Arivale®

Genetic Curation White Paper

Authors: Isabelle Lucas-Beckett, PhD
Niha Zubair, PhD
Matt Conomos, PhD

Created: June 2018

CL-18-0074-WP-REV.1

Introduction

Arivale uses a system approach to improve overall health and wellbeing of its members. Arivale integrates and analyzes individuals' genetics, labs, microbiome data, and behavioral assessments to create personalized lifestyle and behavior change plans that are delivered by licensed healthcare professionals. All dietary and lifestyle coaching recommendations are evidence based and align with recognized clinical guidelines (e.g., American Heart Association, Diabetes Prevention Program, and American College of Sport Medicine).

Herein we describe the methodology for creating genetic insights. Although published studies have shown that genetic variants have only a small impact on non-Mendelian wellness-related phenotypes, curating and reporting relevant genetic variants is possible and can help guide personalized wellness recommendations. While there are currently no standard regulations or guidelines on how to curate and report such variants, at Arivale we have created internal standards to maintain consistency, accountability, and documentation of our variant curation process. This white paper describes the process Arivale uses to determine the validity and strength of variants in both candidate genes and from genome-wide association studies. Arivale only reports variants that meet our quality and evidence-based standards described here.

Selection of Variants in Candidate Genes

We describe the process we created to identify and evaluate the evidence for an association between single variant(s) in candidate genes and a given phenotype. A candidate gene corresponds to a gene of interest that was specifically chosen at the beginning of a study. The gene could be of interest because of its biochemical function, location in the genome, or its specific known genetic variants, which may be linked to the normal or pathological physiology related to the phenotype.

Literature search method

Phenotype of interest

To ensure the most comprehensive literature search, we start our curation process by defining the objective of the search as follows: 1) What is the phenotype of interest? 2) What is the candidate gene of interest (if initially known)? 3) What is the most relevant variant(s) reported for the candidate gene (if initially known)?

We then create an exhaustive list of "keywords" or word permutations that have been used to refer to the phenotype, the gene, the related protein, and the variant(s). We utilize publicly available resources such as Medical Subject Headings (MeSH®), HUGO Gene Nomenclature Committee, GeneCards, UniProtKB/Swiss-Prot, and dbSNP, as well as the expertise of our clinicians and scientists. Keywords, in conjunction with Boolean operators and other special characters, enhance our PubMed searches and assess the overall body of scientific evidence for the specific search. This comprehensive literature search will be further expanded by running new PubMed queries with any additional relevant keywords that may be identified during the curation.

Depending on the objective of the curation, the specific candidate gene(s) and/or genetic variant(s) may not be defined initially. In this case, all of the potential candidate genes and genetic variants potentially associated with the given phenotype will be surveyed during a first broad round of curation using more generic search terms around genetics (e.g. “gene”, “genetics”, “variant”, “SNP”, “polymorphism”, etc.). After this initial review process, the identified candidate gene(s) and genetic variant(s) are then prioritized for a more in-depth curation process using the keyword process described above.

Lastly, for each literature search performed, we select the publications that are relevant for the objective of the curation and retrieve the full text articles. Publications without an available full text are not included in the conclusion making process of the curation. Each literature search is fully documented, including the total number of search results, the number of relevant results, the date the search was run, the exact search terms and logic used for the query, and if any filters (e.g. publication dates, article type) were used for the search results.

Assessing Gene by Environment interactions

The environment to be assessed in the gene by environment (GxE) curation is either identified during an initial broad literature search or is well-established clinical knowledge for the specific phenotype. Of note, the literature search for the GxE is done independently of the search for the potential impact of the genetic variant on the phenotype, i.e. without any keyword related to the GxE. This is done to ensure the most comprehensive review of the potential association between the genetic variant(s) and the phenotype.

The workflow of the curation process for GxE is identical to the one described above for assessing the impact of the candidate variant.

Study selection

Only studies relevant to the phenotype of interest with full text articles are evaluated. We allow for a wide range of study types in our curation process: observational (longitudinal, cross sectional, case-control, prospective), meta-analyses, experimental, functional, and genome-wide association studies (GWAS). If a single study is part of a meta-analysis, we include only the

meta-analysis in order to ensure that no individual study is considered twice in the evaluation process. We do not include reviews or systematic reviews for assessing the level of evidence, but we consider them during our curation process to ensure that all relevant keywords and studies are included in our search process, as well as for providing some background information.

Rating study quality

We defined 5 key criteria for rating the quality of studies (good, moderate, or poor) identified during the curation process. Most of these criteria are similar for the different study types that we routinely curate:

- Adequate study size (depending on the trait and type of study)
- Relevance of the ethnicity of the study population (e.g. not limited to a genetically isolated population such as Hutterites or Sardinians)
- Consideration of the environmental/lifestyle factors that may influence phenotype of interest
- Appropriate statistical analysis
- Replication conducted (for GWAS) OR whole gene/targeted SNP panel (for candidate gene studies)

We also take into account the impact factor of the journal. A good quality study needs to meet at least 4 criteria, a moderate quality publication needs to meet 3 criteria, and a poor quality study meets fewer than 3 criteria. Unless the study was published in a high impact factor (>5) journal, only publications of good and moderate quality are further assessed.

Assessing study evidence

We record the following study details to assess the level of evidence. All of this information will be considered for the overall evaluation of the study findings.

- Study type (e.g. intervention, meta-analysis)
- Study population (e.g. sample size, age, sex, ethnicity)
- Inclusion and exclusion criteria (e.g. pre-existing condition, medications)
- Study design (e.g. intervention regiment, follow-up time frame)
- Measurement method of phenotype (self-reported or directly assessed)
- Statistical analysis (association testing method, multiple-testing correction, controlling for confounding factors if necessary)
- Effect of genotype on phenotype (statistical significance, direction of effect, mode of inheritance, limitations)

Evaluating overall findings

We then evaluate the overall body of findings by tallying the following for the specific genotype-phenotype association in question:

- The number of publications with as well as without a statistically significant association between the genotype and the phenotype.
- The mode of inheritance of the genotype-phenotype association.

We also record possible limitation of the overall findings:

- Gender (e.g. association only found in males).
- Age (e.g. association found only for a certain age group).
- Ethnicities of the populations included in the selected studies.
- If lifestyle (e.g. smoking) or pre-existing medical history (e.g. diabetes) are required for a significant association.

Finally, we assign a confidence rating to the level of evidence for each genotype-phenotype association (for a specific inheritance model). For this purpose, we have established a five-level confidence rating system:

- Confidence Level 1: Lack of significance
- Confidence Level 2: Unknown significance
- Confidence Level 3: Moderate significance
- Confidence Level 4: Strong significance
- Confidence Level 5: Very strong significance

We have defined different case scenarios that can lead to a given confidence level, for example:

- A lack of significance (Confidence level 1) may be due to no association between the variant and the phenotype when assessing good and moderate quality studies. A lack of significance may be due to conflicting results, where there are equal numbers of publications with and without a significant association.
- A strong significance (Confidence level 4) could be due to no conflicted data and at least 4 publications, with a minimum of one good quality study. A strong significance could result where there is some conflicted data but there are a greater number (2 or 3 times) of publications reporting a significant association than not.

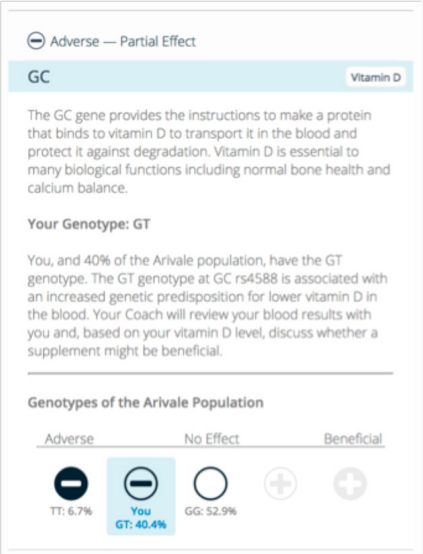
Arivale Confidence Levels reflect the degree of controversy for the findings, as well as the overall number and quality of the relevant publications that were included in the curation process. In general, we only report genetic associations that reach a Confidence Level 3 or higher (moderate to very strong significance). Occasionally, for GxE variants where there are a limited number of studies but the studies are of good quality, we may report a level 2 association.

Reporting data for genetic variants from candidate gene curation to Arivale Members

For associations that reach our Confidence Level threshold, we report the genetic variants as either a single genetic variant or as small polygenic profile.

Single genetic variant

- For a given genetic variant, we define the direction of the impact for each possible genotype (adverse, neutral, or beneficial). Multiple factors are taken into account for defining the direction of the impact: the biological/physiological effect of the different alleles, if known, their frequencies in the general population (1000 Genomes Project, the Exome Aggregation Consortium, and our own internal data), and if a specific environment (amount of fat in the diet, exercise, etc.) can modulate the impact of a given allele on the phenotype.
- We define the level of the impact (partial, full, or no effect) for a specific genotype, based on the most likely inheritance model of the risk allele when taking into account the overall body of evidence from the curated relevant studies. A partial impact is defined as having only one copy of the risk allele when there is an additive model of inheritance. A full impact is defined as having one or two copies of the risk allele when there is a dominant model of inheritance, or as having two copies of the risk allele when there is an additive or recessive inheritance model. No effect is defined as having zero or one copy of the risk allele in the case of a recessive model of inheritance, or as having zero copies of the risk allele in the case of an additive or dominant inheritance model.
- To illustrate further how we report single genetic variants, here is the example of rs4588, a variant in the GC gene, which is associated with lower blood levels of vitamin D. In this case, the individual has only one copy of the risk allele (T), leading to an Adverse--Partial Effect result.

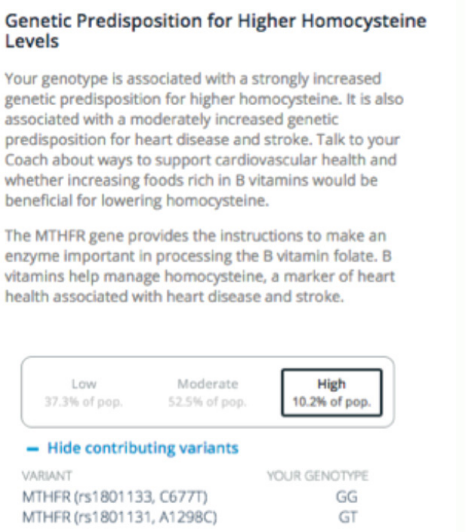


Small polygenic profile

When possible, we create small polygenic profiles by combining multiple variants curated from candidate gene approach studies. We consider creating small polygenic profiles when we have multiple genetic variants associated with the same phenotype or GxE. The following are three scenarios where this may occur:

- Scenario 1: the variants of interest have been studied **together** in the literature for the phenotype of interest. The variants may be in/near the **same gene or different genes**. This is the best case scenario in which we have some knowledge on how these variants may together affect an individual's genetic predisposition for the phenotype or GxE. This information allows us then to create a data-weighted polygenic profile for the different genotype combinations.

- Example: The two genetic variants included in this small polygenic profile, rs1801133 and rs1801131 are located within the same gene, MTHFR and there is published data on how these two variants may together affect blood levels of homocysteine. Depending on the genotype observed for the two genetic variants, an individual may have a “Low”, “Moderate”, or “High” genetic predisposition for higher homocysteine levels. In this display, the percentage of individuals in the Arivale population falling into each of the three different categories is indicated as a reference.



- Scenario 2: the variants have **not been studied together** and are in/near the **same gene**. We consider that we do not have sufficient information to assess the likely epistatic effects of these variants, which, in this case, should be taken into consideration for creating a polygenic profile. In other words, it is unclear how to combine the impact of multiple variants, especially when considering how they interact when found on cis (same copy of the gene) or trans (different copies of the gene) in an individual. In this scenario, we will not create a polygenic profile and report the different genetic variants individually.

  - Example: Two genetic variants, which are both located in/near the CYP1A1 gene, have been found to be associated with an increased genetic predisposition for lung cancer, especially in smokers. However, because we don’t have data on how these two variants would together affect this phenotype, the variants are individually reported using the single genetic variant display.

- Scenario 3: the variants have **not been studied together** and are in/near **different genes.** We create a polygenic profile in which each of the different genotypes are weighted based on the direction of their effect (adverse, neutral, or beneficial), their level of impact (no effect, partial, or full), and their confidence rating for the level of evidence. In this situation, due to the lack of available data, the polygenic profile does not take into account the less likely possibility (in comparison to Scenario 2) of an interaction between the different genetic variants that could affect the phenotype (epistasis).

- Example: The three genetic variants of interest, rs12722, rs143383, and rs679620 are associated with an increased genetic predisposition for Achilles tendinopathy. They are located in or near three different genes, COL5A1, GDF5, and MMP3 and they have not been studied together. In this case, they will be reported together as a small polygenic score by combining the data for each individual variant as described above. Depending on the genotype observed for each of the three variants, an individual's genetic predisposition for Achilles tendinopathy could be "Increased", "Slightly increased", or "No effect".

Curating and Reporting Variants in Genome Wide Association Studies

Below we discuss the process for identifying and evaluating the evidence for an association between GWAS data and a given phenotype.

Literature search method and study selection

We use the GWAS Catalog of published genome-wide association studies, which is curated and maintained by the National Human Genome Research Institute and the European Bioinformatics Institute (NHGRI-EBI) as a one of our main resources for GWAS curation. We also use the PubMed search approach described above. Lastly, we perform a search for published results from consortia focused on the phenotype of interest.

We then select studies that could be used for creating the polygenic profile for the phenotype of interest. Specifically, we look for published studies of good quality (see study quality criteria mentioned above) with either an already created and validated polygenic profile, or with available summary statistics (i.e. rsIDs, allele frequencies, effect sizes, and p-values) for the genetic variants tested.

Rating study quality and assessing study evidence

To assess and rate quality for GWAS studies, we use the same curation process for variants in candidate gene. We evaluate the selected studies to identify the most relevant and comprehensive publication with available data that will be used for the genetic insight. Only the

highest rated study is used, where study date and study size are strongly considered, and meta-analyses are given priority.

Creating the polygenic profile

Published polygenic profile

When the authors created their own polygenic profile and showed a significant association with the phenotype of interest in an independent population, Arivale utilizes the formula for that polygenic profile exactly as published. We are not currently reporting published polygenic profile.

Arivale polygenic profile

When authors did not create a polygenic profile, Arivale utilizes the GWAS summary statistics to identify the genetic variants (single nucleotide polymorphisms or SNPs) potentially associated with the phenotype of interest and constructs a polygenic profile. The creation of the Arivale polygenic profile is based on a more statistical procedure. The SNP selection procedure is as follows:

- Based on the p-values reported in the published GWAS, SNPs are filtered using a false discovery rate (FDR) threshold of 5%.

- Among these FDR filtered SNPs, we check that the alleles reported in the GWAS summary statistics match both the alleles reported in the 1000 Genomes Project reference genotype data for the relevant ancestral population, as well as the alleles observed in the Arivale genotype data. Any SNPs for which the alleles don't match either data set are not included.

- If allele frequencies were reported as part of the GWAS summary statistics, we don't include any SNPs for which the effect allele frequency is very different from either that allele's frequency reported in 1000 genomes for the relevant ancestral population, or that allele's frequency observed in the Arivale population.

- P-value informed linkage disequilibrium pruning of the remaining SNPs is performed using the r2 values reported in the NIH LDlink database (https://analysistools.nci.nih.gov/LDlink/) for the relevant 1000 genomes ancestral population. This results in a set of nearly independent SNPs likely associated with the phenotype.

The polygenic profile for each Arivale member is calculated by taking a weighted sum of his/her observed number of effect alleles across all of the selected SNPs, where the weights for each SNP are given by the effect sizes published in the GWAS summary statistics.

Assessing the validity of the Arivale polygenic profile

Using our database of thousands of members, we look for a correlation of the newly created Arivale polygenic profile with the phenotype of interest in our population. The ability to correlate the polygenic profile with our data strengthen our confidence in the genetic insight. However, it may not be possible to do so if the phenotype is not collected in our database, or our population with the phenotype is too small (inadequate statistical power). In these cases, we either use an external source of publicly-available data for assessing the new polygenic profile or we confirm that the variants selected for the polygenic profiles were found with comparable effect sizes in at least two independent GWAS assessing unrelated populations. If none of the alternative options are possible we do not report the polygenic profile.

Conclusion

Currently, there is no accepted standard in the wellness and disease prevention industry for curating and reporting genetic variants. At Arivale, we are determined to be transparent and rigorous in our processes. We advocate for standardization so that consumers appreciate the limitations and interpretation of their genetic results. Ultimately, we hope that consumers will gain more confidence and understanding the impact of genetics on their wellness. This will empower consumers to make diet and lifestyle choices that positively impact their health.