UNIVERSITY OF CALIFORNIA

Los Angeles

# Uses for High-Throughput Platforms and Big Data in Engineering and Learning Biological Systems

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy
in Chemical Engineering

by

Lemuel M Soh

2018

ABSTRACT OF THE DISSERTATION

# Uses for High-Throughput Platforms and Big Data in Engineering and Learning Biological Systems

by

Lemuel M Soh
Doctor of Philosophy in Chemical Engineering
University of California, Los Angeles, 2018
Professor James C. Liao, Committee Chair

Despite immense growth in our biological knowledge over the past decades, purely knowledge-based rational approaches to metabolic engineering, protein engineering, and cancer prognosis have showed limited success. Instead, tools such as directed evolution and machine learning have greatly accelerated the pace of engineering and learning biological systems in the face of incomplete information. In this work, existing tools to engineer enzymes and shed light on the biochemical basis of cancer prognosis were utilized and built upon. In the first section, the focus is on keto acid decarboxylase (Kdc), a key enzyme in producing keto acid derived higher alcohols such as isobutanol. Kdc has no highly active yet thermostable variant in nature. The only reported Kdc activity is 2 orders of magnitude less active than the most active Kdc's found in

mesophiles. Therefore, isobutanol production temperature is limited by the thermostability of mesophilic Kdc enzyme variants. By configuring a high-throughput platform to parallelize the task of applying our directed evolution scheme on enzyme variants, thermostable 2-ketoisovalerate decarboxylase (Kivd) variants were developed. The top variants were recombined and further computationally directed protein design was applied to improve thermostability. Compared to wild-type Kivd, the final thermostable variant has 10.5-fold increased residual activity after 1h preincubation at 60°C, a 13°C increase in melting temperature and an over 4-fold increase in half-life at 60°C.

In the next section, the focus is on the relationship between current histopathology-based prognostic factors for endometrial cancer and their molecular features. Such information could speed progress on a revised classification system that may provide more accurate prognoses. Starting from predefined biochemical relationships, machine learning classifiers incorporated into a heuristic search strategy were used to identify small gene sets consisting of 3 genes from an endometrial cancer mRNA expression dataset that could predict prognostic factors. Cross-validated prediction accuracies obtained are 80% for overall survival at 5 years, 78% for progression-free survival at 5 years, 77% for European Society for Medical Oncology risk classification, 82% for histological grade, and 91% for histology type among high grade tumors. Predictive accuracy was evaluated on approximately 1.6 to 2 million two-gene and three-gene sets across all five prognostic factors. A statistically significant difference in overall survival and progression-free survival was identified when the most predictive gene sets were

used to separate patient groups in a Kaplan-Meier survival analysis. These small non-canonical gene sets are expected to reveal the underlying endometrial cancer biochemistry and could serve as candidate biomarkers with further investigation and clinical validation. The methods, results and discussion contained in this work contributes to the growing number of uses for high-throughput platforms and big data sets in engineering and learning biological systems.

The dissertation of Lemuel M Soh is approved.

Yi Tang

Robert P. Gunsalus

Jonathan W. Said

James C. Liao, Committee Chair

University of California, Los Angeles

2018

*This dissertation is dedicated to my family*

*Dad and Mom (Francis and Fiona)*

*Brother (Levin)*

*Diana*

*Dale and Helen King*

*Brothers and Sisters in Los Angeles, Houston, New York, Perth and Singapore*

TABLE OF CONTENTS

## LIST OF FIGURES

# LIST OF TABLES

# ACKNOWLEDGEMENTS

I wish to thank Prof. James Liao, my Ph.D. adviser for his invaluable support and timely advice in pushing me to my best and making this dissertation a reality. Many thanks to my Ph.D. committee, especially Dr. Jonathan Said, who went the extra mile in moving my biochemistry-guided search one step closer to clinical relevance. Prof. Richard Korf was instrumental in my learning on heuristic search and for guiding my thinking on its application in cancer. Prof. Robert Damoiseaux contributed tremendously to my practical learning and design of high-throughput experiments.

This research was supported by the DOE BioEnergy Science Center (BESC). This material is based upon research performed in a renovated collaboratory by the National Science Foundation under Grant No. 0963183, which is an award funded under the American Recovery and Reinvestment Act of 2009 (ARRA).

Chapter 2 was originally published under the same title in ACS Synthetic Biology (Soh et al., 2017). The design and execution of most experimental work done by Lemuel Soh. Wai Shun Mak performed the computationally directed protein design. Dr. Paul Lin and Dr. Luo Mi helped in technical advice and preparation of the manuscript. Frederic Chen performed all experiments related to circular dichroism spectroscopy (CD) and helped in writing the section describing CD. Dr. Robert Damoiseaux helped in design of the high-throughput screening. Prof. Justin Siegel supervised Wai Shun Mak, and Prof. James

Liao supervised Lemuel Soh. Ryan Wi, an undergraduate researcher, helped with enzyme purification.

Chapter 3 contains unpublished material and is confidential. The manuscript is in preparation under the same title for submission. Lemuel Soh prepared the dataset, implemented machine learning classifiers, beam search, and evaluated predictive accuracy of all gene sets using parallel processing. Stephen Tran performed the enrichment analysis of biochemical interactions and the Kaplan-Meier survival analysis. Both Lemuel Soh and Stephen Tran contributed to the preparation of the manuscript. Prof. James Liao supervised Lemuel Soh.

I wish to recognize and thank many members of Liao lab who have helped me in my journey. Luo Mi has been a friend, mentor and squash buddy. His contributions are beyond quantification but scientific discussions, an open heart, and a generous spirit are a huge part. Paul and Tony were mentors whom I am grateful to for many big and small things, from plasmid design to finding where the protein purification equipment went. Charlie, with his friendly and welcoming personality made the lab feel like home. Shanshan always brought a smile to my face with her gentle spirit. The undergraduates who worked with me were amazing. Ryan Wi, Justin Trinh, Abraxa Lee, Jonathan Mao and Evan Lee were bright, diligent and always took extra effort to execute experiments successfully.

There are many dear friends who made my time at UCLA and Los Angeles truly special. Stephen Tran, Shawn Tan, Grace Tang, Helen Durand and Amy Morioka made mornings at the botanical gardens refreshing with our devotionals together. Not enough gratitude can be extended to Stephen for making Los Angeles my home in welcoming me to his family, sharing his lab bench, teaching me tennis, and being available for all that Life brings. Jim Wang, Anne Su, Alan Kuei and Aunty Elizabeth Shen have shared many memorable Friday evenings at my apartment feasting on dinner fellowship and studying the Word of Life together. Patrick Ho was there from the beginning of UCLA's engineering orientation to the end of my final oral exam. Alhambra Christian Fellowship has fed me both physically and spiritually in more ways I can remember.

Dale and Helen King provided an oasis in the desert by opening their Arizona home to my frequent visits. Diana is special in my heart for her tender Love and encouragement through every step of Life's narrow way. These brothers and sisters are to me as family.

Finally, and most importantly, I wish to thank my family who has always supported me. In the countless hours of phone calls and prayers while I did my laundry, I received much Love and encouragement that is forever treasured. Without you, this would not be possible. Thank you.

LEMUEL M SOH

VITA

## EDUCATION

**M.S. in Bioengineering**

**B.S. in Bioengineering and B.A. in Economics**

Rice University, Houston, TX | Aug 2009 to May 2013

## RESEARCH EXPERIENCE

**Metabolic Engineering and Synthetic Biology Laboratory**

Ph.D. candidate with Principal Investigator Prof. James Liao

UCLA, Los Angeles, CA | Sep 2013 - Present

- Engineered thermostable keto acid decarboxylase via directed evolution and computational design
- Mentored and managed 4 undergraduates in experimental research
- Pioneered hypotheses-driven approaches in data-driven context
- Engineered heuristic search with machine-learning based metrics to identify candidate gynecological cancer biomarkers consisting of 4 to 6 genes

**Bridgeway Capital Management, Inc.**

Investment Management Quantitative Research Analyst

Houston, TX | May to Sep 2013

- Developed alpha-generating strategies based on corporate action signals
- Deployed models of created strategies that performed well over past 5 years
- Competitive analysis of quantitative strategies by other investment managers
- Studied ways to mitigate price impact of executed trades

**Rice University Senior Design**

Independent Design Team Member of novel pump device in resource limited settings

Houston, TX | Aug 2012 to May 2013

- Formulated idea, designed & built a highly accurate drug delivery device AutoSyp
- Developed novel mechanical dual-pawl system for drug delivery, programmed micro-controller to check flow rate, produced highly accurate schematics and engineered entire device in machine shop.
- Received a USPTO patent US20140303559 A1 for original design
- Published findings on delivery of intravenous thrombolytics
- 3rd place, Design of Medical Devices Conference showcase, Minnesota, 2013
- AutoSyp devices currently deployed in Malawi hospitals, and is in scale-up phase for further deployment

**Rice Undergraduate Scholars Program**
Economics independent research project with advisor Prof. Malcolm Gillis
Houston, TX | Aug 2011 to May 2012

- Studied the provident fund model as a viable alternative to tax-financed pension systems in providing retirement income
- Case studies of Switzerland, Chile and Singapore to illustrate the successes of the provident fund model and to suggest possible hybrid models

**Metabolic Engineering and Systems Biotechnology Laboratory**
Research Assistant with Principal Investigator Prof. Ka-Yiu San
Houston, TX | June 2012 to May 2013

- Design simulations of *E. coli* fatty acid biosynthesis pathways
- Engineered process to extract kinetic information from online databases and used the information in building kinetic models

**Modern Optical Instrumentation and Bio-imaging Laboratory**
Research Assistant with Principal Investigator Prof. Tomasz S. Tkaczyk
Houston, TX | Aug 2011 to Dec 2011

- Characterized the electromechanical characteristics of a multi-angle mirror array (micro-electromechanical system)
- Designed a testing procedure for application of voltage to micro mirrors and imaging using interferometer

## LEADERSHIP EXPERIENCE

**Rice University Student Association**
Treasurer
Houston, TX | Aug 2010 to May 2011

- Planned, oversaw and reported on the Student Association senate budget
- Managed finances of the Student Association

**Rice University Honor Council**

Class representative for freshmen and sophomores

Houston, TX | Aug 2009 to May 2011

- Performed investigations and deliberated on hearings
- Evaluated and initiated changes to consensus penalty structure

# 1 Introduction

"To know what you know and what you do not know, that is true knowledge."

Confucius

## 1.1 Motivation

The understanding and subsequent engineering of biological systems have improved human quality and standard of living over decades. This application of increased biological knowledge has had a profound impact on two key areas, namely in biotechnology and medicine.

In biotechnology, though fermentation has been used in beverage production for thousands of years, it was not until Louis Pasteur's attribution of fermentation to microorganisms in the 1850s, and Eduard Buechner's discovery of enzymes in 1897 that advances in microbiology and biochemistry started building momentum. Using the newly-gained understanding, applications ranging from food preservation to biomanufacturing of fuels, chemicals and pharmaceuticals started to take root in human society. Even though biomanufacturing have transformed pharmaceuticals and specialty chemical production, commodity chemicals and biofuels continue to languish because of seemingly intrinsic properties of biology. These include the lack of enzymes with suitable properties, and difficulties with heterologous enzyme expression.

In medicine, the seminal elucidation of the structure of deoxyribonucleic acid by James Watson and Francis Crick in 1953 opened the door to personalized medicine. As a disease driven primarily by somatic mutations, the treatment of cancer has truly revolutionized because of new molecular understanding. Subtyping with molecular biomarkers for diagnosis, prognosis and treatment of patients have dramatically improved survival when the biomarkers were truly relevant to the underlying biology of the tumor. However, even with the advent of high-throughput sequencing platforms and corresponding bioinformatic analytic tools, the vast majority of candidate biomarkers are never validated or used in the clinic.

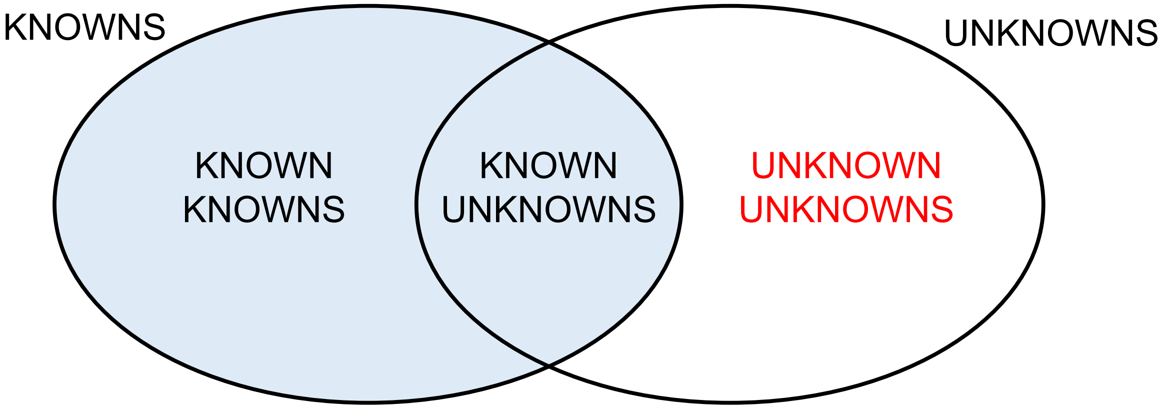


*Figure 1-1: The Utility of Known Unknowns*

## 1.2 Objectives

The overarching goal of my research is to utilize high-throughput platforms and big data in innovative ways that capitalize on our knowledge of biological systems (known knowns) and their known, though not fully understood, complexities (known unknowns) (Figure 1-1). The utility of known unknowns is useful in helping to frame a biological problem. A principle applied in mitigating the inevitable effects of unknown or unmeasured biological phenomenon (unknown unknowns) is to expand the sample size and throughput of experiments. Due to the law of large numbers, the effects of any unknown phenomenon might be averaged out or possibly even discovered if they are significant. My overarching goal is specifically applied in engineering enzymes and identifying the most promising cancer biomarkers.

One known difficulty with rational design of proteins is the lack of complete understanding as to how DNA coding sequence affects enzymatic properties, such as stability, kinetic rates, and substrate specificity. Cancer is thought to be driven by molecular aberrations, but the complexity of biochemical reactions and regulation make it difficult to untangle. Due to this complexity, it becomes difficult to diagnose, prognose and predict the effect of treatments upon patients. Compounding this difficulty, data-driven approaches in cancer biomarker discovery are often divorced from hypotheses of domain experts. This results in many data-driven discoveries that are not applicable in the clinic because the actual practice of medical procedures and other physician domain knowledge were not properly accounted for.

My Ph.D. dissertation contributes by developing a thermostable enzyme using high-throughput directed evolution and by incorporating hypothesis-driven approaches into data-driven tools to identify candidate cancer biomarkers that are biochemically related. By capitalizing on known knowns and known unknowns, it is expected that progress in engineering and learning biological systems without complete knowledge can be accelerated.

## 1.3 Dissertation Overview

Chapter 2 will discuss the engineering of a thermostable Keto Acid Decarboxylase by configuring a high-throughput platform for directed evolution and using computationally directed protein design. Compared to the wild-type, this Keto Acid Decarboxylase has 10.5-fold increased residual activity after 1h preincubation at 60 °C, a 13 °C increase in melting temperature and over 4-fold increase in half-life at 60 °C.

Chapter 3 will discuss the incorporation of hypothesis-driven approaches into data-driven tools, and how searching over predefined biochemical relationships led to the identification of candidate multi-gene endometrial cancer biomarkers that are biochemically related. Such candidate biomarkers are expected to be useful for understanding endometrial cancer biochemistry and for clinical prognosis after further validation. Chapter 5 contains the entire bibliography for this dissertation.

# 2 Engineering a Thermostable Keto Acid Decarboxylase Using Directed Evolution and Computationally Directed Protein Design

This chapter was originally published under the same title in ACS Synthetic Biology (Soh et al., 2017). The design and execution of most experimental work done by Lemuel Soh. Wai Shun Mak performed the computationally directed protein design. Dr. Paul Lin and Dr. Luo Mi helped in technical advice and preparation of the manuscript. Frederic Chen performed all experiments related to circular dichroism spectroscopy (CD) and helped in writing the section describing CD. Dr. Robert Damoiseaux helped in design of the high-throughput screening. Prof. Justin Siegel supervised Wai Shun Mak, and Prof. James Liao supervised Lemuel Soh. Ryan Wi, an undergraduate researcher, helped with enzyme purification.

## 2.1 Abstract

Keto acid decarboxylase (Kdc) is a key enzyme in producing keto acid derived higher alcohols, like isobutanol. The most active Kdc's are found in mesophiles; the only reported Kdc activity in thermophiles is 2 orders of magnitude less active. Therefore, the thermostability of mesophilic Kdc limits isobutanol production temperature. Here, we report development of a thermostable 2-ketoisovalerate decarboxylase (Kivd) with 10.5-fold increased residual activity after 1h preincubation at 60 °C. Starting with mesophilic Lactococcus lactis Kivd, a library was generated using random mutagenesis and approximately 8,000 independent variants were screened. The top single-mutation variants were recombined. To further improve thermostability, 16 designs built using Rosetta Comparative Modeling were screened and the most active was recombined to form our best variant, LLM4. Compared to wild-type Kivd, a 13 °C increase in melting temperature and over 4-fold increase in half-life at 60 °C were observed. LLM4 will be useful for keto acid derived alcohol production in lignocellulosic thermophiles.

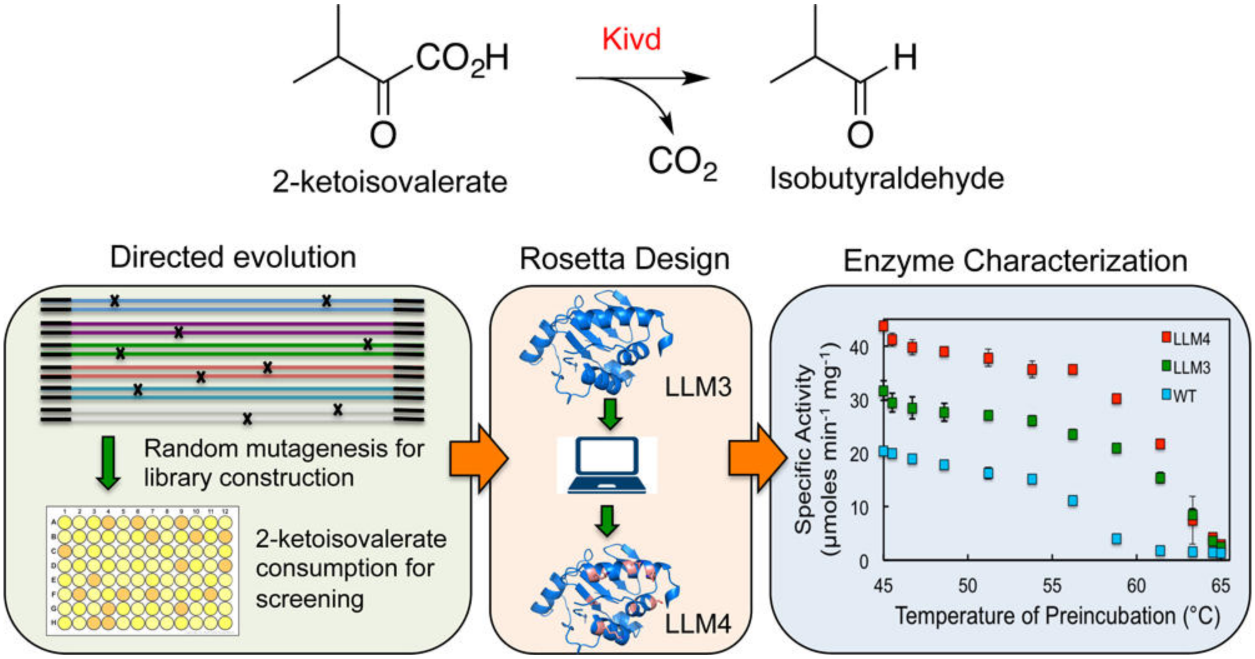

Kivd
$CO_2H$
$CO_2$
H
O
2-ketoisovalerate
Isobutyraldehyde
Directed evolution
Random mutagenesis for library construction
2-ketoisovalerate consumption for screening
Rosetta Design
LLM3
LLM4
Enzyme Characterization
Specific Activity (µmoles min⁻¹ mg⁻¹)
Temperature of Preincubation (°C)
LLM4
LLM3
WT
0
10
20
30
40
45
50
55
60
65

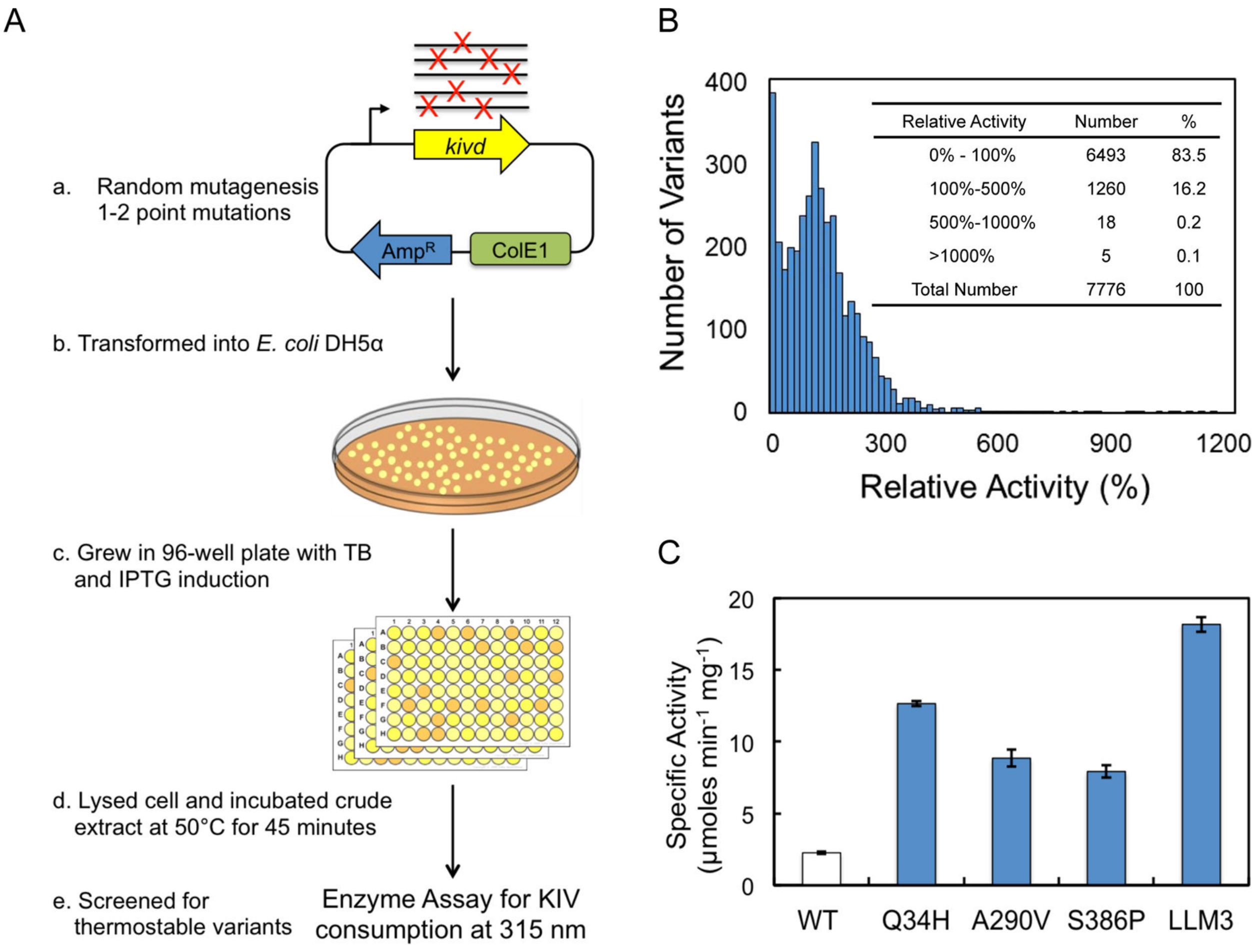
*Figure 2-1: Engineering a Thermostable Keto Acid Decarboxylase Graphical Abstract*

## 2.2 Main Text

Keto acid decarboxylases (Kdc) are an important group of enzymes crucial to the production of keto acid derived alcohols (Atsumi et al., 2008), including 1-propanol (Shen and Liao, 2008), 1-butanol (Shen and Liao, 2008), isobutyl alcohol (Higashide et al., 2011; Li et al., 2012; Lin et al., 2014, 2015; Smith et al., 2010), 2-methyl-1-butanol (Cann and Liao, 2008) and 3-methyl-1-butanol (Connor and Liao, 2008). Despite their importance, there are only two reported enzymes from thermophiles with Kdc activity, both of them homologues of the acetolactate synthase (Als) enzyme in different *Geobacillus* species (Lin et al., 2014). Moreover, the activity of the only characterized, Gtng_0348 from *Geobacillus thermodenitrificans* is about 2 orders of magnitude lower than that of the most active Kdc from mesophile *Lactococcus lactis*, even after a heat treatment at 50 °C for both enzymes (Lin et al., 2014). This *L. lactis* Kdc is a catabolic 2-ketoisovalerate decarboxylase (Kivd) whose main activity is to catalyze the decarboxylation of 2-ketoisovalerate to isobutyr- aldehyde, a crucial enzymatic step needed to divert flux from the valine biosynthesis pathway to isobutyl alcohol production. Kivd shares some structural similarity to pyruvate decarboxylase (Pdc), which also lacks a satisfactory thermostable homologue for heterologous expression at elevated temperatures (Raj et al., 2002; Thompson et al., 2008).

Kivd has been overexpressed to produce isobutyl alcohol in numerous microorganisms (Atsumi et al., 2008; Higashide et al., 2011; Li et al., 2012; Lin et al., 2014, 2015; Smith et al., 2010), most recently at 50 °C in *Clostridium thermocellum* (Lin et al., 2015), a lignocellulolytic thermophile of potential industrial importance. Production at an elevated temperature is desirable when using lignocellulosic feedstocks, as higher temperatures promote cellulose deconstruction (Balsan et al., 2012; Ko et al., 2010; Lee et al., 2008; Liu and Xia, 2006; Schwarz, 2001). Therefore, engineering a highly active and thermostable Kivd is a logical approach toward the further improvement of metabolically engineered isobutyl alcohol production in lignocellulolytic thermophiles.

To increase the thermostability of Kivd, a high-throughput screening process was developed as illustrated in Figure 2-2A. Briefly, we adapted a screening method based on absorbance changes at 315 nm that directly measured the consumption of 2-ketoisovalerate (KIV) (Gocke et al., 2007) to high-throughput automation. The stability of KIV at elevated temperatures was confirmed to allow screening to occur at 50 °C, and proper controls were used to ensure that the change of absorbance is solely due to Kivd activity. For effective automation, the protocol was made simple and robust. To this end, a few criteria had to be fulfilled. First, an appropriate and convenient cell lysis method had to be determined that would ensure quality readout without interference from cell debris. Second, the reaction time of the enzyme assay had to be optimized to balance throughput and accuracy of enzyme activity measured. An end point enzyme assay was chosen instead of measuring a time course to further increase the throughput.



| Relative Activity | Number | % |
|---|---|---|
| 0% - 100% | 6493 | 83.5 |
| 100%-500% | 1260 | 16.2 |
| 500%-1000% | 18 | 0.2 |
| >1000% | 5 | 0.1 |
| Total Number | 7776 | 100 |

*Figure 2-2: Identification of thermostable Kivd variants using high throughput screening*

*(A) summary of mutagenesis and screening; (B) distribution of relative crude extract activities compared to wild-type Kivd after 45 min heat treatment at 50 °C. (Variants with no measurable activity not plotted). (C) Specific activity of purified single mutation variants and LLM3, the recombination of all three single mutations, measured by NADH- coupled assay at 50 °C after 1 h of preincubation at 60 °C. The error bars represent the standard deviation of three independent repeats.*

It is worth noting that during the development of a high throughput screen, the consistency of automated processing should be properly examined. For instance, one challenge observed during screen development was the significant amount of cell debris and precipitate formed after lysis and heat treatment at 50 °C. Despite the availability of more laborious methods such as centrifugation to pellet the debris, it was found that simply allowing the debris to settle and setting the liquid handler to draw from the top at the slowest speed could consistently ensure that no debris is transferred to the final assay solution. The simplicity of the screen is also of practical importance. For example, even though there are other ways to incubate the cell lysate that could be more precise, air incubation was found to be most convenient, and the results were without adverse effects from potential heat transfer limitations.

A random mutagenesis library of Kivd was made using error-prone PCR with only 1 to 2 point substitutions introduced per kivd gene. Such a low mutation rate was chosen because most mutations are deleterious and multiple mutations often completely inactivate the enzyme (Romero and Arnold, 2009). In so doing, the percentage of active enzymes within our library is maximized.

Independent transformants were picked into 96-well plates, induced for protein expression, and subsequently lysed by detergent. The crude cell extract was heat challenged at 50 °C for 45 min before starting the enzyme assay. The screening temperature of 50 °C and incubation time were tested iteratively to ensure that the wild-type Kivd enzyme (positive control) had activity that could be detected by our screen consistently. The change in KIV characteristic UV absorbance before and after the enzyme assay, normalized to initial cell density, was used as an indicator for Kivd activity retained after heat treatment.

Following the aforementioned screening protocol, approximately 2000 variants were picked and screened in each round, with a total of 4 rounds. The activity of these variants, as a percentage of the wild type, is plotted in Figure 2-2B. Two peaks were clearly evident in the histogram. First, there were a large number of variants (about 50%) that did not have any measurable activity, either due to failure to fold or complete loss of thermal stability at 50 °C. Second, many of the variants had activity clustered around the wild type. This cluster most likely consists of Kivd variants that have similar thermal stability as the wild type but some difference in activity. With almost 8000 independent variants screened, about 70% of the single substitutions possible in Kivd would likely have been covered. The screening results mostly corroborate earlier protein evolution studies that show the vast majority of single amino acid mutations are deleterious or at best neutral, with only 0.01 to 1% being beneficial (Aharoni et al., 2005; Bloom et al., 2006; Drummond et al., 2005; Guo et al., 2004; Romero and Arnold, 2009).

From that library of about 8000 independent transformants, 12 of the most active variants were selected and sequenced. Some of the variants had mutations that did not code for amino acid changes, and their increased activity is likely due to an increase in expression of Kivd in E. coli. These were discarded as the thermostable Kivd variant would eventually be used in other thermophiles such as C. thermocellum. From the 12 mutants sequenced, 3 single point substitutions (Q34H, A290V, and S386P) were identified as possibly increasing thermostability. A290V appeared multiple times with the same codon change in independent variants through different rounds of mutagenesis and screening.

Those single-mutation Kivd variants were purified and characterized. Although it is simple to directly measure the consumption of KIV at 315 nm, this assay is not ideal for further detailed characterization due to its low sensitivity (extinction coefficient of KIV is 26.8 $M^{-1}$ $cm^{-1}$ at 315 nm). Therefore, a NADH-coupled enzyme assay (Lin et al., 2014) which is much more sensitive (extinction coefficient of NADH is 6220 $M^{-1}$ $cm^{-1}$ at 340 nm) was chosen for all characterization work of the purified Kivd variants. A 60 °C 1 h heat preincubation was performed in the thermocycler, and activity was measured at 50 °C using the NADH-coupled enzyme assay. The thermal stability effect of all three mutations was confirmed. Compared to the wild type, Q34H had a 5.5-fold increase, A290V had a 4- fold increase, and S386P had a 3.5-fold increase in specific activity after a 1 h preincubation at 60 °C (Figure 2-2C).

Since the effects of thermostable mutations are likely additive (Wu and Arnold, 2013), these three stabilizing mutations were recombined. The recombined kivd variant was named LLM3. The specific activity of LLM3 at 60 °C is 8-fold higher than that of the wild type (Figure 2-2C). The $T_{50}$ of LLM3, which is the temperature at which the enzyme loses half of its maximum activity after an hour incubation, was measured to be 60.3 °C (Figure 2-4A). This represents an increase over the wild-type $T_{50}$ of 56.1 °C by 4.2 °C (Figure 2-4A). The $T_{50}$ of wild-type Kivd measured in this study is similar to that previously reported (Lin et al., 2014). This result indicates that LLM3 has substantial activity at the elevated temperature of 60 °C.

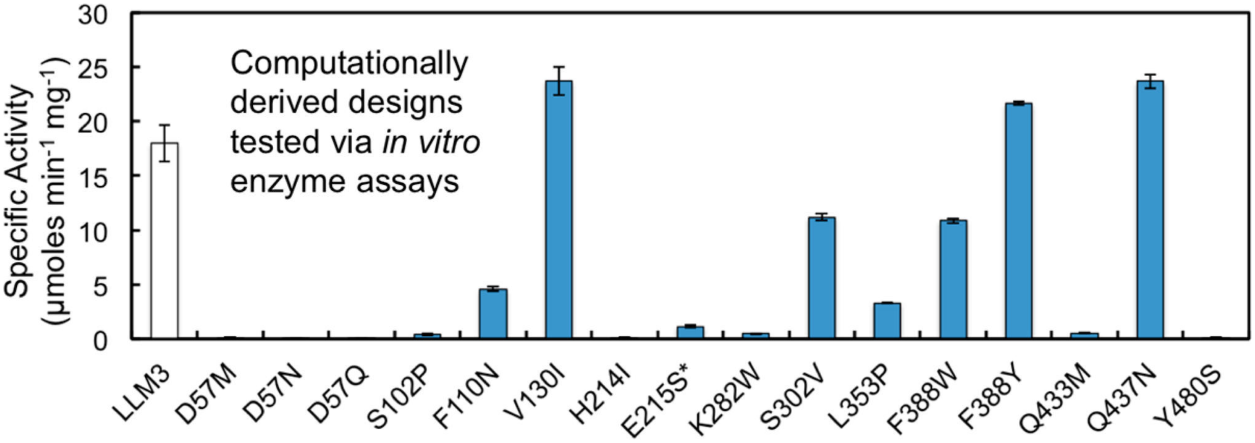


*Figure 2-3: Sixteen mutations derived using Rosetta design were recombined*

*Sixteen mutations derived using Rosetta design were recombined individually with LLM3, and the enzymes were purified. Specific activity was measured by NADH-coupled assay at 50 °C after 1 h preincubation at 60 °C. V130I, F388Y, and Q437N were found to have increased activity over LLM3. *E215S had an added mutation I305P as part of the computational design. LLM3_V130I has the highest activity after preincubation at 60 °C for 1 h and is renamed LLM4.*

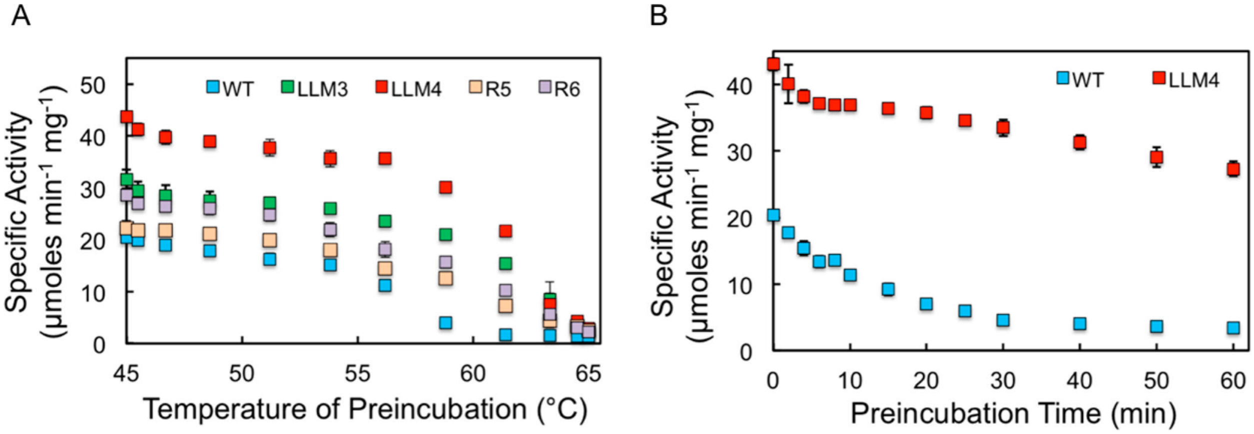


*Figure 2-4: Characterization of purified thermostable Kivd variants*

*(A) Plot of Kivd specific activity after 1 h preincubation at various temperatures. LLM4, best variant after recombining V130I (computationally derived) with LLM3; LLM3, best variant after recombination of Q34H, A290V, and S386P mutations (directed evolution); R6, mutations V130I, Q437N, and F388Y were recombined with LLM3 to form a 6-mutation Kivd variant; R5, mutations V130I and Q437N were recombined with LLM3 to form a 5-mutation Kivd variant; WT, wild-type Kivd. (B) Plot of Kivd LLM4 and WT specific activity at 60 °C heat preincubation for various times. Specific activities measured by NADH consumption at 50 °C. The error bars represent the standard deviation of three independent repeats.*

In addition to high-throughput screening of Kivd mutant libraries, computationally directed protein design of thermo-stabilizing mutations was performed using Foldit (Eiben et al., 2012), a graphical-user-interface to the Rosetta Molecular Modeling Suite. Mutations that improved packing or introduced new hydrogen bonds outside of the enzyme active site were explored based on previous observations that such mutations can significantly increase thermostability (Korkegian, 2005). Sixteen mutants were designed using this approach and experimentally characterized after recombination with LLM3 (Figure 2-3).

While the majority of the mutations significantly decrease enzyme activity, mutations V130I, Q437N, and F388Y increase the activity of LLM3 at 60 °C by 32%, 31%, and 20%, respectively (Figure 2-3). The mutation V130I is predicted to improve core hydrophobic packing with the additional methyl group on isoleucine. The Q437N mutation lies at the Kivd homodimeric interface, and its amide carbon is predicted to be 4.5 Å from the same carbon atom on its dimeric partner. These two partnering Q437N residues are predicted to stabilize the interface by forming a hydrogen bond between their carbonyl oxygen and amide nitrogen. The mutation F388Y introduces a hydroxyl group, which is predicted to satisfy a previously unsatisfied hydrogen bond on the carbonyl oxygen of the L254 backbone. Mutations V130I and Q437N were recombined with LLM3 to form a 5-mutation Kivd variant named R5. Mutations V130I, Q437N, and F388Y were recombined with LLM3 to form a 6-mutation Kivd variant named R6. Kinetic characterization of variant R5 and R6 shows that the effects of these mutations are not additive. They exhibit similar or lower activity and thermostability compared to LLM3 (Figure 2-4A). The origin of these nonadditive effects is unclear as each of these three mutations are 15−30 Å apart from each other. The variant LLM3_V130I (renamed LLM4) has the highest thermostability of all the mutants tested, with its $T_{50}$ being 4.3 °C higher than native Kivd.

Even though the $T_{50}$ of LLM4 did not change significantly from that of LLM3, the large increase in specific activity of LLM4 at 60 °C is a functional advantage for metabolic engineering (Figures 2-3 and Figure 2-4A). The half-life of LLM4 measured at 60 °C was more than 1 h, which is more than a 4-fold increase over the wild type's half-life of 14.0 min (Figure 2-4B). The specific activity of LLM4 at 60 °C is 10.5-fold higher than that of the wild type.

To determine the potential mechanisms by which the mutations identified via directed evolution (Q34H, A290 V and S386P) conferred thermostability, a model of the Kivd protein was acquired via homology modeling (Biasini et al., 2014) by using the crystal structure of L. lactis branched-chain keto acid decarboxylase (Berthold et al., 2007) (PDB ID: 2VBF) which shares an 88% sequence identity with the wild-type Kivd. The acquired Kivd model was visualized and further studied using Pymol software (Figure 2-5) (Schrodinger LLC, 2010). The three amino acid mutations (Q34H, A290V, and S386P) identified via directed evolution are all located near the protein surface and far away from enzyme active site.

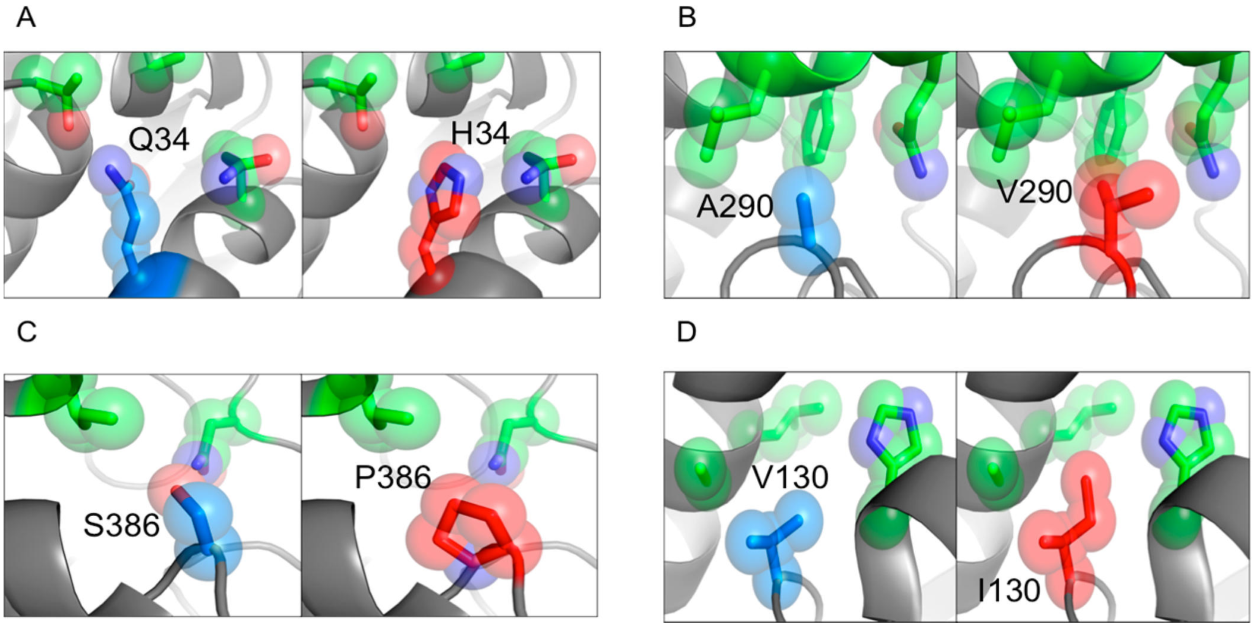


*Figure 2-5: Local environments of beneficial Kivd mutations*

*(A) Q34H, (B) A290V, (C) S386P, and (D) V130I with the native and mutated residue depicted in blue and red, respectively. Mutations Q34H, A290V, and S386P were identified through directed evolution and mutation; V130I was identified via computationally directed protein design. The mutations A290V and V130I are predicted to improve hydrophobic packing of the local environment.*

Besides homology modeling, circular dichroism spectroscopy (CD) of both wild-type Kivd and LLM4 were measured to further analyze the combined structural effects of various point mutations (Figure 2-6). CD spectra results indicate that wild-type Kivd and LLM4 possess similar structures at 25 °C, which is consistent with the fact that both WT and LLM4 are active at physiological conditions (Figure 2-11). Furthermore, results are similar to or without thiamine pyrophosphate (TPP), a crucial cofactor for the enzyme (Plaza et al., 2004), for both WT and LLM4 variant.

Moreover, monitoring ellipticity changes while heating proteins enables the determination of the unfolding temperature $T_m$, defined as the temperature at which 50% of the protein is denatured. The wild-type Kivd without TPP was found to have a lower $T_m$ of 55.8 °C compared to the $T_m$ of wild type with TPP at 61.3 °C. This suggests that wild-type Kivd without TPP is structurally more unstable, which is consistent with the fact that TPP is responsible for proper assembly of the active tetramer in the Pdc family (Furey et al.; Koga et al., 1992; Plaza et al., 2004). In comparison, the $T_m$ values of LLM4 were both higher at 73.8 and 74.6 °C, without and with TPP, respectively. Thus, LLM4 had a significant 13.3 °C improvement in $T_m$ when TPP is included in the buffer (Figure 2-6). Intriguingly, the absence of TPP did not negatively affect the $T_m$ of LLM4 by much, suggesting that the Kivd variant may have stabilized the overall structure of the enzyme so that the active site function would not be affected adversely at higher temperature. Noticeably, all $T_m$ were higher than $T_{50}$ by around 5 to 15 °C, which is typical as enzymes normally first lose their function before degrading (Tian et al., 2010).

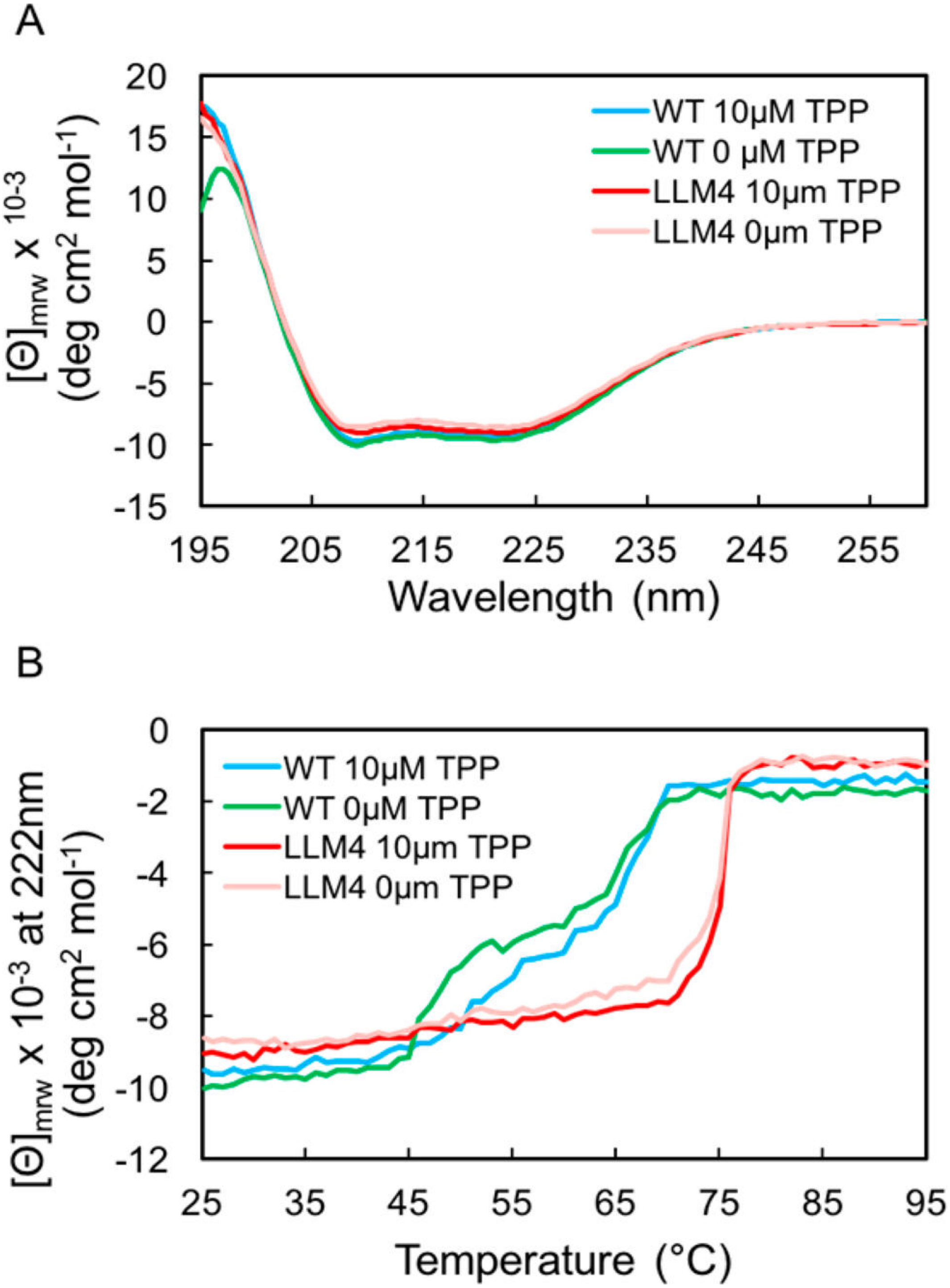


*Figure 2-6: Structure and thermal stability analysis of Kivd WT and LLM4 with and without TPP*

*(A) CD spectra at 25 °C of WT and LLM4; (B) thermal unfolding profiles of WT and LLM4 monitored by change of mean residual ellipticity as a function of temperature at 222 nm.*

Further characterization was also carried out on LLM4 to determine its suitability for other functions. First, kinetic parameters at 50 °C were characterized for LLM4 and wild- type Kivd. Compared to the wild type, LLM4 had a similar $K_m$ and a $k_{cat}$ 2-fold greater (Table 2-1). The $k_{cat}/K_m$ value of LLM4 is 24.38, compared to the 11.15 value of the wild-type enzyme (Table 2-1). Thus, LLM4 has similar affinity but a much higher turnover rate than the wild-type Kivd. Furthermore, it is observed that LLM4 has a higher specific activity after heat incubation at temperatures ranging from 45 to 65 °C (Figure 2-4A). LLM4 specific activities on other substrates similar to 2- ketoisovalerate in structure were also measured in this study. LLM4 had an observably higher specific activity for all the substrates, especially pyruvate, when measured at 50 °C (Figure 2-7). Thus, LLM4 is thermostable at 60 °C while improving catalytic activity on 2-ketoisovalerate or other substrates.

Despite screening at 50 °C for more active variants, we found variants that were substantially more thermostable than the wild type at 60 °C. In fact, a higher specific activity than the wild type was observed at all tested temperatures higher than 45 °C. Besides being useful for isobutyl alcohol production in thermophiles such as C. thermocellum, it is expected that Kivd LLM4 could be useful for the production of higher alcohols in a variety of organisms (Atsumi et al., 2008; Higashide et al., 2011; Smith et al., 2010) and also in cell-free systems (Guterl et al., 2012). The absorbance-based screen used for high throughput evolution is not only simple but can also be modified for the directed evolution of other keto-acid decarboxylases.

*Table 2-1: Kinetic Parameters of the Wild Type and Mutant 2- Ketoisovalerate Decarboxylase Measured at 50°C*

| Kivd Variant | $K_m$ (mM) | $k_{cat}$ ($s^{-1}$) | $k_{cat}$ / $K_m$ ($mM^{-1}$ $s^{-1}$) |
|---|---|---|---|
| LLM4 | 1.55 ± 0.03 | 37.89 ± 0.22 | 24.38 |
| WT | 1.60 ± 0.03 | 17.90 ± 0.28 | 11.15 |

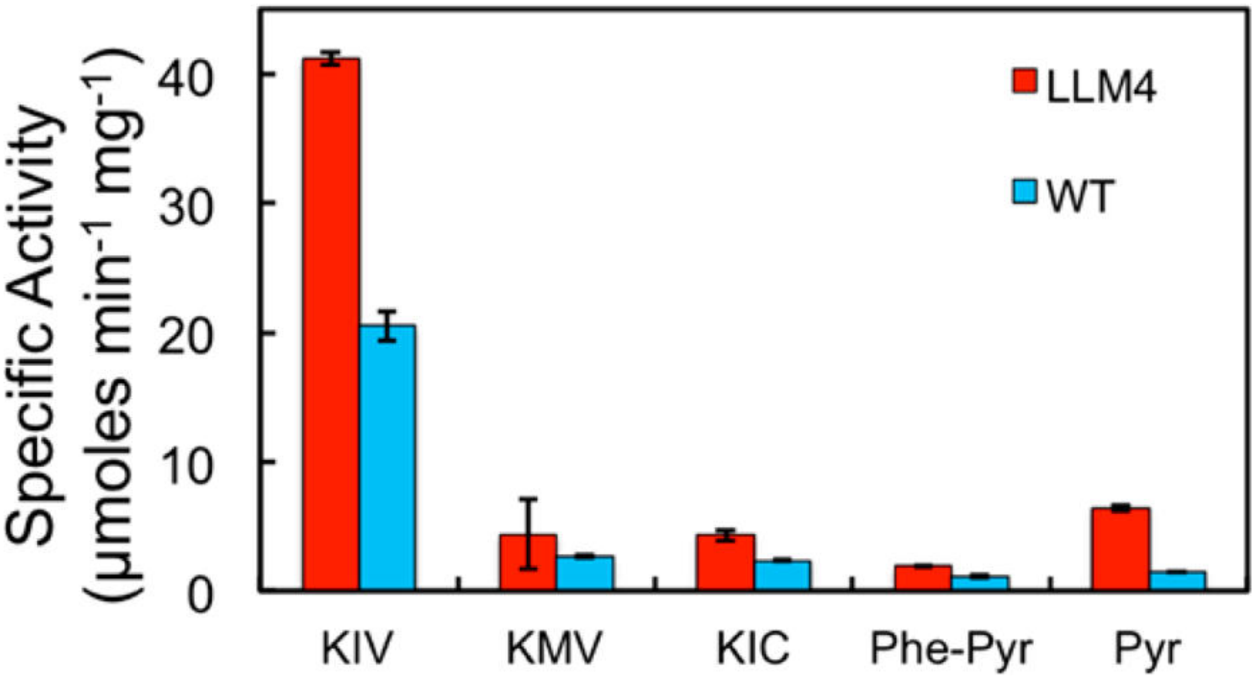


*Figure 2-7: Specific activity of purified LLM4 to different keto acid substrates*

*KIV, 2-ketoisovalerate; KMV, 2-keto-3-methyl-valerate; KIC, 2-keto-4-methyl-pentanoate; Phe-Pyr, phenylpyruvate; Pyr, pyruvate. Specific activities measured by NADH consumption at 50°C. The error bars represent the standard deviation of three independent repeats.*

# 2.3 Methods

## 2.3.1 Bacterial Strains and Plasmid Construction

All plasmid construction and enzyme expression was done using E. coli NEB 5-alpha strain (New England Biolabs, Ipswich, MA). All plasmids were constructed by DNA assembly techniques (Gibson et al., 2009). Oligonucleotides were purchased from IDT technologies (San Diego, CA). Both vector and inserts (target genes) were amplified by PCR using KOD Hot Start DNA polymerase (EMD Millipore, Billerica, MA). The PCR template was digested by DpnI digestion at 37 °C for 1 h (New England Biolabs). PCR products were purified by DNA clean and concentrator kit (Zymo Research, Irvine, CA). The vector and insert were mixed with Gibson Assembly Master Mix (New England Biolabs) and incubated at 50 °C for 1 h. The assembly product was then transformed to the E. coli strain mentioned above. Plasmids meant for Kivd library generation were extracted by miniprep (Qiagen, Hilden, Germany) and sent for DNA sequencing (Laragen, Culver City, CA). kivd (L. lactis) was amplified from the pSA65 plasmid (Atsumi et al., 2010) and a modified pQE9 vector was used to construct plasmid pLS02. All solid and liquid media used for growing strains with pLS02 were supplemented with 200 μg/mL of carbenicillin.

## 2.3.2 Library Construction

Error-prone PCR was performed on kivd using GeneMorph II random mutagenesis kit (Agilent Technologies, San Jose, CA). To achieve 1 to 2 mutations per kivd gene, 300 ng of the template gene was used with 25 cycles of PCR. The kivd random mutagenesis library was used to construct plasmid pLS02 as described above for the wild-type gene. Transformants were plated on Bioassay Q-trays (Molecular Devices, Sunnyvale, CA) with LB agar. Twelve independent transformants were randomly picked and sequenced to ensure that the desired mutation rate was achieved.

### 2.3.3 Cell Lysis for High Throughput Screening

Cells were grown up to an $OD_{600}$ of 0.6 to 0.8 in 96-well plates before being lysed. This $OD_{600}$ is recorded for normalization of the assay results. Despite the efficacy of beads and sonication, they were laborious to use in high throughput screening. Commerical detergent Bugbuster (no. 70921, EMD Millipore) was found to be simple and effective for screening the library. Instead of removing cell debris by centrifugation and resuspension, the process was simplified by adding 80 μL of Bugbuster of 3× Bugbuster reagent to a 150 μL cell culture and incubating for 20 min at room temperature. The cell lysate was then used in the crude extract enzyme assay described in the following section. The 3× Bugbuster reagent is made by diluting Bugbuster 10× Protein Extraction Reagent (no. 70921, EMD Millipore) to a 3× concentration with an appropriate amount of Milli-Q water.

### 2.3.4 KIV Absorbance-Based Assay

A simple Kivd assay (Gocke et al., 2007) was adapted for use in high throughput screening with crude extract cell lysates. Substrate 2-ketoisovalerate had an absorbance peak at 315 nm after a spectral scan on a DU- 800 spectrophotometer (Beckman Coulter, Brea, CA) and substrate disappearance at that wavelength was used as the basis of the screen. The other components, namely magnesium chloride, thiaminpyrophosphate (TPP), and sodium phosphate buffer required in our crude extract enzyme assay were also checked for overlapping absorbance. Only TPP had substantial absorbance in the UV range but this readout was stable at the assay conditions and thus was not a concern. A possibility for signal interference was further reduced by using a lens during screening that aliased lower wavelength UV.

The cell lysate obtained as described in the section Cell Lysis for High Throughput Screening was subject to a 45 min air incubation at 50 °C. The cell debris and precipitate were allowed to settle and only 30 µL of supernatant was transferred to a fresh UV-transparent plate (no. 655801, Greiner, Monroe, NC). A 170 µL sample of enzyme master mix was added to make a final concentration of 5 mM $MgCl_2$, 1.5 mM TPP, and 60 mM 2-ketoisovalerate in 50 mM pH 6.5 sodium phosphate buffer. $OD_{315}$ is measured before air incubating at 50 °C for 2 h. After 2 h, $OD_{315}$ is measured again to obtain the end-point enzyme assay reading. The enzyme thermostability score of each independent transformant is calculated by (beginning $OD_{315}$ − ending $OD_{315}$)/$OD_{595}$. The final results were normalized by $OD_{595}$ of the corresponding cell cultures to minimize the effect of different protein amounts.

### 2.3.5 High Throughput Screening

All screening was per- formed in the Molecular Screening Shared Resource, UCLA. All optical density (OD) measurements were performed on a Victor 3 V plate reader (PerkinElmer, Waltham, MA); 96-well plates (#3370, Corning, Corning, NY) were filled with 150 μL of terrific broth (TB) medium supplemented with 3% (v/v) glycerol. Single colonies were picked into plates using Genetix Qbot colony picker (Molecular Devices, Sunnyvale, CA). The plates were incubated at 37 °C for 5 h. Following this, a copy of all the picked colonies were made into 96-well low-profile plates (X6023, Molecular Devices) containing TB medium supplemented with 10% glycerol (v/v) on Genetix Qbot. This storage copy was grown overnight at 37 °C, covered with aluminum sealing film (no. 6569, Corning) and stored at −80 °C. After the copies were made, 10 μL of 16 mM IPTG was added to make a final concentration of 1 mM for protein induction in the initial plates of the picked colonies. Protein induction continued for 12 h at 37 °C. After this time, an automated ORCA arm (Beckman Coulter, Brea, CA) moved the plates between different stations for screening. The screen was scheduled and controlled using the Sami automation platform (Beckman Coulter, Brea, CA). Plates were transferred from the incubator to BioMek FX (Beckman Coulter, Brea, CA) and shaken for 1 min at 1000 rpm before $OD_{595}$ was measured. Cells were lysed by adding 80 μL Bugbuster of 3× Bugbuster reagent. For 20 min, the cells are lysed at room temperature and cell debris collects to the bottom of the well. A BioMek FX with 200 μL of AP 96 pipetting head (Beckman Coulter, Brea, CA) is used to transfer 30 μL of cell lysate to fresh UV-transparent 96-well plates (no. 655801, Greiner, Monroe, NC). 170 μL of enzyme master mix was added, and the crude extract Kivd enzyme assay is conducted as described above in the section KIV Absorbance-Based Kivd Assay.

### 2.3.6 Computational Simulations

The 3D model of Kivd used for rational design was built using the Rosetta Comparative Modeling protocol (Song et al., 2013). Fragments sets (Gront et al., 2011) and three-dimensional evolutionary constraints (Thompson and Baker, 2011) were generated using crystal structure PDB 2VBG. A total of 1000 models were generated, and the model with the lowest energy was used as the homology model of Kivd. Foldit was used to evaluate the effects of mutations on thermostability. The changes in rosetta energy of the mutated residue and the protein pose after repacking were used to select candidate mutations for experimental characterization.

### 2.3.7 Protein Purification

The highest activity variants obtained from the high throughput screen were selected and sequenced. Kivd variants that resulted in an amino acid substitution were purified. The plasmid construct pLS02 used for screening attaches a His-tag to the 5′ end of Kivd for enzyme purification. Wild-type Kivd and variants were purified as follows: NEB- 5alpha cells containing the pLS02 plasmid from screening were used. The respective cell lines were cultured in 200 mL of LB medium. After the cells reached mid-log phase, IPTG was added to a final concentration of 1 mM to induce protein expression followed by incubation at room temperature with shaking on the Excella E5 platform shaker (New Brunswick Scientific) at 250 rpm overnight. The cells were pelleted by centrifugation for 30 min at 6000g at 4 °C. Recombinant proteins were purified using the Profinia protein purification instrument (Bio-Rad, Hercules, CA) according to the manufacturer's protocol. The Bio-Rad native IMAC protocol was selected at standard flow rate and standard wash time and used in conjunction with the Profinia IMAC purification kit (no. 6200225, Bio-Rad). The buffer of the purified enzyme was changed to a 50% (w/v) ethylene glycol solution (no. 29810, Life Technologies, Carlsbad, CA) with 5 mM $MgCl_2$, 1.5 mM TPP and adjusted to pH6.5. Buffer was changed by centrifugation at 4 °C using an Amicon ultra 15 mL centrifugal filter (EMD Millipore) according to manufacturer's protocol. Protein concentration was measured using the quick start Bradford protein assay kit (Bio-Rad). The purified proteins were analyzed on 15% Mini-PROTEAN TGX stain-free gel (Bio-Rad) and visualized on the Bio-Rad Gel Doc EZ imager to ensure that the purity of Kivd variants are similar. Purified protein was stored at −20 °C and used for characterization as needed according to the protocol described in other sections. Multiple protein purifications were used for each of the experiments to characterize the Kivd variants.

### 2.3.8 Heat Incubation

Purified Kivd was preincubated at respective temperatures for duration as needed by the characterization test in a thermocycler (Mastercycler nexus GSX1 flexlid, Eppendorf, Hauppauge, NY). A 60 µL aliquot of the purified Kivd variant or wild type suspended in 50 mM pH 6.5 sodium phosphate buffer containing 5 mM $MgCl_2$ and 1.5 mM TPP was preincubated in each PCR tube.

### 2.3.9 NADH-Coupled Kivd Enzyme Assay

To characterize the purified Kivd variants, a more sensitive NADH-coupled enzyme assay was adapted from previous work (Lin et al., 2014). The assay was conducted in 200 µL reaction mixtures in a UV-transparent 96-well plate (no. 3635, Corning). A 10 µL aliquot of purified Kivd variant was added to 190 µL of a fresh master mix solution to start the reaction. Similar to that in a previous report (Plaza et al., 2004), final concentrations of the reaction mixture were 5 mM $MgCl_2$, 1.5 mM TPP, 30 mM 2-ketoisovalerate, 140U of commercial alcohol dehydrogenase (A3263, Sigma-Aldrich, St. Louis, MO) and 0.2 mM NADH in 50 mM pH 6.5 sodium phosphate buffer. TPP is prepared fresh at the point of enzyme assays. The master mix was incubated at 50 °C for 5 min while the purified Kivd variant was incubated at room temperature for 5 min before the reaction was started. Reaction mixtures are shaken for 1 min at medium intensity on a PowerWave XS microplate spectrophotometer (Bio-Tek, Winooski, VT) while being incubated at 50 °C. The consumption of NADH is then measured at 340 nm absorbance on the PowerWave XS microplate spectrophotometer at 50 °C (Bio-Tek). Enzyme activity is calculated by using NADH extinction coefficient at 6220 $M^{-1}$ $cm^{-1}$ and a light path of 0.5533 cm. Specific activity is obtained through normalization of the activity by enzyme concentration.

### 2.3.10 Recombination

Recombination was performed after the increased thermostability of the Q34H, A290V, S386P, and V130I mutations were confirmed. Starting with the A290V variant, the other mutations were added with the appropriate primer design sequentially. Primers with a point substitution encoding for the appropriate additional mutation were used to amplify Kivd from the variant in a PCR. The resulting PCR products were assembled as described above in plasmid construction to reconstruct the plasmid pLS02. This pLS02 construct contained Kivd variant LLM4, with mutations Q34H, A290V, S386P, and V130I.

### 2.3.11 $T_{50}$ determination

For the determination of $T_{50}$, the purified wild-type Kivd and variants were preincubated at temperatures ranging from 45 to 65 °C for 1 h using the temperature gradient function on the thermocycler (Master- cycler nexus GSX1 flexlid, Eppendorf, Hauppauge, NY). Following this, the NADH-coupled assay as described before was started by adding 10 µL of the preincubated purified enzyme to 190 µL of the enzyme master mix. Three independent repeats were conducted and the mean specific activities with standard deviation are plotted in Figure 2-4A. $T_{50}$ values are computed as follows: First, a second order polynomial is fitted to the mean values plotted in Figure 2-4A using the MATLAB “fit” function. Next, to calculate the $T_{50}$ for 1 h preincubation, the polynomial fit equation from Figure 2-4A was solved for the temperature at which half the original activity of Kivd would be lost using the appropriate MATLAB code.

### 2.3.12 Half-Life Determination

For the determination of half- life, multiple separate PCR tubes of the purified Kivd variant and wild type were preincubated at 60 °C. Time points from 0 to 60 min were determined and as the respective time point was reached, the appropriate PCR tube was removed from the thermocycler and stored at 4 °C until the purified Kivd enzyme assay (NADH-coupled assay) described before was started. The assay was started within an hour of the first sample being taken out from the thermocycler, and all samples were processed simultaneously. Three independent repeats were conducted and the mean specific activities with standard deviation were plotted in Figure 2-4B. Half-life values are computed as follows: First, a second order polynomial is fitted to the mean values plotted in Figure 2-4B using the MATLAB “fit” function. Next, to calculate the half-life at 60 °C, the polynomial fit equation from Figure 2-4B was solved for the preincubation time at which half the original activity of Kivd would be lost using the appropriate MATLAB code.

### 2.3.13 CD Spectroscopy

Before CD spectroscopy measurements, buffer of the purified protein was exchanged to 50 mM pH 6.5 sodium phosphate buffer with 5 mM $MgCl_2$ via dialysis (MWCO 3500, Spectrumlabs, CA) to ensure total elimination of TPP, imidazole, and glycerol in the solution as these compounds would affect absorption readouts. Samples were then prepared to the appropriate concentration (4.14 μM), with and without 10 μM of fresh TPP, which is around 2.5 mol equiv of the purified protein. CD spectrum was obtained by a JASCO J-815(JASCO, Japan) using 1 mm path length Suprasil quartz cells (Hellma, UK). Full wavelength data were collected at 25 °C, with the wavelengths ranging from 195 to 260 nm at 0.5 nm intervals. A protein unfolding curve was collected with a 1 °C interval at 222 nm after 5 min of incubation at each specific temperature. $T_m$ was computed as follows. First, the acquired data were fitted by a sigmoid curve. Next, the first order derivative of the sigmoid curve was taken. The $T_m$ was then determined by locating the local maxima of the derivative plot.

### 2.3.14 Kinetic Parameters and Substrate Specificity Determination

Kinetic parameter characterization was performed using the NADH-coupled enzyme assay with 2-ketoisovalerate concentrations varying from 0 to 10 mM. Three independent repeats were conducted. A second order polynomial fit was fitted to the mean values computed using the MATLAB “fit” function. Next, to calculate the $K_m$, the polynomial fit equation was solved using the appropriate MATLAB code for the 2-

ketoisovalerate concentration at which half the maximum specific activity of Kivd ($V_{max}$) would be observed. $k_{cat}$ numbers were calculated assuming the theoretical weight of a single Kivd subunit to be 61 kDa (Kivd is a homodimer). To test the specific activity on different substrates, the NADH-coupled enzyme assay was conducted as described before, substituting 30 mM 2-ketoisovalerate with 30 mM of the appropriate substrate. Three independent repeats were conducted, and the mean specific activities with standard deviation were plotted in Figure 2-7.

## 2.4 Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acssynbio.6b00240.

The activity profile of Kivd variants LLM3_F388Y and LLM3_Q437N after preincubation at various temperatures for 1 h (Figure 2-8). The activity profile of all rational designs individually recombined with LLM3 after preincubation at various temperatures for 1 h (Figure 2-9). Comparison of two different buffers on Kivd enzyme activity (Figure 2-10). Specific activities of wild- type KIVD and LLM4 to 2-ketoisovalerate, at physio- logical temperatures of 22, 30, and 37 °C (Figure 2-11). SDS-PAGE of wild-type Kivd and variants (Figure 2-12). Distance between carbanion of TPP ylide and C-alpha of mutated amino acid residue in the proposed Kivd structure model (Table 2-2). Codon changes that did not code for amino acid substitutions (Table 2-3).

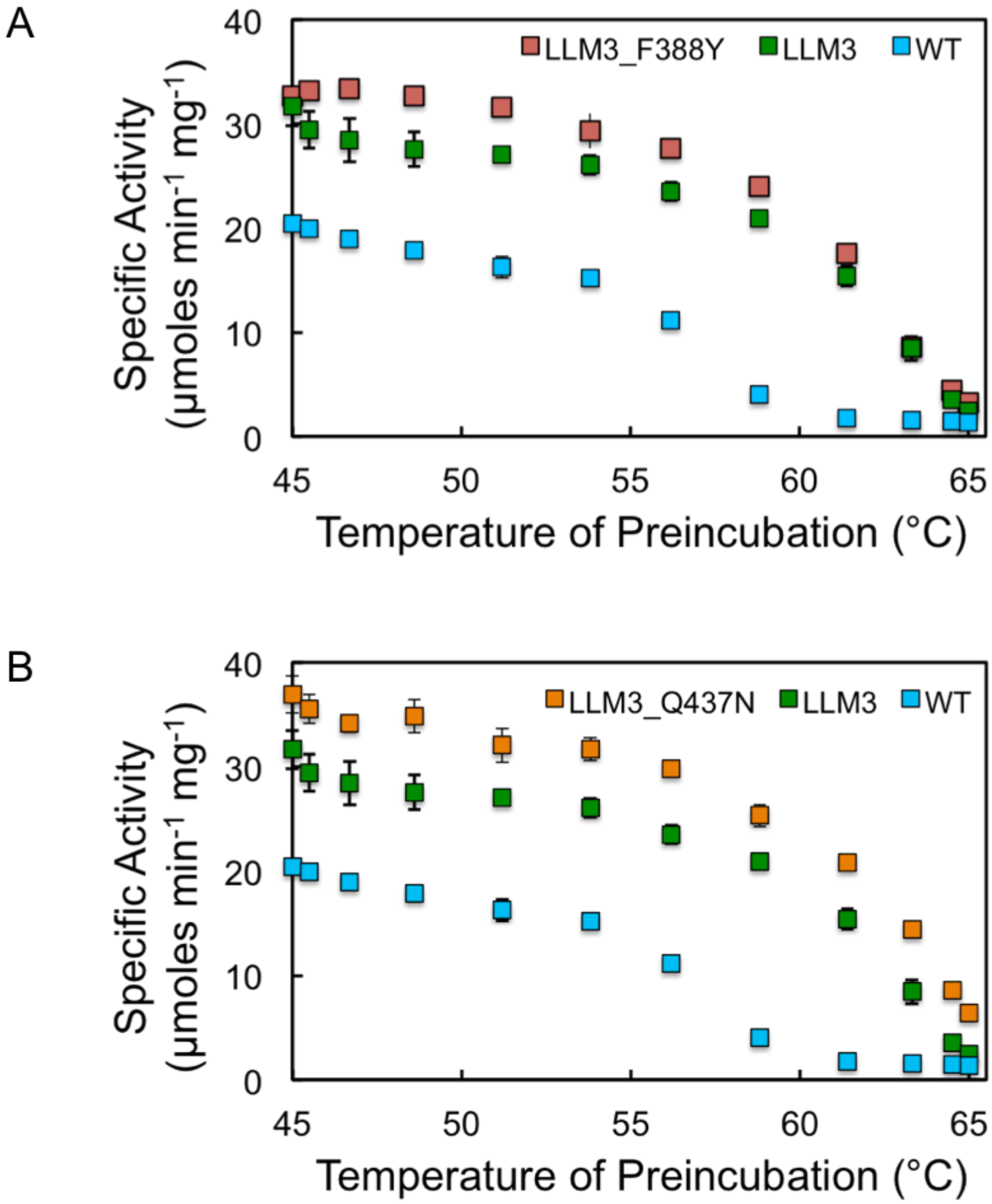


*Figure 2-8: Activity profile of thermostable mutations F388Y and Q437N recombined with Kivd LLM3*

*These mutations were identified via computational design and verified by experiments. Plot of Kivd specific activity measured at 50°C after 1h preincubation at various temperatures with Kivd wild type and LLM3 controls. (A) LLM3_F388Y. (B) LLM3_Q437N. Enzymes were purified and specific activities measured by NADH consumption at 50°C. The error bars represent the standard deviation of three independent repeats.*

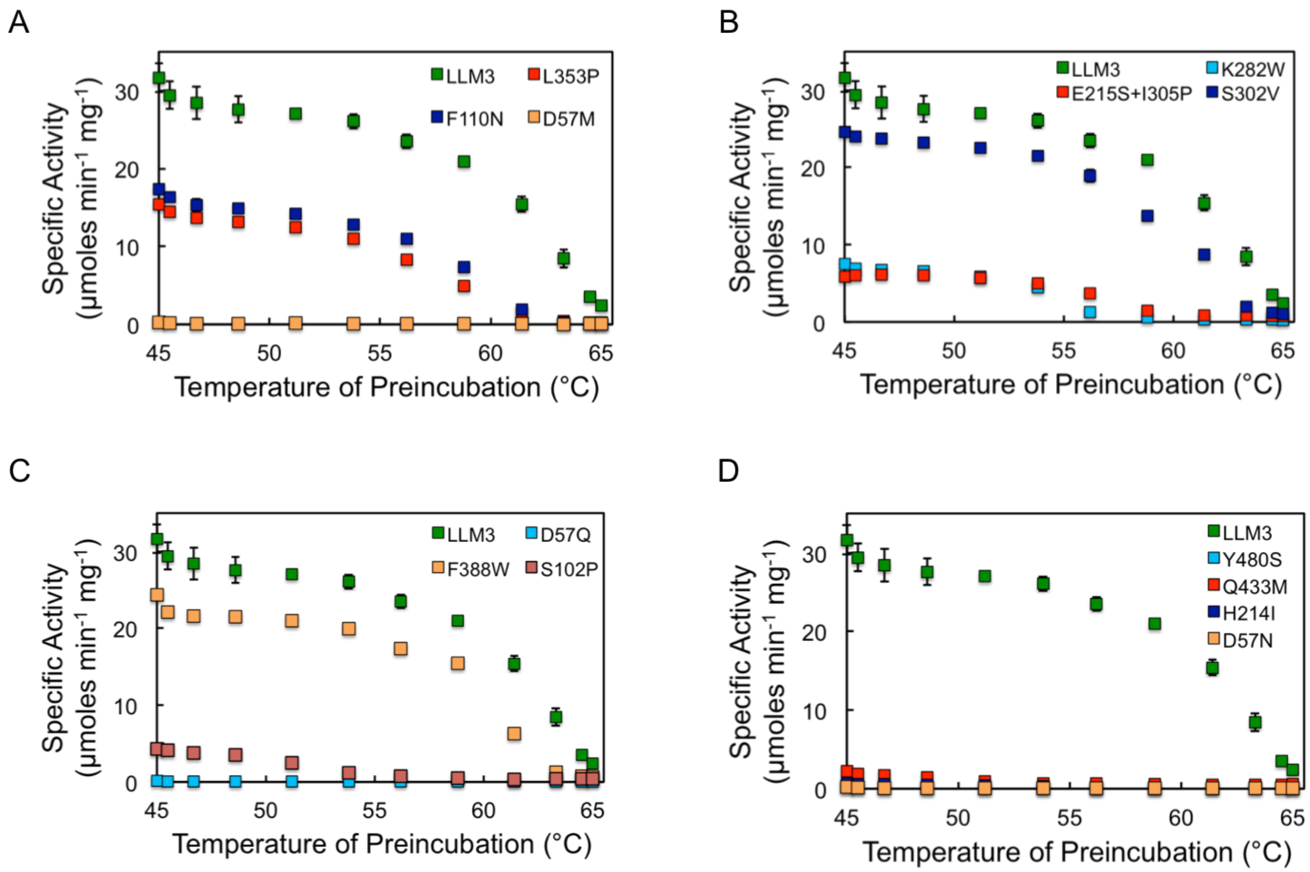


*Figure 2-9: Characterization of other mutations identified via computational design that had no thermostabilizing effect when recombined with Kivd LLM3*

*Plot of Kivd specific activity measured at 50°C after 1h preincubation at various temperatures with LLM3 control. (A) L353P, F110N and D57M (B) K282W, E215S + I305P, and S302V (C) D57Q, F388W and S102P (D) Y480S, Q433M, H214I and D57N. Enzymes were purified and specific activities measured by NADH consumption at 50°C. The error bars represent the standard deviation of three independent repeats.*

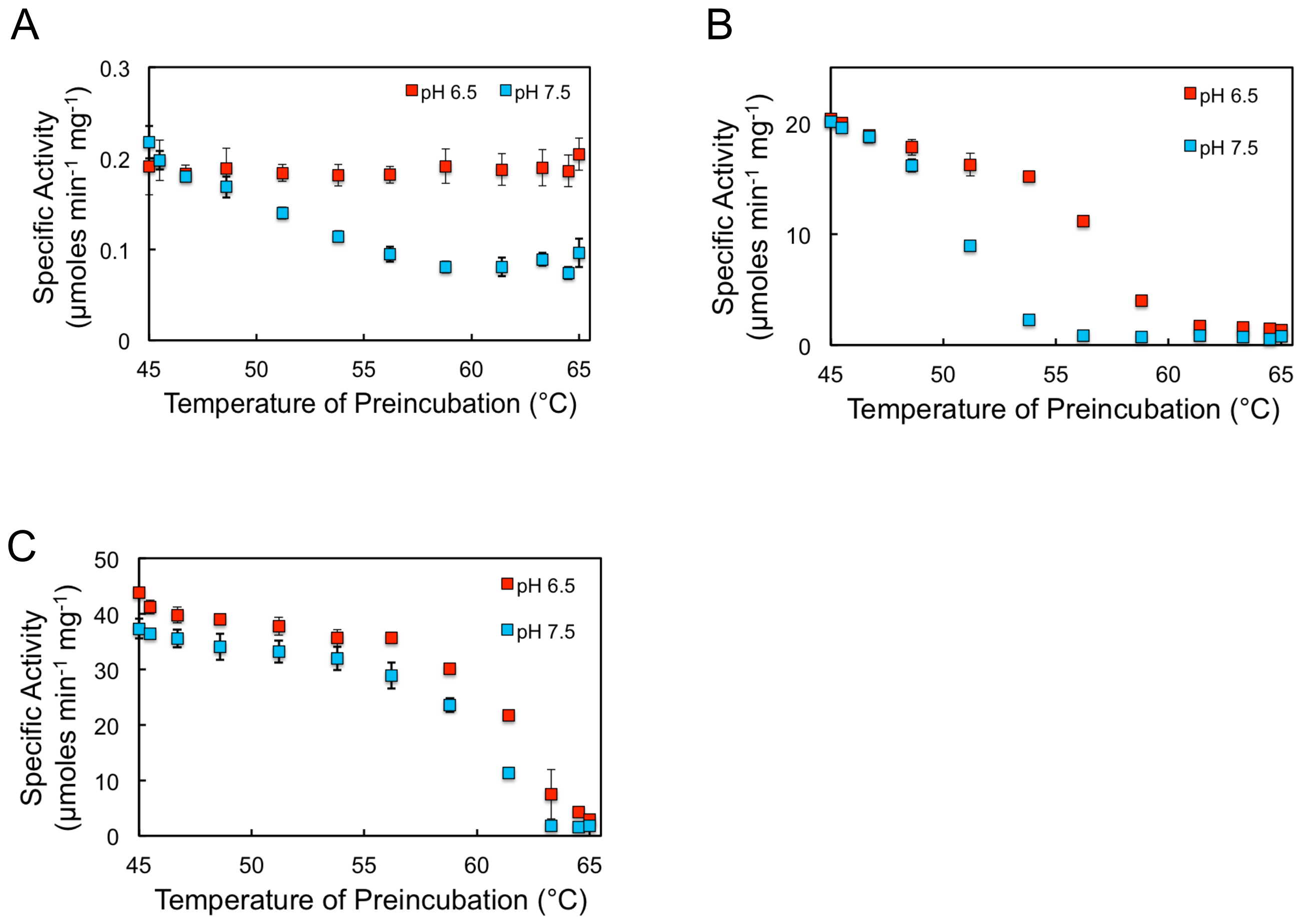


*Figure 2-10: Comparison of the effects of two different buffers on Kivd specific activity and thermal stability. Plot of Kivd specific activity after 1h preincubation at various temperatures*

*(A) Gtng_0348 (B) Wild-type Kivd. (C) LLM4 Kivd variant. Higher specific activity and thermal stability was recorded with the pH 6.5 buffer under different tests for Gtng_0348, wild-type Kivd and Kivd LLM4 variant. Enzyme variants were purified and specific activities measured by NADH consumption at their respective temperatures. Final concentrations of the reaction mixture at pH 6.5 was as follows: 5mM $MgCl_2$, 1.5mM TPP, 30mM 2-ketoisovalerate, 140U of commercial alcohol dehydrogenase and 0.2mM NADH in 50mM pH 6.5 sodium phosphate buffer. Final concentrations of the reaction mixture at pH 7.5 was as follows: 10mM NaCl, 2.5mM $MgSO_4$, 0.1mM TPP, 10mM 2-ketoisovalerate, 140U of commercial alcohol dehydrogenase and 500μM NADH in 10mM Tris buffer at pH 7.5. The error bars represent the standard deviation of three independent repeats.*

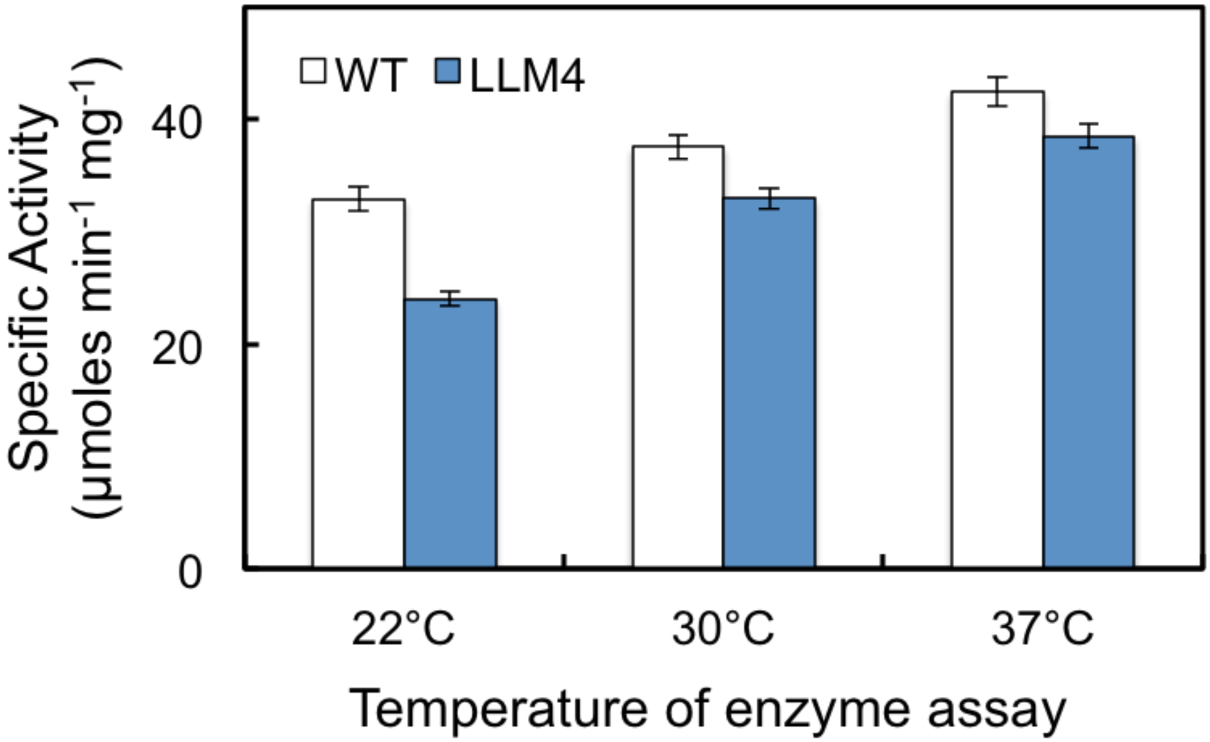


*Figure 2-11: Specific activities of Kivd wild-type and Kivd LLM4 to 2-ketoisovalerate, at physiological temperatures of 22°C, 30°C and 37°C*

*The Kivd wild-type is slightly more active at physiological temperatures than Kivd LLM4 variant. Using purified enzymes, specific activities measured by NADH consumption at their respective temperatures. The error bars represent the standard deviation of three independent repeats.*

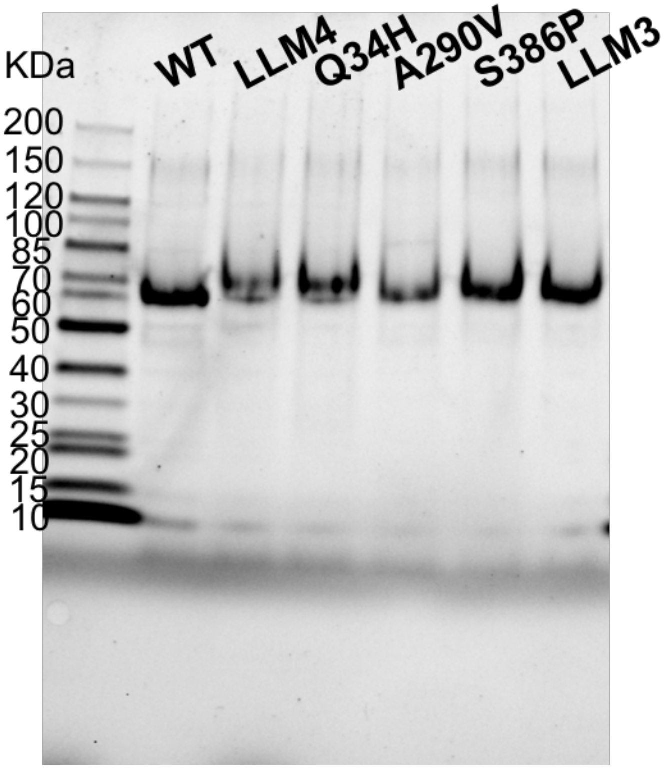


*Figure 2-12: SDS-PAGE of Kivd and different variants*

*LM4: Q34H+A290V+S386P+V130I; LLM3: Q34H+A290V+S386P.*

*Table 2-2: Distance between the carbanion of the TPP ylide and C-alpha of the mutated amino acid residue in the proposed Kivd structure model*

| **Mutation sites** | **Distance (Å)** |
|---|---|
| Rosetta Mutations | |
| D57M | 18.9 |
| D57N | 18.9 |
| D57Q | 18.9 |
| S102P | 20.2 |
| F110N | 16.2 |
| V130I | 21 |
| H214I | 20.6 |
| E215S | 23.6 |
| K282W | 20.3 |
| S302V | 30.2 |
| L353P | 25.4 |
| F388W | 24.1 |
| F388Y | 24.1 |
| Q433M | 12.7 |
| Q437N | 17.6 |
| Y480S | 19.8 |
| Mutations from directed evolution | |
| A290V | 15.8 |
| Q34H | 19.5 |
| S386P | 18.7 |

*Table 2-3: Codon changes observed during initial screen that did not code for amino acid changes*

| Codon Changes | Amino Acid |
|---|---|
| CAC to CAT | 13H |
| CGC to CGT | 38R |
| TAT to TAC | 54Y |
| GAA to GAG | 107E |
| GCA to GCT | 171A |
| TTA to CTA | 441L |

*These could potentially be useful for improving expression of Kivd in E. coli and would require further experiments to verify.*

## 2.5 Acknowledgements

This research was supported by the DOE BioEnergy Science Center (BESC). This material is based upon research performed in a renovated collaboratory by the National Science Foundation under Grant No. 0963183, which is an award funded under the American Recovery and Reinvestment Act of 2009 (ARRA). The authors wish to thank undergraduate Ryan Wi for his help on the experimental work.

# 3 Biochemistry-guided search highlights small non-canonical gene sets predictive of endometrial cancer prognostic factors

This chapter contains unpublished material and is confidential. The manuscript is in preparation under the same title for submission. Lemuel Soh prepared the dataset, implemented machine learning classifiers, beam search, and evaluated predictive accuracy of all gene sets using parallel processing. Stephen Tran performed the enrichment analysis of biochemical interactions and the Kaplan-Meier survival analysis. Both Lemuel Soh and Stephen Tran contributed to the preparation of the manuscript. Prof. James Liao supervised Lemuel Soh.

# 3.1 Abstract

Histopathology-based endometrial cancer subtyping into Type I or II has been utilized for more than 30 years and is important to the prognosis of tumors. Recent studies have proposed distinct molecular subgroups for improved prognosis. However, more complete information about the relationship between current prognostic factors and molecular features could speed progress on a revised classification system. Here, we report the identification of small gene sets consisting of 3 genes that are predictive of prognostic factors using known biochemical interactions between genes and canonical pathways of biochemical relevance. Using a mRNA expression dataset of 548 endometrial cancers, cross-validated prediction accuracies obtained are 80% for overall survival at 5 years, 78% for progression-free survival at 5 years, 77% for European Society for Medical Oncology risk classification, 82% for histological grade, and 91% for histology type among high grade tumors. By building gene sets from fundamental biochemical relationships using machine learning classifiers incorporated into a heuristic search strategy, approximately 1.6 to 2 million two-gene and three-gene sets were evaluated for their predictive accuracy on all of the five prognostic factors. As the gene sets identified were mostly not subsets of curated gene ontologies, they may not be easily identifiable through mainstream bioinformatic tools. Further, Kaplan-Meier survival analysis showed a statistically significant difference in overall survival and progression-free survival between the predicted groups when using the most predictive identified gene sets to separate patient groups. Such small non-canonical gene sets predictive of prognostic factors could serve as candidate biomarkers for further clinical

validation and are expected to shed light on the underlying biochemistry of these factors.

## 3.2 Introduction

Endometrial cancer is the 5th most common cancer in women, with 320,000 new diagnoses and 76,000 mortalities in 2012 alone (Ferlay et al., 2015). Introduced by Bokhman more than 30 years ago, endometrial cancer is grouped into type I or type II, based upon histology, hormone receptor expression, and grade (Bokhman, 1983). However, a growing recognition of significant molecular and morphological heterogeneity in endometrial carcinomas exhorts a refinement of the current histopathology-based classification (Murali et al., 2014). Clinical features like FIGO stage, and pathological features such as histological type and grade, are important prognostic factors (Abu-Rustum et al., 2010; Creasman et al.; Salvesen et al., 2012), and will continue to provide crucial information in any revised classification. Nevertheless, having additional biomarkers, particularly on the molecular level, would help improve prognostic accuracy, and could even provide predictive information preventing overtreatment of patients who could be cured from surgical treatment alone (Murali et al., 2014).

Unfortunately, endometrial cancer has relatively few molecular markers that can accurately predict prognoses (Backes et al., 2016). One prospective avenue is to use messenger RNA (mRNA) because it reflects gene and protein expression levels, and data is readily available as RNA-seq, particularly from consortiums such as The Cancer Genomic Atlas (TCGA). Established bioinformatic tools for RNA-seq, such as differential gene expression analysis (Ritchie et al., 2015; Robinson et al., 2010; Trapnell et al., 2012), unsupervised clustering (Langfelder and Horvath, 2008), and gene set enrichment analysis (Subramanian et al., 2005), have contributed substantially in finding genes differential between prognosis outcomes (Getz et al., 2013), but hold limitations in regards to understanding the underlying biochemical basis.  For example, differential expression analysis can identify genes that have statistically significant differences in expression between prognostic outcomes but does not explain how the multiple genes act cooperatively to produce pathological biochemistry. Unsupervised clustering can identify implicated gene networks in cancer but lacks the resolution to pinpoint the most important sub-pathways within the clusters. Gene set enrichment analysis checks *a priori* known gene sets for enrichment among the most differentially expressed genes but are limited to known gene sets.

In this study we propose a method for identifying tractably sized gene sets that predict endometrial cancer prognostic factors yet share interpretable pathway and biochemical interactions. We apply the method to endometrial cancer data from TCGA and find gene-sets in specific biochemical pathways strongly predictive of a myriad of prognostic factors.

# 3.3 Materials and Methods

## 3.3.1 Dataset Choices and Preparation

A database of 548 endometrial tumor samples with normalized RNA sequencing (RNA-seq) expression and clinical data resource (CDR) provided by TCGA was chosen because of its consistent tissue collection procedures, robust sample processing (Hutter and Zenklusen, 2018), standardization of clinically relevant information (Liu et al., 2018) and reproducible data analytic pipelines (Ellrott et al., 2018; Hoadley et al., 2018). All data was downloaded from the National Cancer Institute (NCI) Genomic Data Commons (GDC).

For each sample, binary labels corresponding to endometrial prognostic factors from the original CDR clinical annotations were assigned, for overall survival, progression-free survival, European Society for Medical Oncology (ESMO) risk (Colombo et al., 2013), histological grade, and histology type (Table 3-1). The discrimination between endometrioid and serous histology type was restricted to high grade tumors as they have overlapping morphological features and are difficult to distinguish histologically (Soslow, 2013). Every prognosis factor was analyzed separately with its classification labels and available samples for all results reported in this study.

For each prognostic factor, we also filtered out samples without RNA-seq or clinical information, matched using TCGA barcodes. 187, 209, 532, 532, 296 samples for OS, PFS, ESMO, grade, and type, were analyzed respectively. Among 20,531 genes, 356 tumor samples had no reads on 3038 genes. Those genes were excluded to ensure similar sample sizes in accuracy comparisons, leaving the majority of known genes (17,493) for analysis.

### 3.3.2 Overview of method to identify gene sets predictive of endometrial prognostic factors

The objective of our approach was to identify pathway or biochemically related gene sets that are predictive of endometrial cancer prognostic factors. Here we explicate the three main components of the method:

1) Choosing a classifier and metric for prognosis prediction
2) Defining canonical pathways and biochemical interactions
3) Integrating canonical pathway and biochemical interactions with the classifier in a heuristic search framework

*Table 3-1: Criteria for Prognosis Factors and Sample Numbers Used in this Study*

| Prognostic Factor | Abbreviation | Label 0 Criteria | Label 0 Samples | Label 1 Criteria | Label 1 Samples | Total Samples |
|---|---|---|---|---|---|---|
| 5 year Overall Survival | OS | Alive after 5 years | 108 | Died from any cause within 5 years | 79 | 187 |
| 5 year Progression Free Survival | PFS | Progression free after 5 years | 94 | New Tumor Event within 5 years | 115 | 209 |
| European Society for Medical Oncology Risk Classification | ESMO | Low Risk: Low and intermediate risk as defined by ESMO | 213 | High Risk: High risk as defined by ESMO | 319 | 532 |
| Histological Grade | Grade | Low Grade: Grade 1 and 2 | 218 | High Grade: Grade 3 | 314 | 532 |
| Histology Type | Type | Endometrioid: Endometrioid Histology Type and Grade 3 | 184 | Serous: Serous Histology Type and Grade 3 | 112 | 296 |

### 3.3.3 Classifier and Metric Choice

Our machine learning objective was to classify samples into their correct binary prognostic labels based on gene expression levels from RNA-seq. Since the prognostic labels were relatively balanced in number (Table 3-1), accuracy was chosen as the evaluation metric because of its simplicity. All reported prediction accuracies in this study refer to the average accuracy of stratified 10-fold cross validation.

To select a classifier, we evaluated the accuracy and average run time of eleven different machine learning algorithms trained using all genes (Table 3-6); calculations were averaged over 3 runs. Since random forest was the only classifier with an average run time of less than 1 second, we additionally tuned the hyperparameters, namely maximum tree depth, number of trees in the forest, and feature number for best split consideration, again using all available genes and samples (Table 3-7). Hyperparameter combinations were sampled in two stages by assigning about three values to each hyperparameter. The values first assigned were broad in range, while those assigned next were narrowly defined around the best first stage combination. Different hyperparameters were chosen for each prognosis factor, and feature number during splits was set to the maximum number of features available at each depth during heuristic search.

Classification was performed using Python Scikit-Learn (Pedregosa et al.). Unless otherwise stated, all computation was performed on a 2016 Macbook Pro, with 2.6GHz Intel Core i7, and 16GB 2133 MHz LPDDR3 memory.

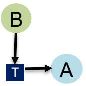
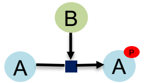
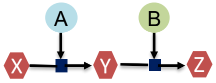
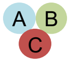
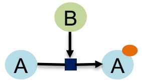
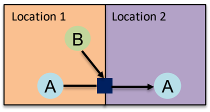
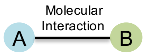
### 3.3.4 Defining Canonical Pathways and Biochemical Relationships

Our heuristic search was constrained to expand gene sets along pre-annotated canonical pathways and biochemical relationships. We defined two genes as part of the same “canonical pathway” if both were annotated as part of a same pathway from the Molecular Signatures Database v6.2 (Subramanian et al., 2005) or Humancyc curated pathways (Harmonizome database (Rouillard et al., 2016)). Only pathways with less than 100 genes were included. 499,573 gene pairs shared at least one canonical pathway.

We defined two genes as sharing biochemical interactions if both were annotated as having a direct biochemical interaction within the  Pathway Commons database (Cerami et al., 2011), which currently holds 632,859 such annotations. Direct biochemical interactions fall into seven different categories: expression control, phosphorylation control, catalysis precedes, in complex with, protein state change control, transport control, and interacts with (Table 3-2). Interactions defined within Pathway Commons that were not found among genes with valid RNA-seq reads in the TCGA dataset were excluded.

Finally, we considered two genes sharing either a canonical pathway or a direct biochemical interaction as broadly having a “biochemical relationship”.

*Table 3-2: Definitions of Direct Biochemical Relationships (Adapted from Pathway Commons)*

| Direct Biochemical Relationship | Figure | Description |
|---|---|---|
| Expression |  | Protein B controls a conversion or a template reaction that changes expression of the protein A. |
| Phosphorylation |  | Protein B controls a reaction that changes the phosphorylation status of protein A. |
| Catalysis |  | Protein A controls a reaction whose output molecule is input to another reaction controlled by protein B. |
| Complex |  | Proteins are members of the same complex. |
| State Change |  | Protein B controls a reaction that changes the state of the protein A. |
| Transport |  | Protein B controls a reaction that changes the cellular location of protein A. |
| Interacts |  | Proteins are participants of the same molecular interaction. |

### 3.3.5 Heuristic Search using Biochemical Relationships and Classification

Beam search (Lowerre and Reddy, 1990; Reddy), a greedy algorithm heuristic search with low memory requirements, was used to identify gene sets relevant to cancer prognosis. It was implemented with variable beam width and reordering based upon node predictive accuracy at every depth. Beam width is infinite up to depth two, where two gene sets are evaluated, and is set to the top 0.1% of gene sets at deeper depths. The goal for beam search was set to the accuracy obtained using all genes, termed the benchmark accuracy in this study.

Beam search begins with an initial single gene and uses our random forest classifier to evaluate the prediction accuracy of two-gene sets for a given prognostic factor, checking if the benchmark accuracy is reached or exceeded. However, the choice of second gene is restricted to those that share at least one biochemical relationship with the initial gene. If the benchmark accuracy is not found amongst the two-gene sets, beam search will expand into three-gene sets using the defined beam width; this time, the choice of third gene is restricted to those that share at least one biochemical relationship with the union of biochemical relationships in the existing two gene set. For instance, given a two-gene set with ATP1B2 and L1CAM, possible three-gene sets are generated by adding any one gene biochemically related to ATP1B2 or L1CAM. A possible three-gene node is ATP1B2, L1CAM and SCN4B. This process continues iteratively according to the same rules until beam search achieves the benchmark for prognostic predictive accuracy, which was set to the predictive accuracy attained using all genes in the genome.

### 3.3.6 Executing heuristic search on TCGA dataset

To achieve greater computational speed, all the 855,400 two-gene sets generated by beam search at depth two based on biochemical relationships were split into a hundred different groups for parallel processing of accuracy on Amazon Web Services (AWS) Elastic Compute Cloud (EC2). Five compute-optimized instances c5.4xlarge were launched, one for each prognostic factor, with 100 different processes executed on each instance. Next, for tractability, we only expanded the top 0.1% (about 600,000 to 900,000 per prognostic factor) of the most predictive two-gene sets into three-gene sets. Three-gene sets were also parallel processed as described before.

### 3.3.7 Using Canonical Pathways for Prognostic Factor Prediction

In general, prediction accuracy of prognostic factors increases with number of genes used in training (Figure 3-7). Thus, when assessing predictive accuracy on prognostic factors of entire canonical pathways (all genes within the pathway), the pathways had to normalized by its size (number of genes). Specifically, the predictive accuracy of each pathway was subtracted against the average of 100 equally sized random gene sets. Only genes with reads in all the samples per prognosis were included for calculating the normalized predictive accuracy.

### 3.3.8 Gene Set Analytics

For both two gene sets and three gene sets proffered by our search (above), we evaluated whether the topmost 1%, in predicting prognostic factors were enriched in any pathway terms or biochemical interactions.

For pathway enrichment, a gene set was considered to contain a pathway term if it was annotated to at least one constituent member gene. Enrichment of each pathway term was then tested using a one-tailed Fisher’s Exact test, which determines a p-value reflecting the strength of the odds ratio, calculated as the number of gene-sets to which the pathway term was held in the top 1% of most predictive gene sets vs bottom 1% least predictive gene sets.

Likewise, each of the 7 types of biochemical interactions was tested for enrichment in the top 1% vs bottom 1%. Any given gene set was considered to contain a given biochemical relationship if any pairwise combination of its member genes had the interaction. The odds ratio, calculated as the number of gene-sets to which the biochemical interaction was held in the top 1% vs bottom 1%, was tested using a two-tailed Fisher’s Exact test, adjusted using Bonferroni correction.

Testing for differences in distributions of overall biochemical relationships was done using a chi-square test, again on the top 1% most predictive gene sets vs bottom 1% least predictive gene sets.

### 3.3.9 Kaplan-Meier Analysis

For a given prognostic factor and gene set, every sample corresponding to a patient was placed into one of two groups based upon their predicted binary classification. Clinical records from the CDR (Liu et al., 2018) were used to determine overall survival and progression free survival for each individual patient. Based on these clinical records of how individual patients fared over time, Kaplan-Meier survival analysis was performed on the two predicted groups of patients. Kaplan-Meier survival analysis and plots were generated using the R package Survival Analysis.

## 3.4 Results

Most systems level studies of cancer are constrained to analyzing established, canonical pathways such as PI3K (Slomovitz and Coleman, 2012) and APC (Moreno-Bueno and Hardisson), despite indications that cancer leverages small, yet critical subsets of genes in non-canonical ways (Miyamoto et al., 2015) and utilizes biochemical interactions between genes across disparate pathways.

To identify salient non-canonical gene sets, we implemented beam search (Lowerre and Reddy, 1990; Reddy) that traverses along known pathways and biochemical interactions (methods) and then identifies “best” subsets of these genes, determined as those that have the highest prediction accuracy of cancer prognoses (e.g. predicting 5 year overall survival, recurrence, etc.) (Figure 3-1) (methods). Cross-validated prediction accuracy was calculated on the endometrial cancer dataset from TCGA using machine learning on RNA-seq derived gene expression levels, a metric found to have relatively high prediction power of cancer prognostic factors (Table 3-6). All further results described are of this dataset, but our approach may be applied to other clinically relevant outcomes or cancer types.

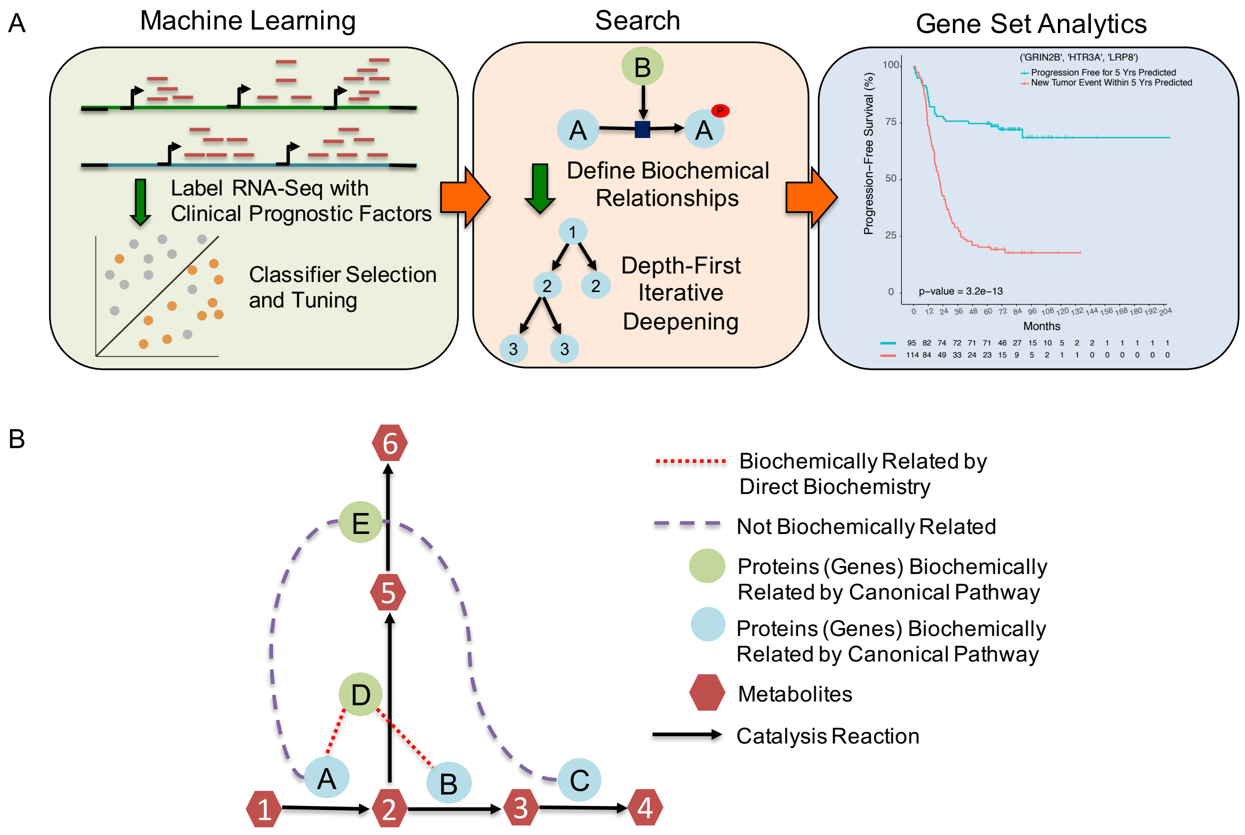


*Figure 3-1: Overview of approach*

*(A) Summary of biochemistry-guided search process. (B) Example of biochemically related genes in catalytic reaction pathways according to the definition in table 3-2 of this study.*

In our initial implementation using AdaBoost ExtraTrees tested on histological grade, calculating a single prediction accuracy took approximately ten seconds (Table 3-6), which would require an untenable 63 years to calculate the predictive accuracy within even the most rudimentary subspace of all 2-gene combinations (about 200 million pairs). Thus, we tested 10 other classifiers and found random forest classification to have a reasonable tradeoff between speed and accuracy (more than 10-fold speed increase with about 0.1 decrease in accuracy for histological grade) (Table 3-6). In particular, given a complete training set of all genes, random forest classification was tuned to achieve 10-fold cross-validated prediction accuracies for histological grade, histology type, overall survival at 5 years, progression-free interval at 5 years, and ESMO risk classification of 81%, 85%, 70%, 67%, and 74% respectively (Table 3-8), which were subsequently considered as the benchmark of prediction accuracy for this dataset.

To assess the utility in finding predictive subsets of gene pathways, we first calculated the predictive accuracy of entire canonical pathways (i.e. all genes within a given pathway). A myriad of canonical pathways had predictive accuracy greater than randomly sampled gene sets (Figure 3-2, Figure 3-8). Interestingly, the strongest predictive accuracy was observed in pathways containing smaller numbers of genes, suggesting that only subsets of canonical pathways are required for accurately classifying prognostic factors.

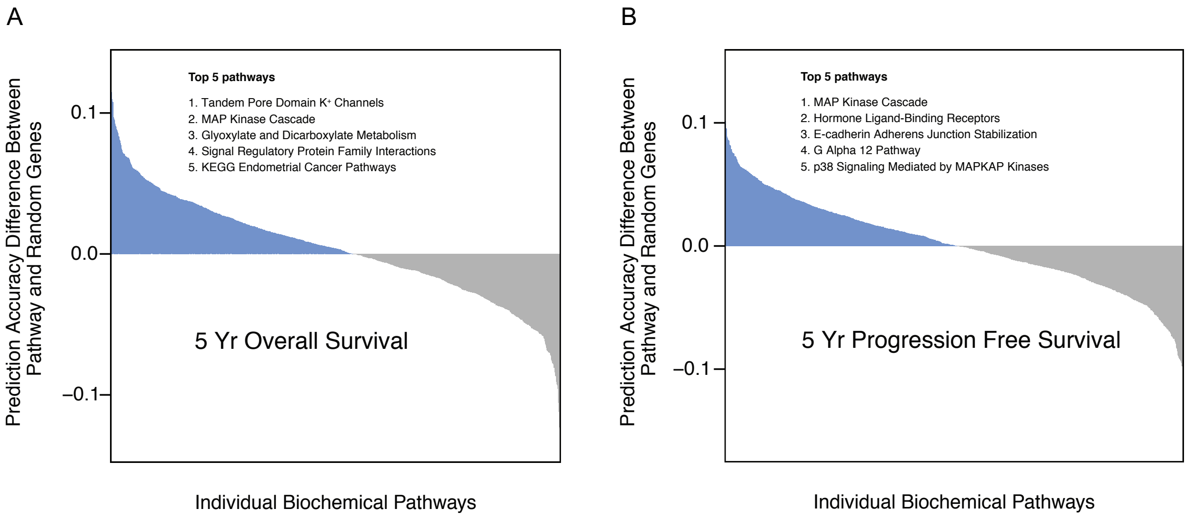


*Figure 3-2: Comparing prediction accuracy of canonical pathways over randomly sampled gene sets of the same size for overall and progression-free survival*

*Shown for (A) Overall survival at 5 years; (B) Progression free survival at 5 years.*

We first traversed our search along 855,400 pairs of genes but did not find any pair reaching the benchmark accuracy established using all genes. Therefore, we further expanded the 0.1% most predictive of two-gene sets (600,000 to 850,000 pairs) into three-gene trios. The most predictive three-gene sets had prediction accuracies exceeding that of the benchmark accuracy (Table 3-3), prompting us to not pursue four-gene sets.

Gene pairs exclusively within the same canonical pathways did not better predict complex prognostic factors (OS, PFS, and ESMO) than gene pairs exclusively sharing direct biochemical interactions (Figure 3-3A), corroborating our original hypothesis that some prognostic factors may not solely be explained by canonical pathways alone.

As expected, individual genes within each three-gene set were mostly from disparate pathways and mostly shared different types of biochemical interactions (Figure 3-3B). Additionally, the most predictive three-gene sets were not just conglomerations of strongly predictive individual genes as they sometimes contained individual genes with relatively low individual predictive power (Table 3-4, Table 3-5). Also, it was verified that there was no correlation between pairwise gene expression and gene set prediction accuracy (Figure 3-9).

*Table 3-3: Most Predictive Gene Sets for Prognostic Factors and Respective Accuracies*

| OS | PFS | ESMO | Grade | Type |
|---|---|---|---|---|
| ADRA1D<br>KHDRBS2<br>SLC5A1<br>(80.1%) | GRIN2B<br>HTR3A<br>LRP8<br>(77.6%) | AURKA<br>SMARCD3<br>YWHAE<br>(76.7%) | ARHGAP32<br>GNAI3<br>PGR<br>(82.2%) | ATP1B2<br>L1CAM<br>SCN4B<br>(90.6%) |
| CHD4<br>PIP5K1C<br>PTEN<br>(79.3%) | GNG3<br>LRP8<br>SLC1A7<br>(77.0%) | ASRGL1<br>GAD1<br>IL1B<br>(76.3%) | GSK3B<br>IHH<br>TACC3<br>(81.6%) | FN1<br>L1CAM<br>THBS3<br>(90.5%) |
| ADRA1D<br>HOMER1<br>SLC22A1<br>(79.1%) | AQP4<br>GNAS<br>LRP8<br>(75.6%) | AURKA<br>CRKL<br>YWHAE<br>(76.3%) | AHR<br>CDC20<br>SLC25A35<br>(81.4%) | BTRC<br>CDKN1A<br>WNT7A<br>(90.5%) |
| ADRA1D<br>CXCR3<br>MGP<br>(78.8%) | ETS2<br>PAK3<br>RAC2<br>(75.6%) | ARHGEF3<br>GNB5<br>IHH<br>(76.3%) | EMR1<br>HHAT<br>IHH<br>(81.4%) | ATP1B2<br>L1CAM<br>SLC5A5<br>(90.2%) |
| DNER<br>MYC<br>TXN<br>(78.7%) | EBF2<br>GNG3<br>LRP8<br>(75.5%) | AURKA<br>RB1<br>STAT5A<br>(76.3%) | CKS1B<br>ESR1<br>IHH<br>(81.2%) | ATP1B2<br>L1CAM<br>UQCRHL<br>(90.2%) |

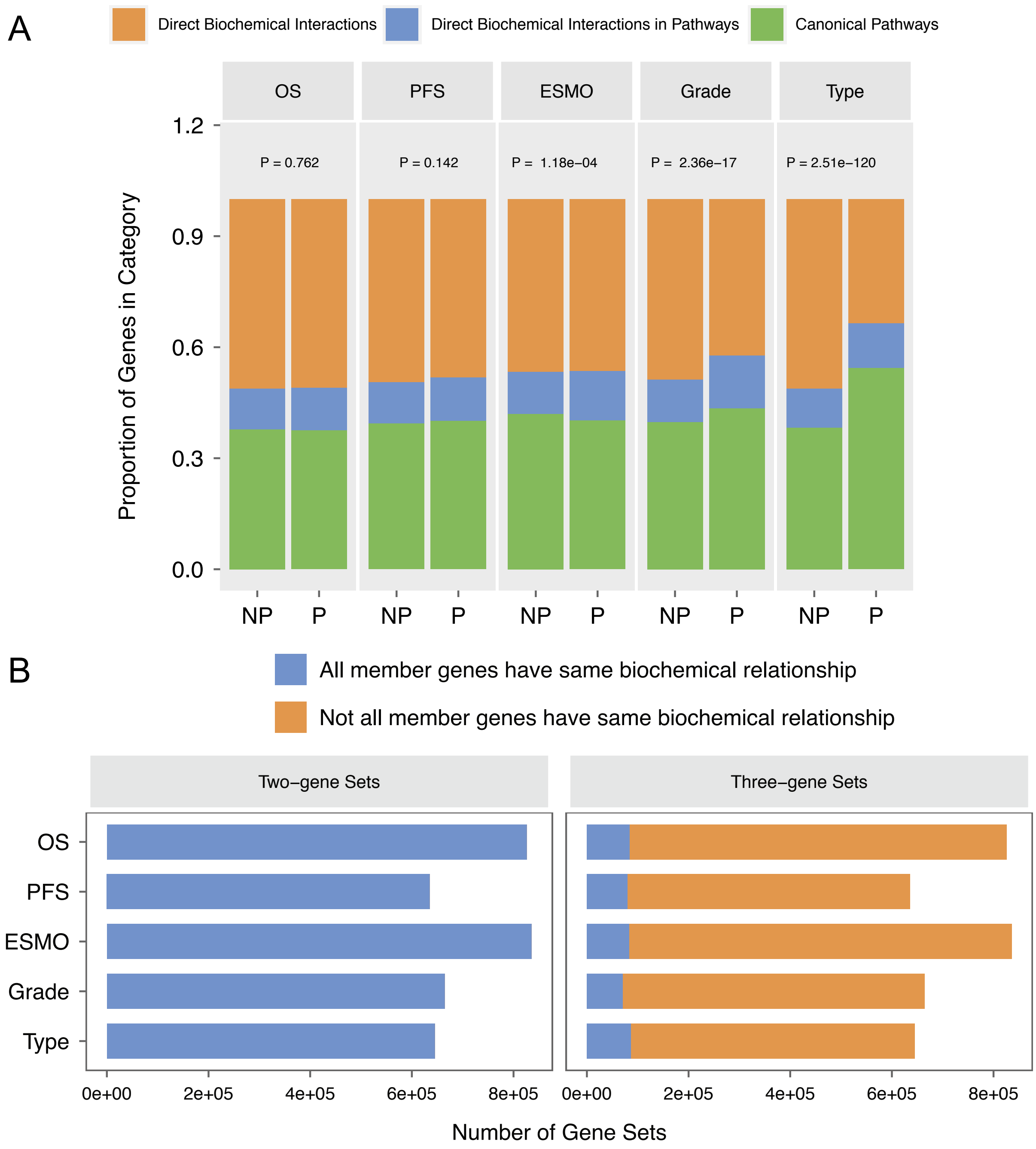


*Figure 3-3: Predictive power and the composition of different biochemical relationships within constructed gene sets*

*(A) Comparing predictive power of canonical pathways versus direct biochemical interactions in 2 gene sets; The top 1% of most predictive gene sets and the least predictive 1% are defined as*

*predictive and not very predictive respectively; P: Predictive, NP: Not very predictive. P-values from chi-square test comparing predictive to non-predictive gene sets (methods) (B) Canonical or non-canonical nature of gene sets. By design, our heuristic search (methods) restricts all 2 gene sets to share the same biochemical relationships (left). In contrast, not all members of 3 gene sets share the same biochemical relationships (right), usually because the 3$^{rd}$ gene introduces a new biochemical relationship. Blue: gene sets where all member genes share the same biochemical relationship; Orange: gene sets where not all member genes share the same biochemical relationship.*

*Table 3-4: Top Ten Most Predictive Individual Genes for All Prognostic Factors*

| Gene Rank | OS | | PFS | | ESMO | | Grade | | Type | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Gene Symbol | Accuracy | Gene Symbol | Accuracy | Gene Symbol | Accuracy | Gene Symbol | Accuracy | Gene Symbol | Accuracy |
| 1 | TUFT1 | 0.706 | MRPL15 | 0.693 | VWA3A | 0.675 | SLC25A35 | 0.701 | L1CAM | 0.779 |
| 2 | KIAA1199 | 0.691 | HTR6 | 0.667 | SERPINA11 | 0.671 | RAE1 | 0.700 | KCP | 0.761 |
| 3 | YWHAQ | 0.690 | FCRLB | 0.660 | OVGP1 | 0.660 | KIAA1324 | 0.696 | OSTF1 | 0.760 |
| 4 | C16orf86 | 0.685 | SNORA71A | 0.659 | ADRA1B | 0.656 | IHH | 0.696 | MYT1 | 0.756 |
| 5 | PDCD1LG2 | 0.685 | NSUN2 | 0.657 | IL20RA | 0.653 | AURKA | 0.690 | MDM2 | 0.752 |
| 6 | EDN2 | 0.683 | GPR141 | 0.656 | C22orf43 | 0.650 | SPDEF | 0.688 | HIF3A | 0.739 |
| 7 | LOC341056 | 0.679 | NOC2L | 0.656 | ANKRD33 | 0.648 | CKS1B | 0.684 | SLC6A12 | 0.736 |
| 8 | C14orf4 | 0.676 | GRM1 | 0.656 | PGR | 0.645 | FADS6 | 0.683 | GALNT10 | 0.735 |
| 9 | IL4I1 | 0.674 | CKMT1A | 0.653 | RARG | 0.645 | SCGB2A1 | 0.682 | FIGNL2 | 0.733 |
| 10 | C6orf142 | 0.674 | CXCR6 | 0.650 | CKS1B | 0.645 | FAM55B | 0.682 | NPR1 | 0.730 |

*Table 3-5: Gene Set Analysis showing Individual, Two-gene and Three-gene Predictive Accuracy and Basic Biochemical Relationships Allowing Identification of the Most Overall Predictive Gene Sets*

| Prognostic Factor | Gene | Individual Accuracy | 1st Pair Accuracy | 2nd Pair Accuracy | Three-Gene Accuracy | 1st Pair Biochemistry | 2nd Pair Biochemistry |
|---|---|---|---|---|---|---|---|
| OS | **ADRA1D** | 0.634 | 0.685 | | 0.802 | ADRA1D catalysis precedes SLC5A1 | |
| | **SLC5A1** | 0.611 | | 0.657 | | | SLC5A1 controls state change of KHDRBS2 |
| | **KHDRBS2** | 0.518 | | | | | |
| PFS | **LRP8** | 0.569 | 0.680 | | 0.776 | • Reelin Pathway<br>• GRIN2B in complex with LRP8 | |
| | **GRIN2B** | 0.588 | | 0.588 | | | HTR3A controls state change of GRIN2B |
| | **HTR3A** | 0.474 | | | | | |
| ESMO | **AURKA** | 0.637 | 0.692 | | 0.767 | • YWHAE controls state change of AURKA<br>• AURKA in complex with YWHAE | |
| | **YWHAE** | 0.500 | | 0.577 | | | LKB1 Pathway |
| | **SMARCD3** | 0.579 | | | | | |
| Grade | **GNAI3** | 0.594 | 0.754 | | 0.822 | Progesterone Mediated Oocyte Maturation | |
| | **PGR** | 0.645 | | 0.709 | | | PGR controls expression of ARHGAP32 |
| | **ARHGAP32** | 0.665 | | | | | |
| Type | **L1CAM** | 0.779 | 0.865 | | 0.906 | • Basigin Interactions<br>• Cell Surface Interactions at the Vascular Wall | |
| | **ATP1B2** | 0.578 | | 0.620 | | | ATP1B2 catalysis precedes SCN4B |
| | **SCN4B** | 0.563 | | | | | |

We next assessed whether our strategy of combining heuristic search with annotated biochemical pathways and interactions enabled more informative gene subsets. We extracted and recombined the 200 most predictive single genes into 19,900 gene pairs. Interestingly, the 186 biochemically related gene pairs had a statistically higher predictive accuracy (70.4%) than the remaining 19,789 biochemically unrelated pairs (69.8%) (two-tailed t-test, p-value=0.007), showing that adding biochemical knowledge augments prediction accuracy of gene sets.

Next, we performed a functional enrichment analysis comparing the top 1% most predictive gene sets to bottom 1% least predictive. Phosphorylation was the most consistently enriched biochemical interaction across the prognostic factors for two-gene sets (Figure 3-4A), whereas phosphorylation was only enriched in ESMO risk classification, grade and type for three-gene sets (Figure 3-4B). The top 1% for histological type was also enriched in pathways such as cell surface interactions, L1CAM interactions, and L1 signal transduction (Figure 3-4C, Figure 3-10). Strikingly, these top pathways were generally not as predictive when their accuracy was measured using all the member genes of the pathway (Figure 3-4D), indicating that subsets of certain canonical pathways are important for predicting prognostic factors, even if the entire set on average is not.

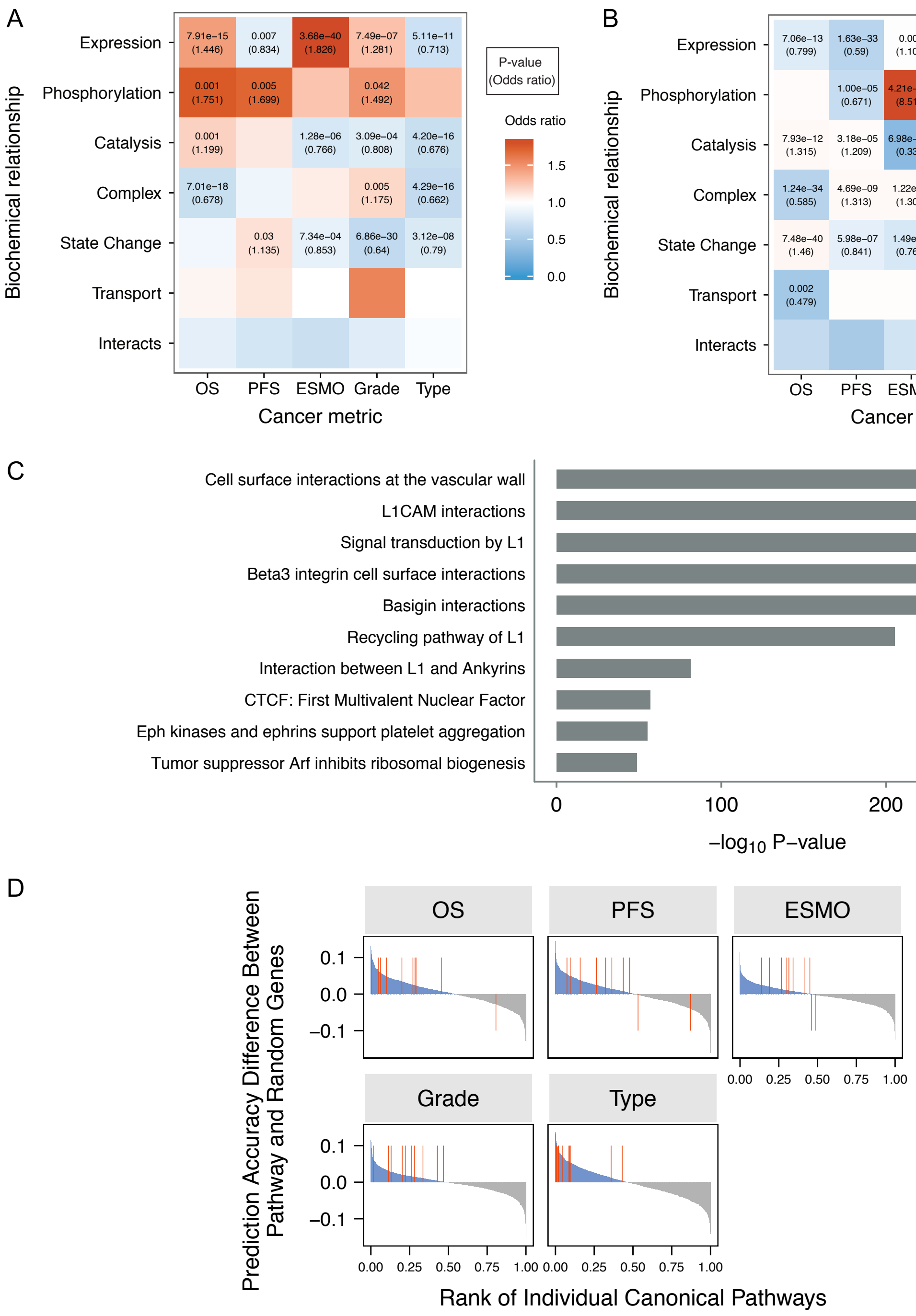


*Figure 3-4: Enrichments of direct biochemical relationships and pathways*

*Enriched direct biochemical interactions amongst the top 1% of gene sets identified as most predictive compared to the bottom 1% for (A) Two-gene sets; (B) Three-gene sets. (C) Enriched canonical pathways amongst the top 1% of gene sets identified as most predictive compared to* *the bottom 1% for three-gene sets predicting histological type. (D) Canonical pathways from analysis in (C) overlaid (red lines) with the canonical pathway ranking from Figure 3-2. Height of red lines are drawn at 0.1 or -0.1 depending on locating in the blue or grey hemisphere respectively.*

Finally, we examined the most predictive three-gene set of each cancer prognosis. The top three-gene sets enabled strong predictions in both of the binary labels (e.g. high vs low risk, high grade vs low grade, etc.) (Figure 3-5, Figure 3-11), indicating non-biased utility for predicting both benign and stark prognostications. Kaplan-Meier plots using these three-gene sets strongly separated differential outcomes in overall survival and progression free survival, suggesting potential salience of these gene sets as prognostic factors (Figure 3-6, Figure 3-12).

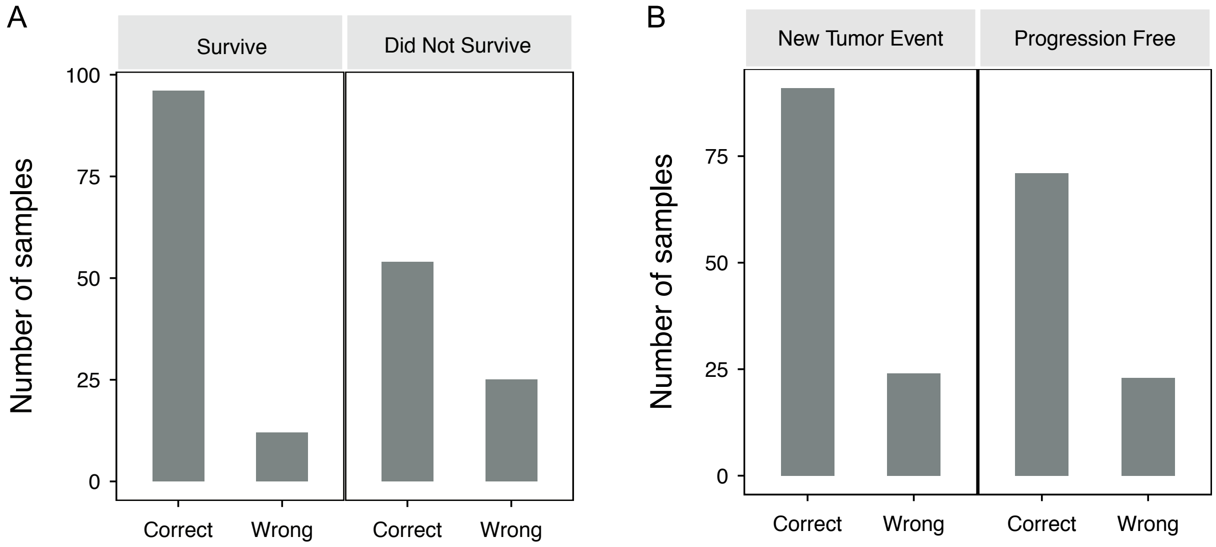


*Figure 3-5: Number of accurate predictions in each binary class using the highest predictive accuracy three-gene set for overall and progression-free survival*

*(A) Overall survival at 5 years; (B) Progression free survival at 5 years. Results indicate non-biased utility for predicting both classes across prognostic factors.*

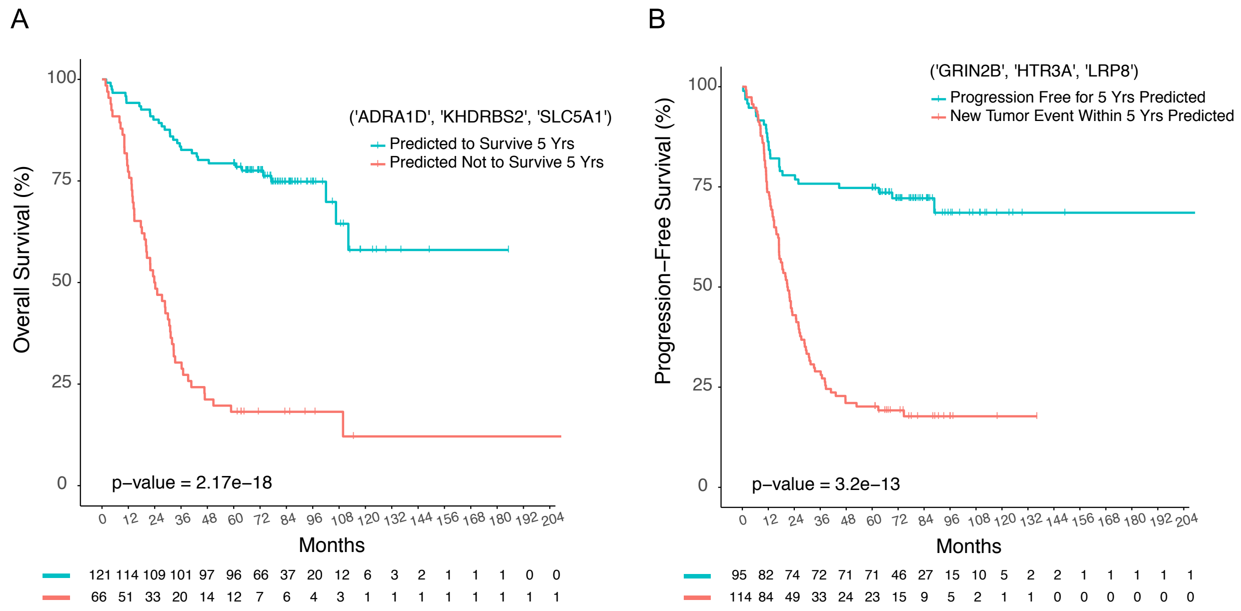


*Figure 3-6: Kaplan-Meier overall and progression-free survival curves for patients segregated into two groups based upon best gene set predictions for 5-year OS and PFS*

*(A) Overall survival for gene set ADRA1D, KHDRBS2 and SLC5A1 predicting 5-year OS; (B) Progression free survival for gene set GRIN2B, HTR3A and LRP8 predicting 5-year PFS.*

## 3.5 Discussion

The value of this study lies primarily in its ability to identify gene sets acting cooperatively to predict prognostic factors. Many contributing genes would not be picked up using traditional bioinformatic tools because their individual contributions are too small to rank highly in differential expression analysis or be found in gene set enrichment analysis. While most tools rely on statistical analysis or fixed gene ontologies, this study defined basic relevant pairwise biochemical relationships that were incorporated into a search strategy. This allows non-canonical gene sets to be identified.

Of the 58 unique genes in the five most predictive sets across all prognostic factors (Table 3-3), 13 genes, namely PTEN (Risinger et al.), AURKA (Umene et al., 2015), SMARCD3 (Bosse et al., 2013), YWHAE (Lee et al., 2012a, 2012b), ASRGL1 (Edqvist et al., 2015), CRKL (Padmanabhan et al., 2011), RB1 (Albitar et al., 2007), PGR (Ma et al., 2004), GSK3B (Moreno-Bueno and Hardisson), AHR (Wormke et al., 2000), ESR1 (Ashton et al., 2009), WNT7A (Carmon and Loose, 2008) and L1CAM (Dellinger et al., 2016) have been implicated in endometrial cancer. Another 8 genes, CHD4 (NIH Intramural Sequencing Center (NISC) Comparative Sequencing Program et al., 2012), CXCR3 (Kawada and Taketo, 2011), DNER (Lawrence et al., 2014), MYC (Doll et al., 2008), GNAS (Gielen et al., 2006), ETS2 (Gutierrez-Hartmann et al., 2007), CDC20 (Wong et al., 2007) and CDKN1A (Decruze and Green, 2007) were identified in screens but not heavily studied. The remaining genes offer opportunities for endometrial cancer biomarker validation and further biochemical study into the relationship within and between the gene sets.

The enrichment results and biochemical functions of genes in the identified sets suggest that each prognostic factor may be driven by different underlying biochemistry. When examining the most predictive three-gene sets, biochemical relationships within some sets could be proposed (Figure 3-13), but the relationships between different sets was not apparent and may require further investigation.

The identified gene sets have high predictive accuracy of prognostic factors comparable to the recombination of individually highly predictive genes, but yet often contain genes that are individually not highly predictive (Table 3-4, Table 3-5). A key importance of these gene sets is that their action of cooperativity can often be discerned without difficulty because the individual biochemical relationships that led to the gene set identification are without ambiguity.

The methods used in this study fall into the broad category of feature selection techniques (Saeys et al., 2007; Wang et al., 2016). Our study is not the first to identify discriminatory power in expressions of very small gene sets. Wang et. al. (Wang et al., 2007) distinguished 14 cancer types using the expression of 28 genes, with 2 genes distinguishing each binary pair of cancer types. However, to the best of our knowledge, this study is the first to use defined biochemical relationships to search for gene sets predictive of endometrial cancer prognostic factors and readily interpretable biochemistry.

In this study, the fundamental biochemical relationship was defined very broadly. All direct biochemical interactions and gene pairs in canonical pathways were considered to be biochemically related. If, however, a specific biochemical hypothesis needed to be tested, the fundamental biochemical relationship can be defined differently. A different definition would restrict the search to biochemical relationships deemed as relevant by the domain expert and use data-driven approaches to identify appropriate gene sets.

Many highly predictive gene sets contain at least one gene previously implicated in endometrial cancer (Albitar et al., 2007; Ashton et al., 2009; Bosse et al., 2013; Edqvist et al., 2015; Lee et al., 2012a, 2012b; Ma et al., 2004; Padmanabhan et al., 2011; Risinger et al.; Umene et al., 2015; Wormke et al., 2000). While this is encouraging to the validity of our approach, our contribution stems mainly from the identification of other genes that may act cooperatively with the previously identified biomarker. A drawback of not using canonical pathways in their entirety is the difficulty of understanding how non-canonical gene sets function to alter endometrial cancer prognosis. The biochemistry linking the gene sets are pre-defined from databases but how the cooperative action of these genes fits into the broader endometrial cancer prognosis requires further investigation and validation.

Close to 2 million gene sets were evaluated for each prognostic factor, but the search space is orders of magnitude greater, with about 200 million two-gene combinations, and about 1.3 trillion three-gene combinations. With limited computational power and data, the true optimal gene sets may not be knowable. Our study provides some evidence that using biochemical relationships to guide search is a viable approach to identifying gene sets that are more interpretable and predictive of endometrial cancer prognostic factors. The approach can also be modified to identify gene sets predictive of other important clinical measures. With further clinical validation, the gene sets identified in our study could serve as useful biomarkers.

## 3.6 Supporting Information

Figure 3-7: Plot of increasing sizes of randomly sampled genes and their predictive accuracies for each of the prognostic factors. Figure 3-8: Prognostic factors prediction accuracy of canonical pathways over randomly sampled gene sets of the same size. Figure 3-9: No statistical association between pairwise gene predictive accuracies and pairwise gene expression correlations. Figure 3-10: Enriched canonical pathways amongst the top 1% of gene sets identified as most predictive compared to the bottom 1%. Figure 3-11: Individual class predictions using the highest predictive accuracy three-gene set for the prognostic factors. Figure 3-12: Kaplan-Meier survival curves for patients segregated into two groups based upon gene set predictions. Figure 3-13: Proposed biochemical interactions between three genes identified as predictive. Table 3-6: Evaluating Different Classifiers Based on Run Time and Average Accuracy. Table 3-7: Random Forest Hyperparameters Chosen During Fine Tuning in Two Stages. Table 3-8: Prediction Accuracy of Prognostic Factors Using All Genes with Tuned Random Forest Classifier.

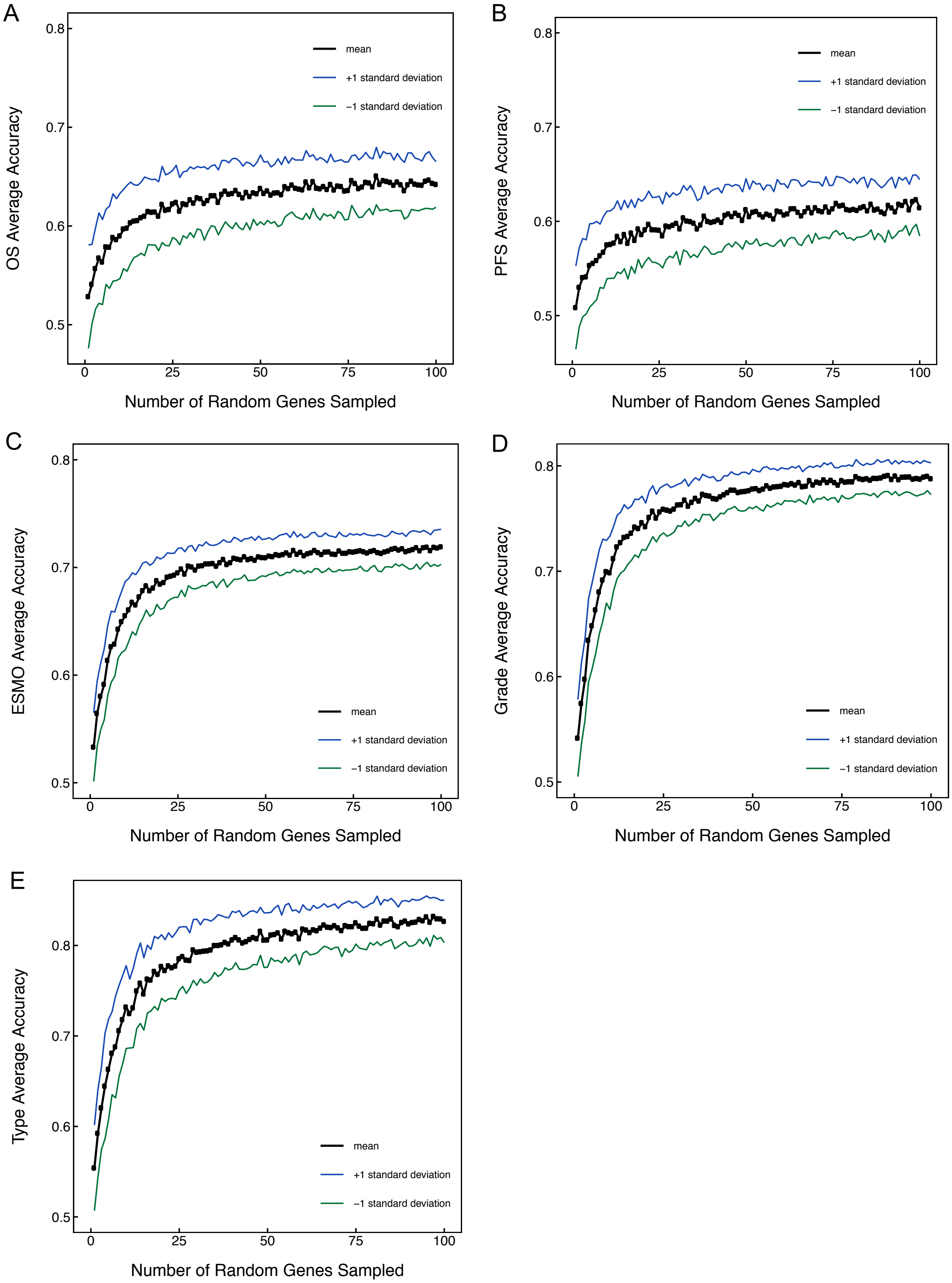

A
OS Average Accuracy
0.5
0.6
0.7
0.8
0
25
50
75
100
Number of Random Genes Sampled
mean
+1 standard deviation
−1 standard deviation
B
PFS Average Accuracy
0.5
0.6
0.7
0.8
0
25
50
75
100
Number of Random Genes Sampled
mean
+1 standard deviation
−1 standard deviation
C
ESMO Average Accuracy
0.5
0.6
0.7
0.8
0
25
50
75
100
Number of Random Genes Sampled
mean
+1 standard deviation
−1 standard deviation
D
Grade Average Accuracy
0.5
0.6
0.7
0.8
0
25
50
75
100
Number of Random Genes Sampled
mean
+1 standard deviation
−1 standard deviation
E
Type Average Accuracy
0.5
0.6
0.7
0.8
0
25
50
75
100
Number of Random Genes Sampled
mean
+1 standard deviation
−1 standard deviation

*Figure 3-7: Plot of increasing sizes of randomly sampled genes and their predictive accuracies for each of the prognostic factors*

*(A) Overall survival at 5 years; (B) Progression free survival at 5 years; (C) ESMO risk classification (D) Histological grade; (E) Histology type.*

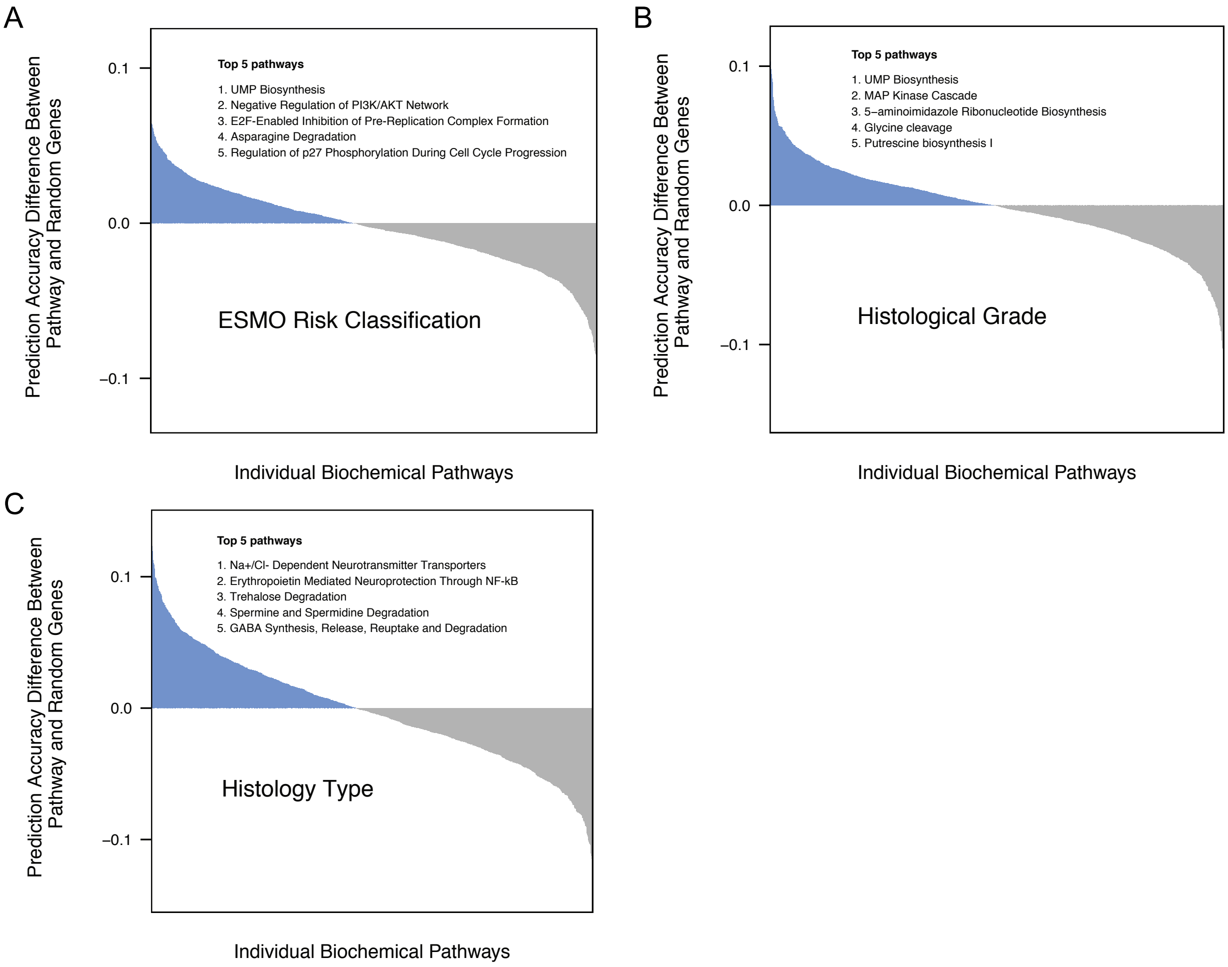


*Figure 3-8: Comparing prediction accuracy of canonical pathways over randomly sampled gene sets of the same size for ESMO risk, histological grade and type*

*(A) ESMO risk classification (B) Histological grade; (C) Histology type.*

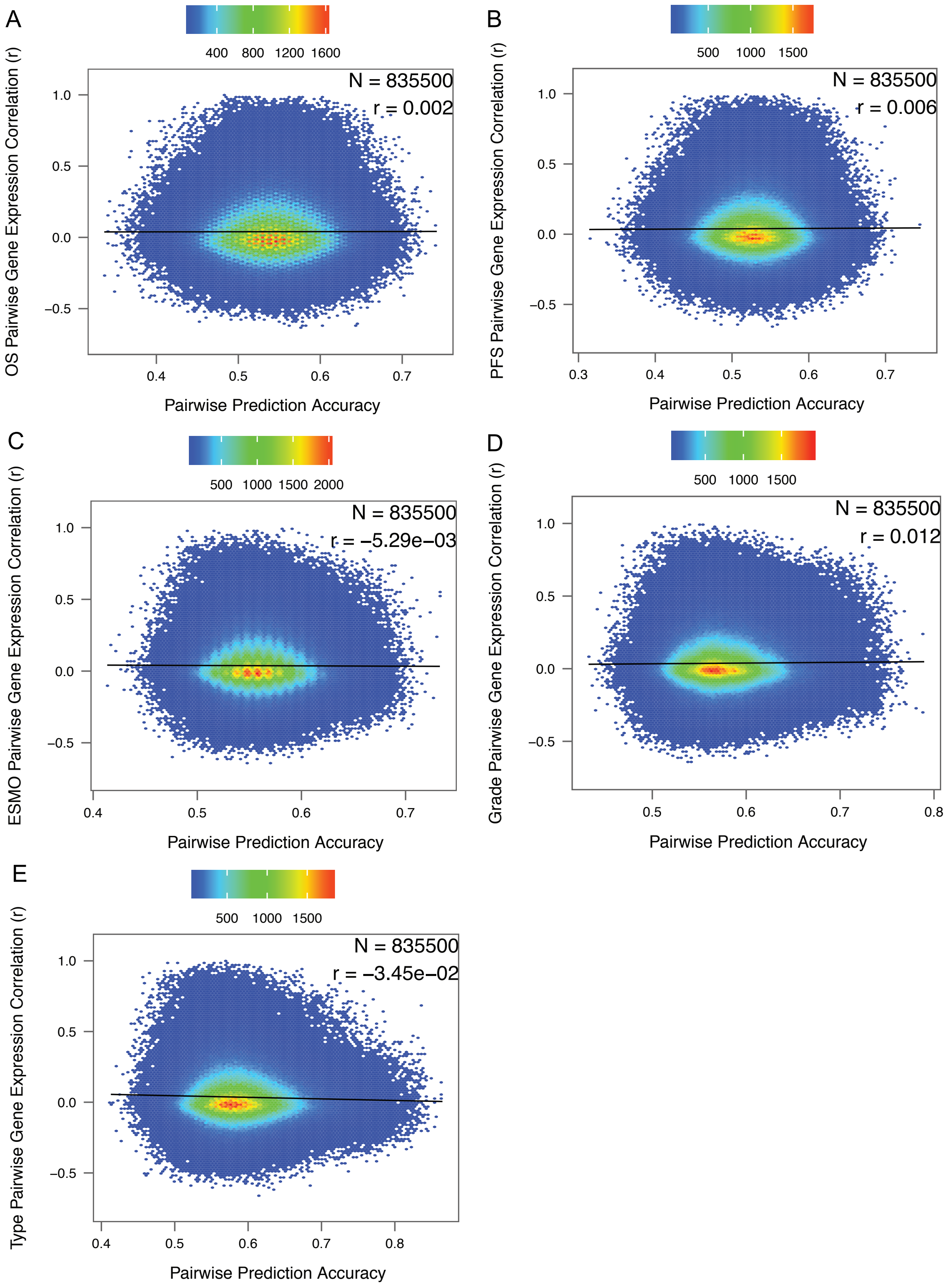

A
400 800 1200 1600
N = 835500
r = 0.002
OS Pairwise Gene Expression Correlation (r)
1.0
0.5
0.0
−0.5
0.4
0.5
0.6
0.7
Pairwise Prediction Accuracy
B
500 1000 1500
N = 835500
r = 0.006
PFS Pairwise Gene Expression Correlation (r)
1.0
0.5
0.0
−0.5
0.3
0.4
0.5
0.6
0.7
Pairwise Prediction Accuracy
C
500 1000 1500 2000
N = 835500
r = −5.29e−03
ESMO Pairwise Gene Expression Correlation (r)
1.0
0.5
0.0
−0.5
0.4
0.5
0.6
0.7
Pairwise Prediction Accuracy
D
500 1000 1500
N = 835500
r = 0.012
Grade Pairwise Gene Expression Correlation (r)
1.0
0.5
0.0
−0.5
0.5
0.6
0.7
0.8
Pairwise Prediction Accuracy
E
500 1000 1500
N = 835500
r = −3.45e−02
Type Pairwise Gene Expression Correlation (r)
1.0
0.5
0.0
−0.5
0.4
0.5
0.6
0.7
0.8
Pairwise Prediction Accuracy

*Figure 3-9: Statistical association between pairwise gene predictive accuracies and pairwise gene expression correlations*

*For all prognostic factors (A) Overall survival at 5 years, (B) Progression free survival at 5 years, (C) ESMO risk classification, (D) Histological grade, and (E) Histology type, there is no statistical association between pairwise gene predictive accuracies and pairwise gene expression correlations. All correlations are calculated using Pearson correlation. N = number of gene pairs; r = Pearson correlation coefficient between pairwise gene predictive accuracies and pairwise gene expression correlations.*

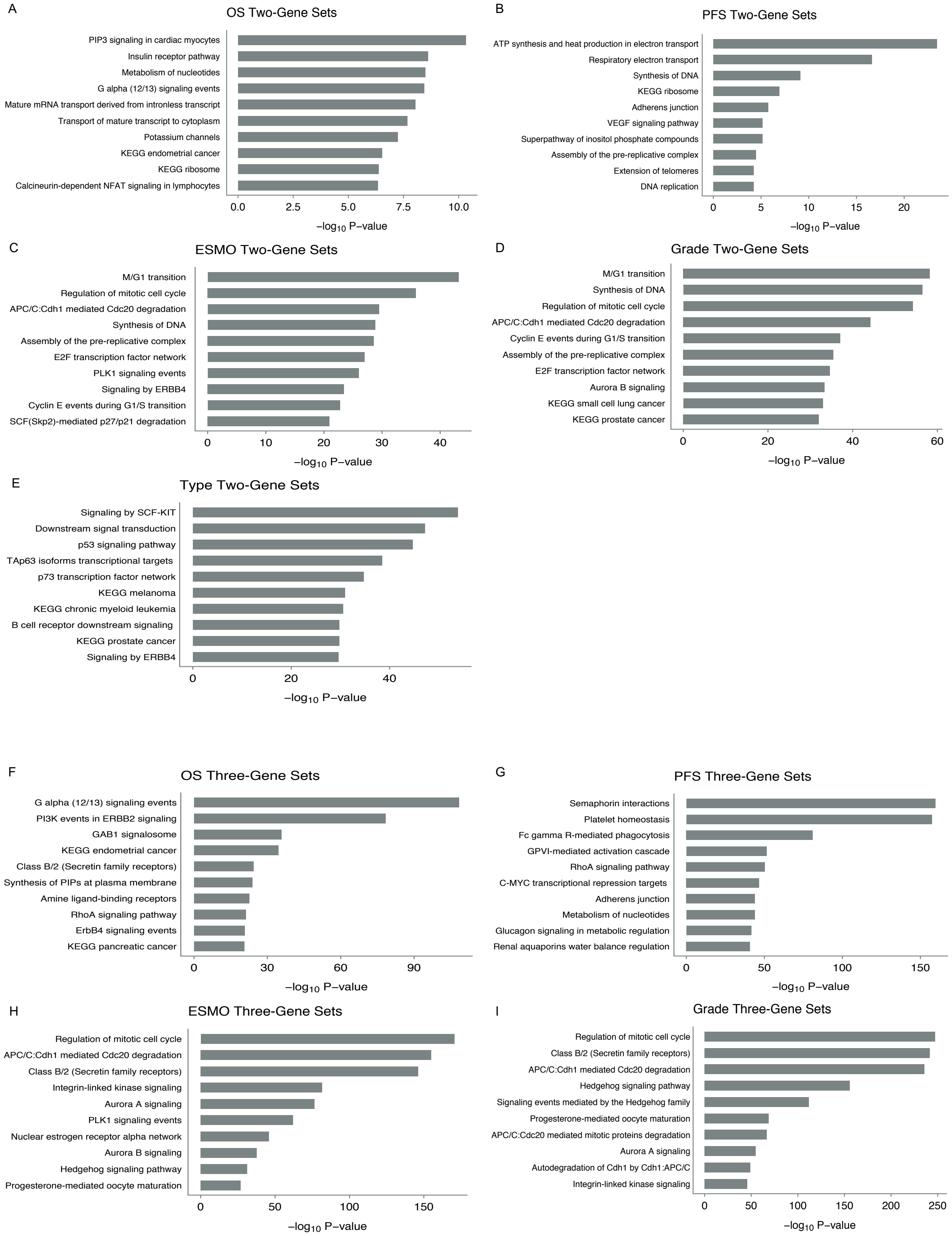

A
OS Two-Gene Sets
PIP3 signaling in cardiac myocytes
Insulin receptor pathway
Metabolism of nucleotides
G alpha (12/13) signaling events
Mature mRNA transport derived from intronless transcript
Transport of mature transcript to cytoplasm
Potassium channels
KEGG endometrial cancer
KEGG ribosome
Calcineurin-dependent NFAT signaling in lymphocytes
0.0
2.5
5.0
7.5
10.0
−log$_{10}$ P−value
B
PFS Two-Gene Sets
ATP synthesis and heat production in electron transport
Respiratory electron transport
Synthesis of DNA
KEGG ribosome
Adherens junction
VEGF signaling pathway
Superpathway of inositol phosphate compounds
Assembly of the pre-replicative complex
Extension of telomeres
DNA replication
0
5
10
15
20
−log$_{10}$ P−value
C
ESMO Two-Gene Sets
M/G1 transition
Regulation of mitotic cell cycle
APC/C:Cdh1 mediated Cdc20 degradation
Synthesis of DNA
Assembly of the pre-replicative complex
E2F transcription factor network
PLK1 signaling events
Signaling by ERBB4
Cyclin E events during G1/S transition
SCF(Skp2)-mediated p27/p21 degradation
0
10
20
30
40
−log$_{10}$ P−value
D
Grade Two-Gene Sets
M/G1 transition
Synthesis of DNA
Regulation of mitotic cell cycle
APC/C:Cdh1 mediated Cdc20 degradation
Cyclin E events during G1/S transition
Assembly of the pre-replicative complex
E2F transcription factor network
Aurora B signaling
KEGG small cell lung cancer
KEGG prostate cancer
0
20
40
60
−log$_{10}$ P−value
E
Type Two-Gene Sets
Signaling by SCF-KIT
Downstream signal transduction
p53 signaling pathway
TAp63 isoforms transcriptional targets
p73 transcription factor network
KEGG melanoma
KEGG chronic myeloid leukemia
B cell receptor downstream signaling
KEGG prostate cancer
Signaling by ERBB4
0
20
40
−log$_{10}$ P−value
F
OS Three-Gene Sets
G alpha (12/13) signaling events
PI3K events in ERBB2 signaling
GAB1 signalosome
KEGG endometrial cancer
Class B/2 (Secretin family receptors)
Synthesis of PIPs at plasma membrane
Amine ligand-binding receptors
RhoA signaling pathway
ErbB4 signaling events
KEGG pancreatic cancer
0
30
60
90
−log$_{10}$ P−value
G
PFS Three-Gene Sets
Semaphorin interactions
Platelet homeostasis
Fc gamma R-mediated phagocytosis
GPVI-mediated activation cascade
RhoA signaling pathway
C-MYC transcriptional repression targets
Adherens junction
Metabolism of nucleotides
Glucagon signaling in metabolic regulation
Renal aquaporins water balance regulation
0
50
100
150
−log$_{10}$ P−value
H
ESMO Three-Gene Sets
Regulation of mitotic cell cycle
APC/C:Cdh1 mediated Cdc20 degradation
Class B/2 (Secretin family receptors)
Integrin-linked kinase signaling
Aurora A signaling
PLK1 signaling events
Nuclear estrogen receptor alpha network
Aurora B signaling
Hedgehog signaling pathway
Progesterone-mediated oocyte maturation
0
50
100
150
−log$_{10}$ P−value
I
Grade Three-Gene Sets
Regulation of mitotic cell cycle
Class B/2 (Secretin family receptors)
APC/C:Cdh1 mediated Cdc20 degradation
Hedgehog signaling pathway
Signaling events mediated by the Hedgehog family
Progesterone-mediated oocyte maturation
APC/C:Cdc20 mediated mitotic proteins degradation
Aurora A signaling
Autodegradation of Cdh1 by Cdh1:APC/C
Integrin-linked kinase signaling
0
50
100
150
200
250
−log$_{10}$ P−value

*Figure 3-10: Enriched canonical pathways amongst the top 1% of gene sets identified as most predictive compared to the bottom 1%*

*(A) two-gene sets predicting overall survival at 5 years; (B) two-gene sets predicting progression-free survival at 5 years; (C) two-gene sets predicting ESMO risk classification; (D) two-gene sets predicting histological grade; (E) two-gene sets predicting histology type; (F) three-gene sets predicting overall survival at 5 years; (G) three-gene sets predicting progression-free survival at 5 years; (H) three-gene sets predicting ESMO risk classification; (I) three-gene sets predicting histological grade.*

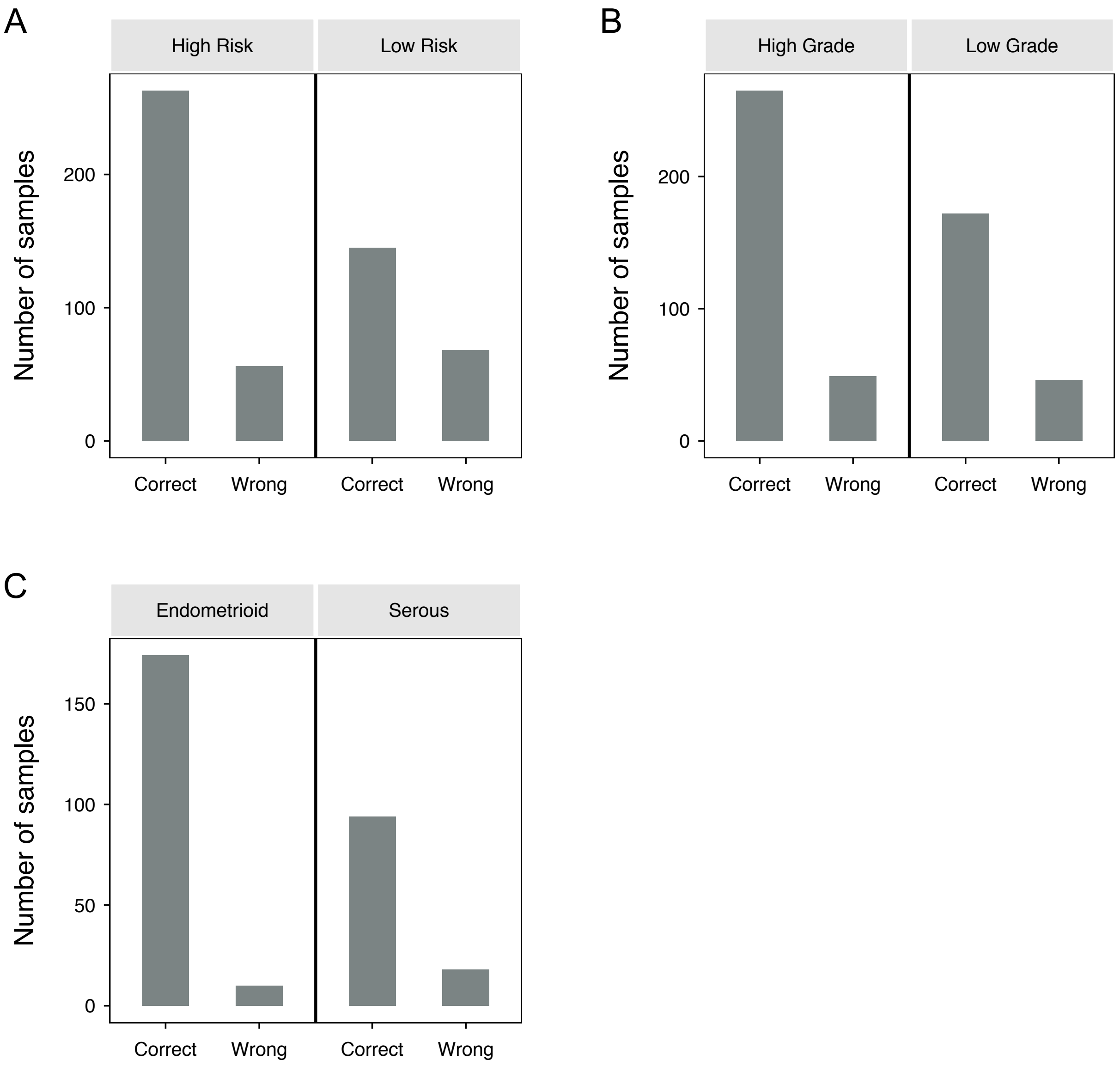
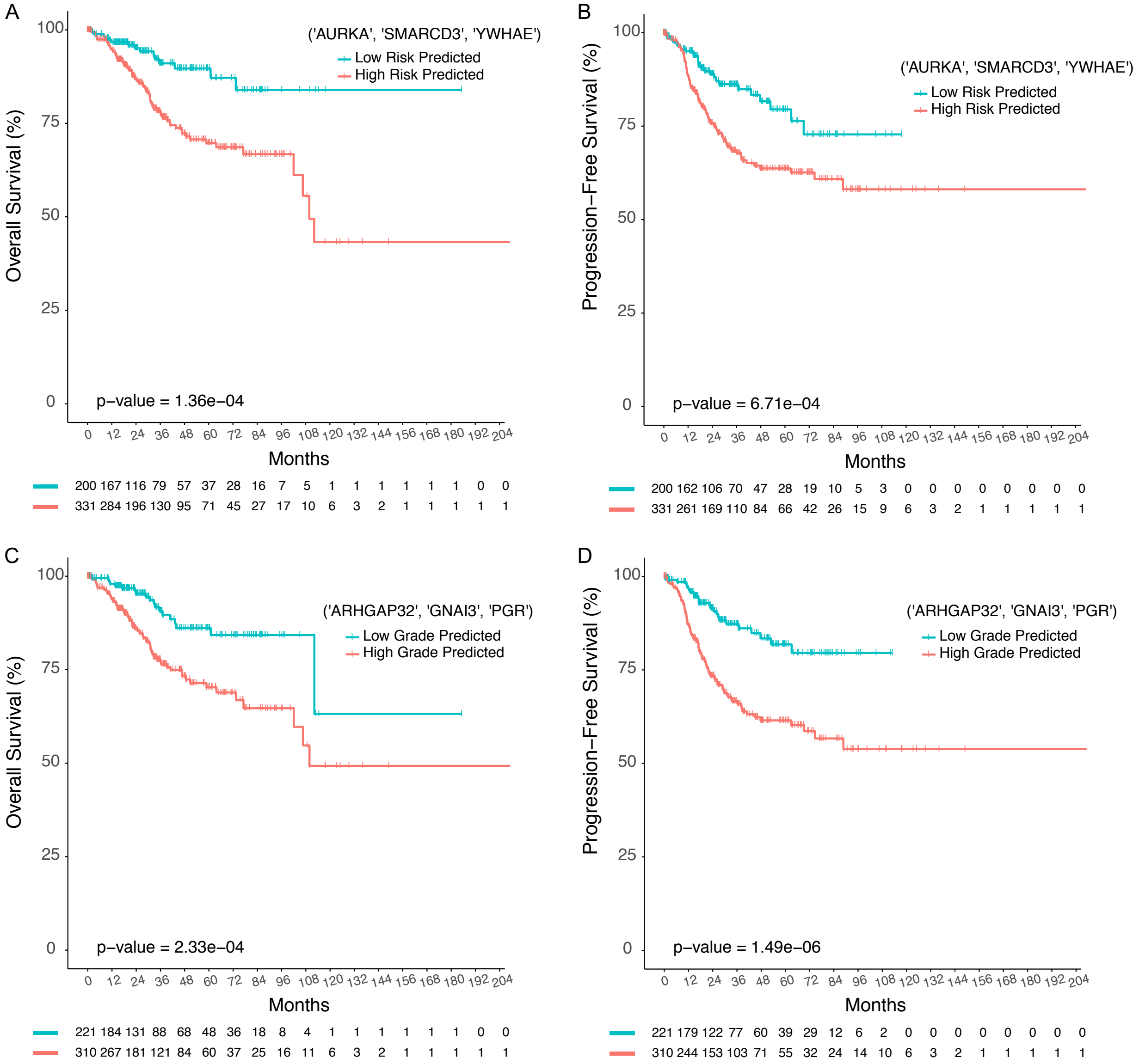


*Figure 3-11: Number of accurate predictions in each binary class using the highest predictive accuracy three-gene set for ESMO risk, histological grade and type*

*(A) ESMO risk classification; (B) Histological grade; (C) Histology type. Results indicate non-biased utility for predicting both classes across prognostic factors.*


A
('AURKA', 'SMARCD3', 'YWHAE')
Low Risk Predicted
High Risk Predicted
Overall Survival (%)
100
75
50
25
0
p−value = 1.36e−04
0 12 24 36 48 60 72 84 96 108 120 132 144 156 168 180 192 204
Months
200 167 116 79 57 37 28 16 7 5 1 1 1 1 1 1 0 0
331 284 196 130 95 71 45 27 17 10 6 3 2 1 1 1 1 1
B
('AURKA', 'SMARCD3', 'YWHAE')
Low Risk Predicted
High Risk Predicted
Progression−Free Survival (%)
100
75
50
25
0
p−value = 6.71e−04
0 12 24 36 48 60 72 84 96 108 120 132 144 156 168 180 192 204
Months
200 162 106 70 47 28 19 10 5 3 0 0 0 0 0 0 0 0
331 261 169 110 84 66 42 26 15 9 6 3 2 1 1 1 1 1
C
('ARHGAP32', 'GNAI3', 'PGR')
Low Grade Predicted
High Grade Predicted
Overall Survival (%)
100
75
50
25
0
p−value = 2.33e−04
0 12 24 36 48 60 72 84 96 108 120 132 144 156 168 180 192 204
Months
221 184 131 88 68 48 36 18 8 4 1 1 1 1 1 1 0 0
310 267 181 121 84 60 37 25 16 11 6 3 2 1 1 1 1 1
D
('ARHGAP32', 'GNAI3', 'PGR')
Low Grade Predicted
High Grade Predicted
Progression−Free Survival (%)
100
75
50
25
0
p−value = 1.49e−06
0 12 24 36 48 60 72 84 96 108 120 132 144 156 168 180 192 204
Months
221 179 122 77 60 39 29 12 6 2 0 0 0 0 0 0 0 0
310 244 153 103 71 55 32 24 14 10 6 3 2 1 1 1 1 1

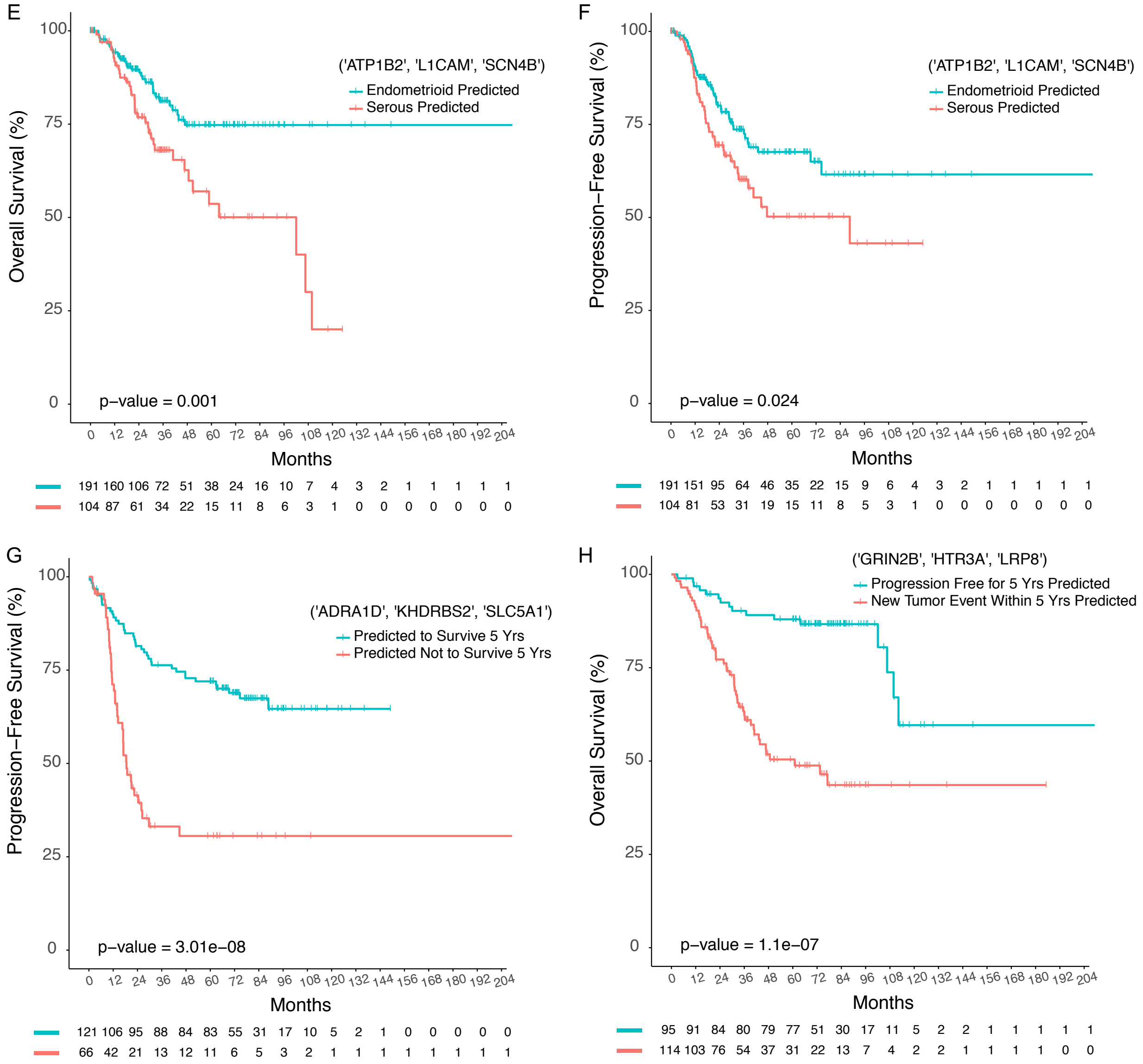


*Figure 3-12: Kaplan-Meier overall and progression-free survival curves for patients segregated into two groups based upon best gene set predictions for all prognostic factors*

*(A) Overall survival and (B) Progression-free survival for gene set AURKA, SMARCD3 and YWHAE predicting ESMO risk classification; (C) Overall survival and (D) Progression-free survival for gene set ARHGAP32, GNAI3 and PGR predicting histological grade; (E) Overall survival and (F) Progression-free survival for gene set ATP1B2, L1CAM and SCN4B predicting*

*histology type; (G) Progression-free survival for gene set ADRA1D, KHDRBS2 and SLC5A1 predicting 5-year OS; (H) Overall survival for gene set GRIN2B, HTR3A and LRP8 predicting 5 year PFS.*

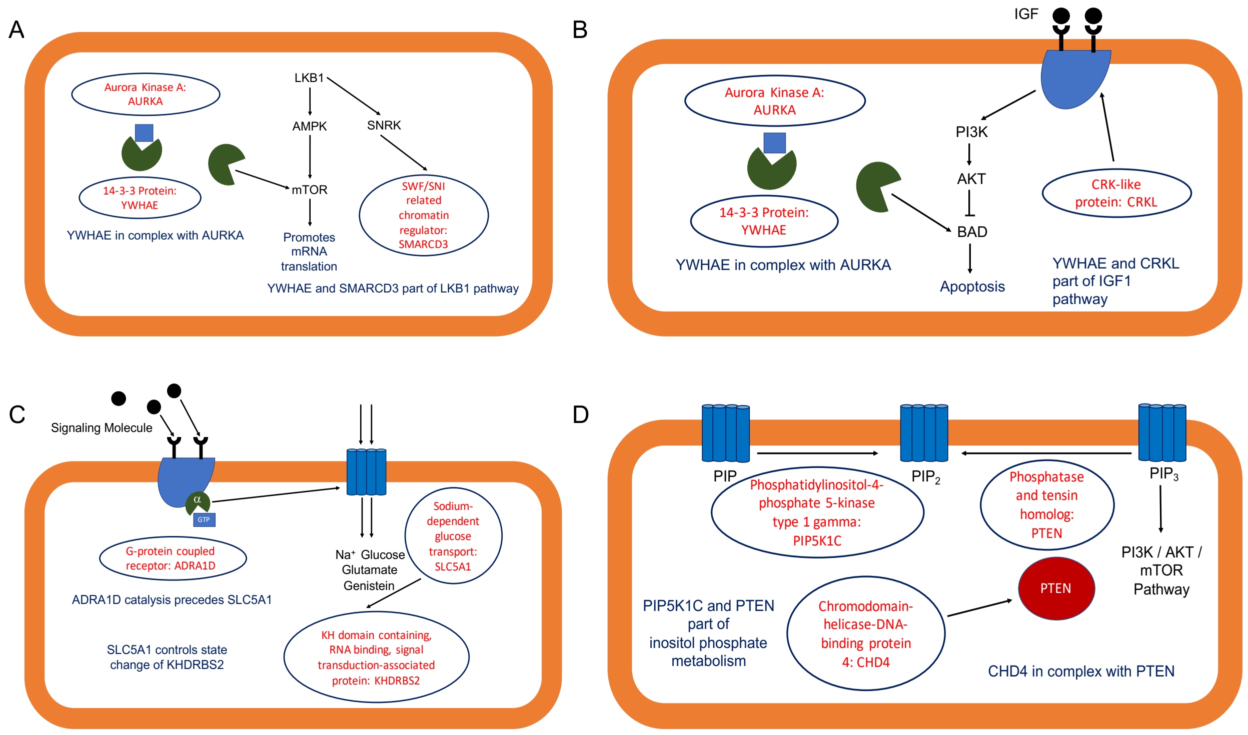

A
Aurora Kinase A: AURKA
14-3-3 Protein: YWHAE
YWHAE in complex with AURKA
LKB1
AMPK
SNRK
mTOR
Promotes mRNA translation
SWF/SNI related chromatin regulator: SMARCD3
YWHAE and SMARCD3 part of LKB1 pathway
B
IGF
Aurora Kinase A: AURKA
14-3-3 Protein: YWHAE
YWHAE in complex with AURKA
PI3K
AKT
BAD
Apoptosis
CRK-like protein: CRKL
YWHAE and CRKL part of IGF1 pathway
C
Signaling Molecule
α
GTP
G-protein coupled receptor: ADRA1D
ADRA1D catalysis precedes SLC5A1
Na+ Glucose Glutamate Genistein
Sodium-dependent glucose transport: SLC5A1
SLC5A1 controls state change of KHDRBS2
KH domain containing, RNA binding, signal transduction-associated protein: KHDRBS2
D
PIP
PIP2
PIP3
Phosphatidylinositol-4-phosphate 5-kinase type 1 gamma: PIP5K1C
Phosphatase and tensin homolog: PTEN
PI3K / AKT / mTOR Pathway
PTEN
PIP5K1C and PTEN part of inositol phosphate metabolism
Chromodomain-helicase-DNA-binding protein 4: CHD4
CHD4 in complex with PTEN

*Figure 3-13: Proposed biochemical interactions within gene sets identified as highly predictive of ESMO risk and 5-year overall survival*

*(A) ESMO risk classification with genes AURKA, SMARCD3 and YWHAE; (B) ESMO risk classification with genes AURKA, CRKL and YWHAE; (C) 5-year overall survival with genes ADRA1D, KHDRBS2 and SLC5A1; (D) 5-year overall survival with genes CHD4, PIP5K1C and PTEN.*

*Table 3-6: Evaluating Different Classifiers Based on Run Time and Average Accuracy for All Prognostic Factors*

| | OS | | PFS | | ESMO | | Grade | | Type | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Run Time (sec) | Accuracy | Run Time (sec) | Accuracy | Run Time (sec) | Accuracy | Run Time (sec) | Accuracy | Run Time (sec) | Accuracy |
| Linear SVM | 7.917 | 0.625 | 8.937 | 0.599 | 53.785 | 0.664 | 32.906 | 0.788 | 8.855 | 0.895 |
| Nearest Neighbors | 1.565 | 0.621 | 1.770 | 0.593 | 8.807 | 0.635 | 9.286 | 0.692 | 3.677 | 0.831 |
| RBF SVM | 9.589 | 0.578 | 9.420 | 0.550 | 59.974 | 0.600 | 63.503 | 0.590 | 19.236 | 0.622 |
| Random Forest | 0.332 | 0.617 | 0.317 | 0.559 | 0.687 | 0.692 | 0.668 | 0.740 | 0.419 | 0.743 |
| AdaBoost | 138.645 | 0.658 | 163.856 | 0.624 | 478.625 | 0.718 | 457.569 | 0.816 | 230.887 | 0.871 |
| ExtraTrees | 4.190 | 0.685 | 5.143 | 0.674 | 13.366 | 0.765 | 11.262 | 0.834 | 5.600 | 0.909 |
| AdaBoost ExtraTrees | 4.389 | 0.692 | 5.273 | 0.660 | 11.840 | 0.756 | 11.610 | 0.838 | 5.610 | 0.895 |
| Neural Net | 4.200 | 0.514 | 5.205 | 0.508 | 23.490 | 0.556 | 24.567 | 0.660 | 12.192 | 0.739 |
| QDA | 2.645 | 0.575 | 3.348 | 0.563 | 7.728 | 0.539 | 6.626 | 0.547 | 4.083 | 0.547 |
| Gaussian Process | 67.123 | 0.422 | 83.475 | 0.550 | 451.743 | 0.400 | 426.376 | 0.410 | 165.177 | 0.622 |
| Decision Tree | 10.289 | 0.630 | 12.205 | 0.563 | 43.810 | 0.673 | 38.169 | 0.735 | 18.700 | 0.821 |

*Table 3-7: Random Forest Hyperparameters Chosen During Fine Tuning in Two Stages*

| Random Forest | OS | | PFS | | ESMO | | Grade | | Type | |
|---|---|---|---|---|---|---|---|---|---|---|
| Parameters | 1st Stage | 2nd Stage | 1st Stage | 2nd Stage | 1st Stage | 2nd Stage | 1st Stage | 2nd Stage | 1st Stage | 2nd Stage |
| Maximum Depth | 10 | 11 | 100 | 100 | 110 | 110 | 100 | 110 | 100 | 90 |
| Number of Estimators | 50 | 49 | 20 | 22 | 20 | 20 | 20 | 21 | 50 | 51 |
| Max Features | 10 | 10 | 50 | 50 | 2 | 3 | 2 | 3 | 2 | 2 |
| Run Time (sec) | 1.118 | 1.106 | 1.035 | 1.161 | 0.841 | 0.865 | 0.780 | 0.845 | 0.861 | 0.978 |
| Accuracy | 0.696 | 0.696 | 0.665 | 0.670 | 0.735 | 0.737 | 0.801 | 0.809 | 0.854 | 0.861 |

*Table 3-8: Prediction Accuracy of Prognostic Factors Using All Genes with Tuned Random Forest Classifier*

| | OS | PFS | ESMO | Grade | Type |
|---|---|---|---|---|---|
| Prediction Accuracy | 0.70 | 0.67 | 0.74 | 0.81 | 0.86 |

## 3.7 Acknowledgements

The authors wish to thank the UCLA Department of Pathology and Dr. Jonathan W. Said for his help in reviewing this study and providing valuable contributions. Much gratitude is extended to the UCLA Computer Science Department and Prof. Richard Korf for the fruitful discussions and his expertise in heuristic search that made this work possible. Dr. Luo Mi is acknowledged for his help in improving this manuscript. Dr. Mirrian Ho is appreciated for the numerous early discussions that developed our thinking for this work.

# 4 References

Abu-Rustum, N.R., Zhou, Q., Gomez, J.D., Alektiar, K.M., Hensley, M.L., Soslow, R.A., Levine, D.A., Chi, D.S., Barakat, R.R., and Iasonos, A. (2010). A nomogram for predicting overall survival of women with endometrial cancer following primary therapy: Toward improving individualized cancer care. Gynecol. Oncol. *116*, 399–403.

Aharoni, A., Gaidukov, L., Khersonsky, O., Gould, S.M., Roodveldt, C., and Tawfik, D.S. (2005). The “evolvability” of promiscuous protein functions. Nat. Genet. *37*, 73–76.

Albitar, L., Carter, M.B., Davies, S., and Leslie, K.K. (2007). Consequences of the loss of p53, RB1, and PTEN: Relationship to gefitinib resistance in endometrial cancer. Gynecol. Oncol. *106*, 94–104.

Ashton, K., Proietto, A., Otton, G., Symonds, I., McEvoy, M., Attia, J., Gilbert, M., Hamann, U., and Scott, R. (2009). Estrogen receptor polymorphisms and the risk of endometrial cancer: Estrogen receptor polymorphisms and the risk of endometrial cancer. BJOG Int. J. Obstet. Gynaecol. *116*, 1053–1061.

Atsumi, S., Hanai, T., and Liao, J.C. (2008). Non-fermentative pathways for synthesis of branched-chain higher alcohols as biofuels. Nature *451*, 86–89.

Atsumi, S., Wu, T.-Y., Eckl, E.-M., Hawkins, S.D., Buelter, T., and Liao, J.C. (2010). Engineering the isobutanol biosynthetic pathway in Escherichia coli by comparison of three aldehyde reductase/alcohol dehydrogenase genes. Appl. Microbiol. Biotechnol. *85*, 651–657.

Backes, F.J., Walker, C.J., Goodfellow, P.J., Hade, E.M., Agarwal, G., Mutch, D., Cohn, D.E., and Suarez, A.A. (2016). Estrogen receptor-alpha as a predictive biomarker in endometrioid endometrial cancer. Gynecol. Oncol. *141*, 312–317.

Balsan, G., Astolfi, V., Benazzi, T., Meireles, M.A.A., Maugeri, F., Di Luccio, M., Dal Prá, V., Mossi, A.J., Treichel, H., and Mazutti, M.A. (2012). Characterization of a commercial cellulase for hydrolysis of agroindustrial substrates. Bioprocess Biosyst. Eng. *35*, 1229–1237.

Berthold, C.L., Gocke, D., Wood, M.D., Leeper, F.J., Pohl, M., and Schneider, G. (2007). Structure of the branched-chain keto acid decarboxylase (KdcA) from *Lactococcus lactis* provides insights into the structural basis for the chemoselective and enantioselective carboligation reaction. Acta Crystallogr. D Biol. Crystallogr. *63*, 1217–1224.

Biasini, M., Bienert, S., Waterhouse, A., Arnold, K., Studer, G., Schmidt, T., Kiefer, F., Cassarino, T.G., Bertoni, M., Bordoli, L., et al. (2014). SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. Nucleic Acids Res. *42*, W252–W258.

Bloom, J.D., Labthavikul, S.T., Otey, C.R., and Arnold, F.H. (2006). Protein stability promotes evolvability. Proc. Natl. Acad. Sci. *103*, 5869–5874.

Bokhman, J.V. (1983). Two pathogenetic types of endometrial carcinoma. Gynecol. Oncol. *15*, 10–17.

Bosse, T., ter Haar, N.T., Seeber, L.M., Diest, P.J. v, Hes, F.J., Vasen, H.F., Nout, R.A., Creutzberg, C.L., Morreau, H., and Smit, V.T. (2013). Loss of ARID1A expression and its relationship with PI3K-Akt pathway alterations, TP53 and microsatellite instability in endometrial cancer. Mod. Pathol. *26*, 1525–1535.

Cann, A.F., and Liao, J.C. (2008). Production of 2-methyl-1-butanol in engineered Escherichia coli. Appl. Microbiol. Biotechnol. *81*, 89–98.

Carmon, K.S., and Loose, D.S. (2008). Secreted Frizzled-Related Protein 4 Regulates Two Wnt7a Signaling Pathways and Inhibits Proliferation in Endometrial Cancer Cells. Mol. Cancer Res. *6*, 1017–1028.

Cerami, E.G., Gross, B.E., Demir, E., Rodchenkov, I., Babur, O., Anwar, N., Schultz, N., Bader, G.D., and Sander, C. (2011). Pathway Commons, a web resource for biological pathway data. Nucleic Acids Res. *39*, D685–D690.

Colombo, N., Preti, E., Landoni, F., Carinelli, S., Colombo, A., Marini, C., Sessa, C., and on behalf of the ESMO Guidelines Working Group (2013). Endometrial cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. Ann. Oncol. *24*, vi33–vi38.

Connor, M.R., and Liao, J.C. (2008). Engineering of an Escherichia coli Strain for the Production of 3-Methyl-1-Butanol. Appl. Environ. Microbiol. *74*, 5769–5775.

Creasman, W., Odicino, F., Maisonneuve, P., Quinn, M., Beller, U., Benedet, J., Heintz, A., Ngan, H., and Pecorelli, S. Carcinoma of the Corpus Uteri. 39.

Decruze, S.B., and Green, J.A. (2007). Hormone therapy in advanced and recurrent endometrial cancer: a systematic review. Int. J. Gynecol. Cancer *17*, 964–978.

Dellinger, T.H., Smith, D.D., Ouyang, C., Warden, C.D., Williams, J.C., and Han, E.S. (2016). L1CAM is an independent predictor of poor survival in endometrial cancer — An analysis of The Cancer Genome Atlas (TCGA). Gynecol. Oncol. *141*, 336–340.

Doll, A., Abal, M., Rigau, M., Monge, M., Gonzalez, M., Demajo, S., Colás, E., Llauradó, M., Alazzouzi, H., Planagumá, J., et al. (2008). Novel molecular profiles of endometrial cancer—new light through old windows. J. Steroid Biochem. Mol. Biol. *108*, 221–229.

Drummond, D.A., Silberg, J.J., Meyer, M.M., Wilke, C.O., and Arnold, F.H. (2005). On the conservative nature of intragenic recombination. Proc. Natl. Acad. Sci. *102*, 5380–5385.

Edqvist, P.-H.D., Huvila, J., Forsström, B., Talve, L., Carpén, O., Salvesen, H.B., Krakstad, C., Grénman, S., Johannesson, H., Ljungqvist, O., et al. (2015). Loss of ASRGL1 expression is an independent biomarker for disease-specific survival in endometrioid endometrial carcinoma. Gynecol. Oncol. *137*, 529–537.

Eiben, C.B., Siegel, J.B., Bale, J.B., Cooper, S., Khatib, F., Shen, B.W., Players, F., Stoddard, B.L., Popovic, Z., and Baker, D. (2012). Increased Diels-Alderase activity through backbone remodeling guided by Foldit players. Nat. Biotechnol. *30*, 190–192.

Ellrott, K., Bailey, M.H., Saksena, G., Covington, K.R., Kandoth, C., Stewart, C., Hess, J., Ma, S., Chiotti, K.E., McLellan, M., et al. (2018). Scalable Open Science Approach for Mutation Calling of Tumor Exomes Using Multiple Genomic Pipelines. Cell Syst. *6*, 271-281.e7.

Ferlay, J., Soerjomataram, I., Dikshit, R., Eser, S., Mathers, C., Rebelo, M., Parkin, D.M., Forman, D., and Bray, F. (2015). Cancer incidence and mortality worldwide: Sources, methods and major patterns in GLOBOCAN 2012: Globocan 2012. Int. J. Cancer *136*, E359–E386.

Furey, W., Arjunan, P., Chen, L., Sax, M., Guo, F., and Jordan, F. Structure-function relationships and £exible tetramer assembly in pyruvate decarboxylase revealed by analysis of crystal structures. Biochim. Biophys. Acta 18.

Getz, G., Gabriel, S.B., Cibulskis, K., Lander, E., Sivachenko, A., Sougnez, C., Lawrence, M., Kandoth, C., Dooling, D., Fulton, R., et al. (2013). Integrated genomic characterization of endometrial carcinoma. Nature *497*, 67–73.

Gibson, D.G., Young, L., Chuang, R.-Y., Venter, J.C., Hutchison, C.A., and Smith, H.O. (2009). Enzymatic assembly of DNA molecules up to several hundred kilobases. Nat. Methods *6*, 343–345.

Gielen, S.C.J.P., Hanekamp, E.E., Hanifi-Moghaddam, P., Sijbers, A.M., van GOOL, A.J., Burger, C.W., Blok, L.J., and Huikeshoven, F.J. (2006). Growth regulation and transcriptional activities of estrogen and progesterone in human endometrial cancer cells. Int. J. Gynecol. Cancer *16*, 110–120.

Gocke, D., Nguyen, C.L., Pohl, M., Stillger, T., Walter, L., and Müller, M. (2007). Branched-Chain Keto Acid Decarboxylase fromLactococcus lactis (KdcA), a Valuable Thiamine Diphosphate-Dependent Enzyme for Asymmetric C-C Bond Formation. Adv. Synth. Catal. *349*, 1425–1435.

Gront, D., Kulp, D.W., Vernon, R.M., Strauss, C.E.M., and Baker, D. (2011). Generalized Fragment Picking in Rosetta: Design, Protocols and Applications. PLoS ONE *6*, e23294.

Guo, H.H., Choe, J., and Loeb, L.A. (2004). Protein tolerance to random amino acid change. Proc. Natl. Acad. Sci. *101*, 9205–9210.

Guterl, J.-K., Garbe, D., Carsten, J., Steffler, F., Sommer, B., Reiße, S., Philipp, A., Haack, M., Rühmann, B., Koltermann, A., et al. (2012). Cell-Free Metabolic

Engineering: Production of Chemicals by Minimized Reaction Cascades. ChemSusChem *5*, 2165–2172.

Gutierrez-Hartmann, A., Duval, D.L., and Bradford, A.P. (2007). ETS transcription factors in endocrine systems. Trends Endocrinol. Metab. *18*, 150–158.

Higashide, W., Li, Y., Yang, Y., and Liao, J.C. (2011). Metabolic Engineering of Clostridium cellulolyticum for Production of Isobutanol from Cellulose. Appl. Environ. Microbiol. *77*, 2727–2733.

Hoadley, K.A., Yau, C., Hinoue, T., Wolf, D.M., Lazar, A.J., Drill, E., Shen, R., Taylor, A.M., Cherniack, A.D., Thorsson, V., et al. (2018). Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer. Cell *173*, 291-304.e6.

Hutter, C., and Zenklusen, J.C. (2018). The Cancer Genome Atlas: Creating Lasting Value beyond Its Data. Cell *173*, 283–285.

Kawada, K., and Taketo, M.M. (2011). Significance and Mechanism of Lymph Node Metastasis in Cancer Progression. Cancer Res. *71*, 1214–1218.

Ko, C.-H., Tsai, C.-H., Lin, P.-H., Chang, K.-C., Tu, J., Wang, Y.-N., and Yang, C.-Y. (2010). Characterization and pulp refining activity of a Paenibacillus campinasensis cellulase expressed in Escherichia coli. Bioresour. Technol. *101*, 7882–7888.

Koga, J., Adachi, T., and Hidaka, H. (1992). Purification and characterization of indolepyruvate decarboxylase. A novel enzyme for indole-3-acetic acid biosynthesis in Enterobacter cloacae. J. Biol. Chem. *267*, 15823–15828.

Korkegian, A. (2005). Computational Thermostabilization of an Enzyme. Science *308*, 857–860.

Langfelder, P., and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics *9*, 559.

Lawrence, M.S., Stojanov, P., Mermel, C.H., Robinson, J.T., Garraway, L.A., Golub, T.R., Meyerson, M., Gabriel, S.B., Lander, E.S., and Getz, G. (2014). Discovery and saturation analysis of cancer genes across 21 tumour types. Nature *505*, 495–501.

Lee, C.-H., Ou, W.-B., Marino-Enriquez, A., Zhu, M., Mayeda, M., Wang, Y., Guo, X., Brunner, A.L., Amant, F., French, C.A., et al. (2012a). 14-3-3 fusion oncogenes in high-grade endometrial stromal sarcoma. Proc. Natl. Acad. Sci. *109*, 929–934.

Lee, C.-H., Zhu, M., Ali, R.H., Gilks, C.B., Debiec-Rychter, M., and Nucci, M.R. (2012b). The Clinicopathologic Features of YWHAE-FAM22 Endometrial Stromal Sarcomas: A Histologically High-grade and Clinically Aggressive Tumor. Am J Surg Pathol *36*, 13.

Lee, Y.-J., Kim, B.-K., Lee, B.-H., Jo, K.-I., Lee, N.-K., Chung, C.-H., Lee, Y.-C., and Lee, J.-W. (2008). Purification and characterization of cellulase produced by Bacillus amyoliquefaciens DL-3 utilizing rice hull. Bioresour. Technol. *99*, 378–386.

Li, H., Opgenorth, P.H., Wernick, D.G., Rogers, S., Wu, T.-Y., Higashide, W., Malati, P., Huo, Y.-X., Cho, K.M., and Liao, J.C. (2012). Integrated Electromicrobial Conversion of CO2 to Higher Alcohols. Science *335*, 1596–1596.

Lin, P.P., Rabe, K.S., Takasumi, J.L., Kadisch, M., Arnold, F.H., and Liao, J.C. (2014). Isobutanol production at elevated temperatures in thermophilic Geobacillus thermoglucosidasius. Metab. Eng. *24*, 1–8.

Lin, P.P., Mi, L., Morioka, A.H., Yoshino, K.M., Konishi, S., Xu, S.C., Papanek, B.A., Riley, L.A., Guss, A.M., and Liao, J.C. (2015). Consolidated bioprocessing of cellulose to isobutanol using Clostridium thermocellum. Metab. Eng. *31*, 44–52.

Liu, J., and Xia, W. (2006). Purification and characterization of a bifunctional enzyme with chitosanase and cellulase activity from commercial cellulase. Biochem. Eng. J. *30*, 82–87.

Liu, J., Lichtenberg, T., Hoadley, K.A., Poisson, L.M., Lazar, A.J., Cherniack, A.D., Kovatich, A.J., Benz, C.C., Levine, D.A., Lee, A.V., et al. (2018). An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics. Cell *173*, 400-416.e11.

Lowerre, B., and Reddy, R. (1990). THE HARPY SPEECH UNDERSTANDING SYSTEM. In Readings in Speech Recognition, A. Waibel, and K.-F. Lee, eds. (San Francisco: Morgan Kaufmann), pp. 576–586.

Ma, B.B.Y., Oza, A., Eisenhauer, E., Stanimir, G., Carey, M., Chapman, W., Latta, E., Sidhu, K., Powers, J., Walsh, W., et al. (2004). The activity of letrozole in patients with advanced or recurrent endometrial cancer and correlation with biological markers - a study of the National Cancer Institute of Canada Clinical Trials Group. Int. J. Gynecol. Cancer *14*, 650–658.

Miyamoto, D.T., Zheng, Y., Wittner, B.S., Lee, R.J., Zhu, H., Broderick, K.T., Desai, R., Fox, D.B., Brannigan, B.W., Trautwein, J., et al. (2015). RNA-Seq of single prostate CTCs implicates noncanonical Wnt signaling in antiandrogen resistance. Science *349*, 1351–1356.

Moreno-Bueno, G., and Hardisson, D. Abnormalities of the APC/b-catenin pathway in endometrial cancer. 10.

Murali, R., Soslow, R.A., and Weigelt, B. (2014). Classification of endometrial carcinoma: more than two types. Lancet Oncol. *15*, e268–e278.

NIH Intramural Sequencing Center (NISC) Comparative Sequencing Program, Le Gallo, M., O’Hara, A.J., Rudd, M.L., Urick, M.E., Hansen, N.F., O’Neil, N.J., Price, J.C., Zhang, S., England, B.M., et al. (2012). Exome sequencing of serous endometrial tumors identifies recurrent somatic mutations in chromatin-remodeling and ubiquitin ligase complex genes. Nat. Genet. *44*, 1310–1315.

Padmanabhan, R.A., Nirmala, L., Murali, M., and Laloraya, M. (2011). CrkL is a Co-Activator of Estrogen Receptor α That Enhances Tumorigenic Potential in Cancer. Mol. Endocrinol. *25*, 1499–1512.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. Scikit-learn: Machine Learning in Python. Mach. Learn. PYTHON 6.

Plaza, M., Fernandez de Palencia, P., Pelaez, C., and Requena, T. (2004). Biochemical and molecular characterization of alpha-ketoisovalerate decarboxylase, an enzyme involved in the formation of aldehydes from amino acids by *Lactococcus lactis*. FEMS Microbiol. Lett. *238*, 367–374.

Raj, K.C., Talarico, L.A., Ingram, L.O., and Maupin-Furlow, J.A. (2002). Cloning and Characterization of the Zymobacter palmae Pyruvate Decarboxylase Gene (pdc) and Comparison to Bacterial Homologues. Appl. Environ. Microbiol. *68*, 2869–2876.

Reddy, R. Foundations and Grand Challenges of Artificial Intelligence. 13.

Risinger, J.I., Hayes, A.K., and Barrett', J.C. PTEN/MMAC1 Mutations in Endometrial Cancers. 4.

Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., and Smyth, G.K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. *43*, e47–e47.

Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics *26*, 139–140.

Romero, P.A., and Arnold, F.H. (2009). Exploring protein fitness landscapes by directed evolution. Nat. Rev. Mol. Cell Biol. *10*, 866–876.

Rouillard, A.D., Gundersen, G.W., Fernandez, N.F., Wang, Z., Monteiro, C.D., McDermott, M.G., and Ma'ayan, A. (2016). The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins. Database *2016*, baw100.

Saeys, Y., Inza, I., and Larranaga, P. (2007). A review of feature selection techniques in bioinformatics. Bioinformatics *23*, 2507–2517.

Salvesen, H.B., Haldorsen, I.S., and Trovik, J. (2012). Markers for individualised therapy in endometrial carcinoma. Lancet Oncol. *13*, e353–e361.

Schrodinger LLC (2010). Schrödinger, L. L. C. (2010) The {PyMOL} Molecular Graphics System, version 1.3r1, Schrodinger, LLC.

Schwarz, W.H. (2001). The cellulosome and cellulose degradation by anaerobic bacteria. Appl. Microbiol. Biotechnol. *56*, 634–649.

Shen, C.R., and Liao, J.C. (2008). Metabolic engineering of Escherichia coli for 1-butanol and 1-propanol production via the keto-acid pathways. Metab. Eng. *10*, 312–320.

Slomovitz, B.M., and Coleman, R.L. (2012). The PI3K/AKT/mTOR Pathway as a Therapeutic Target in Endometrial Cancer. Clin. Cancer Res. *18*, 5856–5864.

Smith, K.M., Cho, K.-M., and Liao, J.C. (2010). Engineering Corynebacterium glutamicum for isobutanol production. Appl. Microbiol. Biotechnol. *87*, 1045–1055.

Soh, L.M.J., Mak, W.S., Lin, P.P., Mi, L., Chen, F.Y.-H., Damoiseaux, R., Siegel, J.B., and Liao, J.C. (2017). Engineering a Thermostable Keto Acid Decarboxylase Using Directed Evolution and Computationally Directed Protein Design. ACS Synth. Biol. *6*, 610–618.

Song, Y., DiMaio, F., Wang, R.Y.-R., Kim, D., Miles, C., Brunette, T., Thompson, J., and Baker, D. (2013). High-Resolution Comparative Modeling with RosettaCM. Structure *21*, 1735–1742.

Soslow, R.A. (2013). High-grade endometrial carcinomas - strategies for typing. Histopathology *62*, 89–110.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., et al. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. *102*, 15545–15550.

Thompson, J., and Baker, D. (2011). Incorporation of evolutionary information into Rosetta comparative modeling. Proteins Struct. Funct. Bioinforma. *79*, 2380–2388.

Thompson, A.H., Studholme, D.J., Green, E.M., and Leak, D.J. (2008). Heterologous expression of pyruvate decarboxylase in Geobacillus thermoglucosidasius. Biotechnol. Lett. *30*, 1359–1365.

Tian, J., Wang, P., Gao, S., Chu, X., Wu, N., and Fan, Y. (2010). Enhanced thermostability of methyl parathion hydrolase from Ochrobactrum sp. M231 by rational engineering of a glycine to proline mutation: Enhanced thermostability of methyl parathion hydrolase. FEBS J. *277*, 4901–4908.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel, H., Salzberg, S.L., Rinn, J.L., and Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat. Protoc. *7*, 562–578.

Umene, K., Yanokura, M., Banno, K., Irie, H., Adachi, M., Iida, M., Nakamura, K., Nogami, Y., Masuda, K., Kobayashi, Y., et al. (2015). Aurora kinase A has a significant role as a therapeutic target and clinical biomarker in endometrial cancer. Int. J. Oncol. *46*, 1498–1506.

Wang, L., Chu, F., and Xie, W. (2007). Accurate Cancer Classification Using Expressions of Very Few Genes. IEEE/ACM Trans. Comput. Biol. Bioinform. *4*, 40–53.

Wang, L., Wang, Y., and Chang, Q. (2016). Feature selection methods for big data bioinformatics: A survey from the search perspective. Methods *111*, 21–31.

Wong, Y.F., Cheung, T.H., Lo, K.W.K., Yim, S.F., Siu, N.S.S., Chan, S.C.S., Ho, T.W.F., Wong, K.W.Y., Yu, M.Y., Wang, V.W., et al. (2007). Identification of molecular markers and signaling pathway in endometrial cancer in Hong Kong Chinese women by genome-wide gene expression profiling. Oncogene *26*, 1971–1982.

Wormke, M., Castro-Rivera, E., Chen, I., and Safe, S. (2000). Estrogen and aryl hydrocarbon receptor expression and crosstalk in human Ishikawa endometrial cancer cells. J. Steroid Biochem. Mol. Biol. *72*, 197–207.

Wu, I., and Arnold, F.H. (2013). Engineered thermostable fungal Cel6A and Cel7A cellobiohydrolases hydrolyze cellulose efficiently at elevated temperatures. Biotechnol. Bioeng. *110*, 1874–1883.