GEDI Strategic Planning – Extramural Workgroup Feedback

Contents

Gustavo Turecki

1. **What are the most important areas of research that NIDA should support in the next 5 years?**

In the context of GEDI, and considering all the knowledge produced recently understanding the regulation of the genome through epigenetic mechanisms as a function of environmental factors, NIDA should strongly support research investigating epigenetic mechanisms of addictions. Studies should explore different epigenetic mechanisms, investigate temporal dimensions and avenues for intervention (i.e, epigenetic mark modification).

Considering that the relevance of studies focusing on peripheral tissues to infer CNS processes remains unclear, before investing heavily on large epidemiological studies investigating epigenetic marks, NIDA should gain better insight into their ability to provide useful data. To this end, it will be useful to conduct extensive animal and human studies investigating both peripheral and CNS tissues obtained from the same individuals as a function to exposure to drugs of abuse. These studies should investigate different epigenetic marks and should be cell-type specific.

NIDA has significantly invested over the last decade on animal studies investigating epigenetic mechanisms of addiction. These studies have produced a wealth of interesting data. Unfortunately we know little about their relevance to humans. NIDA should now invest in the translation of these findings to humans. As such, postmortem studies investigating epigenetic mechanisms should be supported.

There is current excitement over the potential role of circulating small non-coding RNA as inter-cellular signals. Considering their great potential as biomarkers of psychiatric and addictions phenotypes, they should be better investigated.

2. **What resources are needed to achieve these goals?**

Although the epigenomics roadmap, the IHEC, and the ENCODE project generated tremendously useful data, we need to continue generating reference maps from different regions of the brain and at cell-level resolution.

Biobanks of both animal (including non-human primates) and human tissues are essential to move forward. Access to high quality human postmortem brain tissue characterized for addictions phenotypes. Furthermore, increased sequencing capacity and increased bioinformatics/computational resources to concomitantly analyze the effects of multiple marks.

3. **What benchmarks can be measured to track progress?**

Consistency of findings between different labs and across animal models and related human phenotypes.

4. **What training needs should be addressed?**

Increased need in bioinformatics training.

5. **What technologies and innovations can we leverage from other fields?**

   Advances in targeted epigenetic modification through CRISPR and other editing systems coupled with optogenetics and development of more efficient vector systems. Such technologies should help the community gain better insight, among other things, into new treatment avenues.

Eric Johnson

1. **What are the most important areas of research that NIDA should support in the next 5 years?**
    a. One of the keys to success of gene discovery in other fields is very large sample sizes (tens of thousands to hundreds of thousands). One of the most important, low marginal cost, investments for this effort would be **genotyping the many existing samples** that lack funding to genotype.
    b. Great strides have been made in understanding epigenetics in human brain tissue among non-addicted "normal" decedents. These data provide very import tools to understanding potential function of genes and variants specifically in tissue most relevant to addiction. A critical gap in the epigenetics of addiction is the lack of **large-scale genome-wide comparison of epigenetics in brain tissue between those addicted to drugs and non-using controls**. Adequately powder agnostic genome-wide tests of gene expression and methylation coupled with genome-wide genotyping among cases of addiction and non-using controls would provide data are critical to understanding differential expression and methylation specific to addiction that can not be obtained from existing resources. Additionally there is need for such data across ancestry groups, as there are substantial differences from what we know so far and a substantial under representation of groups that are not of European descent. Publically available resources of such epigenetic information would provide both the ability to "look up" results from GWAS and other genetic studies to assess potential gene function relevant for addiction, but also nominate variants for independent tests of association based on their association with differential expression or methylation. Epigenetic data will provide a view on genetics of differential brain function complementary to brain imaging studies, providing data on one set of mechanisms that could help explain observed differences in brain function between those addicted to drugs and those who are not.
    c. Refining phenotypes is one key to genetic discovery. Many existing phenotypes used in addiction genetics are highly heterogeneous, and are often compared heterogeneous controls. The most success in genetic epidemiology has been with well-measured, homogeneous, extreme phenotypes. **Development of biomarkers of addiction** may substantially advance the field in this regard. New technologies such as broad-spectrum metabolomics provide novel opportunities within the field of addiction to identify biological systems perturbed by addiction or cessation of use. Metabolomic profiles that distinguish those addicted and those who are not, or treatment responders and nonresponders could provide novel phenotypes for genetic study, producing insights into biological systems important to these outcomes as well as biomarkers for study of causes of individual differences in those systems.
    d. **Broad sharing of environmental risk factor and phenotype data needs to be required along side current requirements for sharing omics data**. This is the only way to get very large harmonized data sets for testing and replicating measured GxE interactions. The experience with sharing omics data suggests that enforced data sharing encourages actual collaboration too.

e. Although environmental interventions have had great success in lowering incidence rates in the U.S., acquisition among drug users continues here and much more so abroad. Based on *in vitro* work it is estimated that fifty percent of variability in HIV-1 susceptibility is attributable to host genetics. Indeed, mechanisms underlying the only genetic variant conclusively associated with HIV acquisition, a deletion in *CCR5*, gave rise to maraviroc, an antiretroviral drug. Thus identifying such genetic factors holds great promise in advancing our understanding HIV pathogenesis and providing targets drug development. Similarly, **basic research across the spectrum of genomics and other omics can provide unique insights on living with HIV among drug users**. Of particularly interest is the interaction between ongoing drug use and progression of HIV in the context of variability in adherence to HAART or its complete absence.
f. Position NIDA to participate in the Precision Medicine initiative. In particular, it may be useful **bring discovery science tools (e.g., Omics) to real world treatment settings using large numbers of patients to focus discovery on clinical outcomes specifically** (e.g., treatment response), rather than beginning with the disease and then translating those discoveries, hoping the same biological mechanisms for addiction are those that are key to recovery.
g. **Leverage the ABCD Study biospecimens for linking omics to imaging**.

2. What resources are needed to achieve these goals?

a. Funding is the most obvious resource across all these goals, including targeted RFAs
b. To generate the broad sharing of environmental and phenotype data a change in the requirements of dbGaP deposits needs to be put in place. It may also be necessary to create a complementary data repository for non-omics environmental data.

3. What benchmarks can be measured to track progress?

a. The number of new samples genotyped under the NIDA existing samples (Smokescreen) project.
b. Tracking number and success of new awards addressing each targeted area
c. Tracking the impact of data sharing through citation counts for the shared data sets
d. Count the number of newly shared data sets and resources made available to the research community for each targeted area
e. Count the number of new, replicated genetic discoveries

4. What training needs should be addressed?

a. Bioinformatics to integrate data across domains and leverage publically available resources effectively

5. What technologies and innovations can we leverage from other fields?

a. Metabolomics
b. Wearable sensors

Danielle Dick

1. **What are the most important areas of research that NIDA should support in the next 5 years?**
   - Continued genotyping of drug dependence samples in order to advance gene identification efforts parallel to what has been achieved for schizophrenia.
   - Continued attention to the heterogeneity of pathways of risk for substance dependence, as this piece will likely be critical to understanding basic etiology, identifying genes, etc.
   - New mechanisms to enhance connections between basic etiological researchers and prevention/intervention. How are findings from basic etiology being applied? There are not enough bridges or incentives to create these interdisciplinary connections which could have tremendous benefit to affected individuals.

2. **What resources are needed to achieve these goals?**
   - Funding mechanisms in place to explicitly support these areas of research

3. **What benchmarks can be measured to track progress?**

4. **What training needs should be addressed?**

   There has been tremendous emphasis on interdisciplinary training programs. I am no longer sure this is the best way to go. It can produce students with more breadth than depth. Another route is to encourage collaborations between individuals from different fields (inter-individual interdisciplinary collaborations rather than intra-individuals interdisciplinary training). I think this would be more beneficial as it would create breadth by bringing together individuals with depth of experience/training in a given area.

5. **What technologies and innovations can we leverage from other fields?**

   Success in identifying genes involved in schizophrenia definitely suggests a strategy to be emulated in gene identification for other outcomes.

Jane Costello

What are the most important areas of research that NIDA should support in the next 5 years?

1. As the largest funding agency for drug-related GED research, NIDA should keep an eye open for important projects that are too big for other funders.
2. Collecting molecular genetic information is relatively cheap nowadays. Collecting phenotypic and environmental data and doing so in a developmentally informative way, is likely to be much more expensive. But it is also much more difficult.
3. There are things that the USA, with its lack of universal health care and nation-wide records, is particularly poorly placed to study. Examples are very rare disorders, and envirotypes that are not recorded systematically. Perhaps it would make more sense for the US to put its resources into research in countries that have the necessary infrastructure.
4. Perhaps as important are the research studies that NIDA should *not* support. These include underpowered genetic studies unless they are specifically designed to be included with others in meta- or mega-analyses.
5. There is a lot of excitement nowadays about "big data" quarried from administrative data sets like medical records. Before getting too excited about their use, it is important for NIDA to think about their limitations, especially in relation to drug use, where treatment is scarce and "cases" tend not to be representative of the majority of individuals with drug problems.

.

What resources are needed to achieve these goals?

1. Supporting data repositories like DBGaP is a good use of federal funds.
2. A lot of time and money has been put into creating libraries of common phenotype/envirotype measures. These certainly should be supported.
3. But these seem to press for simplicity and brevity. We also need to ensure that more complex and varied constructs are included in studies. Perhaps everyone using some of the PHENX-type measures could be encouraged to examine at least one area in more detail. Otherwise we shall end up, as Konrad Lorenz described scientists, knowing more and more about less and less.

What benchmarks can be measured to track progress?

1. Prediction is the best benchmark**.** Longitudinal research is important.
2. From NIDA's viewpoint, peer-refereed publications from funded grants are the best guarantee of progress.

What training needs should be addressed?

3. Geneticists need to know something about the substantive area they study, and clinical or epidemiologic scientists need to know some genetics: at least enough to talk to one-another. Both need to understand research design.

What technologies and innovations can we leverage from other fields?

1. Statistical methods for protecting data confidentiality.
2. Methods for handling missing data.
3. Methods for modeling complex data including methods for causal inferences.
4. Methods for secure data dissemination.

5. Methods for data federation or integration.

William Iacono

1. **Research to prioritize**
    a. Longitudinal studies of youth using genetically informative designs that make possible better evaluation of causal effects related to development and environment
        i. Twin studies using co-twin control design to identify effects of specific types of nonshared E that are developmentally relevant
            1. E.G. See Irons et al. (Irons, Iacono, & McGue, 2015) which showed that early onset drinking and intoxication (before age 14) were causally linked to age 24 SUD-related outcomes
        ii. Can be used to identify biomarkers that reflect genetic risk vs. consequence of use
    b. Establishment of two kinds of large molecular genetic databases
        i. Because sample size trumps everything else, get the largest possible samples using simple measures of phenotype (like CPG, answer to Q: "do you use marijuana?")
            1. Include whatever environmental measures exist (e.g., self-rating of available social support, life stress)
            2. Use these samples in discovery-based molecular genetic analyses
        ii. Create smaller samples comprising (when possible) longitudinal data with carefully defined phenotypes and E measures to lessen multiple testing burden and improve signal to noise ratio in quality of measures used
            1. Use these samples to follow-up confirmed findings derived from analyses of the large data base to flesh out the nature of confirmed main effects (for the phenotype), genetic risk score effects, and possible GxE effects
            2. Determine the relevance of the findings to different stages of development
        iii. Include sequenced samples whenever possible
    c. Lifespan developmental approach
        i. Much current research is on youth
        ii. Long term effects of substance use, including effects of early use and continued use on physical health and older age adjustment not well researched
        iii. Beyond treatment effects, need to understand better what G and E factors lead to recovery/desistence in adults
    d. In states that are legalizing marijuana:
        i. Continuation of ongoing longitudinal studies
        ii. Encouragement of new studies focused on youth
        iii. Encourage development of parallel investigation in "comparable" states not legalizing marijuana
    e. Molecular genetic GxExD studies in humans where good animal model exists for the GxExD effects with the animal component carried out in parallel

2. **Needed resources**
    a. Administrative support to manage these types of large, complex databases
    b. Require more data sharing and contribution of well documented data to repository
    c. Funding to genotype existing longitudinal samples with quality measures of phenotypes and E
    d. Possible to bring back dual R01s where an integrated proposal is submitted by two (or more) investigators at different universities to enable carrying out research on the scale needed to adequately power studies
    e. Funding support for efforts to use repositories

3. **Benchmarks to track progress**
    a. Monthly pace with which repository samples grow in size
    b. Number of investigators contributing samples to repository
    c. Once samples grow & exceed a certain size (e.g., 100K), number of confirmed findings reported for molecular genetic variants
    d. Number of confirmed main effect findings leading to replicated GxE findings that are linked to developmental stage

4. **Training needs**
    a. More training in genetic methods for those doing longitudinal studies of well phenotyped samples to facilitate follow-up of confirmed effects in their samples
    b. Training in how to develop collaborative/cooperative scientific initiatives in addition to current emphasis on competition
    c. Training in how to leverage digital technology to obtain large samples and
        i. Phenotype them for SUD behaviors
        ii. Identify exposure to risky environments
        iii. Obtain DNA

5. **Technologies to leverage**
    a. Digital use of social media and web site use
    b. Mobile technology that tracks same, and includes geographic location tracking technology
    c. Wearable sensors linked to above

Irons, D. E., Iacono, W. G., & McGue, M. (2015). Tests of the effects of adolescent early alcohol exposures on adult outcomes. *Addiction, 110*(2), 269-278. doi: 10.1111/add.12747

Kenneth Kendler

1. **What are the most important areas of research that NIDA should support in the next 5 years?**
   To fulfill the goals of the GEDI initiative, I would see NIDA as having two somewhat contradictory goals:
   a. Collect really large sample sizes to detect molecular genetic variants. But this approach often means very shallow phenotyping as a cost-saving measure. So once variants are detected, we would expect to have p little in the way of phenotypes, environmental exposures or developmental data to understand the action of those variants. Scientifically, this is a "shallow" victory.
   b. To collect informative epidemiological or high-risk samples with prospective data on a rich set of risk factors, as well as putative mediators and moderators. Ideally, some of these data sets would have good measures of latent risk indices (e.g. from twin or adoption data), others molecular markers (currently GWAS but allow progression to exome or whole genome sequencing if that is merited) and some with both. This will allow modeling at several levels of gene x environment interaction and correlation in a developmental context. But can these collections be big enough to be well powered for molecular variants?
   c. How to resolve this problem? Not entirely clear. Some thoughts: i) cheap but deeper phenotyping – perhaps by web or cell-phone, ii) focus on aggregate molecular measures – networks, polygene scores, GCTA, LD-mapping etc. iii) just buckle down perhaps with collaboration with NIAAA and fund a few studies really large enough to do both? iv) use large samples with good data from registries or other high quality phenotyping to which genotyping could be added at modest cost. Do such studies exist with large enough sample sizes? Assemble many smaller studies – skeptical as very hard to get comparable measures?

2. **What resources are needed to achieve these goals?**
   - Funding for data collection with adequate power and genotyping/sequencing will be expensive. In Scandinavian countries, there is an option to try to add DNA collection to samples that have registry data, although ethical issues may be intractable.
   - It is well worth exploring methods for combining data across samples. But how to produce core common set of phenotypes and risk variables without stifling individual initiative and creativity is an important issue worth pondering.
   - What might be needed to develop high cost efficient ways of gathering both cross-sectional and longitudinal data on large numbers of subjects?

3. **What benchmarks can be measured to track progress?**
   Usual metrics - sample sizes, quality of data collection (e.g. cooperation rates), quality of genetic material.

4. **What training needs should be addressed?**
   Supporting both training and r-grants in methods development for the analysis of rich and informative developmental samples with g-e interaction and covariation.

5. **What technologies and innovations can we leverage from other fields?**

Developments in 1) various polygene methods – GCTA, LD-regression, Polygenic Risk scores, 2) advancements in classes of gene annotations (Encode, expression arrays etc) that can move beyond single variant markers – that may have very low power in the absence of really large samples. 3) developments in inexpensive data collection.

John Rice

1. **What are the most important areas of research that NIDA should support in the next 5 years?**
   - Gene-environment interplay
   - Phenotype refinement for genetic analysis
   - Combining information from multiple domains – genetics, imaging, gene expression and animal models for SUD.

2. **What resources are needed to achieve these goals?**
   - Controlled access, large-scale, harmonized databases of NIDA-related GWAS data
   - Individual level data with detailed genetic and phenotypic variables

3. **What benchmarks can be measured to track progress?**
   - Quantities of available data

4. **What training needs should be addressed?**
   - Support methods related grants to use existing data
   - Support post-doc training aimed at genetic methods and tool development
   - Support meetings/workshops that train scientists to use the above resources

5. **What technologies and innovations can we leverage from other fields?**
   - Expression databases
   - Methylation databases
   - Connectome database
   - iSPCs

Daniele Fallin

What are the most important areas of research that NIDA should support in the next 5 years?

- Integration of genetic risk information with psychological and sociological risk to inform prevention and intervention research
- Integration of omics technologies from “below the skin” to “above the skin”

What resources are needed to achieve these goals?

- Cohorts with appropriately timed biosampling, environmental assessment and phenotyping. This may require new cohorts!
- Bioinformatics infrastructure for sharing and computing

What benchmarks can be measured to track progress?

- For new data collection, the best benchmarks are the quality of data collected, not the findings. A rush to findings often occurs at the sacrifice of deep, rigorous data collection that can inform research for a longer time window. Cohort studies are often criticized for lack of papers in the data collection years. This is a mistake.
- What new work has been inspired by these findings?
- What new insights are being applied to prevention and intervention efforts based on these findings?

What training needs should be addressed?

- Cross-communication between disciplines
- Mental and behavioral health training among bioinformatics or geneticists
- Training in biological focus integrated with psycho-social

What technologies and innovations can we leverage from other fields?

- Brain imaging
- Exposome measures
- Sequencing and pattern recognition analyses

Laura Bierut – NACDA Council Feedback

From: Bierut, Laura [mailto:bierutl@psychiatry.wustl.edu]
Sent: Thursday, May 07, 2015 5:32 PM
To: Pollock, Jonathan (NIH/NIDA) [E]; Weinberg, Naimah (NIH/NIDA) [E]
Cc: Rutter, Joni (NIH/NIDA) [E]; Caulder, Mark (NIH/NIDA) [E]
Subject: Data sharing

Dear Jonathan and Naimah,

I am writing to follow up on the open council presentation that focused on strategic planning. This strategic plan is such an important process to guide NIDA in performing the strongest science possible. I reviewed the slides that Maureen Boyle presented and I wish to emphasize the importance of data sharing that was promoted by Philip Bourne. It is critical to all of science that data be shared with the scientific community. For example phenotypic and genotypic data sharing are now possible with dbGAP and dbGAP has been critical in the success of the discovery of genetic findings through GWAS studies across all the institutes.

My natural scientific focus is in the Gene-Environment-Development Interactions. I hope that this group takes the lead on sharing data with the scientific community. An important example will be the sharing of data on dbGAP from the GEDI initiative where Jane Costello and Bill Iacono where funded. Their projects are unique and it is important to get these data shared with the scientific community. Perhaps at the next NIDA Genetics Consortium meeting in June an update on the sharing from these projects can be given.

Best,
Laura

Eric Johnson – feedback from April 28 WebEx

1. The genome-wide association study (GWAS) method is the most successful approach to identifying replicable variant associations for complex diseases in humans to date: far more successful than candidate gene and linkage studies, both for addiction and for other diseases. For addiction the confirmed variant associations have been limited to nicotine and alcohol so far. It seems likely that heterogeneity in the phenotype (cases and controls), as well as relatively small sample sizes are the primary challenges to success for opioids, cocaine, and possibly marijuana. Thus I think we are still very much in a discovery phase of addiction genetics. A key question is: what can we do beyond increased sample size to facilitate discovery? Some of the answers may be to integrate epigenetics and other functional information with GWAS, as well as improve our phenotypes representation of the biology of addiction.

2. GWAS results have been criticized for the small amount of variance they explain. Explaining a large amount of the variance in addiction with a few variants would be wonderful. However, it seems extremely unlikely this will be the case for such complex diseases. Moreover, explained variance doesn't seem to be the central point of gene discovery. Rather, the point is to identify true associations that point to a biological mechanisms that enhance our understanding of etiology and lead to better treatment. Rs16969968 and nicotine dependence is a strong case in point. This variant explains less than 5% of the variance in ND. However, it's discovery lead to new mouse model experiments by Paul Kenny and Jerry Stitzel that informed us about etiology: demonstrating that the risk mutation increases risk of addiction by reducing the noxious effect of nicotine – removing the brakes on escalating self-administration. Li-Shiun Chen has also demonstrated a replicated interaction between smoking cessation (NRT/bupropion) and rs16969968 wherein high risk smokers benefit much more from treatment than do low risk smokers (number needed to treat being 4 and >1,000 respectively). Dr. Chen most recently showed that the high risk variant at rs16969968 is associated with a four-year delay in smoking cessation, adding significantly to these smokers' risk of lung cancer. Thus, I contend that R^2 is not a good proxy for the importance of a variant or a finding, it is the biology behind the association that is important.

3. Biomarkers of addiction is an important area to think about. Biomarkers may provide stronger, more useful phenotypes than self-reported symptoms given their lower bias and measurement error, as well as closer connection to biology. As a phenotype, the issue of whether the biomarker taps into the state of addiction or the trait that predisposes to addiction may not be strongly important. From the perspective of gene discovery, a biomarker phenotype could represent the current state/level of addiction and/or biological pathways that are perturbed by the state of addiction. Identifying genetic variants associated with this type of phenotype would provide clues to the underlying biology of addiction and potential germ line variation that predisposes to addiction even though the phenotype measures the current state. Genetic associations with the state of addiction may also provide direction for drug discovery.

4. Gene x environment interactions are very difficult to adequately test. A joint 2 degrees of freedom (df) genome-wide GxE approach has greater statistical power than a standard 1df GWAS to identify variants associated with the phenotype in the presence of some degree of GxE, and can identify statistically significant interactions. However, sample sizes for this approach must be large. An alternative is some version of a candidate gene approach. Unfortunately, candidate gene GxE studies have generally not worked unless the genetic variant has been demonstrated to be replicably associated with the phenotype. There are few such variants for addiction phenotypes. These variants should be tested for GxE in available samples. Additionally, a large-scale joint 2df meta-analysis with a harmonized environmental risk factor seems the best way forward. If the field can identify additional true variant – phenotype associations then a candidate gene approach would be feasible.