AN EVALUATION OF BEST PRACTICES FOR CLINICAL LABORATORY IMPROVEMENT AMENDMENTS (CLIA) BIOINFORMATICS PIPELINES

By

Angie B. McGraw

A THESIS

Presented to the Department of Medical Informatics and Clinical Epidemiology
and the Oregon Health & Science University
School of Medicine
in partial fulfillment of
the requirements for the degree of

Master of Science

March 2023

School of Medicine

Oregon Health & Science University

CERTIFICATE OF APPROVAL

This is to certify that the Master's thesis of

Angie McGraw

has been approved

Mentor/Advisor

Member

Member

Member

Member

TABLE OF CONTENTS

C. Methods
    1. Overview
    2. Deltas Computed Based on Gene Expression Differences
        a) Log-Transformed Null Distribution and Separation by Lots
    3. Deltas Computed Based on Rank Differences
D. Results and Discussion
    1. Data Distribution Before Filtering
    2. Move Towards Filtering
    3. Deltas Computed Based on Gene Expression Differences
        a) Log-Transformed Null Distribution Data and Separation by Lots
    4. Deltas Computed Based on Rank Differences
    5. UHRs as a Reference Distribution for Disease Types
    6. Evaluation of the UHRs as an Appropriate Reference Distribution in this Use Case
E. Conclusion

The following table describes the abbreviations that are used throughout the thesis. The page on which each one is defined first is given. Acronyms that are not frequently used in the thesis are not included in the following table.

Abbreviation	Meaning
CLIA	Clinical Laboratory Improvement Amendments
NGS	Next-Generation Sequencing

ACKNOWLEDGEMENTS

I would like to give great thanks to my mentor, Dr. Shannon McWeeney, who has made this thesis possible and who has supported me every step of the way. I would also like to thank my committee members, Dr. Xubo Song and Dr. Sanjay Malhotra, for supporting my defense; and Dr. Christopher Suciu for his discussion on processing of the data. I would like to thank my parents and friends for their continuous support in my career.

ABSTRACT

Next-generation sequencing (NGS) technologies have provided opportunities for developing personalized treatments for patients. Given the rapid advances in technology and computational complexity, there may be gaps in best practices and guidelines to support the current challenges facing Clinical Laboratory Improvement Amendments (CLIA) laboratories. Whole-genome sequencing (WGS), an NGS technology, allows for the exponential generation of human sequencing data. Bioinformatics and computational pipelines have taken advantage of these data to generate new mechanisms to aid in predicting patient treatment responses and adjusting treatments accordingly. We highlight the potential of NGS-based bioinformatics pipelines in precision breast oncology treatments and the challenges of standardization given the historical lack of consensus on pipeline development and validation standards. This is seen as a critical issue given the need for standardization to ensure pipeline accuracy, appropriateness, and efficacy, and those to protect patient safety and ensure quality patient care. We conducted a literature review to assess current best practices for CLIA bioinformatics pipelines and identified potential gaps, and assessed the use case of determining an appropriate reference background distributions for expression signatures for clinical usage. We emphasize the complexity and heterogeneity of NGS-based clinical assays that often necessitates context specific validation of pipelines and reference distributions.

INTRODUCTION

AIM 1. LITERATURE REVIEW TO ASSESS CURRENT BEST PRACTICES FOR CLINICAL LABORATORY IMPROVEMENT AMENDMENTS (CLIA) BIOINFORMATICS PIPELINES.

LITERATURE REVIEW: HISTORY AND BACKGROUND

Based on the first part of our literature review, we briefly summarize the history of CLIA and highlight how sequencing technologies have been incorporated into laboratory testing, as well the accompanying bioinformatics workflows needed to analyze and process the data.

Clinical Laboratory Improvement Amendments (CLIA) Background

In November 1987, a Wall Street Journal article was published, "Lax Laboratories: The Pap Test Misses Much Cervical Cancer Through Labs' Errors," highlighting how patients were receiving incorrect results in routine cancer screening (Bogdanovich 1987). This situation indicated why there needed to be more guidance and regulations for clinical laboratories. In response to growing concerns regarding clinical laboratory testing, and to expand federal oversight to all laboratories conducting testing on human specimens for disease diagnosis and treatment improvement (Schwartz 1999), Clinical Laboratory Improvements Amendments (CLIA) was established in 1988. To ensure standards for laboratories generating clinical tests and analyzing human specimens, the United States federal government mandated that laboratories performing tests on human specimens must meet the guidelines entailed in CLIA (Lyon and Segal 2013). If a laboratory wishes to genetically test samples from human specimens and construct an assay for research purposes, the laboratory must obtain CLIA certification. CLIA certification certifies "that the laboratory has developed a validated clinical assay that is reproducible, precise, and accurate with established sensitivity and specificity" (Gaerig 2012).

Review of the FDA database highlights how assays evolved in complexity with the incorporation of next-generation sequencing (NGS). In the early 1990s, approved tests included ELISA (such as cytomegalovirus (CMV) ELISA, Rubella ELISA or Lyme ELISA tests) or antigen-based tests, such as Cancer antigen 125 (CA 125). In 2006, there was draft guidance on *in vitro* diagnostic multivariate index assays (IVDMIA). IVDMIAs are more complex diagnostics and consist of clinical data, an algorithm, and a threshold provided by the test developer to interpret the result. The 1st IVDMIA, MammaPrint® was approved in 2007. This is a gene expression based signature that can determine which breast cancer patients are at risk of distant recurrence following surgery (providing oncologists with a risk classification based on the patient sample). NGS provides the advantages of broader sequencing coverage and increased identification of biomarkers, but also has additional considerations with regard to cost, data storage, and compute

power (Sboner and Elemento 2016). NGS has enhanced genomics research in areas such as sequencing quality and data production (Behjati and Tarpey 2013). Massively parallel or deep sequencing can be used for an unbiased transcriptomic analysis of mRNAs, small RNAs, noncoding RNAs, genome-wide methylation assays, and high-throughput chromatin immunoprecipitation assays (Reis-Filho 2009). To be CLIA-certified, the laboratory has to maintain data management systems for the generated sequencing and other data types, entailing accurate and quality record keeping (Gaerig 2012). Data maintenance needs to be accurate and ensure governance in a clinical setting (Zhang et al. 2020). NGS data relies heavily on bioinformatic pipelines to uncover genetic variations or expression differences in patient samples. This entails additional emphasis on data infrastructure, algorithm implementation, evaluation and maintenance to support the processing and interpretation of the sequencing data.

Next-Generation Sequencing (NGS) Overview and Whole Genome Sequencing (WGS)

As precision medicine transitions from research to clinical settings, careful consideration must be given to clinical genomics application standards. NGS technologies are key in constructing pipelines that could revolutionize patient care, creating more affordable and efficacious treatments. Using parallel computing, NGS-based whole-genome sequencing (WGS) can sequence the entirety of the human genome rapidly, proving to be a powerful tool for assessing human variation. NGS has shown to be quicker and more accurate than Sanger sequencing (Straiton et al.). Millions of small fragments are sequenced in parallel, and bioinformatic analyses map the fragments to the human genome (Behjati and Tarpey 2013). The sequencing data obtained from NGS can aid in uncovering genetic mutations and markers of various diseases. Mutations include insertions, deletions, substitutions, and translocations with DNA sequence bases. Sanger sequencing is limited to substitutions, small insertions, and deletions (Behjati and Tarpey 2013). Capillary-based cancer sequencing has been utilized for more than a decade and has been limited to a few samples and a few candidate genes (Behjati and Tarpey 2013).This type of sequencing is dependent on prior knowledge of the gene or locus being investigated. In contrast, NGS can be genome-wide (i.e., WGS) allowing for unbiased discovery of disease-associated variants. With increased genome coverage and decreased sequencing time, an individual patient's genome can be sequenced and analyzed to support personalized treatments (precision medicine), with the hope that more individualized treatments can provide more effective treatments.

WGS is a powerful tool for identifying biomarkers and their roles in diseases. This tool has aided researchers in understanding the intricate details of a patient's genome, providing new avenues for personalized medicine, and predictions for how individual patients may respond to changes in therapy. There have also been international efforts to understand and characterize the diversity in human genetic variation, such as the 1000 Genomes Project which sequenced individuals from Europe, East Asia, South Asia, West Africa, and the Americas (Siva 2008). Overall, WGS has

contributed to better monitoring of mutations and disease detection, and improvements in treatments. NGS is highly dependent on complex computational data analysis infrastructure, which is part of the reason why Sanger sequencing and other sequencing technologies are widely utilized for the validation of NGS results (Roy et al. 2018).

Bioinformatics: Background, Pipelines, and Usage in Next-Generation Sequencing (NGS)

Bioinformatics integrates biology, computer science, and statistics, to handle large amounts of data. Bioinformatics algorithms executed in a predefined sequence are known as pipelines, such as the algorithms used to process NGS results (Roy et al. 2018). Pipelines are often run in an automated or batch mode, meaning the pipeline implementation and results need to be accurate and valid. Bioinformatics pipelines in a clinical setting that fail to meet standards and have been improperly validated can detrimentally impact the health and well-being of patients. NGS results are completely reliant on bioinformatics pipelines for processing and analysis. An NGS bioinformatics pipeline to assess genetic variants (one of the most commonly conducted in clinical laboratories) generally consists of the following steps: sequence generation, sequence alignment, variant calling, variant filtering, variant annotation, and variant prioritization (Roy et al. 2018).

Sequence Generation and Alignment. Sequence generation is the process of identifying sequences of nucleotides from short DNA fragments, also known as raw reads, from a sample (Roy et al. 2018). There are currently three widely platforms for massively parallel DNA sequencing read production: (1) the Roche/454 FLX (Margulies et al. 2005), (2) the Illumina/Solexa Genome Analyzer (Bentley et al. 2006), and (3) the Applied Biosystems SOLiD™ System (Pandey et al. 2008). Each DNA sequencing technology aims to amplify single strands of a fragment library and then perform sequencing reactions on the amplified strands (Mardis 2008). Each nucleotide sequence sequenced in the raw reads is given a platform-specific Phred-like quality score (Roy et al. 2018). Reads with low Phred scores are filtered out. Platforms have different thresholds for the Phred scores; thresholds too high or too low may cause loss of data or introduction of errors (Liao et al. 2017). Generated sequences and associated Phred scores are stored in a FASTQ file (Roy et al. 2018). Sequence alignment involves aligning short DNA sequence reads (< 250 base pairs) with a reference genome. This process assigns a Phread-scale mapping quality score to the reads. The higher the score, the more confidence in the alignment process. Sequence alignments are stored in binary alignment map (BAM) file format (Roy et al. 2018).

Variant Calling, Filtering, Annotation, and Prioritization. Variant calling is a process in NGS sequencing that identifies variants from sequencing data. Accurate variant calling is essential in downstream analyses of NGS data, and is dependent on the quality of bases and aligned reads. The input for variant calling is a set of aligned reads stored in BAM or similar format. This is

given to the variant caller to identify sequence variants between the sample and the reference genome sequence. In variant calling, single-nucleotide variants (SNVs), small insertions and deletions (indels), copy number alterations, and large structural alterations such as insertions, inversions, and translocations are utilized (Roy et al. 2018). Koboldt 2020 discusses the current "best practices" for variant calling in clinical sequencing for germline analysis in family trios and somatic analysis of tumor-normal pairs were discussed. The choice of sequencing strategy, NGS read alignment and preprocessing, the combination of multiple variant calling tools, and filtering to remove false positives, are important in accurate execution of NGS pipelines and further analyses. As discussed by Koboldt (2020), the choice between single- or multi-gene panels for sequencing strategy have impacts on cost-effectiveness. Differences in depth and breadth of sequencing coverage impact variant calling. Higher sequence depth in panel and exome sequencing may add to more sensitive detection of variants at low allele frequencies (Koboldt 2020). For alignment, raw sequence data in the form of a FASTQ file are aligned to the reference sequence using an aligner. Alignment results are stored in a binary alignment/map (BAM) file. Samtools is used to work with BAM files. The accuracy of the alignments and associated annotations impacts the quality of the pipeline outcome. When evaluating the accuracy of variant calls, there needs to be access to benchmark datasets where the true variants are known. For instance, Genome in a Bottle (GIAB) (Zook et al. 2014) and the Platinum Genome (Eberle et al. 2017) dataset are often utilized.

Another component in NGS pipelines is variant filtering. Variant filtering involves the flagging and filtering of variants that represent false-positive artifacts of NGS pipelines. Filtering is done based on sequence alignment and variant calling metadata, such as mapping quality and base-calling quality (Roy et al. 2018). With the production of massive amounts of sequencing data, it can be difficult to filter out variants. Since there is no set approach to filtering variants, logical assumptions have to be made about them. Variant annotation can have implications in determining which variants are significant (Sefid Dashti et al. 2017).

Variant annotation involves queries against sequence and variant databases to characterize variants with metadata. The metadata associated with the variants are used to prioritize and filter variants for candidate selection, analysis, and interpretation (Roy et al. 2018). One of which is ANNOVAR, also known as ANNOtate VARiation. ANNOVAR is a software to facilitate variant annotations, involving gene-based, region-based, and filter-based annotations on a variant call format (VCF) file. Oftentimes, the outputs of NGS pipelines are numerous variants. The challenge is determining which variants are clinically relevant and what information is used to make this determination. Clinically insignificant variants are typically flagged on the basis of synonymous, deep intronic variants, and established benign polymorphisms (Roy et al. 2018). While the variant prioritization tools are beneficial, there should not be full reliance on the tool to make prioritization decisions.

Clinical Laboratory Improvement Amendments (CLIA) Genomics

Clinical genomics is a dynamic process as annotation is constantly being updated. Advances in large-scale genomic analyses in individualized care have promising benefits, but due to the lack of consensus about the proper environment and regulatory mechanisms in which clinical genome sequencing and interpretation should be performed, there has not yet been a transition to clinical research. There is an ongoing shift from research discoveries to participant-focused analysis. In response, several CLIA-certified exome sequencing tests have been made to ensure high standards in genetics laboratories. One of the recommendations for maintaining standards is to perform all whole genome sequencing in CLIA-certified laboratories. To lessen concerns with storing and transferring genetic data, producing sequencing data within the CLIA-certified laboratory would help to ensure the generation of high-quality data. Another concern regarding the feasibility of clinical sequencing is cost. The average cost for a CLIA-certified exome is two to three fold higher than what it costs for a typical research exome at the same sequencing depth (Lyon and Segal 2013). Additionally, researchers are not permitted to release non-CLIA-certified results to participants or physicians that will impact diagnosis or management. New samples can be sent to CLIA-certified laboratories for confirmation at about $300 per variant (based on Lyon and Segal 2013). Following CLIA regulations can help to lead towards executing clinical genomics on a large scale (Lyon and Segal 2013).

College of American Pathologists (CAP) Accreditation

A laboratory can choose to be accredited by the approved accrediting organizations including: the American Association of Blood Banks (AABB), the American Osteopathic Association (AOA), the American Society of Histocompatibility and Immunogenetics (ASHI), the College of American Pathologists (CAP), the Commission on Office Laboratory Accreditation (COLA), and the Joint Commission on Accreditation of Healthcare Organizations (JCAHO). After a laboratory meets CLIA requirements, it is issued a certificate of compliance (COC) or a certificate of accreditation (COA) from the chosen accrediting organization to perform complex or moderate testing. Moderate or high complexity testing involves the laboratory to monitor patient tests, to conduct quality assurance and control processes, assess qualifications, and pay required fees (Rivers et al. 2005). In 1961, the CAP created the accreditation program, highlighting standards that are highly important for quality control in clinical laboratories (Lawson and Howanitz 1997). The CAP accreditation program ensures clinical assays are properly validated to ensure the health and safety of patients, and the health of the population as a whole. The goal of the program was to develop standards to evaluate how efficient a clinical laboratory is, and how accurate their assays are. The CAP accreditation program is the first program to evaluate clinical laboratory performance on a national scale (Hamlin and Duckworth 1997). The accreditation program expects laboratories under accreditation consideration to demonstrate their compliance with the Standards for Laboratory Accreditation and CLIA of 1988 regulatory requirements. The

program expects that laboratories are continually making the efforts required to identify and correct areas that may be lacking, as well as efforts to enhance the performance of their clinical assays (Hamlin 1999). The program also allows laboratories to routinely evaluate their performance and their methods.

The CAP Laboratory Accreditation Program has four main standards for accrediting laboratories: evaluation of the laboratory directory, physical facility and safety, quality control, performance improvement, and inspection requirements (AbdelWareth et al. 2018, Hamlin 1999). These standards are from the Standards for Laboratory Accreditation (Hamlin 1999). Laboratory assessment is based on 18 section-specific checklists. The accreditation process follows the All Common Checklist, Team Leader Assessment of Director & Quality, and Laboratory General Checklists (AbdelWareth et al. 2018). The checklists contain questions marked by “yes”, “no”, or “not applicable”. The questions are periodically reviewed at least once each year. Checklist questions are not standards but are tools for inspectors and directors to ensure the laboratory meets the Standards for Laboratory Accreditation. All laboratories that undergo the CAP accreditation process have access to these checklists to routinely and voluntarily check if the laboratory is meeting the quality checks. For the official check, the checklists are used during accreditation inspection every two years. Unmet checklist items are considered “deficiencies” and have to be corrected within 30 days, and documentation of the correction needs to be submitted (Hamlin 1999).

Recommendations for Clinical Laboratories: Association of Molecular Pathology (AMP) with the College of American Pathologists (CAP) and the American Medical Informatics Association (AMIA)

In 2012, five years after the first approved IVDMIA, the Institute of Medicine developed a report on the evolution of translational omics (Omenn et al. 2012) in response to a serious issue of scientific misconduct involving omics-based assays to determine therapy. Researchers at Duke University claimed that they had achieved a genomic technology that could predict with up to 90% accuracy which early stage lung cancer patients were likely to have a recurrence, and would therefore benefit from chemotherapy. When investigators at MD Anderson cancer center attempted to reproduce the results to validate it, they discovered that data was falsified, there were numerous issues in the implementation of the algorithm, all of which led to misassignment of patients in a clinical trial (Barbash 2015). Due to concerns regarding the premature advancement of omics-based tests in clinical trials, the Institute of Medicine conducted a review of the omics field (Omenn et al. 2012), forming the Committee on the Review of Omics-Based Tests for Predicting Patient Outcomes in Clinical Trials (McShane et al. 2013).

The Institute of Medicine report was followed up in 2013 by more guidelines by the United States National Cancer Institute (NCI), with scientists in multiple areas of expertise related to

‘omics’-based test development. They developed criteria that can be used to determine the readiness of omics-based tests in clinical trials for patient treatment decisions (McShane et al. 2013).

To address the need of properly establishing and validating clinical NGS bioinformatics pipelines, the Association of Molecular Pathology (AMP) with organizational representation from the College of American Pathologists (CAP) and the American Medical Informatics Association (AMIA), developed 17 best practices regarding the design, development, and operation of the pipelines (Roy et al. 2018). Analytical validation of NGS tests have been published in medical literature, but there is still a lack of clarity on requirements for NGS assay validation, especially in NGS bioinformatics pipelines (Roy et al. 2018). Kanagal-Shamanna et al. 2016 summarized NGS-based testing recommendations from organizations such as the American College of Medical Genetics, Centers for Disease Control and Prevention, and CAP The study presents the principles of analytical validation and implementation of NGS-based testing in a CLIA-certified laboratory. The focus is on oncologic testing. Under CLIA, each laboratory needs to determine analytical performance characteristics for a test (Kanagal-Shamanna et al. 2016). NGS-based testing consists of two steps: 1) analytical wet-bench and 2) bioinformatics, also called dry-bench. Validation needs to be independently performed on both of the steps (Kanagal-Shamanna et al. 2016, Gargis et al,. 2012, Aziz et al. 2015, Rehm et al. 2013).

Key Considerations in Research versus Clinical Laboratory Improvement Amendments (CLIA)

The ongoing topic of transitioning from bench to bedside illustrates the importance of establishing a consensus on the guidelines and standards for computational framework development. At the “bench”, known conditions can be controlled, predicted, tested, and validated. At the “bedside,” situations may not be able to be predicted or controlled. This indicates why code and pipelines need to be properly and thoroughly tested for code accuracy. In the case of bioinformatics and computational pipelines, the accuracy of the code is a key consideration. For instance, if the code written to support the pipeline produced an inaccurate prediction of how a patient would respond to treatment, adjustments to the treatment could be detrimental to the patient’s health. The safety of the patient and providing the best patient care are of utmost importance. Similarly, the pipeline implementation is expected to work and produce the intended results. In a research setting, pipelines can be tested under known conditions. On the other hand, in a clinical setting, conditions may be unforeseen, making the validation of these workflows even more challenging, especially in evaluating the performance and robustness. This becomes a more crucial point when we consider how rapidly clinical sequencing is evolving. In 2014, a technology review was published that highlighted the rapid shift to clinical sequencing (Curnutte et al. 2014). Companies had been providing support under

the categories of consumables and arrays, supplying sequencing instruments, and sequencing. As clinical sequencing became more utilized, companies started providing bioinformatics support through alignment and annotation, interpretation and reporting, and data storage and bundling -further highlighting the tight coupling of omics assays and bioinformatics workflows (Curnutte et al. 2014). Proprietary software and pipelines are often “black box” (i.e, no transparency on methods, evaluation etc) which further complicates the issue.

A motivating example: Bioinformatics Advances in Breast Cancer

There is a clear synergy between research and CLIA with the desire to move promising results and assays to the clinical setting. There is a tremendous amount of heterogeneity in the types of omic-based signatures being developed which can make translation of them for clinical use difficult. We briefly highlight some recent findings in breast cancer (noting this is not exhaustive).

Breast cancer is a heterogeneous disease impacting a large population of women in the United States. Continued efforts for improving current treatments and treatment specificity are needed. It is crucial to continue to develop better treatments for patients and to develop improved techniques for the early detection of breast cancer. Advanced omics technologies have provided researchers with more opportunities to tackle the challenge of developing detection and treatment methods for patients with advanced stages of breast cancer. The goal of using omics methods is to identify potential biomarkers to be utilized or targeted in new treatment developments. One of the challenges to drug development is identifying the appropriate biomarkers using methods that are time-efficient and accurate. Bioinformatics methods have helped with screening the genes and signaling pathways involved in the prognosis of triple-negative breast cancer (TNBC). These new methods have identified and provided reliable biomarkers for the better diagnosis and treatment of TNBC. Methods to identify new biomarkers vary dramatically and include identifying differentially expressed (DE) genes, enrichment via gene ontology or pathways, network inference, classification and machine learning etc. Ma et al. 2022 computationally identified five hub genes, TOP2A, CCNA2, PCNA, MSH2, and CDK6 that are unique to TNBC and that they considered as prognostic features. Due to the specificity of these biomarkers to TNBC, they have the potential to be therapeutic targets.

In another study, Alam et al. (2022) identified 190 differentially expressed genes (DEGs) between breast cancer and control samples. They then prioritized this to 13 key genes using protein-protein interaction network analysis. For instance, AKR1C1 and AKR1C2 are expressed in carcinoma cells and stromal fibroblasts, having positive correlation and prevalence in primary breast cancer patients (Alam et al. 2022). Prior work had suggested the loss of AKR1C1 and AKR1C2 in breast cancer results in decreased progesterone catabolism. In combination with increased progesterone receptor expression, it can strengthen progesterone signaling by its

nuclear receptors (Ji et al. 2004). Progesterone metabolism is suggested to be involved in the promotion of breast cancer (Singh et al. 2017). Another candidate, NT5E is regulated epigenetically in breast cancer and its status influences metastasis and clinical outcome (Alam et al. 2022). Prior work by Lo Nigro et al. (2012) found NT5E expression is regulated in breast cancer and NT5E methylation is an indicator of favorable clinical outcome. Alam et al. (2022) additional analysis suggested the use of the key genes in seven candidate drugs, NVP-BHG712, Nilotinib, GSK2126458, YM201636, TG-02, CX-5461, and AP-24534. This could lead to the development of a companion diagnostic for therapeutic stratification.

As more biomarkers are found to have potential for breast cancer treatment targeting, an important aspect to consider is how to adjust treatment for individual breast cancer patients at various disease stages. Bioinformatics tools can aid in improving the discovery rate for therapeutic targets. New targets can also be beneficial for vaccine development. Vaccine development for breast cancer has been attractive for treatment. Vaccines can target the use of an individual's immune system, providing the benefit of little to no adverse side effects. Despite the promising impact, using computational methods to develop vaccines, and deploying the use of vaccines in clinical settings has not been approved (Chiang et al. 2013, Parvizpour et al. 2018). This further highlights the challenges in moving these complex assays and their computational pipelines to a clinical setting.

LITERATURE REVIEW: Clinical Bioinformatics, Clinical Genomics, Medical Genomics

The second part of our literature review focused on understanding the use of genomics and bioinformatics in a clinical setting. We began with two related searches of interest, "clinical bioinformatics" and "clinical genomics." There has been a tremendous increase in research in these areas based on the proportion of Pubmed Citations. In 2021, there were 889 results per 100,000 for "clinical bioinformatics"; and 1,836 per 100,000 for "clinical genomics" (Figure 1). The two search terms cover a broad range, due to this, we began to narrow the search to terms and results relevant to this study. However, the search results for both these terms seemed low. We utilized the search term "medical genetics" (see Appendix 1), which provided a much larger and more relevant corpus for our review (Figure 1).

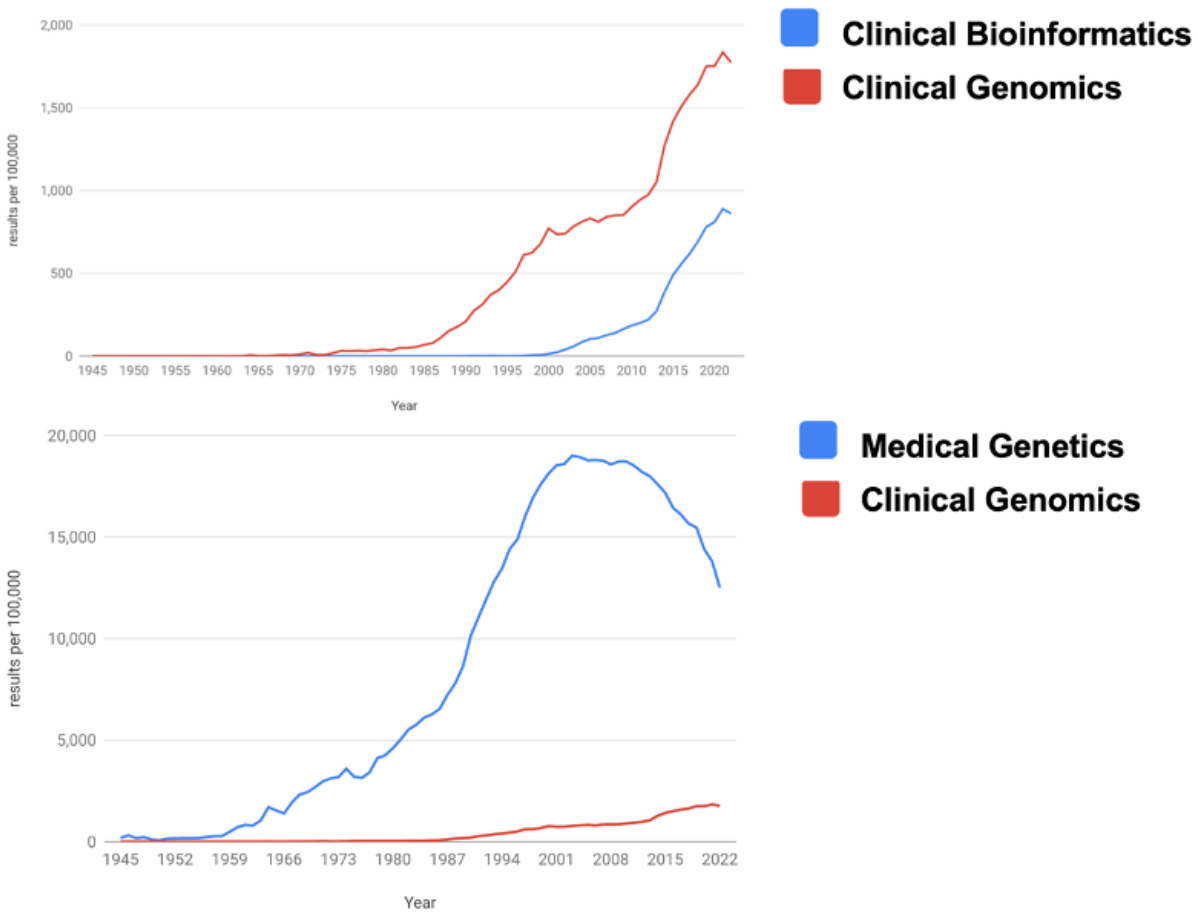

Figure 1. Visualization of Pubmed results for "Clinical Bioinformatics" compared to "Clinical Genomics" per 100,000 citations in Pubmed (Proportion for each search by year 1945 to 2022) (Top); "Medical Genetics" compared to "Clinical Genomics" (Bottom). Note: Y axes are not to the same scale due to difference in magnitude for Medical Genetics.

Given our interest in the literature related to clinical laboratory tests, we evaluated PubMed for "Clinical Laboratory Improvement Amendments" and its acronym, "CLIA". This resulted in a smaller corpus with surprisingly low overlap (18.5%). Inspection of the literature indicated that CLIA was also used as an acronym for chemiluminescent immunoassay at a much higher frequency (30.5%) (Figure 2). A review of the CLIA literature not associated with these two terms found that in the majority of cases, the acronym was not defined in the title or abstract suggesting this term would need to be carefully reviewed if used.

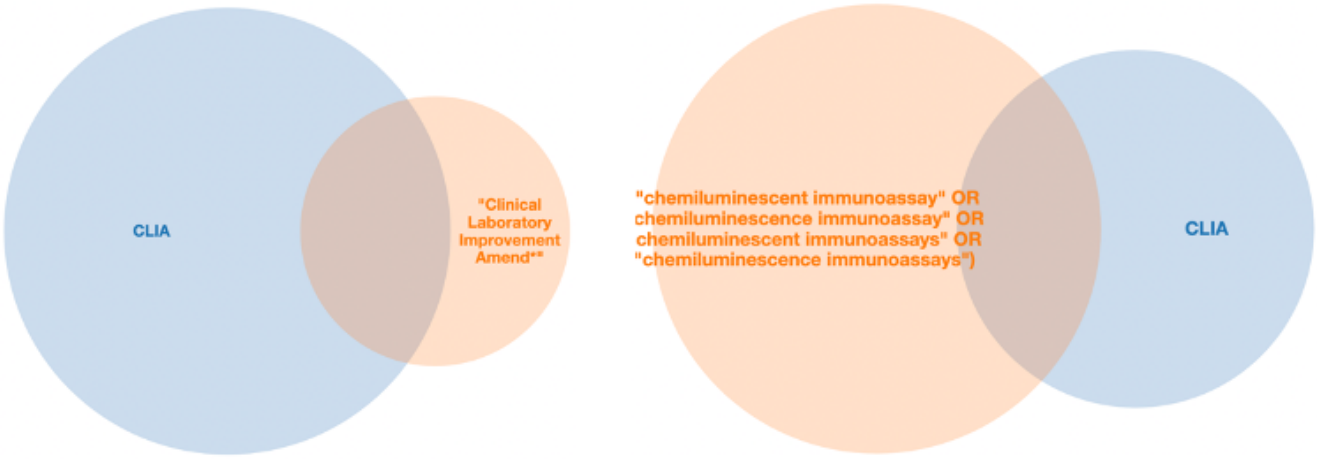

Figure 2. Assessment of literature. Assessment of the literature for the acronym “CLIA” (1595 results) found it was utilized not only for Clinical Laboratory Improvement Amendments (optimized as “Clinical Laboratory Improvement Amend*” with 294 overlapping PubMed results) but also for chemiluminescent immunoassay (optimized as ("chemiluminescent immunoassay" OR "chemiluminescence immunoassay" OR "chemiluminescent immunoassays" OR "chemiluminescence immunoassays") with 487 overlapping citations.

Clinical Bioinformatics

Clinical bioinformatics combines clinical informatics, bioinformatics, medical informatics, information technology, mathematics, and omics science. The field has a large role in clinical applications, such as omics technology, metabolic and signaling pathways, biomarker discovery and development, and computational biology. Clinical bioinformatics differentiates from other types of bioinformatics by its greater focus on clinical informatics (Wang and Liotta 2011). Clinical informatics entails the storage of patient-related information, such as history, clinical symptoms and signs, and vitals (Wang and Liotta 2011, Degoulet and Fieschi 2012). Wang and Liotta 2011 highlights the need for clinical bioinformatics to have integrated analyses, clinical descriptions, and measurements. There is also a need for a communication platform between clinicians and bioinformaticians to improve the quality of patient care (Wang and Liotta 2011). In some instances, large amounts of data can be generated, but can be difficult for bioinformaticians to work with. To help mitigate this problem, analytical platforms have been developed, such as MG-RAST (Keegan et al. 2016), IMG/M (Markowitz et al. 2007), and Qiita (Gonzalez et al. 2018). There is a need to have a comprehensive data processing platform as a majority of the developed platforms are specialized (Shen et al. 2022). MG-RAST, an abbreviation for metagenomics RAST, rapid annotation using subsystems technology, is a metagenomics service for analysis of microbial community structure and function (Keegan et al.

2016). IMG/M, also known as Integrated Microbial Genomics, is a data management and analysis system for microbial community genomes (metagenomes). The database is hosted at the Department of Energy's (DOE) Joint Genome Institute (JGI), and consists of metagenome data integrated with isolated microbial genomes from the IMG system. IGM/M consists of analysis tools for metagenomic data (Markowitz et al. 2007). Qiita is a web-enabled microbiome analysis platform (Gonzalez et al. 2018). In light of the need for a less specialized data analysis platform, Shen et al. 2022 developed Sangerbox 3.0, a web-based user-friendly platform that can be used for pathway enrichment analysis, correlation analysis, among others. The field of clinical bioinformatics provides benefits due to the high volume of biological information obtained and its potential to be utilized in the healthcare system. The field will also aid in moving towards personalized medicine in clinical care (Chang 2005).

Clinical Genomics

Clinical genomics utilizes genomic data to assess and study clinical outcomes. One of the main challenges with clinical genomics is reliably interpreting the multiple and novel variants that can be found through genome sequencing. Genomics offers a large volume of information, and to maximize the benefits, we need to adapt our approaches to analyzing and storing the data. Clinical genetic testing methodologies, such as Sanger gene sequencing, Southern blot, and WES/WGS, are evaluated for validity based on their ability to detect a genetic or genomic variant. To assess this ability, sensitivity, measured as a false-negative rate, and analytical specificity, measured as a false-positive rate, are utilized. Clinical validity is the ability of a test to predict whether or not a clinical condition is present or absent (Katsanis 2013). Vijay et al. 2016 discusses several improvement opportunities in clinical genomics: electronic health records (EHR) data, genomics and chronic illnesses, personalized healthcare and direct-to-consumer genomics, genomics and cancer, genomics and neurobiology, and national and international personalized medicine initiatives. Machine learning and data mining methods have leveraged EHR data as it provides access to large sample sizes and diverse patient cohorts. Transitioning EHR to clinical genomics can help to move towards personalized care. Additionally, genomics data and approaches are beneficial to manage and prevent chronic illnesses and cancer. In the past, cancer had been categorized by the tissue type it affects, but currently, it is increasingly being defined by genetic alterations. Statistical models incorporating family history, age, and other genomic features have contributed to developing personalized care. Along the same lines, the development of consumer genomics has empowered individuals to learn more about their genetics and improve their health. On a broader scale, there is now increased funding for national and international personalized medicine initiatives (Vijay et al. 2016).

Medical Genomics

Medical genomics entails utilizing an individual's genomic information to aid in their clinical care. This field involves translating high throughput genetic methods towards clinical usage (Quintáns et al. 2014). Data sequencing efficiency, and computational and mathematical tools have allowed for the development of tools to understand the functional and regulatory networks of biological systems. This field also contributes to the development of personalized medicine. Medical genomics is evolving from the context of systems biology to systems medicine. This evolution has the potential to work with disease complexity, using molecular diagnostics of patients and diseases (Auffray et al. 2009). A challenge within the field is the interpretation of clinical significance for a single patient, and for patients on a broader scale (Quintáns et al. 2014).

CASE STUDIES OF BIOINFORMATICS WORKFLOWS

From the literature review, we highlight two case studies which each address key considerations for clinical NGS workflows.

Case Study 1: *Use of semantic workflows to enhance transparency and reproducibility in clinical omics*

With the challenges involved in setting pipeline standards, only a few published preclinical omics studies had been translated to a clinical setting (Zheng et al. 2015). Transparency and reproducibility are main components in transitioning an omics analysis pipeline for use in a clinical setting. Within the field of omics analyses, workflow platforms such as Galaxy and Taverna have increased the use, transparency, and reproducibility of omics analysis pipelines (Zheng et al. 2015). Galaxy is an interactive system that utilizes existing genome annotation databases to allow users to search remote resources, combine data from queries, and visualize the outputs (Giardine et al. 2005). Taverna enables users to create and enact scientific workflows (Oinn et al. 2004). Taverna Omics analysis pipelines can contribute to the development of precision medicines. Like many other computationally-based pipelines, omic analysis pipelines can be highly valuable in a clinical setting.

The challenge is that efforts to make omics analysis pipelines more transparent and reproducible are still in the early stages (Zheng et al. 2015). Transparency and reproducibility involve providing all the data used for the project, documentation on the data used for the project, and all the required software downloads to run the pipeline code. Having a clear and direct outline of how the pipeline works and how the pipeline should be run should be available publicly for all users. Addressing how to enhance the transparency and reproducibility of clinical omics

pipelines, they emphasized the use of a checklist , indicating what every pipeline development study should include in their published works. This includes:

- Exact input data used for the analysis.
- Key intermediate data generated from the analysis.
- Third party data (i.e., data from external sources).
- Output data.
- Provenance of all data used.
- All code/software used in the analysis.
- Provenance of all code used.
- Documentation of the computing environment used.
- Veracity checks to ensure analytical validity.
- High-level flow diagram describing the analysis.

Without transparency in the data used, how the pipeline was developed, and how the pipeline should be used, users may find it difficult to get the pipeline to produce the desired results. Having a clear explanation of what data the pipeline can handle as input, and how to obtain the appropriate data type for pipeline usage should be included in the pipeline documentation. Guidelines in output data interpretation should also be included, as in some cases, the pipeline's output might not be clear to the user. In a clinical setting, with users of various backgrounds, computational versus no computational background, it is crucial to have a user-friendly pipeline. Reproducibility is not only in the usage sense, but in the development sense. In the development sense, other researchers may want to understand how the pipeline was constructed. Assistance in running the pipeline, or pipeline troubleshooting, should be easily accessible to the user. With Galaxy and Tavera, "exact input, key intermediate, final output, and relevant external data are all preserved" (Zheng et al. 2015). Galaxy and Taverna also provide high level flow diagrams to guide the users through how to run the workflow (Zheng et al. 2015).

Another challenge is that pipelines and the incorporation of workflow platforms require domain knowledge in the field of translational and clinical omics, making it difficult to transfer an omics analysis pipeline from the research setting to a clinical setting. Without domain knowledge, pipelines can be incorrectly used and negatively impact the patient's care and treatment plan. Users of semantic workflows do not necessarily have to have domain knowledge to utilize the workflow. With semantic enforcement of all datasets, and user-defined methods and constraints, users are guided through each workflow run. The guidance aids in increasing the validity of the workflow, and contributes to patient safety (Zheng et al. 2015). They analyzed the effectiveness of semantic workflows in the fields of translational and clinical omics, and implemented a clinical omics pipeline to annotate DNA sequence variants identified through NGS technologies. Their implementation of a clinical omics pipeline through a semantic workflow allowed for the pipeline to provide transparency, reproducibility, and support analytical validity. They leveraged

the Workflow Instance Generation and Specialization (WINGS) semantic workflow platform. They found that within a semantic framework, using multi-step omics analysis methods resulted in a transparent, reproducible, and semantically validated analysis framework. They suggest semantic workflows have the potential to be beneficial in clinical omics. Unlike other workflow systems, semantic workflow systems can generate semantically validated workflow runs. In these generations, domain knowledge can be embedded with constraints defined by the user. These constraints are then enforced, which helps guide users through running the workflow (Zheng et al. 2015). Zheng et al. 2015 also notes that utilizing the WINGS system addresses four needs in clinical omics analyses:

1) Frequent updates of molecular life science databases.
2) Heterogeneity/consistency of biological data.
3) Rapid development of omics software tools.
4) Processing of large omics data sets.

Analysis of omics data often relies on information in public databases. As biological knowledge increases on a frequent basis, the databases are and must be frequently updated. In clinical settings, having the most updated databases is important and crucial for the best patient care. Analysis also relies on heterogeneous sets of biological data. For instance, RNA-seq analysis protocols typically involve: the genomic sequence used for the alignment of the RNA-seq reads, and the annotated transcript models used for expression quantification (Zheng et al. 2015). Different data types in biological data must be consistent with one another. With rapidly advancing omics software tools, workflow systems need to be able to adapt to the integration of updated and new software tools. Workflow must be able to store and process these large amounts of data. WINGS has the ability to work with frequently updated biological databases, can predefine and constrain the types of datasets, can handle the addition of new, alternative tools, or updated tools, and can execute workflows in a variety of modes, i.e. clusters or cloud (Zheng et al. 2015).

This case study emphasizes that transparency and reproducibility contributes to increasing rigor and reducing errors in pipelines.

Case Study 2: ***A cloud-compatible bioinformatics pipeline for ultrarapid pathogen identification from next-generation sequencing of clinical samples***

Utilizing metagenomic (NGS) data from 237 clinical samples, in a clinical microbiology laboratory setting, Naccache et al. constructed a cloud-compatible bioinformatics pipeline, Sequence-based Ultrarapid Pathogen Identification (SURPI), for ultrarapid comprehensive pathogen identification. NGS technologies have allowed for the development of computational approaches to maintain public health, investigate disease outbreaks, and to diagnose infectious

diseases (Naccache et al. 2014). While NGS has applicability potential in these areas, using these technologies in a clinical setting is difficult due to integration, accuracy, and time efficiency. SURPI utilizes Scalable Nucleotide Alignment Program (SNAP) (Zaharia et al. 2011) and RAPSearch (Ye et al. 2011), two state-of-the-art aligners for accelerated analysis. SURPI consists of two modes, fast and comprehensive. In fast mode, viruses and bacteria are detected by scanning datasets of 7-500 million reads in 11 minutes to 5 hours. In comprehensive mode, all known microorganisms are identified, and de novo assembly and protein homology search for divergent viruses in 50 minutes to 16 hours (Naccache et al. 2014). SURPI addresses the valued time efficiency component in the development of bioinformatics pipelines case study. While a pipeline may be fast in providing results, the pipeline must also be accurate. In the development of unbiased NGS-based clinical assays for combating infectious diseases, rapid turnaround times are in high demand (Naccache et al. 2014). There is high demand for rapid turnaround times when treating infectious diseases as delayed detection and treatment leads to continued transmission, and in some cases uncontrollable transmission, detrimental effects on patient treatment, and increased mortality rates. Metrics that are highly important in the development of aligners for the treatment of infectious diseases are time, sensitivity, accuracy, and throughput (Naccache et al. 2014).

With the development of any bioinformatics pipelines, there are challenges. As listed by Naccache et al. 2014,

(1) Alignment/classification algorithms must contend with massive amounts of sequence data.
(2) Only a small fraction of short NGS reads in clinical metagenomic data corresponds to pathogens.
(3) Novel microorganisms with divergent genomes are not well represented in existing reference databases.

The key challenges can be summarized as the following: volume/scalability, sparsity, and representativeness. NGS technologies can produce > 100 gigabases (Gb) of sequencing reads in a single day (Loman et al. 2012, Naccache et al. 2014). Large amounts of data need to be properly maintained and interpreted. To obtain variants from NGS data, several aligners and variant callers have been developed and incorporated into pipelines. The aligner maps the sequencing reads to a reference genome, and the variant caller identifies variant sites and assigns genotypes (Serrati et al. 2016). With massive amounts of data, various aligners and variant callers, the challenges are, how to maintain this data, which aligners and variant callers should be used, how they should be used in the pipelines, and how the data should be interpreted. In terms of sparsity, Naccache et al. 2014 highlights the needle-in-a-haystack problem with the NGS data, which is also addressed by Kostic et al. 2012, Wylie et al. 2012, and Yu et al. 2012. The problem is only a small fraction of short NGS reads in clinical metagenomic data corresponds to pathogens. The reads must be classified accurately to ensure the data is usable. Even though NGS technologies

produce massive amounts of data, only a small amount of data is usable; relating to how in clinical genomics, there exists the issue of only having a certain amount of variants that can be used for clinical care. In the context of representation in data sets, novel microorganisms with divergent genomes are not well represented in the existing reference databases. Less representation leads to less accurate interpretation of the data. Often, these microorganisms can only be identified through remote amino acid homology (Naccache et al. 2014, Xu et al. 2011, Grard et al. 2012).

The study's cloud-based pipeline, SURPI, begins with raw sequencing reads, goes through preprocessing, SNAP nucleotide alignment to a human reference database, and then branches off based on whether fast or comprehensive mode is chosen. The fast mode involves SNAP nucleotide alignment to a bacterial reference database and SNAP nucleotide alignment to a viral reference database. The comprehensive mode involves SNAP nucleotide alignment to the National Center for Biotechnology Information (NCBI) nucleotide database, followed by de novo contig assembly with ABySS and Minimo, and RAPSearch translated nucleotide alignment to the viral protein or NCBI nucleotide database (Naccache et al. 2014).

To address the problems, SURPI was developed to provide extensive classification of sequencing reads. On the issue of speed, Naccache et al. 2014 compared the computational speed of SNAP, one of the aligners used in SURPI, to BLASTn, BT2, and BWA. Larger *in silico* query data sets of 1.25 million, 25 million, 125 million, and 1.25 billion reads were used for testing. With the 1.25 million reads data set, all of the mentioned aligners performed similarly; with the larger data sets SNAP was 23-87 times faster than BWA and BT2 (Naccache et al. 2014). SURPI's processing times were analyzed with NGS metagenomic data corresponding to 15 data sets ranging from 6.7 to 509 million reads. For SURPI's fast mode, processing times were from 11 minutes to 5 hours; the time increased proportionally to the number of reads. For SURPI's comprehensive mode, processing times were from 59 minutes to 16 hours. To test SURPI's feasibility in real-time clinical analysis, an acute serum sample from a 20-yr-old female patient presenting with a three-day fever to 101.5°C, myalgias, and a headache was utilized. The patient had been exposed to a region of Australia that had mosquito-borne alphaviruses (Knope et al. 2013, Naccache et al. 2014). Within a 13 minute SURPI analysis computational time, sequences that span the genome of human herpesvirus 7 (HHV-7) were detected. SURPI was also tested on various clinical sample types that well represent a variety of infectious diseases (Naccache et al. 2014). The study addresses some of the key challenges when trying to transition a computational pipeline towards CLIA certification. For consideration of usage in a clinical setting, a computational pipeline must be accurate, be well-maintained and well-tested, carry out the intended actions, be sensitive enough to produce the desired outcomes, and meet quality laboratory standards.

This case study illustrates the importance of a pipeline to be both accurate and efficient, and addresses the challenge of sensitivity and false-negatives.

Summary of the Identified Gaps in Pipeline Validation and Development

For the last part of the literature review, we were interested in focusing on the literature since 2018, when the 17 best practices regarding the design, development, and operation of the clinical bioinformatics pipelines (Roy et al. 2018) were published. That paper had emphasized the importance of training and validation. We focused our literature review on the search terms "clinical", "Bioinformatics" and "validation". We observed that there are still challenges in validation which we summarize as three classes of implementations.

For the implementation of existing methods, validation studies were primarily case studies and use case reports. We found tremendous heterogeneity in how algorithms were validated. In some instances, publications were utilized as "surrogates" for actual validation with no in-house validation being reported.

For the implementations where there were no pre-existing methods or workflows, a primary challenge was the lack of gold standards and appropriate reference distributions. This makes evaluation and validation difficult. It was unclear why this is not viewed as a key criteria for assessing if an approach is suitable for translation to the clinic. The lack of gold standards and appropriate reference was the area we had the opportunity to investigate in Aim 2.

For the implementation of artificial intelligence and machine learning, there is a lack of explainability and interpretability, especially for deep learning. There are also training and operational site differences which can be an issue if validation was based on the training site results. Finally there is the issue of distributional drifts (such as data drift), which also impacts the result of the model over time and can be difficult to detect.

AIM 2. USE CASE: DETERMINING BACKGROUND DISTRIBUTIONS FOR EXPRESSION SIGNATURES FOR POTENTIAL CLINICAL USAGE

BACKGROUND

An Ideal Reference/Null Distribution

As noted in Aim 1, a key challenge has been the lack of gold standards and appropriate reference or null distributions. For evaluation of a test or to determine treatment effects, an appropriate reference or null distribution is needed to determine if the difference is significant. For example, if we are assessing gene expression differences pre- and on-treatment, it is crucial to have a null

reference distribution to provide a comparator for when there is no change in gene expression. The differences in gene expression for a null distribution would be due to sample-to-sample variation, and not treatment effects. The goal is to have an appropriate distribution in order to assess outliers due to real treatment effects. We identified four key characteristics for an optimal reference/null distribution:

- The sample gene expression distributions should be appropriate for the disease type.
- There is biological variability sample-to-sample, but is not due to technical artifacts.
- Large enough distribution to avoid sampling error.
- Pairwise differences among samples should reflect normal biological variability, but be less than actual treatment effects, if present.

The gene expression distribution captured by the null reference distribution should well-represent the studied disease type. There should also be biological variability sample-to-sample, however, there should not be batch effects. Not only should the genes appropriately capture the expression distribution for the disease type, but there should be enough genes and samples in the distribution to avoid sampling error. In addition, differences among samples should reflect normal biological validity and be less than the differences observed by actual treatment effects, if there are any present. This would allow us to identify significant changes caused by therapy.

Universal Human Reference (UHR) Replicates

Our use case is to assess the usage of Universal Human References (UHR) replicates as a null reference distribution for assessing differences in gene expression between pre-treatment and on-treatment breast cancer samples. The premise is that UHR replicates can provide appropriate null distributions to assess significant pre- and post-treatment changes. The UHRs are cost effective and consist of multiple cancer cell lines. However, it is noted that they are not exhaustive for all cancer types and that they are not actual patient samples.

Breast Cancer

Breast cancer is one of the more prevalent diseases in the United States and is one of the leading causes of death in the human female population. The 5-year survival rate of breast cancer patients is above 80% due to early prevention (Sun et al. 2017). Approximately 2,261,419 newly diagnosed cases and 684,996 death cases were reported in 2020 (Sung et al. 2021). While advances in breast cancer research and treatment are being made, the disease remains a global health problem, highlighting the great need for treatments to be further improved and individualized. Breast cancer is curable for approximately 70-80% of patients with the early, non-metastatic stage of breast cancer, but there is currently no cure for patients with advanced breast cancer. The heterogeneity of breast cancer has contributed to the difficulties in treatment efficacy (Harbeck et al. 2022). In transitioning countries, the breast cancer incidence rate is

increasing. By 2040, there will be an increase of 3 million new cases and 1 million deaths from breast cancer, solely due to population growth and aging (Arnold et al. 2022).

Breast tumors develop into benign tumors or metastatic carcinomas through constant stimulation from carcinogenic factors. Breast cancer initiation and progression are theorized through two hypotheses, the cancer stem cell theory and the stochastic theory (Sun et al. 2017). The cancer stem cell theory describes tumor growth as being driven by cancer stem cells (Yoo and Hatfield 2008). The theory hypothesizes that tumors originate from a single cell that accumulated mutations and proliferative potential (Polyak 2007). The theory posits that tumor heterogeneity results from an intrinsic hierarchy of cells, not random mutation and clonal evolution. The hierarchy has cancer stem cells at the top (Fábián et al. 2013). The theory also states that cancer stem cells share features with normal stem cells, but traits such as self-renewal, tumor initiation, and maintenance potential, are solely for the cancer stem cells (Reya et al. 2001, Fábián et al. 2013). The second theory, the stochastic theory, as known as the clonal evolution model, hypothesizes that transformation originates from random mutations in breast epithelial cells, such as stem cells, progenitors, or differentiated cells. Additional genetic and epigenetic changes lead to cellular heterogeneity in a tumor (Sgroi 2010).

Current Breast Cancer Therapy

In current breast cancer therapies, with the increased knowledge of biology and understanding of breast cancer, targeted therapies have been developed or are in development. The targets include receptor and non receptor tyrosine kinase inhibitors, intracellular signaling pathways, and DNA repair controls (Alvarez 2010). In a broad sense, there are two types of cancer, non-metastatic and metastatic. Non-metastatic cancer is cancer that has not spread from the primary disease site; metastatic cancer is cancer that has spread from the primary site to other tissues. The goal of non-metastatic cancer therapy is to eradicate the tumor from the breast and regional lymph nodes. The hope is for the prevention of metastatic recurrence. Surgical resection, sampling or removal of axillary lymph nodes, and postoperative radiation are involved in non-metastatic cancer treatment (Waks and Winer 2019). For metastatic cancer, patients are treated according to metastatic cancer subtypes. Ultimately, the goals are to prolong the patient's life, and to provide relief to the patient (Waks and Winer 2019). Current developments have been guided by targeting multiple cells or tissues with receptors for a particular drug. Combining these targets with traditional drugs has great promise for the future of breast cancer therapies.

An approach to breast cancer treatment is radiation. The mammary gland is sensitive to radiation-associated carcinogenesis. Exposure to radiation therapy can be harmful to patients of any age (Ronckers et al. 2004). Utilizing the Surveillance, Epidemiology, and End Results (SEER) Medicare database from January 1st, 1992 through December 31st, 1999 to identify 8724 women aged 70 years and above that were treated for breast cancer, a model was used to

assess the effectiveness of radiation therapy. For older women with early stages of breast cancer, radiation therapy was associated with a lower risk of second ipsilateral breast cancer. It is suggested that radiation therapy may benefit younger breast cancer patients more than older. Patients aged 70-79 with less advanced stages of breast cancer would benefit from radiation therapy, but older patients with more advanced stages of breast cancer would have a lower benefit (Smith et al. 2006). Radiation therapy can pose health problems, such as the risk for other types of diseases or damage to organs (Kamiya et al. 2015).

Another current approach to breast cancer treatment is surgery. Persistent pain post-surgery has been prevalent in 10% to 50% of patients. The pain can be associated with nerve damage or sensory disturbances due to surgery. Gärtner et al. 2009 conducted a questionnaire and database analysis to examine the prevalence of factors associated with persistent pain after breast cancer surgery. The questionnaire involved 3754 women from the ages of 18 through 70 who received primary breast cancer surgery and adjuvant therapy in Denmark. The surgeries were carried out between January 1st, 2005 to December 31st, 2006. The questionnaire was sent between January and April 2008. Of the women who responded to the questionnaire, persistent pain and sensory disturbances post-surgery were found to be clinically significant.

The Move Towards Precision Oncology

Precision medicine follows the concept that since every patient is unique, their cancer is unique. It is the utilization of a patient's molecular and biological features to optimize therapy and drugs that target oncogenic mechanisms. As outlined by Khodadaian et al. 2020, many treatments have been designed for patients with the sample disease, but the following may be observed:

- Symptoms of disease may be reduced (Miyasaki et al. 2002, Patrono et al. 2001)
- Expected responses may fail to occur (Hameed et al. 2018, Wang et al. 2021)
- Side effects may arise (Patrono et al. 2001, Group 1977)

Precision medicine illustrates the move towards studying gene expression in specific organs, tissues, or tumors. With precision medicine, it is important to understand the differences in gene expression between patients, and within patients, pre- and on therapy.

Hallmarks of Cancer

aHanahan et al (2000) indicated that there six key traits that are associated with all cancers: sustaining proliferative signaling, evading growth suppressors, activating invasion and metastasis, enabling replicative immortality, inducing angiogenesis, and resisting cell death. These traits are known as the Hallmarks of Cancer, and are essential to cancer biology. The Hallmarks were updated in 2011, adding two more hallmarks, reprogramming of energy metabolism and evading immune destruction. The update reported that tumors contain a repertoire of cells that contribute to the hallmarks by constructing the tumor microenvironment

(Hanahan et al 2011). The hallmarks were updated again in 2022, stating that phenotypic plasticity and disrupted differentiation are discrete hallmark capabilities. They also state that non-mutational epigenetic reprogramming and polymorphic microbiomes both contribute to hallmark acquisition (Hanahan et al. 2022).

These hallmarks have been translated into representative gene sets. Enrichment and other methods allow evaluation of the gene expression of these hallmarks in a given patient. For the purposes of this use case, the genes to be evaluated for UHR gene expression were restricted to relevant hallmarks. Given overlapping and complementary gene sets, there are a total of 50 gene sets from the Molecular Signatures Database (MSigDB) that represent the hallmarks (Liberzon et al. 2015). Further refinement by clinical collaborators prioritized 27 of the hallmark gene sets as they are related to breast cancer and breast cancer modeling in clinical trials.

The 27 gene sets are: Androgen Response, Angiogenesis, Apoptosis, Bile Acid Metabolism, DNA Repair, E2F Targets, Epithelial Mesenchymal Transition, Estrogen Response Early, Estrogen Response Late, G2-M Checkpoint, Hypoxia, Interleukin-2 (IL2) Signal Transducer and Activator of Transcription 5 (STAT5) Signaling, Interleukin-6 (IL6) Janus Kinase (JAK) Signal Transducer and Activator of Transcription 3 (STAT3) Signaling, Inflammatory Response, Interferon Alpha Response, Interferon Gamma Response, Kras Signaling Downregulated, Kras Signaling Upregulated, Mitotic Spindle, mammalian Target of Rapamycin Complex 1 (mTORC1) Signaling, Myc Targets Version 1, Myc Targets Version 2, Oxidative Phosphorylation, P53 Pathway, Phosphatidylinositol-3-kinase (PI3K) Protein Kinase B (AKT) MTOR Signaling, Transforming Growth Factor beta (TGF-β) Signaling, and Tumor Necrosis Factor alpha (TNFA) Signaling via NFKB.

Gene Set 1: Androgen Response

The androgen receptor (AR) is expressed in more than 70% of breast cancers. When unbound, the AR interacts with chaperone proteins. When bound by a ligand, the AR dissociates from chaperon proteins, forms a homodimer that translocates to the nucleus, and induces a cascade of molecular events. This process results in the activation of target gene transcription. The AR is prevalent in breast tissue and tumors. Androgen signaling pathways have a role in breast cancer development (Gucalp and Traina 2010).

The transcriptional modulation of the cyclin D1 gene (CCND1), a mitogen-regulated cell-cycle control element, has an important role in the growth and progression of breast cancer. An androgen, 5-α-dihydrotestosterone (DHT) inhibits endogenous cyclin D1 expression. DHT is associated with a reduction in cyclin D1 mRNA and protein levels and a decrease in CCND1-promoter activity in MCF-7 cells (Lanzino et al. 2010). MCF-7 cells are the most studied human breast cancer cell line and have contributed greatly to our understanding of the

estrogen response in breast cancer (Lee et al. 2015). MCF-7 cells are used in research for estrogen receptor (ER)-positive breast cancer cell experiments (Comsa et al. 2015). The DHT-dependent inhibition of CCND1 activity requires the involvement of the androgen receptor (AR) DNA-binding domain (Lanzino et al. 2010).

Androgens are commonly associated with males but are expressed in both males and females. In females, androgens are secreted by the adrenal and ovary. Females also secrete higher amounts of androgen, in comparison to estrogen. Mainly, circulating androgens in females are dehydroepiandrosterone sulfate (DHEAS), dehydroepiandrosterone (DHEA), androstenedione (A), testosterone (T), and dihydrotestosterone, listed in descending order of serum concentration. T and dihydrotestosterone bind to the androgen receptor (Burger 2002). In circulation, androgens bind to the steroid hormone-binding globulin (SHBG). This binding controls the availability of hormones to the breast and various tissues. Circulating androgens are risk factors for breast cancer (McNamara et al. 2014).

Gene Set 2: Angiogenesis

Angiogenesis is a main step for breast cancer progression and dissemination (Filho et al. 2010). Tumor growth is known to be dependent on the growth and formation of blood vessels. Pioneer Judah Folkman, was the first to note the association of angiogenesis and cancer (Filho et al. 2010, Folkman et al. 1971). Developing treatments that block angiogenic growth has been supported by oncologists to treat cancer. Tumors have limited capacity to grow without the formation and development of blood vessels, as the vessels provide tumors with the essential nutrients required for tumor growth (Filho et al. 2010). In ongoing clinical trials, an antiangiogenic agent, bevacizumab, has shown promising results. In combination with paclitaxel, in patients with previously untreated metastatic breast cancer, the treatment can improve progression-free survival. Bevacizumab was developed by Genentech in San Francisco, California. It is administered intravenously and is directed against vascular endothelial growth factor A (VEGFA) (Filho et al. 2010). Angiogenesis relies on VEGFA-driven responses (Claesson-Welsh and Welsh 2013).

Gene Set 3: Apoptosis

Apoptosis is defined as normal and controlled cell death. In normal cases, controlled apoptosis is a part of an organism's growth. As outlined by Parton et al. 2001,

- Increased apoptosis with increased proliferation is associated with malignant tumors.
- Breast tumors with increased apoptosis are more likely to be high grade and negative for oestrogen receptors.
- High levels of apoptosis in a breast tumor seem to predict worse survival.
- Measurable increases in apoptosis occur within 24 hours of the start of chemotherapy.

Without apoptotic control, cancers can survive longer and increase in invasiveness due to the increased accumulation of mutations. Tumor growth is a result of uncontrolled proliferation and reduced apoptosis (Parton et al. 2001). More aggressive tumors have higher rates of apoptosis. Under normal conditions, apoptosis occurs to maintain cell populations, and acts as a defense mechanism in immune reactions (Norbury and Hickson 2001). Current cancer treatments act by inducing apoptosis and observing apoptosis to increase treatment efficacy (Parton et al. 2001). The detection of apoptosis in situ has been assessed by electron microscopy or light microscopy, with the assessment of key features such as chromatin condensation and nuclear fragmentation (Parton et al. 2001, Kerr et al. 1994).

Gene Set 4: Bile Acid Metabolism

In regards to metabolism's role in breast cancer, Schramm et al. 2010 indicated that the down-regulation of the bile acid pathway and up-regulation of cholesterol biosynthesis may support steroid biosynthesis, which may support estrogen mediated tumorigenesis of breast cancer cells. The application of large-scale metabolomics, absolute quantification, and a machine-learning based feature selection using Least Absolute Shrinkage and Selection Operator (LASSO) identified metabolites that have an association with tumor development and disease outcomes. LASSO identified the association of tumor glycochenodeoxycholate levels with improved breast cancer survival. Absolute quantification of bile acids revealed the accumulation of bile acids in breast tumors. The studied bile acids indicated an inverse association with proliferation scores in tumors and expression of G2-M checkpoint genes. The findings suggested bile acids may interfere with hormonal pathways in the breast (Tang et al. 2019).

Gene Set 5: DNA Repair

External and internal stressors can cause reversible and irreversible damage to the cells' DNA. DNA damage involves insertions, deletions, DNA mismatch, cross-linking, single-stranded and double-stranded breaks (Hosoya and Miyagawa 2014). Cells respond to DNA damage through cell cycle checkpoints and repair machinery. Cells can either eliminate the damage or trigger apoptosis (Majidinia and Yousefi 2017). DNA repair can be beneficial to developing breast cancer therapeutics. Majidinia and Yousefi 2017 outlines several DNA repair machinery in breast cancer; homologous recombination (HR) pathway, the non-homologous end joining (NHEJ) pathway, the base excision repair (BER) pathway, the nucleotide excision repair (NER) pathway, and the DNA mismatch repair (MMR) pathway. The HR pathway is involved in the reparation of double-stranded breaks. The HR pathway utilizes the sister chromatid as an undamaged homologous template to repair the double-stranded breaks. The NHEJ pathway also works to eliminate double-stranded breaks. NHEJ does not utilize a homologous template for reparations and involves fewer proteins. NHEJ machinery binds the damaged DNA ends without the use of a

homologous template. However, the NHEJ pathway may result in more errors, and in turn, chromosomal damage. Prior studies have indicated the importance of double-stranded break repair mechanisms in breast tumorigenesis (Majidinia and Yousefi 2017). Aside from double-stranded break repair mechanisms, the BER pathway repairs oxidized, alkylated, and deaminated bases (Majidinia and Yousefi 2017, Krokan and Bjørås 2013). DNA glycosylases remove the damaged base by cleaving the N-glycosidic bond (Majidinia and Yousefi 2017, Wallace 2014). BER efficiency is thought to be a determinant of breast cancer risk (Majidinia and Yousefi 2017). NER is another DNA repair pathway, dealing with a variety of DNA helix-distorting lesions. NER factors impact cell metabolism and cell cycle progression (Costa et al. 2003). The deficiency of NER and its potential role in breast cancer development was investigated by Latimer et al. 2010 (Majidinia and Yousefi 2017). It was observed that NER deficiencies were present in stage I breast tumors and that polymorphisms in NER genes may be an indicator of breast cancer risk (Latimer et al. 2010). The MMR pathway is another DNA repair mechanism. The pathway recognizes and eliminates bases that were incorporated incorrectly during replication and recombination. Similar to the NER pathway, polymorphisms in MMR genes may be an indicator of breast cancer risk (Majidinia and Yousefi 2017).

Gene Set 6: E2F Targets

E2F transcription factors have a role in controlling the cell cycle (Hollern et al. 2014). E2F transcription factors have been studied to have biological functions in cancer, but less is known about their function in breast cancer. To further the understanding of the role of E2F transcription factors in breast cancer, Li et al. 2018 conducted a study to analyze the mRNA expression patterns of E2F utilizing the Oncomine and The Cancer Genome Atlas data. The study results indicate factors E2F1, E2F2, E2F3, E2F5, E2F7, and E2F8 were overpressed in patients with breast cancer. The results from the study indicate that E2F targets have the potential to be biomarkers for breast cancer therapeutics. In another study, E2F function was analyzed in mice. Hollern et al. 2014 interbred MMTV-PyMT mice with E2F1, E2F2, or E2F3 knockout mice to test their hypothesis that E2Fs function to regulate tumor development and metastasis. Their results indicate a reduction in metastatic capacity in E2F1 and E2F2 knockouts. With additional gene expression analysis, Hollern et al. 2014 showed the role of E2F factors in tumor development and metastasis.

Gene Set 7: Epithelial Mesenchymal Transition

Epithelial Mesenchymal Transition (EMT) is the transdifferentiation of epithelial cells into motile mesenchymal cells, and has pathological contributions to cancer progression (Lamouille and Derynck 2014). EMT allows a polarized epithelial cell to go through biochemical changes to allow it assume a mesenchymal cell phenotype. Typically, a polarized epithelial cell interacts with the basement membrane through its basal surface. With the mesenchymal cell phenotype,

the cell has increased resistance to apoptosis and increased production of extracellular matrix components (Kalluri and Neilson 2003). EMT is associated with increased aggressiveness and invasiveness in carcinoma cells (Sarrió et al. 2008). Sarrió et al. 2008 conducted a study to assess EMT's presence in human breast tumors by conducting a tissue microarray-based immunohistochemical study in 479 invasive breast carcinomas and 12 carcinosarcomas using 28 different markers. Utilizing unsupervised hierarchical clustering of the tumors, Sarrió et al. 2008 found that up-regulation of EMT markers, vimentin smooth-muscle-actin, N-cadherin, and cadherin-11, along with overexpression of proteins related to extracellular matrix remodeling and invasion, SPARC, laminin, and fascin, occur in breast tumors. The findings from Sarrió et al. 2008 suggest EMT may have relations to breast tumor aggressiveness and progression.

Gene Sets 8 and 9: Estrogen Response Early and Late

Estrogen has an important role in the development of breast cancer. Blocking estrogen receptors is one of the targets of developing therapies. Estrogen is essential in the development of breast cancer. Gustafsson and Warner 2000 conducted a rodent mammary gland study in which they found an estrogen receptor, ERβ, is expressed in approximately 70% of epithelial cells in rodent breast cancer. During pregnancy, they found higher expression of ERβ compared to ERα. Their findings from the rodent mammary gland study suggested the presence of estrogen receptors in epithelial cells to prevent their proliferation, or estrogen has an indirect effect on the breast, possibly taking effect through the immune system. Gustafsson and Warner 2000 suggest the presence of ERβ may not have prognostic value in breast cancer. The overexpression of ERα is observed in early stages of breast cancer, especially in the development of breast cancer tumors. Further understanding the mechanisms of ERα gene expression is beneficial for the development of tools to detect breast cancer early. ERβ is expressed in ERα-positive breast cancers, and ERα and ERβ can be coexpressed in human breast cancer. Suppression of ERα can allow for hormone resistance, but suppression mechanisms are not well understood. Through cell line studies, ERβ was found to result in the inhibition of the growth of ERα-positive breast cancer cells (Hayashi et al. 2003). Studying estrogen receptors gives better insight into understanding the mechanisms in the progression of breast cancer. Early estrogen response may be associated with patient survival and endocrine therapy response in ER-positive breast cancer. Through the usage of a Gene Set Enrichment Analysis (GSEA) algorithm, it was found that estrogen response early scores can be useful in predicting a primary and metastatic breast cancer patient's response to endocrine therapy. Oshi et al. utilized estrogen response early gene sets to obtain gene set enrichment algorithm scores. They hypothesized the score could aid in predicting the response to endocrine therapy and patient survival rates. A low score was found to be significantly associated with a worse response to endocrine therapy and with worse survival in primary and metastatic breast cancer patients (Oshi et al. 2020). Utilizing gene sets entailing genes involved in early and late estrogen responses can help to further understand estrogen reactivity in breast cancer.

Gene Set 10: G2-M Checkpoint

The cell cycle consists of the G1, S, G2, and M stages. The G1, gap 1, stage is where the cell's size increases; the S, synthesis, stage is where the cell's DNA is copied; the G2, gap 2, stage is where the cell undergoes preparation to divide; and the M, mitosis, stage is where the cell divides. The cell cycle contains checkpoints to detect DNA damage. DNA damage response pathways can be used to assess cancer risk. When a single double strand break occurs, damage-induced cell cycle checkpoints are activated, however, the G2-M checkpoint can handle 10-20 double stranded breaks. The G2-M checkpoint has implications in cancer risk (Lobrich 2007). If the damage response pathways were to become abnormal, there is increased risk of developing cancerous cells. Translocations, a type of DNA damage, can be involved in the initiation of carcinogenesis. Translations can be undetected by the damage response mechanisms (Lobrich 2007). As mentioned before, the cell cycle consists of checkpoints at the most crucial cell cycle stages: entry into the S phase, also known as the G1-S checkpoint, entry into the M phase, also known as the G2-M checkpoint, and during replication, also known as the intra-S checkpoints. At these checkpoints, if aberrations are detected, the cell cycle is arrested, and the damage is repaired before moving forward with the cell cycle (Lobrich 2007). BRCA1, the breast cancer tumor-suppressor gene, encodes a protein with a BRCT domain, a feature found in many proteins and is implicated in DNA damage response and in genome stability (Yarden et al. 2002). The role of BRCA1 in DNA damage response has been unclear, Yarden et al. 2002 proposed BRCA1 is involved in Chk1 kinase activation in induced G2-M arrest. Their study indicated BRCA1 controls two proteins essential in the G2-M transition, Cdc25C and Cdc2/cyclin B kinase (Yarden et al. 2002). In 2000, MacLachlan et al. 2000 demonstrated the introduction of BRCA1 in different types of cell lines resulted in the increase of cells with G2-M phase DNA content (MacLachlan et al. 2000, Somasundaram 2003). The implications of BRCA1 with the G2-M checkpoint is an important target in developing breast cancer treatments.

Gene Set 11: Hypoxia

Hypoxia, the decreased availability of oxygen, is a key feature of solid tumors (Brahimi-Horn et al. 2007). Hypoxia makes solid tumors resistant to therapies involving ionizing radiation, some types of chemotherapy, and photodynamic therapy (Vaupel et al. 2005). Hypoxia is involved in poor prognosis of various types of cancers, including breast cancer (Favaro et al. 2011). Hypoxia has been shown to increase patient treatment resistance and contribute to tumor progression. The decreased supply of oxygen induces the hypoxia-inducible transcription factor, which regulates genes that are utilized by tumor cells for survival, treatment resistance, and escape from a nutrient-deprived environment (Brahimi-Horn et al,. 2007). Hypoxia in tumors can be caused by an increase in diffusion distances (> 70 μm), wherein cells receive less oxygen and nutrients. In response to hypoxia, cells reduce their protein synthesis, which leads to cell death (Vaupel et al. 2005). The expression of hypoxia-inducible factor alpha (HIF-1α) and its targets are indicators of

breast cancer prognosis. High HIF-1α expression has been seen with poorer breast cancer prognosis (Favaro et al. 2011).

Gene Set 12: Interleukin-2 (IL2) Signal Transducer and Activator of Transcription 5 (STAT5) Signaling

Interleukin-2 (IL2) is a cytokine that controls the proliferation and differentiation of cells in the immune system (Gesbert et al. 1998). Regulatory T cells require IL2 for homeostasis as it impacts proliferation, survival, and activation (Moro et al. 2022). The IL2 signaling pathway consists of the activation of tyrosine kinases, which leads to the activation of the Jak-STAT pathway, the Ras-MAPK pathway, and the PI3-kinase pathway (Gesbert et al. 1998). The Jak-STAT pathway has been known as a rapid membrane to nuclear signaling pathway (Imada 2000) and controls gene transcription (Gesbert et al. 1998). The Ras-MAPK pathway is a signal transduction pathway, transducing signals for the activation of cell growth, division, and differentiation (Molina and Adjei 2006). The PI3-kinase pathway is involved in cell growth, protein translation, survival, and metabolism. It is also one of the most activated pathways in cancer (Hassan et al. 2013). IL2 binds to specific receptors, IL2R, on the surface of responsive cells, to mediate its activities. IL2R has three subunits, α, β, and γ. IL2R also has various subdomains. One of which, IL2Rβ C-terminal region functions in STAT5 activation (Gesbert et al. 1998). STAT5 is induced by cytokines and growth factors and results in the transcriptional activity of target genes (Buitenhuis et al. 2004). Cytokines control cell survival, death, and differentiation (Atenzi and Sarzi-Puttini 2013). STAT5 is involved with cell proliferation, differentiation, and apoptosis (Buitenhuis et al. 2004). STAT5, under normal conditions, is regulated by prolactin signaling with JAK2/ELF5, EGF signaling networks, and progesterone signaling pathways. Repka et al. 2003 investigated if IL2, in combination with trastuzumab, can increase treatment efficacy. Trastuzumab has clinical activity in metastatic breast cancer. The experiment utilized 10 patients with HER2-overexpression metastatic breast cancer. Each patient was treated with IL2 for 7 weeks and trastuzumab for 6 weeks. *In vitro* immune and clinical responses were assessed. As a result, *in vitro* immune assays showed NK cell expansion, and trastuzumab-mediated increased natural killer cell killing of breast cancer targets. However, there was no correlation with clinical responses (Repka et al. 2003).

Gene Set 13: Interleukin-6 (IL6) Janus Kinase (JAK) Signal Transducer and Activator of Transcription 3 (STAT3) Signaling

IL6 is a cytokine with tumor-promoting and tumor-inhibitory activity. Knüpfer et al. 2007 studied the role of IL6 in *in vitro* studies of breast tumor cells and indicated its potential as a prognostic indicator in breast cancer patients. Through a literature search, IL6 may be a negative prognosticator in breast tumor patients (Knüpfer et al. 2007). Berishaj et al. 2007 investigated IL6 levels in primary breast tumors and Signal Transducer and Activator of Transcription 3

(STAT3) activation mechanisms. STAT3 is present in approximately 50% of primary breast carcinomas. STAT3 activation can be triggered through abnormal activation of receptor tyrosine kinases, Src, and Jaks, which have all been involved in breast cancer. Through analyses of six breast cancer-derived cell lines with high or low levels of tyrosine-phosphorylated STAT3 (pSTAT3), the study found a position correlation between pSTAT3 and IL6 expression (Berishaj et al. 2007). IL6 is capable of having tumor-promoting effects through STAT3, and invasion and metastasis through the JACK/STAT3 and PI3K/AKT pathways (Martínez-Pérez et al. 2021, Johnson et al. 2018, Tawara et al. 2019, Lapeire et al. 2014, Tawara et al. 2019, Winship et al. 2016, Li et al. 2014, Yue et al. 2016, Junk et al. 2017). Siersbæk et al. 2020 showed IL6-activated STAT3 promotes metastasis through ER-FOXA1-STAT3 enhancers. Their results indicate a clinical potential for targeting the IL6/STAT3 pathway in estrogen receptor alpha (ER) positive breast cancer. IL6 and STAT3 have been thought to be connected to ER in breast cancer. Siersbæk et al. 2020 indicate that STAT3 hijacks a subset of ER enhances to drive specific transcriptional activity.

Gene Set 14: Inflammatory Response

Breast cancer can metastasize to the skeleton, as the environment allows for the growth and development of breast cancer cells. Growth factors in the bone are degraded to support tumor cell growth, in a continuing cycle of bone degradation and breast cancer progression. Two important cell types involved with bone development are osteoclasts and osteoblasts. Metastatic breast cancer cells suppress osteoblast differentiation and increase apoptosis. In a study conducted by Kinder et al. 2008, they observed osteoblasts undergo an inflammatory stress response when in the presence of breast cancer cells. Another study indicated sixty-nine percent of patients dying from breast cancer had bone metastases, with the median survival being 24 months in those with the disease, and 85% having widespread skeletal involvement (Coleman and Rubens 1987). Through their findings, they suggest metastatic breast cancer cells can directly induce osteoblasts to express increased levels of inflammatory stress response molecules (Kinder et al. 2008).

Gene Set 15: Interferon Alpha Response

Recent studies have shown that the interaction between autonomous signaling and cytokine networks have implications in inflammatory breast cancer. Type I interferon, more specifically, the interferon alpha signature, has been identified as being upregulated in inflammatory breast cancer. Upregulation is related to apoptosis and cell senescence (Provance and Lewis-Wambi 2019). Interferons are cytokines that affect biological responses. The pathway involved is the Janus kinase/signal transducer and activator of transcription (JAK-STAT) signaling pathway. The pathway involves the interferons and their corresponding receptors, and results in the phosphorylation and activation of STAT1 and STAT2 (Ogony et al. 2016). STAT1 and STAT2

regulate the type I interferon pathway, and are activated after binding to the pathway receptors (Qadir et al. 2020). STAT1 activation has been reported as being tumor suppressive (Qadir et al. 2020, Chan et al. 2012, Schneckenleithner et al. 2011), however in recent studies, STAT1 could also have tumor promotive attributes (Qadir et al. 2020, Khodarev et al. 2010, Greenwood et al. 2012, Tymoszuk et al. 2014, Hix et al. 2013). Bertucci et al. 2014 compared inflammatory breast cancer and non-inflammatory breast cancer groups for pathway and transcription factor activation signatures. Through their analysis of 19 pathways, they found 8 pathways more activated in inflammatory breast cancer, with the interferon alpha response being one of those pathways (Bertucci et al. 2014). Inflammatory breast cancer may have an interferon alpha signature due to chromosomal instability (Provance and Lewis-Wambi 2019). Currently, the interferon alpha response is used in anti-tumor treatment for advanced breast cancer to promote hormone sensitivity and/or to stimulate cellular immunity (Nicolini et al. 2006).

Gene Set 16: Interferon Gamma Response

The interferon gamma response is an inflammatory cytokine. The CCRL2 gene is expressed in breast cancer cells and increased amounts were found in breast tumor tissues with high immune infiltration. Expression of CCRL2 is upregulated by the interferon gamma response. In addition, an alternative transcript of CCRL2, CRAM-A, is expressed under the interferon gamma response. The upregulation of CRAM-A may be a marker of immune response (Sarmadi et al. 2015). García-Tuñón et al. 2007 studied the expression patterns of the interferon gamma response and its receptors through Western blot and immunohistochemistry. Using three breast groups, fibrocystic lesions, *in situ* tumors, and infiltrating tumors, an immunohistochemical and semiquantitative study of interferon gamma response was carried out. The study found within the three groups, the interferon gamma response could be a potential tool in breast cancer. Through Western Blot, they found that the optical density to the interferon gamma response was higher in *in situ* carcinoma than in benign and infiltrating tumors. Additionally, in breast cancer cell lines, treatment involving the interferon gamma response increases p21 (García-Tuñón et al. 2007). p21 has been studied as a factor in breast cancer. p53, the most studied factor in the cancer, transcriptionally upregulates p21 (Elledge and Allred 1998). They discuss the possibility of the interferon gamma response being non-functional and unable to activate p21 to stop the cell cycle (García-Tuñón et al. 2007).

Gene Sets 17 and 18: Kras Signaling Downregulated and *Upregulated*

KRAS belongs to the RAS superfamilies and small-guanosine triphosphate (GTP) binding proteins. KRAS undergoes its inactive state by binding to guanosine diphosphate (GDP), and its active state by binding to GTP, in the cell membrane. The KRAS protein is maintained in oncogenesis. Uprety and Adjei 2020 highlights three rationales for targeting KRAS in cancer therapy:

1. KRAS has a distinct role in tumorigenesis.
2. KRAS mutant cancer cells are KRAS dependent. Preclinical prevention of mutant KRAS inhibits tumor growth.
3. KRAS mutant cancers represent approximately 30% of all human cancers.

Upregulation of KRAS signaling is seen in cancers with higher KRAS mutation rates, such as pancreatic cancer and non-small cell lung cancer. However, less than 2% of breast cancers have mutated KRAS. Mutated KRAS functions as an immune suppressor in other types of cancer, but its effects on the tumor immune microenvironment (TIME) in breast cancer is unknown (Tokumaru et al. 2020). Through the utilization of patient cohorts from the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) and The Cancer Genome Atlas, Tokumaru et al. 2020 hypothesized that KRAS signaling is associated with reduced patient survival and the TIME in triple negative breast cancer.

Gene Set 19: Mitotic Spindle

The mitotic spindle begins to form in prophase of mitosis. In a single cell, two centrosomes will move toward opposite poles during prophase. Then, microtubules will form, connecting the centrosomes. This is the mitotic spindle. The mitotic spindle is a highly dynamic molecular machine that is composed of tubulin, motors, and other molecules. Its purpose is to distribute the duplicated genome to the daughter cells during mitosis (Karsenti and Vernos 2001). Utilizing flow cytometry, Yoon et al. 2002 evaluated the potential role of a defective mitotic spindle checkpoint as the cause of chromosomal instability, a key component of cancers. The study monitored the response of cells to nocodazole-induced mitotic spindle damage. Yoon et al. 2002 indicated that all cell lines with high levels of chromosomal instability have defective mitotic spindle checkpoints, and cell lines with moderate levels of chromosomal instability arrest at the G2 checkpoint of the cell cycle when induced by nocodazole. The study indicated that high levels of chromosomal instability are related to defective mitotic spindle checkpoints (Yoon et al. 2002).

Gene Set 20: Mammalian Target of Rapamycin Complex 1 (mTORC1) Signaling

The mechanistic target of rapamycin (mTOR), which is part of mTORC1 (mTOR complex 1), is involved with metabolic processes for cell growth. mTORC1 is implicated in cancer, as it has a role between signals that control cell growth and metabolic processes associated with growth. mTORC1 can switch between catabolic and anabolic processes; catabolic processes convert macromolecules into nutrients and energy, and anabolic processes convert nutrients and energy into macromolecules (Ben-Sahra and Manning 2017). Preclinical trials have supported inhibition of the PI3K/Akt/mTOR pathway in breast cancer treatments. Phase I to III trials are currently being conducted in solid tumors and breast cancer. mTOR, a serine/threonine protein kinase, is composed of mTORC1 and mTORC2. mTORC1 is attributed to anabolic cell growth, and

mTORC2 is attributed to cellular actin cytoskeleton organization and regulation of AKT phosphorylation (Lee et al. 2015).

Gene Sets 21 and 22: Myc Targets v1 and v2

Myc is an oncogene that is found in most types of cancers. Myc bypasses genetic and epigenetically controlled checkpoint mechanisms when activated. It is also involved in the enforcement of the hallmarks of cancer (Gabay et al. 2014). Myc targets involve genes associated with mitochondrial replication and biogenesis, for instance, POLG, POLG2, and NRF1 (Kim et al. 2008). Rare missense mutations in POLG have been associated with inherited predisposition to breast cancer (Tervasmäki et al. 2018). POLG2 is an accessory subunit of POLG (Singh et al. 2009). Ramos et al. 2020 studied NRF1 activity on molecular signature of breast cancer in Black, White, Asian, and Hispanic women. NRF1 has an important role in estrogen-dependent breast tumorigenesis. Their findings showed that high expressor NRF1 triple-negative breast tumors had an unfavorable prognosis with a high risk of breast cancer mortality in White women, and NRF1's transcriptional activity coupled with target gene signatures contribute to racial differences in breast cancer (Ramos et al. 2020).

Myc targets v1 and Myc targets v2 are obtained from the Molecular Signatures Database Hallmark gene sets (Liberzon et al. 2015). Schulze et al. 2020 utilized Myc targets v1 and Myc targets v2 for gene set variation analysis. The study hypothesized that scores correlate with tumor aggressiveness and survival outcomes. As a result, in estrogen receptor-positive breast cancer, high Myc targets v1 was associated with high mutation load, and high Myc targets v1 and v2 scores were associated with increased infiltration of pro- and anti-cancerous immune cells (Schulze et al. 2020).

Gene Set 23: Oxidative Phosphorylation

Oxidative phosphorylation is where ATP synthesis is coupled to electron movement in the electron transport chain. Cancer cells typically have reduced oxidative phosphorylation in the mitochondria due to reduced flux in the tricarboxylic acid and/or respiration (Solanini et al. 2011). Cancer cells have upregulated glycolysis, leading to the understanding that oxidative phosphorylation is downregulated in all cancers. However, oxidative phosphorylation can be upregulated in some types of cancers. The upregulation of oxidative phosphorylation can be used to alleviate adverse tumor hypoxia and to be a therapeutic target in cancers. In an analysis of gene expression data from 2,000 patients with breast cancer, Whitaker-Menezes et al. 2011 indicated a significant transcriptional upregulation and oxidative phosphorylation. Marizomib, a proteasome inhibitor, has displayed anti-cancer activity. Marizomib, also known as Mzb, inhibits complex II-dependent mitochondrial respiration, which leads to reduced oxidative

phosphorylation. Mzb reduces primary tumor growth and induces apoptosis in human triple-negative breast cancer cell line xenografts (Raninga et al. 2020).

Gene Set 24: p53 Pathway

The p53 pathway is involved in neoplasia. Numerous studies have been conducted on the p53 pathway and have highlighted the p53 transcription factor as a critical point of changes in cellular responses. Under normal conditions, p53 regulates downstream genes involved in cellular responses. Due to p53's critical role in cellular regulation, many diverse regulatory mechanisms control p53. DNA damage activates p53 to carry out processes such as growth arrest and apoptosis. Mutations cause p53 to lose its function, which contributes to the genesis of some tumors and aids in the growth of tumors (Prives and Hall 1999). The p53 pathway is associated with more aggressive breast cancer and low chances of survival. Genetic and epigenetic changes have been identified in regulators of p53 activity (Gasco et al. 2002). Due to the association of p53 with breast cancer, research on p53 provides valuable insight on how to combat breast cancer.

Gene Set 25: Phosphatidylinositol-3-kinase (PI3K) Protein Kinase B (AKT) Mammalian Target of Rapamycin (mTOR) Signaling

The PI3Ks are in a family of lipid kinases involved in the phosphorylation of the 3'-hydroxyl group of phosphoinositides. Phosphatidylinositol-3,4,5-trisphosphate, PIP_3, is a product of the phosphorylation reaction, and a critical second messenger involved in growth and proliferation activation, and survival signaling (Yuan et al. 2016). The dysregulation of the phosphatidylinositol-3-kinase (PI3K) protein kinase B (AKT) mammalian target of rapamycin (mTOR) signaling is involved in cancer growth and progression (Khan et al. 2016). PI3K inhibitors arose due to the high frequency of PI3K pathway alterations in cancer. Tumors can become reliant on PI3K. Genetic alterations in the enzymes involved in the PI3K pathway have made the pathway one of the most frequently dysregulated pathways in cancer (Yuan and Cantley 2008). The PI3K-Akt pathway has been researched in the development of cancer treatments. Serra et al. 2008 researched NVP-BEZ235, a dual inhibitor of the PI3K and the downstream mammalian target of rapamycin (mTOR). They found that this dual inhibitor inhibited the activation of the downstream effectors Akt, S6 ribosomal protein, and 4EBP1 in breast cancer cells. They also found that the inhibitor inhibited PI3K signaling and had antitumor activity (Serra et al. 2008). In another study, Zhang et al. 2020 studied effective specific inhibits for PI3Kα mutants. The study's main focus was PIK3CA, a gene that encodes the p100α catalytic subunit of PI3Kα. PI3Kα contains two subunits, catalytic and inhibitory. PIK3CA is mutated in cancer. The dysregulation of PI3Kα signaling is associated with tumorigenesis and drug resistance (Zhang et al. 2020). Akt inhibitors have been validated as a therapeutic target, and are activated in several types of cancers. Targeting the pathway with drug inhibitors may

result in more effective anticancer treatments (Nitulescu et al. 2016). The Akt signaling cascade is also known to be associated with tumor aggressiveness (Chautard et al. 2014). Inhibitors that inhibit the catalytic activity of mTORC1 and mTORC2 have been developed for anti-tumor activity (Feldman et al. 2009, Hua et al. 2019).

Gene Set 26: Transforming Growth Factor Beta (TGF-β) Signaling

The transforming growth factor beta (TGF-β) signaling regulates tumorigenesis and its associated signaling pathways are modified during tumor progression (Bierie and Moses 2006). TGF-β signaling is controlled extracellularly (Tzavlaki and Moustakas 2020). Studies have shown that TGF-β mediated stromal-epithelial interactions have significantly improved our understanding of cancer regulation. Diverse populations of cell types responding to TGF-β in the tumor microenvironment result in the regulation of cancer initiation, progression, and metastasis (Bierie and Moses 2006). In 1987, Silberstein and Daniel were the first to associate TGF-β with mammary epithelial development (Moses and Barcellos-Hoff 2011). They show that TGF-β can stimulate or inhibit the growth of cells (Silberstein and Daniel 1987). TGF-β is a growth inhibitor, so abnormalities in the signaling pathway result in carcinogenesis. TGF-β is a regulator of epithelial-mesenchymal transition (EMT). EMT has roles in cell motility and cancer cell invasiveness. Due to this, TGF-β signaling may be involved in breast cancer stem cell regulation (Imamura et al. 2012).

Gene Set 27: Tumor Necrosis Factor Alpha (TNFA) Signaling via Nuclear Factor Kappa B (NFKB)

Tumor necrosis factor alpha (TNFA) is involved in inflammatory and immune system responses (Strieter et al. 1993). The immune system is composed of cytokines and tumor necrosis factors, among others. The TNF cytokine family has the ability to induce apoptotic cell death. Sheen-Chen et al. 1997 studied serum TNFA concentration in patients with malignancy. There are increased concentrations in patients with malignancy. The study was designed to evaluate any correlation between the serum TNFA and clinical pathological features. Utilizing 40 patients with invasive breast cancer undergoing radical mastectomy, data pertaining to tumor size, age, estrogen receptor status, lymph node status, and TNM staging was obtained (Sheen-Chen et al. 1997). TNM is a staging system for cancer registries, where T indicates the Tumor, N indicates the Node, and M indicates Metastasis (Piñeros et al. 2019). As a result, the study found that preoperative evaluation of serum TNFA has the potential to be a valuable parameter for evaluating the severity of staging for invasive breast cancer (Sheen-Chen et al. 1997). In a more recent study, Cruceriu et al. 2020 also evaluated the potential of TNFA as a parameter. The tumor microenvironment has a role in breast cancer progression and evolution. Within the tumor microenvironment is TNFA. The study analyzed the correlation between TNFA expression levels at the tumor site and in the plasma or serum of breast cancer patients. They also evaluated the

role of TNFA signaling in estrogen-positive and -negative breast cancer cells. Lastly, they highlighted TNFA's role in epithelial-to-mesenchymal transition and breast cancer cell metastasis. As a result of their analyses, they discuss how TNFA can be a target and a drug when developing breast cancer therapies (Cruceriu et al. 2020). It has been studied that dichloroisocoumarin (DCI) and other serine protease inhibitors are known to block TNFA production (McGeehan et al. 1994). These inhibitors block NFKB, a transcription factor, and subsequent cytokine gene activation (McGeehan et al. 1994).

Increasing Interest in the Application of the Hallmarks

There has been increasing interest in the application of the hallmarks. We identified a variety of applications (see Table 1) and highlighted a few here. In 2018, Cooperberg et al. 2018 analyzed tumor biology and gene expression patterns among men with clinically low-risk prostate cancer. They utilized hallmark gene sets such as Myc targets, G2M checkpoint, and E2F targets. Cooperberg et al. 2018 determined average genomic risk from the hallmark gene sets. They found that genomic risk scores were associated with worse pathology findings, and were associated with prostate-specific antigen recurrence after surgery. They also observed greater genomic diversity among low-risk patients. In addition, they conducted cluster analysis from the hallmark gene sets, identifying 3 main subtypes of prostate cancer. In 2020, Liu et al. (2020) conducted molecular profiling analysis on clinical patient samples. They identified three enriched pathways in relapsed mantle cell lymphoma pathogenesis from hallmarks gene set analysis. As evidenced by these examples and Table 1, the hallmarks are viewed as informative readouts of patient status and response.

Table 1. Studies regarding the 27 hallmarks of cancers in CLIA-certified laboratories.

Hallmark	Study	Year	Disease Type	Summary of Usage
Androgen Response	Analytical Validation of Androgen Receptor Splice Variant 7 Detection in a Clinical Laboratory Improvement Amendments (CLIA) Laboratory Setting (Lokhandwala et al. 2017)	2017	Prostate Cancer	• Patients with castration-resistant prostate cancer can be treated with drugs targeting the androgen receptor (AR) ligand-binding domain. • Active AR splice variant 7 (AR-V7) lacks the ligand-binding domain. • Validation of an AR-V7 assay in a CLIA-certified laboratory.
	Clinical Utility of CLIA-Grade AR-V7 Testing in Patients With Metastatic Castration-Resistant Prostate Cancer (Markowski et al. 2017)	2017	Prostate Cancer	• Analytical validation of AR-V7 assay in a CLIA-certified laboratory.
	Androgen Receptor Immunohistochemistry as a Companion Diagnostic Approach to Predict Clinical Response to Enzalutamide in Triple-Negative Breast Cancer (Kumar et al. 2017)	2017	Breast Cancer	• Comparative analysis of breast carcinoma tissue samples and a validated Clarient CLIA AR Immunohistochemistry (IHC) protocol for AR441 (an AR monoclonal antibody).
Angiogenesis	Circulating baseline plasma cytokines and angiogenic factors (CAF) as markers of tumor burden and therapeutic response in a phase III study of pazopanib for metastatic renal cell carcinoma (mRCC) (Liu et al. 2011)	2011	Renal Cell Carcinoma	• Plasma were analyzed for CAFs by SearchLight multiplex assays in a CLIA-certified laboratory.
	Angiogenic and T-effector subgroups	2021	Clear Cell Renal Cell Carcinoma	• Whole transcriptome sequencing was

	identified by gene expression profiling (GEP) and propensity for PBRM1 and BAP1 alterations in clear cell renal cell carcinoma (ccRCC) (Barata et al. 2021)			performed for ccRCC patient samples submitted to a commercial CLIA-certified laboratory.
	In vivo imaging of eribulin-induced reoxygenation in advanced breast cancer patients: a comparison to bevacizumab (Ueda et al. 2016)	2016	Breast Cancer	● Utilization of a CLIA-certified multiplex protein array from Luminex Multiplex Assays Human Cytokine Magnetic 30-Plex.
Apoptosis	MYC and MCL1 Cooperatively Promote Chemotherapy-Resistant Breast Cancer Stem Cells via Regulation of Mitochondrial Oxidative Phosphorylation (Lee et al. 2017)	2017	Breast Cancer	● Genomic profiling in a CLIA-certified, CAP-accredited reference laboratory.
	A Phase II Trial of Neoadjuvant MK-2206, an AKT Inhibitor, with Anastrozole in Clinical Stage II or III PIK3CA-Mutant ER-Positive and HER2-Negative Breast Cancer (Ma et al. 2017)	2017	Breast Cancer	● After DNA extraction from tumor biopsies, specific exons were PCR amplified and sequenced at a CLIA-certified laboratory using Sanger technology.
	Presence of anaplastic lymphoma kinase in inflammatory breast cancer (Robertson et al. 2013)	2013	Breast Cancer	● Patient tumor samples were analyzed using the FDA approved in situ hybridization (FISH) detection method in a CLIA-certified Genzyme Genetics Laboratory.
Bile Acid Metabolism	Alpha-methylacyl-CoA racemase (AMACR) protein is upregulated in early	2022	Breast Cancer	● AMACR is an enzyme involved in the branched-chain fatty acid and bile

	proliferative lesions of the breast irrespective of apocrine differentiation (Gatalica et al. 2022)			acid metabolism. ● Whole-transcript RNA-Seq was performed. RNA-Seq, also referred to as the Caris WTS (Whole Transcriptome Sequencing) assay is CAP/CLIA validated.
	Intrahepatic Cholestasis of Pregnancy and Serum Bile Acids in HIV-Infected Pregnant Women (Weinberg et al. 2015)	2015	HIV	● Bile acid measurements. ● All assays were conducted in CLIA- and CAP-certified clinical laboratories using FDA-approved methods.
	Bile Acid Profiling Reveals Distinct Signatures in Undernourished Children with Environmental Enteric Dysfunction (Zhao et al. 2021)	2021	Enteric Dysfunction	● Bile acid profiling using ultra-performance liquid-chromatography coupled with tandem mass spectrometry.
DNA Repair	Genomic alterations in DNA repair and chromatin remodeling genes in estrogen receptor-positive metastatic breast cancer patients with exceptional responses to capecitabine (Levin et al. 2015)	2015	Breast Cancer	● Targeted NGS and phosphoprotein analysis using a reversed phase protein microarray (RPMA) platform at a CLIA-certified laboratory.
	Estrogen receptor-positive (ER+) metastatic breast cancer (MBC) patients (pts) with extreme responses (ERs) to capecitabine having tumors with genomic alterations in DNA repair and chromatin remodeling genes (Levin et al. 2014)	2014	Breast Cancer	● Targeted NGS was performed on patients' formalin-fixed paraffin-embedded primary breast cancer specimens at a CLIA-certified laboratory.
	A phase II clinical trial of talazoparib	2021	Breast Cancer	● Assessed the objective response rate (ORR)

	monotherapy for PALB2 mutation-associated advanced breast cancer (Gruber et al. 2021)			of talazoparib monotherapy in patients with PALB2 mutation-associated advanced breast cancer. ● Eligible patients had a deleterious or suspected deleterious mutation in PALB2 on a CLIA-approved commercial germline or next generation sequencing tumor assay.
E2F Targets	Interrogation of Dysregulated Pathways Enables Precision Medicine in Mantle Cell Lymphoma (Liu et al. 2020)	2020	Lymphoma	● Molecular profiling analysis was done on clinical patient samples. ● Hallmarks of cancer such as aberrant apoptosis pathway and E2F targets were identified in patient samples.
	Transcriptomic profiling of patients (pts) with de-novo metastatic castration-sensitive prostate cancer (DN-mCSPC) versus those with mCSPC that have relapsed from prior localized therapy (PLT-mCSPC) (Sayegh et al. 2022)	2022	Prostate Cancer	● Comparison of patients with de-novo metastatic castration-sensitive prostate cancer (DN-mCSPC) and patients with mCSPC. ● RNAseq profiling was performed by a CLIA-certified laboratory. ● Hallmarks of cancers such as E2F Targets, G2-M Checkpoint, Androgen Response, Inflammatory Response, and TNFA Signaling via NFKB were analyzed.
	NeoPalAna: Neoadjuvant Palbociclib, a Cyclin-Dependent Kinase 4/6 Inhibitor, and Anastrozole for Clinical Stage 2 or 3 Estrogen	2017	Breast Cancer	● Single-arm phase II neoadjuvant trial (NeoPalAna) to assess the antiproliferative activity of the CDK4/6 inhibitor palbociclib in primary breast cancer.

	Receptor-Positive Breast Cancer (Ma et al. 2017)			• Analysis of expression of E2F targets in tumor subsets.
Epithelial Mesenchymal Transition	TOPO1 expression in primary and metastatic GI cancers (Castro et al. 2016)	2016	Gastrointestinal Cancers	• Proposed that TOPO1 overexpression is related to metastatic disease as part of the epithelial-mesenchymal-transition (EMT) seen in metastatic phenotypes. • Colorectal (CRC), pancreatic, gastric, and small bowel adenocarcinoma (SBA) patients were tested at a CLIA laboratory.
	Can Patients with Muscle-invasive Bladder Cancer and Fibroblast Growth Factor Receptor-3 Alterations Still Be Considered for Neoadjuvant Pembrolizumab? A Comprehensive Assessment from the Updated Results of the PURE-01 Study (Necchi et al. 2021)	2021	Muscle-Invasive Bladder Cancer	• The PURE-01 study looked at patients with muscle-invasive bladder cancer (MIBC) who had tumor features indicating that immunity may promote response. • Cases involved patients with low epithelial-mesenchymal transition and immune signature scores.
Estrogen Response Early and Late	Estrogen receptor-positive (ER+) metastatic breast cancer (MBC) patients (pts) with extreme responses (ERs) to capecitabine having tumors with genomic alterations in DNA repair and chromatin remodeling genes (Levin et al. 2014)	2014	Breast Cancer	• Analyzed the genomic alterations in tumors of metastatic breast cancer patients who had responses to capecitabine. • Targeted NGS was performed on patients' FFPE primary breast cancer or metastatic breast cancer specimens.
	Metastatic Breast Cancer with ESR1 Mutation: Clinical Management	2016	Breast Cancer	• Reviewed the key considerations involved in clinical decision making.

	Considerations From the Molecular and Precision Medicine (MAP) Tumor Board at Massachusetts General Hospital (Bardia et al. 2016)			• Molecular profiling was performed by an institutional laboratory-developed test, Snapshot-NGS assay (Snapshot-NGS). • The assay was performed in a CLIA-certified laboratory.
	Neratinib Efficacy and Circulating Tumor DNA Detection of HER2 Mutations in HER2 Nonamplified Metastatic Breast Cancer (Ma et al. 2017)	2017	Breast Cancer	• Conducted a single-arm phase II trial for the assessment of the clinical benefit rate (CBR) of neratinib in $HER2^{mut}$ nonamplified metastatic breast cancer. • DNA sequencing of primary and metastatic tumors were performed at a CLIA-certified laboratory. • HER2 mutation testing was conducted at a CLIA laboratory.
G2-M Checkpoint	The Diverse Genomic Landscape of Clinically Low-risk Prostate Cancer (Cooperberg et al. 2018)	2018	Prostate Cancer	• Analyzed the tumor biology among men with clinically low-risk prostate cancer. • Analyzed the gene expression patterns • Utilized hallmark gene sets such as Myc targets, G2-M checkpoint, and E2F targets.
	Molecular Subsets in Renal Cancer Determine Outcome to Checkpoint and Angiogenesis Blockade (Motzer et al. 2020)	2020	Renal Cancer	• Conducted an integrated multi-omics evaluation of tumor specimens from advanced renal cell carcinoma patients. • Identified molecular subsets associated with differential clinical outcomes to

				angiogenesis blockade or with a checkpoint inhibitor. ● Comprehensive genomic profiling was done in a CLIA-certified laboratory.
	Loss of function JAK1 mutations occur at high frequency in cancers with microsatellite instability and are suggestive of immune evasion (Albacker et al. 2017)	2017	Common and rare cancers	● Indicated that loss of function frameshift mutations in JAK1 may have a role in immune evasion. ● Clinical samples were evaluated in a CLIA-certified and CAP-accredited laboratory.
Hypoxia	Phase I study of the Antiangiogenic Antibody Bevacizumab and the mTOR/Hypoxia-Inducible Factor Inhibitor Temsirolimus Combined with Liposomal Doxorubicin: Tolerance and Biological Activity (Moroney et al. 2012)	2012	Advanced malignancies including breast, epithelial ovarian, and colorectal cancer	● Testing for genetic aberrations was conducted in a CLIA-certified molecular diagnostic laboratory.
	Phase 1 study of ARQ 761, a β-lapachone analogue that promotes NQO1-mediated programmed cancer cell necrosis (Gerber et al. 2018)	2018	Advanced solid tumor cancers	● Developed a CLIA-certified IHC assay for the assessment of potential study candidates.
	Clinical proteomics for prostate cancer: understanding prostate cancer pathology and protein biomarkers for improved disease management (Tonry et al. 2020)	2020	Prostate Cancer	● Assays utilized were conducted in a CLIA-certified laboratory.
Interleukin-2 (IL2) Signal Transducer and Activator of	Rapamycin/Il-2 Combination Therapy in Patients with Type 1	2013	Type 1 Diabetes	● Phase 1 clinical trial testing rapamycin/IL-2

Transcription 5 (STAT5) Signaling	Diabetes Augments Tregs yet Transiently Impairs β-Cell Function (Long et al. 2013)			combination therapy in type 1 diabetic (T1D) patients. ● Measured serum cytokine levels in a CLIA-certified contract laboratory. Serum analytes tested: IFN-γ, IL-4, IL-5, IL-6, IL-8, IL-10, IL-1β, IL-1Rα, IL-2, etc.
Inflammatory Response	Immune response profiling of patients with spondyloarthritis reveals signaling networks mediating TNF-blocker function in vivo (Menegatti et al. 2021)	2021	Spondyloarthritis	● Analyzed induced immune responses to define mechanisms of TNF blockers in spondyloarthritis. ● Measured cytokine and chemokines in a CLIA-certified laboratory.
Interferon Alpha Response/ Interferon Gamma Response	Large-scale analysis of KMT2 mutations defines a distinctive molecular subset with treatment implication in gastric cancer (Wang et al. 2021)	2021	Gastric Cancer	● Investigated the distinct molecular features between KMT2-mutant and KMT2-wild-type gastric cancers. ● Compared the distinct molecular features between KMT2-MT and KMT2-WT gastric cancers. ● Datasets utilized were obtained by a commercial CLIA-certified laboratory. ● Study findings were validated with the TCGA cohort. ● Analyzed the interferon alpha response.
Kras Signaling Downregulated and Upregulated	Comparison of KRAS mutation analysis of colorectal cancer samples by standard testing and next-generation sequencing (Kothari et al. 2014)	2014	Colorectal Cancer	● Utilized a colorectal cancer patient population. ● Compared KRAS testing done in CLIA-approved laboratories.

	Whole-exome sequencing of pancreatic cancer defines genetic diversity and therapeutic targets (Witkiewicz et al. 2015)	2015	Pancreatic Cancer	• Micro-dissected pancreatic ductal adenocarcinoma cases underwent whole-exome sequencing. • Indicated that environmental stress and alterations in DNA repair genes associate with distinct mutation spectra.
	KRAS G12C mutations in Asia: a landscape analysis of 11,951 Chinese tumor samples (Loong et al. 2020)	2020	Various types of cancer	• Sequencing data of tumor samples were analyzed for the KRAS mutation. • Samples were analyzed by NGS conducted in a CAP-/CLIA-Accredit ed Laboratory.
Mitotic Spindle	Functional Precision Medicine Identifies Novel Druggable Targets and Therapeutic Options in Head and Neck Cancer (Xu et al. 2018)	2018	Head and Neck Cancer	• Tumor cell culture underwent whole-exome sequencing, RNA sequencing, comparative genome hybridization, and high-throughput phenotyping. • The SEngine Precision Medicine CLIA PARIS test and the CLIA approved UWOnoPlex test were both utilized on the samples.
	PTEN mutations predict benefit from tumor treating fields (TTFields) therapy in patients with recurrent glioblastoma (Dono et al. 2021)	2021	Glioblastoma	• Retrospective review of patients with infiltrating gliomas. • Tumors were evaluated with NGS. • The FoundationOne assay was performed in a CLIA-certified laboratory.
	Small cell lung cancer: Where do we go from here? (Byers et al 2014)	2014	Small Cell Lung Cancer	• Review of the current state of small cell lung cancer treatment. • Discussion of mitotic spindle assembly.

Mammalian Target of Rapamycin Complex 1 (mTORC1) Signaling	Morphoproteomic Profiling of the Mammalian Target of Rapamycin (mTOR) Signaling Pathway in Desmoplastic Small Round Cell Tumor (EWS/WT1), Ewing's Sarcoma (EWS/FLI1) and Wilms' Tumor (WT1) (Subbiah et al. 2013)	2013	Desmoplastic Small Round Cell Tumor (EWS/WT1), Ewing's Sarcoma (EWS/FLI1), and Wilms' Tumor (WT1)	• Assessed patients with DSRCT, Wilms' tumor and Ewing's sarcoma. • Detected p-mTOR, p-Akt, p-ERK1/2, p-STAT3, and cell-cycle related analytes.
Myc Targets v1 and v2	Comprehensive genomic profiling of inflammatory breast cancer cases reveals a high frequency of clinical relevant genomic alterations (Ross et al. 2015)	2015	Breast Cancer	• Conducted comprehensive genomic profiling on specimens using the hybrid capture-based FoundationOne assay. • Assays were conducted at a CLIA-certified and a CAP-accredited laboratory.
	Distinct clinicopathological characteristics, genomic alteration and prognosis in breast cancer with concurrent TP53 mutation and MYC amplification (Lin et al. 2022)	2022	Breast Cancer	• Analyzed breast cancer specimens to discuss the clinical values of concurrent TP53 mutations and MYC alterations.
Oxidative Phosphorylation	Importance of glycolysis and oxidative phosphorylation in advanced melanoma (Ho et al. 2012)	2012	Melanoma	• Utilized analysis of monocarboxylate transporters (MCT) 1 and 4 expression to determine if in advanced melanoma, there exists a link between glycolysis and the oxidative phosphorylation pathways. • All assays were conducted in a CLIA-certified laboratory.
	Therapy resistance: opportunities created by adaptive responses to targeted therapies in	2022	Cancer	• Assessed challenges associated with tumor heterogeneity. • CLIA assays are

	cancer (Labrie et al. 2022)			performed to establish the baseline tumor phenotype and genotype of the tumor.
P53 Pathway	Comprehensive characterization of malignant phyllodes tumor by whole genomic and proteomic analysis: biological implications for targeted therapy opportunities (Jardim et al. 2013)	2013	Breast cancer	● Comprehensive molecular analysis of metastatic malignant phyllodes tumor (uncommon breast tumors) in CLIA-certified laboratories.
	Next generation sequencing of carcinoma of unknown primary reveals novel combinatorial strategies in a heterogeneous mutational landscape (Subbiah et al. 2017)	2017	Carcinoma	● Identified therapeutic strategies through an exploratory analysis of NGS on relapsed and refractory advanced carcinoma of unknown primary (CUP).
	Novel chromatin modifying gene alterations and significant survival association of ATM and P53 in mantle cell lymphoma (Wang et al. 2014)	2014	Lymphoma	● Captured DNA-Seq and RNA-Seq libraries were sequenced to high depth in a CLIA-certified and CAP-accredited laboratory.
Phosphatidylinositol -3-kinase (PI3K) Protein Kinase B (AKT) Mammalian Target of Rapamycin (mTOR) Signaling	Molecular determinants of outcome with mammalian target of rapamycin inhibition in endometrial cancer (Mackay et al. 2014)	2014	Endometrial Cancer	● Identification of molecular markers associated with mTOR inhibitor activity in women with metastatic endometrial cancer. ● Assays were conducted in CAP/CLIA-certified laboratories.
	The PI3K/Akt/mTOR pathways in ovarian cancer: therapeutic opportunities and challenges (Cheaib et al. 2015)	2015	Ovarian Cancer	● Analyzed the PI3K pathway in ovarian cancer. ● Provided a review of clinical trials of novel PI3K inhibitors and inhibitors in combination with

				cytotoxics and ovarian cancer therapies. ● Molecular screening was done with a CLIA-approved targeted sequencing test.
	Landscape of Phosphatidylinositol-3-Kinase Pathway Alterations Across 19,784 Diverse Solid Tumors (Millis et al. 2016)	2016	Cancers	● Conducted a retrospective analysis of 19,784 patients. ● Analyzed aberrations in the PI3K/AKT/mTOR pathway. ● Profiled a large number of diverse solid tumors in a CLIA-certified laboratory.
Transforming Growth Factor Beta (TGF-β) Signaling	Alterations in the Intraocular Cytokine Milieu after Intravitreal Bevacizumab (Forooghian et al. 2010)	2010	Proliferative Diabetic Retinopathy	● Analysis to determine the relationship between cytokine levels pars plana vitrectomy (PPV) and postoperative outcomes. ● Assays on cytokines were done with a sandwich-ELISA multiplex system in a CLIA laboratory.
	Targeting TGF-β for treatment of osteogenesis imperfecta (Song et al. 2022)	2022	Osteogenesis imperfecta (brittle bone disease)	● Histology and RNA-Seq were performed on bones. ● Gene Ontology (GO) enrichment assay, gene set enrichment analysis (GSEA), and Ingenuity Pathway Analysis (IPA) were utilized to identify dysregulated pathways. ● Markers of bone turnover were measured by CLIA- and CAP-certified laboratories.
	Genetics of Pulmonary Arterial Hypertension: Current and Future	2005	Pulmonary Arterial Hypertension	● Discussion of mutations in the gene that codes for activin

	Implications (Elliott et al. 2005)			receptor-like kinase (ALK 1) and TGF-β. ● Discussion of CLIA-approved gene tests.
Tumor Necrosis Factor Alpha (TNFA) Signaling via NFKB	Immune response profiling of patients with spondyloarthritis reveals signalling networks mediating TNF-blocker function in vivo (Menegatti et al. 2021)	2021	Spondyloarthritis	● Analyzed immune responses to microbial and pathway-specific stimuli. ● Utilized peripheral blood samples from 80 patients with axial spondyloarthritis. ● Measured cytokines and chemokines in a CLIA-certified laboratory.
	The Diverse Genomic Landscape of Clinical Low-risk Prostate Cancer (Cooperberg et al. 2018)	2018	Prostate Cancer	● Analyzed prostate cancer cases in the Decipher Genomic Resource Information Database (GRID) and University of California, San Francisco (UCSF) samples in a CLIA/CAP-certified laboratory. ● Samples were analyzed for facilitation of treatment decisions.

DATA

RNA-seq data was utilized from 42 UHR replicates. The UHR is composed of a mixture of 10 human cancer cell lines. Of the 42 UHR replicates, based upon sequencing dates, 38 UHR samples were evaluated. The lots are differentiated by their sequencing date; Lot 1 includes sequencing conducted after November 9th, 2019, and Lot 2 includes sequencing conducted before August 22nd, 2018. For Lot 1, one large fragmentation batch was prepared, and this batch has been used for all subsequent sequencing runs; for Lot 2, one large fragmentation batch was prepared.

For comparison with the null distribution, RNA-seq data from two de-identified patient samples, both with biopsy 1 (pre-treatment) and biopsy 2 (on treatment) data were also utilized.

METHODS

Overview

Rational for Delta Evaluation: As the ultimate focus was to evaluate hallmarks for treatment differences, we are interested in the difference in expression between pre- on-treatment samples (which we denote as delta). The null distribution is pairwise different among the UHR samples.

We utilized prioritized Hallmarks of Cancer (Hanahan et al. 2011) described above as the gene sets to evaluate the gene expression across the 38 UHR samples. All analyses were conducted in the R programming environment and R Studio. The following describes our overall workflow:

- Evaluate gene expression distributions across the UHR lots and in the patient samples
    - Box plots, descriptive statistics for overall expression and pairwise deltas
- Evaluate impact of filtering low expression genes
    - Compare stringent and less-stringent gene filtering criteria using same criteria as above
- Evaluate use of ranks
    - Evaluate the usage of deltas and determine if other ranks would be more appropriate.
- Compute differences in gene expression (deltas) for the patient distribution.
    - Utilize the patient distribution as a comparison point. .
    - UHR expressions should be representative of the patient samples.
- Evaluate the UHR as a null reference distribution.
    - Assess if the UHR distribution meets the key characteristics for an optimal reference/null distribution.

Deltas Computed Based on Gene Expression Differences

Log-transformed Null Distribution Data and Separation by Lots

The UHR TPM expression data was imported into RStudio and NA's were omitted. The columns consist of the ENSG and HUGO gene identifiers, as well as the UHR replicates. This was merged with the run and lot metadata to allow evaluation of batch effects. In addition to the UHR distribution data frame and its associated metadata data frame, we utilize a master gene annotation data frame which includes the Hallmark gene set membership flags.

Once the cutoff metrics are applied, the main data frame is annotated by Lot 1, sequencing conducted after November 9th, 2019, and Lot 2, sequencing conducted prior to August 22nd, 2018. The reference date is set as 2019-11-09, matching the formatting of the dates in the main data frame. The reference data is then compared to the sample date. If the sample's run data is before 11/9/2019, and is annotated as Lot 2. After annotating, there are 26 UHR's in Lot 1 and 12 UHR's in Lot 2. Only UHRs with a run date were used.

The main data frame is then subsetted to produce two new data frames, a data frame containing only Lot 1 samples and another data frame containing only Lot 2 samples. The minimum, median, maximum, and average expression values are obtained for Lot 1 and Lot 2, separately, and together. After these summary statistics, boxplots of the expression values were generated, where the x-axis contains the UHRs and the y-axis contains the expression values. Boxplots were obtained for Lot 1 and Lot 2 separately, and together (Figures 5 and 8).

The coefficients of variation were also computed. Matrices were made for Lot 1 and Lot 2 separately, and together. An annotation summary table was generated, consisting of gene identifier, average expression in Lot 1, average expression in Lot 2, coefficient of variation in Lot 1, coefficient of variation in Lot 2, and overall coefficient of variation.

A function was created to obtain the null deltas which are the pairwise differences of a gene across the UHR's. For instance, if we have a gene with 3 UHR's, the delta would be UHR1 - UHR2, UHR1 - UHR3, and UHR2 - UHR3. The delta distribution was computed for Lot 1 and Lot 2 separated, and together. For Lot 1, there are 26C2 = 325 combinations; 325 pair-wise deltas for a given gene. For Lot 2, there are 12C2 = 66 combinations; 66 pair-wise deltas for a given gene. For Lot 1 and Lot 2 together, there are 38C2 = 703 combinations; 703 pair-wise deltas for a given gene. The R matplot() function is utilized to produce the delta distribution plots (Figures 7 and 10). The maximum deltas, average deltas, and 95th percentiles for the deltas were computed.

Cutoff Metrics Set 1

Following gene list filtering, the NA's are set to 0, and columns with 0's are removed. Genes where average expression < 100 are also removed. The data frame is then log2-transformed. Columns with 0's or negatives are then removed. Columns where log2-average expression > 1 are kept.

Cutoff Metrics Set 2

Following gene list filtering, the NA's are set to 0, and columns with 0's are removed. The data is then log2-transformed. If there are any columns with 0's or negatives, they are removed.

Deltas Computed Based on Rank Differences

The prior section discusses utilizing gene expression differences to compute the deltas, but due to the scaling differences between the UHR null distribution data and the de-identified patient samples used as our patient distribution, we examined deltas as rank-based differences.

Using the main data frame of the UHR distribution, 27 data frames were extracted for the 27 Hallmarks of Cancer. The main data frame was subsetted by the gene lists that are associated with the 27 Hallmarks of Cancer. From here, several filtering methods are examined:

- Keeping genes with 0 expression
- Removing genes with 0 expression
- Keeping genes with > 1 average expression
- Keeping genes with > 10 average expression

Each of the 27 data frames were filtered by each of the five filtering methods for evaluation. The 27 data frames' columns consisted of the ENSG identifier, and the 38 UHRs. The means of the rows were taken and one of the filtering methods listed above was applied.

Once the filtering method was applied, the rankings were computed. The gene with the highest expression would be given a rank of 1, the gene with the second highest expression would be given a rank of 2, and so on. To calculate the rank deltas, the same delta function that was utilized in the prior section was utilized. Once the rank delta distributions were computed, the minimum and maximum rank deltas were computed. The 80th, 85th, 90th, and 95th percentiles were also computed.

RESULTS AND DISCUSSION

Data Distribution Before Filtering

Pre-Filtering

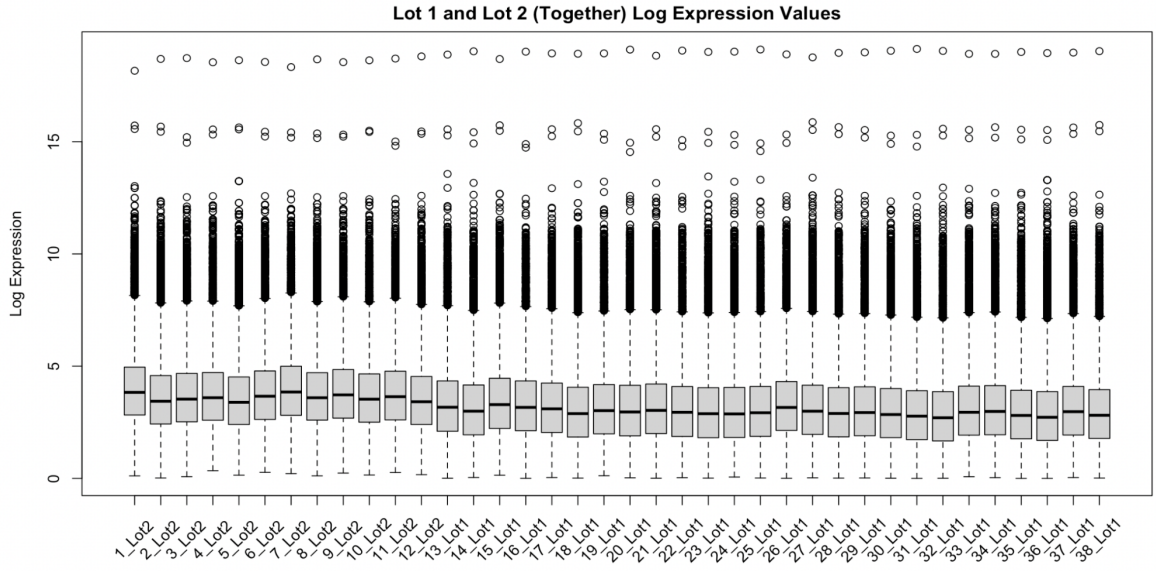

Figure 3. Distribution of the log-transformed gene expression values without filtering. Boxplot of the logged-expression values for Lot 1 and Lot 2 combined. The x-axis indicates the UHR ID number, and the y-axis indicates the log-transformed gene expression value.

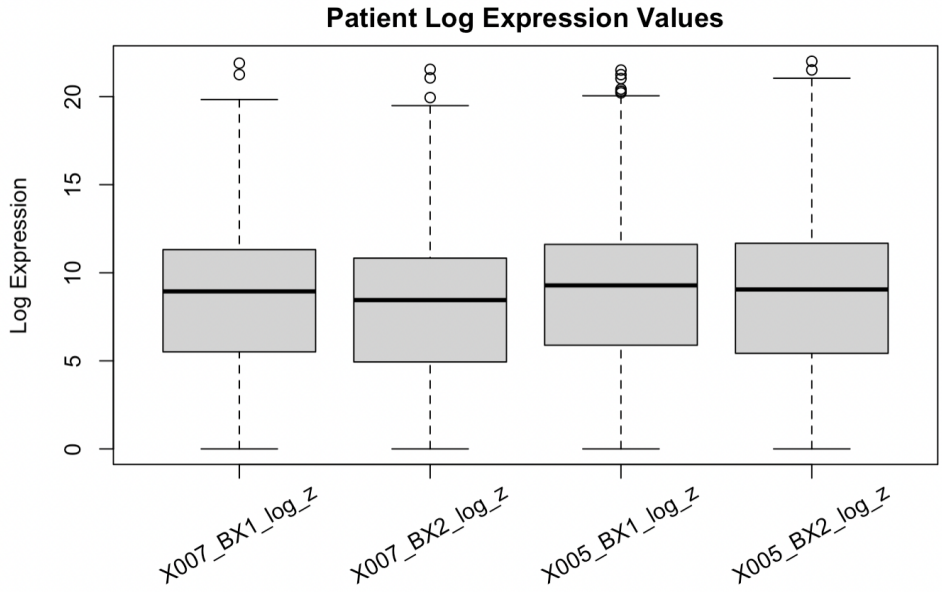

Figure 4. Distribution of the log-transformed gene expression values without filtering for the two paired deidentified patient samples obtained. Boxplot of the logged-expression

values. The x-axis indicates the de-identified ID, and the y-axis indicates the log-transformed gene expression value.

As shown by figures 3 and 4, there exists more variability between the lots than between the patients. This was also reflected in the descriptive statistics as well (Table 2). Given the lot differences and other technical differences, there was concern about combining the UHRs.

Table 2. Descriptive statistics summaries for the pre-filtered UHR distribution and the patient distribution.

	UHR Distribution (Pre-Filtered)	Patient Distribution
Minimum	0.000216388	0
Q1	2.109567	5.438526
Median	3.159304	8.927307
Q3	4.340651	11.360290
Maximum	19.13106	22.00219

In addition to the variability, there was also a remarkable difference in overall expression with the UHRs much lower than the patient samples (Table 2). This led us to assess both filtering (to remove low expressed genes) as well as rank based methods given scaling issues.

Deltas Computed Based on Gene Expression Differences

Log-transformed Null Distribution Data and Separation by Lots

Cutoff Metrics Set 1

The first set of cutoff metrics are as follows:

- Filtered out genes with low non-logged average expression (< 100).
- Filtered out genes with low logged average expression (< = 1).

60554 genes were the starting point, however, after filtering, there are only 533 genes left. Lot 1 consisted of 26 UHRs, 533 genes, and 325 pair-wise deltas for a given gene (26C2). Lot 2 consisted of 12 UHRs, 533 genes, and 66 pair-wise deltas for a given gene (12C2). Overall, there are 38 UHRs, 533 genes, and 703 pair-wise deltas for a given gene (38C2). Lot 1 logged-expression values are slightly higher than Lot 2.When taking Lot 1 and Lot 2 together, it does not appear that normalization within or between lots was successful. Within the UHRs,

there is variation in the expression levels (Table 3). The normalization of the lots was not apparent in the boxplots leading to concerns about the magnitude of the batch effects (Figure 5).

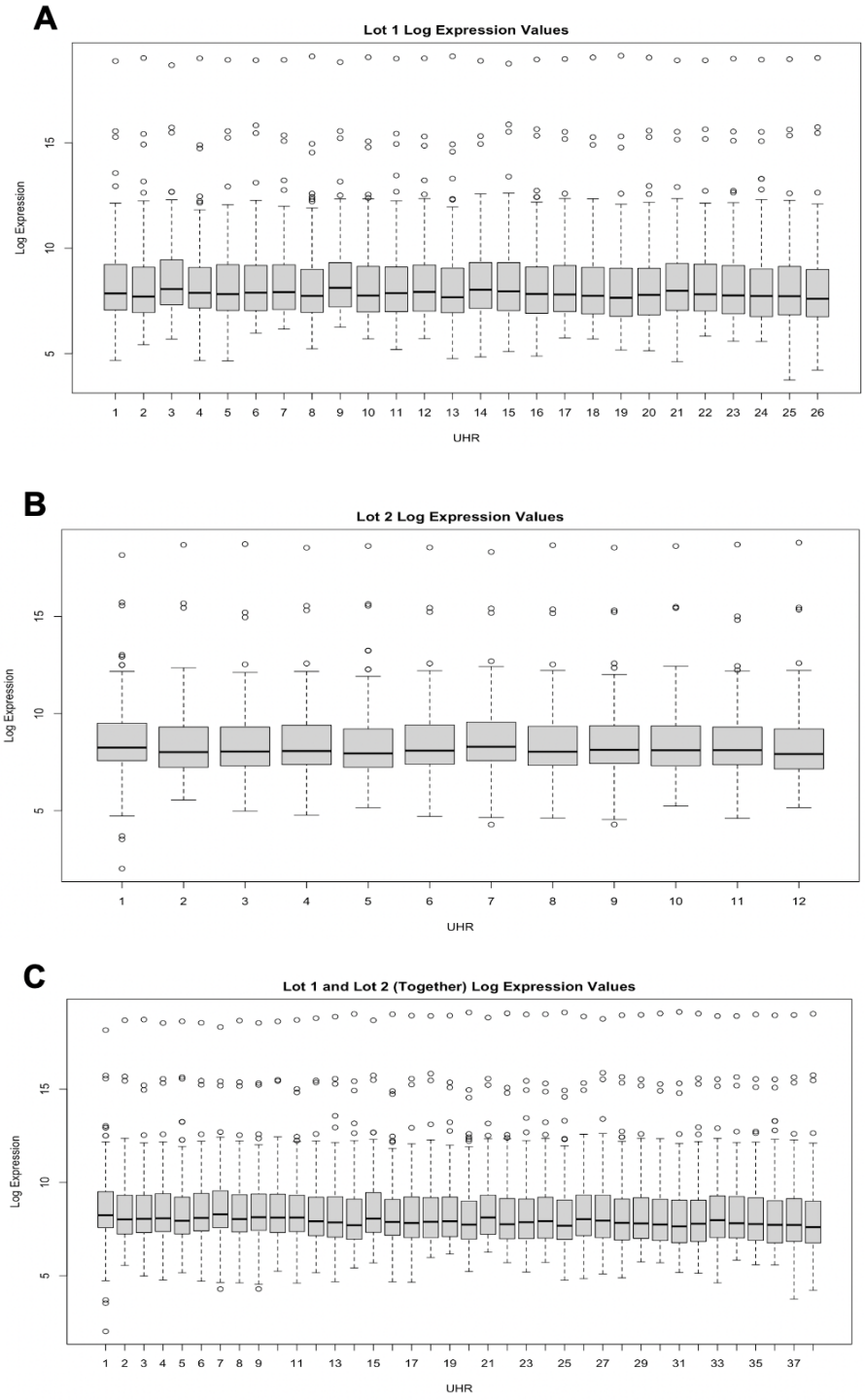

Figure 5. Distribution of the log-transformed gene expression values with the first approach's cutoff metrics. (A) Boxplot of the logged-expression values for Lot 1 only. The

x-axis indicates the UHR ID number, and the y-axis indicates the log-transformed gene expression values. **(B)** Boxplot of the logged-expression values for Lot 2 only. The x-axis indicates the UHR ID number, and the y-axis indicates the log-transformed gene expression values. **(C)** Boxplot of the logged-expression values for Lot 1 and Lot 2 combined. The x-axis indicates the UHR ID number, and the y-axis indicates the log-transformed gene expression values.

Table 3. Descriptive statistics summary for cutoff metrics set 1.

	UHR Distribution (After Cutoff Metrics Set 1)
Minimum	2.015176
Q1	7.118302
Median	7.913196
Q3	9.230560
Maximum	19.13106

Across the genes, there is a slight increase in the coefficients of variation for Lot 1 and Lot 2, (Figure 5). The greatest variability was seen when the lots were combined.

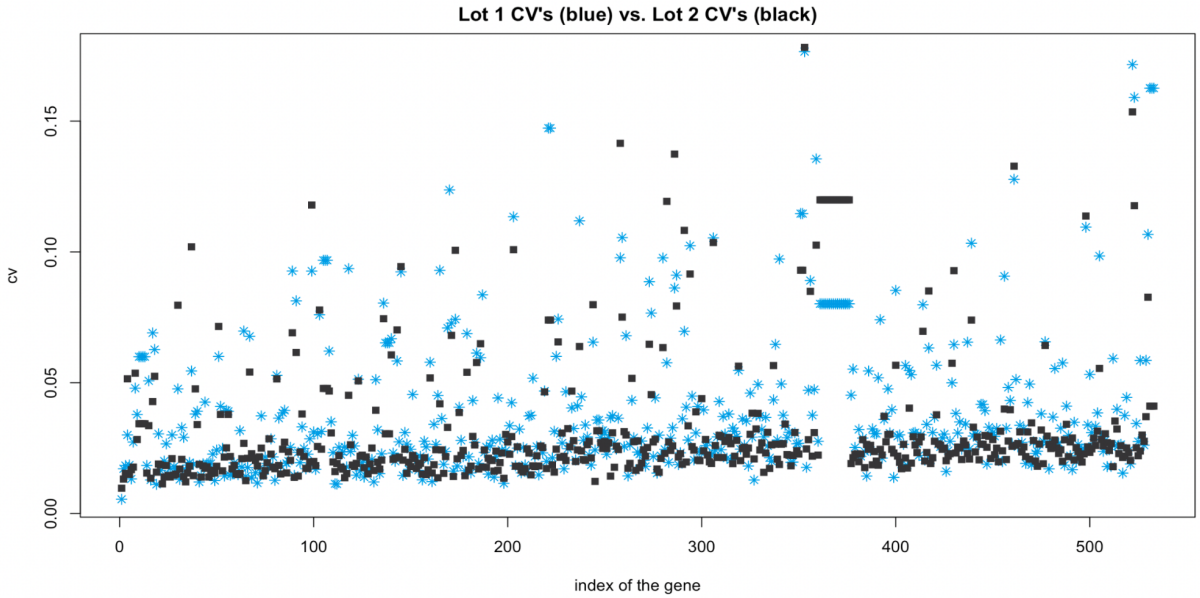

Figure 6. Coefficients of variation distribution. Plot of the distribution of the coefficients of variation, where blue indicates Lot 1 samples and black indicates Lot 2 samples.

There are higher pairwise deltas than expected for the UHR distribution and there is a shift across the two lots. Summary statistics, such as minimum and maximum deltas were computed for Lot 1 and Lot 2. In some cases, Lot 1 and Lot 2 had similar maximum deltas (Table 4). As an example, now deprecated genes, ENSG00000278047.1 and ENSG00000276924.1, had identical maximum deltas across the two lots, 7.704626 (Table 4). However, in other cases, Lot 1 and Lot 2 had very different maximum deltas Table 5). For instance, for small nucleolar RNA genes ENSG00000263934.4 and ENSG00000200087.1 had different maximum deltas across the two lots, 0.401890 and 0.899330, respectively. It is important to remember that these are supposed to be representing null distributions so a doubling of the delta is a concern. There exists very large pairwise differences that would make the detection of biological changes difficult. A stable reference distribution should have less variability in its distribution within lots and between lots.

Table 4. Similar maximum deltas for cutoff metrics set 1.

	Lot 1 Maximum Delta	Lot 2 Maximum Delta	Lot 1 and Lot 2 Maximum Delta
ENSG00000278047.1	5.501958	2.72114	7.704626
ENSG00000276924.1	5.501958	2.72114	7.704626

Table 5. Different maximum deltas for cutoff metrics set 2.

	Lot 1 Maximum Delta	Lot 2 Maximum Delta	Lot 1 and Lot 2 Maximum Delta
ENSG00000263934.4	0.347370	0.401890	0.401890
ENSG00000200087.1	0.899330	0.713850	0.899330

When observing the distribution of the deltas for Lot 1, the deltas are within the range of -5 to 5; for Lot 2, the deltas are within the range of -5 and 5 as well, but taken together, the deltas have a broader range (Figure 7). The variation in the delta distribution indicates that the UHR may not be a stable reference distribution.

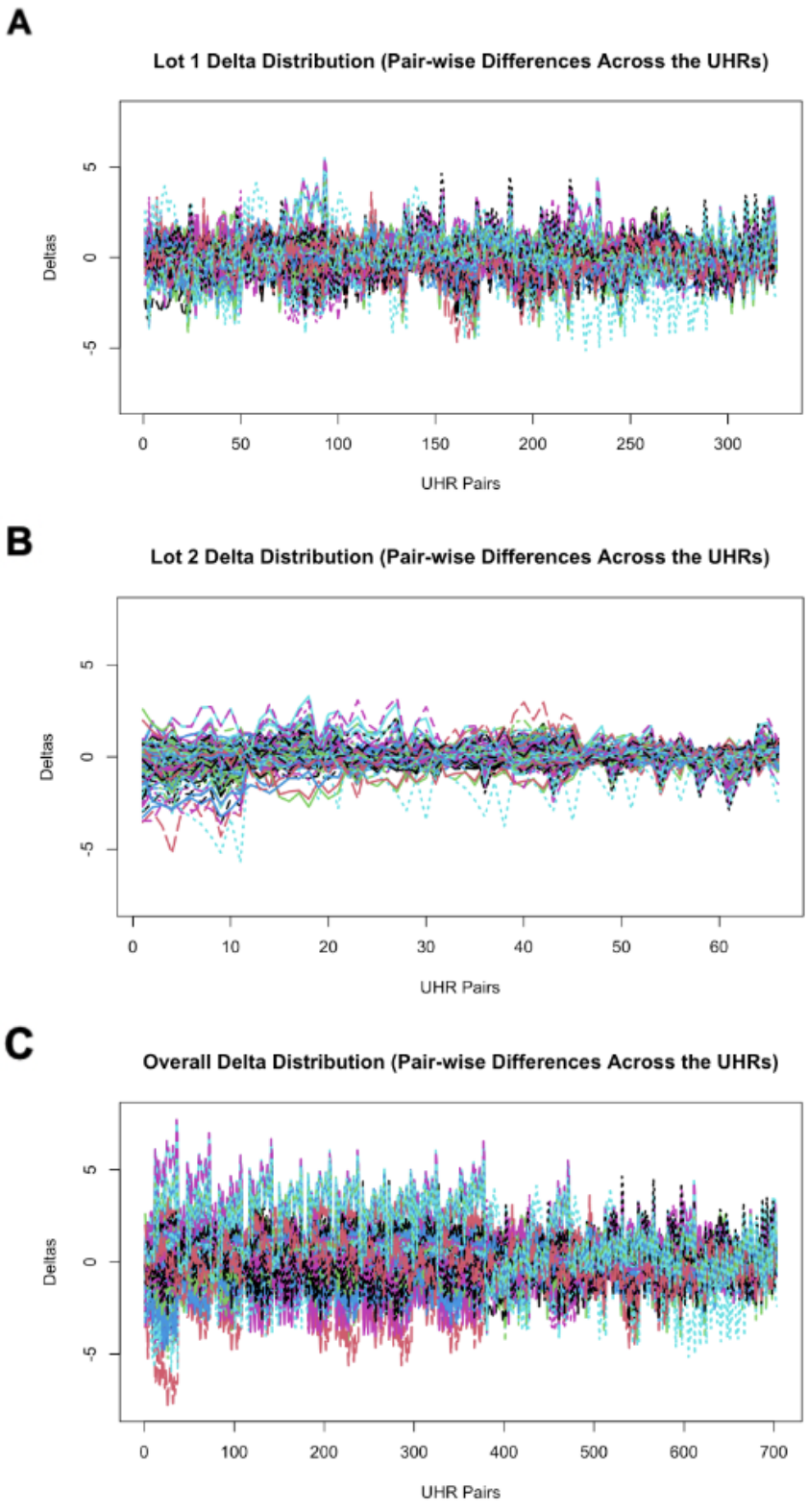

Figure 7. Distribution of the gene expression delta distributions with the first approach's cutoff metrics. (A) Plot of the gene expression delta distribution for Lot 1 only. The x-axis

indicates the UHR pair number, the y-axis indicates the delta value. The delta value is the difference in gene expression between a gene's expression in one UHR and the same gene's expression in another UHR. **(B)** Plot of the gene expression delta distribution for Lot 2 only. The x-axis indicates the UHR pair number, and the y-axis indicates the delta value. The delta value is the difference in gene expression between a gene's expression in one UHR and the same gene's expression in another UHR. **(C)** Boxplot of the gene expression delta distribution for Lot 1 and Lot 2 combined. The x-axis indicates the UHR pair number, and the y-axis indicates the delta value. The delta value is the difference in gene expression between a gene's expression in one UHR and the same gene's expression in another UHR.

Using this filtering method, we observed that many transcripts are not expressed in the UHRs. For those expressed, there was a high level of technical variability among the UHRs. Filtering out non-expressed and highly variable genes reduces the number of genes tremendously but does increase the overall expression closer to the patient distribution.

Cutoff Metrics Set 2

The second set of cutoff metrics are as follows:

- Removed any cells with average expression = 0.
- Removed any cells with 0 or negative logged average expression.

With a starting point of 60554 genes, after filtering with this set of cutoff metrics, there were 10430 genes left, giving a broader gene list than the first set of cutoff metrics. With this set of cutoff metrics, Lot 1 consisted of 26 UHRs and 325 pair-wise deltas for a given gene (26C2); Lot 2 consisted of 12 UHRs and 66 pair-wise deltas for a given gene (12C2); and overall, there were 38 UHRs and 703 pair-wise deltas for a given gene (38C2). With this set of cutoff metrics, the UHRs have a relatively more similar distribution in comparison to the distribution seen with the first set of cutoff metrics (Figure 8 and Table 6). However, the lots do not appear to have been normalized together.

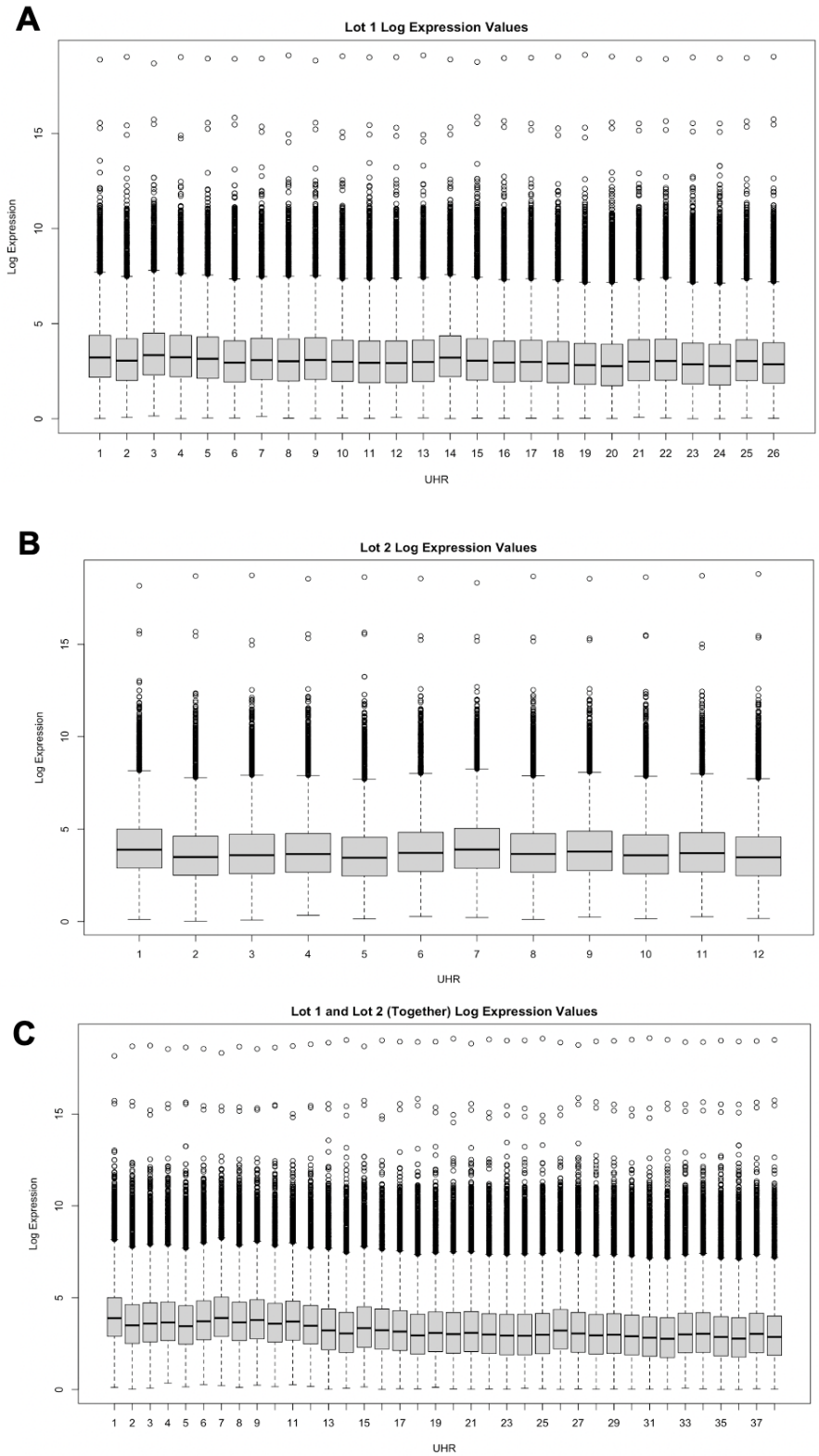

Figure 8. Distribution of the log-transformed gene expression values with the second approach's cutoff metrics. (A) Boxplot of the logged-expression values for Lot 1 only. The

x-axis indicates the UHR ID number, and the y-axis indicates the log-transformed gene expression values. **(B)** Boxplot of the logged-expression values for Lot 2 only. The x-axis indicates the UHR ID number, and the y-axis indicates the log-transformed gene expression values. **(C)** Boxplot of the logged-expression values for Lot 1 and Lot 2 combined. The x-axis indicates the UHR ID number, and the y-axis indicates the log-transformed gene expression values.

Table 6. Descriptive Statistics summary for cutoff metrics set 2.

	UHR Distribution (After Cutoff Metrics Set 1)
Minimum	0.000216388
Q1	2.188729
Median	3.216161
Q3	4.383641
Maximum	19.13106

For Lot 1, there is 1 sample with a coefficient of variation > 1, and 10429 samples with a coefficient of variation <= 1. For Lot 2, there are 0 samples with a coefficient of variation > 1, and 10430 samples with a coefficient of variation <= 1. Combined, there are 0 samples with a coefficient of variation > 1 and 10430 samples with a coefficient of variation <= 1. The higher the coefficient of variation, the greater the level of dispersion around the mean; the lower the coefficient of variation, the better, as the spread of data values is low relative to the mean. The distribution of the coefficients of variation are spread from 0 to 1, indicating high variability between lots and within lots. Across the genes, there is a slight increase in the coefficients of variation, indicating higher variability between the two lots (Figure 9).

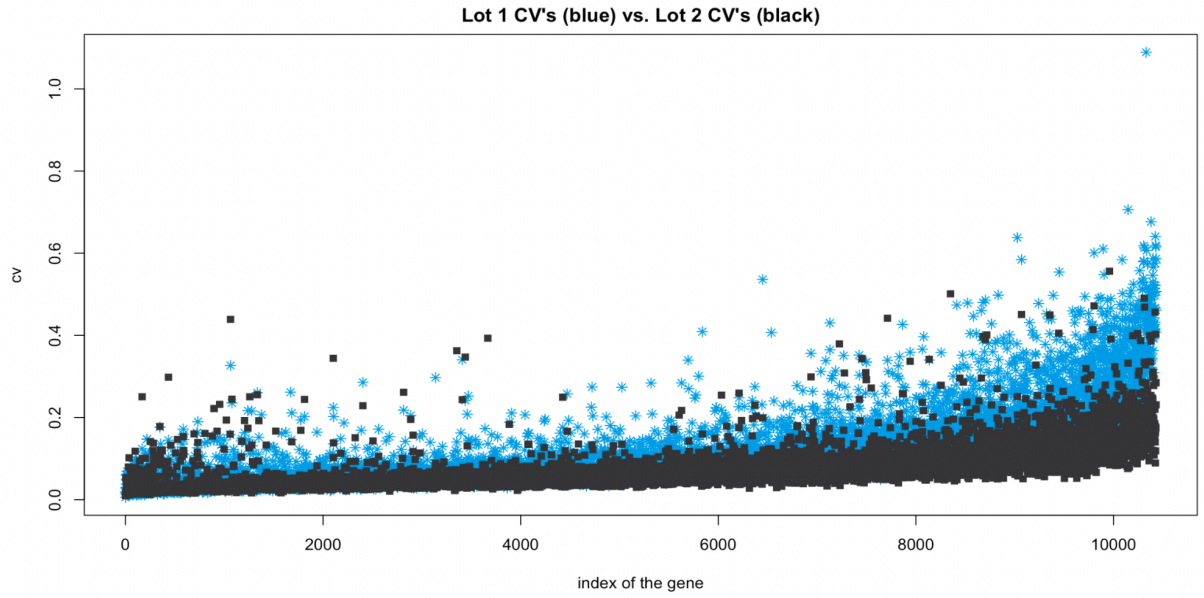

Figure 9. Coefficients of variation distribution. Plot of the distribution of the coefficients of variation, where blue indicates Lot 1 samples and black indicates Lot 2 samples.

Lot 1 minimum delta is -5.548437; Lot 2 minimum delta is -5.716802; and together, the minimum delta is -7.871139. Lot 1 maximum delta is 5.646818; Lot 2 maximum delta is 3.959282; and together, the maximum delta is 7.04626. This set of cutoff metrics is similar to what was observed in the first set of cutoff metrics. In some cases, there were similar maximum deltas. The now deprecated genes, ENSG00000171560.1, ENSG00000117308.14, and ENSG00000182866.16, had similar maximum deltas for Lot 1 only and Lot 2 only, but when considering Lot 1 and Lot 2 together, the maximum deltas varied. For instance, ENSG00000171560.14 had a Lot 1 maximum delta of 0.4132900, a Lot 2 maximum delta of 0.5632920, and then the delta increases to 0.7605690, when considering both Lot 1 and Lot 2 (Table 7). However, in others, the deltas had a noticeable difference. For instance, ENSG00000137285.9 had a Lot 1 maximum delta of 0.3063000, a Lot 2 maximum delta of 0.7170630, and an overall maximum delta of 1.3466430 (Table 9). In addition, for some genes, Lot 1 had higher deltas than Lot 2 (Table 10), but in other genes, Lot 2 had higher deltas than Lot 1 (Table 9). The UHR distribution has variability as a whole and within the lots.

Table 7. Examples of similar maximum deltas for cutoff metrics set 2 (lowest deltas).

	Lot 1 Maximum Delta	Lot 2 Maximum Delta	Lot 1 and Lot 2 Maximum Delta
ENSG00000171560.14	0.4132900	0.5632920	0.7605690

ENSG00000117308.14	0.4134070	0.4859700	1.0419480
ENSG00000182866.16	0.4134930	0.6076070	1.1948270

Table 8. Examples of similar maximum deltas for cutoff metrics set 2 (highest deltas).

	Lot 1 Maximum Delta	Lot 2 Maximum Delta	Lot 1 and Lot 2 Maximum Delta
ENSG00000275127.1	4.304664	3.1155680	5.048557
ENSG00000207263.1	3.954312	3.2132190	4.647363
ENSG00000207279.1	3.916215	2.9740210	4.858417

Table 9. Different maximum deltas for cutoff metrics set 2 (lowest deltas).

	Lot 1 Maximum Delta	Lot 2 Maximum Delta	Lot 1 and Lot 2 Maximum Delta
ENSG00000137285.9	0.3063000	0.7170630	1.3466430
ENSG00000196230.12	0.3063710	0.5641670	1.1350360
ENSG00000073578.16	0.3126200	0.6059340	1.1379830

Table 10. Different maximum deltas for cutoff metrics set 2 (highest deltas).

	Lot 1 Maximum Delta	Lot 2 Maximum Delta	Lot 1 and Lot 2 Maximum Delta
ENSG00000206172.8	5.646818	1.5493750	5.646818
ENSG00000278047.1	5.501958	2.7211400	7.704626
ENSG00000276924.1	5.501958	2.7211400	7.704626

As seen with the first set of cutoff metrics, when considering Lot 1 and Lot 2 together, the deltas have a broader range and greater variability. This again indicates that the UHR distribution may not be suitable as a reference distribution.

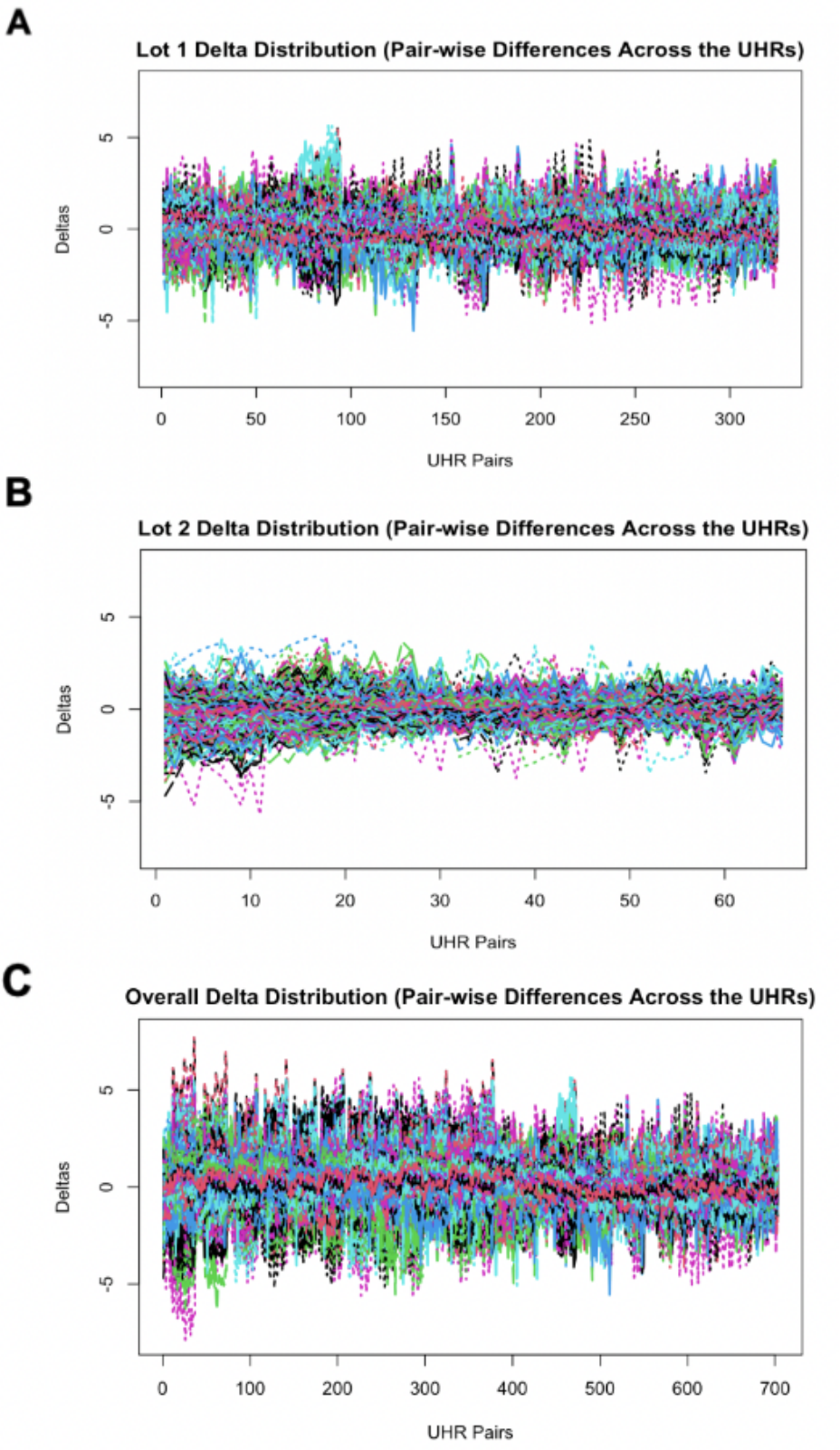

Figure 10. Distribution of the gene expression delta distributions with the second approach's cutoff metrics. (A) Plot of the gene expression delta distribution for Lot 1 only. The

x-axis indicates the UHR pair number, and the y-axis indicates the delta value. The delta value is the difference in gene expression between a gene's expression in one UHR and the same gene's expression in another UHR. **(B)** Plot of the gene expression delta distribution for Lot 2 only. The x-axis indicates the UHR pair number, and the y-axis indicates the delta value. The delta value is the difference in gene expression between a gene's expression in one UHR and the same gene's expression in another UHR. **(C)** Plot of the gene expression delta distribution for Lot 1 and Lot 2 combined. The x-axis indicates the UHR pair number, and the y-axis indicates the delta value. The delta value is the difference in gene expression between a gene's expression in one UHR and the same gene's expression in another UHR.

Deltas Computed Based on Rank Differences

Due to the scaling issues in gene expression data, we evaluated utilizing rank deltas. Using the rankings of genes across the UHRs, the differences between ranks were obtained, and a distribution was made based off of this. Using rank deltas attempts to address the batch and scaling issues that were seen in previous analyses.

Table 11. Impact on Hallmark Gene sets selected filtering evaluation.

Gene Set	Keeping genes with 0 expression	Removing genes with 0 expression	Keeping genes with average expression > 1	Keeping genes with average expression > 10	% genes lost (Removing genes with 0 expression)	% genes lost (Keeping genes with average expression > 1)	% genes lost (Keeping genes with average expression > 10)
Androgen Response	96	94	87	57	2.08%	9.00%	40.63%
Angiogenesis	36	33	28	15	8.33%	22.00%	58.33%
Apoptosis	159	155	144	79	2.52%	9.00%	50.31%
Bile Acid Metabolism	112	106	78	27	5.36%	30.00%	75.89%
DNA Repair	148	148	146	93	0.00%	1.00%	37.16%
E2F Targets	195	195	195	142	0.00%	0.00%	27.18%
Epithelial Mesenchymal Transition	197	196	167	88	0.51%	15.00%	55.33%
Estrogen Response Early	195	194	173	68	0.51%	11.00%	65.13%
Estrogen Response Late	197	195	170	80	1.02%	14.00%	59.39%
G2M Checkpoint	197	195	170	80	1.02%	14.00%	59.39%
Hypoxia	192	187	166	94	2.60%	14.00%	51.04%
IL2 STAT5 Signaling	195	181	145	68	7.18%	26.00%	65.13%
IL6 JAK STAT3 Signaling	87	78	60	23	10.34%	31.00%	73.56%
Inflammatory Response	200	177	112	43	11.50%	44.00%	78.50%
Interferon	96	93	79	37	3.13%	18.00%	61.46%

Alpha Response							
Interferon Gamma Response	200	191	156	67	4.50%	22.00%	66.50%
Kras Signaling Downregulated	194	145	79	12	25.26%	59.00%	93.81%
Kras Signaling Upregulated	196	179	137	51	8.67%	30.00%	73.98%
Mitotic Spindle	198	198	193	115	0.00%	3.00%	41.92%
Mtorc1 Signaling	195	195	192	155	0.00%	2.00%	20.51%
Myc Targets V1	198	196	196	184	1.01%	1.00%	7.07%
Myc Targets V2	58	58	58	36	0.00%	0.00%	37.93%
Oxidative Phosphorylation	186	185	184	155	0.54%	1.00%	16.67%
P53 Pathway	193	190	172	77	1.55%	11.00%	60.10%
PI3K AKT MTOR Signaling	104	99	90	58	4.81%	13.00%	44.23%
TGF Beta Signaling	54	54	51	30	0.00%	6.00%	44.44%
TNFA Signaling via NFKB	199	192	163	62	3.52%	18.00%	68.64%

With the filtering method of keeping genes with average expression > 10 in the UHR data, there are 5 hallmarks where greater than 70% of the genes are lot: Bile Acid Metabolism, IL6 JAK STAT 3 Signaling, Inflammatory Response, Kras Signaling Downregulated, and Kras Signaling Upregulated. There are 2 hallmarks with less than 20 % of the genes being lost: Myc Targets V1 and Oxidative Phosphorylation. When looking at filtering method of keeping genes with average expression > 10 in the UHR data, there are 5 pathways where greater than 70% of the genes were lost: Bile Acid Metabolism, IL6 JAK STAT 3 Signaling, Inflammatory Response, Kras Signaling Downregulated, and Kras Signaling Upregulated. There are 2 pathways where less than 20% of the genes were lost: Myc Targets V1 and Oxidative Phosphorylation.

When assessing the filtering, we wanted to consider if the low expressing genes in the UHR data were also low expressing in the patient samples, which could mitigate the impact.

In addition, we wanted to assess if the low expressing genes in the UHR data were low ranking in the patient samples. For instance, given a pathway, IL6 JAK STAT3 Signaling, unfiltered, there are 87 genes. After filtering for genes with average expression > 10, there are 23 genes left. The 64 genes that were filtered out were examined in the patient data. We hypothesized that the low expressing genes in the UHR would also be low expressing in the patient data. However, the genes that were low expressing in the UHR were both low and high expressing in the patient data. We also hypothesized that these genes would have low rankings in the patient data, but this was not the case. These genes were low and high ranking in the patient data.

Given these filtering assessments, a number of Hallmarks would not be able to be evaluated given the percentage of genes missing if the UHRs were utilized as the reference. In Bennett 2001, it is noted that missing data of 10% or higher can cause bias. The proportion of missing data tolerated varies (often listed as 5% to 20%) as it is dependent on the robustness of the statistical model being applied (Bennett 2001). Many hallmarks had more than 20% of their genes lost after filtering. For instance, the Kras Signaling Downregulated pathway had 93.81% of its genes removed after filtering out genes with < 10 average expression.

Of those genes that are low-expressing in the UHRs, we do not see them as consistently low-expressing in the patients.

Additionally, we need to consider that some genes may be variable due to the disease status of the patient, not just comparing the UHR data with the primary sample tissue, but patient specific considerations. The UHR data also appears to inflate patient outliers if low-expressing genes are not removed.

UHRs as a Reference Distribution for Disease Types

Figure 11. Low-expressing genes in the UHR distribution from the Interleukin-6 (IL6) Janus Kinase (JAK) Signal Transducer and Activator of Transcription 3 (STAT3) Signaling pathway as inputs to the Genotype-Tissue Expression (GTEx) project.

The low-expressing genes in the UHR distribution from the Interleukin-6 (IL6) Janus Kinase (JAK) Signal Transducer and Activator of Transcription 3 (STAT3) Signaling pathway are the following: CD38, TNFRSF12A, FAS, TNFRSF1B, BAK1, IL17RB, IL4R, IL12RB1, HMOX1, CSF2RB, IL7, EBI3, TGFB1, TYK2, DNTT, MAP3K8, CCL7, LTBR, STAM2, REG1A, IL1R2, IL1R1, IL18R1, LEPR, CSF3R, CNTFR, IL9R, IRF1, IL1B, IL13RA1, IL2RA, IL15RA, CD36, ACVR1B, IL6, PIM1, TLR2, CXCL9, INHBE, ACVRL1, PIK3R5, IFNAR1, PDGFC, TNFRSF21, CXCL13, CXCL3, PF4, CXCL1, CCR1, CSF2, CXCL10, CXCL11, CD14, JUN, IL17RA, CSF1, SOCS3, IL3RA, SOCS1, PLA2G2A, PTPN1, CSF2RA, CRLF2, IRF9, LTB, TNF, IL10RB, ITGB3. As a comparison to patient data, we utilized the Genotype-Tissue Expression (GTEx) project, which has samples collected from 54 on-diseased tissue sites from approximately 1000 individuals. This provides us with expression in these genes across tissue types (Figure 11). From the TPM levels observed in Figure 11, the UHR distribution may not be expressing the relevant genes for a number of cancers. Some of the genes in the list have 0 expression in the UHRs, but they are moderately or highly expressed in the GTEx patient distribution. If the UHR data was utilized as a reference distribution for a given tissue, genes would be left out of analyses. For instance, if a study were to be conducted on the lungs, a number of genes would be under-estimated or viewed as non-expressed in the null distribution inflating both normal and aberrant gene expression in patient samples.

CONCLUSION

If UHR replicates were to be used as a reference distribution to evaluate gene expression differences in the patient distributions, a number of the Hallmarks of Cancer would not be able to be evaluated given the large percentage of genes missing from appropriate filtering metrics.

In Bennett (2001) it is noted that missing data of 10% or higher can cause bias. The proportion of missing data tolerated varies (often listed as 5% to 20%) as it is dependent on the robustness of the statistical model being applied. As shown by our evaluation, the majority of the genesets had greater than 40% of their genes lost after appropriate filtering.

Additionally, of those genes that are low-expressing in the UHR replicates, we do not see them as consistently low-expressing the patients. We also need to consider that some genes will be variable due to the disease status of the patient, meaning that we need to consider not only the differences between UHR replicates and primary sample tissues, we also need to account for patient-specific considerations. The UHRs also appear to inflate patient outliers if lowly expressed genes are not accounted for.

Recognizing that the patient samples were limited, we assessed GTEX and found that across tissues the UHRs often under-estimated the expression compared to patient samples.

Referring back to our definition for the key characteristics for an optimal reference/null distribution, the UHR replicates do not appear to meet them, at least based on this data (Table 12).

Table 12. Alignment of the UHR replicates as a reference distribution with our key characteristics for an optimal reference/null distribution.

Key Characteristics	Do the UHR replicates match these criteria?	Reasoning
The sample gene expression distributions should be appropriate for the disease type.	No	UHR replicates are too low; we observed the loss of hallmarks and the replicates are not representative of other disease types (i.e. Lung in the GTEx analysis).
There is biological variability sample-to-sample, but is not due to technical artifacts.	No	Observed batch effects between Lot 1 and Lot 2.
Large enough distribution to	No	Observed batch effects

avoid sampling error.		between Lot 1 and Lots 2; also issues of sampling genes being dependent on gene filtering criteria.
Pairwise differences among samples should reflect normal biological variability, but be less than actual treatment effects, if present.	No	Observed technical variability and low expression in the UHR reference distribution. This led to artificial inflation of patient metrics.

Overall, some disease types would not be able to be evaluated accurately due to the high prevalence of the low-expressing genes in the UHR replicates distribution in the patient distribution.

Through the evolution of laboratory testing, we began to examine how laboratory testing has evolved. We observed an increased emphasis on rigor and appropriateness in terms of model validation, and evaluation of the appropriateness of the validation. As pipelines and data became more complex and heterogeneous, there was increased interest in constructing a null reference distribution. Increasing complexity in laboratory testing led us to consider how to effectively and appropriately address the gaps in pipeline development. Additionally, with the increasing need for validation, it became crucial to determine best practices to evaluate the methods.

Constructing an appropriate reference distribution for gene expression differences to assess therapeutic effects on tumors has been of great interest. However, as examined, the lack of a gold standard is a challenge. Validation will likely be test-specific, and we must consider not only the methods, but how to best evaluate the methods. To effectively transition from research to clinical usage, there is a great need for the validation and hardening of complex and heterogeneous algorithms. Moving computational approaches towards clinical usage has the potential to greatly improve patient outcomes, and the cost and efficacy of therapies.

SUPPLEMENTARY TABLES

Supplementary Table 1. Selected 27 Hallmarks of Cancer percentiles from log-transformed UHR null distribution data. The 27 hallmarks of cancer gene sets are Androgen Response, Angiogenesis, Apoptosis, Bile Acid Metabolism, DNA Repair, E2F Targets, Epithelial Mesenchymal Transition, Estrogen Response Early, Estrogen Response Late, G2-M Checkpoint, Hypoxia, Interleukin-2 (IL2) Signal Transducer and Activator of Transcription 5 (STAT5) Signaling, Interleukin-6 (IL6) Janus Kinase (JAK) Signal Transducer and Activator of Transcription 3 (STAT3) Signaling, Inflammatory Response, Interferon Alpha Response, Interferon Gamma Response, Kras Signaling Downregulated, Kras Signaling Upregulated, Mitotic Spindle, mammalian Target of Rapamycin Complex 1 (mTORC1) Signaling, Myc Targets Version 1, Myc Targets Version 2, Oxidative Phosphorylation, P53 Pathway, Phosphatidylinositol-3-kinase (PI3K) Protein Kinase B (AKT) MTOR Signaling, Transforming Growth Factor beta (TGF-β) Signaling, and Tumor Necrosis Factor alpha (TNFA) Signaling via NFKB.

Gene Set	80%	85%	90%	95%
Hall27_8TPM_3 Cases	0.72737408	0.832910649999 999	0.965429	1.16577631
Androgen Response	0.724289	0.8231581	0.948067	1.1456019
Angiogenesis	0.765911	0.86406155	0.9847476	1.169414665
Apoptosis	0.775832	0.883771	1.017094	1.21569275
Bile Acid Metabolism	0.6839744	0.78451285	0.9136398	1.0963923
DNA Repair	0.6728688	0.777899	0.911773400000 001	1.10796444
E2F Targets	0.6722568	0.7689838	0.8888254	1.0679854
Epithelial Mesenchymal Transition	0.801383	0.90709482	1.04253414	1.2399902
Estrogen Response Early	0.8440108	0.94997652	1.0869838	1.2864356
Estrogen Response Late	0.815328800000 001	0.9264732	1.0649914	1.2677708

G2M Checkpoint	0.815328800000001	0.9264732	1.0649914	1.2677708
Hypoxia	0.75973776	0.859781	0.9881236	1.1867816
IL2 STAT5 Signaling	0.7575114	0.859255275	0.9813087	1.16756195
IL6 JAK STAT3 Signaling	0.8120904	0.91979775	1.0496285	1.2487675
Inflammatory Response	0.7580475	0.855052000000001	0.974345	1.154014
Interferon Alpha Response	0.68399306	0.7801986	0.90189764	1.0888282
Interferon Gamma Response	0.7306946	0.83169055	0.96121617	1.155964462
Kras Signaling Downregulated	0.8496776	0.96214141	1.10766404	1.3322966
Kras SIgnaling Upregulated	0.7403022	0.8439323	0.9720568	1.1604856
Mitotic Spindle	0.719004	0.814107	0.938396999999999	1.119083
MTORC1 Signaling	0.6741474	0.7755839	0.9056416	1.1047935
Myc Targets V1	0.602424	0.6972704	0.8182644	1.0070586
Myc Targets V2	0.702392000000001	0.80156005	0.918627	1.088662
Oxidative Phosphorylation	0.604576	0.699522	0.825072	1.01589875
P53 Pathway	0.604576	0.699522	0.825072	1.01589875
PI3K AKT MTOR Signaling	0.7555876	0.8579957	0.9838778	1.1607478

TGF Beta Signaling	0.849766	0.9497597	1.0739196	1.28278151
TNFA Signaling via NFKB	0.8484084	0.9587466	1.09231997	1.289468455

Supplementary Table 2. Selected Hallmarks of Cancer percentiles from the UHR null distribution data. The 27 hallmarks of cancer gene sets are Androgen Response, Angiogenesis, Apoptosis, Bile Acid Metabolism, DNA Repair, E2F Targets, Epithelial Mesenchymal Transition, Estrogen Response Early, Estrogen Response Late, G2-M Checkpoint, Hypoxia, Interleukin-2 (IL2) Signal Transducer and Activator of Transcription 5 (STAT5) Signaling, Interleukin-6 (IL6) Janus Kinase (JAK) Signal Transducer and Activator of Transcription 3 (STAT3) Signaling, Inflammatory Response, Interferon Alpha Response, Interferon Gamma Response, Kras Signaling Downregulated, Kras Signaling Upregulated, Mitotic Spindle, mammalian Target of Rapamycin Complex 1 (mTORC1) Signaling, Myc Targets Version 1, Myc Targets Version 2, Oxidative Phosphorylation, P53 Pathway, Phosphatidylinositol-3-kinase (PI3K) Protein Kinase B (AKT) MTOR Signaling, Transforming Growth Factor beta (TGF-β) Signaling, and Tumor Necrosis Factor alpha (TNFA) Signaling via NFKB.

Hallmark	80%	85%	90%	95%
HALLMARK_Hall27_8TPM_3Cases	1.8064898184	2.040683279	2.328570468	2.7560888379
Androgen Response	1.862245232	2.083846789	2.3554361313	2.7828646302
Angiogenesis	1.837945834	2.05460191193	2.33359664	2.75632524226
Apoptosis	1.868215792	2.0739212332	2.330296812	2.74721387644
Bile Acid Metabolism	1.755443924	1.99638473735	2.289723455	2.73190025739
DNA Repair	1.838774856	2.062325028	2.330841305	2.742453996
E2F Targets	1.844434846	2.06788338012	2.344990634	2.76448907
Epithelial Mesenchymal Transition	1.866869158	2.0782954875	2.351251534	2.753325389
Estrogen Response Early	1.900927828	2.1060300825	2.3558508593	2.752555438
Estrogen Response Late	1.877595522	2.09443987	2.353831856	2.7562816842
G2M Checkpoint	1.877595522	2.09443987	2.353831856	2.7562816842
Hypoxia	1.87870099	2.0910988	2.34904939	2.75331048

IL2 STAT5 Signaling	1.842645624	2.0700490719	2.349712668	2.7838542957
IL6 JAK STAT3 Signaling	1.82276932	2.056915474	2.3415032977	2.751600495
Inflammatory Response	1.797070153	2.03626169	2.33166671	2.773685445
Interferon Alpha Response	1.812572998	2.047244868	2.3227862216	2.7520969251
Interferon Gamma Response	1.816140186	2.050227218	2.3309535702	2.7734586984
Kras Signaling Downregulated	1.769484958	2.016578381	2.320699056	2.7829069092
Kras Signaling Upregulated	1.800783616	2.0359538808	2.328227822	2.76611229
Mitotic Spindle	1.882997596	2.10814181	2.379096177	2.774779283
MTORC1 Signaling	1.855704736	2.0750123	2.353450986	2.76042797568
Myc Targets V1	1.8285190468	2.0445517555	2.32474921488	2.7579078498
Myc Targets V2	1.889029338	2.1068648007	2.3637815	2.752123158
Oxidative Phosphorylation	1.78321062	2.0160021468	2.299152766	2.730994914
P53 Pathway	1.78321062	2.0160021468	2.299152766	2.730994914
PI3K AKT MTOR Signaling	1.870169498	2.090032386	2.35462838	2.763074448
TGF Beta Signaling	1.8996557108	2.1045169978	2.352185682	2.746576411
TNFA Signaling via NFKB	1.89801849	2.1106782775	2.35807604	2.75313239675

Supplementary Table 3. UHR Rank Delta minimum and maximum distribution.

Hallmark	Keeping genes with 0 expression	Removing genes with 0 expression	Keeping genes with > 1 average expression	Keeping genes with > 10 average expression	X007_delta	X005_delta
Androgen Response	0, 36	0, 36	0, 36	0, 36	0, 30	0, 47
Angiogenesis	0, 9	0, 9	0, 9	0, 7	0, 3	0, 15
Apoptosis	0, 50	0, 50	0, 50	0, 32	0, 3	0, 114
Bile Acid Metabolism	0, 33	0, 33	0, 33	0, 13	0, 22	0, 32
DNA Repair	0, 74	0, 74	0, 74	0, 71	0, 44	0, 80
E2F Targets	0, 74	0, 74	0, 74	0, 74	0, 24	0, 53
Epithelial Mesenchymal Transition	0, 63	0, 63	0, 63	0, 34	0, 65	0, 110
Estrogen Response Early	0, 72	0, 72	0, 72	0, 34	0, 64	0, 92
Estrogen Response Late	0, 60	0, 60	0, 60	0, 41	0, 56	0, 60
G2M Checkpoint	0, 60	0, 60	0, 60	0, 41	0, 56	0, 60
Hypoxia	0, 70	0, 70	0, 70	0, 64	0, 31	0, 130
Il2 STAT5 Signaling	0, 69	0, 69	0, 63	0, 30	0, 40.5	0, 90
IL6 JAK STAT3 Signaling	0, 21	0, 21	0, 21	0, 14	0, 17	0, 46
Inflammatory Response	0, 72.5	0, 60	0, 44	0, 20	0, 36.5	0, 106

Interferon Alpha Response	0, 46	0, 26	0, 29	0, 22	0, 24	0, 69
Interferon Gamma Response	0, 101	0, 68	0, 61	0, 37	0, 62.5	0, 134
Kras Signaling Downregulated	0, 173.5	0, 79	0, 47	0, 9	0, 78	0, 71.5
Kras Signaling Upregulated	0, 70	0, 70	0, 70	0, 30	0, 46	0, 102
Mitotic Spindle	0, 67	0, 67	0, 67	0, 58	0, 24	0, 85
MTORC1 Signaling	0, 81	0, 81	0, 81	0, 81	0, 35	0, 70
Myc Targets V1	0, 118	0, 118	0, 118	0, 112	0, 41	0, 64
Myc Targets V2	0, 24	0, 24	0, 24	0, 15	0, 7	0, 13
Oxidative Phosphorylation	0, 69	0, 69	0, 69	0, 67	0, 33	0, 73
P53 Pathway	0, 70	0, 70	0, 70	0, 43	0, 31	0, 97
PI3K AKT MTOR Signaling	0, 33	0, 33	0, 33	0, 32	0, 10	0, 33
TGF Beta Signaling	0, 19	0, 19	0, 19	0, 12	0, 10	0, 33
TNFA Signaling via NFKB	0, 69	0, 69	0, 62	0, 33	0, 47	0, 106

Supplementary Table 4. UHR Rank delta percentile distribution; 80th, 85th, 90th, and 95th percentiles.

Hallmark	Keeping gene with 0 expression	Removing genes with 0 expression	Keeping genes with > 1 average expression	Keeping genes with > 10 average expression
Androgen Response	5, 6, 7, 10	5, 6, 7, 10	5, 6, 8, 10	5, 6, 8, 10
Angiogenesis	2, 2, 2, 3	2, 2, 2, 3	2, 2, 2, 3	1, 2, 2, 2
Apoptosis	8, 10, 12, 16	8, 10, 12, 16	9, 10, 12, 16	7, 9, 10, 13
Bile Acid Metabolism	6, 7, 9, 11	6, 7, 9, 11	6, 7, 8, 11	3, 3, 4, 5
DNA Repair	12, 15, 18, 23	12, 15, 18, 23	12, 15, 18, 23	12, 14, 17, 22
E2F Targets	13, 15, 18, 23	13, 15, 18, 23	13, 15, 18, 23	13, 15, 18, 23
Epithelial Mesenchymal Transition	9, 11, 13, 17	9, 11, 13, 17	10, 11, 14, 17	7, 9, 11, 14
Estrogen Response Early	11, 13, 16, 21	11, 14, 17, 22	12, 14, 17, 22	7, 8, 10, 13
Estrogen Response Late	11, 12, 15, 19	11, 13, 15, 19	11, 13, 16, 20	8, 10, 12, 15
G2M Checkpoint	11, 12, 15, 19	11, 13, 15, 19	11, 13, 16, 20	8, 10, 12, 15
Hypoxia	10, 12, 14, 19	10, 12, 14, 19	10, 12, 15, 20	10, 12, 15, 19
IL2 STAT5 Signaling	9, 11, 13, 17	9, 11, 13, 17	9, 11, 13, 17	7, 9, 11, 14
IL6 JAK STAT3 Signaling	5, 5, 6, 8	4, 5, 6, 8	4, 5, 6, 8	3, 3, 4, 5
Inflammatory Response	11.5, 14, 16, 22	10, 12, 15, 19	8, 9, 11, 14	5, 6, 8, 10
Interferon Alpha Response	6, 7, 8, 10	5, 6, 8, 10	6, 7, 8, 11	5, 6, 8, 10

Interferon Gamma Response	11, 13, 15, 20	11, 13, 15, 19	11, 13, 15, 20	8, 9, 11, 16
Kras Signaling Downregulated	17, 20, 25, 33	13, 15, 19, 25	10, 12, 14, 19	3, 4, 4, 6
Kras Signaling Upregulated	12, 13, 16, 21	11, 13, 16, 20	11, 13, 15, 20	7, 9, 10, 13
Mitotic Spindle	13, 15, 18, 23	13, 15, 18, 23	13, 15, 18, 24	13, 15, 18, 24
MTORC1 Signaling	12, 15, 18, 24	12, 15, 18, 24	12, 15, 18, 24	14, 16, 20, 26
Myc Targets V1	12, 15, 19, 25	12, 15, 19, 25	12, 15, 19, 25	13, 16, 19, 26
Myc Targets V2	4, 5, 6, 8	4, 5, 6, 8	4, 5, 6, 8	3, 4, 4, 6
Oxidative Phosphorylation	12, 14, 18, 24	12, 14, 18, 24	12, 14, 18, 24	13, 15, 18, 24
P53 Pathway	12, 14, 17, 23	12, 14, 17, 23	12, 15, 18, 24	9, 11, 14, 20
PI3K AKT MTOR Signaling	5, 6, 7, 10	5, 6, 7, 10	5, 6, 8, 10	5, 6, 7, 9
TGF Beta Signaling	3, 4, 5, 6	3, 4, 5, 6	3, 4, 5, 6	3, 4, 4, 6
TNFA Signaling via NFKB	11, 13, 16, 20	11, 13, 16, 20	11, 13, 16, 20	7, 8, 10, 13

APPENDIX

Appendix 1. The following Medical Genetics Search Filters were developed in conjunction with the staff of GeneReviews: Genetic Disease Online Reviews at GeneTests, University of Washington, Seattle and NCBI.
Ref: https://pubmed.ncbi.nlm.nih.gov/help/#medical-genetics-filters

Category	**PubMed equivalent**
Diagnosis	(Diagnosis AND Genetics)
Differential Diagnosis	(Differential Diagnosis[MeSH] OR Differential Diagnosis[Text Word] AND Genetics)
Clinical Description	(Natural History OR Mortality OR Phenotype OR Prevalence OR Penetrance AND Genetics)
Management	(therapy[Subheading] OR treatment[Text Word] OR treatment outcome OR investigational therapies AND Genetics)
Genetic Counseling	(Genetic Counseling OR Inheritance pattern AND genetics)
Molecular Genetics	(Medical Genetics OR genotype OR genetics[Subheading] AND genetics)
Genetic Testing	(DNA Mutational Analysis OR Laboratory techniques and procedures OR Genetic Markers OR diagnosis OR testing OR test OR screening OR mutagenicity tests OR genetic techniques OR molecular diagnostic techniques AND genetics)

Medical Genetics	((Diagnosis AND genetics) OR (Differential Diagnosis[MeSH] OR Differential Diagnosis[Text Word] AND genetics) OR (Natural History OR Mortality OR Phenotype OR Prevalence OR Penetrance AND genetics) OR (therapy[Subheading] OR treatment[Text Word] OR treatment outcome OR investigational therapies AND genetics) OR (Genetic Counseling OR Inheritance pattern AND genetics) OR (Medical Genetics OR genotype OR genetics[Subheading] AND genetics) OR (DNA Mutational Analysis OR Laboratory techniques and procedures OR Genetic Markers OR diagnosis OR testing OR test OR screening OR mutagenicity tests OR genetic techniques OR molecular diagnostic techniques AND genetics))

References

CLIA - Clinical Laboratory Improvement Amendments. https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfCLIA/search.cfm

Carrier/Mutation-specific Testing. (2012). http://www.genedx.com/test-catalog/mutation-specific-testing/

AbdelWareth, L. O., Pallinalakam, F., Ibrahim, F., Anderson, P., Liaqat, M., Palmer, B., Harris, J., Bashir, S., Alatoom, A., & Algora, M. (2018). Fast track to accreditation: an implementation review of College of American Pathologists and International Organization for Standardization 15189 Accreditation. Archives of pathology & laboratory medicine, 142(9), 1047-1053.

Alam, M. S., Sultana, A., Reza, M. S., Amanullah, M., Kabir, S. R., & Mollah, M. N. H. (2022). Integrated bioinformatics and statistical approaches to explore molecular biomarkers for breast cancer diagnosis, prognosis and therapies. PloS one, 17(5), e0268967.

Albacker, L. A., Wu, J., Smith, P.,Warmuth, M., Stephens, P. J., Zhu, P., Yu, L., & Chmielecki J (2017). Loss of function JAK1 mutations occur at high frequency in cancers with microsatellite instability and are suggestive of immune evasion. PloS one, 12(11), e0176181.

Álvarez, R. H. (2010). Present and future evolution of advanced breast cancer therapy. Breast cancer research, 12(2), 1-18.

Arnold, M., Morgan, E., Rumgay, H., Mafra, A., Singh, D., Laversanne, M., Vignat, J., Gralow, J. R., Cardoso, F., & Siesling, S. (2022). Current and future burden of breast cancer: Global statistics for 2020 and 2040. The Breast, 66, 15-23.

Auffray, C., Chen, Z., & Hood, L. (2009). Systems medicine: the future of medical genomics and healthcare. Genome Medicine, 1, 1-11.

Aziz, N., Zhao, Q., Bry, L., Driscoll, D. K., Funke, B., Gibson, J. S., Grody, W. W., Hegde, M. R., Hoeltge, G. A., & Leonard, D. G. (2015). College of American Pathologists' laboratory standards for next-generation sequencing clinical tests. Archives of Pathology and Laboratory Medicine, 139(4), 481-493.

Barata, P. C., Gulati, S., Elliott, A., Rao, A., Hammers, H. J., Quinn, D. I., Gartrell, B. A., Zibelman, M. R., Wei, S., & Geynisman, D. M. (2021). Angiogenic and T-effector subgroups identified by gene expression profiling (GEP) and propensity for PBRM1 and BAP1 alterations in clear cell renal cell carcinoma (ccRCC). In: American Society of Clinical Oncology.

Barbash, F. (2015). Scientist falsified data for cancer research once described as “holy grail,” feds say. The Washington Post. https://www.washingtonpost.com/news/morning-mix/wp/2015/11/09/scientist-falsified-data-for-cancer-research-once-described-as-holy-grail-feds-say

Bardia, A., Iafrate, J. A., Sundaresan, T., Younger, J., & Nardi, V. (2016). Metastatic breast

Cancer with ESR1 mutation: clinical management considerations from the molecular and precision medicine (MAP) tumor Board at Massachusetts General Hospital. The oncologist, 21(9), 1035-1040.

Behjati, S., & Tarpey, P. S. (2013). What is next generation sequencing? Archives of Disease in Childhood-Education and Practice, 98(6), 236-238.

Ben-Sahra, I., & Manning, B. D. (2017). mTORC1 signaling and the metabolic control of cell growth. Current opinion in cell biology, 45, 72-82.

Bennett, D. A. (2001) Australian and New Zealand Journal of Public Health 25 (5), 464-469

Bentley, D. R. (2006). Whole-genome re-sequencing. Current opinion in genetics & development, 16(6), 545-552.

Berishaj, M., Gao, S. P., Ahmed, S., Leslie, K., Al-Ahmadie, H., Gerald, W. L., Bornmann, W., & Bromberg, J. F. (2007). Stat3 is tyrosine-phosphorylated through the interleukin-6/glycoprotein 130/Janus kinase pathway in breast cancer. Breast cancer research, 9, 1-8.

Bertucci, F., Finetti, P., Vermeulen, P., Van Dam, P., Dirix, L., Birnbaum, D., Viens, P., & Van Laere, S. (2014). Genomic profiling of inflammatory breast cancer: a review. The Breast, 23(5), 538-545.

Bierie, B., & Moses, H. L. (2006). TGF-β and cancer. Cytokine & growth factor reviews, 17(1-2), 29-40.

Bogdanovich, W. (1987). The Pap test misses much cervical cancer through lab's errors. Wall Street J.

Brahimi-Horn, M. C., Chiche, J., & Pouysségur, J. (2007). Hypoxia and cancer. Journal of molecular medicine, 85, 1301-1307.

Buitenhuis, M., Baltus, B., Lammers, J.-W. J., Coffer, P. J., & Koenderman, L. (2003). Signal transducer and activator of transcription 5a (STAT5a) is required for eosinophil differentiation of human cord blood–derived CD34+ cells. Blood, The Journal of the American Society of Hematology, 101(1), 134-142.

Burger, H. G. (2002). Androgen production in women. Fertility and sterility, 77, 3-5.

Byers, L. A., & Rudin, C. M. (2015). Small cell lung cancer: where do we go from here? Cancer, 121(5), 664-672.

Castro, M., Feldman, R., & Reddy, S. K. (2016). TOPO1 expression in primary and metastatic GI cancers. In: American Society of Clinical Oncology.

Chan, S. R., Vermi, W., Luo, J., Lucini, L., Rickert, C., Fowler, A. M., Lonardi, S., Arthur, C., Young, L. J., & Levy, D. E. (2012). STAT1-deficient mice spontaneously develop estrogen receptor α-positive luminal mammary carcinomas. Breast cancer research, 14, 1-21.

Chang, P. L. (2005). Clinical bioinformatics. Chang Gung Med J, 28(4), 201-211.

Chautard, E., Ouédraogo, Z. G., Biau, J., & Verrelle, P. (2014). Role of Akt in human malignant glioma: from oncogenesis to tumor aggressiveness. Journal of neuro-oncology, 117, 205-215.

Cheaib, B., Auguste, A., & Leary, A. (2015). The PI3K/Akt/mTOR pathway in ovarian cancer: therapeutic opportunities and challenges. Chinese journal of cancer, 34, 4-16.

Chiang, C. L.-L., Kandalaft, L. E., Tanyi, J., Hagemann, A. R., Motz, G. T., Svoronos, N., Montone, K., Mantia-Smaldone, G. M., Smith, L., & Nisenbaum, H. L. (2013). A dendritic cell vaccine pulsed with autologous hypochlorous acid-oxidized ovarian cancer lysate primes effective broad antitumor immunity: from bench to bedside. Clinical Cancer Research, 19(17), 4801-4815.

Claesson-Welsh, L., & Welsh, M. (2013). VEGFA and tumour angiogenesis. Journal of internal medicine, 273(2), 114-127.

Coleman, R. E., & Rubens, R. D. (1987). The clinical course of bone metastases from breast cancer. British journal of cancer, 55(1), 61-66.

Comşa, Ş., Cimpean, A. M., & Raica, M. (2015). The story of MCF-7 breast cancer cell line: 40 years of experience in research. Anticancer research, 35(6), 3147-3154.

Cooperberg, M. R., Erho, N., Chan, J. M., Feng, F. Y., Fishbane, N., Zhao, S. G., Simko, J. P., Cowan, J. E., Lehrer, J., & Alshalalfa, M. (2018). The diverse genomic landscape of clinically low-risk prostate cancer. European urology, 74(4), 444-452.

Costa, R. M., Chigançcas, V., da Silva Galhardo, R., Carvalho, H., & Menck, C. F. (2003). The eukaryotic nucleotide excision repair pathway. Biochimie, 85(11), 1083-1099.

Cruceriu, D., Baldasici, O., Balacescu, O., & Berindan-Neagoe, I. (2020). The dual role of tumor necrosis factor-alpha (TNF-α) in breast cancer: molecular insights and therapeutic approaches. Cellular Oncology, 43, 1-18.

Curnutte, M. A., Frumovitz, K. L., Bollinger, J. M., McGuire, A. L., & Kaufman, D. J. (2014). Development of the clinical next-generation sequencing industry in a shifting policy climate. Nature Biotechnology, 32(10), 980-982.

Davies, K. (2012). Ambry Genetics Catches the Clinical Sequencing Wave. http://www.bio-itworld.com/news/06/13/12/Ambry-Genetics-catches-clinical-sequencing-wave.html

Degoulet, P., & Fieschi, M. (2012). Introduction to clinical informatics. Springer Science & Business Media.

Dono, A., Mitra, S., Shah, M., Takayasu, T., Zhu, J.-J., Tandon, N., Patel, C. B., Esquenazi, Y., & Ballester, L. Y. (2021). PTEN mutations predict benefit from tumor treating fields (TTFields) therapy in patients with recurrent glioblastoma. Journal of neuro-oncology, 153, 153-160.

Eberle, M. A., Fritzilas, E., Krusche, P., Källberg, M., Moore, B. L., Bekritsky, M. A., Iqbal, Z., Chuang, H.-Y., Humphray, S. J., & Halpern, A. L. (2017). A reference data set of 5.4 million phased human variants validated by genetic inheritance from sequencing a three-generation 17-member pedigree. Genome research, 27(1), 157-164.

Elledge, R. M., & Craig Allred, D. (1998). Prognostic and predictive value of p53 and p21 in breast cancer. Prognostic variables in node-negative and node-positive breast cancer, 169-188.

Elliott, C. G. (2005). Genetics of pulmonary arterial hypertension: current and future implications. Seminars in Respiratory and Critical Care Medicine,

Fábián, Á., Vereb, G., & Szöllősi, J. (2013). The hitchhikers guide to cancer stem cell theory: markers, pathways and therapy. Cytometry Part A, 83(1), 62-71.

Favaro, E., Lord, S., Harris, A. L., & Buffa, F. M. (2011). Gene expression and hypoxia in breast cancer. Genome Medicine, 3, 1-12.

Feldman, M. E., Apsel, B., Uotila, A., Loewith, R., Knight, Z. A., Ruggero, D., & Shokat, K. M. (2009). Active-site inhibitors of mTOR target rapamycin-resistant outputs of mTORC1 and mTORC2. PLoS biology, 7(2), e1000038.

Folkman, J., Merler, E., Abernathy, C., & Williams, G. (1971). Isolation of a tumor factor responsible for angiogenesis. The Journal of experimental medicine, 133(2), 275.

Forooghian, F., Kertes, P. J., Eng, K. T., Agrón, E., & Chew, E. Y. (2010). Alterations in the intraocular cytokine milieu after intravitreal bevacizumab. Investigative ophthalmology & visual science, 51(5), 2388-2392.

Gabay, M., Li, Y., & Felsher, D. W. (2014). MYC activation is a hallmark of cancer initiation and maintenance. Cold Spring Harbor perspectives in medicine, 4(6), a014241.

Gaerig, C. (2012). Beckman Coulter Genomics gears up for genetic sequencing: CLIA certification paves the way for marketing BRAF Sanger-based sequencing assays. Clinical Lab Products, 42(8), 46-47.

García-Tuñón, I., Ricote, M., Ruiz A, A., Fraile, B., Paniagua, R., & Royuela, M. (2007). Influence of IFN-gamma and its receptors in human breast cancer. BMC cancer, 7(1), 1-11.

Gargis, A. S., Kalman, L., Berry, M. W., Bick, D. P., Dimmock, D. P., Hambuch, T., Lu, F., Lyon, E., Voelkerding, K. V., & Zehnbauer, B. A. (2012). Assuring the quality of next-generation sequencing in clinical laboratory practice. Nature Biotechnology, 30(11), 1033-1036.

Gärtner, R., Jensen, M.-B., Nielsen, J., Ewertz, M., Kroman, N., & Kehlet, H. (2009). Prevalence of and factors associated with persistent pain following breast cancer surgery. Jama, 302(18), 1985-1992.

Gasco, M., Shami, S., & Crook, T. (2002). The p53 pathway in breast cancer. Breast cancer research, 4, 1-7.

Gatalica, Z., Stafford, P., & Vranic, S. (2022). Alpha-methylacyl-CoA racemase (AMACR) protein is upregulated in early proliferative lesions of the breast irrespective of apocrine differentiation. Human Pathology, 129, 40-46.

Gerber, D. E., Beg, M. S., Fattah, F., Frankel, A. E., Fatunde, O., Arriaga, Y., Dowell, J. E., Bisen, A., Leff, R. D., & Meek, C. C. (2018). Phase 1 study of ARQ 761, a β-lapachone analogue that promotes NQO1-mediated programmed cancer cell necrosis. British journal of cancer, 119(8), 928-936.

Gesbert, F., Delespine-Carmagnat, M., & Bertoglio, J. (1998). Recent advances in the understanding of interleukin-2 signal transduction. Journal of clinical immunology, 18,

307-320.
Giardine, B., Riemer, C., Hardison, R. C., Burhans, R., Elnitski, L., Shah, P., Zhang, Y., Blankenberg, D., Albert, I., & Taylor, J. (2005). Galaxy: a platform for interactive large-scale genome analysis. Genome research, 15(10), 1451-1455.
Gonzalez, A., Navas-Molina, J. A., Kosciolek, T., McDonald, D., Vázquez-Baeza, Y., Ackermann, G., DeReus, J., Janssen, S., Swafford, A. D., & Orchanian, S. B. (2018). Qiita: rapid, web-enabled microbiome meta-analysis. Nature methods, 15(10), 796-798.
Grard, G., Fair, J. N., Lee, D., Slikas, E., Steffen, I., Muyembe, J.-J., Sittler, T., Veeraraghavan, N., Ruby, J. G., & Wang, C. (2012). A novel rhabdovirus associated with acute hemorrhagic fever in central Africa.
Greenwood, C., Metodieva, G., Al-Janabi, K., Lausen, B., Alldridge, L., Leng, L., Bucala, R., Fernandez, N., & Metodiev, M. V. (2012). Stat1 and CD74 overexpression is co-dependent and linked to increased invasion and lymph node metastasis in triple-negative breast cancer. Journal of Proteomics, 75(10), 3031-3040.
Group, C. D. P. R. (1977). Gallbladder disease as a side effect of drugs influencing lipid metabolism experience in the coronary drug project. New England Journal of Medicine, 296(21), 1185-1190.
Gruber, J. J., Gross, W., McMillan, A., Ford, J. M., & Telli, M. L. (2021). A phase II clinical trial of talazoparib monotherapy for PALB2 mutation-associated advanced breast cancer. In: Wolters Kluwer Health.
Gucalp, A., & Traina, T. A. (2010). Triple-negative breast cancer: role of the androgen receptor. The Cancer Journal, 16(1), 62-65.
Gustafsson, J.-Å., & Warner, M. (2000). Estrogen receptor β in the breast: role in estrogen responsiveness and development of breast cancer. The Journal of steroid biochemistry and molecular biology, 74(5), 245-248.
Hameed, H. A., Islam, M. M., Chhotaray, C., Wang, C., Liu, Y., Tan, Y., Li, X., Tan, S., Delorme, V., & Yew, W. W. (2018). Molecular targets related drug resistance mechanisms in MDR-, XDR-, and TDR-Mycobacterium tuberculosis strains. Frontiers in cellular and infection microbiology, 8, 114.
Hamlin, W. B. (1999). Requirements for accreditation by the college of American pathologists laboratory accreditation program. Archives of Pathology and Laboratory Medicine, 123(6), 465-467.
Hamlin, W. B., & Duckworth, J. K. (1997). The college of American pathologists, 1946-1996. Archives of pathology & laboratory medicine, 121(7), 745.
Hanahan, D. (2022). Hallmarks of cancer: new dimensions. Cancer discovery, 12(1), 31-46.
Hanahan, D., & Weinberg, R. A. (2011). Hallmarks of cancer: the next generation. Cell, 144(5), 646-674.
Harbeck, N., Burstein, H. J., Hurvitz, S. A., Johnston, S., & Vidal, G. A. (2022). A look at current and potential treatment approaches for hormone receptor-positive, HER2-negative early breast cancer. Cancer, 128, 2209-2223.

Hassan, B., Akcakanat, A., Holder, A. M., & Meric-Bernstam, F. (2013). Targeting the PI3-kinase/Akt/mTOR signaling pathway. Surgical Oncology Clinics, 22(4), 641-664.

Hayashi, S., Eguchi, H., Tanimoto, K., Yoshida, T., Omoto, Y., Inoue, A., Yoshida, N., & Yamaguchi, Y. (2003). The expression and function of estrogen receptor alpha and beta in human breast cancer and its clinical application. Endocrine-Related Cancer, 10(2), 193-202.

Hix, L. M., Karavitis, J., Khan, M. W., Shi, Y. H., Khazaie, K., & Zhang, M. (2013). Tumor STAT1 transcription factor activity enhances breast tumor growth and immune suppression mediated by myeloid-derived suppressor cells. Journal of Biological Chemistry, 288(17), 11676-11688.

Ho, J., de Moura, M. B., Lin, Y., Vincent, G., Thorne, S., Duncan, L. M., Hui-Min, L., Kirkwood, J. M., Becker, D., & Van Houten, B. (2012). Importance of glycolysis and oxidative phosphorylation in advanced melanoma. Molecular cancer, 11(1), 1-13.

Hollern, D. P., Honeysett, J., Cardiff, R. D., & Andrechek, E. R. (2014). The E2F transcription factors regulate tumor development and metastasis in a mouse model of metastatic breast cancer. Molecular and cellular biology, 34(17), 3229-3243.

Hosoya, N., & Miyagawa, K. (2014). Targeting DNA damage response in cancer therapy. Cancer science, 105(4), 370-388.

Hui, H., Qingbin, K., Hongying, Z., Jiao, W., Ting, L., & Yangfu, J. (2019). Targeting mTOR for cancer therapy. Journal of Hematology & Oncology, 12(1), 71.

Imada, K., & Leonard, W. J. (2000). The jak-STAT pathway. Molecular Immunology, 37(1-2), 1-11.

Imamura, T., Hikita, A., & Inoue, Y. (2012). The roles of TGF-β signaling in carcinogenesis and breast cancer metastasis. Breast cancer, 19, 118-124.

Jardim, D. L. F., Conley, A., & Subbiah, V. (2013). Comprehensive characterization of malignant phyllodes tumor by whole genomic and proteomic analysis: biological implications for targeted therapy opportunities. Orphanet journal of rare diseases, 8, 1-8.

Ji, Q., Aoyama, C., Nien, Y.-D., Liu, P. I., Chen, P. K., Chang, L., Stanczyk, F. Z., & Stolz, A. (2004). Selective loss of AKR1C1 and AKR1C2 in breast cancer and their potential effect on progesterone signaling. Cancer research, 64(20), 7610-7617.

Johnson, D. E., O'Keefe, R. A., & Grandis, J. R. (2018). Targeting the IL-6/JAK/STAT3 signalling axis in cancer. Nature reviews Clinical oncology, 15(4), 234-248.

Junk, D. J., Bryson, B. L., Smigiel, J. M., Parameswaran, N., Bartel, C. A., & Jackson, M. W. (2017). Oncostatin M promotes cancer cell plasticity through cooperative STAT3-SMAD3 signaling. Oncogene, 36(28), 4001-4013.

Kalluri, R., & Neilson, E. G. (2003). Epithelial-mesenchymal transition and its implications for fibrosis. The Journal of Clinical Investigation, 112(12), 1776-1784.

Kamiya, K., Ozasa, K., Akiba, S., Niwa, O., Kodama, K., Takamura, N., Zaharieva, E. K.,
Kimura, Y., & Wakeford, R. (2015). Long-term effects of radiation exposure on health. The lancet, 386(9992), 469-478.

Kanagal-Shamanna, R., Singh, R. R., Routbort, M. J., Patel, K. P., Medeiros, L. J., & Luthra, R. (2016). Principles of analytical validation of next-generation sequencing based mutational analysis for hematologic neoplasms in a CLIA-certified laboratory. Expert Review of Molecular Diagnostics, 16(4), 461-472.

Karsenti, E., & Vernos, I. (2001). The mitotic spindle: a self-made machine. Science, 294(5542), 543-547.

Katsanis, S. H., & Katsanis, N. (2013). Molecular genetic testing and the future of clinical genomics. Nature Reviews Genetics, 14(6), 415-426.

Keegan, K. P., Glass, E. M., & Meyer, F. (2016). MG-RAST, a metagenomics service for analysis of microbial community structure and function. Microbial environmental genomics (MEG), 207-233.

Kerr, J. F., Winterford, C. M., & Harmon, B. V. (1994). Apoptosis. Its significance in cancer and cancer therapy. Cancer, 73(8), 2013-2026.

Khan, K. H., Wong, M., Rihawi, K., Bodla, S., Morganstein, D., Banerji, U., & Molife, L. R. (2016). Hyperglycemia and phosphatidylinositol 3-kinase/protein kinase B/mammalian target of rapamycin (PI3K/AKT/mTOR) inhibitors in phase I trials: incidence, predictive factors, and management. The oncologist, 21(7), 855-860.

Khodadadian, A., Darzi, S., Haghi-Daredeh, S., Sadat Eshaghi, F., Babakhanzadeh, E., Mirabutalebi, S. H., & Nazari, M. (2020). Genomics and transcriptomics: the powerful technologies in precision medicine. International Journal of General Medicine, 627-640.

Khodarev, N., Ahmad, R., Rajabi, H., Pitroda, S., Kufe, T., McClary, C., Joshi, M. D., MacDermed, D., Weichselbaum, R., & Kufe, D. (2010). Cooperativity of the MUC1 oncoprotein and STAT1 pathway in poor prognosis human breast cancer. Oncogene, 29(6), 920-929.

Kim, J., Lee, J.-h., & Iyer, V. R. (2008). Global identification of Myc target genes reveals its direct role in mitochondrial biogenesis and its E-box usage in vivo. PloS one, 3(3), e1798.

Kinder, M., Chislock, E., Bussard, K. M., Shuman, L., & Mastro, A. M. (2008). Metastatic breast cancer induces an osteoblast inflammatory response. Experimental cell research, 314(1), 173-183.

Knope, K., Whelan, P., Smith, D., Nicholson, J., Moran, R., Doggett, S., Sly, A., Hobby, M.,

Kurucz, N., & Wright, P. (2013). Arboviral diseases and malaria in Australia, 2010-11: annual report of the National Arbovirus and Malaria Advisory Committee. Communicable diseases intelligence quarterly report, 37(1).

Knüpfer, H., & Preiß, R. (2007). Significance of interleukin-6 (IL-6) in breast cancer. Breast cancer research and treatment, 102, 129-135.

Koboldt, D. C. (2020). Best practices for variant calling in clinical sequencing. Genome Medicine, 12(1), 1-13.

Kostic, A. D., Gevers, D., Pedamallu, C. S., Michaud, M., Duke, F., Earl, A. M., Ojesina, A. I., Jung, J., Bass, A. J., & Tabernero, J. (2012). Genomic analysis identifies association of

Fusobacterium with colorectal carcinoma. Genome research, 22(2), 292-298.
Kothari, N., Schell, M. J., Teer, J. K., Yeatman, T., Shibata, D., & Kim, R. (2014). Comparison of KRAS mutation analysis of colorectal cancer samples by standard testing and next-generation sequencing. Journal of clinical pathology, 67(9), 764-767.
Krokan, H. E., & Bjørås, M. (2013). Base excision repair. Cold Spring Harbor perspectives in biology, 5(4), a012583.
Kumar, V., Yu, J., Phan, V., Tudor, I. C., Peterson, A., & Uppal, H. (2017). Androgen receptor immunohistochemistry as a companion diagnostic approach to predict clinical response to enzalutamide in triple-negative breast cancer. JCO Precision Oncology, 1, 1-19.
Labrie, M., Brugge, J. S., Mills, G. B., & Zervantonakis, I. K. (2022). Therapy resistance: opportunities created by adaptive responses to targeted therapies in cancer. Nature Reviews Cancer, 22(6), 323-339.
Lamouille, S., Xu, J., & Derynck, R. (2014). Molecular mechanisms of epithelial–mesenchymal transition. Nature reviews Molecular cell biology, 15(3), 178-196.
Lanzino, M., Sisci, D., Morelli, C., Garofalo, C., Catalano, S., Casaburi, I., Capparelli, C., Giordano, C., Giordano, F., & Maggiolini, M. (2010). Inhibition of cyclin D1 expression by androgen receptor in breast cancer cells—identification of a novel androgen response element. Nucleic acids research, 38(16), 5351-5365.
Lapeire, L., Hendrix, A., Lambein, K., Van Bockstal, M., Braems, G., Van Den Broecke, R., Limame, R., Mestdagh, P., Vandesompele, J., & Vanhove, C. (2014). Cancer-Associated Adipose Tissue Promotes Breast Cancer Progression by Paracrine Oncostatin M and Jak/STAT3 SignalingParacrine Oncostatin M Promotes Breast Cancer Progression. Cancer research, 74(23), 6806-6819.
Latimer, J. J., Johnson, J. M., Kelly, C. M., Miles, T. D., Beaudry-Rodgers, K. A., Lalanne, N. A., Vogel, V. G., Kanbour-Shakir, A., Kelley, J. L., & Johnson, R. R. (2010). Nucleotide excision repair deficiency is intrinsic in sporadic stage I breast cancer. Proceedings of the National Academy of Sciences, 107(50), 21725-21730.
Lawson, N. S., & Howanitz, P. J. (1997). The College of American Pathologists, 1946-1996: Quality Assurance Service. Archives of pathology & laboratory medicine, 121(9), 1000.
Lee, J. J., Loh, K., & Yap, Y.-S. (2015). PI3K/Akt/mTOR inhibitors in breast cancer. Cancer biology & medicine, 12(4), 342.
Lee, K.-m., Giltnane, J. M., Balko, J. M., Schwarz, L. J., Guerrero-Zotano, A. L., Hutchinson, K. E., Nixon, M. J., Estrada, M. V., Sánchez, V., & Sanders, M. E. (2017). MYC and MCL1 cooperatively promote chemotherapy-resistant breast cancer stem cells via regulation of mitochondrial oxidative phosphorylation. Cell metabolism, 26(4), 633-647. e637.
Levin, M. K., Wang, K., Yelensky, R., Cao, Y., Ramos, C., Hoke, G. D., Pippen, J. E., Blum, J. L., Brooks, B. D., & Palmer, G. A. (2014). Estrogen receptor-positive (ER+) metastatic breast cancer (MBC) patients (pts) with extreme responses (ERs) to capecitabine having tumors with genomic alterations in DNA repair and chromatin remodeling genes. In: American Society of Clinical Oncology.

Levin, M. K., Wang, K., Yelensky, R., Cao, Y., Ramos, C., Hoke, N., Pippen Jr, J., Blum, J. L., Brooks, B., & Palmer, G. (2015). Genomic alterations in DNA repair and chromatin remodeling genes in estrogen receptor-positive metastatic breast cancer patients with exceptional responses to capecitabine. Cancer Medicine, 4(8), 1289-1293.

Li, X., Yang, Q., Yu, H., Wu, L., Zhao, Y., Zhang, C., Yue, X., Liu, Z., Wu, H., & Haffty, B. G. (2014). LIF promotes tumorigenesis and metastasis of breast cancer through the AKT-mTOR pathway. Oncotarget, 5(3), 788.

Li, Y., Huang, J., Yang, D., Xiang, S., Sun, J., Li, H., & Ren, G. (2018). Expression patterns of E2F transcription factors and their potential prognostic roles in breast cancer. Oncology letters, 15(6), 9216-9230.

Liao, P., Satten, G. A., & Hu, Y. J. (2017). PhredEM: a phred-score-informed genotype-calling approach for next-generation sequencing studies. Genetic epidemiology, 41(5), 375-387.

Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J. P., & Tamayo, P. (2015). The molecular signatures database hallmark gene set collection. Cell systems, 1(6), 417-425.

Lin, X., Lin, X., Guo, L., Wang, Y., & Zhang, G. (2022). Distinct clinicopathological characteristics, genomic alteration and prognosis in breast cancer with concurrent TP53 mutation and MYC amplification. Thoracic Cancer, 13(24), 3441-3450.

Liu, Y., Tran, H., Lin, Y., Martin, A., Zurita, A., Sternberg, C., Amado, R., Pandite, L., Heymach, J., & Team, V. (2011). Circulating baseline plasma cytokines and angiogenic factors (CAF) as markers of tumor burden and therapeutic response in a phase III study of pazopanib for metastatic renal cell carcinoma (mRCC). Journal of Clinical Oncology, 29(15_suppl), 4553-4553.

Liu, Y., Zhao, S., Jiang, C. C., Yao, Y., McIntosh, J., Jordan, A. A., Li, Y., Che, Y., Jain, P., & Wang, L. (2020). Interrogation of Dysregulated Pathways Enables Precision Medicine in Mantle Cell Lymphoma. Blood, 136, 33.

Lo Nigro, C., Monteverde, M., Lee, S., Lattanzio, L., Vivenza, D., Comino, A., Syed, N., McHugh, A., Wang, H., & Proby, C. (2012). NT5E CpG island methylation is a favourable breast cancer biomarker. British journal of cancer, 107(1), 75-83.

Löbrich, M., & Jeggo, P. A. (2007). The impact of a negligent G2/M checkpoint on genomic instability and cancer induction. Nature Reviews Cancer, 7(11), 861-869.

Lokhandwala, P. M., Riel, S. L., Haley, L., Lu, C., Chen, Y., Silberstein, J., Zhu, Y., Zheng, G., Lin, M.-T., & Gocke, C. D. (2017). Analytical validation of androgen receptor splice variant 7 detection in a clinical laboratory improvement amendments (CLIA) laboratory setting. The Journal of Molecular Diagnostics, 19(1), 115-125.

Loman, N. J., Constantinidou, C., Chan, J. Z., Halachev, M., Sergeant, M., Penn, C. W., Robinson, E. R., & Pallen, M. J. (2012). High-throughput bacterial genome sequencing: an embarrassment of choice, a world of opportunity. Nature Reviews Microbiology, 10(9), 599-606.

Long, S. A., Rieck, M., Sanda, S., Bollyky, J. B., Samuels, P. L., Goland, R., Ahmann, A., Rabinovitch, A., Aggarwal, S., & Phippard, D. (2012). Rapamycin/IL-2 combination therapy in patients with type 1 diabetes augments Tregs yet transiently impairs β-cell function. Diabetes, 61(9), 2340-2348.

Longatto Filho, A., Lopes, J. M., & Schmitt, F. C. (2010). Angiogenesis and breast cancer. Journal of oncology, 2010.

Loong, H. H.-F., Du, N., Cheng, C., Lin, H., Guo, J., Lin, G., Li, M., Jiang, T., Shi, Z., & Cui, Y. (2020). KRAS G12C mutations in Asia: a landscape analysis of 11,951 Chinese tumor samples. Translational Lung Cancer Research, 9(5), 1759.

Lyon, G. J., & Segal, J. P. (2013). Practical, ethical and regulatory considerations for the evolving medical and research genomics landscape. Applied & Translational Genomics, 2, 34-40.

Ma, C. X., Bose, R., Gao, F., Freedman, R. A., Telli, M. L., Kimmick, G., Winer, E., Naughton, M., Goetz, M. P., & Russell, C. (2017). Neratinib Efficacy and Circulating Tumor DNA Detection of HER2 Mutations in HER2 Nonamplified Metastatic Breast CancerNeratinib for HER2-Mutated, Nonamplified Breast Cancer. Clinical Cancer Research, 23(19), 5687-5695.

Ma, C. X., Gao, F., Luo, J., Northfelt, D. W., Goetz, M., Forero, A., Hoog, J., Naughton, M., Ademuyiwa, F., & Suresh, R. (2017). NeoPalAna: Neoadjuvant Palbociclib, a Cyclin-Dependent Kinase 4/6 Inhibitor, and Anastrozole for Clinical Stage 2 or 3 Estrogen Receptor–Positive Breast CancerNeoadjuvant Palbo and Anastrozole for ER+ Breast Cancer. Clinical Cancer Research, 23(15), 4055-4065.

Ma, C. X., Suman, V., Goetz, M. P., Northfelt, D., Burkard, M. E., Ademuyiwa, F., Naughton, M., Margenthaler, J., Aft, R., & Gray, R. (2017). A phase II trial of neoadjuvant MK-2206, an AKT inhibitor, with anastrozole in clinical stage II or III PIK3CA-mutant ER-positive and HER2-negative breast cancerneoadjuvant MK-2206 and anastrozole in ER+ Breast cancer. Clinical Cancer Research, 23(22), 6823-6832.

Ma, J., Chen, C., Liu, S., Ji, J., Wu, D., Huang, P., Wei, D., Fan, Z., & Ren, L. (2022). Identification of a five genes prognosis signature for triple-negative breast cancer using multi-omics methods and bioinformatics analysis. Cancer Gene Therapy, 1-12.

Mackay, H. J., Eisenhauer, E. A., Kamel-Reid, S., Tsao, M., Clarke, B., Karakasis, K., Werner, H. M., Trovik, J., Akslen, L. A., & Salvesen, H. B. (2014). Molecular determinants of outcome with mammalian target of rapamycin inhibition in endometrial cancer. Cancer, 120(4), 603-610.

MacLachlan, T. K., Somasundaram, K., Sgagias, M., Shifman, Y., Muschel, R. J., Cowan, K. H., & El-Deiry, W. S. (2000). BRCA1 effects on the cell cycle and the DNA damage response are linked to altered gene expression. Journal of Biological Chemistry, 275(4), 2777-2785.

Majidinia, M., & Yousefi, B. (2017). DNA repair and damage pathways in breast cancer development and therapy. DNA repair, 54, 22-29.

Maloy, S., & Hughes, K. (2013). Brenner's encyclopedia of genetics. Academic Press.
Mardis, E. R. (2008). Next-generation DNA sequencing methods. Annu. Rev. Genomics Hum. Genet., 9, 387-402.
Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., Berka, J., Braverman, M. S., Chen, Y.-J., & Chen, Z. (2005). Genome sequencing in microfabricated high-density picolitre reactors. Nature, 437(7057), 376-380.
Markowitz, V. M., Ivanova, N. N., Szeto, E., Palaniappan, K., Chu, K., Dalevi, D., Chen, I.-M. A., Grechkin, Y., Dubchak, I., & Anderson, I. (2007). IMG/M: a data management and analysis system for metagenomes. Nucleic acids research, 36(suppl_1), D534-D538.
Markowski, M. C., Silberstein, J. L., Eshleman, J. R., Eisenberger, M. A., Luo, J., & Antonarakis, E. S. (2017). Clinical utility of CLIA-grade AR-V7 testing in patients with metastatic castration-resistant prostate cancer. JCO Precision Oncology, 1, 1-9.
Martínez-Pérez, C., Kay, C., Meehan, J., Gray, M., Dixon, J. M., & Turnbull, A. K. (2021). The IL6-like cytokine family: Role and biomarker potential in breast cancer. Journal of personalized medicine, 11(11), 1073.
McGeehan, G. M., Becherer, J. D., Bast Jr, R. C., Boyer, C. M., Champion, B., Connolly, K. M., Conway, J. G., Furdon, P., Karp, S., & Kidao, S. (1994). Regulation of tumour necrosis factor-α processing by a metalloproteinase inhibitor. Nature, 370(6490), 558-561.
McNamara, K. M., Moore, N. L., Hickey, T. E., Sasano, H., & Tilley, W. D. (2014). Complexities of androgen receptor signalling in breast cancer. Endocrine-Related Cancer, 21(4), T161-T181.
McShane LM, Cavenagh MM, Lively TG, Eberhard DA, Bigbee WL, Williams PM, Mesirov JP, Polley MY, Kim KY, Tricoli JV, Taylor JM. Criteria for the use of omics-based predictors in clinical trials. Nature. 2013 Oct 17;502(7471):317-20.
Menegatti, S., Guillemot, V., Latis, E., Yahia-Cherbal, H., Mittermüller, D., Rouilly, V., Mascia, E., Rosine, N., Koturan, S., & Millot, G. A. (2021). Immune response profiling of patients with spondyloarthritis reveals signalling networks mediating TNF-blocker function in vivo. Annals of the Rheumatic Diseases, 80(4), 475-486.
Millis, S. Z., Ikeda, S., Reddy, S., Gatalica, Z., & Kurzrock, R. (2016). Landscape of phosphatidylinositol-3-kinase pathway alterations across 19 784 diverse solid tumors. JAMA oncology, 2(12), 1565-1573.
Miyasaki, J., Martin, W., Suchowersky, O., Weiner, W., & Lang, A. (2002). Practice parameter: initiation of treatment for Parkinson’s disease: an evidence-based review: report of the Quality Standards Subcommittee of the American Academy of Neurology. Neurology, 58(1), 11-17.
Molina JR, A. A. (2006). The ras/raf/mapk pathway. Journal of Thoracic Oncology, 1(1):7-9.
Moro, A., Gao, Z., Wang, L., Yu, A., Hsiung, S., Ban, Y., Yan, A., Sologon, C. M., Chen, X. S., & Malek, T. R. (2022). Dynamic transcriptional activity and chromatin remodeling of regulatory T cells after varied duration of interleukin-2 receptor signaling. Nature immunology, 23(5), 802-813.

Moroney, J., Fu, S., Moulder, S., Falchook, G., Helgason, T., Levenback, C., Hong, D., Naing, A., Wheler, J., & Kurzrock, R. (2012). Phase I Study of the Antiangiogenic Antibody Bevacizumab and the mTOR/Hypoxia-Inducible Factor Inhibitor Temsirolimus Combined with Liposomal Doxorubicin: Tolerance and Biological ActivityPhase I Study of Bevacizumab and Temsirolimus plus Doxil. Clinical Cancer Research, 18(20), 5796-5805.

Moses, H., & Barcellos-Hoff, M. H. (2011). TGF-β biology in mammary development and breast cancer. Cold Spring Harbor perspectives in biology, 3(1), a003277.

Motzer, R. J., Banchereau, R., Hamidi, H., Powles, T., McDermott, D., Atkins, M. B., Escudier, B., Liu, L.-F., Leng, N., & Abbas, A. R. (2020). Molecular subsets in renal cancer determine outcome to checkpoint and angiogenesis blockade. Cancer Cell, 38(6), 803-817. e804.

Naccache, S. N., Federman, S., Veeraraghavan, N., Zaharia, M., Lee, D., Samayoa, E., Bouquet, J., Greninger, A. L., Luk, K.-C., & Enge, B. (2014). A cloud-compatible bioinformatics pipeline for ultrarapid pathogen identification from next-generation sequencing of clinical samples. Genome research, 24(7), 1180-1192.

Necchi, A., Raggi, D., Giannatempo, P., Marandino, L., Farè, E., Gallina, A., Colecchia, M., Lucianò, R., Salonia, A., & Gandaglia, G. (2021). Can patients with muscle-invasive bladder cancer and fibroblast growth factor receptor-3 alterations still be considered for neoadjuvant pembrolizumab? A comprehensive assessment from the updated results of the PURE-01 study. European urology oncology, 4(6), 1001-1005.

Nicolini, A., Carpi, A., & Rossi, G. (2006). Cytokines in breast cancer. Cytokine & growth factor reviews, 17(5), 325-337.

Nitulescu, G. M., Margina, D., Juzenas, P., Peng, Q., Olaru, O. T., Saloustros, E., Fenga, C., Spandidos, D. A., Libra, M., & Tsatsakis, A. M. (2016). Akt inhibitors in cancer treatment: The long journey from drug discovery to clinical use. International journal of oncology, 48(3), 869-885.

Norbury, C. J., & Hickson, I. D. (2001). Cellular responses to DNA damage. Annual review of pharmacology and toxicology, 41(1), 367-401.

Ogony, J., Choi, H. J., Lui, A., Cristofanilli, M., & Lewis-Wambi, J. (2016). Interferon-induced transmembrane protein 1 (IFITM1) overexpression enhances the aggressive phenotype of SUM149 inflammatory breast cancer cells in a signal transducer and activator of transcription 2 (STAT2)-dependent manner. Breast cancer research, 18, 1-19.

Oinn, T., Addis, M., Ferris, J., Marvin, D., Senger, M., Greenwood, M., Carver, T., Glover, K., Pocock, M. R., & Wipat, A. (2004). Taverna: a tool for the composition and enactment of bioinformatics workflows. Bioinformatics, 20(17), 3045-3054.

Omenn, G. S., Nass, S. J., & Micheel, C. M. (2012). Evolution of translational omics: lessons learned and the path forward.

Oshi, M., Tokumaru, Y., Angarita, F. A., Yan, L., Matsuyama, R., Endo, I., & Takabe, K. (2020). Degree of early estrogen response predict survival after endocrine therapy in primary and

metastatic ER-positive breast cancer. Cancers, 12(12), 3557.
Pagnamenta, A. T., Lise, S., Harrison, V., Stewart, H., Jayawant, S., Quaghebeur, G., Deng, A. T., Murphy, V. E., Akha, E. S., & Rimmer, A. (2012). Exome sequencing can detect pathogenic mosaic mutations present at low allele frequencies. Journal of human genetics, 57(1), 70-72.
Pandey, V., Nutter, R. C., & Prediger, E. (2008). Applied biosystems SOLiD™ system: ligation-based sequencing. Next generation genome sequencing: towards personalized medicine, 29-42.
Parton, M., Dowsett, M., & Smith, I. (2001). Studies of apoptosis in breast cancer. Bmj, 322(7301), 1528-1532.
Parvizpour, S., Razmara, J., & Omidi, Y. (2018). Breast cancer vaccination comes to age: impacts of bioinformatics. BioImpacts: BI, 8(3), 223.
Patrono, C., Coller, B., & James, E. (2001). Dalen JE, FitzGerald GA, Valentin Fuster V, Gent M, Hirsh J and Roth G. Platelet-Active Drugs: The Relationships Among Dose, Effectiveness, and Side effects. Chest, 119, 39-63.
Piñeros, M., Parkin, D. M., Ward, K., Chokunonga, E., Ervik, M., Farrugia, H., Gospodarowicz, M., O'Sullivan, B., Soerjomataram, I., & Swaminathan, R. (2019). Essential TNM: a registry tool to reduce gaps in cancer staging information. The Lancet Oncology, 20(2), e103-e111.
Polyak, K. (2007). Breast cancer: origins and evolution. The Journal of Clinical Investigation, 117(11), 3155-3163.
Prives, C., & Hall, P. A. (1999). The p53 pathway. The Journal of pathology, 187(1), 112-126.
Provance, O. K., & Lewis-Wambi, J. (2019). Deciphering the role of interferon alpha signaling and microenvironment crosstalk in inflammatory breast cancer. Breast cancer research, 21, 1-10.
Qadir, A. S., Stults, A. M., Murmann, A. E., & Peter, M. E. (2020). The mechanism of how CD95/Fas activates the Type I IFN/STAT1 axis, driving cancer stemness in breast cancer. Scientific reports, 10(1), 1310.
Qin, L., Wang, J., Tian, X., Yu, H., Truong, C., Mitchell, J. J., Wierenga, K. J., Craigen, W. J., Zhang, V. W., & Wong, L.-J. C. (2016). Detection and quantification of mosaic mutations in disease genes by next-generation sequencing. The Journal of Molecular Diagnostics, 18(3), 446-453.
Quintáns, B., Ordóñez-Ugalde, A., Cacheiro, P., Carracedo, A., & Sobrido, M. (2014). Medical genomics: The intricate path from genetic variant identification to clinical interpretation. Applied & Translational Genomics, 3(3), 60-67.
Ramos, J., Yoo, C., Felty, Q., Gong, Z., Liuzzi, J. P., Poppiti, R., Thakur, I. S., Goel, R., Vaid, A. K., & Komotar, R. J. (2020). Sensitivity to differential NRF1 gene signatures contributes to breast cancer disparities. Journal of cancer research and clinical oncology, 146, 2777-2815.
Raninga, P. V., Lee, A., Sinha, D., Dong, L.-f., Datta, K. K., Lu, X., Kalita-de Croft, P., Dutt, M.,

Hill, M., & Pouliot, N. (2020). Marizomib suppresses triple-negative breast cancer via proteasome and oxidative phosphorylation inhibition. Theranostics, 10(12), 5259.
Rehm, H. L., Bale, S. J., Bayrak-Toydemir, P., Berg, J. S., Brown, K. K., Deignan, J. L., Friez, M. J., Funke, B. H., Hegde, M. R., & Lyon, E. (2013). ACMG clinical laboratory standards for next-generation sequencing. Genetics in medicine, 15(9), 733-747.
Reis-Filho, J. S. (2009). Next-generation sequencing. Breast cancer research, 11(3), 1-7.
Repka, T., Chiorean, E. G., Gay, J., Herwig, K. E., Kohl, V. K., Yee, D., & Miller, J. S. (2003). Trastuzumab and interleukin-2 in HER2-positive metastatic breast cancer: a pilot study. Clinical Cancer Research, 9(7), 2440-2446.
Reya, T., Morrison, S. J., Clarke, M. F., & Weissman, I. L. (2001). Stem cells, cancer, and cancer stem cells. Nature, 414(6859), 105-111.
Rivers, P. A., Dobalian, A., & Germinario, F. A. (2005). A review and analysis of the clinical laboratory improvement amendment of 1988: compliance plans and enforcement policy. Health Care Management Review, 30(2), 93-102.
Robertson, F. M., Petricoin Iii, E. F., Van Laere, S. J., Bertucci, F., Chu, K., Fernandez, S. V., Mu, Z., Alpaugh, K., Pei, J., & Circo, R. (2013). Presence of anaplastic lymphoma kinase in inflammatory breast cancer. Springerplus, 2(1), 1-12.
Ronckers, C. M., Erdmann, C. A., & Land, C. E. (2004). Radiation and breast cancer: a review of current evidence. Breast cancer research, 7, 1-12.
Ross, J. S., Ali, S. M., Wang, K., Khaira, D., Palma, N. A., Chmielecki, J., Palmer, G. A., Morosini, D., Elvin, J. A., & Fernandez, S. V. (2015). Comprehensive genomic profiling of inflammatory breast cancer cases reveals a high frequency of clinically relevant genomic alterations. Breast cancer research and treatment, 154, 155-162.
Roy, S., Coldren, C., Karunamurthy, A., Kip, N. S., Klee, E. W., Lincoln, S. E., Leon, A., Pullambhatla, M., Temple-Smolkin, R. L., & Voelkerding, K. V. (2018). Standards and guidelines for validating next-generation sequencing bioinformatics pipelines: a joint recommendation of the Association for Molecular Pathology and the College of American Pathologists. The Journal of Molecular Diagnostics, 20(1), 4-27.
Sarmadi, P., Tunali, G., Esendagli-Yilmaz, G., Yilmaz, K. B., & Esendagli, G. (2015). CRAM-A indicates IFN-γ-associated inflammatory response in breast cancer. Molecular Immunology, 68(2), 692-698.
Sarrió, D., Rodriguez-Pinilla, S. M., Hardisson, D., Cano, A., Moreno-Bueno, G., & Palacios, J. (2008). Epithelial-mesenchymal transition in breast cancer relates to the basal-like phenotype. Cancer research, 68(4), 989-997.
Sayegh, N., Chigarira, B., Hernandez, E. J., McFarland, T. R., Li, H., Sahu, K. K., Tripathi, N., Kumar, S. A., Nordblad, B., & Goel, D. (2022). Transcriptomic profiling of patients (pts) with de-novo metastatic castration-sensitive prostate cancer (DN-mCSPC) versus those with mCSPC that have relapsed from prior localized therapy (PLT-mCSPC). In: American Society of Clinical Oncology.
Sboner, A., & Elemento, O. (2016). A primer on precision medicine informatics. Briefings in

bioinformatics, 17(1), 145-153.
Schneckenleithner, C., Bago-Horvath, Z., Dolznig, H., Neugebauer, N., Kollmann, K., Kolbe, T., Decker, T., Kerjaschki, D., Wagner, K.-U., & Müller, M. (2011). Putting the brakes on mammary tumorigenesis: loss of STAT1 predisposes to intraepithelial neoplasias. Oncotarget, 2(12), 1043.
Schramm, G., Surmann, E.-M., Wiesberg, S., Oswald, M., Reinelt, G., Eils, R., & König, R. (2010). Analyzing the regulation of metabolic pathways in human breast cancer. BMC medical genomics, 3(1), 1-10.
Schulze, A., Oshi, M., Endo, I., & Takabe, K. (2020). MYC targets scores are associated with cancer aggressiveness and poor survival in ER-positive primary and metastatic breast cancer. International Journal of Molecular Sciences, 21(21), 8127.
Schwartz, M. K. (1999). Genetic testing and the clinical laboratory improvement amendments of 1988: present and future. Clinical chemistry, 45(5), 739-745.
Sefid Dashti, M. J., & Gamieldien, J. (2017). A practical guide to filtering and prioritizing genetic variants. Biotechniques, 62(1), 18-30.
Serra, V., Markman, B., Scaltriti, M., Eichhorn, P. J., Valero, V., Guzman, M., Botero, M. L., Llonch, E., Atzori, F., & Di Cosimo, S. (2008). NVP-BEZ235, a dual PI3K/mTOR inhibitor, prevents PI3K signaling and inhibits the growth of cancer cells with activating PI3K mutations. Cancer research, 68(19), 8022-8030.
Serratì, S., De Summa, S., Pilato, B., Petriella, D., Lacalamita, R., Tommasi, S., & Pinto, R. (2016). Next-generation sequencing: advances and applications in cancer diagnosis. OncoTargets and therapy, 7355-7365.
Sgroi, D. C. (2010). Preinvasive breast cancer. Annual Review of Pathology: Mechanisms of Disease, 5, 193-221.
Sheen-Chen, S.-M., Chen, W.-J., Eng, H.-L., & Chou, F.-F. (1997). Serum concentration of tumor necrosis factor in patients with breast cancer. Breast cancer research and treatment, 43, 211-215.
Shen, W., Song, Z., Zhong, X., Huang, M., Shen, D., Gao, P., Qian, X., Wang, M., He, X., & Wang, T. (2022). Sangerbox: A comprehensive, interaction-friendly clinical bioinformatics analysis platform. Imeta, 1(3), e36.
Shin, H.-T., Choi, Y.-L., Yun, J. W., Kim, N. K., Kim, S.-Y., Jeon, H. J., Nam, J.-Y., Lee, C., Ryu, D., & Kim, S. C. (2017). Prevalence and detection of low-allele-fraction variants in clinical cancer samples. Nature communications, 8(1), 1377.
Siersbæk, R., Scabia, V., Nagarajan, S., Chernukhin, I., Papachristou, E. K., Broome, R., Johnston, S. J., Joosten, S. E., Green, A. R., & Kumar, S. (2020). IL6/STAT3 signaling hijacks estrogen receptor α enhancers to drive breast cancer metastasis. Cancer Cell, 38(3), 412-423. e419.
Silberstein, G. B., & Daniel, C. W. (1987). Reversible inhibition of mammary gland growth by transforming growth factor-β. Science, 237(4812), 291-293.
Singh, J., Singh, R., Gupta, P., Rai, S., Ganesher, A., Badrinarayan, P., Sastry, G. N., Konwar, R.,

& Panda, G. (2017). Targeting progesterone metabolism in breast cancer with L-proline derived new 14-azasteroids. Bioorganic & Medicinal Chemistry, 25(16), 4452-4463.

Singh, K. K., Ayyasamy, V., Owens, K. M., Koul, M. S., & Vujcic, M. (2009). Mutations in mitochondrial DNA polymerase-γ promote breast tumorigenesis. Journal of human genetics, 54(9), 516-524.

Siva, N. (2008). 1000 Genomes project. Nature Biotechnology, 26(3), 256-257.

Smith, B. D., Gross, C. P., Smith, G. L., Galusha, D. H., Bekelman, J. E., & Haffty, B. G. (2006). Effectiveness of radiation therapy for older women with early breast cancer. Journal of the National Cancer Institute, 98(10), 681-690.

Solaini, G., Sgarbi, G., & Baracca, A. (2011). Oxidative phosphorylation in cancer cells. Biochimica et Biophysica Acta (BBA)-Bioenergetics, 1807(6), 534-542.

Somasundaram, K. (2003). Breast cancer gene 1 (BRCA1): role in cell cycle regulation and DNA repair—perhaps through transcription. Journal of cellular biochemistry, 88(6), 1084-1091.

Song, I.-W., Nagamani, S. C., Nguyen, D., Grafe, I., Sutton, V. R., Gannon, F. H., Munivez, E., Jiang, M.-M., Tran, A., & Wallace, M. (2022). Targeting TGF-β for treatment of osteogenesis imperfecta. The Journal of Clinical Investigation, 132(7).

Straiton, J., Free, T., Sawyer, A., & Martin, J. (2019). From Sanger sequencing to genome databases and beyond. In: Future Science.

Strieter, R. M., Kunkel, S. L., & Bone, R. C. (1993). Role of tumor necrosis factor-alpha in disease states and inflammation. Critical care medicine, 21(10 Suppl), S447-463.

Subbiah, I. M., Tsimberidou, A., Subbiah, V., Janku, F., Roy-Chowdhuri, S., & Hong, D. S. (2017). Next generation sequencing of carcinoma of unknown primary reveals novel combinatorial strategies in a heterogeneous mutational landscape. Oncoscience, 4(5-6), 47.

Subbiah, V., Brown, R. E., Jiang, Y., Buryanek, J., Hayes-Jordan, A., Kurzrock, R., & Anderson, P. M. (2013). Morphoproteomic profiling of the mammalian target of rapamycin (mTOR) signaling pathway in desmoplastic small round cell tumor (EWS/WT1), Ewing's sarcoma (EWS/FLI1) and Wilms' tumor (WT1). PloS one, 8(7), e68985.

Sun, Y.-S., Zhao, Z., Yang, Z.-N., Xu, F., Lu, H.-J., Zhu, Z.-Y., Shi, W., Jiang, J., Yao, P.-P., & Zhu, H.-P. (2017). Risk factors and preventions of breast cancer. International journal of biological sciences, 13(11), 1387.

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., & Bray, F. (2021). Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians, 71(3), 209-249.

Tang, W., Putluri, V., Ambati, C. R., Dorsey, T. H., Putluri, N., & Ambs, S. (2019). Liver-and Microbiome-derived Bile Acids Accumulate in Human Breast Tumors and Inhibit Growth and Improve Patient SurvivalBile Acids and Breast Cancer. Clinical Cancer Research, 25(19), 5972-5983.

Tawara, K., Scott, H., Emathinger, J., Ide, A., Fox, R., Greiner, D., LaJoie, D., Hedeen, D., Nandakumar, M., & Oler, A. J. (2019). Co-expression of VEGF and IL-6 family cytokines is associated with decreased survival in HER2 negative breast cancer patients: subtype-specific IL-6 family cytokine-mediated VEGF secretion. Translational oncology, 12(2), 245-255.

Tawara, K., Scott, H., Emathinger, J., Wolf, C., LaJoie, D., Hedeen, D., Bond, L., Montgomery, P., & Jorcyk, C. (2019). HIGH expression of OSM and IL-6 are associated with decreased breast cancer survival: synergistic induction of IL-6 secretion by OSM and IL-1β. Oncotarget, 10(21), 2068.

Tervasmäki, A., Mantere, T., Hartikainen, J. M., Kauppila, S., Lee, H. M., Koivuluoma, S., Grip, M., Karihtala, P., Jukkola-Vuorinen, A., & Mannermaa, A. (2018). Rare missense mutations in RECQL and POLG associate with inherited predisposition to breast cancer. International Journal of Cancer, 142(11), 2286-2292.

Tokumaru, Y., Oshi, M., Katsuta, E., Yan, L., Satyananda, V., Matsuhashi, N., Futamura, M., Akao, Y., Yoshida, K., & Takabe, K. (2020). KRAS signaling enriched triple negative breast cancer is associated with favorable tumor immune microenvironment and better survival. American journal of cancer research, 10(3), 897.

Tonry, C., Finn, S., Armstrong, J., & Pennington, S. R. (2020). Clinical proteomics for prostate cancer: understanding prostate cancer pathology and protein biomarkers for improved disease management. Clinical Proteomics, 17, 1-31.

Tymoszuk, P., Charoentong, P., Hackl, H., Spilka, R., Müller-Holzner, E., Trajanoski, Z., Obrist, P., Revillion, F., Peyrat, J.-P., & Fiegl, H. (2014). High STAT1 mRNA levels but not its tyrosine phosphorylation are associated with macrophage infiltration and bad prognosis in breast cancer. BMC cancer, 14(1), 1-13.

Tzavlaki, K., & Moustakas, A. (2020). TGF-β signaling. Biomolecules 10, 487. In.

Ueda, S., Saeki, T., Takeuchi, H., Shigekawa, T., Yamane, T., Kuji, I., & Osaki, A. (2016). In vivo imaging of eribulin-induced reoxygenation in advanced breast cancer patients: a comparison to bevacizumab. British journal of cancer, 114(11), 1212-1218.

Uprety, D., & Adjei, A. A. (2020). KRAS: From undruggable to a druggable Cancer Target. Cancer treatment reviews, 89, 102070.

Vaupel, P., Briest, S., & Höckel, M. (2002). Hypoxia in breast cancer: pathogenesis, characterization and biological/therapeutic implications. Wiener Medizinische Wochenschrift, 152(13-14), 334-342.

Vaupel, P., & Mayer, A. (2007). Hypoxia in cancer: significance and impact on clinical outcome. Cancer and Metastasis Reviews, 26, 225-239.

Vijay, P., McIntyre, A. B., Mason, C. E., Greenfield, J. P., & Li, S. (2016). Clinical genomics: challenges and opportunities. Critical Reviews™ in Eukaryotic Gene Expression, 26(2).

Waks, A. G., & Winer, E. P. (2019). Breast cancer treatment: a review. Jama, 321(3), 288-300.

Wallace, S. S. (2014). Base excision repair: a critical player in many games. DNA repair, 19, 14-26.

Wang, J., Xiu, J., Baca, Y., Battaglin, F., Arai, H., Kawanishi, N., Soni, S., Zhang, W., Millstein, J., & Salhia, B. (2021). Large-scale analysis of KMT2 mutations defines a distinctive molecular subset with treatment implication in gastric cancer. Oncogene, 40(30), 4894-4905.

Wang, K., Nahas, M. K., Yelensky, R., Otto, G. A., Lipson, D., He, J., Ross, J., Stephens, P. J., Chow, K. F., & Zielonka, T. (2014). Novel chromatin modifying gene alterations and significant survival association of ATM and P53 in mantle cell lymphoma. Blood, 124(21), 3033.

Wang, L., Ma, L., Xu, F., Zhai, W., Dong, S., Yin, L., Liu, J., & Yu, Z. (2018). Role of long non-coding RNA in drug resistance in non-small cell lung cancer. Thoracic Cancer, 9(7), 761-768.

Wang, X., & Liotta, L. (2011). Clinical bioinformatics: a new emerging science. In (Vol. 1, pp. 1-3): BioMed Central.

Weinberg, A., Allshouse, A., Kinzie, K., Cho, A., Davies, J. K., & Mc Farland, E. J. (2015). Intrahepatic Cholestasis of Pregnancy and Serum Bile Acids in HIV-Infected Pregnant Women. Journal of AIDS & clinical research, 6(6).

Weinberg, R., & Hanahan, D. (2000). The hallmarks of cancer. Cell, 100(1), 57-70.

Whitaker-Menezes, D., Martinez-Outschoorn, U. E., Flomenberg, N., Birbe, R., Witkiewicz, A. K., Howell, A., Pavlides, S., Tsirigos, A., Ertel, A., & Pestell, R. G. (2011). Hyperactivation of oxidative mitochondrial metabolism in epithelial cancer cells in situ: visualizing the therapeutic effects of metformin in tumor tissue. Cell cycle, 10(23), 4047-4064.

Winship, A. L., Van Sinderen, M., Donoghue, J., Rainczuk, K., & Dimitriadis, E. (2016). Targeting Interleukin-11 Receptor-α Impairs Human Endometrial Cancer Cell Proliferation and Invasion In Vitro and Reduces Tumor Growth and Metastasis In VivoRole of IL11 in Endometrial Cancer. Molecular Cancer Therapeutics, 15(4), 720-730.

Witkiewicz, A. K., McMillan, E. A., Balaji, U., Baek, G., Lin, W.-C., Mansour, J., Mollaee, M., Wagner, K.-U., Koduru, P., & Yopp, A. (2015). Whole-exome sequencing of pancreatic cancer defines genetic diversity and therapeutic targets. Nature communications, 6(1), 6744.

Wylie, K. M., Mihindukulasuriya, K. A., Sodergren, E., Weinstock, G. M., & Storch, G. A. (2012). Sequence analysis of the human virome in febrile and afebrile children. PloS one, 7(6), e27735.

Xu, B., Liu, L., Huang, X., Ma, H., Zhang, Y., Du, Y., Wang, P., Tang, X., Wang, H., & Kang, K. (2011). Metagenomic analysis of fever, thrombocytopenia and leukopenia syndrome (FTLS) in Henan Province, China: discovery of a new bunyavirus. PLoS pathogens, 7(11), e1002369.

Xu, C., Nikolova, O., Basom, R. S., Mitchell, R. M., Shaw, R., Moser, R. D., Park, H., Gurley, K. E., Kao, M. C., & Green, C. L. (2018). Functional Precision Medicine Identifies Novel

Druggable Targets and Therapeutic Options in Head and Neck CancerDruggable Target Discovery in Head and Neck Cancer. Clinical Cancer Research, 24(12), 2828-2843.

Yarden, R. I., Pardo-Reoyo, S., Sgagias, M., Cowan, K. H., & Brody, L. C. (2002). BRCA1 regulates the G2/M checkpoint by activating Chk1 kinase upon DNA damage. Nature genetics, 30(3), 285-289.

Yoo, M.-H., & Hatfield, D. L. (2008). The cancer stem cell theory: is it correct? Molecules & Cells (Springer Science & Business Media BV), 26(5).

Yoon, D.-S., Wersto, R. P., Zhou, W., Chrest, F. J., Garrett, E. S., Kwon, T. K., & Gabrielson, E. (2002). Variable levels of chromosomal instability and mitotic spindle checkpoint defects in breast cancer. The American journal of pathology, 161(2), 391-397.

Yu, G., Greninger, A. L., Isa, P., Phan, T. G., Martínez, M. A., de la Luz Sanchez, M., Contreras, J. F., Santos-Preciado, J. I., Parsonnet, J., & Miller, S. (2012). Discovery of a novel polyomavirus in acute diarrheal samples from children. PloS one, 7(11), e49449.

Yuan, T., & Cantley, L. (2008). PI3K pathway alterations in cancer: variations on a theme. Oncogene, 27(41), 5497-5510.

Yue, X., Zhao, Y., Zhang, C., Li, J., Liu, Z., Liu, J., & Hu, W. (2016). Leukemia inhibitory factor promotes EMT through STAT3-dependent miR-21 induction. Oncotarget, 7(4), 3777.

Zaharia, M., Bolosky, W. J., Curtis, K., Fox, A., Patterson, D., Shenker, S., Stoica, I., Karp, R. M., & Sittler, T. (2011). Faster and more accurate sequence alignment with SNAP. arXiv preprint arXiv:1111.5572.

Zhang, M., Jang, H., & Nussinov, R. (2020). PI3K inhibitors: review and new strategies. Chemical science, 11(23), 5855-5865.

Zhao, X., Setchell, K. D., Huang, R., Mallawaarachchi, I., Ehsan, L., Dobrzykowski Iii, E., Zhao, J., Syed, S., Ma, J. Z., & Iqbal, N. T. (2021). Bile acid profiling reveals distinct signatures in undernourished children with environmental enteric dysfunction. The Journal of Nutrition, 151(12), 3689-3700.

Zheng, C. L., Ratnakar, V., Gil, Y., & McWeeney, S. K. (2015). Use of semantic workflows to enhance transparency and reproducibility in clinical omics. Genome Medicine, 7, 1-13.

Zook, J. M., Chapman, B., Wang, J., Mittelman, D., Hofmann, O., Hide, W., & Salit, M. (2014). Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. Nature Biotechnology, 32(3), 246-251.