NTNU
Norwegian University of Science and Technology
Faculty of Natural Sciences
Department of Biotechnology and Food Science

Sigve Strand Landa

Human Phenotype Prediction Through use of Variant Data in Metabolic Modeling

Master's thesis in Biotechnology
Supervisor: Eivind Almaas
Co-supervisor: Christian Schulz, Martina Hall
May 2022

NTNU
Norwegian University of
Science and Technology

Sigve Strand Landa

Human Phenotype Prediction Through use of Variant Data in Metabolic Modeling

Master’s thesis in Biotechnology
Supervisor: Eivind Almaas
Co-supervisor: Christian Schulz, Martina Hall
May 2022

Norwegian University of Science and Technology
Faculty of Natural Sciences
Department of Biotechnology and Food Science

NTNU
Norwegian University of
Science and Technology

Summary

In this project, a method was developed with the purpose of using variant data, primarily SNPs, to perform knockout metabolic modeling, in an attempt to predict diseases caused by the SNPs in question. This approach would not only allow detection of disease-causing SNPs, but also describing the mechanism behind a given disease. The data used was from 2548 individuals from the 1000 Genomes Project, as well as PheWAS data. Combinations of genes for knockout were generated based on SNPs present in the individual, and for PheWAS, SNPs associated with each other. There was a primary focus on nonsense SNPs as they are more likely to cause loss of function. For the individual data, some 2600 such SNPs were identified. These combinations then formed the basis for which genes would be knocked out when performing flux balance analysis (FBA). A number of Human1 based tissue specific models were used for the analysis. Additionally, metabolic tasks were used to further analyze the effects of the knockouts. The purpose of these tasks is to test specific biological functions within the models, and they proved instrumental in detecting issues caused by the knockouts, including many that would have otherwise gone unnoticed. To conclude, for several cases, the approach was successful in predicting disease, with the results matching what was expected based on existing literature.

Oppsummering

Føremålet med dette prosjektet var utvikling av ein metode for bruk av variasjonsdata til å selektere gen for fjerning i metabolsk modellering. Motivasjonen er moglegheita til å kunne oppdage sjukdomsframkallande variantar, samtidig som at mekanismane for sjukdommen kan bli beskrive. Variasjonsdata frå 2548 individ frå 1000 genom prosjektet, samt PheWAS data blei brukt i seleksjonen. Hovudfokuset var på SNPs, og spesielt ikkjesynonyme SNPs som produserer premature stop-kodon. Dette grunna deira meir destruktive natur, som vidare gir breiare grunnlag for fjerning av genet. For individ variantane ga dette 2600 SNPs som blei brukt til å produsere gen kombinasjonar for fjerning, basert på SNP kombinasjonane i individa. Ei rekke organ-spesifikke modellar basert på Human1 modellen blei brukt til FBA analyse. Spesifikke metabolske funksjonar blei testa gjennom bruk av konkrete oppgåver. Disse oppgåvene viste seg å vere svært nyttige i å avdekke problem forårsaka av manglande gen, med mange resultat som ikkje ville ha blitt oppdaga utan. I sum var metoden i ei rekke tilfelle vellykka i å reprodusera effekten av ein SNP, basert på allereie observert sjukdom.

Abbreviations

SNP: Single Nucleotide Polymorphism

SNV: Single Nucleotide Variants

WT: Wild Type

LoF: Loss of Function

GWAS: Genome-Wide Association Study

PheWAS: Phenome-Wide Association Study

EMRs: Electronic Medical Records

GEM: Genome-scale metabolic model

tINIT: Task-driven Integrative Network Inference for Tissues

GTEx: Genotype-Tissue Expression

FBA: Flux Balance Analysis

FVA: Flux Variability Analysis

Table of Contents

1 Introduction

One can argue that our genome is what we are. Our DNA is our blueprint, with each of us carrying a unique code, that from a single first cell, is used to build us up. The uniqueness of our individual code is in part what makes us different from each other. It gives us certain traits, physical as well as psychological[1], that whether we like it or not, we must live with. However, even though we know that our DNA gives us our form, we do not have a complete understanding of how it all works. Why do some people get heart disease, while others have a healthy heart into old age? Why do some people suffer neurodegenerative diseases, while others maintain a sharp mind until the end? Undoubtedly, many of our differences when it comes to such diseases, are due to our lifestyles and other environmental factors.[2] It is, however, also apparent that genetics is an important risk factor for various diseases.[3] Due to our genome having such a big impact on our health, and subsequently our lives, it is a field of intense study; which variants leads to, or increases the risk, for which outcomes.

One approach to gaining knowledge on variants is through genome-wide association studies (GWAS). These studies attempt to associate certain diseases with certain variants through population studies.[4] People who have a certain disease, as well as similar people without the disease are studied for common variants. If certain variants are consistently found in the diseased group, and less so in the non-disease group, that variant may be associated with the disease. While most variants are benign, this method has been successful in finding numerous variants causing or increasing risk for various diseases. Unfortunately, it does not say anything about the underlying mechanism with which the variant has a causal relationship to the disease.[4] Of course, once a variant has been associated with a disease, it can be further scrutinized in that context. Disease causing variants will often be associated with a gene, and it may be that the variant damages the expression or functioning of the gene, or even knocks it out completely. However, even with this information about the direct effects of the variant, it may not be apparent how damage to a certain gene causes a disease; the complete functioning of the gene may not be known or understood. Another possibility is that a variant which damages a gene only causes disease in conjunction with damage to another gene. In this scenario a variant may be associated with a disease, but the fact that it only causes said disease in combination with certain other variants, is not uncovered.[5] Again, such relationships between variants can be uncovered after the fact. In any case, a more holistic approach is needed to get the complete picture of variant(s) to disease, whether only a single gene, or multiple, are involved.

With ever advancing data collection technologies and computers, we can collect and process ever increasing amounts of biological data. And as the amount of knowledge on the functioning and relationships between individual components of the human body has increased, they have been systematized into networks and models.[6] The purpose of these models is to emulate the functioning of the human body, or other organisms, in a computer. By changing the inputs or making changes to the model itself, it is possible to predict real world outcomes of those changes, by running computer simulations.[7] To this purpose, metabolic models have long been developed, with ever increasing complexity and detail. These models attempt to emulate human metabolism, wherein metabolites undergo reactions, the reactions are facilitated by proteins, which again are tied to genes. A simple analysis of a metabolic model is the flux balance analysis (FBA).[8] This analysis treats the model as a mathematical problem in which an objective is maximized. The objective is usually biomass production, although it can be any reaction in the model. Depending on inputs, reactions will have fluxes through them, representing the quantity of metabolites that were consumed and created in the reaction. One use of FBA is to perform knockout analyses, in which FBA is run on a perturbed model, usually the deactivation of one or more genes.[9] By comparing the results of such an analysis to an unperturbed model, it is possible to make a prediction on how those knockouts would affect the real organism. This can then be used to

In this master's project a pipeline has been developed with the purpose of identifying disease causing variants, including combinations of variants, through the use of metabolic modeling. In this setting, an unperturbed model may represent a healthy individual, while a perturbed model may represent a diseased individual. By using variants to select the genes for knockout FBA, underlying metabolic mechanisms for diseases, or risk of disease, may be uncovered. This is because the genes, and their products, are put in the context of each other. Thus, if a variant is damaging to a gene, the larger ramification of that damage will become apparent through analysis of the model. With this approach, the concrete mechanism of the variant(s) on a particular disease may be described, at least in the larger context. In addition, because multiple genes can be knocked out at the same time, multiple variants can be studied at the same time, as specific combinations. This allows for the discovery of variants whose detrimental effects are only present in combination with other variants. Determining which genes to knockout is done through filtering out variants with potential gene damaging effects, and using the combinations of these variants present in actual individuals. Because it is based on variant combinations present in individuals, the combinations can be quite large without requiring insurmountable amounts of FBA runs to run all combinations of that size. The individuals essentially act as filters for which larger gene combinations will be used for knockout. An alternative approach also explored, is to use variants associated with disease through GWAS, and try to identify underlying mechanisms for their association with said disease. Combinations can also be generated here, this time based on associations. SNPs associated with the same disease can be used in combination. **In sum, the aim of this project was to develop a pipeline that can identify disease causing variants, and their underlying mechanisms for causing said disease, through knockout metabolic modeling.**

2 Background

The background for this project is two part: first, genetic variation and how it affects the human body, second, metabolic modeling and how it can be used to predict real world outcomes. For the first part, a detailed look is given on genetic variation, particularly SNPs and how they can affect a gene and subsequent organism. Some insight is also given on how disease-causing SNPs have traditionally been discovered, through high penetrance mutations causing obvious disease, or through GWAS using population studies. Some thought is also given to the advent of individual genomes and personalized medicine. The second part gives insight into the function of metabolic models and various analyses used on them. The main analysis used is Flux Balance Analysis (FBA). Another topic heavily featured in his project is metabolic tasks, whose purpose it is to test specific functions within the metabolic models using FBA.

2.1 The Genome and its Variations

All genetic information is encoded in DNA by the sequence order of the four nucleic acid bases: Adenine, Thymine, Cytosine, and Guanine. The human genome has a length of about 3.1 billion bases, or double that if both copies of the chromosomes are counted.[10] Scattered in the chromosomes are genes, which encode for functional products, such as proteins. Although, while most of the genome does not directly encode functional products, much of the non-coding parts are still thought to have function, for example in gene expression regulation.[11] When a genetic sequence encodes a protein, it is translated by reading three and three bases, which are referred to as codons. Each codon translates into either an amino acid or a stop codon. (Figure 1) Which codon translates to which amino acid is governed by the genetic code.

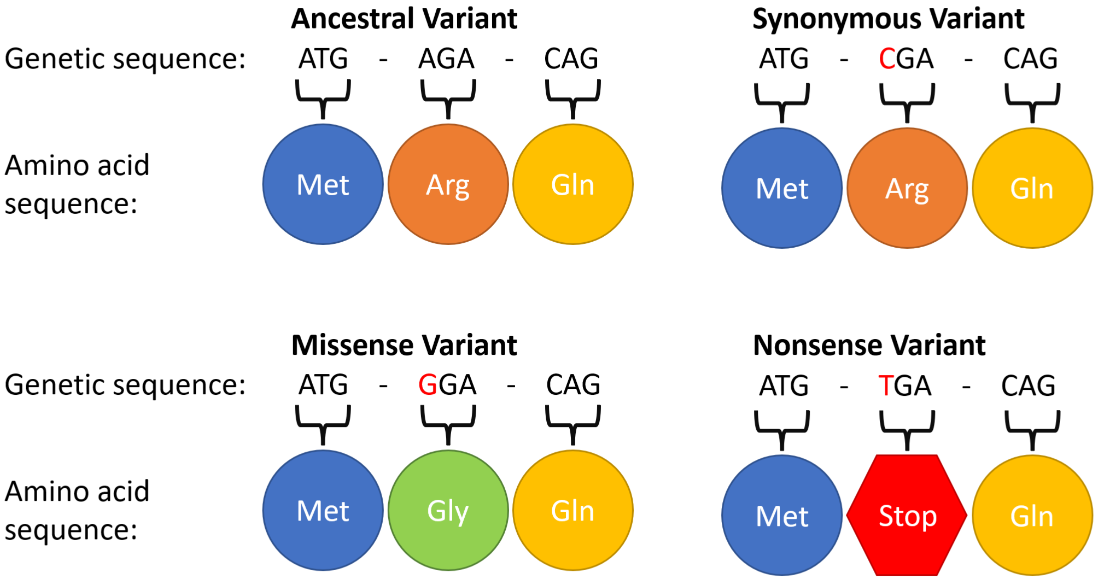


Figure 1: Different SNP types and their potential effect on amino acid sequence if located inside protein coding regions. In each case a single base has been exchanged for another. **Ancestral:** The variant seen as ancestral in that it has persisted in its initial state. In contrast, the other variants are considered divergent in that they have change from the initial state at a later point. **Synonymous:** This variant type does not change the amino acid sequence due to the codon still coding for the same amino acid. Also called a silent mutation. **Missense:** This variant type causes a change in amino acid sequence. Not depicted here is the initiator codon mutation, which is when the start codon (ATG, coding for Methionine) is changed to a different amino acid, and the stop lost mutation, which is when the stop codon is changed to a codon that codes for an amino acid. **Nonsense:** This variant type leads to a premature stop codon, which causes the protein to be truncated.

Genetic variation refers to the differences in base sequence between individuals or populations.[12] On any given locus, a specific position known for variation, there will be an ancestral allele, and one or more divergent alleles. A variant is ancestral in that it has persisted in its initial state, while the divergent variants have changed away from this initial state at some point. It is however possible for multiple alleles to coexist, and different populations may have different alleles as their most common on a given locus, which depending on time frame, could be viewed as ancestral in that group.

The most common genetic variation is the single nucleotide polymorphisms (SNPs).[13] (Figure 1) As the name suggests, this is when a single base is exchanged for another one. Single nucleotide variation (SNV) is the general term for SNP, with SNP referring to a variation that is present in more than 1% of the population. Nevertheless, SNP is commonly used regardless of it fulfilling that criterion.[13] SNPs can be divided into several categories with the first two being whether or not it is located within a protein coding region. If a SNP is located in a coding region, there are several sub-categories: synonymous, missense, and nonsense mutations. Synonymous mutations, also called silent mutations, is when a change in coding sequence does not lead to a change in encoded amino acid. Missense mutations is when one amino acid is substituted for another, while nonsense mutations refer to a premature stop codon. The indel is another common variation type, referring to small insertions or deletions.[13,14] If such a mutation occurs in a coding region, and it does not consist of three bases, there will be a shift in the reading frame. A shift in the reading frame will result in a completely different amino acid sequence, and depending on location, rendering the gene non-functional. If the indel occurs towards the end of the protein, the functional parts may still be intact, and the protein as a whole may still function. When a variant does render a gene nonfunctional, or causes a reduction in function, it is known as a Loss of Function (LoF) mutation.[15] (Figure 2) LoF can be caused by any mutations, even synonymous ones[16] or those located in intergenic regions.[17,18]

2.2 Loss of Function and Genetic Disease

Genetic loss of function, either complete loss or reduced, can be disease causing.[15] (Figure 3 These variants are also not particularly uncommon as it has been shown that all individuals may harbor over 100 potential LoF variants, with potentially over 30 being homozygote.[13,19] However, the frequency of any one variant is still usually low, which means few individuals will be homozygote for that particular variant[20], although there are exceptions. For example, LoF variants do not have to cause disease, even when homozygous. A study from 2009

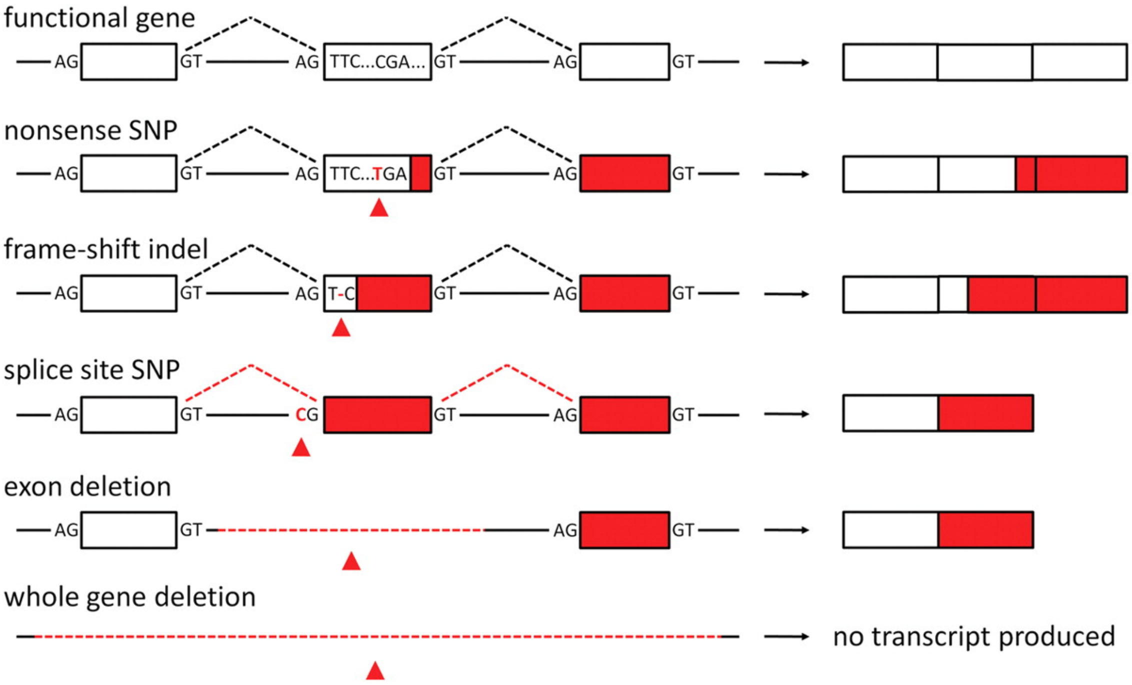


Figure 2: (Acquired from D. G. MacArthur and C. Tyler-Smith [19]) Different variant types causing LoF. Red triangle denotes the specific variant. On the right is a depiction of expressed transcript, with white parts being what can be translated. Any protein functionality located in the red part of the transcript, which is downstream of the variant, is lost.

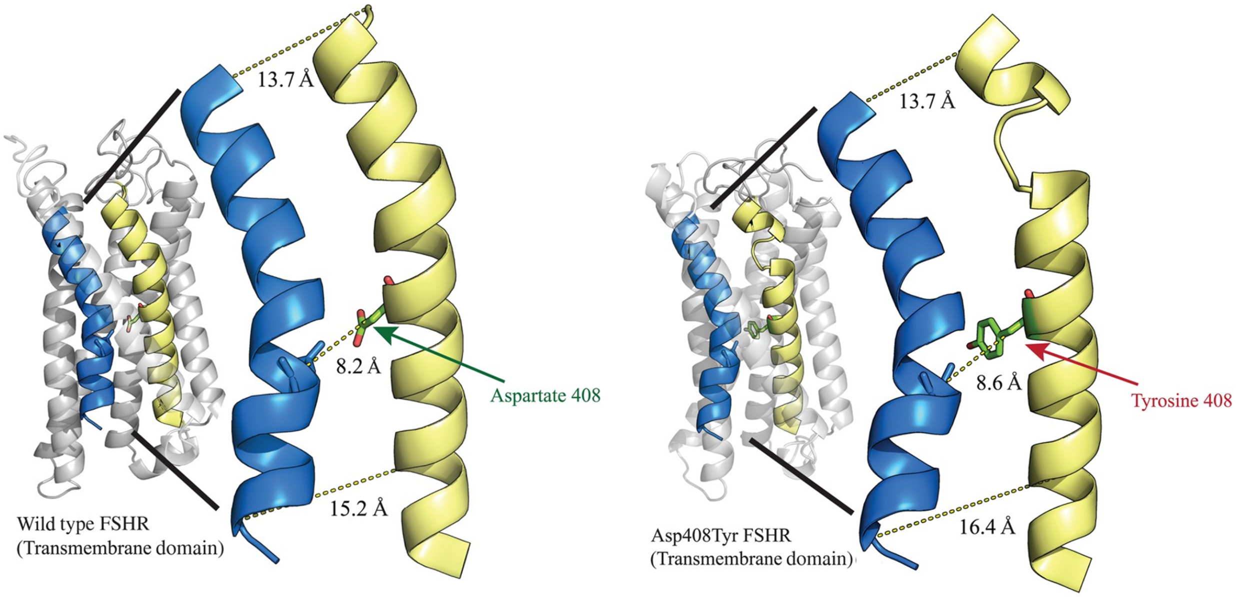


Figure 3: (Modified from M. S. Bramble et. al.[26]) Example of how a SNP can impact protein structure showcasing the *in silico* predicted change in structure of follicle-stimulating hormone receptor protein (FSHR) due to a SNP. On position 408, aspartate in the wild type has been substituted with tyrosine in the mutant. The substitution is predicted to cause a significant change in secondary structure upstream of the mutated amino acid. This apparent small change has been implicated in primary ovarian failure and subsequent reduced fertility.

found that a group of humans differed on average by 24 genes caused by nonsense SNPs alone, indicating that deactivation of certain genes may have no detrimental effect or even be advantageous.[21] Indeed, some genes are known to have a high frequency for nonfunctional variants, with research indicating deactivation may be advantageous. One example is CASP12, a gene that through positive selection, has become predominantly inactive in humans, likely due to reduced sepsis severity.[22,23] Another example, which is much less prevalent in the population, is SLC30A8. The lack of function in this gene, as when inactivated by LoF mutations, is known to reduce risk of type 2 diabetes.[24] Taken together this means that many genes will not cause any issues when inactivated or reduced in function. Although, apparent loss of gene functionality without negative consequence may also be an indication of biological robustness, where certain genes are redundant in their function.[25] In this case specific combinations of variants, where multiple genes are affected, may be still cause disease.

Any mutation can cause LoF, however, particularly prominent in causing disease, due to their disruptive nature, are the nonsense mutations.[15,20,21] (Figure 2) They are not uncommon as the cause of monogenic diseases.[27,28] A genetic disease is monogenic when it is caused by a reduction, or loss, of function in a single gene. This type of diseases will typically have Mendelian inheritance, with offspring either being affected, carrying the disease, or be unaffected. These will typically be high penetrance diseases, due to the variant alone having a significant impact on phenotype. However, in most cases having one functional copy of a gene will be enough for a variant caused disease to be recessive, even with high penetrance. Penetrance is dependent on how certain it is that a particular genotype causes a particular phenotype.[29] Another way to look at penetrance of a variant is as that variant being a risk factor which increases the risk for a disease. If a particular genotype appears to cause disease in some individuals, but not in others, it might have a moderate or low penetrance, where the outcome is dependent on other factors as well. With moderate to low penetrance variants in monogenic disease, it may be more ambiguous how offspring is affected, but they will still be carriers if the variant is inherited. Other factors contributing to a monogenic disease, can be both environmental and genetic. In the case of contributing genetic factors, the disease is still monogenic in root cause, with the effects being exacerbated or reduced by other variants.[30] Although, in the case of LoF caused dominant disease, one defective gene copy is enough to cause the disorder, a phenomenon referred to as haploinsufficiency.[31] In such cases, the one functioning gene copy is insufficient to produce the wild-type phenotype.

There are many well-known monogenic diseases, such as sickle cell disease or Huntington's disease. Huntington's in particular is a good example of varying penetrance. The disease is caused by a single faulty HTT gene copy, with a homozygote being non-viable. A repeat of CAG trinucleotide, leading to a polyglutamine addition to the protein, is the underlying mechanism for the disease. The number of CAG repeats is variable and with higher number more likely to cause disease.[32] A moderate number of CAG trinucleotide repeats have a medium penetrance for the disease, with some individuals developing disease while others do not. At a medium penetrance it is clear that even though the disease has a monogenic root, with mutation of the HTT gene being required, there are other contributing factors to the progression of the disease as well.[33]

In contrast, complex traits or diseases do not have clear patterns of inheritance.[34] In these cases no single gene or variant is the root cause of a disease, but rather certain combinations cause the disease. Many of the most common disease types, such as heart disease, immune disorders, and psychiatric disorders, are typically rooted in complex genetic disorders. With complex diseases, heterogeneity is common.[35] This is when a disease has multiple distinct root causes. In genetic terms this means different genotypes give rise to the same or a similar phenotype. A disease may have a monogenic cause, due to rare high penetrance variants, or a polygenic root, where multiple variants each carry a risk that accumulate.[36] For example, heart diseases may be caused in a monogenic manner through rare variants, but usually occur as complex disorders with multiple factors.[37]

2.3 Genome-wide Association Studies

One important question when it comes to disease causing variants, is how to identify them. In the case of rare variants causing obvious symptoms, case studies might be used to identify the root cause through sequencing for those variants.[38] However, perhaps the most prominent tool disease causing variant discovery is the Genome-wide association study, which aims at identifying loci which are correlated with disease though the use of statistics.[4] If a loci differs significantly between a case group and a control group, the studied loci might have an influence on that case. While this has been successful at discovering many associated with various phenotypes, the predictability of the discoveries has been proven to be rather limited than first thought in the early days of GWAS.[39] Instead of the underlying cause of a disease being high effect variants, many diseases have been shown to instead be best explained by many low impact variants. GWAS identified variants also tend to be regulatory in nature, with most being located outside the coding regions, sometimes without it being clear which genes they are linked to. The regulatory influence might also be subject to cell-type, with certain cell types being affected more than others.[40]

Another consideration are the mechanisms behind the perceived effects. If it is not even clear which gene a variant is linked to, it is obviously not straight forward how the variant causes its effect. However, even when the gene is known, the underlying mechanism might not be easily discernible. Functional genomics seeks to shed more light on the specific biological effects underlying the associations uncovered by GWAS.[41] To highlight this need, Boyle, Li, and Pritchard[34] made the point that, with ever increasing cohort sizes for GWAS studies it is, through the interconnectedness of the regulatory networks, likely that more and more loci will be implicated to disease, with smaller and smaller effect sizes. This will get to a point where, in cells associated with a disease, every locus associated with a gene expressed in those cells, will be implicated with the given disease. This points to the possibility of GWAS reaching a limit of sorts, where heaps of variants are implicated through genes with no direct effect on disease.

To get a better understanding of direct effects of variants, a more functional approach is needed.[41,42] These approaches promises to uncover mechanisms behind the variants disease association such as functional network studies. Another limitation of GWAS which may be alleviated through more functional approaches is the problem of epistasis effects.[43] While detection of these gene-gene effects has improved considerably, there are still difficulties surrounding it. To this end, the aforementioned network approach can be useful, as it allows for uncovering cases where effects might be masked or aggravated based on gene relation.[44]

Another method that has come out of the GWAS approach for discovering disease associated variants is the phenome-wide association study (PheWAS).[45,46] This approach uses electronic medical records (EMRs) or other richly phenotyped data sets, in which genome data is linked to phenotype data (medical data), to perform a sort of reverse GWAS. In essence, for a given genotype, the range of associated clinical phenotypes are determined. Compared to GWAS, where a disease is linked to a wide range of variants, the PheWAS approach can say something about the range of diseases a given variant is associated with. Additionally, phecodes are used to group genetic variants to certain phenotypes with specific phecodes representing clinically meaningful phenotypes.[47] This makes it easy to identify a set of variants, and associated genes, for investigating a specific disease or phenotype.

2.4 Individual Genomes and Personalized Medicine

Personalized medicine seeks to tailor treatments for each individual based on a variety of factors, including their genome. (Figure 4) With complex, heterogenic, diseases being common, it is clear that more research into rare functional variants would be useful to this end.[48] Especially as GWAS is better at identifying common variants, with rare, high impact, variants being more challenging to pinpoint.[49] As whole-genome sequencing is becoming more common, more rare variants with potential disease impacts are discovered.[50,51] As previously stated, common disease might have uncommon causes in that different individuals have different underlying genetic causes. Diseases like diabetes and cardiovascular have many hundreds of implicated genes and associated variants, each individually or even collectively having little predictive power.[52] In addition, due to every individual having unique combinations of variants, with potential epistasis effects governing the genotype to phenotype relationships, the risk contribution of any one variant will vary.[53] If use of individual genomes in personalized medicine is to be achieved at a high level, not only will knowledge of the general risk contributed by a variant be necessary, but also how the impact of that variant is modulated by other variants that individual might carry.[30] Knowing the exact underlying mechanisms,

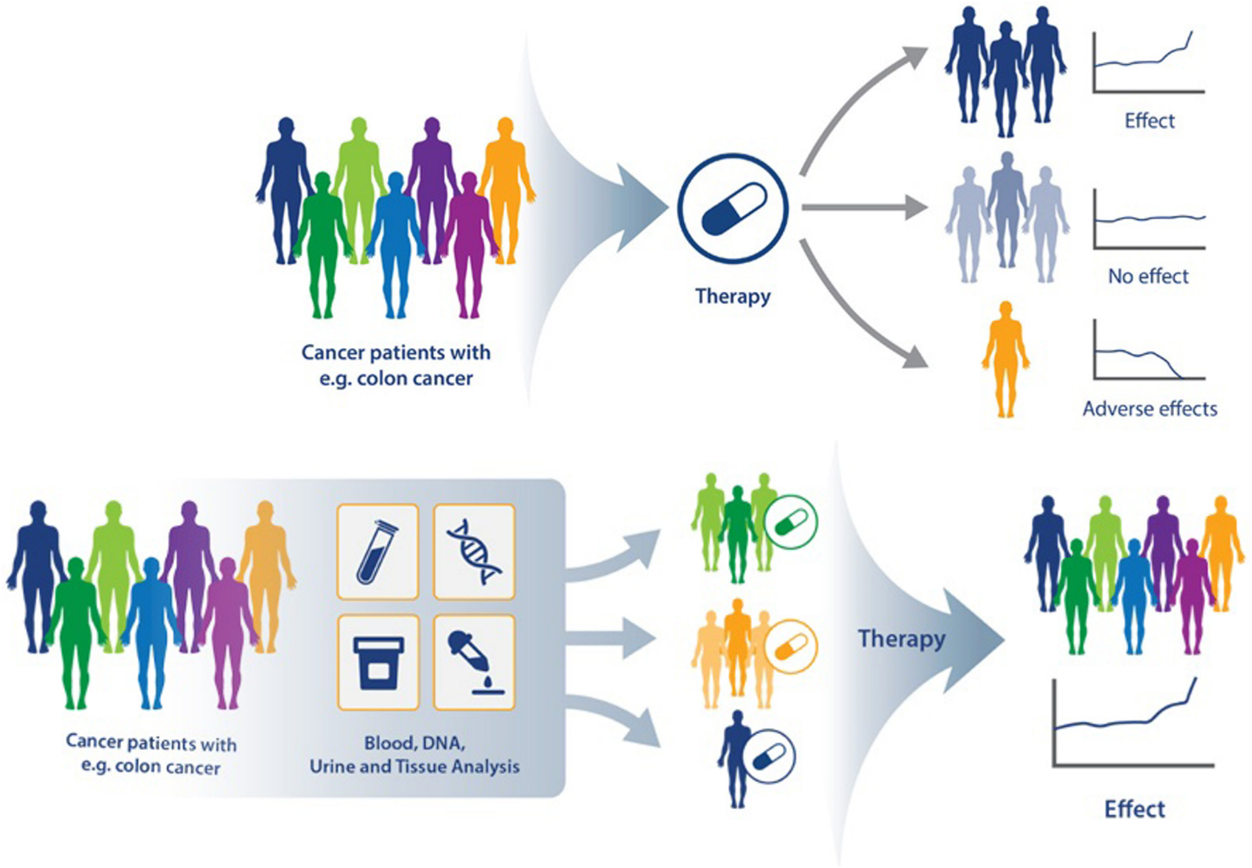


Figure 4: (Modified from J. Barbeau.[57]) Top: Current common medicine where one treatment fits all. This can lead to patent groups getting no benefit from a treatment, or even cause adverse effects. Bottom: With personalized medicine, different patients may get treatments tailored to their specific circumstances, producing an optimal treatment outcome for the individual.

including variant interactions, in an individual can help with selecting effective treatments for a given disease. (Figure 4) Additionally, better knowledge of which mechanisms exist, could help identification of targets for drug development.[54,55] A deeper mechanistic understanding could also allow for drugs which would have otherwise been discarded due to negative effects. For example, if a drug targets protein in a given pathway, and an individual happen to have a LoF variant in that pathway as well, there could be adverse effects. (Figure 4) If such cases can be predicted, an adverse effects due to rare variants could be avoided, as well as additional drugs that would otherwise be discarded, might instead be put to use.[56]

To get a deeper understanding of individual genomics, having whole-genome data available for studying is necessary. One such important source of whole-genome data for a large number of individuals is the 1000 Genome Project.[13,58] This data enables researchers to study genetic biallelic variants of the whole genome from several thousand people, from multiple population groups. In this data set, more than 86 million variants have been identified including their haplotype for different individuals. As a result, unique variant combinations, including their haplotype, existing in real people, can be used for research. One downside, however, is the lack of phenotype data, which is not readily available for the available genotypes. However, many SNPs do already have direct phenotype associations through different sources, for example ClinVar.[59] Additionally, affected genes that are studied might also have existing literature on their functioning and disease. Even if the variants being studied are not the same, the end result of them might be the same or similar.

2.5 Metabolic Modeling

Models are valuable tools for getting a deeper understanding of functional genomics. There are different types of models, with the two main ones being regulatory networks, and metabolic models. Particularly the metabolic models have been extensively built into Genome-scale metabolic models (GEMs) for multi omics data analyses.[60] Metabolism is an integral part of our biological functioning, and many inherited diseases are metabolic in nature.[61] This could be due to the body not being able to produce metabolites necessary for normal function. Or alternatively, it could be a buildup of intermediate metabolites that the body struggles to deal with, or even a combination of the two. The root cause of such issues is often a protein not being expressed, or a LoF mutation, reducing the function of a protein through primary structure change.[28,62]

The newest GEM for humans is the Human1 by Robinson et. al.[6] To create this model, previous parallel modelling efforts in the Recon and Human Metabolic Reaction series, were unified, including incorporation of new research. This resulted in a model of 13,417 reactions, 4,164 unique metabolites, 10,138 total, and 3,625 genes. This is a generalized model of all metabolic functions possible within a human, and as such this model alone will likely have properties that is not possible in a human due to tissues being specialized. To generate context-specific models, Task-driven Integrative Network Inference for Tissues (tINIT)[63] was used with Genotype-Tissue Expression (GTEx)[64] data. (Figure 5) As the name suggests, tINIT uses (essential) metabolic tasks to validate the model. The concept of these tasks is explored in detail in section 2.7. For now, these tasks represent specific metabolic functions, and in the case of generating tissue specific models, tasks considered essential are used to validate the model. Because they are considered essential, the tissue specific models must be

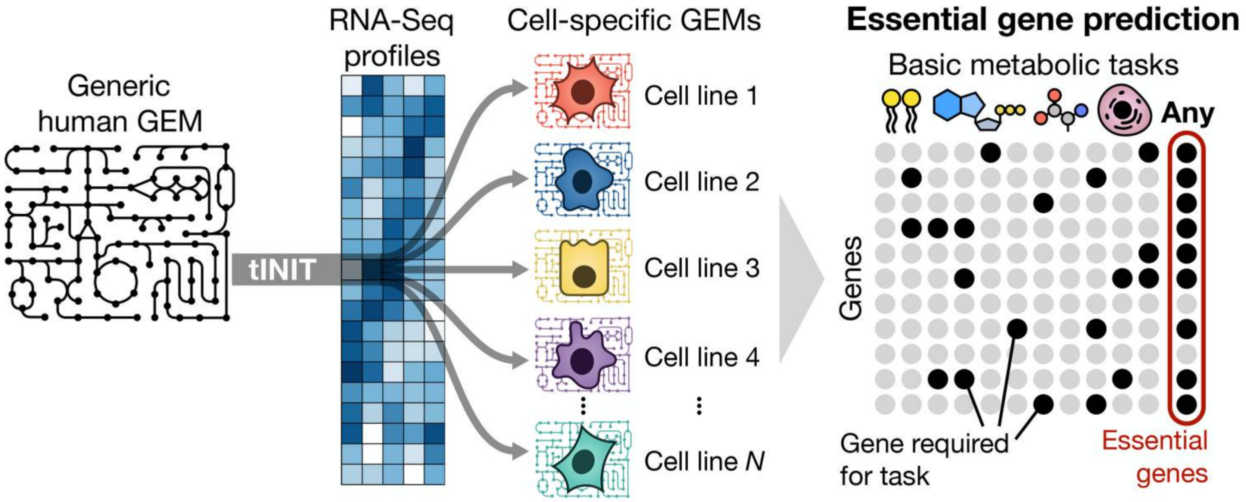


Figure 5: (Modified from J. L. Robinson et. al.[6]) First the generic Human1 metabolic model is used to generate cell specific models through tINIT. tINIT uses RNA-Seq profiles, for example GTEx data, to determine which genes are active in which tissues. Additionally, a number of basic (essential) metabolic tasks are used to ensure the created models have certain essential capabilities. Essential genes for the different tissues are then predicted through the use of essential metabolic tasks, which are also used to validate the models. Different cell types may utilize different genes to accomplish the same task. Not shown here is validation of predictions through CRISPR knockout screens.

able to perform them. One important aspect of these essential tasks is that they allow for a more in dept analysis of essential genes, as any single gene, that when knocked out, causes an essential task to fail, can be considered essential.[65]

2.6 Flux Balance Analysis

Flux balance analysis is a method of analyzing a network, where flow through the network is calculated.[8] It works by treating the network as a mathematical problem and finding the solution that maximizes the flux through an objective reaction. The system is constrained through reactions having a maximum allowed flux, which in turn makes the solution space finite. The analysis then balances, or distributes, flux on various reactions with the propose of maximizing the flux through the objective reaction. (Figure 7A) It is commonly used to study GEMs. When using a GEM, the objective is typically a biomass producing pseudo-reaction, which consumes various metabolites and turns them into a pseudo-metabolite representing biomass. The biomass metabolite represents all metabolites which the cell either needs to get from the environment, or produce from other metabolites it absorbs, in order to grow or maintain itself.[66]

The accuracy of the flux predictions to real organisms is dependent on the constraints imposed on the model. The reactions' flux limits will always be constrained to a certain point, typically an arbitrary number, as well as whether they are reversible or not. When they are constrained with arbitrary (large) values, the model is said to be in an unbound state. In this unbound setup the models are most useful for purely network-based analyses as it cannot give fluxes that are in any way representative of real fluxes.[6]

With humans, or other multi-cellular organisms, properly constrained GEMs are not easy to produce.[6] This is due to humans having multiple tissues, each with their unique model, requiring unique constrains in behavior and in quantity of resources available. Getting the data required for this is incredibly difficult, particularly as human cells do not do well outside the human body. GTEx data and other data sets are available for determining which genes are active in different tissues, but it does not place uptake constraints on the model. Uptakes are generally important flux constraints for constricting a model, as they determine the resources available. This allows for the study of how the behavior of a model changes when different resources are available.

On exception for human cells is various cancer cell-lines, which have been thoroughly studied *in vitro*. Detailed data on different cell lines have been used to create more constrained models, including enzyme constrained models.[6] (Figure 6) These enzyme constrained models have enzymes added as a pseudo-metabolite which is used in reactions, thus putting constraints on how much flux a reaction can have. The constraint is based on enzyme abundance and turnover, and was in the case of Human1, implemented using GECKO.[7] These models are internally restricted, which can be useful for various analyses, especially growth prediction. However, when uptake measurements are available, these internal constraints are less useful. (Figure 6)

A common use of FBA is knockout analysis, where genes are selectively or systematically knocked-out.[9] (Figure 7B) Such an analysis will give varying results depending on the objective and the knocked-out genes relation to the objective. In some cases, there might be no change in outcome. This could be due to redundancy in the genes function, or it is not directly involved in producing flux through the objective. Alternatively, there might be a reduction in objective solution compared to non-perturbed solution. In this case the system is forced to abandon certain pathways, or use alternative, less optimal pathways, to maximize the objective. Finally, the knockout might cause complete breakdown and reduce the solution

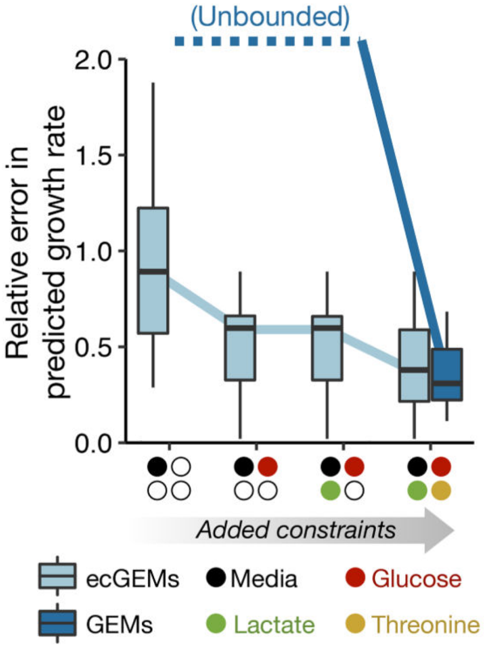


Figure 6: (Modified from J. L. Robinson et. al.[6]) Predicted growth versus added constraints in two set of GEMs. Light blue represents a group of enzyme constrained cancer GEMs. Dark blue are the same GEMs, but unbounded, with the exception of the listed constraints. Boxes represent error rate, with unbound being unquantifiable. Media indicates uptakes were restricted to only nutrients available in Ham's medium, which is used to grow cancer cells. In addition, glucose, lactate, and threonine had measured exchange rates cumulatively added to them. Threonine is one of the essential amino acids, and had its uptake restricted to measured levels. and is alone enough to restrict growth as it is needed in ratio with the other amino acids for protein production. It is clear that cumulative constraints of certain select metabolites can bring an otherwise unbound model to the same growth prediction as an enzyme constrained model.

to zero. In this case, the system has no way of producing flux through the objective reaction. A gene that when knocked out causes a solution of zero might be considered essential if the objective represents an essential function.

Indeed, flux balance analysis lends itself well to the identification of essential genes.[65] A gene is considered essential when loss of its function renders the organism nonviable.[67] In FBA this means when the flux of the objective approaches zero, or deviates sufficiently from optimal, given that the objective can be considered essential. For example, if the typical objective of growth is used, then a gene knockout that produced zero as solution can be considered essential. If that knockout were to occur in the organism, the organism would be non-viable as it cannot grow.

There is also the possibility of uncovering synthetic lethality, which is when a combination of two or more gene knockouts causes cell death, but neither individual knockout is lethal nor essential.[68] This could happen when two genes represent redundant functionality of each other, producing genetic robustness. If they both function in the same metabolic pathway, that pathway may still function with only a single gene knocked out, but with two genes knocked out, it will break down completely.

Related to FBA is the flux variability analysis (FVA).[69] In this analysis one or more reactions are looked at for their ability to vary without changing the objective flux. Essentially, the found solution is not necessarily unique, and there may be a larger solution space that still supports the optimal objective flux. In a knockout analysis, FVA can be used to determine changes to the solution space, even when the objective flux has not changed, which can be useful for investigating underlying issues not directly influencing the objective.

2.7 Metabolic Tasks

Metabolic tasks represents specific metabolic functions, and are used to produce context-specific models.[6,63] When producing tissue specific models it is important that the resulting model is capable of certain functions which are known to occur in that tissue. There are also a number of metabolic functions that can be considered essential to the functioning of a human cell, with few exceptions, which the model must be capable of. Knockout of genes which cause failure of these functions may not always be detected through common biomass optimized FBA. (Figure 8B) This is due to not all essential functions being directly related to growth. Although, biomass production was included in the 57 basic metabolic tasks used to generate tissue-specific models, which means growth cannot be zero without at least one task

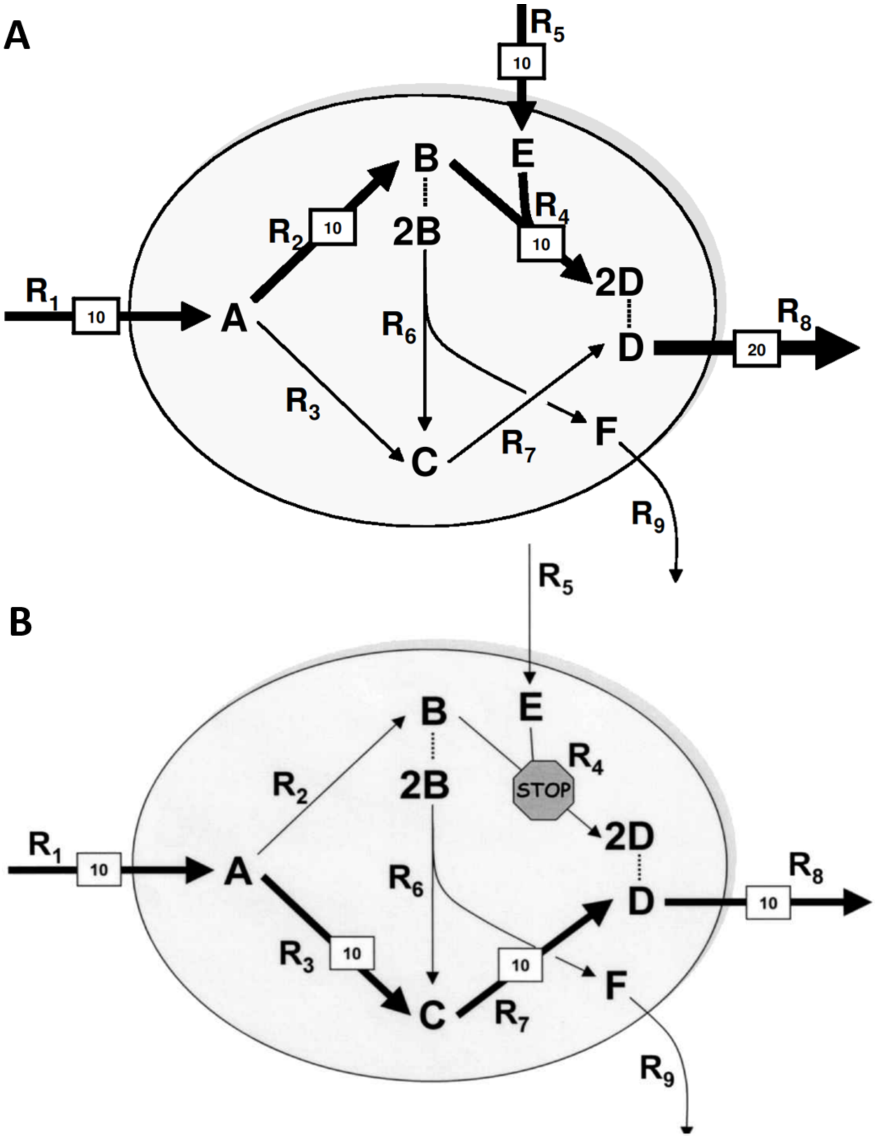
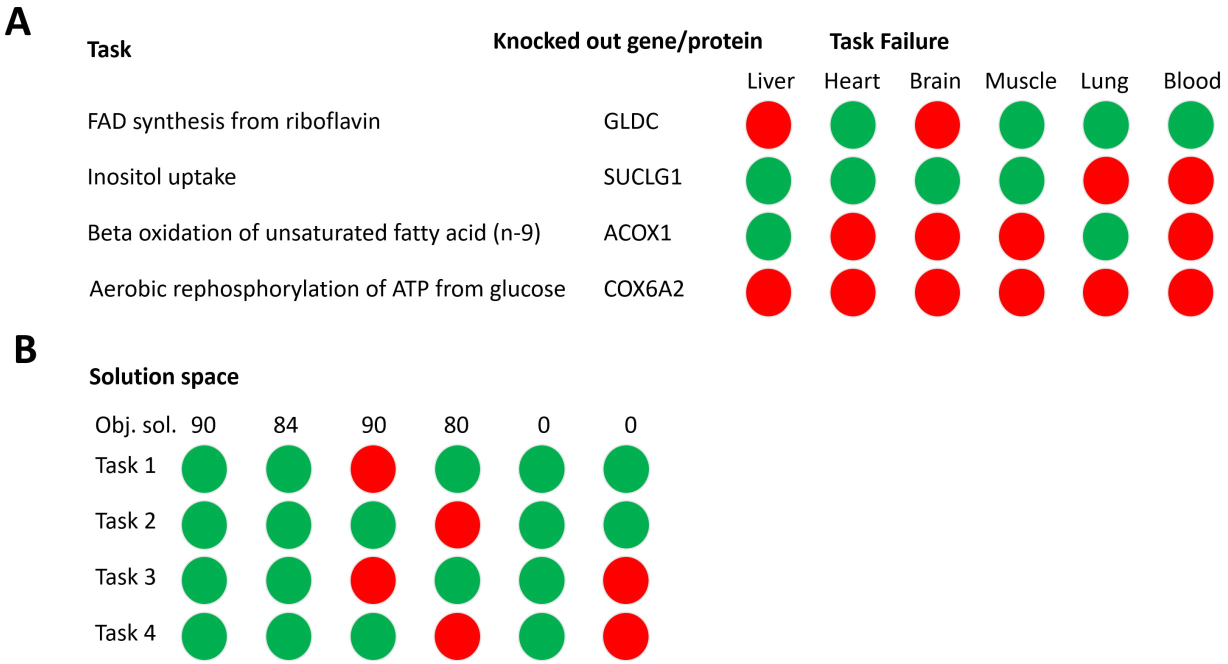


Figure 7: (Modified from D. Segrè et. al.[9]) A simple model for visualization purposes. Reactions are represented by arrows, while metabolites are represented by letters other than R. The thickness of an arrow visualizes flux, which is also shown as a number on the arrows. (**A**) Optimal solution found by FBA, with the objective reaction R_8 being maximized, and input reactions R_1 and R_5 capped at 10. In this case the optimal solution is 20. (**B**) In this case reaction R_4 has been knocked out and can no longer be used by the model. The reason behind this could be a LoF mutation in the gene that codes for the protein that catalyzes R_4. The new optimal solution for the model is now a reduction to 10, with the previous restriction for R_1 still being 10. In this case, alternative possibilities for the same outcome as a knockout of R_4 is if the metabolite E is no longer available or reactions R_2 or R_5 is knocked out. What is apparent is that no single knockout of an internal reaction is enough to bring the solution to zero. A combination of at least two knockouts is needed, for example R_4 and R_7. This shows that the model has a robustness to it, as it can sustain some damage without breaking completely.

failure when using that task set. Due to model differences, different models have different genes considered essential to perform these functions. (Figure 5, 8A) Indeed, knockout of the same gene can cause a task to fail in certain tissues, but not in others.



A

Task	Knocked out gene/protein	Task Failure Liver	Heart	Brain	Muscle	Lung	Blood
FAD synthesis from riboflavin	GLDC	🔴	🟢	🔴	🟢	🟢	🟢
Inositol uptake	SUCLG1	🟢	🟢	🟢	🟢	🔴	🔴
Beta oxidation of unsaturated fatty acid (n-9)	ACOX1	🟢	🔴	🔴	🔴	🟢	🔴
Aerobic rephosphorylation of ATP from glucose	COX6A2	🔴	🔴	🔴	🔴	🔴	🔴

B

Solution space

Obj. sol.	90	84	90	80	0	0
Task 1	🟢	🟢	🔴	🟢	🟢	🟢
Task 2	🟢	🟢	🟢	🔴	🟢	🟢
Task 3	🟢	🟢	🔴	🟢	🟢	🔴
Task 4	🟢	🟢	🟢	🔴	🟢	🔴

Figure 8: (A) Examples of task results with specific gene knockouts for different tissues. Green is pass; red is failure. With certain knockouts some tissues pass a specific task, while others fail. As a result of this, a gene which is essential in one tissue, is nonessential in another. **(B)** Different possible solution categories for running FBA with a number of tasks. For each column a different gene (-combination) was knocked out and run through FBA. If only essential tasks are used, it is enough that any one task fails for the cell to be considered non-viable. This gives six possible result categories. For the tasks, either any one task fails, or all passes. For the objective solution, it is either max, 90 in this case, reduced, or at zero.

Usually, the solver will be free to add or remove metabolites in the extracellular environment however it finds optimally. In this state the model is said to be open. However, when checking if the model is capable of performing specific functions, functions it might not otherwise use during FBA, the model is put in a closed format. When FBA is performed, only specific metabolites will be allowed in or out of the system. (Figure 9) The cell will in most cases need to extract or expel the metabolites from and to the extracellular environment, although, in some tasks, metabolites will be available for addition or removal directly from cytosol or organelles. This allows for testing the production of a metabolite without requiring the cell to be able to excrete it. Some, but not all, of the tasks also have a reaction that get added

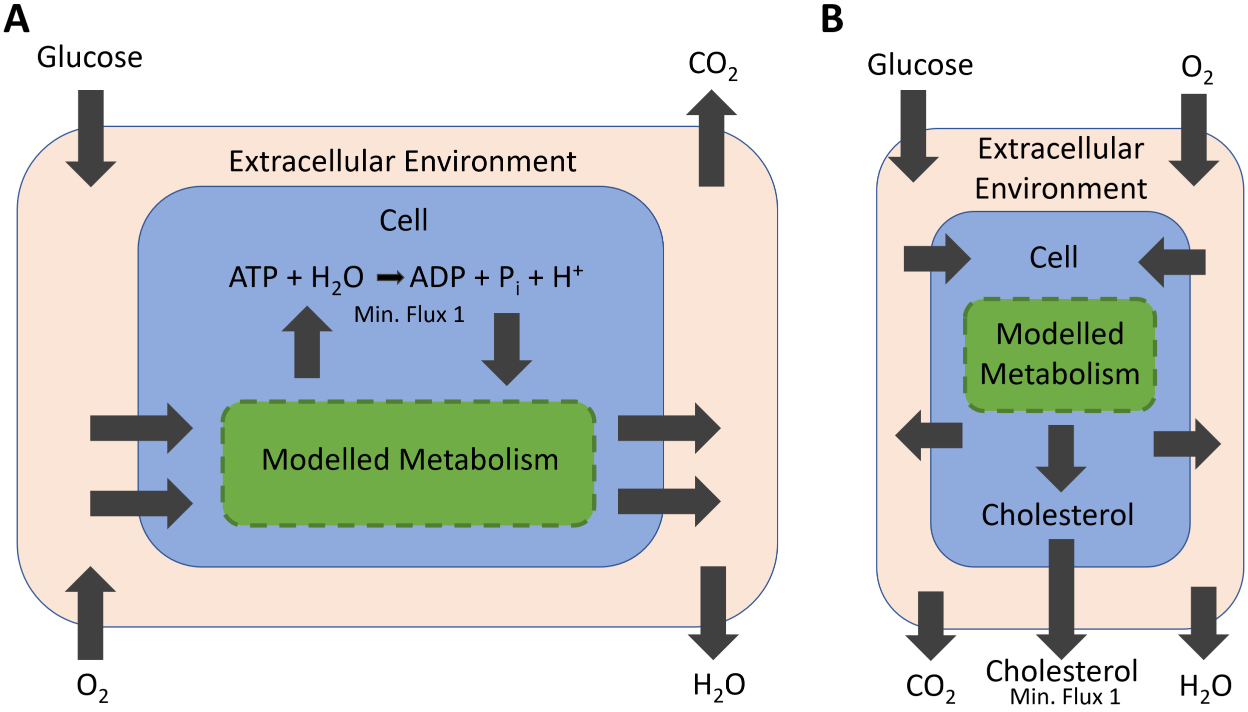


Figure 9: Two metabolic task examples: One with an added metabolic reaction (A) and one without (B). In this case, as both tasks are considered essential to most cells, failure means the model cell is non-viable. **(A)** The essential metabolic function of aerobic phosphorylation of ATP from glucose and oxygen. In this task only glucose and oxygen can be consumed by the system, and only water and carbon dioxide can be removed from the system. In addition, a reaction for consumption of ATP, with a minimum flux of 1, is added. This forces the model to produce ATP from only inputs of glucose and oxygen. The added reaction ensures that there is mass balance as all additional metabolites consumed to produce ATP will be released in the added reaction, essentially creating a circle of ATP production and degradation, fueled by glucose and oxygen. Because only carbon dioxide and water is allowed to exit, complete aerobic breakdown is necessary to achieve the required flux. **(B)** Cholesterol *de novo* synthesis. In this essential metabolic task, the model is tasked with producing cholesterol, using only glucose and oxygen as inputs. The only allowed outputs are cholesterol, carbon dioxide and water. In this task there is no added metabolic reaction. Instead, the model will produce cholesterol and an added sink reaction will remove it directly from cytosol. The purpose of this task is to check whether cholesterol can be produced, and not necessarily if it can be excreted as well. This reaction does have minimum flux of 1, forcing cholesterol production.

to the system. The purpose of adding a metabolic reaction when it is done, is to allow for stoichiometric balance when performing FBA on the model. A minimum of one input or output, or the reaction, will have a minimum flux associated with them. This forces the model to handle that flux, and if it cannot, it is unsolvable, as it will have no valid solution. The major benefit of utilizing these tasks is that they add additional constraints without

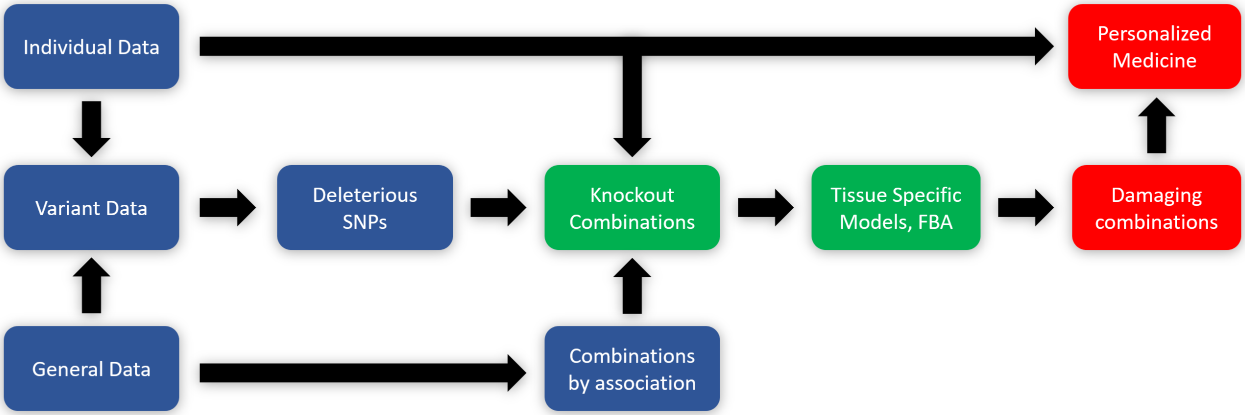


Figure 10: General overview of the approach taken. Variant data is collected and filtered for (potentially) deleterious SNPs. These SNPs as well as individual data or otherwise associated SNPs are used to produce gene combinations for use as knockouts in modeling. Any damaging combinations that are found may be used to predict effects in individuals.

needing detailed exchange and/or enzyme data. As a downside, each task requires a separate FBA run. An alternative approach might be to select a certain reaction as the objective and require a minimum flux, but this has the downside of not having specified allowed inputs and outputs. In such a scenario the system might simply take in a metabolite it would otherwise not be able to produce, or alternatively excrete a metabolite it is otherwise not able to break down.

3 Methods

3.1 Overview

To achieve the goal of using variation/SNP data in metabolic modelling a pipeline was constructed. (Figure 10) This pipeline takes in SNP data, filters it, then uses chosen data to perform FBA with deletions of affected genes. (Figure 11) Everything has been coded in Python and the pipeline has been constructed in a modular way, allowing for easy change of input data and metabolic model. The pipeline has two slightly different modes of function depending on whether general or individual data is used.

As reference genome data, this project used was from Ensembl, specifically all exons from GRCh38.p13.[70] No alternative scaffolds were included. Only 'Ensembl Canonical' transcripts for protein coding genes were included. The genome data was filtered to include only

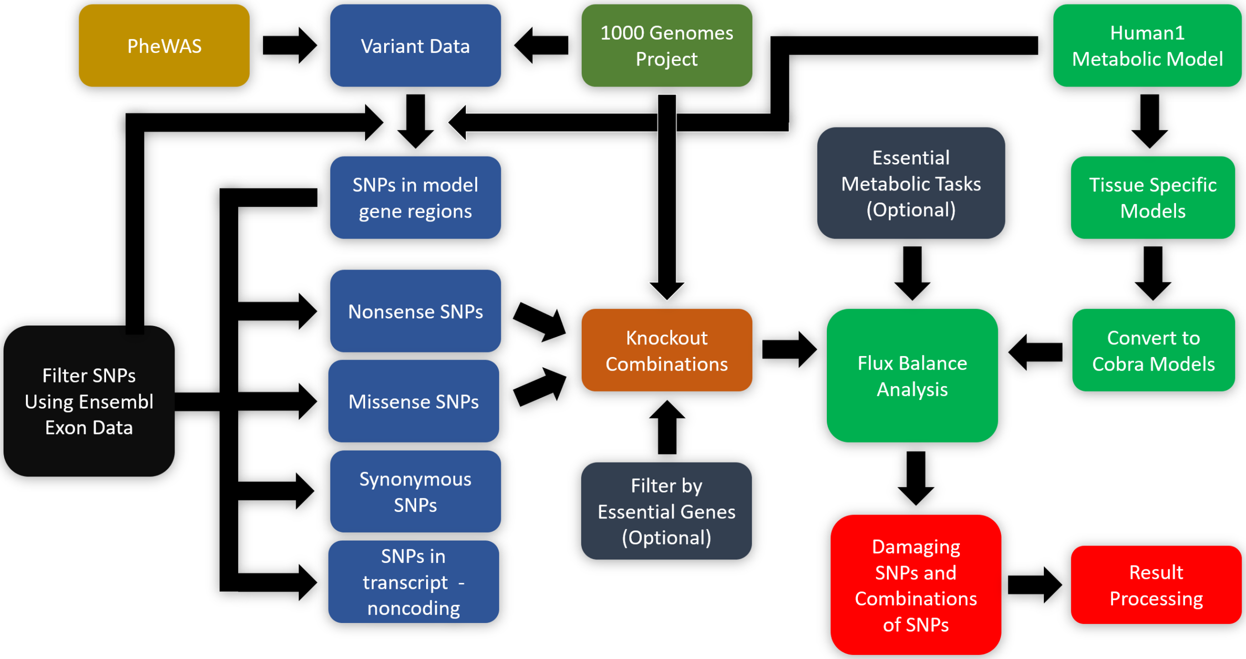


Figure 11: A more detailed overview of the pipeline, and the data and inputs used. The SNPs are divided and filtered based on gene location and effect. A list of genes present in the model is also used to filter the SNPs. Nonsense and missense SNPs are used for gene knockout combinations. Not shown is the use of PheWAS combinations for knockouts. Optionally the combinations can be filtered for essential genes they may contain when running FBA. FBA is run on Human1 tissue specific models. Included in this, and also optional, is the use of metabolic tasks. Any non-optimal results may then be further investigated for the details as to why the result is non-optimal.

genes also present in the model. The general SNP data used was form the PheWAS catalogue[46], which was used to generate gene combinations for knockout grouped by phecodes. As for the individual variant data used, it was from 2548 individuals of the 1000 Genomes Project.[13,58] Relevant SNPs were extracted and used to generate gene combinations for knockout based on individual samples. These were SNPs thought to have a higher probability of causing LoF, and this included premature stop codons, but also start codon loss, and loss of WT stop codon. This group of SNPs will be generally referred to as start/stop SNPs in the remaining text. Distinctions between homozygote and heterozygote SNPs was also used in generating gene combinations for individuals. The combinations were then used for knockout FBA, including a number of tasks, on a variety of tissues. In general, the primary emphasis of the project was on the use of individual data. PheWAS data was included primarily as a showcase for an alternative use of the developed method.

The model used for this project was Human1 version 1.10.0. Specifically, a number of tissue specific models created and published by Robinson et al. using tINIT and GTEx data.[6] A set of 57 metabolic tasks considered essential was used with tINIT to produce these models. These tasks were also used when performing knockout analysis in this project. They are considered essential in that the function they test is considered essential to the viability of human cells, with few exceptions. However, the nine tasks pertaining to essential amino acid uptake were not included. The reasoning behind this being that they were not found to ever fail, and thus to save time, were excluded. An additional set of nonessential tasks was also included after being tailored for each different tissue due to their different capabilities. Additionally, considering that many results were expected to be reoccurring in most tissues, not all 30 GTEx tissue models were used for FBA. Instead, 15 of them, in part arbitrarily but including all major organs, were selected. (Table 1) Again this was to save time on FBA, as 15 models, including major organs were considered to provide wide enough coverage for the purposes of this project. However, for some analyses that did not include FBA, all 30 tissue models were included.

Additional analyses, as well as more detail on the pipeline as a whole is presented in the two sections below. A number of data files is included in the supplementary data folder. All code is available at **https://github.com/SigveLan/Master**. In the Github repository the main 'README' file will contain details on how to use the code produced. This includes which specific scripts or functions are used at each step. Specific sections of the README will be referenced below as to point to the exact code used. Additionally, the code is generally commented with a main comment of what the script or function does, as well as comments

Table 1: The 15 GTEx tissue models used for FBA.

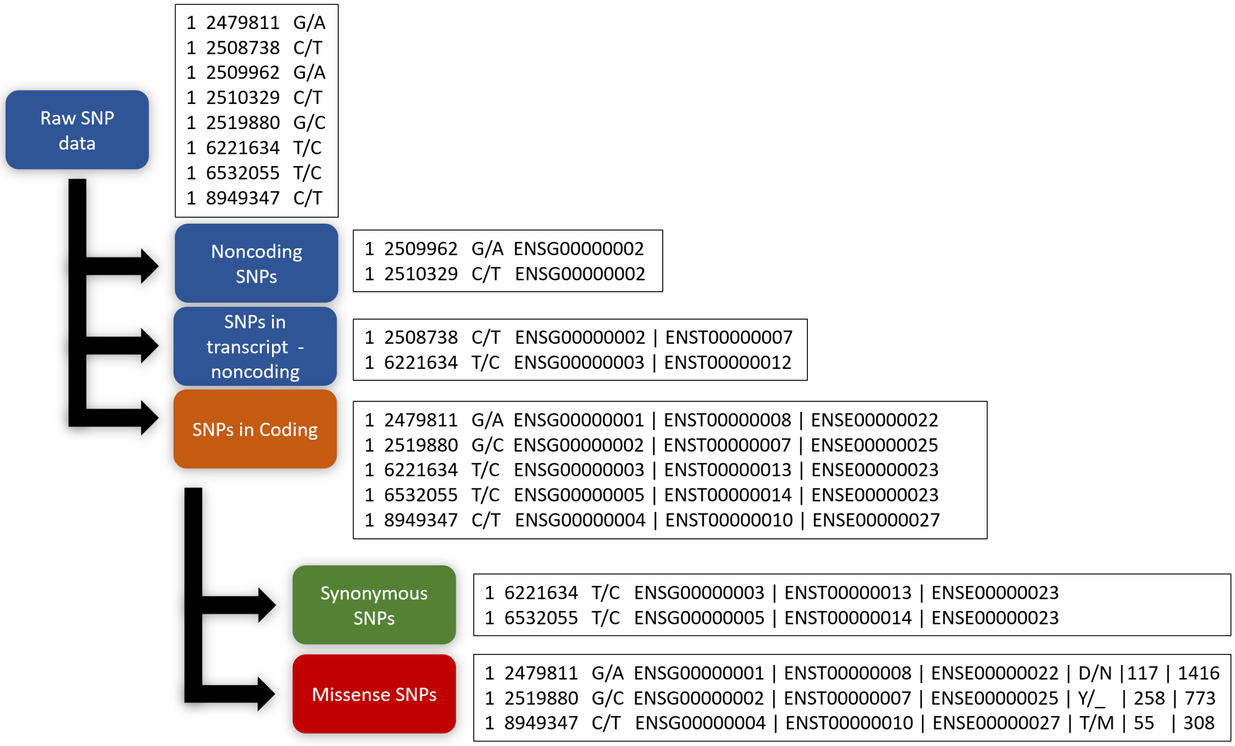
Tissue
Adipose Tissue
Adrenal Gland
Blood
Brain
Heart
Kidney
Liver
Lung
Muscle
Nerve
Pancreas
Pituitary
Skin
Spleen
Thyroid

highlighting settings that can be used, and in some cases, the functioning of the code.

3.2 SNP Filter

3.2.1 SNP Assessment Pipeline

This pipeline takes in lists of SNPs and assesses them based on location and effect. The two methods used in this project is outlined in the sections below, although any list of SNPs with the correct format may be used. The pipeline does support multiple transcripts for each gene, as well as non-protein coding genes/transcripts. To save time and computing resources, the genome data is first filtered by the genes in the metabolic model used for the modeling part. This significantly reduced the amount of genome data the program has to look through. First, the locations of the SNPs are cross referenced with genome data, which divides the SNPs into three categories, depending on location. (Figure 12) First, SNPs inside a gene region, but not inside a transcript. Second, SNPs inside a transcript, but not in a protein coding section. And finally, SNPs inside the protein coding part of the transcript. SNPs in the latter category is then evaluated for any potential effect they may have on the protein, primarily amino acid sequence change. To do this the coding sequence is translated. Evaluated SNPs are here divided into two categories, synonymous and missense SNPs. Included in the output for the missense SNPs is the amino acid change, the position of the affected amino acid, and finally the total length of the protein. This step is accomplish by running a single script as described in the Github README under 'Running the SNP filter'.



```
1  2479811  G/A
1  2508738  C/T
1  2509962  G/A
1  2510329  C/T
1  2519880  G/C
1  6221634  T/C
1  6532055  T/C
1  8949347  C/T
```

```
1  2509962  G/A  ENSG00000002
1  2510329  C/T  ENSG00000002
```

```
1  2508738  C/T  ENSG00000002 | ENST00000007
1  6221634  T/C  ENSG00000003 | ENST00000012
```

```
1  2479811  G/A  ENSG00000001 | ENST00000008 | ENSE00000022
1  2519880  G/C  ENSG00000002 | ENST00000007 | ENSE00000025
1  6221634  T/C  ENSG00000003 | ENST00000013 | ENSE00000023
1  6532055  T/C  ENSG00000005 | ENST00000014 | ENSE00000023
1  8949347  C/T  ENSG00000004 | ENST00000010 | ENSE00000027
```

```
1  6221634  T/C  ENSG00000003 | ENST00000013 | ENSE00000023
1  6532055  T/C  ENSG00000005 | ENST00000014 | ENSE00000023
```

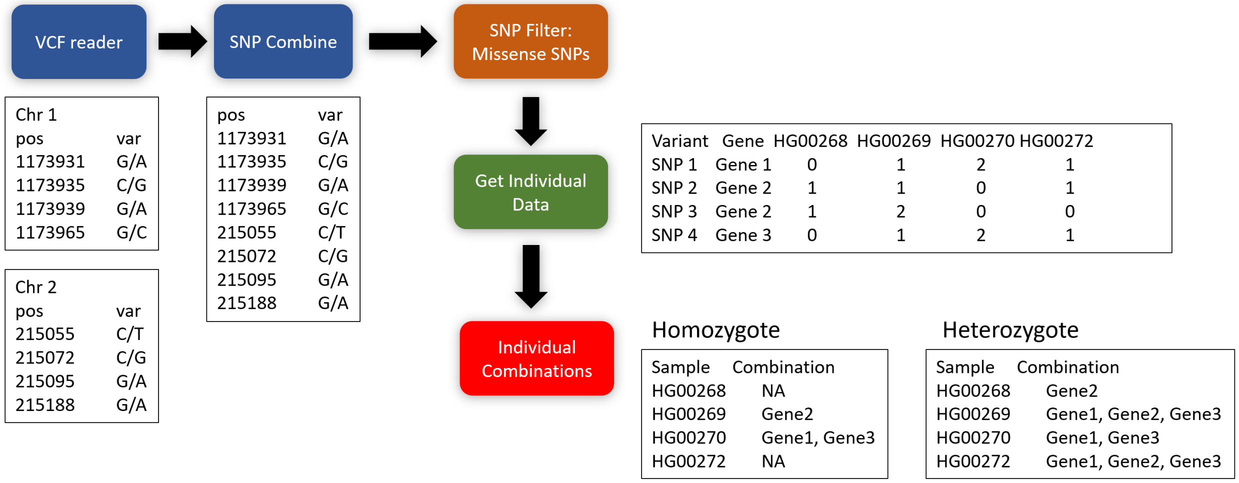
```
1  2479811  G/A  ENSG00000001 | ENST00000008 | ENSE00000022 | D/N |117 | 1416
1  2519880  G/C  ENSG00000002 | ENST00000007 | ENSE00000025 | Y/_  | 258 | 773
1  8949347  C/T  ENSG00000004 | ENST00000010 | ENSE00000027 | T/M | 55   | 308
```

Figure 12: Details of how the SNP filter functions, including output examples. The raw SNP data should be a long list where each SNP has a chromosome, position, and base change listed. Not shown here is the SNP ids, such as a rsIDs if they are available. Raw data is divided into three categories (files) depending on location. All SNPs that are found to be within a gene have the relevant Ensembl gene ID added. For the 'SNPs in transcript - noncoding' and 'SNPs in coding' the relevant transcript ID is added. If there are multiple transcripts, they will be added as a list. SNPs in coding also have their exon ID added. Exons are typically not overlapping, but if they are the SNP will get listed multiple times for the different transcripts, they will also have to be in. SNPs located within coding regions are then further scrutinized. The relevant codon the SNP is affecting is acquired and translated. If there is no change in amino acid, the SNP is synonymous and no further information is added. If the SNP is found to change the amino acid, that information is also added. Additionally, the position of the changed amino acid is included, as well as the total length of the protein. If a SNP is affecting multiple transcripts it will at shis point get listed multiple times, along with the potentially different amino acid and position data. The use of the SNP filter is described in the Github README under 'Running the SNP filter'.

Due to SNP lists potentially being quite large, multiprocessing has been implemented for the first step to increase filter speed. Exon sequence data was chosen because they are used in multiple transcripts, thus there would not be large amounts of overlapping genome data to process. However, this only manifest itself if multiple transcripts are used, and even then, it might not matter much due to only genes of the model being used, substantially reducing the sequence data amount. One quirk of these exons is that a few have multiple coding start or end points depending on if it is the first or last exon in the transcript. To resolve this issue, the coding point that made the exon, and subsequent transcript, the largest, was chosen. The missense output also includes as 'score', with there being a scoring function in the code. The purpose of this was to eventually add a function to score the missense SNP based on how likely it was that it caused a deleterious change in the protein. As an addition to the scoring, there is an option for translating the whole amino acid sequence for the proteins. While this functionality opens up possibilities of future expansion, they have not been used to produce any results at this time.

3.2.2 VCF Extraction and Individual Data and Combinations

To extract relevant SNPs from the 1000 Genomes data, two scripts for reading data from variant call format (VCF) files were developed. The first script extracts all SNPs by filtering them based on which SNPs are located within the genes of the model. (Figure 13) The VCF data is divided into chromosomes and one SNP list file is produced for each of them. This extraction also has the option of dividing the SNPs for each chromosome into multiple files if necessary due to limited computer resources, as this process can be RAM intensive. However, for the 1000 Genomes data this was not necessary. The second script is to be used after running the SNP filter. Then selected SNPs are used to extract which samples contain them, and whether they are hetero- or homozygote for the SNP. In this project only missense SNPs were used for this. The individual data generated can then be used to produce gene combinations for each individual/sample. Included in here is the option of using only homozygotes, heterozygotes or both, as well as further filtering of the SNPs. SNPs can be filtered by any attributed included in the SNP filter results. In this project primarily start/stop SNPs were used, although the use of all missense SNPs was also included as a showcase. The resulting gene combinations can then be used knockout FBA. If several combinations are found to be identical, they are combined into one combination that represents multiple samples, which prevents overlapping FBA runs. The concrete use of this VCF extraction and individual combination generation pipeline is described in the Github README under 'VCF extraction'.



Chr 1	
pos	var
1173931	G/A
1173935	C/G
1173939	G/A
1173965	G/C

Chr 2	
pos	var
215055	C/T
215072	C/G
215095	G/A
215188	G/A

pos	var
1173931	G/A
1173935	C/G
1173939	G/A
1173965	G/C
215055	C/T
215072	C/G
215095	G/A
215188	G/A

Variant	Gene	HG00268	HG00269	HG00270	HG00272
SNP 1	Gene 1	0	1	2	1
SNP 2	Gene 2	1	1	0	1
SNP 3	Gene 2	1	2	0	0
SNP 4	Gene 3	0	1	2	1

Homozygote

Sample	Combination
HG00268	NA
HG00269	Gene2
HG00270	Gene1, Gene3
HG00272	NA

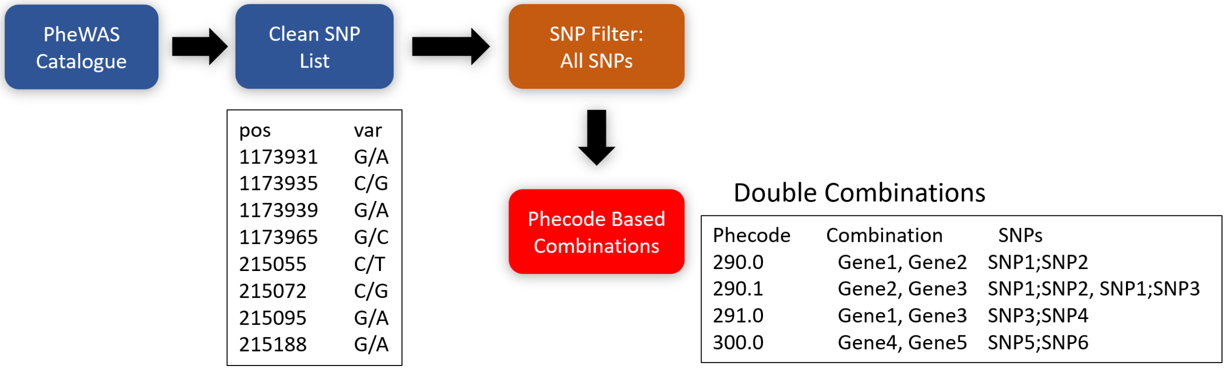
Heterozygote

Sample	Combination
HG00268	Gene2
HG00269	Gene1, Gene2, Gene3
HG00270	Gene1, Gene3
HG00272	Gene1, Gene2, Gene3

Figure 13: Data flow for extracting individual SNP data from the 1000 Genomes Project. At the first step SNPs are extracted from each VCF file, with each file being one chromosome. Applied to this is the filtered exon genome data (same as first step of the SNP filter) to get only SNPs that are in the genes of the model. This step also has the option of creating multiple files for each chromosome. Then the extracted SNPs are combined into a single file for ease of use. The extracted SNPs are then run through the SNP filter. The resulting missense SNPs are then used together with the original VCF files to produce individual allele data. Zero means SNP not present, 1 means one copy, and two means the sample is biallelic for the given SNP. Included in this data but not shown is the full SNP data as provided by the filter. This step can also use other SNP filter results than the missense SNPs. Next individual gene combinations are produced for use in knockout FBA. There are several options here, such as which SNPs to use, and whether the samples should be homozygote or heterozygote. In this case homozygote samples were also used in heterozygote combinations, but that can be adjusted. SNPs are filtered by their properties. The use of the VCF extraction and individual combination generation pipeline is described in the Github README under 'VCF extraction'.

3.2.3 PheWAS Data and Combinations

The extraction and combination generation for PheWAS data is similar to that of individual data. A cleaned-up list that contained only the SNP data was first produced from the phewas-catalogue. (Figure 14) The SNPs are then filter through the SNP filter. Selected SNP outputs, is then used to create combinations of given sizes, based on phecodes. The phecodes are extracted from the phewas-catalogue. Due to the small number of SNPs, options for use of additional SNP filter outputs beyond missense SNPs was added, and used in this project. Specifically, all SNPs were used even if they were not in the coding region or were synonymous. All SNPs found for a given phecode will be included in the gene combinations set for that phecode. The size of the set can be specified. In this project a combination size of 2 was primarily used. Due to generation of larger combinations being slow, multiprocessing was implemented. These gene combinations can then be used for knockout FBA. The processing of PheWAS SNPs to gene combinations is described in the Github README under 'PheWAS extraction'. A general approach nearly identical to PheWAS, but that only has a simple list of SNPs, was also included. This approach does not rely on phecodes, but rather uses the whole SNP list given to generate combinations. No combinations generated in this manner was used to produce results. Description of this process is listed in the Github README under 'Combination preparation using general SNP data'.



pos	var
1173931	G/A
1173935	C/G
1173939	G/A
1173965	G/C
215055	C/T
215072	C/G
215095	G/A
215188	G/A

Phecode	Combination	SNPs
290.0	Gene1, Gene2	SNP1;SNP2
290.1	Gene2, Gene3	SNP1;SNP2, SNP1;SNP3
291.0	Gene1, Gene3	SNP3;SNP4
300.0	Gene4, Gene5	SNP5;SNP6

Figure 14: First a clean list of SNPs needs to be created from the PheWAS-catalogue. This list is then run through the SNP filter. In this case all SNPs found to be within a model gene were used, due to the low number of SNPs in total. Finally, combinations are generated based on all SNPs listed for a given phecode. The size of the combinations is adjustable. In some cases, multiple SNP combinations led to the same gene combinations, all of which were included. The processing of PheWAS SNPs to gene combinations is described in the Github README under 'PheWAS extraction'.

3.3 Metabolic Modeling

3.3.1 Context-Specific and Task Configured Model Preparation

These models were converted to Cobra compatible versions[71] from Raven version[72] in MATLAB using Raven. Additionally, the use of metabolic tasks was previously only available in MATLAB through RAVEN, but was implemented in Python for the purposes of this project. This was to allow for easier use of multiprocessing, and easier use in general, when utilizing metabolic tasks for knockout analysis. Because metabolic models are set up slightly differently in Cobra compared to RAVEN, a process to emulate the RAVEN setup was developed. (Figure 15) Models in Cobra are in an open format, but to utilize metabolic tasks, a closed format is needed. To produce models in a closed format, boundary metabolites were added to all exchange reactions. This prevents the model from adding or removing any extracellular metabolite. Then, for any given metabolic task, exchange reactions with constraints were added for all given metabolites. This ensures the model is only able to exchange metabolites as dictated by the task. Support for one special case was also added, the "ALLMETSIN" keyword, which when placed as an input or output, dictates that all exchange metabolites be allowed into or out of the model respectively. As an example, this is used for the "Growth on Ham's media essential task" as an output, which allows the model to output any metabolite it would have been able to output in an open format. The inputs are however restricted to certain metabolites, based on Ham's media. Comparisons with MATLAB results were made to validate the correct functioning of the metabolic tasks. The process of reading in tasks and producing bound models for their use in FBA is described in the Github README under 'Task Functions'.

3.3.2 Preparation of Additional Metabolic Tasks

Additionally, an expanded, but heavily filtered list of 256 additional metabolic tasks were used for knockout analysis, also from Robinson et al.[6] (Figure 16) The original 256 tasks also included a number of tasks that were already represented in the essential tasks list. These were removed as to not have any overlap that would create unnecessary clutter and FBA runs. The resulting list of tasks are not considered essential, and in general, any given tissue will not be capable of performing all of them. Nevertheless, if a knockout causes a previously functioning metabolic function in a tissue to fail, it could be indicative of a problem. The full list also included tasks that were supposed to fail, which have also not been included here. Those tasks exist to check whether a model behaves as it should, without abnormal metabolic functions that should not be possible. Performing knockout would not induce such behaviors and as such these tasks are unnecessary for this project. Additionally, one task, 'storage of glucose in glycogen', had attributes that were not supported, and was thus removed. The final filtering was done on a tissue-by-tissue basis, with the first step in creating a tissue specific list being whether the metabolites for a task was present in the tissue. Not all tissue models had the metabolites required for a given task, and thus were not included for that tissue. The next step was to determine which tasks each model could pass in an unperturbed state. All tasks were run on all tissue models. Any task that failed was removed for that model. The resulting lists of tasks were then used in knockout analysis in the same manner as the essential tasks. The process of generating the additional task lists is described in the

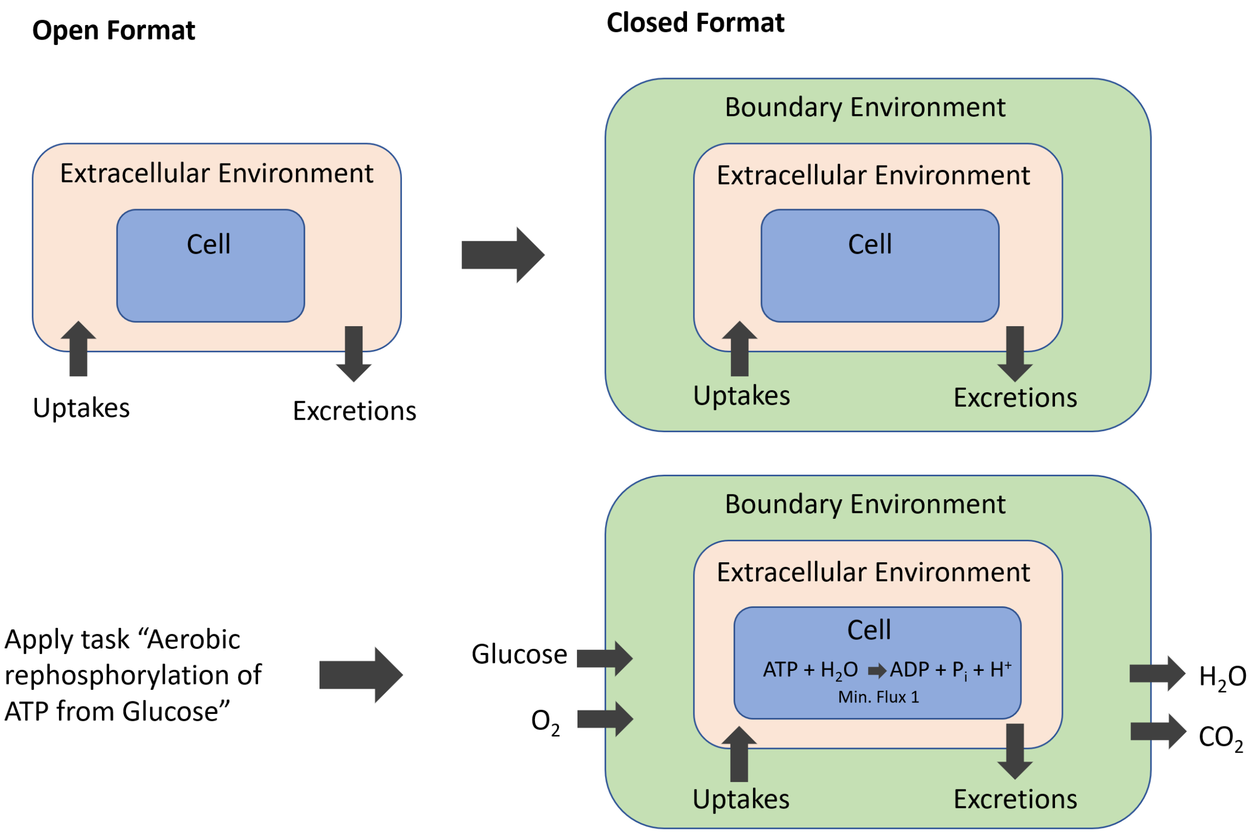

Open Format
Closed Format
Extracellular Environment
Cell
Uptakes
Excretions
Boundary Environment
Extracellular Environment
Cell
Uptakes
Excretions
Apply task “Aerobic rephosphorylation of ATP from Glucose”
Boundary Environment
Extracellular Environment
Cell
ATP + H2O → ADP + Pi + H+
Min. Flux 1
Glucose
O2
H2O
CO2
Uptakes
Excretions

Figure 15: The model in its standard configuration and task ready configuration. To configure the model for use of metabolic tasks, boundary metabolites were added, preventing the model from exchanging any extracellular metabolite at will. Then when a task is implemented, exchanges for the specified metabolites are added. The added exchanges may be added for any environment of the model, it can be directly to or from cytosol or the mitochondria. In this case an internal metabolic reaction is also added, to test the functioning of the cell. The process of reading in tasks and producing bound models for their use in FBA is described in the Github README under ’Task Functions’.

Github README under ’Additional Task Lists Creation’.

Another possibility for use of metabolic tasks is if any additional checks are wanted, a task can be set up for that specific purpose. Using the Metabolic Atlas[6] as a guide, one such task was created for valine degradation. To check the whether the task behaved as expected, a number of genes from in and around the pathway was used for knockout analysis. The task in question is found in the supplementary data.

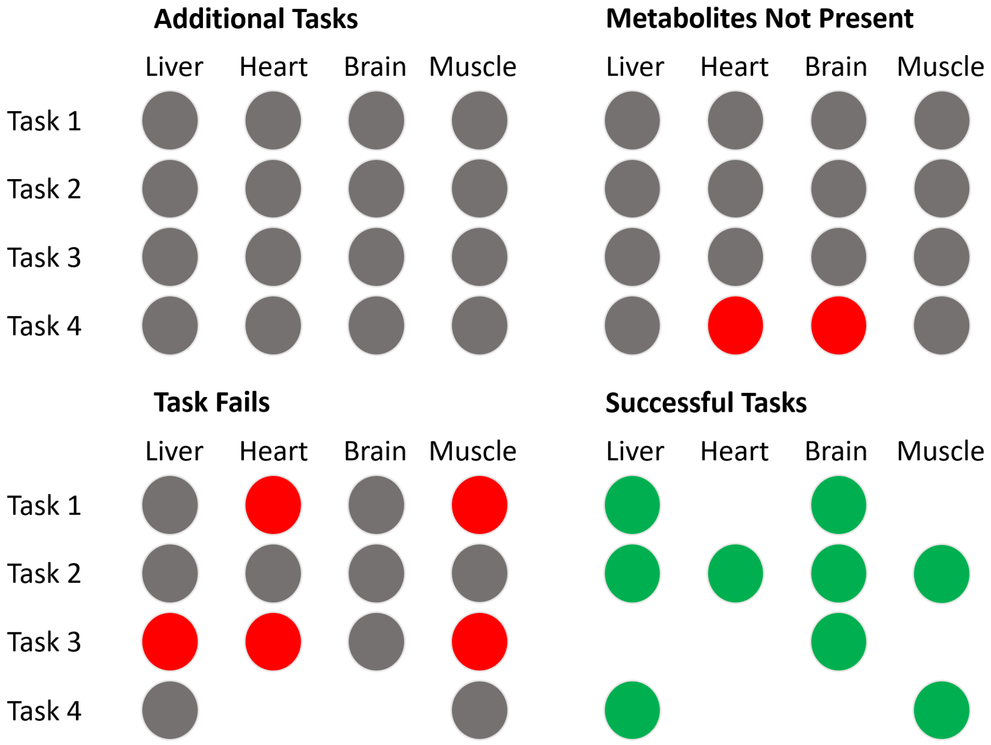


Figure 16: A visual representation for the process of assessing additional metabolic tasks for use in FBA, full list provided by Robinson et. al.[6] **Additional Tasks:** The full list of tasks, not including essential tasks or task that were suppose to fail. **Metabolites Not Present:** Tissues have varying capability in producing and absorbing metabolites. Some tissues will thus not be able to perform certain tasks as they simply do not have the necessary metabolites available. **Task Failures:** The tissue is not able to process the inputs to produce a specific output. Even if a tissue has access to the necessary metabolites, it may not contain all the metabolic functions necessary to perform the task. **Successful Tasks:** These are the tasks found to pass in the various tissues. Due to the variance in tissue capabilities, separate files have been created for each studied tissue model. The process of generating the additional task lists is described in the Github README under 'Additional Task Lists Creation'.

3.3.3 Essential Genes

The definition for essential genes in this project has been set to any singular gene that when knocked out causes any one essential task to fail. Through the task 'Growth on Ham's media' this includes any knockout that causes biomass production to be reduced to zero. Due to differences in the models, they will not have all the same essential genes, and as such a separate list is required for each tissue. RAVEN was used to acquire a list of essential genes for each tissue. An alternative approach to generate a list beforehand, is to use Cobra to find all genes which when knocked out cause zero biomass production. In addition, for use with FBA, analyses were also done on the number essential genes compared to total number of genes for the different models. This was done to highlight tissue model differences, and how they might impact SNP based FBA analysis.

One additional analysis that was undertaken surrounding essential genes and the individual genome data, was how different SNP categories were distributed between essential, and nonessential genes of the generic Human1 model. The four categories were all missense or start/stop SNPs, for both heterozygote and homozygote samples. Number of SNPs as well as the percentage of SNPs in essential and nonessential genes for each category was included.

3.3.4 Flux Balance Analysis

The FBA has been set up to allow for a few different options. Once the gene combinations have been selected, there is the option for which models to use. As many tissues as available can be selected, although, one should be aware of time requirements. Then there is the option to filter the combinations of essential genes. That way, genes that are known to cause non-optimal results will not mask any other results if they are present. If this is not selected, the combinations will still be filtered by the genes that are actually present in the model. If FBA is done with metabolic tasks, there is the option of which tasks to use. The primary options here being universal essential tasks or tissue specific lists of non-essential tasks. In addition, to allow proper system utilization, an option for number of CPU threads to use is also available. Due to the use of 'simple FBA' where only objective solutions are stored, the RAM requirements are relatively low as well. The results produced contain the sample IDs for individual genome data, or alternatively the phecode associated with the gene combination. The objective solution is also included, as well as if any tasks failed, which is presented as a list of 1's and 0's for passed or failed task respectively. Finally, the gene combination used for that run is included. These results will then be printed to file on a tissue-by-tissue basis. Included at the top is also a reference run, where the model was run unperturbed. The process and options when performing FBA is described in the Github README under 'FBA'.

3.3.5 Result Processing

Once FBA output files have been produced, they can be filtered for non-optimal results. Any extracted results with task failures can also be additionally processed to make it clear

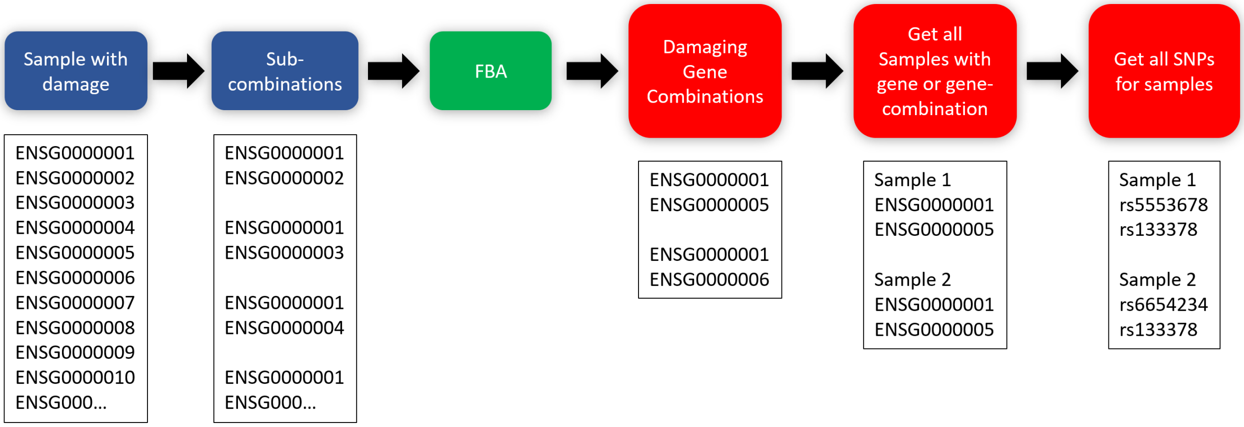


Figure 17: Process to find which genes caused a given FBA result out of a larger combination. Then all samples carrying that gene or gene combination is identified. Finally, the underlying SNPs that produced the gene combination is identified for each sample. Different samples may have different SNPs giving rise to the same gene combination, additionally there may be multiple SNPs for one gene in one sample. The result processing is described in the Github README under 'Result Processing'.

through the task description which tasks failed for which samples/combinations. The result processing is described in the Github README under 'Result Processing'.

In many of the gene deletion combinations found to reduce the biomass output or cause any task to fail, only a small subset or even a single gene, caused the apparent damage. To uncover which specific genes or gene combinations were the cause of the FBA results, sub-combinations of the initial combination were generated and used for FBA. (Figure 17) Finally, the deletions found to cause the non-optimal results were tied back to specific SNPs. In the case of individual data, this additionally included SNPs that were not originally selected for, but are nevertheless present in the individual. For general/PheWAS data, this process was not needed, as the combinations run were small, with generated combinations having a maximum size of 5 genes, although primarily a combination size of two was used.

Once the specific SNPs had been identified, they were further looked at through MutationTaster[73] and ClinVar[59]. MutationTaster was used for checking the direct effects of the SNP on the transcript, whether it might cause LoF or not. ClinVar on the other hand was used to check if a specific SNP has been linked to any diseases. These checks were carried out in order to help confirm whether or not the approach taken could be deemed successful in predicting disease.

3.3.6 Flux Variability Analysis

As a showcase, FVA was performed in order to get a better understanding of what specifically was happening in the model with regards to a specific gene knockout. To help with this, additional constraints were also applied as they were seen fit, in an attempt to prevent the model from simply excreting intermediate metabolites. In part, this was done to get a better understanding of model behavior in general, as well as help uncover exact effects of gene knockouts. The process of using FVA and some other model exploration/testing is described in the Github README under 'FVA and Model Exploration'.

4 Results and Analysis

4.1 The Models: Essential Genes and Metabolic Tasks

While the generic Human1 model contains 3625 genes, the 30 GTEx tissue models were found to only contain a combined total of 3007 genes. This means that there are 623 genes not present in any of the tissue models, and subsequently not used in any gene combination knockouts. In general, it was found that the tissue models had a higher percentage of essential genes, the fewer total genes the model had. (Figure 18) All the tissue models had more essential genes than the generic Human1 model, which in comparison, had 234 essential genes, equating to 6.5% of the genes. All specific genes, their Ensembl ID, and any specific SNPs along with them, mentioned in the results section are listed in the Appendix: Table A.

For the additional metabolic tasks, there were at total of 194 tasks added, with each individual tissue getting around 182. Certain tissues had more, such as the liver with 189. While others had fewer, such as blood with 176, and spleen with 178. The added tasks mostly overlapped between the tissues, with 173 tasks being common to all 15 tissues this was done for.

4.2 Individual Genome Data

For the initial filtering of the 1000 Genomes individual data, 16.5 million SNPs were found to be within the genes found in the generic Human1 model. A total of 121 553 of these were found to be missense SNPs. Of these, 2600 SNPs affected the initiator or terminator codon, or induced a premature terminator codon. 1306 genes were affected by at least one of these SNPs, out of the 3625 genes in the generic model. Including all missense SNPs, a total of

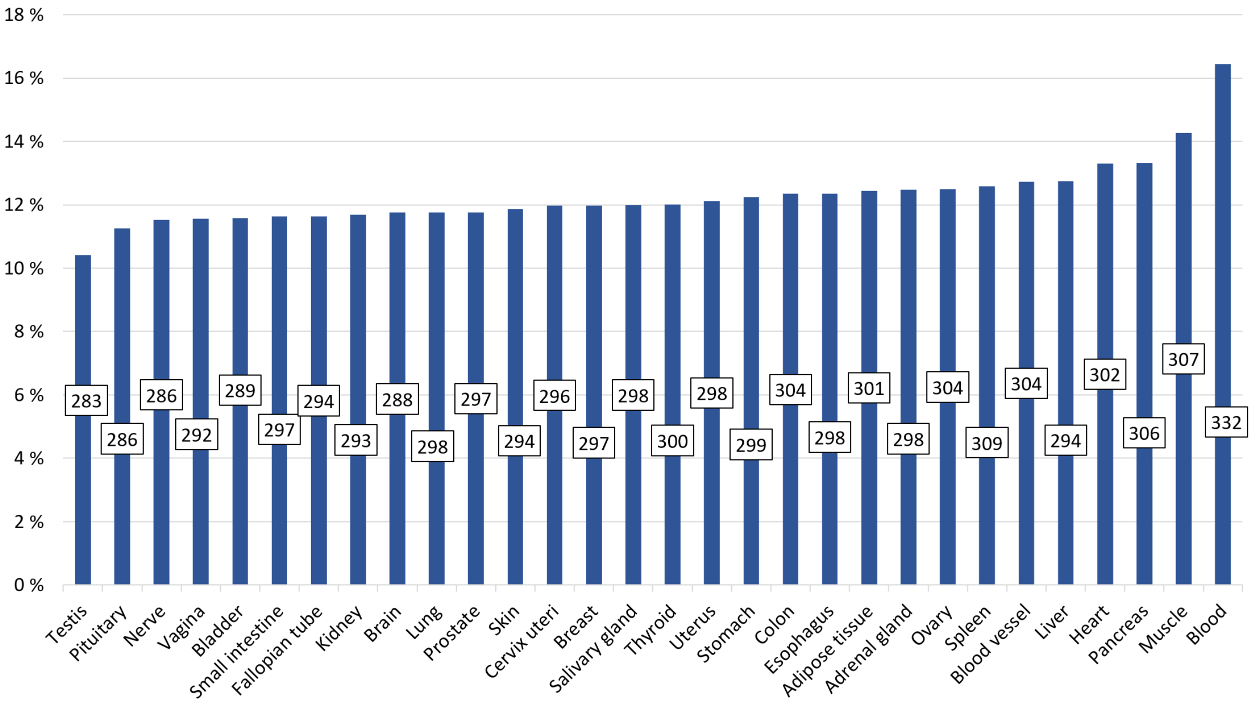


Figure 18: Percent and number of essential genes in the 30 tissue types, including genes necessary to for performing the essential metabolic tasks. Generally, the number of essential genes in a tissue is larger the fewer total genes in the tissue.

3331 genes were affected by at least one SNP. However, due to the missing 623 genes, a total of 19 791 missense SNP, including 405 start/stop SNPs, are not actually represented in the gene knockout combinations used for FBA. These genes are, however, still represented in the summary data below.

The goal of using start/stop SNPs as a crude way of selecting LoF SNPs, and keeping the gene combinations relatively small, was quite successful (Table 2) Using all missense SNPs lead to combination sizes in the hundreds. Even for the tissue specific models, where the number of in-model genes affected would typically be around 50-60% of the average number of genes affected for each sample, there would be over 200 affected genes. Filtering away essential genes would still leave most samples in the 200 range. With the average number of genes in the 30 tissue specific models being 2439, knocking out 200 or so genes would take out 8 to 9% of the genes in the model. A number far too high to produce useful results on its own due to such a large amount of gene loss is practically guaranteed to show a negative impact. Knowing which specific genes contributed to the negative impact would then need sub-combination analysis, which depending on how large the sub-combinations are, could be a huge number. Although not unfeasible, at least for double knockouts, doing this for all samples would require a huge number of FBA runs, even if filtered to not rerun combinations across samples. Given the 3331 genes found to be affected by at least one missense SNP, one could perhaps simply run double knockouts for the entire 3625 genes and cross-reference it with the samples instead. On top of this, running double knockouts would not be guaranteed to recreate the observed FBA results.

SNP type	Num Genes	Non Ess	Ess	Percent Ess	Unique Combs.
All missense heterozygote	903	856	46.3	5.1	2548
All missense homozygote	385	365	20.2	5.3	2548
Start/stop heterozygote	12.3	12.2	0.06	0.5	2538
Start/stop homozygote	4.12	4.11	0.002	0.06	375

Table 2: Overview over average number of affected genes for the individual genomes/samples, for different SNP types. Affected genes have been divided into essential and nonessential categories. Start/stop is included in the all missense data. This data is for the generic Human1 model, which does have significantly fewer essential genes than the tissue specific models. Nevertheless, there is a clear difference in which genes are affected when going from all missense to only start/stop SNPs. Additionally, the number of affected genes included is severely reduced by the more stringent SNP filter criteria. In terms of unique combinations, both all missense categories produced a unique combination for every single sample. For the start/stop combinations there was a slight reduction for heterozygote based combinations, while for the homozygote based combinations there was significant overlap between the samples.

Of course, using all missense SNPs to get gene combinations for knockout is not likely to be representative of LoF in an individual. Using the more disruptive start/stop SNPs, more manageable gene combination sizes were obtained. (Table 2) Additionally it does appear these SNPs are present in a much lower frequency in essential genes than missense SNPs in general. (Figure 19A and B) This indicates that these SNPs are to a greater extent selected against, and can be used as a crude selection of LoF variants. Additionally, whereas all missense SNPs had no significant change in the percent of essential genes that were affected between hetero- and homo-zygote, there was a large change for start/stop SNPs. The reason for this may be that many of the disruptive SNPs that do affect essential genes are rare and therefor primarily exists as heterozygotes. Of course, if they truly are a variant causing LoF in an essential gene, a homozygote carrier may simply not be viable. Due to this, there is a good probability that any start/stop SNP found as a homozygote is in actuality benign or do not cause a complete LoF.

Large differences were also observed between the different tissues. This was not unexpected, as the number of genes as well as essential genes varies between the tissues. (Figure 18) As an example, blood, had across all 2548 samples an average of 65% of the start/stop affected genes not being present in the model. (Figure 19B) Blood also had one of the highest affected essential gene percentages at 0.86%. This was not entirely unexpected as blood was the model with the highest number of essential genes. The tissue which had the highest average of affected essential genes was the pancreas, with 0.96%. (Figure 19D) As compared to blood, this model averaged sample affected genes in the model at around 60%. As previously stated, these numbers do include the genes not present in any tissue model.

In general, all damage observed in FBA could be explained by single genes or double combinations of genes. For larger gene combinations there would typically be multiple genes, which all on their own caused some damage, that compounded into a cumulative output. This could be seen when biomass production was not zero, as well as in the tasks, with different task failures being caused by different gene knockouts. In other cases, the negative effects of certain genes would be masked by other genes. This could happen when biomass output was zero, in which case any gene knockout which simply reduced the output would not be visible on its own. As for the tasks, it is possible for multiple genes to independently cause the same task to fail.

4.2.1 Start/stop homozygote

No tissue had any task failures, essential or not, for this category. Three tissues, the pituitary, muscle, and blood had zero samples with biomass reduction of any kind, and as a result, zero samples with any damage at all. In 11 tissues, four samples: HG01844, HG01970, NA18526, and HG00119, all showed the same slight reduction in biomass production, caused by the knockout of the gene CLYBL. They all had the same two missense SNPs rs41281112 and rs3783185. rs41281112 is a start/stop SNP and causes a premature stop codon on position 259. Despite this, these SNPs were heavily featured in the data set, with an additional 102 heterozygous carriers of rs41281112. The tissue exceptions were the pituitary, muscle, heart, and blood.

In the lung tissue, sample HG00525, was found to have a slight reduction in biomass production dues to rs71581941 which causes a premature stop codon on position 580 in the SLCO1B1 protein. This SNP is registered as pathogenic in ClinVar, and associated with Rotor syndrome.

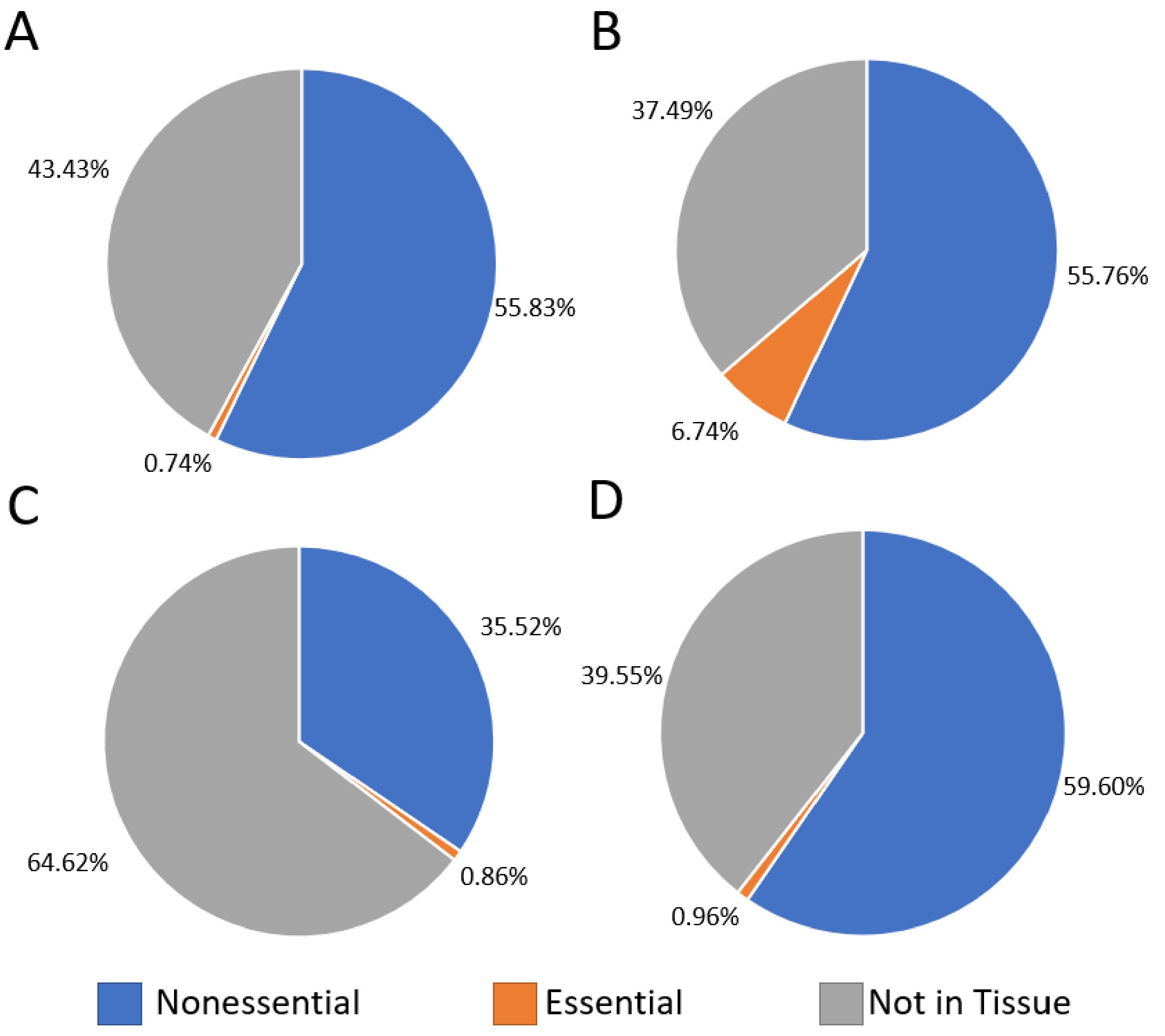


Figure 19: Distribution over affected genes. Average number of nonessential and essential affected genes in a sample.**(A)** Distribution of affected genes in samples on average for all 30 tissue types for start/stop SNPs. **(B)** Distribution of affected genes on average for all 30 tissue types for all missense SNPs. **(C)** Distribution of start/stop SNP affected genes in blood. Second highest in essential genes. The majority of affected genes are not found in this tissue, highest of all tissues. **(D)** Distribution of start/stop SNP affected genes in pancreas, which was the highest percentage of essential genes.

In heart and pancreas tissues, two samples: HG02323 and HG03057, showed a slight reduction in biomass production. This was due a premature stop codon on position 292 of the protein ACSM3, caused by rs52817836. While this SNP is not reported in ClinVar, it is deemed deleterious by MutationTaster. An additional 64 samples are heterozygous for this SNP.

4.2.2 Start/stop homozygote, including essential genes

Now including essential genes, some additional results were found. In all tissues, sample NA19198 showed a multitude of essential task failures, as well as biomass production reduction. These failures were the result of a premature stop codon in position 56 of the protein COX6B2, caused by rs138900187. It was not reported in ClinVar, while MutationTaster deems the SNP to be deleterious. A further 33 samples were heterozygous for this variant.

Also represented in all tissues is sample HG03874, with a SNP in the protein PTPMT1. Knockout of this gene causes biomass production to drop to zero. This SNP changes the original stop codon to a codon coding for serine. This change is however considered benign by ClinVar. A further 22 samples are heterozygous carriers of this SNP.

In the pancreas results, two samples: NA19351, HG02870 were found to fail two essential metabolic tasks: 'Beta oxidation of long-chain fatty acids' and 'Beta oxidation of unsaturated fatty acid (n-9)'. These failures were caused by the start/stop SNP rs114238154 in the protein NQO1, which changes the start codon to a codon coding for threonine. This SNP is not listed in ClinVar, but MutationTaster considers the change to be deleterious due frameshift. 64 other samples were heterozygous for this SNP.

4.2.3 Start/stop heterozygote

These results also picked up the nonessential genes from above, except every sample listed as heterozygote also gave the same result. This led to a considerable increase in non-optimal results. (Table 3)

The most notable result was three double combinations all involving the protein HIBCH, which lead to a multitude of essential task failures depending on tissue. (Table B1) These results were present in a total of six samples: NA21091 and NA20812 in combination with PCCA. NA12045, NA18997, and HG03499 in combination with PCCB, and finally HG00107, in combination with MMUT. (Table 4) None of these genes, when knocked out, causes any reduction in biomass production or essential task failures on their own. The implicated HIBCH SNP is rs291466, which changes the start codon of the transcript. This is however not considered a problem, as another start codon on the same reading frame follows shortly after. ClinVar considers the SNP benign. The majority of the samples do carry at least one copy as well. However, HIBCH deficiency is a disease that does exists, and due to the apparent existence of aggravating effects, a deeper look into its workings has been done in

Table 3: The number of samples having non-optimal FBA results for all 15 tissues using start/stop heterozygote SNPs based gene combinations. Differences for when essential genes are included is also shown. Sorted by non-optimal results. At the bottom are average results. There is a clear difference in how many tissues are affected by what is the same input genes. It is also clear that the majority of non-optimal results are due to knockout of essential genes.

Tissue	Non-optimal	Inc. Essential	Difference
Pancreas	232	488	256
Lung	162	373	211
Liver	159	357	198
Brain	147	344	197
Adipose Tissue	146	356	210
Kidney	143	345	202
Nerve	143	340	197
Thyroid	140	361	221
Adrenal Gland	138	346	208
Spleen	137	372	235
Skin	136	337	201
Heart	111	318	207
Pituitary	23	223	200
Blood	19	268	249
Muscle	19	242	223
Average	124	338	214

section 4.4.

Table 4: Which secondary gene/protein each of the six samples which had a HIBCH based combination had. All listed SNPs cause a premature stop codon in the protein. In the MMUT sample, multiple SNPs were present, the rsID included is for the start/stop SNP. Included is also the position of the acquired stop codon.

Sample	Gene	rsID	Protein Change
All	HIBCH	rs291466	M1T
NA20812	PCCA	rs138149179	R313*
NA21091	PCCA	rs530307529	Q421*
NA12045	PCCB	rs186031457	R113*
NA18997	PCCB	rs768935968	Q403*
HG03499	PCCB	rs191375566	Y405*
HG00107	MMUT	rs200596762	R228*

Present in six tissues are ten samples with stop gain SNPs in the gene XYLB. (Table 5) Knockout of this gene led to decreased biomass production. These SNPs were not reported in ClinVar. All SNPs were considered to be deleterious by MutationTaster. The six tissues were in this case: kidney, liver, lung, pancreas, pituitary, and adrenal gland.

Additionally, here were a number of cases with a few or singular samples showing biomass reduction in this result category. (Table 6) The degree of reduction was varying, but due the model being unbounded it is difficult to say what exactly to make of these results.

4.2.4 Start/stop heterozygote, nonessential tasks

When expanding the scope of the analysis to include additional metabolic tasks, a number of additional results were found. (Table 7) There was a total of 33 samples with at least one failed additional task. In all but one case these failures were due to single gene. The six HIBCH samples were also present in these results, in all tissues, with a multitude of

Table 5: Which samples had which SNP in the XYLB protein. All samples were heterozygous for the relevant SNP. Also included is the positioning of the SNP in the protein.

rsID	Protein Position	Samples
rs200154272	Q19*	HG03066, HG03388
rs140641713	Q450*	HG02239, HG01102, HG01431
rs79233786	R476*	NA18949, NA19088, NA18948, NA18965, NA19064

Table 6: A number of samples that showed a reduction in biomass production as well as the gene associated with that reduction. In all cases the samples were heterozygous for the SNP listed. All SNPs except rs572064049 produced premature stop codons. Instead, rs572064049 affects the start codon causing a reading frame shift. In terms of tissues, their presence was varied. Some were in all tissues, while others in a few. All of the samples were found in the pancreas results.

Sample(s)	Gene	rsID
NA19172, NA18505, HG03565, NA20320	CD38	rs147573494
NA18560	CD38	rs139152162
NA18519	TALDO1	rs202135397
HG02186	DERA	rs572064049
HG03963	CPT2	rs539239516
HG00375, HG00324	NMRK1	rs191532264
HG02887	APRT	rs556666445
HG03640	ABCA1	rs575627531

additional failed tasks. In most cases the knockout combinations did not cause a reduction in biomass production.

An obvious outlier was the liver, with 26 samples failing at least one of the additional tasks. This was primarily due to the task Taurocholate *de novo* synthesis from minimal substrates and excretion, which was not present in any of the other tissues. There were actually multiple different gene knockouts causing this metabolic task to fail. (Table 8) Also failing this additional task (among others) was the six HIBCH samples, although they have not been included here. The three SNPs pertaining to CYP27A1 were found pathogenic or likely pathogenic in ClinVar, as well as deleterious by MutationTaster. The AMACR SNP was not reported in ClinVar, it was however found to be deleterious by MutationTaster. The two AKR1C1 SNPs were also not reported in ClinVar, but both were found deleterious by MutationTaster. On the other hand, the HSD3B7 SNP was found to be benign by MutationTaster, likely due to it affecting the very last codon before the WT stop codon. The SNP was not reported in ClinVar.

A single instance of a combination of knockouts causing a task failure was found. The aforementioned NA18519, with its TALDO1 knockout causing biomass production reduction, would cause task failures in the brain and heart. The task gluconeogenesis from lactate and optionally fatty acid would fail in both, with the heart also having failures in 'release of glucose from glycogen' and 'gluconeogenesis from glycerol'. The two latter tasks were not

Table 7: How many samples failed at least one of the additional metabolic tasks in the different tissues using start/stop heterozygote SNP based gene combinations. This includes the six previously mentioned HIBCH gene combination samples, which also failed essential tasks. A total of 33 unique samples were found to have at least one task failure.

Tissue	Samples with failed tasks
Adipose Tissue	14
Adrenal Gland	12
Blood	9
Brain	15
Heart	14
Kidney	14
Liver	26
Lung	12
Muscle	15
Nerve	14
Pancreas	13
Pituitary	13
Skin	14
Spleen	13
Thyroid	13

Table 8: All samples which failed the metabolic task: Taurocholate *de novo* synthesis from minimal substrates and excretion. This task was only present for the liver, and the only task these samples failed. Also included is the associated start/stop SNP, all of which produces premature stop codons.

Sample(s)	Gene	rsID
NA20892	HSD3B7	rs547474493
HG01142, HG01384, HG01378	AKR1C4	rs191144263
HG02035	AKR1C4	rs558287231
HG03663	AMACR	rs1473596504
HG01989	CYP27A1	rs575064188
HG02152	CYP27A1	rs533885672
NA19771	CYP27A1	rs188850202

present in the brain model, but were present in other tissues where they did not fail. In muscle however, the three tasks did fail, although not due to knockout of TALDO1 alone. The additional knockout of FBP2, through the SNP rs77568573, caused the task failures in this tissue. This SNP causes stop codon loss, and is reported as benign in ClinVar and MutationTaster. While the SNP itself may not be deleterious, it did highlight the situation where a task fails in some tissues due to a single knockout, but not in others. If additional knockouts are introduced however, the same tasks fail in those additional tissues (but not necessarily all).

Three samples failed the metabolic task 'release of glucose from glycogen', caused by three different SNPs producing premature stop codons in the AGL protein. (Table 9) This failure was observed in all 15 tissues. It did not cause a reduction in biomass production. The two SNPs rs193186112 and rs531425980 are both reported in ClinVar as pathogenic in association with glycogen storage disease type III. The third SNP, rs539108137, was not reported in ClinVar. All three SNPs were found deleterious by MutationTaster.

Table 9: The three samples found to fail the metabolic task 'release of glucose from glycogen', due to knockout of the protein AGL. All three SNPs produce premature stop codons, with their position listed. The failure was observed in all 15 tissues.

Sample	rsID	Protein Change
NA20291	rs193186112	Q86*
NA20542	rs531425980	R977*
HG04131	rs539108137	Q1031*

Two samples, HG01805 and HG02769, failed the task creatine *de novo* synthesis from minimal substrate and excretion. This task was only found for the liver model, and was thus only observed in the liver. Reduction in biomass production was not observed. The failures were due to two SNPs causing premature stop codons in the gene MTR. For sample HG01805, the SNP rs536238004 was the cause, while for HG02769, the SNP rs560603359 caused the knockout. Both were found to be deleterious by MutationTaster. In ClinVar, rs560603359 was not reported, while rs536238004 was reported as having uncertain significance for "disorders of intracellular cobalamin metabolism".

There were several task failures only found in single samples. (Table 10) Some were widespread in multiple tissues, with multiple task failures, others were more limited in their scope. There were also a few carrying SNPs reported as pathogenic in ClinVar. A sample that was not reported in ClinVar, but that failed a variation of tasks in all 15 tissues was NA19446, carrying the SNP rs201411926 causing a premature stop codon in the protein GPI. This SNP was considered deleterious by MutationTaster. The failed tasks pertain to gluconeogenesis

Table 10: Various samples which had at least one of the additional metabolic tasks fail in at least one tissue, the gene and SNP that caused the result, and whether or not the SNP was reported on in ClinVar. All SNPs cause premature stop codons, except rs528828174 for the gene MPI, which affects the initiator codon, causing a shift in reading frame. Additionally, all SNPs were found to be deleterious by MutationTaster.

Sample	Gene	rsID	Reported in ClinVar
NA19446	GPI	rs201411926	No
HG02462	ALDH4A1	rs536031585	No
HG01440	CP	rs200363523	No
NA20827	MPI	rs528828174	Yes
HG00704	GNE	rs189454495	No
HG01809	SLC25A15	rs104894429	Yes
NA12546	SLC25A19	rs150354936	No
HG02014	HMGCS2	rs587727388	No
HG03692	ACSS2	rs530201978	No
HG02410	SLCO1A2	rs534910521	No
HG02756	SLCO1B3	rs550941268	No
HG00651	OAT	rs200068769	Yes

from different substrates. In muscle there were additional failed tasks pertaining to anaerobic rephosphorylation of the nucleoside triphosphates: ATP, GTP, CTP, and UTP. Additionally, a slight reduction in biomass production was observed with this knockout.

Sample NA20827 failed the tasks GDP-L-fucose *de novo* synthesis and GDP-mannose *de novo* synthesis due to knockout of the gene MPI. It does not cause biomass production reduction. The implicated SNP, rs528828174, alters the start codon, leading to a shift in reading frame. This is considered deleterious by MutationTaster. In ClinVar it is reported as uncertain significance for Mannose-6-phosphate isomerase - congenital disorders of glycosylation (MPI-CDG). All 15 tissues were affected by this knockout.

Another sample with a SNP reported in ClinVar was HG01809. In this sample the SNP rs104894429 causes a premature stop codon in the protein SLC25A15. This causes a several task failures in the liver, but in no other tissue, even though the gene was present. (Table B2) There was also a slight reduction in biomass production. In ClinVar the SNP is reported as being pathogenic for Hyperornithinemia-hyperammonemia-homocitrullinuria syndrome (HHH). The disease characteristics matches well with the metabolic tasks that failed, with two of them being ornithine degradation and ammonia import and degradation.

The final sample which had its SNP reported on in ClinVar was HG00651. This sample carried the SNP rs200068769 which causes a premature stop codon in the OAT gene. Knockout of this gene causes the task arginine *de novo* synthesis from minimal substrate and excretion to fail in liver tissue. The task was not in use for any other tissues. There was also a slight biomass production reduction for this knockout. In ClinVar, this SNP is reported to be pathogenic for Ornithine aminotransferase (OAT) deficiency disease.

4.2.5 Start/stop heterozygote, including essential genes

When essential genes were included a large increase in samples with knockout combinations showing non-optimal results was produced. (Table 3) Due to every additional result being caused by essential genes, it is already known that the damage observed is by a single gene. Even if there are multiple essential genes in one combination, they will each on their own disrupt essential function. However, in this case each gene is also tied back to one or more SNPs, and as such a few example results have been included.

Present in all tissues was sample NA18552 with the SNP rs199707836 in the protein HMGCR. Knockout of this essential gene causes essential tasks Cholesterol *de novo* synthesis and growth on Ham's media to fail. The general solution also produced zero biomass. The SNP causes a premature stop codon, and was found deleterious by MutationTaster. It was however not reported in ClinVar. Also present in all samples was HG01530, with a premature stop codon inducing SNP, rs546130184, in the protein QARS1. This knockout also caused zero biomass production, failure of growth on Ham's media, as well as failure of the essential metabolic task of protein synthesis from amino acids. This SNP was found deleterious by MutationTaster, but was not reported in ClinVar.

Finally, as an example of how wrong things can go with even a single gene knockout, is sample HG00127. Present in all tissues, with a varying amount of task failures, this sample is homozygous for SNP rs74315366, producing a premature stop codon in the gene SDHB. While there is only a slight reduction in biomass production, up to 22 essential tasks will fail depending on tissue. (Table B3) This includes tasks such as oxidative phosphorylation and Krebs cycle NADH production. The SNP was found deleterious by MutationTaster, and was also reported as pathogenic in ClinVar, which included cancer predisposition and parganglinomas.

4.2.6 All missense homozygote

Due to the now massive number of genes in each combination (Table 2), as well as the fact that a substitution SNP on its own is a poor predictor of LoF, the results themselves are not very interesting. Every single sample will be unique, and show some form of damage. In most cases growth on Ham's media will fail. The six essential tasks listed under the HIBCH task failures also commonly occur together. (Table B1) There are outliers of course, and one such outlier is the sample HG02078, which will be included as an example.

For brain tissue, this sample had 203 genes for which it was homozygote for at least one missense SNP. This combination resulted failure of the tasks Cholesterol *de novo* synthesis and growth on Ham's media, as well as zero biomass production. To uncover which of the 203 knocked out genes were the primary cause of the results, all genes were run singularly and as double knockouts, resulting in 20707 combinations. As this was filtered for essential genes, no single gene caused any essential task failures or reduced the biomass production to zero. There where however a number of double combinations with task failures. (Table 11) Every combination caused growth on Ham's media to fail on its own, however, three combinations involving the gene DBT did not affect biomass production output. The knockout combination of TM7SF2 and LB additionally caused the failure of Cholesterol *de novo* synthesis, and reduced biomass output to zero. In fact, this combination of two genes had the same result as the initial combination of 203 genes. As a result, the effect of these two genes alone is masking the effect of the other deleterious gene knockouts. When it comes to the SNPs for these combinations, all were non-start/stop missense SNPs. (Table B4) All were found benign by MutationTaster, and ClinVar varied between not reported and benign. In this case, due to all SNPs being homozygote, they are also actually less likely to be deleterious, as deleterious SNPs are rarer.

4.3 PheWAS Data Results

The use of the PheWAS Cataoluge led to only a limited number of 46 missense SNPs for the genes found in the generic Human1 model. Only three of these produced premature stop codons. Due to the low number of start/stop SNPs, all missense SNPs were used to produce gene combinations. Additionally, combinations were created based on the phecodes the SNPs were tied to, which did reduced number of combinations significantly, but with great overlap between phecodes. There were a number of essential genes in the knockout screens for the various phecodes.

Table 11: Combinations causing task failure and biomass production reduction in HG02078. There are three combinations involving DBT, which all have the same result. The last combination also causes growth in Ham's media failure, but additionally causes biomass production of zero, and failure of Cholesterol *de novo* synthesis, effectively masking the results of the DBT combinations when combined.

Gene Combination	Solution	Task Failure
DBT, AMACR	Opt	Ham's media
DBT, CYP39A1	Opt	Ham's media
DBT, HSD3B7	Opt	Ham's media
TM7SF2, LBR	Zero	Ham's media, Cholesterol

As an example, some results pertaining to the gene MTHFR will be showcased. This gene was included in knockout combinations due to the SNP rs1801133, which is a missense SNP substituting alanine for valine or glycine on position 222, depending on exact base. This gene, which was present in all 15 tissue models, had varying effects when knocked out. In the heart, blood, and thyroid models, the knockout caused the growth on Ham's media task to fail, although the regular FBA solution remained unchanged. This MTHFR an essential gene in those models. In the pancreas model, it led to a slight decrease in biomass production, something that was not observed in any other tissue model. Finally, in the liver model, this knockout caused the nonessential task of Creatine *de novo* synthesis to fail. As mentioned earlier, this task is only present for the liver model. In ClinVar, rs1801133 is listed with different significances, including pathogenic, for a variety of afflictions. MutationTaster find the substitution of alanine to glycine deleterious for two transcripts of the gene, while being benign for the two other transcripts listed.

4.4 Amino Acid Degradation and HIBCH, A Deeper Look

HIBCH deficiency causes issues with valine degradation. However, when running FBA with knockout of HIBCH alone, no issues are detected for any tissue specific model. One reason is the different models' abilities to compensate in various ways. Even if a model is forced to uptake valine, it may not be forced to break it down completely. One way to test for this is to use a metabolic task, and indeed, the non-essential task list does carry tasks for various amino acid breakdown. However, they do not appear to have a forced uptake or excretion. When uptake is forced, the tasks fail for all tissue specific models. These tasks typically require complete breakdowns, which will include multiple pathways. However, any uptake and excretion of any metabolite can be added to a metabolic task. This makes it possible to pinpoint the task somewhat. Certain other reaction will also have to be used, such as various uptakes, transport reactions, and ATP production as examples. As an example, a metabolic task to test valine degradation was created and tested for the liver model. A number of knockouts of genes somewhat arbitrarily selected from in and around the valine degradation pathway, including the previously mentioned PCCA/PCCB and MMUT, were performed. (Table 12) Four of the genes, including HIBCH hindered valine degradation in this task. Double knockouts were also performed, but revealed no further damage.

Table 12: Single knockout results on a task created to test valine degradation. PCCA/B and MMUT was previously shown to cause essential task failure in all tissues in combination with HIBCH. The rest of the genes, except HIBCH, were partially arbitrarily chosen to test the task, and whether it functioned as expected.

Gene	Pass
HIBCH	No
PCCA	Yes
PCCB	Yes
MMUT	Yes
BCAT2	No
BCKDHA	No
DBT	Yes
HADHA	No

Another way is by exploring the model itself, through different objectives and constraints. FVA can also be used to check what fluxes any given reaction can have. By running FVA on the reactions of valine uptake and the first step of the breakdown of valine, the issues posed by HIBCH deficiency becomes clearer. This does however require some additional constraints on the model to work. The model must not be allowed to excrete intermediates between the first step of breakdown and the step catalyzed by HIBCH. Otherwise, the model was configured to optimize for biomass. While both healthy and HIBCH deficient model could take up as much valine as permitted, only the healthy model could break it down. Another way that was tried was to set the first or an early step of the breakdown pathway as the objective of for FBA. This way the maximum flux through that pathway can be checked. Again, the intermediates must be constrained, otherwise the model will just excrete those intermediates. This is a common occurrence in many metabolic diseases, there is a buildup of a certain metabolite(s). Such cases will not easily be picked up by FBA on unbound models, as the model can just excrete intermediates without issue, or compensate through other pathways in a way that may not be possible *in vivo*.

Unfortunately, which specific metabolites accumulate in a metabolic disease is not something that could be gauged accurately by FVA. The biggest success of the FVA was showing all the different ways the model can compensate. It did prove that the HIBCH pathway could not be used to completion, in that valine could not be completely degraded, but that was also accomplished through the use of a metabolic task. For example, valine could just be incorporated into various peptides and removed, or it could just be excreted as is through the many amino acid exchange reactions.

4.5 The Software Itself

When it comes to the performance of the code it does performs generally quite well. The inclusion of multiprocessing and use of simple FBA allows fast processing with relatively small RAM usage. The biggest issue at current is ease of use. Much of the code is divided into scripts, and while nice for testing and development, may not be as optimal for usage only. There is also a lack of proper documentation, although the code is mostly well commented. The result processing in particular is a bit convoluted and labor intensive, however, making it more automatic may be processing intensive due to potential large number of results and subsequent sub-combinations. (Figure 17) The reason behind all this is that the focus for the time available has been to produce something that works from start to finish, with a decent number of results for use as a proof of concept. Developing the software into an easy-to-use package for publication has not been the focus. In hindsight, there are many relatively small changes that could have been made, that would have accumulated to a better end software product. Then again, how the end product would look was not apparent as much of the overarching process was exploratory.

5 Discussion

5.1 The Genome Data

For this project, only Ensembl canonical transcripts were used for filtering SNPs. This is what will be considered the 'main' transcript of a gene. However, the filter can just as easily accommodate multiple transcript genes and eve non-protein coding genes. In some cases, there may be SNPs that affect one or a few transcripts in a gene, but not all. Being able to detect these SNPs may be useful as different transcripts may represent different functions, or be deferentially used in different tissues. However, canonical Ensembl transcripts were considered enough in that the model has a one gene, one protein, setup. Although, it may still be useful, simply to capture additional missense SNPs for use in combinations.

5.2 Loss of Function SNPs

In general, the SNP filter did perform well as a rudimentary LoF selection. The purpose of creating a SNP filtering mechanism for this project was the characterization of SNPs that have not yet been characterized. With the use of 1000 Genomes and PheWAS data, this was not entirely necessary. Yet, it was found that having the tool at hand was useful, and made for easy use and selection of SNPs in the context of the larger project. The use of start/stop SNPs as indicators of LoF did prove fairly accurate through the use of MutationTaster and ClinVar. This approach was labor intensive, but due to its use on non-optimal results only, did not prove to be too much. Ideally, a more detailed assessment of LoF should be done before FBA, allowing for a better inclusion of all missense SNPs, not just start/stop SNPs. Although, there is still the question of complete LoF and reduction in function.

The heavy focus on missense SNPs, and start/stop SNPs in particular, did prove to work better for the individual data than the PheWAS data. The primary reason behind this being the small number of missense SNPs found in the PheWAS catalogue. Although, due to the nature of GWAS, it is after all more likely to find common SNPs, while highly disruptive SNPs causing loss of function are rarer. These more common SNPs are more likely to affect regulation. Which is not to say that they cannot cause LoF, but it is not likely to be a complete LoF. This means that even if a more comprehensive list of GWAS SNPs were to be used, it would still be limited in the SNP types which would have the most direct impact in this project, meaning how certain a conclusion can be drawn on the effect of the SNPs based on the modeling results.

Highlighting the issues with use of GWAS/PheWAS data are the contrasting results from the individual data. The number of SNPs was of course much higher in the individual data, but adding to this was the fact that many of the start/stop SNPs were rare. For many, only one or a few samples were carrying them, usually as homozygotes, exemplifying the rarity of these SNPs. It should be noted that the number of individuals at 2548 is not massive, but is still sufficient as an example population, especially considering the heterogeneity of the included individuals. From the literature it is known that all individuals carry a number of LoF variants, something which is seen through the SNP filter and sample combinations. Based on this and MutationTaster results, the use of the start/stop SNPs does appear accurate enough to predict complete LoF genes in an individual. While likely not an extensive list of LoF genes in an individual, it is still a sufficient number of genes to produce unique combinations for most individuals. Again, the population size of 2548 is a bit small, and for larger populations there would probably be more overlap. However, the start/stop heterozygote combinations produced 2538 unique combinations, with an average size of 12.2 genes. The question of hetero- vs homo-zygosity still remains. However, if a heterozygote for an allele exists, it is likely that homozygotes also exist, granted they are viable. One thing that was not implemented when looking at heterozygote and homozygote gene combinations, was whether a heterozygote for one allele, was also heterozygote for a different allele on the other gene copy. It is possible that an individual could be homozygote for LoF through two different alleles or variants.

5.3 Essential Genes and Tasks

When it comes to modeling it is important to be clear on the gene being essential to the model, based on some criteria. Which genes are essential to a metabolic model, does not necessarily transfer exactly to the real organism. If the assumption is made that a gene or task is essential, and there is a knockout or task failure based on LoF SNP(s), then that organism would be non-viable. This would of course only be the case when the organism is homozygote for LoF in that gene. It was seen that likely LoF variants were less likely to occur in essential genes, again indicating that the selection method for LoF variants did work fairly well. (Table 2) Conversely, it also indicates that the essential genes are actually essential, as they have a lower frequency of LoF variants. Further exemplifying this was the clear difference between start/stop homozygote and heterozygote based gene combinations, with homozygote SNP based combinations affecting considerably fewer essential genes than the heterozygote ones.

For the heterozygote based gene combinations the majority of results with a non-optimal FBA solution or failed essential tasks, were caused by knockout of essential genes. (Table 3) For one thing this indicates that a large number of individuals carry LoF variants in important genes, and not just genes that are redundant or otherwise not necessary for survival. Secondly it shows that outside of these relatively few genes, there is a rather high tolerance for LoF. Alternatively, the conclusion might be that it is the models, and modeling analyses that fail to capture issues outside of the essential genes. Most likely, reality is a combination of the two.

One interesting aspect of there being different essential genes in different tissue models is that in reality these tissues come together to form a whole organism. As such, it might be prudent to list all essential genes together, and consider the ma unified list of essential genes for the organism. After all, if a gene is essential in even a single tissue, it could be considered essential for the whole organism. This may be useful for modeling as it prevents results which include genes found to be essential in another tissue. As discussed, if a gene is truly essential, LoF will lead to non-viability. Thus, if a gene is non-essential in one tissue, but is essential in another, any results involving that gene in the first tissue may not be usable, due to the gene not actually being non-functional in that individual. Of course, most of what have been discussed here only applies if a complete LoF is assumed. If modeling were to be done with only reduced gene function, inclusion of essential genes may still be useful.

5.4 Additional Tasks and the Models

The additional tasks did uncover several genes and gene combinations, in several tissues with interesting results that would otherwise have gone unnoticed. Exemplifying this was the knockout of SLC25A15 in sample HG01809 through the SNP rs104894429. In this case the SNP was found pathogenic in ClinVar, and it was seen that the disease characteristics matched well with the failed tasks, which could offer an opportunity to uncover mechanisms behind the disease. Another example was sample HG00651 with OAT deficiency. In this case however, it was not as clear how the task failure, which was arginine *de novo* synthesis, was related to the disease. The reason being OAT deficiency causes arginine breakdown issues, with reduced arginine intake as a treatment.[28] However, ornithine is also the precursor for arginine when synthesizing it. It is possible the restricted uptakes limits which pathways can be used; it is after all *de novo* synthesis. Nonetheless, if not for the inclusion of additional tasks such cases may not have been uncovered. In many ways, these additional tasks may prove more useful from a medical point of view, due to them not being essential. Even with homozygote LoF, the organism may still be viable, albeit with potential diseases.

As mentioned, there are also some issues with these additional non-essential metabolic tasks, the main two being their design and which tissues they should be applied to. With the design, the issue is the specificity of the task. Typically, a task will only specify certain basic metabolites, with the addition of a metabolite that is to be either degraded or produced. As was seen in the valine degradation tests under HIBCH (Section 4.4), simply creating a task to test this pathway was not straight forward. Although, an apparent functioning task was produced, there were still some questions regarding why a simpler design would not work. As the tasks become more specific and detailed, it may not be as clear what a passed or failed task actually means.

Originally, the tasks were a tool used by tINIT to help create tissue specific models. In this project however, the additional tasks were determined based on what the models were able to perform, with no regard for the tissues actual *in vivo* biology. The tissue specific models, used in this project, only used the essential tasks for their creation, although, gene expression data was also involved. Creating tissue models with additional tasks, and subsequently more accurate representation of biology, may one improvement point for more accurate results. Also related to the tissue specific models are inter-organic processes. As an example, some tissues may be capable of creatine *de novo* synthesis, but the majority of creatine in a human is likely produced through inter-organic processes.[74] This issue also relates to the first issue of simple tasks as both *de novo* synthesis and degradation of various metabolites may involve multiple organs. One way to get around this is to use a generic metabolic model, but as seen with the number of essential genes, this model will generally be more robust towards perturbations, perhaps beyond true robustness. On the other hand, it may be possible for a different organ to compensate for shortcomings in another, at least to a degree.

5.5 Biomass Production Reduction in Unbound Model

Due to the models being unbound the specific solution numbers does not confer anything except there was a reduction. The exception being if the solution was zero, with such a result indicating non-viability. Although, at that point the growth on Ham's media essential task would also fail. Due to this, producing an objective solution may not have been necessary, and could save some computing time if removed. In some cases, these results with reduced biomass output and nothing else, did lead back to pathogenic SNPs. As an example, is HG00525 with rs7158194 in the SLCO1B protein, which was reported as pathogenic for Rotor Syndrome in ClinVar. Rotor Syndrome is however considered a benign disease that does not require any intervention, at least not initially.[75] Why this SNP lead to biomass production reduction is not known, and it is entirely possible it would not lead to such a reduction in a bound model. This may be supported by the fact that the reduction was only seen in lung tissue, and it may simply be a quirk of that specific model, especially as rotor syndrome is not associated with the lungs. Conversely, it may also be possible, and probably likely, that many gene knockouts which were observed to have no effect, would have had an effect in a bound model. The reason for this being the unbound models having the ability to compensate without affecting the optimal solution. Even if the approach of Robinson et. al.[6] is taken and a few key inputs is limited (Figure 6), there will still be ample opportunity for the model to compensate in various ways in face of knockouts. This is due to the hard limits are really only being placed on biomass production.

5.6 Mechanisms of Disease

One of the promised benefits of this approach is the relatively easy study of the mechanisms underlying the disease, at least as long as affected genes are completely knocked out. If there is only a reduced function, it may be possible for the cells to compensate by increasing expression. It has however become apparent that reconciliation of results and known diseases caused by the SNPs used and their respective genes knockouts, is not straight forward. In the case that there is no disease to reference back on, it is even more difficult to say something about *in vivo* ramifications of the results.

One such example is the aforementioned SLC25A15 knockout. Although, the specific mechanisms were not studied thoroughly in the model, one clear example is one of the tasks that failed was ammonia import and degradation. In the disease HHH, the middle 'H' stands for hyperammonemia, referring excess ammonia. The model does offer the opportunity to study exactly why this task failed, and if done, the exact reason why SLC25A15 knockout causes hyperammonemia may be uncovered through further study of the model. A more clear-cut example of mechanism is sample NA20827 with rs52882817 in the MPI protein, leading to MPI-CDG diseases. In this sample, the two tasks of *de novo* synthesis of GDP-L-fucose and GDP-mannose failed. This disease is treatable with mannose intake, although complications may occur.[76] When mannose is not present, it needs to be produced from glucose or fructose. However, without MPI this is not possible. Mannose based glycosylation precursors can thus not be produced. In terms of mechanism of disease symptoms, they are not explained by this approach. However, as long as it is known that glycosylation issues are the root causes, they can be tracked back to the MPI knockout.

In both of these cases there have been a known disease to referee back to. This makes it possible to attempt to reconcile the results and the already known disease mechanism. As is apparent, it is not straight forward to explain a disease through the FBA results, although it varied. This does not mean it was not possible, just that the effort required in each case varied depending on detail required.

5.6.1 FVA as a Tool to Uncover Mechanisms of Disease

To help with understanding underlying disease mechanisms, FVA may be useful, but will require some often tedious work in establishing proper constraints. The use of HIBCH knockout to test FVA was partly successful. This knockout did not produce any adverse effects on its own, even though its deficiency is known to cause disease.[62] No LoF SNPs were observed. The start/stop SNP that was registered, rs291466, affected the start codon, but did not cause deleterious damage. Nevertheless, it was chosen for a deeper look into its effects due to its role in valine degradation. This was in part due to the given amino acid degradation tasks were apparently not working properly. With the way FBA/FVA works, a buildup of any kind is not allowed, instead a pathway that cannot run to excretion is simply omitted. Not to mention, a model will only take up a metabolite it needs, in the exact proportion it needs it in relation to other metabolites. Using valine as an example, the model will never have an excess of valine, it will simply place its uptake to exactly match what it needs. A task can be setup to test valine degradation, by limiting outputs and inputs, and forcing a valine uptake, it can be tested whether the model is capable of breakdown. This was successful, but it also does not say anything about what the *in vivo* results of this might be, such as which metabolites accumulate.

The core of the issue with FBA/FVA when it comes to finding which metabolites accumulate is that even if a pathway is forced, the model will simply use the first available opportunity to getting rid of the excess metabolites. This may be after the first step of the pathway, while *in vivo* it may continue to the last metabolite before the knocked-out step. There may also be multiple different metabolites that accumulate. Through meticulous use of constraint and FVA it may be possible to learn more of how the model behaves with a knockout, but it is a tedious task.

5.7 The Approach: Known Diseases and New Discoveries

Due to this project being more of a proof of concept for a method development, focus has been kept on what the method has been able to achieve. As such, not all results have been discussed directly, but various results have instead been used as examples. One thing that has become clear however, is with the focus on LoF SNPs in metabolic genes, the potential diseases uncovered are deficiency diseases, which are typically high penetrance. In these diseases, if the protein is part of anabolism, some metabolite is missing, if they are part of catabolism, there is accumulation of some metabolite. Indeed, we have already seen many examples, with HIBCH, OAT, and MPI. These diseases have generally been well characterized, likely due to their disruptive nature, with well-defined mechanisms. For many of the results not mentioned so far this was also true. One example is TALDO1[77], which was the only non-essential tasks result to show task failure as a combination. Although, in certain tissues it did cause task failure on its own. AGL, GPI, and MTR was three other proteins with known deficiency diseases.[78–80] It does appear that many of the gene knockouts disruptive enough to cause non-optimal results, are already associated with disease. This is not a problem in of itself, but instead highlight a limitation in new discoveries, at least with the use of nonsense SNPs. On the other hand, it does to a certain degree validate the approach taken. Raw variant data can be used to predict disease, at least in certain cases.

There are still some possibilities for new discoveries though, even for already known diseases. New SNPs that have not previously been described may be linked to the disease. Although, the disruptiveness would again cause a direct look at the genes known to cause the disease if any individual presents with symptoms. This leads to an additional use case, which is when different genotypes may cause the same phenotype. If two results are similar or even identical, but have different knockouts behind them, it could point to a not known cause of a disease. Then there are the concrete mechanisms. This method may uncover new effects of a disease, through the study of model changes. However, as has been noted this is not a straightforward task. To determine if anything new has been discovered one would not only need to meticulously study the model results, but also already known functions of the disease. To add to this, any potential results would also need to be validated outside of the computer.

One of the main prospects of this method was finding gene combinations which may cause adverse effects when knocked out. The results on these were limited, which may be due to the limited data set. However, some combinations were uncovered, proving they are worth pursuing. These gene combinations do open up a new set of possibilities for new discoveries, and a larger data set may give substantially more results of this type. Diseases caused by combinations would be more complex in nature, and may not be as straight forward to uncover by working backwards from symptoms. As such this approach may be useful for apparent metabolic disease that is not already characterized.

Another aspect of this approach is all the metabolic diseases that do exist, but was not detected. There were likely genes with LoF variants that are known to cause disease, but did not produce any results to suggest that. One example was HIBCH deficiency, where HIBCH knockout on its own did not produce any adverse effects. In this case, once a functioning metabolic task to test valine production was created, it did prove disruptive. However, if not specifically tested, it would have gone by unnoticed, even though it has disease associated with it. The developed method was never going to be extensive, in part due to the limitations of the models. Nevertheless, it highlights that results produced, or lack thereof, cannot be directly translated to biological reality. Instead, results can act as pointers to specific targets for further study.

The way the method was used with LoF nonsense mutation it is clear that disruptive and rare diseases are the main findings. Compared to GWAS, which is suited for more common variants, each with varying disease impacts. One of the issues raised with GWAS is the lack of mechanisms for why a SNP is associated with a disease. While the method developed here does have the capability to uncover mechanisms, it is clear that with the way it was used, it cannot be said to compete with GWAS in terms of which SNPs are discovered. This was exemplified with the PheWAS data, where there was few SNPs suitable for the current use case. Although, some potentially interesting results were produced nevertheless. With more refinement, the developed method may eventually be able to uncover metabolic issues that are less disruptive in of themselves, yet still contribute to disease. A better selection of SNPs, including SNPs which may not cause a complete LoF, as well as use of better models that are also realistically constrained, does promises to expand the range of usable results.

5.8 Personalized Medicine

Finally, in terms of personalized medicine, there are many potential use cases. The current state of the method is certainly not ready for use on individuals in any clinical setting, but increased knowledge of mechanism of disease, and particularly combination may be useful. For example, a LoF in a protein is known to be pathological if another protein is also not functioning properly, a combination. There may be a drug that inhibits the first protein, which may be useful as long as the individual has function in the other. In this scenario a person who carries LoF mutation in the second protein may have adverse effects if given the drug. On the other hand, as a simple illustration, a cancer may stop expressing certain genes, even perhaps the second gene in the previous combination. By giving the patient the drug, the first gene is also shut down, which could disrupt the cancer, while the rest of the body is not affected. These are extremely simple examples, but are not outside the realm of possibility. One issue with these simple examples is however if a combination is non-viable. Such a combination might never occur in an individual, and as such might not be discovered. In addition to the mentioned examples there may be other possible scenarios as well.

6 Conclusions

The purpose of this project was to create a pipeline that could take in variant data, primarily SNPs, and use them to predict disease through the use of metabolic modeling. In general, showcasing the viability of such an approach was successful. Raw variant data could be used to predict disease. In many cases, observed results also matched descriptions of known diseases caused by the variants in question. This was achieved by filtering out LoF SNPs and using them to perform knockout FBA. The use of metabolic tasks was also considered a success in that several of the predicted results did come about through task failures. There were also some issues. In terms of input data, the PheWAS data did not prove particularly useful, primarily due to the small number of SNPs. The individual data was, on the other hand, more successful, producing many interesting results. Predicting exact mechanisms of diseases did also prove challenging in some cases. It was found that many of the results were already well characterized, making the utility of using LoF SNPs uncertain. Nevertheless, with larger sets of input data it is expected that more combinations, and unique results in general, would be produced, increasing the probability of discovering unknown effects. The same is expected with use of a more refined variant filter and modeling. All code used to accomplish these results is available at **https://github.com/SigveLan/Master**. The supplementary data folder includes a wide variety of data used and produced in this project.

7 Outlook

If additional time was to be used from this point on, there are two primary areas that would be worked on. Firstly, effort would be made to compile and clean up the produced code into a more user-friendly format that is better suited for publishing. Especially result processing would be improved upon. Secondly, effort would be made to acquire additional whole-exome and -genome data sets. Even with the 2458 individuals already analyzed there are many interesting results. A larger data set would undoubtedly produce even more. If time allowed it, a deeper dive into these results would be made.

References

[1] C. Montag, R. P. Ebstein, P. Jawinski, and S. Markett, "Molecular genetics in psychology and personality neuroscience: On candidate genes, genome wide scans, and new research strategies," *Neuroscience & Biobehavioral Reviews*, vol. 118, 2020.

[2] S. I. Vrieze, W. G. Iacono, and M. McGue, "Confluence of genes, environment, development, and behavior in a post Genome-Wide Association Study world," *Development and psychopathology*, vol. 24, no. 4, 2012.

[3] W. Satake, Y. Nakabayashi, I. Mizuta, Y. Hirota, C. Ito, M. Kubo, T. Kawaguchi, T. Tsunoda, M. Watanabe, A. Takeda, *et al.*, "Genome-wide association study identifies common variants at four loci as genetic risk factors for parkinson's disease," *Nature genetics*, vol. 41, no. 12, 2009.

[4] P. M. Visscher, N. R. Wray, Q. Zhang, P. Sklar, M. I. McCarthy, M. A. Brown, and J. Yang, "10 years of GWAS discovery: Biology, function, and translation," *American journal of human genetics*, vol. 101, no. 1, 2017.

[5] Q. Huang, "Genetic study of complex diseases in the post-GWAS era," *Journal of Genetics and Genomics*, vol. 42, no. 3, 2015.

[6] J. L. Robinson, P. Kocabas, H. Wang, P.-E. Cholley, D. Cook, A. Nilsson, M. Anton, R. Ferreira, I. Domenzain, V. Billa, A. Limeta, A. Hedin, J. Gustafsson, E. J. Kerkhoven, L. T. Svensson, B. O. Palsson, A. Mardinoglu, L. Hansson, M. Uhlén, and J. Nielsen, "An atlas of human metabolism," *Science Signaling*, vol. 13, no. 624, 2020.

[7] B. J. Sánchez, C. Zhang, A. Nilsson, P.-J. Lahtvee, E. J. Kerkhoven, and J. Nielsen, "Improving the phenotype predictions of a yeast genome-scale metabolic model by incorporating enzymatic constraints," *Molecular systems biology*, vol. 13, no. 8, 2017.

[8] J. D. Orth, I. Thiele, and B. Ø. Palsson, "What is flux balance analysis?," *Nature biotechnology*, vol. 28, no. 3, 2010.

[9] D. Segrè, J. Zucker, J. Katz, X. Lin, P. D'haeseleer, W. P. Rindone, P. Kharchenko, D. H. Nguyen, M. A. Wright, and G. M. Church, "From annotated genomes to metabolic flux models and kinetic parameter fitting," *OMICS A Journal of Integrative Biology*, vol. 7, no. 3, 2003.

[10] S. Nurk, S. Koren, A. Rhie, M. Rautiainen, A. V. Bzikadze, A. Mikheenko, M. R. Vollger, N. Altemose, L. Uralsky, A. Gershman, S. Aganezov, S. J. Hoyt, M. Diekhans, G. A. Logsdon, M. Alonge, S. E. Antonarakis, M. Borchers, G. G. Bouffard, S. Y. Brooks, G. V. Caldas, N.-C. Chen, H. Cheng, C.-S. Chin, W. Chow, L. G. de Lima, P. C. Dishuck, R. Durbin,

T. Dvorkina, I. T. Fiddes, G. Formenti, R. S. Fulton, A. Fungtammasan, E. Garrison, P. G. S. Grady, T. A. Graves-Lindsay, I. M. Hall, N. F. Hansen, G. A. Hartley, M. Haukness, K. Howe, M. W. Hunkapiller, C. Jain, M. Jain, E. D. Jarvis, P. Kerpedjiev, M. Kirsche, M. Kolmogorov, J. Korlach, M. Kremitzki, H. Li, V. V. Maduro, T. Marschall, A. M. McCartney, J. McDaniel, D. E. Miller, J. C. Mullikin, E. W. Myers, N. D. Olson, B. Paten, P. Peluso, P. A. Pevzner, D. Porubsky, T. Potapova, E. I. Rogaev, J. A. Rosenfeld, S. L. Salzberg, V. A. Schneider, F. J. Sedlazeck, K. Shafin, C. J. Shew, A. Shumate, Y. Sims, A. F. A. Smit, D. C. Soto, I. Sović, J. M. Storer, A. Streets, B. A. Sullivan, F. Thibaud-Nissen, J. Torrance, J. Wagner, B. P. Walenz, A. Wenger, J. M. D. Wood, C. Xiao, S. M. Yan, A. C. Young, S. Zarate, U. Surti, R. C. McCoy, M. Y. Dennis, I. A. Alexandrov, J. L. Gerton, R. J. O'Neill, W. Timp, J. M. Zook, M. C. Schatz, E. E. Eichler, K. H. Miga, and A. M. Phillippy, "The complete sequence of a human genome," *Science (American Association for the Advancement of Science)*, vol. 376, no. 6588, 2022.

[11] M. B. Gerstein, R. P. Alexander, G. Fang, J. Rozowsky, and M. Snyder, "Annotating noncoding regions of the genome," *Nature reviews. Genetics*, vol. 11, no. 8, 2010.

[12] A. Chakravarti, "Population genetics—making sense out of sequence," *Nature genetics*, vol. 21, no. 1, 1999.

[13] D. Altshuler, C. Albers, G. Abecasis, and et al, "A global reference for human genetic variation," *Nature (London)*, vol. 526, no. 7571, 2015.

[14] R. E. Mills, C. T. Luttig, C. E. Larkins, A. Beauchamp, C. Tsui, W. S. Pittard, and S. E. Devine, "An initial map of insertion and deletion (INDEL) variation in the human genome," *Genome Research*, vol. 16, no. 9, 2006.

[15] S. Balasubramanian, L. Habegger, A. Frankish, D. G. MacArthur, R. Harte, C. Tyler-Smith, J. Harrow, and M. Gerstein, "Gene inactivation and its implications for annotation in the era of personal genomics," *Genes & development*, vol. 25, no. 1, 2011.

[16] T. Gutman, G. Goren, O. Efroni, and T. Tuller, "Estimating the predictive power of silent mutations on cancer classification and prognosis," *Npj genomic medicine*, vol. 6, no. 1, 2021.

[17] J. Chen and W. Tian, "Explaining the disease phenotype of intergenic SNP through predicted long range regulation," *Nucleic acids research*, vol. 44, no. 18, 2016.

[18] F. Paladini, M. T. Fiorillo, C. Vitulano, V. Tedeschi, M. Piga, A. Cauli, A. Mathieu, and R. Sorrentino, "An allelic variant in the intergenic region between ERAP1 and ERAP2 correlates with an inverse expression of the two genes," *Scientific reports*, vol. 8, no. 1, 2018.

[19] D. G. MacArthur and C. Tyler-Smith, "Loss-of-function variants in the genomes of healthy humans," *Human molecular genetics*, vol. 19, no. R2, 2010.

[20] S. Balasubramanian, Y. Fu, M. Pawashe, P. McGillivray, M. Jin, J. Liu, K. J. Karczewski, D. G. MacArthur, and M. Gerstein, "Using ALoFT to determine the impact of putative loss-of-function variants in protein-coding genes," *Nature communications*, vol. 8, no. 1, 2017.

[21] B. Yngvadottir, Y. Xue, S. Searle, S. Hunt, M. Delgado, J. Morrison, P. Whittaker, P. Deloukas, and C. Tyler-Smith, "A genome-wide survey of the prevalence and evolutionary forces acting on human nonsense SNPs," *American journal of human genetics*, vol. 84, no. 2, 2009.

[22] Y. Xue, A. Daly, B. Yngvadottir, M. Liu, G. Coop, Y. Kim, P. Sabeti, Y. Chen, J. Stalker, E. Huckle, J. Burton, S. Leonard, J. Rogers, and C. Tyler-Smith, "Spread of an inactive form of Caspase-12 in humans is due to recent positive selection," *American journal of human genetics*, vol. 78, no. 4, 2006.

[23] N. Van Opdenbosch and M. Lamkanfi, "Caspases in cell death, inflammation, and disease," *Immunity*, vol. 50, no. 6, 2019.

[24] J. Flannick, G. Thorleifsson, N. L. Beer, S. B. Jacobs, N. Grarup, N. P. Burtt, A. Mahajan, C. Fuchsberger, G. Atzmon, R. Benediktsson, J. Blangero, D. W. Bowden, I. Brandslund, J. Brosnan, F. Burslem, J. Chambers, Y. S. Cho, C. Christensen, D. A. Douglas, R. Duggirala, Z. Dymek, Y. Farjoun, T. Fennell, P. Fontanillas, T. Forsén, S. Gabriel, B. Glaser, D. F. Gudbjartsson, C. Hanis, T. Hansen, A. B. Hreidarsson, K. Hveem, E. Ingelsson, B. Isomaa, S. Johansson, T. Jørgensen, M. E. Jørgensen, S. Kathiresan, A. Kong, J. Kooner, J. Kravic, M. Laakso, J.-Y. Lee, L. Lind, C. M. Lindgren, A. Linneberg, G. Masson, T. Meitinger, K. L. Mohlke, A. Molven, A. P. Morris, S. Potluri, R. Rauramaa, R. Ribel-Madsen, A.-M. Richard, T. Rolph, V. Salomaa, A. V. Segrè, H. Skärstrand, V. Steinthorsdottir, H. M. Stringham, P. Sulem, E. S. Tai, Y. Y. Teo, T. Teslovich, U. Thorsteinsdottir, J. K. Trimmer, T. Tuomi, J. Tuomilehto, F. Vaziri-Sani, B. F. Voight, J. G. Wilson, M. Boehnke, M. I. McCarthy, P. R. Njølstad, O. Pedersen, L. Groop, D. R. Cox, K. Stefansson, and D. Altshuler, "Loss-of-function mutations in SLC30A8 protect against type 2 diabetes," *Nature Genetics*, vol. 46, no. 4, 2014.

[25] J. M. Whitacre, "Biological robustness: paradigms, mechanisms, and systems principles," *Frontiers in genetics*, vol. 3, 2012.

[26] M. S. Bramble, E. H. Goldstein, A. Lipson, T. Ngun, A. Eskin, J. E. Gosschalk, L. Roach, N. Vashist, H. Barseghyan, E. Lee, *et al.*, "A novel follicle-stimulating hormone receptor mutation causing primary ovarian failure: a fertility application of whole exome sequencing," *Human Reproduction*, vol. 31, no. 4, 2016.

[27] M. Jelani, S. Ahmed, M. M. Almramhi, H. S. A. Mohamoud, K. Bakur, W. Anshasi, J. Wang, and J. Y. Al-Aama, "Novel nonsense mutation in the PTRF gene underlies congenital generalized lipodystrophy in a consanguineous saudi family," *European journal of medical genetics*, vol. 58, no. 4, 2015.

[28] R. Montioli, I. Bellezza, M. A. Desbats, C. B. Voltattorni, L. Salviati, and B. Cellini, "Deficit of human ornithine aminotransferase in gyrate atrophy: Molecular, cellular, and clinical aspects," *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics*, vol. 1869, no. 1, 2021.

[29] J. Zlotogora, "Penetrance and expressivity in the molecular age," *Genetics in Medicine*, vol. 5, no. 5, 2003.

[30] A. C. Fahed, M. Wang, J. R. Homburger, A. P. Patel, A. G. Bick, C. L. Neben, C. Lai, D. Brockman, A. Philippakis, P. T. Ellinor, *et al.*, "Polygenic background modifies penetrance of monogenic variants for tier 1 genomic conditions," *Nature communications*, vol. 11, no. 1, 2020.

[31] N. Huang, I. Lee, E. M. Marcotte, and M. E. Hurles, "Characterising and predicting haploinsufficiency in the human genome," *PLoS genetics*, vol. 6, no. 10, 2010.

[32] D. R. Langbehn, R. R. Brinkman, D. Falush, J. S. Paulsen, M. Hayden, and an International Huntington's Disease Collaborative Group, "A new model for prediction of the age of onset and penetrance for huntington's disease based on CAG length," *Clinical genetics*, vol. 65, no. 4, 2004.

[33] G. E. Wright, H. F. Black, J. A. Collins, T. Gall-Duncan, N. S. Caron, C. E. Pearson, and M. R. Hayden, "Interrupting sequence variants and age of onset in huntington's disease: clinical implications and emerging therapies," *The Lancet Neurology*, vol. 19, no. 11, 2020.

[34] E. A. Boyle, Y. I. Li, and J. K. Pritchard, "An expanded view of complex traits: From Polygenic to Omnigenic," *Cell*, vol. 169, no. 7, 2017.

[35] N. R. Wray and R. Maier, "Genetic basis of complex genetic disease: the contribution of disease heterogeneity to missing heritability," *Current Epidemiology Reports*, vol. 1, no. 4, 2014.

[36] A. V. Khera, M. Chaffin, K. G. Aragam, M. E. Haas, C. Roselli, S. H. Choi, P. Natarajan, E. S. Lander, S. A. Lubitz, P. T. Ellinor, *et al.*, "Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations," *Nature genetics*, vol. 50, no. 9, 2018.

[37] E. D. Muse, S.-F. Chen, and A. Torkamani, "Monogenic and polygenic models of coronary artery disease," *Current cardiology reports*, vol. 23, no. 8, 2021.

[38] J. A. Veltman and H. G. Brunner, "*De novo* mutations in human genetic disease," *Nature Reviews Genetics*, vol. 13, no. 8, 2012.

[39] U. M. Marigorta, J. A. Rodríguez, G. Gibson, and A. Navarro, "Replicability and prediction: Lessons and challenges from GWAS," *Trends in genetics*, vol. 34, no. 7, 2018.

[40] S. Edwards, J. Beesley, J. French, and A. Dunning, "Beyond GWASs: Illuminating the Dark Road from Association to Function," *American journal of human genetics*, vol. 93, no. 5, 2013.

[41] M. D. Gallagher and A. S. Chen-Plotkin, "The post-GWAS era: From association to function," *American journal of human genetics*, vol. 102, no. 5, 2018.

[42] F. Lichou and G. Trynka, "Functional studies of GWAS variants are gaining momentum," *Nature communications*, vol. 11, no. 1, 2020.

[43] C. Chatelain, G. Durand, V. Thuillier, and F. Augé, "Performance of epistasis detection methods in semi-simulated GWAS," *BMC bioinformatics*, vol. 19, no. 1, 2018.

[44] R. Cowper-Sal·lari, M. D. Cole, M. R. Karagas, M. Lupien, and J. H. Moore, "Layers of epistasis: genome-wide regulatory networks and network approaches to genome-wide association studies," *Wiley interdisciplinary reviews. Systems biology and medicine*, vol. 3, no. 5, 2011.

[45] J. C. Denny, M. D. Ritchie, M. A. Basford, J. M. Pulley, L. Bastarache, K. Brown-Gentry, D. Wang, D. R. Masys, D. M. Roden, and D. C. Crawford, "PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene–disease associations," *BIOINFORMATICS*, vol. 26, no. 9, 2010.

[46] J. C. Denny, L. Bastarache, M. D. Ritchie, R. J. Carroll, R. Zink, J. D. Mosley, J. R. Field, J. M. Pulley, A. H. Ramirez, E. Bowton, M. A. Basford, D. S. Carrell, P. L. Peissig, A. N. Kho, J. A. Pacheco, L. V. Rasmussen, D. R. Crosslin, P. K. Crane, J. Pathak, S. J. Bielinski, S. A. Pendergrass, H. Xu, L. A. Hindorff, R. Li, T. A. Manolio, C. G. Chute, R. L. Chisholm, E. B. Larson, G. P. Jarvik, M. H. Brilliant, C. A. Mccarty, I. J. Kullo, J. L. Haines, D. C. Crawford, D. R. Masys, and D. M. Roden, "Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data," *Nature biotechnology*, vol. 31, no. 12, 2013.

[47] W.-Q. Wei, L. A. Bastarache, R. J. Carroll, J. E. Marlo, T. J. Osterman, E. R. Gamazon, N. J. Cox, D. M. Roden, and J. C. Denny, "Evaluating phecodes, clinical classification software, and icd-9-cm codes for phenome-wide association studies in the electronic health record," *PloS one*, vol. 12, no. 7, 2017.

[48] Y. Momozawa and K. Mizukami, “Unique roles of rare variants in the genetics of complex diseases in humans,” *Journal of human genetics*, vol. 66, no. 1, 2021.

[49] G. Gibson, “Rare and common variants: twenty arguments,” *Nature Reviews Genetics*, vol. 13, no. 2, 2012.

[50] K. C. Desch, A. B. Ozel, M. Halvorsen, P. M. Jacobi, K. Golden, M. Underwood, M. Germain, D.-A. Tregouet, P. H. Reitsma, C. Kearon, L. Mokry, J. B. Richards, F. Williams, J. Z. Li, D. Goldstein, and D. Ginsburg, “Whole-exome sequencing identifies rare variants in STAB2 associated with venous thromboembolic disease,” *Blood*, vol. 136, no. 5, 2020.

[51] J. Bis, X. Jian, B. Kunkle, Y. Chen, K. Hamilton-Nelson, W. Bush, W. Salerno, D. Lancour, Y. Ma, A. Renton, E. Marcora, J. Farrell, Y. Zhao, L. Qu, S. Ahmad, N. Amin, P. Amouyel, G. Beecham, J. Below, D. Campion, C. Charbonnier, J. Chung, L. Crane, C. Cruchaga, L. Cupples, J.-F. Dartigues, S. Debette, J.-F. Deleuze, L. Fulton, S. Gabriel, E. Genin, R. Gibbs, A. Goate, B. Grenier-Boley, N. Gupta, J. Haines, A. Havulinna, S. Helisalmi, M. Hiltunen, D. Howrigan, A. Ikram, J. Kaprio, J. Konrad, A. Kuzma, E. Lander, M. Lathrop, T. Lehtimäki, H. Lin, K. Mattila, R. Mayeux, D. Muzny, W. Nasser, B. Neale, K. Nho, G. Nicolas, D. Patel, M. Pericak-Vance, M. Perola, B. Psaty, O. Quenez, F. Rajabli, R. Redon, C. Reitz, A. Remes, V. Salomaa, C. Sarnowski, H. Schmidt, M. Schmidt, R. Schmidt, H. Soininen, T. Thornton, G. Tosto, C. Tzourio, S. Lee, C. Duijn, B. Vardarajan, W. Wang, E. Wijsman, R. K. Wilson, D. Witten, K. Worley, X. Zhang, C. Bellenguez, J. Lambert, M. Kurki, A. Palotie, M. Daly, E. Boerwinkle, K. Lunetta, A. DeStefano, J. Dupuis, E. Martin, G. Schellenberg, S. Seshai, A. Naj, M. Fornage, and L. Farrer, “Whole exome sequencing study identifies novel rare and common Alzheimer’s-associated variants involved in immune response and transcriptional regulation,” *Molecular psychiatry*, vol. 25, no. 8, 2018.

[52] M. J. Joyner and N. Paneth, “Promises, promises, and precision medicine,” *The Journal of clinical investigation*, vol. 129, no. 3, 2019.

[53] J. H. Moore and S. M. Williams, “Epistasis and its implications for personal genetics,” *The American Journal of Human Genetics*, vol. 85, no. 3, 2009.

[54] S. J. Andrews, B. Fulton-Howard, and A. Goate, “Interpretation of risk loci from genome-wide association studies of alzheimer’s disease,” *The Lancet Neurology*, vol. 19, no. 4, 2020.

[55] E. A. King, J. W. Davis, and J. F. Degner, “Are drug targets with genetic support twice as likely to be approved? revised estimates of the impact of genetic support for drug mechanisms on the probability of drug approval,” *PLoS genetics*, vol. 15, no. 12, 2019.

[56] M. Ingelman-Sundberg, S. Mkrtchian, Y. Zhou, and V. M. Lauschke, “Integrating rare genetic

variants into pharmacogenetic drug response predictions," *Human genomics*, vol. 12, no. 1, 2018.

[57] J. Barbeau, "PDX and personalized medicine." https://blog.crownbio.com/pdx-personalized-medicine#_ accessed 05.05.2022.

[58] X. Zheng-Bradley, I. Streeter, S. Fairley, D. Richardson, L. Clarke, and P. Flicek, "Alignment of 1000 Genomes Project reads to reference assembly GRCh38," *Gigascience*, vol. 6, no. 7, 2017.

[59] M. J. Landrum, S. Chitipiralla, G. R. Brown, C. Chen, B. Gu, J. Hart, D. Hoffman, W. Jang, K. Kaur, C. Liu, *et al.*, "ClinVar: improvements to accessing data," *Nucleic acids research*, vol. 48, no. D1, 2020.

[60] C. Gu, G. B. Kim, W. J. Kim, H. U. Kim, and S. Y. Lee, "Current status and applications of genome-scale metabolic models," *Genome biology*, vol. 20, no. 1, 2019.

[61] G. S. Hotamisligil, "Inflammation and metabolic disorders," *Nature*, vol. 444, no. 7121, 2006.

[62] M. S. Reuter, J. O. Sass, T. Leis, J. Köhler, J. A. Mayr, R. G. Feichtinger, M. Rauh, I. Schanze, L. Bähr, R. Trollmann, *et al.*, "HIBCH deficiency in a patient with phenotypic characteristics of mitochondrial disorders," *American Journal of Medical Genetics Part A*, vol. 164, no. 12, 2014.

[63] R. Agren, A. Mardinoglu, A. Asplund, C. Kampf, M. Uhlen, and J. Nielsen, "Identification of anticancer drugs for hepatocellular carcinoma through personalized genome-scale metabolic modeling," *Molecular systems biology*, vol. 10, no. 3, 2014.

[64] J. Lonsdale, J. Thomas, M. Salvatore, R. Phillips, E. Lo, S. Shad, R. Hasz, G. Walters, F. Garcia, N. Young, *et al.*, "The genotype-tissue expression (GTEx) project," *Nature genetics*, vol. 45, no. 6, 2013.

[65] F. Gatto, H. Miess, A. Schulze, and J. Nielsen, "Flux balance analysis predicts essential genes in clear cell renal cell carcinoma metabolism," *Scientific reports*, vol. 5, no. 1, 2015.

[66] A. M. Feist and B. O. Palsson, "The biomass objective function," *Current opinion in microbiology*, vol. 13, no. 3, 2010.

[67] I. Bartha, J. di Iulio, J. C. Venter, and A. Telenti, "Human gene essentiality," *Nature Reviews Genetics*, vol. 19, no. 1, 2018.

[68] N. J. O'Neil, M. L. Bailey, and P. Hieter, "Synthetic lethality and cancer," *Nature Reviews Genetics*, vol. 18, no. 10, 2017.

[69] S. Gudmundsson and I. Thiele, "Computationally efficient flux variability analysis," *BMC bioinformatics*, vol. 11, no. 1, 2010.

[70] K. L. Howe, P. Achuthan, J. Allen, J. Allen, J. Alvarez-Jarreta, M. R. Amode, I. M. Armean, A. G. Azov, R. Bennett, J. Bhai, K. Billis, S. Boddu, M. Charkhchi, C. Cummins, L. Da Rin Fioretto, C. Davidson, K. Dodiya, B. El Houdaigui, R. Fatima, A. Gall, C. Garcia Giron, T. Grego, C. Guijarro-Clarke, L. Haggerty, A. Hemrom, T. Hourlier, O. G. Izuogu, T. Juettemann, V. Kaikala, M. Kay, I. Lavidas, T. Le, D. Lemos, J. Gonzalez Martinez, J. C. Marugán, T. Maurel, A. C. McMahon, S. Mohanan, B. Moore, M. Muffato, D. N. Oheh, D. Paraschas, A. Parker, A. Parton, I. Prosovetskaia, M. P. Sakthivel, A. Salam, B. M. Schmitt, H. Schuilenburg, D. Sheppard, E. Steed, M. Szpak, M. Szuba, K. Taylor, A. Thormann, G. Threadgold, B. Walts, A. Winterbottom, M. Chakiachvili, A. Chaubal, N. De Silva, B. Flint, A. Frankish, S. E. Hunt, G. R. IIsley, N. Langridge, J. E. Loveland, F. J. Martin, J. M. Mudge, J. Morales, E. Perry, M. Ruffier, J. Tate, D. Thybert, S. J. Trevanion, F. Cunningham, A. D. Yates, D. R. Zerbino, and P. Flicek, "Ensembl 2021," *Nucleic acids research*, vol. 49, no. D1, 2021.

[71] A. Ebrahim, J. A. Lerman, B. O. Palsson, and D. R. Hyduke, "COBRApy: COnstraints-Based Reconstruction and Analysis for Python," *BMC systems biology*, vol. 7, no. 1, 2013.

[72] H. Wang, S. Marcišauskas, B. J. Sánchez, I. Domenzain, D. Hermansson, R. Agren, J. Nielsen, and E. J. Kerkhoven, "RAVEN 2.0: A versatile toolbox for metabolic network reconstruction and a case study on streptomyces coelicolor," *PLoS computational biology*, vol. 14, no. 10, 2018.

[73] R. Steinhaus, S. Proft, M. Schuelke, D. N. Cooper, J. M. Schwarz, and D. Seelow, "Mutationtaster2021," *Nucleic Acids Research*, vol. 49, no. W1, 2021.

[74] J. T. Brosnan, R. P. Da Silva, and M. E. Brosnan, "The metabolic burden of creatine synthesis," *Amino acids*, vol. 40, no. 5, 2011.

[75] C. P. Strassburg, "Hyperbilirubinemia syndromes (Gilbert-Meulengracht, Crigler-Najjar, Dubin-Johnson, and Rotor syndrome)," *Best Practice & Research Clinical Gastroenterology*, vol. 24, no. 5, 2010.

[76] M. Girard, C. Douillard, D. Debray, F. Lacaille, M. Schiff, S. Vuillaumier-Barrot, T. Dupré, M. Fabre, L. Damaj, A. Kuster, *et al.*, "Long term outcome of MPI-CDG patients on D-mannose therapy," *Journal of Inherited Metabolic Disease*, vol. 43, no. 6, 2020.

[77] T. Moriyama, S. Tanaka, Y. Nakayama, M. Fukumoto, K. Tsujimura, K. Yamada, T. Bamba, Y. Yoneda, E. Fukusaki, and M. Oka, "Two isoforms of TALDO1 generated by alternative

translational initiation show differential nucleocytoplasmic distribution to regulate the global metabolic network," *Scientific reports*, vol. 6, no. 1, 2016.

[78] J. Shen and Y. Chen, "Molecular characterization of glycogen storage disease type III," *Current molecular medicine*, vol. 2, no. 2, 2002.

[79] A. M. Almeida, Y. Murakami, A. Baker, Y. Maeda, I. A. Roberts, T. Kinoshita, D. M. Layton, and A. Karadimitris, "Targeted therapy for inherited GPI deficiency," *New England Journal of Medicine*, vol. 356, no. 16, 2007.

[80] M. H. Vaisbich, A. Braga, M. Gabrielle, C. Bueno, F. Piazzon, and F. Kok, "Thrombotic microangiopathy caused by methionine synthase deficiency: diagnosis and treatment pitfalls," *Pediatric Nephrology*, vol. 32, no. 6, 2017.

A Appendix: Gene List

Table A1: All genes mentioned, their Ensembl ID, and any SNP/rsIDs mentioned along with them. This list includes all SNPs mentioned for that gene, primarily start/stop SNPs. In some cases, the samples/results may have carried more SNPs that were not explicitly mentioned.

Gene Name	Gene ID	rsIDs
CLYBL	ENSG00000125246	rs41281112, rs3783185, rs189643142
SLCO1B1	ENSG00000134538	rs71581941
ACSM3	ENSG00000005187	rs52817836
COX6B2	ENSG00000160471	rs138900187
PTPMT1	ENSG00000110536	rs190036505
NQO1	ENSG00000181019	rs114238154
HIBCH	ENSG00000198130	rs291466
PCCA	ENSG00000175198	rs138149179, rs530307529
PCCB	ENSG00000114054	rs186031457, rs768935968, rs191375566
MMUT	ENSG00000146085	rs200596762
XYLB	ENSG00000093217	rs200154272, rs140641713, rs79233786
CD38	ENSG00000004468	rs147573494, rs139152162
TALDO1	ENSG00000177156	rs202135397
DERA	ENSG00000023697	rs572064049
SLC25A19	ENSG00000125454	rs150354936
CPT2	ENSG00000157184	rs539239516
NMRK1	ENSG00000106733	rs191532264
APRT	ENSG00000198931	rs556666445
ABCA1	ENSG00000165029	rs575627531
HSD3B7	ENSG00000099377	rs547474493
AKR1C4	ENSG00000198610	rs191144263, rs558287231
AMACR	ENSG00000242110	rs1473596504
CYP27A1	ENSG00000135929	rs575064188, rs533885672, rs188850202
FBP2	ENSG00000130957	rs77568573
MTR	ENSG00000116984	rs560603359, rs536238004
GPI	ENSG00000105220	rs201411926
AGL	ENSG00000162688	rs193186112, rs531425980, rs539108137
ALDH4A1	ENSG00000159423	rs536031585
CP	ENSG00000047457	rs200363523
MPI	ENSG00000178802	rs528828174
GNE	ENSG00000159921	rs189454495

SLC25A15	ENSG00000102743	rs104894429
HMGCS2	ENSG00000134240	rs587727388
ACSS2	ENSG00000131069	rs530201978
SLCO1A2	ENSG00000084453	rs534910521
SLCO1B3	ENSG00000111700	rs550941268
OAT	ENSG00000065154	rs200068769
HMGCR	ENSG00000113161	rs199707836
QARS1	ENSG00000172053	rs546130184
SDHB	ENSG00000117118	rs74315366
DBT	ENSG00000137992	
CYP39A1	ENSG00000146233	
TM7SF2	ENSG00000149809	
LBR	ENSG00000143815	
MTHFR	ENSG00000177000	rs1801133
BCAT2	ENSG00000105552	
BCKDHA	ENSG00000248098	
DBT	ENSG00000137992	
HADHA	ENSG00000084754	

B Appendix: Additional Results

B.1 HIBCH Results

Table B1: List of essential tasks that failed pancreas tissue model. There was some variation between tissue models. Combinations used were start/stop heterozygote.

Failed Tasks
Beta oxidation of odd-chain FA
Uptake and beta oxidation of all NEFAs
Phosphatidylcholine *de novo* synthesis
Phosphatidylethanolamine *de novo* synthesis
Phosphatidylserine *de novo* synthesis
Phosphatidylinositol *de novo* synthesis

B.2 Results on HG01809

Table B2: List of additional tasks that failed in sample HG01809. All failures occurred in the liver tissue model. Gene combination used was start/stop heterozygote.

Failed Tasks
Gluconeogenesis from Alanine
Arginine *de novo* synthesis (minimal substrates, minimal excretion)
Ornithine degradation
Urea from alanine
Urea from glutamine
Creatine *de novo* synthesis (minimal substrates, physiological excretion)
NH_3 import and degradation

B.3 Results on HG00127

Table B3: List of essential tasks that failed in multiple tissues for sample HG00127. Some tissues had fewer tasks that failed, but all had some.

Failed Tasks
ATP *de novo* synthesis
CTP *de novo* synthesis
GTP *de novo* synthesis
UTP *de novo* synthesis
dATP *de novo* synthesis
dCTP *de novo* synthesis
dGTP *de novo* synthesis
dTTP *de novo* synthesis
Oxidative phosphorylation
Krebs cycle NADH
Beta oxidation of saturated FA
Beta oxidation of long-chain FA
Beta oxidation of odd-chain FA Beta oxidation of unsaturated fatty acid (n-9)
Beta oxidation of unsaturated fatty acid (n-6)
Uptake and beta oxidation of all NEFAs
Phosphatidylcholine *de novo* synthesis
Phosphatidylethanolamine *de novo* synthesis
Phosphatidylserine *de novo* synthesis
Phosphatidylinositol *de novo* synthesis

B.4 Results on HG02078

Table B4: SNPs for the four gene combinations causing task failures in sample HG02078. Also included is the SNPs amino acid effect on the protein. Gene combination used was all missense homozygote.

rsID	Gene	Protein Pos.
DBT	rs12021720	S384G
AMACR	rs2287939	L201S
AMACR	rs3195676	V9M
AMACR	rs10941112	G175D
AMACR	rs2278008	E277K
CYP39A1	rs7761731	N324K
HSD3B7	rs567144052	P221A
HSD3B7	rs9938550	T250A
TM7SF2	rs11539360	A119V
LBR	rs2230419	S154N

NTNU

Norwegian University of
Science and Technology