Empowered genome community: leveraging a bioinformatics platform as a citizen–scientist collaboration tool

Katherine Wendelsdorf *, Sohela Shah

* Corresponding author.
E-mail addresses: Katherine.Wendelsdorf@qiagen.com (K. Wendelsdorf), sohela.shah@qiagen.com (S. Shah).

1. Introduction

There is on-going effort in the biomedical research community to leverage Next Generation Sequencing (NGS) technology to identify genetic variants that affect our health. The main challenge facing researchers is getting enough samples from individuals either sick or healthy – to be able to reliably identify the few variants that are causal for a phenotype among all other variants typically seen among individuals. At the same time, more and more individuals are having their genome sequenced either out of curiosity or to identify the cause of an illness. These individuals may benefit from a way to view and understand their data. QIAGEN's Ingenuity Variant Analysis is an online application that allows users with and without extensive bioinformatics training to incorporate information from published experiments, genetic databases, and a variety of statistical models to identify variants, from a long list of candidates, that are most likely causal for a phenotype as well as annotate variants with what is already known about them in the literature and databases. Ingenuity Variant Analysis is also an information sharing platform where users may exchange samples and analyses. The Empowered Genome Community (EGC) is a new program in which QIAGEN is making this on-line tool freely available to any individual who wishes to analyze their own genetic sequence. EGC members are then able to make their data available to other Ingenuity Variant Analysis users to be used in research. Here we present and describe the Empowered Genome Community in detail. We also present a preliminary, proof-of-concept study that utilizes the 200 genomes currently available through the EGC. The goal of this program is to allow individuals to access and understand their own data as well as facilitate citizen–scientist collaborations that can drive research forward and spur quality scientific dialogue in the general public.

© 2015 Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).
However, the former ‘sample abundance’ problem still persists. Even with access to all current knowledge about a list of variants, researchers are often in need of additional samples in order to detect disease-causing variants with <100% penetrance as well as those that may be common in an under-sampled healthy subpopulation. Such additional samples are hard to find.

An increasing number of individuals are opting to get their genetic variants through direct-to-consumer (DTC) genotyping such as 23andMe as well as programs such as Illumina’s UVG (http://www.illumina.com/company/events/understand-your-genome.html) and Harvard’s Personal Genome Project (http://www.personalgenomes.org). Often it is curiosity or the need to diagnose a health condition that motivates consumers to directly obtain this genetic information. However, once it is received, they may lack the bioinformatics skills and resources to make sense of their own datasets. Or they may want to connect with trained scientists who may be able to use their data to inform a disease etiology. Such individuals would greatly benefit from a platform to make sense of their data and effectively share it with researchers who can use it to expedite discoveries that lead to treatments and diagnostics.

Here, we introduce QIAGEN’s Empowered Genome Community. This is a program to help people make their own well-sequenced genomes more scientifically useful. In short, we are making Ingenuity Variant Analysis freely available for any individual to explore their own genetic information and share it with researchers in the on-line interface. This will (i) allow individuals to view and explore their own data without the need for bioinformatics skills, (ii) help them better understand the standard genomic interpretation process and its caveats, (iii) enable them to securely share their data with specific researchers of their choice, as well as (iv) create a resource of additional phenotyped genomes for researchers.

2. Results & discussion

2.1. A user-friendly tool for genomic analysis and exploration

Ingenuity Variant Analysis is an online application to filter variants from NGS data and identify a plausible causal pathway from variant to diseases. Users with and without extensive bioinformatics training are able to use the interface to incorporate published experimental observations, variant frequencies across different databases, and a variety of statistical models to define filtering criteria and mechanistic hypotheses. A typical data analysis workflow is outlined in Fig. 1.

(a) Users upload a list of all possible variants – often several thousands to millions – either as an individual sample or as groups of related or unrelated cases/controls samples. (b) Construct a custom filter cascade that brings in different pieces of information to sequentially eliminate candidate variants down to a short list based on different criteria such as:

- **Confidence filter**: data quality such as the read depth at a variant position.
- **Common variants filter**: frequency of the variant in the general population determined by its frequency across different sequence databases (1000 genomes project, NHLBI exomes, and Allele Frequency Community, hosted by QIAGEN).
- **Predicted deleterious filter**: Whether the variant is likely to be pathogenic based on experimental observations in the literature, ACMG guidelines, damaging sequence characteristics (truncating, frame-shift, etc.) defined by in silico modeling algorithms (SIFT, PolyPhen, MaxEntScan, and others), and more.
- **Genetic analysis filter**: whether it is of a genotype one would expect of the disease-causing variant, for example, homozygous in affected individual if the trait is recessive.
- **Statistical association filter**: whether, in a multi-sample study, it is significantly associated with an affected case group versus a group of healthy controls.
- **Biological context filter**: whether the variant occurs in a gene that has been previously implicated in a user-specified disease/phenotype or shown to affect a specific physiological process.

Users start with a default filter cascade and can modify the filters and parameters to create an analysis best suited for their specific scientific question(s). The result is the short list of variants that fit all of the criteria set in each filter. (c) For this shortlist of variants, one can then access peer-reviewed and manually curated experimental information about known gene–gene interactions to identify biological pathways between a mutated gene and a specific phenotype. (d) In the same interface, samples as well as analyses and the filter cascade can be shared with colleagues to view in their own account. This allows users to share/pool samples, collaborate on constructing analyses, as well as effectively communicate about any actionable results.

2.2. Empowered genome community – how it works

As registered members of the Empowered Genome Community (EGC), individuals who have their genomes sequenced can upload and
analyze their variant data using Ingenuity Variant Analysis for free and can also share it with researchers within the application to serve either as cases or controls. With this access, these individuals can explore DNA sequence variants that they have and learn the observed effect of these variants in published studies. Members of the EGC have the option to share their sequence data and health information within Ingenuity Variant Analysis in one of three ways:

i) **Broadly:** an EGC member fills out a health survey and makes the health information and all uploaded sequence data available to all Ingenuity Variant Analysis users. This data is viewable by QIAGEN, but is not sold. It is made freely available on the Ingenuity Variant Analysis platform to researchers who request access to it.

ii) **Privately:** an EGC member shares a specific subset of their genetic variants or health information with specific researchers with whom they are in contact. The data is not viewable by QIAGEN.

iii) **Just genetic variants and no phenotypic data:** an EGC member may submit their variants to the Allele Frequency Community (AFC). This is a community of researchers and laboratories that share anonymized, pooled allele frequency statistics. It is hosted by Ingenuity Variant Analysis and is currently composed of over 100,000 integrated, anonymized human exome/genome variant datasets. This resource is freely available to anyone to annotate and filter their variants with AFC frequencies.

The EGC benefits researchers by giving them access to any of the ‘broadly’ shared genomes with specific phenotypes. Researchers submit the number of samples they need and the phenotypes that must be associated. EGC samples that fit the specified phenotypes will then be provided to them in their account at which point they can use the donated samples in their case-control studies. In addition, they have a platform with which to collaborate and privately exchange data with specific EGC members that they are in contact with.

### 2.3. Value of EGC: access, educate, collaborate

We see the value of the EGC as being three-fold:

**Access:** with Ingenuity Variant Analysis, EGC members are able to use the interface to incorporate published experimental observations, variant information from different databases, and various statistical models to contextualize and identify the potential relevance of their genetic variants. This is all possible without the need for programming experience and big data management skills. In this way, EGC membership allows individuals to view, understand, and explore their own data in a way that is not yet available to them.

**Educating the public:** the EGC provides a means for non-scientists to learn about the caveats of the genome interpretation process. In using the filter cascade and application tutorials to filter and explore their own variants, individuals will be exposed to the concept of variant call confidence and sequencing data quality. They will also be able to see the variety of experimental evidence and criteria used to conclude that a genetic variant is medically relevant. Importantly, they will quickly see how changes to those criteria, through filter cascade modifications, can alter results. This is a small step towards educating the public to be more discerning about claims posed by pharmaceutical companies, DTC sequencing companies, and the media about the impact to be more discerning about claims posed by pharmaceutical companies, DTC sequencing companies, and the media about the impact to be more discerning about claims posed by pharmaceutical companies, DTC sequencing companies, and the media about the impact.

**Fostering collaboration:** through the EGC, members control their own data allowing them to privately view their data on a secure platform and, at the same time, make it scientifically useful by sharing it with specific researchers of their choice. The sharing platform allows them to, not only exchange data, but also better understand and communicate with scientists about what is being done with their data. This is because the researcher can later share any analyses done with the member’s sample for viewing in the interface. This continued communication allows for more active collaboration as both parties desire. EGC members have complete control over the amount of data they share and who they entrust to see it.

### 3. ‘Open’ analysis of EGC data gives insight into genetic causes of hypercholesterolemia

There are currently ~200 phenotyped genomes in the broadly shared pool of EGC genomes – where the sequenced individuals agreed to share the data as well as provide phenotypic health information. The most common health phenotype is hypercholesterolemia. As a proof-of-concept we sought to identify variants specific to hypercholesterolemia using these donated samples. Specifically, we have taken 12,331 shared genomes (36 cases, 95 controls). The analysis was limited to exonic regions of the samples yielding an initial set of 271,840 total variants spanning 18,888 genes. The following filters were then applied:

Variants were kept only if they met the call confidence criteria of a call quality $\geq 20$ in a case or control sample and occur outside the top 5% most exonic variable 100 base windows in healthy public genomes (1000 genomes). Of these, we excluded variants that are observed with an allele frequency $\geq 5\%$ of the genomes in the 1000 genomes project (http://www.1000genomes.org), the public Complete Genomics genomes, the NHLBI ESP exomes, or the Allele Frequency Community (http://www.allelefrequencycommunity.org). Of these we only kept a variant in the list if it met one of the following criteria for predicted deleteriousness: i) the variant was experimentally observed to be associated with a ‘Pathogenic’ or ‘Possibly Pathogenic’ phenotype according to ACMG guidelines, ii) the variant demonstrated gain of function effect on the gene according to a published study, iii) the variant causes a frameshift, in-frame indel, or stop codon change, iv) is a missense mutation that is not predicted to be innocuous by SIFT or Polyphen-2, v) the variant disrupts a splice site up to 2.0 bases into an intron, or vi) is predicted to disrupt splicing by MaxEntScan (Burge et al., 2004). The Statistical Association filter was then applied to the remaining variants to conduct a SKAT-O test that identified genes with a significant variant burden in either the case or control group with a burden test p-value $< 0.01$. We then applied a biological context filter to identify and keep burdened genes that have been previously associated with ‘Hypercholesterolemia’ or ‘cholesterol metabolism’ according to the Ingenuity Knowledge Base.

The final result was a set of 7 variants across 2 genes: TF and LRP5. Interestingly, LRP5 (lipoprotein receptor-related protein) bears a significant burden among the healthy control samples indicating that its mutation may be protective against hypercholesterolemia. Specifically, 5 of the identified 7 of the variants occur in the LRP5 gene across 16 healthy controls, but none of the hypercholesterolemic cases. In particular, c.1999G >> A (p.Val667Met) is an exonic missense variant in LRP5 whose frequency is significantly higher in the control group, occurring in 12 controls and 0 cases (p-value $< 0.05$). Though this variant has not been linked to hypercholesterolemia in humans, it has been shown to affect cholesterol metabolism in mouse models (Fujino et al., 2002). The other 4 variants in LRP5 (c.632G >> A(p.W211*), c.1337A >> G(p.N446S), c.16A >> G(p.T6A), c.2009_2010insC(p.T1252fs*30)) each occurs in only 1 individual.

Of the 2 variants that occur in TF (transferrin), 1 occurs in only 2 cases and the other in only 1 case. Though TF, as a gene, bears a significant variant burden in the hypercholesterolemic case group, these are the only 2 variants that have been previously linked to hypercholesterolemia and neither occurs at a significantly higher frequency in the case group vs. the control group.
This Analysis is freely available for anyone to view and modify in Variant Analysis. All are invited to visit the study and explore this preliminary finding as well as refine the findings with their own insights on hypercholesterolemia etiology, sample structure, and validation.

4. Concluding remarks

Critical to personalized medicine research is the ability to identify the relevant genetic variants in an individual. Having enough samples to detect a variant in the affected population is a key challenge. This 'sample abundance' challenge can be partly reduced by platforms that enable individuals to share their health and biological data with researchers/clinicians in a useful form such that the data/information can be leveraged to make medically relevant discoveries faster. With Ingenuity Variant Analysis we see an opportunity to help individuals access their own data and, at the same time, allow them to readily share it with needy researchers in a ready-to-use form.

The EGC is quite an investment on the part of QIAGEN in terms of compiling, storing and securing the data to provide this free service. We have taken on these responsibilities because it is our philosophy that patients have a right to access and understand their own data. We also believe it is important for commercial and governmental entities to play a role in facilitating citizen-scientist collaborations that can drive research forward and spur quality scientific dialogue in the general public.

More information

Where can I go to get my sequence?

As part of Illumina’s Understand Your Genome Program: http://www.illumina.com/company/events/understand-your-genome.html

Become a member of the Personal Genome Project: http://www.personalgenomes.org/harvard/sign-up

How do I join the EGC?

Once you have your genetic variant data in .vcf format you can register at the EGC website to upload your data and start analyzing it. Please contact your sequencing service to find out how best to receive your data in a .vcf format.

Who will see my private data?

Only you, and anyone you choose to share it with.

Do I have to share my private data?

No. If you wish, you can analyze your own genome, and any genome(s) shared to you by others, without sharing yours.

Will QIAGEN sell, rent, or mine my data?

QIAGEN will not sell or rent any data obtained as part of the EGC. If you actively opt to join the Allele Frequency Community (AFC), then your sequence data will be used to update the anonymized and pooled statistics for variants in the AFC.

How can I view and affect the hypercholesterolemia study?

Visit the Variant Analysis web page to create an account. Once in the application, you can click on the ‘Publications’ tab to open, view, and modify the study.

Links

Ingenuity Variant Analysis: http://www.ingenuity.com/products/variant-analysis
Allele Frequency Community: www.allelefrequencycommunity.com
Empowered Genome Community: https://www.qiagenbioinformatics.com/empowered/

Acknowledgments

We would like to thank Nathaniel Pearson for the initial conception of the Empowered Genome Community. In addition we would also like to acknowledge the work of Rupert Yip, Douglas Bassett, and Daniel Richards for creation of the Allele Frequency Community and their intellectual contributions to the design and implementation of Variant Analysis.

References

Forbes, S.A., Beare, D., Gunasekaran, P., et al., January 2015. COSMIC: exploring the world’s knowledge of somatic mutations in human cancer. Nucleic Acids Res. 43 (Database issue), D805–D811.

Fujino, T., Asaba, H., Kang, M.-J., et al., 2002. Low-density lipoprotein receptor-related protein 5 (LRP5) is essential for normal cholesterol metabolism and glucose-induced insulin secretion. Proc. Natl. Acad. Sci. 100 (1), 229–234.

Hamosh, A., Scott, A.F., Amberger, J.S., et al., January 2005. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res. 33 (Database issue), D514–D517.

Law, V., Knox, C., Djoumbou, Y., et al., January 2014. "DrugBank 4.0: shedding new light on drug metabolism. Nucleic Acids Res. 42 (no. Database issue), D1091–7.