National Advisory Council for Human Genome Research

May 21, 2013

Concept Clearance for RFA

Interpreting Variation in Human Non-Coding Genomic Regions Using Computational Approaches and Experimental Support

Purpose: A new initiative is proposed to support the development of highly innovative computational approaches for interpreting variants in the non-protein-coding regions of the human genome. The goal is to develop methods that combine functional and other datasets to identify or narrow the set of variants that are candidates for affecting organismal function leading to disease risk or other traits. Applications should include development of computational approaches to identify variants that potentially affect organismal function, as well as provide experimental data testing these predictions to allow assessment of the accuracy of the computational methods.

Background: Although the majority of GWAS disease-associated variants and functional elements found so far are not in protein-coding regions, most sequencing projects currently analyze exomes rather than whole genomes. While the cost difference is one reason, a more challenging reason for focusing on exomes is the difficulty of interpreting variation in non-protein-coding regions.

Some variants found to be associated with disease by GWAS studies are located in ENCODE functional elements, but these associations are not sufficient to allow the interpretation of the functional effects of such variants. Linkage disequilibrium (LD) means that multiple genes, genomic elements, and variants in a region may all be statistically associated with a trait, although only one may affect the function. We usually do not know how variants in ENCODE elements affect their molecular functions or whether a molecular phenotype affects an organismal phenotype.

Analysis using additional data is needed to identify which variants affect function at the organismal level. For example, several genes in a region may contain variants associated with a disease, but data showing that only one of those genes has a different RNA expression level in tissues relevant to that disease would provide evidence that the variants affecting organismal function are in the genomic elements regulating that gene. Thus, analyzing multiple data types may further narrow the set of potential variants affecting organismal function.

Proposed Scope and Objectives: This initiative will support research to develop computational methods to combine functional and other datasets to narrow the set of variants that are candidates for affecting organismal-level traits or diseases. The scale of analysis that the initiative seeks to address is genome-wide interpretation of the variants that may contribute to the trait or disease being studied, rather than variants found in a particular gene, gene family, or chromosomal region. The initial approaches for narrowing the set of candidate variants may

include data from other studies such as GWAS or scans for natural selection that start with the entire genome and narrow the focus to sets of regions for more analysis.

The focus of the proposed methods should be on variants in non-protein-coding regions, although the genome-wide analysis results may also include variants in coding regions. The focus may be on variants in specific classes of sites, such as CNVs, transcription-factor binding sites, or CpG islands. The approaches must be generalizable beyond the specific datasets and traits or diseases studied. The approaches might not be able to prove that the variants affect organismal function, but should narrow the set of variants that need further study.

The data types studied could include genome sequence, GWAS genotype and phenotype data, gene-gene or gene-environment interactions, patterns of variation, and various functional data types such as RNA expression, transcription-factor binding sites, and chromatin structure. Other data types may also be proposed. Data from model organisms may be used for interpreting the human variants. Applicants will need to explain what datasets they will use and how they will obtain them. All datasets used should be available to other researchers, at least by the end of the first year of the award. Relevant data may be obtained from public resources such as dbGaP, TCGA, 1000 Genomes, GTEx, ENCODE, Roadmap Epigenomics, the Molecular Signatures Database, and GEO.

Applications may identify one or more organismal traits or diseases to study, such as a human disease, disease resistance, pharmacologic responses, or physiological traits. The variants may have been identified for study by whole-genome sequencing or genotyping arrays. The deliverable should be a robust, generalizable approach to integrate various data types to prioritize genetic variants for their potential to contribute to organismal traits or diseases.

All applications should include some experimental data that provide evidence about the validity of the computational predictions, to assess their effectiveness. Although validating which variants causally contribute to an organismal phenotype can be extremely challenging, a variety of experimental approaches could be used to provide supporting evidence, ranging from high-throughput, less physiologic methods that could evaluate many predictions, to low-throughput methods that provide more complete physiological support for a few predictions. Such data may also be obtained from patients, or from model organisms such as mouse and zebrafish. Creativity in how to produce such evidence cost-effectively is encouraged. Experimental datasets that already exist may also be used.

Aside from the experimental work to provide evidence about the validity of the predictions, this initiative will not support large-scale data production or phenotyping, databases, development of methods to associate variants with disease independent of functional data, or approaches that simply aggregate information on variants.

Relationship to Ongoing Activities:

- The NIGMS RFA “New Methods for Understanding the Functional Role of Human DNA Sequence Variants in Complex Phenotypes” http://grants.nih.gov/grants/guide/rfa-

files/RFA-GM-13-002.html (expired) is much broader than this proposed RFA, because the NIGMS RFA includes a focus on experimental approaches. Only a couple of the awards for this NIGMS RFA are relevant to the proposed RFA, and this topic is large enough that they will not resolve the questions. NIGMS and other ICs may join this proposed RFA.

- The NHGRI RFAs "Technology Development for High-Throughput Functional Genomics (R01, R21, R43/44)" http://grants.nih.gov/grants/guide/rfa-files/RFA-HG-11-013.html (expired) support four projects to develop high-throughput experimental methods to characterize isolated elements using reporter assays, and one project to characterize elements in their native locations.
- The Common Fund RFA "Development and Application of Systems Approaches for Analyzing the Impact of Genomic Variation on Tissue Transcriptomes (R01)" http://grants.nih.gov/grants/guide/rfa-files/RFA-RM-12-019.html (expired) supports statistical analysis of GTEx data, including cis- and trans-acting variants that determine gene expression quantitative trait loci (eQTLs).

Mechanism of Support: R01. Awardees will meet once a year to exchange ideas.

Funds Anticipated: Each award will be up to $500,000 direct costs for each of 3 years, to support computational methods development and experimental testing of the methods. We expect to make 5-6 awards in each of FY 2015 and 2016, for a total of $4.5 million for each of 3 years for each round. Because the topic is challenging and applications need both computational and experimental elements, two rounds of the RFA are proposed. This allows research groups that are ready to develop applications to submit them in January 2014. It also provides notice that applications may be submitted in January 2015, to allow computational and experimental research groups to consider new methods and set up collaborations so that the approaches can have initial tests, allowing applications with well-thought out plans and preliminary data.

Concept clearance	Release RFA	Application receipt	Review	Council	Funding
May 2013	Aug 2013	Jan 2014	June 2014	Sept 2014	Jan 2015
		Jan 2015	June 2015	Sept 2015	Jan 2016

FY 15	FY 16	FY 17	FY 18	
$4.5	$9	$9	$4.5	($ millions)

Examples

Here are some examples of computational approaches that this RFA could support. More innovative approaches are also encouraged.

Layers of -omic data related to disease

1. Use GWAS genotype, sequence, and phenotype data to find genomic regions associated with a disease. Use GTEx, ENCODE, and Roadmap Epigenomic data to reduce this set of genes to ones expressed in cell types known to be related to the disease. Use ENCODE, DNA methylation, pathway, protein interaction, model organism, LINCS, and clinical data to narrow the set of variants that potentially affect organismal function.

Classes of variants

(These examples are more about molecular than organismal function. However, they could be combined with other data types to contribute to inferences on potential organismal function.)

2. Use information on chromatin structure (histone modifications and DNase hypersensitive sites) to predict where indels likely affect gene regulation. For example, indels affect active open chromosome domains for persistent fetal hemoglobin, and affect neighboring gene expression for ALS and type 1 diabetes.

3. Use information on transcription factor binding and RNA expression to predict how variants in promoter sites affect transcription factor binding and RNA expression.

Patterns of variation

4. Use patterns of variation within and among populations to find genomic regions that have undergone directional or balancing selection. Use data on conservation and variation within and among other species to narrow these regions. This approach will identify regions with variants that affect organismal function, but will not indicate their functions; it can be combined with functional approaches.

5. Use information on the variability of epigenomic marks in the population and differences in epigenomic marks among people with and without a disease and among twin pairs discordant for that disease to assess how variability in genes that affect epigenomic marks and in the epigenomic patterns correlates with disease risk. This would show how patterns of epigenomic variability could be used to narrow the set of genomic variants and epigenomic patterns that potentially affect organismal function.