Sequencing-based methods for identifying impactful genomic alterations in cancers

by

Isaac C. Joseph

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Computational Biology

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Lior Pachter, Chair
Professor Joseph Costello
Associate Professor Haiyan Huang
Professor Anthony Joseph

Fall 2016

The dissertation of Isaac C. Joseph, titled Sequencing-based methods for identifying impactful genomic alterations in cancers, is approved:

Chair _____ Date _____

_____ Date _____

_____ Date _____

_____ Date _____

University of California, Berkeley

Sequencing-based methods for identifying impactful genomic alterations in cancers

Abstract

Sequencing-based methods for identifying impactful genomic alterations in cancers

by

Isaac C. Joseph

Doctor of Philosophy in Computational Biology

University of California, Berkeley

Professor Lior Pachter, Chair

Recent advances in collecting sequencing data from tumors is promising for both immediate individual patient treatment and investigation of cancer mechanisms. A resultant central goal is identify changes in tumors that are impactful towards these ends. Here, we develop two tools to identify impactful changes at different levels. We develop both methods in the context of gliomas, a common form of brain cancer. Firstly, we develop and assess a tool for assessing the impact of fusion genes, a type of common mutation using RNA-sequencing data. We validate the tool by working with collaborators in The Cancer Genome Atlas. Secondly, we develop a tool for an overall assessment of patient outcome by integrating data from diverse sequencing platforms. We validate this tool using simulation, data from consortiums, and collaborators at UCSF.

To those that suffer

It is you that have convinced me that there is something still worth fighting for in this world.

Contents

List of Figures

List of Tables

Chapter 1

Goal of collecting genomic data from cancers

1.1 Context

In the start of the second millennium, scientific inquiry as a whole has remained relatively absent from political discourse and relatedly, the collective consciousness of the nation. Part of this may be due to the end of the Cold War around 20 years prior, which removed the impetus for international scientific competition; another part may be due to the sufficient progress science has made in many fields, leading the way to technologies and engineering to take the forefront and to translate discoveries into general comfort.

Health remains an important area, although even here science seems to have sufficiently addressed most aspects – infectious diseases and most non-infectious diseases are now much less deadly than they were in the recent past. To those with sufficient means in the United States and most western countries, developments based on scientific inquiry serves them well.

Arguably the biggest exception to this is cancer; in the last 100 years, cancer mortality as a whole has only very slightly decreased. One reason for this may be the mechanistic uniqueness of the disease in every individual.

Encouragingly, the sequencing of the human genome in 2001 and the subsequent decreasing cost of genomic sequencing methods has yielded some promise in cancer treatment possibilities. In particular, machines that can gather sequencing information from tumors ("genomic data") may be able to be used for an increasing amount of the population, financially, and these might be able to lead to the appropriate handling of uniqueness to decrease mortality.

In 2015, United States' President Barack Obama announced the Precision Medicine initiative during his State of the Union address. During the same address in 2016, he announced the National Cancer Moonshot Initiative. Both were rightfully meant to focus attention on this problem.

In this thesis, I make use of three large-scale multi-institutional consortium-based efforts and one smaller-scale intra-institutional effort to idnetify impactful genomic alternations

using novel tools and approaches. First, I outline the overall goals of the collection of this data, which are twofold:

1.2 Investigation

In order to successfully use genomics data on a personal level for cancer treatment, much needs to be collected in order to increase our understanding of how that genomics data is related to properties of tumors and treatment decisions. There are several ongoing and recently completed large-scale efforts to collect this data in order to achieve this. These can be seen as investigations towards understanding specific areas of unknown about the mechanisms of (1) cancer onset (oncogenesis), (2) progression, and (3) metastasis (spread to distant regions of the body).

Identification of molecular subtypes

One area of unknown is the existence of molecular subtypes of tumors – groups of tumors based on the molecular aberrations they share. This is in contrast with the classical tissue definition of tumors, which is based on the type of tissue from which they originated. The Cancer Genome Atlas (TCGA)[36], a major consortium which gathered 11,000 tumor samples from 33 classically-defined cancer types, aims to help assess this difference. I later describe identification of impactful features from one tumor type contained in TCGA.

Analyses under the TCGA, which gather sequencing data and compare it across tumor types, have made some findings that begin to suggest that a significant fraction tumors that are defined classically would be better assessed based on their molecular similarity with a non-classically equivalent tumor.

One area of success in molecular subtype definition has been gliomas. Gliomas are one of the most common forms of brain cancer, comprising 30% of brain and central nervous system and 80% of all malignant brain tumors. Gliomas are classically classified by World Health Organization (WHO) grade (II through IV for adults), and the glial subtype that the tumor cells most resemble based on examining tumor tissue from biopsies under the microscope (histopathology): oligodendroglioma, astrocytoma, or oligoastrocytoma for tumors appearing similar to oligodendrocytes, astrocytes, or both, respectively.

Somewhat expectedly based on prior related research which suggested yet was not able to provide sufficient evidence for a definitive conclusion, TCGA researchers studying sequencing information from over 300 lower-grade glioma (LGG) tumors identified three molecular subtypes of the tumors based on sequencing information. LGG tumors were defined based on histopathologists' assessments of tumor tissue as being indicative of grade II of any histopathology.

This led to finding consistent subgroups of patterns across all types of sequencing information. Importantly, these subgroups were found to correlate more closely with important patient outcomes than the classical classifications (grade and glial subtype resembled). Inter-

estingly, researchers were also able find that subtypes, while evidenced by much sequencing information across the entire genome, were well defined by a very small number nuclear changes as well, suggesting possible targets for therapies. Researchers also identified that one molecular subtype appeared very similar to tumors whose classical classification would have indicated a higher grade, and these tumors had similarly poor prognosis as these higher-grade tumors.

In addition to finding subtypes within one specific classical tumor diagnosis category and comparing those subtypes with other classical categories in an ad-hoc manner, TCGA has also implemented the search for subtypes with so-called Pan-cancer analysis[66] studies. Based on clustering, researchers assessed classes of tumors in a classical category-agnostic fashion. This is consistent with the molecular understandings of (initiation and onset of tumor tissue (oncogenesis). In particular, the same general mutational patterns involving the same classes of genes are likely to have the ability to initiate tumor-like properties across classical tumor types.

In the context of cancer genomics, mutations are defined as changes in the DNA of a tumor that are not also present in the DNA of non-tumor cells. These are sometimes termed "somatic mutations," to underscore their occurrence within the tumor tissue of the individual in which they are detected, and are contrasted with "germline mutations.

The major finding here in terms of subtypes was the stratification of classical tumor categories on a spectrum from tumors with a high mutational burden (many small aberrations in genomic DNA present in tumor tissue as compared to normal tissue derived from blood) to those with a high copy number change burden (many large-scale aberrations in genomic DNA). Interestingly, no classical category appeared to have high amounts of both type of mutation; this may point to similarities in oncogenesis mechanisms in tissues that are similar in terms of mutational burden type, and two general oncogenic tumor categories.

Assessing clinical utility of molecular subtypes

A further area of investigation is whether subtypes identified based on consistent molecular patterns have clinical utility.

One promising result is in gliomas, where the finding that molecular subtypes were more accurate at distinguishing clinical outcomes can now be used by clinicians to treat patients with the privilege of having had their genome sequenced; the WHO is in the process of updating the standard of care for LGG patients, which will formalize this new ability to treat patients more effectively. In particular, those with a prognosis similar to higher grade tumors based on molecular subtyping can be treated accordingly; this may involve more aggressive use of chemotherapy and/or radiotherapy.

Towards this end, the Cancer Cell Line Encyclopaedia (CCLE)[4] and the Cancer Genome Project (CGP) [19]are testing a variety of new targeted therapies *in vitro* in a relatively knowledge-agnostic approach. I use data from both consortiums towards this goal, as well. Targeted therapies are pharmaceutical agents that interrupt the activity of a particular molecular process, typically by interfering with a protein. As part of the rational development

of cancer therapies, they are typically chosen in a strategic way to interrupt key pathways that are important in particular types. The molecular similarities between tumors of different types justifies the relatively unbiased approaches used by both studies, which contrasts with a one-tissue-type, one-drug approach.

In particular, for example, drugs targeting the epidermal growth factor receptor (EGF) pathway have been shown to be effective against multiple different tumor tissues of origin [4].

Assessing relative impact of genetic and epigeneitc changes

A second area of open investigation is the role of epigenetics in oncogenesis, tumor progression, and metastasis. Epigenetic changes are consist of non-base-pair DNA changes to the genome based on modification of the genome's surroundings; in particular, chromatin can be opened are closed via a myraid of mechanisms, such as histone modifications and DNA methylation[25].

These have an unknown degree of impact; it is not known, for example, whether such changes ("epimutations") are sufficient for oncogenesis [39] [15].

Towards that end, Dr. Joseph Costello, Ph.D. is spearheading a University of California, San Francisco (UCSF)-based approach to gather both genetic and epigenetic data on LGG tumors. He has found already DNA methylation-based epigenetic changes correspond closely with alternations in DNA base pairs[23].

Identification of driver mutations

Another general goal of collecting genomic data from tumors, and one of the first goals, has been the identification of previously unobserved "driver" mutations. Driver mutations are DNA changes that are "causally implicated in oncogenesis"[55]. Driver mutations are typically validated via *in vitro* and *in vivo* experiments. The number of possible driver mutations is as endless as the number of possible mutations, so therefore candidate driver mutations need to be established in order to narrow down the search space of possible drivers.

One common prioritzation technique is to identify somatic mutations that occur more often than what might be expected based on chance ("recurrent mutations"). This technique requires statistical power, necessitating, in turn, the collection of many tumor samples[31]. In particular, power is required to identify candidate driver mutations with either low-impact or rare occurence. Many of these are believed to exist, based on similar properties of variants detected during the conceptually similar genome-wide association studies (GWAS), which search for germline variants associated with a particular phenotype, typically disease-related.

Further prioritization may be obtained based on somatic mutation annotations, which may be based on prior knowledge of a particular mutation's genomic context and its function. In particular, intergenic status, knowledge of domain structure, or knowledge of regulatory regions might be used to prioritize one candidate driver mutation for testing over another. Most somatic mutations are point mutations[29], although many will involve normally disparate genomic regions, which are used in a similar fashion.

Many tools have been developed to harness this data, and many have been applied to the specific datasets I explore later (TCGA,CGP, UCSF), including MuTect [12].

Assessing relative impact of germline and somatic variants

Some cancers are known to have a strong germline genetic component in terms of incidence risk [54]; this is especially true of childhood tumors. However, in general, it's unknown for all cancer types how much of an effect germline variants, rather than somatic mutations, have in terms of oncogenesis, tumor progression, and metastases. This question can be probed with large scale data involving matched normals, which allows the separation of germline variants from somatic mutations and also the tracking of tumor outcomes.

This is one of the aims of collecting sequencing data from glioma patients in the Costello lab; importantly, follow-up information is present for patients, which is nontrivial due to the high mean survival time of such patients (on the scale of 10 years, depending on molecular subtype).

Molecular mechanisms explaining variance in treatment response

In some cancer types, there exist a wide variance in how patients respond to identical treatment and identical cancers (modulo to classical clinical covariates). The molecular reasons behind this are currently unknown in most cases.

Gliomas

This is especially true gliomas, and uncovering this is a major aim for collecting surgical tissue by the Costello lab.

In particular, LGG patients of subtype astrocytoma (identified based on histopathology) vary in their response to current chemotherapeutic agents.

Typically, LGG patients are treated by (1) surgical resection of the tumor, and (2) "adjuvant" chemotherapy. Chemotherapy is typically temozolomide (TMZ), although may also be procarbazine-lomustine-vincristine (PCV). The purpose of adjuvant chemotherapy is to kill remaining tumor tissue that was not extracted during surgery, which is common due to the difficulties of surgery in the context of vital brain areas (eloquent regions) that cannot be easily removed or removed at all without substantially decreasing the patient's quality of life.

LGG has an almost 100% recurrence rate; surgical tissue from the recurrent tumor (second surgery) is also collected by the Costello Lab for further study. Some patients that receive adjuvant TMZ have been found to progress to a higher grade of glioma [23], which is termed undergoing a *malignant transformation* (MT).

Towards the end of understanding why some patients undergo MT whereas others don't, the Costello lab is looking for clues in the genomic information collected on tumor tissue collected during both sugeries.

One lead is a large fraction TMZ-treated patients that undergo MT have a drastically higher rate of DNA mutations than those that do not undergo MT, termed *hypermutation*. This suggests the existence of one molecular pathway relating to the deviating response to TMZ in these patients involving hypermutation, possibly related to the inactivation of particular DNA mismatch repair genes in these tumors.

An open question, then, is if this negative response (MT) can be predicted based on surgical tissue collected during primary tumors; this is one I attempt to investigate, and describe later in this document.

1.3 Making individualized treatment decisions ("precision medicine")

A complementary and ultimate goal of collecting genomic data from tumor samples is to directly influence the course of an individual's cancer treatment regimen. This is motivated by the notion that cancer genomics can collect data that can distinguish previously indistinguishable cases in a way that allows for their more successful treatment.

Methods for treatment of a particular tumor are based on *standards of care*, which are roughly legally-defined and clinically-implemented notions of treatment regimens for a particular diagnosis. These vary a substantial amount between treatment centers, may or may not be formally established legally, and change due to new medical findings being published in medical journals. They may also be set formally by global organizations, such as WHO [38].

Non-genomic methods

Many standard methods for cancer treatment do not rely on the collection of genomic data from a tumor.

clinical covariates are non-tumor, non-genomic attributes such as age and gender which may affect which treatment regimen is proscribed for a particular patient according to a standard of care. For example, a patient of advanced age with a new prostate cancer may not be treated, as it will be assumed the patient will die with, rather than of, the tumor; a younger patient, however, may be treated with surgery and/or cytotoxic chemotherapy.

Treatment decision methods involving properties of the tumor(s) themselves can be divided into *invasive* and *noninvasive* procedures.

noninvasive procedures don't involve breaking the skin, which for tumor inspection means imaging techniques. For gliomas, magnetic resonance encephalopathy (MRI) is used in order to assess tumor grade, stage, and possible recurrence, which in turn influences treatment decisions. Positron emission tomography, electroencephalography, and other similar imaging methods are also used for various tumor types as well.These techniques generally involve analysis of images to assess tumor location, size. Molecular properties can also be inferred

through the use of specific imaging able to detect various metabolites which may be present in a tumor.

invasive procedures involve breaking the skin. All of these are typically considered to be biopses, and involve collecting tumor tissue for physical examination. A typical method by which biopsy tissue can be examined is *histopathology*, which involves examining a slice of tumor tissue under a microscope after staining procedures. This is done in gliomas in order to distinguish grades II, III, and IV based presence of necrotic tissue, cellularity, and other visible properties.

Genomic methods

Increasingly, genomics data is being used clinically and therefore collected for the purpose of making treatment decisions on an individualized basis.

Gene expression

Genomic approaches for treatment prognostication and stratification have focused on gene expression, often estimated through the use of microarray or RNA-sequencing technologies which measure the abundance of RNA in cells through sequencing. The successful discovery of a 231-gene expression signature in 2002 related to breast cancer prognosis[62] spurred the search for signatures in various tumor types for various outcomes. There are currently several genomic signatures from various tumor types that allow for patient stratification. Several metrics of interest exist in order to define *impactful* from the context of making treatment decisions. Essentially, all are a method prognostication of patient outcome, conditioned on specific treatment decisions: (1) survival time, (2) susceptibility to treatment with a specific drug, (3) grading/ malignancy classification of a tumor, and (4) susceptibility to specific events. Events of interest include (4.1) tumor progression, (4.2) tumor metastasis, which is closely related to (4.3) tumor invasiveness, (4.4) functional status of a particular molecular pathway within a tumor (related to (2)), and (4.5) likelihood of recurrence subsequent to tumor resectilevel

In particular, several of these are currently used in the clinic for prognostication, including MammaPrint® 70-gene signature and Genomic Health's OncotypeDX® 21-gene signature for breast cancer, and Veracyte's Affirma® for thyroid cancer. These may be used to decide whether to provide chemotherapy or other further follow-up treatment.

Somatic mutations

Somatic mutations are also increasingly used in the clinic. Specific tests exist for signatures involving a few genes such that they can be queried a patient-level without the use of high-throughput sequencing. Common options include polymerase chain reaction (PCR) or antibody stain of collected tissues. Also, companies such as Foundation Genomics ® perform large-scale targeted sequencing for a handfull of genes known to be generally impactful or

targetable by specific therapies. Specific cancer tests include avian erythroblastic leukemia viral oncogene homolog 2 (ERBB2) for breast cancer via AviaraDx ®.

In gliomas, both of these methods are used; an antibody stain for Isocitrate Dehydrogenase 1 (IDH1) are being used in some clinics for prognostic tests, including UCSF. In particular, IDH-wildtype gliomas have a drastically lower survival time. This is a proximal use of the tissue that is further analyzed by sequencing and by the Costello lab.

DNA methylation

DNA methylation-based collection is relatively new for the use of direct patient stratification; signatures exist for Acute Myeloid Leukemia that relate to overall survival [16], but are not currently used clinically.

Gliomas are at the forefront of this, where the status of the methylation of the promotoer of the O^6-Guanine Methyltransferase (MGMT) gene is queried by antibody, which has found to be prognostic. In particular, lack of methylation here is associated with high MGMT activity and therefore low chemotherapeutic benefit, so is often a major component in treatment decisions [46]. This was a proximal use of tumor tissue further analyzed by the Costello lab and myself later in this document.

Germline variants

Finally, properties of patients' germlines are occasionally used for prognostication. In breast cancer, the appropriately titled breast cancer 1 gene (BRCA1) and breast cancer 2 gene (BRCA2) are queried for prognosis, as there are known heritable deactivating mutations in these tumor-suppressor genes. In gliomas, there is a known heritability of MGMT-methylation response, although this is not currently linked strongly enough to germline variants in order to be used clinically.

Integrated sequencing methods

Integration of features are not currently used for stratification, although multi-type signatures have been collected by CGP relating to response to targeted therapies, as outlined above.

Chapter 2

Properties of analyzed data

Data from several sources were analyzed for the purposes of identifying impactful sequencing-accessible changes. In particular, data from TCGA [13], the Costello Lab at UCSF[23] and the Cancer Cell Line Encyclopaedia[4] were used.

2.1 The Cancer Genome Atlas

Data from TCGA LGG were used to assess the fusion transcript expression estimation tool described in chapter 3. 289 tumors were analyzed by the TCGA LGG Analysis Working Group (AWG).

Raw molecular information

RNA-sequencing

RNA-sequencing protocol was standardized across TCGA data collection centers. In particular, total RNA extraction was performed using 5×10^6 cells using Ambion® Ribopure™ kit. Reads were sequenced with Illumina® HiSeq™ sequencers, collecting 2×50 base-pair paired-end reads.

Exome sequencing

Exome sequencing on tumor tissue was performed using 0.5 to 3 of DNA from tumor and normal blood, respectively. The Agilent®Sure-Select Human All Exon ™v.20, 44 MB kit was used for exome region-targeting, and 2×76 base-pair paired-end sequencing reads were used, again with Illumina®HiSeq™sequencers.

Deep whole-genome sequencing

This was performed on 21 of the 289 samples. 2×101 base-pair paired-end reads were used on the same sequencing platform as exome sequencing.

"Low-pass" whole-genome sequencing

This was performed on 52 of the 289 samples. Here, 0.5 to 0.7μg of DNA were extracted, then KAPA Biosystems®kits were used for preparation. 2×51 base-pair pari-end reads were produced with Illumina®HiSeq™, producing 5.2× average coverage. 39.5 average Phred quality score was achieved, with a 96.5% average mapping rate.

Clinical information

In order assess the impact of sequencing data relative to various outcomes and to also correlate various molecular features with known features, classical clinical information was obtained in addition to the molecular inforation. In particular, histologic type, age of diagnosis, race, year of diagnosis, family history of cancer, extent of tumor resection, tumor location, white matter percentage, and first presenting symptom were collected from all of the 289 patients, with a few omissions. Most notable in terms of fusion expression analysis was survival time, which correlated molecular tumor type and relatedly, the existence of fusions of interest.

Resulting datatypes and details of processing

The above molecular information was processed, resulting in several different high-level genomic features.

DNA rearrangements

Several different DNA rearrangements sets were estimated from data; these were used in order to prioritize resultant FJs.

genome-wide Two sets of genome-wide (as opposed to exome-limited) DNA rearrangements were produced.

deep-sequencing-based

Here, BamBam[50] was used to estimate rearrangements based on exome-sequencing and deep whole-genome sequencing reads. These reads were aligned and processed using the aligners of BWA [33] and Picard [45] Broad Institute Firehose [17]

shallow-sequencing-based

As deep sequencing was only available on a minority of samples, shallow ("low-pass") sequencing was also used to call DNA rearrangements.

For this purpose, the CASAVA [8] tool was used, along with BWA for mapping and BreakDancer[9] and Meerkat[69].

intragenic As more sequencing data was available to support genic region-based rearrangemnets, this was used to estimate rearrangemnets that was directly applicable to FJs.

Criterion	Accepted Value
split reads	≥ 3
spanning reads - split reads	≥ 0
probability from SVM classifier	> 0.65
read through event	`false`
p-value from supporting reads' production from other genomic areas	≤ 0.1
fraction of supporting reads that are repetitive sequence	< 0.78
number of spanning reads	> 1

Table 2.1: Criteria for filtering fusions downstream of deFuse

Criterion	Accepted Value
exon-connecting reads	> 1
spanning reads	> 2
`BLASTN`[1]-computed inter-constituent sequence similarity	< 0.01

Table 2.2: Criteria for filtering fusions as part of PRADA pipeline

MapSplice pipeline MapSplice[64] was used to call FJs as part of the general RNA-seq alignment pipeline.

deFuse pipeline deFuse[37] was used by the Haussler lab at UCSC in order to call FJs from RNA-sequencing data. As part of the deFuse pipeline, TopHat[59] was used, along with a support vector machine (SVM) classifier to assess the probability of fusion existence based on priors collected by a test set.

Reads were further filtered by Drs. Olena Morozova and Sofie Salama at UCSC based on criteria outlined in 2.1. Beyond 2.1, specific FJs bisecting (1) protein kinases, which have a known role in higher-grade gliomas and (2) genes significantly mutated by smaller mutations were prioritized for validation.

PRADA pipeline

PRADA was developed and used by collaborators at the MD Anderson Cancer Center as an independent method of assessing fusions to increase confidence. Briefly, the pipeline involves RNA-seq read alignment, followed by filtering by the following shown in 2.2

Gene expression

Gene expression data was used to validate and prioritize fusions for further fusion-transcript-gnostic expression estimation. Gene expression resulted from RNA-sequencing data and was produced based on the following pipeline:

1. MapSplice for mapping [64]
2. RSEM[32] for expression estimation
3. Filtering to genes expressed in $\geq 70\%$ of samples
4. $\log_2$ transformation
5. Median-centering across samples

2.2 UCSF Department of Neurosurgery glioma samples

Roughly one hundred samples (86 at time of writing) were collected by the UCSF department of neurosurgery and analysed further by the Costello lab. Roughly fifty (48 at time of writing) of these are relevant to the prediction problem that I formulate in chapter 4, based on arising from patients with adjuvant Temozolomide chemotherapy.

Raw molecular information

Matched normal samples

For roughly 40 of the patients, samples from circulating blood were taken and sequenced via a targeted exome sequencing platform with Illumina® reads.

Primary tumor samples

Primary tumor samples were tumor samples collected from the first surgery for LGG patients. This was used in later analyses for the purpose of prognostication regarding outcomes following primary surgery.

On several patients, several tumor sections from one tumor were sampled in order to assess intratumoral heterogeneity. On a minority of these, sections were chosen to maximize variance of metabolic positron-emission tomography (PET)-based imaging as a proxy to maximize genetic variance of sections.

In detail, PET assessed the presence of choline, which is a marker for cell membranes. Cell membrane existence is a marker of cell turnover and therefore a marker fo tumor growth.

Choline is related to lipid isocitrate dehydrogenase 1 (IDH1) operation metabolism, which is known to be altered in some gliomas, so is particularly relevant for glioma metabolic assessment.

Exome sequencing For 27 patients, exome sequencing was performed on at least one section of tumor tissue using identical protocol to 2.2.

RNA-sequencing For around ten patients, RNA-sequencing data was collected using Illumina®paired-end read sequencing protocols. RNA-seq was not performed on more patients due to degradation of RNA.

Human methylation 450k array For roughly 20 patients, the Infinium®HumanMethylation450 BeadChip Kit™(450k) was used to assess CpG methylation at around $500,000$ probe sites.

Secondary tumor samples

Tumor samples were collected from secondary and following surgeries if available; the same datatypes were available as was available on primary tumor samples.

Resulting datatypes and processing details

Germline variants

Germline variants were assessed from exome sequencing of the normal tumor blood samples and processed via the GATK Unified genotype [35]. Briefly, this method works by aligning reads and producing variants at sites differing from an input reference, and results in a list of called variants for every sample on a base-pair resolution.

DNA methylation

Raw probe intensity values were processed by `Methylumi`[14]. In detail, probes were normalized by removing probes with high p-values for detection tests, then background-corrected. Sex chromosomes were filtered, as were probes mapping to SNPs and probes mapping to multiple genomic locatons.

This resulted in per-sample vectors of β values representing average methylation level of CpG site across a sample.

DNA copy number

Genomic copy number segments are called from exome-sequencing information due to newly developed pipelines, and validated with intensity information from 450k.

This results in segments of floating point-numer representing copy number, averaged across cells in a sample. To further process, we took floating point numbers more than 0.5 away from 2.

Somatic mutations

Somatic mutations were called based on combining normal and tumor exome-sequencing information, using GATK. This results each variant of one or more nucleotides on a base-pair resolution between tumor and normal sequencing information.

Type	Unit
age	binned into 5-year intervals
tumor location	$\in$ {frontal, parietal}
extent of resection	$\in$ {gross total, subtotal}
length of TMZ cycles	months
time between surgery and TMZ administration	months
time between TMZ administration and recurrence	months
gender	$\in$ {male, female, neither}

Table 2.3: Clinical information available

There were a group of ampels where the number of mutations was more than $10\times$ higher, on average, than the other group of samples; these samples were considered to having undergone *hypermutation* (HM).

RNA-sequencing

RNA-seq data was analyzed via a pipeline using Kallisto[6] to generate expresion estimates.

Clinical information

A wealth of clinical information is available on the patients. This was used to select an appropriate molecular phenotype that accorded well with clinical outcomes. HM was selected as being related to malignant transformation and relatedly, decreased survival time.

Clinical information was also used as a predictor for the HM. See 2.3 for details.

2.3 Cancer Genome Project

The cancer genome project is a major consortium effort led by the Wellcome Trust Sanger Institute to assess the sensitivity of various cancer types (2.4) to existing pharmaceutical therapies.

Cancer types

At least one cell line each of from 29 cancer types were used (2.4). resulting in $1,001$ different samples.

Names
Adrenocortical carcinoma
Acute lymphoblastic leukemia
Bladder Urothelial Carcinoma
Breast invasive carcinoma
Cervical squamous cell carcinoma and endocervical adenocarcinoma
Chronic Lymphocytic Leukemia
Colon adenocarcinoma and Rectum adenocarcinoma
Lymphoid Neoplasm Diffuse Large B-cell Lymphoma
Esophageal carcinoma
Glioblastoma multiforme
Head and Neck squamous cell carcinoma
Kidney renal clear cell carcinoma
Acute Myeloid Leukemia
Chronic Myelogenous Leukemia
Brain Lower Grade Glioma
Liver hepatocellular carcinoma
Lung adenocarcinoma
Lung squamous cell carcinoma
Medulloblastoma
Mesothelioma
Multiple Myeloma
Neuroblastoma
Ovarian serous cystadenocarcinoma
Pancreatic adenocarcinoma
Prostate adenocarcinoma
Small Cell Lung Cancer
Skin Cutaneous Melanoma
Stomach adenocarcinoma
Thyroid carcinoma
Uterine Corpus Endometrial Carcinoma

Table 2.4: Cancer Genome Project Tumor Types for Derived Cell Lines

Categories
ABL signaling
DNA replication
EGFR signaling
ERK MAPK signaling
Genome integrity
IGFR signaling
JNK and p38 signaling
PI3K signaling
RTK signaling
TOR signaling
WNT signaling
apoptosis regulation
cell cycle
chromain histone acetylation
chromatin histone methylation
chromatin other
cytoskeleton
metabolism
mitosis
other
p53 pathway

Table 2.5: Cancer Genome Project Compound Categories

Compounds information

Raw compound details

A total of 265 compounds were tested, consisting both of drugs under investigation and established therapeutics. The compounds can be considered in several categories based on their molecular targets (2.5) if available. Broadly, compounds were categorized as cytotoxic or targeted.

We focus on analysis of the DNA replication and Cytoskeleton-based categories, both of which are the only cytotoxic classes tested; the rationale for this was that these might be more similar to the phenotype of interest in UCSF datasets.

Resulting sensitivity estimates

For each combination of compound and cell line, a call of `resistant` or `sensitive` was made based on the half maximal inhibitory concentration IC_{50} values.

A heuristic outlier procedure with the following steps was used for each compound:

1. **upsampling** For each cell line, a normal distribution was fit with mean equal to the average IC_{50} over the different samples, and standard deviation equivalent to a 68% estimated confidence interval. 1 000 points were then sampled from the fit normal distribution.

2. **density estimation** A density was estimated from the 1 000 points using kernel density estimation with normal kernels.

3. **modeling resistant population** Based on the shape fo the resultant density estimation, a resistant population of cell lines, for each compound, were identified in an automated fashion. Briefly, a strong prior that the majority of cell lines would be resistant was a major assumption that went into this; the method attempted to find bimodal structure consisting of a mixture of normal distributions; it then assigned the node with the higherIC_{50} as the resistant population.

4. **calling threshold based on cumulative distribution function** Finally, using an empirical cumulative distribution on the estimated sensitive population so as to include most of these, a threshold was called.

Genomic information

Broadly, three different molecular assays were performed on each cell-line-derived sample.

Raw genomic information

exome sequencing A 64-gene panel was sequenced using capillary (Sanger®) sequencing.

copy-number array Affymetrix®SNP6.0™microarrays were used to assess copy number.

expression array Affymetrix®HT-U133A®microarrays were used to gather gene expression information based on RNA transcript abundance-based annealing to predefined probes.

Resulting datatypes and processing details

variants Variants were called within the 64-gene panel as compared to reference. As matched normals were not obtained, variants were a mixture of both somatic mutations and germline variants, likely dominated by somatic mutations.

copy number Copy number information was assessed from the copy-number array using Affymetrix®processing tools.

gene expression Array-based expression estimates were obtained for the 14, 500 genes available on the platform panel.

gene fusions Fusions were called based on the Sanger®sequencing.

Chapter 3

Identifying impactful fusion genes: fusion expression analysis

3.1 Motivation

Acquisition of increasing numbers of somatic mutations on a rapid timescale and in a heterogeneous fashion across different cells within a tumor or pre-tumor tissue, is widespread and occurs in most cancer types. Resulting mutations are sufficient to induce tumorigenesis and also to drive tumor progression and metastasis. This is done classically through the activation of proto-oncogenes and through the deactivation of tumor suppressor genes.

There are several specific pathways by which mutations can be induced and several corresponding classes of mutations[55]. One class is large-scale genomic rearrangement, which involves the breaking and possible rejoining of multiple-megabase sections of DNA.

When large genomic regions dislocate, they may often rejoin in a predictable fashion to regions either on other chromosomes or within the same chromosome (translocations). If there are genic regions spanning the genomic breakpoints, RNA messages may be transcribed that contain two genes from two constituent chromosomes. Such messages, if translated, become fusion genes (FGs).

Due to fragile regions in DNA, specific FGs may be formed to a significant extent in certain tissues during tumorigenesis[72]. In chronic myeloid leukemia (CML), a fusion between breakpoint cluster region (BCR) and Abelson murine leukemia (ABL) virus genes leads to the recurrent BCR-ABL fusion which is present in a large percentage of CML patients. This fusion is known to have potential to transform normal cells into cells with tumor characteristics, and has been successfully targeted by Imatinib, one of the world's first successful targeted therapies in terms of extending patient overall survival time. Several similar examples have been discovered; identifying FGs is thus of primary interest [10].

One goal of major consortium efforts such as The Cancer Genome Atlas (TCGA)[66] and The Cancer Cell Line Encyclopedia (CCLE)[4] is detecting recurrent fusion genes from different caner types based on sequencing information. Recently, for example, recurrent

receptor tyrosine kinase fusions were detected in gliomas[13].

One mechanism whereby fusion genes may have impact on tumor tissue is via the induced expressional deregulation (ED) of the constituents. In particular, a sequence may be placed under the control of regulatory regions that were not originally meant for a sequence, resulting in, for example, the constitutive expression of a growth-factor related protein domain [65], leading to oncogenesis or tumor progression.

3.2 Existing impact-assessment strategies

Detection

Typically, fusions are detected from RNA-sequencing (RNA-seq) reads, which are produced due to its relatively low cost compared to whole genome sequencing and high interpretability. Several computational tools to detect fusions from RNA-seq reads exist. Most identify fusion junctions (FJs), which are the breakpoints of the associated genomic translocations.

Tools identify FJs in three steps: (1) finding chimeric reads (CRs), which is a single read with two portions aligning to two separate genomic locations, (2) aggregating chimeric reads into candidate fusion junctions through realignment-based grouping, and (3) filtering candidate FJs based on heuristic filters.

While simple in concept, identifying FJs is a problematically error-prone process. Firstly, incorrect read mapping leads to spurious CRs. Incorrect mapping is often a result of repetitive regions in the genome such as germline segmental duplications. Secondly, even if reads are correctly aligned, false positives may be generated by read-through events, where genomically adjacent genes are erroneously transcribed into one RNA message or reverse transcriptase template-switching events/ trans-splicing. These produce low but detectable baseline levels of fusion genes in wild-type cells, but are usually not of interest in cancer sequencing efforts as they are not causally involved in tumorigenesis, cancer progression, or metastasis[20].

Thus, candidate FJs must be filtered aggressively by heuristics based on knowledge of the above. One problem is that it's not clear exactly which heuristics to use; many heuristics based on ignoring FGs in repetitive regions, for example, may filter real fusions[27]. Another is that given a set of heuristic filters, it's not clear how to best combine the heuristics in a way that's generalizable to most FG discovery use cases. This is due in part to the wide range of genomic instability that tumors from different tissues have exhibited. A symptom of this is that Machine learning-based classifiers tuned on representative datasets have issues generalizations to new tumor types. The high false-positive rate stymies further discovery and assessment of FGs.

FJ identification is also a very computationally-intensive process, as every read must be split in a number of ways and positions and matched against all possible regions of the genome through alignment during CR-finding steps. One of the reasons for this is that existing fusion discovery methods use computationally intensive first-generation alignment algorithms.

Expression quantification

As ED is a major mechanism by which fusion genes impact tumors and patients, assessing whether an individual fusion detected has also led to ED is a goal for prioritizing fusions for per-patient impact and for further mechanistic study.

Unfortunately, there is no currently established method of doing so in a high-throughput, unbiased fashion. Three methods currently exist:

Constituent gene summing

One approach to assessing whether a fusion gene or transcript is expressionally deregulated is by summing the expression estimates of its constituent transcripts. If the sum of these expression estimates substantially deviates from the same quantity in sample(s) without the fusion, it is concluded that the fusion has led to ED.

However, this is inaccurate due to details involving how much of each constituent transcript makes up the fusion. In particular, a fusion gene can consist of the entire transcripts of the constituents bound, or it can consist of as little as a few bases of each gene. This wide discrepancy leads to the lack of robustness of this method without substantial additional human assessment, which is undesirable as the purpose of high-throughput and automated assessment methods is to minimize human input. The presence of large numbers of false-positive fusions exacerbates this necessity of visual inspection as well.

Per-exon inspection

A related but distinct method of assessing fusion genes is comparison between fusion-present and fusion-absent samples by visually inspecting the per-exon expression of the constituent transcripts. While not susceptible to the problems of the summing method in terms of misestimation, this is even more labor-intensive as it is completely not automated.

Counting junction-spanning reads

One automated approach to assessing fusion expression estimation is by counting the number of reads supporting the fusion junction. In particular, a read can *support* a fusion junction by being a *spanning read* (PR) or *split read* (LR). A PR is a paired-end read wherein one end maps to one transcript and another maps to a genomically distal genomic region, whereas an LR is a read wherein a portions of one read (if applicable) map to at least two distinct genomic regions.

Some method of counting these reads is a common method of prioritizing fusions; however, this is very susceptible to multi-mapping reads. In particular, if a junction region is repeated in another section of the genome, reads could spuriously seem to strongly support a fusion, whereas in reality the reads were produced by different regions altogether. Thus, this method

$$FPKM_{\text{fusion}} \approx FPKM_{\text{fusion}}^{\text{const.estimate}} = FPKM_1 + FPKM_2 \propto \frac{\mu_1'}{l_1} + \frac{\mu_2'}{l_2} = \frac{2}{10} + \frac{2}{20} = 0.3$$

$$FPKM_{\text{fusion}} = \frac{\mu_1' + \mu_2' + \mu_{\text{spanning/split}}}{l_1' + l_2'} \propto \frac{2+2+1}{6+2} = 0.625 \ > 0.3$$

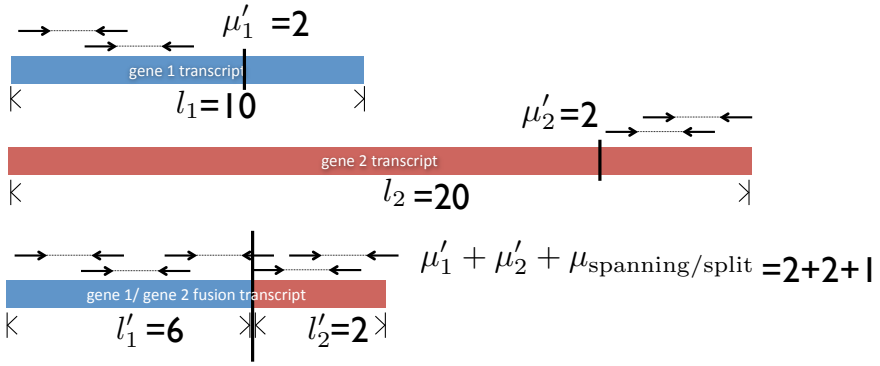


Figure 3.1: Illustration of inaccuracy of constituent summing method ($\text{FPKM}_{\text{fusion}}^{\text{const.estimate}}$) for estimating expression of fusion transcript ($\text{FPKM}_{\text{fusion}}$), resulting in underestimation by a factor of 2. μ_i := number of reads aligning to transcript i; l_i := length of transcript i. $\mu_{\text{spanning/split}}$:=: sum of number of PRs and LRs.

can be inaccurate. One could imagine further difficult-to-set parameters describing heuristics for filtering these reads[37], but this is a complicated and error-prone process in and of itself.

Problems with all existing methods of fusion expression quantification

With all of the above method, a fusions' expression is not directly comparable to the expression of other genes or transcripts. In particular, fusions' expression is not comparable to wild-type constituents' expression, which would be helpful in order to support functional activity of a fusion for prioritization and also for the existence of the fusion. Furthermore, none of the above methods are isoform-specific; all of these merely assess the expression of the fusion gene as a whole, even though isoform-specific expression has found to be impactful in cancer[11][22][56].

3.3 Method

Thus, we propose an improved method for isoform-specific FG discovery and quantification from RNA-seq reads.

In short, the method involves three steps, all of which involve k-mer-based pseudoalignment methods: (1) discovering and aggregating LRs and PRs to predict FJs, (2) construction of candidate fusion transcript isoforms (CFTIs) from FJs, and (3) realignment of reads in order to quantify CFTIs along with all transcripts in the transcriptome.

Predicting Fusion Junctions

The first step of the method is to predict fusion junctions based on LRs and PRs. This is achieved using pseudoalignment methods, which in turn involve the splitting of reads into k-mers, the creation of a De Bruijn-graph transcriptome index for alignment, and aggregation of hashed alignment counts per equivalence class, each of which represent a subset of all possible transcripts.

First, reads with ends aligning to discordant equivalent classes are gathered as candidate PRs. Then, these are aggregated along with non-aligning reads (which include LRs) using a De Bruijn-graph alignment format to assemble candidate fusion junction regions. Finally, these regions themselves are pseudoaligned to the transcriptome in order to find candidate fusion junctions.

Note that this step can be replaced by the use of ordinary RNA-seq-read-based fusion finding tools, such as DeFuse[37] or Tophat-Fusion[24]; results below show data generated by this method.

Construction of Candidate Fusion Transcript Isoforms

Given specific FJs in terms of genomic location, CFTIs are constructed based on enumeration of all possible unique isoforms downstream and upstream of this genomic region. These isoforms can then be added to the alignment index. Note that this can be a De Bruijn-graph-based transcriptomic index[7] for pseudoalignment, or a Bowtie Burrows-Wheeler-based index [28].

Estimating Expression of Fusion Transcripts

Using the CTFI-appended transcriptome index, the classic generative-model based expression estimation approach is used to estimate the expression of all transcripts, including CFTIs.

Briefly, this model[61][49][47] operates via the estimation of parameters representing quantities of interest (transcript interest) and accounts for biases using other parameters, other parameters including per-transcript positional bias and read-composition-based sequence bias.

Importantly, this probabilistic model makes assumptions that assist with the recreation of transcript-wide, isoform specific expression estimation. Crucially, modulo the above biases, the uniform coverage assumption is used in order to deconvolute the extent to which non-junction spanning reads contribute to the expression of the fusion transcript.

The model is a likelihood-based approach which uses hidden data and the expectation-maximization[57] algorithm to estimate parameters.

The result is an estimation of the expression of all transcripts in the CTFI-appended transcriptome, including, importantly, the CTFIs and their constituents.

Checking for Impact of Fusion Transcript

There are three distinct tests that can be performed for the related states of existence, expressional deregulation, and impact of a FJ and its candidate CTFIs:

Overall increase in likelihood of model for including fusion transcript

Since the probabilistic expression model gives a likelihood for a given set of transcripts and a given transcriptome, the likelihood for a fixed transcript set can be seen as a function of a transcriptome. The interpretation of this function is that it will take on a higher value of the transcriptome is more compatible with the sequencing reads; this can be seen as evidence that the transcriptome is the generating transcriptome.

Thus, the likelihood of the model can be compared between a transcriptome with and without CTFIs; this is direct evidence for the existence of a particular FJ and the related CTFIs.

Decrease in bootstrap variance of constituents when including fusions

Using the modern Kallisto[7] expression estimation model, an empirical variance that represents the uncertainty of a particular transcript's expression is estimated. If these quantities decrease for the constituents upon addition of CTFIs, this can be also taken as evidence for FJ and CTFI existence.

Test of nonzero expression for fusion transcripts

Given the expression of all transcripts including CTFIs, one can validate the existence of a particular fusion transcript in a sample based on comparing the expression of the CTFI to the expression of its constituents. In particular, if the CTFI has high expression compared to each constituent, this can be seen as evidence for existence of the fusion transcript. However, if no CTFIs of a given FJ are expressed more than any of their constituent transcripts, this can be seen as evidence against the existence of that fusion transcript.

Regardless, this method allows for the identification of expressionally-deregulated fusion transcripts, in the sense that it allows for direct identification of expressional differences. This can be achieved via comparison of the fusion transcript's expression to the expression of

constituents in a panel of matched normal cells known to be lacking the relevant FJ in their DNA. One can use formal differneital expression tools to gather significance for such a test, such as Cuffdiff[60] or DESeq[2].

3.4 Results

In order to prototype and validate the method, we worked with members of the TCGA LGG AWG. In particular, collaborators at University of California, Santa Cruz and the MD Anderson Cancer Center both ran FJ-discovery software and prioritized samples and fusions for us to assess with our method.

Overview of Group Findings

The full analyses of LGG by the AWG were published in 2015 [13]. The major findings, to which the above method contributed, were that there are three molecular categories of LGG that are effective at predicting prognosis. There was great concordance between sequencing datatypes in this regard. Thehe mutation status, obtained through somatic mutations, of the isocitrate dehydrogenase 1 (IDH1) gene plus the presence of chromosome arms 1p and 19q were sufficient to divide patients into three categories: IDH wildtype, IDH mutant with 1p/19q intact, and IDH mutant with 1p/19 deletions.

This was in concordance with clusters obtained via several unsupervised analyses based on DNA methylation, mRNA, copy number and reverse-phase protein assay. Thus, variance between patients ain the genome, epigenome, transcriptome, and proteome showed high concordance.

Validation of Method

Several fusions were detected using DeFuse[37] and PRADA by collaborators in the group. These fusions were detected within IDH mutant samples and were investigated as possibly explaining within-group variation in prognosis based on known attributes of mechanism. One major category of fusions detected were those involving receptor tyrosine kinases, which are which are involved in cell signalling. As part of the mitogenic circuitry, activity of RTKs leads to mitotic cell division. RTKs are often activated via proximity; thus, overexpression is sufficient to induce activation, as it leads to a higher chance of proximity.

Case 1: FGFR3-TACC3 Fusion

First, we studied a patient with a fusion detected between the Fibroblast Growth Factor Receptor 3 and the Transforming, Acidic, Coiled-coil-containing Protein 3 genes (an FGFR3-TACC3 fusion).

This fusion was a good candidate for validation of our method as it has been identified as undergoing fusion-modulated expressional deregulation [43].

The mechanistic model is as follows: the regulation of FGFR3 is based on the annealing of its 3' untranslated region (UTR) to a micro-RNA (miRNA) (miR-99a, in particular). When miR-99a anneals to the FGFR3 transcript, the transcript is marked for degradation. Thus, FGFR3's expression is downregulated based on its 3' UTR region. However, the fusion is formed such that the FGFR3 constituent lacks its 3' UTR, and therefore loses its wild-type regulatory mechanism. This leads to the temporally and spatially ectopically elevated expression of the FGFR3 sequence, leading to mitosis via mechanisms explained above.

figure from PDF at /Users/ijose Sequencing/

Upon applying the method using eXpress as our expresion quantification module[48] on this specific paper, we were able to validate the results.

Firstly, we noted that isoforms of FGFR3-TACC3 fusion transcripts were among the highest-expressed transcripts in the transcriptome . This was a sanity check offering consistent evidence with the ectopically high expression of FGFR3-TACC3.

figure from page 38 of pdf

Secondly, we noted that we were able to differentiate between the expression of fusion transcript isoforms. In particular, we found that one isoform's expression dominated the expression of all others .

Thirdly, we noted the correct assessment of the FGFR3-TACC3 fusion transcript isoform versus an FGFR3-AC016733.1 transcript, also formed based on the same FJ.

figure from 39 of PDF

Finally, we validated that the expression of the fusion transcript was higher in expression than its constituents in the same cell , finding that the fusion transcript's expression was more than 3 times the expression of FGFR3 and 100 times the expresion of the TACC3 transcript.

figure from 40 of PDF

We also compared the fusion transcripts' expression to its constituents in matched normal tissue. This allowed for further evidence of the fusions' leading to expressional deregulation.

In order to make the comparison between the sample of interest and cerebral cortex samples from the Genotype-Tissue Expression project (GTEx)[34], we transformed expression estimates from fragments per kilobase per million-reads-mapped (FPKM) to transcripts per million (TPM), then applied middle-50th-percentile-normalization[67]. We validated this normalization by using a housekeeping gene, heat-shock protein, 90-kilodaltons, alpha, class b, member 1 (HSP90AB1), for which strong evidence exists of relatively stable expresion temporally, spatially, and inter-cell-type. We saw that the expression in the sample of interest of HSP90AB1 was similar to that of the normal tissue.

We validated that the expresson of the highest-expression FGFR3-TACC3 fusion transcript isoform was much higher than that of its constituents the matched normals – on the order of 10 to 1000 times higher.

EGFR-SEPT14 Fusion

We also studied a fusion between the epidermal growth factor receptor (EFGR) and septin 14 (SEPT14) genes (EGFR-SEPT14 fusion). EGFR is another RTK with similar activity to FGFR3, and was thus another candidate for a fusion having similar oncogenic properties to the FGFR3-TACC3 fusion, although unknown in its ability to expressionally deregulate. EGFR3-

SEPT14 is also known to have oncogenic properties in glioblastoma, which is equivalent to grade IV glioma and related as a progression result to LGGs[42].

Once again, we noticed that the expression of the fusion transcripts were above their constituents in many cases . This suggests expressional deregulation occurred for this fusion.

insert figure from supplement of NEJM

3.5 Discussion

Outlook

The validation of a known expressionally-deregulated fusion and the identification of a candidate expressionally-deregulated fusion is promising that the method will be useful at identifying expressionally-deregulated fusions in the future based on the metric of assessing the expression of the fusions to be significantly higher than their constituents.

Future Work

Much future work could be done on this method; firstly, one could create a formal statistical test of the expression of a fusion transcript being higher than its constituents (a) within sample and (b)a matched panel of normals. This would expedite the identification of expressionally deregulated fusions and would complete the requirement for not requiring manual user assessment of results.

Secondly, one could expand on this method to both identify FJs and lead to their expression. Currently, this is being done by Shannon Hateley and Dr. Páll Melsted, Ph.D. by adapting the Kallisto pseudoalignment algorithm[6] to identify FJs based on discordantly-mapped k-mers.

Thirdly, one could also use modern k-mer-mapping-based expression methods to achieve the expression quantification portion, which are much faster and would be useful as well for providing variance-based evidence for fusion existence.

Fourthly, one could integrate this into one tool, such that pipelining, which is a tedious and error-prone process, would not have to be implemented by the user.

Chapter 4

Integrating genomics data for prediction, with application to gliomas

4.1 Background

One specific prediction problem in gliomas, specifically LGGs, is related to response to chemotherapy. Currently, there is no suggested guideline, according to the World Health Organization and no standard practice between care centers as to whether or not to give chemotherapy following initial resection of the primary tumor in LGG patients.

In particular, temozolomide (TMZ) is given to about 50% of patients being given TMZ following primary therapy. Confounding this decision is the recent finding by Johnson et al. that TMZ appears to be inducing mutations in certain patients which may increase the severity of the recurrent tumor in terms of grade16.

However, it is possible that high-throughput (HT) genomic data might be able to assist in this treatment decision-making problem via predictions. Empirically, several survival time-related HT features have been identified by TCGA; since some of these patients have been treated with TMZ, it's possible that said molecular features might inform recurrence grade decisions for these patients as well, since recurrence grade is associated statistically with survival time. Molecularly (theoretically), the mechanism behind TMZ-induced greater recurrence is partially known.

Candidate molecular mechanism

In particular, TMZ induces cytotoxicity by inducing nuclear genomic mutations, which then cause cells to induce apoptosis. In particular, TMZ adds a methyl group to (methylates) guanine bases in the genome, creating an adduct. The adducts lead to recognition during replication by MMR genes, which recognizes the resulting mismatches but then repeatedly

attempts to repair the region ineffectively by reinsertion of the incorrect base, leading to genomic double-strand breaks and apoptosis.

This apoptosis of quickly replicating tumor cells is the desired effect of TMZ. However, there are several reasons why this might not occur, due to other potentially extant yet testable genomic changes.

Firstly, if apoptotic cell cycle checkpoint machinery is not functioning normally, this leads to consequently genomic instability and large-scale structural rearrangements. Secondly, TMZ-created adducts are removed by MGMT if available and functional. However, if MGMT has a promoter region that is itself methylated, this leads to decreased MGMT expression and persistence of the TMZ-added methylation adducts beyond that which is normal, leading to increasing numbers of mutations possibly beyond even wild-type MMR machinery's ability to detect and control. These two potential issues with TMZ-related apoptosis are thought to be related to the hypermutation phenotype (quantified by much higher mutations per base rates than other tumors), which is correlated with the recurrence as high-grade. Evidence for this is based on TMZ-related mutational patterns having been observed in tumors with hypermutation phenotypes, and the existence of these patterns in regions relating to the above machinery.

4.2 Problem formulation

Underlying this prediction problem in abstraction is a prediction problem: in particular, given high-dimensional clinical and molecular data, can one predict, for a particular patient, whether TMZ-induced hypermutation will occur?

Finding associations can be generalized as a prediction of one random variable y from another, $\vec{x}$, by a function f. The goal is to estimate f. f's performance can be assessed by expected predicted error from true f, which is a function of the complexity of f. f's complexity raises with the dimensionality of its input $\vec{x}$, which leads to an estimate of y that has high variance, which confounds interpretation of any one association as being significant.

In order to address this issue, smoothing parameters can be added to f in order to decrease the variance of f; these are tantamount to making specific assumptions about the underlying probabilistic model that generated the data, which can, in turn, bias the estimate of the association. However, this may still lead to an overall reduction in the expected predicted error, which is a function of both bias and variance of f.

One of the model assumptions that we make is linearity (specifically, that f uses first-order linear terms to manipulate $\vec{x}$ for performing predictions of y, which is a common assumption. This is implemented by linear dimensionality reduction techniques – specifically, the factor association analysis (FAA) method.

Heterogeneity challenge

A fundamental challenge with this prediction problem is the heterogeneity of data types involved the prediction problem. In particular, there is molecular data which may include germline variants of various forms detected by various different exome sequencing pipelines, gnomic copy number data based on arrays and/or DNA-sequencing, expression detected by a RNA-seq and/or microarray platforms, methylation detected by microarrays and/or sequencing-based techniques, and uni-dimensional clinical covariates.

In particular, this will manifest as different underlying distributions that best approximate of each data type, both between patients within a specific modality (due to heterogeneity of platforms) and between modalities within a specific patient.

Unsupervised solutions

Unsupervised machine learning approaches, which seek to more clearly represent variances in a dataset without any explicit inclusion of specific outcomes of interest, have been used to assess multi-platform genomic data.

Consensus clustering *Consensus clustering* or *cluster-of-cluster* [26] methods combine datatypes by hierarchically clustering samples within each datatype, then using the specific cluster assigned per data-type as a feature for a meta-clustering approach. This is best used successfully in situations where variance is shared among multiple different genomic datatypes, *and* that shared variance is mostly or largely a component of the outcome of interest.

factor analysis-based approaches Approaches such as iCluster [52][51] use an explicit probabilistic generative model in order to reduce dimensionality and simultaneously find shared variance among a different number of sequencing datatypes, allowing for parametric assumptions useful for sequencing data type distributions (for example, modelling mutation calls as Bernoulli random variables). This approach has not been heretofore expanded into a supervised setting.

PARADIGM PARADIGM models explicit interactions between sequencing datatypes using existing models of gene interaction networks as well as assumptions about functionality [63] [40]. In particular, the activity of every gene is predicted based on the status of its corresponding copy number, mutation status, and gene expression, assuming, for example, that a deleted gene should not be functional, or a gene that is mutated but is not expressed should not be directly functional. PARADIGM effectively reduces high-dimensional datasets to inferences about the activity level on a per-pathway basis for further manual investigation. PARADIGM suffers from a wealth of parameters, and therefore necessitates a relatively large amount of data to prevent overfitting.

One limitation of all unsupervised approaches is the necessity of the outcome of interest to correspond to the shared variance among the sequencing datatypes, which may not be the case. This is addressed by supervised methods.

Supervised solutions

Many supervised approaches merely add features from multiple platforms without specific inclusion of special related assumptions into the model.

This is appropriate for some models, for example, random survival trees, which are robust to different feature distribution assumptions.

This is less appropriate for linear models. Thus, a few linear supervised machine learning models have been created in order to address the use of heterogeneous sequencing data types as features in order to predict an outcome of interest. This is an area of intense current research.

regression on residuals Yuan, Allen, and Omberg et al. [71] looked at the improvement from adding single molecular datatypes to a clinical variable model of predicting overall survival time in four cancer types with data from TCGA, using a cox multivariate proportional hazards model. In particular, after regressing out the predictive clinical information from the feature set on the outcome, individual genomic features were selected by choosing those that explained the residual variance in the survival time outcome, and added as regular features to the model. However, this was not found to be more accurate than merely using the additional features in random forest models without any special annotation of sequencing data types.

canonical correlation analysis-based approaches Canonical correlation analysis, which is originally an unsupervised technique to find shared variance between two feature sets, has been adapted into supervised models by adding an outcome of interest [53] [3][70][68]. This was used with some success by Drs. Robert Tibshirani and Samuel Gross as a predictor for diffuse B-cell lymphoma prediction of copy number. Recently, an extension was proposed: the *Collaborative Regression* method[21]. This method addressed shortcomings of previous methods by formulating the problem in a way that is convex, and therefore algorithmically efficient.

However, these methods were not specified probabilistically, and therefore are not as amenable to extensions involving explicit data type distribution specifications. In addition, the adaptation of CCA used above was approximate, and thereby is possibly less statistically powerful than a precise (probabilistic) specification. Furthermore, not being probabilistic, these models can not be used to generatively, preventing them using simulation to assess model performance parameters, such as requisite N, p, and effect size.

Dimensionality challenge

A perhaps bigger fundamental challenge with this prediction problem is what's referred to as "the curse of high-dimensionality," that is to say, there are many more covariates/predictors (number of predictors $\coloneqq p$) than observations (N) (i.e., $p \gg N$). This is inherent to this problem due to the wildly high-dimensionality of the molecular datasets, which at their current maximum for even a single modality is about 500 000 with reasonable granularity, but could theoretically be on the order of 10^9. This is manifests as an issue by leading to overfitting or equivalently, high variance in estimators of associations.

Fortunately, these problems are not unique to this particular prediction problem, and consequently, several techniques have been developed to deal with both of these fundamental challenges.

Subset selection approaches

Subset selection involves using only a subset of available features, which must be selected. There are several methods for doing so in order to address the exponential number of possible subsets, including forward and backwards stepwise regression, which iteratively adds or removes predictors based on relationship to the outcome. However, these methods suffer from relatively high prediction error due to the high variance imposed by the discrete inclusion or exclusion of a specific parameter[18].

Shrinkage approaches

A standard method of addressing high-dimensionality and related high variance of any estimated fit parameters that stymie attempts to generalize is by adding a model term that explicitly adds a penalization on the use of each parameter. An optimization is then performed over this adjusted function. Relatedly, a constraint can be placed on the size of the features and solutions within a particular constraint can be sought. Parameters involving the degree of regularization are selected based on a *model selection* step, which typically attempts to assess the expected prediction error of a model with several different parameter settings in order to identify the most accurate combination.

The specific penalty that is applied to each feature may be *Lasso*, which is uses the $L1$-norm of the vector of coefficients: $\sum_{j=1}^{p} |\beta_j| \leq t$, where β_j is each coefficient on the p parameters, and t is a pre-specified constant. Another common approach is *Ridge*, which uses the $L2$-norm of the coefficient vector: $\sum_{j=1}^{p} \beta_j^2 \leq t$. These have different properties, with Lasso generally leading to more sparsity. However, this may be an incorrect assumption that hurts generalizability; to address this, *Elastic Net* regression was created, which involves a hyper-parameter that is specified to address the relative contribution of ridge and lasso constraints.

Dimensionality-reduction approaches

Dimensionality-reduction approaches offer a natural method of dealing with large numbers of correlated features – modeling the high dimensionality of these features by a small number of independent features, and using these as new prediction features. Approaches such as *principal components regression* and *partial least squares* achieve this.

However, existing dimensionality reduction techniques typically do not also directly address heterogeneity in the feature space.

4.3 Developed method

Inspiration from unsupervised dimensionality reduction models

This method's development came about due to the observation by Novembre and Stephens[41] that when taking the principal component analysis (PCA) of the single nucleotide variants present in individuals, the first few components correlated highly with geographic location of the individual. Briefly, PCA finds a ranked list whose entries consist of linear combinations of variables that explain decreasing components of the variance in a given matrix; in this case, that matrix $\mathbf{V}$ (people $\times$ genomic variants).

Novembre's result can be interpreted as much of the first two statistically independent sources of variance in the common gene variants "explains" their geographic location. This knowledge can then be used to decrease geographically-related variance in the germline variants of patients in GWAS studies, increasing their statistical power by decreasing noise.

Recently, gene expression information has also begun to be collected on a large sale, enabled be recent decreases in technology price. Dr. Lior Pachter, Dr. Nicholas Bray, Brielin Brown, and Dr. McCurdy consequently wondered if gene expression information also contained significant geographic signal, and consequently performed PCA on the gene expression matrix $\mathbf{G}$ (people $\times$ genes). However, the first few components did not correlate directly with geography.

This then led to the application of canonical correlation analysis (CCA) to this dataset in order to directly find shared variance between the two-dimensional geographic origin ($\mathbf{R}$: people $\times$ (latitude, longitude)) of the individuals and their high-dimensional associated gene expression vector (i.e., CCA between $\mathbf{G}$ and $\mathbf{R}$). Briefly, CCA is analogous to PCA, except for linear combinations are taken of both matrices (termed "canonical functions" (CFs)) instead of just one matrix, and they are taken in ways that maximize the correlation between the linear combinations of one matrix with the other. The first two canonical components (CCs) in the above analysis therefore, by construction, should have significant correlation with the geographic coordinates, at least inasmuch as the component on the geographic matrix.

This was indeed found to be the case; however, it was quickly realized when doing permutation testing that the $p \gg N$ issue resulted in overfitting (i.e., no CFs were statistically significant), due to the high dimensionality of $\mathbf{G}$. This was addressed with a specific type

of regularization. In particular, $\mathbf{G}$'s dimensionality was first reduced using PCA to $\mathbf{G}'$, and then CCA was performed between $\mathbf{G}'$ and $\mathbf{R}$, resulting in statistically significant CFs. Visualization of these then revealed the desired correlation with geography.

Formulation based on PCA and CCA

Out of this was borne an inspiration to use probabilistic model to create more nuanced dependency relationships between various datatypes.

PCA[58], which can be interpreted as probabilistic graphical model with a latent variable emitting an observed variable based on a dimensionality expansion plus diagonal noise.

CCA can be viewed as a dimensionality reduction technique similar to PCA the sense that both matrices are reduced in dimensionality such that they correlate most with one another. One can also, relatedly, formulate CCA probabilistically as finding a maximum likelihood solution for a probabilistic general model in which a low-dimensional latent space expanded into higher dimensions with added noise of a specific structure – positive semidefinite noise and with two observed variables. Thus, some combination of these two (PCA and CCA) noise constraints was indicated in order to simulate the PCA/CCA combination above in a probabilistic fashion.

One way that these can be combined in a probabilistic model is through simply adding an intermediate node between the observed data and the latent data, and have the intermediate node be generated from the latent node with diagonal noise, and the observed node be generated from the intermediate node with positive semidefinite noise.

As a generalization to more than two datatypes, one can imagine recursively applying this, such that, for example, if $\mathbf{G}'$ and $\mathbf{R}$ are predicted to have been generated from $\mathbf{H}$, a lower-dimensional space, and then performing FAA on $\mathbf{H}$ and $\mathbf{S}$ to find $\mathbf{T}$, etc. We define this as a hierarchical factor association model (HFAM). This lends FAA to more complicated association problems than merely two datasets with two datatypes.

This successful application of HFAM in geographic explanation of gene expression setting and the recognition of its non-setting-specific theoretical nature as a framework led to the idea of applying it in other datasets, in particular those in cancer genomics in prediction problems. Thus came the inspiration for applying it to the prediction problem related to severity of recurrence upon application of TMZ given several HT genomic datatypes.

Formal model specification

In more concrete terms, the proposed HFAM consists of a latent factor model with $n_z > 1$ latent multidimensional variables $\vec{z}_1, ...\vec{z}_{n_z}$ and $n_x > 1$ observed variables $\vec{x}_1, ..., x_{n_x}$.

$\vec{z}_1$ is taken to be generated by a normal prior wtih identity variance: $\vec{z}_1 \sim N(\mathbf{0}, \mathbf{I})$

Variables are generated from other variables as follows: $\vec{\gamma}$ is a dimensionality expansion of $\vec{\delta}$, plus noise: $\vec{\gamma} \sim \mathbf{W}_\gamma \vec{\delta} + \vec{\epsilon}_\gamma$. $\vec{\epsilon}_\gamma \sim N(\vec{0}, \mathbf{\Psi}_\gamma)$, where $\mathbf{\Psi}_\gamma$ can be specified to be diagonal or unconstrained. We are using the normal distribution $\vec{\gamma} \sim N(\mathbf{W}_\gamma \vec{\delta}, \mathbf{\Psi}_\gamma)$, although other distributions could be used as an extension.

One specific observed variable is set to be the outcome of interest: $\vec{x}_i := \vec{y}$.

In general, observed variables are specified to have diagonal noise in their generation from latent variables, whereas latent variables are specified to have unconstrained noise in their generation from other latent variables (other than $\vec{z}_1$).

As a graphical model, this can be viewed as a multifurcating tree, with root $\vec{z}_1$ and leaves $\vec{x}_1, ..., x_{n_x}$, where an arrow from δ to γ indicates the relationship outlined above. As an example, see an instance of the model with three observed nodes and two hidden nodes ($n_z = 2, n_x = 3$): .

insert 2-data figure

This model can be seen as imposing a specific structure on the covariance matrix of the data which is approrpiate for the supervised modeling of sequencing data towards a specific response. In particular, it imposes a block joint covariance structure. See appendix for an example.

Inference is done using the expectation-maximization algorithm in order to estimate the parameters: $\mathbf{\Theta} := \{\mathbf{W}_i, \mathbf{\Psi}_i\}_{i=1}^{n_x+n_z}$ as well as the latent variables $\vec{z}_1, ...\vec{z}_{n_z}$. See the appendix for updates derived for a three-node model.

After learning parameters $\mathbf{\Theta}$, one can predict y from input observed variables. See the appendix for details.

Consistency of model structure and genomics assumptions

The above model specification if used with one multifurcating second-level latent variable and one high-level latent variable of relatively low dimension induces model assumptions we feel are consistent with reasonable assumptions for the datatype, and simultaneously address the challenges of heterogeneity and high-dimensionality. In particular, dimensionality is addressed through the use of relatively low-dimensional latent variables z_i – on the order of 10^0 in our simulations.

Heterogeneity is handled as follows: when diagonal noise variances are used for observed variables, this forces covariance within-feature set to be reflected in the latent variable and consequently sibling variables; thus, the latent variable can be seen as storing covariance that is both within-feature type and between-feature type. This is reflected in the associated variable $\mathbf{W}$. In addition, the multi-level hierarchical nature with unconstrained variance between the latent nodes leads to a variable amount of shared-sequencing variance to be used to explain the outcome of interest.

This is consistent with intuitive and established notions about empirical shared variance between sequencing types and outcomes: *some* of the shared variance between shared data types will be relevant for the outcome, but not all. It can also be seen as using each additional sequencing datatype as further evidence for a particular variance structure. It differs from a single latent-node model, which will expect all covariance within feature-datatype to be explained in the outcome variable. This can be seen using Bayes' ball approach for studying dependencies in probabilistic graphical models .

insert figure

Implementation

The above model was implemented with a set of `python` classes and associated script. It allows for the specification of a general-structure model and its parameters (number of nodes, their associated dimensions, and the diagonal or non-diagonal variance structure of noise). See the `Bitbucket`®repository (`https://bitbucket.org/ijoseph/factorassociationanalysiscancergenomics/`) for details .

make public

Briefly, the `GaussianLatentFactorModel` class includes logic for inference and prediction, and `ThreeNodeGaussianModel` shows an example of a three-node implementation with minimal specification.

4.4 Results

We present the results on three different datasets of increasing difficulty: (1) simulated data with arbitrary parameters, and in order to discover necessary parameters for accurate prediction, (2) data from the Cancer Genome Project[30] with around $1,000$ samples and $100,000$ features, and the Costello Lab (see chapter 2 for details of all).

Assessment

Machine learning assessment is a rich field, with many options available.

Expected prediction error motivation

Essentially, one important metric to measure is expected prediction error, which describes the expected accuracy of a given fit model on predicting from data arising from the same distribution under which it was trained [18].

Under situations with unlimited data size and resources, one would estimate this by using a training/ validation/ test split of the available data, where 50% of the data is used for training, 25% is used for validation, and another 25% is used for testing.

The *training* set would be used to fit the model. The *validation* set would be used to estimate the expected prediction error for a given setting of model tuning parameters (related to model complexity and therefore tendency to over or underfit). Different tuning parameters would be assessed until one with acceptable expected prediction error were found on the validation set. The *testing* set would be used to confirm the expected prediction error.

However, under a relatively limited number of samples, one might perform *cross-validation* in order to assess the prediction error from a limited number of samples. This eschews the need for a validation set, and estimates one based on iteratively holding various small subsets of the data and assessing prediction error on the small subsets.

For expediency and data-size reasons, we chose to use cross-validation in this manner in order to estimate prediction error, and chose to preemptively fix tuning parameters.

In particular, for our model, we chose the dimension of z_1 to be 2 and z_2 to be 3. For logistic regression, we set left the Lasso regularization strength at $C = 1$, its default.

Practically, we assessed the below performance by interfacing with `scikit-learn`[44], a package with built-in assessment tools, and also used existing machine learning models in this package for comparison.

AUC metric

For a given set of predictions on data, one can assess accuracy based on the *area under the [receiver-operating characteristic (ROC)] curve (AUC)*, a common measure of classifier efficacy[18]. Briefly, the ROC curve's axes specify the particular specificity achievable for a given setting of sensitivity. An AUC of 1 indicates perfect classification, whereas 0.5 indicates a classifier that is no better than chance.

Confusion matrix

In order to complement the ROC curve and AUC, we also studied *confusion matrices* based on cross-validation predictions, which show, for a given fixed sensitivity/specificity, the number of true positives, false positives, false negatives, and true negatives.

Simulated data

We sought to validate that the model was performing as expected, and also to assess the necessary effect size for a given setting of N and p. Effect size was approximated through the setting of a specific parameter of element-wise average $\mathbf{e} \coloneqq \frac{\mathbf{W}}{\mathbf{\Psi}}$ for parameter generation, where a larger setting of this quantity indicated higher signal-to-noise ratio and effect size. In essence, this means that the latent variables of the model explain more about the association between the datatypes than the noise specific to each datatype.

Since we are interested in a classification problem and the model as specified generates floating-point numbers as observed data, we made the additional step of binarizing one of the one-dimensional output nodes based on the empirical mean of the simulated distribution and assessed recovery.

$\mathbf{p \ll N}$

First, we sought to assess whether the model would lead to more accurate recovery of its simulated data than other models. We chose logistic regression with lasso penalty as a model for comparison.

$\mathbf{N = 1,000, p = 11}$ We used $N = 1,000$ and $p = 11$ ($n_x = 3$, with two 5-dimensional feature nodes). $N = 1,000$ corresponds to the approximate number of samples in the Cancer Genome Project[30] dataset.

We found that the model was slightly superior from an AUC perspective, especially with relatively low effect size ($e \leftarrow 1.25$).

insert figure showing this

$\mathbf{N = 30, p = 11}$ This number of samples roughly corresponds to that which is available from the Costello Lab. Here, we see a greater advantage of the model in terms of AUC, especially at relatively low effect-sizes.

figure

$\mathbf{p \gg N}$

Secondly, we sought to assess the model's performance for data with dimensions that approximated the dimensions of real genomic data.

First, we assessed its performance on $p = 50\,000$, which is similar to two datatypes' dimensions from the Cancer Genome Project.

Here, we found the model slightly superior than logistic regression, and requisite e to be 1.25 for what we consider adequate prediction accuracy (AUC 0.85).

insert figures

Secondly, we attempted to assess the model performance on $N = 30$, $p = 500\,000$, which is similar to Costello's data size, but found performance prohibitive without further algorithmic optimization. In lieu of this, we opted to preemptively reduce dimensionality for the Costello dataset (see below).

Cancer Genome Project

First, we assessed the efficacy of the model using both expression and mutation information.

Expression information was used in the form of $z - scores$, derived as described earlier. Mutation information was used on a genic level, where any mutation occurring anywhere within a gene body resulted in a binary 1 for that gene in that sample.

As this was the gene level, there were roughly 26 000 features per mutation type.

Missing information for a gene was imputed based on average for that gene across all samples for which information for that gene was present.

Logistic regression performed poorly on this dataset, achieving an AUC of 0.5, which was indistinguishable from chance.

HFAM performed

so how'd it do after I fixed the issues.

Glioma prediction

Here, we used sequencing data collected post-second surgery to see if we could assess hypermutation as having already occurred (versus prediction of future hypermutation), as preliminary attempts to predict outcome from first-surgery tissue were unpromising.

We chose to use methylation and mutation information, as these datatypes were available for the highest number of samples.

We chose the 50 000 most variable methylation probes, pursuant to suggestions from Dr. Matthew Grimmer as to the most cogent way to reduce methylation dimensionality. In

particular, this was deemed superior to binning based on specific regions, as related glioma methylation changes often occur on a single-point basis, rather than on a large scale basis.

.

so how did this turn out

4.5 Discussion

This approach shows promise; we feel that further refinement of the approach could yield to more accurate prediction than what is shown above. In particular, we would like to implement a model tuning phase above in order to assess dimensions; we would like to use the model to integrate more than two feature types – a relatively simple addition on what is currently implemented. We would like to use different distributions for our observed datatypes that work more cogently with mutation information – chiefly, the multinomial distribution.

We would like to assess the model versus existing non-probabilistic methods for heterogeneous, regularized supervised prediction as well, such as collaborative regression.

8

Bibliography

[1] SF Altschul, Warren Gish, and W Miller. "Basic Local Alignment Search Tool". In: *Journal of molecular ...* (1990), pp. 403–410. URL: http://www.sciencedirect.com/science/article/pii/S0022283605803602.

[2] Simon Anders and Wolfgang Huber. "Differential Expression of RNA-Seq Data at the Gene Level–the DESeq Package". In: *Heidelberg, Germany: European Molecular Biology Laboratory (EMBL)* (2012). URL: http://www.genomatix.de/online_help/help_regionminer/DESeq_1.10.1.pdf (visited on 11/11/2016).

[3] Eric Bair et al. "Prediction by Supervised Principal Components". In: *Journal of the American Statistical Association* 101.473 (Mar. 1, 2006), pp. 119–137. ISSN: 0162-1459. DOI: 10.1198/016214505000000628. URL: http://amstat.tandfonline.com/doi/abs/10.1198/016214505000000628 (visited on 01/16/2015).

[4] Jordi Barretina et al. "The Cancer Cell Line Encyclopedia Enables Predictive Modelling of Anticancer Drug Sensitivity". In: *Nature* 483.7391 (Mar. 29, 2012), pp. 603–607. ISSN: 0028-0836. DOI: 10.1038/nature11003. URL: http://www.nature.com/nature/journal/v483/n7391/full/nature11003.html%3FWT.ec_id%3DNATURE-20120329 (visited on 02/19/2016).

[5] Christopher M. Bishop. "Pattern Recognition". In: *Machine Learning* 128 (2006). URL: http://www.academia.edu/download/30428242/bg0137.pdf (visited on 11/22/2016).

[6] Nicolas L. Bray et al. "Near-Optimal Probabilistic RNA-Seq Quantification". In: *Nature Biotechnology* 34.5 (May 2016), pp. 525–527. ISSN: 1087-0156. DOI: 10.1038/nbt.3519. URL: http://www.nature.com/nbt/journal/v34/n5/abs/nbt.3519.html (visited on 09/23/2016).

[7] Nicolas Bray et al. "Near-Optimal RNA-Seq Quantification". In: (May 11, 2015). arXiv: 1505.02710 [cs, q-bio]. URL: http://arxiv.org/abs/1505.02710 (visited on 05/23/2015).

[8] *CASAVA Support*. 2016. URL: http://support.illumina.com/sequencing/sequencing_software/casava.html (visited on 11/22/2016).

[9] Ken Chen et al. "BreakDancer : An Algorithm for High-Resolution Mapping of Genomic Structural Variation". In: 6.9 (2009). DOI: 10.1038/NMETH.1363.

[10] Lynda Chin, Jannik N. Andersen, and P. Andrew Futreal. "Cancer Genomics: From Discovery Science to Personalized Medicine". In: *Nature Medicine* 17.3 (Mar. 2011), pp. 297–303. ISSN: 1078-8956. DOI: 10.1038/nm.2323. URL: http://www.nature.com/nm/journal/v17/n3/abs/nm.2323.html (visited on 06/03/2014).

[11] Heather R. Christofk et al. "The M2 Splice Isoform of Pyruvate Kinase Is Important for Cancer Metabolism and Tumour Growth". In: *Nature* 452.7184 (Mar. 13, 2008), pp. 230–233. ISSN: 0028-0836. DOI: 10.1038/nature06734. URL: http://www.nature.com/nature/journal/v452/n7184/abs/nature06734.html (visited on 11/22/2013).

[12] Kristian Cibulskis et al. "Sensitive Detection of Somatic Point Mutations in Impure and Heterogeneous Cancer Samples". In: *Nature biotechnology* 31.3 (Mar. 2013). ISSN: 1087-0156. DOI: 10.1038/nbt.2514. pmid: 23396013. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3833702/ (visited on 11/06/2014).

[13] "Comprehensive, Integrative Genomic Analysis of Diffuse Lower-Grade Gliomas". In: *New England Journal of Medicine* (June 10, 2015), null. ISSN: 0028-4793. DOI: 10.1056/NEJMoa1402121. pmid: 26061751. URL: http://dx.doi.org/10.1056/NEJMoa1402121 (visited on 06/12/2015).

[14] Sean Davis et al. "Package 'methylumi'". In: (2013). URL: http://master.bioconductor.org/packages/release/bioc/manuals/methylumi/man/methylumi.pdf (visited on 03/03/2015).

[15] Andrew P. Feinberg, Rolf Ohlsson, and Steven Henikoff. "The Epigenetic Progenitor Origin of Human Cancer". In: *Nature Reviews Genetics* 7.1 (Jan. 2006), pp. 21–33. ISSN: 1471-0056. DOI: 10.1038/nrg1748. URL: http://www.nature.com/nrg/journal/v7/n1/abs/nrg1748.html (visited on 06/24/2014).

[16] Maria E. Figueroa et al. "DNA Methylation Signatures Identify Biologically Distinct Subtypes in Acute Myeloid Leukemia". In: *Cancer Cell* 17.1 (Jan. 19, 2010), pp. 13–27. ISSN: 1535-6108. DOI: 10.1016/j.ccr.2009.11.020. URL: http://www.sciencedirect.com/science/article/pii/S1535610809004206 (visited on 06/26/2014).

[17] *Firehose — Www.broadinstitute.org/Cancer/CGA*. 2016. URL: http://archive.broadinstitute.org/cancer/cga/Firehose (visited on 11/22/2016).

[18] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The Elements of Statistical Learning*. Vol. 1. Springer series in statistics Springer, Berlin, 2001. URL: http://statweb.stanford.edu/~tibs/book/preface.ps (visited on 11/18/2016).

[19] Mathew J. Garnett et al. "Systematic Identification of Genomic Markers of Drug Sensitivity in Cancer Cells". In: *Nature* 483.7391 (Mar. 29, 2012), pp. 570–575. ISSN: 0028-0836. DOI: 10.1038/nature11005. URL: http://www.nature.com/nature/journal/v483/n7391/full/nature11005.html (visited on 05/14/2014).

[20] Thomas R. Gingeras. "Implications of Chimaeric Non-Co-Linear Transcripts". In: *Nature* 461.7261 (2009), pp. 206–211. URL: http://www.nature.com/nature/journal/v461/n7261/abs/nature08452.html (visited on 09/23/2016).

[21] Samuel M. Gross and Robert Tibshirani. "Collaborative Regression". In: *Biostatistics* 16.2 (Jan. 4, 2015), pp. 326–338. ISSN: 1465-4644, 1468-4357. DOI: 10.1093/biostatistics/kxu047. pmid: 25406332. URL: http://biostatistics.oxfordjournals.org/326 (visited on 03/06/2015).

[22] Karine Hovanes et al. "-Catenin–sensitive Isoforms of Lymphoid Enhancer Factor-1 Are Selectively Expressed in Colon Cancer". In: *Nature Genetics* 28.1 (May 2001), pp. 53–57. ISSN: 1061-4036. DOI: 10.1038/ng0501-53. URL: http://www.nature.com/ng/journal/v28/n1/abs/ng0501_53.html (visited on 11/22/2013).

[23] Brett E. Johnson et al. "Mutational Analysis Reveals the Origin and Therapy-Driven Evolution of Recurrent Glioma". In: *Science* 343.6167 (Oct. 1, 2014), pp. 189–193. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.1239947. pmid: 24336570. URL: http://www.sciencemag.org/content/343/6167/189 (visited on 06/09/2014).

[24] Daehwan Kim and Steven L. Salzberg. "TopHat-Fusion: An Algorithm for Discovery of Novel Fusion Transcripts". In: *Genome Biology* 12.8 (Aug. 11, 2011), R72. ISSN: 1465-6906. DOI: 10.1186/gb-2011-12-8-r72. pmid: 21835007. URL: http://genomebiology.com/2011/12/8/R72/abstract (visited on 12/09/2013).

[25] Tony Kouzarides. "Chromatin Modifications and Their Function". In: *Cell* 128.4 (Feb. 23, 2007), pp. 693–705. ISSN: 0092-8674. DOI: 10.1016/j.cell.2007.02.005. URL: http://www.sciencedirect.com/science/article/pii/S0092867407001845 (visited on 11/14/2013).

[26] Vessela N. Kristensen et al. "Principles and Methods of Integrative Genomic Analyses in Cancer". In: *Nature Reviews Cancer* 14.5 (May 2014), pp. 299–313. ISSN: 1474-175X. DOI: 10.1038/nrc3721. URL: http://www.nature.com/nrc/journal/v14/n5/full/nrc3721.html (visited on 11/17/2016).

[27] Shailesh Kumar et al. "Identifying Fusion Transcripts Using next Generation Sequencing". In: *Wiley Interdisciplinary Reviews: RNA* (Aug. 1, 2016), n/a–n/a. ISSN: 1757-7012. DOI: 10.1002/wrna.1382. URL: http://onlinelibrary.wiley.com/doi/10.1002/wrna.1382/abstract (visited on 09/23/2016).

[28] Ben Langmead and Steven L Salzberg. "Fast Gapped-Read Alignment with Bowtie 2." In: *Nature methods* 9.4 (Apr. 2012), pp. 357–9. DOI: 10.1038/nmeth.1923. PMID: 22388286.

[29] Michael S. Lawrence et al. "Mutational Heterogeneity in Cancer and the Search for New Cancer-Associated Genes". In: *Nature* 499.7457 (July 11, 2013), pp. 214–218. ISSN: 0028-0836. DOI: 10.1038/nature12213. URL: http://www.nature.com/nature/journal/v499/n7457/full/nature12213.html (visited on 01/17/2014).

[30] Heidi Ledford. "End of Cancer-Genome Project Prompts Rethink". In: *Nature* 517.7533 (Jan. 5, 2015), pp. 128–129. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/517128a. URL: http://www.nature.com/doifinder/10.1038/517128a (visited on 03/10/2016).

[31] Mark D. M. Leiserson et al. "Pan-Cancer Network Analysis Identifies Combinations of Rare Somatic Mutations across Pathways and Protein Complexes". In: *Nature Genetics* 47.2 (Feb. 2015), pp. 106–114. ISSN: 1061-4036. DOI: 10.1038/ng.3168. URL: http://www.nature.com/ng/journal/v47/n2/abs/ng.3168.html (visited on 09/28/2016).

[32] Bo Li and Colin N. Dewey. "RSEM: Accurate Transcript Quantification from RNA-Seq Data with or without a Reference Genome". In: *BMC Bioinformatics* 12.1 (Aug. 4, 2011), p. 323. ISSN: 1471-2105. DOI: 10.1186/1471-2105-12-323. pmid: 21816040. URL: http://www.biomedcentral.com/1471-2105/12/323/abstract (visited on 08/12/2015).

[33] Heng Li and Richard Durbin. "Fast and Accurate Short Read Alignment with Burrows–Wheeler Transform". In: *Bioinformatics* 25.14 (2009), pp. 1754–1760. URL: http://bioinformatics.oxfordjournals.org/content/25/14/1754.short (visited on 11/22/2016).

[34] John Lonsdale et al. "The Genotype-Tissue Expression (GTEx) Project". In: *Nature Genetics* 45.6 (May 2013), pp. 580–585. DOI: 10.1038/ng.2653. URL: http://www.nature.com/doifinder/10.1038/ng.2653.

[35] Aaron McKenna et al. "The Genome Analysis Toolkit: A MapReduce Framework for Analyzing next-Generation DNA Sequencing Data". In: *Genome Research* 20.9 (Jan. 9, 2010), pp. 1297–1303. ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.107524.110. pmid: 20644199. URL: http://genome.cshlp.org/content/20/9/1297 (visited on 11/22/2016).

[36] Roger McLendon et al. "Comprehensive Genomic Characterization Defines Human Glioblastoma Genes and Core Pathways". In: *Nature* 455.7216 (Oct. 23, 2008), pp. 1061–1068. ISSN: 0028-0836. DOI: 10.1038/nature07385. URL: http://www.nature.com/nature/journal/v455/n7216/abs/nature07385.html (visited on 10/16/2014).

[37] Andrew McPherson et al. "deFuse: An Algorithm for Gene Fusion Discovery in Tumor RNA-Seq Data." In: *PLoS computational biology* 7.5 (May 2011), e1001138–e1001138. DOI: 10.1371/journal.pcbi.1001138. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3098195&tool=pmcentrez&rendertype=abstract.

[38] Peter Moffett and Gregory Moore. "The Standard of Care: Legal History and Definitions: The Bad and Good News". In: *Western Journal of Emergency Medicine* 12.1 (Feb. 2011), pp. 109–112. ISSN: 1936-900X. pmid: 21691483. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3088386/ (visited on 10/15/2016).

[39] Raman P. Nagarajan et al. "Recurrent Epimutations Activate Gene Body Promoters in Primary Glioblastoma". In: *Genome Research* (Apr. 7, 2014). ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.164707.113. pmid: 24709822. URL: http://genome.cshlp.org/content/early/2014/04/07/gr.164707.113 (visited on 06/09/2014).

[40] Sam Ng et al. "PARADIGM-SHIFT Predicts the Function of Mutations in Multiple Cancers Using Pathway Impact Analysis". In: *Bioinformatics* 28.18 (Sept. 15, 2012), pp. i640–i646. ISSN: 1367-4803, 1460-2059. DOI: 10.1093/bioinformatics/bts402. pmid: 22962493. URL: http://bioinformatics.oxfordjournals.org/content/28/18/i640 (visited on 01/27/2014).

[41] John Novembre and Matthew Stephens. "Interpreting Principal Component Analyses of Spatial Population Genetic Variation." In: *Nature genetics* 40.5 (May 2008), pp. 646–9. DOI: 10.1038/ng.139. PMID: 18425127.

[42] *OMIM Entry - * 612140 - SEPTIN 14; SEPT14*. 2016. URL: http://www.omim.org/entry/612140?search=sept14&highlight=sept14 (visited on 11/12/2016).

[43] Brittany C. Parker et al. "The Tumorigenic FGFR3-TACC3 Gene Fusion Escapes miR-99a Regulation in Glioblastoma". In: *The Journal of Clinical Investigation* 123.2 (Feb. 1, 2013), pp. 855–865. ISSN: 0021-9738. DOI: 10.1172/JCI67144. pmid: 23298836. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3561838/ (visited on 10/23/2013).

[44] Fabian Pedregosa et al. "Scikit-Learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (Oct. 2011), 28252830. URL: http://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html (visited on 11/18/2016).

[45] *Picard Tools - By Broad Institute*. 2016. URL: http://broadinstitute.github.io/picard/ (visited on 11/22/2016).

[46] Robert W. Rapkins et al. "The MGMT Promoter SNP rs16906252 Is a Risk Factor for MGMT Methylation in Glioblastoma and Is Predictive of Response to Temozolomide". In: *Neuro-Oncology* (Apr. 24, 2015), nov064. ISSN: 1522-8517, 1523-5866. DOI: 10.1093/neuonc/nov064. pmid: 25910840. URL: http://neuro-oncology.oxfordjournals.org/content/early/2015/04/23/neuonc.nov064 (visited on 06/08/2015).

[47] Adam Roberts, Harvey Feng, and Lior Pachter. "Fragment Assignment in the Cloud with eXpress-D". In: *BMC Bioinformatics* 14.1 (Dec. 7, 2013), p. 358. ISSN: 1471-2105. DOI: 10.1186/1471-2105-14-358. pmid: 24314033. URL: http://www.biomedcentral.com/1471-2105/14/358/abstract (visited on 01/25/2014).

[48] Adam Roberts and Lior Pachter. "Streaming Fragment Assignment for Real-Time Analysis of Sequencing Experiments". In: *Nature Methods* 10.1 (Jan. 2013), pp. 71–73. ISSN: 1548-7091. DOI: 10.1038/nmeth.2251. URL: http://www.nature.com/nmeth/journal/v10/n1/full/nmeth.2251.html (visited on 10/08/2013).

[49] Adam Roberts et al. "Improving RNA-Seq Expression Estimates by Correcting for Fragment Bias." In: *Genome biology* 12.3 (Jan. 2011), R22–R22. DOI: 10.1186/gb-2011-12-3-r22. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3129672&tool=pmcentrez&rendertype=abstract.

[50] J. Zachary Sanborn et al. "Double Minute Chromosomes in Glioblastoma Multiforme Are Revealed by Precise Reconstruction of Oncogenic Amplicons". In: *Cancer Research* 73.19 (Jan. 10, 2013), pp. 6036–6045. ISSN: 0008-5472, 1538-7445. DOI: 10.1158/0008-5472.CAN-13-0186. pmid: 23940299. URL: http://cancerres.aacrjournals.org/content/73/19/6036 (visited on 12/14/2013).

[51] Ronglai Shen, Adam B. Olshen, and Marc Ladanyi. "Integrative Clustering of Multiple Genomic Data Types Using a Joint Latent Variable Model with Application to Breast and Lung Cancer Subtype Analysis". In: *Bioinformatics* 25.22 (Nov. 15, 2009), pp. 2906–2912. ISSN: 1367-4803, 1460-2059. DOI: 10.1093/bioinformatics/btp543. pmid: 19759197. URL: http://bioinformatics.oxfordjournals.org/content/25/22/2906 (visited on 02/05/2015).

[52] Ronglai Shen et al. "Integrative Subtype Discovery in Glioblastoma Using iCluster". In: *PLoS ONE* 7.4 (Apr. 23, 2012), e35236. DOI: 10.1371/journal.pone.0035236. URL: http://dx.doi.org/10.1371/journal.pone.0035236 (visited on 02/04/2015).

[53] XiaoBo Shen and QuanSen Sun. "A Novel Semi-Supervised Canonical Correlation Analysis and Extensions for Multi-View Dimensionality Reduction". In: *Journal of Visual Communication and Image Representation* 25.8 (Nov. 2014), pp. 1894–1904. ISSN: 1047-3203. DOI: 10.1016/j.jvcir.2014.09.004. URL: http://www.sciencedirect.com/science/article/pii/S1047320314001448 (visited on 11/17/2016).

[54] Simon N. Stacey et al. "A Germline Variant in the TP53 Polyadenylation Signal Confers Cancer Susceptibility". In: *Nature Genetics* 43.11 (Nov. 2011), pp. 1098–1103. ISSN: 1061-4036. DOI: 10.1038/ng.926. URL: http://www.nature.com/ng/journal/v43/n11/full/ng.926.html (visited on 06/08/2015).

[55] Michael R. Stratton, Peter J. Campbell, and P. Andrew Futreal. "The Cancer Genome". In: *Nature* 458.7239 (Apr. 9, 2009), pp. 719–724. ISSN: 0028-0836. DOI: 10.1038/nature07943. URL: http://www.nature.com/nature/journal/v458/n7239/abs/nature07943.html (visited on 09/22/2016).

[56] "The Insulin Receptor Isoform Exon 11- (IR-A) in Cancer and Other Diseases: A Review". In: *Hormone and Metabolic Research* 35 (11/12 Nov. 2003), pp. 778–785. ISSN: 0018-5043, 1439-4286. DOI: 10.1055/s-2004-814157. URL: https://www.thieme-connect.com/ejournals/html/10.1055/s-2004-814157 (visited on 11/22/2013).

[57] Michael E. Tipping and Christopher M. Bishop. "Mixtures of Probabilistic Principal Component Analyzers". In: *Neural Computation* 11.2 (Feb. 1, 1999), pp. 443–482. ISSN: 0899-7667. DOI: 10.1162/089976699300016728. URL: http://dx.doi.org/10.1162/089976699300016728 (visited on 10/10/2016).

[58] Michael E. Tipping and Christopher M. Bishop. "Probabilistic Principal Component Analysis". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 61.3 (Jan. 1, 1999), pp. 611–622. ISSN: 1467-9868. DOI: 10.1111/1467-9868.00196. URL:

http://onlinelibrary.wiley.com/doi/10.1111/1467-9868.00196/abstract (visited on 05/27/2014).

[59] Cole Trapnell, Lior Pachter, and Steven L. Salzberg. "TopHat: Discovering Splice Junctions with RNA-Seq". In: *Bioinformatics* 25.9 (Jan. 5, 2009), pp. 1105–1111. ISSN: 1367-4803, 1460-2059. DOI: 10.1093/bioinformatics/btp120. pmid: 19289445. URL: http://bioinformatics.oxfordjournals.org/content/25/9/1105 (visited on 11/13/2016).

[60] Cole Trapnell et al. "Differential Analysis of Gene Regulation at Transcript Resolution with RNA-Seq". In: *Nature Biotechnology* 31.1 (Jan. 2013), pp. 46–53. ISSN: 1087-0156. DOI: 10.1038/nbt.2450. URL: http://www.nature.com/nbt/journal/v31/n1/abs/nbt.2450.html (visited on 11/11/2016).

[61] Cole Trapnell et al. "Transcript Assembly and Quantification by RNA-Seq Reveals Unannotated Transcripts and Isoform Switching during Cell Differentiation". In: *Nature Biotechnology* 28.5 (May 2010), pp. 511–515. ISSN: 1087-0156. DOI: 10.1038/nbt.1621. URL: http://www.nature.com/nbt/journal/v28/n5/abs/nbt.1621.html (visited on 10/10/2013).

[62] Laura J. van 't Veer et al. "Gene Expression Profiling Predicts Clinical Outcome of Breast Cancer". In: *Nature* 415.6871 (Jan. 31, 2002), pp. 530–536. ISSN: 0028-0836. DOI: 10.1038/415530a. URL: http://www.nature.com/nature/journal/v415/n6871/abs/415530a.html (visited on 06/04/2014).

[63] Charles J Vaske et al. "Inference of Patient-Specific Pathway Activities from Multi-Dimensional Cancer Genomics Data Using PARADIGM." In: *Bioinformatics (Oxford, England)* 26.12 (June 2010), pp. i237–45. DOI: 10.1093/bioinformatics/btq182. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2881367&tool=pmcentrez&rendertype=abstract.

[64] Kai Wang et al. "MapSplice: Accurate Mapping of RNA-Seq Reads for Splice Junction Discovery." In: *Nucleic acids research* 38.18 (Oct. 2010), e178–e178. DOI: 10.1093/nar/gkq622. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2952873&tool=pmcentrez&rendertype=abstract.

[65] Robert Weinberg. *The Biology of Cancer*. Garland science, 2013. URL: https://books.google.com/books?hl=en&lr=&id=MzMmAgAAQBAJ&oi=fnd&pg=PR4&dq=the+biology+of+cancer&ots=A0OsZ626Z9&sig=910diCymh0-o_ihxtDXF3JD609E (visited on 11/22/2016).

[66] John N. Weinstein et al. "The Cancer Genome Atlas Pan-Cancer Analysis Project". In: *Nature genetics* 45.10 (Oct. 2013), pp. 1113–1120. ISSN: 1061-4036. DOI: 10.1038/ng.2764. pmid: 24071849. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3919969/ (visited on 09/22/2016).

[67] *What the FPKM? A Review of RNA-Seq Expression Units*. 2014-05-08T18:55:06+00:00. URL: https://haroldpimentel.wordpress.com/2014/05/08/what-the-fpkm-a-review-rna-seq-expression-units/ (visited on 11/12/2016).

[68] Daniela M. Witten and Robert J. Tibshirani. "Extensions of Sparse Canonical Correlation Analysis with Applications to Genomic Data". In: *Statistical Applications in Genetics and Molecular Biology* 8.1 (June 9, 2009), pp. 1–27. URL: http://www.degruyter.com/view/j/sagmb.2009.8.1/sagmb.2009.8.1.1470/sagmb.2009.8.1.1470.xml;jsessionid=E7A72A14672827A169BB102DFA3C196B (visited on 08/27/2013).

[69] Lixing Yang et al. "Diverse Mechanisms of Somatic Structural Variations in Human Cancer Genomes". In: *Cell* 153.4 (2013), pp. 919–929. URL: http://www.sciencedirect.com/science/article/pii/S0092867413004510 (visited on 11/22/2016).

[70] Shipeng Yu et al. "Supervised Probabilistic Principal Component Analysis". In: *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '06. New York, NY, USA: ACM, 2006, pp. 464–473. ISBN: 1-59593-339-5. DOI: 10.1145/1150402.1150454. URL: http://doi.acm.org/10.1145/1150402.1150454 (visited on 08/10/2015).

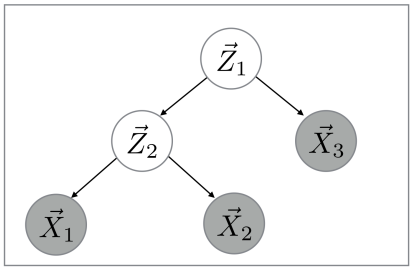
[71] Yuan Yuan et al. "Assessing the Clinical Utility of Cancer Genomic and Proteomic Data across Tumor Types". In: *Nature Biotechnology* 32.7 (July 2014), pp. 644–652. ISSN: 1087-0156. DOI: 10.1038/nbt.2940. URL: http://www.nature.com/nbt/journal/v32/n7/full/nbt.2940.html (visited on 10/21/2014).

[72] J. J. Yunis and A. L. Soreng. "Constitutive Fragile Sites and Cancer". In: *Science* 226.4679 (Dec. 7, 1984), pp. 1199–1204. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.6239375. pmid: 6239375. URL: http://science.sciencemag.org/content/226/4679/1199 (visited on 09/22/2016).

Appendix A

Three-node model formulation

A.1 Model



$$P\left(\{\vec{x}_{i,n}\}_{i=1}^{3}, \{\vec{z}_{j,n}\}_{j=1}^{2} | \{\mathbf{W}_k\}_{k=0}^{3}, \{\mathbf{\Psi}_\ell\}_{\ell=0}^{3}\right) = P(\vec{z}_1)P(\vec{z}_2|\vec{z}_1)P(\vec{x}_1|\vec{z}_2)P(\vec{x}_2|\vec{z}_2)P(\vec{x}_3|\vec{z}_1)$$

$$\begin{aligned}
\vec{z}_1 &\sim N(\vec{0}, \mathbf{I}) \in \mathbb{R}^{m\times 1} \\
\vec{z}_2|\vec{z}_1 &\sim N(\mathbf{W}_0\vec{z}_1, \mathbf{\Psi}_0) \in \mathbb{R}^{p_0\times 1} \\
\vec{x}_1|\vec{z}_2 &\sim N(\mathbf{W}_1\vec{z}_2, \mathbf{\Psi}_1) \in \mathbb{R}^{p_1\times 1} \\
\vec{x}_2|\vec{z}_2 &\sim N(\mathbf{W}_2\vec{z}_2, \mathbf{\Psi}_2) \in \mathbb{R}^{p_2\times 1} \\
\vec{x}_3|\vec{z}_1 &\sim N(\mathbf{W}_3\vec{z}_1, \mathbf{\Psi}_3) \in \mathbb{R}^{p_3\times 1}
\end{aligned}$$

$$\begin{aligned}
\mathbf{W}_0 &\in \mathbb{R}^{p_0\times m} & \mathbf{\Psi}_0 &\in \mathbb{R}^{p_0\times p_0} \\
\mathbf{W}_1 &\in \mathbb{R}^{p_1\times p_0} & \mathbf{\Psi}_1 &\in \mathbb{R}^{p_1\times p_1} \\
\mathbf{W}_2 &\in \mathbb{R}^{p_2\times p_0} & \mathbf{\Psi}_2 &\in \mathbb{R}^{p_2\times p_2} \\
\mathbf{W}_3 &\in \mathbb{R}^{p_3\times m} & \mathbf{\Psi}_3 &\in \mathbb{R}^{p_3\times p_3}
\end{aligned}$$

A.2 Complete log-likelihood

For N samples,

$$\begin{aligned}
\ell_n(\Theta) \coloneqq \sum_{n=1}^{N} \log\left(P\left(\{\vec{x}_{i,n}\}_{i=1}^{3}, \{\vec{z}_{j,n}\}_{j=1}^{2} | \{\mathbf{W}_k\}_{k=0}^{3}, \{\mathbf{\Psi}_\ell\}_{\ell=0}^{3}\right)\right) = \\
\sum_{n=1}^{N} -\frac{p_1}{2}\log(2\pi) - \frac{1}{2}\log(|\mathbf{\Psi}_1|) - \frac{1}{2}\left(\left(\vec{x}_{1,n} - \mathbf{W}_1\vec{z}_{2,n}\right)^T \mathbf{\Psi}_1^{-1}\left(\vec{x}_{1,n} - \mathbf{W}_1\vec{z}_{2,n}\right)\right) + \\
- \frac{p_2}{2}\log(2\pi) - \frac{1}{2}\log(|\mathbf{\Psi}_2|) - \frac{1}{2}\left(\left(\vec{x}_{2,n} - \mathbf{W}_2\vec{z}_{2,n}\right)^T \mathbf{\Psi}_2^{-1}\left(\vec{x}_{2,n} - \mathbf{W}_2\vec{z}_{2,n}\right)\right) + \\
- \frac{p_3}{2}\log(2\pi) - \frac{1}{2}\log(|\mathbf{\Psi}_3|) - \frac{1}{2}\left(\left(x_{3,n} - \mathbf{W}_3\vec{z}_{1,n}\right)^T \mathbf{\Psi}_3^{-1}\left(\vec{x}_{3,n} - \mathbf{W}_3\vec{z}_{1,n}\right)\right) + \\
- \frac{p_0}{2}\log(2\pi) - \frac{1}{2}\log(|\mathbf{\Psi}_0|) - \frac{1}{2}\left(\left(\vec{z}_{2,n} - \mathbf{W}_0\vec{z}_{1,n}\right)^T \mathbf{\Psi}_0^{-1}\left(\vec{z}_{2,n} - \mathbf{W}_0\vec{z}_{1,n}\right)\right) + \\
- \frac{m}{2}\log(2\pi) - \frac{1}{2}\vec{z}_{1,n}^T\vec{z}_{1,n}
\end{aligned} \tag{A.1}$$

$$\begin{aligned}
\Rightarrow \mathbb{E}_{\{\vec{z}_j\}_{j=1}^{2}|\{\vec{x}_i\}_{i=1}^{3}}\left[\sum_{n=1}^{N}\log\left(P\left(\{\vec{x}_{i,n}\}_{i=1}^{3}, \{\vec{z}_{j,n}\}_{j=1}^{2} | \{\mathbf{W}_k\}_{k=0}^{3}, \{\mathbf{\Psi}_\ell\}_{\ell=0}^{3}\right)\right)\right] = \\
\sum_{n=1}^{N} -\frac{p_1}{2}\log(2\pi) - \frac{1}{2}\log(|\mathbf{\Psi}_1|) - \frac{1}{2}\left(\vec{x}_{1,n}^T\mathbf{\Psi}_1^{-1}\vec{x}_{1,n} - 2\vec{x}_{1,n}^T\mathbf{\Psi}_1^{-1}\mathbf{W}_1\mathbb{E}[\vec{z}_{2,n}] + \mathrm{Tr}\left(\mathbb{E}[\vec{z}_{2,n}\vec{z}_{2,n}^T]\mathbf{W}_1^T\mathbf{\Psi}_1^{-1}\mathbf{W}_1\right)\right) \\
- \frac{p_2}{2}\log(2\pi) - \frac{1}{2}\log(|\mathbf{\Psi}_2|) - \frac{1}{2}\left(\vec{x}_{2,n}^T\mathbf{\Psi}_2^{-1}\vec{x}_{2,n} - 2\vec{x}_{2,n}^T\mathbf{\Psi}_2^{-1}\mathbf{W}_2\mathbb{E}[\vec{z}_{2,n}] + \mathrm{Tr}\left(\mathbb{E}[\vec{z}_{2,n}\vec{z}_{2,n}^T]\mathbf{W}_2^T\mathbf{\Psi}_2^{-1}\mathbf{W}_2\right)\right) \\
- \frac{p_3}{2}\log(2\pi) - \frac{1}{2}\log(|\mathbf{\Psi}_3|) - \frac{1}{2}\left(\vec{x}_{3,n}^T\mathbf{\Psi}_3^{-1}\vec{x}_{3,n} - 2\vec{x}_{3,n}^T\mathbf{\Psi}_3^{-1}\mathbf{W}_3\mathbb{E}[\vec{z}_{1,n}] + \mathrm{Tr}\left(\mathbb{E}[\vec{z}_{1,n}\vec{z}_{1,n}^T]\mathbf{W}_3^T\mathbf{\Psi}_3^{-1}\mathbf{W}_3\right)\right) \\
- \frac{p_0}{2}\log(2\pi) - \frac{1}{2}\log(|\mathbf{\Psi}_0|) - \frac{1}{2}\left(\mathrm{Tr}\left(\mathbb{E}[\vec{z}_{2,n}\vec{z}_{2,n}^T]\mathbf{\Psi}_0^{-1}\right) - 2\mathrm{Tr}\left(\mathbb{E}[\vec{z}_{1,n}\vec{z}_{2,n}^T]\mathbf{\Psi}_0^{-1}\mathbf{W}_0\right) + \mathrm{Tr}\left(\mathbb{E}[\vec{z}_{1,n}\vec{z}_{1,n}^T]\mathbf{W}_0^T\mathbf{\Psi}_0^{-1}\mathbf{W}_0\right)\right) \\
- \frac{m}{2}\log(2\pi) - \frac{1}{2}\mathbb{E}[\vec{z}_{1,n}^T\vec{z}_{1,n}]
\end{aligned} \tag{A.2}$$

A.3 M-Step: finding critical points

W

$$\begin{aligned}
&\frac{\partial}{\partial \mathbf{W}_0} \\
&= \sum_{n=1}^{N} \frac{\partial}{\partial \mathbf{W}_0} \mathrm{Tr}\left(\mathbb{E}[\vec{z}_{1,n}\vec{z}_{2,n}^T]\boldsymbol{\Psi}_0^{-1}\mathbf{W}_0\right) - \frac{1}{2}\frac{\partial}{\partial \mathbf{W}_0}\mathrm{Tr}\left(\mathbb{E}[\vec{z}_{1,n}\vec{z}_{1,n}^T]\mathbf{W}_0^T\boldsymbol{\Psi}_0^{-1}\mathbf{W}_0\right) \\
&= \mathbb{E}[\vec{z}_{1,n}\vec{z}_{2,n}^T]\boldsymbol{\Psi}_0^{-1} - \mathbb{E}[\vec{z}_{1,n}\vec{z}_{1,n}^T]\mathbf{W}_0^T\boldsymbol{\Psi}_0^{-1}
\end{aligned}$$

$$\Rightarrow \boxed{\hat{\mathbf{W}}_{0\,\mathrm{MLE}}^T = \left(\sum_{n=1}^{N}\mathbb{E}[\vec{z}_{1,n}\vec{z}_{1,n}^T]\right)^{-1}\left(\sum_{n=1}^{N}\mathbb{E}[\vec{z}_{1,n}\vec{z}_{2,n}^T]\right)}$$

$$\begin{aligned}
&\frac{\partial}{\partial \mathbf{W}_1} \\
&= \sum_{n=1}^{N} \frac{\partial}{\partial \mathbf{W}_1} \vec{x}_{1,n}^T\boldsymbol{\Psi}_1^{-1}\mathbf{W}_1\mathbb{E}[\vec{z}_{2,n}] - \frac{1}{2}\frac{\partial}{\partial \mathbf{W}_1}\mathrm{Tr}\left(\mathbb{E}[\vec{z}_{2,n}\vec{z}_{2,n}^T]\mathbf{W}_1^T\boldsymbol{\Psi}_1^{-1}\mathbf{W}_1\right) \\
&= \mathbb{E}[\vec{z}_{2,n}]\vec{x}_{1,n}^T\boldsymbol{\Psi}_1^{-1} - \mathbb{E}[\vec{z}_{2,n}\vec{z}_{2,n}^T]\mathbf{W}_1^T\boldsymbol{\Psi}_1^{-1}
\end{aligned}$$

$$\Rightarrow \boxed{\hat{\mathbf{W}}_{1\,\mathrm{MLE}}^T = \left(\sum_{n=1}^{N}\mathbb{E}[\vec{z}_{2,n}\vec{z}_{2,n}^T]\right)^{-1}\left(\sum_{n=1}^{N}\mathbb{E}[\vec{z}_{2,n}]\vec{x}_{1,n}^T\right)}$$

By pattern-matching,

$$\boxed{\hat{\mathbf{W}}_{2\,\mathrm{MLE}}^T = \left(\sum_{n=1}^{N}\mathbb{E}[\vec{z}_{2,n}\vec{z}_{2,n}^T]\right)^{-1}\left(\sum_{n=1}^{N}\mathbb{E}[\vec{z}_{2,n}]\vec{x}_{2,n}^T\right)}$$

$$\boxed{\hat{\mathbf{W}}_{3\,\mathrm{MLE}}^T = \left(\sum_{n=1}^{N}\mathbb{E}[\vec{z}_{1,n}\vec{z}_{1,n}^T]\right)^{-1}\left(\sum_{n=1}^{N}\mathbb{E}[\vec{z}_{1,n}]\vec{x}_{3,n}^T\right)}$$

$\boldsymbol{\Psi}$

$$\begin{aligned}
&\frac{\partial}{\partial \boldsymbol{\Psi}_0} = \\
&\sum_{n=1}^{N}\Bigg(-\frac{\partial}{\partial \boldsymbol{\Psi}_0}\frac{1}{2}\log\left(|\boldsymbol{\Psi}_0|\right) - \frac{1}{2}\frac{\partial}{\partial \boldsymbol{\Psi}_0}\mathrm{Tr}\left(\mathbb{E}[\vec{z}_{2,n}\vec{z}_{2,n}^T]\boldsymbol{\Psi}_0^{-1}\right) + \\
&\frac{\partial}{\partial \boldsymbol{\Psi}_0}\mathrm{Tr}\left(\mathbb{E}[\vec{z}_{1,n}\vec{z}_{2,n}^T]\boldsymbol{\Psi}_0^{-1}\mathbf{W}_0\right) - \frac{1}{2}\frac{\partial}{\partial \boldsymbol{\Psi}_0}\mathrm{Tr}\left(\mathbb{E}[\vec{z}_{1,n}\vec{z}_{1,n}^T]\mathbf{W}_0^T\boldsymbol{\Psi}_0^{-1}\mathbf{W}_0\right) \qquad 00 \\
&= \sum_{n=1}^{N}\left(-\frac{1}{2}\boldsymbol{\Psi}_0^{-1} + \frac{1}{2}\boldsymbol{\Psi}_0^{-1}\mathbb{E}[\vec{z}_{2,n}\vec{z}_{2,n}^T]\boldsymbol{\Psi}_0^{-1} - \boldsymbol{\Psi}_0^{-1}\mathbb{E}[\vec{z}_{2,n}\vec{z}_{1,n}^T]\mathbf{W}_0^T\boldsymbol{\Psi}_0^{-1} + \frac{1}{2}\boldsymbol{\Psi}_0^{-1}\mathbf{W}_0\mathbb{E}[\vec{z}_{1,n}\vec{z}_{1,n}^T]\mathbf{W}_0^T\boldsymbol{\Psi}_0^{-1}\right) \\
\Rightarrow \hat{\boldsymbol{\Psi}}_{0\mathrm{MLE}} &= \frac{1}{N}\sum_{n=1}^{N}\left(\mathbb{E}[\vec{z}_{2,n}\vec{z}_{2,n}^T] - 2\mathbb{E}[\vec{z}_{2,n}\vec{z}_{1,n}^T]\mathbf{W}_0^T + \mathbf{W}_0\mathbb{E}[\vec{z}_{1,n}\vec{z}_{1,n}^T]\mathbf{W}_0^T\right)
\end{aligned}$$

By pattern-matching,

$$\begin{aligned}
\hat{\boldsymbol{\Psi}}_{1\mathrm{MLE}} &= \frac{1}{N}\sum_{n=1}^{N}\left(\vec{x}_{1,n}\vec{x}_{1,n}^T - 2\vec{x}_{1,n}\mathbb{E}[\vec{z}_{2,n}^T]\mathbf{W}_1^T + \mathbf{W}_1\mathbb{E}[\vec{z}_{2,n}\vec{z}_{2,n}^T]\mathbf{W}_1^T\right) \\
\hat{\boldsymbol{\Psi}}_{2\mathrm{MLE}} &= \frac{1}{N}\sum_{n=1}^{N}\left(\vec{x}_{2,n}\vec{x}_{2,n}^T - 2\vec{x}_{2,n}\mathbb{E}[\vec{z}_{2,n}^T]\mathbf{W}_2^T + \mathbf{W}_2\mathbb{E}[\vec{z}_{2,n}\vec{z}_{2,n}^T]\mathbf{W}_2^T\right) \\
\hat{\boldsymbol{\Psi}}_{3\mathrm{MLE}} &= \frac{1}{N}\sum_{n=1}^{N}\left(\vec{x}_{3,n}\vec{x}_{3,n}^T - 2\vec{x}_{3,n}\mathbb{E}[\vec{z}_{1,n}^T]\mathbf{W}_3^T + \mathbf{W}_3\mathbb{E}[\vec{z}_{1,n}\vec{z}_{1,n}^T]\mathbf{W}_3^T\right)
\end{aligned}$$

Using a generalized EM algorithm, and therefore substituting $\mathbf{W}$ with $\hat{\mathbf{W}}_{\mathrm{MLE}}$ in $\hat{\boldsymbol{\Psi}}_{\mathrm{MLE}}$:

$$\boxed{\hat{\mathbf{\Psi}}_{0\text{MLE}} = \frac{1}{N}\sum_{n=1}^{N}\left(\mathbb{E}[\vec{z}_{2,n}\vec{z}_{2,n}^{T}] - \mathbb{E}[\vec{z}_{2,n}\vec{z}_{1,n}^{T}]\left(\mathbb{E}[z_{1,n}\vec{z}_{1,n}^{T}]\right)^{-1}\mathbb{E}[\vec{z}_{1,n}\vec{z}_{2,n}^{T}]\right)}$$

$$\boxed{\hat{\mathbf{\Psi}}_{1\text{MLE}} = \frac{1}{N}\sum_{n=1}^{N}\left(\vec{x}_{1,n}\vec{x}_{1,n}^{T} - \vec{x}_{1,n}\mathbb{E}[\vec{z}_{2,n}^{T}]\left(\mathbb{E}[\vec{z}_{2,n}\vec{z}_{2,n}^{T}]\right)^{-1}\mathbb{E}[\vec{z}_{2,n}]\vec{x}_{1,n}^{T}\right)}$$

$$\boxed{\hat{\mathbf{\Psi}}_{2\text{MLE}} = \frac{1}{N}\sum_{n=1}^{N}\left(\vec{x}_{2,n}\vec{x}_{2,n}^{T} - \vec{x}_{2,n}\mathbb{E}[\vec{z}_{2,n}^{T}]\left(\mathbb{E}[\vec{z}_{2,n}\vec{z}_{2,n}^{T}]\right)^{-1}\mathbb{E}[\vec{z}_{2,n}]\vec{x}_{2,n}^{T}\right)}$$

$$\boxed{\hat{\mathbf{\Psi}}_{3\text{MLE}} = \frac{1}{N}\sum_{n=1}^{N}\left(\vec{x}_{3,n}\vec{x}_{3,n}^{T} - \vec{x}_{3,n}\mathbb{E}[\vec{z}_{1,n}^{T}]\left(\mathbb{E}[\vec{z}_{1,n}\vec{z}_{1,n}^{T}]\right)^{-1}\mathbb{E}[\vec{z}_{1,n}]\vec{x}_{3,n}^{T}\right)}$$

Note that in the case of assuming diagonal errors $\mathbf{\Psi}_i$, we simply take the diagonal component of every $\mathbf{\Psi}_i$ when performing each update.

$$\hat{\mathbf{\Psi}}_{i,\text{MLE}} \coloneqq \text{diag}\left(\hat{\mathbf{\Psi}}_{i,\text{MLE}}\right) \tag{A.3}$$

A.4 E-Step: finding posterior conditional distribution

By inspection,

$$\vec{z}_1, \vec{z}_2, \vec{x}_1, \vec{x}_2, \vec{x}_3 \sim N(\vec{\mu}, \mathbf{\Sigma}), \text{where}$$

$$\mathbf{\Lambda} \coloneqq \mathbf{\Sigma}^{-1} =$$

$$\begin{array}{c|cc|ccc} & \vec{z}_1 & \vec{z}_2 & \vec{x}_1 & \vec{x}_2 & \vec{x}_3 \\ \vec{z}_1 & \mathbf{W}_3^T\mathbf{\Psi}_3^{-1}\mathbf{W}_3 + \mathbf{W}_0^T\mathbf{\Psi}_0^{-1}\mathbf{W}_0 + \mathbf{I} & -\mathbf{W}_0^T\mathbf{\Psi}_0^{-1} & \mathbf{0} & \mathbf{0} & -\mathbf{W}_3^T\mathbf{\Psi}_3^{-1} \\ \vec{z}_2 & -\mathbf{\Psi}_0^{-1}\mathbf{W}_0 & \mathbf{W}_1^T\mathbf{\Psi}_1^{-1}\mathbf{W}_1 + \mathbf{W}_2^T\mathbf{\Psi}_2^{-1}\mathbf{W}_2 + \mathbf{\Psi}_0^{-1} & -\mathbf{W}_1^T\mathbf{\Psi}_1^{-1} & -\mathbf{W}_2^T\mathbf{\Psi}_2^{-1} & \mathbf{0} \\ \hline \vec{x}_1 & \mathbf{0} & -\mathbf{\Psi}_1^{-1}\mathbf{W}_1 & \mathbf{\Psi}_1^{-1} & \mathbf{0} & \mathbf{0} \\ \vec{x}_2 & \mathbf{0} & -\mathbf{\Psi}_2^{-1}\mathbf{W}_2 & \mathbf{0} & \mathbf{\Psi}_2^{-1} & \mathbf{0} \\ \vec{x}_3 & -\mathbf{\Psi}_3^{-1}\mathbf{W}_3 & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{\Psi}_3^{-1} \end{array} \tag{A.4}$$

$$\coloneqq \left(\begin{array}{c|c} \mathbf{\Lambda}_{zz} & \mathbf{\Lambda}_{zx} \\ \hline \mathbf{\Lambda}_{xz} & \mathbf{\Lambda}_{xx} \end{array}\right)$$

$$, \text{ where } \vec{x} \coloneqq \begin{pmatrix} \vec{x}_1 \\ \vec{x}_2 \\ \vec{x}_3 \end{pmatrix}, \vec{z} \coloneqq \begin{pmatrix} \vec{z}_1 \\ \vec{z}_2 \end{pmatrix} \tag{A.5}$$

Working with precision matrix

We know that

$$\vec{z}|\vec{x} \sim N(\vec{\mu}_{z|x}, \mathbf{\Sigma}_{z|x}) \text{ ,where} \tag{A.6}$$

$$\vec{\mu}_{z|x} = \vec{\mu}_z - \mathbf{\Lambda}_{zz}^{-1}\mathbf{\Lambda}_{zx}\left(\vec{x} - \vec{\mu}_x\right) \tag{A.7}$$

$$\mathbf{\Sigma}_{z|x} = \mathbf{\Lambda}_{zz}^{-1} \coloneqq \left(\begin{array}{c|c} \mathbf{\Sigma}_{z_1 z_1|x} & \mathbf{\Sigma}_{z_1 z_2|x} \\ \hline \mathbf{\Sigma}_{z_2 z_1|x} & \mathbf{\Sigma}_{z_2 z_2|x} \end{array}\right) \tag{A.8}$$

Finding $\vec{\mu}_{z|x}$ by expanding (5)

$$\vec{\mu} = \vec{0}^{d\times 1}, \text{ where } d \coloneqq m + \sum_{i=0}^{3} p_i \tag{A.9}$$

$$\vec{\mu} \coloneqq \begin{pmatrix} \vec{\mu}_z \\ \vec{\mu}_x \end{pmatrix} \text{ , where } \vec{\mu}_z \in \mathbb{R}^{(m+p_0)\times 1}, \vec{\mu}_x \in \mathbb{R}^{(p_1+p_2+p_3)\times 1}$$

$$\Rightarrow \vec{\mu}_{z|x} = -\mathbf{\Lambda}_{zz}^{-1}\mathbf{\Lambda}_{zx}\vec{x} \tag{A.10}$$

Now we have what we need for the E-step:

$$\boxed{\mathbb{E}[\vec{z}_1|\vec{x}] = \vec{\mu}_{z_1|x}} \tag{A.11}$$

$$\boxed{\mathbb{E}[\vec{z}_2|\vec{x}] = \vec{\mu}_{z_2|x}} \tag{A.12}$$

$$\boxed{\mathbb{E}[\vec{z}_1\vec{z}_1^T|\vec{x}] = \mathbf{\Sigma}_{z_1 z_1|x} + \vec{\mu}_{z_1|x}\vec{\mu}_{z_1|x}^T} \tag{A.13}$$

$$\boxed{\mathbb{E}[\vec{z}_2\vec{z}_2^T|\vec{x}] = \mathbf{\Sigma}_{z_2 z_2|x} + \vec{\mu}_{z_2|x}\vec{\mu}_{z_2|x}^T} \tag{A.14}$$

$$\boxed{\mathbb{E}[\vec{z}_1\vec{z}_2^T|\vec{x}] = \mathbf{\Sigma}_{z_1 z_2|x} + \vec{\mu}_{z_1|x}\vec{\mu}_{z_2|x}^T} \tag{A.15}$$

A.5 Finding conditional distribution of outcome observed variable given input observed variables

For the purposes of classification, we observe $\vec{x}_1$ and $\vec{x}_2$ and would like to predict $\vec{x}_3$ based on parameters learned during fitting on training data $\hat{\mathbf{\Theta}} \coloneqq \left\{ \{\hat{W}_k\}_{k=0}^3, \{\hat{\mathbf{\Psi}}_\ell\}_{\ell=0}^3 \right\}$. Thus, we're interested in obtaining an expression for:

$$\mathbb{E}[\vec{x}_3 | \vec{x}_1, \vec{x}_2, \mathbf{\Theta}] \tag{A.16}$$

We obtain this first based on partitioning the joint covariance matrix and partitioned Gaussian identities (A.5), and then based on the direct integral form of the expectation (A.5).

Partitioning

By equation A.4 and by partitioning,

$$\mathbf{\Sigma} = \mathbf{\Lambda}^{-1} \tag{A.17}$$

$$= \left(\begin{array}{c|c} \mathbf{\Sigma}_{zz} & \mathbf{\Sigma}_{zx} \\ \hline \mathbf{\Sigma}_{xz} & \mathbf{\Sigma}_{xx} \end{array} \right) \tag{A.18}$$

, where $\vec{x}$ and $\vec{z}$ are defined as in A.5.

Since the marginal distribution of a partition of a Gaussian has the joint covariance of that same partition, the joint distribution of the observed variables will have covariance $\mathbf{\Sigma}_{xx}$.

We can then obtain the marginal observed precision matrix ($\mathbf{\Lambda}'$) by inverting, can expand this into blocks for each observed variable, and partition between observed and prediction variables:

$$\mathbf{\Lambda}' \coloneqq \mathbf{\Sigma}_{xx}^{-1} \tag{A.19}$$

$$= \left(\begin{array}{cc|c} \mathbf{\Lambda}'_{x_1x_1} & \mathbf{\Lambda}'_{x_1x_2} & \mathbf{\Lambda}'_{x_1x_3} \\ \mathbf{\Lambda}'_{x_2x_1} & \mathbf{\Lambda}'_{x_2x_2} & \mathbf{\Lambda}'_{x_2x_3} \\ \hline \mathbf{\Lambda}'_{x_3x_1} & \mathbf{\Lambda}'_{x_3x_2} & \mathbf{\Lambda}'_{x_3x_3} \end{array} \right) \tag{A.20}$$

By partitioned Guassian identity (2.97 of Bishop's *Pattern Recognition*[5]), $\vec{x}_3 | \vec{x}_1, \vec{x}_2$ is distributed as a Gaussian with expectation:

$$\mathbb{E}[\vec{x}_3 | \vec{x}_1, \vec{x}_2, \mathbf{\Theta}] = \vec{\mu}_{x_3} - \left(\mathbf{\Lambda}'_{x_3x_3}\right)^{-1} \begin{pmatrix} \mathbf{\Lambda}'_{x_3x_1} & \mathbf{\Lambda}'_{x_3x_2} \end{pmatrix} \left(\begin{pmatrix} \vec{x}_1 \\ \vec{x}_2 \end{pmatrix} - \begin{pmatrix} \vec{\mu}_{x_1} \\ \vec{\mu}_{x_2} \end{pmatrix} \right) \tag{A.21}$$

$$\tag{A.22}$$

, where $\vec{\mu}_{x_i}$ are additional parameters learned during fitting step.

Direct

Starting with the general form for conditional expectation,

$$\mathbb{E}_{\mathbf{\Theta}}\left[\vec{x}_3|\vec{x}_1,\vec{x}_2\right] = \int \vec{x}_3 P(\vec{x}_3|\vec{x}_1,\vec{x}_2)d\vec{x}_3 \tag{A.23}$$

$$= \int \vec{x}_3 \frac{P(\vec{x}_1,\vec{x}_2,\vec{x}_3)}{P(\vec{x}_1,\vec{x}_2)}d\vec{x}_3 \tag{A.24}$$

$$= \int\int\int \vec{x}_3 \frac{P(\vec{x}_1,\vec{x}_2,\vec{x}_3,\vec{z}_1,\vec{z}_2)}{P(\vec{x}_1,\vec{x}_2)}d\vec{z}_1 d\vec{z}_2 d\vec{x}_3 \tag{A.25}$$

$$= \int\int\int \vec{x}_3 \frac{P(\vec{x}_1,\vec{x}_2,\vec{x}_3,\vec{z}_1,\vec{z}_2)}{P(\vec{x}_1,\vec{x}_2)}d\vec{x}_3 d\vec{z}_1 d\vec{z}_2 \tag{A.26}$$

$$= \int\int \frac{P(\vec{x}_1,\vec{x}_2,\vec{z}_1,\vec{z}_2)}{P(\vec{x}_1,\vec{x}_2)}\left(\int \vec{x}_3 P(\vec{x}_3|\vec{z}_1)d\vec{x}_3\right) d\vec{z}_1 d\vec{z}_2 \tag{A.27}$$

$$= \int\int \frac{P(\vec{x}_1,\vec{x}_2,\vec{z}_1,\vec{z}_2)}{P(\vec{x}_1,\vec{x}_2)}\left(\mathbb{E}_{\mathbf{\Theta}}\left[\vec{x}_3|\vec{z}_1\right]\right) d\vec{z}_1 d\vec{z}_2 \tag{A.28}$$

$$= \int\int \frac{P(\vec{x}_1,\vec{x}_2,\vec{z}_1,\vec{z}_2)}{P(\vec{x}_1,\vec{x}_2)}\left(\mathbf{W}_3\vec{z}_1 + \vec{\mu}_3\right) d\vec{z}_1 d\vec{z}_2 \tag{A.29}$$

$$= \int\int \frac{P(\vec{x}_1,\vec{x}_2,\vec{z}_1,\vec{z}_2)}{P(\vec{x}_1,\vec{x}_2)}\left(\mathbf{W}_3\vec{z}_1 + \vec{\mu}_3\right) d\vec{z}_2 d\vec{z}_1 \tag{A.30}$$

$$= \int \left(\mathbf{W}_3\vec{z}_1 + \vec{\mu}_3\right)\left(\int \frac{P(\vec{x}_1,\vec{x}_2,\vec{z}_1,\vec{z}_2)}{P(\vec{x}_1,\vec{x}_2)}d\vec{z}_2\right) d\vec{z}_1 \tag{A.31}$$

$$= \int \left(\mathbf{W}_3\vec{z}_1 + \vec{\mu}_3\right)\left(\frac{P(\vec{x}_1,\vec{x}_2,\vec{z}_1)}{P(\vec{x}_1,\vec{x}_2)}\right) d\vec{z}_1 \tag{A.32}$$

$$= \int \left(\mathbf{W}_3\vec{z}_1 + \vec{\mu}_3\right) P(\vec{z}_1|\vec{x}_1,\vec{x}_2)d\vec{z}_1 \tag{A.33}$$

$$= \mathbf{W}_3 \int \vec{z}_1 P(\vec{z}_1|\vec{x}_1,\vec{x}_2)d\vec{z}_1 + \vec{\mu}_3 \int P(\vec{z}_1|\vec{x}_1,\vec{x}_2)d\vec{z}_1 \tag{A.34}$$

$$= \mathbf{W}_3\mathbb{E}_{\mathbf{\Theta}}[\vec{z}_1|\vec{x}_1,\vec{x}_2] + \vec{\mu}_3 \tag{A.35}$$

, where line A.27 follows based on partial factorization of complete joint probability based on model dependency structures.

A.6 Finding Θ for warm starts where child node is of lower dimension than parent

This ignores all variance contributed to a node by its grandparents.

Using the example of nodes $\vec{x}_3$ and $\vec{z}_1$ above of dimension p_3 and m, respectively.
Decomposing variance of a variable into related parameters:

$$\begin{aligned} \text{Var}(\vec{x}_3) &= \text{Var}(\mathbf{W}_3\vec{z}_1 + \mathbf{\Psi}) && \text{(A.36)} \\ &= \mathbf{W}_3\text{Var}(\vec{z}_1)\mathbf{W}_3^T + \mathbf{\Psi}_3 && \text{(A.37)} \\ &= \mathbf{W}_3\mathbf{I}_m\mathbf{W}_3^T + \mathbf{\Psi}_3 && \text{(A.38)} \\ &= \mathbf{W}_3\mathbf{W}_3^T + \mathbf{\Psi}_3 && \text{(A.39)} \end{aligned}$$

Thus, for a warm start, for a computed variance $\mathbf{A} \leftarrow \text{Var}(\vec{x}_3)$ we need to find $\hat{\mathbf{W}}_{3,\text{warm start}}$ and $\hat{\mathbf{\Psi}}_{3,\text{warm start}}$ such that

$$\hat{\mathbf{W}}_{3,\text{warm start}}\hat{\mathbf{W}}_{3,\text{warm start}}^T + \hat{\mathbf{\Psi}}_{3,\text{warm start}} = \mathbf{A} \tag{A.40}$$

One possible solution is to break half of the diagonal variance into $\hat{\mathbf{W}}_{3,\text{warm start}}$ and half into $\hat{\mathbf{\Psi}}_{3,\text{warm start}}$, and keep all the non-diagonal variance in $\hat{\mathbf{\Psi}}_{3,\text{warm start}}$:
Now, we take:

$$\boxed{\hat{\mathbf{\Psi}}_{3,\text{warm start}} \leftarrow \text{non-diag}(\mathbf{A}) + \frac{1}{2}\text{diag}(\mathbf{A})} \tag{A.41}$$

$$\boxed{\hat{\mathbf{W}}_{3,\text{warm start}} \leftarrow \frac{1}{\sqrt{2}}\left(\begin{array}{c|c} \sqrt{\text{diag}(\mathbf{A})} & \mathbf{0} \end{array}\right)} \tag{A.42}$$

, where $\mathbf{0}$ on line A.42 is $p_3 \times m - p_3$.

Proof

all variance is recovered by this scheme:

$$\begin{aligned}
\mathrm{Var}(\vec{x}_3) &= \mathbf{W}_3\mathbf{W}_3^T + \mathbf{\Psi}_3 \text{ by (A.39)} && \text{(A.43)}\\
&= \hat{\mathbf{W}}_{3,\text{warm start}}\hat{\mathbf{W}}_{3,\text{warm start}}^T + \hat{\mathbf{\Psi}}_{3,\text{warm start}} && \text{(A.44)}\\
&= \frac{1}{\sqrt{2}}\left(\ \sqrt{\mathrm{diag}(\mathbf{A})}\ \middle|\ \mathbf{0}\ \right)\frac{1}{\sqrt{2}}\left(\ \sqrt{\mathrm{diag}(\mathbf{A})}\ \middle|\ \mathbf{0}\ \right)^T + \text{non-diag}(\mathbf{A}) + \frac{1}{2}\mathrm{diag}(\mathbf{A}) && \text{(A.45)}\\
&= \frac{1}{2}\left(\ \mathrm{diag}(\mathbf{A})^{\frac{1}{2}}\ \middle|\ \mathbf{0}\ \right)\begin{pmatrix}\mathrm{diag}(\mathbf{A})^{\frac{1}{2}}\\ \mathbf{0}\end{pmatrix} + \text{non-diag}(\mathbf{A}) + \frac{1}{2}\mathrm{diag}(\mathbf{A}) && \text{(A.46)}\\
&= \frac{1}{2}\mathrm{diag}(\mathbf{A}) + \text{non-diag}(\mathbf{A}) + \frac{1}{2}\mathrm{diag}(\mathbf{A}) && \text{(A.47)}\\
&= \mathrm{diag}(\mathbf{A}) + \text{non-diag}(\mathbf{A}) && \text{(A.48)}\\
&= \mathbf{A} && \text{(A.49)}
\end{aligned}$$

Example

$p_3 = 1, m = 2 \Rightarrow \mathbf{A} = (a_1)$

$$\begin{aligned}
\hat{\mathbf{\Psi}}_{3,\text{warm start}} &\leftarrow \text{non-diag}(\mathbf{A}) + \frac{1}{2}\mathrm{diag}(\mathbf{A}) && \text{(A.50)}\\
&= \frac{1}{2}(a_1) && \text{(A.51)}
\end{aligned}$$

$$\begin{aligned}
\hat{\mathbf{W}}_{3,\text{warm start}} &\leftarrow \frac{1}{\sqrt{2}}\left(\ \sqrt{\mathrm{diag}(\mathbf{A})}\ \middle|\ \mathbf{0}\ \right) && \text{(A.52)}\\
&= \frac{1}{\sqrt{2}}\left(\ \sqrt{a_1}\ \ \mathbf{0}\ \right) && \text{(A.53)}\\
& && \text{(A.54)}
\end{aligned}$$

$$\begin{aligned}
\mathrm{Var}(\vec{x}_3) &= \hat{\mathbf{W}}_{3,\text{warm start}}\hat{\mathbf{W}}_{3,\text{warm start}}^T + \hat{\mathbf{\Psi}}_{3,\text{warm start}} && \text{(A.55)}\\
&= \frac{1}{2}\left(\ \sqrt{a_1}\ \ \mathbf{0}\ \right)\begin{pmatrix}\sqrt{a_1}\\ \mathbf{0}\end{pmatrix} + \frac{1}{2}(a_1) && \text{(A.56)}\\
&= a_1 && \text{(A.57)}\\
&= \mathbf{A} && \text{(A.58)}
\end{aligned}$$