Computational Processing of Omics Data:

Implications for Analysis

by

Ashlee M. Benjamin

Graduate Program in Computational Biology and Bioinformatics
Duke University

Date: _____

Approved:

_____
Joseph Lucas, Co-supervisor

_____
Gregory A. Wray, Co-supervisor

_____
M. Arthur Moseley

_____
Barbara Engelhardt

_____
Raluca Gordan

Dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Graduate Program of Computational Biology and Bioinformatics
in the Graduate School of Duke University
2013

Abstract

Computational Processing of Omics Data: Implications for Analysis

by

Ashlee M. Benjamin

Graduate Program in Computational Biology and Bioinformatics
Duke University

Date: _____

Approved:

_____
Joseph Lucas, Co-supervisor

_____
Gregory A. Wray, Co-supervisor

_____
M. Arthur Moseley

_____
Barbara Engelhardt

_____
Raluca Gordan

An abstract of a dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Graduate Program of Computational Biology and Bioinformatics
in the Graduate School of Duke University
2013

Abstract

In this work, I present four studies across the range of 'omics data types - a Genome-Wide Association Study for gene-by-sex interaction of obesity traits, computational models for transcription start site classification, an assessment of reference-based mapping methods for RNA-Seq data from non-model organisms, and a statistical model for open-platform proteomics data alignment.

Obesity is an increasingly prevalent and severe health concern with a substantial heritable component, and marked sex differences. We sought to determine if the effect of genetic variants also differed by sex by performing a genome-wide association study modeling the effect of genotype-by-sex interaction on obesity phenotypes. Genotype data from individuals in the Framingham Heart Study Offspring cohort were analyzed across five exams. Although no variants showed genome-wide significant gene-by-sex interaction in any individual exam, four polymorphisms displayed a consistent BMI association (P-values .00186 to .00010) across all five exams. These variants were clustered downstream of LYPLAL1, which encodes a lipase/esterase expressed in adipose tissue, a locus previously identified as having sex-specific effects on central obesity. Primary effects in males were in the opposite direction as females and were replicated in Framingham Generation 3. Our data support a sex-influenced association between genetic variation at the LYPLAL1 locus and obesity-related traits.

The application of deep sequencing to map 5' capped transcripts has confirmed

the existence of at least two distinct promoter classes in metazoans: focused promoters with transcription start sites (TSSs) that occur in a narrowly defined genomic span and dispersed promoters with TSSs that are spread over a larger window. Previous studies have explored the presence of genomic features, such as CpG islands and sequence motifs, in these promoter classes, and our collaborators recently investigated the relationship with chromatin features. It was found that promoter classes are significantly differentiated by nucleosome organization and chromatin structure. Here, we present computational models supporting the stronger contribution of chromatin features to the definition of dispersed promoters compared to focused start sites. Specifically, dispersed promoters display enrichment for well-positioned nucleosomes downstream of the TSS and a more clearly defined nucleosome free region upstream, while focused promoters have a less organized nucleosome structure, yet higher presence of RNA polymerase II. These differences extend to histone variants (H2A.Z) and marks (H3K4 methylation), as well as insulator binding (such as CTCF), independent of the expression levels of affected genes.

The application of next-generation sequencing technology to gene expression quantification analysis, namely, RNA-Sequencing, has transformed the way in which gene expression studies are conducted and analyzed. These advances are of particular interest to researchers studying organisms with missing or incomplete genomes, as the need for knowledge of sequence information is overcome. *De novo* assembly methods have gained widespread acceptance in the RNA-Seq community for organisms with no true reference genome or transcriptome. While such methods have tremendous utility, computational complexity is still a significant challenge for organisms with large and complex genomes. Here we present a comparison of four reference-based mapping methods for non-human primate data. We utilize Bowtie2 and Stampy for mapping to the human transcriptome, and TopHat2 and GSNAP for mapping to the human genome. To compare the methods, we explore mapping rates, mapping loca-

tions, number of detected genes, correlations between computed expression values, and the utility of the resulting data for differential expression analysis. We show that reference-based mapping methods indeed have utility in RNA-Seq analysis of mammalian data with no true reference, and that the details of mapping methods should be carefully considered when doing so. We find that shorter seed sequences, allowance of mismatches, and allowance of gapped alignments, in addition to splice junction gaps result in more sensitive alignments of non-human primate RNA-Seq data.

Open-platform proteomics experiments seek to quantify and identify the proteins present in biological samples. Much like differential gene expression analyses, it is often of interest to determine how protein abundance differs in various physiological conditions. Label free LC-MS/MS enables the rapid measurement of thousands of proteins, providing a wealth of peptide intensity information for differential analysis. However, the processing of raw proteomics data poses significant challenges that must be overcome prior to analysis. We specifically address the matching of peptide measurements across samples - an essential pre-processing step in every proteomics experiment. We present a novel method for label-free proteomics data alignment with the ability to incorporate previously unused aspects of the data, particularly ion mobility drift times and product ion information. We compare the results of our alignment method to PEPPeR and OpenMS, and compare alignment accuracy achieved by different versions of our method utilizing various data characteristics. Our method results in increased match recall rates and similar or improved mismatch rates compared to PEPPeR and OpenMS feature-based alignment. We also show that the inclusion of drift time and product ion information results in higher recall rates and more confident matches, without increases in error rates. Based on these results, we argue that the incorporation of ion mobility drift time and product ion information are worthy pursuits. In addition, alignment methods should be

flexible enough to utilize all available data, particularly with recent advancements in experimental separation methods. The incorporation of drift times and/or high energy into alignment methods and accurate mass and time (AMT) tag databases can greatly improve experimenters ability to identify measured peptides, reducing analysis costs and potentially the need to run additional experiments.

For RRV and PSV.

Contents

List of Tables

List of Figures

List of Abbreviations and Symbols

Symbols

d	Index of a proteomics dataset or sample.
i	Index of a measured peptide.
$x_{d,i}$	Measured mass-to-charge ratio, retention time, and drift time vector for peptide i of dataset/sample d.
η_d	Dataset-specific mass-to-charge ratio, retention time, and drift time shift vector for dataset d.
β_d	Dataset-specific mass-to-charge ratio, retention time, and drift time scale vector for dataset d.
$c_{d,i}$	Indicator variable for the latent peptide assignment of $x_{d,i}$.
$z_{c_{d,i}}$	Mass-to-charge ratio, retention time, and drift time vector of the latent peptide to which $x_{d,i}$ is assigned.
$\epsilon_{d,i}$	Residual mass-to-charge ratio, retention time, and drift time vector for $x_{d,i}$ shifted and scaled approximation of $z_{c_{d,i}}$.
Σ	Covariance matrix of the mass-to-charge ratio, retention time, and drift time residuals.
$\mu_{c_{d,i}}$	Mean mass-to-charge ratio, retention time, and drift time vector of latent peptide $z_{c_{d,i}}$.
σ	Covariance matrix of the mass-to-charge ratio, retention time, and drift time of a latent peptide. This covariance is shared among all latent peptides.
a	Prior mean for shift parameters.
b	Prior covariance for shift parameters.
e	Prior mean for scale parameters.

f	Prior covariance for scale parameters.
λ	Prior mean for latent peptide means.
r	Prior covariance for latent peptide means.
g	Prior scale matrix for latent peptide covariance.
h	Prior degrees of freedom for latent peptide covariance.
$y_{d,i}$	Product ion intensity profile vector for measured peptide $x_{d,i}$.
$w_{c_{d,i}}$	Product ion intensity profile vector for latent peptide $z_{c_{d,i}}$.
$\psi_{d,i}$	Sum of squared differences between the measured product ion intensity profile $y_{d,i}$ and the latent product ion intensity profile $w_{c_{d,i}}$.
γ	Rate parameter for the exponential distribution of $\psi_{d,i}$.
a_0	Prior shape for the rate parameter.
b_0	Prior scale for the rate parameter.

Abbreviations

AMT	Accurate mass and time.
auPRC	Area under the PRC.
auROC	Area under the ROC curve.
BMI	Body Mass Index.
BP	Broad with peak.
CAGE	Cap Analysis of Gene Expression.
DDA	Data Dependent Acquisition
DIA	Data Independent Acquisition
DPGMM	Dirichlet process Gaussian mixture model.
GWAS	Genome-wide association study.
GxS	Gene by sex interaction.
HCV	Hepatitis-C virus.

$HDMS^E$	Waters Corporation high-low switching fragmentation mass spectrometry coupled with ion mobility separation.
IM	Ion mobility.
LD	Linkage Disequilibrium.
MAP	Maximum a posteriori.
MCMC	Markov chain Monte Carlo.
ME	Main effect.
MS^E	Waters Corporation high-low switching fragmentation mass spectrometry.
MS/MS	Tandem mass spectrometry.
MZ-IM	Mass-to-charge ratio and ion mobility drift time alignment.
MZ-RT	Mass-to-charge ratio and retention time alignment.
MZ-RT-IM	Mass-to-charge ratio, retention time, and ion mobility drift time alignment.
MZ-RT-IM-HE	Mass-to-charge ratio, retention time, ion mobility drift time, and high-energy data alignment.
NFR	Nucleosome free region.
NGS	Next Generation Sequencing.
NP	Narrow peak.
PCA	Principal Component Analysis
PEAT	Paired end analysis of transcription start sites.
PLGS	ProteinLynx Global SERVER, The Waters informatics platform for processing proteomics data.
pol II	RNA polymerase II.
PRC	Precision recall curve.
QQ	Quantile-quantile.
RTCC	Retention time calibration curve.
ROC	Receiver operating characteristics.

SNP	Single nucleotide polymorphism.
TSS	Transcription start site.
WC	Waist Circumference.
WHR	Waist to Hip Ratio.
WP	Weak peak.

Acknowledgements

I would like to thank the NIH (training Grant T32 GM071340, and Clinical and Translational Science Award 1UL1RR024128-01), and the Defense Advanced Research Projects Agency (DARPA), number lN66001-07-C-0092 (G.S.G.) for generous financial support. I would also like to acknowledge the American Society for Mass Spectrometry for assisting with conference travel expenses, allowing me to present my work.

Special thanks to Sunil Suchindran, Derek Cyr, and Ricardo Henao for their mentorship during my time with the Lucas Lab. Thank you to Will Thompson, Erik Soderblom, and the other members of the Duke Proteomics Core Facility for helping me gain a better understanding of open-platform proteomics data, and providing valuable experimental view points. Thank you to Virginia Kraus for granting me access to information-rich Osteoarthritis proteomics data. Thank you to my fellow CBB students - particularly the entering class of 2009 - for providing a strong student support community. Thank you to Jeanette McCarthy, and Uwe Ohler for outstanding mentorship during my first year at Duke. Thank you to Liz Labriola for always being there to help with all things administrative, and being a sympathetic ear to all problems - graduate school or otherwise. Thank you to all members of IGSP-IT, and the administrators of DSCR. This work would not have been possible without the availability of such computational resources. Thank you to Thomas Kepler and Merlise Clyde for serving on my committee in the early months of my dissertation,

and for my preliminary exam. Thank you to Geoffrey Ginsburg, Tim Veldman, Lori Hudson, Tom Burke, Marshall Nichols, and the rest of the CGSU team for welcoming me in to your exciting work over the past year.

Last but certainly not least, thank you to all of my current committee members - Joe Lucas, Arthur Moseley, Greg Wray, Barbara Engelhardt, and Raluca Gordan for your tremendous guidance and support.

1

Introduction

1.1 Background

Studies of all aspects of biological systems, including genetic variation, transcription initiation, gene expression, and protein abundance, provide a wealth of information advancing our understanding of life science. As these biological data become increasingly high throughput and complex, the need for advanced computational data processing and data analysis methods continues to rise.

1.2 GWAS

The ability to genotype large numbers of single nucleotide polymorphisms (SNPs) via microarrays has paved the way for association studies between phenotypes and genetic variants. Genome-Wide Association Studies (GWAS) have been utilized to study allelic variation underlying many diseases and other phenotypic differences, advancing genetics beyond small scale candidate gene studies McCarthy and Hirschhorn (2008); Wang et al. (2005). Disease phenotypes result from complex interactions between genetic makeup and environmental factors. Identifying genetic variants under-

lying disease risks can greatly advance biological understanding and clinical course of action.

1.2.1 Selecting Tag SNPs

An important consideration of GWAS is the selection of a representative set of SNPs for which genotypes will be obtained. The presence of Linkage Disequilibrium (LD), a nonrandom association of alleles at nearby loci, allows inference of alleles for additional SNPs that are in strong LD with one or more nearby genotyped loci. Selecting these tag SNPs has become a standard practice in GWAS, with the goal of maximizing the represented variation in any genomic region Wang et al. (2005).

1.2.2 Considering Allele Frequency

The power to detect phenotypic associations for complex traits is a function of sample size, genetic effect size, and allele frequency at both causal and marker loci Wang et al. (2005); Stranger et al. (2011). GWAS require larger sample sizes to detect phenotypic associations for more rare variants. Two main hypotheses have driven GWAS - the common disease/common variant (CDCV) hypothesis, and the common disease/rare variant (CDRV) hypothesis. The CDCV hypothesis suggests that common diseases are the result of common variants. The CDRV hypothesis suggests that common diseases are a result of many variants that differ between individuals, and have low overall population frequencies Wang et al. (2005); Stranger et al. (2011). Studies operating under the CDCV hypothesis would need relatively few tag SNPs to represent LD blocks of variation and capture the effect of a common variant Stranger et al. (2011). Rare variants require deeper genotyping. The 1000 genomes project has been driving the effort to generate a resource for common and rare genetic variation studies, now containing genomes from 1092 individuals 1000 Genomes Project Consortium et al. (2012).

1.2.3 Considering Population Structure

Careful study design is necessary to prevent false or missed associations in GWAS. In case-control studies, a powerful and easily collected study group, the cohort may contain population stratification, or systematic differences in allele frequency between subpopulations Wang et al. (2005); McCarthy and Hirschhorn (2008); Stranger et al. (2011). While concerning, several methods have been developed to correct for allele frequency differences due to ancestry. Such methods include a variety of principal component analysis (PCA) based corrections Hoggart et al. (2003); Satten et al. (2001); Price et al. (2006a), as well as the use of ancestry-informative markers Tian et al. (2008).

1.2.4 Multiple Hypothesis Testing

GWAS perform association tests for hundreds of thousands of SNPs with a given phenotype, often utilizing linear regression for continuous traits, and logistic regression for categorical traits. The null hypothesis in these association tests is that a given variant is not associated with the trait of interest Wang et al. (2005); Stranger et al. (2011). Because these tests are performed on such a large number of loci, correction for multiple hypothesis testing is necessary. The conservative Bonferroni correction has been adopted as standard practice in GWAS, requiring a p-value of less than 5e-8 to reach "genome-wide significance" Stranger et al. (2011).

1.2.5 GWAS Results and Missing Heritability

GWAS provide researchers with association signal across the genome. It is often difficult to identify the causal gene, and very rare to identify the causal variant McCarthy and Hirschhorn (2008); Stranger et al. (2011). The associated locus, however, does serve as a marker for the haplotype containing the causal variant. While GWAS have contributed many identifications of important regions of variation,

the majority of associations only account for a small proportion of heritability. This may be due to overestimated heritability, incomplete LD between marker and causal variants, undetected contributions from rare genetic variants, or simply many small undetected contributions Stranger et al. (2011). Another possible explanation for the limited discovery of heritability, is the large focus of GWAS on populations of European decent McCarthy and Hirschhorn (2008); Stranger et al. (2011). In general, association studies have found that the genetic variability underlying common disease are the product of many small effects. In the cases where missing heritability is in fact due to rare variants with large effects, recent advances in sequencing technology and the 1000 genomes project will provide substantial insight Stranger et al. (2011).

1.3 Transcriptomics via NGS

In recent years, Next Generation Sequencing (NGS) technology has been utilized in many ways to obtain high throughput genomics and transcriptomics data - including but not limited to whole small genomes Walker et al. (2013); Didelot et al. (2012), genetic variants Veltman and Brunner (2012); Bamshad et al. (2011); Goldstein et al. (2013), cancer genomes Mwenifumbo and Marra (2013); Meyerson et al. (2010), genomes of microbial communities Cho and Blaser (2012), gene expression quantities Wang et al. (2009); Ozsolak and Milos (2011), RNA-protein interactions König et al. (2011), DNA methlyation patterns Laird (2010), locations of DNA-binding proteins Park (2009); Zhou et al. (2011), and locations of transcription start sites Ng et al. (2005). The development of these massively parallel sequencing methodologies allow experimenters to obtain remarkable amounts of sequence information in very little time.

Identifying transcription start site (TSS) loci is an essential part of understanding transcript regulation and expression. CAGE (Cap Analysis of Gene Expression) is a high-throughput method used to identify TSS in a given sample Shiraki et al.

(2003). The general CAGE protocol includes extraction of total RNA, first strand cDNA synthesis from the RNA library, 5' capture of cDNAs or "cap-trapping", 5' ligation of biotinylated linkers, removal of RNA, second strand cDNA synthesis, cleavage of the first 20 bases by restriction enzymes, 3' ligation of linkers, amplification, and sequencing. CAGE results in DNA tags from the first 20 bases of the 5' end of mRNAs, which can be mapped to a reference genome to identify the TSS and also reflect the expression level of the originating transcript Ng et al. (2005); Nilsson and Virtanen (2006). Extensions of CAGE have been developed to improve mapping specificity and throughput. Such extensions include deepCAGE Valen et al. (2009), Paired End Analysis of Transcription Start Sites (PEAT) Ni et al. (2010), nanoCAGE, and CAGEscan Plessy et al. (2010). The main computational processing of CAGE sequence data includes mapping to a reference genome, and normalization Balwierz et al. (2009).

An important application of NGS data gaining popularity is RNA-Sequencing experiments, or RNA-Seq. RNA-Seq utilizes NGS technologies to obtain expression information for any prepared RNA library. RNA-Seq allows for untargeted gene expression (or non-mRNA expression) analysis in multiple tissues. RNA molecules of interest are isolated from samples, and converted to a cDNA fragment library via reverse transcription and fragmentation. Prior to cDNA synthesis, samples are often depleted for highly abundant ribosomal RNA molecules. This depletion step allows more sequencing reactions to be devoted to mRNA molecules, or other ncRNA targets Martin and Wang (2011). Fragments undergo size selection based on the RNA molecules being targeted, and specific NGS technology. Finally, adaptors are ligated to one or both ends of each cDNA fragment, and each fragment is amplified and sequenced to obtain short reads from one end (single-end sequencing) or both ends (paired-end sequencing) Wang et al. (2009).

RNA-Seq offers several advantages over hybridization-based gene expression quan-

tification approaches. Hybridization-based approaches to transcriptome profiling involve hybridizing fluorescently labeled cDNA to custom-made or commercial microarrays. These methods are high throughput and relatively inexpensive, but rely on knowledge of genomic sequence, result in high background noise levels, and thus have a relatively low dynamic range Wang et al. (2009). RNA-Seq overcomes the limitation to detecting known transcripts, has the ability to reveal sequence variation in transcribed regions, has relatively low background noise, and has a large dynamic range Ozsolak and Milos (2011); Zheng et al. (2011); Pickrell et al. (2010); Li et al. (2010). However, processing of RNA-Seq data presents several bioinformatics challenges, many of which are not fully addressed. As with any other high throughput sequencing data, informaticians must be able to store and process large datasets. Aside from the typical data size challenges, RNA-Seq data requires several computational processing steps including quality control analyses, mapping of reads to a reference genome or transcriptome, transcriptome reassembly, expression quantification, and normalization Wang et al. (2009); Garber et al. (2011); Martin and Wang (2011). These computational data processing protocols for RNA-Seq data are still in their adolescence, with questions raised as to which are the optimal methods for each processing step Oshlack and Wakefield (2009).

1.3.1 Computational Challenges Raised by Library Construction

RNA-Seq data are subject to some biases that may be introduced during library generation. Resulting abundance estimates are a function of true transcript expression levels, and preferential sequence selection in library preparation protocols. Such biases include transcript coverage bias, nucleotide composition bias, GC bias and PCR bias Wang et al. (2012); Roberts et al. (2011); Mortazavi et al. (2008). The order of the reverse transcription and fragmentation steps when generating cDNA fragments can lead to a coverage bias over the length of transcripts. These cDNA fragments are

generated in one of two ways RNA fragmentation followed by reverse transcription, or reverse transcription of whole RNA molecules followed by cDNA fragmentation. In the former, RNA molecules are fragmented prior to reverse transcription resulting in even 5' to 3' coverage of the majority of the transcript, but depleted transcript ends. In cDNA fragmentation, RNA molecules are reverse transcribed prior to fragmentation, resulting in a 3' bias as the reverse transcriptase falls off each transcript Wang et al. (2009). Reverse transcription may introduce a nucleotide composition bias. This step is often initiated with random hexamer primers, which are subject to selective pressure with respect to priming efficacy. As a result, the priming is not truly random and a nucleotide composition bias is observed in the first several bases of each read Hansen et al. (2010). In addition, oligo dT priming of reverse transcription can lead to a 3' coverage bias across transcript lengths. Some recent studies have reported correlation between GC content and expression levels Pickrell et al. (2010); Benjamini and Speed (2012). It is currently thought that PCR amplification is the main source of GC bias Aird et al. (2011); Benjamini and Speed (2012). Data generated in absence of an amplification step show reduced GC bias Aird et al. (2011); Benjamini and Speed (2012). As with genomic sequence libraries, amplification bias exists, although the detection and removal of PCR duplicates is less straightforward with RNA-Seq data. The general assumption is that many identical short reads are the result of abundant RNA fragments, when in fact some may be PCR artifacts. Removing these predicted PCR duplicates is still disputed in RNA-Seq Analysis because it is very challenging to distinguish PCR duplicates from abundant RNA fragments Wang et al. (2009). Correcting for such biases results in improved mapping and gene expression estimates Mortazavi et al. (2008); Roberts et al. (2011).

1.3.2 Read Quality Control

As with any sequence data, and important pre-processing step is quality control assessment. Before proceeding with read mapping, transcriptome reconstruction, and expression quantification, it is essential to ensure reads are high quality and free of contaminantsMartin and Wang (2011). Elegant software exists for QC of raw sequence data Andrews (2010) prior to mapping procedures. Such tools perform several quality checks including sequence quality, GC bias, and nucleotide composition bias. Such quality assessments can indicate poor library preparation or failed sequencing reactions. More recently, QC software aimed at RNA-Seq data has been developed Wang et al. (2012); DeLuca et al. (2012); Lassmann et al. (2011); Planet et al. (2012), including important metrics of mapped reads in quality analysis. For example, RNA-SeQC provides several post-mapping quality metrics including alignment rates, rRNA content, mapping locations (intron, exon, intergenic), transcript coverage, number of detected transcripts, and correlation between samples. Post-mapping metrics are important to consider, as adequate coverage and sequencing depth are necessary to accurately profile transcript expression, detect splice variants, and identify novel isoforms Wang et al. (2012); DeLuca et al. (2012).

1.3.3 Challenges of Mapping

After a set of high quality sequences is obtained, reads are either assigned to transcripts by mapping to a reference genome or transcriptome, or assembled without the guidance of reference sequence. Methods utilizing a reference genome or transcriptome are often termed reference-based alignment approaches, while methods assembling transcripts using only the reads themselves are called de novo assembly approaches.

In reference-based alignment methods, the sequence for each read (and it's mate in paired-end data) is used to find potential mapping locations by exact match or

scoring sequence similarity. A read mapped to a given transcript indicates its origin, and the number of reads originating from a given transcript informs on how much of the transcript was present in the sample. While many short reads will easily map to a contiguous region on the genome, the mapping of other reads is less straightforward. Because RNA-Seq reads originate from spliced transcripts, reads can span exon junctions. The most straightforward approaches utilize "unspliced aligners" to align reads to a reference transcriptome Li et al. (2008, 2009); Li and Durbin (2009); Langmead et al. (2009); Rumble et al. (2009); Lunter and Goodson (2011); Langmead and Salzberg (2012). This alleviates the need to handle splice junctions, but is limited to the analysis of known transcripts. Unspliced read aligners are generally divided into two subcategories based on their methodology seed methods and Burrows-Wheeler transform methods Garber et al. (2011). Seed methods align short subsequences, or seeds, from each read to a reference, requiring a perfect match in the seed subsequence. More sensitive alignment methods are used to eliminate candidate regions where seeds cannot be extended to full read alignments. Burrows-Wheeler transform methods create a Burrows-Wheeler index of the reference genome and efficiently search for perfect matches. Mismatches can be allowed with an exponential increase in computational complexity. In general Burrows-Wheeler transform methods are faster than seed methods, but seed methods provide increased sensitivity Martin and Wang (2011).

A more common reference-based approach is alignment of reads to the reference genome with "spliced aligners". Spliced aligners accommodate junction-spanning reads by splitting them up into smaller segments and determining the best match based on alignment scores and known di-nucleotide splice signals De Bona et al. (2008); Trapnell et al. (2009); Au et al. (2010); Wang et al. (2010); Wu and Nacu (2010). The spliced aligners also fall into two major categories based on their methodology exon first methods and seed and extend methods. Exon first methods begin

by mapping whole reads to the genome using unspliced read aligners, and then search for spliced alignments with the remaining reads. Exon first approaches are efficient, but may miss true spliced alignments when an unspliced alignment is available in a pseudogene Garber et al. (2011). Seed and extend methods break reads into seeds which are mapped to the genome, and much like seed-based unspliced aligners, candidate mapping locations are examined with more sensitive alignment methods. Iterative extension and merging of initial seeds is performed to determine the spliced alignment. As with unspliced aligners, seed and extend methods are slower but more sensitive, and perform better when mapping reads from polymorphic samples Garber et al. (2011).

When spliced aligners are used to map reads to a reference genome, transcripts must be assembled from clusters of reads mapping to different loci. This is accomplished by building a graph to represent all possible isoforms of an expressed feature. Different paths through the graph represent individual isoforms. Two most commonly-used software packages for transcript assembly are Cufflinks Trapnell et al. (2010) and Scripture Guttman et al. (2010). Cufflinks reports the minimum set of transcripts compatible with the set of splice junctions in the reads. Scripture reports all possible transcripts having statistically significant coverage in the read set Martin and Wang (2011).

De novo assembly approaches facilitate transcript reconstruction without the use of a reference genome or transcriptome. De novo assembly of short-reads is a computationally challenging task, and many assemblers have been developed to tackle these challenges, including Velvet Zerbino and Birney (2008), ABYSS Simpson et al. (2009) and ALLPATHS Butler et al. (2008). Typically, de novo assemblers construct a De Bruijn graph using the overlapping reads, and this graph is utilized to re-form transcripts. Trinity Grabherr et al. (2011) and Oases Schulz et al. (2012) traverse the De Bruijn graph to assemble each isoform, while Trans-ABySS Robertson et al. (2010)

and similar tools Martin et al. (2010); Surget-Groba and Montoya-Burgos (2010) build and traverse De Bruijn graphs several times, and merge resulting transcript sets Martin and Wang (2011). Multiple-graph approaches often provide assembled transcripts for a more broad range of expression levels.

Reference-based alignment and de novo assembly each have advantages and disadvantages. The optimal strategy is dependent on the experimental design and available computational resources. Reference-based approaches are far less computationally intensive than de novo approaches. Leveraging a reference genome or transcriptome can also reduce contamination concerns, as it is unlikely to align to the reference in question. Reference-based approaches are well-suited for detection of low abundance transcripts, and the detection of novel transcripts missing from reference annotation Martin and Wang (2011). However, reference-based mapping approaches do rely on the availability and accuracy of reference genomes and transcriptomes. Mapping approaches must also handle single reads that map to more than one location. Mapping uncertainty can arise from paralogous gene families, repetitive sequence, and shared exons of alternatively spliced transcripts Li et al. (2010). Effectively dealing with mapping uncertainty is necessary for accurately measuring gene expression levels Li et al. (2010). Exclusion of ambiguous reads will result in gaps in assembled transcripts, while random assignment can result in false positive transcription detection Martin and Wang (2011). Recent methods have more effectively addressed mapping uncertainty with probabilistic methods Mortazavi et al. (2008); Trapnell et al. (2010); Li et al. (2010). In the case of large and complex transcriptomes, such as plants and mammals, reference-based mapping methods can overcome the computational and isoform-resolving challenges faced by de novo assemblers Martin and Wang (2011). In general, reference-based mapping approaches are the appropriate choice when a reliable reference genome exists.

De novo assemblers have the advantage of not requiring a reference genome. This

allows discovery of transcripts not present in a reference genome or transcriptome, as well as transcripts from a potential external source. De novo assemblers overcome challenges caused by mapping uncertainty, as well as long introns that may be missed by reference-based mapping methods. As previously mentioned, de novo assemblers require large amounts of computational resources, and assembly time. De novo assemblers also require higher sequencing depth than reference-based approaches for adequate transcript assembly. De novo assembly can be very useful for organisms with no reference genome, but is very computationally intensive and is most commonly used in cases of small genomes such as bacteria, archaea, and lower eukaryotes Martin and Wang (2011).

It is possible to combine reference-based mapping and de novo assembly methods. Combined approaches can take advantage of the sensitivity of reference-based mapping, and the flexibility of de novo assembly. This can be accomplished by either aligning the reads to a reference, and then assembling reads that failed to align, or by assembling the reads de novo, then aligning the assembled transcripts to a reference. The "align-then-assemble" approach would allow detection of more transcripts with less computational cost than a de novo assembly alone Martin and Wang (2011).

1.3.4 Challenges of Quantification and Normalization

Once reads have been assigned to transcripts, abundance estimates are calculated for each detected gene or transcript. Prior to normalization, RNA-Seq expression quantities are represented as read counts. The main challenges include low abundance transcripts, shared exons of isoforms for a given gene, and paralogous genes with similar sequences that result in multiple mapping loci for a single read. In the case of transcript-level quantification from reference-based approaches, ambiguous read mappings must be handled as described above. In some cases, ambiguous read assignment, expression quantification, and normalization are combined into one step

with model-based expression estimation Trapnell et al. (2010). If the objective is to quantify gene-level expression rather than transcript-level expression, there are two main approaches to combining read counts from all isoforms of a given gene - the "exon union" method, and the "exon intersection" method Garber et al. (2011) The exon union method counts reads from all exons in all isoforms of a particular gene, while the exon intersection method counts reads only from exons shared by all isoforms.

In order to make inferences with RNA-Seq gene expression data systematic variability must be corrected as with microarray data. In RNA-Seq the two main sources of systematic variability include transcript coverage bias from library construction, and the total number of reads produced in each run. There are many existing normalization techniques for RNA-Seq Data, and much debate as to which techniques are best. Choice of normalization method has been shown to have an impact on subsequent differential expression analyses Bullard et al. (2010). A common approach is to utilize a global per-lane scaling factor to correct for differences between samples in total read counts. Examples include total count normalization, median count normalization, upper-quartile count normalization Bullard et al. (2010), and housekeeping gene count normalization. It has been shown that total count normalization is largely affected by a small number of highly expressed genes Bullard et al. (2010). Housekeeping gene count normalization has the disadvantage of requiring a priori knowledge of a gene that is not differentially expressed, or investigation to find an appropriate candidate Bullard et al. (2010). The length of the transcripts being sequenced has also been shown to impact downstream differential expression analyses, particularly in lowly expressed genes Oshlack and Wakefield (2009); Bullard et al. (2010); Zheng et al. (2011). Specifically, longer transcripts are over-represented in differentially expressed gene sets. Methods such as reads per kilobase of exon model per million reads mapped (RPKM) Mortazavi et al. (2008) or fragments per kilobase

of exon model per million reads mapped (FPKM) Trapnell et al. (2010) have been utilized to normalize read counts by gene length. However, this normalization increases the variance of shorter transcripts, and reduces the statistical power for differential expression analyses for these transcripts Zheng et al. (2011); Oshlack and Wakefield (2009). Recent studies have shown that raw read counts and RPKM/FPKM normalization methods are ineffective Bullard et al. (2010); Zheng et al. (2011); Dillies et al. (2012). A recent study utilized a combination approach, scaling each read count by the total reads, and performing quantile normalization - as is often performed in microarray experiments. Expression estimates were then corrected for unknown biases with principal components a second quantile normalization was performed Pickrell et al. (2010). It has been shown that quantile-based normalization procedures result in improved differential expression analyses Bullard et al. (2010).

1.3.5 Challenges of differential expression analyses

Once transcript expression has been quantified and normalized, many RNA-Seq experiments seek to find differential expression between physiological conditions. In theory, the statistical approaches developed to analyze microarray gene expression data are applicable to RNA-Seq gene expression data (provided appropriate normalization techniques are used) Garber et al. (2011). As an alternative to parametric analysis based on assumptions of normality, additional methods have been developed to accommodate the count-based nature of RNA-Seq data, employing Poisson Marioni et al. (2008); Wang et al. (2010) or Negative Binomial Anders and Huber (2010); Robinson et al. (2010) distributions. The validity of the Poisson distribution has been called into question in the presence of biological replicates Bullard et al. (2010); Anders and Huber (2010). Because it only has a single parameter - an equal mean and variance - it predicts smaller than observed variations, resulting in type I error Anders and Huber (2010).

A recent comparison of differential expression tests compared a likelihood ratio test, t-test, and fisher's exact test Bullard et al. (2010). The authors point out that likelihood ratio statistics have the advantage of being general enough to compare many biological sample types, as well as adjust for potential covariates,t-statistics are only applicable for testing differences between two groups, and Fisher's exact test makes no assumptions about sample size, but only adjusts for global experiment effects. This study also found that technical variations resulting from flow cell or library preparation differences do not have a substantial impact on differential expression analysis results. In some cases, genes will have zero counts in one or more samples. This is a potentially interesting biological observation, but t-statistics cannot adequately test in this case. Fisher's exact test and likelihood ratio tests are more appropriate for such situations. T-statistics also have low sensitivity if either sample has low read counts. If both samples have reasonable gene counts, any of these tests perform well Bullard et al. (2010). As previously mentioned, transcript length has been shown to impact differential expression calls Oshlack and Wakefield (2009); Bullard et al. (2010); Zheng et al. (2011). The over-representation of longer transcripts in resulting DE gene lists may be approached by weighting test statistics by transcript length, however an in-depth analysis of specific methodology has yet to be done Bullard et al. (2010).

1.4 Proteomics

An organism's proteome consists of the set of gene products present as proteins, in constantly varying levels and compositions in different tissues. Untargeted proteomics experiments seek to determine which proteins are present in biological samples, at what concentrations, and how these concentrations differ in various physiological conditions. Analysis of protein expression plays an essential role in biomarker discovery Service (2008), and the elucidation of the Human Proteome Pearson (2008).

The comparative nature of biomarker studies is highly dependent on an accurate assessment of differential protein levels in biological samples, requiring peptide peaks that are accurately matched across LC-MS/MS runs Jeffries (2005). The complexity of biological samples poses significant computational challenges in data processing and analysis steps. Most samples contain tens of thousands of peptides, only a fraction of which are identified Prince and Marcotte (2006). Here we briefly describe label-free LC-MS/MS methods, data processing steps, and their associated challenges.

1.4.1 Open-Platform Proteomics Experiments

In a typical label-free proteomics experiment, proteins are digested by a proteolytic enzyme into peptides. The peptide mixture is physically separated by Liquid Chromatography, providing a column retention time for each analyte. Eluting peptides are converted to gas phase ions, which are then separated in the Mass Spectrometer by mass-to-charge ratio. The relative abundance of each separated ion is measured by a detector. LC-MS experiments utilize a single mass analyzer. In LC-MS/MS, select peptide ions (precursor ions) are further fragmented into product ions and sent through a second mass analyzer. The product ions are analyzed to determine a peptide sequence, which is used to query a database and identify the parent protein. Recently, a variation of LC-MS/MS has been introduced, provides product ion measurements for nearly every precursor ion. Data Independent Acquisition (DIA), also known as MS^E SJ et al. (2009) utilizes a high-low switching fragmentation method to accomplish this. More recently, a new methodology was introduced - $HDMS^E$ Waters (2011) incorporating Ion Mobility (IM) spectrometry. IM spectrometry is added for separation of peptide ions after LC, and before MS^E, separating ionized peptides based on charge and three-dimensional cross-sectional area. Figure 1.1 illustrates the experimental workflow of a typical $HDMS^E$ experiment.

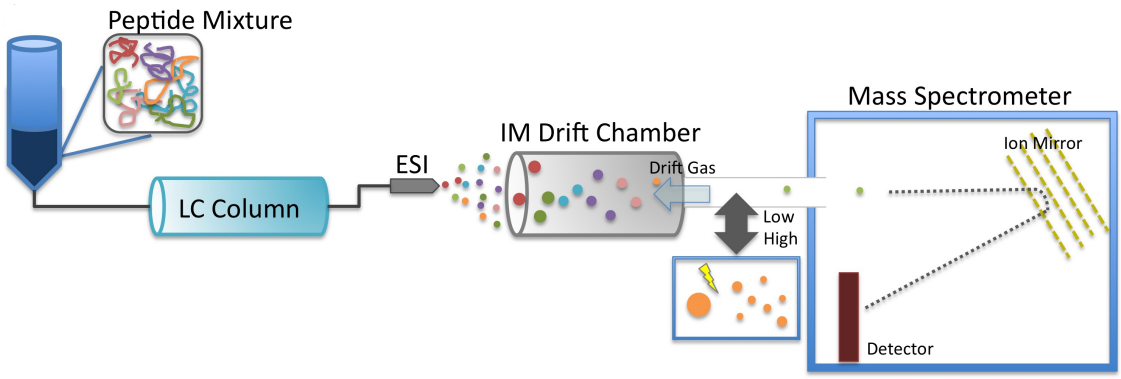

FIGURE 1.1: Modern Open-Platform Proteomics Experiment. Proteins are digested by a proteolytic enzyme into peptides, peptides are separated by hydrophobicity in liquid chromatography, converted to gas phase ions, separated by cross-sectional area and charge by ion mobility, potentially fragmented in the collision cell, and separated in the Mass Spectrometer by mass-to-charge ratio. The relative abundance of each separated ion is measured by a detector.

1.4.2 Open-Platform Proteomics Data Processing

The generation of a peptide-by-sample intensity array requires several data processing steps. Beginning with raw data, true peptide peaks must be discerned from noise, precursor ion charge states must be determined, multiple isotopes of a single peptide must be identified and often combined, and intensity measurements of peptides must be matched across samples. A recent test study by Bell et al. illustrated the challenging nature of computational processing steps. A sample of 20 proteins was distributed to 27 different labs, which used a variety of data-processing methods. There were significant discrepancies in reported proteins, however, all raw data was sufficient to identify all 20 proteins when centrally re-processed Bell et al. (2009). The lack of standard data processing procedures may be the source of much irreproducibilty.

Peak Detection, Charge State Determination, and De-Isotoping

Many data processing pipelines begin with peak detection, charge state determination, and de-isotoping, though the order of data processing can vary between packages. Raw data contains peptide and noise peaks, with each peptide presenting as several peaks due to multiply charged ions and the presence of different isotopes (i.e. the presence of one or more ^{13}C). A common peak detection and de-isotoping technique is to repeatedly determine the most intense peak in the dataset, and determine the charge state and isotopic distribution from the frequency and intensity of the neighboring peaks. Many methods utilize MS intensity and isotope patterns Monroe et al. (2007); Mueller et al. (2007); Sturm et al. (2008); Li et al. (2005). Other peak detection methods utilize LC peak shape Katajamaa et al. (2006); Hastings et al. (2002); Andreev et al. (2003). Some methods utilize both the LC and the MS dimensions of the raw data Leptos et al. (2006); Bellew et al. (2006); Du et al. (2007). Current peak detection and de-isotoping methods are described in Dowsey et al. (2010) and Zhang et al. (2009).

De-Warping

After peak detection, charge state determination, and de-isotoping, multiple LC-MS/MS runs are aligned. As with any laboratory experiment, LC-MS/MS data are subject to variability. The LC retention times often shift between runs. Pressure fluctuations, changes in column temperature, column manufacturing differences, and peptide interactions can cause changes in the elution time, and/or the elution order of peptides Vandenbogaert et al. (2008). These experimental variations are typically called warp. De-warping may performed on raw profile data (prior to or independent of peak detection and de-isotoping), or on feature data (detected peptide peaks). Many de-warping methods exist, performing linear or non-linear (or both) corrections of two or more samples Listgarten and Emili (2005).

Alignment

After de-warping, peptide features are matched, or aligned, across samples to generate a peptide-by-sample intensity array. Alignment methods utilize various aspects of the data (i.e. charge, retention time, mass-to-charge ratio) to make match assignments. Matching is complicated by variations in retention time and/or mass-to-charge ratio, errors in peak detection, and overlapping peptides. Many alignment methods exist, utilizing different aspects of the data most commonly mass-to-charge ratio and retention time Katajamaa et al. (2006); Pluskal et al. (2010); Mueller et al. (2007); Bellew et al. (2006); Li et al. (2005); Silva et al. (2005). It is also possible to utilize intensity measurements or peptide identifications Jaffe et al. (2006); Fischer et al. (2006); Tang et al. (2011); Mueller et al. (2007). Not surprisingly, incorporating additional data provides a higher degree of specificity Listgarten and Emili (2005).

Normalization

As with gene expression data, protein intensities must be normalized prior to comparative analyses, correcting for systematic variations such as injection volumes Riley et al. (2010). This can be achieved by choosing a single run to serve as a reference, and normalizing all other runs to that reference, simply computing the overall normalization constant as the median intensity ratio between matches peptides of the reference and sample in question Wang et al. (2003). Another approach is to divide the intensity at each m/z value by the mean intensity of the whole m/z range Zhu et al. (2003). Quantile-based normalization techniques have also been adopted from microarray analyses, and other sophisticated methods have been developed with proteomics data in mind Karpievitch et al. (2012). A noteable characteristic of such proteomics-specific normalization methods is the ability to handle the prevalence of missing values in these data sets Karpievitch et al. (2009); Wang et al. (2006).

2

GWAS for Measures of Adiposity

2.1 Background

Obesity presents a major public health challenge in the developed world and is a primary focus of preventative healthcare. Rates of both overall adiposity, measured by body mass index (BMI), as well as central (intra- abdominal) adiposity, measured by waist circumference (WC) or waist to hip ratio (WHR) have been steadily rising during the past several decades, accompanied by increased rates of diabetes mellitus, cardiovascular disease, and other morbidities Wang and Beydoun (2007). In the United States, regional, racial, and sex differences in adiposity have been noted, but the patterns are complex and changing over time Wang and Beydoun (2007). According to U.S. national health survey data, men on average have had a higher BMI than women, but since the mid 1990s the average BMI in women has been higher than men Zhang and Wang (2004). Men also tend to have larger abdominal girth than women, and this disparity has persisted over time Okosun et al. (2004, 2003). Obesity is a heritable trait and recent genome-wide association studies have identified dozens of loci influencing measures of adiposity Speliotes et al. (2010); Willer

et al. (2009); Thorleifsson et al. (2009); Heid et al. (2010). Sex differences in the heritability of obesity-related traits have been noted as well in several studies Zillikens et al. (2008). In addition, linkage analysis in both rodent models and humans have found evidence of sex-specific loci affecting obesity-related traits Sammalisto et al. (2009); Atwood et al. (2006). Framingham Heart Study investigators found widespread evidence for sex-specific effects of genetic loci on BMI, identifying several chromosomal regions with suggestive linkage to BMI in one sex, but not the other Atwood et al. (2006). Indeed some effects were only seen in sex-stratified analyses and were not at all evident in the combined cohort of men and women. More recently, two genome-wide association study meta-analyses of WHR examined their top loci for sex differences and identified sex-specific effects for several loci Heid et al. (2010); Lindgren et al. (2009). We sought evidence for significant differences in SNP effects on adiposity traits in men and women across the genome by carrying out a genome-wide association study modeling gene by sex interaction for WHR, WC, and BMI in the population-based Framingham Heart Study. Genome- wide association analysis of SNPs having main effects (as opposed to gene-by-sex interaction) on obesity were reported earlier in the Framingham Heart Study using 100 K SNPs, but gene by sex interactions were not considered at that time Fox et al. (2007). Subsequently, the full genotype data (>500K SNPs) have been pooled with other studies and reported in large meta-analyses, which found evidence of gene by sex interaction for WHR but not BMI among the SNPs with main effects Speliotes et al. (2010); Heid et al. (2010).

2.2 Methods

2.2.1 Study Population

We conducted this research using data from the Framingham Heart Study, a population-based, longitudinal study of families living in the town of Framingham, Massachusetts

collected over three-generations beginning in 1948. An overview of the study is provided at the dbGap website (http://www.ncbi.nlm.nih.gov/sites/entrez?db=gap) and detailed descriptions are available elsewhere Govindaraju et al. (2008); Cupples et al. (2007). Briefly, the original study (Generation 1) enrolled 5209 individuals, primarily Caucasian, and it later added the offspring of the original cohort (Generation 2), and the grandchildren (Generation 3) of the original cohort. Primary analyses were carried out using data from the five first exams of subjects in Generation 2, collected between 1971 and 1994. Obesity-related traits evaluated in this study included BMI, measured at exams 1, 2, 3, 4, and 5, WHR, measured at exams 4 and 5, and WC measured at exams 4 and 5. We limited our analyses to these exams due to a drop in sample size at subsequent exams. Replication of genome wide association study (GWAS) results was sought in subjects from Generation 3 (data collected from 2002 to 2005). Individuals with diabetes (n = 92, 94, 59, 27, 116, and 136 for generation 2 exams 1, 2, 3, 4, 5 and generation 3 exam 1, respectively) or thyroid disorder (n = 117, 94, 9, 36, 265, and 72 for generation 2 exams 1, 2, 3, 4, 5 and generation 3 exam 1, respectively) were removed because these diseases have an effect on both BMI and fat distribution. The data were further trimmed, excluding individuals with outlier trait values determined by taking the mean of the phenotype (independently for each exam and each sex) and adding/subtracting three standard deviations. Removal of outlier values in the BMI GWAS data was performed with weight, height, and BMI. WC and hip circumference (HC) outliers were also eliminated in the waist phenotype GWAS. Finally, we restricted our analysis to premenopausal women and individuals under the age of 50 to enhance differences related to estrogen-mediated gene-by-sex interaction and to reduce as much as possible the age-related differences in association that may occur across exams. The total sample sizes for the BMI GWAS after genotype quality control and trait outlier removal were 3150, 1991, 1630, 1330, 990, and 2872 for generation 2 exams 1, 2, 3, 4, 5, and generation 3 exam 1, respec-

tively. The sample sizes for the waist phenotype GWAS were 1330, 984, and 2872 for generation 2 exams 4, 5 and generation 3 exam 1, respectively.

2.2.2 Genotype Data and Quality Control

Genome-wide genotypes and detailed clinical data have been made accessible to the research community through the SHARe project (SNP-Health Association Resource). The study protocol was approved by Duke Universitys Institutional Review Board and the Framingham SHARe Data Access Committee. The unfiltered genotype data contained 9215 individuals (all generations) genotyped for 549782 SNPs. This included 500568 SNPs from the Affymetrix 500K mapping array and 49214 SNPs from the Affymetrix 50K supplemental array (Affymetrix, Santa Clara, CA, USA). We used the toolset PLINK Purcell et al. (2007) to perform quality control. Individuals were excluded if genotyping rates were less than 97%. Markers were excluded if genotyping rates were less than 97%, minor allele frequencies were less than 0.05, or if Hardy-Weinberg P-values were less than .001. All SNP exclusions were made sequentially in the preceding order. Using this filtered data, we checked for Mendel errors using a 5% cutoff per family, and a 10% cutoff per SNP (as defined in PLINK), but none were detected. Individuals were also excluded if the predicted sex-based on X-chromosome genotypes did not match the recorded sex. Pairwise identity-by-descent measures were calculated to detect replicated samples and unknown interfamilial relationships. We detected 4 identical twins and randomly selected one member of each pair for the analytic sample. After quality controls, the remaining sample consisted of genotype data on 360811 SNPs, attaining a genotyping rate of 99.5%.

2.2.3 Statistical Analyses

Analysis of WHR and WC were based on data obtained at exam 4 (n = 1330) and exam 5 (n = 984) of subjects from Generation 2. The gene-by-sex GWAS was run

on data from each exam separately. We ran the full model for both WHR and WC regressed on BMI, age, age-squared, genotype, sex, and the genotype-by-sex cross product. BMI was available at all exams, with adequate sample sizes on the first five exams. Five separate GWAS were run using the full model of BMI regressed on age, age-squared, genotype, sex, and the genotype-by-sex cross product - one each for exams 1, 2, 3, 4, and 5 of Generation 2. SNPs were evaluated for associations in an additive genetic model. A main effect GWAS was also run for BMI across the five exams, using the model specifications above without the cross product term. Sex-specific associations were tested using the full model of BMI regressed on age, age-squared, and genotype on each sex. To account for relatedness, we used generalized estimating equations while accounting for sibling correlation in the Yags package Vince (2004) of the R statistical language. The P-values of the covariates were obtained via the Wald test using robust standard errors. The Framingham population has been studied extensively, and evidence for considerable population stratification has not been detected. To test this assumption, we estimated the inflation factor by dividing the median of the observed χ^2 statistics for each GWAS, by the expected median in the absence of stratification (0.456) Devlin and Roeder (1999); Bacanu et al. (2002). Also, adjusted for population stratification with the scores of the first 10 principal components, computed with Eigenstrat Price et al. (2006b). We defined genome-wide significance using a Bonferroni cutoff of $1.4x10^7$, which corrects for 360811 tests. Following genome-wide analysis, we annotated results using the WGAViewer package Ge et al. (2008), Ensembl Hubbard et al. (2009), and the UCSC genome browser. We generated plots using the Gap package Zhao (2007) of the R statistical language and Haploview software Barrett et al. (2005). To enrich for true positive associations, we took a strategy whereby associations that appeared in all exams were considered to have a higher likelihood of being true associations. We expected earlier exams to have greater power due to larger sample sizes, but other

factors, including decreased heritability with age Atwood et al. (2006) may affect the results as well. This strategy required us to make some decisions about what cutoff to use when comparing results across exams. We took the consensus across exams of the top most significant 10, 100, 1000, and 10000 hits and found 0, 0, 4, and 105 SNPs, respectively, and focused our analysis on the four SNPs from the top 1000 consensus.

2.3 Results

Characteristics of the subjects from Generations 2 and 3 of the Framingham Study used in the current analyses are presented in Table 2.1, broken down by exam. For each exam, we restricted our analyses to men and women <50 years of age, resulting in a decrease in sample size over time, above and beyond the loss due to death or non participation.

Table 2.1: Mean ± standard deviation for obesity-related traits in Framingham subjects <50 years old. * - Significant Difference between Generation 2 and Generation 3 after controlling for age and age-squared (* - p <0.001, ** - p <1e-5, *** - p <1e-10, **** - p <1e-20, ***** - p <1e-50).

	Population	Generation 2, Exam 1	Generation 2, Exam 2	Generation 2, Exam 3	Generation 2, Exam 4	Generation 2, Exam 5	Generation 3, Exam 1
Sample Size (N)	All	3150	1991	1630	1330	990	2872
	Men	1478	958	776	640	463	1388
	Women	1672	1033	854	690	527	1484
Age (years)	All	33.71 ± 8.59	38.39 ± 6.70	40.46 ± 5.77	42.27 ± 5.11	43.68 ± 4.44	37.63 ± 7.27
	Men	33.81 ± 8.63	38.38 ± 6.88	40.47 ± 5.92	42.14 ± 5.23	43.56 ± 4.65	37.79 ± 7.30
	Women	33.62 ± 8.56	38.39 ± 6.54	40.45 ± 5.63	42.40 ± 5.00	43.78 ± 4.24	37.47 ± 7.24
BMI (kg/m^2)	All	24.85 ± 3.77***	25.10 ± 3.79****	25.40 ± 4.08***	25.97 ± 4.37*	26.42 ± 4.48	26.23 ± 4.70
	Men	26.42 ± 3.47*	26.57 ± 3.44**	26.72 ± 3.49**	27.23 ± 3.62	27.55 ± 3.92	27.42 ± 4.03
	Women	23.48 ± 3.47***	23.74 ± 3.59***	24.20 ± 4.20**	24.80 ± 4.66*	25.42 ± 4.70	25.11 ± 4.99
Height (cm)	All	167.62 ± 9.36****	168.59 ± 9.57***	169.66 ± 9.20*	169.67 ± 9.12*	169.56 ± 9.06	170.91 ± 9.18
	Men	174.95 ± 6.79****	175.99 ± 6.77***	176.75 ± 6.66*	176.65 ± 6.52*	176.69 ± 6.41	177.88 ± 6.41
	Women	161.15 ± 5.90*****	161.73 ± 6.02****	163.23 ± 5.86*	163.20 ± 5.81	163.29 ± 5.85	164.39 ± 6.06
Weight (kg)	All	69.77 ± 14.52****	71.25 ± 14.62****	72.89 ± 14.97***	74.50 ± 15.55**	75.72 ± 15.79	76.32 ± 16.55

	Population	Generation 2, Exam 1	Generation 2, Exam 2	Generation 2, Exam 3	Generation 2, Exam 4	Generation 2, Exam 5	Generation 3, Exam 1
	Men	80.26 ± 11.76****	81.63 ± 11.34***	82.78 ± 11.82***	84.27 ± 12.24**	85.34 ± 13.02	86.06 ± 13.76
	Women	60.50 ± 9.58****	61.63 ± 10.01****	63.90 ± 11.45***	65.45 ± 12.51*	67.27 ± 12.91	67.20 ± 13.47
WC (cm)	All	-	-	-	86.70 ± 13.92****	88.95 ± 13.30***	91.03 ± 13.33
	Men	-	-	-	95.85 ± 9.93**	96.49 ± 10.34	96.50 ± 11.15
	Women	-	-	-	78.21 ± 11.50*****	82.36 ± 12.04***	85.90 ± 13.18
HC (cm)	All	-	-	-	100.07 ± 9.08	101.37 ± 8.63	-
	Men	-	-	-	101.31 ± 7.18	101.83 ± 6.85	-
	Women	-	-	-	98.92 ± 10.42	100.98 ± 9.93	-
WHR	All	-	-	-	0.86 ± 0.10	0.88 ± 0.09	-
	Men	-	-	-	0.95 ± 0.06	0.95 ± 0.06	-
	Women	-	-	-	0.79 ± 0.06	0.81 ± 0.07	-

2.3.1 Genome-Wide Association Analysis of Gene-by-Sex Interaction for WHR and WC and BMI

None of the gene-by-sex interaction GWAS revealed genome-wide significant loci. For BMI we noted marked heterogeneity in quantile-quantile (QQ) plots between exams (Figure 2.1), which does not appear to be a function of sample size (which decreases with exam). There is also some evidence of inflation in the QQ plots, which was not alleviated after controlling for population stratification. In sex-stratified analysis, the inflation appeared to be restricted to men. The top 1000 hits from each exam for each trait (ordered by the P-value of the gene by sex interaction term) were extracted (Supplementary Tables available online on doi:10.1155/2011/329038), and the intersection of those datasets was sought for each trait.

For WHR, we identified 43 SNPs (28 unique loci) and for WC, we identified 43 SNPs (27 unique loci) appearing among the top 1000 in both exams 4 and 5. When examining loci across these two traits, SNPs near SPOCK3, OSTF1, RAB31, and RPF1 appear in the top 1000 consensus for WC and WHR. SPOCK3 stands out as appearing among the top 100 hits across both exams 4 and 5 for WC (P = $5.33x10^7$ and P = $2.45x10^5$) and WHR (P = 1.8510^4 and P = 7.9510^5). For BMI, only four SNPs appeared among the top 1000 hits in all five exams. All four SNPs localized to the same linkage disequilibrium block on chromosome 1 - 100kb downstream of LYPLAL1. We were most intrigued by these findings as the LYPLAL1 locus has been reported as a sex-specific locus affecting central adiposity in two prior genome-wide association meta-analyses Heid et al. (2010); Lindgren et al. (2009). The extent of linkage disequilibrium (LD) surrounding the associated SNPs in the region of LYPLAL1 was determined in the Hap Map phase 3 CEU population by identifying the farthest SNP away in each direction that had r^2 >0.5 for each of the four SNPs. The LD block extends over 330 kb from position 217,321,833 to 217,655,426, and encompasses the LYPLAL1 gene (Figure 2.2). The block does not

include the SNPs from Lindgren et al. (2009) or Heid et al. (2010), which are in moderate linkage disequilibrium with each other and located an additional 55kb and 258kb downstream of LYPLAL1, respectively.

2.3.2 Replication of LYPLAL1 SNP Association with BMI in Framingham Generation 3 Subjects

We next sought to replicate the observed association in subjects from Generation 3 of the Framingham Study. Again, we restricted our analyses to those less than 50 years of age. A comparison of results by sex in the five exams of Generation 2 and in Generation 3 are shown in Figure 2.3 for the top associated LYPLAL1 SNP. The SNP-by-sex interaction for LYPLAL1 was significant in all Generation 2 exams, but not significant in Generation 3 subjects. However, when stratified by sex, the minor allele showed a consistent increase in BMI in men across generations (Figure 2.3). In contrast, in women the minor allele was associated with lower BMI in Generation 2 but not in Generation 3.

2.3.3 Association of LYPLAL1 SNPs with Obesity-Related Traits

To understand the relationship between LYPLAL1 SNPs and obesity in greater detail, we examined the top SNP from the present study (rs7552206) along with SNPs from the Lindgren et al. (2009) and Heid et al. (2010) studies for association with related phenotypes, including height, weight, WC, and WHR. The rs7552206-by-sex interaction for BMI tracked with weight in all five exams, and with WC and HC in the two exams that had these data available. However, the waist and hip associations were completely or nearly completely attenuated when controlling for BMI. For rs2605100 Lindgren et al. (2009), no compelling evidence of gene by sex interaction in central adiposity was found. Heid et al. (2010) independently found a female-biased WHR association with LYPLAL1 (rs4846567), an SNP in moderate linkage disequilibrium with the Lindgren et al. SNP. We analyzed an available proxy

for this SNP (rs2820446, HapMap CEU $r^2 = 1$) and found a borderline significant gene-by-sex interaction with WHR (P = 0.09).

2.3.4 Genome-Wide Association Analysis of Gene Main Effects for BMI

We also explored our cross-exam consensus approach for detecting significant main effects for BMI, using the same age-restricted datasets as the gene by sex interaction analyses. As with our gene by sex interaction analyses, the QQ plots show marked heterogeneity between exams (Figure 2.1) and modest inflation, which was not accounted for by population stratification. Only one SNP, located approximately 26 kb upstream of DUSP10 on chromosome 1, appeared among the top 1000 hits in all five exams of Generation 2 and was borderline significant in Generation 3 subjects (Figure 2.4). Interestingly, this locus is approximately 2.4 Mb away from the gene by sex interaction LYPLAL1 SNPs. No SNPs from prior genome-wide association studies of BMI showed up among our top 1000 consensus, including SNPs in the genes INSIG2, FTO Fox et al. (2007); Herbert et al. (2006), and MC4R Renstrom et al. (2009) (Figure 2.4). Surprisingly, the SNPs identified with the consensus approach yielded more significant P values than other loci.

2.4 Discussion

We carried out a genome-wide assessment of gene-by-sex interaction for standard measures of obesity in men and women less than 50 years of age in the Framingham Heart Study. We took advantage of longitudinal data from multiple exams to identify loci showing consistent evidence of SNP-by-sex interaction across exams. Among the most prominent was a region approximately 100 kb downstream of LYPLAL1, encoding the lysophospholipase-like 1 protein. We found evidence across five exams, spanning a 20-year time frame, of opposite effects of genetic variants in this region on BMI in men and women. An attempt to replicate this finding in a later genera-

tion of Framingham Heart Study subjects found a consistent, significant association in men, but not in women, possibly indicating a male-specific association. Ours is not the first study to link LYPLAL1 to obesity: two other genome-wide association meta-analyses identified this locus as having a sex-specific effect on WHR Heid et al. (2010); Lindgren et al. (2009). While neither SNP is in linkage disequilibrium with the region identified in our study, the coincidental discovery of two distinct regions near the LYPLAL1 locus associated with obesity-related traits in a sex-specific fashion warrants further attention. Moreover, a prior linkage analysis of BMI in Generation 2 of the Framingham Heart Study identified a male-biased linkage for BMI in the vicinity of LYPLAL1 on chromosome 1q41 Atwood et al. (2006). None of the other sex-specific obesity loci from Heid et al. (2010) were found in our study. LYPLAL1 is a member of the lysophospholipase gene family (EC number 3.1.1.5). It was initially identified as a gene on chromosome 1 found incidentally during investigation of a familial chromosomal translocation David et al. (2003). It was named on the basis of approximately 30% predicted amino acid sequence homology with lysophospholipases I and II Wang and Dennis (1999). The sequence suggests an α/β hydrolase fold typically found in many lipases and esterases. LYPLAL1 was subsequently identified as one of 23 esterolytic/lipolytic proteins extracted from mouse adipose tissue. The presence of an active site serine was determined by activity tagging with a fluorescent probe of broad specificity, resembling a single-chain carboxylic acid ester. Similar probes modeling triglyceride and cholesteryl ester did not tag LYPLAL1 Birner-Gruenberger et al. (2005). LYPLAL1 protein has not yet been isolated, however, and its substrate specificity is unknown. Along with the gene for adipocyte triglyceride lipase and several others related to lipolysis, LYPLAL1 mRNA was expressed more abundantly in abdominal subcutaneous adipose tissue from obese versus lean humans Steinberg et al. (2007). Given the minimal characterization of LYPLAL1, we can only speculate about its sex-specific role in

adiposity. It might be involved in triglyceride synthesis or lipolysis, similar to some of the proteins with which it is co-expressed Zechner et al. (2009). If indeed it is a lysophospholipase, it might play a role along with autotaxin, a secreted phospholipase D, in regulating extracellular levels of lysophosphatidic acid in adipose tissue. Via specific G protein-coupled receptors, lysophosphatidic acid has been shown to have varying effects on adipocyte differentiation and growth Pages et al. (2000); van Meeteren and Moolenaar (2007); Simon et al. (2005). Another possibility relates to the endocannabinoid system, which has been a recent pharmacologic target for investigative obesity treatments. The monoglyceride, 2-arachidonoyl glycerol, as well as other esters or amides of long-chain polyunsaturated fatty acids belong to a family of compounds that are natural ligands for cannabinoid receptors. These endogenous signaling molecules affect physiologic and behavioral processes governing appetite and energy metabolism Zechner et al. (2009). Interestingly lipolysis control has been shown to vary by sex in some studies Williams (2004) but not others Bulow et al. (2006). The aforementioned study showing support for sex differences in lipolysis suggests that women show greater sensitivity to lipolysis in abdominal subcutaneous fat. The authors argue that the differences in lipolysis sensitivity are due to the presence of fewer inhibitory alpha-adrenergic receptors in the abdominal subcutaneous adipose tissue. This area of lipid metabolism is not well understood, but recent discoveries and conflicting opinions warrant further studies on LYPLAL1 and its potential roles and sex-specific effects in lipid metabolism and obesity. Our analysis revealed marked heterogeneity of effects across different exams of the study, both in gene-by-sex interaction and main effect analyses, even among established loci from other genome-wide association studies of BMI. The consensus approach appears to be robust, identifying a locus with strong prior evidence of gene by sex interaction for obesity-related traits. Using this approach, we also identified a possible novel candidate locus for BMI, located approximately 26kb upstream of DUSP10,

encoding a dual specificity protein phosphatase. The DUSPs are a subclass of the protein tyrosine phosphatase gene superfamily that controls MAP kinase function Camps et al. (2000). Our study was carried out in the Framingham Heart Study Offspring cohort, a longitudinal, population-based study. Although no loci reached genome-wide significance in gene-by-sex interaction analyses, the longitudinal nature of the data allowed us to prioritize SNPs based on consistency of effect across exams. However, data on waist circumference were available only at two exams, limiting the effectiveness of our approach for these traits. Nonetheless, for BMI, this approach yielded a plausible candidate sex-specific locus and another sex-independent locus. Interestingly, in both of these cases, results from Generation 3 were not as significant as in Generation 2, possibly reflecting a cohort effect: Generation 2 subjects were enrolled nearly a decade or more prior to Generation 3 subjects. Generation 3 subjects were on average more overweight than Generation 2 subjects at comparable ages, consistent with temporal trends of increasing obesity observed in other population-based studies. These differences, driven in large part by changes in diet and physical activity over time, may impact the heritability over time and thus, the ability to detect genetic effects.

2.5 Conclusions

Few studies have systematically modeled gene by sex interaction for obesity-related traits on a genome-wide level. We confirm in our study that SNPs in the vicinity of LYPLAL1 may exhibit sex-specific effects on obesity-related traits. By utilizing a well-designed population-based study, and taking advantage of longitudinal data, we were able to demonstrate this effect using a much smaller sample size than the original meta-analysis that identified this locus. This has implications for the design of GWAS, where large samples sizes are often sought sometimes at the expense of population homogeneity. We suggest that smaller epidemiologically sound

population-based studies may be more powerful than larger heterogeneous metacohorts. We also highlight the importance of considering longitudinal robustness of association within a cohort as another means of prioritizing loci and reducing false positive associations. Future studies of LYPLAL1 are needed to determine the basis of the apparent sex-specific effect on obesity.

2.6 Acknowledgements

We would like to thank the National Heart, Lung and Blood Institute (NHLBI) for creating the open- access Framingham SHARe resource, and all of the study participants and Framingham Heart Study investigators who helped to create this tremendously valuable resource. The Framingham Heart Study is conducted and supported by the NHLBI in collaboration with Boston University. This study was not prepared in collaboration with investigators of the Framingham Heart Study and does not necessarily reflect the opinions or views of the Framingham Heart Study, Boston University, or the NHLBI. This work was supported in part by National Institutes of Health Grants: RO1 HL085191 to J. McCarthy, PI. Jennifer Rowell was supported by NIH training Grant T32-DK007012-31 (Feinglos, PI), and Ashlee Benjamin by training Grant T32 GM071340 (Harer, PI).

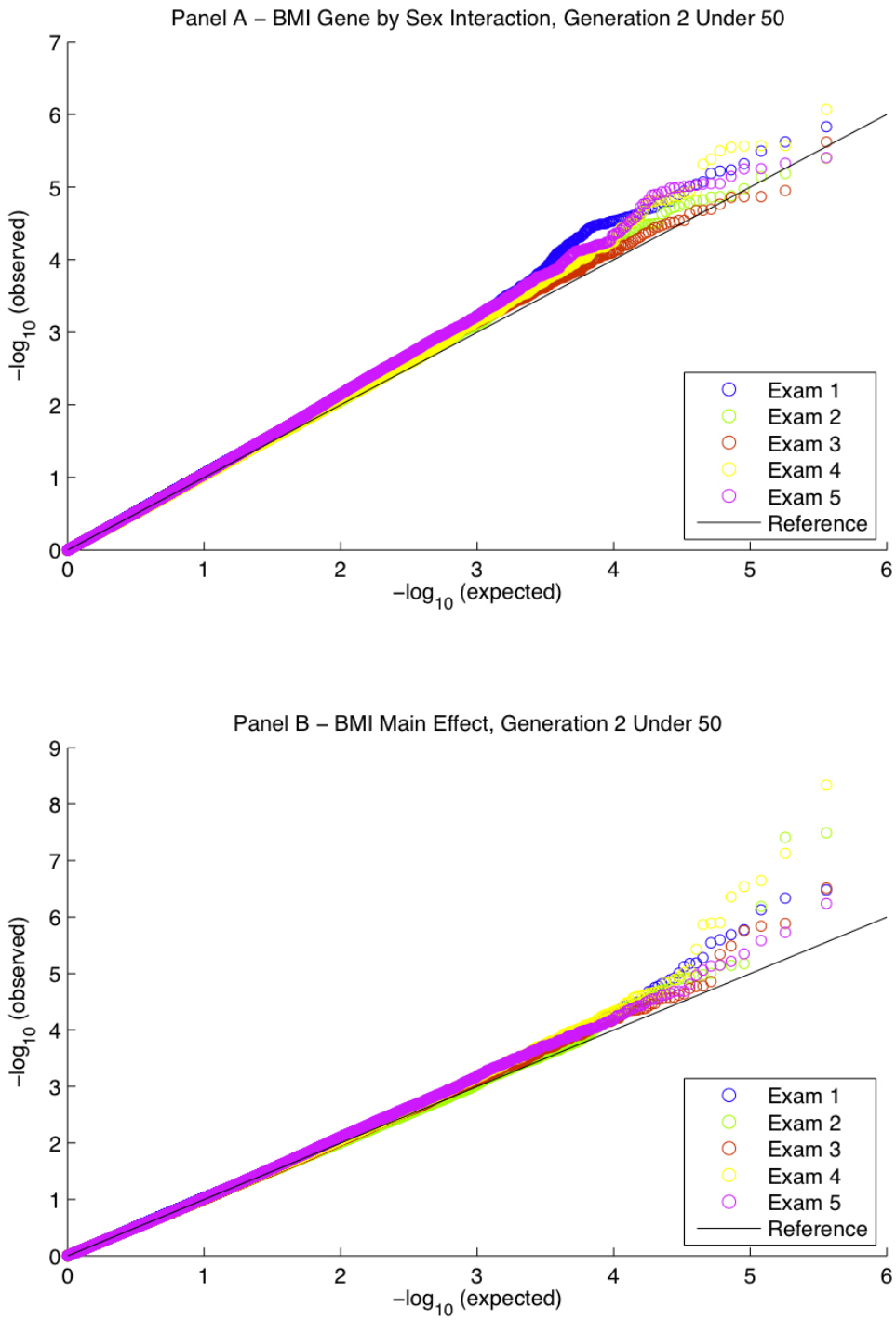

FIGURE 2.1: QQ plots for gene by sex interaction (a) and main effect (b) GWAS for body mass index (BMI) in Generation 2, exams 1, 2, 3, 4, and 5.

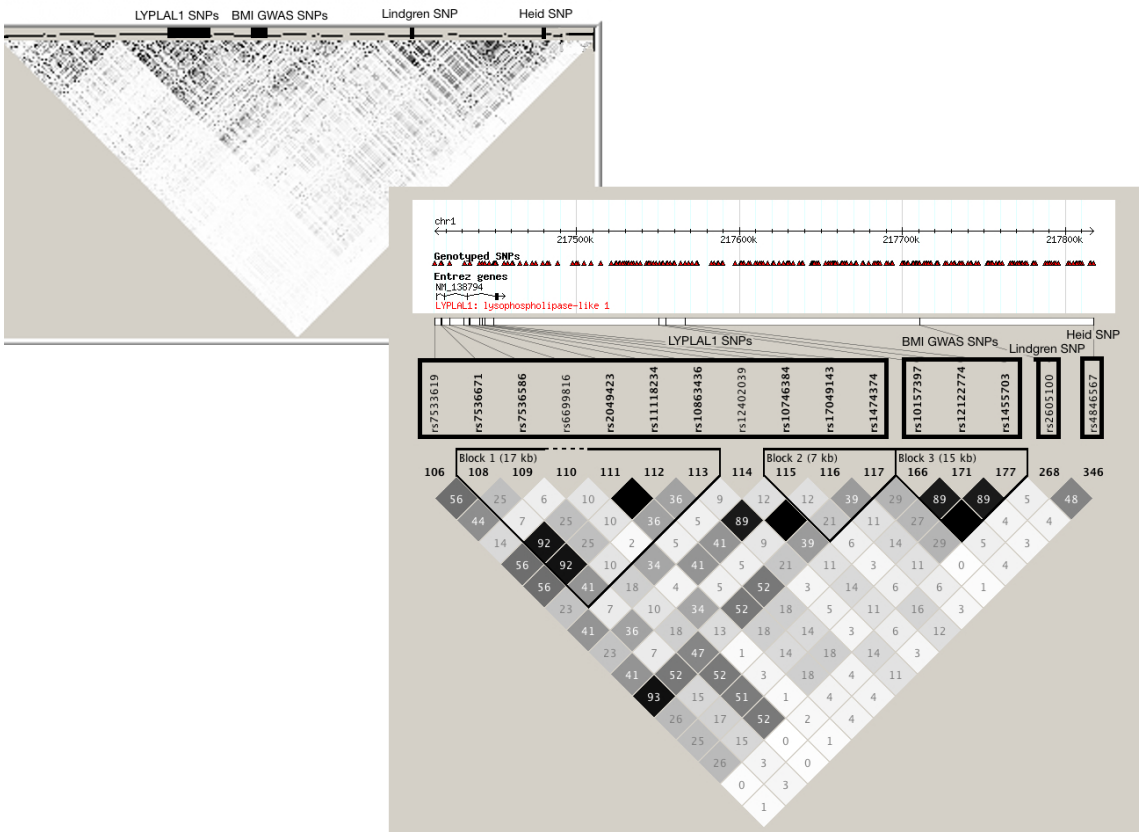

FIGURE 2.2: Linkage disequilibrium (shown as r^2) in the region encompassing LYPLAL1, the consensus SNPs associated with body mass index (BMI) in our gene by sex interaction GWAS, and the sex-specific SNPs associated with waist to hip ratio (WHR) in recent GWAS meta-analyses.

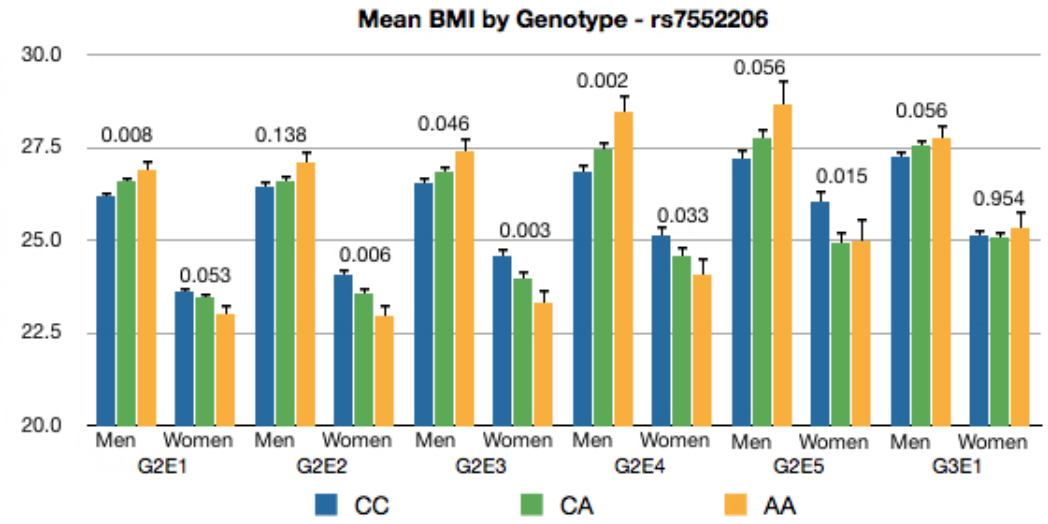

FIGURE 2.3: Mean body mass index (BMI) by genotype and sex across exams for the top associated SNP in LYPLAL1 (rs7552206) with Standard Error Bars and SNP P-values.

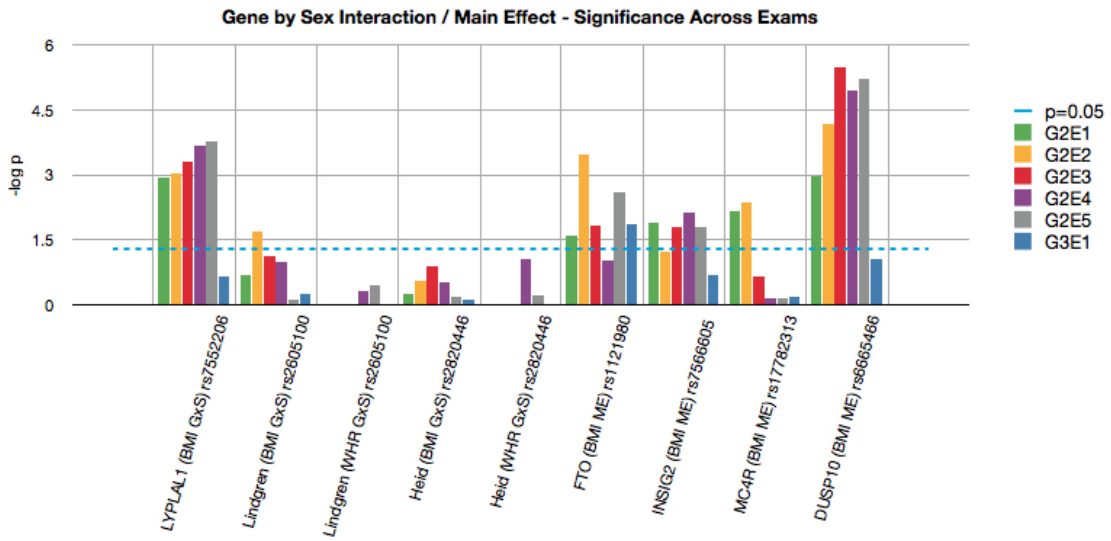

FIGURE 2.4: Significance level of main effect (ME) and/or gene by sex interaction (GxS) associations with body mass index (BMI) and/or waist to hip ratio (WHR) for various loci of interest.

3

Classifying Transcription Start Sites

3.1 Background

The development of high-throughput sequencing strategies, which generate millions of 5' sequence tags from capped RNAs transcribed by RNA polymerase II (pol II), has enabled obtaining fine-grained pictures of transcription initiation. Each of the tags originates from a transcription start site (TSS), and mapping the tags to the genome identifies tag clusters for individual genes. In particular, the application of Cap Analysis of Gene Expression (CAGE) produces comprehensive data sets for mammalian promoters Carninci et al. (2006), and an extension of this methodology to Paired End Analysis of Transcription Start Sites (PEAT) was used to map and cluster millions of paired reads from *Drosophila melanogaster* embryos Ni et al. (2010). Tag clusters exhibit different initiation patterns, i.e. distributions of tags within a cluster, and have been used to define distinct promoter classes, generally falling into two basic groups: Both flies and mammals have focused promoters in which transcription occurs within a narrow genomic window of a few nucleotides, and dispersed promoters in which TSSs spread out over a larger genomic region on the order of a hundred

nucleotides. Promoter classes have distinct associations to core promoter motifs and functional roles Juven-Gershon and Kadonaga (2010); Ohler and Wassarman (2010), and evidence has pointed towards enriched pausing, or stalling, of Drosophila pol II at focused promoters Nechaev et al. (2010).

Many studies have shown a generic pattern of chromatin organization in promoters, in which a nucleosome free region (NFR) upstream of the TSS is surrounded by periodic arrangements of nucleosomes within the transcript and further upstream Mavrich et al. (2008); Schones et al. (2008), illustrating the connection between chromatin features and the accessibility of the DNA to transcription factors (TFs). Nucleosomes containing H2 and H3 histone variants provide particularly strong signals for the beginnings of genes in eukaryotes Mavrich et al. (2008); Jin et al. (2009); Raisner et al. (2005), as they are preferentially incorporated in or near areas of active transcription. Data on frequent modifications to the N-terminal histone tails have furthermore supported a histone code specifying functional domains in the genome; for instance, the tri-methylation of H3K4 has been shown to mark the promoter regions surrounding TSSs Barski et al. (2007). In addition, individual instances of insulator elements have been shown or suggested to play a role in chromatin remodeling near promoter regions Tsukiyama et al. (1994); Fu et al. (2008).

Given that the distinct promoter classes are widely conserved throughout metazoans, and nucleosomes are correlated with the accessibility of the DNA, it may be surprising that virtually no analysis so far has directly examined whether focused or dispersed promoters are associated with different nucleosome organization and chromatin structure. Instead, the majority of reports have taken the approach of dividing genes according to chromatin or insulator patterns, and then associating the promoters in each group with sequence features Mavrich et al. (2008); Ioshikhes et al. (2006) or function Engström et al. (2007); Ganapathi et al. (2005). One of the main limitations of this approach has been that these characteristics are present in only a

fraction of promoters. For instance, the TATA box motif is present in only 10-20% of all eukaryotic promoters, and 35% of focused promoters Ohler (2006). On the other hand, CpG islands are a very frequent sequence feature of mammalian regulatory regions Saxonov et al. (2006); Tillo et al. (2010) and have been repeatedly associated with dispersed promoters. Yet, this property is by far not unique to one initiation pattern: depending on the definition, 70-80% of dispersed promoters coincide with the presence of a CpG island, but 50-60% of focused promoters do so as well (Table 3.1). Furthermore, while chromatin features and initiation patterns are conserved at least in metazoans, CpG islands do not exist in the fruit fly genome Ponger et al. (2001), suggesting that specific sequence features may lead to enrichments but not be the sole or primary indicators of the underlying process.

Studies in different metazoans have identified several promoter classes based on the size of the initiation region and the distribution of initiation events within each region Carninci et al. (2006). In previous work in Drosophila Ni et al. (2010), three specific classes were defined. Narrow Peak (NP) promoters are typical focused promoters with high occurrences of initiation at one location. They typically contain one or more canonical position-specific core promoter motifs such as the TATA box, which have been found in genes with developmental regulation and tissue-specific functions. Conversely, Weak Peak (WP) promoters are dispersed promoters, in which transcription is distributed over a larger genomic span and lacks a clear preference for a single start site. In flies, WP promoters are associated with distinct core promoter sequence elements but largely lack the canonical eukaryotic-wide core promoter motifs, and are frequently associated with housekeeping genes Engström et al. (2007); Rach et al. (2009). CpG islands, long stretches of CpG dinucleotides that play a role in chromatin packing and nucleosome organization Davey et al. (1997, 2004), are a feature of most mammalian promoters and are more frequently present in WP promoters Carninci et al. (2006) . Finally, an intermediate class, Broad with Peak (BP)

promoters, displays both a preference for a narrow location as in NP promoters, yet with tags covering a larger genomic span as in WP promoters.

Table 3.1: Distribution of Promoters Classes

Class	CpG/total (Frommer)	CpG/total (Jones)
NP	827/1409 (58.7%)	689/1409 (48.9%)
BP	1375/1759 (78.2%)	1130/1759 (64.2%)
WP	6510/7656 (85.0%)	5244/7656 (68.5%)

Table 3.1 lists the number of promoters in each class, and indicates the presence of CpG islands within TSS classes. As the table shows, individual sequence features are enriched in certain promoter classes, but any single feature does not cover any of the classes completely. CpG islands were defined using two sets of criteria: the classic definition of Gardiner-Garden and Frommer (1987), and the more stringent definition of Takai and Jones (2002) which aims at a better separation from Alu-repetitive elements.

In this work, we show that computational models of promoter classes defined on patterns of transcription initiation show class-specific enrichment of chromatin and sequence features. This supports the presence of divergent strategies of transcription, as recently proposed for yeast and for special functional classes of mammalian genes Ramirez-Carrozzi et al. (2009); Tirosh and Barkai (2008).

3.2 Methods

3.2.1 Selection of Human Transcription Start Sites

Promoters of tag clusters in the initiation region as defined in ENSEMBL, were classified as NP, BP, and WP based on the shape of their tag distributions. We utilized the published alignments of 29 million tags in human, generated by the FANTOM consortium Nechaev et al. (2010); Kawaji et al. (2009). Promoters were classified by means of two features, genomic span of initiation events (as defined by

the size of distinct 5' tag clusters), and localization of initiation. For NP promoters, tag clusters have to be smaller than 25 nt, and at least 50% of tags align at the peak location (defined as the mode of the cluster 62 nt). BP promoters exceed the 50% tag cutoff at the mode, but are spread out over a genomic range >25 nt. WP promoters are those which meet neither genomic span nor peak location cutoffs; they do however still show a distinct albeit lower peak, frequently associated with the presence of a minimal initiator sequence motif. This classification resulted in 1,409 NP, 1,759 BP, and 7,656 WP promoters falling in the initiation region that contained more than 100 reads. The modes of the tag distributions were used as representative TSS locations for all promoter classes.

3.2.2 Computing Nucleosome Profiles

Bulk nucleosome, H2A.Z, and H3K4 mono, di, and tri-methyl variant data from human CD4+ T cells were obtained from Schones et al. (2008); Barski et al. (2007). The nucleosome occupancy score for H2A.Z, H3K4 methylation, and bulk profiles was calculated according to Schones et al, using raw short aligned reads mapping to 5' or 3' nucleosome boundaries Schones et al. (2008). We divided each somatic chromosome into 10 bp non-overlapping windows, and read counts for a window were calculated by summing the number of reads that aligned in the 80 bp upstream (on the sense strand) or 80 bp downstream (on the anti-sense strand) windows, assuming that 5' and 3' reads mapping to the ends of the same nucleosome would be 140-160 bp apart. Promoters were analyzed in windows from -21 kb to +1 kb of the TSSs identified by tag clustering; to reduce the noise in the bulk data, promoters with outlier read counts less than 8 or greater than 2,400 were removed from the analysis.

3.2.3 Computational TSS Models Using Chromatin and Sequence Features

To evaluate the contribution of chromatin and sequence features to the definition of different promoter classes, separate linear classifiers for NP and WP promoters were trained on chromatin features, or combinations of sequence and chromatin features. These classifiers were then tested to determine how well they were able to distinguish between TSSs from the three promoter classes and other genomic locations.

Training and test data

NP and WP TSSs were divided into training and test data, using two-thirds of each set for training and the remaining samples for testing. For each TSS in the training set, 20 intergenic locations were drawn at random from -24000 to -2100 bp relative to the TSS. Additionally, one location was drawn from annotated CDS of human UCSC Known Genes, and two locations from annotated CpG islands without evidence of transcript activity (i.e. those without human CAGE aligned reads). CpG islands were initially taken from the UCSC Genome Browser annotation, which follows the definition by Gardiner-Garden and Frommer (1987): a >200 bp stretch with a G+C content of at least 50% and an observed vs expected ratio of CG dinucleotides of >0.6. We then filtered this initial set by the more recent criteria of Takai and Jones (2002), which led to a strict subset of regions with length >500 bp, G+C content >55%, and CG ratio >0.65. Intergenic, CDS, and CpG island locations together comprised the negative examples. For each of the remaining independent TSSs in the test set, we further randomly selected 100,000 CpG island locations (again sampled from those without human CAGE tags) as well as locations from anywhere in the genome. To ensure that each sample contained sufficient data for chromatin feature extraction, all positive and negative training and test samples passed a filter of at least eight aligned reads of the bulk nucleosome data. All analyses were also performed using unfiltered data, with consistent results (data not shown).

Feature generation

Chromatin or epigenetic features were designed to reflect similarity to the typical nucleosome profile surrounding a TSS. Epigenetic features were calculated as the inner product of an examples profile and a reference profile obtained from the respective training set. Reference profiles were generated by averaging the profiles of the respective TSS training set, split at the TSS in 2 kb upstream and 2 kb downstream regions. A total of 10 profiles were thus generated for each model, corresponding to Bulk, H2A.Z, and H3K4 monomethyl, dimethyl, and trimethyl profiles. The processed chromatin data was binned into 10 bp intervals, and the closest datapoint to the TSS location was used as the 0 location for relative profile coordinates. Each epigenetic profile was smoothed using a Discrete Fourier Transform Low Pass Filter with a low pass limit of 150 bp, eliminating noise at frequencies higher than an average nucleosome size. To select informative sequence features, position weight matrices (PWMs) of transcription factors were obtained from the JASPAR Core Vertebrate and RNA pol II datasets Portales-Casamar et al. (2010). We then followed the protocol described in Megraw et al. (2009), in which we previously described a classifier for murine NP promoters. Briefly, for each promoter class, TFs were filtered to those exhibiting match score enrichments in specific regions relative to the TSSs; these factor- specific enriched regions were each subdivided into seven windows. For every selected factor, background-normalized cumulative PWM scores were computed for each of the windows and used as features.

Model training, testing, and evaluation.

Further following the example of Megraw et al. (2009), we used L1-regularized logistic regression to learn a sparse linear classifier for each promoter class, as implemented in the l1 logreg package Rach et al. (2011). Sparse logistic regression selects features by assigning coefficients to each, while penalizing the use of large numbers of features.

Thus, coefficients of features that are not important for the classification problem are driven to zero and effectively excluded from the model. L1-regularized logistic regression uses the L1 penalty parameter to set the balance of including features. We performed 10-fold cross-validation to select the optimal L1 parameter for each model. The training data was divided into 10 parts, each part having an equal number of positive, negative intergenic, negative CDS, and negative CpG island examples. For each round of cross-validation, 8 parts were used for training, one for testing and selection of the optimal L1 parameter, and one for independent testing with the optimal L1 parameter. The range of L1 parameters for each cross- validation ranged from 0.0001 to 0.01. All training was performed using the l1 logreg data standardization option, normalizing for potentially different scales between features. After cross-validation, a final model was created by training on the entire training set with the mean optimal L1 parameter. The models were tested on the independent test data of each of the three classes, using the final NP and WP models generated on the full respective training sets. Classification performance was evaluated with two standard metrics: the receiver operating characteristics (ROC) and the precision recall curves (PRC), and the area under ROC (auROC) and PRC curves (auPRC), which summarize classifier performance when varying the true positive rate. While ROC effectively normalizes for differences in size of positives and negatives, PRC is sensitive to imbalanced datasets as is the case for promoters in which a small number of TSS locations are outnumbered by the non-TSS locations in the genome. This implies that ROC curves are comparable for different classifiers (e.g. NP and WP), while PRC curves will reflect differences in the relative size of the positive class. This partially explains the larger differences we observed for auPRC values, which reflects the harder problem of identifying fewer NP than BP promoters within a large genomic background. To visualize the importance of features for each class, a modified version of l1 logreg was used to obtain standardized coefficients, representing

input features normalized to the same scale. From these standardized coefficients, we determined which features were consistently present during the ten-fold cross-validation training step. For each model, we determined the features whose absolute value was greater than 0.05 in at least 8 of the 10-fold cross-validations.

3.3 Results

We determined TSS clusters from available human CAGE tags in the FANTOM4 database Kawaji et al. (2009) (see Methods). 13% of promoter clusters fell into the NP class, 16% into the BP class, and 71% were classified as WP. As core promoters have traditionally been characterized and identified by the presence of regulatory sequence elements, we sought to quantify how informative chromatin features would be to define human TSSs. Specifically, we were interested in how strongly the different promoter classes were defined by sequence versus chromatin features. To this end, we trained and applied computational models to classify between TSS versus non-promoter genomic locations. Our goal was to identify potential differences between promoter classes when comparing models under the same assumptions side-by-side, similar in spirit to recent splicing simulators integrating sequence and chromatin features Spies et al. (2009). We computed average profiles of the 2 kb upstream and downstream regions of each TSS for bulk and H2A.Z nucleosomes as well as H3K4 mono-, di-, and trimethylation, for a total of 10 representative profiles for each promoter class. The inner products of the representative profiles with those of a genomic test location were used as input features for sparse linear classifiers, trained separately for WP and NP promoters. Each model was then tested on independent data of WP, NP, and BP promoters (Figure 3.1), as well as negative samples from other genomic locations, including CpG islands without evidence of transcription. WP and BP classification was much more accurate than NP.

Inspection of the model features showed that each class relied on similar features,

selecting an informative subset of nucleosome profiles (Figure 4). The highest weight was assigned to the H3K4 trimethylation downstream profile, followed by the H2A.Z profiles, likely due to the strong periodic signal especially within the transcript. In fact, applying the WP model for the recognition of NP promoters was more successful than using the model trained on NP promoters themselves. Overall however, results stayed well below those obtained on both WP and BP promoters. When adding Fourier-transform based features to reflect the periodicity of nucleosomes, results were slightly improved but highly consistent (Figure 3.2).

It has been demonstrated that NP promoters could be characterized with great success by ensembles of transcription factor binding sites based on their enrichment at specific locations relative to the TSS, using features beyond the strict core promoter sequence motifs (including factors such as E2F, CREB, YY1, etc) Megraw et al. (2009). Following this example, and using the performance of the chromatin models as baseline, WP classifiers built on sequence features performed considerably worse than the WP chromatin model (Figure 3.3). The opposite was true for NP promoters, for which sequence models achieved higher success rates on NP and BP promoters than chromatin models. Combining sequence and chromatin features increased accuracy on all test sets, and demonstrated that WP TSSs relied much more on chromatin features than NP TSSs. This was seen in both the relative changes of classification accuracy as well as in the relative strength of features within the combined models, in which chromatin features accounted for stronger contributions for the WP compared to the NP model (Figure 3.3).

3.4 Discussion

The high-throughput sequencing of 5' capped sequence tags has clearly shown that eukaryotic promoters separate into at least two classes defined by focused and dispersed distributions of initiation events. Many recent studies have reported on the

chromatin structure in eukaryotic genomes; our approach differed from most of these efforts by assessing chromatin features from the basis of transcription initiation as derived from 5' tag data. In one exception, work concurrent to ours found differences on H3K9 acetylation based on different promoter classes Kratz et al. (2010). Here, we provide support that promoters from different classes not only contain different core promoter sequence features, but also reflect distinct patterns of nucleosome organization, and chromatin structure.

A separation of mammalian promoters has frequently been proposed based on the presence of CpG islands. Differential regulation of some promoters with CpG islands has been shown to result from unstable nucleosomes, contrary to the involvement of chromatin remodelers at non-CpG island promoters Ramirez-Carrozzi et al. (2009). Somewhat differently, we found that CpG islands are present across all initiation patterns, which indicates that CpG islands are not a homogeneous class and do not all encode constitutively unstable arrangements of nucleosomes. The work by Ramirez-Carrozzi et al. Ramirez-Carrozzi et al. (2009) focused on a specific set of promoters, those adjacent to stimulus-response genes, in which nucleosomes are pre-organized to facilitate a regulated primary response. Such genes may form an intermediate class between constitutively expressed genes typically associated with CpG islands, and NP promoter genes, which contain genes like developmental TFs that are expressed in a precisely determined and highly regulated order. Multiple aspects may contribute to the relationship between the promoter classes and chromatin features. First, differences in chromatin architecture may be directly reflected in distinct initiation patterns, as illustrated by the nucleosome organization in constitutive versus regulated genes in yeast Cairns (2009). Thus, dispersed promoters result from a well-defined NFR increasing the accessibility of the DNA to the polymerase, causing initiation to occur at multiple locations over a large region. In turn, the lower accessibility of focused promoters provides for a more regulated transcription initiation

due to the lack of a common NFR. Instead, TSSs of focused promoters are well-defined by position-specific sequence elements including the canonical core promoter motifs Ni et al. (2010); Megraw et al. (2009), which serve to actively recruit the core complex to precise TSS locations. Our computational models clearly support this idea: chromatin features contribute to NP promoter definition, but much less so than for other classes, and with little improvement on sequence information. As more data becomes available through large-scale efforts such as the modENCODE and ENCODE projects, the presence of high-level divergent strategies of gene regulation established at the basal promoter will become better characterized throughout development and differentiation in model organisms as well as in human. Promoter classes may have associations to epigenetic inheritance, cellular memory, evolvability, and the development of disease Tirosh et al. (2009); Bernstein et al. (2007). Understanding initiation patterns does not only help deepening our knowledge of the core promoter sequence, but also provide insight into the epigenetic architecture of regulatory regions. Together, they illustrate the interplay between chromatin and sequence information to encode divergent strategies for gene expression.

3.5 Acknowledgements

We would like to thank Kevin White and his lab at the University of Chicago for making the H3K4 and insulator data available ahead of publication, as part of the modENCODE Consortium; Gregory Crawford at Duke University for use and assistance with DHS data generated as part of the ENCODE Consortium; and Molly Megraw at Duke University for assisting with software for computational promoter models.

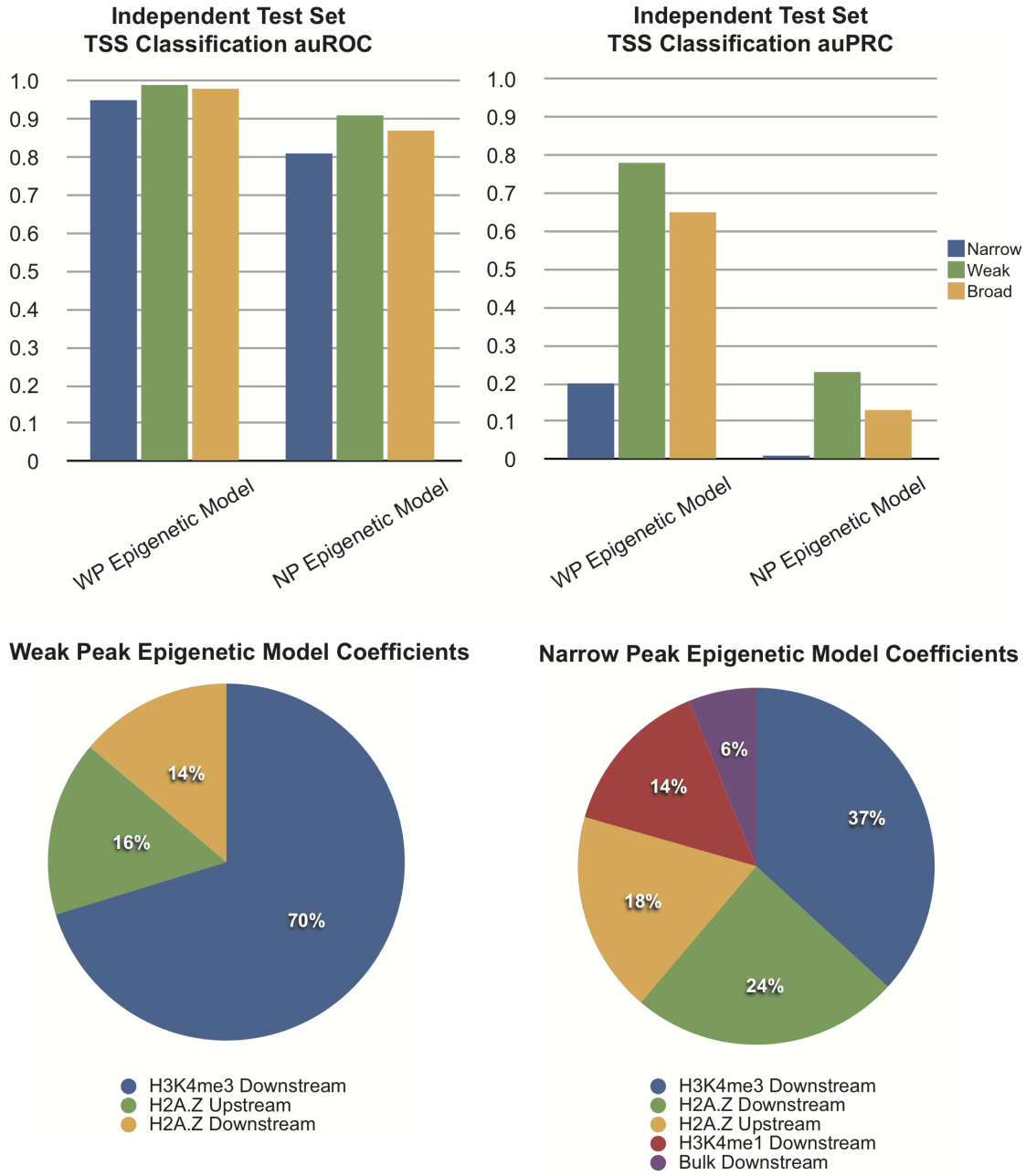

FIGURE 3.1: Computational Models Using Chromatin Features Show Different Accuracy for Promoter Classes. Classification accuracy of two epigenetic models (i.e., using chromatin features) was evaluated on test sets for each promoter class (evaluated with auROC and auPRC). Values of 1 indicate perfect classification; auROC values close to 0.5 and auPRC values close to 0 reflect random results. At the bottom, relative weights of chromatin profile features included in each model are depicted.

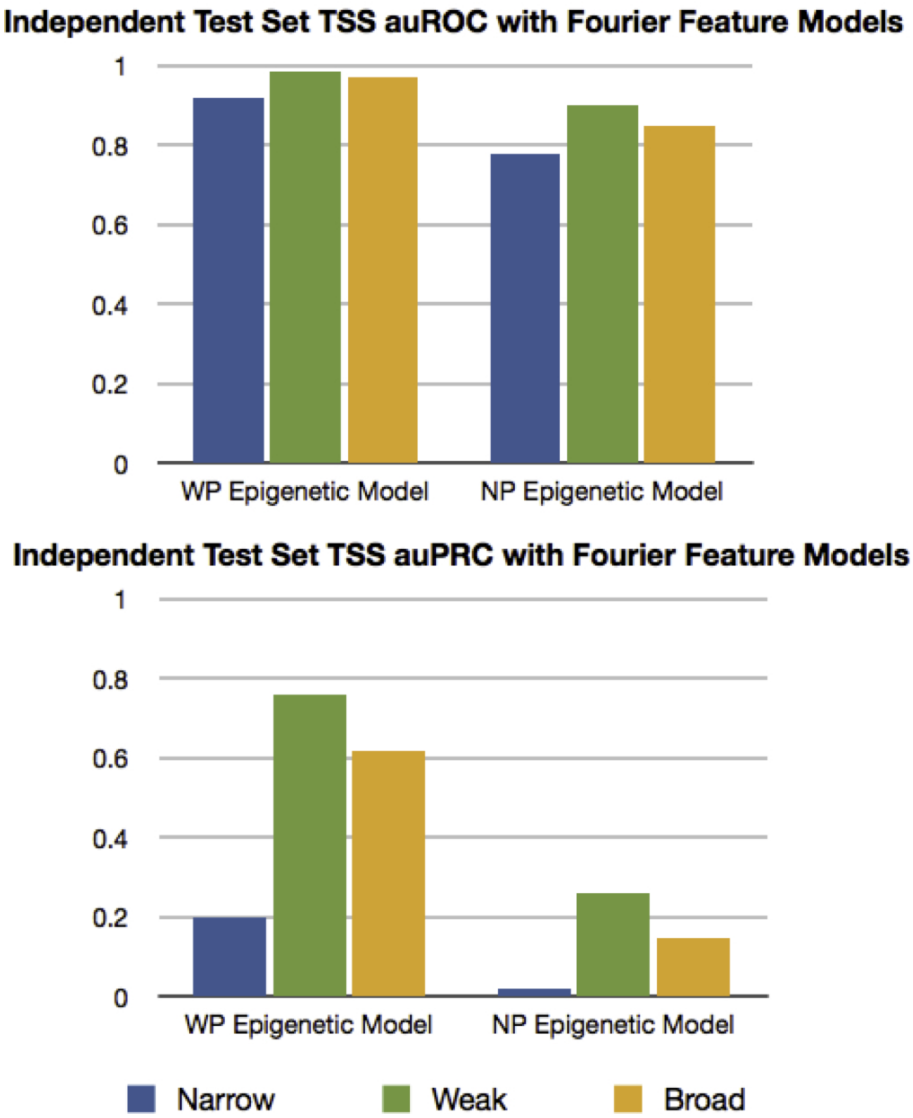

FIGURE 3.2: Including Fourier TransformBased Chromatin Features in a Computational TSS Model. We explored the effect of adding Discrete Fourier Transform (DFT) coefficients as features, in addition to the epigenetic profile features. The Fourier transform decomposes a signal into its spectral components, and coefficients reflect the presence of periodicities within the data. The DFT was computed in Matlab, on the data pre-processed as described in the main text. As with the profile features, DFT coefficients were computed for the 2 kb upstream and 2 kb downstream regions relative to the TSS, for the whole 2 kb windows as well as smaller 500 bp sliding windows, moved within the 2 kb regions 250 bp at a time. DFT coefficients were computed for Bulk, H2A.Z, and H3K4 monomethyl, dimethyl, and trimethyl profiles, and coefficients reflecting periodicity in the range of a nucleosome turn were added to the features for model training.

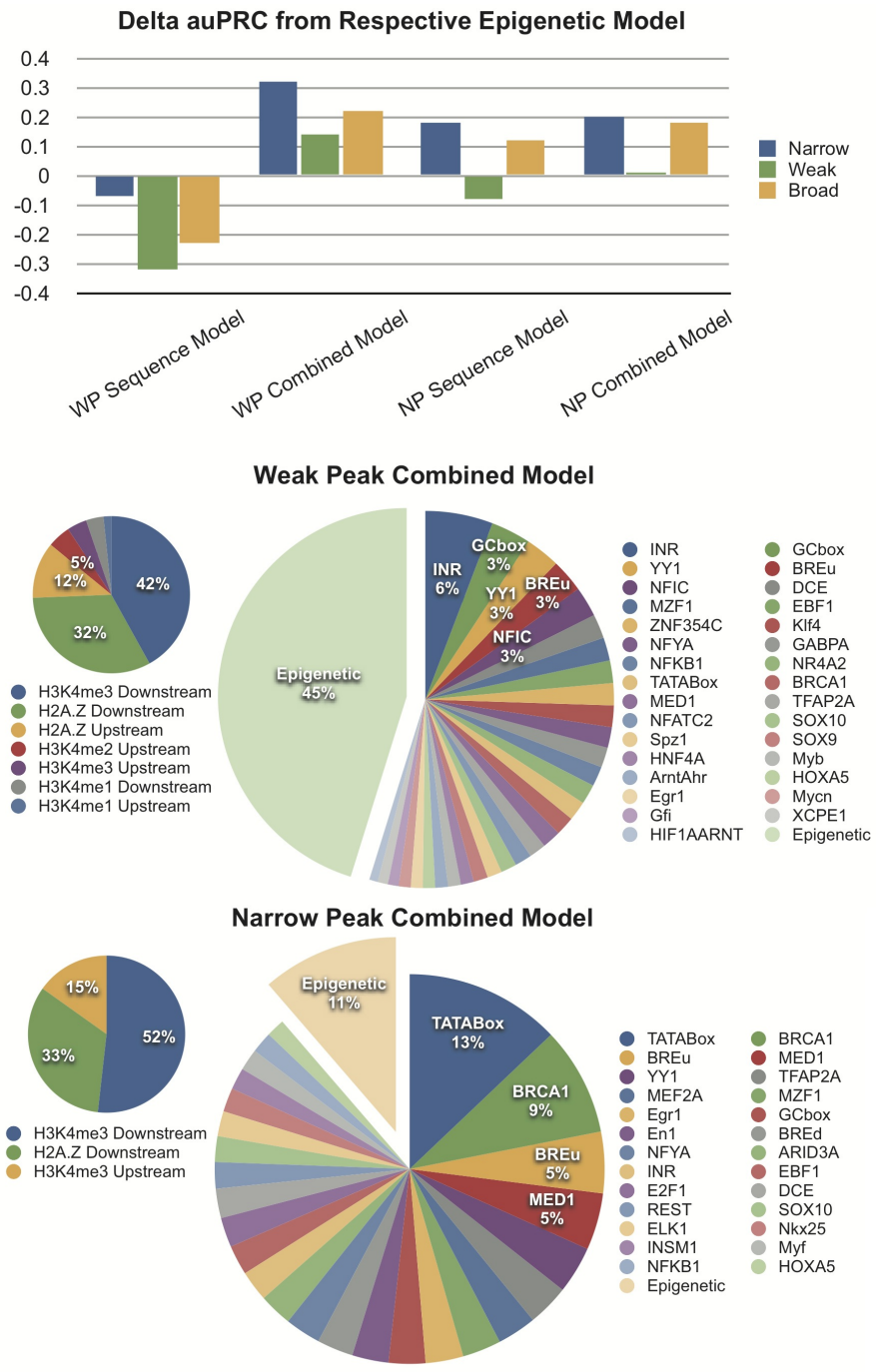

FIGURE 3.3: Computational Models Using Chromatin Features Show Different Accuracy for Promoter Classes. Classification accuracy of two epigenetic models (i.e., using chromatin features) was evaluated on test sets for each promoter class (evaluated with auROC and auPRC). Values of 1 indicate perfect classification; auROC values close to 0.5 and auPRC values close to 0 reflect random results. At the bottom, relative weights of chromatin profile features included in each model are depicted.

4

RNA-Seq Mapping of Non-Human Primate Data to Build Human Clinical Models

4.1 Background

For the past decade, microarray gene expression data has revolutionized all areas of life science allowing quantification of thousands of genes in several samples simultaneously, and paving the way for countless research studies. With the advent of RNA-Seq data, researchers now have the ability to perform untargeted gene expression analysis via next generation sequencing (NGS) technology, obtaining qualitative sequence information as well as quantitative gene expression data. RNA-Seq provides a comprehensive gene expression profile of each sample with the potential to quantify and annotate all genes and isoforms. This untargeted approach proves particularly useful when quantifying gene expression in polymorphic cell lines and in organisms with a nonexistent or provisional reference genome where the sequence of features to be quantified is unknown Wang et al. (2009); Hornett and Wheat (2012). Due to the limited genomic resources available for under-characterized species, RNA-Seq is a popular method of choice for differential gene expression analyses. We present a

comparison of reference-based mapping methods for RNA-Seq data originating from Non-Human Primates (NHPs), and their implications in downstream differential expression analysis.

4.1.1 Mapping and Assembly

RNA-Seq obtains gene expression estimates by assigning next-generation sequencing (NGS) reads to transcripts, either by mapping to a reference sequence, or assembling into contiguous stretches of sequence data (contigs) utilizing overlapping sequence amongst the reads themselves. These methods are termed reference-based alignment approaches, and *de novo* assembly approaches, respectively. Reference-based alignment and *de novo* assembly each have advantages and disadvantages. The optimal strategy likely depends on the experimental design and available genomic and computational resources. Reference-based approaches are far less computationally intensive than *de novo* approaches, and are well-suited for detection of low abundance transcripts Martin and Wang (2011). However, reference-based mapping approaches do rely on the availability and accuracy of reference genomes and transcriptomes. Mapping approaches must also handle reads with more than one potential mapping location. Such uncertainty can arise from paralogous gene families, repetitive sequences, and shared exons of alternatively spliced transcripts Li et al. (2010). *De novo* assemblers have the advantage of not requiring a reference sequence. This allows discovery of transcripts not present in a reference genome or transcriptome. *De novo* assemblers overcome challenges caused by mapping uncertainty, as well as long introns that may be missed by reference-based mapping methods. However, *de novo* assemblers require large amounts of computational resources and assembly time. In addition, higher sequencing depth is needed for adequate *de novo* transcript assembly than is required by reference-based approaches Martin and Wang (2011). *De novo* assembly can be very useful for organisms with no reference genome, but

due to its computational demands is most commonly used in cases of small genomes such as bacteria, archaea, and lower eukaryotes Martin and Wang (2011). In the case of large and complex transcriptomes, such as plants and mammals, reference-based mapping methods can overcome the computational and isoform-resolving challenges faced by *de novo* assemblers Martin and Wang (2011). In general, reference-based mapping approaches are the appropriate choice when a reliable reference genome exists. The choice becomes less clear when working with complex mammalian species with little to no genomic resources, such as under-characterized NHPs. *De novo* assembly-based approaches are computationally intensive, show reduced sensitivity for genes expressed at low levels, and concerns have been raised over the quality of *de novo* transcriptomes in comparison to reference-based approaches Martin and Wang (2011). In large-scale NHP studies, *de novo* assembly may be cost and time prohibitive.

We propose the use of a human reference genome (or transcriptome) for reference-based mapping and expression quantification of NHP RNA-Seq data. An elegant study by Hornett et al. Hornett and Wheat (2012) evaluated the utility of divergent species gene sets for annotation of *de novo* assembly. When utilizing reference gene sets from divergent species, the authors found that there was little bias in expression levels and strong correlation in gene expression up to approximately a 100 million year window. More divergent species (greater than 100 million years apart) suffered from incorrect assignment of assembled contigs to genes. The authors found little difference in the number of genes having contigs assigned, when using chimpanzee, orangutan, macaque, or marmoset vs. human. In addition, the authors compared the use of *de novo* assembled transcriptomes to mapping directly to the reference predicted gene set for quantifying gene expression. When comparing the mapping of the reads directly to the predicted gene set vs. the *de novo* assemblies, the authors found that mapping to the gene set recovered expression data for more genes, and

that within the shared detected gene sets, correlation was high Hornett and Wheat (2012).

We examine the mapping efficacy and utility for differential expression analyses of various reference-based approaches when the reference sequence originates from a related species. Specifically, we utilize the human genome and transcriptome as a reference for non-human primate RNA-Seq data from yellow baboons, *Papio cynocephalus*. We compare four different reference-based mapping methods one representative method from four different mapping method categories Garber et al. (2011).

4.1.2 Reference-Based Mapping Methods

Reference-based alignment methods utilize the sequence for each read (and it's mate in paired-end data) to find potential mapping locations by exact match or scoring sequence similarity. Mapping locations indicate transcripts of origin, and the number of reads originating from a given transcript informs on how much of the transcript was present in the sample. We briefly describe four categories of reference-based mapping methods, and subsequently show the results of one representative method from each category when mapping RNA-Seq reads from yellow baboon to a human reference. A summary of the four categories and the representative methods tested in this study is shown in Table 4.1.

Table 4.1: Reference-Based Mapping Methods Overview - Summary of the four categories of reference-based mapping methods compared in this study. * - Bowtie2 may be considered a hybrid BWT-Seed Method, as multiple substrings are taken from each read for the BWT lookup of candidate mapping loci, and the alignment at each candidate loci is extended.

Method Category	**Reference**	**Method Subcategory**	**Representative Method**	**Notes**
Unspliced Aligners	Transcriptome	Burrows-Wheeler Transform Method	Bowtie2*	Index reference sequence, Rapidly look up candidate mapping loci. Typically faster and less sensitive than Seed Methods.
		Seed Method	Stampy	Align short subsequences of reads to find candidate mapping loci, Narrow candidates by extending alignments. Typically slower and more sensitive than Burrows-Wheeler Transform Methods.
Spliced Aligners	Genome	Exon First Method	TopHat2	Align whole reads with Unspliced Aligners, Search for spliced alignments in remaining reads. Typically faster and less sensitive than Seed and Extend Methods.
		Seed and Extend Method	GSNAP	Align short subsequences of reads to find candidate mapping loci, Narrow candidates by extending alignments. Typically slower and more sensitive than Exon First Methods.

The first major category has been referred to as unspliced read aligners, which align reads to a reference without allowing large gaps Li et al. (2008, 2009); Li and Durbin (2009); Langmead et al. (2009); Rumble et al. (2009); Lunter and Goodson (2011); Langmead and Salzberg (2012). Unspliced aligners may be used to align reads to a reference transcriptome. This alleviates the need to handle splice junctions, but is limited to the analysis of known transcripts. Unspliced read aligners are generally divided into two subcategories based on their methodology seed methods and Burrows-Wheeler transform (BWT) methods Garber et al. (2011). Seed methods align short subsequences, or seeds, from each read to a reference, requiring a perfect match in the seed subsequence. More sensitive alignment methods are used to eliminate candidate regions where seeds cannot be extended to full read alignments. The unspliced seed method we chose to test is Stampy Lunter and Goodson (2011). BWT methods create a Burrows-Wheeler index of the reference genome and efficiently search for perfect matches. Mismatches can be allowed with an exponential increase in computational complexity. In general, Burrows-Wheeler transform methods are faster than seed methods, but seed methods provide increased sensitivity Martin and Wang (2011). The unspliced BWT method we chose to test is Bowtie2 Langmead and Salzberg (2012). A similar analysis reported that when the true reference transcriptome is available BWT methods are faster with minimal differences in alignment specificity. When the reference transcriptome of a distant species is available, seed methods result in large increases in sensitivity Garber et al. (2011). These increases in sensitivity have also been observed when aligning reads to polymorphic regions Degner et al. (2009).

Methods in the unspliced read aligner category are limited to known exons and splice sites. The second major category of mapping methods, spliced aligners, align reads to the whole genome, with intron-spanning reads requiring large gaps. Spliced aligners accommodate junction-spanning reads by splitting them up into smaller

segments and determining the best match based on alignment scores and known dinucleotide splice signals De Bona et al. (2008); Trapnell et al. (2009); Au et al. (2010); Wang et al. (2010); Wu and Nacu (2010). The spliced aligners also fall into two major categories based on their methodology exon first methods and seed and extend methods. Exon first methods begin by mapping whole reads to the genome using unspliced read aligners, and then search for spliced alignments with the remaining reads. Exon first approaches are efficient, but may miss true spliced alignments when an unspliced alignment is available in a pseudogene Garber et al. (2011). Seed and extend methods break reads into seeds which are mapped to the genome, and much like seed-based unspliced aligners, candidate mapping locations are examined with more sensitive alignment methods. Iterative extension and merging of initial seeds is performed to determine the spliced alignment. As with unspliced aligners, seed and extend methods are slower, but more sensitive, and perform better when mapping reads from polymorphic samples. Garber et al. (2011) The exon-first and seed and extend methods we chose to test are TopHat2 Kim et al. (2013), and GSNAP Wu and Nacu (2010), respectively.

After mapping to a reference genome, a transcriptome reconstruction step is required to appropriately assign reads to transcipts. Aligned reads spanning splice junctions are connected, and read counts to various isoforms of each gene are reported. This is accomplished by building a graph to represent all possible isoforms of an expressed feature. Different paths through the graph represent individial isoforms. Two commonly-used software packages for transcript assembly are Cufflinks Trapnell et al. (2010) and Scripture Guttman et al. (2010).

We present a comparison of reference-based mapping methods for RNA-Seq data having no true reference, but that of a closely related model organism. We assess the various methods using mapping rates, mapping locations, correlation of gene expression, as well as the utility of the data for differential expression analyses and

building predictive models for a phenotype of interest.

4.2 Methods

Study Samples

We obtained data from 12 adult male yellow baboons (*Papio cynocephalus*), inoculated with five different levels of *Streptococcus pneumoniae*, a bacterial pathogen causing pneumonia. The bacterial doses administered to the participants included 10^9 colony forming units (CFU) (n = 4), 10^8 CFU (n = 3), 10^7 CFU (n = 1), 10^6 CFU (n = 1), and 0 CFU (n = 3). Peripheral blood samples for gene expression analysis were taken at five different time points immediately before inoculation, and 6, 24, 48, and 168 hours following inoculation. Antibiotics were administered immediately following the 48 hour time point. Three of the 60 samples did not meet RNA quality standards. Each participant was evaluated to determine clinical pneumonia status using pre-determined criteria. A participant was classified as developing clinical pneumonia if each of the three following conditions were met:

- A white blood cell count of greater than 15,400, less than 3,400, a 2fold change from baseline measurement, or greater than or equal to 90% neutrophilia at 24 or 48 hours.
- A positive culture of Streptococcus pneumoniae from BAL or blood samples at 48 hours.
- Any one or more of the following at 24 or 48 hours:
  - Heart rate of greater than 100 bpm
  - A 25
  - Positive indication of infiltrate on a chest X-ray
  - Decreased Activity

- Decreased Food Intake
- A fever of 38.2 degrees Celsius or greater
- Cough
- Nasal Discharge

RNA Isolation and Library Preparation

Samples were collected in PreAnalytiX PAXgene Blood RNA collection tubes, and total RNA was isolated using the PreAnalytiX PAXgene Blood RNA miRNA isolation kit. RNA quality was assessed using Agilent Bioanalyzer and Nanodrop spectrophotometry, and samples with RNA integrity number greater than or equal to 7, and greater than or equal to 1 microgram of RNA were deemed sufficient for RNA-Seq. Abundant globin transcripts were depleted with the GlobinClear Globin RNA Reduction for RNA-Seq protocol. The fragment library was prepared with the Illumina TruSeq RNA Seq protocol, and Illumina HiSeq RNA Sequencing was performed, run in 6-plex per flow cell lane, obtaining 50 bp paired-end reads.

Read Quality Control and Trimming

Read quality analysis was performed on the raw data using FastQC version 0.10.1 Andrews (2010). Quality trimming and adapter clipping were performed using Trimmomatic version 0.25 Lohse et al. (2012), trimming trailing bases below quality 20, clipping Illumina adapters, and discarding clipped reads shorter than 25 bp. FastQC was used to re-assess the integrity of the clipped reads prior to subsequent mapping and analysis. Reads whose mates were discarded due to quality trimming and length constraints were removed from the fastq files used for mapping.

Read Mapping

The UCSC hg19 human reference genome and annotation was used as a reference, from the Illumina iGenomes download, March 2013. To generate a fasta file of transcripts for unspliced mapping, the RSEM prepare reference tool was used citeli2011rsem. Clipped reads were mapped to the hg19 transcriptome using Bowtie2 version 2.0.6. Langmead and Salzberg (2012), and Stampy version 1.0.17. Lunter and Goodson (2011). Clipped reads were mapped to the hg19 human genome using Tophat version 2.0.7 Kim et al. (2013), and GMAP (GSNAP) version 2013-03-12 Wu and Nacu (2010). With Tophat, the unmapped reads were merged into the mapped SAM file. Default parameter settings were used for all methods. SAM/BAM conversions, sorting, indexing, and marking of PCR duplicates were performed with SAMtools version 0.1.18 Li et al. (2009) and Picard version 1.83 None. (2013).

Quantification and Normalization

Read counts for each transcript were obtained with HTSeq Planet et al. (2012), specifically the intersection-nonempty mode of htseq-count. Conditional Quantile Normalization was used to obtain normalized gene expression estimates Hansen et al. (2012).

Mapping Comparisons

To compare the four reference-based mapping methods, we examined several mapping metrics including mapping rates, mapping locations, transcripts detected, mate pair concordance, and coverage over transcripts. Mapping metrics were computed with RNA-SeQC version 1.1.7 DeLuca et al. (2012), for the entire gene set, as well as subsets of genes in Gene Ontology Ashburner et al. (2000) functional groupings, and evolutionary distance groupings. See the corresponding methods sections for more details on the generation of these gene lists.

Functional Groups

To assess the number of detected genes within different functional groupings, we took the set of genes present within each of the 23 top-level biological process ontology terms. BEDtools version 2.17.0 Quinlan and Hall (2010) was used to select reads mapping to each gene list from each of the BAM files, which were then analyzed with RNA-SeQC.

Evolutionary Distance Groups

To assess the number of detected genes at various evolutionary distances, we obtained the mRNA sequence of all olive baboon (*Papio anubis*) RefSeq genes from the UCSC genome browser, and determined the human ortholog and evolutionary distance of each reference baboon gene. Orthologs were found by performing a BLASTN Altschul et al. (1990) search against the human transcriptome, and taking the top hit by percent identity. Jukes-Cantor distances Matsubara et al. (1968) were then computed between the orthologous sequences. Five subset gene lists were created using evolutionary distance: greater than or equal to 0.1 (n = 19), less than 0.1 and greater than or equal to 0.075 (n = 20), less than 0.075 and greater than or equal to 0.05 (n = 62), less than 0.05 and greater than or equal to 0.025 (n = 202), and less than 0.025 (n = 155). As with the functional gene lists, BEDtools version 2.17.0 Quinlan and Hall (2010) was used to select reads mapping to each gene list from each of the BAM files.

Differential Expression Analysis

For each of the four mapping methods, we determined the differentially expressed transcripts using edgeR Robinson et al. (2010). We report the genes exhibiting significant differential expression between healthy participants, and participants with clinical pneumonia at maximal symptoms (48 hours) after Bonferonni correction. A

4-way venn diagram web tool was used to generate the figure Oliveros (2007).

Predictive Model from Gene Expression Data

Using the normalized gene expression estimates, we built a statistical classifier for clinical pneumonia using the 7 participants meeting clinical criteria for pneumonia at maximal symptoms (48 hours), vs. all baseline samples, and all control participants at 48 hours. The normalized expression estimates for the top 100 differentially expressed genes were used as predictive variables, and elastic net regularized regression was used to fit the model. Leave one out cross-validation (LOOCV) was used to obtain classification results for each sample, and to obtain an estimate of how well the classifier would perform on an independent test set.

4.3 Results and Discussion

We mapped RNA-Seq reads from 57 yellow baboon (*Papio cynocephalus*) peripheral blood samples to a human reference using four different mapping methods - Bowtie2 Langmead and Salzberg (2012), Stampy Lunter and Goodson (2011), TopHat2 Kim et al. (2013), and GSNAP Wu and Nacu (2010). We then assessed the utility of each mapping method with mapping rates, base mismatch rates, mapping locations, detected transcripts, correlation of gene expression estimates, differential expression analysis results, and the predictive utility of gene expression data. The 57 baboon samples were taken from 12 different animals at 5 time points, where the phenotype of interest is clinical bacterial pneumonia infection (see Methods for more details). When mapping RNA-Seq reads from a distant NHP species to human reference sequence, we see differences in the utility of various mapping methods with respect to mapping rates, detected genes, correlation of expression values, and differentially expressed genes. To better understand the differences between the methods, we examine more closely the default behaviors of the four mapping methods used.

Bowtie2 is a BWT-based unspliced aligner, with the recent addition of supported gapped alignments. This method may actually be considered a combination BWT and seed unspliced mapper. Bowtie2 extracts multiple substrings or seeds from each read and aligns them using a BWT approach with no gaps, then extends alignments using a Smith-Waterman-like scoring scheme. By default, seeds are 22bp and no mismatches are allowed within the seed. Base call quality scores are incorporated by assigning more severe mismatch penalties at high-quality read positions. Gap initiation and extension penalties are also utilized, while the number and lengths of gaps within extended alignments are not restricted. Bowtie2 does not guarantee that the alignment reported is the best in terms of alignment score, and when there is more than one potential mapping location of equal score, one reported location is selected at random.

Stampy is a seed-based unspliced aligner that uses a hash table to store the locations of 15-mers in the reference sequence. For each read, candidate alignment locations are identified with a hash lookup of the 15-mers in the read, allowing for one mismatch. Candidate mapping locations are screened for sequence similarity with the read, and then full alignments are attempted at each remaining candidate location. As with Bowtie2, Stampy also considers base quality calls, and allows gaps in this alignment step. Stampy also allows the use of BWT as a "pre-mapping" step to increase speed, however the manual does not recommend this for paired-end data. For this reason, and for method-comparison purposes, we did not use the BWT option. Stampy uses a Bayesian probabilistic model to represent mapping quality, and reports the single best alignment location.

TopHat2 is an exon-first spliced read aligner that uses Bowtie2 as a base algorithm. TopHat2 has the recent additions of the ability to align reads across fusion breaks. Like the original TopHat, potential splice sites are detected within candidate alignment locations, and used in a subsequent step to align reads spanning

exon-junctions. TopHat2 first maps to the transcriptome with Bowtie2. Remaining whole reads are then mapped to the reference genome, and then spliced alignments are attempted. Most of the default Bowtie2 parameters when run within Tophat2 are the same as the default standalone Bowtie2 parameters, with the exception of seed length and intervals between seeds. TopHat2 seeds within Bowtie2 are 20bp, and the interval between seeds is longer. TopHat2 reduces alignment to pseudo-genes by aligning reads preferentially to genes within provided annotation. This use of annotation by TopHat2 has been shown to increase sensitivity and accuracy of mapping. We provided TopHat2 with annotation information for this purpose. In addition to gapped alignment in the Bowtie2 step, TopHat2 also allows indels in the spliced alignment detection step.

GSNAP is a seed-extend spliced aligner that allows for long and even chromosome spanning gaps, likely resulting from gene-fusion events. GSNAP uses all 12-mers in the read to identify candidate mapping locations, not favoring positions within short reads. Alignments are extended at candidate loci, requiring that the read has a consecutive stretch of 14 nucleotides exactly matching the reference sequence. GSNAP allows multiple mismatches and long indels, but only allows one splice or indel per read. Splicing is identified in two different ways - searching surrounding sequence for splice signals, or a user-provided set of known exon-intron boundaries.

Mapping Statistics

For each of the four mapping methods, we computed the average and standard deviation of several mapping statistics across the 57 samples. Table 4.2 shows a summary of these mapping statistics.

Table 4.2: Reference-Based Mapping Results Overview - Summary of mapping metrics results for the four reference-based mapping methods assessed in this study.

Comparison Metric	**Value**	**Bowtie2**	**Stampy**	**TopHat2**	**GSNAP**
Mapping Rate	Mean	0.673	0.885	0.618	0.472
	Standard Deviation	0.016	0.021	0.008	0.007
Base Mismatch Rate	Mean	0.113	0.128	0.020	0.017
	Standard Deviation	0.001	0.002	0.000	0.00
Intragenic Mapping Rate	Mean	1	1	0.964	0.896
	Standard Deviation	N/A	N/A	0.007	0.012
Intergenic Mapping Rate	Mean	0	0	0.034	0.104
	Standard Deviation	N/A	N/A	0.007	0.012
rRNA Mapping Rate	Mean	0.011	0.011	0.011	0.011
	Standard Deviation	0.003	0.003	0.003	0.003
Detected Transcripts	Mean	31615	37585	31338	27264
	Standard Deviation	803.209	815.801	807.139	945.805
Correlation Between Baseline Samples	Mean	0.976	0.970	0.980	0.977
	Standard Deviation	0.006	0.011	0.005	0.005
Differentially Expressed Genes	Number	1283	1333	450	1019
Predictive Utility	auROC	1	1	1	1

We first examine overall mapping rates and locations of the four methods, illustrated in Figure 4.1.

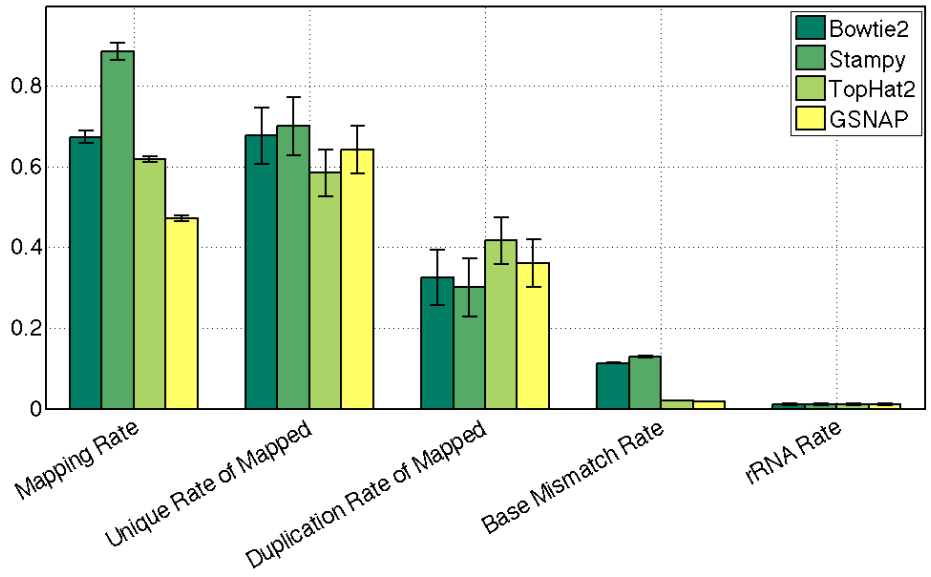

FIGURE 4.1: Mapping Statistics for Reference Transcriptome and Reference Genome Methods - Mapping, unique, duplication, base mismatch, and rRNA rate for each of the four mapping methods. Error bars show plus/minus one standard deviation. Mapping rate is computed as mapped reads divided by total reads, unique rate is computed as unique mapped reads divided by mapped reads, duplication rate is computed as duplicate mapped reads divided by mapped reads, base mismatch rate is computed as the number of bases not matching the reference divided by the number of aligned bases, and rRNA rate is computed as the number of reads mapping to ribosomal RNA divided by the total reads.

Examining the overall mapping rates of the four methods, we see that GSNAP obtains the lowest mapping rates, ranging from 0.4497 to 0.4857. This is likely due to the 14 nucleotide exact match requirement, and the single gap or splice restriction. Any reads spanning exon junctions that would align well with an additional small gap on either side of the intron will not be considered. Other reads mapping to divergent regions may not have a continuous stretch of 14 conserved nucleotides. TopHat2 was

reported to be more sensitive and accurate than GSNAP Kim et al. (2013), which was of similar sensitivity as the original TopHat Wu and Nacu (2010). GSNAP was found to perform poorly on short-anchored reads (small "anchor" on either end of a splice junction) Kim et al. (2013). TopHat2 and GSNAP were found to perform similarly on single-end reads with small indels (1-3bp), but GSNAP performed better on indel alignments with paired-end reads Kim et al. (2013). TopHat2 obtains significantly higher mapping rates ranging from 0.5940 to 0.6365. Bowtie2 obtains higher still mapping rates of 0.6269 to 0.7052. We see slightly lower mapping rates with TopHat2 than Bowtie2, regardless of the fact that Bowtie2 is the underlying algorithm of TopHat2. As previously mentioned, the only difference in default parameters is the seed length and interval between seeds. TopHat2 runs Bowtie2 with a shorter seed, which would suggest increased sensitivity, but a longer interval between seeds, which would lead to decreased sensitivity. Because the interval between seeds is longer, less seeds are used to identify candidate mapping locations, and with the default of no mismatches allowed within a seed some correct alignment loci would not be considered. Finally, Stampy achieves the highest mapping rates ranging from 0.8199 to 0.9358. This is likely due to the shorter seed length, and the allowance for a single mismatch in seeds. This allows more candidate loci to be considered. All increases in mapping rates were significant.

We also observe higher base mismatch rates in the transcriptome mapping methods. This is likely due to the fact that these methods are more sensitive, and so reads may be successfully mapped to more divergent regions. Another possible explanation is the presence of splice variants in baboon not present in human. The human reference transcripts may contain additional or missing exons with respect to baboon, causing mapped reads that span the true exon junction to have a high mismatch rate for a short stretch of sequence. Unique, duplication, and rRNA rates are all similar across the four mapping methods.

We also examined the mapping locations - intergenic and intragenic (exonic and intronic). These results are shown in Figure 4.2

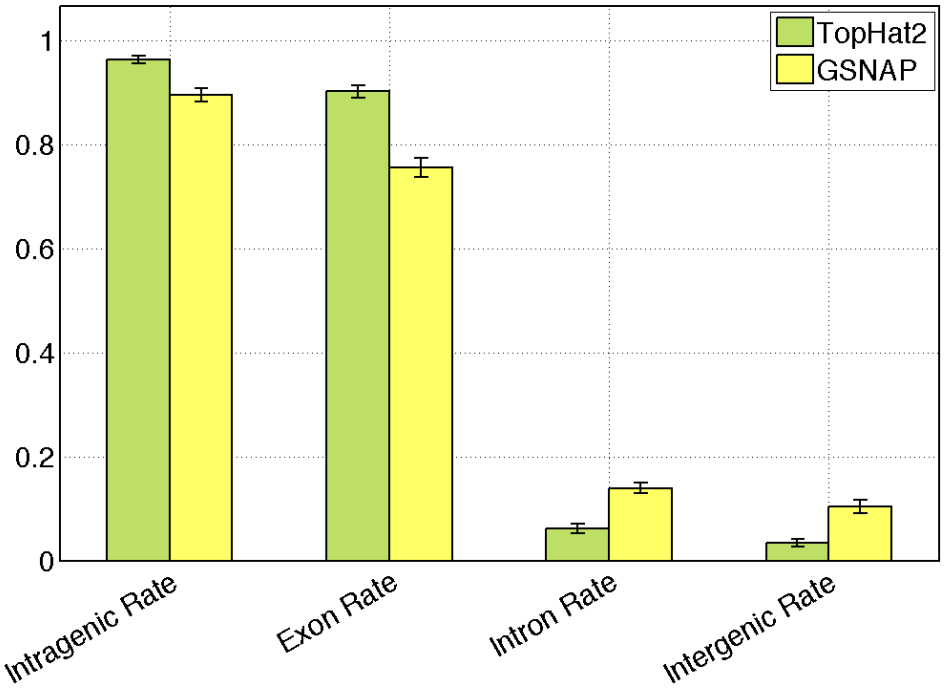

FIGURE 4.2: Mapping Locations Reference Genome Methods - Mapping locations for the two reference genome mapping methods. Each value is computed as the number of reads mapping to a type of region divided by the total reads mapped.

The transcriptome mapping methods obtain an intragenic mapping rate of 1, simply due to the nature of the mapping procedure. When comparing the mapping locations of the reference genome methods, we see that GSNAP obtains a significantly higher intergenic mapping rate than TopHat2, ranging from 0.0856 to 0.1537, and 0.0248 to 0.0654, respectively. Conversely, TopHat2 obtains a significantly higher exonic mapping rate than GSNAP, ranging from 0.8705 to 0.9258, and 0.7023 to 0.7938, respectively. This is likely due to TopHat2's preferential mapping to the reference transcriptome prior to exploration of genomic locations.

4.3.1 Detected Transcripts

We computed the number of human transcripts detected by each method, defining a transcript as detected if at least 5 reads mapped to an exon region. We found that TopHat2 and Bowtie2 detect significantly more transcripts than GSNAP, and Stampy detects significantly more transcripts than TopHat2 and Bowtie2. There was no significant difference between number of detected transcripts between TopHat2 and Bowtie2. The number of detected transcripts range from 25,069-28,585 for GSNAP, 29,427-32,556 for TopHat, 29,631-32,817 for Bowtie2, and 35,308-39,234 for Stampy. We do not see 100 % of transcripts represented, but that is to be expected. The transcripts present in a single tissue type will not contain an organisms full repertoire of transcripts Hornett and Wheat (2012); Weber et al. (2007). There were a total of 44,312 human transcripts in our annotation set (iGenomes download). We also examined the detected genes, collapsing all splice variants, in functional groupings determined by Gene Ontology Ashburner et al. (2000) annotations. Figure 4.3 shows the percent of genes detected within each gene list.

We observe little differences in the ability of each mapping method to detect genes in different functional groups compared to the full gene set. It is worth noting that for some functional groups, the differences seem less or more pronounced. For example, Stampy's increased sensitivity seems less pronounced within immune system genes. This may be due to the nature of our samples - we might expect immune system genes to be highly expressed. Similarly, we examined the ability of each mapping method to detect genes at varying evolutionary distances. Using *Papio anubis* RefSeq genes as a surrogate for *Papio cynocephalus*, we identified human orthologs, computed Jukes-Cantor evolutionary distance, and examine the percent of genes detected within evolutionary distance strata (see methods). Figure 4.4 shows the mean and standard deviation for percent of genes detected at increasing evolutionary distances, up until

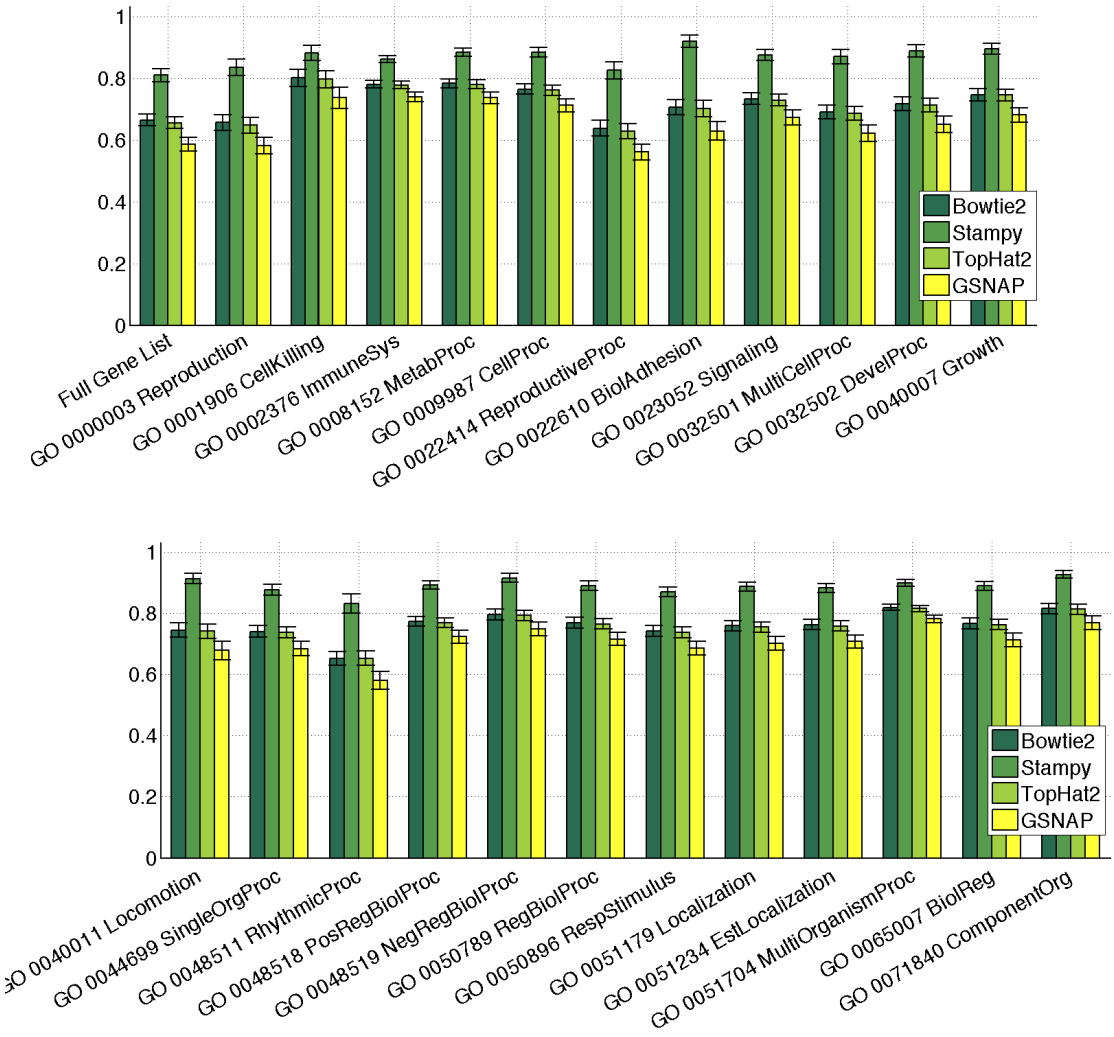

FIGURE 4.3: Detected Genes by Function - Mean and standard deviation of percent detected genes (computed as detected genes within a list divided by the number of genes within the list) for the full gene set, and 23 different Gene Ontology Biological Process groupings.

the highest evolutionary distances, where the difference in sensitivity is minimized as all methods lose the ability to detect genes.

We still see that Stampy detects the most genes, followed by Bowtie2, TopHat2, and GSNAP.

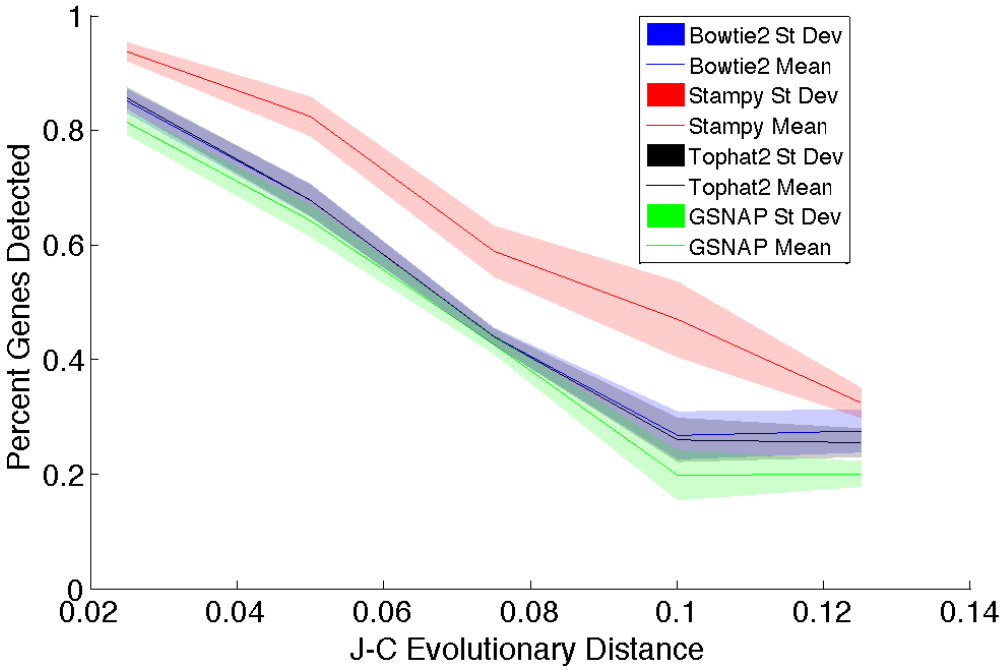

FIGURE 4.4: Detected Genes by Evolutionary Distance - Mean and standard deviation of percent detected genes at increasing evolutionary distance.

4.3.2 Correlation of Gene Expression

We examined correlations of normalized gene expression between baseline samples within the same mapping method, and correlations between normalized gene expression levels computed with results from the different mapping methods. Figure 4.5 shows a heat map of the Pearson correlations between all samples, and all methods.

We see strong correlation between samples for all four methods (the large blocks along the diagonal). Similarly, Figure 4.6 shows the correlation between biological replicates of baseline samples.

These correlations are very strong, with none falling below 0.9314, and the mean of each method above 0.97. Comparing the correlations between methods, we see that Bowtie2 obtains the highest correlation between samples, significantly higher than TopHat2, GSNAP, and Stampy. Stampy and GSNAP both obtain significantly

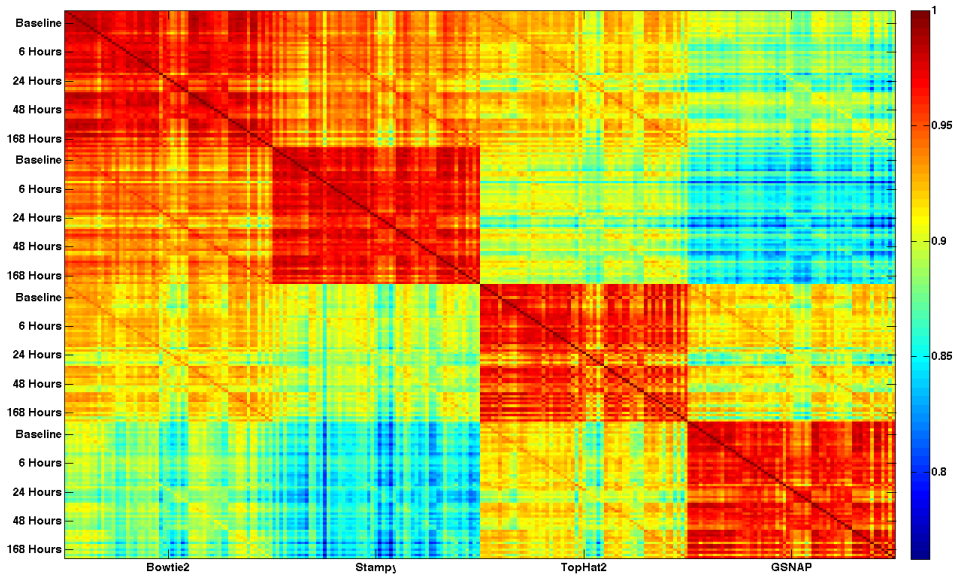

FIGURE 4.5: Correlation of Gene Expression - Heat map of all pairwise Pearson correlations between gene expression of each sample computed with each of the four mapping methods.

higher correlations than TopHat2, with no significant difference between GSNAP and Stampy correlations.

We also computed the Pearson correlation between the different methods for identical samples. This is illustrated by the blocks off the diagonal Figure 4.5, and by the boxplots in Figure 4.7.

Bowtie2 and Stampy had the highest correlations, ranging from 0.9343 to 0.9712. TopHat2 and Bowtie2 had the next highest correlations, ranging from 0.9265 to 0.9539, followed by GSNAP and TopHat2 correlations ranging from 0.9076 to 0.9585. TopHat2 and Stampy had correlations ranging from 0.8875 to 0.9320, GSNAP and Bowtie2 between 0.8749 and 0.9219, and finally GSNAP and Stampy from 0.8227 to 0.8940.

To further illustrate the concordance in gene expression estimates obtained with each mapping method, we constructed a dendrogram, computing the Euclidean dis-

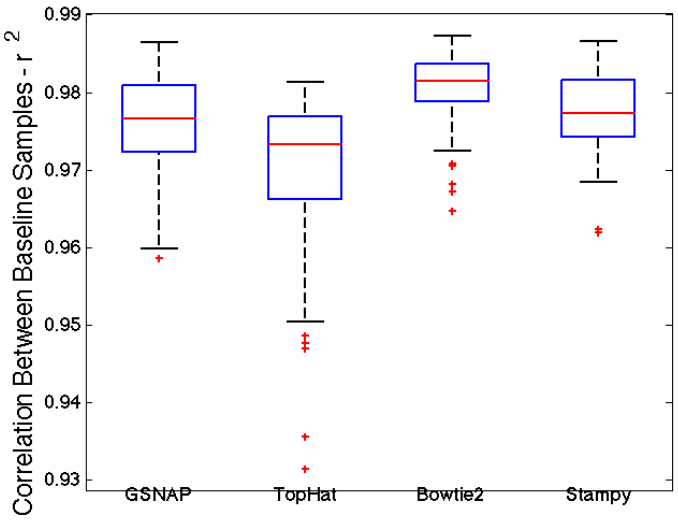

FIGURE 4.6: Correlation of Baseline Sample Gene Expression - Boxplots of the correlations between gene expression of baseline samples (0 hours), within each method.

tance in gene expression between methods for each of the 57 samples, and then averaging the distances. Examining the dendrogram in Figure 4.8, we observe Bowtie2 and Stampy have the shortest distance between gene expression estimates, followed by TopHat2, and then GSNAP. These results are in accordance with the correlations shown above. The weakest correlations are seen between the most and least sensitive methods, Stampy and GSNAP, respectively. This difference is less pronounced for Bowtie2 and Stampy, the two most sensitive mapping methods. For the less sensitive methods, it is likely that reads from divergent regions of transcripts are not successfully mapped, affecting expression estimates.

4.3.3 Differential Expression Analysis

We used edgeR to identify differentially expressed genes between healthy and sick participants from the expression estimates computed with each mapping method. Sick participants were considered to be animals meeting clinical criteria for bacterial pneumonia infection, at the time of maximal symptoms (48 hours). We define healthy

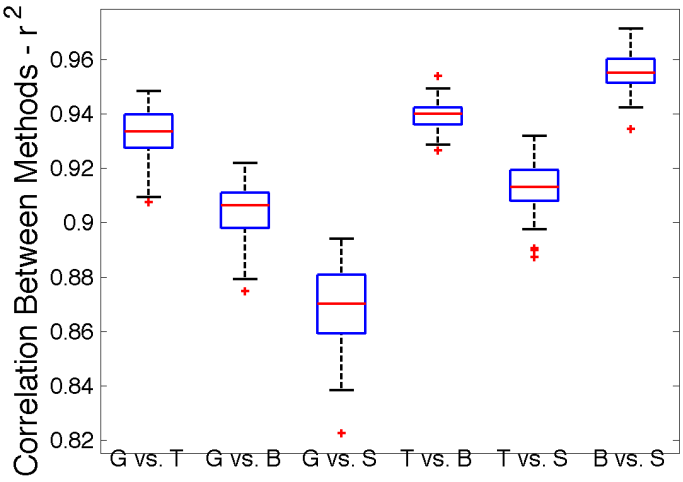

FIGURE 4.7: Correlation of Gene Expression Between Methods -Boxplots of the correlations between gene expression of identical samples between methods.

partipicants as all baseline samples, as well as the control doses (0 CFU) at 48 hours. Figure 4.9 shows a 4-way venn diagram of the significant differentially expressed genes after bonferonni correction, resulting from the four mapping methods.

There are a large number of differentially expressed (DE) genes shared by all 4 methods. Most of the other DE genes are shared between 3 methods, GSNAP, Bowtie2 and Stampy. TopHat2 results in the fewest DE genes. Bowtie2 and Stampy yield many additional differentially expressed genes, which is not surprising given their increased sensitivity. We utilized the gene expression data from each method to build a classifier of bacterial pneumonia status, using the groups described above. Figure 4.10 shows the LOOCV results of the classifiers built on gene expression data from each method.

We see that in all four mapping methods, the gene expression data is sufficient to build an accurate classifier, obtaining perfect LOOCV classification. The ability to detect differential gene expression and build predictive models will be less dependent on mapping method, as the ability to map reads with a single method will be constant

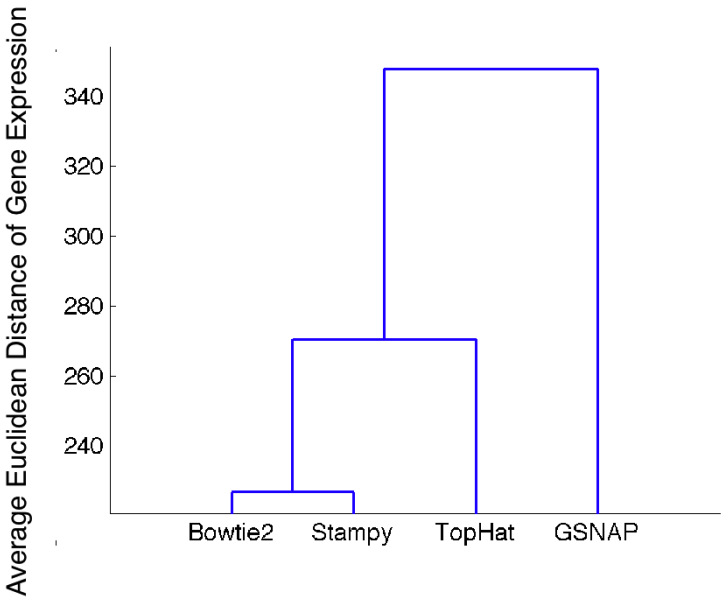

FIGURE 4.8: Dendrogram of Gene Expression -Average dendrogram of gene expression, computed with the average Euclidean distance between gene expression estimates for each sample.

across samples. More sensitive methods, however, may identify more differentially expressed genes simply because these genes are detected by these methods, and not others.

4.3.4 Read Counts by Evolutionary Distance

Using the orthologous genes stratified by evolutionary distance described above, we compared the number of reads mapping to homologous human genes of varying evolutionary distance by each of the four mapping methods. Figure 4.11 shows read count comparisons between methods for 453 orthologous genes, colored by Jukes-Cantor evolutionary distance.

Points above the diagonal indicate a higher read count from the mapping method indicated on the Y-axis, while points below the diagonal indicate a higher read count from the mapping method indicated on the X-axis. Most notable is the read count differences between GSNAP with the others. All three methods obtain higher read

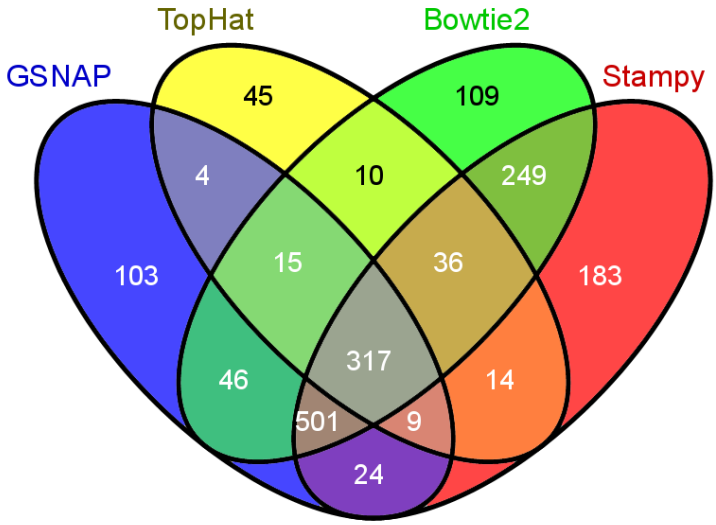

FIGURE 4.9: Shared Differentially Expressed Genes - Venn diagram of the number of differentially expressed genes found using each of the four mapping methods.

counts for all evolutionary distances. As with the differences in mapping rates, these differences are likely due to the constraints GSNAP places on alignments - the requirement of 14 base identical stretches, and the allowance of a single gap or splice site. We see strong concordance in read counts between Bowtie2 and Stampy, with Bowtie2 obtaining slightly more read counts for more conserved genes, and Stampy obtaining slightly higher read counts for less conserved genes. TopHat2 and Bowtie2 also show strong concordance, with fairly even "spread" of read count changes. TopHat2 and Stampy are similar, with Stampy obtaining slightly higher gene counts for the most divergent genes. These results are in accordance with the mapping rate differences.

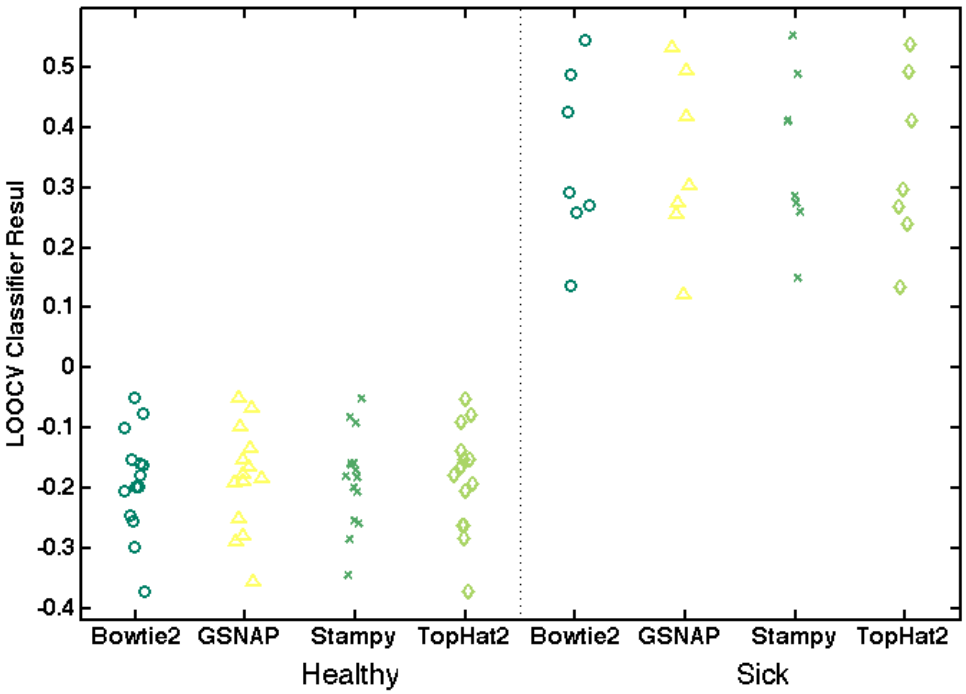

FIGURE 4.10: Predictive Utility - Leave One Out Cross-Validation results of the Top K, Elastic Net classifiers built on the gene expression data from the four mapping methods.

4.4 Conclusions

We present a comparison of reference-based mapping methods for mapping Non-Human Primate RNA-Seq data to a human reference. Four different mapping approaches were assessed using mapping rates, mapping locations, detected transcripts, correlation of gene expression, differential expression analysis, and predictive utility. Table 4.3 shows a summary of our comparison of the four reference-based mapping methods, when default parameter settings are used.

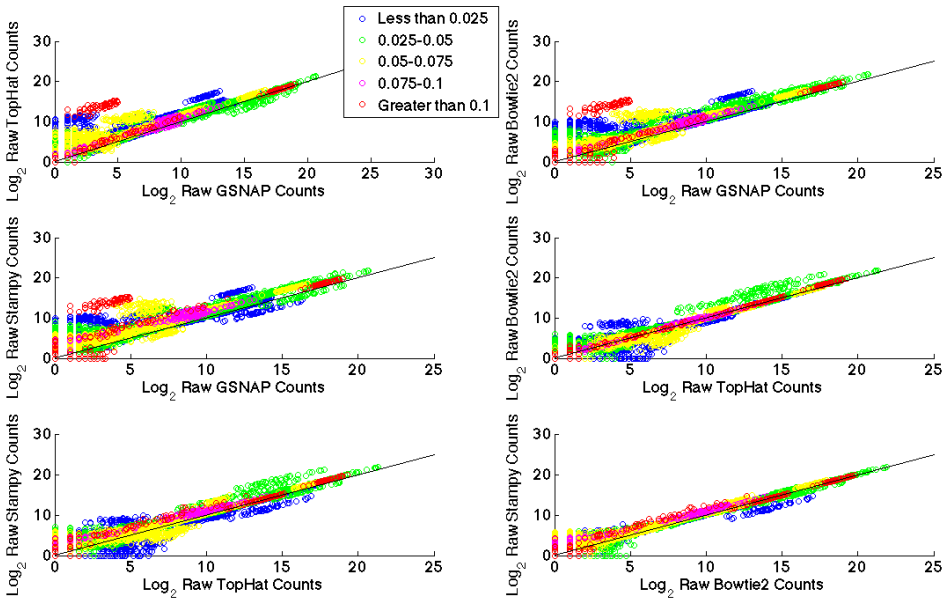

FIGURE 4.11: Read Count Comparison by Evolutionary Distance - Figure 4.11 compares the number of reads assigned to genes by each of the four mapping methods, stratified by evolutionary distance. Each panel shows a pairwise comparison of read counts between two methods. Each point indicates a particular gene in a single sample, the log_2 raw read count in two methods. Points above the diagonal indicate higher read counts in the Y-axis method, while points below indicate higher read counts in the X-axis method.

Table 4.3: Summary of Comparison Results - Summary of the comparison results for the four reference-based mapping methods examined in this study. ○- Good, ◐- Better, ●- Best.

Method	**Mapping Rate**	**Intergenic Mapping Rate**	**Detected Transcripts**	**Detected Divergent Transcripts**	**Gene Expression** r^2	**Predictive Utility**
Transcriptome Mapping with Bowtie2	◐	●	◐	●	●	●
Transcriptome Mapping with Stampy	●	●	●	●	◐	●
Genome Mapping with TopHat2	◐	◐	◐	●	○	●
Genome Mapping with GSNAP	○	○	◐	◐	◐	●

We show that when aligning RNA-Seq reads to a surrogate reference, it is important to consider the intricate details of the methods used. We have found that the use of shorter seeds, allowance of mismatches within seeds, and the allowance for alignment gaps in addition to splice junctions are essential for sensitive read mapping. Specifically, a seed length of approximately 15bp, allowing a single mismatch within seeds, and allowing for at least 2 gaps and/or splice junctions in spliced alignments will facilitate effective read mapping. When utilizing the default behaviors of the methods compared here for mapping NHP data to a human reference, we recommend Stampy for maximum sensitivity within known genes, and TopHat2 if the detection of novel (or non-human) transcripts is desired. However, it should be noted that all methods have adjustable alignment parameters, and in most cases could be optimized for more sensitive alignment. With the proper parameter settings, each may achieve suitable sensitivity for reference-based mapping of NHP species. We also point out that a major limitation of reference-based mapping of NHP data to a human reference will not identify all genes that may be implicated in a particular phenotype. This will likely depend on the evolutionary distance between species, and on the conservation of the genes of interest. However, for exploratory purposes and hypothesis generation, reference-based mapping of NHP data may have great utility.

4.5 Acknowledgements

We would like to acknowledge the support of the NIH CTSA (Clinical and Translational Science Award) 1UL1RR024128-01, and the Defense Advanced Research Projects Agency (DARPA), number lN66001-07-C-0092 (G.S.G.).

5

Proteomics Alignment Model

5.1 Background

The goal of many proteomics experiments is to determine the abundance of proteins in biological samples, and the variation thereof in various physiological conditions. High-throughput quantitative proteomics, specifically label-free LC-MS/MS, allows rapid measurement of thousands of proteins, enabling large-scale studies of various biological systems. Prior to analyzing these information-rich datasets, raw data must undergo several computational processing steps. We present a method to address one of the essential steps in proteomics data processing - the matching of peptide measurements across samples.

5.1.1 Open-Platform Proteomics

In a standard "bottom-up" proteomics experiment, proteins are first digested into peptides by a proteolytic enzyme. Peptides in this mixture are then physically separated by Chromatography, often Liquid Chromatography (LC). Eluting peptides are converted to gas phase ions, which are separated in a Mass Spectrometer (MS) by mass-to-charge ratio, and the relative abundance of each ion is measured by a

detector. LC-MS experiments utilize a single mass analyzer, resulting in a retention time, mass-to-charge ratio, and intensity for each analyte. In LC, tandem MS experiments, or LC-MS/MS, select precursor ions are further fragmented into product ions, resulting in an additional level of information for each peptide ion. The product ions are analyzed to determine a peptide sequence, which is used to identify the parent protein. A recent variation of LC-MS/MS - Data Independent Acquisition (DIA) - generates product ions for virtually every precursor ion, providing tremendous utility for quantification and identification in a single data set. Examples of DIA include SWATH Gillet et al. (2012) and MS^E SJ et al. (2009). In MS^E, precursor ions enter a collision cell, rapidly alternating between high and low kinetic energy states. This high-low switching fragmentation enables the measurement of both precursor and product ions in a single experiment. An even more recent DIA approach to bottom-up proteomics experiments - $HDMS^E$ incorporates Ion Mobility (IM) spectrometry, an additional separation of peptide ions after LC, and before MS^E. IM spectrometry separates ionized peptides based on charge and three-dimensional cross-sectional area.

5.1.2 Open-Platform Proteomics Data Processing

Several data processing steps are required to elucidate individual peptide intensities from raw label-free proteomics data. A typical data processing pipeline for a label-free proteomics experiment with multiple samples is illustrated in Figure 5.1.

Peptide peaks must be discerned from noise, charge states determined, and isotopic distributions identified and often combined into peptide features. Further details regarding current peak detection, de-isotoping, and charge state detection methods are described in Dowsey et al. Dowsey et al. (2010) and Zhang et al. Zhang et al. (2009). The LC retention times and elution order of peptides often shift between runs. Such variations in retention time are typically called warp. The process

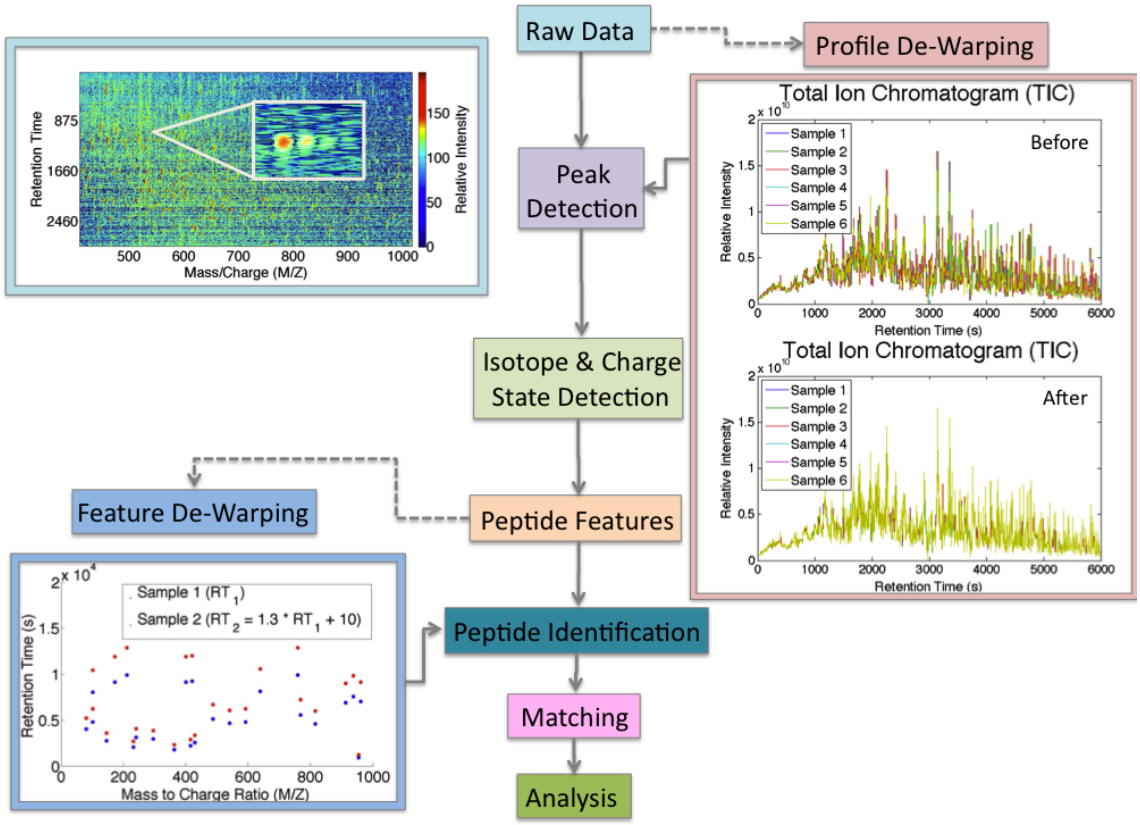

FIGURE 5.1: Processing Open-Platform Proteomics Data. Raw proteomics data requires several data processing steps, including peak detection, de-isootoping, charge state determination, collapsing peaks into peptide features, data de-warping, peptide identification, and peptide features matching.

of correcting these distortions to allow accurate matching across runs is called de-warping. Many de-warping methods exist, performing linear or non-linear (or both) corrections of two or more samples Listgarten and Emili (2005). This de-warping step is either performed on raw profile data (prior to or independent of peak detection and de-isotoping), or on feature data (detected peptide features). After generation of a peptide feature set, peptide identifications are made wherever possible, and intensity measurements of both identified and unidentified peptides are grouped across runs, creating a peptide-by-sample intensity array for subsequent analyses. It should be noted that the order of data processing steps may vary within different

pipelines. These data processing steps pose significant computational challenges, and are thought to be the source of much irreproducibility. This was illustrated by a recent test study by Bell et al. Bell et al. (2009). In the study, a sample of 20 proteins was distributed to 27 different labs, experimentally analyzed, and subjected to a variety of computationsl data-processing methods. There were significant discrepancies in reported proteins, however, all raw data was sufficient to identify all 20 proteins when centrally re-processed.

5.1.3 Label-Free Proteomics Data Alignment

We focus on the problems of simple data de-warping, and matching peptide intensities across multiple high-throughput proteomics runs - a combined processing step we call alignment. Our analysis emphasizes the data matching step. Accurate alignment is essential in large-scale proteomics experiments, particularly in biomarker discovery where the comparative nature of these studies require intensities of the same peptide to be compared across samples Jeffries (2005); Service (2008). In addition, accurate matching across samples can increase identifications as information can be leveraged from all individual runs Prince and Marcotte (2006). The complexity of biological samples, however, poses significant computational challenges for both data alignment and peptide identification. Most samples contain tens of thousands of peptides and measurements often reflect overlapping peptides, or co-eluting peptides having nearly the same mass-to-charge ratio. These overlapping peptides complicate and often prevent identification. However, recent experimental advancements provide additional separation information that has not yet been leveraged in data alignment - namely comprehensive product ion information with DIA, and IM drift times with $HDMS^E$. Product ions have been used extensively in database matching for peptide identification Vissers et al. (2007); Silva et al. (2006), but are not widely used in proteomics data alignment. Matching across samples is typically performed using

experimentally measured and inferred characteristics of each peptide feature. Measured characteristics include precursor ion retention time, mass-to-charge ratio, and intensity. Depending on the experimental methods, some or all peptide features may have additional measured characteristics including the intensities, mass-to-charge ratios, and retention times of product ions. In DIA experiments, virtually every peptide feature has measured product ion data. In $HDMS^E$ experiments, all precursor and product ions have a measured IM drift time as well. Inferred characteristics include charge state for nearly every peptide feature, and an amino acid sequence for some peptide features. Modern high-throughput proteomics experiments offer a great deal of information, not all of which is currently utilized in data processing - specifically in data alignment steps.

5.1.4 Previous Alignment Approaches

Matching methods utilize various aspects of the data to group peptide measurements across samples. Previously utilized characteristics include retention time, mass-to-charge ratio, intensity, and amino acid sequence. Incorporating additional data provides a higher degree of specificity when making matches Listgarten and Emili (2005). The majority of existing alignment techniques utilize mass-to-charge ratio and retention time information. Such methods include SpecArray Li et al. (2005), AMT tag approaches Silva et al. (2005), Xalign Zhang et al. (2005), MZmine Katajamaa et al. (2006), msInspect Bellew et al. (2006), XCMS Smith et al. (2006), PETAL Wang et al. (2007), OpenMS Lange et al. (2007); Sturm et al. (2008), apLCMS Yu et al. (2009), and MZmine2 Pluskal et al. (2010). Semi-supervised approaches such as PEPPeR Jaffe et al. (2006) and a method by Fischer et al. Fischer et al. (2006), take advantage of existing MS/MS peptide identifications. More recent alignment methods by Tang et al. Tang et al. (2011) and Zhang et al. Zhang (2012) utilize peptide intensity along with mass-to-charge ratio and retention time. SuperHirn

Mueller et al. (2007) indirectly incorporates intensity information during multiple alignments by performing pairwise alignments in a specific order based on LC-MS similarity and intensity correlations.

In addition to utilizing novel data characteristics, our method is designed to avoid common pitfalls of other approaches including static distance cutoffs, elution order assumptions, and the selection of a single reference sample. Methods should not rely on the fact that peptides always elute in the same order from the LC column as this is a false assumption and can introduce matching errors Nielsen et al. (1998); Prakash et al. (2006); Lange et al. (2008). Multiple alignment methods often require a reference run to which all other runs are aligned. While this approach is successful for the de-warping step, choosing a single reference run for the matching step can be problematic. If the measurement variability is high in the reference sample, this can result in incorrect matches. We present a novel statistical alignment method that corrects for linear global variation, is not restricted by static distance cutoffs, and has the ability to utilize retention time, mass-to-charge ratio, peptide identifications, and previously ignored aspects of proteomics data - ion mobility drift times, and product ion data. Our method is an adapted Bayesian Dirichlet Process Gaussian Mixture Model (DPGMM) West (1992); Rasmussen (2000), adding sample-specific shift and scale parameters. The proposed method is also easily extensible to incorporate additional dimensions, such as a second LC separation. We present the results of our alignment model on various datasets, comparing alignment accuracies with the inclusion of various data characteristics.

5.2 Materials and Methods

5.2.1 Alignment Model

We present a statistical model for the alignment of label-free proteomics data to match peptide features across multiple samples after peak-detection and de-isotoping.

Unlike any existing proteomics alignment method, our model has the ability to utilize ion mobility drift time from HDMS^{E} experiments, and product ion spectra from traditional LC-MS/MS Data Dependent Aquisition (DDA), or DIA (MS^{E} or HDMS^{E}) experiments, along with the typical parent ion mass-to-charge ratio and retention time - increasing the individuality of each peptide feature and providing a better alignment. At the time of publication, no open-source proteomics file format was capable of storing ion mobility separation data. In order to allow incorporation of this data into our alignment method, we wrote a small data-processing script to read Waters spectrum.xml and finalfrag.csv files into a Matlab data frame. The Matlab data frame format and the data processing script are available in the Appendix and Supplemental Files, respectively, and can be easily adapted to incorporate any additional separation dimensions similar to ion mobility drift times and liquid chromatography retention times, including retention times from multidimensional LC.

Model

We adapt a DPGMM West (1992); Rasmussen (2000) by adding sample-specific shift and scale parameters. Gaussian Mixture Models lend themselves well to the problem of proteomics data alignment. Each peptide existing in nature has a theoretical mass-to-charge ratio, retention time, etc. within a specific experimental condition, and is represented in our model as a mixture component. We expect the measurement of a peptide to have the same mass-to-charge ratio, retention time, etc., with two different types of measurement error: systematic error and random error. As with any laboratory experiment, LC-MS/MS data are subject to variability. The LC retention times often shift between runs. Pressure fluctuations, changes in column temperature, column manufacturing differences, and peptide interactions can cause changes in the elution time, and/or the elution order of peptides Silva et al. (2005). The mass-to-charge ratios are also subject to measurement error, albeit to a lesser

degree than the LC dimension. We account for systematic error with a global shift and scale. Such a transformation would most likely be the result of variations in LC protocols (total run times), or in the time it takes for the first peptide to elute from the column (gradient delays due to different tubing volumes). The remaining random error is assumed to be a sum of small variations from many independent sources of variation, and therefore have a Gaussian distribution. In addition, Gaussian distributions are closed under linear transformations, allowing straightforward computation of posterior distributions with the addition of the shift and scale parameters. Measurements assigned to the same mixture component, or latent peptide, by the model are considered to be matched. Seed peptide matches are determined with identified peptide sequences and charge states. To avoid introducing error with incorrect identifications, outliers with respect to mass-to-charge ratio and retention time are discarded. These matches are used to initialize hyperparameters, and remain matched at all iterations of the MCMC. Mixture component assignments are given a Chinese Restaurant Process prior, allowing the addition of a new latent peptide if no suitable match exists. Our model addresses simple linear de-warping of the data, however, any preferred de-warping method may be applied prior to utilizing our algorithm. We first describe the model for peptide-level alignment, and then describe the extension to include product ions.

Peptide-Level Model A sample-specific linear shift (η_d) and scale (β_d) is used for de-warping. In the formulas that follow, samples or datasets are indexed by d, individual measured peptides within a sample are indexed by i, and latent (theoretical) peptides are indexed by j. As shown in equation 5.1, we assume that a measured peptide feature, $x_{d,i}$, is a shifted and scaled noisy measurement of a true peptide feature, $z_{c_{d,i}}$. Let $c_{d,i}$ be an indicator variable for the latent peptide assignment of measurement

$x_{d,i}$, taking on values $j = 1 \ldots J$ where J is unbounded.

$$x_{d,i} = \eta_d + z_{c_{d,i}}\beta_d + \epsilon_{d,i} \tag{5.1}$$

$$\epsilon_{d,i} \sim Normal(0, \Sigma) \tag{5.2}$$

$$z_{c_{d,i}} \sim Normal(\mu_{c_{d,i}}, \sigma) \tag{5.3}$$

Where $\epsilon_{d,i}$ are the residuals between the measured values ($x_{d,i}$) and the shifted and scaled latent values ($\eta_d + z_{c_{d,i}}\beta_d$), having a multivariate normal distribution - equation 5.2. Let Σ be the covariance of these residuals, and let each latent peptide $z_{c_{d,i}}$ be a draw from DPGMM mixture component $c_{d,i} = 1 \ldots J$ having mean $\mu_{c_{d,i}}$ and covariance σ, as shown in equation 5.3. Note that we make the simplifying assumption of shared covariance across all latent peptides. This would suggest that the measured values of each latent peptide show the same variation across the entire mass-to-charge ratio, retention time, and drift time range. We acknowledge that this is not likely the case, although we find this assumption works well in practice. Conjugate priors are used for all model parameters as follows:

$$\eta_d \sim Normal(a_d, b_d) \tag{5.4}$$

$$\beta_d \sim Normal(e_d, f_d) \tag{5.5}$$

$$\Sigma \sim Inverse - Wishart(s, t) \tag{5.6}$$

$$\mu_j \sim Normal(\lambda, r) \tag{5.7}$$

$$\sigma \sim Inverse - Wishart(g, h) \tag{5.8}$$

Normal priors are assigned to the shift and scale parameters as shown in equations 5.4 and 5.5. The seed matches, for each peptide feature are averaged to generate a list of implied-identified peptides. Robust fit linear regression is performed for each dataset using the implied-identified peptides as predictors, and the measured identified peptides as response. The resulting intercept is taken as the mean hyper-parameter in the shift prior distribution (a_d). Similarly, the coefficient is taken as

the mean hyperparameter in the scale prior distribution (e_d). The variance parameters on the shift and scale priors (b_d and f_d) are set tightly to the variance of the regression estimate. This allows an optimal solution to be reached as latent peptides are updated and added, while reducing shift and scale identifiability issues. Both the match covariance (Σ), and latent peptide covariance (σ) matrices are given conjugate inverse-Wishart priors as shown in equations A.2 and 5.8. The residuals of the shifted and scaled identified peptide measurements, and their respective implied-identified peptides are used set hyperparameters. The degrees of freedom parameters (h and t) are set to the number of identified matches minus one, and the inverse-scale matrix is set to the sum of squared residuals. The mean of each latent peptide (μ_j) is given a conjugate normal prior with as shown in equation 5.7. The prior mean (λ) is set to the empirical mean of all measured peptide features in all datasets, and the prior covariance (r) is set to the sum of squared differences between this empirical mean and all measured data. We express the likelihood of $x_{d,i}$ as follows:

$$P(x_{d,i} \mid \quad \eta, \beta, \Sigma, z_1 ... z_J, c_{d,i}) = \quad Normal(x_{d,i} \mid \eta_d + z_{c_{d,i}} \beta_d, \Sigma) \tag{5.9}$$

Where $z_1 ... z_J$ indicates all existing latent peptides. We may also integrate out $z_{c_{d,i}}$ and re-express the likelihood as:

$$\begin{aligned} P(x_{d,i} \mid \quad \eta, \beta, \Sigma, \mu_1 ... \mu_J, \sigma, c_{d,i}) = \quad & Normal(x_{d,i} \mid A_{c_{d,i}}, B_d) \\ A_{c_{d,i}} \quad & = \eta_d + \mu_{c_{d,i}} \beta_d \\ B_d \quad & = \beta_d^T \sigma \beta_d + \Sigma \end{aligned} \tag{5.10}$$

The prior probability of an observation, $x_{d,i}$, being assigned to latent peptide component j given all other assignment indicators, $c_{-(d,i)}$, is given in equation 5.11. The notation $c_{-(d,i)}$ refers to all component indicators from all features in all datasets, except d, i. Similarly the prior probability of observation $x_{d,i}$ being assigned to a new

latent peptide component is shown in equation 5.12.

$$p(c_{d,i} = j \mid c_{-(d,i)}, \alpha) = \frac{n_{-(d,i),j}}{N-1+\alpha} \times I(!\exists i, c_{d,-i} = c_{d,i}) \quad (5.11)$$

$$p(c_{d,i} \neq c_{-(d,i)} \mid c_{-(d,i)}, \alpha) = \frac{\alpha}{N-1+\alpha} \quad (5.12)$$

Let α be the DPGMM concentration parameter, N the total number of observed peptide features across all samples, and $n_{-(d,i)}$ the number of observed peptide features other than $x_{d,i}$ assigned to latent peptide j. A constraint is imposed such that only one measurement per dataset may be assigned to a given latent peptide feature. The concentration parameter of the Dirichlet Process (α) is set to the number of peptide feature observations across all samples being aligned. Latent peptide feature assignments are updated from their full conditional posterior distributions, as shown in equation 5.13.

$$\begin{aligned} P(c_{d,i} = j \mid -) \propto & \frac{\alpha}{N-1+\alpha} \times \int P(x_{d,i} \mid -) \times P(\mu_j \mid \lambda, r) d\mu_j \\ & + \sum_j \frac{n_{-(d,i),j}}{N-1+\alpha} Normal(x_{d,i} \mid A_{c_{d,i}}, B_d) \times I(!\exists i, c_{d,-i} = c_{d,i}) \end{aligned} \quad (5.13)$$

The integral above is tractable due to the shared covariance across latent peptide components. Further details and full conditional distributions are available in the Appendix. Figure 5.2 shows a plate diagram of the peptide-level alignment model.

Product Ion Model Extension To incorporate product ion data, we select up to the 50 most intense product ions for each peptide feature measurement, $x_{d,i}$. We then generate a K-dimensional product ion intensity profile for each $x_{d,i}$. Each position, y_{d,i_k}, in the product ion intensity profile, $y_{d,i}$, is computed as:

$$y_{d,i_k} = \frac{\sum_p \Omega_p \times I(M_p \leqslant B_k)}{\sum_p \Omega_p} \quad (5.14)$$

where $k = 1 \ldots K$, $p = 1 \ldots 50$, Ω is a 50-dimensional vector of intensities, M is a 50-dimensional vector of product ion mass-to-charge ratios, and B is a K-dimensional

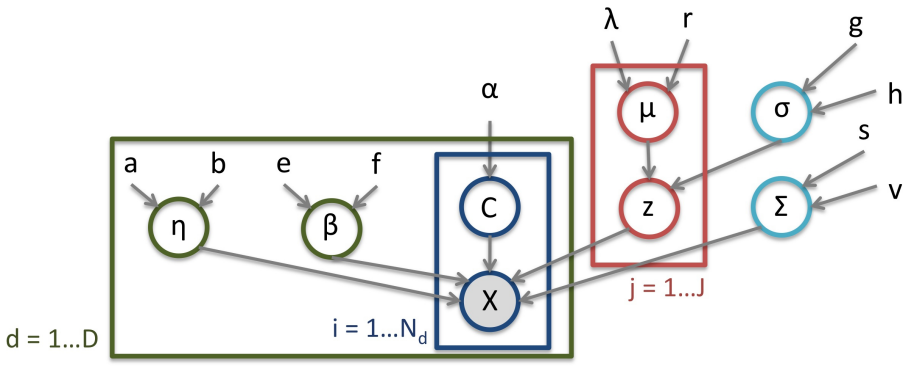

FIGURE 5.2: Alignment Model. A Plate Diagram of our Peptide-Level Model. We adapt a Dirichlet Process Gaussian Mixture Model to address open-platform proteomics data alignment.

vector of product ion profile mass-to-charge ratio bin upper limits. All values in $y_{d,i}$ sum to one. The mass-to-charge ratio ranges, or bins (B_k), are determined at the initialization of the alignment, such that bin boundaries fall on mass-to-charge ratio deserts. In open-platform proteomics data, there are subsets of the mass-to-charge ratio dimension not occupied by any peptide features. We term these subsets mass-to-charge ratio deserts and utilize them to split the data for parallelization as well as building product ion profiles. The deserts are empirically determined using all datasets in the alignment. Figure 5.3 shows a histogram of measured m/z values for E. coli lysate data, the empirical m/z deserts are indicated with vertical red lines.

These deserts are also used to build product ion profiles. The boundaries of the product ion profile bins are placed within the m/z deserts, and each product ion profile position is set to the sum of all intensities of product ions belonging to that bin. A small example using a profile size of K = 10 is shown below the m/z desert histogram. See the Parallelization section and for more details on bin boundary determination.

In the experiments described here, we set K = 250 to ensure most product ions

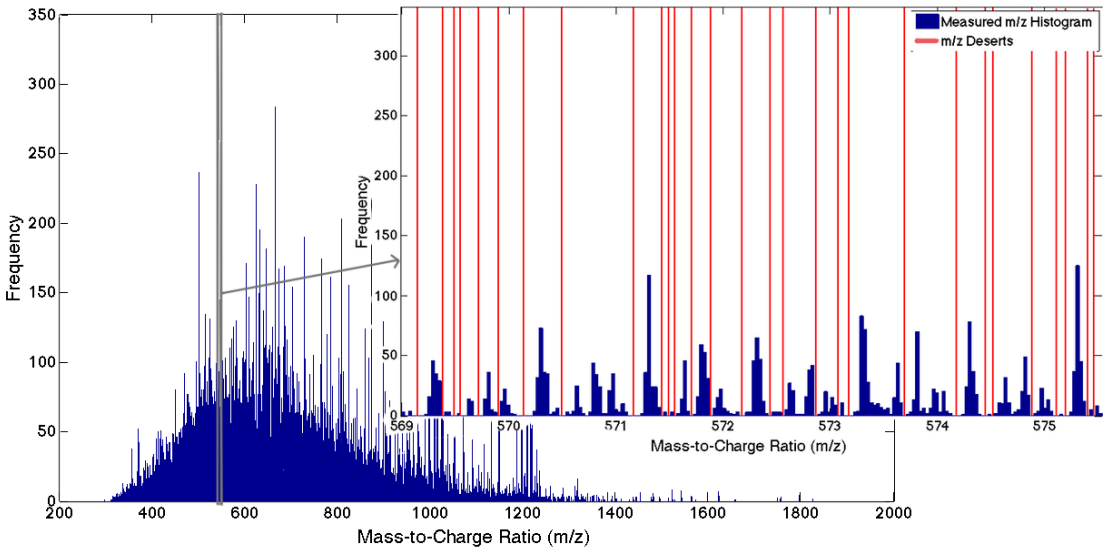

FIGURE 5.3: Mass-to-Charge Ratio Deserts. Mass-to-Charge Ratio Deserts are utilized to split the data for parallelization, and to build product ion profiles.

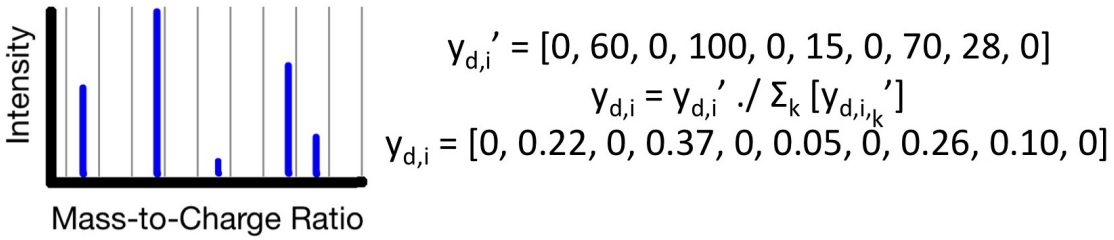

FIGURE 5.4: Sample Product Ion Profile. Product ion profiles are contructed by summing the intensity of product ions in mass-to-charge ratio bins, and then normalizing by the total intensity of product ions assigned to a given peptide.

would be assigned to their own bin in the product ion profile, and to avoid additional computational complexity. Each existing latent peptide feature is given a K-dimensional product ion profile (w_j for latent peptide feature z_j). To assess the similarity of a measured product ion profile, $y_{d,i}$, and a latent product ion profile, $w_{c_{d,i}}$, we introduce a similarity score, ψ, which is computed as the sum of squared differences of the two product ion intensity profiles, and is assumed to have an exponential distribution to encourage distances close to zero, as shown in equations 5.15

and 5.16.

$$\psi_{d,i} = (y_{di} - w_{c_{d,i}})^T (y_{d,i} - w_{c_{d,i}}) \tag{5.15}$$

$$\psi_{d,i} \sim Exponential(\gamma) \tag{5.16}$$

We assign a conjugate gamma prior to the rate parameter, as shown in equation 5.17. The hyperparameters for profile scores are set to one.

$$\gamma \sim Gamma(a_0, b_0) \tag{5.17}$$

At each iteration of the MCMC, the product ion profile, w_j of an existing latent peptide is updated empirically the product ion profile is set to the average of the measured product profiles assigned to that latent peptide. The latent product ion profile, w_0, of a new latent peptide (one that currently does not exist) is a blank profile - a uniform vector of size K with each element having value 1/K. Combining the product ion model with the peptide-level model, we have the following likelihood (equation 5.18) and conditional posterior (equation 5.19):

$$P(x_{d,i}, y_{d,i} \mid -) = Normal(x_{d,i} \mid A_{c_{d,i}}, B_d) \times Exponential(\psi_{d,i} \mid \gamma) \tag{5.18}$$

$$\begin{aligned} P(c_{d,i} = j \mid -) &\propto \tfrac{\alpha}{N-1+\alpha} \times \textstyle\int P(x_{d,i} \mid -) \times P(\mu_j \mid \lambda, r) d\mu_j \\ &\quad \times Exponential((y_{di} - w_0)^T (y_{d,i} - w_0) \mid \gamma) \\ &+ \textstyle\sum_j \tfrac{n_{-(d,i),j}}{N-1+\alpha} \times Normal(x_{d,i} \mid A_{c_{d,i}}, B_d) \times I(!\exists i, c_{d,-i} = c_{d,i}) \\ &\quad \times Exponential(\psi_{d,i} \mid \gamma) \end{aligned} \tag{5.19}$$

Further details and full conditional distributions are available in the Appendix. We explored additional values of K, as well as implementations of different high energy models, the results and discussions of which can also be found in the Appendix.

Model Fitting

Posterior Match Probabilities As our primary goal is obtaining a list of matches, we are only interested in maximum a posteriori (MAP) estimates of the parameters. We employ simulated annealing on all parameters after the initial burn-in period of the MCMC. In addition, with the exception of the latent peptide means, the model parameters are being updated with a large number of observations, and will have fairly tight posterior distributions. Our model assumes that the product ion match likelihoods are independent from the peptide-level match likelihoods. New peptide match indicators are sampled from the full conditionals for each measured peptide in a random order. We sample match indicators in accordance with Algorithm 2 for sampling mixture component indicators in DPGMM from Neal, 2000. We impose a restriction on match assignment such that only one measured peptide per dataset may be assigned to a given latent peptide. All other parameters are sampled from their full conditional distributions. After obtaining MAP estimates for all parameters, we then iteratively re-sample only the component assignment indicators, keeping track of how often each measurement is assigned to each latent peptide. These assignment proportions are used to make final matches.

Estimating the Best Alignment Utilizing the assignment proportions from the final assignment-only MCMC iterations, we use a greedy algorithm to determine the final alignment. The best match (latent peptide-measurement pair) across the entire alignment is selected, and then the assignment proportions for measurements in each of the remaining datasets are examined for the current latent peptide. For each dataset, the measurement with the maximum assignment proportion is selected. All remaining match probabilities for the assigned measurements and the current latent peptide are set to zero. This process is repeated until no non-zero assignment

proportions remain. These assignment probabilities represent the probability that a given measurement arises from a certain latent peptide. To compute the match probability of two measurements from two datasets, we compute the probability that both measurements are assigned to the same latent peptide the product of the two individual latent peptide assignment probabilities. Users may utilize these match probabilities to examine matches of varying confidence. An illustration of the algorithm steps is shown in Figure 5.5.

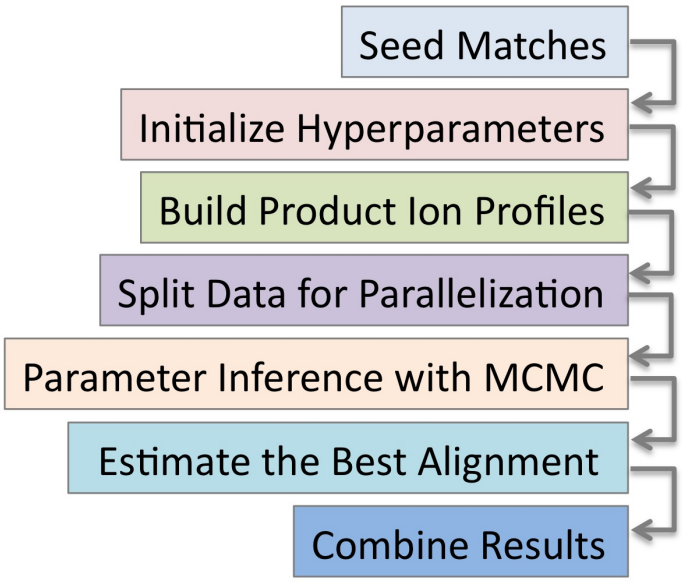

FIGURE 5.5: Algorithm Overview. This figure illustrates an overview and the order of the alignment algorithm steps.

Seed matches are determined using peptide identifications and charge states from each dataset. These seed matches are used to initialize hyperparameters for prior distributions. A product ion profile is then computed for each peptide measurement, and data is split for parallelization of the alignment. For each split in the data, MAP estimates of all parameters are obtained with Gibbs Sampling and the best alignment is approximated with assignment proportions and a greedy algorithm. Finally, the

results of each split are combined to report matches for the entire dataset.

Parallelization In order to make alignment of large datasets tractable, we split the datasets being aligned in the mass-to-charge ratio dimension, and perform separate alignments of each split in parallel. The boundaries of these splits fall only on mass-to-charge ratio deserts, which are empirically determined using all datasets being aligned. There exist gaps - also called "forbidden zones" in the mass distribution of all possible tryptic peptides AV et al. (2011). These gaps have been utilized to improve peak de-noising techniques Mitra et al. (2012). We calculate these gaps empirically based on the data in question, and utilize them to split the data for alignment. Such mass-to-charge ratio deserts are shown in Figure 5.3. When determining these mass-to-charge ratio deserts, we utilize the given matches to determine an approximate shift, scale, and match standard deviation. We then obtain the number of measured peptides in each mass-to-charge ratio bin the size of match standard deviation, and split the datasets at mass-to-charge ratio deserts defined as stretches of five or more empty bins. This ensures that any measured peptide features with the potential for being aligned to the same latent peptide feature (any measurements that should match one another) will be in the same alignment split. The hyperparameters set in the model initialization are shared across all alignment splits. The boundaries of the product ion profiles are determined in a similar way. Utilizing the product ion annotations of the given matches, we obtain a mass-to-charge ratio match standard deviation of product ions. We then obtain the number of measured product ions in each mass-to-charge ratio bin the size of the match standard deviation, and set the product ion profile boundaries on mass-to-charge ratio deserts defined as stretches of three or more empty bins. The size of the product ion profile boundaries is as close to 1/K of the spanning product ion mass-to-charge ratio range as possible, given the boundaries are set within mass-to-charge ratio deserts.

5.2.2 Data

All data used in this analysis was obtained under MS^E and/or $HDMS^E$ conditions (SYNAPT HDMS G2, Waters), and subject to Waters ProteinLynx Global SERVER (PLGS) processing. We utilize peptide features that have already been subject to peak detection, de-isotoping, charge state determination, and tentative identification (although not all identifications are utilized for alignment). All samples were separated by 1D nanoscale capillary ultraperformance Liquid Chromatography in a 90-minute gradient using a 5-40% acetonitrile/water (0.1% formic acid in each).

Three HCV cohorts were utilized in the alignment of serum samples from HCV patients to urine samples from OA patients. The first cohort included 47 patients ages 5 to 18 years from a clinical trial for HCV treatmentSchwarz et al. (2011). The two additional HCV cohorts (n = 41,55) were selected from the Duke Hepatology Clinical Research (DHCR) database Patel et al. (2011). The pediatric clinical trial study was approved by the institutional review boards of the participating sites. Written informed consent was provided by all parents or guardians, and written assent was provided by all participants over 12 years of age. All patients present in the DHCR database cohorts, as well as all OA patients, provided written informed consent, and all study procedures were approved by the Duke University Institutional Review Board.

5.2.3 Analysis

All alignments were performed using Matlab on the Duke Shared Computing Resource, a cluster of Intel x86 compute notes running Linux. Each alignment was partitioned into a maximum of 250 splits, and each alignment partition was run on a single node with at least 8GB of memory. Our method does require considerable computation time - approximately 2 to 6 hours for the *E. coli* Lysate and HCV-OA alignments not utilizing product ions, and approximately 20-24 hours for the *E. coli*

Lysate alignments utilizing product ions. Times vary by the number and size of datasets being aligned.

5.3 Results

5.3.1 E. coli Lysate Data

To assess both the performance of our alignment method, and the utility of various data characteristics, we aligned technical replicates of *E. coli* lysate data. Three technical replicates of 500ng of *E. coli* lysate were analyzed with Waters MS^E and $HDMS^E$. The MS^E data was used to compare our alignment method with PEPPeR Jaffe et al. (2006), and the feature matching functionality of OpenMS Lange et al. (2007); Sturm et al. (2008). The PEPPeR PeakMatch module was downloaded from GenePattern Reich et al. (2006) and run locally using default parameter settings. Similarly, the MapAlignerPoseClustering, and FeatureLinkerUnlabeled functions of OpenMS version 1.11.0 were run using default parameter settings to align peptide features of MS^E data. The $HDMS^E$ data was used to compare the results of our alignment method when utilizing various data characteristics. Table 5.1 shows the four different combinations of data characteristics utilized for alignment. We assess alignment performance using held-out peptide identifications resulting from product ion spectra.

Table 5.1: Alignments and Data Utilization. The alignment type names and data utilized that were compared in the analysis.

Alignment Name	Parent Ion Monoisotopic Mass-to-Charge Ratio	Parent Ion Peak Centroid Time	Parent Ion Drift Time	Product Ion Profile
MZ-RT	✓	✓		
MZ-IM	✓		✓	
MZ-RT-IM	✓	✓	✓	
MZ-RT-IM-HE	✓	✓	✓	✓

The MS^{E} *E. coli* lysate data was aligned using PEPPeR, OpenMS, and our alignment method, utilizing precursor ion mass-to-charge ratio and retention time (MZ-RT). Each alignment method was provided with the same 15 given matches to initialize model hyperparameters the 15 identifications shared by all three replicates having the highest average ProteinLynx Global SERVER (PLGS) Identity^{E} Li et al. (2009) peptide score. When assessing alignment performance, we examine matches between each replicate pair ID0821901 vs. ID0821902, ID0821901 vs. ID0821903, and ID0821902 vs. ID0821903. Correct matches occur when two identified peptides shared between the pair of technical replicates are aligned. Mismatches occur when two identified peptides with conflicting identifications are aligned. Since PEPPeR allows multiple peptides from a single sample to be present in a "cluster", there may be more than one identification from a single replicate. In these cases, we counted a correct match if shared identifications were present in the same cluster. Similarly, we counted a mismatch if conflicting identifications were present in the same cluster. As there are varying levels of confidence in peptide identifications, we present the results of each alignment method considering identifications having a PLGS Identity^{E} Li et al. (2009) peptide score of five, six, and seven or greater. Figure 5.6 shows the recall rate for each alignment method.

Similarly, Figure 5.7 shows the mismatch rate for each alignment method.

As PEPPeR does not directly report match confidence, recall and mismatch rates were computed from all reported matches for each method. We see that our alignment method obtains significantly higher recall rates than OpenMS FeatureLinker and PEPPeR, for identified peptides of each confidence level. When comparing the mismatch rates, computed as the number of mismatches divided by the total matches, we see that our method obtains mismatch rates comparable with OpenMS, while PEPPeR obtains significantly higher mismatch rates, particularly at lower peptide scores. It should be noted that this mismatch rate is not a false positive rate.

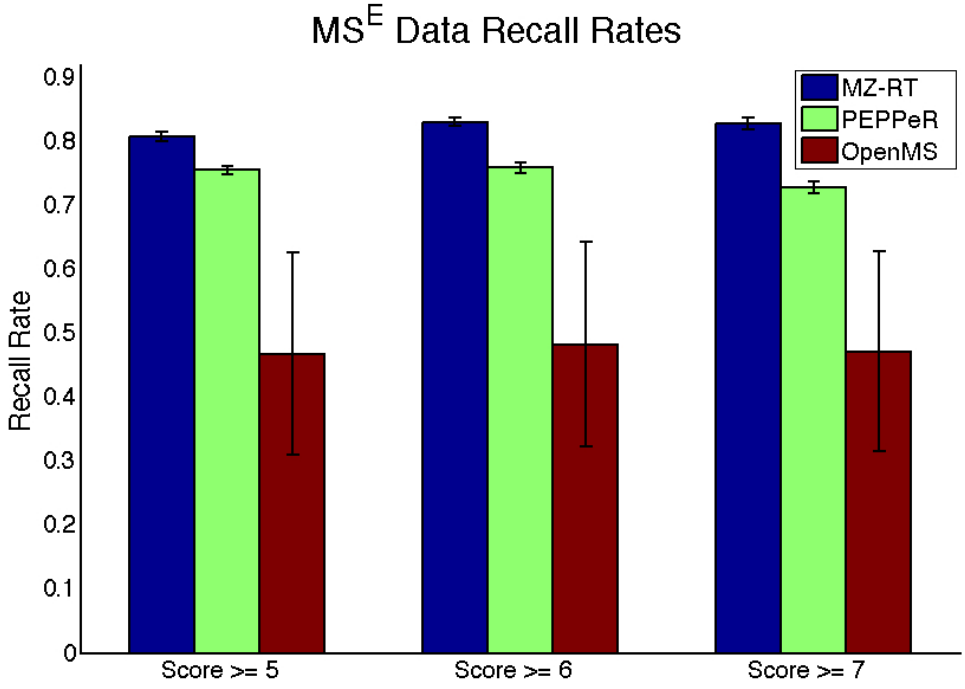

FIGURE 5.6: MS^E Alignment Recall Rates. This figure shows the recall rates considering identifications having peptide score 5, 6, and 7 or greater for our MZ-RT method, PEPPeR, and OpenMS.

The total match count includes many matches which cannot be identified as correct or incorrect, as neither peptide has a putative peptide sequence. When examining the total match counts in Figure 5.8, we see that our method obtains match counts comparable to PEPPeR, and significantly more matches than OpenMS.

The $HDMS^E$ *E. coli* lysate data was aligned using different data characteristics to compare their utility for alignment. As with the MS^E data, each alignment was provided with the same 15 given matches to initialize model hyperparameters and the remaining identifications were used to assess alignment performance. We present the results of each alignment only considering identifications having a PLGS $Identity^E$ Li et al. (2009) peptide score of five or greater. Results at additional peptide score thresholds are provided in the Appendix. Comparing the $HDMS^E$ alignments with various data combinations informs about the utility of each dimension in data align-

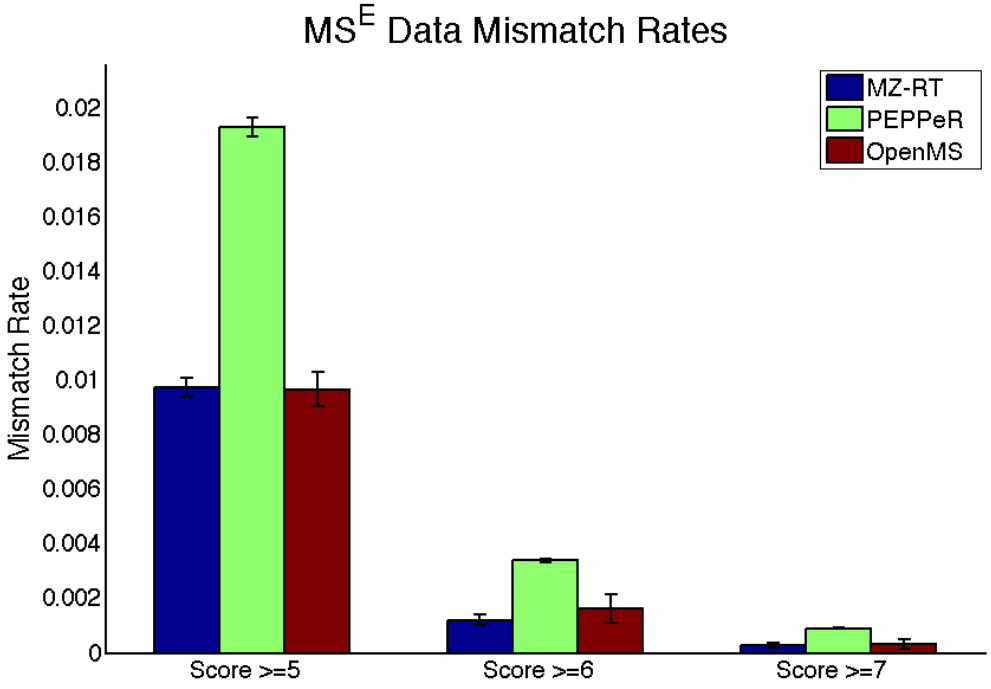

FIGURE 5.7: MS^E Alignment Mismatch Rates. This figure shows the mismatch rates considering identifications having peptide score 5, 6, and 7 or greater for our MZ-RT method, PEPPeR, and OpenMS. The mismatch rate is computed as the number of mismatches (pairwise match with conflicting identifications) divided by the total matches.

ment. We examine matches between each replicate pair ID0822001 vs. ID0822002, ID0822001 vs. ID0822003, and ID0822002 vs. ID0822003. A match across two samples occurs when a measured peptide feature from each sample is assigned to the same latent peptide feature. Figures 5.9 and 5.10 show the recall rate and the number of mismatches, respectively, for each of the four alignments.

When examining the results of the two-dimensional parent ion alignments, MZ-RT and MZ-IM, we see that the MZ-RT alignment obtains significantly higher recall rates for matches of all confidence levels. Utilizing all parent ion data characteristics obtained via $HDMS^E$ with a three-dimensional alignment (MZ-RT-IM) results in significantly higher recall rates than the two-dimensional alignments, with a small

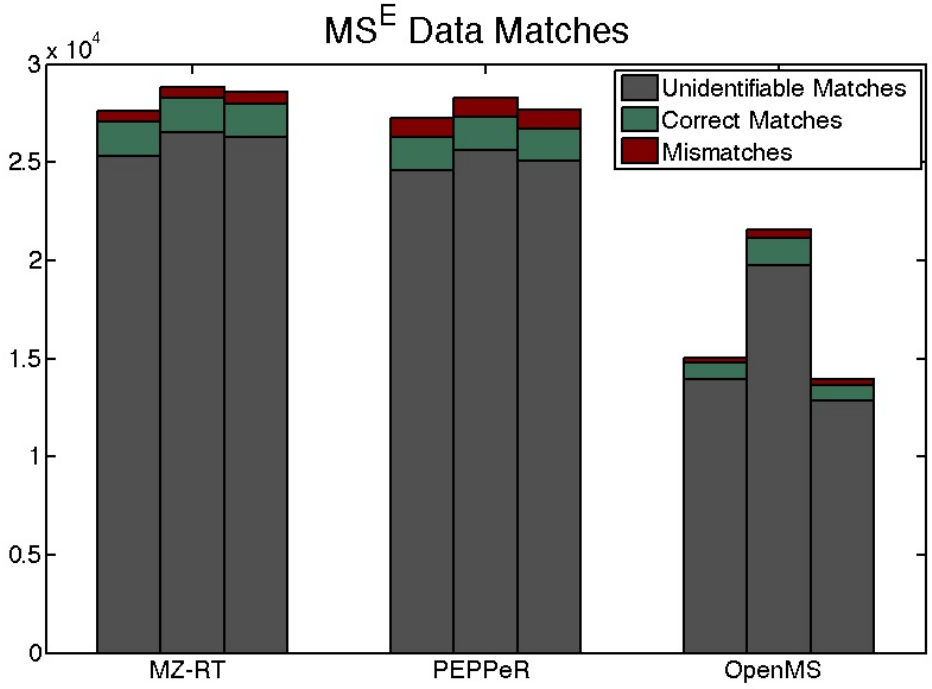

FIGURE 5.8: MS^E Alignment Match Counts. This figure showsa bar plot of all correct, incorrect and unidentifiable matches for each method. Unidentifiable matches are pairwise matches where neither peptide has a putative peptide sequence, and so the accuracy cannot be inferred.

increase in mismatches from the MZ-RT alignment, and a significant increase in mismatches from the MZ-IM alignment. When including product ion profiles, we see significant increases in recall rates, particularly with more confident matches. We see an insignificant increase in mismatches from the MZ-RT and MZ-RT-IM alignments. The alignment including product ion profiles results in much more confident matches overall. It should be noted that the results presented for the *E. coli* lysate analysis assume that all peptide identifications having an $Identity^E$ Li et al. (2009) peptide score 5 or greater are correct. We also assume that a match of two peptides differing only by a leucine vs. isoleucine amino acid call or by amino acid order in the peptide sequences, still represents an mismatch. The resulting p-values across the range of match probability stringencies are available inthe Appendix. We also examined

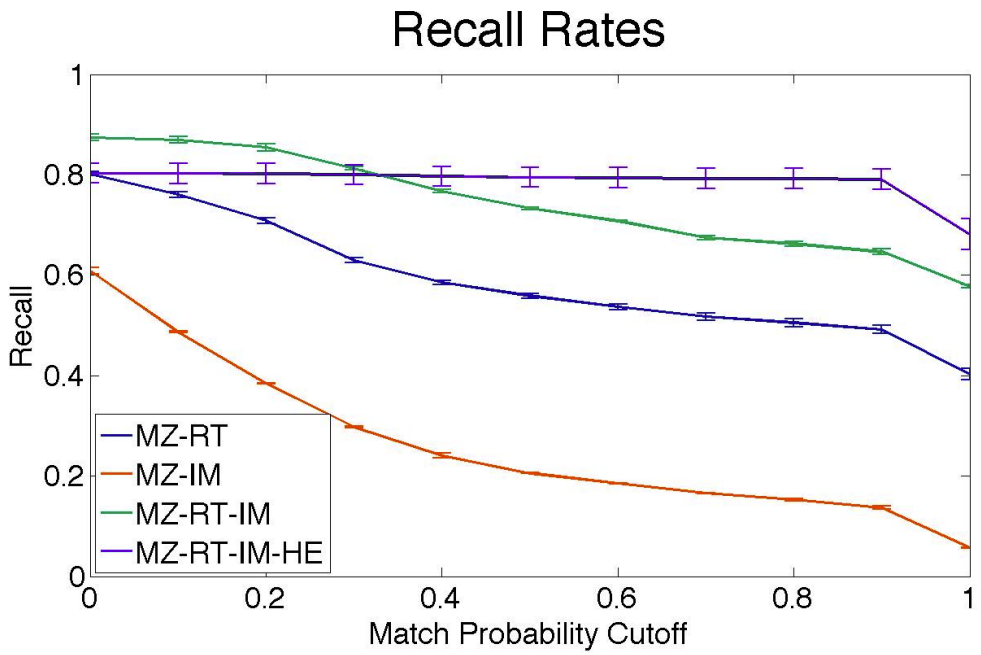

FIGURE 5.9: HDMS^{E} Alignment Recall Rates. This figure shows the recall rates considering identifications having peptide score 5 or greater across a range of match probability cutoffs.

the match probabilities of all shared identifications the peptides that should be matched - between the replicates, for each of the four alignments. Histograms of the match probabilities are shown in Figure 5.11. We see many more confident match probabilities (near 1) of the shared identifications for the MZ-RT-IM when comparing to both two-dimensional alignments. When including product ion profiles with the MZ-RT-IM-HE alignment, we also see an increase in confident match probabilities, with a migration of all intermediate match probabilities to near-zero.

We see many more confident match probabilities (near 1) of the shared identifications for the MZ-RT-IM when comparing to both two-dimensional alignments. When including product ion profiles with the MZ-RT-IM-HE alignment, we also see an increase in confident match probabilities, with a migration of all intermediate match probabilities to near-zero.

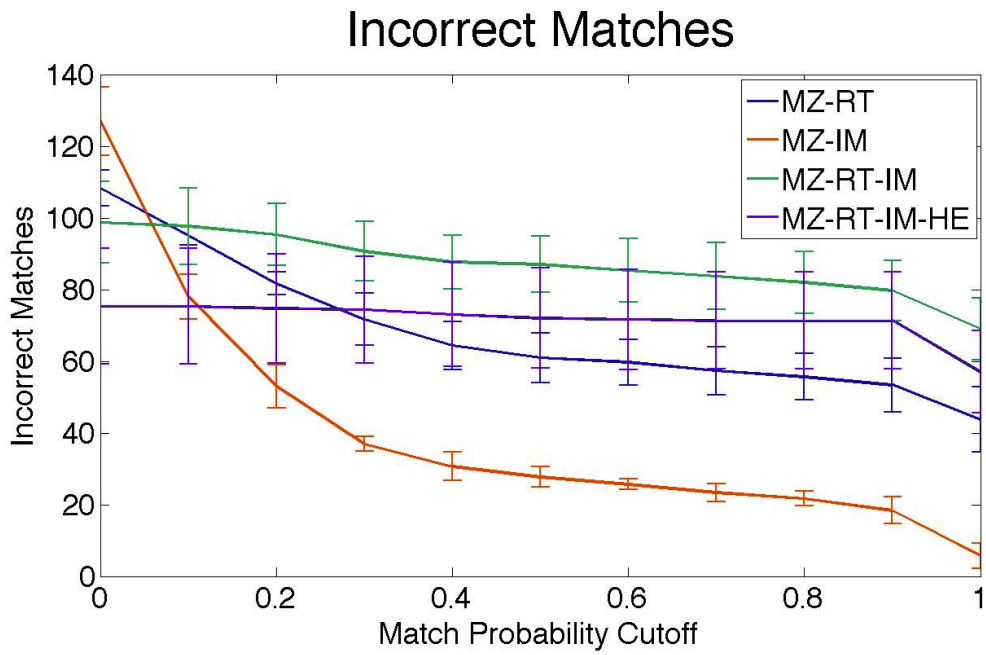

FIGURE 5.10: HDMS^{E} Alignment Mismatch Counts. This figure shows the number of incorrect matches considering identifications having peptide score 5 or greater across a range of match probability threshold.

5.3.2 Human with E. coli *Lysate Decoy*

In order to assess alignment performance without experimental identification bias, we aligned technical replicates of *E. coli* lysate data with a Human plasma decoy, and technical replicates of Human plasma with an *E. coli* lysate decoy. One technical replicate of the *E. coli* lysate sample was aligned with another *E. coli* lysate technical replicate combined in silico with a decoy Human plasma sample. To combine the samples, we append the human plasma peak list onto the peak list of one of the *E. coli* replicates. The Human plasma was a pooled sample from 20 individuals, run with 2 technical replicates. Similarly, one replicate of the human plasma was aligned with another plasma technical replicate combined with an *E. coli* lysate sample. As with the *E. coli* lysate data, we performed the four different alignments described in Table 5.1. Each alignment was provided with the same 15 given matches the 15 identifications shared by the technical replicates, and having the highest average

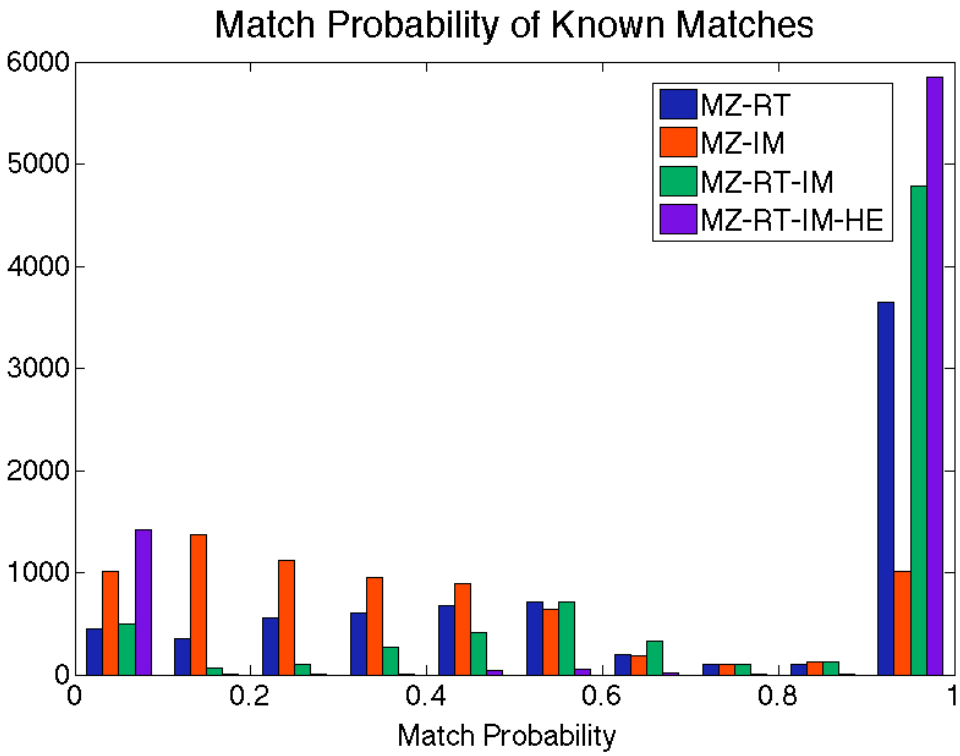

FIGURE 5.11: HDMS^{E} Alignment Known Match Probabilities. A histogram of the match probabilities of all shared identifications having peptide score 5 or greater, for each of the four alignments of the *E. coli* lysate data.

PLGS Identity^{E} Li et al. (2009) peptide score. To assess alignment performance, we determine the proportion of incorrect species alignments. This provides a false discovery rate that avoids experimental identification bias - assuming the set of shared peptides across species is negligible Wilkins and Williams (1997). Figure 5.12 shows the correct and incorrect species match counts for each of the four alignments.

We see that the MZ-RT-IM and MZ-IM alignments yield similar results, while the MZ-RT alignment obtains fewer incorrect species matches, but fewer matches overall. The MZ-RT-IM-HE alignment obtains the least incorrect species matches, and many more correct species matches at reasonably confident match levels. The results of the inverse decoy analysis (aligning technical replicates of *E. coli* lysate with a Human plasma decoy) are available in the Appendix.

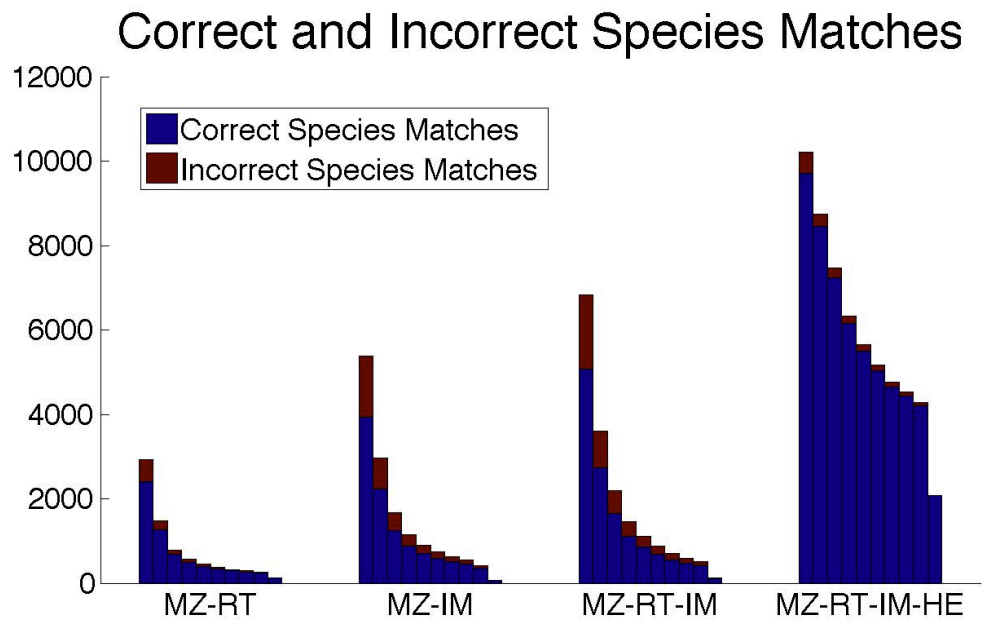

FIGURE 5.12: Decoy Alignment Results. This figure shows the number of matches made to the correct species, and the number of matches made to the incorrect species for each of the four alignments, across a range of increasing match confidence thresholds from 0.1 to 1 in 0.1 intervals.

5.3.3 Hepatitis-C and Osteoarthritis Data

To illustrate the utility of aligning datasets obtained from two different tissues, we aligned MS^E serum samples of Hepatitis-C patients with MS^E human urine samples from an Osteoarthritis cohort using peptide ion mass-to-charge ratio and retention time. Peptide identifications from the urine samples were carried over to the serum, and from the serum to the urine via alignment. For each of the datasets, we examined the 500 peptides having the most significant differential expression with a phenotype of interest treatment response in Hepatitis-C, and disease progression in Osteoarthritis. Significance was assessed with two-sample t-tests assuming equal variance, on peptide intensities that were log-transformed, mean-centered by sample, and standardized by peptide. We explored the inferred identifications that were carried over via alignment from the other tissue, in order to identify potential

biomarkers. Specifically, we looked at the previously unidentified peptides exhibiting significant differential expression having an inferred identification from the alignment results. We analyzed these inferred identifications using GATHER Chang and Nevins (2006) and DAVID Dennis et al. (2003) to search for functional and pathway enrichment. The lists of proteins from these inferred identifications and their GATHER and DAVID analysis results are available in the Additional File 2. The inferred identifications of peptides exhibiting differential expression for Hepatitis-C treatment response totaled 42 corresponding proteins. These proteins were significantly enriched for defense response (GO0006952 - 15 proteins), immune response (GO0006955 15 proteins), and response to biotic stimulus (GO0009607 14 proteins), having GATHER p-values 1.56e-9, 3.76e-9, and 9.81e-9, respectively. These functional annotations are very much in accordance with what one would expect with a response to viral infection, suggesting that useful identifications were obtained with the alignment. In addition, the genes encoding 6 of these proteins were located at 14q32, indicating significant chromosome location enrichment with GATHER p-value 1.27e-5. This is the location of the immunoglobulin heavy locus a region containing genes encoding the heavy chains of antibodies. Differential expression of genes in this region is also in accordance with what one would expect for HCV treatment response. Similarly, the inferred identifications of peptides exhibiting significant differential expression for Osteoarthritis progression totaled 47 corresponding proteins. These proteins were significantly enriched for complement activation (GO0006956 - 14 proteins), response to pest, pathogen, or parasite (GO0009613 20 proteins), and response to external biotic stimulus (GO0043207 20 proteins), having GATHER p-values 2.39e-24, 5.71e-16, and 1.97e-15, respectively. In addition, 16 of these proteins are involved in the Complement and Coagulation Pathway with a DAVID p-value of 3.0e-24. These findings are in accordance with recent evidence that the complement pathway has been found to play a critical role in the pathogenesis of osteoarthritis

Wang et al. (2011).

5.4 Discussion

We have developed a novel method for label-free proteomics data alignment that incorporates aspects of the data ignored by other open-source data alignment methods. Our alignment method incorporates ion mobility separation data and MS/MS product ion data. Our results suggest that the inclusion of more data characteristics increases alignment sensitivity, and increases matching robustness.

When comparing to OpenMS, our method obtains significantly higher recall rates (Figure 5.6), as well as more overall matches. This is likely due to the density of the data we use for comparison, and the matching technique of OpenMS. After the de-warping step, OpenMS makes pairwise matches between samples, or "maps", if the putative match is the nearest neighbor and the distance to the second-nearest neighbor is significantly greater. This results in low false positive rates, as seen in Figure 5.7. However, in dense datasets this appears to result in lower recall rates as many true matches are not considered. In Figure 5.8 , we also observe that for OpenMS, one pairwise combination of technical replicates shows significantly more matches than the other pairwise combinations. This may be the result of selecting a single reference sample to which all other samples are aligned. OpenMS first selects the sample with the most features as the reference. Each remaining sample is then aligned to the reference, estimating a "consensus" with each pairwise alignment. The density and number of features within the consensus increases with each pairwise alignment, resulting in fewer matches meeting the nearest-neighbor criteria at each step. It should be noted that we only evaluated the peptide feature-based functionality of OpenMS as a comparison to our feature-based alignment method. The ability to work with raw data, alignment specificity and ease-of-use of OpenMS are advantageous for many applications.

Our alignment model before the addition of the product ion component is very similar to PEPPeR Jaffe et al. (2006) both methods are built on the principles of Gaussian Mixture Models. Our results in the MS^E comparison reflect the similarity of our approaches. The three main differences between PEPPeR and our MZ-RT method are the technique for inferring the number of mixture components, PEPPeRs splitting of the data by charge state, and our constraint allowing only one measured peptide per sample in a given mixture component. We chose to ignore charge state information to avoid propagation of errors from earlier data-processing steps, although alignments can easily be stratified by charge state with our method.

Figures 5.9, 5.10, and 5.11 illustrate the significant improvements resulting from the inclusion of ion mobility and product ion data, while maintaining low levels of mismatches. In addition, the inclusion of more data particularly the product ion profile information results in increased confidence and robustness of alignment matching. We note that none of the alignments reach a recall rate of 1. This is likely due to to the tendency of our method to generate new latent peptide when a confident match to an existing latent peptide does not exist. This same behaviour avoids large numbers of false positive matches.

In our decoy experiment, we observed that the addition of product ion data results in a dramatic decrease in false matches, this is likely due to the lack of confounding product ion assignments to precursor ions as the decoy data is a separate experiment. However, if one were utilizing product ion data to align measurements to an AMT tag-like database, we would expect a comparable situation. These results also speak to the importance of accurate product ion to precursor ion assignment in DIA if peptides were well separated experimentally and accurate product ion assignments were made, alignment accuracy would increase dramatically. Aligning data from different experiments can actually yield additional identifications, as illustrated by the alignment of human urine data to human serum. Due to the diverse protein compo-

sition of different types of samples, specific peptides may be identified more easily in certain types of samples. It is worth noting that this behavior is much like spectral library searching Lam (2011), because surrounding peptides will not confound the product ion assignments to precursor ions. We show that the alignment of data from different tissues (even when only utilizing precursor ion data) has utility for inferring peptide identifications. If this were extended to a database and data from many tissues were used to update the database, it could have a comprehensive identification set of measured peptides, and be utilized as an additional resource or replacement for de novo identification. This is particularly useful in biomarker analyses when performing a label free experiment for an initial analysis, and then identifying proteins of interest for a subsequent targeted analysis. Also, the addition of product ion data will provide more confident alignments, and thus more confidence in identifications that may be carried over. Although we argue that the incorporation of product ion data can result in more matches of increased confidence, it should be noted that the method in which these data are incorporated has importance. If the presence of additional product ions, or the lack of product ions is highly penalized, alignments are likely to obtain fewer matches due to the variability in measurement of product ions. Conversely, if differences in product ions are not penalized enough, alignments are likely to obtain more matches, and more incorrect matches particularly because nearby peptides with respect to mass-to-charge ratio, retention time and drift time, will be those with incorrect, but similar product ion profiles. When incorporating product ion data, researchers should consider the penalization of extra and missing product ions within the data being aligned. We found that similarity functions based on sums rather than products worked well, specifically, the sum of squared differences. Our exploration of other product ion profile similarity functions is described in the Appendix.

5.5 Conclusions

Presented here is a novel method for open-platform proteomics data alignment with the ability to incorporate previously unused aspects of the data, particularly ion mobility drift times and MS/MS product ion data. Our method results in increased match recall rates and similar or improved mismatch rates compared to PEPPeR and OpenMS feature-based alignment. We also show that the incorporation of product ion data when aligning to a different dataset (or a database) can improve results dramatically. This is likely due to the lack of confounding by incorrect product ion assignments to nearby precursor ions. Based on these results, we argue that the incorporation of ion mobility drift times and product ion information are worthy pursuits. The addition of drift times and/or high energy to AMT tag databases can greatly improve experimenters ability to identify measured peptides, reducing analysis costs and the need to run additional experiments. In addition, alignment methods should be flexible enough to utilize all potential data, particularly with the recent advancements in experimental separation methods. When incorporating high-energy data, researchers should consider the penalization of extra and missing product ions from data being aligned. The results presented here provide motivation for further exploration of incorporating additional separation information into proteomics data processing, particularly as experimental advancements are made in the field.

5.6 Acknowledgements

We would like to acknowledge the support of the Measurement to Understand Re-Classification of Disease of Cabarrus and Kannapolis (MURDOCK) Study, the NIH CTSA (Clinical and Translational Science Award) 1UL1RR024128-01, and the Defense Advanced Research Projects Agency (DARPA), number lN66001-07-C-0092 (G.S.G.).

Appendix A

Proteomics Alignment Model Supplemental Information

A.1 Software and Data Formatting

A.1.1 Data Processing and Formatting

Data Processing Script

We have written a data-processing script to read Waters spectrum.xml and final-frag.csv files into a Matlab data frame for subsequent alignment with our software. The script and accompanying functions are in the software archive. The prepare-Datasets function takes a variety of input file formats, pre-processes intensities, and creates a data structure for the alignment software. It can be called as follows from the Matlab command window: datasets = prepareDatasets(file_names, options, data_range, peptide_list, sample_list) The variable file_names should be a cell array of strings. Each row should be a dataset (or single sample), with each column containing a different file for that dataset. This function is designed to handle 4 different input types:

1. 2 files listed in this order: spectrum.xml, finalfrag.csv (Water's pipeline output)

2. 1 file: spectrum.xml (Water's pipeline output)

3. 1 file: finalfrag.csv (Water's pipeline output)

4. 1 file: data.mat (un-processed matlab data frame).

The variable options should be a 7x1 or 1x7 vector of numbers indicating a user's selections for the following options (in order): set very low intensity values to missing (0 for off, or a raw intensity lower limit), impute the missing intensity values (0 for off and 1 for on), combine replicates (0 for off and 1 for on), the maximum number of product ions to store per peptide (most intense stored first), a peptide annotation quality limit (1 for "Pass 1", 2 for "Pass 1" and "Pass 2"), an option for log-transforming the intensities (0 for off and 1 for on), and an option for mean-centering the intensities for each sample. This input vector is optional, but must be included if any of the subsequent input parameters are used. To use the default values, the user may simply specify an empty vector [] in its place. The variable data_range should be a 3x2 matrix of numbers indicating the minimum and maximum (inclusive) data ranges for the mass-to-charge ratio, retention time, and drift time dimensions, respectively. The first column should contain the minimum values and the second, the maximum values. This input matrix is optional, but must be included if any of the subsequent input parameters are used. To use the entire data range, the user may simply specify an empty matrix [] in it's place. To only specify a minimum OR a maximum, or only limit certain dimensions, use NaNs in the unlimited slots. The variable peptide_list should be a cell array of peptide ids to include in the prepared data. Each row should correspond to a dataset, and each column a peptide id. This input cell array is optional, but must be included if any of the subsequent input parameters are used. To use all peptides, the user may simply specify an empty vector [] in it's place. The variable sample_list should be a cell array of sample ids to include in the prepared data. Each row should correspond to a dataset, and each

column a sample id. With single-sample files, the corresponding sample ID in this list will be used in the resulting data frame. With aggregate data, the sample IDs in this list will be used to select specific samples to include in the data frame. This input cell array is optional. To use all samples for aggregate data, the user may simply specify an empty vector [] in its place.

Data Frame Format

The output, "datasets", will be a list of data frames containing the following fields:

sids Sample IDs

pids Peptide IDs

data Mass-to-Charge Ratio, Retention Time, and Ion Mobility

xpr Intensities (Rows-Peptides, Columns-Samples)

key Sample Key Information

keyHead Key Header

anno Peptide Annotation Information

annoHead Annotation Header

e Peptide Charge State

pep Modified Peptide Sequence (or "-" if missing)

pCode Protein Code (or "-" if missing)

productmz Product Ion Mass-to-Charge Ratios

productints Product Ion Intensities (Raw)

productanno Product Ion Annotations (y1, y2, b7, etc.)

The prepareDatasets function can be easily adapted to incorporate any additional separation dimensions similar to ion mobility and liquid chromatography by appending additional columns into the data field. If you wish to use your own files in other formats, you may adapt the prepareDatasets function, or write your own function to generate a data frame as specified above.

A.1.2 Alignment

Alignment Setup

To run your own alignment, you should set up a config file much like sampleconfig.txt. In your config file, there should be the following specifications in order:

%DESCRIPTION: On the line following this tag, include a description of your alignment. This is for your record-keeping purposes only and does not affect the alignment.

%FILENAMES: On the lines following this tag, you should include the file(s) for your datasets being aligned. These should be one dataset (or sample) per line, and may be specified 4 different ways:

- Two files separated by commas and listed in this order: spectrum.xml,finalfrag.csv (Water's pipeline output)
- One file: spectrum.xml (Water's pipeline output)
- One file: finalfrag.csv (Water's pipeline output)
- One file: data.mat (file containing a matlab data frame in the format specified above)

%DIMNAMES: On the line following this tag, you should list the peptide-level separation dimensions you wish to use in your alignment, separated by commas.

If you are using HDMSE data and want to use all three dimensions, the line should read: Monoisotopic m/z,Peak Centroid Time,Ion Mobility

%NGIVEN: This is the number of shared identifications that the model will trust. These will remain matched at all iterations of the MCMC and be used as the seed matches to set hyperparameters. If you would like to trust all identifications in your data, specify inf.

%MEASHE: The value on the line following this tag should be the maximum number of measured product ions to use per peptide (most intense used first). If you dont want to run a HE alignment, set this to 0.

%HEVECT: The value on the line following this tag should be the profile size K of the product ion profiles.

%ITERATIONS: There should be three value lines following this tag. The first is the number of burn-in iterations for the Gibbs Sampler, the second is the number of iterations post burn-in, and the third is the number of assignment-only iterations used to estimate match probabilities and the final alignment.

%NSPLIT: The value following this tag is the maximum number of splits to use to parallelize data alignment. If you are running your alignment on a cluster, each split will be set up as a separate job. If you are running your alignment locally, each split will be run consecutively after the previous split finishes.

%CODE: The line following this tag should contain the path of the directory containing the alignment software.

%OUT: The line following this tag should contain the path of the desired directory for results files.

%START_COMBINE: The line following this tag should contain either START, or COMBINE. The keyword START instructs the software to either start the alignment (if run locally) or set up the alignment (if run on a cluster. The keyword COMBINE instructs the software to combine the alignment splits this should be run after all alignment partitions are finished and is only necessary if NSPLIT was set to a value greater than one.

%LOCAL_CLUSTER: The line following this tag should contain either LOCAL, or CLUSTER. The keyword LOCAL instructs the software to start the alignment on the local machine. The keyword CLUSTER sets up as many as NSPLIT qsub files and a submission bash script submitall.sh to be submitted to a SGE queue.

Running Your Alignment

If you are running your alignment locally using only 1 split, you can simply use the Align function provided in the software with your config file: Align(config.txt). If you are running your alignment locally using multiple splits, you should do the following:

1. Make sure the value line of the START_COMBINE tag in your config file says START, and then run Align(config.txt) in your Matlab command window.

2. Modify the value line of the START_COMBINE tag in your config file to say COMBINE.

3. After all alignment partitions have finished, run Align(config.txt) in your Matlab command window once more.

If you are running your alignment in many partitions on a cluster, you should make a queue file for setting up your alignment like the file sampleconfig.q, and do the following:

1. Run qsub sampleconfig.q to set up your alignment.

2. Once the job in 1 finishes, run the submission script to start your alignment: bash submitall.sh.

3. Modify the value line of the START_COMBINE tag in your config file to say COMBINE.

4. After all alignment partitions have finished, run qsub sampleconfig.q once more to collect your alignment results.

Alignment Results Format

After your alignment, There will be 6 files containing various aspects of your alignment results. Each file is a matlab data frame and their contents are described in detail below.

split_datasets.mat This file contains your original datasets as they were split into mass-to-charge ratio positions

split_datasets A data structure of size NSPLIT x the number of datasets. Each element in the structure is a subset of the original data within a determined m/z range, having the format specified in section 1.1.2.

db.mat This file contains the new inferred dataset, combining the information from your aligned datasets into a single data frame.

db A data structure containing inferred peptide sequences, protein names, charge states, peptide intensities, and peptide ids for the latent peptides. The peptide intensities are collected from measured peptides aligned to the specific latent peptide, and are in the same order as the input datasets column-wise.

latentpeps A data structure containing the inferred monoisotopic m/z, retention time, and drift time, and the inferred product ion profiles of each latent peptide.

matchinfo.mat This file contains the original datasets and match information from the alignment.

datasets The original datasets in the format specified in section 1.1.2.

matches For each latent peptide (row), the indices of the assigned measured peptide from each dataset (column).

matched_pids For each latent peptide (row), the peptide ID (pids) of the assigned measured peptide from each dataset (column).

peps For each latent peptide (row), the peptide sequences of the assigned measured peptide from each dataset (column).

pcodes For each latent peptide (row), the protein name of the assigned measured peptide from each dataset (column)

accuracy.mat This file contains information about the number of matches, correct matches, and incorrect matches based on the identifications in your data. This is only relevant if the number of seed matches given to the model was fewer than the number of shared identifications.

nummatches The number of matches obtained between dataset i (dimension 1) and dataset j (dimension 2) at match probability cutoff k (dimension 3) or greater. Values of match probability cutoffs are 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, and 1.

case1 The number of correct matches obtained between dataset i (dimension 1) and dataset j (dimension 2), considering identifications of peptide score

cutoff k (dimension 3) or greater, at match probability cutoff l (dimension 3) or greater. Values of peptide score cutoffs are 3, 4, 5, 6, 7, 8, 9 and 10, and values of match probability cutoffs are as listed above.

case2 The number of incorrect matches obtained between dataset i (dimension 1) and dataset j (dimension 2), considering identifications of peptide score cutoff k (dimension 3) or greater, at match probability cutoff l (dimension 3) or greater. Values of peptide score cutoffs are 3, 4, 5, 6, 7, 8, 9 and 10, and values of match probability cutoffs are as listed above.

case3 The number of matches of unknown accuracy, resulting from neither dataset having an identification, between dataset i (dimension 1) and dataset j (dimension 2), considering identifications of peptide score cutoff k (dimension 3) or greater, at match probability cutoff l (dimension 3) or greater. Values of peptide score cutoffs and values of match probability cutoffs are as listed above.

case4 The number of matches of unknown accuracy, resulting from one dataset having an identification that is not known to exist in the other, between dataset i (dimension 1) and dataset j (dimension 2), considering identifications of peptide score cutoff k (dimension 3) or greater, at match probability cutoff l (dimension 3) or greater. Values of peptide score cutoffs and values of match probability cutoffs are as listed above.

case5 The number of incorrect matches, resulting from one dataset having an identification that exists elsewhere in the other, between dataset i (dimension 1) and dataset j (dimension 2), considering identifications of peptide score cutoff k (dimension 3) or greater, at match probability cutoff l (dimension 3) or greater. Values of peptide score cutoffs and values of match probability cutoffs are as listed above.

possible_case1 The number of possible correct matches (shared identifications) between dataset i (dimension 1) and dataset j (dimension 2), considering identifications of peptide score cutoff k (dimension 3) or greater. Values of peptide score cutoffs are as listed above.

model.mat This file contains information about the estimated model parameters.

mcmcparams A structure of size NSPLIT containing MAP estimates for the shift, scale, match covariance, and latent peptide covariance parameters for each partition of the alignment.

parameters A structure of size NSPLIT containing hyperparameters set by the seed matches, and alignment parameters as specified in the config file.

matchprobs For each latent peptide (row), the probability of assignment for the final matched measured peptide from each dataset (column).

matchinit A structure of size NSPLIT containing seed matches for each alignment split. For each latent peptide (row), the indices of the seed matched peptide from each dataset (column).

maxpepperds The maximum number of measured peptides from the datasets being aligned (the size of the largest dataset).

matchaverages.mat This file contains the results from the assignment-only iterations of the sampler.

matchaverages A sparse matrix containing the match probability of each measured peptide to each latent peptide, from the assignment-only iterations of the sampler.

A.2 Full Conditional Distributions

The full conditional posterior distributions used to update the shift, scale, residual covariance, latent peptides, latent peptide covariance, and match indicators are shown below.

Given D open-platform proteomics datasets/samples $X = x_{1,1}...x_{D,N_D}$ of N total peptide features, where $\sum_{d=1}^{D} N_d = N$ and x_{d,N_d} is k-dimensional, we fit the DPGMM, matching each measurement to one of J latent peptide features, $Z = z_1, ..., z_J$ where J is unbounded. A global, linear shift and scale of peptide-level data is simutaneously estimated to de-warp experimental variation. We first describe the model assuming for peptide-level, or precursor ion data, and then describe the extension to incorporate product ion information. The measurement $x_{d,i}$, assigned to latent peptide feature j is expected to be a shifted and scaled approximation of z_j, with dataset or sample-specific shift and scale parameters with Gaussian noise. The latent peptide features are modeled as components of the DPGMM - with z_j having mean μ_j and covariance σ, making the simplifying assumption that latent peptide feature precision is the shared across all latent peptide features. Conjugate priors are used

for all model parameters.

$$\begin{aligned} x_{d,i} &= \eta_d + z_j\beta_d + \epsilon_{d,i} \\ \epsilon_{d,i} &\sim Normal(0, \Sigma) \\ z_j &\sim Normal(\mu_j, \sigma) \end{aligned} \tag{A.1}$$

$$\begin{aligned} \eta_d &\sim Normal(a_d, b_d) \\ \beta_d &\sim Normal(e_d, f_d) \\ \mu_j &\sim Normal(\lambda, r) \\ \sigma &\sim iWishart(g, h) \\ \Sigma &\sim Inverse - Wishart(s, t) \end{aligned} \tag{A.2}$$

We adapt the Chinese Restaraunt Process formulation of the DPGMM, introducing indicator variables $c_{d,i}$ for latent peptide feature assignment, where $d = 1...D$, and $i = 1...N_d$ (one for each measurement) and $c_{d,i} \in \{1, 2, ...J\}$. Occupation numbers n_j for $j = 1...J$ are also introduced, where n_j is the number of $c = j$. We express the likelihood conditioned on the indicators.

$$\begin{aligned} &P(X \mid \eta, \beta, \Sigma, z_1...z_J, c_{1,1}...c_{D,N_D}) = \\ &\prod_{d=1}^{D}\prod_{i=1}^{N_d} Normal(x_{d,i} \mid \eta_d + z_{c_{d,i}}\beta_d, \Sigma) \end{aligned} \tag{A.3}$$

We may also integrate out Z and re-express the likelihood as follows:

$$\begin{aligned} &P(X \mid \eta, \beta, \Sigma, \mu_1...\mu_J, \sigma, c_{1,1}...c_{D,N_D}) = \\ &\prod_{d=1}^{D}\prod_{i=1}^{N_d} Normal(x_{d,i} \mid \eta_d + \mu_{c_{d,i}}\beta_d, \beta_d^T\sigma\beta_d + \Sigma) \end{aligned} \tag{A.4}$$

Given the above likelihood and prior distribution, we obtain the full conditional distribution of the global shift parameter, η_d, as follows:

$$
\begin{aligned}
P(\eta_d \mid -) &\propto P(X \mid \eta_d, \beta_d, \Sigma, \mu_1 ... \mu_J, \sigma, c_{d,1} ... c_{d,N_d}) \times P(\eta_d \mid a_d, b_d) \\
&\propto \prod_{i=1}^{N_d} Normal(x_{d,i} \mid \eta_d + \mu_{c_{d,i}} \beta_d, \beta_d^T \sigma \beta_d + \Sigma) \times Normal(\eta_d \mid a_d, b_d) \\
&\propto (2\pi)^{-\frac{k}{2}} \mid b_d \mid^{-\frac{1}{2}} e^{-\frac{1}{2}(\eta_d - a_d)^T b_d^{-1} (\eta_d - a_d)} \times \prod_{i=1}^{N_d} (2\pi)^{-\frac{k}{2}} \mid (\beta_d^T \sigma \beta_d + \Sigma) \mid^{-\frac{1}{2}} \\
&\times exp\left(-\frac{1}{2} (x_{d,i} - \eta_d - \mu_{c_{d,i}} \beta_d)^T (\beta_d^T \sigma \beta_d + \Sigma)^{-1} (x_{d,i} - \eta_d - \mu_{c_{d,i}} \beta_d) \right) \\
&\propto Normal\left(W^{-1} \left[b_d^{-1} a_d + \left(\beta_d^T \sigma \beta_d + \Sigma \right)^{-1} \left(\sum_{i=1}^{N_d} x_{d,i} - \mu_{c_{d,i}} \beta_d \right) \right], W^{-1} \right) \\
&W = \left[b_d^{-1} + N_d (\beta_d^T \sigma \beta_d + \Sigma)^{-1} \right]
\end{aligned}
\tag{A.5}
$$

Similarly, we obtain the full conditional distribution of the global scale parameter, β_d, as follows:

$$
\begin{aligned}
P(\beta_d \mid -) &\propto P(X \mid \eta_{d,k}, \beta_{d,k}, \Sigma_{k,k}, Z_{1,k} ... Z_{J,k}, c_{d,1} ... c_{d,N_d}) \times P(\beta_{d,k} \mid c_{d,k}, d_{d,k}) \\
&\propto \prod_{i=1}^{N_d} Normal\left(\frac{x_{d,i,k} - \eta_{d,k}}{Z_{c_{d,i},k}} - \beta_{d,k} \mid 0, \frac{\Sigma_{k,k}}{Z_{c_{d,i},k}^2} \right) \times Normal(\beta_{d,k} \mid c_{d,k}, d_{d,k}) \\
&\propto (2\pi)^{-\frac{1}{2}} d_{d,k}^{-\frac{1}{2}} e^{-\frac{1}{2}(\beta_{d,k} - c_{d,k})^2 d_{d,k}^{-1}} \times \prod_{i=1}^{N_d} (2\pi)^{-\frac{1}{2}} \frac{\Sigma_{k,k}}{Z_{c_{d,i},k}^2}^{-\frac{1}{2}} e^{-\frac{1}{2}(\frac{x_{d,i,k} - \eta_{d,k}}{Z_{c_{d,i},k}} - \beta_{d,k})^2 (\frac{\Sigma_{k,k}}{Z_{c_{d,i},k}^2})^{-1}} \\
&\propto Normal\left(\frac{\frac{c_{d,k}}{d_{d,k}} + \sum_{i=1}^{N_d} \frac{Z_{c_{d,i},k}}{\Sigma_{k,k}} (x_{d,i,k} - \eta_{d,k})}{\frac{1}{d_{d,k}} + \sum_{i=1}^{N_d} \frac{Z_{c_{d,i},k}^2}{\Sigma_{k,k}}}, \frac{1}{\frac{1}{d_{d,k}} + \sum_{i=1}^{N_d} \frac{Z_{c_{d,i},k}^2}{\Sigma_{k,k}}} \right)
\end{aligned}
\tag{A.6}
$$

We obtain the full conditional posterior distribution for the mean of each latent peptide as follows:

$$
\begin{aligned}
P(\mu_j \mid -) &\propto P(X \mid \eta_d, \beta_d, \Sigma, \mu_j, \sigma, c_{1,1}...c_{D,N_D}) \times p(\mu_j \mid \lambda, r) \\
&\propto \prod_{d,i:c_{d,i}=j} Normal((x_{d,i}-\eta_d)\beta_d^{-1} \mid \mu_{c_{d,i}}, \sigma + \beta_d^{-1}\Sigma\beta_d^{-1}) \times Normal(\mu_j \mid \lambda, r) \\
&\propto \prod_{d,i:c_{d,i}=j} (2\pi)^{-\frac{k}{2}} \mid \sigma + \beta_d^{-1}\Sigma\beta_d^{-1} \mid^{-\frac{1}{2}} e^{-\frac{1}{2}((x_{d,i}-\eta_d)\beta_d^{-1}-\mu_j)^T(\sigma+\beta_d^{-1}\Sigma\beta_d^{-1})^{-1}((x_{d,i}-\eta_d)\beta_d^{-1}-\mu_j)} \\
&\times (2\pi)^{-\frac{k}{2}} \mid r \mid^{-\frac{1}{2}} e^{-\frac{1}{2}(\mu_j-\lambda)^T r^{-1}(\mu_j-\lambda)} \\
&\propto Normal\left(R^{-1}\left[r^{-1}\lambda + \sum_{d,i:c_{d,i}=j}(\sigma + \beta_d^{-1}\Sigma\beta_d^{-1})^{-1}(x_{d,i}-\eta_d)\beta_d^{-1}\right], R^{-1}\right) \\
R &= \left[r^{-1} + \sum_{d,i:c_{d,i}=j}(\sigma + \beta_d^{-1}\Sigma\beta_d^{-1})^{-1}\right]^{-1}
\end{aligned}
\tag{A.7}
$$

We derive the full conditional distribution for the residual covariance as follows:

$$
\begin{aligned}
P(\Sigma \mid -) &\propto P(X \mid \eta, \beta, \Sigma, Z_1...Z_J, c_{1,1}...c_{D,N_D}) \times P(\Sigma \mid s, \nu) \\
&\propto \prod_{d=1}^{D}\prod_{i=1}^{N_d} Normal(x_{d,i} - \eta_d - z_{c_{d,i}}\beta_d \mid 0, \Sigma) \times inverseWishart(\Sigma \mid s, \nu) \\
&\propto \prod_{d=1}^{D}\prod_{i=1}^{N_d}(2\pi)^{-\frac{k}{2}} \mid \Sigma \mid^{-\frac{1}{2}} e^{-\frac{1}{2}(x_{d,i}-\eta_d-z_{c_{d,i}}\beta_d)^T\Sigma^{-1}(x_{d,i}-\eta_d-z_{c_{d,i}}\beta_d)} \\
&\times \left[2^{\nu k/2}\pi^{\binom{k}{2}/2} \mid s \mid^{-\nu/2} \prod_{k=1}^{K}\Gamma((\nu+1-k)/2)\right]^{-1} \mid \Sigma \mid^{-(\nu+k+1)/2} e^{-tr(s\Sigma^{-1})/2} \\
&\propto iWishart([s+s_\theta]^{-1}, \nu+N) \\
s_\theta &= \sum_{d=1}^{D}\sum_{i=1}^{N_d}(x_{d,i}-\eta_d-z_{c_{d,i}}\beta_d)^T(x_{d,i}-\eta_d-z_{c_{d,i}}\beta_d)
\end{aligned}
\tag{A.8}
$$

The full conditional distribution for the shared latent peptide covariance is ob-

tained as follows:

$$
\begin{aligned}
P(\sigma \mid -) &\propto P(Z \mid \mu_1 ... \mu_j, \sigma) \times p(\sigma \mid g, h) \\
&\propto \prod_{j=1}^{J} Normal(z_j \mid \mu_j, \sigma) \times inverseWishart(\sigma \mid g, h) \\
&\propto \prod_{j=1}^{J} (2\pi)^{-\frac{k}{2}} \mid \sigma \mid^{-\frac{1}{2}} e^{-\frac{1}{2}(z_j - \mu_j)^T \Sigma^{-1} (z_j - \mu_j)} \\
&\times \left[2^{hk/2} \pi^{\binom{k}{2}/2} \mid g^{-1} \mid^{-h/2} \prod_{k=1}^{K} \Gamma((h+1-j)/2) \right]^{-1} \mid \sigma \mid^{-(h+k+1)/2} e^{-tr(g^{-1}\sigma^{-1})/2} \\
&\propto \mid \sigma \mid^{-\frac{J}{2}} e^{-\frac{tr(g_\theta \sigma^{-1})}{2}} \times \mid \sigma \mid^{-(h+k+1)/2} e^{-tr(g^{-1}\sigma^{-1})/2} \\
&\propto iWishart([g + g_\theta]^{-1}, h + J) \\
&g_\theta = \sum_{j=1}^{J} (z_j - \mu_j)^T (z_j - \mu_j)
\end{aligned}
\tag{A.9}
$$

Conjugate Gaussian priors are given to the latent peptide feature means, with each latent peptide feature sharing the same hyperparameters, and an inverse-Wishart prior to the universal component covariance. As in Rasmussen (2000) the prior for a single assignment indicator given the rest is:

$$
p(c_{d,i} = j \mid c_{-(d,i)}, \alpha) = \frac{n_{-(d,i),j} + \alpha/J}{N - 1 + \alpha} \tag{A.10}
$$

Where $-(d,i)$ indicates all indices from all LC-MS runs except d, i, and $n_{-(d,i),j}$ is the number of measurements besides measurement i, in sample d assigned to latent peptide feature j. To prevent multiple features from the same run being assigned to the same latent peptide feature, we impose a restriction on measurement assigment to latent peptide features such that only one measurement per LC-MS run may be

assigned to a given latent peptide feature, making the conditional prior for a single assignment indicator given the rest:

$$p(c_{d,i} = j \mid c_{-(d,i)}, \alpha) = \frac{n_{-(d,i),j} + \alpha/J}{N - 1 + \alpha} \times I(\nexists\, c_{d,-i} = c_{d,i}) \tag{A.11}$$

where $c_{d,-i}$ indicates all indices except i from LC-MS run d. If we consider the case where the number of latent peptide features is unknown, we take the limit as J approaches ∞. Taking the limit for the conditional prior distribution on the latent peptide feature assignment indicators yields:

$$p(c_{d,i} = j \mid c_{-(d,i)}, \alpha) = \frac{n_{-(d,i),j}}{N - 1 + \alpha} \times I(\nexists\, c_{d,-i} = c_{d,i}) \tag{A.12}$$

$$p(c_{d,i} \neq c_{-(d,i)} \mid c_{-(d,i)}, \alpha) = \frac{\alpha}{N - 1 + \alpha} \tag{A.13}$$

Where A.12 represents the prior probability of introducing a new latent peptide feature. Combining these priors with the likelihood conditioned on the assignment indicators, we obtain the following conditional posteriors on peptide feature assignment:

$$\begin{aligned} P(c_{d,i} = j \mid -) \propto & \frac{n_{-(d,i),j}}{N - 1 + \alpha} \times \mid (\beta_d^T \sigma \beta_d + \Sigma) \mid^{-\frac{1}{2}} \\ & \times e^{-\frac{1}{2}(x_{d,i} - \eta_d - \mu_{c_{d,i}}\beta_d)^T (\beta_d^T \sigma \beta_d + \Sigma)^{-1} (x_{d,i} - \eta_d - \mu_{c_{d,i}}\beta_d)} \\ & \times I(\nexists\, c_{d,-i} = c_{d,i}) \end{aligned} \tag{A.14}$$

$$\begin{aligned} P(c_{d,i} \neq c_{-(d,i)} \mid -) \propto & \frac{\alpha}{N - 1 + \alpha} \times \mid (\sigma + \beta_d^{-1}\Sigma(\beta_d^{-1})^T) \mid^{-\frac{1}{2}} \mid r \mid^{-\frac{1}{2}} \\ & \times \mid (r^{-1} + (\sigma + \beta_d^{-1}\Sigma(\beta_d^{-1})^T)^{-1}) \mid^{-\frac{1}{2}} \\ & \times e^{-\frac{1}{2}\left((x_{d,i} - \eta_d)\beta_d^{-1}\right)^T \left(\sigma + \beta_d^{-1}\Sigma(\beta_d^{-1})^T\right)^{-1} (x_{d,i} - \eta_d)\beta_d^{-1} + \lambda^T r^{-1}\lambda - B^T A^{-1} B} \\ & A = (r^{-1} + (\sigma + \beta_d^{-1}\Sigma(\beta_d^{-1})^T)^{-1}) \\ & B = (r^{-1}\lambda + (\sigma + \beta_d^{-1}\Sigma(\beta_d^{-1})^T)^{-1}[x_{d,i} - \eta_d]\beta_d^{-1}) \end{aligned} \tag{A.15}$$

The computation of A.15 is possible because we assume all latent peptide features share the same covariance. We now describe the extension of the model to incorporate product ion spectra. To incorporate product ion data, we select up to the 50 most intense product ions for each peptide feature measurement, $x_{d,i}$. We then generate a K-dimensional product ion intensity profile for each $x_{d,i}$. Each position, y_{d,i_k}, in the product ion intensity profile, $y_{d,i}$, is computed as:

$$y_{d,i_k} = \frac{\sum_p \Omega_p \times I(M_p \leqslant B_k)}{\sum_p \Omega_p} \tag{A.16}$$

where $k = 1 \ldots K$, $p = 1 \ldots 50$, Ω is a 50-dimensional vector of intensities, M is a 50-dimensional vector of product ion mass-to-charge ratios, and B is a K-dimensional vector of product ion profile mass-to-charge ratio bin upper limits. All values in $y_{d,i}$ sum to one. The mass-to-charge ratio ranges, or bins, are determined at the initialization of the alignment, such that bin boundaries fall on mass-to-charge ratio deserts. See section 2.1.2.3 and Figure 5.4 for a detailed description of bin boundary determination. To assess the similarity of a measured product ion profile and a latent product ion profile, w_j for latent peptide feature z_j, we introduce a similarity score, ψ , which is computed as the sum of squared differences of the two product ion intensity profiles, and is assumed to have an exponential distribution to encourage distances close to zero.

$$\begin{aligned} \psi_{d,i} &= (y_{di} - w_{c_{d,i}})^T (y_{d,i} - w_{c_{d,i}}) \\ \psi_{di,j} &\sim Exponential(\gamma) \end{aligned} \tag{A.17}$$

We assign a conjugate gamma prior to the rate parameter. The hyperparameters for profile scores are set to one.

$$\gamma \sim Gamma(a_0, b_0) \tag{A.18}$$

The likelihood for the high energy component of the alignment model is expressed

as:

$$P(Y \mid -) = \prod_{d=1}^{D} \prod_{i=1}^{N_d} Exponential((y_{di} - w_{c_{d,i}})^T (y_{d,i} - w_{c_{d,i}}) \mid \gamma) \tag{A.19}$$

At each iteration of the MCMC, the product ion profile, w_j, of an existing latent peptide is updated empirically. The latent product ion profile is set to the average of the measured product profiles assigned to that latent peptide feature. The latent product ion profile, w_0, of a new latent peptide (one that currently does not exist) is a blank profile - a uniform vector of size K with each element having value 1/K. We update the rate parameter of the distribution on the similarity score as follows:

$$\begin{aligned} P(\gamma \mid \psi, a_0, b_0) &\propto P(\psi \mid \gamma) \times p(\gamma \mid a_0, b_0) \\ &\propto \prod_{d=1}^{D} \prod_{i=1}^{N_d} Exponential(\psi_{d,i} \mid \gamma) \times Gamma(\gamma \mid a_0, b_0) \\ &\propto \prod_{d=1}^{D} \prod_{i=1}^{N_d} \gamma e^{-\gamma \psi_{d,i}} \times \frac{1}{b_0^{a_0}} \frac{1}{\Gamma(a_0)} \gamma^{a_0 - 1} e^{-\frac{\gamma}{b_0}} \\ &\propto Gamma\left(a_0 + N, b_0 + \sum_{d=1}^{D} \sum_{i=1}^{N_d} \psi_{d,i}\right) \end{aligned} \tag{A.20}$$

Combining the product ion model with the peptide-level model, we have the following conditional posterior:

$$\begin{aligned} P(c_{d,i} = j \mid -) \propto & \frac{n_{-(d,i),j}}{N - 1 + \alpha} \times \mid (\beta_d^T \sigma \beta_d + \Sigma) \mid^{-\frac{1}{2}} \\ & \times e^{-\frac{1}{2}(x_{d,i} - \eta_d - \mu_{c_{d,i}} \beta_d)^T (\beta_d^T \sigma \beta_d + \Sigma)^{-1} (x_{d,i} - \eta_d - \mu_{c_{d,i}} \beta_d)} \\ & \times I(\nexists\, c_{d,-i} = c_{d,i}) \\ & \times \gamma e^{-\gamma (y_{di} - w_{c_{d,i}})^T (y_{d,i} - w_{c_{d,i}})} \end{aligned} \tag{A.21}$$

$$
\begin{aligned}
P(c_{d,i} \neq c_{-(d,i)} \mid -) \propto & \frac{\alpha}{N-1+\alpha} \times \mid (\sigma + \beta_d^{-1}\Sigma(\beta_d^{-1})^T) \mid^{-\frac{1}{2}} \mid r \mid^{-\frac{1}{2}} \\
& \times \mid (r^{-1} + (\sigma + \beta_d^{-1}\Sigma(\beta_d^{-1})^T)^{-1}) \mid^{-\frac{1}{2}} \\
& \times e^{-\frac{1}{2}\left((x_{d,i}-\eta_d)\beta_d^{-1}\right)^T\left(\sigma+\beta_d^{-1}\Sigma(\beta_d^{-1})^T\right)^{-1}(x_{d,i}-\eta_d)\beta_d^{-1}+\lambda^T r^{-1}\lambda - B^T A^{-1} B} \\
& \times \gamma e^{-\gamma(y_{di}-w_0)^T(y_{d,i}-w_0)} \\
& A = (r^{-1} + (\sigma + \beta_d^{-1}\Sigma(\beta_d^{-1})^T)^{-1}) \\
& B = (r^{-1}\lambda + (\sigma + \beta_d^{-1}\Sigma(\beta_d^{-1})^T)^{-1}[x_{d,i} - \eta_d]\beta_d^{-1})
\end{aligned}
\tag{A.22}
$$

A.3 Exploration of Other HE Models

We explored additional values of K, as well as implementations of different high energy models. To assess the utility and computational complexity of various high energy models, we utilized shared identifications having peptide score 5 or greater, among the first two replicates of the E. coli Lysate data. We constructed product ion profiles of various sizes, K = {50, 100, 250, 500}, and assessed correct matches, nearby incorrect matches, random incorrect matches, a blank profile, and an empirical profile using eight different scoring schemes. Correct matches are shared identifications, nearby incorrect matches are the 3 closest peptides in 3-dimentionsional space (m/z, retention time, drift time) excluding the correct match, random incorrect matches are a random peptide excluding the correct match and nearby incorrect matches, a blank profile is a uniform vector of size K with each element having value 1/K, and an empirical profile is the mean of all measured product ion profiles. The eight metrics assessed were dot product, 1-norm, 2-norm, Pearson correlation, Spearman correlation, Kendall correlation, K-dimensional multivariate normal PDF, and the sum of squared differences. For each metric, we examined box plots of the scores from different match types, and measured the CPU time it took to compute the scores for

all match types. In order to the alignment model to perform as desired that is to make correct matches, and avoid incorrect matches the appropriate metric should favor correct matches above incorrect matches and new latent peptides (blank and empirical profiles), but favor new latent peptides above incorrect matches. Figure A.1 shows boxplots from three of the metrics for two different profile sizes, illustrating different scenarios.

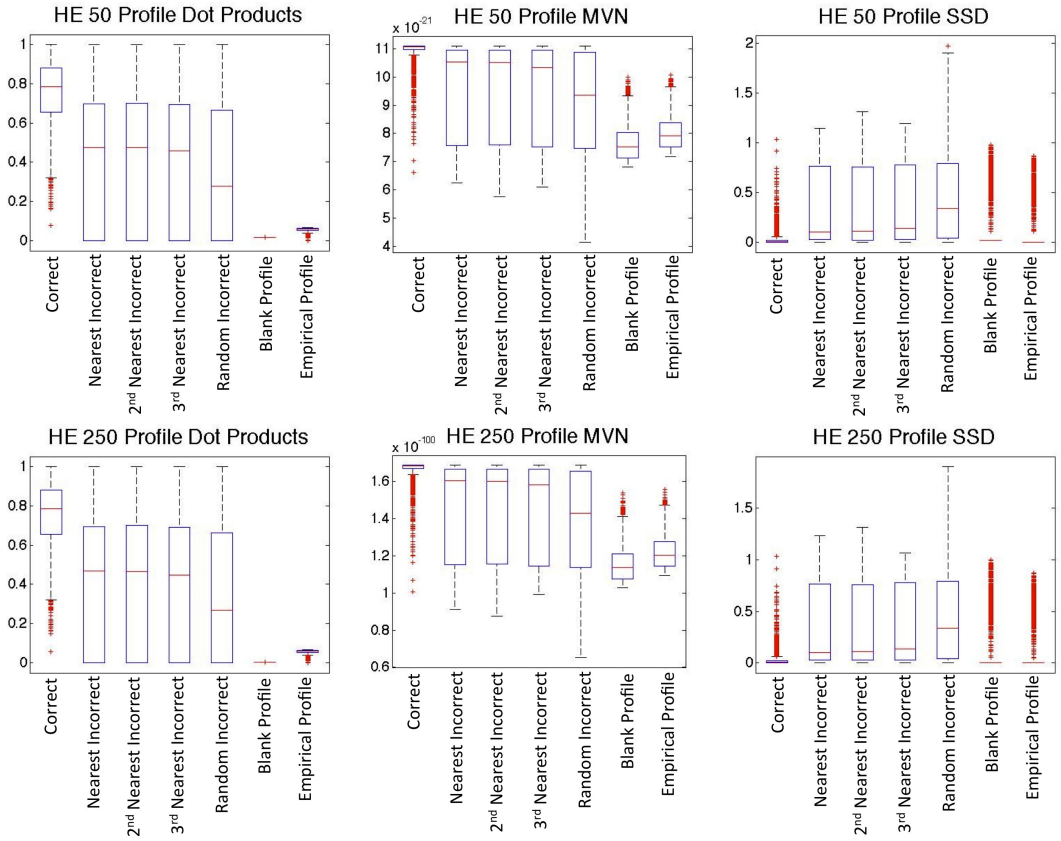

FIGURE A.1: Results of Additional High Energy Model Assessment. This figure shows boxplots of match scores from the Dot Product, Multivariate Normal PDF (MVN) and the sum of squared differences (SSD) metrics for two different profile sizes.

We see that in all three of the presented metrics, correct matches are favored above incorrect matches (larger values in the dot product and multivariate normal PDF, closest to zero in the sum of squared differences). However, in the dot product

and multivariate normal box plots, we see that the incorrect matches appear to be more favorable than the addition of a new latent peptide. This would be detrimental to our alignment results by encouraging mismatches. The correlation coefficients and norms gave similar results. The sum of squared differences metric gives us a desirable result, where correct matches and the addition of new latent peptides are more favorable than incorrect matches. With regard to product ion profile size, we saw only minor differences in performance with regards to match scores and compute times (Figure A.2).

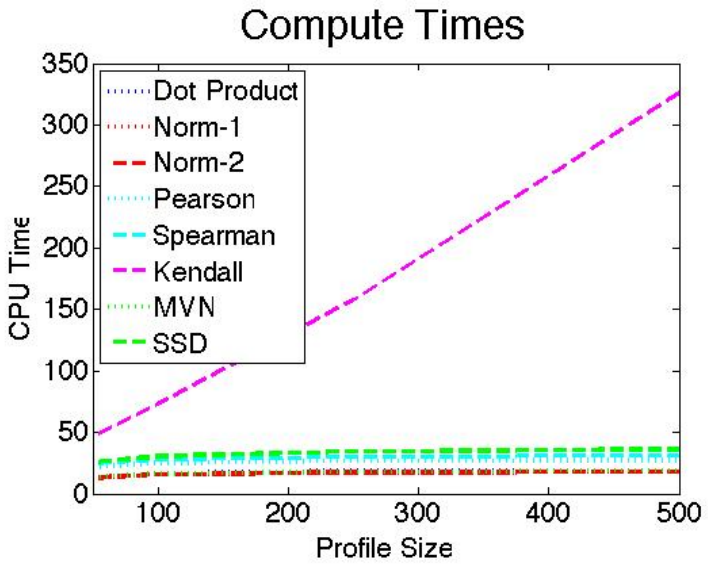

FIGURE A.2: Compute Times from High Energy Model Assessment. This figure shows the CPU time it took to compute the scores for the various metrics: Dot Product, 1-Norm, 2-Norm, Pearson Correlation, Spearman Correlation, Kendall Correlation, Multivariate Normal PDF (MVN) and the sum of squared differences (SSD) metrics for different profile sizes.

We decided to use K = 250 for our analyses to minimize the potential for overlapping product ions within a single peptides product ion profile, without a drastic increase in computational overhead.

A.4 Supplemental Results

A.4.1 Supplemental Results for E. coli Lysate Alignment

The recall rates and number of mismatches for the E. coli alignment have been presented considering identifications having peptide score 5 or greater. To assess the significance of the differences in these measures, we performed a series of two-sample t-tests assuming equal variance for the recall rates and incorrect matches to assess differences between alignments. The resulting p-values across the range of match probability stringencies are shown in Supplemental Tables A.1 and A.2.

Table A.1: Resulting P-values Comparing Recall Rates of Alignments.

Match Probability Cutoff	MZ-RT vs. MZ-IM	MZ-RT vs. MZ-RT-IM	MZ-IM vs. MZ-RT-IM	MZ-RT vs. MZ-RT-IM-HE	MZ-IM vs. MZ-RT-IM-HE	MZ-RT-IM vs. MZ-RT-IM-HE
$\geqslant 0$	1.4E-06	8.5E-05	8.9E-07	9.4E-01	8.3E-05	4.0E-03
$\geqslant 0.1$	1.8E-07	3.0E-05	6.9E-08	2.5E-02	1.1E-05	5.2E-03
$\geqslant 0.2$	6.6E-08	1.0E-05	3.6E-08	1.4E-03	3.4E-06	1.3E-02
$\geqslant 0.3$	4.8E-08	7.4E-07	2.6E-09	1.2E-04	1.5E-06	3.3E-01
$\geqslant 0.4$	1.0E-07	5.4E-07	1.5E-08	5.2E-05	1.2E-06	5.8E-02
$\geqslant 0.5$	2.2E-08	5.6E-07	3.7E-10	3.7E-05	8.7E-07	6.1E-03
$\geqslant 0.6$	4.3E-08	7.4E-07	7.3E-11	2.8E-05	8.0E-07	1.8E-03
$\geqslant 0.7$	1.6E-07	6.0E-06	2.3E-09	2.4E-05	6.8E-07	5.5E-04
$\geqslant 0.8$	2.6E-07	9.7E-06	1.1E-08	2.2E-05	6.8E-07	4.1E-04
$\geqslant 0.9$	3.2E-07	1.4E-05	2.2E-08	2.0E-05	6.6E-07	2.9E-04
$\geqslant 1$	8.9E-07	1.6E-05	1.7E-09	1.2E-04	3.7E-06	4.2E-03

Table A.2: Resulting P-values Comparing Mismatches of Alignments.

Match Probability Cutoff	MZ-RT vs. MZ-IM	MZ-RT vs. MZ-RT-IM	MZ-IM vs. MZ-RT-IM	MZ-RT vs. MZ-RT-IM-HE	MZ-IM vs. MZ-RT-IM-HE	MZ-RT-IM vs. MZ-RT-IM-HE
$\geqslant 0$	1.7E-02	1.1E-01	2.5E-03	2.2E-02	3.3E-03	6.8E-02
$\geqslant 0.1$	1.2E-02	9.4E-01	5.7E-03	8.1E-02	6.7E-01	7.2E-02
$\geqslant 0.2$	7.3E-03	1.1E-01	2.2E-04	4.3E-01	7.0E-02	6.9E-02
$\geqslant 0.3$	9.8E-04	1.8E-02	5.1E-05	8.7E-01	9.7E-03	1.1E-01
$\geqslant 0.4$	1.1E-03	8.2E-03	2.5E-05	4.5E-01	6.7E-03	1.4E-01
$\geqslant 0.5$	1.4E-03	6.9E-03	3.3E-05	2.9E-01	4.6E-03	1.4E-01
$\geqslant 0.6$	1.5E-03	1.1E-02	8.5E-05	2.8E-01	4.4E-03	1.9E-01
$\geqslant 0.7$	1.7E-03	8.2E-03	1.8E-04	1.9E-01	4.2E-03	2.5E-01
$\geqslant 0.8$	1.3E-03	5.8E-03	1.3E-04	1.4E-01	3.5E-03	3.2E-01
$\geqslant 0.9$	2.2E-03	1.0E-02	9.7E-05	1.2E-01	2.8E-03	4.7E-01
$\geqslant 1$	3.1E-03	1.4E-02	7.9E-05	1.6E-01	1.9E-03	2.5E-01

Figures A.3, A.4, A.5, and A.6 show the recall rates and mismatch counts when considering identifications having peptide score 6 or greater, and 7 or greater, respectively.

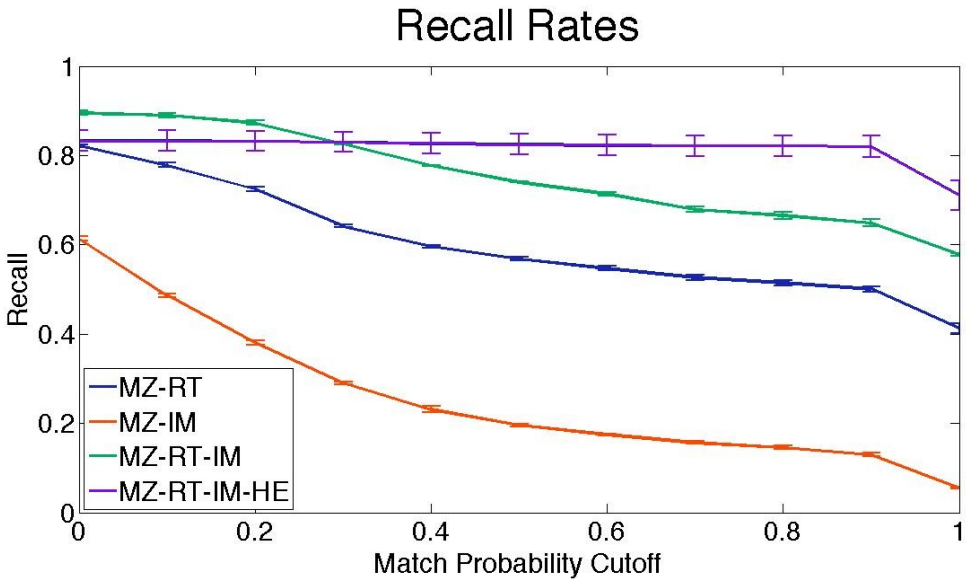

FIGURE A.3: Recall Rates for E. coli Lysate Data. This figure shows the recall rate considering identifications having peptide score 6 or greater, for each of the four alignments across a range of match probability cutoffs.

The results considering identifications of higher stringencies are consistent with those considering identifications having peptide score 5 or greater.

A.4.2 Supplemental Results for Decoy Experiment

We presented the results of an alignment of two technical replicates of human plasma samples with an E. coli lysate decoy. Figure A.7 shows the results of the inverse decoy analysis (aligning technical replicates of E. coli lysate with a Human plasma decoy).

A.4.3 Supplemental Results for Identification Carryover

By aligning datasets from different human tissues, we were able to infer identifications for several of the top peptides exhibiting differential expression for a phenotype of

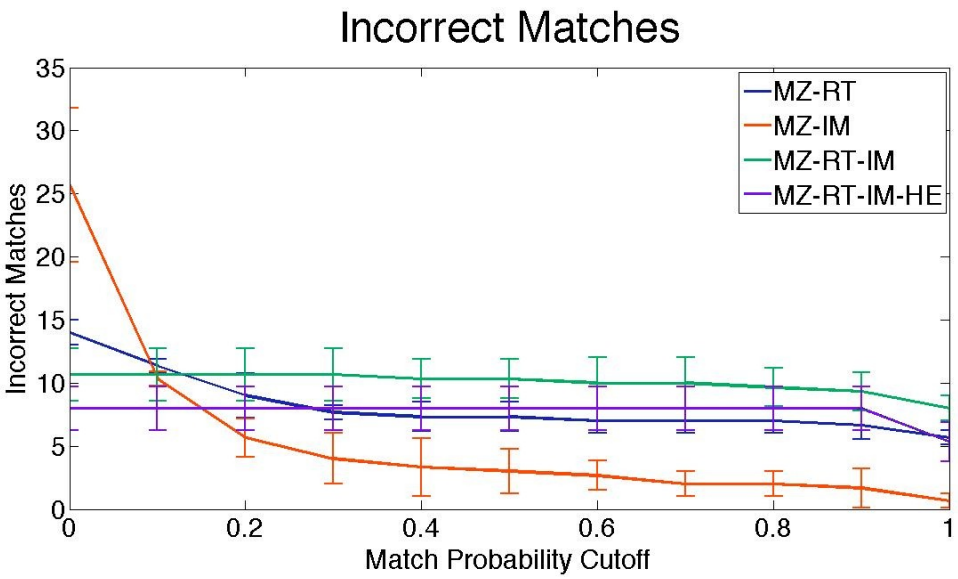

FIGURE A.4: Incorrect Matches for E. coli Lysate Data. This figure shows the number of incorrect matches considering identifications having peptide score 6 or greater, for each of the four alignments across a range of match probability cutoffs.

interest treatment response for the hepatitis-C data, and disease progression for the osteoarthritis data. The lists of proteins inferred by this analysis are shown in Tables A.3 and A.4.

We searched for functional enrichment of these proteins using GATHER and DAVID. The Top 15 GATHER Gene Ontology results on each of these protein sets are shown in Supplemental Tables 5 and 6. The Top 15 DAVID Biological Process Gene Ontology results are shown in Supplemental Tables 7 and 8. The GATHER chromosome location enrichment results for the Hepatitis-C inferred protein set are shown in Supplemental Table 9, and the DAVID KEGG Pathway results for the Osteoarthritis inferred protein set are shown in Supplemental Table 10.

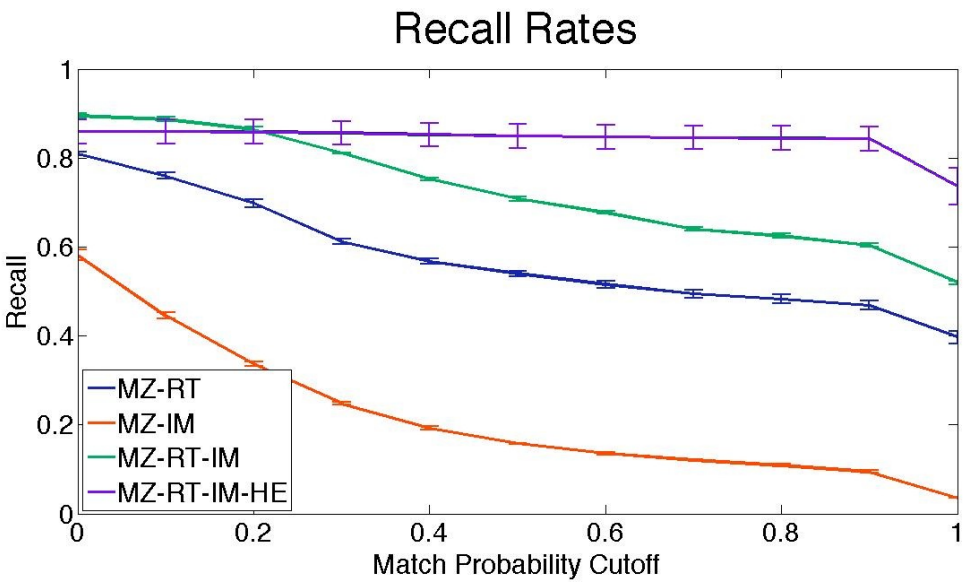

FIGURE A.5: Recall Rates for E. coli Lysate Data. This figure shows the recall rate considering identifications having peptide score 7 or greater, for each of the four alignments across a range of match probability cutoffs.

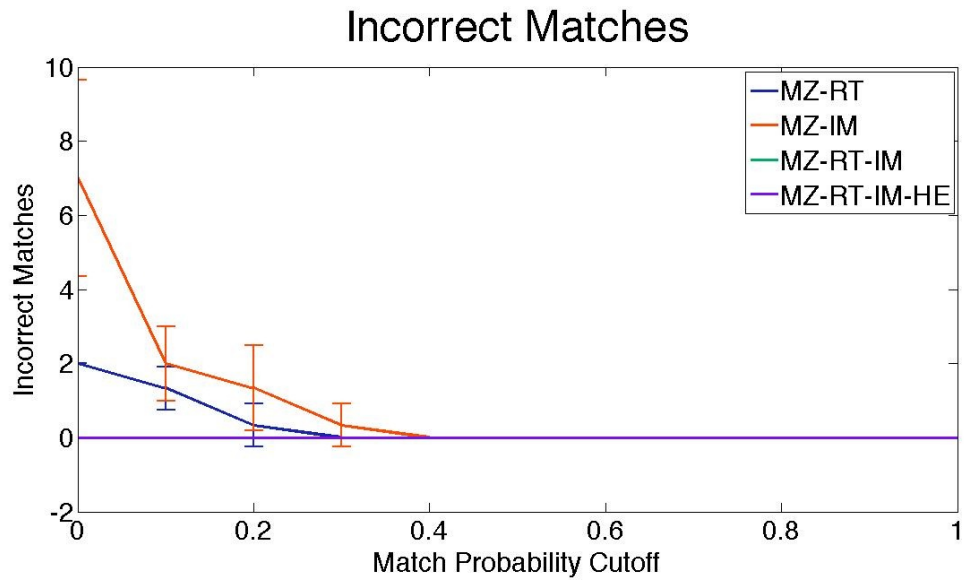

FIGURE A.6: Incorrect Matches for E. coli Lysate Data. This figure shows the number of incorrect matches considering identifications having peptide score 7 or greater, for each of the four alignments across a range of match probability cutoffs.

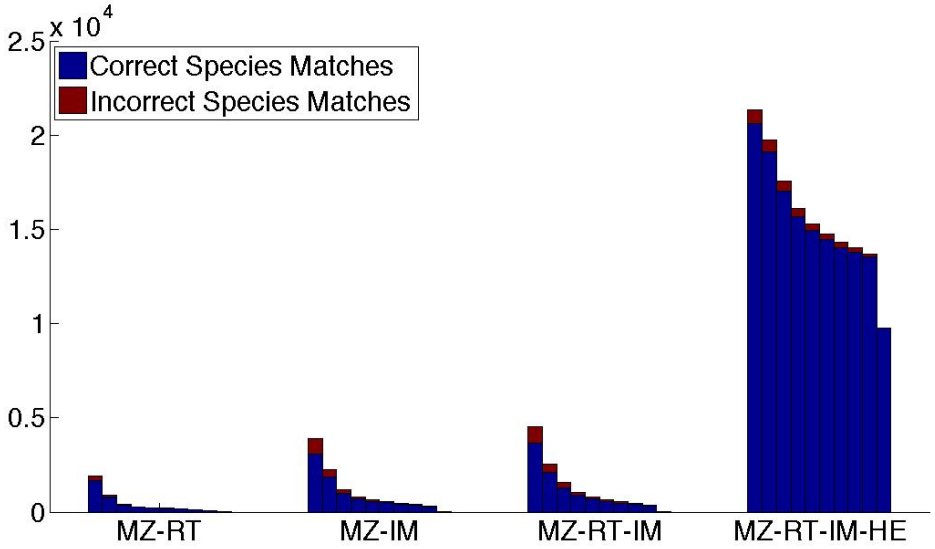

FIGURE A.7: Correct and Incorrect Matches for Decoy Analysis. This figure shows the number of matches made to the correct species (E. coli), and the number of matches made to the incorrect species (Human) for each of the four alignments, across increasing match confidence thresholds from 0.1 to 1 in 0.1 intervals.

Table A.3: List of Inferred Proteins Associated with Osteoarthritis Progression.

A1BG HUMAN	AACT HUMAN	AFAM HUMAN
ALS HUMAN	ANT3 HUMAN	APOA4 HUMAN
APOB HUMAN	APOC3 HUMAN	APOE HUMAN
APOH HUMAN	BTD HUMAN	C1QC HUMAN
C1R HUMAN	C1S HUMAN	CERU HUMAN
CFAB HUMAN	CFAH HUMAN	CFAI HUMAN
CLUS HUMAN	CO2 HUMAN	CO4B HUMAN
CO5 HUMAN	CO7 HUMAN	CO9 HUMAN
CPN2 HUMAN	FETUA HUMAN	FINC HUMAN
GELS HUMAN	HEMO HUMAN	HEP2 HUMAN
HRG HUMAN	IC1 HUMAN	ITIH1 HUMAN
ITIH2 HUMAN	ITIH3 HUMAN	ITIH4 HUMAN
K2C1 HUMAN	KNG1 HUMAN	PLMN HUMAN
RET4 HUMAN	SAMP HUMAN	TETN HUMAN
TEX15 HUMAN	VTDB HUMAN	ZA2G HUMAN
BP1 HUMAN		

Table A.4: List of Inferred Proteins Associated with Hepatitis-C Treatment Response.

A1AG1 HUMAN	A1AT HUMAN	A1BG HUMAN
A2GL HUMAN	ALBU HUMAN	AMYP HUMAN
APOD HUMAN	ARC HUMAN	B2MG HUMAN
CATD HUMAN	CD59 HUMAN	COFA1 HUMAN
CRNN HUMAN	CYTB HUMAN	DNAS1 HUMAN
EGF HUMAN	FETUA HUMAN	GUC2A HUMAN
HEMO HUMAN	HPT HUMAN	IBP7 HUMAN
IGHA1 HUMAN	IGHA2 HUMAN	IGHG1 HUMAN
IGHG2 HUMAN	IGHG4 HUMAN	IGKC HUMAN
ITIH4 HUMAN	KI26B HUMAN	KV206 HUMAN
LG3BP HUMAN	LYAG HUMAN	NID1 HUMAN
P3IP1 HUMAN	PIGR HUMAN	QPCT HUMAN
SAP3 HUMAN	SCTM1 HUMAN	SH3L3 HUMAN
TRFE HUMAN	UROM HUMAN	VASN HUMAN

Table A.5: GATHER Gene Ontology Results for Inferred Osteoarthritis Proteins.

Annotation	Your Genes (With Ann)	Genome (With Ann)	ln(Bayes factor)	neg ln(p value)	FE: neg ln(p value)	FE: neg ln(FDR)	Genes
GO:0006956 (6): complement activation	14	25	54.39	11.17	58.15	52.62	BF C1QG C1R C1S C2 C4A C5 C7 C9 CFH CLU IF KRT1 SERPING1
GO:0006958 (7): complement activation, classical pathway	11	19	42.05	11.17	45.82	40.99	C1QG C1R C1S C2 C4A C5 C7 C9 CLU IF SERPING1
GO:0009613 (5): response to pest, pathogen or parasite	20	451	35.1	11.17	38.84	34.41	AHSG APCS BF C1QG C1R C1S C2 C4A C5 C7 C9 CFH CLU FN1 IF ITIH4 KNG1 KRT1 SERPINA3 SERPING1
GO:0043207 (5): response to external biotic stimulus	20	482	33.86	11.17	37.6	33.45	AHSG APCS BF C1QG C1R C1S C2 C4A C5 C7 C9 CFH CLU FN1 IF ITIH4 KNG1 KRT1 SERPINA3 SERPING1
GO:0006959 (5): humoral immune response	14	149	32.56	11.17	36.32	32.4	BF C1QG C1R C1S C2 C4A C5 C7 C9 CFH CLU IF KRT1 SERPING1
GO:0006952 (5): defense response	23	815	32.04	11.17	35.76	32.02	AHSG APCS APOH AZGP1 BF C1QG C1R C1S C2 C4A C5 C7 C9 CFH CLU FN1 IF ITIH1 ITIH4 KNG1 KRT1 SERPINA3 SERPING1
GO:0009607 (4): response to biotic stimulus	24	934	31.81	11.17	35.52	31.98	AHSG APCS APOE APOH AZGP1 BF C1QG C1R C1S C2 C4A C5 C7 C9 CFH CLU FN1 IF ITIH1 ITIH4 KNG1 KRT1 SERPINA3 SERPING1

Annotation	Your Genes (With Ann)	Genome (With Ann)	ln(Bayes factor)	neg ln(p value)	FE: neg ln(p value)	FE: neg ln(FDR)	Genes
GO:0006955 (4): immune response	22	725	31.7	11.17	35.43	31.98	AHSG APCS AZGP1 BF C1QG C1R C1S C2 C4A C5 C7 C9 CFH CLU FN1 IF ITIH1 ITIH4 KNG1 KRT1 SERPINA3 SERPING1
GO:0006950 (4): response to stress	21	829	26.39	11.17	30.1	26.76	AHSG APCS APOE BF C1QG C1R C1S C2 C4A C5 C7 C9 CFH CLU FN1 IF ITIH4 KNG1 KRT1 SERPINA3 SERPING1
GO:0050874 (3): organismal physiological process	28	1816	26.15	11.17	29.83	26.61	AHSG APCS APOA4 APOB APOE AZGP1 BF C1QG C1R C1S C2 C4A C5 C7 C9 CFH CLU FN1 IF ITIH1 ITIH4 KNG1 KRT1 PLG SERPINA3 SERPINC1 SERPIND1 SERPING1
GO:0016064 (6): humoral defense mechanism (sensu Vertebrata)	11	108	25.51	11.17	29.26	26.13	C1QG C1R C1S C2 C4A C5 C7 C9 CLU IF SERPING1
GO:0009605 (4): response to external stimulus	21	1242	18.8	11.17	22.46	19.42	AHSG APCS BF C1QG C1R C1S C2 C4A C5 C7 C9 CFH CLU FN1 IF ITIH4 KNG1 KRT1 SERPINA3 SERPIND1 SERPING1
GO:0050896 (3): response to stimulus	25	1919	18.36	11.17	22.01	19.04	AHSG APCS APOE APOH AZGP1 BF C1QG C1R C1S C2 C4A C5 C7 C9 CFH CLU FN1 IF ITIH1 ITIH4 KNG1 KRT1 SERPINA3 SERPIND1 SERPING1
GO:0006957 (7): complement activation, alternative pathway	5	7	18.08	11.17	21.86	18.97	BF C5 C7 C9 CFH

Annotation	Your Genes (With Ann)	Genome (With Ann)	ln(Bayes factor)	neg ln(p value)	FE: neg ln(p value)	FE: neg ln(FDR)	Genes
GO:0030212 (8): hyaluronan metabolism	4	2	16.28	11.17	20.06	17.24	ITIH1 ITIH2 ITIH3 ITIH4

Table A.6: GATHER Gene Ontology Results for Inferred Hepatitis-C Proteins.

Annotation	Your Genes (With Ann)	Genome (With Ann)	ln(Bayes factor)	neg ln(p value)	FE: neg ln(p value)	FE: neg ln(FDR)	Genes
GO:0006952 (5): defense response	15	823	16.82	11.17	20.28	15.03	AHSG B2M CD59 HP IGHA1 IGHA2 IGHG1 IGHG2 IGHG4 IGKC ITIH4 LGALS3BP ORM1 SERPINA1 UMOD
GO:0006955 (4): immune response	14	733	15.94	11.17	19.4	14.84	AHSG B2M CD59 IGHA1 IGHA2 IGHG1 IGHG2 IGHG4 IGKC ITIH4 LGALS3BP ORM1 SERPINA1 UMOD
GO:0009607 (4): response to biotic stimulus	15	943	15	11.17	18.44	14.29	AHSG B2M CD59 HP IGHA1 IGHA2 IGHG1 IGHG2 IGHG4 IGKC ITIH4 LGALS3BP ORM1 SERPINA1 UMOD
GO:0006953 (5): acute-phase response	4	19	11.14	11.17	14.69	10.82	AHSG ITIH4 ORM1 SERPINA1
GO:0006879 (8): iron ion homeostasis	3	16	7.37	10.47	10.91	7.27	HP HPX TF
GO:0046916 (8): transition metal ion homeostasis	3	20	6.78	9.78	10.32	6.86	HP HPX TF
GO:0050874 (3): organismal physiological process	15	1829	6.67	9.78	9.99	6.68	AHSG ALB B2M CD59 IGHA1 IGHA2 IGHG1 IGHG2 IGHG4 IGKC ITIH4 LGALS3BP ORM1 SERPINA1 UMOD
GO:0050896 (3): response to stimulus	15	1929	6.05	9.09	9.36	6.19	AHSG B2M CD59 HP IGHA1 IGHA2 IGHG1 IGHG2 IGHG4 IGKC ITIH4 LGALS3BP ORM1 SERPINA1 UMOD

Annotation	Your Genes (With Ann)	Genome (With Ann)	ln(Bayes factor)	neg ln(p value)	FE: neg ln(p value)	FE: neg ln(FDR)	Genes
GO:0009987 (2): cellular process	22	10925	5.53	8.77	0	0	AHSG ALB APOD CD59 COL15A1 CTSD DNASE1 EGF GAA GM2A GUCA2A HP HPX IGFBP7 ITIH4 LGALS3BP NID ORM1 QPCT SECTM1 TF UMOD
GO:0042592 (3): homeostasis	4	98	5.17	8.53	8.67	5.61	ALB HP HPX TF
GO:0044237 (4): cellular metabolism	8	6557	4.65	8.4	0	0	CTSD DNASE1 EGF GAA GM2A HP ITIH4 QPCT
GO:0030005 (7): di-, trivalent inorganic cation homeostasis	3	46	4.5	8.28	8.02	5.07	HP HPX TF
GO:0006875 (7): metal ion homeostasis	3	51	4.22	8.12	7.74	4.88	HP HPX TF
GO:0008152 (3): metabolism	10	6955	3.55	7.11	0	0	AMY2A APOD CTSD DNASE1 EGF GAA GM2A HP ITIH4 QPCT
GO:0030003 (6): cation homeostasis	3	67	3.47	6.93	6.97	4.21	HP HPX TF

Table A.7: DAVID Gene Ontology Biological Process Results for Inferred Osteoarthritis Proteins.

Term	PValue	Genes	Fold Enrichment	Bonferroni	Benjamini	FDR
GO:0002526-acute inflammatory response	1.77E-28	FETUA HUMAN, CO5 HUMAN, C1QC HUMAN, IC1 HUMAN, CLUS HUMAN, ITIH4 HUMAN, CFAB HUMAN, CO2 HUMAN, CO7 HUMAN, CO4B HUMAN, SAMP HUMAN, C1S HUMAN, CFAH HUMAN, CO9 HUMAN, AACT HUMAN, CFAI HUMAN, K2C1 HUMAN, FINC HUMAN, C1R HUMAN	60.99	1.56E-25	1.56E-25	2.76E-25
GO:0006956-complement activation	7.54E-24	CO5 HUMAN, C1QC HUMAN, IC1 HUMAN, CLUS HUMAN, CO2 HUMAN, CFAB HUMAN, CO7 HUMAN, CO4B HUMAN, C1S HUMAN, CFAH HUMAN, CO9 HUMAN, CFAI HUMAN, K2C1 HUMAN, C1R HUMAN	104.87	6.63E-21	3.31E-21	1.17E-20

Term	PValue	Genes	Fold Enrichment	Bonferroni	Benjamini	FDR
GO:0002541-activation of plasma proteins involved in acute inflammatory response	1.08E-23	CO5 HUMAN, C1QC HUMAN, IC1 HUMAN, CLUS HUMAN, CO2 HUMAN, CFAB HUMAN, CO7 HUMAN, CO4B HUMAN, C1S HUMAN, CFAH HUMAN, CO9 HUMAN, CFAI HUMAN, K2C1 HUMAN, C1R HUMAN	102.43	9.48E-21	3.16E-21	1.68E-20
GO:0009611-response to wounding	1.83E-23	GELS HUMAN, ITIH4 HUMAN, CFAB HUMAN, CO2 HUMAN, CO7 HUMAN, CO4B HUMAN, KNG1 HUMAN, SAMP HUMAN, PLMN HUMAN, C1S HUMAN, CFAH HUMAN, CO9 HUMAN, ANT3 HUMAN, AACT HUMAN, K2C1 HUMAN, C1R HUMAN, FETUA HUMAN, CO5 HUMAN, HEP2 HUMAN, C1QC HUMAN, IC1 HUMAN, CLUS HUMAN, APOH HUMAN, CFAI HUMAN, FINC HUMAN	14.84	1.61E-20	4.02E-21	2.84E-20

Term	PValue	Genes	Fold Enrichment	Bonferroni	Benjamini	FDR
GO:0051605-protein maturation by peptide bond cleavage	2.64E-21	CO5 HUMAN, C1QC HUMAN, IC1 HUMAN, CLUS HUMAN, CO2 HUMAN, CFAB HUMAN, CO7 HUMAN, CO4B HUMAN, APOH HUMAN, C1S HUMAN, CFAH HUMAN, CO9 HUMAN, CFAI HUMAN, K2C1 HUMAN, C1R HUMAN	54.87	2.32E-18	4.64E-19	4.10E-18
GO:0006954-inflammatory response	2.67E-20	FETUA HUMAN, CO5 HUMAN, C1QC HUMAN, IC1 HUMAN, CLUS HUMAN, ITIH4 HUMAN, CFAB HUMAN, CO2 HUMAN, CO7 HUMAN, CO4B HUMAN, KNG1 HUMAN, SAMP HUMAN, C1S HUMAN, CFAH HUMAN, CO9 HUMAN, AACT HUMAN, CFAI HUMAN, K2C1 HUMAN, FINC HUMAN, C1R HUMAN	19.36	2.34E-17	3.91E-18	4.14E-17
GO:0006959-humoral immune response	7.20E-20	CO5 HUMAN, C1QC HUMAN, IC1 HUMAN, CLUS HUMAN, CO2 HUMAN, CFAB HUMAN, CO7 HUMAN, CO4B HUMAN, C1S HUMAN, CFAH HUMAN, CO9 HUMAN, CFAI HUMAN, K2C1 HUMAN, C1R HUMAN	55.75	6.33E-17	9.04E-18	1.12E-16

Term	PValue	Genes	Fold Enrichment	Bonferroni	Benjamini	FDR
GO:0016485-protein processing	1.32E-19	CO5 HUMAN, C1QC HUMAN, IC1 HUMAN, CLUS HUMAN, CO2 HUMAN, CFAB HUMAN, CO7 HUMAN, CO4B HUMAN, APOH HUMAN, C1S HUMAN, CFAH HUMAN, CO9 HUMAN, CFAI HUMAN, K2C1 HUMAN, C1R HUMAN	42.13	1.16E-16	1.45E-17	2.05E-16
GO:0051604-protein maturation	4.61E-19	CO5 HUMAN, C1QC HUMAN, IC1 HUMAN, CLUS HUMAN, CO2 HUMAN, CFAB HUMAN, CO7 HUMAN, CO4B HUMAN, APOH HUMAN, C1S HUMAN, CFAH HUMAN, CO9 HUMAN, CFAI HUMAN, K2C1 HUMAN, C1R HUMAN	38.68	4.05E-16	4.50E-17	7.16E-16
GO:0006958-complement activation, classical pathway	5.00E-19	CO5 HUMAN, C1QC HUMAN, IC1 HUMAN, C1S HUMAN, CLUS HUMAN, CO9 HUMAN, CO2 HUMAN, CO7 HUMAN, CO4B HUMAN, CFAI HUMAN, C1R HUMAN	119.33	4.39E-16	4.39E-17	7.77E-16

Term	PValue	Genes	Fold Enrichment	Bonferroni	Benjamini	FDR
GO:0002253-activation of immune response	7.96E-19	CO5 HUMAN, C1QC HUMAN, IC1 HUMAN, CLUS HUMAN, CO2 HUMAN, CFAB HUMAN, CO7 HUMAN, CO4B HUMAN, C1S HUMAN, CFAH HUMAN, CO9 HUMAN, CFAI HUMAN, K2C1 HUMAN, C1R HUMAN	46.86	6.99E-16	6.36E-17	1.24E-15
GO:0002455-humoral immune response mediated by circulating immunoglobulin	1.10E-18	CO5 HUMAN, C1QC HUMAN, IC1 HUMAN, C1S HUMAN, CLUS HUMAN, CO9 HUMAN, CO2 HUMAN, CO7 HUMAN, CO4B HUMAN, CFAI HUMAN, C1R HUMAN	111.63	9.68E-16	8.07E-17	1.71E-15
GO:0045087-innate immune response	2.75E-18	CO5 HUMAN, C1QC HUMAN, IC1 HUMAN, CLUS HUMAN, CO2 HUMAN, CFAB HUMAN, CO7 HUMAN, CO4B HUMAN, C1S HUMAN, CFAH HUMAN, CO9 HUMAN, CFAI HUMAN, APOA4 HUMAN, K2C1 HUMAN, C1R HUMAN	34.20	2.41E-15	1.86E-16	4.27E-15

Term	PValue	Genes	Fold Enrichment	Bonferroni	Benjamini	FDR
GO:0050778-positive regulation of immune response	5.60E-18	CO5 HUMAN, C1QC HUMAN, IC1 HUMAN, CLUS HUMAN, CO2 HUMAN, CFAB HUMAN, HEMO HUMAN, CO7 HUMAN, CO4B HUMAN, C1S HUMAN, CFAH HUMAN, CO9 HUMAN, CFAI HUMAN, K2C1 HUMAN, C1R HUMAN	32.55	4.92E-15	3.51E-16	8.70E-15
GO:0002252-immune effector process	9.58E-17	CO5 HUMAN, C1QC HUMAN, IC1 HUMAN, CLUS HUMAN, CO2 HUMAN, CFAB HUMAN, CO7 HUMAN, CO4B HUMAN, C1S HUMAN, CFAH HUMAN, CO9 HUMAN, CFAI HUMAN, K2C1 HUMAN, C1R HUMAN	32.87	9.76E-14	6.55E-15	1.78E-13

Table A.8: DAVID Gene Ontology Biological Process Results for Inferred Hepatitis-C Proteins.

Term	PValue	Genes	Fold Enrichment	Bonferroni	Benjamini	FDR
GO:0006953-acute-phase response	2.52E-06	FETUA HUMAN, A1AT HUMAN, ITIH4 HUMAN, A1AG1 HUMAN, TRFE HUMAN	49.74	1.15E-03	1.15E-03	3.58E-03
GO:0002526-acute inflammatory response	9.02E-05	FETUA HUMAN, A1AT HUMAN, ITIH4 HUMAN, A1AG1 HUMAN, TRFE HUMAN	20.30	4.02E-02	2.03E-02	1.28E-01
GO:0006952-defense response	5.89E-04	FETUA HUMAN, A1AT HUMAN, ITIH4 HUMAN, A1AG1 HUMAN, HPT HUMAN, UROM HUMAN, TRFE HUMAN, LG3BP HUMAN	5.18	2.35E-01	8.55E-02	8.34E-01
GO:0006955-immune response	1.16E-03	IGHA1 HUMAN, IGHG1 HUMAN, SCTM1 HUMAN, IGHG2 HUMAN, IGHA2 HUMAN, IGHG4 HUMAN, IGKC HUMAN, B2MG HUMAN	4.61	4.11E-01	1.24E-01	1.64E+00
GO:0006879-cellular iron ion homeostasis	2.57E-03	HPT HUMAN, HEMO HUMAN, TRFE HUMAN	38.50	6.90E-01	2.09E-01	3.59E+00
GO:0055072-iron ion homeostasis	3.45E-03	HPT HUMAN, HEMO HUMAN, TRFE HUMAN	33.16	7.93E-01	2.31E-01	4.79E+00
GO:0006954-inflammatory response	7.72E-03	FETUA HUMAN, A1AT HUMAN, ITIH4 HUMAN, A1AG1 HUMAN, TRFE HUMAN	6.12	9.71E-01	3.96E-01	1.04E+01

Term	PValue	Genes	Fold Enrichment	Bonferroni	Benjamini	FDR
GO:0009611-response to wounding	8.67E-03	FETUA HUMAN, A1AT HUMAN, ITIH4 HUMAN, A1AG1 HUMAN, CD59 HUMAN, TRFE HUMAN	4.50	9.81E-01	3.91E-01	1.16E+01
GO:0019725-cellular homeostasis	2.58E-02	LYAG HUMAN, SH3L3 HUMAN, HPT HUMAN, HEMO HUMAN, TRFE HUMAN	4.27	1.00E+00	7.34E-01	3.10E+01
GO:0014070-response to organic cyclic substance	3.50E-02	IBP7 HUMAN, A1AT HUMAN, TRFE HUMAN	9.86	1.00E+00	8.02E-01	3.97E+01
GO:0018298-protein-chromophore linkage	3.83E-02	IGHA1 HUMAN, NID1 HUMAN	49.74	1.00E+00	8.02E-01	4.26E+01
GO:0002824-positive regulation of adaptive immune response based on somatic recombination of immune receptors built from immunoglobulin superfamily domains	7.07E-02	HEMO HUMAN, B2MG HUMAN	26.53	1.00E+00	9.38E-01	6.47E+01
GO:0009311-oligosaccharide metabolic process	7.30E-02	LYAG HUMAN, SAP3 HUMAN	25.67	1.00E+00	9.30E-01	6.59E+01
GO:0002821-positive regulation of adaptive immune response	7.30E-02	HEMO HUMAN, B2MG HUMAN	25.67	1.00E+00	9.30E-01	6.59E+01
GO:0009267-cellular response to starvation	7.53E-02	CATD HUMAN, ALBU HUMAN	24.87	1.00E+00	9.21E-01	6.71E+01

Table A.9: GATHER Chromosome Location Results for Inferred Hepatitis-C Proteins.

Annotation	Your Genes (With Ann)	Genome (With Ann)	ln(Bayes factor)	neg ln(p value)	FE: neg ln(p value)	FE: neg ln(FDR)	Genes
14q32	6	413	7.69	10.09	11.27	7.98	IGHA1 IGHA2 IGHG1 IGHG2 IGHG4 SERPINA1
17q25	3	213	2.4	6.01	5.99	3.39	GAA LGALS3BP SECTM1
1p35	2	99	1.34	4.93	4.94	2.74	GUCA2A SH3BGRL3
16p13	3	351	1.07	4.64	4.62	2.71	DNASE1 MRX85 UMOD
4q11	1	16	0.21	3.79	3.86	2.17	ALB
1q43	1	29	-0.33	0	3.3	1.82	NID
4q12	1	49	-0.81	0	2.8	1.82	IGFBP7
11p13	1	55	-0.92	0	2.69	1.82	CD59
2p22	1	56	-0.94	0	2.67	1.82	QPCT
4q25	1	56	-0.94	0	2.67	1.82	EGF
3q22	1	58	-0.94	0	2.64	1.82	TF
1q31	1	59	-0.99	0	2.62	1.82	PIGR
1p21	1	64	-1.04	0	2.55	1.82	AMY2A
3q27	1	75	-1.23	0	2.4	1.74	AHSG
11p15	2	451	-1.26	0	2.21	1.72	CTSD HPX
9q31	1	87	-1.34	0	2.26	1.72	ORM1
9q21	1	94	-1.42	0	2.19	1.72	COL15A1
15q21	1	112	-1.53	0	2.02	1.69	B2M
3q26	1	109	-1.53	0	2.05	1.69	APOD
2p12	1	116	-1.56	0	1.99	1.69	IGKC
16q22	1	171	-1.89	0	1.64	1.39	HP
8q24	1	222	-2.08	0	1.41	1.2	ARC
5q31	1	248	-2.16	0	1.31	1.16	GM2A

Annotation	Your Genes (With Ann)	Genome (With Ann)	ln(Bayes factor)	neg ln(p value)	FE: neg ln(p value)	FE: neg ln(FDR)	Genes
3p21	1	257	-2.2	0	1.28	1.16	ITIH4
21q22	1	295	-2.3	0	1.17	1.09	CSTB
19p13	1	658	-2.64	0	0.57	0.53	LRG1
19q13	1	966	-2.64	0	0.34	0.34	A1BG

Table A.10: DAVID KEGG Pathway Results for Inferred Osteoarthritis Proteins.

Term	PValue	Genes	Fold Enrichment	Bonferroni	Benjamini	FDR
hsa04610:Complement and coagulation cascades	3.05E-24	HEP2 HUMAN, CO5 HUMAN, C1QC HUMAN, IC1 HUMAN, CFAB HUMAN, CO2 HUMAN, CO7 HUMAN, CO4B HUMAN, KNG1 HUMAN, PLMN HUMAN, C1S HUMAN, CFAH HUMAN, CO9 HUMAN, ANT3 HUMAN, CFAI HUMAN, C1R HUMAN	51.27	4.57E-23	4.57E-23	2.12E-21
hsa05322:Systemic lupus erythematosus	1.15E-07	CO5 HUMAN, C1QC HUMAN, C1S HUMAN, CO9 HUMAN, CO2 HUMAN, CO7 HUMAN, CO4B HUMAN, C1R HUMAN	17.87	1.73E-06	8.63E-07	8.00E-05
hsa05020:Prion diseases	4.21E-04	CO5 HUMAN, C1QC HUMAN, CO9 HUMAN, CO7 HUMAN	25.27	6.29E-03	2.10E-03	2.92E-01

Bibliography

1000 Genomes Project Consortium, G. R. Abecasis, A. Auton, L. D. Brooks, M. A. DePristo, R. M. Durbin, R. E. Handsaker, H. M. Kang, G. T. Marth, and G. A. McVean (2012, Nov). An integrated map of genetic variation from 1,092 human genomes. *Nature 491*(7422), 56–65.

Aird, D., M. G. Ross, W.-S. Chen, M. Danielsson, T. Fennell, C. Russ, D. B. Jaffe, C. Nusbaum, and A. Gnirke (2011). Analyzing and minimizing pcr amplification bias in illumina sequencing libraries. *Genome Biol 12*(2), R18.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman (1990, Oct). Basic local alignment search tool. *J Mol Biol 215*(3), 403–10.

Anders, S. and W. Huber (2010). Differential expression analysis for sequence count data. *Genome Biol 11*(10), R106.

Andreev, V. P., T. Rejtar, H. S. Chen, E. V. Moskovets, A. R. Ivanov, and B. L. Karger (2003). A universal denoising and peak picking algorithm for lc-ms based on matched filtration in the chromatographic time domain. *Anal Chem 75*(22), 6314–26. Andreev, Victor P Rejtar, Tomas Chen, Hsuan-Shen Moskovets, Eugene V Ivanov, Alexander R Karger, Barry L GM 15847/GM/NIGMS NIH HHS/United States HG 02033/HG/NHGRI NIH HHS/United States Research Support, U.S. Gov't, P.H.S. United States Analytical chemistry Anal Chem. 2003 Nov 15;75(22):6314-26.

Andrews, S. (2010). Fastqc. a quality control tool for high throughput sequence data. *URL http://www. bioinformatics. babraham. ac. uk/projects/fastqc*.

Ashburner, M., C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock (2000, May). Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nat Genet 25*(1), 25–9.

Atwood, L. D., N. L. Heard-Costa, C. S. Fox, C. E. Jaquish, and L. A. Cupples (2006). Sex and age specific effects of chromosomal regions linked to body mass index in the framingham study. *BMC Genet 7*, 7. Atwood, Larry D Heard-Costa, Nancy

L Fox, Caroline S Jaquish, Cashell E Cupples, L Adrienne HC-25195/HC/NHLBI NIH HHS/ R01 DK066241/DK/NIDDK NIH HHS/ England BMC Genet. 2006 Jan 26;7:7.

Au, K. F., H. Jiang, L. Lin, Y. Xing, and W. H. Wong (2010, Aug). Detection of splice junctions from paired-end rna-seq data by splicemap. *Nucleic Acids Res 38*(14), 4570–8.

AV, N., M. I, B. AR, and S. RG (2011). Examining troughs in the mass distribution of all theoretically possible tryptic peptides. *J Proteome Res.. 10*, 4150–4157.

Bacanu, S. A., B. Devlin, and K. Roeder (2002). Association studies for quantitative traits in structured populations. *Genet Epidemiol 22*(1), 78–93. Bacanu, Silviu-Alin Devlin, Bernie Roeder, Kathryn MH 56193/MH/NIMH NIH HHS/ MH 57881/MH/NIMH NIH HHS/ R01 MH057881-15/MH/NIMH NIH HHS/ Genet Epidemiol. 2002 Jan;22(1):78-93.

Balwierz, P. J., P. Carninci, C. O. Daub, J. Kawai, Y. Hayashizaki, W. Van Belle, C. Beisel, and E. van Nimwegen (2009). Methods for analyzing deep sequencing expression data: constructing the human and mouse promoterome with deepcage data. *Genome Biol 10*(7), R79.

Bamshad, M. J., S. B. Ng, A. W. Bigham, H. K. Tabor, M. J. Emond, D. A. Nickerson, and J. Shendure (2011, Nov). Exome sequencing as a tool for mendelian disease gene discovery. *Nat Rev Genet 12*(11), 745–55.

Barrett, J. C., B. Fry, J. Maller, and M. J. Daly (2005). Haploview: analysis and visualization of ld and haplotype maps. *Bioinformatics 21*(2), 263–5. Barrett, J C Fry, B Maller, J Daly, M J England Oxford, England Bioinformatics. 2005 Jan 15;21(2):263-5. Epub 2004 Aug 5.

Barski, A., S. Cuddapah, K. Cui, T.-Y. Roh, D. E. Schones, Z. Wang, G. Wei, I. Chepelev, and K. Zhao (2007). High-resolution profiling of histone methylations in the human genome. *Cell 129*(4), 823–837.

Bell, A. W., E. W. Deutsch, C. E. Au, R. E. Kearney, R. Beavis, S. Sechi, T. Nilsson, and J. J. Bergeron (2009). A hupo test sample study reveals common problems in mass spectrometry-based proteomics. *Nat Methods 6*(6), 423–30. Bell, Alexander W Deutsch, Eric W Au, Catherine E Kearney, Robert E Beavis, Ron Sechi, Salvatore Nilsson, Tommy Bergeron, John J M HUPO Test Sample Working Group N01-HV-28179/HV/NHLBI NIH HHS/United States P01-008111/PHS HHS/United States P41RR018627/RR/NCRR NIH HHS/United States R01 CA120393-03/CA/NCI NIH HHS/United States RR020843/RR/NCRR NIH HHS/United States Canadian Institutes of Health Research/Canada Evaluation Studies Research Support, N.I.H., Extramural Research Support, Non-U.S. Gov't United States Nature methods Nat Methods. 2009 Jun;6(6):423-30.

Bellew, M., M. Coram, M. Fitzgibbon, M. Igra, T. Randolph, P. Wang, D. May, J. Eng, R. H. Fang, C. W. Lin, J. Z. Chen, D. Goodlett, J. Whiteaker, A. Paulovich, and M. McIntosh (2006). A suite of algorithms for the comprehensive analysis of complex protein mixtures using high-resolution lc-ms. *Bioinformatics 22*(15), 1902–1909. 075QG Times Cited:98 Cited References Count:25.

Benjamini, Y. and T. P. Speed (2012, May). Summarizing and correcting the gc content bias in high-throughput sequencing. *Nucleic Acids Res 40*(10), e72.

Bernstein, B. E., A. Meissner, and E. S. Lander (2007). The mammalian epigenome. *Cell 128*(4), 669–681.

Birner-Gruenberger, R., H. Susani-Etzerodt, M. Waldhuber, G. Riesenhuber, H. Schmidinger, G. Rechberger, M. Kollroser, J. G. Strauss, A. Lass, R. Zimmermann, G. Haemmerle, R. Zechner, and A. Hermetter (2005). The lipolytic proteome of mouse adipose tissue. *Mol Cell Proteomics 4*(11), 1710–7. Birner-Gruenberger, Ruth Susani-Etzerodt, Heidrun Waldhuber, Markus Riesenhuber, Gernot Schmidinger, Hannes Rechberger, Gerald Kollroser, Manfred Strauss, Juliane G Lass, Achim Zimmermann, Robert Haemmerle, Guenter Zechner, Rudolf Hermetter, Albin Mol Cell Proteomics. 2005 Nov;4(11):1710-7. Epub 2005 Jul 26.

Bullard, J. H., E. Purdom, K. D. Hansen, and S. Dudoit (2010). Evaluation of statistical methods for normalization and differential expression in mrna-seq experiments. *BMC Bioinformatics 11*, 94.

Bulow, J., K. Gjeraa, L. H. Enevoldsen, and L. Simonsen (2006). Lipid mobilization from human abdominal, subcutaneous adipose tissue is independent of sex during steady-state exercise. *Clin Physiol Funct Imaging 26*(4), 205–11. Bulow, Jens Gjeraa, Kirsten Enevoldsen, Lotte H Simonsen, Lene England Clin Physiol Funct Imaging. 2006 Jul;26(4):205-11.

Butler, J., I. MacCallum, M. Kleber, I. A. Shlyakhter, M. K. Belmonte, E. S. Lander, C. Nusbaum, and D. B. Jaffe (2008, May). Allpaths: de novo assembly of whole-genome shotgun microreads. *Genome Res 18*(5), 810–20.

Cairns, B. R. (2009). The logic of chromatin architecture and remodelling at promoters. *Nature 461*(7261), 193–198.

Camps, M., A. Nichols, and S. Arkinstall (2000). Dual specificity phosphatases: a gene family for control of map kinase function. *FASEB J 14*(1), 6–16. Camps, M Nichols, A Arkinstall, S FASEB J. 2000 Jan;14(1):6-16.

Carninci, P., A. Sandelin, B. Lenhard, S. Katayama, K. Shimokawa, J. Ponjavic, C. A. Semple, M. S. Taylor, P. G. Engström, M. C. Frith, et al. (2006). Genome-wide analysis of mammalian promoter architecture and evolution. *Nature genetics 38*(6), 626–635.

Chang, J. T. and J. R. Nevins (2006). Gather: a systems approach to interpreting genomic signatures. *Bioinformatics 22*(23), 2926–33. Chang, Jeffrey T Nevins, Joseph R 5-U54-CA112952/CA/NCI NIH HHS/ England Oxford, England Bioinformatics. 2006 Dec 1;22(23):2926-33. Epub 2006 Sep 25.

Cho, I. and M. J. Blaser (2012, Apr). The human microbiome: at the interface of health and disease. *Nat Rev Genet 13*(4), 260–70.

Cupples, L. A., H. T. Arruda, E. J. Benjamin, S. D'Agostino, R. B., S. Demissie, A. L. DeStefano, J. Dupuis, K. M. Falls, C. S. Fox, D. J. Gottlieb, D. R. Govindaraju, C. Y. Guo, N. L. Heard-Costa, S. J. Hwang, S. Kathiresan, D. P. Kiel, J. M. Laramie, M. G. Larson, D. Levy, C. Y. Liu, K. L. Lunetta, M. D. Mailman, A. K. Manning, J. B. Meigs, J. M. Murabito, C. Newton-Cheh, G. T. O'Connor, C. J. O'Donnell, M. Pandey, S. Seshadri, R. S. Vasan, Z. Y. Wang, J. B. Wilk, P. A. Wolf, Q. Yang, and L. D. Atwood (2007). The framingham heart study 100k snp genome-wide association study resource: overview of 17 phenotype working group reports. *BMC Med Genet 8 Suppl 1*, S1. Cupples, L Adrienne Arruda, Heather T Benjamin, Emelia J D'Agostino, Ralph B Sr Demissie, Serkalem DeStefano, Anita L Dupuis, Josee Falls, Kathleen M Fox, Caroline S Gottlieb, Daniel J Govindaraju, Diddahally R Guo, Chao-Yu Heard-Costa, Nancy L Hwang, Shih-Jen Kathiresan, Sekar Kiel, Douglas P Laramie, Jason M Larson, Martin G Levy, Daniel Liu, Chun-Yu Lunetta, Kathryn L Mailman, Matthew D Manning, Alisa K Meigs, James B Murabito, Joanne M Newton-Cheh, Christopher O'Connor, George T O'Donnell, Christopher J Pandey, Mona Seshadri, Sudha Vasan, Ramachandran S Wang, Zhen Y Wilk, Jemma B Wolf, Philip A Yang, Qiong Atwood, Larry D 1R01 AG028321/AG/NIA NIH HHS/ 1S10RR163736-01A1/RR/NCRR NIH HHS/ 5R01-AG08122/AG/NIA NIH HHS/ 5R01-AG16495/AG/NIA NIH HHS/ HL54776/HL/NHLBI NIH HHS/ K24 HL 04334/HL/NHLBI NIH HHS/ N01-HC 25195/HC/NHLBI NIH HHS/ England BMC Med Genet. 2007;8 Suppl 1:S1.

Davey, C., S. Pennings, and J. Allan (1997). Cpg methylation remodels chromatin structure in vitro. *Journal of molecular biology 267*(2), 276–288.

Davey, C. S., S. Pennings, C. Reilly, R. R. Meehan, and J. Allan (2004). A determining influence for cpg dinucleotides on nucleosome positioning in vitro. *Nucleic acids research 32*(14), 4322–4331.

David, D., J. Cardoso, B. Marques, R. Marques, E. D. Silva, H. Santos, and M. G. Boavida (2003). Molecular characterization of a familial translocation implicates disruption of hdac9 and possible position effect on tgfbeta2 in the pathogenesis of peters' anomaly. *Genomics 81*(5), 489–503. David, Dezso Cardoso, Joana Marques, Barbara Marques, Ramira Silva, Eduardo D Santos, Heloisa Boavida, Maria G Genomics. 2003 May;81(5):489-503.

De Bona, F., S. Ossowski, K. Schneeberger, and G. Rätsch (2008, Aug). Optimal spliced alignments of short sequence reads. *Bioinformatics 24*(16), i174–80.

Degner, J. F., J. C. Marioni, A. A. Pai, J. K. Pickrell, E. Nkadori, Y. Gilad, and J. K. Pritchard (2009, Dec). Effect of read-mapping biases on detecting allele-specific expression from rna-sequencing data. *Bioinformatics 25*(24), 3207–12.

DeLuca, D. S., J. Z. Levin, A. Sivachenko, T. Fennell, M.-D. Nazaire, C. Williams, M. Reich, W. Winckler, and G. Getz (2012, Jun). Rna-seqc: Rna-seq metrics for quality control and process optimization. *Bioinformatics 28*(11), 1530–2.

Dennis, G., J., B. T. Sherman, D. A. Hosack, J. Yang, W. Gao, H. C. Lane, and R. A. Lempicki (2003). David: Database for annotation, visualization, and integrated discovery. *Genome Biol 4*(5), P3. Dennis, Glynn Jr Sherman, Brad T Hosack, Douglas A Yang, Jun Gao, Wei Lane, H Clifford Lempicki, Richard A N01-C0-56000/PHS HHS/ England Genome Biol. 2003;4(5):P3. Epub 2003 Apr 3.

Devlin, B. and K. Roeder (1999). Genomic control for association studies. *Biometrics 55*(4), 997–1004. Devlin, B Roeder, K NIMH 56193/MH/NIMH NIH HHS/ NIMH 57881/MH/NIMH NIH HHS/ R01 MH057881-15/MH/NIMH NIH HHS/ Biometrics. 1999 Dec;55(4):997-1004.

Didelot, X., R. Bowden, D. J. Wilson, T. E. A. Peto, and D. W. Crook (2012, Sep). Transforming clinical microbiology with bacterial genome sequencing. *Nat Rev Genet 13*(9), 601–12.

Dillies, M.-A., A. Rau, J. Aubert, C. Hennequet-Antier, M. Jeanmougin, N. Servant, C. Keime, G. Marot, D. Castel, J. Estelle, et al. (2012). A comprehensive evaluation of normalization methods for illumina high-throughput rna sequencing data analysis. *Briefings in Bioinformatics*.

Dowsey, A. W., J. A. English, F. Lisacek, J. S. Morris, G. Z. Yang, and M. J. Dunn (2010). Image analysis tools and emerging algorithms for expression proteomics. *Proteomics 10*(23), 4226–57. Dowsey, Andrew W English, Jane A Lisacek, Frederique Morris, Jeffrey S Yang, Guang-Zhong Dunn, Michael J CA107304/CA/NCI NIH HHS/United States R01 CA107304-05/CA/NCI NIH HHS/United States R01 CA107304-06/CA/NCI NIH HHS/United States R01 CA107304-07/CA/NCI NIH HHS/United States Research Support, N.I.H., Extramural Research Support, Non-U.S. Gov't Review Germany Proteomics Proteomics. 2010 Dec;10(23):4226-57.

Du, P., R. Sudha, M. B. Prystowsky, and R. H. Angeletti (2007). Data reduction of isotope-resolved lc-ms spectra. *Bioinformatics 23*(11), 1394–400. Du, Peicheng Sudha, Rajagopalan Prystowsky, Michael B Angeletti, Ruth Hogue CA101150/CA/NCI NIH HHS/United States CA103547/CA/NCI NIH

HHS/United States Research Support, N.I.H., Extramural England Bioinformatics (Oxford, England) Bioinformatics. 2007 Jun 1;23(11):1394-400. Epub 2007 May 11.

Engström, P. G., S. J. H. Sui, Ø. Drivenes, T. S. Becker, and B. Lenhard (2007). Genomic regulatory blocks underlie extensive microsynteny conservation in insects. *Genome research 17*(12), 1898–1908.

Fischer, B., J. Grossmann, V. Roth, W. Gruissem, S. Baginsky, and J. M. Buhmann (2006). Semi-supervised lc/ms alignment for differential proteomics. *Bioinformatics 22*(14), e132–40. Fischer, Bernd Grossmann, Jonas Roth, Volker Gruissem, Wilhelm Baginsky, Sacha Buhmann, Joachim M Research Support, Non-U.S. Gov't England Bioinformatics (Oxford, England) Bioinformatics. 2006 Jul 15;22(14):e132-40.

Fox, C. S., N. Heard-Costa, L. A. Cupples, J. Dupuis, R. S. Vasan, and L. D. Atwood (2007). Genome-wide association to body mass index and waist circumference: the framingham heart study 100k project. *BMC Med Genet 8 Suppl 1*, S18. Fox, Caroline S Heard-Costa, Nancy Cupples, L Adrienne Dupuis, Josee Vasan, Ramachandran S Atwood, Larry D 1S10RR163736-01A1/RR/NCRR NIH HHS/ N01-HC-25195/HC/NHLBI NIH HHS/ R01 DK066241/DK/NIDDK NIH HHS/ England BMC Med Genet. 2007 Sep 19;8 Suppl 1:S18.

Fu, Y., M. Sinha, C. L. Peterson, and Z. Weng (2008). The insulator binding protein ctcf positions 20 nucleosomes around its binding sites across the human genome. *PLoS genetics 4*(7), e1000138.

Ganapathi, M., P. Srivastava, S. Sutar, K. Kumar, D. Dasgupta, G. P. Singh, V. Brahmachari, and S. Brahmachari (2005). Comparative analysis of chromatin landscape in regulatory regions of human housekeeping and tissue specific genes. *BMC bioinformatics 6*(1), 126.

Garber, M., M. G. Grabherr, M. Guttman, and C. Trapnell (2011). Computational methods for transcriptome annotation and quantification using rna-seq. *Nature methods 8*(6), 469–477.

Gardiner-Garden, M. and M. Frommer (1987). Cpg islands in vertebrate genomes. *Journal of molecular biology 196*(2), 261–282.

Ge, D., K. Zhang, A. C. Need, O. Martin, J. Fellay, T. J. Urban, A. Telenti, and D. B. Goldstein (2008). Wgaviewer: software for genomic annotation of whole genome association studies. *Genome Research 18*(4), 640–3. Ge, Dongliang Zhang, Kunlin Need, Anna C Martin, Olivier Fellay, Jacques Urban, Thomas J Telenti, Amalio Goldstein, David B Genome Res. 2008 Apr;18(4):640-3. doi: 10.1101/gr.071571.107. Epub 2008 Feb 6.

Gillet, L. C., P. Navarro, S. Tate, H. Röst, N. Selevsek, L. Reiter, R. Bonner, and R. Aebersold (2012, Jun). Targeted data extraction of the ms/ms spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis. *Mol Cell Proteomics 11*(6), O111.016717.

Goldstein, D. B., A. Allen, J. Keebler, E. H. Margulies, S. Petrou, S. Petrovski, and S. Sunyaev (2013, Jul). Sequencing studies in human genetics: design and interpretation. *Nat Rev Genet 14*(7), 460–70.

Govindaraju, D. R., L. A. Cupples, W. B. Kannel, C. J. O'Donnell, L. D. Atwood, S. D'Agostino, R. B., C. S. Fox, M. Larson, D. Levy, J. Murabito, R. S. Vasan, G. L. Splansky, P. A. Wolf, and E. J. Benjamin (2008). Genetics of the framingham heart study population. *Adv Genet 62*, 33–65. Govindaraju, Diddahally R Cupples, L Adrienne Kannel, William B O'Donnell, Christopher J Atwood, Larry D D'Agostino, Ralph B Sr Fox, Caroline S Larson, Marty Levy, Daniel Murabito, Joanne Vasan, Ramachandran S Splansky, Greta Lee Wolf, Philip A Benjamin, Emelia J AG028321/AG/NIA NIH HHS/ N01 HC025195/HC/NHLBI NIH HHS/ N01-HC 25195/HC/NHLBI NIH HHS/ R01 AG028321-01/AG/NIA NIH HHS/ R01 HL076784/HL/NHLBI NIH HHS/ R01 HL076784-01/HL/NHLBI NIH HHS/ Adv Genet. 2008;62:33-65. doi: 10.1016/S0065-2660(08)00602-0.

Grabherr, M. G., B. J. Haas, M. Yassour, J. Z. Levin, D. A. Thompson, I. Amit, X. Adiconis, L. Fan, R. Raychowdhury, Q. Zeng, Z. Chen, E. Mauceli, N. Hacohen, A. Gnirke, N. Rhind, F. di Palma, B. W. Birren, C. Nusbaum, K. Lindblad-Toh, N. Friedman, and A. Regev (2011, Jul). Full-length transcriptome assembly from rna-seq data without a reference genome. *Nat Biotechnol 29*(7), 644–52.

Guttman, M., M. Garber, J. Z. Levin, J. Donaghey, J. Robinson, X. Adiconis, L. Fan, M. J. Koziol, A. Gnirke, C. Nusbaum, J. L. Rinn, E. S. Lander, and A. Regev (2010, May). Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincrnas. *Nat Biotechnol 28*(5), 503–10.

Hansen, K. D., S. E. Brenner, and S. Dudoit (2010, Jul). Biases in illumina transcriptome sequencing caused by random hexamer priming. *Nucleic Acids Res 38*(12), e131.

Hansen, K. D., R. A. Irizarry, and Z. Wu (2012, Apr). Removing technical variability in rna-seq data using conditional quantile normalization. *Biostatistics 13*(2), 204–16.

Hastings, C. A., S. M. Norton, and S. Roy (2002). New algorithms for processing and peak detection in liquid chromatography/mass spectrometry data. *Rapid Commun Mass Spectrom 16*(5), 462–7. Hastings, Curtis A Norton, Scott M Roy, Sushmita England Rapid communications in mass spectrometry : RCM Rapid Commun Mass Spectrom. 2002;16(5):462-7.

Heid, I. M., A. U. Jackson, J. C. Randall, T. W. Winkler, L. Qi, V. Steinthorsdottir, G. Thorleifsson, M. C. Zillikens, E. K. Speliotes, R. Magi, T. Workalemahu, C. C. White, N. Bouatia-Naji, T. B. Harris, S. I. Berndt, E. Ingelsson, C. J. Willer, M. N. Weedon, J. Luan, S. Vedantam, T. Esko, T. O. Kilpelainen, Z. Kutalik, S. Li, K. L. Monda, A. L. Dixon, C. C. Holmes, L. M. Kaplan, L. Liang, J. L. Min, M. F. Moffatt, C. Molony, G. Nicholson, E. E. Schadt, K. T. Zondervan, M. F. Feitosa, T. Ferreira, H. Lango Allen, R. J. Weyant, E. Wheeler, A. R. Wood, K. Estrada, M. E. Goddard, G. Lettre, M. Mangino, D. R. Nyholt, S. Purcell, A. V. Smith, P. M. Visscher, J. Yang, S. A. McCarroll, J. Nemesh, B. F. Voight, D. Absher, N. Amin, T. Aspelund, L. Coin, N. L. Glazer, C. Hayward, N. L. Heard-Costa, J. J. Hottenga, A. Johansson, T. Johnson, M. Kaakinen, K. Kapur, S. Ketkar, J. W. Knowles, P. Kraft, A. T. Kraja, C. Lamina, M. F. Leitzmann, B. McKnight, A. P. Morris, K. K. Ong, J. R. Perry, M. J. Peters, O. Polasek, I. Prokopenko, N. W. Rayner, S. Ripatti, F. Rivadeneira, N. R. Robertson, S. Sanna, U. Sovio, I. Surakka, A. Teumer, S. van Wingerden, V. Vitart, J. H. Zhao, C. Cavalcanti-Proenca, P. S. Chines, E. Fisher, J. R. Kulzer, C. Lecoeur, N. Narisu, C. Sandholt, L. J. Scott, K. Silander, K. Stark, M. L. Tammesoo, et al. (2010). Meta-analysis identifies 13 new loci associated with waist-hip ratio and reveals sexual dimorphism in the genetic basis of fat distribution. *Nat Genet 42*(11), 949–60. Heid, Iris M Jackson, Anne U Randall, Joshua C Winkler, Thomas W Qi, Lu Steinthorsdottir, Valgerdur Thorleifsson, Gudmar Zillikens, M Carola Speliotes, Elizabeth K Magi, Reedik Workalemahu, Tsegaselassie White, Charles C Bouatia-Naji, Nabila Harris, Tamara B Berndt, Sonja I Ingelsson, Erik Willer, Cristen J Weedon, Michael N Luan, Jian'an Vedantam, Sailaja Esko, Tonu Kilpelainen, Tuomas O Kutalik, Zoltan Li, Shengxu Monda, Keri L Dixon, Anna L Holmes, Christopher C Kaplan, Lee M Liang, Liming Min, Josine L Moffatt, Miriam F Molony, Cliona Nicholson, George Schadt, Eric E Zondervan, Krina T Feitosa, Mary F Ferreira, Teresa Lango Allen, Hana Weyant, Robert J Wheeler, Eleanor Wood, Andrew R MAGIC Estrada, Karol Goddard, Michael E Lettre, Guillaume Mangino, Massimo Nyholt, Dale R Purcell, Shaun Smith, Albert Vernon Visscher, Peter M Yang, Jian McCarroll, Steven A Nemesh, James Voight, Benjamin F Absher, Devin Amin, Najaf Aspelund, Thor Coin, Lachlan Glazer, Nicole L Hayward, Caroline Heard-Costa, Nancy L Hottenga, Jouke-Jan Johansson, Asa Johnson, Toby Kaakinen, Marika Kapur, Karen Ketkar, Shamika Knowles, Joshua W Kraft, Peter Kraja, Aldi T Lamina, Claudia Leitzmann, Michael F McKnight, Barbara Morris, Andrew P Ong, Ken K Perry, John R B Peters, Marjolein J Polasek, Ozren Prokopenko, Inga Rayner, Nigel W Ripatti, Samuli Rivadeneira, Fernando Robertson, Neil R Sanna, Serena Sovio, Ulla Surakka, Ida Teumer, Alexander van Wingerden, Sophie Vitart, Veronique Zhao, Jing Hua Cavalcanti-Proenca, Christine Chines, Peter S Fisher, Eva Kulzer, Jennifer R Lecoeur, Cecile Narisu, Narisu Sandholt, Camilla Scott, Laura J Silander, Kaisa Stark, Klaus Tammesoo, Mari-Liis Teslovich, Tanya M Timpson, Nicholas John Watanabe, Richard M Welch, Ryan Chasman, Daniel I Cooper, Matthew N Jansson, John-Olov Kettunen, Johannes

Lawrence, Robert W Pellikka, Niina Perola, Markus Vandenput, Liesbeth Alavere, Helene Almgren, Peter Atwood, Larry D Bennett, Amanda J Biffar, Reiner Bonnycastle, Lori L Bornstein, Stefan R Buchanan, Thomas A Campbell, Harry Day, Ian N M Dei, Mariano Dorr, Marcus Elliott, Paul Erdos, Michael R Eriksson, Johan G Freimer, Nelson B Fu, Mao Gaget, Stefan Geus, Eco J C Gjesing, Anette P Grallert, Harald Grassler, Jurgen Groves, Christopher J Guiducci, Candace Hartikainen, Anna-Liisa Hassanali, Neelam Havulinna, Aki S Herzig, Karl-Heinz Hicks, Andrew A Hui, Jennie Igl, Wilmar Jousilahti, Pekka Jula, Antti Kajantie, Eero Kinnunen, Leena Kolcic, Ivana Koskinen, Seppo Kovacs, Peter Kroemer, Heyo K Krzelj, Vjekoslav Kuusisto, Johanna Kvaloy, Kirsti Laitinen, Jaana Lantieri, Olivier Lathrop, G Mark Lokki, Marja-Liisa Luben, Robert N Ludwig, Barbara McArdle, Wendy L McCarthy, Anne Morken, Mario A Nelis, Mari Neville, Matt J Pare, Guillaume Parker, Alex N Peden, John F Pichler, Irene Pietilainen, Kirsi H Platou, Carl G P Pouta, Anneli Ridderstrale, Martin Samani, Nilesh J Saramies, Jouko Sinisalo, Juha Smit, Jan H Strawbridge, Rona J Stringham, Heather M Swift, Amy J Teder-Laving, Maris Thomson, Brian Usala, Gianluca van Meurs, Joyce B J van Ommen, Gert-Jan Vatin, Vincent Volpato, Claudia B Wallaschofski, Henri Walters, G Bragi Widen, Elisabeth Wild, Sarah H Willemsen, Gonneke Witte, Daniel R Zgaga, Lina Zitting, Paavo Beilby, John P James, Alan L Kahonen, Mika Lehtimaki, Terho Nieminen, Markku S Ohlsson, Claes Palmer, Lyle J Raitakari, Olli Ridker, Paul M Stumvoll, Michael Tonjes, Anke Viikari, Jorma Balkau, Beverley Ben-Shlomo, Yoav Bergman, Richard N Boeing, Heiner Smith, George Davey Ebrahim, Shah Froguel, Philippe Hansen, Torben Hengstenberg, Christian Hveem, Kristian Isomaa, Bo Jorgensen, Torben Karpe, Fredrik Khaw, Kay-Tee Laakso, Markku Lawlor, Debbie A Marre, Michel Meitinger, Thomas Metspalu, Andres Midthjell, Kristian Pedersen, Oluf Salomaa, Veikko Schwarz, Peter E H Tuomi, Tiinamaija Tuomilehto, Jaakko Valle, Timo T Wareham, Nicholas J Arnold, Alice M Beckmann, Jacques S Bergmann, Sven Boerwinkle, Eric Boomsma, Dorret I Caulfield, Mark J Collins, Francis S Eiriksdottir, Gudny Gudnason, Vilmundur Gyllensten, Ulf Hamsten, Anders Hattersley, Andrew T Hofman, Albert Hu, Frank B Illig, Thomas Iribarren, Carlos Jarvelin, Marjo-Riitta Kao, W H Linda Kaprio, Jaakko Launer, Lenore J Munroe, Patricia B Oostra, Ben Penninx, Brenda W Pramstaller, Peter P Psaty, Bruce M Quertermous, Thomas Rissanen, Aila Rudan, Igor Shuldiner, Alan R Soranzo, Nicole Spector, Timothy D Syvanen, Ann-Christine Uda, Manuela Uitterlinden, Andre Volzke, Henry Vollenweider, Peter Wilson, James F Witteman, Jacqueline C Wright, Alan F Abecasis, Goncalo R Boehnke, Michael Borecki, Ingrid B Deloukas, Panos Frayling, Timothy M Groop, Leif C Haritunians, Talin Hunter, David J Kaplan, Robert C North, Kari E O'Connell, Jeffrey R Peltonen, Leena Schlessinger, David Strachan, David P Hirschhorn, Joel N Assimes, Themistocles L Wichmann, H-Erich Thorsteinsdottir, Unnur van Duijn, Cornelia M Stefansson, Kari Cupples, L Adrienne Loos, Ruth J F Barroso, Ines McCarthy, Mark I Fox, Caroline S Mohlke, Karen L Lindgren, Cecilia M Wellcome Trust United Kingdom

2010 Nov;42(11):949-60. doi: 10.1038/ng.685. Epub 2010 Oct 10.

Herbert, A., N. P. Gerry, M. B. McQueen, I. M. Heid, A. Pfeufer, T. Illig, H. E. Wichmann, T. Meitinger, D. Hunter, F. B. Hu, G. Colditz, A. Hinney, J. Hebebrand, K. Koberwitz, X. Zhu, R. Cooper, K. Ardlie, H. Lyon, J. N. Hirschhorn, N. M. Laird, M. E. Lenburg, C. Lange, and M. F. Christman (2006). A common genetic variant is associated with adult and childhood obesity. *Science 312*(5771), 279–83. Herbert, Alan Gerry, Norman P McQueen, Matthew B Heid, Iris M Pfeufer, Arne Illig, Thomas Wichmann, H-Erich Meitinger, Thomas Hunter, David Hu, Frank B Colditz, Graham Hinney, Anke Hebebrand, Johannes Koberwitz, Kerstin Zhu, Xiaofeng Cooper, Richard Ardlie, Kristin Lyon, Helen Hirschhorn, Joel N Laird, Nan M Lenburg, Marc E Lange, Christoph Christman, Michael F CA87969/CA/NCI NIH HHS/ K23DK067288/DK/NIDDK NIH HHS/ P30DK46200/DK/NIDDK NIH HHS/ R01 HD060726/HD/NICHD NIH HHS/ R01GM046877/GM/NIGMS NIH HHS/ R01HL074166/HL/NHLBI NIH HHS/ R01HL54485/HL/NHLBI NIH HHS/ R01HL66289/HL/NHLBI NIH HHS/ R01MH59532/MH/NIMH NIH HHS/ U01HL65899/HL/NHLBI NIH HHS/ New York, N.Y. Science. 2006 Apr 14;312(5771):279-83.

Hoggart, C. J., E. J. Parra, M. D. Shriver, C. Bonilla, R. A. Kittles, D. G. Clayton, and P. M. McKeigue (2003, Jun). Control of confounding of genetic associations in stratified populations. *Am J Hum Genet 72*(6), 1492–1504.

Hornett, E. A. and C. W. Wheat (2012). Quantitative rna-seq analysis in non-model species: assessing transcriptome assemblies as a scaffold and the utility of evolutionary divergent genomic reference species. *BMC Genomics 13*, 361.

Hubbard, T. J., B. L. Aken, S. Ayling, B. Ballester, K. Beal, E. Bragin, S. Brent, Y. Chen, P. Clapham, L. Clarke, G. Coates, S. Fairley, S. Fitzgerald, J. Fernandez-Banet, L. Gordon, S. Graf, S. Haider, M. Hammond, R. Holland, K. Howe, A. Jenkinson, N. Johnson, A. Kahari, D. Keefe, S. Keenan, R. Kinsella, F. Kokocinski, E. Kulesha, D. Lawson, I. Longden, K. Megy, P. Meidl, B. Overduin, A. Parker, B. Pritchard, D. Rios, M. Schuster, G. Slater, D. Smedley, W. Spooner, G. Spudich, S. Trevanion, A. Vilella, J. Vogel, S. White, S. Wilder, A. Zadissa, E. Birney, F. Cunningham, V. Curwen, R. Durbin, X. M. Fernandez-Suarez, J. Herrero, A. Kasprzyk, G. Proctor, J. Smith, S. Searle, and P. Flicek (2009). Ensembl 2009. *Nucleic Acids Res 37*(Database issue), D690–7. Hubbard, T J P Aken, B L Ayling, S Ballester, B Beal, K Bragin, E Brent, S Chen, Y Clapham, P Clarke, L Coates, G Fairley, S Fitzgerald, S Fernandez-Banet, J Gordon, L Graf, S Haider, S Hammond, M Holland, R Howe, K Jenkinson, A Johnson, N Kahari, A Keefe, D Keenan, S Kinsella, R Kokocinski, F Kulesha, E Lawson, D Longden, I Megy, K Meidl, P Overduin, B Parker, A Pritchard, B Rios, D Schuster, M Slater, G Smedley, D Spooner, W Spudich, G Trevanion, S Vilella, A Vogel, J White, S

Wilder, S Zadissa, A Birney, E Cunningham, F Curwen, V Durbin, R Fernandez-Suarez, X M Herrero, J Kasprzyk, A Proctor, G Smith, J Searle, S Flicek, P 062023/Wellcome Trust/United Kingdom 077198/Wellcome Trust/United Kingdom BBE0116401/Biotechnology and Biological Sciences Research Council/United Kingdom WT062023/Wellcome Trust/United Kingdom Biotechnology and Biological Sciences Research Council/United Kingdom Medical Research Council/United Kingdom England Nucleic Acids Res. 2009 Jan;37(Database issue):D690-7. doi: 10.1093/nar/gkn828. Epub 2008 Nov 25.

Ioshikhes, I. P., I. Albert, S. J. Zanton, and B. F. Pugh (2006). Nucleosome positions predicted through comparative genomics. *Nature genetics 38*(10), 1210–1215.

Jaffe, J. D., D. R. Mani, K. C. Leptos, G. M. Church, M. A. Gillette, and S. A. Carr (2006). Pepper, a platform for experimental proteomic pattern recognition. *Mol Cell Proteomics 5*(10), 1927–41. Jaffe, Jacob D Mani, D R Leptos, Kyriacos C Church, George M Gillette, Michael A Carr, Steven A R01 CA126219/CA/NCI NIH HHS/United States Research Support, N.I.H., Extramural Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, Non-P.H.S. United States Molecular and cellular proteomics : MCP Mol Cell Proteomics. 2006 Oct;5(10):1927-41. Epub 2006 Jul 19.

Jeffries, N. (2005). Algorithms for alignment of mass spectrometry proteomic data. *Bioinformatics 21*(14), 3066–73. Jeffries, Neal Comparative Study Evaluation Studies England Bioinformatics (Oxford, England) Bioinformatics. 2005 Jul 15;21(14):3066-73. Epub 2005 May 6.

Jin, C., C. Zang, G. Wei, K. Cui, W. Peng, K. Zhao, and G. Felsenfeld (2009). H3. 3/h2a. z double variant–containing nucleosomes mark'nucleosome-free regions' of active promoters and other regulatory regions. *Nature genetics 41*(8), 941–945.

Juven-Gershon, T. and J. T. Kadonaga (2010). Regulation of gene expression via the core promoter and the basal transcriptional machinery. *Developmental biology 339*(2), 225–229.

Karpievitch, Y. V., A. R. Dabney, and R. D. Smith (2012). Normalization and missing value imputation for label-free lc-ms analysis. *BMC Bioinformatics 13 Suppl 16*, S5.

Karpievitch, Y. V., T. Taverner, J. N. Adkins, S. J. Callister, G. A. Anderson, R. D. Smith, and A. R. Dabney (2009, Oct). Normalization of peak intensities in bottom-up ms-based proteomics using singular value decomposition. *Bioinformatics 25*(19), 2573–80.

Katajamaa, M., J. Miettinen, and M. Oresic (2006). Mzmine: toolbox for processing and visualization of mass spectrometry based molecular profile data. *Bioinformatics 22*(5), 634–6. Katajamaa, Mikko Miettinen, Jarkko Oresic, Matej Research

Support, Non-U.S. Gov't England Bioinformatics (Oxford, England) Bioinformatics. 2006 Mar 1;22(5):634-6. Epub 2006 Jan 10.

Kawaji, H., J. Severin, M. Lizio, A. Waterhouse, S. Katayama, K. M. Irvine, D. A. Hume, A. Forrest, H. Suzuki, P. Carninci, et al. (2009). The fantom web resource: from mammalian transcriptional landscape to its dynamic regulation. *Genome Biol 10*(4), R40.

Kim, D., G. Pertea, C. Trapnell, H. Pimentel, R. Kelley, and S. L. Salzberg (2013). Tophat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome biology 14*(4), R36.

König, J., K. Zarnack, N. M. Luscombe, and J. Ule (2011, Feb). Protein-rna interactions: new genomic technologies and perspectives. *Nat Rev Genet 13*(2), 77–83.

Kratz, A., E. Arner, R. Saito, A. Kubosaki, J. Kawai, H. Suzuki, P. Carninci, T. Arakawa, M. Tomita, Y. Hayashizaki, et al. (2010). Core promoter structure and genomic context reflect histone 3 lysine 9 acetylation patterns. *BMC genomics 11*(1), 257.

Laird, P. W. (2010). Principles and challenges of genome-wide dna methylation analysis. *Nature Reviews Genetics 11*(3), 191–203.

Lam, H. (2011, Dec). Building and searching tandem mass spectral libraries for peptide identification. *Mol Cell Proteomics 10*(12), R111.008565.

Lange, E., C. Gropl, O. Schulz-Trieglaff, A. Leinenbach, C. Huber, and K. Reinert (2007). A geometric approach for the alignment of liquid chromatography-mass spectrometry data. *Bioinformatics 23*(13), i273–81. Lange, Eva Gropl, Clemens Schulz-Trieglaff, Ole Leinenbach, Andreas Huber, Christian Reinert, Knut Evaluation Studies Research Support, Non-U.S. Gov't England Bioinformatics (Oxford, England) Bioinformatics. 2007 Jul 1;23(13):i273-81.

Lange, E., R. Tautenhahn, S. Neumann, and C. Gropl (2008). Critical assessment of alignment procedures for lc-ms proteomics and metabolomics measurements. *BMC Bioinformatics 9*, 375. Lange, Eva Tautenhahn, Ralf Neumann, Steffen Gropl, Clemens Cancer Research UK/United Kingdom Research Support, Non-U.S. Gov't England BMC bioinformatics BMC Bioinformatics. 2008 Sep 15;9:375.

Langmead, B. and S. L. Salzberg (2012, Apr). Fast gapped-read alignment with bowtie 2. *Nat Methods 9*(4), 357–9.

Langmead, B., C. Trapnell, M. Pop, and S. L. Salzberg (2009). Ultrafast and memory-efficient alignment of short dna sequences to the human genome. *Genome Biol 10*(3), R25.

Lassmann, T., Y. Hayashizaki, and C. O. Daub (2011, Jan). Samstat: monitoring biases in next generation sequencing data. *Bioinformatics 27*(1), 130–1.

Leptos, K. C., D. A. Sarracino, J. D. Jaffe, B. Krastins, and G. M. Church (2006). Mapquant: open-source software for large-scale protein quantification. *Proteomics 6*(6), 1770–82. Leptos, Kyriacos C Sarracino, David A Jaffe, Jacob D Krastins, Bryan Church, George M Research Support, U.S. Gov't, Non-P.H.S. Germany Proteomics Proteomics. 2006 Mar;6(6):1770-82.

Li, B., V. Ruotti, R. M. Stewart, J. A. Thomson, and C. N. Dewey (2010, Feb). Rna-seq gene expression estimation with read mapping uncertainty. *Bioinformatics 26*(4), 493–500.

Li, G.-Z., J. P. Vissers, J. C. Silva, D. Golick, M. V. Gorenstein, and S. J. Geromanos (2009). Database searching and accounting of multiplexed precursor and product ion spectra from the data independent analysis of simple and complex peptide mixtures. *Proteomics 9*, 1696–1719.

Li, H. and R. Durbin (2009, Jul). Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics 25*(14), 1754–60.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, et al. (2009). The sequence alignment/map format and samtools. *Bioinformatics 25*(16), 2078–2079.

Li, R., Y. Li, K. Kristiansen, and J. Wang (2008, Mar). Soap: short oligonucleotide alignment program. *Bioinformatics 24*(5), 713–4.

Li, R., C. Yu, Y. Li, T.-W. Lam, S.-M. Yiu, K. Kristiansen, and J. Wang (2009, Aug). Soap2: an improved ultrafast tool for short read alignment. *Bioinformatics 25*(15), 1966–7.

Li, X. J., E. C. Yi, C. J. Kemp, H. Zhang, and R. Aebersold (2005). A software suite for the generation and comparison of peptide arrays from sets of data collected by liquid chromatography-mass spectrometry. *Mol Cell Proteomics 4*(9), 1328–40. Li, Xiao-jun Yi, Eugene C Kemp, Christopher J Zhang, Hui Aebersold, Ruedi 1R21CA114852/CA/NCI NIH HHS/United States N01-CO-12400/CO/NCI NIH HHS/United States N01-HV-28179/HV/NHLBI NIH HHS/United States U01-ES-11045/ES/NIEHS NIH HHS/United States Comparative Study Research Support, N.I.H., Extramural Research Support, U.S. Gov't, P.H.S. United States Molecular and cellular proteomics : MCP Mol Cell Proteomics. 2005 Sep;4(9):1328-40. Epub 2005 Jul 26.

Lindgren, C. M., I. M. Heid, J. C. Randall, C. Lamina, V. Steinthorsdottir, L. Qi, E. K. Speliotes, G. Thorleifsson, C. J. Willer, B. M. Herrera, A. U. Jackson, N. Lim, P. Scheet, N. Soranzo, N. Amin, Y. S. Aulchenko, J. C. Chambers,

A. Drong, J. Luan, H. N. Lyon, F. Rivadeneira, S. Sanna, N. J. Timpson, M. C. Zillikens, J. H. Zhao, P. Almgren, S. Bandinelli, A. J. Bennett, R. N. Bergman, L. L. Bonnycastle, S. J. Bumpstead, S. J. Chanock, L. Cherkas, P. Chines, L. Coin, C. Cooper, G. Crawford, A. Doering, A. Dominiczak, A. S. Doney, S. Ebrahim, P. Elliott, M. R. Erdos, K. Estrada, L. Ferrucci, G. Fischer, N. G. Forouhi, C. Gieger, H. Grallert, C. J. Groves, S. Grundy, C. Guiducci, D. Hadley, A. Hamsten, A. S. Havulinna, A. Hofman, R. Holle, J. W. Holloway, T. Illig, B. Isomaa, L. C. Jacobs, K. Jameson, P. Jousilahti, F. Karpe, J. Kuusisto, J. Laitinen, G. M. Lathrop, D. A. Lawlor, M. Mangino, W. L. McArdle, T. Meitinger, M. A. Morken, A. P. Morris, P. Munroe, N. Narisu, A. Nordstrom, P. Nordstrom, B. A. Oostra, C. N. Palmer, F. Payne, J. F. Peden, I. Prokopenko, F. Renstrom, A. Ruokonen, V. Salomaa, M. S. Sandhu, L. J. Scott, A. Scuteri, K. Silander, K. Song, X. Yuan, H. M. Stringham, A. J. Swift, T. Tuomi, M. Uda, P. Vollenweider, G. Waeber, C. Wallace, G. B. Walters, M. N. Weedon, et al. (2009). Genome-wide association scan meta-analysis identifies three loci influencing adiposity and fat distribution. *PLoS Genet 5*(6), e1000508. Lindgren, Cecilia M Heid, Iris M Randall, Joshua C Lamina, Claudia Steinthorsdottir, Valgerdur Qi, Lu Speliotes, Elizabeth K Thorleifsson, Gudmar Willer, Cristen J Herrera, Blanca M Jackson, Anne U Lim, Noha Scheet, Paul Soranzo, Nicole Amin, Najaf Aulchenko, Yurii S Chambers, John C Drong, Alexander Luan, Jian'an Lyon, Helen N Rivadeneira, Fernando Sanna, Serena Timpson, Nicholas J Zillikens, M Carola Zhao, Jing Hua Almgren, Peter Bandinelli, Stefania Bennett, Amanda J Bergman, Richard N Bonnycastle, Lori L Bumpstead, Suzannah J Chanock, Stephen J Cherkas, Lynn Chines, Peter Coin, Lachlan Cooper, Cyrus Crawford, Gabriel Doering, Angela Dominiczak, Anna Doney, Alex S F Ebrahim, Shah Elliott, Paul Erdos, Michael R Estrada, Karol Ferrucci, Luigi Fischer, Guido Forouhi, Nita G Gieger, Christian Grallert, Harald Groves, Christopher J Grundy, Scott Guiducci, Candace Hadley, David Hamsten, Anders Havulinna, Aki S Hofman, Albert Holle, Rolf Holloway, John W Illig, Thomas Isomaa, Bo Jacobs, Leonie C Jameson, Karen Jousilahti, Pekka Karpe, Fredrik Kuusisto, Johanna Laitinen, Jaana Lathrop, G Mark Lawlor, Debbie A Mangino, Massimo McArdle, Wendy L Meitinger, Thomas Morken, Mario A Morris, Andrew P Munroe, Patricia Narisu, Narisu Nordstrom, Anna Nordstrom, Peter Oostra, Ben A Palmer, Colin N A Payne, Felicity Peden, John F Prokopenko, Inga Renstrom, Frida Ruokonen, Aimo Salomaa, Veikko Sandhu, Manjinder S Scott, Laura J Scuteri, Angelo Silander, Kaisa Song, Kijoung Yuan, Xin Stringham, Heather M Swift, Amy J Tuomi, Tiinamaija Uda, Manuela Vollenweider, Peter Waeber, Gerard Wallace, Chris Walters, G Bragi Weedon, Michael N Wellcome Trust Case Control Consortium Witteman, Jacqueline C M Zhang, Cuilin Zhang, Weihua Caulfield, Mark J Collins, Francis S Davey Smith, George Day, Ian N M Franks, Paul W Hattersley, Andrew T Hu, Frank B Jarvelin, Marjo-Riitta Kong, Augustine Kooner, Jaspal S Laakso, Markku Lakatta, Edward Mooser, Vincent Morris, Andrew D Peltonen, Leena Samani, Nilesh J Spector, Timothy D Strachan, David

P Tanaka, Toshiko Tuomilehto, Jaakko Uitterlinden, Andre G van Duijn, Cornelia M Wareham, Nicholas J Hugh Watkins Procardis Consortia Waterworth, Dawn M Boehnke, Michael Deloukas, Panos Groop, Leif Hunter, David J Thorsteinsdottir, Unnur Schlessinger, David Wichmann, H-Erich Frayling, Timothy M Abecasis, Goncalo R Hirschhorn, Joel N Loos, Ruth J F Stefansson, Kari Mohlke, Karen L Barroso, Ines McCarthy, Mark I Giant Consortium 0600705/Medical Research Council/United Kingdom 064890/Wellcome Trust/United Kingdom 068545/Z/02/Wellcome Trust/United Kingdom 081682/Wellcome Trust/United Kingdom 086596/Z/08/Z/Wellcome Trust/United Kingdom DK062370/DK/NIDDK NIH HHS/ DK067288/DK/NIDDK NIH HHS/ DK07191/DK/NIDDK NIH HHS/ DK072193/DK/NIDDK NIH HHS/ DK075787/DK/NIDDK NIH HHS/ DK079466/DK/NIDDK NIH HHS/ DK080145/DK/NIDDK NIH HHS/ F32 DK079466-01/DK/NIDDK NIH HHS/ G0000649/Medical Research Council/United Kingdom G0000934/Medical Research Council/United Kingdom G02651/PHS HHS/ G0500539/Medical Research Council/United Kingdom G0601261/Medical Research Council/United Kingdom G9521010D/Medical Research Council/United Kingdom GR069224/Wellcome Trust/United Kingdom GR072960/Wellcome Trust/United Kingdom GR076113/Wellcome Trust/United Kingdom HL084729/HL/NHLBI NIH HHS/ HL087679/HL/NHLBI NIH HHS/ K23 DK080145-01/DK/NIDDK NIH HHS/ R01 DK029867/DK/NIDDK NIH HHS/ R01 DK072193-04/DK/NIDDK NIH HHS/ Biotechnology and Biological Sciences Research Council/United Kingdom British Heart Foundation/United Kingdom PLoS Genet. 2009 Jun;5(6):e1000508. doi: 10.1371/journal.pgen.1000508. Epub 2009 Jun 26.

Listgarten, J. and A. Emili (2005). Statistical and computational methods for comparative proteomic profiling using liquid chromatography-tandem mass spectrometry. *Molecular and Cellular Proteomics 4*(4), 419–434.

Lohse, M., A. M. Bolger, A. Nagel, A. R. Fernie, J. E. Lunn, M. Stitt, and B. Usadel (2012). Robina: a user-friendly, integrated software solution for rna-seq-based transcriptomics. *Nucleic acids research 40*(W1), W622–W627.

Lunter, G. and M. Goodson (2011, Jun). Stampy: a statistical algorithm for sensitive and fast mapping of illumina sequence reads. *Genome Res 21*(6), 936–9.

Marioni, J. C., C. E. Mason, S. M. Mane, M. Stephens, and Y. Gilad (2008, Sep). Rna-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res 18*(9), 1509–17.

Martin, J., V. M. Bruno, Z. Fang, X. Meng, M. Blow, T. Zhang, G. Sherlock, M. Snyder, and Z. Wang (2010). Rnnotator: an automated de novo transcriptome assembly pipeline from stranded rna-seq reads. *BMC Genomics 11*, 663.

Martin, J. A. and Z. Wang (2011, Oct). Next-generation transcriptome assembly. *Nat Rev Genet 12*(10), 671–82.

Matsubara, H., T. H. Jukes, and C. R. Cantor (1968, Jun). Structural and evolutionary relationships of ferredoxins. *Brookhaven Symp Biol 21*(1), 201–16.

Mavrich, T. N., C. Jiang, I. P. Ioshikhes, X. Li, B. J. Venters, S. J. Zanton, L. P. Tomsho, J. Qi, R. L. Glaser, S. C. Schuster, et al. (2008). Nucleosome organization in the drosophila genome. *Nature 453*(7193), 358–362.

McCarthy, M. I. and J. N. Hirschhorn (2008, Oct). Genome-wide association studies: past, present and future. *Hum Mol Genet 17*(R2), R100–1.

Megraw, M., F. Pereira, S. T. Jensen, U. Ohler, and A. G. Hatzigeorgiou (2009). A transcription factor affinity-based code for mammalian transcription initiation. *Genome research 19*(4), 644–656.

Meyerson, M., S. Gabriel, and G. Getz (2010, Oct). Advances in understanding cancer genomes through second-generation sequencing. *Nat Rev Genet 11*(10), 685–96.

Mitra, I., A. V. Nefedov, A. R. Brasier, and R. G. Sadygov (2012). Improved mass defect model for theoretical tryptic peptides. *Anal Chem. 84*, 3026–3032.

Monroe, M. E., N. Tolic, N. Jaitly, J. L. Shaw, J. N. Adkins, and R. D. Smith (2007). Viper: an advanced software package to support high-throughput lc-ms peptide identification. *Bioinformatics 23*(15), 2021–3. Monroe, Matthew E Tolic, Nikola Jaitly, Navdeep Shaw, Jason L Adkins, Joshua N Smith, Richard D RR018522/RR/NCRR NIH HHS/United States Y1-AI-4894-01/AI/NIAID NIH HHS/United States Research Support, N.I.H., Extramural Research Support, U.S. Gov't, Non-P.H.S. England Bioinformatics (Oxford, England) Bioinformatics. 2007 Aug 1;23(15):2021-3. Epub 2007 Jun 1.

Mortazavi, A., B. A. Williams, K. McCue, L. Schaeffer, and B. Wold (2008, Jul). Mapping and quantifying mammalian transcriptomes by rna-seq. *Nat Methods 5*(7), 621–8.

Mueller, L. N., O. Rinner, A. Schmidt, S. Letarte, B. Bodenmiller, M. Y. Brusniak, O. Vitek, R. Aebersold, and M. Muller (2007). Superhirn - a novel tool for high resolution lc-ms-based peptide/protein profiling. *Proteomics 7*(19), 3470–3480. 222GF Times Cited:78 Cited References Count:37.

Mwenifumbo, J. C. and M. A. Marra (2013, May). Cancer genome-sequencing study design. *Nat Rev Genet 14*(5), 321–32.

Nechaev, S., D. C. Fargo, G. dos Santos, L. Liu, Y. Gao, and K. Adelman (2010). Global analysis of short rnas reveals widespread promoter-proximal stalling and arrest of pol ii in drosophila. *Science 327*(5963), 335–338.

Ng, P., C.-L. Wei, W.-K. Sung, K. P. Chiu, L. Lipovich, C. C. Ang, S. Gupta, A. Shahab, A. Ridwan, C. H. Wong, E. T. Liu, and Y. Ruan (2005, Feb). Gene identification signature (gis) analysis for transcriptome characterization and genome annotation. *Nat Methods 2*(2), 105–11.

Ni, T., D. L. Corcoran, E. A. Rach, S. Song, E. P. Spana, Y. Gao, U. Ohler, and J. Zhu (2010). A paired-end sequencing strategy to map the complex landscape of transcription initiation. *Nature methods 7*(7), 521–527.

Nielsen, N. P. V., J. M. Carstensen, and J. Smedsgaard (1998). Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimised warping. *Journal of Chromatography A 805*(1-2), 17–35. Zp037 Times Cited:258 Cited References Count:14.

Nilsson, P. and A. Virtanen (2006, Aug). Expression and purification of recombinant poly(a)-specific ribonuclease (parn). *Int J Biol Macromol 39*(1-3), 95–9.

None. (2013). Picard tools. *http://picard.sourceforge.net*.

Ohler, U. (2006). Identification of core promoter modules in drosophila and their application in accurate transcription start site prediction. *Nucleic acids research 34*(20), 5943–5950.

Ohler, U. and D. A. Wassarman (2010). Promoting developmental transcription. *Development 137*(1), 15–26.

Okosun, I. S., K. M. Chandra, A. Boev, J. M. Boltri, S. T. Choi, D. C. Parish, and G. E. Dever (2004). Abdominal adiposity in u.s. adults: prevalence and trends, 1960-2000. *Prev Med 39*(1), 197–206. Okosun, Ike S Chandra, K M Dinesh Boev, Angel Boltri, John M Choi, Simon T Parish, David C Dever, G E Alan Prev Med. 2004 Jul;39(1):197-206.

Okosun, I. S., S. T. Choi, J. M. Boltri, D. C. Parish, K. M. Chandra, G. E. Dever, and A. Lucas (2003). Trends of abdominal adiposity in white, black, and mexican-american adults, 1988 to 2000. *Obes Res 11*(8), 1010–7. Okosun, Ike S Choi, Simon T Boltri, John M Parish, David C Chandra, K M Dinesh Dever, G E Alan Lucas, Amy Obes Res. 2003 Aug;11(8):1010-7.

Oliveros, J. (2007). Venny. an interactive tool for comparing lists with venn diagrams. *BioinfoGP, CNB-CSIC*.

Oshlack, A. and M. J. Wakefield (2009). Transcript length bias in rna-seq data confounds systems biology. *Biol Direct 4*(1), 14.

Ozsolak, F. and P. M. Milos (2011, Feb). Rna sequencing: advances, challenges and opportunities. *Nat Rev Genet 12*(2), 87–98.

Pages, G., A. Girard, O. Jeanneton, P. Barbe, C. Wolf, M. Lafontan, P. Valet, and J. S. Saulnier-Blache (2000). Lpa as a paracrine mediator of adipocyte growth and function. *Ann N Y Acad Sci 905*, 159–64. Pages, G Girard, A Jeanneton, O Barbe, P Wolf, C Lafontan, M Valet, P Saulnier-Blache, J S Ann N Y Acad Sci. 2000 Apr;905:159-64.

Park, P. J. (2009). Chip–seq: advantages and challenges of a maturing technology. *Nature Reviews Genetics 10*(10), 669–680.

Patel, K., J. E. Lucas, J. W. Thompson, L. G. Dubois, H. L. Tillmann, A. J. Thompson, D. Uzarski, R. M. Califf, M. A. Moseley, G. S. Ginsburg, J. G. McHutchison, J. J. McCarthy, and MURDOCK Horizon 1 Study Team (2011, Jun). High predictive accuracy of an unbiased proteomic profile for sustained virologic response in chronic hepatitis c patients. *Hepatology 53*(6), 1809–18.

Pearson, H. (2008). Biologists initiate plan to map human proteome. *Nature 452*(7190), 920–1. Pearson, Helen News England Nature Nature. 2008 Apr 24;452(7190):920-1.

Pickrell, J. K., J. C. Marioni, A. A. Pai, J. F. Degner, B. E. Engelhardt, E. Nkadori, J.-B. Veyrieras, M. Stephens, Y. Gilad, and J. K. Pritchard (2010, Apr). Understanding mechanisms underlying human gene expression variation with rna sequencing. *Nature 464*(7289), 768–72.

Planet, E., C. S.-O. Attolini, O. Reina, O. Flores, and D. Rossell (2012, Feb). htseqtools: high-throughput sequencing quality control, processing and visualization in r. *Bioinformatics 28*(4), 589–90.

Plessy, C., N. Bertin, H. Takahashi, R. Simone, M. Salimullah, T. Lassmann, M. Vitezic, J. Severin, S. Olivarius, D. Lazarevic, N. Hornig, V. Orlando, I. Bell, H. Gao, J. Dumais, P. Kapranov, H. Wang, C. A. Davis, T. R. Gingeras, J. Kawai, C. O. Daub, Y. Hayashizaki, S. Gustincich, and P. Carninci (2010, Jul). Linking promoters to functional transcripts in small samples with nanocage and cagescan. *Nat Methods 7*(7), 528–34.

Pluskal, T., S. Castillo, A. Villar-Briones, and M. Oresic (2010). Mzmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinformatics 11*, 395. Pluskal, Tomas Castillo, Sandra Villar-Briones, Alejandro Oresic, Matej Research Support, Non-U.S. Gov't England BMC bioinformatics BMC Bioinformatics. 2010 Jul 23;11:395.

Ponger, L., L. Duret, and D. Mouchiroud (2001). Determinants of cpg islands: expression in early embryo and isochore structure. *Genome research 11*(11), 1854–1860.

Portales-Casamar, E., S. Thongjuea, A. T. Kwon, D. Arenillas, X. Zhao, E. Valen, D. Yusuf, B. Lenhard, W. W. Wasserman, and A. Sandelin (2010). Jaspar 2010: the greatly expanded open-access database of transcription factor binding profiles. *Nucleic acids research 38*(suppl 1), D105–D110.

Prakash, A., P. Mallick, J. Whiteaker, H. Zhang, A. Paulovich, M. Flory, H. Lee, R. Aebersold, and B. Schwikowski (2006). Signal maps for mass spectrometry-based comparative proteomics. *Mol Cell Proteomics 5*(3), 423–32. Prakash, Amol Mallick, Parag Whiteaker, Jeffrey Zhang, Heidi Paulovich, Amanda Flory, Mark Lee, Hookeun Aebersold, Ruedi Schwikowski, Benno N01-HV-28179/HV/NHLBI NIH HHS/United States Comparative Study Research Support, N.I.H., Extramural United States Molecular and cellular proteomics : MCP Mol Cell Proteomics. 2006 Mar;5(3):423-32. Epub 2005 Nov 3.

Price, A. L., N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, and D. Reich (2006a, Aug). Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet 38*(8), 904–9.

Price, A. L., N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, and D. Reich (2006b). Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet 38*(8), 904–9. Price, Alkes L Patterson, Nick J Plenge, Robert M Weinblatt, Michael E Shadick, Nancy A Reich, David Nat Genet. 2006 Aug;38(8):904-9. Epub 2006 Jul 23.

Prince, J. T. and E. M. Marcotte (2006). Chromatographic alignment of esi-lc-ms proteomics data sets by ordered bijective interpolated warping. *Anal Chem 78*(17), 6140–52. Prince, John T Marcotte, Edward M Research Support, N.I.H., Extramural Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, Non-P.H.S. United States Analytical chemistry Anal Chem. 2006 Sep 1;78(17):6140-52.

Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. de Bakker, M. J. Daly, and P. C. Sham (2007). Plink: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet 81*(3), 559–75. Purcell, Shaun Neale, Benjamin Todd-Brown, Kathe Thomas, Lori Ferreira, Manuel A R Bender, David Maller, Julian Sklar, Pamela de Bakker, Paul I W Daly, Mark J Sham, Pak C EY-12562/EY/NEI NIH HHS/ R03 MH73806-01A1/MH/NIMH NIH HHS/ U01 HG004171/HG/NHGRI NIH HHS/ Am J Hum Genet. 2007 Sep;81(3):559-75. Epub 2007 Jul 25.

Quinlan, A. R. and I. M. Hall (2010, Mar). Bedtools: a flexible suite of utilities for comparing genomic features. *Bioinformatics 26*(6), 841–2.

Rach, E. A., D. R. Winter, A. M. Benjamin, D. L. Corcoran, T. Ni, J. Zhu, and U. Ohler (2011). Transcription initiation patterns indicate divergent strategies for gene regulation at the chromatin level. *PLoS Genetics 7*(1), e1001274.

Rach, E. A., H.-Y. Yuan, W. H. Majoros, P. Tomancak, and U. Ohler (2009). Motif composition, conservation and condition-specificity of single and alternative transcription start sites in the drosophila genome. *Genome Biol 10*(7), R73.

Raisner, R. M., P. D. Hartley, M. D. Meneghini, M. Z. Bao, C. L. Liu, S. L. Schreiber, O. J. Rando, and H. D. Madhani (2005). Histone variant h2a. z marks the 5 ends of both active and inactive genes in euchromatin. *Cell 123*(2), 233–248.

Ramirez-Carrozzi, V. R., D. Braas, D. M. Bhatt, C. S. Cheng, C. Hong, K. R. Doty, J. C. Black, A. Hoffmann, M. Carey, and S. T. Smale (2009). A unifying model for the selective regulation of inducible transcription by cpg islands and nucleosome remodeling. *Cell 138*(1), 114–128.

Rasmussen, C. E. (2000). The infinite gaussian mixture model. *Advances in Neural Information Processing Systems 12 12*, 554–560. Bq93f Times Cited:95 Cited References Count:9 Advances in Neural Information Processing Systems.

Reich, M., T. Liefeld, J. Gould, J. Lerner, P. Tamayo, and J. P. Mesirov (2006). Genepattern 2.0. *Nature genetics 38*(5), 500–501.

Renstrom, F., F. Payne, A. Nordstrom, E. C. Brito, O. Rolandsson, G. Hallmans, I. Barroso, P. Nordstrom, and P. W. Franks (2009). Replication and extension of genome-wide association study results for obesity in 4923 adults from northern sweden. *Hum Mol Genet 18*(8), 1489–96. Renstrom, Frida Payne, Felicity Nordstrom, Anna Brito, Ema C Rolandsson, Olov Hallmans, Goran Barroso, Ines Nordstrom, Peter Franks, Paul W GIANT Consortium Wellcome Trust/United Kingdom England Hum Mol Genet. 2009 Apr 15;18(8):1489-96. doi: 10.1093/hmg/ddp041. Epub 2009 Jan 22.

Riley, C. P., E. S. Gough, J. He, S. S. Jandhyala, B. Kennedy, S. Orcun, M. Ouzzani, C. Buck, A. M. Roumani, and X. Zhang (2010). The proteome discovery pipeline-a data analysis pipeline for mass spectrometry-based differential proteomics discovery. *The Open Proteomics Journal 3*, 8–19.

Roberts, A., C. Trapnell, J. Donaghey, J. L. Rinn, and L. Pachter (2011). Improving rna-seq expression estimates by correcting for fragment bias. *Genome Biol 12*(3), R22.

Robertson, G., J. Schein, R. Chiu, R. Corbett, M. Field, S. D. Jackman, K. Mungall, S. Lee, H. M. Okada, J. Q. Qian, M. Griffith, A. Raymond, N. Thiessen, T. Cezard, Y. S. Butterfield, R. Newsome, S. K. Chan, R. She, R. Varhol, B. Kamoh, A.-L. Prabhu, A. Tam, Y. Zhao, R. A. Moore, M. Hirst, M. A. Marra, S. J. M. Jones, P. A. Hoodless, and I. Birol (2010, Nov). De novo assembly and analysis of rna-seq data. *Nat Methods 7*(11), 909–12.

Robinson, M. D., D. J. McCarthy, and G. K. Smyth (2010, Jan). edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics 26*(1), 139–40.

Rumble, S. M., P. Lacroute, A. V. Dalca, M. Fiume, A. Sidow, and M. Brudno (2009, May). Shrimp: accurate mapping of short color-space reads. *PLoS Comput Biol 5*(5), e1000386.

Sammalisto, S., T. Hiekkalinna, K. Schwander, S. Kardia, A. B. Weder, B. L. Rodriguez, A. Doria, J. A. Kelly, G. R. Bruner, J. B. Harley, S. Redline, E. K. Larkin, S. R. Patel, A. J. Ewan, J. L. Weber, M. Perola, and L. Peltonen (2009). Genome-wide linkage screen for stature and body mass index in 3.032 families: evidence for sex- and population-specific genetic effects. *Eur J Hum Genet 17*(2), 258–66. Sammalisto, Sampo Hiekkalinna, Tero Schwander, Karen Kardia, Sharon Weder, Alan B Rodriguez, Beatriz L Doria, Alessandro Kelly, Jennifer A Bruner, Gail R Harley, John B Redline, Susan Larkin, Emma K Patel, Sanjay R Ewan, Amy J H Weber, James L Perola, Markus Peltonen, Leena AI024717/AI/NIAID NIH HHS/ AR04246/AR/NIAMS NIH HHS/ AR062277/AR/NIAMS NIH HHS/ DK55523/DK/NIDDK NIH HHS/ N01 AR062277/AR/NIAMS NIH HHS/ R01 AR042460-14/AR/NIAMS NIH HHS/ R01 DK055523-09/DK/NIDDK NIH HHS/ R37 AI024717-20/AI/NIAID NIH HHS/ England Eur J Hum Genet. 2009 Feb;17(2):258-66. doi: 10.1038/ejhg.2008.152. Epub 2008 Sep 10.

Satten, G. A., W. D. Flanders, and Q. Yang (2001, Feb). Accounting for unmeasured population substructure in case-control studies of genetic association using a novel latent-class model. *Am J Hum Genet 68*(2), 466–77.

Saxonov, S., P. Berg, and D. L. Brutlag (2006). A genome-wide analysis of cpg dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proceedings of the National Academy of Sciences of the United States of America 103*(5), 1412–1417.

Schones, D. E., K. Cui, S. Cuddapah, T.-Y. Roh, A. Barski, Z. Wang, G. Wei, and K. Zhao (2008). Dynamic regulation of nucleosome positioning in the human genome. *Cell 132*(5), 887–898.

Schulz, M. H., D. R. Zerbino, M. Vingron, and E. Birney (2012, Apr). Oases: robust de novo rna-seq assembly across the dynamic range of expression levels. *Bioinformatics 28*(8), 1086–92.

Schwarz, K. B., R. P. Gonzalez-Peralta, K. F. Murray, J. P. Molleston, B. A. Haber, M. M. Jonas, P. Rosenthal, P. Mohan, W. F. Balistreri, M. R. Narkewicz, L. Smith, S. J. Lobritto, S. Rossi, A. Valsamakis, Z. Goodman, P. R. Robuck, B. A. Barton, and Peds-C Clinical Research Network (2011, Feb). The combination of ribavirin and peginterferon is superior to peginterferon and placebo for children and adolescents with chronic hepatitis c. *Gastroenterology 140*(2), 450–458.e1.

Service, R. F. (2008). Proteomics. proteomics ponders prime time. *Science 321*(5897), 1758–61. Service, Robert F News United States Science (New York, N.Y.) Science. 2008 Sep 26;321(5897):1758-61.

Shiraki, T., S. Kondo, S. Katayama, K. Waki, T. Kasukawa, H. Kawaji, R. Kodzius, A. Watahiki, M. Nakamura, T. Arakawa, S. Fukuda, D. Sasaki, A. Podhajska, M. Harbers, J. Kawai, P. Carninci, and Y. Hayashizaki (2003, Dec). Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proc Natl Acad Sci U S A 100*(26), 15776–81.

Silva, J. C., R. Denny, C. Dorschel, M. V. Gorenstein, G. Z. Li, K. Richardson, D. Wall, and S. J. Geromanos (2006). Simultaneous qualitative and quantitative analysis of the escherichia coli proteome: a sweet tale. *Mol Cell Proteomics 5*(4), 589–607. Silva, Jeffrey C Denny, Richard Dorschel, Craig Gorenstein, Marc V Li, Guo-Zhong Richardson, Keith Wall, Daniel Geromanos, Scott J United States Molecular and cellular proteomics : MCP Mol Cell Proteomics. 2006 Apr;5(4):589-607. Epub 2006 Jan 5.

Silva, J. C., R. Denny, C. A. Dorschel, M. Gorenstein, I. J. Kass, G. Z. Li, T. McKenna, M. J. Nold, K. Richardson, P. Young, and S. Geromanos (2005). Quantitative proteomic analysis by accurate mass retention time pairs. *Anal Chem 77*(7), 2187–200. Silva, Jeffrey C Denny, Richard Dorschel, Craig A Gorenstein, Marc Kass, Ignatius J Li, Guo-Zhong McKenna, Therese Nold, Michael J Richardson, Keith Young, Phillip Geromanos, Scott United States Analytical chemistry Anal Chem. 2005 Apr 1;77(7):2187-200.

Simon, M. F., D. Daviaud, J. P. Pradere, S. Gres, C. Guigne, M. Wabitsch, J. Chun, P. Valet, and J. S. Saulnier-Blache (2005). Lysophosphatidic acid inhibits adipocyte differentiation via lysophosphatidic acid 1 receptor-dependent down-regulation of peroxisome proliferator-activated receptor gamma2. *J Biol Chem 280*(15), 14656–62. Simon, Marie Francoise Daviaud, Daniele Pradere, Jean Philippe Gres, Sandra Guigne, Charlotte Wabitsch, Martin Chun, Jerold Valet, Philippe Saulnier-Blache, Jean Sebastien J Biol Chem. 2005 Apr 15;280(15):14656-62. Epub 2005 Feb 14.

Simpson, J. T., K. Wong, S. D. Jackman, J. E. Schein, S. J. M. Jones, and I. Birol (2009, Jun). Abyss: a parallel assembler for short read sequence data. *Genome Res 19*(6), 1117–23.

SJ, G., V. JP, S. JC, D. CA, L. GZ, G. MV, B. RH, and L. JI (2009). The detection, correlation, and comparison of peptide precursor and product ions from data independent lc-ms with data dependent lc-ms/ms. *Proteomics 9*, 1683–1695.

Smith, C. A., E. J. Want, G. O'Maille, R. Abagyan, and G. Siuzdak (2006). Xcms: Processing mass spectrometry data for metabolite profiling using nonlinear peak

alignment, matching, and identification. *Analytical Chemistry 78*(3), 779–787. 010LC Times Cited:304 Cited References Count:23.

Speliotes, E. K., C. J. Willer, S. I. Berndt, K. L. Monda, G. Thorleifsson, A. U. Jackson, H. Lango Allen, C. M. Lindgren, J. Luan, R. Magi, J. C. Randall, S. Vedantam, T. W. Winkler, L. Qi, T. Workalemahu, I. M. Heid, V. Steinthorsdottir, H. M. Stringham, M. N. Weedon, E. Wheeler, A. R. Wood, T. Ferreira, R. J. Weyant, A. V. Segre, K. Estrada, L. Liang, J. Nemesh, J. H. Park, S. Gustafsson, T. O. Kilpelainen, J. Yang, N. Bouatia-Naji, T. Esko, M. F. Feitosa, Z. Kutalik, M. Mangino, S. Raychaudhuri, A. Scherag, A. V. Smith, R. Welch, J. H. Zhao, K. K. Aben, D. M. Absher, N. Amin, A. L. Dixon, E. Fisher, N. L. Glazer, M. E. Goddard, N. L. Heard-Costa, V. Hoesel, J. J. Hottenga, A. Johansson, T. Johnson, S. Ketkar, C. Lamina, S. Li, M. F. Moffatt, R. H. Myers, N. Narisu, J. R. Perry, M. J. Peters, M. Preuss, S. Ripatti, F. Rivadeneira, C. Sandholt, L. J. Scott, N. J. Timpson, J. P. Tyrer, S. van Wingerden, R. M. Watanabe, C. C. White, F. Wiklund, C. Barlassina, D. I. Chasman, M. N. Cooper, J. O. Jansson, R. W. Lawrence, N. Pellikka, I. Prokopenko, J. Shi, E. Thiering, H. Alavere, M. T. Alibrandi, P. Almgren, A. M. Arnold, T. Aspelund, L. D. Atwood, B. Balkau, A. J. Balmforth, A. J. Bennett, Y. Ben-Shlomo, R. N. Bergman, S. Bergmann, H. Biebermann, A. I. Blakemore, T. Boes, L. L. Bonnycastle, S. R. Bornstein, M. J. Brown, T. A. Buchanan, et al. (2010). Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nat Genet 42*(11), 937–48. Speliotes, Elizabeth K Willer, Cristen J Berndt, Sonja I Monda, Keri L Thorleifsson, Gudmar Jackson, Anne U Lango Allen, Hana Lindgren, Cecilia M Luan, Jian'an Magi, Reedik Randall, Joshua C Vedantam, Sailaja Winkler, Thomas W Qi, Lu Workalemahu, Tsegaselassie Heid, Iris M Steinthorsdottir, Valgerdur Stringham, Heather M Weedon, Michael N Wheeler, Eleanor Wood, Andrew R Ferreira, Teresa Weyant, Robert J Segre, Ayellet V Estrada, Karol Liang, Liming Nemesh, James Park, Ju-Hyun Gustafsson, Stefan Kilpelainen, Tuomas O Yang, Jian Bouatia-Naji, Nabila Esko, Tonu Feitosa, Mary F Kutalik, Zoltan Mangino, Massimo Raychaudhuri, Soumya Scherag, Andre Smith, Albert Vernon Welch, Ryan Zhao, Jing Hua Aben, Katja K Absher, Devin M Amin, Najaf Dixon, Anna L Fisher, Eva Glazer, Nicole L Goddard, Michael E Heard-Costa, Nancy L Hoesel, Volker Hottenga, Jouke-Jan Johansson, Asa Johnson, Toby Ketkar, Shamika Lamina, Claudia Li, Shengxu Moffatt, Miriam F Myers, Richard H Narisu, Narisu Perry, John R B Peters, Marjolein J Preuss, Michael Ripatti, Samuli Rivadeneira, Fernando Sandholt, Camilla Scott, Laura J Timpson, Nicholas J Tyrer, Jonathan P van Wingerden, Sophie Watanabe, Richard M White, Charles C Wiklund, Fredrik Barlassina, Christina Chasman, Daniel I Cooper, Matthew N Jansson, John-Olov Lawrence, Robert W Pellikka, Niina Prokopenko, Inga Shi, Jianxin Thiering, Elisabeth Alavere, Helene Alibrandi, Maria T S Almgren, Peter Arnold, Alice M Aspelund, Thor Atwood, Larry D Balkau, Beverley Balmforth, Anthony J Bennett, Amanda J Ben-Shlomo, Yoav Bergman, Richard N Bergmann, Sven Biebermann,

Heike Blakemore, Alexandra I F Boes, Tanja Bonnycastle, Lori L Bornstein, Stefan R Brown, Morris J Buchanan, Thomas A Busonero, Fabio Campbell, Harry Cappuccio, Francesco P Cavalcanti-Proenca, Christine Chen, Yii-Der Ida Chen, Chih-Mei Chines, Peter S Clarke, Robert Coin, Lachlan Connell, John Day, Ian N M den Heijer, Martin Duan, Jubao Ebrahim, Shah Elliott, Paul Elosua, Roberto Eiriksdottir, Gudny Erdos, Michael R Eriksson, Johan G Facheris, Maurizio F Felix, Stephan B Fischer-Posovszky, Pamela Folsom, Aaron R Friedrich, Nele Freimer, Nelson B Fu, Mao Gaget, Stefan Gejman, Pablo V Geus, Eco J C Gieger, Christian Gjesing, Anette P Goel, Anuj Goyette, Philippe Grallert, Harald Grassler, Jurgen Greenawalt, Danielle M Groves, Christopher J Gudnason, Vilmundur Guiducci, Candace Hartikainen, Anna-Liisa Hassanali, Neelam Hall, Alistair S Havulinna, Aki S Hayward, Caroline Heath, Andrew C Hengstenberg, Christian Hicks, Andrew A Hinney, Anke Hofman, Albert Homuth, Georg Hui, Jennie Igl, Wilmar Iribarren, Carlos Isomaa, Bo Jacobs, Kevin B Jarick, Ivonne Jewell, Elizabeth John, Ulrich Jorgensen, Torben Jousilahti, Pekka Jula, Antti Kaakinen, Marika Kajantie, Eero Kaplan, Lee M Kathiresan, Sekar Kettunen, Johannes Kinnunen, Leena Knowles, Joshua W Kolcic, Ivana Konig, Inke R Koskinen, Seppo Kovacs, Peter Kuusisto, Johanna Kraft, Peter Kvaloy, Kirsti Laitinen, Jaana Lantieri, Olivier Lanzani, Chiara Launer, Lenore J Lecoeur, Cecile Lehtimaki, Terho Lettre, Guillaume Liu, Jianjun Lokki, Marja-Liisa Lorentzon, Mattias Luben, Robert N Ludwig, Barbara MAGIC Manunta, Paolo Marek, Diana Marre, Michel Martin, Nicholas G McArdle, Wendy L McCarthy, Anne McKnight, Barbara Meitinger, Thomas Melander, Olle Meyre, David Midthjell, Kristian Montgomery, Grant W Morken, Mario A Morris, Andrew P Mulic, Rosanda Ngwa, Julius S Nelis, Mari Neville, Matt J Nyholt, Dale R O'Donnell, Christopher J O'Rahilly, Stephen Ong, Ken K Oostra, Ben Pare, Guillaume Parker, Alex N Perola, Markus Pichler, Irene Pietilainen, Kirsi H Platou, Carl G P Polasek, Ozren Pouta, Anneli Rafelt, Suzanne Raitakari, Olli Rayner, Nigel W Ridderstrale, Martin Rief, Winfried Ruokonen, Aimo Robertson, Neil R Rzehak, Peter Salomaa, Veikko Sanders, Alan R Sandhu, Manjinder S Sanna, Serena Saramies, Jouko Savolainen, Markku J Scherag, Susann Schipf, Sabine Schreiber, Stefan Schunkert, Heribert Silander, Kaisa Sinisalo, Juha Siscovick, David S Smit, Jan H Soranzo, Nicole Sovio, Ulla Stephens, Jonathan Surakka, Ida Swift, Amy J Tammesoo, Mari-Liis Tardif, Jean-Claude Teder-Laving, Maris Teslovich, Tanya M Thompson, John R Thomson, Brian Tonjes, Anke Tuomi, Tiinamaija van Meurs, Joyce B J van Ommen, Gert-Jan Vatin, Vincent Viikari, Jorma Visvikis-Siest, Sophie Vitart, Veronique Vogel, Carla I G Voight, Benjamin F Waite, Lindsay L Wallaschofski, Henri Walters, G Bragi Widen, Elisabeth Wiegand, Susanna Wild, Sarah H Willemsen, Gonneke Witte, Daniel R Witteman, Jacqueline C Xu, Jianfeng Zhang, Qunyuan Zgaga, Lina Ziegler, Andreas Zitting, Paavo Beilby, John P Farooqi, I Sadaf Hebebrand, Johannes Huikuri, Heikki V James, Alan L Kahonen, Mika Levinson, Douglas F Macciardi, Fabio Nieminen, Markku S Ohlsson, Claes Palmer, Lyle J Ridker, Paul M Stumvoll, Michael Beckmann, Jacques S Boeing, Heiner Boerwinkle, Eric

Boomsma, Dorret I Caulfield, Mark J Chanock, Stephen J Collins, Francis S Cupples, L Adrienne Smith, George Davey Erdmann, Jeanette Froguel, Philippe Gronberg, Henrik Gyllensten, Ulf Hall, Per Hansen, Torben Harris, Tamara B Hattersley, Andrew T Hayes, Richard B Heinrich, Joachim Hu, Frank B Hveem, Kristian Illig, Thomas Jarvelin, Marjo-Riitta Kaprio, Jaakko Karpe, Fredrik Khaw, Kay-Tee Kiemeney, Lambertus A Krude, Heiko Laakso, Markku Lawlor, Debbie A Metspalu, Andres Munroe, Patricia B Ouwehand, Willem H Pedersen, Oluf Penninx, Brenda W Peters, Annette Pramstaller, Peter P Quertermous, Thomas Reinehr, Thomas Rissanen, Aila Rudan, Igor Samani, Nilesh J Schwarz, Peter E H Shuldiner, Alan R Spector, Timothy D Tuomilehto, Jaakko Uda, Manuela Uitterlinden, Andre Valle, Timo T Wabitsch, Martin Waeber, Gerard Wareham, Nicholas J Watkins, Hugh Procardis Consortium Wilson, James F Wright, Alan F Zillikens, M Carola Chatterjee, Nilanjan McCarroll, Steven A Purcell, Shaun Schadt, Eric E Visscher, Peter M Assimes, Themistocles L Borecki, Ingrid B Deloukas, Panos Fox, Caroline S Groop, Leif C Haritunians, Talin Hunter, David J Kaplan, Robert C Mohlke, Karen L O'Connell, Jeffrey R Peltonen, Leena Schlessinger, David Strachan, David P van Duijn, Cornelia M Wichmann, H-Erich Frayling, Timothy M Thorsteinsdottir, Unnur Abecasis, Goncalo R Barroso, Ines Boehnke, Michael Stefansson, Kari North, Kari E McCarthy, Mark I Hirschhorn, Joel N Ingelsson, Erik Loos, Ruth J F Wellcome Trust United Kingdom. Nat Genet. 2010 Nov;42(11):937-48. doi: 10.1038/ng.686. Epub 2010 Oct 10.

Spies, N., C. B. Nielsen, R. A. Padgett, and C. B. Burge (2009). Biased chromatin signatures around polyadenylation sites and exons. *Molecular cell 36*(2), 245–254.

Steinberg, G. R., B. E. Kemp, and M. J. Watt (2007). Adipocyte triglyceride lipase expression in human obesity. *Am J Physiol Endocrinol Metab 293*(4), E958–64. Steinberg, Gregory R Kemp, Bruce E Watt, Matthew J Am J Physiol Endocrinol Metab. 2007 Oct;293(4):E958-64. Epub 2007 Jul 3.

Stranger, B. E., E. A. Stahl, and T. Raj (2011, Feb). Progress and promise of genome-wide association studies for human complex trait genetics. *Genetics 187*(2), 367–83.

Sturm, M., A. Bertsch, C. Gropl, A. Hildebrandt, R. Hussong, E. Lange, N. Pfeifer, O. Schulz-Trieglaff, A. Zerck, K. Reinert, and O. Kohlbacher (2008). Openms - an open-source software framework for mass spectrometry. *BMC Bioinformatics 9*, 163. Sturm, Marc Bertsch, Andreas Gropl, Clemens Hildebrandt, Andreas Hussong, Rene Lange, Eva Pfeifer, Nico Schulz-Trieglaff, Ole Zerck, Alexandra Reinert, Knut Kohlbacher, Oliver Research Support, Non-U.S. Gov't England BMC bioinformatics BMC Bioinformatics. 2008 Mar 26;9:163.

Surget-Groba, Y. and J. I. Montoya-Burgos (2010, Oct). Optimization of de novo transcriptome assembly from next-generation sequencing data. *Genome Res 20*(10), 1432–40.

Takai, D. and P. A. Jones (2002). Comprehensive analysis of cpg islands in human chromosomes 21 and 22. *Proceedings of the national academy of sciences 99*(6), 3740–3745.

Tang, Z., L. Zhang, A. K. Cheema, and H. W. Ressom (2011). A new method for alignment of lc-maldi-tof data. *Proteome Sci 9 Suppl 1*, S10. Tang, Zhiqun Zhang, Lihua Cheema, Amrita K Ressom, Habtom W England Proteome Sci. 2011 Oct 14;9 Suppl 1:S10.

Thorleifsson, G., G. B. Walters, D. F. Gudbjartsson, V. Steinthorsdottir, P. Sulem, A. Helgadottir, U. Styrkarsdottir, S. Gretarsdottir, S. Thorlacius, I. Jonsdottir, T. Jonsdottir, E. J. Olafsdottir, G. H. Olafsdottir, T. Jonsson, F. Jonsson, K. Borch-Johnsen, T. Hansen, G. Andersen, T. Jorgensen, T. Lauritzen, K. K. Aben, A. L. Verbeek, N. Roeleveld, E. Kampman, L. R. Yanek, L. C. Becker, L. Tryggvadottir, T. Rafnar, D. M. Becker, J. Gulcher, L. A. Kiemeney, O. Pedersen, A. Kong, U. Thorsteinsdottir, and K. Stefansson (2009). Genome-wide association yields new sequence variants at seven loci that associate with measures of obesity. *Nat Genet 41*(1), 18–24. Thorleifsson, Gudmar Walters, G Bragi Gudbjartsson, Daniel F Steinthorsdottir, Valgerdur Sulem, Patrick Helgadottir, Anna Styrkarsdottir, Unnur Gretarsdottir, Solveig Thorlacius, Steinunn Jonsdottir, Ingileif Jonsdottir, Thorbjorg Olafsdottir, Elinborg J Olafsdottir, Gudridur H Jonsson, Thorvaldur Jonsson, Frosti Borch-Johnsen, Knut Hansen, Torben Andersen, Gitte Jorgensen, Torben Lauritzen, Torsten Aben, Katja K Verbeek, Andre L M Roeleveld, Nel Kampman, Ellen Yanek, Lisa R Becker, Lewis C Tryggvadottir, Laufey Rafnar, Thorunn Becker, Diane M Gulcher, Jeffrey Kiemeney, Lambertus A Pedersen, Oluf Kong, Augustine Thorsteinsdottir, Unnur Stefansson, Kari HL072518/HL/NHLBI NIH HHS/ HL087698/HL/NHLBI NIH HHS/ M01-RR000052/RR/NCRR NIH HHS/ Nat Genet. 2009 Jan;41(1):18-24. doi: 10.1038/ng.274. Epub 2008 Dec 14.

Tian, C., R. M. Plenge, M. Ransom, A. Lee, P. Villoslada, C. Selmi, L. Klareskog, A. E. Pulver, L. Qi, P. K. Gregersen, and M. F. Seldin (2008, Jan). Analysis and application of european genetic substructure using 300 k snp information. *PLoS Genet 4*(1), e4.

Tillo, D., N. Kaplan, I. K. Moore, Y. Fondufe-Mittendorf, A. J. Gossett, Y. Field, J. D. Lieb, J. Widom, E. Segal, and T. R. Hughes (2010). High nucleosome occupancy is encoded at human regulatory sequences. *PloS one 5*(2), e9129.

Tirosh, I. and N. Barkai (2008). Two strategies for gene regulation by promoter nucleosomes. *Genome research 18*(7), 1084–1091.

Tirosh, I., N. Barkai, and K. J. Verstrepen (2009). Promoter architecture and the evolvability of gene expression. *J Biol 8*(11), 95.

Trapnell, C., L. Pachter, and S. L. Salzberg (2009, May). Tophat: discovering splice junctions with rna-seq. *Bioinformatics 25*(9), 1105–11.

Trapnell, C., B. A. Williams, G. Pertea, A. Mortazavi, G. Kwan, M. J. van Baren, S. L. Salzberg, B. J. Wold, and L. Pachter (2010, May). Transcript assembly and quantification by rna-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol 28*(5), 511–5.

Tsukiyama, T., P. B. Becker, and C.-G. Wu (1994). Atp-dependent nucleosome disruption at a heat-shock promoter mediated by binding of gaga transcription factor. *Nature*, 525–532.

Valen, E., G. Pascarella, A. Chalk, N. Maeda, M. Kojima, C. Kawazu, M. Murata, H. Nishiyori, D. Lazarevic, D. Motti, T. T. Marstrand, M.-H. E. Tang, X. Zhao, A. Krogh, O. Winther, T. Arakawa, J. Kawai, C. Wells, C. Daub, M. Harbers, Y. Hayashizaki, S. Gustincich, A. Sandelin, and P. Carninci (2009, Feb). Genome-wide detection and analysis of hippocampus core promoters using deepcage. *Genome Res 19*(2), 255–65.

van Meeteren, L. A. and W. H. Moolenaar (2007). Regulation and biological activities of the autotaxin-lpa axis. *Prog Lipid Res 46*(2), 145–60. van Meeteren, Laurens A Moolenaar, Wouter H England Prog Lipid Res. 2007 Mar;46(2):145-60. Epub 2007 Mar 16.

Vandenbogaert, M., S. Li-Thiao-Te, H. M. Kaltenbach, R. X. Zhang, T. Aittokallio, and B. Schwikowski (2008). Alignment of lc-ms images, with applications to biomarker discovery and protein identification. *Proteomics 8*(4), 650–672. 271MD Times Cited:20 Cited References Count:105.

Veltman, J. A. and H. G. Brunner (2012). De novo mutations in human genetic disease. *Nature Reviews Genetics 13*(8), 565–575.

Vince, C. (2004). yags: Yet another gee solver.

Vissers, J. P. C., J. I. Langridge, and J. M. F. G. Aerts (2007). Analysis and quantification of diagnostic serum markers and protein signatures for gaucher disease. *Molecular and Cellular Proteomics 6*(5), 755–766. 168LU Times Cited:59 Cited References Count:34.

Walker, T. M., C. L. C. Ip, R. H. Harrell, J. T. Evans, G. Kapatai, M. J. Dedicoat, D. W. Eyre, D. J. Wilson, P. M. Hawkey, D. W. Crook, J. Parkhill, D. Harris, A. S. Walker, R. Bowden, P. Monk, E. G. Smith, and T. E. A. Peto (2013, Feb). Whole-genome sequencing to delineate mycobacterium tuberculosis outbreaks: a retrospective observational study. *Lancet Infect Dis 13*(2), 137–46.

Wang, A. and E. A. Dennis (1999). Mammalian lysophospholipases. *Biochim Biophys Acta 1439*(1), 1–16. Wang, A Dennis, E A GM 20501/GM/NIGMS NIH HHS/ GM 51606/GM/NIGMS NIH HHS/ HD 26171/HD/NICHD NIH HHS/ NETHERLANDS Biochim Biophys Acta. 1999 Jul 9;1439(1):1-16.

Wang, K., D. Singh, Z. Zeng, S. J. Coleman, Y. Huang, G. L. Savich, X. He, P. Mieczkowski, S. A. Grimm, C. M. Perou, J. N. MacLeod, D. Y. Chiang, J. F. Prins, and J. Liu (2010, Oct). Mapsplice: accurate mapping of rna-seq reads for splice junction discovery. *Nucleic Acids Res 38*(18), e178.

Wang, L., Z. Feng, X. Wang, X. Wang, and X. Zhang (2010, Jan). Degseq: an r package for identifying differentially expressed genes from rna-seq data. *Bioinformatics 26*(1), 136–8.

Wang, L., S. Wang, and W. Li (2012, Aug). Rseqc: quality control of rna-seq experiments. *Bioinformatics 28*(16), 2184–5.

Wang, P., H. Tang, M. P. Fitzgibbon, M. McIntosh, M. Coram, H. Zhang, E. Yi, and R. Aebersold (2007). A statistical method for chromatographic alignment of lc-ms data. *Biostatistics 8*(2), 357–367. 154MP Times Cited:26 Cited References Count:19.

Wang, P., H. Tang, H. Zhang, J. Whiteaker, A. G. Paulovich, and M. Mcintosh (2006). Normalization regarding non-random missing values in high-throughput mass spectrometry data. *Pac Symp Biocomput*, 315–26.

Wang, Q., A. L. Rozelle, C. M. Lepus, C. R. Scanzello, J. J. Song, D. M. Larsen, J. F. Crish, G. Bebek, S. Y. Ritter, T. M. Lindstrom, I. Hwang, H. H. Wong, L. Punzi, A. Encarnacion, M. Shamloo, S. B. Goodman, T. Wyss-Coray, S. R. Goldring, N. K. Banda, J. M. Thurman, R. Gobezie, M. K. Crow, V. M. Holers, D. M. Lee, and W. H. Robinson (2011). Identification of a central role for complement in osteoarthritis. *Nat Med 17*(12), 1674–9. Wang, Qian Rozelle, Andrew L Lepus, Christin M Scanzello, Carla R Song, Jason J Larsen, D Meegan Crish, James F Bebek, Gurkan Ritter, Susan Y Lindstrom, Tamsin M Hwang, Inyong Wong, Heidi H Punzi, Leonardo Encarnacion, Angelo Shamloo, Mehrdad Goodman, Stuart B Wyss-Coray, Tony Goldring, Steven R Banda, Nirmal K Thurman, Joshua M Gobezie, Reuben Crow, Mary K Holers, V Michael Lee, David M Robinson, William H K08 AR057859/AR/NIAMS NIH HHS/ K08 AR057859-02/AR/NIAMS NIH HHS/ N01 HV 28183/HV/NHLBI NIH HHS/ N01 HV028183/HV/NHLBI NIH HHS/ NS069375/NS/NINDS NIH HHS/ R01 AR051749/AR/NIAMS NIH HHS/ R01 DK076690/DK/NIDDK NIH HHS/ T32 AR007530/AR/NIAMS NIH HHS/ Nat Med. 2011 Nov 6;17(12):1674-9. doi: 10.1038/nm.2543.

Wang, W., H. Zhou, H. Lin, S. Roy, T. A. Shaler, L. R. Hill, S. Norton, P. Kumar, M. Anderle, and C. H. Becker (2003, Sep). Quantification of proteins and metabo-

lites by mass spectrometry without isotopic labeling or spiked standards. *Anal Chem 75*(18), 4818–26.

Wang, W. Y. S., B. J. Barratt, D. G. Clayton, and J. A. Todd (2005, Feb). Genome-wide association studies: theoretical and practical concerns. *Nat Rev Genet 6*(2), 109–18.

Wang, Y. and M. A. Beydoun (2007). The obesity epidemic in the united states–gender, age, socioeconomic, racial/ethnic, and geographic characteristics: a systematic review and meta-regression analysis. *Epidemiol Rev 29*, 6–28. Wang, Youfa Beydoun, May A R01 DK 63383/DK/NIDDK NIH HHS/ Epidemiol Rev. 2007;29:6-28. Epub 2007 May 17.

Wang, Z., M. Gerstein, and M. Snyder (2009). Rna-seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics 10*(1), 57–63.

Waters (2011, June 6, 2011). New high definition mass spectrometer gives researchers the most sensitive and selective discovery tool. Technical report.

Weber, A. P. M., K. L. Weber, K. Carr, C. Wilkerson, and J. B. Ohlrogge (2007, May). Sampling the arabidopsis transcriptome with massively parallel pyrosequencing. *Plant Physiol 144*(1), 32–42.

West, M. (1992). Mixture-models, monte-carlo, bayesian updating and dynamic-models. *Computing Science and Statistics : Vol 24*, 325–333. Bz23b Times Cited:0 Cited References Count:0.

Wilkins, M. R. and K. L. Williams (1997). Cross-species protein identification using amino acid composition, peptide mass fingerprinting, isoelectric point and molecular mass: A theoretical evaluation. *Journal of Theoretical Biology 186*(1), 7–15. Wz934 Times Cited:87 Cited References Count:20.

Willer, C. J., E. K. Speliotes, R. J. Loos, S. Li, C. M. Lindgren, I. M. Heid, S. I. Berndt, A. L. Elliott, A. U. Jackson, C. Lamina, G. Lettre, N. Lim, H. N. Lyon, S. A. McCarroll, K. Papadakis, L. Qi, J. C. Randall, R. M. Roccasecca, S. Sanna, P. Scheet, M. N. Weedon, E. Wheeler, J. H. Zhao, L. C. Jacobs, I. Prokopenko, N. Soranzo, T. Tanaka, N. J. Timpson, P. Almgren, A. Bennett, R. N. Bergman, S. A. Bingham, L. L. Bonnycastle, M. Brown, N. P. Burtt, P. Chines, L. Coin, F. S. Collins, J. M. Connell, C. Cooper, G. D. Smith, E. M. Dennison, P. Deodhar, P. Elliott, M. R. Erdos, K. Estrada, D. M. Evans, L. Gianniny, C. Gieger, C. J. Gillson, C. Guiducci, R. Hackett, D. Hadley, A. S. Hall, A. S. Havulinna, J. Hebebrand, A. Hofman, B. Isomaa, K. B. Jacobs, T. Johnson, P. Jousilahti, Z. Jovanovic, K. T. Khaw, P. Kraft, M. Kuokkanen, J. Kuusisto, J. Laitinen, E. G. Lakatta, J. Luan, R. N. Luben, M. Mangino, W. L. McArdle, T. Meitinger, A. Mulas, P. B. Munroe, N. Narisu, A. R. Ness, K. Northstone, S. O'Rahilly,

C. Purmann, M. G. Rees, M. Ridderstrale, S. M. Ring, F. Rivadeneira, A. Ruokonen, M. S. Sandhu, J. Saramies, L. J. Scott, A. Scuteri, K. Silander, M. A. Sims, K. Song, J. Stephens, S. Stevens, H. M. Stringham, Y. C. Tung, T. T. Valle, C. M. Van Duijn, K. S. Vimaleswaran, P. Vollenweider, et al. (2009). Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. *Nat Genet 41*(1), 25–34. Willer, Cristen J Speliotes, Elizabeth K Loos, Ruth J F Li, Shengxu Lindgren, Cecilia M Heid, Iris M Berndt, Sonja I Elliott, Amanda L Jackson, Anne U Lamina, Claudia Lettre, Guillaume Lim, Noha Lyon, Helen N McCarroll, Steven A Papadakis, Konstantinos Qi, Lu Randall, Joshua C Roccasecca, Rosa Maria Sanna, Serena Scheet, Paul Weedon, Michael N Wheeler, Eleanor Zhao, Jing Hua Jacobs, Leonie C Prokopenko, Inga Soranzo, Nicole Tanaka, Toshiko Timpson, Nicholas J Almgren, Peter Bennett, Amanda Bergman, Richard N Bingham, Sheila A Bonnycastle, Lori L Brown, Morris Burtt, Noel P Chines, Peter Coin, Lachlan Collins, Francis S Connell, John M Cooper, Cyrus Smith, George Davey Dennison, Elaine M Deodhar, Parimal Elliott, Paul Erdos, Michael R Estrada, Karol Evans, David M Gianniny, Lauren Gieger, Christian Gillson, Christopher J Guiducci, Candace Hackett, Rachel Hadley, David Hall, Alistair S Havulinna, Aki S Hebebrand, Johannes Hofman, Albert Isomaa, Bo Jacobs, Kevin B Johnson, Toby Jousilahti, Pekka Jovanovic, Zorica Khaw, Kay-Tee Kraft, Peter Kuokkanen, Mikko Kuusisto, Johanna Laitinen, Jaana Lakatta, Edward G Luan, Jian'an Luben, Robert N Mangino, Massimo McArdle, Wendy L Meitinger, Thomas Mulas, Antonella Munroe, Patricia B Narisu, Narisu Ness, Andrew R Northstone, Kate O'Rahilly, Stephen Purmann, Carolin Rees, Matthew G Ridderstrale, Martin Ring, Susan M Rivadeneira, Fernando Ruokonen, Aimo Sandhu, Manjinder S Saramies, Jouko Scott, Laura J Scuteri, Angelo Silander, Kaisa Sims, Matthew A Song, Kijoung Stephens, Jonathan Stevens, Suzanne Stringham, Heather M Tung, Y C Loraine Valle, Timo T Van Duijn, Cornelia M Vimaleswaran, Karani S Vollenweider, Peter Waeber, Gerard Wallace, Chris Watanabe, Richard M Waterworth, Dawn M Watkins, Nicholas Wellcome Trust Case Control Consortium Witteman, Jacqueline C M Zeggini, Eleftheria Zhai, Guangju Zillikens, M Carola Altshuler, David Caulfield, Mark J Chanock, Stephen J Farooqi, I Sadaf Ferrucci, Luigi Guralnik, Jack M Hattersley, Andrew T Hu, Frank B Jarvelin, Marjo-Riitta Laakso, Markku Mooser, Vincent Ong, Ken K Ouwehand, Willem H Salomaa, Veikko Samani, Nilesh J Spector, Timothy D Tuomi, Tiinamaija Tuomilehto, Jaakko Uda, Manuela Uitterlinden, Andre G Wareham, Nicholas J Deloukas, Panagiotis Frayling, Timothy M Groop, Leif C Hayes, Richard B Hunter, David J Mohlke, Karen L Peltonen, Leena Schlessinger, David Strachan, David P Wichmann, H-Erich McCarthy, Mark I Boehnke, Michael Barroso, Ines Abecasis, Goncalo R Hirschhorn, Joel N Genetic Investigation of ANthropometric Traits Consortium 01-HG-65403/HG/NHGRI NIH HHS/ 068545/Z/02/Wellcome Trust/United Kingdom 076113/Wellcome Trust/United Kingdom 076467/Z/05/Z/Wellcome Trust/United Kingdom 077011/Wellcome Trust/United Kingdom 077016/Wellcome Trust/United King-

dom 079557/Wellcome Trust/United Kingdom 082390/Wellcome Trust/United Kingdom 089061/Wellcome Trust/United Kingdom 1RL1MH083268/MH/NIMH NIH HHS/ 1Z01 HG000024/HG/NHGRI NIH HHS/ 5UO1CA098233/CA/NCI NIH HHS/ CA49449/CA/NCI NIH HHS/ CA50385/CA/NCI NIH HHS/ CA65725/CA/NCI NIH HHS/ CA67262/CA/NCI NIH HHS/ CA87969/CA/NCI NIH HHS/ DK062370/DK/NIDDK NIH HHS/ DK072193/DK/NIDDK NIH HHS/ DK075787/DK/NIDDK NIH HHS/ F32 DK079466/DK/NIDDK NIH HHS/ F32 DK079466-01/DK/NIDDK NIH HHS/ FS/05/061/19501/British Heart Foundation/United Kingdom G0000649/Medical Research Council/United Kingdom G0000934/Medical Research Council/United Kingdom G0601261/Medical Research Council/United Kingdom HG02651/HG/NHGRI NIH HHS/ HL084729/HL/NHLBI NIH HHS/ HL087679/HL/NHLBI NIH HHS/ K23 DK067288/DK/NIDDK NIH HHS/ K23 DK080145/DK/NIDDK NIH HHS/ K23 DK080145-01/DK/NIDDK NIH HHS/ MC U147585819/Medical Research Council/United Kingdom N01-AG-1-2109/AG/NIA NIH HHS/ R01 DK029867/DK/NIDDK NIH HHS/ R01 DK072193-03/DK/NIDDK NIH HHS/ T32DK07191/DK/NIDDK NIH HHS/ U.1475.00.003.00010.02 (85819)/Medical Research Council/United Kingdom Nat Genet. 2009 Jan;41(1):25-34. doi: 10.1038/ng.287. Epub 2008 Dec 14.

Williams, C. M. (2004). Lipid metabolism in women. *Proc Nutr Soc 63*(1), 153–60. Williams, Christine M England Proc Nutr Soc. 2004 Feb;63(1):153-60.

Wu, T. D. and S. Nacu (2010, Apr). Fast and snp-tolerant detection of complex variants and splicing in short reads. *Bioinformatics 26*(7), 873–81.

Yu, T., Y. Park, J. M. Johnson, and D. P. Jones (2009). aplcms–adaptive processing of high-resolution lc/ms data. *Bioinformatics 25*(15), 1930–6. Yu, Tianwei Park, Youngja Johnson, Jennifer M Jones, Dean P 1P01ES016731-01/ES/NIEHS NIH HHS/United States 1UL1RR025008-01/RR/NCRR NIH HHS/United States 2P30A1050409/PHS HHS/United States Research Support, N.I.H., Extramural Research Support, Non-U.S. Gov't England Bioinformatics (Oxford, England) Bioinformatics. 2009 Aug 1;25(15):1930-6. Epub 2009 May 4.

Zechner, R., P. C. Kienesberger, G. Haemmerle, R. Zimmermann, and A. Lass (2009). Adipose triglyceride lipase and the lipolytic catabolism of cellular fat stores. *J Lipid Res 50*(1), 3–21. Zechner, Rudolf Kienesberger, Petra C Haemmerle, Guenter Zimmermann, Robert Lass, Achim F 3001-B19/Austrian Science Fund FWF/Austria F 3002-B19/Austrian Science Fund FWF/Austria W 901-B12/Austrian Science Fund FWF/Austria J Lipid Res. 2009 Jan;50(1):3-21. doi: 10.1194/jlr.R800031-JLR200. Epub 2008 Oct 23.

Zerbino, D. R. and E. Birney (2008). Velvet: algorithms for de novo short read assembly using de bruijn graphs. *Genome research 18*(5), 821–829.

Zhang, J. Q., E. Gonzalez, T. Hestilow, W. Haskins, and Y. F. Huang (2009). Review of peak detection algorithms in liquid-chromatography-mass spectrometry. *Current Genomics 10*(6), 388–401. 487AB Times Cited:7 Cited References Count:41.

Zhang, Q. and Y. Wang (2004). Trends in the association between obesity and socioeconomic status in u.s. adults: 1971 to 2000. *Obes Res 12*(10), 1622–32. Zhang, Qi Wang, Youfa 1 R01 DK63383-01/DK/NIDDK NIH HHS/ Obes Res. 2004 Oct;12(10):1622-32.

Zhang, X. A., J. M. Asara, J. Adamec, M. Ouzzani, and A. K. Elmagarmid (2005). Data pre-processing in liquid chromatography-mass spectrometry-based proteomics. *Bioinformatics 21*(21), 4054–4059. 978VM Times Cited:21 Cited References Count:9.

Zhang, Z. (2012, Apr). Retention time alignment of lc/ms data by a divide-and-conquer algorithm. *J Am Soc Mass Spectrom 23*(4), 764–72.

Zhao, J. (2007). gap: Genetic analysis package.

Zheng, W., L. M. Chung, and H. Zhao (2011). Bias detection and correction in rna-sequencing data. *BMC Bioinformatics 12*, 290.

Zhou, V. W., A. Goren, and B. E. Bernstein (2011, Jan). Charting histone modifications and the functional organization of mammalian genomes. *Nat Rev Genet 12*(1), 7–18.

Zhu, W., X. Wang, Y. Ma, M. Rao, J. Glimm, and J. S. Kovach (2003, Dec). Detection of cancer-specific markers amid massive mass spectral data. *Proc Natl Acad Sci U S A 100*(25), 14666–71.

Zillikens, M. C., M. Yazdanpanah, L. M. Pardo, F. Rivadeneira, Y. S. Aulchenko, B. A. Oostra, A. G. Uitterlinden, H. A. Pols, and C. M. van Duijn (2008). Sex-specific genetic effects influence variation in body composition. *Diabetologia 51*(12), 2233–41. Zillikens, M C Yazdanpanah, M Pardo, L M Rivadeneira, F Aulchenko, Y S Oostra, B A Uitterlinden, A G Pols, H A P van Duijn, C M Germany Diabetologia. 2008 Dec;51(12):2233-41. doi: 10.1007/s00125-008-1163-0. Epub 2008 Oct 7.

Biography

Ashlee Marie Benjamin was born on October 31, 1986 in Syracuse, NY. Raised in Bernhard's Bay, NY, she graduated from Paul V. Moore High School in 2004. Ashlee attended Rochester Institute of Technology from 2004-2009, attaining a BS and MS in Bioinformatics. In the fall of 2009, Ashlee entered the Computational Biology and Bioinformatics PhD program at Duke University. During her first year at Duke, Ashlee completed research rotations in the laboratories of Jeanette McCarthy, Uwe Ohler, and Joseph Lucas. In the fall of 2010, Ashlee joined the lab of Joseph Lucas for the duration of her graduate studies. Ashlee's research has resulted in the publications listed below:

1. Genome-wide association study of Lp-PLA(2) activity and mass in the Framingham Heart Study. Suchindran S, Rivedal D, Guyton JR, Milledge T, Gao X, **Benjamin A**, Rowell J, Ginsburg GS, McCarthy JJ. PLoS Genet. 2010 Apr 29;6(4):e1000928. doi: 10.1371/journal.pgen.1000928.

2. Gene by sex interaction for measures of obesity in the framingham heart study. **Benjamin AM**, Suchindran S, Pearce K, Rowell J, Lien LF, Guyton JR, McCarthy JJ. J Obes. 2011;2011:329038. doi: 10.1155/2011/329038. Epub 2010 Dec 26.

3. Transcription initiation patterns indicate divergent strategies for gene regulation at the chromatin level. Rach EA, Winter DR, **Benjamin AM**, Corcoran

DL, Ni T, Zhu J, Ohler U. PLoS Genet. 2011 Jan 13;7(1):e1001274. doi: 10.1371/journal.pgen.1001274.

4. A Flexible Statistical Model for Alignment of Open-Platform Proteomics Data Incorporating Ion Mobility and High Energy Information. **Benjamin AM**, Thompson JW, Soderblom EJ, Geromanos SJ, Henao R, Kraus VB, Moseley MA, Lucas JE. Under Review at BMC Bioinformatics.

5. Comparing Reference-Based RNA-Seq Mapping Methods for Non-Human Primate Data. **Benjamin AM**, Nichols M, Burke T, Ginsburg GC, Lucas JE. Under Review at BMC Genomics.

Ashlee will remain at Duke University as a Post Doctoral Associate with Geoffrey Ginsburg and Barbara Engelhardt.