Science Culture & Accountability Plan Center for Genomic and Computational Biology
Duke University School of Medicine

An epidemic of mistrust of science exists because of a few instances of serious, highly publicized scientific misconduct and several examples of unreliable findings evidenced by "failures to replicate." In recent years, events at Duke and elsewhere have highlighted the ongoing need for explicit processes and mechanisms to manage scientific research data in a responsible manner. At the same time, most investigators have received little, if any, training or guidance in this area. GCB is committed to ensuring that policies and procedures are in place to maintain the highest level of professional conduct:

- Create and maintain a culture in which scientific results are critically reviewed.
- Demand accountability for data integrity.
- Voice concerns without hesitation.
- Establish mechanisms for addressing these concerns fairly and transparently.

We recognize this requires appropriate aptitude and active participation of all parties in the research mission, from faculty and trainees to our center director and to the many departments and centers we collaborate with.

Scientific rigor ensures the proper application of the scientific method using the highest standards in the field. We commit to following four general principles:

- Know where your data are.
- Know what has been done to acquire and modify your data.
- Make all efforts to ensure that data collection and analysis are at least unbiased, blinded if possible.
- Follow proper statistical procedures.

All center and associated departmental staff should understand why these principles exist and should be incentivized to follow the principles

Overview of Best Practices

- Insofar as possible given the nature of the research, best practices in scientific rigor, including statistics, should be followed.
- In recognition that no one size fits all, each research group and shared resource should establish its own specific plan for scientific accountability and scientific rigor, according to established standards of its field, integrating industry or other perspectives when appropriate.
- Record keeping should track all primary data and should provide a way to "audit" the data for each figure of each paper readily and transparently.
- All modifications of raw data should be performed on copies of the original data, if possible, and should be tracked, dated, and documented fully.
- The PI or core facility Director should avoid allowing his/her stresses about funding and publication to affect the attitudes, expectations, or behavior of the laboratory staff.

- Scientific accountability and scientific rigor should be a frequent discussion between PI or core facility Director and the laboratory staff, to establish a sense of common purpose and a shared goal to discover the truth.

Best practices in experimental design

- Employ systematic random sampling for data collection as appropriate, including selection of subjects and cells.
- Strive to eliminate bias in experimental procedures and analysis. If practical, experimenters should be blinded to treatment. The timing of experiments might be balanced to account for sources of bias over time (e.g., evolution of surgical skills, fatigue, change in personnel; test order effects, circadian rhythms in experimental animals).
- Use positive and negative controls.
- Use replicate samples (including both technical and biological replicates) for experimental groups, as appropriate.
- Use validated and/or well-characterized reagents (such as antibodies and pharmacological agents), or conduct full validation.
- Consider limitations of behavioral, animal, or cellular models including possible contributions of genetic background, age, and gender.
- Find a proper balance between increasing numbers of animals for replication and the goals of "replacement, reduction, and refinement" in animal research.

Best practices in data analysis and statistics

- Consult with our bioinformaticians from our shared resources both before and after data collection if analysis is needed.
- Determine sample size by pre-experiment power analyses when possible. Identify stopping points a priori to avoid testing to a foregone conclusion.
- Obtain and examine the raw data for any results provided by shared research cores for indicators of data quality and any evidence of mislabeled samples.
- Conduct a thorough characterization of experimental effects.
- Repeat experiments within the laboratory to reduce likelihood of statistical flukes.
- Use quantitative analysis when possible to put bounds on the size of effects.
- Use care in pooling data across experiments done at different times, multiple time points, or different experimental groups. Apply statistical tests for batch effects.
- Avoid data exclusion. If necessary, define and report objective procedures for dealing with attrition or other missing data and data exclusion. Unless there is a compelling, transparent reason to exclude data, include all runs of each experimental procedure. This applies to exclusion of individual points or complete data sets.
- Perform theoretically-correct analysis of data using appropriate statistics and sample sizes.
- Don't mislead with statistics. Take advantage of resources that provide professional statistical expertise such as GCB's Genomic Analysis shared resource.
    - Perform statistical tests to validate what is seen in the data rather than reveal effects that may be statistically significant but functionally non-significant.

- Select appropriate statistical tests, including testing statistical assumptions, such as adherence of data to or departure from a normal distribution.
- Control for multiple comparisons.
- Avoid "significance chasing" or "p-value hacking," such as interpreting the data in different ways so it passes the statistical test of significance, or analyzing different measures until finding one in which groups differ.

Best practices in data management

- The complete primary data should be retained, backed up, and protected against alterations.
- Alterations and modifications of the primary data should be performed on copies of the data whenever possible and should be tracked, dated, and described.
- Data notebooks should be open for viewing.
- Digital archives should be properly organized and labeled so each dataset has the equivalent of a "unique web address."
- Every figure of every paper should be cross-referenced with the location of the original data that contributed to the figure.
- We will continue to develop mechanisms to monitor provenance for any data that comes from GCB's core facilities.
- The level of information security should be appropriate for the material, especially for human subject protection and PHI.
- Data should be readily accessible to all data owners and available to outside investigators if necessary.

Best practices in publication

- Report full details on methods and experimental design, including technical and biological replicates, methods for randomization and blinding, and self-replication efforts.
- Report complete results of all analyses done as part of an experiment, including statistical controls for multiple comparisons and identification of pre- and post-hoc analyses. Methods sections should be too long, rather than too short.
- Avoid "rushing" findings into publication without a full investigation and proper self-replication.
- Target appropriate venues for publication. Avoid pressure to publish in the most glamorous journal at the expense of following the best practices for experimental design, data analysis, statistics, and publication. If a paper requires a long methods section or many figures to document the science thoroughly, do not try to compress it into a short format, no matter how "important" the results seem. Publish well-controlled negative, "uninteresting," or "not novel" results in appropriate venues.
- Resist the emerging trend in which the peer review process demands additional experiments on an abbreviated timeline with the associated pressure for results to be interpreted to conform to previously-reached conclusions.

Creating a functional and proactive scientific culture

- We will strive to create a culture of getting it right, with the expectation of open conversation and a lack of retribution for calling results or procedures into question either within the group, or to the lab head in confidence.
- All Center and associated staff in GCB should know that they may bring any and all concerns to the attention of the center Director, Chair of their home department, or Ombudsperson in confidence without fear of retaliation or retribution. Staff should also be aware of the Duke Integrity Line to report concerns anonymously.
- PIs must minimize incentives or pressures (or the appearance thereof) that drive their staff to perform for reasons other than pursuit of truth. It is critical to avoid the real danger that staff will respond to the PI's concerns about academic promotions, choice of publication venue, or competition with other labs.
- Issues of proper scientific conduct and scientific rigor should be discussed regularly with laboratory personnel in both private and group meetings.
- PIs and core facility Directors should be involved in laboratory procedures, should oversee some of the actual experimental work, and should "know" how things are done in their laboratory.
- Meetings with staff should include inspection of some primary data and discussion of detailed analysis procedures, as well as discussion of final publication-style figures.

Concrete steps to be taken by the Center

1. We will continue to discuss proper scientific conduct at all levels: faculty meetings, lab meetings, and courses.
2. We will strive to create a culture where we talk about the incentives for poor conduct openly, and try to address the pressures that create those incentives.
3. We will endeavor to create a culture where people are evaluated on the basis of what they have done rather than metrics that may be weakly correlated with accomplishment.
4. Each shared resource should develop a "Data Management Standard Operating Procedure (SOP)" that will provide specific guidelines for data acquisition, storage, and transparency. The SOP should cover four basic components of data management:
    a. How is the data collected and stored?
    b. How are notes taken and stored?
    c. How is analysis done, tracked and, if intermediate steps are saved, stored?
    d. How are figures made and linked to both the analysis steps and the original data?
5. Each faculty member's home department is responsible for ensuring that they develop a "Data Management Standard Operating Procedure (SOP)" as outlined immediately above. We will coordinate with the home departments of GCB's faculty to ensure that they complete training through in-person or web-based modules as required by the Duke SOM ASIST office.
6. The laboratory's Data Management SOP should be discussed with the center Director (for shared resources) or department Chair (for faculty) when it has been completed, and compliance measurement will be a topic in the annual 1-on-1 meetings.
7. All research staff in all laboratories must read the Center's Action Plan and the laboratory's Data Management SOP and sign an affirmation that they have done so.
8. GCB will offer workshops to the research community across campus through its GCB Academy series on the subject of data integrity and reproducibility in research.

9. Researchers will be encouraged to consult with GCB's Genomic Analysis shared resource about correct methods for data analysis. Faculty who are expert in data analysis are available to advise students, postdocs, and faculty on how to analyze their data.
10. The Center's IT staff will work with each laboratory and our shared resources to implement their chosen procedures for data storage, backup, and tracking.