Statement of research interests

Jason Flannick, PhD

Complex human diseases are influenced by lifestyle, environmental, and genetic factors. For most, we lack adequate means for prediction, prevention, or treatment. To advance our understanding of disease biology, genetic mapping studies have identified many genomic regions causally linked to disease. Today, a central challenge in complex disease genetics is to translate these associations to biological or clinical insights.

To define and test hypotheses suggested by genetic associations, molecularly manipulable model systems (e.g. animals or cells) are necessary. Defining appropriate model systems and validating their relevance to human pathophysiology remains a major obstacle impeding biological translation. In recent years, however, high-throughput genomic and perturbational studies have begun to yield sufficient data to envision computationally facilitating the design and interpretation of experimental characterizations.

My research aims to use computational methods to speed translation of genetic associations to biological or clinical insights. To this end, I hope to develop computational models of disease mechanisms that can aid selection and interpretation of experimental manipulations and appropriate model systems. These models will be founded upon genetic and cellular characterization of "high impact" variants from large-scale sequencing studies – that is, mutations of large molecular or phenotypic effect – and over time incorporate additional genomic datasets and experimental perturbations. My work will focus on type 2 diabetes (T2D), a complex disease of persistent hyperglycemia and leading cause of global morbidity and mortality.

Past and current research

My research interests have been formed by a background in computer science, statistical genetics, and more recent close biological and clinical collaborations.

Method and algorithm development. As a computer science PhD student, I developed efficient algorithms [1] and machine learning approaches [2,3] to identify conserved regions of DNA [4] and "modules" within protein interaction networks [5]. This work led to one of the first practical algorithms for large-scale network comparison, as well as a new paradigm for improving accuracy through principled data integration.

Statistical analysis of DNA sequence data. As a postdoctoral fellow, I studied DNA sequence data for insights into T2D and other complex diseases [6]. Early work led to a genotyping method integrating multiple technologies [7] and software to process tens of thousands of exome sequences through various analytical pipelines, used in numerous genetic studies [8–14].

Work I co-led also used a combination of large-scale sequencing and genotyping experiments to produce a comprehensive characterization of the genetic basis of T2D [14], using computational models I co-developed [15]. Deeper analysis of the same or similar data also (a) showed that presumed damaging variants in genes for rare forms of diabetes in fact have limited effects in the general population, quantifying some limits of genetic risk prediction [16,17]; (b) identified protective loss-of-function mutations in a gene (*SLC30A8*) encoding a pancreatic islet zinc transporter, suggesting a novel potential therapeutic hypothesis for T2D [18]; and (c) demonstrated the feasibility of integrating cellular model assay readouts with genetic associations, identifying broader T2D "allelic series" within *PPARG* [13] and *HNF1A* [19].

Integrating genetic datasets. With *SLC30A8* as a model, the genetic datasets produced for these and other studies were recognized to be of great value to advance biological discovery or drug development for T2D. Today I lead the team building the data coordinating center (DCC) for the Type 2 Diabetes Knowledge Portal (T2DKP) [20], an integrated collection of T2D-relevant genomic datasets accessible through a public web portal [6]. Several projects fund this activity, including building a central genetics knowledgebase and portal, improving methods for and visualizations of genomic analyses, implementing a federated knowledgebase of datasets from around the world, exploring new techniques for privacy-preserving distributed genetic analysis, and instantiating analogous knowledgebases for other diseases.

Future research

In recent years, the critical challenges in complex disease genetics have begun to shift. No longer are we limited by our ability to identify disease associations or explain disease inheritance – genetic mapping has yielded thousands of associations and suggests that genetic risk is mainly influenced by common variants across hundreds of genes. Instead, we are limited by our ability to deliver biological and clinical value from these advances: common variant associations are challenging to localize to genes or disease mechanisms, and biological insights have not yet emerged from large-scale sequencing studies. Increasingly, it is clear that translating associations to disease insights requires rich data, diverse experiments, and hard work.

To accelerate this translation, we need more efficient means to suggest, test, and refine biological hypotheses related to a genetic association. In my research I aim to address this knowledge and technology gap. I will first develop the use of high impact variants as an effective means to relate human physiological effects to molecular and cellular perturbations. I will then, over time, incorporate other genomic datasets into computational models of disease mechanisms, in which natural and perturbed molecular, cellular, and physiological states are statistically related and made broadly interrogable. I hope that these advances could one day help yield biological insights from genetic associations more rapidly and with more regularity.

To advance this goal, my future research aims to:

Develop methods to use genomic datasets to understand disease biology To meaningfully advance clinical or biological translation of genetic associations, a computational methodology should (a) support interpretable queries by diverse communities; (b) quantitatively link molecular, cellular, and physiological states; (c) predict the contextual impact of perturbing one state on many other states; and (d) evolve in tandem with relevant datasets.

In the first several years of my independent research, I will use high impact variants to address this challenge. I will work to efficiently identify these variants from large-scale exome sequencing studies for T2D; collaborate with experimentalists to link their molecular, cellular, and physiological effects; and enable them to be broadly queryable through a portal such as the T2DKP. I will develop new methods to prioritize high impact variants for further study, even if they do not achieve robust statistical significance. For example, I will develop statistical models to identify patterns of nominally significant associations across phenotypes that could capture disease-relevant "endophenotypes", as well as methods to integrate epigenomic, transcriptomic, and interactomic annotations to identify genes with an increased prior likelihood of carrying high impact variants. In my first R01 application, I plan to propose development of these techniques and their application to a 52K sample T2D exome sequencing dataset.

In the longer term, I will incorporate these methods into more comprehensive computational models of disease mechanisms. One research direction will be the machinery to represent, learn, and query these more general models; probabilistic graphical models (e.g. bayesian networks) offer one promising approach, as they can efficiently represent relationships among many variables while supporting relatively interpretable queries of them. Within this framework, I will aim to integrate traditional genetic analyses and large-scale biological datasets to enable complex inferences and easy interpretation by biologists. Key questions will be how to convert traditional statistics into probabilities of phenotypic association or functional impact, as well as how to link between experimental readouts (e.g. cellular phenotypes measured by CRISPR/Cas9 knockout screens or physiological phenotypes impacted in knockout mice) and clinical outcomes. My contribution to the Biomedical Translator [21], on which I serve as the technical lead of the Broad Institute team, was designed as an introduction to this research.

Connect and organize genomic datasets to calibrate these computational methods The biological utility of computational methods will be greatly determined by the data available for analysis. In my early independent research I will thus aim to expand the T2DKP to establish it as a comprehensive resource on complex disease-relevant genomic datasets. This will entail continued collaborations with current T2DKP data contributors, expansion to new collaborators through non-central "federated" knowledgebases, and quantification and resolution of privacy limitations that might limit the datasets available to the knowledge portal. Six grants on which I have a leadership role provide funding for these activities over the next 2-4 years, including two on which I serve as principal investigator.

In the longer term, I will seek to contribute my work and ideas for the T2DKP to all research communities to which they are applicable. Many complex disease research communities have similar datasets and goals as that for T2D, while consortia like the Global Alliance for Genomics and Health aim to establish broad standards for genomic and clinical data sharing. If these efforts could be organized and integrated using common access and computational standards, possibly via established computer science techniques for "data warehouses" or distributed programming, they could have a transformative impact on our ability to understand disease.

Develop experimental approaches to test and refine computational predictions Finally, I will seek to use my work to help study the biological mechanisms of genetic associations. In my early independent research, I will work with my established collaborators to investigate T2D associations using cell differentiation and glucose uptake assays in SGBS adipocyte models or insulin secretion and ion transport assays in Ins1e and EndoC-βh1 beta-cell models. Additionally, a small grant on which I am principal investigator provides funding for a project evaluating the use of zinc transport assays to characterize all *SLC30A8* missense mutations, and I aim in my first R01 to apply similar approaches in adipocyte models.

In the longer term, I am interested to develop or help develop experimental frameworks and analyses that can be systematically applied to many different variants or genes. For example, assays that distinguish disease-causing from benign variants could be developed at single-cell resolution for use in genome-wide knockout screens, while systematic measurements of cell type or assay similarities might enable reference lookup tables to "impute" a variant's effect on several model systems. I would seek to analyze data from these approaches not only to characterize genetic associations, but also to further calibrate and improve the computational disease models that will be the ongoing research focus of mine.

Summary

In my research, I hope to help address a current critical challenge in complex disease genetics: how to accelerate the translation of genetic associations to insights into disease. I believe that computational methods, with the right focus, can help make genetic and experimental data more integrated, accessible, and useful in pursuit of this goal.

References

[1] Flannick J, Novak A, Srinivasan BS, McAdams HH, Batzoglou S (2006) Graemlin: general and robust alignment of multiple large interaction networks. Genome Res : 1169–1181.

[2] Flannick J, Novak AF, Do CB, Srinivasan BS, Batzoglou S (2008) Automatic parameter learning for multiple network alignment. In: RECOMB'08. pp. 214-231.

[3] Flannick J, Novak A, Do CB, Srinivasan BS, Batzoglou S (2009) Automatic parameter learning for multiple local network alignment. J Comput Biol : 1001–1022.

[4] Flannick J, Batzoglou S (2005) Using multiple alignments to improve seeded local alignment algorithms. Nucleic Acids Res : 4563–4577.

[5] Flannick J (2009) Algorithms for biological network alignment. Ph.D. thesis, Stanford University.

[6] Flannick J, C FJ (2016) Type 2 diabetes âĂŞ genetic data sharing to advance complex disease research. Nat Rev Genet .

[7] Flannick J, Korn JM, Fontanillas P, Grant GB, Banks E, et al. (2012) Efficiency and power as a function of sequence coverage, SNP array density, and imputation. PLoS Comput Biol : e1002604.

[8] Jimenez NL, Flannick J, Yahyavi M, Li J, Bardakjian T, et al. (2011) Targeted 'next-generation' sequencing in anophthalmia and microphthalmia patients confirms SOX2, OTX2 and FOXE3 mutations. BMC Med Genet : 172.

[9] Neale BM, Kou Y, Liu L, Ma'ayan A, Samocha KE, et al. (2012) Patterns and rates of exonic de novo mutations in autism spectrum disorders. Nature : 242–245.

[10] Bick AG, Flannick J, Ito K, Cheng S, Vasan RS, et al. (2012) Burden of rare sarcomere gene variants in the Framingham and Jackson Heart Study cohorts. Am J Hum Genet .

[11] Estrada K, Aukrust I, Bjørkhaug L, Burtt NP, Mercader JM, et al. (2014) Association of a low-frequency variant in HNF1A with type 2 diabetes in a Latino population. JAMA .

[12] Lim ET, Wurtz P, Havulinna AS, Palta P, Tukiainen T, et al. (2014) Distribution and medical impact of loss-of-function variants in the Finnish founder population. PLoS Genet .

[13] Majithia AR, Flannick J, Shahinian P, Guo M, Bray MA, et al. (2014) Rare variants in PPARG with decreased activity in adipocyte differentiation are associated with increased risk of type 2 diabetes. Proc Natl Acad Sci USA .

[14] The T2D-GENES consortium, The GoT2D consortium (2016) The genetic architecture of type 2 diabetes.

[15] Agarwala V, Flannick J, Sunyaev S, Altshuler D (2013) Evaluating empirical bounds on complex disease genetic architecture. Nat Genet .

[16] Flannick J, Beer NL, Bick AG, Agarwala V, Molnes J, et al. (2013) Assessing the phenotypic effects in the general population of rare variants in genes for a dominant Mendelian form of diabetes. Nat Genet .

[17] Flannick J, Johansson S, Njølstad PR (2016) Common and rare forms of diabetes mellitus: towards a continuum of diabetes subtypes. Nat Rev Endocrinol .

[18] Flannick J, Thorleifsson G, Beer NL, Jacobs SB, Grarup N, et al. (2014) Loss-of-function mutations in SLC30A8 protect against type 2 diabetes. Nat Genet .

[19] Najmi L, I A, Flannick J, Molnes J, Burtt NP, et al. (Submitted) Functional investigations of hnf1a identify rare variants as risk factors for type 2 diabetes in the general population.

[20] Type 2 Diabetes Genetics. URL `http://www.type2diabetesgenetics.org`.

[21] Biomedical Data Translator Program. URL `https://ncats.nih.gov/translator`.