ENCODE Consortia Data Release, Data Use, and Publication Policies

Draft 9-11-08

Summary

NHGRI has designated the Encyclopedia of DNA Elements (ENCODE) and model organism ENCODE (modENCODE) Projects as community resource projects to accelerate access to and use of the data by the entire scientific community, while ensuring the widespread availability of all information that is generated by the ENCODE and modENCODE Projects. ENCODE/modENCODE resource producers will release data, as soon as they have been verified and prior to publication, to public databases. At the same time, until the data are published upon in a peer-reviewed journal, NHGRI asks resource users to consider them to be unpublished and to follow standard scientific etiquette regarding the use of unpublished data. Specifically, resource users are asked to respect the ability of the producers to publish an initial analysis of the data they have generated in a timely manner. To facilitate this compromise between unrestricted use of the data and unavailability of the data until publication, NHGRI will promote observation of a 9-month period during which resource users may freely use the ENCODE/modENCODE data to design and carry out their own research programs, but not to submit publications that use unpublished ENCODE/modENCODE data without prior consent. (NHGRI recognizes that there may be some exceptions to this blanket request; examples are discussed more fully below.) After the expiration of this 9-month period or publication of the data (whichever comes first), resource users should continue to properly acknowledge the ENCODE or modENCODE Project and resource producer(s) as the source of the data in any publication.

NHGRI recognizes that the examples below do not address all potential uses of ENCODE/modENCODE data, nor are all of the terms in this document (e.g., locus) precisely defined. It is not possible to be exhaustive in anticipating all of the ways that ENCODE/modENCODE data may be used and all situations that might arise. Therefore, this policy document emphasizes the goals of the proposed data release policy and requests that both resource users and resource producers attempt to behave within the spirit of the policy. In specific situations, resource users who have questions about the appropriate use of ENCODE/modENCODE data are encouraged to contact the resource producers directly or NHGRI staff (contact information available at http://www.genome.gov/ENCODE).

Background

In 2007, NHGRI funded two research consortia for the purpose of comprehensive discovery of all sequence-based functional elements in the human genome through the Encyclopedia of DNA Elements (ENCODE) program, and in the genomes of *C. elegans* and *D. melanogaster* through the model organism ENCODE (modENCODE) program. A single policy on data release, data use, and publication has been developed by NHGRI for these programs. The policy (below) was developed in consultation with the members of the two research consortia, the members of the ENCODE/modENCODE External Consultants Panel, and members of the research community. For the purpose

of simplicity, this policy will be referred to as the "ENCODE Consortia data release policy."

Data Release Principles and Standards

The NHGRI is committed to the principle of rapid data release to the scientific community. This principle was initially implemented during the Human Genome Project and has been recognized as one of the most effective ways of promoting the use of the human genome sequence and subsequent genomic data sets to advance scientific knowledge and application to human health. At a meeting in Ft. Lauderdale, FL, co-sponsored by the Wellcome Trust and NHGRI in January 2003, the principles of rapid release of genomic sequence data by large-scale producers were reaffirmed. The attendees furthermore strongly recommended applying the practice to other types of data produced by "community resource projects," defined as research projects specifically devised and implemented to create a set of data, reagents, or other material whose primary utility is as a resource for the broad scientific community (recognizing that different issues, particularly with respect to data verification and validation, would be involved in the development of appropriate release practices for different types of data). A second important conclusion of the Ft. Lauderdale meeting was that the success and utility of community resource projects is based on mutual and independent responsibilities for the production and use of the resource by the resource producers, the resource users, and the funding agencies (http://www.genome.gov/Pages/Research/WellcomeReport0303.pdf).

The Ft. Lauderdale principles call upon the resource producers to release data rapidly, prior to publication, once they have been established to be reproducible (verified), even if the data have not been sampled to determine if there is biochemical or biological significance (validated). The resource users are, in turn, called upon to recognize the source of the data and to respect the legitimate interest of the resource producers to publish an initial report of their work. The resource producers also are expected to publish on their data in a timely manner in order to avoid undue delay in the ability of users to publish results of experiments done with data obtained from the resource. Consistent with the ethical use of unpublished data, resource users should not publish a paper using unpublished data from a community resource project that will diminish the ability of the resource producers to publish an initial report of their findings in a peer-reviewed journal. Finally, the funding agencies are called upon to recognize the need for funding to support the analysis and dissemination of the data. The NHGRI has based its data release policies for several projects, including the ENCODE Pilot Project, the International HapMap Project, the Mammalian Gene Collection, and genome sequencing projects, on these principles.

ENCODE Consortia Data Release Policy

The NHGRI has designated ENCODE and modENCODE as community resource projects. The pipeline for data production and analysis of most components of these two programs involves at least of two levels of data. The first is the primary dataset as obtained directly from the experimental platform. The second is an interpreted dataset

that is the result of the initial analysis of the primary dataset. In practice, the ENCODE Consortia data release policy will be affected by three important considerations: (1) several different primary data types will be generated, as a variety of experimental approaches are being taken in the project to identify functional sequence elements; (2) the criteria for verification for each data type will vary and need to be taken into account in developing appropriate data release standards for each data type; and (3) each primary data set will be further processed to produce an interpreted data set of functional elements. This interpreted data set will be subjected to an additional level of experimentation to validate the identified functional elements.

In this data release policy, NHGRI makes a distinction between data verification and data validation. "Data verification" is understood to refer to assessing the reproducibility of an experiment. A practical definition for verification is when a dataset has been demonstrated to be of sufficient quality to merit follow-up experiments. "Data validation," or "biochemical validation," is understood to refer to confirmation, often of a subset of the initial verified data, of the biochemical event by other, independent methods. Data validation is a measure of the accuracy of the verified dataset. Further biological characterization of the putative functional elements also will be conducted by some of the experimental groups.

The data release policy is intended to ensure early deposition of data in public databases, and defines "early" to mean as soon as data are verified – even if they have not yet been validated. For each data type, the ENCODE and modENCODE Consortia will identify a minimal verification standard necessary for public release; those standards will be described on the ENCODE and modENCODE websites (http://www.genome.gov/ENCODE and http://www.genome.gov/modENCODE, respectively). Based on these standards, the resource producers will release primary data along with an initial interpretation, in the form of genome features, to the appropriate public databases as soon as the data are verified. The Consortia members will also identify validation standards that will be applied in subsequent analyses of the data or with additional experimentation where appropriate. Completed validation datasets, which will be defined by each ENCODE/modENCODE research group for each data type that it generates, will be released to the appropriate public databases in a timely manner consistent with the policy for verified data. When possible, estimates of the false positive and false negative rates for the particular experimental approach based on data validation studies will be included in the data releases as a measure of data quality. All data will be deposited to public databases, such as GenBank or ENCODE/modENCODE Consortia databases, such as the Data Coordination Centers (DCCs) and these pre-publication data will be available for all to use (see Appendix A).

ENCODE Publication Policy

In October 2004 the ENCODE Consortium published an initial manuscript, a so-called "marker paper," describing the goals of the project, its data release practices, and its publication policies. The modENCODE Consortium intends to publish a similar "marker paper" describing the project in the fall of 2008. One purpose of the marker papers is to provide a literature reference to those who use the data produced by the consortia.

Updated information about plans for the ENCODE and modENCODE Projects are regularly provided on the NHGRI's ENCODE website (http://www.genome.gov/ENCODE) as well as on the public websites developed by the Consortium.

To balance the interests of all stakeholders, resource users are asked to respect the ability of the resource producers to publish an initial analysis of their own data, while the resource producers are asked to do so in a timely manner so as not to slow the progress of science. NHGRI defines "timely," for the purpose of the ENCODE Consortia data release policy, as an initial period of nine months after the **release** of the data into public databases. This will provide time for the resource producers to have a protected opportunity to **publish** initial analyses of the data they have generated. The nine-month period will be established for each submitted dataset by the creation of a timestamp at the time the data are posted for public access and will apply to all data types, including primary, interpreted (as defined above), validation, and biological characterization data. The timestamp will be maintained by the respective DCC where the dataset will be publicly displayed. During this time period, resource users may work without restriction to analyze, and otherwise use the data in their own work, but they are requested not to submit their analyses or conclusions for publication. The publication moratorium by resources users ends either at the expiration of the nine-month protected period or when the data have been published, whichever is shorter (see below for more information on acceptable use of ENCODE/modENCODE data). During the protected period, resource users and resource producers are encouraged to communicate about their activities for purposes of either establishing collaborations or organizing simultaneous publications.

The community can expect that the individual research groups in the ENCODE and modENCODE Consortia will publish the results of their own efforts in independent publications. In these individual papers, Consortia participants will not be restricted to describing the methods developed for the project, but can and should expand into describing biological insights that arise from their analyses. Consortia research groups who are doing analyses that are similar to each other are encouraged, but not required, to establish collaborations and/or coordinate the publication of research results. To facilitate comparison of data between different groups involved in ENCODE and modENCODE, all publications by Consortia members should, to the extent possible, include data on the common reference set of reagents that has been agreed upon by the Consortia; these include common cell lines or common antibodies. Information about the reference reagents is available on the Consortia web sites and all members of the scientific community are encouraged to include data on these reagents for comparison to ENCODE/modENCODE data.

At the appropriate time, each Consortium expects to publish one or more "synthesis" paper(s) in a peer-reviewed journal, which will highlight the conclusions of integrative analyses of the ENCODE/modENCODE data. In advance of these publications, each Consortium will establish data freezes consisting of a snapshot of all datasets that have been verified by the freeze date in order to provide a uniform platform on which all research groups can base their analyses. Each paper will synthesize the information

from all aspects of the project in order to bring out the key biological insights gained from the project. Every contributor will be indexed as an author in the PubMed record. The Consortia also plan to publish sets of “companion” papers, which will be series of papers published by subgroups of each Consortium containing detailed analyses of individual datasets and combinations of datasets, that were not a part of the ”synthesis” paper. Ideally the “companion” papers will appear together, and either be published in the same edition of the journal in which the “synthesis” paper appears, or be published in another journal at roughly the same time as the “synthesis” paper. Investigators who do not participate in the Consortia are welcome to include their work in the companion paper set, and are encouraged to communicate with the Consortia to arrange for this cooperation. It is possible that the two Consortia will also write one or more papers on integrated analyses across both projects, but there are no formal plans to do so.

Using ENCODE/modENCODE Project Data

Investigators outside of the ENCODE and modENCODE Consortia are free to use the ENCODE/modENCODE Consortium data, either en masse or as specific subsets, but are asked to follow the guidelines elaborated in this ENCODE Consortia data release policy. Specifically, users of Consortia data (including Consortia members) should be aware of the publication status of the data they use and treat them accordingly. NHGRI will monitor the use of unpublished ENCODE/modENCODE data and will modify this policy if there is evidence of significant misuse of the data by resource users.

For use of unpublished data within the initial nine-month period: Resource users who perform analyses of unpublished data from the ENCODE and modENCODE Consortia and are interested in publishing a report before the expiration of the nine-month protected period should discuss their plans to use the pre-publication data with the resource producer(s) and should obtain their consent before using unpublished data in their individual publications (obtaining and giving consent to use unpublished data are considered to be accepted scientific etiquette). Collaborations are encouraged; however, the Consortia members are not required to collaborate with any outside investigator(s). The time stamp-related moratorium on publication is expected to apply to **submission** of manuscripts for publication by resource users. Resource users are expected to acknowledge the resource producers; reference the ENCODE (2007 publication of the results of the pilot project) or modENCODE (2008 marker paper) Consortia papers, as relevant; the funding organization(s) that supported the work; and the respective DCC in all resulting oral or written presentations, disclosures, or publications of the analyses. All investigators, through their roles as journal and grant reviewers, and journal editors, are asked to help to maintain a high standard of respect for the scientific contribution of the resource producers.

For use of published data, or unpublished data for which the nine-month time stamp has expired: Following expiration of the protected publication period, any investigator may submit manuscripts without restriction, including integrated analyses using data from multiple research groups. However, for unpublished data, as above, resource users are expected to acknowledge the resource producers; reference the ENCODE (2007 publication of the results of the pilot project) or modENCODE (2008 marker paper)

Consortia papers, as relevant; the funding organization(s) that supported the work; and the respective DCC in all resulting oral or written presentations, disclosures, or publications of the analyses. For examples of appropriate use of the ENCODE/modENCODE data, see Appendix B.

Consortium members will not have privileged access to data from other members of the Consortium. Rather, all data used by Consortium members will be obtained from data that have been released to public databases.

This discussion of the ENCODE Consortium data release policy has been primarily directed at issues concerning the use of ENCODE/modENCODE data in scientific publications. The intent of the policy is to accelerate the use of the data by the scientific community. To facilitate this goal, the resource producers agree not to restrict the use of the data by others for all purposes other than those described above, while the resource users are encouraged to act in a manner that is consistent with this data access policy. The associated issue of intellectual property as it pertains to the ENCODE/modENCODE data is addressed in Appendix C.

Appendix A: Data Release Standard for Verified Data

The ENCODE and modENCODE Consortia will establish a well-articulated description of a verification standard for each data type produced by Consortia members and those standards will be described on the ENCODE and modENCODE websites (http://www.genome.gov/ENCODE and http://www.genome.gov/modENCODE, respectively). ENCODE/modENCODE research groups will release, to an appropriate public database, data obtained in experiments at the time that this standard is met. In most cases, it is anticipated that additional efforts for further verification and validation of the data will be carried out, but these should not delay the initial release of data. NHGRI acknowledges that releasing preliminary data may not be the first choice of the resource producers. However, on the assumption that such data can be useful to the scientific community, NHGRI has adopted the policy for the ENCODE and modENCODE Projects to make such data available in a timely manner. This policy is consistent with the Institute's commitment to rapid data release to the scientific community.

All of the data generated by the ENCODE and modENCODE Projects will be linked to the respective genome sequence. Data from the ENCODE Project that can be directly displayed on the human genome sequence will be stored and delivered by the DCC at the University of California, Santa Cruz (http://genome.ucsc.edu/ENCODE); other ENCODE Project data will be stored and delivered by the appropriate databases, such as GenBank, and GEO or Array Express. Similarly, for modENCODE, data that can be displayed on the genome sequences of *C. elegans* and *D. melanogaster* will be stored and delivered by the modENCODE DCC at the Ontario Institute for Cancer Research (http://www.modencode.org). All ENCODE/modENCODE data must have the associated information (metadata) on how the experiments were performed and how the raw data were analyzed to generate the conclusions (i.e., sequence elements) to be displayed. As data are deposited into public databases, individual tracks will be created

to display these data on genome browsers of the DCCs. Where applicable, the primary data underlying any sequence elements will be linked directly to the individual browser tracks. As additional data validations are performed, the investigators can modify the submitted data or even withdraw the data if further tests call into question the validity of the released data. All data will be accompanied by prominent caveats to notify users of the level of verification of the data and that frequent data release and updates will be forthcoming as further validation and analyses are performed.

Appendix B: Appropriate use of ENCODE/modENCODE Data

As noted above, NHGRI recognizes that the following examples do not address all potential uses of ENCODE/modENCODE data, nor are all of the terms in this document (e.g., locus) precisely defined. Resource users who have questions about the appropriate use of ENCODE/modENCODE data are encouraged to contact the resource producers directly. Specific questions can also be sent to NHGRI (contact information available at http://www.genome.gov/ENCODE).

Example 1: Analysis of a single gene locus. Resource users interested in analysis of a single gene locus can download all available ENCODE/modENCODE data for the locus and use the data to develop hypotheses about the individual gene locus. In this case, the resource users can publish on the data at any time by acknowledging the resource producers, the funding agency, and the respective DCC, along with referencing the appropriate ENCODE paper (as defined above). Resource users are encouraged to discuss their plans to use pre-publication data and to establish collaborations with the resource producers, but are not required to do so.

Example 2: Analysis of a single transcription factor's binding sites in the genome. Resource users interested in the analysis of the binding sites for a single transcription factor (or histone modification) can download all available ENCODE/modENCODE data for the transcription factor and perform any analysis of the data. However, prior to the expiration of the nine-month restricted period, the resource users should obtain the consent of the respective resource producers prior to submission of a manuscript. After the exclusivity period has expired or the data have been published, resource users are free to submit any manuscript with proper acknowledgement of the source of the data.

Example 3: Integrative analysis of all ENCODE/modENCODE data. Resource users can obtain all ENCODE/modENCODE data and integrate the different datasets to identify correlations of functional elements and possible biological insights from the data. These users are free to submit any manuscript on these analyses provided that the ENCODE/modENCODE data previously have been published and that the source of the data is properly acknowledged. If any data used for the analysis are derived from unpublished data prior to the expiration of the nine-month protected period, then the resource users should obtain the consent of respective resource producers prior to submission of a manuscript.

Appendix C: ENCODE/modENCODE Intellectual Property Issues

Since the inception of the Human Genome Project, NHGRI policy has encouraged the rapid release and ready accessibility of genomic data to the broad research community. A related issue of availability pertains to any intellectual property rights that might be sought by data generators, and the effects that the exercise of such rights have on access to the data.

The Bayh-Dole Act of 1980 provides a statutory mandate to NIH grantees and contractors to seek patent protection, when appropriate, on inventions made using government funds and to license those inventions with the goal of promoting their utilization, commercialization, and public accessibility. While the NHGRI has, in accordance with that law, encouraged grantees to seek patent protection for genomic technologies that have been developed with grant funds, the Institute has been concerned about the claims and exercises of those claims in the case of large-scale genomic data sets because of the Institute's belief that broad accessibility to the data is of paramount importance, and that such data are generally pre-competitive, i.e., a considerable amount of work would need to be performed beyond the initial data production to demonstrate utility. For genomic sequence data, for example, NHGRI indicated its opinion that raw data, in the absence of additional experimental biological information, lack demonstrated specific utility and therefore are inappropriate materials for patent filing. The grantees participating in the NHGRI large-scale sequencing program and other large-scale data generation projects have been monitored for whether they filed patent claims and, to date, none have.

In the case of the International HapMap Project, the participants (including the NHGRI grantees) agreed not to file for patents on the bulk data from the project. However, there was a complication because the raw data produced by the project (SNPs and individual genotypes) had to be processed to generate the project's ultimate output (haplotypes). In considering the issue of data release, HapMap participants were concerned about the possibility that researchers outside of the project could add some of their own data to the raw project data, develop haplotypes prior to the project's ability to do so, file patent claims based on the combined data, and then potentially restrict access by others to the HapMap data (a so-called parasitic patent). To deal with this concern, a click-wrap license was imposed on the individual genotype data; to gain access to the data, researchers were required to agree not to restrict the access of others to the data and not to share the data with anyone who has not agreed to the click-wrap license.

In some respects, the cases of genomic sequence data and haplotype data were relatively easy to deal with because the data themselves do not have "utility" (in the patent law sense of the term). As a result, grantees did not express concern about the NHGRI policies on data release. In the case of the ENCODE and modENCODE Projects, however, the applicability of this argument is not as obvious. The ENCODE and modENCODE Consortia will include both members funded by NHGRI ENCODE/modENCODE grants and those funded by other sources. The purpose of the ENCODE and modENCODE Projects is to generate data that identify or define genomic DNA sequence elements that have biological function, and therefore might be considered to have utility and be able to be patented. Therefore, the use of patents in

ways that might restrict access to large amounts or broad categories of data, e.g., all transcription factor binding sites, is an issue that needs to be addressed.

NHGRI's primary interest is to ensure the widespread availability of all information and any inventions that are generated during the ENCODE and modENCODE Projects. NHGRI, therefore, encourages all ENCODE/modENCODE resource producers to consider placing all information generated from their project-related efforts in the public domain and to address the NIH guidelines on the sharing of research tools (http://sharing.nih.gov). In the cases in which the Consortium members elect to exercise their intellectual property rights, NHGRI encourages consideration of maximal use of non-exclusive licensing of patents to allow for broad access and stimulate the development of multiple products as outlined in the NIH Best Practices for Licensing Genomic Inventions (http://www.ott.nih.gov/pdfs/70FR18413.pdf). As a criterion for joining the ENCODE and modENCODE Consortia, investigators have agreed to abide by the project's data release policy.

NHGRI also encourages users of the ENCODE/modENCODE data to act responsibly and share the effort involved in maintaining unrestricted access to the data. Thus, for example, if a data user were to incorporate ENCODE/modENCODE data into an invention, the subsequent license should not restrict the access of others to the ENCODE/modENCODE data. For this purpose, the term "resource users" is meant to include both researchers who are members of the ENCODE and modENCODE Consortia and researchers who are not.

The ENCODE pilot phase provided NHGRI with an opportunity to observe data producer and data user practices with respect to intellectual property and the ENCODE Project. NHGRI grantees are reminded that the grantee institution is required to disclose each subject invention to the Federal Agency providing research funds within two months after the inventor discloses it in writing to grantee institution personnel responsible for patent matters. NHGRI has monitored grantee activity in this area and it does not appear that attempts are being made to patent large amounts of information derived from the ENCODE Project. If, in the future, circumstances arise that convince NHGRI that additional measures are needed to achieve the goal of widespread access to the results of the project, the Institute reserves the right to consider a Determination of Exceptional Circumstance to restrict or eliminate the right of parties, under future grants, to elect to retain title. Similarly, NHGRI will continue to monitor the activity of resource users to attempt to determine whether access to the ENCODE/modENCODE data is being encumbered by any restrictive licenses. If the policy of reliance on data user responsibility to maintain unrestricted data access is not effective, the NHGRI will consider adopting a click-wrap license similar to that used by the International HapMap Project to protect the ENCODE/modENCODE data and to ensure unrestricted access to the use of the data.