DTT, and 3.7 μM heparin, pH 7.5) up to a total volume of 30 μL. We incubated the reactions 30 min at 30 °C, added Laemli sample buffer, heat-denatured the samples, and ran 10% Bis-Tris SDS/PAGE. We cut gel slices to detect GST-CDK10 and Strep‖-CycM by Western blotting. We stained the gel slices containing the substrate with Coomassie (R-250, Bio-Rad), dried them, and detected the incorporated radioactivity by autoradiography. We identified four unrelated girls with anogenital and renal malformations, dysmorphic facial features, normal intellect and syndactyly of toes. A similar combination of features had been reported previously in a mother–daughter pair1 (Table 1 and Supplementary Note online). These authors noted clinical overlap with Townes-Brocks syndrome but suggested that the phenotype represented a separate autosomal dominant entity (MIM601446). Here we define the cardinal features of this syndrome as a characteristic facial appearance with apparent telecanthus and broad tripartite nasal tip, variable syndactyly of toes 2–5, hypoplastic labia, anal atresia and urogenital malformations (Fig. 1a–h). We also observed a variety of other features (Table 1). Figure 1: Clinical and molecular characterization of STAR syndrome. Figure 1 : Clinical and molecular characterization of STAR syndrome. (a–f) Facial appearances of cases 1–3 (apparent telecanthus, dysplastic ears and thin upper lips; a,c,e), and toe syndactyly 2–5, 3–5 or 4–5 (b,d,f) in these cases illustrate recognizable features of STAR syndrome (specific parental consent has been obtained for publication of these photographs). Anal atresia and hypoplastic labia are not shown. (g,h) X-ray films of the feet of case 2 showing only four rays on the left and delta-shaped 4th and 5th metatarsals on the right (h; compare to clinical picture in d). (i) Array-CGH data. Log2 ratio represents copy number loss of six probes spanning between 37.9 and 50.7 kb, with one probe positioned within FAM58A. The deletion does not remove parts of other functional genes. Schematic structure of FAM58A and position of the mutations. FAM58A has five coding exons (boxes). The cyclin box domain (green) is encoded by exons 2–4. The horizontal arrow indicates the deletion extending 5′ in case 1, which includes exon 1 and, whereas the horizontal line below exon 5 indicates the deletion found in case 3, which removes exon 5 and some 3′-sequence. The pink horizontal bars above the boxes indicate the amplicons used for qPCR and sequencing (one alternative exon 5 amplicon is not indicated because of space constraints). The mutation 201dupT (case 4) results in an immediate stop codon, and the 555+1G>A and 555-1G>A splice mutations in cases 2, 5 and 6 are predicted to be deleterious because they alter the conserved splice donor and acceptor site of intron 4, respectively. Full size image (97 KB) Table 1: Clinical features in STAR syndrome cases Table 1 - Clinical features in STAR syndrome cases Full table On the basis of the phenotypic overlap with Townes-Brocks, Okihiro and Feingold syndromes, we analyzed SALL1 (ref. 2), SALL4 (ref. 3) and MYCN4 but found no mutations in any of these genes (Supplementary Methods online). Next, we carried out genome-wide high-resolution oligonucleotide array comparative genomic hybridization (CGH)5 analysis (Supplementary Methods) of genomic DNA from the most severely affected individual (case 1, with lower lid coloboma, epiblepsy and syringomyelia) and identified a heterozygous deletion of 37.9–50.7 kb on Xq28, which removed exons 1 and 2 of FAM58A (Fig. 1i,j). Using real-time PCR, we confirmed the deletion in the child and excluded it in her unaffected parents (Supplementary Fig. 1a online, Supplementary Methods and Supplementary Table 1 online). Through CGH with a customized oligonucleotide array enriched in probes for Xq28, followed by breakpoint cloning, we defined the exact deletion size as 40,068 bp (g.152,514,164_152,554,231del(chromosome X, NCBI Build 36.2); Fig. 1j and Supplementary Figs. 2,3 online). The deletion removes the coding regions of exons 1 and 2 as well as intron 1 (2,774 bp), 492 bp of intron 2, and 36,608 bp of 5′ sequence, including the 5′ UTR and the entire KRT18P48 pseudogene (NCBI gene ID 340598). Paternity was proven using routine methods. We did not find deletions overlapping FAM58A in the available copy number variation (CNV) databases. Subsequently, we carried out qPCR analysis of the three other affected individuals (cases 2, 3 and 4) and the mother-daughter pair from the literature (cases 5 and 6). In case 3, we detected a de novo heterozygous deletion of 1.1–10.3 kb overlapping exon 5 (Supplementary Fig. 1b online). Using Xq28-targeted array CGH and breakpoint cloning, we identified a deletion of 4,249 bp (g.152,504,123_152,508,371del(chromosome X, NCBI Build 36.2); Fig. 1j and Supplementary Figs. 2,3), which removed 1,265 bp of intron 4, all of exon 5, including the 3′ UTR, and 2,454 bp of 3′ sequence. We found heterozygous FAM58A point mutations in the remaining cases (Fig. 1j, Supplementary Fig. 2, Supplementary Methods and Supplementary Table 1). In case 2, we identified the mutation 555+1G>A, affecting the splice donor site of intron 4. In case 4, we identified the frameshift mutation 201dupT, which immediately results in a premature stop codon N68Xfs×1. In cases 5 and 6, we detected the mutation 556-1G>A, which alters the splice acceptor site of intron 4. We validated the point mutations and deletions by independent rounds of PCR and sequencing or by qPCR. We confirmed paternity and de novo status of the point mutations and deletions in all sporadic cases. None of the mutations were seen in the DNA of 60 unaffected female controls, and no larger deletions involving FAM58A were found in 93 unrelated array-CGH investigations. By analyzing X-chromosome inactivation (Supplementary Methods and Supplementary Fig. 4 online), we found complete skewing of X inactivation in cases 1 and 3–6 and almost complete skewing in case 2, suggesting that cells carrying the mutation on the active X chromosome have a growth disadvantage during fetal development. Using RT-PCR on RNA from lymphoblastoid cells of case 2 (Supplementary Fig. 2), we did not find any aberrant splice products as additional evidence that the mutated allele is inactivated. Furthermore, FAM58A is subjected to X inactivation6. In cases 1 and 3, the parental origin of the deletions could not be determined, as a result of lack of informative SNPs. Case 5, the mother of case 6, gave birth to two boys, both clinically unaffected (samples not available). We cannot exclude that the condition is lethal in males. No fetal losses were reported from any of the families. The function of FAM58A is unknown. The gene consists of five coding exons, and the 642-bp coding region encodes a protein of 214 amino acids. GenBank lists a mRNA length of 1,257 bp for the reference sequence (NM_152274.2). Expression of the gene (by EST data) was found in 27 of 48 adult tissues including kidney, colon, cervix and uterus, but not heart (NCBI expression viewer, UniGene Hs.496943). Expression was also noted in 24 of 26 listed tumor tissues as well as in embryo and fetus. Genes homologous to FAM58A (NCBI HomoloGene: 13362) are found on the X chromosome in the chimpanzee and the dog. The zebrafish has a similar gene on chromosome 23. However, in the mouse and rat, there are no true homologs. These species have similar but intronless genes on chromosomes 11 (mouse) and 10 (rat), most likely arising from a retrotransposon insertion event. On the murine X chromosome, the flanking genes Atp2b3 and Dusp9 are conserved, but only remnants of the FAM58A sequence can be detected. FAM58A contains a cyclin-box-fold domain, a protein-binding domain found in cyclins with a role in cell cycle and transcription control. No human phenotype resulting from a cyclin gene mutation has yet been reported. Homozygous knockout mice for Ccnd1 (encoding cyclin D1) are viable but small and have reduced lifespan. They also have dystrophic changes of the retina, likely as a result of decreased cell proliferation and degeneration of photoreceptor cells during embryogenesis7, 8. Cyclin D1 colocalizes with SALL4 in the nucleus, and both proteins cooperatively mediate transcriptional repression9. As the phenotype of our cases overlaps considerably with that of Townes-Brocks syndrome caused by SALL1 mutations1, we carried out co-immunoprecipitation to find out if SALL1 or SALL4 would interact with FAM58A in a manner similar to that observed for SALL4 and cyclin D1. We found that FAM58A interacts with SALL1 but not with SALL4 (Supplementary Fig. 5 online), supporting the hypothesis that FAM58A and SALL1 participate in the same developmental pathway. How do FAM58A mutations lead to STAR syndrome? Growth retardation (all cases; Table 1) and retinal abnormalities (three cases) are reminiscent of the reduced body size and retinal anomalies in cyclin D1 knockout mice7, 8. Therefore, a proliferation defect might be partly responsible for STAR syndrome. To address this question, we carried out a knockdown of FAM58A mRNA followed by a proliferation assay. Transfection of HEK293 cells with three different FAM58A-specific RNAi oligonucleotides resulted in a significant reduction of both FAM58A mRNA expression and proliferation of transfected cells (Supplementary Methods and Supplementary Fig. 6 online), supporting the link between FAM58A and cell proliferation. We found that loss-of-function mutations of FAM58A result in a rather homogeneous clinical phenotype. The additional anomalies in case 1 are likely to result from an effect of the 40-kb deletion on expression of a neighboring gene, possibly ATP2B3 or DUSP9. However, we cannot exclude that the homogeneous phenotype results from an ascertainment bias and that FAM58A mutations, including missense changes, could result in a broader spectrum of malformations. The genes causing the overlapping phenotypes of STAR syndrome and Townes-Brocks syndrome seem to act in the same pathway. Of note, MYCN, a gene mutated in Feingold syndrome, is a direct regulator of cyclin D2 (refs. 10,11); thus, it is worth exploring whether the phenotypic similarities between Feingold and STAR syndrome might be explained by direct regulation of FAM58A by MYCN. FAM58A is located approximately 0.56 Mb centromeric to MECP2 on Xq28. Duplications overlapping both MECP2 and FAM58A have been described and are not associated with a clinical phenotype in females12, but no deletions overlapping both MECP2 and FAM58A have been observed to date13. Although other genes between FAM58A and MECP2 have been implicated in brain development, FAM58A and MECP2 are the only genes in this region known to result in X-linked dominant phenotypes; thus, deletion of both genes on the same allele might be lethal in both males and females."

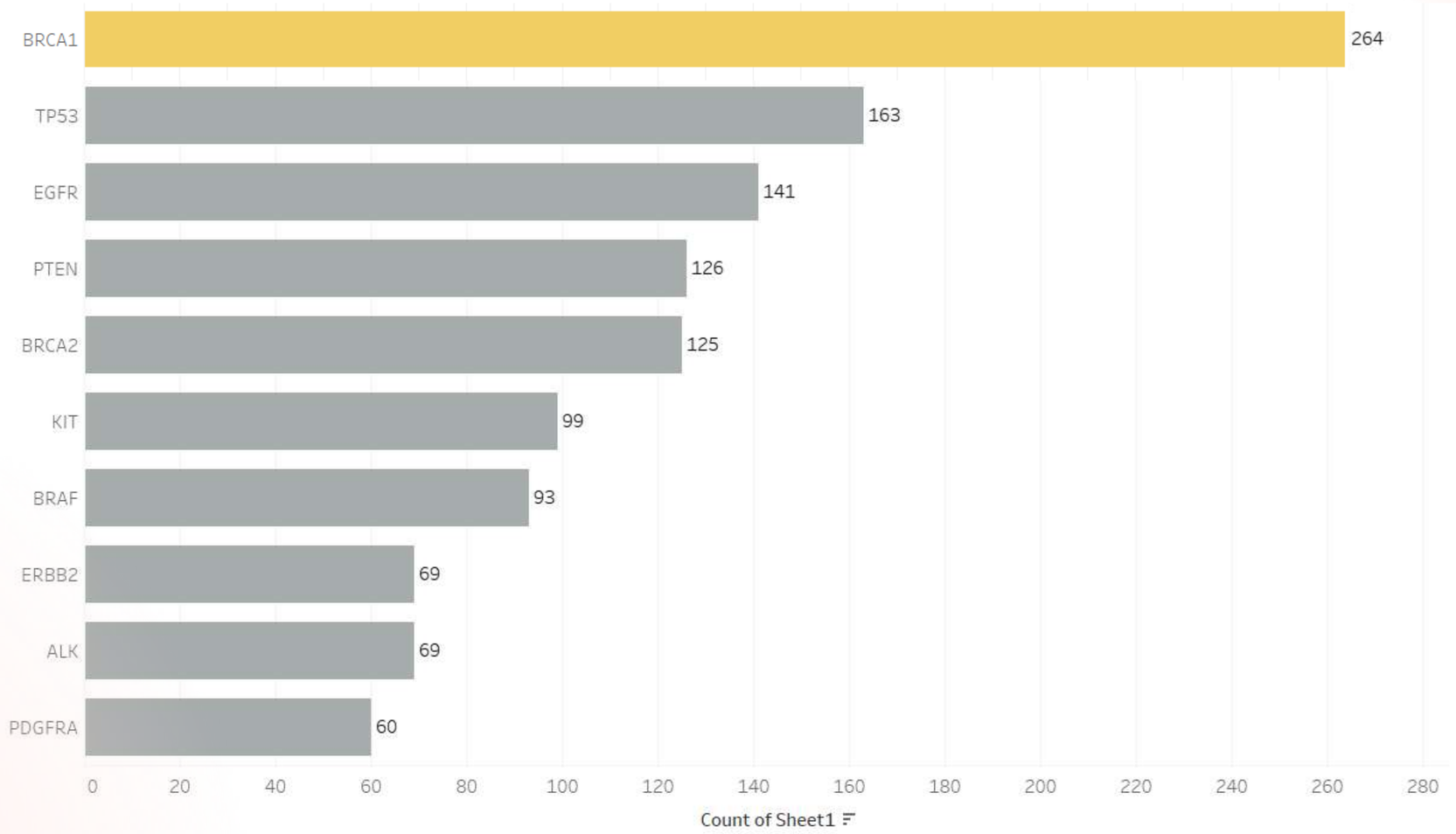
01 Data Understanding: Exploratory Data Analysis (EDA)

- จำนวนแต่ละ Class ในข้อมูล



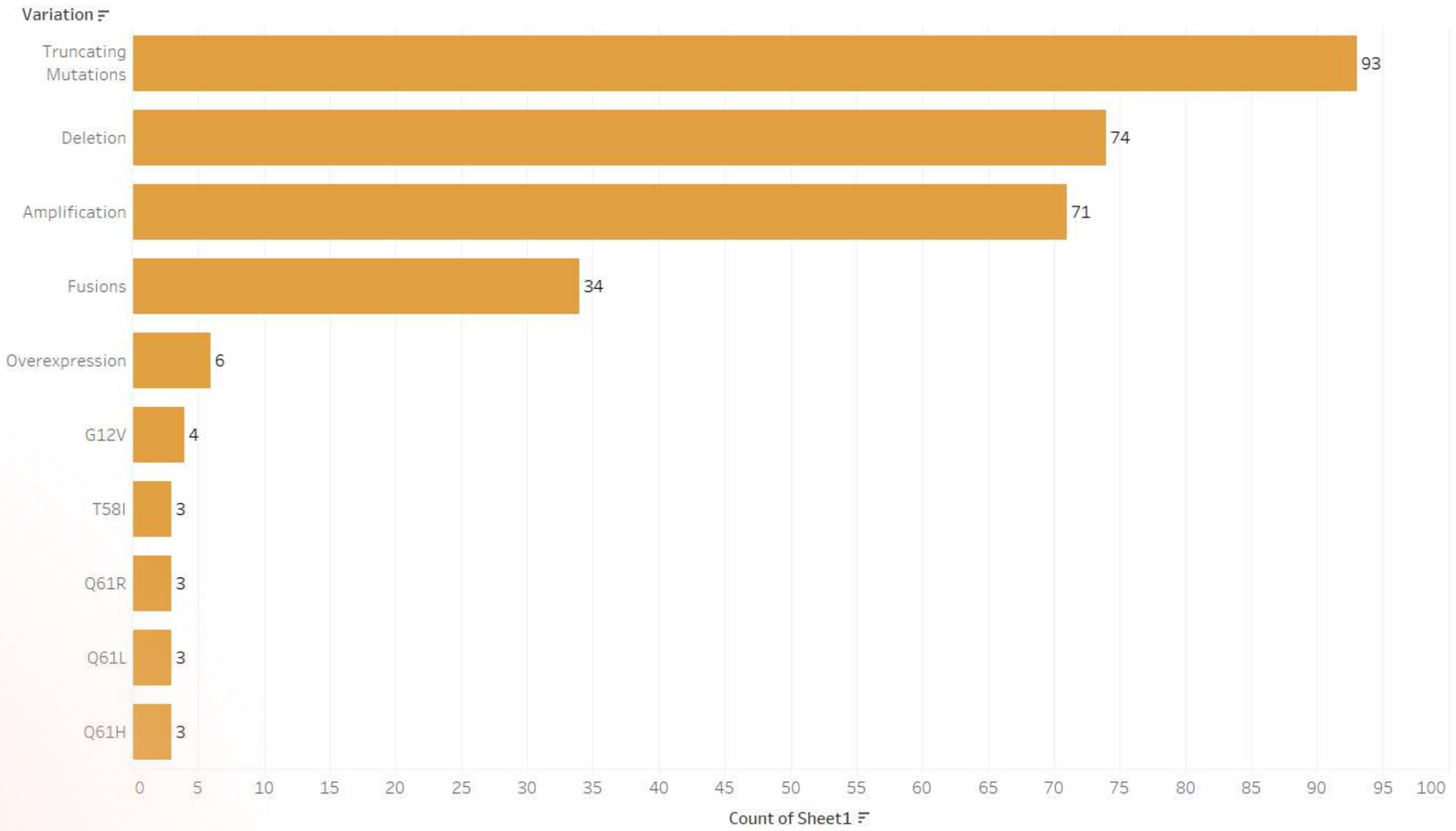
01 Data Understanding: Exploratory Data Analysis (EDA)

- **จำนวน Gene ที่พบมาก 10 อันดับ**



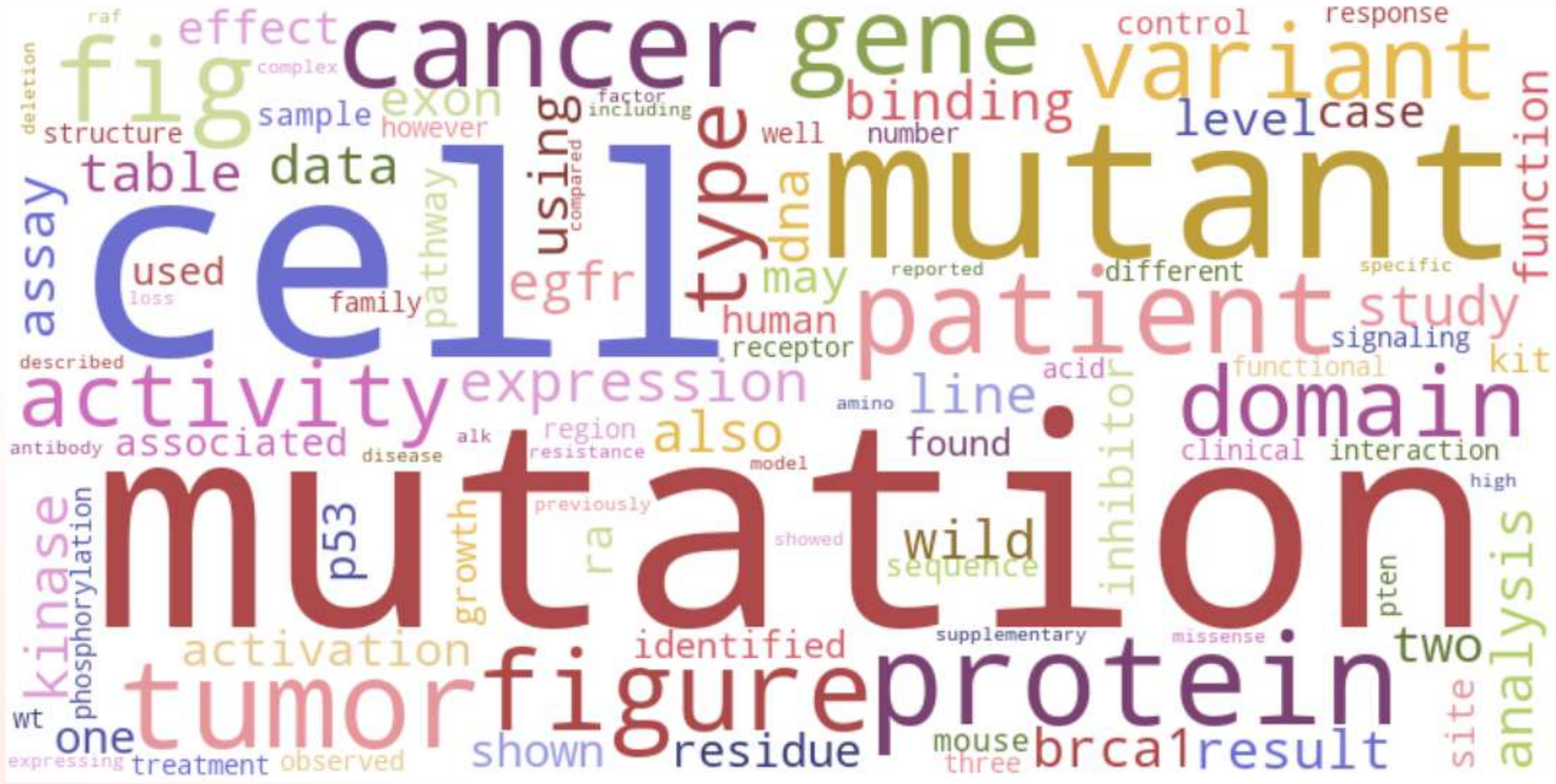
01 Data Understanding: Exploratory Data Analysis (EDA)

- **จำนวน Variation ที่พบมาก 10 อันดับ**



01 Data Understanding: Exploratory Data Analysis (EDA)

- **Wordcloud**

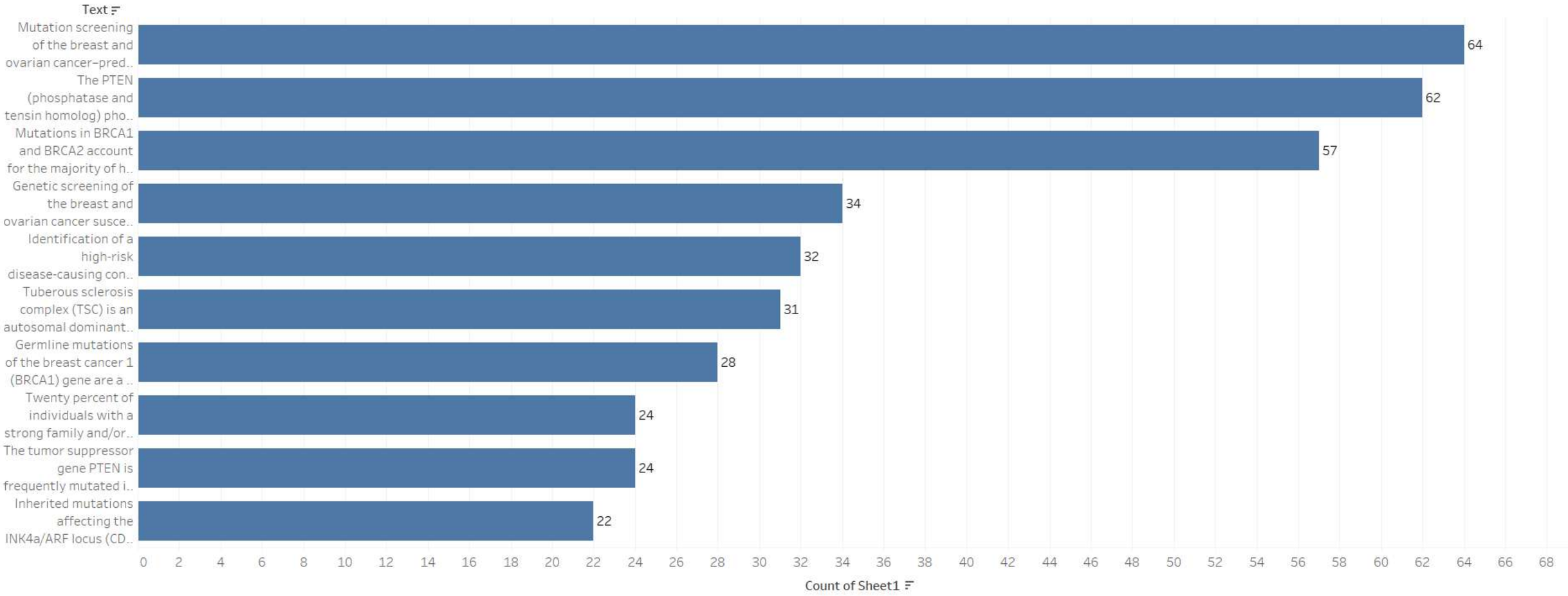


10 อันดับ

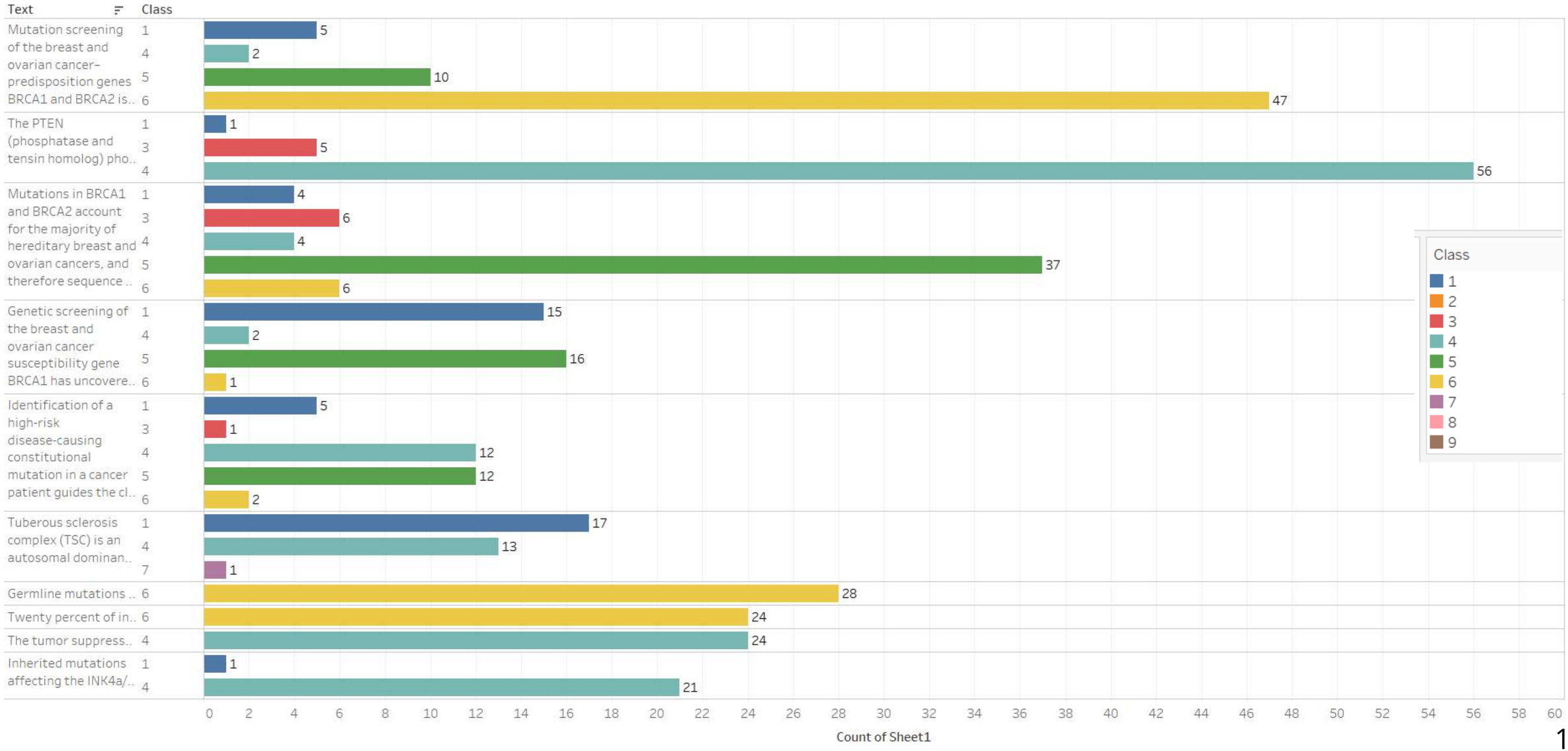
```
1 freq_text.head(10)
```

token	freq
mutation	343839
cell	312489
mutant	126577
protein	125359
tumor	122770
cancer	120094
patient	111373
fig	107416
figure	104767
gene	102268

จำนวน Text ที่พบมาก 10 อันดับ



จำนวน Text ที่พบมาก 10 อันดับ ในแต่ละ Class



01 Data Understanding: Exploratory Data Analysis (EDA)

- มี Text ที่เหมือนกัน แต่อยู่คนละ Class

Text	Class	Gene	Variation
Mutation screening of the breast and ovarian cancer–predisposition genes BRCA1 and BRCA2 is becoming an increasingly important part of clinical practice. Classification of rare nontruncating sequence variants in these genes is problematic, because it is not known whether these subtle changes alter function sufficiently to predispose cells to cancer development. Using data from the Myriad Genetic	1	BRCA2	I2627F
			G2748D
			E2663V
			D3095E
			D2723G
	4	BRCA2	R2659K
			R2336H
	5	BRCA2	V2908G
			S1172L
			R2973C
			P1819S
			K513R
			K2950N
			K2729N
			I2285V
			E462G
			D2665G
	6	BRCA2	Y3098H
			Y3092C
			W2626C
			V894I
			V3079I
			V2969M
The PTEN (phosphatase and tensin homolog) phosphatase is unique in mammals in terms of its tumor suppressor activity, exerted by dephosphorylation of the lipid second messenger PIP3 (phosphatidylinositol 3,4,5-trisphosphate), which activates the phosphoinositide 3-kinase/Akt/mTOR (mammalian target of rapamycin) oncogenic pathway. Loss-of-function mutations in the PTEN gene are frequent in human	1	PTEN	G129A
	3	PTEN	V369G
			T131S
			K125R
			I122V
			A121V
	4	PTEN	Y68D
			Y155C
			V343L
			V217D
			T167A
			T160I
			T131L
			T131I
			T131A
			R161G
			R15K
			R130K
			R130G
			R130A
			P96Q
			P95L
			P169H
			N48K

01 Data Understanding: Exploratory Data Analysis (EDA)

สรุปสิ่งที่ได้จากการทำ EDA

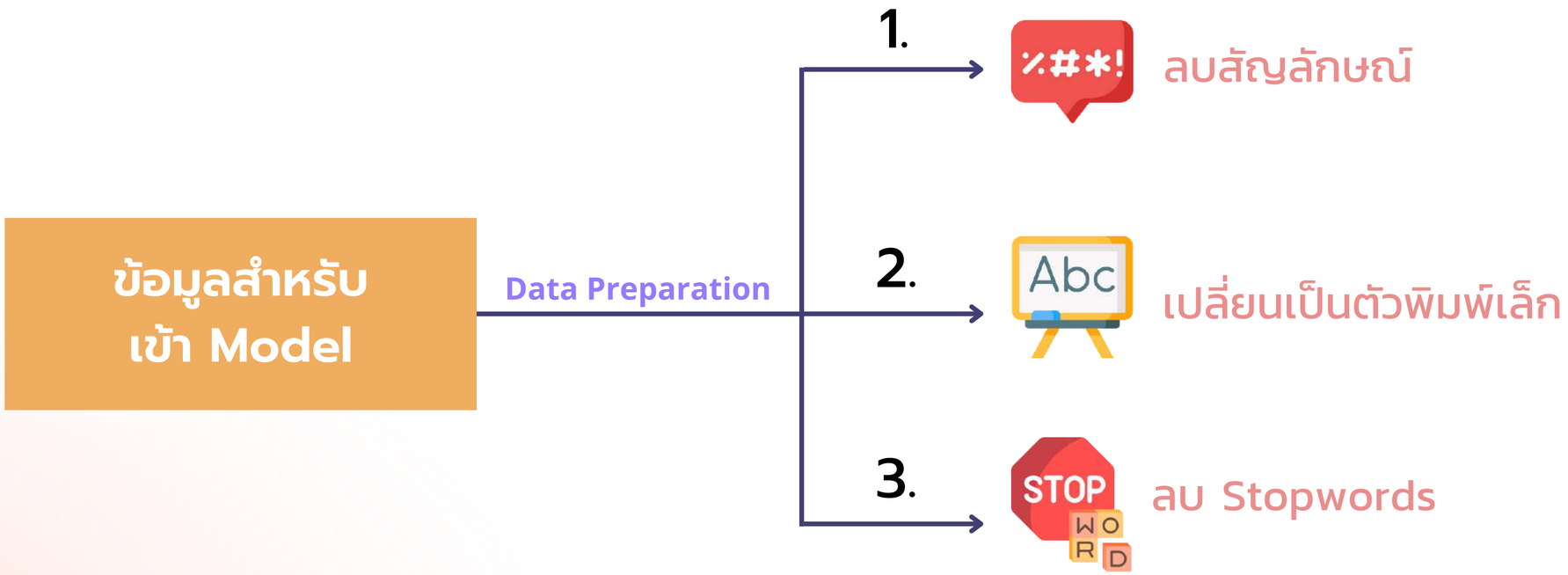
- จำนวนข้อมูลในแต่ละ class ของข้อมูล train มีไม่เท่ากัน
- ข้อมูลในบาง Class มีจำนวนน้อยเกินไป
- column text ในบาง row เหมือนกันทั้งหมด
- ความหมายของ class ไม่ชัดเจน

Objective

ใน Dataset นี้ แต่ละ Class (9 classes) สื่อถึงอะไร

โดยการทำ Topic Modeling

02 Data Preparation



02 Data Preparation

- **ตัวอย่าง Stopwords**

```
nltk.download('stopwords')
stopwords = set(nltk.corpus.stopwords.words('english'))
stopwords
```

```
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Unzipping corpora/stopwords.zip.
{'a',
 'about',
 'above',
 'after',
 'again',
 'against',
 'ain',
 'all',
 'am',
 'an',
 'and',
 'any',
 'are',
 'aren',
 "aren't",
 'as',
 'at',
 'be',
 'because',
 'been',
 'before',
 'being',
 'below',
 'between',
 'both',
 'but',
 'by',
 'can',
 'couldn',
 "couldn't",
 'd',
 'did',
 'didn',
 "didn't",
 'do',
 'does',
 'doesn',
 "doesn't",
 'doing',
```

02 Data Preparation

- **ข้อมูลที่ได้จาก Data Cleansing**

```
train_merge
```

	ID	Gene	Variation	Class	TEXT	TEXT_cleaned
0	0	FAM58A	Truncating Mutations	1	Cyclin-dependent kinases (CDKs) regulate a var...	cyclin dependent kinases cdks regulate variety...
1	1	CBL	W802*	2	Abstract Background Non-small cell lung canc...	abstract background non small cell lung cance...
2	2	CBL	Q249E	2	Abstract Background Non-small cell lung canc...	abstract background non small cell lung cance...
3	3	CBL	N454D	3	Recent evidence has demonstrated that acquired...	recent evidence demonstrated acquired uniparen...
4	4	CBL	L399V	4	Oncogenic mutations in the monomeric Casitas B...	oncogenic mutations monomeric casitas b lineag...
...	...	...	...	...	...	...
3316	3316	RUNX1	D171N	4	Introduction Myelodysplastic syndromes (MDS) ...	introduction myelodysplastic syndromes mds het...
3317	3317	RUNX1	A122*	1	Introduction Myelodysplastic syndromes (MDS) ...	introduction myelodysplastic syndromes mds het...
3318	3318	RUNX1	Fusions	1	The Runt-related transcription factor 1 gene (...	runt related transcription factor 1 gene runx1...
3319	3319	RUNX1	R80C	4	The RUNX1/AML1 gene is the most frequent targe...	runx1 aml1 gene frequent target chromosomal tr...
3320	3320	RUNX1	K83E	4	The most frequent mutations associated with le...	frequent mutations associated leukemia recurre...

3321 rows × 8 columns

02 Data Preparation

	ข้อมูลสำหรับเข้า Model	ข้อมูลสำหรับทำ Wordcloud
ลบสัญลักษณ์	✓	✓
เปลี่ยนเป็นตัวพิมพ์เล็ก	✓	✓
ลบ Stopwords	✓	✓
ทำ Lemmatisation	X	✓
แยกคำ	X	✓

02 Data Preparation

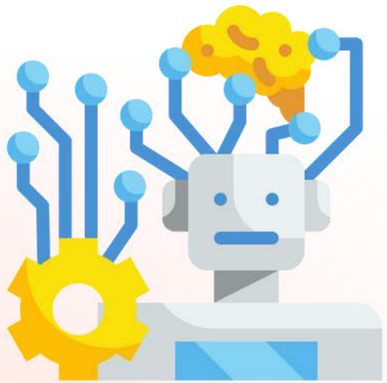
ข้อมูลสำหรับทำ Wordcloud

train_merge

	ID	Gene	Variation	Class	TEXT	length	TEXT_cleaned	tokens_word	tokens_word_lemma
0	0	FAM58A	Truncating Mutations	1	Cyclin-dependent kinases (CDKs) regulate a var...	39672.0	cyclin dependent kinases cdks regulate variety...	[cyclin, dependent, kinases, cdks, regulate, v...	[cyclin, dependent, kinase, cdks, regulate, va...
1	1	CBL	W802*	2	Abstract Background Non-small cell lung canc...	36691.0	abstract background non small cell lung cance...	[abstract, background, non, small, cell, lung,...	[abstract, background, non, small, cell, lung,...
2	2	CBL	Q249E	2	Abstract Background Non-small cell lung canc...	36691.0	abstract background non small cell lung cance...	[abstract, background, non, small, cell, lung,...	[abstract, background, non, small, cell, lung,...
3	3	CBL	N454D	3	Recent evidence has demonstrated that acquired...	36238.0	recent evidence demonstrated acquired uniparen...	[recent, evidence, demonstrated, acquired, uni...	[recent, evidence, demonstrated, acquired, uni...
4	4	CBL	L399V	4	Oncogenic mutations in the monomeric Casitas B...	41308.0	oncogenic mutations monomeric casitas b lineag...	[oncogenic, mutations, monomeric, casitas, b, ...	[oncogenic, mutation, monomeric, casitas, b, l...
...	...	...	...	...	...	...	...	...	...
3316	3316	RUNX1	D171N	4	Introduction Myelodysplastic syndromes (MDS) ...	73895.0	introduction myelodysplastic syndromes mds het...	[introduction, myelodysplastic, syndromes, mds...	[introduction, myelodysplastic, syndrome, md, ...
3317	3317	RUNX1	A122*	1	Introduction Myelodysplastic syndromes (MDS) ...	40127.0	introduction myelodysplastic syndromes mds het...	[introduction, myelodysplastic, syndromes, mds...	[introduction, myelodysplastic, syndrome, md, ...
3318	3318	RUNX1	Fusions	1	The Runt-related transcription factor 1 gene (...	36200.0	runt related transcription factor 1 gene runx1...	[runt, related, transcription, factor, 1, gene...	[runt, related, transcription, factor, 1, gene...
3319	3319	RUNX1	R80C	4	The RUNX1/AML1 gene is the most frequent targe...	32520.0	runx1 aml1 gene frequent target chromosomal tr...	[runx1, aml1, gene, frequent, target, chromoso...	[runx1, aml1, gene, frequent, target, chromoso...
3320	3320	RUNX1	K83E	4	The most frequent mutations associated with le...	67136.0	frequent mutations associated leukemia recurre...	[frequent, mutations, associated, leukemia, re...	[frequent, mutation, associated, leukemia, rec...

3321 rows × 9 columns

03 Modeling



UNSUPERVISED LEARNING

TOPIC MODELING

"Latent Dirichlet Allocation (LDA)"

04 Result

Result of Topic Modeling

Class 1

```
Topic #0:
ure domain binding proteins mutant dna activity wild mutants residues
Topic #1:
patients exon cancer deletion cases dna family genes ure associated
Topic #2:
tumor tumors cancer ure genes loss lines supplementary dna human
Topic #3:
p53 dna binding wild domain mutant ure site mutants levels
Topic #4:
binding smad3 complex proteins tsc2 wild ure amino domain activity
```

04 Result

Result of Topic Modeling

Class 2

```
Topic #0:
gene expression cancer fusion genes ure tumor tumors protein cases
Topic #1:
kinase mutations mutant activity domain mutants protein activation type mutation
Topic #2:
braf ras raf met mutations activation kinase ure cancer activity
Topic #3:
patients patient mutations kinase clinical treatment domain fusion response gene
Topic #4:
resistance ure mutations mutation tumor pathway tumors treatment clinical identified
```

04 Result

Result of Topic Modeling

Class 3

```
Topic #0:
flt3 kinase dna sequence signaling activation expressing patients phosphorylation identified
Topic #1:
alk variants brca1 activating variant patients functional kinase tsc1 assays
Topic #2:
brca1 variants erbb4 assay ssa recombination repair variant missense mutants
Topic #3:
brca1 variants assay vus functional neutral assays variant dna function
Topic #4:
mtor activating supplementary signaling amino mutants identified brca1 variants wt
```

04 Result

Result of Topic Modeling

Class 4

```
Topic #0:
dna patients cancer tumors genes binding domain site ure tumor
Topic #1:
activity catalytic residues kinase acid amino proteins mutant function phosphatase
Topic #2:
ure tumor genes cancer supplementary tumors mutant identified domain loss
Topic #3:
brca1 ras cancer activation activity assay ure function loss dna
Topic #4:
variants variant functional missense assays wild assay binding function activity
```

04 Result

Result of Topic Modeling

Class 5

```
Topic #0:
brca1 vus variants neutral deleterious brca2 functional assays assay sequence
Topic #1:
variants variant classification class pathogenic clinical family missense functional assay
Topic #2:
ure dna structure residues variants kinase binding supplementary proteins studies
Topic #3:
domain dna breast binding residues studies domains proteins number wild
Topic #4:
variants brca1 brct binding functional assays structure missense wild author
```

04 Result

Result of Topic Modeling

Class 6

```
Topic #0:
cells ure dna erbb2 cell associated resistance activity binding expression
Topic #1:
brca1 variants assay cells functional missense variant activity expression breast
Topic #2:
variants brca1 deleterious brca2 vus family sequence variant neutral history
Topic #3:
patients expression cells tumor tumors ure cell clinical mutant 12
Topic #4:
binding site kinase resistance variants ure cells residues activation structure
```

04 Result

Result of Topic Modeling

Class 7

```
Topic #0:
egfr patients gefitinib lung response treatment cancer exon clinical tumor
Topic #1:
patients exon kit patient 11 clinical response study disease tumor
Topic #2:
egfr exon patients gefitinib response clinical treatment tumor survival study
Topic #3:
cancer tumor gene tumors genes ure dna pathway human binding
Topic #4:
ure mutants mutant signaling phosphorylation wt expressing receptor wild growth
```

04 Result

Result of Topic Modeling

Class 8

```
Topic #0:
idh1 idh2 2hg mutant idh alleles wild tumor mutation ure
Topic #1:
akt ure kinase activation mutant activity lines binding supplementary mutation
Topic #2:
mutant erbb2 genes tumors mutation kinase idh1 sf3b1 expressing associated
Topic #3:
erbb2 mutation kinase associated growth binding expressing mutant activity activation
Topic #4:
genes gene tumors methylation ure tumor observed rna supplementary lines
```

04 Result

Result of Topic Modeling

Class 9

```
Topic #0:
sf3b1 genes splicing cell gene u2af35 exon mds supplementary rna
Topic #1:
ezh2 cell ure histone h3k27 levels methylation activity wt cancer
Topic #2:
sf3b1 splicing ure genes mds patients rna aberrant sf3b1mut cell
Topic #3:
ag cell alternative ezh2 methylation sf3b1 levels lines idh1 histone
Topic #4:
sf3b1 idh1 idh2 2hg splicing cell idh ure patients genes
```

05 Conclusion & Discussion

จากการทำ Topic Modeling พบว่า

- หลาย ๆ Class มีเนื้อหาที่ใกล้เคียงกัน
- แต่ในบาง Class มีเนื้อหาไปในทิศทางเดียวกัน เช่น Class 9 พูดถึง sf3b1 ชัดเจน
- ต้องอาศัยความรู้ในการตีความ Topic ค่อนข้างมาก

05 Conclusion & Discussion

ในการศึกษาครั้งหน้าจึงอยาก

- ลองใช้ grid search กับ LDA
- ลองนำข้อมูลที่ทำ Lemmatisation เข้า model ดู
- ลองใช้ Model อื่น ๆ ในการทำ Topic Modeling
- หาหน้าที่ และความเกี่ยวข้องกันของ Gene หรือ Variation ในแต่ละ Class

Responsible

Data Understanding
Modeling
Data interpretation

Data Understanding
Data Preparation
Data interpretation

Data Understanding
Modeling
Data interpretation

Thanks for your attention

Back Slides

02 Data Preparation

1.  ลบสัญลักษณ์

```
[18] def tokenize(text):
         return re.sub('[^a-zA-Z0-9\n]', ' ', text) # เท่าที่เข้าใจคือจะแทนที่ทุกตัวที่ไม่ใช่ a-z, A-Z, 0-9, \n ด้วย ' '
```

```
[19] pipeline = [str, tokenize]

     def prepare(text, pipeline):
         tokens = text
         for transform in pipeline:
             tokens = transform(tokens)
         return tokens
```

```
[20] train_merge['TEXT_cleaned'] = train_merge['TEXT'].apply(prepare, pipeline = pipeline) # ใช้ prepare() function
```

02 Data Preparation

2.  เปลี่ยนเป็นตัวพิมพ์เล็ก

```
[21] # ทำให้ str เป็นตัวพิมพ์เล็ก
     train_merge['TEXT_cleaned'] = train_merge['TEXT_cleaned'].str.lower()
```

3.  ลบ Stopwords

```
[22] # ลบ stopwords ออก
     train_merge['TEXT_cleaned'] = train_merge['TEXT_cleaned'].str.split(' ').apply(lambda x: ' '.join(k for k in x if k not in stopwords))
```

02 Data Preparation

4.  อื่น ๆ

```
# ลบ spacebar หลาย ๆ ครั้งออก
train_merge['TEXT_cleaned'] = train_merge['TEXT_cleaned'].str.replace(r'\s+', ' ', regex=True)
```

```
[24] # ตัดแบ่งค่า ไว้ใน col ใหม่ tokens_word
     train_merge['tokens_word'] = train_merge['TEXT_cleaned'].str.split()
```

ดูความยาวของ Text ในข้อมูล

```
# เพิ่ม col length
train_merge['length'] = train_merge['TEXT'].str.len()
```

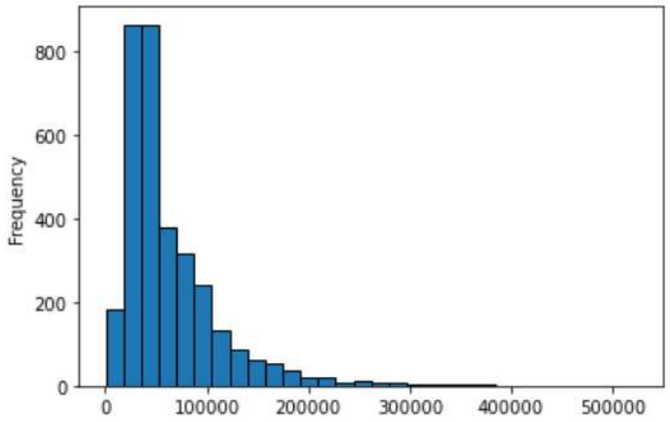
```
train_merge
```

	ID	Gene	Variation	Class	TEXT	length
0	0	FAM58A	Truncating Mutations	1	Cyclin-dependent kinases (CDKs) regulate a var...	39672.0
1	1	CBL	W802*	2	Abstract Background Non-small cell lung canc...	36691.0
2	2	CBL	Q249E	2	Abstract Background Non-small cell lung canc...	36691.0
3	3	CBL	N454D	3	Recent evidence has demonstrated that acquired...	36238.0
4	4	CBL	L399V	4	Oncogenic mutations in the monomeric Casitas B...	41308.0
...	...	...	...	...	...	...
3316	3316	RUNX1	D171N	4	Introduction Myelodysplastic syndromes (MDS) ...	73895.0
3317	3317	RUNX1	A122*	1	Introduction Myelodysplastic syndromes (MDS) ...	40127.0
3318	3318	RUNX1	Fusions	1	The Runt-related transcription factor 1 gene (...	36200.0
3319	3319	RUNX1	R80C	4	The RUNX1/AML1 gene is the most frequent targe...	32520.0
3320	3320	RUNX1	K83E	4	The most frequent mutations associated with le...	67136.0

3321 rows × 6 columns

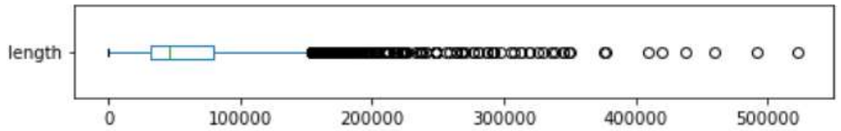
```
train_merge['length'].plot(kind='hist', bins=30, figsize=(6, 4), edgecolor='k')
```

```
<AxesSubplot:ylabel='Frequency'>
```




```
train_merge['length'].plot(kind='box', vert=False, figsize=(8, 1))
```

```
<AxesSubplot:>
```



จำนวนToken ในแต่ละ Class

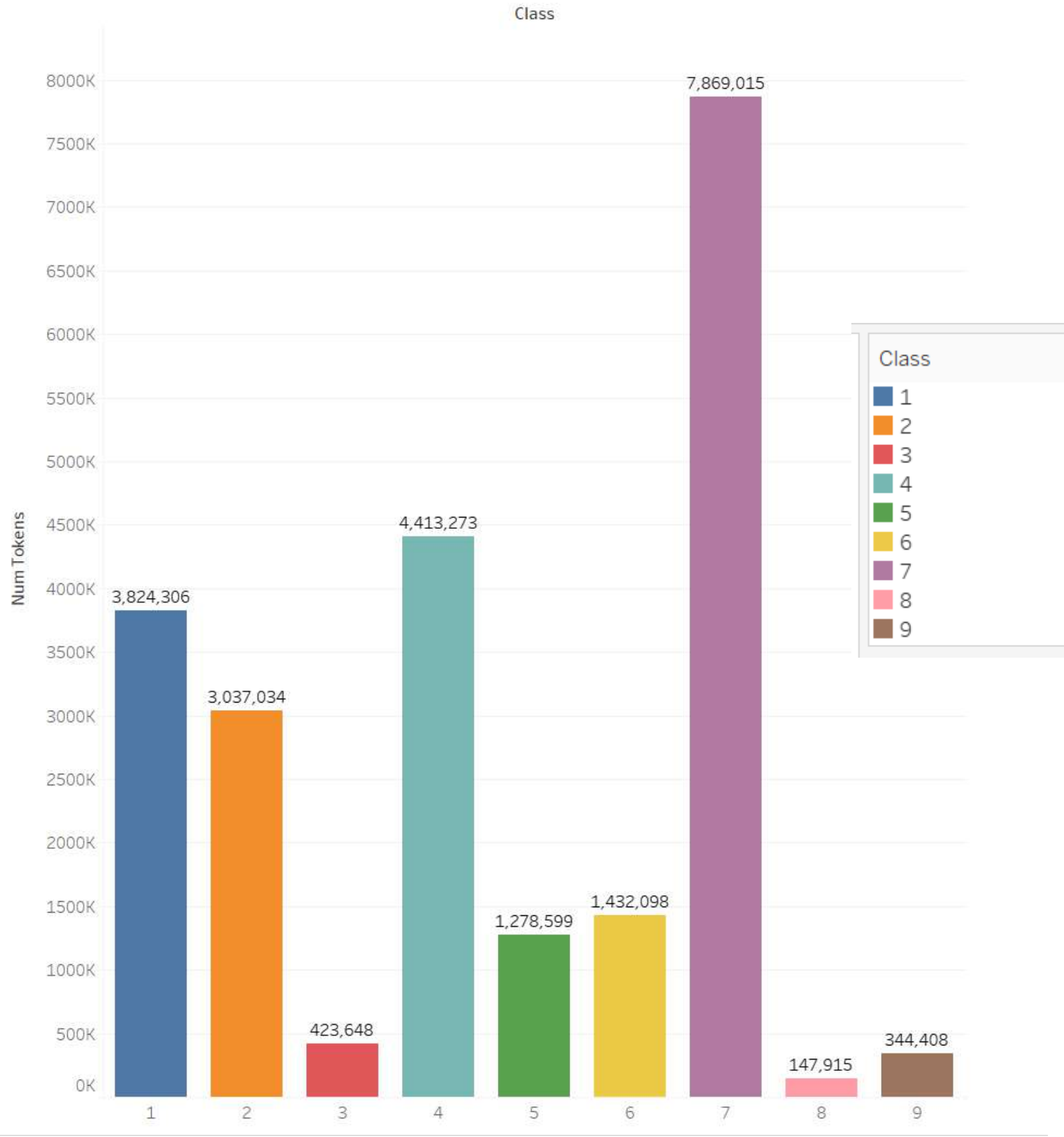
word len

```
train_merge['num_tokens'] = train_merge['tokens_word_lemma'].map(len)
```

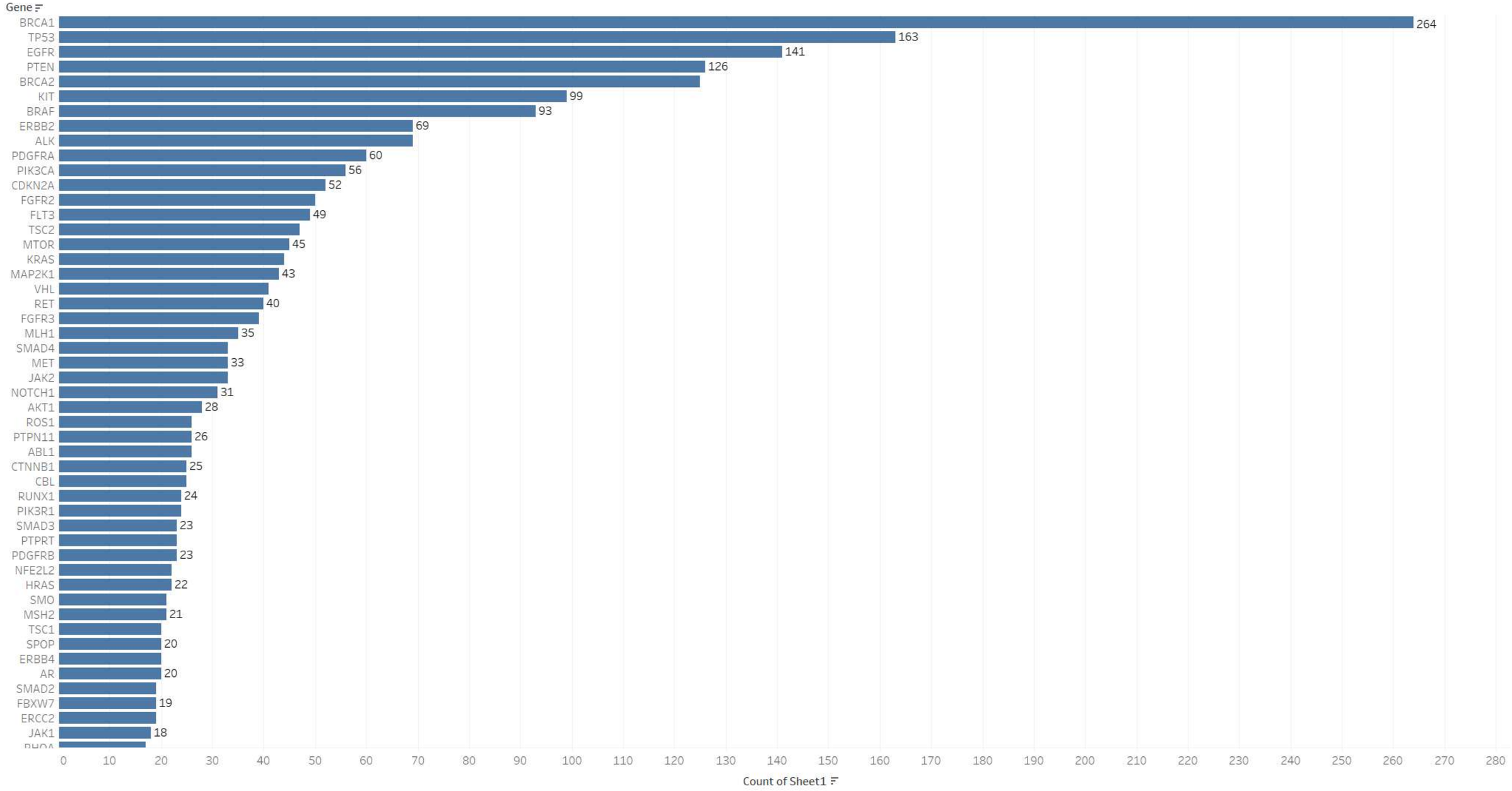
```
train_merge
```

	ID	Gene	Variation	Class	TEXT	length	TEXT_cleaned	tokens_word	tokens_word_lemma	num_tokens
0	0	FAM58A	Truncating Mutations	1	Cyclin-dependent kinases (CDKs) regulate a var...	39672.0	cyclin dependent kinases cdks regulate variety...	[cyclin, dependent, kinases, cdks, regulate, v...	[cyclin, dependent, kinase, cdks, regulate, va...	4370
1	1	CBL	W802*	2	Abstract Background Non-small cell lung canc...	36691.0	abstract background non small cell lung cance...	[abstract, background, non, small, cell, lung,...	[abstract, background, non, small, cell, lung,...	4139
2	2	CBL	Q249E	2	Abstract Background Non-small cell lung canc...	36691.0	abstract background non small cell lung cance...	[abstract, background, non, small, cell, lung,...	[abstract, background, non, small, cell, lung,...	4139
3	3	CBL	N454D	3	Recent evidence has demonstrated that acquired...	36238.0	recent evidence demonstrated acquired uniparen...	[recent, evidence, demonstrated, acquired, uni...	[recent, evidence, demonstrated, acquired, uni...	3841
4	4	CBL	L399V	4	Oncogenic mutations in the monomeric Casitas B...	41308.0	oncogenic mutations monomeric casitas b lineag...	[oncogenic, mutations, monomeric, casitas, b, ...	[oncogenic, mutation, monomeric, casitas, b, l...	4254
...	...	...	...	...	...	...	...	...	...	...
3316	3316	RUNX1	D171N	4	Introduction Myelodysplastic syndromes (MDS) ...	73895.0	introduction myelodysplastic syndromes mds het...	[introduction, myelodysplastic, syndromes, mds...	[introduction, myelodysplastic, syndrome, md, ...	8153
3317	3317	RUNX1	A122*	1	Introduction Myelodysplastic syndromes (MDS) ...	40127.0	introduction myelodysplastic syndromes mds het...	[introduction, myelodysplastic, syndromes, mds...	[introduction, myelodysplastic, syndrome, md, ...	4495
3318	3318	RUNX1	Fusions	1	The Runt-related transcription factor 1 gene (...	36200.0	runt related transcription factor 1 gene runx1...	[runt, related, transcription, factor, 1, gene...	[runt, related, transcription, factor, 1, gene...	4593
3319	3319	RUNX1	R80C	4	The RUNX1/AML1 gene is the most frequent targe...	32520.0	runx1 aml1 gene frequent target chromosomal tr...	[runx1, aml1, gene, frequent, target, chromoso...	[runx1, aml1, gene, frequent, target, chromoso...	3465
3320	3320	RUNX1	K83E	4	The most frequent mutations associated with le...	67136.0	frequent mutations associated leukemia recurre...	[frequent, mutations, associated, leukemia, re...	[frequent, mutation, associated, leukemia, rec...	7013

3321 rows × 10 columns



จำนวน Gene ที่พบในข้อมูล



จำนวน Gene ที่พบในข้อมูล

Sort fields | Data source order

ID	Gene	Variation	Class	Text
0	FAM58A	Truncating Muta...	1	Cyclin-dependen...
1	CBL	W802*	2	Abstract Backgr...
2	CBL	Q249E	2	Abstract Backgr...
3	CBL	N454D	3	Recent evidence ...
4	CBL	L399V	4	Oncogenic mutat...
5	CBL	V391I	4	Oncogenic mutat...
6	CBL	V430M	5	Oncogenic mutat...
7	CBL	Deletion	1	CBL is a negative ...
8	CBL	Y371H	4	Abstract Juvenil...
9	CBL	C384R	4	Abstract Juvenil...
10	CBL	P395A	4	Oncogenic mutat...
11	CBL	K382E	4	Noonan syndrom...
12	CBL	R420Q	4	Noonan syndrom...
13	CBL	C381A	4	Noonan syndrom...
14	CBL	P428L	5	Oncogenic mutat...
15	CBL	D390Y	4	Noonan syndrom...
16	CBL	Truncating Muta...	1	To determine if r...

Describe Field

Gene

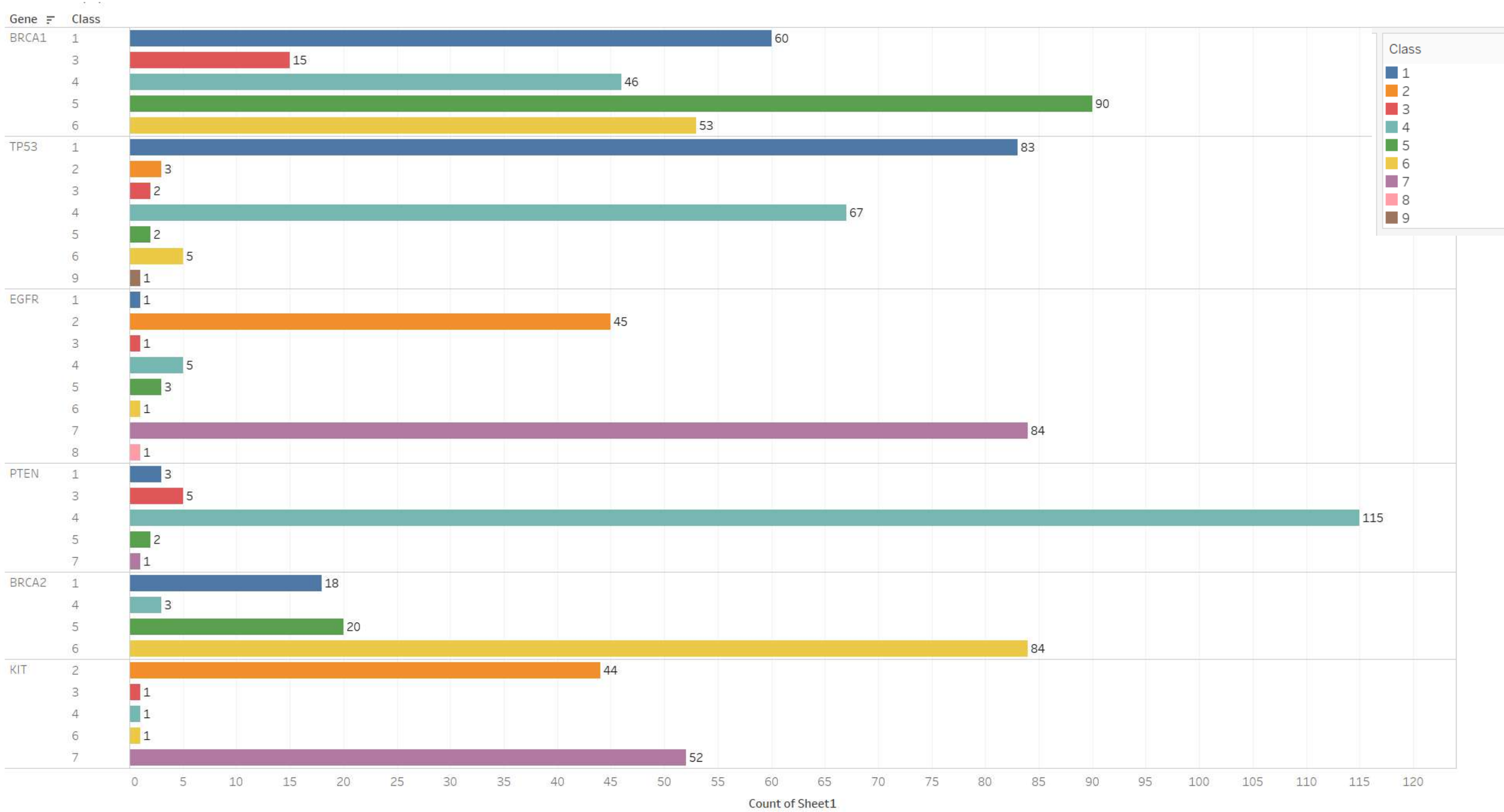
Role: Discrete Dimension
Type: Database column
Remote column: [Sheet1].[Gene]
Remote type: Unicode character string
Contains NULL: No
Locale: United States(English)
Sort flags: Case-insensitive
Column width: 8
Status: Valid

Domain (20 of 264 members)

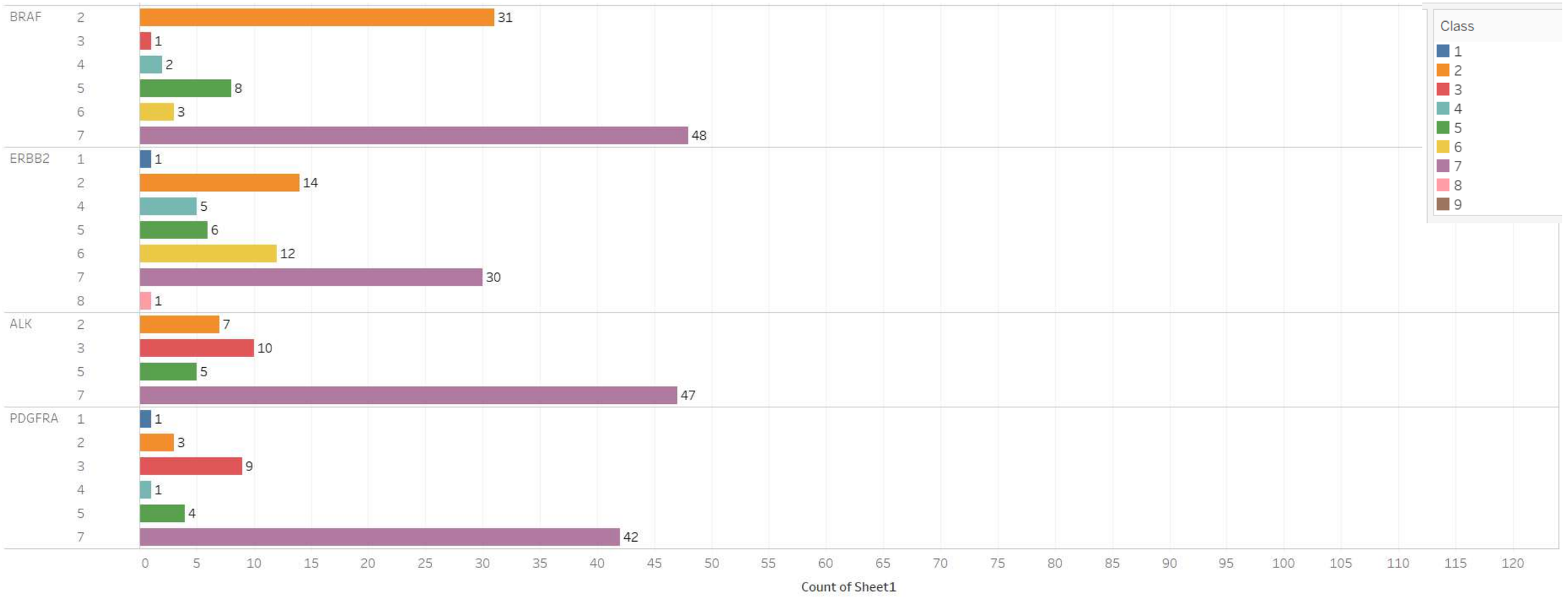
ABL1
ACVR1
AGO2
AKT1
AKT2
AKT3
ALK
APC
AR
ARAF
ARID1A
ARID1B
ARID2

Load | Copy | Close

จำนวน Gene ที่พบมาก 10 อันดับ ในแต่ละ Class



จำนวน Gene ที่พบมาก 10 อันดับ ในแต่ละ Class (ต่อ)



จำนวน Variation ที่พบในข้อมูล

Sort fields | Data source order

# Sheet1 ID	Abc Sheet1 Gene	Abc Sheet1 Variation	Abc Sheet1 Class	Abc Sheet1 Text
0	FAM58A	Truncating Muta...	1	Cyclin-dependen...
1	CBL	W802*	2	Abstract Backgr...
2	CBL	Q249E	2	Abstract Backgr...
3	CBL	N454D	3	Recent evidence ...
4	CBL	L399V	4	Oncogenic mutat...
5	CBL	V391I	4	Oncogenic mutat...
6	CBL	V430M	5	Oncogenic mutat...
7	CBL	Deletion	1	CBL is a negative ...
8	CBL	Y371H	4	Abstract Juvenil...
9	CBL	C384R	4	Abstract Juvenil...
10	CBL	P395A	4	Oncogenic mutat...
11	CBL	K382E	4	Noonan syndrom...
12	CBL	R420Q	4	Noonan syndrom...
13	CBL	C381A	4	Noonan syndrom...
14	CBL	P428L	5	Oncogenic mutat...
15	CBL	D390Y	4	Noonan syndrom...
16	CBL	Truncating Muta...	1	To determine if r...
17	CBL	Q367P	4	Acquired uniparе...

Describe Field

Variation

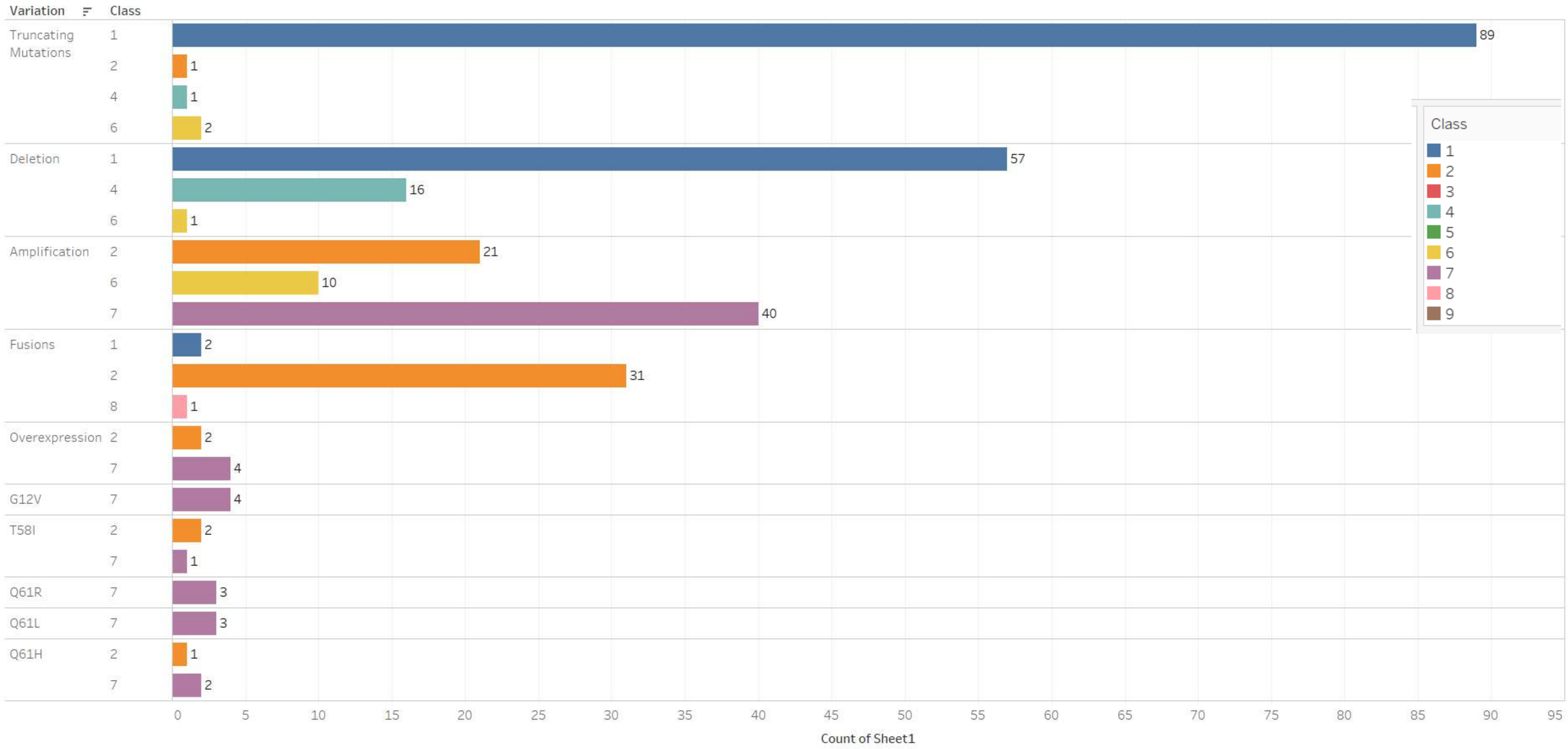
Role:	Discrete Dimension
Type:	Database column
Remote column:	[Sheet1].[Variation]
Remote type:	Unicode character string
Contains NULL:	Unknown
Locale:	United States(English)
Sort flags:	Case-insensitive
Column width:	Unknown
Status:	Valid

Domain (20 of 2,996 members)

1_2009trunc
2010_2471trunc
256_286trunc
3' Deletion
385_418del
422_605trunc
533_534del
534_536del
550_592del
560_561insER
596_619splice
963_D1010splice
981_1028splice
A1020V
A1022E
A1065T
A1066V
A1099T
A11_G12insGA
A111P
...

Load | Copy | Close

จำนวน Variation ที่พบมาก 10 อันดับ ในแต่ละ Class



จำนวน Text ที่พบในข้อมูล

Sort fields | Data source order

# ID	Abc Gene	Abc Variation	Abc Class	Abc Text
	Sheet1	Sheet1	Sheet1	Sheet1
0	FAM58A	Truncating Muta...	1	Cyclin-dependen...
1	CBL	W802*	2	Abstract Backgr...
2	CBL	Q249E	2	Abstract Backgr...
3	CBL	N454D	3	Recent evidence ...
4	CBL	L399V	4	Oncogenic mutat...
5	CBL	V391I	4	Oncogenic mutat...
6	CBL	V430M	5	Oncogenic mutat...
7	CBL	Deletion	1	CBL is a negative ...
8	CBL	Y371H	4	Abstract Juvenil...
9	CBL	C384R	4	Abstract Juvenil...
10	CBL	P395A	4	Oncogenic mutat...
11	CBL	K382E	4	Noonan syndrom...
12	CBL	R420Q	4	Noonan syndrom...
13	CBL	C381A	4	Noonan syndrom...
14	CBL	P428L	5	Oncogenic mutat...
15	CBL	D390Y	4	Noonan syndrom...
16	CBL	Truncating Muta...	1	To determine if r...
17	CBL	Q367P	4	Acquired uniparе...

Describe Field

Text

Role: Discrete Dimension
Type: Database column
Remote column: [Sheet1].[TEXT]
Remote type: Unicode character string
Contains NULL: Unknown
Locale: United States(English)
Sort flags: Case-insensitive
Column width: Unknown
Status: Valid

Domain (20 of 1,610 members)

Null
A 57-year-old man with gastrointestinal stromal tumor (GIST) of the stomach with peritoneal dissemination underwent gastrectomy. After surgery, he was treated with 400mg/ day of imatinib, without recurrence, for 26 months. At 26 months, the imatinib dose was reduced because of nausea, and 4 months after the dose reduction, recurrence of GIST was detected, for which surgical resection was performed again. The first surgical specimen had a mutation of exon 11 in the ckit receptor gene. Intriguingly, the second surgical specimen had a novel mutation of exon 17, in addition to the abovementioned mutation, in the c-kit receptor gene. Based on the result of molecular analysis, the novel mutation of exon 17, induced by longterm chemotherapy, was judged to have been responsible for the recurrence, which perhaps was triggered by the dose reduction of imatinib. Introduction Gastrointestinal stromal tumors (GISTs) account for the majority of gastrointestinal mesenchymal tumors, including leiomyosarcomas and leiomyoblastomas. Most GISTs express c-kit receptor tyrosine kinase (KIT) [1]. Although surgery used to be the only valid therapy for GISTs, imatinib mesylate, a multitargeted tyrosine kinase inhibitor, has recently been reported to be effective therapy for GIST [2]. Nevertheless, several reports to date indicate that the acquisition of c-kit gene mutations can result in resistance to imatinib treatment [3,4]. We report a patient who developed recurrent GIST, in association with the development of a secondary mutation in the c-kit receptor gene. A 57-year-old man was referred to our hospital because of a submucosal tumor in the stomach. Upper gastrointestinal endoscopy revealed an elevated lesion, with a deep central depression, in the greater curvature of the gastric body. Biopsy specimens indicated GIST. He underwent laparotomy, in which an extramural tumor 14cm in diameter, was found in the lower gastric body; there was also a great account of peritoneal dissemination. He underwent distal gastrectomy without lymphadenectomy. Two tumors, each approximately 5cm, in the peritoneum were also resected. GIST in the stomach was confirmed by pathological and

Load | Copy | Close