SCIENCE & TECHNOLOGY POLICY INSTITUTE

Informatics Technologies for Cancer Research (ITCR) Case Study Report

Brian L. Zuckerman
Ian D. Simon
Katherine M. Kowal
William E. J. Doane

October 2018

Approved for public release; distribution is unlimited.

IDA Document D-10311

Log: H 18-000443

IDA SCIENCE & TECHNOLOGY POLICY INSTITUTE
1701 Pennsylvania Ave. NW, Suite 500
Washington, DC 20006-5805

The Institute for Defense Analyses is a non-profit corporation that operates three federally funded research and development centers to provide objective analyses of national security issues, particularly those requiring scientific and technical expertise, and conduct related research on other national challenges.

About This Publication

This work was conducted by the IDA Science and Technology Policy Institute under under contract NSFOIA0408601, Project NC-20-4482, "Evaluation of the NCI Informatics Technology for Cancer Research (ITCR) Program," for the National Cancer Institute (NCI). The views, opinions, and findings should not be construed as representing the official positions of the National Science Foundation or the sponsoring office.

Acknowledgements

Thanks to Erika Tildon and Mark Mancuso for their editorial assistance in completing this report. Thanks also to Daniel Bernstein, loather of broken links in public documents, for bringing fresh eyes to the final draft.

For More Information

Brian L. Zuckerman, Project Leader
bzuckerm@ida.org, 202-419-5485

Mark J. Lewis, Director, IDA Science and Technology Policy Institute
mjlewis@ida.org, 202-419-5491

SCIENCE & TECHNOLOGY POLICY INSTITUTE

IDA Document D-10311

Informatics Technologies for Cancer Research (ITCR) Case Study Report

Brian L. Zuckerman
Ian D. Simon
Katherine M. Kowal
William E. J. Doane

Executive Summary

Introduction

The Informatics Technology for Cancer Research (ITCR) Program is a trans-National Cancer Institute (NCI) program supporting investigator-initiated informatics technology development driven by critical needs in cancer research. The program began in 2012 and was first renewed in 2015. In support of a second renewal request, NCI requires an independent evaluation of the program. In February 2018, NCI asked the IDA Science and Technology Policy Institute (STPI) to conduct three activities: (1) facilitate an expert panel process intended to provide evaluative insights regarding the program's rationale and impact to date; (2) survey current and former ITCR awardees to identify the program's role in fostering collaboration; and (3) conduct a set of case studies of ITCR impact at an award level. This report responds to the third and final charge.

STPI researchers began by working with NCI staff to identify a set of cases balanced across three selection criteria: (1) Award activity code (Exploratory/Developmental Research Grant or "R21" awards for algorithm development versus Research Project Cooperative Agreement or "U01" awards for early-stage development versus Resource-Related Research Projects–Cooperative Agreements or "U24" awards for late-stage development and sustainment); (2) whether awardees have participated in multiple rounds of ITCR funding or a single round of funding; and (3) the nature of the informatics technologies supported. A final contribution to case selection is that STPI chose solely from among the 46 principal investigators (PIs) (out of 56 total ITCR awardees through 2017) who had returned the STPI-administered collaboration surveys—the second of the three STPI activities described above. STPI developed nine cases concerning software platforms (U01 and U24 awards) and two cases from among the R21 algorithm development awards.

Cases cover six topics: (1) the goals and needs underlying tool development; (2) the state of tool development to date; (3) interactions with other tools and funded ITCR teams, including specifically the funding ITCR U01 and U24 awardees are required to reserve for collaborations ("set-aside funding"); (4) measuring impact with respect to users, uses, and user communities, including descriptions of translational research and clinical uses and publications and citations as measures of impact; (5) future development and long-term tool sustainability; and (6) final considerations. Data were collected through PI interviews (following the set of case study topics) and reviews of ITCR-supported publications, presentations, and internet sites.

Cross-Case Findings

Cross-case summary findings were identified with respect to collaborations; uses, users and impact; and future development and sustainability.

Collaborations

In nine of the 11 cases, ITCR-supported tools have already been linked to at least one ITCR-supported effort. These reported collaborations follow a bimodal distribution, with five teams reporting a single, or two, ITCR-ITCR collaborations; four teams reported four or more ITCR-ITCR collaborations. Even in the two cases when ITCR-supported tools have not yet been linked to other ITCR-supported efforts, there is interest in linking tools in the future. In six of the nine cases, the links are made to other tools described in these case studies. The genomics and proteomics (collectively referred to by the program as "-omics") tools increasingly are coming to function as a single, interconnected suite of capabilities—users may enter through a particular portal or analysis tool but then can call upon the tools developed by multiple ITCR-funded groups. In many cases, the collaborations were described as being fostered through network activities, including ITCR set-aside funding and annual investigator meetings.

Interactions with industry represented a second notable aspect of collaboration. Four of the teams reported industry collaborations, though the nature of those collaborations varied. In three cases software firms provide support for tool development and maintenance. In one case companies are also involved in the development of the algorithms underlying the tool itself. In a single case there were already reports of industry users of the tool as of summer 2018 (although multiple teams expect their tools to have industry users eventually).

Uses, Users, and Impact

There is no single common measure of the size of user communities (e.g., based upon number of internet site hits, number of user profiles, or number of downloads) across the 11 ITCR case studies. ITCR teams vary with respect to the extent to which they have embedded mechanisms for tracking users and uses into their software platforms; they also vary with respect to how they describe their user communities. Some teams measure communities by the number of users, while others measure by the number of downloads. Some teams reported monthly or annual data, while others reported cumulative data. Some teams cannot measure use of the ITCR-supported tool directly (e.g., ITCR-supported functionality is embedded in another tool; ITCR has supported multiple individual algorithms not packaged as a single piece of software). Any estimates of impact based on reports of users, therefore, may understate the research impact of at least some ITCR tools. This challenge is especially acute with respect to the integrated -omics tools, where a single user may interact with multiple ITCR-developed tools unknowingly. With these caveats,

there appears to be a range of sizes of user communities associated with the 11 ITCR-supported tools analyzed. Some of the newer tools that are still under development report tens of users, while more established tools—especially those that are building upon pre-existing platforms—report thousands of users per month.

ITCR-supported tools are intended for use by the cancer research community. One measure of the influence tools may have with respect to research is their acknowledgement (e.g., through the citation of an ITCR-supported publication that describes the tools, through their mention in the body of a publication authored by non-ITCR investigators without citation, or the acknowledgement of the funding source for the tool in the publication by ITCR investigators). ITCR investigators reported challenges in relying upon acknowledgements as a measure of tool use and influence. One consideration is that some tools' internet sites ask users to reference a single suggested citation, while in other cases PIs described multiple publications that users acknowledge, making it difficult to rely upon simple citation-based measures to identify publications that report use of a tool.

ITCR-supported teams themselves publish results of research that builds upon or makes use of their tools and acknowledge their ITCR awards in those publications. To the extent to which publications that acknowledge ITCR support are a meaningful measure of research impact, there appears to be a bimodal distribution of publications acknowledging ITCR awards. Three teams' ITCR awards acknowledge more than 20 publications, while three teams' awards acknowledge five or fewer publications. Two ITCR teams have placed multiple articles in very-high-impact journals such as *Science*, *Nature*, and the *New England Journal of Medicine.*

An additional category of use and impact considered is the extent to which ITCR-supported tools are being used for translational research or clinical purposes. Seven of the 11 cases identified translational or clinical uses, although their nature differs. In four cases the tools are being used for or are specifically designed for clinical decision support. In two cases investigators report that tools are being used for "bedside-to-bench" uses—deriving basic research insights from clinical data. In a final example, the tool is being used for drug discovery research. Investigators also report that these translational and clinical uses are expected to accelerate as the tools mature.

Future Development and Sustainability

Awardees described various paths they expect or intend to take with respect to the future development and sustainability of their ITCR-supported tools. All awardees would certainly welcome additional ITCR funds for long-term sustainment. While none of the PIs mentioned that their efforts would immediately cease at the close of their funding without additional NCI support, all PIs noted that continuing development would be required to keep their tools current for them to remain valuable to users. Some awardees prefer to keep

software development in-house (even in the long-term) while others work with (or have considered working with) private firms for support.

Contents

1. Informatics Technologies for Cancer Research (ITCR) Case Study Report

A. Introduction

The Informatics Technology for Cancer Research (ITCR) Program is a trans-National Cancer Institute (NCI) program supporting investigator-initiated informatics technology development driven by critical needs in cancer research. The program was initiated in 2012 and was first renewed in 2015. The program is currently supported through four funding opportunities:

- PAR-15-334 (Exploratory/Developmental Research Grant or "R21"): Development of Innovative Informatics Methods and Algorithms for Cancer Research and Management
- PAR-15-332 (Research Project Cooperative Agreement or "U01"): Early-Stage Development of Informatics Technologies for Cancer Research and Management
- PAR-15-331 (Resource-Related Research Projects—Cooperative Agreements or "U24"): Advanced Development of Informatics Technologies for Cancer Research and Management
- PAR-15-333 (U24): Sustained Support for Informatics Resources for Cancer Research and Management

In support of a second renewal request, NCI requires an independent evaluation of the program. The renewal request was submitted in September 2018 to NCI Scientific Program Leadership and the program evaluation provided important input for preparing and submitting this request.[1]

In February 2018, NCI asked the IDA Science and Technology Policy Institute (STPI) to conduct three activities: (1) facilitate an expert panel process intended to provide evaluative insights regarding the program's rationale and impact to date; (2) survey current and former ITCR awardees to identify the program's role in fostering collaboration; and

[1] Program leadership is considering whether to shift the funding vehicle from a program announcement to requests for applications (RFA). If approved to renew the program funding opportunity announcements as RFAs, NCI's management practices will also require the proposal to be approved by the NCI Board of Scientific Advisors (BSA). Although the program is not currently funded through RFAs, the fact that it has been running for several years suggests that an evaluation is appropriate in support of an RFA request to the BSA.

(3) conduct a set of case studies of ITCR impact at an award level. This report responds to the third and final charge.

B. Methodology

1. Case Selection

In selecting cases, STPI researchers aimed for balance across multiple dimensions, including:

- Award activity code (R21 awards for algorithm development versus U01 awards for early-stage development versus U24 awards for late-stage development and sustainment),
- Whether awardees have participated in multiple rounds of ITCR funding or a single round of funding, and
- The nature of the informatics technologies supported (e.g., -omics[2] technologies versus imaging technologies versus other technology development efforts).

A final contribution to case selection is that STPI chose solely from among the 46 principal investigators (PIs) (out of 56 total ITCR awardees through 2017) who had returned the STPI-administered collaboration surveys—the second of the three STPI activities described above. The decision to limit case selection to survey participants reflected two considerations. First, it was judged that PIs who chose not to respond to the survey (after multiple email and telephonic requests) would also be unlikely to participate in the case study process. Second, given that the survey provided information regarding award progress and collaborations to date, STPI researchers could build upon completed surveys to reduce the burden of the case development process.

2. Case Conduct

STPI researchers began by working with NCI staff to identify a set of cases balanced across the selection criteria. A first round of cases concerned software platforms (U01 and U24 awards) and selected PIs were contacted in April 2018 to arrange interviews, which were conducted with nine PIs between April and June 2018. Interviews covered six topics: (1) the goals and needs underlying tool development; (2) the state of tool development to date; (3) interactions with other tools and funded ITCR teams, including specifically the funding ITCR U01 and U24 awardees are required to reserve for collaborations ("set-aside

[2] The program refers to genomics and proteomics tools collectively as "-omics" tools

funding")[3]; (4) measuring impact with respect to users, uses, and user communities, including descriptions of translational research and clinical uses and publications and citations as measures of impact; (5) future development and long-term tool sustainability; and (6) final thoughts to give PIs the opportunity to add information and perspectives that were not otherwise touched upon during the interview process. STPI staff provided participating PIs with summaries of the interviews for review during July 2018, and then used the corrected versions as part of case development, supplemented by public information (e.g., award abstracts, publications, internet references, and materials from the 2018 ITCR PI meeting) and survey responses. A second wave of interviews in August-September 2018 targeted R21 awardees. Case write-ups follow the format of the interviews.

C. List of Cases and Case Breakdown

1. List of Cases

Nine cases were completed in the first wave:

- Cancer-Related Analysis of VAriants Toolkit (CRAVAT) and the Mutation Position Imaging Toolbox (MuPIT)
- Clinical Interpretation of Variants in Cancer (CIViC)
- Trinity Cancer Transcriptome Analysis Toolkit (CTAT)
- Deep Phenotype Extraction tool (DeepPhe)
- Galaxy for Proteomics project (Galaxy-P)
- Network Data Exchange (NDEx)
- Pathology Image Informatics Platform (PIIP)
- Quantitative Image Informatics for Cancer Research (QIICR)
- Xena genomics visualization tool

Two additional cases, focusing on R21 algorithm development awards, were completed in the second wave:

- AMARETTO regulatory networks analysis
- BayesGO and GAIL cancer subtypes analysis

[3] See for example PAR-15-331, "Collaborative Activities: Applicants must set aside 10 percent of their annual budget (Direct Costs) to support collaborative or joint activities within or beyond ITCR projects, initiated post-award." Available from: https://grants.nih.gov/grants/guide/pa-files/PAR-15-331.html

STPI researchers focused on awards that were in at least their second year and had published by summer 2018 at least one peer-reviewed journal article describing the approach being researched.

2. Cross-Case Breakdown

The cases assembled and in progress were drawn from across the range of ITCR awards. Of the 11 completed cases:

- Two involve R21 ITCR awards, three involve U01 awards, and six involve U24 awards.
- Two involve PIs who received multiple, related awards.
- Seven are -omics-focused, two are imaging-focused, and two represent other informatics technologies.

D. Cross-Case Summary Findings

1. Collaborations

In nine of the 11 cases, ITCR-supported tools have already been linked to at least one ITCR-supported effort (Figure 1). These reported collaborations follow a bimodal distribution, with five teams reporting a single, or two, ITCR-ITCR collaborations; four teams reported four or more ITCR-ITCR collaborations. The MuPIT/CRAVAT and CIViC teams reported collaborations with six other ITCR teams each; the Galaxy-P team reported five; and the NDEx team reported collaborations with four other ITCR teams. Even in the two cases when ITCR-supported tools have not yet been linked to other ITCR-supported efforts, there is interest in linking tools in the future.

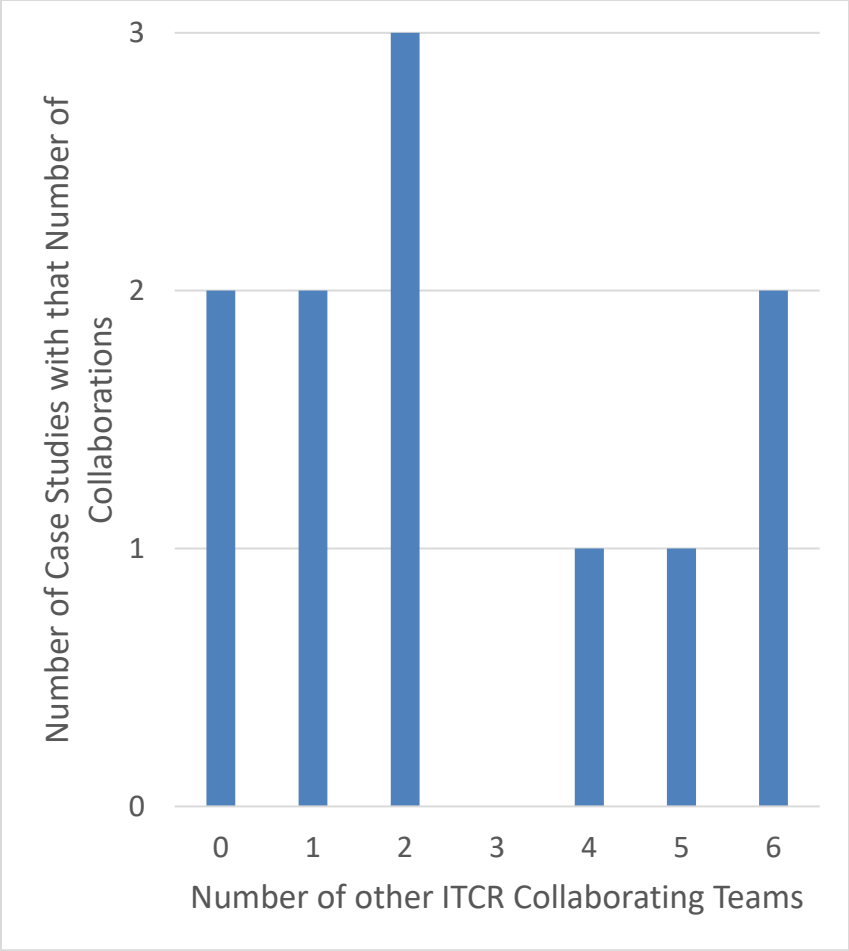

Figure 1: Reported ITCR-ITCR Collaborations

In six of the nine cases, these reported links are made to other tools described in the case studies. The -omics tools increasingly are coming to function as a single, interconnected suite of capabilities—users may enter through a particular portal or analysis tool but then can call upon the tools developed by multiple ITCR-funded groups. In many cases, the collaborations were described as being fostered through network activities, including ITCR set-aside funding and annual investigator meetings. These case study results complement findings from the ITCR collaboration survey showing high levels of satisfaction with network activities and the role they play in fostering new collaborations among ITCR researchers and in strengthening existing ties.[4]

Interactions with industry represented a second notable discussion of collaboration. Four of the teams reported industry collaborations, though the nature of those

[4] For more detail on ITCR network activities and their benefits, see chapter 5 of *Collaboration in the National Cancer Institute (NCI) Informatics Technologies for Cancer Research (ITCR) Initiative*; Cassidy A. Pomeroy-Carter, Brian L. Zuckerman, Justin C. Mary, and Xueying Han, IDA Paper P-9200, July 2018.

collaborations varied. In three cases (MuPIT/CRAVAT, PIIP, QIICR) software firms provide support for tool development and maintenance. In one case (QIICR) companies are also involved in the development of the algorithms underlying the tool itself. In a single case (CIViC) there were already reports of industry users of the tool as of summer 2018 (although multiple teams expect their tools to have industry users eventually).

2. Uses, Users, and Impact

There is no single common measure of the size of user communities (e.g., based upon number of internet site hits, number of user profiles, or number of downloads) across the 11 ITCR case studies. ITCR teams vary with respect to the extent to which they have embedded mechanisms for tracking users and uses into their software platforms. Some teams measure communities by the number of users, while others measure by the number of downloads. Some teams reported monthly or annual data, while others reported cumulative data. Some teams cannot measure use of the ITCR-supported tool directly (e.g., ITCR-supported functionality is embedded in another tool; ITCR has supported multiple individual algorithms not packaged as a single piece of software.) Any estimates of impact based on reports of users, therefore, may understate the research impact of at least some ITCR tools. This challenge is especially acute with respect to the integrated -omics tools, where a single user may interact with multiple ITCR-developed tools unknowingly.

With these caveats, there appears to be a range of sizes of user communities associated with the 11 ITCR-supported tools analyzed. Some of the newer tools that are still under development (e.g., DeepPhe, Galaxy-P, PIIP) report tens of users, while more established tools—especially those that are building upon pre-existing platforms (e.g., Trinity CTAT, UCSC Xena)—report thousands of users per month (Table 1).

Table 1: Reported Users or Users of ITCR-Supported Tools

ITCR Team	Reported Uses or Users
Tools that appear to be used more frequently as of summer 2018	
QIICR	100,000 3-D Slicer downloads/year, including all tools—not just QIICR-supported tools
Xena	7,000-8,000 users/month; 1M hits on hub/month
CTAT	3,000 users/month of Trinity, including CTAT
CIViC	2,500 users, 1M API requests/month
MuPIT/CRAVAT	No user statistics, 1,200 Docker image downloads cumulatively
NDEx	~1,000 users cumulatively
Tools that appear to be more developmental as of summer 2018	
PIIP	~70 users
Galaxy-P	~10-12 users
DeepPhe, AMARETTO, BayesGO/InGRiD	Used primarily within research teams at present

ITCR-supported tools are intended for use by the cancer research community. One measure of the influence tools may have with respect to research is their acknowledgement (e.g., through the citation of an ITCR-supported publication that describes the tools, through their mention in the body of a publication authored by non-ITCR investigators without citation, or the acknowledgement of the funding source for the tool in the publication by ITCR investigators). ITCR investigators reported challenges in relying upon acknowledgements as a measure of tool use and influence. One consideration is that some tools' internet sites ask users to reference a single suggested citation, while in other cases PIs described multiple publications that users acknowledge, making it difficult to rely upon simple citation-based measures to identify publications that report use of a tool.

ITCR-supported teams themselves publish results of research that builds upon or makes use of their tools and acknowledge their ITCR awards in those publications.[5] To the

[5] All of the usual caveats associated with inferring utility or impact from publication acknowledgements remain the case with respect to ITCR-supported research. Some teams have nearly completed their ITCR award periods while others are partway through, making comparisons across teams difficult to interpret. Not all publications properly acknowledge awards, so that some relevant publications may not include their ITCR support. Many publications acknowledge multiple sources of funding, so the specific contribution of the ITCR funding to the research may vary from publication to publication. Some investigators may acknowledge funding at the very beginning of an award period (when the

extent to which publications that acknowledge ITCR support are a meaningful measure of research impact, there appears to be a bimodal distribution of publications acknowledging ITCR awards (Figure 2). Three teams' ITCR awards acknowledge more than 20 publications (PIIP, QIICR, Xena). Two ITCR teams (CTAT, Xena) have placed multiple articles in very-high-impact journals such as *Science*, *Nature*, and the *New England Journal of Medicine.*

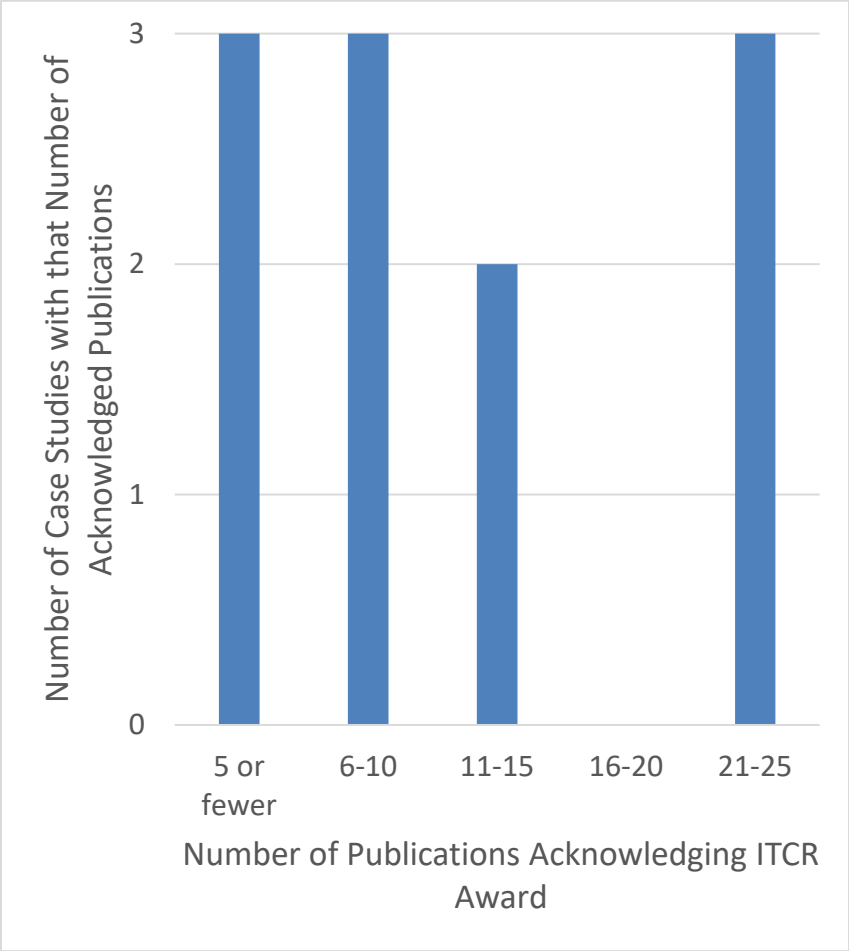

Figure 2: Publications Acknowledging ITCR Awards, as of September 2018

An additional category of use and impact considered is the extent to which ITCR-supported tools are being used for translational research or clinical purposes. Seven of the 11 cases identified translational or clinical uses, although their nature differs. In four cases (CIViC, DeepPhe, PIIP, QIICR), the tools are being used for or are specifically designed for clinical decision support. In two cases (MuPIT/CRAVAT, Xena), investigators report

maturity of the work supported specifically through that award may be limited) or well after the conclusion of an award period.

that tools are being used for "bedside-to-bench" uses—deriving basic research insights from clinical data. In a final example (AMARETTO), the tool is being used for drug discovery research. Investigators also report that these translational and clinical uses are expected to accelerate as the tools mature.

3. Future Development and Sustainability

Awardees described various paths they expect or intend to take with respect to the future development and sustainability of their ITCR-supported tools. All awardees would certainly welcome additional ITCR funds for long-term sustainment. While none of the PIs mentioned that their efforts would immediately cease at the close of their funding without additional NCI support, all PIs noted that continuing development would be required to keep their tools current for them to remain valuable to users. Some awardees prefer to keep software development in-house (even in the long-term) while others have contacted (or have considered contacting) private firms for support.

4. Summary Points

Looking across the cases, CIViC appears to be the best example of a newly developed ITCR tool that: (1) has a substantial user base; (2) includes clinical decision support capacity; (3) has current industry users; and (4) involves many ITCR-ITCR collaborations. NDEx is another example of a newly developed tool with substantial use and ITCR-ITCR collaborations (though to a lesser extent than CIViC). Three other tools (CTAT, QIICR, Xena) built atop pre-ITCR platforms appear to have wide use. The other tools appear still to be maturing as of September 2018.

2. Cancer-Related Analysis of VAriants Toolkit (CRAVAT)/Mutation Position Imaging Toolbox (MuPIT)

A. Background and Goals

One challenge facing cancer researchers is that while there has been an explosion of available genomic data, there has been a lack of software that could annotate and interpret large datasets of cancer mutations from genomic data and was usable by scientists who were not bioinformatics experts (or who did not have in-house access to a bioinformatics team). Rachel Karchin's research group at Johns Hopkins University developed two tools (released in 2012, described in 2013 publications) for high-throughput analysis of mutations potentially related to carcinogenesis—the Cancer-Related Analysis of VAriants Toolkit (CRAVAT) and the Mutation Position Imaging Toolbox (MuPIT).[6] CRAVAT maps mutations to transcriptome and proteomic information, while MuPIT maps mutations to X-ray crystal structures of proteins stored in the Protein Data Bank (PDB). Professor Karchin's team has received two awards from ITCR for further development of these tools, U01CA180956 (active 2013–2016) and U24CA204817 (active 2016–2021). An additional challenge facing the community is that many investigators are developing machine learning classifiers, visualizers, and annotators of genomic data, so that software tools, to be useful, need to be able to provide a unified view across many analytical tools for a multi-dimensional perspective.

B. State of Development

Over the course of the two awards, the Karchin group has developed a suite of tools—machine learning classifiers, annotation tools, data aggregation tools, and visualization widgets—that can be run sequentially as part of a computational pipeline or run individually. Interactivity features make it easy for users who are not computer scientists and who do not have a professional bioinformatics team to get detailed and graphical reports that visualize information about the mutations that are hard to derive from a text or table format. The tools allow filtering and sorting of a large volume of mutations, followed

---

6 C. Douville et al., "CRAVAT: Cancer-Related Analysis of Variants Toolkit," *Bioinformatics* 29, no. 5 (March 1, 2013): 647–8, doi: 10.1093/bioinformatics/btt017; and N. Niknafs et al., "MuPIT Interactive: Webserver for Mapping Variant Positions to Annotated, Interactive 3D Structures," *Human Genetics* 132, no. 11 (November 2013): 1235–43, doi: 10.1007/s00439-013-1325-0. Both were supported by R21CA152432.

by in-depth exploration. The result is a selected number of mutations that users may want to prioritize for functional studies. Tools are available from the Karchin laboratory's internet site (http://karchinlab.org/). Tutorials are available from the site as well. Installation instructions, especially for use in a cloud environment, can be found at: https://hub.docker.com/r/karchinlab/cravatmupit/. The team has conducted several workshops at NCI and conducted invited workshops at the American Society of Human Genetics annual meetings in 2017 and 2018.

The group's efforts have been shifting from a fully web-based implementation to a more portable, local installation, the Python package OpenCRAVAT. OpenCRAVAT allows lightweight, modular installations of selected tools in the CRAVAT suite to a user's local machine. It includes mechanisms for scientists outside the Karchin team to contribute and distribute tools. In addition, a Docker image of the system is available for download for users with protected data. Galaxy tools have also been developed to incorporate CRAVAT into Galaxy pipelines. The group has developed a new machine learning tool to predict cancer driver mutations. This is the first method with a tool for each cancer type, since different cancer types have different driver mutations.

C. Interactions with Other Tools

An overarching theme regarding collaboration in the context of this ITCR award is that collaborations have arisen in response to emerging opportunities. The set of collaborations identified below is as of May 2018; additional collaborations are being pursued and a large number of potential future data clients have been identified. The software is released under an open MIT license to facilitate re-use. The well-developed web services interfaces and ability to link to CRAVAT and MuPIT with structured uniform resource locators (URLs) facilitates use of these tools by others. The benefits of linking their tools with those of other ITCR groups is that the team's most current version of the tool propagates to each new integration with others' tools.

Collaborations with other ITCR-funded teams:

- *NDEx*: CRAVAT was an early adopter of NDEx. The latest version of CRAVAT takes network data as an input and maps and visualizes pathway variants. The interaction was supported through a 2015 administrative supplement.
- *Galaxy-P*: Galaxy-P is linked to CRAVAT. The intent is that when a user runs a workflow in Galaxy, the tool will provide results to CRAVAT in order to better understand what is already known about a particular gene and how it encodes a protein. The Galaxy-P collaboration resulted in spontaneous pipelines when the two tools were integrated; Galaxy-P then incorporated NDEx network

capabilities. The Karchin group has recently submitted a paper with the Galaxy-P group.

- *UCSC Xena:* MuPIT visualizes mutations in three dimensions; the collaborating teams have developed a workflow so that MuPIT can accept Xena data to map mutations using MuPIT's method for visualizing protein structure. The interaction was supported through a 2014 administrative supplement.
- *Trinity CTAT:* CRAVAT maps and visualizes mutations; data from Trinity are fed into CRAVAT to identify and prioritize mutations of interest. If variants are in the MuPIT dataset, a three-dimensional visualization can be provided. These functionalities are fully integrated into CTAT. The interaction was supported through a 2014 administrative supplement.
- *CIViC*: CRAVAT incorporates CIViC data so that if a mutation analyzed by CRAVAT is listed in CIViC, data on that mutation can be displayed.
- *NG-CHM*. NG-CHM incorporates MuPIT data so that users can view the structure of genes included in a heat map produced by the NG-CHM tool.

Collaborations with other academic investigators: The Karchin group has a large number of research collaborators. One advantage of the OpenCRAVAT approach is that it will make it easy for other developers to plug in their tools and build on the existing infrastructure that handles mapping from the genome to RNA transcripts, to protein sequence, and finally protein structure, so developers do not have to rebuild that for themselves.

ITCR set-aside funding is being used to link the BRCA Exchange with MuPIT, to enable BRCA variant curators to assess the importance of missense variants identified in BRCA genetic testing in the context of three-dimensional protein structure/function relationships. The Karchin team has also integrated their tools with the Kaviar Genomic Variant Database and PeptideAtlas.

Collaborations with industry: In Silico Solutions provides software development support to the MuPIT and CRAVAT projects.

D. Measuring Impact: Uses, Users, and User Communities

1. Tracking Users

The Karchin laboratory tracks user submissions to the web portal and the analysis tools users request. The Docker images of MuPIT and CRAVAT are reported to have approximately 1,200 downloads to date; additionally, these tools are integrated with many other genomics tools and accessible through those platforms.

2. Translational Research and Clinical Use

No current direct clinical uses have been reported. As an example of uses of the tool in translational research, users are submitting data from medium-size and large studies with 20–50 and up to several hundred patients in order to conduct "bedside-to-bench" cancer research.

3. Publication and Citation as Measures of Use

As of September 2018, the combination of the two awards has been acknowledged in eight publications. Team members report that there have been multiple important publications describing the MuPIT and CRAVAT tools. Therefore, there is no single "founding" publication that can be used for tracking citations as a measure of use in the academic, basic research community. There have been hundreds of citations in total of these tools spread across many publications.

E. Future Development and Sustainability

The Karchin team plans to include a large community of developers into OpenCRAVAT and to integrate with an additional half-dozen ITCR tools. In the near future, the team is hoping to hire a consultant and a new postdoc to focus on OpenCRAVAT outreach to developers and users with the goal of widening their developer and user base in preparation for a potential eventual ITCR U24 sustainment application. In addition, the team is developing modules that are relevant outside NCI, which will open new funding opportunities. Other options include possible commercialization efforts.

F. Final Thoughts

- ITCR set-asides to support collaboration have been helpful.
- Annual ITCR PI meetings have been helpful to facilitate collaboration.
- ITCR is one of a kind; it has become the lifeblood of the computational genomics community.
- There would be a huge crisis in this country in terms of computational cancer genomics without ITCR. It has made a substantial difference.

3. Clinical Interpretation of Variants in Cancer (CIViC)

A. Background and Goals

Cancer researchers and clinicians are challenged by variant interpretation in cancer medicine. High-throughput sequencing has been largely automated, allowing rapid identification of somatic and germline variants in tumors. Many mutations have been identified. In some cases it is understood that particular mutations are clinically actionable (predisposing patients to develop cancer, diagnostic of tumor subtype, prognostic of survival change, or predicting therapeutic response), but the meaning and clinical relevance of others is unclear. Clinical interpretation of genomic alterations—especially given that studies describing potential variants of significance are growing exponentially—remains a major bottleneck for realizing precision medicine. To address this challenge, in 2016 NCI made award U01CA209936 to a team led by Obi Griffith at Washington University, St. Louis, to develop Clinical Interpretation of Variants in Cancer (CIViC). CIViC is an open-source tool intended to assist clinicians in assessing the clinical import of tumor mutations. While there are other databases of tumor mutations available, CIViC is designed to make use of crowdsourcing (a group of "curators" are responsible for ensuring that there is evidence underlying statements of clinical relevance) to address potential redundancies and manual efforts at interpreting data in a software tool designed to be easily accessible to clinicians. It is based on an open-source web framework called Ruby on Rails.

B. State of Development

The CIViC tool was first released in 2015 and described in a 2017 *Nature Genetics* issue.[7] CIViC is available via a dedicated internet site (https://civicdb.org/) and through the GitHub software repository of the award PI.[8] The CIViC data are publicly available under a Creative Commons license (https://creativecommons.org/publicdomain/zero/1.0/) while the source code is released under the open MIT License (https://github.com/griffithlab/civic-client/blob/master/LICENSE).

---

7 M. Griffith et al., "CIViC is a Community Knowledgebase for Expert Crowdsourcing the Clinical Interpretation of Variants in Cancer," *Nature Genetics* 49, no. 2 (January 31, 2017): 170–4, doi: 10.1038/ng.3774.

8 "Web client for CIViC: Clinical Interpretations of Variants in Cancer," GitHub, https://github.com/griffithlab/civic-client.

Curated evidence statements are the foundational unit of CIViC. Evidence is curated with structured data (Level, Type, Direction, Variant Origin, and Clinical Significance) and evidence statements are linked to the publications (in PubMed) that represent the source of the information. Assertions summarizing the evidence for clinical significance of genes and variants can draw on multiple lines of evidence with differing levels of significance and quality. Gene-level and variant-level summary pages provide descriptions and links to canonical data sources such as gene-level detail from mygene.info or variant-level detail from myvariant.info. As of May 2018, CIViC incorporated 4,719 interpretations building upon 1,761 papers curated for 1,717 variants, 331 genes, and 205 cancer types. More than 100 contributors have participated. CIViC knowledge represents most tumor types (germline and somatic) but with variable coverage (strong coverage for lung cancer, more limited coverage for pancreatic cancer).

Although the site is live and being utilized, major future enhancements are planned before the award period ends in 2019. Extensive interactions between team members and the user community have identified complexities that require enhancements to the data structure and user interface. Another challenge identified is that given the large number of variants and mutations and the rapid expansion of knowledge, they expect the number of variants incorporated into CIViC to continue to grow rapidly for the foreseeable future. The curation problem represents a critical challenge for CIViC's growth, and the team is exploring methods (e.g., through CIViCmine, working with the Jones group at Michael Smith Genome Sciences Centre in Vancouver, British Columbia and associated with the University of British Columbia) for automating at least some aspects of the process.[9] The team has placed extensive help documentation similar to a tutorial on the CIViC internet site and has begun to add YouTube tutorials. A series of additional videos will be forthcoming.

C. Interactions with Other Tools

An overarching theme regarding collaboration in the context of this ITCR award is that collaborations have arisen as opportunities have emerged. As of July 2018, the team reports that more than 25 software suites are CIViC data clients; additional collaborations are being pursued and a large number of potential future data clients have been identified. Some examples of these interactions and collaborations are described below.

Collaborations with other ITCR-funded teams:

- *cBioPortal*: Provides clinical interpretation of variants alongside pan-cancer genomic information from cBioPortal.

[9] For more information, see "Research," BioNLP@GSC, http://bionlp.bcgsc.ca/.

- *CRAVAT*: CRAVAT incorporates CIViC data so that if a mutation analyzed by CRAVAT is listed in CIViC, data on that mutation can be displayed.
- *GEMINI*: Set-aside funds are being used to develop an application program interface (API) to enable queries to CIViC from GEMINI and SuperSeeker (U24CA209999, MONITORING TUMOR SUBCLONAL HETEROGENEITY OVER TIME AND SPACE, PIs Gabor Marth and Aaron Quinlan) to improve the annotation and interpretation of tumor variants/mutations in the study of tumor subclonal heterogeneity.
- *JBrowse*: Set-aside funds are being used to establish connections between JBrowse (U24CA220441) and the CIViC community knowledgebase through incorporating myvariant.info data into JBrowse and implementing a ProteinPaint feature into JBrowse using protein-level CIViC features as the driving use case.
- *NG-CHM*: NG-CHM incorporates CIViC data so that if a mutation analyzed in a heat map is listed in CIViC, data on that mutation can be displayed.
- *ITCR-Innovative Molecular Analysis Technologies (IMAT) supplement*: The CIViC team has an ITCR supplement, entitled "DEVELOPMENT OF A KNOWLEDGE-DRIVEN SMMIP ASSAY FOR ULTRA-SENSITIVE DETECTION OF CLINICALLY RELEVANT VARIANTS IN CANCER." The supplement was described in the survey as a collaboration with IMAT-funded investigator Stephen Salipante at the University of Washington to develop a cancer sequencing panel based on his smMIP technology, driven by knowledge of clinically actionable variants from the CIViC database. Investigators report that work on the assay has led to development of new features in the CIViC database to support panel designs.

Collaborations with other academic investigators: One aspect of CIViC's development to date identified in the ITCR collaboration survey has been collaboration on both resource development and content creation. The CIViC team has worked with non-ITCR funded researchers to incorporate myvariant.info and mygene.info data into CIViC and incorporate CIViC data into BioGPS,[10] incorporate Disease Ontology (Lynn Schriml)[11] to standardize and—where necessary—expand the representation of cancer subtypes, use an ITCR set-

[10] BioGPS, http://biogps.org/#goto=welcome; C. Wu et al., "BioGPS: Building Your Own Mash-Up of Gene Annotations and Expression Profiles," *Nucleic Acids Research* 44, no. D1 (January 4, 2016): D313–6, doi: 10.1093/nar/gkv1104.

[11] S. M. Bello et al., "Disease Ontology: Improving and Unifying Disease Annotations across Species," *Disease Models & Mechanisms* 11, no. 3 (March 12, 2018): pii: dmm032839, doi: 10.1242/dmm.032839. The publication identifies CIViC specifically as a tool that will benefit from incorporating the Disease Ontology toolset and approach.

aside award to work with the ClinGen Somatic Working Group[12] to develop the CIViC software platform further for use as the standard portal for somatic cancer variant curation for submission to ClinVar by ClinGen (Subha Madhavan),[13] and to work with the Personalized OncoGenomics Program[14] to incorporate their knowledgebase into CIViC.

Collaborations with industry: One example of industry collaboration is the incorporation of CIViC into the Cartagenia Bench Lab product of Agilent Technologies. The Agilent software is used by molecular pathologists to automate their variant filtration and classification workflow. Incorporating CIViC allows pathologists using the Bench Lab to assess the molecular profile of a sample and automatically flag the presence of prognostic, diagnostic, and therapeutic evidence in the CIViC database. As CIViC data are being incorporated, in this case, into a commercial product, the CIViC team does not have direct feedback regarding how the application is being used or the number of users. The CIViC team interacts with Agilent staff, which represents a mechanism for indirect feedback regarding how the tool is being used. ITCR survey responses identified other companies, non-governmental organizations (NGOs), and university investigator groups with whom the CIViC team has worked to help them make use of CIViC data: Cambridgene, CancerStop Android Application, NEXUS Personalized Health Technologies of ETH Zurich—Swiss Variant Interpretation Platform for Oncology, Euformatics OmnomicsNGS, GeneCards, LifeMap Sciences TGex NGS Analysis & Interpretation Platform, MolecularMatch, SolveBio, VarSome, and Wikidata.

D. Measuring Impact: Uses, Users, and User Communities

1. Tracking Users

The team uses Google Analytics to identify the number of unique IP addresses/ individuals that access the site and the number of users of the API; this provides a sense of traffic and impact. To date, growth in use has been steady—there were more than 2,500 web users and more than one million API requests per month as of July 2018. Google Analytics provides the location of their users based on their IP addresses. The team can measure the size of the curation community directly. As of August 2018, there are 42 curators listed on the CIViC internet site. A goal of the CIViC effort is for users to become curators and identify and remedy deficiencies in the knowledge base if they find them. Few curators are not also users. One mechanism for outreach is to discuss CIViC at conferences

---

12 "Somatic Cancer," ClinGen, https://www.clinicalgenome.org/working-groups/somatic/.

13 S. Madhavan et al., "ClinGen Cancer Somatic Working Group – Standardizing and Democratizing Access to Cancer Molecular Diagnostic Data to Drive Translational Research," *Pacific Symposium on Biocomputing* 23 (2018): 247–58, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5728662/.

14 "Personalized OncoGenomics Program," BC Cancer Genome Sciences Centre, http://personalizedoncogenomics.org/.

and meetings, and team members have been invited to present regularly. Growth in use also arises through forming key collaborations in the community.

2. Translational Research and Clinical Use

The tool is intended to have clinical use and impact. Clinicians are using CIViC data (e.g., through the Agilent Cartagenia Bench Lab) for clinical decision support—for example to identify a patient who has a particular variant that makes him or her eligible for a new treatment option or a clinical trial—but that impact will be difficult to capture and quantify. More than 10 letters of support were recently provided for a CIViC U24 application describing clinical uses of CIViC across four continents. It is possible that at some future date CIViC may be incorporated into the workflow of clinical research studies, but that has not yet occurred.

3. Publication and Citation as Measures of Use

As of September 2018, four publications (including the *Nature Genetics* article previously mentioned) acknowledge the ITCR award. Also as of September 2018, the *Nature Genetics* article has been cited 31 times in PubMed Central. Team members report that tracking citations as a measure of use may be valuable in assessing CIViC's impact in the academic, basic research community, but not in the clinical community.

E. Future Development and Sustainability

The team reports that long-term sustainment is, of course, desirable; an ITCR U24 would be of interest and an application has been submitted. They expect to continue development internally rather than relying on a software firm for long-term sustainment and have a long-term development plan and a list of improvements that they would like to undertake. Given the goals of CIViC, they expect to continue with an open-source approach into the future.

F. Final Thoughts

- Quantifiable metrics (e.g., number of paper citations) understate CIViC's impact.
- It is difficult, therefore, to compare CIViC's impact with the impact of bench science research efforts.

4. Cancer Transcriptome Analysis Toolkit (CTAT)

A. Background and Goals

Trinity is an open-source ribonucleic acid (RNA) sequencing/transcriptome sequencing data analysis tool first released in 2011.[15] Unlike tools that compare extant RNA to a reference genome, Trinity instead assembles identified RNA into a reconstructed transcriptome. This approach allows for the development of reference genomes from previously unsequenced organisms. When first released, however, Trinity was not optimized for use by cancer researchers. To address this gap, in 2013 NCI made award U24CA180922 to a team led by Aviv Regev at the Broad Institute, to take the core capability of Trinity and add to it a Cancer Transcriptome Analysis Toolkit (CTAT). In addition, funds were intended to maintain and enhance the core Trinity algorithms and computing infrastructure to ensure that the software is world-class and to conduct outreach and training for community use. A final project goal is to build the community of single-cell RNA-Seq cancer researchers, which was a nascent cancer research approach at the time the award was funded.

B. State of Development

The Trinity team used collaborations with cancer researchers ("driving projects") to identify CTAT needs and to refine approaches. Several driving projects were defined pre-award, while others have begun organically in response to new opportunities. CTAT includes tools to analyze mutations, fusion transcripts, transcript expression, noncoding RNAs, alternative splicing, transcripts from viruses and microbes, and single cell tumor heterogeneity. Trinity (and CTAT) is accessible as a multi-component software suite available to the public via the GitHub software repository;[16] older versions also are available from the SourceForge site.[17] CTAT can be accessed through a portal hosted by the Indiana University National Center for Genome Analysis Support. CTAT is also available through the Bioconda, Docker, and FireCloud open-source cloud computing

---

[15] M. G. Grabherr et al., "Full-Length Transcriptome Assembly from RNA-Seq Data without a Reference Genome," *Nature Biotechnology* 29, no. 7 (May 15, 2011): 644–52, doi: 10.1038/nbt.1883.

[16] See "RNA-Seq De novo Assembly Using Trinity," GitHub, https://github.com/trinityrnaseq/trinityrnaseq/wiki.

[17] "Trinity RNA-Seq Assembly," SourceForge, https://sourceforge.net/projects/trinityrnaseq/files/PREV_CONTENTS/previous_releases/.

resources. Team members comment that outreach occurs through Google Forum and participation in workshops. Online training modules are available.[18] Trinity is licensed under the BSD 3-Clause "New" or "Revised" License (https://github.com/trinityrnaseq/trinityrnaseq/blob/master/LICENSE), which limits promotion of the software by third parties without written consent. The current award period is scheduled to end in August 2018.

C. Interactions with Other Tools

Collaborations with other ITCR-funded teams:

- *MuPIT/CRAVAT*: CRAVAT maps and visualizes mutations; data from Trinity are fed into CRAVAT to identify and prioritize mutations of interest. If variants are in the MuPIT dataset, a three-dimensional visualization can be provided. These functionalities are fully integrated into CTAT. The collaboration arose from an ITCR PI meeting.
- *IGV*: Historically, the Trinity team has been collaborating with the IGV team to develop a FusionInspector tool, which provides evidence of fusion transcripts in an IGV interface integrated into Galaxy. They are working as well with the IGV team to develop standalone reports that can be generated not only through Galaxy but also through a web browser using FireCloud.

Collaborations with other academic investigators: As part of the driving projects, the team reports having scientifically productive collaborations with groups in and outside the Broad Institute. Long-standing collaborations include, for example, work on glioblastoma with the Suva and Bernstein groups at Massachusetts General Hospital.[19] One example of a collaboration arising during the award period has been investigating chronic lymphocytic leukemia transcriptomes with Cathy Wu at Dana-Farber.[20]

Collaborations with industry: No collaborations reported as of September 2018.

---

[18] "Trinity Screencast," Broad Institute, https://www.broadinstitute.org/broade/trinity-screencast; "RNA-Seq Analysis Workshop," GitHub, https://github.com/trinityrnaseq/RNASeq_Trinity_Tuxedo_Workshop/wiki.

[19] See, for example, A. P. Patel et al., "Single-cell RNA-Seq Highlights Intratumoral Heterogeneity in Primary Glioblastoma," *Science* 344, no. 6190 (June 20, 2014): 1396–401, doi: 10.1126/science.1254257.

[20] See, for example, L. Wang et al., "Somatic Mutation as a Mechanism of Wnt/β-catenin Pathway Activation in CLL," *Blood* 124, no. 7 (August 14, 2014): 1089–98, doi: 10.1182/blood-2014-01-552067.

D. Measuring Impact: Uses, Users, and User Communities

1. Tracking Users

Trinity team members report that the tool has approximately 3,000 users per month, including users at 486 institutions in 51 countries as of May 2018. There is no current functionality to identify users of CTAT specifically within the overall Trinity framework, so it is not feasible to identify the number of CTAT users. Team members report that they are considering developing such tracking tools in the future.

2. Translational Research and Clinical Use

No current clinical uses have been reported, but Trinity team members expect that in the future CTAT will be used clinically; they have initiated pilot collaborations with clinicians and industry to explore use of the tools in clinical settings.

3. Publication and Citation as Measures of Use

As of September 2018, 15 publications acknowledge the ITCR award, including four articles in *Science,* two in *Nature,* and two in *Nature Biotechnology*. Team members report that tracking citations as a measure of use may be valuable in assessing CTAT's impact in the academic, basic research community. The team tracks publications using Google Scholar and can identify the cancer-specific subset of papers; approximately 20 percent of the approximately 8,000 publications acknowledging Trinity are cancer-related.

E. Future Development and Sustainability

Additional support will be valuable to continue CTAT development and to continue to enhance the software to optimize it for single cell transcriptome analyses; an ITCR renewal would be appropriate for this purpose. One specific area of enhancement is that Trinity was designed for *de novo* transcriptomic reconstructions, but integrating existing reference genomes into Trinity where available will enhance the quality of CTAT analyses.

F. Final Thoughts

- It is critical for ITCR projects to remain nimble throughout the award period, given the rapid change over time in information technologies.
- An ITCR programmatic success has been to enable awards to interact with other NCI-funded efforts (e.g., FireCloud, collaborative ITCR projects through set-aside funding). Collaboration facilitates the realization of synergies so that tools can be world-class with limited additional expenditure, which is a very efficient approach.

- It may be valuable in the future of the network to identify one or more large-scale driving research projects that would provide incentives for groups to collaborate and connect their tools under the aegis of a group of ITCR investigators or the network Steering Committee, who would jointly manage the project.

5. Deep Phenotype Extraction (DeepPhe)

A. Background and Goals

Linking phenotypic data with molecular data has been a challenge for the cancer research community. To address this challenge, in 2014 NCI made award U24CA184407 to a team led by Guergana Savova at Children's Hospital in Boston and Rebecca Jacobson at the University of Pittsburgh to develop a deep phenotype extraction tool (DeepPhe). The ITCR-supported team is building methods and software to enhance the electronic medical record (EMR) documentation for cancer patients by extracting the cancer and the tumor for that patient and the attributes of the tumor and the cancer (e.g., temporality—historical vs. current and pathology, tumor type, biomarkers, location, laterality, quadrant position for breast cancer). The intent is to support a high throughput approach that processes and annotates data at multiple levels (from mention to phenotype) and across time. The tool is designed with two current use cases in mind:

- Translational research: Currently, translational researchers who extract EMR data to enrich their analyses rely on manual processing methods, which limits the number of patients who can be included in studies. Automated extraction will increase sample size and therefore study power.
- Supporting tumor registries: Tumor registries currently rely on manual processing of medical records to identify and tabulate cancer patients and their medical outcomes, which has led to significant backlogs. Automated extraction (with manual validation by physicians) has the potential to increase throughput without sacrificing accuracy. DeepPhe is also designed to be able to visualize results.

B. State of Development

The two ITCR PIs had collaborated previously and used the program as a chance to renew and continue collaborative research efforts. The University of Pittsburgh Medical Center also serves as a translational research use case site for ovarian cancer applications. In advance of the start of the project, the team forged collaborations with additional sites for individual use cases (e.g., Dana-Farber for melanoma applications, Vanderbilt for breast cancer applications, NCI's Surveillance, Epidemiology, and End Results (SEER) program, Kentucky and Louisiana state tumor registries).

DeepPhe is designed to be a comprehensive extraction system. Currently the tool is under active development and is capable of processing data from multiple EMR systems

with minimal degradation of performance across systems. A 2017 *Cancer Research* article describes the tool.[21] Code associated with DeepPhe is available from a GitHub repository (https://github.com/DeepPhe/DeepPhe-Release) and from a project page (https://healthnlp.hms.harvard.edu/deepphe/wiki/index.php/Main_Page). Licensing is based on Apache Source licenses (for the software) and Creative Commons licenses (for the content and the models).[22]

The team is planning to release version 2 in 2018, which will be applied to all three tumor types associated with the driving projects (breast, ovarian, and melanoma) and will incorporate episode classification, treatment regimens, and clinical genomics, while improving the visualization tool relative to version 1 to visualize both data for individual patients (e.g., treatment types over time) and cohorts (e.g., grouping patients by cancer stage or age). Version 2 will expose 28 distinct cancer and tumor attributes. Goals for future development include automated treatment regimen extraction and the expansion of extracted information to include clinical genomic observations, such as somatic variants (e.g., BRAF status) or gene rearrangements (e.g., ALK, NTRK).

The desired accuracy of the tool remains under discussion based on the particular use case. The team's goal is for the software to be at least as accurate as or somewhat more accurate than a human coder (i.e., 65–75 percent). Clinical/registry use cases may require higher degrees of accuracy (i.e., 98 percent)—especially if the intent is to use DeepPhe with limited or no human review. The award period of performance runs through 2019.

C. Interactions with Other Tools

Collaborations with other ITCR-funded teams: There are no current links to other ITCR-supported tools as the system is under development. The PIs have discussed potential future linkages with David Hanauer at the University of Michigan (EMERSE) and Andrey Fedorov at Brigham and Women's Hospital (QIICR/3-D Slicer).

Collaborations with other academic investigators: As mentioned above, the award was designed with academic and registry collaborations being integral to the award. Researchers at collaborating institutions are using DeepPhe to pursue their own translational research goals with respect to breast cancer, ovarian cancer, and melanoma. The DeepPhe team's involvement of researchers and potential users in the design, development, and testing of the tools promotes success.

---

[21] G. K. Savova et al., "DeepPhe: A Natural Language Processing System for Extracting Cancer Phenotypes from Clinical Records," *Cancer Research* 77, no. 21 (November 1, 2017): e115–18, doi: 10.1158/0008-5472.CAN-17-0615.

[22] For more detail, see "Licensing," Health NLP, https://healthnlp.hms.harvard.edu/deepphe/wiki/index.php/Licensing.

Collaborations with industry: No collaborations with industry reported as of September 2018.

D. Measuring Impact: Uses, Users, and User Communities

1. Tracking Users

There is no need to track users at this time as the tool is actively under development and its use limited to a small group of investigators, but team members expect usage to grow dramatically in the next phase.

2. Translational Research and Clinical Use

The tool is intended to have clinical use and impact. As mentioned above, the level of accuracy required for clinical use cases may serve as a limit on the speed with which the software can become useful for clinicians.

3. Publication and Citation as Measures of Use

As of September 2018, seven publications (including the *Cancer Research* article previously cited) acknowledge the ITCR award. Also as of September 2018, the source publication has not been cited in PubMed Central.

E. Future Development and Sustainability

The team reports that long-term sustainment is, of course, desirable; renewal of ITCR funding would represent a natural mechanism for continued development. Once the current award is complete, the team expects to move to a stage at which the user community would expand and links to additional ITCR-supported tools would occur.

F. Final Thoughts

- The Savova group is funded through PA-14-156 (R01GM114355), a National Institutes of Health (NIH)-wide solicitation for software development, hardening, and verification; that group also has an unsolicited R01 for methods development (R01LM010090). Team members report that collaborative relationships built through the ITCR network (monthly PI calls with presentations on scientific topics, annual meetings) are different from traditional R01-supported awards.

6. Galaxy-P

A. Background and Goals

Genomics tools are in common use. The Galaxy genomics analysis tool that allows users to combine and visualize data from multiple, independent queries using a graphics-driven interface that reduces dependence upon specialized data science and bioinformatics skills was first described in the literature in a 2005 publication by investigators at Pennsylvania State University.[23] Researchers intending to understand cancer also need sophisticated tools to conduct combined multi-omics analyses to link genes to protein expression data (e.g., derived from mass spectroscopy), which will provide a more complete picture at the molecular level. To address these challenges, in 2016 NCI made award U24CA199347 to a team led by Timothy Griffin at the University of Minnesota, intended to develop a proteomics extension to integrate into Galaxy to allow for multi-omics analyses. Specific award goals are to: (1) support metabolomics studies including development of a platform that takes high-throughput data, extracts thousands of features from metabolomics experiments, identifies them, and analyzes them statistically looking for metabolites of interest; and (2) use RNA-Seq data to build precision protein databases, match proteomics data to confirm expression of variant protein sequences, and correlate transcript and protein abundance data for assessing post-transcriptional regulatory mechanisms.

B. State of Development

The investigators are using the core Galaxy framework to develop Galaxy-P; the Galaxy-P team interacts regularly with the Galaxy community. Halfway through the U24 award, the tool is still in development. The Galaxy proteomics-genomics joint analysis tool was developed[24] and described in a 2015 publication;[25] it is available at http://galaxyp.org. The tool utilizes RNA-Seq data to build "precision" protein databases, matches mass spectroscopy-based proteomics data to these databases to confirm the expression of variant protein sequences, and correlates transcript and protein abundance for assessing post-transcriptional regulatory mechanisms. The software includes a Galaxy visualization plug-

---

[23] B. Giardine et al., "Galaxy: a Platform for Interactive Large-scale Genome Analysis," *Genome Research* 15, no. 10 (October 2005): 1451–5, doi: 10.1101/gr.4086505.

[24] NSF also provided funding, through Awards 1147079 and 1458524.

[25] J. Boekel et al., "Multi-omic Data Analysis using Galaxy," *Nature Biotechnology* 33, no. 2 (February 2015): 137–9, doi: 10.1038/nbt.3134.

in for "protein-centric" viewing of validated peptide sequence variants. The Galaxy-P project's GitHub repository is accessible at https://github.com/galaxyproteomics/.

An additional effort underway is the development of tools for microbiome analysis, given that microbial-derived signals may modulate hallmarks of cancer through diverse mechanisms. The team is working to use genomics, transcriptomics, and proteomics to find microbiota in samples and to understand their interactions with cancer; they are developing a user-friendly Galaxy workflow for mass spectrometry-based metabolite quantification and identification within the Galaxy framework. The team is leveraging and contributing to the Galaxy-based community of metabolomic informatics developers (Workflow4metabolomics, W4M). Galaxy-P is intended to be a user-driven resource that meets requirements for reproducibility, throughput, and making comparisons.

Along the way, the team reports that it has hit its benchmarks for developing tools. By the end of the funding period in 2020, the team expects to have a resource to help investigators analyze their data in the multi-omics realm. There will be a deliverable that helps with workflows for investigators to conduct high-level integrative multi-omic analysis of data. Components will be available as Docker containers for portable cloud computing applications.

In addition to the PI's research group, the Minnesota Supercomputing Institute provides expert personnel (e.g., Galaxy expertise, high-performance computing expertise) as well as computational infrastructure to support the Galaxy-P effort.

The team presents regularly at the American Society for Mass Spectrometry, the Association of Biomolecular Research Facilities, the Galaxy Community Conference, and other meetings to present Galaxy-P to investigators and to train them in its use.

C. Interactions with Other Tools

Collaborations with other ITCR-funded teams:

- *Globus Genomics*: The team has used the Globus Genomics team's services, mainly integrating the Galaxy workflow with the Globus genomics workflow.[26]
- *CRAVAT*: The team is undertaking current development work to link Galaxy-P to CRAVAT. The intent is that when a user runs a workflow in Galaxy, the tool will provide results to CRAVAT in order to better understand what is already known about a particular gene and how it encodes a protein. Specifically, users select variants with peptide-level confirmation in the CRAVAT Query Tool. A Galaxy plugin is used for automated display of CRAVAT visualizations in the

[26] See, for example, "Galaxy-P: Recent Developments and Emerging Applications," University of Minnesota, http://cbs.umn.edu/sites/cbs.umn.edu/files/public/downloads/ASMS_Griffin_poster_FINAL.pdf.

Galaxy interface. The approach leverages the full suite of CRAVAT tools for impact prediction of peptide variants (including both functional and pathogenic effects). A joint publication between the two teams has been submitted for journal review.

- *IGV*: The team is undertaking ongoing development work to link Galaxy-P to IGV. The intent is to use the new IGV JavaScript functionality to call IGV once sequences are identified to visualize where a particular protein maps to a genome.
- *NDEx*: The NDEx team is in discussions with the Galaxy-P team as to how to integrate the NDEx network-viewing functionality with Galaxy-P.
- *Trinity*: Galaxy-P currently uses reference genomes; the Galaxy-P team hopes to collaborate in the future to integrate with Trinity to be able to use that tool for *de novo* DNA and RNA assembly.

Collaborations with other academic investigators: The team is working directly in collaboration with 10–12 investigators who are trained and running Galaxy-P during its developmental stage. The Griffin group has an IMAT-ITCR collaboration supplement (EXPANDING MULTIPLEXED KINASE BIOSENSOR ANALYSIS TO SWATH-MS). The ITCR-IMAT supplement will be used by the Galaxy-P team to develop a new approach to analyzing proteomic data by using predefined segmented isolation windows so that low-abundant peptides can be reproducibly identified and quantified; the approach will be applied to a new assay developed by the IMAT PI (isotope-coded peptide biosensors for multiplexed profiling of kinase activity response) and used in a clinical context (assessing response in leukemia treatment).

Collaborations with industry: The team is discussing collaborations with industry users (e.g., Persistent Systems, Intero Life Sciences).

D. Measuring Impact: Uses, Users, and User Communities

1. Tracking Users

The tool is under active development and has a core group of 10–12 trained users who are collaborating with the PI's research group during the developmental phase. The team expects to have ~100 regular users in 3–5 years. Training/workshop attendance is one measure of dissemination to date. Over the last two years, the team has conducted six workshops and trained 20–50 people per workshop.

2. Translational Research and Clinical Use

The Galaxy-P tool is intended for use (in the near term) by discovery researchers. No clinical use is expected in the near future.

3. Publication and Citation as Measures of Use

The U24CA199347 award is acknowledged by five publications as of September 2018, all of which describe tool development and functionality associated with aspects of the Galaxy-P approach. The 2015 article in *Nature Biotechnology* previously mentioned has been cited 15 times in PubMed Central as of September 2018. The Galaxy-P team considers citations to be an insufficient measure of community size at this time.

E. Future Development and Sustainability

The Griffin team reports that they would benefit from an additional round of ITCR funding to continue development and enhance dissemination, as well as to harden validated workflows and tools and to further extend functionalities via collaboration with other ITCR groups.

F. Final Thoughts

- The convening efforts of ITCR have been valuable. Bringing together awardees facilitates sharing information and leads to improvements to the ITCR-supported tools. The ITCR-IMAT supplement also has built synergies between physical tool developers with a new approach to using mass spectrometry to collect proteomic data and software tool developers with a new approach to analyzing proteomic data.
- The ITCR team used NSF funding for core algorithm development, which was synergistic with ITCR funding.

7. Network Data Exchange (NDEx)

A. Background and Goals

Cancer is a disease of pathways and interactions (protein binding interactions, phosphorylation events). These interactions can be expressed as nodes and links among them—the definition of a biological network. Current representations are usually static visuals rather than computable and searchable objects. One challenge facing cancer researchers is that there is no standard software tool for representing, storing, and sharing biological network information (e.g., within publications). To address this challenge, in 2014 NCI made award U24CA184427 to a team led by Trey Ideker at the University of California to develop the Network Data Exchange (NDEx). The award was renewed in 2017 for an additional five years. NDEx is an open-source tool intended to be a "Dropbox for networks"—a single, common, repository. It is a sister project to the Cytoscape (http://cytoscape.org/) network visualization tool developed by an investigator team whose members overlap with the NDEx group.

B. State of Development

The first iteration of the ITCR award was used to develop the functional site. The initial version of NDEx was described in a 2015 journal article.[27] During the last year of that funding period, capstone core technology building occurred; the team rebuilt the core data system and replaced it with higher-performance approaches. A second version was released in December 2016 and described in a 2017 journal article.[28] The team reports that they have an operational system, and developers will continue adding features and optimizing—ongoing updates occur approximately quarterly (the NDEx internet home page, http://www.home.ndexbio.org/index/, lists three updates in 2017). The tool has search features plus filtering to find networks of interest. There is a drill-down interface whereby a network is visualized as a graph so that investigators can focus on particular edges and nodes; there is also a tabular representation of the network available. The tool also has an interface to grant permissions—to lock access, to make public or limit access, and to create shareable links to networks of interest.

---

[27] D. Pratt et al., "NDEx, the Network Data Exchange," *Cell Systems* 1, no. 4 (October 28, 2015): 302–5, doi: 10.1016/j.cels.2015.10.001.

[28] D. Pratt et al., "NDEx 2.0: A Clearinghouse for Research on Cancer Pathways," *Cancer Research* 77, no. 21 (November 1, 2017): e58–61, doi: 10.1158/0008-5472.CAN-17-0606.

The NDEx source code is available on GitHub at https://github.com/ndexbio. NDEx is licensed under the BSD 3-Clause "New" or "Revised" License (https://github.com/ndexbio/ndex-rest/blob/master/LICENSE), which limits promotion of the software by third parties without written consent. The software has client libraries in several languages (Python, R, Java, JavaScript), and has recently released an updated version of the Python client (collaborators produced an R client). The software includes a REST API for searching, sharing, uploading, and updating networks.

Team members report that an eventual goal is to make NDEx functionality available generally—any software that generates networks allows users to open those networks in NDEx. Currently, the team is focused on incorporating large numbers of networks into the system to ensure that potential users are interested in sharing their content and are optimizing collaboration mechanisms. Pursuant to that goal, NDEx has an approved stable repository at the University of California, San Diego (UCSD) so the software can make digital object identifiers (DOIs) available as unique identifiers for networks. NDEx is an approved official recommended repository for the Nature/Springer family of journals.

Another current activity is to develop a standard set of workshop materials to present NDEx to potential users as well as to programmers who might be interested in building NDEx into their applications.

C. Interactions with Other Tools

Collaborations with other ITCR-funded teams:

- *CRAVAT*: CRAVAT was an early adopter of NDEx. The latest version of CRAVAT takes network data as an input and maps pathway variants.
- *IGV, UCSC Xena*: Analyses conducted using these tools frequently connect sets of relevant genes, implicitly generating networks. Collaborations are underway to connect them to NDEx, which will allow those networks to be generated on the fly, create reusable objects, and thereby add value to users' analyses. If applications or researchers can express their data in a readily computable format, others are more likely to take advantage of their data and reuse them.
- *TCPA*: The investigator team reports that they are working with TCPA investigators to link the tools so that TCPA users can generate network-analyzable data and then open and visualize those networks in Cytoscape and store them in NDEx.

Collaborations with other academic investigators:

- *Cytoscape:* Cytoscape is an open source software platform for visualizing complex networks and integrating these with any type of attribute data. Linking Cytoscape with NDEx allows an NDEx user to identify proteins of interest and

have a set of relationships populate in Cytoscape; Cytoscape users can open their networks using the NDEx functionality.

- *Bioconductor:* Bioconductor is a library of many software packages that are written in the R language. The R language is commonly used in biological network analysis tools, especially for networks expressed as matrices. Team members report that collaborators in Germany developed a package that enables programs to communicate with NDEx with minimal application programming, which facilitates NDEx's use by an R programmer.

Collaborations with industry: No current industry collaborations have been identified. The NDEx team had initial funding from three companies (Roche, Janssen, and Pfizer) and acknowledges their contributions.

D. Measuring Impact: Uses, Users, and User Communities

1. Tracking Users

NDEx is designed to track its use base; counts/lists of users and stored networks are made available by the NDEx tool itself as part of the "Browse" feature.[29] As of May 2018, there were 995 NDEx users (including 45 user groups) and 3,267 stored networks on the NDEx site.

Team members report that early motivated users (e.g., Cytoscape users) are people who are already engaged in biological research that revolves around networks or that uses networks as an important component. As Cytoscape is well established, its users can see NDEx as adding value to augment their experience. A second group of users is those who are accessing NDEx through another tool (e.g., CRAVAT users who are using NDEx to analyze generated networks). The team sees as an eventual goal having NDEx embedded in enough interesting applications so there are users who are investigational biologists. Collecting data on the background of the users, however, will be difficult.

2. Translational Research and Clinical Use

Translational and clinical use of NDEx will depend upon the eventual inclusion of network biology in these applications. Future clinical users will be able to download their own copies of NDEx, maintain network data behind their local firewalls, and incorporate networks as part of their Health Insurance Portability and Accountability Act (HIPAA)-protected patient data.

---

[29] See NDEx, http://www.ndexbio.org/#/search?searchType=All&searchString=&searchTermExpansion=false.

3. Publication and Citation as Measures of Use

As of September 2018, 14 publications (including the previously cited *Cell Systems* article) acknowledge the ITCR award. As of September 2018, the *Cell Systems* article has been cited 29 times in PubMed Central. Team members report that the identification of publications incorporating NDEx-stored networks (e.g., networks incorporated as supplementary data in a publication or whose DOI numbers are acknowledged in a publication) will become an alternative mechanism for tracking impact. The team expects to disseminate a publication describing use of NDEx-stored networks in the future as more data are incorporated.

E. Future Development and Sustainability

The NDEx team has a long-term commitment from UCSD for archiving purposes, so all stored networks will remain discoverable even post-award and even in the absence of additional development funding.

F. Final Thoughts

- A goal of NDEx is for the software to bridge communities of researchers who are conducting different types of cancer research. Having public repositories of networks should facilitate collaboration formation as investigators discover others' work.
- The combination of tool development and collaboration formation goals makes the ITCR initiative unique.

8. Pathology Image Informatics Platform (PIIP)

A. Background and Goals

Computational pathology requires tools and resources for investigators to carry out quantitative analyses on whole-slide images for the digital pathology community to grow. Commercially available products are expensive, so there is a need for creating an open-source, freely available tool. Commercial products are also difficult to customize to users' needs, whereas open-source software enables individual investigators to adapt tools for their particular purposes. To address these challenges, in 2015 NCI made award U24CA199374 to a team led by Anant Madabhushi at Case Western Reserve University (CWRU) to develop the Pathology Image Informatics Platform (PIIP). PIIP is built atop the Sedeen virtual slide viewer platform. Sedeen was chosen because it was a good viewer with a large user community, yet one where it was easy to create applications and APIs that could be integrated into it and provided to the community. The award team is a multi-institutional consortium of computational pathologists and clinicians (other key personnel are Metin Gurcan at the Ohio State University, Anne Martel at Sunnybrook Research Institute affiliated with the University of Toronto, and Pathcore; collaborators at the University of Michigan, the University of Pennsylvania, and the University at Buffalo provide access to additional curated databases of imaging information and assist with validation) that builds on more than a decade of interactions and collaborations. The multi-disciplinary combination provides critical input to the design of the platform while helping to promulgate the technology within their respective communities.

B. State of Development

The PIIP platform is intended to support the visualization of web service interfaces (WSIs) from multiple vendors, annotation tools for pathologists, plug-in architecture to allow integration of algorithms, multimodality support, creation of an archive of richly annotated datasets, and the evaluation and validation of algorithms on benchmarked datasets. PIIP-developed tools (written in MATLAB and C++) are incorporated into the Sedeen viewer rather than being compiled as a separate software package or tool. PIIP was described in a 2017 article in *Cancer Research.*[30] PIIP supports Visual Studio 2015 and

[30] A. L. Martel et al., "An Image Analysis Resource for Cancer Research: PIIP-Pathology Image Informatics Platform for Visualization, Analysis, and Management," *Cancer Research* 77, no. 21 (November 1, 2017): e83–6, doi: 10.1158/0008-5472.CAN-17-0323.

2017 and a range of industry image formats (e.g., PerkinElmer qptiff, Olympus VSI). Other recent enhancements including adding x64bit architecture support, adding more input parameters (including FileDialog input), and improved support for MATLAB-based plugins. PIIP source code is available from GitHub at http://github.com/sedeen-piip-plugins/. The project has a dedicated internet site (http://pathiip.org) that describes the PIIP tools in the context of the Sedeen viewer and points potential users to where Sedeen may be downloaded. The team has incorporated human factors engineering into the design of the interface; they have identified cognitive challenges that inform the design of the interface with respect to its potential use cases and have improved interface usability by modifying it at useful leverage points.

Although PIIP tools are built (through multiple new releases per year) and being utilized, the project team is working on enhancements before the award period ends in 2020. The team reports that one area for future development is to enable the user community to add algorithms to the PIIP repository and have them be compiled in future versions of Sedeen.[31] Additional enhancements under development are improved support for MATLAB routines, mechanisms to call Python procedures from plugins, distribution of the Sedeen developer toolkit to a wider research community, creation of MacOS and Linux versions, support for web-based image tile servers, collection of datasets for validation, and integration of a deep learning framework into Sedeen.

Team members actively conduct outreach and are presenting on the tool often (i.e., once per month or more) at meetings such as the International Society for Optics and Photonics (SPIE) Medical Imaging Conference, the European Congress on Digital Pathology meeting, and the Machine Intelligence in Medical Imaging meeting. The top level of the PIIP internet site includes online tutorials and links to workshop materials.

C. Interactions with Other Tools

Collaborations with other ITCR-funded teams: The PIIP team has not engaged in any collaborations with other ITCR teams to date.

Collaborations with other academic investigators: In their survey response, team members identify forming collaborations with David Rimm (Yale University) and Richard Levenson (University of California, Davis) on account of their ITCR work. The purpose is to develop new computational approaches for the analysis and interrogation of whole slide images.

Collaborations with industry: While there have not been any collaborations (beyond Pathcore's role as the Sedeen developer) with industry during the ITCR award itself, the ITCR award has seeded collaborations between the CWRU team and General Electric

[31] "Software Downloads," PIIP, http://pathiip.org/?q=software-downloads.

(GE), Philips, Bristol Meyers Squibb, and Inspirata Inc. The collaborations with GE and Inspirata have been formalized in NCI academic-industry translational partnership awards (R01CA208236, R01CA202752, R01CA216579, and R01CA220581). In their survey response, the GE collaboration was described by PIIP team members as having been initiated through ITCR, as the project team met the GE collaborators at an ITCR PI meeting.

D. Measuring Impact: Uses, Users, and User Communities

1. Tracking Users

Downloads have increased from approximately 100 in 2014 to a projection of more than 500 in 2018. The team estimates that the number of users has increased from approximately 20 in 2016 to an expected 70 in 2018. Team members expect that the number of users will grow with the digital pathology community, which has been growing rapidly over the last 3–4 years and is expected to continue to do so.

Team members consider a growing number of academic and industrial collaborations to be a useful potential measure of the impact of the PIIP effort.

2. Translational Research and Clinical Use

The tool is intended to have clinical use and impact. The FDA's approval in 2017 of the Philips IntelliSite Pathology Solution as the first whole slide imaging system for digital pathology is expected to foster growth of the clinical user community.

3. Publication and Citation as Measures of Use

As of September 2018, 25 publications (including the previously cited *Cancer Research* article) acknowledge the ITCR award. As of September 2018, the *Cancer Research* article has been cited twice in PubMed Central. Team members report that tracking citations as a measure of use will be valuable.

E. Future Development and Sustainability

The team reports that ITCR has helped to seed more than $10 million in new awards made in FY 2017 and FY 2018 (e.g., the academic-industry partnership awards identified above) that will be used for further development and enhancement of digital pathology tools.

F. Final Thoughts

- ITCR is complemented by NCI efforts such as the academic-industry translational partnerships (e.g., PAR-18-530). ITCR develops general-purpose

algorithms and tools that can then be incorporated into specific partnerships for particular purposes (e.g., diagnostic solutions, patient prognosis tools).

- A final measure of success (though a difficult one to capture) for the program as a whole is that this ITCR project (and ITCR more broadly) has drawn more attention to digital pathology. NCI is funding more awards and there are higher impact articles related to digital and computational pathology being published in journals such as *Cell* and *Clinical Cancer Research*.

9. Quantitative Image Informatics for Cancer Research (QIICR)

A. Background and Goals

Imaging tools such as magnetic resonance imaging (MRI), positron emission tomography (PET), and computed tomography (CT) provide clinicians and researchers with information regarding patients' health status, but historically have not provided quantitative results that lead to reliable and reproducible analyses across patients, especially given the biases and variability introduced by the multiplicity of imaging instruments and measurement protocols. NCI supports a Quantitative Imaging Network (QIN) to validate quantitative imaging software, but this network is not designed to produce new methods, tools, and software platforms. To address this gap, in 2013 NCI made award U24CA180918 to a team led by Ron Kikinis and Andrey Fedorov at Brigham and Women's Hospital, to support the Quantitative Image Informatics for Cancer Research (QIICR) project. The overall goal of QIICR is to develop a set of open-source tools, including those built upon the 3-D Slicer platform for medical image informatics, image processing, and three-dimensional visualization, to develop quantitative approaches to interpreting images as it applies to cancer research. The team has three collaborative clinical projects deriving from the NCI-funded QIN and develops tools that fulfill those needs and makes those tools usable by other groups as well. A second goal is to ensure that analysis results produced by the tools are reusable, machine readable, and standardized (based on the Digital Imaging and Communications in Medicine (DICOM) standard) so that investigators can build upon them and compare different analysis techniques. A third goal is to conduct training and outreach activities.

B. State of Development

The clinical software development projects are at different stages of development. The first project is relevant to head and neck cancer. In that project (semiautomated segmentation in 18F-FDG PET scans, in collaboration with clinical investigators at the University of Iowa), the team has developed tools and complete workflows and has published multiple papers and a public dataset (TCIA QIN-HEADNECK). Participation in

multi-site evaluation studies has identified that these open source tools are at least as effective as proprietary tools developed in-house by other participants in the QIN.[32]

In a second project, relevant to prostate adenocarcinoma (in collaboration with clinical investigators at Brigham and Women's Hospital), the team has developed a collection of tools (extracting quantitative measures, analysis of diffusion) that are available through 3-D Slicer. They have published multiple papers evaluating the technology, including multi-site collaborative projects conducted by the QIN, and the team is in the final stages of making the dataset available (TCIA QIN-PROSTATE-Repeatability).[33]

In a third project, relevant to glioblastoma (in collaboration with clinical investigators at Massachusetts General Hospital), there are ongoing efforts to develop diffusion MRI and dual contrast MRI tools; the QIICR team is working with MGH collaborators to validate them. They have published a paper evaluating the technology in a multi-site setting as part of a collaborative project.[34]

In addition to these tools' incorporation into 3-D Slicer directly, the team maintains a GitHub repository at https://github.com/QIICR where the tools' source code is shared and from which individual tools may be downloaded; the tools employ a variety of license

---

[32] R. R. Beichel et al., "Semiautomated Segmentation of Head and Neck Cancers in 18F-FDG PET Scans: A Just-Enough-Interaction Approach," *Medical Physics* 43, no. 6 (June 2016): 2948–64, doi: 10.1118/1.4948679.

R. R. Beichel et al., "Multi-Site Quality and Variability Analysis of 3D FDG PET Segmentations Based on Phantom and Clinical Image Data," *Medical Physics* 44, no. 2 (February 2017): 479–96, doi: /10.1002/mp.12041.

A. Fedorov et al., "DICOM for Quantitative Imaging Biomarker Development: A Standards Based Approach to Sharing Clinical Data and Structured PET/CT Analysis Results in Head and Neck Cancer Research," *PeerJ* 4:e2057 (May 2016), doi: /10.7717/peerj.2057.

D. C. Newitt et al., "Multisite Concordance of Apparent Diffusion Coefficient Measurements across the NCI Quantitative Imaging Network," *Journal of Medical Imaging* 5, no. 1 (January-March 2018): 011003-1–011003-9, accessed October 10, 2017, doi: /10.1117/1.JMI.5.1.011003.

W. Huang et al., "Variations of Dynamic Contrast-Enhanced Magnetic Resonance Imaging in Evaluation of Breast Cancer Therapy Response: A Multicenter Data Analysis Challenge," *Translational Oncology* 7, no. 1 (February 2014): 153–66, doi: /10.1593/tlo.13838.

[33] F. Langkilde et al., "Evaluation of Fitting Models for Prostate Tissue Characterization using Extended-Range b-factor Diffusion-Weighted Imaging," *Magnetic Resonance in Medicine* 79, no. 4 (April 2018): 2346–58, doi: 10.1002/mrm.26831.

A. Fedorov et al., "Multiparametric Magnetic Resonance Imaging of the Prostate: Repeatability of Volume and Apparent Diffusion Coefficient Quantification," *Investigative Radiology* 52, no. 9 (September 2017): 538–46, doi: /10.1097/RLI.0000000000000382.

[34] K. M. Schmainda et al., "Multisite Concordance of DSC-MRI Analysis for Brain Tumors: Results of a National Cancer Institute Quantitative Imaging Network Collaborative Project," *American Journal of Neuroradiology* 39, no. 6 (June 2018): 1008–16, doi: /10.3174/ajnr.A5675.

formats (including the MIT license, the BSD-3-Clause, and the Apache 2.0 license). The tools also have Docker images for cloud computing use.

With respect to standards development, the team has extended the DICOM standard to support the use cases of the collaborating QIN projects. They have also established reference datasets demonstrating the appropriate use of the DICOM standard for quantitative imaging biomarker development and created reusable open source tools and libraries accompanied by scientific papers to support adoption of the DICOM standard in quantitative imaging research.[35] Finally, the team has incorporated the DICOM toolkit (DCMTK) and the dcqmi conversion tool into 3-D Slicer for converting imaging data in a variety of digital formats into the DICOM standard format. This extension is intended to facilitate the sharing and analysis of imaging data.

Outreach and training activities include: (1) Annually since 2015, an organized demonstration of DICOM and connectathon applied to quantitative imaging at the Radiological Society of North America (RSNA) conference. Participants include commercial vendors and open source and academic developers (see https://qiicr.gitbooks.io/dicom4qi/); (2) conducting a DICOM tutorial at the Medical Image Computing and Computer Assisted Intervention (MICCAI) Society conference in 2017 and 2018 (see http://qiicr.org/dicom4miccai/); (3) conducting demonstrations at DICOM4QI, RSNA, MICCAI, and other conferences; (4) having the project co-PI serve as the chair of the training and outreach working group of ITCR; and (5) organizing a workshop, "The Role of Academic Technology Development in Cancer Research" at the Spring 2016 Cancer Informatics for Cancer Center (CI4CC; http://qiicr.org/CI4CC-Spring2016-ASW/) meeting, co-chaired by Andrey Fedorov and Mary Goldman.

C. Interactions with Other Tools

Collaborations with other ITCR-funded teams:[36]

- *The Cancer Imaging Archive (TCIA)*: TCIA has adopted the DICOM standard. 3-D Slicer can be used to access and visualize TCIA-stored data.

---

35 C. Herz et al., "dcmqi: An Open Source Library for Standardized Communication of Quantitative Image Analysis Results Using DICOM," *Cancer Research* 77, no. 21 (November 1, 2017): e87–90, doi: /10.1158/0008-5472.CAN-17-0336.

36 The QIICR group, along with two other ITCR-funded teams, received an administrative supplement in 2015 to extend 3-D Slicer to support ground truth pathology tissue segmentation data and use FeatureDB to support the management of combined pathology and radiology features. STPI staff did not identify a continuing research collaboration or a linkage made between two ITCR-funded groups' tools stemming from this particular supplement.

- QIICR-developed tools are integrated with several other ITCR projects, as evident from the ITCR tool connectivity map.[37]

Collaborations with other academic investigators: In addition to the QIICR clinician-collaborators associated with the driving research projects, the team includes an investigator at the Queen's University in Canada. In their survey response, investigators identified a large number of new collaborations formed with academic investigators over the course of the award. One collaboration is with the cancer imaging research group at MAASTRO (Maasrtricht, Netherlands) to leverage ITCR work on the development of the DICOM standard and supporting tools with the development of radiomics ontologies and to enable interoperable structured communication for the results of radiomics studies. Another reported collaboration is with an investigator from the Dana-Farber Cancer Institute to adapt and expand the capability of the 3-D Slicer pharmacokinetic analysis tool to do perfusion analysis using dynamic contrast-enhanced MRI. A third reported collaboration is with Daniel Rubin and Sandy Napel, investigators at Stanford, to integrate the capabilities of the DICOM standard and conversion tools that have been developed through QIICR into their analysis tools. A fourth reported collaboration is with Marco Nolden at the German Cancer Research Center (DKFZ) in Heidelberg. The groups have established interoperability of 3-D Slicer and their MITK tool based on DICOM, have worked together on organizing a demonstration/connectathon at RSNA, and have co-organized a tutorial on the use of the standard and supporting tools at MICCAI. A final reported collaboration is with Sheila Reynolds from one of the Cancer Genomics Cloud Pilots at the Institute for Systems Biology. The scope of the collaboration was to apply QIICR-developed tools to enable integration and exploration of TCIA-stored imaging and image analysis data.

Collaborations with industry: The QIICR team itself includes collaborators from Kitware, PixelMed Publishing, Isomics, and OFFIS (in Germany). The DICOM standard and work to enhance it involves widespread industry participation.

D. Measuring Impact: Uses, Users, and User Communities

1. Tracking Users

The investigator team reports that they have two groups of intended users: individual researchers and research groups in the short term, and industry in the long term. The impact of tools is magnified and sustained when tools are incorporated by the vendor community. 3-D Slicer supports a large user community (100,000 downloads per year). The tool is

[37] See ITCR connectivity map at http://www.ndexbio.org/#/network/04c0a7e8-af92-11e7-94d3-0ac135e8bacf. QIICR tools are also integrated with a number of non-ITCR tools, including ePAD and MITK.

maintained by a private company (Kitware). Kitware describes its approach as developing open-source platforms that can be customized for commercial customers.[38] Most users access the QIICR tools through 3-D Slicer itself or though the Docker images. It is also possible to capture downloads of individual tools and pulls of the Docker images. Investigators report that tools that are broadly applicable tend to be the most highly downloaded. TCIA Browser is the most popular over the lifetime of the award (with nearly 10,000 downloads as of May 2018 and approximately 2,500 downloads in the last two years), while dcmqi has been downloaded approximately 2,000 times in the last two years. Investigators report that while downloads can be captured, there is no mechanism for attempting to track how the tools are used.

2. Translational Research and Clinical Use

The QIICR tools, deriving from the QIN's goals, are developed for the purpose of measurement or prediction of tumor response to therapies in clinical trial settings, with the overall goal of facilitating clinical decision making.[39] The expectation is that as they are completed and fully validated they will be used by clinicians.

3. Publication and Citation as Measures of Use

Unlike other ITCR projects, because this award is developing individual tools, QIICR as a whole does not necessarily have a single, founding publication. As of September 2018, 25 publications acknowledge funding from the ITCR award, including both publications that introduce individual tools and research publications describing the validation and research use of QIICR-supported tools.

E. Future Development and Sustainability

As described above, QIICR tools are incorporated into 3-D Slicer, which is maintained outside of the ITCR award. This facilitates long-term sustainability. DCMTK is maintained by a separate entity outside of the ITCR award. DICOM standard adoption and implementation is independent from the ITCR award and is expected to have lasting impact due to broader applicability beyond the ITCR award.

F. Final Thoughts

- There is value in the QIN approach of comparing performance across multiple tools using common data sets.

---

[38] Kitware, Inc. "About Kitware." Accessed September 1, 2018. https://www.kitware.com/about/

[39] National Cancer Institute. "Quantitative Imaging Network." Accessed September 1, 2018. https://imaging.cancer.gov/programs_resources/specialized_initiatives/qin/about/default.htm

10. Xena

A. Background and Goals

David Haussler's group at the University of California, Santa Cruz (UCSC) has been involved with providing bioinformatics support first to the Human Genome Project (HGP) and then to The Cancer Genome Atlas (TCGA); he has been a principal investigator on HGP and TCGA-related awards since 2001. The UCSC Genome Browser was first released in 2003,[40] and in a 2008 R21 award, the Haussler group proposed to expand the Genome Browser's functionality to "integrate, visualize and analyze genomic and clinical data generated by the TCGA project."[41] While the Genome Browser facilitates analysis of TCGA data, there are challenges accessing combining genomic datasets across laboratories and research projects. To address this gap, in 2013 NCI made award U24CA180951 to the Haussler group to create the Xena platform. Xena is designed as a mechanism for comparing and analyzing tens of thousands of patients' genomic data across many datasets.

B. State of Development

Team members report that it took approximately three years to build the browser and hub software before the team made Xena publicly available online in 2016. Xena is both a web browser interface and a federated data hub platform that includes support for seven public-facing data hubs (the NCI Genomic Data Commons, TCGA, the PanCan Atlas that is the final repository of TCGA information, International Cancer Genome Consortium, PanCancer Analysis of Whole Genomes, Global Alliance for Genomics and Health, and Treehouse Childhood Cancer Initiative) in the cloud. Researchers can explore data on any of these hubs using the Xena Browser. Using the Xena Browser, anyone can access data in public-facing hubs, while only people with controlled access can access data on private hubs. Each individual hub owner decides who has access, making Xena a decentralized data platform that can both enable data sharing and protect data access based upon users' preferences and data access policies. A Python API allows users to pull just a slice of the data should they so choose, while users can download whole datasets from individual URLs. Xena is designed to scale with genomics resources—all that is required is for UCSC to add more data hubs. The tool can visualize data from tens of thousands of samples in

[40] D. Karolchik et al., "The UCSC Genome Browser Database," *Nucleic Acids Research* 31, no. 1 (January 1, 2003): 51–4, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC165576/.

[41] From the Abstract of R21CA135937.

seconds, and has been used for analyses of up to 1 million cells' worth of single-cell RNA-Seq data.

The tool allows researchers to visualize many types of data—single-nucleotide polymorphisms, structural variants, copy number, methylation data, protein expression data, and survival data. Data can be represented as a visual spreadsheet—rows are samples, columns are genomics data—and users can build spreadsheets column by column, adding data types as desired. Users can find, filter, and group samples to create subgroups for further analysis; if datasets incorporate survival data, users can run Kaplan-Meier analyses to compare subgroups statistically and visually.

The UCSC Xena project has a GitHub repository, accessible at https://github.com/ucscXena. The source code is licensed under the Apache 2.0 license.[42]

C. Interactions with Other Tools

Collaborations with other ITCR-funded teams:

- *MuPIT*: MuPIT visualizes mutations in protein three-dimensional structures; the collaboration is working to develop a workflow so that MuPIT can accept Xena data to map mutations using MuPIT's method for visualizing protein structure.
- *TIES*: TIES was developed to process and annotate biospecimen and imaging data to facilitate sharing of de-identified samples for research purposes while protecting patient privacy and confidentiality. The collaboration is intended to incorporate sample information stored in TIES into the back end of Xena so that Xena users can see pathology/sample data associated with data returned from Xena searches. The work has been completed and a prototype is built and available online.

Collaborations with other academic investigators:

- *TCGA*: The Xena team has continued to work to incorporate TCGA's genomic and clinical data into Xena, which can add visualization and analysis functionalities to these data (e.g., survival analysis, pathway visualization). Xena contains all the latest TCGA open-access genomics data, and the team continues to update the resource as new data are published.
- *UCSC TumorMap* (funded through a TCGA Genome Data Analysis Center). TumorMap visualizes samples based on their molecular similarities. A web-based click-through has been developed to allow data viewed on Xena to be transferred and displayed in TumorMap. Additional workflows are being developed to allow for comparisons between the TumorMap and Xena

[42] Github. "Apache License 2.0." https://github.com/ucscXena/ucsc-xena-client/blob/master/LICENSE.

visualizations of gene expression (e.g., in order to divide samples into sub-groups or clusters) upon which Xena can conduct additional analyses (e.g., comparative survival analyses).

- *Genomic Data Commons (GDC)*: The UCSC team is part of the group helping NIH to analyze and visualize data in the GDC and are exploring the possibility of incorporating Xena (or similar functionality) into the future Commons.

Collaborations with industry: No collaborations have been reported as of September 2018.

D. Measuring Impact: Uses, Users, and User Communities

1. Tracking Users

Currently there are 7,000–8,000 users per month (from approximately 6,000 institutions), and the growth rate has been 250 percent from 2016 to 2017, then 70 percent from 2017 to 2018 (last 12 months).

Website hits and data downloads are also tracked. There are approximately 1 million hits on the hub per month. In April 2018 for example, approximately 600 GB of data were downloaded—and that does not take into account uses by experimental biologists who use the browser to visualize data without downloading.

2. Translational Research and Clinical Use

The UCSC Xena team reported that there are already known uses of the tool related to analyzing data derived from clinical trials. They are learning lessons from each translational application for future efforts.

- Treehouse Childhood Cancer Initiative: The Treehouse initiative has deployed two Xena hubs of their own, one for public sharing and one for private sharing within the Treehouse group. They have incorporated de-identified data from four clinical trials into their own Xena dataset (British Columbia Children's Hospital, Orange County Children's Hospital, Pacific Pathway Neuron consortium, and the Children's Hospital at Stanford).
- I-SPY2 clinical trial: As part of the trial, clinicians want to analyze gene expression data correlated with clinical information such as drug treatment and response for the full set (and subsets) of patients. The UCSC team is working with clinicians to set up their own local hub where they can run analyses using Xena while keeping patient data local and confidential.
- ITOMIC trial: ITOMIC is a pilot trial collecting and analyzing multi-omics data from breast cancer patients. Investigators are using Xena's data and tools to compare trial-collected data with TCGA data in the ITOMIC Xena hub.

3. Publication and Citation as Measures of Use

As of September 2018, 22 publications acknowledge the ITCR award, including five publications in *Nature* and two in the *New England Journal of Medicine.* Articles either describe annual updates to the Genome Browser (including references to the Xena tool) or research that makes use of genomic data. As of early 2018, there was not a designated article investigators could use to cite use of the Xena tool.[43] At this time, therefore, citation is not considered to be a meaningful indicator of research use.

E. Future Development and Sustainability

Team members indicated that future support through ITCR would be desirable. They have other federal funding and philanthropic support as well. They indicate that it is important that the data remain free and unencumbered.

F. Final Thoughts

- The fact that ITCR concentrates on engineering the actual mechanisms of data analysis rather than focusing on a specific project (e.g., breast cancer) is valuable. It is important to have a cross-cutting program that creates integrated infrastructure that spans NCI.

[43] See "UCSC Xena and Cancer Genomics Browser," Google Groups, https://groups.google.com/forum/#!topic/ucsc-cancer-genomics-browser/Dv3BL_gfRCE.

11. R21: AMARETTO Regulatory Networks Analysis

A. Background and Goals

One challenge facing cancer researchers lies in integrating and interpreting multi-omics data collected through large studies, such as TCGA, to identify novel mechanisms of carcinogenesis and uncover new therapeutic targets. Researchers have been developing algorithms to mine these large-scale, high-dimensional data sets in search of new insights; in 2013 AMARETTO was released by investigators at Stanford to aid in the discovery process.[44] The tool has been used to infer regulatory networks and sub-networks associated with cancer, but historically had not been applied to understand the role of viruses in carcinogenesis, in cancers such as hepatocellular carcinoma. To address this challenge, in 2017 NCI made award R21CA209940 to Nathalie Pochet, an investigator at Brigham and Women's Hospital, to extend AMARETTO for this purpose.

The R21 award had two goals. First, to develop and disseminate an algorithm and software tool for inferring regulatory networks via multi-omics (genomics, epigenetics, proteomics) data fusion and learning communities across networks with applications in cancer. Second, to use the package to learn the regulatory networks underlying cancers, with an initial focus on virus-induced cancers specifically, to study the role of hepatitis C and B virus infections progressing to hepatocellular carcinoma.

B. State of Development

The AMARETTO software package is implemented in R and the source code is available at: https://bitbucket.org/gevaertlab/pancanceramaretto. The AMARETTO and Community-AMARETTO tools were developed within a collaboration between Pochet and Olivier Gevaert (Stanford University). They are being used in research as laid out in the original goals of the application. New research applications are also expanding beyond the original research plan along multiple dimensions. Through a set-aside collaboration with the Carey ITCR group, the team is conducting research related to identifying cancer drivers and regulatory networks from multi-omics and imaging associated with glioblastoma. Second, the team is exploring translational applications. Identifying cancer

[44] Gevaert O, Villalobos V, Sikic BI, Plevritis SK. Identification of ovarian cancer driver genes by using module network integration of multi-omics data. *Interface Focus*. 2013 Aug 6;3(4):20130013. doi: 10.1098/rsfs.2013.0013. Erratum in: *Interface Focus.* 2014 Jun 6;4(3):20140023. PubMed PMID: 24511378; PubMed Central PMCID: PMC3915833.

drivers and potential drug targets simultaneously can be useful for drug discovery purposes; AMARETTO is making predictions that are being validated with empirical data. Third, the team is experimenting with using these techniques to identify networks associated with neurological and immune-mediated diseases (e.g., multiple sclerosis).

C. Interactions with Other Tools

Collaborations with other ITCR-funded teams:

- *Bioconductor*: In 2018, ITCR funded a set-aside collaboration with U01CA214846 (PI: Vincent Carey). The set-aside project is being used to incorporate the AMARETTO software tools and downstream functionalities into Bioconductor. Carey and Pochet entered ITCR in the same cohort of awards and met for the first time at the 2017 ITCR PI meeting. Carey suggested the groups collaborate and developed the set-aside collaboration project plan. The two groups are located at Boston-area universities, thus facilitating collaboration.
- *GenePattern:* The team is collaborating with the Mesirov group to incorporate the AMARETTO software tools and downstream functionalities into GenePattern and GenePattern Notebook.

The Pochet group intends to write a joint paper with the Carey and Mesirov groups describing the incorporation of AMARETTO into these two tools.

Collaborations with other academic investigators: The award has allowed deepening of pre-existing collaborations with Olivier Gevaert at Stanford University and Thomas Baumert at Inserm, Strasbourg University. Dr. Pochet reported in her survey response that the purpose of these collaborations is to link her research to other algorithms and tools not funded by ITCR that have similar goals and to apply her research to study human disease beyond cancer, such as infectious, neurological and immune-mediated diseases.

Collaborations with industry: There are potential industry-related, drug discovery applications of the Pochet team's work with AMARETTO. These potential applications may lead to future collaborations.

D. Measuring Impact: Uses, Users, and User Communities

1. Tracking Users

At this time, AMARETTO is a research tool rather than a publicly-disseminated software package. The Pochet team is giving demonstrations to collaborators and planning to demonstrate the tool to the ITCR community later in 2018. They have created private accounts for collaborators to use (including for debugging purposes) while AMARETTO is being incorporated into Bioconductor and GenePattern Notebook. Both GenePattern

Notebook and Bioconductor have a Faculty of 1000 channel, and plans are underway to publish application notes there to describe AMARETTO's integration.

2. Translational Research and Clinical Use

The team is exploring AMARETTO's potential use in translational research and in drug discovery applications.

3. Publication and Citation as Measures of Use

As of September 2018, three publications acknowledge the ITCR award.

E. Future Development and Sustainability

The team is planning to submit a U01 award for early-stage software development of AMARETTO as well as to develop further the applications of the tool described above. The funding will be used as well to continue to identify and develop new applications for AMARETTO. They have also submitted a NCI U01 systems biology grant.

F. Final Thoughts

- ITCR has provided access to a network of collaborators and leaders in the field that is incredibly exciting for an early-career investigator. The focused solicitation also offers greater opportunities to early-career investigators through focused guidance and support by the ITCR program than might be available in the unsolicited pool.

12. R21: BayesGO and GAIL Cancer Subtypes Analysis

A. Background and Goals

One challenge facing cancer researchers is to identify distinct and clinically relevant cancer subtypes in order to develop hypotheses for potential therapeutic targets or to stratify cancer patients who might be helped by already-existing therapeutic approaches. To address this challenge, NCI funded a team led by Dongjun Chung at the Medical University of South Carolina (award R21CA209848) to develop new statistical methods to identify cancer subtypes and to apply those methods in cancer research. The award has three aims. The first is to develop new statistical methods to improve pathway information by integrating publicly available pathway databases with the pathway information derived from the literature. The second is to develop new statistical methods to incorporate systems biology and cancer pathway data into cancer subtype identification approaches based on multiple types of genomic data. The final aim is to apply these techniques to ovarian cancer and other related cancer types.

B. State of Development

The team is funded through 2019 and therefore is entering the last year of their award. To date, in response to the first specific aim the Chung group has developed bayesGO as a publicly available R package and has published a paper describing it, and has developed GAIL as a web-based user interface to improve access to the biomedical literature mining data used by bayesGO. The group is working on a paper describing the GAIL interface. With respect to the second aim, the team has developed InGRiD as a publicly available R package and has published a paper describing it. The tools do not yet have embedded trackers to collect information on uses and users, but the team is currently working on adding them. Their goal for the final year of the award is to work toward the application of the techniques that they have developed to ovarian cancer and other cancer types, which corresponds to the final specific aim of the award.

The team has developed training materials as part of their efforts to disseminate their tools. They have developed YouTube videos to introduce InGRiD that will be enhanced in the future, while the InGRiD R package provides a vignette describing the analysis workflow with screenshots. The GAIL internet page includes a tutorial and user guide with screenshots. All of the packages have detailed tutorials to help users to interpret their output. In addition to their online documentation, the Chung group has published a book

chapter describing the analysis workflow using GAIL and bayesGO with sample results and advice for how to interpret outputs.

C. Interactions with Other Tools

The tools that the Chung group develops make use of GWAS and other systems biology and pathway-based genomic data. To the extent to which ITCR awards make GWAS data more readily available or more interpretable, the Chung team makes use of those data as part of the analyses they undertake.

Collaborations with other ITCR-funded teams:

- *CLIP-Seq*: The Chung group has expertise with high-throughput sequencing data analysis while the Xiao group (funded through ITCR award U01CA204695) has expertise with CLIP-seq and miRNA experiments and analysis. The teams are working together on multi-read analysis of CLIP-seq data in order to enhance the signal from the data. The teams have jointly developed algorithms and are working toward developing software that they hope to formalize and disseminate through a publication next year.

Collaborations with other academic investigators: The Chung group developed a collaboration with the Neelon group at MUSC for the development of InGRiD.

Collaborations with industry: No industry collaborations were reported.

D. Measuring Impact: Uses, Users, and User Communities

1. Tracking Users

The team considers downloads and user tracking to be a useful measure of impact. Once trackers are developed and implemented for the software they have developed, they expect to collect information on the number of users and how the tools are being used. The team further expects that having the GAIL web interface will facilitate use and increase the number of users.

2. Translational Research and Clinical Use

No current uses have been reported.

3. Publication and Citation as Measures of Use

As of September 2018, six publications acknowledge the ITCR award. Team members consider publications (impact factor of journals, citations) to be a reasonable measure of impact, while noting that there is a time lag in using citations as an indicator.

E. Future Development and Sustainability

The team is working on follow-on grant applications to NCI making use of their ITCR-developed tools for secondary analysis of genomic data. There is potential for the team to submit eventual follow-on applications (through ITCR or as unsolicited research) to develop their algorithms further.

F. Final Thoughts

- ITCR is considered to be an excellent initiative and investigators' participation to date has been enjoyable.
- One value of ITCR is that it is an interdisciplinary initiative that brings together cancer researchers and computer scientists—although the initiative could make more of an effort to attract statisticians.
- The ITCR annual meeting was very helpful, especially in forming collaborations—the collaboration with Dr. Xiao's group arose during discussions at the PI meeting.
- There may be value in adding a R01 component for larger-scale/longer-term funding for algorithm development.

Appendix A.
References

Beichel, R. R., B. J. Smith, C. Bauer, E. J. Ulrich, P. Ahmadvand, M. M. Budzevich, R. J. Gillies et al. "Multi-Site Quality and Variability Analysis of 3D FDG PET Segmentations Based on Phantom and Clinical Image Data." *Medical Physics* 44, no. 2 (February 2017): 479–96. doi: /10.1002/mp.12041.

Beichel, R. R., M. Van Tol, E. J. Ulrich, C. Bauer, T. Chang, K. A. Plichta, B. J. Smith et al. "Semiautomated Segmentation of Head and Neck Cancers in 18F-FDG PET Scans: A Just-Enough-Interaction Approach." *Medical Physics* 43, no. 6 (June 2016): 2948–64. doi: 10.1118/1.4948679.

Bello, S. M., M. Shimoyama, E. Mitraka, S. J. F. Laulederkind, C. L. Smith, J. T. Eppig, and L. M. Schriml. "Disease Ontology: Improving and Unifying Disease Annotations across Species." *Disease Models & Mechanisms* 11, no. 3 (March 12, 2018): pii: dmm032839. doi: 10.1242/dmm.032839.

Boekel, J., J. M. Chilton, I. R. Cooke, P. L. Horvatovich, P. D. Jagtap, L. Käll, J. Lehtiö et al. "Multi-omic Data Analysis using Galaxy." *Nature Biotechnology* 33, no. 2 (February 2015): 137–9. doi: 10.1038/nbt.3134.

Douville, C., H. Carter, R. Kim, N. Niknafs, M. Diekhans, P. D. Stenson, D. N. Cooper et al. "CRAVAT: Cancer-Related Analysis of Variants Toolkit." *Bioinformatics* 29, no. 5 (March 1, 2013): 647–8. doi: 10.1093/bioinformatics/btt017.

Fedorov, A., D. Clunie, E. Ulrich, C. Bauer, A. Wahle, B. Brown, M. Onken et al. "DICOM for Quantitative Imaging Biomarker Development: A Standards Based Approach to Sharing Clinical Data and Structured PET/CT Analysis Results in Head and Neck Cancer Research." *PeerJ* 4:e2057 (May 2016). doi: /10.7717/peerj.2057.

Fedorov, A., M. G. Vangel, C. M. Tempany, and F. M. Fennessy. "Multiparametric Magnetic Resonance Imaging of the Prostate: Repeatability of Volume and Apparent Diffusion Coefficient Quantification." *Investigative Radiology* 52, no. 9 (September 2017): 538–46. doi: /10.1097/RLI.0000000000000382.

Giardine, B., C. Riemer, R. C. Hardison, R. Burhans, L. Elnitski, P. Shah, Y. Zhang et al. "Galaxy: a Platform for Interactive Large-Scale Genome Analysis." *Genome Research* 15, no. 10 (October 2005): 1451–5. doi: 10.1101/gr.4086505.

Grabherr, M. G., B. J. Haas, M. Yassour, J. Z. Levin, D. A. Thompson, I. Amit, X. Adiconis et al. "Full-Length Transcriptome Assembly from RNA-Seq Data without a Reference Genome." *Nature Biotechnology* 29, no. 7 (May 15, 2011): 644–52. doi: 10.1038/nbt.1883.

Griffith, M., N. C. Spies, K. Krysiak, J. F. McMichael, A. C. Coffman, A. M. Danos, B. J. Ainscough et al. “CIViC is a Community Knowledgebase for Expert Crowdsourcing the Clinical Interpretation of Variants in Cancer.” *Nature Genetics* 49, no. 2 (January 31, 2017): 170–74. doi: 10.1038/ng.3774.

Herz, C., J. C. Fillion-Robin, M. Onken, J. Riesmeier, A. Lasso, C. Pinter, G. Fichtinger et al. “dcmqi: An Open Source Library for Standardized Communication of Quantitative Image Analysis Results Using DICOM.” *Cancer Research* 77, no. 21 (November 1, 2017): e87–90. doi: /10.1158/0008-5472.CAN-17-0336.

Huang, W., Xin Li, Y. Chen, Xia Li, M-C. Chang, M. Oborski, D. I. Malyarenko et al. “Variations of Dynamic Contrast-Enhanced Magnetic Resonance Imaging in Evaluation of Breast Cancer Therapy Response: A Multicenter Data Analysis Challenge.” *Translational Oncology* 7, no. 1 (February 2014): 153–66. doi: /10.1593/tlo.13838.

Karolchik, D., R. Baertsch, M. Diekhans, T. S. Furey, A. Hinrichs, Y. T. Lu, K. M. Roskin et al. “The UCSC Genome Browser Database.” *Nucleic Acids Research* 31, no. 1 (January 1, 2003): 51–4. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC165576/.

Langkilde, F., T. Kobus, A. Fedorov, R. Dunne, C. Tempany, R. V. Mulkern, and S. E. Maier. “Evaluation of Fitting Models for Prostate Tissue Characterization using Extended-Range b-factor Diffusion-Weighted Imaging.” *Magnetic Resonance in Medicine* 79, no. 4 (April 2018): 2346–58. doi: 10.1002/mrm.26831.

Madhavan, S., D. Ritter, C. Micheel, S. Rao, A. Roy, D. Sonkin, M. McCoy et al. “ClinGen Cancer Somatic Working Group – Standardizing and Democratizing Access to Cancer Molecular Diagnostic Data to Drive Translational Research.” *Pacific Symposium on Biocomputing* 23 (2018): 247–58. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5728662/.

Martel, A. L., D. Hosseinzadeh, C. Senaras, Y. Zhou, A. Yazdanpanah, R. Shojaii, E. S. Patterson et al. “An Image Analysis Resource for Cancer Research: PIIP-Pathology Image Informatics Platform for Visualization, Analysis, and Management.” *Cancer Research* 77, no. 21 (November 1, 2017): e83–6. doi: 10.1158/0008-5472.CAN-17-0323.

Newitt, D. C., D. Malyarenko, T. L. Chenevert, C. C. Quarles, L. C. Bell, A. Fedorov, F. M. Fennessy et al. “Multisite Concordance of Apparent Diffusion Coefficient Measurements across the NCI Quantitative Imaging Network.” *Journal of Medical Imaging* 5, no. 1 (January–March 2018): 011003-1–011003-9. Accessed October 10, 2017. doi: /10.1117/1.JMI.5.1.011003.

Niknafs, N., D. Kim, R. Kim, M. Diekhans, M. Ryan, P. D. Stenson, D. N. Cooper, and R. Karchin. “MuPIT Interactive: Webserver for Mapping Variant Positions to Annotated, Interactive 3D Structures.” *Human Genetics* 132, no. 11 (November 2013): 1235–43. doi: 10.1007/s00439-013-1325-0.

Patel, A. P., I. Tirosh, J. J. Trombetta, A. K. Shalek, S. M. Gillespie, H. Wakimoto, D. P. Cahill et al. “Single-Cell RNA-Seq Highlights Intratumoral Heterogeneity in

Primary Glioblastoma." *Science* 344, no. 6190 (June 20, 2014): 1396–401. doi: 10.1126/science.1254257.

Pratt, D., J. Chen, R. Pillich, V. Rynkov, A. Gary, B. Demchak, and T. Ideker. "NDEx 2.0: A Clearinghouse for Research on Cancer Pathways." *Cancer Research* 77, no. 21 (November 1, 2017): e58–61. doi: 10.1158/0008-5472.CAN-17-0606.

Pratt, D., J. Chen, D. Welker, R. Rivas, R. Pillich, V. Rynkov, K. Ono et al. "NDEx, the Network Data Exchange." *Cell Systems* 1, no. 4 (October 28, 2015): 302–5. doi: 10.1016/j.cels.2015.10.001.

Savova, G. K., E. Tseytlin, S. Finan, M. Castine, T. Miller, O. Medvedeva, D. Harris et al. "DeepPhe: A Natural Language Processing System for Extracting Cancer Phenotypes from Clinical Records." *Cancer Research* 77, no. 21 (November 1, 2017): e115–18. doi: 10.1158/0008-5472.CAN-17-0615.

Schmainda, K. M., M. A. Prah, S. D. Rand, Y. Liu, B. Logan, M. Muzi, S. D. Rane et al. "Multisite Concordance of DSC-MRI Analysis for Brain Tumors: Results of a National Cancer Institute Quantitative Imaging Network Collaborative Project." *American Journal of Neuroradiology* 39, no. 6 (June 2018): 1008–16. doi: /10.3174/ajnr.A5675.

Wang, L., A. K. Shalek, M. Lawrence, R. Ding, J. T. Gaublomme, N. Pochet, P. Stojanov et al. "Somatic Mutation as a Mechanism of Wnt/β-catenin Pathway Activation in CLL." *Blood* 124, no. 7 (August 14, 2014): 1089–98. doi: 10.1182/blood-2014-01-552067.

Wu, C., X. Jin, G. Tsueng, C. Afrasiabi, and A. I. Su. "BioGPS: Building Your Own Mash-Up of Gene Annotations and Expression Profiles." *Nucleic Acids Research* 44, no. D1 (January 4, 2016): D313–6. doi: 10.1093/nar/gkv1104.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
XX-10-2018	Final	Aug 2018 - Oct 2018

4. TITLE AND SUBTITLE

Informatics Technologies for Cancer Research (ITCR) Case Study Report

5a. CONTRACT NUMBER

NSFOIA-0408601

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Zuckerman, Brian L.
Simon, Ian D.
Kowal, Katherine M.
Doane, William E. J.

5d. PROJECT NUMBER

NC-20-4482

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

IDA Science and Technology Policy Institute
1899 Pennsylvania Avenue, NW, Suite 520
Washington, DC 20006-3602

8. PERFORMING ORGANIZATION REPORT NUMBER

IDA Document D-10311

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

National Cancer Institute
9609 Medical Center Drive
Rockville, MD 20850

10. SPONSOR/MONITOR'S ACRONYM(S)

NCI

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited (29 January 2019).

13. SUPPLEMENTARY NOTES

14. ABSTRACT

The Informatics Technology for Cancer Research (ITCR) Program is a trans-National Cancer Institute (NCI) grant program supporting investigator-initiated informatics technology development driven by critical needs in cancer research. In February 2018, NCI asked the IDA Science and Technology Policy Institute (STPI) to conduct three activities: (1) facilitate an expert panel process intended to provide evaluative insights regarding the program's rationale and impact to date; (2) survey current and former ITCR awardees to identify the program's role in fostering collaboration; and (3) conduct a set of case studies of ITCR impact at an award level. This document responds to the third and final charge and presents the results of 11 case studies of ITCR-supported technologies.

15. SUBJECT TERMS

Case Study, Evaluation, genomics, information technology, technology transfer

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Klemm, Juli D.
Unclassified	Unclassified	Unclassified	Same as Report	63	19b. TELEPHONE NUMBER *(Include area code)* 301-480-5778

Reset

Standard Form 298 (Rev. 8/98)
Prescribed by ANSI Std. Z39.18