# The current status of the human DREAM/Calsenilin/KChIP3 interactome: focus on arrays for the screening of novel protein interactions.

Casado-Vela, J.[1]; Fuentes, M.[2]; Dopazo, J.[1]; Perez, P. [1]; Dasilva N. [2]; Gonzalez-Gonzalez M [2]; Mellström, B.[1]; Naranjo, J.R.[1]

[1]Centro Nacional de Biotecnología (CNB). Darwin 3, Campus de Cantoblanco. Spanish National Research Council (CSIC). 28049 Madrid, Spain.
[2]Centro de Investigación del Cáncer de Salamanca, IBSAL, University of Salamanca-Spanish National Research Council (CSIC). 37007 Salamanca, Spain.

## Introduction.

The human **DREAM/Calsenilin/KChIP3** is a neuronal protein that localizes to three different cellular compartments (membrane, cytoplasm and nucleus). In the membrane DREAM plays a role as a K channel (KChIP3). In the cytosol it acts as a $Ca^{2+}$-binding protein [1] and may also shuttle to the nucleus, where it acts as a gene silencer of DRE-responsive genes. At structural level, DREAM harbors four EF-hand domains binding $Ca^{+2}$ ions that provoke structural changes that may lead to changes in its protein-binding capabilities. Thus, it is foreseeable that the DREAM interactome significantly varies depending on both: sub-cellular localization and $Ca^{+2}$ concentration.

In this report we show a multi-step workflow developed in our lab that facilitates the task of discerning novel from previous information on protein-protein interactions. The applicability of this workflow is demonstrated here through its application to address the study of the **DREAM/Calsenilin/KChIP3** interactome.

For the screening of previously published information on protein interactomes we combined database searches and literature mining using three multiple web-based platforms. The bulk of information on protein interactions deposited in databases and peer-reviewed published literature is constantly growing. Multiple databases interfaced from user-friendly web tools recently emerged to facilitate the task of protein interaction data retrieval and data integration. Nevertheless, we evidenced that the information retrieved by *in-silico* approaches is not error-free. Indeed, the probability of including false protein interactions using gene symbols is relatively high. The main cause acting as a source of errors is the occurrence of gene **acronyms redundancy**.

At experimental level, for the identification of novel **DREAM/Calsenilin/KChIP3** interactions were we used Nucleic Acid Programmable Protein Arrays [2, 3] incubated with purified recombinant DREAM Protein arrays enabled not only large screening of protein interactors [4] but also differential screening of interaction changes due to changes in buffer composition. This is a key feature to understand DREAM interactions, since this protein is strongly affected by $Ca^{+2}$.

## Results.

### A. Combination of experimental and *in-silico* database strategies facilitating the screening of protein interactions.

A multi-step workflow developed in our lab to screen for novel and previously describe protein interactions is shown in **Figure 1**. Our strategy consists on **(i)** retrieval of interactions previously described in the literature and *in-silico* databases **(ii)** gathering experimental information (in this case using **Nucleic Acid Programmable Protein Arrays –NAPPA- arrays**, and **(iii)** data curation and integration. This strategy proposed here was applied to compile *bona fide* information on human DREAM protein interactome, which constitutes liable training datasets that can be used to improve computational predictions. More interestingly, this strategy is general and can also be applied to compile interactomic information from other proteins of interest.

### B. The redundancy of gene acronyms.

In previous reports [2, 4, 5] we demonstrated that one gene acronym may refer to different genes or gene products and such redundancy may lead to errors when identifying protein-protein interactions through automated database searches. Thus, we hypothesized that redundancy leads to ambiguity that constitutes the causal origin of mistakes, introducing erroneous protein interactors through *in-silico* searches. Therefore, we were prompted to investigate the frequency of gene acronym redundancy and its effect on the identification of protein-protein interactions. To that aim, we built a file containing the description and the gene acronyms of the **20,488 human protein-coding genes** (according to the HUGO gene nomenclature: www.genenames.org). It is important to note that acronyms from the nomenclature authority and synonyms of each gene found in the literature are also included in the same file. The full list of human gene acronyms can be downloaded in from our lab data repository (www.cnb.csic.es/~naranjo/), which constitutes the best resource for measuring the frequency of redundancy. We calculated the redundancy of gene names and synonyms and plotted the number of names referencing N different genes versus N genes referred (**Figure 2**).

Interestingly, **the repetition of gene acronyms is a frequent event** and a significant portion of the genes displayed redundancies ranging from two and up to ten. This plot evidences that one gene acronym may designate multiple genes and/or protein. Importantly, a significant percentage of gene acronyms may refer to two or more different genes. Obviously, such redundancy leads to undesired ambiguities and errors introduced by database retrieval algorithms, which are unable to discern the attribution of a gene acronym to a certain gene or protein product. As explained above, the lack of consensus regarding the use of gene symbols may constitute a significant hurdle in the process of retrieving *bona-fide* protein interactions from public repositories. As a result, we show that **the probability of including false protein interactions after *in-silico* searches using gene symbols is relatively high** (see [5] for further information and dedicated examples). This means that it is probable that using gene abbreviations as the only information included in database searches may end up showing interactions that do not strictly correspond to the protein of interest. Consequently, in order to discard potential errors, manual curation of the list of interactors retrieved using gene symbols against published data seems highly recommendable, especially in those cases where the frequency of the gene symbols used for a protein in the literature is ≥2.

Probably the best way to circumvent the ambiguity problem caused by gene acronym redundancies would be using consensus lists of gene acronyms exclusively attributed to single genes (not shared by any other gene). Nevertheless, as exemplified above, **manual curation of metadata seems necessary to preclude unnecessary errors**. In the meanwhile, we provide here an open-access standalone software tool termed **'Gene Symbol redundancy checker'** to facilitate validation of gene redundancies (https://dl.dropboxusercontent.com/u/77276631/SymbolRedundancy.zip). This software runs under Windows and calculates the degree of symbol redundancy for one or more genes of interest. This can of example be a list of genes in a network provided by an online tool. The output also summarizes all alternative symbols for all the genes that share a given gene symbol. The output can be subsequently exported or copy-pasted to any spreadsheet data processing software for further analysis.

### C. The current DREAM/CALSENILIN interactome fails to include the dynamic changes on protein networks.

The **retrieval of the overall list of interactors for a specific protein of interest is basic to discern novel from previous findings** and also serves to identify the pathways involved. In sections above we quoted some problems affecting the number and quality of interactors retrieved by searching in multiple databases and the need to complete and curate that information by comparing the list of interactors with published literature. However, the interactome of a given protein of interest is not static and adapts to changes in the environmental conditions. Thus, understanding the dynamics or protein interaction networks is crucial to unravel the role and the regulation of proteins under different cellular conditions [6]. Here we briefly point to the fact that *in-silico* database searches typically fail to provide information on protein interaction changes as a response to modifications in the experimental conditions.

It is important to underline that **none of the databases tested**-including 22 databases focused on protein-protein interactions available at http://www.pathguide.org/ [7]- or integrated web-based platforms like PSICQUIC [8], DASMI [9] and BIPS [10] **offered direct information about DREAM interactome dynamics**, changes due to changes in $Ca^{+2}$ concentration or provided clues on factors modifying or affecting DREAM networking. Conversely, a good deal of information can be extracted from published literature (for review see [11]). **Thus, we decided to compile the current status of DREAM interactome (Table 1)** including supporting references, year of publication, *in-vitro* and *in-vivo* models used, human gene acronyms (including synonyms), UniprotKB/Swiss-Prot accessions and entries, description, techniques used for detection of the interactions. The potential effect of $Ca^{+2}$ and other post-translational modifications on DREAM interactome is also included.

The bulk of information presented here serves as a basis for the screening of novel binary DREAM interactors. In this regard, we carried out NAPPA experiments comparing the interactome of soluble GST-DREAM versus GST-tag used as a control (**Figure 3**). Preliminary data led to the identification more than 35 of novel DREAM interactions using this approach.

## Conclusions.

1. Understanding protein interaction networks and their dynamic changes still constitutes a major challenge in modern biology.

2. Despite the current efforts towards data integration, the quality of the information on protein interactions retrieved by *in-silico* approaches is frequently incomplete and may even list false interactions.
3. We showed that the probability of including false protein interactions after *in-silico* searches using gene symbols is relatively high. We provide here an open-access standalone bioinformatic tool termed **'Gene Symbol redundancy checker'** that facilitates validation of gene redundancies.
4. Updated information on DREAM interactions currently reported was compiled and false positives due to redundancies in gene acronyms were removed and manually curated. Thus, the **DREAM/Calsenilin/KChIP3** interactome reported here serves as a *bona-fide* 'training-set' for future improvements of protein-protein prediction algorithms.
5. It is important to underline that none of the databases tested or integrated web-based platforms used offered direct information about DREAM interactome dynamics, changes due to changes in Ca$^{+2}$ concentration or provided clues on factors modifying or affecting DREAM networking.
6. Further details on our platform and related literature can be found at **www.cnb.csic.es/~bmyc/protein_array_platform.html**.
7. High-density protein arrays for the screening of protein interactions and ourbioinformatic platform is currently offered to researchers interested in this field.

**Web resources and bioinformatic tools.**

**Gene Symbol redundancy checker tool**: www.cnb.csic.es/~bmyc/naranjo.html.

**HUGO gene nomenclature**: www.genenames.org.

**Lab data repository** (http://www.hupo.org/2013/)

www.pathguide.org/

www.cnb.csic.es/~bmyc/protein_array_platform.html

**Table 1.** Human DREAM interactome: publication year, experimental model used, corresponding human gene acronyms, Swiss-Prot protein accessions, protein entries, protein descriptions and synonyms, detection method used and other additional relevant information is also included.

**Figure 1.** Multistep workflow consisting on **(i)** retrieval of interactions previously described in the literature and *in-silico* databases **(ii)** gathering experimental information (in this case using NAPPA arrays, and **(iii)** data curation and integration.

The strategy proposed above was applied to compile *bona fide* information on human DREAM protein interactome. The zoomed region on the array of proteins represents a schematic view of the protein synthesis taking place directly on the surface of cell-free protein arrays.

**Figure 2. Gene acronym redundancy checker**. Barplot demonstrating the redundancy displayed by human gene acronyms and their synonyms.

**Figure 3.** Schematic view of protein arrays enabling the screening of binary protein interactors. High-density protein array showing 23,232 different positions (22×22×48). (A) A regular slide (7.5×2.5 cm) may harbor >9000 human proteins spotted in duplicates plus controls. The array on the left was probed with GST-DREAM and the panel of interactors revealed after the addition of primary and secondary antibody cocktail against DREAM. The array on the right corresponds to a control experiment probed with GST. (B) Close-up view (~1.0×1.0 cm and 22×22 different positions) corresponds to a single sub-array sections showing the occurrence green spots (Cy3 fluorescence on specific proteins printed on the array surface). Differential image analysis showed the novel identification of a novel DREAM interaction not present in GST controls. Positive and negative controls, buffer, BSA and GST were also printed in each sub-array.

**Acknowledgements.** J. Casado-Vela is a JAE-DOC (CSIC) holder supported by Ministerio de Economía y Competitividad (Spain) and the European Social Fund (European Commission). This work is funded by FIS02114 and JCYLSA198-A12-2.

## References.

[1] Carrion, A. M., Link, W. A., Ledo, F., Mellstrom, B., Naranjo, J. R., DREAM is a Ca2+-regulated transcriptional repressor. *Nature* 1999, *398*, 80-84.

[2] Casado-Vela, J., Cebrian, A., Gomez del Pulgar, M. T., Sanchez-Lopez, E.*, et al.*, Lights and shadows of proteomic technologies for the study of protein species including isoforms, splicing variants and protein post-translational modifications. *Proteomics* 2011, *11*, 590-603.

[3] Casado-Vela, J., Gonzalez-Gonzalez, M., Matarraz, S., Martinez-Esteso, M.*, et al.*, Protein arrays: recent achievements and their application to study the human proteome. . *Current Proteomics* 2013, *10*, 83-87.

[4] Casado-Vela, J., Cebrian, A., Gomez del Pulgar, M. T., Lacal, J. C., Approaches for the study of cancer: towards the integration of genomics, proteomics and metabolomics. *Clinical & translational oncology*. 2011, *13*, 617-628.

[5] Casado-Vela, J., Matthiesen, R., Sellés, S., Naranjo, J. R., Protein-protein interactions: gene acronym redundancies and current limitations precluding fully automated data integration. . *Proteomes* 2013, *1*, 3-24.

[6] Hegde, S. R., Manimaran, P., Mande, S. C., Dynamic changes in protein functional linkage networks revealed by integration with gene expression data. *PLoS computational biology* 2008, *4*, e1000237.

[7] Klingstrom, T., Plewczynski, D., Protein-protein interaction and pathway databases, a graphical review. *Briefings in bioinformatics* 2011, *12*, 702-713.

[8] Aranda, B., Blankenburg, H., Kerrien, S., Brinkman, F. S.*, et al.*, PSICQUIC and PSISCORE: accessing and scoring molecular interactions. *Nat Methods* 2011, *8*, 528-529.

[9] Blankenburg, H., Finn, R. D., Prlic, A., Jenkinson, A. M.*, et al.*, DASMI: exchanging, annotating and assessing molecular interaction data. *Bioinformatics* 2009, *25*, 1321-1328.

[10] Garcia-Garcia, J., Schleker, S., Klein-Seetharaman, J., Oliva, B., BIPS: BIANA Interolog Prediction Server. A tool for protein-protein interaction inference. *Nucleic acids research* 2012, *40*, W147-151.

[11] Rivas, M., Villar, D., Gonzalez, P., Dopazo, X. M.*, et al.*, Building the DREAM interactome. *Science China. Life sciences* 2011, *54*, 786-792.