# Candidate Cancer Gene Database Project

# History

The idea for the CCGD database originated from Tim Starr, David Largaespada and Vincent Keng in 2009. Due to the increasing number of transposon mutagenesis screens it was difficult to keep track of which genes were identified by which studies. Our idea was to have a web-based database that could be used to easily query genes or studies to quickly identify candidate cancer genes found in these mutagenesis screens.

Previously, in 2003, members of David Largaesapada's lab created a database of transposon insertions called Mouse Transposon Insertion Database (MTID), but it did not get used or updated {Roberg-Perez, 2003 #102}. A few studies were also entered into to the Retroviral tagged cancer gene database {Akagi, 2004 #184}, which was later renamed the Retrovirus and Transposon tagged Cancer Gene Database (RTCGD). The RTCGD is maintained by Keiko Akagi at the Ohio State University, and contains lists of transposon and viral insertions mapping to the mouse genome. The CCGD is complementary to these two databases, in that the database does not consist of transposon insertions, but is instead based on the cancer genes that were identified in the mutagenesis screens.

In 2013 two students, Erik Nyre and Kenneth Abbott, developed the CCGD. Erik extracted the datasets from all the currently published transposon mutagenesis screens. Ken wrote all the software for the web-based application.

The database is hosted by the Office of Information Technology at the University of Minnesota. The website is: <http://ccgd-starrlab.oit.umn.edu/search.php>

In 2019 The University upgraded from RedHat Enterprise Linus 6 to RedHat Enterprise Linux 7.

# Email and Google Groups

Set up a [ccgd@umn.edu](mailto:ccgd@umn.edu) email account.

8/6/14: Set up another email account in order to get Google Analytics back.

ccgd.starrlab@gmail.com

Password: Candidate!Genes

7/28/17 Password updated to: Candidate!Genes!

The sponsored account you have requested for " CCGD StarrLab" is ready.

If you requested this account for someone who will have an appointment in the future, you will need to let us know when that person's official entry appears in the electronic directory (http://umn.edu/lookup) so that we can convert this account to a staff/faculty account and discontinue the sponsored account. If you have requested this account in error, please email us to have it removed (sponsor@umn.edu). That's especially important if your request included a U Card - with its associated charge.

If you have questions and/or need to make changes to the above account

please send an email to sponsor@umn.edu. Please include the new

Internet ID in your request.

Expiration date: 20180701 (YYYYMMDD format)

Type of account: Basic

InternetID: starr211

\*Sign in: Up49kWkW

8/7/14: Ken Abbott changed the password to the one we used prevop

\*NOTE: This is a temporary value and must be reset the first time the user logs in.

The University utilizes Google Apps for Higher Education to make collaboration easier for staff/faculty. Please inform your affiliate that he/she can sign up for Google Apps for the University of Minnesota after he/she has created a permanent password.

Setup a CCGD google group through OIT. See Incident # INC0500212. To sign into the group, go to groups.google.com and enter only your email address in the login screen, then you will be directed to the UM sign-in screen, enter x500 and password. The current members are Ken Abbott and Erik Nyrex, while Tim Starr is the manager. We are not considered, owners, however, which must be the UM OIT.

# Box storage: NETFILES

Setup a NETFILEs account. We store the xcel spreadsheet uploads, the manuals and the manuscript versions in a folder called CCGD. To login to netfiles just google Netfiles and then sign in with x500 and UM password.

3/28/19: NETFILES was retired from UMN and replaced with Box storage. I set up an account and the files are now stored in a folder called CCGD in my account. When they transferred them from NETFILES there were no files in the CCGD folder, but there were several in the trash, so they transferred the trash also.

# Linux Hosting

## 6/20/13: Email from OIT

The following Linux host has been provisioned per your request.

HOST: [ccgd-starrlab.oit.umn.edu](http://ccgd-starrlab.oit.umn.edu/)

ssh access is avilable via:

1. Gateway server

Non OIT staff: [ale03.oit.umn.edu](http://ale03.oit.umn.edu/)

OIT staff: [ale.oit.umn.edu](http://ale.oit.umn.edu/) [ale02.oit.umn.edu](http://ale02.oit.umn.edu/)

Run "sudo -l" to see your current sudo rights.

Please install custom software and configurations in /swadm

For technical OIT linux documentation, please see:

<https://wiki.umn.edu/OitVirtualServerHosting/Linux>

For general information about the Server Hosting Service, please see:

<http://www.oit.umn.edu/hosting/>

For assistance with the VPN service, please see:

<http://www.oit.umn.edu/vpn/index.htm>

Please send all additional configuration requests as new tickets to [oialinux@umn.edu](mailto:oialinux@umn.edu)

Unfortunately, we don't install specific versions of packages for customers. We use the latest available packages from the RHEL repository. This helps us keep things up to date and minimizes security holes. Here's what's installed from your list:

[root@ccgd-starrlab ~]# rpm -qa | grep subversion

subversion-1.6.11-9.el6\_4.x86\_64

[root@ccgd-starrlab ~]# rpm -qa | grep gzip

gzip-1.3.12-19.el6\_4.x86\_64

[root@ccgd-starrlab ~]# rpm -qa | grep postfix

postfix-2.6.6-2.2.el6\_1.x86\_64

[root@ccgd-starrlab ~]# rpm -qa | grep http-2

rpm [root@ccgd-starrlab ~]# rpm -qa | grep httpd-2

httpd-2.2.15-28.el6\_4.x86\_64

[root@ccgd-starrlab ~]# rpm -qa | grep php-5

php-5.3.3-22.el6.x86\_64

[root@ccgd-starrlab ~]# rpm -qa | grep mysql-s

mysql-server-5.1.69-1.el6\_4.x86\_64

[root@ccgd-starrlab ~]# rpm -qa | grep subversion

subversion-1.6.11-9.el6\_4.x86\_64

[root@ccgd-starrlab ~]# rpm -qa | grep gzip

gzip-1.3.12-19.el6\_4.x86\_64

If you need a newer version of any of these, you're welcome to install them under /swadm.

We only have two of your requested Perl packages available. These are installed system-wide:

[root@ccgd-starrlab ~]# rpm -qa | grep DBD

perl-DBD-MySQL-4.013-3.el6.x86\_64

[root@ccgd-starrlab ~]# rpm -qa | grep -i ParseExcel

perl-Spreadsheet-ParseExcel-0.5900-1.el6.x86\_64

You'll have to install the others under /swadm.

The "swadm" user has the ability to run "sudoedit /etc/my.cnf", which will allow you to customize MySQL as requested. (User "swadm" can also start/stop the DB.)

Finally, task scheduling can be accomplished with crontab with either your x.500 account or the "swadm" user.

Please note that I was not able to find a userid "abbot195" in the directory. Can you confirm the spelling, and if correct, give us the individual's first and last name for our records? We'll then create an account.

I've added a user for you on [ale03.oit.umn.edu](http://ale03.oit.umn.edu/), which is our security gateway. Note that you'll need an M-Key for both the ale03 and ccgd-starrlab hosts (once you're logged in, you can set up an SSH keypair.)

Good luck and let us know at [oialinux@umn.edu](mailto:oialinux@umn.edu) if you have any questions or run into any trouble.

--Darby

We value your input. Please help us by taking the time to fill out this short survey:

Click [here](https://umnprd.service-now.com/nav_to.do?uri=survey_take.do?sysparm_survey=UMN%20IT%20Customer%20Survey%26sysparm_task_survey=44a5e3b1448681c0ec80c3c417ecca58) to take the survey. If the "Click here" link does not work please copy the following URL in your browser:

<https://umnprd.service-now.com/nav_to.do?uri=survey_take.do?sysparm_survey=UMN%20IT%20Customer%20Survey%26sysparm_task_survey=44a5e3b1448681c0ec80c3c417ecca58>

We believe we have resolved your incident, [INC0453652](https://umnprd.service-now.com/nav_to.do?uri=incident.do%3Fsys_id=84aa684044c2cd80ec80c3c417eccacc%26sysparm_stack=incident_list.do%3Fsysparm_query=active=true). If this issue is not resolved to your satisfaction, please reply to this message within 5 days and your incident will be reopened.

Thank you!

Ref:MSG2059564

M-Key (M Key) Serial number L293058

P\*--N = 0 6 0 2

Note: I have also set up Duo authentication using an iphone app for my cell phone (3/24/14).

## 3/25/19: John Trammell phone call

John will send a provision request to get our space on the new Red Hat 7 server.

Red Hat is a company that packages a set of Linux programs that work together well. Red Hat is an operating system for the server.

We will need to update php, possibly need to update perl, possibly new versions of other systems.

## 4/23/19: Sonya Sustacek, Tom Kell and John Trammell internet call and email

Tom Krell from HST can request a server on our behalf. They will set up with storage and database access.

Fill out provisioning form from Sonya and send to Tom Krell with info. Include info on the MySQL requirements

Email on 4/10/19 Questions to answer:

1. Where is the best location to move this site? Drupal? A new RHEL 7 server?

2. Who will migrate the info to the new location?

3. When can this happen?

4. Issues:

a. Outdated Python version

b. Not following U of M branding standards

Please feel free to invite others you think should be involved or send a representative from you team if you cannot attend.

Info sent to us on 10/31/18 from Tom Kell re: this server. Tom do all of these options still exist?

1. If Starrlab wants to continue to have a free OIT hosted server they will be responsible for transferring all data and apps, as OIT will not assist in this. HST (formerly AHC) will approve a request for a new server with a current OS as the server does not contain PHI/HIPAA. AHC-IS will be granted administrator rights to the new server per OIT policy. We will only exercise these rights when requested by OIT.

2. If Starrlab wants to continue to have a free OIT hosted server but need help moving apps and data we can do so at an hourly rate. Again, HST will be granted administrative rights and will use these during setup. We can be available on an ongoing basis for future support needs.

3. If they want a hosted HST server there will be a monthly charge and depending on effort involved we will assist at an hourly rate. We will manage the system in our environment to COE standards.

Ken Abbot response email:

Hi folks. I built this site for Tim in 2012. I’m a medical student doing 100-hour weeks (until December) so I cannot make it to this meeting, but I hope we can exchange some information via email. I think the best bet is just another RHEL server—with only a few webpages here, I think we don’t need a platform, and it could be a pain to adapt what we have now to work with a platform. I may be able to help with this, but I cannot really speculate much about how much I will be able to accomplish, according to any timetable; I would appreciate a labor estimate (and description of any known hurdles) from someone who has experience completing these RHEL migrations for other University entities.

I see the newer version of this invite has items 4a (Python) and 4b (branding). The Python is no big deal because our scripts are PHP and Perl. Regarding branding, I built these pages with a switch I could flip so University banners would pop up, if someone required branding, but those banners are probably 4-5 years old now. Does anyone really care about the branding?

## 4/23/19 Sonya email:

I put together a spreadsheet for you to complete. Thought it would be easier. Tom would need this info in order to request a server on your behalf. Since Tom will be submitting the request, his name and number would be on the form.

[Info Needed to submit VM Request](https://docs.google.com/spreadsheets/d/1cCw7jIa1W_DMcqD-43H-1y66dq_xs_Ifgq8ARoVuffY/edit?usp=sharing) (for Tim to complete)

[Linux VM Request Form](https://docs.google.com/a/umn.edu/forms/d/e/1FAIpQLSen1wM1Gxbz6M0egEZRMDNLs_Qj_H1TPyNIhKPXJFwDXBkEHw/viewform)  (for Tom to complete)

Below is where you can find info on UMN web branding standards.

<http://brand.umn.edu>    --> Brand has requirements for our websites.

<http://folwell.umn.edu> --> The brand-compliant website resource for University sites both Drupal and non-Drupal.

Please let me know if you have any questions about anything.

# Pathway analysis 2/25/14

### Overview

The basic idea is to use a modified version of gene set enrichment to see if the candidate cancer gene list is enriched for canonical pathways, if any individual study is enriched for canonical pathways, and if any group of studies based on cancer type is enriched for canonical pathways. The canonical pathways are the KEGG human pathways that are listed in the NCBI Biosystems database.

### Methods

**List of studies:**

A list of 28 studies was generated by combining CCGD studies that were essentially replicates, and by removing 2 studies that identified less than 10 CISs. See the CCGD Project excel worksheet for a detailed description of the method used to create this list.

**List of CIS human genes:**

A list of 11,820 human genes was created by selecting all CISs from the CCGD database that have a human symbol, expanding records that have multiple human symbols, adding in predisposing mutations based on the study, and eliminating any duplicate entries. This list is created by the perl script tks\_CCGD\_Human.pl. The two input files are the CCGD\_export.csv and the list of predisposing mutations (predisposing\_mutations.txt). The export file was downloaded on 2/26/14 and the predisposing mutations file was created using the CCGD Project excel worksheet. The final list is generated by the script and saved as CCGD\_Human\_output.txt. This is a text file with four tab-separated columns (Human symbol, Gene ID, Combined study name, Cancer Type).

Screen output from tks\_CCGD\_Human.pl script run on 2/26/14

12726 Number of input records

11967 Number of Records with Human Symbols

24 Number of records with multiple Human Symbols

11999 Number of records with human symbols after expanding

19 Number of predisposing mutations added

12018 Total number of Human symbols after adding predisposing mutations

11820 Number of records after eliminating duplicates

The list of CIS human genes has been output to the CCGD\_human\_output.txt file

The list generated above was then reduced to 10,558 records by combining studies that were replicates except for either predisposing mutations or method of identifying CISs. This was done using the perl script tks\_CCGD\_combine\_studies.pl

Screen output from tks\_CCGD\_combine\_studies.pl script run on 2/26/14

"Van der Weyden 2012-01" and "Van der Weyden 2012-02" were combined to create "Van der Weyden 2012 combined"

"Keng 2009-01" and "Keng 2013-01" were combined to create "Keng 2009 combined"

"Koso 2012-01" and "Koso 2012-02" were combined to create "Koso 2012 combined"

"Wu 2012-01" and "Wu 2012-02" were combined to create "Wu 2012 combined"

"Perez Mancera 2012-01" and "Perez Mancera 2012-02" were combined to create "Perez Mancera 2012 combined"

"Van der Weyden 2013-01" and "Van der Weyden 2013-02" and "Van der Weyden 2013-03" were combined to create "Van der Weyden 2013 combined"

"Rahrmann 2013-01" and "Rahrmann 2013-02" and "Rahrmann 2013-03" and "Rahrmann 2013-04" were combined to create "Rahrmann 2013 combined"

10566 Number of records after combining

"Ni 2013-01" and "Rahrmann 2009-01" were eliminated due to fewer than 10 CISs reported

10558 Number of records after combining and deleting

Data has been saved to CCGD\_combined\_output.txt

The CCGD\_combined\_output.txt file is also a text file with four tab-separated columns (Human symbol, Gene ID, Combined study name, Cancer Type). This is the final list of CISs for pathway analysis.

**List of pathways:**

The list of pathways was generated from downloads from the NCBI Biosystems database website (biosystems\_gene.txt and bsid2info.txt) and the HUGO gene symbol website. For a detailed description of the method see the worksheet NCBI\_Biosystems\_pathways.xlsx.

Screen output from tks\_NCBI\_Biosystems\_genes.txt run on 2/26/14

Number of pathways 443

Number of Entrez IDs 38105

Number of genes 16145501

The list of bsids and the gene symbols associated with those bsids has been saved as pathway\_genes.txt

The final list of pathways and genes has been saved as final\_output.txt

The pathway file is called final\_output.txt and is a tab delimited file containing the NCBI bsid, KEGG ID, KEGG pathway name, and all the gene symbols associated with the pathway.

2/26/14: Sent the three files above to Rachel along with a brief description of what to do.

# Pathway analysis 12/16/13

### Methods

Data from the 34 studies included in the CCGD were included in this pathway analysis. A total of 5,258 genes were identified in at least one of these studies, including predisposing mutations (Table 1). Among these, 1,736 genes were mapped and grouped into 225 KEGG curated pathways categories. After excluding pathways with 10 or less genes, 142 KEGG pathways remained for analysis. The distribution of resulting sizes (number of genes) of each KEGG pathway is shown in Figure 1.

Pathways across studies were analyzed following a modified individualized gene set analysis method. 1 For each study, a pathway was considered altered if 10% of genes in the pathway were mutated. The number of studies in which a particular pathway was altered represented a score for that pathway. Random permutation methods were used to assess whether each pathway score was statistically significant. A total of 2x108 permutations were conducted to obtain the distribution of score under the null for each pathway and calculate associated p-values.

For pathways by cancer type, two cancer types were explored which had data from multiple studies (Blood cancers – 13 studies and Nervous system cancers – 9 studies). Data were analyzed using Cochran-Mantel-Haenszel chi-squared tests.2

Reported p-values were adjusted to reduce the false discovery rate using the Benjamini and Hochberg method.3 All analyses were completed in R version 2.15.2.

Unsupervised clustering was performed (ask Rachel for method).

### Results

A total of 45 KEGG pathways were enriched with CIS-associated genes with adjusted p-values <0.01 (Table 2).

Most pathways are statistically significant in blood and nervous system cancers, though the top pathways differ (Table 3).

### References

1. Boca SM, Kinzler KW, Velculescu VE, Vogelstein B, Parmigiani G. “Patient oriented gene-set analysis for cancer mutation data.” Genome Biol., 11: R112, 2010.

2. Agresti A. Categorical data anlsysis (second edition). 2002. New York: Wiley.

3. Benjamini Y, Hochberg Y. “Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.” Journal of the Royal Statistical Society. Series B (Methodological), Vol. 57, No. 1 (1995), 289-300.

# Pathway Analysis 4/7/14

## Methods:

Data from the 28 studies included in the CCGD were included in this pathway analysis. A total of 6,459 genes were identified in at least one of these studies, including predisposing mutations (Table 1). Among these, 2,445 genes were mapped and grouped into 280 KEGG curated pathways categories (KEGG pathway version updated on March 14, 2014). After excluding pathways with 10 or less genes, 220 KEGG pathways remained for analysis. The resulting sizes of each KEGG pathway are shown in Figure 1.

A determination of statistically significant pathways across studies were analyzed following a modified individualized gene set analysis method (2). For each study, a pathway was considered altered if 10% of genes in the pathway were mutated. The number of studies in which a particular pathway was altered represented a score for that pathway. Random permutation methods were used to assess whether each pathway score was statistically significant. A total of 105permutations were conducted to obtain the distribution of score under the null for each pathway and calculate associated p-values. P-values were additional adjusted to reduce the false discovery rate using the Benjamini and Hochberg method (1995) (1).

For pathways by cancer type, two cancer types were explored which had data from multiple studies (Blood cancers – 12 studies and Nervous system cancers – 5 studies). The table for each study was obtained through hypergeometric test and the tables from multiple studies were combined using Cochran-Mantel-Haenszel chi-squared tests.

Reported p-values were adjusted to reduce the false discovery rate using the Benjamini and Hochberg method.1 All analyses were completed in R version 3.0.2

## References:

(1) Benjamini Y, Hochberg Y. “Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.” Journal of the Royal Statistical Society. Series B (Methodological), Vol. 57, No. 1 (1995), 289-300.

(2) Boca SM, Kinzler KW, Velculescu VE, Vogelstein B, Parmigiani G. “Patient oriented gene-set analysis for cancer mutation data.” Genome Biol., 11: R112, 2010.

## Results:

A total of 23 KEGG pathways were enriched with CIS-associated genes with permutated p-values <10-5; 46 additional with p-values <0.05 (Table 2; see Excel file for full list).

The enriched pathways for blood and nerve system cancer are listed in Table 3 and 4 respectively.

# Pathway Analysis 5/13/14

## Methods:

Data from the 28 studies included in the CCGD were included in this pathway analysis. A total of 6,459 genes were identified in at least one of these studies, including predisposing mutations (Table 1). Among these, 2,432 genes were mapped and grouped into 285 KEGG curated pathways categories (KEGG pathway version updated on May 03, 2014). After excluding pathways with 10 or less genes, 266 KEGG pathways remained for analysis. The resulting sizes of each KEGG pathway are shown in Figure 1.

## Pathway Enrichment Analysis

A determination of statistically significant pathways across studies were analyzed following a modified individualized gene set analysis method (2). For each study, a pathway was considered altered if 10% of genes in the pathway were mutated. The number of studies in which a particular pathway was altered represented a score for that pathway. Random permutation methods were used to assess whether each pathway score was statistically significant. A total of 10e6 permutations were conducted to obtain the distribution of score under the null for each pathway and calculate associated p-values. P-values were additional adjusted to reduce the false discovery rate using the Benjamini and Hochberg method (1995) (1). The results of this pathway enrichment permutation analysis are shown in Table 2.

In addition, we also try a second definition where if there is at 1 gene is mutated in the pathway, then the pathway is disrupted. Then we repeated the pathway enrichment analysis as described above. The analysis results are shown in Table 3.

### Individual Cancer Type Pathway Enrichment Analysis

For pathways by cancer type, two cancer types were explored which had data from multiple studies (Blood cancers – 12 studies and Nervous system cancers – 5 studies). The table for each study was obtained through hypergeometric test and the tables from multiple studies were combined using Cochran-Mantel-Haenszel chi-squared tests. Reported p-values were adjusted to reduce the false discovery rate using the Benjamini and Hochberg method.1 All analyses were completed in R version 3.0.2

### Pathways that Differentiate Blood Cancer from Other Type of Cancers

For each pathway, we perform logistic regression to examine whether mutation event in the pathway under consideration is associated whether the study is Blood cancer. The results of pathways that significantly associated with blood cancer are shown in Table 6. In Table 6, positive t statistic value indicates that higher mutation frequency in the pathway under consideration is associated with the occurrence of blood cancer. In other words, we can observe deeper blue colors in blood cancer studies than studies of other cancer types in the Heatmap2. Similar analysis results for nerve cancer are shown in Table 7.

## References:

(1) Benjamini Y, Hochberg Y. “Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.” Journal of the Royal Statistical Society. Series B (Methodological), Vol. 57, No. 1 (1995), 289-300.

(2) Boca SM, Kinzler KW, Velculescu VE, Vogelstein B, Parmigiani G. “Patient oriented gene-set analysis for cancer mutation data.” Genome Biol., 11: R112, 2010.

## Results:

A total of 20 KEGG pathways were enriched with CIS-associated genes with permutated p-values <10-6; 48 additional with adjusted p-values <0.05 (Table 2; see Excel file for full list). Using the second definition (at least 1 gene are mutated), there are 20 pathways with permuted p-value < <10-6, and 66 pathways with adjusted p value < 0.05. Among the pathways reported in Table 2, 38% of them belong to KEGG Human Disease category while the average frequency of human disease pathways are 26% in the total 266 pathways considered. Hence there is a significantly increased number of human disease pathways reported Table 2 (p value-0.03). In addition, among the pathways reported in Table 2, there are 22% pathways belong to cancer disease pathways while the average frequency of cancer disease pathways are 7.1% in the 266 pathways considered. Hence there are an increased number of human cancer pathways reported in Table 2 (p value=0.0003)

The enriched pathways for blood and nerve system cancer are listed in Table 4 and 5 respectively.

# Pathway Analysis 6/11/14

## Methods

Data from the 28 studies included in the CCGD were included in this pathway analysis. A total of 6,459 genes were identified in at least one of these studies, including predisposing mutations (Table 1). Among these, 2,432 genes were mapped and grouped into 285 KEGG curated pathways categories (KEGG pathway version updated on May 03, 2014). After excluding pathways that only had 10 or fewer genes identified in the CCGD, 266 KEGG pathways remained for analysis. The resulting sizes of each KEGG **pathway are shown in Figure 1.**

## Pathway Enrichment Analysis

Statistically significant pathways across studies were analyzed following a modified individualized gene set analysis method (2). For each study, a pathway was considered disrupted if at least 1 gene in the pathway were mutated. The number of studies in which a particular pathway was altered represented a score for that pathway. Random permutation methods were used to assess whether each pathway score was statistically significant. A total of 10e7 permutations were conducted to obtain the distribution of score under the null for each pathway and calculate associated p-values. P-values were additionally adjusted to reduce the false discovery rate using the Benjamini and Hochberg method (1995) (1). The results of this pathway enrichment permutation **analysis are shown in Table 2.**

### Individual Cancer Type Pathway Enrichment Analysis

For pathways by cancer type, two cancer types were explored which had data from multiple studies (Blood cancers – 12 studies and Nervous system cancers – 5 studies). The table for each study was obtained through hypergeometric test and the tables from multiple studies were combined using Cochran-Mantel-Haenszel chi-squared tests. Reported p-values were adjusted to reduce the false discovery rate using the Benjamini and Hochberg method.1 All analyses we**re completed in R version 3.0.2**

### Pathways that Differentiate Blood Cancer from Other Type of Cancers

For each pathway, we perform logistic regression to examine whether mutation event in the pathway under consideration is associated when the study is Blood cancer. The results of pathways that significantly associated with blood cancer are shown in Table 5. In Table 5, positive t statistic value indicates that higher mutation frequency in the pathway under consideration is associated with the occurrence of blood cancer. In other words, we can observe deeper blue colors in blood cancer studies than studies of other cancer types in the Heatmap2.

## References:

(1) Benjamini Y, Hochberg Y. “Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.” Journal of the Royal Statistical Society. Series B (Methodological), **Vol. 57, No. 1 (1995), 289-300.**

(2) Boca SM, Kinzler KW, Velculescu VE, Vogelstein B, Parmigiani G. “Patient oriented gene-set analysis for cancer mutation data.” **Genome Biol., 11: R112, 2010.**

## Results:

Using the definition (at least 1 gene is mutated in the pathway), there are 20 pathways with permuted p-value < <10-7, and 67 pathways with adjusted p value < 0.05. Among the pathways reported in Table 2, 26 (41.5%) of them belong to KEGG Human Disease category while the average frequency of human disease pathways are 24.2% (69/285) in the total 285 pathways considered. Hence there is a significantly increased number of human disease pathways repo**rted Table 2 (p value=0.003).**

Among all the 69 human disease pathways, 19 (19/69=27.5%) belong to cancer disease pathways. There is an increased frequency of cancer disease pathways reported in Table 2 since 18 (18/26=69.2%) of them are listed in Table 2 (p value < 0.001)

The enriched pathways for blood and nerve system cancer are listed in Table 3 and 4 respectively.

# Pathway Analysis for paper

Figure/Table/Text 1) The KEGG database lists 284 annotated human pathways, of which 220 contain at least 10 member genes. Of these 220 annotated human pathways, 69 (?probably less) are "disease-specific" associated pathways while the remaining 151(? estimate) are annotated as normal "healthy" human pathways. Of the 69 "disease-specific" pathways, 20 of them were cancer pathways, while the remaining 49 were other types of diseases. GSEA analysis of the CIS-associated genes indicated that 23 of 220 pathways were enriched for CIS-genes (p<0.0005)(? what is appropriate value). Of those 23 enriched pathways, 11 of them were cancer pathways (1 was a non-cancer disease pathway and the remaining 11 were healthy pathways). Chances of this distribution occurring by random chance are highly unlikely (2-Tail : p-value = 1.21e-7 Fisher's Exact Test?). This finding indicates that transposon screens for cancer genes are finding known cancer drier genes, and by inference, are also finding previously unknown cancer driver genes.

Figure/Table/Text 2) Of the 151(?) KEGG annotated pathways associated with human health, GSEA analysis found that several known pathways (ErbB, Jak-STAT, Wnt, and Notch) were highly associated with CIS-genes. Interestingly, there were some other pathways that were unexpected (Tight junction, Dorso-ventral axis formation, T & B cell receptor signaling etc.)

Figure/Table/Text 3) Of the 28 studies analyzed, 12 were conducted in hematopoietic cancers. Unsupervised hierarchical clustering of the studies indicated that the blood cancers clustered together (with 3 exceptions).

Figure/Table/Text 4) Analysis of the unsupervised hierarchical clustering allowed us to identify a subset of pathways that were specifically mutated in blood cancers, but not other cancers and vice versa. These pathways are shown in the Table.

Choose some blood-specific pathways and mention them.

# Reis-Filho help

Reis-Filho has a post-doc "Ng" who can run algorithms to pull out enriched or lost pathways in cancer. She would ned lots of information. Should contact them by email.

They use MEMO to discern pathways from mutation/CNA data from sets of tumors. This program can find cooperating and mutually exclusive pathways apparently.

They also use STRING from embo

They use NetGO for overlaying onto canonical pathways

They use Meta Core also for pathway analysis.

# Uploading new versions of CCGD

Connect to the server using your FTP client (See instructions below "Accessing the CCGD Server")

Migrate from the star0044 folder back to root, and then migrate to /swadm/var/archive/spreadsheet

Download the latest copy of the spreadsheet. At this point you can close FUGU because you will not upload anything into this directory using the FTP client. After modifying the file or adding a new study the file will be uploaded using the web browser.

Rename the spreadsheet with the current date “CCGD YYYY-MM-DD”

Add new records to the bottom of the "Study" sheet and use the Format painter to match the format to previous entries. Add the "Study Guide" information. Save the file and close it.

Go to: <http://ccgd-starrlab.oit.umn.edu/login.php>

Enter username and password below:

Username - ccgd

Password – Candidate!Genes

Click on "Choose file" button at bottom of window. Browse for the correct file and once selected, hit Upload. Verification of the upload should appear on the screen, as well as a log of past uploads.

Note: The CCGD makes the backup copy in the archive.

## Pre-6/30/15 Instructions

1. Save latest version of the database as “CCGD Eversion YYYY-MM-DD” and upload to netfiles. Note, the file should be .xls NOT .xlsx.

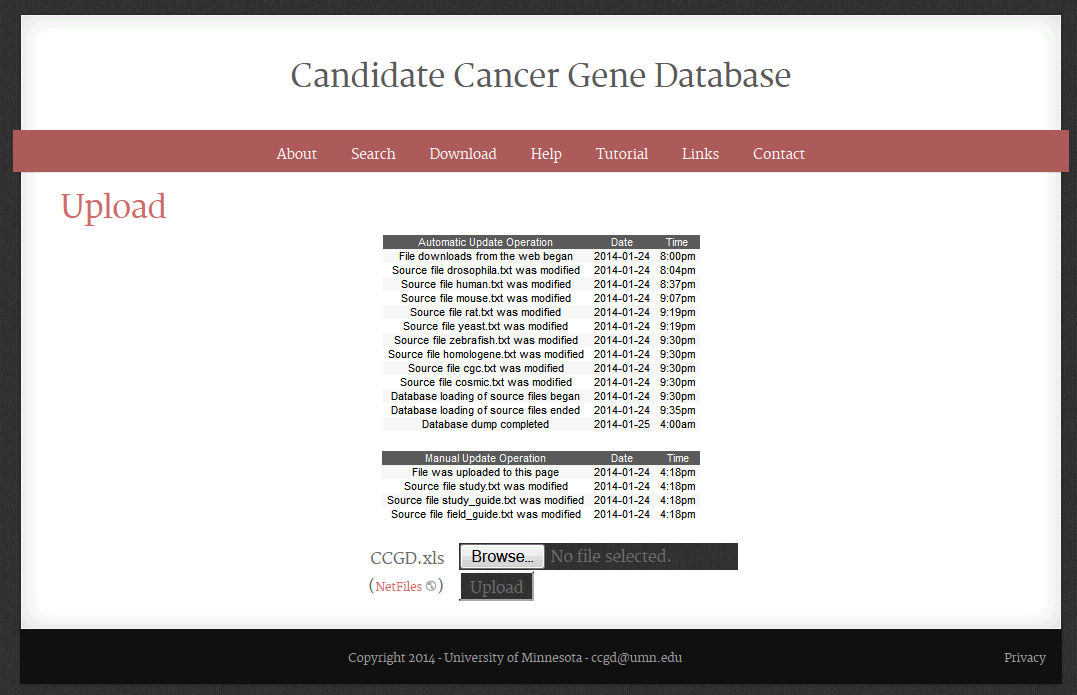
2. Go to: <http://ccgd-starrlab.oit.umn.edu/login.php>

3. Enter username and password below:

Username - ccgd

Password – Candidate!Genes

4. Browse for the correct file and once selected, hit Upload. Verification of the upload should appear on the screen, as well as a log of past uploads.



# Accessing the CCGD server

The CCGD server resides on the OIT server. The server name is:

ccgd-starrlab.oit.umn.edu

Before accessing the server, I had to set up a VPN using the CISCO AnyConnect Secure Mobility Client. I have to enter my x500 and corresponding password.

## Access the CCGD server using Terminal Window

Generally, to access the server you need to use Duo authentication, which means they call your cell phone or leave a code on your cell phone.

Open the terminal window and type

tim-starrs-mbp:~ star0044old$ ssh star0044@ccgd-starrlab.oit.umn.edu

Password:

Enter your x-500 password, then you will see the following

Duo two-factor login for star004

Enter a passcode or select one of the following options:

1. Duo Push to XXX-XXX-1118

2. Phone call to XXX-XXX-1118

3. SMS passcodes to XXX-XXX-1118 (next code starts with: 1)

Passcode or option (1-3): 2

At this point I am using option "2". My cell phone will ring and a female voice will ask me to hit any key. Once I hit the key, it logs me in

Success. Logging you in...

Last login: Wed Jun 17 22:37:10 2015 from x-10-21-37-148.vpn.umn.edu

[star0044@ccgd-starrlab ~]$

## Accessing the server using FUGU

Initiate a VPN connection using Cisco AnyConnect Secure Mobility Client

In the FUGU SFTP window enter the following:

Connect to: ccgd-starrlab.oit.umn.edu

Username: star0044

Port: 22

This works without Duo authentication.

In order to get this to work I did the following:

I followed the instructions that Ken Abbott wrote in the CCGD manual starting on page 48 (Production Environment > Configuration > SSH).

Generated a key for accessing the CCGD server using FUGU

tim-starrs-mbp:~ star0044old$ ssh-keygen -t rsa -b 1024

Generating public/private rsa key pair.

Enter file in which to save the key (/Users/star0044old/.ssh/id\_rsa):

Enter passphrase (empty for no passphrase):

Enter same passphrase again:

Your identification has been saved in /Users/star0044old/.ssh/id\_rsa.

Your public key has been saved in /Users/star0044old/.ssh/id\_rsa.pub.

The key fingerprint is:

45:1e:c7:4d:46:6a:e9:b3:f0:85:2a:50:9e:fe:1e:6d star0044old@tim-starrs-mbp

The key's randomart image is:

+--[ RSA 1024]----+

| o..++ |

| o o.+. |

| . o + |

| o o o . |

| . S . + . |

| o = + |

| o o E |

| o o |

| .o |

+-----------------+

tim-starrs-mbp:~ star0044old$

Then I followed the instructions on page 48 to set up the key.

The key is found in the file id\_rsa.pub in the /Users/star0044old/.ssh/ folder

ssh-rsa AAAAB3NzaC1yc2EAAAADAQABAAAAgQChiooaj60GYfmRAN7/9Q1LHNTkyD0FIpCvGd8YQaAfEnJOOQ/sOkIixMmedgtSvU5PAW096fL6MlMqyeH7v9VZxPn8gwEoyce/Z7mj8bAsMuZsdYvUaHjFlGagG6wInZ2lr8D7kQ3Avtv5M90m1j6PZfE5E5ELLg1doYww6AGTjQ== star0044old@tim-starrs-mbp

x-160-94-122-229:.ssh star0044old$

To login to the CCGD server using FUGU I entered the following in FUGU

Connect to: ccgd-starrlab.oit.umn.edu

Username: star0044

Port: 22

Note: Originally Ken Abbott said I should put the following command in the advanced SFTP options box. But I don't need to do this and I still am able to get connected

-i /Users/star0044old/.ssh/id\_rsa.pub

To get it to work I had to change some of the permissions using chmod in the terminal window. Here are the permissions that work

[star0044@ccgd-starrlab ~]$ ls -la

total 40

drwx------ 3 star0044 star0044 4096 Jun 17 22:42 .

drwxr-xr-x. 22 root root 4096 May 31 08:10 ..

-rw------- 1 star0044 star0044 422 Jun 22 09:21 .bash\_history

-rw-r--r-- 1 star0044 star0044 18 Aug 29 2012 .bash\_logout

-rw-r--r-- 1 star0044 star0044 176 Aug 29 2012 .bash\_profile

-rw-r--r-- 1 star0044 star0044 124 Aug 29 2012 .bashrc

-rw-r--r-- 1 star0044 star0044 500 Feb 27 2012 .emacs

-rw-r--r-- 1 star0044 star0044 121 Apr 11 2013 .kshrc

drwx------ 2 star0044 star0044 4096 Jun 17 22:40 .ssh

-rw------- 1 star0044 star0044 644 Jun 17 22:40 .viminfo

[star0044@ccgd-starrlab ~]$ ls -la .ssh

total 12

drwx------ 2 star0044 star0044 4096 Jun 17 22:40 .

drwx------ 3 star0044 star0044 4096 Jun 17 22:42 ..

-rw------- 1 star0044 star0044 240 Jun 17 22:40 authorized\_keys

[star0044@ccgd-starrlab ~]$ cd ..

[star0044@ccgd-starrlab home]$ ls -la

total 100

drwxr-xr-x. 22 root root 4096 May 31 08:10 .

dr-xr-xr-x. 26 root root 4096 May 31 08:10 ..

drwx------ 5 abbot195 abbot195 4096 Jun 22 07:07 abbot195

drwx------. 3 amun0155 amun0155 4096 Jun 19 2013 amun0155

drwx------. 2 ander662 ander662 4096 Jun 21 2011 ander662

drwx------. 3 baker074 baker074 4096 Dec 17 2013 baker074

drwx------. 3 bogg0010 bogg0010 4096 Jul 2 2013 bogg0010

drwx------. 3 brown345 brown345 4096 Aug 26 2011 brown345

drwx------. 3 dack dack 4096 Jan 5 2014 dack

drwx------. 3 jschnide jschnide 4096 Nov 11 2011 jschnide

drwx------. 3 515 515 4096 May 11 14:31 lixxx211

drwx------. 2 root root 16384 Jun 19 2013 lost+found

drwx------. 2 lxadm lxadm 4096 Jun 21 2011 lxadm

drwx------. 3 mejh mejh 4096 Nov 27 2013 mejh

drwx------. 2 oracle oinstall 4096 Jun 19 2013 oracle

drwx------. 3 p-vitk p-vitk 4096 Nov 13 2013 p-vitk

drwx------. 4 schnide schnide 4096 Dec 12 2013 schnide

drwx------ 3 shouston shouston 4096 Apr 24 2014 shouston

drwx------ 3 star0044 star0044 4096 Jun 23 16:51 star0044

drwx------. 5 swadm swadm 4096 Jun 4 09:38 swadm

drwx------. 3 tbrown tbrown 4096 Jun 21 2011 tbrown

drwx------. 3 tkunz tkunz 4096 Jul 15 2013 tkunz