The handle http://hdl.handle.net/1887/38350 holds various files of this Leiden University dissertation.

Author: Dharuri, Harish
Title: Bioinformatic approaches to identify genomic, proteomic and metabolomic biomarkers for the metabolic syndrome
Issue Date: 2016-03-02
Summary

Advances in technology have turned modern biology into a data-intensive enterprise. The advent of high-output technologies like Microarrays and Next-generation sequencing technologies has resulted in researchers grappling not just with huge volumes but also multiple types of data. While generation and storage of high-quality data are an important research focus, it is increasingly recognized that translating data into actionable information and insight is a critical research challenge. To infer reliable conclusions from the data, it is often necessary to integrate large amounts of heterogeneous data with different formats and semantics. Given the breadth and volume of data involved, this goal is best achieved through automated methods and tools for data integration and workflow management. This thesis presents automated strategies that combine bioinformatics and statistical methods to identify novel biomarkers in high-throughput OMICs datasets pertaining to the metabolic syndrome and to gain mechanistic insight into the underlying biological processes. An underlying theme in this thesis is data-driven approaches that generate plausible hypothesis which is followed by experimental verification.

The main findings in each of the chapters are summarized below.

Genome-wide association studies of metabolite profiles explain a higher percentage of genetic variation and have larger effect sizes than clinical phenotypes and traits. However, given the large number of metabolites measured, these studies come with a large multiple testing penalty. In Chapter 2 we present an automated workflow approach that utilizes prior knowledge of biochemical pathways present in databases like KEGG and BioCyc to generate smaller gene sets relevant to the metabolite. To retrieve a prioritized list of candidate genes associated with metabolite levels, gene sets were generated for each metabolite by identifying genes that participate in pathways and reactions relevant to the synthesis or degradation of the metabolite. For every gene set, a corresponding SNP set was generated by retrieving SNPs within the flanking 50 kb of every gene. Re-analysis of a published GWAS dataset using the metabolite specific SNP sets confirmed previously identified hits and identified a new locus of human metabolic individuality, associating Aldehyde dehydrogenase family 1 L1 (ALDH1L1) with serine / glycine ratios in blood. The workflow paradigm described in chapter 2 is gaining ground in bioinformatics as the technology of choice for
recording the steps of computational experiments. In a typical workflow, data outputs are generated from data inputs via a set of (potentially distributed) computational tasks that are coordinated following a workflow definition. However, workflows do not provide a complete solution for aggregating all data and all meta-data that are necessary for understanding the full context of an experiment. Encapsulating all aspects of an in silico analysis and communicating it to the scientific community is a key challenge in a computational experiment. Chapter 3 explores the utility of semantic web technologies in the preservation of computational experiments. Semantic web technologies facilitate the integration of heterogenous data on the World Wide Web by making the semantics of the data explicit through formal ontologies. More specifically, the chapter discusses the Research Object (RO) model, where a research object is defined as a resource that aggregates other resources, e.g. datasets, software, spreadsheets, text, etc. The overarching goal of the RO model is to facilitate transparency and reproducibility of scientific studies. The RO model was applied to a study where the goal was to facilitate the interpretation of the results of a GWAS of metabolite profiles. Applying a workflow-centric RO model to aggregate and annotate the resources used in the bioinformatics experiment, allowed us to retrieve the conclusions of the experiment in the context of the driving hypothesis, the executed workflows and their input data.

Obesity results in decreased life expectancy due to associated metabolic and cardiovascular disorders, as well as several types of cancer. A majority of obese individuals develop insulin resistance and type-2 diabetes. However, approximately 10-25% of these individuals seem to remain insulin sensitive and Normal Glucose Tolerant (NGT). Studies have shown that the expanded adipose tissue serves as an important pathogenic site in the development of type 2 diabetes. Chapter 4 presents a study designed to investigate the role of the adipose tissue in development of T2DM in severely obese subjects, by performing RNA-Sequencing of the subcutaneous (SAT) and visceral adipose tissue (VAT) samples obtained during bariatric surgery. The sets of expressed genes were subjected to a gene network-based approach to distinguish obese individuals with NGT from obese individuals with type 2 diabetes. This identified acetyl-CoA metabolic network down-regulation as an important feature in the pathophysiology of obese individuals with type 2 diabetes. In general, genes within two reaction steps of acetyl-CoA were found to be down-regulated in the VAT and SAT of individuals with type 2 diabetes. Upon weight loss and amelioration of metabolic abnormalities three months following bariatric surgery, the expression level of these genes recovered to
levels seen in NGT individuals. We report four novel genes associated with type-2 diabetes and recovery upon weight loss: acetyl-CoA acetyltransferase 1 (ACAT1), acetyl-CoA carboxylase alpha (ACACA), aldehyde dehydrogenase 6 family, member A1 (ALDH6A1) and methylenetetrahydrofolate dehydrogenase (MTHFD1). In addition to confirming earlier findings by other groups on the role of branched-chain amino acid degradation, fatty acid oxidation and citrate cycle in type-2 diabetes, we show through a network analysis that acetyl-CoA metabolism is the unifying principle and that its dysregulation distinguishes between obese women with type-2 diabetes and those with NGT.

Next generation RNA-sequencing technology has made it possible to quantify gene expression but also to use the sequence itself to identify expressed alleles by calling haplotypes of an individual based on heterozygosity of SNPs in expressed loci. Allele-specific expression studies help to understand the cis-regulatory basis of variation in gene expression. In Chapter 5, we investigate the hypothesis that cis-regulatory variants differentially affect gene expression in visceral and subcutaneous adipose tissue. We investigated differential allele-specific expression between visceral and subcutaneous adipose tissue of very obese individuals (BMI>40) with and without type 2 diabetes mellitus with the aim of identifying regulatory variants that could explain the pathophysiological differences observed in the two tissues. The objective of the study was to identify from a panel of known genome-wide association hits the subset of common variants that are under the control of cis-regulatory elements and to assess the consequence of such variants on the T2DM phenotype. We identified a single nucleotide polymorphism (SNP) rs1049174, in the 3' untranslated region (3' UTR) in KLRK1 (Killer cell lectin like receptor subfamily K, family member 1) gene that displays a significant differential allelic expression between VAT and SAT, and for which expression is different between individuals with normal glucose tolerance (NGT) and T2D. The differential allele-specific expression of KLRK1 between visceral and subcutaneous adipose tissue and the increased expression of KLRK1 in visceral adipose tissue of very obese individuals with type 2 diabetes provides evidence for a role of KLRK1 in the susceptibility to type-2 diabetes.

Very low calorie diets (VLCD) with and without exercise programs lead to major metabolic improvements in obese type 2 diabetes patients. In Chapter 6, we investigate the mechanisms of a VLCD with or without exercise to uncover possible biomarkers associated with these interventions. In the first step, targeted multiple reaction monitoring (MRM) analysis was conducted
for 13 abundant proteins hypothesized to be associated with T2DM and obesity, including apolipoproteins and markers of inflammation and coagulation. Subsequently, a large scale isobaric tag for relative and absolute quantification (iTRAQ) approach was utilized to uncover differences between the VLCD with and without exercise groups for less abundant proteins. Using proteomic analysis several potential disease state and intervention associated markers were found distinguishing T2DM patients from obese and lean controls and showing a VLCD effect.

In Chapter 7 we explore the prolonged effects of niacin on gene expression profile in adipocytes of a hyperlipidemic mouse model. We applied bioinformatic and statistical analyses to the gene expression data and showed that prolonged niacin treatment led to an increase in the poly unsaturated fatty acid (PUFA) synthesis pathways. To investigate whether PUFA levels and possible derivatives thereof (i.e. oxylipins) were functionally affected, we determined the fatty acid composition in the adipose tissue. These analyses revealed increased n-3 PUFA secretion from the adipocytes and an increased plasma level of n-3 PUFAs and their anti-inflammatory oxylipins. Together with the up-regulation of the PUFA biosynthesis pathway in gWAT, this point towards an atheroprotective plasma profile induced by prolonged niacin treatment.

In Chapter 8, we present a global review of the current status of metabolomic GWAS (mGWAS). The first waves of metabolomics and genetic analyses by mGWAS have provided a wealth of insight into the genetic basis of metabolic individuality and risk factors for common metabolic disorders. However, we still face many hurdles in the interpretation of mGWAS data. Metabolomics platforms generally yield information on the levels of one to several hundreds of metabolites. Consideration of all metabolites results in a severe multiple testing burden. This precludes genuine SNP-metabolite pairs from being considered when they fail to reach the stringent statistical threshold for significance. Pathway analysis is exquisitely suited to increase the statistical power to identify biologically plausible loci and simultaneously improve our understanding of the underlying biological mechanisms. In addition, the next step in pathway analysis is to include stoichiometric and kinetic parameters and complement the statistical analysis with a more comprehensive systems biology based approach using mathematical modelling. The application of a priori knowledge present in databases and the potential of mathematical models in enhancing the interpretation of mGWAS are presented.
The thesis concludes with Chapter 9 where future developments in the discipline are outlined.