**-------------------------------------------------------------------------------------**

**Gene Card and Variant Card Text Mining Error Analysis Report**

**-------------------------------------------------------------------------------------**

Hui Li

Dianwei Han

Jinlian Wang

Robert

**IT Team**

**Chen Lab**

**Mount Sinai Medicine School**

We performed two tests on the pdf files and generated mutation information from them.

**PART I**

The First Round:

Number of pdf files: **7081**

Number of files that had been processed successfully: **6456**

625 failed: 548 empty text files and 77 bad text files.

The reasons for failures:

#1. The pdf files were named in wrong format. [**230]** cases resulted in 230 empty text files.:q

Correct format: PMID.pdf. Example: **12809638.pdf**

* 1. Wrong format: PMID .pdf: There is one space after PMID. Example: **12813757 .pdf**
  2. Wrong format: PMID + other information.pdf.

**Examples: 12815606 (2).pdf, 16774583\_forward.pdf, 16774583\_updates.pdf**

Detailed file names are listed as follows.

|  |  |  |  |
| --- | --- | --- | --- |
| 1 10338101 .pdf    2 10408784 .pdf    3 10612839 .pdf    4 11102988 .pdf    5 11754108 .pdf    6 11754113 .pdf    7 11793481 .pdf    8 11793483 .pdf    9 11857743 .pdf   10 11857755 .pdf   11 11933210 .pdf   12 12007221 .pdf   13 12813757 .pdf   14 12815589 .pdf   15 12815590 .pdf   16 12815591 .pdf   17 12815592 .pdf   18 12815602 .pdf   19 12815606 (2).pdf   20 12815607 .pdf   21 12872253 .pdf   22 12872256 .pdf   23 12879434 .pdf   24 1301194 .pdf   25 1301932 .pdf   26 1301960 .pdf   27 15569173 .pdf   28 15682412 .pdf   29 15735794 .pdf   30 15810915 .pdf   31 16086318 .pdf   32 16173970 .pdf   33 16972227 .pdf   34 17498081 .pdf   35 17498085 .pdf   36 17610549 .pdf   37 17610560 .pdf   38 17973853 .pdf   39 18371166 .pdf   40 19678855 .pdf   41 20015216 .pdf   42 20028422 .pdf   43 20101161 .pdf   44 20232353 .pdf   45 20236128 .pdf   46 20236351 .pdf   47 20331761 .pdf   48 20333770 .pdf   49 20358604 .pdf   50 20358613 .pdf   51 20402754 .pdf   52 20403973 .pdf   53 20425798 .pdf   54 20431853 .pdf   55 20437613 .pdf   56 20447141 .pdf   57 20560959 .pdf   58 20574985 .pdf   59 20578160 .pdf   60 20597900 .pdf | 61 20597900 \_Japenese.pdf   62 20602486 .pdf   63 20618352 .pdf   64 20618355 .pdf   65 20635367 .pdf   66 20635399 .pdf   67 20653736 .pdf   68 20662851 .pdf   69 20681997 .pdf   70 20683995 .pdf   71 20695019 .pdf   72 20695909 .pdf   73 20718790 .pdf   74 20799318 .pdf   75 20809529 .pdf   76 20814951 .pdf   77 20825431 .pdf   78 20827784 .pdf   79 20841309 '.pdf   80 20848650 .pdf   81 20854437 .pdf   82 20883264 .pdf   83 20890085 .pdf   84 20931554 .pdf   85 20979233 .pdf   86 21031596 .pdf   87 21061399 .pdf   88 21070499 .pdf   89 21077202 .pdf   90 21091672 .pdf   91 21118338 .pdf   92 21140503 .pdf   93 21143337 .pdf   94 21143469 .pdf   95 21149811 .pdf   96 21162090 .pdf   97 21166991 .pdf   98 21182502 .pdf   99 21198797 .pdf  100 21204211 .pdf  101 21208200 .pdf  102 21214702 .pdf  103 21275939 .pdf  104 21291454 .pdf  105 21291455 .pdf  106 21309765 .pdf  107 21332469 .pdf  108 21342251 .pdf  109 21344540 .pdf  110 21382890 .pdf  111 21394825 .pdf  112 21394827 .pdf  113 21404287 .pdf  114 21410451 .pdf  115 21412940 .pdf  116 21416585 .pdf  117 21424115 .pdf  118 21448560 .pdf  119 21457405 .pdf  120 21472225 .pdf | 121 21480867 .pdf  122 21484994 .pdf  123 21484995 .pdf  124 21517827 .pdf  125 21517828 .pdf  126 21520339 .pdf  127 21520340 .pdf  128 21542064 .pdf  129 21545677 .pdf  130 21560188 .pdf  131 21560189 .pdf  132 21567908 .pdf  133 21574979 .pdf  134 21592259 .pdf  135 21598378 .pdf  136 21605293 .pdf  137 21645175 .pdf  138 21645214 .pdf  139 21649796 .pdf  140 21658164 .pdf  141 21668430 .pdf  142 21668437 .pdf  143 21689372 .pdf  144 21696385 .pdf  145 21699652 .pdf  146 21707886 .pdf  147 21713987 .pdf  148 21752152 .pdf  149 21762381 .pdf  150 21792934 .pdf  151 21801158 .pdf  152 21844455 .pdf  153 21895644 .pdf  154 21905166 .pdf  155 21906148 .pdf  156 21910785 .pdf  157 21913903 .pdf  158 21955034 .pdf  159 21987686 .pdf  160 21990100 .pdf  161 21999818 .pdf  162 22004070 .pdf  163 22004506 .pdf  164 22007850 .pdf  165 22009552 .pdf  166 22161967 .pdf  167 22211794 .pdf  168 22220808 .pdf  169 22223181 .pdf  170 22243284 .pdf  171 22246954 .pdf  172 22334403 .pdf  173 22367672 .pdf  174 22393170 .pdf  175 22409273 .pdf  176 22416021 .pdf  177 22434185 .pdf  178 22438180 .pdf  179 22462673 .pdf  180 22471768 .pdf | 181 22494447 .pdf  182 22495880 .pdf  183 22507161 .pdf  184 22612257 .pdf  185 22644586 .pdf  186 22672365 .pdf  187 22678886 .pdf  188 22681893 .pdf  189 22693042 .pdf  190 22958177 .pdf  191 22998390 .pdf  192 23014764 .pdf  193 23024633 .pdf  194 23034915 .pdf  195 23042628 .pdf  196 23064986 .pdf  197 23067021 .pdf  198 23076989 .pdf  199 23088352 .pdf  200 23093250 .pdf  201 23106673 .pdf  202 23106979 .pdf  203 23110775 .pdf  204 23129629 .pdf  205 23130880 .pdf  206 23143442 .pdf  207 23167694 .pdf  208 23240968 .pdf  209 23252888 .pdf  210 23278109 .pdf  211 23278291 .pdf  212 23278509 .pdf  213 23279303 .pdf  214 23279342 .pdf  215 23281160 .pdf  216 23305485 .pdf  217 23306324 .pdf  218 23306409 .pdf  219 23356216 .pdf  220 23420618 .pdf  221 23472758 .pdf  222 23559858 .pdf  223 23601268 .pdf  224 23787031 .pdf  225 24317041 .pdf  226 27292088 .pdf  227 7627196 .pdf  228 7914875 .pdf  229 9450898 .pdf  230 9603440 .pdf |

#2. Papers were written in Chinese (no solution so far). **[21]**

PMIDs are listed as follows.

|  |  |  |  |
| --- | --- | --- | --- |
| 15347444  24915922  25537548  25877244  25190158 | 26791927  27256227  25876686  25748404  27143074 | 27510877  25748408  25190167  25190157  27122458 | 25619355  25190159  26887550  15651669  20077893  21868336 |

#3. Pdf files are scanned (suggestion to download html or full text file) [**115]**

|  |  |  |  |
| --- | --- | --- | --- |
| 7683628  8406510  8288243  8297373  8406498  8526905  10698280  7575533  7581448  7670469  7679367  7773290  7794769  7987396  8088783  8252036  8299233  8348157  8484765  8490650  8581357  8589715  8896563  9171832  9326927  9354784  9398847  9590298 | 10611909  7713925  8048932  8135797  8260194  8297359  7626071  7550325  7581449  7670500  7682196  7811247  7842009  8012393  8088811  8258146  8358442  8401582  8513326  8528239  8630485  8640223  8896564  9018031  9326931  9398839  9462743  9661653 | 7695647  8096693  7999071  7177196  9139825  8361539  9661653  7551830  7521937  7670461  7698759  7842019  7894480  8012394  8088845  8298639  8406445  8406453  8528242  8528245  8693491  8696334  9090381  9171833  9326935  9398840  9500541  9858861 | 9858861  9018031  12934488  8316268  16596947  7493035  7518679  7522211  7626145  7663524  7698774  7719340  7914875  8028609  8088810  8298641  8439335  8449505  8528251  8530087  8696348  8896559  9140401  9288094  9326941  9398842  9537412 |

**#4. Unknown reason (need to further investigate). [257] A lot of papers are from “J Hum Genet”.**

**Some paper said “ downloaded from** [**http://jamanetwork.com/pdfaccess.ashx?url=/data/journals/neur/13512/by**](http://jamanetwork.com/pdfaccess.ashx?url=/data/journals/neur/13512/by) **a Icahn School of Medicine and Mount Sinai Medical Center User on07/14/2017” or other copy right …**

|  |  |  |  |
| --- | --- | --- | --- |
| 7871171  10586284  15515206  7637805  10083738  10319579  10376574  10404764  10495930  11381253  11405812  11487575  11493200  11522031  11532986  11532985  12406895  12413376  12537648  12464226  12625412  12820708  12885342  14596636  14599123  15069170  16415886  16415970  16416207  16435195  16467727  17895567  18520120  19015585  19074751  19168821  19289697  19295657  19398866  19461158  20023066  20046049  20124734  20350966  20424301  24944099  25288723  25312915  25316729  25373420  25388789  25404053  25425405  25447658  25503412  25491489  25503412  27375219  27500519  27683759  8639447  9233564  9401068  9747027  9929978 | 7782284  9610619  9678418  10024914  10083731  10319589  10570905  10570908  10636421  11556849  11571214  11572889  11594966  11689492  11704758  11708998  11741828  11743515  12873855  12947561  12879953  15937070  16087766  16087855  16244706  16270648  16344347  16471174  16476814  16523049  16601827  16601852  16601870  20453399  20532556  20558929  20647579  20823027  20838461  21115745  21285037  21386772  21447811  21487335  21745802  21921422  21940685  21940687  21986928  25504019  25613935  25876185  25922261  25948085  26018748  26075876  26077908  26102279  26114819  27701760  8703815  9233593  9569180  9767370 | 10721669  10766140  10794365  10807540  10807541  10896236  10910910  10910913  10944847  11043508  11043517  11216666  11231775  11336405  11341489  15477515  15967879  16118345  16151907  16183799  16601880  16641580  16648375  16679490  16738945  16802107  22008442  22041377  22089611  22156789  22159456  22185990  22214629  22378672  22687851  22824774  22892490  22892508  22960877  23018980  25008049  25030574  25044645  25110155  25117010  25152992  25170860  25222778  25236789  25243733  25273610  25277229  26115733  26115735  26223439  26408288  26462614  26499107  27935012  28182336  9163595  9326186  9569189  9767373 | 11788090  11788826  11799392  11843659  11883825  12008755  12227864  12925341  12963616  14568816  15256764  15262743  15313851  15710861  15750329  15933521  15998910  16838077  16897190  16924389  16937026  16971480  17003641  17030669  17030721  17047887  17066260  17160617  17172620  17294068  17309986  17347910  17361009  17410460  17426749  17526799  18075239  23034980  23112236  23162007  23271757  23299470  24706876  24827497  24838951  24850569  24934643  26583491  26593283  26609033  26655004  26693850  26789422  26990548  26997013  27037498  27066551  27167370  27171515  8594556  9213168  9399845  9621520  9852677 |

**NOTEs: Could not open:**

25748406, 26310554, 9450791, 9666834

#5 specific cases. (solution, re-download the whole files).

1. **16774583\_foreward.pdf: correctly processed, and obtained content of file. But this file is not very useful. Should download the whole file.**
2. **16167124.pdf: is not the whole paper. Should download the whole file.**

#6. Extracted info could not be inserted into Data base. (solution: update code to cover this case).

(1). 9678418-9678418-9678418.txt

output of this file:

Full Terms & Conditions of access and use can be found at

http://www.tandfonline.com/action/journalInformation?journalCode=icey20

Download by: [Mount Sinai Health System Libraries] Date: 13 July 2017, At: 12:38

Current Eye Research

ISSN: 0271-3683 (Print) 1460-2202 (Online) Journal homepage: http://www.tandfonline.com/loi/icey20

Altered rep-1 expression due to substitution at

position +3 of the IVS13 splice-donor site of the

choroideremia (CHM) gene

Laurent Beaufr?re, St?phanie Rieu, Jean-Claude Hache, Viviane Dumur,

Mireille Claustres & Sylvie Tuffery

To cite this article: Laurent Beaufr?re, St?phanie Rieu, Jean-Claude Hache, Viviane Dumur,

Mireille Claustres & Sylvie Tuffery (1998) Altered rep-1 expression due to substitution at position

+3 of the IVS13 splice-donor site of the choroideremia (CHM) gene, Current Eye Research, 17:7,

726-729, DOI: 10.1080/02713689808951249

To link to this article:  http://dx.doi.org/10.1080/02713689808951249

Published online: 02 Jul 2009.

Submit your article to this journal

Article views: 8

View related articles

Citing articles: 1 View citing articles

The information from txt file is not complete for database.

(2). 8316268-8316268-8316268.txt

output of this file:

The New England Journal of Medicine

Downloaded from nejm.org at MOUNT SINAI SCHOOL OF MEDICINE on July 11, 2017. For personal use only. No other uses without permission.

 Copyright ' 1993 Massachusetts Medical Society. All rights reserved.

This information from txt file is not useful for data base. So it could not be inserted into the Data base.

(3). 16596947-16596947-16596947.txt

output of this file:

Reproduced with permission of the copyright owner.  Further reproduction prohibited without permission.

Spectrum of CFTR Mutations on R?union Island: Impact on Neonatal Screening

Bienvenu, T;Viel, M;Leroy, C;Cartault, F;et al

Human Biology; Oct 2005; 77, 5; ProQuest

pg. 705

Reproduced with permission of the copyright owner.  Further reproduction prohibited without permission.

This information from txt file is not useful for data base. So it could not be inserted into the Data base.

(4). 27143074-27143074-27143074.txt

Output of this file:

?!"#$%&'()\*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^\_`abcdefghijklmnopqrstuvwxyz{|}~^??????????????????????????????????????????????????????????????????????????????????

9-U?/,G+VT1Y0=7WXZ@;H$>Q"BC!)<3'%(\*AEbM:LK2F5&D`#d.N?????4????????s??R??????????????????????P????w??????????????????????????x???O???a6z8??JSI[\^\_?ce?k?fpq?hi]gjlmno??y{??~rtu?v?|}^??????????????????????????????????

??????\*?'?F`5%>??XY(?$?de3\_-SBIJNO??H&"KLM2)?=???????^Z???[\?khipqrs#!Pbo????gR????????;Gxy~C???????????????????????}A?+4@7Q????????????az{f??????T???U?????^?????????|???1?.?9/Vc

M#QB!????|}a3@A$\_IJNO??<=>?`-??pqrsK?;GHLfghibo~?tuvwX??'()2%"xy?4?z{F+S5&,?^??????^?Z?????????C???????????k?D??????9T8???[??\*????./?E01?????????R67???VYc

M#QB!????|}a3A$\_;GHL?xypqrsJK<=>?@`-CDEboI~?fghi(NO?"+%)R??9:&FP??c??5,??'?4??12?V?0??^??\???SuZ??jXY[???????????78w??????t??????./T???????U?W?\*k]^????6lmn??????d???e????????????v??????????

M#QB!????|}a3A$\_;GHLXxypqrsJK?<=>?@`-CDEboI~?fghi(NO?"+)'R??%?9:&??????F\*????????w??????^?,Z?5???????Y??12VWj4?]T?v?8?c?????????S???z/7????????{????^??te??????0????6??.?[???

h???VW="?S???o?.w?B?\_^t???R???????'C??XNOb????I~??J??)$?????}ALM5???D?u?4G(fg01??K?H?ij!EY?;F??v??????pqrs????xy??????PQ?:?6789+,d/a-T&@3%2U\*`>Z#\k??<???c????[]lmn???

This messy information is not useful. So it could not be inserted into the Data base.

(5). 25190157-25190157-25190157.txt

?!"#$%&'()\*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^\_`abcdefghijklmnopqrstuvwxyz{|}~^???????????????????????????????????????????????????????????????????????????????????????????????????????????????????????????????

,=:Z/QU+X0Y>1N-@<(&dHE[^"'%$J56#\*Lc\32)SFD9`aG4]Pe8RBTK!;IVWAC.MO???\_???k????st^?f|????????7????????????????mlnu?vxy?ijb?z{gh???????wopqr?????}????~??????????????????????????????????????????????????????????????

??????(?P?\?Q???$S5?\_???????z???????'s?e?????????n????0?AEFD9!CI???X?7?o???"2g#?H)63G@BRTUVW&^uk`a}~m??y??+,|vw>\*LN???1:%<?4K-?8/;?MOYZ[]??????{???t?f?????????????????????????????.?b??????p??r?Jq=??c

1?ABCD!e5??K???E????a?(?)???\*???`"4?s?#$9?0??I'??:;%?&?o???@S??F}678??]???????p2M??zh<????+VHQT??g???G3?R\_?v???W???????L?????/U?d??X??w?P\t?b?-?=??rx>?NZ,im?Jf.^O?k?????yY????[????????c?????????jln^??q????u????

'%??y?:z^??Q]Re0J\_-\@m??^{??TU?(3?I???9?6??|7+$,BCD!????5?o??f}?"?#2??1ASEF&????~???????????M????????Pk????????????hx??????X?WHYa??????`??????=?????;??wst????????V?q8/G??n???[pr)????d>vu?\*NO4?????Zc????K?L??g<????b

R]:+$),'Q&H(F?\x??=?f?GSe5?\*6Bo?`aTU?|??I2???%EN">??????h9???J???!DK????\_/1;???L??v0?r?????-?34t^Z?C??A?W#V[8{?g}wd????????????qs?X?c???????y?????z???????????????P????^?b<??k??????u????jMO7???@Ylm.n

This messy information is not useful. So it could not be inserted into the Data base.

(6) 8430317-8430317-8430317.txt

on July 14, 2017  http://science.sciencemag.org/ Downloaded from

on July 14, 2017 http://science.sciencemag.org/Downloaded from

on July 14, 2017  http://science.sciencemag.org/ Downloaded from

on July 14, 2017 http://science.sciencemag.org/Downloaded from

The skipping of constitutive exons in vivo induced by nonsense mutations

HC Dietz, D Valle, CA Francomano, RJ Kendzior Jr, RE Pyeritz and GR Cutting

DOI: 10.1126/science.8430317

 (5095), 680-683.259Science

ARTICLE TOOLS

http://science.sciencemag.org/content/259/5095/680

REFERENCES

http://science.sciencemag.org/content/259/5095/680#BIBL

This article cites 29 articles, 17 of which you can access for free

PERMISSIONS

http://www.sciencemag.org/help/reprints-and-permissions

Terms of ServiceUse of this article is subject to the

 is a registered trademark of AAAS.ScienceWorks. The title

DC 20005. 2017 ' The Authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original U.S. Government

(print ISSN 0036-8075; online ISSN 1095-9203) is published by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington,Science

This messy information is not useful. So it could not be inserted into the Data base.

(7) 12966700-12966700-12966700.txt

See     discussions,    stats,  and     author  profiles        for     this    publication     at:     https://www.researchgate.net/publication/10573583

Gentamicin      administration  in      Duchenne

patients        with    premature       stop    codon.

Preliminary     results

Article         in              Acta    myologica:      myopathies      and     cardiomyopathies:       official        journal of      the     Mediterranean   Society of

Myology /       edited  by      the     Gaetano Conte   Academy for     the     study   of      striated        muscle  diseases        ?       June    2003

Source: PubMed

CITATIONS

124

READS

142

7       authors,        including:

Some    of      the     authors of      this    publication     are     also    working on      these   related projects:

Dystrophinopathy        View    project

Limb    girdle  and     congenital      dystrophy:genotype/phenotype    correlations.   View    project

Luisa   Politano

Second  University      of      Naples

249     PUBLICATIONS                    4,283   CITATIONS

SEE     PROFILE

Gerardo Nigro

Second  University      of      Naples

152     PUBLICATIONS                    2,221   CITATIONS

SEE     PROFILE

Giulio  Piluso

Universit?      degli   Studi   della   Campania        "Luigi  ?

94      PUBLICATIONS                    2,199   CITATIONS

SEE     PROFILE

Orlando Paciello

University      of      Naples  Federico        II

128     PUBLICATIONS                    1,122   CITATIONS

SEE     PROFILE

All     content following       this    page    was     uploaded        by      Luisa   Politano        on      11      November        2014.

The     user    has     requested       enhancement     of      the     downloaded      file.

This info is not useful. So it could not be inserted into the Data base.

**PART II**

The Second Round:

Number of pdf files: **1120**

Number of files that had been processed successfully: **1086**

34 failed

**#1. Other languages**

20597900

**#2. Scanned pdf. [4]**

1284639, 1363812, 1586746, 2306516

**#3. Unknown reason (need to further investigate). [29]**

|  |  |  |  |
| --- | --- | --- | --- |
| 10601093  19652879  23129629  23306409  23430926  23430927  23430940  23545674 | 10875853  20447141  23479201  23615275  23625533  23700290  23857568 | 11440996  23014764  23877479  24056153  24190153  24691507  9285778 | 15735794  2306523  23093250  24270469  24282183  24452774  24584348 |