ITCR PI Meeting – May 27-28, 2015
UC San Diego
Day 1

Panel 1: U24 Genomics Project Perspective

Talk 1: Brian Haas, Trinity Project
- Trinity Galaxy portal now available: https://galaxy.ncgas-trinity.indiana.edu
- Tools for mining fusions and mutations are to be released soon.

Talk 2: Martin Morgan, Integrative & scalable solution in *R*/*Bioconductor*
CCGRR: Curated Cancer Genomics Resources in R
1. AnnotationHub - "wrangles" data and gets it in a format that's ready for R analysis
2. BiocParallel - consistent interface to parallel computation on cores, computers, clusters, clouds
3. GenomicFiles - for managing on-disk collections of genomic files
4. AMI & docker images

User needs:
1. Priorities: statistical insights, integrative comprehension including visualization, data management
2. Leading edge analyses
3. Fast exploratory analyses that transition to production environment

Talk 3: Trey Ideker, Dexter Pratt, NDEx
- NDex: Aimed at addressing the need to aggregate and use broadly-generated network data
- Currently project focus is on content acquisition and network building
- Moving toward a model of "NDex-enabled software for cancer biologists"

Talk 4: David Haussler, UCSC Xena
- Genomics is the future to cancer treatment. To be successful, we need to aggregate data for statistical power
- Get involved in GA4GH. Massive open source project, repos on GitHub. Make sure it meets your needs.
    - Driving Projects: Genomics Matchmaker, Beacon, BRCA Challenge, Treehouse Childhood Cancer project (future), trial of oMics
- http://xena.ucsc.edu
- GalaxyXena tools for integrating with Galaxy
- Collaborations with ClinVar, LOVD
- Driving project: Supporting the BRCA Challenge - carefully annotate the inherited BRCA1/2 mutations and determine which are pathogenic

Panel Discussion

Q: What did you do right to get your project to where it is, supporting the needs of a significant user community?
- Martin (Bioconductor) Driving principles about quality, inspired leadership

- Aviv (Trinity) Addressed an unmet, burning need; principle of quality (does a certain job well); articulating what the tool does - this is what you can get from using the tool; provide support & be responsive to users
- Trey (NDEx) - Echo meeting the unmet need is key. How do I impact the largest number of people I possibly can? Pitch it right. E.g. Elsevier collaboration (embedding networks in journal articles)
- David (Xena) Get involved in an international effort, using standards. Supports coupling tools; interoperability.

Q: How do you work with your user community?

- Dexter (NDEx) We changed our staffing and hired a dedicated staff member devoted to outreach. Also, Dexter has a number of industry connections
- David (Xena) Work with the large projects (e.g. TCGA)
- Martin (BioC) User and Developer mailing lists. Now a stack-overflow type of approach, which has shifted the dynamic. Bioconductors strength is not its focus (unlike other projects). Ability to work with standard file formats has been key.
- Trey - Getting the developer community on board is equally important
- Aviv (Trinity) - 1) email requests; 2) more formal, in person outreach; 3) exploring online approach

Q: (Yantian Zhang) How do we generate resources that are greater than the sum of their parts? Plans for sustainability of these resources beyond the duration of funding - what are the opportunities?

- David: Modularity is key. Fund the "glue" that allows these resources to work together. Settle on the same concepts that are machine-readable and machine exchangeable.
- Aviv: Echo modularity. This is a fast-moving field and if tools are over-engineered, they are too hard to adapt. ITCR should think about supporting and incentivizing the "glue"
- Trey: But be careful when you build the glue - don't overfit this. "Glue can go wrong"
- David: It's important to know the right time to retire your tool so that it can be replaced with one more relevant

Q: (Rachel Karchin) How does ITCR fit into the Cancer Genomics Cloud Pilot project? How open are they for tools other than that developed by the project teams?

- David Haussler: Moving to the cloud like putting your stuff in a moving van and settling in a new home.
- Martin: A challenge, but one that has to be tackled.
- Juli: An important objective of the Cloud Pilots is to develop an infrastructure that can host tools like CRAVAT, MuPIT
- Bill Barnett: Other considerations for sustainability: Licensing, quality, credit for contributors; strong, dedicated project leadership
- David: GA4GH is using Apache license. Credit is very important and moving to GitHub helps with this. E.g. # of commits can go on a CV. NCI needs to acknowledge the power of the crowd for building software
- Martin: Make my tool work for everybody
- Trey: How much does the program want to drive ITCR 'translationally'?

Keynote: Stephen Friend: "Open collaborative research approaches to multidimensional biomedical problems"

Open collaborative research includes:

- **Data sharing:**
  - Synapse: improve transparency & reproducibility
  - Examples: NIA cross-project data sharing (NIH data portal for Alzheimer's)
  - Learning principles of sharing from the software world GitHub >> Synapse
  - Successful collaborative communities need an enabler who is not the PI, pushing on people to work together ("blowing on the embers")

- **Wisdom of the crowd:** other people you don't know have access & contribute
  - Dream Challenges - Can this be a vehicle for working together, not just a particular prize.
  - Drug synergy prediction challenge

- **Federated Approaches**
  - AML Federation: coordinated generation and analysis of functional and genomic data to Delivery of clinical care - 'Virtual Tumor Boards'; drug screen harmonization workflows

- **Consortia**
  - Share & compare analyses from 4 different groups (papers)
  - Required active encouragement/facilitation

- **Citizen engagement**: Apple's Research Kit
  - Incorporate open data and patient wisdom
  - Smartphone to boost large-scale health study
  - Example: Parkinson mPower study app
  - Participant centered studies = benefits to individual

U24 Clinical Research Project Perspectives

Talk 1: Rebecca Crowley Jacobsen: TIES Cancer Research Network
http://ties.dbmi.pitt.edu (sourceforge)

- Governance and SOPs may be the most important part of the project
- Providing a VM was *huge* for getting over the threshold of gaining use
- Documentation is absolutely critical
- Interact with users through discussions on SourceForge
- GoogleHangout office hours after releases
- Integration: finding the resources to source experiments that will be analyzed with other tools including ITCR tools

Talk 2: Jayashree Kalpathy-Cramer: Imaging Biomarker Tools

- QIN challenges (> 10 challenges underway)
- Make imaging data more accessible to non-imaging scientists
- Docker containers for image analysis and metrology tools in progress
- International challenge - share algorithms (data cannot leave country)

Talk 3: Guergana Savova - Cancer Deep Sequencing

- Breast cancer, ovarian cancer and melanoma are the driving use cases
- cTAKES: Started in 2006 at Mayo Clinic, NLP system for EMR text
- Early decision to be open source, be modular, use existing standards and conventions. Now an Apache project (since 2012) with a thriving community
- Using the HL7 FHIR framework for data exchange
- Integration with i2b2 and TranSMART
- Three user communities: NLP'ers, developers, end users (biomed investigators, point of care clinicians)
- SPARK parallelization of cTAKES - enable prototype to be more broadly used software

Talk 4: Andrey Fedorov, 3D Slicer project: QIICR http://qiicr.org

- Motivated by 3 QIN projects to evaluate technology
- QIN challenges: data sharing, tool sharing, indiv sites not tasked with platform. Some groups willing to share but they don't have a platform
- Users: clinical and pre-clinical researchers, engineers, commercial entities
- Project aims, paraphrased - 1: I want to repeat what you did. 2: I want to know what your result is and where it came from 3: I want to find what other people have done
- Using the BSD-3 license
- AppStore model of extensibility
- Users: clinical & pre-clinical researchers, engineers, commercial (product development)
- 119647 downloads, 205 publications as of 2012
- Found that it was important to have a journal article for citation.
- In GitHub, have merged >800 pull requests
- Weekly Google HangOuts
- Integrated with TCIA

Talk 5: Joel Saltz, Multi-scale analysis

- at the intersection of Pathology, omics, radiology, patient outcome
- quip.bmi.stonybrook.edu (quantitative imaging in Pathology)
- broad use cases addressed
- Integrative search linking pathology & omic

Industry Perspective: Jadwiga Bienkowska, Pfizer
"Applying Fresh from the Oven Computational Methods in an Industrial Setting"

Need: 'Home Depot' (tool store) for tools useful e.g BioConductor

- Inventory of tools with short description
- Maintenance with test examples
- Quality checks on the tools

Approach: CBDD Program - Computational Biology for Drug Discovery, by Thomson Reuters

- Not an exclusive club, anyone can join. Price is the same for all members.
- Standardized library in R (R-script library with computationally intense functions in java)
- Primarily network analysis tools, including node prioritization, edges prioritization, subnetwork ID, etc.
- Deliverables include development, documentation, training, unit tests

Top algorithms collected, annotated and prioritized by Thompson Reuters based on applicability, development time, performance, popularity, validation

Audience comment: Can you really set up a "Home Depot"? The only tools that are really useful are those getting uptake (publications)

Audience comment: Incentives in academia are for coming up with new tools. Few incentives for maintenance.

Q: It would be great if these activity was more of a "two-way street", rather than a "one-way-valve". Would be great if pharma could work directly with the academic groups on this.

- It's a practical problem. All of the companies were doing this evaluation independently. Makes more sense to do it once and share the information. Understood that none of these tools are perfect.

Q: How do you interact with the tool developers? Are you contributing back to the community?

- Currently, no obligation to communicate back the errors.

Q: Could you be violating licenses if you're not contributing back?

- Thomson Reuters is monitoring compliance with the licenses

Q: How could this be done in a more open way? That would be ideal. We would all be interested in participating.

- What are the incentives? Thompson Reuters is providing a service and they are very responsive. We are not opposed to a collaboration, it is just more difficult.

Q: Is Thompson Reuters' Metabase part of the collaboration?

- No.

Q: Would it be possible to pay a third party to make the documentation you are interested in come together and organize

- A brokering role is required to achieve coordination and economy of scale

Audience comment: There is some parallel here with NCI's drug evaluation program.

ITCR PI Meeting – May 27-28, 2015
UC San Diego
Day 2

Update on the ITCR Program - Juli Klemm
- Renewal of the ITCR program has been approved
- There will be some modifications
    - Support for sustainment of existing resources (U24)
    - Support for development of innovative computational methods and algorithms (R21)

Panel: Common considerations for early-stage development projects

Talk 1: Bobbie-Jo Webb-Robertson - Interactive Informatics for Research-Driven Cancer Proteomics
The need:
- Explore proteomic data associated with cancer
- Address the realities of high instrument variability and biological variability

Goal - enable biomarker discovery using online software tools = **P-Mart**

Talk 2: Rachel Karchin - Informatics Tools for High-throughput Analysis of Cancer Mutations
Project goals:
- Scale of mutation analysis for larger data sets
- Concise and interpretable results
- Enable hypotheses generation
- Interoperabilty with community tools and pipelines

CRAVAT + MuPIT:
Funnel for automated mutation analysis to get prioritized list of relevant/interesting variants and genes
Bioinformatics scoring made easy
Examples:
1. Familial pancreatic cancer: 70 million SNVs reduced to 8 SNVs
2. Metastatic triple negative breast cancer cisplatin treatment

Unique users: 6056
Planning mutation analysis services

Talk 3: Ben Berman - Tools for regulatory analysis of large cancer methylome datasets
Epigenomic organization reflects enhancer activity
International Human Epigenome Consortium

Need: Many public epigenome data sets require integrated analysis for biological interpretation

ELMER: Enhancer Linking by Methylation/Expression Relations
BioConductor package and GitHub Repository

U01 Plan: Implement Bioconductor package in Galaxy; support comparisons between cancer subtypes or clinical covariates

Talk 4: Josh Stuart, Kyle Ellrott - BMEG: Biomedical Evidence Graph

- Building the evidence graph from combined data sets
- Same algorithm run in different places gives different results. Need portable scientific analysis

Talk 5: Peng Jiang (Shirly Liu's lab) - Developing Informatics Technologies to Model Cancer Gene Regulation

1. Cistrome: Analysis pipeline
2. Cistrome: Data collection
3. RABIT: Cancer modeling rabit.dfci.harvard.edu
4. Xena: Visualization

Talk 6: John Quackenbush, MeV (Multi experiment Viewer)

- Positioning: Bridge translational scientists and bioinformaticists/data scientists
- Interface to tool in BioConductor
- Web-based MeV currently beta testing (source code in GitHub)
- Next steps: Deploy on AWS and enable data and result sharing and discussion through cloud storage

Panel Discussion

Q: How do you work with users to understand their needs?

- Conduct workshops to get feedback
- Collect responses from Github
- Cannot understate the importance of responding to email inquires. Users reporting problems or making requests can be turned into partnering opportunities
- Be proactive with documentation (work with developers)
- Talk to collaborators, especially local users - need to jump on ideas
- Biologists don't know always know what they need Can't just present people with a blank slate - give them some initial ideas and ask whether it is useful or not.

Q: How do you prioritize activities? Resources are usually provided for new work and not necessarily for sustainment. How do you balance these needs?

- Partner with software company for implementing tools as high-quality software
- Nothing better than watching somebody use the software. Having developers watch people try to use their software can be a valuable learning experience.
- Recognize different user groups and understanding their needs
- Keep in mind that even the largest software projects (Microsoft, Oracle) don't respond to all feature requests. Ok no to do everything and to stay focused.

Q: What provenance information to you believe is important and how do you determine this?

- Workflow systems such as Synapse and Galaxy have approaches to capturing and storing workflow provenance. Unclear how harmonized their approaches are.
- GA4GH has a working group (project?) focused on provenance/containers/workflows. Includes a workflow description language.

Q: How do you deal with the lack of citations for software tools?

- Jill's example: See an IGV figure in a paper, but no citation. Difficult for an editor to pick this up.
- Idea: Journals could provide a checklist to reviewers (and/or submitters?) to ask for confirmation that all tools used to generate results for the paper have been cited.

- Jerry: Analogy is the provision of structure factors by x-ray crystallographers. Used to be that these were not required to be submitted. Users began demanding this and PDB and publishers began requiring it. Can the bioinformatics software community take the same approach - grassroots approach?
- Can the NIH get involved here and require grantees to provide information on the tools they are using? This could be a resource that could be mined by the developer community.
- This group should set an example. For every paper, provide a container with the data, software, etc.
- Keep in mind the distinction between citability and reproducibility

Q: For those of you using cloud resources, you have the opportunity to track usage in a very comprehensive way. Are you taking advantage of this?

- Understanding usage beyond just downloads is very important. Someone who downloads the software doesn't necessarily use the software.

Q. What can the ITCR program do to help meet the goals of your project?

- Allocate funding for collaboration - multiple people supported this idea
- Have scientists participate and present problems to solve - proactive matchmaking
    - Consider an AACR Educational Session
- John: rather than the NIH acting as matchmaker, provide a mechanism for the tool developers to host workshops (R13?)
- Rachel: more opportunities to get together, on focused topics.

Training & Outreach Working Group Report - Rebecca Crowley Jacobsen

Top 3 activities of the working group thus far:

1. NCI Center for Cancer Research F2F training sessions
2. Publish 'explainer' videos
3. Presentation at conferences, including CI4CC

Highlighted success stories:

1. 3D Slicer: Documentation; cite your users; project week
2. Bioconductor: Everything! BioConductor: course materials, mailing list, see youtube on graph based viz
3. NDEx: Innovative use of LinkedIn

What should TOW do next? Discussion

- Consider a working group to discuss core software/infrastructure. across projects (this would be distinct from the Training and Outreach WG)
- Do much more collaborative work together
- Identify specific conferences to target
- Get these diverse spectrum of tools works together - flow of tools; string all of the analysis you can do. Create an end-to-end use cases using as many ITCR tools as possible.
- Create focus group: Determine goals of working group, scientific focus to address a biological problem
- TCGA -based working groups - PanCanAtlas and PCAWG. ITCR investigators should join these!
    - Josh will follow up with an email to ITCR participant list
- Short papers on different tools