emerge network
ELECTRONIC MEDICAL RECORDS & GENOMICS

Summary of Steering Committee Meeting: Summer 2019

June 20th-21st, 2019 in Seattle, WA

TABLE OF CONTENTS

eMERGE DAY 1: THURSDAY

- **NHGRI program official report |** Robb Rowley *(NIH/NHGRI)*
    - NHGRI is developing a strategic plan for the future, and looking for input. There is a new eMERGE RFA; questions can be directed to emergerfa@mail.nih.gov
    - Several funding opportunities are available. The NHGRI is looking for great proposals from clinical genomics applications.
        - The Electronic Medical Records and Genomics (eMERGE) Genomic Risk Assessment and Management Network- Coordinating Center (U01 Clinical Trial Required)
            - Application Due Date: August 2, 2019. Expiration Date: August 3, 2019
        - Investigator-Initiated Genomic Medicine Research (R01 Clinical Trial Optional)
            - Application Due Date(s): October 21st, 2019; June 19, 2020; October 20, 2020 Expiration Date: January 23, 2021
        - Limited Competition Centers of Excellence in Ethical, Legal and Social Implications (ELSI) Research (CEER) (RM1 Clinical Trial Optional)
            - Application Due Date(s): July 23, 2019 Expiration Date: July 24, 2019
        - Genomic Innovator Award (R35 Clinical Trial Optional)
            - Application Due Dates: October 30, 2019; October 30, 2020; Expiration Date: October 31, 2020
            - eMERGE investigators are encouraged to submit to the innovation award.
        - Novel Approaches for Relating Genetic Variation to Function and Disease (R21 Clinical Trial Not Allowed)
            - Application Due Date(s): July 16, 2019; November 16, 2019; July 16, 2020; November 16, 2020; July 16, 2021; Expiration Date: September 8, 2021
        - Ethical, Legal, and Social Implications (ELSI) of Genomic Research Regular Research Program (R01)
            - Application Receipt Date(s): Standard dates apply. Expiration date: September 8, 2020
        - Genomic Community Resources (U24)
            - Application Due Date(s): First due date is July 13, 2017; Expiration Date: January 20, 2020
    - Specific research questions can become notices of special interests (NOSIs). NOSIs are ways to increase efficiency in expressing interest in a scientific area while reducing the administrative burden of traditional program announcements.
    - The eMERGEseq Freeze1 (Phs001616) was submitted to dbGaP on 5/1/2019.
    - There are several datasets committed to deposit into dbGaP before 2020, including the GWAS dataset (N= 83, 717) with the updated case/control file, the GWAS dataset with additional Harvard and CCHMC samples (N= 105,000) and the eMERGEseq Freeze V2 (N=25,000).
    - The Network should pursue how to explore discrepancies in ROR between sites and how to interpret non-responder data to help understand the impact that this will have on outcomes data
    - The Network should prioritize publishing lessons learned papers as Year 5 funding continues, specifically focusing on Network-wide lessons learned.
    - Establishing XML standards for genetic testing results will help implement genomics into medical care. The eMERGE Network has been instrumental in helping develop FHIR standards that will allow the broader adoption of structured genetic test data into the EMR.
- **Announcements, opening remarks |** Rex Chisholm *(SC Chair, Northwestern)* |
    - AnVIL was confirmed as an affiliate member for pilot testing of eMERGE data.

- eMERGE will be one of the first Networks to share data using AnVIL and will help shape the AnVIL data repository for discovery moving forward.
- The final stages of ROR is in progress and four sites have already returned their results.
- Four sites have completed ROR. There have been 19 Outcomes forms harmonized and deployed for data collection and abstraction guides. There is a significant amount of Outcomes data available over 640 forms.
- EHRI group is working with HL7 clinical genomics workgroup to map eMERGE scheme to FHIR standards.
- The ESP would like for the Network to focus on lessons learned. The Network will address penetrance moving forward and specifically during the October 2019 lessons learned panel.
- The Network should review in development projects and prioritize which projects are feasible to complete within this phase.
- The Network has a good history of publications but there are concerns about the number of manuscripts in development. The Network should prioritize publishing existing manuscripts, and lessons learned papers.
- There have been 1350 external downloads of dbGaP as of May 2019.
- Goals of this meeting include: genomics lessons learned panel, update on AnVIL plans for eMERGE dataset utilization, demonstrate progress of the seven network wide milestones and lessons learned manuscripts, and discussion of any major barriers for work completion in Year 5, including return of results, loss of follow up of participants, and impact of penetrance studies.

- **Genomic data update |** David Crosslin & Ian Stanaway *(UW/CC)*
  - The GWAS V3 SNV imputation is almost complete.. Network investigators can find an updated report here.
  - The structural variation imputation set QC metrics have been looked at to release. There is about a 90% concordance with CNV calling, and about 60% of them compute with high concordancy.
  - There are multiple MCS's for discovery analyses planned with the 105,108 eMERGE array samples.
  - NT314: Ian will be imputing structural variations from the GWAS array samples, and he is looking for potential collaborations.
  - The eMERGEseq V2 sample (N = 24,956) has been compiled. Ian is running PCA analyses.
- **CSG Updates |** Eric Venner (BCM/HGSC) & Hana Zouk *(Partners/Broad)*
  - A new CNV method that improves the sensitivity and resolution of CNV detection has been recently published in the paper "Atlas-CNV: a validated approach to call single-exon CNVs in the eMERGESeq gene panel" (Chiang et al, 2019).
  - Second set of reports are coming from ClinVar, that are 'mashed' up clinical variations. In the past month, BCM has identified 158 variants with a new pathogenic classification. There are eight remaining.
  - The ReVU (reanalysis of variants and updater) tool was developed to help with ClinVar reanalysis.
  - 148 remain VUS, 10 passed on for second review (eight remaining), second review ongoing.
  - To manage updated reports, ARBoR (Authenticated Resources in Hashede Block Registry) was recently released. All reports have a barcode; barcodes are scanned and verified to check validity using ARBoR Client Python API and ARBoRScan iOS and Android app. Report histories are recorded in encrypted ledger. A paper titled "ARBoR: an identity and security solution for clinical reporting" was recently released. (Venner et al, 2019).
  - All of the VUS that are likely pathogenic are pulled out of the sample (n= 79). 50% are cardiomyopathy or arrhythmia genes.
  - There were sex discrepancies found in SNP data (n= 73) and metadata that was reported in January and most of the samples were removed.

- As part of the QC for the full eMERGE dataset, the CC reviewed heterozygosity in X and Y variants and matching to the sample metadata (submitted sex). Initially over 2400 samples were flagged due to a CHEK2 duplication but when issues were resolved there were only 222 still discrepant.
- FHIR implementation is led by Larry Babb and Mullai Murugan. The EHRI WG is mapping the eMERGE reports to the FHIR standards to create a pilot implementation at a clinical site.
- There is a discrepancy resolution group within ClinGen that prioritizes medical data. It is the intent to prioritize anything eMERGE-relevant in the group.
- Alerts have been in reporting system, and 11 out of 565 reported variants that have issued reports that have affected 23 reports.
- BCM and Broad have not compared the specific criteria that they are using to reclassify the VUS. Comparison of the specific criteria used for each lab is hard to determine. In MRA7 there are very specific rules. For others as long as we use the same classification is fine but in others this may not be the case.
- Participant and provider responses to reclassified variants is currently not being captured, but there is an opportunity to do this on the 12-month outcomes form. This could be discussed in the Outcomes workgroup.
    - This was not built into the ROR protocol and how this will be done is important.

- **Update on an investigation of somatic mutations in eMERGE datasets |** Ken Kaufman *(CCHMC)*
    - Ken gave update on ongoing somatic mutation project at CCHMC. Particularly relevant given the update with P/B showing that our understanding of somatic mutation can affect health (Yizhah et al, 2019).
    - Unlike germline mutations, somatic mutations are spontaneous and can only be detected in a small subset of cells.
    - Throughout the data, there are places that do not match. This can be sequencing artification or somatic mutations.
    - CCHMC processed two large datasets, including: eMERGE elll set A (16,170 samples) and PGx (10,000 samples screened) and found 6% somatic mutation out of both datasets. 8% of reads have a polymorphism.
    - Cardiomyopathy genes appear on the list of most somatic mutations.
    - Findings demonstrated somatic variation in 178 genes, with 61 severe polymorphisms. Some polymorphisms appear to be germline. 225 appear in the VCF file, however only 75 correspond to the sample.
    - One of the underlying goals of the project is to identify how damaging the variant will be. Findings showed that cells with somatic mutation *are* more likely to have a biological impact. Looking at functional predictions of the variants (how damaging) could be a way to prioritize the variants.
    - Alternate allele frequency is seen about 1-3% of the time indicating that they would be present in up to 6% of the cells. Some are present in up to 30% of the cells. A lot of alternate alleles have variants in very low frequencies.
    - Investigators obtained DNA from VUMV and NU, and tested 11 samples for validation using digital droplet PCR in order to validate results and were able to validate nine out of the 11. The samples that had digital droplet PCR had results close to that of the sequencing data.
    - It is important to collect as many samples that have been sequenced in order to validate. These include samples from PGx data and the eMERGE III eMERGEseq freeze data.
    - The group has been working with iGENOMX Riptide to sequence quickly, they use primers that look specifically at SNPs that investigators are interested in.
    - The problem with a lot of samples coming in for validation is that the cost becomes prohibitive. Using this technology they will be able to look at 96-960 candidates for about $15-$20.
    - Currently in the process of obtaining samples for validation and finalizing the validation strategy.

- ACTION ITEM: Investigators with questions or samples for validation regarding the somatic mutation project, NT171: *Possible somatic mutation in targeted sequence data,* should contact Ken Kaufman (kenneth.kaufman@cchmc.org) and Paul Gecaine (paul.gecaine@cchmc.org).
- With regards to which samples or criterion was used to re-test, everything has been data driven. Primary goal is to develop a pipeline that identifies a cost-effective method for identifying somatic variation from NGS; criteria includes many samples with a range of variants, including a range of alternate alleles to determine limitations to identify calling mechanisms for different variants.
- The group discussed implications for identifying different somatic mutations in different alleles
  - All three top mutations are in conditions including cardiomyopathy, which was also discussed earlier today. Potential for looking at differences in age, or other biological impacts.

- **AnVIL Progress & Discussion** | James Taylor *(JHU)*
  - The goal for AnVIL is to create a scalable and computable cloud-based resource that supports data access and storing of large Network datasets driven by NHGRI programs. AnVIL will promote greater cost-effectiveness, and be more secure and compliant than prior data repositories.
  - AnVIl will be collaborative for datasets and analysis workflows and a hub for data access.
  - Terra, which houses AnVIL, facilitates analyses for workflows and mapping. WDL, CDL, Jupyter Notebook, Python, and R will all be supported.
  - Dockstore is a hosting service for different tools and workflows that can hold the work of Terra.
  - Galaxy, a web-based analysis environment for running analysis tools and building workflows for users with little to no programming or bioinformatics experience. Galaxy Toolshed has over 6,000 tools.
  - AnVIL needs a complete system to query data. Gen3 is currently filling this need, as it creates data models, indexes and queries.
  - AnVIL is working on creating a complete bioconductor environment for storing data.
  - AnVIL's security configurations and controls for data management are equivalent to those of a Federal Information Security Management Act (FISMA) Moderate System environment and built on top of the Google Cloud platform.
  - AnVIL will roll out data models incrementally; a soft launch is expected at the end of June. The eMERGE data is not going to be included in the initial June rollout.
  - In 2019/2020, AnVIL will add additional analysis environments for researchers to expand and use.
  - New 1000 Genomes High Coverage Data platform is open access; investigators can look directly at the data objects that are located in AnVILG. The data objects can be mapped into workflows to be used for notebooks.
  - Other initial datasets include GTEx Versions 7 & 8 and selected datasets from the Centers for Common Disease Genomics (CCDG) and Centers for Mendelian Genomics.
  - eMERGE's adjacent datasets are different than the current GWAS data stored in AnVIL which will require a bit more work to prepare.
  - eMERGE will be the driver in building models due to facets such as structured phenotype models. Data can be analyzed in the platform using a wide variety of tools supported by AnVIL.
  - The hope is as additional projects, tools, and datasets are brought in, this can be coupled with training materials for those who will be coming in to use this data. By providing this platform, they can train investigators using the best next methods on data repositories.
  - If you have approval in dbGap to use a project you will be able to use that project in the AnVIL.
  - Broad has a system that the Network is piloting as part of the launch, called the Data Use Oversight System (DUOS).. This will check if data access requests are compatible with the data use restrictions to automate use of data.

- Workshops and tutorials are planned for upcoming scientific meetings. There is a workshop at ASHG onhow to use AnVIL.
- For some datasets, access will be free, but analysis will incur a charge.
- All data access and usage on AnVIL will be audited, via the GCP auditing system. Every person that accesses the AnVIL data layer is documented. The AnVIL team is aware of who accesses the data and users are required to go through the authorization process.
- There is an access control framework built into the platform through Google compute. Rights form dbGaP will transfer to AnVIL users. This will be accomplished with users using their eCommons ID.
- Regarding DNAcommons project migration to other cloud environments, investigators would need to build an extraction layer in order to push data into another cloud environment.
    - There are some google services AnVIL will take advantage of. The intent is to use big query for large multi-sample VCF calling.
- AnVIL storage needs to have identifiers from data model mapped into data file. Each graph has figures mapped to ontology terms. Multiple data models are possible, however, integrative approaches across projects require mapping. Single data models are more simplified. AnVIL would simplify data model to facilitate querying.
- AnVIL is funded by two NIH U24-grant currently, with five years of support. Storage for datasets that are made available to the public are currently paid for by the grant. Ingress fees (add data) are covered by Google. Egress fees (pull data down) are paid by investigators toGoogle, but hopefully the data will be computed on the platform and users only egress derived data.

- **Challenges in Implementing Genomic Medicine in a Federally Qualified Community Health Center: Insights from Mayo Clinic's RAVE-Phoenix Study** | Gabriel Shaibi & Valentina Hernandez *(Mayo)*
    - Arizona Return of Actionable Variants Empirical (RAVE) study enrolled 500 Latino participants from SPS biobank with either hyperlipidemia or colon polyps. Participants consented to have their sample sequenced and the results, actionable and negative, to be returned to them...
    - Dr. Laney Lindor, program director spoke with participants explaining the implications of ROR.
    - Arizona's ROR process is detailed in the publication "Developing a process for returning medically actionable genomic variants to latino patients in a federally qualified health center" (Shaibi et al, 2018).
    - 486 letters were sent for negative/neutral results. There were 926 follow up phone calls, 57 participants reported they did not receive the initial letter and requested a second letter which was sent by certified mail.here were no requests for consultations with medical geneticists after receiving result.
    - There were 50 individuals who were brought back in with neutral or negative findings about four months after the site received theirresults and were interviewed.
    - Participant motivations for becoming involved with the study included personal, family, and resource benefit and contributing to their community or public good. Major concerns included fears or worries about a positive diagnosis, unfamiliarity with research studies, and a distrust of medical institutions and genomic screening.
    - Although participants were recruited for hyperlipidemia and colorectal cancer, only one participant had an actionable result of a lipidemia gene; four results were shown as P/LP.
    - The average time from enrollment to ROR was 549 days. There were 10 actionable results. There were significant logistical challenges with many individuals. It required 3.2 calls and letters to re-engage, 6/10 were uninsured, 3 certified letters were marked "unclaimed" or returned, and one participant never responded to calls or certified letters.
    - Eight participants only spoke Spanish. To account for these socio lingual differences, the program ensured there was always an interpreter available.

- The researchers highlight that it is important to think about how to approach participants with positive results, and the emotional impact of those results will have on the patient.
- There is a need to meet the community where they are in regards to location, timing, cultural preferences, needs, and practices, as well as financial burdens and insurance status.
- Speeding up the turnaround time in returning results would be helpful in eMERGE IV.
- Primary care providers need as much support as can be given (resources/educational materials).
- This is a time where researchers can be forward-thinking and determine how the communities can be best help by genetic testing.

- **Meharry site experiences with recruitment and ROR** | Raj Singh *(Meharry)*
  - Meharry's role in eMERGE is to enroll 500 African American (AA) participants who are high-risk for cancer or with selected cancer. Evidence shows that the incidence of cancer is highest among AA compared to all racial populations. This is particularly poignant for AA males. Participants were recruited for breast cancer, prostate cancer, colorectal cancer, and lung cancer.
  - Investigators obtained blood samples for DNA extraction for germ-line sequencing as well as RNA genomic analysis and proteomics, captured socio demographic survey info and past and future health outcomes from EMR access.
  - Recruitment was started in April 2017, and enrollment began in September of 2017. 500 individuals were enrolled by February 2018. Meharry began returning results in May 2019.
  - Participants were predominantly recruited as at-risk for colon and prostate cancer. Interestingly, the least number of patients recruited were patients who had lung and/or colon cancer.
  - Results showed that of the 500 total results, 496 were received. Of these, 19/496 were positive.
  - Only two patients out of the 19 were able to be reached to discuss the results; one scheduled with a genetic counselor and one declined.
  - The results in the BRCA group showed six positive results. There were five positive results in the lung cancer group, four at risk. In the CRC group, there were four positive results.
  - ROR plan for negative results includes contact up to three times until the participant is reached followed by a letter. Challenges included not being able to extractDNA , had to amend the protocol to allow for return of non-cancer related genes, managing the large turnover in staff, including lead investigator of the study and coordinators, and Meharry had a poor rate of reaching participants in the early ROR phase.
  - Several notes regarding patients with positive results who lack insurance and need further follow-up:
    - Meharry suggested they meet with a genetic counselor and follow up with their PCP.
    - There is an antigen program available for participants to get tested at no cost.
    - Other sites have noticed that these same challenges. There is a network of providers that sites work with, but many do this on a cash price. This is a challenge that eMERGE could address moving forward.
- **Exploring the Impact of Family and Personal History on the Perceived Value and Usefulness of Negative Genetic Test Results** | Sharon Aufox *(NU)*
  - Northwestern has been interested in ROR of negative results. There are concerns about receiving negative genomic results, including participants not understanding negative results or misunderstanding results, and participants not engaging with follow-up preventative screening.
  - Questions were focused on gathering data on usefulness and value on negative results.
  - 178/336 (53%) participants completed the survey..
  - First question that participants were asked was related to family history. Most participants (~60%) were interested in how their family history affected their results.
  - Reactions to the negative results were collected in the qualitative results. Most participants (~65%) felt that they were less worried about their families reactions.

- Most participants (~70%) felt a 'peace of mind' about receiving negative results.
- How participants altered their life was also tested; it was found that the majority of participants (~90%) did not really change their lifestyles based on their negative results.
- Most participants (~90%) said that they would agree to additional genetic testing in the future even if it resulted innegative results.
- Limitations included a small number of participants, minority populations were not well represented, and the population was highly educated.
- Questions about concern of residual risks in negatives results were considered but were not included in the survey due to complexity.

eMERGE DAY 2: FRIDAY

- **Opening remarks |** Robb Rowley *(NIH/NHGRI)*
  - The Network is trying to wrap up ROR, the next 10 months will be focused on outcomes and penetrance.
- **Workgroup updates on milestones & discussion|** Moderated by Rex Chisholm *(SC ChqiNorthwestern)*, led by Workgroup co-chairs
  - Each workgroup was given two sides to summarize the current work that is ongoing in the respective workgroups, challenges encountered, and lessons learned.
  - ROR Workgroup: Iftikhar Kullo and Ingrid Holm
    - Many sites are in the process of returning negative results. ROR has been completed at four sites, and anticipate completion in other sites within the next few months.
    - The workgroup has different manuscripts with a focus on ongoing work and lessons learned, including Sharon Aufox's project presented yesterday.
      - Leadership has noted that the importance of focusing on completing the projects in development as well as lessons learned manuscripts. These prioritized projects can be decided upon in the next nine months.
      - The Participant Survey and Healthcare Provider Survey subgroups also have manuscripts in progress.
    - The group is currently working to define the subsets of Non-Responders (e.g. decliners, deceased participants, transition to adulthood). Prioritization has been placed on understanding non-responder status and the impact of non-responders on outcomes and penetrance analysis work.
    - To help consider geographic and environmental factors of non-response, the group plans to partner with the Genomics group for a geocoding project on non-responders.
    - Initial work can be done on variants with three or more instances to give insight on the time it takes.
    - ACTION ITEM: BCM and LMM will share variant list in order to determine the quantity of P/LP variants for a given disease and help to prioritize efforts.
    - BCM plans to have combined list of VUSs that might require reclassification by late-June, and the combined list from LMM and Baylor can be planned to be available by early-July.
    - ACTION ITEM: The Clinical Annotation workgroup and CSGs will provide details of the flow from a laboratory standpoint for VUS's that are reclassified.
    - ACTION ITEM: The ROR/ELSI workgroup will discuss common themes among sites regarding providing reclassification information to participants.
  - Clinical Annotation Workgroup: Gail Jarvik and Heidi Rehm
    - Clinical Annotation is focused on incidental findings across the network.

        - The preferred order for obtaining the outcomes data for penetrance analysis has been defined.
        - The group is waiting for the 6-months outcomes data to be entered for the Tier 1 diseases.
        - There are several VUS that have been upgraded to L/LP and reports are being generated for return.
        -
        - The group is discussing ways to provide feedback on the outcome forms penetrance questions.Collecting "non-responders" and "non-consenters" data should still be captured for penetrance analysis.
        - Communication between the leaders of Clinical Annotation, ROR, and Outcomes will continue to ensure the Network can maximize the data collected and analyses produced.
    - Outcomes Workgroup: Josh Peterson, Hakon Hakonarson and Marc Williams
    - As of June 2019, the group has completed outcomes data for over 500 participants, many that have more than one phenotype form completed.
    - There are approximately 1430 total outcomes forms to be completed.
        - The group plans to work with the ROR/ELSI workgroup to determine how this total number is affected by "non-responders" and "non-consenters." In addition, there are subsets of eMERGE patients without returned results.
        - According to the ROR Progress Tracker, sites received 1427 positive results and had 267 non-responders (some including PGx results). 1006 positive results were returned to participants.
    - To ensure the correct forms are being used for outcomes analysis (returned results) and penetrance analysis (returned and not returned results), a separate penetrance analysis form will be created for results not returned.
    - ACTION ITEM: The CC will work with the Clinical Annotation workgroup to create penetrance analysis forms for participants with results not returned.
    - For data analysis and QC, two data freezes are anticipated: the first will be released in September 2019 for the October SC/ESP meeting and a second in January for the Winter SC meeting.
        - The first freeze analyses are solely for the steering committee meeting presentations and initial QC of the data.
    - A general consensus is that the date of return should be defined as the date the result was added to the EHR.
    - ACTION ITEM: The CC will add the break down of returned results, non-responders, and non-consented individuals regarding Outcomes & Penetrance analysis topic to the July 2019 PI call for discussion.
    - ACTION ITEM: The CC will circulate the ROR tracker to the Network for reference.
    - EHRI Workgroup: Sandy Aronson and Casey Overby Taylor
        - There is progress being made on the transformation of work from XML format to FHIR.
        - A subgroup has been created for FHIR integration.
        - ACTION ITEM: Investigators interested in joining the FHIR Integration EHRI Subgroup should contact Michelle Stone (timoethia.m.stone@vumc.org) at the CC.
        - .
        - The FHIR integration subgroup will continue to identify use cases and technical implementations.
        - Two main papers under review:

        - NT352: Lessons from eMERGE on readiness for genomic clinical decision support implementation
        - NT301: IGNITE Clinical Informatics Working Group: Genetic Data Pipeline Project
    - The EHRI workgroup is in-process of performing a hazard analysis. The plan to use Hazard analysis to mitigate some of the hazards with integrating genomic information into the EMR that were identified.
- Genomics Workgroup: Megan Roy-Puckelwartz, David Crosslin, and Patrick Sleiman
    - The 105,000 GWAS and eMERGEseq V2 datasets are almost released.
    - All datasets are moving along quickly.
    - The SPHINX focus group discusses the new tool 'eMERGENT', David Crosslin has submitted a U24 for this work.
    - Collecting and compiling data is a challenge for this workgroup.
    - The CC will release the eMERGEseq V2 dataset shortly, there is just one more issue regarding 222 sex mismatches, that likely occured when the site sent demographic data to the CC, not necessarily contamination or sample swap issues. This will be described fully in the QC report. The CSGs and CC also realized there was duplicate mapping to a CHEK2 region and this caused some variants to map improperly to the Y chromosome, this region has been excluded from analysis. .
- Phenotyping Workgroup: Chunhua Weng and Wei-Qi Wei
    - Development of 23 of the 25 main algorithms have been completed. Many of them have already been implemented across the Network. Of the five NLP algorithms, two have already been developed and are currently being validated.
    - Small sample sizes and low PPV results has led to issues in algorithm development.
    - Lessons learned papers are the focus of the workgroup as well as learning how to implement NLP to improve portability of workgroup.
    - The Goal is for all original algorithms to be released by the next SC meeting, the group will continue to work on the NLP algorithms throughout Year 5.
- PGx Workgroup: Laura Rasmussen-Torvik and Cindy Prows
    - eMERGE PGx was prominently featured at CPIC in June.
    - Implementation of PGx is particularly active. The group is looking for Network-wide projects to continue PGx analysis.
    - ROR of PGx is being examined and completed at VUMC and CHOP.
    - NT260 is in-progress and will be submitted for publication soon.
    - One area for feedback would be what is the best way to influence EPIC leadership to make it more user-friendly. PGx is interested in publishing implementation strategies for EPIC as it is difficult to work with, and relay feedback in literature.
    - There is a small eMERGE group that works on implementing family health history into the EMR.
    - Family history or genetic information should be able to be inputted into EPIC in a standardized fashion in order to be used in the EHR.

- **Outcomes Workgroup Science Presentation: Case Studies |** Josh Peterson *(VUMC/CC)* Maureen Smith *(Northwestern), Alanna Rahm (Geisinger), & Cindy Prows (CCHMC)*
    - The Outcomes workgroup has collected over 500 records of outcomes data from the EHR including several interesting cases of diagnosis and treatment.
    - Outcomes data collection tells the story of return of results to participants and how participants are affected by the result.

- Maureen presented a case study of a Northwestern participant.
    - The participant's primary care provider was also part of the Northwestern healthcare system.
    - When enrolled, the participant stated that he had no family history to suggest genetic testing. The results came back with two positive mutations of two forms of cystic fibrosis.
        - Results were returned over the phone.
        - He was thrilled to receive result and began relaying information based on his health history including chronic sinusitis, digestive issues, and infertility.
    - Participant was referred to the adult cystic fibrosis clinic located in the children's hospital and received full series of tests: chest x-ray, vitamin-d test, sweat chloride. Histherapy, resulting from his genetic testing, has made a big difference in his care and his outlook.
- Alanna described case findings of a 61-year-old MyCode participant from Geisinger with multiple comorbidities: type 2 diabetes, kidney disease, hypothyroidism, PAD, anemia, and obesity.
    - Recently before receiving result, the patient had undergone an amputation of her right foot which led to ischemic stroke as a complication. An LDLR mutation result was returned in August 2017 while she was in the nursing home. She did not keep the appointment with the genetic counselor due to recovery issues. Shehad another amputation after return of result which led to another stroke. The patient also has extensive coronary artery disease. The participant is unable to tolerate statins and now has a BRCA diagnosis. Her son had a heart attack at 39 and was unable to take care of himself.
- Cindy presented the case report of an adolescent participant from CCHMC. The adolescent and parent independently chose to learn all possible results.
    - Participant's father and sister had a history of migraine and temporary muscle weakness/paralysis. They were evaluated in the genetics clinic. The mother reported periodic numbness, falling, and word choice issues, however previous neurology evaluations were found to be negative. The father had hemiplegic migraines and engaged in post strenuous exercise when younger.
- These three cases indicate the challenges for outcomes research but for the healthcare system in general in synthesizing information.

- **Genomics Workgroup lessons learned panel & discussion |** David Crosslin *(UW/CC)*Megan Roy-Puckelwartz *(Northwestern)*, Ian Stanaway (UW/CC), & Mullai Murugan (Baylor)
    - Megan Roy-Puckelwartz " *How sample size, diversity , and target genes affect implementation and discovery of dataset* s"
    - Genomics is widely successful when you have a very large, rich dataset. Genome-wide data genotyping creates a richer discovery set, but requires extensive phenotyping algorithms.
    - Panel-based sequencing create a richer implementation data set. Panels can select genes to understand how genes interact, to understand penetrance, CNV analysis, mosaicism. This approach is not likely to result in new gene discovery. When you have panel-based sequencing, you need to be more thoughtful of how to create discovery hypotheses. Need to be vigilant about goals, and clear hypothesis need to be made.
    - Adding additional data is costly and time-consuming. Need to weigh benefits and costs.
    - continuing to add data and recreating multisample VCFs results in delays in data analysis. Groups need to be mindful that there will always be another dataset that is bigger and better.. The challenge for Networks is to understand the importance of data freezes and establishingclear goals and objectives to maximize the science for a given amount of funding.

- An additional challenge is when groups create new requirements resulting in delays of analysis. This has been evident when NHGRI pushed for diverse datasets; this additional requirement lead to delays while we were waiting on the data from these samples to be available.
- eMERGE has very large datasets, and the group needs to understand how size affects the speed and efficiency of analysis and QC. There needs to be better prioritization vs. speed and efficiency.
- Investigators should limit the time for comments and suggestions when consensus is required. Make decisions quickly, and learn to prioritize are important elements regarding compilation and analysis of large data sets.
- An example of the importance of organizing and prioritizing dataset analysis was with the Genomics Workgroup. They did not have the necessary power to run certain data because the group did not know the phenotype status. This is important especially in the future if anydata is not present or available.
- Ian Stanaway "*Lessons Learned and Earned: Compute Time with Big Genomic Data and Naming Conventions and IDs*"
- To manage the array data (83k), it takes 96 hours of compute time for the largest batches. It took about six months to actually impute and merge the 78 sets. This does not include the QC. The compute time is about a month and a half when a change is made to put everything back together again. Ian cannot finish the genetic QC until the genetic and demographic files have been uploaded to the CC. This would then require a nine day rewrite.
- Collecting and finalizing the demographics file at the beginning of the process and freezing the dataset at that point, would help with frequent withdrawals and additions which add significant time . Potentially could move the workflow to Github to help upload data.
- For the GWAS set, there are several different ways the files were titled, including some without eIDs, he had to write code to detect differences and standardize.
- Investigators should consider to enlist IT in streamlining heterogeneity of naming and organizing data prior to initiating a large sequencing or genotyping project.
- The Network needs to be more disciplined about data freezes. Even once the pipeline is built, adding and subtracting data take significant amounts of time and effort.
- It is unknown how much this kind of work would cost on AnVIL, or a cloud computing environment.
    - Each restart and test there will be computation cost, so it will be more important to plan out analyses and data freezes on the front end.
- Mullai Murugan: "*eMERGE Commons and Genomic Analysis*"
- eMERGE is a complex ecosystem and requires a storage and analysis platform. The eMERGE Data commons primary goal is storage, but gives a portal that also allows for analysis. As there is so much data, there is a need for an analysis platform that gives access for data mining and research.
- The PHI partition allows access to authenticated users of clinical reports and other related files. The non-PHI partition focus contains raw data, access to all clinical sites, the analysis tools, and the BCM/LMM data. eDAP portal serves for PHI partition storage.
- eCAP is the eMERGE commons access portal, and serves as one of the analysis tools.
- There is a structural variation project with parliament 2. Aims are to increase resolution and sensitivity of CNV calls, identify novel copy variants. This would aid in interpretation and pathogenicity assignment.
    - Once the BAMs were constructed, they were pushed into parliament 2. It has taken 115,400 hours to compute the data.

- Data are just now starting to come from this project. CNV analyses were completed and showed some variation across chromosomes. The group tried to identify the number of CNVs across of all of the sample including deletions and duplications.
- Lessons learned include a common need for a cloud platform that helps process optimization and management that gives a standard model for operations.
- Next steps include finishing of the SV calling project and dissemination of results, as well as consideration of the AnVIL plan and how data can move from one environment to another without wasting too many resources.
- There can be orthogonal validation between the 4000 in eMERGEseq and the GWAS cohort.
- ACTION ITEM: The Genomics Workgroup will write up a paper on the lessons learned from eIII.
- The commons lessons are part of the harmonization paper.

- **The impact of return of unsolicited results on health care providers (HCPs) in eMERGE III: Preliminary findings** | Ingrid Holm *(BCH)*
  - The health care providers survey (HCP) supplement R01 was one of the first to identify the clinicians point of view as health care providers face different challenges as part of disseminating ROR.
  - The eMERGE network is one of the first to return unsolicited genomic results to providers.
  - The HCP RO1 aimed to survey participants' health care providers one month after receiving an unsolicited positive result, conduct qualitative interviews at four eMERGE sites, and identify approaches to returning unsolicited genomic sequencing results.
  - The group interviewed providers six months after results were given to determine their thoughts/actions/feelings on delivering a positive result and actionable items.
  - The group also interviewed providers when negative results were given to assess any actions in delivering these results, as well as any feedback in how this may have affected clinical care.
  - There were 306 enrollment invitations sent to HCPs with a 38% response rate for email, 8% response rate for mail with a29% overall response rate.
  - About 80% of respondents are in primary care and practice in a hospital (41%) or outpatient (44%) setting.
  - Generally, initial reactions to receiving genomic result were positive (33.3% strongly agree) and found it to be informative (35.7% as strongly agree).
  - Most (77%) HCPs planned to spend less than 15 minutes with patients discussing the results.
  - Most patients were referred to a specialist (55%) and 41.2% were referred to geneticists for testing.
  - Most HCPs found that the genetic testing was important to the patient's healthcare (46% strongly agree). However, 14% of participants did not find the results useful and 25% of the patients wanted a letter containing results.
  - Statements from HCPs showed that while the idea of genetic testing is exciting, it is difficult to recommend it when they are not sure who will be providing the counseling or follow-up.
- **EHRI Workgroup Science Presentation: Hazard analysis of CDS |** Sandy Aronson (Harvard) & Casey Overby Taylor *(Geisinger/JHU)*
  - Hazard analysis must be completed before introducing a new medical device into clinical care. A hazard analysis identifies possible issues with the device, classifying potential hazards, identifying mitigations, and assessing whether the overall risk profile of the device is acceptable..
  - As eMERGE sites move more towards CDS of electronics, a new working group has been introduced to perform the hazard analysis. The scope of the analysis included focusing on two types of genomic decision support:

        - The first type of genetic decision support involves a geneticist constantly signing out cases and providing clinicians with knowledge updates
        - Second scenario is PGx alert. PGx status may be known and the CDS is triggered when a provider needs ordering guidance.
    - Because patient PGx status is not known, using an ancillary genomic status approach is useful.
    - During the workgroup breakout session, 25 hazards were identified, including inappropriate alert firing context (nine), technical issues (eight), user experience problems that included issues that were disseminated to the clinician but not the patient (five), and knowledge management that included issues that were disseminated to the laboratory, not the clinician/patient (three).
    - A risk index table was created measuring occurrence by severity and then sorted by type. Although there are challenges in standardizing hazard analysis, it can provide great value in implementation. Any group that is implementing the hazard analysis must take into account local variation in practice that might not be recognized in a generic hazard analysis. It would be helpful to have examples that can help assisting how to predict hazards and label appropriately.
- **Clinical Annotations Workgroup Science Presentation: Incidental Findings |** Adam Gordon *(Northwestern)*
    - For Adam Gordon's PGRNseq project (NT179), an incidental finding is a L/LP variant on the consensus actionable list, and unrelated to site-submitted participant indication. A finding is still incidental if the participant is discovered to have a relevant but unsubmitted indication. If a person had no indication, then there are no incidental findings, they are just findings.
    - There is difficulty in harmonizing indications, and this needs to be clarified in future publications.
    - Adam is working on the incidental findings committee for ACMG so the eMERGE incidental findings are contributing to the ACMG gene list. There was about 3.02% positive variants that had Incidental Findings (IF). By site, rates are consistent at about 2%, with the exception of Geisinger which showed 10% IF.
    - Incidental finding rates by self-reported race and ethnicity showed that certain genes may be more typical in White populations and Ashkenazi Jews. This may be due to founder effects.
    - Cancer was the highest disease domain to find a likely pathogenic and pathogenic variant. The pathogenic variant result was the second most common finding. However, LP BRCA2 and LP LDR2 results were also the most commonly found incidental findings.
    - The most commonly found individual findings were BRCA2 pathogenic gene in Ashkenazi Jews (1.52% IF%), and the MYBPC3 variant that was found in 13 individuals (11 Asian, 2 Unknown). LDLR had eight CNV incidental findings which was the most common.
    - RYR1 is a founder allele and common in the Midwest. Investigators would be interested to see if RYR1 is more common in midwestern sites.
    - ATM is not on the consensus list and not returned by all sites.
    - ACTION ITEM: Investigators should contact Adam Gordon (adam.gordon@northwestern.edu) by the end of July if they are interested in collaborating on the PGRNseq paper.
- **Closing remarks |** Rex Chisholm *(SC Chair, Northwestern)*
    - Rex would like to thank Gail and David for hosting the SC Meeting in Seattle and organizing the Harbor cruise. Thank you to the CC for the support. The Network is pleased with how smoothly the meeting went.

Meeting Action Items

- **Coordinating Center**
    - The CC will work with the Clinical Annotation workgroup to create penetrance analysis forms for participants with results not returned.

    - The CC will add the break down of returned results, non-responders, and non-consented individuals regarding Outcomes & Penetrance analysis topic to the August 2019 PI call for discussion.
    - The CC will create a rough estimate about how many results we expect to obtain for 6 month and 12 month Outcomes
    - The CC will circulate the ROR tracker to the Network for reference.
- **ROR Workgroup**
    - The ROR/ELSI workgroup will discuss common themes among sites regarding providing reclassification information to participants.
- **Clinical Annotation Workgroup**
    - The Clinical Annotation workgroup and CSGs will provide details of the flow from a laboratory standpoint for VUS's that are reclassified.
    - Investigators should contact Adam Gordon (adam.gordon@northwestern.edu) by the end of July if they are interested in collaborating on the PGRNseq paper.
- **CSG Operations Group**
    - BCM and LMM will share variant list in order to determine the quantity of P/LP variants for a given disease and help to prioritize efforts.
- **Network**
    - Investigators with questions or samples for validation regarding the somatic mutation project, NT171: *Possible somatic mutation in targeted sequence data,* should contact Ken Kaufman (kenneth.kaufman@cchmc.org) and Paul Gecaine (paul.gecaine@cchmc.org).
- **Genomics**
    - The Genomics Workgroup will write up a paper on the lessons learned from eIII. The commons lessons are part of the harmonization paper.
- **EHRI**
    - Investigators interested in joining the FHIR Integration EHRI Subgroup should contact Michelle Stone (timoethia.m.stone@vumc.org) at the CC.