Final Progress Report for Research Projects Funded by Health Research Grants

Instructions: Please complete all of the items as instructed. Do not delete instructions. Do not leave any items blank; responses must be provided for all items. If your response to an item is "None", please specify "None" as your response. "Not applicable" is not an acceptable response for any of the items. There is no limit to the length of your response to any question. Responses should be single-spaced, no smaller than 12-point type. The report **must be completed using MS Word**. Submitted reports must be Word documents; they should not be converted to pdf format. Questions? Contact Health Research Program staff at 717-231-2825.

1. **Grantee Institution:** The Pennsylvania State University

2. **Reporting Period (start and end date of grant award period):** 1/1/2011 - 12/31/2014

3. **Grant Contact Person (First Name, M.I., Last Name, Degrees):** John Anthony, MPA

4. **Grant Contact Person's Telephone Number:** 814 935 1081

5. **Grant SAP Number:** 4100054865

6. **Project Number and Title of Research Project:** 8. *Resolving the Structure and Function of Proteomes*

7. **Start and End Date of Research Project:** 1/1/2011 – 12/31/2014

8. **Name of Principal Investigator for the Research Project:** Damian B. van Rossum, Ph.D.

9. **Research Project Expenses**.

   9(A) Please provide the total amount of health research grant funds spent on this project for the entire duration of the grant, including indirect costs and any interest earned that was spent:

   Total: $1,085,134

   9(B) Provide the last names (include first initial if multiple individuals with the same last name are listed) of **all** persons who worked on this research project and were supported with health research funds. Include position titles (Principal Investigator, Graduate Assistant, Post-doctoral Fellow, etc.), percent of effort on project and total health research funds expended for the position. For multiple year projects, if percent of effort varied from year to year, report in the % of Effort column the effort by year 1, 2, 3, etc. of the project (x% Yr 1; z% Yr 2-3).

Last Name, First Name	Position Title	% of Effort on Project	Cost
Babu, Jagadish	Research Assistant	100%	$97,203.23 (Sal+FB+FA)
Ho, Ngai Lam	Graduate Student	25% Part of Year 1 50% Year 2	$7,851.28 (Sal+FB+FA) $55,799.75 (Sal+FB+FA)
van Rossum, Damian	PI, Research Assistant Professor	12%	$17,409.16 (Sal+FB+FA)

9(C) Provide the names of **all** persons who worked on this research project, but who *were not* supported with health research funds. Include position titles (Research Assistant, Administrative Assistant, etc.) and percent of effort on project. For multiple year projects, if percent of effort varied from year to year, report in the % of Effort column the effort by year 1, 2, 3, etc. of the project (x% Yr 1; z% Yr 2-3).

Last Name, First Name	Position Title	% of Effort on Project
Anishkin, Andriy	Research Associate	25%
Chintapalli, Sree	Research Associate	12%

9(D) Provide a list of **all** scientific equipment purchased as part of this research grant, a short description of the value (benefit) derived by the institution from this equipment, and the cost of the equipment.

Type of Scientific Equipment	Value Derived	Cost
Dionex Ultimate 3500RS nano pump and Top level, LTQ Orbitrap Velos	Top-of-the-line MS instruments that are capable of identifying thousands of proteins from a single sample.	$730,255
Zeiss PALM Laser Capture Microdissection microscope	Images a tissue slice and is used to cut out mm-size areas of interest from the slice using a laser. These areas of interest can be collected in a microcentrifuge tube and analyzed by proteomics, for example.	$180,000

10. Co-funding of Research Project during Health Research Grant Award Period. Did this research project receive funding from any other source during the project period when it was supported by the health research grant?

Yes____X_____ No_____

If yes, please indicate the source and amount of other funds:

Six months of support ($20,220 Sal) for Research Associate Sree V. Chintapalli, Ph.D. was derived from NIH RO1 GM107652-01, a grant which was leveraged from the Health Research Grant Award (described below).

11. Leveraging of Additional Funds

11(A) As a result of the health research funds provided for this research project, were you able to apply for and/or obtain funding from other sources to continue or expand the research?

Yes____X___ No_____

If yes, please list the applications submitted (column A), the funding agency (National Institutes of Health—NIH, or other source in column B), the month and year when the application was submitted (column C), and the amount of funds requested (column D). If you have received a notice that the grant will be funded, please indicate the amount of funds to be awarded (column E). If the grant was not funded, insert "not funded" in column E.

Do not include funding from your own institution or from CURE (tobacco settlement funds). Do not include grants submitted prior to the start date of the grant as shown in Question 2. If you list grants submitted within 1-6 months of the start date of this grant, add a statement below the table indicating how the data/results from this project were used to secure that grant.

A. Title of research project on grant application	B. Funding agency (check those that apply)	C. Month and Year Submitted	D. Amount of funds requested:	E. Amount of funds awarded:
"The bacterial mechanosentitive channel as a multimodal sensor device" R01 GM107652-01	X NIH ☐ Other federal (specify:_____) ☐ Nonfederal source (specify:_)	May 2012	$ 1,583,172 Total Direct and Indirect costs for PSU $440,466	$1,095,259
"The tension coupling mechanism in the mechanosensitive channel MscS" R01 GM106298-01	X NIH ☐ Other federal (specify:_____) ☐ Nonfederal source (specify:_)	May 2012	$1,386,066 Total Direct and Indirect costs for PSU $458,000	NOT FUNDED

"Spatial representation and enhanced analytical inference of high throughput cancer-related data from genomic, proteomic, metabolomic, and pharmacological screens."	☐NIH ☐ Other federal (specify:_____) X Nonfederal source (specify: NVIDIA FOUNDATION Compute the Cure Project Proposal)	Oct 2013	$200,000	NOT FUNDED

11(B) Are you planning to apply for additional funding in the future to continue or expand the research?

Yes____X_____ No _____

If yes, please describe your plans:

We are planning to submit grants to extend the Spatial Dynamic 3D positioning approach described below and in Section 17.

12. Future of Research Project. What are the future plans for this research project?

The investment made from the Health Research Funds for the top-of-the-line MS instrumentation will continue to pay dividends in the future in the form of educational outreach, research collaborations, and routine analyses performed by the excellent support staff of the Penn State MS core-facility.

For the Spatial Dynamic 3D positioning approach described in Section 17, we will use Tcl scripting, as a very flexible cross-platform tool, to develop software (both in standalone and web service versions) for automated preparation of the raw data for spatial representation, submission of the computational jobs to molecular simulation programs on supercomputer clusters (when needed), and preparation of the visualization results (using molecular visualization software) in formats suitable for publication and presentations (including movies and embeddable 3D structures).

There is always a question how much of the user time and learning is required to use a visualization approach. Quite often, with researcher being focused on another field (e.g. clinical therapy) there is so much commitment needed that it prevents the researcher from attempting the new visualization. To avoid this, our important contribution will be developing very straightforward and universal user interface, with high degree of automation in raw data preparation, suggesting the visualization parameters, submitting the job to spatial positioning and visualization software, up to the final step of publication-quality presentation of the results in the common formats. The user should be able to obtain meaningful results, while starting just with the typical raw data and minimal knowledge of this visualization approach. At the same time, when needed we will provide complete transparency and option

of complete customization.

There are several competing approaches to visualize gene expression arrays, including “heat maps”, clustering trees, even representation as a network of spatial clusters, each with their own benefits and drawbacks. They typically struggle with both visual representation and speed at transition to larger datasets, require significant experience from the user, depend on numerous arbitrary parameters, and often “overprocess” data with emphasized clustering and loss of important details. The distinct features of our future approach would be:

- Ability to quickly process very large datasets due to the use of the existing parallelized MD software on supercomputers.

- Minimal processing of the data to reveal undistorted and detailed picture, including the ability to show continuous trends without the necessity to cluster.

- Enhanced representation of multidimensional data through optional dynamic representation such that time-averaged position approximate pairwise distances incompatible with a static 3D structure.

- Robustness of spatial representation facilitating compatibility of various modalities of data (e.g. combination of gene expression with pharmacological data and clinical information) and additivity of the spatial representations (addition of the new data do not distort significantly the existing representation).

- High degree of automation allowing users with no experience in data clustering or visualization to benefit from visualization of their complex datasets.

- Flexibility and transparency of settings allowing complete customization of the spatial positioning.

- Access to powerful capabilities of the existing molecular visualization software.

- Choice of a standalone software or a web server interface (to avoid the need for software installation).

- Automated production of various types of visualization material – 3D representations embedded in Adobe Acrobat PDF, movies, flash animations, optimized 2D projections etc.

- For users with accounts on supercomputers (e.g. resources of their universities) – automated submission of the computational tasks to their clusters.

A high-level timeline for actualization of this proposal is as follows:
Optimization (memory use, performance) of the current Tcl-based algorithm for larger datasets (4-months)

Adaptation of the approach to the format compatible with the leading parallelized MD software (e.g. NAMD, GROMACS) (8-months)

Development of software for automated conversion of input data to simulation-ready format (6-months)

Development of the software for automated presentation of the spatial results in the convenient formats compatible with the common software (molecular structure viewers, VRML viewers, 3D objects embeddable into Adobe Acrobat PDF files etc.) (3-months)

Development of the web server integrating automated data preparation, submission to the clusters, and visualization (3-months)

13. **New Investigator Training and Development**. Did students participate in project supported internships or graduate or post-graduate training for at least one semester or one summer?

Yes X____ No_____

If yes, how many students? Please specify in the tables below:

	Undergraduate	Masters	Pre-doc	Post-doc
Male		2		
Female				
Unknown				
Total		**2**		

	Undergraduate	Masters	Pre-doc	Post-doc
Hispanic				
Non-Hispanic		2		
Unknown				
Total		**2**		

	Undergraduate	Masters	Pre-doc	Post-doc
White				
Black				
Asian		1		
Other		1		
Unknown				
Total		**2**		

14. Recruitment of Out-of–State Researchers. Did you bring researchers into Pennsylvania to carry out this research project?

Yes_____ No X

If yes, please list the name and degree of each researcher and his/her previous affiliation:

15. Impact on Research Capacity and Quality. Did the health research project enhance the quality and/or capacity of research at your institution?

Yes X____ No_____

If yes, describe how improvements in infrastructure, the addition of new investigators, and other resources have led to more and better research.

Prior to the acquisition of the LTQ Orbitrap Velos ETD, the majority of established faculty laboratories were using proteomics services outside of Penn State as the proteomics support was unavailable to the University Park researchers. As of January 2015, 50 research groups at Penn State are using Orbitrap MS services, of these 12 groups are using the Orbitrap on a regular basis; 3 groups have their own MS instruments but require Orbitrap for high-resolution, high-mass-accuracy analyses; 5 groups have established a collaborative relationship with the Core Facility and depend on the Orbitrap for completing their externally-funded research projects or for securing new external funding (see Table 1 at the end of Question 17).

Penn State's primary mission is education, and the Proteomics Core Facility has truly become a center for learning and exchanging ideas for students, faculty, and the Facility staff. In addition to providing access to the advanced MS instrumentation, the Facility staff offers expert input in experiment design and implementation, data processing and analysis. The Facility's website (http://sites.psu.edu/msproteomics/) provides updates on the instrument and analysis status, posts standard operating procedures, sample preparation tips, instructions on data interpretation, and other technical information in a user-friendly, mobile-device-compatible format. In 2012 and 2014, the Core Facility hosted the high-resolution MS practicum as part of biannual Bioinorganic workshop held at Penn State. The 5-day, 2 sessions per day hands-on workshop demonstrated the utility of the Orbitrap's high resolving power in distinguishing between an apo and holo states of a metalloenzyme.

The acquisition of the Thermo LTQ Orbitrap Velos ETD positioned the Proteomics Core as the state-of-the-art instrumentation facility dedicated to education and actively promoting the use of advanced technology in the life sciences research. The well-equipped Core Facility gives Penn State a competitive edge in hiring world-class faculty and negotiating optimal start-up packages. This investment has already started paying in external funding and, most importantly, in the number of graduate students who can add the advanced MS-based proteomics technology to the list of their competencies.

16. Collaboration, business and community involvement.

16(A) Did the health research funds lead to collaboration with research partners outside of your institution (e.g., entire university, entire hospital system)?

Yes____X_____ No_____

If yes, please describe the collaborations:

The Health Research Funds led to collaboration and a leveraged subcontract RO1 grant award with Sergei I. Sukharev, Ph.D. (University of Maryland College Park). In addition, research was initiated and is ongoing with John Weinstein's group at The University of Texas, MD Anderson Cancer Center. This collaboration focuses on increased analytical inference of cancer-related multimodal datasets through 3D spatial dynamic representation and is described in Section 17.

16(B) Did the research project result in commercial development of any research products?

Yes_____ No ____X_____

If yes, please describe commercial development activities that resulted from the research project:

16(C) Did the research lead to new involvement with the community?

Yes_____ No ____X_____

If yes, please describe involvement with community groups that resulted from the research project:

17. Progress in Achieving Research Goals, Objectives and Aims.

List the project goals, objectives and specific aims (as contained in the grant agreement). Summarize the progress made in achieving these goals, objectives and aims for the period that the project was funded (i.e., from project start date through end date). Indicate whether or not each goal/objective/aim was achieved; if something was not achieved, note the reasons why. Describe the methods used. If changes were made to the research goals/objectives/aims, methods, design or timeline since the original grant application was submitted, please describe the changes. Provide detailed results of the project. Include evidence of the data that was generated and analyzed, and provide tables, graphs, and figures of the data. List published abstracts, poster presentations and scientific meeting presentations at the end of the summary of progress; peer-reviewed publications should be listed under item 20.

This response should be a DETAILED report of the methods and findings. It is not sufficient to state that the work was completed. Insufficient information may result in an unfavorable performance review, which may jeopardize future funding. If research findings are pending publication you must still include enough detail for the expert peer reviewers to evaluate the progress during the course of the project.

Health research grants funded under the Tobacco Settlement Act will be evaluated via a performance review by an expert panel of researchers and clinicians who will assess project work using this Final Progress Report, all project Annual Reports and the project's strategic plan. After the final performance review of each project is complete, approximately 12-16 months after the end of the grant, this Final Progress Report, as well as the Final Performance Review Report containing the comments of the expert review panel, and the grantee's written response to the Final Performance Review Report, will be posted on the CURE Web site.

There is no limit to the length of your response. Responses must be single-spaced below, no smaller than 12-point type. If you cut and paste text from a publication, be sure symbols print properly, e.g., the Greek symbol for alpha (α) and beta (ß) should not print as boxes (□) and include the appropriate citation(s). DO NOT DELETE THESE INSTRUCTIONS.

The overall goal of this project is to develop a world-class proteomics center at Penn State. By bridging the current strengths at PSU in computational biology (The Center for Computational Proteomics (CCP)) and the existing mass-spectrometry (MS) core-facility, an interdisciplinary center will be formed capable of high-throughput computational proteomics.

Specific Aim 1: Develop/Refine high-throughput computational approaches capable of accurate assessment of protein characteristics. Results obtained from the recent Critical Assessment of Techniques for Protein Structure Prediction (CASP) competition by CCP indicate that the Adaptive-BLAST algorithm developed by CCP is capable of performing high-throughput structural homology modeling; overcoming a major hurdle to proteomics research. However, proteins that have less than 30% identity with known protein folds are still extremely difficult to accurately predict and model. Our preliminary data demonstrates that with sufficiently large and divergent position specific scoring matrix (PSSM) libraries, that protein folds can be predicted with >90% accuracy. To obtain acceptable accuracy levels for all structural folds, structural libraries comprised of divergent fold-PSSMs will be generated and benchmarked. Computational and wet lab experiments will be performed to determine the accuracy of the libraries. CCP will also interact with the MS Center to develop all the necessary tools and infrastructure (i.e. website, wiki-base, etc.) such that the resources developed during the course of this project will be open access.

Specific Aim 2: Outfit the current MS core-facility with cutting-edge instrumentation/improved data collection services, and synergize the pipeline with our computational protein models. To excel at proteomics research, the technology within the current MS center must be updated. We plan to purchase a Thermo LTQ Orbitrap Velos

(ETD) mass spectrometer, a Dionex Ultimate 3000 nanoflow LC, and a Proxeon ion source; all of which are currently top-of-the-line instruments that are capable of identifying thousands of proteins from a single sample. Most importantly, we will work closely with the experimental biologists using the MS facility to translate our predictions to the benchtop. In these experiments we will critically evaluate our assessment of soluble protein fragments, ligand-binding pockets (e.g. protein-protein, protein-lipid, and/or protein-pharmacophore binding), and putative loss-of-function or gain-of-misfunction mutants. These combined aims will result in a pipeline which provides researchers with protein identification and cutting edge computational models.

Progress Achieved

Overall, both Specific Aims 1 and 2 were highly successful. Actualization of the project resulted in novel findings and mechanistic insight, therapeutic approaches related to human health and disease, new productive collaborations with faculty at Penn State and outreach to outside institutions, opportunities for educational experiences and training for students, and procurement of extramural funding. The following sections will describe the individual progress of each Specific Aim.

Specific Aim 1

While this research aim was successful from the perspective of multiple objective metrics, some of the original proposed experiments and analyses were refocused due to the loss of collaborators at UC Davis. However, we still pursued refined structural/functional models in extreme divergence as a key part of the research endeavors, with a specific focus on ion channels and protein-ligand interactions. In addition, we developed a computational method which can increase the visualization and analytic inference of multimodal data from genetic sequence datasets, massive high throughput screens and clinical trials, and people-parameter data, in general. This is a relevant, timely, and important tool for biomedical researchers and social scientists as there is a necessity for efficient high-resolution visualization of data, in its integrity, to reveal complex patterns and identify trends important for mechanistic insight and treatment of disease. Such an approach will likely have a substantial impact to the study of human health and disease in general.

Spatial Dynamic 3D Positioning

We are visual creatures. For centuries graphical tools have helped scientists grasp the essence of mathematic relationships. When we hear “parabola” we associate it with an image of a U-shaped curve rather than a dry table of x and y coordinates. However with the development of computers and high throughput experimental techniques (e.g. metabolomics, proteomics, genomics, lipidomics, transcriptomics, glycomics) biology has entered an era of “mega-data”. Instead of a few points relating two variables we have millions of numbers linking together thousands of independent variables. Conventional tools of simple plots and histograms are failing to visualize the big picture. To show a 1000-dimensional variable space, our best option is resorting to native 3D space- which the human brain has evolved to comprehend. A straightforward and minimally “invasive” way to map the multi-dimensional space is by use of pairwise distances between data-points, trying to preserve this mathematical measure during the transition from N to 3D as much as possible. Being a

challenging mathematical problem, computer simulations could start with a random placement of the points in 3D and iteratively optimize their positions until the pairwise distances approach the ideal "target" values of the source table. This is exactly the approach we developed to represent large scale, related data from genomic, proteomic, metabolomics, and pharmacological screens.

Our overreaching goal is to enable researchers to grasp the "big picture" of their raw data using straightforward, quick, and qualitative visual representation to see the general trends, outliers, to pinpoint the relationships for more rigorous statistical analysis, to generate hypotheses and explore "what if" scenarios by manipulating subgroups of data and interactively observing the response from the rest of the dataset linked by the experimentally defined relationships. In the course of the award period we made significant advances creating a universal tool that can be used in semi-automated mode, while still allowing the full control and customization of the visualization parameters when needed.

No step in the Scientific Method is as vital to the discovery process as visualization of the data space. Indeed, even the highest quality & error-free data, systematically derived with elegant controls and multiple independent experimental techniques is rendered essentially useless unless the results can be observed and related, intuitively and interactively, without overprocessing and oversimplifying of the data. Millions of research dollars and countless human hours are devoted to generate and analyze high throughput physiological measures of cancer-related screens and trials. Our approach can help maximize this investment by enabling scientists, through enhanced visualization and analytics, to better understand, illustrate, communicate, and derive meaningful insight into their data. This in turn will speed up the discovery process and lead to increased rate of successful treatment outcomes.

Currently, the visualization options for multimodal cancer-related data are neither sufficient enough nor efficient enough to handle the inherent magnitude and complexity. Take for example a larger dataset from Pan-Cancer project developed with participation of our collaborators at MD Anderson Cancer Center and other researchers of The Cancer Genome Atlas (http://cancergenome.nih.gov/). It contains expression levels of 181 genes in tissue samples of 12 types of cancer from 3467 patients, accompanied by 12 different fields of clinical information (~670,000 data points). What current tools or technology can meaningfully visualize such a massive dataset? Our computational method seeks to fill this critical void while accounting for the inherent mathematical issue representing multi-dimensional data in just 3D. Importantly, our approach includes a solution for more accurate and adequate representation of any number of dimensions in 3D. We can use TIME (i.e. motion) to represent all the relationships as TIME-AVERAGED pairwise distances. This feature improves the pairwise distance reproduction through motion as much as we, or any similar approach, have the limitation of just 3 dimensions.

Our approach is to visualize the relationships between all the variables (e.g. cancer patient, gene expression levels, tumor characteristics, responsiveness to treatment) through clear and straightforward 3D representation. Each variable is shown as a point in 3D space. The pairwise distance between two points is based on the numerical value linking them in the original input data. For example, distance between the point representing a particular patient

and the point representing a specific gene will be inversely proportional to the expression level of this gene in this patient. If two patients have very similar gene expression patterns they will tend to position close together, while different expression profiles will move them apart. To make the quantitative interpretation of the visual picture intuitive and straightforward the transformation formula between the raw data and spatial distance should be simple and straightforward. Currently we use linear transformation. However, in future implementations the transformation formula could be automatically suggested based on the statistical analysis of the input data and clearly communicated to user. At the same time, we would provide an option of a custom transformation when needed by user.

Through iterative search, the final 3D visualization attempts to preserve all the pairwise relationships from the input data as close as possible. Such a representation provides a comprehensive direct view of the raw data with minimal processing and distortion, with no simplification, under-sampling, artificial clustering etc. by default (they can be available as post-processing options). From our experience with applying this very approach to the gene expression in cancer (as well as in other fields, e.g. phylogenetic analysis, demographic data, behavioral characteristics), such a representation immediately reveals the structure of the data: whether there is a natural tendency to cluster (such as similar gene expression profiles) or there is a continuum distribution of the variables, whether there are some gradients in values (consistent gradual increase in survival with higher expression of certain genes, as shown on Figure. 1), which combinations are over- or under-sampled, conspicuous outliers, typical representatives of the groups, uniform samples to cover the whole range of data, as well as many others. Importantly, 3D visualization does not intend to replace the robust statistical analysis, but instead allows the selection of the data regions for further analysis and enables researchers to choose the most appropriate statistical technique based on visualization of the data patterns, distribution.

We intend visualization of the tens of thousands of linked data points positioned in 3D space. However, there is already an established field representation of similar arrangements – visualization of molecular structures. There are several leading programs capable of fast representation of up to millions of interconnected atoms. Currently we have implemented Tcl scripting for Visual Molecular Dynamics (www.ks.uiuc.edu/Research/vmd) because of its extremely flexible scripting interface, rich graphical tools, ability to use GPUs for extremely fast visualization, and popularity in the scientific community. Nevertheless, our use of the common molecular structure formats makes the visualization compatible with any molecular visualization software for convenient exploration.

Prior to visual representation, the most challenging part of the approach, especially for genome-wide datasets, is iterative calculation of the optimal 3D positions approximating pairwise relationships in the raw data. In our pilot implementation it was done using Tcl scripts in VMD for the ease of the development and portability. Importantly, all the restraints on pairwise distances were expressed as harmonic restraints – the same force law that governs inter-atom bonds in most of the molecular dynamics programs, which are specifically developed for dynamic spatial positing of the large (up to millions) linked points in space, with structural output directly compatible with molecular visualization software.

This enables us to adapt our approach for the use of existing MD software and strongly benefit from the scale and speed of calculations.

The leading MD programs (NAMD, GROMACS, Amber etc.) are developed and maintained by the top research and programming teams, the software is highly parallelized for the use of tens of thousands of CPUs and GPUs. They are already installed on the majority of supercomputer clusters available for scientific community and will undoubtedly grow in software and hardware capabilities. Therefore one of our future goals is to provide the protocol for spatial representation of massive multidimensional data arrays that would be readily compatible with the flagship resources for the molecular simulations. This will ensure that the approach will not be obsolete in quickly approaching era of whole-genome databases times thousands of samples times thousands of pharmacological substances.

Spatial Dynamic 3D Positioning of Clinical Data

Most relevant for the present Health Research Grant, we actively worked to enhance the visualization of and analytical inference within cancer-related data. We have an ongoing collaboration with Dr. John Weinstein's group at M.D. Anderson Cancer Center to analyze data sets from The Cancer Genome Atlas project. The TCGA collects, analyzes and distributes massive amounts of genomic and proteomic data for more than 20 types of cancer; a series of TCGA papers has been published in Nature Genetics. The data visualized on the included Figure 1 were derived from TCGA database, they contain reverse phase protein array (RPPA) data from 254 kidney cancer patients on 172 key proteins in various tissue samples (+/- phosphorylation). The data are integrated with 50 types of clinical information for those patients (e.g. tumor stage, survival status, age, gender etc.), thus producing ~56,000 data entries in total, relating 476 variables with each other. We have also performed visualization and exploratory analysis on another data set (not shown), relating effect of 788 gene silencing siRNA molecules on the response of the cultured ovarian cancer cells to three types of conditions testing resistance to anti-cancer therapeutic enzyme Type-II Asparaginase (7092 data points connecting 794 independent variables). Importantly, the researchers who generated the empirical data preferred the resultant 3D spatial representations over conventional 2D gene expression heatmaps in multiple analytical criteria. This demonstrates that further research and development is needed and warranted to refine both approach and implementation so that more researchers can actually use it and derive its inherent added value.

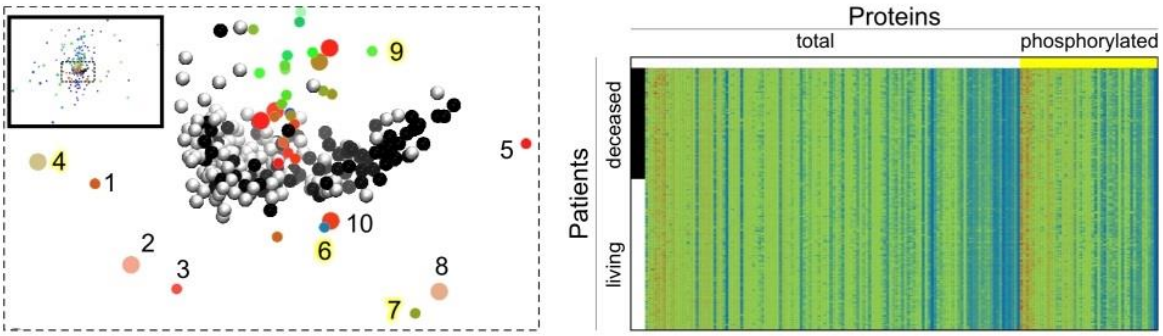


Figure 1 Legend: Meta-analyzed visualization of the protein expression and phosphorylation levels for kidney cancer patients using spatial positioning approach (left panel) comparing to the heatmap table (right panel) of the original data. Individual patients are represented as white (living) or black (deceased) spheres on 3D panel or rows in the heatmap. Proteins are presented as colored dots on the 3D panel or columns on the heatmap. Proteins are colored by the highest total expression (smaller dots) or phosphorylation (bigger dots) level from red (high), through green (medium) to blue (low) observed in patients. Spatial distance between the protein and the patient is inverse proportional to this value. 3D positioning was done only based on these protein levels, with no additional information used (e.g. the live/deceased status). The insert on 3D panel shows the whole "protein universe", the framed high expression region was used for the main 3D panel. The positioning reveals grouping of the patients by the live/deceased status and suggests the proteins whose high expression (#1 and 3) or phosphorylation (#2) correlates with the survival, while others (#5, 8, 10) are more typical for the deceased. Proteins with lower peak levels but consistent elevation over the patients pool (#4, 6, 7, 9) might be promising therapeutic targets and outcome prediction tools, although they can be easily overlooked by conventional search approaches.

Spatial Dynamic 3D Positioning of Sequence Data

We have used the same VMD-based approach as described above to represent sequence similarity networks in 3D and 2D space. Each sequence is shown as a point in space. The pairwise distance between two points is directly proportional to the numerical value linking them in the original input data of pairwise similarity. If two sequences are similar (low Evolutionary Divergence value) they will tend to position close together, if different (high Evolutionary Divergence) they will move apart. The benefit of such visualization is that the direct visual proximity between the sequences represents their similarity, whereas in conventional evolutionary trees one has to make a "round trip" to the closest common split point to estimate the proximity, and the contact region between the big branches often brings two remotely related sequences in close proximity, which is misleading.

To make the quantitative interpretation of the visual picture intuitive and simple the conversion formula between the raw Evolutionary Divergence data (M_{ij} , in our sample ranged from 0.04 to 4.046) and spatial distance (T_{ij}, measured in Å for 3D and 2D structures in molecular visualization software) should be straightforward. We used linear transformation: $T_{ij} = 1000 * M_{ij}$

It is impossible in a general case to reproduce exactly in 3D or 2D an arbitrary combination of pairwise distances from multi-dimensional space. Through iterative search, our final spatial visualization attempts to approximate all pairwise relationships from the input data as close as possible using the following algorithm. For every step, displacement of any given node in search for the best match of the spatial pairwise displacement was calculated as a vector sum of the displacements towards the target distance for each pair with the given node ($\vec{R}$ is the vector of the current position of the node in space, so the vector difference of two positions divided by its length represents a unit vector directed from the current node I to its neighbor j, n= number of sequences in representation):

$$\vec{D}_i = \frac{\sum_{j=1}^{n} \frac{(\vec{R}_j - \vec{R}_i)(|\vec{R}_j - \vec{R}_i| - T_{ij})}{|\vec{R}_j - \vec{R}_i|^2}}{\sum_{j=1}^{n} \frac{1}{|\vec{R}_j - \vec{R}_i|}}$$

This algorithm uses the close-neighbor weighting scheme that was found to reproduce well close- and long-range evolutionary relationships for our sample with moderate divergence. To provide a physical analogy, connection between all the nodes are analogous to harmonic springs or rubber threads of the same elasticity, with the length proportional to pairwise Evolutionary Divergence.

The effectiveness of this close-neighbor weighting scheme might arise for a number of reasons, including: (i) the information for spatial positioning is progressively more reliable as similarity increases, while remote relatives bear the highest fraction of noise in the derived distance values, (ii) the proper spatial positioning of remote relatives is more precise when the local closer- relative structure is built first (i.e. there is a hierarchy to the tree assembly), and/or (iii) the spatial positioning of remote families integrates information from numerous pairwise contacts of comparable distances, while the local structure is dominated by just a few close-neighbor relationships. Therefore with respect to the latter case, the under-weighting of local relationships could disrupt proper spatial positioning of the close-neighbors, while up-weighting close-relationships might resolve them without negatively affecting the spatial positioning of remote relatives.

We have used this approach to model a variety of protein families such as RecA/RAD51 Recombinases, RNA-dependent RNA polymerases, and multiple subfamilies of ion channels. As an example of the applicability of spatial dynamic 3D positioning for sequence data we show here the analysis for a diverse set of Shaker voltage-gated K^+ channels (see Figure 2 reproduced from the supplemental material of Li X, Liu H, Luo JC, Rhodes SA, Trigg LM, van Rossum DB, Anishkin A, Diatta FH, Sassic JK, Simmons DK, Kamel B, Medina M, Martindale MQ, Jegla T. A Major Diversification of Voltage-gated K^+ Channels Occurred in Ancestral Parahoxozoans. PNAS 2015 (In Press)).

The Evolutionary Divergence between the sequences was quantified as the number of amino acid substitutions per site from between sequences. Analyses of substitution pattern and rates were conducted using the JTT (Jones-Taylor-Thornton) matrix-based model (+G). The rate variation among sites was modeled with a gamma distribution. The analysis involved 134

amino acid sequences. All positions containing gaps and missing data were eliminated. There were a total of 240 positions in the final dataset. Evolutionary analyses were conducted in MEGA6. The estimated value of the shape parameter for the discrete Gamma Distribution was 1.0736. A discrete Gamma distribution was used to model evolutionary rate differences among sites (5 categories, [+G]). Mean evolutionary rates in these categories were 0.12, 0.38, 0.72, 1.23, 2.55 substitutions per site. The amino acid frequencies are 7.69% (A), 5.11% (R), 4.25% (N), 5.13% (D), 2.03% (C), 4.11% (Q), 6.18% (E), 7.47% (G), 2.30% (H), 5.26% (I), 9.11% (L), 5.95% (K), 2.34% (M), 4.05% (F), 5.05% (P), 6.82% (S), 5.85% (T), 1.43% (W), 3.23% (Y), and 6.64% (V). For estimating ML values, a tree topology was automatically computed. The maximum Log likelihood for this computation was -42455.645.

We repeated spatial positioning for 3D space five times, each one starting with random initial positions. All the simulations have produced topologically similar structures with reproducible features. As the final 3D structure, we chose the one with the lowest average deviation of the spatial distances from the Evolutionary Distance (2.04%). For that structure, we produced a matching 2D representation by continuing the same annealing with added restraint of using just two dimensions. This two-dimensional “map” allows easier view in a still image suitable for publication as the nodes do not visually overlap with each other, however it comes at a cost of somewhat worse approximation of the ideal set of Evolutionary Distance table (2.82% deviation).

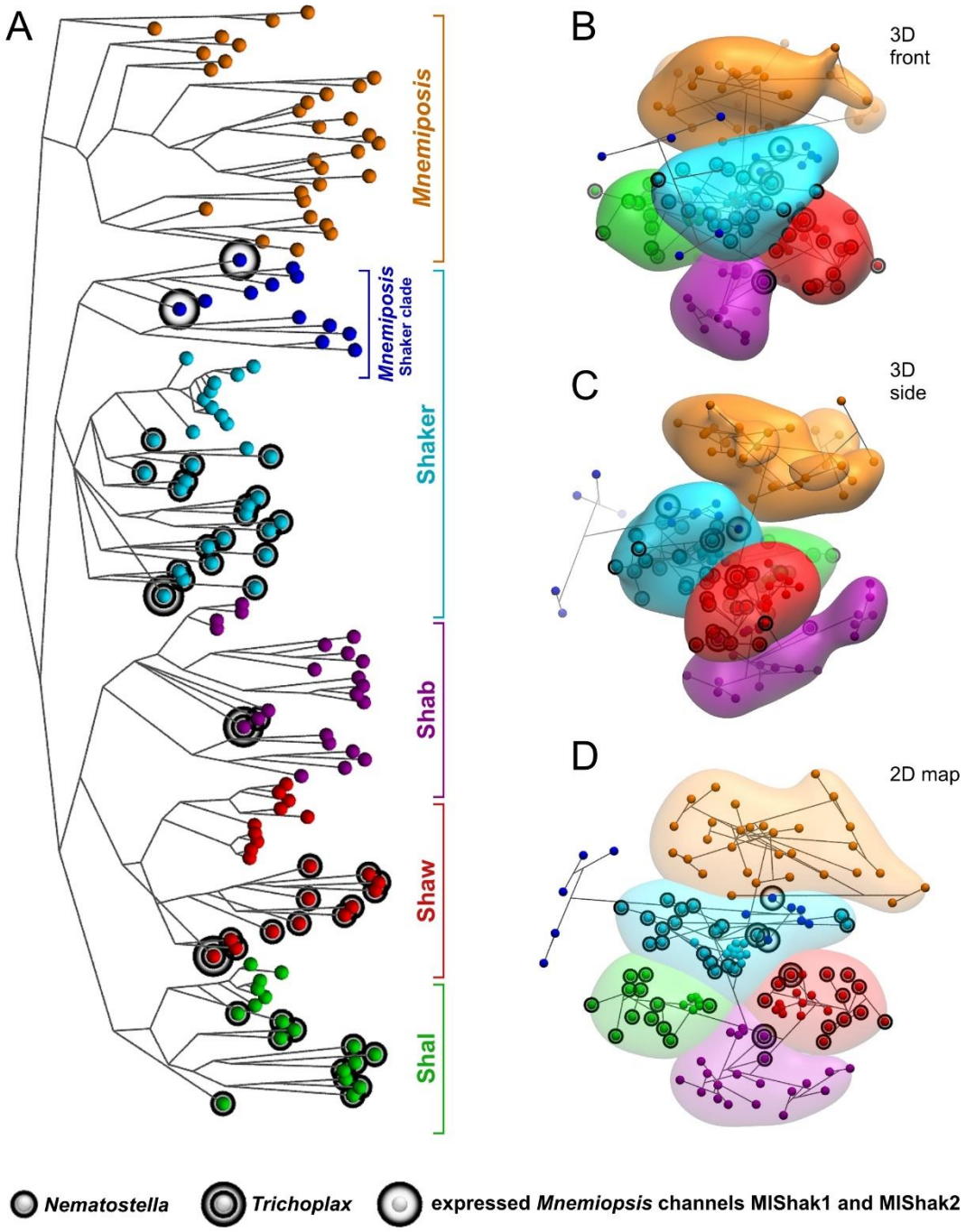


Figure 2 Legend: Sequence Similarity Networks (SSN) analysis supports inclusion of a subset of ctenophore Shaker family channels in the Shaker gene subfamily. (A) Representation of the conventional Bayesian phylogeny, colored to help interpretation of 2D and 3D representation. Colored dots are used to indicate the Shaker subfamily (parahoxozoan, cyan; Mnemiopsis, blue), the Shab subfamily (purple), the Shaw subfamily (red), the Shal subfamily (green), and an outgroup of Mnemiopsis channels (orange). Nematostella sequences have a small black outline, Trichoplax sequence are double outlined, and the expressed Mnemiopsis channels MlShak1 and MlShak2 have a large black outline.

(B,C) The same set of channels shown as sequence similarity network in two 3D views with pairwise distance between the dots directly proportional to Evolutionary Distance between the sequences. It allows straightforward interpretation of spatial proximity and clustering. Channels are labeled as in (A), lines show connections from the phylogeny, and transparent surfaces are used to highlight gene subfamilies. The phylogenetic connectivity was not used in calculations of spatial positions. Note a few of the most divergent sequences lie outside the surfaces, but connect to the network through the appropriate subfamily. (D) 2D representation of the network to allow simultaneous visualization of all nodes without occlusion, while still maintaining the same direct spatial relationships.

Structural/Functional Modeling in Extreme Divergence

Over the course of the award period we conducted structural and functional modeling on a number of proteins including:

- Fatty acid-binding to oxy-myoglobin (Chintapalli SV, Bhardwaj G, Patel R, Shah N, Anishkin A, van Rossum DB, Patterson RL, Adams SH. Molecular Dynamic Simulations Reveal the Structural Determinants of Palmitic Acid Binding to Oxy-Myoglobin. (In Review 2015, PLoS ONE)),

- Novel lipid-association domain in BRCA1 which impact pro-apoptotic mechanism (Hedgepeth SC, Garcia MI, Wagner LE II, Rodriquez AM, Chintapalli SV, Snyder RR, Hankins GDV, Henderson BR, Brodie KM, Yule DI, van Rossum DB, Boehning D. The BRCA1 Tumor Suppressor Binds to Inositol 1,4,5-Trisphosphate Receptors to Stimulate Apoptotic Calcium Release. J Biol Chem (2015) Feb 2. [Epub ahead of print] PMID: 25645916

- N-terminal Per-Arndt-Sim (PAS) protein interaction domain and cyclic nucleotide binding homology domain (CNBHD) of ERG Voltage-gated K^+ channels (Martinson AS, van Rossum DB, Diatta FH, Layden MJ, Rhodes SA, Martindale MQ, Jegla T. Functional evolution of Erg potassium channel gating reveals an ancient origin for IKr. Proc Natl Acad Sci U S A (2014) Apr 15; 111(15):5712-7. PMID: 24706772),

- Transient Receptor Potential (TRP) ion channel pore forming unit (Lindy A, Parekh PK, Zhu R, Kanju P, Chintapalli SV, Tsvilovskyy V, Patterson RL, Anishkin A, van Rossum DB, Liedtke WB. TRPV-Channel-Mediated Calcium-Transients in Nociceptor Neurons are Dispensable for Avoidance Behavior. Nature Communications (2014) Sep 2; 5:4734. PMID: 25178952)

Perhaps the most challenging modeling project was the TRP ion channel due to the extreme sequence divergence relative to structurally resolved PDB templates (~12% identity between template and target). We curated a set of 88 channel sequences from Genbank representative of TRPV sequences from various mammalian and invertebrate taxa and 1 TRPA1 sequence from marine sponge, used as outgroup. These sequences were analyzed to assess the extent of primary amino acid sequence similarity between OSM-9, structural templates, and mammalian TRPV1-6 across pore forming regions and the conservation of secondary

structure over individual channel subdomains.

Amino acid secondary structure was predicted utilizing the algorithm JNet8 implemented by Jalview v2.8 9. JNet uses several different neural networks and decides on the most likely prediction via a jury network. This analysis revealed that secondary structure is generally conserved across channel subtypes with insertion/deletions occurring primarily in loop regions between TM1 and TM2 and the turret region between TM5 and the pore helix.

To estimate the degree of divergence between amino acid sequences we aligned the 88 channel sequences with Muscle and parsed out the sequence regions representing individual channel subdomains. We utilized MEGA5 to quantify the number of amino acid differences per site between sequences with all ambiguous positions removed for each sequence pair. As one might expect we found that TM4, TM5, and TM6 are the most conserved subdomains in OSM-9, although they are still highly divergent. We limited the OSM-9 sequence boundaries between amino acids 523-649, encoding TM5-TM6 (i.e. the pore-forming unit). This estimation was informed by: (i) reported boundaries from previous TRP channel models, (ii) various predictions of secondary structure and transmembrane domain topology for OSM-9 and (iii) multiple sequence alignment of OSM-9 with the TRPV subfamily.

Within the pore-forming unit we found a high degree of divergence with structural templates and TRP homologues. For example, we compared the pairwise alignment between structurally resolved KcsA and TM5-TM6 of OSM-9. We calculated the sequence identity for a variety of sequence fragments: TM5-TM6, Percentage identity 10.7% (alignment length: 122 residues); Pore-Helix and Selectivity Filter, Percentage identity 26.3% (19 residues); Selectivity Filter, Percentage identity 42.9% (7 residues). We also compared pairwise percentage identity between TM5-TM6 of OSM-9 versus Kv and KIR potassium channel templates and mouse versions of TRPV1-6. These data show a high degree of sequence divergence in the pore forming unit. Overall, the channels are much less conserved by primary amino acid sequence than by secondary structure, and presumably tertiary structure.

A main factor that determines the quality of a homology model is the alignment accuracy between template and target. It is generally considered that the quality of the model stands questionable when the sequence identity falls in the twilight zone (<20-35% identity) of sequence similarity. In our case both potassium channel templates are highly divergent in sequence (the KIR, 3JYC template and OSM-9 share sequence identity of 13.1% id; the Kv, 2R9R template and OSM-9 share sequence identity of 12.3% id; 3JYC and 2R9R share sequence identity of 20.35% id). With respect to gaps in both template-target alignments, insertions/deletions in one are mostly absent in the other. While the overall percentage identity in the alignment of OSM-9 against either structure is low (<25% identity), there are several regions with significant blocks of similarity. Previous reports also found low identity alignments between TRP and potassium channel templates in a similar range of identity and similarity. Based upon the low degree of sequence identity we took extra care in all analyses and refinement steps to try and achieve the highest level of precision one could expect given the divergence.

Since OSM-9 and other TRP channels have low sequence identity (<25% id) to potassium channel templates, conclusions based solely on these homology models should be taken with a grain of salt. Quite recently, three cryo-EM structures of rat TRPV1 (PDB: 3J5P, 3J5Q, and 3J5R) were reported with crystallographic resolution 3.3, 3.8, and 4.2 Å respectively. Pairwise sequence identity between OSM-9 and 3J5Q was analyzed (~31% id with turret removed). These new structural references afforded us the opportunity to assess models of OSM-9 based upon potassium channel templates and a TRP channel template, and to evaluate the extent to which we can reconcile mutational analysis in a structural context.

Specifically, we relied on structurally resolved potassium channels Kv1.2-2.1 Paddle Chimera Channel (PDB ID: 2R9R), the chicken inward-rectifier K+ channel Kir2.2 (PDB ID: 3JYC), and TRPV1 cryo-EM structures (PDB ID: 3J5P) as templates. The TRPV1 and Kv monomers possess six transmembrane domains while the full length Kir monomer contains only two transmembrane domains. For all templates and the OSM-9 target, four monomers combine to form the channel pore domain. One rationale for the use of multiple templates is based on our desire to track any variation in orientation, conformation or match to the membrane environment that may indicate higher quality of one model over the other. Both potassium channel templates used in this study have a comparable degree of sequence similarity to OSM-9. However they have certain differences in orientation and position of the transmembrane helices. In addition, similar to OSM-9, 3JYC structure has a turret region, which is removed or absent in TRPV1 and 2R9R, respectively, and was modeled *de-novo* for those templates. Turrets have been structurally resolved in multiple potassium channels and the presence of a turret in OSM-9 is consistent with recent reports modeling turrets in TRPV channels and their potential role in temperature sensation.

We used MODELLER (9v8) to generate 30 initial models of OSM-9 based on Kv or Kir templates or based on the three TRPV1 cryostructures. MODELLER uses a variety of methods (e.g. Z-DOPE, molpdf, GA341) to assess the model. These scores are associated with the estimated accuracy of the model; the molpdf and DOPE scores can be used to rank models. Overall, the unliganded (3J5P) TRPV1-based starting structure has the lowest Z-DOPE score (0.44) of all models. We refined this TRPV1-based model and compare it with refinements of the top candidate model from both potassium channel templates. With respect to the latter, we selected the top candidate model guided by the molecular PDF score (molpdf), which is the sum of all spatial restraints achieved when transferring structural regions from the template to the target. These candidates also had the lowest Z-DOPE scores of the 30 initial models generated for each potassium channel template.

We next sought to generate tetramers representing the putative pore-forming unit. For each monomer (Kv- or Kir-based), one hundred candidate symmetrical tetramer models were generated using SYMMDOCK which predicts multimers using a geometry-based docking algorithm. SYMMDOCK predicts the cyclically symmetric complex given the input symmetric unit structure and the symmetry order ‘n’ (n =4 tetramer). The algorithm produces top hundred cyclically symmetric transformations in both its search and clustering by exploiting local feature matching. We visually inspected each tetramer and selected the best models whose domain organization accords to both Kv and Kir-based templates and previous TRP channel models. The criteria included the tightness of the inter-subunit contacts, the

width of the conductive pathway, and the expected match to the lipid bilayer. For the TRPV1-based monomer we superimposed the structure over the cryotemplate in VMD to generate the putative tetramer for refinement.

We implemented a refinement strategy using symmetry-driven annealing in explicit media. In this manner we derived important information about the hydration of the protein and ions, dynamics of the ions in the filter, interaction of the channel with lipids, de-wetting of the hydrophobic gate and partial de-wetting of the ion chamber. These MD simulations and annealings refine models through interactions with the medium and provide informed, detailed and reliable insight into a model's quality and stability.

To perform the MD Simulations, we used the NAMD package. Assembly of the simulation cells, analysis and visualization of the results were performed using custom written Tcl scripts in VMD v1.9.1. All MD Simulations were performed in the NPT ensemble using CHARMM36 force field parameters and TIP3P water model. The acidic and basic residues were set in two variants for each model: one variant in their default protonation state, and another “Kir2.2-like, predicted charges” and “Kv1.2-like, predicted charges”. For the latter two structures the dissociation state of the ionizable residues was estimated using PROPKA for the neutral pH.

Based on that, residues D605, and K619 for Kir2.2-derived model and E591, D605, E607, E608, and K619 for Kv1.2-derived model were set non-dissociated. The N- and C-termini were set to a truncated non-dissociated state since the real protein is longer than the simulated fragment in both directions. All the models were embedded in a POPE (1-[(2)H(31)] palmitoyl-2-oleoyl-sn-glycero-3-phosphoethanolamine) bilayer membrane (pre-equilibrated for 40 ns in the Laboratory of Molecular & Thermodynamic Modeling (PI Jeffery Klauda), University of Maryland, College Park) surrounded by water, 247 and 228 POPE molecules for simulations with Kv1.2 and Kir2.2 templates respectively.

The presence and the location of cations might significantly influence the structure of the selectivity filter. Two calcium ions were placed in the filter in two (S1 and S3) out of four positions found in the template crystal structures. Two water molecules were placed in the remaining two positions, alternating with the calcium ions. To maintain the electroneutrality of the system Na^+ and Cl^- ions were added up to equivalent of 150 mM salt concentration – 52 Na^+ and 48 Cl^- for all Kir2.2-based systems, 64 Na^+ and 72 Cl^- for Kv1.2-based system with PROPKA-predicted dissociation states, and 68 Na^+ and 64 Cl^- for Kv1.2-based system with default protonation. Langevin Dynamics was used to maintain constant pressure (1 atm) and constant temperature (303.15^o K) in a flexible simulation cell to allow relaxation of the in-plane area. Periodic boundary conditions were maintained and the particle mesh Ewald method with a real space cutoff distance of 12 Å and a grid width of 0.97 Å. Energy Minimization steps were performed using the steepest descent in the first 2000 steps and then a conjugate gradient in the subsequent 2000 steps. To attain the equilibrium, the system was subjected to gradual heating until it reached to 303.15^o K (30°C), first with the protein backbone harmonically restrained (1 kcal/mol/Å^2 per backbone atom) to the initial coordinates, while the lipids were allowed to pack around and water to stably solvate the protein for 5 ns.

In the TRPV1 cryo-EM structures no ions were resolved, just protein. Therefore, in the starting TRPV1-based model we have placed the calcium ions manually, based on similarity of ion placement in potassium channels in the vicinity of the pore-facing backbone oxygens of the filter. Importantly, the side chains of D605 in the starting model are facing away ("upwards") from the filter, they are not sufficiently close to coordinate ions to reasonable degree. Unlike the highly selective potassium channels, where the filter is tight and potassium ions are stripped of most of the hydration shell, a low-selective filter as in OSM-9 is likely to allow Ca^{2+} ions to keep most of the hydration shell, and therefore is probably wider and more dynamic.

As a family, TRP channels are generally considered non-selective cation channels. Hence, one might expect that the filters in these channels be sufficiently flexible to accommodate ions of multiple sizes (Pauling radii/pm: K^+, 133; Na^+, 95; Ca^{2+}, 99; Mg^{2+}, 65) and differential hydrations states of the ion (e.g. Ca^{2+} can have up to 9 H20 molecules in its first hydration layer). In support of this, it has been shown that TRPV1 and TRPV5 channels have wide pores and dilate in the presence of agonists. Furthermore, the kinetics of pore widening demonstrates a dependence of the rate of widening as a function of solute radius. Therefore, the position of the ions in such a filter is probably not very restricted and might allow certain degrees of freedom - they do not necessarily have to be exactly at the pore axis. Along the permeation pathway, more energetically favorable positions at some steps might be near the acidic side chains. Because of that, we did not put any restraints on the position of the ions during the refinement process (i.e. only the protein is symmetry-restrained).

Starting imprecision in the modeled conformation of the selectivity filter region might cause the loss of coordinated ions at the beginning of the unrestrained simulation, which in turn might distort the filter conformation and hinder its refinement. To avoid this, we used Tcl scripting in NAMD to apply soft boundary conditions for ions that would prevent them from leaving the 12 Å vicinity of the center of mass of the selectivity filter, but do not apply any force within this sphere. The boundary conditions were present through all the simulations, however the ions were found to settle at the binding points within the filter after the first few nanoseconds of simulation and after that did not attempt to leave the selectivity filter and did not approach the boundary, therefore no forces were applied to ions through the most of the simulation time.

To refine the channel structure in an explicit all-atom medium and study the network of protein-protein, protein-lipid, and protein-ion interactions across the channel, we applied a simulation protocol that combines cycles of unrestrained relaxation with a recently developed symmetry-restrained annealing step. Previous results have shown that unrestrained simulations are useful in refining small and medium size proteins but have limited competency in improving the refinement of larger low-resolution models. When compared to unrestrained models alone, symmetry-restrained annealing has been shown to improve accuracy (by up to 50%) in homology models of potassium channels.

In the symmetry-driven simulated annealing protocol, a rotational symmetric average conformation is calculated for the four monomers composing OSM-9 channel. A weak

harmonic force (spring constant starting at 0.001 kcal/mol/Å^2) is applied to every atom of a monomer to drive it slowly towards the symmetric average. The average is continuously updated (every 10 fs), so that the final conformation is not pre-defined at the beginning, but is gradually "chosen" by the system. One of the principles underlying the refinement is that if one of the monomers finds a more stable conformation than others, then driving all of them to symmetric average will prompt the unstable monomers to change first and adopt the more stable conformation. The strength of the harmonic restraints gradually increases to produce near-linear decrease in RMS deviation of the structure from symmetric average (at start it is usually on the order of a few Å) over 1 ns annealing simulation.

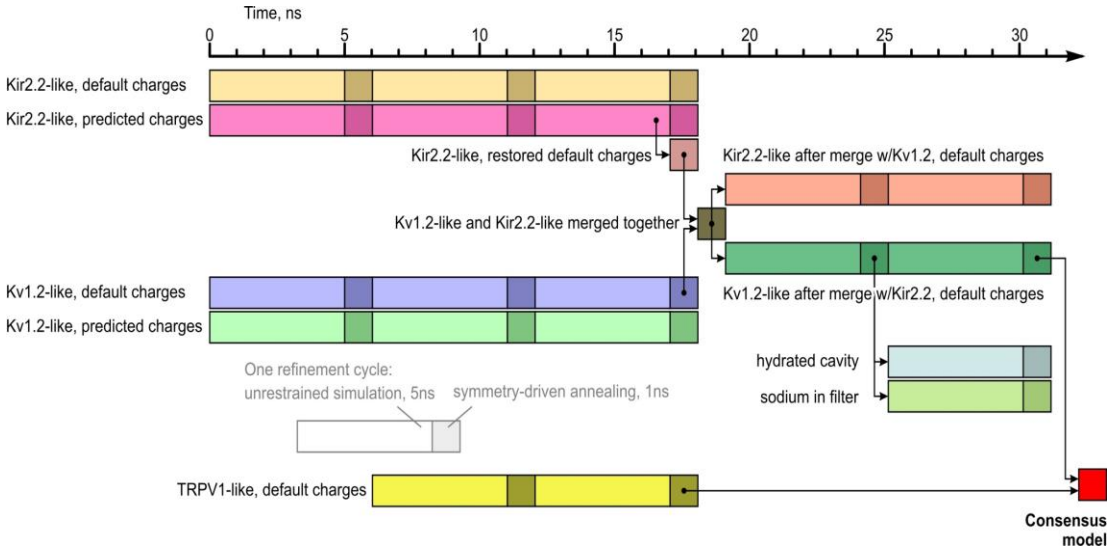


Figure 3 Legend: Graphical representation of Molecular Dynamics refinement protocol. Schematic overview of the steps involved in the MD refinement protocol for the OSM-9 structure. The protocol shows different simulation time steps carried on Kir2.2-, Kv1.2-, and TRPV1-derived models. Kir2.2-, and Kv1.2- models were simulated in two variants of protonation state for certain amino acids. After first 3 cycles of Kir2.2-, and Kv1.2-like models, the best structures were converged to an updating average structure, followed by 2 cycles of refinement of the merged model. The final refined merged model was converged with 2x refined TRPV1-like mode following the same protocol. We also compared the wild type Kir2.2-Kv1.2 merged 1x refined model with Na^+ placed instead of Ca^{2+} in the selectivity filter, and with the system with fully-rehydrated ion chamber.

Figure 3 depicts the protocol employed for the molecular dynamic refinement for all simulations that were conducted for this study. In total, we simulated more than 154 ns of molecular dynamics for the OSM-9 pore. From the initial TRPV1-based starting model we implemented two cycles of 5 ns relaxation and 1 ns annealing with default charges for the ionization state. From the initial starting models (Kv- and Kir-based) we implemented three cycles of 5 ns relaxation and 1 ns annealing. These simulations were repeated for two different ionization states- default charges and PROPKA predicted charges. After

performing three cycles of 5 ns relaxation and 1 ns annealing we re-analyzed the dissociation state of amino acids in all the potassium channel-based models using PROPKA. Unlike the starting conformation, all the residues (including the residues D605, and K619 for Kir2.2-derived model and E591, D605, E607, E608, and K619 for Kv1.2-derived model that were simulated non-dissociated) were estimated to have their default dissociation at neutral conditions. This fact likely indicates that the starting unconventional dissociation state was reflecting some minor conflicts in the starting conformation rather than the essential features of the modeled structures.

For all the subsequent relaxation/annealing cycles we simulated the default dissociation state for all the potassium channel-based models. Nevertheless, using two alternative dissociation sets have provided certain benefits: the analysis of the quality of the refined models indicates that the OSM-9 structure based on Kir2.2 with the predicted charges ranks higher than that of Kir2.2 with the default dissociation state. One can speculate that in some conditions varying the dissociation state of amino acids can be useful for exploring the conformational space of the models as it can allow the structure to temporarily overcome energy barriers not possible in the default dissociation state (e.g. if the charged residue is trapped in a “wrong” salt bridge or needs to switch from one hydrophilic pocket to another through a hydrophobic region).

We assessed refinements to the TRPV1-, Kv- and Kir-based starting structures through comparative analyses of protein-lipid contacts and hydration states between structures and cycles of the simulation. Specifically, by tracking the distribution of water molecules in the vicinity of the channel, one can identify potential structural issues that need to be resolved such as leaking chamber walls, leaking filter, and streams of water between helices. Conversely, the lack (or repair) of issues through cycles of refinement may give confidence to a particular structure over another.

All four variants (default and predicted dissociation states of Kv- and Kir-based models) pull significant amount of water into the region near the selectivity filter. In the Kv-based models at the first stage of refinement water spontaneously forms and maintains a few strings through the gate region; however, by the second refinement both Kv-based models "evolve" to eliminate water strings in the hydrophobic gate. Initially in one of the Kv-based models fwith a dissociated K619, water tends to form strings in the space between S5 and S6 helices, either to the cytoplasmic or external sites, although during the second refinement K619 repositions more closely to the hydrated region near the selectivity filter and water strings interrupt. We observe that the channel chamber seems to be easily dehydrated in the absence of ions; however, presence of even one calcium ion holds water in the chamber.

The Kir-based models have a different hydration pattern. These models initially show strings of water crossing the S6 bundle from the chamber to the periplasmic bulk side. Notably, this negative feature was not completely eliminated in future refinements. A point of distinction between Kir- and Kv-based models is that the putative gate region seems to be more tightly closed in Kir-based models. Comparatively, the Kv-based OSM-9 models appear to be more stable, both in transmembrane helices and the loops. In this model, the selectivity filter is positioned somewhat higher than the Kir-based structure; however, the Kv-based filter is about at the same height as it was after the very first annealing. Notably, both structures have

ions in similar positions and coordinations.

Taken together, these simulations show that the initial potassium channel-based starting models had some minor structural issues such as leaking chamber walls, leaking filter, water strings between helices 5 and 6, and exposure of some polar residues to the lipids. The fact that the simulation protocol employed solved most of them demonstrates that symmetry-driven annealing improves the quality of the models rather than just testing them in simulations. With respect to the turret in Kv-based model, there is still a conflict even after three cycles of simulation. In the initial starting model the turret is largely unstructured and sticks straight up into the extracellular media. After three cycles of refinement the turret has become more compact and perpendicular to the membrane; however, its loops erroneously cross in an impossible conformation. Conversely, the turret in the Kir-based model was more structured and compact in the initial starting structure versus the model after the third cycle of refinement. Hence there is a tendency for the turret to be extended in both models.

For the TRPV1-based starting structure, we found the ion chamber to be narrow even at start, and the water content of the ion chamber not to increase. As in our potassium channel based models it tended to dehydrate. The fit to the membrane is reasonably good and there is no water permeation between the channel and the lipids. The gate is dry. These features indicate that the starting model is of suitable good quality. However, a more troublesome feature in the starting TRPV1-based model is the filter which quickly collapsed (flattened) in simulation; it is even dry in the very middle. Conversely, our K-channel based model had a stable water string and ion presence in the middle of the filter. The TRPV1-based structure is derived from a channel with higher % sequence identity – yet still highly divergent. One of the reasons the filter in the TRPV1-based model, which shows several advantageous features, is collapsed and apparently non-conductive could be that the template is a cryo-EM structure. The filter region is heavily hydrated and water is a critical structural element. Low temperature might affect the aggregate state of water, and therefore likely the structure of the filter region.

In the 1x refined TRPV1-based model the top Ca^{2+} ion stays at the side, near two D605 side chains, while the lower ion is close to symmetry axis being coordinated by four partial charges of backbone oxygens. The 2x refined TRPV1-based model of OSM-9 improves, (with some room for further refinement). Ca^{2+} ions stay in similar positions where they were after the first round - above and below the filter. After the unrestrained stage we observe a continuous string of water in the filter, however after symmetry annealing the filter has de-wetted again. The filter got narrower compared to the previous cycle. Some of the charged residues in the loops have aggregated in a salt-bridged ring-like structure, an interesting conformation. These salt-bridges pulled the top of the channel together, to a narrower conformation.

After each cycle of refinement, we manually inspected each model to assess stability and any differences that may inform preferences to one model over the other. After four rounds of refinement the two best potassium channel-based models were selected based on structural fitness criteria. The dissociation state of the ionizable side chains was re-evaluated using PROPKA and found to be the default dissociation for the neutral pH for both models.

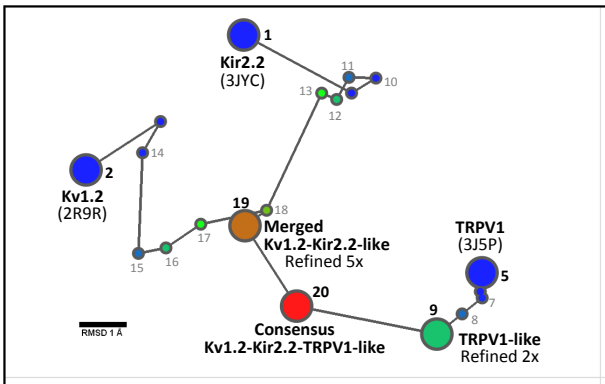
Importantly, even before the models were merged, we found that in the course of multiple cycles of refinement both models were converging towards a common structure (Figure 4). Only after convergence had leveled off did we apply our protocol for gradual steering two alternative models towards a common conformation. Similar to symmetry-driven annealing, an instantaneous average of the two structures was computed and the soft harmonic forces were applied to every protein atom towards continuously updated average to ensure gradual convergence over 1 ns. Since more stable regions will persist longer than unstable regions in the presence of a small uniform force, the merged model tends to combine the most stable features of the two starting models.

After converging the Kv- and Kir-based models together, we performed two cycles of relaxation/annealing. In addition, we conducted a cycle of refinement with either a hydrated cavity, sodium in the filter, or mutations to aromatic tyrosine residues at amino acid positions 604 or 606. The structures obtained after the five simulations can be considered as the most promising potassium channel-based models of OSM-9, in slightly different states - the top one has the driest inner cavity, the second one is more hydrated (although with less water than was added at the beginning of the simulation cycle; there is certainly some tendency towards dehydration in this model), and the third one has sodium ions in the selectivity filter, which organizes it to some extent. Notably, the strings of water crossing the S6 bundle from the chamber to the periplasmic bulk side in the 3x refined Kir-based model have been corrected in the merged model. In addition, conflict in the turret from the 3x refined Kv-based model has been resolved in the merged structure though it is still extended and largely unstructured. Nevertheless, we observe dynamic motion in the turret and changes in the extent of hydration in the different conformations it adopts.

Notably, we also found that in the course of refinement the TRPV1-based model spontaneously “evolved” from the initial TRPV1 conformation to become closer to the Kv-based models (Figure 4). Specifically, this change occurred in overall structure and particularly in the way calcium ions are coordinated by acidic side chains at the filter entrance. Hence, we also conducted one final simulation - merging the TRPV1-based model and our Kv/KIR-based models together to a consensus conformation. In this way we addressed which structural features will dominate. We found that the TRPV1-based model converges to the Kv/KIR-based model, at least in the filter region. For quantitative evaluation of the entire set of homology models we calculated root-mean-square deviation (RMSD in Å, lower number indicates better superimposition) for backbone atoms of superimposed protein regions. These analyses represent the transmembrane helices, pore helix and selectivity filter derived from comparison of the potassium channel templates, TRPV1 EM cryo-structures, a published TRPV1 model based on a potassium channel template 19, and starting and refined models of OSM-9 from the current work. We compared more than 36 structures, four channel regions in different combinations (S5, S6, pore helix and filter) totaling thousands of pairwise comparisons (A subset of these comparisons are shown in Figure 4).

We found that the most representative measure is if we consider all the structurally stable domains of the pore together - S5, S6, pore helix and filter simultaneously (i.e. not loops, turrets and linkers). With this metric, our best potassium channel-based model is the final

merged and refined structure with the dry pore. It approaches the closest with the structure 3J5Q (TRPV1 with bound toxins. Other expectations about the course of model refinement are supported by these RMSD data. For example, OSM-9 models from both potassium channel templates do approach the TRPV1 cryo-EM structure in the course of refinement.



Name	#	1 Kir2.2 (3JYC)	2 Kv1.2 (2R9R)	3 TRPV1 (3J5P)	4 TRPV1 (3J5Q , with bound toxins)	5 TRPV1 (3J5R, with Capsaicin)	6 TRPV1 model (Fernandez-Ballester et al.)	7 TRPV1-like OSM-9 model	8 TRPV1-like OSM-9 model, refined 1x	9 TRPV1-like OSM-9 model, refined 2x	10 Kir2.2-like OSM-9 model, predicted charges	11 Kir2.2-like OSM-9 model, predicted charges, refined 1x	12 Kir2.2-like OSM-9 model, predicted charges, refined 2x	13 Kir2.2-like OSM-9 model, predicted charges, refined 3x	14 Kv1.2-like OSM-9 model, default charges	15 Kv1.2-like OSM-9 model, default charges, refined 1x	16 Kv1.2-like OSM-9 model, default charges, refined 2x	17 Kv1.2-like OSM-9 model, default charges, refined 3x	18 merged_Kv1.2-Kir2.2-like OSM-9 model, default charges, refined 4x	19 merged_Kv1.2-Kir2.2-like OSM-9 model, default charges, refined 5x	20 consensus_Kv1.2-Kir2.2-TRPV1-like OSM-9 model, default charges
Kir2.2 (3JYC)	1		5.0	6.6	6.6	6.4	6.2	6.6	6.7	6.6	3.6	4.2	4.3	4.1	4.9	5.4	5.4	5.3	5.5	5.0	5.5
Kv1.2 (2R9R)	2	5.0		6.9	6.9	6.9	4.2	6.8	6.8	6.4	5.2	5.9	5.8	5.9	2.2	3.3	3.8	4.2	4.8	4.3	5.3
TRPV1 (3J5P)	3	6.6	6.9		1.6	1.4	5.4	0.8	1.4	1.8	5.6	5.8	5.7	5.7	6.3	6.1	5.6	5.2	4.9	4.8	4.1
TRPV1 (3J5Q , with bound toxins)	4	6.6	6.9	1.6		1.7	5.5	1.9	2.0	2.2	5.5	5.6	5.5	5.5	6.4	6.1	5.6	5.1	4.8	4.7	4.0
TRPV1 (3J5R, with Capsaicin)	5	6.4	6.9	1.4	1.7		5.2	1.7	2.0	2.0	5.4	5.6	5.4	5.5	6.2	5.9	5.5	5.0	4.7	4.6	3.8
TRPV1 model (Fernandez-Ballester et al.)	6	6.2	4.2	5.4	5.5	5.2		5.4	5.3	4.8	5.8	6.4	6.1	6.1	3.7	3.2	3.4	3.8	3.9	3.6	4.0
TRPV1-like OSM-9 model	7	6.6	6.8	0.8	1.9	1.7	5.4		1.4	1.7	5.6	5.8	5.7	5.7	6.2	6.0	5.5	5.1	4.8	4.7	4.0
TRPV1-like OSM-9 model, refined 1x	8	6.7	6.8	1.4	2.0	2.0	5.3	1.4		1.1	5.8	5.9	5.7	5.8	6.2	5.9	5.5	5.0	4.6	4.5	3.7
TRPV1-like OSM-9 model, refined 2x	9	6.6	6.4	1.8	2.2	2.0	4.8	1.7	1.1		5.7	5.9	5.7	5.8	5.8	5.6	5.2	4.8	4.3	4.2	3.3
Kir2.2-like OSM-9 model, predicted charges	10	3.6	5.2	5.6	5.5	5.4	5.8	5.6	5.8	5.7		2.2	2.8	2.5	4.7	5.1	5.1	5.0	5.1	4.7	4.9
Kir2.2-like OSM-9 model, predicted charges, refined 1x	11	4.2	5.9	5.8	5.6	5.6	6.4	5.8	5.9	5.9	2.2		1.2	1.2	5.4	5.2	4.9	4.6	4.6	4.3	4.7
Kir2.2-like OSM-9 model, predicted charges, refined 2x	12	4.3	5.8	5.7	5.5	5.4	6.1	5.7	5.7	5.7	2.8	1.2		0.9	5.4	4.8	4.5	4.2	4.1	3.9	4.4
Kir2.2-like OSM-9 model, predicted charges, refined 3x	13	4.1	5.9	5.7	5.5	5.5	6.1	5.7	5.8	5.8	2.5	1.2	0.9		5.4	4.9	4.6	4.4	4.3	4.0	4.4
Kv1.2-like OSM-9 model, default charges	14	4.9	2.2	6.3	6.4	6.2	3.7	6.2	6.2	5.8	4.7	5.4	5.4	5.4		2.8	3.4	3.7	4.3	3.8	4.5
Kv1.2-like OSM-9 model, default charges, refined 1x	15	5.4	3.3	6.1	6.1	5.9	3.2	6.0	5.9	5.6	5.1	5.2	4.8	4.9	2.8		1.4	2.1	2.7	2.6	3.6
Kv1.2-like OSM-9 model, default charges, refined 2x	16	5.4	3.8	5.6	5.6	5.5	3.4	5.5	5.5	5.2	5.1	4.9	4.5	4.6	3.4	1.4		1.1	1.9	2.0	3.1
Kv1.2-like OSM-9 model, default charges, refined 3x	17	5.3	4.2	5.2	5.1	5.0	3.8	5.1	5.0	4.8	5.0	4.6	4.2	4.4	3.7	2.1	1.1		1.5	1.5	2.7
merged_Kv1.2-Kir2.2-like OSM-9 model, default charges, refined 4x	18	5.5	4.8	4.9	4.8	4.7	3.9	4.8	4.6	4.3	5.1	4.6	4.1	4.3	4.3	2.7	1.9	1.5		1.1	2.2
merged_Kv1.2-Kir2.2-like OSM-9 model, default charges, refined 5x	19	5.0	4.3	4.8	4.7	4.6	3.6	4.7	4.5	4.2	4.7	4.3	3.9	4.0	3.8	2.6	2.0	1.5	1.1		2.1
consensus_Kv1.2-Kir2.2-TRPV1-like OSM-9 model, default charges	20	5.5	5.3	4.1	4.0	3.8	4.0	4.0	3.7	3.3	4.9	4.7	4.4	4.4	4.5	3.6	3.1	2.7	2.2	2.1	

Figure 4. Structural comparison of the starting crystallographic templates and three refined homology models of OSM-9. Root-Mean-Square Deviation (RMSD, in Å) was calculated for the homologous regions covering transmembrane helices S5 and S6, the pore helix and the selectivity filter. Refinement of Kir2.2- and Kv1.2-based models makes them more similar to TRPV1 cryo structure, comparing to the starting templates (green frames in the table). Also, even before the merging was performed, Kir2.2- and Kv1.2-based models become more similar to each other during the independent cycles of the refinement (blue frames). Refinement of the TRPV1-based model drives structure away from TRPV1 and towards the Kv1.2 template to certain extent (cyan frames) and makes it more similar to the refined Kir2.2-Kv1.2-based model (red frames). Slow convergence of the refined Kir2.2-Kv1.2- and TRPV1-based models to a continuously updated average have produced a consensus model that is somewhat more similar to Kv- rather than TRPV1-based model (yellow frame). An insert on the top left shows a 2D "map" of the refinement process reflects pairwise structure similarity from RMSD table as spatial proximity.

Overall, the comparative values in the table are quite encouraging: they show that in the course of refinement the TRPV1-based model becomes more similar to our 5x refined Kv/KIR-based model. Inversely, the TRPV1-based models become less similar to the starting TRPV1 conformation. In other words, two remote Kv templates and 5x refinement produce a more stable model than one TRPV1 template with 2x refinement. Theoretically, by the same logic, the final consensus model is even closer to reality. Taken together, these data argue that the potassium channel-based model had suitable features. Combined with additional observations from the table, that models from two Kv templates converge more similar to each other spontaneously, and that they evolve to be more similar to the TRPV1 cryo-EM structure we have reason to believe that the refinement methods are working and the former models are valuable, along with the new TRPV1 models.

Figure 5 depicts OSM-9 models versus template structures. Compared to the TRPV1 cryostructure, the Merged Kv1.2-Kir2.2-like Refined 5x model correctly predicts the start and end of pore-lining TM domains, their axial rotation (i.e. what side looks into the pore), overall tilt, and the pore helix. This model also successfully identified the residues comprising the gate region, with only minor differences, and "second gate" region positioned right under the filter formed by methionine in both channels. The acidic residues above the filter form a similar negative funnel. In both channels the filter seems to be wide enough to path the ions with most of the hydration shell, hence the low selectivity. In both filters the lowest ring of the acidic amino acids looks into the pore with its side-chains, enabling a direct contact of four negative charges with ions. The wider filter in TRPV1 vs. OSM-9 calcium-in-filter model may not necessarily be a structural feature that differentiates OSM-9, but a function of the degree to which divalent cations are coordinated in the cryo preparations. Notably, in the OSM-9 model that has Na^+ ions present instead of Ca^{2+} the filter is exactly as wide as the one in TRPV1 structure.

The main discrepancy between the potassium channel based models and the TRPV1 reference structure is that the S6 helix starts one helical turn earlier and is positioned one turn lower, towards the cytoplasmic/gate side. This might have caused the narrowing of the upper gate-forming ring of the isoleucines in our model. However, one cannot say that the 1-turn vertical shift of S6 is necessarily wrong - one can imagine such a variation in two family members such as TRPV1 and OSM9. Indeed TRPV1 is likely to show critical structural differences from OSM-9 since it is unable to rescue osmotic avoidance behavior in worms but can confer de-novo capsaicin sensitivity. This difference in S6 could be tested experimentally (although it might be pretty challenging to distinguish – it is a minor shift, with the same side looking the same way). If the potassium channel based model is correct, then hydrophilic mutations in the isoleucine ring should make gain-of-function channels comparable to mutations in the lower ring, while in TRPV1 analogous mutations might be more mild. If experiments indicate that the OSM-9 helix is too high, one could easily displace it one turn up, most of functional interpretations would still hold.

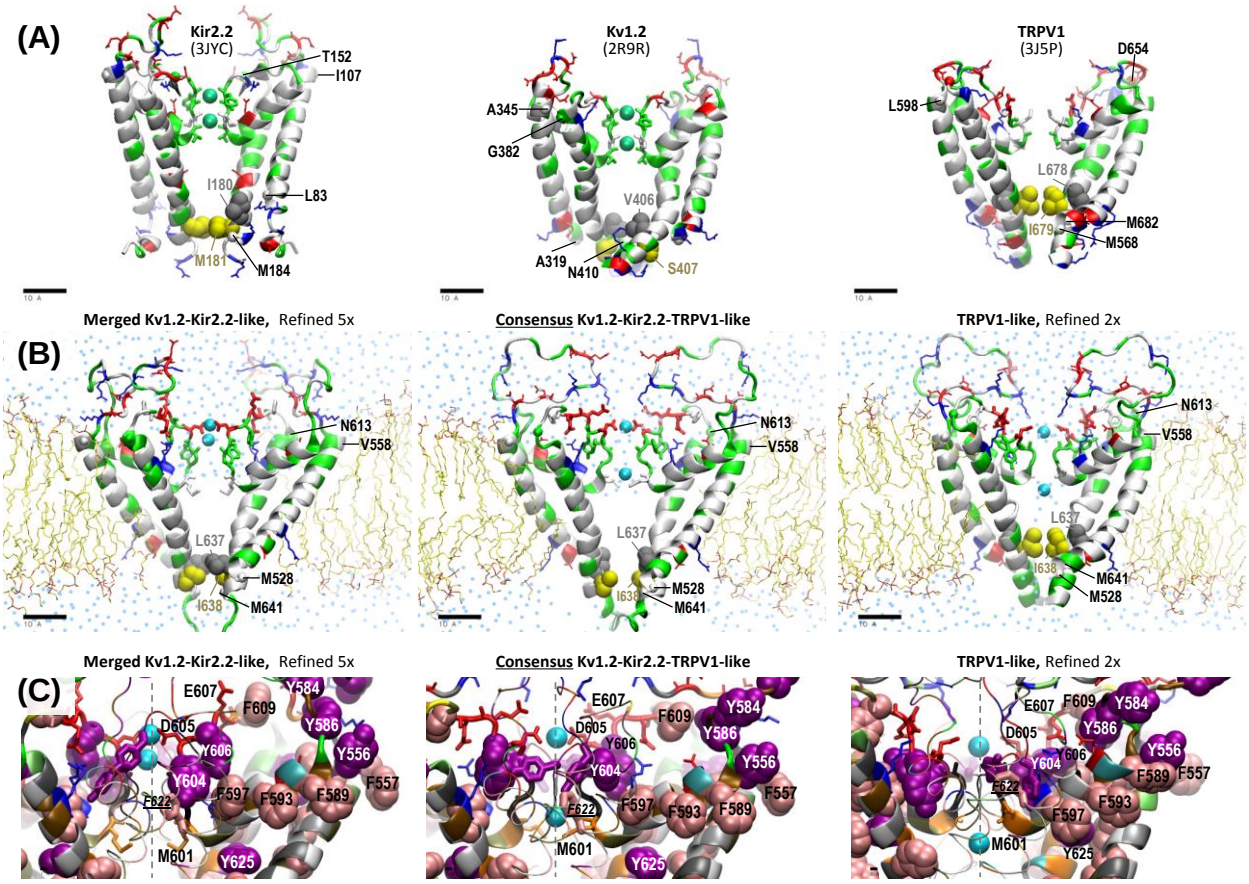


Figure 5 Legend: Comparison of templates and targets. Row (A) Full-length view of the templates used for building the homology models. Protein residues are colored by type (white – hydrophobic, green – hydrophilic, red – acidic, blue – basic). For Kir2.2 and Kv1.2 template structures two of the K^+ ions are shown in crystallographic positions as green spheres, they were replaced by Ca^{2+} ions in the homology models at the start of the refinement. As no ions were resolved in the cryo-EM TRPV1 structure, the Ca^{2+} ions in the homology models were originally placed above and below the narrowest part of the filter. The residues labeled in black on the panels in rows (A) and (B) mark the homologous positions for the beginning and the end of the transmembrane helices S5 and S6 in the starting OSM-9 models. The residues homologous to L637 and I638 (forming the hydrophobic gate in OSM-9 models) are shown in gray and yellow color respectively, in VDW representation. Row (B) Refined models based on Kv1.2-Kir2.2-like merged structures (left), TRPV1 (right), and the converged consensus model (center). A slice of water oxygens (cyan dots) through the simulation cell shows stably dehydrated ion chamber, while a continuous water pathway is retained in the selectivity filter of the refined Kv1.2-Kir2.2-like model and the consensus model, but not in TRPV1-based model. Row (C) Stacking in aromatic residues near the selectivity filter and the pore helix is shown for the three refined models. Side chains of tyrosines and phenylalanines are represented as purple and pink space fill respectively. Aromatic residues, shown as "thick licorice" rather than van der Waals are those that belong to another subunit, shown in front as thin tubes. Acidic and basic protein residues are colored red and blue respectively.

When considering the impact of residues Y604, Y606, and F609 on the selectivity filter we also examined them in the context of the homology models. In these models it appears that residues with large hydrophobic side chains are critical to pore geometry, and therefore ion permeability. We found that Y604 plays an important role in ion selectivity, which is demonstrated by the complete elimination of Ca^{2+} influx into ASH for the Y604F mutation. Yet ASH is still activated in Y604F mutant worms, albeit not fully to WT levels. It is well established that tyrosine residues are critical contributors to the Kv channel selectivity signature sequence, defined as TXXTXGYG. TRPV1-6 channels do not typically have tyrosine residues in their predicted selectivity filters, although invertebrate TRPV channels are replete with aromatics in the presumed selectivity filters. We find it attractive to hypothesize that Y604 could function in a manner similar to that of the tyrosine residue in Kv channels and serve a structural role in maintaining the orientation of the filter. This concept is in keeping with our homology modeling data.

The finding that mutation of Y604F partially maintains the worm's avoidance behavior indicates that the aromatic ring at position 604 is important to pore geometry. Y604 could possibly function as a master scaffold via its aromatic side chain. This concept implies that Y604 is at the core of the OSM-9 selectivity filter and could influence the local positions of residues D605, Y606, and E607, among others. The aromatic ring and hydroxyl group of the tyrosine residue facilitates Ca^{2+} permeation. The aromatic ring without the hydroxyl group is more hydrophobic and does not allow Ca^{2+} to permeate, so that presumably only Na^{+} can enter the cell. Changing the charge landscape of the pore by removing the hydroxyl group from the aromatic ring leads to reduced Ca^{2+} permeability, but not Na^{+} permeability, allowing for partial activation of ASH. In that case, Na^{+} mediated depolarization would be the activation mechanism of ASH that drives the appreciable avoidance behavior of Y604F worms.

Modeling data also suggest that the aromatic side chain of F609 is not a pore lining residue per se, but may contribute significantly to the size and hydration of the pore and have a considerable range of motion in various channel conformations. Integrating our experimental data and computational modeling, we propose that mutations of aromatic residues that drastically alter both behavior and Ca^{2+} influx (Y604 and F609) are amino acids that are critical to pore geometry. Alterations to pore geometry invariably result in changes to ion permeability. This can be further extended in future molecular dynamics studies.

Specific Aim 2

Of the > 50 research groups using the MS facility, multiple Orbitrap users (e.g. PIs Sundstrom, Lindner, and Booker) conducted research related to human health. For example, Dr. Sundstrom's research is focused on age- and disease-related changes in the eye at the protein level and developing treatments for macular degeneration and diabetic retinopathy. Lindner is characterizing the *Plasmodium* proteome at various stages of development of this malaria-causing parasite with the ultimate goal of finding a cure for this disease. Dr. Lindner's decision to join Penn State in July 2013 was largely due to the availability of the state-of-the-art core facilities, including the Proteomics Core. Another faculty hire for the Penn State's Center for Malaria Research, Dr. Manuel Llinas made his decision considering availability of the Orbitrap for the analysis of the protein post-translational modifications. Dr.

Booker's lab is investigating the mechanisms of enzyme catalysis in pathogenic microorganisms which have evolved to resist even the most potent antibiotics. This work requires a high resolution MS as the Cys oxidation states within a peptide might differ by less than 0.01% of the peptide mass.

The Core Facility has provided 11 letters of support (LOS) for external funding proposals, 2 of which have been funded (see Table 1). Moreover, the Huck Institutes of the Life Sciences have an internal Seed Grant program, funded separately, for new faculty which encourages the young investigators to familiarize themselves with the available technology and collect preliminary data for the external funding proposals. The Proteomics Facility has participated in this program since 2012 which resulted in a number of new collaborations.

Table 1. LTQ Orbitrap Velos recurrent users. Principle Investigators (PI) highlighted in bold rely on the Orbitrap in at least one project. Names of investigators who are in the process of submitting a grant proposal or have submitted grant proposals requiring the use of Orbitrap are italicized.

PI	**Affiliation***	**Research interests, LOS, and funding**
Booker, Squire	BMB, Chem	Mechanisms of cofactor action in enymatic reactions (**Seed Grant, $10K**; Oxidation states of quinolinate synthetase (NadA) from Mycobacterium tuberculosis)
Lindner, Scott	BMB	Molecular parasitology and structural biology of malaria
Llinas, Manuel	BMB	Gene regulation and metabolism in the malaria parasite Plasmodium falciparum
Showalter, Scott	Chem	Function of partially disordered proteins; Defining the features of protein-RNA interactions
Sundstrom, Jeffrey	Hershey Eye Center	Development of biomarkers and novel therapeutic targets for age-related macular degeneration (2 LOS pending)
Bollinger, Martin	BMB, Chem	Mechanisms of metalloenzymes and metallofactor assembly
Bryant, Donald	BMB	Physiology, biochemistry, genetics, and genomics of photosynthetic bacteria
Krebs, Carsten	BMB, Chem	Bioinorganic chemistry- spectroscopic and kinetic studies on the mechanisms of iron-containing enzymes
Luscher, Bernhard	BMB, Biology	Understanding of the role and function of GABAergic transmission in health and disease

Demirel, Melik	Eng Sci and Mechanics	Biosensors and nanomaterials
Babitzke, Paul	BMB	Regulation of gene expression by RNA structure and RNA-binding proteins
Keiler, Kenneth	BMB	Protein quality control and new antibiotics.
Krasilnikov, Andrey	BMB	Structural biology of RNA and RNA-protein complexes (**LOS NIH R01, $313K**; Proteins in RNA-based Enzymes)
Murakami, Katsu	BMB	Structural and Mechanistic Enzymology of Prokaryotic RNA Polymerases
Tien, Ming	BMB	Cellulose synthesis in a variety of organisms; Fungal degradation of lignin
Harvill, Eric	VBS	Interactions between microbial pathogens and host immunity in the mouse model
Luthe, Dawn	Plant Sci	Plants response to environmental stresses at the physiological, biochemical and molecular levels
Patterson, Andrew	VBS	Impact of the manipulation of gut bacteria through xenobiotic exposure on bile acid pools, metabolism, and interactions with host nuclear receptors
Liu, Wangsheng	Animal Sci	Functional characterization of the PRAME/PRAMEY gene family during spermatogenesis in cattle and mice
Baumrucker, Craig	Animal Sci	Endocrine regulation of mammary cell proliferation, differentiation and mechanisms of cellular transport
Ades, Sarah	BMB	Signal transduction and antibiotic induced stress responses in bacteria
Cameron, Craig	BMB	RNA polymerases and RNA-binding proteins in viral infection and mitochondrial disease
Ferry, Greg	BMB	Enzymology and molecular biology of anaerobic microbes from the Archaea domain
Gilmour, David	BMB	Transcriptional regulation of the hsp70 heat shock gene in Drosophila
Golbeck, John	BMB, Chem	Photosystem I of cyanobacteria and plants (**LOS, NIH R01** Characterization of the putative ferredoxin binding sites on Photosystem I)
Nixon, Tracy	BMB	Signal transduction and gene regulation by AAA+ATPases in bacteria; Degradation of

		lignocellulose by cellulases; Synthesis of cellulose
Postle, Kathleen	BMB	Signal transduction and iron transport in bacteria (**LOS, NIH R01** Energy Transduction Between Membranes)
Tan, Song	BMB	Structural biology of eukaryotic gene regulation
Wood, Thomas	BMB, ChemE	Genetic basis of biofilm formation with E. coli and P. aeruginosa; The role of toxin/antitoxin systems and cryptic prophages
Benkovic, Stephen	Chem	Enzyme catalysis
Bevilacqua, Philip	Chem	Role of RNA-activated protein kinase in innate immunity
Boehr, David	Chem	Biophysical chemistry of enzymes and its applications in rational drug design and protein enginnering (**Seed Grant, $2K**; Peptide library screening for a tryptophane dehydrogenase inhibitor)
Green, Michael	Chem	Role of thiolate ligands in oxidative heme chemistry; Selenium-containing proteins
Cosgrove, Daniel	Biology	Mechanisms of plant cell growth
Guiltinan, Mark	Biology	Plant biotechnology and molecular biology (**NSF, coPI,** Exploring the Genetic Diversity of Theobroma cacao Germplasm Resources for Disease Resistance Genes)
Hughes, David	Biology	Role of behavior in disease transmission (**Seed Grant, $10K**; The proteomic landscape of parasite manipulation of host behavior)
Mao, Yingwei	Biology	Regulation of neurogenesis using mouse models and human stem cells; Therapies for mental illness
Medina, Monica	Biology	Reproductive biology
Diaz, Francisco	Animal Sci	Reproductive biology (**NIH R03, $149K**; Biochemical and proteomic analysis of cumulus cell secreted zinc inhibitory factor)
Kuldau, Gretchen	Plant Path	Mycotoxigenic fungi and their interactions with plants
Korzick, Donna	Kinesiology	Post-menopausal women, aging and

		cardiovascular disease
Broach, James	Hershey BMB	Cellular regulation of growth and development in response to environmental conditions, e.g. nutritional status and external stress
Boal, Amie	BMB, Chem	Structural biology of metalloproteins
Reese, Joseph	BMB	Chromatin structure and gene expression, DNA damage resistance pathways
Ott, Troy	Animal Sci	Physiology of reproduction; Reproductive immunology
Pate, Joy	Animal Sci	Regulation of luteolysis, prostaglandin production by the corpus luteum, and the interactions between the immune and the reproductive systems
Vijay-Kumar, Mantam	Nutritional Sci	Gut microbiotal interactions in metabolic diseases
Mastro, Andrea	BMB	Breast cancer and immune system (**LOS,** Characterization of compositional changes in extracellular matrix in the presence and absence of metastatic prostate cancer cells)
Lambert, Joshua	Food Sci	Food toxicology, disease prevention by dietary components (**2 LOS, USDA AFRI and AICR**; Suppression of inflammatory bowel disease by dietary soy)
Meredith, Tim	BMB	Lipid metabolism in key human pathogens including Staphylococcus aureus and Pseudomonas aeruginosa.

*BMB – Biochemistry and Molecular Biology; Chem – Chemistry; Eng Sci – Engineering Science; VBS – Veterinary and Biomedical Sciences; Plant Sci – Plant Science; ChemE – Chemical Engineering; Animal Sci – Animal Science; Plant Path – Plant Pathology; Nutritional Sci – Nutritional Sciences; Food Sci – Food Science

18. Extent of Clinical Activities Initiated and Completed. Items 18(A) and 18(B) should be completed for all research projects. If the project was restricted to secondary analysis of clinical data or data analysis of clinical research, then responses to 18(A) and 18(B) should be "No."

18(A) Did you initiate a study that involved the testing of treatment, prevention or diagnostic procedures on human subjects?

_____Yes
__X___No

18(B) Did you complete a study that involved the testing of treatment, prevention or diagnostic procedures on human subjects?

_____Yes
__X___No

If "Yes" to either 18(A) or 18(B), items 18(C) – (F) must also be completed. (Do NOT complete 18(C-F) if 18(A) and 18(B) are both "No.")

18(C) How many hospital and health care professionals were involved in the research project?

_____Number of hospital and health care professionals involved in the research project

18(D) How many subjects were included in the study compared to targeted goals?

_____Number of subjects originally targeted to be included in the study
_____Number of subjects enrolled in the study

Note: Studies that fall dramatically short on recruitment are encouraged to provide the details of their recruitment efforts in Item 17, Progress in Achieving Research Goals, Objectives and Aims. For example, the number of eligible subjects approached, the number that refused to participate and the reasons for refusal. Without this information it is difficult to discern whether eligibility criteria were too restrictive or the study simply did not appeal to subjects.

18(E) How many subjects were enrolled in the study by gender, ethnicity and race?

Gender:
_____Males
_____Females
_____Unknown

Ethnicity:
_____Latinos or Hispanics
_____Not Latinos or Hispanics
_____Unknown

Race:
_____American Indian or Alaska Native
_____Asian
_____Blacks or African American
_____Native Hawaiian or Other Pacific Islander
_____White
_____Other, specify:_____
_____Unknown

18(F) Where was the research study conducted? (List the county where the research study was conducted. If the treatment, prevention and diagnostic tests were offered in more than one county, list all of the counties where the research study was conducted.)

19. Human Embryonic Stem Cell Research. Item 19(A) should be completed for all research projects. If the research project involved human embryonic stem cells, items 19(B) and 19(C) must also be completed.

19(A) Did this project involve, in any capacity, human embryonic stem cells?
_____Yes
___X_ No

19(B) Were these stem cell lines NIH-approved lines that were derived outside of Pennsylvania?
_____Yes
_____ No

19(C) Please describe how this project involved human embryonic stem cells:

20. Articles Submitted to Peer-Reviewed Publications.

20(A) Identify all publications that resulted from the research performed during the funding period and that have been submitted to peer-reviewed publications. Do not list journal abstracts or presentations at professional meetings; abstract and meeting presentations should be listed at the end of item 17. **Include only those publications that acknowledge the Pennsylvania Department of Health as a funding source** (as required in the grant agreement). List the title of the journal article, the authors, the name of the peer-reviewed publication, the month and year when it was submitted, and the status of publication (submitted for publication, accepted for publication or published.). Submit an electronic copy of each publication or paper submitted for publication, listed in the table, in a PDF version 5.0.5 (or greater) format, 1,200 dpi. Filenames for each publication should include the number of the research project, the last name of the PI, and an abbreviated title of the publication. For example, if you submit two publications for Smith (PI for Project 01), one publication for Zhang (PI for Project 03), and one publication for Bates (PI for Project 04), the filenames would be:

Project 01 – Smith – Three cases of isolated
Project 01 – Smith – Investigation of NEB1 deletions
Project 03 – Zhang – Molecular profiling of aromatase
Project 04 – Bates – Neonatal intensive care

If the publication is not available electronically, provide 5 paper copies of the publication.

Note: The grant agreement requires that recipients acknowledge the Pennsylvania Department of Health funding in all publications. Please ensure that all publications listed

acknowledge the Department of Health funding. If a publication does not acknowledge the funding from the Commonwealth, do not list the publication.

Title of Journal Article:	Authors:	Name of Peer-reviewed Publication:	Month and Year Submitted:	Publication Status (check appropriate box below):
1. PHYRN: A Robust Method for Phylogenetic Analysis of Highly Divergent Sequences	Bhardwaj G, Ko K, Hong Y, Zhang Z, Ho NL, Chintapalli SV, Kline L, Gotlin M, Hartranft D, Patterson M, Dave F, Smith EJ, Holmes E, Patterson RL, van Rossum DB	PLOS ONE (2012) 7(4):e34261 . PMID: 22514627	November 2011	☐Submitted ☐Accepted X Published
2. Reevaluation of the evolutionary events within recA/RAD51 phylogeny	Chintapalli SV, Bhardwaj G, Babu J, Hadjiyianni L, Hong Y, Todd GK, Boosalis CA, Zhang Z, Zhou X, Ma H, Anishkin A, van Rossum DB, Patterson RL	BMC Genomics (2013) Apr 10; 14(1):240. PMID: 23574621	September 2012	☐Submitted ☐Accepted X Published
3. Functional evolution of Erg potassium channel gating reveals an ancient origin for IKr	Martinson AS, van Rossum DB, Diatta FH, Layden MJ, Rhodes SA, Martindale MQ, Jegla T	Proc Natl Acad Sci USA (2014) Apr 15; 111(15):5712-7. PMID: 24706772	November 2013	☐Submitted ☐Accepted X Published
4. TRPV-Channel-Mediated Calcium-Transients in Nociceptor Neurons are Dispensable for Avoidance Behavior	Lindy A, Parekh PK, Zhu R, Kanju P, Chintapalli SV, Tsvilovskyy V, Patterson RL, Anishkin A, van Rossum DB, Liedtke WB.	Nature Communications (2014) Sep 2; 5:4734. PMID: 25178952	August 2013	☐Submitted ☐ Accepted X Published
5. The BRCA1 Tumor Suppressor Binds To Inositol 1,4,5-Trisphosphate Receptors To Stimulate Apoptotic Calcium Release	Hedgepeth SC, Garcia MI, Wagner LE II, Rodriquez AM, Chintapalli SV, Snyder RR, Hankins GDV, Henderson BR, Brodie KM, Yule DI, van Rossum DB, Boehning D	Journal of Biological Chemistry 2015 Feb 2. PMID: 25645916	September 2014	☐Submitted ☐ Accepted X Published

20(B) Based on this project, are you planning to submit articles to peer-reviewed publications in the future?

Yes___X_____ No_____

If yes, please describe your plans:

The manuscript titles listed below are in various stages of preparation. Each was supported by the CURE grant.

Li X, Liu H, Luo JC, Rhodes SA, Trigg LM, van Rossum DB, Anishkin A, Diatta FH, Sassic JK, Simmons DK, Kamel B, Medina M, Martindale MQ, Jegla T. A Major Diversification of Voltage-gated K+ Channels Occurred in Ancestral Parahoxozoans .

Chintapalli SV, Bhardwaj G, Patel R, Shah N, Anishkin A, van Rossum DB, Patterson RL, Adams SH. Molecular Dynamic Simulations reveal the structural determinants of fatty acid binding to oxy-myoglobin.

Chintapalli SV, Bhardwaj G, Anishkin A, van Rossum DB, Knotts TA, Adams SH. Mechanistic Insight into Acylcarnitine Binding to Myoglobin Revealed Through Molecular Dynamics.

Zachos NC, Chakraborty M, van Rossum DB, Patterson RL, Donowitz M. A Peptide Mimicking the NHE3 C-terminal Domain Prevents Luminal Fluid Accumulation Induced by Cholera Toxin in the Intact Mouse Jejunum: A Potential Model for Drug Therapy for Diarrhea.

Bhardwaj G, Chintapalli SV, Ho NL, Zhang Z, Hong Y, Babu J, Rabba M, Hadjiyianni L, Anishkin A, Patterson RL, van Rossum DB. Enhanced Measures of Phylogenetic Signal across the Landscape of Viral RNA-dependent RNA Polymerases.

Anishkin, A, Chintapalli SV, Bhardwaj G, Babu J, Hadjiyianni L, Hong Y, Zhang Z, van Rossum DB. Spatial Dynamic 3D Representation of Sequence Similarity Networks using Molecular Visualization Software.

21. Changes in Outcome, Impact and Effectiveness Attributable to the Research Project. Describe the outcome, impact, and effectiveness of the research project by summarizing its impact on the incidence of disease, death from disease, stage of disease at time of diagnosis, or other relevant measures of outcome, impact or effectiveness of the research project. If there were no changes, insert “None”; do not use “Not applicable.” Responses must be single-spaced below, and no smaller than 12-point type. DO NOT DELETE THESE INSTRUCTIONS. There is no limit to the length of your response.

None

22. Major Discoveries, New Drugs, and New Approaches for Prevention Diagnosis and Treatment. Describe major discoveries, new drugs, and new approaches for prevention, diagnosis and treatment that are attributable to the completed research project. If there were no major discoveries, drugs or approaches, insert "None"; do not use "Not applicable." Responses must be single-spaced below, and no smaller than 12-point type. DO NOT DELETE THESE INSTRUCTIONS. There is no limit to the length of your response.

Diarrheal diseases are the second leading cause of infant mortality world-wide. They kill due to dehydration from intestinal loss of water and electrolytes. A critical lack in current diarrheal therapy is availability of a drug that can stimulate intestinal Na+ absorption. In the human small intestine, the Na^+/H^+ Exchanger Isoform 3 (NHE3) protein accounts for the majority of Na^+ absorption and is inhibited in both inflammatory and enterotoxigenic diarrheas. While chloride secretion leads to massive fluid loss in some diarrheal diseases such as cholera, inhibition of Na^+ absorptions occurs in almost all diarrheal diseases. There still exists, therefore, an unmet need for improved treatments for gastrointestinal diseases such as diarrhea. Using a combination of sequence analysis methods and physiological assays we developed a bioactive peptide therapeutic based on a soluble portion of NHE3 which ameliorates cholera toxin in mouse model.

23. Inventions, Patents and Commercial Development Opportunities.

23(A) Were any inventions, which may be patentable or otherwise protectable under Title 35 of the United States Code, conceived or first actually reduced to practice in the performance of work under this health research grant? Yes____X____ No_____

If "Yes" to 23(A), complete items a – g below for each invention. (Do NOT complete items a - g if 23(A) is "No.")

a. Title of Invention: Compositions which regulate the epithelial brush border Na^+/H^+ exchanger NHE3 and their methods of use in gastrointestinal disease

b. Name of Inventor(s): Damian B. van Rossum (PSU), Randen L. Patterson (PSU), Nick C. Zachos (Johns Hopkins University), Mark Donowitz (JHU)

c. Technical Description of Invention (describe nature, purpose, operation and physical, chemical, biological or electrical characteristics of the invention):

In accordance with an embodiment, the present invention provides a sodium/hydrogen Exchanger 3 (NHE3) Inhibitory Regulatory Complex ICRX Competing Peptide (ICRX-CP), or a functional portion or analog or derivative thereof. In accordance with another embodiment, the present invention provides a peptide wherein the peptide stimulates basal NHE3 activity and prevents elevated Ca^{2+} and cAMP inhibition of NHE3. In accordance with a further embodiment, the present invention provides a pharmaceutical composition comprising the peptide, and a pharmaceutically acceptable carrier. In accordance with an embodiment, the present invention provides a pharmaceutical

composition comprising the peptide, and a second therapeutic agent, and a pharmaceutically acceptable carrier. In accordance with a further embodiment, the present invention provides a method of treatment of a gastrointestinal disease in a subject comprising administering to the subject the ICRX-CP peptides described herein, or the pharmaceutical compositions described herein, in an effective amount to modulate or treat the gastrointestinal disease. In accordance with an embodiment, the present invention provides the use of the ICRX-CP peptide heretofore described above, in preparation of a medicament, preferably a medicament for use in treatment of a gastrointestinal disease in a mammal.

d. Was a patent filed for the invention conceived or first actually reduced to practice in the performance of work under this health research grant?
Yes____X___ No____

If yes, indicate date patent was filed: Provisional US patent application filed on May 17, 2012

e. Was a patent issued for the invention conceived or first actually reduced to practice in the performance of work under this health research grant?
Yes____X___ No____

If yes, indicate number of patent, title and date issued:
Patent number: US2013041514

Title of patent: Compositions which regulate the epithelial brush border Na^+/H^+ exchanger NHE3 and their methods of use in gastrointestinal disease
Date issued: International Patent issued 12/22/14

f. Were any licenses granted for the patent obtained as a result of work performed under this health research grant? Yes_____ No ____X_____

If yes, how many licenses were granted?_____

g. Were any commercial development activities taken to develop the invention into a commercial product or service for manufacture or sale? Yes___ No__X___

If yes, describe the commercial development activities:

23(B) Based on the results of this project, are you planning to file for any licenses or patents, or undertake any commercial development opportunities in the future?

Yes_____ No__X___

If yes, please describe your plans:

24. Key Investigator Qualifications. Briefly describe the education, research interests and experience and professional commitments of the Principal Investigator and all other key investigators. In place of narrative you may insert the NIH biosketch form here; however, please limit each biosketch to 1-2 pages.

Damian B. van Rossum, Ph.D. is a Research Assistant Professor in the Department of Biology at The Pennsylvania State University and the Director of the Center of Computational Proteomics in the Huck Institutes of the Life Sciences. He conducted his Ph.D. dissertation in the laboratory of Donald L. Gill, Ph.D., at the University of Maryland, Baltimore School of Medicine, graduating in 2003. Following his doctoral work in Biochemistry and Molecular Biology, Dr. van Rossum was supported by a postdoctoral fellowship in the Department of Neuroscience at Johns Hopkins University School of Medicine, laboratory of Solomon H. Snyder, MD. His research focused on the elucidation of non-canonical functions for signaling proteins using a combination of computational and biochemical approaches. He came to Penn State University in 2006 where he continued to leverage his expertise to derive enhanced measures of cellular systems. His main focus is modeling proteins involved in the signal transduction of short-term and long-term cellular processes; specifically those underlying human health and disease states. He has a specific interest in structural and functional annotation of primary amino-acid sequences, phylogenetic inference in highly divergent protein families, and bioengineering of therapeutic molecules.

Andriy Anishkin, Ph.D. obtained his terminal degree in Biophysics in 1999 (Kharkiv National University, Ukraine), and received postdoctoral training in electrophysiology and computational mechanics of mechanosensitive ion channels in the laboratory of Dr. Sukharev (University of Maryland, College Park). Dr. Anishkin is a Biophysicist and Computational Biologist with 14-years of experience in steered molecular-dynamics and structure refinement in proteins, specifically the large-scale motion of membrane-associated proteins and spatial representation of multidimensional data. He has advanced expertise in predictive and descriptive modeling and algorithmic design, specifically coding, implementation, and dissemination. Dr. Anishkin has developed several novel computational techniques that critically advanced exploration of the conformational spaces for multimeric molecular assemblies. He used these approaches in collaborative projects on bacterial mechanosensitive channels (MscL and MscS), channels from the TRP channel superfamily (TRPY1, OSM9, TRPV4), two-pore channels (TREK-1, TRAAK), proapoptotic ceramide-formed channels, assemblies of amyloid beta oligomers, and Asparaginase type-2 enzymes used in anticancer therapy. His training allows him to combine wet and dry-lab approaches towards refining models and providing testable predictions

Tatiana Laremore, Ph.D. graduated from the College of Saint Rose, Albany, NY in 2003 with a B.A. in Chemistry (*summa cum laude*) and continued her education at the Rensselaer Polytechnic Institute, Troy, NY. In 2006, Tatiana completed her M.S. degree in Chemistry with a research thesis entitled ‘An evaluation of the in vacuo methylation procedure in mass spectrometry-based proteomics’, adviser C.T. Choma. In 2007, she finished her graduate studies at Rensselaer with a Ph.D. dissertation entitled ‘MALDI-MS applications in

glycomics of glycosaminoglycans', adviser R.J. Linhardt. Dr. Laremore continued her work on glycosaminoglycan characterization in Dr. Linhardt's group as a post-doctoral researcher using a combination of enzymology and high-resolution Fourier-Transform mass spectrometry until 2010 when she joined the Huck Institutes of the Life Sciences as the Proteomics and Mass Spectrometry Core Facility Director. Dr. Laremore's research interests include general proteomics, protein and proteoglycan characterization by FT MS, biological mass spectrometry, and MALDI-TOF and TOF-TOF method development (e.g. microorganism characterization, tissue imaging, matrix-free laser desorption).